Rostocker Mathematisches Kolloquium

Heft 49

Gewidmet
den Herren

Prof. Dr. rer. nat. habil. Lothar Berg
Prof. Dr. rer. nat. habil. Wolgang Engel
Prof. Dr. rer. nat. habil. Gerhard Pazderski
Prof. Dr.-Ing. habil. Hans–Wolfgang Stolle

Universität Rostock

Fachbereich Mathematik

1995

https://doi.org/10.18453/rosdok_id00003275

Wissenschaftliche Leitung: Prof. Dr. Wildenhain
(Sprecher des Fachbereichs Mathematik)
Dr. Werner Plischke

Redaktionelle Bearbeitung: Dr. Werner Plischke

Herstellung der Druckvorlage: S. Dittmer, W. Hartmann, H. Schubert

Universität Rostock
Fachbereich Mathematik
18051 Rostock
Bundesrepublik Deutschland

Redaktionsschluß: 28. Oktober 1995

Das **Rostocker Mathematische Kolloquium** erscheint dreimal im Jahr und ist im Rahmen des Schriftentausches über die Universität Rostock, Universitätsbibliothek, Tauschstelle, 18051 Rostock, Bundesrepublik Deutschland, zu beziehen.

Außerdem bestehen Bezugsmöglichkeiten für Bestellungen aus Deutschland und dem Ausland über die Universität Rostock, Abteilung Wissenschaftspublizistik, 18051 Rostock, Bundesrepublik Deutschland

Zitat–Kurztitel: Rostock. Math. Kolloq. (1995) **49**

Universität Rostock
Abteilung Wissenschaftspublizistik
18051 Rostock, Bundesrepublik Deutschland
Telefon 4 40 55 20
Druck: Universitätsdruckerei 104/96
01000

Inhalt

Günther Wildenhain	*Generationswechsel am Fachbereich Mathematik der Universität Rostock*	5
Manfred Krüppel	*Ein asymptotischer Fixpunktsatz für Lipschitz-stetige Operatoren in uniform konvexen Banach-Räumen*	11
Konrad Engel Gabriele Sauerbier	*An application of* Dilworth*'s Theorem to a problem on free* Lie*-algebras*	25
Ingo Steinke	*Asymptotic optimal classification of three populations by ML-estimators*	31
Klaus–Dieter Drews	*Eine vereinfachende Behandlung der Konstruierbarkeit von p^k– Ecken*	44
Harry Poppe	*A theorem on summable families in normed groups*	51
Dieter Schott	*Basic properties of Fejer monotone sequences*	57
Günter Mayer	*On a Unified Representation of Some Interval Analytic Algorithms*	75
Reinhard Knörr	*Auslander–Reiten sequences and a certain ideal in mod–FG*	89
Reinhard Strecker	*$\mathcal{H}-$commutative $\Delta-$semigroups*	98
Jürgen Prestin Kathi Selig	*On the Gram Matrix of Translates of De la Vallée Poussin Kernels*	105
Gerlind Plonka	*Approximation Properties of Multi–Scaling Functions: A Fourier Approach*	115

Raimond Strauss	*Numerische Integration von hypersingulären Integralen*	127
M. F. Newman G. Sauerbier J. Wisliceny	*Groups of prime-power order with a small number of relations*	141
F. G. Boese	*Stabilität dynamischer Systeme unter Parameterunsicherheit*	155
Peter Takáč	*Dynamics on the Attractor for the Complex Ginzburg-Landau Equation*	163

Generationswechsel am Fachbereich Mathematik der Universität Rostock

Am Fachbereich Mathematik der Universität Rostock vollzieht sich derzeit ein sanfter, aber auf lange Sicht mit Sicherheit deutlich spürbarer Generationswechsel. Nach den personellen Veränderungen im Zusammenhang mit dem Übernahmeverfahren im Jahre 1992 zieht das altersbedingte Ausscheiden der Professoren Lothar Berg, Wolfgang Engel, Gerhard Pazderski und Hans-Wolfgang Stolle eine weitere einschneidende Änderung in der Personalstruktur nach sich.

Die genannten Kollegen, die alle in den Jahren 1959 bzw. 1965 nach Rostock berufen wurden, haben ausnahmslos Verdienste um die Mathematik in Rostock. Daher soll ihnen mit einem gemeinsamen Festkolloquium und der Widmung eines Heftes der Schriftenreihe „Rostocker Mathematisches Kolloquium“ eine besondere Ehrung zuteil werden.

Die frühere Sektion Mathematik der Universität Rostock genoß sowohl innerhalb der Universität als auch über die Grenzen der Universität hinaus den Ruf, durch eine menschlich angenehme Atmosphäre und ein politisch sehr gemäßigtes Klima geprägt zu sein. Nach Einsicht in früher geheime Akten stellt sich heute heraus, daß dieser von der großen Mehrheit der Sektionsangehörigen zweifellos als sehr wohltuend empfundene Zustand von den damaligen politischen Instanzen schlichtweg als „politische Unzuverlässigkeit“ eingestuft wurde. In der Tat war die Situation an der Sektion Mathematik dadurch charakterisiert, daß zwei Drittel der Lehrstuhlinhaber nicht Mitglied der staatstragenden SED waren.

Diese Situation hatte DDR-weit gewiß Seltenheitswert und bot daher übergeordneten Gremien immer wieder Anlaß zum Mißtrauen. Für die SED war es unter diesen Umständen schwerer als anderswo, ihren Machtanspruch uneingeschränkt durchzusetzen. Die Leitungen der Sektion und ihrer Wissenschaftsbereiche bemühten sich spürbar, rein wissenschaftliche Prinzipien und menschlichen Anstand hochzuhalten. Die jetzt bzw. demnächst aus dem Dienst ausscheidenden Kollegen, die alle der oben beschriebenen Zweidrittelmehrheit angehörten, haben einen entscheidenden Anteil daran, daß dies weitgehend gelungen ist. Die Ausübung von Leitungsfunktionen war damals nicht leicht und für Wissenschaftler mit einer kritischen Grundhaltung zum Staat nicht immer ohne Gewissenskonflikte und Kompromisse zu bewältigen. Die genannten Kollegen haben - auf zum Teil sehr unterschiedlichen Ebenen und in sehr unterschiedlichem Umfang - Leitungsfunktionen wahrgenommen. Man kann ihnen heute bescheinigen, daß sie die ihnen übertragenen Funktionen im Interesse der Angehörigen der

Sektion im oben beschriebenen Sinne ausgeübt haben. H.-W. Stolle war der erste Sektionsdirektor in den Jahren 1968 bis 1971. Es war ein Glück für die Sektion, daß die personellen und strukturellen Maßnahmen der von Staat und Partei verordneten 3. Hochschulreform unter seiner Leitung in sehr moderater und gemilderter Form umgesetzt werden konnten, bevor dann 1971 bis Anfang 1974 eine Phase folgte, die als ungebremste SED-Herrschaft bezeichnet werden muß. Offenbar war aber die Sektion dafür nicht „reif", und daher sah sich die Universitätsleitung Anfang 1974 veranlaßt, W. Engel die Leitung zu übertragen. Er hat sie dann 13 Jahre lang mit viel Diplomatie und Geschick, immer auf das Wohl der Sektion bedacht, in der Hand behalten, bevor aus verschiedenen Anlässen in den letzten Jahren der DDR seine „Entmachtung" erfolgte. Dennoch mußte auch sein Nachfolger Rücksicht auf die besondere politische Situation nehmen, die grundsätzlich nicht beseitigt werden konnte. Die Aufgaben der Wendezeit stellten sich für die frühere Sektion und den jetzigen Fachbereich Mathematik wie folgt dar. Einerseits sollte das erfolgreiche wissenschaftliche Profil und das angenehme menschliche Klima erhalten werden, andererseits mußten Konsequenzen aus vorhandenen Defiziten, Fehlleistungen und menschlichem Versagen gezogen und die uneingeschränkte Freiheit von Lehre und Forschung hergestellt werden. Daran, daß beides ohne gravierende Rückschläge gelungen ist, haben die genannten vier Kollegen ihren Anteil. Bevor eine Würdigung im Detail erfolgt, soll noch ein übergreifender Aspekt hervorgekehrt werden. Zurückgehend nicht zuletzt auf Professoren wie R. Kochendörffer hat die mathematische Lehrkultur in Rostock traditionell einen guten Standard. Unsere jetzt ausscheidenden Kollegen haben - jeder auf seine Weise - dazu maßgeblich beigetragen. Zur Lehrkultur gehört dabei auch, daß Studenten zum eigenen Nachdenken angeregt und zur eigenen Anstrengung gezwungen werden. Ein besonderer Aspekt ist die Vermittlung mathematischen Wissens und Könnens an Studenten anderer Studienrichtungen. Gerade in dieser Hinsicht hat H.-W. Stolle während seiner jahrzehntelangen Lehrtätigkeit für Ingenieurstudenten Maßstäbe gesetzt. Das, was die ausscheidenden Kollegen ausnahmslos durch ihre langjährige, sehr intensive und sehr vielseitige Lehrtätigkeit für die mathematische Kultur in Rostock geleistet haben, könnte leicht als selbstverständlich unterschätzt werden. Es sei daher hier nochmals ausdrücklich und mit Dank an die Akteure hervorgehoben, insbesondere im Hinblick auf die sehr verbreiteten Klagen über das mangelhafte Engagement in der Lehre vieler Professoren in den alten Bundesländern. Im folgenden soll (in alphabetischer Reihenfolge) versucht werden, die Verdienste der vier genannten ausscheidenden Kollegen in kurzen Zügen zusammenfassend zu würdigen.

Lothar Berg, der am 28. 07. 1995 seinen 65. Geburtstag feierte, ist z. Zt. noch im Amt. Er studierte 1949 bis 1953 und promovierte 1955 jeweils in Rostock. Nach den Stationen Ilmenau (Oberassistent und Dozent, Habilitation 1957) und Halle (1959 mit 29 Jahren Berufung zum Professor) kehrte er im Jahre 1965 als Professor mit Lehrstuhl an das damalige Mathematische Institut nach Rostock zurück. In den 30 Jahren seines Wirkens in Rostock hat L. Berg das Profil der Mathematik an dieser Universität ganz wesentlich mitgeprägt. Bestechend ist seine außergewöhnliche, bis heute nicht nachlassende wissenschaftliche Kreativität, die sich nach gegenwärtigem Stand in etwa 220 Publikationen in wissenschaftlichen Zeitschriften und 8 Büchern niederschlägt. Ausgehend von fundamentalen Beiträgen zur Operatorenrechnung umspannen diese Arbeiten einen großen Bogen über die Asymptotik, die Funktionalanalysis, die Algebraische Analysis, die Operator-Funktionalgleichungen bis zur Numerischen Analysis. Oft wurde er zu Vorträgen eingeladen, und - soweit ihm die Annahme der Einladungen erlaubt wurde - hat er seine wissenschaftlichen Ergebnnisse in vielen Ländern der Welt erfolgreich präsentiert. Lothar Berg hat Generationen junger Mathematiker ausgebildet und sich große Verdienste um die Heranbildung des wissenschaftlichen Nachwuchses erworben. Er hat 28 Doktoranden betreut und 14 davon bis zur Habilitation geführt. Viele seiner Schüler sind heute selbst anerkannte Hochschullehrer. Seine wissenschaftlichen Leistungen und seine Kompetenz haben in der Fachwelt vielfältige Anerkennung gefunden. L. Berg ist Mitglied der Deutschen Akademie der Naturforscher Leopoldina, Mitherausgeber mehrerer mathematischer Zeitschriften, wie zum Beispiel der Zeitschrift für Angewandte Mathematik und Mechanik. Er arbeitet mit in Auswahlkommissionen der Studienstiftung des deutschen Volkes sowie des DAAD und ist ein weltweit gefragter Gutachter. An der Rostocker Universität wurden seine Leistungen 1978 mit der Verleihung des Universitätspreises für Forschung gewürdigt. L. Berg zeichnet sich durch geistige Unbestechlichkeit und Offenheit aus, was ihm in der DDR-Zeit gelegentlich Probleme bis hin zu Maßregelungen einbrachte. Er scheute sich auch nicht, unabhängige und eigenständige politische Ansichten zu vertreten. Die dabei demonstrierte Zivilcourage verdient heute in Erinnerung gerufen zu werden. Nach der deutschen Wiedervereinigung hat sich L. Berg mit Engagement für den Neuaufbau der Universität Rostock zur Verfügung gestellt. Er hat in der Ehrenkommission der Universität mitgearbeitet und ist zur Zeit gewähltes Mitglied des Akademischen Senats der Universität Rostock, des Rates der Mathematisch-Naturwissenschaftlichen Fakultät sowie des Fachbereichsrates Mathematik. Für diesen Zeit- und Kraftaufwand, den er neben seinem nach wie vor hohen Einsatz in Lehre und Forschung investiert, gebührt ihm Dank und Anerkennung.

Wolfgang Engel, geboren am 10. 04. 1928, wurde nach Ablegung des Staatsexamens in den Fächern Mathematik und Physik 1950, der Promotion 1953 und der Habilitation 1957 an der Martin-Luther-Universität in Halle im Jahre 1959 als Professor für Mathematik an die Universität Rostock berufen. Sein Arbeitsgebiet war zunächst die Algebraische Geometrie, insbesondere die Theorie der Cremona-Transformationen. Später wandte er sich immer mehr Fragen der Schulmathematik sowie allgemeinen Problemen der mathematischen Bildung und der Geschichte der Mathematik zu. 65 Artikel in wissenschaftlichen Zeitschriften und Sammelwerken legen davon Zeugnis ab. Problemen der Schule sowie der Lehrerbildung galt immer sein besonderes Interesse. In zahlreichen Gremien engagierte er sich dafür. W. Engel war einer der Begründer der Mathematik-Olympiaden Junger Mathematiker sowie zweimal Präsident der Jury der Internationalen Mathematik-Olympiade. Das System der Schul-, Kreis-, Bezirks- und DDR-Olympiaden entwickelte sich unter seiner Leitung mehr und mehr zu einem Instrument der Förderung mathematisch talentierter Schüler, die sich in der Regel an den Universitäten und Hochschulen fortsetzte. Es handelte sich hierbei um eine fachlich motivierte und dominierte Nachwuchsförderung, bei der die ideologische Komponente fast völlig zurückgedrängt war. So war es möglich, daß W. Engel in das Committee of the World Federation of National Mathematics Competitions und nach der Wende in das Kuratorium des Bundeswettbewerbs Mathematik gewählt wurde. Sein Engagement für den wissenschaftlichen Nachwuchs fand auch Ausdruck in seinem Einsatz für die Gründung und Entwicklung von Spezialklassen und -schulen für Mathematik und Naturwissenschaften/Technik zur Hochbegabtenförderung. W. Engel war während der DDR-Zeit 13 Jahre Direktor der damaligen Sektion Mathematik. Sein Wirken in diesem Amt verdient auch aus der Rückschau Respekt und Anerkennung. Sein Hauptanliegen galt stets der Pflege der Wissenschaft in Lehre und Forschung sowie der Förderung des wissenschaftlichen Nachwuchses. Es gelang ihm, die Sektion und die Wissenschaftler soweit wie irgend möglich von ideologischen Einflüssen abzuschirmen. Dadurch konnte an der Sektion ein Klima erhalten werden, das durch wissenschaftliches Ethos und Kollegialität geprägt war. Eine ähnliche Wertung kann für sein langjähriges Wirken als Vorsitzender der Mathematischen Gesellschaft der DDR gegeben werden. Nicht unerwähnt darf bleiben, daß W. Engel die Schriftenreihe „Rostocker Mathematisches Kolloqium“ begründete und als leitender Herausgeber der Studienbücherei Mathematik für Lehrer fungierte, einem 20-bändigen Lehrwerk für die Mathematik-Ausbildung für Lehrerstudenten. Sein Wirken fand in der DDR eine Würdigung, z. B. durch den Titel Verdienter Lehrer des Volkes, den Vaterländischen Verdienstorden in Bronze und die Berufung zum Mitglied der Akademie der Pädagogischen Wissenschaften.

Wie W. Engel, so kommt auch Gerhard Pazderski von der Martin-Luther-Universität Halle-Wittenberg. Geboren am 11. 01. 1928, studierte er in Halle Mathematik und im Nebenfach Physik, legte 1953 das Diplom ab, promovierte 1958 und habilitierte sich 1963, jeweils zu Fragen der Gruppentheorie. Nach einer einjährigen Dozentur in Halle wurde G. Pazderski im Jahre 1965 als Professor nach Rostock berufen. Er ist ein Leben lang der Gruppentheorie treu geblieben und entwickelte sich immer mehr zu einem international bekannten und anerkannten Spezialisten auf diesem zentralen Gebiet der Algebra. In der ehemaligen DDR war er der einzige ausgewiesene Fachmann der Gruppentheorie. Er setzte damit eine bereits von R. Kochendörffer begründete Rostocker Tradition fort. Es gelang ihm, eine Forschungsgruppe aufzubauen, die eine erfreuliche Ausstrahlung besitzt und durch ein enges mathematisches Beziehungsgeflecht mit vielen Forschungseinrichtungen der Bundesrepublik und der ganzen

Welt verbunden ist. Leider war es G. Pazderski in der DDR-Zeit verwehrt, seine vielfältigen Kontakte zu ausländischen Mathematikern sowie aus Westdeutschland persönlich zu pflegen. Zahlreich vorliegende Einladungen zu Tagungen und Kolloquiumsvorträgen in das westliche Ausland durfte er nicht annehmen. Nach der Wende konnten diese Kontakte jedoch wesentlich belebt werden, und der gute Ruf der Rostocker Gruppentheorie trug dazu bei, nach dem Übergang von G. Pazderski in den Ruhestand den Lehrstuhl mit Prof. Dr. R. Knörr so zu besetzen, daß die Rostocker gruppentheoretische Tradition gewahrt bleibt. Das wissenschaftliche Werk von G. Pazderski spiegelt sich u. a. in etwa 30 Publikationen sowie in einem sehr aktiven, von ihm ausgebildeten und betreuten wissenschaftlichen Nachwuchs. Insgesamt betreute er 10 Promotionen und Habilitationen. G. Pazderski war und ist auf Grund seiner Fachkompetenz sehr gefragt als Gutachter für Dissertationen, Habilitationen, Berufungen, Verlagsmanuskripte u. a., sowie als Mitglied von wissenschaftlichen Gremien. Zum Beispiel war er Mitglied des Redaktionskollegiums „Mitteilungen der Mathematischen Gesellschaft der DDR". Von 1973 bis 1990 leitete er den damaligen Wissenschaftsbereich Theoretische Mathematik an der Sektion Mathematik der Universität Rostock. Bei allem, was er anpackt, zeichnet er sich durch Solidität, Kollegialität und Zuverlässigkeit aus. Eine Würdigung des Wirkens von G. Pazderski wäre aber unvollständig, würde man seine populärwissenschaftliche Tätigkeit vergessen, die Problemen der bildenden Kunst gewidmet ist und sich sowohl in Publikationen als auch in einzelnen Vorträgen, insbesondere für die frühere URANIA-Gesellschaft, niederschlägt. Dies rundet das Bild seiner stillen und unaufdringlichen, aber dennoch nicht weniger erfolgreichen und eindrucksvollen Persönlichkeit ab.

Hans-Wolfgang Stolle, geboren am 07. 07. 1927, ist der „älteste" der hier zu würdigenden Kollegen. Er ist im Sommer dieses Jahres aus dem Dienst ausgeschieden. Im März 1991 wurde er in das Amt des Fachbereichssprechers gewählt und hat es dankenswerter Weise noch bis zum Sommer 1994 wahrgenommen. Mit diesem Amt war in der schwierigen Phase der personellen Erneuerung unserer Universität eine ganz besondere Verantwortung verbunden. H.-W. Stolle ist dieser Verantwortung in hohem Maße gerecht geworden, einerseits durch sein menschliches Einfühlungsvermögen, andererseits durch Konsequenz bei der Durchsetzung des Erneuerungsgedankens. Auch durch seine Mitarbeit in der Ehrenkommission hat sich H.-W. Stolle verdient gemacht. Es ist ganz wesentlich seinem Einfluß zu verdanken, daß der Umstrukturierungsprozeß am Fachbereich Mathematik unter Wahrung einer höchstmöglichen Gerechtigkeit und Effizienz vollzogen werden konnte. H.-W. Stolle studierte von 1947 bis 1951 Mathematik und Physik in Rostock mit dem Abschluß als Diplom-Mathematiker 1951. In den Jahren 1952 bis 1961 war er als Assistent bzw. Oberassistent am Fachbereich Statik und Dynamik der damaligen Schiffbautechnischen Fakultät Rostock tätig. Seine Promotion zum Dr.-Ing. erfolgte 1956 und seine Habilitation an derselben Fakultät im Jahre 1961, und zwar zu Themen der Angewandten Mechanik. 1961 wechselte H.-W. Stolle als Dozent an das damalige Mathematische Institut der Universität Rostock und wurde hier 1965 zum Professor berufen. In insgesamt über 30 Publikationen widmete er sich zunächst vor allem Problemen der Kontinuumsmechanik und der Anwendung der Mathematik in der Schiffbaumechanik, später schwerpunktmäßig den Integralgleichungen, insbesondere den singulären Integralgleichungen und ihrer numerischen Behandlung. Aus seinen Publikationen ragt das gemeinsam mit I. Fenyö geschriebene 4-bändige Werk „Theorie und Praxis der linearen Integralgleichungen“ heraus. Band 4 ist den Anwendungen gewidmet und umfaßt allein 700 Seiten. Mit dem verdienstvollen Werk werden nicht nur theoretisch interessierte Mathematiker, sondern auch Naturwissenschaftler und Techniker angesprochen. Die pädagogischen Fähigkeiten und die menschlichen Qualitäten von H.-W. Stolle zogen immer wieder Studenten an. Obwohl der Hauptakzent seiner Lehrtätigkeit auf der Mathematikausbildung der Ingenieurstudenten lag, führte er eine Vielzahl von Mathematikern zum Diplom und zur Promotion. Er war ein gefragter Gutachter und Mitglied in verschiedenen wissenschaftlichen Gremien. In den verschiedenen Leitungsämtern, in denen er tätig war (Sektionsdirektor, stellvertretender Sektionsdirektor, Wissenschaftsbereichsleiter, Vorsitzender der Bezirkssektion der Mathematischen Gesellschaft), hat er vorgelebt, wie man zu DDR-Zeiten Leitungsfunktionen ausüben konnte, ohne zweifelhafte und nicht mit dem eigenen Gewissen zu vereinbarende Kompromisse einzugehen.

Aus den voranstehenden Darlegungen wird deutlich, daß der Fachbereich Mathematik allen Anlaß hat, den Professoren L. Berg, W. Engel, G. Pazderski und H.-W. Stolle Dank zu sagen für ihre über viele Jahrzehnte hinweg erbrachten Leistungen in Lehre und Forschung für die Universität Rostock und im Dienste der Wissenschaft. Mit ihrer Berufung in den Jahren 1959 bzw. 1965 wurden seinerzeit entscheidende Akzente für das fachliche Profil der Mathematik in Rostock gesetzt. Dieses Profil hat sich bewährt, wurde seither systematisch weiterentwickelt und blieb auch über die Wende und die grundlegende strukturelle und personelle Erneuerung 1992 im wesentlichen erhalten. Durch die inzwischen erfolgten bzw. in naher Zukunft geplanten Nachfolgeberufungen ist einerseits die Kontinuität gewahrt, andererseits die Gewähr für den Erfolg in der Zukunft gegeben.

Der Fachbereich sieht sich in der Pflicht, den Kontakt zu den ausscheidenden Kollegen nicht abreißen zu lassen. Ihr Rat und ihre Hilfe werden weiterhin gefragt und geschätzt bleiben. Wir wünschen den Professoren Lothar Berg, Wolfgang Engel, Gerhard Pazderski und Hans-Wolfgang Stolle für den bereits angetretenen bzw. bevorstehenden Ruhestand Gesundheit und persönliches Wohlergehen.

Prof. Dr. G. Wildenhain
Sprecher des Fachbereichs Mathematik

Rostock. Math. Kolloq. **49**, 11–24 (1995)

Subject Classification (AMS)
47H10

Manfred Krüppel

Ein asymptotischer Fixpunktsatz für Lipschitz-stetige Operatoren in uniform konvexen Banach-Räumen

Gewidmet den Herren Professoren
L. Berg, W. Engel, G. Pazderski *und* H.-W. Stolle.

ABSTRACT. For every uniformly convex Banach space X and for every $p > 1$ there exists a constant $\gamma_p > 1$ such that holds: If $C \subset X$ is nonempty, bounded closed and convex and $T : C \to C$ is a Lipschitzian mapping such that the Lipschitzian norms $\|T^n\|$ of the iterates T^n fulfil the inequality

$$\varliminf_{n\to\infty} (\|T\|^p + \|T^2\|^p + \ldots + \|T^n\|^p)/n < \gamma_p$$

then T has a fixed point in C. In Hilbert space $\gamma_2 \geq 2$ and in L^p-space with $p \geq 2$ the constant is $\gamma_p \geq 1 + 1/(2^{p-1} - 1)$.

Es sei C eine nichtleere, beschränkte, abgeschlossene und konvexe Teilmenge eines uniform konvexen Banach-Raumes X und $T : C \to C$ eine Lipschitz-stetige Abbildung. Mit $\|T\|$ bezeichnen wir die Lipschitz-Norm von T, d.h., es ist

$$\|T\| = \sup\left(\frac{\|Tx - Ty\|}{\|x - y\|} \;:\; x \neq y\right).$$

In der vorliegenden Arbeit wird das Problem untersucht, welche Bedingungen bzgl. der Lipschitz-Normen $\|T^n\|$ die Existenz eines Fixpunktes von T in C sichern. Unter Verwendung eines neuen Konvexitätsmoduls in uniform konvexen Banach-Räumen und einer Ungleichung für Banach-Limites beweisen wir einen Fixpunktsatz, der mehrere bekannte Fixpunktaussagen als Spezialfälle enthält (vgl. F.E. Browder [3], D. Göhde [7], W.A. Kirk [10], K. Goebel und W.A. Kirk [5], J.B. Baillon [1] und M. Krüppel [13]).

1 Der Konvexitätsmodul $d_p(\varepsilon)$

Der Banach-Raum X heißt uniform konvex, wenn zu jedem ε mit $0 < \varepsilon \leq 2$ ein $\delta(\varepsilon) > 0$ existiert, so daß für $\|x\| \leq 1, \|y\| \leq 1$ und $\|x - y\| \geq \varepsilon$

$$\left\| \frac{x+y}{2} \right\| \leq 1 - \delta(\varepsilon)$$

gilt (vgl. G. Köthe [11], S. 353). Beispiele für uniform konvexe Banach-Räume sind die Hilbert-Räume sowie die Räume l^p und L^p für $1 < p < \infty$. Da jeder uniform konvexe Banach-Raum reflexiv ist, sind z.B. der Raum der summierbaren Funktionen L^1 und der Raum der stetigen Funktionen $C[0,1]$ nicht uniform konvex (vgl. G. Köthe [11] und E. Zeidler [17]).

In [13] wurde folgendes gezeigt: Ist X ein uniform konvexer Banach-Raum und p eine reelle Zahl mit $1 < p < \infty$, dann hat die für $0 < \varepsilon \leq 2$ definierte Funktion

$$\delta_p(\varepsilon) = \inf \left(\frac{\|x\|^p + \|y\|^p}{2} - \left\| \frac{x+y}{2} \right\|^p \ : \ \|x\| \leq 1, \ \|y\| \leq 1, \ \|x - y\| \leq \varepsilon \right)$$

die folgenden Eigenschaften:

(i) $\delta_p(\varepsilon) > 0$ für $\varepsilon > 0$,

(ii) $\delta_p(\lambda\varepsilon) \leq \lambda^p \delta_p(\varepsilon)$ für $0 \leq \lambda \leq 1$.

(iii) Für beliebige $x, y \in X$ mit $\|x\| \leq 1, \|y\| \leq 1$ gilt

$$\left\| \frac{x+y}{2} \right\|^p \leq \frac{\|x\|^p + \|y\|^p}{2} - \delta_p(\|x - y\|).$$

Weiter gilt (vgl. [13]): Für beliebige $x, y \in X$mit $\|x\| \leq d$, $\|y\| \leq d \, (d > 0)$ und $0 \leq t \leq 1$ ist

$$\|tx + (1-t)y\|^p \leq t\|x\|^p + (1-t)\|y\|^p - 2 \sum_{n=0}^{\infty} \varphi(2^n t) \delta_p \left(\frac{\|x - y\|}{2^n d} \right) d,$$

wobei $\varphi(t)$ gleich dem Abstand des Punktes t vom nächsten ganzzahligen Punkt ist (Sägezahnkurve).

Bemerkung 1 Auf Grund der Parallelogrammgleichung erhält man für den Hilbert-Raum $\delta_2(\varepsilon) = (\varepsilon/2)^2$. In den Räumen l^p und L^p gilt (vgl. [13]):

$$\delta_p(\varepsilon) \geq (2^{p-1} - 1)(\varepsilon/2)^q \quad \text{für } 1 < p \leq 2$$

und

$$\delta_p(\varepsilon) \geq (\varepsilon/2)^p \quad \text{für } 2 \leq p < \infty,$$

wobei q der zu p konjugierte Exponent ist.

Definition 1 *Es sei p eine reelle Zahl mit* $1 < p < \infty$. *Für* $0 < \varepsilon \leq 2$ *definieren wir den Konvexitätsmodul* $d_p(\varepsilon)$ *durch*

$$d_p(\varepsilon) = \sup\{k(\varepsilon) : k(\varepsilon) \textit{ ist konvex und } k(\varepsilon) \leq \delta_p(\varepsilon^{1/p})\}.$$

Auf Grund der Eigenschaften von $\delta_p(\varepsilon)$ und der Tatsache, daß das Supremum einer Menge konvexer Funktionen auch eine konvexe Funktion ist, ergeben sich für den Konvexitätsmodul $d_p(\varepsilon)$ die folgenden Eigenschaften:

(i) $d_p(\varepsilon)$ ist konvex, streng monoton wachsend und stetig für $0 < \varepsilon \leq 2^p$,

(ii) $0 < d_p(\varepsilon) \leq \delta_p(\varepsilon^{1/p})$,

(iii) $d_p(\lambda\varepsilon) \leq \lambda d_p(\varepsilon)$ für $0 \leq \lambda \leq 1$.

(iv) Ist $\|x\| \leq d, \|y\| \leq d$ mit $d > 0$ und $0 \leq t \leq 1$, dann gilt

$$\|tx + (1-t)y\|^p \leq t\|x\|^p + (1-t)\|y\|^p - 2\sum_{n=0}^{\infty} \varphi(2^n t) d_p\left(\frac{\|x-y\|^p}{2^{np}d^p}\right) d^p.$$

Bemerkung 2 Ist $\delta_p(\varepsilon^{1/p})$ als Funktion von ε konvex, dann ist $d_p(\varepsilon) = \delta_p(\varepsilon^{1/p})$. Dies ist im Hilbert-Raum der Fall. Daher gilt im Hilbert-Raum $d_2(\varepsilon) = \varepsilon/4$. In den Räumen l^p und L^p erhalten wir (vgl. Bemerkung 1)

$$d_p(\varepsilon) \geq \frac{2^{p-1}-1}{2^q}\varepsilon^{q/p} \quad \text{für } 1 < p < 2$$

und

$$d_p(\varepsilon) \geq \frac{\varepsilon}{2^p} \quad \text{für } 2 \leq p < \infty,$$

wobei q wieder den zu p konjugierten Exponenten bezeichnet.

Bemerkung 3 Ist $d(t)$ für $t > 0$ konvex, dann ist auch $f(t) = td(1/t)$ konvex.
Hierzu ist für beliebige $t_1, t_2 > 0$ zu zeigen:

$$f\left(\frac{t_1+t_2}{2}\right) \leq \frac{f(t_1)+f(t_2)}{2}.$$

Da $d(t)$ konvex ist, gilt

$$d(tu + (1-t)v) \leq td(u) + (1-t)d(v)$$

für beliebige $u, v > 0$ und $0 < t < 1$. Setzen wir

$$t = \frac{t_1}{t_1+t_2}, \; u = \frac{1}{t_1}, \; v = \frac{1}{t_2},$$

so erhalten wir

$$d\left(\frac{2}{t_1+t_2}\right) \leq \frac{t_1}{t_1+t_2}\, d\left(\frac{1}{t_1}\right) + \frac{t_2}{t_1+t_2}\, d\left(\frac{1}{t_2}\right)$$

und somit

$$f\left(\frac{t_1+t_2}{2}\right) = \frac{t_1+t_2}{2}\, d\left(\frac{2}{t_1+t_2}\right) \leq \frac{t_1 d\left(\frac{1}{t_1}\right) + t_2 d\left(\frac{1}{t_2}\right)}{2} = \frac{1}{2}\,[f(t_1)+f(t_2)],$$

d. h., $f(t)$ ist eine konvexe Funktion.

2 Der Banach-Limes

Zum Beweis des Fixpunktsatzes benötigen wir den Begriff des Banach-Limes (vgl. S. Banach [2], L.W. Kantorowitsch und G.P. Akilow [9], M. Krüppel [12], [14] und G.G. Lorentz [15]). Ein Banach-Limes ist ein auf dem Raum der beschränkten reellen Zahlenfolgen (x_n) definiertes lineares Funktional, das mit LIM x_n bezeichnet wird und die folgenden Eigenschaften besitzt:

1°. LIM $x_n \geq 0,$ falls $x_n \geq 0$ für alle n.

2°. LIM $(ax_n + by_n) = a\,\text{LIM}\,x_n + b\,\text{LIM}\,y_n$

3°. LIM $x_{n+1} = \text{LIM}\,x_n$

4°. LIM $x_n = 1,$ falls $x_n = 1$ für alle n.

Als Folgerung aus den Eigenschaften 1° bis 4° ergibt sich die weitere Eigenschaft (vgl. [14])

5°. LIM $(x_n y_n) = x_0\,\text{LIM}\ y_n,$ falls $\lim\limits_{n\to\infty} x_n = x_0$ ist .

Zum Beweis des Fixpunktsatzes im nächsten Abschnitt benötigen wir die folgende Ungleichung für Banach-Limites (vgl. Krüppel [14]): Ist $g(t)$ eine konvexe Funktion, dann gilt für jeden Banach-Limes

$$\text{LIM}\ g(x_n) \geq g(\text{LIM}\ x_n). \tag{1}$$

Diese Ungleichung steht in enger Beziehung zur Jensenchen Ungleichung für konvexe Funktionen (vgl. Natanson [16]).

3 Asymptotische Fixpunktsätze

Satz 1 *Es sei C eine nichtleere, beschränkte abgeschlossene und konvexe Teilmenge des uniform konvexen Banach-Raumes X und $T : C \to C$ eine Lipschitzstetige Abbildung.*

Genügt die Zahl $\gamma_p\,(p>1)$ der Gleichung

$$\frac{1}{\gamma_p}+4\sum_{n=0}^{\infty}2^n d_p\left(\frac{1}{2^{np+1}\gamma_p}\right)=1 \tag{2}$$

und gilt für ein p

$$k_p:=\varliminf_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\|T^k\|^p<\gamma_p,$$

dann hat T einen Fixpunkt in C.

Beweis: Ist $k_p<1$, dann gibt es ein k mit $\|T^k\|<1$ und T hat nach dem verallgemeinerten Fixpunktsatz von Banach einen Fixpunkt (vgl. z.B. Zeidler [17]).
Im folgenden sei $k_p\geq 1$. Den Beweis des Satzes werden wir in mehreren Schritten führen.

1) Es sei (n_i) eine solche Teilfolge der natürlichen Zahlen, daß

$$k_p=\lim_{i\to\infty}\frac{1}{n_i}\sum_{k=1}^{n_i}\|T^k\|^p \tag{3}$$

und LIM irgendein Banach-Limes. Es sei u ein festes Element aus C. Für alle x aus C definieren wir

$$r(x):=\operatorname*{LIM}_i\frac{1}{n_i}\sum_{k=1}^{n_i}\|x-T^ku\|^p. \tag{4}$$

Offensichtlich ist die Funktion $r(x)\geq 0$. Wir zeigen, daß $r(x)$ stetig und konvex ist.
Stetigkeit: Nach der Dreiecksungleichung ist

$$\|x-T^ku\|^p\leq(\|y-T^ku\|+\|x-y\|)^p.$$

Setzen wir $a=\|y-T^ku\|$ und $b=\|x-y\|$, dann gilt nach dem Mittelwertsatz der Differentialrechnung: $(a+b)^p=a^p+bp\xi^{p-1}$ mit einer Zahl $\xi\in(a,a+b)$. Bezeichnet d den Durchmesser der beschränkten Menge C, dann erhalten wir wegen $\xi<a+b\leq 2d$

$$\|x-T^ku\|^p\leq\|y-T^ku\|^p+\|x-y\|)p(2d)^{p-1}.$$

Somit ist

$$\frac{1}{n_i}\sum_{k=1}^{n_i}\|x-T^ku\|^p\leq\frac{1}{n_i}\sum_{k=1}^{n_i}\|y-T^ku\|^p+\|x-y\|p(2d)^{p-1}.$$

Auf Grund der Eigenschaften 1° und 2° eines Banach-Limes erhalten wir:

$$r(x)\leq r(y)+\|x-y\|\,p(2d)^{p-1}.$$

Vertauschen wir x und y, so folgt

$$|r(x) - r(y)| \leq p(2d)^{p-1}\|x - y\|,$$

d.h., die Funktion $r(x)$ ist auf C sogar Lipschitz-stetig.
Konvexität: Nach der Dreiecksungleichung und der Jensenschen Summationsungleichung für konvexe Funktionen (vgl. [16], S. 344) gilt für $0 \leq t \leq 1$

$$\begin{aligned}\|tx + (1-t)y - T^k u\|^p &= \|t(x - T^k u) + (1-t)(y - T^k u)\|^p \\ &\leq (t\|x - T^k u\| + (1-t)\|y - T^k u\|)^p \\ &\leq t\|x - T^k u\|^p + (1-t)\|y - T^k u\|^p.\end{aligned}$$

Durch Summation bekommen wir

$$\frac{1}{n_i}\sum_{k=1}^{n_i}\|tx + (1-t)y - T^k u\|^p \leq t\frac{1}{n_i}\sum_{k=1}^{n_i}\|x - T^k u\|^p + (1-t)\frac{1}{n_i}\sum_{k=1}^{n_i}\|y - T^k u\|^p.$$

Nach den Eigenschaften 1° und 2° eines Banach-Limes ergibt sich

$$r(tx + (1-t)y) \leq tr(x) + (1-t)r(y).$$

Die Funktion $r(x)$ ist also nichtnegativ, stetig und konvex. Auf Grund der Reflexivität des uniform konvexen Banach-Raumes X existiert daher ein $z = z(u) \in C$ mit $r(z) \leq r(x)$ für alle $x \in C$ (vgl. etwa E. Zeidler [17], S. 94).

2) Behauptung: Ist $r(z) = 0$, dann gilt $Tz = z$.

2.1) Zunächst zeigen wir: Zu $\varepsilon > 0$ gibt es eine natürliche Zahl n, so daß sowohl $\|z - T^n u\| < \varepsilon$ als auch $\|z - T^{n+1}u\| < \varepsilon$ gilt. Ist dies nämlich nicht der Fall, dann enthält die Menge $M = \{n : \|z - T^n u\| \geq \varepsilon\}$ von zwei aufeinanderfolgenden natürlichen Zahlen mindestens eine Zahl. Folglich gilt für alle n

$$\frac{1}{n}\sum_{k=1}^{n}\|z - T^k u\|^p \geq \frac{1}{n}\frac{n-1}{2}\varepsilon^p.$$

Wir erhalten daher

$$\operatorname*{LIM}_i \frac{1}{n_i}\sum_{k=1}^{n_i}\|z - T^k u\|^p \geq \varliminf_{i\to\infty}\frac{1}{n_i}\sum_{k=1}^{n_i}\|z - T^k u\|^p \geq \frac{\varepsilon^p}{2}.$$

Dies steht im Widerspruch zu $r(z) = 0$.

2.2) Sei (ε_k) eine streng monoton fallende Nullfolge. Nach 2.1) gibt es zu ε_k ein n_k derart, daß

$$\|z - T^{n_k}u\| < \varepsilon_k \quad \text{und} \quad \|z - T^{n_k+1}u\| < \varepsilon_k$$

gilt. Wegen $\varepsilon_k \to 0$ folgt

$$\lim_{k\to\infty} T^{n_k}u = z \quad \text{und} \quad \lim_{k\to\infty} T^{n_k+1}u = z.$$

Auf Grund der Stetigkeit von T ist

$$z = \lim_{k\to\infty} T^{n_k+1}u \;=\; \lim_{k\to\infty} T(T^{n_k}u) = Tz.$$

Folglich gilt $Tz = z$.

3) Im folgenden sei $r(z) > 0$. Wir zeigen, daß für alle $x \in C$ gilt:

$$\frac{r(z)}{r(x)} + 4\sum_{n=0}^{\infty} 2^n d_p\left(\frac{\|z-x\|^p}{2^{np+1}r(x)}\right) \le 1. \tag{5}$$

Zum Nachweis gehen wir von der Eigenschaft (iv) der Funktion $d_p(\varepsilon)$ aus. Danach gilt für jedes $x \in C$ und $t \in (0,1)$ die Ungleichung

$$\|tx + (1-t)z - T^k u\|^p \le t\|x - T^k u\|^p + (1-t)\|z - T^k u\|^p - 2\sum_{n=0}^{m} \varphi(2^n t) d_p\left(\frac{\|z-x\|^p}{2^{np}m_k}\right) m_k,$$

wobei m eine beliebige natürliche Zahl ist und m_k den Bedingungen $\|x - T^k u\|^p \le m_k$ und $\|z - T^k u\|^p \le m_k$ genügt. Dies ist sicher erfüllt für $m_k = c(\|x - T^k u\|^p + \|z - T^k u\|^p)$ mit $c = 2r(x)/(r(x) + r(z))$, wobei wir beachten, daß $c \ge 1$ ist wegen $r(x) \ge r(z)$. Summieren wir von $k = 1$ bis $k = n_i$, dividieren durch n_i und bilden dann den Banach-Limes bzgl. i, so erhalten wir unter Berücksichtigung der Eigenschaften 1° und 2° eines Banach-Limes

$$r(tx + (1-t)z) \le tr(x) + (1-t)r(z) - \operatorname*{LIM}_i \frac{2}{n_i}\sum_{k=1}^{n_i}\sum_{n=0}^{m} \varphi(2^n t) d_p\left(\frac{\|z-x\|^p}{2^{np}m_k}\right) m_k.$$

Vertauschen wir die Reihenfolge der Summationen im letzten Term, dann erhalten wir wegen der Eigenschaft 2° eines Banach-Limes

$$\operatorname*{LIM}_i \frac{2}{n_i}\sum_{k=1}^{n_i}\sum_{n=0}^{m} \varphi(2^n t) d_p\left(\frac{\|z-x\|^p}{2^{np}m_k}\right) m_k = \sum_{n=0}^{m} \varphi(2^n t) \operatorname*{LIM}_i \frac{2}{n_i}\sum_{k=1}^{n_i} d_p\left(\frac{\|z-x\|^p}{2^{np}m_k}\right) m_k.$$

Zur Abkürzung setzen wir

$$M_i = \frac{1}{n_i}\sum_{k=1}^{n_i} m_k.$$

Da mit $d_p(\varepsilon)$ auch $f(\varepsilon) = d_p(c/\varepsilon)/\varepsilon$ konvex ist (vgl. Bemerkung 3 in Abschnitt 1), gilt nach der Jensenschen Summationsungleichung

$$\frac{1}{n_i}\sum_{k=1}^{n_i} d_p\left(\frac{\|z-x\|^p}{2^{np}m_k}\right) m_k \ge d_p\left(\frac{\|z-x\|^p}{2^{np}M_i}\right) M_i.$$

Nochmals wegen der Konvexität der Funktion $f(\varepsilon) = d_p(c/\varepsilon)\varepsilon$ gilt nach der Ungleichung (1) für Banach-Limites für $n = 0, 1, 2, \ldots$

$$\operatorname*{LIM}_i d_p\left(\frac{\|z-x\|^p}{2^{np}M_i}\right) M_i \geq d_p\left(\frac{\|z-x\|^p}{2^{np}\operatorname*{LIM}_i M_i}\right) \operatorname*{LIM}_i M_i.$$

Wegen $m_k = c(\|x - T^k u\|^p + \|z - T^k u\|^p)$ und der Definition von $r(x)$ ist

$$\operatorname*{LIM}_i M_i = \operatorname*{LIM}_i c\left(\frac{1}{n_i}\sum_{k=1}^{n_i}\|x - T^k u\|^p + \frac{1}{n_i}\sum_{k=1}^{n_i}\|z - T^k u\|^p\right) = c(r(x) + r(z)).$$

Wegen $c = 2r(x)/(r(x)+r(z))$ ist also LIM $M_i = 2r(x)$. Fassen wir die obigen Ungleichungen zusammen, dann erhalten wir

$$r(tx + (1-t)z) \leq tr(x) + (1-t)r(z) - 2\sum_{n=0}^{m}\varphi(2^n t)d_p\left(\frac{\|z-x\|^p}{2^{np}2r(x)}\right)2r(x).$$

Wegen $r(z) \leq r(tx + (1-t)z)$ folgt

$$tr(z) \leq tr(x) - 4\sum_{n=0}^{m}\varphi(2^n t)d_p\left(\frac{\|z-x\|^p}{2^{np+1}r(x)}\right)r(x).$$

Setzen wir $t = 2^{-m}$, so erhalten wir unter Beachtung von $\varphi(2^{n-m}) = 2^{n-m}$ für $n = 0, 1, \ldots, m$

$$2^{-m}r(z) \leq 2^{-m}r(x) - 4\sum_{n=0}^{m}2^{n-m}d_p\left(\frac{\|z-x\|^p}{2^{np+1}r(x)}\right)r(x).$$

Multiplizieren wir mit $2^m/r(x)$, dann folgt für $m \to \infty$ die Behauptung (5).

4) Für alle $x \in C$ definieren wir die Funktion

$$d(x) := \operatorname*{LIM}_i \frac{1}{n_i}\sum_{k=1}^{n_i}\|x - T^k x\|^p \tag{6}$$

und zeigen, daß die folgenden Ungleichungen gelten:

(i) $\|z - u\| \quad \leq \quad 2d^{1/p}(u)$,

(ii) $d(z) \qquad \leq \quad \alpha d(u) \qquad (\alpha < 1)$.

Zu (i): Wegen der Dreiecksungleichung und der Konvexität der Funktion t^p für $t > 0$ ist

$$\left(\frac{\|z-u\|^p}{2}\right) \leq \left(\frac{\|z - T^k u\| + \|u - T^k u\|}{2}\right)^p \leq \frac{\|z - T^k u\|^p + \|u - T^k u\|^p}{2},$$

d.h.

$$\|z-u\|^p \leq 2^{p-1}(\|z-T^k u\|^p + \|u-T^k u\|^p).$$

Durch Summation bekommen wir

$$\|z-u\|^p \leq 2^{p-1}\left[\frac{1}{n_i}\sum_{k=1}^{n_i}\|z-T^k u\|^p + \frac{1}{n_i}\sum_{k=1}^{n_i}\|u-T^k u\|^p\right].$$

Durch Anwendung des Banach-Limes erhalten wir unter Beachtung von $r(z) \leq r(u) = d(u)$

$$\|z-u\|^p \leq 2^{p-1}(r(z)+d(u)) \leq 2^p d(u).$$

Zu (ii): Die Abschätzung (ii) wird sich unmittelbar aus

$$\frac{1}{k_p} + 4\sum_{n=0}^{\infty} 2^n d_p\left(\frac{d(z)}{2^{np+1}k_p r(z)}\right) \leq 1 \tag{7}$$

und der Voraussetzung (2) ergeben. Zum Nachweis dieser Ungleichung zeigen wir zunächst, daß $r(T^k z) \leq \|T^k\|^p r(z)$ für jede natürliche Zahl k gilt. Für $n_i > k$ ist

$$\frac{1}{n_i}\sum_{j=1}^{n_i}\|T^k z - T^j u\|^p \leq \frac{1}{n_i}\sum_{j=1}^{k}\|T^k z - T^j u\|^p + \|T^k\|^p \frac{1}{n_i}\sum_{j=k+1}^{n_i}\|z-T^{j-k}u\|^p.$$

Für die letzte Summe können wir auch schreiben:

$$\frac{1}{n_i}\sum_{j=1}^{n_i-k}\|z-T^j u\|^p = \frac{1}{n_i}\sum_{j=1}^{n_i}\|z-T^j u\|^p - \frac{1}{n_i}\sum_{j=n_i-k+1}^{n_i}\|z-T^j u\|^p.$$

Für $i \to \infty$ streben die beiden Summen

$$\frac{1}{n_i}\sum_{j=1}^{k}\|T^k z - T^j u\|^p \quad \text{und} \quad \frac{1}{n_i}\sum_{j=n_i-k+1}^{n_i}\|z-T^j u\|^p$$

gegen 0. Durch Anwendung des Banach-Limes erhalten wir daher die Ungleichung $r(T^k z) \leq \|T^k\|^p r(z)$ (vgl. Definition (4)).

Setzen wir in (5) $x = T^k z$, so erhalten wir unter Beachtung der Monotonie von $d_p(t)$

$$1 + 4\|T^k\|^p \sum_{n=0}^{m} 2^n d_p\left(\frac{\|z-T^k z\|^p}{2^{np+1}\|T^k\|^p r(z)}\right) \leq \|T^k\|^p,$$

wobei m eine beliebige natürliche Zahl ist. Summieren wir von $k=1$ bis $k=n_i$ und dividieren durch n_i, dann erhalten wir nach Vertauschung der Reihenfolge der Summationen

$$1 + 4\sum_{n=0}^{m} 2^n \frac{1}{n_i}\sum_{k=1}^{n_i}\|T^k\|^p d_p\left(\frac{\|z-T^k z\|^p}{2^{np+1}\|T^k\|^p r(z)}\right) \leq s_i, \tag{8}$$

wobei wir zur Abkürzung

$$s_i = \frac{1}{n_i} \sum_{k=1}^{n_i} \|T^k\|^p$$

gesetzt haben. Wegen der Konvexität der Funktion $d_p(t)$ gilt wieder nach der Jensenschen Ungleichung

$$\sum_{k=1}^{n_i} \frac{\|T^k\|^p}{s_i n_i} d_p\left(\frac{\|z - T^k z\|^p}{2^{np+1}\|T^k\|^p r(z)}\right) \geq d_p\left(\sum_{k=1}^{n_i} \frac{1}{s_i n_i} \frac{\|z - T^k z\|^p}{2^{np+1} r(z)}\right).$$

Aus (8) bekommen wir daher

$$\frac{1}{s_i} + 4\sum_{n=0}^{m} 2^n d_p\left(\frac{1}{2^{np+1} r(z)} \frac{1}{s_i n_i} \sum_{k=1}^{n_i} \|z - T^k z\|^p\right) \leq 1. \tag{9}$$

Nach (3) ist $\lim s_i = k_p$. Auf Grund der Eigenschaft 5° eines Banach-Limes ist unter Beachtung von (6)

$$\underset{i}{\mathrm{LIM}} \frac{1}{s_i n_i} \sum_{k=1}^{n_i} \|z - T^k z\|^p = \frac{1}{k_p} d(z). \tag{10}$$

Wenden wir auf (9) den Banach-Limes an, so erhalten wir unter Berücksichtigung von Satz 1 ($d_p(t)$ ist ja konvex) und (10)

$$\frac{1}{k_p} + 4\sum_{n=0}^{m} 2^n d_p\left(\frac{d(z)}{2^{np+1} k_p r(z)}\right) \leq 1. \tag{11}$$

Für $m \to \infty$ folgt die Ungleichung (7). Wir betrachten nun die Funktion

$$f(t) = \frac{1}{k_p} + 4\sum_{n=0}^{\infty} 2^n d_p\left(\frac{t}{2^{np+1} k_p}\right), \quad 0 \leq t \leq 1.$$

Die Stetigkeit und die strenge Monotonie von $d_p(t)$ überträgt sich auf $f(t)$. Es ist

$$f(0) = \frac{1}{k_p} \leq 1.$$

Nach den Voraussetzungen $k_p < \gamma_p$ und (2) ist $f(1) > 1$. Somit gibt es ein $\alpha < 1$ mit $f(\alpha) = 1$. Aus (11) folgt $d(z) \leq \alpha r(z)$. Wegen $r(z) \leq r(u) = d(u)$ erhalten wir schließlich (ii).

5) Ausgehend von einem beliebigen Punkt $u_0 \in C$ definieren wir eine Folge (u_m) auf folgende Weise: $u_{m+1} = z(u_m), m = 0, 1, 2, \ldots$. Ist $r(u_m) = 0$ für irgendein m, dann ist u_m

nach Punkt 2) ein Fixpunkt von T. Ist dies nicht der Fall, dann gilt nach (i) und (ii) aus Punkt)

$$\|u_{m+1} - u_m\| \leq 2d^{1/p}(u_m) \leq 2(\alpha^{1/p})^m d^{1/p}(u_0).$$

Folglich ist (u_m) eine Cauchy-Folge und somit konvergent gegen einen Punkt $y \in C$. Es ist

$$\|y - T^k y\| \leq \|y - u_m\| + \|u_m - T^k u_m\| + \|T^k u_m - T^k y\|.$$

Wegen der Konvexität der Funktion t^p für $t > 0$ gilt nach der Jensenschen Ungleichung

$$\begin{aligned} \frac{\|y - T^k y\|^p}{3^p} &\leq \tfrac{1}{3}(\|y - u_m\|^p + \|u_m - T^k u_m\|^p + \|T^k u_m - T^k y\|^p) \\ &\leq \tfrac{1}{3}(1 + \|T^k\|^p)\|y - u_m\|^p + \tfrac{1}{3}\|u_m - T^k u_m\|^p \end{aligned}$$

Durch Summation folgt

$$\frac{1}{n_i}\sum_{k=1}^{n_i} \|y - T^k y\|^p \leq 3^{p-1}\left(1 + \frac{1}{n_i}\sum_{k=1}^{n_i} \|T^k\|^p\right)\|y - z_m\|^p + \frac{3^{p-1}}{n_i}\sum_{k=1}^{n_i} \|u_m - T^k u_m\|^p.$$

Durch Anwendung des Banach-Limes erhalten wir

$$d(y) \leq 3^{p-1}(1 + k_p)\|y - u_m\|^p + 3^{p-1} d(u_m).$$

Für $m \to \infty$ folgt $d(y) = 0$, d.h., y ist nach Punkt 2) ein Fixpunkt von T. Damit ist der Satz vollständig bewiesen.

Satz 2 *Es sei C eine nichtleere, beschränkte, abgeschlossene und konvexe Teilmenge des uniform konvexen Banach-Raumes X und $T : C \to C$ eine Lipschitzstetige Abbildung. Genügt die Zahl γ_p $(p > 1)$ der Gleichung*

$$\frac{1}{\gamma_p} + 4\sum_{n=0}^{\infty} 2^n d_p\left(\frac{1}{2^{np+1}\gamma_p}\right) = 1$$

und gilt für ein p

$$k_p = \varliminf_{n\to\infty} \inf_m \frac{1}{n}\sum_{k=1}^{n} \|T^{m+k}\|^p < \gamma_p, \tag{12}$$

dann hat T einen Fixpunkt in C.

Der Beweis verläuft völlig analog dem des vorigen Satzes. Er unterscheidet sich nur in folgendem: Wir betrachten zwei Folgen (n_i) und (m_i) mit $n_i \to \infty$ für $i \to \infty$, so daß statt (3)

$$k_p = \lim_{i\to\infty} \frac{1}{n_i}\sum_{k=1}^{n_i} \|T^{m_i+k}\|^p$$

ist, und statt (4) setzen wir

$$r(x) = \operatorname*{LIM}_{i} \frac{1}{n} \sum_{k=1}^{n_i} \|x - T^{m_i+k}\|^p.$$

Ansonsten führen wir die gleiche Schlußweise wie beim Beweis des vorigen Satzes.

Beispiel 1 Im Hilbert-Raum ergibt die Gleichung (2) wegen $d_2(t) = t/4$ für γ_2 den Wert 2. Die hinreichende Fixpunktbedingung (12) lautet

$$\varliminf_{n\to\infty} \inf_{m} \frac{1}{n} \sum_{k=1}^{n} \|T^{m+k}\|^2 < 2.$$

Dieses Ergebnis verallgemeinert Fixpunktaussagen von J.B. Baillon [1], E.A. Lifschitz (vgl. [6], S. 36), W. J. Downing, W.O. Ray [4] und M. Krüppel [13].

Beispiel 2 Im Raum L^p mit $2 \leq p < \infty$ erhalten wir unter Beachtung von $d_p(t) \geq t/2^p$ aus (2) für γ_p die Abschätzung

$$\gamma_p \geq 1 + \frac{1}{2^{p-1} - 1}.$$

Unsere Fixpunktbedingung (12) lautet

$$\varliminf_{n\to\infty} \inf_{m} \frac{1}{n} \sum_{k_p=1}^{n} \|T^{m+k}\|^p < 1 + \frac{1}{2^{p-1} - 1}$$

(vgl. Krüppel [13]).

Literatur

[1] **Baillon, J. B. :** *Quelques aspects de la théorie des points fixés dans les espaces de Banach I.* Séminaire d'Analyse Fonctionelle de Í École Polytechnique, VII, 1978-79

[2] **Banach, S. :** *Théorie des opérations linéaires.* Warszawa 1932

[3] **Browder, F. E. :** *Nonexpansive nonlinear operators in a Banach space.* Proc. Nat. Acad. Sci. U.S.A. **54**, 1041–1044 (1965)

[4] **Downing, W. J.** und **Ray, W. O. :** *Uniformly Lipschitzian semigroups in Hilbert space.* Canad. Math. Bull. **25**, 210–214 (1982)

[5] **Goebel, K.** und **Kirk, W. A. :** *A fixed point theorem for transformations whose iterates have uniformly Lipschitz constant.* Studia Math. **17**, 135–140 (1973)

[6] **Goebel, K.** und **Reich, S. :** *Uniform Convexity, Hyperbolic Geometry and Nonexpansive Mappings.* New York 1984

[7] **Göhde, D. :** *Zum Prinzip der kontraktiven Abbildungen.* Math. Nachr. **30**, 251–258 (1965)

[8] **Goebel, K.** und **Kirk, W. A. :** *A fixed point theorem for mappings which do not increase distances.* Amer. Math. Monthly **72**, 1004–1006 (1965)

[9] **Kantorowitsch, L. W.** und **Akilow, G. P. :** *Funktionalanalysis in normierten Räumen.* Berlin 1978

[10] **Kirk, W. A. :** *Fixed point theory for nonexpansive mappings.* Proc. Conf. Sherbrooke/Canad. 1980. Lecture Notes in Math. **886**, 484–505 (1981)

[11] **Köthe, G. :** *Topologische lineare Räume.* Berlin 1960

[12] **Krüppel, M. :** *Beiträge zur Theorie der universellen Maße und Integrale.* Diss. B, Univ. Rostock 1977

[13] **Krüppel, M. :** *Ungleichungen für den asymptotischen Radius in uniform konvexen Banach-Räumen mit Anwendung in der Fixpunkttheorie.* Rostock. Math. Kolloq. **48**, 59–74 (1995)

[14] **Krüppel, M. :** *Eine Ungleichung für Banach-Limites beschränkter Zahlenfolgen.* Rostock. Math. Kolloq. **48**, 75–79 (1995)

[15] **Lorenz, G. G. :** *A contribution to the theory of divergent sequences.* Acta Math. **80**, 167–190 (1948)

[16] **Natanson, I. P. :** *Theorie der Funktionen einer reellen Veränderlichen.* Berlin 1969

[17] **Zeidler, E. :** *Vorlesungen über nichtlineare Funktionalanalysis I, -Fixpunktsätze-.* Leipzig 1976

eingegangen: 05. Juli 1994

Autor:

Prof. Dr. M. Krüppel
Universität Rostock
Fachbereich Mathematik
Universitätsplatz 1
18051 Rostock
Deutschland

Rostock. Math. Kolloq. **49**, 25–30 (1995)

Subject Classification (AMS)
20C20, 16G70

Konrad Engel; Gabriele Sauerbier

An application of Dilworth's Theorem to a problem on free Lie-algebras

Dedicated to the professors of mathematics
L. Berg, W. Engel, G. Pazderski, *and* H.- W. Stolle.

1 Introduction

A generalization of a theorem of Golod, Šafarevič [3] after an idea by Koch [5] gives the following lower bound for the number r of relations given by sums of Lie-monoms of degree m which are necessary to obtain a nilpotent Lie-algebra as a factor-algebra of a free Lie-algebra with d generators [8]:

$$\frac{(m-1)^{m-1}}{m^m} d^m < r.$$

In order to study the sharpness of this bound one has to construct nilpotent Lie-algebras with a small number of relations. Wisliceny [9] described a class of systems of relations, called increasing systems (Erhöhungssysteme), which yield nilpotent Lie-algebras. This was first done for the case $m = 2$ [9] and later this was generalized by him [10] and Sauerbier [7], [8] to any $m \geq 2$.

In the next section we will explain that increasing systems can be described in a purely combinatorial way. Thus we are led to the combinatorial problem of the minimization of the size of increasing systems for given parameters d and m. (Of course, in general there may exist other systems than increasing systems yielding nilpotent Lie-algebras, and these systems may have still smaller sizes. One of them has been constructed by Sauerbier in the case $m = 3$ [8]. Another example is known in the case $m = 2$ [11]. But we regard only increasing systems).

After some preparations we will see that we have to find the minimum number of chains in a chain decomposition of a certain interval order which equals by Dilworth's Theorem [3] the maximum size of an antichain in that order. We are able to determine this size in a short and elementary way. Thus we obtain a new short proof for the minimum size of increasing systems and a simpler formula for this number than given in [8].

2 The order-theoretic formulation

Let $\mathcal{L}(X)$ be the free LIE-algebra with the set $X = \{x_1, \dots, x_d\}, d \in \{3, 4, \dots\}$, of free generators over an arbitrary field K. We write $\mathcal{L}(X)$ as a graded algebra $\mathcal{L}(X) = \sum_{m=1}^{\infty} \mathcal{L}^m(X)$ where $\mathcal{L}^m(X)$ denotes the K-vector space which is generated by the LIE-monoms of degree m. For a subset $R \subseteq \mathcal{L}^m(X)$ let $I(R)$ be the ideal in $\mathcal{L}(X)$ generated by R and let $L_R := \mathcal{L}(X)/I(R)$ be the corresponding factor algebra. Let B be any basis of the K-vector space $\mathcal{L}(X)$ which contains LIE-monoms only. Clearly, $B^m := B \cap \mathcal{L}^m(X)$ is then a basis of $\mathcal{L}^m(X)$. Now we come to the main definition (cf. [8]). A subset R of $\mathcal{L}^m(X)$ is called an *increasing system of degree m relative to the basis B^m* if the following two conditions hold:

1. Any element of R is a finite sum of elements of B^m and the summands can be ordered in such a way that every index of a generator appearing as a factor in one summand is smaller than every index in the succeeding summand.

2. For every $b \in B^m$ there exists exactly one sum in R containing b as a summand.

As a specified basis we take the HALL-basis $B^m = H^m$ of $\mathcal{L}^m(X)$ [2]. We will not give its original definition. Instead we use the fact that there exists a bijection between H^m and the set of non-periodic cyclic words with m letters over the alphabet X [5]. To be precise we will define such words. For the sake of brevity we work only with the indices, i.e. we put $X := \{1, 2, \dots, d\}$. Let $W_{m,d}$ be the set of all words with m letters from X. For $\boldsymbol{a} = a_1 a_2 \cdots a_m \in W_{m,d}$ let $\{\boldsymbol{a}\} = \{a_1, a_2, \dots, a_m\}$. In $W_{m,d}$ we introduce an equivalence relation by $a_1 \cdots a_m \sim b_1 \cdots b_m$ iff there is some $p \in \{0, \dots, m-1\}$ such that $a_i = b_{i+p}$ holds for all i where the addition of the indices is modulo m. Clearly, for $\boldsymbol{a} \sim \boldsymbol{b}$ we have $\{\boldsymbol{a}\} = \{\boldsymbol{b}\}$. The class containing $\boldsymbol{a}$ is denoted by $[\boldsymbol{a}]$. Let $\min[\boldsymbol{a}] := \min\{l : l \in \{\boldsymbol{a}\}\}$ and $\max[\boldsymbol{a}]$ be defined analogously. Let $C_{m,d}$ be the set of all equivalence classes of size m. The elements of this set can be interpreted as the non-periodic cyclic words with m letters from X.

Given $b \in H^m$ one obtains a representative $\boldsymbol{a}$ of the bijectively associated element $[\boldsymbol{a}] \in C_{m,d}$ by omitting the brackets. Thus the bijection between $C_{m,d}$ and H^m has the property that for $[\boldsymbol{a}] \in C_{m,d}$ the corresponding basis element b in H^m has exactly $\{\boldsymbol{a}\}$ as the set of indices of the factors whose LIE-product is b. Standard application of MÖBIUS-inversion (cf. [1])

gives WITT's formula (cf. [2]):

$$\mid C_{m,d} \mid = \frac{1}{m} \sum_{t|m} \mu(t) d^{\frac{m}{t}} =: f_m(d)$$

where μ is the number theoretic MÖBIUS-function. Our set $C_{m,d}$ becomes a poset (in particular an interval order, but we do not need this fact) if we define an ordering by $[\boldsymbol{a}] < [\boldsymbol{b}]$ iff $\max[\boldsymbol{a}] < \min[\boldsymbol{b}]$.

By condition 1. of an increasing system R, every element of R corresponds to a chain in $C_{m,d}$ (the summands correspond to the elements of the chain). Moreover, condition 2. says that the whole set R corresponds to a decomposition of $C_{m,d}$ into $\mid R \mid$ chains.

Theorem *In the case of d generators, the minimum size of an increasing system of degree m relative to the* HALL*-basis H^m equals*

$$f_m(d) - f_m\left(\left\lfloor \frac{d-1}{2} \right\rfloor\right) - f_m\left(\left\lceil \frac{d-1}{2} \right\rceil\right).$$

Proof: From above we know that the minimum size of an increasing system equals the minimum number of chains in a chain decomposition of $C_{m,d}$ and by DILWORTH' Theorem [3] this equals the maximum size of an antichain in $C_{m,d}$. Thus we study in the following only the size of antichains.

For $u \in \{1, \ldots, d\}$ let

$$A_u := \{[\boldsymbol{a}] \in C_{m,d} : \min[\boldsymbol{a}] \leq u \leq \max[\boldsymbol{a}]\}.$$

Here both inequalities cannot be satisfied simultaneously with equality since we are dealing with "non-periodic words" and thus need at least two different letters. It is easy to see that A_u is an antichain in $C_{m,d}$ for each u. Moreover, if A is any antichain in $C_{m,d}$, let

$$u := \max\{\min[\boldsymbol{a}] : [\boldsymbol{a}] \in A\}.$$

Then, for all $[\boldsymbol{a}] \in A$,

$$\begin{aligned} \min[\boldsymbol{a}] &\leq u \qquad \text{(by definition of } u\text{)}, \\ \max[\boldsymbol{a}] &\geq u \qquad \text{(since } A \text{ is an antichain)}, \end{aligned}$$

thus $A \subseteq A_u$. Consequently, it is enough to look for the maximum size of the antichains $A_u, u = 1, 2, \ldots, d$. For that we introduce the following sets (omitting the parameters m and d):

$$\begin{aligned} C^{\leq}[i,j] &:= \{[\boldsymbol{a}] \in C_{m,d} : i \leq \min[\boldsymbol{a}] \quad \text{and} \quad j \geq \max[\boldsymbol{a}]\}, \\ C^{=}[i,j] &:= \{[\boldsymbol{a}] \in C_{m,d} : i = \min[\boldsymbol{a}] \quad \text{and} \quad j = \max[\boldsymbol{a}]\}. \end{aligned}$$

Since $C^{\leq}[i,j]$ is constructed using the $j-i+1$ letters $\{i, i+1, \ldots, j\}$ we have

$$\mid C^{\leq}[i,j] \mid = f_m(j-i+1). \tag{1}$$

Moreover, in view of

$$C^{=}[i,j] = C^{\leq}[i,j] \setminus \left(C^{\leq}[i+1,j] \cup C^{\leq}[i,j-1] \right)$$

and

$$C^{\leq}[i+1,j] \cap C^{\leq}[i,j-1] = C^{\leq}[i+1,j-1]$$

it holds

$$\mid C^{=}[i,j] \mid = f_m(j-i+1) - 2f_m(j-i) + f_m(j-i-1), \tag{2}$$

i. e. the size of $C^{=}[i,j]$ depends only on the difference $j-i$.
Obviously,

$$\begin{aligned} A_u \setminus A_{u-1} &= \{[\boldsymbol{a}] \in C_{m,d} : \min[\boldsymbol{a}] = u \leq \max[\boldsymbol{a}]\} \\ &= C^{=}[u,u+1] \mathbin{\dot\cup} C^{=}[u,u+2] \mathbin{\dot\cup} \ldots \mathbin{\dot\cup} C^{=}[u,d] \end{aligned}$$

and

$$\begin{aligned} A_{u-1} \setminus A_u &= \{[\boldsymbol{a}] \in C_{m,d} : \min[\boldsymbol{a}] \leq u-1 = \max[\boldsymbol{a}]\} \\ &= C^{=}[u-2,u-1] \mathbin{\dot\cup} C^{=}[u-3,u-1] \mathbin{\dot\cup} \ldots \mathbin{\dot\cup} C^{=}[1,u-1] \end{aligned}$$

where $\dot\cup$ denotes a disjoint union.
Because of (2) we have

$$\begin{aligned} \mid C^{=}[u,u+1] \mid &= \mid C^{=}[u-2,u-1] \mid, \\ \mid C^{=}[u,u+2] \mid &= \mid C^{=}[u-3,u-1] \mid, \ldots \end{aligned}$$

Moreover, in the partition of $A_u \setminus A_{u-1}$ there are $d-u$ sets, and in the partition of $A_{u-1} \setminus A_u$ there are $u-2$ sets. Thus $\mid A_u \mid \geq \mid A_{u-1} \mid$ iff $d-u \geq u-2$ iff $u \leq \dfrac{d+2}{2}$, and $\mid A_u \mid$ is maximum if $u = \left\lfloor \dfrac{d+1}{2} \right\rfloor$.
Finally,

$$A_u = C_{m,d} \setminus \left(C^{\leq}[1,u-1] \mathbin{\dot\cup} C^{\leq}[u+1,d] \right),$$

consequently, using (1),

$$\mid A_u \mid = f_m(d) - f_m(u-1) - f_m(d-u),$$

and with the optimal $u = \left\lfloor \dfrac{d+1}{2} \right\rfloor$ we obtain the formula from the theorem.

Remark: The theorem remains true if the HALL-basis H^m is replaced by any other basis B^m which contains LIE-monoms only. This follows from [7, Theorem 6].

References

[1] **Aigner, M. :** *Combinatorial Theory.* Berlin 1979

[2] **Bourbaki, N. :** *Liesche Gruppen und Liesche Algebren. (russ. Übers. aus dem Franz.)* Moskau 1976

[3] **Dilworth, R. P. :** *A decomposition theorem for partially ordered sets.* Ann. of Math. **51**, 161–166 (1950)

[4] **Golod, E. S.** and **Šafarevič, I. R. :** *Über Klassenkörpertürme (russ.).* Izv. Akad. Nauk SSSR, Ser. Mat. **28**, 261–272 (1964)

[5] **Hall, M. Jr. :** *The Theory of Groups.* New York 1959

[6] **Koch, H. :** *Erzeugenden- und Relationenrang für endlich dimensionale nilpotente Liesche Algebren.* Algebra and Logik **16**, 3, 364–374 (1977)

[7] **Sauerbier, G. :** *Untersuchungen zum Erzeugenden- und Relationenrang endlicher Pro-p-Gruppen und endlichdimensionaler nilpotenter Lie-Algebren.* Dissertation A. Güstrow 1987

[8] **Sauerbier, G. :** *Zur Konstruktion nilpotenter Lie-Algebren.* Wiss. Z. Pädagog. Hochsch. Güstrow, Math.- Nat.- Fak. **2**, 237 - 246 (1989)

[9] **Wisliceny, J. :** *Zur Darstellung von Pro-p-Gruppen und Lieschen Algebren durch Erzeugende und Relationen.* Dissertation B. Güstrow 1980

[10] **Wisliceny, J. :** *Eine Methode zur Konstruktion nilpotenter Liescher Algebren.* Math. Nachr. **118**, 209 - 214 (1984)

[11] **Newman, M. F., Sauerbier, G.** and **Wisliceny, J. :** *Groups of prime-power order with a small number of relations.* Rostock. Math. Kolloq. **49**, 141 - 154 (1995)

received: January 4, 1995 **revised:** September 11, 1995

Author:

Prof. Dr. K. Engel; Dr. G. Sauerbier
Universität Rostock
Fachbereich Mathematik
Universitätsplatz 1
18051 Rostock
Germany

Rostock. Math. Kolloq. **49**, 31–43 (1995)

Subject Classification (AMS)
62F05, 62F12

Ingo Steinke

Asymptotic optimal classification of three populations by ML-estimators

Dedicated to the professors of mathematics
L. Berg, W. Engel, G. Pazderski, *and* H.- W. Stolle.

ABSTRACT. Let π_1 and π_2 be independent populations distributed according to Q_{ϑ_1} and Q_{ϑ_2} with unknown parameters ϑ_1, $\vartheta_2 \in \mathcal{H}$ and let π_3 be a third population whose distribution Q_{ϑ_3} coincides with that of either π_1 or π_2. Assume there are given samples of size $n_1 = n_2$ and n_3, it is to be decided whether $Q_{\vartheta_3} = Q_{\vartheta_1}$ or $Q_{\vartheta_3} = Q_{\vartheta_2}$. The decision problem is described by a sequence of localized experiments. The corresponding lower Hajek-LeCam-bound is established and there is constructed a sequence of decision rules based on ML-estimators which attains this bound and is therefore considered to be asymptotic optimal.

KEY WORDS. Classification of 3 populations, Hajek-LeCam bound, weakly convergent experiments

1 Introduction

Let $\{Q_\vartheta, \vartheta \in \mathcal{H}\}$ be a family of probability measures defined on a measurable space $(\mathcal{X}, \mathfrak{A})$ and π_1, π_2 be independent populations distributed according to Q_{ϑ_1} and Q_{ϑ_2}, respectively, where the parameters ϑ_1, $\vartheta_2 \in \mathcal{H}$ are unknown. π_3 is a third population that is distributed according to either Q_{ϑ_1} or Q_{ϑ_2}. Taking samples from each population one has to decide whether π_3 has distribution Q_{ϑ_1} or Q_{ϑ_2}.

To identify the correct distribution it is natural to apply the following classificatory procedure: By means of samples of π_1, π_2 and π_3 the parameters ϑ_1, ϑ_2 and ϑ_3 are estimated. Let $\widehat{\vartheta}_i$ be an estimator for ϑ_i, $i = 1, 2, 3$. Then one decides for the population π_1 if $\widehat{\vartheta}_3$ lies closer to $\widehat{\vartheta}_1$ than to $\widehat{\vartheta}_2$ and, conversely, it is decided for population π_2 if $\widehat{\vartheta}_3$ is closer to $\widehat{\vartheta}_2$ than to $\widehat{\vartheta}_1$.

It is clear that in this approach, "good" estimators will lead to "good" classification rules. As maximum likelihood estimators (MLE) are efficient under some regularity conditions one conjectures that classification rules based on ML-estimators are best in some sense. It is the motivation of this paper to show that this natural classificatory procedure is under certain conditions an optimal one. It turns out that the necessary conditions for this optimality essentially coincide with that needed for asymptotic normality of ML-estimators. Using methods from asymptotic decision theory for sequences of localized models an asymptotic lower bound for the risk of sequences of classification rules is established. Then optimality of the classification rule based on maximum likelihood estimators is shown to be asymptotically optimal in the sense that the sequence of risks tends to the universal lower bound.

2 Notations and results

Because of asymptotic normality of maximum likelihood estimators we first consider the case that the underlying distributions are univariate normal, i.e. $\mathcal{H} = I\!R_1$ and $Q_h = N(h, \sigma^2)$, where $N(h, \sigma^2)$ represents the univariate normal distribution with mean h and variance σ^2. Let π_i, $i = 1, 2, 3$, be distributed according to $N(h_i, \sigma^2)$, where $h_3 = h_1$ or $h_3 = h_2$, and denote the sample from π_i, $i = 1, 2$, by $X_{i,1}, \dots, X_{i,m}$. Let furthermore $Y_1, \dots, Y_n$ be the sample from π_3, i.e. $\mathcal{L}(Y_i) = N(h_\tau, \sigma^2)$ and $\tau = 1$ or $\tau = 2$. For fixed σ^2 the arithmetic means $\overline{X}_{i,m} = \sum_{j=1}^m X_{i,j}/m$, $\overline{Y}_n = \sum_{j=1}^n Y_j/n$ are sufficient statistics for the distributions $\mathcal{L}(X_{i,1}, \dots, X_{i,m})$ and $\mathcal{L}(Y_1, \dots, Y_n)$, respectively, with $\mathcal{L}(\overline{X}_{i,m}) = N(h_i, \sigma^2/m)$, $\mathcal{L}(\overline{Y}_n) = N(h_\tau, \sigma^2/n)$. Therefore the decision problem may be reduced to the case of sample sizes 1 but, if $m \neq n$, unequal variances for the reference populations π_1, π_2 on one side and π_3 on the other side.

Hence we may assume in the following that there is given a sample X_1, X_2, Y of independent random variables such that $\mathcal{L}(X_i) = N(h_i, \sigma^2)$, $\mathcal{L}(Y) = N(h_\tau, \sigma_0^2)$, $i, \tau \in \{1, 2\}$, where the h_i are unknown and one has to decide whether the mean of Y corresponds to the mean of X_1 or X_2.

To perform the statistical procedure of classification a classification function is applied, as was proposed by Kudo [2], in analogy to that of randomized test functions in hypothesis testing. Let $\mathfrak{B}_k$ denote the σ-algebra of Borel sets of $I\!R_k$.

Definition 1 *A* **classification function** q (*for the classification problem discussed above*) *is a two-dimensional vector function,* $q : I\!R_3 \to I\!R_2$, $q = (q_1, q_2)$, *satisfying the following conditions:*

1. q *is* $(I\!R_3, \mathfrak{B}_3)$*-*$(I\!R_2, \mathfrak{B}_2)$*-measurable,*
2. $q_1 + q_2 = 1$ *and* $q_i \geq 0$, i=$1, 2$.

$q_i(x_1, x_2, y)$ is the probability to decide for population π_i if x_1, x_2, and y are observed from π_1, π_2, and π_3, respectively. From the decision theoretical point of view the classification problem may be stated in the following way:

Let $P_\theta = N(h_1, \sigma^2) \times N(h_2, \sigma^2) \times N(h_\tau, \sigma_0^2)$, $\theta = (h_1, h_2, \tau) \in I\!R_2 \times \{1, 2\}$, denote the common multivariate distribution of X_1, X_2, and Y. Then

$$E = (I\!R_3, \mathfrak{B}_3, P_\theta, \theta \in I\!R_3 \times \{1, 2\})$$

is called the corresponding **experiment** of the decision problem and $I\!D = \{1, 2\}$ is said to be the **decision space**. Let $\mathcal{C}(E, I\!D)$ denote the set of all classification functions for this decision problem. If we introduce the loss function, $\theta = (h_1, h_2, \tau) \in I\!R_2 \times I\!D$,

$$L(\theta, d) = L(h_1, h_2, \tau, d) = \begin{cases} 1 \ , if \ \tau \neq d \\ 0 \ , if \ \tau = d \end{cases} , \tag{1}$$

we get for the risk function of the classification rule q

$$R(\theta, q) = R(h_1, h_2, \tau, q) = \int \sum_{i=1}^{2} L(\theta, i) \, q_i \, dP_\theta = 1 - \int q_\tau \, dP_\theta \ . \tag{2}$$

By $\overline{q} = (\overline{q}_1, \overline{q}_2)$,

$$\overline{q}_1(x_1, x_2, y) = \begin{cases} 1 \ , if \ |x_1 - y| < |x_2 - y| \ , \\ \frac{1}{2} \ , if \ |x_1 - y| = |x_2 - y| \ , \\ 0 \ , if \ |x_1 - y| > |x_2 - y| \ , \end{cases} \tag{3}$$

$\overline{q}_2 = 1 - \overline{q}_1$, is denoted the so-called "next-neighbour-rule". Note that x_1, x_2, and y represent in this case the estimators for the means h_1, h_2, and h_3, respectively, of the populations. Kudo [2] proved that $\overline{q}$ is the uniformly best classification function within a certain class of classification functions:

Theorem 2 (*Kudo* [2])

Let $\mathcal{C}$ be the class of classification functions q satisfying the following conditions

1. $q(x_1 + b, x_2 + b, y + b) = q(x_1, x_2, y)$ *for all* $x_1, x_2, y \in I\!R_1$,
2. $R(h_1, h_2, 1, q) = R(h_1, h_2, 2, q)$ *for all* $h_1, h_2 \in I\!R_1$,
3. $R(h_1, h_2, \tau, q)$ *depends only through* $|h_1 - h_2|$ *on* h_1 *and* h_2.

Then the "next-neighbour-rule" $\overline{q}$ belongs to $\mathcal{C}$ and is uniformly the best in $\mathcal{C}$, that means

$$R(h_1, h_2, \tau, \overline{q}) \leq R(h_1, h_2, \tau, q) \ \ \textit{for all } h_1, h_2 \in I\!R_1, \ \tau \in \{1, 2\}. \tag{4}$$

Similar conditions to that of Theorem 2 can be expressed by transformation groups. A classification function is said to satisfy the invariance condition (T), (S), or (P), respectively, iff

$$\begin{array}{ll} \text{(T)} & q_i(x_1,x_2,y) = q_i(x_1+b,x_2+b,y+b) \ \text{ for all } x_1,x_2,y,b \in I\!R_1, \\ \text{(S)} & q_i(x_1,x_2,y) = q_i(-x_1,-x_2,-y) \ \text{ for all } x_1,x_2,y \in I\!R_1, \\ \text{(P)} & q_1(x_1,x_2,y) = q_2(x_2,x_1,y) \ \text{ for all } x_1,x_2,y \in I\!R_1. \end{array}$$

Lemma 3 *If $q \in \mathcal{C}(E, I\!D)$ satisfies $(T), (S)$,and (P) then q satisfies the conditions 1.- 3. of Theorem 2 and (4) holds true.*

The proof of Lemma 3 will be given in section 3.

The uniform optimality of $\overline{q}$ with respect to the parameter space holds only in the class $\mathcal{C}$. But one can ask whether $\overline{q}$ satisfies a weaker minimax optimality in $\mathcal{C}(E, I\!D)$, the family of all classification rules on E. In order to obtain reasonable, non-trivial minimax decision rules we follow the indifference zone approach and assume a lower bound $\Delta > 0$ for the distance between the expected values of the populations π_1 and π_2. Let $c > \Delta$ be a positive real number and

$$\Theta_\Delta := \{(h_1,h_2,\tau) \in I\!R_2 \times I\!D : \ |h_1 - h_2| \geq \Delta\} \ , \tag{5}$$

$$K_c := \{(h_1,h_2,\tau) \in I\!R_2 \times I\!D : \ |h_1| \leq c, |h_2| \leq c\} \ . \tag{6}$$

Theorem 4 *Given the decision problem described above, it holds*

$$\inf_{q \in \mathcal{C}(E,I\!D)} \sup_{\theta \in \Theta_\Delta} R(\theta, q) = \sup_{\theta \in \Theta_\Delta} R(\theta, \overline{q}) = \lim_{c \uparrow \infty} \inf_{q \in \mathcal{C}(E,I\!D)} \sup_{\theta \in \Theta_\Delta \cap K_c} R(\theta, q) \ . \tag{7}$$

It should be noted that Theorem 4 may be derived from the generalized Theorem of Hunt and Stein (e.g. see Strasser [6], chapter 48). But in order to avoid several additional sophisticated notions from the theory of amenable groups, we give a direct proof in section 3.

Now we will consider the situation of more general underlying distributions. Then, in general, it is only possible to provide asymptotic results. Let $\{Q_\vartheta, \vartheta \in \mathcal{H}\}$, $\mathcal{H} \subseteq I\!R_1$, be a family of probability distributions on a measurable space $(\mathcal{X}, \mathfrak{A})$ and π_i, $i = 1, \ldots, 3$, independent populations distributed according to Q_{ϑ_i} where ϑ_i belongs to $\mathcal{H}$ and $\vartheta_3 = \vartheta_1$ or $\vartheta_3 = \vartheta_2$. Assume that there are samples $X_{n,1}^{(i)}, \ldots, X_{n,n}^{(i)}$ from π_i, $i = 1, 2$, of size n and a sample $Y_{n,1}, \ldots, Y_{n,k_n}$ from π_3 of size k_n. Then it is to be decided whether π_3 should be assigned to either π_1 or π_2. In addition, we demand that

$$\lim_{n \to \infty} \frac{k_n}{n} = \alpha > 0 \ . \tag{8}$$

For a suffiently large number of observations and a reasonable classificatory procedure it may be assumed that the corresponding error probabilities will tend to zero. Therefore, in order to measure the power of a sequence of classification rules, a localized parametrization is introduced. Let ϑ_0 belong to the interior kernel $\overset{\circ}{\mathcal{H}}$ of $\mathcal{H}$ and put $\mathcal{H}_n(\vartheta_0) = \{h \in I\!R_1 | \vartheta_0 + \frac{h}{\sqrt{n}} \in \mathcal{H}\}$, $\theta = (h_1, h_2, \tau) \in \mathcal{H}_n(\vartheta_0) \times \mathcal{H}_n(\vartheta_0) \times I\!D$ and $P_{n,\theta} = Q^n_{\vartheta_0 + h_1/\sqrt{n}} \times Q^n_{\vartheta_0 + h_2/\sqrt{n}} \times Q^{k_n}_{\vartheta_0 + h_\tau/\sqrt{n}}$. Then the experiment

$$E_n = (\Omega^{2n+k_n}, \mathfrak{A}^{2n+k_n}, P_{n,\theta}, \theta \in \mathcal{H}_n(\vartheta_0) \times \mathcal{H}_n(\vartheta_0) \times I\!D)$$

describes the distribution of the underlying decision problem. Here and in the following let $Q^n_\vartheta = Q_\vartheta \times \cdots \times Q_\vartheta$, $\Omega^n = \Omega \times \cdots \times \Omega$, and $\mathfrak{A}^n = \mathfrak{A} \otimes \cdots \otimes \mathfrak{A}$ denote the n-fold direct product of the probability measures Q_ϑ, sample sets Ω, and σ-algebras $\mathfrak{A}$, respectively. In analogy to Definition 1 a map $q_n : \Omega^{2n+k_n} \to I\!R_2$, $q_n = (q_{n,1}, q_{n,2})$, is said to be a **classification function of the experiment $\boldsymbol{E_n}$** if q_n is $\mathfrak{A}^{2n+k_n}$-$\mathfrak{B}_2$- measurable and $q_{n,i} \geq 0, i = 1, 2$, as well as $q_{n,1} + q_{n,2} = 1$. Then $\mathcal{C}(E_n, I\!D)$ denotes the set of all classification functions for the experiment E_n. Using the loss function given in (1) the risk function is defined to be

$$R_n(\theta, q_n) = R_n(h_1, h_2, \tau, q_n) = 1 - \int q_{n,\tau} \, dP_{n,\theta} \ . \tag{9}$$

Impose the following conditions on the family of probability distributions $\{Q_\vartheta, \vartheta \in \Theta\}$:

(A1) $\{Q_\vartheta, \vartheta \in \mathcal{H}\}$ is dominated by a σ-finite measure μ and there exist a version f_ϑ of the density $dQ_\vartheta/d\mu$ such that $f_\vartheta > 0$ μ- a.s. and $\partial f_\vartheta / \partial \vartheta$ exists μ- a.s in an neighbourhood $U(\vartheta_0)$ of $\vartheta_0 \in \mathcal{H}$.

(A2) For all $\vartheta' \in U(\vartheta_0)$ holds

$$0 < I(\vartheta') := \int \Big(\frac{\partial}{\partial \vartheta} \ln f_\vartheta \Big|_{\vartheta = \vartheta'} \Big)^2 f_{\vartheta'} d\mu < \infty$$

and $I(\vartheta)$ is continuous in ϑ in $U(\vartheta_0)$.

$I(\vartheta)$ is called the **Fisher-Information** of $\{Q_\vartheta, \vartheta \in \mathcal{H}\}$ in ϑ. If **(A1)** and **(A2)** are fulfilled we define

$$\begin{aligned} Z_n \;&:=\; \Big(\frac{1}{\sqrt{nI(\vartheta_0)}} \sum_{i=1}^{n} \frac{\partial}{\partial \vartheta} \ln f_\vartheta \Big|_{\vartheta = \vartheta_0} (X^{(1)}_{n,i}) \, , \; \frac{1}{\sqrt{nI(\vartheta_0)}} \sum_{i=1}^{n} \frac{\partial}{\partial \vartheta} \ln f_\vartheta \Big|_{\vartheta = \vartheta_0} (X^{(2)}_{n,i}), \\ &\qquad \frac{1}{\sqrt{n\alpha I(\vartheta_0)}} \sum_{i=1}^{k_n} \frac{\partial}{\partial \vartheta} \ln f_\vartheta \Big|_{\vartheta = \vartheta_0} (Y_{n,i}) \Big) \ . \end{aligned} \tag{10}$$

For any $n \in I\!N$ let P_n and Q_n probability measures defined on the same measurable spaces $(\Omega_n, \mathfrak{A}_n)$. Q_n is said to be **contiguous** w.r.t. P_n ($Q_n \lhd P_n$) iff for any sequence $(A_n)_{n \in I\!N}$, $A_n \in \mathfrak{A}_n$, holds:

$$\text{if} \quad P_n(A_n) \underset{n\to\infty}{\longrightarrow} 0 \ , \quad \text{then} \quad Q_n(A_n) \underset{n\to\infty}{\longrightarrow} 0 \ .$$

If $X_n : (\Omega_n, \mathfrak{A}_n) \to (I\!R_1, \mathfrak{B}_1)$ is a sequence of random variables and $Q_n \lhd P_n$ then $X_n \xrightarrow{P_n} 0$, i.e. for any $\varepsilon > 0$ holds: $\lim_{n\to\infty} P_n(|X_n| > \varepsilon) = 0$ implies $X_n \xrightarrow{Q_n} 0$.

Let $P_{n,0} = Q_{\vartheta_0}^{2n+k_n} = Q_{\vartheta_0} \times \cdots \times Q_{\vartheta_0}$ the $(2n+k_n)$-fold direct product of Q_{ϑ_0}, $\Theta = I\!R_2 \times I\!D$, $\Theta_n = \mathcal{H}_n(\vartheta_0) \times I\!D$,

$$P_\theta = N(h_1, I^{-1}(\vartheta_0)) \times N(h_2, I^{-1}(\vartheta_0)) \times N(h_\tau, \tfrac{1}{\alpha} I^{-1}(\vartheta_0)) \ , \tag{11}$$

$\theta = (h_1, h_2, \tau) \in \Theta$, and

$$E = (I\!R_3, \mathfrak{B}_3, P_\theta, \theta \in \Theta) \ .$$

Then under **(A1)** and **(A2)** the asymptotic decision theory of LeCam provides the following result:

Theorem 5 *Under* (**A1**) *and* (**A2**)*, for any sequence $(q_n)_{n\in I\!N}$ of classification functions and any subset $\mathcal{H}_0$ of Θ, it holds*

$$\liminf_{n\to\infty} \sup_{\theta\in\Theta_n\cap\mathcal{H}_0} R_n(\theta, q_n) \geq \inf_{q\in\mathcal{C}(E,D)} \sup_{\theta\in\mathcal{H}_0} R(\theta, q) \ , \tag{12}$$

where $R(\theta,q)$ and $R_n(\theta,q)$ are defined as in (2) *and* (9)*. Furthermore, for any convergent sequence $(\theta_n)_{n\in I\!N}$, $\theta_n = (h_{n,1}, h_{n,2}, \tau) \in \Theta_n$, $h_{n,1} \to h_1$, $h_{n,2} \to h_2$, $\theta = (h_1, h_2, \tau) \in I\!R_2 \times I\!D$,*

$$\mathcal{L}(Z_n | P_{n,\theta_n}) \Longrightarrow P_\theta \quad \text{and} \quad P_{n,\theta_n} \lhd P_{n,0} \tag{13}$$

Here "⇒" denotes the weak convergence of the corresponding probability measures.

For a proof of Theorem 5 we refer to Strasser [6], chapter 10 and chapter 13, dealing with asymptotic decision theory and asymptotic normality of sequences of experiments , respectively, in a more general way. Or see Witting [7], Theorem 1.182, to convince yourself that **(A1)** and **(A2)** are sufficient for the L_2-differentiablility of the family $\{Q_\vartheta, \vartheta \in \mathcal{H}\}$ and Rüschendorf [5], chapter 4, Theorem 4.2, for the asymptotic normality of the sequence of experiments E_n and (13), where the different sample sizes have to be taken into account. Finally Strasser [6], Theorem 62.5, gives (12).

Note that $\mathcal{H}_0$ is any subset of Θ and that the lower bound for a sequence of decision rules depends on the choice of $\mathcal{H}_0$. In order to construct classification rules which attain the lower bound in (12), a sequence of statistics is used whose the distributions converge weakly under $P_{n,\theta}$ to the normal distributions given in (11). Note that Z_n fulfils this condition but it still depends on ϑ_0 which is assumed to be unknown.

(12) provides a lower bound for a sequence of classification functions for the experiments E_n. A sequence $(q_n)_{n\in I\!N}$, $q_n \in \mathcal{C}(E_n, I\!D)$, of classification rules is now called to be **optimal**

w.r.t. $\mathcal{H}_0$ iff

$$\lim_{n\to\infty} \sup_{\theta\in\Theta_n\cap\mathcal{H}_0} R_n(\theta, q_n) = \inf_{q\in\mathcal{C}(E,D)} \sup_{\theta\in\Theta\cap\mathcal{H}_0} R(\theta, q)$$

attains this lower bound.

Let $\widehat{\theta}_{n,1}, \widehat{\theta}_{n,2}$, and $\widehat{\theta}_{n,3}$ be a sequences of consistent ML-estimators, based on $Q^n_{\vartheta_0}, Q^n_{\vartheta_0}$, and $Q^{k_n}_{\vartheta_0}$, respectively, for $\vartheta_0 \in \mathcal{H}$. By

$$\begin{aligned} \overline{q}_n(X^{(1)}_{n,1}, \dots, X^{(2)}_{n,1}, \dots, Y_{n,k_n}) &= \overline{q}(\widehat{\theta}_{n,1}, \widehat{\theta}_{n,2}, \widehat{\theta}_{n,3}) \\ = \ &\overline{q}(\sqrt{n}(\widehat{\theta}_{n,1} - \vartheta_0), \sqrt{n}(\widehat{\theta}_{n,2} - \vartheta_0), \sqrt{n}(\widehat{\theta}_{n,3} - \vartheta_0)) \end{aligned} \tag{14}$$

is given a sequence of decision rules for our localized problem. Let Θ_Δ and K_c be defined as in (5) and (6).

Theorem 6 *Let the family of probability measures $\{Q_\vartheta, \vartheta \in \mathcal{H}\}$ fulfil the following conditions:* (**A1**), **(A2)**, *and there is a neighbourhood $U(\vartheta_0)$ of ϑ_0, such that*

(A3) $\frac{\partial^2}{\partial\vartheta^2} f_\vartheta$ *exists and is continuous in ϑ, $\vartheta \in U(\vartheta_0)$,*

(A4) $\int \frac{\partial}{\partial\vartheta} f_\vartheta d\mu = \int \frac{\partial^2}{\partial\vartheta^2} f_\vartheta d\mu = 0 \ \forall\vartheta \in U(\vartheta_0)$.

(A5) *For any $\vartheta \in U(\vartheta_0)$ there is a $\delta_\vartheta > 0$ with $U_{\delta_\vartheta}(\vartheta) \subset U(\vartheta_0)$ and a measurable function $M_\vartheta(z)$ with $\int |M_\vartheta(X_1)|\, dQ_\vartheta < \infty$, such that*

$$\left| \frac{\partial}{\partial\vartheta} \ln f_\vartheta(z)\Big|_{\vartheta=\vartheta'} \right| \le M_\vartheta(z) \quad \textit{for all } \vartheta' \in U_{\delta_\vartheta}(\vartheta).$$

Then for consistent ML-estimators $\widehat{\theta}_{n,i}$ for ϑ_0 and $\overline{q}_n$ from (14) *holds for any $c > \Delta$*

$$\lim_{n\to\infty} \sup_{\theta\in\Theta_n\cap K_c\cap\Theta_\Delta} |R_n(\theta, \overline{q}_n) - R(\theta, \overline{q})| = 0 \tag{15}$$

and

$$\lim_{c\to\infty} \lim_{n\to\infty} \sup_{\theta\in\Theta_\Delta\cap K_c\cap\Theta_n} R_n(\theta, \overline{q}_n) = \lim_{c\to\infty} \inf_{q\in\mathcal{C}(E,D)} \sup_{\theta\in\Theta_\Delta\cap K_c} R(\theta, q). = \sup_{\theta\in\Theta_\Delta} R(\theta, \overline{q}). \tag{16}$$

(15) means that the risk function $R_n(\theta, \overline{q}_n)$ converge uniformly on the compact parameter set $K_c \cap \Theta_\Delta$ to the risk function of the next-neighbour-rule for normal distributed populations where c may be chosen arbitrarily large while (16) states the convergence of the maximal risks of $\overline{q}_n$ to the minimax risk of the normal distributed decision problem if c tends to infinity.

The conditions **(A1)** - **(A5)** correspond to that of Witting, Noelle [8] and guarantee that any consistent sequence of ML-estimators is even asymptotically normal distributed. In any special situation the conditions of Theorem 6 have to be established. Especially, it is easy to show that these conditions are met by an one-dimensional exponential family with natural parametrization, i.e. if $\{Q_\vartheta, \vartheta \in \mathcal{H}\}$ is dominated and there is a dominating measure μ such that its density w.r.t. μ may be represented in the form

$$\frac{dQ_\vartheta}{d\mu}(x) = \exp\{x\vartheta - K(\vartheta)\} \ .$$

3 Proofs of the auxiliary results

Proof of Lemma 3: For $\theta = (h_1, h_2, \tau) \in I\!R_2 \times I\!D$, let

$$N_\theta = N(h_1, \sigma^2) \times N(h_2, \sigma^2) \times N(h_\tau, \sigma_0^2) \ ,$$

$\theta' = (h_2, h_1, \tau)$, $\overline{\theta} = (-h_1, -h_2, \tau)$, and for any $b \in I\!R_1$: $\theta_b = (h_1 + b, h_2 + b, h_\tau + b)$. Then (T) implies

$$\begin{aligned} R(h_1+b, h_2+b, \tau, q) &= 1 - \iiint q_\tau(x_1, x_2, y) N_{\theta_b}(dx_1, dx_2, dy) \\ &= 1 - \iiint q_\tau(x_1+b, x_2+b, y+b) N_\theta(dx_1, dx_2, dy) \\ &= R(h_1, h_2, \tau, q) \end{aligned} \tag{17}$$

for all $h_1, h_2, b \in I\!R_1$ and $\tau \in I\!D$. If q fulfils condition (S), then holds

$$\begin{aligned} R(-h_1, -h_2, \tau, q) &= 1 - \iiint q_\tau(x_1, x_2, y) N_{\overline{\theta}}(dx_1, dx_2, dy) \\ &= 1 - \iiint q_\tau(-x_1, -x_2, -y) N_\theta(dx_1, dx_2, dy) = R(h_1, h_2, \tau, q), \end{aligned} \tag{18}$$

and, moreover, if q satisfies (P), then

$$\begin{aligned} R(h_2, h_1, 3-\tau, q) &= 1 - \iiint q_{3-\tau}(x_1, x_2, y) N_{\theta'}(dx_1, dx_2, dy) \\ &= 1 - \iiint q_\tau(x_2, x_1, y) N_\theta(dx_2, dx_1, dy) = R(h_1, h_2, \tau, q) \end{aligned} \tag{19}$$

for all $h_1, h_2 \in I\!R_1$, $\tau \in I\!D$.

It follows: if $q \in \mathcal{C}(E, I\!D)$ satisfies (T), (S) and (P) then holds for all $h_1, h_2 \in I\!R_1$, $\tau \in I\!D$,

$$\begin{aligned} R(h_1, h_2, 1, q) &\overset{(19)}{=} R(h_2, h_1, 2, q) \overset{(18)}{=} R(-h_2, -h_1, 2, q) \overset{(17)}{=} R(h_1, h_2, 2, q) \\ R(h_1, h_2, \tau, q) &\overset{(17)}{=} R(h_1 - h_2, 0, \tau, q) \overset{(18)}{=} R(h_2 - h_1, 0, \tau, q). \end{aligned}$$

Hence $R(h_1, h_2, \tau, q) = R(|h_1 - h_2|, 0, \tau, q)$. ■

Proof of Theorem 4: The proof is carried out into 3 steps.

1. Let first $\widetilde{q}$ denote a classification function that is invariant w.r.t. translations, i.e. it satisfies (T). We define by

$$\widehat{q}_1(x_1,x_2,y) = \frac{\widetilde{q}_1(x_1,x_2,y)+\widetilde{q}_1(-x_1,-x_2,-y)+\widetilde{q}_2(x_2,x_1,y)+\widetilde{q}_2(-x_2,-x_1,-y)}{4} \ ,$$

$\widehat{q}_2 = 1-\widehat{q}_1$, a classification function that fulfils all assumptions of Lemma 3. Therefore

$$\sup_{\theta\in\Theta_\Delta} R(\theta,\overline{q}) \le \sup_{\theta\in\Theta_\Delta} R(\theta,\widehat{q}) \le \sup_{\theta\in\Theta_\Delta} R(\theta,\widetilde{q}).$$

Hence $\overline{q}$ is minimax-optimal within all classification functions satisfying (T).

2. Let now q be any classifcation function. Then $(q_n)_{n\in I\!\!N}$,

$$q_{n,i}(x_1,x_2,y) = \frac{1}{2n}\int\limits_{-n}^{n} q_i(x_1+t,x_2+t,y+t)dt \ ,$$

is a sequence of classification functions. Let $b \in I\!\!R_1$ be fixed.

$$\begin{aligned}
&|q_{n,i}(x_1+b,x_2+b,y+b)-q_{n,i}(x_1,x_2,y)|\\
=&\left|\frac{1}{2n}\int\limits_{-n}^{n}(q_{n,i}(x_1+t+b,x_2+t+b,y+t+b)-q_{n,i}(x_1+t,x_2+t,y+t))dt\right|\\
=&\frac{1}{2n}\left|\int\limits_{n}^{n+b} q_{n,i}(x_1+t,x_2+t,y+t)dt - \int\limits_{-n}^{-n+b} q_{n,i}(x_1+t,x_2+t,y+t)dt\right|\\
=&\frac{|b|}{n}\longrightarrow 0 \quad , \ n\to\infty \ .
\end{aligned} \tag{20}$$

By Witting [7], Corollary 2.15, the class of test functions which corresponds to the class of classification functions is weakly sequential compact, i.e. for any sequence of classification functions (p_n) there is a subsequence (p_{n_k}) and a classification function p_0 such that

$$\lim_{k\to\infty}\int p_{n_k}d\mu \longrightarrow \int p_0 d\mu \tag{21}$$

for any finite measure μ on $(I\!\!R_3,\mathfrak{B}_3)$ dominated by the Lebesgue-measure λ_3. Let for any $\underline{h}=(h_1,h_2,h_3)\in I\!\!R_3$

$$N_{\underline{h}} = N(h_1,\sigma^2)\times N(h_2,\sigma^2)\times N(h_3,\sigma_0^2) \ .$$

Then there is a classification function $\check{q}$ and a subsequence (q_{n_k}) of (q_n) such that

$$\lim_{k\to\infty}\iiint q_{n_k,i}(x_1,x_2,y)N_{\underline{h}}(dx_1,dx_2,dy) = \iiint \check{q}_i(x_1,x_2,y)N_{\underline{h}}(dx_1,dx_2,dy)$$

for all $h_1, h_2, h_3 \in \mathbb{R}_1$. Therefore, by (20), (21) and Lebesgue's Theorem

$$\begin{aligned}
&\left| \iiint (\check{q}_i(x_1+b, x_2+b, y+b) - \check{q}_i(x_1, x_2, y)) N_{\underline{h}}(dx_1, dx_2, dy) \right| \\
&= \lim_{k\to\infty} \left| \iiint \left[q_{n_k,i}(x_1+b, x_2+b, y+b) - q_{n_k,i}(x_1, x_2, y) \right] N_{\underline{h}}(dx_1, dx_2, dy) \right| = 0.
\end{aligned}$$

Put $\mathcal{Q} = \{N(h_1, \sigma^2) \times N(h_2, \sigma^2) \times N(h_3, \sigma_0^2), h_1, h_2, h_3 \in \mathbb{R}_1\}$. Because the family $\mathcal{Q}$ of distributions is complete, i.e. $\int h(x)Q(dx) = 0$ for any $Q \in \mathcal{Q}$ implies $h = 0$ $\mathcal{Q}$-a.s., we have a $\mathcal{Q}$-almost-sure invariance: for all $b \in \mathbb{R}_1$

$$\check{q}(x_1, x_2, y) = \check{q}(x_1+b, x_2+b, y+b) \quad \mathcal{Q} - a.s.$$

According to Witting [7], Theorem 3.109, for $\mathcal{Q}$ is closed w.r.t. translations and $\mathcal{Q}$ is equivalent to the Lebesgue-measure λ_3 there is a classification function $\widetilde{q}_0$ such that

$$\widetilde{q}_0(x_1, x_2, y) = \check{q}(x_1, x_2, y) \quad \mathcal{Q} - a.s.$$

Consequently, for any $\theta = (h_1, h_2, \tau) \in \Theta_\Delta$

$$\begin{aligned}
R(h_1, h_2, \tau, \widetilde{q}_0) &= \lim_{k\to\infty} R(h_1, h_2, \tau, q_{n_k}) \\
&= 1 - \lim_{k\to\infty} \frac{1}{2n_k} \iiint \int_{-n_k}^{n_k} q_\tau(x_1+t, x_2+t, y+t)\, dt\, N_\theta(dx_1, dx_2, dy) \\
&= \lim_{k\to\infty} \frac{1}{2n_k} \int_{n_k}^{n_k} R(h_1+t, h_2+t, \tau) dt \le \sup_{\theta\in\Theta_\Delta} R(\theta, q).
\end{aligned}$$

Hence for any $q \in \mathcal{C}(E, D)$ there is a classification function $\widetilde{q}_0$ that is invariant w.r.t. translations such that

$$\sup_{\theta\in\Theta_\Delta} R(\theta, q) \ge \sup_{\theta\in\Theta_\Delta} R(\theta, \widetilde{q}_0)$$

and by Lemma 3 we get the left equality in (7).

3. For the second equality first note that "$\ge$" holds trivially. Now we assume that there is a sequence $(c_n)_{n\in\mathbb{N}}$, $c_n \in \mathbb{R}_1$, $c_n \uparrow \infty$, such that

$$\lim_{n\to\infty} \inf_{q\in\mathcal{C}(E,\mathbb{D})} \sup_{\theta\in\Theta_\Delta\cap K_{c_n}} R(\theta, q) < \inf_{q\in\mathcal{C}(E,\mathbb{D})} \sup_{\theta\in\Theta_\Delta} R(\theta, q) =: A\,, \tag{22}$$

i.e. there is a sequence of classification functions $(q_n)_{n\in\mathbb{N}}$ and a positive real number ε so that

$$\sup_{\theta\in\Theta_\Delta\cap K_{c_n}} R(\theta, q_n) \le A - \varepsilon\ .$$

Because $\mathcal{C}(E, I\!\!D)$ is weakly sequential compact there is a classification function $\widetilde{q}$ and a subsequence (q_{n_k}) such that

$$R(\theta, \widetilde{q}) = \lim_{k\to\infty} R(\theta, q_{n_k}) \quad \text{for all } \theta \in \Theta_\Delta.$$

And therefore for all $\theta \in \Theta_\Delta$, $c_n \uparrow \infty$,

$$R(\theta, \widetilde{q}) \le \limsup_{n\to\infty} \sup_{\theta\in\Theta_\Delta\cap K_{c_n}} R(\theta, q_n) \le A - \varepsilon$$

and, in particular, $\sup\{R(\theta, \widetilde{q}) : \theta \in \Theta_\Delta\} < A$ which contradicts the definition of A in (22).

■

Proof of Theorem 6: According to Witting, Nölle [8] the conditions **(A1)** to **(A5)** imply for any consistent sequence of maximum liklihood estimators

$$\mathcal{L}(\sqrt{n}(\widehat{\theta}_{n,i} - \vartheta_0)|Q^n_{\vartheta_0}) \Longrightarrow N(0, I^{-1}(\vartheta_0))$$

and

$$\sqrt{n}(\hat{\theta}_{n,i} - \vartheta_0) - Z_{n,i} \overset{P_{n,0}}{\longrightarrow} 0 \ , i = 1, 2.$$

Similarily, for $\widehat{\theta}_{n,3}$ which is based on a sample of size k_n

$$\mathcal{L}(\sqrt{\frac{n}{k_n}}\sqrt{k_n}(\widehat{\theta}_{n,3} - \vartheta_0)|Q^{k_n}_{\vartheta_0}) \Longrightarrow N(0, \frac{1}{\alpha}I^{-1}(\vartheta_0))$$

and

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_{n,3} - \vartheta_0) - Z_{n,3} \quad &= \quad \sqrt{\frac{n}{k_n}}\big(\sqrt{k_n}(\widehat{\theta}_{n,3} - \vartheta_0) - \frac{k_n}{\alpha n}\frac{1}{\sqrt{k_n}I(\vartheta_0)}\sum_{i=1}^{k_n}\frac{\partial}{\partial\vartheta}\ln f_\vartheta\Big|_{\vartheta=\vartheta_0}(Y_{n,i}) \\ &\overset{Q^{k_n}_{\vartheta_0}}{\longrightarrow} 0,\end{aligned}$$

i.e. for $T_n = (T_{n,1}, T_{n,2}, T_{n,3}) = (\sqrt{n}(\widehat{\theta}_{n,1} - \vartheta_0), \sqrt{n}(\widehat{\theta}_{n,2} - \vartheta_0), \sqrt{n}(\widehat{\theta}_{n,3} - \vartheta_0))$

$$Z_n - T_n \overset{P_{n,0}}{\longrightarrow} 0.$$

Let $(\theta_n)_{n\in I\!\!N}$, $\theta_n = (h_{n,1}, h_{n,2}, \tau)$, be any convergent sequence, $(h_{n,1}, h_{n,2}, \tau) \to (h_1, h_2, \tau) =: \theta \in I\!\!R_2 \times I\!\!D$ and $P_\theta = N(h_1, I^{-1}(\vartheta_0)) \times N(h_2, I^{-1}(\vartheta_0)) \times N(h_\tau, \frac{1}{\alpha}I^{-1}(\vartheta_0))$ then the contiguity of P_{n,θ_n} w.r.t. $P_{n,0}$ (according to (13)) implies

$$Z_n - T_n \overset{P_{n,\theta_n}}{\longrightarrow} 0. \tag{23}$$

Hence, by $\mathcal{L}(Z_n|\, P_{n,0}) \Rightarrow P_\theta$, (23) and Slutsky's Theorem $\mathcal{L}(T_n|P_{n,\theta_n}) \Rightarrow P_\theta$. Let $\overline{q}_i$ be defined as in (3) and $D_{\overline{q}_i}$ denote the set of all discontinuity points of $\overline{q}_i$, then holds

$$P_\theta(D_{\overline{q}_i}) = 0 \quad \text{for all } \theta \in \Theta \ . \tag{24}$$

According to Billingsley [1], that implies the weak convergence $\mathcal{L}(\overline{q}_i|\, P^{T_n}_{n,\theta_n}) \Rightarrow \mathcal{L}(\overline{q}_i|P_\theta)$ and because of the boundedness of $\overline{q}_i$, $i = 1, 2$,

$$\int \overline{q}_i(T_{n,1}, T_{n,2}, T_{n,3}) dP_{n,\theta_n} \longrightarrow \int \overline{q}_i dP_\theta \tag{25}$$

for any $\theta_n \to \theta$.

Let $\varphi^i(\theta) = \int \overline{q}_i dP_\theta$ and $\varphi^i_n(\theta) = \int q_i(\widehat{\theta}_{n,1}, \widehat{\theta}_{n,2}, \widehat{\theta}_{n,3}) dP_{n,\theta_n}$ for a compact parameter set $K \subseteq \Theta$. For $\{P_\theta, \theta \in \Theta\}$ is as a subfamily of normal multivariate distributions it is weakly continuous, i.e. for any $\theta_n \to \theta$ holds $P_{\theta_n} \Rightarrow P_\theta$, and by (24) and Billingsley [1] $\varphi^i(\theta)$ is continuous. Standard arguments from analysis now say that, if $\varphi^i(\theta)$ is a continuous, real-valued function, defined on a compact set K, and if for a sequence of real-valued functions φ^i_n, defined on the same set K, holds $\varphi^i_n(\theta_n) \to \varphi^i(\theta)$ for any $\theta_n \to \theta$ then (φ^i_n) converges even uniformly on K. Therfore by (25)

$$\lim_{n\to\infty} \sup_{\theta \in K} |\varphi^i_n(\theta) - \varphi^i(\theta)| = 0 \ ,$$

and, hence for $K_n = \Theta_n \cap K_c \cap \Theta_\Delta \uparrow K = K_c \cap \Theta_\Delta$

$$\lim_{n\to\infty} \sup_{\theta \in \Theta_n \cap K_c \cap \Theta_\Delta} |R_n(\theta, \overline{q}_n) - R(\theta, \overline{q})| = 0 \tag{26}$$

for any $c > \Delta$. Hence for $c \uparrow \infty$

$$\begin{aligned} \lim_{c\uparrow\infty} \lim_{n\to\infty} \sup_{\theta \in \Theta_\Delta \cap \Theta_n \cap K_c} R_n(\theta, \overline{q}_n) &= \lim_{c\uparrow\infty} \sup_{\theta\in\Theta_\Delta} R(\theta, \overline{q}) \\ = \sup_{\theta\in\Theta_\Delta} R(\theta, \overline{q}) &= \lim_{c\uparrow\infty} \inf_{q \in \mathcal{C}(E, I\!D)} \sup_{\theta \in \Theta_\Delta \cap \Theta_n \cap K_c} R(\theta, q) \end{aligned}$$

according to Theorem 4. ■

References

[1] **Billingsley, P. :** *Convergence of Probability Measures.* New York 1969

[2] **Kudo, A. :** *The classificatory problem viewed as a two-decision problem.* Mem. Fac. Sci. Kyushu Univ. Ser. A (Math.) **13**, 96–125 (1959)

[3] **LeCam, L. :** *Asymptotic Methods in Statistical Decision Theory.* Berlin 1986

[4] **Neuhaus, G. :** *Einige Kapitel der finiten und asymptotischen Entscheidungstheorie von LeCam.* Skripten zur Mathem. Statistik Nr.17, Hamburg 1989

[5] **Rüschendorf, L. :** Asymptotische Statistik. Skripten zur Mathem. Statistik Nr.13, Münster 1987

[6] **Strasser, H. :** *Mathematical Theory of Statistics.* Berlin 1985

[7] **Witting, H. :** *Mathematische Statistik I.* Stuttgart 1985

[8] **Witting, H.** and **Nölle, G. :** *Angewandte Mathematische Statistik.* Leipzig 1970

received: June 18, 1995

Author:

Dipl.-Math. I. Steinke
Fachbereich Mathematik
Universität Rostock
Universitätsplatz 1
18051 Rostock
Germany

Rostock. Math. Kolloq. **49**, 44–50 (1995)

Subject Classification (AMS)
51M15,12F05,00A35

Klaus–Dieter Drews

Eine vereinfachende Behandlung der Konstruierbarkeit von p^k– Ecken

Gewidmet den Herren Professoren
L. Berg, W. Engel, G. Pazderski *und* H.-W. Stolle.

Die Lösung des Konstruierbarkeitsproblem von regulären p^k–Ecken $-\ p$ hier stets ungerade Primzahl, $k = 1, 2, \dots$, Konstruktionen mit Lineal und Zirkel – läßt sich bekanntlich zusammenfassen in folgende drei Aussagen:

1. Ist ein p–Eck konstruierbar, so hat p die Form $2^m + 1$. Zur Primzahleigenschaft muß hierin $m = 2^n$, p eine FERMATsche Primzahl sein. (Bisher sind 5 solche bekannt: 3, 5, 17, 257, 65537.)

2. Hat p die Form $2^m + 1$, so ist das p–Eck konstruierbar.

3. Kein p^k–Eck mit $k > 1$ ist konstruierbar.

Eine Behandlung dieses Themas kann man nur auf der Basis gewisser Vorleistungen angehen, und es seien einige Begriffe und Tatsachen genannt, auf die wir uns berufen wollen:
Formulierung der Konstruierbarkeit mittels Einheitswurzeln; Körpererweiterungen durch irreduzible Polynome; Konstruierbarkeit genau aller Elemente aus Körpern über den rationalen Zahlen Q, die durch eine endliche Anzahl von Erweiterungen mit quadratischen Gleichungen entstehen (d. h., aller Qudratwurzelterme über Q); Zyklizität der primen Restklassengruppe mod p; EISENSTEINsches Irreduzibilitätskriterium.

Sehr sorgfältig widmen sich beispielweise [1] und [2] diesen Vorbereitungen, Bücher, die mit dem von uns beabsichtigten Abstraktionsniveau korrelieren; in [2] wird auch die obige Aussage 1 bewiesen, wir gehen auf sie hier nicht weiter ein. Bedauerlich fand ich es, als ich abseits einer Algebra–Vorlesung einen Zugang zu dem Problem benötigte, daß für die Aussagen 2 und 3 im dortigen Rahmen, unter Hinweis auf zusätzlich erforderliche Gruppen- und Körpertheorie, kein Weg angeboten wurde.

Alle mir bekannt gewordenen (modernen) Darstellungen ziehen – mit Ausnahme der ersten von zwei Varianten aus [3] - die GALOISsche Theorie zur Lösung heran. So ist es mein Bestreben, die gewünschten Aussagen auf eingangs abgesteckter Basis ohne wesentliche begriffliche Vertiefung zu erhalten. Moralische Unterstützung möchte ich mir z. B. aus [4] holen, wo es zu nahe verwandter Thematik heißt (S. 204): "Man soll ja nicht mit Kanonen nach Spatzen schießen."
Geniale Grundlagen zum Thema rühren bekanntlich von GAUSS [5] her, wir verwenden sie mit heutigen Werkzeugen. Wie in der erwähnten Variante aus [3] sind auch für nachfolgende Ausführungen die GAUSSschen Perioden wesentlich, dort in einer beanspruchenden Rechnung ("the proof is rather intricate", S. 104), wohingegen hier vergleichbare Argumente aus elememtaren Kenntnissen über Körper und Vektorräume gewonnen werden sollen.

Zu Aussage 2.

Wegen $x^p - 1 = (x-1)(x^{p-1} + x^{p-2} + \ldots + x + 1)$ enthält

$$f(x) := x^{p-1} + x^{p-2} + \ldots + x + 1$$

alle p–ten Einheitswurzeln $\neq 1$ als Nullstellen; für $\varepsilon := e^{i\frac{2\pi}{p}}$ und jede Zahl g aus $\{1, \ldots, p-1\}$ gilt somit

$$\varepsilon^p = 1,\ f(\varepsilon) = 0, \quad f(\varepsilon^g) = 0.$$

Mittels des über Q irreduziblen $f(x)$ sei der Erweiterungskörper K von Q konstruiert:

$$K = Q(\varepsilon) = \{r(\varepsilon)\ :\ r(x) \in Q[x]\ \wedge\ \text{Grad } (r) \leq p-2\} \cup \{0\}.$$

K ist ein Vektorraum der Dimension $p-1$ über Q mit der Basis

$$1,\ \varepsilon,\ \varepsilon^2,\ \ldots\ ,\ \varepsilon^{p-2}. \tag{1}$$

Zu Polynomen $r_1(x), r_2(x)$ aus $Q[x]$ vom Grad $\leq p-2$ gibt es in $Q[x]$ eindeutig bestimmte Polynome $q(x),\ r_3(x)$ vom Grad $\leq p-2$ mit

$$r_1(x) \cdot r_2(x) \ = \ q(x) \cdot f(x) + r_3(x); \tag{2}$$

die Multiplikation im Körper K verläuft (für Elemente $\neq 0$) gemäß

$$r_1(\varepsilon) \cdot r_2(\varepsilon) = r_3(\varepsilon).$$

Wir behandeln nun unter

a) Die Existenz von Teilköpern des Körpers K, sodann unter

b) deren Vektorraum–Dimensionen, die erkennen lassen, daß ε Quadratwurzelterm über Q ist.

a) Durch $\varphi_0(\varepsilon) := \varepsilon^g,\ \varphi_0(r(\varepsilon)) := r(\varphi_0(\varepsilon)),\ \varphi_0(0) := 0$
($g \in \{1, \dots, p-1\}$, über g wird unten noch geeignet verfügt) ist eine Abbildung φ_0 von K in K definiert. Insbesondere gilt $\varphi_0(a) = a$ für die Elemente a aus Q und
$\varphi_0(\varepsilon^v) = (\varphi_0(\varepsilon))^\nu$ ($\nu = 0, 1, \dots, p-2$).

Die nachfolgenden Überlegungen (für Elemente $\neq 0$) zeigen, daß φ_0 einen *Automorphismus* von K bewirkt:

$\varphi_0(r_1(\varepsilon) + r_2(\varepsilon)) = \varphi_0(r_1(\varepsilon)) + \varphi_0(r_2(\varepsilon))$ ist klar;

$\varphi_0(r_1(\varepsilon) \cdot r_2(\varepsilon)) = \varphi_0(r_3(\varepsilon)) = r_3(\varphi_0(\varepsilon)) = r_1(\varphi_0(\varepsilon)) \cdot r_2(\varphi_0(\varepsilon)) = \varphi_0(r_1(\varepsilon)) \cdot \varphi_0(r_2(\varepsilon))$

gilt nach (2) wegen $f((\varphi_0(\varepsilon)) = f(\varepsilon^g) = 0$;

$\varphi_0(r(\varepsilon)) = r(\varphi_0(\varepsilon))$ verschwindet nicht, denn $\varphi_0(\varepsilon) = \varepsilon^g$ ist nicht Nullstelle eines Polynoms vom Grad $< p-1$.

Das Element 1 der Basis (1) werde nunmehr gegen das Element $-1-\varepsilon-\varepsilon^2-\dots-\varepsilon^{p-2} = \varepsilon^{p-1}$ ausgetauscht. Außerdem sei g ein erzeugendes Element der (zyklischen !) primen Restklassengruppe mod p, so daß die Potenzen $g^0, g^1, g^2, \dots, g^{p-2}$ mod p die Menge $\{1, 2, \dots, p-1\}$ durchlaufen, und wir notieren die neue Basis $\varepsilon, \varepsilon^2, \dots, \varepsilon^{p-1}$ in der Anordnung

$$\varepsilon^{g^0},\ \varepsilon^{g^1},\ \dots,\ \varepsilon^{g^{p-3}},\ \varepsilon^{g^{p-2}}. \tag{3}$$

Für die Darstellung

$$\alpha = \sum_{\nu=0}^{p-2} c_\nu \varepsilon^{g^\nu} \quad (c_\nu \in Q) \tag{4}$$

der Elemente aus K bez. dieser Basis gilt wegen $\varepsilon + \varepsilon^2 + \dots + \varepsilon^{p-1} = -1$ speziell:

$$\alpha \in Q \iff c_0 = c_1 = \dots = c_{p-2}\ (= -\alpha). \tag{5}$$

Neben φ_0 betrachten wir weitere Automorphismen φ_i von K: Für $\alpha \in K$ sei

$$\varphi_i(\alpha) := \varphi_{i-1}(\varphi_{i-1}(\alpha)) \quad (i = 1, 2, \dots, m).$$

Wichtig sind im folgenden solche Elemente aus K, die bei φ_i auf sich abgebildet werden (*Fixelelemente*).

Für $i = 0, 1, \dots, m$ bilden die Fixelemente von φ_i jeweils einen Körper K_i.

Sind nämlich α und β bei φ_i invariant, so sind dies auch $\alpha+\beta, \alpha-\beta, \alpha\cdot\beta$ und α/β $(\beta \neq 0)$, weil φ_i ein Automorphismus ist.
Ersichtlich bestehen die Inklusionen $Q \subseteq K_{i-1} \subseteq K_i \subseteq K$ $(i \neq 0)$.

b) Bei der Anwendung von φ_0 gehen die Elemente (3) der Reihe nach über in

$$\varepsilon^{g^1},\ \varepsilon^{g^2},\ \dots,\ \varepsilon^{g^{p-2}},\ \varepsilon^{g^{p-1}} = \varepsilon^{g^0},$$

denn es ist $g^{p-1} = jp + 1$ und $\varepsilon^{jp} = (\varepsilon^p)^j = 1$ (j ganz), d.h., die Basiselemente (3) werden beim Automorphismus φ_0 zyklisch versetzt, und zwar um eine Position (um 2^0 Positionen). Dann gilt, wie man induktiv sofort sieht:

> Bei Anwendung von φ_i werden die $p-1$ Basiselemente (3) zyklisch um 2^i Positionen versetzt $(i \in \{0, 1, \dots, m\})$.

Weil aber 2^i ein Teiler von $p-1$ $(= 2^m)$ ist, geschieht diese Vertauschung auf 2^i elementefremden Zyklen (Orbits) mit jeweils 2^{m-i} Basiselementen. Die Summen der Elemente auf jedem dieser Orbits sind die GAUSS*schen Perioden*

$$\begin{array}{rcccccccl}
\eta_0 & := & \varepsilon^{g^0} & + & \varepsilon^{g^{2^i}} & + & \varepsilon^{g^{2\cdot 2^i}} & + & \dots, \\
\eta_1 & := & \varepsilon^{g^1} & + & \varepsilon^{g^{1+2^i}} & + & \varepsilon^{g^{1+2\cdot 2^i}} & + & \dots, \\
 & & \dots & & \dots & & \dots & & \\
\eta_{2^i-1} & := & \varepsilon^{g^{2^i-1}} & + & \varepsilon^{g^{2^i-1+2^i}} & + & \varepsilon^{g^{2^i-1+2\cdot 2^i}} & + & \dots
\end{array}$$

der Länge 2^{m-i}. Sie liegen (für jeweils festes i) im Körper K_i und sind offensichtlich linear unabhängig über Q.

Darüber hinaus aber gehört ein Element α genau dann zu K_i, bleibt es invariant bei φ_i, wenn es Linearkombination der η_ν ist, denn nach der oben angemerkten Wirkung von φ_i auf die Basis (3) muß für solches α in der eindeutigen Darstellung (4) immer $c_\nu = c_\mu$ sein, falls ν und μ um Vielfache von 2^i differieren, d.h.:

Für $i = 0, 1, \dots, m$ ist der Körper K_i ein Vektorraum der Dimension 2^i über Q.

Insbesondere gilt $K_0 = Q$ (vgl. auch (5)) und $K_m = K$.

Zusammen mit K_i betrachten wir jetzt K_{i-1} (für fixiertes $i > 0$). Eine Basis dieses Vektorraums der Dimension 2^{i-1} über Q bilden die 2^{i-1} GAUSSschen Perioden der Länge 2^{m-i+1}, sie seien mit

$$\zeta_0,\ \zeta_1,\ \dots\ ,\ \zeta_l$$

bezeichnet (l steht für $2^{i-1}-1$). Zu K_{i-1} gehört η_0 ausdrücklich nicht, weil sich in der Darstellung (4) für η_0 die Koeffizienten erst nach 2^i und nicht nach 2^{i-1} Gliedern wiederholen. Andererseits liegen die 2^i Elemente

$$\zeta_o,\ \zeta_1,\ \dots\ ,\ \zeta_l,\ \zeta_0\eta_0,\ \zeta_1\eta_0,\ \dots\ ,\ \zeta_l\eta_0$$

in K_i, und überdies bilden sie eine Basis von K_i über Q, denn sie sind linear unabhängig über Q: Verschwindet nämlich die Linearkombination

$$\sum_{\nu=0}^{l} d'_\nu\zeta_\nu + \left(\sum_{\nu=0}^{l} d''_\nu\zeta_\nu\right)\eta_0 \quad (d'_\nu, d''_\nu \in Q), \tag{6}$$

so gilt $\sum\limits_{\nu=0}^{l} d''_\nu\zeta_\nu = 0$ (sonst folgte $\eta_0 \in K_{i-1}$) und dann müssen alle Koeffizienten d'_ν, d''_ν verschwinden.

Die Elemente aus K_i können deshalb in der Form (6), d. h. als $\beta' + \beta''\eta_0$ ($\beta', \beta'' \in K_{i-1}$) dargestellt werden, und dies besagt:

Für $i = 1, 2, \dots, m$ ist K_i ein Vektorraum der Dimension 2 über K_{i-1}.

Jedes Element α aus K_i genügt somit einer Gleichung höchstens zweiten Grades mit Koeffizienten aus K_{i-1}, denn $1, \alpha, \alpha^2$ sind linear abhängig über K_{i-1}. Wegen $K_0 = Q$ und $\varepsilon \in K_m = Q(\varepsilon)$ erhalten wir abschließend:

Für Primzahlen p von der Form $2^m + 1$ ist die p–te Einheitswurzel ε Quadratwurzelterm über Q und folglich konstruierbar.

Hierbei sind Konstruktionen (z. B. v. d. WAERDEN [6]) als solche in der *komplexen Ebene* aufgefaßt. Aus $g^{\frac{p-1}{2}} \equiv -1 \pmod p$ (für das erzeugende Element g der primen Restklassengruppe mod p) folgt jedoch

$$\varphi_{m-1}(\varepsilon) = \varepsilon^{g^{2^{m-1}}} = \varepsilon^{g^{\frac{p-1}{2}}} = \varepsilon^{-1} = \overline{\varepsilon},$$

so daß φ_{m-1} jedes Element aus K in sein konjugiert komplexes abbildet. Die Fixelemente hiervon, der Körper K_{m-1} und seine Teilkörper, sind somit *reell*: In $K_{m-1}(= K_i)$ liegt

$$\eta_0 = \varepsilon^{g^0} + \varepsilon^{g^{2^{m-1}}} = \varepsilon + \overline{\varepsilon} = 2\cos\frac{2\pi}{p}$$

als reell aufgebauter Quadratwurzelterm über Q.

Zu Aussage 3.

Es genügt zu zeigen, daß p^2–Ecke nicht konstruierbar sind. Wegen

$$x^{p^2}-1=(x^p)^p-1=(x^p-1)((x^p)^{p-1}+(x^p)^{p-2}+\ldots+x^p+1)$$

ist eine primitive p^2–te Einheitswurzel γ Nullstelle von

$$f_1(x):=(x^p)^{p-1}+(x^p)^{p-2}+\ldots+x^p+1.$$

Wir werden $f_1(x)$ als über Q irreduzibel nachweisen. Demnach hat der kleinste Körper $Q(\gamma)$, in dem γ liegt, die Vektorraumdimension $p(p-1)$ über Q; dies ist keine Potenz von 2 und folglich γ nicht konstruierbar.
Vorausgeschickt sei, daß für $\mu=1,\ldots,p-1$ und $\kappa=1,\ldots,\mu$ stets

$$\binom{p\cdot\mu}{p\cdot\kappa}\equiv\binom{\mu}{\kappa}\pmod p$$

gilt; denn in

$$\binom{p\mu}{p\kappa}=\frac{p\mu(p\mu-1)\ldots(p\mu-p+1)\;\;p(\mu-1)\ldots\;\;p(\mu-\kappa+1)\ldots(p\mu-p\kappa+1)}{p\kappa(p\kappa-1)\ldots(p\kappa-p+1)\;\;p(\kappa-1)\ldots\;\;p\qquad\qquad\ldots 1}$$

kürze man alle Faktoren p sowie mudolo p die zwischen den Vielfachen von p stehenden Reste $p-1,p-2,\ldots,1$.

Sei nun

$$f_1(x+1)=(x+1)^{p(p-1)}+(x+1)^{p(p-2)}+\ldots+(x+1)^p+1=:\sum_{\nu=0}^{p(p-1)}b_\nu x^\nu.$$

Man sieht $b_{p(p-1)}=1$, $b_0=p$ und für die übrigen Koeffizienten

$$b_\nu=\binom{p(p-1)}{\nu}+\binom{p(p-2)}{\nu}+\ldots+\binom{p}{\nu}.$$

Gilt $\nu=p\kappa$ $(\kappa=1,\ldots,p-2)$, so folgt mit der Vorbemerkung

$$b_\nu\equiv\binom{p-1}{\kappa}+\binom{p-2}{\kappa}+\ldots+\binom{\kappa}{\kappa}=\binom{p}{\kappa+1}\equiv 0 \pmod p;$$

ist ν dagegen kein Vielfaches von p, so teilt p schon jeden Summanden von b_ν. Nach dem EISENSTEINschen Kriterium ist hiernach $f_1(x+1)$ und dann auch $f_1(x)$ irreduzibel über Q.

Literatur

[1] **Jones, A., Morris, S. A.** und **Pearson, K. A. :** *Abstract Algebra and Famous Impossibilities.* New York 1991

[2] **Gilbert, W. J. :** *Modern Algebra with Applications.* New York 1976

[3] **Hadlock, C. R. :** *Field Theory and its Classical Problems.* Washington, D.C. 1978

[4] **Laugwitz, D. :** *Unlösbarkeit geometrischer Konstruktionsaufgaben - Braucht man dazu moderne Algebra ?* In: Fuchssteiner, B. u. a. (Eds.) : *Jahrbuch Überblicke Mathematik 1976.* S. 201–204, Mannheim 1976

[5] **Gauss, C. F. :** *Disquisitiones Arithmeticae (1801).* Werke Bd.1, 2. Abdruck, Göttingen 1870

[6] **van der Waerden, B. L. :** *Algebra I.* 4. Aufl., Berlin 1955

eingegangen: 29. August 1995

Autor:

Dr. K.-D. Drews
Universität Rostock
Fachbereich Mathematik
Universitätsplatz 1
18051 Rostock
Deutschland

Rostock. Math. Kolloq. **49**, 51–56 (1995)

Subject Classification (AMS)
46B15, 54A20, 54E15

Harry Poppe

A theorem on summable families in normed groups

Dedicated to the professors of mathematics
L. Berg, W. Engel, G. Pazderski, *and* H.- W. Stolle.

ABSTRACT. In functional analysis the notion of a summable family (with sum x) is well-known. If $(x_i)_{i\in I}$ is a family of points from a normed space, (x_i) is called summable to x iff for each $\varepsilon > 0$ there exits a finite set $F_0 \subset I$ such that $\|x - \sum\limits_{i\in F} x_i\| < \varepsilon$ for each finite set F, $F_0 \subset F \subset I$. But we can interprete this definition as the convergence of a suitable net. In a normed commutative group we characterize the fact that this net is a Cauchy net.

KEY WORDS. Commutative topological group, summable family, special nets, normed (commutative) group, Cauchy nets

1 Introduction

We consider a normed space $(X, \|\cdot\|)$ and let $(x_i)_{i\in I}$ be a family of points from $X, x \in X$; the notion that " $(x_i)_{i\in I}$ is a summable family with the sum x " is well–known; one defines:

1.1 For each $\varepsilon > 0$ there exists a finite set $J_0 \subset I$ such that $\|x - \sum\limits_{i\in J} x_i\| < \varepsilon$ for each finite $J \subset I$, $J_0 \subset J$. See for instance [2], [3], [5], [7], [8].

One can prove a "Cauchy criterion", and probably at first in Hilbert spaces it has been found that for a summable family $(x_i)_{i\in I}$ (with sum x) for all but countably many $i \in I$ we have $x_i = 0$. But one can observe that definition 1.1 means that the following net (Moore–Smith–sequence) converges to x with respect to the usual notion of convergence of nets in a topological space: let $\underline{F} = \underline{F}(I)$ be the collection of all finite nonempty subsets of I and let be $x_F := \sum\limits_{i\in F} x_i$ where $F \in \underline{F}$. If $\underline{F}$ is ordered by inclusion then $\underline{F} = (\underline{F}, \subseteq)$ is a poset which is directed: for F', $F'' \in \underline{F}$, $F' \cup F'' \in \underline{F}$ and $F' \cup F''$ is an upper bound for F', F''. Hence $\underline{F}$ can serve as an index set for a net and the desired net is $(x_F)_{F\in\underline{F}}$. And of course "$x_F \to x$ in $(X, \|\cdot\|)$" is equivalent to definition 1.1. Clearly, the limit point x is unique since $(X, \|\cdot\|)$

is a Hausdorff topological space. In [2] and [8] this net is mentioned but from this net the corresponding filter is constructed and then this filter is used. But in my opinion it is more intiutive to use directly the net $(x_F)_{F\in\underline{F}}$.

Our aim in this short note is first to emphasize that we should distinguish between the notion that the net $(x_F)_{F\in\underline{F}}$ converges and the notion that this net is a Cauchy net and second to establish the precise relationship between the convergence of the net $(x_F)_{F\in\underline{F}}$, the situation when (x_F) is a Cauchy net and the fact that most members of the family $(x_i)_{i\in I}$ will vanish. The answer is given by the theorem (in section 3) and its corollary. In a second corollary we consider a short application to orthonormal families in an inner product space. Of course concerning the concrete subject of the paper both the assertions and the proofs are in some sense elementary.

2 Summable families of points in a commutative topological group

Let us start with a set X and a family $(x_i)_{i\in I}$ of points of X; we consider the index set $\underline{F}=\underline{F}(I)$ as defined in section 1. If $(X,+)$ is a group we can define $x_F=\sum\limits_{i\in F}x_i$ for each F; if $(X,+)$ is commutative then x_F is uniquely defined and $(x_F)_{F\in\underline{F}}$ is a net in $(X,+)$.

Now let $(X,+)$ be a commutative topological group and we denote by $\underline{N}(0)$ the neighbourhood filter of the zeroelement $0\in X$. Then $\underline{V}:=\{V_U|\ U\in\underline{N}(0)\}$, where $V_U=\{(x,y)\in X\times X|\ x-y\in U\}$, is a base of a natural (diagonal) uniformity for X. Hence we can use Cauchy nets in X. Let us briefly recall this notion: If $(X,\underline{V})$ is an arbitrary uniform space with diagonal structure $\underline{V}$ then a net $(x_i)_{i\in I}$ from X is called Cauchy net iff for each $V\in\underline{V}$ there exists $i_V\in I$ such that $i_1,i_2\geq i_V$ implies $(x_{i_1},x_{i_2})\in V$. But since for $\underline{V}$ the triangle inequality holds we also can define a Cauchy net by: For $V\in\underline{V}$ we find $i_V\in I$: for each $i\geq i_V$ holds: $(x_i,x_{i_V})\in V$. Cauchy nets are used for instance in [4], [6]. For the proof of the theorem which we want to state in section 3 we need some lemmas. For the statements of the lemmas we assume that $(X,+)$ is a commutative topological group, $(x_i)_{i\in I}$ is a family from X and $\underline{F}$ is defined as above. The main technical argument in the proofs of the lemmas is the simple fact: if F_1, $F_2\subset I$ and $F_1\cap F_2=\emptyset$ then $x_{F_1\cup F_2}=x_{F_1}+x_{F_2}$. For this reason we prove only the first lemma (as an example).

Lemma 1 *Equivalent are:*

(1) $(x_F)_{F\in\underline{F}}$ *is a Cauchy net.*

(2) *For each* $V\in\underline{V}$ *there exists* $F_0\in\underline{F}$ *such that for all* $F\in\underline{F}$, $F\cap F_0=\emptyset$ *implies* $(x_F,0)\in V$.

Proof: (1) $\Longrightarrow$ (2): Let $V = V_U$, $U \in \underline{N}(0)$, be a basic element from $\underline{V}$; then by (1) we find $F_0 \in \underline{F}$ such that $F \in \underline{F}$ and $F_0 \subset F$ imply $(x_F, x_{F_0}) \in V$ and hence $x_F - x_{F_0} \in U$; now let $H \in \underline{F}$ and $H \cap F_0 = \emptyset$; then since $F_0 \subset F_0 \cup H$ we get: $x_{F_0 \cup H} - x_{F_0} = (x_{F_0} + x_H) - x_{F_0} = x_H \in U$ and thus $(x_H, 0) \in V$.

(2) $\Longrightarrow$ (1): Let $V = V_U \in \underline{V}$; by (2) we find $F_0 \in \underline{F}$ such that $(x_F, 0) \in V$ for each $F \in \underline{F}$, $F \cap F_0 = \emptyset$; now we consider an arbitrary $H \in \underline{F}$ such that $F_0 \subset H$; we can assume $F_0 \neq H$; $F_0 \cap (H \backslash F_0) = \emptyset$ implies $(x_{H \backslash F_0}, 0) \in V$ and hence $x_{H \backslash F_0} - 0 \in U$; now $x_H - x_{F_0} = x_{H \backslash F_0} + x_{F_0} - x_{F_0} = x_{H \backslash F_0} \in U$ and hence $(x_H, x_{F_0}) \in V$ showing that $(x_F)_{F \in \underline{F}}$ is a Cauchy net.

Remarks:

1. This result is also proved in [2], [8], where the authors used the (Cauchy) filter corresponding to (x_F).
2. Since $i \notin F_0$ means $\{i\} \cap F_0 = \emptyset$ from 1 follows that if (x_F) is a Cauchy net then for each neighbourhood U of 0 there exists $F \in \underline{F}$ such that $x_i \in U$ for each $i \in I \backslash F$.
3. In [1] families which fulfill assertion (2) of lemma 1 are called summable, that means, $(x_i)_{i \in I}$ is summable by the definition of the author iff the net $(x_F)_{F \in \underline{F}}$ is a Cauchy net.
4. We call (x_i) a Cauchy family if (x_F) is a Cauchy net.

Lemma 2 *Let A be a nonempty subset of I such that $A \neq I$ and $x_i = 0$ for each $i \in I \backslash A$; let $\underline{A} = \{B \subset A \mid B$ finite $\}$, hence $\underline{A} \subset \underline{F}$. If $(x_F)_{F \in \underline{A}}$ is a Cauchy net then $(x_F)_{F \in \underline{F}}$ is a Cauchy net, too.*

Lemma 3 *Let I be infinite; let $A \subset I$ such that $A \neq \emptyset$, $A \neq I$ and $x_i = 0$ for each $i \in I \backslash A$; let $\underline{A}$ be defined as in lemma 2. If $(x_F)_{F \in \underline{F}}$ converges in X, $x_F \to x$, then $(x_F)_{F \in \underline{A}} \to x$.*

3 Characterizing the fact that $(x_F)_{F \in \underline{F}}$ is a Cauchy net

Here we want to consider normed groups. We define (see [8]):

3.1 Let $(X, +)$ be a group; we call $\| \cdot \| : X \to \mathbb{R}$ a norm for $(X, +)$ iff the axioms hold:

1. $\|x\| \geq 0$ for each $x \in X$,
2. $\|x\| = 0$ iff $x = 0$,

3. $\| -x\| = \|x\|$,

4. $\|x+y\| \leq \|x\| + \|y\|$.

Then as is it easy to see, $d : d(x,y) = \|x-y\|$ defines a metric for X.

Lemma 4 $(X, +, \| \cdot \|) = (X, +, d)$ *is a topological group.*

Proof: Each translation $x \to x + a$ is continuous and hence the topological space X is homogenous; now let be $f : X \to X$, $f(x) = -x$; then if (x_i) is a net from X and $x_i \to 0$ we get: $\|f(x_i) - f(0)\| = \| -x_i - 0\| = \| - x_i\| = \|x_i\| = \|x_i - 0\|$ showing that f is continuous; since convergence in $X \times X$ means coordinatewise convergence and by axiom 4. we see that the map $(x, y) \to x + y$ is continuous, too.

Now we want to state the main result of our paper.

Theorem 5 *Let* $(X, +, \| \cdot \|)$ *be a commutative normed group; let* $(x_i)_{i \in I}$ *be a family of elements of* X *and let* I *be infinite. We consider the net* $(x_F)_{F \in \underline{F}}$*, where* $\underline{F} = \underline{F}(I)$*. Then the following statements are equivalent:*

(1) (x_F) *is a Cauchy net.*

(2) (a) *There exists a countable set* $I^* \subset I$ *such that* $x_i = 0$ *for each* $i \in I \backslash I^*$*.*
(b) $(x_F)_{F \in \underline{I}^*}$ *is a Cauchy net.*

Proof: By lemma 4 we see that we can use the lemmas of section 2 for the proof.
(1) $\Longrightarrow$ (2): By lemma 1 we get: for each $n \in \mathbb{N}$, $n \geq 1$ there exists $F_n \in \underline{F}$ such that $d(x_F, 0) = \|x_F\| < \dfrac{1}{n}$ for each $F \in \underline{F}$, $F \cap F_n = \emptyset$; hence: $i \in T \backslash \bigcup\limits_{n=1}^{\infty} F_n$ implies $\{i\} \cap F_n = \emptyset$ for each n and thus $\|x_i\| = \|x_{\{i\}}\| < \dfrac{1}{n}$ for each n showing that $x_i = 0$ holds. $I^* = \bigcup\limits_{n=1}^{\infty} F_n$ is countable and so (a) holds. To prove also (b) let $\varepsilon > 0$ be given. We assume that $I^* \neq I$, otherwise there is nothing to prove. Since (x_F) is a Cauchy net we find $F_\varepsilon \in \underline{F}$: $\|x_F - x_{F_\varepsilon}\| < \varepsilon$ for each $F \in \underline{F}$, $F_\varepsilon \subset F$; now $F_\varepsilon \cap I^* = \emptyset$ or $F_\varepsilon \cap I^* \neq \emptyset$; if $F_\varepsilon \cap I^* = \emptyset$ let $F_1 \in \underline{I}^*$ be fixed and $H \in \underline{I}^*$, $F_1 \subset H$; then $F_\varepsilon \subset F_1 \cup F_\varepsilon \subset H \cup F_\varepsilon$ and hence: $\|x_H - x_{F_1}\| = \|x_{H \cup F_\varepsilon} - x_{F_1 \cup F_\varepsilon}\| = \|x_{H \cup F_\varepsilon} - x_{F_\varepsilon} + x_{F_\varepsilon} - x_{F_1 \cup F_\varepsilon}\| < 2\varepsilon$; if $F_\varepsilon \cap I^* \neq \emptyset$ let be $H \in \underline{I}^*$ and $F_\varepsilon \cap I^* \subset H$; then $F_\varepsilon \subset H \cup F_\varepsilon \backslash I^*$ implies: $\|x_H - x_{F_\varepsilon \cap I^*}\| = \|x_{H \cup (F_\varepsilon \backslash I^*)} - x_{(F_\varepsilon \cap I^*) \cup (F_\varepsilon \backslash I^*)}\| = \|x_{H \cup (F_\varepsilon \backslash I^*)} - F_\varepsilon\| < \varepsilon$.
(2) $\Longrightarrow$ (1): Setting $A = I^*$ in lemma 2 we get the proof.

Corollary 6 *Under the assumptions of the theorem it holds: If the net* $(x_F)_{F \in \underline{F}}$ *converges in* X*,* $x_F \to x$*, then we get:*

(a) *There exists a countable set* $I^* \subset I$ *such that* $x_i = 0$ *for each* $i \in I \backslash I^*$.

(b) $(x_F)_{F \in \underline{I}^*} \to x$

(c) *For each enumeration of* I^*, $I^* = \{i_0, i_1, \dots\}$ *holds that the sequence* $\left(\sum\limits_{j=0}^{k} x_{i_j} \right)_{k \in \mathbb{N}} = (x_{\{i_0, i_1, \dots, i_k\}})_{k \in \mathbb{N}}$ *converges to* x.

Proof: $x_F \to x$ implies that (x_F) is a Cauchy net and hence we get (a) by theorem 1; then we get assertion (b) by lemma 3 if we substitute $A = I^*$ in the lemma. If $I^* = \{i_0, i_1, \dots\}$ and $F_k = \{i_0, \dots, i_k\}$ for each $k \in \mathbb{N}$ then $(\{F_k\}, \subseteq)$ clearly is a cofinal subset of $\underline{I}^*$ and hence we get (c).

Remark: Assertion (a) of corollary 6 is proved in [2], [8] and is stated (without proof) in [3].

Corollary 7 *Let* $(X, (\cdot, \cdot))$ *be an inner product space and* $(x_i)_{i \in I}$ *an arbitrary orthonormal family in* X*; let* x *be an arbitrary point in* X*; then there exists a countable set* $I^* \subset I$ *such that* $(x, x_i) = 0$ *for each* $i \in I \backslash I^*$.

Proof: In [7] it is shown that $((x, x_i)x_i)_{i \in I}$ is a Cauchy family; since $(X, +, \|\cdot\|)$ is a commutative normed group by theorem 1 we find a countable $I^* \subset I$ such that $(x, x_i)x_i = 0$ for each $i \in I \backslash I^*$; but x_i being a member of an orthonormal family yields $x_i \neq 0$ and hence $(x, x_i) = 0$.

Remark: This result one finds in [5] (with another proof).

References

[1] **Banaszczyk, W. :** *Summable families in nuclear groups.* Studia Math. **105**, 271–282 (1993)

[2] **Bourbaki, N. :** *Topologie générale. Chap. III, Groups topologiques (Théorie élémentaire).* Paris 1960

[3] **Heuser, H. :** *Funktionalanalysis.* Stuttgart 1992

[4] **Kelley, J. L. :** *General Topology.* Princeton, N.J. 1957

[5] **Naylor, A. W.** and **Sell G. R. :** *Linear Operator Theory in Engineering and Science.* New York 1982

[6] **Poppe, H. :** *Compactness in General Function Spaces.* Berlin 1974

[7] **Scheja, G.** and **Storch, U. :** *Lehrbuch der Algebra. Teil* 2. Stuttgart 1988

[8] **Warner, S. :** *Topological Fields.* Amsterdam 1989

received: August 31, 1995

Author:

Prof. Dr. H. Poppe
Universität Rostock
Fachbereich Mathematik
Universitätsplatz 1
18051 Rostock
Germany

Rostock. Math. Kolloq. **49**, 57–74 (1995)

Subject Classification (AMS)
65J05, 49D20

Dieter Schott

Basic properties of Fejer monotone sequences

Dedicated to the professors of mathematics
L. Berg, W. Engel, G. Pazderski, *and* H.- W. Stolle.

Summary. *We consider certain classes of Fejer monotone sequences and their basic properties. These properties are the starting point of a convergence theory for corresponding iterative methods which are widely used to solve convex problems.*

1 Introduction

The theory of Fejer monotone sequences and mappings has a long history. It started already in the sixteeth of our century and was closely connected with the Russian mathematician Eremin ([4],[5]). The monography [6] contains a summarizing description for finite-dimensional spaces with various applications to the iterative solution of convex problems. Parallel to this development the basic ideas of Fejer theory also occured in concrete applications (e.g. [9],[1],[7],[12]). This process continues up to present time. Often the authors do not know the original literature and the developed theory (e.g.[18],[10],[8],[3]). The great interest in Fejer methods is caused by the wide range of applications. At present they play an important part in computerized tomography and image recovery (see e.g. [19]). Consequently, there are strong requirements to expand and to complete the Fejer theory. My paper [13] supplies a unifying concept for the application of Fejer methods on the level of Hilbert space assuming somewhat different conditions than in [6]. Finally I develop in [17] a theory of strong convergent Fejer methods using some basic properties of Fejer monotone sequences given here. Such sequences (x_k) can be generated by Fejer monotone mappings g ([16],[17]). The aim is to win iterative Fejer methods

$$x_{k+1} \in g(x_k)$$

which converge to an element x^* of a convex and closed problem set M.

2 Fejer monotone sequences

Let H be a Hilbert space. We consider a sequence (x_k) in H. Its set of weak and strong accumulation values is denoted by $A^w(x_k)$ and $A^s(x_k)$, respectively. As usual the number

$$\rho(y, M) = \inf\{\|y - x\| : x \in M\}$$

gives the distance between $y \in H$ and $M \subseteq H$. If M is convex and closed, then there is a unique metric (orthogonal) projector P_M onto M with

$$P_M y = \operatorname{argmin}\{\|y - x\| : x \in M\}.$$

Moreover, we have the relation $\rho(y, M) = \|y - P_M y\|$. Finally $B(x, r)$ denotes the **closed ball** with midpoint x and radius r.

Definition 2.1 *Let M be a nonempty subset of H. The sequence (x_k) is said to be* M-**Fejer-monotone** *iff*

$$\|x_{k+1} - x\| \leq \|x_k - x\| \quad \forall k \in \mathbb{N},\ \forall x \in M.$$

It is said to be **regularly** M-**Fejer monotone** *iff additionally*

$$x_{k+1} \neq x_k \quad \forall k \in \mathbb{N} \, : \, x_k \notin M.$$

It is said to be **strictly** M-**Fejer monotone** *iff additionally*

$$\|x_{k+1} - x\| < \|x_k - x\| \quad \forall k \in \mathbb{N} \, : \, x_k \notin M \, , \ \forall x \in M.$$

It is called **(regularly, strictly) Fejer monotone** *iff it is (regularly, strictly) M-Fejer monotone for any subset M. The set*

$$C(x_k) = \{x \in H \, : \, \|x_{k+1} - x\| \leq \|x_k - x\| \quad \forall k \in \mathbb{N}\}$$

is called the **Fejer carrier** *of (x_k).*

Remark 2.2 Originally strictly M-Fejer monotone sequences (x_k) with $x_k \notin M$ were called Fejer monotone (see [6, subsection 1.1]). But we want to start with the above weaker version of the concept (see also [2, p. 22],[13, p. 886]). So the distance between the iterates of a M-Fejer monotone sequence and each fixed element in M does not expand. In the strict case this distance is reduced step by step for all iterates outside of M.

The above definitions have some easy consequences. The assertions listed in the first Lemma need not be proven. They are obvious.

Lemma 2.3 *The following properties hold:*

(P1) (x_k) *Fejer monotone* $\Longleftrightarrow C(x_k) \neq \emptyset$,

(P2) (x_k) *M-Fejer monotone* $\Longrightarrow M \subseteq C(x_k)$,

(P3) (x_k) *Fejer monotone* $\Longrightarrow (x_k)$ *$C(x_k)$-Fejer monotone,*

(P4) (x_k) *M-Fejer monotone* , $\emptyset \neq N \subseteq M \Longrightarrow (x_k)$ *N-Fejer monotone,*

(P5) (x_k) *regularly M-Fejer monotone* $\Longrightarrow (x_k)$ *M-Fejer monotone,*

(P6) (x_k) *(strictly) Fejer monotone* $\Longrightarrow$ *subsequence* $(x_{k'})$ *(strictly) Fejer monotone.*

If a sequence (x_k) is M-Fejer monotone, but not strictly, it can be strictly Fejer monotone relative to a subset N. If (x_k) is strictly M-Fejer monotone, it need not be strictly $C(x_k)$-Fejer monotone (compare (P3)).

Lemma 2.4 *Let be $k_0 \in \mathbb{N}$ fixed. Then there holds*

(P7) (x_k) *M-Fejer monotone,* $x_{k_0} \in M \Longrightarrow x_{k_0+1} = x_{k_0}$,

(P8) (x_k) *regularly M-Fejer monotone,* $x_{k_0+1} = x_{k_0} \Longrightarrow x_{k_0} \in M$,

(P9) (x_k) *strictly M-Fejer monotone* $\Longrightarrow$ (x_k) *regularly M-Fejer monotone.*

Proof: Let (x_k) be M-Fejer monotone. Assuming $x_{k_0} \in M$ the property (P7) follows immediately if $k = k_0$ and $x = x_{k_0}$ are chosen in the defining inequality. Let (x_k) be regularly M-Fejer monotone. Then $x_{k_0+1} = x_{k_0}$ generates a contradiction if x_{k_0} is supposed to be not in M. So (P8) arises. If (x_k) is strictly M-Fejer monotone and x_{k_0} is not in M for any index k_0 then $\|x_{k_0+1} - x\| < \|x_{k_0} - x\|$ and consequently $x_{k_0+1} \neq x_{k_0}$. Thus (x_k) is regularly M-Fejer monotone as asserted in (P9). ■

By Lemma 2.4 the set M of a M-Fejer monotone sequence (x_k) contains at most stationary points of (x_k) . In the regular case all stationary points of (x_k) lie in M. Now we introduce the abbreviation

$$d_k(x) := \|x_k - x\|^2 - \|x_{k+1} - x\|^2 \qquad (1)$$

and list some properties of $d_k(x)$.

Lemma 2.5 *The functional $d_k(x)$ has the representations*

$$\begin{aligned} d_k(x) &= (x_k - x_{k+1}, x_k + x_{k+1} - 2x) & (2)\\ &= 2(x_k - x_{k+1}, x_k - x) - \|x_k - x_{k+1}\|^2. & (3)\end{aligned}$$

The sequence (x_k) is M-Fejer monotone iff

$$d_k(x) \geq 0 \quad \forall k \in \mathbb{N},\ \forall x \in M.$$

Besides, it is

$$C(x_k) = \{x \in H : d_k(x) \geq 0 \quad \forall k \in \mathbb{N}\}.$$

Proof: The transformation of (1) supplies

$$\begin{aligned} d_k(x) &= \|x_k\|^2 - \|x_{k+1}\|^2 - 2(x_k, x) + 2(x_{k+1}, x)\\ &= (x_k - x_{k+1}, x_k + x_{k+1}) - 2(x_k - x_{k+1}, x)\\ &= (x_k - x_{k+1}, x_k + x_{k+1} - 2x)\\ &= (x_k - x_{k+1}, 2(x_k - x) - (x_k - x_{k+1}))\\ &= 2(x_k - x_{k+1}, x_k - x) - \|x_k - x_{k+1}\|^2.\end{aligned}$$

The next assertions are immediate consequences of Fejer monotony (see Definition 2.1). ■

We turn to a remarkable property of the set $C(x_k)$.

Theorem 2.6 *The Fejer carrier $C(x_k)$ of (x_k) is convex and closed.*

Proof: By Lemma 2.5 the Fejer carrier of (x_k) can be represented in the form

$$C(x_k) = \{x \in H : (x_k - x_{k+1}, x_k + x_{k+1} - 2x) \geq 0 \quad \forall k \in \mathbb{N}\}.$$

Now we fix arbitrary elements $x, x' \in C(x_k)$ and arbitrary numbers α, α' satisfying $\alpha \geq 0,\ \alpha' \geq 0$, $\alpha + \alpha' = 1$. Using the abbreviations $u_k = x_k - x_{k+1}$ and $v_k = x_k + x_{k+1}$ we get at first $\alpha(u_k, v_k - 2x) \geq 0,\ \alpha'(u_k, v_k - 2x') \geq 0$ and then by addition of both inequalities

$$(u_k, (\alpha + \alpha')v_k - 2\alpha x - 2\alpha' x') = (u_k, v_k - 2(\alpha x + \alpha' x')) \geq 0.$$

This means $\alpha x + \alpha' x' \in C(x_k)$. Hence $C(x_k)$ is convex. Finally we consider a sequence (y_n) in $C(x_k)$ converging to $y \in H$. Consequently we conclude with the above abbreviations

$$(u_k, v_k - 2y_n) \geq 0 \quad \forall k, n \in \mathbb{N}$$

and, since the scalar product is continuous, also

$$(u_k, v_k - 2y) \geq 0 \quad \forall k \in \mathbb{N}.$$

This means $y \in C(x_k)$. Hence $C(x_k)$ is closed. ■

Theorem 2.7 *Let M be nonempty and (x_k) be M-Fejer monotone. Then the following statements hold:*

a) *(x_k) is bounded and therefore weakly precompact ($A^w(x_k) \neq \emptyset$),*

b) $\|x' - x\| = \inf_k \ \|x_k - x\| = \lim_k \ \|x_k - x\| \quad \forall x' \in A^s(x_k), \ \forall x \in M,$

c) *$A^s(x_k)$ ist a singleton in $M \iff A^s(x_k) \cap M \neq \emptyset$,*

d) *$\lim_k \ x_{k'} = x^* \in M$ for any subsequence $(x_{k'}) \implies \lim_k \ x_k = x^* \in M$,*

e) $\lim_k \ d_k(x) = 0 \quad \forall x \in M.$

Proof: Let (x_k) be M-Fejer monotone.
a) We choose a fixed $x \in M$. Then we get by repeated use of definition

$$\|x_k - x\| \leq \|x_0 - x\|$$

and hence

$$\|x_k\| \leq \|x_k - x\| + \|x\| \leq \|x_0 - x\| + \|x\|$$

for all $k \in \mathbb{N}$. Thus (x_k) is bounded and weakly precompact.
b) We consider the sequence $r_k = \|x_k - x\|$ for a certain fixed $x \in M$. Obviously (r_k) is monotone decreasing and bounded from below (by 0). Consequently, (r_k) converges, and we have

$$r := \lim_k \ r_k = \inf_k \ r_k \, .$$

Let be $x' \in A^s(x_k)$ and choose any subsequence $(x_{k'})$ of (x_k) with limit x'. Since the norm is continuous, we obtain

$$r = \lim_{k'} \ \|x_{k'} - x\| = \|x' - x\| \, .$$

c) Let be $x^* \in A^s(x_k) \cap M$. Then b) implies for $x' = x = x^*$

$$0 = \|x^* - x^*\| = \lim_k \ \|x_k - x^*\| \, .$$

This means $\lim_k \ x_k = x^* \in M$. The reversion is trivial.
d) Since the sequence (r_k) with members $r_k = \|x_k - x\|$ is convergent for each $x \in M$, we get immediately

$$\lim_k \ \|x_{k'} - x\| = \lim_k \ \|x_k - x\|$$

for any subsequence $(x_{k'})$ of (x_k). The assertion follows if we substitute the limit x^* of $(x_{k'})$ for x.
e) Evidently, $d_k(x) = r_k^2 - r_{k+1}^2$ tends to 0 because (r_k) is convergent. ■

Theorem 2.8 *Let M be nonempty, convex and closed. Further, let (x_k) be M-Fejer monotone. Then the following conditions are equivalent:*

a) *$A^s(x_k)$ is a singleton in M.*

b) $\lim_k \ \rho(x_k, M) = 0$.

Moreover, under a) *or* b) *the estimates*

$$\|P_M x_k - x^*\| \leq \rho(x_k, M) \quad , \quad \|x_k - x^*\| \leq 2\rho(x_k, M)$$

hold, where P_M is the metric projector onto M and $x^* \in M$ is the limit of (x_k).

Proof: We consider a M-Fejer monotone sequence (x_k). At first we show the implication b) ⇒ a). The reversion is obvious. By repeated use of definition we get

$$\|x_m - x\| \leq \|x_k - x\| \quad \forall m \geq k \ , \ \forall x \in M \ .$$

If we put $x = P_M x_k$, it follows

$$\|x_m - P_M x_k\| \leq \|x_k - P_M x_k\| = \rho(x_k, M) \quad \forall m \geq k \ .$$

Hence

$$\|x_k - x_m\| \leq \|x_k - P_M x_k\| + \|P_M x_k - x_m\| \leq 2\rho(x_k, M) \quad \forall m \geq k \ .$$

Let us now suppose $\lim_k \ \rho(x_k, M) = 0$. Then (x_k) turns out to be a Cauchy sequence as the last relation supplies $\|x_k - x_m\| \to 0$ for $k \to \infty$. Thus (x_k) converges to a certain element x^* in H because H is complete. Observing $\lim_k \ \rho(x_k, M) = 0$ and the closedness of M , $x^* \in M$ follows. If we finally let m tend to infinity in the second and third above estimate, then also the asserted error estimates for (x_k) are shown. ■

Remark 2.9 The convergence assertion of Theorem 2.8 without error estimates is already given in [7, Lemma 6]. But the proof is more troublesome. Moreover, the well-known theorem had to be used that the intersection of a descending sequence of convex, closed and bounded balls is nonempty. Corresponding error estimates as in Theorem 2.8 under more general and more special conditions can be found in [14] and [15].

Theorem 2.10 *Let (x_k) be M-Fejer monotone. Then (x_k) converges weakly to an element x^* in M iff $A^w(x_k) \subseteq M$.*

Proof: The first part of the equivalence is trivial. We show the second. We consider a M-Fejer monotone sequence (x_k) with $A^w(x_k) \subseteq M$. By Theorem 2.7a) the set $A^w(x_k)$ is nonempty. We assume that there are at least two different elements x' and x'' in $A^w(x_k)$ which lie consequently also in M . Hence the limits

$$r' = \lim_k \|x_k - x'\| \quad , \quad r'' = \lim_k \|x_k - x''\|$$

exist in view of Theorem 2.7b). Without loss of generality we suppose $r'' \leq r'$. We choose a subsequence $(x_{k'})$ of (x_k) with $\operatorname{wlim}_k x_{k'} = x'$. Using the so-called Opial's condition (see [11]) which is satisfied in H we attain

$$\begin{aligned} r' &= \lim_k \|x_k - x'\| = \lim_k \|x_{k'} - x'\| \\ &< \lim_k \|x_{k'} - x''\| = \lim_k \|x_k - x''\| = r'' . \end{aligned}$$

Generally Opial's condition holds for the limit superior. But in view of the Fejer monotony the limit superior coincides with the usual limit. So the last estimate leads to a contradiction. Hence all weak accumulation values of (x_k) are equal (to the weak limit x^* in M).

3 Geometrical aspects

If we want to study the geometrical aspects of Fejer monotone sequences, there are two possible points of view. The first point of view assumes that the set M and an iterate x_k are given. Then we are interested in the **restriction set** which describes the admissible location of the next iterate x_{k+1}. The second point of view starts from the given sequence (x_k) and looks for sets which enclose and generate $C(x_k)$. Both aspects are considered successively.

Lemma 3.1 *Let (x_k) be Fejer monotone. Then there holds with $M = C(x_k)$*

$$x_{k+1} \in \bigcap_{x \in M} B(x, \|x_k - x\|) \subseteq B(x, \rho(x_k, M)) \quad \forall k \in \mathbb{N} .$$

Proof: By Lemma 2.3 (x_k) is M-Fejer monotone with $M = C(x_k)$. This implies immediately the ball condition

$$x_{k+1} \in B(x, \|x_k - x\|)$$

for all k and all $x \in M$. So x_{k+1} must lie also in the intersection of these balls taken over all $x \in M$. One special ball is obtained for $x = P_M x_k$, namely $B(x, \rho(x_k, M))$. ■

Remark 3.2 Obviously, the first restriction set given in Lemma 3.1 is convex and closed, since all balls have this property. Without doubt the shape of this set depends on the special properties of (x_k) and the shape of $C(x_k)$, respectively.

Now we turn to the second point of view. We assume the sequence (x_k) to be given and look for the generation of $C(x_k)$. We consider the sets

$$\begin{aligned} Q_k &:= \{x \in H : \|x_{k+1} - x\| \leq \|x_k - x\|\} , \\ H_k &:= \{x \in H : \|x_{k+1} - x\| = \|x_k - x\|\} , \end{aligned}$$

which are only of interest in the case $x_k \neq x_{k+1}$, since they otherwise coincide with the whole space H. If the definition of $C(x_k)$ is observed, we get the relation

$$C(x_k) = \bigcap_k Q_k .$$

In the following we use the abbreviation

$$y_k := \frac{1}{2}(x_k + x_{k+1}) . \tag{4}$$

.

Theorem 3.3 *Let (x_k) be Fejer monotone.*
Then there holds for any k with $x_k \neq x_{k+1}$: The set Q_k is a nonempty halfspace with the boundary hyperplane H_k. H_k has the normal direction $x_k - x_{k+1}$ and contains the element y_k which is the orthogonal projection of x_k onto H_k. The iterate x_{k+1} is the mirror point of x_k with respect to H_k.

Proof: Let (x_k) be Fejer monotone and consider an index k with $x_k \neq x_{k+1}$. Then Q_k is nonempty as this holds already for $C(x_k)$. The expression (1) can be written in the form

$$d_k(x) = (x_k - x_{k+1}, x_k + x_{k+1} - 2x)$$

(see (2) in Lemma 2.5). Just for the elements of Q_k both sides of the equation are nonnegative. Hence we can write by use of (4)

$$Q_k = \{x \in H : (x_k - x_{k+1}, x) \leq (x_k - x_{k+1}, y_k)\} , \tag{5}$$

$$H_k = \{x \in H : (x_k - x_{k+1}, x) = (x_k - x_{k+1}, y_k)\} . \tag{6}$$

The condition for k ensures that Q_k is a halfspace whose boundary hyperplanes H_k contains the element y_k and has the normal direction $x_k - x_{k+1}$ which is orthogonal to H_k. In view of $x_k - y_k = \frac{1}{2}(x_k - x_{k+1})$ this difference is also orthogonal to H_k. Consequently, y_k is the orthogonal projection of x_k onto H_k. Besides y_k turns out to be the midpoint of the segment generated by the convex hull of $\{x_k, x_{k+1}\}$. Hence x_{k+1} is the mirror point of x_k with respect to H_k. ■

Lemma 3.4 *Let (x_k) be M-Fejer monotone. Then there holds*

a) $(x_{k+1} - x_k, x - x_k) \geq 0 \quad \forall k \in \mathbb{N}\ ,\ \forall x \in M,$

b) $(x_{k+1} - x_k, x - x_k) > 0 \quad \forall k \in \mathbb{N} : x_k \neq x_{k+1}\ ,\ \forall x \in M,$

c) $\|x_{k+1} - x_k\| \leq 2\,\|x_k - x\| \quad \forall k \in \mathbb{N}\ ,\ \forall x \in M,$

d) $\|x_{k+1} - x_k\| \leq 2\,\rho(x_k, M) \quad \forall k \in \mathbb{N}.$

Proof: Let (x_k) be M-Fejer monotone. Then we have in view of Lemma 2.5 for all k and all x in M

$$2(x_k - x_{k+1}, x_k - x) \geq \|x_k - x_{k+1}\|^2\ .$$

This inequality leads immediately to the assertions a) and b). For $x_k = x_{k+1}$ the assertion c) holds trivially. In the other case we get by Schwarz's inequality

$$2\|x_k - x_{k+1}\|\|x_k - x\| \geq \|x_k - x_{k+1}\|^2$$

and therefore the assertion c). Putting $x = P_M x_k$ in assertion c) assertion d) follows. ■

Remark 3.5 The relations a) and b) in Lemma 3.4 show that the angle between the elements $x_{k+1} - x_k$ and $x - x_k$ is not obtuse and for $x_k \neq x_{k+1}$ even acute considering any of the elements x in M. The relations c) and d) can be expressed as ball conditions

$$x_{k+1} \in \bigcap_{x \in M} B(x_k, 2\,\|x_k - x\|) \subseteq B(x_k, 2\,\rho(x_k(M))\ .$$

Obviously, they are weaker than the ball conditions in Lemma 3.1. But c) and d) show that for iterates x_k in a small neighbourhood of any $x \in M$ or of M itself the new iterates x_{k+1} are nearby the old.

4 Uniformly Fejer monotone sequences

Definition 4.1 *Let M be a nonempty subset of H and (τ_k) a sequence of nonnegative numbers. The sequence (x_k) is said to be* **uniformly M-Fejer monotone relative to** *(r.t.) (τ_k) iff*

$$d_k(x) = \|x_k - x\|^2 - \|x_{k+1} - x\|^2 \geq \tau_k \quad \forall k \in \mathbb{N}\ ,\ \forall x \in M$$

and

$$\tau_k > 0 \quad \forall k \in \mathbb{N} : x_k \notin M.$$

It is said to be **uniformly Fejer monotone** *iff it is uniformly M-Fejer monotone r.t. (τ_k) for any admissible sequence (τ_k) and for any subset M. The set*

$$C(x_k, \tau_k) := \{x \in H : d_k(x) \geq \tau_k \quad \forall k \in \mathbb{N}\}$$

is called the **Fejer carrier** *of (x_k)* **r.t.** *(τ_k).*

Lemma 4.2 *Let (x_k) be uniformly M-Fejer monotone r.t. (τ_k). Then holds*

a) *(x_k) is strictly M-Fejer monotone,*

b) *(x_k) is $C(x_k, \tau_k)$-Fejer monotone r.t. (τ_k),*

c) $\lim_k \ \tau_k = 0$.

Proof: a) Let (x_k) be uniformly M-Fejer monotone r.t. (τ_k). This means by definition

$$d_k(x) = \|x_k - x\|^2 - \|x_{k+1} - x\|^2 \geq \tau_k \geq 0$$

and implies

$$\|x_k - x\| \geq \|x_{k+1} - x\|$$

for all k and all $x \in M$. Hence (x_k) is M-Fejer monotone. Assuming $x_k \notin M$ we know $\tau_k > 0$. This leads to $\|x_k - x\| > \|x_{k+1} - x\|$. So (x_k) is strictly M-Fejer monotone.
b) Assume that $x_k \notin C(x_k, \tau_k)$. Then also $x_k \notin M$ because of $M \subseteq C(x_k, \tau_k)$. Hence $\tau_k > 0$. Now assertion b) is obvious.
c) Considering the first relation in part a) the assertion follows by Theorem 2.7e). ■

Lemma 4.2 shows that we can also put $M = C(x_k, \tau_k)$ in assertion a). The comparision of the corresponding types of Fejer carriers leads to the relation

$$C(x_k, \tau_k) \subseteq C(x_k) \ . \tag{7}$$

Already simple examples show that the equality need not to hold. The relative carrier $C(x_k, \tau_k)$ turns out to be convex and closed, too. This can be shown directly in an analogue way as in Theorem 2.6 for $C(x_k)$. But it is also a consequence of geometric facts given below (see Corollary 4.6).

Lemma 4.3 *Let (μ_k) and (ν_k) be two number sequences satisfying $\mu_k \geq \nu_k \geq 0$. Let (x_k) be a sequence with $\nu_k > 0$ for $x_k \notin M$. Then there holds*

a) *(x_k) uniformly M-Fejer monotone r.t. $(\mu_k) \Longrightarrow (x_k)$ uniformly M-Fejer monotone r.t. (ν_k),*

b) $C(x_k, \mu_k) \subseteq C(x_k, \nu_k)$.

Proof: The sequences (μ_k) and (ν_k) satisfy the conditions of Definition 4.1. Assertion a) follows immediately by the relations

$$\|x_k - x\|^2 - \|x_{k+1} - x\|^2 \geq \mu_k \geq \nu_k \ ,$$

which are fulfilled for all k and all $x \in M$. Since the weaker estimate with ν_k on the right-hand side admits possibly more elements x than the estimate with μ_k, assertion b) is also obvious. ■

If (x_k) is uniformly M-Fejer monotone r.t. (τ_k), then (τ_k) is not uniquely determined. The maximal choice is given by

$$\tau_k = \tau_k^m := \inf\nolimits_{x \in M} \ d_k(x).$$

Now we turn again to geometrical aspects. First we are interested in restriction sets of x_{k+1} for given M and x_k.

Lemma 4.4 *Let (x_k) be uniformly M-Fejer monotone r.t. (τ_k). Then*

$$x_{k+1} \in \bigcap_{x \in M} B(x, \sqrt{\|x_k - x\|^2 - \tau_k}) \subseteq B(x, \sqrt{\rho^2(x_k, M) - \tau_k}) \quad \forall k \in \mathbb{N} \ .$$

Proof: Observing Definition 4.1 we get for uniformly M-Fejer monotone sequences r.t. (τ_k) the ball condition

$$x_{k+1} \in B\left(x, \sqrt{\|x_k - x\|^2 - \tau_k}\right)$$

for all k and all $x \in M$. Hence x_{k+1} lies also in the intersection of these balls taken over all $x \in M$. A special ball is obtained for $x = P_M(x_k)$, namely $B(x, \sqrt{\rho^2(x_k, M) - \tau_k})$. ■

The above stated ball conditions for x_{k+1} have a crucial disadvantage if τ_k depends for its part on x_{k+1} as it is the case for strongly Fejer monotone sequences (see section 5). Then x_{k+1} occurs on both sides of the element relation causing a reflexivity. This effect can be avoided for the mentioned special case by a rearrangement of the defining estimate (see Lemma 5.3 and Remark 5.5).

Now we assume that the given sequence (x_k) is uniformly M-Fejer monotone r.t. (τ_k) and look for the generation of M. We consider the sets

$$\begin{aligned} Q'_k &:= \{x \in H : \|x_k - x\|^2 - \|x_{k+1} - x\|^2 \geq \tau_k\} \ , \\ H'_k &:= \{x \in H : \|x_k - x\|^2 - \|x_{k+1} - x\|^2 = \tau_k\} \ . \end{aligned}$$

Consulting the definition of $C(x_k, \tau_k)$ we can conclude

$$M \subseteq \bigcap_k Q'_k = C(x_k, \tau_k) \ .$$

Theorem 4.5 *Let (x_k) be uniformly M-Fejer monotone r.t. (τ_k). Then there holds for all k with $x_k \notin M$: The set Q'_k is a nonempty halfspace contained in Q_k with the boundary hyperplane H'_k parallel to H_k. The orthogonal projection y'_k of x_k onto H'_k is given by*

$$y'_k = \frac{1}{2}\left(1 - \frac{\tau_k}{\|x_k - x_{k+1}\|^2}\right) x_k + \frac{1}{2}\left(1 + \frac{\tau_k}{\|x_k - x_{k+1}\|^2}\right) x_{k+1}.$$

The new iterate x_{k+1} has the representation

$$x_{k+1} = \frac{\tau_k - \|x_k - x_{k+1}\|^2}{\tau_k + \|x_k - x_{k+1}\|^2}\, x_k \; + \; 2\, \frac{\|x_k - x_{k+1}\|^2}{\tau_k + \|x_k - x_{k+1}\|^2}\, y'_k \; .$$

Proof: We suppose a sequence (x_k) which is uniformly M-Fejer monotone r.t. (τ_k). Then the nonempty set M is contained in the intersection of all sets Q'_k. Hence these sets are nonempty. Using the abbreviation (1) we can write

$$Q'_k = \{x \in H : d_k(x) - \tau_k \geq 0\} \; , \; H'_k = \{x \in H : d_k(x) - \tau_k = 0\} \; .$$

By relation (2) in Lemma 2.5 we get

$$d_k(x) - \tau_k = (x_k - x_{k+1}, x_k + x_{k+1} - 2x) - \tau_k \geq 0$$

and by rearrangement

$$(x_k - x_{k+1}, x) \leq \frac{1}{2}\,(x_k - x_{k+1}, x_k + x_{k+1}) - \frac{1}{2}\,\tau_k \leq (x_k - x_{k+1}, \frac{1}{2}\,(x_k + x_{k+1})) \qquad (8)$$

for all elements x in Q'_k. As (x_k) is regularly M-Fejer monotone, we know $x_k - x_{k+1} \neq 0$. Consequently, the sets Q'_k are halfspaces whose boundary hyperplanes H'_k possess the normal direction $x_k - x_{k+1}$. A comparision with (5), (6) shows $Q'_k \subseteq Q_k$ and $H'_k \parallel H_k$ if (4) is taken into account. The projection y'_k of x_k onto H'_k satisfies

$$y'_k = x_k - \beta\,(x_k - x_{k+1}) \in H'_k \; .$$

If we substitute $x = y'_k$ in the equation for H'_k corresponding to (8), we obtain

$$(x_k - x_{k+1}, x_k) - \beta\,\|x_k - x_{k+1}\|^2 = \frac{1}{2}\,(x_k - x_{k+1}, x_k + x_{k+1}) - \frac{1}{2}\,\tau_k$$

and

$$\beta\,\|x_k - x_{k+1}\|^2 = \frac{1}{2}(\|x_k - x_{k+1}\|^2 + \tau_k) \; .$$

This leads to

$$\beta = \frac{1}{2}(1 + \frac{\tau_k}{\|x_k - x_{k+1}\|^2}) \; .$$

Since y'_k has the equivalent representation $y'_k = (1-\beta)x_k + \beta x_{k+1}$, the asserted form of the projection arises. If we separate x_{k+1} in the last equation, we win

$$x_{k+1} = (1-\lambda)x_k + \lambda y'_k \ , \tag{9}$$

where

$$\lambda = \frac{1}{\beta} = 2\,\frac{\|x_k - x_{k+1}\|^2}{\tau_k + \|x_k - x_{k+1}\|^2} \ . \tag{10}$$

The calculation of $1-\lambda$ completes the proof for the asserted representation of x_{k+1}. ■

Denoting the orthogonal projector onto H'_k by P'_k the equation (9) has the form

$$x_{k+1} = (1-\lambda)x_k + \lambda P'_k x_k \ ,$$

where λ lies in the open interval $(0,2)$ because of (10). This kind of iteration is called **relaxation** or more precisely **relaxed projection** of x_k onto H'_k. The value $\lambda = 1$ supplies just the projection while $\lambda < 1$ and $\lambda > 1$ lead to **underrelaxation** and **overrelaxation**, respectively. In the first case the new iterate stops before the hyperplane H'_k, in the second case it passes H'_k and stops before the mirror point.

Corollary 4.6 *The relative carrier $C(x_k, \tau_k)$ is convex and closed.*

Proof: The set $C(x_k, \tau_k)$ is the intersection of all sets Q'_k which are by Theorem 4.5 half-spaces or for $x_k = x_{k+1}$ the whole space H. Hence $C(x_k, \tau_k)$ is convex and closed because the sets Q'_k are convex and closed. ■

5 Strongly Fejer monotone sequences

We introduce an important special case of uniformly Fejer monotone sequences.

Definition 5.1 *Let M be a nonempty subset of H and α a positive number. The sequence (x_k) is said to be α-**strongly** M-**Fejer monotone** iff*

$$d_k(x) = \|x_k - x\|^2 - \|x_{k+1} - x\|^2 \geq \alpha \, \|x_k - x_{k+1}\|^2 \quad \forall k \in \mathbb{N} \ , \ \forall x \in M$$

and

$$x_{k+1} \neq x_k \quad \forall k \in \mathbb{N} : x_k \notin M \ .$$

It is called **strongly Fejer monotone** *iff it is α-strongly Fejer monotone for any α and any M. The set*

$$C^\alpha(x_k) := \{x \in H : d_k(x) \geq \alpha \, \|x_k - x_{k+1}\|^2 \quad \forall k \in \mathbb{N}\}$$

*is called the α-**strong Fejer carrier** of (x_k).*

Remark 5.2 Evidently, α-strongly M-Fejer monotone sequences are uniformly M-Fejer monotone r.t. the special sequences

$$\tau_k = \tau_k^s(\alpha) := \alpha \|x_k - x_{k+1}\|^2 \ . \tag{11}$$

Consequently, we get $C^\alpha(x_k) = C(x_k, \tau_k^s(\alpha))$. This carrier is also convex and closed by Corollary 4.6. Additionally there holds

$$C^\alpha(x_k) \subseteq C^\beta(x_k) \subseteq C(x_k) \quad \text{for} \quad \alpha \geq \beta > 0 \ .$$

The first relation follows by Lemma 4.3b). The second relation is a special case of (7). The maximal choice for the parameter α is given by

$$\alpha = \alpha^m := \inf_{k: x_k \notin M,\, x \in M} \frac{d_k(x)}{\|x_k - x_{k+1}\|^2} \ .$$

For α-strongly M-Fejer monotone sequences (x_k) Lemma 4.2c) leads with $\tau_k = \tau_k^s(\alpha)$ to

$$\lim_k \|x_k - x_{k+1}\| = 0 \ .$$

Now we turn to results which can not be obtained by specialization of statements in section 4.

Lemma 5.3 *Let (x_k) be α-strongly M-Fejer monotone. Then there holds*

a) $(1+\alpha)\|x_k - x_{k+1}\|^2 \leq 2\,(x_k - x_{k+1}, x_k - x) \quad \forall k \in \mathbb{N}\ ,\ \forall x \in M\ ,$

b) $\|x_{k+1} - x_k + \dfrac{1}{1+\alpha}\,(x_k - x)\| \leq \dfrac{1}{1+\alpha}\,\|x_k - x\| \quad \forall k \in \mathbb{N}\ ,\ \forall x \in M\ .$

Proof: We consider the expression

$$f_k(x) := d_k(x) - \tau_k^s(\alpha) \ . \tag{12}$$

In view of Definition 5.1 and (11) this expression is nonnegative for α-strongly M-Fejer monotone sequences (x_k). Using (2) and (11) this means explicitely

$$f_k(x) = 2\,(x_k - x_{k+1}, x_k - x) - \|x_k - x_{k+1}\|^2 - \alpha\|x_k - x_{k+1}\|^2 \geq 0$$

and consequently

$$(1+\alpha)\|x_k - x_{k+1}\|^2 \leq 2\,(x_k - x_{k+1}, x_k - x)$$

for all k and all $x \in M$. Hence assertion a) is fulfilled. Finally we obtain by rearrangement

$$\|x_{k+1} - x_k\|^2 + 2\,(x_{k+1} - x_k,\, \frac{1}{1+\alpha}\,(x_k - x)) \leq 0$$

and quadratic completion

$$\begin{aligned}
&\|x_{k+1} - x_k + \frac{1}{1+\alpha}\,(x_k - x)\|^2 \\
&\quad = \|x_{k+1} - x_k\|^2 + 2\,(x_{k+1} - x_k, \frac{1}{1+\alpha}(x_k - x)) + \frac{1}{(1+\alpha)^2}\,\|x_k - x\|^2 \\
&\quad \leq \frac{1}{(1+\alpha)^2}\,\|x_k - x\|^2\ .
\end{aligned}$$

This leads to assertion b). ■

Corollary 5.4 *Let (x_k) be α-strongly M-Fejer monotone. Then there holds*

a) $(1+\alpha)\|x_k - x_{k+1}\| \leq 2\,\|x_k - x\| \quad \forall k \in \mathbb{N}\ ,\ \forall x \in M\ ,$

b) $(1+\alpha)\|x_k - x_{k+1}\| \leq 2\,\rho(x_k, M) \quad \forall k \in \mathbb{N}\ .$

Proof: Starting with the estimate a) of Lemma 5.3 and applying Schwarz's inequality to the right-hand side the assertion a) occurs if $\|x_k - x_{k+1}\|$ is cancelled in the case $x_k \neq x_{k+1}$. But assertion a) is trivially fulfilled for $x_k = x_{k+1}$. The special choice $x = P_M(x)$ in assertion a) supplies assertion b). ■

Remark 5.5 The statement b) of Lemma 5.3 corresponds to the ball condition

$$x_{k+1} \in \bigcap_{x\in M} B(x_k - \frac{1}{1+\alpha}(x_k - x), \frac{1}{1+\alpha}\,\|x_k - x\|) \quad \forall k \in \mathbb{N}\ ,$$

whereas the ball with the minimal radius arises for $x = P_M x_k$. The midpoints of the balls can be written in the form

$$x' = \frac{\alpha}{1+\alpha}\,x_k + \frac{1}{1+\alpha}\,x\ .$$

Hence they are convex combinations of x_k and x. The statement a) of Corollary 5.4 is equivalent to the ball condition

$$x_{k+1} \in \bigcap_{x\in M} B(x_k, \frac{2}{1+\alpha}\,\|x_k - x\|) \quad \forall k \in \mathbb{N}\ .$$

Evidently, we have

$$B(x', \frac{1}{1+\alpha}\,\|x_k - x\|) \subseteq B(x_k, \frac{2}{1+\alpha}\,\|x_k - x\|)\ .$$

By the way, both compared balls have the same nearest point to x, namely

$$\begin{aligned}
x'' &= x' - \frac{1}{1+\alpha}(x_k - x) = x_k - \frac{2}{1+\alpha}(x_k - x) \\
&= \frac{\alpha - 1}{\alpha + 1}\,x_k + \frac{2}{\alpha+1}\,x\ .
\end{aligned}$$

Hence x'' is as x' a convex combination of x_k and x.

Theorem 5.6 *Let (x_k) be α-strongly M-Fejer monotone. Then*

$$y_k^\alpha := \frac{1-\alpha}{2}\, x_k + \frac{1+\alpha}{2}\, x_{k+1}$$

is the projection of x_k onto the halfspace

$$Q_k^\alpha := \{x \in H : \|x_k - x\|^2 - \|x_{k+1} - x\|^2 \geq \alpha \, \|x_k - x_{k+1}\|^2\}$$

and onto the corresponding boundary hyperplane H_k^α, respectively. The new iterate x_{k+1} satisfies the equation

$$x_{k+1} = \frac{\alpha - 1}{\alpha + 1}\, x_k + \frac{2}{\alpha + 1}\, y_k^\alpha \; .$$

It realizes with respect to H_k^α overrelaxation for $\alpha < 1$ and underrelaxation for $\alpha > 1$. Moreover, it is

$$M \subseteq \bigcap_k Q_k^\alpha = C^\alpha(x_k) \; .$$

Proof: The results follow from Theorem 4.5 and the preceding set relation if we put there $\tau_k = \tau_k^s(\alpha)$. Especially, we get $\lambda = \dfrac{2}{\alpha+1}$ and $\alpha = \dfrac{2-\lambda}{\lambda}$. This implies also the statements about overrelaxation and underrelaxation (see remarks at the end of section 4). ■

References

[1] **Agmon, S. :** *The relaxation method for linear inequalities.* Canad. J. Math. **6**, 382-392 (1954)

[2] **Cegielski, A. :** *Metody relaksacyjne w problemach optymalizacji wypukłej.* Monografie 67, Wyzsza Szkoła Inzynierska, Zielona Gora 1993

[3] **Elsner, L., Koltracht, I.** and **Neumann, M. :** *Convergence of sequential and asynchronous nonlinear paracontractions.* Numer. Math. **62**, 305-319 (1992)

[4] **Eremin, I.I. :** *The relaxation method of solving systems of inequalities with convex functions on the left-hand sides.* Dokl. Akad. Nauk SSSR **160**, 994-996 (1965)

[5] **Eremin, I.I. :** *Fejer mappings and problems of convex optimization.* Sibirsk. Mat. Zh. **10**, 1034-1047 (1969)

[6] **Eremin, I.I.** and **Mazurov, V.D. :** *Nestacionarnye Processy Programmirovanija* (Russian). Moskva 1979

[7] **Gurin, L.G., Poljak, B.T.** and **Raik, E.V. :** *The method of projections for finding the common point of convex sets.* Zh. Vychisl. Mat. i Mat. Fiz. **7**, 1211-1228 (1967)

[8] **He, B.** and **Stoer, J. :** *Solution of projection problems over polytopes.* Numer. Math. **61**, 73-90 (1992)

[9] **Kaczmarz, S. :** *Angenäherte Auflösung von Systemen linearer Gleichungen.* Bull. Internat. Acad. Polon. Sci. Cl. A **35**, 355-357 (1937)

[10] **Oettli, W. :** *Symmetric duality, and a convergent subgradient method for discrete, linear, constrained approximation problems with arbitrary norms appearing in the objective function and in the constraints.* J. Approx. Theory **14**, 43-50 (1975)

[11] **Opial, Z. :** *Weak convergence of succession approximations for nonexpansive mappings.* Bull. Amer. Math. Soc. **73**, 591-597 (1967)

[12] **Poljak, B.T. :** *Minimization of nonsmooth functionals.* Zh. Vychisl. Mat. i Mat. Fiz. **9**, 509-521 (1969)

[13] **Schott, D. :** *A general iterative scheme with applications to convex optimization and related fields.* Optimization **22**, 885-902 (1991)

[14] **Schott, D. :** *About geometrical convergence of general iterative methods applied to nonunique solvable convex problems - part I .* J. Comput. Appl. Math. **54**, 1-14 (1994)

[15] **Schott, D. :** *About geometrical convergence of general iterative methods applied to nonunique solvable convex problems - part II.* J. Comput. Appl. Math. **54**, 133-150 (1994)

[16] **Schott, D. :** *Basic properties of Fejer monotone mappings.* (submitted to Rostock. Math. Kolloq.)

[17] **D. Schott, D. :** *Iterative solution of convex problems by Fejer monotone methods.* Preprint 95/5, Universität Rostock (to appear in Numer. Funct. Anal. Optim.)

[18] **Schumacher, K. :** *Iterative Optimierungsverfahren, die unter schwachen Voraussetzungen konvergieren.* Numer. Math. **24**, 443-456 (1975)

[19] **Stark, H. (Ed.) :** *Image recovery: Theory and application.* New York 1987

received: September 6, 1995

Author:

Prof. Dr. Dieter Schott
Neubrandenburger Str. 49a
D-18055 Rostock
Germany

Rostock. Math. Kolloq. **49**, 75–88 (1995)

Subject Classification (AMS)
65F15, 65G10, 65H10

Günter Mayer

On a Unified Representation of Some Interval Analytic Algorithms

Dedicated to the professors of mathematics
L. Berg, W. Engel, G. Pazderski, *and* H.- W. Stolle.

ABSTRACT. In this article we show that several interval analytic algorithms for verifying solutions x^* of various mathematical problems can be viewed as special cases of some iterative method involving the first and the second derivative of the underlying function. Starting with an approximation of x^* a way is given how to construct interval vectors which contain x^*, and how to improve these interval bounds.

KEY WORDS. Systems of nonlinear equations, second order methods, enclosure methods, verification methods, interval methods for nonlinear systems, algebraic eigenproblem, singular value decomposition, quadratic systems, invariant supspaces

1 Introduction

The solutions of many mathematical problems can be expressed as zeros of some function $f : D \subseteq \boldsymbol{R}^n \to \boldsymbol{R}^n$. Among these problems are the algebraic eigenproblem, the generalized eigenproblem, the singular value problem and the generalized singular value problem, see [1]–[4], [8]. For example, let A, $B \in \boldsymbol{R}^{n\times n}$, $\zeta \in \boldsymbol{R}\backslash\{0\}$ and $i_0 \in \{1, \dots , n\}$ be given. Then with $x := (v^T, \lambda)^T$, $v \in \boldsymbol{R}^n$, $\lambda \in \boldsymbol{R}$ the zeros $x^* = ((v^*)^T, \lambda^*)^T$ of the function

$$f(x) := \begin{pmatrix} Av - \lambda Bv \\ v_{i_0} - \zeta \end{pmatrix} \tag{1}$$

are obviously eigenpairs of the generalized eigenproblem $Av = \lambda Bv$ with the eigenvector $v^* = (v_i^*)$ being normalized by $v_{i_0}^* = \zeta \neq 0$. We will show how certain algorithms for verifying and enclosing solutions of the above problems can be derived from one verification method for general systems of nonlinear equations. To this end we will consider the interval function

$$[g]([x], \widetilde{x}) := t(\widetilde{x}) + \{t'(\widetilde{x}) + [H]([x], \widetilde{x})\}([x] - \widetilde{x}) \tag{2}$$

for which the function $t : D \subseteq \boldsymbol{R}^n \to \boldsymbol{R}^n$ is assumed to be twice continuously differentiable on a given open set D. The matrix $t'(\widetilde{x})$ is the Jacobian of t at some fixed vector $\widetilde{x} \in D$. The function $[H] = [H]([x], \widetilde{x})$ is defined for all interval vectors $[x] \subseteq D$ and has $n \times n$ interval matrices as values. It is supposed to be continuous and inclusion monotone with respect to $[x]$.

For t, $[H]$ and all $[x] \subseteq D$ we require

$$t(x) \in [g]([x], \widetilde{x}) \quad \text{for all } x \in [x] \tag{3}$$

and

$$\| \, |[H]([x], \widetilde{x})| \, \| \leq \gamma \| \, |[x] - \widetilde{x}| \, \|. \tag{4}$$

The constant γ is positive and fixed for all $[x] \subseteq D$; it may depend on $\widetilde{x}$. Throughout the paper $\| \cdot \|$ denotes the maximum norm of a real vector and the row sum norm of a real matrix, respectively; $| \cdot |$ is the absolute value which we shall define for interval quantities in our next section. There, we also show how t and $[H]$ normally are related to the given function f, the zeros of which we are interested in. We shall derive criteria (Theorem 1) for guaranteeing the existence of a fixed point x^* of t which then turns out to be a zero of f. In verifying x^* we will construct an interval vector $[x]^0$ which contains x^* and which in a natural way provides lower and upper bounds for it. Some of our criteria will yield the subset property $[g]([x]^0, \widetilde{x}) \subseteq [x]^0$ which, together with Brouwer's fixed point theorem, forms the basis for many verification algorithms. We also will improve the enclosure $[x]^0$ of x^* by considering the iteration

$$[x]^{k+1} := [g]([x]^k, \widetilde{x}) \cap [x]^k, \quad k = 0, 1, \dots ,$$

where the intersection can be dropped if the above–mentioned subset property holds. Under slight additional assumptions on $[H]$ we will show that x^* is the unique fixed point of t within $[x]^0$, that all the iterates $[x]^k$ contain it as element and that they contract to it for $k \to \infty$.

For the standard choice of t and $[H]$ it will turn out that the function $[g]$ reduces to that in Platzöder [14] and Alefeld [5]. Therefore, parts of our criteria are generalizations of results of these authors.

2 Results

In order to formulate our results we first list some notations needed later on. By $\boldsymbol{IR}$, $\boldsymbol{IR}^n$, $\boldsymbol{IR}^{m \times n}$, respectively, we denote the set of real compact intervals, the set of vectors with n

interval components, and the set of $m \times n$ matrices with entries from $\boldsymbol{IR}$. For degenerate intervals $[a,a]$ we simply write a identifying in this way $\boldsymbol{R}$ with $\{[a,a] \mid a \in \boldsymbol{R}\} \subseteq \boldsymbol{IR}$. We proceed similarly for degenerate interval vectors and degenerate interval matrices. Examples are the null matrix O, the identity matrix I, the i–th column $e^{(i)}$ of I and the vector $e := (1,1,\ldots,1)^T$. In order to indicate $I \in \boldsymbol{R}^{n\times n}$ we sometimes write I_n instead of I. As usual, we equip $\boldsymbol{R}^n$ and $\boldsymbol{R}^{m\times n}$, respectively, with the natural semi–ordering '$\leq$' which is defined to hold entrywise. We use the notation $[A] = ([a]_{ij}) \in \boldsymbol{IR}^{m\times n}$ simultaneously without further reference, and we assume the same for the elements of $\boldsymbol{R}^n, \boldsymbol{R}^{m\times n}, \boldsymbol{IR}$ and $\boldsymbol{IR}^n$. For $[a] = [\underline{a}, \overline{a}] \in \boldsymbol{IR}$ we define the *absolute value* $|[a]|$ by $|[a]| := \max\{|a| \mid a \in [a]\} = \max\{|\underline{a}|, |\overline{a}|\}$ and the *diameter* $d([a])$ by $d([a]) := \overline{a} - \underline{a}$, and we denote the *convex hull* of $[a]$, $[b] \in \boldsymbol{IR}$ by $[a]\underline{\cup}[b]$. For interval vectors and interval matrices these terms are applied entrywise. If $f(x)$ is an expression for some function f, we write $f([x])$ for the interval arithmetic evaluation of this expression (cf. [5] assuming that $f([x])$ exists. For further details on interval analysis we refer to [5] or [13].

We start by presenting the main result of our paper. Among others, this result lists criteria for proving the existence of a fixed point x^* of t from (2) and for guaranteeing the subset property

$$[g]([x]^0, \widetilde{x}) = t(\widetilde{x}) + \{t'(\widetilde{x}) + [H]([x]^0, \widetilde{x})\}([x]^0 - \widetilde{x}) \subseteq [x]^0 \tag{5}$$

for some interval vector $[x]^0$.

Theorem 1 *With D, $[g]$, $[H]$, t, $\widetilde{x}$ as in (2) – (4) and with $\gamma > 0$ from (4) choose $r \in \boldsymbol{R}$ with $r > 0$ such that $[x]^0 := \widetilde{x} + [-r, r]e \subseteq D$, and define α, β by*

$$\alpha := \|\, t(\widetilde{x}) - \widetilde{x}\,\|, \qquad \beta := \|\, t'(\widetilde{x})\,\|.$$

Let $\beta < 1$, $\quad \Delta := (1-\beta)^2 - 4\alpha\gamma \geq 0$ and let

$$r^- := (1 - \beta - \sqrt{\Delta})/(2\gamma), \quad r^+ := (1 - \beta + \sqrt{\Delta})/(2\gamma).$$

a) If $r \geq r^-$ then t has at least one fixed point $x^* \in [x]^0$. The iteration

$$[x]^{k+1} := [g]([x]^k, \widetilde{x}) \cap [x]^k, \quad k = 0, 1, \ldots$$

converges to some interval vector $[x]^*$ with

$$x^* \in [x]^* \subseteq [x]^k \subseteq [x]^{k-1} \subseteq \ldots \subseteq [x]^0, \quad k \in \boldsymbol{N}.$$

b) If $r \in [r^-, r^+]$ then t has at least one fixed point $x^* \in [x]^0$. In addition, $[g]([x]^0, \widetilde{x}) \subseteq [x]^0$ holds and the iteration

$$[x]^{k+1} := [g]([x]^k, \widetilde{x}), \quad k = 0, 1, \ldots$$

converges to some interval vector $[x]^*$ *with*

$$x^* \in [x]^* \subseteq [x]^k \subseteq [x]^{k-1} \subseteq \ldots \subseteq [x]^0, \quad k \in \boldsymbol{N}.$$

c) *In addition to* (4), *let* $[H]$ *fulfill*

$$\| \, d\,([H]([x],\widetilde{x})) \, \| \leq \delta \, \| \, d([x]) \, \| \tag{6}$$

for all interval vectors $[x] \subseteq D$ *and for some positive number* δ *which is independent of* $[x]$ *but which may depend on* $\widetilde{x}$. *Define* $\hat{\Delta}$, $\hat{r}^-$, $\hat{r}^+$ *as* Δ, r^-, r^+, *with* γ *being replaced by* $\hat{\gamma} := \max\{\gamma, \delta\}$. *If* $\hat{\Delta} \geq 0$ *and if* $r \in [\hat{r}^-, (\hat{r}^- + \hat{r}^+)/2)$ *then the function* t *has exactly one fixed point* $x^* \in [x]^0$; $[g]([x]^0, \widetilde{x}) \subseteq [x]^0$ *holds, and the iteration*

$$[x]^{k+1} := [g]([x]^k, \widetilde{x}), \quad k = 0, 1, \ldots$$

converges to x^* *with*

$$x^* \in [x]^k \subseteq [x]^{k-1} \subseteq \ldots \subseteq [x]^0, \quad k \in \boldsymbol{N}.$$

Proof: First we remark that the assumptions $\beta < 1$ and $\Delta \geq 0$ guarantee $0 \leq r^- \leq r^+$.

b) Let $[x]^0 := \widetilde{x} + [-r, r]e$. Then (5) is equivalent to

$$[g]([x]^0, \widetilde{x}) - \widetilde{x} \subseteq [x]^0 - \widetilde{x}. \tag{7}$$

Property (7) certainly holds if

$$|t(\widetilde{x}) - \widetilde{x}| + \big(|t'(\widetilde{x})| + |[H]([x]^0, \widetilde{x})|\big)\, re \leq re,$$

and this, in turn, is true if

$$\alpha + (\beta + \gamma r) r \leq r, \tag{8}$$

where we used (4). Now (8) can be rewritten as

$$\gamma r^2 + (\beta - 1) r + \alpha \leq 0 \tag{9}$$

with equality for $r = r^-$ and $r = r^+$. Hence (9) is fulfilled for each $r \in [r^-, r^+]$, and by (3)

$$t(x) \in [g]([x]^0, \widetilde{x}) \subseteq [x]^0$$

holds for any $x \in [x]^0$. Therefore, Brouwer's fixed point theorem guarantees that t has at least one fixed point $x^* \in [x]^0$. Since $[H]$ was assumed to be inclusion monotone the iterates $[x]^k$ decrease monotonically with respect to the semi–ordering '$\subseteq$'; hence they are convergent to some limit $[x]^*$, and

$$x^* = t(x^*) \in [g]([x]^k, \widetilde{x}) = [x]^{k+1}$$

holds for $k = 0, 1, \ldots$ by induction.

a) Choose $r' \leq r$ such that $r' \in [r^-, r^+]$. Then b) applies with $[\hat{x}]^0 := \widetilde{x} + [-r', r']e \subseteq [x]^0$ yielding a fixed point $x^* \in [\hat{x}]^0 \subseteq [x]^0$ of t. By the intersection in the definition of $[x]^{k+1}$ the iterates $[x]^k$ decrease monotonically with respect to '$\subseteq$', thus the proof of a) terminates analogously to that of b).

c) Since $\hat{\gamma} \geq \gamma$ we have $\hat{\Delta} \leq \Delta$ and

$$r^- = \frac{2\alpha}{1 - \beta + \sqrt{\Delta}} \leq \hat{r}^- \leq \hat{r}^+ \leq r^+ = \frac{2\alpha}{1 - \beta - \sqrt{\Delta}} \ .$$

Therefore, r is contained in $[r^-, r^+]$, and c) is proved by b) with the exception of the uniqueness of x^* and of the degeneracy of $[x]^* = [x^*, x^*]$. In order to show $d([x]^*) = 0$ apply $d(\cdot)$ to the equality $[x]^* = [g]([x]^*, \widetilde{x})$. Then by the subdistributivity of the interval arithmetic and by elementary rules for the diameter (cf. [5] for example) one obtains

$$\begin{aligned} d([x]^*) &\leq d\left(t'(\widetilde{x})([x]^* - \widetilde{x}) + [H]([x]^*, \widetilde{x})\left([x]^* - \widetilde{x}\right)\right) \\ &= |t'(\widetilde{x})|\, d([x]^*) + d\left([H]([x]^*, \widetilde{x})\left([x]^* - \widetilde{x}\right)\right) \\ &\leq |t'(\widetilde{x})|\, d([x]^*) + d([H]([x]^*, \widetilde{x}))\, |[x]^* - \widetilde{x}| + |[H]([x]^*, \widetilde{x})|\, d([x]^*) \\ &\leq |t'(\widetilde{x})|\, d([x]^*) + d([H]([x]^*, \widetilde{x}))\, |[x]^0 - \widetilde{x}| + |[H]([x]^0, \widetilde{x})|\, d([x]^*) \end{aligned}$$

Let $d^* := \|d([x]^*)\|$ and apply $\|\cdot\|$ to this inequality in order to get

$$d^* \leq \beta d^* + r\delta d^* + r\gamma d^* \leq \beta d^* + 2r\hat{\gamma} d^*.$$

If $d^* > 0$, we obtain $1 \leq \beta + 2r\hat{\gamma}$ which yields the contradiction

$$r \geq \frac{1 - \beta}{2\hat{\gamma}} = \frac{\hat{r}^- + \hat{r}^+}{2}.$$

Therefore, $d^* = 0$, and $x^* \in [x]^*$ implies $[x]^* = [x^*, x^*]$. In particular, this proves uniqueness.

□

Note that if $\widetilde{x}$ is a sufficiently good approximation of a fixed point x^* of t then α will be small and the assumption $\Delta \geq 0$ will certainly be fulfilled provided that $\beta < 1$.

In practical applications, one often chooses

$$[H]([x], \widetilde{x}) := \frac{1}{2} t''([x] \underline{\cup} \widetilde{x})([x] - \widetilde{x}) \in \boldsymbol{IR}^{n \times n} \tag{10}$$

where

$$t''(x) : \begin{cases} \boldsymbol{R}^n \times \boldsymbol{R}^n & \rightarrow \quad \boldsymbol{R}^n \\ \quad (y, z) & \mapsto \quad t''(x)(y, z) \end{cases} \tag{11}$$

is the second derivative of $t = (t_i)$ at $x \in D$. In (10) we assume, that $t(x)''(y)$ is defined by

$$t''(x)(y) := \begin{pmatrix} y^T t_1''(x) \\ \vdots \\ y^T t_n''(x) \end{pmatrix} \in \boldsymbol{R}^{n\times n} \text{ for } x \in D \text{ and } y \in \boldsymbol{R}^n,$$

and in (11) we define $t(x)''(y,z)$ by

$$t''(x)(y,z) := \left(y^T t_1''(x) z, \dots , y^T t_n''(x) z\right) \in \boldsymbol{R}^n,$$

with

$$t_i''(x) := \left(\frac{\partial^2 t_i(x)}{\partial x_l\, \partial x_k}\right) = \begin{pmatrix} \frac{\partial^2 t_i(x)}{\partial x_1^2} & \frac{\partial^2 t_i(x)}{\partial x_2\, \partial x_1} & \cdots & \frac{\partial^2 t_i(x)}{\partial x_n\, \partial x_1} \\ \frac{\partial^2 t_i(x)}{\partial x_1\, \partial x_2} & \frac{\partial^2 t_i(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 t_i(x)}{\partial x_n\, \partial x_2} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 t_i(x)}{\partial x_1\, \partial x_n} & \frac{\partial^2 t_i(x)}{\partial x_2\, \partial x_n} & \cdots & \frac{\partial^2 t_i(x)}{\partial x_n^2} \end{pmatrix} = (\operatorname{grad} t_i(x))' \in \boldsymbol{R}^{n\times n}$$

being the Hessian associated with $t_i(x)$. Note that k counts the rows while l counts the columns.

The reason behind the choice of $[H]$ according to (10) is the Taylor expansion of t at $\widetilde{x} \in [x]$ which we write in the form

$$t(x) = t(\widetilde{x}) + \{t'(\widetilde{x}) + R(x,\widetilde{x})\}(x - \widetilde{x})$$

with the remainder term $R(x,\widetilde{x}) \cdot (x - \widetilde{x})$, where $R(x,\widetilde{x}) \in \boldsymbol{R}^{n\times n}$. We remind that according to [7], p. 284, and by applying the extended mean value theorem, the entries $r_{ij}(x,\widetilde{x})$ of $R(x,\widetilde{x})$ can be expressed as

$$\begin{aligned} r_{ij}(x,\widetilde{x}) &= (x-\widetilde{x})^T \left(\int_0^1 \frac{\partial^2 t_i}{\partial x_j\, \partial x_k}(\widetilde{x} + \tau(x-\widetilde{x}))\,(1-\tau)\,d\tau\right)_{k=1,\dots,n} \\ &= \int_0^1 (1-\tau)\,d\tau\,(x-\widetilde{x})^T \left(\frac{\partial^2 t_i}{\partial x_j\, \partial x_k}(\xi^{(ijk)})\right)_{k=1,\dots,n} \\ &= \frac{1}{2}\,(x-\widetilde{x})^T \left(\frac{\partial^2 t_i}{\partial x_j\, \partial x_k}(\xi^{(ijk)})\right)_{k=1,\dots,n} \end{aligned}$$

with $\xi^{(ijk)} \in \boldsymbol{R}^n$ between x and $\widetilde{x}$ for i, j, $k = 1, \dots n$. Hence

$$R(x,\widetilde{x}) \in \frac{1}{2}\,\left(([x]-\widetilde{x})^T t_i''([x]\underline{\cup}\widetilde{x})\right) = [H]([x],\widetilde{x})$$

and

$$t(x) \in [g]([x], \widetilde{x})$$

for all $x \in [x]$ and $[H]$ from (10). In addition, $[H]([x], \widetilde{x})$ is continuous and inclusion monotone since the function $c([x]) := [x] \underline{\cup} \widetilde{x}$ has these properties and since $[H]([x], \widetilde{x})$ can be interpreted as the interval arithmetic evaluation of the expression $\frac{1}{2}\left((x-\widetilde{x})^T t''(c(x))\right)$; cf.[5] or [13] for details.

Assume now that a function $f : D \subseteq \boldsymbol{R}^n \to \boldsymbol{R}^n$ is given and that $[x] \subseteq D$. We are interested in the zeros $x^* \in [x]$ of f. To this end use the transformation

$$t(x) := x - Cf(x), \quad C \in \boldsymbol{R}^{n\times n} \text{ nonsingular and independent of } x. \tag{12}$$

Then the zeros of f are the fixed points of t and vice versa. With $t'(x) = I - Cf'(x)$ and with $[H]$ from (10) we get

$$[g]([x], \widetilde{x}) = \widetilde{x} - Cf(\widetilde{x}) + \left\{ I - Cf'(\widetilde{x}) - \frac{1}{2}(Cf)''([x]\underline{\cup}\widetilde{x})([x] - \widetilde{x}) \right\} ([x] - \widetilde{x}). \tag{13}$$

For $\widetilde{x} \in [x]$ this is just the function k_3 in [5], p. 239, and k_7 in [14], p. 30. Therefore, it is not astonishing that Theorem 1a) and b) reduces to similar results as in [14], § 4, and in [5], § 19. For a comparison take into account the factor $\frac{1}{2}$ in (12). But note that we will also use $[H]$ and $[g]$ in Example 2 in a different meaning as in (10) and (13). This is caused by the possibility of representing the remainder term $R(x, \widetilde{x})(x - \widetilde{x})$ in different ways, as is also shown in the following example.

Example 1 Let $f(x) := \begin{pmatrix} x^2 - 2xy + y^2 \\ 0 \end{pmatrix} \in \boldsymbol{R}^2$, $C := I$, $t(x) := x - Cf(x)$. Then

$$f''(x)(y, z) = \begin{pmatrix} y^T \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix} z \\ 0 \end{pmatrix} = \begin{pmatrix} 2y_1(z_1 - z_2) + 2y_2(-z_1 + z_2) \\ 0 \end{pmatrix}$$

for all x, y, $z \in \boldsymbol{R}^2$, whence

$$f''(x)(y) = \begin{pmatrix} y^T \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix} \\ \begin{matrix} 0 & 0 \end{matrix} \end{pmatrix} \in \boldsymbol{R}^{2\times 2}. \tag{14}$$

Let

$$A(y) := \begin{pmatrix} y^T \begin{pmatrix} 2 & -4 \\ 0 & 2 \end{pmatrix} \\ \begin{matrix} 0 & 0 \end{matrix} \end{pmatrix} \in \boldsymbol{R}^{2\times 2}. \tag{15}$$

Then apparently $f''(x)(y,y) = \begin{pmatrix} 2y_1^2 - 4y_1y_2 + 2y_2^2 \\ 0 \end{pmatrix} = A(y)y$ holds although $f''(x)(y) \neq A(y)$ and therefore $f''(x)(y,z) \neq A(y)z$, in general. So we can represent $t(x)$ by means of (14) as well as by means of (15). Choosing $[H]([x],\widetilde{x}) := \frac{1}{2}(Cf)''([x]\underline{\cup}\widetilde{x})([x]-\widetilde{x})$ and $[H]([x],\widetilde{x}) := \frac{1}{2}CA([x]-\widetilde{x})$, respectively, yields two different admissible interval functions $[g]$, where the second one differs from (13).

□

If $\widetilde{x}$ is a sufficiently good approximation of a zero of f, if $f'(\widetilde{x})^{-1}$ exists, and if $C \approx f'(\widetilde{x})^{-1}$ then $\alpha \approx 0$, $\beta \approx 0$, $\Delta \approx 1$ and $r^- \approx 0$, $r^+ \approx 1/\gamma$ for the quantities in Theorem 1. In particular, the assumptions $\beta < 1$ and $\Delta \geq 0$ of this theorem are fulfilled.

We want to apply now Theorem 1 with (2) and (12) to various problems of numerical analysis.

Example 2 (The generalized eigenproblem with a simple real eigenvalue)
Consider the generalized algebraic eigenproblem $Av = \lambda Bv$ as in § 1. Let $\widetilde{v} \in \boldsymbol{R}^n$ be an approximation of an eigenvector which belongs to an algebraic simple eigenvalue. Let $\widetilde{\lambda}$ be an approximation of this eigenvalue and use f from (1), $C \in \boldsymbol{R}^{(n+1)\times(n+1)}$ nonsingular, $\widetilde{x} := (\widetilde{v}^T, \widetilde{\lambda})^T \in \boldsymbol{R}^{n+1}$, and $[v] \in \boldsymbol{IR}^n$. In [15] the interval function

$$[g]([x],\widetilde{x}) := \widetilde{x} - C\,f(\widetilde{x}) + \left\{ I_{n+1} - C \begin{pmatrix} A - \widetilde{\lambda}B & -B[v]) \\ (e^{(i_0)})^T & 0 \end{pmatrix} \right\} ([x] - \widetilde{x}) \qquad (16)$$

was applied in order to verify eigenpairs of the generalized eigenproblem. With $t(x) = x - Cf(x)$ as in (13) one gets

$$t'(\widetilde{x}) = I_{n+1} - C \begin{pmatrix} A - \widetilde{\lambda}B & -B\widetilde{v} \\ (e^{(i_0)})^T & 0 \end{pmatrix}.$$

In [12] it was mentioned that for degenerate interval vectors $[x] \equiv x$ the expression $[g](x,\widetilde{x})$ from (16) is the complete Taylor expansion of $t(x)$ at $\widetilde{x}$ even if $\widetilde{x} \notin [x]$. Therefore, $t(x) \in [g]([x],\widetilde{x})$ holds trivially for all $x \in [x]$. Nevertheless $[g]([x],\widetilde{x})$ is not in the form (13), i. e., $[H]([x],\widetilde{x})$ is not given by (10). In fact, in order to obtain the representation (2) we have to define

$$[H]([x],\widetilde{x}) := C \begin{pmatrix} O & B([v] - \widetilde{v}) \\ O & 0 \end{pmatrix} \in \boldsymbol{IR}^{(n+1)\times(n+1)}. \qquad (17)$$

One can recover this function if one expresses the last, i. e. third, Taylor summand $\frac{1}{2}t''(x)\,(x - \widetilde{x}, x - \widetilde{x})$ appropriately and if one evaluates this expression interval arithmetically. We want to show that $[H]$ fulfills (4) and (6). From (17) we get

$$|[H]([x],\widetilde{x})| \le |C| \begin{pmatrix} O & |B|\,|[v]-\widetilde{v}| \\ O & 0 \end{pmatrix}$$

and

$$d([H]([x],\widetilde{x})) = |C| \begin{pmatrix} O & |B|d([v]) \\ O & 0 \end{pmatrix}.$$

With $\gamma := \|\, |C| \begin{pmatrix} |B| \\ 0 \end{pmatrix} \|$ this implies

$$\begin{aligned} \|\,|[H]([x],\widetilde{x})|\,\| &\le \max_i \sum_{k=1}^{n} |c|_{ik} \sum_{j=1}^{n} |b|_{kj} |[v]-\widetilde{v}|_j \\ &\le \max_i \sum_{j=1}^{n} \left(\sum_{k=1}^{n} |c|_{ik} |b|_{kj} \right) \|\,|[x]-\widetilde{x}|\,\| \\ &= \gamma \,\|\,|[x]-\widetilde{x}|\,\| \end{aligned}$$

and, analogously,

$$\|\, d([H]([x],\widetilde{x}))\,\| \le \gamma \|\, d([x])\,\|.$$

Therefore, Theorem 1 applies with $\hat{\gamma} := \gamma$.

For practical applications there is also a modification of $[g]$ which one gets by choosing C in the form

$$C = \begin{pmatrix} C_{11} & C_{12} \\ 0 & 1 \\ C_{21} & C_{22} \end{pmatrix} \tag{18}$$

with $C_{11} \in \boldsymbol{R}^{(n-1)\times n}$, $C_{12} \in \boldsymbol{R}^{n-1}$, $C_{21}^T \in \boldsymbol{R}^{n-1}$, $C_{22} \in \boldsymbol{R}$. Theorem 1 then reduces to a result in [4]. For details see [12].

□

Example 3 (The algebraic eigenproblem with a simple real eigenvalue)
Here we start again with f as in (1), where this time we choose $B := I$. The results in Example 2 remain true, of course, and are therefore omitted. They can be found, in [9]–[12] where they have been derived in a different way. For the particular choice of C in (18) they are already contained in [1].

We assumed that the eigenvalue λ^* to be enclosed is an algebraic simple one. This is due to the fact that only in this case $f'(x^*)^{-1}$ exists where $x^* = (v^{*T}, \lambda^*)^T$ is a corresponding eigenpair; cf. Theorem 2 in [12] for details. Thus, for a sufficiently good approximation $\widetilde{x}$ the inverse of $f'(\widetilde{x})$ exists, and $C \approx f'(\widetilde{x})^{-1}$ can be chosen as in the remark preceding the Example 2. □

Example 4 (Two–dimensional invariant subspaces)
In order to enclose double or nearly double eigenvalues, Alefeld and Spreuer verify in [6] a basis of a two–dimensional subspace of $\boldsymbol{R}^n$ which is invariant with respect to the linear mapping given by $A \in \boldsymbol{R}^{n\times n}$. To this end they start with the function

$$f(x) := \begin{pmatrix} Au - m_{11}u - m_{21}v \\ u_{i_1} - \varepsilon \\ u_{i_2} - \zeta \\ Av - m_{12}u - m_{22}v \\ v_{i_1} - \eta \\ v_{i_2} - \theta \end{pmatrix} \in \boldsymbol{R}^{2n+4}$$

where $x = (u^T, m_{11}, m_{21}, v^T, m_{12}, m_{22})^T \in \boldsymbol{R}^{2n+4}$, $i_1 \neq i_2 \in \{1, \ldots, n\}$ and $\varepsilon\theta - \zeta\eta \neq 0$. It is obvious that the vectors u^*, v^*, which are part of a zero $x^* = ((u^*)^T, m_{11}^*, m_{21}^*, (v^*)^T, m_{12}^*, m_{22}^*)^T$ of f, form a basis of such an invariant subspace. Note that u^*, v^* are unique within a fixed subspace because of the four normalization conditions which are hidden in f. Again we set $t(x) := x - Cf(x)$ with a nonsingular matrix $C \in \boldsymbol{R}^{(2n+4)\times(2n+4)}$, and we choose $\widetilde{x} = (\widetilde{u}^T, \widetilde{m}_{11}, \widetilde{m}_{21}, \widetilde{v}^T, \widetilde{m}_{12}, \widetilde{m}_{22})^T$ as an approximation of x^*. With

$$B := \begin{pmatrix} A - \widetilde{m}_{11}I_n & -\widetilde{u} & -\widetilde{v} & -\widetilde{m}_{21}I_n & 0 & 0 \\ (e^{(i_1)})^T & 0 & 0 & 0 & 0 & 0 \\ (e^{(i_2)})^T & 0 & 0 & 0 & 0 & 0 \\ -\widetilde{m}_{12}I_n & 0 & 0 & A - \widetilde{m}_{22}I_n & -\widetilde{u} & -\widetilde{v} \\ 0 & 0 & 0 & (e^{(i_1)})^T & 0 & 0 \\ 0 & 0 & 0 & (e^{(i_2)})^T & 0 & 0 \end{pmatrix} \in \boldsymbol{R}^{(2n+4)\times(2n+4)},$$

$$[T] := \begin{pmatrix} O & [u] - \widetilde{u} & [v] - \widetilde{v} & O & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ O & 0 & 0 & O & [u] - \widetilde{u} & [v] - \widetilde{v} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \boldsymbol{IR}^{(2n+4)\times(2n+4)}$$

and

$$[g]([x], \widetilde{x}) := \widetilde{x} - Cf(\widetilde{x}) + \{I_{2n+4} - C(B - [T])\}([x] - \widetilde{x})$$

we again have $t'(\widetilde{x}) = I_{2n+4} - CB$ and $t(x) \in [g]([x], \widetilde{x})$ for all x, $\widetilde{x} \in \boldsymbol{R}^n$. Note that $[g](x, \widetilde{x})$ is again the Taylor expansion of $t(x)$ at $\widetilde{x}$; cf. [12], for example. Therefore, Theorem 1 applies with

$$[H]([x], \widetilde{x}) := C[T]$$

and with

$$\hat{\gamma} := \gamma := \delta := \left\| \, 2|C|(e^T, 0, 0, e^T, 0, 0)^T \, \right\| \leq 2\| \, C \, \|, \quad e := (1, \ldots, 1)^T \in \boldsymbol{R}^n.$$

The results coincide with those in [6] and [12].

□

Example 5 (The singular value problem)

Each rectangular real matrix $A \in \boldsymbol{R}^{m\times n}$ can be represented as

$$A = V\Sigma U^T \iff AU = V\Sigma \iff A^TV = U\Sigma^T$$

with orthogonal matrices $U \in \boldsymbol{R}^{n\times n}$, $V \in \boldsymbol{R}^{m\times m}$ and with a rectangular diagonal matrix $\Sigma \in \boldsymbol{R}^{m\times n}$, where

$$(\Sigma)_{ij} := \begin{cases} 0 & \text{for } i \neq j \\ \sigma_i & \text{for } i = j \end{cases}, \ \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_{\min\{m,n\}} = 0.$$

The product $V\Sigma U^T$ is called a *singular value decomposition* of A, the positive values σ_i are called (non–trivial) *singular values* of A. Note that Σ is unique while U and V are not. The i–th singular value and the i–th columns u^i, v^i of U and V – so–called *singular vectors* – can be expressed as the zeros of the function

$$f(x) := \begin{pmatrix} Au - \sigma v \\ A^Tv - \sigma u \\ u^Tu - 1 \end{pmatrix},$$

where $x := (u^T, v^T, \sigma)^T$. If $x^* = ((u^*)^T, (v^*)^T, \sigma^*)^T$ is a zero of f with $\sigma^* \neq 0$ then

$$(v^*)^Tv^* = (v^*)^T\frac{1}{\sigma^*}Au^* = \frac{1}{\sigma^*}\left(A^Tv^*\right)^T u^* = (u^*)^Tu^* = 1.$$

Let $\widetilde{x} = (\widetilde{u}^T, \widetilde{v}^T, \widetilde{\sigma})^T$. In [12] a slight modification of the interval function

$$[g]([x], \widetilde{x}) := \widetilde{x} - Cf(\widetilde{x}) + \left\{ I_{m+n+1} - C\left(B - \begin{pmatrix} O & O & [v] - \widetilde{v} \\ O & O & [u] - \widetilde{u} \\ ([u] - \widetilde{u})^T & 0 & 0 \end{pmatrix} \right) \right\} ([x] - \widetilde{x})$$

was used with

$$B := \begin{pmatrix} A & -\widetilde{\sigma} I_m & -\widetilde{v} \\ -\widetilde{\sigma} I_n & A^T & -\widetilde{u} \\ 2\widetilde{u}^T & 0 & 0 \end{pmatrix}$$

in order to verify singular values σ and corresponding singular vectors u, v of A. It is an easy task to prove that $[g](x, \widetilde{x})$ is again the complete Taylor expansion of $t(x) := x - Cf(x)$ at $x = \widetilde{x}$ with $t'(\widetilde{x}) = I - CB$. As in Example 2 one easily checks that (4) and (6) hold for

$$[H]([x], \widetilde{x}) := C \begin{pmatrix} O & O & [v] - \widetilde{v} \\ O & O & [u] - \widetilde{u} \\ ([u] - \widetilde{u})^T & 0 & 0 \end{pmatrix}.$$

with

$$\gamma := \delta := \| \, |C| \cdot (1, \ldots, 1, n)^T \, \|.$$

Therefore, Theorem 1 applies with $\hat{\gamma} = \gamma$. Cf. also [2].

□

Example 6 (Quadratic systems)

In this example we consider systems of equations of the form

$$t(x) := b + Ax + T(x, x) = x$$

with b, $x \in \boldsymbol{R}^n$, $A \in \boldsymbol{R}^{n \times n}$ and with

$$T(x, y) := \left(\sum_{j=1}^{n} \sum_{k=1}^{n} t_{ijk} x_k y_j \right)_{i=1,\ldots,n}.$$

Note that $t(0) = b$, $t'(0) = A$ and $t''(x)(y, z) = 2T(y, z)$. Let

$$\begin{aligned} [H]([x], 0) &:= T([x]) := \left(\sum_{k=1}^{n} t_{ijk} [x]_k \right) \in \boldsymbol{IR}^{n \times n}, \\ [g]([x], 0) &:= b + (A + T([x]))[x]. \end{aligned}$$

In particular, $\widetilde{x} = 0$ in this example. One easily sees that (4) and (6) hold with

$$\gamma := \delta := \| \left(\sum_{k=1}^{n} |t_{ijk}| \right) \| = \max_{1 \le i \le n} \left\{ \sum_{j=1}^{n} \sum_{k=1}^{n} |t_{ijk}| \right\} \tag{19}$$

Therefore, Theorem 1 applies with $\alpha := \|b\|$, $\beta := \|A\|$, with γ, δ as in (19) and with $\hat{\gamma} = \gamma$. Its results coincide with those in [3] and [12].

□

References

[1] **Alefeld, G. :** *Berechenbare Fehlerschranken für ein Eigenpaar unter Einschluß von Rundungsfehlern bei Verwendung des genauen Skalarprodukts.* Z. Angew. Math. Mech. **67**, 145–152 (1987)

[2] **Alefeld, G. :** *Rigorous error bounds for singular values of a matrix using the precise scalar product.* In: Kaucher, E., Kulisch, U., Ullrich, Ch. (eds.): *Computerarithmetic.* pp. 9–130. Stuttgart 1987

[3] **Alefeld, G. :** *Errorbounds for quadratic systems of nonlinear equations using the precise scalar product.* In: Kulisch, U., Stetter, H. J. (eds.): *Scientific Computation with Automatic Result Verification.* Computing, Suppl. **6**, 59–68 (1988)

[4] **Alefeld, G. :** *Berechenbare Fehlerschranken für ein Eigenpaar beim verallgemeinerten Eigenwertproblem.* Z. Angew. Math. Mech. **68**, 181–184 (1988)

[5] **Alefeld, G.** and **Herzberger, J. :** *Introduction to Interval Computations.* New York 1983

[6] **Alefeld, G.** and **Spreuer, H. :** *Iterative improvement of componentwise errorbounds for invariant subspaces belonging to a double or nearly double eigenvalue.* Computing **36**, 321–334 (1986)

[7] **Heuser, H. :** *Lehrbuch der Analysis. Teil 2.* Stuttgart 1981

[8] **Hoffmann, R. :** *Konstruktion von Fehlerschranken bei der verallgemeinerten Singulärwertzerlegung und ihre iterative Verbesserung.* Thesis, Universität Karlsruhe 1993

[9] **Mayer, G. :** *Enclosures for eigenvalues and eigenvectors.* In: Atanassova, L., Herzberger, J. (eds.): *Computer Arithmetic and Enclosure Methods.* pp. 49–68. Amsterdam 1992

[10] **Mayer, G. :** *Taylor-Verfahren für das algebraische Eigenwertproblem.* Z. Angew. Math. Mech. **73**, T857–T860 (1993)

[11] **Mayer, G. :** *A unified approach to enclosure methods for eigenpairs.* Z. Angew. Math. Mech. **74**, 115–128 (1994)

[12] **Mayer, G. :** *Result verification for eigenvectors and eigenvalues.* In: Herzberger, J. (ed.): *Topics in Validated Computations.* Studies in Computational Mathematics **5**, pp. 209–276. Amsterdam 1994

[13] **Neumaier, A. :** *Interval Methods for Systems of Equations.* Cambridge 1990

[14] **Platzöder, L. :** *Einige Beiträge über die Existenz von Lösungen nichtlinearer Gleichungssysteme und Verfahren zu ihrer Berechnung.* Thesis, Berlin 1981

[15] **Rump, S. M. :** *Guaranteed inclusions for the complex generalized eigenproblem.* Computing **42**, 225–238 (1989)

received: September 11, 1995

Author:

Prof. Dr. G. Mayer
Universität Rostock
Fachbereich Mathematik
Universitätsplatz 1
18051 Rostock
Germany
e-mail: guenter.mayer@mathematik.uni-rostock.de

Rostock. Math. Kolloq. **49**, 89–97 (1995)

Subject Classification (AMS)
20C20, 16G70

Reinhard Knörr

Auslander–Reiten sequences and a certain ideal in mod–FG

Dedicated to the professors of mathematics
L. Berg, W. Engel, G. Pazderski, *and* H.- W. Stolle.

In [1], Auslander and Reiten defined an almost split sequence as a certain type of non–split short exact sequence of modules over artin algebras (the definition is repeated here in (19) below). Since then, this concept has been extensively used in the general representation theory of finite dimensional algebras, but also — for example by Erdmann [2] — in the much more restricted case of group algebras. In this situation, the most basic property of Auslander–Reiten sequences, namely their existence, turns out to be closely related to an old acquaintance, the so called projective maps. The purpose of the present note is to exhibit this relation, thereby giving a rather explicit construction of Auslander–Reiten sequences for group algebras. In fact, the construction can be relativized up to a point (Theorem 17). The reader not interested in this can always take $H = 1$ in what follows. For another treatment of the topic, see Greens paper [3].

Let G be a finite group, H a subgroup and F a field of characteristic $p > 0$. We consider finitely generated, right FG-modules $K, L, M, \ldots$ If L and M are two such modules, denote $(L, M)_G = Hom_{FG}(L, M)$ and $(L, M)_H = Hom_{FH}(L, M)$. We will use Greek letters $\alpha, \beta, \ldots$ for FG-homomorphims and Latin letters $a, b, \ldots$ for FH-homomorphisms; both are written on the right. So if $\alpha : L \to M$ and $b : M \to K$, then $\alpha b \in (L, K)_H$.

The relative trace map $T_H^G : (L, M)_H \to (L, M)_G$ by $T_H^G(a) = \sum_g g^{-1}ag$, where $G = \dot{\bigcup}_g Hg$, is simply denoted by T since G and H are fixed. We will, however, sometimes use the notation $T = T(L, M)$ to emphasize the pair of FG-modules involved. Note that T is well defined, i. e. independent of the choice of coset representatives; also, $T(\alpha b\gamma) = \alpha T(b)\gamma$ whenever $\alpha b\gamma$ is defined.

Finally, the trace map from $(M, M)_1$ into F is denoted by $tr = tr_M$. Note that for $a : M \to L$ and $b : L \to M$, one has $tr_M(ab) = tr_L(ba)$.

Now we define

(1) $I(M,L) = \{\alpha : M \to L \mid tr(\alpha b) = 0 \ \forall b \in Ker\ T(L,M)\}$.

It is clear that $I(M,L)$ is an F-subspace of $(M,L)_G$ and it is a trivial exercise to show:

(2) I is a categorial ideal, i. e. for all modules K, L, M, N, one has

$$I(M,L)\ (L,K)_G \subseteq I(M,K)$$

and

$$(N,M)_G\ I(M,L) \subseteq I(N,L).$$

It is well known that $Im\ T$ is also a categorial ideal; one has

(3) $Im\ T \subseteq I$.

Proof: Fix modules M and L and let $a : M \to L$. Then $T(a)$ is a typical element in $Im\ T(M,L)$, so we have to show that $tr\,(T(a)b) = 0$ whenever $b : L \to M$ and $T(b) = 0$. Choose coset representatives $G = \dot{\bigcup}_g Hg$ such that also $G = \dot{\bigcup}_g Hg^{-1}$; this is always possible as is well known. Then $tr\,(T(a)b) = \sum_g tr(g^{-1}agb) = \sum_g tr(agbg^{-1}) = tr\,(aT(b)) = 0$ since $T(b) = 0$.

(4) If $H = 1$, then $Im\ T = I$.

Proof: Let M and L be . Denote

$$\begin{aligned} n &= \dim\ (M,L)_1 = \dim(L,M)_1, \\ k_1 &= \dim\ Ker\ T(M,L) \qquad \text{and} \\ k_2 &= \dim\ Ker\ T(L,M). \end{aligned}$$

Note that $(a,b) = tr(ab)$ defines a non-degenerate bilinear map

$$(\ \ ) :\ (M,L)_1 \times (L,M)_1 \to F$$

and that by definition $I(M,L) = (M,L)_G\ \cap\ [Ker\ T(L,M)]^{\perp}$. Therefore

$$\begin{aligned} n - k_1 &= \dim(M,L)_1 - \dim\ Ker\ T(M,L) \\ &= \dim\ Im\ T(M,L) \\ &\leq \dim\ I(M,L) \qquad \text{by (3)} \\ &\leq \dim[\text{Ker}\ T(L,M)]^{\perp} \\ &= n - k_2 \qquad \text{, since } (\ ,\ ) \text{ is non-degenerate.} \end{aligned}$$

By symmetry, equality holds and in particular, $Im\ T(M,L) = I(M,L)$.

[For arbitrary $H \leq G$, equality does not hold in (3) in general.]

As usual, denote $sl(M) = \{f \in (M,M)_1 \mid trf = 0\}$.

(5) $I(M,M) = (M,M)_G$ iff $Ker\ T(M,M) \subseteq sl(M)$.

In particular, this is the case if M is H–projective.

Proof: Since $I(M,M)$ is an ideal in the ring $(M,M)_G$ by (2), one has

$$\begin{aligned} I(M,M) = (M,M)_G &\Leftrightarrow 1_M \in I(M,M) \\ &\Leftrightarrow 0 = tr(1_M\ b) = tr(b)\ \forall b \in Ker\ T(M,M) \\ &\Leftrightarrow Ker\ T(M,M) \subseteq sl(M). \end{aligned}$$

If M is H–projective, then $Im\ T(M,M) = (M,M)_G$ and the assertion follows from (3).

(6) M is projective iff $Ker\ T_1^G \subseteq sl(M)$.

Proof: Clear from (4) and (5).

(7) Let U be H–projective. For every $\alpha : U \rightarrow$, there is an $a : U \rightarrow U$ such that $T(a) = \alpha$. Define $t_0(\alpha) = tr(a)$; by (5), t_0 is well defined, i. e. independent of the choice of a. Clearly $t_0 : (U,U)_G \rightarrow F$ is F-linear.

Now lef M be an FG-module and let

(8)

$$0 \longrightarrow L \xrightarrow{\alpha} U \xrightarrow{\beta} M \longrightarrow 0$$

and

$$0 \longleftarrow L \xleftarrow{a} U \xleftarrow{b} M \longleftarrow 0$$

be the H–projective cover of M, i. e. the sequences are exact, L contains no H–projective direct summand $\neq 0$, U is H–projective and $a\alpha = 1_L\ \ b\beta = 1_M$. We fix this situation for the rest of this note.

(9) For any $\lambda : M \rightarrow L$, define $t(\lambda) = t_0(\beta\lambda\alpha)$, so by (7), $t : (M,L)_G \rightarrow F$ is a well defined F-linear map.

(10) (a) Given

$$\begin{tikzcd} 0 \arrow[r] & L \arrow[r, "\alpha"] & U \arrow[r, "\beta"] & M \arrow[r] \arrow[d, "\rho"] & 0 \\ & & & K & \end{tikzcd}$$

one has $\rho \in I(M,K)$ iff $t(\rho\sigma) = 0 \quad \forall \sigma : K \to L$.

(b) Given

$$\begin{tikzcd}
0 \arrow[r] & L \arrow[r, "\alpha"] & U \arrow[r, "\beta"] & M \arrow[r] & 0 \\
& K \arrow[u, "\sigma"'] & & &
\end{tikzcd}$$

one has $\sigma \in I(K,L)$ iff $t(\rho\sigma) = 0 \quad \forall \rho : M \to K$.

Proof: (a) ”$\Rightarrow$” Let $\sigma : K \to L$. Since U is H–projective, there is an $f : U \to M$ such that $T(f) = \beta$. Define $r = \sigma\alpha f : K \to M$. Then $T(r) = \sigma\alpha T(f) = \sigma\alpha\beta = 0$. By assumption, $\rho \in I(M,K)$, hence $tr(\rho r) = 0$. Therefore $t(\rho\sigma) = t_0(\beta\rho\sigma\alpha) = tr(f\rho\sigma\alpha) = tr(\rho\sigma\alpha f) = tr(\rho r) = 0$.
”$\Leftarrow$” Let $r : K \to M$ with $T(r) = 0$. We have to show that $tr(\rho r) = 0$. Note that $rb : K \to U$, so $\eta = T(rb) : K \to U$ and $\eta\beta = T(rb\beta) = T(r) = 0$; hence $\sigma = \eta\alpha^{-1} : K \to L$ is well defined. By assumption, $0 = t(\rho\sigma) = t_0(\beta\rho\sigma\alpha) = t_0(\beta\rho\eta) = tr(\beta\rho rb) = tr(\rho rb\beta) = tr(\rho r)$.
(b) Similar.

(11) For any two modules, denote

$$\overline{(K,N)} = (K,N)_G \;/\; I(K,N)\ .$$

Write $S_K = \overline{(K,K)}$ and note that by (2), S_K is a ring and $\overline{(K,N)}$ is an S_K–S_N–bimodule.

(12) Given

$$\begin{tikzcd}
0 \arrow[r] & L \arrow[r, "\alpha"] \arrow[d, dotted, "\varphi^*"] & U \arrow[r, "\beta"] \arrow[d, dotted, "\varphi_1"] & M \arrow[r] \arrow[d, "\varphi"] & 0 \\
0 \arrow[r] & L \arrow[r, "\alpha"] & U \arrow[r, "\beta"] & M \arrow[r] & 0
\end{tikzcd}\quad ,$$

the map $f = \beta\varphi b : U \to U$ satisfies $f\beta = \beta\varphi$. Since U is H–projective, there exists $\varphi_1 : U \to U$ such that $\varphi_1\beta = \beta\varphi$. Then $\varphi^* = \alpha\varphi_1\alpha^{-1} : L \to L$ is well defined, and the above diagram commutes.
Note that φ is H–projective iff φ^* is H–projective:
If $\varphi = T(f)$ is H-projective, then $\varphi_2 = \beta T(fb) : U \to U$ and $\alpha\varphi_2 = 0$. Moreover, $\varphi_2\beta = \beta T(fb\beta) = \beta T(f) = \beta\varphi = \varphi_1\beta$, so $\eta = (\varphi_1 - \varphi_2)\alpha^{-1} : U \to L$ is well defined and H-projective since U is H-projective. It follows that $\alpha\eta\alpha = \alpha(\varphi_1 - \varphi_2) = \alpha\varphi_1 = \varphi^*\alpha$, so $\alpha\eta = \varphi^*$, since α is a monomorphism. Therefore φ^* is H-projective. The converse is shown

similarly.
Therefore, the map $\varphi \mapsto \varphi^*$ induces a ring isomorphism

$$\hat{\omega} : \ \hat{S}_M = (M,M)_G \ /Im\ T(M,M) \to \hat{S}_L.$$

By (2) and (3),

$$\hat{I}_M = I(M,M) \ / \ Im\ T(M,M)$$

is an ideal of $\hat{S}_M$. Not surprisingly,

(13)

$$\hat{\omega}(\hat{I}_M) = \hat{I}_L.$$

We show first

(14) Given the commutative diagram in (12) and $\lambda : M \to L$, one has

$$t(\varphi\lambda) = t(\lambda\varphi^*).$$

Proof: Take $f : U \to U$ such that $T(f) = \varphi_1$. Then $t(\varphi\lambda) = t_0(\beta\varphi\lambda\alpha) = t_0(\varphi_1\beta\lambda\alpha) = tr(f\beta\lambda\alpha) = tr(\beta\lambda\alpha f) = t_0(\beta\lambda\alpha\varphi_1) = t_0(\beta\lambda\varphi^*\alpha) = t(\lambda\varphi^*)$.

Proof: of (13). By (10a) with $K = M$, we have $\varphi \in I(M,M)$ iff $tr(\varphi\lambda) = 0$ for all $\lambda : M \to L$. By (14), this is equivalent to $t(\lambda\varphi^*) = 0$ for all such λ, and by (10b) with $K = L$, this is the case iff $\varphi^* \in I(L,L)$. The assertion follows.

From (13), it is clear that
(15) $S_M \simeq S_L$; the map $\varphi \mapsto \varphi^*$ induces a ring isomorphism $\omega : S_M \to S_L$.

(16) As usual, we denote $V^* = Hom_F(V,F)$ the dual of the F–space V. Note that if V is an R–S–bimodule for rings R and S, then V^* is an S–R–bimodule by

$$v(sw^*) = (vs)w^*$$

and

$$v(w^*r) = (rv)w^*$$

for $v \in V,\ w^* \in V^*,\ r \in R$ and $s \in S$.

(17) Theorem Let (8) be the H–projective cover of M and let K be an FG–module. Then

$$[\ ,\ ] : \ \overline{(M,K)} \times \overline{(K,L)} \to F \qquad \text{by} \qquad [\overline{\gamma}, \overline{\delta}] = t(\gamma\delta)$$

for $\gamma \in (M,K)_G$ and $\delta \in (K,L)_G$ is a well defined non–degenerate bilinear map. The map

$$\zeta : \overline{(K,L)} \to \overline{(M,K)}\ ^* \qquad \text{by} \qquad (\overline{\gamma})(\overline{\delta}\zeta) = [\overline{\gamma}, \overline{\delta}]$$

is an S_K–S_M–bimodule isomorphism. Also

$$\xi : \overline{(M,K)} \to \overline{(K,L)}\ ^* \qquad \text{by} \qquad (\overline{\delta})(\overline{\gamma}\xi) = [\overline{\gamma}, \overline{\delta}]$$

is an S_M–S_K–bimodule isomorphism.

Proof: The first statement is immediate from (10a) and (10b). It is then obvious that ζ is an F–isomorphism . The S_K–S_M structure needs some comment: $\overline{(M,K)}$ is an S_M–S_K–bimodule by (11), hence $\overline{(M,K)}\ ^*$ is an S_K–S_M–bimodule by (16). Again by (11), $\overline{(K,L)}$ is an S_K–S_L–bimodule; now $S_M \simeq S_L$ by (15), and we can use the isomorphism $\omega : S_M \to S_L$ to define an S_K–S_M–structure on $\overline{(K,L)}$.

Now let $\gamma \in (M,K)_G,\ \delta \in (K,L)_G,\ \eta \in (K,K)_G$ and $\varphi \in (M,M)_G$. Then

$$(\overline{\gamma})[(\overline{\eta}\overline{\delta})\zeta] = [\overline{\gamma}, \overline{\eta}\overline{\delta}] = t(\gamma\eta\delta)$$

and

$$\begin{aligned} (\overline{\gamma})[\overline{\eta}(\overline{\delta}\zeta)] &= (\overline{\gamma}\ \overline{\eta})(\overline{\delta}\zeta) \qquad \text{[see (16)]} \\ &= [\overline{\gamma\eta}, \overline{\delta}] \\ &= t(\gamma\eta\delta) \end{aligned}$$

so $(\overline{\eta}\overline{\delta})\zeta = \overline{\eta}(\overline{\delta}\zeta)$ and ζ is S_K–linear.

To see the S_M–linearity, note that by definition

$$\overline{\delta}\overline{\varphi} = \overline{\delta}(\overline{\varphi}\omega) = \overline{\delta}\overline{\varphi^*}\ .$$

Therefore

$$\begin{aligned} (\overline{\gamma})[(\overline{\delta}\overline{\varphi})\zeta] &= (\overline{\gamma})[(\overline{\delta}\overline{\varphi^*})\zeta] \\ &= [\overline{\gamma}, \overline{\delta}\overline{\varphi^*})] \\ &= t(\gamma\delta\varphi^*) \end{aligned}$$

while

$$\begin{aligned} (\overline{\gamma})[(\overline{\delta}\zeta)\overline{\varphi}] &= (\overline{\varphi\gamma})(\overline{\delta}\zeta) \qquad \text{[see (16)]} \\ &= [\overline{\varphi\gamma}, \overline{\delta}] \\ &= t(\varphi\gamma\delta)\ . \end{aligned}$$

Equality follows now from (14) with $\lambda = \gamma\delta : M \to L$.
The proof for ξ is similar.

(18) Corollary Let M be indecomposable and non–projective. Then there exists a non–projective map $\tau : M \to \Omega M$ such that for every module K and every map $\eta : K \to M$ which is not a split–epimorphism, the map $\eta\tau : K \to \Omega M$ is projective.

Proof: Take $H = 1$ in the above discussion; so in particular, in (8), $L = \Omega M$. By (4), $S_M \neq 0$ since M is not projective. But M is indecomposable, so S_M is local. Choose $0 \neq f \in (S_M)^*$ such that $J(S_M)f = 0$. By the theorem — with $K = M$ — , there is a $\tau : M \to \Omega M$ such that $\overline{\tau}\zeta = f$. Clearly, τ is not projective, since otherwise $\overline{\tau} = 0$, hence $f = 0$. If $\eta : K \to M$ is not a split–epimorphism and $\rho : M \to K$, then $\rho\eta : M \to M$ and in fact $\rho\eta \in J(M,M)_G$. Hence

$$0 = (\overline{\rho\eta})f = (\overline{\rho\eta})(\overline{\tau}\zeta) = [\overline{\rho\eta}, \overline{\tau}] = t(\rho\eta\tau)\ .$$

This holds for all $\rho : M \to K$, so by (10b) with $\sigma = \eta\tau : K \to \Omega M = L$, we conclude that $\eta\tau \in I(K, \Omega M)$. By (4), this means that $\eta\tau$ is projective.

(19) An *Auslander–Reiten sequence* is a non–split short exact sequence

$$0 \longrightarrow N \xrightarrow{\varphi} E \xrightarrow{\psi} M \longrightarrow 0$$

such that every map $\eta : K \to M$ which is not a split–epimorphism admits a factorization over ψ and such that M and N are indecomposable.

(20) Theorem (Auslander–Reiten) Let M be indecomposable and non-projective. Then there exists a unique (up to isomorphism of short exact sequences) Auslander–Reiten sequence

$$0 \longrightarrow N \longrightarrow E \longrightarrow M \longrightarrow 0\ .$$

Moreover, $N \simeq \Omega^2 M$.

Proof: The uniqueness is an easy exercise, so it is enough to produce an Auslander –Reiten sequence.

Consider

$$0 \longrightarrow \Omega M \longrightarrow P \longrightarrow M \longrightarrow 0$$

and

$$0 \longrightarrow \Omega^2 M \xrightarrow{\mu} Q \xrightarrow{\pi} \Omega M \longrightarrow 0$$

the projective covers, so $\Omega^2 M$ is indecomposable. Let τ be as in (18) and consider

$$\begin{tikzcd}
Q \arrow[r, "\pi"] & \Omega M \\
E \arrow[u, "\tau_1"] \arrow[r, "\psi"] & M \arrow[u, "\tau"]
\end{tikzcd}$$

the pull–back of (π, τ) , i. e.

$$\begin{aligned}
E &= \{(q,m) \in Q \oplus M \mid q\pi = m\tau\} \\
(q,m)\tau_1 &= q \\
(q,m)\psi &= m\ .
\end{aligned}$$

Note that ψ is an epimorphism since π is. Define $\varphi : \Omega^2 M \to E$ by $x\varphi = (x\mu, 0)$. Then the diagram

$$\begin{tikzcd}
 & 0 \arrow[r] & \Omega^2 M \arrow[r, "\mu"] \arrow[d, equal] & Q \arrow[r, "\pi"] & \Omega M \arrow[r] & 0 \\
(*) & 0 \arrow[r] & \Omega^2 M \arrow[r, "\varphi"] & E \arrow[u, "\tau_1"] \arrow[r, "\psi"] & M \arrow[u, "\tau"] \arrow[r] & 0
\end{tikzcd}$$

is exact and commutative. Moreover, $(*)$ is non–split, since otherwise, there is a $\sigma : M \to E$ such that $\sigma\psi = 1_M$, but then $(\sigma\tau_1)\pi = \sigma\psi\tau = \tau$, i. e. τ factors over Q, hence is projective, a contradiction. Finally, if $\eta : K \to M$ is not a split–epimorphism, then $\eta\tau : K \to \Omega M$ is projective by (18), so $\eta\tau$ factors over π :

$$\begin{tikzcd}
 & Q \arrow[r, "\pi"] & \Omega M \\
 & E \arrow[u, "\tau_1"] \arrow[r, "\psi"'] & M \arrow[u, "\tau"] \\
K \arrow[uur, "\gamma"] \arrow[ur, dotted, "\delta"] \arrow[urr, "\eta"']
\end{tikzcd}$$

Since E is the pull–back, there is $\delta : K \to E$ such that $\delta\psi = \eta$ (and $\delta\tau_1 = \gamma$) . This shows that $(*)$ is an Auslander–Reiten sequence.

References

[1] **Auslander, M.** and **Reiten, I. :** *Representation theory of Artin algebras III, almost split sequences.* Comm. Algebra **3**, 239–294 (1975)

[2] **Erdmann, K. :** *Blocks whose defect groups are Klein four groups : a correction.* J. Algebra **76** , 505–518 (1982)

[3] **Green, J. A. :** *Notes on almost split sequences I.* Preprint

received: September 13, 1995

Author:

Prof. Dr. R. Knörr
Universität Rostock
Fachbereich Mathematik
Universitätsplatz 1
18051 Rostock
Germany

Rostock. Math. Kolloq. **49**, 98–104 (1995)

Subject Classification (AMS)
20M10

Reinhard Strecker

$\mathcal{H}-$commutative $\Delta-$semigroups

Dedicated to the professors of mathematics
L. Berg, W. Engel, G. Pazderski, *and* H.- W. Stolle.

ABSTRACT. A $\Delta-$semigroup is a semigroup whose lattice of congruences is a chain with respect to inclusion. Schein [9] and Tamura [11] have investigated commutative $\Delta-$semigroups, Trotter [13] exponential, Nagy [4] weakly exponential and Bonzini and Cherubini Spoletini [2] finite $\Delta-$semigroups. The pupose of the present paper is to investigate archimedean weakly commutative and $\mathcal{H}-$commutative $\Delta-$semigroups.

KEY WORDS. weakly commutative semigroups, archimedean semigroups

1 Weakly commutative semigroups

The theory of congruences on semigroups is fundamental and the study of semigroups, whose lattice of congruences is simple (in some sense), is therefore relevant. Important semigroups such as the full transformation semigroup, symmetric inverse semigroups and symmetric groups, on finite sets, are all $\Delta-$semigroups (see [3], vol. II, page 247 and [12], examples 1 and 2).

A semigroup S is weakly commutative if, for each $a, b \in S$, there are $x, y \in S$ and an integer $k > 0$ such that $(ab)^k = xa = by$ ([6]). A semigroup S is $\mathcal{H}-$commutative if, for each $a, b \in S$, there are $w, z \in S^1$ with $ab = baw = zba$ [14].

Remark: A $\mathcal{H}-$commutative semigroup is weakly commutative. Groups are $\mathcal{H}-$commutative semigroups and nilsemigroups are weakly commutative.
A semigroup S is archimedean if for each $a, b \in S$ there is an integer $n > 0$, such that $a^n \in SbS$. An archimedean semigroup is semilattice indecomposable. A semigroup S is right archimedean [8] if for each $a, b \in S$ there is an element $u \in S$ and an integer i such that $a^i = bu$. Left archimedean semigroups are defined analogously. Clearly, every right archimedean semigroup is archimedean.

Lemma 1 *S is a weakly commutative archimedean semigroup iff S is right and left archimedean.*

Proof: See [1], exercise 4, page 34.

Lemma 2 *A weakly commutative archimedean semigroup S has at most one idempotent.*

Proof: Let e and f be idempotents in S. By Lemma 1 there is an element x with $e = fx$ and therefore $e = fx = ffx = fe$. By Lemma 1, S is left archimedean, hence there is an element y with $ye = f$. It follows that $f = ye = yee = fe$ and $e = f$.

Theorem 3 *A semigroup S is a weakly commutative archimedean semigroup with an idempotent iff S is an ideal extension of a group G by a nilsemigroup N.*

Proof: 1) Suppose that S is a weakly commutative archimedean semigroup with an idempotent e. Then, by Lemma 2, S has only one idempotent e. Let G be the maximal subgroup of S. For each $g \in G$ and $a \in S$ we have $ag = age$. Since S is weakly commutative, there is an element $w \in S$ and an integer $n > 0$ such that $(ag)^n = gw$. Then $e(ag)^n = egw = gw = (ag)^n$. By Lemma 1, S is right archimedean, hence there is an element u such that $agu = e$. By multiplikation on the right by u it follows from $e(ag)^n = (ag)^n$ and $age = ag$ that $e(ag)^{n-1} = (ag)^{n-1}$. Repeating the argument gives the equation $e(ag) = ag$. We have that eue is the right inverse of ag. Since S ist left archimedean, a similar argument gives that eue is the left inverse of ag and $ag \in G$, then G is a left ideal of S. In a similar way it follows that G is a right ideal. By the archimedean property and Lemma 1 for each element $a \in S$ there is a $v \in S$ and an integer $n > 0$ such that $a^n = ev \in G$. Thus G is an ideal, therefore the Rees factor semigroup S/G is a nilsemigroup.

2) Conversely, suppose that S is an ideal extension of a group G by a nilsemigroup N. Then for each $a, b \in S$ there are intergers $k, l, m \geq 2$ such that $a^k = g_1$, $b^l = g_2$, $(ab)^m = g_3$, $g_1, g_2, g_3 \in G$. Note that $b^l = g_2 = g_1 g_1^{-1} g_2 = a(a^{k-1}) g_1^{-1} b^l$, therefore S is right archimedean. We can prove, in a similar way, that S is left archimedean. Note that $(ab)^m = g_3 = g_2 g_2^{-1} g_3 = b^l g_2^{-1} g_3 = b(b^{l-1} g_2^{-1} g_3)$ and $(ab)^m = g_3 = g_3 g_1^{-1} g_1 = (g_3 g_1^{-1} a^{k-1})a$, therefore S is weakly commutative.

Theorem 4 *A semigroup S is a $\mathcal{H}-$commutative archimedean semigroup with an idempotent iff S is an ideal extension of a group G by a commutative nilsemigroup.*

Proof: Suppose that S is a $\mathcal{H}-$commutative archimedean semigroup with idempotent e. Then for each $a, b \in S$ there are $w, z \in S^1$ such that $ab = baw$, $ba = abz$. Hence $ab = baw = abzw$ and $ab = ab(zw)^n$ for any n. Since the Rees factor semigroup S/G is a nilsemigroup, by Theorem 3, we have $ab = ba$ or ab and $ba \in G$, so S/G is commutative.

Conversely, suppose S is an ideal extension of a commutative nilsemigroup by a group G. Then, by Theorem 3, we have to prove only that S is $\mathcal{H}-$commutative. If $ab \notin G$ for $a, b \in S$ then $ab = ba$. If $ab \in G$ then $ba \in G$ and we have $ab = b(a(ba)^{-1}a)b = (ab(ba)^{-1}b)a$ and S is $\mathcal{H}-$commutative.

Proposition 5 *A weakly commutative archimedean $\Delta-$semigroup with idempotent e is either a $\Delta-$nilsemigroup or a $\Delta-$group.*

Proof: By Theorem 3, S is an ideal extension of a group G by a nilsemigroup N. Assume S is not a group, then $|N| > 1$. Define a relation ρ by $a\rho b$ iff $ae = be$. Then ρ is an equivalence relation which is left compatible and $a\rho b$ implies $ag = bg$ for all $g \in G$. For all $x \in S$, $xe = g \in G$ and therefore $a\rho b$ implies $axe = bxe$, thus ρ is right compatible. The restriction of ρ to G is the equality. By definition of a $\Delta-$semigroup the Rees congruence ρ_G determined by G and the relation ρ are comparable. Assume $|G| > 1$. Then ρ is contained in ρ_G and so ρ is the equality on S/G and therefore ρ is the equality. For $s \notin G$ we have $se = h \in G$. Therefore $se = he$, $s\rho h$, $s = h$, in contradiction to $s \notin G$, $h \in G$. Therefore $|G| = 1$, $G = \{e\}$ and S is a nilsemigroup.

Lemma 6 *If a weakly commutative archimedean semigroup S contains an element a and elements $x, y \in S^1$ with $xy \in S$ such that $a = xay$, then S contains an idempotent.*

Proof: Assume $x, y \in S$. The proof is similar if one of x or y is 1. If $a = xay$ then $a = x^n a y^n$ for all integers $n > 0$. By Lemma 1, S is right and left archimedean. Hence there are elements z, w with $az = x^n$ and $wa = y^n$ for a suitable n. We have $a = azawa$, a is a regular element and $azaw$ is idempotent.

Corollary 7 *A $\mathcal{H}-$commutative archimedean semigroup is commutative or contains an idempotent.*

Proof: If $ab \neq ba$ then there are $z_1, z_2 \in S$ such that $ab = baz_1$, $ba = abz_2$. This implies $ab = abz_2z_1$ and by Lemma 6, S contains an idempotent.

Lemma 8 *In a weakly commutative archimedean semigroup S without idempotent,* Green*'s relation $\mathcal{J}$ is the equality.*

Proof: If $S^1aS^1 = S^1bS^1$ then $a = ubv$, $b = xab$ for some $x, y, u, v \in S^1$. Hence $a = uxayv$. Assume $a \neq b$ then $uv \in S$, $xy \in S$, therefore $uxyv \in S$. By Lemma 6, S contains an idempotent, which is a contradiction.

2 Weakly commutative and $\mathcal{H}-$commutative $\Delta-$semigroups

Proposition 9 *A weakly commutative archimedean $\Delta-$semigroup S without an idempotent is cancellative.*

Proof: In [7] it was shown that if the relation σ is defined by $a\sigma b$ iff $a^{n+1} = a^n b = ba^n$ and $b^{n+1} = b^n a = ab^n$ for some n, then σ is a congruence on a weakly commutative semigroup. We show that in our case σ is the equality. Denote by σ_x the Rees congruence modulo the ideal $I = S^1xS^1$. σ_x is comparable with σ. Assume $\sigma_x \subseteq \sigma$. Since $x\sigma_x x^2$ we have $x\sigma x^2$ and $x^{n+1} = x^n x^2 = x^{n+2}$ for some n. Then x^{n+1} is idempotent, a contradiction. Therefore we have $\sigma \subset \sigma_x$ for all $x \in S$. If $x\sigma y$ and $x \neq y$ then $y \in S^1xS^1$ and $x \in S^1yS^1$, so $y = uxv$, $x = wyz$ and $y = uwyzv$, where $u, v, w, z \in S^1$ and $uvwz \in S$. By Lemma 6, S contains an idempotent, a contradiction, so σ is the equality. Therefore S is separative (in the sense of [6]), archimedean and by [6] cancellative.

Corollary 10 *Any $\mathcal{H}-$commutative archimedean $\Delta-$semigroup S has an idempotent.*

Proof: If S has no idempotent then S is commutative. By [9] or [11] then S has idempotents, a contradiction.

In [7] it was shown that a weakly commutative semigroup is a semilattice of archimedean semigroups, and in [11] it was shown that a $\Delta-$semigroup is either semilattice indecomposable or the set-theoretical union of two semilattice indecomposable semigroups S_0 and S_1 with $S_0S_1 \subseteq S_0$, $S_1S_0 \subseteq S_0$, $S_0 \cap S_1 = \emptyset$. In the following we denote by S a weakly commutative semilattice decomposable $\Delta-$semigroup and by S_0, S_1 the archimedean components as above.

Lemma 11

a) *S_1 is a weakly commutative archimedean $\Delta-$semigroup.*

b) *S_0 is weakly commutative and archimedean.*

Proof: a) S_1 is a $\Delta-$semigroup because the Rees factor semigroup S/S_0 is a $\Delta-$semigroup. b) For each $a, b \in S$ there is an integer n and an element $w \in S$ with $(ab)^n = bw$. By multiplikation it follows that $(ab)^{n+1} = bwab$, and S_0 is an ideal, therefore $wab \in S_0$.

Remark. S_0 is not necessary a $\Delta-$semigoup.

Proposition 12 *If S_1 is a nilsemigroup, then $|S_1| = 1$.*

Proof: Let e be the idempotent of S_1. Then $S_0 \cup \{e\}$ is an ideal I of S. The Rees congruence ρ modulo I is comparable with the semilattice congruence. This implies that ρ is the universal ralation and $S_1 = \{e\}$.

Corollary 13 *If S_1 has an idempotent e, then S_1 is a $\Delta-$group.*

Proof: This follows immediately from Proposition 5 , Lemma 11 and Proposition 12.

Proposition 14 *If S_0 contains an idempotent, then S_0 is a nilsemigroup.*

Proof: By theorem 3 S_0 is an ideal extension of a group G_e by a nil semigroup (e is the unit of G_e). G_e is an ideal in S, since $G_e = (G_e)^2$. We denote the REES congruence of S modulo G_e by ρ and define the relation σ as follows

$$a\sigma b \quad \text{iff} \quad ae = be.$$

Then σ is an equivalence relation and is left compatible with multiplikation. For all $x \in S$, xe is an element from G_e, therefore $xe = xee = exe$ and $a\sigma b$ implies $axe = aexe = bexe = bxe$, so σ is right compatible. Suppose $\rho \subseteq \sigma$. For all $x \in S_1$, $xe = xee$ which means that $x\sigma xe$. Then for all $g, h \in G_e$ we have $ge = he = g = h$, $G_e = \{e\}$ and S_0 is a nilsemigroup.

Proposition 15 *If S is $\mathcal{H}-$commutative, then S_1 is a $\Delta-$group and his unit is the unit of S.*

Proof: By Corollaries 10 and 13 S_1 is a $\Delta-$group. Let e be the unit of S_1. Then eSe is an ideal I of S, because S is $\mathcal{H}-$commutative. $e \in S$ and the REES congruence is comparable with the semilattice congruence. Therefore $eSe = S$ and e is the unit of S.

Lemma 16 *If S is $\mathcal{H}-$commutative, then S_0 contains an idempotent.*

Proof: GREEN's relation $\mathcal{H}$ in a $\mathcal{H}-$commutative semigroup is a congruence [14] and S_1 is a group, hence S_1 has only one $\mathcal{H}-$class. If the REES congruence modulo S_0 is comparable with $\mathcal{H}$, then S_0 is a $\mathcal{H}-$class or $|S_1| = 1$. If S_0 is a $\mathcal{H}-$class, then $S^1a = S^1a^2$ for all $a \in S$. Then there exists $s \in S$ with $a = sa^2$ and $y \in S$ with $sa = ay = sa^2y = sasa$ and sa is idempotent. If $|S_1| = 1$ then S_0 is a $\mathcal{H}-$commutative archimedean $\Delta-$semigroup. By Corollary 10, S_0 has an idempotent.

Theorem 17 *Let S be a $\mathcal{H}-$commutative semilattice decomposable $\Delta-$semigroup. Then $S = G^0$, a $\Delta-$group with zero adjoined, or $S = N^1$, a commutative nilsemigroup with unit adjoined.*

Proof: By Proposition 15, S_1 is a group and the identity of S_1 is the identity of S and, by Lemma 16, S_0 contains an idempotent 0 and S_0 is a nilsemigroup [see Proposition 14]. Consider the relation

$$a\tau b \text{ iff there is a } g \in S_1 \text{ with } a = gb.$$

The relation τ is reflexive by Proposition 15, symmetric, transitive and right compatible. For every $x \in S$ there are $u, v \in S = S^1$ with $xg = ugx$, $gx = vgx$. Hence $xg = ugx = uvxg = (uv)^n xg$, and S_0 is nilpotent, therefore u, v can be chosen from S_1, and we have $xa = xgb = ugxb$ with $u \in S_1$, so τ is left compatible. S_1 is a congruence class modulo τ and τ is comparable with the REES congruence modulo S_0. Because S_0 is a nilsemigroup, we have $|S| = 1$ or $S_0 = N$ is a single class modulo τ. In the last case for every $s_0 \in S_0$ there is an $g \in S^1$ with $gs_0 = g^{-1}gs_0 = es_0 = s_0$. Therefore $|S_0| = 1$ and $S = G^0$.

If $S = N^1$ then for $a, b \in N$ there are x and x' with $ab = xba$ and $ba = x'ab$. This implies $ab = xx'ab$ and $ab = (xx')^n ab$. If x or $x' \in S_0$ then $(xx')^n = 0$ for suitable n, $ab = ba = 0$. If x and $x' \in S_1$, then $x = x' = e$ and S_0 is commutative.

References

[1] **Bogdanovič, S. :** *Semigroups with a System of Subsemigroups.* Novi Sad 1985

[2] **Bonzini** and **Cherubini Spoletini :** *Sui $\Delta-$semigruppi di Putcha.* Instituto Lombardo (rend. Sc.) A **114**, 179–194 (1980)

[3] **Clifford, A.H.** and **Preston, G.B. :** *The Algebraic Theory of Semigroups. I, II.* Providence, 1961 and 1967

[4] **Nagy, A. :** *Weakly exponential $\Delta-$semigroups.* Semigroup Forum **40**, 297–313 (1990)

[5] **Nagy, A. :** *The least separative congruence on a weakly commutative semigroup.* Czechoslovak. Math. J. **32**, 630–632 (1982)

[6] **Petrich, M. :** *Introduction to semigroups.* Columbus 1973

[7] **Pondělíček, B. :** *On weakly commutative semigroups.* Czechoslovak. Math. J. **25**, 20–23 (1975)

[8] **Putcha, M.S. :** *Band of t-archimedean semigroups.* Semigroup Forum **6**, 232–239 (1973)

[9] **Schein, B.M. :** *Commutative semigroups where congruences form a chain.* Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys. **17**, 523–527 (1969)

[10] **Strecker, R. :** *Über das Radikal $\mathcal{H}-$kommutativer Halbgruppen.* Math. Nachr. **68**, 49–57 (1975)

[11] **Tamura, T. :** *Commutative semigroups whose lattice of congruences is a chain.* Bull. Soc. Math. France **97**, 369–380 (1969)

[12] **Tamura, T.** and **Trotter, P.G. :** *Completely semisimple inverse $\Delta-$semigroups admitting principal series.* Pacific J. Math. **68**, 515–525 (1977)

[13] **Trotter, P.G. :** *Exponential $\Delta-$semigroups.* Semigroup Forum **12**, 313–331 (1976)

[14] **Tully, E.J.jr. :** *$\mathcal{H}-$commutative semigroups in which each homomorphism is uniquely determined by its kernel.* Pacific J. Math. **45**, 669–681 (1973)

received: September 13, 1995

Author:

Prof. Dr. R. Strecker
Universität Rostock
Fachbereich Mathematik
Universitätsplatz 1
18051 Rostock
Germany

Rostock. Math. Kolloq. **49**, 105–114 (1995)

Subject Classification (AMS)
42A10, 42A15

Jürgen Prestin; Kathi Selig

On the Gram Matrix of Translates of De la Vallée Poussin Kernels

Dedicated to the professors of mathematics
L. Berg, W. Engel, G. Pazderski, *and* H.- W. Stolle.

ABSTRACT. Extending a result of A. A. Privalov [2] we prove the diagonal dominance of the Gram matrix of bases of translates of de la Vallée poussin kernels. This can be used to show the uniform boundedness of a corresponding orthogonal projection operator.

KEY WORDS. De la Vallée Poussin Kernels, Gram Matrix, Orthogonal Projection

1 Introduction

In this note we investigate arithmetic means of Dirichlet kernels, namely the de la Vallée Poussin kernels for $N, M \in \mathbb{N}$ and $1 \le M \le N$, defined by

$$\begin{aligned}\varphi_N^M(x) \;&:=\; \frac{1}{\sqrt{2N}} + \sqrt{\frac{2}{N}} \sum_{\ell=1}^{N-M} \cos \ell x + \sum_{\ell=N-M+1}^{N+M-1} \frac{N+M-\ell}{M\sqrt{2N}} \cos \ell x \\ &=\; \begin{cases} \dfrac{\sin Nx \sin Mx}{2M\sqrt{2N}\sin^2 \frac{x}{2}}, & \text{for} \quad x \notin 2\pi\mathbb{Z}, \\ \sqrt{2N}, & \text{for} \quad x \in 2\pi\mathbb{Z}. \end{cases}\end{aligned}$$

For $s = 0, \dots, 2N-1$, the translates

$$\varphi_{N,s}^M(x) \;:=\; \varphi_N^M\Big(x - \frac{s\pi}{N}\Big)$$

satisfy the interpolatory condition $\varphi_{N,s}^M\left(\frac{k\pi}{N}\right) = \sqrt{2N}\delta_{k,s}$, for $k = 0, \dots, 2N-1$. Hence, the space of translates

$$V_N^M \;:=\; \operatorname{span}\left\{\varphi_{N,s}^M : s = 0, \dots, 2N-1\right\}$$

has the dimension $\dim V_N^M = 2N$ which is independent of M. For a further discussion of these functions and spaces and for corresponding multiresolution analyses and wavelet spaces see e.g. A. A. Privalov [2] and the authors [1].

2 The Gram matrix

Since the basis of translates $\{\varphi_{N,s}^M\}_{s=0}^{2N-1}$ is not an orthogonal one, we are interested in its Gram matrix

$$\boldsymbol{G}_N^M \;:=\; \left(\left\langle\varphi_{N,r}^M\,,\varphi_{N,s}^M\right\rangle\right)_{r,s=0}^{2N-1}\,,$$

with the inner product

$$\langle f\,,g\rangle \;:=\; \frac{1}{2\pi}\int\limits_0^{2\pi} f(x)\,g(x)\,dx\,.$$

As discussed in [1] the circulant matrix $\boldsymbol{G}_N^M$ can be diagonalized by means of the $2N$-th Fourier matrix; i.e., $\boldsymbol{G}_N^M = \overline{\boldsymbol{F}}_{2N}\,\boldsymbol{D}_N^M\,\boldsymbol{F}_{2N}$, where

$$\boldsymbol{F}_{2N} \;:=\; \frac{1}{\sqrt{2N}}\left(e^{-\frac{2\pi irs}{2N}}\right)_{r,s=0}^{2N-1}, \qquad \overline{\boldsymbol{F}}_{2N} \;=\; \left(\boldsymbol{F}_{2N}\right)^{-1}. \tag{1}$$

The diagonal matrix $\boldsymbol{D}_N^M = \operatorname{diag}\left(d_{N,r}^M\right)_{r=0}^{2N-1}$ contains the eigenvalues

$$d_{N,r}^M \;=\; \begin{cases} \dfrac{1}{2}+\left(\dfrac{N-r}{2M^2}\right)^2, & \text{if} \quad |N-r|<M\,,\\ 1\,, & \text{otherwise.}\end{cases}$$

From this representation one rewrites immediately

$$\langle\varphi_{N,r}^M\,,\varphi_{N,s}^M\rangle \;=\; \delta_{s,r}-\frac{(-1)^{s-r}}{4NM^2}\sum_{k=-M+1}^{M-1}(M^2-k^2)\,\cos\frac{k(s-r)\pi}{N}\,.$$

A. A. Privalov [2] computed the entries of $\boldsymbol{G}_N^M$ in a closed form, namely

$$\left\langle\varphi_{N,r}^M\,,\varphi_{N,r}^M\right\rangle \;=\; 1-\frac{M}{3N}+\frac{1}{12NM} \tag{2}$$

and, for $r\neq s$,

$$\left\langle\varphi_{N,r}^M\,,\varphi_{N,s}^M\right\rangle \;=\; (-1)^{r-s}\,\frac{2M\cos\frac{M(r-s)\pi}{N}\sin\frac{(r-s)\pi}{2N}-\sin\frac{M(r-s)\pi}{N}\cos\frac{(r-s)\pi}{2N}}{8NM^2\sin^3\frac{(r-s)\pi}{2N}}\,. \tag{3}$$

In order to investigate the quantity of the main diagonal dominance, we will estimate the difference

$$\rho(N,M) \;:=\; \left\langle \varphi_{N,r}^{M}\,,\varphi_{N,r}^{M}\right\rangle - \sum_{\substack{s=0\\ s\neq r}}^{2N-1} \left|\left\langle \varphi_{N,r}^{M}\,,\varphi_{N,s}^{M}\right\rangle\right| \,,$$

which is independent of r due to the circulant structure of the Gram matrix.

Our main objective of this paper is to prove the following inequality.

Theorem 1 *For $N,M \in \mathbb{N}$, with $4M|N$, we have*

$$\rho(N,M) \;>\; 0.28\,.$$

For special choices of the indices N,M the value of ρ can be exactly determined.

Remark: In case of the modified Dirichlet kernel ($M=1$), we have

$$\langle \varphi_{N,r}^{1}\,,\varphi_{N,s}^{1}\rangle \;=\; \delta_{s,r} - \frac{(-1)^{s-r}}{4N}\,,$$

which gives $\rho(N,1)=\frac{1}{2}$, for every N. For the Fejer kernel ($N=M$), from

$$\begin{aligned} \langle \varphi_{N,r}^{N}\,,\varphi_{N,r}^{N}\rangle \;&=\; \frac{2}{3}+\frac{1}{12N^2}\,,\\ \langle \varphi_{N,r}^{N}\,,\varphi_{N,s}^{N}\rangle \;&=\; \frac{1}{4N^2\sin^2\frac{(r-s)\pi}{2N}} \qquad \text{for } r\neq s\,, \end{aligned}$$

and from (15) one obtains $\rho(N,N)=\dfrac{1}{3}+\dfrac{1}{6N^2}$.

Note that from numerical computations one can suggest that

$$\frac{1}{3} \;<\; \rho(N,M) \;\le\; \frac{1}{2}$$

for all $N,M\in\mathbb{N}$, and that for fixed N the function $\rho(N,M)$ is monotone decreasing with respect to M, for $1\le M\le N$. In [2] A. A. Privalov proved

$$\rho(3M,M) \;>\; \frac{1}{3}\,.$$

Theorem 1 is an extension of his result to the case $4M|N$.

The main diagonal dominance of the Gram matrix shows the almost orthogonality of the basis $\{\varphi_{N,s}^{M}\}_{s=0}^{2N-1}$. We need Theorem 1 in order to estimate the boundedness of the orthogonal projection operator $P_N^M : C_{2\pi} \to V_N^M$. This is in particular interesting for $N=2^j$ and $M=2^{j-\lambda}$ for large $j\in\mathbb{N}$ and fixed $\lambda\in\mathbb{N}$.

Corollary 2 *The orthogonal projection P_N^M onto V_N^M, with $N/M = 2^\lambda \geq 4$, satisfies*

$$\|P_N^M\|_{C_{2\pi}\to C_{2\pi}} \;<\; (3\lambda+10)^2 \,.$$

Proof of the Corollary: Writing

$$P_N^M f \;=\; \sum_{k=0}^{2N-1} \epsilon_k \, \varphi_{N,k}^M \,, \tag{4}$$

we obtain by Hölder's inequality and by a well-known inequality for trigonometric polynomials (see A. F. Timan [3], Chap. 4.9)

$$\|P_N^M f\|_\infty \;\leq\; 2N(1+\frac{5\pi}{4})\,\|\varphi_{N,0}^M\|_1 \max_{0\leq k<2N} |\epsilon_k| \;=:\; 2N(1+\frac{5\pi}{4})\,\|\varphi_{N,0}^M\|_1\,|\epsilon_\ell|\,.$$

Taking the inner product with $\varphi_{N,\ell}^M$ in (4), we estimate

$$\begin{aligned}|\langle P_N^M f\,,\,\varphi_{N,\ell}^M\rangle| \;&=\; \left|\sum_{k=0}^{2N-1} \epsilon_k \langle \varphi_{N,k}^M\,,\,\varphi_{N,\ell}^M\rangle\right| \\ &\geq\; |\epsilon_\ell| \left(\langle \varphi_{N,\ell}^M\,,\,\varphi_{N,\ell}^M\rangle - \sum_{\substack{k=0\\k\neq\ell}}^{2N-1} |\langle \varphi_{N,k}^M\,,\,\varphi_{N,\ell}^M\rangle|\right).\end{aligned}$$

Applying Theorem 1 yields

$$|\epsilon_\ell| \;<\; 3.6\,|\langle P_N^M f\,,\,\varphi_{N,\ell}^M\rangle| \;=\; 3.6\,|\langle f\,,\,\varphi_{N,\ell}^M\rangle| \;\leq\; 3.6\,\|f\|_\infty\,\|\varphi_{N,\ell}^M\|_1\,.$$

Hence,

$$\|P_N^M f\|_\infty \;<\; 7.2\,N(1+\frac{5\pi}{4})\,\|\varphi_{N,0}^M\|_1^2\,\|f\|_\infty\,.$$

Using known results on the Lebesgue constants with respect to the de la Vallée Poussin kernel (see e.g. [1]) the Corollary is proved. ■

3 Proof of the Theorem

The proof is based on three Lemmata concerning the function

$$f_m(x) \;:=\; \frac{2m\cos 2mx\sin x - \sin 2mx\cos x}{\sin^3 x} \tag{5}$$

for any $m \in \mathbb{N},\ m \geq 2$. Since $\cos k(\pi - x) = (-1)^k \cos kx$ and $\sin k(\pi-x) = (-1)^{k-1}\sin kx$ for all $k \in \mathbb{N}$, we have

$$f_m(\pi - x) \;=\; f_m(x)\,, \tag{6}$$

i.e., f_m is symmetric in $[0,\pi]$ with respect to $\frac{\pi}{2}$.
Using the rule of l'Hôpital, we compute

$$f_m(0) \quad = \quad \frac{2m(1-4m^2)}{3} \,. \tag{7}$$

The first Lemma describes the behaviour of f_m in the interval $\left[0, \frac{\pi}{2m}\right]$.

Lemma 3 *For all $m \in \mathbb{N},\ m \geq 2$, the function f_m is negative and monotone increasing in $\left[0, \frac{\pi}{2m}\right]$.*

Proof: In order to prove the monotonicity of f_m we show that for all $0 < x < \frac{\pi}{2m}$

$$f_m^{'}(x) \quad > \quad 0\,.$$

We have

$$f_m^{'}(x) \quad = \quad \frac{g_m(x)}{\sin^4 x}\,,$$

with

$$g_m(x) \quad = \quad 3\sin 2mx\cos^2 x - 6m\cos 2mx\cos x\sin x - (4m^2-1)\sin 2mx\sin^2 x\,,$$

which has the Taylor expansion

$$g_m(x) \quad = \quad \frac{8m(4m^4-5m^2+1)}{15}\,x^5 + \frac{g_m^{(7)}(\theta)}{5040}\,x^7$$

for a certain $\theta \in \left(0, \frac{\pi}{2m}\right)$. We have to verify that $g_m(x) > 0$ in $\left(0, \frac{\pi}{2m}\right)$. If $g_m^{(7)}(\theta) \geq 0$, then it is evident. If $g_m^{(7)}(\theta) < 0$, then with

$$\begin{aligned} g_m^{(7)}(\theta) \quad = \quad & (2m^2-2)\big\{(11(2m)^6+100(2m)^4-48(2m)^2-64)\sin 2m\theta\sin 2\theta \\ & +(-48(2m)^5-80(2m)^3+128(2m))\cos 2m\theta\cos 2\theta + 2(2m)^7\cos 2m\theta\sin^2\theta\big\}, \end{aligned}$$

we estimate

$$-g_m^{(7)}(\theta) \quad < \quad \begin{cases} (2m^2-2)(2m)^5(48+5)\,, & \text{for } 0 < \theta \leq \frac{\pi}{4m}\,, \\ (2m^2-2)(2m)^5 2\pi^2\,, & \text{for } \frac{\pi}{4m} < \theta < \frac{\pi}{2m}\,. \end{cases}$$

Hence,

$$\begin{aligned} g_m(x) \quad & > \quad \frac{8m(4m^4-5m^2+1)}{15}x^5 - \frac{53(2m)^5(2m^2-2)}{5040}x^7 \\ & = \quad \frac{m(m^2-1)x^5}{1260}\big(672(4m^2-1) - 53(2m)^4x^2\big) \\ & \geq \quad \frac{m(m^2-1)x^5}{1260}\big(672(4m^2-1) - 53\pi^2(2m)^2\big) \quad > \quad 0, \end{aligned}$$

which proves that f_m is monotone increasing in the interval $\left[0, \frac{\pi}{2m}\right]$.

Since $f_m(0) < 0$ (see (7)) and

$$f_m\left(\frac{\pi}{2m}\right) \;=\; -\frac{2m}{\sin^2\frac{\pi}{2m}} \;<\; 0\,,$$

the monotonicity of f_m assures further that f_m is strictly negative in the interval $\left[0, \frac{\pi}{2m}\right]$. ■

In the second Lemma, we analyze the sign behaviour of f_m in the interval $\left[\frac{\pi}{2m}, \pi - \frac{\pi}{2m}\right]$.

Lemma 4 *Let $m \in \mathbb{N}$, with $m \geq 2$. Then for $\ell = 1, \ldots, m-1$, we have*

$$\begin{aligned} \left|f_m\left(\tfrac{\ell\pi}{2m}+x\right)\right| &= (-1)^\ell\, f_m\left(\tfrac{\ell\pi}{2m}+x\right) && \textit{for } 0 \leq x \leq \tfrac{\pi}{8m}\,, && (8)\\ \left|f_m\left(\tfrac{\ell\pi}{2m}+x\right)\right| &\leq (-1)^\ell\, f_m\left(\tfrac{\ell\pi}{2m}+x\right) + \tfrac{4m}{\pi\sin^2\left(\frac{\ell\pi}{2m}+x\right)} && \textit{for } \tfrac{\pi}{8m} < x \leq \tfrac{\pi}{4m}\,, && (9)\\ \left|f_m\left(\tfrac{\ell\pi}{2m}+x\right)\right| &= (-1)^{\ell+1}\, f_m\left(\tfrac{\ell\pi}{2m}+x\right) && \textit{for } \tfrac{\pi}{4m} \leq x \leq \tfrac{\pi}{2m}\,, && (10) \end{aligned}$$

and for $\ell = m, \ldots, 2m-2$, we have

$$\begin{aligned} \left|f_m\left(\tfrac{\ell\pi}{2m}+x\right)\right| &= (-1)^\ell\, f_m\left(\tfrac{\ell\pi}{2m}+x\right) && \textit{for } 0 \leq x \leq \tfrac{\pi}{4m}\,, && (11)\\ \left|f_m\left(\tfrac{\ell\pi}{2m}+x\right)\right| &\leq (-1)^{\ell+1}\, f_m\left(\tfrac{\ell\pi}{2m}+x\right) + \tfrac{4m}{\pi\sin^2\left(\frac{\ell\pi}{2m}+x\right)} && \textit{for } \tfrac{\pi}{4m} < x < \tfrac{3\pi}{8m}\,, && (12)\\ \left|f_m\left(\tfrac{\ell\pi}{2m}+x\right)\right| &= (-1)^{\ell+1}\, f_m\left(\tfrac{\ell\pi}{2m}+x\right) && \textit{for } \tfrac{3\pi}{8m} \leq x \leq \tfrac{\pi}{2m}\,. && (13) \end{aligned}$$

Proof: By (5) we can write

$$f_m\left(\frac{\ell\pi}{2m}+x\right) \;=\; \frac{(-1)^\ell\,(a-b)}{\sin^3\left(\frac{\ell\pi}{2m}+x\right)}\,,$$

with

$$a = 2m\;\cos 2mx\;\sin\left(\tfrac{\ell\pi}{2m}+x\right), \quad b = \sin 2mx\;\cos\left(\tfrac{\ell\pi}{2m}+x\right).$$

For all $\ell = 1, \ldots, m-1$ and $0 \leq x \leq \frac{\pi}{2m}$, we have

$$\frac{\pi}{2m} \;\leq\; \frac{\ell\pi}{2m} + x \;\leq\; \frac{\pi}{2}\,.$$

We use that $\sin x \geq \frac{2x}{\pi}$, and consider three intervals for x.

(i) For $0 \leq x \leq \frac{\pi}{8m}$, we estimate

$$\begin{aligned} a - b &= 2m\;\cos 2mx\;\sin\left(\frac{\ell\pi}{2m}+x\right) - \sin 2mx\;\cos\left(\frac{\ell\pi}{2m}+x\right)\\ &\geq 2m\;\sin 2mx\;\sin\frac{\pi}{2m} - \sin 2mx\;\cos\left(\frac{\ell\pi}{2m}+x\right)\\ &\geq \sin 2mx\;\left(2 - \cos\left(\frac{\ell\pi}{2m}+x\right)\right) \;>\; 0\,, \end{aligned}$$

which gives (8).
(ii) For $\frac{\pi}{8m} < x \le \frac{\pi}{4m}$, we have that $a, b > 0$ and so we get

$$\begin{aligned}\left|f_m\left(\frac{\ell\pi}{2m}+x\right)\right| &= \frac{|a-b|}{\sin^3\left(\frac{\ell\pi}{2m}+x\right)} \le \frac{a-b+2b}{\sin^3\left(\frac{\ell\pi}{2m}+x\right)}\\ &\le (-1)^\ell\, f_m\left(\frac{\ell\pi}{2m}+x\right)+\frac{2\cos\left(\frac{\ell\pi}{2m}+x\right)}{\sin^3\left(\frac{\ell\pi}{2m}+x\right)}\,.\end{aligned}$$

Further we estimate

$$\cot\left(\frac{\ell\pi}{2m}+x\right) < \frac{1}{\frac{\ell\pi}{2m}+x} \le \frac{2m}{\pi}\,,$$

from which (9) follows.
(iii) For $\frac{\pi}{4m} \le x \le \frac{\pi}{2m}$, we have $a \le 0$ and $b \ge 0$, i.e., $a - b \le 0$, which yields (10).

In order to obtain (11)–(13), by means of the symmetry (6) we rewrite

$$\begin{aligned}f_m\left(\frac{\ell\pi}{2m}+x\right) &= f_m\left(\pi-\left(\frac{\ell\pi}{2m}+x\right)\right)\\ &= f_m\left(\frac{(2m-(\ell+1))\pi}{2m}+\left(\frac{\pi}{2m}-x\right)\right).\end{aligned}$$

Thereby, (11), (12) and (13) follow from (10), (9) and (8), respectively. ■

The third Lemma contains special sums of values of f_m .

Lemma 5 *For all $m \in \mathbb{N}$, $m \ge 2$, and for $0 \le x \le \frac{\pi}{2m}$, we have*

$$\sum_{\ell=1}^{2m-2}(-1)^\ell\, f_m\left(\frac{\ell\pi}{2m}+x\right) = |f_m(x)| - \left|f_m\left(\frac{\pi}{2m}-x\right)\right|,$$

and

$$\sum_{\ell=1}^{2m-1}\left|f_m\left(\frac{\ell\pi}{2m}\right)\right| = |f_m(0)|\,.$$

Proof: We use the well-known equality

$$\sum_{\ell=0}^{2m-1}\frac{1}{\sin^2(\frac{x}{2}+\frac{\ell\pi}{2m})} = \frac{4m^2}{\sin^2 mx}\,, \tag{14}$$

from which, by differentiation,

$$\sum_{\ell=0}^{2m-1}\frac{\cos(\frac{x}{2}+\frac{\ell\pi}{2m})}{\sin^3(\frac{x}{2}+\frac{\ell\pi}{2m})} = \frac{8m^3\cos mx}{\sin^3 mx}$$

follows. Then,

$$\begin{aligned}
&\sum_{\ell=0}^{2m-1} (-1)^\ell\, f_m\left(\frac{\ell\pi}{2m}+x\right) \\
&= \sum_{\ell=0}^{2m-1} (-1)^\ell\, \frac{2m\cos(\ell\pi+2mx)\sin\left(\frac{\ell\pi}{2m}+x\right)-\sin(\ell\pi+2mx)\cos\left(\frac{\ell\pi}{2m}+x\right)}{\sin^3\left(\frac{\ell\pi}{2m}+x\right)} \\
&= 2m\cos 2mx \sum_{\ell=0}^{2m-1} \frac{1}{\sin^2\left(\frac{\ell\pi}{2m}+x\right)} - \sin 2mx \sum_{\ell=0}^{2m-1} \frac{\cos\left(\frac{\ell\pi}{2m}+x\right)}{\sin^3\left(\frac{\ell\pi}{2m}+x\right)} \\
&= \frac{2m\cos 2mx\ 4m^2}{\sin^2 2mx} - \frac{\sin 2mx\ 8m^3\cos 2mx}{\sin^3 2mx} \quad = \quad 0\,.
\end{aligned}$$

From Lemma 3 and from the symmetry (6) we further deduce

$$\begin{aligned}
\sum_{\ell=1}^{2m-2} (-1)^\ell\, f_m\left(\frac{\ell\pi}{2m}+x\right) &= f_m\left(\frac{(2m-1)\pi}{2m}+x\right) - f_m(x) \\
&= |f_m(x)| - \left|f_m\left(\frac{\pi}{2m}-x\right)\right|.
\end{aligned}$$

Using Lemma 3 and (8), (11) and (13) from Lemma 4, we conclude

$$\sum_{\ell=1}^{2m-1} \left|f_m\left(\frac{\ell\pi}{2m}\right)\right| = \sum_{\ell=1}^{2m-1} (-1)^\ell\, f_m\left(\frac{\ell\pi}{2m}\right) = -f_m(0) = |f_m(0)|\,. \qquad \blacksquare$$

Proof of Theorem 1 From (3) and definition (5) of f_m, for all $s = 1, \ldots, 2N-1$, we have

$$|\langle \varphi_{N,0}^M, \varphi_{N,s}^M\rangle| = \left|\frac{2M\cos\frac{Ms\pi}{N}\sin\frac{s\pi}{2N} - \sin\frac{Ms\pi}{N}\cos\frac{s\pi}{2N}}{8NM^2\sin^3\frac{s\pi}{2N}}\right| = \frac{1}{8NM^2}\left|f_M\left(\frac{s\pi}{2N}\right)\right|.$$

In view of the Remark on page 107 we can assume $M \geq 2$, and we set $K := N/M$. Replacing $s = K\ell + k$, where $k, \ell \in \mathbb{N}_0,\ 0 \leq k \leq K-1$ and $0 \leq \ell \leq 2M-1$, we can split

$$\begin{aligned}
\rho(N,M) &= \left\langle \varphi_{N,0}^M, \varphi_{N,0}^M\right\rangle - \sum_{s=1}^{2N-1} \left|\left\langle \varphi_{N,0}^M, \varphi_{N,s}^M\right\rangle\right| \\
&= \left\langle \varphi_{N,0}^M, \varphi_{N,0}^M\right\rangle - \sum_{\ell=1}^{2M-1} |\langle \varphi_{N,0}^M, \varphi_{N,K\ell}^M\rangle| - \sum_{k=1}^{K-1}\sum_{\ell=0}^{2M-1} |\langle \varphi_{N,0}^M, \varphi_{N,K\ell+k}^M\rangle|\,.
\end{aligned}$$

By Lemma 5 we have

$$\sum_{\ell=1}^{2M-1} |\langle \varphi_{N,0}^M, \varphi_{N,K\ell}^M\rangle| = \frac{1}{8NM^2}\sum_{\ell=1}^{2M-1}\left|f_M\left(\frac{\ell\pi}{2M}\right)\right| = \frac{1}{8NM^2}\,|f_M(0)|\,.$$

From Lemmata 4 and 5 and by symmetry (6) we obtain

$$\begin{aligned}
8NM^2 \sum_{k=1}^{K-1} \sum_{\ell=0}^{2M-1} |\langle \varphi_{N,0}^M, \varphi_{N,K\ell+k}^M \rangle| \;&=\; \sum_{k=1}^{K-1} \sum_{\ell=0}^{2M-1} \left| f_M\left(\frac{\ell\pi}{2M} + \frac{k\pi}{2N}\right)\right| \\
=\; \sum_{k=1}^{K-1} &\left(\left| f_M\left(\frac{k\pi}{2N}\right)\right| + \left| f_M\left(\frac{(2M-1)\pi}{2M} + \frac{k\pi}{2N}\right)\right| \right) + \sum_{k=1}^{K-1} \sum_{\ell=1}^{2M-2} \left| f_M\left(\frac{\ell\pi}{2M} + \frac{k\pi}{2N}\right)\right| \\
\leq\; \sum_{k=1}^{K-1} &\left(\left| f_M\left(\frac{k\pi}{2N}\right)\right| + \left| f_M\left(\frac{\pi}{2M} - \frac{k\pi}{2N}\right)\right| \right) \\
&+ \sum_{k=1}^{\frac{K}{2}} \sum_{\ell=1}^{2M-2} (-1)^\ell \, f_M\left(\frac{\ell\pi}{2M} + \frac{k\pi}{2N}\right) - \sum_{k=\frac{K}{2}+1}^{K-1} \sum_{\ell=1}^{2M-2} (-1)^\ell \, f_M\left(\frac{\ell\pi}{2M} + \frac{k\pi}{2N}\right) \;+\; S \\
=\; \sum_{k=1}^{K-1} &\left(\left| f_M\left(\frac{k\pi}{2N}\right)\right| + \left| f_M\left(\frac{\pi}{2M} - \frac{k\pi}{2N}\right)\right| \right) + \sum_{k=1}^{\frac{K}{2}} \left(\left| f_M\left(\frac{k\pi}{2N}\right)\right| - \left| f_M\left(\frac{\pi}{2M} - \frac{k\pi}{2N}\right)\right| \right) \\
&- \sum_{k=\frac{K}{2}+1}^{K-1} \left(\left| f_M\left(\frac{k\pi}{2N}\right)\right| - \left| f_M\left(\frac{\pi}{2M} - \frac{k\pi}{2N}\right)\right| \right) \;+\; S \\
=\; 4 \sum_{k=1}^{\frac{K}{2}-1} &\left| f_M\left(\frac{k\pi}{2N}\right)\right| + 2\left| f_M\left(\frac{\pi}{4M}\right)\right| \;+\; S \quad\leq\quad 2(K-1)\,|f_M(0)| \;+\; S\,,
\end{aligned}$$

where

$$S \;=\; \frac{4M}{\pi} \left(\sum_{\ell=1}^{M-1} \sum_{k=\frac{K}{4}+1}^{\frac{K}{2}} \frac{1}{\sin^2\left(\frac{\ell\pi}{2M} + \frac{k\pi}{2N}\right)} + \sum_{\ell=M}^{2M-2} \sum_{k=\frac{K}{2}+1}^{\frac{3K}{4}-1} \frac{1}{\sin^2\left(\frac{\ell\pi}{2M} + \frac{k\pi}{2N}\right)} \right).$$

From (14) we deduce

$$\sum_{\ell=1}^{2M-1} \frac{1}{\sin^2 \frac{\ell\pi}{2M}} \;=\; \frac{4M^2-1}{3} \;=\; 1 + 2\sum_{\ell=1}^{M-1} \frac{1}{\sin^2 \frac{\ell\pi}{2M}}\,, \tag{15}$$

which gives

$$\sum_{\ell=1}^{M-1} \frac{1}{\sin^2 \frac{\ell\pi}{2M}} \;=\; \frac{2}{3}(M^2-1)\,.$$

Hence,

$$S \;\leq\; \left(\frac{K}{2} - 1\right) \frac{8M(M^2-1)}{3\pi}\,.$$

In view of the monotonicity of f_M in $\left[0, \frac{\pi}{2M}\right]$ (see Lemma 3), together with (2), (7) and

$K \geq 4$, we finally estimate

$$\begin{aligned}\rho(N,M) &\geq 1-\frac{M}{3N}+\frac{1}{12NM}-\frac{1}{8NM^2}\left((2K-1)|f_M(0)|+S\right)\\ &\geq 1-\frac{M}{3N}+\frac{1}{12NM}-\frac{(2K-1)(4M^2-1)}{12NM}-\frac{(K-2)(M^2-1)}{6\pi NM}\\ &= \frac{1}{3}-\frac{1}{6\pi}+\frac{1}{6M^2}+\frac{M}{3\pi N}+\frac{K-2}{6\pi NM} \quad > \quad 0.28\,. \end{aligned}$$ ■

References

[1] **Prestin, J.** and **Selig, K. :** *Interpolatory and orthonormal trigonometric wavelets.* In: Zeevi, J. and Coifman, R. (Eds.): *Signal and Image Representation in Combined Spaces.* (to appear)

[2] **Privalov, A.A. :** *Ob odnom ortogonalnom trigonometričeskom basisje.* Mat. Sb. **182**, 3, 384-394 (1991)

[3] **Timan, A. F. :** *Theory of Approximation of Functions of a Real Variable.* Oxford 1963

received: September 15, 1995

Authors:

Dr. J. Prestin; Dipl.-Math. K. Selig
Fachbereich Mathematik
Universität Rostock
Universitätsplatz 1
18051 Rostock
Germany
e-mail: prestin@mathematik.uni-rostock.d400.de

Rostock. Math. Kolloq. **49**, 115–126 (1995)

Subject Classification (AMS)
41A25, 41A30, 42A99

Gerlind Plonka

Approximation Properties of Multi–Scaling Functions: A Fourier Approach

Dedicated to the professors of mathematics
L. Berg, W. Engel, G. Pazderski, *and* H.- W. Stolle.

ABSTRACT. In this paper, we consider approximation properties of a finite set of functions ϕ_ν ($\nu = 0, \dots, r-1$) which are not necessarily compactly supported, but have a suitable decay rate. Assuming that the function vector $\boldsymbol{\phi} = (\phi_\nu)_{\nu=0}^{r-1}$ is refinable, we sketch a new way, how to derive necessary and sufficient conditions for the refinement mask in Fourier domain.

KEY WORDS. Approximation Order, Refinement Mask, Strong-Fix Conditions

1 Introduction

For applications of multi–wavelets in finite element methods, the problem occurs, how to construct refinable vectors $\boldsymbol{\phi} := (\phi_\nu)_{\nu=0}^{r-1}$ ($r \in \mathbb{N}$) of functions with short support, such that algebraic polynomials of degree $< m$ ($m \in \mathbb{N}$) can be exactly reproduced by a linear combination of integer translates of ϕ_ν ($\nu = 0, \dots, r-1$).

In Heil, Strang and Strela [9] and in Plonka [13], the approximation properties of refinable function vectors $\boldsymbol{\phi} := (\phi_\nu)_{\nu=0}^{r-1}$ were studied in some detail. In particular, new necessary and sufficient conditions for the refinement mask of $\boldsymbol{\phi}$ could be derived. In [13], it could even be shown that the function vector $\boldsymbol{\phi}$ can only provide approximation order m if its refinement mask factorizes in a certain manner. For finding these results, [9] as well as [13] strongly used properties of doubly infinite matrices determined by the matrix coefficients occuring in the refinement equation (in time domain).

Now we want to sketch a way, how the necessary and sufficient conditions for the refinement mask of $\boldsymbol{\phi}$ can completely be derived in the Fourier domain.

As in [13], the functions ϕ_ν are allowed to have a noncompact support if they have a suitable decay rate. The main tool of our new approach is the so called superfunction, which is contained in the span of the integer translates of ϕ_ν ($\nu = 0, \ldots, r-1$) and already provides the same approximation order as $\boldsymbol{\phi}$. The results are applied to some multi–scaling functions ϕ_0, ϕ_1 first considered by Donovan, Geronimo, Hardin and Massopust [6, 7].

2 Notations

Let us introduce some notations. Consider the Hilbert space $L^2 = L^2(\mathbb{R})$ of all square integrable functions on $\mathbb{R}$. The Fourier transform of $f \in L^2(\mathbb{R})$ is defined by $\hat{f} := \int_{-\infty}^{\infty} f(x) e^{-ix\cdot}\, \mathrm{d}x$. The function vector $\boldsymbol{\phi}$ with elements in $L^2(\mathbb{R})$ is *refinable*, if $\boldsymbol{\phi}$ satisfies a refinement equation of the form

$$\boldsymbol{\phi} = \sum_{l\in\mathbb{Z}} \boldsymbol{P}_l\, \boldsymbol{\phi}(2\cdot - l) \qquad (\boldsymbol{P}_l \in \mathbb{R}^{r\times r}),$$

or equivalently, if $\boldsymbol{\phi}$ satisfies the Fourier transformed refinement equation

$$\hat{\boldsymbol{\phi}} = \boldsymbol{P}(\cdot/2)\, \hat{\boldsymbol{\phi}}(\cdot/2) \tag{1}$$

with $\hat{\boldsymbol{\phi}} := (\hat{\phi}_\nu)_{\nu=0}^{r-1}$ and with the *refinement mask* (*two–scale symbol*)

$$\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{\phi}} := \frac{1}{2} \sum_{l\in\mathbb{Z}} \boldsymbol{P}_l\, e^{-il\cdot}. \tag{2}$$

Note that $\boldsymbol{P}$ is an $(r \times r)$–matrix of 2π–periodic functions. The components ϕ_ν of a refinable function vector $\boldsymbol{\phi}$ are called *multi–scaling functions.*

Let $BV(\mathbb{R})$ be the set of all functions which are of bounded variation over $\mathbb{R}$ and normalized by

$$\lim_{|x|\to\infty} f(x) = 0, \quad f(x) = \frac{1}{2} \lim_{h\to 0}(f(x+h) + f(x-h)) \quad (-\infty < x < \infty).$$

If $\hat{f} \in L^1(\mathbb{R}) \cap BV(\mathbb{R})$, then the Poisson summation formula

$$\sum_{l\in\mathbb{Z}} f(l)\, e^{-iul} = \sum_{j\in\mathbb{Z}} \hat{f}(u + 2\pi j)$$

is satisfied (cf. Butzer and Nessel [3]). By $C(\mathbb{R})$, we denote the set of continuous functions on $\mathbb{R}$. For a measurable function f on $\mathbb{R}$ and $m \in \mathbb{N}$ let

$$\begin{aligned} \|f\|_p &:= \Big(\int_{-\infty}^{\infty} |f(x)|^p \;\mathrm{d}x\Big)^{1/p}, \\ |f|_{m,p} &:= \|\mathrm{D}^m f\|_p, \qquad \|f\|_{m,p} := \sum_{k=0}^{m} \|\mathrm{D}^k f\|_p. \end{aligned}$$

Here and in the following, D denotes the differential operator with respect to x $\mathrm{D} := \mathrm{d}/\mathrm{d}x$. Let $W_p^m(\mathbb{R})$ be the usual Sobolev space with the norm $\|\cdot\|_{m,p}$. The l^p–norm of a sequence $\boldsymbol{c} := \{c_l\}_{l\in\mathbb{Z}}$ is defined by $\|\boldsymbol{c}\|_{l^p} := (\sum_{l\in\mathbb{Z}} |c_l|^p)^{1/p}$.

For $m \in \mathbb{N}$, let $E_m(\mathbb{R})$ be the space of all functions $f \in C(\mathbb{R})$ with the decay property

$$\sup_{x\in\mathbb{R}}\{|f(x)|\,(1+|x|)^{1+m+\epsilon}\} < \infty \qquad (\epsilon > 0).$$

Let $l^2_{-m} := \{\boldsymbol{c} := (c_k) : \sum_{k=-\infty}^{\infty}(1+|k|^2)^{-m}\,|c_k|^2 < \infty\}$ be a weighted sequence with the corresponding norm

$$\|\boldsymbol{c}\|_{l^2_{-m}} := \left(\sum_{l=-\infty}^{\infty}(1+|l|^2)^{-m}\,|c_l|^2\right)^{1/2}.$$

Considering the functions $\phi_\nu \in E_m(\mathbb{R})$ $(\nu = 0,\ldots,r-1)$, we call the set $\mathcal{B}(\boldsymbol{\phi}) := \{\phi_\nu(\cdot - l) : l \in \mathbb{Z},\, \nu = 0,\ldots,r-1\}$ L^2_{-m}*–stable* if there exist constants $0 < A \le B < \infty$ with

$$A\sum_{\nu=0}^{r-1}\|\boldsymbol{c}_\nu\|^2_{l^2_{-m}} \le \|\sum_{\nu=0}^{r-1}\sum_{l\in\mathbb{Z}} c_{\nu,l}\,\phi_\nu(\cdot - l)\|^2_{L^2_{-m}} \le B\sum_{\nu=0}^{r-1}\|\boldsymbol{c}_\nu\|^2_{l^2_{-m}}$$

for any sequences $\boldsymbol{c}_\nu = \{\boldsymbol{c}_{\nu,l}\}_{l\in\mathbb{Z}} \in l^2_{-m}$ $(\nu = 0,\ldots,r-1)$. Here L^2_{-m} denotes the weighted Hilbert space $L^2_{-m} = \{f : \|f\|_{L^2_{-m}} := \|(1+|\cdot|^2)^{-m/2}\,f\|_2 < \infty\}$. Note that, if the functions ϕ_ν are compactly supported, then the (algebraic) linear independence of the integer translates of ϕ_ν $(\nu = 0,\ldots,r-1)$ yields the L^2_{-m}–stability of $\mathcal{B}(\boldsymbol{\phi})$. For $m = 0$, we obtain the well–known L^2*–stability* (*Riesz stability*).

For $\phi_\nu \in E_m(\mathbb{R})$ $(\nu = 0,\ldots,r-1)$, we say that $\boldsymbol{\phi}$ provides *controlled* L^p*–approximation order* m $(1 \le p \le \infty)$, if the following three conditions are satisfied:

For each $f \in W_p^m(\mathbb{R})$ there are sequences $\boldsymbol{c}_\nu^h = \{c_{\nu,l}^h\}_{l\in\mathbb{Z}}$ $(\nu = 0,\ldots,r-1; h > 0)$ such that for a constant c independent of h we have:

1^0 $$\|f - h^{-1/p}\sum_{\nu=0}^{r-1}\sum_{l\in\mathbb{Z}} c_{\nu,l}^h\,\phi_\nu(\cdot/h - l)\|_p \le c\,h^m\,|f|_{m,p}.$$

2^0 Furthermore,

$$\|\boldsymbol{c}_\nu^h\|_{l^p} \le c\,\|f\|_p \quad (\nu = 0,\ldots,r-1).$$

3^0 There is a constant δ independent of h such that for $l \in \mathbb{Z}$

$$\mathrm{dist}\,(lh, \mathrm{supp}\,f) > \delta \quad \Rightarrow \quad c_{\nu,l}^h = 0 \quad (\nu = 0,\ldots,r-1).$$

This notation of controlled L^p–approximation order, first introduced in Jia and Lei [11], is a generalization of the well–known definition of approximation order for compactly supported functions. In [11], the strong connection of controlled approximation order provided by $\boldsymbol{\phi}$ and the Strang–Fix conditions for $\boldsymbol{\phi}$ was shown. Note that, instead of using the definition of Jia and Lei [11], we also could take the definition of local approximation order by Halton and Light [8]. For our considerations the equivalence to the Strang–Fix conditions is important.

The theory of closed shift–invariant subspaces of $L^2(\mathbb{R})$, spanned by integer translates of a finite set of functions has been extensively studied (cf. e.g. de Boor, DeVore and Ron [1, 2]; Jia [10]). In particular, it has been shown that the approximation order provided by a vector $\boldsymbol{\phi}$ can already be realized by a finite linear combination

$$f = \sum_{\nu=0}^{r-1} \sum_{l \in \mathbb{Z}} a_{\nu l}\, \phi_\nu(\cdot - l). \qquad (a_{\nu l} \in \mathbb{R}).$$

We call f *superfunction* of $\boldsymbol{\phi}$.

3 Approximation by refinable function vectors

In this section we shall give a new approach to necessary and sufficient conditions for the refinement mask of a refinable vector $\boldsymbol{\phi}$ ensuring controlled L^p–approximation order m. In particular, we show, how a superfunction f of $\boldsymbol{\phi}$ (providing the same approximation order as $\boldsymbol{\phi}$) can be constructed by the coefficients which occur in the linear combinations of ϕ_ν reproducing the monomials.

In the following, let $r \in \mathbb{N}$ and $m \in \mathbb{N}$ be fixed. First we want to recall the result in [13] dealing with the connection between controlled L^p–approximation order, reproduction of polynomials and Strang–Fix conditions.

Theorem 1 (cf. [13]) *Let $\boldsymbol{\phi} = (\phi_\nu)_{\nu=0}^{r-1}$ be a vector of functions $\phi_\nu \in E_m(\mathbb{R})$, $\hat{\phi}_\nu \in L^1(\mathbb{R}) \cap BV(\mathbb{R})$. Further, let $\mathcal{B}(\boldsymbol{\phi})$ be L^2_{-m}–stable. Then the following conditions are equivalent:*
(a) *The function vector $\boldsymbol{\phi}$ provides controlled approximation order m $(m \in \mathbb{N})$.*
(b) *Algebraic polynomials of degree $< m$ can be exactly reproduced by integer translates of ϕ_ν, i.e., there are vectors $\boldsymbol{y}_l^n \in \mathbb{R}^r$ $(l \in \mathbb{Z};\ n = 0, \ldots, m-1)$ such that the series $\sum_{l \in \mathbb{Z}} (\boldsymbol{y}_l^n)^T \boldsymbol{\phi}(\cdot - l)$ are absolutely and uniformly convergent on any compact interval of $\mathbb{R}$ and*

$$\sum_{l \in \mathbb{Z}} (\boldsymbol{y}_l^n)^T \boldsymbol{\phi}(x - l) = x^n \qquad (x \in \mathbb{R};\ n = 0, \ldots, m-1).$$

(c) *The function vector $\boldsymbol{\phi}$ satisfies the Strang–Fix conditions of order m, i.e., there is a finitely supported sequence of vectors $\{\boldsymbol{a}_l\}_{l \in \mathbb{Z}}$, such that*

$$f := \sum_{l \in \mathbb{Z}} \boldsymbol{a}_l^T \boldsymbol{\phi}(\cdot - l)$$

satisfies

$$\hat{f}(0) \neq 0; \quad \mathrm{D}^n \hat{f}(2\pi l) = 0 \quad (l \in \mathbb{Z} \setminus \{0\}; n = 0, \dots, m-1).$$

The equivalence of (a) and (c) is already shown in Jia and Lei [11], Theorem 1.1. Further, (b) follows from (c) by [11], Corollary 2.3. For showing that (b) yields (c), in [13] the function

$$f := \sum_{k=0}^{m-1} \boldsymbol{a}_k^{\mathrm{T}} \, \boldsymbol{\phi}(\cdot + k), \tag{3}$$

is introduced. Here, the coefficient vectors $\boldsymbol{a}_k$ are determined by

$$(\boldsymbol{a}_0, \dots, \boldsymbol{a}_{m-1}) := (\boldsymbol{y}_0^0, \dots, \boldsymbol{y}_0^{m-1}) \, \boldsymbol{V}^{-1}$$

with the Vandermonde matrix $\boldsymbol{V} := (k^n)_{k,n=0}^{m-1}$. Hence we have

$$\boldsymbol{y}_0^n = \sum_{k=0}^{m-1} k^n \, \boldsymbol{a}_k \qquad (n = 0, \dots, m-1). \tag{4}$$

By Fourier transform of (3) we obtain

$$\hat{f}(u) = \boldsymbol{A}(u)^{\mathrm{T}} \, \hat{\boldsymbol{\phi}}(u)$$

with

$$\boldsymbol{A}(u) := \sum_{k=0}^{m-1} \boldsymbol{a}_k \, e^{iuk}. \tag{5}$$

That means, $\boldsymbol{A}(u)$ is an $(r \times r)$–matrix of trigonometric polynomials. Observe that by (4)

$$(\mathrm{D}^n \boldsymbol{A})(0) = \sum_{k=0}^{m-1} (ik)^n \, \boldsymbol{a}_k = i^n \, \boldsymbol{y}_0^n \quad (n = 0, \dots, m-1).$$

Using the Poisson summation formula it can be shown that f satisfies the conditions

$$(\mathrm{D}^\mu \hat{f})(2\pi l) = \delta_{0,l} \, \delta_{0,\mu} \quad (l \in \mathbb{Z}; \; \mu = 0, \dots, m-1)$$

and hence the Strang–Fix conditions of order m (cf. [13]). Observe that f in (3) is a superfunction of $\boldsymbol{\phi}$.

In the new proof for the following theorem, this superfunction will be the main tool.

Theorem 2 *Let $\boldsymbol{\phi} = (\phi_\nu)_{\nu=0}^{r-1}$ be a refinable vector of functions $\phi_\nu \in E_m(\mathbb{R})$, $\hat{\phi}_\nu \in L^1(\mathbb{R}) \cap BV(\mathbb{R})$. Further, let $\mathcal{B}(\boldsymbol{\phi})$ be L^2_{-m}–stable. Then the function vector $\boldsymbol{\phi}$ provides L^p–controlled approximation order m if and only if the refinement mask $\boldsymbol{P}$ of $\boldsymbol{\phi}$ in* (2) *satisfies the following*

conditions:
There are vectors $\boldsymbol{y}_0^k \in \mathbb{R}^r$; $\boldsymbol{y}_0^0 \neq \mathbf{0}$ $(k = 0, \ldots, m-1)$ *such that for* $n = 0, \ldots, m-1$ *we have*

$$\begin{aligned} \sum_{k=0}^{n} \binom{n}{k} (\boldsymbol{y}_0^k)^{\mathrm{T}} \, (2i)^{k-n} \, (\mathrm{D}^{n-k} \boldsymbol{P})(0) &= 2^{-n} \, (\boldsymbol{y}_0^n)^{\mathrm{T}}, && (6) \\ \sum_{k=0}^{n} \binom{n}{k} (\boldsymbol{y}_0^k)^{\mathrm{T}} \, (2i)^{k-n} \, (\mathrm{D}^{n-k} \boldsymbol{P})(\pi) &= \mathbf{0}^{\mathrm{T}}, && (7) \end{aligned}$$

where $\mathbf{0}$ *denotes the zero vector.*

Proof: Note that the conditions (6)–(7) can also be written in the form

$$\begin{aligned} \mathrm{D}^n [\boldsymbol{A}^T(2u) \, \boldsymbol{P}(u)]|_{u=0} &= (\mathrm{D}^n \boldsymbol{A})^T(0) \quad (n = 0, \ldots, m-1), && (8) \\ \mathrm{D}^n [\boldsymbol{A}^T(2u) \, \boldsymbol{P}(u)]|_{u=\pi} &= \mathbf{0}^T \quad (n = 0, \ldots, m-1), && (9) \end{aligned}$$

where $\boldsymbol{A}$, defined in (5), is the symbol of a superfunction f of $\boldsymbol{\phi}$ in (3). From Theorem 1 we know that $\boldsymbol{\phi}$ provides controlled approximation order m if and only if f satisfies the Strang–Fix conditions of order m. Hence we only have to prove:
The relations (8)–(9) are satisfied if and only if f satisfies the Strang–Fix conditions of order m, i.e.,

$$(\mathrm{D}^n \hat{f})(2\pi l) = c_n \, \delta_{0,l} \quad (n = 0, \ldots, m-1) \tag{10}$$

with constants $c_n \in \mathbb{R}$ and $c_0 \neq 0$.

1. We show that the relations (8)–(9) are satisfied if we have (10).

Note that by (1)

$$\hat{f}(2u) = \boldsymbol{A}^T(2u) \, \hat{\boldsymbol{\phi}}(2u) = \boldsymbol{A}^T(2u) \, \boldsymbol{P}(u) \hat{\boldsymbol{\phi}}(u).$$

Taking the derivatives, it follows on the one hand

$$\begin{aligned} (\mathrm{D}^n \hat{f})(u) &= \mathrm{D}^n [\boldsymbol{A}^T(u) \, \hat{\boldsymbol{\phi}}(u)] \\ &= \sum_{k=0}^{n} \binom{n}{k} (\mathrm{D}^k \boldsymbol{A}^T)(u) \, (\mathrm{D}^{n-k} \hat{\boldsymbol{\phi}})(u) \end{aligned}$$

and on the other hand

$$\begin{aligned} 2^n \, (\mathrm{D}^n \hat{f})(2u) &= \mathrm{D}^n [\boldsymbol{A}^T(2u) \, \boldsymbol{P}(u) \, \hat{\boldsymbol{\phi}}(u)] \\ &= \sum_{k=0}^{n} \binom{n}{k} \mathrm{D}^k [\boldsymbol{A}^T(2u) \, \boldsymbol{P}(u)] \, (\mathrm{D}^{n-k} \hat{\boldsymbol{\phi}})(u). \end{aligned}$$

2. Let us first show that the conditions (9) are satisfied. For all $l \in \mathbb{Z}$ we find by (10) that

$$0 = \hat{f}(4\pi l + 2\pi) = \boldsymbol{A}^T(4\pi l + 2\pi)\, \boldsymbol{P}(2\pi l + \pi)\, \hat{\boldsymbol{\phi}}(2\pi l + \pi) \; = \boldsymbol{A}^T(0)\, \boldsymbol{P}(\pi)\, \hat{\boldsymbol{\phi}}(2\pi l + \pi).$$

Hence, linear independence of the sequences $\{\hat{\boldsymbol{\phi}}_\nu(\pi + 2\pi l)\}_{l\in\mathbb{Z}}$ for $\nu = 0, \dots, r-1$ gives

$$\boldsymbol{A}^T(0)\, \boldsymbol{P}(\pi) = \boldsymbol{0}^T.$$

Note that the linear independence of the sequences $\{\hat{\boldsymbol{\phi}}_\nu(u + 2\pi l)\}_{l\in\mathbb{Z}}$ for all $u \in \mathbb{R}$ and for $\nu = 0, \dots, r-1$ is satisfied if and only if the integer translates of ϕ_ν form a L^2–stable basis of their closed span (cf. Jia and Micchelli [12]). This was the first step of the induction proof.

Let now $\mathrm{D}^\mu[\boldsymbol{A}^T(2u)\, \boldsymbol{P}(u)]|_{u=\pi} = \boldsymbol{0}^T$ be satisfied for $\mu = 0, \dots, n-1$ $(n < m)$, and observe that by assumption (10) $(\mathrm{D}^n\hat{f})(4\pi l + 2\pi) = 0$ for all $l \in \mathbb{Z}$. Then, by the linear independence of $\{\hat{\boldsymbol{\phi}}_\nu(\pi + 2\pi l)\}_{l\in\mathbb{Z}}$ for $\nu = 0, \dots, r-1$ and by

$$\begin{aligned} 0 &= 2^n\, (\mathrm{D}^n\hat{f})(4\pi l + 2\pi) = \sum_{k=0}^{n} \binom{n}{k} \mathrm{D}^k[\boldsymbol{A}^T(2u)\, \boldsymbol{P}(u)]|_{u=\pi}\, (\mathrm{D}^{n-k}\hat{\boldsymbol{\phi}})(\pi + 2\pi l) \\ &= \mathrm{D}^n[\boldsymbol{A}^T(2u)\, \boldsymbol{P}(u)]|_{u=\pi}\, \hat{\boldsymbol{\phi}}(\pi + 2\pi l) \end{aligned}$$

it follows that

$$\mathrm{D}^n[\boldsymbol{A}^T(2u)\, \boldsymbol{P}(u)]|_{u=\pi} = \boldsymbol{0}^T.$$

Thus, the relations (9) are satisfied.

3. Now we show that (10) yields (8). Let $u = 2\pi l$ $(l \in \mathbb{Z})$. Then we have on the one hand by the Strang–Fix conditions

$$\hat{f}(4\pi l) = \boldsymbol{A}^T(0)\, \boldsymbol{P}(0)\, \hat{\boldsymbol{\phi}}(2\pi l) = c_0\, \delta_{0,l}$$

and on the other hand

$$\hat{f}(2\pi l) = \boldsymbol{A}^T(0)\, \hat{\boldsymbol{\phi}}(2\pi l) = c_0\, \delta_{0,l}.$$

By linear independence of $\{\hat{\boldsymbol{\phi}}_\nu(2\pi l)\}_{l\in\mathbb{Z}}$ for $\nu = 0, \dots, r-1$ we obtain

$$\boldsymbol{A}^T(0)\, \boldsymbol{P}(0) = \boldsymbol{A}^T(0).$$

Again, we proceed by induction. Let now $\mathrm{D}^\mu[\boldsymbol{A}^T(2u)\, \boldsymbol{P}(u)]|_{u=0} = (\mathrm{D}^\mu\boldsymbol{A}^T)(0)$ be satisfied for $\mu = 0, \dots, n-1$ $(n < m)$ and observe that by assumption $(\mathrm{D}^n\hat{f})(2\pi l) = c_n\, \delta_{0,l}$ $(l \in \mathbb{Z},\ c_0 \neq 0)$. Then we find for all $l \in \mathbb{Z}$

$$\begin{aligned} 2^n(\mathrm{D}^n\hat{f})(4\pi l) &= \sum_{k=0}^{n} \binom{n}{k} \mathrm{D}^k[\boldsymbol{A}(2u)^T\, \boldsymbol{P}(u)]|_{u=0}\, (\mathrm{D}^{n-k}\hat{\boldsymbol{\phi}})(2\pi l) \\ &= \sum_{k=0}^{n-1} \binom{n}{k} (\mathrm{D}^k\boldsymbol{A}^T)(0)\, (\mathrm{D}^{n-k}\hat{\boldsymbol{\phi}})(2\pi l) \\ &\quad + \mathrm{D}^n[\boldsymbol{A}^T(2u)\, \boldsymbol{P}(u)]|_{u=0}\, \hat{\boldsymbol{\phi}}(2\pi l) = c_n\, \delta_{0,l}\,. \end{aligned}$$

On the other hand, for $l \in \mathbb{Z}$

$$(\mathrm{D}^n \hat{f})(2\pi l) = \sum_{k=0}^{n} \binom{n}{k} (\mathrm{D}^k \boldsymbol{A}^T)(0)\, (\mathrm{D}^{n-k} \hat{\boldsymbol{\phi}})(2\pi l) = c_n\, \delta_{0,l}.$$

Hence, a comparison yields

$$\mathrm{D}^n[\boldsymbol{A}^T(2u)\, \boldsymbol{P}(u)]|_{u=0}\, \hat{\boldsymbol{\phi}}(2\pi l) = (\mathrm{D}^n \boldsymbol{A}^T)(0)\, \hat{\boldsymbol{\phi}}(2\pi l)\,.$$

By linear independence of $\{\hat{\boldsymbol{\phi}}_\nu(2\pi l)\}_{l\in\mathbb{Z}}$ for $\nu = 0, \ldots, r-1$ we obtain

$$\mathrm{D}^n[\boldsymbol{A}^T(2u)\, \boldsymbol{P}(u)]|_{u=0} = (\mathrm{D}^n \boldsymbol{A}^T)(0).$$

Now the proof by induction is complete.

4. We are going to prove the reverse direction. Assume that the relations (8)–(9) are satisfied. We show that then the conditions $(\mathrm{D}^n \hat{f})(2\pi l) = c_n\, \delta_{0,l}$ $(n = 0, \ldots, m-1)$ hold, where $c_0 \neq 0$. For the μ-th derivative of $\hat{f}$ we find

$$\begin{aligned} 2^\mu\, (\mathrm{D}^\mu \hat{f})(4\pi l) &= \sum_{k=0}^{\mu} \binom{\mu}{k} \mathrm{D}^\mu[\boldsymbol{A}^T(2u)\, \boldsymbol{P}(u)]|_{u=0}\, (\mathrm{D}^{\mu-k} \hat{\boldsymbol{\phi}})(2\pi l) \\ &= \sum_{k=0}^{\mu} \binom{\mu}{k} (\mathrm{D}^\mu \boldsymbol{A})(0)\, (\mathrm{D}^{\mu-k} \hat{\boldsymbol{\phi}})(2\pi l) \\ &= (\mathrm{D}^\mu \hat{f})(2\pi l) \end{aligned}$$

and

$$\begin{aligned} 2^\mu\, (\mathrm{D}^\mu \hat{f})(4\pi l + 2\pi) &= \sum_{k=0}^{\mu} \binom{\mu}{k} \mathrm{D}^\mu[\boldsymbol{A}^T(2u)\, \boldsymbol{P}(u)]|_{u=\pi}\, (\mathrm{D}^{\mu-k} \hat{\boldsymbol{\phi}})(2\pi l + \pi) \\ &= 0. \end{aligned}$$

Thus, we indeed obtain $(\mathrm{D}^n \hat{f})(2\pi l) = c_n\, \delta_{0,l}$. It only remains to show that $c_0 \neq 0$. By Poisson summation formula and using the L^2–stability of $\boldsymbol{\phi}$ we have

$$\hat{\phi}(0) = \boldsymbol{A}^T(0)\, \hat{\boldsymbol{\phi}}(0) = (\boldsymbol{y}_0^0)^T\, \hat{\boldsymbol{\phi}}(0) = (\boldsymbol{y}_0^0)^T \sum_{l\in\mathbb{Z}} \boldsymbol{\phi}(\cdot - l) \neq 0.$$

Hence f satisfies the Strang–Fix conditions of order m. ■

Remark: For proving the second direction in Theorem 2 we do not need any stability condition if we assume that $(\boldsymbol{y}_0^0)^T\, \hat{\boldsymbol{\phi}}(0) \neq 0$. Since $\boldsymbol{y}_0^0$ and $\hat{\boldsymbol{\phi}}(0)$ are a left and a right eigenvector of $\boldsymbol{P}(0)$, respectively, this assumption is satisfied if the eigenvalue 1 of $\boldsymbol{P}(0)$ is simple.

4 The GHM–multi–scaling functions

We consider the example of a vector of two multi–scaling functions $\boldsymbol{\phi} := (\phi_0, , \phi_1)^T$ treated in Donovan, Geronimo, Hardin and Massopust ([6, 7]). In the special case $s = s_0 = s_1$ (with $s \in [-1, 1]$) of their construction, the refinement equation of $\boldsymbol{\phi}$ is given by

$$\boldsymbol{\phi}(x) = \boldsymbol{P}_0\, \boldsymbol{\phi}(2x) + \boldsymbol{P}_1\, \boldsymbol{\phi}(2x-1) + \boldsymbol{P}_2\, \boldsymbol{\phi}(2x-2) + \boldsymbol{P}_3\, \boldsymbol{\phi}(2x-3), \tag{11}$$

where

$$\begin{aligned}
\boldsymbol{P}_0 &:= \begin{pmatrix} -\dfrac{s^2-4s-3}{2\,(s+2)} & 1 \\ -\dfrac{3(s-1)(s+1)(s^2-3s-1)}{4\,(s+2)^2} & \dfrac{3s^2+s-1}{2\,(s+2)} \end{pmatrix}, \\
\boldsymbol{P}_1 &:= \begin{pmatrix} -\dfrac{s^2-4s-3}{2\,(s+2)} & 0 \\ -\dfrac{3(s-1)(s+1)(s^2-s+3)}{4\,(s+2)^2} & 1 \end{pmatrix}, \\
\boldsymbol{P}_2 &:= \begin{pmatrix} 0 & 0 \\ -\dfrac{3(s-1)(s+1)(s^2-s+3)}{4\,(s+2)^2} & \dfrac{3s^2+s-1}{2\,(s+2)} \end{pmatrix}, \\
\boldsymbol{P}_3 &:= \begin{pmatrix} 0 & 0 \\ -\dfrac{3(s-1)(s+1)(s^2-3s-1)}{4\,(s+2)^2} & 0 \end{pmatrix}.
\end{aligned}$$

For the refinement mask $\boldsymbol{P}$ we have

$$\boldsymbol{P}(u) := \frac{1}{2}\,(\boldsymbol{P}_0 + \boldsymbol{P}_1\, e^{-iu} + \boldsymbol{P}_2\, e^{-2iu} + \boldsymbol{P}_3\, e^{-3iu}).$$

Applying the result of Theorem 2 we can show that $\boldsymbol{\phi}$ provides the controlled L^p–approximation order $m = 2$:

Observing that

$$\begin{aligned}
\boldsymbol{P}(0) &= \begin{pmatrix} \dfrac{-s^2+4s+3}{2(s+2)} & \dfrac{1}{2} \\ \dfrac{-3(s-1)^3(s+1)}{2(s+2)^2} & \dfrac{3s^2+2s+1}{2(s+2)} \end{pmatrix}, \\
(\mathrm{D}\boldsymbol{P})(0) &= i \begin{pmatrix} \dfrac{s^2-4s-3}{4(s+2)} & 0 \\ \dfrac{9(s-1)^3(s+1)}{4(s+2)^2} & \dfrac{-(3s^2+2s+1)}{2(s+2)} \end{pmatrix},
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{P}(\pi) &= \begin{pmatrix} 0 & \frac{1}{2} \\ 0 & \frac{3(s^2-1)}{2(s+2)} \end{pmatrix}, \\
(\mathrm{D}\boldsymbol{P})(\pi) &= i \begin{pmatrix} \frac{-s^2+4s+3}{4(s+2)} & 0 \\ \frac{3(s^2-1)(-s^2+4s+3)}{4(s+2)^2} & \frac{-3(s^2-1)}{2(s+2)} \end{pmatrix},
\end{aligned}$$

we find with

$$\boldsymbol{y}_0^0 = \left(\frac{-3(s^2-1)}{s+2}, \, 1 \right), \qquad \boldsymbol{y}_0^1 = \left(\frac{-3(s^2-1)}{2(s+2)}, \, 1 \right)$$

the relations

$$(\boldsymbol{y}_0^0)^T \, \boldsymbol{P}(0) = (\boldsymbol{y}_0^0)^T, \quad (\boldsymbol{y}_0^0)^T \, \boldsymbol{P}(\pi) = \mathbf{0}^T$$

and

$$\begin{aligned}
(2i)^{-1} \, (\boldsymbol{y}_0^0)^T \, (\mathrm{D}\boldsymbol{P})(0) + (\boldsymbol{y}_0^1)^T \, \boldsymbol{P}(0) &= 2^{-1} \, (\boldsymbol{y}_0^1)^T, \\
(2i)^{-1} \, (\boldsymbol{y}_0^0)^T \, (\mathrm{D}\boldsymbol{P})(\pi) + (\boldsymbol{y}_0^1)^T \, \boldsymbol{P}(\pi) &= \mathbf{0}^T.
\end{aligned}$$

Hence, (8)–(9) are satisfied for $m = 2$. Knowing $\boldsymbol{y}_0^0$ and $\boldsymbol{y}_0^1$, we can construct a superfunction f of $\boldsymbol{\phi}$ (as defined in (3)) by

$$f(x) = (\boldsymbol{y}_0^0 - \boldsymbol{y}_0^1)^T \, \boldsymbol{\phi}(x) + (\boldsymbol{y}_0^1)^T \, \boldsymbol{\phi}(x+1)$$

obtaining

$$f(x) = \frac{3(1-s^2)}{2(s+2)} (\phi_0(x) + \phi_0(x+1)) + \phi_1(x+1).$$

Application of the refinement equation (11) on the right hand side yields

$$\begin{aligned}
f(x) &= \frac{9(1-s^2)}{4(s+2)} (\phi_0(2x) + \phi_0(2x+1)) + \frac{3(1-s^2)}{4(s+2)} (\phi_0(2x-1) + \phi_0(2x+2)) \\
&\quad + \frac{1}{2} (\phi_1(2x+2) + \phi_1(2x)) + \phi_1(2x+1) \\
&= \frac{1}{2} f(2x-1) + f(2x) + \frac{1}{2} f(2x+1).
\end{aligned}$$

That means, f itself satisfies the refinement equation of the hat–function $h(x) := \max\{(1 - |x|), \, 0\}$. Hence, taking a proper normalization constant, the superfunction f coincides with the hat function h. Indeed, in [6] the approximation order 2 provided by $\boldsymbol{\phi}$ was derived by showing that the hat–function h lies in the span of the integer translates of ϕ_0, ϕ_1.

References

[1] **de Boor, C., DeVore, R. A.** and **Ron, A. :** *Approximation from shift–invariant subspaces of $L_2(\mathbb{R}^d)$.* Trans. Amer. Math. Soc. **341**, 787–806 (1994)

[2] **de Boor, C., DeVore, R. A.** and **Ron, A. :** *The structure of finitely generated shift–invariant spaces in $L_2(\mathbb{R}^d)$.* J. Funct. Anal. **119(1)**, 37–78 (1994)

[3] **Butzer, P. L.** and **Nessel, R. J. :** *Fourier Analysis and Approximation.* Basel 1971

[4] **Chui, C. K. :** *An Introduction to Wavelets.* Boston 1992

[5] **Daubechies, I. :** *Ten Lectures on Wavelets.* Philadelphia 1992

[6] **Donovan G., Geronimo, J. S., Hardin, D. P.** and **Massopust, P. R. :** *Construction of orthogonal wavelets using fractal interpolation functions.* Preprint, Georgia Institute of Technology, Atlanta 1994

[7] **Geronimo, J. S., Hardin, D. P.** and **Massopust, P. R. :** *Fractal functions and wavelet expansions based on several scaling functions.* J. Approx Theory **78**, 373 – 401 (1994)

[8] **Halton, E. J.** and **Light, W. A. :** *On local and controlled approximation order.* J. Approx. Theory **72**, 268–277 (1993)

[9] **Heil, C., Strang, G.** and **Strela, V. :** *Approximation by translates of refinable functions.* Numer. Math. (to appear)

[10] **Jia, R. Q. :** *Shift–invariant spaces on the real line.* Proc. Amer. Math. Soc. (to appear)

[11] **Jia, R. Q.** and **Lei, J. J. :** *Approximation by multiinteger translates of functions having global support.* J. Approx. Theory **72**, 2–23 (1993)

[12] **Jia, R. Q.** and **Micchelli, C. A. :** *Using the refinement equations for the construction of pre–wavelets II: Powers of two.* In: Laurent, P. J., Le Méhauté, A., and Schumaker, L. L. (eds.): *Curves and Surfaces.* pp. 209–246, Boston 1991

[13] **Plonka, G. :** *Approximation order provided by refinable function vectors.* Constr. Approx. (to appear)

[14] **Strang, G.** and **Strela, V. :** *Short wavelets and matrix dilation equations.* IEEE Trans. Acoust. Speech Signal Process. **43**, 108–115 (1995)

received: September 13, 1995

Author:

Dr. G. Plonka
Universität Rostock
Fachbereich Mathematik
Universitätsplatz 1
18051 Rostock
Germany

Rostock. Math. Kolloq. **49**, 127–140 (1995)

Subject Classification (AMS)
65D30, 41A55, 65A10

Raimond Strauss

Numerische Integration von hypersingulären Integralen

Gewidmet den Herren Professoren
L. Berg, W. Engel, G. Pazderski *und* H.-W. Stolle.

Einleitung

Es wird ein Verfahren zur numerischen Integration von Integralen des Hadamardschen Typs

$$\mathop{=\!\!\!\!\!\!\int}_a^b \frac{f(t)}{(t-x)^2}\,dt\,, \quad a<x<b \tag{1}$$

angegeben, wobei nur die Kenntnis von Funktionswerten $f(t_i)$ an den Stellen $a=t_0<t_1<\ldots<t_N=b$ vorausgesetzt wird.

Die Quadraturgewichte werden aus bekannten Gewichten ([6], S. 183f) zur Quadratur von Cauchy-Hauptwert-Integralen gewonnen. Das Verfahren konvergiert unter schwachen Voraussetzungen. Die maximal erreichbare Konvergenzordnung ist $O(N^{-2}\ln N)$ auf jedem Teilintervall $[a_1,b_1]\subset(a,b)$. Sind neben den Funktionswerten auch die Ableitungen $f'(a)$ und $f'(b)$ bekannt, kann man eine Quadratur konstruieren, die auf (a,b) mit der gleichen Ordnung gleichmäßig konvergiert. Die Konvergenzordnung ist nur um den Faktor $\ln N$ schlechter als die maximal mögliche (vgl. [6], [3]). Anstelle der hier verwendeten Ausgangsgewichte für Cauchy-Hauptwert-Integrale kann man auch andere nutzen und durch analoges Vorgehen Quadraturverfahren für Hadamardsche Integrale herleiten, die eine höhere Konvergenzordnung haben. Eine Übertragung auf Integrale

$$\mathop{=\!\!\!\!\!\!\int}_a^b \frac{f(t)}{(t-x)^m}\,dt\,, \quad a<x<b \tag{2}$$

mit ganzzahligem $m > 2$ ist möglich und geschieht in einer weiterführenden Arbeit [8].

Vorausgesetzt wird die Hölderstetigkeit der ersten Ableitung von $f(t)$. Sie ist hinreichend für die Existenz des Integrals (1). Der Stetigkeitsmodul einer Funktion $f(t)$ wird mit $\omega(f, \delta)$ bezeichnet.

Für die behandelten Verfahren werden die Quadraturfehler abgeschätzt und die Fehlerordnung mit der optimalen verglichen. Schließlich werden Testrechnungen für die verschiedenen Quadraturen ausgeführt.

Eine Quadraturformel für Hadamardsche Integrale

Man kann das Integral (1) partiell integrieren und erhält (vgl. [5])

$$\int\!\!\!\!=_a^b \frac{f(t)}{(t-x)^2}\,dt = -\left[\frac{f(b)}{(b-x)} - \frac{f(a)}{(a-x)}\right] + \int\!\!\!\!-_a^b \frac{f'(t)}{t-x}\,dt\,. \tag{3}$$

Um den Wert für das gesuchte Hadamard Integral zu ermitteln, muß man das Cauchy-Hauptwert-Integral von $f'(t)$ auf der rechten Seite von (3) berechnen. Die dafür bekannten Methoden ließen sich sofort ausnutzen, wenn anstelle der Funktionswerte $f(t_i)$ an den Stellen $a = t_0 < t_1 < \ldots < t_N = b$ die entsprechenden Werte der Ableitung $f'(t_i)$ gegeben wären.

Zunächst wird das Cauchy-Hauptwert-Integral für $f(t)$ betrachtet.

Die Bezeichnungen

$$h_i = t_i - t_{i-1}, \quad \bar{h} = \max_i h_i \quad \text{und} \quad \underline{h} = \min_i h_i$$

werden vereinbart. Die Abhängigkeit der Größen $\bar{h}$ und $\underline{h}$ von N wird nicht gekennzeichnet. Für $N \to \infty$ soll $\bar{h}$ gegen Null gehen. Weiter wird verlangt, daß für jede Zerlegungsfolge der Quotient $\bar{h}/\underline{h}$ beschränkt bleibt, d.h., bei fortgesetzter Verfeinerung $(N \to \infty)$ soll eine von N unabhängige Konstante $H > 0$ mit der Eigenschaft

$$\frac{\bar{h}}{\underline{h}} \leq H$$

existieren.

Für $i = 0, 1, \ldots, N$ werden mit

$$\phi_i(t) = \begin{cases} (t - t_{i-1})/h_i & \text{für } a \le t_{i-1} \le t \le t_i \\ (t_{i+1} - t)/h_{i+1} & \text{für } t_i \le t \le t_{i+1} \le b, \\ 0 & \text{sonst} \end{cases}$$

die „Dachfunktionen“ bezeichnet. Ersetzt man die Funktion $f(t)$ durch die stückweise lineare Interpolation durch die Punkte $(t_i, f(t_i))$, $i = 0, 1, \ldots, N$, so ergibt sich die Quadraturformel

$$\int_a^b \frac{f(t)}{t - x}\, dt = \sum_{i=0}^N f(t_i) \int_a^b \frac{\phi_i(t)}{t - x}\, dt + r_N(f, x) = \sum_{i=0}^N f(t_i) A_i(x) + r_N(f, x) \tag{4}$$

mit den Gewichten $A_i(x)$ und dem Fehler $r_N(f, x)$.

Für $x \neq t_j$, $j = i-1,\ i,\ i+1$ erhält man die Gewichtsfunktionen $A_i(x)$ als

$$\begin{aligned} A_0(x) &= \frac{t_1 - x}{h_1} \ln \left| \frac{t_1 - x}{t_0 - x} \right| - 1, \\ A_i(x) &= \frac{t_{i+1} - x}{h_{i+1}} \ln \left| \frac{t_{i+1} - x}{t_i - x} \right| - \frac{t_{i-1} - x}{h_i} \ln \left| \frac{t_i - x}{t_{i-1} - x} \right|, \\ A_N(x) &= \frac{x - t_{N-1}}{h_N} \ln \left| \frac{t_N - x}{t_{N-1} - x} \right| + 1 \quad (i = 1, \ldots, N-1). \end{aligned} \tag{5}$$

An den Stellen $x = t_j$, $j = i-1,\ i,\ i+1$ existieren die Grenzwerte der Funktionen $A_i(x)$, $1 \le i \le N-1$:

$$\begin{aligned} A_i(t_{i-1}) &= \frac{h_i + h_{i+1}}{h_{i+1}} \ln \frac{h_i + h_{i+1}}{h_i}\,, \\ A_i(t_{i+1}) &= \frac{h_i + h_{i+1}}{h_i} \ln \frac{h_{i+1}}{h_i + h_{i+1}}\,, \\ A_i(t_i) &= \ln \frac{h_{i+1}}{h_i}\,. \end{aligned}$$

Genauso werden die Werte $A_0(t_1) = -1$ und $A_N(t_{N-1}) = 1$ festgelegt. Für $i = 1, 2, \ldots, N-1$ sind die Funktionen $A_i(x)$ für jedes $x \in \mathbb{R}$ definiert und beschränkt. Die Gewichtsfunktionen $A_0(x)$ und $A_N(x)$ sind auf jedem Intervall $[a_1, b_1] \subset (a, b)$, beschränkt.

Hilfssatz 1 *Für $1 \leq i \leq N-1$ sind die Gewichte $A_i(x)$ beschränkt. Die folgenden Abschätzungen sind für jedes $x \in \mathbb{R}$ gültig:*

$$|A_i(x)| \leq A$$

mit

$$A = \max\{2H\ln(2H), 3+2H\}\,.$$

Mit der Parametrisierung $x = t_0 + sh_1$ bzw. $x = t_N - sh_N$, $0 < s \leq 1$, erhält man für die Randgewichte

$$|A_0(x)| \leq \ln\frac{1}{s} + 1 \quad bzw. \quad |A_N(x)| \leq \ln\frac{1}{s} + 1.$$

Beweis: Aus den Grenzwerten ergeben sich die Schranken:

$$\begin{aligned} |A_i(t_{i-1})| &\leq 2H\,|\ln(2H)|\,, \\ |A_i(t_{i+1})| &\leq 2H\left|\ln\frac{H}{2}\right| \quad \text{und} \\ |A_i(t_i)| &\leq |\ln H|\,. \end{aligned}$$

Für jedes x mit $a \leq x \leq t_{i-1}$ bzw. $x \geq t_{i+1}$ und jedes i findet man durch Einsetzen der Funktionen $\phi_i(t)$ die Ungleichung

$$|A_i(x)| = \left| -\!\!\!\!\!\!\int_a^b \frac{\phi_i(t)}{t-x}\,dt \right| \leq \sup_{t_{i-1}\leq t\leq t_i}\left|\frac{\phi_i(t)}{t-t_{i-1}}\right| h_i + \frac{1}{h_i}h_{i+1}^2 \leq 1 + H\overline{h}\,.$$

Analog wird für $x \geq t_{i+1}$ vorgegangen.

Für $t_{i-1} < x < t_i$ ist

$$\begin{aligned} |A_i(x)| &= \left| -\!\!\!\!\!\!\int_{t_{i-1}}^{t_{i+1}} \frac{\phi_i(t)-\phi_i(x)}{t-x}\,dt + \phi_i(x)\ln\left|\frac{t_{i+1}-x}{t_{i-1}-x}\right| \right| \\ &\leq \left| -\!\!\!\!\!\!\int_{t_{i-1}}^{t_i} \frac{\phi_i(t)-\phi_i(x)}{t-x}\,dt + -\!\!\!\!\!\!\int_{t_i}^{t_{i+1}} \frac{\phi_i(t)-\phi_i(x)}{t-x}\,dt \right| + \left| \frac{x-t_{i-1}}{h_i}\left(\frac{t_{i+1}-x}{x-t_{i-1}} - 1\right) \right| \\ &\leq 1 + \left| -\!\!\!\!\!\!\int_{t_i}^{t_{i+1}} \frac{\phi_i(t)-\phi_i(t_i)}{t-x}\,dt \right| + \left| -\!\!\!\!\!\!\int_{t_i}^{t_{i+1}} \frac{\phi_i(t_i)-\phi_i(x)}{t-x}\,dt \right| + \frac{h_i+h_{i+1}}{h_i} \\ &\leq 1 + 1 + \frac{h_{i+1}}{h_i} + 1 + \frac{h_{i+1}}{h_i} \leq 3 + 2H \end{aligned}$$

Die beiden letzten Ungleichungen sind offenbar. ■

Der Hilfssatz 1 läßt sich folgendermaßen ergänzen.

Hilfssatz 2 *Für jedes $x \in [a_1, b_1] \subset (a, b)$ ist die Ungleichung*

$$\sum_{i=0}^{N} |A_i(x)| \leq K_1 \ln(N-1)$$

erfüllt. Die Konstante K_1 ist aus den im Beweis angegebenen Ungleichungen (6), (7) *und* (8) *zu ermitteln. Werden in der Summe der Gewichte der erste und der letzte Summand nicht berücksichtigt, erhält man für $x \in [a, b]$*

$$\sum_{i=1}^{N-1} |A_i(x)| \leq K_2 \ln(N-1)$$

mit einer geeigneten Konstante K_2, die der Ungleichung $K_2 \ln(N-1) \geq 4A + 2H \ln(N-1)$ genügt.

Beweis: Es sei $x \in [t_k, t_{k+1}]$ mit $1 \leq k \leq N-2$. Nach Hilfssatz 1 sind die Gewichte $A_j(x)$, $j = k-1, k, k+1, k+2$ durch die Konstante A beschränkt. Im weiteren wird $\sum_{i=k}^{l} = 0$ gesetzt, falls $k > l$ gilt.

$$\sum_{\substack{i=0 \\ i \neq k-1,k,k+1,k+2}}^{N} |A_i(x)| = \sum_{\substack{i=0 \\ i \neq k-1,k,k+1,k+2}}^{N} \left| \fint_a^b \frac{\phi_i(t)}{t-x}\, dt \right|$$

$$\leq \bar{h} \left[\sum_{i=0}^{k-2} \frac{1}{|t_{i+1} - x|} + \sum_{i=k+3}^{N} \frac{1}{|t_{i-1} - x|} \right] \leq \bar{h} \left[\sum_{i=0}^{k-2} \frac{1}{|t_{i+1} - t_k|} + \sum_{i=k+3}^{N} \frac{1}{|t_{i-1} - t_{k+1}|} \right]$$

$$\leq \bar{h} \left[\sum_{i=0}^{k-2} \frac{1}{\sum\limits_{j=i+2}^{k} h_j} + \sum_{i=k+3}^{N} \frac{1}{\sum\limits_{j=k+2}^{i-1} h_j} \right] \leq \frac{\bar{h}}{\underline{h}} \left[\sum_{i=0}^{k-2} \frac{1}{k-i-1} + \sum_{i=k+3}^{N} \frac{1}{i-k-2} \right]$$

$$\leq 2H \sum_{i=1}^{N-2} \frac{1}{i} \leq 2H \ln(N-1).$$

Insgesamt findet man für diesen Fall die Ungleichung

$$\sum_{i=0}^{N} |A_i(x)| \leq 4A + 2H \ln(N-1)\,. \tag{6}$$

Für hinreichend großes N ist $t_1 < a_1$ bzw. $b_1 < t_{N-1}$ und der Beweis ist vollständig.

Der Fall $a_1 < t_1$ und $x \in [a_1, t_1]$, d.h. $x = a + sh_1$, $0 < \dfrac{a_1 - a}{h_1} \leq s \leq 1$, kann ähnlich behandelt werden:

$$\begin{aligned}
|A_0(x)| + \sum_{i=1}^{N} |A_i(x)| &\leq (1-s)(|\ln(1-s)| + |\ln(s)|) + \sum_{i=1}^{N} |A_i(x)| \\
&\leq \frac{1}{e} + |\ln \frac{a_1 - a}{h_1}| + 2A + \sum_{i=3}^{N} |A_i(t_1)| \\
&\leq \frac{1}{e} + |\ln \frac{a_1 - a}{h_1}| + 2A + H\ln(N-1). \qquad (7)
\end{aligned}$$

Analog findet man für $x \in [t_{N-1}, b_1]$

$$\sum_{i=0}^{N} |A_i(x)| \leq \frac{1}{e} + \left|\ln \frac{b - b_1}{h_N}\right| + 2A + H\ln(N-1). \qquad \blacksquare \qquad (8)$$

Den Fehler $r_N(f,x)$ in Gleichung (4) erhält man als Cauchy–Hauptwert–Integral des Interpolationsfehlers. Er ist in verschiedenen Arbeiten angegeben worden ([7], [6], [3]).

Hilfssatz 3 *Wenn $f(t) \in C^{k,\lambda}$, $k \in \{0,1\}$, $0 < \lambda \leq 1$, gilt für den Fehler $r_N(f,x)$ aus* (4)

$$|r_N(f,x)| \leq c_k N^{-(k+\lambda)} \ln N \quad \textit{für jedes} \quad x \in (a,b)\,.$$

Die Konstanten c_0 und c_1 sind von N unabhängig.

Bemerkung 4 Die Ordnung in Hilfssatz 3 läßt sich nicht verbessern. Die maximale Konvergenzordnung $N^{-2}\ln N$ wird für $f(t) \in C^{1,1}$ erreicht. Diethelm hat in [3] für $f(t) \in C^k$, $k \in \{1,2\}$, die im Hilfssatz 3 auftretenden Konstanten c_k asymptotisch scharf angegeben.

Jetzt wird das Cauchy-Hauptwert-Integral über $f'(t)$ aus (3) behandelt. Bezeichnet $Q_i(t)$ für $i = 1, \ldots, N-1$ die Parabel durch die Punkte $(t_{i-1}, f(t_{i-1}))$, $(t_i, f(t_i))$ und $(t_{i+1}, f(t_{i+1}))$, so werden die Werte $f'(t_i)$ durch $f_i' = Q_i'(t_i)$ approximiert. Weiter wird $f_0' = Q_1'(t_0)$ und $f_N' = Q_{N-1}'(t_n)$ gesetzt. Im einzelnen erhält man für $i = 1, 2, \ldots, N-1$

$$f_i' = \frac{1}{h_i h_{i+1}(h_i + h_{i+1})} \left[-h_{i+1}^2 f(t_{i-1}) + (h_{i+1}^2 - h_i^2) f(t_i) + h_i^2 f(t_{i+1})\right], \qquad (9)$$

für $i = 0$

$$f_0' = \frac{1}{h_1 h_2 (h_1 + h_2)} \left[-h_2(2h_1 + h_2) f(t_0) + (h_1 + h_2)^2 f(t_1) - h_1^2 f(t_2) \right] \tag{10}$$

und $i = N$

$$\begin{aligned} f_N' &= \frac{1}{h_{N-1} h_N (h_{N-1} + h_N)} [h_N^2 f(t_{N-2}) \\ &- (h_{N-1} + h_N)^2 f(t_{N-1}) + h_{N-1}(h_{N-1} + 2h_N) f(t_N)] \,. \end{aligned} \tag{11}$$

Ferner wird der Fehler der numerischen Differentiation (9), (10) und (11) benötigt:

Hilfssatz 5 *Sei $f(t) \in C^{k+1}$ für $k \in \{0, 1\}$, dann gelten für $i = 0, \ldots, N$ die Ungleichungen*

$$|f'(t_i) - f_i'| \le C_k h^k \omega(f^{(k+1)}, h)$$

mit $C_0 = 1 + H$ und $C_1 = 2$.

Beweis: Für $f(t) \in C^1$ existieren Punkte $\xi_1 \in (t_0, t_1)$ und $\xi_2 \in (t_1, t_2)$ mit

$$\begin{aligned} |f'(t_0) - f_0'| &= \left| f'(t_0) + \frac{h_1}{h_1 + h_2} \frac{f(x_2) - f(x_1)}{h_2} - \frac{2h_1 + h_2}{h_1 + h_2} \frac{f(x_1) - f(x_0)}{h_1} \right| \\ &= \left| f'(t_0) + \frac{h_1}{h_1 + h_2} f'(\xi_2) - \frac{2h_1 + h_2}{h_1 + h_2} f'(\xi_1) \right| \\ &\le |f'(t_0) - f'(\xi_1)| + \frac{h_1}{h_1 + h_2} |f'(\xi_2) - f'(\xi_1)| \\ &\le \omega(f', \overline{h}) + \frac{H}{2} \omega(f', 2\overline{h}) \le (1 + H) \omega(f', \overline{h}) \,. \end{aligned}$$

Analog zeigt man

$$|f'(t_N) - f_N'| \le (1 + H)\, \omega(f', \overline{h})$$

und für $1 \le i \le N-1$

$$|f'(t_i) - f_i'| \le H \omega(f', \overline{h}) \,.$$

Im Falle $f(t) \in C^2$ wird für $t_{i-1} \le t \le t_{i+1}$ die Parabel $Q_i(t)$ betrachtet. Bezeichnet x^* eine Nullstelle von $f'(t) - Q_i'(t)$ und x^{**} die Nullstelle von $f''(t) - Q_i''(t)$, so gilt für jedes

$t \in [t_{i-1}, t_{i+1}]$ die Darstellung

$$\begin{aligned} |f'(t) - Q_i'(t)| &= \left| \int_{x^*}^{t} (f''(s) - Q_i''(s))\, ds \right| \\ &\leq |t - x^*| \max_{s\in[x^*,t]} |f''(s) - Q_i''(s)| \\ &\leq |t - x^*| \max_{s\in[x^*,t]} (|f''(s) - f''(x^{**})| + |Q_i''(s) - Q_i''(x^{**})|) \\ &\leq 2\overline{h}\omega(f'', \overline{h}). \end{aligned}$$

■

Mit den Gewichten $A_i(x)$ aus (5) erhält man aus (3) und (4)

$$\fint_a^b \frac{f(t)}{(t-x)^2}\, dt = -\left[\frac{f(b)}{(b-x)} - \frac{f(a)}{(a-x)}\right] + \sum_{i=0}^{N} f'(t_i) A_i(x) + r_N(f', x)\,.$$

Da nur die Kenntnis der Funktionswerte $f(t_i)$ vorausgesetzt wird, werden die $f'(t_i)$ durch die Werte f_i' aus (9), (10) und (11) ersetzt:

$$\fint_a^b \frac{f(t)}{(t-x)^2}\, dt = -\left[\frac{f(b)}{(b-x)} - \frac{f(a)}{(a-x)}\right] + \sum_{i=0}^{N} f_i' A_i(x) + R_N(f, x)\,. \tag{12}$$

Es werden die Vektoren $\underline{f} = (f(t_0), f(t_1), \ldots, f(t_N))^T$, $\underline{f'} = (f_0', f_1', \ldots, f_N')^T$, $\underline{a} = (A_0, A_1, \ldots, A_N)^T$ und $\underline{b} = (B_0, B_1, \ldots, B_N)^T$ und die Matrix $D = (d_{ij})_0^N$ mit den Elementen

$$d_{00} = \frac{-(2h_1 + h_2)}{h_1(h_1 + h_2)} + \frac{1}{a - x}\,, \qquad d_{01} = \frac{h_1 + h_2}{h_1 h_2}\,, \qquad d_{02} = \frac{-h_1}{h_2(h_1 + h_2)}\,,$$

$$d_{i,i-1} = \frac{-h_{i+1}}{h_i(h_i + h_{i+1})}\,, \qquad d_{i,i} = \frac{h_{i+1} - h_i}{h_i h_{i+1}}\,, \qquad d_{i,i+1} = \frac{h_i}{h_{i+1}(h_i + h_{i+1})}$$

für $1 \leq i \leq N-1$,

$$d_{N,N-2} = \frac{h_N}{h_{N-1}(h_{N-1} + h_N)}\,, \qquad d_{N,N-1} = -\frac{h_{N-1} + h_N}{h_{N-1} h_N}\,,$$

$$d_{N,N} = \frac{h_{N-1} + 2h_N}{h_N(h_{N-1} + h_N)} - \frac{1}{b - x}\,, \quad d_{i,j} = 0 \quad \text{sonst}$$

eingeführt. Dann ist für

$$\underline{b} = D^T \underline{a}$$

die folgende Gleichung erfüllt:

$$-\left[\frac{f(b)}{(b-x)} - \frac{f(a)}{(a-x)}\right] + \sum_{i=0}^{N} f_i' A_i(x) = \sum_{i=0}^{N} f(t_i) B_i(x)\,. \tag{13}$$

Die Gleichung (12) wird zu

$$\int\!\!\!\!\!\!=_a^b \frac{f(t)}{(t-x)^2}\,dt = \sum_{i=0}^{N} f(t_i) B_i(x) + R_N(f,x)\,. \tag{14}$$

Für den Fehler $R_N(f,x)$ gilt wegen Hilfssatz 2 die Abschätzung:

$$\begin{aligned} |R_N(f,x)| &= \left| \int\!\!\!\!\!\!=_a^b \frac{f(t)}{(t-x)^2}\,dt - \sum_{i=0}^{N} f(t_i) B_i(x) \right| = \left| \int\!\!\!\!\!-_a^b \frac{f'(t)}{t-x}\,dt - \sum_{i=0}^{N} f_i' A_i(x) \right| \\ &\le \left| \int\!\!\!\!\!-_a^b \frac{f'(t)}{t-x}\,dt - \sum_{i=0}^{N} f'(t_i) A_i(x) \right| + \max_i \{|f'(t_i) - f_i'|\} \sum_{i=0}^{N} |A_i(x)| \qquad (15) \\ &\le \left| r_N(f', \overline{h}) \right| + \max_i \{|f'(t_i) - f_i'|\} K_1 ln(N-1)\,. \end{aligned}$$

Daraus ergibt sich mit den Hilfssätzen 3 und 5

Satz 6 *Für* $a < a_1 \le x \le b_1 < b$, $f \in C^{k+1,\lambda}$ *mit* $k \in \{0,1\}$, $0 < \lambda \le 1$ *gilt für den Fehler der Quadraturformel* (14) *die Abschätzung*

$$|R_N(f,x)| \le (c_k + K_1 C_k)\ \overline{h}^{k+\lambda} \ln N$$

mit den Konstanten c_k *aus Hilfssatz* 3*,* K_1 *aus Hilfssatz* 2 *und* C_k *aus Hilfssatz* 5*.*

Verwendet man äquidistante Stützstellen $t_i = a + ih$ mit $h = (b-a)/N$, dann vereinfachen sich die f_i' aus (9), (10) und (11) zu

$$f_i' = \frac{1}{2h}(f(t_{i+1}) - f(t_{i-1})) \quad \text{für} \quad i = 1, \ldots, N-1,$$

$$f_0' = \frac{1}{2h}(-3f(t_0) + 4f(t_1) - f(t_2)) \quad \text{und}$$

$$f_N' = \frac{1}{2h}(f(t_{N-2}) - 4f(t_{N-1}) + 3f(t_N)).$$

Für diesen Fall sollen die Gewichte $B_i(x)$ für $0 \leq i \leq N$ explizit angegeben werden:

$$\begin{aligned}
B_0(x) &= \frac{1}{2h}(-3A_0(x) - A_1(x) + \frac{2h}{a-x}) \\
B_1(x) &= \frac{1}{2h}(4A_0(x) - A_2(x)) \\
B_2(x) &= \frac{1}{2h}(-A_0(x) + A_1(x) - A_3(x)) \\
B_i(x) &= \frac{1}{2h}(A_{i-1}(x) - A_{i+1}(x)) \quad \text{für} \quad 3 \leq i \leq N-3 \qquad (16) \\
B_{N-2}(x) &= \frac{1}{2h}(A_{N-3}(x) - A_{N-1}(x) + A_N(x)) \\
B_{N-1}(x) &= \frac{1}{2h}(A_{N-2}(x) - 4A_N(x)) \\
B_N(x) &= \frac{1}{2h}(A_{N-1}(x) + 3A_N(x) - \frac{2h}{b-x})) \,.
\end{aligned}$$

Bemerkung 7 Aus [6], S. 185, Theorem 4 erkennt man, daß für $f(t) \in C^{r+m-1,\lambda}$ die maximal erreichbare (optimale) Konvergenzordnung einer Quadratur mit $N+1$ Stützstellen für das hypersinguläre Integral $⨎_a^b \frac{f(t)}{(t-x)^m}\, dt$, $a < x < b$, $m > 1$, von der Ordnung $O(1/N^{r+\lambda})$ ist. Ein Vergleich dieser Aussage mit Satz 6 zeigt den Verlust der Ordnung $O(\ln N)$.

Bemerkung 8 Wenn die Werte $f'(a)$ und $f'(b)$ zur Verfügung stehen, kann man das Verfahren so modifizieren, daß es nicht nur in jedem abgeschlossenen Teilintervall $[a_1, b_1]$ von (a, b), sondern auch in (a, b) mit gleicher Konvergenzordnung konvergiert. Ausgehend von (12) erhält man wegen $f'(t_0) = f_0'$ und $f'(t_N) = f_N'$ unter Berücksichtigung von der zu (13) analogen Bedingung

$$-\left[\frac{f(b)}{(b-x)} - \frac{f(a)}{(a-x)}\right] + \sum_{i=1}^{N-1} f_i' A_i(x) = \sum_{i=0}^{N} f(t_i) B_i(x)$$

die Quadraturformel

$$⨎_a^b \frac{f(t)}{(t-x)^2}\, dt = f'(a)A_0(x) + f'(b)A_N(x) \qquad (17)$$

$$+ \sum_{i=0}^{N} f_i B_i(x) + R_N(f, x) \,.$$

Die Gewichte berechnen sich jetzt nach

$$\begin{aligned}
B_0(x) &= \frac{1}{2h}(-A_1(x) + \frac{2h}{a-x}) \\
B_1(x) &= \frac{1}{2h}(-A_2(x)) \\
B_i(x) &= \frac{1}{2h}(A_{i-1}(x) - A_{i+1}(x)) \quad \text{für} \quad 2 \leq i \leq N-2 \\
B_{N-1}(x) &= \frac{1}{2h}A_{N-2}(x) \\
B_N(x) &= \frac{1}{2h}(A_{N-1}(x) - \frac{2h}{b-x})\,.
\end{aligned} \tag{18}$$

In der Fehlerdarstellung (15) entfallen wegen $f'(t_0) = f'_0$ und $f'(t_N) = f'_N$ der erste und der letzte Summand. Man erhält die Ungleichung

$$|R_N(f,x)| \leq |r_N(f',x)| + \max_i |f'(t_i) - f'_i| \sum_{i=1}^{N-1} |A_i(x)|. \tag{19}$$

Nach Hilfssatz 1 sind alle Gewichte $A_i(x)$, $1 \leq i \leq N-1$, im Intervall $[a,b]$ beschränkt. Daraus folgt die Konvergenz des modifizierten Verfahrens mit gleicher Konvergenzordnung auf dem Intervall (a,b). Das Randverhalten des hypersingulären Integrals wird durch die angegebene Formel exakt widergespiegelt.

Beispielrechnungen

Es werden Beispielrechnungen vorgestellt, die die oben angegebenen Abschätzungen praktisch bestätigen. Die zweite und die dritte Spalte der Tabellen enthalten die Fehler der Quadraturformel (14) mit den Gewichten (16) und der Quadraturformel (17) aus Bemerkung 8 mit den Gewichten (18). Im ersten Beispiel, das in [2] ebenfalls berechnet wurde, ist ein gewöhnliches Integral als Hadamardsches Integral geschrieben und dann numerisch berechnet worden. Das Beispiel 5 zeigt als einziges bessere Resultate in der zweiten Spalte als in der ersten. Hier wirkt sich die gleichmäßige Konvergenz des zweiten Verfahrens aufgrund der verwendeten exakten Werte $f'(a)$ und $f'(b)$ aus. Im 6. Beispiel wurde ein uneigentliches Integral berechnet. Da die Singularität $(x = 0)$ an der unteren Grenze des Intervalls liegt und somit $A_0(0)$ nicht existiert, mußte das Integrationsintervall dort vergrößert werden. Die Wahl von $a = -0,1$ als untere Grenze ist willkürlich.

Beispiel 1

$$\neq\!\!\!\!\!\!\int_0^1 \frac{(t-0,2)^3}{(t-0,2)^2}\,dt = 0,3$$

N	**Quadratur (14)**	**Quadratur (17)**
45	0,0010916022	0,0010484255
90	0,0002646182	0,0002593515
180	0,0000651559	0,0000645054
360	0,0000161663	0,0000160855

Beispiel 2

$$\fint_{-1}^{1} \frac{(1,21 - t^2)^{-\frac{1}{2}}}{(t - 0,25)^2}\, dt = -0,89562283$$

N	**Quadratur (14)**	**Quadratur (17)**
45	0,213472745	0,13842829
90	0,04729129	0,03928024
180	0,01151341	0,01052917
360	0,002860037	0,00273817

Beispiel 3

$$\fint_{0}^{1} \frac{t^6}{t - \frac{1}{2}}\, dt = \fint_{0}^{1} \frac{t^6 (t - \frac{1}{2})}{(t - \frac{1}{2})^2}\, dt = 0,4\bar{3}$$

N	**Quadratur (14)**	**Quadratur (17)**
45	0,0102968314	0,0108143596
90	0,0026823830	0,0027489929
180	0,0006823664	0,0006908153
360	0,0001720917	0,0001731555

Beispiel 4

$$\fint_{0}^{1} \frac{t^6}{(t - \frac{1}{2})^2}\, dt = 1,2$$

N	**Quadratur (14)**	**Quadratur (17)**
45	0,0116577349	0,0120815182
90	0,0030126239	0,0030665368
180	0,0007623189	0,0007691175
360	0,0001917366	0,0001925902

Beispiel 5

$$\neq\!\!\!\!\!\!\int_0^1 \frac{t^6}{(t-0,999999999)^2}\,dt = -1000000110,62385$$

N	Quadratur (14)	Quadratur (17)
45	0,2452501174	0,0537552676
90	0,0573006670	0,0160168610
180	0,0130679748	0,0046476019
360	0,0029215371	0,00013225671

Beispiel 6

$$\int_0^1 t^{\frac{1}{2}}\,dt = \neq\!\!\!\!\!\!\int_{-0,1}^1 \frac{f(t)}{t^2}\,dt \quad \text{mit} \quad f(t) = \begin{cases} 0 & \text{für } t \le 0 \\ t^{\frac{5}{2}} & \text{für } t > 0 \end{cases}$$

N	Quadratur (14)	Quadratur (17)
45	0,0084331287	0,0084377737
90	0,0030551119	0,0030556874
180	0,0010135985	0,0010136701
360	0,0002943699	0,0002943789

Literatur

[1] **Delbourgo, D.** und **Elliott, D. :** *On the approximate evaluation of Hadamard finite-part integrals.* IMA J. Numer. Anal. **14**, 485-500 (1994)

[2] **Delbourgo, D. :** *On the numerical evaluation of Hadamard finite-part integrals.* B.Sc. (Hons) Thesis, Mathematics Department, University of Tasmania 1994

[3] **Diethelm, K. :** *Asymptotically sharp error bounds for a quadrature rule for Cauchy Principal Value Integrals based on piecewise linear interpolation.* Hildesheimer Informatikberichte 19/94 (1994)

[4] **Linz, P. :** *On the approximate computation of certain strongly singular integrals.* Computing **35**, 345-353 (1985)

[5] **Monegato, G. :** *Numerical evaluation of hypersingular integrals.* J. Comp. Appl. Math. **50**, 9-31 (1994)

[6] **Stolle, H.-W.** und **Strauß, R. :** *On the numerical integration of certain singular integrals.* Computing **48**, 177-189 (1992)

[7] **Strauß, R. :** *Eine Interpolationsquadratur für Cauchy-Hauptwert-Integrale.* Rostock. Math. Kolloq. **22**, 57-63 (1983)

[8] **Strauß, R. :** *Quadratur von Hadamardschen Integralen ganzzahliger Ordnung.* (in Vorbereitung)

[9] **Volkov, E. A. :** *Numerical Methods.* Moscow 1986

eingegangen: 15. September 1995

Autor:

Dr. R. Strauß
Universität Rostock
Fachbereich Mathematik
Universitätsplatz 1
18051 Rostock
Deutschland

Rostock. Math. Kolloq. **49**, 141–154 (1995)

Subject Classification (AMS)
17 B 30, 20 F 05

M. F. Newman, G. Sauerbier, J. Wisliceny

Groups of prime-power order with a small number of relations

Dedicated to the professors of mathematics
L. Berg, W. Engel, G. Pazderski, *and* H.- W. Stolle.

1 Introduction

The Theorem of Golod/Shafarevich (1964) (briefly GST) gives a lower bound for the number of relations needed to define a group of prime-power order. Specifically let G be a group of prime-power order and d the generator number of G, then every presentation for G has more than $d^2/4$ relations. (The generator number is the size of a minimal generating set.) The GST actually says that every pro-p-presentation for G has more than $d^2/4$ relations. The same inequality also holds for presentations of nilpotent Lie algebras with finite generator number d (Koch 1977). It is natural to ask how sharp this inequality is (see for example Kostrikin 1965). Wisliceny (1981) has shown that, for odd primes p, there are pro-p-presentations with d generators and $d^2/4 + d/2 - (7 + (-1)^d)/8$ relations which define finite pro-p-groups (and also that there are presentations for Lie algebras with the same numbers of generators and of relations which define nilpotent Lie algebras) with generator number d. The corresponding result for $p = 2$ was proved by Sauerbier (1986). Thus for $d = 3$ and $d = 4$ the bound is reached for all primes p. With the help of the p-Quotient-Program [4] Newman and Wisliceny (unpublished) showed that the bound $\lfloor d^2/4\rfloor + 1$ is reached for finite pro-p-groups with $5 \leq d \leq 8$ for some specific primes. On the basis of these results they conjectured that the GST is sharp in the sense that:

for every prime p and every positive integer d there is a pro-p-presentation with d generators and $\lfloor d^2/4\rfloor + 1$ relations which defines a finite pro-p-group with generator number d.

Direct use of the p-Quotient Program can only give results for a finite number of primes and generator numbers.

The main results of this paper are:

For every prime p there is a pro-p-presentation with 5 *generators and* 7 *relations which defines a finite pro-p-group with generator number* 5. (Theorem 4.2 and Remark 4.3)

For every prime p there is a pro-p-presentation with 6 *generators and* 10 *relations which defines a finite pro-p-group with generator number* 6. (Theorem 4.4 and Remark 4.5)

The results of Wisliceny (1981) are obtained by first constructing corresponding Lie algebra presentations and then modifying them to construct suitable pro-p-presentations. We follow a similar path here though with a significant variation for $d = 5$. The main problem is *finding* suitable presentations. Showing they have the required properties is then relatively straight-forward. The search for presentations for Lie algebras was, in part, done with a program written in the computer algebra system GAP (M. Schönert et al. 1995).

We have found a nilpotent Lie algebra with generator number 6 defined by a presentation with 6 generators and 10 relations (3.3).

For $d = 5$ we have been unable to find a nilpotent Lie algebra with generator number 5 which has a presentation with 5 generators and 7 relations. In fact we expect none exist - at least none with homogeneous relations. We prove a partial result in this direction in Section 3. The Lie algebras associated with groups given by some presentations with 5 generators and 7 relations can be shown to have additional relations which are enough to prove nilpotence.

The relevant definitions are recalled briefly in Section 2 (more detail is given, for instance, in Wisliceny 1981). We deal with results for Lie algebras in Section 3. In Section 4 we exhibit the finite pro-p-groups.

Acknowledgements. We are indebted to DAAD, DFG, ANU and RWTH-Aachen for support to work on this project in Canberra, Güstrow and Aachen.

2 Preliminaries

Let $\mathcal{L}(X)$ or simply $\mathcal{L}$ be the free K-Lie algebra or $K[\pi]$-Lie algebra with the set $X = \{x_1, \dots, x_d\}$ of free generators over an arbitrary field K given with the usual gradation $\mathcal{L} = \sum_{m=1}^{\infty} \mathcal{L}^m$ ($\deg(x_i) = \deg(\pi) = 1$). In the case of the free K-Lie algebra the vector space $\mathcal{L}^m$ is generated by all the left-normed Lie products of the form $x_{i_1} \dots x_{i_m}$; in the case of the free $K[\pi]$-Lie algebra $\mathcal{L}^m$ is generated by all products of the form $\pi^k x_{i_1} \dots x_{i_l}$ with $l > 0$, $k \geq 0$ and $k + l = m$. Let $I(R)$ be the ideal in $\mathcal{L}$ which is generated by a set $R \subseteq \mathcal{L}^2$. Then the factor Lie algebra $L = \mathcal{L}/I(R)$ is a graded K-algebra or $K[\pi]$-algebra (respectively) $L = \sum_{m=1}^{\infty} L^m$ with the homogeneous components L^m which are generated by products of degree m. We denote the dimension of L^m by $\dim L^m$. We define the dimension-vector of L by $\dim L := (\dim L^1, \dim L^2, \dots)$. We call a Lie algebra L *dimension-periodic*, if $\dim L$ is eventually periodic. We consider nilpotent Lie algebras as special cases of dimension-periodic Lie algebras.

Let $\mathcal{F}(X)$ or simply $\mathcal{F}$ be the free pro-p-group freely generated by the set X. Further let $\{X; \mathcal{R}_\mathcal{F}\}$ be a pro-p-presentation with $\mathcal{R}_\mathcal{F} \subseteq \mathcal{F}_2 = \mathcal{F}^p(\mathcal{F}, \mathcal{F})$. Let G be the pro-p-group given by this presentation. We will, in the usual way, use the graded object $gr\, G = \sum_{m=1}^{\infty} G_m/G_{m+1}$ with respect to the p-central series

$$G_1 := G, \qquad G_{m+1} := G_m^p(G_m, G)$$

of G as an $F_p[\pi]$-Lie algebra where F_p is the field with p elements.

There is an isomorphism between $gr\, \mathcal{F} = \sum_{m=1}^{\infty} \mathcal{F}_m/\mathcal{F}_{m+1}$ and the free $F_p[\pi]$-Lie algebra $\mathcal{L} = \sum_{m=1}^{\infty} \mathcal{L}^m$. Let $R = h(\mathcal{R}_\mathcal{F})$ be the image under the canonical homomorphism $h : \mathcal{F}_2 \to \mathcal{F}_2/\mathcal{F}_3 \cong \mathcal{L}^2$. This map h assigns group products to sums, p-th-powering to multiplication by π and commutators to Lie products. For example $h(x_1 \cdot x_2^p \cdot (x_3, x_4)) = x_1 + \pi x_2 + x_3x_4$. Let J be the ideal in $\mathcal{L}$ with $gr\, G \cong \mathcal{L}/J$. Then we have $I(R) \subseteq J$.

We will use that the structure of the m-th homogeneous component of $gr\, G$ is

$$G_m/G_{m+1} = V^m \oplus \pi V^{m-1} \oplus \pi^2 V^{m-2} \oplus \ldots \oplus \pi^{m-2} V^2 \oplus \pi^{m-1} V,$$

where V^n $(1 \leq n \leq m)$ denotes the vector space which is generated by Lie monomials $x_{i_1} \cdots x_{i_n}$.

3 Lie Algebras

Theorem 3.1 *Let K be a field with characteristic zero. Then the K-Lie algebra $L5$ defined by*

$$\begin{aligned}&\{x_1, x_2, x_3, x_4, x_5;\\ &x_1x_2, x_1x_3, x_3x_5, x_4x_5, x_1x_4 - x_2x_3, x_1x_5 - x_2x_4, x_2x_5 - x_3x_4\}\end{aligned}$$

is dimension-periodic and $\dim L5 = (5, 3, 5, 3, \ldots)$.

Remark 3.2 If $char(K) = p$, then the dimension-vector depends on p. It seems to be dimension-periodic with period $2p$: $\dim L5 = (5, 3, ..., 5, 3, 5, 4, ...)$. In the proof below we simply show that $\dim L5 = (5, 3, 5, 3, 5, 3, 5, 3, ...)$ for $p \geq 5$ which is used later.

Proof: of Theorem 3.1 and Remark 3.2.
We have to show that for m odd $\dim L5^m = 5$ and for m even $\dim L5^m = 3$.
For $i = 1, \ldots , 5$ we set

$$i^{(1)} := x_i.$$

For $k \geq 1$ we set recursively:

$$\begin{array}{rcl} a^{(k)} & = & 2^{(k)}3^{(1)}, \\ b^{(k)} & = & 2^{(k)}4^{(1)}, \\ c^{(k)} & = & 4^{(k)}3^{(1)}, \end{array} \qquad \begin{array}{rcl} 1^{(k+1)} & = & a^{(k)}2^{(1)}, \\ 2^{(k+1)} & = & a^{(k)}3^{(1)}, \\ 3^{(k+1)} & = & a^{(k)}4^{(1)}, \\ 4^{(k+1)} & = & c^{(k)}3^{(1)}, \\ 5^{(k+1)} & = & c^{(k)}4^{(1)}. \end{array}$$

The system of relations of $L5$ is invariant with respect to the following transformation:
$1^{(k)} \longleftrightarrow 5^{(k)}$, $2^{(k)} \longleftrightarrow 4^{(k)}$, $3^{(k)} \longleftrightarrow 3^{(k)}$,
$a^{(k)} \longleftrightarrow c^{(k)}$, $b^{(k)} \longleftrightarrow -b^{(k)}$.
Our aim is to show that $\{1^{(k)}, 2^{(k)}, 3^{(k)}, 4^{(k)}, 5^{(k)}\}$ is a basis of $L5^m$ for $m = 2k-1$ and $\{a^{(k)}, b^{(k)}, c^{(k)}\}$ is a basis of $L5^m$ for $m = 2k$.
It is easy to see this for $m = 1$ and $m = 2$. By induction, for $m > 2$ we get the following equations.
$m = 2k$, $k \geq 2$, $i + j = k$:

$$a^{(i)}a^{(j)} = b^{(i)}b^{(j)} = c^{(i)}c^{(j)} = 0.$$

$m = 2k$, $k \geq 2$, $i + j = k + 1$:

$$1^{(i)}1^{(j)} = 2^{(i)}2^{(j)} = 3^{(i)}3^{(j)} = 4^{(i)}4^{(j)} = 5^{(i)}5^{(j)} = 0,$$

$$1^{(i)}2^{(j)} = 1^{(i)}3^{(j)} = 5^{(i)}4^{(j)} = 5^{(i)}3^{(j)} = 0.$$

$m = 2k$, $k \geq 2$:

$$\begin{array}{rccccccccc} a^{(k)} & = & 2^{(k)}3^{(1)} & = & -\frac{1}{3}*2^{(k-1)}3^{(2)} & = & \cdots & = & -\frac{1}{3}*2^{(1)}3^{(k)} \\ & = & \frac{1}{2}*1^{(k)}4^{(1)} & = & \cdots & = & \frac{1}{2}*1^{(2)}4^{(k-1)} & = & 1^{(1)}4^{(k)} \\ & = & -a^{(k-1)}b^{(1)} & = & \cdots & = & -a^{(1)}b^{(k-1)}, & & \\ b^{(k)} & = & 2^{(k)}4^{(1)} & = & \cdots & = & 2^{(1)}4^{(k)} & = & \frac{1}{2}*1^{(k)}5^{(1)} \\ & = & \frac{1}{4}*1^{(k-1)}5^{(2)} & = & \cdots & = & \frac{1}{4}*1^{(2)}5^{(k-1)} & = & \frac{1}{2}*1^{(1)}5^{(k)} \\ & = & -a^{(k-1)}c^{(1)} & = & \cdots & = & -a^{(1)}c^{(k-1)}, & & \\ c^{(k)} & = & 4^{(k)}3^{(1)} & = & -\frac{1}{3}*4^{(k-1)}3^{(2)} & = & \cdots & = & -\frac{1}{3}*4^{(1)}3^{(k)} \\ & = & \frac{1}{2}*5^{(k)}2^{(1)} & = & \cdots & = & \frac{1}{2}*5^{(2)}2^{(k-1)} & = & 5^{(1)}2^{(k)} \\ & = & c^{(k-1)}b^{(1)} & = & \cdots & = & c^{(1)}b^{(k-1)}. & & \end{array}$$

$m = 2k + 1$, $k \geq 2$, $i + j = k + 1$:

$$a^{(i)}1^{(j)} = b^{(i)}3^{(j)} = c^{(i)}5^{(j)} = 0.$$

$m = 2k+1$, $k \geq 2$:

$$\begin{aligned}
1^{(k+1)} &= a^{(k)}2^{(1)} = \cdots = a^{(1)}2^{(k)} = b^{(k)}1^{(1)} \\
&= \frac{1}{2} * b^{(k-1)}1^{(2)} = \cdots = \frac{1}{2} * b^{(1)}1^{(k)}, \\
2^{(k+1)} &= a^{(k)}3^{(1)} = -\frac{1}{3} * a^{(k-1)}3^{(2)} = \cdots = -\frac{1}{3} * a^{(1)}3^{(k)} \\
&= b^{(k)}2^{(1)} = \cdots = b^{(1)}2^{(k)} = -c^{(k)}1^{(1)} = -\frac{1}{2} * c^{(k-1)}1^{(2)} \\
&= \cdots = -\frac{1}{2} * c^{(1)}1^{(k)}, \\
3^{(k+1)} &= a^{(k)}4^{(1)} = \cdots = a^{(1)}4^{(k)} = c^{(k)}2^{(1)} = \cdots = c^{(1)}2^{(k)}, \\
4^{(k+1)} &= c^{(k)}3^{(1)} = -\frac{1}{3} * c^{(k-1)}3^{(2)} = \cdots = -\frac{1}{3} * c^{(1)}3^{(k)} \\
&= -b^{(k)}4^{(1)} = \cdots = -b^{(1)}4^{(k)} = -a^{(k)}5^{(1)} \\
&= -\frac{1}{2} * a^{(k-1)}5^{(2)} = \cdots = -\frac{1}{2} * a^{(1)}5^{(k)}, \\
5^{(k+1)} &= c^{(k)}4^{(1)} = \cdots = c^{(1)}4^{(k)} = -b^{(k)}5^{(1)} \\
&= -\frac{1}{2} * b^{(k-1)}5^{(2)} = \cdots = -\frac{1}{2} * b^{(1)}5^{(k)}.
\end{aligned}$$

It is routine to check that there are no further relations among the given generators.

Here we will only prove the equation

$$b^{(i)}b^{(j)} = 0 \quad \text{for} \quad m = 2k, \quad k \geq 2 \quad \text{and} \quad i + j = k.$$

In this part of the proof we need that $char(K) = 0$. Moreover the proof of the other equations is easier.

For $k = 2$, and thus $m = 4$, we have $b^{(1)}b^{(1)} = 0$ because of anticommutativity. For $k = 3$, and thus $m = 6$:

$$\begin{aligned}
2^{(3)}4^{(1)} &= b^{(2)}2^{(1)}4^{(1)} = b^{(2)}4^{(1)}2^{(1)} + 4^{(1)}2^{(1)}b^{(2)} \\
&= -4^{(3)}2^{(1)} - b^{(1)}b^{(2)} = b^{(1)}4^{(2)}2^{(1)} + b^{(2)}b^{(1)} \\
&= b^{(1)}2^{(1)}4^{(2)} + 2^{(1)}4^{(2)}b^{(1)} + b^{(2)}b^{(1)} \\
&= 2^{(2)}4^{(2)} + 2 * b^{(2)}b^{(1)} = -4^{(2)}2^{(2)} + 2 * b^{(2)}b^{(1)} \\
&= b^{(1)}4^{(1)}2^{(2)} + 2 * b^{(2)}b^{(1)} \\
&= b^{(1)}2^{(2)}4^{(1)} + 2^{(2)}4^{(1)}b^{(1)} + 2 * b^{(2)}b^{(1)} \\
&= 2^{(3)}4^{(1)} + 3 * b^{(2)}b^{(1)}
\end{aligned}$$

and therefore $b^{(2)}b^{(1)} = 0$. This also holds for $char(K) > 3$ which justifies the claim in Remark 3.2.

Assuming that the given equations are true for all $m < 2k$ we show $b^{(i)}b^{(j)} = 0$ for $i + j = k$, $k \geq 3$. We get

$$\begin{aligned} 2^{(k)}4^{(1)} &= b^{(i)}2^{(j)}4^{(1)} \\ &= b^{(i)}4^{(1)}2^{(j)} + 4^{(1)}2^{(j)}b^{(i)} \\ &= -4^{(i+1)}2^{(j)} - b^{(j)}b^{(i)}, \end{aligned}$$

moreover, if $\mu + \nu = k + 1$, $i < \mu$:

$$\begin{aligned} 2^{(\mu)}4^{(\nu)} &= b^{(i)}2^{(\mu-i)}4^{(\nu)} \\ &= b^{(i)}4^{(\nu)}2^{(\mu-i)} + 4^{(\nu)}2^{(\mu-i)}b^{(i)} \\ &= -4^{(i+\nu)}2^{(\mu-i)} - b^{(\mu+\nu-i-1)}b^{(i)} \\ &= 2^{(\mu-i)}4^{(i+\nu)} + b^{(i)}b^{(j)}, \end{aligned}$$

on the other hand, if $i \geq \mu$ and consequently $j < \nu$:

$$\begin{aligned} 2^{(\mu)}4^{(\nu)} &= -4^{(\nu)}2^{(\mu)} \\ &= b^{(j)}4^{(\nu-j)}2^{(\mu)} \\ &= b^{(j)}2^{(\mu)}4^{(\nu-j)} + 2^{(\mu)}4^{(\nu-j)}b^{(j)} \\ &= 2^{(j+\mu)}4^{(\nu-j)} + b^{(i)}b^{(j)}. \end{aligned}$$

Thus there are k_1, k_2 with $k_1 < k_2$ and s, t with $s + t = k + 1$ so, that

$$2^{(s)}4^{(t)} + k_1 * b^{(i)}b^{(j)} = 2^{(s)}4^{(t)} + k_2 * b^{(i)}b^{(j)}$$

and since $char(K) = 0$ follows $b^{(i)}b^{(j)} = 0$.

With the help of a program written in the computer algebra system GAP we have found that the following holds

3.3 *Let K be a field with characteristic different from 2. Then the K-Lie algebra $L6$ defined by the presentation*

$$\begin{aligned} &\{x_1, x_2, x_3, x_4, x_5, x_6; \\ &x_1x_2, x_1x_3, x_3x_4, x_4x_6, x_5x_6, x_1x_4 + x_2x_3, \\ &x_1x_5 + x_2x_4, x_1x_6 + x_2x_5, x_2x_6 + x_3x_5, x_3x_6 + x_4x_5\} \end{aligned}$$

is nilpotent of class 5 and

$$\dim L6 = (6, 5, 10, 12, 10, 0) \quad for \quad char(K) = 3,$$

$$\dim L6 = (6, 5, 10, 12, 8, 0) \quad for \quad char(K) > 3 \quad \text{or} \quad = 0.$$

We now examine the question of the existence of nilpotent K-Lie algebras L with $d(L) = 5$ and $r(L) = 7$.

In the papers of Brahana (1951,1958) quotient groups G/N with $N \subseteq (G, G)$ of the group G of the order p^{15} which can be given by the presentation $\{X; R\}$ with

$$\begin{aligned} X &= \{x_1, x_2, \ldots, x_{15}\} \quad \text{and} \\ R &= \{x_i^p \ (i = 1, \ldots, 15), \\ &\quad (x_2, x_1) = x_6, \ldots, (x_5, x_4) = x_{15}, \\ &\quad (x_j, x_i) \ (15 \geq j > i \geq 1, \ j > 5)\}. \end{aligned}$$

are studied. Using Brahana's classification of types of invariant subgroups N we can prove the following

Theorem 3.4 *There are no nilpotent Lie algebras with generator number* 5 *which can be defined by* 7 *homogeneous relations of degree* 2*.*

Proof: Brahana's classification of the planes in a 9-dimensional projective space into 22 types yields that we have only to investigate the nilpotence of the Lie algebras with relations:

$$\begin{aligned} a)\, L_a &: x_1x_2 - x_3x_4 + x_3x_5, x_1x_3 - x_2x_5, x_2x_4 - x_3x_5, x_1x_4, x_1x_5, x_2x_3, x_4x_5, \\ b)\, L_b &: x_1x_2 - x_3x_4, x_1x_3 - x_2x_5, x_1x_4 - x_2x_3, x_1x_5, x_2x_4, x_3x_5, x_4x_5, \\ c)\, L_c &: x_1x_2 - x_3x_4, x_1x_3 - x_2x_5, x_2x_3 - x_4x_5, x_1x_4, x_1x_5, x_2x_4, x_3x_5. \end{aligned}$$

Again let $i := x_i$ in the following calculations.
a) For $n \geq 1$ we consider the Lie products

$$\begin{aligned} ak_1l_1 &:= 12 \cdots 23 \cdots 3\,, \\ bk_2l_2 &:= 42 \cdots 23 \cdots 3\,, \\ ck_3l_3 &:= 52 \cdots 23 \cdots 3\,, \end{aligned}$$

where k_i denotes the number of the factors „2“ and l_i denotes the number of the factors „3“ with $k_i + l_i = n$, $k_i = 0, \ldots, n$. The number of these Lie products is $3n + 3$. From the relation $23 = 0$ and the Jacobi identity we get the following system of equations:

$$ak_1l_1 + b(k_1 - 1)(l_1 + 1) - c(k_1 - 1)(l_1 + 1) = 0, \text{ where } k_1 \geq 1, k_1 + l_1 = n\,,$$

$$ak_2l_2 + c(k_2 + 1)(l_2 - 1) = 0, \text{ where } l_2 \geq 1, k_2 + l_2 = n\,,$$

$$ck_3l_3 - b(k_3+1)(l_3-1) = 0, \text{ where } l_3 \geq 1, k_3 + l_3 = n.$$

This is a system of $3n$ homogeneous linear equations in $3n+3$ unknowns. Therefore the dimension of the space of solutions is at least 3 and consequently the Lie algebra L_a is not nilpotent.

b) Mapping $5 \to 0$ we get the Lie algebra $12 = 34$, $14 = 23$, 13, 24 with $d = 4$ and $r = 4$ as a homomorphic image of L_b. It is a consequence of GST that this Lie algebra, and therefore the Lie algebra L_b, is not nilpotent.

c) The transformation $1 \to 3$, $2 \to 2$, $3 \to 4$, $4 \to 1$, $5 \to 5$ shows that $L_c \cong L5$.

4 Finite pro-p-groups

Let G be the group defined by the following presentation:

$$\begin{aligned}&\{x_1,\ x_2,\ x_3,\ x_4,\ x_5;\\&(x_1,x_2),\ (x_1,x_3),\ (x_5,x_4),\ (x_5,x_3),\\&(x_1,x_4) = (x_2,x_3),\ (x_1,x_5) = (x_2,x_4),\\&(x_2,x_5) = (x_3,x_4)\}.\end{aligned}$$

The Nilpotent Quotient Program (Nickel 1993) shows that G has a largest nilpotent quotient and it has class 6; moreover modulo its 2-torsion it has class 5. Hence for p odd the pro-p-group G_p defined by the above presentation, viewed as a pro-p-presentation, is nilpotent of class 5. It follows that the Lie ring associated with the p-central series of G_p is nilpotent. Since by Theorem 3.1 this is not the case for $L5$ we need to find additional relations in the associated Lie ring.

4.1 *For $p \geq 3$ the space V^n in $gr\,G_p$ is zero for $n \geq 7$.*

Proof: The F_p-Lie algebra $gr\,G_p$ is a homomorphic image of the dimension-periodic Lie algebra $L5$ for $p \geq 3$.

If $gr\,G_p = \mathcal{L}/J$ and $L5 = \mathcal{L}/I$ then $I \subseteq J$. Let $\mathcal{L}_n := \sum_{m=n}^{\infty} \mathcal{L}^m$. We want to find elements of $J \setminus I$. For this we have to calculate in $gr\,G_p$ with regard to higher commutators. Instead of the Jacobi identity we use the rules:

$$abc = acb + cba + (ab)(bc) + (bc)(ca) + (ca)(ab),$$

$$a_1 \cdots a_k(ba) = a_1 \cdots a_k ba - a_1 \cdots a_k ab + a_1 \cdots a_k b(ab) + a_1 \cdots a_k a(ab)$$

$(k > 1)$ and calculate modulo $\mathcal{L}_{m+2} \cup I$ if the product on the left is an element of $\mathcal{L}_m$. We use again the symbols from the proof of (3.1). We get:

$$\begin{aligned}
0 &= 2^{(1)}3^{(1)}1^{(1)}5^{(1)}1^{(1)} \\
&= 1^{(1)}4^{(1)}1^{(1)}5^{(1)}1^{(1)} \\
&= 1^{(1)}4^{(1)}(1^{(1)}5^{(1)})1^{(1)} + 1^{(1)}4^{(1)}5^{(1)}1^{(1)}1^{(1)} \\
&\quad + 1^{(1)}4^{(1)}5^{(1)}(1^{(1)}5^{(1)})1^{(1)} + 1^{(1)}4^{(1)}1^{(1)}(1^{(1)}5^{(1)})1^{(1)} \\
&= 1^{(1)}4^{(1)}(1^{(1)}5^{(1)})1^{(1)} + 1^{(1)}4^{(1)}5^{(1)}1^{(1)}1^{(1)} + a^{(3)} + 0 \\
&= 5^{(1)}1^{(1)}(2^{(1)}3^{(1)})1^{(1)} + 2^{(1)}3^{(1)}5^{(1)}1^{(1)}1^{(1)} + a^{(3)}\,,
\end{aligned}$$

further

$$\begin{aligned}
(5^{(1)}1^{(1)})(2^{(1)}3^{(1)})1^{(1)} &= 5^{(1)}1^{(1)}1^{(1)}(2^{(1)}3^{(1)}) + 1^{(1)}(2^{(1)}3^{(1)})(5^{(1)}1^{(1)}) \\
&= 5^{(1)}1^{(1)}1^{(1)}(2^{(1)}3^{(1)}) + 0 \\
&= 5^{(1)}1^{(1)}1^{(1)}2^{(1)}3^{(1)} - 5^{(1)}1^{(1)}1^{(1)}3^{(1)}2^{(1)} \\
&\quad + 5^{(1)}1^{(1)}1^{(1)}2^{(1)}(3^{(1)}2^{(1)}) + 5^{(1)}1^{(1)}1^{(1)}3^{(1)}(3^{(1)}2^{(1)}) \\
&= 5^{(1)}1^{(1)}1^{(1)}2^{(1)}3^{(1)} - 5^{(1)}1^{(1)}1^{(1)}3^{(1)}2^{(1)} + 0 + 0 \\
&= 4^{(1)}3^{(1)}1^{(1)}1^{(1)}3^{(1)} + 0 \\
&= 4^{(1)}1^{(1)}3^{(1)}1^{(1)}3^{(1)} + (1^{(1)}4^{(1)})(4^{(1)}3^{(1)})1^{(1)}3^{(1)} \\
&= 3^{(1)}2^{(1)}3^{(1)}1^{(1)}3^{(1)} + 0 = 0.
\end{aligned}$$

Therefore we have

$$\begin{aligned}
0 &= 2^{(1)}3^{(1)}5^{(1)}1^{(1)}1^{(1)} + a^{(3)} \\
&= 2^{(1)}5^{(1)}3^{(1)}1^{(1)}1^{(1)} + (5^{(1)}2^{(1)})(2^{(1)}3^{(1)})1^{(1)}1^{(1)} + a^{(3)} \\
&= 3^{(1)}4^{(1)}3^{(1)}1^{(1)}1^{(1)} + a^{(3)} \\
&= 3^{(1)}4^{(1)}1^{(1)}1^{(1)}3^{(1)} + a^{(3)} \\
&= 1^{(1)}4^{(1)}3^{(1)}1^{(1)}3^{(1)} + (3^{(1)}4^{(1)})(4^{(1)}1^{(1)})1^{(1)}3^{(1)} + a^{(3)} \\
&= 2^{(1)}3^{(1)}3^{(1)}1^{(1)}3^{(1)} + 0 + a^{(3)} = a^{(3)}.
\end{aligned}$$

The symmetry of the presentation of G_p gives also $c^{(3)} = 0$. It follows $1^{(4)} = 2^{(4)} = \ldots = 5^{(4)} = 0$ and thus $V^7 = 0$.
(We can also show $b^{(3)} = 0$ so even $V^6 = 0$.)

Theorem 4.2 *The pro-p-presentation*

$$\begin{aligned}
\{x_1, x_2, x_3, x_4, x_5; \quad & (x_1, x_2),\ (x_1, x_3) = x_3^{-p}, \\
& (x_5, x_4) = x_4^{p},\ (x_5, x_3) = x_1^{-p}, \\
& (x_1, x_4) = (x_2, x_3) \cdot x_2^{p} \cdot x_3^{p}, \\
& (x_1, x_5) = (x_2, x_4), \\
& (x_2, x_5) = (x_3, x_4) \cdot x_5^{-p}\}
\end{aligned}$$

defines a finite group for all primes $p > 2$.

Remark 4.3 We denote the group H_p. We can see with the help of the p-Quotient Program:

$$\dim gr\, H_2 = (5, 8, 13, 16, 11, 4, 3, 3, 3, 3, \dots),$$

$$\dim gr\, H_3 = (5, 8, 13, 16, 10, 4, 0),$$

$$\dim gr\, H_p = (5, 8, 13, 16, 10, 3, 0) \quad \text{for} \quad p = 5,\ 7,\ 11\,.$$

This is presumably true for all $p \geq 5$; that, however, would require more work to show. It seems likely that H_2 is not a finite 2-group. The p-Quotient Program shows that the following slightly modified presentation does define a 2-group with class 6 and order 2^{57}:

$$\begin{array}{ll} \{x_1, x_2, x_3, x_4, x_5; & (x_1, x_2),\ (x_1, x_3) = x_3^{-2}, \\ & (x_5, x_4) = x_4^{-2},\ (x_5, x_3) = x_1^{-2}, \\ & (x_1, x_4) = (x_2, x_3) \cdot x_2^{-2} \cdot x_3^2, \\ & (x_1, x_5) = (x_2, x_4), \\ & (x_2, x_5) = (x_3, x_4) \cdot x_5^{-2}\} \end{array}$$

Proof: of Theorem 4.2: Clearly H_p is finite iff $gr\, H_p$ is finite. Using (4.1), Remark 3.2 and the proof of Theorem 3.1 it is enough to show that there is a natural number l with $\pi^l x_i = 0$ for $i = 1, 2, \dots, 5$ in $gr\, H_p$. Let L be the $F_p[\pi]$-Lie algebra defined by the following relations:

$$\begin{aligned} x_1x_2 &= 0, && (1) \\ x_3x_1 &= \pi x_3, && (2) \\ x_5x_4 &= \pi x_4, && (3) \\ x_3x_5 &= \pi x_1, && (4) \\ x_1x_4 &= x_2x_3 + \pi x_2 + \pi x_3, && (5) \\ x_1x_5 &= x_2x_4, && (6) \\ x_2x_5 &= x_3x_4 - \pi x_5. && (7) \end{aligned}$$

We show that in L and therefore also in $gr\, H_p$:

$$\pi^5 x_1 = 0,\ \pi^6 x_3 = 0,\ \pi^7 x_2 = 0,\ \pi^8 x_5 = 0,\ \pi^9 x_4 = 0.$$

Again let $i := x_i$ in the following calculations. By (5), (1) and (2), it follows that

$$1412 = \pi 232 + \pi^2 32. \tag{8}$$

Further we have $4112 = 4211 = 5111$ and

$$\begin{aligned}
51113 &= 511(13) + 51131 = -\pi 5113 + 51(13)1 + 51311 \\
&= -\pi 51(13) - \pi 5131 - \pi 5131 + 5(13)11 + 53111 \\
&= \pi^2 513 - 2 * 5(13)1 - 2 * \pi 5311 - \pi 1111 \\
&= \pi^2 5(13) + \pi^2 531 + 2 * \pi^2 531 - \pi 5311 \\
&= -\pi^3 53 = \pi^3 35 = \pi^4 1.
\end{aligned}$$

Therefore

$$14123 = -\pi^4 1. \tag{9}$$

By (8) and (9) we have

$$\pi 23231 - \pi^2 2331 = 0. \tag{10}$$

Further

$$\begin{aligned}
2331 &= 23(31) + 2313 = \pi 233 + 2(31)3 = 2 * \pi 233 \\
&\text{and} \\
23231 &= 232(31) + 23213 = \pi 2323 + 23123 = \pi 2323 + \pi 2323 = 2 * \pi 2323.
\end{aligned}$$

By (10), (8) and (9) it follows

$$0 = 2 * \pi^2 2323 - 2 * \pi^3 233 = 2 * \pi(\pi 2323 - \pi^2 233) = 2 * \pi 14123 = -2 * \pi^5 1.$$

Therefore it is $\pi^5 1 = 0$. Further it follows $\pi^6 3 = 0$ by (2), $\pi^7 2 = 0$ by (5), $\pi^8 5 = 0$ by (7) and $\pi^9 4 = 0$ by (3).
Consequently, we can choose $l = 9$.

Theorem 4.4 *The pro-p-presentation*

$$\begin{aligned}
\{x_1, x_2, x_3, x_4, x_5, x_6; \quad & (x_1, x_2), \\
& (x_1, x_3) = x_3^{-p},\ (x_3, x_4) = x_1^{-p}, \\
& (x_1, x_4) = (x_3, x_2) \cdot x_2^p \cdot x_3^p,\ (x_1, x_5) = (x_4, x_2), \\
& (x_2, x_5) = (x_6, x_1) \cdot x_6^p,\ (x_3, x_5) = (x_6, x_2), \\
& (x_4, x_5) = (x_6, x_3),\ (x_4, x_6) = x_4^p, \\
& (x_5, x_6) = x_5^p\}
\end{aligned}$$

defines a finite group for all $p > 2$.

Proof: We denote the group M_p. The Lie algebra $gr\, M_p$ is a homomorphic image of the following $F_p[\pi]$-Lie algebra L:

$$\begin{aligned}
x_2x_1 &= 0, && (11)\\
x_3x_1 &= \pi x_3, && (12)\\
x_4x_3 &= \pi x_1, && (13)\\
x_4x_1 &= x_2x_3 - \pi x_2 - \pi x_3, && (14)\\
x_1x_5 &= x_4x_2, && (15)\\
x_2x_5 &= x_6x_1 + \pi x_6, && (16)\\
x_3x_5 &= x_6x_2, && (17)\\
x_4x_5 &= x_6x_3, && (18)\\
x_4x_6 &= \pi x_4, && (19)\\
x_5x_6 &= \pi x_5. && (20)
\end{aligned}$$

We show that in L and therefore also in $gr\, M_p$:

$$\pi^5x_1 = 0, \quad \pi^6x_3 = 0, \quad \pi^7x_2 = 0, \quad \pi^8x_6 = 0, \quad \pi^9x_4 = \pi^9x_5 = 0.$$

Again let $i := x_i$ in the following calculation. By the Jacobi identity, (12) and (13) it is $413 = 4(13) + 431 = -\pi 43 = -\pi^2 1$. It follows $4131 = 0$. On the other hand we have $413 = 233 - \pi 23$ by (14) and so

$$\begin{aligned}
0 &= 2331 - \pi 231 && (21)\\
&= 23(31) + 2313 - \pi 2(31) - \pi 213\\
&= \pi 233 + 2(31)3 + 2133 - \pi^2 23 = \pi 233 + \pi 233 - \pi^2 23\\
&= \pi 233 + \pi(233 - \pi 23) = \pi 233 - \pi^3 1.
\end{aligned}$$

That is

$$\pi^3 1 = \pi 233, \qquad 0 = \pi 2331 \qquad (22)$$

and so $0 = \pi 2331 = \pi^2 231 = \pi^2 2(31) = \pi^3 23$ by (21). By (22) we then have $\pi^5 1 = \pi^3 233 = 0$. It follows that $\pi^6 3 = 0$ by (12), $\pi^7 2 = 0$ by (14), $\pi^8 6 = 0$ by (16), $\pi^9 4 = 0$ by (19) and $\pi^9 5 = 0$ by (20).
Now Theorem 4.4 follows from (3.3).

Remark 4.5 Here again the p-Quotient Program gives the group M_p has class 5 and order p^{94} for small odd primes. It can also be used to show that the following pro-2-presentation defines a finite 2-group with class 6 and order 2^{125}.

$$\begin{aligned}\{x_1, x_2, x_3, x_4, x_5, x_6; \quad & (x_1, x_2) \cdot x_2^2,\ (x_1, x_3) \cdot x_3^2,\ (x_4, x_1) = (x_3, x_2) \cdot x_6^2,\\ & (x_5, x_1) = (x_4, x_2),\ (x_5, x_2) = (x_4, x_3),\\ & (x_5, x_3) = (x_6, x_2),\ (x_5, x_4) = (x_6, x_3),\\ & (x_1, x_6) = x_1^2,\ (x_4, x_6) = x_4^2,\ (x_5, x_6) = x_5^2\}\end{aligned}$$

References

[1] **Brahana, H.R. :** *Finite metabelian groups and the lines of a projective four-space.* Amer. J. Math. **73**, 539-555 (1951)

[2] **Brahana, H.R. :** *Metabelian p-groups with five generators and orders p^{12} and p^{11}.* Illinois J. Math. **2**, 641-717 (1958)

[3] **Golod, E.S.** and **Shafarevich, I.R. :** *On the class field tower.* Izv. Akad. Nauk SSSR Ser. Mat. **28**, 261-272 (1964)

[4] **Havas, G., Newman, M.F.** and **O'Brien, E.A. :** *ANU p-Quotient Program (Version* 1.3*), written in* `C` *(1995), available from* `maths.anu.edu.au` *by* `anonymous ftp` *in the directory* `pub/PQ`*, as a share library with* GAP 3.4 *and as part of* `Magma`*.*

[5] **Koch, H. :** *Erzeugenden- und Relationenrang für endlich dimensionale nilpotente Liesche Algebren.* Algebra i Logika **16**, 364-374 (1977)

[6] **Kostrikin, A.I. :** *On presenting groups by generators and defining relations.* Izv. Akad. Nauk SSSR Ser. Mat. **29**, 1119-1122 (1965)

[7] **Nickel, W. :** *A Nilpotent Quotient Program (Version* 1.1*), written in* `C`*.* (1993)

[8] **Sauerbier, S. :** *Zur Darstellung von Pro-2-Gruppen durch Erzeugende und Relationen.* Wiss. Z. Päd. Hochsch. Güstrow, Math.-Nat.Fak. **1**, 27-38 (1986)

[9] **Schönert, M. et al. :** *GAP – Groups, Algorithms, and Programming.* (Release 3.4, 1995). Lehrstuhl D für Mathematik, Rheinisch-Westfälische Technische Hochschule Aachen, Germany

[10] **Wisliceny, J. :** *Zur Darstellung von Pro-p-Gruppen und Lieschen Algebren durch Erzeugende und Relationen.* Math. Nachr. **102**, 57-78 (1981)

received: September 29, 1995

Authors:

Prof. Dr. M. F. Newman
Australian National University
Mathematics, IAS
Canberra ACT 0200
Australia

Dr. G. Sauerbier
Universität Rostock
Fachbereich Mathematik
18051 Rostock
Germany

Prof. Dr. J. Wisliceny
Ernst-Moritz-Arndt-Universität
FR Mathematik/Informatik
17489 Greifswald
Germany

Rostock. Math. Kolloq. **49**, 155–162 (1995)

Subject Classification (AMS)
primär 93D20,
sekundär 30C15

F. G. Boese

Stabilität dynamischer Systeme unter Parameterunsicherheit

Gewidmet den Herren Professoren
L. Berg, W. Engel, G. Pazderski *und* H.-W. Stolle.

ABSTRAKT. In den letzten fünfzehn Jahren etwa sind in einem Gebiet der Theorie der Stabiliät dynamischer Systeme bemerkenswerte Fortschritte gemacht worden. Ziel dieses Beitrages ist es, den interessierten mathematischen Nicht-Spezialisten darauf aufmerksam zu machen und eine bestimmte Entwicklungslinie nachzuzeichnen.

SCHLAGWORTE. Dynamische Systeme, asymptotische Stabilität, charakteristische Funktionen, intervallwertige Koeffizienten, Satz von Charitonow, lineare intervallwertige Parameter

1 Problem und Motivation

Von den Anwendungen ausgehend, ist es eher die Regel als die Ausnahme, daß die Dynamik der zu untersuchenden Systeme nur ungenau bekannt ist. Ein eingeschränkter Fall hiervon ist, daß nur Systemparameter unsicher sind. Eine Systemeigenschaft von Wichtigkeit ist die *asymptotische Stabilität* (im Sinne von Ljapunow) des betrachteten Systemzustandes. Im einfachsten Fall wird man (nach Linearisierung) auf die Differentialgleichung mit konstanten reellen Koeffizienten für den skalaren Zustand $x \ : \mathbf{R}_+ \to \mathbf{R}$ geführt

$$x^{(n)}(t) + a_1(p)x^{(n-1)}(t) + \cdots + a_n(p) = 0, \qquad n \in \mathbf{N}. \tag{1}$$

Die reellen Koeffizienten $a_k(p)$, $k = 1, \cdots, n$, sind stetige Funktionen eines reellen Parametervektors $p \in \mathbf{P} \subset \mathbf{R}^m$, wobei der Parameterraum $\mathbf{P}$ kompakt ist. Es ist bekannt, daß alle Lösungen von (1) genau dann asymptotisch stabil sind, wenn das zu (1) gehörige *charakteristische Polynom* f_p in der komplexen Variablen z

$$f_p(z) := z^n + a_1(p)z^{n-1} + \cdots + a_n(p), \qquad p \in \mathbf{P}, \qquad z \in \mathbf{C}, \tag{2}$$

ein *Hurwitz-Polynom* ist, d.h., wenn alle seine Nullstellen negativen Realteil besitzen. Im Hinblick auf (1) wird f_p dann *stabil* genannt. Ferner heisst die *Familie* $\mathcal{F}_{\mathbf{P}} := \{f_p : \ p \in \mathbf{P}\}$ von charakteristischen Polynomen *stabil*, falls jedes f_p, $p \in \mathbf{P}$, stabil ist. Das Problem der Entwicklung von *Stabilitätstests* für eine gegebene (nicht notwendigerweise polynomiale) Familie $\mathcal{F}$ von charakteristischen Funktionen ist das folgende:

Finde Unterfamilien $\mathcal{F}' \subset \mathcal{F}$ möglichst kleiner Mächtigkeit, so daß die Stabiliät von $\mathcal{F}'$ die von $\mathcal{F}$ nach sich zieht. (3)

Je kleiner die Mächtigkeit von $\mathcal{F}'$ ist, je weniger Tests sind auszuführen, um die Frage der Stabilität von $\mathcal{F}$ zu entscheiden.

In Verallgemeinerung der Hurwitz-Stabilität tritt an die Stelle der geforderten Nullstellenfreiheit in der rechten abgeschlossenen Halbebende $\mathbf{H} := \{z \in \mathbf{C} : \ \mathrm{Re}(z) \geq 0\}$ der komplexen Ebene Nullstellenfreiheit in anderen Mengen $\mathbf{G} \subset \mathbf{C}$. Im Falle der Schur-Stabiliät ist $\mathbf{G}$ beispielsweis das abgeschlossende Äußere $\mathbf{G} := \{z \in \mathbf{C} : \ |z| \geq 1\}$ der offenen Einheitskreisscheibe $\mathbf{D} := \{z \in \mathbf{C} : \ |z| < 1\}$. Die unterliegenden dynamischen Systeme sind dabei die linearen zeitdiskreten, die auch in der Numerik bedeutsam sind.

2 Der Satz von Charitonow

Im Jahre 1978 wurde von V. L. Charitonow [3] eine Polynomfamilie $\mathcal{F}$ behandelt, die eine Unterfamilie $\mathcal{F}'$ sehr kleiner Mächtigkeit ermöglicht. Es war dieser Satz, der die stürmische Entwicklung in diesem Feld auslöste, nachdem er nach Jahren der Nichtbeachtung Mitte der 80er Jahre einem grossen Kreis bekannt wurde, ja fast zum mathematischen Allgemeingut wurde.

Die betrachtete Familie $\mathcal{F} := \mathcal{F}_{\mathbf{A}}$ ist die der reellen monischen Polynome mit intervallwertigen Koeffizienten,

$$\begin{aligned} \mathcal{F}_{\mathbf{A}} &:= \{f_a : \ a \in \mathbf{A}\}, \\ f_a(z) &:= z^n + a_{n-1}z^{n-1} + a_{n-2}z^{n-2} + \cdots + a_0, \quad a_k \in [\underline{a}_k, \bar{a}_k], \quad k = 0, 1, \cdots, n-1, \\ a &:= (a_0, a_1, \cdots, a_{n-1}), \\ \mathbf{A} &:= [\underline{a}_0, \bar{a}_0] \times [\underline{a}_1, \bar{a}_1] \times \cdots \times [\underline{a}_{n-1}, \bar{a}_{n-1}], \quad \underline{a}_k \leq \bar{a}_k, \quad k = 0, \cdots, n-1. \end{aligned} \tag{4}$$

Anstelle der Darstellung von $\mathcal{F}_{\mathbf{A}}$ in (4) mittels Funktionen $f_a(z)$ mit punktwertigen Koeffizienten kann man sich mit Vorteil der Auffassung der Intervallmathematik bedienen. In dieser hat man es mit einer einzigen mengenwertigen Funktion $f_{\mathbf{a}}(z)$, der Intervallerweiterung von $f_a(z)$ bezüglich der Koeffizienten, zu tun,

$$\begin{aligned} f_{\mathbf{a}}(z) &:= z^n + [\underline{a}_{n-1}, \bar{a}_{n-1}]z^{n-1} + [\underline{a}_{n-2}, \bar{a}_{n-2}]z^{n-2} + \cdots + [\underline{a}_0, \bar{a}_0], \\ \mathbf{a} &:= \left([\underline{a}_0, \bar{a}_0], [\underline{a}_1, \bar{a}_1], \cdots, [\underline{a}_{n-1}, \bar{a}_{n-1}]\right), \end{aligned} \tag{5}$$

Die Menge $f_{\mathbf{a}}(z) \subset \mathbf{C}$ ist durch die Auswertungsvorschriften

$$[a,b] := \operatorname{conv}(\{a,b\}), \quad w \cdot [a,b] := [wa, wb], \quad \mathbf{A}+\mathbf{B} := \{a+b \in \mathbf{C}: \ a \in \mathbf{A},\, b \in \mathbf{B}\} \tag{6}$$

wohlbestimmt. Durch conv in (6) wird der Menge $\mathbf{A} \subset \mathbf{C}$ ihre konvexe Hülle $\operatorname{conv}(\mathbf{A}) \subset \mathbf{C}$ zugeordnet. Das zu lösende Problem ist: Wie kann man entscheiden, ob $f_{\mathbf{a}}(z)$ nullstellenfrei in $\operatorname{Re}(z) \geq 0$ ist? Eine Strategie zur Problemlösung ist in (3) angegeben.

Die Unterfamilie $\mathcal{F}'_{\mathbf{A}}$ wird von vier explizit bekannten Extremalpolynomen gebildet,

$$\begin{aligned} \mathcal{F}'_{\mathbf{A}} &:= \{\, f_{1,1}, f_{1,2}, f_{2,1}, f_{2,2} \,\}\,, \\ f_{1,1}(z) &:= \underline{a}_0 + \underline{a}_1 z + \bar{a}_2 z^2 + \bar{a}_3 z^3 + \underline{a}_4 z^4 + \cdots\,, \\ f_{2,1}(z) &:= \bar{a}_0 + \underline{a}_1 z + \underline{a}_2 z^2 + \bar{a}_3 z^3 + \bar{a}_4 z^4 + \cdots\,, \\ f_{1,2}(z) &:= \underline{a}_0 + \bar{a}_1 z + \bar{a}_2 z^2 + \underline{a}_3 z^3 + \underline{a}_4 z^4 + \cdots\,, \\ f_{2,2}(z) &:= \bar{a}_0 + \bar{a}_1 z + \underline{a}_2 z^2 + \underline{a}_3 z^3 + \bar{a}_4 z^4 + \cdots\,. \end{aligned} \tag{7}$$

Nach den in (7) angegebenen ersten vier Unter- und Überstreichungen wiederholt sich das angegebene Muster jeweils. Mit den obigen Definitionen von $\mathcal{F}_{\mathbf{A}}$ und $\mathcal{F}'_{\mathbf{A}}$ gilt der

Satz 2.1 (von Charitonow) *$\mathcal{F}_{\mathbf{A}}$ von* (4) *ist genau dann stabil, wenn $\mathcal{F}'_{\mathbf{A}}$ von* (7) *stabil ist.*

Es bereitet keine Schwierigkeiten, die vier Polynome aus $\mathcal{F}'_{\mathbf{A}}$, etwa mittels der Routh'schen Kettenbruchentwicklung, auf Stabilität zu testen. Der usprüngliche Beweis von Charitonow verwandte den Satz von Hermite-Biehler und ein bekanntes Induktionsargument nach dem Polynomgrad n. Heute sind die Beweise einfacher und von größerer Tragweite und nicht ohne ästhetischen Reiz. Vor einem Beweis werden seine Bausteine vorgestellt.

Lemma 2.2 *Ist $f(z) := z^n + a_{n-1}z^{n-1} + \cdots + a_0$ ein reelles Hurwitz-Polynom, so sind alle Koeffizienten a_k, $k = 0, \cdots, n-1$, positiv. Ferner ist* $\arg[f(iy)]$ *eine strikt wachsende Funktion in $y \in \mathbf{R}$ mit* $\arg[f(0)] = 0$ *und* $\arg[f(+i\infty)] = n\pi/2$.

Beide Eigenschaften sind wohlbekannt und folgen unschwer aus der faktorisierter Form von $f(z)$.

Es sei $\mathcal{F}$ eine Familie holomorpher Funktionen. Als *Wertmenge* von $\mathcal{F}$ im Punkt z des Holomorphiegebietes wird das Bild von $\mathcal{F}$ unter der Auswertungsabbildung $E: \ \mathcal{F} \to \mathbf{C}$,

$$\mathbf{W}(z) := \{\, f(z): \ f \in \mathcal{F} \,\} \tag{8}$$

bezeichnet. Für die Spezialisierung $f(z) := f_{\mathbf{a}}(z)$ ist also $\mathbf{W}(z) := f_{\mathbf{a}}(z)$. Im Falle einer parametrischen Familie wie $\mathcal{F}_{\mathbf{A}}$ kann die Wertmenge auch als das Bild des Parameterraumes

$\mathbf{P}$, also des Intervalles $\mathbf{A}$ des Koeffizienten-Parameterraumes $\mathbf{R}^n$ im vorliegenden Fall, in $\mathbf{C}$ angesehen werden. Wie noch zu sehen sein wird, wird $\mathbf{W}(z)$ nur längs des Randes $\partial\mathbf{H}$ der rechten Halbebene $\mathbf{H}$, der imaginären Achse also, benötigt.

Es stellt sich heraus, daß in unserem Falle $\mathbf{W}(iy)$, $y \in \mathbf{R}$, achsenparallele Rechtecke in der komplexen w-Bildebene sind.

Lemma 2.3 *Für $y \geq 0$ gilt*

$$\mathbf{W}(iy) := \operatorname{conv}\left(\{f_{1,1}(iy), f_{2,1}(iy), f_{1,2}(iy), f_{2,2}(iy)\}\right). \tag{9}$$

Beweis: Für $f \in \mathcal{F} := \mathcal{F}_{\mathbf{A}}$ und $y \geq 0$ gilt mit $a_n := 1$ und positiven restlichen Koeffizienten $a_k > 0$

$$\begin{aligned}
\min_{f\in\mathcal{F}} \operatorname{Re}\{f(iy)\} \leq \operatorname{Re}\{f(iy)\} &\leq \max_{f\in\mathcal{F}} \operatorname{Re}\{f(iy)\}, \\
\sum_{0\leq 2k\leq n} y^{2k} \min_{\underline{a}_{2k}\leq a_{2k}\leq \bar{a}_{2k}} \operatorname{Re}\{i^{2k}a_{2k}\} \leq \operatorname{Re}\{f(iy)\} &\leq \sum_{0\leq 2k\leq n} y^{2k} \max_{\underline{a}_{2k}\leq a_{2k}\leq \bar{a}_{2k}} \operatorname{Re}\{i^{2k}a_{2k}\}, \\
\sum_{0\leq 2k\leq n} y^{2k} \min_{a_{2k}\in\{\underline{a}_{2k},\bar{a}_{2k}\}} (-1)^k a_{2k} \leq \operatorname{Re}\{f(iy)\} &\leq \sum_{0\leq 2k\leq n} y^{2k} \max_{a_{2k}\in\{\underline{a}_{2k},\bar{a}_{2k}\}} (-1)^k a_{2k}, \\
\sum_{0\leq 2k\leq n} y^{2k}(-1)^k a_{2k,(k+1)\mathrm{mod}2} \leq \operatorname{Re}\{f(iy)\} &\leq \sum_{0\leq 2k\leq n} y^{2k}(-1)^k a_{2k,k\mathrm{mod}2}, \\
\operatorname{Re}\{f_{1,j}(iy)\} \leq \operatorname{Re}\{f(iy)\} &\leq \operatorname{Re}\{f_{2,j}(iy)\}, \qquad j = 1,2.
\end{aligned} \tag{10}$$

In (10) ist $a_{2k,0} := \bar{a}_{2k}$ und $a_{2k,1} := \underline{a}_{2k}$. Für die Imaginärteile gilt die analoge Ungleichung

$$\operatorname{Im}\{f_{j,1}(iy)\} \leq \operatorname{Im}\{f(iy)\} \leq \operatorname{Im}\{f_{j,2}(iy)\}, \qquad j = 1,2. \tag{11}$$

Damit ist $\mathbf{W}(iy)$ als Teilmenge der rechten Seite von (9) erkannt. Ist nun für gegebenes $y \geq 0$ der Punkt $w \in \mathbf{W}(iy)$ gegeben, so lassen sich, wie man (10) entnehmen kann, Koeffizienten a_0 bis a_{n-1} so finden, daß w von der zugehörigen Funktion f_a angenommen wird, $w = f(iy)$. Damit ist die Gleichheit der beiden Mengen in (9) gezeigt. ■

Unter passenden Voraussetzungen ändert sich die Anzahl der Nullstellen, die eine holomorphe, parameterabhängige Funktion $f_p(z)$ bei Parameteränderung $p \in \mathbf{P}$ in einem Bereich $\mathbf{G}$ hat, nur, wenn Nullstellen durch den Bereichsrand $\partial\mathbf{G}$ hindurchtreten. Das nächste Lemma klärt, unter welchen Vorausssetzungen das so ist.

Von $\mathbf{P}$ wird verlangt:

> $\mathbf{P}$ ist eine kompakte, einfach zusammenhängende Menge eines topologischen Hausdorffraumes. (12)

Die Menge $\mathbf{G} \subset \mathbf{C}$ hat die Eigenschaften:

1^0 $\mathbf{G}$ ist einfach zusammenhängend, beschränkt und abgeschlossen,
2^0 Die Randkurve $\partial\mathbf{G}$ besteht aus endlich vielen stückweise differenzierbaren Kurven. (13)

Demnach kann $\partial\mathbf{G}$ höchstens endlich viele Ecken haben.

Von den Funktionen f_p der Familie $\mathcal{F} := \{\, f_p(z) : \; p \in \mathbf{P} \,\}$ wird verlangt:

1^0 $f_p(z)$ ist holomorph bezüglich z in $\mathbf{G}$ für alle $p \in \mathbf{P}$, (14)
2^0 $f_p(z)$ ist stetig bezüglich p für $p \in \mathbf{P}$ für alle $z \in \mathbf{G}$.

Lemma 2.4 *Es sei* $\mathcal{F} := \{f_p(z) : \; p \in \mathbf{P}\}$ *eine parametrische Familie von in* $\mathbf{G}$ *holomorphen Funktionen. Dabei erfülle* $\mathbf{P}$ (12), $\mathbf{G}$ (13) *und die* f_p (14). *Dann hängt die Anzahl der Nullstellen von* f_p *in* $\mathbf{G}$ *nicht von* $p \in \mathbf{P}$ *ab, falls gilt:*

$$f_p(z) \neq 0 \quad \textit{längs} \quad \partial\mathbf{G} \quad \textit{für alle} \quad p \in \mathbf{P}. \tag{15}$$

Einen Beweis findet man in [2]. Die Stetigkeit von f_p im Parameter p führt über den Satz über die implizite Funktion zur Stetigkeit der Nullstellen in p.

Wie schon im Falle der Hurwitz-Stabilität benötigt man Lemma 2.4 auch für nichtkompakte $\mathbf{G}$. Falls aber die Familie $\mathcal{F}$ nur Nullstellen in einem kompakten Teil von $\mathbf{G}$ hat, genügt es, nur $\mathbf{G}$ wie in (13) zu betrachten.

Beweis zum Satz von Charitonow

Ist $\mathcal{F}_{\mathbf{A}}$ stabil, so ist es auch $\mathcal{F}'_{\mathbf{A}}$, da letztere Familie zur ersteren gehört. Zu zeigen ist, daß auch die Umkehrung gilt. Aus $f_a(z)/z^n \to 1$ für $|z| \to \infty$ für alle $a \in \mathbf{A}$ folgt, daß mögliche Nullstellen von f_a in $\mathbf{H}$ in $\mathbf{H} \cap \mathbf{D}_R$, wobei R in $\mathbf{D}_R := \{z \in \mathbf{C} : \; |z| \leq R\}$ groß genug ist. Wir wählen nun $\mathbf{P} := \mathbf{A}$, $f_p := f_a$, $\mathbf{G} := \mathbf{H} \cap \mathbf{D}_R$ in Lemma 2.4. Ist R genügend groß, so ist (15) auf dem Halbkreis von $\partial\mathbf{H} \cap \mathbf{D}_R$ erfüllt. Wir zeigen, daß (15) auch längs des auf der imaginären Achse liegenden Stückes von $\partial\mathbf{G}$ gilt. Nach Voraussetzung gilt (15) für $f_a \in \mathcal{F}'_{\mathbf{A}}$. Zu zeigen ist dies auch für die restlichen f_a.

Da $f_a(z)$ reelle Funktionen sind, genügt es, $z := iy$, $y \geq 0$, zu wählen. Wir nehmen an, es gäbe ein $y \geq 0$ und eine Kante von $\mathbf{W}(iy)$, die den Ursprung der komplexen w-Bildebene im Inneren enthält. Da das Kantenendpunktpaar durch Hurwitz-Polynome $f_a \in \mathcal{F}'_{\mathbf{A}}$ gebildet

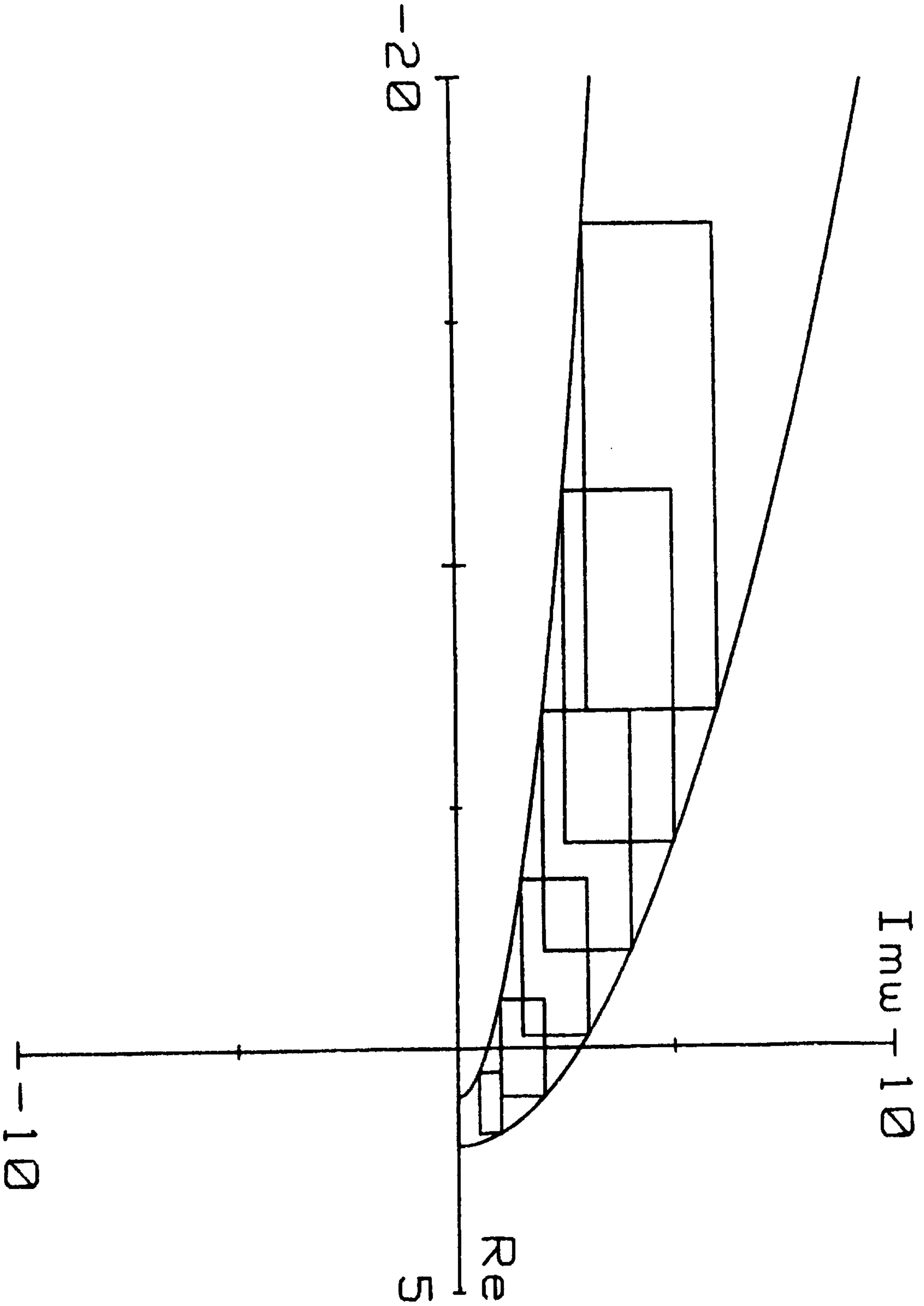

Fig. 1: Die Wertmengen $\mathbf{W}(iy) := [1 - 2y^2, 2 - y^2] \times [iy, 2iy]$ für $y = k/2$, $k = 1(1)6$, für $f_{\mathbf{a}}(z)$ aus (16) nebst der Enveloppe-Parabeln $e_1(y) := 1 + iy - 2y^2$ und $e_2(y) := 2 + i2y - y^2$.

wird, muß nach Lemma 2.2, für $\delta y > 0$ klein genug, die benachbarte Wertmenge $\mathbf{W}(iy+i\delta y)$ eine nicht achsenparallele Kante besitzen. Nach Lemma 2.3 kann das aber nicht sein, da $\partial\mathbf{W}(iy)$ für alle $y \geq 0$ ein achsenparalleles Rechteck ist. Wir müssen daher die Annahme fallen lassen, daß es ein $\mathbf{W}(iy)$ gibt, das den Ursprung $w = 0$ überdeckt. Nach Lemma 2.4 hat $\mathcal{F}_{\mathbf{A}}$ in $\mathbf{H}$ eine konstante Anzahl von Nullstellen. Da $\mathcal{F}'_{\mathbf{A}} \subset \mathcal{F}_{\mathbf{A}}$ stabil ist, ist es auch $\mathcal{F}_{\mathbf{A}}$.

■

Figur 1 zeigt für

$$f_{\mathbf{a}}(z) := [1,2] + [1,2]z + [1,2]z^2 \tag{16}$$

die Wertmengen

$$\mathbf{W}(iy) := [1-2y^2, 2-y^2] \times [iy, 2iy] \tag{17}$$

für $y = k/2$, $k = 1(1)6$, in der w-Bildebene. Zusätzlich sind die Enveloppe-Parabeln $e_1(y) := 1+iy-2y^2$ und $e_2(y) := 2+i2y-y^2$ gezeichnet. Ersichtlich überdeckt keine der Wertmengen den Ursprung $w = 0$. Aus der Positiviät der Koeffizientenintervalle in (16) folgt nach dem Hurwitz-Kriterium sofort die Stabiliät von $f_{\mathbf{a}}$ aus (16). Teilt man in (16) durch das führende Intervall, so erhält $f_{\mathbf{a}}$ die bisher vorausgesetzte monische Form

$$f_{\mathbf{a}}(z) := [1/2, 2] + [1/2, 2]z + z^2. \tag{18}$$

Wenn der führende Intervallkoeffizient den Ursprung von $\mathbf{R}$ nicht enthält, kann die Form (5) immer hergestellt werden. In den Fällen niedriger Grade $n < 6$ umfasst $\mathcal{F}'_{\mathbf{A}}$ weniger als 4 Extremalpolynome. Die Erweiterung auf komplexwertige Intervalle in Form von achsenparallelen Rechtecken in $\mathbf{C}$ wurde ebenfalls von Charitonow [4] vorgenommen,

$$\mathbf{a} := \left([\underline{a}_0, \bar{a}_0] \times i\,[\underline{b}_0, \bar{b}_0], [\underline{a}_1, \bar{a}_1] \times i\,[\underline{b}_1, \bar{b}_1], \cdots, [\underline{a}_{n-1}, \bar{a}_{n-1}] \times i\,[\underline{b}_{n-1}, \bar{b}_{n-1}] \right),$$
$$\underline{a}_k \leq \bar{a}_k, \quad \underline{b}_k \leq \bar{b}_k. \tag{19}$$

Hierbei geht die Symmetrie zur imaginären Achse verloren. Das führt zu einem $\mathcal{F}'_{\mathbf{A}}$ mit 8 anstatt der 4 Extremalpolynomen im reellen Fall. Ist man ausschließlich am Beweis des Satzes von Charitonow interessiert, so läßt sich die Beweisanlage noch straffen.

3 Charakteristische Funktionen mit linearen Parametern

Im allgemeinen Falle von charakteristischen Funktionen mit reellen intervallwertigen linearen Parametern tritt an die Stelle von $f_{\mathbf{a}}(z)$ aus (5) die Funktion

$$\begin{aligned} f_{\mathbf{p}}(z) &: = f_0(z) + [\underline{p}_1, \bar{p}_1]f_1(z) + [\underline{p}_2, \bar{p}_2]f_2(z) + \cdots + [\underline{p}_n, \bar{p}_n]f_n(z), \\ \mathbf{p} &: = \Big([\underline{p}_1, \bar{p}_1], [\underline{p}_2, \bar{p}_2], \cdots, [\underline{p}_n, \bar{p}_n] \Big). \end{aligned} \tag{20}$$

Die $f_k(z)$, $k = 0, \cdots, n$ sind holomorph in einer Umgebung von $\mathbf{G}$.

Im generischen Fall ist die Wertmenge $\mathbf{W}(iy)$ zu $f_{\mathbf{p}}$ aus (20) ein konvexer, zentralsymmetrischer Polygonbereich mit $2n$ Seiten. Von diesem generischen Fall aus beurteilt, sind die im Satz von Charitonow auftretenden achsenparallelen Rechtecke hochgradig degeneriert. Alle in (20) auftretenden n Richtungen $\arg[f_k(iy)]$, $k = 1(1)n$, fallen in die reelle oder die imaginäre Richtung. Wächst $|f_0(z)|$ für $|z| \to \infty$ schneller als die $|f_k(z)|$, $k = 1(1)n$, so ist die Enthaltenseinsrelation $0 \notin \mathbf{W}(iy)$ im Falle der Hurwitz-Stabilität für $|y|$ aus einem beschränkten Intervall zu testen. Der Fall von zu testenden Unterfamilien $\mathcal{F}'$ mit endlicher Mächtigkeit ist eher als Ausnahme anzusehen. Eine eigene Theorie ist entstanden, um solche für die Praxis willkommenen Fälle aufzuspüren.

Buchdarstellungen des Gebietes sind jüngst von Barmish [1] und Kogan [5] vorgelegt worden.

Literatur

[1] **Barmish, R. B. :** *New Tools for Robustness of Linear Systems.* New York 1994

[2] **Boese, F. G. :** *On the stability of interval families of characteristic functions depending linearly on parameters.* J. Math. Anal. Appl. **188**, 2, 472-499 (1994)

[3] **Charitonow, V. L. :** *Über die asymptotische Stabilität der Gleichgewichtslage einer Familie linearer Differentialgleichungen.* Differentsial'nye Uravneniya **14**, 11, 2086-2089 (1978)

[4] **Charitonow, V. L. :** *Über eine Verallgemeinerung des Stabilitätskriteriums.* Isv. Akad. Nauk Kazakh. SSR Ser. Fiz. Mat. **1**, 53-57 (1978)

[5] **Kogan, J. :** *Robust Stability and Convexity.* Lecture Notes in Control and Information Sciences **201**, Berlin 1995

eingegangen: 11. Oktober 1995

Autor: Dr. F. G. Boese
Ganghoferstr. 81
D-81373 München
Germany

e-mail: gub@mpe-garching.mpg.de

Rostock. Math. Kolloq. **49**, 163–184 (1995)

Subject Classification (AMS)
Primary 76E30;
Secondary 58F12

Peter Takáč

Dynamics on the Attractor for the Complex Ginzburg-Landau Equation[1]

Dedicated to the professors of mathematics
L. Berg, W. Engel, G. Pazderski, *and* H.- W. Stolle.

ABSTRACT. We present a numerical study of the large-time asymptotic behavior of solutions to the one-dimensional complex Ginzburg-Landau equation with periodic boundary conditions. Our parameters belong to the Benjamin-Feir unstable region. Our solutions start near a pure-mode rotating wave that is stable under sideband perturbations for the Reynolds number R ranging over an interval (R_{sub}, R_{sup}). We find sub- and super-critical bifurcations from this stable rotating wave to a stable 2-torus as the parameter R is decreased or increased past the critical value R_{sub} or R_{sup}. As $R > R_{sup}$ further increases, we observe a variety of dynamical phenomena, such as a local attractor consisting of three unstable manifolds of periodic orbits or 2-tori cyclically connected by manifolds of connection orbits. We compare our numerical simulations to both rigorous mathematical results and experimental observations for binary fluid mixtures.

KEY WORDS. Periodic orbit; 2- and 3-tori; stability; local attractor; psedo-spectral method; Fourier modes

1 Introduction

An important tool in the theory of phase transitions and instability waves is the time-dependent complex Ginzburg-Landau equation (CGL, for short)

$$\partial_t A = (1 + i\nu)\partial_{xx}^2 A + (R - (1 + i\mu)|A|^2)A, \quad -\infty < x < \infty,\ t \geq 0, \tag{1}$$

[1] This work was supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research, United States Department of Energy, under Contract W-31-109-Eng-38 through Argonne National Laboratory. The author expresses his appreciation to the Mathematics and Computer Science Division at Argonne National Laboratory for the hospitality during his 1990–91 academic leave.

where x and t are the spatial and temporal variables, respectively, and the unknown complex-valued function $A(x,t)$ represents an order parameter or a wave function. Here, μ, ν and $R \in (-\infty, \infty)$ are parameters. Originally discovered by GINZBURG AND LANDAU [6] for a phase transition in superconductivity, this equation was subsequently derived for instability waves in hydrodynamics such as the nonlinear growth of RAYLEIGH-BÉNARD convective rolls (NEWELL AND WHITEHEAD [17]), the appearance of TAYLOR vortices in the COUETTE flow between counter-rotating cylinders (STUART AND DI PRIMA [21]), and the development of TOLLMIEN-SCHLICHTING waves in plane POISEUILLE flows (BLENNERHASSETT [1]). Also instability waves for perturbation concentration in chemically reacting and diffusing systems are described by the CGL equation (KURAMOTO AND TSUZUKI [12]). In these applications, Eq. (1) describes the small and slowly varying (in space and time) amplitude and phase of a mode that bifurcates via an oscillatory instability from a homogeneous basic state (NEWELL [16]). The parameter R corresponds to a REYNOLDS number; we use it as the bifurcation (or control) parameter. We impose periodic boundary conditions

$$A(x+1,t) = A(x,t), \quad -\infty < x < \infty,\ t \geq 0. \tag{2}$$

For example, periodic boundary conditions are appropriate for experiments with RAYLEIGH-BÉNARD convective rolls in a binary fluid mixture contained in a cell of annular geometry, where the complex amplitude $A(x,t)$ describes the wave moving along the boundary between two concentric convective rolls, cf. JANIAUD ET AL. [8, 9]. The oscillatory instability of the two convective rolls develops as the control parameter R increases and crosses a critical value $R_0 > 0$. At $R = R_0$ the boundary between the two concentric rolls is a circle. A generic point on this circle is determined by its azimuthal angle θ, $0 \leq \theta < 2\pi$. Hence $x = \theta/2\pi$ in Eqs. (1, 2). For $R > R_0$ near R_0, after a transient, a wave rotating either clockwise or counter-clockwise settles along the circular boundary between the two rolls. The complex amplitude $A(x,t)$ of this *rotating wave* is a solution of Eqs. (1, 2) given by

$$A(x,t) = a_n e^{i(k_n x - \omega_n t)}, \quad n = 0, \pm 1, \pm 2, \cdots, \tag{3}$$

where $k_n = 2n\pi$ is the wavenumber of the annular spatial pattern with $|n|$ wavelengths, a_n is a complex constant satisfying $|a_n|^2 = R - k_n^2$, and $\omega_n = \mu R + (\nu - \mu)k_n^2$. The real constants μ and ν are determined by measuring the angular frequency ω_n for several values of R. As R further increases from $R_0 \equiv R_{0,n}$ and crosses a critical value $R_1 \equiv R_{1,n} > R_0$, the spatial pattern of the rotating wave starts to exhibit an *amplitude modulation* of the form

$$|A(x,t)| = a(x - ct), \quad a \not\equiv \text{const}, \tag{4}$$

with a second frequency $c \equiv c(R)$ near $c_n = c(R_1) \neq 0$ for $R > R_1$ near R_1. This temporally *biperiodic pattern* is characterized by the complex wave function

$$A(x,t) = (1 + U(x - ct))a_n e^{i[k_n(x-ct)-\omega t]} \equiv V(x - ct)e^{-i\omega t}, \tag{5}$$

where U is a small relative perturbation of the rotating wave $a_n e^{i[k_n(x-ct)-\omega t]}$ and ω is a small perturbation of ω_n, for $R - R_1 > 0$ small (cf. Takáč [22]).

In our numerical simulations we fix the constants $\mu = -1$ and $\nu = 5$, and thus, by the NEWELL criterion [15], we are in the BENJAMIN-FEIR *unstable* region $1 + \mu\nu < 0$. We focus on the large-time asymptotic behavior of the solutions $A(x,t)$ to Eqs. (1, 2) starting from an initial distribution $A(x,0)$ near the rotating wave (3) for $n = 1$. In particular, we investigate the sideband instability of this rotating wave, cf. ECKHAUS [5]. We have performed numerical simulations for a large range of values of the bifurcation parameter $R \in (0,\infty)$. We report here only those cases which clearly exhibit local attractors of special interest. For particular values of R, we have obtained stable rotating waves, 2-tori, and periodic orbits that have a temporally constant modulus which becomes time-periodic as R passes a critical value. We determine several critical values of R at which a bifurcation from a stable periodic orbit to a stable 2-torus takes place. We have also obtained local attractors consisting of two or three unstable manifolds of periodic orbits or 2-tori cyclically connected by manifolds of connection orbits, where the manifolds of periodic orbits or 2-tori have distinct dimensions. We compare these periodic orbits and 2-tori with those ones previously obtained by rigorous analysis and/or numerical simulations, cf. NEWTON AND SIROVICH [18, 19], SIROVICH AND NEWTON [20], and TAKÁČ [22]. This comparison provides some analytic background for our numerical simulations. We emphasize that our simulations using periodic boundary conditions (2) are different from those performed in MOON ET AL. [14], KEEFE [10], and DOELMAN [3] with Neumann boundary conditions. Since we are also interested in the *stability* of the solutions to Eq. (1) in a suitable Banach or Hilbert space of smooth 1-periodic functions $f\colon \mathbb{R} \to \mathbb{C}$, we do *not* presuppose any *special form* of these solutions. The numerical simulations presented in this article are the first ones studying the stability of 2-tori for the full system (1, 2). As usual, we write $\mathbb{R} = (-\infty,\infty)$ and $\mathbb{C} = \mathbb{R} \oplus i\mathbb{R}$ to denote the real line and the complex plane, respectively.

Our numerical integration of the evolution equation (1) uses a pseudo-spectral method (PSM, for short) for sufficiently long time (from $t = 0$ up to $t = 1000$). This PSM consists of an approximation by FOURIER series in space (with $-N^{\text{th}}$ through N^{th} FOURIER modes, $N = 15$, 31 and 63) combined with a fourth-order RUNGE-KUTTA discretization in time (with the time step $h = 0.0001$, 0.0002 and 0.0005). We have implemented this method on an Ardent Titan Supercomputer with four parallel processors using a FORTRAN code with both vector and parallel optimizations. All computations have been carried out in double precision complex arithmetics.

This article is organized as follows. In Section 2 we study bifurcations from rotating waves (3) to the 2-tori (5) together with their stability. In Section 3 we observe certain periodic orbits that have a temporally constant modulus which becomes time-periodic as R passes a critical value. In Section 4 we obtain local attractors consisting of two or three unstable manifolds of periodic orbits or 2-tori cyclically connected by manifolds of connection orbits. Finally, Section 5 contains a discussion.

2 Bifurcations from Rotating Waves

In this section we investigate bifurcations from stable rotating waves of the form (3) to the 2-tori (5) together with the exchange of stability, for $n = 1$. We find a *stable* 2-torus (5) for $137.9 \le R \le 139.7$ which bifurcates *supercritically* from a stable rotating wave (3). A perturbation of the rotating wave (3) can be written as

$$A(x,t) = (1 + B(x,t))a_n e^{i(k_n x - \omega_n t)}, \tag{6}$$

where B is small enough. The sideband instability analysis is carried out by inserting Eq. (6) into (1) and retaining only the terms of the first order in B (cf. Doering et al. [4] or Takáč [22]):

$$\partial_t B = (1 + i\nu)\partial_{xx}^2 B + 2i(1 + i\nu)k_n \partial_x B - (1 + i\mu)|a_n|^2(B + B^*), \tag{7}$$

where $|a_n|^2 = R - k_n^2$. The asterisk * denotes the complex conjugate. The sideband stability of the zero solution of Eq. (7) to perturbations of a discrete wavenumber $k_m = 2\pi m$, $m = \pm 1, \pm 2, \cdots$, is determined by writing

$$B(x,t) = b^+(t)e^{ik_m x} + (b^-(t))^* e^{-ik_m x}. \tag{8}$$

The evolution equations for the complex amplitudes $b^+(t)$ and $b^-(t)$ have the form

$$\frac{d}{dt}\begin{pmatrix} b^+ \\ b^- \end{pmatrix} = -\begin{pmatrix} C^+ & (1 + i\mu)|a_n|^2 \\ (1 - i\mu)|a_n|^2 & (C^-)^* \end{pmatrix} \begin{pmatrix} b^+ \\ b^- \end{pmatrix} \tag{9}$$

where

$$C^\pm = (1 + i\nu)k_m^2 \pm 2(1 + i\nu)k_m k_n + (1 + i\mu)|a_n|^2. \tag{10}$$

The *neutral stability curves* (R, k_n^2) for the zero solution of Eq. (9) satisfy the relation (cf. [4, 22])

$$4k^2 \left[1 + \left(\frac{\nu k_m^2 + \mu(R - k^2)}{k_m^2 + R - k^2} \right)^2 \right] = 2(1 + \mu\nu)(R - k^2) + (1 + \nu^2)k_m^2 \tag{11}$$

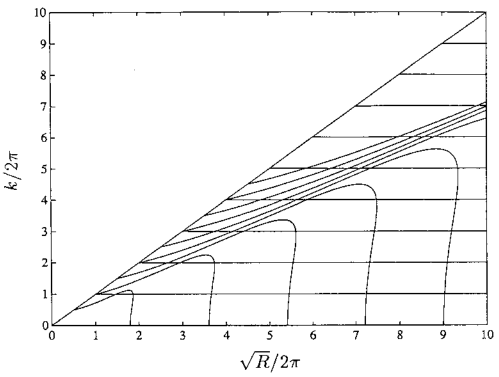

Figure 1: Neutral stability curves in the $(R^{1/2}, k)$-plane for the parameter values $(\mu, \nu) = (-1, 5)$.

where the continuous variable k replaces the discrete wavenumber $k = k_n$. For $(\mu, \nu) = (-1, 5)$ these curves are plotted in the $(\sqrt{R}, k)$-plane in Fig. 1. Each curve corresponds to a fixed value of $|m| = 1, 2, \cdots$. For a fixed value of $R > 0$, the rotating wave (3) is stable to sideband perturbations (8) if and only if exactly one solution k of Eq. (11) satisfies $k > |k_n|$. In particular, the rotating wave for $n = 1$ is stable to *all* sideband perturbations if and only if the REYNOLDS number R belongs to an interval (R_{sub}, R_{sup}). From a magnification of the graph in Fig. 1 (cf. TAKÁČ [22]) for $m = n = 1$, we have found the numerical values (precise up to computer round-off errors)

$$R_{sub} = 84.96 \quad \text{and} \quad R_{sup} = 137.90. \tag{12}$$

The existence and uniqueness (up to shifts in space and time) of the 2-tori having the form (5) has been proved by TAKÁČ [22] using bifurcation theory for $R_{sub} - R > 0$ and $R - R_{sup} > 0$ small enough. The second frequency $c = c_n$ for the critical values $R = R_{sub}$ and $R = R_{sup}$ can be computed in a similar way as the neutral stability curves. We insert Eq. (5) into (1) and retain only the terms of first order in U (cf. TAKÁČ [22])

$$(1 + i\nu)U'' + (c_n + 2i(1 + i\nu)k_n)U' - (1 + i\mu)|a_n|^2(U + U^*) = 0, \tag{13}$$

where $|a_n|^2 = R - k_n^2$. This equation has a nonzero solution of the form

$$U(x) = b^+ e^{ik_m x} + (b^-)^* e^{-ik_m x} \tag{14}$$

if and only if the linear system

$$\begin{pmatrix} C^+ - ik_m c_n & (1 + i\mu)|a_n|^2 \\ (1 - i\mu)|a_n|^2 & (C^-)^* - ik_m c_n \end{pmatrix} \begin{pmatrix} b^+ \\ b^- \end{pmatrix} = 0 \tag{15}$$

has a nonzero solution, where $C^{\pm}$ is defined by Eq. (10). This is the case if and only if R satisfies Eq. (11) and

$$c = 2k(\nu - \mu)\frac{R - k^2}{k_m^2 + R - k^2}. \tag{16}$$

The numerical values (12) for $m = n = 1$ yield the respective frequencies

$$c_{sub} = 40.36 \quad \text{and} \quad c_{sup} = 53.81. \tag{17}$$

To study the stability of the 2-tori from Eq. (5) we have used numerical integration of the evolution equation (1). Applying our pseudo-spectral method (PSM, for short, described in Section 1) for sufficiently long time (from $t = 0$ up to $t = 1000$), we have obtained the following numerical results for $\mu = -1$ and $\nu = 5$:

2.1 A supercritical bifurcation at $R = R_{sup}$

The numerical value of R_{sup} obtained by PSM coincides with its value from (12) within the indicated precision, $R_{sup}^{num} = 137.9$. For $R = R_{sup}^{num}$ the numerical solution of (1) slowly looses all FOURIER modes except for the first one, thus converging towards the pure-mode rotating wave (3) with $n = 1$. Various initial values at $t = 0$ led to a fast transition before $t = 1$ into a state with the first FOURIER mode more than 10^3-times larger than the remaining ones which also decayed exponentially with the increasing wavenumber. The decay of these sideband FOURIER modes ($n \neq 1$) then slowly continued throughout the entire evolution ($0 \leq t \leq 1000$). On the other hand, for $R = 138.0$ a similar transition was observed, but after $t = 1$ the decaying sideband modes started to settle away from zero until $t = 100$ when their decay ceased completely. The first mode was on the order of 10^5-times larger than the zeroth and second modes. From the graph of the amplitude modulation (4) the frequency c was found to be $c_{sup}^{num} = 1/0.0186 = 53.76$. Increasing R up to the value $R = 139.7$ we observe increasing sideband modes. The amplitudes of these modes are constant in time with nearly linear logarithmic decay (Fig. 2(a)). The graph of the amplitude modulation (4) in Fig. 2(b) gives the frequency $c = 1/0.0194 = 51.55$ at $R = 139.7$. For $137.9 \leq R \leq 139.7$ the stability of this temporally biperiodic motion (2-torus) was tested by various small perturbations. At $R = 139.8$ this 2-torus becomes unstable. The zeroth FOURIER mode slowly increases and becomes the dominant one. Eventually the amplitudes of all modes become constant in time with nearly linear logarithmic decay again (Fig. 3(a)). The graph of the amplitude modulation (4) in Fig. 3(b) is time-independent ($c = 0$). During the transition from a 2-torus to a limit-cycle, the amplitude (4) decreases to zero at one point where a phase slip occurs.

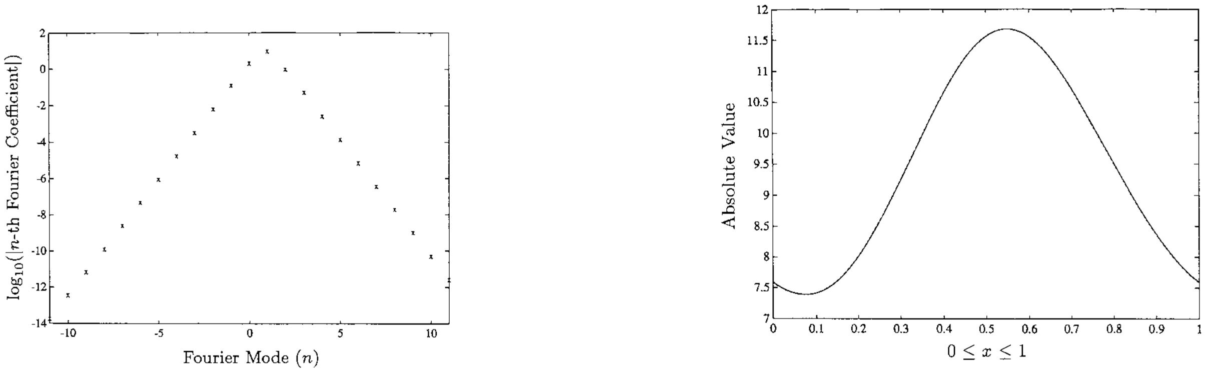

Figure 2: For $R = 139.7$, (a) logarithmic decay of Fourier modes; (b) amplitute modulation.

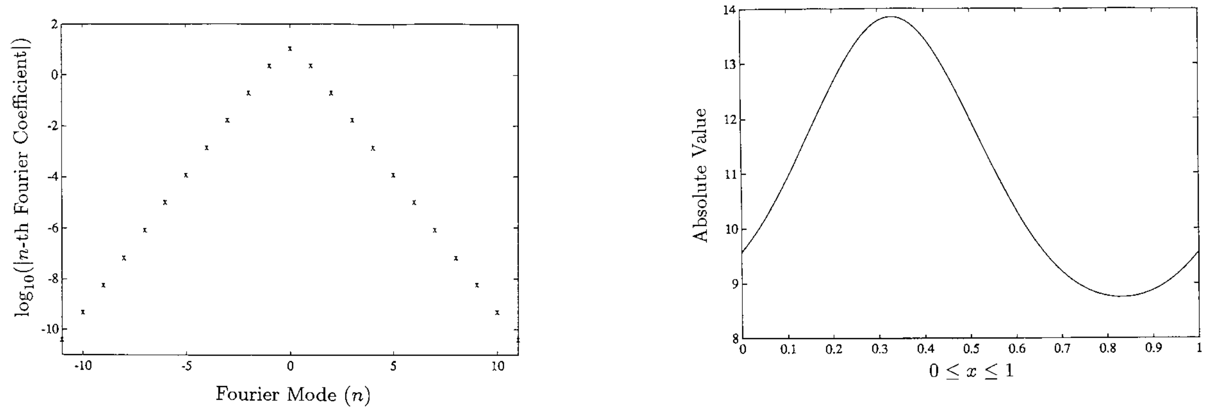

Figure 3: For $R = 139.8$, (a) logarithmic decay of Fourier modes; (b) amplitude modulation.

2.2 A subcritical bifurcation at $R = R_{sub}$

The numerical value of R_{sup} obtained by PSM differs from its value from (12), $R_{sub}^{num} = 84.99$. Similarly to the case $R = R_{sup}^{num}$, for $R = R_{sub}^{num}$ the numerical solution of (1) slowly looses all FOURIER modes except for the first one, thus converging towards the pure-mode rotating wave (3) with $n = 1$. For $R = 84.98$ a stable 2-torus (5) was observed with $c_{sub}^{num} = 1/0.0248 = 40.32$. At $R = 84.97$ this 2-torus becomes unstable. Again, the zeroth FOURIER mode slowly increases and becomes the dominant one with the remaining FOURIER modes approaching zero, and thus the rotating wave (3) with $n = 0$ is obtained.

To summarize the results of this section, we have found a *stable* 2-torus (5) for $137.9 \leq R \leq 139.7$ which bifurcates *supercritically* from a stable rotating wave (3) for $n = 1$. Although several previous articles [2, 8, 9, 11, 13] studied solutions similar to (5), their stability under the temporal evolution was not known.

3 Bifurcations from Some Periodic Orbits

In this section we are concerned with primary and secondary bifurcations from the STOKES *wave* $A(x,t) = a_0 e^{-i\omega_0 t}$ which is the special case of a rotating wave (3) for $n = 0$. In the next section we will need these bifurcations in order to be able to explain the dynamics on some local attractors for the CGL equation. Some analytic and computational background for these simulations can be found in NEWTON AND SIROVICH [18, 19] and SIROVICH AND NEWTON [20]. Therefore we present here only those results that are either new or necessary for a proper understanding of the next section. The reader is referred to GUCKENHEIMER AND HOLMES [7] for general facts about dynamical systems and bifurcations.

3.1 Separable periodic orbits

We begin with the stability of the STOKES wave for $(\mu,\nu) = (-1,5)$. Similarly as in the previous section, the STOKES wave ($n = 0$) is stable to *all* sideband perturbations if and only if the REYNOLDS number R belongs to an interval $(0, R_I)$, where

$$R_I = -2\pi^2(1+\nu^2)/(1+\mu\nu) \doteq 128.3049 \tag{18}$$

is the value of R satisfying Eq. (11) for $k = 0$ and $m = \pm 1$, see Fig. 1. According to NEWTON AND SIROVICH [18], as the parameter R crosses R_I, the STOKES wave bifurcates to another periodic orbit of the separable form

$$A(x,t) = F(x)e^{-i\Omega t}. \tag{19}$$

Here Ω is a real number, and $F\colon \mathbb{R} \to \mathbb{C}$ is a 1-periodic function satisfying

$$(1+i\nu)F'' + (R + i\Omega - (1+i\mu)|F|^2)F = 0, \quad -\infty < x < \infty, \tag{20}$$

by (1, 2) and (19). Our numerical simulations show that the periodic orbit (19) is stable for all R from an interval (R_I, R_{II}), where we have observed the value

$$R_{II} = 162.4. \tag{21}$$

In particular, we have $F \not\equiv \text{const}$. As R crosses R_{II}, the periodic orbit (19) loses its stability and bifurcates to a stable 2-torus. An analytic approximation of the value R_{II} is presented in NEWTON AND SIROVICH [19]. More precisely, we have obtained a local attractor $\mathcal{A}$ of real dimension 2 consisting of all functions $f\colon \mathbb{R} \to \mathbb{C}$ such that

$$f(x) = F(x - x_0)e^{i\Omega t_0} \quad \text{for some} \quad x_0, t_0 \in \mathbb{R}. \tag{22}$$

Hence, $\mathcal{A}$ is a 2-torus. Furthermore, for all

$$R \in (139.7, R_{II}) = (139.7, 162.4)$$

(see §2.1 and (21) above), our simulations show that the solutions $A(x,t)$ to Eqs. (1, 2) starting from an initial distribution $A(x,0)$ near the rotating wave (3) for $n = 1$ are attracted to $\mathcal{A}$.

3.2 2-tori with periodic modulus

As R increases past the critical value R_{II}, the separable periodic orbit (19) bifurcates *supercritically* to a 2-torus. Our simulations show that also *exchange of stability* takes place. The corresponding stability analysis for this bifurcation has been carried out in Sirovich and Newton [20]. The 2-torus has the form

$$A(x,t) = (F(x) + G(x,t))e^{-i\Omega t} \equiv \Phi(x,t)e^{-i\Omega t}, \tag{23}$$

where the perturbation function $G\colon \mathbb{R} \times \mathbb{R} \to \mathbb{C}$ takes small values for $R - R_{II} > 0$ small, and it is 1-periodic in x and τ-periodic in t, for some $\tau > 0$. The half-period $\tau/2$ coincides with the time-period of the zeroth Fourier mode of the square modulus $|A(x,t)|^2 = |\Phi(x,t)|^2$. As this mode is always real and nonnegative, τ can easily be determined from the graph of this mode versus the time t. For all

$$R \in (R_{II}, R_{II} + 0.1) = (162.4, 162.5),$$

we have observed approximately the same value of τ,

$$\tau_{II} = 1.23. \tag{24}$$

The 2-torus (23) then remains stable for all values $R \in (R_{II}, 180.0]$. More precisely, we have obtained a local attractor $\mathcal{A}$ of real dimension 3 consisting of all functions $f\colon \mathbb{R} \to \mathbb{C}$ such that

$$f(x) = \Phi(x - x_0, -t_1)e^{i\Omega t_0} \quad \text{for some} \quad x_0, t_0, t_1 \in \mathbb{R}. \tag{25}$$

Hence, $\mathcal{A}$ is a 3-torus. Furthermore, for all

$$R \in (R_{II}, 180.0] = (162.4, 180.0],$$

our simulations show that the solutions $A(x,t)$ to Eqs. (1, 2) starting from an initial distribution $A(x,0)$ near the rotating wave (3) for $n = 1$ are attracted to $\mathcal{A}$.

It is remarkable that the angular frequency Ω in both cases (19) and (23) *stays near* the angular frequency of the Stokes wave

$$\omega_0 = \mu R + (\nu - \mu)k_0^2 = -R$$

with a relative error of less than 1.5%, for all $R \in (R_I, 180.0] = (128.3, 180.0]$.

3.3 A 2-torus with periodic modulus for $R = 180.0$

Now we will closer examine a transparent case of a 2-torus (23) for $R = 180.0$. We write

$$A(x,t) = r(x,t)e^{i[\theta(x,t)-\omega_0 t]}, \tag{26}$$

where $r(x,t) = |A(x,t)|$, $\theta(x,t) \in \mathbb{R}$, and $\omega_0 = -180.0$ is the angular frequency of the Stokes wave. In our figures below we limit the values of the phase angle $\theta(x,t)$ to the interval $(-\pi, \pi]$, thus admitting jumps of size $\pm 2\pi$ in the values of $\theta(x,t)$. Given a real function $f(x,t)$, we make its 3-dimensional plot as follows. The origin of our coordinate system is in the middle of the left side of our window, the axis x (axis t, respectively) runs from the origin towards the right lower (right upper) corner, and the values of f are plotted on the vertical axis.

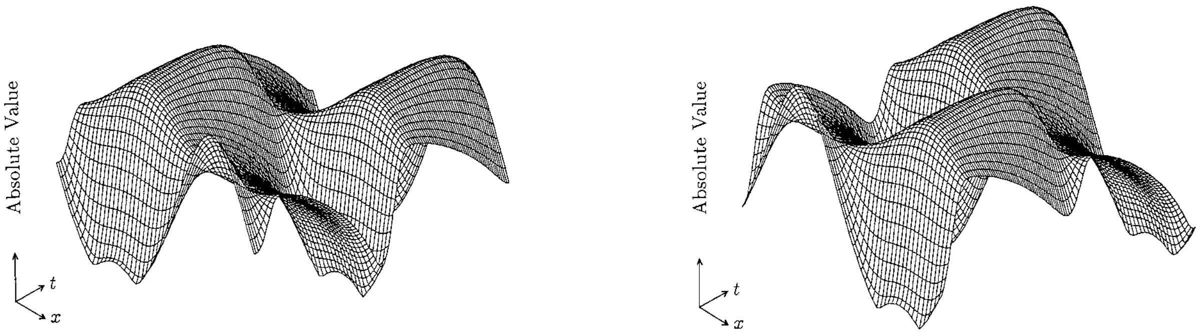

Figure 4: For $R = 180.0$, the modulus $|A(x,t)|$ for (a) $0 \le t \le 0.116$; (b) $0.058 \le t \le 0.174$.

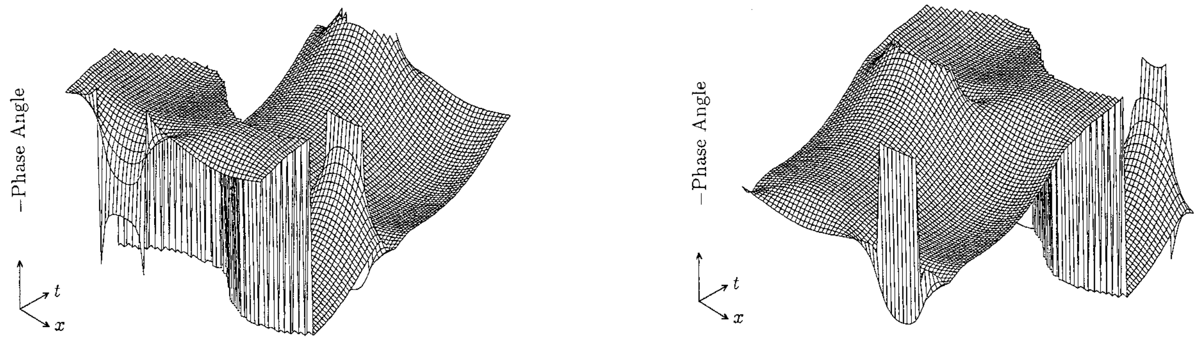

Figure 5: For $R = 180.0$, the negative phase angle $-\theta(x,t)$ for (a) $0 \le t \le 0.116$; (b) $0.058 \le t \le 0.174$.

By (23), the modulus $r(x,t)$ is 1-periodic in x and τ-periodic in t. The graphs of $r(x,t)$ for $(x,t) \in [0,1] \times [0,\tau]$ and $(x,t) \in [0,1] \times [\tau/2, 3\tau/2]$, respectively, are plotted in Fig. 4(a,b). From magnifications of these plots we have obtained the period $\tau = 0.116$. Analogously, the graphs of the negative phase angle $-\theta(x,t)$ are plotted in Fig. 5(a,b). Again, from magnifications of these plots we have deduced that the relative error

$$\frac{|\Omega - \omega_0|}{|\omega_0|} = \frac{|\theta(x,t+\tau) - \theta(x,t)|}{|\omega_0|\tau}$$

is less than 1.5%. We will derive this estimate also from the following study of the complex FOURIER modes $c_n(t)$ of the function $A(x,t)$ of $x \in [0,1]$.

It follows from (23) that $c_n(t)e^{i\Omega t}$ is a τ-periodic function of t. In Fig. 6(a,b) we plot the trajectories of the functions $c_n(t)e^{i\omega_0 t}$ $(0 \leq t < 25\tau)$ in the complex plane for each $n = 0$ and $n = 1$, respectively. The phase angle of the function

$$c_n(t)e^{i\Omega t}/c_n(t)e^{i\omega_0 t} = e^{i(\Omega - \omega_0)t}$$

is equal to $(\Omega - \omega_0)t$. By (an animation leading to) Fig. 6(b), it takes approximately 21 time-periods of length τ for this angle to increase or decrease from 0 to $\pm 2\pi$. Thus, the relative difference between Ω and ω_0 can be estimated by

$$\frac{|\Omega - \omega_0|}{|\omega_0|} \leq \frac{2\pi}{21\tau|\omega_0|} = \frac{2\pi}{21 \cdot 0.116 \cdot 180.0} < 0.015.$$

Consequently, the period $2\pi/|\Omega| \doteq 2\pi/180 \doteq 0.035$ corresponding to Ω is more than 3-times smaller than $\tau = 0.116$.

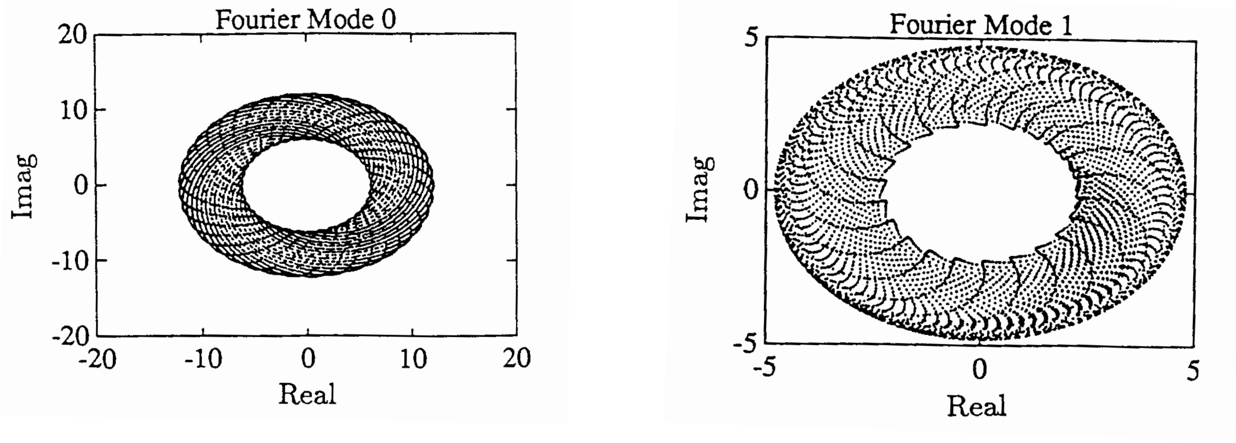

Figure 6: For $R = 180.0$, the complex plots of $c_n(t)e^{i\omega_0 t}$ $(0 \leq t < 2.9)$ for (a) $n = 0$; (b) $n = 1$.

To summarize the results of this section, we have found a *stable* 2-torus (23) for $162.4 < R \leq 180.0$ which bifurcates *supercritically* from a stable periodic orbit (19) at $R_{II} = 162.4$.

4 A Few Complicated Local Attractors

We have observed that the 2-torus (23) collapses as R is increased from 180.0 to 181.0. We have not detected any signs of a bifurcation from this stable 2-torus to a stable 3-torus for $R \in (180.0, 181.0)$. What we have detected for a number of values of $R \in [181.0, 185.0]$ is an interesting *local attractor* which we analyze below.

4.1 A local attractor for $R = 185.0$

We start our simulations from an initial distribution $A(x, t_0)$ near the rotating wave (3), where $n = 1$, $k_1 = 2\pi$,

$$|a_1| = (185.0 - (2\pi)^2)^{1/2} \doteq 12.0632, \tag{27}$$

and $\omega_1 = -185.0 + 6 \cdot (2\pi)^2 \doteq 51.8701$. For a later comparison, the period τ_1 corresponding to the angular frequency ω_1 has the value

$$\tau_1 = 2\pi/\omega_1 \doteq 0.1211. \tag{28}$$

After a transient time interval of 2 units we set our time variable t to $t = 0$ (i.e. $t_0 = -2$) and begin our observations. Our simulations show that the modulus $|A(x, t)|$ remains both spatially and temporally constant for a relatively long time,

$$|A(x, t)| = 12.063 \quad \text{for all} \quad x \in [0, 1] \text{ and } t \in [0, 0.37].$$

Examining also the phase angle $\theta(x, t)$ from (26), we conclude that $A(x, t)$ is very close to the rotating wave (3) with $n = -1$, for all $t \in [0, 0.37]$, see (27).

As the time t increases past 0.37, the instability of the rotating wave causes both spatial and temporal deviations of the modulus $|A(x, t)|$ from the value 12.063. By Fig. 7(a,b), these deviations have the form of an increasing travelling wave with a sine-like shape moving from the right to the left with the velocity $c_{-1} = -61$ (approximately). Although the present parameter $R = 185.0$ is far beyond the bifurcation value $R_{sup} = 137.90$ from (12), formula (16) from our instability analysis in Section 2 yields a comparable value of the velocity (for $m = \pm 1$, $n = -1$)

$$\hat{c}_{-1} = 2 \cdot (-2\pi) \cdot 6 \cdot \frac{185.0 - (2\pi)^2}{185.0} \doteq -59.31.$$

The deviations continue to increase up to $t = 0.744$ when a phase slip occurs at $x = 0.40$ where $A(x, t) = 0$, see Fig. 8(a,b). After this moment, $A(x, t)$ starts approaching a 2-torus apparently having the form (23). This 2-torus can clearly be seen for $0.80 \le t \le 0.95$ in

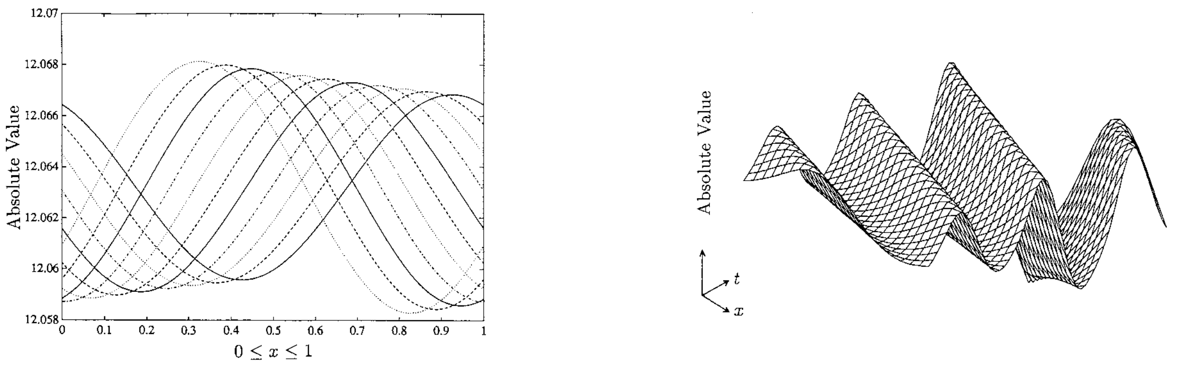

Figure 7: For $R = 185.0$, the deviations $|A(x,t)| - 12.063$ for (a) $0.49 \leq t \leq 0.50$; (b) $0.70 \leq t \leq 0.75$.

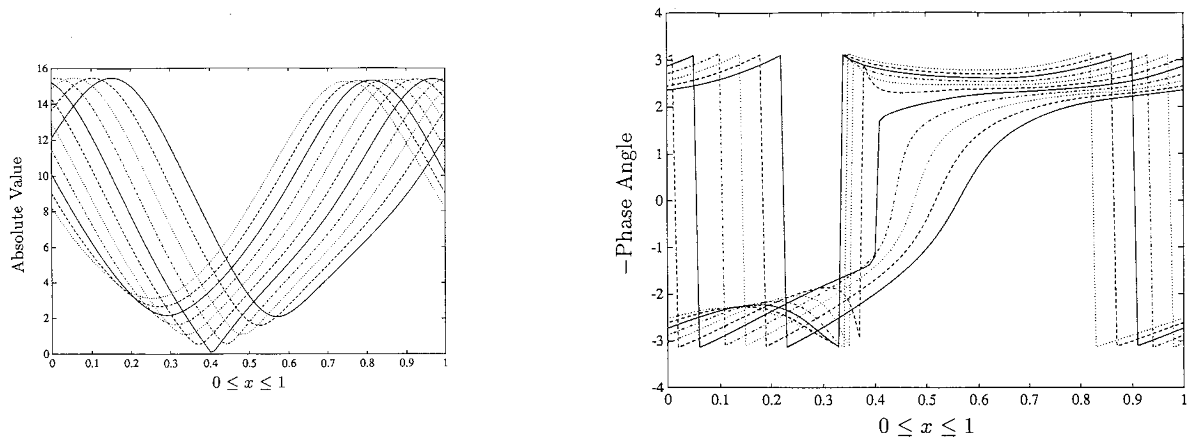

Figure 8: The phase slip for $R = 185.0$, (a) modulus; (b) negative phase angle.

Fig. 9(a,b). It is very similar to the 2-torus in Fig. 4(a,b), but its period $\tau = 0.06$ is about twice smaller than in the latter case ($\tau = 0.116$).

As the time t increases past 0.95, the instability of the 2-torus becomes visible. For $0.95 \leq t \leq 1.06$ we observe a connection orbit from this 2-torus towards a periodic orbit having the form (19), see Fig. 10(a,b) for the modulus $|A(x,t)|$. This modulus remains temporally constant for all $t \in [1.06, 1.29]$ (Fig. 11(a)). It has a shape close to that of the function $|\sin k_1(x - x_0)|$, where $k_1 = 2\pi$ and $x_0 = 0.31$. Inspecting also the negative phase angle $-\theta(x,t)$ defined in (26) (Fig. 11(b)), we conclude that $A(x,t)$ has a form close to

$$A_1(x,t) = \varepsilon e^{-i\Omega t} \sin k_1(x - x_0), \tag{29}$$

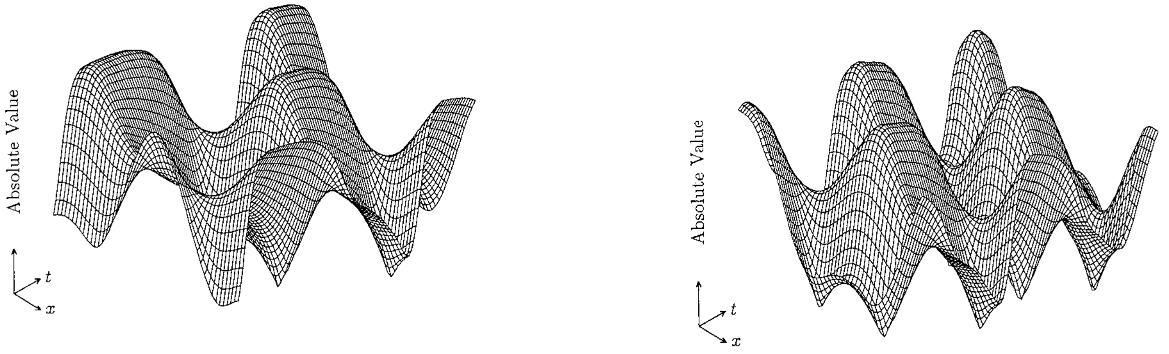

Figure 9: For $R = 185.0$, the modulus $|A(x,t)|$ for (a) $0.80 \le t \le 0.90$; (b) $0.90 \le t \le 1.00$.

where $|\varepsilon| = 14.0$ and

$$\Omega = \omega_0 - \frac{\theta(x,t+T) - \theta(x,t)}{T} = -185.0 + \frac{2\pi}{0.025} \doteq 66.33.$$

The corresponding period is $2\pi/\Omega \doteq 0.095$. It is worth a mention that this solution $A(x,t)$ very well agrees with the following analytic solution to Eqs. (1, 2) obtained in TAKÁČ [22] by rigorous bifurcation analysis,

$$\begin{aligned}&\frac{1}{\varepsilon} e^{i\Omega t} A(x,t) = \\ &\sin k_n x - \frac{\varrho|\varepsilon|^2}{32k_n^2} \sin 3k_n x + \left(\frac{\varrho|\varepsilon|^2}{32k_n^2}\right)^2 (3\sin 3k_n x + \sin 5k_n x) + \mathcal{O}(|\varepsilon|^6),\end{aligned} \tag{30}$$

where ε is a complex bifurcation parameter whose square modulus $|\varepsilon|^2 > 0$ measures the smallness of $R - k_n^2 > 0$, $k_n = 2n\pi \neq 0$, $\varrho \equiv \frac{1+i\mu}{1+i\nu}$, and $R + i\Omega$ is related to $|\varepsilon|^2$ by

$$\frac{R + i\Omega - (1+i\nu)k_n^2}{1+i\mu} = \frac{3}{4}|\varepsilon|^2 \left(1 - \frac{\varrho}{32k_n^2}|\varepsilon|^2 + \mathcal{O}(|\varepsilon|^4)\right). \tag{31}$$

Together with the rotating wave (3), the solution (30) bifurcates from the zero solution as R increases past k_n^2.

As the time t increases past 1.29, also this separable periodic orbit shows its instability. For $1.29 \le t \le 1.43$ we observe a connection orbit from this periodic orbit towards a rotating wave (3) with $n = -1$, see Fig. 12(a,b). This rotating wave is identical (up to multiplication by a complex unit reflecting a phase shift) with the one we have started from at $t = 0$. Again, our simulations show

$$|A(x,t)| = 12.063 \quad \text{for all} \quad x \in [0,1] \ \text{ and } \ t \in [1.43, 1.91].$$

From this point on the entire scenario repeats as we have just described it. This can also be seen from the complex plots of the zeroth and first FOURIER modes of the function $|A(x,t)|^2$

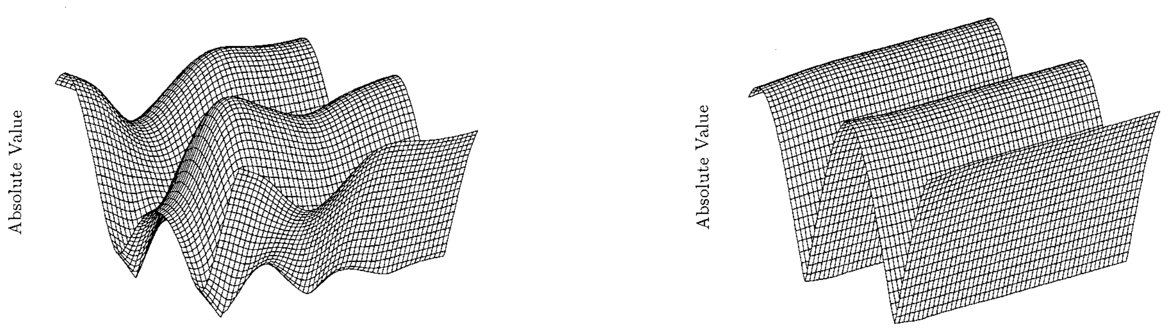

Figure 10: For $R = 185.0$, the modulus $|A(x,t)|$ for (a) $1.00 \leq t \leq 1.05$; (b) $1.05 \leq t \leq 1.10$.

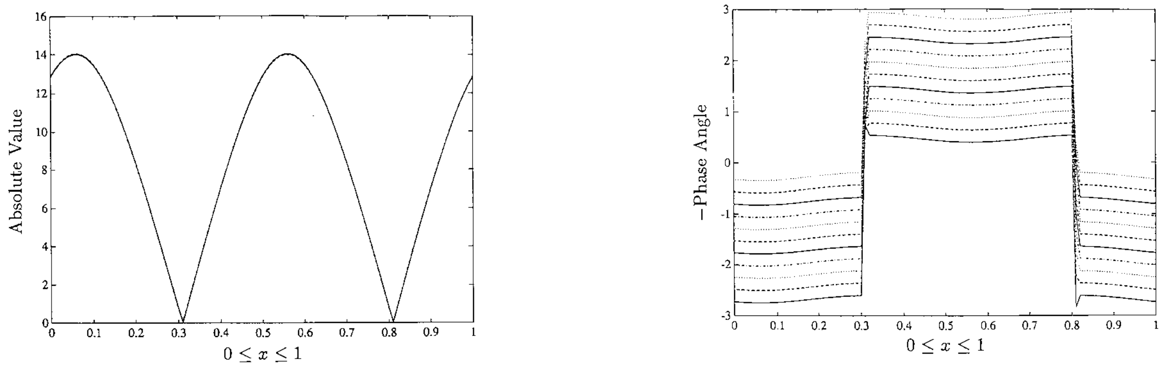

Figure 11: For $R = 185.0$, (a) modulus for $1.06 \leq t \leq 1.29$; (b) negative phase angle for $1.09 \leq t \leq 1.10$.

of $x \in [0,1]$, see Fig. 13(a,b). It takes a time interval of (approximate) length 1.55 units for this scenario to repeat.

It is somewhat amazing that the scenario repeats after approximately the same time interval. This fact strongly suggests that our numerical simulations have produced a homoclinic periodic orbit (of period $T = 1.55$) which *shadows* three heteroclinic connection orbits, the first one running from the rotating wave (3) ($n = -1$) towards the 2-torus (23), the second one running from this 2-torus towards the periodic orbit (30), and the third one returning from this periodic orbit towards the rotating wave (3) ($n = -1$). Furthermore, even if we have started our simulations from $A(x, t_0)$ near the rotating wave (3) with $n = 1$, the simulated shadowing periodic orbit has never returned to this rotating wave again. We are unable to explain why our approximating dynamical system "prefers" the rotating wave (3) with $n = -1$ to the one with $n = 1$. Notice that the approximating dynamical system uses the same number of positive and negative complex FOURIER modes $e^{ik_n x}$, $|n| \leq N$, for the

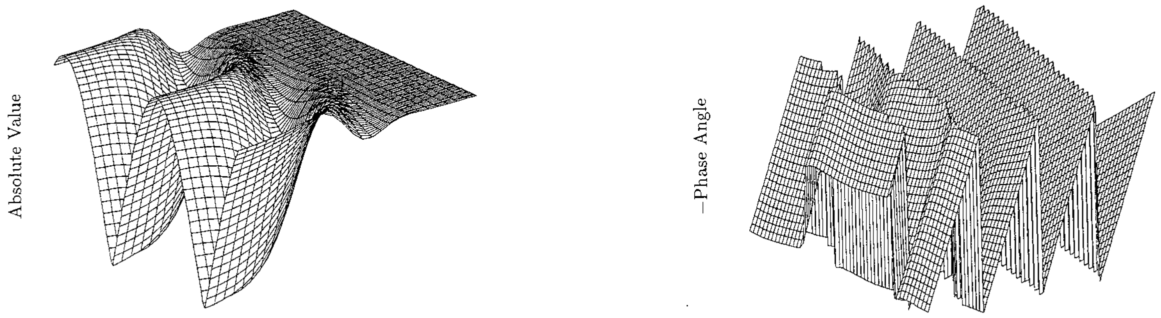

Figure 12: A connection orbit for $R = 185.0$ and $1.30 \leq t \leq 1.40$, (a) modulus; (b) negative phase angle.

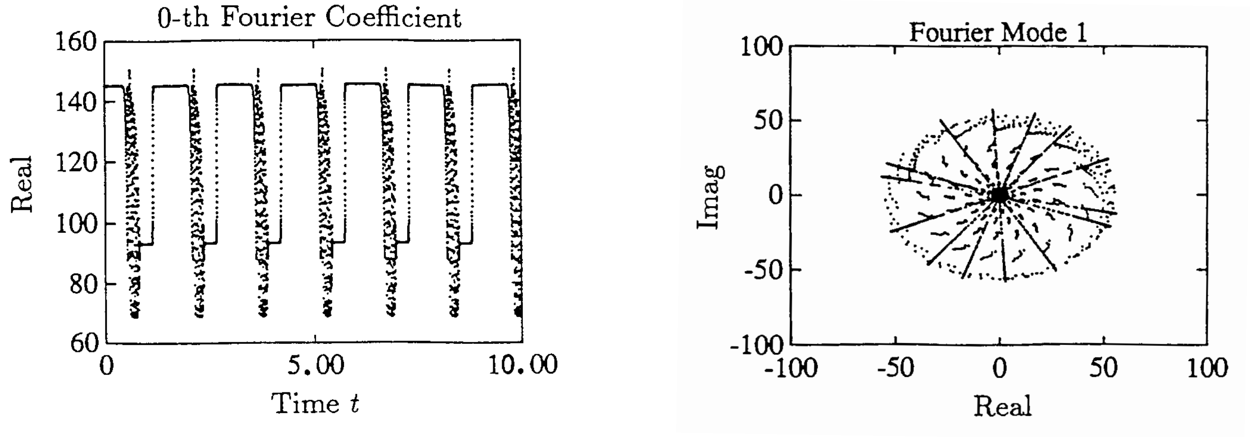

Figure 13: For $R = 185.0$ and $0 \leq t \leq 10.0$, (a) zeroth Fourier mode (real, on vertical axis) versus time (on horizontal axis); (b) first Fourier mode (complex plot) of the square modulus $|A(x,t)|^2$.

spatial approximation, i.e. $A(x,t)$ is approximated by the truncated FOURIER series

$$A_N(x,t) = \sum_{|n| \leq N} c_n(t) e^{ik_n x}. \tag{32}$$

In particular, the corresponding truncation of Eq. (1) yields a system of $2N+1$ ordinary differential equations (ODE's, for short) for $c_n(t)$, $t \geq 0$, cf. DOELMAN [3]. Thus, the only possible reason for $n = -1$ being preferred to $n = 1$ we can think of are some round-off computer errors, despite of double precision complex arithmetics used in our computations. In any event, the *local attractor* $\mathcal{A}$ suggested by our simulations must be symmetric with respect to the reflection $x \mapsto -x$ (i.e. $n \mapsto -n$) because it contains the periodic orbit (30) having this reflection symmetry, and this periodic orbit is connected to all, the 2-torus (23) and the rotating waves (3) with $n = \pm 1$.

More precisely, given the symmetries of the CGL equation, we have obtained a local attractor $\mathcal{A}$ consisting of

(i) two 1D manifolds (circles) $\mathcal{A}'_{\infty}$ and $\mathcal{A}'_{-\infty}$ of rotating waves (3) for $n = \pm 1$, respectively;

(ii) a 3D manifold (a 3-torus) $\mathcal{A}''$ of 2-tori (23);

(iii) a 2D manifold (a 2-torus) $\mathcal{A}'''$ of separable periodic orbits (30); and

(iv) manifolds of connection orbits, $C'_{\pm 1}$, C'' and $C'''_{\pm 1}$. Here, $C'_{\pm 1}$ connects from $\mathcal{A}'_{\pm\infty}$ to $\mathcal{A}''$ and has real dimension 3, C'' connects from A'' to A''' and has 3D, and $C'''_{\pm 1}$ connects from $\mathcal{A}'''$ to $\mathcal{A}'_{\pm\infty}$ and has 2D.

4.2 A local attractor for $R = 250.0$

This case is a simplification of the previous one from §4.1. Again, our simulations start from an initial distribution $A(x, t_0)$ near the rotating wave (3), where $n = 1$, $k_1 = 2\pi$,

$$|a_1| = (250.0 - (2\pi)^2)^{1/2} \doteq 14.51, \tag{33}$$

and $\omega_1 = -250.0 + 6 \cdot (2\pi)^2 \doteq -13.13$. After a transient time interval of about 2 units we set t to $t = 0$ and begin our observations. For $0 \le t \le 0.005$, the modulus $|A(x,t)|$ stays within the interval $(14.4, 14.6)$ for all $x \in [0, 1]$, see Fig. 14(a). Examining also the phase angle $\theta(x,t)$ from (26) in Fig. 14(b), we conclude that $A(x,t)$ is very close (with relative precision $< 1\%$) to the rotating wave (3) with $n = -1$, for all $t \in [0, 0.005]$, see (33). However, Fig. 14(a,b) suggest that an unstable 2-torus may be present. This 2-torus may be the reason why the computed orbit does not get closer to the rotating wave.

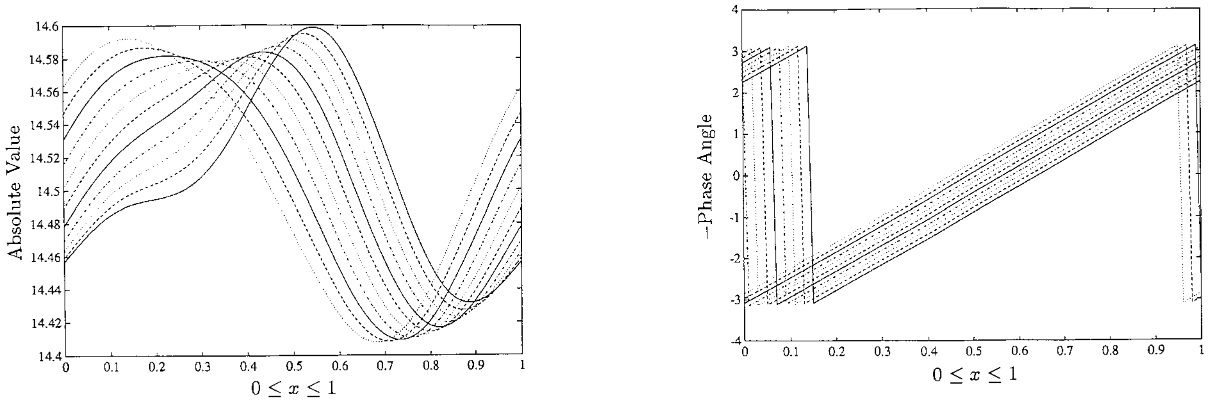

Figure 14: For $R = 250.0$ and $0 \le t \le 0.005$, (a) modulus; (b) negative phase angle.

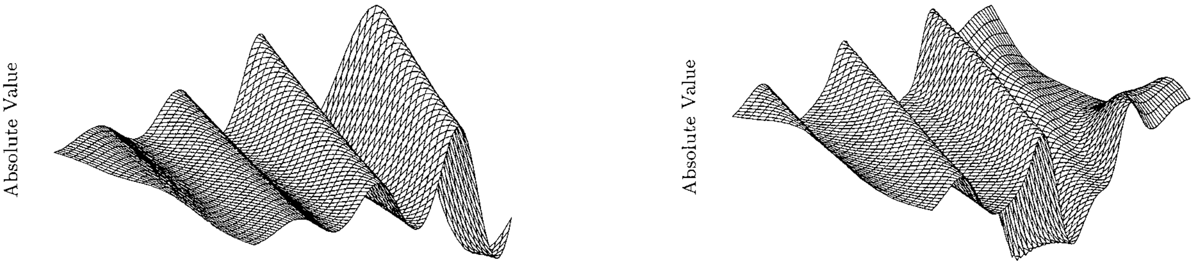

Figure 15: For $R = 250.0$, the deviations $|A(x,t)| - 14.5$ for (a) $0 \le t \le 0.05$; (b) $0.05 \le t \le 0.10$.

As the time t increases past 0.005, we observe both spatial and temporal deviations of $|A(x,t)| - 14.5$ from 0 forming an increasing travelling wave with a sine-like shape moving from the right of the left with the velocity $c_1 = -50$ (approximately), see Fig. 15(a,b). The deviations continue to increase up to $t = 0.085$ when a phase slip occurs at $x = 0.51$ where $A(x,t) = 0$. In this case we observe no sign of $A(x,t)$ approaching a 2-torus (23). Rather, after the phase slip occurs, $A(x,t)$ starts approaching a periodic orbit having the form (30), see Fig. 16(a,b). The modulus $|A(x,t)|$ remains temporally constant for all $t \in [0.23, 0.25]$. Moreover, in (29) we have $|\varepsilon| = 17$ and

$$\Omega = \omega_0 - \frac{\theta(x,t+T) - \theta(x,t)}{T} = -250.0 + \frac{2\pi}{0.025} \doteq 1.33.$$

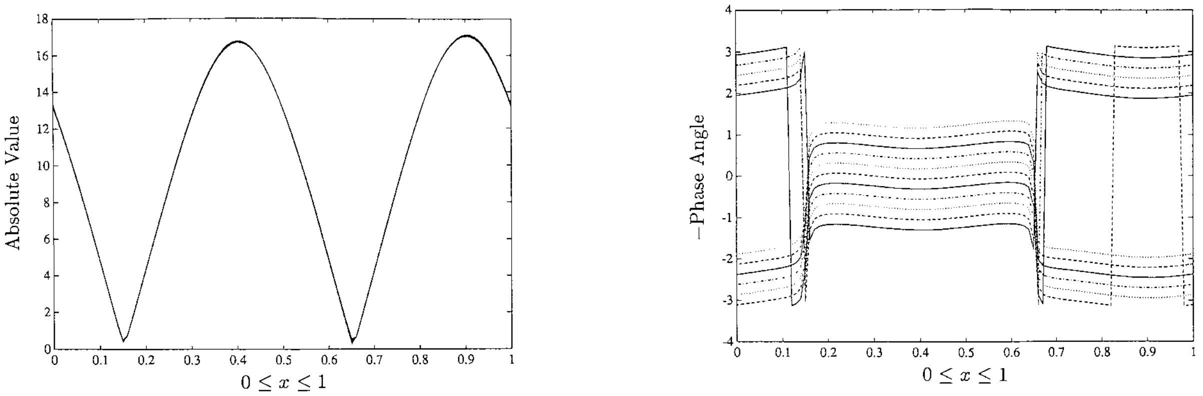

Figure 16: For $R = 250.0$, (a) modulus for $0.23 \le t \le 0.25$; (b) negative phase angle for $0.24 \le t \le 0.25$.

As the time t increases past 0.25, also this separable periodic orbit shows its instability. For $0.25 \le t \le 0.315$ we observe a connection orbit from this periodic orbit towards a rotating wave (3) with $n = -1$, see Fig. 17(a,b). This rotating wave is identical (up to multiplication

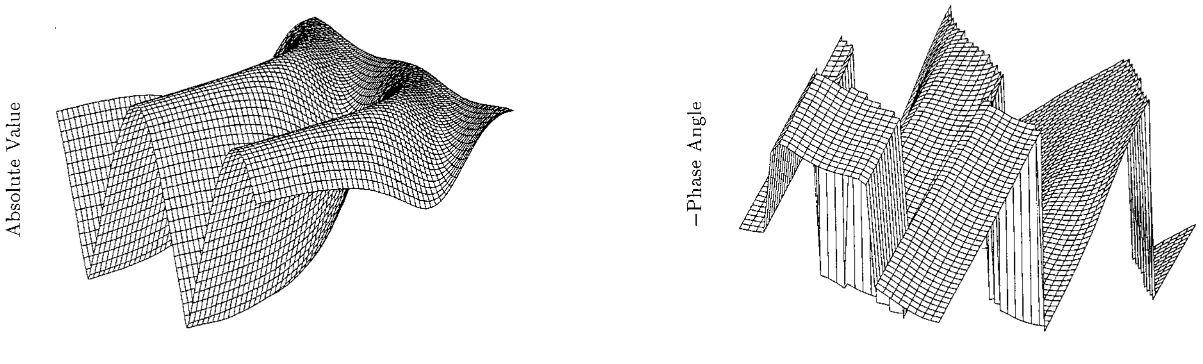

Figure 17: A connection orbit for $R = 250.0$ and $0.25 \le t \le 0.30$, (a) modulus; (b) negative phase angle.

by a complex unit reflecting a phase shift) with the one we have started from at $t = 0$. Again, our simulations show

$$\big||A(x,t)| - 14.5\big| < 0.1 \quad \text{for all} \quad x \in [0,1] \;\text{ and }\; t \in [0.315, 0.320].$$

From this point on the entire scenario repeats as we have just described it. This can also be seen from the complex plots of the zeroth and first FOURIER modes of the function $|A(x,t)|^2$ of $x \in [0,1]$. It takes a time interval of (approximate) length 0.316 units for this scenario to repeat.

Given the symmetries of the CGL equation, we have obtained a local attractor $\mathcal{A}$ consisting of

(i) two 1D manifolds (circles) $\mathcal{A}'_\infty$ and $\mathcal{A}'_{-\infty}$ of rotating waves (3) for $n = \pm 1$, respectively;

(ii) a 2D manifold (a 2-torus) $\mathcal{A}''$ of separable periodic orbits (30); and

(iii) four 2D manifolds of connection orbits, $C'_{\pm 1}$ and $C''_{\pm 1}$. Here, $C'_{\pm 1}$ connects from $\mathcal{A}'_{\pm\infty}$ to $\mathcal{A}''$, and $C''_{\pm 1}$ connects from $\mathcal{A}''$ to $\mathcal{A}'_{\pm\infty}$.

4.3 Separable periodic orbits for $R = 300.0$ and $R = 350.0$

Both these cases are analogous to the case $R \in (R_I, R_{II}) = (128.3049, 162.4)$ from §3.1. The separable periodic orbits (19) have the angular frequencies $\Omega = -161.30$ for $R = 300.0$ and $\Omega = -257.87$ for $R = 350.0$. The shape of their modulus $|A(x)|$ and the phase angle $|\theta(x)|$ from (26) suggest that these periodic orbits should lie on the same bifurcation branch of separable periodic orbits as those in §3.1.

To summarize the results of this section, for each $R = 185.0$ and 250.0, we have found an interesting local attractor $\mathcal{A}$. For $R = 185.0$, $\mathcal{A}$ consists of two 1D manifolds (circles) of rotating waves (3) for $n = \pm 1$, a 3D manifold (a 3-torus) of 2-tori (23), a 2D manifold (a 2-torus) of separable periodic orbits (30), and manifolds of connection orbits. For $R = 250.0$, $\mathcal{A}$ consists of two 1D manifolds (circles) of rotating waves (3) for $n = \pm 1$, a 2D manifold (a 2-torus) of separable periodic orbits (30), and four 2D manifolds of connection orbits.

5 Discussion

For the supercritical bifurcation at $R = R_{sup} = 137.90$ we have obtained a stable 2-torus for $137.9 < R < 139.7$. For the subcritical bifurcation at $R = R_{sub}$ our results are qualitatively similar to the experimental and numerical results obtained by Janiaud et al. [8, 9] who observed a 2-torus given by Eq. (5) persisting for long time before it collapsed. We have observed its instability outside the tiny interval $84.97 < R < 84.99$; the stability for $R = 84.98$ may be due to numerical errors, cf. (12).

Finally, the numerical simulations studying the stability of the 2-tori having the forms (5) and (23) suggest that these 2-tori may bifurcate into (possibly unstable) 3-tori having the following form,

$$A(x,t) = B(x - ct, t/\tau)e^{-i\Omega t}, \quad -\infty < x < \infty,\ t \geq 0, \tag{34}$$

where $c \neq 0$, $\Omega \neq 0$, $\tau > 0$, and $B(x,\xi)$ is 1-periodic in both $x, \xi \in \mathbb{R}$. Namely, given the symmetries of the CGL equation, the 2-tori (23) form a 3D manifold which itself is a 3-torus (25) invariant under the semiflow generated by Eqs. (1, 2). Hence, it is very reasonable to expect the possibility of a bifurcation that preserves this 3-torus, but the 2-tori (23) bifurcate into a *triperiodic motion* (34). In particular, inserting (34) into (1) we obtain the following boundary value problem for the unknown $(1,1)$-periodic complex-valued function $B(x,\xi)$,

$$\tau^{-1}\partial_\xi B - (1 + i\nu)\partial_{xx}^2 B - c\partial_x B - (R + i\Omega)B = -(1 + i\mu)|B|^2 B, \quad x, \xi \in \mathbb{R}. \tag{35}$$

Here, suitable values for the real constants $\mu \neq 0$, $\nu \neq 0$, $\tau > 0$, $c \neq 0$, $R > 0$, and $\Omega \neq 0$ are to be found so that Eq. (35) admits a $(1,1)$-periodic solution $B(x,t)$ such that its modulus $|B(x,t)|$ is nonconstant in both x and t. A different 3-torus motion was obtained in Moon et al. [14] by numerical simulations.

References

[1] **Blennerhassett, P. J.** : *On the generation of waves by wind.* Philos. Trans. Roy. Soc. London, Ser. A, **298**, 451–494 (1980)

[2] **Doelman, A.** : *Slow time-periodic solutions to the Ginzburg-Landau equation.* Physica D **40**(2), 156–172 (1989)

[3] **Doelman, A.** : *Finite-dimensional models of the Ginzburg-Landau equation.* Nonlinearity **4**(2), 231–250 (1991)

[4] **Doering, C. A.**, **Gibbon, J. D.**, **Holm, D. D.** and **Nicolaenko, B.** : *Low-dimensional behaviour in the complex Ginzburg-Landau equation.* Nonlinearity **1**(2), 279–309 (1988)

[5] **Eckhaus, W.** : *Studies in Nonlinear Stability Theory.* Berlin 1965

[6] **Ginzburg, V. L.** and **Landau, L. D.** : *On the theory of superconductivity.* Zh. Eksp. Teor. Fiz. (USSR) **20**, 1064 (1950) English transl. in: Ter Haar, D. (ed.): *Men of Physics: L. D. Landau.* Vol. I, pp. 546–568. New York 1965

[7] **Guckenheimer, J.** and **Holmes, P.** : *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields.* New York 1983

[8] **Janiaud, B.**, **Guyon, E.**, **Bensimon, D.** and **Croquette, V.** : *Sideband instability waves with periodic boundary conditions.* In: Busse, F. H. and Kramer, L. (eds.): *Nonlinear Evolution of Spatio-Temporal Structures in Dissipative Continuous Systems.* NATO ASI Series B; pp. 45–50. New York 1990

[9] **Janiaud, B.**, **Pumir, A.**, **Bensimon, D.**, **Croquette, V.**, **Richter, H.** and **Kramer, L.** : *The Eckhaus instability for travelling waves.* Physica D **55**(3&4), 269–286 (1992)

[10] **Keefe, L. R.** : *Dynamics of perturbed wavetrain solutions to the Ginzburg-Landau equation.* Stud. Appl. Math. **73**, 91–153 (1985)

[11] **Kuramoto, Y.** : *Phase dynamics of weakly unstable periodic structures.* Prog. Theor. Phys. **71**(6), 1182–1196 (1984)

[12] **Kuramoto, Y.** and **Tsuzuki, T.** : *On the formation of dissipative structures in reaction-diffusion systems.* Prog. Theor. Phys. **54**(3), 687–699 (1975)

[13] **Landman, M. J.** : *Solutions of the Ginzburg-Landau equation of interest in shear flow transition.* Stud. Appl. Math. **76**, 187–237 (1987)

[14] **Moon, H. T.**, **Huerre, P.** and **Redekopp, L. G.** : *Transitions to chaos in the Ginzburg-Landau equation.* Physica D **7**(1-3), 135–150 (1983)

[15] **Newell, A. C.** : *Envelope equations.* Lectures in Appl. Math. **15**, 157–163 (1974)

[16] **Newell, A. C.** : *The dynamics of patterns: A survey.* In: Wesfreid, J. E., Brand, H. R., Manneville, P., Albinet, G. and Boccara, N.: *Propagation in Systems Far from Equilibrium.* pp. 122–155. Berlin 1988

[17] **Newell, A. C.** and **Whitehead, J. A.** : *Finite bandwidth, finite amplitude convection.* J. Fluid Mech. **38**(2), 279–303 (1969)

[18] **Newton, P. K.** and **Sirovich, L.** : *Instabilities in the Ginzburg-Landau equation: periodic solutions.* Quart. Appl. Math. **44**(1), 49–58 (1986)

[19] **Newton, P. K.** and **Sirovich, L.** : *Instabilities in the Ginzburg-Landau equation II. Secondary bifurcation.* Quart. Appl. Math. **44**(2), 367–374 (1986)

[20] **Sirovich, L.** and **Newton, P. K.** : *Periodic solutions of the Ginzburg-Landau equation.* Physica D **21**(1), 115–125 (1986)

[21] **Stuart, J. T.** and **Di Prima, R. C.** : *The Eckhaus and Benjamin-Feir resonance mechanisms.* Proc. Roy. Soc. London Ser. A **362**, 27–41 (1978)

[22] **Takáč, P.** : *Invariant 2-tori in the time-dependent Ginzburg-Landau equation.* Nonlinearity **5**(2), 289–321 (1992)

received: September 6, 1995

Author:

Prof. Dr. Peter Takáč, Ph.D.
Fachbereich Mathematik
Universität Rostock
Universitätsplatz 1
D–18055 Rostock, Germany
e-mail: peter.takac@mathematik.uni-rostock.de

Hinweise für Verfasser

Um die redaktionelle Bearbeitung und die Herstellung der Druckvorlage zu erleichtern, wären wir den Verfassern dankbar, sich betreffs der Form der Manuskripte an den in **Rostock. Math. Kolloq.** (ab Heft **43**) veröffentlichten Beiträgen zu orientieren.
Insbesondere beachte man:

1. Manuskripte sollten grundsätzlich **maschinengeschrieben** (Schreibmaschine, Drucker) in **deutscher oder englischer Sprache** abgefaßt sein.

2. Zur inhaltlichen Einordnung der Arbeit sind **1-2 Klassifizierungsnummern** (entsprechend der „1980 Mathematics Subject Classification" der Mathematical Reviews) anzugeben.

3. **Textbreite/Texthöhe** des Manuskripts sollten sich an den Maßen **160mm/230mm** orientieren.

4. Der Manuskripttext ist **eineinhalbzeilig, linksbündig**, wenn möglich links-und rechtsbündig zu schreiben. Beim Auftreten von Formeln im laufenden Text ist der Zeilenabstand entsprechend zu vergrößern.

5. Der Platz für **Abbildungen** sollte beim Schreiben ausgespart werden; die Abbildungen selbst sind in der dem ausgesparten Platz entsprechenden Größe gesondert beizufügen.

6. **Literaturzitate** sind im Text durch laufende Nummern (vgl. [3], [4]; [7, 8, 10]) zu kennzeichnen und am Schluß der Arbeit unter der Zwischenüberschrift **Literatur** bzw. **References** zusammenzustellen. Hierbei ist die durch die nachfolgenden Beispiele veranschaulichte Form einzuhalten (die Zeitschriftenabkürzungen erfolgen nach Mathematical Reviews).

   [3] **Zariski, O.**, and **Samuel, P.:** *Commutative Algebra.* Prinston 1958

   [4] **Steinitz, E.:** *Algebraische Theorie der Körper.* J. Reine Angew. Math. **137**, 167-309 (1920)

   [8] **Gnedenko, B.W.:** *Über die Arbeiten von C.F. Gauß zur Wahrscheinlichkeitsrechnung.* In: Reichard, H. (Ed.): C.F. Gauß, Gedenkband anläßlich des 100. Todestages. S. 193-204, Leipzig 1967

   Die Angaben erfolgen in Originalsprache; bei kyrillischen Buchstaben sollte die bibliothekarische Transkription bzw. eine Übersetzung lt. Mathematical Reviews verwendet werden.

7. Die aktuelle, vollständige Adresse des Verfassers sollte enthalten: Titel / Vornamen Name / Institution / Struktureinheit / Straße Hausnummer / Postleitzahl Ort / Land.

Weiterhin besteht die Möglichkeit, mit dem **Satzsystem LaTeX** oder unter anderen Textprogrammen erstellte Manuskripte auf unter **MS-DOS** formatierten Disketten (**3.5″, 0.72MB, 1.44MB**) einzureichen oder per e-mail an eine der folgenden Adressen zu schicken:

`susann.dittmer@mathematik.uni-rostock.de`

`heike.schubert@mathematik.uni-rostock.de`