

Courant's Nodal Line Theorem and its Discrete Counterparts

by

Hongmei Zhu

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Applied Mathematics

Waterloo, Ontario, Canada, 2000

© Hongmei Zhu 2000



National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-51246-0

Canada

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

Courant's Nodal Line Theorem (CNLT) relates to the Dirichlet/Neumann eigenfunctions $u(\mathbf{x})$ of elliptic equations, the simplest and most important of which is the Helmholtz equation $\Delta u + \lambda \rho u = 0$ for $D \subset \mathbb{R}^m$. CNLT states that if the eigenvalues are ordered increasingly, the nodal set of the n th eigenfunction u_n divide D into *no more than n nodal domains* in which u_n has a fixed sign. We investigate whether the numerical solutions approximated by finite element method (FEM) retain this sign characteristic stated in CNLT. We derive various properties of the FEM solutions, then formulate and prove discrete analogues of CNLT for piecewise linear FEM solutions on a triangular/tetrahedral mesh.

For linear combinations of eigenfunctions, CNLT is replaced by Courant-Herrmann conjecture (CHC). CHC states that the nodal set of a combination $v = \sum_{i=1}^n c_i u_i$ also divides D into at most n nodal domains. We exhibit numerical counterexamples. We find that even linear combinations of the first two eigenfunctions can have three, four or more nodal domains. Also, we show that the discrete version of CHC is false in general. A restricted theorem is proved, which holds for both continuous and discrete cases. Although CHC is false in general, We conjecture that CHC is true for some convex domains, particularly for rectangles.

Acknowledgements

I would like to express my deepest gratitude to my supervisor Professor Graham M.L. Gladwell for suggesting the topic of this thesis and providing excellent guidance, invaluable encouragement and constant support throughout the entire course of my graduate studies. I would also like to thank his wife Joyce Gladwell for her friendship and encouragement over the years.

I wish to thank the thesis committee, Professors David Siegel, Moody Chu, Sivabal Sivaloganathan, and Jiahua Chen whose comments and suggestions greatly improved the quality of this thesis. I would specially thank Professor Siegel who has been available for many invaluable discussions.

Special thanks also go to the faculty and staff in the department of Applied Mathematics for providing a friendly and supportive research environment.

In the last few years, I have received the care and support of many individuals who have contributed to make my time in Waterloo enjoyable. I welcome this opportunity to extend my sincere thanks:

- To my husband Joe who has believed in me and supported me wholeheartedly. His love and understanding is an essential ingredient for the completion of this work.
- To my parents who have challenged me academically. Their unconditional love is always the source of my strength.
- To my extended families who have showered me with their love and caring.
- To all my friends who have touched my life in numerous ways.

I dedicate this thesis to the memory of my brother.

Hongwen Zhu

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Survey on the CNLT and CHC	6
2	Qualitative Properties of Eigenfunctions	12
2.1	The first three properties	12
2.2	Courant's Nodal Line Theorem	19
3	Qualitative Properties of FEM Solutions	25
3.1	Finite element counterpart	26
3.1.1	The constraints on finite element mesh	26
3.1.2	The necessity of K having the right signs	35
3.2	Graph theory notation and properties of FEM eigenvectors	38
3.2.1	Graph theory notation	38
3.2.2	Properties of FEM eigenvectors	42
3.3	Sign Graphs	46
3.4	Discrete CNLT: simple eigenvalues	50

3.5	Discrete CNLT: multiple eigenvalues	55
3.6	More on discrete analogues of CNLT	63
3.6.1	Discussion of Fiedler's result	63
3.6.2	Discussion of Duval and Reiner's result	71
4	Linear Combinations of Eigenfunctions	80
4.1	Nodal sets of the linear combinations	81
4.2	CHC is false in general	89
4.3	A revised CHC	91
4.4	CHC on square domains	100
4.4.1	The first fourteen eigenfunctions	102
4.4.2	Special linear combinations	111
5	Conclusions and Further Research	122
	Bibliography	124

List of Figures

3.1	An arbitrary element in a triangular mesh.	30
3.2	A tetrahedral finite element.	31
3.3	The angles between the outward normals are abused.	34
3.4	The triangulation of the rectangle where some obtuse-angled triangle appear.	36
3.5	The nodal lines of an approximated eigenfunction corresponding to λ_3 , divide the domain into six nodal domains. But $\lambda_3 = \lambda_4 = \lambda_5 <$ λ_6 . Thus, discrete CNLT fails.	38
3.6	Basic concepts in graph theory and matrix correspondences.	41
3.7	A FEM solution can have zero (shaded) polygons.	42
3.8	The graph $\mathcal{G}$ is associated with a pair of symmetric matrices $(\mathbf{K}, \mathbf{M})$. The subgraphs that are indicated with thick lines or vertices are sign graphs. There are three negative sign graphs and two positive sign graphs. Sign graph 2 is adjacent to sign graphs 1 and 3, but is disconnected from sign graphs 4 and 5.	47
3.9	The sign domains of the FEM continuous solution $u_I(\mathbf{x})$ constructed from vector $\mathbf{u}$ in the example shown in Figure 3.8.	49

3.10	Inconsistency of sign domains of $u(x, y)$ and the sign graphs of $\mathbf{u}$ occurs in rectangular mesh.	50
3.11	There exist four M -orthogonal eigenvectors associated with the multiple eigenvalue λ_3 , each of which has 6 sign graphs.	56
3.12	There exist four M -orthogonal eigenvectors $\{\mathbf{u}_j\}_3^6$ associated with the multiple eigenvalue λ_3 such that $SG(\mathbf{u}_j) \leq j$	57
3.13	There also exist four linearly independent eigenvectors $\{\mathbf{u}_j\}_3^6$ associated with the multiple eigenvalue λ_3 , each of which has 3 sign graphs.	58
3.14	A star with five vertices. There exists a set of linearly independent eigenvectors of $\mathbf{K}$ corresponding to λ_2 , each of which has three positive sign graphs.	68
3.15	For matrix pair $(\mathbf{K}, \mathbf{M})$ generated by FEM, there also exists a set of linearly independent eigenvectors corresponding to λ_2 , each of which has three positive sign graphs.	69
3.16	λ_3 is a 2-fold eigenvalue. An eigenvector $\mathbf{u}_3$ corresponding to λ_3 has four strict sign graphs, but all of them are adjacent to zero vertices only.	75
3.17	λ_4 is an eigenvalue with multiplicity 3. There is an eigenvector $\mathbf{u}_4$ corresponding to λ_4 such that every sign graph of $\mathbf{u}_4$ is adjacent to another sign graph. But $\mathbf{u}_4$ has six sign graphs.	76
4.1	The digram shows the contours of the combination $2 * \sin x \sin y + \sin x \sin 3y + \sin y \sin 3x$, where point $(\pi/2, \pi/2)$ is an isolated nodal point.	82

4.2	The two nodal lines of the linear combination w in (4.1) do not meet at right angles.	84
4.3	Nodal lines of a linear combination of eigenfunctions can have cusps.	85
4.4	The nodal lines of the linear combination $w = (u_{31} - u_{13}) + (u_{32} - u_{23}) + 2(u_{34} - u_{43})$ meet at a common nodal point of the eigenfunctions, but do not form an equiangular system.	88
4.5	The nodal lines of the linear combinations $t * u_1 + u_2$ for $t = 0, 0.96$ and 1.2. Linear combination of the first two eigenfunctions may have more than 2 nodal domains	90
4.6	The nodal lines of the linear combination $1.35 * u_1 + u_2$ divide the domain into four nodal domains: one positive and three negative.	92
4.7	The nodal lines of the linear combination $1.46 * u_1 + u_2$ divide the domain into five nodal domains: one positive and four negative.	93
4.8	The nodal lines of the linear combinations $t * u_1 + u_2$ for $t = 0, 0.5, 1$ and 1.3. Linear combination of the first two Neumann eigenfunctions can have more than 2 nodal domains.	94
4.9	The linear combination $0.7u_1 + u_2$ of the first eigenvectors has three sign graphs. CHC does not hold in discrete case either.	95
4.10	$\cos(\frac{8}{25}\pi)u_1 + u_2 + \sin(\frac{8}{25}\pi)u_3$ has three positive nodal domains.	98
4.11	$\cos(\frac{3}{10}\pi)u_1 + 0.525u_2 + \sin(\frac{3}{10}\pi)u_3$ has four positive nodal domains.	99
4.12	An ellipse can divide a square into 5 subregions.	104
4.13	The diagram shows the situation in which the most number of nodal domains can occur for each possible factorization.	106

4.14	An ellipse with a line can divide a square into 8 subdomains. . . .	107
4.15	Two ellipses can divide a square into 13 subdomains.	111
4.16	This table illustrates that CHC is true for the first N eigenfunctions on the square, $N = 1, 2, \dots, 14$. The second column gives all the new high order terms involved in the polynomial $P(X, Y)$ for each N , and the figures in the last column provides a linear combination that can produce the maximum number of nodal domains.	112
4.17	The shapes of G_n for $n = 2$ and $n = 3$	119

Chapter 1

Introduction

1.1 Motivation

There are two different views on discretization methods for the numerical solutions of boundary value problems: one focuses on the convergence analysis of the methods used, the other investigates how the numerical solutions reflect basic properties of the continuous solutions. Here, using the second point of view, we study the numerical solutions, approximated by *finite element method* (FEM), of second-order elliptic equations with Dirichlet boundary conditions. Such a problem is often related to a vibrating string, a vibrating membrane, etc.

For vibrating systems, one can uncover crucial information about a system from the places where nothing happens. (Hald and McLaughlin [16], Gladwell [13]). We call them “*nodal places*”. Because of their importance, the characteristics of nodal places have been studied for centuries. A number of interesting and important properties of the continuous solutions have been discovered. They include the unique

continuation property, the regularity of nodal places, the equiangular property and the sign property stated by Courant's Nodal Line Theorem (CNLT). It is of interest to investigate whether the FEM solutions also retain those properties of eigenfunctions.

More specifically, let us consider the Dirichlet eigenfunctions $u(\mathbf{x})$ of the Helmholtz equation

$$\Delta u + \lambda \rho u = 0, \quad \mathbf{x} \in \Omega, \quad (1.1)$$

the simplest and most important case of the self-adjoint second order elliptic equation. Here, Δ is the Laplacian operator, $\rho(\mathbf{x})$, the mass density, is positive and bounded, and Ω is a bounded connected domain in $\mathbb{R}^m$. In equation (1.1), the values of the parameter λ are the eigenvalues, and a nontrivial solution $u(\mathbf{x})$ is called an eigenfunction associated with the eigenvalue λ . The eigenfunctions are the spatial eigenmodes of a vibrating system in $\mathbb{R}^2$, and acoustic standing waves in $\mathbb{R}^3$.

The *nodal places* or *nodes* of an eigenfunction u are those points in the domain Ω at which u vanishes. The nodal set of u is denoted as $\mathcal{N}(u)$. It is well known that the nodes of the eigenfunctions of (1.1) satisfy the following properties:

- A. The unique continuation property (Jerison and Kenig [18], Müller [20]);
- B. If $\Omega \subset \mathbb{R}^m$ and $\rho \in C^\infty(\Omega)$, the nodal set $\mathcal{N}(u)$ is locally an $(m - 1)$ -dimensional manifold except for a closed set of dimensions less than $m - 1$ (Cheng [6]). In particular, the nodal places of an eigenfunction in two-

dimensions are curves; the nodal places of an eigenfunction in three-dimensions consist of surfaces; the nodal places of an eigenfunction in m -dimensions ($m \geq 4$) are a set of hypersurfaces;

Property B is stated for background information only, and is not going to be used in the analysis of Chapter 3 and 4. Also, ρ need not be analytic in our analysis.

- C. **Equiangular property (in two-dimensional case):** Suppose that ρ is analytic. If several nodal lines of u intersect at an interior point in Ω , then they form an equiangular system of rays (Cheng [6]).

Also, if Ω is a convex domain, ρ is analytic and the nodal lines intersects the boundary at a point, they also have the equiangular property (Alessandrini [1]);

- D. **Courant's Nodal Line Theorem (CNLT):** *If the eigenvalues are ordered increasingly, then the nodal places of the n -th eigenfunction divide the domain into at most n subdomains. We call those subdomains nodal domains, i.e., a nodal domain of u is a connected subdomains G of Ω such that u has a fixed sign in G and vanishes on the boundary of G . (Courant and Hilbert [7])*

As it is often difficult to get the exact solutions of a PDE, approximated solutions are often computed by FEM. In the FEM, the simplest implementation is to subdivide the region Ω into triangles in $\mathbb{R}^2$ or tetrahedra in $\mathbb{R}^3$, and to use piecewise continuous linear basis functions. FEM reduces equation (1.1) to a generalized

eigenvalue problem of the form

$$(\mathbf{K} - \lambda \mathbf{M})\mathbf{u} = \mathbf{0}. \quad (1.2)$$

where $\mathbf{K}$ is the *stiffness matrix*, $\mathbf{M}$ the *mass matrix*, λ an eigenvalue of $(\mathbf{K}, \mathbf{M})$ and $\mathbf{u} \neq \mathbf{0}$ an *eigenvector* corresponding to λ . The matrices $\mathbf{K}$ and $\mathbf{M}$ are both symmetric and positive definite. It is easy to show that $\mathbf{M}$ is nonnegative, i.e., $m_{ij} \geq 0$ for all $i, j = 1, \dots, n$; its diagonal entries are positive. $\mathbf{K}$ has positive diagonal entries, but the signs of its off-diagonal entries depend on the characteristic of the FEM mesh.

One of our aims in this thesis is to investigate if the approximated FEM solutions, corresponding to some refined or crude mesh in two or higher dimensions, have properties analogous to those of the continuous solutions. Especially, we are interested in whether FEM solutions have the discrete analogue of CNLT. This is motivated by existing results in one dimension for a spring-mass vibrating system with either fixed or free ends. The eigenmodes of such a system have the following property: the k -th eigenmode divides the system into *exactly* k parts by its nodes (Gantmakher [11]). This property provides the essential condition for the reconstruction of a spring-mass system given one or more eigenmodes (Gladwell [13]). The study of the nodes of eigenvectors, the discretized eigenfunctions, can be thus of great relevance in inverse eigenproblems. Our main results extend the one-dimensional results of Gantmakher's to higher dimensional discretized eigenproblems.

In a footnote on page 454, Courant and Hilbert [7] states that Herrmann proved

the following in his 1932 Göttingen dissertation:

Courant-Herrmann Conjecture (CHC): *Any linear combination of the first n eigenfunctions of (1.1) has at most n nodal domains.*

However there is no such proof in his dissertation nor in his later publications. In fact, Arnol'd [2] was the first to notice that CHC is false, although he did not present a counterexample. We exhibit interesting numerical counterexamples, computed by MATLAB PDE Toolbox. We find that even linear combinations of the first two eigenfunctions can have three, four or five nodal domains. We conjecture that combinations of the first two eigenfunctions can have arbitrarily many nodal domains for some non-convex domains. Also, we show that the discrete version of CHC is false in general.

We conjecture that CHC may be true for some simple domains, especially rectangular domains. We examine CHC for a square and establish it for combinations $w = \sum_{i=1}^n c_i u_i$ for $n \leq 14$. In addition, for certain types of linear combinations, we prove that CHC holds for square domain. We conjecture that CHC is true for rectangular domains.

The thesis is organized as follows. In § 1.2, we survey various results on CNLT and CHC. Proofs and connections between property A-D of the nodal places in equation (1.1) are explored in Chapter 2. In Chapter 3, qualitative properties of FEM solutions of (1.2) are studied. Among them, we show that a straight forward analogue of the unique continuation property for an FEM solution does not hold, in the sense that an FEM solution can be zero in one or more complete elements without vanishing identically. We then formulate and prove a discrete

CNLT for piecewise linear finite element discretization on a triangular/tetrahedral mesh. The discussion of CHC and discrete CHC is provided in Chapter 4. The investigation reveals that a linear combination of the eigenfunctions behaves quite differently from a single eigenfunction. The unique continuation property still holds for combinations of eigenfunctions while other properties do not. In Chapter 5, we summarize our results and state some possible extensions to our current research.

1.2 Survey on the CNLT and CHC

For one-dimensional systems, the Helmholtz equation is known as the Sturm-Liouville equation. The CNLT is now replaced by a stronger result, see Courant and Hilbert ([7], p.454)

Theorem 1.1 (Sturm) *The zeros of the n -th eigenfunction divide the domain into exactly n nodal intervals.*

In the one-dimensional case, the CHC is replaced by

Theorem 1.2 *The number S of nodal intervals of a linear combination of the eigenfunctions $u_p, u_{p+1}, \dots, u_q$ satisfies*

$$p \leq S \leq q.$$

This theorem was conjectured by Sturm, but proved by Liouville and Rayleigh. The CHC is a weak analogue of Theorem 1.2 for the special case $p=1$. See also Gantmakher and Krein [12] and Gladwell [13].

The matrix analogues of CNLT and CHC in one dimension are based on the theory of oscillatory matrices. Following Berman and Plemmons [3], we say that $\mathbf{A} \in \mathbb{R}^{N \times N}$, not necessarily symmetric, is *oscillatory* if $\mathbf{A}$ is nonsingular, all the minors of $\mathbf{A}$ are nonnegative and $a_{i,i+1} > 0$, $a_{i+1,i} > 0$ for $i = 1, 2, \dots, N - 1$.

Two of the fundamental theorems of matrix algebra, used here, are named after Perron and Frobenius, and Binet and Cauchy.

Theorem 1.3 (Perron-Frobenius) *(see Berman and Plemmons [3]) Let $\mathbf{A}_{N \times N}$ be a non-negative matrix and $\rho(\mathbf{A})$ be the greatest eigenvalue of $\mathbf{A}$. Then there is a non-negative eigenvector $\mathbf{u}_\rho$ corresponding to $\rho(\mathbf{A})$. $\mathbf{A}^T$ also has a nonnegative eigenvector corresponding to $\rho(\mathbf{A})$.*

Furthermore, if $\mathbf{A}$ is irreducible, then $\rho(\mathbf{A})$ is simple and $\mathbf{u}_\rho$ is the unique positive eigenvector of $\mathbf{A}$ apart from scalar multiples.

A matrix $\mathbf{A}$ is *reducible* if for some permutation matrix $\mathbf{P}$ such that

$$\mathbf{P}^T \mathbf{A} \mathbf{P} = \begin{pmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{C} & \mathbf{D} \end{pmatrix},$$

where $\mathbf{B}$ and $\mathbf{D}$ are square matrices. (When $N = 1$, $\mathbf{A}$ is reducible iff $\mathbf{A} = 0$.)

Otherwise, $\mathbf{A}$ is *irreducible* (Berman and Plemmons [3]).

The Binet-Cauchy theorem relates to *compound* matrices. Suppose $\mathbf{A} \in \mathbb{R}^{M \times N}$, $p \leq \min(M, N)$, $S = \binom{M}{p}$, $T = \binom{N}{p}$, then $\mathbf{A}_p \in \mathbb{R}^{S \times T}$ is the matrix composed of the p -th order minors of $\mathbf{A}$ arranged in lexical order.

Theorem 1.4 (Binet-Cauchy) (see Gantmakher [11]) *If $\mathbf{A} \in \mathbb{R}^{M \times N}$, $\mathbf{B} \in \mathbb{R}^{N \times q}$, $p \leq \min(M, N, q)$ and $\mathbf{AB} = \mathbf{C}$, then*

$$\mathbf{A}_p \mathbf{B}_p = \mathbf{C}_p.$$

A consequence of this theorem is

Theorem 1.5 *If $\mathbf{A} \in \mathbb{R}^{N \times N}$ has eigenvalues $(\lambda_i)_1^N$ and eigenvectors $(\mathbf{x}^{(i)})_1^N$, $\mathbf{X} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}]$, $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_N)$ and $\mathbf{AX} = \mathbf{X}\mathbf{\Lambda}$, then*

$$\mathbf{A}_p \mathbf{X}_p = \mathbf{X}_p \mathbf{\Lambda}_p.$$

This means that the eigenvalues of $\mathbf{A}_p$ are the products of p eigenvalues of $\mathbf{A}$.

From these theorems Gantmakher [11] proved

Theorem 1.6 *Let $\mathbf{A} = (a_{ij})_{N \times N}$ be an oscillatory matrix. $\mathbf{A}$ has distinct eigenvalues, which we order so that:*

$$\lambda_1 > \lambda_2 > \dots > \lambda_N > 0.$$

The k -th eigenvector $\mathbf{u}_k = (u_{k1}, u_{k2}, \dots, u_{kN})$ has exactly $k - 1$ sign changes in the sequence of its coordinates.

Moreover, let $\mathbf{u}$ be any linear combination of the eigenvector $\mathbf{u}_p, \mathbf{u}_{p+1}, \dots, \mathbf{u}_q$, then the number of sign changes in the sequence of the coordinates of $\mathbf{u}$ is

$$p - 1 \leq S_{\mathbf{u}}^- \leq S_{\mathbf{u}}^+ \leq q - 1.$$

Note that since we can give arbitrary signs to the zero coordinates of $\mathbf{u}$, $S_{\mathbf{u}}^-$ and $S_{\mathbf{u}}^+$ denote the *minimum* and *maximum* numbers of sign changes in the sequence of the coordinates of $\mathbf{u}$.

With this result, we can study the behavior of the discrete eigenmodes of a vibrating string. In the FEM discretization with a linear interpolation, the eigenvector $\mathbf{u}$ should satisfy equation (1.2) for some eigenvalue λ . $\mathbf{K}$ and $\mathbf{M}$ are positive definite tridiagonal matrices; $\mathbf{K}$ has negative quasi-diagonal and $\mathbf{M}$ has positive quasi-diagonal. Equation (1.2) may be reduced to the standard form

$$(\mathbf{A} - \mu \mathbf{I})\mathbf{u} = \mathbf{0} \quad (1.3)$$

where

$$\mathbf{A} = \mathbf{K}^{-1}\mathbf{M}, \quad \mu = \frac{1}{\lambda}.$$

It can be shown that $\mathbf{K}^{-1}$ and $\mathbf{M}$ are oscillatory. A product of two oscillatory matrices is oscillatory; thus $\mathbf{A}$ is oscillatory. Theorem 1.6 holds for the eigenvectors of $\mathbf{A}$. Thus the matrix analogues of CNLT and CHC hold for the FEM model in the one-dimensional case.

Gantmakher's theorem on oscillatory matrices cannot be applied to FEM eigenvectors in higher dimensions. In such cases the matrices are no longer tridiagonal and the eigenvalues are not necessarily distinct; so $\mathbf{K}^{-1}\mathbf{M}$ is not oscillatory. However, we can show that the eigenvectors do obey a discrete counterpart of CNLT if $\mathbf{K}$ is an M-matrix. A matrix $\mathbf{A}$ is called a (non-singular) *M-matrix* if it is positive

definite with $a_{ii} > 0$ and $a_{ij} \leq 0$ for $i \neq j$. (Berman and Plemmons [3]). If there are some positive off-diagonal entries of $\mathbf{K}$, then the discrete CNLT may not be true. To formulate discrete counterparts of CNLT, the concept of nodal domains, which appears in the continuous case, is replaced by that of *sign graphs* or *strict sign graphs* (see §3.3).

The inverse of a nonsingular $\mathbf{M}$ -matrix is nonnegative. Thus, $\mathbf{A} = \mathbf{K}^{-1}\mathbf{M}$ is non-negative. When a FEM mesh is connected, $\mathbf{K}$ and $\mathbf{M}$ are irreducible; so is $\mathbf{A}$. The Perron-Frobenius Theorem states that the *highest* eigenvalue of $\mathbf{A}$, and hence the *lowest* eigenvalue of $(\mathbf{K} - \lambda\mathbf{M})\mathbf{u} = \mathbf{0}$ is simple, and the corresponding eigenvector is *positive*. Hence, $\mathbf{u}_1$ has one sign graph and any eigenvector of $(\mathbf{K}, \mathbf{M})$ corresponding to higher eigenvalue must have more than one sign graph.

Fiedler [10], Duval and Reiner [8] studied CNLT for eigenvectors of a real symmetric matrix with non-positive off-diagonal elements. We can generalize their results to eigenvectors of matrix pair $(\mathbf{K}, \mathbf{M})$. Fiedler [10] proved that the n th eigenvector has no more than $n - 1$ non-negative sign graphs. Non-negative or non-positive sign graphs are termed as *loose sign graphs*. The only certain conclusion we can draw from that statement is that the n th eigenvector has no more than $2n - 2$ loose sign graphs, which is loose compared to the upper bound n in CNLT. Gladwell [14] recently reduced the upper bound $2n - 2$ to n . He proved that the n -th eigenvector has at most n loose sign graphs. Regarding the global upper bound for the number of strict sign graphs, Duval and Reiner proved a much stronger result than Fiedler: when λ_n is distinct, the n th eigenvector has no more than n sign graphs. But their result does not hold for multiple eigenvalues. For multiple

eigenvalues, there is a difference between the continuous case and the discrete case.

Let λ_n be a r -fold eigenvalue such that

$$\lambda_{n-1} < \lambda_n = \lambda_{n+1} = \cdots = \lambda_{n+r-1} < \lambda_{n+r},$$

CNLT implies that any eigenfunction corresponding to λ_n has no more than n nodal domains, the least upper bound. However, counterexamples in Chapter 3 show that an eigenvector corresponding to λ_n may have more than n sign graphs. This fundamental difference makes the multiple eigenvalue case complicated.

We will make further comparison with the results of Fiedler, and Duval and Reiner in Chapter 3.

Chapter 2

Qualitative Properties of Eigenfunctions

In Chapter 1, we listed four basic properties that describe the behavior of the nodal places of the eigenfunctions. In this chapter, we examine these properties and discuss their relations. We will indicate in later chapters how these properties differ from those of discrete solutions and finite linear combinations of eigenfunctions.

2.1 The first three properties

Jerison and Kenig [18] proved that

A. The unique continuation property: *Suppose that $\Omega \subset \mathbb{R}^m$ is open and connected, $c \in L_{loc}^{m/2}(\Omega)$ and $q = 2m/(m + 2)$. If any solution $u \in H_{loc}^{2,q}(\Omega)$ of $\Delta u + c(\mathbf{x})u = 0$ vanishes on a non-empty open subset of Ω , then $u \equiv 0$ in Ω .*

Since $\rho(\mathbf{x})$ is bounded, the unique continuation property holds for the solutions of

(1.1) in $H_{loc}^{2,q}(\Omega)$. In particular, if u is a classical solution of (1.1), i.e., $u \in C^2(\Omega)$, then u has the unique continuation property (see also Müller [20]). If $\rho \equiv 1$ or ρ is analytic in Ω , then a solution to (1.1) is analytic so that the unique continuation property follows from the following result for analytic functions.

Proposition *If u is analytic in Ω and u vanishes on an open subset of Ω then $u \equiv 0$ on Ω .*

The proof follows the basic idea of John [19].

Proof: Let M denote the interior of the set $\{\mathbf{x} \in \Omega : u(\mathbf{x}) = 0\}$. By hypothesis, $M \neq \emptyset$. Then M is an open subset of Ω . All the derivatives of u vanish identically on M . This implies that M is closed in Ω : Let a Cauchy sequence $\{\mathbf{x}_k\}$ in M converge to $\mathbf{x}^*$ as $k \rightarrow \infty$. By the continuity of the derivatives, all derivatives of u vanish at $\mathbf{x}^*$. That is, the power series of u at $\mathbf{x}^*$ is identically zero; hence $\mathbf{x}^* \in M$. Since M is non-empty, $M = \Omega$, i.e., u is identically zero in Ω , a contradiction to u being an eigenfunction. ■

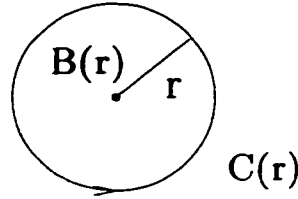
The unique continuation property says that an eigenfunction cannot vanish on any non-empty open subset of Ω . Hence, the nodal set $\mathcal{N}$ of an eigenfunction must belong to a lower dimensional space.

B. *If $\Omega \subset \mathbb{R}^m$ and $\rho \in C^\infty(\Omega)$, the nodal set $\mathcal{N}(u)$ is locally an $(m-1)$ -dimensional manifold except for a closed set of dimensions less than $m-1$ (Cheng [6]).*

This follows from the maximum principle and the unique continuation property.

Theorem 2.1 (Maximum Principle) *Let $\Delta u \geq 0$ (or $\Delta u \leq 0$) in a domain D . If u attains a maximum (or a minimum) at an interior point of D , then u is constant in D .*

Proof: [Protter and Weinberger [22]] Suppose that u attains its maximum M at some interior point $\mathbf{p}_1 \in D$ and u is not identically equal to M in D , say $u(\mathbf{p}_2) < M$ where $\mathbf{p}_2 \in D$. Connect $\mathbf{p}_1$ and $\mathbf{p}_2$ by a curve in D . Let $\mathbf{p}$ be the first point along the curve such that $u(\mathbf{p}_2) = M$. u is not identically equal to M on any sufficiently small circle centered at $\mathbf{p}$. Let $B(r)$ be a neighborhood of $\mathbf{p}$ in D with radius r . Denote the boundary of $B(r)$ by $C(r)$.



Applying the divergence theorem to Δu gives

$$\int_{B(r)} \Delta u d\mathbf{x} = \int_{B(r)} \text{div}(\nabla u) d\mathbf{x} = \oint_{C(r)} \frac{\partial u}{\partial \mathbf{n}} dS,$$

where $\partial u / \partial \mathbf{n}$ is the normal derivative taken on the boundary $C(r)$. Transforming the coordinates into polar coordinates, we have $dS = r^{m-1} \sin^{m-2} \theta_1 \sin^{m-3} \theta_2 \cdots \sin \theta_{m-2} d\theta$ and

$$\int_{B(r)} \Delta u d\mathbf{x} = r^{m-1} \int \frac{\partial u}{\partial r} \sin^{m-2} \theta_1 \cdots \sin \theta_{m-2} d\theta,$$

in which $\theta = (\theta_1, \dots, \theta_{m-2})$. It follows from $\Delta u \geq 0$ in D that

$$\int \frac{\partial u}{\partial r} \sin^{m-2} \theta_1 \cdots \sin \theta_{m-2} d\theta \geq 0. \quad (2.1)$$

Fix a radius R sufficiently small so that $B(R) \subset D$ and u is not identically equal to M on $C(R)$. Integrating (2.1) from 0 to R and interchanging the order of integration, we obtain

$$\begin{aligned} & \int \int_0^R \frac{\partial u}{\partial r} \sin^{m-2} \theta_1 \cdots \sin \theta_{m-2} dr d\theta \\ &= \int u(R, \theta) \sin^{m-2} \theta_1 \cdots \sin \theta_{m-2} d\theta - \omega_m u(\mathbf{p}) \geq 0, \end{aligned}$$

where ω_m is a positive constant, depending on m only. (Note that $\omega_2 = 2\pi$ and $\omega_3 = 4\pi$). That is,

$$\begin{aligned} u(\mathbf{p}) &\leq \frac{1}{\omega_n R^{m-1}} \int u(R, \theta) R^{m-1} \sin^{m-2} \theta_1 \cdots \sin \theta_{m-2} d\theta \\ &= \frac{1}{\omega_n R^{m-1}} \oint_{C(R)} u dS. \end{aligned}$$

Since $u \leq M$ and is not identically equal to M on $C(R)$, the average of u over $C(R)$ is less than M . We have

$$M = u(\mathbf{p}) \leq \frac{1}{\omega_n R^{m-1}} \oint_{C(R)} u dS < M,$$

which is a contradiction. If u attains a minimum in Ω , we can apply the same argument to $-u$. ■

We are not going to go through the proof of property B, but will state and prove property B' which is closely related to property B.

B' . *For every point $\mathbf{p} \in \mathcal{N}$ and every $\epsilon > 0$ then there exist at least two points $\mathbf{p}_1, \mathbf{p}_2 \in B(\mathbf{p}, \epsilon)$ so that $u(\mathbf{p}_1)$ and $u(\mathbf{p}_2)$ have opposite signs.*

Proof: Suppose that this is not true, i.e., there exists a nodal point $\mathbf{p}$ and an open ball $B(\mathbf{p})$ in Ω such that u has a fixed sign in $B(\mathbf{p})$. Without loss of generality, let u be non-positive in $B(\mathbf{p})$. Hence $\Delta u = -\lambda \rho u \geq 0$ in $B(\mathbf{p})$ and u attains its maximum, 0, in $B(\mathbf{p})$. By the maximum principle, $u \equiv 0$ in $B(\mathbf{p})$. The unique continuation property then shows that $u \equiv 0$ in Ω , which contradicts the statement that u is an eigenfunction. ■

Thus, those $(m - 1)$ -dimensional hypersurfaces are either closed, or begin and end at the boundary.

For $m = 2$, an eigenfunction cannot have an isolated nodal point; its nodal set consists of continuous nodal lines, which are either closed, or begin and end at the boundary. It is then of interest to know what happens when several nodal lines intersect at a point? Property C answers this question.

C. Equiangular property (in two dimensional case): *Let ρ be analytic. If several nodal lines of u intersect at an interior point in Ω , then they form an equiangular system of rays.*

Also, if Ω is a convex domain, ρ is analytic and the intercept of the nodal lines is on the boundary, they also have the equiangular property.

Property C states that the nodes of an eigenfunction behave locally as the nodes of a harmonic function. We call this the equiangle behaviour of nodal lines. We first prove it when the intercept of the nodal lines is an interior point.

Proof: Suppose that there are m nodal lines of an eigenfunction u intersecting at a point $\mathbf{p} = (0, 0)$. Expand u in a power series at some neighborhood of $\mathbf{p}$

$$u = \sum_{j=0}^{\infty} v_j,$$

where

$$v_j = \frac{1}{j!} \left(x \frac{\partial}{\partial x} + y \frac{\partial}{\partial y} \right)^j (u(\mathbf{p})).$$

By assumption, $v_0 = v_1 = \dots = v_{m-1} = 0$ in $B_{\mathbf{p}}$. That is,

$$u = \sum_{j=m}^{\infty} v_j. \tag{2.2}$$

Substituting (2.2) into equation (1.1) gives

$$\sum_{j=m}^{\infty} \Delta v_j + \lambda \sum_{j=m}^{\infty} v_j = 0,$$

where the degree of the polynomial v_j is j and the degree of Δv_j is $j - 2$. It then gives

$$\Delta v_m = -\lambda v_{m-2} = 0$$

in $B(p)$, i.e., v_m behaves like a harmonic function near p . Thus we have

$$\begin{aligned} v_m &= \frac{1}{2}[A\operatorname{Re}(z^m) + B\operatorname{Im}(z^m)] \\ &= \frac{1}{2}[Ar^m\cos(m\theta) + Br^m\sin(m\theta)] \\ &= \frac{1}{2}r^m\sqrt{A^2 + B^2}\sin(m(\theta - \beta)). \end{aligned}$$

for some constants A , B and β . This leads to

$$u = \tilde{A}r^m\sin(m(\theta - \beta)) + o(r^m). \quad (2.3)$$

Equating (2.3) to zero gives

$$\sin(m(\theta - \beta)) = 0,$$

the solutions of which are

$$\theta_k = \beta + \frac{k\pi}{m}, \quad k = 1, 2, \dots, 2m.$$

Intuitively, the tangents of the nodal lines of u at p form an equiangular system of rays. For further details, see Cheng [6]. ■

We can prove the boundary equiangle behaviour of eigenfunctions in a similar fashion, see Alessandrini [1]. The basic idea is to use conformal mappings to transform the sector $\Gamma(p) = \Omega \cap B_p$ into a half disk, and perform an odd reflection of the eigenfunction across the boundary; then taking the advantage of the conformal

invariance of elliptic equations in divergence form, we transform the problem of boundary equiangle behaviour into interior equiangle behaviour.

Property C is a local property of nodal set of an eigenfunction. In the next section, we will focus on a global property, namely, Courant's Nodal Line Theorem.

2.2 Courant's Nodal Line Theorem

The nodal hypersurfaces of an eigenfunction divide the domain into a number of subdomains, called nodal domains. CNLT gives a global upper bound for the number of nodal domains of an eigenfunction.

D. (CNLT) *If the eigenvalues are ordered increasingly, then the n -th eigenfunction has at most n nodal domains.*

Proofs of CNLT can be found in Courant and Hilbert [7], Herrmann [17] or Pleijel [21], etc. Among them, Herrmann or Pleijel's proof is simpler, and will be given here. The proof is essentially based on two tools: the variational characterization of eigenvalues, and the unique continuation property.

It is well known (Evans [9]) that (1.1) has infinitely many positive eigenvalues

$$0 < \lambda_1 < \lambda_2 \leq \lambda_3 \leq \dots .$$

The corresponding eigenfunctions $\{u_i\}_1^\infty$, where the u_i are orthonormal, form a complete set of $L^2(\Omega)$. Hence for any function $f \in H_0^1(\Omega)$, we can write

$$f = \sum_{i=1}^{\infty} c_i u_i.$$

which converges in $L^2(\Omega)$. We say that $u \in H_0^1(\Omega)$ is a weak solution of (1.1) if

$$\int_{\Omega} -\nabla u \cdot \nabla v + \lambda \rho u v d\mathbf{x} = 0, \quad (2.4)$$

for all $v \in H_0^1(\Omega)$.

The eigenvalues can be characterized recursively.

Theorem 2.2 (Minimax principle) *Suppose the eigenvalues of (1.1) are ordered increasingly. Then*

$$\lambda_k = \max_{\dim(S)=k-1} \min_{u \in S^\perp, u \neq 0} \frac{\int_{\Omega} \nabla u \cdot \nabla u d\mathbf{x}}{\int_{\Omega} \rho u^2 d\mathbf{x}}, \quad (2.5)$$

where S is any $(k-1)$ -dimensional subspace of $H_0^1(\Omega)$ and $S^\perp$ is the orthogonal space of S .

The ratio of the quadratic form on the right hand side of (2.5) is called the *Rayleigh quotient* of function u , denoted as $R(u)$. Define $(u, v) = \int_{\Omega} \rho u v d\mathbf{x}$. u and v are orthogonal iff $(u, v) = 0$.

Proof: We shall first prove that for an arbitrarily $(k-1)$ -dimensional subspace S of $H_0^1(\Omega)$,

$$\min_{u \in S^\perp, u \neq 0} R(u) \leq \lambda_k.$$

Let $\{v_i\}_1^{k-1}$ be an orthonormal basis for S . We can find a function $u = \sum_{i=1}^k c_i u_i$,

where the c_i can be determined so that

$$\begin{cases} (u, v_i) = 0, & \text{for } i = 1, \dots, k-1 \\ (u, u) = \sum_{i=1}^k c_i^2 = 1. \end{cases}$$

Hence, $u \in S^\perp$. Since the u_i satisfy (2.4),

$$\begin{aligned} \int_{\Omega} \nabla u \cdot \nabla u dx &= \sum_{i=1}^k c_i \int_{\Omega} \nabla u_i \cdot \nabla u dx \\ &= \sum_{i=1}^k c_i \lambda_i \int_{\Omega} \rho u_i u dx \\ &= \sum_{i=1}^k \lambda_i c_i^2 \int_{\Omega} \rho u_i^2 dx \\ &= \sum_{i=1}^k \lambda_i c_i^2 \\ &\leq \lambda_k. \end{aligned}$$

That is, $R(u) \leq \lambda_k$ which implies that

$$\min_{u \in S^\perp, u \neq 0} R(u) \leq \lambda_k. \quad (2.6)$$

On the other hand, take S_1 to be the $(k-1)$ -dimensional space spanned by eigenfunctions $\{u_i\}_1^{k-1}$. For any $u \in S_1^\perp$, u can be written as $u = \sum_{i=k}^\infty c_i u_i$. Calculate the Rayleigh quotient similarly as before

$$R(u) = \frac{\sum_{i=k}^\infty \lambda_i c_i^2}{\sum_{i=k}^\infty c_i^2} \geq \lambda_k.$$

With (2.6), this shows that

$$\min_{u \in S_1^\perp, u \neq 0} R(u) = \lambda_k,$$

which proves the result. ■

In particular, the equality holds iff u is an eigenfunction of λ_k .

Proof: [CNLT]: Assume that the n -th eigenfunction u_n divides Ω into m regions.

say $\{\Omega_i\}_1^m$ ($\cup_{i=1}^m \Omega_i = \Omega$). Define a sequence of functions $\{v_i\}_1^m$ such that

$$v_i(\mathbf{x}) = \begin{cases} u_n(\mathbf{x}), & \mathbf{x} \in \Omega_i \\ 0, & \text{otherwise} . \end{cases}$$

We note that each v_i satisfies the equation (2.4) with $\lambda = \lambda_n$ in Ω_i and $(v_i, v_j) = 0$ for $i \neq j$.

If the eigenvalue λ_n is simple, take a function

$$u(\mathbf{x}) = \sum_{i=1}^m c_i v_i(\mathbf{x}).$$

Choose the c_i so that $(u, u_j) = 0$ for $j = 1, 2, \dots, m-1$. Calculate

$$\begin{aligned} \int_{\Omega} \nabla u \cdot \nabla u d\mathbf{x} &= \sum_{i=1}^m c_i^2 \int_{\Omega_i} \nabla v_i \cdot \nabla v_i d\mathbf{x} \\ &= \sum_{i=1}^m c_i^2 \int_{\Omega_i} \lambda_n \rho v_i^2 d\mathbf{x} \\ &= \lambda_n \int_{\Omega} \rho u^2 d\mathbf{x}. \end{aligned}$$

Thus, $R(u) = \lambda_n$. By the recursive characterization, we have $\lambda_m \leq R(u) = \lambda_n$. Since $\lambda_n < \lambda_{n+1}$, we have $\lambda_m < \lambda_{n+1}$, i.e., $m < n + 1$ or $m \leq n$. This proves the assertion.

However, when λ_n is a r -fold eigenvalue, so that

$$\lambda_{n-1} < \lambda_n = \lambda_{n+1} = \cdots = \lambda_{n+r-1} < \lambda_{n+r}, \quad (2.7)$$

the above analysis gives that $\lambda_m \leq \lambda_n < \lambda_{n+r}$ so that $m < n + r$ or $m \leq n + r - 1$. Hence any eigenfunctions corresponding to λ_n has at most $n + r - 1$ nodal domains; a different treatment is required in order to reduce this to n . We take a function

$$u(\mathbf{x}) = \sum_{i=1}^{m-1} c_i v_i(\mathbf{x}).$$

It is zero on nodal domain Ω_m . Choose the coefficients c_i so that $(u, u_j) = 0$ for $j = 1, 2, \dots, m - 2$. Again, we can conclude that $\lambda_{m-1} \leq R(u) = \lambda_n$. If $\lambda_{m-1} = \lambda_n$, by the variational characterization, $u \in H_0^1(\Omega)$ must be a weak solution of the differential equation. By interior H^2 -regularity (Evans [9]), $u \in H_{loc}^{2,2}(\Omega) \subset H_{loc}^{2,q}(\Omega)$ where $q = 2m/(m + 2) < 2$. In addition, ρ is bounded which implies that $\rho \in L_{loc}^{m/2}(\Omega)$. Therefore, u has the unique continuation property. Since $u \equiv 0$ in Ω_m , u vanishes identically in Ω . This contradiction implies that $\lambda_{m-1} < \lambda_n$, i.e. $m \leq n$. ■

Notice that the unique continuation property is used in the proof. However, we will show later that the analogue of the unique continuation property does not hold for the eigenvectors in the discrete case. Therefore, we cannot simply apply the same argument to derive a matrix analogue of CNLT.

Another remarkable difference between continuous and discrete versions of CNLT appears when λ_n is an eigenfunction of multiplicity r as in (2.7), CNLT states that any eigenfunction associated with λ_n has no more than n nodal domains. Examples shown later indicate that this is false in the discrete case. Duval and Reiner [8] failed to notice this difference.

Chapter 3

Qualitative Properties of FEM Solutions

Applying the FEM procedure to the eigenvalue problem (1.1) gives the matrix eigenvalue problem

$$(\mathbf{K} - \lambda \mathbf{M})\mathbf{u} = \mathbf{0}$$

where both $\mathbf{K}$ and $\mathbf{M}$ are symmetric and positive definite. The off-diagonal entries of $\mathbf{M}$ are non-negative, *i.e.*, $m_{ij} \geq 0$ for $i \neq j$. The sign of the off-diagonal elements of $\mathbf{K}$ depend on the mesh.

For a triangular mesh in $\mathbb{R}^2$ or a tetrahedral mesh in $\mathbb{R}^3$, we find conditions on the finite elements so that $\mathbf{K}$ has non-positive off-diagonal entries, *i.e.*, $\mathbf{K}$ is an M-matrix. In this case, we show that a discrete counterpart of CNLT holds for the FEM solutions. Otherwise, it is easy to construct a counterexample of a mesh with

some obtuse angled triangles, for which the discrete CNLT fails.

Then we explore the behaviour of the eigenvectors of such a pair of matrices $(\mathbf{K}, \mathbf{M})$, which helps us to formulate the discrete CNLT properly. The definitions of sign graphs are then formally introduced. As in the continuous case, there are two parts of the discrete CNLT: when eigenvalues are simple, and when some are multiple. The latter case requires different treatment due to the differences between continuous and discrete solutions. A discussion and a comparison with Fiedler's and Duval and Reiner's results are given in § 3.6.

3.1 Finite element counterpart

3.1.1 The constraints on finite element mesh

The FEM method in our discussion is a Rayleigh-Ritz method with piecewise linear basis functions. In the FEM procedure, we first subdivide Ω into regular shaped elements. The simplest case is a triangular mesh in $\mathbb{R}^2$ or a tetrahedral mesh in $\mathbb{R}^3$. The collection of the finite elements is denoted by D , and the mesh points of these elements are called *vertices*. There are three kinds of vertices. The ones on the boundary are called *boundary vertices*. Vertices adjacent to boundary vertices are called *near-boundary vertices*. The rest are defined as *interior vertices*. The interior vertices belong only to the elements totally in the interior of D . With each vertex i , we associate a basis function $f_i(\mathbf{x})$, which is non-zero and linear in the elements that contain vertex i ; f_i is one at i and zero at other vertices. The FEM

method seeks an approximation u_I with the form

$$u_I(\mathbf{x}) = \sum_{i=1}^M u_i f_i(\mathbf{x}).$$

The solution u_I takes the value u_i at vertex i ; in particular, $u_i = 0$ for $i = N + 1, \dots, M$, because of the boundary condition.

Recall that the eigenvalues of the Helmholtz Equation are also the stationary points of the Rayleigh quotient $R(u)$

$$\lambda = \frac{\int_{\Omega} \nabla u \cdot \nabla u d\Omega}{\int_{\Omega} \rho u^2 d\Omega}. \quad (3.1)$$

Applying the FEM procedure to (1.1) gives the generalized eigenvalue problem

$$(\mathbf{K} - \lambda \mathbf{M})\mathbf{u} = \mathbf{0}, \quad (3.2)$$

where

$$\int_{\Omega} \nabla u \cdot \nabla u d\Omega \simeq \int_D \nabla u_I \cdot \nabla u_I d\Omega = \mathbf{u}^T \mathbf{K} \mathbf{u} \quad (3.3)$$

and

$$\int_{\Omega} \rho u^2 d\Omega \simeq \int_D \rho u_I^2 d\Omega = \mathbf{u}^T \mathbf{M} \mathbf{u}. \quad (3.4)$$

The global matrices $\mathbf{K}$ and $\mathbf{M}$ are generated by assembling the entries of the element matrices $\mathbf{K}_e$ and $\mathbf{M}_e$ on each finite element. Note that indices of $\mathbf{u}$ include all the

vertices except the boundary ones.

Next we derive the sufficient conditions on the finite element mesh so that $\mathbf{K}_e$ has non-positive off-diagonals.

Theorem 3.1 *In the triangulation of the domain, if every triangle is acute angled, then the local stiffness matrix $\mathbf{K}_e$ has non-positive off-diagonal entries.*

Proof: In an arbitrary triangle $\Delta P_1 P_2 P_3$ as shown in Figure 3.1, the FEM solution u takes the linear form

$$u = a + bx + cy.$$

This leads to

$$\nabla u \cdot \nabla u = b^2 + c^2. \quad (3.5)$$

u takes the values u_i at the vertex P_i for $i = 1, 2, 3$, i.e.,

$$u_i = a + bx_i + cy_i, \quad i = 1, 2, 3.$$

Solving the above linear system, we have

$$b\Delta = u_1(y_2 - y_3) + u_2(y_3 - y_1) + u_3(y_1 - y_2) \quad (3.6)$$

and

$$c\Delta = u_1(x_2 - x_3) + u_2(x_3 - x_1) + u_3(x_1 - x_2) \quad (3.7)$$

where Δ is the determinant of

$$\begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{pmatrix}.$$

Notice that $|\Delta| = 2 * \text{Area}(P_1P_2P_3)$. Substituting (3.6) and (3.7) to (3.5), we find that the coefficient of u_1u_2 in

$$\int \int_{\Delta P_1P_2P_3} \nabla u \cdot \nabla u dx dy = \mathbf{u}^T \mathbf{K}_e \mathbf{u}$$

is

$$-\{(x_3 - x_1)(x_3 - x_2) + (y_3 - y_1)(y_3 - y_2)\}/|\Delta| = -|P_1P_3||P_2P_3|\cos\gamma/|\Delta| < 0,$$

as γ is acute. Similarly, because the angles α and β are acute, the coefficients of u_2u_3 and u_1u_3 are negative as well. Therefore, the signs of the element stiffness matrix $\mathbf{K}_e$ are

$$\mathbf{K}_e = \begin{pmatrix} + & - & - \\ - & + & - \\ - & - & + \end{pmatrix}.$$

■

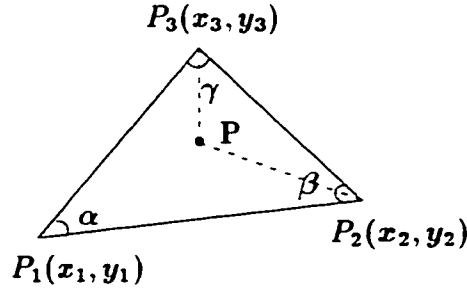


Figure 3.1: An arbitrary element in a triangular mesh.

Theorem 3.1 can be generalized to right-angled triangular elements. In this case, the coefficient corresponding to the right-angle is zero. $\mathbf{K}_e$ still has non-positive off-diagonals.

In the finite element $P_1P_2P_3$ as shown in Figure 3.1, u also can be written as

$$u(x, y) = u_1\phi_1(x, y) + u_2\phi_2(x, y) + u_3\phi_3(x, y),$$

where the ϕ_i are the areal coordinates of the triangle (Carey and Oden [5], Vol. 2).

For instance, let $\mathbf{p} = (x, y)$ be any point in the triangle, then

$$\phi_i = \frac{\text{Area}(PP_2P_3)}{\text{Area}(P_1P_2P_3)} \geq 0.$$

The ϕ_i are positive inside $P_1P_2P_3$ so that if

$$\int \int_{\Delta P_1P_2P_3} \rho u^2 dx dy = \mathbf{u}^T \mathbf{M}_e \mathbf{u},$$

then the element mass matrix $\mathbf{M}_e$ has the form

$$\mathbf{M}_e = \begin{pmatrix} + & + & + \\ + & + & + \\ + & + & + \end{pmatrix}.$$

There is a similar pattern for tetrahedral elements in three-dimensions. We first introduce some notation about the tetrahedron.

Consider an arbitrary tetrahedron $P_1P_2P_3P_4$ shown in Figure 3.2. The coordinates of vertex P_i are (x_i, y_i, z_i) , for $1 \leq i \leq 4$. The outward normal to triangle $P_2P_3P_4$ (may not be unit vector) is $\overrightarrow{P_2P_3} \times \overrightarrow{P_2P_4} = \tilde{\mathbf{n}}^{(1)}$, to triangle $P_1P_3P_4$ is $\overrightarrow{P_1P_4} \times \overrightarrow{P_1P_3} = \tilde{\mathbf{n}}^{(2)}$, to triangle $P_1P_2P_4$ is $\overrightarrow{P_1P_2} \times \overrightarrow{P_1P_4} = \tilde{\mathbf{n}}^{(3)}$, and to triangle $P_1P_2P_3$ is $\overrightarrow{P_1P_3} \times \overrightarrow{P_1P_2} = \tilde{\mathbf{n}}^{(4)}$.

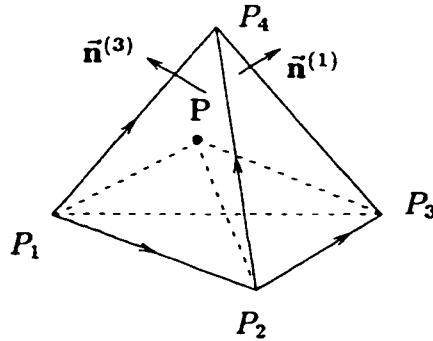


Figure 3.2: A tetrahedral finite element.

Let $\mathbf{T}$ be the coordinate matrix of tetrahedron $P_1P_2P_3P_4$, i.e.,

$$\mathbf{T} = \begin{pmatrix} 1 & x_1 & y_1 & z_1 \\ 1 & x_2 & y_2 & z_2 \\ 1 & x_3 & y_3 & z_3 \\ 1 & x_4 & y_4 & z_4 \end{pmatrix}.$$

Define $\Delta = |\mathbf{T}|$ where $|\Delta| = 6 * \text{Volume}(P_1P_2P_3P_4)$. For each $i = 1, 2, 3, 4$, we define X_i , Y_i and Z_i : X_i is the determinant of the submatrix of $\mathbf{T}$ obtained by deleting the x -column and i -th row; Y_i and Z_i are defined in a similar fashion. Notice that

$$X_1 = \begin{vmatrix} 1 & y_2 & z_2 \\ 1 & y_3 & z_3 \\ 1 & y_4 & z_4 \end{vmatrix} = (y_3 - y_2)(z_4 - z_2) - (z_3 - z_2)(y_4 - y_2) = n_1^{(1)}.$$

Following the similar computations, we can establish the relation between X_i , Y_i , Z_i and the outward normals:

$$X_i = (-1)^{i+1}n_1^{(i)}, \quad Y_i = (-1)^i n_2^{(i)}, \quad \text{and} \quad Z_i = (-1)^{i+1}n_3^{(i)}.$$

With these notation, we can easily prove that

Theorem 3.2 *Let $\Omega \subset \mathbb{R}^3$. If Ω is divided into tetrahedra whose angles between exterior normals are obtuse, then the local stiffness matrix $\mathbf{K}_e$ has negative off-diagonal.*

Proof: Assume that u takes the linear form

$$u = a + bx + cy + dz$$

in each tetrahedron element. Then

$$\nabla u \cdot \nabla u = b^2 + c^2 + d^2.$$

Since u takes the value u_i at vertex P_i , $i = 1, 2, 3, 4$, we have

$$\begin{aligned} -b\Delta &= u_1X_1 - u_2X_2 + u_3X_3 - u_4X_4 \\ &= u_1n_1^{(1)} + u_2n_1^{(2)} + u_3n_1^{(3)} + u_4n_1^{(4)}; \\ -c\Delta &= -u_1Y_1 + u_2Y_2 - u_3Y_3 + u_4Y_4 \\ &= u_1n_2^{(1)} + u_2n_2^{(2)} + u_3n_2^{(3)} - u_4n_2^{(4)}; \\ -d\Delta &= u_1Z_1 - u_2Z_2 + u_3Z_3 - u_4Z_4 \\ &= u_1n_3^{(1)} - u_2n_3^{(2)} + u_3n_3^{(3)} - u_4n_3^{(4)}. \end{aligned}$$

Substituting the above equations into $\nabla u \cdot \nabla u$, we find that the sign of the coefficient of $u_i u_j$ ($i \neq j$) is fixed by $\vec{n}^{(i)} \cdot \vec{n}^{(j)}$. That is, the signs of the off-diagonal entries in K_e are determined by the inner products of the outward normals.

Consider $\vec{n}^{(2)} \cdot \vec{n}^{(4)}$. Both of $\vec{n}^{(2)}$ and $\vec{n}^{(4)}$ are orthogonal to $\overrightarrow{P_1P_3}$ and can thus be placed in the plane orthogonal to $\overrightarrow{P_1P_3}$. Assume this plane cuts $\overrightarrow{P_1P_2}$, $\overrightarrow{P_1P_3}$, and $\overrightarrow{P_1P_4}$ at A, B, C as shown in Figure 3.3. Since $\angle ABC$ is acute, the angle between $\vec{n}^{(2)}$ and $\vec{n}^{(4)}$ is obtuse. Thus $\vec{n}^{(2)} \cdot \vec{n}^{(4)} < 0$. Applying the same reasoning to other

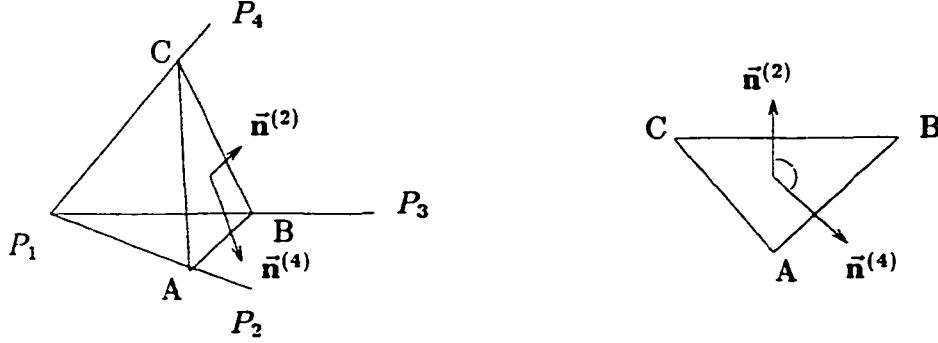


Figure 3.3: The angles between the outward normals are abused.

inner products, we know that $\mathbf{K}_e$ has negative off-diagonal entries. ■

Again, this theorem is applicable to right-angled tetrahedral finite elements. The off-diagonal entries in $\mathbf{K}_e$ associated with the right-angles are zero, the others are negative, and thus the off-diagonals of $\mathbf{K}_e$ are non-positive.

In each tetrahedron $P_1 P_2 P_3 P_4$, u can be expressed as

$$u(x, y, z) = u_1 \phi_1 + u_2 \phi_2 + u_3 \phi_3 + u_4 \phi_4$$

where the ϕ_i 's are the volume coordinates for the tetrahedron. For example,

$$\phi_4 = \frac{\text{Volume}(P P_1 P_2 P_3)}{\text{Volume}(P_1 P_2 P_3 P_4)} \geq 0.$$

Therefore, $\mathbf{M}_e$ is positive.

To summarize, if the finite element mesh satisfies the required constraints, then the global matrices $\mathbf{K}$ and $\mathbf{M}$ possess the following properties.

- $\mathbf{K}$, $\mathbf{M}$ are symmetric and positive definite;
- $\mathbf{K}$ has non-positive off-diagonals, *i.e.*, $\mathbf{K}$ is an M-matrix;
- $\mathbf{M}$ has non-negative off-diagonals, where $m_{ij} > 0$ iff there is a mesh line connecting vertices i and j .

In this thesis, we will always assume that $\mathbf{K}$ and $\mathbf{M}$ have these properties unless specified otherwise.

3.1.2 The necessity of $\mathbf{K}$ having the right signs

The constraint that $\mathbf{K}$ has non-positive off-diagonal entries is essential. If not, the discrete counterpart of CNLT may not hold. We present an example where CNLT fails when $\mathbf{K}$ has some positive off-diagonal entries.

Example 3.1: Consider a simple case of (1.1) where $\rho \equiv 1$ and $\Omega = [-3, 3] \times [-4, 4]$. The domain Ω is divided into triangles where obtuse-angled triangles appear, as shown in Figure 3.4.

The stiffness matrix $\mathbf{K}$ and the mass matrix $\mathbf{M}$ obtained by FEM have the same

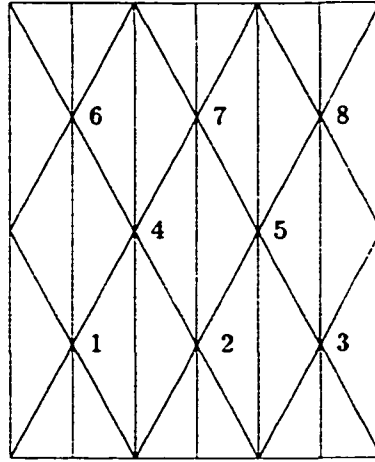


Figure 3.4: The triangulation of the rectangle where some obtuse-angled triangle appear.

structure

$$A = \begin{pmatrix} a_1 & & & & & & & \\ & a_1 & & & & & & \\ & & a_1 & & & & & \\ a_3 & a_3 & & a_2 & & a_3 & a_3 & \\ & a_3 & a_3 & & a_2 & & a_3 & a_3 \\ a_4 & & & a_3 & & a_1 & & \\ & a_4 & & a_3 & a_3 & & a_1 & \\ & & a_4 & & a_3 & & & a_1 \end{pmatrix}.$$

The nonzero entries of $\mathbf{M}$ are positive. $\mathbf{K}$ has positive diagonal; however, among the off-diagonal elements of $\mathbf{K}$, $a_4 > 0$ due to the contributions from the obtuse

angles while $a_3 < 0$.

The third eigenvalue of $(\mathbf{K}, \mathbf{M})$ is multiple where $\lambda_3 = \lambda_4 = \lambda_5$. The eigenmodes corresponding to λ_3 have the form

$$\mathbf{u} = (x, y, z, w, -w, -x, -y, -z)^T.$$

$(\mathbf{K} - \lambda\mathbf{M})\mathbf{u} = \mathbf{0}$ can be reduced to an equivalent four by four linear system:

$$(k_1 - \lambda m_1)x - (k_4 - \lambda m_4)x + (k_3 - \lambda m_3)w = 0 \quad (3.8)$$

$$(k_1 - \lambda m_1)y - (k_4 - \lambda m_4)y = 0 \quad (3.9)$$

$$(k_1 - \lambda m_1)z - (k_4 - \lambda m_4)z = 0 \quad (3.10)$$

$$(k_2 - \lambda m_2)w = 0 \quad (3.11)$$

Taking $\lambda = \lambda_3 = \frac{k_1 - k_4}{m_1 - m_4}$, we have $w = 0$ and x, y and z satisfy the same equation

$$(k_1 - \lambda m_1)x - (k_4 - \lambda m_4)x = 0, \quad i.e., \quad 0 \cdot x = 0.$$

Therefore, the eigenvectors corresponding to $\lambda_3 = \lambda_4 = \lambda_5$ can have the form

$$\mathbf{u} = (x, y, z, 0, 0, -x, -y, -z)^T,$$

in which x, y and z are any real numbers as long as not all of them are zero. Hence, we can choose the eigenmode $\{-1, +1, -1, 0, 0, +1, -1, +1\}$. It divides the rectangle into *six* regions as shown in Figure 3.5. The FEM solutions fail to have

discrete CNLT property.

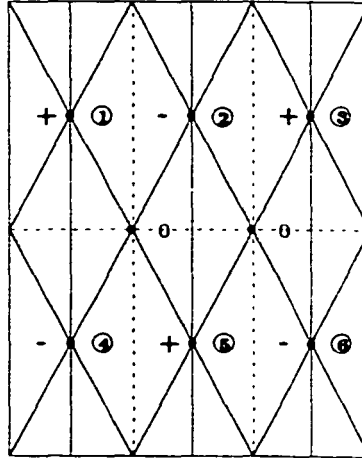


Figure 3.5: The nodal lines of an approximated eigenfunction corresponding to λ_3 , divide the domain into six nodal domains. But $\lambda_3 = \lambda_4 = \lambda_5 < \lambda_6$. Thus, discrete CNLT fails.

Therefore, if $\mathbf{K}$ does not have non-positive off-diagonal, *i.e.*, $\mathbf{K}$ is not an M-matrix, then the FEM solutions do not necessarily have the CNLT property.

3.2 Graph theory notation and properties of FEM eigenvectors

3.2.1 Graph theory notation

To obtain the discrete counterparts of CNLT, we first introduce some basic terminologies in graph theory.

A graph $\mathcal{G} = (V, E)$ consists of a finite set of vertices V together with a set E of

edges, which are unordered pairs of vertices. Such a graph is useful in representing the structure of symmetric matrices. Let $\mathbf{A}$ be an $N \times N$ symmetric matrix. The associated/underlying graph of $\mathbf{A}$, denoted by $\mathcal{G}(\mathbf{A}) = (V(\mathbf{A}), E(\mathbf{A}))$, is one for which the N vertices of $\mathcal{G}(\mathbf{A})$ are numbered from 1 to N , and $(i, j) \in E(\mathbf{A})$ iff $a_{ij} = a_{ji} \neq 0$ ($i \neq j$).

As $\mathcal{G}(\mathbf{M})$ reflects exactly the structure and connectivity of the FEM mesh (triangular or tetrahedral mesh), we consider $\mathcal{G}(\mathbf{M})$ as the associated graph of the symmetric matrix pair $(\mathbf{K}, \mathbf{M})$. For example, in an acute-angled triangular mesh, $\mathbf{K}$ and $\mathbf{M}$ will have the same non-zero structure, i.e., $k_{ij} < 0$ and $m_{ij} > 0$ iff i and j are adjacent, i.e., i and j are the end-vertices of the same mesh line. (If an element is right angled, the element stiffness matrix $\mathbf{K}_e$ has a zero off-diagonal entry. Thus k_{ij} may be zero even though there is a mesh line connecting i and j . Hence, $\mathcal{G}(\mathbf{K})$ is a subgraph of $\mathcal{G}(\mathbf{M})$; $\mathcal{G}(\mathbf{K})$ may not be exactly the same as the mesh.)

For distinct vertices i and j in $\mathcal{G}$, a *path* from i to j is an ordered set of vertices $(i_1, i_2, \dots, i_{p+1})$ such that $(i_k, i_{k+1}) \in E(\mathcal{G})$, $k = 1, 2, \dots, p$ with $i_1 = i$ and $i_{p+1} = j$. A graph is *connected* if every pair of distinct vertices is linked by at least one path. It is well known that $\mathcal{G}(\mathbf{A})$ is a connected graph if and only if $\mathbf{A}$ is irreducible (see Busacker and Saaty [4], p.111). If Ω is connected, then the FEM mesh is a connected graph and hence $\mathbf{K}$ and $\mathbf{M}$ are irreducible. It is easy to see that $\mathbf{M}$ is irreducible because $\mathcal{G}(\mathbf{M})$ is exactly the mesh. But it is not so obvious for $\mathbf{K}$ when $\mathcal{G}(\mathbf{K})$ is only a subgraph of the mesh, such as in the case that there are right-angled triangular elements. If an element is right angled, $\mathbf{K}_e$ has a zero on the off-diagonal; $\mathcal{G}(\mathbf{K}_e)$ is not a triangle, but a chain connecting all three vertices. Hence $\mathcal{G}(\mathbf{K}_e)$ is

still connected and $\mathbf{K}_e$ is irreducible. In general, to prove that $\mathbf{K}$ is irreducible, it is sufficient to show that any two vertices i and j are connected by a path in $\mathcal{G}(\mathbf{K})$. If the mesh is connected, then i and j are connected by a path in the FEM mesh, say $(i_1, i_2, \dots, i_{p+1})$ where $i_1 = i$ and $i_{p+1} = j$. Suppose that some edge in the path, say (i_k, i_{k+1}) , is not in $\mathcal{G}(\mathbf{K})$. Then i_k and i_{k+1} lie in a right-angled triangle, say $\triangle i_k i_{k+1} t$. In the triangle, both (i_k, t) and (t, i_{k+1}) are edges in $\mathcal{G}(\mathbf{K})$. So we can replace (i_k, i_{k+1}) by a path (i_k, t, i_{k+1}) in $\mathcal{G}(\mathbf{K})$. Repeating this procedure for the other edges not in $\mathcal{G}(\mathbf{K})$, we can find a new path connecting i and j that is contained in $\mathcal{G}(\mathbf{K})$, which implies that $\mathcal{G}(\mathbf{K})$ is connected. Hence $\mathbf{K}$ is irreducible.

A subgraph $\mathcal{G}' = (V', E')$ of $\mathcal{G}$ is a graph for which $V' \subseteq V$ and $E' \subseteq E$. We also call $\mathcal{G}'$ the *induced subgraph* on the vertex subset V' . Note that $(i, j) \in E'$ implies that both i and j belong to V' . The concept of subgraphs is related to submatrices. In matrix terms, the subgraph $\mathcal{G}(V')$ of $\mathcal{G}(\mathbf{A})$ is the graph of the matrix obtained by deleting the rows and columns from $\mathbf{A}$ that correspond to $V \setminus V'$. This is illustrated in Figure 3.6.

An eigenvector $\mathbf{u}$ of $\mathbf{A} \in \mathbb{R}^{N \times N}$ specifies a sequence of real numbers $\{u_i\}_1^N$. We can associate $\mathbf{u}$ with the graph $\mathcal{G}(\mathbf{A})$ by assigning signs to the vertices: the i -th vertex is positive if $u_i > 0$, negative if $u_i < 0$, or, zero if $u_i = 0$ ($1 \leq i \leq N$). See the example in Figure 3.6. By doing so, we then can compare the characteristics of eigenvectors with those of eigenfunctions.

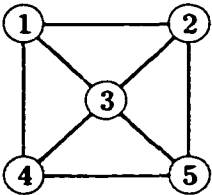
$$\mathbf{A} = \begin{bmatrix} \textcircled{1} & \text{X} & \text{X} & \text{X} \\ \text{X} & \textcircled{2} & \text{X} & \\ \text{X} & \text{X} & \textcircled{3} & \text{X} & \text{X} \\ \text{X} & & \text{X} & \textcircled{4} & \text{X} \\ & \text{X} & \text{X} & \text{X} & \textcircled{5} \end{bmatrix}$$

Matrix $\mathbf{A}$

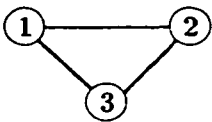
$$V_1 = \{ 1, 2, 3 \}$$

$$\begin{bmatrix} \textcircled{1} & \text{X} & \text{X} \\ \text{X} & \textcircled{2} & \text{X} \\ \text{X} & \text{X} & \textcircled{3} \end{bmatrix}$$

Submatrix associated with $\mathcal{G}(V_1)$



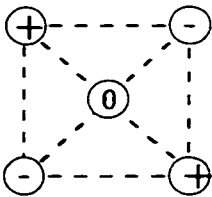
Graph $\mathcal{G}(\mathbf{A})$



Induced subgraph $\mathcal{G}(V_1)$

$$\mathbf{u} = \{ 1, -1, 0, -1, 1 \}$$

An eigenvector of matrix $\mathbf{A}$



The vertices in $\mathcal{G}(\mathbf{A})$ with assigned signs with respect to $\mathbf{u}$.

Figure 3.6: Basic concepts in graph theory and matrix correspondences.

3.2.2 Properties of FEM eigenvectors

Recall that one of the main theorems about the solution of (1.1) is the unique continuation property: $u(\mathbf{x})$ cannot vanish in any non-empty open subset of Ω . However, a discrete analogue of the result does not hold for FEM models, in the sense that a FEM solution may have one or more elements that are completely zero, as illustrated in Example 3.2.

Example 3.2: In the triangulation of a square shown in Figure 3.7, the fifth eigenvector of the discretized system is zero in four complete triangles.

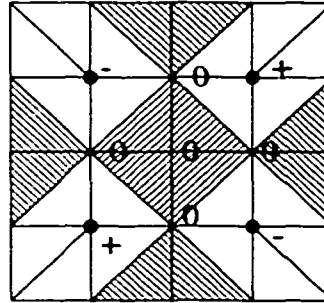


Figure 3.7: A FEM solution can have zero (shaded) polygons.

Because of the sign structures of $\mathbf{K}$ and $\mathbf{M}$, the FEM eigenvectors do inherit some characteristics of eigenfunctions, such as the maximum principle: $u(\mathbf{x})$ cannot have an interior negative maximum or interior positive minimum. In the matrix eigenvalue problem of (1.1), the indices of the coordinates in $\mathbf{u}$ are composed of interior vertices and near-boundary vertices. If i is an interior vertex, all the elements to which i belongs, are interior elements. Since the Helmholtz equation with Neumann boundary conditions has the first eigenfunction $u_1 \equiv 1$, the FEM approximation would have the first eigenvector $\mathbf{u}_1 = \{1, \dots, 1\}$. Hence, the stiffness

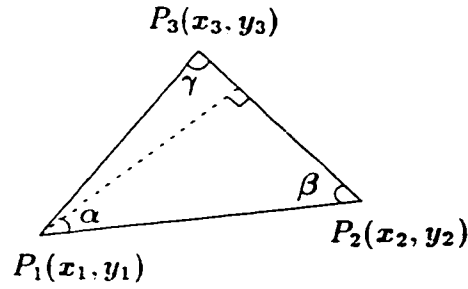
matrix $\mathbf{K}_e$ of an interior element admits a rigid body mode $\{1, \dots, 1\}$, i.e.,

$$\mathbf{K}_e\{1, \dots, 1\} = \mathbf{0}.$$

This can be also seen from the calculation in § 3.1.1. For instance, in the triangular element as shown Figure 3.1, the sum of the off-diagonal terms in the first row of $\mathbf{K}_e$ is

$$k_{e12} + k_{e13} = -(|P_1 P_3| \cos \gamma + |P_1 P_2| \cos \beta) |P_2 P_3| / |\Delta| = -|P_2 P_3|^2 / |\Delta| = -k_{e11}.$$

It is also true for the sums of the rest rows. Thus, $\mathbf{K}_e\{1, 1, 1\} = \mathbf{0}$.



This means that if i is an interior vertex, after assembling $\mathbf{K}$, we have

$$\sum_{j=1}^n k_{ij} = 0.$$

If i is a near-boundary vertex, it is adjacent to boundary vertices. Thus, after

assembling $\mathbf{K}$, we have

$$\sum_{j=1}^n k_{ij} > 0.$$

We then have the modified maximum principle for the eigenvectors.

Theorem 3.3 *The eigenvectors of $(\mathbf{K}, \mathbf{M})$ cannot attain a negative maximum or a positive minimum at an interior vertex.*

Proof: A contradiction can be derived directly from the i -th equation of $\mathbf{K}\mathbf{u} = \lambda\mathbf{M}\mathbf{u}$

$$\sum_j k_{ij}u_j = \lambda \sum_j m_{ij}u_j, \quad (3.12)$$

i.e.,

$$\sum_j k_{ij}(u_j - u_i) + (\sum_j k_{ij})u_i = \lambda \sum_j m_{ij}u_j. \quad (3.13)$$

Suppose that there is a local positive minimum at an interior vertex i , such that $u_i > 0$ and $u_j - u_i \geq 0$ for $(i, j) \in E(\mathbf{M})$. On the left hand side of (3.13), the first sum is non-positive and the second is zero due to $\sum_j k_{ij} = 0$. Hence, the left hand side is non-positive. But the sum on the right hand side is positive as $\lambda \sum m_{ij}u_j > m_{ii}u_i > 0$. This is a contradiction. Therefore, there is no interior positive minimum. Similarly, there is no interior negative maximum. ■

A particular case of (3.13) is when i is a node, *i.e.*, $u_i = 0$. We conclude that i cannot be adjacent only to vertices j for which the values of u_j are all non-negative,

or all non-positive, unless they are all zero. Otherwise, suppose that u_i is adjacent to all positive vertices. The left hand side of (3.13) satisfies $\sum_{j \neq i} k_{ij} u_j < 0$ while the right hand side $\lambda \sum_{j \neq i} m_{ij} u_j > 0$. This is impossible. Thus, a zero of any eigenvector of $(\mathbf{K}, \mathbf{M})$, no matter whether it is an interior vertex or near-boundary vertex, is either

- (a) connected only to zeros; or
- (b) connected to both positive and negative vertices.

For example, consider an acute-angled triangular mesh in $D \in \mathbb{R}^2$ with $u = 0$ on the boundary. Let $u_I(x, y)$ be a FEM solution constructed from an eigenvector of (3.2). The second case implies that a node on a nodal line of $u_I(x, y)$ must have positive and negative neighbours. The first case gives us a zero polygon. A zero polygon, unless all its vertices are on the boundary (namely boundary zero polygon), must have both positive and negative neighbours, see Figure 3.7. Therefore, $u_I(x, y)$ cannot have isolated interior nodal point nor isolated interior zero polygon. The nodal set of $u_I(x, y)$ consists of nodal lines and nodal polygons. Inside each triangle, $u_I(x, y)$ is given by linear interpolation. Thus, nodal lines are straight lines inside any one triangle. They are either closed or begin and end at the boundary.

Hence, the nodal set of $u_I(x, y)$ will divide D into polygonal sign domains. Inside of each sign domain, $u_I(x, y)$ will be either positive or negative. There may be also some nodal polygons.

This argument can be extended to tetrahedral meshes in $\mathbb{R}^3$ which satisfies the condition of Theorem 3.2. A FEM solution $u_I(\mathbf{x})$ cannot have isolated interior nodal points or lines going through interior points or interior nodal volumes. The

nodal set of $u_I(\mathbf{x})$ will consist of piecewise nodal surfaces or nodal volumes. The nodal surfaces will be either closed or begin and end at the boundary. The nodal set of $u_I(\mathbf{x})$ will divide D into a number of sign domains.

3.3 Sign Graphs

In order to formulate the discrete counterpart of CNLT formally, we need to introduce the discrete analogue of a nodal domain. We name it a sign graph while Duval and Reiner [8] continue to call it a nodal domain. The term nodal domain can potentially create unnecessary confusions in the discrete case. This can be seen through the two dimensional case. Since nodal lines refer to lines formed by zero points, one may be tempted to term nodal domains as subregions formed by zero points. In fact, a nodal domain (from CNLT) refers to a subregion encircled by nodal lines or boundaries. Since in the continuous setting, by the unique continuation property of eigenfunctions, there cannot exist any open subset of positive measure formed by zero points, the CNLT style definition of nodal domain cannot cause confusion. However, in the discrete setting, there do exist subregions of positive measure formed by zero points, nodal polygons, in the FEM solutions and thus there is potential confusion of terminology. Since CNLT describes a sign characteristic of eigenfunctions/eigenvectors, using the notation of sign graph is preferable.

The nodal set of a FEM solution $u_I(\mathbf{x})$ subdivides Ω into positive, negative sign domains and nodes. This subdivision of Ω defines a subdivision of the vertices and the edges of the mesh. In general, let $\mathcal{G}$ be the associated graph of a symmetric

matrix pair $(\mathbf{K}, \mathbf{M})$ of order N . Given a vector $\mathbf{u} \in \mathbb{R}^{N \times 1}$, the *sign graphs* of $\mathbf{u}$ are the connected components in the graph obtained from $\mathcal{G}$ by

- (1) deleting all the nodes i where $u_i = 0$ and their incident edges;
- (2) deleting all the edges (i, j) for which $u_i u_j < 0$.

In each of these connected components, $\mathbf{u}$ takes a fixed sign: if $\mathbf{u}$ is positive (negative), we call such a connected component a positive (negative) sign graph. The definition of sign graphs implies that two adjacent vertices in $\mathcal{G}$, having the same sign, lie in the same sign graph. Thus, a positive sign graph is adjacent to either zeros or negative sign graphs. We say that two sign graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are adjacent if there is an edge (i, j) in $\mathcal{G}$ such that $i \in V(\mathcal{G}_1)$ and $j \in V(\mathcal{G}_2)$. Figure 3.8 illustrates the concept of sign graphs with respect to a vector $\mathbf{u}$.

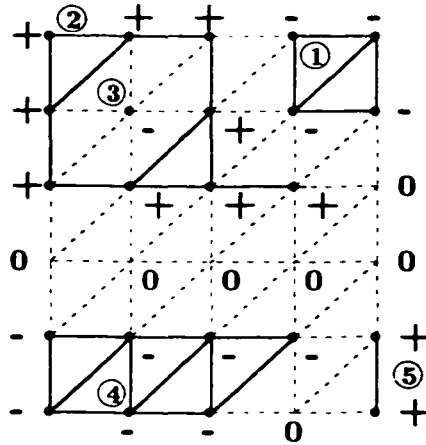


Figure 3.8: The graph $\mathcal{G}$ is associated with a pair of symmetric matrices $(\mathbf{K}, \mathbf{M})$. The subgraphs that are indicated with thick lines or vertices are sign graphs. There are three negative sign graphs and two positive sign graphs. Sign graph 2 is adjacent to sign graphs 1 and 3, but is disconnected from sign graphs 4 and 5.

Let $u_I(\mathbf{x})$ be the FEM continuous solution constructed from the vector $\mathbf{u}$. In

an arbitrary element of a triangular or tetrahedral mesh, $u_I(\mathbf{x})$ takes a linear form. Hence, inside of any element where there are at least two vertices having opposite signs, a nodal line or surface appears and is linear. In addition, since zeros of $\mathbf{u}$ cannot be surrounded by all positive vertices or all negative vertices, the nodal lines or surfaces separate positive and negative sign domains of $u_I(\mathbf{x})$. We then can conclude that the number of the sign graphs of $\mathbf{u}$ is equal to that of the sign domains of $u_I(\mathbf{x})$. First of all, since each positive sign graph is adjacent to either negative sign graphs or zeros, a nodal line of the FEM approximation will be formed in the elements that are adjacent to this positive sign graph; hence, each positive (negative) sign graph of $\mathbf{u}$ is embedded in one positive (negative) sign domain of $u_I(\mathbf{x})$. On the other hand, inside each positive (negative) sign domain, there will be only one positive (negative) sign graph; otherwise, suppose that a positive sign domain Ω^+ of $u_I(\mathbf{x})$ includes several sign graphs of $\mathbf{u}$. Those sign graphs have to be positive; otherwise, $u_I(\mathbf{x})$ changes sign in Ω^+ , which implies that Ω^+ is not a positive sign domain. Thus, Ω^+ is composed of several positive sign graphs and zeros (the appearance of zeros is due to the fact that positive sign graphs are disjoint). In this case, there exists at least one zero that connects to all positive vertices, which contradicts the property of zeros. Therefore, there is a one-to-one correspondence between sign domains of FEM solutions and sign graphs of eigenvectors. Because of the consistency, we will connect the nodal places of u_I to separate positive and negative sign graphs of $\mathbf{u}$, see Figure 3.9.

Note that the rectangular mesh may not provide such consistency. Inside each

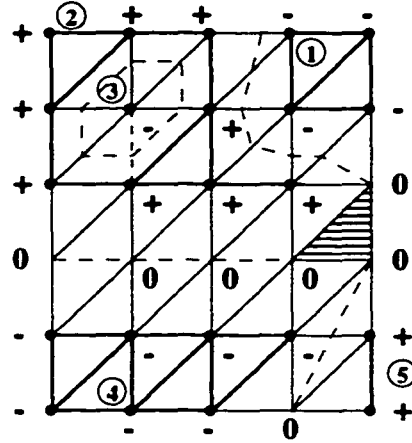


Figure 3.9: The sign domains of the FEM continuous solution $u_I(\mathbf{x})$ constructed from vector $\mathbf{u}$ in the example shown in Figure 3.8.

rectangle, $u_I(x, y)$ takes the bilinear form

$$u_I(x, y) = a + bx + cy + dxy.$$

Its nodal lines may be hyperbolic inside one element. But all the vertices of the rectangle are adjacent to one another. A paradox may happen in this case. The example in Figure 3.10 shows that vertex A and B belong to different sign domains of $u_I(x, y)$, but are in the same sign graph of $\mathbf{u}$ as they are adjacent.

We are also interested in studying another kind of sign graph which was studied by Fiedler [10]. We term them *weak sign graphs*, i.e., sign graphs of vertices i where $u_i \geq 0$ ($u_i \leq 0$). Strictly speaking, the *non-negative (non-positive) sign graphs* of $\mathbf{u}$ are the connected components in the graph obtained from $\mathcal{G}(\mathbf{A})$ by deleting all the vertices i for which $u_i < 0$ ($u_i > 0$) and their incident edges. Each non-negative (non-positive) sign graph contains at least one positive (negative) sign

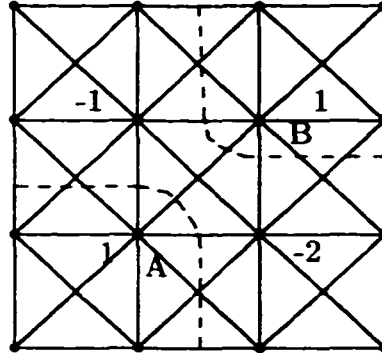


Figure 3.10: Inconsistency of sign domains of $u(x, y)$ and the sign graphs of $\mathbf{u}$ occurs in rectangular mesh.

graph. Clearly, given a graph $\mathcal{G}(\mathbf{A})$ and a vector $\mathbf{u}$, the number of weak sign graphs cannot exceed the number of strict sign graphs. In the example in Figure 3.8, there are three weak sign graphs: a non-negative sign graph consisting of positive sign graphs 2 and 5 connected by zeros, a non-positive sign graph consisting of negative sign graphs 1 and 4 connected by zeros and a non-positive sign graph consisting of only a negative sign graph 3. Two weak sign graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are defined to be adjacent if there is one strict sign graph in $\mathcal{G}_1$ adjacent to a strict sign graph in $\mathcal{G}_2$.

We next state and prove the discrete versions of CNLT.

3.4 Discrete CNLT: simple eigenvalues

With the definitions we are able to state and prove discrete analogues of CNLT for the eigenvectors of $(\mathbf{K}, \mathbf{M})$ derived from Dirichlet boundary condition and satisfying the conditions mentioned in § 3.1.1. We first consider the case that the eigenvalues

of (3.2) are distinct. Assume that $\mathbf{K}, \mathbf{M} \in \mathbb{R}^{N \times N}$.

Theorem 3.4 *If the eigenvalues of $(\mathbf{K}, \mathbf{M})$ are distinct and are ordered increasingly, then the n -th eigenvector $\mathbf{u}_n$ has at most n strict sign graphs.*

The proof of Theorem 3.4 is based on two lemmas.

Lemma 3.5 (Minimax Principle) *Let $\mathbf{B}$ and $\mathbf{C}$ be symmetric matrices with $\mathbf{C}$ positive definite. If the eigenvalues of $(\mathbf{B}, \mathbf{C})$ are labeled increasingly, then*

$$\lambda_k = \min_{\dim(S)=k} \max_{\substack{\mathbf{x} \in S \\ \mathbf{x} \neq \mathbf{0}}} \frac{\mathbf{x}^T \mathbf{B} \mathbf{x}}{\mathbf{x}^T \mathbf{C} \mathbf{x}},$$

where S is a k -dimensional subspace of $\mathbb{R}^N$ (see Stewart [23]).

The proof of the minimax principle is similar to that of Theorem 2.2. In the context of matrix pair $(\mathbf{B}, \mathbf{C})$, $\mathbf{x} \perp \mathbf{w}_i$ means that $\mathbf{x}$ is $\mathbf{C}$ -orthogonal to $\mathbf{w}_i$, i.e., $\mathbf{x}^T \mathbf{C} \mathbf{w}_i = 0$. When $\mathbf{C} = \mathbf{I}$, we just say that $\mathbf{x}$ is orthogonal to $\mathbf{w}_i$. For simplicity, we sometimes define $(\mathbf{x}, \mathbf{w}_i)_{\mathbf{C}} = \mathbf{x}^T \mathbf{C} \mathbf{w}_i$.

Let $\mathbf{A} \geqslant \mathbf{0}$, i.e., $\mathbf{A}$ is a non-negative matrix. By the Perron-Frobenius theorem, an eigenvector corresponding to the spectral radius, the greatest eigenvalue, of an *irreducible* non-negative matrix is strictly positive. We next show that a strictly positive eigenvector of a non-negative matrix, *not* necessarily irreducible, must correspond to the spectral radius of the matrix (Berman and Plemmons [3]).

Lemma 3.6 *Let $\mathbf{A} \geqslant \mathbf{0}$. If $\mathbf{u}$ is a positive vector, such that $\mathbf{A} \mathbf{u} = \lambda \mathbf{u}$, then $\lambda = \rho(\mathbf{A})$.*

Proof: Let $\mathbf{y}$ be an eigenvector of $\mathbf{A}^T$ corresponding to $\rho(\mathbf{A})$. By the Perron-Frobenius theorem, $\mathbf{y}$ is nonnegative. We then have

$$\lambda \mathbf{y}^T \mathbf{u} = \mathbf{y}^T \mathbf{A} \mathbf{u} = (\mathbf{u}^T \mathbf{A}^T \mathbf{y})^T = \rho(\mathbf{A})(\mathbf{u}^T \mathbf{y})^T = \rho(\mathbf{A}) \mathbf{y}^T \mathbf{u}.$$

But $\mathbf{u} > \mathbf{0}$ and $\mathbf{y} \geq \mathbf{0}$, thus $\mathbf{y}^T \mathbf{u} > 0$. Therefore, $\lambda = \rho(\mathbf{A})$. ■

With the two lemmas, we are ready to prove Theorem 3.4.

Proof: Suppose that $\mathbf{u}_n$ has m sign graphs, say $\{\mathcal{G}_k\}_1^m$, where $m \leq N$. Without loss of generality, the vertex i runs consecutively through $\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_m$ and then through any vertices for which $u_i = 0$. Thus

$$\mathbf{u}_n^T = \{\mathbf{v}_1^T, \mathbf{v}_2^T, \dots, \mathbf{v}_m^T, \mathbf{0}^T\}, \quad (3.14)$$

where the terms in $\mathbf{v}_j$ are labeled $v_{j,k}$. Now construct the m vectors $\mathbf{w}_j = \mathbf{w}_j(N \times 1)$ so that

$$\mathbf{w}_j^T = \{\mathbf{0}^T, \mathbf{0}^T, \dots, \mathbf{v}_j^T, \mathbf{0}^T, \dots, \mathbf{0}^T\}. \quad (3.15)$$

As there are no overlaps among the vertex sets of $\{\mathcal{G}_k\}_1^m$, $\{\mathbf{w}_j\}_1^m$ have non-overlapping entries.

With $\{\mathbf{w}_j\}_1^m$, we can consider the eigenvalue problem (3.2) on the m -dimensional space spanned by $\{\mathbf{w}_j\}_1^m$. Consider

$$\mathbf{u} = \sum_{j=1}^m \alpha_j \mathbf{w}_j = \mathbf{W} \boldsymbol{\alpha}, \quad (3.16)$$

where $\mathbf{W} = \mathbf{W}(N \times m) = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m]$ and $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_m\}$. In this notation

$$\mathbf{u}_n = \mathbf{W}\mathbf{e}, \quad (3.17)$$

where $\mathbf{e} = \{1, 1, \dots, 1\}$. The Rayleigh Quotient of $\mathbf{u}$ is

$$\lambda_R(\mathbf{u}) = \frac{\mathbf{u}^T \mathbf{K} \mathbf{u}}{\mathbf{u}^T \mathbf{M} \mathbf{u}} = \frac{\alpha^T \mathbf{W}^T \mathbf{K} \mathbf{W} \alpha}{\alpha^T \mathbf{W}^T \mathbf{M} \mathbf{W} \alpha} = \frac{\alpha^T \mathbf{B} \alpha}{\alpha^T \mathbf{C} \alpha}. \quad (3.18)$$

Since $\mathbf{K}$ and $\mathbf{M}$ are symmetric and positive-definite, $\mathbf{B}$ and $\mathbf{C}$ are also symmetric and positive-definite. The terms in $\mathbf{B}$ and $\mathbf{C}$ are

$$b_{ij} = \mathbf{w}_i^T \mathbf{K} \mathbf{w}_j, \quad c_{ij} = \mathbf{w}_i^T \mathbf{M} \mathbf{w}_j. \quad (3.19)$$

So the associated graph of $(\mathbf{B}, \mathbf{C})$ describes the connections of the sign graphs of $\mathbf{u}_n$. Any two sign graphs $\mathcal{G}_i, \mathcal{G}_j$ ($i \neq j$), are *adjacent* only if they have opposite signs. This implies that $\mathbf{w}_i$ and $\mathbf{w}_j$ have opposite signs. In addition the off-diagonal terms of $\mathbf{K}$ are non-positive and of $\mathbf{M}$ are non-negative. Thus

$$b_{ij} \geq 0 \quad \text{and} \quad c_{ij} \leq 0$$

for $i \neq j$. This means that $\mathbf{B}$ is positive definite and non-negative; $\mathbf{B} \geq \mathbf{0}$. Matrix $\mathbf{C}$, being a positive definite matrix with non-positive off-diagonal terms, is a nonsingular M-matrix; the inverse of such a matrix is non-negative (Berman and Plemmons, p.13 (N98)), *i.e.*, $\mathbf{C}^{-1} \geq \mathbf{0}$.

In the space spanned by $\{\mathbf{w}_j\}_1^m$, the eigenvalue problem is

$$(\mathbf{B} - \mu\mathbf{C})\alpha = 0. \quad (3.20)$$

We know one eigenvalue of (3.20), namely λ_n ; the corresponding eigenvector is $\mathbf{e} = \{1, 1, \dots, 1\}$, which gives $\mathbf{u}_n$ in the original space. Multiplying $\mathbf{C}^{-1}$ on both sides of the equation (3.20) yields

$$(\mathbf{C}^{-1}\mathbf{B} - \mu\mathbf{I})\alpha = 0.$$

i.e.,

$$(\mathbf{A} - \mu\mathbf{I})\alpha = 0. \quad (3.21)$$

Since $\mathbf{C}$ is an M-matrix, $\mathbf{C}^{-1} \geq 0$. This implies $\mathbf{C}^{-1}\mathbf{B} \geq 0$, *i.e.*, $\mathbf{A}$ is a non-negative matrix. Hence, by Lemma 3.6, the eigenvector $\mathbf{e}$ corresponds to the greatest eigenvalue of (3.21), namely μ_m . Then $\lambda_n = \mu_m$. On the other hand, μ_m is the m th eigenvalue of (3.2) on the space spanned by $\{\mathbf{w}_j\}_1^m$ which is a subspace of R^N . The m th eigenvalue on a subspace is not less than the m th eigenvalue on the whole space, *i.e.*, $\mu_m \geq \lambda_m$. We deduce that $\lambda_m \leq \lambda_n$. Since, by hypothesis, the eigenvalues are distinct, $\lambda_m \leq \lambda_n$ implies $m \leq n$. ■

3.5 Discrete CNLT: multiple eigenvalues

Recall that in the continuous version of CNLT, the proof of multiple eigenvalue is based on the unique continuation theorem. Assume that λ_n is a multiple eigenvalue and u_n has more than n nodal domains. A new eigenfunction u corresponding to λ_n can be constructed from u_n , where u is defined on n nodal domains of u_n and identically zero on the rest. Hence, by the unique continuation theorem, u has to be identically zero in Ω . This contradiction shows that any eigenfunction corresponding to λ_n has at most n nodal domains.

In § 3.2.2., we showed that there is not a discrete analogue of the unique continuation theorem. That is, an eigenvector can be zero in one or more complete elements without being identically zero. Thus, a direct translation of the proof of CNLT does not hold for the matrix eigenvalue problem.

In fact, when λ_n is a multiple eigenvalue, it is not always true that any eigenvector corresponding to λ_n has at most n sign graphs. As the counterexamples in Figure 3.11, the eigenspace of the multiple eigenvalue $\lambda_3 = \lambda_4 = \lambda_5 = \lambda_6$ consists of those vectors $\mathbf{u}$ satisfying the equations

$$\begin{cases} u_4 = u_5 = 0; \\ u_1 + u_6 = u_3 + u_8 = -(u_2 + u_7). \end{cases}$$

With these conditions, we can find a set of M -orthogonal eigenvectors corresponding to the multiple eigenvalue, each of which has six sign graphs, see Figure 3.11. This shows that Theorem 6 of Duval and Reiner [8] is false, in which it implies that any eigenvector corresponding to λ_n , no matter whether λ_n is simple or multiple, has at

most n sign graphs. This is a major difference between the continuous and discrete CNLT.

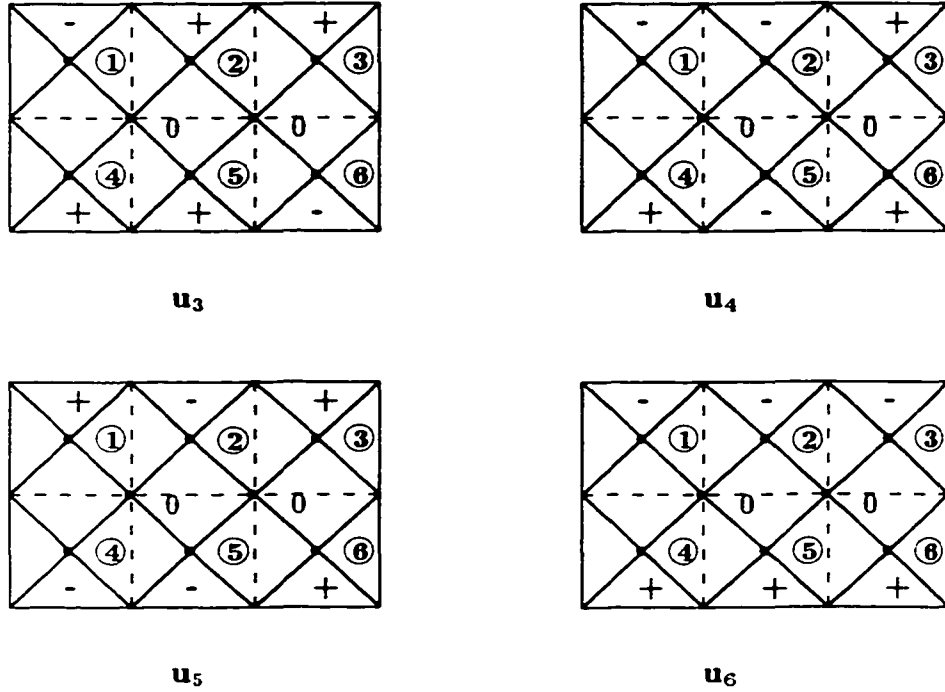


Figure 3.11: There exist four M -orthogonal eigenvectors associated with the multiple eigenvalue λ_3 , each of which has 6 sign graphs.

Hence, when λ_n is a r -fold eigenvalue, i.e.,

$$\lambda_{n-1} < \lambda_n = \lambda_{n+1} = \cdots = \lambda_{n+r-1} < \lambda_{n+r}, \quad (3.22)$$

some of its eigenvectors can have more than n sign graphs. A direct conclusion from Theorem 3.4 tells that any eigenvector corresponding to the multiple eigenvalue λ_n has at most $n + r - 1$ sign graphs. The question now is if these

eigenvectors can have a stronger property. For the multiple eigenvalue λ_3 in Figure 3.11, we exhibit four M -orthogonal eigenvectors $\{\mathbf{u}_j\}_3^6$, so that $\mathbf{u}_j$ has at most j sign graphs for $j = 3, 4, 5, 6$, see Figure 3.12; we also exhibit four linearly independent eigenvectors, all of which have at most three sign graphs, see Figure 3.13. In fact, these are generally true. To prove them, we first state two lemmas.

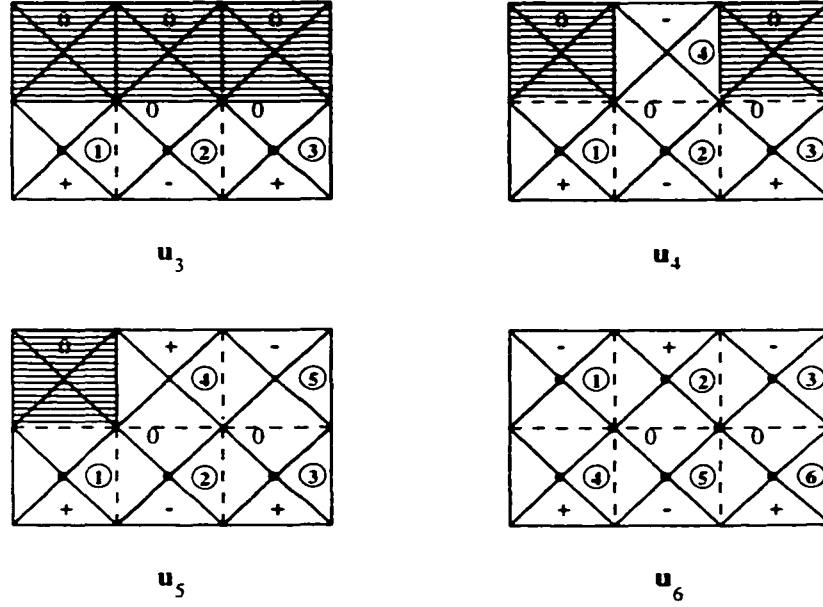


Figure 3.12: There exist four M -orthogonal eigenvectors $\{\mathbf{u}_j\}_3^6$ associated with the multiple eigenvalue λ_3 such that $SG(\mathbf{u}_j) \leq j$.

Corollary 3.7 Suppose $\mathbf{u}_n$ is an eigenvector corresponding to λ_n and the $\{\mathbf{w}_j\}_1^m$ are defined as in equation (3.15). In the space spanned by $\{\mathbf{w}_j\}_1^m$, no eigenvalue exceeds λ_n .

Proof: In the space spanned by $\{\mathbf{w}_j\}_1^m$, λ_n is the largest eigenvalue of $(\mathbf{B}, \mathbf{C})$ in the equation (3.20) as in the argument of Theorem 3.4. ■

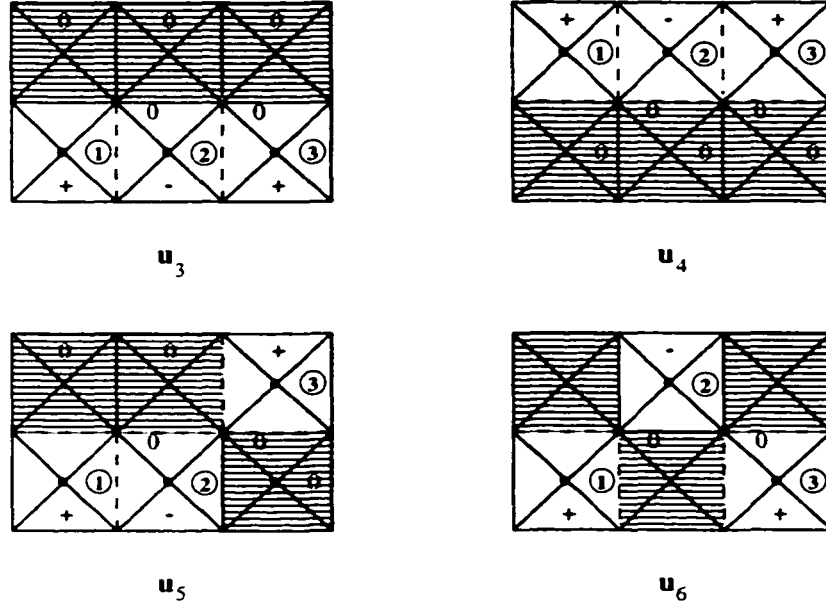


Figure 3.13: There also exist four linearly independent eigenvectors $\{\mathbf{u}_j\}_3^6$ associated with the multiple eigenvalue λ_3 , each of which has 3 sign graphs.

For simplicity, we now introduce the notation

$$SG(\mathbf{u}) = \text{number of sign graphs of } \mathbf{u}.$$

Given an eigenvector $\mathbf{u}$ corresponding to λ_n with $SG(\mathbf{u}) > n$, we can also construct another eigenvector $\mathbf{v}$ from $\mathbf{u}$ as in the continuous case, so that $SG(\mathbf{v}) \leq n$.

Corollary 3.8 *If $\mathbf{u}$ is an eigenvector corresponding to λ_n and $SG(\mathbf{u}) = m > n$, then, in the notation of equation (3.15), we can find*

$$\mathbf{v} = \sum_{j=1}^n \alpha_j \mathbf{w}_j$$

such that $\mathbf{v}$ is an eigenvector corresponding to λ_n and $SG(\mathbf{v}) \leq n$.

Proof: Choose $\{\alpha_j\}_1^n$ not all zero, such that $\mathbf{v}$ is M -orthogonal to $\{\mathbf{u}_i\}_1^{n-1}$. By the minimax theorem, $\lambda_R \geq \lambda_n$, and by Corollary 3.7, $\lambda_R \leq \lambda_n$. Thus $\lambda_R = \lambda_n$. By its construction $SG(\mathbf{v}) \leq n$. ■

In fact, there will be many such $\mathbf{v}$ which may be formed from a given $\mathbf{u}$. For simplicity, we call any such M -orthogonal one $\mathbf{v} = T(\mathbf{u}, \{\mathbf{u}_j\}_1^{n-1})$. In general, $\mathbf{v} = T(\mathbf{u}, \{\mathbf{u}_j\}_1^{n-1})$ implies that

- (a) $SG(\mathbf{u}) > n$;
- (b) $\mathbf{v} \neq \mathbf{0}$ is constructed from n of the $\mathbf{w}_j$, where the $\mathbf{w}_j$ are defined from $\mathbf{u}$ as in (3.15). By the construction, $SG(\mathbf{v}) \leq n$.
- (c) $\mathbf{v}$ is M -orthogonal to $\mathbf{u}_j$ for $j = 1, 2, \dots, n-1$.

We now state and prove a discrete version of CNLT for multiple eigenvalues.

Theorem 3.9 *If λ_n is a r -fold eigenvalue of $(\mathbf{K}, \mathbf{M})$ as in (3.22), then we may find r M -orthogonal eigenvectors $\{\mathbf{u}_j\}_n^{n+r-1}$ such that $SG(\mathbf{u}_j) \leq j$.*

Proof: Take an M -orthonormal basis $\{\mathbf{v}_j\}_n^{n+r-1}$ in the r -dimensional eigenspace V of λ_n . If $SG(\mathbf{v}_n) \leq n$, take $\mathbf{u}_n = \mathbf{v}_n$; otherwise, take $\mathbf{u}_n = T(\mathbf{v}_n, \{\mathbf{u}_j\}_1^{n-1})$ and construct a new M -orthogonal basis $\mathbf{u}_n, \{\mathbf{v}^{(1)}_j\}_{n+1}^{n+r-1}$ for V . If $SG(\mathbf{v}^{(1)}_{n+1}) \leq n+1$, take $\mathbf{u}_{n+1} = \mathbf{v}^{(1)}_{n+1}$; otherwise, take $\mathbf{u}_{n+1} = T(\mathbf{v}^{(1)}_{n+1}, \{\mathbf{u}_j\}_1^n)$ and construct a new orthogonal basis $\{\mathbf{u}_j\}_n^{n+1}, \{\mathbf{v}^{(2)}_j\}_{n+1}^{n+r-1}$. Repeat the process until we find a new M -orthogonal basis $\{\mathbf{u}_j\}_n^{n+r-1}$ for V . As a result, $SG(\mathbf{u}_j) \leq j$ for $j = n, n+1, \dots, n+r-1$. ■

We can also prove a stronger result.

Theorem 3.10 *If λ_n is a r -fold eigenvalue of $(\mathbf{K}, \mathbf{M})$ as in (3.22), then we may find r linearly independent eigenvectors $\{\mathbf{v}_j\}_n^{n+r-1}$ corresponding to λ_n such that $SG(\mathbf{v}_j) \leq n$ for $j = n, n+1, \dots, n+r-1$.*

Proof: We start with an M -orthogonal basis $\{\mathbf{u}_j\}_n^{n+r-1}$ of the eigenspace V corresponding to λ_n . $\{\mathbf{u}_j\}_n^{n+r-1}$ are labeled so that $SG(\mathbf{u}_k) \leq SG(\mathbf{u}_\ell)$ for $k \leq \ell$. Let $\mathbf{u}_s$ be the first eigenvector in $\{\mathbf{u}_j\}_n^{n+r-1}$ such that $SG(\mathbf{u}_s) = n+t > n$. Take $\mathbf{v}_j = \mathbf{u}_j$ for $j = n, n+1, \dots, s-1$.

Consider the nonzero function $\mathbf{y} = \sum_{j=1}^{n+t} \alpha_j \mathbf{w}_j$, where $\{\mathbf{w}_j\}_1^{n+t}$ is constructed from $\mathbf{u}_s$ as (3.15). Choose the α_j so that

$$(\mathbf{u}_i, \mathbf{y})_{\mathbf{m}} = 0, \quad \text{for } i = 1, 2, \dots, n-1,$$

i.e.,

$$\sum_{j=1}^{n+t} (\mathbf{u}_i, \mathbf{w}_j)_{\mathbf{m}} \alpha_j = 0, \quad i = 1, 2, \dots, n-1. \quad (3.23)$$

These are $n-1$ homogeneous linear equations in $n+t$ unknown variables, which has $m \geq t+1$ independent variables. A basis of the solution space of (3.23) can be obtained by taking only one of the independent variables nonzero at a time. It then gives a basis $\{\mathbf{y}_j\}_1^m$ for the λ_n eigenspace V' spanned by $\{\mathbf{w}_j\}_1^{n+t}$. Note that $SG(\mathbf{y}_j) \leq n+t-m+1 \leq n$ for $j = 1, 2, \dots, m$. As $\mathbf{u}_s$ is linearly independent of $\{\mathbf{v}_j\}_n^{s-1}$, there exists at least one of the $\mathbf{y}_j$ that is linearly independent of $\{\mathbf{v}_j\}_n^{s-1}$, say $\mathbf{y}_1$. Take $\mathbf{v}_s = \mathbf{y}_1$ ($SG(\mathbf{v}_s) \leq n$). Construct a new basis $\{\mathbf{v}_j\}_n^s$, $\{\mathbf{u}^{(1)}_j\}_{s+1}^{n+r-1}$ for V where $\{\mathbf{u}^{(1)}_j\}_{s+1}^{n+r-1}$ are labeled so that $SG(\mathbf{u}^{(1)}_k) \leq SG(\mathbf{u}^{(1)}_\ell)$ for $k \leq \ell$. We

continue this process and thus construct a linearly independent basis $\{\mathbf{v}_j\}_n^{n+r-1}$ for V with $SG(\mathbf{v}_j) \leq n$ for all j . ■

The above argument can be extended to the eigenvectors of $(\mathbf{K}, \mathbf{M})$ derived from Neumann boundary conditions. In the FEM procedure, the element stiffness matrix $\mathbf{K}_e$ of an element admits a rigid body mode $\mathbf{e} = \{1, \dots, 1\}$, i.e.,

$$\mathbf{K}_e \mathbf{e} = \mathbf{0}.$$

With Neumann boundary conditions, the values of boundary vertices in the mesh need to be computed as well, while those for Dirichlet boundary condition are known. Thus, with Neumann boundary condition, the indices of the assembled global stiffness and mass matrices $(\mathbf{K}, \mathbf{M})$ include all the mesh vertices. In this case, $\mathbf{K}$ is singular and satisfies

$$\mathbf{K} \mathbf{e} = \mathbf{0};$$

hence, zero is the first eigenvalue of $(\mathbf{K}, \mathbf{M})$ and $\mathbf{e}$ is the first eigenvector. If the mesh satisfies the conditions stated in § 3.1.1 (no right angles in the mesh are allowed), $\mathbf{K}$ and $\mathbf{M}$ satisfy

- $\mathbf{M}$ are symmetric, positive definite and has non-negative off-diagonals (the same as in the Dirichlet boundary conditions);
- $\mathbf{K}$ are symmetric, semi-positive definite and has non-positive off-diagonals; $\mathbf{K}$ is a singular M-matrix. In addition, $\mathbf{K} \mathbf{e} = \mathbf{0}$; $\mathbf{K}$ is also a Laplacian matrix as

in Fiedler [10];

- $\mathbf{K}$ and $\mathbf{M}$ have the same non-zero structure.

However, the proofs of Theorem 3.4, 3.9 and 3.10 require that $\mathbf{K}$ is positive definite. In order to apply the same argument for Neumann boundary condition, we add a reasonable positive amount to $\mathbf{K}$ to avoid the singularity but still keep the off-diagonal entries non-positive. To do so, we choose a positive number s such that

$$0 < s \leq \min_{\substack{i \neq j \\ m_{ij} \neq 0}} \{-k_{ij}/m_{ij}\}.$$

Instead of considering $\mathbf{K}\mathbf{u} = \lambda\mathbf{M}\mathbf{u}$, we consider

$$(\mathbf{K} + s\mathbf{M})\mathbf{u} = (\lambda + s)\mathbf{M}\mathbf{u}.$$

Denote $\tilde{\mathbf{K}} = \mathbf{K} + s\mathbf{M}$ and $\mu = \lambda + s$. We then have

$$\tilde{\mathbf{K}}\mathbf{u} = \mu\mathbf{M}\mathbf{u}.$$

Now $\mathbf{M}$ is positive definite, by the choice of s , $\tilde{\mathbf{K}}$ is also positive definite with non-positive off-diagonal terms. Thus, the analysis for Dirichlet boundary conditions is also applicable to the case of Neumann boundary conditions.

Also note that irreducibility of $\mathbf{K}$ and $\mathbf{M}$ is not required in the theorems on discrete CNLT. The discrete analogues of CNLT need not to be restricted to FEM eigenvectors, but can be also applied to the eigenvectors of a single symmetric $\mathbf{M}$ -

matrix. Hence, the results of discrete CNLT can be regarded as a matrix theory, which gives a sign characteristic of the eigenvectors of matrix pair $(\mathbf{K}, \mathbf{M})$ or matrix $\mathbf{K}$ satisfying the properties mentioned.

3.6 More on discrete analogues of CNLT

As mentioned in Chapter 1, Fiedler [10], Duval and Reiner [8] also have studied the matrix analogue of CNLT. They dealt with a single matrix $\mathbf{K}$, where $\mathbf{K}$ is real symmetric with non-positive off-diagonal entries. Note that there is no requirement for the diagonal entries of $\mathbf{K}$. This is because $\mathbf{K} + s\mathbf{I}$ will be positive definite for sufficiently large $s > 0$, i.e., $\mathbf{K} + s\mathbf{I}$ can be an M-matrix. Although they dealt with a single matrix, their results can be generalized to matrix pair $(\mathbf{K}, \mathbf{M})$, as we will show in this section. We also prove more theorems on discrete CNLT.

3.6.1 Discussion of Fiedler's result

Let $\mathbf{K}$ be a real symmetric matrix with non-positive off-diagonal entries. Fiedler [10] proved that the n -th eigenvector of $\mathbf{K}$ has no more than $n - 1$ non-negative sign graphs. His proof is based on the minimax principle and the following lemma.

Lemma 3.11 ((Fiedler [10], Theorem 1.2)) *Any principal submatrix of a symmetric matrix $\mathbf{A} - \lambda_n \mathbf{I}$ has no more than $n - 1$ negative eigenvalues.*

Proof: Let the eigenvalues of $\mathbf{A}$ be

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N.$$

Then the eigenvalues of $\mathbf{A} - \lambda_n \mathbf{I}$ are $\lambda_i - \lambda_n$ for $i = 1, 2, \dots, N$; among them, there are $n - 1$ negative eigenvalues. By the minimax principle of eigenvalues, the i -th eigenvalue μ_i of any principal submatrix is no less than the i -th eigenvalue of $\mathbf{A} - \lambda_n \mathbf{I}$, i.e., $\mu_i \geq \lambda_i - \lambda_n$. Thus, the statement holds. ■

This lemma is also applicable to the eigenvalues of $(\mathbf{K}, \mathbf{M})$ as the generalized eigenvalue problem can be transformed to the standard one by using the Cholesky factorization of $\mathbf{M}$ [23] :

$$\begin{aligned} (\mathbf{K} - \lambda \mathbf{M})\mathbf{u} &= \mathbf{0} \\ &\downarrow \mathbf{M} = \mathbf{L}\mathbf{L}^T \\ (\mathbf{L}^{-1}\mathbf{K}(\mathbf{L}^{-1})^T - \lambda \mathbf{I})\mathbf{L}^T\mathbf{u} &= \mathbf{0} \\ &\downarrow \mathbf{A} = \mathbf{L}^{-1}\mathbf{K}(\mathbf{L}^{-1})^T, \quad \mathbf{v} = \mathbf{L}^T\mathbf{u} \\ (\mathbf{A} - \lambda \mathbf{I})\mathbf{v} &= \mathbf{0}. \end{aligned}$$

With Lemma 3.11, we can generalize Fiedler's result ([10], Theorem 2.1) to the eigenvectors of $(\mathbf{K}, \mathbf{M})$.

Theorem 3.12 *Let $\mathbf{K}$ and $\mathbf{M}$ be irreducible. If $\mathbf{K}\mathbf{u}_n = \lambda_n \mathbf{M}\mathbf{u}_n$, then $\mathbf{u}_n$ has at most $n - 1$ non-negative sign graphs.*

Proof: Suppose that $\mathbf{u}_n$ has at least n non-negative sign graphs., say $r \geq n$. Without loss of generality, we can assume that $\mathbf{u}_n$ can be partitioned into

$$\mathbf{u}_n^T = \{\mathbf{v}_1^T, \mathbf{v}_2^T, \dots, \mathbf{v}_r^T, \mathbf{v}_{r+1}^T\}, \quad (3.24)$$

where $\mathbf{v}_i \geq \mathbf{0}$, for $i = 1, 2, \dots, r$ and $\mathbf{v}_{r+1} < \mathbf{0}$. $\mathbf{K}$ and $\mathbf{M}$ are partitioned conformally, in which each $\mathbf{K}_{ii}$ and $\mathbf{M}_{ii}$ are irreducible.

As $\mathbf{K}\mathbf{u}_n = \lambda_n \mathbf{M}\mathbf{u}_n$, we have

$$(\mathbf{K}_{ii} - \lambda_n \mathbf{M}_{ii})\mathbf{v}_i = -(\mathbf{K}_{i,r+1} - \lambda_n \mathbf{M}_{i,r+1})\mathbf{v}_{r+1}, \quad (3.25)$$

for $i = 1, 2, \dots, r$. Now consider a principal submatrix of $\mathbf{K} - \lambda_n \mathbf{M}$

$$\begin{pmatrix} \mathbf{K}_{11} - \lambda_n \mathbf{M}_{11} & & & \\ & \mathbf{K}_{22} - \lambda_n \mathbf{M}_{22} & & \\ & & \ddots & \\ & & & \mathbf{K}_{rr} - \lambda_n \mathbf{M}_{rr} \end{pmatrix}.$$

From Lemma 3.11, there is at least one of the matrices $\mathbf{K}_{ii} - \lambda_n \mathbf{M}_{ii}$ whose eigenvalues are non-negative, say $\mathbf{K}_{11} - \lambda_n \mathbf{M}_{11}$.

Suppose that $\mathbf{K}_{11} - \lambda_n \mathbf{M}_{11}$ is nonsingular. Then $\mathbf{K}_{11} - \lambda_n \mathbf{M}_{11}$ is positive definite and has non-positive off-diagonal. Thus $\mathbf{K}_{11} - \lambda_n \mathbf{M}_{11}$ is a nonsingular M-matrix. This gives $(\mathbf{K}_{11} - \lambda_n \mathbf{M}_{11})^{-1} > 0$. From the equation (3.25) for $i = 1$, we have

$$(\mathbf{K}_{11} - \lambda_n \mathbf{M}_{11})\mathbf{v}_1 = -(\mathbf{K}_{1,r+1} - \lambda_n \mathbf{M}_{1,r+1})\mathbf{v}_{r+1},$$

i.e.,

$$\mathbf{v}_1 = -(\mathbf{K}_{11} - \lambda_n \mathbf{M}_{11})^{-1}(\mathbf{K}_{1,r+1} - \lambda_n \mathbf{M}_{1,r+1})\mathbf{v}_{r+1} \leq 0,$$

as $\mathbf{K}_{1,r+1} - \lambda_n \mathbf{M}_{1,r+1} \leq 0$ and $\mathbf{v}_{r+1} < 0$. But $\mathbf{v}_1 \geq 0$, so that $\mathbf{v}_1 = 0$, and hence

$$(\mathbf{K}_{1,r+1} - \lambda_n \mathbf{M}_{1,r+1})\mathbf{v}_{r+1} = 0.$$

Since $\mathbf{v}_{r+1} < \mathbf{0}$, $\mathbf{K}_{1,r+1} \leq \mathbf{0}$ and $-\lambda_n \mathbf{M}_{1,r+1} \leq \mathbf{0}$, we get

$$\mathbf{K}_{1,r+1} = \mathbf{M}_{1,r+1} = \mathbf{0},$$

a contradiction to $\mathbf{K}$ and $\mathbf{M}$ being irreducible.

Hence $\mathbf{K}_{11} - \lambda_n \mathbf{M}_{11}$ must be singular. $\mathbf{K}_{11} - \lambda_n \mathbf{M}_{11}$ is a singular irreducible M-matrix, and there is a positive vector $\mathbf{x}_1$ corresponding to the zero eigenvalue, so that

$$\mathbf{x}_1^T (\mathbf{K}_{11} - \lambda_n \mathbf{M}_{11}) = \mathbf{0}.$$

This leads to

$$\mathbf{x}_1^T (\mathbf{K}_{11} - \lambda_n \mathbf{M}_{11}) \mathbf{v}_1 = -\mathbf{x}_1^T (\mathbf{K}_{1,r+1} - \lambda_n \mathbf{M}_{1,r+1}) \mathbf{v}_{r+1} = \mathbf{0}.$$

Since $(\mathbf{K}_{1,r+1} - \lambda_n \mathbf{M}_{1,r+1}) \mathbf{v}_{r+1} \geq \mathbf{0}$ and $\mathbf{x}_1^T > \mathbf{0}$, then

$$(\mathbf{K}_{1,r+1} - \lambda_n \mathbf{M}_{1,r+1}) \mathbf{v}_{r+1} = \mathbf{0}.$$

Again, because of $\mathbf{v}_{r+1} < \mathbf{0}$, we have

$$(\mathbf{K}_{1,r+1} - \lambda_n \mathbf{M}_{1,r+1}) = \mathbf{0}.$$

Hence $\mathbf{K}_{1,r+1} = \mathbf{M}_{1,r+1} = \mathbf{0}$. $\mathbf{K}$ and $\mathbf{M}$ are reducible, a contradiction. ■

Theorem 3.12 holds for the non-positive sign graphs as well. Although we do not expect that $\mathbf{u}_n$ has at most n sign graphs in general due to the discussion in § 3.5, a question is raised – whether $\mathbf{u}_n$ has at most n *weak* sign graphs. The upper bound we could get from Theorem 3.12 is $2(n - 1)$, which is far larger than n . However, Gladwell [14] recently extended Theorem 3.12 and has proved that $\mathbf{u}_n$ has at most n weak sign graphs; that is the closest interpretation of CNLT in discrete case. The proof of Gladwell [14] is very interesting but rather technical; so we will not go through it here.

However, the results of weak sign graphs cannot produce an upper bound for positive (negative) sign graphs of $\mathbf{u}_n$, because a non-negative (non-positive) sign graph may contain more than one positive (negative) sign graphs. Thus, it is also of interest to give an upper bound for the number of positive (negative) sign graphs in $\mathbf{u}_n$.

When λ_n ($n \geq 2$) is simple, it is a direct consequence of the discrete CNLT for distinct eigenvalue that $\mathbf{u}_n$ has at most $n - 1$ positive sign graphs.

However, when λ_n is multiple, $\mathbf{u}_n$ may have more than $n - 1$ positive sign graphs. The simplest example is a star with N vertices. Matrix $\mathbf{K}$ is the adjacency matrix of the star, *i.e.*,

$$k_{ij} = \begin{cases} -1, & \text{if } i \neq j \text{ and } (i, j) \in E; \\ 0, & \text{if } i \neq j \text{ and } (i, j) \notin E; \\ -\sum_{t \neq i} k_{it}, & \text{if } i = j. \end{cases}$$

$\mathbf{M} = \mathbf{I}$ in this case. The second eigenvalue of $\mathbf{K}$ is an eigenvalue with multiplicity $N - 2$. The eigenvectors corresponding to λ_2 are zero at the center vertex of the star, and the sum of the values on the remainder of the vertices is zero. Hence, we can construct a second eigenvector with $N - 2$ positive sign graphs and one negative sign graph. It has more than one positive sign graph if $N > 3$, see Figure 3.14. Figure 3.15 shows a counterexample for a matrix pair $(\mathbf{K}, \mathbf{M})$, generated by

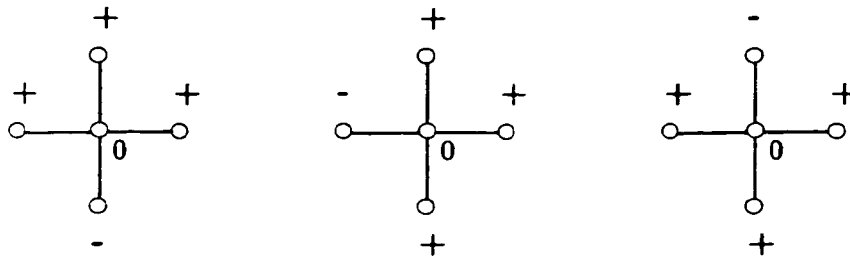


Figure 3.14: A star with five vertices. There exists a set of linearly independent eigenvectors of $\mathbf{K}$ corresponding to λ_2 , each of which has three positive sign graphs.

FEM method for a square with Dirichlet boundary conditions. In this example, the underlying graph of $(\mathbf{K}, \mathbf{M})$ is a star with 4 vertices. Similarly, the second eigenvalue is 3-fold, and its corresponding eigenvectors satisfy the same property as the eigenvectors for a single matrix $\mathbf{K}$. $\mathbf{u}_2$ can have more than one positive sign graph.

However, under the condition that there is at least an edge connecting a positive and a negative vertex, i.e., a positive sign graph is adjacent to a negative sign graph, then $\mathbf{u}_n$ has no more than $n - 1$ positive sign graphs.

Theorem 3.13 *Let $\mathbf{K}\mathbf{u}_n = \lambda_n\mathbf{M}\mathbf{u}_n$ where $\mathbf{K}$ and $\mathbf{M}$ need not be irreducible. If there exists a positive sign graph adjacent to a negative sign graph, then $\mathbf{u}_n$ has at*

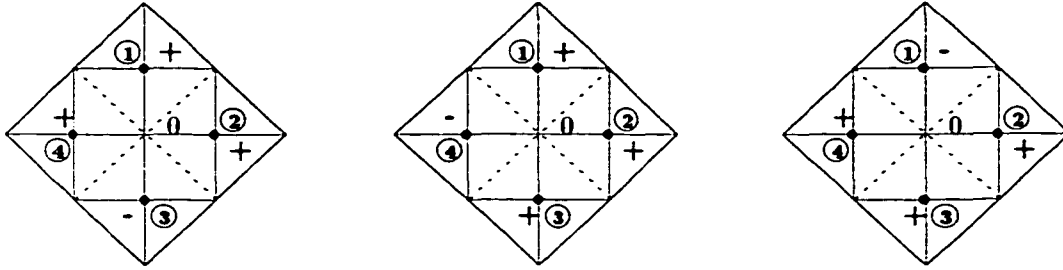


Figure 3.15: For matrix pair $(\mathbf{K}, \mathbf{M})$ generated by FEM, there also exists a set of linearly independent eigenvectors corresponding to λ_2 , each of which has three positive sign graphs.

most $n - 1$ positive sign graphs.

Proof: Suppose that $\mathbf{u}_n$ has r positive sign graphs where $r \geq n$. Partition $\mathbf{u}_n$ in the same way as (3.24), where $\mathbf{v}_i > 0$, $i = 1, 2, \dots, r$ and $\mathbf{v}_{r+1} \leq 0$. $\mathbf{K}$ and $\mathbf{M}$ are partitioned accordingly.

Construct a vector $\mathbf{u} = \sum_{i=1}^r c_i \mathbf{v}_i$. The coefficients c_i are chosen so that $\mathbf{u}$ is $\mathbf{M}$ -orthogonal to the first $n - 1$ eigenvectors. By the minimax principle,

$$R(\mathbf{u}) \geq \lambda_n.$$

On the other hand, since $\mathbf{K}\mathbf{u}_n = \lambda_n \mathbf{M}\mathbf{u}_n$, we can have

$$\mathbf{K}_{ii}\mathbf{v}_i + \mathbf{K}_{i,r+1}\mathbf{v}_{r+1} = \lambda_n(\mathbf{M}_{ii}\mathbf{v}_i + \mathbf{M}_{i,r+1}\mathbf{v}_{r+1}),$$

for $i = 1, 2, \dots, r$. Since $\mathbf{K}_{i,r+1} \leq 0$, $\mathbf{M}_{i,r+1} \geq 0$ and $\mathbf{v}_{r+1} \leq 0$, then $\mathbf{K}_{i,r+1}\mathbf{v}_{r+1} \geq 0$ and $\mathbf{M}_{i,r+1}\mathbf{v}_{r+1} \leq 0$. So, for $i = 1, 2, \dots, r$, we get

$$\mathbf{K}_{ii}\mathbf{v}_i \leq \lambda_n \mathbf{M}_{ii}\mathbf{v}_i. \quad (3.26)$$

In addition, there is a positive sign graph adjacent to a negative sign graph. Let the positive sign graph be $\mathcal{G}_1$. As $\mathcal{G}_1$ is connected to a negative sign graph, $\mathbf{K}_{1,r+1}\mathbf{v}_{r+1} > 0$ and $\mathbf{M}_{1,r+1}\mathbf{v}_{r+1} < 0$. Hence, for $i = 1$, the inequality holds strictly

$$\mathbf{K}_{11}\mathbf{v}_1 < \lambda_n \mathbf{M}_{11}\mathbf{v}_1. \quad (3.27)$$

Since any two positive sign graphs are not adjacent, we have

$$\begin{aligned} \mathbf{u}^T \mathbf{K} \mathbf{u} &= \sum_{i=1}^r c_i^2 \mathbf{v}_i^T \mathbf{K}_{ii} \mathbf{v}_i; \\ \mathbf{u}^T \mathbf{M} \mathbf{u} &= \sum_{i=1}^r c_i^2 \mathbf{v}_i^T \mathbf{M}_{ii} \mathbf{v}_i; \end{aligned}$$

As $\mathbf{v}_i > 0$ for $i = 1, 2, \dots, r$, from equations (3.26) and (3.27), we get

$$\mathbf{u}^T \mathbf{K} \mathbf{u} = \sum_{i=1}^r c_i^2 \mathbf{v}_i^T \mathbf{K}_{ii} \mathbf{v}_i < \lambda_n \sum_{i=1}^r c_i^2 \mathbf{v}_i^T \mathbf{M}_{ii} \mathbf{v}_i = \lambda_n \mathbf{u}^T \mathbf{M} \mathbf{u}.$$

Hence $\mathbf{u}^T \mathbf{K} \mathbf{u} < \lambda_n \mathbf{u}^T \mathbf{M} \mathbf{u}$, i.e., $R(\mathbf{u}) < \lambda_n$, a contradiction to $R(\mathbf{u}) \geq \lambda_n$. ■

3.6.2 Discussion of Duval and Reiner's result

If λ_n is simple, Duval and Reiner [8] proved that the n -th eigenvector of $\mathbf{K}$ has at most n sign graphs, where $\mathbf{K}$ is real symmetric with non-positive off-diagonal entries. Their approach is different from ours and is based on the minimax principle and the following key calculation lemma.

Lemma 3.14 ([8], Lemma 5) *Let $\mathbf{u} \in \mathbb{R}^{N \times 1}$. Suppose that $\mathbf{u}$ has the sign graphs $\mathcal{G}_1, \dots, \mathcal{G}_m$. Define $\mathbf{v} = \sum_{i=1}^m c_i \mathbf{w}_i$ in the notation of (3.15). Then*

$$\mathbf{v}^T \mathbf{K} \mathbf{v} - \lambda \mathbf{v}^T \mathbf{v} = \sum_i c_i^2 \mathbf{w}_i^T (\mathbf{K} \mathbf{u} - \lambda \mathbf{u}) - \sum_{i < j} (c_i - c_j)^2 \mathbf{w}_i^T \mathbf{K} \mathbf{w}_j. \quad (3.28)$$

When $\mathbf{u}$ is an eigenvector corresponding to λ , i.e., $\mathbf{K} \mathbf{u} - \lambda \mathbf{u} = \mathbf{0}$ and (3.28) can be simplified to

$$\mathbf{v}^T \mathbf{K} \mathbf{v} - \lambda \mathbf{v}^T \mathbf{v} = - \sum_{i < j} (c_i - c_j)^2 \mathbf{w}_i^T \mathbf{K} \mathbf{w}_j. \quad (3.29)$$

Since the off-diagonal entries of $\mathbf{K}$ are non-positive and two adjacent sign graphs have opposite signs, $\mathbf{w}_i^T \mathbf{K} \mathbf{w}_j \geq 0$. So (3.29) implies that $R(\mathbf{v}) \leq \lambda$. When $\mathbf{v}$ is an eigenvector corresponding to λ , $R(\mathbf{v}) = \lambda$, i.e., the summation in (3.29) must vanish. Since all the terms $\mathbf{w}_i^T \mathbf{K} \mathbf{w}_j$ are non-negative, each term in the sum must be zero. i.e., if $\mathbf{w}_i^T \mathbf{K} \mathbf{w}_j > 0$, then $c_i = c_j$. This implies that whenever two sign graphs are adjacent, $\mathbf{v}$ either does not vanish on both sign graphs or vanishes on both.

However, their result on multiple eigenvalues:

Theorem 6 ([8], p. 264) *Let $\mathbf{K}$ be a real symmetric irreducible matrix with non-positive off-diagonal entries. If $\mathbf{K}\mathbf{u}_n = \lambda_n\mathbf{u}_n$, then $\mathbf{u}_n$ has at most n sign graphs.*

is not correct. Their theorem implies that even when λ_n is multiple, any eigenvector corresponding to λ_n has at most n sign graphs. But the star example in Figure 3.14 shows that Theorem 6 of Duval and Reiner's is false in general; an eigenvector corresponding to a multiple eigenvalue λ_n can have more than n sign graphs.

We first quote their proof and then show where the proof breaks down. In the end, we extend Lemma 3.14 to matrix pair $(\mathbf{K}, \mathbf{M})$, which sheds more light on the discrete CNLT.

Proof: [[8]] Suppose that $\mathbf{u}_n$ has $n + t$ sign graphs, say $\{\mathcal{G}_j\}_1^{n+t}$ ($t \geq 1$). A new eigenvector $\mathbf{v} = \sum_{i=1}^n c_i \mathbf{w}_i$ corresponding to λ_n can be constructed by eliminating remaining t sign graphs. Then Duval and Reiner claimed that

A. “there must exist some pair of vertices i and j ($i \neq j$) so that $k_{ij}u_i u_j > 0$. That is, there must exist a positive sign graph adjacent to a negative sign graph (where u_i presents the i -th coordinate of $\mathbf{u}_n$).”

If not, there does not exist a pair of adjacent sign graphs. It means that every sign graph connects to zeros only. Hence, each non-zero vertex connects to either zeros or vertices with the same sign. For i such that $u_i \neq 0$, we have

$$k_{ii}u_i + \sum_{j \neq i} k_{ij}u_j = \lambda_n u_i.$$

Multiplying u_i on both sides of the equation, we get

$$k_{ii}u_i^2 + \sum_{j \neq i} k_{ij}u_i u_j = \lambda_n u_i^2.$$

Because $u_i u_j \geq 0$ for any j adjacent to i , we get

$$|u_i| \{ (k_{ii}|u_i| + \sum_{j \neq i} k_{ij}|u_j|) - \lambda_n |u_i| \} = 0.$$

This implies that $|\mathbf{w}_i|^T (\mathbf{K}|\mathbf{u}| - \lambda|\mathbf{u}|) = 0$ for $i = 1, 2, \dots, m$. Hence, after substituting $|\mathbf{v}| = \sum_{i=1}^n |c_i| |\mathbf{w}_i|$ into (3.28), the first sum on the right vanishes. Since all the sign graphs of $\mathbf{u}_n$ connect to zeros only, $|\mathbf{w}_i|^T \mathbf{K} |\mathbf{w}_j| = 0$; thus, the second sum also vanishes. Therefore, from Lemma 3.14, $|\mathbf{v}|^T \mathbf{K} |\mathbf{v}| - \lambda_n |\mathbf{v}|^T |\mathbf{v}| = 0$, i.e., $R(|\mathbf{v}|) = \lambda_n$. Then Duval and Reiner [8] claimed that

B. “ $\mathbf{K}|\mathbf{v}| = \lambda_n |\mathbf{v}|$. ”

Note that $|\mathbf{v}|$ has all non-negative coordinates. Since $\mathbf{K}$ is irreducible, the Perron-Frobenius theorem says that $\lambda_n = \lambda_1$ and $|\mathbf{v}|$ must be strictly positive, contradicting the vanishing of $|\mathbf{v}|$ on $\mathcal{G}_{n+1}$.

Hence we may assume that $\mathcal{G}_n$ and $\mathcal{G}_{n+1}$ are adjacent. There exists at least one edge (i, j_o) in $\mathcal{G}(\mathbf{K})$ such that $i \in V(\mathcal{G}_{n+1})$, $j_o \in V(\mathcal{G}_n)$. Since the eigenvector $\mathbf{v}$ vanishes on $\mathcal{G}_{n+1}$, then $v_i = 0$. The i -th equation of $\mathbf{K}\mathbf{v} = \lambda_n \mathbf{v}$ is

$$0 = \lambda_n v_i = k_{ii}v_i + \sum_{j, (i,j) \in E} k_{ij}v_j = \sum_{j, (i,j) \in E} k_{ij}v_j. \quad (3.30)$$

Duval and Reiner [8] claimed that

C. “the terms v_j in (3.30), which belongs to sign graphs $\mathcal{G}_1, \dots, \mathcal{G}_n$, do not vanish and all have the same sign as v_{j_0} .” If so, let $v_{j_0} < 0$ without loss of generality. From equation (3.30), we have

$$0 = \sum_{j, (i,j) \in E} k_{ij} v_j > 0. \quad (3.31)$$

a contradiction. ■

However, there are flaws in the proof which lie in statements A, B and C. First of all, statement B is not true; $R(|\mathbf{v}|) = \lambda_n$ does not imply that $\mathbf{K}|\mathbf{v}| = \lambda_n|\mathbf{v}|$. The reason is as follows: since $\mathbf{u}_1^T |\mathbf{v}| > 0$, $|\mathbf{v}|$ is not orthogonal to $\mathbf{u}_1$; hence $|\mathbf{v}|$ is not orthogonal to the first $n - 1$ eigenvectors. Thus $|\mathbf{v}|$ is not an eigenvector of λ_n even though $R(|\mathbf{v}|) = \lambda_n$. The false statement B implies that statement A is false. Figure 3.16 illustrates a case where statement A fails. In this example, matrix $\mathbf{K}$ is the adjacency matrix of the graph shown in Figure 3.16. $\mathbf{K}$ has a 2-fold eigenvalue, $\lambda_3 = \lambda_4$. An eigenvector $\mathbf{u}_3$ corresponding to λ_3 has *four* strict sign graphs, but all the sign graphs are adjacent to zero vertices only. This shows that when $\mathbf{u}_n$ has more than n sign graphs, it is not necessarily to have a positive sign graph adjacent to a negative sign graph.

Even in the situation that statement A is true, the conclusion in Theorem 6 is still false. This is because statement C is wrong. Since $\mathbf{v}$ vanishes on $\mathcal{G}_{n+1}$, by Lemma 3.14, $\mathbf{v}$ vanishes on $\mathcal{G}_n$ as well, i.e., $c_n = c_{n+1} = 0$ which implies $v_{j_0} = 0$. The same reasoning can be applied to other terms v_j in (3.30). Thus, all the terms

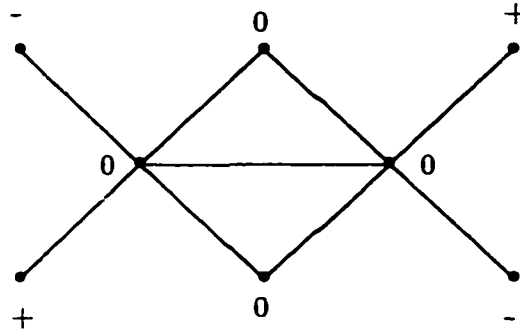


Figure 3.16: λ_3 is a 2-fold eigenvalue. An eigenvector $\mathbf{u}_3$ corresponding to λ_3 has *four* strict sign graphs, but all of them are adjacent to zero vertices only.

v_j in (3.30) are zero; statement C is wrong. The contradiction (3.31) does not occur. The eigenvector whose signs are shown in Figure 3.17 is a counterexample to statement C. This eigenvector corresponds to the fourth eigenvalue of the matrix $\mathbf{A}$

$$\mathbf{A} = \begin{pmatrix} 3 & 0 & 0 & -1 & 0 & -1 & 0 & 0 \\ 0 & 3 & 0 & -1 & -1 & 0 & -1 & 0 \\ 0 & 0 & 3 & 0 & -1 & 0 & 0 & -1 \\ -1 & -1 & 0 & 4 & 0 & -1 & -1 & 0 \\ 0 & -1 & -1 & 0 & 4 & 0 & -1 & -1 \\ -1 & 0 & 0 & -1 & 0 & 3 & 0 & 0 \\ 0 & -1 & 0 & -1 & -1 & 0 & 3 & 0 \\ 0 & 0 & -1 & 0 & -1 & 0 & 0 & 3 \end{pmatrix}.$$

Note that $\mathbf{A}$ is a symmetric, irreducible matrix with non-positive off-diagonal entries, which satisfies the condition in Theorem 6 of [8]. $\mathbf{A}$ has a 3-fold eigenvalue,

$\lambda_4 = \lambda_5 = \lambda_6 = 4$. The eigenspace of λ_4 consists of those vectors $\mathbf{u}$ satisfying the equations

$$u_4 = 0 = u_5, \quad u_1 + u_6 = 0 = u_2 + u_7 = u_3 + u_8.$$

Figure 3.17 shows that an eigenvector $\mathbf{u}_4$ of $\mathbf{A}$ has *six* sign graphs, each of which is adjacent to another sign graph. This shows that even if there is a positive sign graph adjacent to a negative sign graph, $\mathbf{u}_n$ can still have more than n sign graphs.

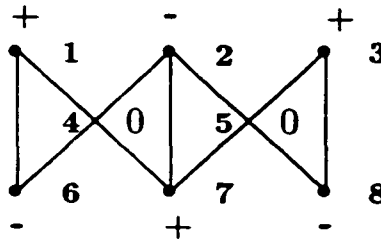


Figure 3.17: λ_4 is an eigenvalue with multiplicity 3. There is an eigenvector $\mathbf{u}_4$ corresponding to λ_4 such that every sign graph of $\mathbf{u}_4$ is adjacent to another sign graph. But $\mathbf{u}_4$ has six sign graphs.

Therefore, due to the flaws in their argument, the statement of Theorem 6 in Duval and Reiner [8] does not hold.

Although Duval and Reiner [8] made a false conclusion concerning the discrete CNLT for multiple eigenvalues, We can extend Lemma 3.14 to matrix pair $(\mathbf{K}, \mathbf{M})$.

Lemma 3.15 *Let $\mathbf{u} \in \mathbb{R}^{N \times 1}$. Suppose that $\mathbf{u}$ has the sign graphs $\mathcal{G}_1, \dots, \mathcal{G}_m$. Define $\mathbf{v} = \sum_{i=1}^m c_i \mathbf{w}_i$ in the notation of (3.15). Then*

$$\mathbf{v}^T \mathbf{K} \mathbf{v} - \lambda \mathbf{v}^T \mathbf{M} \mathbf{v} = \sum_i c_i^2 \mathbf{w}_i^T (\mathbf{K} \mathbf{u} - \lambda \mathbf{M} \mathbf{u}) - \sum_{i < j} (c_i - c_j)^2 (\mathbf{w}_i^T \mathbf{K} \mathbf{w}_j - \lambda \mathbf{w}_i^T \mathbf{M} \mathbf{w}_j) \quad (3.32)$$

Proof: We compute $\mathbf{v}^T \mathbf{K} \mathbf{v}$ first.

$$\begin{aligned} \mathbf{v}^T \mathbf{K} \mathbf{v} &= \mathbf{v}^T \sum_j c_j \mathbf{K} \mathbf{w}_j \\ &= \sum_{i,j} c_i c_j \mathbf{w}_i^T \mathbf{K} \mathbf{w}_j \\ &= \sum_i c_i^2 \mathbf{w}_i^T \mathbf{K} \mathbf{w}_i + \sum_{i \neq j} c_i c_j \mathbf{w}_i^T \mathbf{K} \mathbf{w}_j. \end{aligned}$$

Since $\mathbf{K} \mathbf{u} = \sum_j \mathbf{K} \mathbf{w}_j$, $\mathbf{K} \mathbf{w}_i = \mathbf{K} \mathbf{u} - \sum_{j \neq i} \mathbf{K} \mathbf{w}_j$. Then we have

$$\begin{aligned} \mathbf{v}^T \mathbf{K} \mathbf{v} &= \sum_i c_i^2 \mathbf{w}_i^T (\mathbf{K} \mathbf{u} - \sum_{j \neq i} \mathbf{K} \mathbf{w}_j) + \sum_{i \neq j} c_i c_j \mathbf{w}_i^T \mathbf{K} \mathbf{w}_j \\ &= \sum_i c_i^2 \mathbf{w}_i^T \mathbf{K} \mathbf{u} - \sum_i c_i^2 \mathbf{w}_i^T \sum_{j \neq i} \mathbf{K} \mathbf{w}_j + \sum_{i \neq j} c_i c_j \mathbf{w}_i^T \mathbf{K} \mathbf{w}_j \\ &= \sum_i c_i^2 \mathbf{w}_i^T \mathbf{K} \mathbf{u} - \sum_{i \neq j} c_i^2 \mathbf{w}_i^T \mathbf{K} \mathbf{w}_j + \sum_{i \neq j} c_i c_j \mathbf{w}_i^T \mathbf{K} \mathbf{w}_j \\ &= \sum_i c_i^2 \mathbf{w}_i^T \mathbf{K} \mathbf{u} - \sum_{i < j} (c_i - c_j)^2 \mathbf{w}_i^T \mathbf{K} \mathbf{w}_j. \end{aligned}$$

Similarly, $\mathbf{v}^T \mathbf{M} \mathbf{v}$ has the same form as $\mathbf{v}^T \mathbf{K} \mathbf{v}$. Thus, the lemma follows by subtracting $\lambda \mathbf{v}^T \mathbf{M} \mathbf{v}$ from $\mathbf{v}^T \mathbf{K} \mathbf{v}$. ■

Particularly, when $\mathbf{Ku} = \lambda \mathbf{Mu}$, the first sum in (3.32) is zero. We have a simpler formula of (3.32)

$$\mathbf{v}^T \mathbf{K} \mathbf{v} - \lambda \mathbf{v}^T \mathbf{M} \mathbf{v} = - \sum_{i < j} (c_i - c_j)^2 (\mathbf{w}_i^T \mathbf{K} \mathbf{w}_j - \lambda \mathbf{w}_i^T \mathbf{M} \mathbf{w}_j). \quad (3.33)$$

Recall that when $SG(\mathbf{u}_n) = m > n$, we can construct a new eigenvector $\mathbf{v}$ of λ_n from n of those sign graphs and $\mathbf{v}$ vanishes on the remaining sign graphs. Since $\mathbf{K} \mathbf{v} - \lambda_n \mathbf{M} \mathbf{v} = \mathbf{0}$, the sum in (3.33) is zero. Since $\mathbf{w}_i^T \mathbf{K} \mathbf{w}_j \geq 0$ and $-\lambda \mathbf{w}_i^T \mathbf{M} \mathbf{w}_j \geq 0$, each term in the sum is zero. That is, if two sign graphs of $\mathbf{u}_n$ are adjacent. Lemma 3.15 says that either both remain sign graphs of $\mathbf{v}$ or both become zero.

With Lemma 3.15, we are able to deduce more on sign graphs of the eigenvectors. We define *connected sets* of sign graphs of $\mathbf{u}_n$ to be the connected components in the new graph obtained from $\mathcal{G}$ by removing zero vertices and their incident edges, and denote those connected sets as $\mathcal{C}_1, \mathcal{C}_2, \dots$. A connected set is a set of sign graphs, in which any two sign graphs are connected by a path consisting of nonzero vertices. Thus if $\mathbf{v}$ vanishes on one sign graph in a connected set $\mathcal{C}$, $\mathbf{v}$ must vanish identically on $\mathcal{C}$. We can immediately conclude the following.

Corollary 3.16 *If the sign graphs of $\mathbf{u}_n$ form only one connected set, then $\mathbf{u}_n$ has at most n sign graphs.*

Proof: If not, $SG(\mathbf{u}_n) = m > n$. By Corollary 3.8, a new eigenvector $\mathbf{v} = \sum_{j=1}^n c_j \mathbf{w}_j$ can be constructed. $\mathbf{v}$ vanishes on sign graphs $\{\mathcal{G}_j\}_{n+1}^m$; but all the sign graphs of $\mathbf{u}_n$ are in one connected set, Lemma 3.14 implies that $\mathbf{v} = \mathbf{0}$, a contradiction. ■

Corollary 3.17 *If $\mathbf{u}_n$ has $n + i$ ($i \geq 1$) sign graphs, then $\mathbf{u}_n$ must have at least $i + 1$ connected sets of sign graphs, and each connected set contains at most n sign graphs.*

Proof: If not, we can construct a new eigenvector $\mathbf{v}$ which vanishes on at least one sign graph in each connected set. Since there are at most i connected sets, from Lemma 3.14, $\mathbf{v}$ is identically zero in $\mathcal{G}$, a contradiction. The second statement comes directly from Corollary 3.16. ■

If $\mathbf{u}_n$ has more than n sign graphs, then Corollary 3.17 implies that $\mathbf{u}_n$ has at least two connected sets of sign graphs. If $\mathcal{G}(\mathbf{K}, \mathbf{M})$ is connected, there exist some zeros that connect those connected sets. Suppose that $\mathbf{u}_n$ has two connected sets $\mathcal{C}_1$ and $\mathcal{C}_2$ and one of them consists of n sign graphs, say $\mathcal{C}_1$. By Corollary 3.8, a new eigenvector $\mathbf{v}$ with at most n sign graphs can be constructed by eliminating $\mathbf{v}$ on some sign graphs of $\mathcal{C}_1$. Lemma 3.14 says that $\mathbf{v}$ has to be zero in $\mathcal{C}_1$; $\mathbf{v}$ is non-zero on $\mathcal{C}_2$. There must exist a zero in $\mathbf{v}$ that is adjacent to at least one positive vertex and one negative vertex in $\mathcal{C}_2$. Thus, $\mathcal{C}_2$ must contain at least 2 sign graphs, which implies $SG(\mathbf{u}_n) \geq n + 2$.

As we can see, when λ_n is multiple, the sign property of the eigenvectors becomes more complicated as not all of the eigenvectors corresponding to λ_n have at most n sign graphs. When $SG(\mathbf{u}_n) > n$, there is also certain pattern for the structures of the sign graphs in $\mathbf{u}_n$. Thus, certainties and uncertainties make the case of multiple eigenvalue more interesting to study.

Chapter 4

Qualitative Properties of the Linear Combinations of Eigenfunctions/FEM solution

In this chapter, we will discuss the qualitative properties of finite linear combinations of eigenfunctions of the Helmholtz equation. The linear combinations of eigenfunctions inherit some basic properties of eigenfunctions such as the unique continuation property; however, they are quite different from the eigenfunctions. We will show the similarities and differences between a linear combination and a single eigenfunction in Section 4.1.

Note that for the linear combinations, the study of CNLT property is replaced by that of CHC property. We present some interesting counterexamples showing that CHC is generally false for linear combinations of the eigenfunctions and also for those of the FEM solutions. A restricted theorem is proved, which holds for

both continuous and discrete cases.

Even though CHC is not true in general, we conjecture that CHC is true for certain convex domains, such as rectangles and circles. We study CHC on square domains. Without loss of generality, we consider the square membrane $R_{[0,\pi]} \equiv [0, \pi] \times [0, \pi]$ with mass density $\rho \equiv 1$. It is well known that the eigenvalues of (1.1) on $R_{[0,\pi]}$ with Dirichlet boundary conditions are

$$\lambda_{n,m} = n^2 + m^2$$

for $n, m = 1, 2, \dots$; the associated (not normalized) eigenfunctions are the products

$$u_{n,m} = \sin(nx)\sin(my).$$

We are able to prove that CHC is true for linear combinations of the first fourteen eigenfunctions and also for certain linear combinations of the eigenfunctions.

4.1 Nodal sets of the linear combinations

Let w denote an arbitrary linear combination of the first n eigenfunctions $\{u_i\}_1^n$, *i.e.*

$$w = \sum_{i=1}^n c_i u_i,$$

where the c_i 's are arbitrary real numbers.

Since each eigenfunction u_i is analytic, w is analytic in Ω also. The proof of the unique continuation property for the finite combinations is then identical to that for the eigenfunctions.

Suppose $\Omega \subset \mathbb{R}^m$. The nodal set of a combination of eigenfunctions exhibits greater variety than that of a single eigenfunction. The following example shows that in $\mathbb{R}^2$, the nodal set of a combination can contain isolated nodal points.

Example 4.1 Consider linear combinations of the eigenfunctions on $R_{[0,\pi]}$. Take a function as $w = au_{11} + bu_{13} + cu_{31}$. That is,

$$\begin{aligned} w &= a \sin x \sin y + b \sin x \sin 3y + c \sin y \sin 3x \\ &= \sin x \sin y \{a + b(4\cos^2 y - 1) + c(4\cos^2 x - 1)\} \end{aligned}$$

Choose $a = 2$, $b = 1$ and $c = 1$. Then $w = 4 \sin x \sin y (\cos^2 x + \cos^2 y)$ is zero only at $(\pi/2, \pi/2)$, which is an isolated nodal point. In $\mathbb{R}^2$, the nodal set of w can

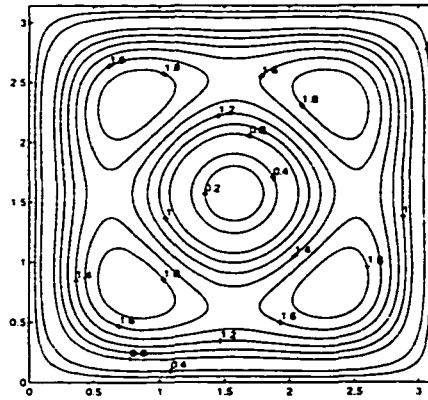


Figure 4.1: The digram shows the contours of the combination $2 * \sin x \sin y + \sin x \sin 3y + \sin y \sin 3x$, where point $(\pi/2, \pi/2)$ is an isolated nodal point.

have isolated nodal points in addition to nodal lines. In $\mathbb{R}^3$, the nodal set may

contain isolated nodal points and nodal lines in addition to nodal surfaces. In $\mathbb{R}^m$, the nodal set of a combination may consist of hypersurfaces of various dimensions, $i = 1, 2, \dots, m - 1$. Thus, the property B' of the eigenfunctions does not always hold for the linear combinations.

In $\mathbb{R}^2$. If ρ is analytic and an eigenfunction u has m nodal lines ($m \geq 2$) meeting at an interior point $\mathbf{p}$, then the first m terms in power series of u at $\mathbf{p}$ are zero, i.e., $v_0 = 0 = v_1 = \dots = v_{m-1}$ in a small neighborhood $B(\mathbf{p})$ of $\mathbf{p}$. $\Delta v_m(x, y) = -\lambda \rho v_{m-2}(x, y) = 0$ in $B(\mathbf{p})$, i.e., u behaves like a harmonic function near $\mathbf{p}$. This forced us to conclude that the nodal lines meeting at an interior point form an equiangular system. But this may not be true for a linear combination of eigenfunctions. When m nodal lines of a combination $w = \sum_{i=1}^n c_i u_i$ meet at an interior point $\mathbf{p}$, then the first m terms in power series of w at $\mathbf{p}$ are zero, i.e., $w_0 = 0 = w_1 = \dots = w_{m-1}$, in a small neighborhood $B(\mathbf{p})$ of $\mathbf{p}$. But this does not imply that $\Delta w_m(x, y) = -\rho \sum_{i=1}^n \lambda_i c_i u_{i,m-2}(x, y)$ is zero in some neighborhood of $\mathbf{p}$, where $u_{i,m-2}(x, y)$ is the $(m - 2)$ th term in the power series of the i -th eigenfunction u_i . Hence the nodal lines may not have normal crossings as those of an eigenfunction. For instance, when two nodal lines of an eigenfunction meet at an interior point, they meet at right angles; however, when two nodal lines of a combination meet at an interior point, the nodal lines can have various types of crossings.

Example 4.2 Take the Dirichlet eigenfunctions on the square $R_{[0,\pi]}$ as an example. First, consider the following combination

$$w = 4\sin x \sin(3y) - \sin(5x) \sin y - 3\sin(3x) \sin y + 2\sin x \sin y. \quad (4.1)$$

Expanding the sine in terms of cosines multiplied by $\sin x$ or $\sin y$, we have

$$\begin{aligned} w &= \sin x \sin y \{4(4\cos^2 y - 1) - (16\cos^4 x - 12\cos^2 x + 1) - 3(4\cos^2 x - 1) + 2\} \\ &= 16\sin x \sin y (\cos^2 y - \cos^4 x). \end{aligned}$$

The nodal lines of w are $\cos y = \pm \cos^2 x$, which meet at the point $\mathbf{p} = (\pi/2, \pi/2)$. Note that w , its first derivatives, w_{xx} and w_{xy} vanish at $\mathbf{p}$, but $w_{yy}(\mathbf{p}) = 32$. Hence $\Delta w(\mathbf{p}) \neq 0$. These two nodal lines do not meet at right angles, see Figure 4.2.

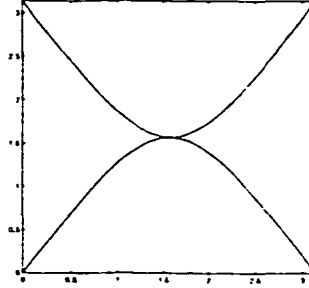


Figure 4.2: The two nodal lines of the linear combination w in (4.1) do not meet at right angles.

Apart from that, nodal lines of a combination can even have cusps. Take a linear combination

$$\begin{aligned} w &= \sin x \sin(4y) - 2\sin(3x) \sin y + 2\sin(x) \sin(2y) - 2\sin x \sin y \quad (4.2) \\ &= \sin x \sin y \{(8\cos^3 y - 4\cos y) - 2(4\cos^2 x - 1) + 4\cos y - 2\} \\ &= 8\sin x \sin y (\cos^3 y - \cos^2 x). \end{aligned}$$

We can see that the nodal set of w is determined by $\cos^3 y = \cos^2 x$. In this case, w , its first derivatives, w_{yy} and w_{xy} vanish at $\mathbf{p}$, but $w_{xx}(\mathbf{p}) = -16$. which has a

cuspidal at $(\pi/2, \pi/2)$. Thus, in some neighborhood of $\mathbf{p}$,

$$w(x, y) = -8(x - \frac{\pi}{2})^2 + o(|(x, y) - \mathbf{p}|^2).$$

So the nodal line $\cos^3 y = \cos^2 x$ does not cross itself at $\mathbf{p}$ but instead it bends in such a fashion that it is tangent to itself. see Figure 4.3. Thus, the nodal line has a cusp.

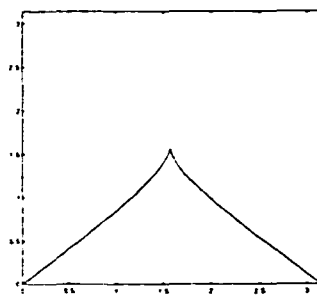


Figure 4.3: Nodal lines of a linear combination of eigenfunctions can have cusps.

However, there is one case that nodal lines meeting at a point do form an equiangular system. We classify the intercepts of nodal lines of a combination into two kinds: the intercept at which every u_i is zero is called a *common nodal point*; otherwise, it is a *uncommon nodal point*. In Example 4.2, point $(\pi/2, \pi/2)$ is a uncommon nodal point of the linear combination in (4.1), and the nodal lines meeting at $(\pi/2, \pi/2)$ do not meet at right angles. Thus, nodal lines meeting at a uncommon point may not form an equiangular system. However, when two nodal lines of w meet at a common nodal point, the nodal lines do meet at right angles.

Theorem 4.1 *Let $\rho(\mathbf{x})$ in (1.1) be analytic. If two nodal lines of a combination $w = \sum_{i=1}^n c_i u_i$ meet at a common nodal point, they must meet at right angles.*

Proof: Without loss of generality, let $\mathbf{p} = (0, 0) \in \Omega$. Each u_i has a Taylor expansion in a small neighborhood of $\mathbf{p}$, say $B_{r_i}(\mathbf{p})$. Then $w = \sum_{i=1}^n c_i u_i$ also has a Taylor expansion around $\mathbf{p}$

$$w = \sum_{j=0}^{\infty} w_j = \sum_{j=0}^{\infty} \frac{1}{j!} \left(x \frac{\partial}{\partial x} + y \frac{\partial}{\partial y} \right)^j w(\mathbf{p}),$$

for $(x, y) \in B_r(\mathbf{p}) = \cap_{i=1}^n B_{r_i}(\mathbf{p})$. By assumption, $\mathbf{p}$ is a double singular point, i.e., $w_0 \equiv 0 \equiv w_1$ in $B_r(\mathbf{p})$. As $\Delta u_i + \lambda_i \rho u_i = 0$ in Ω , w satisfies

$$\Delta w + \rho \sum_{i=1}^n \lambda_i c_i u_i = 0.$$

Expanding w and the u_i in power series in $B_r(\mathbf{p})$, we have a polynomial in x and y on the left hand side. The constant term of the polynomial is

$$\Delta w_2 + \rho(\mathbf{p}) \sum_{i=1}^n \lambda_i c_i u_i(\mathbf{p}) = 0$$

in $B_r(\mathbf{p})$. Since $\mathbf{p}$ is a common nodal point of the u_i , $\sum_{i=1}^n \lambda_i c_i u_i(\mathbf{p}) = 0$. Thus, in $B_r(\mathbf{p})$,

$$\Delta w_2 = 0, \quad \text{i.e.,} \quad w_{xx}(\mathbf{p}) + w_{yy}(\mathbf{p}) = 0. \quad (4.3)$$

Again, w_2 is a harmonic function in $B_r(\mathbf{p})$; the nodal lines of w behave like the nodal lines of a harmonic function around $\mathbf{p}$. Thus, two nodal lines of w passing through $\mathbf{p}$ must meet at right angles. ■

Furthermore, when three or more nodal lines meet at a point, the behaviour becomes even more irregular. The following example illustrates that even when three nodal lines meet at a common nodal point, they do not necessarily form an equiangular system.

Example 4.3 Again, in the square domain $R_{[0,\pi]}$, take the linear combination

$$\begin{aligned} w &= (\sin(3x)\sin(y) - \sin(x)\sin(3y)) + (\sin(3x)\sin(2y) - \sin(2x)\sin(3y)) \\ &\quad + 2(\sin(4x)\sin(3y) - \sin(3x)\sin(4y)) \\ &= 2\sin(x)\sin(y)(2\cos(x) - 1)(2\cos(y) - 1)(\cos(y) - \cos(x)) \\ &\quad (8\cos(x)\cos(y) + 4\cos(y) + 4\cos(x) + 3). \end{aligned}$$

Equating w to 0, we see that there are three nodal lines passing through the point $(\frac{\pi}{3}, \frac{\pi}{3})$. They are

$$x = \frac{\pi}{3}, \quad y = \frac{\pi}{3}, \quad \text{and } x = y.$$

However, these nodal lines do not form an equiangular system even though $(\frac{\pi}{3}, \frac{\pi}{3})$ is a common nodal point of these eigenfunctions, see Figure 4.4.

If Ω is a convex domain, Theorem 4.1 can be generalized to the case when a nodal line intersects the boundary at a point. For any $\mathbf{p} \in \partial\Omega$, we denote by $\Gamma(\mathbf{p})$ the smallest open infinite sector contained in Ω with vertex at $\mathbf{p}$.

Corollary 4.2 *Let $\Omega \subset \mathbb{R}^2$ be convex and ρ be analytic. If a nodal line of $w = \sum_{i=1}^n u_i$ intersects the boundary at point $\mathbf{p}$, then the tangent line of the nodal line divide $\Gamma(\mathbf{p})$ into 2 sectors of equal amplitude.*

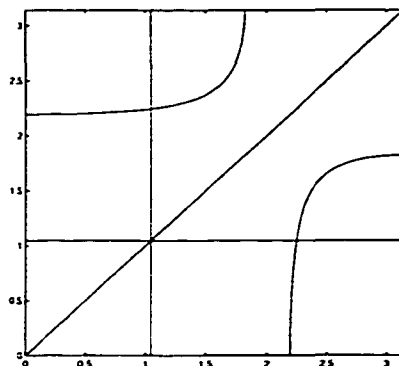


Figure 4.4: The nodal lines of the linear combination $w = (u_{31} - u_{13}) + (u_{32} - u_{23}) + 2(u_{34} - u_{43})$ meet at a common nodal point of the eigenfunctions, but do not form an equiangular system.

Proof: We use the same technique as in Alessandrini [1]. Again, let $\mathbf{p} = (0, 0) \in \partial\Omega$. Define $D_R = \Gamma(\mathbf{p}) \cap B_R(\mathbf{p})$, where $R > 0$ is sufficiently small so that $\partial D_R \cap \Omega \neq \emptyset$. Perform the conformal mapping so that D_R is transformed to a half disk centered at the origin and contained in the upper half plane. By reflecting w oddly across the real axes, the result follows immediately from Theorem 4.1. ■

As a conclusion, linear combinations of the eigenfunctions cannot vanish on any non-empty open subset of $\Omega \subset \mathbb{R}^m$. The nodal set of a combination has more varieties than that of an eigenfunction. It may consist of hypersurfaces of various dimensions, $i = 1, 2, \dots, m-1$. In $\mathbb{R}^2$, nodal set of a combination can have isolated nodal points or cusps while nodal set of an eigenfunction cannot. For a combination, when two nodal lines meet at a common nodal point or one nodal line meets at the boundary of a convex domain, an equiangular system is formed; in other cases, the nodal lines do not necessarily form an equiangular system when they meet at a point. Therefore, as we see, the behaviour of finite linear combinations is quite

different from that of eigenfunctions.

4.2 CHC is false in general

It is also of interest to study CHC property of linear combinations of the eigenfunctions. As mentioned in Chapter 1, Arnol'd [2] first noticed that CHC is false, but no counterexample was provided. We here present a few interesting numerical counterexamples, computed by MATLAB PDE Toolbox, which shows that $tu_1 + u_2$ ($t \geq 0, u_1 > 0$) can have three, four or five nodal domains.

Let us first start with a vibrating membrane with fixed boundary. It consists of two similar rectangles with the ratio 2 : 1 as shown in Figure 4.5. Using MATLAB, we observe that when $t = 0$, the nodal line of the combination $tu_1 + u_2$ is that of the second eigenfunction, which divides the domain into two nodal domains; as t increases, the nodal line of $tu_1 + u_2$ goes upwards and still divides the domain into two subdomains; however, when $t \doteq 0.96$, the nodal lines breaks into two and $tu_1 + u_2$ has now three nodal domains; as t keeps increasing, both nodal lines move towards the boundary as in Figure 4.5 and the combination has **three** nodal domains; eventually, when t is sufficiently large, $tu_1 + u_2$ behaves more like u_1 and the nodal lines vanish at the boundary leaving $tu_1 + u_2$ with one nodal domain. Hence, linear combinations of the first two eigenfunctions can have more than two nodal domains.

The domains in Figure 4.6 and 4.7 are similar to that in Figure 4.5, but consist of more rectangles. The smaller rectangles are attached to the bigger ones clockwise. With three rectangles, we found a linear combination of the first two eigenfunctions

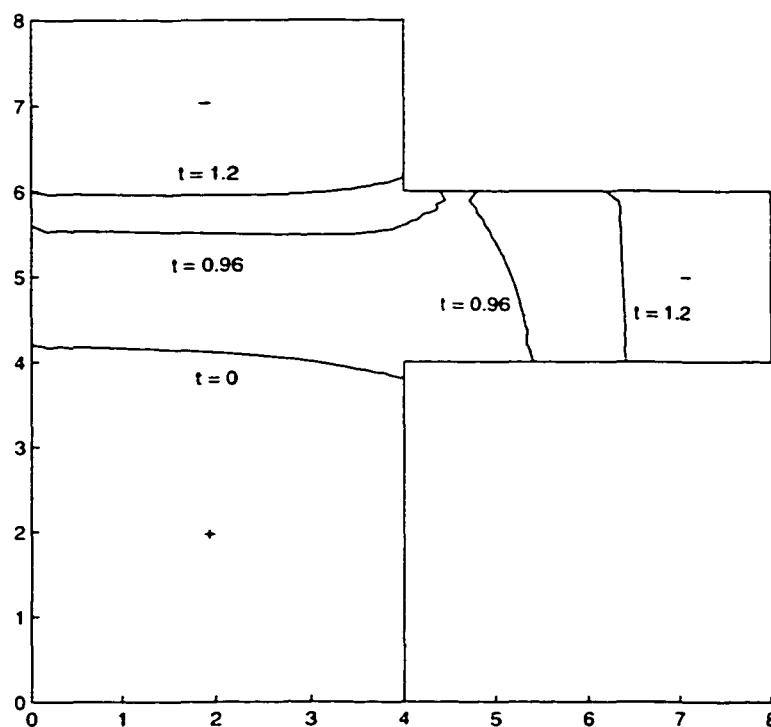


Figure 4.5: The nodal lines of the linear combinations $t \cdot u_1 + u_2$ for $t = 0, 0.96$ and 1.2 . Linear combination of the first two eigenfunctions may have more than 2 nodal domains

which has **four** nodal domains: one positive, three negative. With four rectangles, the linear combination $1.46u_1 + u_2$ has **five** nodal domains: one positive and four negative. Those examples show that CHC is false. In fact, if more rectangles are added in similar fashion, we conjecture that a linear combination of the first two eigenfunctions may have arbitrarily many nodal domains.

A counterexample can also be found for Neumann boundary conditions. Figure 4.8 shows that there is a linear combination of the first two Neumann eigenfunctions that has more than two nodal domains.

Notice that the domains in the counterexamples are all non-convex, as are all the known counterexamples to Mark Kac's question: can one hear the shape of a drum? As we have not found a counterexample with a convex domain, we conjecture that CHC may be true for convex domains, particularly circles or rectangular domains.

The discrete analogue of CHC does not hold either. A simple counterexample is the Dirichlet eigenvectors on the FEM mesh with four interior vertices, illustrated in Figure 4.9. The linear combination $0.7u_1 + u_2$ has three sign graphs.

4.3 A revised CHC

However, there is a restricted theorem for the particular linear combination $tu_1 + u_n$ ($t \geq 0$, $u_1 > 0$). We consider the theorem as a substitute of CHC, which hold in both continuous and discrete cases. Without loss of generality, we can always assume that the first eigenfunction u_1 in continuous case and the first eigenvector u_1 in discrete case are strictly positive.

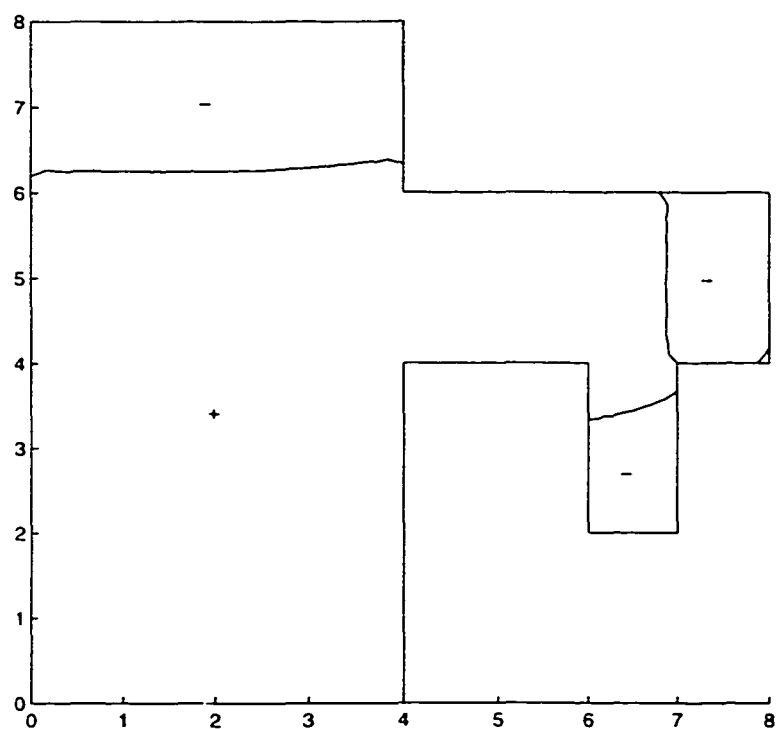


Figure 4.6: The nodal lines of the linear combination $1.35 * u_1 + u_2$ divide the domain into four nodal domains: one positive and three negative.

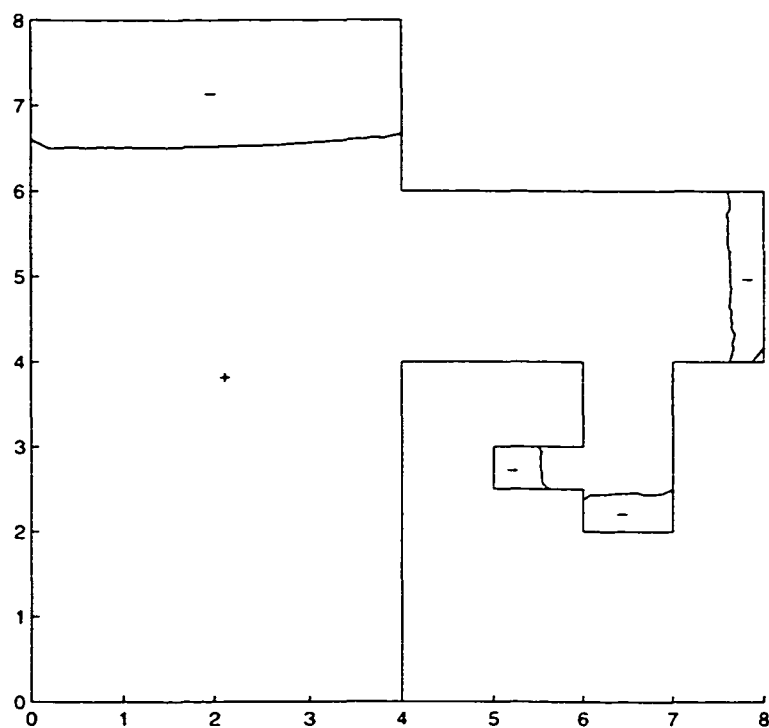


Figure 4.7: The nodal lines of the linear combination $1.46 * u_1 + u_2$ divide the domain into five nodal domains: one positive and four negative.

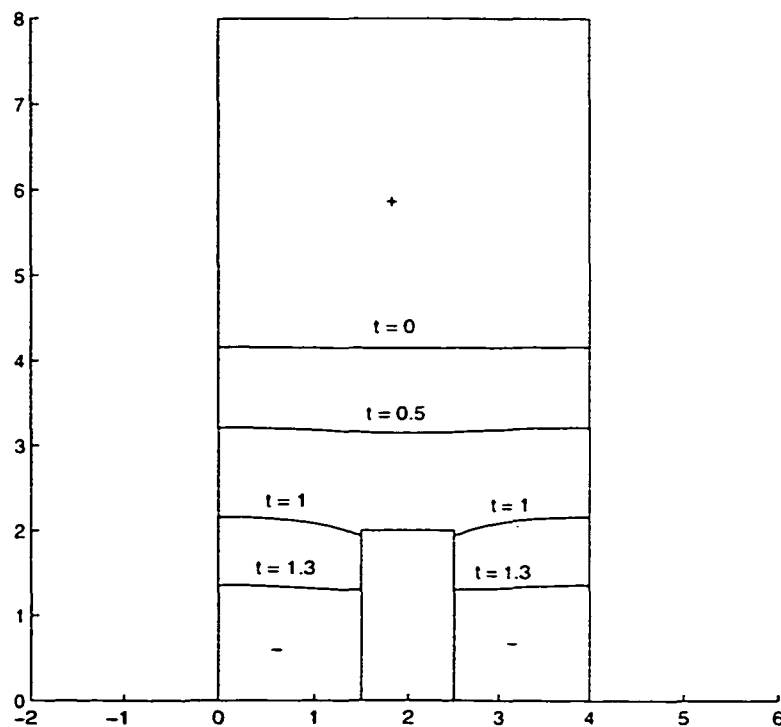


Figure 4.8: The nodal lines of the linear combinations $t \cdot u_1 + u_2$ for $t = 0, 0.5, 1$ and 1.3 . Linear combination of the first two Neumann eigenfunctions can have more than 2 nodal domains.

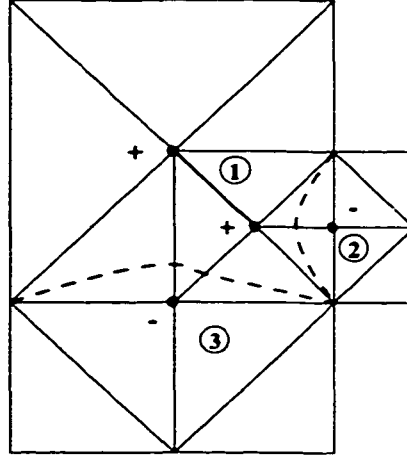


Figure 4.9: The linear combination $0.7u_1 + u_2$ of the first eigenvectors has three sign graphs. CHC does not hold in discrete case either.

Theorem 4.3 *Suppose that Ω is connected. The linear combination $w = tu_1 + u_n$ ($t \geq 0$, $n \geq 2$) cannot have more than $n - 1$ **positive** nodal domains.*

Proof: Without loss of generality, let $t > 0$ (when $t = 0$, it follows directly from CNLT). Suppose that w has r positive nodal domains, say $\{\Omega_i\}_1^r$ ($r \geq n$). Define

$$v_i(x, y) = \begin{cases} w(x, y), & (x, y) \in \Omega_i \\ 0, & \text{otherwise} \end{cases}$$

for $i = 1, 2, \dots, r$. Note that the v_i satisfy

$$\int_{\Omega_i} -\nabla v_i \cdot \nabla v_i + \rho(tu_1 + u_n)(t\lambda_1 u_1 + \lambda_n u_n) d\mathbf{x} = 0,$$

for $i = 1, \dots, r$. Consider a linear combination of $\{v_j\}_1^n$

$$u = \sum_{i=1}^n a_i v_i,$$

where the coefficients a_i are chosen so that u is orthogonal to the first $n - 1$ eigenfunctions. By the minimax principle for the eigenvalues,

$$R(u) \geq \lambda_n. \quad (4.4)$$

On the other hand, we compute

$$\begin{aligned} \int_{\Omega} \nabla u \cdot \nabla u d\mathbf{x} &= \sum_{i=1}^n a_i^2 \int_{\Omega_i} \nabla v_i \cdot \nabla v_i d\mathbf{x} \\ &= \sum_{i=1}^n a_i^2 \int_{\Omega_i} \rho(tu_1 + u_n)(t\lambda_1 u_1 + \lambda_n u_n) d\mathbf{x} \\ &= \sum_{i=1}^n a_i^2 \int_{\Omega_i} \rho(t^2 \lambda_1 u_1^2 + t(\lambda_1 + \lambda_n)u_1 u_n + \lambda_n u_n^2) d\mathbf{x}, \end{aligned}$$

and calculate

$$\begin{aligned} \int_{\Omega} \rho u^2 d\mathbf{x} &= \sum_{i=1}^n a_i^2 \int_{\Omega_i} \rho v_i^2 d\mathbf{x} \\ &= \sum_{i=1}^n a_i^2 \int_{\Omega_i} \rho(t^2 u_1^2 + 2tu_1 u_n + u_n^2) d\mathbf{x}. \end{aligned}$$

Then we compute

$$\begin{aligned}
 (\lambda_n - R(u)) \int_{\Omega} \rho u^2 d\mathbf{x} &= \lambda_n \int_{\Omega} \rho u^2 d\mathbf{x} - \int_{\Omega} \nabla u \cdot \nabla u d\mathbf{x} \\
 &= t(\lambda_n - \lambda_1) \sum_{i=1}^n a_i^2 \int_{\Omega_i} \rho (tu_1^2 + u_1 u_n) d\mathbf{x} \\
 &= t(\lambda_n - \lambda_1) \sum_{i=1}^n a_i^2 \int_{\Omega_i} \rho u_1 (tu_1 + u_n) d\mathbf{x} \\
 &= t(\lambda_n - \lambda_1) \sum_{i=1}^n a_i^2 \int_{\Omega_i} \rho u_1 u d\mathbf{x}
 \end{aligned}$$

By the definition of the v_i , we get

$$(\lambda_n - R(u)) \int_{\Omega} \rho u^2 d\mathbf{x} = t(\lambda_n - \lambda_1) \sum_{i=1}^n a_i^2 \int_{\Omega_i} \rho u_1 v_i d\mathbf{x}. \quad (4.5)$$

For convenience, we define

$$b_i \equiv \int_{\Omega_i} \rho u_1 v_i d\mathbf{x} > 0$$

for $i = 1, \dots, n$. Rewrite (4.5) as

$$(\lambda_n - R(u)) \int_{\Omega} \rho u^2 d\mathbf{x} = t(\lambda_n - \lambda_1) \sum_{i=1}^n a_i^2 b_i > 0,$$

This leads to $R(u) < \lambda_n$, a contradiction to (4.4). ■

This theorem cannot be generalized to linear combinations of the first n ($n > 2$) eigenfunctions. In other words, $\sum_{i=1}^n c_i u_i$ ($c_i > 0$, $u_i > 0$) may have more than $n - 1$ positive nodal domains. Figure 4.10 and 4.11 show that linear combinations

of the first three eigenfunctions with $c_1 > 0$ and $u_1 > 0$ can have three or four positive nodal domains.

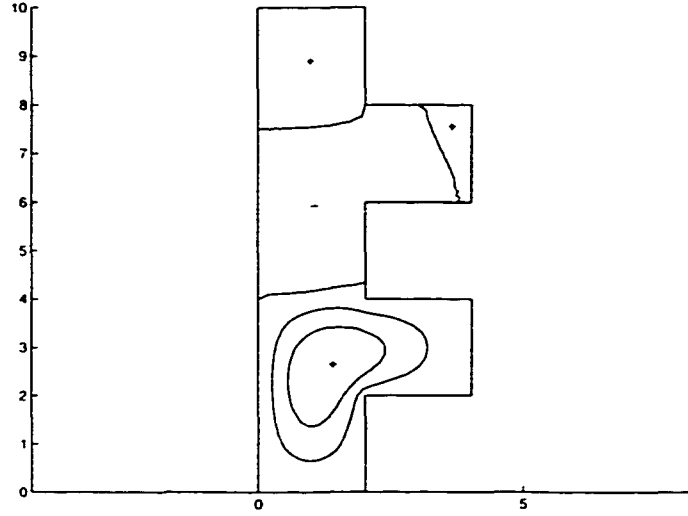


Figure 4.10: $\cos(\frac{8}{25}\pi)u_1 + u_2 + \sin(\frac{8}{25}\pi)u_3$ has three positive nodal domains.

There is also a matrix analogue of Theorem 4.3 for FEM models.

Theorem 4.4 *Let $\mathbf{K}$ and $\mathbf{M}$ in Theorem 3.4 be irreducible. For any $t > 0$ and $n \geq 2$, the combination $t\mathbf{u}_1 + \mathbf{u}_n$ has at most $n - 1$ positive sign graphs, where $\mathbf{u}_1 > 0$.*

Proof: The idea of the proof is similar to that of Theorem 3.13. Suppose that $\mathbf{u} = t\mathbf{u}_1 + \mathbf{u}_n$ has r positive sign graphs where $r \geq n$. Partition $\mathbf{u}$ in the same way as in Theorem 3.13, where $\mathbf{v}_i > 0$, $i = 1, 2, \dots, r$ and $\mathbf{v}_{r+1} \geq 0$. Take the linear combination

$$\mathbf{v} = \sum_{i=1}^r a_i \mathbf{v}_i.$$

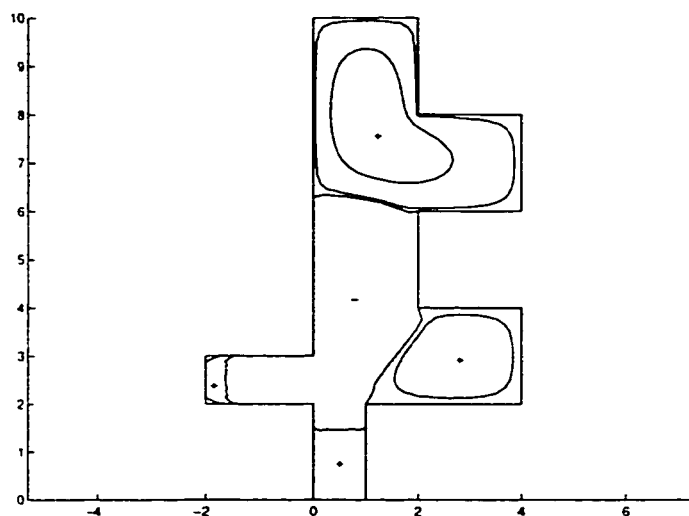


Figure 4.11: $\cos(\frac{3}{10}\pi)u_1 + 0.525u_2 + \sin(\frac{3}{10}\pi)u_3$ has four positive nodal domains.

Choose the coefficients a_i such that $\mathbf{v}$ is M -orthogonal to the first $n-1$ eigenvectors.

Thus

$$R(\mathbf{v}) \geq \lambda_n. \quad (4.6)$$

However, we compute

$$\begin{aligned} \mathbf{K}(t\mathbf{u}_1 + \mathbf{u}_n) &= t\mathbf{K}\mathbf{u}_1 + \mathbf{K}\mathbf{u}_n \\ &= t\lambda_1\mathbf{M}\mathbf{u}_1 + \lambda_n\mathbf{M}\mathbf{u}_n \\ &< \lambda_n\mathbf{M}(t\mathbf{u}_1 + \mathbf{u}_n). \end{aligned}$$

In the same notation as in Theorem 3.13, the above inequality gives us

$$\mathbf{K}_{ii}\mathbf{v}_i + \mathbf{K}_{i,r+1}\mathbf{v}_{r+1} < \lambda_n(\mathbf{M}_{ii}\mathbf{v}_i + \mathbf{M}_{i,r+1}\mathbf{v}_{r+1}).$$

Since $\mathbf{K}_{i,r+1}\mathbf{v}_{r+1} \geq \mathbf{0}$ and $\mathbf{M}_{i,r+1}\mathbf{v}_{r+1} \leq \mathbf{0}$, we have

$$\mathbf{K}_{ii}\mathbf{v}_i < \lambda_n \mathbf{M}_{ii}\mathbf{v}_i, \quad \text{i.e.,} \quad \mathbf{v}_i^T \mathbf{K}_{ii} \mathbf{v}_i < \lambda_n \mathbf{v}_i^T \mathbf{M}_{ii} \mathbf{v}_i,$$

for $i = 1, 2, \dots, r$. Now consider

$$\begin{aligned} \mathbf{v}^T \mathbf{K} \mathbf{v} &= \sum_{i=1}^r a_i^2 \mathbf{v}_i^T \mathbf{K}_{ii} \mathbf{v}_i \\ &< \lambda_n \sum_{i=1}^r a_i^2 \mathbf{v}_i^T \mathbf{M}_{ii} \mathbf{v}_i = \lambda_n \mathbf{v}^T \mathbf{M} \mathbf{v}. \end{aligned}$$

This implies that $R(\mathbf{v}) < \lambda_n$, which contradicts (4.6). ■

Note that Duval and Reiner [8] proved a similar result to Theorem 4.4. [8] states that if $\mathbf{K}$ is irreducible, then $t\mathbf{u}_1 + \mathbf{u}_n$ ($t \geq 0$) has at most $n - 1$ *non-negative* sign graphs. It was considered as a corollary of Fiedler's result. When $t > 0$, we can get a stronger result: $t\mathbf{u}_1 + \mathbf{u}_n$ has at most $n - 1$ *positive* sign graphs.

4.4 CHC on square domains

Although CHC does not hold in general, we conjecture that CHC may be true for some convex domains, especially rectangular or circular domains. In this section, we study CHC on square domains since the eigenvalues and the eigenfunctions have explicit and simple forms, and the ordering of the eigenvalues is clearer, compared with that of general rectangles of arbitrary dimensions a, b . Without loss of generality, we choose the square to be $R_{[0,\pi]} = [0, \pi] \times [0, \pi]$. For $R_{[0,\pi]}$, we will show

that CHC is true for at least the first few eigenfunctions and also for some special combinations.

Note that $\sin(nx) = \sin x U_{n-1}(\cos x)$ where U_{n-1} is Chebyshev polynomial of second kind with degree $n - 1$. Thus each eigenfunction $u_{n,m} = \sin(nx) \sin(my)$ can be regarded as a polynomial in $\cos x$ and $\cos y$ multiplied by $\sin x \sin y$, i.e.,

$$u_{n,m} = \sin x \sin y U_{n-1}(\cos x) U_{m-1}(\cos y).$$

From this point of view, any linear combination of the first t eigenfunctions can be written as

$$\sum_{i=1}^t c_i \sin(n_i x) \sin(m_i y) = \sin x \sin y \bar{P}_{n_t+m_t-2}(\cos x, \cos y).$$

The nodal set of the linear combination is then determined by that of $\bar{P}_{n_t+m_t-2}(\cos x, \cos y)$.

For simplicity, we make a transformation

$$\begin{cases} X = \cos x, & 0 \leq x \leq \pi \\ Y = \cos y, & 0 \leq y \leq \pi \end{cases},$$

so that $R_{[0,\pi]}$ is transformed to $R_{[-1,1]} = [-1, 1] \times [-1, 1]$. As this transformation is one to one, it does not change the number of nodal domains of $\bar{P}_{n_t+m_t-2}(X, Y)$. Hence, we study the nodal set of the new polynomial $P(X, Y)$ on the square $R_{[-1,1]}$ instead.

4.4.1 The first fourteen eigenfunctions

With knowledge of polynomials with lower order, we can actually verify that CHC is true for the first fourteen eigenfunctions on the square.

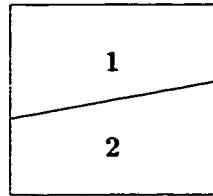
In the following discussion, we use $(1, X, Y)$ to represent an arbitrary polynomial consisting of the terms 1, X and Y ; the notation

$$P = (1, X)(1, Y)$$

means that the polynomial P has two factors $c_1 + c_2X$ and $d_1 + d_2Y$ for some numbers c_1 , c_2 , d_1 and d_2 . By analyzing all the possible forms of $P(X, Y)$, we can find the maximum number of nodal domains of $P(X, Y)$; so we can verify CHC for linear combinations of the first N eigenfunctions up to $N = 14$.

Certainly the first eigenfunction has *one* nodal domain. This follows immediately from CNLT or from the fact that $P(X, Y)$ just has a constant term.

As $\lambda_2 = \lambda_3$, linear combinations of the first two eigenfunctions are associated with polynomials consisting of 1, X and Y . In this case, the maximum number of nodal domains occurs when there is a nodal line which cuts the square into two parts. Thus, any linear combination of the first two/three eigenfunctions has at most *two* nodal domains.



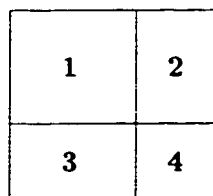
When $N = 4$, $\lambda_4 = 2^2 + 2^2$. Thus, the polynomial $P(X,Y)$ associated with the combinations of the first four eigenfunctions is

$$P_2 = (1, X, Y, XY),$$

which is either a hyperbola or can be factorized as

$$P_2 = (1, X)(1, Y).$$

Thus, we can see that the maximum nodal domains of combinations of the first four eigenfunctions is *four*.



As $\lambda_5 = \lambda_6$, the polynomial $P(X,Y)$ of combinations of the first five/six eigen-

functions have the form

$$P_2(X, Y) = (1, X, Y, XY, X^2, Y^2).$$

The nodal lines of this quadratic are either two straight lines, an ellipse, a hyperbola or a parabola. The maximum number of nodal domains occurs when the quadratic is an ellipse, the center of the ellipse is inside the square, and the ellipse cuts each side of the square twice or touches each side of the square; this gives *five* nodal domains shown in Figure 4.12. CHC is true in this case.

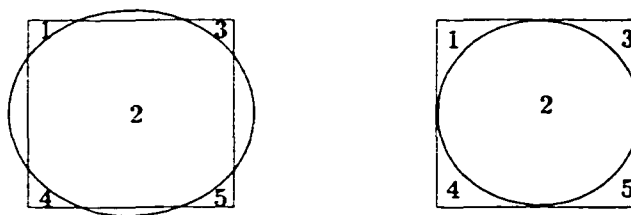


Figure 4.12: An ellipse can divide a square into 5 subregions.

Again, since λ_7 is an eigenvalue with multiplicity two, we look at linear combinations of the first seven/eight eigenfunctions. The associated polynomial consists of the terms

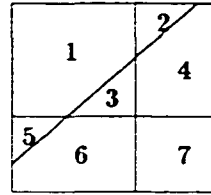
$$P_3 = (1, X, Y, XY, X^2, Y^2, X^2Y, XY^2).$$

It can have the following different factorizations.

i) A product of three line equations. That is

$$P_3 = (1, X)(1, Y)(1, X, Y),$$

the nodal lines of which can divide the square into at most seven subdomains.



ii) A product of a line and an irreducible quadratic. In this case, P_3 can have the following forms only

$$P_3 = (1, X)(1, X, Y, XY, Y^2),$$

or

$$P_3 = (1, Y)(1, X, Y, XY, X^2).$$

Figure 4.13 shows all the possible cases of the nodal curves of such a P_3 , which cuts the square into at most six subregions.

iii) An irreducible cubic equation. Weinberg [24] topologically classified the plane cubic curves into 21 equivalence classes; the irreducible cubics fall into 15 of them.

the last case, i.e.,

$$P = (1, X, Y)(1, X, Y, XY, X^2, Y^2).$$

With this form, the maximum number of nodal domains is eight. For example, Figure 4.14 shows a case where an ellipse with a line can divide a square into eight subdomains. As a result, the linear combination of the first nine/ten eigenfunctions have no more than *eight* nodal domains.

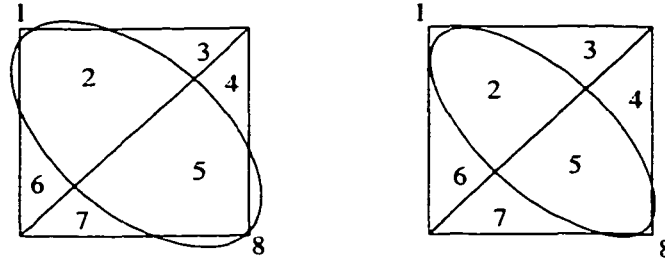


Figure 4.14: An ellipse with a line can divide a square into 8 subdomains.

For combinations of the first eleven eigenfunctions, the associated polynomial has a term of order four, X^2Y^2 . The possibilities for the factorization of P_4 are as follows:

- (i) A product of four line equations. Due to the high order term X^2Y^2 , we can have

$$P_4 = (1, X)^2(1, Y)^2,$$

the nodal lines of which divide the square into nine pieces.

1	2	3
4	5	6
7	8	9

ii) A product of two line equations and an irreducible quadratic equation. In this case, the polynomial can be factorized into the form

$$P_4 = (1, X)(1, Y)(1, X, Y, XY),$$

which has at most eight nodal domains.

1	2	3
4	5	6
	7	8

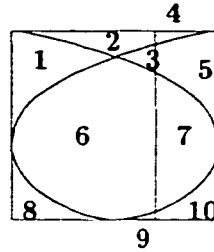
iii) A product of a line and an irreducible cubic. P_4 can be factorized into

$$(1, X)(1, X, Y, XY, X^2, Y^2, XY^2)$$

or

$$(1, Y)(1, X, Y, XY, X^2, Y^2, X^2Y).$$

The maximum number of nodal domains in this case is ten.



iv) An irreducible quartic equation. Gudkov, Utkin and Tai [15] classified the irreducible quartic plane curves into 99 types. According to the classification, it has no more than ten nodal domains.

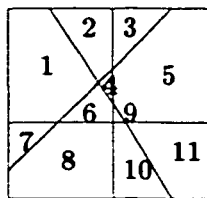
As a result, linear combinations of the first 11 eigenfunctions have at most *ten* nodal domains.

For combinations of the first twelve/thirteen eigenfunctions, the associated polynomial has new fourth order terms, X^3Y and XY^3 , which can have the following factorizations.

i) A product of four line equations, *i.e.*,

$$P_4 = (1, X)(1, Y)(1, X, Y)^2,$$

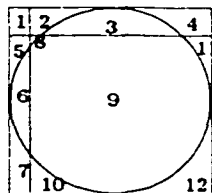
which has 11 nodal domains.



ii) A product of two line equations and an irreducible quadratic. In this case, P_4 has the form

$$P_4 = (1, X)(1, Y)(1, X, Y, XY, X^2, Y^2),$$

which has at most 12 nodal domains.



iii) The cases that the polynomial is a product of a line and an irreducible cubic or an irreducible quartic, are the same as those for the first eleven eigenfunctions. The maximum nodal domain is 10.

Thus, linear combinations of the first twelve/thirteen eigenfunctions has at most *twelve* nodal domains.

For linear combinations of the first fourteen eigenfunctions, the only case different from that of the first twelve/thirteen is that P_4 is a product of two irreducible quadratics. Actually, the maximum number of nodal domains can occur when two ellipses cross each other at four different points and each ellipse either intersects each side of the square twice or touches each side. It cuts the square into 13 nodal domains. Thus, combinations of the first fourteen eigenfunctions have at most *thirteen* nodal domains.

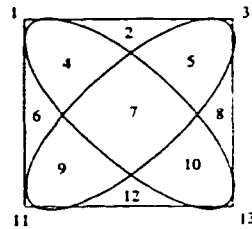


Figure 4.15: Two ellipses can divide a square into 13 subdomains.

As a conclusion, CHC is true for combinations $w = \sum_{i=1}^N c_i u_i$ up to $N = 14$.

4.4.2 Special linear combinations

We are also able to prove that CHC is true for certain types of linear combinations of the eigenfunctions on a square, such as when the associated polynomial $P(X,Y)$ can be factorized into a product of linear equations or a product of ellipses or with a line.

Let $P(X,Y)$ be a polynomial generated by a linear combination of the first N eigenfunctions. The appearance of each $X^i Y^j$ term in $P(X,Y)$ implies that eigenfunction $u_{i+1,j+1}(x,y)$ or $u_{j+1,i+1}(x,y)$, corresponding to eigenvalue $\lambda_{i+1,j+1}$, has appeared, and that eigenfunctions $u_{k,\ell}(x,y)$, for k, ℓ such that $1 \leq k \leq i+1$,

N	New High Order Terms		Maximum No. of Nodal domains	
1	1		1	
2, 3	X	Y	2	
4	XY		4	
5, 6	X^2	Y^2	5	
7, 8	X^2Y	XY^2	7	
9, 10	X^3	Y^3	8	
11	X^2Y^2		10	
12, 13	X^3Y	XY^3	12	
14	X^4	Y^4	13	

Figure 4.16: This table illustrates that CHC is true for the first N eigenfunctions on the square, $N = 1, 2, \dots, 14$. The second column gives all the new high order terms involved in the polynomial $P(X, Y)$ for each N , and the figures in the last column provides a linear combination that can produce the maximum number of nodal domains.

$1 \leq \ell \leq j+1$, have appeared (this is because $k^2 + \ell^2 \leq (i+1)^2 + (j+1)^2$, i.e., eigenvalues $\lambda_{k,\ell}$ appear before $\lambda_{i+1,j+1}$ for $1 \leq k \leq i+1$, $1 \leq \ell \leq j+1$). If $P(X,Y)$ is a complete polynomial with order m , then $P(X,Y)$ has $N_1 = \frac{1}{2}(m+1)(m+2)$ terms. It implies that there are N_1 eigenvalues involved (some may be equal). Among them, the highest eigenvalue is $\lambda_{m+1,1}$, which appears twice as both X^m and Y^m correspond to the highest eigenvalue. Since some eigenvalues, corresponding to higher order terms than X^m and Y^m , may be less than $\lambda_{m+1,1}$, they will come before the eigenvalue $\lambda_{m+1,1}$. Thus, the eigenvalue $\lambda_{m+1,1}$ is labeled as at least $(N_1 - 1)$ -th eigenvalue, i.e., $N \geq N_1 - 1$. So if we can prove that the number of nodal domains of $P(X,Y)$ is no greater than $N_1 - 1$ which is no greater than N , then these linear combinations of the first N eigenfunctions have no more than N nodal domains; CHC holds in this case.

I. Products of lines

Assume that the polynomial $P(X,Y)$, associated with linear combinations of the first N eigenfunctions, is a product of n linear equations. Generally, the nodal set of such a $P(X,Y)$ consists of n_1 vertical lines, n_2 horizontal lines and n_3 slanting lines, in which $n_1 + n_2 + n_3 = n$. If there is no slanting line, i.e., $n_3 = 0$, the square is divided into $N_2 = (n_1 + 1)(n_2 + 1)$ subregions by n_1 vertical lines and n_2 horizontal lines. The number of $X^i Y^j$ terms in $P(X,Y)$ is $N_1 = (n_1 + 1)(n_2 + 1)$. In this case, the maximum eigenvalue involved is λ_{n_1+1,n_2+1} , which corresponds to $X^{n_1} Y^{n_2}$, and appears only once; hence, $N \geq N_1$. Therefore, we have $N_2 = N_1 \leq N$. CHC is true in this case.

When there is at least one slanting line, i.e., $n_3 \geq 1$, then it is easy to see that

the number of terms involved in $P(X, Y) = (1, X)^{n_1} (1, Y)^{n_2} (1, X, Y)^{n_3}$ is

$$\begin{aligned} N_1 &= (n_1 + 1)(n_2 + 1) + n_3(n_1 + n_2 + 1) + \frac{1}{2}n_3(n_3 + 1) \\ &= n_1n_2 + n_2n_3 + n_1n_3 + n_1 + n_2 + 1 + \frac{1}{2}n_3(n_3 + 3). \end{aligned}$$

The term corresponding to the highest eigenvalue is either $X^{n_1+n_3}Y^{n_2}$ or $X^{n_1}Y^{n_2+n_3}$, depending on whether $n_1 \geq n_2$ or $n_1 \leq n_2$. In either case, $N \geq N_1 - 1$.

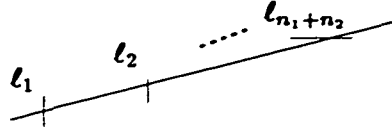
Theorem 4.5 *The nodal lines of the polynomial*

$$P(X, Y) = (1, X)^{n_1} (1, Y)^{n_2} (1, X, Y)^{n_3}$$

can divide the square $R_{[-1,1]}$ into no more than N_2 nodal domains, where

$$N_2 = N_1 - n_3.$$

Proof: Clearly, n_1 vertical lines and n_2 horizontal lines can divide the square into $(n_1 + 1)(n_2 + 1)$ subregions. Add one slanting line on top of the n_1 vertical lines and n_2 horizontal lines. The maximum number of subregions, which the lines cut the square into, occurs only when the slanting line intercepts the other $n_1 + n_2$ lines at different points. Since two lines can intercept only once, the slanting line is divided into $n_1 + n_2 + 1$ line segments.



The new subregions appearing would be the ones having one of such line segments as boundaries. There are $2(n_1 + n_2 + 1)$ such regions. Hence, the number of regions is now increased by $n_1 + n_2 + 1$. So, the number of nodal domains in this case is

$$(n_1 + 1)(n_2 + 1) + (n_1 + n_2 + 1).$$

The same reasoning is applicable to the rest of the slanting lines. Therefore, after adding n_3 slanting lines, we compute the maximum number of nodal domains

$$N_2 = (n_1 + 1)(n_2 + 1) + \sum_{i=1}^{n_3} (n_1 + n_2 + i) = N_1 - n_3.$$

■

Notice that $N_2 \leq N_1 - 1 \leq N$, i.e., the number of nodal domains is no greater than N . CHC holds in this case.

II. Product of ellipses or with one line

Let the degree of $P(X,Y)$ be n . Suppose that $P(X,Y)$ is a product of ellipses when n is even; $P(X,Y)$ is a product of $\frac{n-1}{2}$ ellipses and a line equation when n is odd.

Again, we first count the number of terms involved in $P(X,Y)$. When n is even, the number of terms involved in $(1, X, Y)^n$ is $\frac{1}{2}(n+1)(n+2)$. When n is odd, we can have $\frac{n-1}{2}$ ellipses with a horizontal/vertical line or a slanting line. As long as this line meets every ellipse twice and at different points, the numbers of nodal domains of $P(X,Y)$ with either a horizontal/vertical line or a slanting line will be the same. But if the nodal set of P consists of $\frac{n-1}{2}$ ellipses and a vertical or horizontal line, the number of terms involved is less than that for ellipses with a slanting line. Thus, the number of terms in $(1, X)(1, X, Y)^{n-1}$ is $\frac{1}{2}n(n+3)$. In total, the number of X and Y terms in $P(X,Y)$ is at least

$$N_1 = \begin{cases} \frac{1}{2}(n+1)(n+2), & \text{if } n \text{ is even;} \\ \frac{1}{2}n(n+3), & \text{if } n \text{ is odd} \end{cases}$$

Next, we examine the maximum number of nodal domains that the nodal lines of such a $P(X,Y)$ can divide the square into.

Theorem 4.6 *Let $P(X,Y)$ be a polynomial as in the above. Then the nodal lines of P divide the square into no more than N_2 nodal domains, where*

$$N_2 = \begin{cases} \frac{(n+1)^2+1}{2} & \text{if } n \text{ is even;} \\ \frac{(n+1)^2}{2}, & \text{if } n \text{ is odd} \end{cases}$$

Proof: Observe that with one ellipse, the maximum number of nodal domains is obtained when the ellipse either cuts each side of the square twice or touches each side. It produces $\frac{(2+1)^2+1}{2} = 5$ nodal domains, see Figure 4.12. With two ellipses, the maximum number of nodal domains occurs when both of them intersect each

side of the square twice (or touch each side), and they meet each other at four different points. In this case, the number of nodal domains is 13, see Figure 4.15.

Suppose that it is true for $n - 2$, i.e., $\frac{n-2}{2}$ ellipses divide a square into at most $\frac{(n-1)^2+1}{2}$ subdomains. We will show that it is true for $n/2$ ellipses. In this case, the maximum number of nodal domains can occur when the $\frac{n}{2}$ -th ellipse, with the diagonals of the square as its axes,

- (a) intercepts every other ellipse at four different points (in total, it meets the $\frac{n-2}{2}$ ellipses at $2(n - 1)$ different points),
- (b) and intercepts each side of square twice or touches each side.

With condition (a), it creates $2(2\frac{n-2}{2} + 1) = 2n - 2$ new regions; with condition (b), two new regions are created, which are bounded by the $\frac{n}{2}$ -th ellipse and two corners of the square. In total, there are $2n$ more subregions created. Thus, the maximum number of nodal domains is

$$N_2 = \frac{(n-1)^2+1}{2} + 2n = \frac{(n+1)^2+1}{2}.$$

When n is odd, we add a line so that it is cut by $\frac{n-1}{2}$ ellipses into n line segments, see Figure 4.14. This way, we get the maximum number of nodal domains, which is

$$N_2 = \frac{n^2+1}{2} + n = \frac{(n+1)^2}{2}.$$

■

Therefore, by comparison, we have $N_2 \leq N_1 - 1 \leq N$. CHC is true in the second

case as well.

Using the properties of polynomials, we are able to prove that CHC is true on a square membrane for certain linear combinations of eigenfunctions and for any linear combination of the first fourteen eigenfunctions. This approach appears to be efficient when N is small, as the classification of polynomials of lower order is known and simple. However, due to the lack of the knowledge of higher order polynomials, as N becomes larger, it is increasingly more difficult to verify CHC: even when the classification of the plane curves is known, it may still be painful to go through the large number of equivalent classes. Thus, to prove CHC for general N , this is not a preferable way.

There is another possible approach that is used to prove CHC in one-dimensional case, see Gladwell ([13], chapter 8). Let $\{u_i\}$ denote the eigenfunctions of

$$\begin{cases} u''(x) + \lambda u(x) = 0, & x \in (0, \pi); \\ u(0) = u(\pi) = 0. \end{cases}$$

In fact, $\lambda_i = i^2$ and $u_i = \sin(ix)$, $i = 1, 2, \dots$. It is well known that the sequence of functions

$$U = U \begin{pmatrix} i_1 & i_2 & \cdots & i_n \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} = \begin{vmatrix} u_{i_1}(x_1) & u_{i_1}(x_2) & \cdots & u_{i_1}(x_n) \\ \vdots & \vdots & \cdots & \vdots \\ u_{i_n}(x_1) & u_{i_n}(x_2) & \cdots & u_{i_n}(x_n) \end{vmatrix}$$

are the eigenfunctions for

$$\begin{cases} \Delta_n u + \lambda u = 0, & \text{in } G_n \\ u = 0, & \text{on } \partial G_n \end{cases}$$

where G_n is a subdomain of $[0, \pi]^n$ such that $x_1 < x_2 < \cdots < x_n$. Particularly, $U_1 = U \begin{pmatrix} 1 & 2 & \cdots & n \\ x_1 & x_2 & \cdots & x_n \end{pmatrix}$ is the first eigenfunction in G_n , i.e., it is strictly positive for all $(x_i)_1^n$ such that $x_1 < x_2 < \cdots < x_n$. For instance, when $n = 2$, $U \begin{pmatrix} 1 & 2 \\ x_1 & x_2 \end{pmatrix}$ maintains strictly fixed sign and is the first eigenfunction on the half of the square where $x_1 < x_2$, as shown in Figure 4.17(a). When $n = 3$, U_1 is the first eigenfunction on one-sixth of a cube where $x_1 < x_2 < x_3$ shown in Figure 4.17(b), and is strictly positive in G_3 .

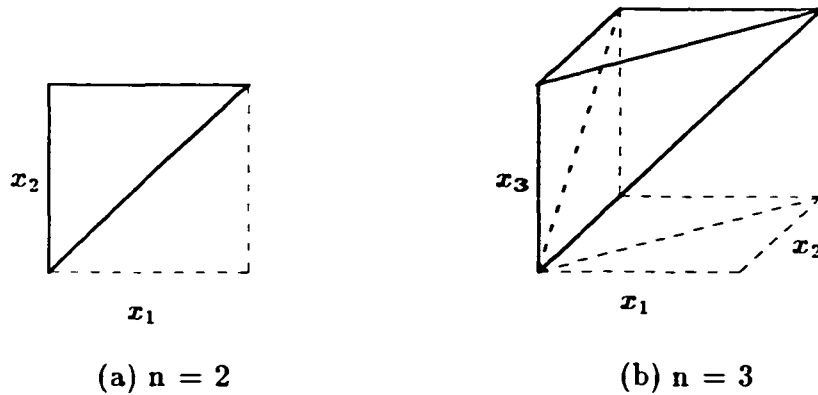


Figure 4.17: The shapes of G_n for $n = 2$ and $n = 3$.

Now suppose that a linear combination $w = \sum_{i=1}^n c_i u_i(x)$ has more than n nodal

intervals. Then w has at least n different zeros in $(0, \pi)$, say $y_1 < y_2 < \cdots < y_n$, such that

$$\sum_{i=1}^n c_i u_i(y_j) = 0, \quad j = 1, 2, \dots, n.$$

Since not all c_i 's are zero,

$$U_1(y_1, \dots, y_n) = U \begin{pmatrix} 1 & 2 & \cdots & n \\ y_1 & y_2 & \cdots & y_n \end{pmatrix} = 0$$

which contradicts the statement that U_1 is strictly positive in G_n . Therefore, CHC holds in the one-dimensional case. This proof is elegant.

In two dimensions, let $\{u_i(\mathbf{p})\}$ be the eigenfunctions of the Helmholtz equation on the square $R_{[0, \pi]}$, where $\mathbf{p} = (x, y) \in R_{[0, \pi]}$. The products of the eigenfunctions

$\{u_{i_1}(\mathbf{p}_1)u_{i_2}(\mathbf{p}_2) \cdots u_{i_n}(\mathbf{p}_n)\}$ form a complete set on $R_{[0, \pi]}^n$; so $U = U \begin{pmatrix} 1 & 2 & \cdots & n \\ \mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_n \end{pmatrix}$

is still an eigenfunction on $R_{[0, \pi]}^n$. However, unlike the U 's defined in one-dimensional

case, it becomes hard to tell in which subregion of $R_{[0, \pi]}^n$, $U \begin{pmatrix} 1 & 2 & \cdots & n \\ \mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_n \end{pmatrix}$

would be the first eigenfunction. This is because the region that $\mathbf{p}_i = \mathbf{p}_j$ is a

$(2n - 2)$ -dimensional nodal hypersurface (not a $(2n - 1)$ -dimensional nodal hypersurface) of $R_{[0, \pi]}^n$, which does not separate positive and negative nodal domains.

Thus, we cannot follow the exact approach used in one-dimensional case to prove

CHC in two dimensional case. But, since the U 's are the eigenfunctions on $R_{[0, \pi]}^n$,

the U 's have CNLT property. If we can prove the following

“Suppose that a combination $w = \sum_{i=1}^n c_i u_i(p)$ have more than n nodal domains.

Then the eigenfunction $U \begin{pmatrix} 1 & 2 & \cdots & n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$ will not obey CNLT.”,

then we will get a contradiction and CHC will hold for the eigenfunctions on squares.

The difficulty, though, is to have a clear visualization of objects in four or higher dimensions.

Whether CHC is true for square domains is left unsolved in this thesis.

Chapter 5

Conclusions and Further Research

The main contributions of the thesis are contained in Chapter 3 and 4, where the sign property of the FEM solutions, of finite linear combinations of the eigenfunctions and of finite linear combinations of the FEM eigenvectors are studied.

In Chapter 3, we found that when the stiffness matrix $\mathbf{K}$ has non-positive off-diagonal entries, there are discrete versions of CNLT holding for the FEM eigenvectors, which extends Gantmakher's one-dimensional result to higher dimensions. Restraints on triangular finite elements in $\mathbb{R}^2$ or tetrahedral elements in $\mathbb{R}^3$ are found so that the nonzero off-diagonals of $\mathbf{K}$ are negative. We proved that

- (i) when λ_n is distinct, then $\mathbf{u}_n$ has at most n sign graphs;
- (ii) when λ_n is an eigenvalue with multiplicity r , there exists a set of M -orthogonal eigenvectors $\{\mathbf{u}_j\}_n^{n+r-1}$ corresponding to λ_n such that

$$SG(\mathbf{u}_j) \leq j, \quad j = n, n+1, \dots, n+r-1;$$

- (iii) when λ_n is an eigenvalue of multiplicity r , there also exists a set of linearly independent eigenvectors $\{\mathbf{v}_j\}_n^{n+r-1}$ such that

$$SG(\mathbf{v}_j) \leq n, \quad j = n, n+1, \dots, n+r-1.$$

Regarding the applications, the findings in this thesis provide a necessary condition for a mode to be the n -th eigenmode of a vibration system. But what are the necessary and sufficient conditions for the inverse eigenmode problem would be of interest for further research.

Chapter 4 mostly discussed the CHC property in continuous and discrete cases. We provided a few interesting numerical counterexamples showing that CHC is generally false in both continuous and discrete cases. However, we formulated and proved a restricted CHC which holds in both continuous and discrete cases. As all the counterexamples of CHC are with non-convex domains, we conjectured that CHC is true for certain convex domains, such as rectangles or circle. We verified that for square domains, CHC is true for any combinations of the first fourteen eigenfunctions and for certain special linear combinations of any eigenfunctions. However, whether CHC is true in general for the eigenfunctions on rectangular domains is left as a remaining problem. Also, another question is raised – if there exists a convex counterexample to CHC. It is the author's opinion that the answer may be 'No'. If the answer is 'No', why does the convexity play such an important role in CHC?

Many potential research topics remain on this interesting subject.

Bibliography

- [1] G. Alessandrini. Nodal lines of eigenfunctions of the fixed membrane problem in general convex domains. *Comment. Math. Helv.* **69**(1) 1994, 142–154.
- [2] V.I. Arnol'd. The topology of real algebraic curves (the works of Petrovskii and their development), *Uspekhi Mat. Nauk*, **28**(5) 1973, 260–262 (in Russian).
- [3] A. Berman and R.J. Plemmons. *Nonnegative Matrices in the Mathematical Science*. Classics in applied mathematics, V. 9, New York: Academic Press, 1979.
- [4] R.G. Busacker and T.L. Saaty. *Finite Graphs and Networks: an introduction with applications*. International Series in Pure and Applied Mathematics, New York: McGraw-Hill, 1965.
- [5] G.F. Carey and J.T. Oden. *Finite Elements: a second course*, Vol. II. Englewood Cliffs, NJ: Prentice-Hall, 1981.
- [6] S.Y. Cheng. Eigenfunctions and Nodal Sets. *Comment. Math. Helvetici*, **51** 1976, 43–55.

- [7] R. Courant and D. Hilbert. *Methods of Mathematical Physics*, Vol. 1. Interscience, New York, 1953.
- [8] A.M. Duval and V. Reiner, Perron-Frobenius type results and discrete versions of nodal domain theorems, *Lin. Alg. Appl.*, **294** 1999, 259–268.
- [9] L.C. Evans. *Partial Differential Equations*, American Mathematical Society, Providence, Rhode Island, 1998.
- [10] M. Fiedler, A property of eigenvectors of non-negative symmetric matrices and its application to graph theory, *Czech. Math. J.*, **25** 1975, 619–633.
- [11] F.R. Gantmakher. *The Theory of Matrices*, Vols I & II. New York: Chelsea, 1959.
- [12] F.R. Gantmakher and M.G. Krein. *Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems*. Moscow-Leningrad: State Publishing House of Technical-Theoretical Literature, 1950. (Translation: US Atomic Energy Commission. Washington, DC, 1961.)
- [13] G.M.L. Gladwell. *Inverse Problems in Vibration*. Martinus Nijhoff Publishers, Dordrecht, 1986.
- [14] G.M.L. Gladwell. Sign properties of eigenvectors: an extension of a result due to M. Fiedler, to appear.
- [15] D.A. Gudkov, G.A. Utkin and M.L. Tai. A complete classification of indecomposable curves of fourth order. *Mat. Sb. (N.S.)*, **69**(111) 1966, 222–256.

- [16] O.H.Hald and J.R. McLaughlin, Inverse problems using nodal position data – uniqueness results, algorithms and bounds. Special program on Inverse Problems (Proceedings of the centre for Math Analysis, ANU) Vol. 17 1988, 32–59.
- [17] H. Herrmann. Beziehungen zwischen den Eigenwerten und Eigenfunktionen verschiedener Eigenwertprobleme. *Math. Z.*, **40** 1935, 221–241.
- [18] D. Jerison and C. Kenig. Unique continuation and absence of positive eigenvalues for Schrödinger operators. *Ann. Math.* **2**(121) 1985, 463–494.
- [19] F. John. *Partial Differential Equations*, fourth edition. Applied mathematical sciences, v. 1, New York: Springer. 1981.
- [20] C. Müller, On the behavior of the solutions of the differential equation $\Delta u = F(x, u)$ in the neighborhood of a point. *Comm. Pure Appl. Math.*, **7** 1954, 505–551.
- [21] A. Pleijel. Remarks on Courant’s nodal line theorem, *Commun. Pure Appl. Math.*, **9** 1956, 543–550.
- [22] M.H. Protter and H.F. Weinberger. *Maximum Principles in Differential Equations*, New York: Springer, 1984.
- [23] G.W. Stewart, *Introduction to Matrix Computations*, New York: Academic Press, 1973.
- [24] D.A. Weinberg. The affine classification of cubic curves. *Rocky Mountain J. of Math.* **18**(3) 1988, 655–679.