International Series of Numerical Mathematics
Internationale Schriftenreihe zur Numerischen Mathematik
Série Internationale d'Analyse Numérique
Vol. 96

ISNM 96

Numerical Treatment of Eigenvalue Problems Vol. 5
Numerische Behandlung von Eigenwertaufgaben Band 5

Edited by
Herausgegeben von
J. Albrecht
L. Collatz
P. Hagedorn
W. Velte

Springer Basel AG

ISNM 96:
International Series of Numerical Mathematics
Internationale Schriftenreihe zur Numerischen Mathematik
Série Internationale d'Analyse Numérique
Vol. 96

Edited by
K.-H. Hoffmann, Augsburg; H. D. Mittelmann, Tempe;
J. Todd, Pasadena

Springer Basel AG

Numerical Treatment of Eigenvalue Problems Vol. 5

Workshop in Oberwolfach, February 25 – March 3, 1990

Numerische Behandlung von Eigenwertaufgaben Band 5

Tagung in Oberwolfach, 25. Februar – 3. März 1990

Edited by
Herausgegeben von
J. Albrecht
L. Collatz
P. Hagedorn
W. Velte

1991 Springer Basel AG

Editors

J. Albrecht
Technische Universität Clausthal
Institut für Mathematik
Erzstrasse 1
D–3392 Clausthal-Zellerfeld

P. Hagedorn
TH Darmstadt
Institut für Mechanik
Hochschulstrasse 1
D–6100 Darmstadt

W. Velte
Universität Würzburg
Institut für Angewandte Mathematik und Statistik
Am Hubland
D–8700 Würzburg

Library of Congress Cataloging-in-Publication Data

Numerische Behandlung von Eigenwertaufgaben.
(Internationale Schriftenreihe zur numerischen Mathematik ; v. 24, 43, 69, 96)
Includes bibliographies.
Contents: [1] Tagung über Numerische Behandlung von Eigenwertaufgaben vom 19. bis 24. November 1972 / Tagungsleiter, L. Collatz und K. P. Hadeler – Bd. 2. Tagung an der Technischen Universität Clausthal vom 18. bis 20. Mai 1978 / herausgegeben von J. Albrecht. L. Collatz – [etc.] – v. 5. Workshop in Oberwolfach, February 25 – March 3, 1990 = Tagung in Oberwolfach, 25. February – 3. März 1990 / edited by J. Albrecht . . . et al.
1. Differential equations–Numerical solutions –Data processing–Congresses. 2. Matrices–Data processing–Congresses. 3. Eigenvalues–Data processing–Congresses. I. Collatz, L. (Lothar), 1910– . II. Hadeler, K. P. (Karl Peter), 1926– . III. Albrecht, J. (Julius), 1926– Tagung über Numerische Behandlung von Eigenwertaufgaben (1972 : Oberwolfach, Germany). IV. Numerical treatment of eigenvalue problems. Series: International series of numerical mathematics ; v. 24, etc.
QA371.N84 512.9'434 74-235221

Deutsche Bibliothek Cataloging-in-Publication Data

Numerical treatment of eigenvalue problems: workshop . . . = Numerische Behandlung von Eigenwertaufgaben. – Basel ; Boston ; Berlin : Birkhäuser.
Beitr. teilw. dt., teilw. engl.
Bis Bd. 2 (1979) u.d.T.: Numerische Behandlung von Eigenwertaufgaben
NE: PT
Vol. 5. In Oberwolfach, February 25 – March 3, 1990. – 1991
(International series of numerical mathematics ; Vol. 96)

NE: GT

Originally published by Birkhäuser Verlag Basel in 1991.
Softcover reprint of the hardcover 1st edition 1991
Printed from the authors' camera-ready manuscripts on acid-free paper in Germany
ISBN 978-3-0348-6334-6 ISBN 978-3-0348-6332-2 (eBook)
DOI 10.1007/978-3-0348-6332-2

Preface

The present volume contains the lectures held at the conference on "Eigenvalue problems in the natural and engineering sciences, and their numerical treatment" sponsored by the Mathematical Research Institute in Oberwolfach, 25 February -3 March 1990. The conference was attended by participants from seven European countries and from the United States.

A first group of papers concerns various methods for reaching eigenvalue boundaries in the case of general eigenvalue problems for (partial) differential equations, including those of rational approximation, approximation with finite elements and domain decomposition. One contribution treats eigenvalue problems that occur when one studies the singularities at corners and edges of the solutions to partial differential equations.

A second group of papers draws on concrete eigenvalue problems in engineering and the natural sciences; included here is the problem of oscillation of a wheel rolling on a rail, a nonselfadjoint problem for a vibrating plate, problems in quantum chemistry, as well as a problem taken from the theory of ARMA models.

A third area is devoted to the numerics of eigenvalue problems for large, sparse matrices, and more specifically, how they arise from discretization of eigenvalue problems in partial differential equations. Preconditioning and stability of algorithms and new variants in iterative methods, including parallel algorithms, are among the topics considered.

J. Albrecht (Clausthal)
L. Collatz (Hamburg)
P. Hagedorn (Darmstadt)
W. Velte (Würzburg)

Vorwort

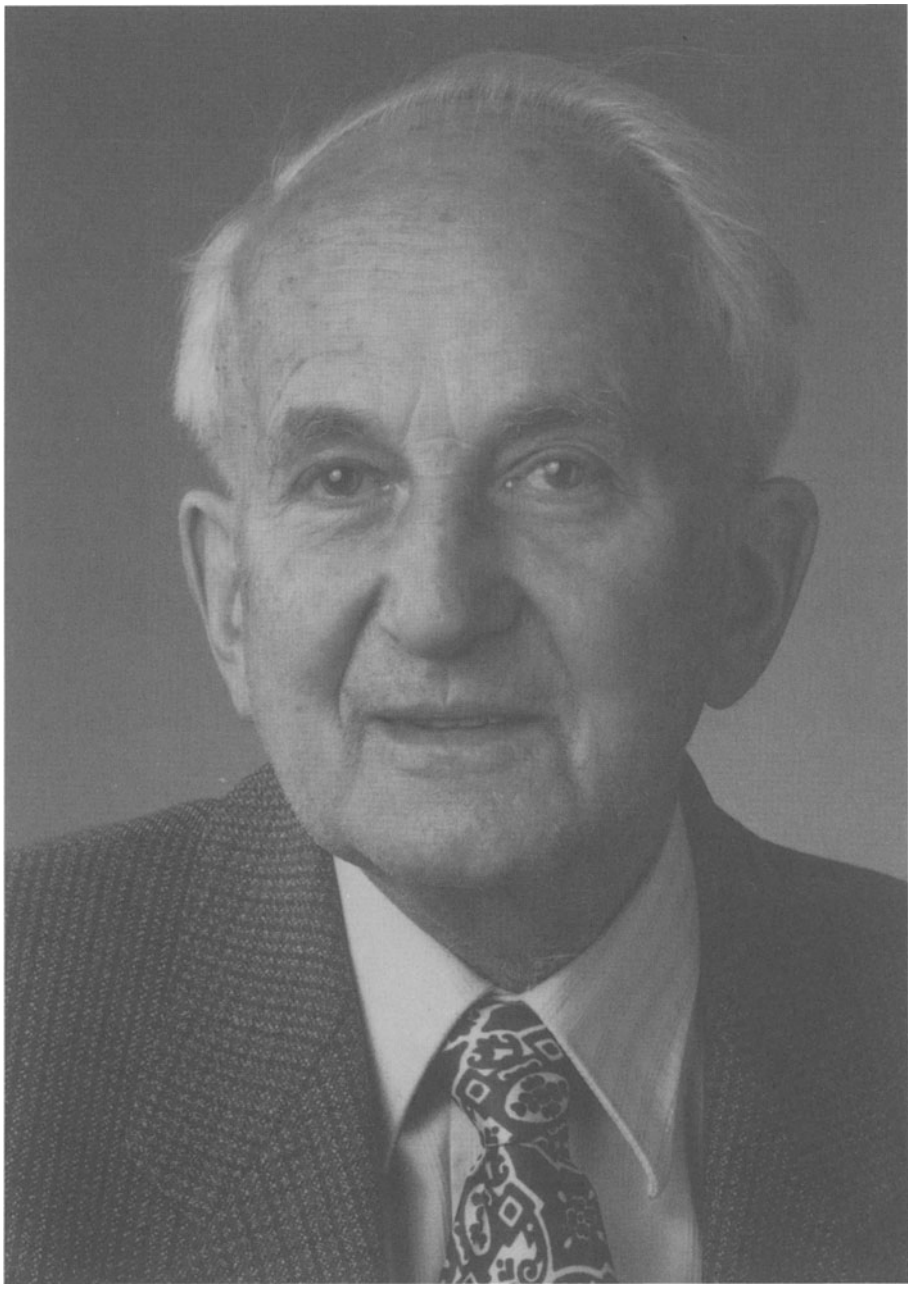
Der vorliegende Band enthält die Manuskripte von Vorträgen, die im Mathematischen Forschungsinstitut Oberwolfach auf der Tagung "Eigenwertaufgaben in den Natur- und Ingenieurwissenschaften und ihre numerische Behandlung" (25. Februar bis 3. März 1990) gehalten wurden. Die Tagung wurde von Teilnehmern aus sieben europäischen Ländern sowie aus den USA besucht.

Eine erste Gruppe von Beiträgen befasste sich mit verschiedenen Methoden zur Gewinnung von Eigenwertschranken bei allgemeinen Eigenwertaufgaben für (partielle) Differentialgleichungen, und zwar einschliesslich rationaler Approximation, Approximation mit finiten Elementen sowie Gebietszerlegung. Ein Beitrag handelte von Eigenwertproblemen, die beim Studium von Ecken- und Kantensingularitäten der Lösungen partieller Differentialgleichungen auftreten.

In einer zweiten Gruppe von Vorträgen wurden konkrete Eigenwertaufgaben aus den Ingenieur- und Naturwissenschaften abgehandelt, darunter das Problem der Eigenschwingungen eines rollenden Rades auf einer Schiene, ein nichtselbstadjungiertes Problem bei einer schwingenden Platte, Probleme aus der Quantenchemie, sowie ein Problem aus der Theorie der ARMA-Modelle.

Ein dritter, ausgedehnter Themenkreis war der Numerik von Eigenwertaufgaben bei grossen, dünn besetzten Matrizen gewidmet, wie sie durch Diskretisierung von Eigenwertaufgaben bei partiellen Differentialgleichungen entstehen. Behandelt wurden u.a. Fragen der Vorkonditionierung und der Stabilität von Algorithmen sowie neue Varianten iterativer Verfahren einschliesslich paralleler Algorithmen.

J. Albrecht (Clausthal)
L. Collatz (Hamburg)
P. Hagedorn (Darmstadt)
W. Velte (Würzburg)



Lothar Collatz 6.7.1910 - 26.9.1990

Lothar Collatz †

Dr.phil. Dr.h.c.mult. Lothar Collatz, ordentlicher Professor für Angewandte Mathematik an der Universität Hamburg, starb am 26.9.1990 während einer Tagung in Varna (Bulgarien) nach einem Herzinfarkt. Er wurde mitten aus dem wissenschaftlichen Leben abberufen: Im August weilte er zu Vorträgen in Japan, Mitte September erschien in der DMV-Festschrift "Ein Jahrhundert Mathematik" sein Beitrag "Numerik", am 24.9. hielt er in Varna die Eröffnungsansprache und einen Übersichtsvortrag.

Lothar Collatz war einer der bedeutendsten Forscher und Lehrer auf dem Gebiet der Angewandten und Numerischen Mathematik; seine Ideen und seine Forschungsergebnisse, die er in seinen Büchern und in weit mehr als 200 Zeitschriftenaufsätzen niederlegte, begleiteten und prägten über Jahrzehnte hinweg den wissenschaftlichen Weg vieler Mathematiker und Ingenieure.

Einladungen zu Vorträgen, die aus der ganzen Welt an ihn ergingen, und die Verleihung der Ehrendoktorwürde durch deutsche und ausländische Universitäten, zuletzt durch die Technische Universität Dresden unmittelbar bevor Lothar Collatz am 6.7.1990 sein achtzigstes Lebensjahr vollendete, sind herausragende Zeichen der Anerkennung seines wissenschaftlichen Werkes und der Verehrung, die ihm entgegengebracht wurde.

Seit 1964 hat Lothar Collatz gemeinsam mit Kollegen und Schülern im Mathematischen Forschungsinstitut Oberwolfach Tagungen geleitet und Berichte darüber in der ISNM-Reihe des Birkhäuser-Verlages herausgegeben. Der vorliegende Band, der letzte, an dem Lothar Collatz beteiligt ist, schließt nun den Ring zu seinem ersten großen Werk, dem 1945 erschienenen Buch "Eigenwertaufgaben mit technischen Anwendungen".

J. Albrecht P. Hagedorn W. Velte

Contents

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

EIGENWERTAUFGABEN MIT FUNKTIONAL-DIFFERENTIALGLEICHUNGEN

J. Albrecht
Technische Universität Clausthal, Institut für Mathematik
Clausthal-Zellerfeld, Germany

Herrn Prof. Dr. Dr. h.c. Lothar Collatz zum 80. Geburtstag gewidmet.

In seinen Büchern [2], [3] hat L. Collatz nicht nur eine Fülle von Eigenwertaufgaben aus Physik und Technik ausführlich behandelt, sondern auch einige Eigenwertaufgaben mit Funktional-Differentialgleichungen; auf diese sollen im folgenden - neben dem Ritzschen Verfahren - Einschließungssätze angewendet werden, die zuerst von Temple und Collatz angegeben und später von Lehmann, Goerisch und anderen verallgemeinert worden sind.

Liu und Ortiz [7] haben die Collatzschen Funktional-Differentialgleichungen kürzlich ebenfalls wieder aufgegriffen und gute Näherungen für die Eigenwerte bestimmt.

1. Die Eigenwertaufgabe [3]

$$-\phi''(x) = \lambda\phi(1-x) \quad \text{in } [0,1]; \quad \phi(0) = 0, \ \phi'(1) = 0 \qquad (1)$$

ist der linksdefiniten Eigenwertaufgabe

$$M(f,\phi) = \lambda N(f,\phi), \quad \phi \in D \quad \text{für alle } f \in D \qquad (1^*)$$

mit den symmetrischen Bilinearformen

$$M(f,g) := \int_0^1 f'(x)g'(x)dx \quad \text{für } f,g \in D,$$

$$N(f,g) := \int_0^1 f(x)g(1-x)dx \quad \text{für } f,g \in D$$

und mit

$$D := \{f \in C^1[0,1]: f(0) = 0\}$$

äquivalent. Aus der Schwarzschen Ungleichung folgt

$$\left|\int_0^1 f(x)f(1-x)dx\right| \leq \int_0^1 f(x)^2 dx \quad \text{für } f \in D;$$

daher kann die Theorie regulärer symmetrischer Eigenwertaufgaben [10] auf (1*) angewendet und die Eigenwertaufgabe

$$-\tilde{\phi}''(x) = \tilde{\lambda}\tilde{\phi}(x) \quad \text{in } [0,1]; \quad \tilde{\phi}(0) = 0, \ \tilde{\phi}'(1) = 0$$

bzw.

$$\tilde{M}(f,\tilde{\phi}) = \tilde{\lambda}\tilde{N}(f,\tilde{\phi}), \quad \tilde{\phi} \in D \text{ für alle } f \in D$$

mit

$$\tilde{M}(f,g) := \int_0^1 f'(x)g'(x)dx \quad \text{für } f,g \in D,$$

$$\tilde{N}(f,g) := \int_0^1 f(x)g(x)dx \quad \text{für } f,g \in D$$

als Vergleichsaufgabe herangezogen werden.
Obere Schranken für die ersten positiven und untere Schranken für die ersten negativen Eigenwerte liefert nun das Verfahren von Ritz, untere Schranken für die ersten positiven und obere Schranken für die ersten negativen Eigenwerte das Verfahren von Lehmann [6], [4]. Verwendet man paarweise die Ansatzfunktionen

$$v_{2p-1}(x) = x^{4p-3}, \ v_{2p}(x) = x^{4p-2} \quad (p=1,\ldots,n),$$

für das Verfahren von Lehmann außerdem

$$w_{2p-1}(x) = \frac{1-(1-x)^{4p-1}}{(4p-1)(4p-2)}, \quad w_{2p}(x) = \frac{1-(1-x)^{4p}}{4p(4p-1)} \ (p=1,\ldots,n)$$

sowie $\tilde{\lambda}_2 = 2.25\pi^2$ als grobe untere Schranke für den zweitkleinsten positiven Eigenwert λ_2, so ergeben sich die in Tabelle 1 zusammengestellten Einschließungsintervalle für λ_1.

Zur Berechnung von Lehmann-Schranken für weitere Eigenwerte ist der Einsatz des Stufenverfahrens von Goerisch [5] mit der Schar von Vergleichsaufgaben

$$-\phi_\alpha''(x) = \lambda_\alpha\{\alpha\phi_\alpha(1-x)+(1-\alpha)\phi_\alpha(x)\} \text{ in } [0,1];$$

$$\phi_\alpha(0) = 0, \ \phi_\alpha'(1) = 0 \quad (0 \leq \alpha \leq 1)$$

erforderlich.

Tabelle 1 Nach den Verfahren von Ritz und Lehmann berechnete Schranken für den kleinsten positiven Eigenwert λ_1 der Aufgabe 1.

n	Lehmann	Ritz
1	3.5O6	3.553
2	3.516 O15 266	3.516 O15 281
3	3.516 O15 268 5OO 151 17	3.516 O15 268 5OO 151 26

Aufgabe (1) ist auch geschlossen lösbar: Aus den - jeweils gemäß

$$0 < \rho_1 < \rho_2 < \dots$$

angeordneten - Nullstellen ρ_ν ($\nu \in \mathbb{N}$) der Gleichungen

$$\tan\frac{\rho}{2}\cdot\tanh\frac{\rho}{2} - 1 = 0 \quad \text{bzw.} \quad \tan\frac{\rho}{2}\cdot\tanh\frac{\rho}{2} + 1 = 0$$

ergeben sich die Eigenwerte zu

$$\lambda_\nu = \rho_\nu^2 \ (\nu \in \mathbb{N}) \qquad \text{bzw.} \quad \lambda_{-\nu} = -\rho_\nu^2 \ (\nu \in \mathbb{N})$$

und die zugehörigen Eigenfunktionen zu

$$\phi_\nu(x) = \sin\rho_\nu\{\cosh\rho_\nu x - \cosh\rho_\nu(1-x)\} + \sinh\rho_\nu\{\cos\rho_\nu x + \cos\rho_\nu(1-x)\} \ (\nu \in \mathbb{N})$$

bzw.

$$\phi_{-\nu}(x) = \sin\rho_\nu\{\sinh\rho_\nu x + \sinh\rho_\nu(1-x)\} + \sinh\rho_\nu\{\sin\rho_\nu x - \sin\rho_\nu(1-x)\} \ (\nu \in \mathbb{N}).$$

2. Die Eigenwertaufgabe [3]

$$-(\Delta\phi)(x,y) = \lambda\phi(-x,-y) \text{ in } \Omega;\ \phi = 0 \text{ auf } \partial\Omega^+,\ \frac{\partial\phi}{\partial n} = 0 \text{ auf } \partial\Omega^-, \quad (2)$$

wobei

$$\Omega := \{(x,y) \in \mathbb{R}^2 : |x|<1, |y|<1\},$$
$$\Omega^+ := \{(x,y) \in \mathbb{R}^2 : x=1,\ |y|\le 1\} \cup \{(x,y) \in \mathbb{R}^2 : |x|\le 1,\ y=1\},$$
$$\Omega^- := \{(x,y) \in \mathbb{R}^2 : x=-1, |y|\le 1\} \cup \{(x,y) \in \mathbb{R}^2 : |x|\le 1,\ y=-1\}$$

ist, lautet in schwacher Form

$$M(f,\phi) = \lambda N(f,\phi),\ \phi \in D \quad \text{für alle } f \in D; \qquad (2^*)$$

dabei ist

$$M(f,g) := \int_\Omega \operatorname{grad} f \cdot \operatorname{grad} g \, dxdy \quad \text{für } f,g \in D,$$

$$N(f,g) := \int_\Omega f(x,y)g(-x,-y)dxdy \quad \text{für } f,g \in D$$

und

$$D := \{f \in W^{1,2}(\Omega): f = 0 \text{ auf } \partial\Omega^+\}.$$

Aus der Schwarzschen Ungleichung folgt

$$\left|\int_\Omega f(x,y)f(-x,-y)dxdy\right| \leq \int_\Omega f(x,y)^2 dxdy \quad \text{für } f \in D;$$

daher kann die Theorie regulärer symmetrischer Eigenwertaufgaben [10] auf (2*) angewendet und die Eigenwertaufgabe

$$-(\Delta\tilde{\phi})(x,y)=\tilde{\lambda}\tilde{\phi}(x,y) \text{ in } \Omega;\ \tilde{\phi}=0 \text{ auf } \partial\Omega^+,\ \frac{\partial\tilde{\phi}}{\partial n}=0 \text{ auf } \partial\Omega^-$$

als Vergleichsaufgabe herangezogen werden.

Das Ritzsche Verfahren führt in der Symmetrieklasse

$$D^S := \{f \in D: f(y,x) = f(x,y)\}$$

mit den Ansatzfunktionen

$$v^S_{p,q}(x,y)=(1-x)^p(1-y)^q+(1-x)^q(1-y)^p$$
$$(q=1,\ldots,p \text{ für } p=1,\ldots,n)$$

und in der Symmetrieklasse

$$D^A := \{f \in D: f(y,x) = -f(x,y)\}$$

mit den Ansatzfunktionen

$$v^A_{p,q}(x,y)=(1-x)^{p+1}(1-y)^q-(1-x)^q(1-y)^{p+1}$$
$$(q=1,\ldots,p \text{ für } p=1,\ldots,n)$$

auf die in Tabelle 2 wiedergegebenen Schranken $\bar{\lambda}^S_{1,n}$, $\underline{\lambda}^S_{-1,n}$ für λ^S_1, λ^S_{-1} und $\underline{\lambda}^A_{-1,n}$, $\bar{\lambda}^A_{1,n}$ für λ^A_{-1}, λ^A_1.

Entsprechende Schranken $\underline{\lambda}^S_{1,n}$, $\bar{\lambda}^S_{-1,n}$ für λ^S_1, λ^S_{-1} sind mit dem Stufenverfahren von Goerisch [5] und der von Schellhaas [9] angegebenen Variante des Verfahrens von Lehmann berechnet worden. Man erhält so z. B. die Einschließung

$$2.170\,851 \leq \lambda^S_1 \leq 2.170\,872\,.$$

3. Die Eigenwertaufgabe [2]

$$-\phi''(x) = \lambda\phi(\tfrac{x}{2}) \text{ in } [-1,1];\ \phi(-1) = 0,\ \phi(1) = 0 \qquad (3)$$

ist der Eigenwertaufgabe

$$\phi(x) = \lambda P\phi(x) \text{ in } [-\tfrac{1}{2},\tfrac{1}{2}] \qquad (3^*)$$

Tabelle 2 Nach dem Verfahren von Ritz berechnete Schranken für die ersten Eigenwerte der Aufgabe 2.

n	$\overline{\lambda}^S_{1,n}$	$\underline{\lambda}^S_{-1,n}$	$\underline{\lambda}^A_{-1,n}$	$\overline{\lambda}^A_{1,n}$
1	6		-26	
3	2.175 1	-6.687	-8.381	18.52
5	2.170 92	-6.522 7	-8.259 78	17.641
7	2.170 875	-6.520 89	-8.259 188	17.628 16
9	2.170 871 6	-6.520 811	-8.259 184 7	17.628 088 4
11	2.170 871 12	-6.520 800 2	-8.259 184 52	17.628 088 11

mit dem Integraloperator

$$Pf(x) := \int_{-1/2}^{1/2} K(x,2\eta)f(\eta)d\eta \text{ in } [-\tfrac{1}{2},\tfrac{1}{2}] \text{ für } f \in C[-\tfrac{1}{2},\tfrac{1}{2}],$$

wobei

$$K(x,\xi) := \begin{cases} (1-x)(1+\xi) & \text{für } \xi \le x \\ (1+x)(1-\xi) & \text{für } x \le \xi \end{cases}$$

ist, äquivalent. In beiden Symmetrieklassen,

$D^S := \{f \in C[-\tfrac{1}{2},\tfrac{1}{2}]: f(-x)=f(x)\}$ und $D^A := \{f \in C[-\tfrac{1}{2},\tfrac{1}{2}]: f(-x)=-f(x)\}$,

wird durch

"$f \le g$ bedeutet $f(x) \le g(x)$ für $x \in [0,\tfrac{1}{2}]$"

eine Halbordnung definiert; P ist dann positiv, so daß eine Verallgemeinerung [8] des Collatzschen Quotientensatzes für positive Matrizen [1] anwendbar ist:

"Wähle $v_0 \in \begin{Bmatrix} D^S \\ D^A \end{Bmatrix}$, $v_0 \ge 0$, berechne $v_{k+1} := Pv_k$ $(k=0,1,2,\ldots)$

und bestimme

$$m_k := \inf_{x \in (0,\frac{1}{2}]} \frac{v_k(x)}{v_{k+1}(x)}, \quad M_k := \sup_{x \in (0,\frac{1}{2}]} \frac{v_k(x)}{v_{k+1}(x)} \quad (k=0,1,2,\ldots).$$

Dann gilt die Einschließung

$$m_0 \le m_1 \le m_2 \le \ldots \le \begin{Bmatrix} \lambda^S_1 \\ \lambda^A_1 \end{Bmatrix} \le \ldots \le M_2 \le M_1 \le M_0."$$

Einige Ergebnisse:

$$v_o = 1,\ v_1 = \frac{1-x^2}{2},\ v_2 = \frac{(1-x^2)(23-x^2)}{96},\ \ldots$$

$$2.0905 < m_2 \le \lambda_1^S \le M_2 < 2.0915,$$

$$2.090657 < m_3 \le \lambda_1^S \le M_3 < 2.090695,$$

$$2.0906632 < m_4 \le \lambda_1^S \le M_4 < 2.0906648;$$

$$v_o = x,\ v_1 = \frac{x(1-x^2)}{12},\ v_2 = \frac{x(1-x^2)(37-3x^2)}{5760},\ \ldots$$

$$12.97 < m_2 \le \lambda_1^A \le M_2 < 13.25,$$

$$13.0543 < m_3 \le \lambda_1^A \le M_3 < 13.0560,$$

$$13.05481 < m_4 \le \lambda_1^A \le M_4 < 13.05494.$$

Potenzreihenansätze [2] in D^S und D^A führen auf

$$\phi^S(x) = \sum_{k=0}^{\infty} (-1)^k \frac{\lambda^k}{2^{k^2-k}} \frac{x^{2k}}{(2k)!} \text{ und } \phi^A(x) = \sum_{k=0}^{\infty} (-1)^k \frac{\lambda^k}{2^{k^2}} \frac{x^{2k+1}}{(2k+1)!}$$

und weiter auf die Gleichungen

$$P^S(\lambda) := \sum_{k=0}^{\infty} (-1)^k \frac{\lambda^k}{2^{k^2-k}(2k)!} = 0 \text{ und } P^A(\lambda) := \sum_{k=0}^{\infty} (-1)^k \frac{\lambda^k}{2^{k^2}(2k+1)!} = 0.$$

Die Nullstellen

$$\lambda_{\nu,n}^S \ (\nu=1,\ldots,n) \text{ und } \lambda_{\nu,n}^A \ (\nu=1,\ldots,n)$$

der Polynome

$$P_n^S(\lambda) := \sum_{k=0}^{n} (-1)^k \frac{\lambda^k}{2^{k^2-k}(2k)!} \text{ und } P_n^A(\lambda) := \sum_{k=0}^{n} (-1)^k \frac{\lambda^k}{2^{k^2}(2k+1)!}$$

sind dann Näherungen (s. Tabelle 3) für die Eigenwerte

$$\lambda_\nu^S \ (\nu=1,\ldots,n) \text{ und } \lambda_\nu^A \ (\nu=1,\ldots,n).$$

Herrn Prof. Dr. F. Goerisch und Herrn Dr. P. Klein danke ich herzlich für wertvolle Hinweise, Frau Z. He (Shanghai) für die sorgfältige Bearbeitung von Beispielen.

Tabelle 3 Durch Potenzreihenansätze gewonnene (alternierende) Näherungen für die ersten Eigenwerte der Aufgabe 3.

n	$\lambda^S_{1,n}$	$\lambda^A_{1,n}$
1	2	12
2	2.091 ...	13.06 ...
3	2.090 663 2 ...	13.054 83 ...
4	2.090 663 486 77 ...	13.054 850 017 ...
5	2.090 663 486 756 268 1 ...	13.054 850 013 176 3 ...

n	$\lambda^S_{2,n}$	$\lambda^A_{2,n}$
2	45. ...	146. ...
3	51.7 ...	170.3 ...
4	51.608 ...	169.726 ...
5	51.609 092 5 ...	169.728 650 ...
6	51.609 092 458 43 ...	169.728 649 376 5 ...

Literatur

[1] Collatz, L. (1942) Einschließungssatz für die charakteristischen Zahlen von Matrizen. Math. Z. 48, 221-226.

[2] Collatz, L. (1949) Eigenwertaufgaben mit technischen Anwendungen (Akademische Verlagsgesellschaft Geest und Portig, Leipzig).

[3] Collatz, L. (1950) Numerische Behandlung von Differentialgleichungen (Springer-Verlag, Berlin-Göttingen-Heidelberg).

[4] Goerisch, F. (1985) Eigenwertschranken für Eigenwertaufgaben mit partiellen Differentialgleichungen. Z. Angew. Math. Mech. 65, 129-135.

[5] Goerisch, F. (1987) Ein Stufenverfahren zur Berechnung von Eigenwertschranken. Int. Ser. Numer. Math. 83, 104-114.

[6] Lehmann, N.J. (1963) Optimale Eigenwerteinschließungen. Numer. Math. 5, 246-272.

[7] Liu, K.M., Ortiz, E.L. (1989) Numerical Solution of Ordinary and Partial Functional-Differential Eigenvalue Problems. Computing 41, 205-217.

[8] Mewborn, A.C. (1960) Generalizations of some theorems on positive matrices to completely continuous linear transformations on a normed linear space. Duke Math. J. 27, 273-281.

[9] Schellhaas, H. (1968) Ein Verfahren zur Berechnung von Eigenwertschranken mit Anwendung auf das Beulen von Rechteckplatten. Ing.-Archiv 37, 243-250.

[10] Stummel, F. (1969) Rand- und Eigenwertaufgaben in Sobolewschen Räumen (Lecture Notes in Mathematics 102; Springer-Verlag, Berlin-Heidelberg-New York).

Prof. Dr. Julius Albrecht, Institut für Mathematik, Technische Universität Clausthal, Erzstraße 1,
D 3392 Clausthal-Zellerfeld

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

AN ENCLOSURE METHOD WITH HIGHER ORDER OF CONVERGENCE-APPLICATIONS TO THE ALGEBRAIC EIGENVALUE PROBLEM

G. Alefeld , University of Karlsruhe

B. Illg , University of Karlsruhe

F. Potra , University of Iowa

Dedicated to L. Collatz on the occassion of his 80th birthday

1. Introduction

In [1] we have considered the nonlinear equation $f(x) = 0$ where f is a continuous differentiable real function of a real variable. We suppose that f is strictly monotone on an interval X^0 . Without loss of generality we may assume that f is strictly increasing on X^0 . We assume that by using interval arithmetic methods it is possible to compute two positive numbers ℓ_1 , ℓ_2 such that $0 < \ell_1 \leq f'(x) \leq \ell_2$ for all $x \in X^0$. Let us denote by L the interval $[\ell_1,\ell_2]$. We suppose that the derivative $f'(x) \in \mathbb{R}$, $x \in X^0$, has an interval extension $f'(X)$, $X \subseteq X^0$, satisfying the following conditions

$$f'(x) \in f'(X) , \qquad x \in X \subseteq X^0$$
$$f'(X) \subseteq f'(Y) , \qquad X \subseteq Y \subseteq X^0$$
$$d(f'(X)) \leq c\, d(X) , \qquad X \subseteq X^0$$

where c is a constant independent of X and where d denotes the diameter of an interval. Furthermore we assume that these three relations also hold for the second derivative of f. Together with f and its derivatives we consider its divided differences

$$f[x,y] = \begin{cases} \dfrac{f(x)-f(y)}{x-y} & \text{if } x \neq y \\ f'(x) & \text{if } x = y \end{cases} ,$$

$$f[x,y,z] = \begin{cases} \dfrac{f[x,z]-f[y,z]}{x-y} & \text{if } x \neq y \\ \dfrac{f''(x)}{2} & \text{if } x = y \end{cases} .$$

Then for any nonnegative integer p we can define the following iterative procedure.

Algorithm S_p : For $k = 0,1,...$ DO through ES

$x^k = m(X^k)$

if $k = 0$ then $Q^k = L$, $Y^k = X^0$ & GOTO E1

else

$M^k = \{ f[x^k, x^{k-1,p}] + \frac{1}{2} f''(X^{k-1}) (X^k - x^{k-1,p}) \} \cap L$

$Y^k = \{ x^k - f(x^k)/M^k \} \cap Y^k$

$Q^k = \{ f[x^k, x^{k-1,p}] + \frac{1}{2} f''(X^{k-1}) (Y^k - x^{k-1,p}) \} \cap L$

E1 $X^{k,1} = \{ x^k - f(x^k)/Q^k \} \cap X^k$

if $p = 0$ then $x^{k,p} = x^k$ & GOTO ES

else

$x^{k,1} = m(X^{k,1})$

$M^{k,1} = \{ f[x^k, x^{k,1}] + \frac{1}{2} f''(X^k)(X^{k,1} - x^k) \} \cap L$

$Y^{k,1} = \{ x^{k,1} - f(x^{k,1})/M^{k,1} \} \cap X^{k,1}$

$Q^{k,1} = \{ f[x^k, x^{k,1}] + \frac{1}{2} f''(X^k)(Y^{k,1} - x^k) \} \cap L$

$X^{k,2} = \{ x^{k,1} - f(x^{k,1})/Q^{k,1} \} \cap Y^{k,1}$

if $p = 1$ then GOTO ES

else for $i = 2,3,...,p$ DO through E2

$x^{k,i} = m(X^{k,i})$

$M^{k,i} = \{ f[x^{k,i-1}, x^{k,i}] + \frac{1}{2} f''(X^k)(X^{k,i} - x^{k,i-1}) \} \cap L$

E2 $X^{k,i+1} = \{ x^{k,i} - f(x^{k,i})/M^{k,i} \} \cap X^{k,i}$

ES $X^{k+1} = X^{k,p+1}$ □

For S_p we have the following result.

Theorem. Assume that $f(x) = 0$ has a zero x^* in X^0. Moreover assume that the assumptions mentioned before hold. Then the sequence $\{ X^k \}$ generated by S_p is convergent to x^*. Moreover the sequence of diameters $\{ d(X^k) \}$ converges to zero with R-order ω_p defined as

$$\omega_p = (f_p + 2f_{p+2} - 1 + \sqrt{12f_{p+2}^2 + 9f_p^2 - 20f_p f_{p+2} - 4f_{p+2} + 2f_p + 1})/2$$

where f_j denotes the j-th Fibonacci number, i.e.

$$f_0 = 0 \ , \quad f_1 = 1 \ , \quad f_{j+1} = f_j + f_{j-1} \ , \quad j = 1,2,... \quad .$$

□

A proof can be found in [1].

Under the assumption that the cost of an interval evaluation of the second derivative is about the same as a function evaluation, the efficiency index of the algorithm in the sense of Ostrowski is given by

$$\mathrm{eff}(S_p) = \sqrt[p+2]{\omega_p} \quad .$$

In table I we give the values of ω_p and $(\omega_p)^{1/(p+2)}$ for $p = 0,1,2,...,10$.

Table I. Order and efficiency index of S_p

p	ω_p	$\sqrt[p+2]{\omega_p}$
0	2.00000000000 E+00	1.41421356237 E+00
1	3.73205087570 E+00	1.55113351807 E+00
2	6.46410161514 E+00	1.59450925267 E+00
3	1.10000000000 E+01	1.61539426620 E+00
4	1.82736184955 E+01	1.62294608383 E+00
5	3.00996688705 E+01	1.62638403519 E+00
6	4.92032386541 E+01	1.62741835990 E+00
7	8.01372644808 E+01	1.62756060099 E+00
8	1.30176682947 E+02	1.62724640258 E+00
9	2.11151549918 E+02	1.62677223759 E+00
10	3.42166585524 E+02	1.62624684244 E+00

It can be proved that

$$\omega_p > \left(\frac{1+\sqrt{5}}{2}\right)^{p+2} \quad \text{for } p \geq 4 \ ,$$

$$\lim_{p\to\infty} \sqrt[p+2]{\omega_p} = \frac{1+\sqrt{5}}{2} \ ,$$

$$\max_{p\geq 0} \sqrt[p+2]{\omega_p} = \sqrt[9]{\omega_7} = 1.627... \ .$$

2. The Method for Systems

In the present paper we show how the method repeated in the introduction can be generalized to systems of equations and give some applications to the algebraic eigenvalue problem. For the formulation of the method we need some definitions. Assume that $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is a mapping which has continuous partial derivatives. Then for every pair $x,y \in D$ we define an $n\times n$ matrix $f[x,y]$ by

$$(1) \qquad f[x,y]_{ij} = \begin{cases} \dfrac{1}{x_j - y_j} \{ f_i(x_1,\ldots,x_j,y_{j+1},\ldots,y_n) - \\ \qquad - f_i(x_1,\ldots,x_{j-1},y_j,\ldots,y_n) \} & \text{for } x_j \neq y_j \\ \dfrac{\partial f_i}{\partial x_j}(x_1,\ldots,x_j,y_{j+1},\ldots,y_n) & \text{for } x_j = y_j \ . \end{cases}$$

The matrix $f[x,y]$ is called a "Steigung" or a divided difference operator. It was used by J.W. Schmidt in [3] where the generalization of the Regula falsi to systems of equations was investigated.

For a given mapping $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ and a fixed $z \in D$ we define $\varphi_z : D \subseteq \mathbb{R}^n \to L(\mathbb{R}^n)$ by

$$\varphi_z(x) = f[x,z]$$

and similarly $\psi_x : D \subseteq \mathbb{R}^n \to L(\mathbb{R}^n)$ by

$$\psi_x(z) = f[x,z] \ .$$

We define the divided difference operator of the second order $f[x,y,z]^1$ by applying (1) to the columns of ψ_x. $f[x,y,z]^3$ is defined similarly. The divided difference operators of the second order are bilinear operators.

For a mapping $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ which has second order partial derivatives with an interval arithmetic evaluation for an interval vector $[u] \subseteq D$, we define three dimensional arrays of intervals $\Delta_s([u]) = (\Delta_s([u])_{ijk})$, $s = 1,2,3$, as follows:

$$\Delta_1([u])_{ijk} = \begin{cases} \frac{1}{2} \dfrac{\partial^2 f_i([u])}{\partial x_j^2} & \text{for} \quad j = k \\ \dfrac{\partial^2 f_i([u])}{\partial x_j \partial x_k} & \text{for} \quad k > j \\ 0 & \text{otherwise} \end{cases} ,$$

$$\Delta_2([u])_{ijk} = \begin{cases} \frac{1}{2} \dfrac{\partial^2 f_i([u])}{\partial x_j^2} & \text{for} \quad j = k \\ 0 & \text{for} \quad k > j \\ \dfrac{\partial^2 f_i([u])}{\partial x_j \partial x_k} & \text{for} \quad k < j \end{cases} ,$$

$$\Delta_3([u])_{ijk} = \frac{1}{2}\frac{\partial^2 f_i([u])}{\partial x_j \partial x_k} \qquad \text{for} \quad i,j,k = 1(1)n \quad .$$

Using these arrays it can be shown that

$$(f[x,y,z]^1(y-z))_{ij} \in (\Delta_1(x \cup y \cup z)(y-z))_{ij} \ ,$$

$$(f[y,z,x]^3(y-z))_{ij} \in (\Delta_2(x \cup y \cup z)(y-z))_{ij} \ ,$$

$$((f[x,y,z]^1 + f[y,z,x]^3)(y-z))_{ij} \in (\Delta_3(x \cup y \cup z)(y-z))_{ij}$$

where $x \cup y \cup z$ denotes the smallest interval vector containing x, y and z .

Now we set

$$\delta_s(x,u) = \begin{cases} f[x,u] & \text{for} \quad s = 1 \\ f[u,x] & \text{for} \quad s = 2 \\ \frac{1}{2}(f[x,u] + f[u,x]) & \text{for} \quad s = 3 \quad . \end{cases}$$

Assume that the mapping $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ has partial derivatives of second order which can be evaluated in the interval arithmetic sense. Assume that $[L]$ is an interval matrix with $f'(x) \in [L]$ for all $x \in [x]^0$ and that Gaussian elimination can be performed with $[L]$ and an arbitrary interval vector $[u]$. (The result is denoted by IGA $([L],[u])$.) For a given interval vector $[x]$ we denote by $m[x]$ the center of $[x]$. Let $p \geq 0$ be some fixed integer. Then we consider the following iteration methods for $s = 1,2,3$.

Algorithm S_p^s : For k = 0,1,... DO through ES

$x^k = m([x]^k)$

if k = 0 then $[Q]^k = [L]$, $[y]^k = [x]^0$ & GOTO E1

else

$[M]^k = \{\delta_S(x^k,x^{k-1,p}) + \Delta_S([x]^{k-1})([x]^k - x^{k-1,p})\} \cap [L]$

$[y]^k = \{x^k - IGA([M]^k,f(x^k))\} \cap [x]^k$

$[Q]^k = \{\delta_S(x^k,x^{k-1,p}) + \Delta_S([x]^{k-1})([y]^k - x^{k-1,p})\} \cap [L]$

E1 $[x]^{k,1} = \{x^k - IGA([Q]^k,f(x^k))\} \cap [y]^k$

if p = 0 then $x^{k,p} = x^k$ & GOTO ES

else

$x^{k,1} = m[x]^{k,1}$

$[M]^{k,1} = \{\delta_S(x^{k,1},x^k) + \Delta_S([x]^k)([x]^{k,1} - x^k)\} \cap [L]$

$[y]^{k,1} = \{x^{k,1} - IGA([M]^{k,1},f(x^{k,1}))\} \cap [x]^{k,1}$

$[Q]^{k,1} = \{\delta_S(x^{k,1},x^k) + \Delta_S([x]^k)([y]^{k,1} - x^k)\} \cap [L]$

$[x]^{k,2} = \{x^{k,1} - IGA([Q]^{k,1},f(x^{k,1}))\} \cap [y]^{k,1}$

if p = 1 then GOTO ES else

for i = 2,3,...,p DO through E2

$x^{k,i} = m[x]^{k,i}$

$[M]^{k,i} = \{\delta_S(x^{k,i},x^{k,i-1}) + \Delta_S([x]^k)([x]^{k,i} - x^{k,i-1})\} \cap [L]$

E2 $[x]^{k,i+1} = \{x^{k,i} - IGA([M]^{k,i},f(x^{k,i}))\} \cap [x]^{k,i}$

ES $[x]^{k+1} = [x]^{k,p+1}$ □

If $f(x^*) = 0$ for some $x^* \in [x]^0$ (under our assumptions x^* is unique) then $x^* \in [x]^k$, $k \geq 0$.

Furthermore it has been proved under appropriate assumptions that the R-order of convergence of S_p^s is the same as in the one dimensional case.

4. Applications

In general the method S_p^s seems not to be very attractive for systems in n unknowns since one needs approximately n^3 interval arithmetic evaluations for the second order partial derivatives. However, there are some important cases in which one needs less work.

a) Consider the nonlinear integral equation

$$\int_0^1 K(t,s,x(t))dt = x(s) , \quad s \in [0,1]$$

for the unknown function $x(s)$. For the numerical solution of this equation we choose equidistant points $s_i = \frac{i}{n}$, $i = 0(1)n$, and use one of the well known numerical integration formulas. Omitting the discretization error, we get a nonlinear system $f(x) = 0$ with

$$f_i(x_0,\ldots,x_n) = x_i - \sum_{j=0}^{n} w_j K(\tfrac{j}{n},\tfrac{i}{n},x_j) , \quad i = 0(1)n ,$$

for the unknowns x_i , $i = 0(1)n$, which are considered as approximations to $x(\frac{i}{n})$, $i = 0(1)n$. It follows that

$$\frac{\partial f_i}{\partial x_j} = \delta_{ij} - w_j K_u(\tfrac{j}{n},\tfrac{i}{n},x_j) \quad , \quad i = 0(1)n \quad , \quad j = 0(1)n$$

and

$$\frac{\partial^2 f_i}{\partial x_j \partial x_k} = \begin{cases} -w_j K_{uu}(\tfrac{j}{n},\tfrac{i}{n},x_j) & \text{for} \quad j = k \\ 0 & \text{otherwise} \end{cases} \quad , \; i = 0(1)n \, , \; j = 0(1)n \, , \; k = 0(1)n \, .$$

(δ_{ij} denotes the Landau-symbol, K_u and K_{uu} denote the first and second order partial derivative with respect to the third variable of K). Hence in this case f'' has only $(n+1)^2$ elements different from zero.

b) A similar result as in a) holds if a solution of the boundary value problem

$$y'' = f(t,y)$$
$$y(a) = \alpha \, , \quad y(b) = \beta$$

is approximated by the usual method of differences.

c) Even more spectacular than in the two proceeding examples is the saving of arithmetic operations for the algebraic eigenvalue problem. Consider the eigenvalue problem for the matrix A . If we define the vector $x = (z^T, \lambda)^T$ then

$$Az = \lambda z$$
$$z^T z = 1$$

is equivalent to the nonlinear system

$$f(x) = 0$$

where

$$f_i(x) = (a_{ii} - x_{n+1})\, x_i + \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}\, x_j \quad , \qquad i = 1\,(1)\,n$$

$$f_{n+1}(x) = \sum_{i=1}^{n} z_i^2 - 1 \quad .$$

In this case we get

$$\frac{\partial f_i(x)}{\partial x_j} = \begin{cases} a_{ij} - \delta_{ij}\, x_{n+1} & , \quad 1 \le i \le n \quad , \quad 1 \le j \le n \\ -\, x_i & , \quad 1 \le i \le n \quad , \quad j = n+1 \\ 2\, x_j & , \quad i = n+1 \quad , \quad 1 \le j \le n \\ 0 & , \quad i = j = n+1 \end{cases} ,$$

$$\frac{\partial^2 f_i(x)}{\partial x_j\, \partial x_k} = \begin{cases} 0 & , \quad 1 \le i \le n \quad , \quad 1 \le j \le n \quad , \quad 1 \le k \le n \\ -\, \delta_{ij} & , \quad 1 \le i \le n \quad , \quad 1 \le j \le n \quad , \quad k = n+1 \\ -\, \delta_{ik} & , \quad 1 \le i \le n \quad , \quad j = n+1 \quad , \quad 1 \le k \le n+1 \\ 2\, \delta_{jk} & , \quad i = n+1 \quad , \quad 1 \le j \le n \quad , \quad 1 \le k \le n+1 \\ 0 & , \quad i = j = n+1 \quad , \quad 1 \le k \le n+1 \end{cases} ,$$

$$f[x,y]_{ij} = \begin{cases} a_{ij} - \delta_{ij}\, y_{n+1} & , \quad 1 \le i \le n \quad , \quad 1 \le j \le n \\ -\, x_i & , \quad 1 \le i \le n \quad , \quad j = n+1 \\ x_j + y_j & , \quad i = n+1 \quad , \quad 1 \le j \le n \\ 0 & , \quad i = j = n+1 \end{cases} .$$

Hence the second derivative is constant and many of its elements are equal to zero. Similarly f[x,y] can be formed nearly without arithmetical work. Hence S_p^s can be performed with simple available operators.

4. Numerical Examples

a) The matrix

$$A = \begin{bmatrix} 33 & 16 & 72 \\ -24 & -10 & -57 \\ -8 & -4 & -17 \end{bmatrix}$$

has an eigenpair $x = (z^T, \lambda)^T$ which is contained in the intervalvector

$$[x]^o = \begin{bmatrix} [-0.765, -0.764] \\ [0.611, 0.612] \\ [0.203, 0.204] \\ [0.991, 1.001] \end{bmatrix} .$$

The following table II contains the numerical results obtained by applying S_p^3 for different values of p. For a fixed p the integer k denotes the number of iteration steps until the lower and upper bounds of the iterates $[x]^k$ differ by at most one unit of the last digit in the mantissa. (We are using a computer with 12 decimal digits in the mantissa.) f denotes the number of function evaluations and IGA is the number of applications of Gaussian elimination.

Table II			
p	k	f	IGA
0	2	3	4
1	1	3	4
2	1	3	4
3	0	3	4
4	0	3	4
5	0	3	4
6	0	3	4

b) The matrix

$$A = \begin{bmatrix} -2 & 1 & 0 & 27 & -18 & -6 \\ -8 & 4 & 0 & 54 & -36 & -12 \\ -8 & -5 & 6 & 81 & -54 & -18 \\ -8 & -5 & -6 & 117 & -72 & -24 \\ -8 & -5 & -6 & 129 & -78 & -30 \\ -8 & -5 & -6 & 129 & -60 & -48 \end{bmatrix}$$

has an eigenpair $x = (z^T,\lambda)^T$, which is contained in the intervalvector

$$[x]^0 = \begin{bmatrix} [\ 0.127\ ,\ 0.128\ ] \\ [\ 0.254\ ,\ 0.255\ ] \\ \vdots \\ [\ 0.508\ ,\ 0.509\ ] \\ [11.991\ ,\ 12.01\ ] \end{bmatrix} .$$

The values in table III have the analogous meaning as in table II .

Table III			
p	k	f	IGA
0	2	3	4
1	1	3	4
2	1	3	4
3	0	3	4
4	0	3	4
5	0	3	4
6	0	3	4
7	0	3	4

References

[1] Alefeld, G., Potra, F.: A new class of interval methods with higher order of convergence. Computing 42, 69-80 (1989).

[2] Illg, B.: Über einige Verfahren höherer Ordnung zur iterativen Einschließung bei nichtlinearen Gleichungssystemen. Diplomarbeit. Universität Karlsruhe 1989. (Not available).

[3] Schmidt, J. W.: Eine Übertragung der Regula falsi auf Gleichungen in Banachräumen II. ZAMM 43, 97-110 (1963).

G. Alefeld , B. Illg
Institut für Angewandte Mathematik
Universität Karlsruhe
Kaiserstraße 12
D-7500 Karlsruhe
Deutschland

F. Potra
Department of Mathematics
University of Iowa
Iowa City , 117 522 42
U. S. A.

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

SOME REMARKS CONCERNING CLOSURE RATES FOR ARONSZAJN'S METHOD

C. A. Beattie and W. M. Greenlee

1. Introduction.

In various problem settings occurring in engineering and scientific applications, it is desirable to compute a tight bracketing interval of some predetermined (small) width guaranteed to contain a selected eigenvalue of a differential operator. We restrict ourselves here to the consideration of a selfadjoint operator A, densely defined on a Hilbert space $\mathcal{H}$, that is furthermore semi–bounded below with the lower portion of the spectrum consisting of discrete eigenvalues with finite multiplicity.

Obtaining such an inclusion bracket is evidently equivalent to calculating upper and lower bounds to the eigenvalue of interest. High quality upper bounds are often available via the Rayleigh–Ritz method. This approach combines great flexibility and economy in application with good theoretical properties both consistent with scientific intuition and conducive to the use of highly refined numerical techniques. For the most part, it is without peer for resolving upper bounds.

The situation regarding lower bound calculation is far less clear cut. If eigenvalue bracketing is to be guaranteed in diverse circumstances for any finite calculation, we must focus attention on variational methods and neglect otherwise useful methods such as defect minimization and Ritz–related residual bounds. All methods capable of producing guaranteed eigenvalue lower bounds require some sort of *a priori* spectral information about A — methods differ primarily in the type of information required and how use is made of it. Temple–Lehmann bounds, for example, can be quite effective if a parameter value separating two adjacent eigenvalues of A with known indices can be obtained. Important extensions to the basic Temple–Lehmann method are reviewed and developed in [14]. Intermediate problem methods such as those of Aronszajn (cf. [1,23,24]) have comparatively relaxed requirements for *a priori* spectral information and it will be those methods that we choose to emphasize in this paper.

Generally speaking, in considering methods available for bracketing eigenvalues of selfadjoint operators, there is a common perception that lower bounds are far more expensive to compute than are the corresponding upper bounds. The expense of any such scheme is related to a complex constellation of competing factors: the difficulty in

assembling the basic approximating data (e.g. the evaluation of inner product matrices); the underlying rate of convergence of the approximation method; the structure of the resulting numerical problem; and the efficiency with which the algorithm selected can utilize this structure. Acknowledging that no brief study such as this can adequately address all these issues, our work here is motivated by the far simpler question: How fast can an inclusion interval be contracted? For our purposes here we interpret this question as: What convergence rates are attainable for Aronszajn's method and how do they compare with corresponding rates for the Rayleigh–Ritz method?

2. The Rayleigh–Ritz Method.

In this section, we fix our notation and briefly summarize some basic facts related to the Rayleigh–Ritz method that we will refer to later. For a thorough development one should consult the excellent books of Chatelin [11] and Babuška and Osborn [2].

Let $\mathcal{H}$ be a separable complex Hilbert space with norm $\|u\|$ and inner product $\langle u, v\rangle$. Let A be a selfadjoint operator with domain $Dom\,(A)$ dense in $\mathcal{H}$. We suppose that A is bounded below with spectrum that begins with isolated eigenvalues of finite multiplicity, $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_\infty$, and corresponding orthonormal eigenvectors $u_1, u_2, \ldots$. Here λ_∞ denotes the least point of the essential spectrum of A, where by convention we set $\lambda_\infty = \infty$ if the essential spectrum of A is empty. The closure of the quadratic form $\langle Au, u\rangle$ is denoted by $a(u)$ with $a(u, v)$ representing the associated sesquilinear form. By an appropriate shift in the spectrum we may assume without loss of generality that $a(u) \geq \|u\|^2$.

The Rayleigh–Ritz method begins with the selection of a family of trial vectors $\{v_i\}_{i=1}^n \subset Dom\,(a)$ and then proceeds with the set up and resolution of the generalized matrix eigenvalue problem

$$[a(v_i, v_j)]\mathbf{x} = \Lambda[\langle v_i, v_j\rangle]\mathbf{x}. \tag{2.1}$$

If the resulting matrix eigenvalues are denoted as $\Lambda_1^{(n)} \leq \Lambda_2^{(n)} \leq \cdots \leq \Lambda_n^{(n)}$, the first monotonicity principle guarantees that for each i, $\lambda_i \leq \Lambda_i^{(n)}$, providing us with upper bounds to the lowest n eigenvalues of A.

The convergence rates for the Rayleigh–Ritz methods can be quite rapid. In order to lay out these and subsequent convergence rate results we introduce the following notation (similar to [20]): For any densely defined closed positive coercive quadratic form $c(u)$ on $\mathcal{H}$, $\mathcal{H}_c$ will denote the Hilbert space $Dom\,(c)$ equipped with the norm $\|u\|_c = c(u)^{\frac{1}{2}}$. Given such a $c(u)$, let $\mathcal{M}$ and $\mathcal{N}$ be subspaces of $Dom\,(c)$ with $\dim\mathcal{N} > 0$. The

containment gap relative to $c(u)$ for the approximation of $\mathcal{M}$ by $\mathcal{N}$ is

$$\delta_c(\mathcal{N},\mathcal{M}) = \sup_{0\neq u\in\mathcal{N}} \inf_{v\in\mathcal{M}} \frac{\|u-v\|_c}{\|u\|}.$$

Notice that $\delta_c(\mathcal{N},\mathcal{M})$ is not symmetric in $\mathcal{N}$ and $\mathcal{M}$ and $\delta_c(\mathcal{N},\mathcal{M}) = 0$ if and only if $\mathcal{M} \supset \mathcal{N}$. This is an adaptation of the notation of [20] (and [7]) to accommodate the use of norms induced by quadratic forms.

The basic convergence rate for the Rayleigh–Ritz method appears as

THEOREM 2.1. *Let $\mathcal{V}_k = span(v_1, v_2, \ldots, v_k) \subset Dom\,(a)$ be given for $k = 1, 2, \ldots$. If $\bigcup_n \mathcal{V}_n$ is a core for $a(u)$ then $\Lambda_i^{(n)} \underset{n\to\infty}{\longrightarrow} \lambda_i$ for each i such that $\lambda_i < \lambda_\infty$ and furthermore*

$$0 \leq \frac{\Lambda_i^{(n)} - \lambda_i}{\lambda_i} \leq C_i \delta_a^2(\mathcal{U}_i, \mathcal{V}_n) \tag{2.2}$$

where $\mathcal{U}_i$ is the span of all eigenvectors corresponding to λ_i and C_i is a constant independent of n.

One may refer to [2] or [11] for a thorough development of these and related ideas. Babuška and Osborn [2] derive refined estimates for the case that λ_i is a multiple eigenvalue. Chatelin [11] gives optimal zero–order estimates for C_i.

In order to make these considerations concrete we pause in our development to give a simple illustration. Let $\mathcal{H} = L^2(0,1)$ and for $u \in H_0^1(0,1)$ — the closure of $C_0^\infty(0,1)$ in the Sobolev space $H^1(0,1)$ — let

$$a(u) = \int_0^1 \left(|u'|^2 + q|u|^2\right)\, dx,$$

where q is a nonnegative function in $C^3[0,1]$. Then $Dom\,(A) = H^2(0,1) \cap H_0^1(0,1)$ and the eigenvalue problem for A means

$$Au = -u'' + qu = \lambda u \text{ in } (0,1) \tag{2.3}$$

$$u(0) = u(1) = 0.$$

If we select $v_k(x) = \sin(k\pi x)$ notice that $-v_k'' = (k\pi)^2 v_k$ and $\{v_k\}_{k=1}^\infty$ forms a "perturbed basis" for the Rayleigh–Ritz problem for A (in the parlance of Birkhoff and Fix [8]). For this choice of projecting vectors, the gap of (2.2) may be bounded as in [8] or using the more general approach of [7] to find:

$$0 \leq \frac{(\Lambda_i^{(n)} - \lambda_i)}{\lambda_i} \leq C_i\, n^{-7}.$$

3. Aronszajn's Method.

In order to apply the method of intermediate problems, we require knowledge of another closed quadratic form $a_0(u)$, satisfying $a_0(u) \leq a(u)$ for all $u \in Dom\,(a)$. Furthermore, we require that the spectral problem for the selfadjoint operator A_0 corresponding to a_0 is solved explicitly and that the spectrum of A_0 also begins with isolated eigenvalues of finite multiplicity, $\lambda_1^0 \leq \lambda_2^0 \leq \lambda_3^0 \leq \cdots \leq \lambda_\infty^0$ with corresponding orthonormal eigenvectors $u_1^0, u_2^0, \ldots$. The second monotonicity principle implies that $\lambda_\infty^0 \leq \lambda_\infty$ and for each i such that $\lambda_i < \lambda_\infty^0$, λ_i^0 exists and $\lambda_i^0 \leq \lambda_i$. As before, we assume without loss of generality that $a_0(u) \geq \|u\|^2$ for all $u \in Dom\,(a_0)$.

Define the difference between $a(u)$ and $a_0(u)$ as $b(u) = a(u) - a_0(u) \geq 0$ with $Dom\,(b) = Dom\,(a) \subset Dom\,(a_0)$. Suppose that $b(u)$ is closable in $\mathcal{H}$ and denote its closure henceforth with no change in notation. There is a densely defined selfadjoint operator, B, associated with $b(u)$ so that $b(u,v) = \langle Bu, v\rangle$ for all $u \in Dom\,(B)$ and $v \in Dom\,(b)$.

Aronszajn's method proceeds by selecting a set of trial vectors $\{p_k\}_{k=1}^\infty \subset Dom\,(B)$ and defining for each n,

$$P_n u = \sum_{i,j=1}^{n} \langle u, Bp_i\rangle \beta_{ij} p_i$$

where $[\beta_{ij}]$ is the matrix inverse to $[b(p_i, p_j)]$. P_n is the b–orthogonal projection onto $\mathcal{P}_n = span\,\{p_1, \ldots, p_n\}$.

Now for each n, define the intermediate quadratic form

$$a_n(u) = a_0(u) + b(P_n u) \tag{3.1}$$

for all $u \in Dom\,(a_n) = Dom\,(a_0)$ with the corresponding selfadjoint operator

$$A_n = A_0 + BP_n \tag{3.2}$$

each with spectrum beginning with eigenvalues $\lambda_1^{(n)} \leq \lambda_2^{(n)} \leq \ldots$. By construction, $a_0(u) \leq a_n(u) \leq a_{n+1}(u) \leq a(u)$ for all $u \in Dom\,(a)$, and thus the second monotonicity principle provides

$$\lambda_i^0 \leq \cdots \leq \lambda_i^{(n)} \leq \lambda_i^{(n+1)} \leq \cdots \leq \lambda_i$$

for all i such that $\lambda_i < \lambda_\infty^0 = \lambda_\infty^{(n)}$. Hence, the intermediate operators, $\{A_n\}$, have eigenvalues that provide improving lower bounds to the eigenvalues of A as the dimension index n is increased. Some practical issues surrounding the computational resolution of these intermediate eigenvalue problems are presented in Section 5. For a complete discussion, one may refer to [13].

Criteria guaranteeing convergence of the intermediate eigenvalues, $\lambda_i^{(n)}$, to the corresponding eigenvalues of A date back to Aronszajn [1] and Bazley and Fox [3] for problems with compact resolvent and relatively bounded perturbations. For eigenvalue problems that admit nontrivial essential spectra and perturbations that are not relatively bounded — as occur for example in many quantum mechanical eigenvalue problems — convergence rates are relatively recent, including those of Beattie [5], Beattie and Greenlee [6,7], Brown [9,10] and Greenlee [18]. We present and prove a result representative of this recent work in a generality that encompasses some of the earlier results of [7] and [10] yet with a somewhat more direct proof.

Recall that we have assumed that $a(u)$ and $a_0(u)$ are closed lower semibounded quadratic forms on $\mathcal{H}$ which satisfy (perhaps after a shift in the spectrum): $a(u) \geq a_0(u) \geq \|u\|^2$. It follows directly from definitions that $b(u) = a(u) - a_0(u) \geq 0$ is a closed quadratic form in $\mathcal{H}_{a_0}$ and hence there exists an a_0–selfadjoint operator $\mathbf{B}$ corresponding to the form $b(u)$: $b(u,v) = a_0(\mathbf{B}u, v)$ for all $u \in Dom(\mathbf{B})$ and $v \in Dom(b)$. While it is true that the operator $\mathbf{B}$ may be difficult to obtain explicitly, it is often the case that $Dom(\mathbf{B})$ can be characterized or at least estimated from within (cf. [15,16,17]), and for the setting we later consider this will suffice to deduce convergence.

THEOREM 3.1. *If projecting subspaces, $\mathcal{P}_n$ for the Aronszajn intermediate problems (3.2) are chosen so that*

$$\mathfrak{D} = \bigcup_u \mathcal{P}_n + Ker(b)$$

is a core for $\mathbf{B}$ then $\lim_n \lambda_i^{(n)} = \lambda_i$ for each i such that $\lambda_i < \lambda_\infty^0$.

PROOF: Consider an index k such that $\lambda_k < \lambda_\infty^0$ and pick ϵ such that $0 < \epsilon < (\lambda_\infty^0 - \lambda_k)/\lambda_\infty^0$. Define $b_+(u) = b(u) + \epsilon a_0(u)$. Then b_+ is a closed quadratic form on both $\mathcal{H}_{a_0}$ and $\mathcal{H}$ with corresponding operators B_+ and $\mathbf{B}_+$ satisfying $b_+(u) = \langle B_+ u, u\rangle$ in $\mathcal{H}$ and $b_+(u) = a_0(\mathbf{B}_+ u, u)$ in $\mathcal{H}_{a_0}$. Note that $b_+(u) \geq \epsilon a_0(u) \geq \epsilon\|u\|^2$. Define also a family of b_+–orthogonal projections, $\hat{P}_n$, so that $Ran(\hat{P}_n) = \mathcal{P}_n + Ker(b)$ and $b_+(\hat{P}_n u, v) = b_+(u, \hat{P}_n v)$ — note that $\mathcal{P}_n + Ker(b)$ need not be finite–dimensional but since $Ker(b)$ is a closed subspace of $\mathcal{H}_{b_+}$ and $\mathcal{P}_n$ is finite–dimensional, $\hat{P}_n$ is well–defined. From an easy modification of [6, Lemma 3.5], we have

$$\hat{a}_n(u) = (1-\epsilon)a_0(u) + b_+(\hat{P}_n u) \leq a_0(u) + b(P_n u) = a_n(u) \leq a(u).$$

Since all $\{\hat{a}_n\}$ are closed (in $\mathcal{H}$), there are corresponding densely defined selfadjoint operators $\{\hat{A}_n\}$ satisfying $(1-\epsilon)A_0 \leq \cdots \leq \hat{A}_n \leq \hat{A}_{n+1} \leq \cdots \leq A$. Because $\{\hat{A}_n\}$ is a monotone increasing family of positive definite selfadjoint operators, $\hat{A}_n$ converges in the strong resolvent sense to a selfadjoint operator $\hat{A}_\infty$ satisfying $\hat{A}_n \leq \hat{A}_\infty \leq A$. Since $\lambda_k < inf\, \sigma_{ess}(\hat{A}_k) = (1-\epsilon)\lambda_\infty^0$, a theorem of Weidmann [22] yields the conclusion

$\lim_n \lambda_k^{(n)} = \lambda_k$ provided we can show that $\hat{A}_\infty = A$. Denoting the closed quadratic form corresponding to $\hat{A}_\infty$ as $\hat{a}_\infty$, it will clearly be sufficient to show $\hat{a}_\infty = a$.

Define now an auxiliary quadratic form $\hat{a}_*(u) = \lim_n \hat{a}_n(u)$ for all u in

$$Dom\,(\hat{a}_*) = \left\{ u \in \bigcap_n Dom\,(\hat{a}_n) | \sup_n \hat{a}_n(u) < \infty \right\} = \left\{ u \in Dom\,(a_0) | \lim_n b_+(\hat{P}_n u) < \infty \right\}$$

Then $Dom\,(\hat{a}_*)$ is a linear subspace of $\mathcal{H}$ and $\hat{a}_*(u)$ is a positive quadratic form on $Dom\,(\hat{a}_*)$. In fact, $\hat{a}_*$ is a least upper bound for the family of forms $\{\hat{a}_n\}$ in the sense that any closed real–valued quadratic form that dominates each $\hat{a}_n$ also dominates $\hat{a}_*$. Hence in particular $Dom\,(\hat{a}_*) \supset Dom\,(a)$ and $\hat{a}_* \leq \hat{a}_\infty \leq a$. We will show $\hat{a}_\infty = a$ by showing $\hat{a}_* = a$.

Since $\mathfrak{D}$ is a core for $\mathbf{B}$ it is additionally a core for $\mathbf{B}_+ = \mathbf{B} + \epsilon I$ and hence for b_+ as well. Thus $\hat{P}_n \to I$ strongly in $\mathcal{H}_{b_+}$ and for $u \in Dom\,(a) = Dom\,(b_+)$, $b_+(u - \hat{P}_n u) = b_+(u) - b_+(\hat{P}_n u) \longrightarrow 0$ as $n \to \infty$. This implies that $\hat{a}_*(u) = a(u)$ for each $u \in Dom\,(a)$. Now we need only show $Dom\,(\hat{a}_*) \subset Dom\,(a)$.

Take $u \in Dom\,(\hat{a}_*)$. Then $\{b_+(\hat{P}_n u)\}$ is a Cauchy sequence in $\mathbb{R}$ and the Pythagorean Theorem implies

$$b_+(\hat{P}_n u - \hat{P}_m u) = \left| b_+(\hat{P}_n u) - b_+(\hat{P}_m u) \right| \longrightarrow 0 \text{ as } n, m \to \infty.$$

Since b_+ is coercive in both $\mathcal{H}_{a_0}$ and $\mathcal{H}$, $\hat{P}_n u$ converges in each of $\mathcal{H}_{b_+}$, $\mathcal{H}_{a_0}$ and $\mathcal{H}$ to a (single) vector w say. Additionally since $b_+ \geq b \geq 0$, $a(\hat{P}_n u - \hat{P}_m u) = a_0(\hat{P}_n u - \hat{P}_m u) + b(\hat{P}_n u - \hat{P}_m u) \longrightarrow 0$ as $n, m \to \infty$. Since a is a closed quadratic form in $\mathcal{H}$ we may conclude that $u \in Dom\,(a)$ once we ascertain that $w = u$.

Let $\hat{p} \in \mathcal{P}_n + Ker\,(b) \subset \mathfrak{D}$ for some index n. Then $a_0(u - w, \mathbf{B}_+ \hat{p}) = \lim_{m \to \infty} a_0([I - \hat{P}_m]u, \mathbf{B}_+ \hat{p}) = \lim_{m \to \infty} a_0(u, \mathbf{B}_+ [I - \hat{P}_m]\hat{p}) = 0$. By allowing $\hat{p}$ to range over $\mathfrak{D}$, which is a core for $\mathbf{B}_+$, we must conclude $w = u$. □

COROLLARY 3.2. *Let the selfadjoint operator B be relatively form bounded with respect to A_0. If $\bigcup_n \mathcal{P}_n + Ker\,(B)$ is a form core for A_0 (i.e., a core for a_0) then $\lim_n \lambda_i^{(n)} = \lambda_i$ for each i satisfying $\lambda_i < \lambda_\infty^0$.*

PROOF: Relative form boundedness implies the existence of $M > 0$ such that $0 \leq b(u) \leq M a_0(u)$ for all $u \in Dom\,(a_0)$. Thus $\mathbf{B}$ is an a_0–bounded operator and any core for a_0 will be a core for $\mathbf{B}$. □

Weaker convergence criteria are available if B is actually a bounded operator on $\mathcal{H}$ as well. The previous method of proof may be repeated defining $b_+(u) = b(u) + \epsilon \|u\|^2$ and with related analysis done in $\mathcal{H}$ now instead of $\mathcal{H}_{a_0}$. The result is

THEOREM 3.3. *Let the selfadjoint operator B be bounded on $\mathcal{H}$. If $\bigcup_n \mathcal{P}_n + Ker\,(B)$ is dense in $\mathcal{H}$ then $\lim_n \lambda_i^{(n)} = \lambda_i$ for each i satisfying $\lambda_i < \lambda_\infty^0$.*

Under a number of additional conditions, Poznyak obtained the following convergence rate for Aronszajn's method.

THEOREM 3.4. *(Poznyak [21]) Let A_0 and A have compact inverses in $\mathcal{H}$. Suppose B is relatively bounded with respect to A_0 and is coercive on $\mathcal{H}$. If $\bigcup_n \mathcal{P}_n$ is a core for b, then $\lambda_i^{(n)} \to \lambda_i$ as $n \to \infty$ and*

$$0 \leq \frac{\lambda_i - \lambda_i^{(n)}}{\lambda_i} \leq c_i \delta_b^2(\mathcal{U}_i, \mathcal{P}_n)$$

where $\mathcal{U}_i$ is the space of all eigenvectors corresponding to λ_i and c_i is a constant independent of n.

Referring to our previous illustrative example (2.3) we may set up Aronszajn's method by selecting $A_0 u = -\dfrac{d^2u}{dx^2}$ for u in $Dom\,(A_0) = H^2(0,1) \cap H_0^1(0,1)$ and $Bu = qu$ for all u in $\mathcal{H} = L_2(0,1)$. Since B is bounded on $\mathcal{H}$, Theorem 3.3 indicates that any choice of $\{p_i\}$ for (3.2) that spans a dense subspace of $L_2(0,1)$ will produce convergent estimates. This is a weaker requirement than is needed for convergence of Rayleigh–Ritz estimates, so the original "perturbed basis" choice $p_k(x) = \sin(k\pi x)$ will produce convergent lower bounds. If additionally $q(x) \geq \epsilon > 0$ for all $x \in (0,1)$ then Theorem 3.4 may be applied (note that A_0 has compact resolvent) to obtain the asymptotic closure rates

$$0 \leq \frac{\lambda_i - \lambda_i^{(n)}}{\lambda_i} \leq c_i\, n^{-9}$$

— faster than Rayleigh–Ritz!

4. Extension of Poznyak's Convergence Rate.

We are able to show here that the convergence rate given by Poznyak [21] remains valid for a much larger class of problems than he had considered originally. With hypotheses sufficient to make sense of the operator sum $A_0 + B$, we find that B need only be relatively form bounded with respect to A_0 and neither compactness of A, A_0 or their inverses, nor coerciveness of B is needed. Thus operators with nontrivial essential spectrum are admissible, allowing a variety of quantum mechanical problems that would otherwise be excluded. We make fundamental use of the Kato–Temple inequalities:

LEMMA 4.1. *(Kato [19]) Let T be a selfadjoint operator on $\mathcal{H}$ and let $\{v_i\}_{i=1}^r \subset Dom\,(T)$ be an orthonormal family of vectors so that $[\langle v_i, Tv_j\rangle] = \underset{k=1,\ldots,r}{diag}\,(\eta_k)$ with $\eta_1 \leq \eta_2 \leq \cdots \leq \eta_r$.*

Suppose an open interval (α, β) *is known such that* $\{\eta_k\}_{k=1}^r \subset (\alpha, \beta)$ *and such that* $\sigma(T) \cap (\alpha, \beta)$ *consists of at most* r *(not necessarily distinct) eigenvalues:* $\tau_1 \leq \tau_2 \leq \ldots$. *If* $\theta_k = \|Tv_k - \eta_k v_k\|$ *denotes the residual corresponding to the approximate eigenvalue* η_k *and if*

$$\sum_{k=1}^{r} \frac{\theta_k^2}{(\eta_k - \alpha)(\beta - \eta_k)} < 1,$$

then there are exactly r *eigenvalues of* T *in* (α, β) *and they are bounded by*

$$\eta_k - \sum_{i=k}^{r} \frac{\theta_i^2}{\beta - \eta_i} \leq \tau_k \leq \eta_k + \sum_{i=1}^{k} \frac{\theta_i^2}{\eta_i - \alpha} \tag{4.1}$$

for each $k = 1, \ldots, r$.

Additionally, we require a technical lemma that extends the validity of a resolvent identity to the form bounded case.

LEMMA 4.2. *Let* A_0 *be selfadjoint and boundedly invertible on* $\mathcal{H}$ *and let* B *be a nonnegative selfadjoint operator with* $Dom(A_0) \cap Dom(B)$ *dense in* $\mathcal{H}$. *Suppose that*

(i) B *is relative form bounded with respect to* A_0, *and*

(ii) $A_0 + B$ *is essentially selfadjoint with unique selfadjoint extension* A.

Then

$$A_n^{-1} - A^{-1} = C_n^* B^{\frac{1}{2}} (I - P_n) A^{-1} \tag{4.2}$$

where

$$C_n = B^{\frac{1}{2}} (I - P_n) A_n^{-1}$$

PROOF: Denote $\mathfrak{A} = Dom(A_0) \cap Dom(B)$ and observe for any $q \in \mathfrak{A}$, $Aq = A_0 q + Bq$ and $A_n q = A_0 q + BP_n q$ are well defined. Thus, $(A - A_n)q = B(I - P_n)q$.

Now pick $p \in A\mathfrak{A}$ arbitrarily and note that

$$(A - A_n)A^{-1}p = B(I - P_n)A^{-1}p$$

so

$$\begin{aligned}(A_n^{-1} - A^{-1})p &= A_n^{-1}(A - A_n)A^{-1}p = A_n^{-1}B(I - P_n)A^{-1}p \\ &= A_n^{-1}(I - P_n^*)B(I - P_n)A^{-1}p \\ &= \left[A_n^{-1}(I - P_n^*)B^{\frac{1}{2}}\right]\left[B^{\frac{1}{2}}(I - P_n)A^{-1}\right]p\end{aligned}$$

Notice that since $Dom(B^{\frac{1}{2}}) = Dom(b) \supset Dom(a_0) \supset Dom(A_0) = Dom(A_n)$ and $Dom(B^{\frac{1}{2}}) \supset Dom(B)$, C_n must be continuous. Similarly $Dom(B^{\frac{1}{2}}) \supset Dom(a_0) =$

$Dom\,(a) \supset Dom\,(A)$ implies that $B^{\frac{1}{2}}(I-P_n)A^{-1}$ is continuous. Thus, $C_n^* B^{\frac{1}{2}}(I-P_n)A^{-1}$ is a continuous operator defined everywhere on $\mathcal{H}$. The proof is completed by noting that since $\mathfrak{A}$ is a core for A, (4.2) holds pointwise on a dense subset, $A\mathfrak{A}$, and hence by continuity, everywhere on $\mathcal{H}$. □

We can now state our extension of Poznyak's convergence theory.

THEOREM 4.3. *Suppose A_0+B is essentially selfadjoint with unique selfadjoint extension A, where A_0 is selfadjoint and coercive with $A_0 \geq \epsilon > 0$ and B is nonnegative and selfadjoint. If B is relatively form-bounded with respect to A_0 and $\bigcup_n \mathcal{P}_n + Ker\,(B)$ forms a core for a_0 then $\lambda_\ell^{(n)} \to \lambda_\ell$ as $n \to \infty$ for each ℓ that satisfies $\lambda_\ell < \lambda_\infty^0$ and*

$$0 \leq \frac{\lambda_\ell - \lambda_\ell^{(n)}}{\lambda_\ell} \leq K_\ell \delta_b^2(\mathcal{U}_\ell, \mathcal{P}_n) \tag{4.3}$$

for some constant K_ℓ independent of n.

PROOF: Let λ_ℓ be an eigenvalue of A with multiplicity r and let $\mathcal{U}_\ell$ be the corresponding r-dimensional eigenspace. For each n, we may choose an orthonormal basis, $\{v_k^{(n)}\}_{k=1}^r$ for $\mathcal{U}_\ell$ so that

$$\left[\langle A_n^{-1} v_i^{(n)}, v_j^{(n)}\rangle\right] = \underset{k=1,\ldots,r}{diag}\,(\eta_k^{(n)})$$

with $0 < \eta_1^{(n)} \leq \eta_2^{(n)} \leq \cdots \leq \eta_r^{(n)}$. Notice from (4.2) that

$$\begin{aligned}\eta_k^{(n)} = \langle A_n^{-1} v_k^{(n)}, v_k^{(n)}\rangle &= \langle A^{-1} v_k^{(n)}, v_k^{(n)}\rangle + \langle (A_n^{-1} - A^{-1}) v_k^{(n)}, v_k^{(n)}\rangle] \\ &= \lambda_\ell^{-1} + \lambda_\ell^{-1} \langle B^{\frac{1}{2}}(I-P_n) v_k^{(n)}, C_n v_k^{(n)}\rangle.\end{aligned}$$

Thus

$$\left|\eta_k^{(n)} - \lambda_\ell^{-1}\right| \leq \lambda_\ell^{-1} \left\|B^{\frac{1}{2}}(I-P_n) v_k^{(n)}\right\| \cdot \left\|C_n v_k^{(n)}\right\|. \tag{4.4}$$

The second normed factor may be bounded as

$$\begin{aligned}\left\|C_n v_k^{(n)}\right\| &= \left\|B^{\frac{1}{2}}(I-P_n)A^{-1} v_k^{(n)} + B^{\frac{1}{2}}(I-P_n)(A_n^{-1} - A^{-1}) v_k^{(n)}\right\| \\ &\leq \lambda_\ell^{-1} \left(\left\|B^{\frac{1}{2}}(I-P_n) v_k^{(n)}\right\| + \left\|\left[B^{\frac{1}{2}} A_n^{-1} B^{\frac{1}{2}}\right] \cdot B^{\frac{1}{2}}(I-P_n) v_k^{(n)}\right\|\right) \\ &\leq \lambda_\ell^{-1} \left(1 + \left\|B^{\frac{1}{2}} A_0^{-\frac{1}{2}}\right\|^2\right) \left\|B^{\frac{1}{2}}(I-P_n) v_k^{(n)}\right\|,\end{aligned}$$

where we have used the fact

$$B^{\frac{1}{2}} A_n^{-1} B^{\frac{1}{2}} \leq B^{\frac{1}{2}} A_0^{-1} B^{\frac{1}{2}} \leq \|B^{\frac{1}{2}} A_0^{-\frac{1}{2}}\|^2.$$

Substituting into (4.4) yields

$$\left|\eta_k^{(n)} - \lambda_\ell^{-1}\right| \leq m_\ell \cdot b\left([I-P_n] v_k^{(n)}\right) \tag{4.5}$$

where $m_\ell = \lambda_\ell^{-2}(1 + \|B^{\frac{1}{2}} A_0^{-\frac{1}{2}}\|^2)$.

The residuals corresponding to $\eta_k^{(n)}$ may be estimated in a similar way:

$$\begin{aligned}
\theta_k^{(n)} &= \left\| A_n^{-1} v_k^{(n)} - \eta_k^{(n)} v_k^{(n)} \right\| \\
&= \left\| (A_n^{-1} - A^{-1}) v_k^{(n)} + (\lambda_\ell^{-1} - \eta_k^{(n)}) v_k^{(n)} \right\| \\
&\leq \lambda_\ell^{-1} \left\| C_n^* B^{\frac{1}{2}} (I - P_n) v_k^{(n)} \right\| + \left| \lambda_k^{-1} - \eta_k^{(n)} \right| \\
&\leq \lambda_\ell^{-1} \|C_n\| \left[b(I - P_n) v_k^{(n)} \right]^{\frac{1}{2}} + m_\ell b \left[(I - P_n) v_k^{(n)} \right].
\end{aligned}$$

In order to deduce behavior for large n, we must bound $\|C_n\|$. Note for any $u \in \mathcal{H}$:

$$\begin{aligned}
\|C_n u\|^2 &= \langle B(I - P_n) A_n^{-1} u, A_n^{-1} u \rangle \\
&\leq \langle B A_n^{-1} u, A_n^{-1} u \rangle = \left\| B^{\frac{1}{2}} A_0^{-\frac{1}{2}} A_0^{\frac{1}{2}} A_n^{-1} u \right\|^2 \\
&\leq \left\| B^{\frac{1}{2}} A_0^{-\frac{1}{2}} \right\|^2 \langle A_0 A_n^{-1} u, A_n^{-1} u \rangle \leq \left\| B^{\frac{1}{2}} A_0^{-\frac{1}{2}} \right\|^2 \langle A_n A_n^{-1} u, A_n^{-1} u \rangle \\
&= \left\| B^{\frac{1}{2}} A_0^{-\frac{1}{2}} \right\|^2 \langle A_n^{-1} u, u \rangle \leq \left\| B^{\frac{1}{2}} A_0^{-\frac{1}{2}} \right\|^2 \langle A_0^{-1} u, u \rangle \\
&\leq \left\| B^{\frac{1}{2}} A_0^{-\frac{1}{2}} \right\|^2 \cdot \left\| A_0^{-1} \right\| \cdot \|u\|^2.
\end{aligned}$$

Hence $\|C_n\|$ is bounded uniformly in n. Thus, we find

$$\theta_k^{(n)} \leq M_\ell \left[b([I - P_n] v_k^{(n)}) \right]^{\frac{1}{2}} + m_\ell b([I - P_n] v_k^{(n)}) \tag{4.6}$$

where $M_\ell = \lambda_\ell^{-1} \cdot \|B^{\frac{1}{2}} A_0^{-\frac{1}{2}}\| \cdot \|A_0^{-\frac{1}{2}}\|$.

Now, since $\bigcup_n \mathcal{P}_n + Ker(B)$ forms a core for a_0, conditions guaranteeing the convergence of eigenvalues of A_n^{-1} to those of A^{-1} are satisfied from Corollary 3.2. Hence, we may pick an interval (α, β) with $\alpha, \beta \notin \sigma(A^{-1})$ containing only those eigenvalues of A^{-1} coinciding with λ_ℓ^{-1} and then take n sufficiently large so that only r eigenvalues of A_n^{-1} remain in (α, β) as n increases further — namely those that will eventually converge to λ_ℓ^{-1}. For each such n, label those eigenvalues in $\sigma(A_n^{-1}) \cap (\alpha, \beta)$ in increasing order as $\tau_1^{(n)} \leq \tau_2^{(n)} \leq \cdots \leq \tau_r^{(n)}$. Let k be assigned $1 \leq k \leq r$ so that $\tau_k^{(n)} = [\lambda_\ell^{(n)}]^{-1}$.

The conditions of convergence guarantee in particular that $b\left([I - P_n] v_k^{(n)}\right) \longrightarrow 0$ and $\theta_i^{(n)} \longrightarrow 0$ as $n \to \infty$, so the hypotheses of Lemma 4.1 will hold for all n sufficiently big. Thus, assigning $T = A_n^{-1}$, we find

$$\eta_k^{(n)} - \sum_{i=k}^{r} \frac{[\theta_i^{(n)}]^2}{\beta - \eta_i^{(n)}} \leq \tau_k^{(n)} = \left[\lambda_\ell^{(n)} \right]^{-1} \leq \eta_k^{(n)} + \sum_{i=1}^{r} \frac{[\theta_i^{(n)}]^2}{\eta_i^{(n)} - \alpha}.$$

Asymptotically, then

$$\left| \left[\lambda_\ell^{(n)} \right]^{-1} - \eta_k^{(n)} \right| \leq (\text{constant}) \max_{i=1,\ldots,r} \left[\theta_i^{(n)} \right]^2 \tag{4.7}$$

as $n \to \infty$.

Since

$$\left|\left[\lambda_\ell^{(n)}\right]^{-1} - \lambda_\ell^{-1}\right| \leq \left|\left[\lambda_\ell^{(n)}\right]^{-1} - \eta_k^{(n)}\right| + \left|\eta_k^{(n)} - \lambda_\ell^{-1}\right|,$$

we may combine (4.5), (4.6) and (4.7) to obtain

$$\left|\left[\lambda_\ell^{(n)}\right]^{-1} - \lambda_\ell^{-1}\right| \leq (\text{constant}) \max_{i=1,\ldots,r} b\left([I - P_n]v_i^{(n)}\right).$$

the final result (4.3) follows from rearrangement and

$$b\left([I - P_n]v_k^{(n)}\right) \leq \max_{v \in \mathcal{U}_\ell} \frac{b([I - P_n]v)}{\|v\|^2} \leq \delta_b^2\left(\mathcal{U}_\ell, \mathcal{P}_n\right).$$

□

5. Computational Comparisons.

In this final section we illustrate convergence rates attainable with Aronszajn's method for an elementary Sturm–Liouville problem and show that, for the most part, predicted rates are roughly comparable with rates actually observed.

We return first to the example (2.3) in order to derive from Theorem 3.4 the rate for Aronszajn's method given at the end of Section 3. To this end define the orthogonal projection, E_n^0, taking $\mathcal{H}$ onto $span\,\{u_1^0, u_2^0, \ldots, u_n^0\} = \mathcal{P}_n$. Then we have for any v in $Dom\,(b) = \mathcal{H}$,

$$\begin{aligned} b((I - P_n)v) &= \|B^{\frac{1}{2}}(I - P_n)v\|^2 = \|B^{\frac{1}{2}}(I - P_n)(I - E_n^0)v\|^2 \\ &\leq \|B^{\frac{1}{2}}\| \cdot \|(I - E_n^0)v\|^2 = \|B^{\frac{1}{2}}\| \sum_{i=n+1}^{\infty} |\langle v, p_i\rangle|^2. \end{aligned}$$

And if $v \in Dom\,(A_0^\tau)$ for some $\tau > 0$,

$$\begin{aligned} \sum_{i=n+1}^{\infty} |\langle v, p_i\rangle|^2 &= \sum_{i=n+1}^{\infty} |\langle A_0^\tau v, A_0^{-\tau} p_i\rangle|^2 \\ &= \sum_{i=n+1}^{\infty} [\lambda_i^0]^{-2\tau} |\langle A_0^\tau v, p_i\rangle|^2 = (constant)\; n^{-4\tau}. \end{aligned}$$

Now in particular, for $v \in \mathcal{U}_\ell$ we find

$$A_0 v = -v'' = \lambda_\ell v - qv \in H^2(0,1) \cap H_0^1(0,1).$$

So

$$A_0^2 v = -\lambda_\ell v'' + q''v + q'v' + qv'' \in H^1(0,1)$$

and thus $\mathcal{U}_\ell \subset Dom\,(A_0^\tau)$ for all $\tau < \frac{9}{4}$ (cf. [15]). It follows that $b((I - P_n)v) = O(n^{-\tau})$ for all $\tau < 9$ so that from Theorem 3.4 $(\lambda_\ell - \lambda_\ell^{(n)})/\lambda_\ell = O(n^{-\tau})$ for all $\tau < 9$.

In order to illustrate the derived convergence rate in this setting we consider a parabolic cylinder equation obtained by setting the potential $q(x) = 10x^2$. Aronszajn's method may be applied directly since the corresponding Weinstein–Aronszajn matrix may be obtained in closed form:

$$\begin{aligned} W(\lambda) &= \left[\langle p_i + (A_0 - \lambda)^{-1} B p_i, B p_j \rangle\right] \\ &= \left[\langle p_i, B p_j \rangle\right] + \left[\langle s_i, B p_j \rangle\right] \end{aligned}$$

where $s_i(x)$ solves

$$-s_i'' - \lambda s_i = 10x^2 \sin(i\pi x)$$

$$with\ s_i(0) = s_i(1) = 0$$

and may be found using elementary techniques. Intermediate problem eigenvalues were then found using bisection to solve $det\,(w(\lambda)) = 0$. The results are summarized in Table 1.

Complementary upper bounds were computed with the Rayleigh–Ritz method with the same projecting vectors (i.e., a perturbed basis). A selection of results is listed in Table 2. Comparative closure rates are summarized graphically in Figures 1, 2, 3 and 4. In order to gauge rates of convergence the figures present the relative error in both Rayleigh–Ritz and Aronszajn approximations plotted against approximating subspace dimension, on a log–log scale. Calculations were all performed using MATLAB on a VAXstation 3800. Note that for each eigenvalue, Aronszajn's method is more rapidly convergent then the Rayleigh–Ritz method. For the second, third and fourth eigenvalues the observed asymptotic convergence rates appear to be consistent with what is predicted from the theory developed here, while for the first eigenvalue, a slower rate ($\sim n^{-8}$) is suggested by the data, thus providing equivocal evidence in this case for the derived asymptotic convergence rate.

REFERENCES

[1] Aronszajn, N. (1951) Approximation methods for eigenvalues of completely continuous symmetric operators. Proc. of the Symposium on Spectral Theory and Differential Problems, Stillwater, Oklahoma; 179–202.

[2] Babuška, I. and Osborn, J. (1990) *Eigenvalue Problems*, in *Handbook of Numerical Analysis*, (ed: P.G. Ciarlet and J. L. Lions). North-Holland Press, Amsterdam.

[3] Bazley, N. W. and Fox, D. W. (1961) Truncations in the method of intermediate problems for lower bounds to eigenvalues. J. Res. Nat. Bur. Standards Sect. B **65**, 105–111.

[4] Bazley, N. W. and Fox, D. W. (1966) Methods for lower bounds to frequencies of continuous elastic systems. J. Appl. Math. and Phys. (ZAMP) **17** (1), 1–37.

[5] Beattie, C. (1982) Some convergence results for intermediate problems with essential spectra. M. S. E. Research Center preprint series **65**, Johns Hopkins University Applied Physics Laboratory, Laurel.

[6] Beattie, C. and Greenlee, W. M. (1985) Convergence theorems for intermediate problems. Proc. Roy. Soc. Edinburgh **100A**, 107–122. Also corrigendum: Proc. Roy. Soc. Edinburgh **104A** (1986), 349–350.

[7] Beattie, C. and Greenlee, W. M. (1987) Convergence rates for intermediate problems. Manuscripta Mathematica **59**, 209–227.

[8] Birkhoff, G. and Fix, G. (1970) Accurate eigenvalue computations for elliptic problems. Proc. Numerical Solution of Field Problems in Continuum Physics. Symp. on Appl. Math., Vol 2. American Mathematical Society, 111–151.

[9] Brown, R. D. (1984) Variational approximation methods for eigenvalues: Convergence theorems. Banach Center Publications **13** Warsaw, 543–558.

[10] Brown, R. D. (1988) Convergence criterion for Aronszajn's method and for the Bazley–Fox method, Proc. Roy. Soc. Edinburgh **108A**, 91–108.

[11] Chatelin, F. (1983) *Spectral Approximation of Linear Operators*, Academic Press, New York.

[12] Fix, G. (1968) Orders of convergence of the Rayleigh–Ritz and Weinstein–Bazley methods. Proc. Nat. Acad. Sci. U.S.A. **61**, 1219–1223.

[13] Fox, D. W. and Rheinboldt, W. C. (1966) Computational methods for determining lower bounds for eigenvalues of operators in Hilbert space. SIAM Rev. **8**, 427–462.

[14] Goerisch, F. and Haunhorst, H. (1985) Eigenwertschranken für Eigenwertaufgaben mit partiellen Differentialgleichungen, J. Appl. Math. and Mech. (ZAMM) **3**, 129–135.

[15] Greenlee, W. M. (1968) Rate of convergence in singular perturbations. Ann. Inst. Fourier (Grenoble) **18**, 135–191.

[16] Greenlee, W. M. (1969) Singular perturbation of eigenvalues. Arch. Rational Mech. Anal. **34**, 143–164.

[17] Greenlee, W. M. (1978) Perturbation of bound states of the radial equation by repulsive singular potentials — first order asymptotics. J. Funct. Anal. **27**, 185–202.

[18] Greenlee, W. M. (1983) A convergent variational method of eigenvalue approximation. Arch. Rational Mech. Anal. **81**, 279–287.

[19] Kato, T. (1955) Quadratic forms in Hilbert spaces and asymptotic perturbation series. Tech. Rep. 7, University of California, Berkeley.

[20] Kato, T. (1976) *Perturbation Theory for Linear Operators*, 2nd edn. (Berlin: Springer).

[21] Poznyak, L. T. (1968) Estimation of the rate of convergence of a variant of the method of intermediate problems. Ž. Vyčisl. Mat. i Mat. Fiz. **8**, 1117—1126; USSR Computational Math. and Math. Phys. **8** (1968), 246–260.

[22] Weidmann, J. (1980) Monotone continuity of the spectral resolution and the eigenvalues. Proc. Roy. Soc. Edinburgh Sect. A **85**, 131–136.

[23] Weinberger, H. (1974) *Variational Methods for Eigenvalue Approximation.* Philadelphia: SIAM.

[24] Weinstein, A. and Stenger, W. (1972) *Methods of Intermediate Problems for Eigenvalues.* New York: Academic Press.

Christopher Beattie
Department of Mathematics
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061
U.S.A.

W. M. Greenlee
Department of Mathematics
University of Arizona
Tucson, AZ 85721
U.S.A.

Table 1. Parabolic Cylinder Equation: Lower Bounds

Index	Base Problem	Aronszajn's Method $n = 5$	$n = 10$	$n = 15$
1	9.8696	12.587555357_{950}	12.58755540_{1932}	12.58755540_{2112}
2	39.478	42.71338839_{2138}	42.7133872_{9925}	42.71338873_{0920}
3	88.826	92.12318195_{3966}	92.12318380_{4803}	92.12318380_{7709}
4	157.91	161.2277543_{9486}	161.227769_{70114}	161.227769_{70852}

Script–size digits represent a contribution smaller than *a posteriori* estimates of accuracy.

Table 2. Parabolic Cylinder Equation: Upper Bounds

	Index	Rayleigh–Ritz Method $n = 5$	$n = 10$	$n = 30$
	1	12.58755984_{9706}	12.58755545_{1862}	12.58755540_{2147}
	2	42.71342130_{1772}	42.71338900_{3346}	42.71338873_{1096}
	3	92.12334750_{2829}	92.12318458_{1200}	92.12318380_{8163}
	4	161.12289336_{7650}	161.22777154_{533}	161.227769_{70852}

Script–size digits represent a contribution smaller than *a posteriori* estimates of accuracy.

Figure 1. Parabolic Cylinder Equation: First Eigenvalue

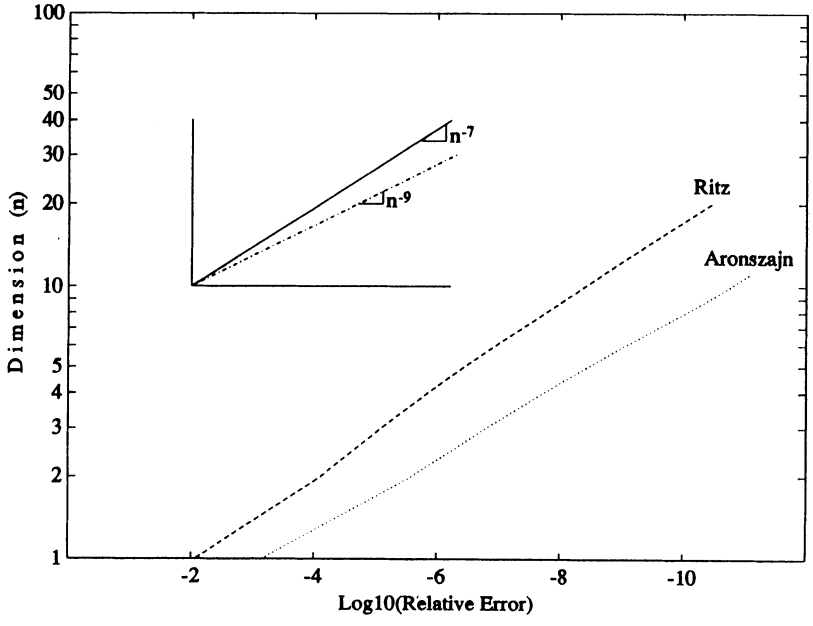


Figure 2. Parabolic Cylinder Equation: Second Eigenvalue

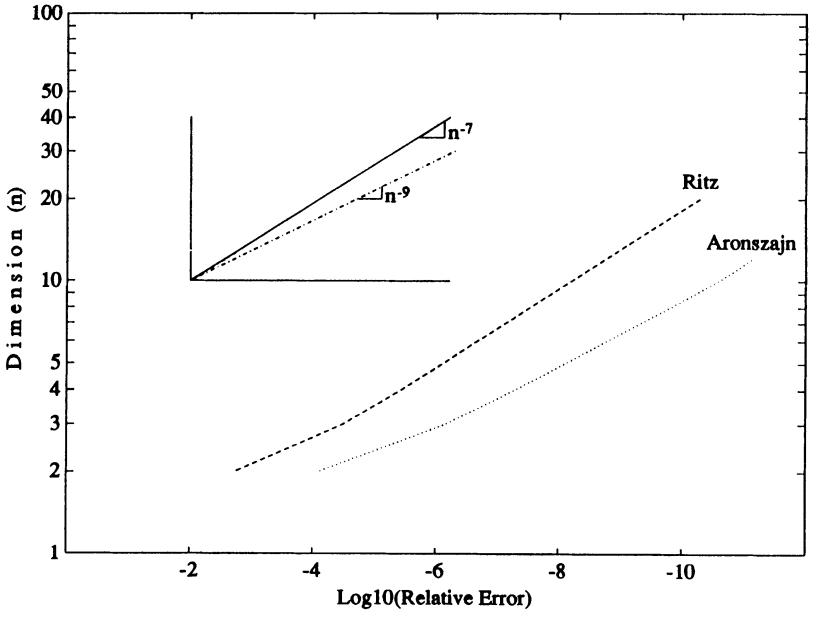


Figure 3. Parabolic Cylinder Equation: Third Eigenvalue

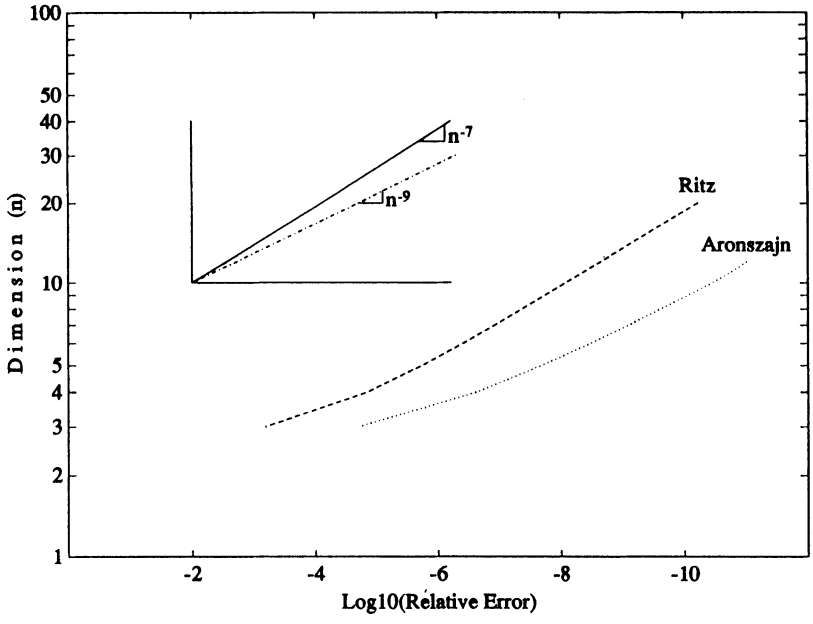
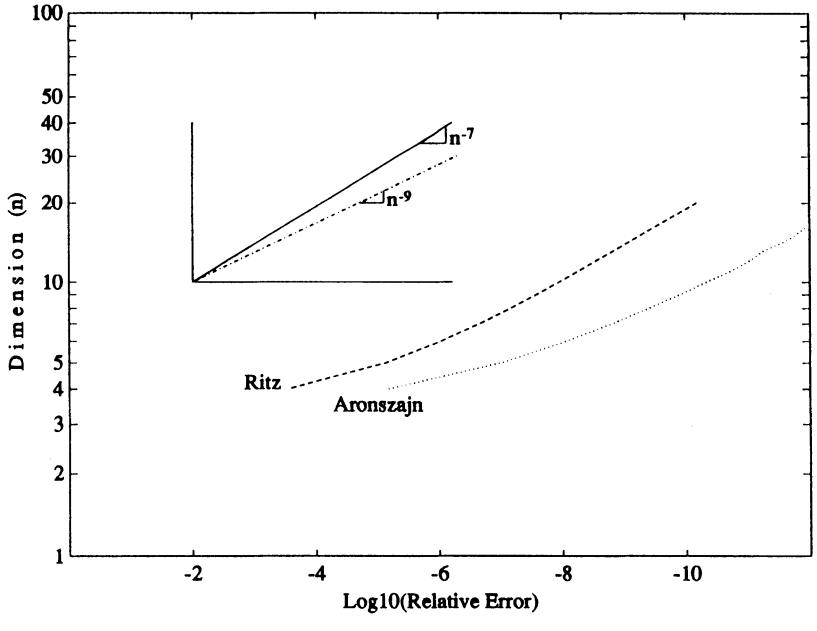


Figure 4. Parabolic Cylinder Equation: Fourth Eigenvalue



International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

AN EIGENVALUE PROBLEM OF THE THEORY OF ARMA MODELS AND MULTISTEP ITERATION PROCEDURES

Leonhard Bittner

Arma–models are recursive equations of the kind

$$x_k = A_0 + \sum_{i=1}^{p} A_i x_{k-i} + \sum_{i=0}^{q} B_i z_{k-i} + \ldots \tag{1}$$

and are used to describe objects of unknown mathematical structure approximately. The main task is to determine appropriate (n,n)–matrices A_i, B_i. In order to guarantee that the n–vectors x_k form a stationary time series, it is required that the polynomial matrix

$$\lambda^p I - \sum_{i=1}^{p} \lambda^{p-i} A_i$$

has only singular values λ with absolute values $|\lambda| < 1$. Sometimes it is also required that $\sum \lambda^{q-i} B_i$ has only singular values within the unit circle. After having computed estimations A_i, B_i by means of concrete measured values x_k (k = 0,...,N) and some numerical method, f.e. the YULE–WALKER procedure, it is desirable to check – with as little effort as possible – whether the singular value condition mentioned above is satisfied and whether it is advisable to accept or to reject the computed results.

A quite similar question concerning the singular values of a polynomial operator arises in the theory of (fixed point) equations in constructing multistep procedures, as will be shown below.

1. Reduction to a HURWITZ problem

With the aid of the (real) (n,n)–matrices A_i determine

$$\det(\lambda^p I - \sum_{i=1}^{p} \lambda^{p-i} A_i) = \lambda^l (b_l + b_{l+1}\lambda + \ldots + b_{l+m}\lambda^m) = \lambda^l P(\lambda), \tag{2}$$

where b_l and $b_{l+m} \neq 0$, $l + m = n \cdot p$.

Check whether $P(-1) = 0$. In the affirmative case the singular values of the polynomial matrix are not contained in the unit circle completely. In the negative case perform the substitutions

$$\lambda = \frac{1+z}{1-z}, \quad z = \frac{\lambda - 1}{\lambda + 1}, \quad Q(z) = P\left[\frac{1+z}{1-z}\right](1-z)^m. \tag{3}$$

The substitutions transform a real polynomial P again into a real polynomial Q, the interior (exterior) of the λ–unit–circle is mapped onto the left (right) z–half–plane one–to–one, whereby $\lambda = -1$, a regular value, is lost, because it is mapped into $z = \infty$, and $z = 1$ is gained, since it is the image of $\lambda = \infty$, but is no root of Q. Hence all singular values of the polynomial matrix lie within the unit circle if and only if all roots of Q lie within the left half plane. Now the criteria of HURWITZ, ROUTH etc. allow to decide the question. However, in order to execute the necessary instructions one has to do a lot of additional calculations.

2. Companion matrix methods

With the aid of the (n,n)–matrices A_i and n–vectors x_i, y_i, e_i (i = 1,...,p) define the following hypermatrix A and hypervectors x, y, e

$$A = \begin{bmatrix} 0 & I & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & I \\ A_p & A_{p-1} & \cdots & A_1 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ x_p \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ y_p \end{bmatrix}, \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ e_p \end{bmatrix}. \tag{4}$$

Let (x_i, y_i) be the original scalar product and [x,y] be

$$[x,y] = \sum_{i=1}^{p} (x_i, y_i). \tag{5}$$

Since $Ax = \lambda x$, if $(\lambda^p I - \sum \lambda^{p-i} A_i)\, x_1 = 0$, $x_i = \lambda^{i-1} x_1$, the set of eigenvalues and singular values coincide. If there is a single singular value λ_0 of the maximal absolute value ρ, a so called dominant singular value, then ρ, the spectral radius of A, can be defined by some "power" method. F.e. let x, e be appropriate (hyper) vectors, then – as is well known – the sequences

$$\left| \frac{[e,A^{k+1}x]}{[e,A^{k}\ x]} \right| , \ | \ [e,A^{k}\ x]^{\frac{1}{k}} \ | \ , \left| \frac{[AA^{k}\ x,\ A^{k}\ x]}{[A^{k}\ x,\ A^{k}\ x]} \right| \tag{6}$$

converge to ρ (individually).

The meaning of the loose saying that x, e be appropriate vectors is most briefly explained by means of the resolvent matrix which has a LAURENT expansion with the pole λ_0

$$(\lambda I - A)^{-1} = (\lambda - \lambda_0)^{-q}R_{-q} + \ldots + (\lambda - \lambda_0)^{-1}R_{-1} + \sum_{j=0}^{\infty} (\lambda - \lambda_0)^{j}R_j \tag{7}$$

in the neighbourhood of the dominant eigenvalue λ_0. Let L_ϵ be the periphery of a circle with center λ_0 and sufficient small radius ϵ, let L be the periphery of a circle with center 0 and radius r, containing the spectrum of A apart from λ_0, then L and L_ϵ do not intersect, $r < |\lambda_0|$ and for $k \geq q$

$$\begin{aligned} A^{k}x &= \frac{1}{2\pi i}\int_{L} \lambda^{k}(\lambda I - A)^{-1}x\, d\lambda + \frac{1}{2\pi i}\lim_{\epsilon \to 0}\int_{L_\epsilon} \lambda^{k}(\lambda I - A)^{-1}x\, d\lambda \\ &= 0(r^{k}) + \binom{k}{q-1}\lambda_0^{k-q+1}R_{-q}x + \ldots + \binom{k}{0}\lambda_0^{k}R_{-1}x. \end{aligned} \tag{8}$$

Appropriate vectors e, x for the first two sequences are such vectors, for which at least one of the numbers

$$\left[e, R_{-q}x\right] , \ldots , \left[e, R_{-1}x\right] \tag{9}$$

is different from 0. An appropriate vector x for the third sequence is a vector, for which at least one of the vectors

$$R_{-q}x, \ldots , R_{-1}x \tag{10}$$

is different from 0.

Now let us translate the companion matrix formulation into original expressions. Starting with the "components" $x_1,\ldots,x_p$ of x, generate recursively

$$x_k = \sum_{i=1}^{p} A_i x_{k-i} \qquad (k = p+1, p+2, \ldots) \tag{11}$$

and the corresponding hypervectors

$$x^k = \begin{bmatrix} x_{k-p+1} \\ \cdot \\ \cdot \\ \cdot \\ x_k \end{bmatrix}. \tag{12}$$

Obviously

$$\begin{aligned} &x^p = x,\ Ax^k = x^{k+1},\ A^k x = x^{k+p}, \\ &\left[e, A^k x\right] = (e_1, x_{k+1}) + (e_2, x_{k+2}) + \ldots + (e_p, x_{k+p}), \\ &\left[A^{k+1}x, A^k x\right] = (x_{k+2}, x_{k+1}) + (x_{k+3}, x_{k+2}) + \ldots + (x_{k+1+p}, x_{k+p}). \end{aligned} \tag{13}$$

Hence we obtain

Theorem 1

If the polynomial matrix corresponding to $A_1,\ldots,A_p$ has a dominant singular value, if $x_1,\ldots,x_p$, $e_1,\ldots,e_p$ are appropriate vectors and if the x_k are defined by (11), then the sequences

$$\left| \frac{\sum_{j=1}^{p} (e_j, x_{k+1+j})}{\sum_{j=1}^{p} (e_j, x_{k+j})} \right|, \left| \sum_{j=1}^{p} (e_j, x_{k+j}) \right|^{\frac{1}{k}}, \frac{\left| \sum_{j=1}^{p} (x_{k+1+j}, x_{k+j}) \right|}{\sum_{j=1}^{p} |x_{k+j}|^2}$$

converge (individually) to some number ρ. If $\rho < 1$, then all singular values lie within the unit circle.

With respect to Arma–model applications we may state that the assumption of a dominant singular value is a reasonable one and the generation of the x_k is almost a byproduct of necessary calculations.

3. Construction of multistep iteration procedures

This section deals with iteration processes of the kind

$$x_k = g(x_{k-p},\dots,x_{k-1}) \qquad (k = p+1,\dots) \tag{14}$$

for equations $h(x_0) = 0$ in a BANACH space X_0. At first let us suppose that such an iteration operator g is available though the main problem is to construct g for a given h. Each relevant solution x_0 of the equation should be a fixed point of g, i.e.

$$h(x_0) = 0 \implies x_0 = g(x_0,\dots,x_0).$$

Let X be the p–fold Cartesian product $X_0 \times \dots \times X_0$ with the elements x and norm $\|x\|_m$

$$x = (x_1,\dots,x_p), \quad \|x\|_m = \max \|x_i\|, \quad x_i \in X_0. \tag{15}$$

Assume that g maps X into X_0 and has continuous partial FRECHET derivatives $g'_i(x)$ with respect to all components x_i of x. Pose

$$x^0 = (x_0,\dots,x_0), \quad A_i = g'_{p+1-i}(x^0) \tag{16}$$

and by A denote the hypermatrix operator in X defined as follows

$$Ax = (x_2,\dots,x_p, \sum_{i=1}^{p} A_{p+1-i}x_i). \tag{17}$$

Let ρ be the spectral radius of A which will also be called spectral radius of the polynomial operator

$$\lambda^p I - \sum_{i=1}^{p} \lambda^{p-i} A_i \tag{18}$$

since the spectrum of A coincides with the set of all singular values of it.

Theorem 2

Suppose that x_0 is a fixed point of g, that x^0 and A_i are defined by (16) and that the spectral radius of the polynomial operator (18) is $\rho < 1$. If q is any number statisfying $\rho < q < 1$, then there is a positive r such that the process (14) converges towards x_0 for any initial system x_i with $\|x_i - x_0\| < r$ (i = 1,...,p) and

$$\|x_i - x_0\| \leq cq^k \quad (k \geq p)$$

with some positive number c.

Proof. Let ϵ be positive so that $\rho < \rho + \epsilon < q$. As is well known, there is a norm $\|\cdot\|_\epsilon$ equivalent to the original $\|\cdot\|_m$ in X so that the corresponding sup–norm of an operator A differs from the spectral radius ρ only by a value smaller than ϵ, which means

$$\|A\|_\epsilon < \rho + \epsilon.$$

From the relation

$$a\|x\|_m \leq \|x\|_\epsilon \leq b\|x\|_m, \; x \in X,$$

follows the corresponding sup–norm relation

$$\frac{a}{b}\|A\|_m \leq \|A\|_\epsilon \leq \frac{b}{a}\|A\|_m.$$

From the continuity of the derivatives it follows the existence of a positive r_ϵ so that

$$\sum_{i=1}^{p} \|g_i'(x) - g_i'(x^0)\| < \frac{a}{b}(q-\rho-\epsilon)$$

for $\|x - x^0\|_\epsilon < r_\epsilon$.

Pose with the iterates generated by (14)

$$x^k = (x_{k-p+1},\ldots,x_k),$$

$$T_k y = (y_2,\ldots,y_p, \sum_{i=1}^{p} \int_0^1 g_i'(x^0 + t(x^k - x^0))\, y_i\, dt$$

for all

$$y = (y_1,\ldots y_p) \in X,$$
$$T_0 = A.$$

Then we obtain

$$\begin{aligned} x^k - x^0 &= (x_{k-p+1} - x_0,\ldots,x_{k-1} - x_0, g(x^{k-1}) - g(x^0)) = \\ &= T_{k-1}(x^{k-1} - x^0) = A(x^{k-1} - x^0) + (T_{k-1} - T_0)(x^{k-1} - x^0), \\ \|x^k - x^0\|_\epsilon &\le (\rho + \epsilon)\, \|x^{k-1} - x^0\|_\epsilon + \|T_{k-1} - T_0\|_\epsilon \cdot \|x^{k-1} - x^0\|_\epsilon, \\ \|T_{k-1} - T_0\|_\epsilon &\le \frac{b}{a}\, \|T_{k-1} - T_0\|_m \le \\ &\le \frac{b}{a} \sum_{i=1}^{p} \max_{0\le t\le 1} \|g_i'(x^0 + t(x^{k-1} - x^0)) - g_i'(x^0)\| < q-\rho-\epsilon, \end{aligned}$$

provided $\|x^{k-1} - x^0\|_\epsilon < r_\epsilon$. If this holds, then moreover

$$\|x^k - x^0\|_\epsilon \le q\, \|x^{k-1} - x^0\|_\epsilon \le \|x^{k-1} - x^0\|_\epsilon < r_\epsilon.$$

Assume $\|x^p - x^0\|_\epsilon < r_\epsilon$, which is assured for $\|x^p - x^0\|_m < r = b^{-1} r_\epsilon$. This implies $\|x^k - x^0\|_\epsilon < r_\epsilon$ for all $k \ge p$ and consequently

$$\|x^k - x^0\|_m = a^{-1}\, \|x^k - x^0\|_\epsilon \le (a^{-1} q^{-p} \|x^p - x^0\|_\epsilon)\, q^k.$$

Now we are able to make some proposals concerning the choice of a multistep operator g.

As a rule the original equation $h(x_0) = 0$ ought to be modified by introducing additional items and by writing $x_1,\ldots,x_p$ at different spots instead of x_0. Let us consider some

examples for equations of the type $h(x_0) = x_0 - f(x_0) = 0$:

$$1)\ g(x_1,\ldots,x_p) = x_p + \sum_{i=1}^{p-1} B_i(f(x_{p-i}) - x_{p-i+1})$$

$$2)\ g(x_1,\ldots,x_p) = x_p + \sum_{i=1}^{p-1} B_i(f^{(i)}(x_{p-i}) - x_{p-i+1}),$$

where $f^{(i)}$ denotes the i−th iterate of f,

$$3)\ g(x_1,\ldots,x_p) = x_p + \sum_{i=1}^{p} B_i H_i,$$

where

$$H_1 = h(x_1),\ H_2 = h(x_2 + B_{21}H_1),$$
$$H_3 = h(x_3 + B_{31}H_1 + B_{32}H_2),\ldots \text{ (RUNGE−KUTTA−scheme).}$$

The $B_i,\ldots$ have to map 0 into 0. They are yet undefined operators or constants. Evidently each solution of the original equation turns out to be fixed point of g, the converse should be valid too. Theorem 2 now suggests to choose the free "parameters" $B_i,\ldots$ in such a way that the spectral radius of the polynomial operator, generated by the partial derivatives of g at x^0 or at some first guess for x^0, will become as small as possible, at least smaller than 1.

References

1. Andel, J. (1984) Statistische Analyse von Zeitreihen (Akademie−Verlag, Berlin)
2. Faddejew, D. K. (1973) Numerische Methoden der linearen Algebra, 3. Auflage (Deutscher Verl. Wiss., Berlin)

L. Bittner, Sektion Mathematik der Universität, L.−Jahn−Str. 15a, Greifswald, DDR−O−2200.

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

OSCILLATIONS AND STABILITY OF A WHEEL ROLLING ON A FLEXIBLE RAIL

Dedicated to Professor Dr. Drs h.c. Lothar Collatz
on the occasion of his eightieth birthday

Eberhard Brommundt, Technische Universität Braunschweig, Germany

1. Introduction and Summary

Frequently rails of railway systems undergo corrugation due to an oscillatory contact between wheel and rail, cf. KNOTHE and VALDIVIA. This paper studies the stability of the stationary motion of a rigid wheel rolling on a straight (uncorrugated) flexible rail. Starting from geometrically nonlinear equations of motion, a set of variational equations for oscillations about a simplified stationary solution is established. Its spectrum consists of a continous and a discrete part. The discrete eigenvalues are calculated for parameters close to realistic ones. Several numerical problems occur.
In certain parameter regions the eigenvalues get positive real parts which lead to instability of the stationary motion. This supports the author's hypothesis that self-excitation might initiate corrugation.

2. Equations of Motion

Figure 1 gives a general view of the model: The body is horizontally driven with constant velocity $v > 0$. The center of the

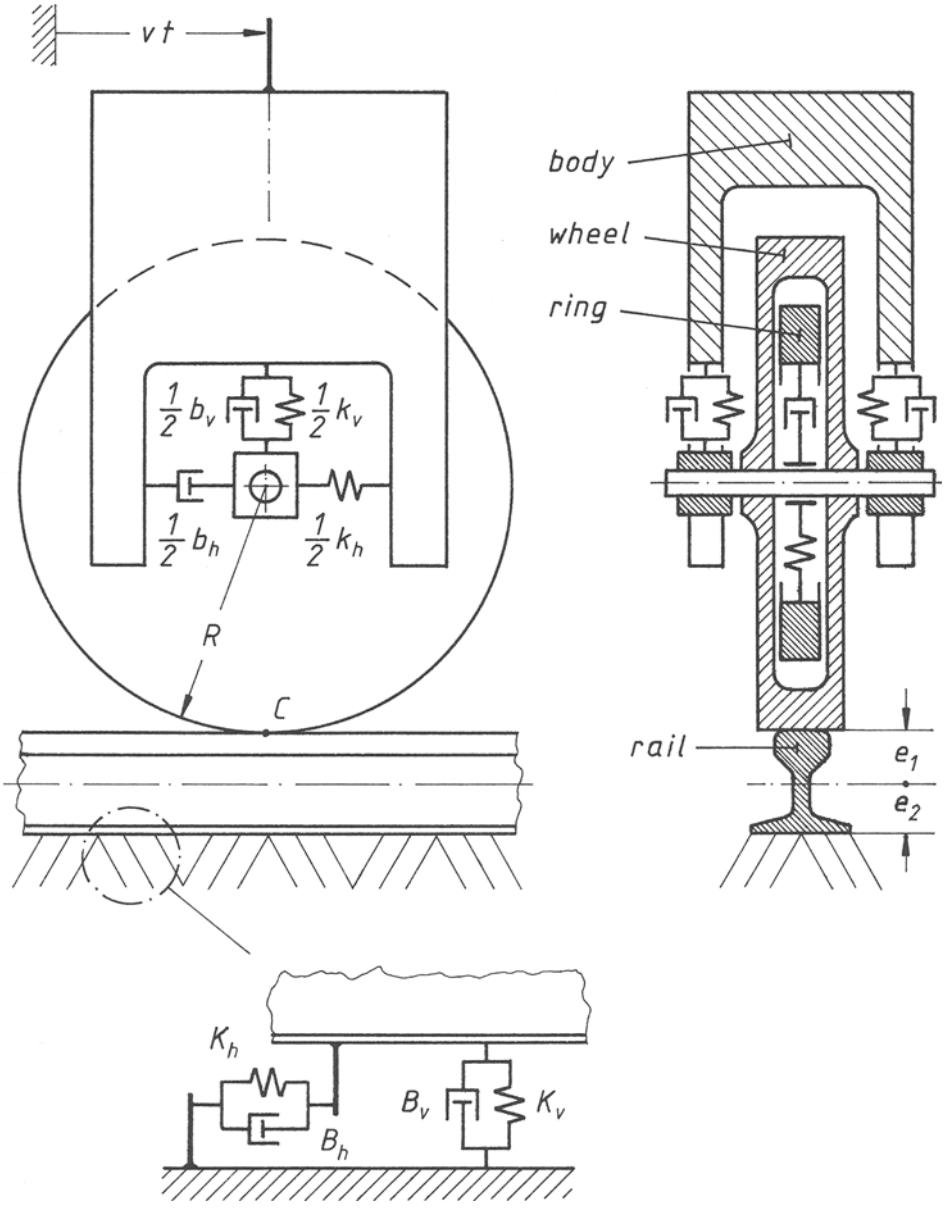


Fig. 1. The System

rigid wheel is flexibly connected to the body. The wheel contains as a constituent part a flexibly attached ring (which serves to mimic translational inertial effects due to deformations) and rolls on top of the rail. The rail is flexible and is horizontally as well as vertically bedded on damped Winkler foundations. Motions parallel to the plane of Figure 2a are studied, x and z are the reference coordinates for the longitudinal direction of the rail and the vertical direction, respectively; t is the time; $(.)' := \partial(.)/\partial x$, $(.)^{\cdot} = \partial(.)/\partial t$. All horizontal deflections with respect to the undeformed state are denoted by u(.), all vertical ones by w(.). They are distinguished by the subscripts: b - body, wh - wheel, r - ring, no subscript - rail. The masses are denoted by m (for the rail: distributed mass μ), the weights by W, the stiffness and damping constants are denoted by k, b and subscripted by h (horizontal), v (vertical), or r (ring); the corresponding constants for the Winkler foundation are K and B. Further variables and parameters are obvious from the figures or explained in the text.

2.1 Equations for the Rail

The rail is modeled as a slender Euler-Bernoulli beam, including a center strain $\varepsilon(x,t)$. For the angle $\varphi(x,t)$, cf. Fig.2b, holds

$$\sin\varphi = w'/(1 + \varepsilon) \ , \quad \cos\varphi = (1 + u')/(1 + \varepsilon). \tag{2.1}$$

The strain distribution $\varepsilon^*(x,z,t)$ in the rail satisfies

$$\varepsilon^* = \varepsilon - z\,\varphi', \tag{2.2}$$

which leads via the linear visco-elastic law

$$\sigma = E\varepsilon^* + B\dot{\varepsilon}^* + \sigma_o \tag{2.3}$$

to the normal force

$$N = \int_A \sigma \, dA = N_o + EA\,\varepsilon + BA\,\dot{\varepsilon} \tag{2.4}$$

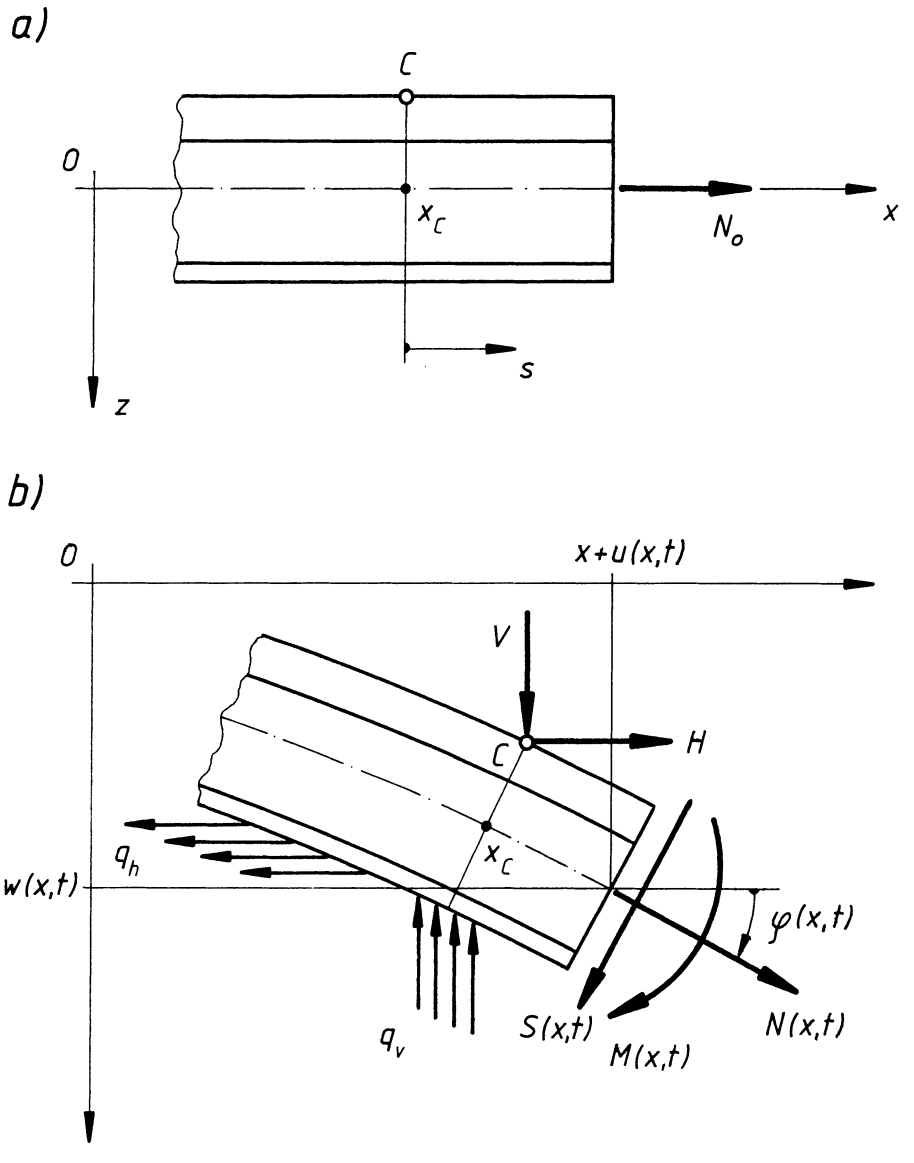


Fig. 2. a) Reference configuration of the rail
b) Deflected rail with loads

and the bending moment

$$M = -\int_A z\,\sigma\,dA = EI\,\varphi' + BI\,\dot{\varphi}' \quad , \tag{2.5}$$

cf. Fig.2b. Here denote: N_o - preload; EA, BA - tensional stiffness and damping coefficient; EI, BI - bending stiffness and damping coefficient. The deformation due to the shear force S, cf. Fig.2b, is neglegted.
The transmitted loads N(x,t), S(x,t), M(x,t), together with the distributed forces from the roadbed onto the rail,

$$q_h = K_h\,(u - e_2\,\sin\varphi) + B_h\,(u - e_2\,\sin\varphi)^{\cdot} \quad , \tag{2.6}$$

$$q_v = K_v\,[w - e_2\,(1 - \cos\varphi)] + B_v\,[w - e_2\,(1 - \cos\varphi)]^{\cdot} , \tag{2.7}$$

and the contact forces H(t), V(t) from the wheel onto the rail at C, the point of contact (reference coordinate $x_c(t)$, cf. Fig.2b), must satisfy the equilibrium conditions

$$\begin{aligned} N' - S\,\varphi' - (q_h + \mu\,\ddot{u})\,\cos\varphi - (q_v + \mu\,\ddot{w})\,\sin\varphi \\ + (H\,\cos\varphi + V\,\sin\varphi)\,\delta(x - x_c) = 0 \quad , \end{aligned} \tag{2.8}$$

$$\begin{aligned} N\,\varphi' + S' + (q_h + \mu\,\ddot{u})\,\sin\varphi - (q_v + \mu\,\ddot{w})\,\cos\varphi \\ + (-H\,\sin\varphi + V\,\cos\varphi)\,\delta(x - x_c) = 0 \quad , \end{aligned} \tag{2.9}$$

$$\begin{aligned} M' + S\,(1 + \varepsilon) + q_h\,e_2\,\cos\varphi + q_v\,e_2\,\sin\varphi \\ + (H\,\cos\varphi + V\,\sin\varphi)\,e_1\,\delta(x - x_c) = 0 \quad ; \end{aligned} \tag{2.10}$$

$\delta(x-x_c)$ is Dirac's delta function.

2.2 Equations for Wheel, Body and Ring

There hold the equilibriumn conditions for the body:

$$m_b\,\ddot{w}_b + b_v\,(\dot{w}_b - \dot{w}_{wh}) + k_v\,(w_b - w_{wh}) = W_b \quad ; \tag{2.11}$$

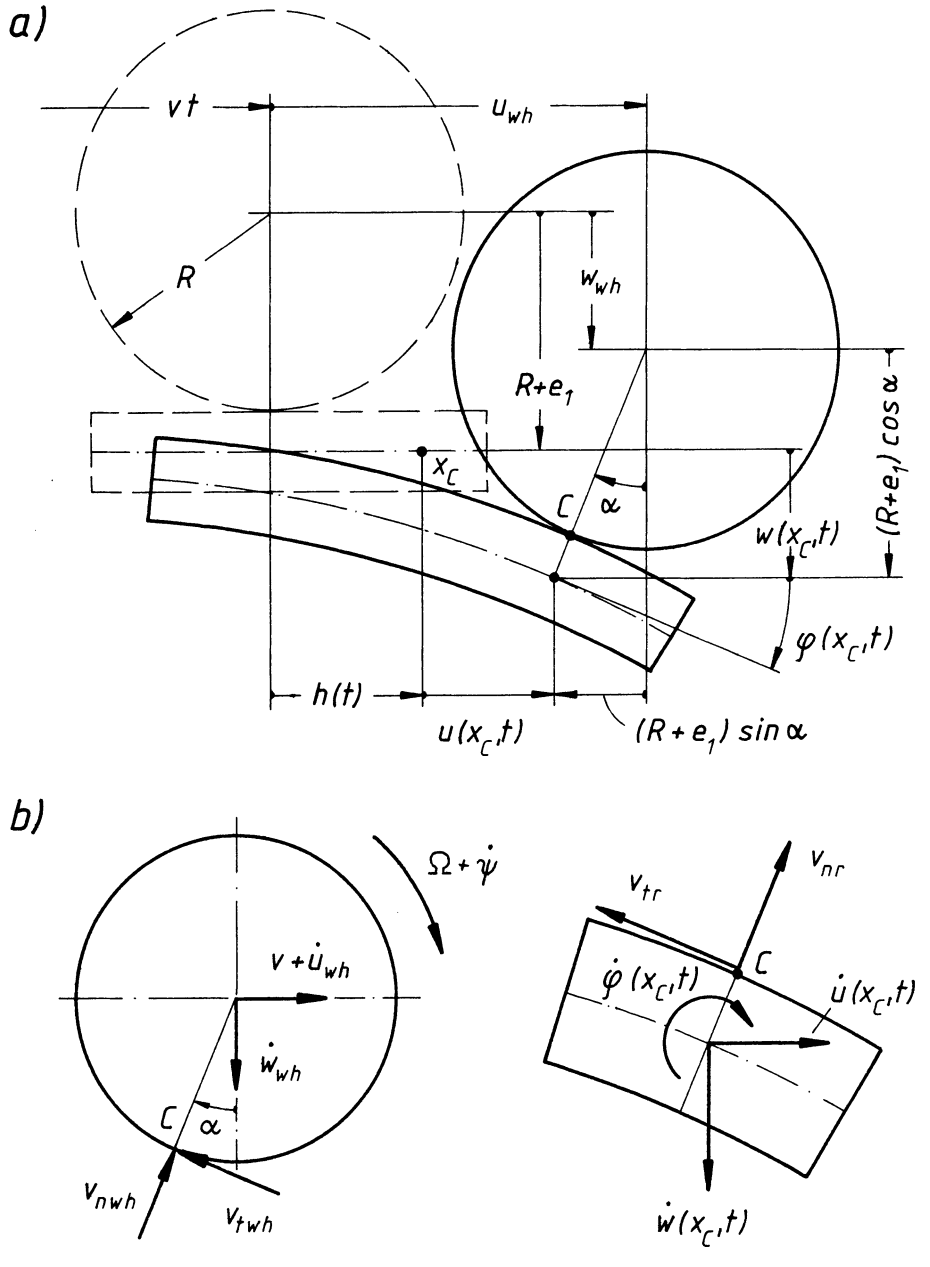


Fig. 3. Contact between wheel and rail
a) Geometry b) Kinematics

the wheel:

$$m_{wh}\,\ddot{u}_{wh} + b_h\,\dot{u}_{wh} + b_{rh}(\dot{u}_{wh} - \dot{u}_r) + k_h\,u_{wh} + k_{rh}(u_{wh} - u_r) + H = 0\ , \tag{2.12}$$

$$m_{wh}\,\ddot{w}_{wh} + b_v(\dot{w}_{wh} - \dot{w}_b) + b_{rv}(\dot{w}_{wh} - \dot{w}_r) + k_v(w_{wh} - w_b) + k_{rv}(w_{wh} - w_r) + V = W_{wh}\ , \tag{2.13}$$

$$\Theta_{wh}\,\ddot{\psi}_{wh} - H\,R\,\cos\alpha - V\,R\,\sin\alpha = 0\ , \tag{2.14}$$

Θ_{wh} - moment of inertia, $\Omega t + \psi_{wh}$ - angle of rotation, Ω - mean angular velocity, R - radius, $\alpha(t)$ see (2.17);

the ring:

$$m_r\,\ddot{u}_r + b_{rh}(\dot{u}_r - \dot{u}_{wh}) + k_{rh}(u_r - u_{wh}) = 0\ , \tag{2.15}$$

$$m_r\,\ddot{w}_{wh} + b_{rv}(\dot{w}_r - \dot{w}_{wh}) + k_{rv}(w_r - w_{wh}) = W_r\ . \tag{2.16}$$

2.3 Geometry and Kinematics at the Point of Contact

From Fig.3a follow the conditions of geometric consistency of the coordinates:

$$\alpha(t) = \varphi(x_c(t),t) =: \varphi_c\ , \tag{2.17}$$

$$v\cdot t + u_{wh}(t) - (R + e_1)\,\sin\alpha(t) = x_c(t) + u(x_c(t),t)\ , \tag{2.18}$$

$$w_{wh}(t) - (1 - \cos\alpha(t))\,(R + e_1) = w(x_c(t),t)\ . \tag{2.19}$$

Furthermore, Fig.3b shows the tangential velocities at C, the point of contact between wheel and rail:

$$v_{twh} = -(v + \dot{u}_{wh})\,\cos\alpha - \dot{w}_{wh}\,\sin\alpha + R\,(\Omega + \dot{\psi}_{wh}) \tag{2.20}$$

and

$$v_{tr} = -\dot{u}(x_c,t)\,\cos\alpha - \dot{w}(x_c,t)\,\sin\alpha - e_1\,\dot{\varphi}(x_c,t)\ . \tag{2.21}$$

We assume rolling contact:

$$v_{twh} = v_{tr} \,. \tag{2.22}$$

Due to (2.18), (2.19) the normal velocities satisfy $v_{nwh} = v_{nr}$.

3. Transition to a Moving Coordinate System

3.1 Transformations

To remove the time dependent coordinate of the point of contact, $x_c(t)$, from the distribution $\delta(x-x_c)$ in (2.8) - (2.10) we introduce the moving coordinate s whose origin, s = 0, is affixed to x_c (in the reference configuration, cf. Fig. 2a). This leads to

$$s := x - x_c = x - (vt + h) \tag{3.1}$$

where h = h(t) denotes the horizontal deviation of the point of contact with respect to the body measured in the reference configuration (cf. Fig.3a). This transforms $\delta(x-x_c)$ into $\delta(s)$, and $s \gtrless 0$ denote the regions in front of and behind C, the point of contact, respectively.
All variables depending on (x,t), e.g. u(x,t), are transformed into functions of (s,t), e.g. $\bar{u}(s,t)$. From the identity

$$u(x,t) \equiv \bar{u}(x - (vt + h),t) \tag{3.2}$$

follow with $(.)' = \partial(.)/\partial s$

$$u' = \bar{u}', \quad u'' = \bar{u}'' \,, \text{ etc.} \tag{3.3}$$

$$\dot{u} = -(v + \dot{h})\,\bar{u}' + \dot{\bar{u}} \,,$$

$$\dot{u}' = -(v + \dot{h})\,\bar{u}'' + \dot{\bar{u}}', \tag{3.4}$$

$$\ddot{u} = (v + \dot{h})^2\,\bar{u}'' - 2(v + \dot{h})\,\dot{\bar{u}}' + \ddot{h}\,\bar{u}' + \ddot{\bar{u}} \,;$$

etc.

After the transformation the overbars are dropped, $\bar{u}(s,t) \to u(s,t)$, $\bar{u}'(s,t) \to u'(s,t)$, etc.

3.2 Conditions at s = 0 and s → ±∞

Let [u] denote a "jump" of the function u(s,t) at s = 0:

$$[u] := \lim_{\nu \downarrow 0} \{u(\nu,t) - u(-\nu,t)\} \ . \tag{3.5}$$

From the transformed equations (2.1) - (2.10) the following continuity and jump conditions for u, w, ε, φ can be deduced ($\varphi_c := \varphi(0,t)$ etc.):

$$\begin{aligned}
&[u] = 0,\ [w] = 0,\ [\varepsilon] = 0,\ [\varphi] = 0,\ [\varphi'] = 0 \ , \\
&- BA\ (v + \dot{h})\ [\varepsilon'] + H \cos\varphi_c + V \sin\varphi_c = 0 \ , \\
&- BI\ (v + \dot{h})\ [\varphi''] + e_1\ (H \cos\varphi_c + V \sin\varphi_c) = 0 \ , \\
&EI\ [\varphi''] + BI \{ - (v + \dot{h})\ [\varphi'''] + [\dot{\varphi}''] \} \\
&\qquad + (1 + \varepsilon_c)(H \sin\varphi_c - V \cos\varphi_c) = 0 \ .
\end{aligned} \tag{3.6}$$

From the transformed equations (2.17) - (2.22) follow

$$\begin{aligned}
&u_{wh} - (R + e_1) \sin\varphi_c - h = u_c \ , \\
&w_{wh} - (1 - \cos\varphi_c)(R + e_1) = w_c \ , \\
&(v + \dot{h})\{(1 + u_c')\cos\varphi_c + w_c' \sin\varphi_c + e_1 \varphi_c'\} \\
&\qquad - R\ \{\Omega + \dot{\psi}_{wh} - \dot{\varphi}_c\} = 0 \ .
\end{aligned} \tag{3.7}$$

Because of the damping terms introduced we are permitted to require as boundary conditions that all functions of (s,t) vanish for s → ±∞.

After stating explicitly the above conditions the field equations, i.e. the transformed equations (2.1) - (2.10) for $s \neq 0$, are treated separately for s > 0 and s < 0. For lack of space we cannot list them here.

4. Parameters and Normalization

4.1 Parameters

We choose the following parameters:

m_{wh} = 500 kg, W_{wh} = 5 000 N, Θ_{wh} = 50 kg m^2, R = 0.45 m,

m_b = 5 000 kg, w_b = 50 000 N,

$k_h = k_v = 6.25 \cdot 10^5$ N/m, $b_h = b_v = 1.12 \cdot 10^4$ Ns/m,

$K_h = K_v = 1 \cdot 10^8$ N/m^2, $B_h = B_v$ = 14.0 Ns/m^2,

$EA = 1.31 \cdot 10^9$ N, $BA = 2.60 \cdot 10^3$ Ns, $EI = 3.78 \cdot 10^6$ Nm^2,

$BI = 7.49$ Nsm^2, μ = 49 kg/m, e_1 = 75.1 mm, e_2 = 72.9 mm;

m_r, W_r are chosen as fractions of m_{wh} and W_{wh}; k_r and b_r are treated corresponding to k_h and b_h, respectively.

4.2 Reference Quantities, Normalization

To nondimensionalize variables and parameters we choose the following reference quantities

Length: L = R = 0.45 m; Force: $P = W_b$ = 50 000 N;

Frequency (time): $\Omega_R = 60\ s^{-1} \approx 1/L \cdot 100$ km/h.

Let $\beta = 1/33$ serve as a small parameter.

By multiplications of the form $q \cdot \beta^{n_1} \cdot L^{n_2} \cdot P^{n_3} \cdot \Omega_R^{n_4} =: \tilde{q}$, n_i integers, we nondimensionalize all parameters q and scale them into the interval $\sqrt{\beta} < \tilde{q} < 1/\sqrt{\beta}$.

Normalized Parameters

$\tilde{EA} = EA/P \cdot \beta^3 \approx 0.73$, $\tilde{EI} = EI/(L^2 P) \cdot \beta^2 \approx 0.34$,

$\tilde{BA} = BA\ \Omega_R/P \approx 3.18$, $\tilde{BI} = BI\ \Omega_R/(L_R^2 P\ \beta) \approx 1.46$,

$\tilde{K}_v = K_v\ L^2/P \cdot \beta^2 \approx 0.37$, $\tilde{K}_h = K_h\ L^2/P \cdot \beta^2 \approx 0.37$,

$\tilde{B}_v = B_v\ L^2\ \Omega\ /P \approx 0.34$, $\tilde{B}_h = B_h\ L^2\ \Omega\ /P \approx 0.34$,

$$\tilde{\mu} = \mu L^2 \Omega_R^2/P \approx 0.71, \quad \tilde{v} = v/(L\,\Omega_R) \approx 1.03,$$

$$\tilde{e}_1 = e_1/(L\,\beta) \approx 5.51, \quad \tilde{e}_2 = e_2/(L\,\beta) \approx 5.35,$$

$$\tilde{e}_{22} = e_2^2/(L^2\,\beta) \approx 0.87, \quad \tilde{m}_{wh} = m_{wh}\,L\,\Omega_R^2/P\cdot\beta \approx 0.49,$$

$$\tilde{k}_v = k_v\,L/P \approx 5.63, \quad \tilde{k}_h = k_h\,L/P \approx 5.63,$$

$$\tilde{b}_v = b_v\,L\,\Omega_R/P\cdot\beta \approx 0.18, \quad \tilde{b}_h = b_h\,L\,\Omega_R/P\cdot\beta \approx 0.18,$$

$$\tilde{W}_{wh} = W_{wh}/(P\,\beta) \approx 3.30, \quad \tilde{W}_b = W_b/P \approx 1.00,$$

$$\tilde{\Theta}_{wh} = \Theta_{wh}\,\Omega_R^2/(P\,L)\cdot\beta \approx 0.24, \quad \tilde{m}_b = m_b\,L\,\Omega_R^2/P\cdot\beta \approx 4.91,$$

$$\tilde{R} = R/L \approx 1.00, \quad \tilde{\Omega} = \Omega/\Omega_R \approx 1.00,$$

$$\tilde{N}_o = N_o/P = 0 \text{ (free)}.$$

Nondimensional Variables

The variables are brought into nondimensional forms, e.g. $\tilde{s} = s/L$, $\tilde{u} = u/L$, $\tilde{S} = S/P$, $\tilde{M} = M/(L\,P)$, etc.; the time only is scaled, $\tilde{t} = \Omega_R\,t/\beta$, such that a nondimensional circular frequency $\tilde{v} \approx 1$ corresponds to 300 Hertz (the range we are interested in).

5. Stationary Motion

For the parameters of section 4 the equations of section 3 possess a time independent solution: Looked onto the system from the moving body the deflection is "stationary". These equations have not yet been solved. But preliminary estimates show that all deflections are small (of some positive order of the small parameter β). Near C this does not hold for the loads S, M and V of the rail. But, for simplicity and to obtain a preliminary result for the title problem, let us neglect the stationary values of S and M, too, and take into account only the vertical load at the point of contact (in nondimensional form, tilde dropped):

$$V = V_{stat} = W_b + \beta\,W_{wh}\,. \tag{5.1}$$

We call this approximation "simplified stationary motion".

6. Variational Equations About the Simplified Stationary Motion

It is not difficult to derive from the equations (2.1) - (2.22), transformed according to section 3 and section 4, the variational equations for small (linear) oscillations of the system about the simplified stationary motion. (There is no space to give the details.)
From the field equations (cf. end of section 3.2) the loads can easily be eliminated and one arrives at the two partial differential equations for the variations $\delta u(s,t)$ and $\delta w(s,t)$, describing the oscillations of rail, in nondimensional form (tilde dropped):

$$
\begin{aligned}
&-\beta^3 v\, BA\, \delta u''' + (EA - \beta^3 v^2 \mu)\, \delta u'' + \beta^2 BA\, \delta\dot{u}'' \\
&+\beta^3 v\, B_h\, \delta u' + 2\beta^2 v \mu\, \delta\dot{u}' - \beta K_h\, \delta u \\
&-\beta^2 B_h\, \delta\dot{u} - \beta \mu\, \delta\ddot{u} \\
&-\beta^4 v\, e_2\, B_h\, \delta w'' + \beta^2 e_2 K_h\, \delta w' + \beta^3 e_2 B_h\, \delta\dot{w}' = 0 ,
\end{aligned}
\tag{6.1}
$$

$$
\begin{aligned}
&-\beta^3 v\, e_2\, B_h\, \delta u'' + \beta e_2 K_h\, \delta u' + \beta^2 e_2 B_h\, \delta\dot{u}' \\
&-\beta^3 v\, BI\, \delta w^{v} + EI\, \delta w^{iv} + \beta^2 BI\, \delta\dot{w}^{iv} + \beta^3 v\, e_{22} B_h\, \delta w''' \\
&-\beta (e_{22} K_h - \beta v^2 \mu + \beta^2 N_o)\, \delta w'' - \beta e_{22} B_h\, \delta\dot{w}'' \\
&-\beta^2 v\, B_v\, \delta w' - 2\beta v \mu\, \delta\dot{w}' + K_v\, \delta w + \beta B_v\, \delta\dot{w} + \mu\, \delta\ddot{w} = 0 .
\end{aligned}
\tag{6.2}
$$

From the equations (3.6) and (3.7) follow:

$$
\begin{aligned}
&[\delta u] = 0,\ [\delta w] = 0,\ [\delta u'] = 0,\ [\delta w'] = 0,\ [\delta w''] = 0 , \\
&\delta H - v\, BA\, [\delta u''] + (W_b + \beta W_{wh})\, \delta w_c' = 0 , \\
&\delta H\, e_1 - v\, BI\, [\delta w'''] + (W_b + \beta W_{wh})\, e_1\, \delta w_c' = 0 , \\
&\beta^2 \delta V - EI\, [\delta w'''] - \beta^2 BI\, [\delta\dot{w}'''] + \beta^3 v\, BI\, [\delta w^{iv}] \\
&\qquad + \beta^2 (W_b + \beta W_{wh})\, \delta u_c' = 0 ,
\end{aligned}
\tag{6.3}
$$

$$
\begin{aligned}
&(R + \beta\, e_1)\ \delta w_c' - \delta u_{wh} + \delta h = 0 \ , \\
&\qquad \delta w_c - \delta w_{wh} \qquad = 0 \ , \\
&\beta\, v\, \delta u_c' + \beta^2\, v\, e_1\, \delta w_c'' + R\, \delta \dot{w}_c' - R\, \delta \dot{\psi}_{wh} + \delta \dot{h} = 0 \ .
\end{aligned}
\tag{6.4}
$$

The equations (2.11) - (2.16) lead to

$$
\begin{aligned}
&m_b\, \delta \ddot{w}_b + \beta\, b_v(\delta \dot{w}_b - \delta \dot{w}_{wh}) + \beta^3\, k_v(\delta w_b - \delta w_{wh}) = 0 \ , \\
&m_{wh}\, \delta \ddot{u}_{wh} + \beta(b_h + b_{rh})\, \delta \dot{u}_{wh} - \beta\, b_{rh}\, \delta \dot{u}_r \\
&\qquad + \beta^3(k_h + k_{rh})\, \delta u_{wh} - \beta^3\, k_{rh}\, \delta u_r + \beta^3\, \delta H = 0 \ , \\
&m_{wh}\, \delta \ddot{w}_{wh} + \beta(b_v + b_{rv})\, \delta \dot{w}_{wh} - \beta\, b_v\, \delta \dot{w}_b - \beta\, b_{rv}\, \delta \dot{w}_r \\
&\qquad + \beta^3(k_v + k_{rn})\, \delta w_{wh} - \beta^3\, k_v\, \delta w_{wh} - \beta^3\, k_{rv}\, \delta w_r + \beta^3\, \delta V = 0 \ , \\
&\Theta_{wh}\, \delta \ddot{\psi}_{wh} - \beta^3\, R\, (W_b + \beta\, W_{wh})\, \delta w_c' - \beta^3\, R\, \delta H = 0 \ , \\
&m_r\, \delta \ddot{u}_r + \beta\, b_{rh}(\delta \dot{u}_r - \delta \dot{u}_{wh}) + \beta^3\, k_{rh}(\delta u_r - \delta u_{wh}) = 0 \ . \\
&m_r\, \delta \ddot{w}_r + \beta\, b_{rv}(\delta \dot{w}_r - \delta \dot{w}_{wh}) + \beta^3\, k_{rv}(\delta w_r - \delta w_{wh}) = 0 \ .
\end{aligned}
\tag{6.5}
$$

7. The Eigenvalue Problem Arising From the Variational Equations

Since the set of equations (6.1) - (6.5) is linear and autonomous (the coefficients do not depend on the time), we separate the time-dependency and look for a solution of the form

$$
\begin{aligned}
&\{\delta u,\ \delta w,\ \delta w_b,\ \delta u_{wh},\ \delta w_{wh},\ \delta \psi_{wh},\ \delta u_r,\ \delta w_r,\ \delta h,\ \delta H,\ \delta V\} \\
&= e^{\lambda t}\ \{u,\ w,\ w_b,\ u_{wh},\ w_{wh},\ \psi_{wh},\ u_r,\ w_r,\ h,\ H,\ V\}
\end{aligned}
\tag{7.1}
$$

where the variables in the braces on the right side do not depend on the time. When (7.1) is put into (6.1) - (6.5) we arrive at a homogeneous system of equations for $\{u, w, \ldots, H, V\}$ which con-

tains the parameter λ. Question: For which eigenvalues λ_i has this system nontrivial solutions?
Remark: The spectrum of this operator equation has a continuous and a discrete part. Here we look for the discrete part, the eigenvalues, only.

8. Solution of the Eigenvalue Problem

8.1 Solution of the Ordinary Differential Equations

After the substitution (7.1) the equations (6.1), (6.2) become ordinary differential equations for u(s) and w(s). We look for solutions of the form

$$(u,w) = e^{\varkappa s}(\hat{u},\hat{w}) =: e^{\varkappa s}\,\underline{x} \tag{8.1}$$

where $\varkappa$ is a parameter and $\hat{u}$, $\hat{w}$ are constant. By (7.1), (8.1) we get from (6.1), (6.2) the eigenvalue problem

$$\underline{A}(\varkappa,\lambda)\,\underline{x} = \underline{0}. \tag{8.2}$$

For assumed (complex) values λ this is an eighth order eigenvalue problem having the solutions $(\varkappa_i(\lambda), \underline{x}_i(\lambda))$, $i = 1,\ldots,8$. Let $\varkappa_i$, $i = 1,\ldots,I$ be the eigenvalues with $\mathrm{Re}\,\varkappa_i < 0$ and let $\mathrm{Re}\,\varkappa_i > 0$ for $i > I$. Then the solutions of the differential equations

$$\begin{bmatrix}u\\w\end{bmatrix}^{+} := \sum_{i=1}^{I} a_i\,\underline{x}_i\,e^{\varkappa_i s} \quad \text{and} \quad \begin{bmatrix}u\\w\end{bmatrix}^{-} := \sum_{I+1}^{8} a_i\,\underline{x}_i\,e^{\varkappa_i s} \tag{8.3}$$

satisfy the boundary conditions for decay for $s \to +\infty$ and for $s \to -\infty$, respectively (cf. the end of section 3.2).

8.2 Solution of the Complete Problem

Via (8.3) the two variables u, w in (7.1) are replaced by the 8 arbitrary constants a_i. Thus, together with the other 9 constants

of the right side of (7.1) there are 17 constants which are put into the 17 equations of the sets (6.3) - (6.5). We arrive at an eigenvalue problem of the form

$$\underline{B}(\lambda)\ \underline{y} = \underline{0}\ . \tag{8.4}$$

Some of the elements of the 17×17 matrix $\underline{B}(\lambda)$ depend explicitly on λ, other implicitly via the $\varkappa_i(\lambda)$ and $\underline{x}_i$ of the eigenvalue problem (8.2). Since no special procedure seems to be available for the solution of (8.4), its characteristic equation

$$\Delta(\lambda) = \det \underline{B} \tag{8.5}$$

was established numerically and solved by Newton's procedure.

For example, the normalized parameters listed in section 4.2, complemented by $m_{wh} = 0.35$, $m_r = 0.15$, $k_{rh} = k_{rv} = 2\,000.0$, $b_{rh} = b_{rv} = 0.01$, lead to the 6 couples of complex conjugate eigenvalues:

$$\lambda_1 \approx -5.50\cdot10^{-4} \pm j\ 5.61\cdot10^{-3},\ \lambda_2 \approx -3.74\cdot10^{-3} \pm j\ 1.42\cdot10^{-2},$$

$$\lambda_3 \approx -5.27\cdot10^{-3} \pm j\ 2.37\cdot10^{-1},\ \lambda_4 \approx -3.87\cdot10^{-3} \pm j\ 6.14\cdot10^{-1},$$

$$\lambda_5 \approx -1.07\cdot10^{-3} \pm j\ 7.24\cdot10^{-1},\ \lambda_6 \approx -7.38\cdot10^{-3} \pm j\ 7.47\cdot10^{-1}.$$

8.3 Some Numerical Problems

Numerical difficulties originate, except from the complexity of the problem, from two sources. First, the equations (6.1), (6.2) are nearly singular: The order of the eigenvalue problem (8.2) reduces to 6 for $\beta \to 0$.
The second difficulty arises from the search for parameter values for which the system becomes unstable, i.e. Re $\lambda_{i*} > 0$ for one or more i*. Most promising for that search are parameter sets where two λ_i coincide ("bifurcate") or, at least, come close together (above, k_{rh}, k_{rv} were chosen such that $|\lambda_5 - \lambda_6| \ll 1$).

But then all problems of (nearly) multiple roots occur; especially, it is difficult to secure (numerically) that the solutions of (8.2) contribute linearly independent lines to (8.4).
Example: For $B_h \approx 75$ we get Re $\lambda_6 > 0$, the stationary solution becomes unstable; but it is extremely difficult to trace the corresponding unstable region in the parameter space.

Reference

Knothe, K.; Valdivia, A. (1988) Riffelbildung auf Eisenbahnschienen - Wechselpiel zwischen Kurzzeitdynamik und Langzeit-Verschleißverhalten. ZEV-Glas. Ann. 112, 50-57.

Address: Prof. Dr. E. Brommundt, Institut für Technische Mechanik Technische Universität, PF 3329, D-3300 Braunschweig, Germany

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

RATIONAL APPROXIMATION FOR CALCULATION OF EIGENVALUES

Lothar Collatz

Abstract For the numerical calculation of eigenvalues rational approximation is sometimes more adequate than polynomial approximation, especially for the use of "quotient-inclusion"-theorems. These theorems permit for certain types of eigenvalue problems a rough inclusion sometimes of all eigenvalues. The paper shows the preparation for the numerical treatment and testing of examples.

1. Introduction, the quotient-inclusion-theorem

Sometimes rational approximation (shortly R. A.) gives more accurate values as polynomial approximation (shortly P. A.) (compare f. i. Meinardus [67], Collatz [68] [90], a. o.). P. A. can fail in unbounded domains and in the neighbourhood of singularities f. i. of poles (Werner [62/63] a. o.).

For eigenvalue problems the R. A. is often quite natural, because the eigenvalues λ_n can be written as quotients (f. i. Φ in (1.3)).

The eigenvalue problem may be given by a differential equation (1.1) and boundary conditions (1.2):

(1.1) $\quad Mu = \lambda Nu \quad$ in B,

(1.2) $\quad Su = 0 \quad$ on ∂B.

(usual notations: eigenfunction u(x), eigenvalue λ, open connected domain B in the point space R^n, given operators M,N, S; for details compare f. i. Courant-Hilbert [68], Collatz [63][68] a. o.).

For an "admissible" function w(x) one can introduce the quotient:

(1.3) $\quad \Phi(x) = \frac{Mw}{Nw}$

Under certain conditions on the operators M,N,S the quotient ϕ in (1.3) gives immediately inclusion theorems for the eigen-

values; if for a chosen function w(x) the quotient ϕ in (1.3) lies between finite bounds ϕ_{min} and ϕ_{max} (Collatz [41],[63]), then there exists at least one eigenvalue λ_s with

(1.4) $$\phi_{min} \leq \lambda_s \leq \phi_{max} .$$

This "quotient-inclusion theorem" was generalized considerably by Lehmann [50],[63], Albrecht [68],[87], Goerisch [78],[84] and Velte [76] a. o., and illustrated by numerous examples. The purpose of this paper is to bring for certain classes of problems the theorem into an immediately applicable form.

2. An ordinary differential equation

The following very special class of eigenvalue problems occurs frequently in various fields of application (vibrations, bending a. o.)

(2.1) $$-y''(x) = \lambda p(x) y(x) \quad \text{in } [a,b]$$

with boundary conditions

(2.2) $$y(a) = y(b) = 0$$

or with more general boundary conditions of Sturm's type.

The method works if the prescribed function p(x) is not too complicated.

For many problems of type (2.1), (2.2) eigenvalues $\hat{\lambda}_n$ and eigenfunctions are completely known. We mention some cases (Collatz [63] p. 454):

Problem $-y''(x)=\hat{\lambda}\hat{p}(x)y(x)$, $y(a)=y(b)=0$	Eigenvalues $\hat{\lambda}_n$ $(n \in \mathbb{N})$
(2.3) $-y''(x)=\hat{\lambda}\frac{1}{(x+c)^2}y(x)$ with $x+c\neq 0$ in $[a,b]$	$\left[\frac{n\pi}{\ln\frac{c+b}{c+a}}\right]^2 + \frac{1}{4}$
(2.4) $-y''(x)=\hat{\lambda}\frac{1}{(x^2+c^2)^2}y(x)$	$\left[\frac{cn\pi}{\operatorname{arctg}\frac{b}{c}-\operatorname{arctg}\frac{a}{c}}\right]^2 - c^2$
(2.5) $-y''(x)=\hat{\lambda}\frac{1}{(x^2-c^2)^2}y(x)$ with $x^2-c^2\neq 0$ in $[a,b]$	$\left[\frac{2cn\pi}{\ln\frac{c+b}{c+a}\frac{c-a}{c-b}}\right]^2 + c^2$

We now try to approximate a prescribed function p(x) by two functions $k_1\hat{p}(x)$, $k_2\hat{p}(x)$ with

(2.6) $\quad k_1\hat{p}(x) \leq p(x) \leq k_2\hat{p}(x)$,

then the inclusion holds

(2.7) $\quad \frac{\hat{\lambda}_n}{k_2} \leq \lambda_n \leq \frac{\hat{\lambda}_n}{k_1}$ $(n \in \mathbb{N})$.

If (2.6) is satisfied, then all eigenvalues λ_n are included with the same procentual error $\sigma = 2(k_2-k_1)/(k_2+k_1)$. For a prescribed function $p(x)$ one has to choose the suitable $\hat{p}(x)$.

The function $\hat{p}(x) = \frac{1}{(x+c)^2}$ in (2.3) is suitable for functions of the form of fig. 1, analogously fig. 2 for $\hat{p}(x) = \frac{1}{(x^2+c^2)^2}$ and fig. 3 for $\hat{p}(x) = \frac{1}{(x^2-c^2)^2}$.

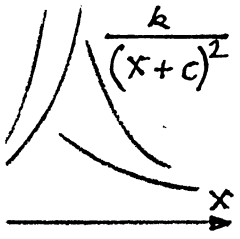


Fig. 1

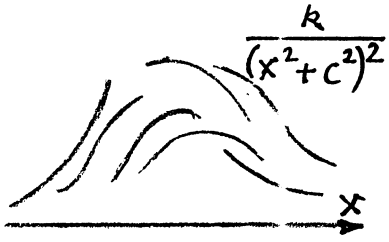


Fig. 2

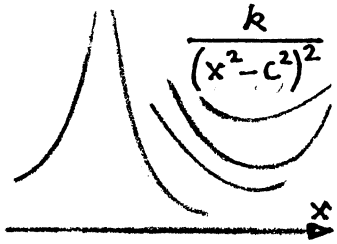


Fig. 3

Example I: $-y''(x) = \lambda e^{-\frac{1}{2}x} y(x)$; $y(0) = y(1) = 0$.

We compare $p(x) = e^{-\frac{1}{2}x}$ with $k\cdot\hat{p}(x) = \frac{k}{(x+c)^2}$; from

$p(0) = k\hat{p}(0)$ and $p(1) = k\hat{p}(1)$

we get $k\hat{p}(x) \leq p(x)$ with

$c = \frac{1}{e^{1/4}-1} = 3.520...$ and $k = c^2 = 12.39...$.

We multiply $k\hat{p}(x)$ with a factor τ such that $p(x) \leq \tau k\hat{p}(x)$ (fig. 4, symbolic scetch); this holds for $\tau = 1.015...$, $\tau k = 12.59...$ and gives the inclusion

$$\frac{1}{12.60}\{16n^2\pi^2+\tfrac{1}{4}\} \leq \lambda_n \leq \frac{1}{12.39}\{16n^2\pi^2+\tfrac{1}{4}\} \quad (n \in \mathbb{N}) \text{ with } \sigma < 0.016.$$

Example II: $p(x)$ may be a symmetric function, $p(x) = p(-x)$ in $(-b,b)$; boundary conditions $y(\pm b) = 0$. We only need the values of $p_0 = p(0)$ and $p_1 = p(b)$ to get a fixed approximating function

$k\hat{p}(x)$. Suppose $p_o > p_1$ (the case $p_o < p_1$ can be treated analogously), fig. 5.

We take $\hat{p}(x) = \frac{1}{(c^2+x^2)^2}$; from $p_o = \frac{k}{c^4}$ and $p_1 = \frac{k}{(c^2+b^2)^2}$ we get

$$\frac{1}{c^2} = \frac{1}{b^2}\left(\sqrt{\frac{p_o}{p_1}} - 1\right) \quad \text{and} \quad k = p_o c^4.$$

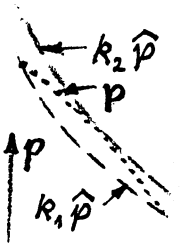


Fig. 4

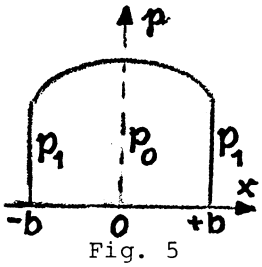


Fig. 5

Numerical example: $p(x) = 2+\cos x$, $b=\pi/2$; $p_o=3$, $p_1=2$;

$\frac{1}{c^2}=\frac{2\sqrt{6}-4}{\pi^2}=0.09108...$, $k=\frac{6+3\sqrt{6}}{4}\pi^4=361.5...$, $\tau=1.008...$; $\sigma<0.0081$.

$$0.342 = \lambda_1 \le 0.352, \qquad 3.33 \le \lambda_3 \le 3.42,$$
$$1.465 \le \lambda_2 \le 1.499, \qquad 5.95 \le \lambda_4 \le 6.09.$$

3. A partial differential equation (P.D.E.)

We consider a very special class: Radialsymmetric vibrations of a membrane of variable thickness; polarcoordinates r,φ, distribution of displacement $u(r,\varphi)$ in $B = \{(r,\varphi): 0 \le r < R, 0 \le \varphi \le 2\pi\}$ (compare Wetterling [77]):

$$(3.1) \qquad -\Delta u = -\left(u_{rr} + \frac{1}{r}u_r\right) = \lambda g(r)u(r) \text{ in } B,$$

$$(3.2) \qquad u = 0 \text{ on } \partial B.$$

We shortly explain a method which is immediately applicable for every given positive continuous function $g(r)$. We use for (1.3), (1.4) the function (R=1):

$$(3.3) \qquad w(r) = (1-r^2)\{(3-r^2)-\alpha(5-r^4)\},$$

$$(3.4) \qquad -\Delta w(r) = (1-r^2)\{16-4\alpha(5+9r^2)\}.$$

If the quotient (1.3)

$$\phi(r) = \frac{-\Delta w(r)}{g(r)w(r)}$$

lies between finite bounds the inclusion (1.4)

$$\Phi_{min} \leq \lambda_s \leq \Phi_{max}$$

holds. For instance for $g(r) = (1+\frac{r^2}{2})^{-1}$ one gets immediately:

$\alpha = 0$ $\qquad \frac{16}{3} \leq \lambda_1 \leq 12,$

$\alpha = 0.2$ $\qquad \frac{6}{=} \leq \lambda_1 \leq \frac{32\sqrt{15}-90}{5} < 6.79,$

$\alpha = \frac{535+15\sqrt{6}}{2279} = 0.250...$ $\qquad 2.93 < \frac{282-30\sqrt{6}}{71} \leq \lambda_1 \leq 10\sqrt{6}-18 < \underline{\underline{6.50}}.$

A generalization of this inclusion theorem is given for P.D.E. of 4. order (Collatz [43] a. o.).

I thank Prof. Dr. J. Albrecht, Dr. P. Klein (Clausthal) for valuable hints and Mr. Th. Schiemann (Hamburg) for numerical calculations.

References

Albrecht, J. [68]: Verallgemeinerung eines Einschließungssatzes von L. Collatz. ZAMM 48 (1968), T43 - T46

Albrecht, J. und F. Goerisch [87]: Anwendungen des Verfahrens von Lehmann auf Schwingungsprobleme. ISNM 83 (1987), 1 - 9

Collatz, L. [41]: Einschließungssatz für die Eigenwerte von Integralgleichungen. Math. Zeitschr. 47 (1941), 395 - 398

Collatz, L. [41]: Einschließungssatz für die charakteristischen Zahlen von Matrizen. Math. Zeitschr. 48 (1942), 221 - 226

Collatz, L. [43]: Einschließung für Eigenwerte bei partiellen Differentialgleichungen 2. und 4. Ordnung. ZAMM 43 (1963), 277 - 280

Collatz, L. [63]: Eigenwertaufgaben mit technischen Anwendungen. Leipzig (1963), 2. Aufl.

Collatz, L. [68]: Funktionalanalysis und numerische Mathematik. Springer (1968)

Collatz, L. [90]: Rational and algebraic approximation for initial- and boundary-value-problems. ISNM 90 (1989), 103 - 106

Courant, R. und D. Hilbert [68]: Methoden der mathematischen Physik. Springer (1968)

Goerisch, F. [78]: Über Quotienten-Einschließungssätze bei allgemeinen Eigenwertaufgaben. ISNM 39 (1978), 86 - 100

Goerisch, F. und J. Albrecht [84]: Eine einheitliche Herleitung von Einschließungssätzen für Eigenwerte. ISNM 69 (1984), 58 - 88

Lehmann, N. J. [50]: Beiträge zur numerischen Lösung linearer Eigenwertprobleme. ZAMM 29 (1949), 341 - 356

Lehmann, N. J. [50]: Beiträge zur numerischen Lösung linearer Eigenwertprobleme. ZAMM 30 (1950), 1 - 16

Lehmann, N. J. [63]: Optimale Eigenwerteinschließungen. Num. Math. 5 (1963), 246 - 272

Meinardus, G. [67]: Approximation of Functions: Theory and Numerical Methods. Springer (1967)

Velte, W. [76]: Zur Eigenwerttheorie Steklovscher Probleme. Mitt. Math. Sem. Giessen, Heft 121, 125 - 137

Werner, H.[62]: Die konstruktive Ermittlung der Tschebyscheff-Approximierenden im Bereich der rationalen Funktionen. Arch. Rat. Mech. Anal. 11 (1962), 368 - 384

Werner, H. [63]: Rationale Tschebyscheff-Approximation, Eigenwerttheorie und Differenzenrechnung. Arch. Rat. Mech. Anal. 13 (1963), 330 - 347

Wetterling, W. [77]: Quotienteneinschließung bei Eigenwertaufgaben mit partieller Differentialgleichung. ISNM 38 (1977), 213 - 218

Prof. Dr. L. Collatz
Institut für Angewandte Mathematik
der Universität Hamburg
Bundesstraße 55

D-2000 Hamburg 13

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

SOME QUESTIONS IN EIGENVALUE PROBLEMS IN ENGINEERING

Isaac Elishakoff

Florida Atlantic University, Center for Applied Stochastics Research and Department of Mechanical Engineering, Boca Raton, Florida, USA

Abstract

This paper deals with some questions arising in dealing with engineering eigenvalue problems. The emphasis is placed on the coincident or closely spaced natural frequencies from the standpoint of the normal mode method applied for determining the structural response; it concentrates on the degree of refinement in the determination theories which should be introduced, in order to obtain accurate predictions of the response of structures subjected to high-frequency excitation. Nonsymmetric eigenvalue problems arising in composite plates are considered. Some pertinent questions associated with the eigenvalue problems arising for small vibrations superimposed on the basic nonlinear state, as well as those for nonprobabilistic treatment of uncertainty are posed. Special considerations associated with interval arithmetic and convex models of uncertainty are elucidated. Each section is concluded with open problems, awaiting for their resolution by the combined efforts of mathematicians and engineers.

1. Introduction

I am very pleased to be among mathematicians, as one of the few engineers at this meeting. This allows me to continue my learning process. Noting Professor Collatz's statement at this conference that "we live from applications", I thought it would be useful to this mathematicians' forum to see what are some of the problems we engineers are facing, and from which angle eigenvalue problems are viewed in our professional lives. Accordingly,

I will discuss some of the problems which are of interest to me, in the hope that this review promotes interaction between "real life" and mathematics.

2. Questions Associated with Closely Spaced or Coalescent Eigenvalues

The problem of coincident or close eigenvalues has been extensively discussed during the conference. How does this problem appear in a real-life situation? Does it have any engineering significance?

In trying to answer this question, consider a 2 degrees-of-freedom system consisting of two masses, connected with springs and dashpots. We introduce a control parameter ϵ which permits simultaneous consideration of a number of familiar systems. The first mass is subjected to a random excitation which is a weakly stationary random process, with zero mean function and the white noise autocorrelation function. The mean square responses of this system can be put in the following form [1]:

$$E\left(|X_1|^2\right) = R_{1,1} + R_{1,2} + R_{2,1} + R_{2,2} \tag{1}$$

$$E\left(|X_2|^2\right) = R_{1,1} - R_{1,2} - R_{2,1} + R_{2,2} \tag{2}$$

where $R_{1,1}$ could be interpreted as the part of the response associated with the contribution of the first natural mode of vibration, $R_{2,2}$ - it's counterpart for the second natural mode, and $R_{1,2}$, $R_{2,1}$ are the cross-correlation terms stemming from interaction of the two modes. These have the form

$$R_{1,2} = R_{2,1} \sim \frac{1}{\left(\omega_1^2 - \omega_2^2\right)^2 + \alpha\epsilon\omega_1^{\beta}}, \tag{3}$$

where ω_j are the free vibration frequencies, α and β are positive coefficients. Inspection of Eq. (1) shows that if the natural frequencies are far apart, $R_{1,2}$ and $R_{2,1}$ can be neglected in comparison to $R_{1,1}$. However, if the natural frequencies happen to be closely spaced or coincide, then

$$R_{1,2} = R_{2,1} \sim R_{11} \tag{4}$$

in which case the cross-correlations may contribute up to 50% of the total response.

Consider, for example, the two-degrees-of-freedom system, whose motion is governed by the differential equations:

$$\begin{aligned} &m\ddot{X}_1 + (1+\epsilon)c\dot{X}_1 - \epsilon c\dot{X}_2 + (1+\epsilon)kX_1 - \epsilon kX_2 = F_1(t) , \\ &m\ddot{X}_2 - \epsilon c\dot{X}_1 + (1+\epsilon)c\dot{X}_2 - \epsilon kX_1 + (1+\epsilon)kX_2 = 0 , \end{aligned} \tag{5}$$

where $F_1(t)$ is a random excitation with zero mean representing an ideal white noise with spectral density $S(\omega) = S_0$, m denotes the mass, k - spring constant, c - damping coefficient. The control parameter ϵ characterizes the degree of coupling of the two masses. The problem consists of finding the mean-square values of the displacements $X_1(t)$ and $X_2(t)$, and

those of velocities $\dot{X}_1(t)$ and $\dot{X}_2(t)$. We immediately note that at $\epsilon \to 0$, we should find the following result (see Ref. 1, Eq. 9.67)

$$E(X^2) = \frac{S_0 \pi}{ck} \tag{6}$$

valid for one-degree-of-freedom system, whereas for the mean-square value of the displacement $X_2(t)$, we should have a vanishing result, as the response of an unexcited system.

The natural frequencies of the system are

$$\omega_1 = \left[\frac{k}{m}\right]^{1/2} , \qquad \omega_2 = \left[\frac{k}{m}(1+2\epsilon)\right]^{1/2} \tag{7}$$

The following expressions are valid (see Ref. 1, Eqs. 9.205) for the meansquare responses:

$$E(|X_1|^2) = \frac{\pi S_0}{4kc}\left[1 + \frac{1}{(1+2\epsilon)^2} + \frac{2\alpha}{\epsilon^2/(1+\epsilon)+\alpha(1+2\epsilon)}\right] \tag{8}$$

$$E(|X_2|^2) = \frac{\pi S_0}{4kc}\left[1 + \frac{1}{(1+2\epsilon)^2} - \frac{2\alpha}{\epsilon^2/(1+\epsilon)+\alpha(1+2\epsilon)}\right] \tag{9}$$

$$E(|\dot{X}_1|^2) = \frac{\pi S_0}{4mc}\left[1 + \frac{1}{(1+2\epsilon)^2} + \frac{2\alpha}{\epsilon^2/(1+\epsilon)+\alpha(1+\epsilon)}\right] \tag{10}$$

$$E(|\dot{X}_2|^2) = \frac{\pi S_0}{4mc}\left[1 + \frac{1}{(1+2\epsilon)^2} - \frac{2\alpha}{\epsilon^2/(1+\epsilon)+\alpha(1+\epsilon)}\right] \tag{11}$$

where $\alpha = c^2/km$. For ϵ tending to zero, $E(|X_1|^2) \to \pi S_0/kc$. This is, we recover the result given in Eq. (6) for the single-degree-of-freedom system. Moreover, $E(|X_2|^2) \to 0$, since the second mass becomes a separate unexcited system. These results are expected as noted above.

The first two terms in Eqs. (8) - (11) are associated with the modal auto-correlations $R_{1,1} + R_{2,2}$, whereas the third term represents modal cross-correlations $R_{j,k}$ $(j \neq k)$. Significantly, contributions of the two correlations for $\epsilon=1$ and $\alpha=0.01$ are 96.72% and 3.28%, respectively. However, for ϵ tending to zero, they tend to contribute equally, so that the error incurred by disregarding the cross-correlations reaches 50%. In addition, with the cross-correlations omitted, we can no longer arrive at the result in Eq. 6 associated with the single degree-of-freedom system.

The reason for this rather dramatic contribution is revealed by Eq. 3 for the cross-correlation term: at $\epsilon \to 0$, the natural frequencies crowd together, as is seen in Eq. 7, and in these circumstances the contribution of the cross-correlation term is of the same order of magnitude as that of its autocorre-

lation counterpart. By contrast, when the natural frequencies are far apart, the cross-correlation term is of order of magnitude of $\zeta^2/(\omega_1-\omega_2)^2$ times the autocorrelation term and may be omitted. Here ζ denotes the nondimensional damping coefficient $\zeta_j = c\omega_j/2k$. In other words, the cross-correlation term can be omitted of the following strong inequality holds [2]:

$$\zeta_1^2 \ll \left| 1 - \frac{\omega_1^2}{\omega_2^2} \right| , \qquad \zeta_2^2 \ll \left| 1 - \frac{\omega_1^2}{\omega_2^2} \right| \tag{12}$$

Consider now the n-degrees-of-freedom system [3]. The system consists of n identical masses m, and identical spring stiffnesses k. Each mass is connected both with the ground and with the other masses. Damping proportional to the mass is provided by a deshpot attached to each mass, the damping coefficient being c. The first mass is subjected to the ideal white noise loading with intensity S_0. We are interested in the mean-square responses of the system. The equations of motion read:

$$\begin{aligned}
&m\ddot{x}_1 + c\dot{x}_1 + kx_1 + k\sum_{j=2}^{n}(x_1-x_j) = f_1 \\
&m\ddot{x}_2 + c\dot{x}_2 + k(x_2-x_1) + k\sum_{j=3}^{n}(x_2-x_j) = 0 \\
&m\ddot{x}_3 + c\dot{x}_3 + kx_3 + k(x_3-x_1) + k(x_3-x_2) + k\sum_{j=4}^{n}(x_3-x_j) = 0 , \\
&\vdots \\
&m\ddot{x}_{n-1} + c\dot{x}_{n-1} + kx_{n-1} + k\sum_{j=1}^{n-2}(x_{n-1}-x_j) + k(x_{n-1}-x_n) = 0 , \\
&m\ddot{x}_n + c\dot{x}_n + kx_n + k\sum_{j=1}^{n-1}(x_n-x_j) = 0 ,
\end{aligned} \tag{13}$$

The natural frequencies are represented by two sets of eigenvalues: one comprising the separate first natural frequency

$$\omega_1 = \left[\frac{k}{m}\right]^{1/2}, \tag{14}$$

which can be interpreted as that of a single-degree-of-freedom systems, and the other remaining (coalescent) n-1 natural frequencies

$$\omega_2 = \omega_3 = \ldots = \omega_n \left[\frac{(n+1)k}{m}\right]^{1/2}. \tag{15}$$

For the mean-square displacements, we have

$$E\left[|X_j|^2\right] = \sum_{\alpha=1}^{n} v_{j\alpha} R_{\alpha\alpha} + 2 \sum_{\substack{\alpha=1 \\ \alpha<\beta}}^{n} \sum_{\beta=1}^{n} v_{j\alpha}\, v_{j\beta}\, R_{\alpha\beta} \tag{16}$$

where $v_{j\alpha}$ are elements of the modal matrix, and $R_{\alpha\beta}$ are the elements of the variance-covariance matrix.

For the first mass the error, incurred by neglecting the cross-correlations, denoted by the second, double sum in Eq. 16, reaches 70%. In conclusion, the coincidence of n-1 natural frequencies results in the highly intensified response. Therefore, omission of the cross-correlation between different modes yields misleading result both qualitatively and quantitatively.

For square plates, or rectangular ones with integer side ratio, there may be double, triple, quadruple, etc. frequencies, due to the specific symmetries involved. For example, for a plate simply supported all round, the natural frequencies have the form

$$\omega_{mn}^2 = \frac{D}{\rho h} \frac{\pi^4}{a^4} \left(m^2+n^2\right)^2, \tag{17}$$

where m and n are integers, representing the number of half sine waves in the mode shape of the plate in the x, y-directions, respectively:

$$X_{mn} = \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{a} . \qquad (18)$$

In addition to the double frequencies $\omega mn = \omega nm$, there are triple frequencies, e.g. $\omega_{17} = \omega_{71} = \omega_{55}$, quadruple frequencies, such as $\omega_{19,7} = \omega_{17,11} = \omega_{11,17} = \omega_{7,19}$ or $\omega_{21,1} = \omega_{19,9} = \omega_{1,21} = \omega_{9,19}$, etc.

Analogously, cylindrical shells may exhibit a "cluster" of close (though not coincident) natural frequencies. Such condensation leads to a highly intensified response at the zone of load application. Fig. 5 in Ref. 4 shows the relevant response variation in the axial direction $\xi = x/L$ (L denoting shell length) for a shell under ring loading $q(x,t) = \delta(\xi-0.3)\, q(t)$, where q(t) is a stationary random process with white noise autocorrelation function. Curve S_1, is associated with the contribution of like modes and S_2 with the cross-correlations, R_{mn} $(m \neq n)$ for closely spaced modes. Under these circumstances, the contribution of the cross-correlations may exceed that of the autocorrelations by a factor of 3/2, so that the total response on the excited cross-section may exceed the share of the autocorrelations by a factor of 3.

Now, consider a fully clamped square plate, i.e., one with boundary conditions

$$w = \frac{\partial w}{\partial n} = 0 , \qquad (19)$$

where w is the plate displacement and n a normal to the plate boundary. In the absence of an exact solution to this problem, engineers resort to approximate techniques like the finite difference method or the finite element method. This gives rise to the following practical questions:

Does the plate have multiple natural frequencies of multiplicity 2, 3, 4 and more as it occurs in the all round simply supported plates? An approximate technique developed by the author [5] suggests that there are coincident natural frequencies of multiplicity 2. No natural frequencies of higher

multiplicity could be detected by the numerical methods presently known to the author.

Can the answers to these questions be established without performing actual numerical computations? Note that symmetry property was utilized by Gorman [6] to establish the presence of double frequencies.

In this connection the following questions arise:

1. Are there triple, quadruple, et. natural frequencies in a square all-round damped plate?

2. How many symmetry classes are possessed by a fully-damped square plate?

3. How many symmetry classes are possessed by a fully-damped rectangular plates with commensurable side lengths?

4. How many natural frequencies satisfy the inequality

$$|\omega_{ij} - \omega_{mn}| < \gamma \tag{20}$$

for specified γ, in a specific frequency band

$$\Omega_1 < \omega_{ij}, \omega_{mn} < \Omega_2 \tag{21}$$

there Ω_1 and Ω_2 representing the lower and upper levels of the frequency band, respectively?

3. Which Governing Differential Equations to Use?

Some of the papers of this conference are dedicated to the accurate numerical evaluation of large number of eigenvalues. Then some sophisticated procedures are developed to evaluate there eigenvalues. The question arises: Which governing differential equations to use? For example, if we are interested in the first, say, 150 natural frequencies of the plate, the questions which arise are:

1. Should we use the classical, Kirchoff-Love theory of plates?

2. Should we use the refined, Mindlin theory of plates?

3. Should we use the elastokinetic equations?

The answer to these questions may depend on the geometry of the structure, namely the thickness-to-side length ratio. However, often it is overlooked, that the type of the excitation to which the structure is subjected, is of paramount importance.

Consider for example a beam, subjected to the random wide-band excitation. The following question arises: Which governing differential equations should be used for describing the motion of the beam?

The classical Bernoulli-Euler equation of free vibration of uniform beam reads

$$EI \frac{\partial^4 w}{\partial x^4} + \rho A \frac{\partial^2 w}{\partial t^2} = 0 \tag{22}$$

where E is the modulus of elasticity, I is the moment of inertia of the cross-section, A is the cross-sectional area, ρ is the mass density of the

beam material, w is the deflection, x is the spatial coordinate along the beam axis and t-time.

In 1877, Lord Rayleigh [7] refined Eq. (22) by taking into account the rotatory movement of the beam elements in addition to translatory ones. This resulted in the following equation:

$$EI \frac{\partial^4 w}{\partial x^4} + \rho A \frac{\partial^2 w}{\partial t^2} + \rho I \frac{\partial^4 w}{\partial x^2 \partial t^2} = 0 \tag{23}$$

where I is the polar moment of inertia.

Timoshenko [8] in 1921 further refined the theory by taking into account the shear deformation of the beam, and derived the following equation:

$$EI \frac{\partial^4 w}{\partial x^4} + \rho A \frac{\partial^2 w}{\partial t^2} + \rho I \left[1 + \frac{E}{kG}\right] \frac{\partial^4 w}{\partial x^2 \partial t^2} + \frac{m^2 r^2}{kAG} \frac{\partial^4 w}{\partial t^4} = 0 \tag{24}$$

where $r = (I/A)^{1/2}$ is the radius of gyration, and $m = \rho A$.

It is known [9,10] that the rotary inertia and shear deformation effects become more and more important with the increase in the frequency of vibration. They also become more important when the slenderness ratio L/r decreases, L being the length of the beam. Samuels and Eringen [11] concluded that the mean-square displacements of the beam, subjected to wide-band excitation, derived via the classical Bernoulli-Euler and refined Timoshenko theories, differ, in the pure transverse damping case, by less than five percent. This finding caused a "calm" in random vibration research in the framework of refined theories.

In Ref. 12, Elishakoff and Lubliner studied the random vibrations of beams based on the Timoshenko theory. Two types of excitation were considered - distributed loading represented by spacewise white noise, and concentrated

point loading. Timewise, both excitations were given by band-limited white noise, namely

$$S(\omega) = \begin{cases} S_0, & \text{for } \Omega_{c1} \leq |\omega| \leq \Omega_{c2} \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

where Ω_{c1} is a lower cutoff frequency and Ω_{c2} is upper cutoff frequency.

A number of interesting findings were uncovered in Ref. 12: first - there are "important" modes of vibration (i.e. those which contribute considerably to formulation of the random vibration response) and unimportant modes of vibration. Important modes of vibration are associated with the eigenvalues ω_j falling within the excitation band.

$$\Omega_{c1} < \omega_j < \Omega_{c2} \tag{26}$$

Now, for $\Omega_{c1} \to 0$, the classical and higher order theories predict coincidental or very close results (like in Ref. 11), depending on upper cutoff frequency Ω_{c2}. For higher values of Ω_{c2}, when higher modes are excited and the contribution of lower modes is negligible, the mean square response predicted by the refined Timoshenko theory may exceed its classical counterpart, found via Bernoulli-Euler theory, by as much as a factor 2! The lesson learned from Ref. 12, is that the number of frequencies sought, as well as the governing equations for the boundary value problem, should be made consistent with the type and the frequency content of the excitation.

In this connection, the following questions arise:

1. How to determine the number of "important" eigenvalues dependng on the geometry and excitation of the system?

2. How to adjust the accuracy of the determination of the eigenvalues with that of the response of the system?

Indeed, engineers probably are not interested just in the natural frequencies per se. We are looking at the natural frequencies and the modes of vibration as tools to determine the response of the engineering system, be it deterministic or probabilistic.

4. NONSYMMETRIC EIGENVALUE PROBLEMS

In the subsequent analysis the distribution of the displacement field across the composite plate's thickness is taken as

$$U_1 = z\psi_x - \frac{4}{3h^2} z^3(\psi_x + W,_x)$$

$$U_2 = z\psi_y - \frac{4}{3h^2} z^3(\psi_y + W,_y) \qquad (27)$$

$$U_3 = W$$

Where U_1, U_2 and U_3 are the components of the three-dimensional displacement vector in the x -, y - and z directions respectively. ψ_x and ψ_y denote the rotation of the normals to mid-plane about the y - and x - axes, respectively, while $(.),_j$ denotes the partial derivative of (.) with respect to the indicated coordinate.

The above representation of the displacement field yields a parabolic distribution of transverse shear strains across the plate thickness and the condition of zero in-plane loads on the bounding planes of the plane. By this means, the need to introduce a transverse shear correction factor k (see e.g. Eq. 24 appearing in the Timoshenko beam theory [13] or Mindlin plate theory [11], or as in the case of the first order shear deformation theory for the composite plates, is circumvented.

For an orthotropic material, on which the elastic axes of the layer coincide with the geometrical ones, the pertinent constitutive equations may be expressed as

$$\begin{aligned}
\sigma_1 &= Q_{11}\,\epsilon_1 + Q_{12}\,\epsilon_2 + R_{11}\sigma_3 \\
\sigma_2 &= Q_{12}\,\epsilon_1 + Q_{22}\,\epsilon_2 + R_{22}\sigma_3 \\
\sigma_4 &= Q_{44}\,\epsilon_4 \\
\sigma_5 &= Q_{55}\,\epsilon_5 \\
\sigma_6 &= Q_{66}\,\epsilon_6
\end{aligned} \tag{28}$$

where

$$\begin{aligned}
&Q_{11} = E_1/\Omega,\ \Omega_{22} = E_2/\Omega,\ \Omega_{12} = \nu_{12}\,E_2/\Omega \\
&\Omega = 1 - \nu_{12}\nu_{21},\ Q_{44} = G_{23},\ Q_{55} = G_{13},\ Q_{66} = G_{12}
\end{aligned} \tag{29}$$

and

$$\begin{aligned}
R_{11} &= (E_1/E_3)\,(\nu_{31} + \nu_{21}\nu_{32})/\Omega \\
R_{22} &= (E_2/E_3)\,(\nu_{32} + \nu_{12}\nu_{31})/\Omega
\end{aligned} \tag{30}$$

are the reduced elastic constants.

The distribution of the transverse normal stress, σ_3 can be obtained by integration across the segment (0,z), of the equation of motion of the three-dimensional elasticity theory, written in the absence of the body forces as

$$\sigma_{i3},i = \rho\,\ddot{U}_3\ ,\ i = (1,2,3) \tag{31}$$

where ρ is the mass density and dots denote the time derivative. This yields finally the differential equations

$$\begin{aligned}
L_{11}\psi_X + L_{12}\psi_y + L_{13}\,W &= 0 \\
L_{21}\psi_X + L_{22}\psi_y + L_{23}\,W &= o \\
L_{33}\psi_X + L_{32}\psi_y + L_{33}\,W &= -P_3
\end{aligned} \tag{32}$$

where L_{ij} are differential operators [15].

Now, it turns out that the operators L_{ij} are not symmetric, i.e., $L_{ij} \neq L_{ij}$. The solution functions are represented in the form that exactly satisfies the

boundary conditions

$$\psi_x(x,y,t) = \sum_{m,n} X_{mn} \cos \alpha x \sin \beta y \, T_{mn}(t)$$

$$\psi_y(x,y,t) = \sum_{m,n} Y_{mn} \sin \alpha x \cos \beta y \, T_{mn}(t) \tag{33}$$

$$W(x,y,t) = \sum_{m,n} W_{mn} \sin \alpha x \sin \beta y \, T_{mn}(t)$$

where $\alpha = m\pi/a$, $\beta = n\pi/b$, m and n are integers representing the number of half-waves in the x and y directions, respectively; a and b are side lengths of the plate. In the free vibration problem

$$p_3 = 0, \; c = 0, \; T_{mn}(t) = \exp(i\omega_{mn} t), \tag{34}$$

where ω_{mn} is the natural frequency. The eigenvalue problem becomes

$$\{[K] - \omega_{mn}^2 [M]\} \{\Delta\}_{mn} = \{0\} \tag{35}$$

where

$$\{\Delta\}_{mn}^T = \{ X_{mn}, Y_{mn}, W_{mn}\} \tag{36}$$

In Eq. 35, both K and M are real and nonsymmetric matrices.

The following mathematical questions arise:

1. Are the matrices K and M symmetrizable?

2. If yes, how?

One of the prominent mathematicians at this conference remarked that on one hand, these questions are presently unanswerable, and on the other hand, these questions are mathematically totally legitimate.

Of course one can bring examples of nonsymmetric symmetrizable matrices. For example [16]:

$$A = \begin{bmatrix} 5 & 4 \\ 16 & 7 \end{bmatrix} \tag{37}$$

Then with matrices B and C

$$B = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} \tag{38}$$

we obtain

$$BAC = D , \tag{39}$$

where D is a symmetric matrix

$$D = \begin{bmatrix} 5 & 8 \\ 8 & 7 \end{bmatrix} . \tag{40}$$

In our analysis (Ref.15) we applied the left-and-right-eigenvectors, and the associated biorthogonality condition

$$(\omega_{mn}^2 - \tilde{\omega}_{mn}^2) \int_0^a \int_0^b \{\Delta_{mn}\} \{\tilde{\Delta}_{pq}\} \, dxdy = 0 \tag{41}$$

where Δ_{mn} are the eigenvectors of the original problem and $\tilde{\Delta}_{pq}$ are those of the adjacent operator.

5. Eigenvalue Problems Arising For Small Vibrations Superimposed on the Basic Nonlinear State

Since the classical dissertation [17] by Koiter, in 1945, numerous studies have been performed to study the influence of initial imperfections on the buckling load of structures, whose load-carrying capacity is often reduced by tens of percent. These studies were extended to vibration problems by Rosen and Singer [18], Lipovski and Tokarenko [19], and Nash [20] for isotropic cylindrical shells, while stiffened shells were treated by Singer and Prucz [21]. An interesting conclusion of these papers was that the natural frequency is also substantially lowered or raised, due to initial imperfections.

Let us consider a simple model structure to demonstrate imperfection sensitivity in vibrations [22]. Consider a simple model structure suggested by Budiansky and Hutchinson [23]: a three-hinge, rigid rod-system constrained laterally by a nonlinear spring. We suppose that the restoring force F is related to the additional displacement by

$$F = k_1\xi + k_2\xi^2 + k_3\xi^3 \tag{42}$$

where $\xi = x/L$. Denoting the nondimensional initial imperfection by $\bar{\xi} = \bar{x}/L$, the the equilibrium dictates

$$\lambda(\xi+\bar{\xi}) = \frac{1}{2} F \sqrt{1-(\xi+\bar{\xi})^2}$$

$$= \frac{1}{2} (k_1\xi+k_2\xi^2+k_3\xi^3) \sqrt{1-(\xi+\bar{\xi})^2} \tag{43}$$

and for small values of ξ, the following result is obtained

$$\lambda(\xi+\bar{\xi}) = \lambda_c\,[\xi+a\xi^2+b\xi^3+\ldots] + O(\xi^2\bar{\xi}) \tag{44}$$

where $\lambda_c = k_1/2$ is the classical buckling load, which is found via the linear theory; moreover

$$a = k_2/k_1\ , \qquad b = k_3/k_1 - \tfrac{1}{2} \tag{45}$$

Structure with a=0, b=0 is designated as symmetric, that with a≠0, b=0 as asymmetric, and that with a≠0, b≠0 as nonsymmetric. For simplicity, let us consider the symmetric structures. General derivation for asymmetric or nonsymmetric structures can be found in Ref. 22. As was shown by Budiansky and Hutchinson [24] symmetric structure is imperfection-sensitive (in the sense that the presence of imperfections reduces the maximum load the structure can support)

both for $\bar{\xi} > 0$ and $\bar{\xi} < 0$, for b<0, and imperfection-insensitive for b>0. For an imperfection-sensitive structure with respect to stability the following formula holds, relating the static buckling load λ, to the initial

imperfection parameter $\bar{\xi}$:

$$\left[1 - \frac{\lambda_s}{\lambda_d}\right]^{3/2} - \frac{3\sqrt{3}}{2}\,|\bar{\xi}|\,\sqrt{-b}\,\frac{\lambda_s}{\lambda_d} = 0 \tag{46}$$

In what follows, formula analogous to Eq. (46) is derived for the natural frequency.

The static state described by Eq. (46) will be referred to as a basic state andindicated by a circumflex, namely

$$\lambda\,(\hat{\xi}+\bar{\xi}) = \lambda_{c1}\,(\hat{\xi}+b\hat{\xi}^3+\ldots) \tag{47}$$

The small vibrations are superimposed on the basic state, so that now

$$x(t) = \hat{\xi} + w(t) \tag{48}$$

The corresponding dynamic equation is obtained from Eq. 30 of Ref. 22 :

$$\left[\frac{ML}{k_1}\right] \frac{d^2x}{dt^2} + \left[1 - \frac{\lambda}{\lambda_{cl}}\right]\lambda + bx^3 = \frac{\lambda}{\lambda_{cl}} \xi \tag{49}$$

Substituting Eq. 48 into Eq. 49, omitting nonlinear terms with respect to w(t),we obtain

$$\frac{1}{\omega_1^2} \frac{d^2w}{dt^2} + \left(1 - \frac{1}{\lambda_{cl}}\right) w + 3b\hat{\xi}^2 w = 0 \tag{50}$$

where

$$\omega_1 = \left[\frac{k_1}{ML}\right]^{1/2} \tag{51}$$

is the natural frequency of the corresponding linear structural model. For freevibrations, we set

$$w(t) = We^{i\omega t} \tag{52}$$

which upon substituting into Eq. (50) yields

$$\left[1 - \frac{\lambda}{\lambda_{cl}} + 3b\hat{\xi}^2 - \frac{\omega^2}{\omega_1^2}\right] W = 0 \tag{53}$$

where ω is the natural frequency of the small vibrations around the nonlinear state $\hat{\epsilon}$. Eq. (53) must be solved in conjunction with Eq. (47).

The final formula for the nondimensional natural frequency

$$\Omega = \frac{\omega}{\omega_1} \tag{54}$$

related to initial imperfection $\bar{\xi}$ is

$$\left[1 - \frac{\lambda}{\lambda_{c1}} + \frac{\Omega^2}{2}\right]\left[\frac{1}{3b}(\Omega^2 - 1 + \frac{\lambda}{\lambda_{c1}})\right]^{\frac{1}{2}} = \frac{3}{2}\frac{\lambda}{\lambda_{c1}}|\bar{\xi}| \tag{55}$$

This equation leads to some remarkable conclusions: it implies that for the structure to be vibrationally imperfection-sensitive (in the sense that the vibration frequencies are reduced in the presence of imperfections), b must be negative, as in the case of static buckling. For b>0, the natural frequency increases with $\bar{\xi}$. For a perfect structure, the Eq. 55 reduces to

$$\Omega^2 = 1 - \frac{\lambda}{\lambda_{c1}} \tag{56}$$

i. e. the natural frequency squared varies linearly with the axial load.

In connection with this problem, the following questions arise:

1. Is it possible to generalize the results for this single-degree-of-freedom model to the continuous structures, such as beams, plates and shells?

2. In one-degree-of-freedom structure, the natural frequency of small vibrations superimposed upon the basic nonlinear state vanishes when the axial load λ tends to the buckling load of the structure λ_s (limit load). This opens the possibility of the non-destructive evaluation of the buckling load λ_s. In shells the vibration and buckling modes do generally not coincide. Is the non-destructive evaluation of the buckling load possible also for shells?

3. How to determine the higher eigenvalues of structures with small vibrations around the basic nonlinear state?

6. Nonprobabilistic Treatment of Uncertainty

In a number of recent mathematical studies, interval analysis was developed [25-27]. Interval of "real" numbers can be thought of as a new kind of number, represented by a pair of real numbers, namely its endpoints. An arithmetic-interval- for such numbers has been introduced [28]. When the endpoints of an interval coincide, we obtain a degenerate interval - when the real number equals to its coalescing endpoints. As a consequence, interval arithmetic reduces to the real arithmetic. The interval analysis is utilized usually for obtaining arbitrary narrow intervals containing exact real arithmetic results by carrying enough data.

The "fathers" of interval analysis have devised it in order computations in properly rounded interval arithmetic to produce a completely general mechanism for bounding the accumulation of roundoff error in any machine computation. If roundoff is the only error present, then the widths of the interval results will go to zero as the length of the machine word increases. However, there is another avenue to powerfully use interval mathematics: namely, to deal with uncertain phenomena. Usually, the theory of probability and random processes is applied to deal with uncertainties. One has to have a full knowledge of the probability distributions of the involved random variables, in order to derive those of the output quantities. However, the probabilistic information on the initial data is not always present. We will provide an illustrative an example, in which results of the probabilistic calculations may turn to be highly sensitive to imprecise data.

Consider a bar under central tensile force N, which is an uncertain variable. We will treat it as a random variable with the probability distribution $F_N(n)$. For simplicity, we will assume that N is an exponentially distributed random variable, with distribution

$$F_N(n) = \text{Prob}\,(N \leq n) = 1 - \exp\,[\,-\,n/E(N)] \tag{57}$$

We are interested in the reliability of the bar; in other words, in probability that the following random event holds

$$\sum \leq \sigma_y \tag{58}$$

where $\sum$ is a random stress and σ_y is a yield stress, which is treated here as a deterministic variable. Now

$$\sum = \frac{N}{g} \tag{59}$$

where g is an area of the cross-section of the bar. Reliability of the system, defined as the probability of the successful performance of the system, equals

$$\begin{aligned} R &= \text{Prob}\,(\textstyle\sum \leq \sigma_y) = \text{Prob}\,(N \leq \sigma_y g) \\ &= 1 - \exp\,[\,-\,\sigma_y g/E(N)] \end{aligned} \tag{60}$$

We visualize that the design problem is to be solved. The cross-sectional area should be determined so that the reliability is not lower than some specified, codified value

$$R \geq r$$

The minimum required cross-sectional area g^* is found then from the requirement

$$R = r \tag{61}$$

or

$$1 - \exp\,[\,-\,\sigma_y g^*/E(N)] = r \tag{62}$$

This requirement yields

$$g^* = \frac{E(N)}{\sigma_y} \ln \frac{1}{1-r} \tag{63}$$

In actuality, the data characterizing the distribution function, should be derived from the experiments unless some solid theoretical ground exists for their justification; hence both E(N) and σ_y have some errors, say α% of error is associated with E(N) and β% with σ_y. Then the actual cross-sectional area, which will be used in the actual design, is

$$g_{act} = \frac{E(N)}{\sigma_y} \frac{(1+\alpha/100)}{(1+\beta/100} \ln \frac{1}{1-r} \tag{64}$$

We may ask the question on what is the exact reliability the system is operating with. To obtain this result, we substitute Eq. (64) into Eq. (60). The result is actual reliability

$$\begin{aligned} R_{act} &= 1 - \exp [- \sigma_y g_{act}/E(N)] \\ &= 1 - \exp \left[- \frac{100+\alpha}{100+\beta} \ln \frac{1}{1-r} \right] \end{aligned} \tag{65}$$

The final expression for the actual probability of failure

$$P_{f,act} = 1 - R_{act} \tag{66}$$

reads

$$P_{f,act} = \left[P_{f,all} \right]^{(100+\alpha)/(100+\beta)} \tag{67}$$

where $P_{f,all}$ is an "allowed" probability of failure

$$P_{f,all} = 1-r. \tag{68}$$

One can deduct from formula (67) that asymptotically

$$\ln P_{f,act} = \ln P_{f,all} \left[1 + \frac{\alpha}{100} - \frac{\beta}{100}\right] \tag{69}$$

and for negative α and positive β the actual probability of failure may reach values, which are much larger than the allowed probabilities of failure.

This example shows vividly that one may "accept" the structure for its actual use, based on the requirement that the probability of failure will be not higher than its "allowed" level, but in actuality due to errors in data, the system will be in the "adverse" or "failed" state.

The interval mathematics may become the possible alternative to classical probabilistic methods, in these circumstances.

In the following example, we can visualize that instead of the probability distribution of N, we know just the interval for the applied load

$$N = [\underline{N}, \overline{N}] \tag{70}$$

as well as for the yield stress σ_y

$$\sigma_y = [\underline{\sigma}_y, \overline{\sigma}_y] \cdot \tag{71}$$

Then, if zero is not contained in the above interval

$$\frac{1}{\sigma_y} = \left[\frac{1}{\bar{\sigma}_y}, \frac{1}{\underline{\sigma}_y}\right] \tag{72}$$

and

$$g = \left[\underline{N} \cdot \frac{1}{\bar{\sigma}_y}, \bar{N} \cdot \frac{1}{\underline{\sigma}_y}\right] \tag{73}$$

We are interested in the variation of the eigenvalues if the system's mass distribution and the stiffness distribution are uncertain; namely, the problem is formulated as follows: find

$$\omega_j = [\underline{\omega}_j, \bar{\omega}_j] \tag{74}$$

if the elements of the mass matrix and the elements of the stiffness matrix belong to some specified intervals

$$m_{ij} = [\underline{m}_{ij}, \bar{m}_{ij}] \tag{75}$$

$$k_{ij} = [\underline{k}_{ij}, \bar{k}_{ij}] \tag{76}$$

for the N-degree of freedom system, governes by the equations

$$[M]\{\ddot{X}\} + [K]\{X\} = \{0\} \tag{77}$$

where

$$\{X\} = \{A_j\} \sin\omega_j t \tag{78}$$

M is the mass matrix, K stiffness matrix, $\{A_j\}$jth normal mode.

More generally, we are interested in the problem: find the response

$$X(t) = [\underline{X}(t), \bar{X}(t)] \tag{79}$$

for the system governed by the equation

$$[M]\{\ddot{X}\} + [K]\{X\} + [C]\{\dot{X}\} = \{f(t)\} \tag{80}$$

where [M] and [K] are interval matrices, and moreover

$$c_{ij} = \alpha m_{ij} + \beta k_{ij} \tag{81}$$

α and β are proportionality coeffiicents, and

$$f_j(t) = [\underline{f}_j(t), \bar{f}_j(t)] \tag{82}$$

Present writer is most interested in cooperating with the applied mathematicians in solution of these applied mechanics problems, dealing with uncertainty. This new interpretation of the goals of the interval analysis to deal with uncertainty, rather than with carrying out in rounder interval arithmetic, appears to open ample possibilities for the solution of many important boundary value problems. As Richtmyer mentions [29], "although interval analysis is in a sense just a new language for inequalities it is a very powerful language and is one that has direct applicability to the important problem of significance on large computations". We are confident that the interval analysis has a considerable potential to be used in dealing with uncertain phenomena.

7. Convex Models of Uncertainty

Independently of these developments, a number of mechanicists approached the problem of uncertain phenomena non-probabilistically. Apparently, Drenick in 1968 was the first to realize this possibility [30-32]. This work was generalized (see Ref. 33) by Shinozuka. Recently, Shinozuka [31] presented upper and lower bounds on the response variability. These results [30-34] provide important physical, as well as numerical, insight into both the earthquake response and response variability issues. In Ref. 35 new, convex models of uncertainty have been developed, for applied mechanics applications in a quite general context. This monograph utilizes the imperfect, scanty knowledge on uncertain quantities, instead of precise information on the probability contents of random events. This study uses a representation of uncertain phenomena by convex sets. The approach itself is referred to a convex modelling.

Consider, for example, dynamics and failure of isotropic, unstiffened cylindrical shells with uncertain imperfections. The differential equation governing the axisymmetric motion of this cylindrical shell reads

$$D \frac{\partial^4 w}{\partial x^4} + N \frac{\partial^2 w}{\partial x^2} + \rho h \frac{\partial^2 w}{\partial t^2} + \frac{Eh}{R^2} w = N \frac{\partial^2 w_0}{\partial x^2} \tag{83}$$

where $w_0(x)$ is the initial imperfection function, $w(x,t)$ is the additional shell displacement, x is the axial coordinate, t - time, $D = Eh^3/12(1-\nu^2)$ - the flexural stiffness, E is - Young's modulus, h - shell thickness, R - shell radius, ν - Poisson's ratio, N - axial loading, ρ - shell material density. The shell is simply supported at its ends. The boundary conditions are

$$w(x,t) = \frac{\partial^2 w}{\partial x^2} = 0, \quad \text{at } x = 0, \; x = L; \tag{84}$$

initial conditions are

$$w(x,t) = \frac{\partial w}{\partial t} = 0 \quad \text{at} \quad t = 0 \tag{85}$$

For the simply supported shell we expand the initial imperfection profile in a Fouries series

$$w_0(x) = \sum_{i=1}^{\infty} A \sin \frac{i\pi x}{L} \tag{86}$$

Similarly, we expand the additional displacement of the shell in a series as

$$w(x,t) = \sum_{i=1}^{\infty} G(t) \sin \frac{i\pi x}{L} \tag{87}$$

Solution of Eq. 83, in view of Eqs. (86) and (87) reads:

$$G_i(t) = A_i \psi_i(t) \tag{88}$$

where $\psi_i(t)$ satisfies and equation

$$\rho h \frac{d^2\psi_i}{dt^2} + \left[D \left(\frac{i\pi}{L}\right)^4 - N \left(\frac{i\pi}{L}\right)^2 + \frac{Eh}{R^2} \right] \psi_i = - N \left(\frac{i\pi}{L}\right)^2 \tag{89}$$

Let us assume that we have only limited information for charactering the initial imperfections. In particular, the only information is that the dominant N initial imperfection Fourier coefficients on Eq. (86) fall within ellipsodal set

$$Z(\Omega,\theta) = \{A^T = (A_1, A_2, \ldots, A_N): \; A^T \Omega A \leq \theta^2\} \tag{90}$$

where Ω is a positive definite symmetric matrix, θ^2 is a positive constant and N is the number of dominant Fourier coefficients.

The total displacement is

$$v(x,t) = w(x,t) + w_0(x) = \sum_{i=1} A_i \left[1+\psi_i(t)\right] \sin \frac{i\pi x}{L} \tag{91}$$

and is rewritten as follows

$$\delta(x,t) = \delta(x,t)^T A \tag{92}$$

where δ is an N-vector whose i-th element is

$$\delta_i(x,t) = [1 + \psi_i(t)] \sin \frac{i\pi x}{L} \tag{93}$$

The problem is formulated as follows: given an imperfection ellopsoid of the initial imperfections, find the initial imperfection vector which maximizes the total displacement. We will denote this maximizing vector A_m. This maximum is represented as

$$\hat{v}(x,t) = \max_{A \in Z(\Omega,\theta)} v(x,t) \tag{94}$$

The set of extreme points of $Z(\Omega,\theta)$ is the ellipsoidal shell

$$C(\Omega,\theta) = \{A: A^T \Omega A = \theta^2\} \tag{95}$$

The set $Z(\Omega,\theta)$ is a convex set and v(x,t) is a linear function of A. Hence the maximum deflection will be reached on the extreme points of $Z(\Omega,\theta)$, i.e. on the ellipsoidal shell $C(\Omega,\theta)$; or

$$\hat{v}(x,t) = \max_{A \in C(\Omega,\theta)} v(x,t) \tag{96}$$

A closed form expression for $\hat{v}(x,t)$ is obtained by the method of Lagrange multipliers [36]:

$$A_w = \frac{}{\sqrt{\delta^T \Omega^{-1} \delta}} \Omega^{-1}\delta \tag{97}$$

$$\hat{v}(x,t) = \delta^T A_w = \theta \sqrt{\delta(x,t)^T \Omega^{-1}\, \delta(x,t)} \tag{98}$$

It is remarkable that A_w is a function of vector δ (x,t), which means that A_w depends on both time t and space coordinate x. This implies that different initial imperfections will maximize the total displacement at the different x and t. For numerical examples, the reader should consult Ref. 35.

In the connection with convex modelling the following questions are posed:

1. What are other acceptable convex models from an engineering point of view?

2. How to obtain a general tool for the prediction of the worst response when the uncertain phenomenon is described by an intersection of a number of convex sets?

3. How to combine probabilistic and convex models of uncertainty?

8. Best and Worst Lyapunov Exponents

Convex modelling could be successfully applied to treatment of uncertain parameters of the system.

In this section we will consider a viscoelastic plate, subjected to harmonic inplane excitation. The governing differential equation reads:

$$\nabla^4 w\,[D(t)f(0) + \int_{0^+}^{t} D(t-\tau)\, f(\tau)\, d\tau] \tag{99}$$

$$+ (N_{x,s} + N_{x,d}\cos\beta t)\, w_{,xx} + (N_{y,s} + N_{y,d}\cos\beta t)\, w_{,yy}\, f(t)$$

$$+ \rho h \ddot{f}(t) = 0$$

For the notations see Ref. 37).

We are interested in the stability of the unperturbed equilibrium of the viscoelastic plate. For the treatment of ordinary differential equations with time dependent coefficients, Lyapunov introduced the concept of characteristic numbers, the sign of which determines whether the unperturbed motion is stable (see, e.g. Hahn [38]). The negative values of these characteristic numbers are presently referred to as Lyapunov exponents. According to Lyapunov, if all the exponents are negative, the unperturbed motion is stable. In addition, Chetaev [39] proved that if one of the Lyapunov exponents is positive, then the unperturbed motion is unstable. Lyapunov exponents recently became a powerful tool in the study of chaotic motion [40]. From above it follows that it suffices to compute the largest Lyapunov exponent for the determination of the stability of the unperturbed motion of the viscoelastic plate in question. In Ref. 29 the stability of the Voigt-Kelvih model as well as the standard linear solid model was considered.

Let us concentrate now on the uncertain material parameter case. Assume that the relaxation function in the unidirectional case is given by

$$E(t) = a + b\exp(-ct) \tag{100}$$

where a, b and c are uncertain parameters, let z be a vector whose components a, b and c are the uncertain parameters. Furthermore, let N(z) represent the buckling load for a plate whose uncertain parameters are represented by the vector z. Let z_0 be a nominal uncertain vector. For example, z_0 may correspond to the average values of the parameters a_0, b_0 and c_0. The dynamic instability loads for material parameters $a_0 + \alpha$, $b_0 + \beta$, $c_0 + \gamma$, to the first order of vector ζ, whose elements are α, β and γ is:

$$N(z_0+\zeta) = N(z_0) + \sum_{i=1}^{3} \frac{\partial N}{\partial \zeta_i} \zeta_i \tag{101}$$

We shall evaluate the lower limit of the buckling load as ζ varies on the ellipsoidal set

$$Z(\theta,\omega) = \left[\frac{\alpha}{\omega_1}\right]^2 + \left[\frac{\beta}{\omega_2}\right]^2 + \left[\frac{\gamma}{\omega_3}\right]^2 \leq \theta^2 \tag{102}$$

where the site parameter θ and the oemiaxes ω_1, ω_2 and ω_3 are based on experimental data. Thus $Z(\alpha,\omega)$ can be chosen to represent a realistic ensemble of plates. The lowest dynamic buckling load which can be obtained from any of the plates in this ensemble is

$$\mu(\theta,\omega) = \min_{\zeta \epsilon Z} [N(z_0) + \phi^T \zeta] \tag{103}$$

where $\mu(\theta,\omega)$ is the buckling load of the weakest plate in the ensemble Z. The extreme value will occur, due to convexity of Z, on the set of extreme points of Z, which is colletion of vectors $\xi = (\xi_1,\xi_2,\xi_3)$ in the set

$$C(\theta,\omega) = \xi: \sum_{i=1}^{3} \frac{\xi_i^2}{\omega_i} = \theta^2 \tag{104}$$

Thus the minimum dynamic buckling load becomes

$$\mu(\theta,\omega) = \min_{\xi \epsilon C} [N(z_0) + \phi^T \zeta] \tag{105}$$

The solution is given as follows:

$$\mu(\theta,\omega) = N(z_0) - \theta\sqrt{\phi^T \Omega^{-1} \phi} \tag{106}$$

and is similar to the solution of other nonlinear boundary-value problems ([ ]).

The mathematical questions which are posed in this context are:

1. How to find natural frequencies of the nearly circular plates with smooth irregularities satisfying the inequalities

$$| w_0(\theta) | \leq \delta \tag{107}$$

or

$$| w_0(\theta) | \leq \delta_1 \quad , \quad | w'(\theta) | \leq \delta_2$$

where the boundary in the polar coordinates reads

$$r^{\epsilon}(\theta) = R + \epsilon\, w_0(\theta) \quad ? \tag{108}$$

2. How to solve the above problem for

$$w_0(\theta) = \sum_{i=1}^{I} (A_m \cos m\theta + B_m \sin m\theta) \tag{109}$$

when, in addition the following restriction is imposed

$$\int_0^{2\pi} [w_0(\theta)]^2 \, d\theta \leq \delta_3 \quad ? \tag{110}$$

In Eq.109, "I" denotes the number of harmonics describing the irregularity.

9. Conclusion

A brief review of some questions of mathematical nature, arising in applied mechanics is given. Some pertinent comments could be made on other various nagging problems, but these are left for future occassions.

10. Acknowledgement

The research reported in this paper was supported by the National Science Foundation MSM-9015371. Any opinions, findings, and conclusions or recommendations expressed by this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Elishakoff, I., (1983) Probabilistic methods in the theory of structures, Wiley-Interscience, New York, pp. 361-379.

2. Bolotin, V. V., (1969) Statistical methods in structural mechanics, Holden-Day, San Francisco, p. 122.

3. Lubliner, E. and Elishakoff, I., (1986) Random viabration of system with finitely many degrees of freedom and several coalescent natural frequencies, International Journal of Engineering Sciences, 24 (4), pp. 461-470.

4. Elishakoff, I., A. Th. van Zanten and S. H. Crandall, (1979) Wide-band random axisymmetric vibration of cylindrical shells, Journal of Applied Mechanics, 96, pp. 417-422.

5. Elishakoff, I. (1974) Vibration analysis of clamped square orthotropic plate, American Institute of Aeronautics and Astronautics Journal, 12, pp. 921-924.

6. Gorman, D. J. (1982) Free vibration analysis of rectangular plates, Elsevier, New York.

7. Rayleigh, J.W.S., The theory of sound, London, Macmillan and Co.

8. Timoshenko, S.P., (1921) On the correction factor for shear of the differential equation for transverse vibrations of prismatic bar, Philosophical Magazine, 6, (41/295), pp. 799-796.

9. Weaver, W.Jr., Timoshenko, S.P. and Young, D.H., (1990) Vibration problems in engineering, John Wiley and Sons, New York, pp. 433-436.

10. Clough, R.W. and Penzien, J., (1975) Dynamics of structures, McGraw Hill, Auckland, pp. 318-320.

11. Samuels, J.C. and Eringen, A.C., (1956) Response of a simply supported Timoshenko beam to a purely random Gaussian process, Journal of Applied Mechanics, 25, pp. 496-500.

12. Elishakoff, I. and Lubliner, E., (1985) Random vibration of a structure via classical and nonclassical theories, in S. Eggwertz and N.C. Lind, "Probabilistic methods in the mechanics of solids and structures," Springer, Berlin, pp. 455-468.

13. Cowper, G.R., (1966) The shear coefficient in Timoshenko's beam theory, Journal of Applied Mechanics, 35, pp. 335-340.

14. Mindlin, R. D., (1951) Influence of rotary inertia and shear on flexural motions of isotropic plates, Journal of Applied Mechanics, 18, pp.31-38.

15. Cederbaum, G., Librescu, L. and Elishakoff, I., (1989) Remarks on a dynamical higher-order theory of laminated plates and its application in random vibration response, International Journal of Solids and Structures, 25, pp. 515-526.

16. Private communication, Oberwolfach, 1990.

17. Koiter, W.T., (1967) (1970) On the stability of elastic equilibrium, Ph.D. thesis, Delft University of Technology, H. J. Paris, Amsterdam (in Dutch); English translations: (a) NASA TTF - 10, 833; (b) AFFDL-TR-70-25.

18. Rosen, A. and Singer, J. (1974) Effect of axisymmetric imperfections on the vibrations of cylindrical shells under axial compression, American Institute of Aeronautics and Astronautics Journal, 17, pp. 995-997.

19. Lipovski, D.E. and Tokarenko, V.M. (1966) Effect of initial imperfections on free oscillation frequencies of cylindrical shells, 6th All-Union Conference on the Theory of Shells and Plates, "Nauka" Pub., pp. 547-547 (in Russian).

20. Nash, W.A, (1981) Free vibrations of initially imperfect cylindrical shells Proceedings of the Eighth Canadian Congress of Applied Mechanics, Moncton, pp. 365-366.

21. Singer, J. and Prucz, J., (1982) Influence of initial geometrical imperfections on vibrations of axially compressed stiffened cylindrical shells, Journal of Sound and Vibration, 80, pp. 117-143.

22. Elishakoff, I., Birman, V. and Singer, J., (1984) Effect of imperfections on the vibrations of loaded structures, Journal of Applied Mechanics, 51, pp. 191-193.

23. Budiansky, B. and Hutchinson, J.W., (1964) Dynamic Buckling of imperfection-sensitive structures, in "Proceedings of the Eleventh International Congress of Applied Mechanics," Goertler, M., ed., Munich, pp. 636-651.

24. Budiansky, B., (1967) Dynamic buckling of elastic structures: criteria and estimates, in "Dynamic stability of structures," G. Herrmann, ed., Pergamon, pp. 83-106.

25. Moore, R.E. (1979) Methods and applications of interval analysis, SIAM, Philadelphia.

26. Alefeld, G. and Herzberger, J., (1983) Introduction to interval computations, Academic Press, New York.

27. Kulisch, U. and Miranker, W., (1981) A new approach to scientific computation, Academic Press, New York.

28. Moore, R.E., (1966) Interval analysis, Prentice Hall, Englewood Cliffs, NJ.

29. Richtmeyer, R.D., (1968) Math. Comput., 22, p. 221.

30. Drenick, R.F., (1968) Functional analysis of effects of earthquakes, 2nd joint US-Japan Seminar on Applied Stochastics, Washington, DC, (Sept. 19-29).

31. Drenick, R.F., (1970), Model-free design of aseismic structures, Journal of Engineering Mechanics Division, 96, pp. 483-493.

32. Drenick, R.F., (1977), On a class of non-robust problems in stochastic dynamics, in B.L. Clarkson, ed., Stochastic Problems in Dynamics, Pitman, London, pp. 237-255.

33. Shinozuka, M., (1970), Maximum structural response to seismic excitations, Journal of the Engineering Mechanics Division, 96, pp. 729-738.

34. Shinozuka, M., (1987), Structural response variability, Journal of the Engineering Mechanics Division, 113, pp. 825-842.

35. Ben-Haim, Y. and Elishakoff, I., (1990) Convex models of uncertainty in applied mechanics, Elsevier, Amsterdam.

36. Elishakoff, I. and Ben-Haim, Y., (1990) Dynamics of a thin cylindrical shell under impact with limited deterministic information on its initial imperfections, in Casciati, F., Elishakoff, I. and Roberts, J.B. (1990) Nonlinear mechanical systems under stochastic conditions, Elsevier, Amsterdam.

37. Aboudi, J., Cederbaum, G. and Elishakoff, I., (1990) Dynamic stability analysis of viscoelastic plates by Lyapunov exponents, Journal of Sound and Vibration, 139, pp. 459-468.

38. Hahn, W., (1967) Stability of motion, Springer, Berlin., pp. 308-309.

39. Chetaev, N.G., (1960) On certain questions related to the problem of stability of unsteady motion, Prikladnaya Matematika i Mekhanika, 24 (1), pp. 5-22.

40. Moon, F.C., (1987) Chaotic Vibration, Wiley, New York.

Professor Isaac Elishakoff, Center for Applied Stochastics Research and the Department of Mechanical Engineering, Boca Raton, FL 33431-0991, (U.S.A.)

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

COMPUTING THE MINIMUM EIGENVALUE OF A SYMMETRIC POSITIVE DEFINITE TOEPLITZ MATRIX WITH SPECTRAL TRANSFORMATION LANCZOS METHODS

Thomas Huckle
Institut für Angewandte Mathematik und Statistik
Universität Würzburg
D-8700 Würzburg, Federal Republic of Germany

1. Introduction

A matrix T is Toeplitz if the elements on each diagonal are all equal. Thus, for the real symmetric case T is of the form

$$T = T(t_0, t_1, ..., t_{n-1}) := \begin{pmatrix} t_0 & t_1 & \cdots & \cdots & t_{n-1} \\ t_1 & t_0 & t_1 & & \vdots \\ \vdots & t_1 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & t_1 \\ t_{n-1} & \cdots & \cdots & t_1 & t_0 \end{pmatrix}.$$

There is considerable interest in the problem of finding the minimal eigenvalue of a symmetric positive definite Toeplitz matrix T [16]. Cybenko and van Loan have presented an algorithm for computing $\lambda_{min}(T)$ based on the Levinson-Durbin algorithm and a combination of bisection and Newton's method [6]. The aim of this paper is to display the behaviour of Lanczos type methods for computing $\lambda_{min}(T)$ using good estimates for the minimal eigenvalue and its corresponding eigenvector.

Toeplitz systems of linear equations arise from many sources (see [2] and the literature quoted therein). There are several algorithms for obtaining the solution in $O(n^2)$ operations, the so called fast Toeplitz solvers [13,19]. More recently, superfast Toeplitz solvers have been presented for Toeplitz systems of equations using $O(n \log^2(n))$ arithmetic operations (see e.g. [10,1]). Applied to general Toeplitz systems, current fast and superfast Toeplitz solvers are numerically unstable. For some of the superfast Toeplitz solvers stability can be expected for positive definite matrices [1,2].

Circulant matrices $C = T(c_0, c_1, ..., c_m, ..., c_2, c_1)$ form a special subclass of Toeplitz

matrices. For circulant matrices a product $C*b$ can be evaluated by the Fast Fourier Transform in $O(n\log(n))$ operations [7], and therefore for a Toeplitz matrix $T = T(t_0,...,t_{n-1})$ the product $T*b$ can also be computed in $O(n\log(n))$ operations by considering $C := T(t_0,t_1,...,t_{n-1},...,t_1)$. For the Frobenius norm, the circulant matrix C_f, defined by $\|C_f - T\|_F$ is minimal over all circulant matrices, turns out to be a good approximation to the Toeplitz matrix T [5]. If C_f is positive definite it can be used as a preconditioner in the preconditioned conjugate gradient method for solving $Tx = b$ [18,3,4]. Furthermore, the eigenvectors of C_f are good approximations to the eigenvectors of T. In a similar way, we get lower and upper bounds for $\lambda_{min}(T)$, considering circulant approximations to T [11]. In section 2 we formulate algorithms for computing $\lambda_{min}(T)$ by Lanczos methods, using the above approximations as start values; numerical results are given in section 3.

2. Lanczos methods for computing $\lambda_{min}(T)$

For the following denote by ρ the lower bound for $\lambda_{min}(T)$ and by x_1 the eigenvector approximation for the corresponding eigenvector of T, derived by the circulant approximations to T. In [18] Strang proposed the Rayleigh quotient iteration for computing eigenvalues of a Toeplitz matrix T, using x_1 as start vector. The generated sequence of eigenvalue approximations converges very fast, in some cases however not to $\lambda_{min}(T)$, but to any other eigenvalue of T [14]. In the course of the algorithm, there have to be solved at least 3 indefinite Toeplitz linear equations, which may be unstable. If we use the inverse iteration method with start vector x_1 and the iteration matrix $(T-\rho I)^{-1}$, we have to compute only one inverse of a positive definite Toeplitz matrix in $O(n\log^2(n))$; in order to solve $(T-\rho I)*x_{i+1} = x_i$ in every step we have to carry out only products between Toeplitz matrices and vectors in $O(n\log(n))$. But depending on the shift ρ and the spectrum of T the convergence of the inverse iteration method may be very slow [14].

Now let us consider the Lanczos method [12] for computing the minimum eigenvalue of a symmetric matrix A. Define by $K_p(A,x_1) := \text{span}\{x_1, Ax_1, ..., A^{p-1}x_1\}$ the Krylov subspace of order p to A and x_1, with QR-decomposition $(\,x_1 \quad Ax_1 \quad \cdots \quad A^{p-1}x_1\,) = Q_pR$. Then, the orthogonal projection of A on $K_p(A,x_1)$ has the form

$$A_p = Q_p^T A Q_p = \begin{pmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{p-1} \\ 0 & & \beta_{p-1} & \alpha_p \end{pmatrix}. \tag{1}$$

The eigenvalues and corresponding eigenvectors of A_p give approximations to eigen-

values and eigenvectors of A. Let μ_{min} be the minimal eigenvalue of A_p with corresponding eigenvector x_{min}; then the Ritz pair $(\mu_{min}, Q_p x_{min})$ converges to the minimum eigenvalue and corresponding eigenvector of A. Moreover, we get the inequality (see e.g. [9,14])

$$\min_{\mu \in spectrum(A)} |\mu_{min} - \mu| \leq |\beta_p x_{min,p}| =: \delta_p. \tag{2}$$

Thus, the circle with center μ_{min} and radius δ_p contains at least one eigenvalue of A, but not in every case the minimal eigenvalue. This depends on the angle between the start vector and the eigenvector to $\mu_{min}(A)$ [15]. In the same way the Lanczos algorithm can be used to give estimates to $\lambda_{max}(A)$. Applied to a Toeplitz matrix T and start vector x_1, we have to compute only products between Toeplitz matrices and vectors, each taking $O(n\log(n))$ operations, but for some examples the convergence is too slow.

In order to get faster convergence we apply the Lanczos algorithm not on T itself but on $(T - \rho I)^{-1}$ according to the Spectral Transformation Lanczos method introduced by Ericsson and Ruhe [8]. But again, for large $\rho - \lambda_{min}$ the convergence can get very slow. For the following, let us denote the eigenvalues of T by $\lambda_1 < \lambda_2 < ... < \lambda_k$. Then, the convergence of the Spectral Lanczos Transformation method depends on the magnitude of $(\rho - \lambda_1)/(\lambda_1 - \lambda_2)$ [17]. Hence, in the case of poor convergence, we have to improve the shift ρ. To get a better shift we can use (2) and define $\rho_{new} = \rho + 1/(\mu^{(p)} + \delta_p)$, whereby $\mu^{(p)}$ is the eigenvalue estimate given by the Spectral Lanczos Transformation after the p-th step.

To give reasons for the choice of this shift let us first analyse the first Lanczos steps for computing the maximum eigenvalue of a diagonal matrix $D = \mathrm{diag}(d_1, ..., d_n)$ with start vector $q_1 = (y_1, ..., y_n)^T$ of Euclidean norm 1 and $d_1 \leq d_2 \leq ... \leq d_{n-1} < d_n$. With the notation

$$s_i := \sum_{j=1}^{n-1} d_j^i y_j^2/(1 - y_n^2) \quad , i = 1, 2, 3$$

we get

$$\mu^{(1)} = \alpha_1 = d_n(y_n^2 + (1 - y_n^2)\frac{s_1}{d_n}),$$

$$\delta_1^2 = \beta_1^2 = d_n^2 y_n^2 (1 - y_n^2)(1 - 2\frac{s_1}{d_n} - \frac{1 - y_n^2}{y_n^2}\frac{s_1^2}{d_n^2} + \frac{1}{y_n^2}\frac{s_2}{d_n^2}),$$

$$\alpha_2 = d_n(1 - y_n^2 + y_n^2\frac{s_1}{d_n} + \frac{1 + y_n^2}{y_n^2}\frac{s_1^2 - s_2}{d_n^2} + O((\frac{d_{n-1}}{d_n})^3)),$$

$$\mu^{(2)} = d_n(1 + \frac{1 - y_n^2}{y_n^2}\frac{s_1^2 - s_2}{d_n^2} + O((\frac{d_{n-1}}{d_n})^3)),$$

$$x^{(2)} = (0,...,0,1)^T + O((\frac{d_{n-1}}{d_n})),$$

$$\beta_2^2 = \frac{d_n^2}{y_n^2}(\frac{s_2 - s_1^2}{d_n^2} + O((\frac{d_{n-1}}{d_n})^3)),$$

$$\delta_2^2 = d_n^2(\frac{1 - y_n^2}{y_n^2}\frac{s_2 - s_1^2}{d_n^2} + O((\frac{d_{n-1}}{d_n})^3)).$$

Thereby, $\mu^{(p)}$ and $x^{(p)}$ denote the eigenvalue and eigenvector approximations given by the Lanczos method in the p-th step, δ_p is the radius of the circle (2), and α_i, β_i as in (1).

The formulas above show, that for a good start vector x_1 it holds $0 \leq \mu^{(1)} - d_1 \leq \delta_1$ for the Lanczos method and $\rho \leq \rho + 1/(\mu^{(1)} + \delta_1) \leq d_1$ for the Spectral Transformation Lanczos method. Hence, if we use the Lanczos method and choose $\mu^{(p)} - \delta_p$ as new shift, or if we use the Spectral Lanczos method and choose $\rho + 1/(\mu^{(p)} + \delta_p)$ as new shift, then we can expect this new shift to be a good lower bound for the minimal eigenvalue of T.

Thus we are led to the following algorithms:
Iterated Spectral Transformation Lanczos method (ISTL): Apply a few Lanczos steps to $(T - \rho I)^{(-1)}$ to get a new shift ρ and start vector $x^{(p)}$; repeat this until a stopping criterion is fulfilled. If we always use m Lanczos steps, then the analysis above shows that the generated eigenvalue estimates converge of order m. In order to avoid costly inversions and nearly singular linear systems the Iterated Spectral Transformation Lanczos method should only restart once.
Lanczos with Spectral Transformation Lanczos method (L-STL): Apply a few Lanczos steps on T in order to get a shift ρ and start vector $x^{(p)}$; then start the Spectral Lanczos Transformation method.

Nevertheless, the new shift can get larger than λ_1. Then $T - \rho I$ is indefinite, and in the computation of the inverse of this matrix we can read of the inertia of $T - \rho I$ [1]. Thus, we can either keep the shift and use the information about the inertia of $T - \rho I$ to compute estimates for λ_1 or we choose a new shift; but in practice this doesn't occur.

3. Numerical results

For testing the eigenvalue algorithms described above we consider the following examples:

1. $t_i = 1/(i+1)^2$ for $i = 0,...,19$, cf. [3,5];
2. $t_0 = 1$ and $t_i = random(-.2,.2)$ for $i = 1,...,19$;
3. $t_i = \cos(i)/(i+1)$ for $i = 0,...,14$, cf [3,5];

4. $t_0 = 6.2$, $t_i = random(-1,1)$ for $i = 1,...,10$ and $t_i = 0$ for $i = 11,...,19$;
5. $T = \sum_{k=1}^{40} w_k T_{2\pi\Theta_k}$ a 40×40 matrix with $(T_\Theta)_{i,j} = \cos(\Theta(i-j))$ for $i,j = 1,...,40$ and w_k and Θ_k are uniformly distributed random numbers taken from $[0,1]$, cf. [6];
6. $T = \sum_{k=1}^{80} w_k T_{2\pi\Theta_k}$ a 40×40 matrix with $(T_\Theta)_{i,j} = \cos(\Theta(i-j))$ for $i,j = 1,...,40$ and w_k and Θ_k are uniformly distributed random numbers taken from $[0,1]$, cf. [6];
7. $t_i = .8^{i^2}$ for $i = 0,...,39$, cf. [1].

The stopping criterion in the Lanczos method was $\delta_p < 10^{-8} * \mu^{(p)}$ with δ_p from (2) and $\mu^{(p)}$ the eigenvalue estimate; in the Spectral Lanczos Transformation method the stopping criterion was $\delta_p < 10^{-8} * \mu^{(p)}$ with $\mu^{(p)}$ the estimate for the largest eigenvalue of $(T - \rho I)^{-1}$ (see [8]).

Nr.	Lanczos	STL	ISTL	L-STL
1		4	2+2	2+4
2		8	3+3	3+5
3		4	2+2	2+4
4		6	2+3	2+5
5	$> n$	15	3+6	3+8
6	$> n$	15	2+5	2+8
7	$> n$	36	3+8	4+8

Table 1. Iteration numbers for the considered eigenvalue algorithms

For very good start values, table 1 shows that the Spectral Transformation Lanczos method converges fast; in the case of not so good start values one should use the ISTL method or the L-STL method. The L-STL method needs only one inversion and thus seems to be preferable.

4. Conclusions

We have presented algorithms for estimating the minimum eigenvalue of a symmetric positive definite Toeplitz matrix, which need only one or two inverses of positive definite Toeplitz matrices. The methods are based on good initial approximations to the corresponding eigenvector which can be derived by approximating the Toeplitz matrix by circulant matrices. The algorithms seem also to be applicable to find eigenvalues of symmetric matrices A for which the $L^T D L$ decomposition can be computed very fast, e.g. band matrices.

References

[1] Ammar,G.S.,Gragg,W.B. (1989) Numerical experience with a superfast real Toeplitz solver. Linear Algebra and Appl. 121, 185-206

[2] Bunch,J.R. (1985) Stability of methods for solving Toeplitz systems of equations. SIAM J. Sci. Stat. Comput. 6 (2), 349-364

[3] Chan,R.H. (1989) Circulant preconditioners for Hermitian Toeplitz systems. SIAM J. Matrix Anal. Appl. 10 (4), 542-550

[4] Chan,R.H.,Strang,G. (1989) Toeplitz equations by conjugate gradients with circulant preconditioner. SIAM J. Sci. Stat. Comput. 10 (1), 104-119

[5] Chan,T.F. (1988) An optimal circulant preconditioner for Toeplitz systems. SIAM J. Sci. Stat. Comput. 9 (4), 766-771

[6] Cybenko,G.,Van Loan,C. (1986) Computing the minimum eigenvalue of a symmetric positive definite Toeplitz matrix. SIAM J. Sci. Stat. Comput. 7 (1), 123-131

[7] Davis,P.J. (1979) Circulant matrices (Wiley, New York, N.Y.).

[8] Ericsson,T.,Ruhe,A. (1980) The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems. Math. Comp. 35, 1251-1268

[9] Golub,G.H.,Van Loan,C. (1983) Matrix computations (Johns Hopkins Univ. Press, Baltimore).

[10] de Hoog, F. (1987) A new algorithm for solving Toeplitz systems of equations. Linear Algebra and Appl. 88/89, 123-138

[11] Huckle,T. (1990) Circulant and skewcirculant matrices for solving Toeplitz matrix problems. To appear in: SIAM J. Sci. Stat. Comput.

[12] Lanczos,C. (1950) An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. J. Res. Nat. Bur. Standards 45, 255-282

[13] Levinson,N. (1947) The Wiener RMS error criterion in filter design and prediction. J. Math. Phys. 25, 261-278

[14] Parlett,B.N. (1980) The symmetric eigenvalue problem (Prentice-Hall, Englewood Cliffs, NJ)

[15] Parlett,B.N.,Simon,H.,Stringer,L.M. (1982) On estimating the largest eigenvalue with the Lanczos algorithm. Math. Comp. 38, 153-165

[16] Pisarenko,V.P. (1973) The retrieval of harmonics from a covariance function. Geophys. J. R. Astr. Soc. 33, 347-366

[17] Saad,Y. (1980) On the rates of convergence of the Lanczos and the block-Lanczos methods. SIAM J. Numer. Anal. 17 (5), 687-706

[18] Strang,G. (1986) A proposal for Toeplitz matrix calculations. Stud. Appl. Math. 74, 171-176

[19] Trench,W.F. (1964) An algorithm for the inversion of finite Toeplitz matrices. J. SIAM 12 (3), 515-522

Dr. Thomas Huckle, Institut für Angewandte Mathematik und Statistik, Universität Würzburg, Am Hubland, D-8700 Würzburg, Federal Republic of Germany.

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

NUMERICAL TREATMENT OF NONSELFADJOINT PLATE VIBRATION PROBLEMS*

Peter Paul Klein

Rechenzentrum der TU Clausthal, Clausthal-Zellerfeld,
Bundesrepublik Deutschland

A class of eigenvalue problems is stated comprising two nonself-adjoint plate problems given in a rectangular domain. The dependance of the real part of the eigenvalue λ upon two real parameters is investigated, one being the length of one side of the domain, the other being a load parameter in the differential equation. The numerical methods employed are similar to those used in KLEIN (1990), where the Orr-Sommerfeld problem is treated.

1. A Class of Eigenvalue Problems

A linear eigenvalue problem of the form

$$(A_o+\mu K)u = \lambda L_o u, \quad u \in D(A_o) \tag{1}$$

is considered, where μ is a given real parameter and A_o, K and L_o are linear operators, the domains of which are linear subspaces of a complex Hilbert space $(H,(\cdot,\cdot))$ such that $D(A_o) \subset D(L_o)$, $D(A_o) \subset D(K)$, $\overline{D(A_o)} = H$ holds. The operators A_o and L_o are symmetric and positive definite; K is not assumed to be a symmetric operator. Let H_o denote the closure of $D(A_o)$ with respect to the energy norm of the operator A_o. There exists a mapping $G: H \to H_o$, which is symmetric and positive, satisfying $A_o \subset G^{-1}$. Applying G to eigenvalue problem (1) yields

$$(I+\mu GK)u = \lambda GL_o u, \quad u \in H_o . \tag{2}$$

Under the assumptions

(A1) $\lambda = 0$ is no eigenvalue of (2)

(A2) G, GK, GL_o are compact operators on H_o

eigenvalue problem (2) has a countably infinite number of eigenvalues with no accumulation point in the complex plane. In order to treat eigenvalue problem (2) numerically it is transformed

---

Dedicated to Prof. Dr. L. Collatz on the occasion of his 80th birthday

into an equivalent eigenvalue problem with infinite dimensional matrices. Therefore it is assumed that for the following eigenvalue problem

$$A_o f = \lambda L_o f, \quad f \in D(A_o) \tag{3}$$

the eigenvalues $\lambda_j (j \in \mathbb{N})$ and corresponding eigenfunctions $f_j (j \in \mathbb{N})$ are known. The eigenfunctions shall be orthonormalized such that $(L_o f_j, f_k) = \delta_{jk}$ $(j,k \in \mathbb{N})$ holds. In case the system of eigenfunctions of (3) is complete in H_o eigenvalue problem (2) is equivalent to an eigenvalue problem in the Hilbert sequence space l_2

$$(I+\mu B)v = \lambda \Lambda^{-1} v, \quad v \in l_2 \tag{4}$$

with infinite dimensional matrices

$$B = (b_{jk}) = \left(\frac{(Kf_k, f_j)}{\sqrt{\lambda_k}\,\sqrt{\lambda_j}} \right)_{j,k \in \mathbb{N}}, \quad \Lambda^{-1} = \operatorname{diag}\left(\frac{1}{\lambda_j}\right)_{j \in \mathbb{N}}.$$

In the sequel the eigenfunctions $f_j (j \in \mathbb{N})$ of (3) will also be called coordinate functions.

2. Two Nonselfadjoint Plate Problems

The class of eigenvalue problems described above comprises two nonselfadjoint eigenvalue problems derived from plate vibration problems stated by LEIPHOLZ (1975):

Plate problem 2.1

$$\alpha \Delta^2 w + \gamma(a-x_1) w_{11} = \lambda p w \quad \text{in } \Omega = (0,a)\times(0,b)$$
$$w = \Delta w = 0 \text{ on } x_1 = 0,\ x_1 = a;\ x_2 = 0,\ x_2 = b$$

Plate problem 2.2

$$\alpha \Delta^2 w + \beta w_{11} = \lambda p w \quad \text{in } \Omega = (0,a)\times(0,b)$$
$$w = \Delta w = 0 \text{ on } x_1 = 0;\ x_2 = 0,\ x_2 = b$$
$$\Delta w - (1-\nu) w_{22} = 0,\ \Delta w_1 + (1-\nu) w_{122} = 0 \text{ on } x_1 = a$$

using the denotations $w_1 = \frac{\partial}{\partial x_1} w$, $w_2 = \frac{\partial}{\partial x_2} w$ and the constants: rigidity α, load γ, mass density p, boundary load β and Poisson ratio ν. As is shown in KLEIN (1987) plate problems 2.1 and 2.2 fulfill the above assumptions (A1) and (A2) as well as the completeness assumption concerning the eigenfunctions of eigenvalue problem (3) by choosing

$H = L_2(\Omega)$, $A_o u = \alpha\Delta^2 u$, $L_o u = pu$

$D(A_o) = \{u \in C^4(\bar{\Omega}) \mid u$ fulfills boundary conditions of 2.1 or 2.2$\}$

$\mu = \gamma$, $Ku = (a-x_1)u_{11}$ for plate problem 2.1

$\mu = \beta$, $Ku = u_{11}$ for plate problem 2.2.

Keeping the parameters α, p, b, ν constant the eigenvalues λ of eigenvalue problem (4) corresponding to one of the plate problems may be considered to be dependant upon the parameters a and μ only. In order to investigate the dependance of $Re\lambda$ upon a and μ the following definition is given:

Definition 2.3: A pair of parameters (a,μ) belongs to domain I, iff all eigenvalues λ of eigenvalue problem (4) satisfy $Re\lambda \geq 0$. A pair of parameters (a,μ) belongs to domain II, iff there exists an eigenvalue λ of eigenvalue problem (4) with $Re\lambda < 0$.

Remark 2.4: For two nonselfadjoint eigenvalue problems arising in the linear theory of hydrodynamic stability, the Taylor problem and the Orr-Sommerfeld problem, domain I is equivalent to the domain of stability and domain II is equivalent to the domain of instability. Both of these problems are derived from initial boundary value problems involving a first order partial derivative with respect to the time (see for instance LIN (1955)).

The interpretation of domains I and II for plate vibration problems is somewhat different owing to the second order partial derivative with respect to the time occurring in the plate vibration equation

$$\alpha\Delta^2\tilde{w} + \mu K\tilde{w} = -p\frac{\partial^2}{\partial t^2}\tilde{w} \quad . \tag{5}$$

Separating the time dependance from the spatial variables x_1, x_2

$$\tilde{w}(x_1,x_2,t) = w(x_1,x_2)\exp(i\sqrt{\lambda}t), \quad \sqrt{\lambda} = \rho+i\sigma \tag{6}$$

with the imaginary unit $i = \sqrt{-1}$, one ends up with:

$$\lambda = (\rho+i\sigma)^2 = \rho^2-\sigma^2+i2\rho\sigma, \quad Re\lambda = \rho^2-\sigma^2, \quad Im\lambda = 2\rho\sigma \tag{7}$$

$$\tilde{w} = w\exp((i\rho-\sigma)t) = w(\cos\rho t+i\sin\rho t)\exp(-\sigma t). \tag{8}$$

In case $\rho \neq 0$ holds, $\tau = |\rho|t$ may be defined leading to:

$$\tilde{w} = w(\cos\tau+\mathrm{sgn}(\rho)\,i\sin\tau)\exp(-\frac{\sigma}{|\rho|}\tau) \tag{8a}$$

where $\mathrm{sgn}(\rho)$ denotes the sign of ρ. Assuming $t \geq 0$ for $\sigma \geq 0$ $(\sigma < 0)$

the ratio $\sigma/|\rho|$ specifies the amount of exponential decrease (increase). Now the following interpretation may be given to domains I and II:

Domain I: For all eigenvalues λ of eigenvalue problem (4) $\mathrm{Re}\lambda = \rho^2-\sigma^2 \geq 0$ holds, i. e. $\rho^2 \geq \sigma^2$. In case $\rho = 0$ this entails $\sigma = 0$, such that both the oscillation term and the exponential term in (8) are constant; in case $\rho \neq 0$ the inequality $1 \geq (\frac{\sigma}{\rho})^2$ holds for the exponent in (8a).

Domain II: For one eigenvalue λ of eigenvalue problem (4) $\mathrm{Re}\lambda = \rho^2-\sigma^2 < 0$ holds, i. e. $\rho^2 < \sigma^2$. In the special case $\rho = 0$ the oscillation term in (8) is constant; for $\rho \neq 0$ the inequality $1 < (\frac{\sigma}{\rho})^2$ is valid for the exponent in (8a).

3. Locating Domains I and II in the (a,μ)-Plane

Let λ be an eigenvalue of eigenvalue problem (4), v a corresponding eigenvector and μ a real number, then $\mathrm{Re}\lambda$ may be represented by

$$\mathrm{Re}\lambda = \frac{\bar{v}^T v+\mu\bar{v}^T\frac{1}{2}(B+\bar{B}^T)v}{\bar{v}^T\Lambda^{-1}v} . \tag{9}$$

A sufficient condition for all eigenvalues λ of (4) to satisfy $\mathrm{Re}\lambda \geq 0$, i. e. for (a,μ) to belong to domain I, would be

$$\inf_{v \in l_2-\{0\}} \frac{\bar{v}^T v+\mu\bar{v}^T\frac{1}{2}(B+\bar{B}^T)v}{\bar{v}^T\Lambda^{-1}v} \geq 0. \tag{10}$$

This condition is also necessary in case $B = \bar{B}^T$ holds. An equivalent condition is given by

$$\frac{1}{\mu} \geq -\inf_{v \in l_2-\{0\}} \frac{\bar{v}^T\frac{1}{2}(B+\bar{B}^T)v}{\bar{v}^T v} = \sup_{v \in l_2-\{0\}} -\frac{\bar{v}^T\frac{1}{2}(B+\bar{B}^T)v}{\bar{v}^T v} . \tag{11}$$

The operator $-\frac{1}{2}(B+\bar{B}^T)$ is symmetric and compact. Thus

$$s(a) = \sup_{v \in l_2-\{0\}} \frac{\bar{v}^T(-\frac{1}{2}(B+\bar{B}^T))v}{\bar{v}^T v} \tag{12}$$

is the largest eigenvalue of the following eigenvalue problem

$$-\frac{1}{2}(B+\bar{B}^T)v = \kappa v, \quad v \in l_2 \,. \tag{13}$$

A sufficient condition for domain I is given by:
A pair (a,μ) satisfying $\frac{1}{\overline{s(a)}} \geq \mu$ belongs to domain I, where $\overline{s(a)}$ denotes an upper bound of $s(a)$ defined in (12).
On the other hand a criterion for domain II is furnished by:
A pair (a,μ) belongs to domain II, if for (a,μ) the eigenvalue inclusion for eigenvalue problem (4) yields a Gerschgorin disk, being isolated from the other inclusion sets for eigenvalues and lying completely in the left half plane (Rez < 0), so that the included eigenvalue has a negative real part.

4. Numerical Methods

Let N be natural number; defining the vector norm in l_2

$$v = (\zeta_j)_{j \in \mathbb{N}}, \quad \|v\| = \max\{\max_{1 \leq j \leq N} |\zeta_j|, \left(\sum_{j=N+1}^{\infty} |\zeta_j|^2\right)^{\frac{1}{2}}\}, \tag{14}$$

given by DONELLY (1974), an appropriate matrix norm can be derived

$$\begin{aligned}\|A\| = \max\{&\max_{1 \leq i \leq N} \Big[\sum_{j=1}^{N} |a_{ij}| + \left(\sum_{j=N+1}^{\infty} |a_{ij}|^2\right)^{\frac{1}{2}}\Big], \\ &\sum_{j=1}^{N} \left(\sum_{i=N+1}^{\infty} |a_{ij}|^2\right)^{\frac{1}{2}} + \left(\sum_{i,j=N+1}^{\infty} |a_{ij}|^2\right)^{\frac{1}{2}}\}.\end{aligned} \tag{15}$$

Using this matrix norm and an assumption for matrix B in (4)

$$\sum_{j,k=1;\, j \neq k}^{\infty} |b_{jk}|^2 < \infty, \quad \sup_{j \in \mathbb{N}} |b_{jj}| < \infty \tag{16}$$

eigenvalue inclusions for eigenvalue problem (4) and (13) may be derived after a suitable transformation of an N-dimensional subproblem of (4) and (13) respectively. For eigenvalue problem (4) this is shown in DONELLY (1974) or KLEIN (1987); for eigenvalue problem (13) the procedure described in KLEIN (1989) may be adapted to matrices $-\frac{1}{2}(B+\bar{B}^T)$ having elements different from zero on the main diagonal, as will be shown in the sequel. Using the denotation $C = -\frac{1}{2}(B+\bar{B}^T)$ the infinite dimensional eigenvalue problem (13) is subdivided into an N-dimensional subproblem and the rest

$$\begin{pmatrix} C_{11} - \kappa I_1 & C_{12} \\ C_{21} & C_{22} - \kappa I_2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} \Theta \\ \Theta \end{pmatrix} \tag{17}$$

with $N\times N$ matrices C_{11}, I_1 and an N-dimensional vector v_1. As C_{11} is a symmetric matrix it is similar to the diagonal matrix of its eigenvalues. Calculating eigenvalues and eigenvectors numerically will lead to an "almost" diagonal matrix M, having possibly small off-diagonal entries:

$$C_{11}U = UM, \qquad U \text{ regular} .$$

Denoting $D_1 = \operatorname{diag}(m_{ii})_{i=1}^{N}$, $\hat{v}_1 = U^{-1}v_1$, $\hat{v} = (\hat{v}_1, v_2)^T$

$$S = \begin{pmatrix} M-D_1 & U^{-1}C_{12} \\ C_{21}U & C_{22} \end{pmatrix} \qquad D = \begin{pmatrix} D_1 & \Theta \\ \Theta & \Theta \end{pmatrix}$$

eigenvalue problem (17) is equivalent to

$$(S+D-\kappa I)\hat{v} = \Theta \quad \text{or} \quad S\hat{v} = (\kappa I - D)\hat{v}. \tag{18}$$

Applying a positive diagonal similarity transformation

$$X = \begin{pmatrix} X_1 & \Theta \\ \Theta & I_2 \end{pmatrix}, \quad X_1 = \operatorname{diag}(x_i)_{i=1}^{N}, \quad x_i > 0 \ (i=1,2,\ldots,N)$$

to (18) leads to

$$X^{-1}SX\tilde{v} = (\kappa I - D)\tilde{v}, \quad \tilde{v} = X^{-1}\hat{v} \tag{19}$$

Thus for all eigenvalues $\kappa \neq m_{ii}$ $(i=1,2,\ldots,N)$ the inequality $1 \leq \|A\|$ with $A = (\kappa I - D)^{-1}X^{-1}SX$ and matrixnorm (15) holds, leading to the eigenvalue inclusion

$$\kappa \in \bigcup_{i=1}^{N+1} G_i(x) \text{ with } x = (x_1, x_2, \ldots, x_N)^T, \ x_i > 0 \ (i=1,2,\ldots,N)$$

$$G_i(x) = \{z \in \mathbb{C} \mid |z - m_{ii}| \leq r_i(x)\}, \quad r_i(x) = \frac{1}{x_i}\Big(\sum_{j=1, j\neq i}^{N} |m_{ij}| x_j + \delta_i\Big) \ (i=1,2,\ldots,N)$$

$$G_{N+1}(x) = \{z \in \mathbb{C} \mid \inf_{j \geq N+1} |z - c_{jj}| \leq r_{N+1}(x)\}, \quad r_{N+1}(x) = \sum_{j=1}^{N} \eta_j x_j + \eta_{N+1}$$

with $\eta_j = \sum_{l=1}^{N} |u_{lj}| \beta_l$, $\beta_l^2 = \sum_{i=N+1}^{\infty} |c_{il}|^2$, $U = (u_{ij})_{1 \leq i,j \leq N}$

$$\delta_j = \eta_j \ (j=1,2,\ldots,N); \quad \eta_{N+1}^2 = \sum_{i,j=N+1; i\neq j}^{\infty} |c_{ij}|^2 .$$

The convergence of the infinite series involved is guaranteed by (16) and the triangle inequality for the Euclidean vector norm. The relation $\delta_j = \eta_j$ $(j=1,2,\ldots,N)$ is due to the symmetry of $C = -\frac{1}{2}(B+\bar{B}^T)$ and $U^{-1} = \bar{U}^T$. Let c be an arbitrary complex number. Introducing the disk

$$K_c(x) = \{z \in \mathbb{C} \mid |z-c| \leq r_c(x)\},\quad r_c(x) = r_{N+1}(x) + \sup_{j\geq N+1} |c-c_{jj}|$$

the inclusion $G_{N+1}(x) \subset K_c(x)$ is valid, because for all $z \in G_{N+1}(x)$ the following inequality holds:

$$|z-c| - \sup_{j\geq N+1} |c-c_{jj}| \leq \inf_{j\geq N+1} |z-c_{jj}| \leq r_{N+1}(x).$$

If for a positive vector $x = (x_1,x_2,\ldots,x_N)$ the i-th Gerschgorin disk with $i \in \{1,2,\ldots,N\}$ is separated from the other disks such that the separation conditions

(S1) $\overset{\circ}{G}_i(x) \cap K_c(x) = \emptyset$

(S2) $\overset{\circ}{G}_i(x) \cap G_j(x) = \emptyset$ for $j=1,2,\ldots,N;\ j\neq i$

are fulfilled ($\overset{\circ}{G}_i(x)$ denoting the interior of $G_i(x)$), then $G_i(x)$ contains at least one eigenvalue of eigenvalue problem (13). If $G_i(x)$ has no boundary points in common with any other disk, then $G_i(x)$ contains exactly one eigenvalue of eigenvalue problem (13). An equivalent formulation for the separation conditions (S1), (S2) can be given in the form of inequalities:

(S1') $r_i(x) + r_{N+1}(x) + \sup_{j\geq N+1} |c-c_{jj}| \leq |m_{ii}-c|$

(S2') $r_i(x) + r_j(x) \leq |m_{ii}-m_{jj}|$ $(j=1,2,\ldots,N;\ j\neq i)$.

Under additional assumptions (S1') can be stated in a more explicit form. Assuming

$$\underline{c} \leq c_{jj} \leq \bar{c} \quad \text{for all } j \geq N+1 \tag{20}$$

and choosing $c = \frac{1}{2}(\bar{c}+\underline{c})$ leads to $\sup_{j\geq N+1} |c-c_{jj}| \leq \frac{1}{2}(\bar{c}-\underline{c})$.

Then the following condition is sufficient for (S1'):

(S1") $r_i(x) + r_{N+1}(x) + \frac{1}{2}(\bar{c}-\underline{c}) \leq |m_{ii}-c|$.

If $0 \leq m_{ii} - \bar{c}$ is valid, implying $c \leq \bar{c} \leq m_{ii}$, (S1") reduces to

$$r_i(x) + r_{N+1}(x) \leq m_{ii} - \bar{c} \quad . \tag{21a}$$

If instead $0 \leq \underline{c} - m_{ii}$ is valid, implying $m_{ii} \leq \underline{c} \leq c$, (S1") turns out to be

$$r_i(x) + r_{N+1}(x) \leq c - m_{ii} \,. \tag{21b}$$

The algorithm stated in KLEIN (1988) investigating whether there exists a single Gerschgorin disk, being isolated from the other inclusion sets for eigenvalues, in the case of eigenvalue problem (4), may be adapted to the separation conditions (S1"), (S2') for eigenvalue problem (13).

5. Numerical Examples

For plate problem 2.1 with constants α, γ and p

$$\alpha\Delta^2 w + \gamma(a-x_1)w_{11} = \lambda p w \text{ in } \Omega = (0,a)\times(0,b)$$
$$w = \Delta w = 0 \text{ on } x_1 = 0,\ x_1 = a;\ x_2 = 0,\ x_2 = b$$

the following proposition holds:

Proposition 5.1 Plate problem 2.1 has only real eigenvalues.

Proof: All eigenvalues of plate problem 2.1 are also eigenvalues of the following selfadjoint eigenvalue problem:

$$\begin{aligned} &\alpha\Delta^2 u + \gamma((a-x_1)u_1)_1 = \lambda p u \text{ in } \Omega = (0,a)\times(0,b) \\ &u = \Delta u = 0 \quad \text{on } x_2 = 0,\ x_2 = b \\ &u_1 = \Delta u_1 = 0 \quad \text{on } x_1 = 0,\ x_1 = a. \end{aligned} \tag{22}$$

On the other hand every eigenvalue λ of eigenvalue problem (22) with $\lambda \neq \frac{\alpha}{p}(\frac{k\pi}{b})^4$, $k \in \mathbb{N}$, is an eigenvalue of plate problem 2.1. In case $\lambda = \frac{\alpha}{p}(\frac{k\pi}{b})^4$, $k \in \mathbb{N}$, $u(x_1,x_2) = \sin\frac{k\pi}{b}x_2$ is a corresponding eigenfunction of eigenvalue problem (22).

Corollary 5.2 For plate problem 2.1 the following interpretation may be given to domains I and II using (8) and Im $\lambda = 2\rho\sigma = 0$ for all eigenvalues λ of 2.1: In domain I, where $\rho^2 \geq \sigma^2$ holds for all eigenvalues λ, $\rho\cdot\sigma = 0$ leads to $\sigma = 0$ for all eigenvalues λ, implying undamped oscillations for the corresponding solutions of plate vibration problem (5). In domain II, where $\rho^2 < \sigma^2$ holds for one eigenvalue λ, $\rho\cdot\sigma = 0$ leads to $\rho = 0$ for one eigenvalue λ, implying no oscillation for the corresponding solution of plate vibration problem (5).

For the numerical treatment of plate problem 2.1 in the case of $\alpha = p = 1$ the following eigenvalue problem

$$\Delta^2 u = \lambda u \text{ in } \Omega,\ u = \Delta u = 0 \text{ on } \partial\Omega \tag{23}$$

has to be solved for eigenvalues and eigenfunctions. Both can be given in closed form:

$$\lambda_{jk} = [(\frac{j\pi}{a})^2 + (\frac{k\pi}{b})^2]^2, \quad u_{jk} = \frac{1}{\sqrt{ab}} \sin\frac{j\pi}{a}x_1 \sin\frac{k\pi}{b}x_2 \quad (j,k \in \mathbb{N}).$$

Matrix B of eigenvalue problem (4) is real and nonsymmetric. Matrix $-\frac{1}{2}(B+B^T)$ of eigenvalue problem (13) is positive definite, ensuring that every finite dimensional subproblem has positive eigenvalues.

The numerical treatment of plate problem 2.2

$$\alpha\Delta^2 w + \beta w_{11} = \lambda p w \text{ in } \Omega = (0,a)\times(0,b)$$
$$w = \Delta w = 0 \text{ on } x_1 = 0;\ x_2 = 0,\ x_2 = b$$
$$\Delta w - (1-\nu)w_{22} = 0,\ \Delta w_1 + (1-\nu)w_{122} = 0 \text{ on } x_1 = a$$

in the case $\alpha = p = 1$, using the above mentioned theory, necessitates the solution of the following selfadjoint eigenvalue problem:

$$\Delta^2 u = \lambda u \text{ in } \Omega, \text{ boundary conditions as in plate problem 2.2.} \tag{24}$$

Starting with eigenfunctions of the form

$$u(x_1,x_2) = f(x_1)\sin\frac{k\pi x_2}{b} \text{ for fixed } k \in \mathbb{N} \tag{25}$$

leads to an eigenvalue problem with an ordinary differential equation, having constant coefficients, for $f(x_1)$ as an eigenfunction and λ as an eigenvalue (see KLEIN (1987)). The eigenvalues of this problem may be obtained by solving a transcendental equation given in LEIPHOLZ (1975). The corresponding eigenfunctions are linear combinations of hyperbolic and trigonometric sin-functions. Matrix B of eigenvalue problem (4) again turns out to be real and nonsymmetric; matrix $-\frac{1}{2}(B+B^T)$ is not positive definite in general, since finite dimensional submatrices for fixed a, b, ν are having negative eigenvalues.

There exist two symmetry classes of eigenfunctions for plate problem 2.2 (and plate problem 2.1 as well): eigenfunctions being even and eigenfunctions being odd as functions of x_2 with respect to $x_2 = \frac{b}{2}$. The matrix eigenvalue problem corresponding to plate problem 2.2 may be split up into two subproblems using only coordinate functions of the symmetry class under consideration, thus reducing the amount of numerical labour for each subproblem.

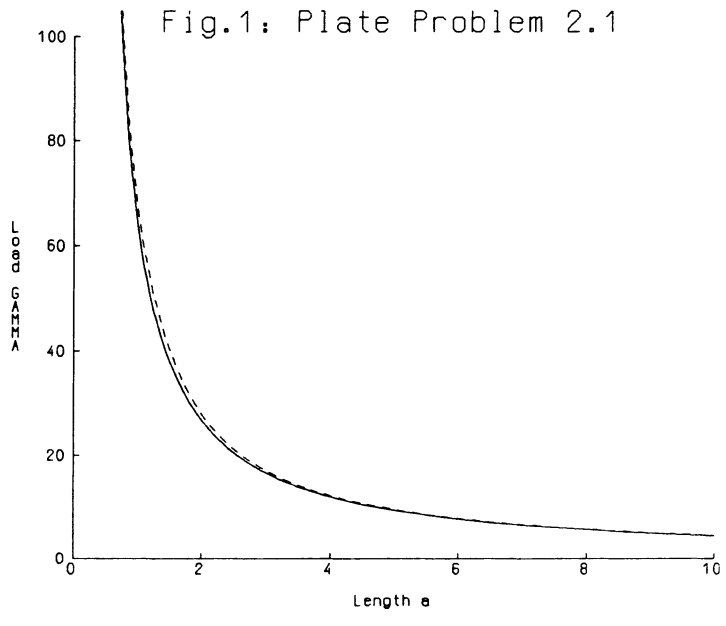


Fig.1: Plate Problem 2.1
Load GAMMA
100
80
60
40
20
0
0
2
4
6
8
10
Length a

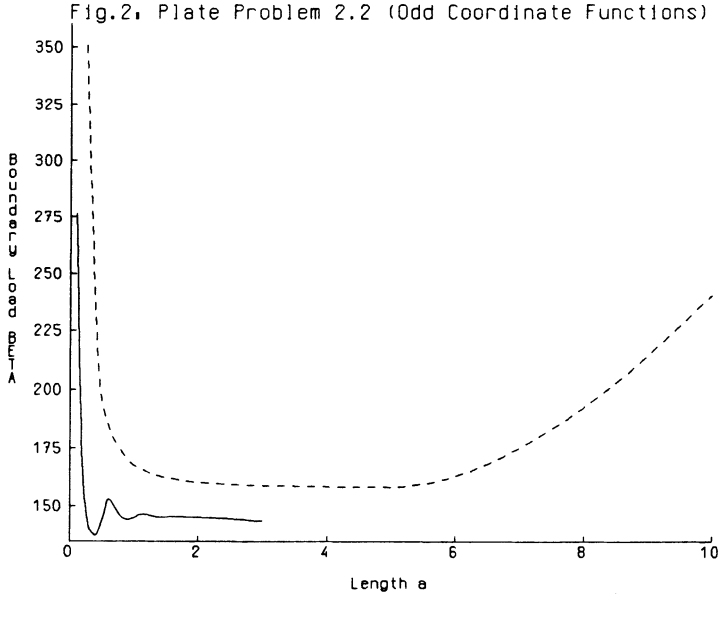


Fig.2: Plate Problem 2.2 (Odd Coordinate Functions)
Boundary Load BETA
350
325
300
275
250
225
200
175
150
0
2
4
6
8
10
Length a

Numerical results were obtained for $b = 1$, $\nu = 0.3$.
Explanation of figures 1, 2 and 3: In figure 1 the solid line curve interpolates points (a,γ), in figures 2 and 3 it interpolates points (a,β), which are obtained according to the sufficient condition for domain I, given in chapter 3. The broken line curves interpolate points (a,γ) in figure 1 and points (a,β) in figures 2 and 3 obtained according to the criterion for domain II, when it was first fulfilled. In figure 1 the broken line curve and the solid line curve are almost coinciding, indicating that plate problem 2.1 is "not far away" from a selfadjoint eigenvalue problem. Comparing figures 2 and 3 both the broken line curve and the solid line curve of figure 2 are above the respective curves of figure 3, showing that the curves of figure 3 are the relevant ones. In the case of the solid line curves in figures 2 and 3 results could not be obtained for values of length a much bigger than 2, because it would have been necessary, for the eigenvalue inclusion process to converge, to work with larger finite dimensional matrices, for which there was not enough virtual storage on the central computer used.
Explanation of figure 4 and table 1: Figure 4 and table 1 are meant to give an understanding of the oscillating behaviour of the solid line curve of figure 3. This curve interpolates lower bounds obtained for the reciprocal of the largest eigenvalue of eigenvalue problem (13). The three curves in figure 4: solid, broken and dotted are interpolating upper bounds to the reciprocals of the three largest eigenvalues of eigenvalue problem (13) for different values of a; the solid line curve refers to the largest eigenvalue, the broken line curve refers to the second largest eigenvalue and the dotted curve to the third largest eigenvalue of eigenvalue problem (13). According to the principle of Ritz (see for instance COLLATZ (1968)) the three largest eigenvalues of eigenvalue problem (13) are bounded from below by the three largest eigenvalues of any finite dimensional subproblem of (13). When going over to reciprocals lower bounds turn into upper bounds. In figure 4 the marks below each curve are giving lower bounds to the reciprocals of the three largest eigenvalues of (13) for

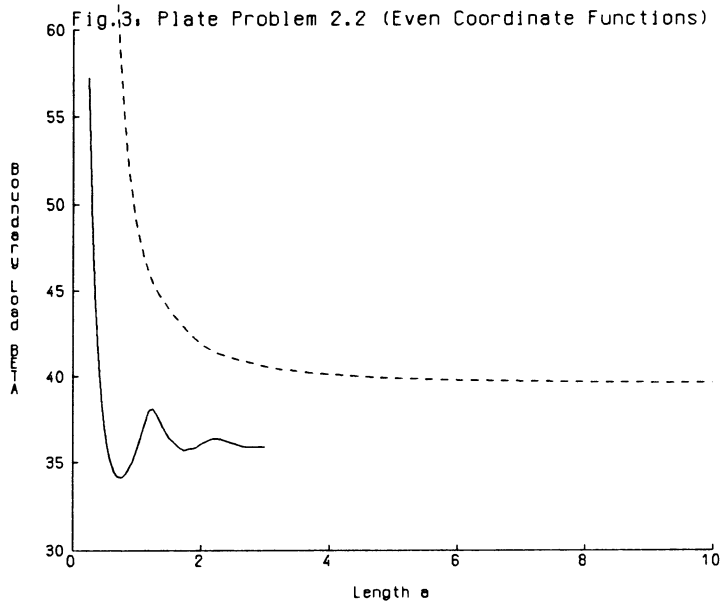


Fig.3, Plate Problem 2.2 (Even Coordinate Functions)
Boundary Load BETA
60
55
50
45
40
35
30
0
2
4
6
8
10
Length a

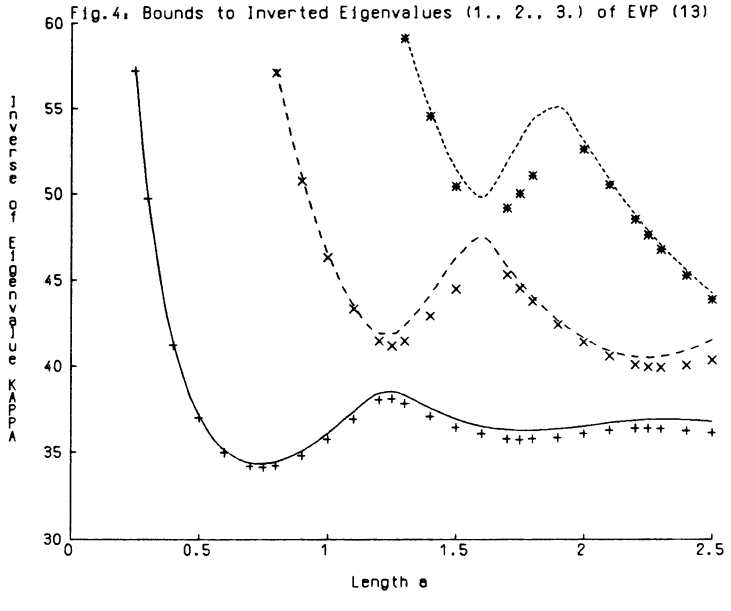


Fig.4, Bounds to Inverted Eigenvalues (1., 2., 3.) of EVP (13)
Inverse of Eigenvalue KAPPA
60
55
50
45
40
35
30
0
0.5
1
1.5
2
2.5
Length a

Table 1. Number of sign changes of approximate eigenfunctions belonging to the three largest eigenvalues of the eigenvalue problem in H_o being equivalent to eigenvalue problem (13). The eigenfunctions are considered as functions of x_1 in the interval (0,a).

a	3rd largest eigenvalue	2nd largest eigenvalue	largest eigenvalue
0.70	3	2	1
0.75	3	2	1
0.80	3	2	1
0.90	3	2	1
1.00	3	2	1
1.10	3	2	1
1.20	3	2	2
1.25	3	2	2
1.30	3	1	2
1.40	3	1	2
1.50	3	1	2
1.60	1	3	2
1.70	1	3	2
1.75	1	3	2
1.80	1	3	2
1.90	4	3	2
2.00	4	3	2
2.10	4	3	2
2.20	4	3	2
2.25	4	3	3
2.30	4	3	3
2.40	4	2	3
2.50	4	2	3

different values of a, when an eigenvalue inclusion by the method of Gerschgorin was possible.

Comparing figure 4 and table 1 with respect to the largest eigenvalue of (13) the first arc of the solid line curve is characterized by one sign change of the corresponding approximate eigenfunction, the second arc is characterized by two sign changes and the third arc is characterized by three sign changes. The number of sign changes of the approximate eigenfunction belonging to the largest eigenvalue of (13) varies whenever the solid line curve in figure 4 referring to the same eigenvalue is approached by the broken line curve referring to the second largest eigenvalue of (13). Similarly the number of sign changes of the

approximate eigenfunction belonging to the second largest eigenvalue of (13) varies whenever the broken line curve in figure 4 referring to the same eigenvalue is approached by the solid line curve referring to the largest eigenvalue of (13) and the dotted curve referring to the third largest eigenvalue of (13).

6. References

Collatz, L. (1968) Funktionalanalysis und numerische Mathematik, 1st edn (Springer, Berlin)

Donnelly, J. D. P. (1974) Bounds for the eigenvalues of non-self-adjoint differential operators. J. Inst. Math. Appl. 13, 249 - 261

Klein, P. P. (1987) Zur Eigenwerteinschließung bei nichtselbstadjungierten Eigenwertaufgaben mit Differentialgleichungen. In Albrecht, J. u. a. (eds): ISNM 83, Numerische Behandlung von Eigenwertaufgaben, Band 4. Basel: Birkhäuser Verlag, pp. 130 - 144

Klein, P. P. (1988) Inclusion of eigenvalues of nonselfadjoint eigenvalue problems. Proceedings of 2nd Int. Symp. on Numerical Analysis (ISNA) Prag 1987. Teubner Texte Band 107, pp. 205 - 209

Klein, P. P. (1989) Including eigenvalues of nonselfadjoint eigenvalue problems. Proceedings of the International Conference on Numerical Methods and Applications, Sofia 1988. Publishing House of the Bulgarian Academy of Sciences, pp. 217 - 221

Klein, P. P. (1990) Eigenwerteinschließung bei nichtselbstadjungierten Eigenwertaufgaben. Z. angew. Math. Mech. 70, T560 - T562

Leipholz, H. H. E. (1975) Die direkte Methode der Variationsrechnung und Eigenwertprobleme der Technik (G. Braun, Karlsruhe)

Lin, C. C. (1955) The theory of hydrodynamic stability (Cambridge University Press)

Dr. Peter Paul Klein, Rechenzentrum der TU Clausthal, Erzstraße 51, D-3392 Clausthal-Zellerfeld, Bundesrepublik Deutschland.

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

MINIMIZATION OF THE VARIANCE. A METHOD FOR TWO-SIDED BOUNDS FOR EIGENVALUES OF SELFADJOINT OPERATORS.

Heinz Kleindienst, Rainer Emrich

To Prof. Dr. Dr. h.c. Lothar Collatz on the occasion of his 80^{th} birthday

1. THEORY

Let H be a selfadjoint operator in a Hilbert-Space $\mathcal{H}$, D_H the domain of H, $\sigma(H)$ the spectrum and $\varrho(H)$ the resolvent set of H. From the spectral theorem of selfadjoint operators for each $v \in \mathcal{H}$ and $u \in D_A$ with $A = f(H)$ and continuous f follows

$$(v|f(H)u) = \int_{-\infty}^{\infty} f(\lambda) d(v|E_\lambda u) \tag{1}$$

with $\{E_\lambda\}$ as a spectral family of H. Let λ^* be an arbitrary real number and r the distance between λ^* and $\sigma(H)$, i.e.

$$r = \inf_{\lambda \in \sigma(H)} |\lambda - \lambda^*| \quad .$$

It follows for every $u \in D_H$

$$\begin{aligned} r^2 ||u||^2 &= \int_{-\infty}^{\infty} r^2 d||E_\lambda u||^2 \\ &\leq \int_{-\infty}^{\infty} |\lambda - \lambda^*|^2 d||E_\lambda u||^2 \\ &= ||(H - \lambda^*)u||^2 \end{aligned}$$

taking into account $E_\lambda u$ being a constant function on $\varrho(H)$ with respect to λ and $r^2 \leq |\lambda - \lambda^*|^2$ for all $\lambda \in \sigma(H)$. The inequality obtained

$$\inf_{\lambda \in \sigma(H)} |\lambda - \lambda^*| \leq ||(H - \lambda^*)u|| \quad , \quad u \in D_H \quad , \quad ||u|| = 1 \tag{2}$$

leads for a n-dimensional subspace $V_n \subset D_H$ to the principle of variance minimization :

THEOREM 1: Minimization of the non-negative functional

$$F[u] = ||Hu||^2 - (Hu|u)^2 \quad , \quad ||u|| = 1$$

i.e. the determination of

$$F_o^* = \inf_{u \in V_n} F[u] = F[u_o^*]$$

gives an $u_o^* \in V_n$ and further a $\lambda_o^* = (Hu_o^*|u_o^*)$ with

$$\inf_{\lambda \in \sigma(H)} |\lambda - \lambda_o^*|^2 \leq F_o^* \quad .$$

Then λ_o^* is an approximate value for a spectral point λ_o with

$$\lambda_o^* - \sqrt{F_o^*} \leq \lambda_o \leq \lambda_o^* + \sqrt{F_o^*} \quad .$$

PROOF: The left side of inequality (2) is independent of u and therefore the following inequality is true

$$\inf_{\lambda \in \sigma(H)} |\lambda - \lambda^*| \leq \inf_{u \in V_n} ||(H - \lambda^*)u|| \quad , \quad ||u|| = 1 \quad . \tag{3}$$

From

$$||(H - \lambda^*)u||^2 = (\lambda^* - (Hu|u))^2 + ||Hu||^2 - (Hu|u)^2 \quad , \quad ||u|| = 1 \tag{4}$$

it is obvious (because $||Hu|| \geq |(Hu|u)|$) that the minimal value of the left side of eq. (4) is F_o^*, i.e. there is an $u_o^* \in V_n$ and a λ_o^* with

$$\lambda_o^* = (Hu_o^*|u_o^*) \quad .$$

Because $\sigma(H)$ is a closed set on the real line considered as a metric space, there is at least one point $\lambda = \lambda_o \in \sigma(H)$ for which the infimum in eq. (3) is assumed. Therefore we have

$$|\lambda_o - \lambda_o^*|^2 \leq F_o^*$$

and λ_o^* is an approximate value for λ_o with the error $\sqrt{F_o^*}$, i.e.

$$\lambda_o^* - \sqrt{F_o^*} \leq \lambda_o \leq \lambda_o^* + \sqrt{F_o^*} \quad . \tag{5}$$

Because of

$$\lambda_o^* \geq \lambda_o^r = \inf_{u \in V_n} (Hu|u) \quad , \quad ||u|| = 1$$

with λ_o^r being the Ritz value, the upper bound of inequality (5) can be improved by λ_o^*, i.e.

$$\lambda_o^* - \sqrt{F_o^*} \leq \lambda_o \leq \lambda_o^r \leq \lambda_o^* \quad .$$

For selfadjoint operators bounded below with a discrete spectrum $\sigma_d(H)$ below the bottom of the continuum the approximation (5) can be sharpened. Using Temple's formula with

$$\sigma_d(H) = \{E_i | E_0 < E_1 < E_2 < \ldots\}$$

we have

$$E_i \geq E_i^* = \lambda_i^* - \frac{F_i^*}{\rho - \lambda_i^*} \quad , \quad E_i < \rho < E_{i+1} \quad .$$

Because $\lambda_i^* \geq \lambda_i^r$ but $F_i^* \leq F_i^r$ with

$$\lambda_i^r = (Hu_i^r | u_i^r) \quad \text{and} \quad F_i^r = ||Hu_i^r||^2 - (Hu_i^r | u_i^r)^2$$

the lower bound E_i^* is sharpened by both terms against the lower bound E_i^r with

$$E_i \geq E_i^r = \lambda_i^r - \frac{F_i^r}{\rho - \lambda_i^r}$$

obtained from Temple's formula with the Ritz values λ_i^r and F_i^r. The relations are demonstrated for $i = 0$ in Table 2. Obviously Temple's formula needs a lower bound for E_{i+1} to determine ρ. Usually the correct sequence of the eigenvalues is known from either upper bound calculations or experimental data available. It is just an advantage of the following "linearized version" of the variance minimization to yield suitable ρ's:
To minimize $F_n^*(u)$ from (3) the problem is reduced to matrix diagonalization of quadratic forms. This is established by an iterative procedure starting with the Rayleigh-Quotient

$$R[\lambda, u] = \frac{||(H - \lambda)u||^2}{||u||^2}$$

with

$$u = \sum_{p=1}^{n} c_p v_p \quad , \quad c_p \in \mathbb{R}$$

and v_p as a basis of $V_n \subset D_H$. The determination of the extremal values of $R[\lambda, u]$ by

$$\frac{\partial R}{\partial \lambda} = 0 \quad , \quad \frac{\partial R}{\partial c_j} = 0 \quad (j = 1, \ldots, n)$$

yields a nonlinear system of equations

$$\lambda = \frac{\sum_{i,k} H_{ik} c_i c_k}{\sum_{i,k} S_{ik} c_i c_k} \tag{6}$$

$$\sum_{k=1}^{n}(A_{ik}(\lambda) - \Lambda S_{ik})c_k = 0 \quad (i = 1, \ldots, n) \tag{7}$$

with

$$\begin{aligned} S_{ik} &= (v_i|v_k) = S_{ki} \\ H_{ik} &= (v_i|Hv_k) = H_{ki} \\ K_{ik} &= (Hv_i|Hv_k) = K_{ki} \\ A_{ik}(\lambda) &= K_{ik} - 2\lambda H_{ik} + \lambda^2 S_{ik} \end{aligned}$$

and

$$\Lambda = \inf_{\lambda,c} R[\lambda, u] \quad .$$

Using an arbitrary starting value for $\lambda^{(o)}$ in (7) we obtain from the secular determinant

$$det(A_{ik}^{(1)} - \Lambda^{(1)} S_{ik}) = 0 \quad \text{with} \quad A_{ik}^{(1)} = A_{ik}(\lambda^{(o)})$$

a sequence of eigenvalues $\Lambda_p^{(1)}$, $(p = 1, \ldots, n)$. The latter are the extremal values of $R[\lambda, u]$ for $\lambda = \lambda^{(o)}$ with the corresponding eigenvectors $c_p^{(1)}$. By substituting $c_p^{(1)}$ in (6) we get a $\lambda_p^{(1)}$ for each $\Lambda_p^{(1)}$. Via eq. (7) any new $\lambda^{(1)}$ from the set of $\lambda_p^{(1)}$ yields another sequence $\Lambda_p^{(2)}$ etc. . For this iterative process the following statements can be proved concerning the lowest eigenvalue E_0 [1]:

(1) The sequence $\Lambda^{(i)} = \Lambda^{(i)}(n)$ is monotonously decreasing and converges to a $\Lambda^{(\infty)}(n) \geq F_n^*$ for any n.

(2) The sequence $\lambda^{(i)} = \lambda^{(i)}(n)$ is monotonous and bounded for $i > i_o(n)$ and converges to a $\lambda^* = \lambda^{(\infty)}(n)$ for any n.

(3) A sequence of starting values $\lambda_m^{(o)} > E_0$ with $\lambda_m^{(o)} < \lambda_{m+1}^{(o)}$ gives a sequence of monotonously increasing lower bounds for E_0, as long as the interval

$$[\lambda_m^{(o)} - \sqrt{\Lambda_m^{(1)}}, \lambda_m^{(o)} + \sqrt{\Lambda_m^{(1)}}]$$

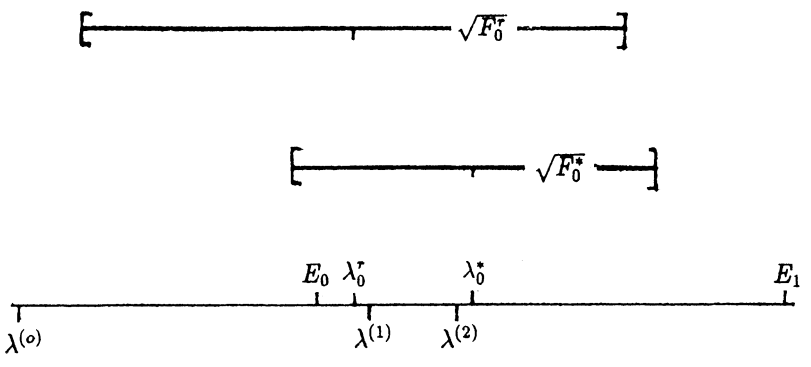
contains the eigenvalue E_0 only.

(4) A sequence of starting values $\lambda_m^{(o)}$ with $E_0 > \lambda_m^{(o)} > \lambda_{m+1}^{(o)}$ produces a sequence of corresponding upper bounds for E_0, which is monotonously decreasing and converges to the Ritz value λ_0^r with

$$\lambda_0^r = \inf_{u \in V_n} (Hu|u) \quad , \quad ||u|| = 1$$

for $\lambda_m^{(o)} \underset{m \to \infty}{\longrightarrow} -\infty$.

The relations for the lowest eigenvalue E_0 are illustrated on the real axis



2. APPLICATIONS

The method of variance minimization was originally developed to get two-sided bounds for eigenvalues of Schrödinger operators. The latter are selfadjoint operators bounded below with a discrete spectrum below the bottom of the continuum [2]. The corresponding operator H is [3]

$$H = T + V \tag{8}$$

with

$$\begin{aligned} T &= -\frac{\hbar^2}{2}\sum_{a=1}^{Q}\frac{\Delta_a}{M_a} - \frac{\hbar^2}{2m}\sum_{k=1}^{N}\Delta_k \\ V &= \sum_{a<b}\frac{Z_a Z_b e^2}{|R_a - R_b|} - \sum_{k,a}\frac{Z_a e^2}{|R_a - r_k|} + \sum_{i<k}\frac{e^2}{|r_i - r_k|} \quad . \end{aligned}$$

The fundamental works of *T. Kato* [4], *B. Simon* [5] *et al.* show that the Schrödinger equation

$$H\psi = E\psi$$

is a correct eigenvalue problem with a selfadjoint Schrödinger operator H (8). The domain D_H of H is characterized as a subspace of a Sobolev space [6]. The efficiency of the method of chapter 1 will be demonstrated at the He atom. Using the so called infinite mass approximation, i.e. the neglect of the motion of the nucleus (Born-Oppenheimer Approximation), the Schrödinger operator for this system is given (in atomic units a.u.) by

$$H = -\frac{1}{2}(\Delta_1 + \Delta_2) - \frac{2}{r_1} - \frac{2}{r_2} + \frac{1}{r_{12}} \quad . \tag{9}$$

The results of the linearized procedure are shown in Table 1. After two iterations the final value λ_o^* and the minimal error Λ_o^* are already reached.

Table 1

Convergence of the iteration procedure

Iterations	$\lambda^{(i)}$	$\Lambda^{(i+1)}$
$i = 0$	− 100	942.768
$i = 1$	− 2.9037236	$5.19735 \cdot 10^{-5}$
$i = 2$	− 2.9037219	$5.19729 \cdot 10^{-5}$
$i = 3$	− 2.9037219	$5.19729 \cdot 10^{-5}$

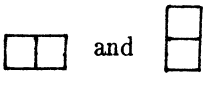
For the variation we used functions of the type

$$\psi = \sum_{p,q,s=0}^{M} c_{pqs} r_1^p r_2^q r_{12}^s e^{-\alpha(r_1+r_2)} \tag{10}$$

with $p + q + s \leq M$ and $c_{pqs} = c_{qps}$. Because $||H\psi|| < \infty$ it follows that $\psi \in D_H$. With $M = 8$ it is easy to see that $\dim V_n = 95$.
REMARK: The choice of functions symmetric in r_1 and r_2 is admissible because the symmetry group of the Hamiltonian H from eq. (9) is the symmetric group S_2. Following from the Pauli principle the eigensolutions of physical relevance of H (eq. (8)) exist only in symmetry adapted subspaces $V_\Gamma \subset D_H$. The V_Γ's transform like the irreducible representations (IR's)

$$\Gamma = [2^k 1^{N-2k}] \qquad (k = 0, \ldots, [\frac{N}{2}])$$

of the symmetric group S_N. For our example (eq. (9)) the symmetry of the spatial functions (eq. (10)) is given by the Young diagrams



The experiment shows that the ground state E_0 (the lowest eigenvalue of H) is a ^1S state. Therefore a symmetric eigenfunction belongs to E_0.
Table 2 shows that it is worth-while minimizing the variance.

Table 2

Comparison of lower bounds E_0^r and E_0^* in a.u.

$\dim V_n$	$\rho = \lambda_1^* - \sqrt{F_1^*}$	E_0^r	E_0^*	$\Delta_E = E_0^* - E_0^r$ in cm^{-1}
7	−2.35320	−2.926296582	−2.916383343	2175.00
34	−2.20449	−2.904764047	−2.904226724	117.92
95	−2.16836	−2.903875234	−2.903792612	18.13
161	−2.15711	−2.903780296	−2.903749015	6.86
203	−2.15355	−2.903760485	−2.903740122	4.47
252	−2.15105	−2.903748447	−2.903734775	3.00
444	−2.14736	−2.903732653	−2.903727879	1.05

$$E_0^r = \lambda_0^r - \frac{F_0^r}{\rho - \lambda_0^r} \quad , \quad E_0^* = \lambda_0^* - \frac{F_0^*}{\rho - \lambda_0^*}$$

The lower bounds were calculated via Temple's formula on the one hand with λ_0^r and the corresponding variance Λ_0^r, on the other hand with λ_0^* and Λ_0^*.
REMARK: In order to compare the calculated eigenvalues with the experimental results one has to take into account the conversion factor for the energy units (Hartree and cm^{-1}). It is

$$1H \;\hat{=}\; 2.1947462 \cdot 10^5 \text{ cm}^{-1} \quad .$$

Because of the experimental accuracy of ± 0.1 cm^{-1} it is necessary to calculate the eigenvalues with an absolute error less than 10^{-6}.

The method of chapter 1 was used for a systematical study of the quantum-mechanical three-body system.

A. The atomic system

The Hamiltonian for the atomic case (one nucleus, two electrons) is given (in atomic units a.u.) by

$$H = -\frac{1}{2}\frac{m}{M_K}\Delta_K - \frac{1}{2}\sum_{k=1}^{2}\Delta_k - \frac{2}{r_1} - \frac{2}{r_2} + \frac{1}{r_{12}} \quad . \tag{11}$$

The calculations were done for He, H^-, D^- and T^- in Born-Oppenheimer approximation ($M_K = \infty$) as well as for finite M_K. For the isoelectronic series up to $Z = 10$ see [7]. The quality of two-sided bounds depends essentially on the coordinates used and the type of basis functions of these coordinates. In accordance with Frankowski/Pekeris [8] and Freund/Huxtable/Morgan [9] we used Hylleraas' coordinates s, t, u as

$$s = r_1 + r_2 \quad , \quad t = r_2 - r_1 \quad , \quad u = r_{12} \tag{12}$$

and obtained $H = H_\infty + H_\kappa + V$ with

$$\begin{aligned} H_\infty &= -\frac{2}{u}\frac{\partial}{\partial u} - \frac{\partial^2}{\partial u^2} + \frac{4t}{s^2-t^2}\frac{\partial}{\partial t} - \frac{\partial^2}{\partial t^2} - \frac{4s}{s^2-t^2}\frac{\partial}{\partial s} - \frac{\partial^2}{\partial s^2} \\ &\quad + \frac{2u^2t - 2s^2t}{u(s^2-t^2)}\frac{\partial^2}{\partial u\,\partial t} - \frac{2u^2s - 2st^2}{u(s^2-t^2)}\frac{\partial^2}{\partial u\,\partial s} \qquad (13) \\ H_\kappa &= -\frac{1}{\kappa}\left(\frac{4s^2-4u^2}{s^2-t^2}\frac{\partial^2}{\partial s^2} + \frac{8s}{s^2-t^2}\frac{\partial}{\partial s} - \frac{4t^2-4u^2}{s^2-t^2}\frac{\partial^2}{\partial t^2} - \frac{8t}{s^2-t^2}\frac{\partial}{\partial t}\right) \\ V &= \frac{1}{u} - \frac{4Zs}{s^2-t^2} \\ d\tau &= \frac{1}{8}u(s^2-t^2)\,du\,dt\,ds \end{aligned}$$

with $\kappa = \frac{2M_K}{m}$ and the domain of integration is given by

$$0 \le t \le u \quad , 0 \le u \le s \quad , 0 \le s \le \infty \quad .$$

In anology to [9] our ansatz for the wave function was

$$\psi(s,t,u) = \sum_{\mathbf{k,l,m,n}}^{\infty} c_{\mathbf{k,l,m,n}}\, e^{-\alpha s} s^{\mathbf{k}} t^{2\mathbf{l}} u^{\mathbf{m}} (\ln s)^{\mathbf{n}} \, . \qquad (14)$$

The following restrictions for the extreme limits for summation were made:

$$-7 \le \mathbf{k} \le 7 \quad , \quad 0 \le \mathbf{l} \le 8 \quad , \quad 0 \le \mathbf{m} \le 11 \quad , \quad 0 \le \mathbf{n} \le 4 \quad ,$$

where $|\mathbf{k}| \le \mathbf{l}$. These conditions imply basis functions with logarithmic terms and negative powers in s. The necessity of logarithmic terms for the ground state wave functions was shown by V. Fock [10] and J.D. Morgan [11] proved the convergence of Fock's expansion. It is still an open question, whether Fock's solution of the He - eigenvalue problem, considered as a partial differential equation, contains the true ground state wave function.

Although already shown by Freund et al. [9] we found the advantage of negative **k**-values still surprising since it is known, that the true wave function is continuous and even satisfies a Hölder-condition for all arguments. The latter was first proved by F. Stummel [12]. Without the use of negative **k**-values the accuracy mentioned could not be obtained with basis sets of comparable dimension. Using the Hamiltonian H and the wave function ψ given by eqs. (13) and (14) all integrals resulting from $||H\psi||$ and $(H\psi|\psi)$ are of the type

$$I_{n,k}(\alpha) = \int_0^\infty e^{-2\alpha s} s^{k-1} \ln^n s \, ds \qquad (15)$$

with $n = 0, 1, \ldots$ and $k = 1, 2, \ldots$. A closed solution can be determined by aid of the Gamma-, the Psi- and the generalized Zeta-function, see [7]. The results for the nonrelativistic ground state energies of He, H^-, D^- and T^- are given in Table 3, where $^\infty$He and $^\infty H^-$ represent the case of Born-Oppenheimer approximation.

Table 3

Lower and upper bounds for He- and H^--isotopes at $\kappa < \infty$ and infinite κ

Species	dim	F_0^*	$\lambda_0 \in$
^{4}He	448	$1.084 \cdot 10^{-14}$	$-2.90330455747670^{70}_{55}$
H^-	402	$1.635 \cdot 10^{-11}$	-0.5274458^{16}_{09}
D^-	402	$1.638 \cdot 10^{-11}$	-0.527598^{52}_{46}
T^-	402	$1.640 \cdot 10^{-11}$	-0.52764904^{73}_{14}
$^\infty$He	448	$1.088 \cdot 10^{-14}$	-2.90372437703^{34}_{19}
$^\infty H^-$	402	$1.642 \cdot 10^{-11}$	-0.5277510^{72}_{65}

B. The molecular system

The Hamiltonian for the molecular case (two nuclei, one electron) is given (in atomic units a.u.) by

$$H = -\frac{m}{2}\sum_{a=1}^{2}\frac{\Delta_a}{M_a} - \frac{1}{2}\Delta + \frac{Z_1 Z_2}{|R_1 - R_2|} - \sum_{a=1}^{2}\frac{Z_a}{|R_a - r|} \quad . \tag{16}$$

The calculations were done for H_2^+, HD^+, D_2^+, DT^+ and HT^+ in the nonrelativistic case. For fixed nuclei (Born-Oppenheimer approximation) the results are well known because the Hamiltonian is separable in confocal elliptic coordinates. Even generally for diatomics it is of advantage to use these elliptic coordinates given by

$$\xi = \frac{r_a + r_b}{R} \quad , \quad \eta = \frac{r_a - r_b}{R} \quad , \quad \varphi$$

with $R = |R_1 - R_2|, r_a = |R_1 - r|, r_b = |R_2 - r|$, see [13]. The corresponding Hamiltonian is given by

$$H = -\frac{1}{2}\Delta_e - \left(\frac{1}{r_a} + \frac{1}{r_b}\right) + \frac{1}{R} - \frac{1}{2\mu}\Delta_R - \frac{1}{8\mu}\Delta_e - \frac{1}{2\mu_a}\nabla_e \nabla_R \tag{17}$$

with

$$\Delta_e = \frac{4}{R^2(\xi^2 - \eta^2)}\left(\frac{\partial[(\xi^2 - 1)\frac{\partial}{\partial \xi}]}{\partial \xi} + \frac{\partial[(1 - \eta^2)\frac{\partial}{\partial \eta}]}{\partial \eta}\right)$$

$$\begin{aligned}
\frac{1}{r_a}+\frac{1}{r_b} &= \frac{4\xi}{R(\xi^2-\eta^2)} \\
\Delta_R &= \left\{\frac{\partial^2}{\partial R^2}+\frac{2}{R}\frac{\partial}{\partial R}-\frac{2}{R^2(\xi^2-\eta^2)}\left(\left[\xi(\xi^2-1)\frac{\partial}{\partial\xi}+\eta(1-\eta^2)\frac{\partial}{\partial\eta}\right]\right.\right. \\
&\quad \left.\left.\cdot\left[1+R\frac{\partial}{\partial R}\right]\right)\right\}+\left[\frac{\xi^2+\eta^2-1}{R^2(\xi^2-\eta^2)}\left(\frac{\partial[(\xi^2-1)\frac{\partial}{\partial\xi}]}{\partial\xi}+\frac{\partial[(1-\eta^2)\frac{\partial}{\partial\eta}]}{\partial\eta}\right)\right] \\
\nabla_e\nabla_R &= \left[\frac{2}{R^2(\xi^2-\eta^2)}\left(\eta(\xi^2-1)\frac{\partial}{\partial\xi}+\xi(1-\eta^2)\frac{\partial}{\partial\eta}\right)\left(1+R\frac{\partial}{\partial R}\right)\right] \\
&\quad +\left[\frac{2\xi\eta}{R^2(\xi^2-\eta^2)}\left(\frac{\partial[(\xi^2-1)\frac{\partial}{\partial\xi}]}{\partial\xi}+\frac{\partial[(1-\eta^2)\frac{\partial}{\partial\eta}]}{\partial\eta}\right)\right]
\end{aligned} \tag{18}$$

with $\mu=\frac{M_1\cdot M_2}{M_1+M_2}$ and $\mu_a=\frac{M_1\cdot M_2}{M_1-M_2}$ and

$$d\tau = R^3(\xi^2-\eta^2)\,d\xi\,d\eta\,d\varphi\,dR \quad .$$

Moreover the domain of integration is given by

$$1\le\xi\le\infty \quad , \quad -1\le\eta\le 1 \quad , \quad 0\le\varphi\le 2\pi \quad , \quad 0\le R\le\infty \quad .$$

For our calculations we chose the following basis set

$$v_i = e^{-\alpha\xi}\xi^{\lambda_i}\eta^{\nu_i}e^{-\frac{c^2}{2}(R-d)^2}H_{t_i}[c(R-d)] \tag{19}$$

where H_{t_i} is a Hermite polynomial and α,λ_i,ν_i,c and d are constants. Reasons for this choice are given in [14] and [15]. The resulting integrals are elementary except

$$I=\int_1^\infty\int_{-1}^1\frac{e^{-\alpha\xi}\xi^p\eta^q}{\xi^2-\eta^2}d\eta\,d\xi$$

which is solved in [15][1].

REMARK: In the general case (heteronuclear molecules) an additional difficulty arises. Applying H from eq. (17) on the basis functions v_i (see eq. (19)), i.e. Hv_i, yields a sum of 58 terms, differing at least in one of the exponents (λ_i,ν_i) or in the order t_i of the Hermite polynomial. Thus calculating $(Hv_i|Hv_i)$ leads to 3364 terms for each element of the K-Matrix. In order to handle this enormous number of terms a computer procedure was designed in

[1] In [15] the correct form of I_∞ is

$$I_\infty=\sum_{k=0}^\infty\frac{2}{q+2k+1}\int_2^\infty e^{-\alpha\xi}\xi^{p-2-2k}d\xi \quad .$$

analogy to the method described in [15]. The results for the nonrelativistic ground states of H_2^+, D_2^+ (see [15]) and HD^+, HT^+, DT^+ (see [16]) are given in Table 4.

Table 4

Nonadiabatic lower and upper bounds for some diatomic cationic Hydrogen-molecules

Species	dim	F_0^*	$\lambda_0 \in$
H_2^+	300	$6.892 \cdot 10^{-10}$	-0.59713^{907}_{899}
HD^+	666	$4.072 \cdot 10^{-11}$	-0.5978979^{72}_{67}
HT^+	666	$4.112 \cdot 10^{-11}$	-0.59817613^{85}_{35}
D_2^+	300	$3.521 \cdot 10^{-10}$	-0.598788^{75}_{38}
DT^+	666	$3.152 \cdot 10^{-11}$	-0.59913066^{69}_{21}

3. REFERENCES

[1] H. Kleindienst, W. Altmann, Int. J. Quantum Chem. **10**, 873 (1976)

[2] G. M. Žislin, A. G. Sigalov, Amer. Math. Soc. Transl. (2) **91**, 263 (1970)

[3] J. A. Pople, D. L. Beveridge, *Approximate Molecular Orbital Theory*, 1970, page 5

[4] T. Kato, Trans. Am. Math. Soc. **70**, 195 (1951)

[5] B. Simon, Arch. Rat. Mech. Math. **52**, 44 (1973)

[6] B. Klahn, Adv. Quantum Chem. **13**, 155 (1981)

[7] H. Kleindienst, R. Emrich, Int. J. Quantum Chem. **37**, 257 (1990)

[8] K. Frankowski, C.L. Pekeris, Phys. Rev. **146**(1), 46 (1966)

[9] D.E. Freund, B.D. Huxtable, J.D. Morgan III, Phys. Rev. A **29**(2), 980 (1984)

[10] V. Fock, D K N V S Forh. Bd. **31**, No's. 22,23 (1958)

[11] J.D. Morgan III, Theor. Chim. Acta **69**, 181 (1986)

[12] F. Stummel, Math. Ann. **132**, 150 (1956)

[13] H. Eyring, J. Walter, G.E. Kimball, *Quantum Chemistry*, John Wiley and Sons, Inc. , New York, 1944, page 367

[14] D. Bishop, Molecular Physics **28**, 1397 (1974)

[15] H. Kleindienst, D. Hoppe, Theor. Chim. Acta **70**, 221 (1986)

[16] H. Kleindienst, A. Müller, Chem. Phys. Lett. **157**, 426 (1989)

Prof. Dr. Heinz Kleindienst, Dipl. Chem. Rainer Emrich,
Institut für Physikalische Chemie und Elektrochemie der Universität Düsseldorf,
Universitätsstrasse 1, D – 4000 Düsseldorf

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

A PRECONDITIONED CONJUGATE GRADIENT METHOD FOR EIGENVALUE PROBLEMS AND ITS IMPLEMENTATION IN A SUBSPACE

A. V. Knyazev

Department of Numerical Mathematics, USSR Academy of Sciences, Moscow, USSR

Abstract - We treat systematically the preconditioned steepest ascent method and the preconditioned conjugate gradient method for eigenvalue problems and present convergence rate estimates. We also suggest a modification of the methods, that makes it possible to implement them in a subspace (such as that of mesh functions, defined in the mesh-points on the dividing line for the domain decomposition methods). We discuss as an example an eigenvalue problem for $-\Delta_h$ (a mesh discretisation of Laplacian) and show that the rate of convergence does not slow as $h \longrightarrow 0$.

Introduction

The paper deals with the preconditioned conjugate gradient (CG) method for generalized eigenvalue problems and its modification for domain decomposition.

Section 1 gives two convergence rate estimates of the locally optimal variant of the method. The first of the two is just the known estimate of the preconditioned steepest ascent method. The second one is of the asymptotic character and shows the peculiarities of CG algorithm, however no proof is supplied in this paper so the asymptotic estimate could be considered as a hypothesis only.

The preconditioned steepest ascent method shows in Section 2 what changes are necessary for its implementation in a subspace for the domain decomposition technique. Then the analogous modification of the preconditioned CG method is given. It is noted that for these implementations

of the methods in a subspace the convergence rate estimates of Section 1 are valid, so it allows to state that the convergence of iterations in a subspace is not slowing as the mesh gets finer for a number of mesh problems.

The construction of the domain decomposition conjugate gradient type method for the eigenvalue problems with such a convergence property seems to be essentially new and there are no references on this subject.

Section 3 deals with the example of computing the minimal eigenvalue of the Laplace mesh operator in L-shaped domain under the first type boundary condition, for which all the required suppositions are valid.

1. THE PRECONDITIONED CONJUGATE GRADIENT METHOD FOR EIGENVALUE PROBLEMS.

We shall consider in a Euclidean space H equipped with scalar product $(\cdot,\cdot)$ an eigenvalue problem

$$Mu = \lambda Lu \ , \quad M = M^* , \ L = L^* > 0 \ . \tag{1.1}$$

We seek the maximal eigenvalue λ_1 (which is assumed to be simple) and the corresponding eigenvector. Let $\lambda(\cdot) = (M\cdot,\cdot)/(L\cdot,\cdot)$ be the Rayleigh quotient. In fact, it is implied, that the problem (1.1) is a mesh discretisation of a differential eigenvalue problem with a compact operator $L^{-1}M$. For instance, M can be an operator of multiplication by a mesh function and $L = -\Delta_h$, i.e. a mesh discretisation of Laplacian with some boundary conditions. Such a problem can hardly be solved numerically, using standard programs (such as QR algorithm) and requires the development of special iterative techniques. The Lanczos method also loses its attractivness for this problem because of necessity to solve linear systems of the type $Lu = f$ or $Mu = f$ on each iteration. Moreover, in the case $Mu = f$ the convergence of the method slows down when the mesh is getting finer.

Of primary importance are iterative methods which require the solution on each iteration of an auxiliary system of simultaneous equations $Bu = f$ with a preconditioner $B = B^* > 0$. In general case, B does not coincide with any linear combination of M and L.

If B is a 'good' preconditioner, i.e. the constant δ^* from the inequalities

$$0 < \delta_0^* B \leqslant L \leqslant \delta_1^* B \ , \quad \delta^* \equiv \delta_0^*/\delta_1^*$$

is close to 1 (does not depend of the mesh size h for grid problems), and the system $Bu = f$ can be solved sufficiently effectively, then these preconditioned methods allow us to compute the eigenpair we seek at almost optimal computational cost, see for instance [1].

Some of these methods can be treated as a gradient type maximization of the Rayleigh quotient with the gradient

$$\operatorname{grad} \lambda(u) = 2B^{-1}\{Mu - \lambda(u)Lu\}/(Lu, u)$$

to be computed in a special scalar product $(\cdot,\cdot)_B = (B\cdot,\cdot)$. The preconditioned conjugate gradient method

$$p^n = g^n + \beta^n p^{n-1}, \quad u^{n+1} = u^n + \alpha^n p^n, \quad n = 0,1,\ldots$$

where $g^n = \operatorname{grad} \lambda(u^n)$, the parameters α^n are chosen to ensure the maximization of the value $\lambda(u^{n+1})$, while the parameters β^n (except $\beta^0 = 0$) can be chosen in several different ways (see, for example [2,3]), is of great interest. Evidently, the function $\lambda(\cdot)$ to be maximized in these methods is not quadratic, and the method described lacks many properties of the traditional conjugate gradient method. For example, various versions of the method are not found to be equivalent to each other, and none of them seems to achieve a global maximization of the Rayleigh quotient in general case, even when $B = L = I$. In the author's opinion the most favourable variant of the method is the so-called 'locally optimal' scheme, which is given by

$$\lambda^n \equiv \lambda(u^n), \quad u^{n+1} = B^{-1}(\lambda^n Lu^n - Mu^n) + \alpha^n u^n + \beta^n u^{n-1}. \tag{1.2}$$

Both α^n and β^n are chosen to maximize $\lambda(u^{n+1})$, $\beta^0 = 0$ (see [4-6]).

For the last method some important results concerning its global behaviour can be proved by comparing this method with the well-known [1,5,6] preconditioned steepest ascent technique (formula (1.2) with all $\beta^n \equiv 0$).

Theorem 1.1. Let us choose a number $p \geq 1$, $p \equiv p(n)$ such that in (1.2) $\lambda_{p+1} < \lambda^n \leq \lambda_p$. Then either $\lambda^{n+1} > \lambda_p$, $p \neq 1$, or

$$0 \leq \frac{\lambda_p - \lambda^{n+1}}{\lambda^{n+1} - \lambda_{p+1}} \leq \left\{ 1 - \delta^* \frac{\lambda_p - \lambda_{p+1}}{\lambda_p - \lambda_{\min}} \right\} \frac{\lambda_p - \lambda^n}{\lambda^n - \lambda_{p+1}}.$$

Specifically, for $\lambda^0 > \lambda_2$ the value $(\lambda_1 - \lambda^n)/(\lambda^n - \lambda_2)$ decreases at least as fast as a geometric progression with the ratio $1 - \delta^*(\lambda_1 - \lambda_2)/(\lambda_1 - \lambda_{\min})$.

We have introduced here the notation for eigenvalues which is unlikely to require any comments.

Proof. The assertion of the theorem for the case $\beta^n \equiv 0$ (when (1.2) becomes the steepest ascent method) was proved in [5]. In (1.2) β^n is choosen to maximize $\lambda(u^{n+1}) \equiv \lambda^{n+1}$ so the quotient $(\lambda_p - \lambda^{n+1})/(\lambda^{n+1} - \lambda_{p+1})$ for the method (1.2) is not greater than for the previous case $\beta^n \equiv 0$.

It is very important that the convergence rate estimate in Theorem 1.1 is not asymptotic. This allows us to state the fast convergence of (1.2) for the fine mesh problem (1.1).

Local properties of the conjugate gradient method are described by

Proposition 1.1. Let $\{u^0\}$ be a set of initial guesses, such that $\lambda^0 \equiv \lambda(u^0) \to \lambda_1$. Then the asymptotic estimate

$$0 \leqslant \lambda_1 - \lambda^n \leqslant \left\{ \left| T_n\left(\frac{1+\xi}{1-\xi} \right) \right|^{-2} + o(1) \right\} (\lambda_1 - \lambda^0) \tag{1.3}$$

$$n = 1, 2, \ldots, \quad \xi = \delta^*(\lambda_1 - \lambda_2)/(\lambda_1 - \lambda_{\min})$$

holds for (1.2), where $T_n(\cdot)$ is the Chebyshev polynomial of the degree n and $o(1) \to 0$ as $\lambda^0 \to \lambda_1$.

For $n = 1, 2$ this estimate was proved in [5] with $O(\sqrt{\lambda_1 - \lambda^0})$ except $o(1)$.

The proposition can be considered as a consequence of a more profound fact, namely, the asymptotic equivalence (under $\lambda^0 \to \lambda_1$) of the method (1.2) and the traditional conjugate gradient method for the solution of the linear system $B^{-1}\{\lambda_1 L - M\}u = 0$.

The functionals to be minimized in these methods, specifically,

$$\lambda_1 - \lambda(u) \quad \text{and} \quad (\{\lambda_1 L - M\}u, u)$$

are asymptotically in close proximity to each other, provided that the vector u is properly normalized [5,6].

It is beyond the scope of this paper to investigate the asymptotic behaviour of the conjugate gradient method and to prove Proposition 1.1 as well, so it could be considered as a hypothesis.

The simplest way to reformulate Proposition 1.1 in terms of $n \to \infty$ instead of $\lambda^0 \to \lambda_1$ is to consider method (1.2) with stopping and restarting after each k iterations, putting $\beta^n = 0$. It is the so-called 'cyclic' conjugate gradient method. If the quantity λ^n is frosen within each k-iteration cycle we arrive at the two-stage iterative method analogous to one in the book [5] where it is treated theoretically.

It is still unclear whether the stopping and restarting procedure is good in practice.

The results mentioned above are likely to hold for other versions of the conjugate gradient method (cf. [7]).

2. IMPLEMENTATION OF THE MODIFIED PRECONDITIONED CONJUGATE GRADIENT METHOD IN A SUBSPACE.

Let the original space H be represented as a sum of two orthogonal subspaces H_1 and H_2, i.e.

$$H = H_1 \oplus H_2 .$$

This decomposition of H leads to the unique representation of vectors $u \in H$:

$$u = u_1 + u_2 \equiv \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \equiv (u_1; u_2)^T , \quad u_1 \in H_1, \ u_2 \in H_2$$

and linear operators A: $H \to H$:

$$A = \begin{bmatrix} A_1 & A_{12} \\ A_{21} & A_2 \end{bmatrix}$$

$$A_i : H_i \to H_i , \quad A_{ij} : H_j \to H_i , \quad i,j = 1,2, \ j \neq i .$$

Definition. Matrix S_A, defined by the formula

$$S_A \equiv A_2 - A_{21}A_1^{-1}A_{12}$$

is called a Schur complement of A_1 in A, if A_1 is invertible. Now consider the problem (1.1), making use of the block representation of M and L. Assuming, that the value λ_1 is known, it can be easily seen that the component u_2 of eigenvector satisfies the equation $S_{\lambda\, L-M}\, u_2 = 0$, which can be solved using some preconditioned iterative method, for instance, the simple iterative method

$$u_2^{n+1} = u_2^n - \gamma S_B^{-1} S_{\lambda_1 L-M}\, u_2^n .$$

Suppose, that the multiplication by the operator $S_{\lambda\, L-M}$ can be done using only elements from H_2. Then the method of this kind is precisely method in the subspace H_2.

The theory of iterative methods in subspace, when applied to solving systems of simultaneous equations is well-developed [8-16]. These methods are highly effective in solving Poisson's equation in regions composed of slabs. For this case algorithms of 'partial solutions' that enable to implement the multiplication by Schur complement in subspace, have been worked out. 'Good' preconditioners S_B are known as well, so the rate of convergence does not decrease as a mesh size tends to zero, i.e. as mesh becomes finer.

When applied to ordinary eigenvalue problems, similar results have been obtained recently [17]. In previous papers [18,19] the reported convergence rate worsens as mesh grows finer.

In this section the implementation of the CG method in a subspace is derived for solving (1.1). Necessary details can be found in [17], where implementation of simpler methods, such as the steepest ascent method, in a subspace is proposed.

Let parameter λ satisfy the condition $\lambda \geqslant \bar{\lambda}$, where $\bar{\lambda}$ is the maximal eigenvalue of the eigenvalue problem $M_1 u_1 = \bar{\lambda} L_1 u_1$. We denote $S_\lambda \equiv S_{\lambda L-M}$ and set

$$B \equiv B(\lambda) \equiv \lambda L - M + \begin{bmatrix} 0 & 0 \\ 0 & S_B - S_\lambda \end{bmatrix}$$

where $S_B: H_2 \to H_2$ is an operator, $S_B = S_B^* > 0$. Let us note, that $B = B^* > 0$ and the Schur complement of B_1 in B is equal to S_B.

At first, following [17], we consider the steepest ascent method. In order to implement this method in a subspace, it is sufficient to modify it as follows:

Modified Preconditioned Steepest Ascent.

Step 1. Choose vector $u^0 \neq 0$ such, that the inequality

$$\lambda^0 \equiv \lambda(u^0) > \bar{\lambda}$$

is valid.

Step 2. For $n = 0,1,\ldots$:

(a) For the second component u_2^n of u^n, set

$$v_{\lambda^n} \equiv (-(\lambda^n L_1 - M_1)^{-1}(\lambda^n L_{12} - M_{12})u_2^n; u_2^n)^T .$$

(b) Construct the spectral problem

$$\hat{M}X = \nu\hat{L}X , \quad X \in \mathbb{R}^2$$

with the 2×2 matrices $\hat{M}$ and $\hat{L}$, defined by

$$\hat{M} \equiv \{(Mw_i, w_j)\}$$

$$\hat{L} \equiv \{(Lw_i, w_j)\}$$

$$i,j = 1,2$$

where

$$w_1 \equiv v_{\lambda^n}$$

$$w_2 \equiv B^{-1}(\lambda^n)(\lambda^n L - M)v_{\lambda^n} .$$

Compute its maximal eigenvalue ν and corresponding eigenvector $X = (x_1, x_2)^T$. Compute $u^{n+1} = x_1 w_1 + x_2 w_2$.

(c) Set $\lambda^{n+1} \equiv \nu = \lambda(u^{n+1})$.

The formulated method differs from the traditional one in the following.

Firstly, the parameter λ^n in formula for w_2 does not coincide in general with Rayleigh quotient $\lambda(v_\lambda)$. One can only prove that the inequality

$\lambda^n \leq \lambda(v_\lambda)$ holds [17]. By that, vector w_2 can not be equal to $B^{-1}\{\lambda(v_\lambda)L - M\}v_\lambda$, i.e. w_2 is not parallel to **grad** $\lambda(v_\lambda)$ and the method suggested can not, strictly speaking, be called the steepest ascent method (along gradient).

Secondly, in (a) traditionally it is set $v_\lambda = u^n$.

Due to these peculiarities alongside with special form of the preconditioner B one can assert, that

$$w_1 \equiv v_{\lambda^n}, \quad w_2 \text{ and } u^{n+1} \in B^{-1}(\lambda^n)H_2 .$$

Now it is sufficient to notice, that the subspace $B^{-1}(\lambda^n)H_2$ is invariant with respect to operator $B^{-1}(\lambda^n)(\lambda^n L - M)$ and elements of matrices $\hat{M}$ and $\hat{L}$ can be computed using only the second components of vectors w_1 and w_2, which are equal to u_2^n and $S_B^{-1}S_\lambda u_2^n$, respectively. Then the implementation of the proposed method in the subspace H_2 reduced to the 'damping' of the first components of all the vectors involved.

This implementation we will call the modified preconditioned steepest ascent method in a subspace. For this method the assertion of Theorem 1.1 with the quantity δ^* for the operator B from (2.1) is valid, as it follows from [20], see [17].

In the case, when $M = I$ in paper [17] an explicit lower bound for δ^* in terms of λ^n, λ_1, $\bar{\lambda}$ and **cond** $S_B^{-1}S_L$ is obtained. The estimate enables us to assert, that as the mesh becomes finer, the quantity δ^* does not tend to zero, i.e. the convergence rate does not get slower.

Let us proceed now with the CG-method. As well as the steepest ascent method, it can not be implemented in a subspace directly, some modifications are needed.

The Modified Preconditioned Conjugate Gradient Method.

Step 1. See Step 1 of the previous method.

Step 2. For $n = 0$ coincides with Step 2 with $n = 0$ of the previous method.

Step 3. For $n = 1,2,\ldots$:

(a) Set

$$v_{\lambda^n} \equiv (-(\lambda^n L_1 - M_1)^{-1}(\lambda^n L_{12} - M_{12})u_2^n; u_2^n)^T .$$

(b) Compose a generalized eigenvalue problem

$$\hat{M}X = \nu\hat{L}X\ , \quad X \in \mathbb{R}^3$$

with the **3×3** matrices $\hat{M}$ and $\hat{L}$

$$\hat{M} \equiv \{(Mw_i, w_j)\}$$

$$\hat{L} \equiv \{(Lw_i, w_j)\}$$

$$i, j = 0, 1, 2$$

where

$$w_1 \equiv v_{\lambda^n}$$

$$w_2 \equiv B^{-1}(\lambda^n)(\lambda^n L - M)v_{\lambda^n}\ .$$

and

$$w_0 \equiv (-(\lambda^n L_1 - M_1)^{-1}(\lambda^n L_{12} - M_{12})u_2^{n-1}; u_2^{n-1})^{\mathrm{T}}$$

Compute its maximal eigenvalue ν and corresponding eigenvector $X \equiv (x_0; x_1; x_2)^{\mathrm{T}}$. Compute vector $u^{n+1} = x_0 w_0 + x_1 w_1 + x_2 w_2$.

(c) Set $\lambda^{n+1} \equiv \nu = \lambda(u^{n+1})$.

There is only one difference between the described modified CG method and the traditional generalized method with preconditioner $B(\lambda^n)$ additional to those between the modified and traditional steepest ascent pointed out earlier. Namely, on step 3 (b) one traditionally sets $w_0 \equiv u^{n-1}$. In the proposed modification

$$w_0, w_1 \equiv v_{\lambda^n}\ , \quad w_2 \quad \text{and} \quad u^{n+1} \in B^{-1}(\lambda^n)H_2\ .$$

Then, to compute scalar products on step 3 (b) one should know only the second component of the vectors w_0, w_1 and w_2, which are equal to u_2^{n-1}, u_2^n, and $S_B^{-1}S_\lambda u_2^n$, respectively. Consequently, the first components can be brushed off as earlier and we obtain the modified preconditioned conjugate gradient method in a subspace. Evidently, for this method, as well as for the modified preconditioned steepest ascent method, Theorem 1.1 holds. Proposition 1.1 is likely to be valid too, since when $\lambda^0 \to \lambda_1$ the traditional and the modified methods are asymptotically equivalent.

If several eigenvalues are needed, block versions of the steepest ascent conjugate gradient methods can be used [2,5,6,10]. After appropriate modification these methods can be implemented in a subspace too.

3. EXAMPLES

The test problem is the model problem with $M = I$ and $L = -\Delta_h$, where Δ_h is a mesh approximation of Laplacian with homogeneous Dirichlet boundary conditions on the L-shaped region, composed of three unit squares. We use the standard five point stencil on a uniform square mesh [21].

We use the domain decomposition method, cutting off one of the side squares. The number of mesh points on the dividing line we denote N. We decompose the space H, consisting of mesh functions, in orthogonal sum $H = H_1 \oplus H_2$, where H_1 consists of the functions, vanishing in the mesh points on the dividing line. Then H_2 consists of the functions, vanishing in all the mesh points but the mesh points on the dividing line, and $\dim H_2 = N$. Under this decomposition of H block L_1 in the matrix L is a 2×2 block-diagonal matrix, its diagonal blocks being mesh approximation of Laplacian in a rectangle with the number of internal nodes $2N\times N$ and on a unit square with the number of nodes $N\times N$.

The preconditioner S_B can be chosen for instance,

$$S_B = (N+1)^2\{\Lambda + \Lambda^2/4\}^{1/2}, \quad \text{where} \quad \Lambda \equiv \operatorname{tridiag}\{-1;2;-1\}_1^N .$$

Then the quantity $\operatorname{cond} S_B^{-1}S_L$ has as an upper bound a constant, independent of N [16].

Clearly, the convergence of all methods described in this paper does not deteriorate, as $N \longrightarrow \infty$.

Multiplication by the Schur complement S_λ in this example can be performed making use of algorithm of partial solutions, that requires $O(N \ln N)$ arithmetic operations and only N memory locations. Algorithms with similar characteristics are known to exist [17], that enables to compute scalar products (needed to form the elements of matrices $\hat{M}$ and $\hat{L}$) with vectors from the subspace $B^{-1}(\lambda^n)H_2$, when only the second components are known. Thus, to carry out one iteration of the modified steepest ascent and

conjugate gradient methods in a subspace only $O(N \ln N)$ arithmetic operations and N memory locations are needed. This is also true for general case of domain decomposition method for domains composed of rectangles except one multiplication $S_\lambda u_2$ costs $O(N^2)$ operations.

The numerical test was carried out by my colleague Dr. A. L. Skorokhodov for N from 1 to 8191 using the method of steepest ascent. For various appropriate choices of the initial vector, the rate of convergence to an eigenvalue was typically found to be a gain of two decimal places per iteration, independent of N . Therefore we did not apply the conjugate gradient method to accelerate convergence.

For $N = 8191$ the time cost of one iteration was two minutes for the computer HP-3000 (double precision), and after five iterations the eigenvalue approximation **1/9.639737073** has been achieved.

For $N = 29$ it is equal to **1/9.657433677** as in [21].

REFERENCES

1. D'yakonov E. G. and Orekhov M. Yu. (1980) Minimization of the computational labor in determining the first eigenvalues of differential operators. *Math. Notes*, **27**, 382-391.

2. Bradbury W. W. and Fletcher R. (1966) New iterative methods for solution of eigenproblem. *Num. Math.*, **9**, 259-266.

3. Cullum J. K. and Willoughby R. A. (1985) Lanczos Algorithms for Large Symmetric Eigenvalue Computations. Vol. 1. (Birkhäuser, Boston).

4. D'yakonov E. G. and Orekhov M Yu. (1983) On some iterative methods for eigenvalue problems. *Math. Notes*, **34**, 879-912.

5. Knyazev A. V. (1986) Calculation of eigenvalues and eigenvectors in mesh problems: algorithms and error estimates. Dept. Numer. Math., USSR Acad. Sci., Moscow (in Russian).

6. Knyazev A. V. (1987) Convergence rate estimates for iterative methods for a mesh symmetric eigenvalue problem. *Sov. J. Numer. Anal. Math. Modelling* **2**, 371-396.

7. Savinov G. V. (1979) Studing of the convergence rate of the preconditioned conjugate gradient method for some algebraic problems. *Ph. D. Th.*, Shipbuilding Inst., Leningrad (in Russian).

8. Kuznetsov Yu. A. (1985) Computational methods in subspaces. In: *Vychisl. Protsessy i Sistemy*, Nauka, Moscow, **2**, 265-350 (in Russian).

9. Marchuk G. I., Kuznetsov Yu. A. and Matsokin A. M. (1986) Fictitious

domain and domain decomposition methods. *Sov. J. Numer. Anal. Math. Modelling* **1**, 1-82.

10. Kuznetsov Yu. A. and Trufanov O. D. (1987) Domain decomposition methods for the wave Helmholtz equation. *Sov. J. Numer. Anal. Math. Modelling* **2**, 113-136.

11. Matsokin A. M. and Nepomnyashchikh S. V.(1983) An employment of the bordering for solving mesh simultaneous equations. In: *Vychisl. algoritmy v zadachah matem. phys.* (Novosibirsk, USSR Acad. Sci.), 99-109 (in Russian).

12. Lebedev V. I. (1986) Composition method. Dept. Numer. Math., USSR Acad. Sci., Moscow (in Russian).

13. Bramble J. H., Pasciak J. E. and Schatz A. M. (1986) An iterative method for elliptic problems on regions partitioned into substructures. *Math. Comp.* **46**, 361-370.

14. Bjorstad P. E. and Widlund O. B. (1986) Iterative methods for the solution of elliptic problems on regions partitioned into substructures. *SIAM. J. Numer. Anal.* **23**, 1097-1120.

15. Dryja M. (1984) A finite element - capacitance matrix method for elliptic problems on regions partitioned into substructures. *Numer. Math.* **44**, 153-168.

16. Golub G. H. and Meyers D. (1983) The use of preconditioning over irregular regions. *Proc. 6th. Intern. Conf. Comput. Meth. Sci. Engng.*, (Versailles, France).

17. Knyazev A. V. and Skorokhodov A. L. (1989) Preconditioned iterative methods in subspace for solving linear systems with indefinite coefficient matrices and eigenvalue problems. *Sov. J. Numer. Anal. Math. Modelling* **4**, 283-310.

18. Kuznetsov Yu. A. (1986) Iterative methods in subspaces for eigenvalue problems. In: *Vistas in Applied Math.*, (Opt. Soft. Inc., New York), 96-113.

19. Bespalov A. N. and Kuznetsov Yu. A. (1987) RF cavity computations by the domain decomposition and fictitious components methods. *Sov. J. Numer. Anal. Math. Modelling* **2**, 259-278.

20. D'yakonov E. G. and Knyazev A. V. (1982) Group iterative method for finding lower-order eigenvalues. Moscow Univ., Ser.15, Comput. Math. Cyber, **2**, 32-40.

21. Forsythe G. E. and Wasow W. R. (1960) Finite-difference methods for partial differential equations. J. Wiley Inc., New York.

Dr. A. V. Knyazev, Department of Numerical Mathematics, USSR Academy of Sciences. Ryleev str.,29, Moscow, 119034, USSR.

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

AGGREGATION METHODS OF COMPUTING STATIONARY DISTRIBUTIONS OF MARKOV PROCESSES

Ivo Marek
Charles University
Prague, Czechoslovakia

Dedicated to Prof. Dr. Lothar Collatz, Dr.h.c., on the occasion of his 80 th birthday.

Abstract. A class of general aggregation methods is presented and analyzed in order to compute elements $x \in K$ such that $Tx = x$ and $[x,\hat{x}'] = 1$, where K is a closed normal generating cone in a Banach space E and $\hat{x}'$ is a strictly positive linear form on E.

1. Definitions and Notation

Let E and F be real Banach spaces, E' and F' be their duals respectively. Let $B(E,F)$ be the space of bounded linear operators mapping E into F. In particular, $B(E,E) = B(E)$. If $T \in B(E,F)$ then T' denotes the dual of T, i.e. $T' \in B(F',E')$.

Let E and F be generated by closed normal cones K and H respectively (see [2]). In such case the dual space E' is generated by the dual cone K', where $K' = \{x' \in E' : [x,x'] \geq 0 \text{ for all } x \in K\}$. Let us denote by K^d the dual interior of K i.e. $K^d = \{x \in K : [x,x'] > 0$ for all $x' \in K'$, $x' \neq 0\}$. Note that $K^d = \text{Int } K$ whenever $\text{Int } K \neq \emptyset$. It may happen however that K^d is not empty whilst $\text{Int } K$ is.

An operator $T \in B(E)$ is called *K-nonnegative*, if $TK \subset K$ ([2]).

A K-nonnegative operator $T \in B(E)$ is called *K-irreducible*, if for every pair $x \in K$, $x \neq 0$ and $x' \in K'$, $x' \neq 0$, there is an index $p = p(x,x')$ such that $[T^p x, x'] > 0$ ([9]). An element $x' \in K'$, is called *strictly positive*, if $[x,x'] > 0$ whenever $x \in K$, $x \neq 0$.

An operator $T \in B(E)$ is called *average convergent*, if the sequence $\{S_N\}$ is convergent, where

$$S_N = N^{-1} \sum_{k=1}^{N} T^k , \quad N = 1,2,\ldots .$$

A K-nonnegative operator T is called *Markov operator*, or more precisely, *K-Markov operator corresponding to element* $\hat{x}'$, if $\hat{x}' \in K'$ is strictly positive and $T'\hat{x}' = \hat{x}'$.

Let $\hat{E}$ denote the complex extension of E, i.e. $\hat{E} = E + iE = \{z = x + iy, \ x,y \in E\}$ with the norm

$$\|z\|_{\hat{E}} = \sup\{\|x \cos \vartheta + y \sin \vartheta\|_E : 0 \leq \vartheta \leq 2\pi\}$$

and let $\hat{E}'$ denote its dual.

If E is a Hilbert space then we can extend the E-inner product to $\hat{E}$ as follows. Let $z = x + iy$ and $w = u + iv$, where $x, y, u, v \in E$. We set

$$[z,w] = (x,u) + (y,v) + i[(y,u) - (x,v)] .$$

In this manner $\hat{E}$ becomes a Hilbert space as well.

Let $T \in B(E)$. We construct a complex extension $\hat{T}$ of T by setting

$$\hat{T}z = Tx + iTy , \text{ where } x, y \in E .$$

Let $T \in B(\hat{E})$ (we may write also $\hat{T}$ in place of T) and let λ be a complex scalar. Let $R(\lambda,T) = (\lambda I - T)^{-1}$ be the resolvent operator of T. The set $\{\lambda \in C : (\lambda I - T)^{-1} \in B(\hat{E})\}$ is called *resolvent set of* T and is denoted by $\rho(T)$, where C denotes the field of complex numbers. The complement of $\rho(T)$ is called *spectrum of* T and is denoted by $\sigma(T)$. We let $r(T) = \text{Max}\{|\lambda| : \lambda \in \sigma(T)\}$ and call it *spectral radius of* T. In particular, if $T \in B(E)$, we let by definition $\sigma(T) = \sigma(\hat{T})$.

We say that operator $T \in B(E)$ *has property "p"*, if every $a \in \sigma(T)$,

$|a| = r(T)$, is a pole of the resolvent operator $R(\lambda,T)$. It is well known [10, p. 305] that

$$(1.1) \qquad (\lambda I - T)^{-1} = \sum_{k=1}^{\infty} A_k(a)(\lambda - a)^k + \sum_{k=1}^{q(a)} B_k(a)(\lambda - a)^{-k}$$

where $A_k(a)$ and $B_{k+1}(a)$, $k = 0,1,2,\ldots$ belong to $B(\hat{E})$ and $q(a) < +\infty$ is the multiplicity of a as a pole of the resolvent operator. It is also known [10, p. 306] that the operators $B_k(a)$ in the main part of the Laurent series (1.1) satisfy the identities

$$(1.2) \qquad B_1^2(a) = B_1(a) \ , \quad B_{k+1}(a) = (T - aI)B_k(a) \ , \quad k = 1,2,\ldots \ .$$

An operator $T \in B(\hat{E})$ is said to be *Radon - Nikolskii operator*, if $T = U + V$, where U and V belong to $B(\hat{E})$, U is compact and $r(V) < r(T)$ ([6]).

The cone K converts E into a partially oredered space by letting $x \leqq y$, or equivalently, $y \geqq x$ if and only if $(y - x) \in K$, $x, y \in E$. In a similar manner $x' \leqq y'$ $(y' \geqq x')$ iff $[x,x'] = x'(x) \leqq y'(x) = [x,y']$ for all $x \in K$, $x', y' \in E'$. Similarly for T and S in $B(E)$ let $S \leqq T$ $(T \geqq S)$ iff $(T - S)K \subset K$.

A K-nonnegative operator $T \in B(E)$, $TK \subset K$, has *property "d"* if there exists an $x' \in K'$, $x' \neq 0$ such that $T'x' = r(T)x'$.

Remarks. Obviously, if $T \in B(E)$, $TK \subset K$, has property "p" then T has property "d" as well.

On the other hand, T may have property "d" but not property "p" as the next example shows.

Let $E = C([0,1])$, $K = \{x \in C([0,1]) : x(s) \geqq 0 \ , \ s \in [0,1]\}$ and let $T \in B(C([0,1]))$ be defined by

$$(Tx)(s) = sx(s) \ , \quad s \in [0,1] \ .$$

It is easy to see that $r(T) = 1$ and that value 1 is not an eigenvalue of T , but x' , defined by setting $x'(x) = x(1)$ (the Dirac functional), is an eigenelement of T' : $[x, T'x'] = [Tx, x'] = [x,x']$, $x \in C([0,1])$.

Let T be a Markov operator corresponding to the element $\hat{x}' \in K'$. A vector $x \in K$ is called *stationary distribution of* T if $Tx = x$ and $[x,\hat{x}'] = 1$.

2. Auxiliary Assertions

2.1 Lemma. *Let* $T \in B(E)$ *be K-nonnegative and let* T *have property "d". Let* $u \in K$, $u \neq 0$ *be such that*

$$Tu \leqq au \tag{2.1}$$

for some $a > 0$. *Let* $x_0' \in K'$, *be such that* $T'x_0' = r(T)x_0'$. *If*

$$[u,x_0'] \neq 0 \tag{2.2}$$

then

$$r(T) \leqq a . \tag{2.3}$$

In particular, relation (2.2) *holds whenever either* (i) u *in* (2.1) *belongs to* K^d *or* (ii) x_0' *is a strictly positive element of* K'.

If at least one condition of the above alternative holds, $a = r(T)$ *and* $r(T)$ *is a pole of the resolvent operator* $R(\lambda,T)$, *then* $a = r(T)$ *is a simple pole.*

Proof. It follows from (2.1) that $[Tu,x_0'] = a[u,x_0']$ and because of (2.2) relation (2.3) holds.

Let $a = r(T)$ be a pole of the resolvent operator and let $q = q\big(r(T)\big) > 1$ be its multiplicity. Let $B_q = B_q\big(r(T)\big)$ be the leading term in the main part of the Laurent expansion (1.1) about $a = r(T)$. It is known that $B_q K \subset K$ ([5]).

In case (i) $r(T)u = Tu$. Let $y' \in K'$ be such that $B_q'y' = x_0'$. Then because of (1.2)

$$0 < [u, B_q'y'] = [Tu - r(T)u, B_{q-1}'y'] = 0$$

a contradiction.

In case (ii) $T'x_0' = r(T)x_0'$. Let $y \in K$ be such that $B_q y = u \neq 0$. Then by (1.2)

$$0 < [B_q y, x_0'] = [B_{q-1}y, T'x_0' - r(T)x_0'] = 0$$

again a contradiction completing the proof of Lemma 2.1.

2.2 Corollary. *Let* $T \in B(E)$ *be a Markov operator. Let value* 1 *be a pole of the resolvent operator. Then* $r(T) = 1$ *and this value is a simple pole of the resolvent operator.*

3. Aggregation

Let the spaces E and F be generated by the cones K and H respectively. It is assumed that both K and H are closed and normal. Let $R \in B(E,F)$ and let

$$D = \{x \in K : Rx \in K^d\} .$$

Let $S(u) \in B(F,E)$, $u \in D$. The map R is called *reduction operator* and $S(u)$ *prolongation operator* if

(3.1) $$RS(u) = I , \quad u \in D .$$

From definition it follows that $P(u) \in B(E)$, where

(3.2) $$P(u) = S(u)R , \quad u \in D ,$$

is a projection operator.

Let $T \in B(E)$ and let $TK \subset K$. We let

$$A(u) = RTS(u) , \quad u \in D ,$$

and call it *aggregation operator*. Obviously, $A(u) \in B(F)$. It is assumed that

(3.3) $$RK \subset H$$

and

(3.4) $$S(u)H \subset K , \quad u \in D$$

and this implies that

$$A(u)H \subset H .$$

Let T be a Markov operator corresponding to element $\hat{x}'$. We assume that

(3.5) $$P(u)u = u , \quad u \in D ,$$

and

$$[P(u)x, \hat{x}'] = [x, \hat{x}'] , \quad u \in D . \tag{3.6}$$

3.1 Lemma. *Let conditions* (3.1) *through* (3.6) *hold. Then there exists* $\tilde{p} \in H$ *such that*

$$RTS(u)\tilde{p} = \tilde{p}$$

and

$$[\tilde{p}, \tilde{x}'] = [S(u)\tilde{p}, \tilde{x}'] = 1 , \quad u \in D ,$$

where $\tilde{x}'$ *is strictly positive on* H .

Proof. Let $\tilde{x}' = [S(u)]'\hat{x}'$. Then by our hypotheses

$$[A(u)]'\tilde{x}' = \tilde{x}' .$$

Moreover, for $0 \neq w \in H$ we have that

$$[w, \tilde{x}'] = [S(u)w, \hat{x}'] > 0$$

and this completes the proof.

To compute stationary distributions of Markov operators we present the following aggregation - disaggregation method.

Let $L \in B(E)$ and $\varepsilon > 0$ be given. The algorithm is as follows.

1° Choose an approximation $u^k \in D$, $k = 0$.
2° Construct the aggregation operator $A(u^k) = RTS(u^k)$.
3° Compute z^{k+1} as a solution to $z = A(u^k)z$, $[S(u^k)z, \hat{x}'] = 1$.
4° Let $v^{k+1} = S(u^k)z^{k+1}$.
5° Perform a relaxation with L : $u^{k+1} = Lv^{k+1}$.
6° Test whether $\|u^{k+1} - u^k\| < \varepsilon$.
7° If NO in 6° then $k + 1 \to k$ and GO TO 2°. If YES in 6° then GO TO 8°.
8° Stop.

Remarks. The presentation given here is a generalized version of a method published in [4] for the case of solving input - output systems of equations of the form $x = Ax + b$ with an $n \times n$ Leontieff matrix A . A similar method is applied to compute stationary distributions of finite Markov chains in [8].

3.2 Lemma. *Let* $Q \in B(E)$, $QK \subset K$ *and let* $r(P(u)Q) < 1$ *for* $u \in D$. *Then for* $u \in D$

$$[RQS(u)]^k = R[P(u)Q]^k S(u) \tag{3.7}$$

and

$$S(u)[I - RQS(u)]^{-1}R = [I - P(u)Q]^{-1}P(u) , \quad k = 1,2,\ldots . \tag{3.8}$$

Proof. Easy.

3.3 Lemma. *Let the following hypotheses hold.*

1. $Q \in B(E)$, $QK \subset K$, $r(Q) < 1$.
2. $r(P(u)Q) < 1$, $u \in D$.
3. *There exists an operator* $C \in B(E,F)$ *such that* $L = Q + CR(I - Q)$.

Then

$$L[I - P(u)Q]^{-1}[I - P(u)] = Q[I - P(u)Q]^{-1}[I - P(u)] .$$

Proof. We see that

$$Q[I - P(u)Q]^{-1}[I - P(u)] =$$
$$[L - CR(I - Q)][I - P(u)Q]^{-1}[I - P(u)] .$$

However, since $R(I - P(u)) = 0$,

$$R(I - Q)[I - P(u)Q]^{-1}[I - P(u)] =$$
$$R[I - P(u)Q + P(u)Q - Q][I - P(u)Q]^{-1}[I - P(u)] =$$
$$R[I - P(u)] = 0 .$$

The proof is complete.

3.4 Lemma. *Let* $p \in K$ *be a stationary distribution of a Markov operator* T *corresponding to element* $\hat{x}'$. *Let* $Q \in B(E)$ *and* $b \in E$ *satisfy the following hypotheses.*

1. $QK \subset K$, $r(Q) < 1$.
2. $r(P(u)Q) < 1$.
3. $p - Qp = b$.
4. *The relations*

$$RTS(u)\tilde{p} = \tilde{p} , \quad \tilde{p} \in H , \quad [S(u)\tilde{p}, \hat{x}'] = 1 ,$$

where $\hat{x}'$ *comes from the Markov property of* T, *imply that*

$$\tilde{p} - RQS(u)\tilde{p} = Rb \ .$$

Then

$$p^N = L[I - P(p^S)Q]^{-1}P(p^S)b \ , \tag{3.9}$$

where $p^S \in D$ *and* $p^N = LS(p^S)\tilde{p}$.

Proof. We see that for $u = p^S$, $[I - P(p^S)Q]^{-1}Rb = \tilde{p}$ and by (3.7) and (3.8) the required relation follows.

3.5 Lemma. *Let* $u \in D$ *and let* p *be a stationary distribution of a Markov operator corresponding to element* $\hat{x}' \in K'$ *and let* T *have property* "p" . *Furthermore, let*

1. $Q \in B(E)$, $QK \subset K$, $r(Q) < 1$.
2. $r\big(P(u)Q\big) < 1$, $u \in D$.
3. $p - Qp = b$.
4. $Lp = p$.

Then

$$p = L[I - P(u)Q]^{-1}P(u)b + L[I - P(u)Q]^{-1}[I - P(u)]p \ .$$

Proof. We see that

$$\begin{aligned} p = Lp &= L[I - P(u)Q]^{-1}[p - P(u)Qp] = \\ &L[I - P(u)Q]^{-1}[P(u)(I - Q) + p - P(u)p] = \\ &L[I - P(u)Q]^{-1}\{P(u)b + [p - P(u)p]\} = \\ &L[I - P(u)Q]^{-1}P(u)b + L[I - P(u)Q]^{-1}[I - P(u)]p \end{aligned}$$

and this is the required relation.

3.6 Theorem. *Let* $Q \in B(E)$, $b \in E$ *and* $p \in K$ *fulfil the hypotheses of Lemma 3.5. Moreover, let* $Lp = p$.

Then

$$p^N - p = L[I - P(p^S)Q]^{-1}[I - P(p^S)](p^S - p) \ .$$

Proof. Let $u = p^S$. By Lemma 3.4 and Lemma 3.5

$$\begin{aligned} p^N - p = {} & L[I - P(p^S)Q]^{-1}P(p^S)b + L[I - P(p^S)Q]^{-1}[I - P(p^S)]p^S \\ & - L[I - P(p^S)Q]^{-1}P(p^S)b - L[I - P(p^S)Q]^{-1}[I - P(p^S)]p \end{aligned}$$

$$= L\left[I - P(p^s)Q\right]^{-1}\left[I - P(p^s)\right](p^s - p)$$

and this completes the proof.

3.7 Theorem. *Besides the relations* (3.1) *through* (3.6) *let the following hypotheses hold.*

1. *The spaces* E *and* F *are generated by closed normal cones* K *and* H *respectively and* $W \subset E$ *is a Banach space equipped by a norm majorizing the original norm in* E .
2. *Operator* $T \in B(E)$ *has property* "p" .
3. *There is a stationary distribution* p *such that* $Rp \in H^d$.
4. $Q \in B(E)$, $QK \subset K$, $r(Q) < 1$ *and* $b \in E$ *are such that* $r\left(P(u)Q\right) < 1$ *for* $u \in D$ *and* $p - Qp = b$. *Moreover,* $TS(u)\tilde{p} \in D$ *whenever* $u \in D$ *and* $\tilde{p} = RTS(u)\tilde{p} \in H$.
5. *There is a* W*-open neighborhood* $B \subset D \subset W$ *of* $p = Tp$ *such that* $b + QS(u)\left[I - RQS(u)\right]^{-1}Rb \in B$ *whenever* $u \in B$.
6. *There is a real number* α , $0 < \alpha < 1$, *such that* $r\left(P(u)Q\right) \leqq \alpha$, $u \in D$.
7. *There is a* $\beta \in (0,1)$ *such that* $r\left(J(p)\right) = \beta$, *where*
$$J(u) = T\left[I - P(u)Q\right]^{-1}\left[I - P(u)\right] .$$
8. *The operator-function* $S = S(u)$ *is* W*-continuous with respect to* $u \in D$.
9. *The element* $b \in K$ *is such that* $0 \neq b \leqq Tx$ *for all* $x \in ExK$, $\left[x,\hat{x}'\right] = 1$, *where* ExK *denotes the set of all extremal elements of* K.

Then the aggregation - disaggregation method described above is locally convergent to p , *i.e. there exists a* W*-open neighborhood* U *of* p *such that*

$$\lim_{k\to\infty} \|u^k - p\|_W = 0 \quad \textit{for} \quad u^0 \in U .$$

The speed of convergence is bounded above by the estimate

$$\|u^k - p\|_W \leqq \kappa\tau^k \tag{3.10}$$

where $\tau = r\left(J(p)\right) + \varepsilon + \eta < 1$ *with some* $\eta > 0$ *and* κ *independent of* k .

Proof. We let

$$Qx = Tx - [x, \hat{x}']b \ .$$

We see that all assumptions of Lemmas 3.4 and 3.5 and of Theorem 3.6 are fulfilled with T in place of L.

Let $|||\cdot|||$ be a norm in W-equivalent with the original one and such that

$$|||J(p)||| < r(J(p)) + \varepsilon \ .$$

The existence of such a norm is guaranteed by the well known Krasnoselskii Lemma see [3, p. 155].

According to Theorem 3.6 we get that

$$|||u^{k+1} - p||| \leq |||J(u^{k+1})||| \ |||u^{k+1} - p||| \leq \ldots \leq$$
$$|||J(u^{k})||| \ \ldots \ |||J(u^{0})||| \ |||u^{0} - p||| \ .$$

Since the operator-function $S = S(u)$ is W-continuous with respect to u so is $J = J(u)$. It follows that there exists a $|||\cdot|||$ - neighborhood $U \subset D$ of p such that if $u^0 \in U$ then

$$|||J(u^{k})||| < r(J(p)) + \varepsilon + \eta$$

with some $\eta > 0$ independent of k. As a consequence we obtain the required estimate (3.10). Theorem 3.7 is thus proved.

4. Examples

(i) Stationary distributions of finite Markov chains.

Let n and N be positive integers, $n < N$. Let $E = W = R^N$, $F = R^n$, $K = R^N_+$ and $H = R^n_+$. Let

$$T = (t_{jk}) \ , \quad 0 \leq t_{jk} \ , \quad j, k = 1,2,\ldots,N \ ,$$

$$\sum_{j=1}^{N} t_{jk} = 1 \ , \quad k = 1,2,\ldots,N \ .$$

Let $\hat{x}' = \hat{e} = (1,\ldots,1) \in R^N$ and let

(4.1) $$t_{j_0 k} > 0 \quad \text{for some} \quad j_0 \ , \quad 1 \leq j_0 \leq N \ .$$

Let G map the set $\{1,2,\ldots,N\}$ onto $\{1,2,\ldots,n\}$. Let us denote by (j) the set $G^{-1}(j) = \{m : G(m) = j\}$.

Define

$$Ru = z$$

by setting

$$z_{(j)} = \sum_{G(m)=j} u_m , \quad j = 1,\ldots,n$$

and

$$S(u)z = v$$

by

$$[S(u)z]_j = [(Ru)_{(j)}]^{-1} z_{(j)} u_j , \quad j = 1,\ldots,N .$$

Here $u \in D$, where

$$D = \{u \in R^N : \sum_{G(m)=j} u_m > 0 , \quad j = 1,\ldots,n\} .$$

If we define the matrices Q and C and the vector $b = (b_1,\ldots,b_N)$ by

$$b_j = \min \{(Te_m)_j : e_m = (0,\ldots,1,\ldots 0) \in R^N , \quad m = 1,\ldots,N\} ,$$

$$Qx = Tx - [x, \hat{e}_N]b$$

and

$$Cz = \frac{1}{[b,\hat{e}_N]} [z,\hat{e}_N]b$$

we observe that all the hypotheses of Theorem 3.7 are fulfilled.

In particular, $0 < [b, \hat{e}_N] \leq 1$ and

$$Q'\hat{e}_N = T'\hat{e}_N - [b, \hat{e}_N]\hat{e}_N = \hat{e}_N - [b, \hat{e}_N]\hat{e}_N = (1 - [b, \hat{e}_N])\hat{e}_N$$

and thus

$$r(Q) \leq \|Q\|_{\hat{e}_N} = \sup \{a \in R^1_+ : Q'\hat{e}_N \leq a\hat{e}_N\} = 1 - [b,\hat{e}_N] = \alpha < 1 .$$

In a manner invented in [3] one can show that

$$\|P(u)\|_{\hat{e}_N} = 1 , \quad u \in D ,$$

$$\|QP(u)\|_{\hat{e}_N} \leq \alpha < 1 ,$$

and

$$r\big(J(p)\big) \leq \sqrt{\alpha} = \beta < 1 \ .$$

The elements of the aggregation matrix have the form

$$a_{(j)(k)}(u) = \Big[\sum_{G(m)=k} u_m\Big]^{-1} \Big[\sum_{G(t)=j} \sum_{G(m)=k} t_{tm} u_m\Big]$$

and thus,

$$\big[A(u)\big]' \, \hat{e}_n = \hat{e}_n \ .$$

4.1 Theorem. *Let* T *be a transition matrix of a finite Markov chain, let* (4.1) *hold and let there exist a stationary distribution* p *such that* $Rp \in H^d$.

Then the iterative aggregation-disaggregation method converges locally to a stationary distribution.

(ii) Stationary distributions of infinite Markov chains

Let $E = W = \ell^1$, $F = R^n$, $K = \ell^1_+$ and $H = R^n_+$. Let $T = (t_{jk})$, $0 \leq t_{jk}$, $j, k = 1,2,\ldots$ and let

$$\sum_{j=1}^{\infty} t_{jk} = 1 \ , \quad j = 1,2,\ldots \quad \text{i.e.} \quad T'\hat{e} = \hat{e} \ ,$$

where $\hat{e} = (1,1,\ldots) \in \ell^\infty = (\ell^1)'$.

Let $T \in B(\ell^1) \cap B(\ell^2)$ and let there exist an index j_0 such that

(4.2) $$0 < \delta \leq t_{j_0 k} \ , \quad k = 1,2,\ldots$$

with δ independent of k .

Let G be a map of the set of positive integers $N = \{1,2,\ldots\}$ onto the set $\{1,2,\ldots n\}$.

Denote by

$$(j) = \{m \in N : G(m) = j\}$$

and define R by setting

$$Ru = z \Leftrightarrow (z)_{(j)} = \sum_{G(m)=j} u_m$$

and $S(u)$

$$S(u)z = v \ , \quad u \in D$$

by

$$v_j = \frac{1}{[(Ru)_{(j)}]} z_{(j)} v_j \ , \quad j = 1,2,\ldots$$

where

$$D = \{u \in l_+^1 : \sum_{G(m)=j} u_m > 0 \ , \quad j = 1,2,\ldots,n\} \ .$$

The elements of the aggregation matrix $A(u)$ have the form

$$a_{(j)(k)}(u) = \left[\sum_{G(m)=k} u_m\right]^{-1} \sum_{G(m)=j} \sum_{G(s)=k} t_{ms}$$

and hence

$$[A(u)]' \hat{e}_n = \hat{e}_n \ .$$

According to point 3° of the aggregation-disaggregation algorithm we have to find a stationary distribution of a finite Markov chain at each sweep of the iterative procedure.

For the sake of simplicity let us assume that operator T is l_+^1-irreducible.

After defining the matrices Q, C and the vector $b = (b_1, b_2, \ldots)$ by setting

$$b_j = \min \{(Te_m) : e_m = (0,0,\ldots 1,0,\ldots) \ , \quad m = 1,2,\ldots\}$$

$$Qx = Tx - [x, \hat{e}]b$$

and

$$Cz = ([b,\hat{e}])^{-1} [z,\hat{e}_m] \ ,$$

we are going to show that the hypotheses required in Theorem 3.7 are fulfilled.

As in the case of finite Markov chains the most difficult problem is to check conditions 4, 6 and 7.

Condition 2 is fulfilled if for every $\varepsilon > 0$ there is an index n_0 such that

$$\sum_{j>n_0} \sum_{k=1} t_{jk} |x_k| < \varepsilon \tag{4.3}$$

where $x \in l^1$, $\|x\|_1 \leq 1$. In such case, since $t_{jk} \leq 1$, $j, k = 1,2,\ldots$, T is compact (see [3, p. 225]), thus has property "p". The irreducibility

of T implies that condition 3 holds.

Further $p = Qp + b$ and $TS(u)\tilde{p} = \tilde{p} \in \operatorname{Int} R_+^n$. By applying an approach from [4] we are able to show that

$$\|P(u)Q\|_{\hat{e}} \leq \alpha < 1 , \quad u \in D$$

and also

$$r\big(J(p)\big) \leq (\alpha)^{1/2} = \beta < 1 .$$

We see that all hypotheses of Theorem 3.7 are fulfilled and we can formulate our result as a Theorem.

4.2 Theorem. *Let* $T \in B(l^1)$ *be a transition matrix of an infinite Markov chain. Let* T *be* l^1*-irreducible, let* (4.2) *and* (4.3) *hold.*

Then the iterative aggregation-disaggregation method locally converges to the unique stationary distribution.

References

[1] F. CHATELIN, W.L. MIRANKER: Acceleration by aggregation of successive approximation method. Linear Algebra Appl. 43: 17-47 (1980).

[2] M.G. KREIN, M.A. RUTMAN: Linear operators leaving invariant a cone i a Banach space. Usp. mat. nauk III (1948), Nr.1, 3-95 (In Russian). English translation in Amer. Math. Soc. Translations Nr.26 (1958), 128 pp.

[3] L.A. LUSTERNIK, V.I. SOBOLEV: Elements of Functional Analysis. Gos. Izd. Tech. Teor. Liter., Moscow 1951 (In Russian).

[4] J. MANDEL, B. SEKERKA: A local proof for the iterative aggregation method. Linear Algebra Appl. 51: 163-172 (1983).

[5] I. MAREK: Frobenius theory of positive operators. Comparison theorems and applications. SIAM J. Appl. Math. 19: 607-628 (1970).

[6] I. MAREK: On some spectral properties of Radon-Nikolskii operators and their generalizations. Comment. Math. Univ. Carol. 3, Nr.1, 20-30 (1962).

[7] I. MAREK, K. ZITNY: Matrix Analysis for Applied Sciences Vol. 1. Teubner Texte zur Mathematik, Band 60, Leipzig 1983.

[8] J. POLAK: Aggregation and Disaggregation in Markov Chains. Ph. D. Thesis. Charles University, Prague 1990. (In Czech).

[9] I. SAWASHIMA .: On spectral properties of some positive operators. Natur. Sci. Rep. Ochanomizu University 15: 53-64 (1964).

[10] A.E. TAYLOR: Introduction to Functional Analysis. J. Wiley, New York 1958.

Prof. Dr. Ivo Marek, katedra numerické matematiky na matematicko-fyzikální fakultě University Karlovy, Malostranské nám. 25, 118 00 Praha 1, Czechoslovakia

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

STABILITY OF A VIBRATING AND ROTATING BEAM

Jiří Neustupa

Czech Technical University, Prague, Czechoslovakia

Deformations of a rotating beam (of a circular cross-section and the lenght $L = 1$) may be described (after an appropriate choice of a time scale) by the differential equation

$$(1)\quad z_{tt} + z_{xxxx} + a(t,x)\, z_{xxx} + b(t,x)\, z_{xx} + c(t,x)\, z_x + d(t,x)\, z + e(t,x)\, z_t = f(t,x,z,z_t,z_x,z_{xx},z_{xxx})$$

where z is a complex function of variables t and x defined for $t \geq 0$ and $x \in \langle 0,1\rangle$ (see [2], [3]). The coefficients a, b, c, d, e as well as the nonlinear function f represent forces and moments acting on the beam. The function z is supposed to satisfy boundary conditions

$$(2_0)\quad z(t,0) = \nu_1 z_{xx}(t,0) + \nu_2 z_x(t,0) = 0,$$

$$(2_1)\quad z(t,1) = \nu_3 z_{xx}(t,1) + \nu_4 z_x(t,1) = 0$$

for $t \geq 0$. We assume that $|\nu_1| + |\nu_2| > 0$, $|\nu_3| + |\nu_4| > 0$. The function f is assumed to be continuous and bounded together with its second derivatives with respect to $x, z, z_t, z_x, z_{xx}, z_{xxx}$ on the set $A = \langle 0,+\infty) \times \langle 0,1\rangle \times \langle -R, R\rangle^5$ (where R is a positive real number). Moreover, we suppose that it satisfies the estimate

$$|f(t,x,z,z_t,z_x,z_{xx},z_{xxx})| \leq K_1 [\max\{|z|,|z_t|,|z_x|,|z_{xx}|,|z_{xxx}|\}]^2$$

on the set A. These properties of the function f together with sufficient regularity of a, b, c, d, e guarantee that the uniform asymptotic stability of the zero solution of the problem (1), (2_0), (2_1) is a consequence of the uniform asymptotic stability of the zero solution of the linear problem given by

$$(3)\quad z_{tt} + z_{xxxx} + a(t,x)\, z_{xxx} + b(t,x)\, z_{xx} + c(t,x)\, z_x + d(t,x)\, z + e(t,x)\, z_t = 0$$

and the boundary conditions (2_0) and (2_1). It can be proved by methods explained in [1], the stability can be considered for example

with respect to the norm

$$\|z(t,.)\| = \left\{ \int_0^1 [\, z(t,x)^2 + z_t(t,x)^2 + z_x(t,x)^2 + z_{tx}(t,x)^2 + z_{xx}(t,x)^2 + z_{xxx}(t,x)^2 + z_{xxxx}(t,x)^2 \,] \, dx \right\}^{1/2}.$$

Let us study the uniform asymptotic stability of the zero solution of the problem (3), (2_0), (2_1) now. Suppose that

(4) $a \equiv i \cdot \alpha$, $b \equiv \beta$, $c \equiv 0$, $d \equiv -\delta^2$, $e \equiv \varepsilon$,

(where α, β, δ, ε are real constants) and

(5) $-\nu_1 + \overline{\nu}_1 = i \cdot \alpha \cdot \nu_2$, $-\nu_3 + \overline{\nu}_3 = i \cdot \alpha \cdot \nu_4$.

The equation (3) is equivalent to the system

$$u_{1,t} = u_2$$
$$u_{2,t} = -u_{1,xxxx} - i \cdot \alpha \, u_{1,xxx} - \beta \, u_{1,xx} + \delta^2 u_1 - \varepsilon \, u_2.$$

The stability depends on eigenvalues λ of the problem

$$u_2 = \lambda u_1$$
$$-u_{1,xxxx} - i \cdot \alpha \, u_{1,xxx} - \beta \, u_{1,xx} + \delta^2 u_1 - \varepsilon \, u_2 = \lambda u_2.$$

Substituting $u_2 = \lambda u_1$ into the second equation and writing u instead of u_1, we obtain

(6) $-u_{xxxx} - i \cdot \alpha \, u_{xxx} - \beta \, u_{xx} + \delta^2 u = (\lambda^2 + \varepsilon \cdot \lambda) u$.

The eigenfunction u must satisfy boundary conditions

(7_0) $u(0) = \nu_1 u_{xx}(0) + \nu_2 u_x(0) = 0$,

(7_1) $u(1) = \nu_3 u_{xx}(1) + \nu_4 u_x(1) = 0$.

The zero solution of the problem (3), (2_0), (2_1) is uniformly asymptotically stable if there exists $\Delta > 0$ such that all eigenvalues λ of the problem (6), (7_0), (7_1) satisfy the condition $\mathrm{Re}\,\lambda < -\Delta$. (This statement follows from the expression of solutions of (3), (2_0), (2_1) by means of the Fourier method.)

Let us denote by σ the expression $\lambda^2 + \varepsilon \lambda$. It can be shown that all λ satisfying the equation $\sigma = \lambda^2 + \varepsilon \lambda$ have a real part less than $-\Delta$ if and only if

(8) $(\varepsilon - 2\Delta)^2 \cdot \mathrm{Re}\,\sigma + (\mathrm{Im}\,\sigma)^2 < (\varepsilon - 2\Delta)(\Delta - \varepsilon)\Delta$, $\varepsilon - 2\Delta > 0$.

Using (5), it can be proved that all numbers σ (equal to $\lambda^2 + \varepsilon \lambda$, where λ is an eigenvalue of (6), (7_0), (7_1)) are real. Thus, we obtain the simple sufficient condition for the uniform asymptotic stability of

the zero solution of (3), (2_0), (2_1): $\varepsilon > 0$ and $\mathrm{Re}\,\sigma < 0$ (for all numbers $\sigma \equiv \lambda^2 + \varepsilon\lambda$, where λ is an eigenvalue of (6), (7_0), (7_1)). (If these inequalities hold then we can easily show that there exists $\Delta > 0$ such that (8) is satisfied.) Let us suppose that $\varepsilon > 0$ in the following.

The equation (6) can be solved analytically. Using this analytic solution, we can show that if $\alpha = \beta = \delta = 0$ then there exists an infinite countable set of the eigenvalues λ of the problem (6), (7_0), (7_1) and each of the appropriate numbers σ is negative. If α, β, δ are growing then numbers σ are moving on the real axis from left to right. The zero solution of the problem (4), (2_0), (2_1) looses the stability at the moment when the greatest of the numbers σ crosses the point 0. The set of corresponding points $[\alpha, \beta, \delta]$ is a surface in R^3. This surface can be numerically obtained in the following way: We find the fundamental system of solutions u^{I}, u^{II}, u^{III}, u^{IV} of the equation (6), corresponding to $\sigma = 0$ (it is advantageous from practical reasons to choose such a system that $u^{I}(0) = 1$, $(d/dx)\,u^{I}(0) = \ldots = (d^3/dx^3)\,u^{I}(0) = 0$, $u^{II}(0) = 0$, $(d/dx)\,u^{II}(0) = 1$, $(d^2/dx^2)\,u^{II}(0) = (d^3/dx^3)\,u^{II}(0) = 0$, etc.). The solution of (6) satisfying (7_0), (7_1) has the form $C_1 u^{I} + C_2 u^{II} + C_3 u^{III} + C_4 u^{IV}$. We substitute it into the boundary conditions and we obtain a system of four linear algebraic equations for unknowns C_1, C_2, C_3, C_4. The system has a nontrivial solution if and only if its determinant is equal to zero. The determinant depends on α, β, δ. It is not difficult to find the surface as a set of such points $[\alpha, \beta, \delta]$ that the determinant is equal to zero. In fact, we can obtain not only one surface in this way. The first surface (which is the nearest to the point $[\alpha, \beta, \delta] = [0, 0, 0]$ and which is dividing the domain of stability from the domain of instability) corresponds to the case when the greatest of the numbers σ crosses the point 0, the second surface corresponds to the case when the second of the numbers σ crosses the point 0, etc. The Fig. 1 shows intersections of this surface in the region $\alpha \geq 0$, $\beta \geq 0$, $\delta \geq 0$ with planes $\beta = 0$, $\beta = 4$, $\beta = 8$, ..., $\beta = 36$. The intersection of the surface with the β axis is the point $[0, 4\pi^2, 0]$.

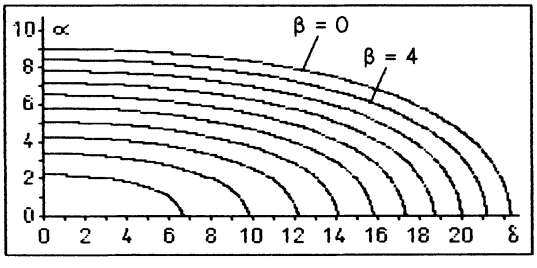


Fig. 1

The same method can be used also in the case of other boundary conditions, for instance if $\nu_1 = 1$, $\nu_2 = i \cdot \alpha/2$, $\nu_3 = 1$ and $\nu_4 = i \cdot \alpha/2$. We can see the intersections of the surface dividing the domain of stability and the domain of instability with planes $\beta = 0$, $\beta = 1, \ldots, \beta = 9$ in the region $\alpha \geq 0$, $\beta \geq 0$, $\delta \geq 0$ on the Fig. 2. The intersection of the surface with the β axis is the point $[0, \pi^2, 0]$.

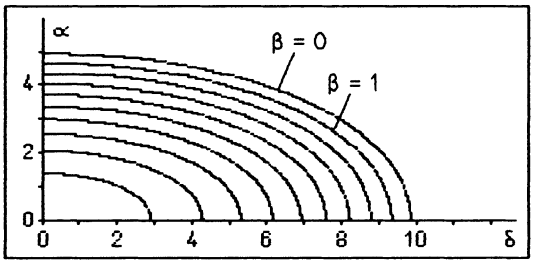


Fig. 2

The zero solution of (3), (2_0), (2_1) is only stable in the sense of Lyapunov (and not asymptotically stable) for $[\alpha, \beta, \delta]$ laying on the surfaces (see Fig. 1, Fig. 2). Similarly, we obtain only the stability of the zero solution of (3), (2_0), (2_1) if $\varepsilon = 0$ and $[\alpha, \beta, \delta]$ is on or under the mentioned surfaces. We cannot make the conclusion about the stability of the zero solution of the nonlinear problem (1), (2_0), (2_1) in these cases.

We can obtain similar results also for instance if $\nu_1 = \nu_3 = 1$, $\nu_2 = \nu_4 = i \cdot \alpha$ or $\nu_1 = \nu_3 = 1$, $\nu_2 = \nu_4 = 0$, although the condition (5) is not satisfied in these cases.

If we consider for example the boundary conditions corresponding to $\nu_1 = 0$, $\nu_2 = 1$, $\nu_3 = 1$, $\nu_4 = i \cdot \alpha$ and if we assume that $\varepsilon > 0$ then the transition from stability of the zero solution of (3), (2_0), (2_1) to instability (for α, β, δ growing) is not caused by the first number $\sigma = \lambda^2 + \varepsilon\lambda$ (where λ is an eigenvalue of the problem (6), (7_0), (7_1)) crossing the point 0 on the real axis as in the preceeding cases. If $\alpha = \beta = \delta = 0$ then all numbers σ lay on a negative part of the real axis but they leave the real axis if α is growing. But the set of the numbers σ is infinite and we are not able to study the movement of all of them in the complex domain if parameters α, β, δ are changing. That is why our attention was paid to three numbers σ which were the nearest to the point 0 only. We were looking for the case when the first of these numbers is crossing the parabola

$$\text{(10)} \qquad \operatorname{Re}\sigma \cdot \varepsilon^2 + (\operatorname{Im}\sigma)^2 = 0.$$

(If $\operatorname{Re}\sigma \cdot \varepsilon^2 + (\operatorname{Im}\sigma)^2 < 0$ for all of our three numbers σ then there exists $\Delta > 0$ so that these numbers satisfy (8).) The surface which

is the set of corresponding points $[\alpha, \beta, \delta]$ for $\varepsilon = 1$ can be seen on the Fig. 3. The curves show intersections of the surface with planes $\beta = 0$, $\beta = 2, \ldots, \beta = 20$; the intersection with the β axis is the point $[0, 20.18, 0]$. We must confess that there is no theory saying that for all combinations of α, β, δ, one of our three numbers σ is really the first one from the set of all numbers σ which causes the transition to instability.

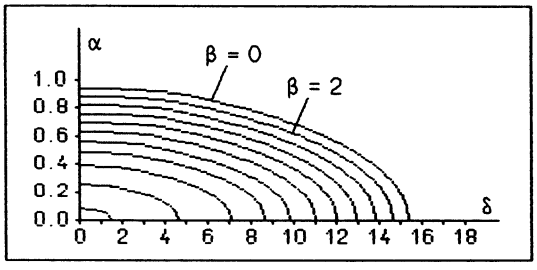


Fig. 3

In a real situation, the coefficients α, β and δ have the following meaning: $\alpha = M/(EI)$, $\beta = N/(EI)$, $\delta^2 = \varrho A \Omega^2/(EI)$, where M is a twisting moment, E is a modulus of elasticity, I is the moment of inertia of the cross section, N is the normal force, ϱ is a density of the material, A is the area of the cross section and Ω the angular velocity of the rotation.

The method which makes possible to find the eigenvalues λ of the problem (6), (7_0), (7_1)) can be used also if coefficients a, b, c, d, e in the equation (4) depend on the variable x. The only essential difference is that we must use some numerical method to obtain an appropriate fundamental system of solutions of the equation (6).

References:

[1] Neustupa J., (1983) The linearized uniform asymptotic stability of evolution differential equations. Czech. Math. J. 34, 257 - 284

[2] Vejvoda O. et al., (1981) Partial differential solutions. Sijthoof & Noordhoff International Publishers B.V., Alphen aan den Rijn, Netherlands

[3] Volmir A.S., (1963) Stability of elastic systems. Gos. Izdat. Fyz.- Mat. Lit., Moscow, USSR (Russian)

Authors address: Czech Technical University, Faculty of Mechanical Engineering, Technická 4, 166 07 Praha 6, CSFR

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

QR; ITS FORWARD INSTABILITY AND FAILURE TO CONVERGE

Beresford N. Parlett

Department of Mathematics
and Computer Science Division of EECS
University of California
Berkeley, CA 94720, USA

Jian Le

EEG Inc.
San Francisco, CA, USA

ABSTRACT

Examples are given of forward instability in the QR algorithm. Analysis shows that for symmetric tridiagonal matrices this occurs when the shift is close to an eigenvalue whose normalized eigenvector has a tiny last entry. Forward instability occurs after a condition we call premature deflation. A condition number for deflation from Hessenberg matrices is proposed.

An example is given in which the QR algorithm with Rayleigh quotient shift fails to converge on a 3×3 nonsymmetric matrix.

1. Introduction and Summary

The QR algorithm computes a sequence of unitary similarity transformations designed to drive a given matrix into an upper triangular (or Schur) form. The convergence properties depend on the shift strategy that must be supplied as part of the algorithm. It is well known that the algorithm is backward stable, see [Wilkinson, 1965, Chap.3]. This means that the final triangular, or diagonal, matrix is exactly unitarily similar to a matrix that is very close (in norm) to the original one. It is also known to the experts, see [Stewart, 1970] that sometimes the algorithm exhibits forward instability. This means that the sequence of orthogonal transformations is far from the sequence used in exact arithmetic. To compute eigenvalues backward stability suffices. However, the algorithm has other uses (in deflation and in inverse eigenvalue problems) and then forward stability is most desirable.

Our goal is to investigate the conditions for forward instability and to see when its presence is detectable. Nevertheless at first glance the solution is easy and may be summarized as follows.

The QR factorization of $B \in C^{m\times n}$, $m \geq n$, is a compact representation of the Gram-Schmidt orthonormalization of the columns of B taken in the natural order $b_1, b_2, ..., b_n$. Q holds the new columns and upper triangular R holds the intermediate coefficients. If $b_k \in$ span $(b_1, ..., b_{k-1})$ for some $k<n$, then q_k, the k-th column of Q, is not determined by the data; uniqueness fails. This is the only way it can fail. In finite precision arithmetic forward instability in the computed Q and R occurs if, and only if, B is very close to a matrix for which QR factorization is not unique. The QR algorithm simply inherits any instability from one or more of the factorizations involved.

Perhaps, this simple argument is the reason why there has been no previous study of forward instability. There is nothing more to be said - in general.

For reasons of economy the QR algorithm is usually applied to upper Hessenberg matrices (the (i, j) entry vanishes if $i>j+1$) and this structure plays an important role. Suppose that B is upper Hessenberg and has a strong subdiagonal. Let $B_k := (b_1, b_2, ..., b_k)$. At first sight it is not clear how b_{k+1} could ever come close to *range* (B_k) since b_{k+1} has one more nonzero entry than b_k. Nevertheless this is possible although the strong subdiagonals ensure that stability is lost little by little during the factorization. A careful analysis reveals that the sensitivity of the Q and the R factors of B_{k+1} depend only on the condition number of B_k. Moreover these R factors and their leading principal submatrices conceal their ranks - any ill condition hides behind substantial diagonal entries.

Automatic deflation, which is an attractive feature of the QR algorithm, is a casualty of forward instability. We discovered that this instability is the aftermath of what we call **premature deflation**. Programs in EISPACK do not look for this phenomenon and so it has not been observed. In other words, if unreduced Hessenberg B is singular then, in exact arithmetic, one single QR transformation allows deflation of the last row and column, since the new last row will vanish. If forward instability occurs then, in finite precision, the tiny row emerges earlier than expected, goes unnoticed, and then gets overwritten. See Section 4.2.

The good news is that forward instability of the QR algorithm is rare. If the shift σ is not close to an eigenvalue there is no danger at all. Even if σ is very close it is still possible for the transformation to be forward stable. However for symmetric tridiagonal matrices forward instability will occur when σ is very close to an eigenvalue whose normalized eigenvector has a tiny last entry. See Theorem 4.1. Moreover this situation occurs if, and only if, σ is almost an eigenvalue of two successive leading principal submatrices (Section 4.2).

For an upper Hessenberg matrix which is a smooth function of a real parameter Section 4.3 presents tight bound on the derivative of the last subdiagonal entry $\hat{\beta}$ of the QR transform. Further studies of forward instability are in progress.

Section 2 presents background information on QR and some illuminating numerical examples. Section 3 analyzes the QR factorization of a symmetric tridiagonal T and establishes needed results.

Section 4 presents bounds on derivatives and explains premature deflation. Finally, in Section 5, we present a recent result due to Batterson and Smillie on the failure of a reasonable shift strategy for non-symmetric QR. It is not known whether or not the shift strategies used in practice converge for almost all matrices.

We use the Householder conventions: lower case Greek letters for scalars, lower case Roman letters for column vectors, and upper case Roman for matrices.

2. Background Information

This section covers the definition of the QR transformation and its relation to eigenvalue deflation and eigenvector calculation for a symmetric tridiagonal matrix T. The examples in Section 2.3 show that the QR transformation on T can sometimes be violently unstable in the forward sense.

2.1. The QR transformation

This well established procedure is described in several books; e.g. [Wilkinson, 1965, Chap.8], [Stewart, 1973, Chap.7], [Parlett, 1980, Chap.8], [Golub & Van Loan, 1983, Chap.7]. For any square complex matrix A and any scalar $\{\sigma\}$ (called the shift) that excludes A's eigenvalues, the associated *QR transformation* $A \rightarrow \hat{A}$ is defined as follows:

i) let $A - \sigma I = QR$, *the unique unitary upper triangular decomposition with the diagonal elements of* R *being positive.*

ii) define $\hat{A} = RQ + \sigma I = Q^* A Q$.

One important property of the QR transformation is that both the upper Hessenberg form $(A = (a_{ij})$ with $a_{ij} = 0$ if $i>j+1)$ and the Hermitian form $(A = (a_{ij})$ with $\bar{a}_{ij} = a_{ji})$ are preserved.

2.2. Eigenvalue deflation and eigenvector calculation

A well known result (see [Wilkinson, 1965, pp 469-471]) connects QR with eigenvalue deflation and eigenvector computation.

Definition 2.1: *A symmetric tridiagonal matrix* T *is called* **unreduced** *if its subdiagonal elements are nonzero.*

Remark: *When* T *is unreduced the QR transformation is well defined for all shifts* σ *because the first* $n-1$ *columns of* $T - \sigma I$ *are linear independent for all* σ.

Lemma 2.1: (QR and deflation)

Let T *be* **unreduced** *and* $\hat{T}$ *be the QR transform of* T *with shift* σ, *i.e.*

$$\hat{T} := Q^* T Q = RQ + \sigma I \tag{2.2.1}$$

where $T - \sigma I = QR$. *If* $\sigma = \lambda$, *an eigenvalue of* T, *then*

(1) last row of $\hat{T}$ *has the form* $(0, \ldots, 0, \lambda)$;

(2) last column of Q, *namely* $\mathbf{q}_n$, *satisfies*

$$T\mathbf{q}_n = \mathbf{q}_n \lambda. \tag{2.2.2}$$

Here Q *is orthogonal and* R *is upper triangular.*

Since $\hat{T} = \bar{T} \oplus \lambda$ where $\bar{T}$ has order one less than T, we say that λ has been **deflated from** T in one sweep of QR transformation. It is clear that the spectrum of $\bar{T}$ consists of the remaining eigenvalues of T. Also from Lemma 2.1, we see that when λ is deflated from T, its corresponding eigenvector is revealed in Q, namely its last column $\mathbf{q}_n$.

2.3. Some examples

In this subsection, we show, by example, that Lemma 2.1 is not a reliable guide to results in finite precision computation. Example 2.1 will show a successful deflation and Example 2.2 will show a failure. Example 2.3 will exhibit the success of deflation on the failed case in Example 2.2 after two sweeps of QR have been applied. Example 2.4 is an interesting case of success despite having a shift σ that is an exact eigenvalue of several of T's leading principal submatrices.

The data given in the following matrices have been multiplied by 10^4 for the purpose of better presentation. The transformed $\hat{T}$ here is generated by the numerical implementation of QR called **TQR**, which will be described in Section 3.2.

Example 2.1: (the successful case)

$$T_6 = \begin{bmatrix} 6683.3333 & 14899.672 & & & & & \\ 14899.672 & 33336.632 & 34.640987 & & & & \\ & 34.640987 & 20.028014 & 11.832164 & & & \\ & & 11.832164 & 20.001858 & 10.141851 & & \\ & & & 10.141851 & 20.002287 & 7.5592896 & \\ & & & & 7.5592896 & 20.002859 & \end{bmatrix} \tag{2.3.1}$$

$$\lambda_1 = 0, \quad \lambda_2 = 10, \quad \lambda_3 = 20, \quad \lambda_4 = 30, \quad \lambda_5 = 40, \quad \lambda_6 = 40000.$$

The shift is $\lambda_1 = 0$. The matrix $\hat{T}$ after one QR sweep is:

$$\hat{T}_6 = \begin{bmatrix} 39999.925 & 54.726511 & & & & & \\ 54.726511 & 33.404823 & 8.3017268 & & & & \\ & 8.3017268 & 24.730751 & 8.8065994 & & & \\ & & 8.8065994 & 21.646903 & 7.2175779 & & \\ & & & 7.2175779 & 20.292461 & -7.943\mathbf{d}-12 & \\ & & & & -7.943\mathbf{d}-12 & -2.344d-15 & \end{bmatrix}$$

The last row of $\hat{T}_6$ is negligible as we expected. For comparison, here is $\hat{T}_6$ computed by a method other than **TQR**.

$$\hat{T}_6 = \begin{bmatrix} 39999.925 & 54.726511 & & & & & \\ 54.726511 & 33.404823 & 8.3017268 & & & & \\ & 8.3017268 & 24.730751 & 8.8065994 & & & \\ & & 8.8065994 & 21.646903 & 7.2175779 & & \\ & & & 7.2175779 & 20.292461 & -1.113\mathbf{d}-14 & \\ & & & & -1.113\mathbf{d}-14 & 9.520d-13 & \end{bmatrix}$$

The matrix elements of these two transformed T_6 are almost identical except the bottom ones. However, they are negligible.

Example 2.2: (the failed case)

The matrix T is the same as the one in Example 2.1. The shift is λ_6. The matrix $\hat{T}$ after one QR sweep is:

$$\hat{T}_6^{(1)} = \begin{bmatrix} 19.989995 & 14.142133 & & & & & \\ 14.142133 & 20.003002 & 11.832160 & & & & \\ & 11.832160 & 20.001858 & 10.141851 & & & \\ & & 10.141851 & 20.002287 & 7.5593584 & & \\ & & & 7.5593584 & 20.730517 & -170.56153 & \\ & & & & -170.56153 & 39999.272 & \end{bmatrix} \quad (2.3.2)$$

The last subdiagonal element is not negligible. For comparison, here is $\hat{T}_6$ computed by a method other than **TQR**.

$$\hat{T}_6 = \begin{bmatrix} 19.989995 & 14.142133 & & & & & \\ 14.142133 & 20.003002 & 11.832160 & & & & \\ & 11.832160 & 20.001858 & 10.141851 & & & \\ & & 10.141851 & 20.002287 & 7.5592896 & & \\ & & & 7.5592896 & 20.002859 & -1.608\mathbf{d}{-13} & \\ & & & & -1.608\mathbf{d}{-13} & 40000.000 & \end{bmatrix} \quad (2.3.3)$$

Examples 2.1, 2.2 have shown that the transformation with $\sigma = \lambda_1$ is stable and the one with $\sigma = \lambda_6$ is unstable. Examination of the eigenvalues of the leading principal submatrices of T_6 reveals that λ_6 matched the biggest eigenvalues of T_3, T_4, and T_5 to almost full working precision. On the other hand λ_1 is not close to any eigenvalues of T_3, T_4 and T_5.

Example 2.3:

The tridiagonal matrix T_6 is the same as that in Example 2.2. We applied the QR transformation once more to the " $\hat{T}_6^{(1)}$ " exhibited in Example 2.2 keeping the same shift $\lambda_6 = 40000$. The resulting matrix is:

$$\hat{T}_6^{(2)} = \begin{bmatrix} 19.979990 & 14.142125 & & & & & \\ 14.142125 & 20.006003 & 11.832161 & & & & \\ & 11.832161 & 20.003716 & 10.141851 & & & \\ & & 10.141851 & 20.004574 & 7.5592897 & & \\ & & & 7.5592897 & 20.005717 & 8.425\mathbf{d}{-15} & \\ & & & & 8.425\mathbf{d}{-15} & 40000.000 & \end{bmatrix} \quad (2.3.4)$$

It is important to notice that $\hat{T}_6^{(2)}$ is clearly different from $\hat{T}_6$. Thus the second QR transform with the same shift σ is not a viable way to obtain $\hat{T}_6$.

Example 2.3 has shown that, in one case at least, **TQR** will take two sweeps to get the deflated $\hat{T}_6$. Examination of the eigenvalues of the leading principal submatrices of $\hat{T}_6^{(1)}$ obtained by the first QR sweep with $\sigma = \lambda_6$ reveals that the first QR sweep does no more than destroy the closeness of the eigenvalues of T_6's leading principal submatrices.

Example 2.4: (successful deflation in an interesting case)

The matrix T in this example is the well known *second difference matrix.*

$$T_5 = \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}$$

$\lambda_1 = -2-\sqrt{3},\ \lambda_2 = -3,\ \lambda_3 = -2,\ \lambda_4 = -1,\ \lambda_5 = -2+\sqrt{3}.$

The shift is $\lambda_3 = -2$. The matrix $\hat{T}_5$ after one QR sweep is:

$$\hat{T}_5 = \begin{bmatrix} -2.0000000 & 1.4142136 & & & \\ 1.4142136 & -2.0000000 & 0.70710678 & & \\ & 0.70710678 & -2.0000000 & 1.2247449 & \\ & & 1.2247449 & -2.0000000 & 0.000000 \\ & & & 0.0000000 & -2.00000000 \end{bmatrix}$$

One can verify that λ_3 is also an eigenvalue of the first and the third leading principal submatrices of T_5. This example shows that even if the shift is an eigenvalue of some of the leading principal submatrices, nevertheless QR deflates T in one sweep.

3. Implementation of the QR transformation

This section develops the usual implementation of the QR algorithm applied to a symmetric tridiagonal matrix T. Most of the material is standard, see [Wilkinson, 1965, Chap.8], [Stewart, 1973, Chap.7], [Parlett, 1980, Chap.8] and [Golub & Van Loan, 1983, Chap.7]. However all the results are needed in the next section.

In the following discussion, we assume that all the tridiagonal matrices in question are *unreduced* since otherwise the problem decouples. We also assume that the off-diagonal elements β_k $(2 \le k \le n)$ of T are *positive*.

3.1. QR factorization of $T - \sigma I$

The desired QR decomposition can be carried out by pre-multiplying the tridiagonal matrix $T - \sigma I$ by a sequence of plane rotation matrices R_k ($2 \le k \le n$) with c_k, s_k in row $k-1$, $-s_k$, c_k in row k.

The duty of R_k ($2 \le k \le n$) is to annihilate the ($k, k-1$) position of the matrix on the way to an upper triangular form. The formulae in step (k) are important for the analysis in Section 4.

Let

$$T - \sigma I = \begin{bmatrix} \alpha_1 - \sigma & \beta_2 & & & \\ \beta_2 & \alpha_2 - \sigma & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \beta_n \\ & & & \beta_n & \alpha_n - \sigma \end{bmatrix}. \tag{3.1.1}$$

It can be shown that at step (k) ($k<n$):

$$R_k R_{k-1} \cdots R_2(T - \sigma I) = \begin{bmatrix} \xi_2 & \zeta_2 & \times & & & \\ & \ddots & \ddots & & & \\ & & \ddots & \ddots & & \\ & & & \xi_k & \zeta_k & \times \\ & & & & \pi_k & c_k\beta_{k+1} \\ & & & & \beta_{k+1} & \alpha_{k+1}-\sigma \\ & & & & & \ddots \end{bmatrix} \tag{3.1.2}$$

where

$$\xi_k = (\pi_{k-1}^2 + \beta_k^2)^{1/2} \qquad \zeta_k = c_{k-1}c_k\beta_k + s_k(\alpha_k - \sigma)$$
$$c_k = \pi_{k-1}/\xi_k \qquad s_k = \beta_k/\xi_k$$
$$\pi_k = -s_k\beta_k c_{k-1} + c_k(\alpha_k - \sigma). \tag{3.1.3}$$

At step (n):

$$R_n R_{n-1} \cdots R_2(T - \sigma I) = \begin{bmatrix} \xi_2 & \zeta_2 & \times & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \xi_n & \zeta_n \\ & & & & \pi_n \end{bmatrix} = R. \tag{3.1.4}$$

with

$$Q = R_2^t R_3^t \cdots R_n^t.$$

We collect some direct consequences of this decomposition that are used later.

(I): $\chi_k = \det[T_k - \sigma I_k]$,

$$\chi_1(\sigma) = \alpha_1 - \sigma = \pi_1,$$
$$\chi_k(\sigma) = \xi_2 \cdots \xi_k \pi_k, \qquad 2 \le k \le n. \tag{3.1.5}$$

(II): *If T is unreduced, then*

$$\xi_k \neq 0, \quad s_k \neq 0, \qquad 2 \le k \le n. \tag{3.1.6}$$

(III): *If T is unreduced, then*

$$\pi_k(\sigma) = 0 \quad \text{if and } \textit{only} \text{ if} \quad \chi_k(\sigma) = 0, \quad 1 \le k \le n. \tag{3.1.7}$$

Lemma 3.1: *The detailed structure of matrix Q is:*

$$Q = R_2^t R_3^t \cdots R_n^t = \begin{bmatrix} c_1c_2 & c_1(-s_2)c_3 & c_1(-s_2)(-s_3)c_4 & \cdots & c_1(-s_2)\ldots(-s_n) \\ s_2 & c_2c_3 & c_2(-s_3)c_4 & \cdots & \vdots \\ & s_3 & c_3c_4 & \ddots & \vdots \\ & & s_4 & \ddots & \vdots \\ & & & c_{n-1}c_n & c_{n-1}(-s_n) \\ & & & s_n & c_n \end{bmatrix}. \tag{3.1.8}$$

3.2. An example revisited

In the following, the computed c_k's, π_k's from **TQR** in Example 2.2 (see Section 2.3) are listed. The resulting $\hat{T}$ is exhibited in (2.3.2). For comparison, the correct c_k's and π_k's are also listed. The resulting $\hat{T}$ is exhibited in (2.3.3).

	computed	*correct*
k	c_k	c_k
1	+1.0000000000000d+00	+1.0000000000000d+00
2	−9.1287087206375d−01	−9.1287087206375d−01
3	+7.9096456588243d−04	+7.9096456594300d−04
4	−2.3388300212821d−07	−2.3408762727625d−07
5	−8.0658942646820d−07	+5.9381746504385d−11
6	+4.2662106420853d−03	−1.1223688023196d−14

k	π_k	π_k
1	−3.3316666666667d+00	−3.3316666666667d+00
2	+2.7399802074030d−06	+2.7399802074446d−06
3	−2.7673419903274d−10	−2.7697632729029d−10
4	−8.1803099953432d−10	+6.0232682537306d−14
5	+3.2249815432264d−06	−1.9058339689554d−17
6	−1.7056308317811d−02	−1.6080662764449d−17

3.3. Analysis of the QR transformation

Section 3.2 reveals the relations between the quantities of c_k, s_k, π_k and the matrix elements α_k, β_k in one step. This is inadequate if the analysis in terms of σ is needed. In the next lemma, we present several matrix-vector relations between all the intermediate quantities generated in the QR process and the matrix T. It is these relations which will help us understand QR more deeply. These relations also tell us the structure of an eigenvector when σ happens to be an eigenvalue of T.

Recall the partial reduction of $T-\sigma I$ to upper triangular form as it appears at step (k). It is given in (3.1.2).

The product of the plane rotation matrices $R_2, \ldots, R_k$ satisfies:

$$R_2^t R_3^t \cdots R_k^t = \begin{bmatrix} c_2 & -s_2c_3 & \cdot & \cdot & \cdot & c_1(-s_2)\ldots(-s_k) \\ s_2 & c_2c_3 & \cdot & \cdot & \cdot & \cdot \\ & s_3 & \cdot & \cdot & \cdot & \cdot \\ & & \cdot & \cdot & \cdot & \cdot \\ & & & \cdot & c_{k-1}c_k & c_{k-1}(-s_k) \\ & & & & s_k & c_k \\ & & & & & \\ & & & & & I_{n-k} \end{bmatrix}. \tag{3.3.1}$$

The $k-th$ column of (3.3.1) plays a crucial role in our analysis.

Definition 3.2: *We denote by* $\mathbf{y}_k$ *the following vector in* $\mathbf{R}^k$.

$$\mathbf{y}_k: = \begin{bmatrix} c_1(-s_2)\cdots(-s_k) \\ c_2(-s_3)\cdots(-s_k) \\ \vdots \\ c_{k-1}(-s_k) \\ c_k \end{bmatrix}. \tag{3.3.2}$$

Definition 3.3: *We denote by* $\bar{\mathbf{y}}_k$ *the following vector in* $\mathbf{R}^n$.

$$\bar{\mathbf{y}}_k: = R_2^t \cdots R_k^t \, \mathbf{e}_k = \begin{bmatrix} \mathbf{y}_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{3.3.3}$$

Equating the $k-th$ row of (3.1.2), the following matrix-vector relations emerge.

Lemma 3.2: *If* T *is a symmetric tridiagonal matrix of order* n *and* σ *is an arbitrary real number, then the quantities derived up to step* k *in* **TQR satisfy**

$$T\bar{\mathbf{y}}_k - \bar{\mathbf{y}}_k\sigma = \pi_k \mathbf{e}_k + c_k\beta_{k+1}\mathbf{e}_{k+1}, \qquad k < n, \tag{3.3.4}$$

$$\| T\bar{\mathbf{y}}_k - \bar{\mathbf{y}}_k\sigma \|^2 = \pi_k^2 + c_k^2\beta_{k+1}^2, \qquad k < n, \tag{3.3.5}$$

$$(\bar{\mathbf{y}}_k)^t(T\bar{\mathbf{y}}_k - \sigma\bar{\mathbf{y}}_k) = \pi_k c_k = \gamma_k = (\bar{\mathbf{y}}_k)^t T\bar{\mathbf{y}}_k - \sigma, \qquad k \le n, \tag{3.3.6}$$

$$T\bar{\mathbf{y}}_n - \bar{\mathbf{y}}_n\sigma = \pi_n \mathbf{e}_n. \tag{3.3.7}$$

Proof: Equate the $k-th$ row on each side of (3.1.2) and transpose to get

$$(T - \sigma I)(R_2^t \cdots R_k^t)\, \mathbf{e}_k = (T - \sigma I)\, \bar{\mathbf{y}}_k = \pi_k \mathbf{e}_k + c_k\beta_{k+1}\mathbf{e}_{k+1}.$$

Since (3.1.2) holds for $k < n$, (3.3.4) is true for $k < n$. (3.3.5), (3.3.6) are the direct results of (3.3.4). (3.3.7) is a special case of (3.3.4) since $\beta_{n+1} = 0$ when $k = n$. □

We use notation $\bar{\mathbf{y}}_k$ instead of $\mathbf{q}_k$ since the former is slightly different from the $k-th$ column $\mathbf{q}_k$ of matrix Q. When $k = n$, $\bar{\mathbf{y}}_n = \mathbf{q}_n$. □

Lemma 3.2 used T and $\bar{\mathbf{y}}_k$. Equally important is the relation between T_k, the leading $k \times k$ submatrix of T, and $\mathbf{y}_k$.

Corollary 1 of Lemma 3.2: *For any real* σ,

$$T_k\mathbf{y}_k - \mathbf{y}_k\sigma = \pi_k \mathbf{e}_k^{(k)}, \qquad 1 \le k \le n. \tag{3.3.8}$$

$$(\mathbf{y}_k)^t(T_k\mathbf{y}_k - \sigma\mathbf{y}_k) = \pi_k c_k = \gamma_k = (\mathbf{y}_k)^t T_k\mathbf{y}_k - \sigma, \qquad k \le n. \tag{3.3.9}$$

Proof: Taking the first k rows of (3.3.4) and using the notation in (3.3.2), the formula in (3.3.8) is obtained for $k < n$. When $k = n$ it is the case in (3.3.7).

(3.3.9) is the direct results of (3.3.8) and (3.3.2). □

4. Sensitivity of Deflation

4.1. Dependence on the shift

The QR transform of T is given by $\hat{T} = RQ + \sigma I$. Reference to Equations (3.1.4) and (3.1.8) shows that

$$\hat{\beta}_n(\sigma) = \pi_n(\sigma)s_n(\sigma)$$

where $s_n = \sin\theta_n$ and θ_n is the last rotation angle, and, from (3.3.7), $(T - \sigma I)y_n = e_n\pi_n$. Thus at an eigenvalue λ,

$$\pi_n(\lambda) = 0 \ , \tag{4.1.1}$$

and $c_n = \cos\theta_n$ is the last entry of the eigenvector y_n. For general σ let f' denote the derivative $f(\sigma)$.

Lemma 4.1: *For all real σ and $k \le n$,*

$$\pi_k c_k' - c_k \pi_k' = 1 \ . \tag{4.1.2}$$

Proof. Differentiate (3.3.8)

$$-y_k + (T_k - \sigma I_k)y_k' = \pi_k' e_k^{(k)} \ .$$

Multiply by y_k^t and recall (3.3.2) to find

$$-y_k^t y_k + y_k^t(T_k - \sigma I_k)y_k' = c_k\pi_k' \ .$$

By (3.3.8) transposed $y_k^t(T_k - \sigma I_k) = \pi_k e_k^{(k)t}$ and so

$$-1 + \pi_k c_k' = c_k \pi_k' \ . \qquad \square$$

Theorem 4.1: *When $\sigma = \lambda_i$, an eigenvalue of T, then*

$$\hat{\beta}_n'(\lambda_i) = -\tan\theta_n \tag{4.1.3}$$

where $\cos\theta_n$ is the last entry of λ_i's normalized eigenvector.

Proof. For all σ, $\hat{\beta}_n' = \pi_n' s_n + \pi_n s_n'$. By (4.1.1) $\pi_n(\lambda_i) = 0$ and by (4.1.2) $\pi_n'(\lambda_i) = -1/c_n$. Thus $\hat{\beta}_n'(\lambda_i) = -s_n/c_n$. $\square$

Theorem 4.2 indicates those eigenvalues that will provoke forward instability in the QR algorithm. Of course there exist tridiagonal matrices for which no normalized eigenvector has a tiny last entry and in such cases the QR transformation is forward stable for all shifts σ.

4.2. Premature deflation

Next we exhibit the active part of the matrix $R_k R_{k-1} \ldots R_2 T R_2^t \ldots R_k^t$. It is derived from (3.1.2)

$$\begin{bmatrix} \cdot & \beta_{k-1} & & & \\ \beta_{k-1} & \alpha_{k-1} & s_k\pi_k & s_k\beta_{k+1} & \\ & \pi_k s_k & \pi_k c_k + \sigma & c_k\beta_{k+1} & \\ & \beta_{k+1}s_k & c_k\beta_{k+1} & \alpha_{k+1} & \beta_{k+2} \\ & & & \beta_{k+2} & \cdot \end{bmatrix}$$

The bulge $\beta_{k+1}s_k$ occurs in position $(k-1,k+1)$ and $(k+1,k-1)$. If both $s_k\pi_k$ and $c_k\beta_{k+1}$ are negligible then $c_k\pi_k + \sigma$ may be taken as an eigenvalue. In addition if row k and column k are deleted then the

remaining matrix is tridiagonal. We refer to this as **premature deflation**. There is no need to complete the QR transformation since the goal has been achieved. Current implementations of the QR algorithm do not monitor the quantity $\pi_k^2 s_k^2 + c_k^2 \beta_{k+1}^2$ and so never know whether this event occurs. When subsequent similarities by R_{k+1}, R_{k+2} are applied, even in exact arithmetic (but σ is not exactly an eigenvalue), the Schur parameters $c_{k+1}, c_{k+2}, \ldots$ increase slowly and deletion of row and column l, $k < l \leq n$, is not warranted. By failing to recognize the stage at which deflation is justified the opportunity is lost.

Observe that, from (3.3.5),

$$\pi_k^2 s_k^2 + \beta_{k+1}^2 c_k^2 \leq \pi_k^2 + \beta_{k+1}^2 c_k^2 = \| T\bar{y}_k - \bar{y}_k \sigma \|^2 .$$

Thus premature deflation occurs precisely when $(\sigma, \bar{y}_k)$ is an acceptable approximate eigenpair for T. Equation (3.3.3) defines $\bar{y}_k$. Recall from (3.3.8) that

$$|\pi_k| = \| T_k y_k - y_k \sigma \| .$$

Removal of superfluous rows in (3.3.5) yields

$$\pi_k^2 + \beta_{k+1}^2 c_k^2 = \| T_{k+1} \bar{y}_k - \bar{y}_k \sigma \|^2$$

and so premature deflation at step k implies that σ is an acceptable eigenvalue of T_k and T_{k+1}.

4.3. General Perturbations

It is illuminating to analyze the QR factorization of $T - \sigma I$ as a function of the shift σ but the roundoff errors in execution are best interpreted as being the result of exact QR on a matrix close to $T - \sigma I$. So we turn to general perturbations and consider not just tridiagonal but upper Hessenberg matrices from which σI has already been subtracted. Partition the QR factorization as indicated;

$$\tilde{B} = \begin{bmatrix} B & b \end{bmatrix} = \begin{bmatrix} Q & q \end{bmatrix} \begin{bmatrix} R & r \\ 0^t & \pi \end{bmatrix} = \tilde{Q}\tilde{R}, \tag{4.3.1}$$

where B and $Q \in \mathbf{R}^{n \times (n-1)}$, $R \in \mathbf{R}^{(n-1)\times(n-1)}$, $b, q \in \mathbf{R}^n$, $r \in \mathbf{R}^{n-1}$.

Lemma 4.4: *If $\tilde{B}$ is a smooth function of τ in an open interval $I \subset \mathbf{R}$ and if the QR factorization is given by (4.3.1) with Q having full rank then, for $\tau \in I$,*

$$|\dot{\pi}| \leq \frac{\| b \|}{\sigma_{\min}} \| \dot{B} \| + \| \dot{b} \|$$

where $\sigma_{\min} = \sigma_{\min}[B]$ and $\dot{B}$ is the τ derivative of B.

Proof: Equate final columns in $\tilde{B} = \tilde{Q}\tilde{R}$ to find

$$q\pi = b - Qr = b - QQ^t b \ .$$

It is well known that when Q has full rank then $\tilde{R}$, and hence $\tilde{Q}$, are as smooth as $\tilde{B}$. So differentiate $q\pi$,

$$\dot{q}\pi + q\dot{\pi} = \dot{b} - (QQ^t)^{\cdot} b - (QQ^t)\dot{b} \ ,$$

and recall that $q^t \dot{q} = 0$, $q^t Q = 0^t$ so that on premultiplying by q^t,

$$\dot{\pi} = q^t \dot{b} - q^t (QQ^t)^{\cdot} b - 0^t .$$

Moreover

$$\frac{d}{dt}(QQ^t) = \frac{d}{dt}(BB^+) = B(B^+)^{\cdot} + \dot{B}B^+$$

where B^+ is the Moore-Penrose inverse of B. Since by construction, q is perpendicular to

range(B),

$$\dot{\pi} = q^t\dot{b} - q^t\dot{B}B^tb$$

and the result follows from Cauchy-Schwarz and the standard result $\|B^+\| = 1/\sigma_{min}(B)$. □

The analogue of Theorem 4.1 follows immediately.

Theorem 4.2: *Let unreduced upper Hessenberg $\tilde{B}$ be a smooth function of τ in I. if $\tilde{B}(\tau_0)$ is singular while $B(\tau_0)$ has full rank then $\hat{\beta}_n$, the last subdiagonal entry of $\tilde{R}\tilde{Q}$ satisfies*

$$|\dot{\hat{\beta}}_n(\tau_0)| \le |\dot{\pi}_n(\tau_0)| \le \|\dot{B}\|\,\|b\|/\sigma_{min}(B) + \|\dot{b}\| \quad .$$

This result suggests that the appropriate condition number **for** QR **deflation** is

$$(1 + \|b\|^2/\sigma^2_{min}[B])^{1/2} \quad .$$

The analysis in [Stewart, 1977] did not cover special entries in the QR factors as we have done here.

5. Failure of The Rayleigh Quotient Shift Strategy

There is no proof of convergence for the shift strategies used in the EISPACK implementations of the QR algorithm for unreduced Hessenberg matrices. The actual strategy used there is complicated so we consider now a simple but reasonable procedure: shift by the last diagonal entry of the current matrix. This is called the Rayleigh quotient shift. The figure shown below shows a 3×3 matrix which induces a cycle of length four in QR with these shifts. The cycle is stable and attractive. Moreover other 3×3 Hessenberg matrices can be found which provoke stable cycles of lengths 8, 16, 32, and so on.

This class of examples was developed in [Batterson and Smillie, 1989].

1.052	0.113	0.378		1.093	−0.110	0.110
0.296	0.132	−1.735	→	0.151	−0.416	−0.712
0.	0.572	−0.083		0.	1.616	0.423
	↑				↓	
1.049	−0.158	0.448		1.028	0.148	0.414
0.257	0.539	−0.797	←	0.397	−0.067	−1.716
0.	1.468	−0.489		0.	0.579	0.138

References

Batterson, S. and Smillie, J. **(1990)** Rayleigh Quotient Iteration for Nonsymmetric Matrices, *Math. Comp., vol. 55*, pp. 169-178.

Golub, G. H. and Van Loan, C. F. **(1983)** "Matrix Computation", *The Johns Hopkins University Press, Baltimore, Maryland.*

Parlett, B. N. **(1980)** "The Symmetric Eigenvalue Problem", *Prentice-Hall, Englewood Cliffs, N.J.*

Stewart, G. W. **(1970)** Incorporating Origin Shifts into the QR Algorithm for Symmetric Tridiagonal Matrices, *Comm. Assoc. Comp. Mach., vol. 13*, pp. 365-367.

Stewart, G. W. **(1973)** "Introduction to Matrix Computation", *Academic Press, New York.*

Stewart, G. W. **(1977)** Perturbation Bounds for the QR Factorization of a Matrix, *SIAM J. Numer. Anal., vol. 14*, pp. 509-518.

Wilkinson, J. H. **(1965)** "The Algebraic Eigenvalue Problem", *Oxford University Press, London.*

Prof. Beresford N. Parlett, Department of Mathematics and Computer Science Division of EECS, University of California, Berkeley, CA 94720, USA.

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

EIGENVALUE PROBLEMS AND PRECONDITIONING

Hans-Rudolf Schwarz

Institute of Applied Mathematics, University, Zurich, Switzerland

1. Introduction

Problems of vibration in engineering applications, that are treated by the finite element method using consistent mass matrices, lead to a general eigenvalue problem

$$\mathbf{A}\,\mathbf{x} = \lambda\,\mathbf{B}\,\mathbf{x}\,, \tag{1}$$

where **A** and **B** are the stiffness and mass matrices, respectively. The order n of the matrices is in general large, corresponding to the number of nodal variables. The matrices **A** and **B** are symmetric, and without loss of generality we can assume **A** and **B** to be positive definite. The matrices are sparse and have the same, in general irregular, sparsity structure. We look for the m smallest eigenvalues

$$0 < \lambda_1 < \lambda_2 \leq \lambda_3 \leq \ldots \leq \lambda_m \tag{2}$$

and for the corresponding eigenvectors $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m$ of (1) such that

$$\mathbf{A}\mathbf{z}_j = \lambda_j \mathbf{B}\mathbf{z}_j, \quad \mathbf{z}_j^T \mathbf{B}\mathbf{z}_j = 1, \quad (j=1,2,\ldots,m). \tag{3}$$

The number m of the desired eigenpairs $(\lambda_j, \mathbf{z}_j)$ is small compared with the order n of the matrices.

For solving the partial eigenvalue problem we present in the following an efficient procedure that exploits the sparsity of the matrices to full extent. The algorithm should be able to solve the problem with a minimum of memory space so that problems as large as possible can be treated on small computers such as PC's or workstations. The usual methods such as vector iteration, bisection or the famous Lanczos algorithm require the decomposition of at least one matrix, and as a consequence the memory requirements for this step may be so high that a treatment of

the eigenvalue problem gets simply impossible on the computer at hand without specific measures. Especially three-dimensional structures, for example a radar dome, give rise to eigenvalue problems with matrices **A** and **B** having a relatively wide, variable band width but being extremely sparse inside the band. The profile of the matrix to be decomposed may simply be too big for storing the matrix.

Therefore only algorithms based on the minimization of the Rayleigh quotient are adequate, and the method of conjugate gradients will be considered for finding the minimum of the Rayleigh quotient to obtain the eigenpair (λ_1, z_1). The convergence rate of this elementary method can be improved by a preconditioning of the given eigenvalue problem in full analogy to systems of linear equations [10]. If a preconditioner with the same sparsity pattern is used and if the preconditioning is performed in an implicit way the computational effort per iteration step is only increased by about fifty percent. Moreover, the preconditioning can be maintained for the computation of the higher eigenpairs if the deflation process is accompanied by the technique of rank-one modifications. A proper choice of the preconditioner makes the resulting algorithm at least competitive with all methods that exploit the sparsity structure of the matrices.

2. The basic algorithm

The minimum of the Rayleigh quotient $R[\mathbf{x}]$ corresponding to (1) is equal to λ_1 and is attained at z_1:

$$\min_{\mathbf{x}\neq\mathbf{0}} R[\mathbf{x}] = \min_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{B}\mathbf{x}} = \lambda_1 = \frac{z_1^T\mathbf{A}z_1}{z_1^T\mathbf{B}z_1} \tag{4}$$

The minimum of $R[\mathbf{x}]$ is determined iteratively by means of the method of conjugate gradients [3,5,7,10,11,13]. For an iterate $\mathbf{x}_k$ the corresponding gradient of $R[\mathbf{x}]$

$$\mathbf{g}_k = \mathbf{g}(\mathbf{x}_k) = \operatorname{grad} R[\mathbf{x}_k] = \frac{2}{\mathbf{x}_k^T\mathbf{B}\mathbf{x}_k} \left\{\mathbf{A}\mathbf{x}_k - R[\mathbf{x}_k]\mathbf{B}\mathbf{x}_k\right\} \tag{5}$$

is used to fix the direction of descent $\mathbf{p}_{k+1}$ in which $R[\mathbf{x}]$ is minimized. If $\mathbf{x}_0$ denotes the initial vector, the directions of descent are defined as follows

$$\mathbf{p}_1 = -\mathbf{g}_0, \quad \mathbf{p}_{k+1} = -\mathbf{g}_k + \frac{\mathbf{g}_k^T\mathbf{g}_k}{\mathbf{g}_{k-1}^T\mathbf{g}_{k-1}} \mathbf{p}_k, \quad (k=1,2,3,\ldots). \tag{6}$$

The subsequent iterate

$$x_{k+1} = x_k + \delta_{k+1} p_{k+1}, \qquad (k=0,1,2,...) \tag{7}$$

is obtained from, setting $\delta_{k+1} = \delta$,

$$R[x_{k+1}] = \frac{x_k^T A x_k + 2\delta p_{k+1}^T A x_k + \delta^2 p_{k+1}^T A p_{k+1}}{x_k^T B x_k + 2\delta p_{k+1}^T B x_k + \delta^2 p_{k+1}^T B p_{k+1}} = \text{min!} \tag{8}$$

With the auxiliary quantities

$$\alpha := x_k^T A x_k, \quad \beta := p_{k+1}^T A x_k, \quad \gamma := p_{k+1}^T A p_{k+1} \tag{9}$$

$$\rho := x_k^T B x_k, \quad \sigma := p_{k+1}^T B x_k, \quad \tau := p_{k+1}^T B p_{k+1} \tag{10}$$

the value δ_{k+1}, that yields the minimal value $R[x_{k+1}]$, is defined by one of the two solutions of a quadratic equation derived from (8) and is given by the following, numerically stable formula that avoids cancellation of digits

$$\delta_{k+1} = \begin{cases} [\alpha\tau - \gamma\rho + \sqrt{\Delta}\,]/[2(\gamma\sigma - \beta\tau)] & \text{if } \alpha\tau - \gamma\rho > 0 \\ 2(\beta\rho - \alpha\sigma)/[\alpha\tau - \gamma\rho - \sqrt{\Delta}\,] & \text{otherwise} \end{cases} \tag{11}$$

where

$$\Delta := (\alpha\tau - \gamma\rho)^2 - 4(\gamma\sigma - \beta\tau)(\beta\rho - \alpha\sigma) > 0. \tag{12}$$

If we define the vectors $v_k := Ax_k$, $\hat{v}_k := Bx_k$, which will be recursively computed, as well as the vectors $w_k := Ap_k$, $\hat{w}_k := Bp_k$ the basic algorithm for finding the minimum of the Rayleigh quotient and hence the eigenpair (λ_1, z_1) by means of the method of conjugate gradients can be summarized as follows:

Start: Choose $x_0 \neq 0$;

$$v_0 = Ax_0, \quad \hat{v}_0 = Bx_0; \quad \alpha_0 = x_0^T v_0, \quad \rho_0 = x_0^T \hat{v}_0;$$

$$q_0 = \alpha_0/\rho_0 ;$$

Iteration $(k=1,2,3, \ldots)$:

$$g_{k-1} = 2(v_{k-1} - q_{k-1}\,\hat{v}_{k-1})/\rho_{k-1}$$

if $k=1$: $p_k = -g_{k-1}$;

if $k>1$: $\varepsilon_{k-1} = g_{k-1}^T g_{k-1} / g_{k-2}^T g_{k-2}$;

$$p_k = -g_{k-1} + \varepsilon_{k-1} p_{k-1} ;$$

$$w_k = Ap_k , \quad \hat{w}_k = Bp_k ;$$

$$\beta = x_{k-1}^T w_k, \quad \gamma = p_k^T w_k, \quad \sigma = x_{k-1}^T \hat{w}_k, \quad \tau = p_k^T \hat{w}_k ;$$

(13)

compute δ_k from (11) ;

$x_k = x_{k-1} + \delta_k p_k$,

$v_k = v_{k-1} + \delta_k w_k$, $\hat{v}_k = \hat{v}_{k-1} + \delta_k \hat{w}_k$;

$\alpha_k = x_k^T v_k$, $\rho_k = x_k^T \hat{v}_k$; $q_k = \alpha_k / \rho_k$;

test on convergence

A typical iteration step requires two multiplications of a sparse matrix with a vector, seven scalar products, five triads and a multiplication of a vector by a scalar.

3. Convergence and preconditioning

The convergence of the sequence of iterates x_k of the algorithm (13) towards the direction of z_1 essentially depends on the condition number of the Hessian matrix $H(x)$ of $R[x]$

$$H(x) = \frac{2}{x^T Bx} \left\{ A - R[x]B - g(x)(Bx)^T - (Bx)g(x)^T \right\}$$

evaluated at z_1 [11]. Due to $z_1^T B z_1 = 1$ and $g(z_1) = 0$ $H(z_1) = 2(A - \lambda_1 B)$ holds. Because the eigenvectors z_1 of (1) satisfy

$$H(z_1)z_j = 2(A - \lambda_1 B)z_j = 2(\lambda_j - \lambda_1)Bz_j, \quad (j=1,2, \ldots ,n), \tag{14}$$

the norm of $H(z_1)$, subordinate to the appropriate pair of vector norms $\|x\|_B^2 := (x^T Bx)$ and $\|x\|_{B^{-1}}^2 := (x^T B^{-1} x)$, is given by [4]

$$\|H(z_1)\|_{B,B^{-1}} = 2(\lambda_n - \lambda_1). \tag{15}$$

Since $H(z_1)$ is positive semidefinite the corresponding condition number is defined by [11]

$$\kappa_{B,B^{-1}}(H(z_1)) = (\lambda_n - \lambda_1)/(\lambda_2 - \lambda_1) , \tag{16}$$

where we have tacitely assumed $\lambda_2 > \lambda_1$. The number of steps required to compute z_1 with a relative accuracy ε can be estimated by the reasonable formula

$$\mu_e \approx \frac{\ln(2/\varepsilon)}{2} \sqrt{\kappa_{B,B^{-1}}(H(z_1))} . \tag{17}$$

We adopt the estimate (17) from [1] that bounds the number of steps needed to

solve systems of linear equations with a relative accuracy ε by means of the method of conjugate gradients, because the algorithm (13) behaves asymptotically like that for solving linear equations. Obviously, the estimate (17) yields prohibitively large values μ_e for large condition numbers of $H(z_1)$.

However, the condition number of the Hessian matrix can be essentially decreased by means of a preconditioning. For this aim we set

$$y := C^T x, \quad x = C^{-T} y, \quad C \text{ nonsingular}, \tag{18}$$

and transform the Rayleigh quotient (4) into

$$R[x] = \frac{y^T C^{-1} A\, C^{-T} y}{y^T C^{-1} B\, C^{-T} y} = \frac{y^T \tilde{A} y}{y^T \tilde{B} y} =: \tilde{R}[y], \tag{19}$$

where the resulting matrices $\tilde{A} := C^{-1}AC^{-T}$ and $\tilde{B} := C^{-1}BC^{-T}$ are symmetric and positive definite. The given eigenvalue problem (1) is thus transformed to $\tilde{A}y = \lambda \tilde{B}y$ having the eigenvectors $y_j = C^T z_j$, $(j=1,2, \ldots ,n)$ that correspond to the eigenvalues λ_j of (1). The Hessian matrix of $\tilde{R}[y]$, evaluated at the eigenvector y_1, is

$$\tilde{H}(y_1) = 2(\tilde{A} - \lambda_1 \tilde{B}) = 2C^{-1}(A - \lambda_1 B)C^{-T} = C^{-1}H(z_1)C^{-T}.$$

It is congruent to $H(z_1)$ and, moreover, similar to

$$K := C^{-T}\tilde{H}(y_1)C^T = 2(CC^T)^{-1}(A - \lambda_1 B) = (CC^T)^{-1}H(z_1). \tag{20}$$

The matrix $M := (CC^T)$ is called the preconditioner. An indication of its proper choice is given by the following

Theorem If $M = A$ holds, the condition numbers of $\tilde{H}(y_1)$ and $H(z_1)$ are related as follows

$$\kappa_{B,B}(\tilde{H}(y_1)) = \frac{\lambda_2}{\lambda_n}\, \kappa_{B,B^{-1}}(H(z_1)), \quad \lambda_2 > \lambda_1 . \tag{21}$$

Proof From the assumption on M we get from (20)

$$K = 2A^{-1}(A - \lambda_1 B) = 2(I - \lambda_1 A^{-1}B),$$

and hence we have for the eigenvectors z_j of (1)

$$Kz_j = 2(I - \lambda_1 A^{-1}B)z_j = 2\left(1 - \frac{\lambda_1}{\lambda_2}\right) z_j, \quad (j=1,2, \ldots , n) .$$

Due to the similarity of the matrices K and $\tilde{H}(y_1)$ the eigenvalues of $0.5\tilde{H}(y_1)$ are, ordered by magnitude,

$$0 = 1 - \frac{\lambda_1}{\lambda_1} < 1 - \frac{\lambda_1}{\lambda_2} \leq 1 - \frac{\lambda_1}{\lambda_3} \leq \ldots \leq 1 - \frac{\lambda_1}{\lambda_n} < 1 . \qquad (22)$$

Consequently, the condition number of the Hessian matrix $\tilde{\mathbf{H}}(\mathbf{y}_1)$, now corresponding to the appropriate B-vector norm, is given by

$$\kappa_{B,B}(\tilde{\mathbf{H}}(\mathbf{y}_1)) = \frac{1 - \lambda_1/\lambda_n}{1 - \lambda_1/\lambda_2} = \frac{\lambda_2}{\lambda_n} \frac{\lambda_n - \lambda_1}{\lambda_2 - \lambda_1} = \frac{\lambda_2}{\lambda_n} \kappa_{B,B^{-1}}(\mathbf{H}(\mathbf{z}_1)).$$

□

Apart from the fact that different norms occur in (21), the condition number is substantially decreased whenever $\lambda_2 << \lambda_n$ holds. But this property applies in general to practical applications. Moreover, from (22) it is unlikely that the eigenvalues of $0.5\tilde{\mathbf{H}}(\mathbf{y}_1)$ show a special distribution if those of (1) do not, and hence the condition number of $\tilde{\mathbf{H}}(\mathbf{y}_1)$ is the decisive factor with respect to the convergence rate.

From the above theorem it is seen that any preconditioner $\mathbf{M} = (\mathbf{C}\mathbf{C}^T) = \mathbf{A}$ would provide an excellent choice to improve the convergence behaviour. Thus the Cholesky decomposition $\mathbf{A} = \mathbf{C}\mathbf{C}^T$, where $\mathbf{C}$ is a left triangular matrix, would be adequate. However, such a decomposition has to be avoided in order to exploit the sparsity structure. Therefore a nonoptimal, appropriate preconditioner must be chosen, for instance one that is based on a variant of an incomplete Cholesky decomposition [1,8,14].

4. Reformulation of the algorithm

The straightforward implementation of the preconditioning in the algorithm (13) by replacing $\mathbf{A}$ and $\mathbf{B}$ by $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$, respectively, and $\mathbf{x}_k$ by $\mathbf{y}_k$ is not appropriate. It is more adequate to perform the preconditioning in an implicit way. For this purpose the algorithm (13) is first formulated in terms of $\tilde{\mathbf{A}}$, $\tilde{\mathbf{B}}$ and $\mathbf{y}$, and then the formulae are treated as usual [1,14]. We use the relations

$$\begin{aligned} \tilde{\mathbf{v}}_k &:= \tilde{\mathbf{A}}\mathbf{y}_k = \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-T}\mathbf{C}^T\mathbf{x}_k = \mathbf{C}^{-1}(\mathbf{A}\mathbf{x}_k) = \mathbf{C}^{-1}\mathbf{v}_k \\ \tilde{\hat{\mathbf{v}}}_k &:= \tilde{\mathbf{B}}\mathbf{y}_k = \mathbf{C}^{-1}\mathbf{B}\mathbf{C}^{-T}\mathbf{C}^T\mathbf{x}_k = \mathbf{C}^{-1}(\mathbf{B}\mathbf{x}_k) = \mathbf{C}^{-1}\hat{\mathbf{v}}_k \end{aligned} \qquad (23)$$

to obtain for the gradient

$$\tilde{\mathbf{g}}_{k-1} = (\tilde{\mathbf{v}}_{k-1} - q_{k-1}\tilde{\hat{\mathbf{v}}}_{k-1})(2/\rho_{k-1}) = \mathbf{C}^{-1}(\mathbf{v}_{k-1} - q_{k-1}\hat{\mathbf{v}}_{k-1})(2/\rho_{k-1}) = \mathbf{C}^{-1}\mathbf{g}_{k-1}, \qquad (24)$$

where the vector $\mathbf{g}_{k-1}$, defined in (24), denotes the gradient corresponding to the iterate $\mathbf{x}_{k-1}$ of the preconditioned algorithm and will in general be different from

the vector $\mathbf{g}_{k-1}$ of the algorithm (13). In order to simplify the numerator of $\tilde{\varepsilon}_{k-1}$ we define the vector $\mathbf{h}_k$ from

$$(\mathbf{C}\mathbf{C}^T)\mathbf{h}_k = \mathbf{g}_k \tag{25}$$

to get

$$\zeta = \tilde{\mathbf{g}}_{k-1}{}^T\tilde{\mathbf{g}}_{k-1} = \mathbf{g}_{k-1}{}^T\mathbf{C}^{-T}\mathbf{C}^{-1}\mathbf{g}_{k-1} = \mathbf{g}_{k-1}{}^T(\mathbf{C}\mathbf{C}^T)^{-1}\mathbf{g}_{k-1} = \mathbf{g}_{k-1}{}^T\mathbf{h}_{k-1}\,.$$

The formula for the direction of descent is multiplied by $\mathbf{C}^{-T}$ to obtain

$$\mathbf{C}^{-T}\tilde{\mathbf{p}}_k = -\mathbf{C}^{-T}\tilde{\mathbf{g}}_{k-1} + \tilde{\varepsilon}_{k-1}\mathbf{C}^{-T}\tilde{\mathbf{p}}_{k-1} = -(\mathbf{C}\mathbf{C}^T)^{-1}\mathbf{g}_{k-1} + \tilde{\varepsilon}_{k-1}(\mathbf{C}^{-T}\tilde{\mathbf{p}}_{k-1})$$

a recursion formula for the vectors $\mathbf{s}_k = \mathbf{C}^{-T}\tilde{\mathbf{p}}_k$ of the form

$$\mathbf{s}_k = -\mathbf{h}_{k-1} + \tilde{\varepsilon}_{k-1}\mathbf{s}_{k-1}.$$

Furthermore we define vectors $\mathbf{w}_k$ and $\hat{\mathbf{w}}_k$ from

$$\tilde{\mathbf{w}}_k = \mathbf{A}\tilde{\mathbf{p}}_k = \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-T}\tilde{\mathbf{p}}_k = \mathbf{C}^{-1}(\mathbf{A}\mathbf{s}_k) = \mathbf{C}^{-1}\mathbf{w}_k$$
$$\hat{\tilde{\mathbf{w}}}_k = \mathbf{B}\tilde{\mathbf{p}}_k = \mathbf{C}^{-1}\mathbf{B}\mathbf{C}^{-T}\tilde{\mathbf{p}}_k = \mathbf{C}^{-1}(\mathbf{B}\mathbf{s}_k) = \mathbf{C}^{-1}\hat{\mathbf{w}}_k\,.$$

Thus the vector $\mathbf{s}_k$ has to be multiplied by the matrices $\mathbf{A}$ and $\mathbf{B}$ instead of the vector of descent $\mathbf{p}_k$.

By simple substitutions the representations of the scalars $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\gamma}$, $\tilde{\rho}$, $\tilde{\sigma}$ and $\tilde{\tau}$ can be simplified and the recursion formulae for $\tilde{\mathbf{x}}_k$, $\tilde{\mathbf{v}}_k$ and $\hat{\tilde{\mathbf{v}}}_k$ can be reformulated in an obvious way. Hence the RQPCG algorithm of minimizing the Rayleigh quotient $R[\mathbf{x}]$ with implicit preconditioning is as follows:

Choose the preconditioner $\mathbf{M}$

Start: Choose $\mathbf{x}_o \neq \mathbf{0}$; $\mathbf{v}_o = \mathbf{A}\mathbf{x}_o$, $\hat{\mathbf{v}}_o = \mathbf{B}\mathbf{x}_o$; $\mathbf{s}_o = \mathbf{0}$;

$\alpha_o = \mathbf{x}_o{}^T\mathbf{v}_o$, $\rho_o = \mathbf{x}_o{}^T\hat{\mathbf{v}}_o$; $q_o = \alpha_o/\rho_o$; $\zeta_a = 1$;

Iteration $(k=1,2,3,\ldots)$:

$\mathbf{g}_{k-1} = (\mathbf{v}_{k-1} - q_{k-1}\hat{\mathbf{v}}_{k-1})\,(2/\rho_{k-1})$;

$\mathbf{M}\mathbf{h}_{k-1} = \mathbf{g}_{k-1}$; (preconditioning step)

$\zeta = \mathbf{g}_{k-1}{}^T\mathbf{h}_{k-1}$; $\varepsilon_{k-1} = \zeta/\zeta_a$;

$\mathbf{s}_k = -\mathbf{h}_{k-1} + \varepsilon_{k-1}\mathbf{s}_{k-1}$;

$\mathbf{w}_k = \mathbf{A}\mathbf{s}_k$, $\hat{\mathbf{w}}_k = \mathbf{B}\mathbf{s}_k$; (26)

$\beta = \mathbf{x}_{k-1}^T \mathbf{w}_k$, $\gamma = \mathbf{s}_k^T \mathbf{w}_k$, $\sigma = \mathbf{x}_{k-1}^T \hat{\mathbf{w}}_k$, $\tau = \mathbf{s}_k^T \hat{\mathbf{w}}_k$;

compute δ_k from (11) ;

$\mathbf{x}_k = \mathbf{x}_{k-1} + \delta_k \mathbf{s}_k$,

$\mathbf{v}_k = \mathbf{v}_{k-1} + \delta_k \mathbf{w}_k$, $\hat{\mathbf{v}}_k = \hat{\mathbf{v}}_{k-1} + \delta_k \hat{\mathbf{w}}_k$,

$\alpha_k = \mathbf{x}_k^T \mathbf{v}_k$, $\rho_k = \mathbf{x}_k^T \hat{\mathbf{v}}_k$; $q_k = \alpha_k / \rho_k$; $\zeta_a = \zeta$;

test on convergence

Each iteration step of the algorithm (26) requires the determination of the vector $\mathbf{h}_{k-1}$ from $\mathbf{M}\mathbf{h}_{k-1} = \mathbf{g}_{k-1}$ and the subsequent multiplication of $\mathbf{s}_{k-1}$ by $\mathbf{A}$ and $\mathbf{B}$. Moreover, there are again seven scalar products, five triads and one multiplication of a vector by a scalar. If the computation of $\mathbf{h}_{k-1}$ requires the same amount of work as a multiplication of a vector by a matrix $\mathbf{A}$, then the amount of computational work per iteration step does only increase by a matrix-vector multiplication in comparison with that of the algorithm (13) without preconditioning. This holds if the preconditioner $\mathbf{M}$ is defined by a lower triangular matrix $\mathbf{C}$ that has the same sparsity structure as the lower part of $\mathbf{A}$.

5. Higher eigenvalues and preconditioning

The algorithm (26) produces the smallest eigenvalue and a corresponding eigenvector of the matrix pair $(\mathbf{A},\mathbf{B})$. In order to exploit the numerical property of the Rayleigh quotient for the computation of the higher eigenvalues and to transfer the preconditioning to this problem the subsequent eigenpairs are determined from modified eigenvalue problems using a deflation based on a partial shift of the spectrum. On the assumption that the (r-1) first eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \lambda_{r-1}$ and the B-normalized eigenvectors $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{r-1}$ are approximately known, the next eigenpair $(\lambda_r, \mathbf{z}_r)$ can be determined by minimizing the Rayleigh quotient of the matrix pair $(\mathbf{A}_r,\mathbf{B})$, where $\mathbf{A}_r$ is defined by

$$\mathbf{A}_r := \mathbf{A} + \sum_{\nu=1}^{r-1} \sigma_\nu \, (\mathbf{B}\mathbf{z}_\nu)(\mathbf{B}\mathbf{z}_\nu)^T \tag{27}$$

with shifts $\sigma_\nu > 0$ that satisfy $\lambda_\nu + \sigma_\nu > \lambda_r$, $(\nu = 1,2, \ldots, r\text{-}1)$. This is clear from the fact that the eigenvalues and eigenvectors of $\mathbf{A}_r\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$ satisfy, because of

the B-orthonormality of the z_j,

$$A_r z_j = Az_j + \sum_{\nu=1}^{r-1} \sigma_\nu (Bz_\nu)(Bz_\nu)^T z_j$$

$$= \begin{cases} (\lambda_j + \sigma_j)\, Bz_j\,, & j = 1,2, \ldots, r\text{-}1\ ; \\ \lambda_j\, Bz_j\,, & j = r, r+1, \ldots, n. \end{cases}$$

Hence, (λ_r, z_r) is obtained by means of the algorithm (26) by replacing A by A_r. The required multiplication of a vector x by A_r is realized by adding to Ax scalar multiples of (Bz_ν), where the scalars are essentially given by the scalar products $(Bz_\nu)^T x$. This procedure requires the storing of the auxiliary vectors (Bz_ν), ($\nu = 1, 2, \ldots, r\text{-}1$), and the computational effort per iteration step is increased, so far, by (r-1) scalar products and triads, each.

We show next the numerical stability of the deflation process (27). We assume that an approximation $\hat{z}_1$ of the first eigenvector z_1 has been determined with a relative accuracy ε and being B-normalized. We want to study the influence of $\hat{z}_1$ to the higher eigenpairs (λ_i, z_i) for $i \geq 2$. The computed approximation $\hat{z}_1$ can be assumed to have the following representation

$$\hat{z}_1 = c_1 z_1 + \varepsilon \sum_{i=2}^{n} c_i z_i \quad \text{with} \quad \| \sum_{i=2}^{n} c_i z_i \|_B = 1. \tag{28}$$

Hence the coefficients c_i satisfy

$$c_1^2 + \varepsilon^2 \sum_{i=2}^{n} c_i^2 = c_1^2 + \varepsilon^2 = 1 \quad \text{and} \quad c_1 \cong 1 - 0.5\varepsilon^2. \tag{29}$$

The matrix $\hat{A}_2$, that is used in the deflation process instead of A_2, is given by

$$\hat{A}_2 = A + \sigma_1 (B\hat{z}_1)(B\hat{z}_1)^T$$

$$= A + \sigma_1 [c_1 Bz_1 + \varepsilon \sum_{i=2}^{n} c_i Bz_i][c_1 Bz_1 + \varepsilon \sum_{j=2}^{n} c_j Bz_j]^T$$

$$= A_2 + \varepsilon\sigma_1 \sum_{j=2}^{n} c_j [(Bz_1)(Bz_j)^T + (Bz_j)(Bz_1)^T] \tag{30}$$

$$+ \varepsilon^2 \sigma_1 [-(Bz_1)(Bz_1)^T + \sum_{i=2}^{n} \sum_{j=2}^{n} c_i c_j (Bz_i)(Bz_j)^T] + O(\varepsilon^3).$$

For estimating the influence of the error in $\hat{z}_1$ we use a theorem from [9].

Theorem Let **C** be a symmetric matrix with eigenpairs (λ_j, y_j). Let **z** be a normed vector with $\|z\|_2 = 1$, having the Ritz-value $\Theta := R[z] = z^T Cz$ and the residual $\rho := Cz - \Theta z$. Then there exists an eigenpair (λ_j, y_j) of **C**, assumed to be simple, with $|\lambda_j - \Theta| = \min_k |\lambda_k - \Theta|$ satisfying

$$|\sin \psi| \le \|\rho\|_2/d \quad \text{and} \quad |\Theta - \lambda_j| \le \|\rho\|_2^2/d ,$$

where $\psi := \sphericalangle(z, y_j)$ and $d := \min_{k \neq j} |\lambda_j - \lambda_k|$.

The straight-forward generalization to the general eigenvalue problem yields the

Theorem Let **A** and **B** be symmetric matrices and **B** positive definite and (λ_j, z_j) be the eigenpairs of $Ax = \lambda Bx$. If **x** is any vector with $\|x\|_B = 1$, whose Ritz-value is $\Theta := R[x] = x^T Ax$ and whose residual is $r := Ax - \Theta Bx$, then there exists an eigenpair (λ_j, z_j) of the matrix pair **(A, B)** with $|\lambda_j - \Theta| = \min_k |\lambda_k - \Theta|$ satisfying the inequalities

$$|\sin \psi| \le \|r\|_{B^{-1}}/d \quad \text{and} \quad |\Theta - \lambda_j| \le \|r\|^2_{B^{-1}}/d, \tag{31}$$

where $\psi := \sphericalangle(x, z_j)_B$ and $d := \min_{j \neq k} |\lambda_j - \lambda_k|$.

This theorem is applied to $\hat{A}_2 x = \lambda Bx$ with $x = z_k$, $(k \ge 2)$. From (30) we get

$$\hat{A}_2 z_k = A_2 z_k + \varepsilon\sigma_1 \sum_{j=2}^{n} c_j[(Bz_1)(Bz_j)^T + (Bz_j)(Bz_1)^T]z_k$$

$$+ \varepsilon^2\sigma_1[-(Bz_1)(Bz_1)^T + \sum_{i=2}^{n}\sum_{j=2}^{n} c_i c_j (Bz_i)(Bz_j)^T]z_k + O(\varepsilon^3)$$

$$= A_2 z_k + \varepsilon\sigma_1 c_k (Bz_1) + \varepsilon^2\sigma_1 c_k \sum_{i=2}^{n} c_i (Bz_i) + O(\varepsilon^3) ;$$

$$\Theta := \hat{R}[z_k] = z_k^T \hat{A}_2 z_k = z_k^T A_2 z_k + \varepsilon^2\sigma_1 c_k^2 + O(\varepsilon^3)$$

$$= \lambda_k + \varepsilon^2\sigma_1 c_k^2 + O(\varepsilon^3) ;$$

$$\begin{aligned} \mathbf{r} &:= \hat{\mathbf{A}}_2 z_k - \Theta \mathbf{B} z_k \\ &= \lambda_k \mathbf{B} z_k + \varepsilon\sigma_1 c_k (\mathbf{B} z_1) + \varepsilon^2 \sigma_1 c_k \sum_{i=2}^{n} c_i (\mathbf{B} z_i) - (\lambda_k + \varepsilon^2 \sigma_1 c_k^2) \mathbf{B} z_k + O(\varepsilon^3) \\ &= \varepsilon\sigma_1 c_k (\mathbf{B} z_1) + O(\varepsilon^2) \; ; \end{aligned}$$

$$\|\mathbf{r}\|_{B^{-1}} = \|\varepsilon\sigma_1 c_k (\mathbf{B} z_1) + O(\varepsilon^2)\|_{B^{-1}} = \varepsilon\sigma_1 |c_k| + O(\varepsilon^2) \; .$$

Hence we get from (31) for the angles $\psi_k := \sphericalangle(\hat{z}_k, z_k)_B$ between the eigenvectors $\hat{z}_k$ of the perturbed eigenvalue problem and the exact eigenvectors z_k and for the corresponding eigenvalues $\hat{\lambda}_k$ the following estimates

$$|\sin \psi_k| \le \varepsilon\sigma_1 |c_k| / d_k \quad \text{and} \quad |\hat{\lambda}_k - \lambda_k| \le \varepsilon^2 \sigma_1^2 c_k^2 / d_k, \quad (k \ge 2)$$

with $d_k := \min\limits_{i \ne k} |\lambda_k - \lambda_i|$.

The errors induced to z_k from $\hat{z}_1$ are of the same order and depend on the magnitude $|c_k| \le 1$ of the component of z_k present in $\hat{z}_1$, and are only critical in case of neighbouring eigenvalues. Moreover, if a proper scaling of the matrices **A** and **B** is used, the shift σ_1 is less than 1, too. This result can be sharpened by a different analysis.

If the preconditioner **M** is kept fixed for minimizing the Rayleigh quotient of the matrix pair $(\mathbf{A}_r, \mathbf{B})$, the preconditioning effect is lost for increasing r. Moreover, the shifts σ_ν should be chosen in such a way that even $\lambda_\nu + \sigma_\nu > \lambda_{r+1}$ hold for $\nu = 1,2, \ldots, r\text{-}1$, but too a large choice of the shifts diminishes the preconditioning effect quite drastically [10,13] or can even be definitely lost, because the constant matrix **M** does no more represent a meaningful approximation of $\mathbf{A}_r$. Therefore it is necessary to use an equivalent preconditioner for the matrix $\mathbf{A}_r$ that takes into account the deflation steps. But $\mathbf{A}_r$ results from **A** by a sequence of rank- one modifications, and hence it is almost natural to define the corresponding preconditioner $\mathbf{M}_r$, that is the exact one in case of $\mathbf{M} = \mathbf{A}$, as follows

$$\mathbf{M}_r := \mathbf{M} + \sum_{\nu=1}^{r-1} \sigma_\nu (\mathbf{B} z_\nu)(\mathbf{B} z_\nu)^T \; . \tag{32}$$

The major preconditioning step of the algorithm (26), namely the solution of the system of equations $\mathbf{M} h_{k-1} = g_{k-1}$ for h_{k-1}, has to be replaced by

$$\left\{\mathbf{M} + \sum_{\nu=1}^{r-1} \sigma_\nu (\mathbf{B}\mathbf{z}_\nu)(\mathbf{B}\mathbf{z}_\nu)^T\right\} \mathbf{h} = \mathbf{g} , \tag{33}$$

where we omitted the indices for simplicity. To solve (33) for $\mathbf{h}$ we can apply the well known technique of rank-one modifications of systems of linear equations. Let us consider the case $r=2$ for deriving the essential steps. Thus we have to solve the system of equations

$$\left\{\mathbf{M} + \sigma_1 (\mathbf{B}\mathbf{z}_1)(\mathbf{B}\mathbf{z}_1)^T\right\} \mathbf{h} = \mathbf{g} . \tag{34}$$

We assume that the solution of $\mathbf{M}\mathbf{h}^{(o)} = \mathbf{g}$ can easily be obtained. Then the solution vector $\mathbf{h}$ of (34) is given by means of the Sherman-Morrison formula [4]

$$\mathbf{h} = \mathbf{h}^{(o)} - \frac{\mathbf{v}^T\mathbf{h}^{(o)}}{1 + \mathbf{v}^T\mathbf{y}} \mathbf{y} , \tag{35}$$

where $\mathbf{v} := (\mathbf{B}\mathbf{z}_1)$ and $\mathbf{y}$ is the solution of $\mathbf{M}\mathbf{y} = \sigma_1(\mathbf{B}\mathbf{z}_1)$. To apply (35) we have to solve the system

$$\mathbf{M}\mathbf{y}_1 = (\mathbf{B}\mathbf{z}_1) \tag{36}$$

for the auxiliary vector $\mathbf{y}_1$, that depends on the preconditioner $\mathbf{M}$, to get for

$$\mathbf{v}^T\mathbf{y} = \sigma_1 \mathbf{v}^T\mathbf{y}_1 = \sigma_1 (\mathbf{B}\mathbf{z}_1)^T\mathbf{y}_1 = \sigma_1 \cdot c_1 . \tag{37}$$

Hence, for a known eigenvector $\mathbf{z}_1$ we have to compute $\mathbf{y}_1$ from (36) and the scalar product $c_1 := (\mathbf{B}\mathbf{z}_1)^T\mathbf{y}_1$, that can be used in the denominator of (35) with variable value of σ_1. With these preliminaries the formula (35) finally consists essentially of computing the scalar product $(\mathbf{B}\mathbf{z}_1)^T\mathbf{h}^{(o)}$ and of subtracting a multiple of $\mathbf{y}_1$ from $\mathbf{h}^{(o)}$. For the general case (33) the computation of $\mathbf{h}$ can be summarized as follows, if the vectors $\mathbf{y}_\nu$ are at hand as the solutions of $\mathbf{M}\mathbf{y}_\nu = \mathbf{B}\mathbf{z}_\nu$, ($\nu = 1,2, \ldots, r-1$) as well as the corresponding constants $c_\nu := (\mathbf{B}\mathbf{z}_\nu)^T\mathbf{y}_\nu$:

$$\begin{aligned} &\mathbf{M}\,\mathbf{h}^{(o)} = \mathbf{g} \quad \rightarrow \quad \mathbf{h}^{(o)} \\ &\text{for } \nu = 1,2, \ldots, r-1 : \\ &\qquad \mathbf{h}^{(\nu)} = \mathbf{h}^{(\nu-1)} - \frac{\sigma_\nu}{1 + c_\nu\sigma_\nu} \left[(\mathbf{B}\mathbf{z}_\nu)^T\mathbf{h}^{(\nu-1)}\right] \mathbf{y}_\nu \end{aligned} \tag{38}$$

The computational effort of an iteration step of the algorithm (26) is increased by

(r-1) additional scalar products and triads, due to the preconditioning step (33). For an efficient implementation of (38) the vectors $\mathbf{y}_\nu$ must be stored, too.

The difficulties mentioned in [10,13] caused by too a large choice of the shifts σ_ν are largely eliminated by the improved handling of the preconditioning (38), and the additional computational work pays out. Experience still indicates that too large shifts σ_ν of the computed eigenvalues have some negative influence on the convergence, so that they must be chosen in a reasonable way, taking into account the distribution of the eigenvalues. A possible strategy defines the σ_ν to be a multiple of the lastly computed eigenvalue λ_{r-1}, where the factor decreases from a starting value to a limiting value larger than one with increasing index r. However, in case of pathological eigenvalue distributions every rule for defining the shifts may happen to fail. A potential wrong choice with $\lambda_1+\sigma_{r-1}<\lambda_r$ can easily be discovered because the resulting minimal value of the Rayleigh quotient will be equal to $\lambda_1+\sigma_{r-1}$. A bad choice with $\lambda_r < \lambda_1+\sigma_{r-1} < \lambda_{r+1}$ results in a slow convergence.

In the previous discussion we assumed a general preconditioner **M**. Apart from the above mentioned variants of incomplete Cholesky decompositions of **A** or incomplete LDU factorizations other preconditioners **M** are possible which yield even better approximations of **A** or solve the aspects of vectorization or parallelization more appropriately. Thus it can be advantageous to implement the multiplications of a vector by **A** and **B** on the base of the finite element matrices without compiling the stiffness and mass matrices **A** and **B** explicitely. The analogous idea can be adopted to the preconditioning step for which the matrix **C** is defined by means of complete Cholesky decompositions of element stiffness matrices that are made nonsingular by adding certain contributions of the corresponding diagonal entries of **A** to the diagonal elements [6]. An even better preconditioner is obtained by means of complete Cholesky decompositions of submatrices of **A** that correspond to the elements [2]. The matrix vector multiplications as well as the forward and backward substitutions of the preconditioning step can be performed in a parallel fashion over the set of element matrices, thus guaranteeing a high degree of vectorization also for these two steps of the algorithm (26).

6. Numerical examples

A first test example is presented for showing the efficiency of the RQPCG method, using an incomplete Cholesky decomposition of **A** to define the preconditioner, compared with vector iteration and the bisection method.

We consider the elliptic eigenvalue problem

$$u_{xx} + u_{yy} + \lambda u = 0 \quad \text{in} \quad \Omega = [0,5] \times [0,4] \tag{39}$$

subject to Dirichlet boundary conditions $u=0$ on the two horizontal sides of the rectangular domain, to a Neumann condition $\partial u/\partial n = 0$ on the left vertical side and to the Cauchy condition $\partial u/\partial n + u = 0$ on the right vertical side. The exact eigenvalues are given by $\lambda = \mu_i^2 + \nu_j^2$, (i,j=1,2,3,...), where μ_i are the positive solutions of $\mu\tan(5\mu) = 1$ and $\nu_j = j\pi/4$, and hence the smallest eight eigenvalues are approximately

$$\lambda_1 = 0.685897, \quad \lambda_2 = 1.267637, \quad \lambda_3 = 2.526551, \quad \lambda_4 = 2.536448,$$

$$\lambda_5 = 3.118188, \quad \lambda_6 = 4.377102, \quad \lambda_7 = 4.531512, \quad \lambda_8 = 5.620699.$$

The eigenvalue problem (39) is treated by means of the finite element method using a discretization of Ω into 240 rectangular elements of the size 0.3125 by 0.26667. The quadratic element of the Serendipity class with eight nodes yields a total of 783 nodal points, and the order of the eigenvalue problem is n=717, after taking into account the 66 Dirichlet boundary conditions. The number of nonzero elements of the lower part of the matrices **A** and **B** is $N_A = 6156$, each.

The computations have been done on the computer AS/XL V60 of the University of Zurich in scalar mode. The initial vectors have been chosen by random numbers from (-1,+1). The iteration is stopped when the Rayleigh quotient $R[\mathbf{x}_k]$ attains its stationary value to computer accuracy, and hence the relative accuracy of the computed eigenvectors $\mathbf{z}_\nu$ is of the order of $\varepsilon = 10^{-6}$. Table 1 contains information on the number of iteration steps per eigenpair and the computing time in seconds. The high number of steps, needed for the third eigenpair, is caused by the extremely near next higher eigenvalue λ_4. An improvement of this situation can possibly be achieved by a simultaneous iteration of several vectors [12].

The largest eigenvalue of the matrix pair **(A,B)** is $\lambda_{717} \cong 955.579$, and thus the condition number (16) of $\mathbf{H}(z_1)$ is $\kappa_{B,B^{-1}}(\mathbf{H}(z_1)) \cong 1641$, yielding an estimate (17) of $\mu_e \cong 294$ iteration steps for the unconditioned algorithm. The reduction to 39 steps is indeed remarkable if a preconditioner **M** on the base of the partial Cholesky decomposition is used. A better preconditioner $\mathbf{M} = \mathbf{A}$ with a matrix **C** that is equal to a complete Cholesky decomposition would reduce the condition number to $\kappa_{B,B}(\tilde{H}(y_1)) \cong 2.18$, and hence the corresponding estimate would be $\tilde{\mu}_e \cong 11$. This estimate is verified by a corresponding calculation (see table 1). However, a

preconditioner **M** on the base of a complete Cholesky decomposition increases the total computing time, although the number of iteration steps is reduced, due to the fact that the profile of the matrix **C** is p = 26614, thus increasing the amount of work of the preconditioning step by a factor of about four.

Table 1 Numerical results, test example

i	λ_i	partial Cholesky steps	partial Cholesky CPU	complete Cholesky steps	complete Cholesky CPU
1	0.685899	39	0.314	12	0.198
2	1.267642	28	0.238	11	0.186
3	2.526632	147	1.285	92	1.541
4	2.536553	38	0.352	20	0.345
5	3.118296	29	0.276	15	0.266
6	4.377289	58	0.569	40	0.717
7	4.532231	29	0.298	18	0.329
8	5.621877	46	0.490	28	0.519
total		414	3.868	236	4.199

For further comparison the eight smallest eigenvalues have been computed by the simultaneous inverse vector iteration (working with 12 vectors) and by the bisection method. The total computing times were 7.3 and 11.1 seconds, respectively.

A similar picture is obtained, when the first 30 eigenpairs are computed: Rayleigh quotient minimization with partial Cholesky decomposition as preconditioner requires a total time of 21.6 seconds. 28.8 seconds were needed in case of a complete Cholesky decomposition. The bisection method solved the problem in 32.8 seconds, and the simultaneous vector iteration with 40 vectors required 54.5 seconds. Thus it is seen that the presented algorithm solves the considered eigenvalue problem in quite an efficient way, especially with respect to storage requirements, because the sparsity of the matrices **A** and **B** is used to full extent.

The second example stems from a realistic, practical problem in engineering. It should illustrate the superiority of the RQPCG algorithm over other methods for solving the eigenvalue problem with respect to memory requirements. We consider the problem of computing the lowest eigenfrequencies and acoustic waves in the interior of an existing car. After separation of one coordinate variable we have to

solve an elliptic eigenvalue problem $u_{xx} + u_{yy} + \lambda u = 0$ in a region Ω, that is defined by the longitudinal section of a car, under Neumann boundary conditions. Because the back seats can be turned forward to gain space, it was intended to study the influence of the different regions to the acoustic behaviour. Figures 1 and 2 show the two different regions as well as the triangulations into finite elements, for which quadratic functions are used. The number of nodal variables is $n=933$ and $n = 957$, and the number of nonzero matrix elements is $N_A=N_B=5511$ and $N_A=N_B=5691$, respectively. The smallest eight eigenvalues of the two problems have been computed with a required relative accuracy of the eigenvalues of $\varepsilon=10^{-10}$. Due to the singularity of the matrix **A** a shift of the spectrum has to be applied in order to satisfy the condition of **A** to be positive definite. The computation was done on an IBM Personal System 2, model 55, with a restricted available memory of about 580 kB. The computed approximate eigenvalues and the number of iterations per eigenvalue for the two cases are listed in table 2.

Table 2 Results, acoustic eigenfrequencies

Case 1, n = 933			Case 2, n = 957		
i	λ_i	n_{it}	i	λ_i	n_{it}
1	0	102	1	0	100
2	0.009698	73	2	0.008162	52
3	0.046659	74	3	0.050943	78
4	0.052992	52	4	0.080814	48
5	0.107027	64	5	0.132043	63
6	0.159441	85	6	0.163548	61
7	0.172882	46	7	0.175573	45
8	0.233522	62	8	0.246367	57
Total iterations:		558			504

The total computing time was 1082 and 1004 seconds, respectively. The total number of iterations may look very high, but it should be mentioned that the two eigenvalue problems of medium size could not be treated on this personal computer by one of the classical methods due to their high storage requirements! Thus the RQPCG method solved the problems at least within a reasonable time on a small computer.

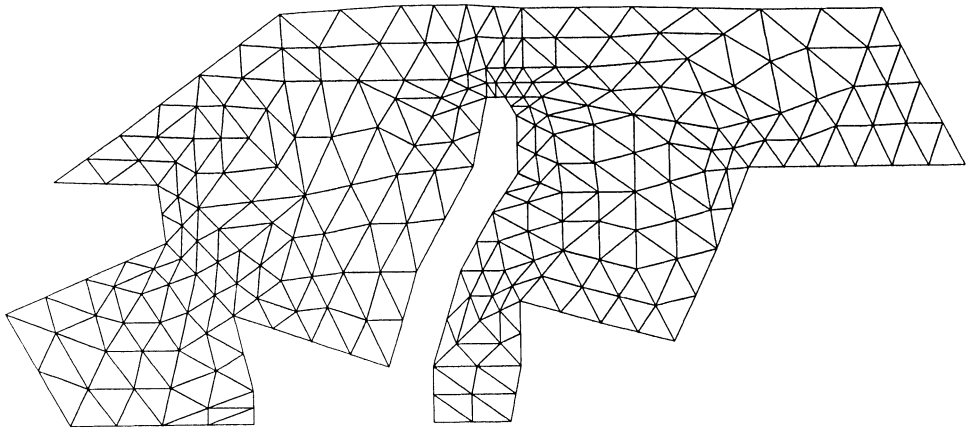


Figure 1 Region Ω and triangulation, case 1

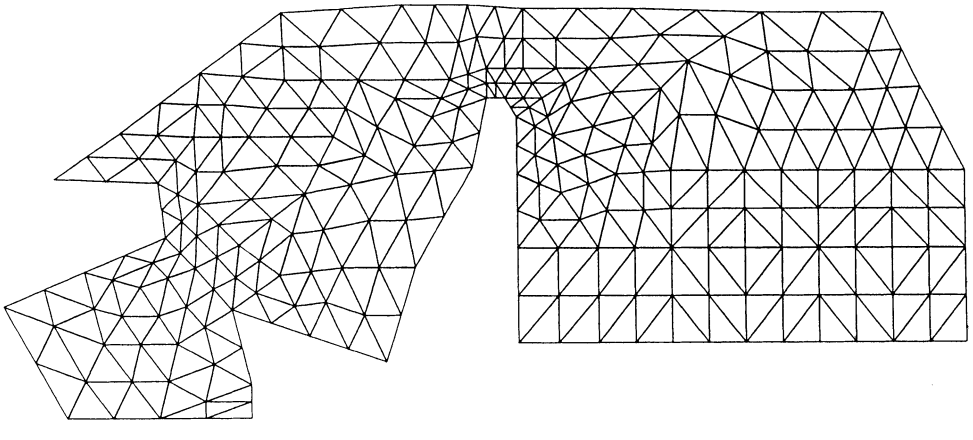


Figure 2 Region Ω and triangulation, case 2

References

[1] Axelsson, O.; Barker, V.A. (1984) Finite element solution of boundary value problems. Theory and computation (Academic Press, New York).

[2] Bartelt, P. (1989) Finite element procedures on vector/tightly coupled parallel-computers. Ph.D. Dissertation, ETH Zuerich.

[3] Bradbury, W.W.; Fletcher, R. (1966) New iterative methods for the solution of the eigenproblem. Numer. Math. **9**, 259-267.

[4] Golub, G.H.; van Loan, Ch.F. (1989) Matrix computations, 2nd edn (The Johns Hopkins University Press, Baltimore and London).

[5] Hestenes, M.R.; W. Karush, W. (1951) Solutions of $Ax = \lambda Bx$. J. Res. Nat. Bur. Standards **49**, 471-478.

[6] Hughes, T.J.R.; Ferencz, R.M.; Hallquist, J.O. (1987) Large-scale vectorized implicit calculations in solid mechanics on a CRAY-XMP/48 utilizing EBE preconditioned conjugate gradients. Comput. Meth. Appl. Mech. Engrg. **61**, 215-248.

[7] Longsine, D.E.; McCormick, S.F. (1980) Simultaneous Rayleigh quotient minimization methods for $Ax = \lambda Bx$. Lin. Alg. Appl. **34**, 195-234.

[8] Manteuffel, T.A. (1980) An incomplete factorization technique for positive definite linear systems. Math. Comput. **34**, 473-480.

[9] Parlett, B.N. (1980) The symmetric eigenvalue problem. (Prentice-Hall, Englewood Cliffs).

[10] Perdon, A.; Gambolati, G. (1986) Extreme eigenvalues of large sparse matrices by Rayleigh quotient and modified conjugate gradients. Comput. Meths. Appl. Mech. Engrg. **56**, 251-264.

[11] Ruhe, A. (1977) Computation of eigenvalues and eigenvectors, in: Barker,V.A., ed., Sparse matrix techniques, Lecture Notes Math. **572** (Springer, Berlin), 130-184.

[12] Sartoretto, F.; Pini, G.; Gambolati, G. (1989) Accelerated simultaneous iterations for large finite element eigenproblems. J. Comput. Physics **81**, 53-69.

[13] Schwarz, H.R. (1987) Rayleigh-Quotient-Minimierung mit Vorkonditionierung, in: Collatz L. et al, ed., ISNM, Vol. 81 (Birkhäuser, Basel) 229-245.

[14] Schwarz, H.R. (1988) Method of finite elements (Academic Press, London). (1991) Methode der finiten Elemente, 3. Aufl. (Teubner, Stuttgart).

Prof. Dr. H. R. Schwarz, Institute of Applied Mathematics, University of Zurich, Rämistrasse 74, CH - 8001 Zürich, Switzerland

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

A NUMERICAL COMPARISON OF TWO APPROACHES TO COMPUTE MEMBRANE EIGENVALUES BY DEFECT MINIMIZATION

G. Still[1], E. Haaren-Retagne[2] and R. Hettich[2]

[1] Universität Twente, toegepaste wiskunde, Enschede, Netherlands

[2] Universität Trier, Fachbereich IV – Mathematik, Trier, Germany

Abstract: In this paper we compare two approaches to the computation of eigenvalues by defect minimization which differ principally in the choice of norm to measure the resp. defect. One of them, basically developed by Kuttler and Sigillito, leads to the consideration of parametric (matrix–) eigenvalue problems, whereas the other requires the treatment of parametric linear programming problems. It will be shown that the latter allows a very flexible choice of trial functions, giving good rates of approximation. Numerical results for rhombical membranes are given.

Key words: Membrane eigenvalues, defect minimization, elliptic equations

AMS (MOS) subject classifications: 65N25, 65N35, 35J05

1. Introduction

For the eigenvalue problem

$$\begin{aligned} \mathcal{L}u + \lambda u &= 0 \text{ in } G \\ u &= 0 \text{ on } \Gamma = \partial G \end{aligned} \tag{1}$$

with $G \subset \mathbb{R}^2$ an open connected bounded region and $\mathcal{L}$ a symmetric elliptic differential

operator, defect minimization methods have been considered by various authors (for instance [2,4,8]). These methods depend on theorems of the following type:

Given an approximation (λ,u) of an eigenpair (λ^*,u^*) of (1). Let $d_G = \|\mathcal{L}u + \lambda u\|_G$, $d_\Gamma = \|u\|_\Gamma$ be the defects in the differential equation and in the boundary condition measured in some norms $\|\ \|_G$, $\|\ \|_\Gamma$. Then, under appropriate conditions on $\mathcal{L}$ and G one can show that in a neighborhood of λ^* a bound

$$\left|\frac{\lambda_k - \lambda}{\lambda_k}\right| \leq \left\{\frac{c_1 d_G^2 + c_2 d_\Gamma^2}{\|u\|_G}\right\}^{\frac{1}{2}} =: \delta(\lambda,u) \tag{2}$$

with constants c_1,c_2 is valid. A similar bound holds for $\|u - u^*\|_G$. Now, defect minimization methods for solving (1) proceed as follows:

Given a $\lambda > 0$. Choose a finite-dimensional linear function space

$$V^n(\lambda) = \{v(p,\lambda,\bullet) \mid p \in \mathbb{R}^n\} \subset C^2(G) \cap C(G \cup \Gamma).$$

Let

$$\varepsilon(\lambda) := \min_{v \in V^n(\lambda)} \delta(\lambda,v) \tag{2'}$$

be the minimal error bound (2) available in $V^n(\lambda)$. Then the local minima $\tilde{\lambda}_\ell$ of $\varepsilon(\lambda)$ are taken as approximations to eigenvalues of (1).

In this paper we will compare two approaches corresponding to different choices of norms $\|\ \|_G$, $\|\ \|_\Gamma$ in (2) and spaces $V^n(\lambda)$:

1. Taking $V^n(\lambda)$ such that $d_G = 0$ and the supremum norm for $\|\ \|_\Gamma$, approximations to eigenvalues are computed by minimizing the error bound $\varepsilon(\lambda)$ derived from Theorem 1 below. This amounts to finding local minima of the value-function of a parametric linear programming problem (cf. Section 3). This approach has also been used in [4,5,6].

2. Choosing alternatively L_2-norms for $\|\ \|_G$ and $\|\ \|_\Gamma$, the minimization of $\varepsilon(\lambda)$ (derived from Theorem 2 below) leads to a parametric eigenvalue problem (Section 4) which can be treated with reasonable effort only for very special spaces $V^n(\lambda)$, such as spaces of polynomials or finite element spaces. We note that the method presented in Sections 4,5 is an improved and modified version of that considered by Kuttler and Sigillito [8,9], which is based on the same error bound, but only considers a local improvement of a given approximation of an eigenvalue by means of a rather inefficient minimization process applied to $\varepsilon(\lambda)$.

2. Error bounds

In this section we state without proof the two basic error bounds in Theorems 1 and 2. These are specializations of much more general results which can be found in [12] for instance.

From now on we consider the problem

$$\begin{aligned} \Delta u + \lambda u &= 0 \text{ in } G \\ u &= 0 \text{ on } \Gamma \end{aligned} \tag{3}$$

with Δ the Laplacian and Γ given by finitely many twice differentiable arcs joined in corners with interior angles ω, $0 < \omega < 2\pi$. In the sequel λ_k, $k \in \mathbb{N}$, always denotes an eigenvalue of (3) and u_k a corresponding eigenfunction. By V we denote the space

$$V = C^2(G) \cap C(G \cup \Gamma) \tag{4}$$

and define for $v \in V$ the following "norms over G and Γ" resp.:

$$\|v\|_G = \frac{1}{A}\left\{\int_G v^2\, dG\right\}^{\frac{1}{2}},\ \|v\|_\Gamma = \frac{1}{A}\left\{\int_\Gamma v^2\, d\Gamma\right\}^{\frac{1}{2}},\ A = \left\{\int_G dG\right\}^{\frac{1}{2}},$$

and

$$\|v\|_{\Gamma,\infty} = \max_{x \in \Gamma} |v(x)|\,. \tag{5}$$

Theorem 1. Given $\lambda > 0$, $u \in V$, $\|u\|_G = 1$, such that

$$\Delta u + \lambda u = 0 \text{ in } G\,.$$

Let

$$\delta = \delta(\lambda,u) = \|u\|_{\Gamma,\infty} \tag{6}$$

and assume $\delta < 1$.

Then there exists an eigenvalue λ_k of (3) such that

$$\frac{\lambda}{1+\delta} \le \lambda_k \le \frac{\lambda}{1-\delta} \tag{7}$$

implying

$$\left|\frac{\lambda_k - \lambda}{\lambda_k}\right| \le \delta\,. \tag{8}$$

Theorem 1 corresponds to the case that u is chosen such that (in the notation of Section 1) $d_G = 0$ and the defect $d_\Gamma = \delta$ on the boundary is measured in the maximum-norm.

Theorem 2. (cf. [8]) Given $\lambda > 0$, $u \in V$, $\Delta u \not\equiv 0$. Then there exists λ_k, an eigenvalue of (3), such that

$$\left|\frac{\lambda - \lambda_k}{\lambda_k}\right| \leq \left[\frac{2\|\Delta u + \lambda u\|_G^2 + 2C^2\lambda^2\|u\|_\Gamma^2}{\|\Delta u\|_G^2}\right]^{\frac{1}{2}} \tag{9}$$

with $C = q_1^{-\frac{1}{2}}$ where q_1 is (an upper bound of) the least eigenvalue of the Stekloff-problem

$$\Delta^2 v = 0 \text{ in } G, \quad v = \Delta v - q\frac{\partial v}{\partial u} = 0 \text{ on } \Gamma.$$

Remarks.

- Kuttler and Sigillito [8,9] use also a slightly different bound with $\|u\|_G^2$ in the denominator. (9) proved to give better bounds at least in case of the rhombical membranes considered in Section 6.
- In addition to bounds (8),(9) on the error of λ, bounds for $\|u - u_k\|_G$, u_k an eigenfunction to λ_k, are available, too. For an extensive discussion cf. [12].
- In more explicit form (9) can be written as

$$\left|\frac{\lambda - \lambda_k}{\lambda_k}\right| \leq \left[\frac{2\int_G (\Delta u + \lambda u)^2 dG + 2C^2\lambda^2 \int_\Gamma u^2 d\Gamma}{\int_G \Delta u^2 dG}\right]^{\frac{1}{2}}. \tag{9'}$$

3. A parametric semi-infinite optimization problem

To apply Theorem 1 we choose an n-dimensional function space $V^n(\lambda) \subset V$,

$$V^n(\lambda) = \{u_n(p,\lambda,x) = \sum_{\nu=1}^{n} p_\nu \varphi_\nu(\lambda,x) \mid p = (p_1,\dots,p_n)^T \in \mathbb{R}^n\}$$

with basis functions $\varphi_\nu(\lambda,\bullet)$, $\nu = 1(1)n$, satisfying $\Delta\varphi_\nu + \lambda\varphi_\nu = 0$ on G for all $\lambda > 0$. To minimize the error bound (8) for a given λ means that the following nonlinear optimization problem $Q(\lambda)$ must be solved.

$Q(\lambda)$ Minimize $\|u_n\|_{\Gamma,\infty}$ subject to $u_n \in V^n(\lambda)$ and $\|u_n\|_G = 1$.

In practice we considered instead of $Q(\lambda)$ the following easier linear problem $P(\lambda)$ which proved to give bounds almost as good as those obtained with $Q(\lambda)$.

$P(\lambda)$ Minimize $\phi(p,e) := e$ subject to the constraints $p_1 \geq 1$ and

$$|u_n(p,\lambda,x)| \leq e \quad \text{for all } x \in \Gamma$$

$$\left(\text{i.e. } \sum_{\nu=1}^{n} p_\nu \varphi_\nu(\lambda,x) \leq e \text{ and } -\sum_{\nu=1}^{n} p_\nu \varphi_\nu(\lambda,x) \leq e,\ x \in \Gamma\right).$$

The constraint $p_1 \geq 1$ is added to exclude the trivial solution $p = 0$. In some cases, other normalizations may be appropriate. Let $(p(\lambda), e(\lambda))$ be a solution of $P(\lambda)$. Then, taking

$$\varepsilon(\lambda) := \frac{e(\lambda)}{\|u_n(p(\lambda),\lambda,\bullet)\|}, \tag{10}$$

and assuming $\varepsilon(\lambda) < 1$, (8) guarantees the existence of an eigenvalue λ_k of (3) such that

$$\left|\frac{\lambda_k - \lambda}{\lambda_k}\right| \leq \varepsilon(\lambda).$$

Therefore, an obvious way to obtain approximations to the required eigenvalues, is by computing local minima of $\varepsilon(\lambda)$. For a detailed theoretical and numerical discussion of this approach we refer to [4,5,6].

Trial functions $\varphi(\lambda;r,\tau)$ satisfying the equation $\Delta\varphi + \lambda\varphi = 0$ (transformed to polar coordinates (r,τ)) are given for example by

$$J_\gamma(\sqrt{\lambda}\, r)\sin\gamma\tau,\ J_\gamma(\sqrt{\lambda}\, r)\cos\gamma\tau,\ \gamma \in \mathbb{R}$$

where J_γ denotes the Bessel function of first kind and order γ. For the sake of efficiency it is very important to choose spaces $V^n(\lambda)$ of trial functions which have good approximation properties w.r.t. eigenfunctions. For problem (3) it is well–known that in corners of G the eigenfunctions have singularities in their derivatives. These singularities depend on the interior angle ω at the corner.

Taking $\alpha = \frac{\pi}{\omega}$, the eigenfunctions can be (asymptotically) expanded in terms of functions

$$J_{\alpha\nu}(\sqrt{\lambda}\, r)\cos\nu\alpha\tau,\ J_{\alpha\nu}(\sqrt{\lambda}\, r)\sin\nu\alpha\tau,\ \nu \in \mathbb{N}, \tag{11}$$

where (r,τ) denote polar coordinates w.r.t the corner in question. (For rational numbers α and curved edges of the corner also logarithmic terms may occur, cf. [1,10,13]). Using Vekua's theory ([15]), results from approximation theory, and these expansions the following has been shown by Eisenstat [10] and Still [13] :

Given an arbitrary number $q > 0$. Then there exist spaces $V^n(\lambda)$, $n \in \mathbb{N}$, (depending on q) such that for the k–th local minimum $\lambda_k^{(n)}$ of the function $\varepsilon_n(\lambda)$ ($= \varepsilon(\lambda)$ in (10)) we have for $n \in \mathbb{N}$ sufficiently large

$$\left| \frac{\lambda_k - \lambda_k^{(n)}}{\lambda_k} \right| \leq \varepsilon_n(\lambda) \leq \frac{c}{n^q} \tag{12}$$

where c depends on λ_k (but not on n).

In Section 6 we will demonstrate the construction of such $V^n(\lambda)$ for the case of rhombical membranes.

4. A parametric eigenvalue problem

Now let us consider the error bound (9) (or (9')) in Theorem 2. Proceeding as in Section 3, we again choose a subspace $V^n(\lambda) \subset V$ with basis functions $\varphi_1(\lambda;\bullet),..,\varphi_n(\lambda;\bullet)$. Minimization of (9') for given λ subject to $u \in V^n(\lambda)$ is obviously equivalent to the following problem

$$\text{Minimize } p^T A(\lambda) p \text{ subject to } p^T B(\lambda) p = 1 \tag{13}$$

with

$$A(\lambda) = 2(B(\lambda) + \lambda A^1(\lambda) + \lambda^2 A^2(\lambda))$$

and $B(\lambda)$, $A^1(\lambda)$, and $A^2(\lambda)$ the n×n–matrices with elements given resp. by

$$b_{ij}(\lambda) = \int_G \Delta\varphi_i \, \Delta\varphi_j dG \tag{14}$$

$$a^1_{ij}(\lambda) = \int_G (\Delta\varphi_i \varphi_j + \varphi_i \Delta\varphi_j) dG \tag{15}$$

$$a^2_{ij}(\lambda) = \int_G \varphi_i \varphi_j dG + C^2 \int_\Gamma \varphi_i \varphi_j \, d\Gamma . \tag{16}$$

Applying the Lagrange–Multiplier–Theorem shows that (13) is solved by every eigenvector $p(\lambda)$, $p^T(\lambda) B(\lambda) p(\lambda) = 1$, corresponding to the least eigenvalue of the generalized eigenvalue problem

$$E(\lambda) \qquad A(\lambda)x = \mu B(\lambda)x \tag{17}$$

and that the minimal value of (13) is given by $\mu(\lambda)$. Thus we have: There exists an eigenvalue λ_k of problem (3) such that

$$\left|\frac{\lambda - \lambda_k}{\lambda_k}\right| \leq \varepsilon(\lambda) = \sqrt{\mu(\lambda)}\,. \tag{18}$$

Again, to compute approximations to $\lambda_1,\lambda_2\ldots$, we determine local minima $\tilde{\lambda}_1,\tilde{\lambda}_2,\ldots$ of $\varepsilon(\lambda)$ or, equivalently, $\mu(\lambda)$. This, of course, requires to evaluate $\mu(\lambda)$ for a large number of values λ. However, if B,A^1,A^2 really depend on λ, then for every λ we have to compute the matrices B,A^1,A^2, i.e. a large number of integrals over G and Γ have to be computed with a rather high accuracy to enable a reasonable minimization algorithm for $\mu(\lambda)$. Even if $\Delta\varphi_i + \lambda\varphi_i = 0$, we still have to evaluate $B(\lambda)$ and the integrals over Γ in (16). To avoid this, Kuttler and Sigillito [9] choose spaces V^n independent of λ. Then B,A^1,A^2 are independent of λ and the integrals (14 – 16) must be evaluated only once and this can be done analytically if simple functions, such as polynomials, are taken, for instance

$$V^n = \text{span}\,\{\varphi_{ij}(x,y) = x^i\,y^j \mid i,j = 0,..,M\} \tag{19}.$$

5. Computation of the least eigenvalue for the parametric eigenvalue problem

Let us now consider the parametric eigenvalue problem

$$E(\lambda): \qquad A(\lambda)x = \mu Bx \quad \text{with} \quad A(\lambda) = 2(B + \lambda\cdot A^1 + \lambda^2\cdot A^2)$$

where B,A^1,A^2 are symmetric $n\times n$–matrices. Thus, $A(\lambda)$ depends analytically on λ. The following Lemma states that the eigenvalues and eigenfunctions of $E(\lambda)$ also depend analytically on λ.

Lemma. Suppose the matrix B is regular. Then there exist functions $\mu_\nu(\lambda),x_\nu(\lambda)$, $\nu = 1,\ldots,n$, analytic for $\lambda \in \mathbb{R}$, such that $x_\nu(\lambda)$ are independent eigenvectors corresponding to the eigenvalues $\mu_\nu(\lambda)$ of $E(\lambda)$.

Proof. With an orthogonal matrix U such that

$$U^T BU = D \quad \text{with} \quad D = \begin{bmatrix} \mu_1 & & 0 \\ & \mu_2 & \\ 0 & & \ddots\ \mu_n \end{bmatrix}$$

let

$$W = \begin{bmatrix} \sqrt{\mu_1} & & 0 \\ & \sqrt{\mu_2} & \\ 0 & & \ddots \sqrt{\mu_n} \end{bmatrix}, \quad V = UW\,U^T .$$

Hence $B = V \cdot V$ and with the transformation $y = Vx$ problem $E(\lambda)$ is equivalent to the parametric common eigenvalue problem

$$C(\lambda)y = \mu y \quad \text{with} \quad C(\lambda) = V^{-1}\,A(\lambda)\,V^{-1} . \tag{20}$$

Since $C(\lambda)$ is symmetric and analytic for all $\lambda \in \mathbb{R}$ the result follows from a corresponding statement for (20) which can be found in Kato [7, p.71].

The derivatives of the functions $\mu_\nu(\lambda)$, $x_\nu(\lambda)$, $\nu = 1,...,n$, can be obtained by considering the system of equations

$$\begin{aligned} (A(\lambda) - \mu(\lambda)B)\,x(\lambda) &= 0 \\ x^T(\lambda)B\,x(\lambda) &= 0 . \end{aligned} \tag{21}$$

Let $x_\nu(\lambda)$, $\mu_\nu(\lambda)$, $\nu = 1,...,n$, denote the eigenpairs of $E(\lambda)$ such that $x_\nu^T(\lambda)B\,x_\rho(\lambda) = \delta_{\nu\rho}$. Then by differentiating (21) it is not hard to see that the derivatives of the eigenpairs are given by

$$\frac{d}{d\lambda}\mu_\nu(\lambda) = -\frac{x_\nu^T(\lambda)(\frac{d}{d\lambda}A(\lambda))x_\nu(\lambda)}{x_\nu^T(\lambda)Bx_\nu(\lambda)} \tag{22}$$

and

$$\frac{d}{d\lambda}x_\nu(\lambda) = \sum_{\rho=1}^{n} c_{\nu\rho}(\lambda)\,x_\rho(\lambda)$$

where

$$c_{\nu\rho}(\lambda) = \frac{x_\rho^T(\lambda)(\frac{d}{d\lambda}A(\lambda))\,x_\nu(\lambda)}{\mu_\nu - \mu_\rho} \quad \text{for} \quad \rho \text{ with } \mu_\rho \neq \mu_\nu$$

and $c_{\nu\rho}(\lambda) = 0$ for ρ with $\mu_\rho = \mu_\nu$. Similarly formulas for higher derivatives of the eigenpairs can be obtained.

Now, we want to determine the local minima of $\mu(\lambda)$, where $\mu(\lambda)$ is the least eigenvalue of the parametric eigenvalue problem $E(\lambda)$. Let $x(\lambda)$ be the normalized eigenvector corresponding to $\mu(\lambda)$.

We compute values of $\mu(\lambda)$ in an interval $[\lambda_a, \lambda_e]$ for $\lambda^{(\nu)} = \lambda_a + \nu \cdot \Delta\lambda$. If

$$\mu(\lambda^{(\nu-1)}) > \mu(\lambda^{(\nu)}) < \mu(\lambda^{(\nu+1)})$$

a local minimum of $\mu(\lambda)$ in $[\lambda^{(\nu-1)}, \lambda^{(\nu+1)}]$ is determined with an algorithm given by Mifflin [11]. This algorithm has the following convergence properties :

Let μ be a (locally) Lipschitz function defined on an interval [a,c] which has a local minimum on (a,c). Then the algorithm converges to a stationary point $\lambda^* \in (a,c)$. Moreover, the convergence is superlinear (in a certain sense (cf. [11])) if either μ is convex on [a,c] or $\mu \in C^2[a,\lambda^*) \cap C^2(\lambda^*,c]$ and $\lim \mu''(\lambda)$ exists for $\lambda \to \lambda_+^*$ and $\lambda \to \lambda_-^*$. The algorithm requires the computation of generalized gradients $\mu'(\bullet)$ in every iteration–point. In view of Theorem 3 and (22) this is no difficulty in our problem.

For the computation of $\mu(\lambda)$ we use inverse iteration with shift $\tilde{\mu}$:

$$\begin{aligned} (A - \tilde{\mu}B)\, x^{(\nu+1)} &= Bx^{(\nu)}) \\ x^{(\nu+1)} &= x^{(\nu+1)}/\|x^{(\nu+1)}\| \end{aligned} \tag{23}$$

$x^{(0)}$ an arbitrary vector (see for example [16]).

To solve $E(\lambda^{(\nu+1)})$, our algorithm makes use of the solutions $\mu(\lambda^{(j)})$, $x(\lambda^{(j)})$, $j < \nu + 1$, of the foregoing problems:

Algorithm

(1) shift with an estimation $\tilde{\mu}$ for $\mu(\lambda^{(\nu+1)})$ (in our implementation we use quadratic extrapolation on $\mu(\lambda^{(j)})$, $j = \nu - 2, \nu - 1, \nu$, to determine $\tilde{\mu}$)

(2) use $x^{(0)} = x(\lambda^{(\nu)})$ as an initial approximation in the scheme (23) for $E(\lambda^{(\nu+1)})$

(3) if the calculation of $\mu(\lambda^{(\nu+1)})$ with shift requires more iterations "than usual", the value of μ for $\lambda^{(\nu+1)}$ will be determined again without shift (i.e. $\tilde{\mu} = 0$ in (23)).

Let μ_i be the eigenvalues of the problem $(A - \tilde{\mu}B)x = \mu Bx$. Assuming that $|\mu_1 - \tilde{\mu}| < |\mu_k - \tilde{\mu}|$ for all $k \neq 1$, the convergence rate for the inverse iteration is linear with a convergence factor

$$\frac{|\mu_1 - \tilde{\mu}|}{\min\limits_{k \neq 1} |\mu_k - \tilde{\mu}|}. \tag{24}$$

Problems may arise when the least eigenvalue of $E(\lambda)$ is close to the next eigenvalue. Usually, this is the case when $\mu(\lambda)$ passes a local maximum. Then, the eigenvector $x(\lambda - \Delta\lambda)$ is almost perpendicular to $x(\lambda)$ and the shift could happen to be closer to the second eigenvalue of $E(\lambda)$ than to the least one. For that reason, in some cases a

higher (i.e. not the least) eigenvalue was computed.

To circumvent these difficulties a vector consisting of the sum of $x(\lambda - \Delta\lambda)$ and a random vector of equal length was used as an initial approximation for the inverse iteration. In addition, for $\lambda^{(k)}$, $\lambda^{(2k)}$,... (for instance $k = 10$), the iteration has been performed without shift and a randomly chosen starting vector.

Although there is no guarantee that we always will find the least eigenvalue, we recommend this procedure for the following reasons:

(1) the bounds for λ_k

$$\left|\frac{\lambda^{(\nu)} - \lambda_k}{\lambda^{(\nu)}}\right| \leq \varepsilon(\lambda^{(\nu)}) = \sqrt{\mu(\lambda^{(\nu)})}$$

obviously also hold for higher eigenvalues

(2) the computation of the least eigenvalue with shift is far less costly than without shift: to compute $\mu(\lambda)$ to a relative accuracy of 10^{-6}, 3 to 4 iterations are needed when shifting is used, in comparison to 15 to 20 iterations without shift.

6. Numerical results for rhombical membranes

With the method outlined in Section 3 (Method 1) and alternatively with the method described in Sections 4 and 5 (Method 2), we have computed approximations of eigenvalues (and eigenfunctions) of (3) for some rhombical regions $G = R(\Theta)$. Here, for given Θ, $0 < \Theta \leq \frac{\pi}{2}$, $R(\Theta)$ denotes the rhombus with edges of length 1 and angle Θ at the vertices v_1, v_3 and $\pi - \Theta$ at v_2, v_4 (cf. Fig. 1).

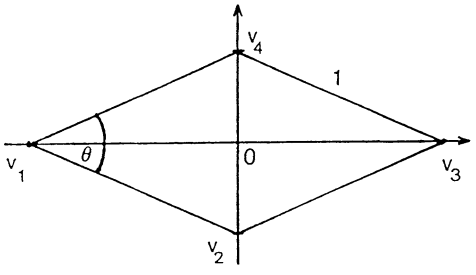


Fig. 1

Due to the symmetry of the region $R(\Theta)$ w.r.t. the x_1- and x_2- axis, for every eigenspace there exists a basis which can be subdivided into four classes of symmetry: one class symmetric w.r.t. both axes (SS), another class symmetric w.r.t. the x_1- axis and antisymmetric w.r.t. the x_2- axis (SA) and accordingly classes (AS) and (AA). Consequently for the sake of efficiency it is favorable to use trial functions with corresponding symmetries. In Method 1 we have taken spaces V^n which contain singular functions of the type (11) at the corners $v_1..v_4$ as suggested in Section 3. Taking the symmetries into consideration appropriate choices are

$$V_{SS}^{m,k,\ell} = \{ \sum_{\nu=0}^{m-1} p_\nu J_{2\nu}(\sqrt{\lambda}\, r) \cos 2\nu\tau$$

$$+ \sum_{\nu=1}^{k} p_{\nu+m} [J_{(2\nu-1)\alpha_1}(\sqrt{\lambda}\, r_1) \sin \alpha_1(2\nu-1)\tau_1 + J_{(2\nu-1)\alpha_1}(\sqrt{\lambda}\, r_3) \sin \alpha_1(2\nu-1)\tau_3]$$

$$+ \sum_{\nu=1}^{\ell} p_{\nu+m+k} [J_{(2\nu-1)\alpha_2}(\sqrt{\lambda}\, r_2) \sin \alpha_2(2\nu-1)\tau_2 + J_{(2\nu-1)\alpha_2}(\sqrt{\lambda}\, r_4) \sin \alpha_2(2\nu-1)\tau_4]\}$$

$$V_{SA}^{m,k,\ell} = \{ \sum_{\nu=1}^{m} p_\nu J_{(2\nu-1)}(\sqrt{\lambda}\, r) \cos (2\nu-1)\tau$$

$$+ \sum_{\nu=1}^{k} p_{\nu+m} [J_{(2\nu-1)\alpha_1}(\sqrt{\lambda}\, r_1) \sin \alpha_1(2\nu-1)\tau_1 + J_{(2\nu-1)\alpha_1}(\sqrt{\lambda}\, r_3) \sin \alpha_1(2\nu-1)\tau_3]$$

$$+ \sum_{\nu=1}^{\ell} p_{\nu+k+m} [J_{2\nu\alpha_2}(\sqrt{\lambda}\, r_2) \sin \alpha_2 2\nu\, \tau_2 + J_{2\nu\alpha_2}(\sqrt{\lambda}\, r_4) \sin \alpha_2 2\nu\, \tau_4]\}$$

and the spaces $V_{AS}^{m,k,\ell}$, $V_{AA}^{m,k,\ell}$ defined analoguously. Here $\alpha_1 = \frac{\pi}{\Theta}$, $\alpha_2 = \frac{\pi}{\pi-\Theta}$, and (r,τ), and (r_i,τ_i) are polar coordinates w.r.t. the origin and the vertices v_i, $i = 1,..,4$, resp. (cf. Fig. 1). Thus, m denotes the number of regular trial functions and k,ℓ the number of singular functions corresponding to the vertices with angle Θ, $\pi-\Theta$. Thus, the whole number of trial functions is $n = m + k + 1$. To demonstrate the approximation properties of these spaces consider the $V_{SS}^{m,k,\ell}$. Obviously the differentiability properties of the errors

$$U - v \ , \ v \in V_{SS}^{m,k,\ell}$$

in approximating the eigenfunction u of (3) are determined by the first neglected singular terms

$$J_{(2k+1)\alpha_1}(\cdot\,\cdot) \sin \alpha_1(\cdot\,\cdot) \tag{25}$$

and

$$J_{(2\ell+1)\alpha_2}(\cdot\,\cdot)\sin\alpha_2(\cdot\,\cdot)\,. \tag{25}$$

Let $\min\{(2k+1)\alpha_1, (2\ell+1)\alpha_2\} = p+\gamma$, $p \in \mathbb{N}$, $0 \le \gamma < 1$. In [13] it is shown that for $\varepsilon > 0$ arbitrarily small we have

$$\varepsilon_n(\tilde{\lambda}^{(n)}) \le \frac{c(\varepsilon)}{m^{p+\gamma-\varepsilon}}$$

($m \to \infty$, $n = m+k+\ell$, k,ℓ fixed) for the approximate eigenvalues $\tilde{\lambda}^{(n)}$ obtained by minimization of $\varepsilon_n(\lambda)$ (cf. (12)).

For $\Theta = 55°$ this gives

(k,ℓ)	(0,0)	(1,0)	(0,1)	(0,2)	(1,1)	(1,2)
$p+\gamma$	1.44	1.44	$3.2\overline{7}$	$\overline{3.27}$	4.32	7.2

The results confirm this behavior. For instance for the first eigenvalue $\lambda_1 \approx 27.265\,004\,705\,9$ we get

(k,ℓ)	(0,0)	(1,1)	(1,2)
m = 4	.31-1	.29-4	.78-5
6	.12-1	.23-6	.21-6
8	.63-2	.14-6	.16-8
10	.36-2	.37-7	.15-9
12	.22-2	.13-7	.53-10
14	.19-2	.48-8	
16	.17-2		

Note, that the cases (1,0), (0,2) shouldn't be chosen. For Method 2 as explained in Section 5 we use spaces V^n of polynomials (cf. (19)).

Exploiting again the symmetries of R appropriate choices for V^n are

$$V^n_{SS} = \text{span}\,\{\varphi_{ij}(x,y) = x^{2i}y^{2j} \mid i,j = 0,..,M\}$$

$$V^n_{SA} = \text{span}\,\{\varphi_{ij}(x,y) = x^{2i+1}y^{j} \mid i,j = 0,..,M\}$$

and V^n_{AS}, V^n_{AA} defined analoguously. Note that $n = (M+1)^2$.

In the following tables λ^* denotes the computed approximations to eigenvalues λ_k of (3), n denotes the number of trial-functions and $\varepsilon_n = \varepsilon_n(\lambda^*)$ the corresponding error bound (8),(18) resp.

To compare the methods, we have determined the values of $\varepsilon(\lambda)$ for the discrete set $\{\lambda^{(\nu)} = 10 + 2\nu \mid \nu = 0,1,..,95\}$ in [10,200]. Then, using Mifflin's algorithm, the local minima of $\varepsilon(\lambda)$ in this interval have been computed.

The approximate eigenvalues obtained by these algorithms and the required CPU-time on a Micro-VAX II are given in the following tables.

n	(m,k,ℓ)	λ^*	$\varepsilon_n(\lambda^*)$
10	(10,0,0)	20.8687	0.988 E-4
		79.044	0.196 E-3
		108.89	0.575 E-3
		168.99	0.497 E-3

Table I. Method 1, type (SS)
$\Theta = 75°$, CPU: 5 min 5 sec

n	(m,k,ℓ)	λ^*	$\varepsilon_n(\lambda^*)$
9	(7,1,1)	----	----
		79.04285	0.183 E-5
		108.8852	0.379 E-5
		168.9839	0.378 E-5

Table II. Method 1, type (SS)
$\Theta = 75°$, CPU: 9 min 35 sec

n	λ^*	$\varepsilon_n(\lambda^*)$
100	20.868	0.194 E-3
100	79.043	0.319 E-3
100	108.89	0.707 E-3
100	168.98	0.258 E-3

Table III. Method 2, type (SS)
$\Theta = 75°$, CPU: 27 min 46 sec

We note, that for the calculation corresponding to (Table II) the chosen discretisation of [10,200] was not fine enough to determine the first eigenvalue. A finer discretisation led to the approximation $\lambda^* = 20.8684839$ with relative error bound $\varepsilon_n(\lambda^*) = 0.189E - 7$.

In the following tables we have listed some approximations for eigenvalues computed with both methods.

n	(m,k,ℓ)	λ^*	$\varepsilon_n(\lambda^*)$
9	(8,0,1)	62.4090	0.144 E-4
10	(8,1,1)	147.66	0.188 E-3
10	(8,0,2)	248.820	0.482 E-4

Table IV. Method 1
type (SS), $\Theta = 30°$

n	λ^*	$\varepsilon_n(\lambda^*)$
100	62.40	0.732 E-2
100	147.7	0.528 E-2
100	248.7	0.225 E-1

Table V. Method 2
type (SS), $\Theta = 30°$

n	(m,k,ℓ)	λ^*	$\varepsilon_n(\lambda^*)$
13	(8,0,5)	199.280511	0.291 E-6
10	(7,0,3)	358.333	0.156 E-3
11	(7,0,4)	521.18	0.902 E-3

Table VI. Method 1
type (SS), $\Theta = 15°$

n	λ^*	$\varepsilon_n(\lambda^*)$
100	199.	0.349 E-1
100	359.	0.631 E-1
100	539.	0.162

Table VII. Method 2
type (SS), $\Theta = 15°$

n	(m,k,ℓ)	λ^*	$\varepsilon_n(\lambda^*)$
10	(8,1,1)	48.20481699	0.225 E-8
10	(8,1,1)	120.47621	0.765 E-6
10	(8,1,1)	178.6069	0.464 E-5

Table VIII. Method 1
type (SA), $\Theta = 75°$

n	λ^*	$\varepsilon_n(\lambda^*)$
100	48.2048	0.360 E-4
100	120.476	0.969 E-4
100	178.607	0.475 E-4

Table IX. Method 2
type (SA), $\Theta = 75°$

n	(m,k,ℓ)	λ^*	$\varepsilon_n(\lambda^*)$
10	(8,0,2)	104.9545	0.210 E-5
10	(8,1,1)	196.293	0.550 E-4
10	(8,1,1)	307.26	0.255 E-3

Table X. Method 1
type (SA), $\Theta = 30°$

n	λ^*	$\varepsilon_n(\lambda^*)$
100	104.95	0.947 E-3
100	196.3	0.334 E-2
100	307.	0.327 E-1

Table XI. Method 2
type (SA), $\Theta = 30°$

n	(m,k,ℓ)	λ^*	$\varepsilon_n(\lambda^*)$
9	(6,0,3)	283.06	0.201 E-3
11	(8,0,3)	438.77	0.132 E-3
13	(8,0,5)	609.091	0.985 E-4

Table XII. Method 1
type (SA), $\Theta = 15°$

n	λ^*	$\varepsilon_n(\lambda^*)$
100	283.	0.232 E-1
100	440.	0.796 E-1
100	640.	0.185

Table XIII. Method 2
type (SA), $\Theta = 15°$

7. Conclusion

The results reported in Section 6 give rise to the following conclusions:

- Method 1 has the advantage, that for special problems spaces $V^n(\lambda)$ of trial func-

tions can be chosen which really depend on λ and, by that, allow to incorporate functions which reproduce analytical properties of the eigenfunctions (singularities, etc.) known from theoretical investigations. In this way the asymptotic rate of the error can be determined according to (12). In comparison with Method 2, this allows to keep the dimension n of $V^n(\lambda)$ small. This results in a high accuracy with comparatively modest computational effort. Note, that the choice of a polynomial space as in Method 2 gives a rate corresponding to $k = \ell = 0$ in (25) resulting in a bad rate of approximation.

– On the other hand, in case of differential operators for which function spaces satisfying $\mathcal{L}u + \lambda u = 0$ are not available, Method 2 seems undoubtedly preferable as it allows to use high–dimensional spaces of trial functions as long as these do not depend on λ. In this sense, Method 2 is more robust and is likely to have a wider range of applicability.

References

[1] Eisenstat, S.C. (1974) On the rate of convergence of the numerical solutions of elliptic boundary value problems, SIAM J. Numer. Anal. 11, 654–680

[2] Fox, L., Henrici, P., Moler, C. (1967) Approximation and bounds for eigenvalues of elliptic operators, SIAM J. Numer. Anal. 4, 89–102

[3] Haaren, E. (1987) Lösung von Eigenwertproblemen durch Defektminimierung, Diplomarbeit, Universität Trier (FRG)

[4] Hettich, R. (1985) On the computation of membrane–eigenvalues by semi–infinite programming, in E.J. Anderson and A.P. Philpott (eds), Infinite programming, Springer Lecture Notes, 79–89

[5] Hettich, R., Haaren, E., Ries, M., Still, G. (1987) Accurate numerical approximations of eigenfrequencies and eigenfunctions of elliptic membranes, ZAMM 67, 589–597

[6] Hettich, R., Still, G. (1987) Local aspects of a method for solving membrane-eigenvalue problems by parametric semi–infinite programming, in J. Guddat et al (eds), Parametric optimization and related topics, Akademie Verlag Berlin, 183–195

[7] Kato, T. (1980) Perturbation for linear operators, Springer Verlag, Berlin, Heidelberg, New York

[8] Kuttler, J.R., Sigillito, V.G. (1978) Bounding eigenvalues of elliptic operators, SIAM J. Math. Anal. 9, 768–773

[9] Kuttler, J.R., Sigillito, V.G. (1985) Estimating eigenvalues with a posteriori/a priori inequalities, Pitman, Boston–London–Melbourne

[10] Lehman, S.R. (1959) Developments at an analytic corner of solutions of elliptic partial differential equations, J. of Math. a. Mech. 8, 727–760

[11] Mifflin, R. (1984) Stationarity and superlinear convergence of an algorithm for univariate locally Lipschitz constrained minimization, Math. Programming 28, 50–71

[12] Still, G. (1988) Computable bounds for eigenvalues and eigenfunctions of elliptic differential operators, Num. Math. 54, 201–223

[13] Still, G. (1989) Defektminimierungsmethoden zur Lösung elliptischer Rand– und Eigenwertaufgaben – spezielle Klassen von Ansatzfunktionen und ihre Approximationseigenschaften, Habilitationsschrift, Universität Trier (FRG)

[14] Still, G. (in preparation) On the approximation of eigenvalues and eigenfunctions of polygonal membranes

[15] Vekua, I.N. (1967) New methods for solving elliptic equations, North Holland, Amsterdam – New York

[16] Wilkinson, J.H. (1965) The algebraic eigenvalue problem, Clarendon Press, Oxford

Georg Still, Universiteit Twente, toegepaste wiskunde, postbus 217
NL – 7500 AE Enschede, Netherlands

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

CONVERGENCE AND ERROR ESTIMATES FOR A FINITE ELEMENT METHOD WITH NUMERICAL QUADRATURE FOR A SECOND ORDER ELLIPTIC EIGENVALUE PROBLEM

M. Vanmaele and R. Van Keer

Faculty of Engineering Sciences, State University of Ghent, Belgium

1. Introduction

This paper deals with a FE-numerical quadrature method for a class of 2nd order elliptic eigenvalue problems on a bounded rectangular domain $\Omega \subset \mathbb{R}^2$, viz.

$$\text{Find } \lambda \in \mathbb{R},\ u \in V : a(u,v) = \lambda.(u,v) \qquad \forall\, v \in V \tag{1.1}$$

where

$$V = \{v \in H^1(\Omega) \mid v = 0 \ \text{ on } \Gamma_1 \subset \partial\Omega = \text{boundary of } \Omega\}$$

(Γ_1 consisting of an integer number of sides)

$(.,.) = L_2(\Omega)$-inner product

$$a(u,v) = \int_\Omega \Big[\sum_{i,j=1}^{2} a_{ij}(x) \cdot \frac{\partial u}{\partial x_i} \cdot \frac{\partial v}{\partial x_j} + a_0(x).u.v \Big].dx, \quad x = (x_1, x_2).$$

Here the coefficients a_{ij} and $a_0 \in L_\infty(\Omega)$ obey the usual ellipticity conditions [4], with moreover $a_{12} = a_{21}$ a.e. in Ω. Then $a : V \times V \to \mathbb{R}$ is a bounded, symmetric bilinear form, which is moreover strongly coercive if $V \neq H^1(\Omega)$ or if the positive constant bounding below $a_0(x)$ in Ω is strictly positive.

To define a FE-approximation of (1.1), let $(\tau_h)_{h \to 0}$ be a regular family of partitions of Ω in subrectangles K, with mesh parameter $h = \max_K h_K$ (h_K diameter of K). Let, for a given integer $k \geq 1$,

$$V_h = \{v \in C^0(\bar{\Omega}) \mid v_{|K} \in Q_k(K),\ \forall\, K \in \tau_h\ ;\ v = 0 \ \text{ on } \ \Gamma_1\} \tag{1.2}$$

where $Q_k(K)$ is the set of polynomials of degree k in each variable on K.
The corresponding consistent mass FE-approximation then reads

$$\text{Find } \lambda_h \in \mathbb{R},\ u_h \in V_h : a(u_h,v) = \lambda_h.(u_h,v) \qquad \forall\, v \in V_h \tag{1.3}$$

Let λ_ℓ, u_ℓ be the ℓ-th exact eigenpair of (1.1), the eigenvalues being numbered in increasing order of magnitude, counted with their multiplicity, and the eigenfunctions being (ortho)normalized in $L_2(\Omega)$. The approximation properties of the corresponding eigenpair λ_ℓ^h, u_ℓ^h of (1.3) are well understood. See e.g. [2] for a recent and very general treatise on eigenvalue problems.

In a recent paper [1], a lumped mass FE-approximation of (1.1) is considered, differing from (1.3) in that at the RHS a discrete inner product $(.,.)_h$ on V_h is used, corresponding to a suitable Lobatto quadrature formula for the involved integrals on each element K. The main result for the approximate eigenpairs $\tilde{\lambda}_\ell^h$, $\tilde{u}_\ell^h$ is

Theorem 1.1 Let $(\tau_h)_{h\to 0}$ be a regular family of 'triangulations' of Ω, satisfying an inverse assumption, i.e.

$$\exists\ \nu > 0 : \frac{h}{h_K} \le \nu\ , \quad \forall\ K \in \tau_h, \qquad \forall h \tag{1.4}$$

Let λ_ℓ be a single exact eigenvalue, $\ell \ge 1$. Then

(i) $\tilde{\lambda}_\ell^h \to \lambda_\ell$ if $h \to 0$, viz. $|\tilde{\lambda}_\ell^h - \lambda_\ell| < C(\lambda_\ell).h^2$ for h sufficiently small

(ii) When moreover the exact eigenfunctions $u_1, u_2, \ldots, u_\ell$ are in $H^{k+1}(\Omega)$ and $|u_\ell| = 1$, $|\tilde{u}_\ell^h|_h = 1$, then for $k \ge 2$

$$\|\tilde{u}_\ell^h - u_\ell\| \le C(\lambda_\ell).h^{k-1}, \qquad |\tilde{\lambda}_\ell^h - \lambda_\ell| \le C(\lambda_\ell).h^{2k-1} \tag{1.5}$$

If, in addition, the bilinear form a is 'regular', then for $k \ge 2$

$$|\tilde{\lambda}_\ell^h - \lambda_\ell| \le C(\lambda_\ell).h^{2k} \tag{1.6}$$

In this theorem $|.|$ and $|.|_h$ stand for the $L_2(\Omega)$-norm and its discrete counterpart, while $\|.\|$ denotes the $H^1(\Omega)$-norm. Contrary to the pioneering work [7] no use is made of perturbation theory of compact, self adjoint operators. Leaning on this theory we may obtain $\|\tilde{u}_\ell^h - u_\ell\| = O(h)$ for $k = 1$, along similar lines as in [8]. We recall that (1.6) is optimal, as the rate of convergence is exactly the one of the consistent mass case, while, compared to that case, there is one unit less in the rate of convergence in (1.5).

In this paper we extend the results of [1] in two respects. First we take into account an approximation $a_h(.,.)$ of the bilinear form $a(.,.)$ on $V_h \times V_h$, resulting from a deliberately chosen Gauss-Legendre quadrature formula for the involved integrals on the elements K. Thus we consider

$$\text{Find } \hat{\lambda}_h \in \mathbb{R},\ \hat{u}_h \in V_h : a_h(\hat{u}_h, v) = \hat{\lambda}_h . (\hat{u}_h, v)_h \quad \forall\, v \in V_h \tag{1.7}$$

Secondly, we also allow for the case of multiple exact eigenvalues.

An outline of the paper is now in order. In Section 2 the quadrature formulas used are stated and investigated. Section 3 deals with the convergence of the approximate eigenvalue $\hat{\lambda}_\ell^h \to \lambda_\ell$ if λ_ℓ is a simple exact eigenvalue. In Section 4 we obtain error estimates for both the eigenvalues and eigenfunctions in the simple exact eigenvalue case, while the case of a multiple exact eigenvalue is taken up in Section 5. A numerical example is given in Section 6.

2. Some results on the numerical quadrature formulas

2.1. Preliminaries (cfr. [4] and [6])

Consider the affine invertible mapping

$$F_K : \hat{K} = [-1,1]^2 \text{ (master square)} \to K,\ \hat{x} \to x = B_K . \hat{x} + b_K$$

where without loss of generality the Jacobian of F_K is assumed to be positive. Introduce the Lobatto quadrature formula on $\hat{K}$

$$I_{\hat{K}}(\hat{\phi}) = \sum_{r=1}^{(k+1)^2} \hat{\omega}_r . \hat{\phi}(\hat{b}_r) \simeq \int_{\hat{K}} \hat{\phi}(\hat{x}) . d\hat{x} \quad \forall\, \hat{\phi} \in C^0(\bar{\hat{K}}) \tag{2.1}$$

where $\hat{b}_r$ and $\hat{\omega}_r$ are the Lobatto quadrature nodes and weights respectively. Recall that $\{\hat{b}_r\} = \{\xi_1, \xi_2, \ldots, \xi_{k+1}\}^2$ where $\xi_1 = -1$, $\xi_{k+1} = 1$ and $\xi_2, \ldots, \xi_k$ are the zero's of $P_k'(\xi)$ (P_k Legendre polynomial of degree k). Putting $\phi(x) = \hat{\phi}(\hat{x})$, whenever $x = F_K(\hat{x})$, $\hat{x} \in \hat{K}$, we define the Lobatto quadrature formula for $\int_K \phi(x)dx$ by

$$I_K(\phi) = (\det B_K) . I_{\hat{K}}(\hat{\phi}) \simeq \int_K \phi(x) . dx \tag{2.2}$$

We will use the well known property

$$E_K(\phi) \equiv \int_K \phi(x) . dx - I_K(\phi) = 0 \quad \forall\, \phi \in Q_{2k-1}(K) \tag{2.3}$$

Henceforth the nodes of the finite element mesh, corresponding to V_h, are chosen to be just the Lobatto quadrature nodes $b_{r,K} = F_K(\hat{b}_r)$, $1 \le r \le (k+1)^2$, in each element K.

Next we define the approximate inner product and associated norm in V_h by

$$(v,w)_h = \sum_{K \in \tau_h} I_K(v.w) \quad , \quad |v|_h = (v,v)_h^{1/2} \qquad \forall\, v,w \in V_h \tag{2.4}$$

Then the corresponding mass matrix $\tilde{M}$ is readily seen to be diagonal.

To obtain a suitable approximation $a_h(.,.)$ of $a(.,.)$, we will need a quadrature formula on each element K which is exact for polynomials of $Q_{2k+1}(K)$. We use the Gauss-Legendre formula, defining $I^G_{\hat{K}}(\hat{\phi})$ and $I^G_K(\phi)$ by similar relations as (2.1) and (2.2), where now the Gauss-Legendre quadrature nodes $\hat{b}^G_r$ and weights $\hat{\omega}^G_r$ are introduced. Then, indeed

$$E^G_K(\phi) \equiv \int_K \phi(x).dx - I^G_K(\phi) = 0 \quad \forall\, \phi \in Q_{2k+1}(K) \tag{2.5}$$

We define the discrete analogon of $a(.,.)$ on $V_h \times V_h$ by

$$a_h(v,w) = \sum_{K\in\tau_h} a^G_K(v,w), \quad a^G_K(v,w) = \sum_{i,j=1}^{2} I^G_K(a_{ij}.\partial_i v.\partial_j w) + I^G_K(a_0.v.w) \qquad \forall\, v,w \in V_h \tag{2.6}$$

assuming a_{ij}, $1 \le i,j \le 2$ and $a_0 \in C^0(\bar{\Omega})$, (denoting $\partial_i v = \frac{\partial v}{\partial x_i}$ etc.).

We use the standard notation for the Sobolev spaces $W^{m,p}(R)$ $(= H^m(R)$ if $p = 2)$ and their corresponding (semi)norms, [4].

2.2. Auxiliary results

Proposition 2.1. For $|.|_h$, (2.4), and $a_h(.,.)$,(2.6), we have : there exist constants C_1, C_2 and C_3, independent of h, such that :

(i) $$C_1.|v| \le |v|_h \le C_2.|v| \qquad \forall\, v \in V_h \tag{2.7}$$

(ii) $$C_3.\|v\|^2 \le a_h(v,v) \qquad \forall\, v \in V_h \tag{2.8}$$

The proof of (i) may be found in [1], while the proof of (ii) may proceed along similar lines as in [4], p.187. Note that (2.8) implies the positive definiteness of the stiffness matrix $\hat{S}$, associated to $a_h(.,.)$ and appearing in the algebraic version of (1.7). As $\hat{S}$ is symmetric by the symmetry of $a_h(.,.)$, the eigenvalues of (1.7) indeed obey $\hat{\lambda}^h_\ell > 0$, $1 \le \ell \le I$, $I = \dim V_h$.

Lemma 2.1 ([4],p.141).
Let (1.4) hold. Then there exists a constant C, independent of h, such that

$$|w|_{k,K} \le C.h_K^{-i}.|w|_{k-i,K}\,, \quad 0 \le i \le k, \quad \forall\, w \in Q_k(K),\ \forall\, K \in \tau_h \tag{2.9}$$

Finally, the estimation of quadrature errors will heavily rest upon the version of the Bramble Hilbert lemma, [3], for elements of the dual space of $W^{m+1,r}(R)$ which vanish on $Q_m(R)$, $m \in \mathbb{N}_0$.

2.3. First estimates

From now on we assume that the (regular) family $(\tau_h)_{h\to 0}$ obeys (1.4).

Theorem 2.1. Let $a \in W^{k+2,\infty}(\Omega)$. Then there exists a constant C, independent of $K \in \tau_h$, h and a, such that

$$|E_K^G(a.p.q)| \le C.h_K^{k+2}.\|a\|_{k+2,\infty,K}.\|p\|_{k,K}.|q|_{0,K} \quad \forall\ p,q \in Q_k(K) \tag{2.10}$$

When $a \in W^{2k+2,\infty}(\Omega)$ this can be improved to

$$|E_K^G(a.p.q)| \le C.h_K^{2k+2}.\|a\|_{2k+2,\infty,K}.\|p\|_{k,K}.\|q\|_{k,K} \quad \forall\ p,q \in Q_k(K) \tag{2.11}$$

Proof. For (2.10) apply the Bramble Hilbert lemma on the mapping $\hat{\phi} \in H^{k+2}(\hat{K}) \mapsto E_{\hat{K}}^G(\hat{\phi}.\hat{q})$, where $\hat{q} \in Q_k(\hat{K})$ is given (recall (2.5)). Take $\hat{\phi} = \hat{a}.\hat{p}$, with $\hat{p} \in Q_k(\hat{K})$. Relate the involved seminorms on Sobolev spaces on $\hat{K}$ to seminorms on the corresponding Sobolev spaces on K by means of standard inequalities, [4], p.194. Finally, use Hölder's inequality in $\mathbb{R}^2$. For (2.11) consider the mapping $\hat{\phi} \in W^{2k+2,1}(\hat{K}) \mapsto E_{\hat{K}}^G(\hat{\phi})$ and proceed similarly with $\hat{\phi} = \hat{a}.\hat{p}.\hat{q}$.■

Theorem 2.2. Let a_{ij}, $a_0 \in W^{k+2,\infty}(\Omega)$, i and j = 1,2. Then there exists a constant C independent of h such that

$$|\varepsilon_a(u,v)| \equiv |a(u,v) - a_h(u,v)| \le C.h^2.\|u\|.\|v\| \quad \forall\ u,v \in V_h \tag{2.12}$$

Proof. From (1.1) and (2.5)-(2.6) one has

$$\varepsilon_a(u,v) = \sum_{K\in\tau_h}\left[\sum_{i,j=1}^{2} E_K^G(a_{ij}.\partial_i u.\partial_j v) + E_K^G(a_0.u.v)\right] \tag{2.13}$$

Then invoke (2.10), combined with (2.9).■

3. Convergence of the eigenvalues (case of simple exact eigenvalue)

Theorem 3.1. Let λ_ℓ be a simple eigenvalue of (1.1). The corresponding approximate eigenvalue $\hat{\lambda}_\ell^h$ of (1.7) obeys

$$\hat{\lambda}_\ell^h \to \lambda_\ell \quad \text{if} \quad h \to 0 \tag{3.1}$$

Moreover, if the exact eigenfunctions u_i, $1 \le i \le \ell$, are in $H^{k+1}(\Omega)$, then

$$|\hat{\lambda}_\ell^h - \lambda_\ell| \le C(\ell).h^2 \qquad (C \text{ independent of } h) \tag{3.2}$$

Proof. Leave from the standard Rayleigh-Ritz characterisation of $\hat{\lambda}_\ell^h$ and $\tilde{\lambda}_\ell^h$

$$\hat{\lambda}_\ell^h = \min_{E_\ell \subset V_h} \max_{v\in E_\ell} \frac{a_h(v,v)}{(v,v)_h}, \qquad \tilde{\lambda}_\ell^h = \min_{E_\ell \subset V_h} \max_{v\in E_\ell} \frac{a(v,v)}{(v,v)_h}$$

Here E_ℓ is any ℓ-dimensional subspace of V_h. (2.12) implies $|\hat{\lambda}_\ell^h - \tilde{\lambda}_\ell^h| \leq C.h^2$. Then invoke part (i) of Theorem 1.1 together with the triangle inequality.■

The convergence (3.1) easily leads to

Corollary 3.1. We have, for sufficiently small h,

$$\rho_{\ell,h} \equiv \max_{1\leq i\leq I, i\neq \ell} \frac{\lambda_\ell}{|\hat{\lambda}_i^h - \lambda_\ell|} \leq \gamma_\ell \qquad \text{(independent of h)} \tag{3.3}$$

4. Error estimates for eigenvalues and eigenfunctions (simple exact eigenvalue)

Throughout this section λ_ℓ is a simple eigenvalue of (1.1) with corresponding eigenfunction u_ℓ, $|u_\ell| = 1$. $\hat{\lambda}_i^h$, $\hat{u}_i^h$ is the i-th eigenpair of (1.7), $|\hat{u}_i^h|_h = 1$. Throughout h is sufficiently small. We put

$$E(v,w) = (v,w) - (v,w)_h \qquad \forall\, v,w \in V_h \tag{4.1}$$

As in [1] we introduce the elliptic projection operator $P : V \to V_h$, defined by

$$a(v - Pv, w) = 0 \qquad \forall\, v \in V, \quad \forall\, w \in V_h \tag{4.2}$$

with properties

$$\|v - Pv\| \leq C.h^k.\|v\|_{k+1,\Omega}\,, \qquad \forall\, v \in H^{k+1}(\Omega)\cap V \tag{4.3}$$

$$\left(\sum_{K\in\tau_h} \|Pv\|_{m,K}^2\right)^{1/2} \leq C.\|v\|_{k+1,\Omega}\,, \quad m = 1,\ldots,k+1 \quad \forall\, v \in H^{k+1}(\Omega)\cap V \tag{4.4}$$

Lemma 4.1. There is a constant $C(\ell)$, independent of h, such that

$$\sum_{i=1, i\neq\ell}^{I} |(Pu_\ell, \hat{u}_i^h)_h|^2 \leq C(\ell).\sum_{i=1, i\neq\ell}^{I} \Big[|(u_\ell - Pu_\ell, \hat{u}_i^h)|^2 + |E(Pu_\ell, \hat{u}_i^h)|^2 + |\varepsilon_a(Pu_\ell, \hat{u}_i^h)|^2\Big] \tag{4.5}$$

Proof. We proceed similarly as in [1]. The definitions of E, ε_a and P lead to

$$(\hat{\lambda}_i^h - \lambda_\ell).(Pu_\ell, \hat{u}_i^h)_h = \lambda_\ell.[(u_\ell - Pu_\ell, \hat{u}_i^h) + E(Pu_\ell, \hat{u}_i^h) - \varepsilon_a(Pu_\ell, \hat{u}_i^h)]$$

It is then sufficient to use (3.3).■

Theorem 4.1. Let a_{ij} and $a_0 \in W^{k+2,\infty}(\Omega)$, i and j = 1,2. Assume that the first ℓ eigenfunctions $u_1,\ldots,u_\ell \in H^{k+1}(\Omega)$. Then one has, for $k \geq 2$,

$$|u_\ell - \hat{u}_\ell^h| \leq C(\ell).h^{k-1} \qquad \text{(C(\ell) independent of h)} \tag{4.6}$$

Proof. We again proceed similarly as in [1], the difference coming from the term $\varepsilon_a(Pu_\ell, \hat{u}_i^h)$ in (4.5). Using (2.13) and the inequalities (2.9)-(2.10) we

may arrive at

$$|\varepsilon_a(Pu_\ell, \hat{u}_i^h)| \le C(\ell) \sum_{K\in\tau_h} h_K^{k+1} . \|Pu_\ell\|_{k+1,K} . |\hat{u}_i^h|_{0,K} \qquad (C(\ell) \text{ independent of } h)$$

From Hölder's inequality and (4.4),(2.7) we obtain $|\varepsilon_a(Pu_\ell, \hat{u}_i^h)| < C(\ell).h^{k+1}$.■

(4.6) allows us to improve the estimate (3.2) for $k \ge 2$, under stronger conditions on a_0 and a_{ij} (to may apply (2.11)). Introduce the piecewise Lagrange interpolator π_h: $C^0(\bar{\Omega}) \to X_h$ by

$$v \to \pi_h v \in X_h \ , \quad \pi_h v \text{ coincides with } v \text{ in all nodes of the mesh.}$$

Here X_h is defined as V_h without the homogeneous Dirichlet B.C. on Γ_1.
We have, with C a constant independent of h, cfr. [4], p.246,

$$\|v - \pi_h v\|_{m,\Omega} \le C.h^{k+1-m} . |v|_{k+1,\Omega} \qquad m = 0,1 \qquad \forall\, v \in H^{k+1}(\Omega) \tag{4.7}$$

$$(\sum_{K\in\tau_h} \|v - \pi_h v\|_{m,K}^2)^{1/2} \le C.h^{k+1-m} . |v|_{k+1,\Omega} \qquad 2 \le m \le k+1 \tag{4.8}$$

Theorem 4.2. Let a_{ij} and $a_0 \in W^{2k+2,\infty}(\Omega)$, i and j = 1,2. Assume again $u_1, \ldots, u_\ell \in H^{k+1}(\Omega)$. Then, for $k \ge 2$,

$$|\lambda_\ell - \hat{\lambda}_\ell^h| \le C(\ell).h^{2k-1} \ , \qquad (C(\ell) \text{ independent of } h). \tag{4.9}$$

Proof. Proceeding similarly as in [1], we obtain

$$|\lambda_\ell - \hat{\lambda}_\ell^h| \le C(\lambda_\ell).[|(u_\ell - Pu_\ell, \hat{u}_\ell^h)| + |E(Pu_\ell, \hat{u}_\ell^h)| + |\varepsilon_a(Pu_\ell, \hat{u}_\ell^h)|] \tag{4.10}$$

where the first 2 terms may be found to be of $O(h^{2k-1})$. The additional term is

$$\begin{aligned}\varepsilon_a(Pu_\ell, \hat{u}_\ell^h) = {} & \varepsilon_a(Pu_\ell - \pi_h u_\ell, \hat{u}_\ell^h - \pi_h u_\ell) + \varepsilon_a(Pu_\ell - \pi_h u_\ell, \pi_h u_\ell) \\ & + \varepsilon_a(\pi_h u_\ell, \hat{u}_\ell^h - \pi_h u_\ell) + \varepsilon_a(\pi_h u_\ell, \pi_h u_\ell)\end{aligned} \tag{4.11}$$

As in the proof of the previous theorem we get, using the triangle inequality,

$$\begin{aligned}|\varepsilon_a(\pi_h u_\ell, \hat{u}_\ell^h - \pi_h u_\ell)| \le C(\lambda_\ell).h^{k+1} . \sum_{K\in\tau_h} & [\|u_\ell - \pi_h u_\ell\|_{k+1,K} + \|u_\ell\|_{k+1,K}] \\ & . [|\hat{u}_\ell^h - u_\ell|_{0,K} + |u_\ell - \pi_h u_\ell|_{0,K}]\end{aligned}$$

Invoking Hölder's inequality, as well as (4.7)-(4.8) and (4.6), the RHS is found to be bounded above by $C(\lambda_\ell).h^{2k}$. The other terms in (4.11) are estimated in a similar way, resulting in $|\varepsilon_a(Pu_\ell, \hat{u}_\ell^h)| \le C(\lambda_\ell).h^{2k}$.■

(4.9) lead to an estimate for the eigenfunctions in the $H^1(\Omega)$-norm.

Theorem 4.3. Under the assumptions of the previous theorem we have, for $k \ge 2$,

$$\|u_\ell - \hat{u}_\ell^h\| \leq C(\ell).h^{k-1} \quad (C(\ell) \text{ a constant independent of } h) \tag{4.12}$$

Proof. Using the definitions of ε_a, E, and P, we have

$$\beta.\|u_\ell - \hat{u}_\ell^h\|^2 \leq [\lambda_\ell.(u_\ell, u_\ell - \hat{u}_\ell^h) - \hat{\lambda}_\ell^h.(\hat{u}_\ell, Pu_\ell - \hat{u}_\ell^h) + \hat{\lambda}_\ell^h.E(\hat{u}_\ell^h, Pu_\ell - \hat{u}_\ell^h)]$$

$$- \varepsilon_a(\hat{u}_\ell^h, Pu_\ell - \hat{u}_\ell^h) \quad (\beta \text{ constant of strong coercivity of a})$$

Using (4.3), (4.6) and (4.7)-(4.8) the expression into brackets may readily be estimated by $C(\lambda_\ell).h^{2(k-1)}$, proceeding as in [1]. Moreover

$$\varepsilon_a(\hat{u}_\ell, Pu_\ell - \hat{u}_\ell^h) = \varepsilon_a(\hat{u}_\ell - \pi_h u_\ell, Pu_\ell - \hat{u}_\ell^h) + \varepsilon_a(\pi_h u_\ell, Pu_\ell - \hat{u}_\ell^h)$$

Estimating both terms as above we get $|\varepsilon_a(\hat{u}_\ell, Pu_\ell - \hat{u}_\ell^h)| \leq C(\lambda_\ell).h^{2(k-1)}$.■

Finally the estimate (4.12) leads to an improvement of the estimate (4.9) when the bilinear form a(.,.) on V x V is regular in the sense of [4], p.138. This will be the case if $\Gamma_1 = \phi$ or if $\Gamma_1 = \partial\Omega$. Then, by the Aubin-Nitsche argument, (4.3) implies

$$|v - Pv| \leq C.h^{k+1}.\|v\|_{k+1,\Omega} \qquad \forall\, v \in H^{k+1}(\Omega) \cap V \tag{4.13}$$

Remark 4.1. By the method of [8] we may obtain $\|u_\ell - \hat{u}_\ell^h\| = O(h)$ for $k = 1$.

Theorem 4.4. Let the bilinear form a(.,.) on V x V be 'regular'. Then, retaining the conditions of the previous theorem, we have for $k \geq 2$,

$$|\lambda_\ell - \hat{\lambda}_\ell^h| \leq C(\ell).h^{2k} \qquad (C(\ell) \text{ independent of } h) \tag{4.14}$$

Proof. Return to (4.10). Following [1], we may improve the estimate of $|E(Pu_\ell, \hat{u}_\ell^h)|$ by a factor h, when (4.12) is used instead of (4.6). Also the estimate of $|(u_\ell - Pu_\ell, \hat{u}_\ell^h)|$ may be improved by a factor h, using (4.13) instead of (4.3) for $|u_\ell - Pu_\ell|$.■

5. Case of a multiple exact eigenvalue

Let λ_ℓ be an (L+1)-fold eigenvalue of (1.1), i.e. $\lambda_{\ell-1} < \lambda_\ell = \lambda_{\ell+1} = \ldots = \lambda_{\ell+L} < \lambda_{\ell+L+1}$. By u_ℓ, $u_{\ell+1}, \ldots, u_{\ell+L}$ we denote a set of (L+1) associated eigenfunctions, orthonormal in $L_2(\Omega)$. Let $\hat{\lambda}_{\ell+r}^h$, $\hat{u}_{\ell+r}^h$, $r = 0, \ldots, L$, be the corresponding approximate eigenpairs of (1.7), the eigenfunctions being orthonormal with respect to $(.,.)_h$.

To obtain estimates for $(\hat{\lambda}_{\ell+r}^h - \lambda_\ell)$ and $(\hat{u}_{\ell+r}^h - u_\ell)$, we retain a similar chain of steps as in Sections 3-4.

(1) $\hat{\lambda}^h_{\ell+r} \to \lambda_\ell, \qquad r = 0,1,\ldots,L \qquad \text{if} \quad h \to 0 \qquad (5.1)$

(2) Leaning on this convergence we prove that, for a suitable set of exact eigenfunctions $U^*_{\ell+r}$, $0 \le r \le L$, for $k \ge 2$,

$$|U^*_{\ell+r} - \hat{u}^h_{\ell+r}| \le C(\ell).h^{k-1} \qquad r = 0,1,\ldots,L \qquad (5.2)$$

(3) This estimate will lead us to, for $k \ge 2$,

$$|\hat{\lambda}^h_{\ell+r} - \lambda_\ell| \le C(\ell).h^{2k-1} \qquad r = 0,1,\ldots,L \qquad (5.3)$$

(4) This estimate for the eigenvalues in turn will imply, for $k \ge 2$,

$$\|U^*_{\ell+r} - \hat{u}^h_{\ell+r}\| \le C(\ell).h^{k-1} \qquad r = 0,1,\ldots,L \qquad (5.4)$$

(5) Finally, if the bilinear form a is regular, from the previous estimate we may improve (5.3), for $k \ge 2$,

$$|\hat{\lambda}^h_{\ell+r} - \lambda_\ell| \le C(\ell).h^{2k} \qquad r = 0,1,\ldots,L \qquad (5.5)$$

Again $C(\ell)$ is a generic constant independent of h and h is sufficiently small.

The precise formulation and the proof of the theorems dealing with (5.3)-(5.5) require only simple adaptations of Theorems 4.2-4.4 and their proofs, and will be skipped here. We begin with step (1) and its corollaries.

Proposition 5.1. The convergence (5.1) holds. Moreover, if the exact eigenfunctions u_i, $1 \le i \le \ell+L$, are in $H^{k+1}(\Omega)$, then

$$|\hat{\lambda}^h_{\ell+r} - \lambda_{\ell+r}| \le C(\ell).h^2 \qquad r = 0,1,\ldots,L.$$

Proof. It is sufficient to leave from the Rayleigh-Ritz characterisations of $\tilde{\lambda}^h_{\ell+r}$ and $\lambda_{\ell+r}(=\lambda_\ell)$, $0 \le r \le L$ to first extend part (i) of Theorem 1.1, and next of $\hat{\lambda}^h_{\ell+r}$ and $\tilde{\lambda}^h_{\ell+r}$, $0 \le r \le L$, proceeding then similarly as in Theorem 3.1■ This result implies a direct analogon of (3.3) and hence also of (4.5), differing from these inequalities in that now the index $i \ne \ell, \ell+1,\ldots,\ell+L$.

Theorem 5.1. Let a_{ij} and $a_0 \in W^{k+2}(\Omega)$, i and j = 1,2. Assume that $u_i \in H^{k+1}(\Omega)$, $1 \le i \le \ell+L$. Then for $k \ge 2$ there exists a set $U^*_{\ell+r}$, $0 \le r \le L$, of eigenfunctions of (1.1) corresponding to λ_ℓ, orthonormal in $L_2(\Omega)$, such that

$$|U^*_{\ell+r} - \hat{u}^h_{\ell+r}| \le C(\ell).h^{k-1}, \qquad 0 \le r \le L \qquad (5.6)$$

Proof. The proof contains 3 major steps and proceeds similarly as in [5],p.908, but is more involved.

(i) We start with

$$|u_{\ell+s} - \textstyle\sum_{r=0}^{L} (Pu_{\ell+s}, \hat{u}^h_{\ell+r})_h . \hat{u}^h_{\ell+r}| \le C(\ell).h^{k-1}, \qquad 0 \le s \le L \qquad (5.7)$$

This follows from

$$|Pu_{\ell+s} - \sum_{r=0}^{L} (Pu_{\ell+s}, \hat{u}^h_{\ell+r})_h . \hat{u}^h_{\ell+r}|^2_h = \sum_{i=1;\, i\neq \ell,\ldots,\ell+L}^{I} |(Pu_{\ell+s}, \hat{u}^h_i)_h|^2$$

by invoking the analogon of (4.5), proceeding similarly as in the proof of Theorem 4.1 and using (4.3) and (2.7).

(ii) By a reasoning ex absurdo, the matrix $\beta = (\beta_{sr})$, $\beta_{sr} = (Pu_{\ell+s}, \hat{u}^h_{\ell+r})_h$, $0 \le r,s \le L$, is readily seen to be non singular. Then introduce

$$U_{\ell+t} = \sum_{s=0}^{L} (\beta^{-1})_{ts} . u_{\ell+s} \qquad 0 \le t \le L \tag{5.8}$$

From (5.7) and (2.7) we infer that

$$|U_{\ell+t} - \hat{u}^h_{\ell+t}| \le C(\ell).h^{k-1} \qquad 0 \le t \le L \tag{5.9}$$

Moreover, from the definition of E, (4.1), and the properties of the Lobatto quadrature formula, (2.1)-(2.2), we deduce that

$$||U_{\ell+t}| - 1| \le C(\ell).h^{k-1} \tag{5.10}$$

(iii) Out of the set (5.8) we construct a set of (L+1) exact eigenfunctions $U^*_{\ell+t}$, $0 \le t \le L$, corresponding to λ_ℓ and being orthonormal in $L_2(\Omega)$, by the Gram-Schmidt procedure. By a proof by induction on $t = 0,1,\ldots,L$, using (5.9)-(5.10), we readily get (5.6) for this new set.■

Remark 5.1. For the case $k = 1$ we have a similar result as in Remark 4.1.

6. Numerical example

Consider the eigenvalue problem

$$\begin{cases} -\partial_1(e^{x_1 . \ell n 4} . \partial_1 u) - \partial_2^2 u = \lambda . u & \text{in }]0,1[^2 \\ u = 0 \text{ on } x_1 = 0 \text{ and } x_1 = 1 \; ; \; \partial_2 u = 0 \text{ on } x_2 = 0 \text{ and } x_2 = 1 \end{cases}$$

By the method of separation of variables the eigenvalues are easily found to be $\lambda_{(m,n)} = n^2.\pi^2 + (\xi_m . \ell n\, 4)^2$, $n \in \mathbb{N}$, $m \in \mathbb{N}_0$, where $\xi_m > 0$ is the m-th zero of $(J_1(\xi).Y_1(2\xi) - Y_1(\xi).J_1(2\xi))$ (with the usual notation of Bessel's functions).

We use a biquadratic mesh (i.e. $k = 2$ in the analysis above) with 4^N identical rectangular elements, $N = 1,2,\ldots$. From the mixed B.C's, (4.13) and hence (4.14) need not to be valid. The relative error $|\hat{\lambda}^h_\ell - \lambda_\ell|/\lambda_\ell$ is expected to decrease by a factor 2^{-3} when increasing N by 1, according to (4.9). This

is confirmed by the results of (table I) (mostly even a factor 2^{-4} is observed). CM and NQ stand for the consistent mass and numerical quadrature method respectively, i.e. refer to (1.3) and (1.7) respectively.

Table I. CM and NQ finite element eigenvalues. NQ compared to the exact values

N	CM finite element	NQ finite element	Relative error NQ
	Eigenvalue No 1 : "exact" value = 1.963729E+01		
1	1.97548E+01	1.92480E+01	1.98E-02
2	1.96513E+01	1.96227E+01	7.44E-04
3	1.96383E+01	1.96365E+01	3.83E-05
4	1.96374E+01	1.96372E+01	2.23E-06
5	1.96373E+01	1.96373E+01	7.62E-08
	Eigenvalue No 2 : "exact" value = 2.950690E+01		
1	2.96986E+01	2.85139E+01	3.37E-02
2	2.95260E+01	2.94587E+01	1.63E-03
3	2.95082E+01	2.95041E+01	9.45E-05
4	2.95070E+01	2.95067E+01	5.69E-06
5	2.95069E+01	2.95069E+01	2.91E-07
	Eigenvalue No 3 : "exact" value = 5.911571E+01		
1	6.77548E+01	6.56383E+01	1.10E-01
2	5.94267E+01	5.86003E+01	8.72E-03
3	5.91369E+01	5.90850E+01	5.19E-04
4	5.91171E+01	5.91138E+01	3.20E-05
5	5.91158E+01	5.91156E+01	1.80E-06
	Eigenvalue No 4 : "exact" value = 7.657605E+01		
1	8.44450E+01	6.79858E+01	1.12E-01
2	7.71732E+01	7.55551E+01	1.33E-02
3	7.66308E+01	7.65388E+01	4.86E-04
4	7.65797E+01	7.65739E+01	2.75E-05
5	7.65763E+01	7.65759E+01	1.36E-06

7. References

1. Andreev, A.B., Kaschieva, V.A. and Vanmaele, M. (1990) Some results in lumped mass finite element approximation of eigenvalue problems using numerical quadrature. (submitted)
2. Babuška, I. and Osborn, J.E. (1989) Finite element-Galerkin approximation of the eigenvalues of selfadjoint problems. Math. Comp. 52, 275-297.
3. Bramble, J.H. and Hilbert, S. (1971) Bounds for a class of linear functionals with applications to Hermite interpolation. Numer. Math. 16, 362-369.
4. Ciarlet, P.G. (1976) The finite element method for elliptic problems (North-Holland, Amsterdam).
5. Dautray, R. and Lions, J.L. (1985) Analyse mathématique et calcul numérique pour les sciences et les techniques, Tome 2 (Masson, Paris).
6. Davis, P.J. and Rabinowitz, P. (1975) Methods of numerical integration (Academic Press, N.Y.).
7. Fix, G.J. (1972) Effects of quadrature errors in finite element approximation of steady state, eigenvalue and parabolic problems. In : Aziz, A.K. (editor) The mathematical foundations of the finite element method with applications to partial differential equations (Academic Press, N.Y.).
8. Vanmaele M. and Van Keer R. (1990) On an internal approximation of a class of elliptic eigenvalue problems. In : Kurzweil, J. (editor) Equadiff 7 (Teubner, Leipzig) (to appear).

Acknowlegdements

We thank M. Bernadou (INRIA) for clarifying a result in [5], p.909, similarly to which the matrix β, entering (5.8), may be seen to be regular. We also thank F. Van Der Weeën (RUG) for his help with the implementation of the FE-method used in Section 6.

Michèle Vanmaele and Roger Van Keer, Seminar of Mathematical Analysis, Faculty of Engineering Sciences, State University of Ghent, Sint Pietersnieuwstraat 39 9000 Gent, Belgium.

International Series of Numerical Mathematics, Vol. 96, © 1991 Birkhäuser Verlag Basel

CALCULATION OF THE FORMS OF SINGULARITIES IN ELLIPTIC BOUNDARY VALUE PROBLEMS

J.R. Whiteman
BICOM, Institute of Computational Mathematics,
Department of Mathematics and Statistics,
Brunel University, Uxbridge, England.

1. INTRODUCTION

It is well known that, when a boundary value problem contains a boundary singularity, the accuracy and rate of convergence of approximations produced by *standard* numerical methods are worse than those for the smooth cases. As a result great effort has been expended by many researchers over the last half century in devising numerical schemes which overcome these deficiencies, see e.g. [24], [35], [16]. In the main these numerical schemes have been produced using knowledge of the forms of the singularities, and the production of theoretical error estimates for the numerical approximations demands knowledge of the regularity of the (weak) solutions of the problems. As a result the evaluation of the forms of the singularities and their effective treatment both constructively and numerically collectively now constitute a major field of research, see e.g. the books of Grisvard, Wendland, Whiteman [16], Grisvard [14], [15], Dauge [9], Leguillon and Sanchez-Palencia [20], Kufner and Sändig [19] and the habilitation or doctoral theses of Abdel-Messieh [1], Beagles [3], Becker [6], Dobrowolski [11], von Petersdorff [25], Rank [27] and Volk [30]. The solving of *eigenvalue* problems is a continually occurring feature of the methods for determining the forms of the singularities.

The importance of the "singularity" field is apparent from the wide range of problems in which singularities occur; examples are potential problems, semiconductors, linear elastic and nonlinear solid mechanics, and Newtonian and non-Newtonial fluid mechanics, with particular applications in linear elastic and nonlinear fracture mechanics, delamination in composite structures, and forming processes. Inevitably the areas with established mathematical theory lag far behind those used in engineering practice.

This short survey article is based on the opening lecture which was presented to start this Oberwolfach meeting. The purpose of the lecture was first to give some indication of the part played by eigenvalue problems in the singularity field and to outline some of the recent results for singular problems of potential theory and linear elasticity in two and three dimensions. A second aim was to consider some recent superconvergent finite element recovered gradient error estimates for such problems and to review some more complicated problems with singularities from continuum mechanics to which finite elements are now being applied, in the hope that eventually for these also rigorous mathematical theory will be developed.

2. SINGULARITIES IN ELLIPTIC BOUNDARY VALUE PROBLEMS

2.1 *Poisson Problems in Two and Three Dimensions*

For simplicity we consider the Dirichlet problem for the Poisson equation in which $u \equiv u(\boldsymbol{x})$ satisfies

$$-\Delta[u(\boldsymbol{x})] = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega \tag{2.1}$$

$$u(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \partial\Omega \tag{2.2}$$

where Ω is either a bounded domain in $\mathbb{R}^2$ with a polygonal boundary $\partial\Omega$ or a bounded domain in $\mathbb{R}^3$ with polyhedral boundary $d\Omega$. The given function f is assumed to be as smooth as required.

With the usual Sobolev spaces $H^1(\Omega)$ and $\mathring{H}^1(\Omega)$ we define the weak formulation of problem (2.1) - (2.2) to be that in which $u \in \mathring{H}^1(\Omega)$ satisfies

$$a(u,v) = (f,v) \quad \forall \ v \in \mathring{H}^1(\Omega) \tag{2.3}$$

where

$$a(u,v) \equiv \int_{\Omega} \nabla u \, \nabla v \, dx \tag{2.4}$$

$$(f,v) \equiv \int_{\Omega} f \, v \, dx \ . \tag{2.5}$$

If in (2.3) $f \in H^{S-1}(\Omega)$, $s \geq 0$, then when $\partial\Omega$ is a regular boundary $u \in H^{S+1}(\Omega)$. Unfortunately for the types of problems of interest here, where $\partial\Omega$ contains (re-entrant) corners, u will contain corner singularities which cause the solution to have lower regularity over Ω.

In order to see this we first take Ω to be a two-dimensional polygonal domain as in Fig. 1, where the corners t_i have internal angles α_i, $i = 1,2,...,N$. Local polar coordinates (ρ_i,θ_i) are defined for each corner

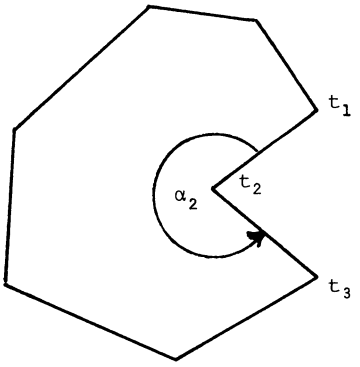


Fig. 1

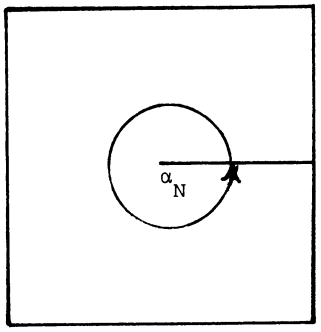


Fig. 2

t_i, together with a cut-off function $\chi_i \equiv \chi_i(\rho_i)$ such that $\chi_i \in C^{\infty}(\Omega)$ with $\chi_i(\rho_i) = 1$, $0 \leq \rho_i < \rho_0$ and $\chi_i(\rho_i) = 0$, $\rho_0 \leq \rho_i$, for some ρ_0.

If local to each corner the formal procedure is followed whereby separation of variables is used to write $u(\rho,\theta) = R(\rho)F(\theta)$, (where the subscripts on the ρ and θ have been omitted and we assume normalisation to the unit circle surrounding each vertex), then substitution into (2.1) leads to the eigenvalue problem

$$F_{\theta\theta} = -\lambda F \quad in \ G \tag{2.6}$$

$$F = 0 \qquad on \ \partial G \tag{2.7}$$

where (2.6) is the one-dimensional Laplace-Beltrami equation with eigenvalue λ and G is that part of the unit circle centred at the vertex cut off by the lines which form the angle; the points of intersection form ∂G. The first eigenvalue in (2.6) must be found in order that the (singular) behaviour of u may be described locally to the vertex.

The above simple technique has been extensively developed by several authors, see for example Grisvard [14], [15] and Dauge [8], [9], so that for the two-dimensional case there is the following result, see [14], [8] and [26].

Lemma 1 If $f \in H^{s-1}(\Omega)$, $s > 0$, $\nu_i = \pi/\alpha_i$, with $s \neq m\nu_i$ for m integer and m is odd for $\nu_i = 1/2$, then with the cut-off functions as above the solution of the weak problem (2.3) can be written in terms of a smooth part $w \in H^{s+1}(\Omega)$ and singular functions $S_i^m(\rho_i,\theta_i)$ as

$$u = \sum_{i=1}^{N} \chi_i(\rho_i) \sum_{0<m\nu_i<s} a_i^m S_i^m(\rho_i,\theta_i) + w \tag{2.8}$$

where the a_i^m are constants depending on f, and for $\nu_i \neq 1/2$ the m are integers. For the special case $\nu_i = 1/2$ the second summation is over odd values of m. In (2.8) the singular functions $S_i^m(\rho_i,\theta_i)$ are given by

$$S_i^m(\rho_i,\theta_i) = \rho_i^{m\nu_i} \sin m\nu_i \theta_i \quad \text{for } m\nu_i \notin N$$

$$= \rho^{m\nu_i} (\log \rho_i \sin m\nu_i \theta_i + \theta_i \sin m\nu_i \theta_i) \text{ , otherwise .} \tag{2.9}$$

The often quoted example of (2.8) is that for problem (2.6) - (2.7) in a rectangular domain containing a slit, $\alpha_N = 2\pi$, as in Fig. 2. Here, with $f \in L_2(\Omega)$, $s = 3/2 - \epsilon$, the dominant singular term in u arises at the re-entrant corner and has the form $\rho^{1/2}$ so that $u \in H^{3/2-\epsilon}(\Omega)$ and $w \in H^{5/2-\epsilon}(\Omega)$. Alternatively we may write $u \in W^2_{4/3-\epsilon}(\Omega)$).

For problem (2.1), (2.2) in three dimensions when Ω is a polyhedral domain, the boundary $\partial\Omega$ contains both edges and vertices and singularities can be associated with each of these. The simple idea set out above of separating the variables at a *vertex* can be extended to three dimensions, see e.g. Stephan and Whiteman [29] and Beagles and Whiteman [5]. Suppose that the spherical polar coordinates local to each vertex are (ρ_i,θ_i,ϕ_i), then the singular functions in u associated with the vertices have the form $\rho_i^\gamma F_i(\theta_i,\phi_i)$, where the γ depends on the eigenvalues of the two-dimensional Laplace -Beltrami problem

$$\Delta_{\theta\phi} F = -\lambda F \quad in \quad G \tag{2.10}$$

$$F = 0 \qquad on \quad \partial G \tag{2.11}$$

where now G is the spherical surface cut off from the unit ball centred on the vertex by the faces of $\partial\Omega$ which make up the corner and the curves of intersection constitute ∂G. Provided the vertex is (θ,ϕ)-separable, see [5], a further separation of variables for the $F(\theta,\phi)$ can be undertaken. In certain cases this leads to well known equations, such as the generalised Legendre equation, which can be solved in closed form, see [5] and [29].

Based upon the above analysis and the two-dimensional vertex decomposition (2.8), and accepting that we expect singularities associated with both vertices and edges in the three-dimensional context, it should be the general aim to produce a decomposition of $u = u(x,y,z)$ of the form

$$u = \underset{\substack{i=1 \\ (vertices)}}{\sum^{N}} \chi_i(\ldots) + \underset{\substack{j=1 \\ (edges)}}{\sum^{M}} \tilde{\chi}_j(\ldots) + w\,, \tag{2.12}$$

where the χ_i are as before and the $\tilde{\chi}_j$ are cut-off functions based on the radial distance orthogonal to an edge. The establishing of decompositions of the form (2.12) is highly technical, but results of this type are now being evaluated, see Dauge [8]. [9], von Petersdorff and Stephan [26].

Important applications of Poisson problems in three dimensions occur for much more general domains than polyhedra, and for other boundary conditions than (2.2). The domains involve curved faces, curved edges and conical corners, and many problems have mixed boundary conditions. The forms of singularities and the regularity of the solutions are needed for all these cases. Important work for conical corners has been done by Kondratiev [17], [18].

2.2 Problems of Linear Elasticity in two and three dimensions

We next consider problems of two- or three-dimensional linear elasticity defined in Ω, where Ω is as described in Section 2.1, in which the displacement $\boldsymbol{u} \equiv \boldsymbol{u}(\boldsymbol{x})$ satisfies the Lamé equations

$$-\mu\Delta\boldsymbol{u} - (\lambda+\mu)\,grad\ div\ \boldsymbol{u} = \boldsymbol{f}\,, \quad \boldsymbol{x} \in \Omega\,, \tag{2.13}$$

together with boundary conditions

$$\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{0}\,, \quad \boldsymbol{x} \in \partial\Omega_C, \tag{2.14}$$

$$\sum_{j}^{2(3)} \sigma_{ij}(u(\boldsymbol{x}))n_j = g_i(\boldsymbol{x})\,, \quad \boldsymbol{x} \in \partial\Omega_T\,,\ 1 \le i \le 2(3)\,, \tag{2.15}$$

where $\partial\Omega = \partial\Omega_C \cup \partial\Omega_T$, $\boldsymbol{n}$ is the unit outward normal vector to the boundary, the tractions $\boldsymbol{g} \in (L_2(\partial\Omega_T))^{2(3)}$ and the volumic forces $\boldsymbol{f} \in (L_2(\Omega))^{2(3)}$.

For admissible displacement vectors $\boldsymbol{v} \in (H^1(\Omega))^{2(3)}$ we define

$$V \equiv \left\{\boldsymbol{v} : \boldsymbol{v} \in (H^1(\Omega))^{2(3)},\ \boldsymbol{v}\Big|_{\partial\Omega_C} = 0\right\}$$

and set up the weak form of (2.13) - (2.15), which is

$$find\ \boldsymbol{u} \in V \ni a(\boldsymbol{u},\boldsymbol{v}) = F(\boldsymbol{v}) \quad \forall\ \boldsymbol{v} \in V\,, \tag{2.16}$$

where

$$a(\boldsymbol{u},\boldsymbol{v}) = \int_{\Omega} \Big\{\lambda div\ \boldsymbol{u}\ div\ \boldsymbol{v} + 2\mu \sum_{i,j} \epsilon_{ij}(\boldsymbol{u})\,\epsilon_{ij}(\boldsymbol{v})dx$$

$$F(\boldsymbol{v}) = \int_{\Omega} \boldsymbol{f}\,.\,\boldsymbol{v}\,dx + \int_{\partial\Omega_T} \sum_{i} g_i\,v_i\,ds$$

with

$$\epsilon_{ij}(\boldsymbol{v}) \equiv \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right).$$

The case of problem (2.16) in which Ω is a two-dimensional polygonal domain as in Fig. 1, with g in (2.15) set to zero, and again where the boundary $\partial\Omega$ contains corners t_i, has been considered by von Petersdorff [25]. In [25] (Theorem 1.1) a decomposition of $\boldsymbol{u}$ of the form (2.8) is given and the appropriate singular functions $S_i^m(\rho_i,\theta_i)$ are presented. As is usually the case it is necessary to solve transcendental equations to obtain these.

Again for the case $\boldsymbol{g} = 0$ on $\partial\Omega_T$ the three-dimensional form of problem (2.16) in a polyhedron is considered in [25] and a decomposition of the solution $\boldsymbol{u}$ comparable to (2.12) is given. The same problem, (2.16), but for the case of a circular conical domain and with non-homogeneous boundary conditions in both (2.14) and (2.15), is considered by Beagles and Sändig [4]. Singular functions are derived which give the form of the solution near the vertex so that again a decomposition of $\boldsymbol{u}$ into a smooth part and a series involving the singular functions is possible. Finally mention must be made of the work of Maz'ja and Plamanevskii [21], [22] who have treated extensively problems of three-dimensional linear elasticity containing boundary singularities.

3. FINITE ELEMENT METHODS AND ERROR ESTIMATES

For the Poisson problems and problems of linear elasticity considered in Section 2 the finite element method is applied respectively via the weak formulations (2.3) and (2.16). This is done using Galerkin techniques in which the test and trial functions are piecewise polynomials defined over partitions of the domain Ω. These are so defined that the Galerkin approximations u_h and $\boldsymbol{u}_h$ are respectively conforming approximations to the weak solutions u and $\boldsymbol{u}$ of (2.3) and (2.16). If the partitions consist of triangular or tetrahedra elements and piecewise linear or piecewise quadratic functions are used, then for h the generic mesh size, the appropriate energy error norms $|u - u_h|_E$ and $|\boldsymbol{u} - \boldsymbol{u}_h|_E$ can be estimated in terms of a power of h multiplied by a seminorm of u or $\boldsymbol{u}$. The power of h is dependent on the degree of the polynomial used and, very importantly on the regularity of the solution which determines what seminorm is available. The energy norms involve gradients of u and u_h, ($\boldsymbol{u}$ and $\boldsymbol{u}_h$). It has been found that for very regular triangular and tetrahedral meshes, using averaging techniques, see [13], [32], [33], gradients of the approximations can be recovered at the vertices, and for piecewise quadratics at the remaining nodes in each element, and hence recovered gradient functions can be defined over Ω. These have the property that the rate of convergence of the recovered gradients to the gradients of the solutions of the problems is in general higher than that without recovery; this is the phenomenon of *superconvergence* and is again dependent on the regularity of the solution. The method of recovery was first analysed for Poisson problems involving boundary singularities by Wheeler and Whiteman [31], but has since been extended to elasticity problems with singularities by Goodsell and Whiteman [13], [33]. Knowledge of the regularity of solutions is thus vital to the success of the recovery techniques, as indeed it is in general for error estimates, and this demonstrates the need for the analysis of singularities as in Section 2.

Another aspect of finite element analysis that is becoming increasingly important is that of *adaptivity*. The recovered gradient functions above have a part to play here as they can be used for producing error estimates which determine where a mesh must be adapted. Closely allied to this are a posteriori error estimates which again demand knowledge of regularity of the solutions.

4. MORE PRACTICAL ASPECTS

The analysis of Section 2 was done in the theoretical context of Poisson and Lamé problems. As was said earlier these problems have important practical applications, and the determination of the forms of singularities, particularly for linear elastic fracture, has been a topic of research for engineers. In the

neighbourhood of a crack tip in a linear elastic material the near-tip displacement and stress fields are needed for determining stress intensity factors and fracture criteria, see Rice [28], Whiteman and Thompson [34]. For two-dimensional linear elastic fracture problems these fields are well understood. However, in three dimensions the situation is much lass satisfactory and contains open questions, see Folias [12], Bazant [2].

All the discussion to date has concerned linear problems for which much rigorous theory is available. If one considers, for example, a solid containing a crack for which the deformation is elastic-plastic, then a totally new situation arises. Before the form of the singularity in a ductile zone near a crack tip can be considered, it is necessary to decide which model of plasticity, for example flow theory or deformation theory, will be assumed. The analytic treatment of nonlinear problems of this type is much less advanced mathematically than that for linear cases.

Finally, problems of singularities arise for time dependent deformation as in viscoelastic materials, and for the Newtonian and non-Newtonian flow of liquids, see e.g. Moffat [23], Davies [10], Crochet, Davies and Walters [7]. All these contexts demand knowledge of the forms of the singularities.

ACKNOWLEDGEMENT

Much of the work behind this lecture has been done in collaboration with colleagues. It is a great pleasure to acknowledge the helpful discussions and benefits that I have had. In particular for the field of singularities I am indebted to Adam Beagles, Pierre Grisvard, Anna-Margarete Sändig, Ernst Stephan, Jay Walton, Michael Warby and Wolfgang Wendland.

REFERENCES

1. ABDEL-MESSIEH, Y.S. (1989) A Numerical Assessment of the Form of Singularities in Elliptic Problems. Ph.D. Thesis, University of Manchester.
2. BAZANT, Z.P. (1974) Three-dimensional harmonic functions near termination or intersection of gradient singularity lines. Int. J. Eng. Sci. **12**, 221-243.
3. BEAGLES, A.E. (1987) Theoretical and Numerical Treatment of Singularities in Elliptic Boundary Value Problems. Ph.D. Thesis, Brunel University.
4. BEAGLES, A.E.and SÄNDIG, A.-M. , Singularities of rotationally symmetric solutions of boundary value problems for the Lamé equations. (To appear in ZAMM.)
5. BEAGLES, A.E. and WHITEMAN, J.R. (1989) General conical singularities in three-dimensional Poisson problems. Math. Meth. Appl. Sci. 11, 215-235.
6. BECKER, I. (1989) Numerische Behandlung von Ecken - und Kantensingularitäten elastischer Felder für dreidimensionale Rissprobleme. Ph.D. Thesis, University of Karlsruhe.
7. CROCHET, M.J., DAVIES, A.R. and WALTERS, K. (1984) Numerical Simulation of Non-Newtonian Flow. Elsevier, Amsterdam.
8. DAUGE, M. (1986) Regularitiés et Singularities des Solutions de Problemes aux limites Elliptiques sur les Domaines Singular de Type à Coins. Ph.D. Thesis, University of Nantes.
9. DAUGE, M. (1988) Elliptic Boundary Value Problems on Corner Domains. Lectures Notes in Mathematics 1341, Springer Verlag, Berlin.
10. DAVIES, A.R. (1988) Re-entrant corner singularities in non-Newtonian flow. Part I: Theory. J. Non-Newtonian Fluid Mech. **29**, 269-293.
11. DOBROWOLSKI, M. (1981) Numerical Approximation of Elliptic Interface and Corner Problems. Habilitationsschrif, University of Bonn.
12. FOLIAS, E.S. (1980) Method of solution of a class of three-dimensional elastostatic problems under Mode I loading. Int. J. Fracture **16**, 335-348.
13. GOODSELL, G. and WHITEMAN, J.R. (1990) Pointwise superconvergence of recovered gradients for piecewise linear finite element approximations to problems of planar linear elasticity. Numer. Meth. Partial Differential Equations **6**, 59-74.
14. GRISVARD, P. (1985) Elliptic Problems in Nonsmooth Domains. Pitman, London.

15. GRISVARD, P. (1983) Singular solutions of elliptic boundary problems in polyhedra. Portugalae Math. **41**, 4-20.
16. GRISVARD, P., WENDLAND, W. and WHITEMAN, J.R. (eds) (1985) Singularities and Constructive Methods for Their Treatment. Lecture Notes in Mathematics 1121, Springer Verlag, Berlin.
17. KONDRATIEV, V.A. (1967) Boundary problems for elliptic equations in domains with conical or angular points. Trans. Moscow Math. Soc. **16**, 227-313.
18. KONDRATIEV, V.A. (1970) The smoothness of a solution of Dirichlet's problem for 2nd order elliptic equations in a region with a piecewise smooth boundary. Diff. Eqns. **6**, 1392-1401.
19. KUFNER, A. and SÄNDIG, A.-M. (1987) Some Applications of Weighted Sobolev Spaces. Teubner, Leipzig.
20. LEGUILLON, D. and SANCHEZ-PALENCIA, E. (1987) Computation of Singular Solutions in Elliptic Problems and Elasticity, Masson, Paris.
21. MAZ'JA, V.G. and PLAMENEVSKII, B.A. (1976) On the coefficients in the asymptotics of elliptic boundary value problems near the edge. Trans. Moscow Math. Soc. **17**, 970-974.
22. MAZ'JA, V.G. and PLAMENEVSKII, B.A. (1977) About the coefficients in the asymptotics of the solutions of elliptic boundary value problems in domains with conical points. Math. Nachr. **76**, 29-41.
23. MOFFAT, H.K. (1964) Viscous and resistive eddies near a sharp corner. J. Fluid Mech. **18**, 1-18.
24. MOTZ, H. (1946) The treatment of singularities of partial differential equations by relaxation methods. Q. Appl. Math. **4**, 371-377.
25. von PETERSDORFF, T. (1989) Randwertprobleme der Elastizitätstheorie für Polyeder-Singularitäten und Approximation mit Randelementmethoden. Ph.D. Thesis, Technische Hochschule Darmstadt.
26. von PETERSDORFF, T. and STEPHAN, E.P., Decompositions in Edge and Corner Singularities for the Solution of the Dirichlet Problem of the Laplacian in a Polyhedron. (To appear in Math. Nachr.)
27. RANK, E. (1985) A-posteriori-Fehlerabschätzungen und adaptive Netzverfeinerung für Finite-Element - und Randintegralelement-Methoden. Ph.D. Thesis, University of Munich.
28. RICE, J.R. (1968) Mathematical analysis in the mechanics of fracture. pp.191-311 of H. Liebowitz (ed.). Fracture, Vol.II. Academic Press, New York.
29. STEPHAN, E.P. and WHITEMAN, J.R. (1988) Singularities of the Laplacian at corners and edges of three-dimensional domains and their treatment with finite element methods. Math. Meth. Appl. Sci. **10**, 339-350.
30. VOLK, K. (1989) Zur Berechnung von Singularfunktionen dreidimensionaler elastischer Felder. Ph.D. Thesis, University of Stuttgart.
31. WHEELER, M.F. and WHITEMAN, J.R. (1987) Superconvergent recovery of gradients on subdomains for piecewise linear finite element approximations. Numer. Meth. Partial Differential Equations **3**, 65-82.
32. WHITEMAN, J.R. and GOODSELL, G. (1989) Some gradient superconvergence results in the finite element method. pp.182-260 of P.R. Turner (ed.) Numerical Analysis and Parallel Processing. Lecture Notes in Mathematics 1397, Springer Verlag, Berlin.
33. WHITEMAN, J.R. and GOODSELL, G. A survey of gradient superconvergence for finite element approximations to second order elliptic problems on triangular and tetrahedral meshes. (To appear in J.R. Whiteman (ed.) The Mathematics of Finite Elements and Applications VII, MAFELAP 1990. Academic Press, London.)
34. WHITEMAN, J.R. and THOMPSON, G. (1985) Analysis of strain representation in linear elasticity by both singular and nonsingular finite elements. Numer. Meth. Partial Differential Equations **2**, 85-104.
35. WOODS, L.C. (1953) The relaxation treatment of singular points in Poisson's equation. Q.J. Mech. Appl. Math. **6**, 163-185.