42 Stoker, J. J., Water Waves, Wiley-Interscience, New York (1957).
43 Bürger, W., A Note on the Breaking of Waves on Non-uniformly Sloping Beaches **16**, pp. 1131—1142 (1967).
44 Eringen, A. C. and Suhubi, E. S., Elastodynamics Vol. 1, Finite Motion, Academic Press, New York (1974).
45 John, F., Formation of Singularities in One-dimensional Nonlinear Hyperbolic Partial Differential Equations, Comm. Pure Appl. Math. **27**, pp. 377—405 (1974).
46 Jeffrey, A., The Exceptional Condition and Unboundedness of Solutions of Hyperbolic Systems of Conservation Type, Proc. Roy. Soc. Edinburgh, Sect. A. 77, pp. 1—8 (1977).
47 Jeffrey, A., Breakdown of the Solution to a Completely Exceptional System of Hyperbolic Equations, J. Math. Anal. and Applics. **45**, pp. 375—381 (1974).
48 Glimm, J., Solution in the Large for Nonlinear Hyperbolic Systems of Equations, Comm. Pure Appl. Math. **18**, pp. 697—715 (1965).
49 Glimm, J., and Lax, P. D., Decay of Solutions of Systems of Nonlinear Hyperbolic Conservation Laws, Am. Math. Soc. Memoir **101** (1970).
50 Lax, P. D., Weak Solutions of Nonlinear Hyperbolic Equations and Their Numerical Computation, Comm. Pure Appl. Math. **7**, pp. 159—193 (1954).
51 Germain, P., Shock Waves, Jump Relations and Structure, Advances in Applied Mechanics [Ed. Chia-Shun Yih], vol. **12**, pp. 131—144. Academic Press, New York (1972).
52 Lax, P. D., Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves, SIAM Regional Conference Series in Applied Mathematics, No. 11 (1973).
53 Oleinik, O., Discontinuous Solutions of Nonlinear Differential Equations, Uspekhi Mat. Nauk. **12**, pp. 3—73 (1957).
54 Friedrichs, K. O. and Lax, P. D., Systems of Conservation Equations with a Convex Extension, Proc. Nat. Acad. Sci. U.S.A. **68**, pp. 1686—1688 (1971).
55 Conley, C. C. and Smoller J. A., Shock Waves as Limits of Progressive Wave Solutions of Higher Order Equations, Comm. Pure Appl. Math. **24**, pp. 459—472 (1971).
56 Gel'fand, I. M., Some Problems in the Theory of Quasilinear Equations, Uspekhi Mat. Nauk, **14**, pp. 87—158 (1959), English translation in Am. Math. Soc. Trans., Ser. 2, No. **29**, pp. 295—381 (1963).
57 Kruzhkov, S. N., Generalised Solutions of the Cauchy Problems in the Large for Nonlinear Equations of First Order, Soviet Math. Dokl. **10**, pp. 785—788 (1969).
58 Suhubi, E. S. and Jeffrey, A., Propagation of Weak Discontinuities in a Layered Hyperelastic Half-Space, Proc. Roy. Soc. Edinburgh, Sect. A. **75**, pp. 209—221 (1976).
59 Jeffrey, A. and Tin, Saw, Waves over Obstacles on a Shallow Seabed, Proc. Roy. Soc. Edinburgh, Sect. A. **71**, pp. 181—192 (1973).
60 Jeffrey, A., The Propagation of Weak Discontinuities in Quasilinear Hyperbolic Systems with Discontinuous Coefficients, Part I — Fundamental Theory, Applicable Anal. **3**, pp. 79—100 (1973); Part II — Special Cases and Application, Applicable Anal. **3**, pp. 359—375 (1974).
61 Donato, A., The Propagation of Weak Discontinuities in Quasilinear Hyperbolic Systems when a Characteristic Shock Occurs, Proc. Roy. Soc. Edinburgh, Sect. A. (in press).
62 Taniuti, T. and Yajima, N., Perturbation Method for a Nonlinear Wave Modulation — I, J. Math. Phys. **10**, pp. 1369—1372 (1969).
63 Asano, N., Taniuti, T. and Yajima, N., Perturbation Method for Nonlinear Wave Modulation — II, J. Math. Phys. **10**, pp. 2020—2024 (1969).
64 Lax, P. D., Almost Periodic Solutions of the KdV Equation, SIAM Review, **18**, pp. 351—375 (1976).
65 Miura, R. M. (Ed.), Bäcklund Transformations, the Inverse Scattering Method, Solutions and their Applications, Lecture Notes in Mathematics, Springer, Berlin (1974).
66 Benjamin, T. B., Bona, J. and Mahony, J. J., Model Equations for Long Waves in Nonlinear Dispersive Systems, Phil. Trans. Roy. Soc. A. **272**, pp. 47—78 (1972).
67 Korteweg, D. J. and de Vries, G., On the Change of Form of Long Waves Advancing in a Rectangular Canal, and on a New Type of Long Stationary Waves, Phil. Mag. (V), **39**, pp. 422—443 (1895).
68 Miura, R. M., Gardner, C. S. and Kruskal, M. D., The Korteweg-de Vries Equation and Generalisation II; Existence of Conservation Laws and Constants of Motion, J. Mathematical Phys. **9**, pp. 1204—1209 (1968).

Address: Prof. Alan Jeffrey, University of Newcastle upon Tyne, Department of Engineering Mathematics, Stephenson Building, Claremont Road, Newcastle upon Tyne, NE1 7Ru, Great Britain

ZAMM 58, T 56 – T 65 (1978)

K. Magnus

Kreiselmechanik

Es wird ein Überblick über Ergebnisse der Kreiseltheorie gegeben, die meist im Zusammenhang mit der Untersuchung technischer Probleme der Raumfahrt, der Kreiselgerätetechnik oder der Rotordynamik erarbeitet worden sind. Einflüsse von Zusatzmassen, Dämpfung und Nachgiebigkeit von Teilsystemen werden behandelt und es werden Ansätze einer Theorie von Kreiselsystemen mit drehzahlgeregelten Rotoren angegeben.

1. Einführendes

Unter dem Begriff Kreiselmechanik faßt man üblicherweise die Probleme der Drehbewegungen starrer Körper zusammen. Klassische Beispiele hierfür sind die bekannten Fälle von Euler und Lagrange (kräftefreier unsymmetrischer bzw. schwerer symmetrischer Kreisel). Ausgehend davon haben in der Folgezeit vorwiegend Mathematiker die Bewegungsgleichungen der Kreisel unter den verschiedensten Nebenbedingungen und zusätzlichen Annahmen zu lösen versucht. Einen viel beachteten Höhepunkt dieser Bemühungen bildet zweifellos die Dissertation der Kowalewskaja, weil dort das Instrumentarium der elliptischen Funktionen virtuos zur Lösung eines Sonderfalles

(spezieller symmetrischer Kreisel mit exzentrischer Schwerpunktslage) eingesetzt und dabei zugleich weiterentwickelt wurde. Wenngleich hier ein brillantes Beispiel in angewandter Mathematik und Mechanik präsentiert wurde, so haben doch Physiker und Ingenieure kaum davon Notiz genommen, da weder ein phänomenologisch besonders interessantes noch ein technisch verwertbares Ergebnis erkennbar war. Dennoch haben die klassischen Arbeiten wesentlich dazu beigetragen, Methoden zur Behandlung auch technisch interessierender Probleme zu entwickeln.

Kreiselmechanische Aufgaben spielen heute bei allen Maschinen und Geräten eine Rolle, die drehende Teile enthalten. Sondergebiete sind außer der hochentwickelten Kreiselgerätetechnik die Rotordynamik und in den letzten Jahrzehnten vor allem die Satellitentechnik. Gerade aktuelle Probleme der Raumfahrt haben zu einem Aufleben des Interesses auch an den von den Klassikern behandelten Fragen geführt. Gleichzeitig erwies es sich aber als notwendig, weit über den Rahmen klassischer Ergebnisse hinauszugehen, um aktuelle Probleme lösen zu können.

Verallgemeinerungen sind in vielfacher Hinsicht vorgenommen worden. So hat man bei der Untersuchung der Drehbewegungen starrer Körper die folgenden Erweiterungen berücksichtigt:

— innere und/oder äußere Zusatzmassen,
— innere und/oder äußere Dämpfungen,
— elastische Teilsysteme, z. B. angebaute Antennen,
— Hohlräume, die ganz oder teilweise mit idealem oder viskosem Fluid gefüllt sind,
— spezielle Momentenfunktionen, wie sie bei selbsterregten Kreiseln oder bei Raumfahrzeugen durch den Schweregradienten hervorgerufen werden,
— Regelvorrichtungen, die entweder zu aktiver Dämpfung oder zu Lageregelungen, meist aber für Optimierungsprobleme eingesetzt werden.

Über einige der dabei erhaltenen Ergebnisse soll berichtet werden. Dabei muß auf das Zitieren aller einschlägigen Veröffentlichungen verzichtet werden, weil es hier vor allem darauf ankommt, einen Überblick über mögliche Verhaltensformen von Kreiseln und Kreiselsystemen zu geben.

Um die Zusammenhänge besser sichtbar werden zu lassen, soll — soweit möglich — eine einheitliche Darstellung gewählt werden. Bei Stabilitätsproblemen hat sich dafür das Formdreieck, eine geometrisch durchsichtige Darstellung für Körper beliebiger Massenverteilungen, bewährt. Seine Konstruktion sei kurz erläutert (Bild 1): Man denke sich die Zahlwerte für die Hauptträgheitsmomente A, B, C eines starren Körpers auf den Achsen 1, 2, 3 eines kartesischen Dreibeins aufgetragen. Sie bilden die Koordinaten eines im ersten Oktanten des Koordinatensystems gelegenen Punktes P, der als Bildpunkt des Körpers angesehen werden kann, weil er in eindeutiger Weise sein Trägheitsellipsoid kennzeichnet. Da jedoch selten die absoluten Größen der Trägheitsmomente, sondern im allgemeinen nur ihre Verhältnisse interessieren, kann die dreidimensionale Mannigfaltigkeit der Bildpunkte auf eine zweidimensionale reduziert werden. Hierzu wird der Durchstoßpunkt der Verbindungslinie von P zum Koordinatenursprung mit der durch $A + B + C = 1$ definierten Ebene bestimmt. Wegen der bekannten Relationen zwischen den Hauptträgheitsmomenten eines starren Körpers ($A + B \geqq C$, ...) liegen dann die Bildpunkte realer Körper stets in dem in Bild 1 schattierten gleichseitigen Dreieck, das als Formdreieck bezeichnet wird.

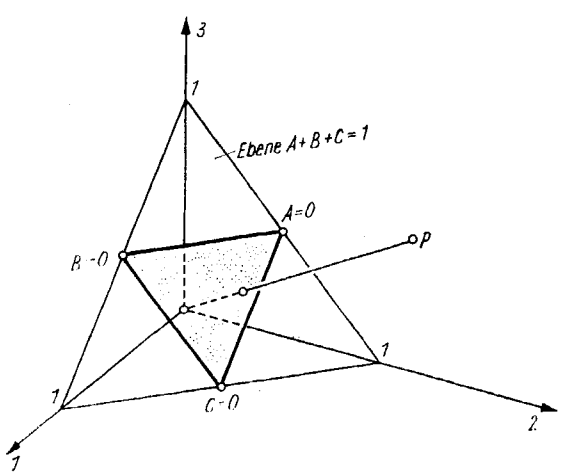


Bild 1. Konstruktion des Formdreiecks

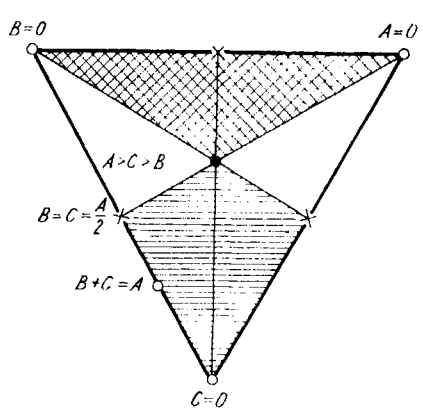


Bild 2. Das Formdreieck

Eckpunkte des Formdreiecks (Bild 2) entsprechen stabförmigen Körpern; sie sind durch $A = 0$ oder $B = 0$ oder $C = 0$ gekennzeichnet. Bildpunkte auf den Dreieckseiten entsprechen scheibenförmigen Körpern ($A + B = C$, $B + C = A$, $C + A = B$). Punkte auf den Mittelsenkrechten stehen für Körper, bei denen jeweils zwei Hauptträgheitsmomente gleich groß sind (symmetrische Kreisel). Der Mittelpunkt wird für Körper mit kugelförmigem Trägheitsellipsoid erhalten ($A = B = C$). Den sechs, durch die Mittelsenkrechten abgeteilten Teildreiecken sind jeweils eindeutige Größenbeziehungen, z. B. wie eingetragen $A > C > B$, zugeordnet. Demnach können die Bereiche der bezüglich jeder der drei Hauptachsen gestreckten oder abgeplatteten Körper unmittelbar erkannt werden: so ist bezüglich der 3-Achse, zu der das Hauptträgheitsmoment C gehört, der untere, einfach schraffierte Bereich der Ort aller gestreckten, der obere, doppelt schraffierte Bereich der Ort aller abgeplatteten Körper; die Achse 3 bildet

dann die „lange" bzw. die „kurze" Achse des Kreisels. Ist dagegen C mittleres Hauptträgheitsmoment, dann liegen die zugehörigen Bildpunkte in den beiden nicht schattierten Teildreiecken.

2. Verallgemeinerungen klassischer Fälle

Der momentenfreie starre Körper mit drei Freiheitsgraden der Drehung (EULER-Kreisel) kann um alle drei Hauptachsen stationäre Drehbewegungen ausführen. Sie sind stabil, wenn sie um die Achsen des größten oder kleinsten Hauptträgheitsmomentes erfolgen; Drehungen um die mittlere Hauptachse sind instabil. Dieses klassische Ergebnis ist in Bild 3a im Formdreieck für den Fall dargestellt, daß der Körper stationär um die 3-Achse dreht. Für Körper, deren Bildpunkte in die beiden schraffierten Dreiecke fallen, ist die Drehung instabil, weil sie um die „mittlere Achse" erfolgt.

Eine erste Verallgemeinerung dieses Ergebnisses zeigt Bild 3b. Hier ist viskose äußere Dämpfung berücksichtigt worden, wie sie etwa bei einem in viskosem Fluid drehenden Körper vorliegt. Die instabilen Bereiche werden mit wachsender Stärke der Dämpfung kleiner. Auf diese Weise können Körper mit nicht zu starker Unsymmetrie bezüglich der 3-Achse auch dann stabil rotieren, wenn die Drehachse mittlere Hauptachse ist. Dieses Ergebnis ist eine Art Abfallprodukt von Untersuchungen, die zur Klärung des Verhaltens schnellaufender Zentrifugen durchgeführt worden sind.

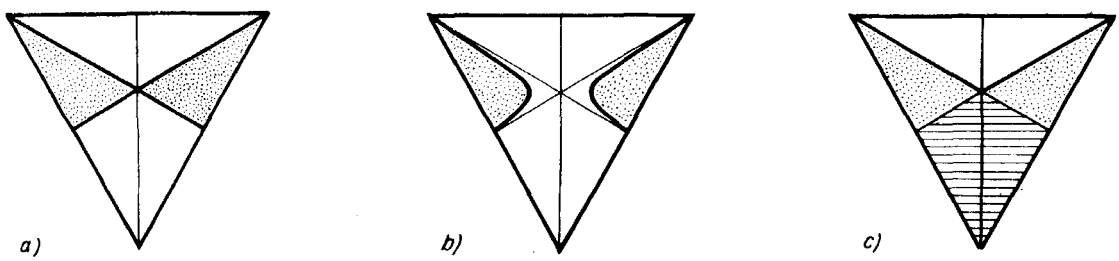


Bild 3. Der EULER-Kreisel; a) ohne ... , b) mit äußerer ... , c) mit innerer Dämpfung. Instabilitätsbereiche sind schraffiert

Ein völlig anderes Ergebnis wird bei Vorhandensein „innerer Dämpfungen" erhalten (Bild 3c). Hier wird der instabile Bereich größer als im ungedämpften Fall: stabile Drehungen sind nur noch um die Achse des größten Hauptträgheitsmomentes, d. h. um die kurze Achse möglich. Dieser Sachverhalt wurde zuerst experimentell durch Beobachtung des Verhaltens der ersten künstlichen Satelliten entdeckt: die nach dem Start um die lange Achse drallstabilisierten Satelliten fingen an zu taumeln und drehten nach einigen Tagen stabil um die kurze Achse. Eine einleuchtende Erklärung für dieses Verhalten wurde zunächst durch die heuristische „energy-sink"-Hypothese gegeben: wegen innerer Energiedissipation kann die kinetische Energie nur abnehmen. Da aber der Drall wegen des Fehlens äußerer Momente konstant bleibt, strebt der Kreisel schließlich dem stationären Bewegungszustand zu, bei dem die Energie den kleinsten möglichen Wert annimmt. Das aber ist die Drehung um die kurze Achse, weil hier die zur Erreichung des konstanten Dralls $\boldsymbol{L}$ erforderliche Drehgeschwindigkeit $\boldsymbol{\omega}$ am kleinsten wird, also auch die kinetische Energie $T = \frac{1}{2}\boldsymbol{L}^T\boldsymbol{\omega}$ den kleinsten Wert annimmt.

Über die Ursache der Energiedissipation durch innere Dämpfung ist viel diskutiert worden. Bei den ersten Satelliten war sie zweifellos in der Tatsache zu sehen, daß sich die dünnen Antennen bei den Taumelbewegungen verformten und damit Werkstoffdämpfung wirksam wurde. Es sind aber vielerlei andere Möglichkeiten denkbar und zum Teil auch durchgerechnet worden. Ihnen allen ist gemeinsam, daß das Grundkonzept der klassischen Kreiseltheorie, die Annahme eines „einzelnen starren Körpers", verlassen werden muß. Es müssen vielmehr stets irgendwelche inneren oder äußeren Zusatzmassen vorhanden sein, die sich relativ zum Hauptkörper bewegen können. Dabei können Fessel- und/oder Dämpfungs-Kräfte zwischen Haupt- und Zusatz-Körper wirken. Einige mögliche Konfigurationen sind in Bild 4 skizziert worden. Der mit körperfester Achse im Kreisel liegende symmetrische Rotor sowie ein vollkommen mit homogenem Fluid gefüllter Hohlraum im Kreisel sind dadurch gekenn-

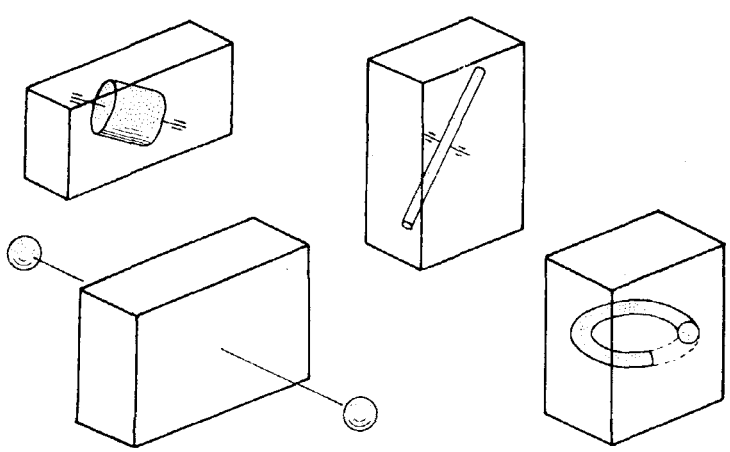


Bild 4. Beispiele für starre Körper mit Zusatzmassen

zeichnet, daß die Trägheitsmomente des Gesamtsystems nicht von der Stellung der inneren Massen abhängen. Bei einem im oder am Körper drehbar gelagerten Stab sowie bei relativ zum Körper bewegten Zusatzmassen innerhalb oder außerhalb des Körpers ist das nicht der Fall. Dennoch kann als gemeinsames Ergebnis der Analyse vieler verschiedener Varianten der in Bild 4 skizzierten Fälle festgestellt werden, daß für den Fall, daß die Zusatzmassen klein gegenüber der Masse des Hauptkörpers sind, eine vorhandene innere Dämpfung zu dem in Bild 3c dargestellten Stabilitätsverhalten führt. Bei größeren Zusatzmassen kann es zu weiteren Einschränkungen für die Stabilitätsbereiche kommen, wofür später zwei Beispiele näher betrachtet werden sollen.

Die bisher betrachteten Fälle bezogen sich ausnahmslos auf den kräftefreien Fall, bei dem die äußeren Momente vernachlässigbar klein sind. Für den „schweren Kreisel" entsteht das äußere Moment durch eine Abweichung des Schwerpunktes vom Fixpunkt, um den der Kreisel frei drehbar gelagert ist. Allgemeinere, d. h. für beliebige Anfangsbedingungen gültige, explizite Lösungen sind aber nur unter zusätzlichen Voraussetzungen gefunden worden. So haben LAGRANGE und KOWALEWSKAJA bei ihren klassischen Untersuchungen Einschränkungen sowohl bezüglich der Form des Trägheitsellipsoides ($A = B$) als auch bezüglich der Lage des Schwerpunktes (S auf einer Hauptachse) eingeführt. STAUDE hingegen, der eine beliebige Form für das Trägheitsellipsoid zuließ, mußte empfindliche Einschränkungen bezüglich der Anfangsbedingungen fordern. Dennoch hat gerade der STAUDEsche Fall in den letzten Jahren ein durchaus praktisches Interesse gewonnen, weil sich herausstellte, daß bestimmte Zentrifugentypen angenähert einem STAUDE-Kreisel entsprechen.

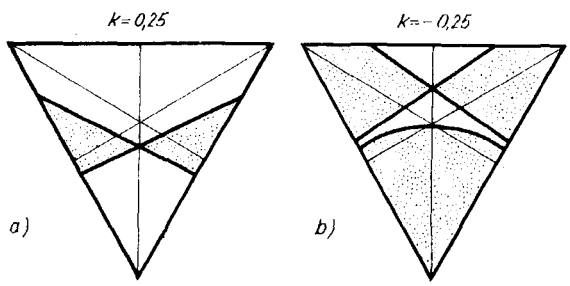


Bild 5. Der STAUDE-Kreisel ohne Dämpfung; a) statisch stabile ..., b) statisch instabile Fesselung

Der klassische STAUDE-Kreisel bildet eine Art Kombination der Fälle von EULER und LAGRANGE. Sein Stabilitätsverhalten wurde in einer ebenfalls schon als klassisch zu bezeichnenden Arbeit von GRAMMEL geklärt. Diese Ergebnisse lassen sich anschaulich im Formdreieck (Bild 5) darstellen. Dabei wird ein dimensionsloser, von der Fesselungsstärke und der Drehgeschwindigkeit Ω abhängiger Beiwert

$$k = \frac{c - mgs}{C\Omega^2} \tag{1}$$

als Parameter eingeführt. Über die STAUDE/GRAMMELschen Ergebnisse hinausgehend, wurde neben der Schwerefesselung (mgs) auch noch eine bei Zentrifugen vorhandene elastische Fesselung (c) berücksichtigt. Im dämpfungsfreien Fall $d = 0$ werden die instabilen Bereiche bei statisch stabiler Fesselung ($k > 0$, Bild 5a) kleiner, bei statisch instabiler Fesselung ($k > 0$, Bild 5b) größer als im Fall des astatischen Kreisels (Bild 3a). Sie sind weiterhin durch Geraden begrenzt, jedoch tritt für $k < 0$ ein zusätzlicher instabiler Bereich auf, der die Anwendung gestreckter Rotoren stark einschränkt. Die folgenden bemerkenswerten Feststellungen können aus den Diagrammen abgelesen werden:

1. Ein statisch stabiler, um die Achse des kleinsten Hauptträgkeitsmomentes drehender Rotor, bei dem man globale Stabilität erwarten würde, kann in bestimmten Drehzahlbereichen instabil werden;
2. ein statisch instabiler, um die mittlere Hauptachse drehender Rotor, bei dem man globale Instabilität erwarten würde, kann in bestimmten Drehzahlbereichen dennoch stabil rotieren.

Für Zentrifugen ist der dämpfungsfreie Fall uninteressant. Deshalb untersuchten SCHIEHLEN und WEBER [1] die Veränderung der Stabilitätsbereiche infolge äußerer Dämpfung. Zwei der dabei erhaltenen Ergebnisse sind in Bild 6 dargestellt. Der Vergleich mit dem in Bild 5 wiedergegebenen Fall $d = 0$ zeigt, daß die Dämpfung bei statisch stabiler Fesselung stabilisierend wirkt — kleine Unsymmetrien werden unschädlich gemacht — während die instabilen Bereiche bei $k < 0$ vergrößert werden. Dieses Ergebnis kann als eine Verallgemeinerung des schon vom LAGRANGE-Kreisel mit Dämpfung bekannten Verhaltens angesehen werden.

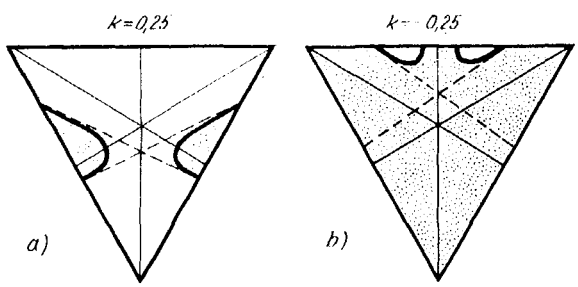


Bild 6. Der STAUDE-Kreisel mit äußerer Dämpfung; a) statisch stabile ..., b) statisch instabile Fesselung

Anforderungen der Kreiselgerätetechnik haben dazu geführt, daß auch Einzelkörper mit äußeren Zusatzmassen genauer untersucht worden sind. Die rahmengelagerten Kreisel mit zwei oder drei Freiheitsgraden sind Beispiele dafür. Auch hier haben sich interessante Verallgemeinerungen klassischer Fälle ergeben:

1. Die Drallachse eines astatisch gelagerten Kardankreisels kann bei schwingendem Kreisel, also z. B. bei Vorhandensein von Nutationsschwingungen, ihre Richtung im Raum ändern („kinetische Drift", s. [2] Kap. 4.4).
2. Für Stabilität eines kardanisch gelagerten Kreiselpendels muß nicht nur die klassische Stabilitätsbedingung des LAGRANGE-Kreisels modifiziert werden, sondern es ist noch eine weitere Bedingung hinzuzufügen, die den Bewegungszustand des Rahmensystems einschränkt ([2], Kap. 4.3).
3. Bei einem kardanisch gelagerten Kreisel mit unsymmetrischem Rotor können je nach den Massenverhältnissen der Rahmen Abweichungen vom Verhalten des EULER-Kreisels auftreten: Es können die Drehbewegungen um alle drei Hauptachsen stabil sein, oder es können stationäre Drehungen um zwei Hauptachsen instabil werden ([2], Kap. 4.5).

3. Gyrostaten

Der in Bild 4 skizzierte Körper mit einem symmetrischen Rotor im Innern wird nach KELVIN als Gyrostat bezeichnet. Das Interesse an Theorie und Phänomenologie der Gyrostaten ist in den letzten Jahrzehnten stark gewachsen, da viele Satelliten zum Zwecke der Lageregelung mit Rotoren versehen werden, also Gyrostaten bilden. Aus den zahlreichen neuerlichen Veröffentlichungen über Gyrostaten sollen hier nur einige, vom kreiseltechnischen Standpunkt besonders interessierende, aufgeführt werden: es soll der Einfluß der Rotordrehgeschwindigkeit ω^R behandelt und damit die Möglichkeit einer Stabilisierung untersucht werden, und es soll der Energietransfer von der Achse des kleinsten zur Achse des größten Hauptträgheitsmomentes am Beispiel eines Gyrostaten mit frei drehbarem, aber relativ zum Gehäuse viskos gebremsten Rotor gezeigt werden.

Wesentliche Ergebnisse auf dem Gebiet der Gyrostatentheorie sind in der Monographie [3] von ROBERSON, WILLEMS und WITTENBURG zusammengestellt worden. Für den Fall eines Gyrostaten mit konstanter Rotordrehzahl hat insbesondere WITTENBURG [4] eine fast erschöpfende Darstellung der analytisch gelösten Fälle gegeben. Er hat durch Ausrechnen der Polkurven und Auftragen auf dem Energieellipsoid gezeigt, wie sich das Verhalten des Gyrostaten im gesamten Bereich von ω^R, vom festgeklemmten Rotor ($\omega^R = 0$) bis zum Gyrostaten mit dominierendem Rotordrall ($\omega^R \to \infty$) ändert. Eine noch elegantere, auch auf Gyrostaten mit nicht konstanter Rotordrehzahl anwendbare Darstellung desselben Übergangs hat H. H. MÜLLER [5] gegeben; das soll hier gezeigt werden.

Bei einem momentenfreien Gyrostaten bleibt nach dem Drallsatz, unabhängig von eventuellen Relativbewegungen des Rotors, der Vektor $\boldsymbol{L}$ des Gesamtdralls nach Größe und Richtung konstant. Sein Endpunkt beschreibt deshalb bei Drehbewegungen des Gyrostaten auf einer mit dem Hüllkörper fest verbunden gedachten Kugel vom Radius L, der „Drallkugel", eine Drall-Polkurve, deren Verlauf, völlig analog zu den bekannten Drehgeschwindigkeits-Polkurven auf dem Energie-Ellipsoid, das grundsätzliche Verhalten der Gyrostaten erkennen läßt (Bild 7a). In den Bildern 7b, 8 und 9 wird gezeigt, wie sich die Polkurven im Bereich $0 \leqq \omega^R < \infty$ verändern und welche stationären Drehbewegungen möglich sind. Bild 7b gilt für den Gyrostaten mit festgeklemmtem Rotor — also für einen starren Einzelkörper. Es gibt sechs Durchstoßpunkte der Hauptachsen durch die Drallkugel; sie entsprechen

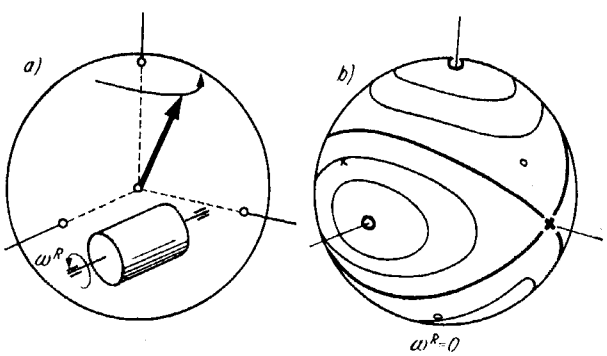


Bild 7. Polkurven auf der Drallkugel eines momentenfreien Gyrostaten. a) Entstehung der Drall-Polkurven, b) Kurven für festgeklemmten Rotor

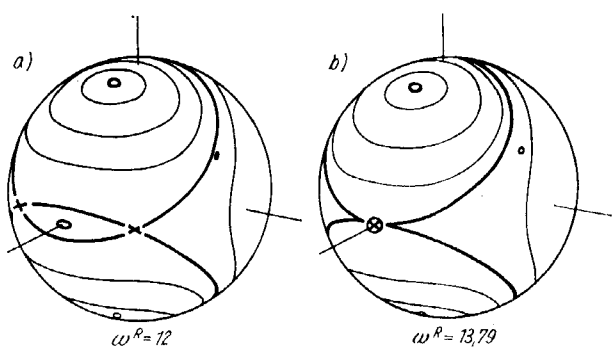


Bild 8. Polkurven auf der Drallkugel

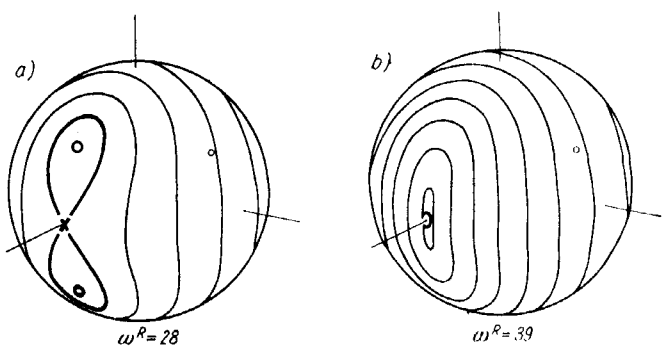


Bild 9. Polkurven auf der Drallkugel

ebensovielen möglichen stationären Drehbewegungen des Gyrostaten, nämlich Drehungen um die drei Hauptachsen mit jeweils verschiedenem Drehsinn. Die Durchstoßpunkte bilden singuläre Punkte des Polkurvenfeldes; vier von ihnen sind stabile Wirbelpunkte (durch Ringe gekennzeichnet), zwei sind instabile Sattelpunkte (durch Kreuze gekennzeichnet). In dem hier dargestellten Fall eines Gyrostaten mit einer zur 1-Achse parallelen Rotorachse behalten die Durchstoßpunkte der 1-Achse ihren Ort auf der Drallkugel unverändert bei. Die anderen Punkte wandern jedoch in Abhängigkeit von der Rotordrehgeschwindigkeit. Es zeigt sich, daß Wirbelpunkte und Sattelpunkte bei anwachsendem ω^R zweimal aufeinander zulaufen; je ein Wirbelpunkt und ein Sattelpunkt verschmelzen dabei und heben sich gegenseitig auf. Die Zahl der möglichen stationären Drehbewegungen wird auf diese Weise von 6 bei kleinem ω^R über 4 auf 2 bei dominierendem Rotordrall verringert. Dieser Übergang kann aus dem Kurvenverlauf der Polkurven und insbesondere aus dem Verlauf der durch die Sattelpunkte laufenden Separatrizen in den Bildern 8a, b und 9a, b im einzelnen verfolgt werden. In Bild 8b ist der Grenzfall skizziert, bei dem gerade zwei Sattel- und ein Wirbelpunkt zusammenfallen. Übrig bleibt ein Sattelpunkt, der jedoch für $\omega^R > 28$ (die Zahlen sind dimensionslose Bezugsgrößen) durch Verschmelzen mit zwei Wirbelpunkten wieder zu einem Wirbelpunkt wird. Für $\omega^R \to \infty$ geht das Polkurvenbild in das bekannte Bild für einen einzelnen symmetrischen Rotor über. Der Hüllkörper hat dann praktisch keinen Einfluß mehr.

Die Stabilisierungsmöglichkeiten für einen Gyrostaten lassen sich in dem für einen Ersatzkörper konstruierten Formdreieck darstellen (Bild 10). Der Ersatzkörper entspricht dem Gyrostaten mit eingefrorenem Rotor, wobei jedoch das Rotorträgheitsmoment bezüglich der Rotordrehachse für diese Achse abgezogen wird. Man kann zeigen, daß bei Variation des Verhältnisses L^R/L (Rotordrall zu Gesamtdrall) für Drehungen des Gyrostaten um die zur Rotorachse parallele Hauptachse („dual-spin-satellites") jeder Punkt des Formdreiecks in einen stabilen oder auch in einen instabilen Bereich gelegt werden kann. Also kann die Achse permanenter Drehungen des Gyrostaten auch bei sonst beliebiger Massenverteilung durch geeignete Wahl des Drallverhältnisses stabilisiert werden.

Die bisherigen Überlegungen gelten für Gyrostaten mit verschiedenem, aber konstantem ω^R. Für Anwendungen bei Satelliten interessiert jedoch auch der Fall veränderlicher Rotordrehgeschwindigkeit. Insbesondere läßt sich damit der Übergang der Drehenergie zur Achse des größten Hauptträgheitsmomentes in allen Einzelheiten verfolgen. H. H. Müller [5] hat hierzu eindrucksvolle Versuche durchgeführt und völlige Übereinstimmung mit theoretischen Ergebnissen erhalten. Zwei seiner Ergebnisse sind in den Bildern 11 und 12 skizziert: es sind die Übergänge des Drallvektors von den stationären Drehungen um die 2- bzw. 1-Achse (mittlere bzw. lange Achse) auf die 3-Achse (kurze Achse) als Drallpolkurven auf der Drallkugel dargestellt. Wegen der Abwesenheit äußerer Momente bleibt der Gesamtdrall auch bei gebremstem Rotor konstant; die Energie dagegen nimmt ab. Deshalb wäre eine Darstellung dieser Übergänge auf dem Energieellipsoid nicht möglich, weil dieses im Verlauf der Bewegung schrumpft.

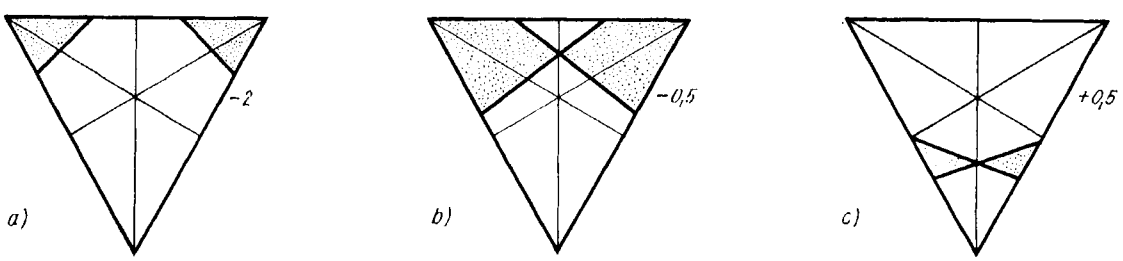


Bild 10. Stabilitätsdiagramme für Gyrostaten bei verschiedenen Werten von L^R/L

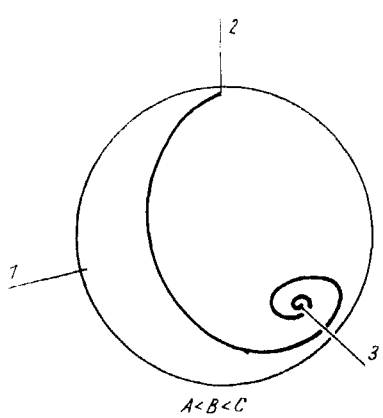


Bild 11. Übergang der Drehbewegung von der 2- auf die 3-Achse für einen Gyrostaten mit viskos gebremstem Rotor

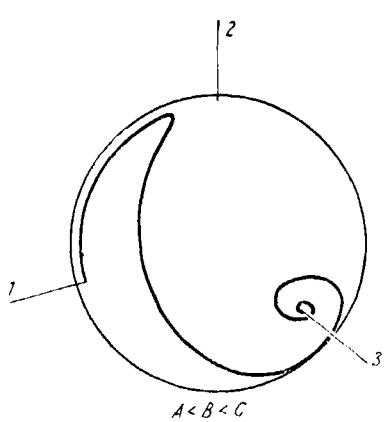


Bild 12. Übergang der Drehbewegung von der 1- auf die 3-Achse für einen Gyrostaten mit viskos gebremstem Rotor

4. Raumflugkörper

Bei Raketen, Satelliten, Raumsonden und Raumstationen sind bezüglich des Kreiselverhaltens zwei Fälle zu unterscheiden: Bei drallstabilisierten Körpern, deren Eigendrehung groß gegenüber der Drehgeschwindigkeit des Bahnumlaufs ist, kann in guter Näherung wie bei kräftefreien Kreiseln gerechnet werden. Dagegen muß bei den

viel verwendeten erdorientierten Satelliten stets das vom Schweregradienten herrührende Moment berücksichtigt werden. Es verschwindet nur bei Körpern mit kugelförmigem Trägheitsellipsoid.

Rotierende Raumstationen mit hinreichend großem Drall, der z. B. zur Simulation eines künstlichen Schwerefeldes vorgeschlagen worden ist, sind vielfach untersucht worden. Bei dem bereits erfolgreich geflogenem Skylab ist auch eine Variante mit Eigendrall durchgerechnet worden, bei der aus konstruktiven Gründen die Eigendrehung Ω gerade um die mittlere Hauptachse erfolgen sollte (Bild 13). Zur Stabilisierung wurden deshalb zwei Zusatzmassen m an langen, elastischen Auslegern parallel zur kurzen Hauptachse angebracht. Sie waren so dimensioniert, daß die Achse der Eigendrehung bei starr angenommenen Auslegern Achse des größten Hauptträgheitsmomentes ist. Es ist demnach zu erwarten, daß bei kleinen Drehgeschwindigkeiten Ω, für die die Nutationsfrequenzen deutlich unter den elastischen Eigenfrequenzen der Ausleger bleiben, Stabilität vorhanden ist. Andererseits haben die Zusatzmassen bei sehr großem Ω, also schnellen Nutationsschwingungen, nur geringen Einfluß auf das Kreiselverhalten, so daß dann Instabilität zu erwarten ist. Die Stabilitätsgrenze hängt demnach von der Elastizität der Ausleger, von Ω und natürlich von der Massenverteilung des Gesamtsystems ab. Theoretische Ergebnisse hierzu sind u. a. von SELTZER, PATEL und SCHWEITZER [6] sowie von LOHMEIER [7] angegeben worden. Ihr wesentliches Ergebnis ist in Bild 14 im Formdreieck dargestellt worden. Zugleich zeigt Bild 14 die von LOHMEIER behandelte Raumstation mit elastischen Auslegern. Bei quasistarren Systemen mit Energiedissipation wäre Stabilität bei Drehungen um die kurze Achse des Gesamtsystems, also für Bildpunkte in den oberen beiden Teildreiecken des Formdreiecks, zu erwarten. Infolge der elastischen Nachgiebigkeit wird dieser Bereich jedoch so eingeschränkt, wie es Bild 14 zeigt. Die neue Stabilitätsgrenze ist eine vom rechten oberen Eckpunkt ausgehende Gerade, deren Neigung von Ω und der Eigenfrequenz ν der Massen m an den Auslegern von der Länge a abhängt. Als Stabilitätsbedingung wird

$$\frac{\Omega^2}{\nu^2 + \Omega^2} > 1 + \frac{C - B}{2ma^2} \tag{2}$$

erhalten. Daraus folgt, daß nachgiebige Satelliten bezüglich der Achse des Eigendralls hinreichend stark abgeplattet sein müssen.

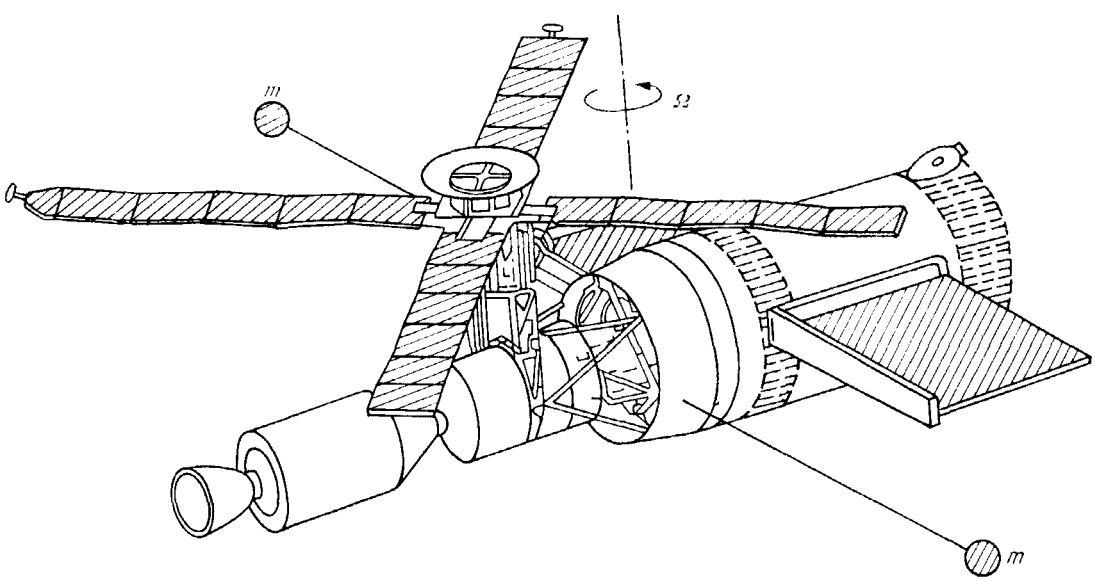


Bild 13. Das Skylab mit Eigendrall und Stabilisierungsmassen

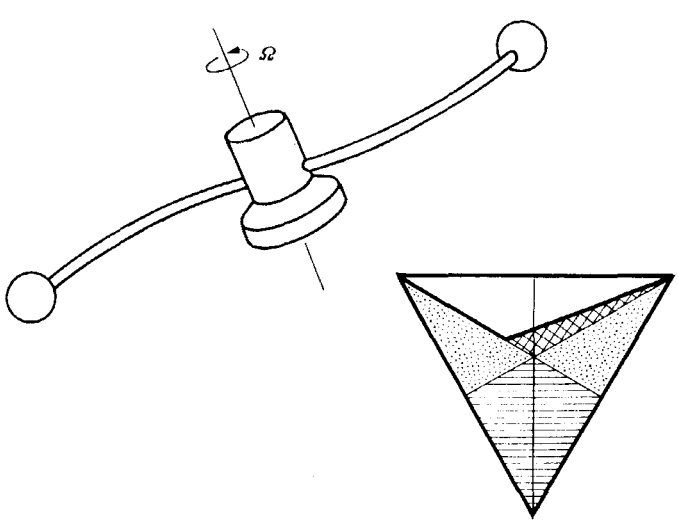


Bild 14. Eine rotierende, elastische Raumstation und ihr Stabilitätsdiagramm

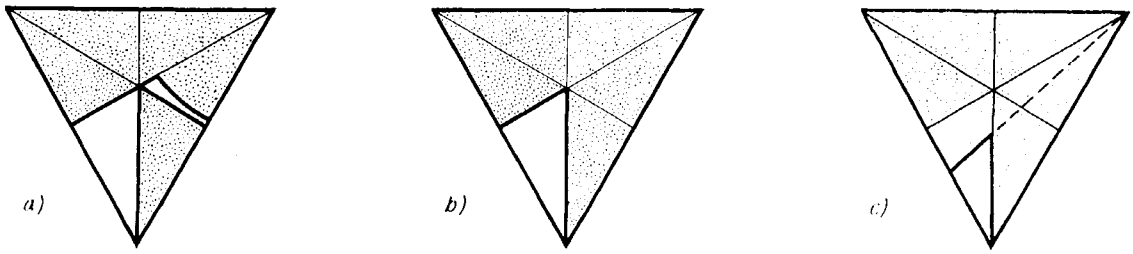


Bild 15. Stabilitätsdiagramm für erdorientierte Raumstationen. a) starr, ohne Dämpfung, b) starr, mit innerer Dämpfung, c) elastisch, mit innerer Dämpfung

Auch für erdorientierte Raumstationen lassen sich Einschränkungen der Stabilitätsbereiche sowohl durch Energiedissipation wie auch durch elastische Nachgiebigkeit feststellen. Hierzu zeigt Bild 15 Beispiele: Der einzelne starre Körper (Bild 15a) kann seine Orientierung zum Anziehungszentrum stabil beibehalten, wenn seine lange Achse zum Anziehungszentrum (Erdmittelpunkt), seine kurze Achse senkrecht zur Bahnebene und die mittlere Achse tangential zur Bahn zeigt. Dieser Orientierung, die z. B. auch von dem natürlichen Erdsatelliten, dem Mond, eingehalten wird, entspricht der durch das linke untere Teildreieck gekennzeichnete Bereich (LAGRANGE-Bereich). In diesem Bereich ist der Körper sowohl statisch infolge der Fesselung durch das Moment des Schweregradienten wie auch kinetisch stabil. Es existiert jedoch noch ein schmaler Stabilitätsbereich im mittleren rechten Teildreieck (DELP-Bereich), für den die Stabilität gyroskopischen Charakter besitzt: der hier statisch instabile Körper wird infolge der durch die Umlaufdrehung bewirkten Kreiselkräfte stabilisiert. Entsprechend dem Stabilitätsgesetz von THOMSON-TAIT-CHETAEV geht diese Stabilität verloren, wenn Energiedissipation durch innere Dämpfung vorhanden ist (Bild 15b). Schließlich wird der Stabilitätsbereich weiter eingeschränkt, wenn elastisch nachgiebige Teilsysteme vorhanden sind. Das ist am Beispiel von zwei elastisch miteinander verbundenen Teilsystemen von POPP [8] ausführlich untersucht worden. Er hat das in Bild 15c skizzierte Stabilitätsdiagramm angegeben. Die neue obere Grenze des Stabilitätsbereiches hängt von der Nachgiebigkeit der elastischen Teile ab. Demnach verbieten flexible Teilsysteme die Verwendung von Raumflugkörpern mit fast kugelförmigem Trägheitsellipsoid. Günstig sind vielmehr Körper mit starker Streckung in Richtung zum Anziehungszentrum.

5. Kreiselsysteme

Gyrostaten sind besonders einfache Zwei-Körper-Systeme. Kardanisch gelagerte Kreisel mit voller Drehfreiheit sind bereits Drei-Körper-Systeme, bei denen ein repräsentierender „Ersatzkörper" schon nicht mehr allgemein definiert werden kann. Reale Systeme, wie sie bei Satelliten oder in der Kreiselgerätetechnik viel verwendet werden, sind noch erheblich komplizierter aufgebaut. Als Beispiele seien Raumflugkörper mit kardangelagerten Regler-Kreiseln oder die für Navigationszwecke verwendeten Trägheitsplattformen oder Zentrifugen-Systeme genannt. Bei genauerer Berechnung des Bewegungsverhaltens derartiger Systeme müssen vielfach dieselben Einflüsse berücksichtigt werden, von denen zuvor berichtet wurde, vor allem Dämpfungs- und Verformungs-Effekte. Hinzu kommen vielfach noch kinematische Zwangsbedingungen, die sich z. B. aus der topologischen Struktur ergeben. So haben die Anforderungen der technischen Praxis zu einer Präzisierung und Erweiterung der klassischen Theorien für Kreiselsysteme geführt, die in zahlreichen Veröffentlichungen dokumentiert wurde. Bewegungsgleichungen vom LAGRANGEschen Typ, bei denen von Energieausdrücken Gebrauch gemacht wird, sind dabei fast ebenso häufig verwendet worden, wie solche vom EULER-NEWTONschen Typ, die auf Impuls- und Drallsatz basieren. Wichtiges Ziel bei der Entwicklung der Untersuchungsmethoden war in jedem Fall eine Formulierung, die das Ausrechnen der Lösungen auf modernen Rechenanlagen erleichtert. Meist sind dabei Darstellungen in Matrizenform verwendet worden.

Über eine für praktische Anwendungen wichtige Erweiterung des von KELVIN-TAIT und ROUTH angegebenen Algorithmus soll hier noch berichtet werden. Sie betrifft die Berücksichtigung von Systemen mit drehzahlgeregelten Kreiseln. Unter der Voraussetzung, daß die Regelung ideal arbeitet, läßt sich die Theorie nämlich erheblich einfacher als im klassischen Fall formulieren. Man geht dabei am besten von einem Ausdruck für die kinetische Energie

$$T = \tfrac{1}{2}\,\dot{\boldsymbol{q}}^T \boldsymbol{M}_S \dot{\boldsymbol{q}} + \tfrac{1}{2}\,\boldsymbol{\omega}^T \boldsymbol{I}_R \boldsymbol{\omega} \tag{3}$$

aus, bei dem der Energieanteil der Rotoren abgespalten wird. $\boldsymbol{I}_R$ ist dabei die Diagonalmatrix der Rotorträgheitsmomente,

$$\boldsymbol{\omega} = \dot{\boldsymbol{\varphi}} + \boldsymbol{B}\dot{\boldsymbol{q}} \tag{4}$$

ist der aus den Drehgeschwindigkeitskomponenten der Rotoren gebildete Spaltenvektor. Er kann aus dem Vektor $\dot{\boldsymbol{\varphi}}$ der Relativgeschwindigkeiten und dem Führungsanteil $\boldsymbol{B}\dot{\boldsymbol{q}}$ zusammengesetzt werden, wobei die Matrix $\boldsymbol{B}$ zugleich von der allgemeinen Struktur des Systems, wie auch von den Einbaurichtungen der Rotorachsen abhängt. Bei nicht geregelten Rotoren und Momentengleichgewicht bezüglich der Rotorachsen sind die φ_i zyklische Koordinaten. Sie können nach KELVIN-TAIT durch Einführen der ROUTHschen Funktion

$$R = T - \boldsymbol{r}^T \dot{\boldsymbol{\varphi}} \tag{5}$$

eliminiert werden, wobei p_i die zu φ_i gehörenden zyklischen Impulse sind. Als Ergebnis werden die nur noch für die nichtzyklischen Koordinaten $\boldsymbol{q}$ geltenden bekannten KELVIN-TAITschen Bewegungsgleichungen erhalten.

Wenn die Rotoren durch eine ideal arbeitende Regelung auf konstanter Relativdrehzahl gehalten werden, dann hat $\dot{\boldsymbol{\varphi}}$ konstante Komponenten. Das wird erreicht durch Antriebsmomente um die Rotorachsen, deren Größe

im einzelnen garnicht bekannt zu sein braucht. Obwohl daher die Energie der Rotoren verändert wird, kann man dennoch wie bei konservativen Systemen rechnen. Die ideale Regelung wirkt sich nämlich so aus, als seien die Rotoren — bei Erhaltung ihres Relativdralls — in ihren Gehäusen eingefroren. Sie können einfach als Zusatzmassen zum Gehäuse hinzugeschlagen werden. Diese physikalisch einleuchtende Tatsache läßt sich unter Berücksichtigung der Gleichungen für die Bewegung um die Rotorachsen und Elimination der Antriebsmomente nachprüfen.

Mit $\dot{\varphi} = \text{const}$ läßt sich jetzt der Energieausdruck (3) mit (4) nach Einführung des konstanten „Drallvektors"

$$\boldsymbol{L} = \boldsymbol{I}_R \dot{\boldsymbol{\varphi}} \tag{6}$$

in die Form

$$\begin{aligned} T &= \underbrace{\tfrac{1}{2}\dot{\boldsymbol{q}}^T \boldsymbol{M}\dot{\boldsymbol{q}}}_{T_2} + \underbrace{\boldsymbol{L}^T \boldsymbol{B}\dot{\boldsymbol{q}}}_{T_1} + \underbrace{\tfrac{1}{2}\boldsymbol{L}^T \boldsymbol{I}_R^{-1}\boldsymbol{L}}_{T_0} \\ T &= T_2 + T_1 + T_0 \end{aligned} \tag{7}$$

mit der Massenmatrix

$$\boldsymbol{M} = \boldsymbol{M}_S + \boldsymbol{B}^T \boldsymbol{I}_R \boldsymbol{B} \tag{8}$$

überführen. Sie enthält die Anteile der wegen der Regelung eingefroren zu denkenden Rotoren. Jetzt zerfällt bereits die kinetische Energie T selbst — und nicht erst die Routhsche Funktion R — in drei Anteile, die quadratisch, linear bzw. unabhängig bezüglich der nichtzyklischen Koordinaten $\boldsymbol{q}$ sind. Deshalb erhält man jetzt ohne weitere Umformung durch Anwendung des bekannten Lagrangeschen Formalismus Bewegungsgleichungen, die völlig analog zu den klassischen Gleichungen von Kelvin und Tait sind:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial T_2}{\partial \dot{\boldsymbol{q}}}\right) - \frac{\partial T_2}{\partial \boldsymbol{q}} = \boldsymbol{Q}^T + \boldsymbol{F}_G^T \tag{9}$$

mit den verallgemeinerten äußeren Kräften $\boldsymbol{Q}$ sowie den gyroskopischen Kräften

$$\boldsymbol{F}_G = \boldsymbol{G}_R \dot{\boldsymbol{q}} = \left\{\frac{\partial(\boldsymbol{B}^T\boldsymbol{q})}{\partial \boldsymbol{q}} - \left[\frac{\partial(\boldsymbol{B}^T\boldsymbol{q})}{\partial \boldsymbol{q}}\right]^T\right\}\dot{\boldsymbol{q}}\,. \tag{10}$$

Im Gegensatz zu den Kelvin-Tait-Gleichungen treten in (9) keine Kräfte infolge der sogenannten kinetischen Fesselungen auf, da die Elemente von $\boldsymbol{I}_R$ von den $\boldsymbol{q}$ unabhängig sind. Nach Ausrechnung der linken Seite kann (9) auch in die Form

$$\boldsymbol{M}\ddot{\boldsymbol{q}} + \boldsymbol{G}_S\dot{\boldsymbol{q}} + \boldsymbol{G}_R\dot{\boldsymbol{q}} = \boldsymbol{Q} \tag{11}$$

gebracht werden, wobei neben den Beschleunigungskräften noch Kräfte $\boldsymbol{G}_S\dot{\boldsymbol{q}}$ vom Coriolis-Typ mit

$$\boldsymbol{G}_S = \frac{\partial(\boldsymbol{M}\dot{\boldsymbol{q}})}{\partial \boldsymbol{q}} - \frac{1}{2}\left[\frac{\partial(\boldsymbol{M}\dot{\boldsymbol{q}})}{\partial \boldsymbol{q}}\right]^T \tag{12}$$

auftreten, die vom Rotordrall unabhängig sind.

In der Theorie von Kreiselsystemen spielen die für schnelle Kreisel oder dominierende gyroskopische Kräfte geltenden Näherungen eine besondere Rolle. So kann man als Näherungsgleichungen einer Präzessionstheorie, die vor allem zur Berechnung langsamer Präzessionsbewegungen geeignet ist, die folgenden, verkürzten Bewegungsgleichungen verwenden:

im Fall $\boldsymbol{\varphi}$ zyklisch:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial R_1}{\partial \dot{\boldsymbol{q}}}\right) - \frac{\partial R_1}{\partial \boldsymbol{q}} = \boldsymbol{Q}^T \quad \text{mit} \quad R_1 = \boldsymbol{p}^T\boldsymbol{B}\dot{\boldsymbol{q}}\,, \tag{13}$$

im Fall $\dot{\boldsymbol{\varphi}} = \text{const}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial T_1}{\partial \dot{\boldsymbol{q}}}\right) - \frac{\partial T_1}{\partial \boldsymbol{q}} = \boldsymbol{Q}^T \quad \text{mit} \quad T_1 = \boldsymbol{L}^T\boldsymbol{B}\dot{\boldsymbol{q}}\,. \tag{14}$$

Wenn jedoch schnelle Nutationsschwingungen ausgerechnet werden sollen, dann können Näherungsgleichungen aus dem Kräftegleichgewicht von Kreisel- und Beschleunigungskräften abgeleitet werden.

Ohne auf Einzelheiten einzugehen, soll noch erwähnt werden, daß in den letzten Jahren einige bemerkenswerte Ergebnisse zu Fragen der Stabilität von Kreiselsystemen erarbeitet worden sind. Zum Beispiel sind Hagedorn [8] Umkehrungen für die klassischen Stabilitätssätze von Lagrange-Dirichlet und Routh gelungen. Anstelle des von Routh betrachteten kinetischen Potentials $W(\boldsymbol{q}) = V - T_0$ mit der potentiellen Energie V und dem von $\dot{\boldsymbol{q}}$ unabhängigen Teil T_0 der kinetischen Energie hat er den von den zyklischen Impulsen unabhängigen Teil H_0 der Hamilton-Funktion als Testfunktion verwendet. Er konnte zeigen, daß die Existenz eines Maximums von $H_0(\boldsymbol{q})$ hinreichend für die Instabilität der Grundlösung eines gyroskopischen Systems ist. Dagegen ist die Existenz eines Minimums von H_0 weder notwendig noch hinreichend für die Stabilität der Lösung. Hierfür ist vielmehr nach Routh die Existenz eines Minimums von $W(\boldsymbol{q})$ hinreichend.

Eine sehr vollständige Zusammenstellung sowie Erweiterungen von Stabilitätssätzen für lineare, gyroskopische Systeme, bei denen auch Dämpfungskräfte sowie nichtkonservative Lagekräfte vorkommen können, sind von P. C. Müller [9] gegeben worden. Er hat zugleich durch Übertragen von Begriffen und Methoden der Regeltheorie den Zusammenhang zwischen Regel- und Kreiselproblemen für die Klärung von Stabilitätsproblemen nutzbar

gemacht. Auf diesem Wege konnten z. B. Probleme der aktiven Dämpfung und der Optimierung von Kreiselsystemen erfolgreich behandelt werden. Insbesondere hat sich die Einführung des Begriffes der „durchdringenden Dämpfung" als überaus nützlich erwiesen. Man versteht darunter solche Dämpfungskräfte, die auch die Bewegungen in den nicht unmittelbar gedämpften Freiheitsgraden beeinflussen können.

Aktuelle Gebiete, auf denen zur Zeit gearbeitet wird und auf denen sicher auch weiterhin Forschungsarbeiten lohnend erscheinen, sind:

- Anpassung der analytischen Methoden an die numerischen Möglichkeiten modernen Rechenanlagen,
- Computereinsatz für ein zumindest teilweises Ableiten der Bewegungsgleichungen, besonders bei Vorliegen komplizierter kinematischer Bedingungen,
- Untersuchung aktiver Kreiselsysteme mit geregelten Teilsystemen,
- Erforschung von Starrkörper-Systemen mit kontinuierlichen, festen oder flüssigen Teilsystemen,
- Berücksichtigung nichtlinearer Effekte, insbesondere Ausweitung der für lineare Systeme erhaltenen Ergebnisse auf nichtlineare.

Ganz allgemein sollten globale Ergebnisse gegenüber den nur für Sonderfälle geltenden punktuellen Ergebnissen bevorzugt werden, damit der Einblick in das physikalische Verhalten komplizierter Systeme vertieft wird. Als Beispiel dieser Art seien die globalen Stabilitätssätze genannt, deren Brauchbarkeit auch zur Lösung konkreter Probleme in der letzten Zeit eindrucksvoll bestätigt werden konnte.

Literatur

1 Schiehlen, W. O.; Weber, M. I., On the stability of Staudes permanent rotations of a gyroscope with damping, Ing.-Arch. **46**, S. 281–292, 1977.
2 Magnus, K., Kreisel, Theorie und Anwendungen, Springer-Verlag, Berlin-Heidelberg-New York, 1971.
3 Roberson, R.; Willems, P.; Wittenburg, J., Rotational dynamics of orbiting gyrostats, Courses and lectures No. 102 of the International Centre for Mechanical Sciences, Udine 1971.
4 Wittenburg, J., Beiträge zur Dynamik von Gyrostaten, Habilitationsschrift, TU Hannover, 1972.
5 Müller, H. H., Zur Bewegung des Gyrostaten mit variablem Rotordrall, Dissertation, TU München, 1976.
6 Seltzer, S. M.; Patel, J. S.; Schweitzer, G., Attitude control of a spinning flexible spacecraft, Comput. & Elect. Engng. Vol. 1, pp. 323–339, 1973.
7 Lohmeier, P., On the stability of spinning flexible satellites, in "Satellite-Dynamics", Springer-Verlag, Berlin-Heidelberg-New York 1975, S. 289–303.
8 Hagedorn, P., Über die Stabilität konservativer Systeme mit gyroskopischen Kräften, Arch. Rat. Mech. Anal. **58**, S. 1–9, 1975.
9 Müller, P. C., Stabilität und Matrizen — Matrizenverfahren in der Stabilitätstheorie linearer dynamischer Systeme, Springer-Verlag, Berlin-Heidelberg-New York, 1977.

Anschrift: Prof. Dr. K. Magnus, Germeringer Straße 13, D-8035 Gauting, BRD

ZAMM 58, T 65 – T 71 (1978)

H. K. Moffatt

Some Problems in the Magnetohydrodynamics of Liquid Metals

When electric currents are caused to flow in an electrically conducting fluid, either by the external application of time-periodic magnetic fields or by the application of large electric potential gradients at the boundary, the associated Lorentz force is in general rotational and a fluid motion, which may be laminar or turbulent, is in general established. Three prototype problems, on which some progress has been made over the last decade, are reviewed: (i) the problem of the generation of rotation in a liquid metal by the application of a rotating magnetic field; (ii) the generation of cellular motion by the application of an alternating field of fixed direction; and (iii) the problem of the generation of fluid motion by the injection of steady current at a point electrode on the fluid boundary. All three problems are of importance in molten metal technology.

1. Introduction

I would like first to thank the Organising Committee for inviting me to give this general lecture. The title that was proposed to me was 'Magnetohydrodynamics' but I felt that it would be hard to do justice to such a wide subject in a single lecture, and I thought it more useful to select a few special problems of some practical importance within the field of liquid metal MHD and to give a brief review of progress on these problems.

It is usual to think of magnetohydrodynamics as a relatively young subject, and it has of course developed enormously over the last 20 years under the stimulus of thermonuclear fusion research, and of parallel research programmes in many centres on astrophysical and geophysical fluid dynamics. The problems that I propose to

discuss are however much more mundane and were in fact known to metallurgists long before the word 'magneto-hydrodynamics' was invented. Significant theoretical progress on these problems has been made only over the last ten years or so, and much remains to be done in bridging the gap between viable mathematical models and the practical realities of the situations considered.

Liquid metals such as molten aluminium or steel are good conductors of electricity and the electric current distribution $\boldsymbol{j}(\boldsymbol{x}, t)$ can interact with the associated magnetic field distribution $\boldsymbol{B}(\boldsymbol{x}, t)$ to give a LORENTZ force distribution $\boldsymbol{F} = \boldsymbol{j} \wedge \boldsymbol{B}$ which may have very strong dynamic effects. In general, this force is rotational and it therefore necessarily drives a rotational velocity field $\boldsymbol{u}(\boldsymbol{x}, t)$ whose distribution is of course controlled by inertia and viscous effects. Let u_0 be a velocity scale characterising $\boldsymbol{u}$, and L a length scale characterising the geometry of the container. Then the magnetic REYNOLDS number is defined by

$$R_m = \mu_0 \sigma L u_0 , \tag{1}$$

where σ is the electric conductivity of the fluid and μ_0 the permeability of free space. If $R_m \ll 1$, then the velocity field has only a weak perturbing effect on the magnetic field distribution, and this may be determined to good approximation by neglecting the fluid motion, i.e. by treating the conductor as if it were solid.

The current $\boldsymbol{j}$ and field $\boldsymbol{B}$ are related by AMPERE's law

$$\mu_0 \boldsymbol{j} = \nabla \wedge \boldsymbol{B} , \tag{2}$$

and if conditions are such that $\boldsymbol{B}$ (and so $\boldsymbol{j}$) are known functions of position, then $\boldsymbol{F}$ is also known. Suppose that the fluid is contained in a finite volume V with fixed rigid surface S, and that $\boldsymbol{F}(\boldsymbol{x})$ is steady. Then, if $\boldsymbol{u}(\boldsymbol{x})$ is the corresponding steady velocity field and $\boldsymbol{\omega} = \nabla \wedge \boldsymbol{u}$ the corresponding vorticity distribution, we have, from the NAVIER-STOKES equation for incompressible steady flow,

$$-\boldsymbol{u} \wedge \boldsymbol{\omega} = -\frac{1}{\varrho} \nabla \left(p + \frac{1}{2} \varrho u^2 \right) + \boldsymbol{F} - \nu \nabla \wedge \boldsymbol{\omega} . \tag{3}$$

The streamlines within V may be closed, or they may cover surfaces (it is easy for example to imagine a situation in which each streamline covers a member of a family of nested toroids). Suppose first that C is a closed streamline. The line integral of (3) round C gives

$$\int_C \boldsymbol{F} \cdot \mathrm{d}\boldsymbol{x} = \nu \int_C \mathrm{d}\boldsymbol{x} \cdot (\nabla \wedge \boldsymbol{\omega}) . \tag{4}$$

It is evident from this that, no matter how small the viscosity of the fluid may be, it is viscous effects alone that can limit the growth of circulation round C when $\boldsymbol{F}$ is rotational.

More generally, if J is a closed surface entirely within V, on which $\boldsymbol{u} \cdot \boldsymbol{n} = 0$, we may easily deduce from (3) that

$$\int_J \boldsymbol{u} \cdot \boldsymbol{F} \, \mathrm{d}S = \nu \int_J \boldsymbol{u} \cdot (\nabla \wedge \boldsymbol{\omega}) \, \mathrm{d}S , \tag{5}$$

and if the left-hand side is non-zero it again follows that the kinetic energy of the motion generated is limited only by viscous effects.

2. The rotating field problem

Fluid contained within a closed surface S can be set in rotation by the application of a rotating magnetic field in the exterior region. This phenomenon was investigated by BRAUNBECK [1] (1932) with the object of devising a method for the measurement of liquid conductivity. The sample, enclosed in a small cylindrical container, is suspended with its axis vertical, and a rotating horizontal field is applied. When suitably calibrated, the rotation of the cylinder about the vertical axis provides a measure of the liquid conductivity (OZELTON and WILSON [2] 1966). The advantage of this method, over the more direct conventional method of simply passing a direct current through the sample, is that it avoids the need for contact between the liquid and inserted solid electrodes.

The rotating field can be regarded as the superposition of two uniform alternating fields out of phase and at right angles. When the field frequency ω is large (compared with $(\mu_0 \sigma L^2)^{-1}$), it penetrates only a small distance $\delta = O(\mu_0 \sigma \omega)^{-1/2}$ into the conductor (the skin effect). The associated LORENTZ force $\boldsymbol{F}(\boldsymbol{x}, t)$, which in general has a mean component $\boldsymbol{F}_0(\boldsymbol{x}) = \langle \boldsymbol{F}(\boldsymbol{x}, t) \rangle$ and a periodic component with frequency 2ω, is then confined to this thin magnetic boundary layer.

The situation is very easily described for the idealised situation in which the cylinder containing the liquid is of circular cross-section and of infinite length (figure 1). The magnetic field is best represented in terms of its vector potential $A\boldsymbol{k}$, where $\boldsymbol{k}$ is the unit vector (0, 0, 1) parallel to the cylinder axis. If a is the radius of the cylinder, then the equations and boundary conditions determining A are simply

$$\begin{aligned} &\partial A/\partial t = \lambda \nabla^2 A && (r < a) \\ &\nabla^2 A = 0 && (r > a) \\ &A \sim B_0 r \sin(\theta - \omega t) \quad \text{as} \quad r \to \infty \\ &[A] = [\partial A/\partial r] = 0 \quad \text{across} \quad r = a , \end{aligned} \tag{6}$$

where $\lambda = (\mu_0 \sigma)^{-1}$ is the magnetic diffusivity of the fluid, and the solution, when $\omega a^2/\lambda \gg 1$, (MOFFATT [3] 1965), is

$$A = \begin{cases} B_0 (r - a^2/r) \sin(\theta - \omega t) & (r > a) \\ 2^{1/2} B_0 \delta \, \mathrm{e}^{-k(a-r)/\delta} \sin(\theta - \omega t + k(a - r) + \pi/4) & (r < a) \end{cases} \tag{7}$$

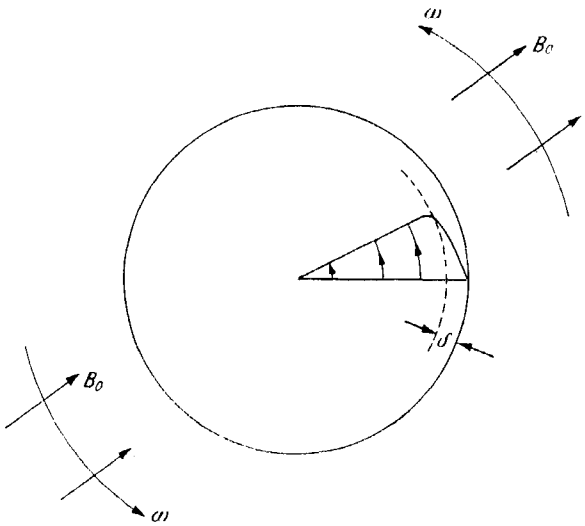


Fig. 1. The velocity distribution inside a cylinder generated by an externally applied magnetic field rotating at high frequency

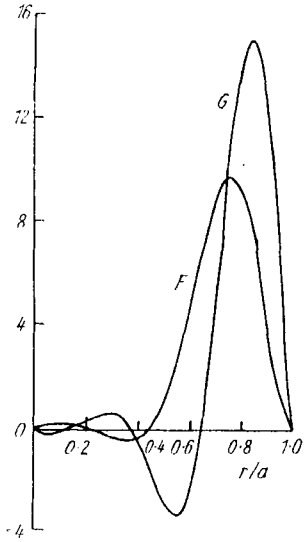


Fig. 2. Radial structure of radial and azimuthal components of the most unstable disturbance of the rotating flow (Richardson [7] 1974)

where $\delta = (\omega/2\lambda)^{-1/2}$, exhibiting the expected boundary layer structure. From this solution, $\boldsymbol{B}$ and $\boldsymbol{j}$ and hence $\boldsymbol{F} = \boldsymbol{j} \wedge \boldsymbol{B}$ (in $r < a$) may be readily calculated. The rate of production of vorticity is given by

$$\nabla \wedge \boldsymbol{F} = -\frac{4B_0^2}{\mu_0 a \delta} e^{-2(a-r)/\delta} \boldsymbol{k} . \tag{8}$$

The periodic ingredient vanishes for this idealised geometry. Dahlberg (1972) has shown that this result remains true even if the frequency ω is low and a boundary layer approach is not applicable.

The velocity field satisfying (3) in this situation is very simple; its streamlines are circular, and the radial distribution is given by

$$v(r) = \Omega(r - a\, e^{-2(a-r)/\delta}) , \tag{9}$$

where

$$\Omega = B_0^2 \lambda / \mu_0 \varrho \nu a^2 \omega . \tag{10}$$

Inside the boundary layer, the fluid rotates rigidly with angular velocity Ω. Note that when ν is small, Ω is large, since viscosity is the only mechanism limiting the angular acceleration of the fluid.

The analysis as described here neglects the effect of the fluid motion on the field. If the applied field is very strong (so that Ω as given by (10) becomes of the same order as ω) this neglect is no longer justified. In the limit of an infinitely strong field, it is clear that (again except in boundary layers which are now of the Hartmann layer type) the fluid effectively acquires rigidity because of the infinite tension in the field lines, and field and fluid then rotate with the same angular velocity ω. This situation has been investigated by Alemany [4] (1976), who also considers effects associated with applied rotating fields of more complicated structure.

The low-frequency weak-field situation was studied by Smith [5] (1964) and by Dahlberg [6] (1972); the resulting velocity field, analogous to (9), but now valid for $\omega a^2/\lambda \ll 1$, is given by

$$v(r) = \Omega_1 (r - r^3/a^2) , \tag{11}$$

where

$$\Omega_1 = a^2 B_0^2 \omega / 16 \mu_0 \lambda \varrho \nu . \tag{12}$$

Note that

$$\frac{\mathrm{d}}{\mathrm{d}r}(rv(r)) = \Omega_1 (2r - 4r^3/a^2) , \tag{13}$$

so that the circulation is a decreasing function of r in the outer region $0.707 < r/a < 1$. The flow is therefore potentially unstable in this outer region. The stability of the profile (11) has been investigated by Richardson [1] (1973). Defining a Taylor number

$$T_1 = a \Omega_1^2 \delta_1^3 / \nu^2 , \tag{14}$$

where $\delta_1 = 0.293a$ represents the radial extent of the unstable region, Richardson's criterion for instability (obtained numerically) is

$$T_1 > T_{1c} \approx 3344 , \tag{15}$$

and the length of the unstable cell in the z-direction under critical conditions is $0.476a$. The radial and azimuthal perturbations have radial structures given by the functions $F(r)$, $G(r)$ reproduced in figure 2. Note the expected concentration of the disturbance in the unstable region, the maximum for $G(r)$ occurring almost precisely at the centre of this region. The disturbance does of course penetrate weakly into the stable region $r \lessapprox 0.707a$, but vortices here must clearly be regarded as 'driven' rather than spontaneous.

The profile (9) may likewise be expected to be unstable if the relevant TAYLOR number

$$T = a\Omega^2\delta^3/\nu^2$$

exceeds a value of order 10^3. This speculation (MOFFATT [3] 1965) has not yet been subjected to analytical or numerical verification.

There are three principal industrial applications of the centrifuging action of a rotating magnetic field: (i) as a straightforward centrifuge for liquid sodium (or other liquid metal) to remove gas bubbles or other contaminants (HAYES et al. [8] 1971); (ii) as a large scale stirrer in the metal casting process (KAPUSTA [9] 1969); and (iii) as a generator of turbulence to accelerate mixing in metallurgical reactions. The scale of these operations is such that the flows generated are almost inevitably turbulent, and laminar theories can at best provide only a qualitative indication of the results to be expected. An experiment under fully turbulent conditions has been carried out by ROBINSON [10] (1973), using a constant-temperature hot-film anemometer to measure both mean and fluctuating components $V(r)$ and $v(r)$ of the azimuthal velocity. A semi-empirical description of the turbulence was devised by LARSSON [10] (1973), and shows the right qualitative trends, although the difference between predicted and measured values of V and v range up to about 40%.

3. Induction furnace problems

Similar semi-empirical methods have been adopted by TARAPORE and EVANS [11] (1976) in a study of the velocities generated in the melt of an induction furnace. Similar calculations have been carried out by HODGKINS [12] (1972). The principle is illustrated in figure 3. An alternating current in the external coils induces a vertical component of magnetic field which diffuses into the melt. The primary purpose is to heat the melt by joule dissipation, and this purpose is of course helped by convection currents driven by buoyancy forces and by the LORENTZ force. The latter is predominantly radial and is maximal near the centre of the system, as indicated in the figure, and a two-cell axisymmetric flow is generated. The upper free surface is generally perturbed, an effect that can be a limiting factor in the operation of such furnaces.

A simpler prototype problem has been studied by SNEYD [13] (1971). Again the fluid domain is idealised by the infinite cylindrical geometry, to which an alternating transverse field is applied. At high frequencies, the skin effect again allows a simple analysis, The LORENTZ force is radially inwards and is greatest at the points of the cylinder where its tangents are parallel to the applied field. This generates a flow with a four-cell structure as illustrated in figure 4. Again the result (4) holds on each closed streamline C.

It is difficult to carry out any analysis (other than computational) for any geometry other than the simple cylinder as described above. In some circumstances however, a local analysis is possible and illuminating. In particular, if the rigid boundaries of the fluid domain have any sharp corners (as they do have in a typical induction furnace) then a local analysis near the corner is indicated. Suppose for example that the fluid is bounded by plane walls $\theta = \pm\alpha$ (figure 5). Then in the notation of § 2, the general solution of $\nabla^2 A = 0$ in the external region for which $A = 0$ on $\theta = \pm\alpha$ is

$$A = Cr^p \cos p(\theta - \pi)\, e^{i\omega t}, \qquad p = \frac{\pi}{2(\pi - \alpha)} > 0\,, \tag{16}$$

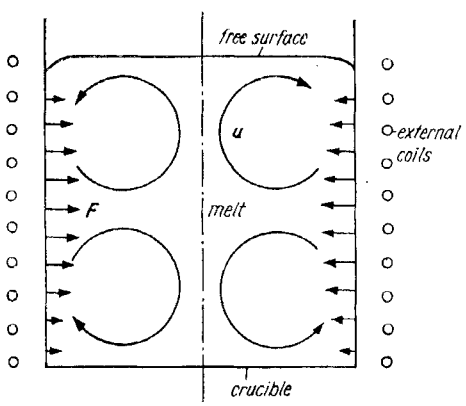


Fig. 3. Sketch of the induction furnace configuration, as studied by TARAPORE & EVANS [11] (1976)

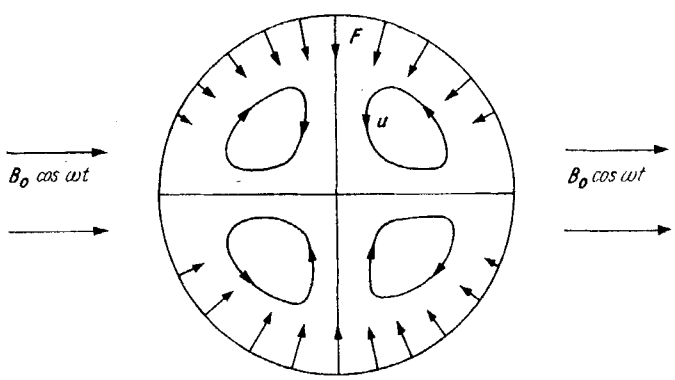


Fig. 4. The idealised model of SNEYD [13] (1971); the LORENTZ force distribution near the circumference of the cylinder drives a flow with a four-cell structure

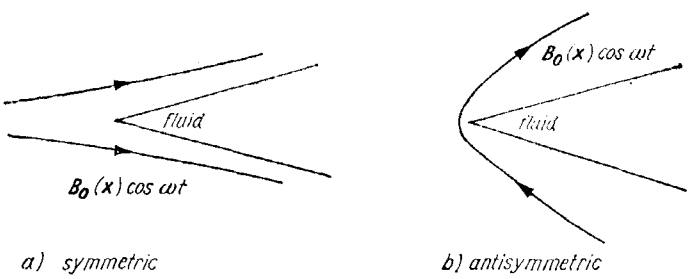


Fig. 5. Symmetric and asymmetric configurations for high frequency field near a sharp corner

or

$$A = Cr^p \sin p(\theta - \pi)\, e^{i\omega t}, \qquad p = \frac{\pi}{\alpha - \pi} < 0 . \tag{17}$$

Equation (16) gives field lines (A = const.) that are symmetrically disposed with respect to the wedge bisector, while (17) gives the antisymmetric configuration. In either case, the equation $\partial A/\partial t = \lambda \nabla^2 A$ may be readily solved in the fluid domain coupled with the condition that the tangential component of $\boldsymbol{B}$ be continuous across $\theta = \pm\alpha$. We again have a skin effect except in the immediate neighbourhood ($r \lessapprox (\lambda/\omega)^{1/2}$) of the vertex. Outside this region the rate of production of vorticity $\nabla \wedge \boldsymbol{F}$ may be calculated; we find

$$\nabla \wedge \boldsymbol{F} = -\frac{1}{\delta} |C|^2 p^2 (p-1)\, x^{2p-3}\, e^{-2y/\delta}, \tag{18}$$

where $\delta = (2\lambda/\omega)^{1/2}$ as before, and y is a coordinate normal to the boundary into the fluid. In the symmetric case given by (16),

$$2p - 3 = \frac{3\alpha - \pi}{\pi - \alpha} \gtrless 0 \quad \text{acc. as} \quad \alpha \gtrless \frac{\pi}{3}. \tag{19}$$

Vorticity production apparently increases as the corner is approached if $\alpha < \pi/3$, a singularity being thwarted only in the small excluded region $r \lessapprox (\lambda/\omega)^{1/2}$.

In the antisymmetric case,

$$2p - 3 = \frac{3\alpha - 4\pi}{\pi - \alpha} < 0 \quad \text{for all } \alpha \tag{20}$$

and in this case vorticity production inevitably increases as the corner is approached for all values of α.

4. The weld-pool problem

A closely related class of problem is that in which a steady current is injected into a volume of conducting fluid by prescription of the electrostatic potential distribution φ over its boundary. Again neglecting the weak perturbing effect of the fluid motion, the current field is then simply given by

$$\boldsymbol{j} = -\sigma \nabla \varphi \tag{21}$$

and if this current is the only source of magnetic field, $\boldsymbol{B}$ is determined by

$$\nabla \cdot \boldsymbol{B} = 0, \qquad \nabla \wedge \boldsymbol{B} = \mu_0 \boldsymbol{j} = -\mu_0 \sigma \nabla \varphi . \tag{22}$$

The prototype situation, which has been studied by ZHIGULEV (1960) and SHERCLIFF (1970), is that in which current is injected from a point electrode into a half-space of conducting fluid (figure 6). This can be regarded as a primitive model of what happens in the neighbourhood of the contact both in the arc welding process, and in the arc furnace in which a container of liquid metal is heated by just this method, viz. the injection of a large steady current at a point on its surface. In this latter context, uniform heating of the melt depends critically on the convection currents induced, and for this reason again the dynamics of the system have to be considered.

In the problem as formulated by SHERCLIFF, in spherical polar coordinates (r, θ, φ) the current is given by

$$\boldsymbol{j} = (J/2\pi r^2, 0, 0) \tag{23}$$

and, by AMPERE's law, the field $\boldsymbol{B}$ is then purely azimuthal and has the form

$$\boldsymbol{B} = (0, 0, \mu_0 J \sin\theta/2\pi r\,(1 + \cos\theta)). \tag{24}$$

Hence

$$\boldsymbol{F} = \boldsymbol{j} \wedge \boldsymbol{B} = (0, -\mu_0 J^2 (\sin\theta)/4\pi^2 r^3 (1 + \cos\theta), 0), \tag{25}$$

and

$$\nabla \wedge \boldsymbol{F} = \left(0, 0, \frac{\mu_0 J^2}{2\pi^2 r^4} \frac{\sin\theta}{(1 + \cos\theta)}\right). \tag{26}$$

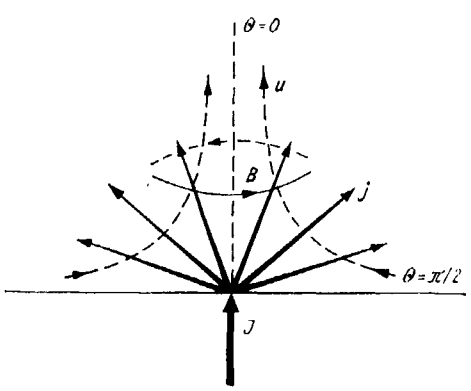


Fig. 6. Current injection into a half-space, as analysed by SHERCLIFF [15] (1970)

Note that $F_\theta < 0$ for all $\theta < \pi/2$, and the magnitude of F_θ decreases as θ decreases from $\pi/2$ to 0. The force field therefore drives a jet-type motion along the axis of symmetry, the fluid being drawn in laterally by a 'pinch effect' which is naturally strongest near to the current source.

The fluid particles that pass very near the source singularity acquire a very large vorticity (due to the r^{-4} dependence in (26)). SHERCLIFF showed, on the basis of an inviscid analysis that if the flow 'upstream' is irrotational, then the flow 'downstream' is necessarily singular on the axis of symmetry. As pointed out by SOZOU (1971), viscous effects must be included if a well-behaved solution is to be found; this may be seen again from equation (1): although the streamlines are not closed (the flow domain being infinite) the result (1) still holds on any streamline C provided the pressure is uniform and the velocity zero at infinity; a steady state satisfying these conditions is therefore not possible unless viscous effects are taken into account.

Dimensional analysis led SOZOU to seek a similarity solution to the problem in which the STOKES stream function is given by

$$\psi = \nu r g(\mu, K)\,, \tag{27}$$

where $\mu = \cos\theta$ and

$$K = \mu_0 J_0^2/\varrho\nu^2\,. \tag{28}$$

This form of solution is suggested by the fact that there is no natural length-scale in the problem, and K is the sole dimensionless parameter that can be constructed. The situation is closely analogous to the classical round jet problem, for which a flow is generated by the application of a point force (or equivalently a point source of momentum) on a plane boundary. In the present context, the force is distributed rather than concentrated at a point, but its radial dependence ($\sim r^{-3}$) is just such as to be in possible equilibrium with both the inertia force $\varrho \boldsymbol{u}.\nabla\boldsymbol{u}$ and the viscous force $\varrho\nu\nabla^2\boldsymbol{u}$ when $\boldsymbol{u} \propto r^{-1}$, i.e. when ψ is given by (27).

There is however a difficulty in pursuing this analogy that does not appear to have been fully appreciated. For, on the one hand, a streamfunction of the form (27) is associated with a definite flux of momentum F_0 in the direction of the axis of symmetry (see, for example, BATCHELOR 1967, § 4.6). If $g(\mu, k)$ is known, then integration of the associated momentum flux over any hemisphere $r = R$, $0 < \pi/2$, gives a definite relationship of the form

$$F_0 = 2\pi\varrho\nu^2 G(K) \tag{29}$$

for some function G, and this value of F_0 must be interpreted as the equivalent point force at $r = 0$ which will generate a jet flow having the same momentum flux as the flow given by (27).

On the other hand, we can easily calculate the total force imparted to the fluid in the region $r_0 < r < R$ on the basis of (25), viz.

$$\hat{F}_0 = -\int_{r_0}^{R}\int_{0}^{\pi/2} F_\theta \sin\theta \cdot 2\pi r^2 \sin\theta \,\mathrm{d}r\,\mathrm{d}\theta = \frac{\mu_0 J^2}{4\pi}\log\left(\frac{R}{r_0}\right), \tag{30}$$

which diverges logarithmically as $R/r_0 \to \infty$. Unless this divergence is compensated by similar divergence in the suction exerted by the plane $\theta = \pi/2$ as $R \to \infty$, this implies an unbounded flux of momentum in the fluid as $r \to \infty$. SOZOU did in fact find, by integration of the non-linear ordinary differential equation for $g(\mu, K)$, that singularities appeared on the axis $\mu = 1$ for large values of K (> 300). The above argument raises questions concerning the physical realisability of solutions of the form (27) for any value of K.

The most reasonable way to resolve this sort of difficulty is of course to return to the problem of a finite fluid domain, and at the same time to replace the point electrode by an electrode of finite size. A step in this direction has been taken by SOZOU & PICKERING (1976) who consider the effect of the force distribution (25) within a hemispherical container $r < R$, $\theta < \pi/2$, and obtain steady streamline patterns by integrating the nonlinear equation for the stream-function $\psi(r, \theta)$ numerically. There is still however in this situation a difficulty in the overall momentum balance. The force $\hat{F}_0$ given by (30) becomes infinite as $r_0 \to 0$ for fixed R, and this infinite volume force imparted to the fluid must be imparted (by momentum conservation) to the boundaries containing the fluid. The only point of the boundary at which this infinity can reasonably be accounted for is the point electrode itself; at this point, the boundary must exert an infinite suction, a situation that would inevitably lead to local cavitation, and intermittency in the resulting current passed to the liquid. SOZOU & PICKERING actually suppose that the surface $\theta = \pi/2$ is a free surface (as is appropriate in the technological applications mentioned above); but the assumption of a point source of current on the boundary must then inevitably lead to a singularity in the surface deformation at this point also. (This may equally be appreciated in terms of the infinite magnetic pressure at the origin.)

The alternative way to try to resolve the difficulty is to accept that where the velocity is large, neglect of its effect on the magnetic field distribution may no longer be tenable. Allowance for field convection and diffusion introduces one new physical parameter into the problem, viz. the magnetic diffusivity λ. We now have the very curious situation of a problem defined in terms of three dimensional parameters $(\mu_0/\varrho)^{1/2} J$, ν and λ all having the same dimensions (length)^2 (time)$^{-1}$, from which we still cannot construct a natural length-scale. The current lines must therefore still be radial, so that instead of (23) we can have only

$$\boldsymbol{j} = ((J/2\pi r^2)\, f(\theta),\, 0,\, 0) \tag{31}$$

for some function $f(\theta)$ satisfying

$$\int_0^{\pi/2} f(\theta)\sin\theta\,\mathrm{d}\theta = 1\,. \tag{32}$$

An easy calculation now leads to the appropriate modification of (30) viz

$$\hat{F}_0 - \frac{\mu_0 J^2}{2\pi} \log\left(\frac{R}{r_0}\right) \int_0^{\pi/2} f(\theta) \cos\theta \sin\theta \, d\theta \,. \tag{33}$$

The only way that $\hat{F}_0$ can remain finite as $R/r_0 \to \infty$ is if

$$\int_0^{\pi/2} f(\theta) \cos\theta \sin\theta \, d\theta = 0 \,. \tag{34}$$

This can be satisfied (in conjunction with (32)) only if $f(\theta)$ is negative for some values of θ in the range $0 < \theta < \pi/2$. For example, the function

$$f(\theta) = 6 \cos\theta - 4 \tag{35}$$

satisfies both (32) and (34). Indications of reversed current flow have in fact been found in numerical computation incorporating induction (or 'field sweeping') effects by SOZOU & ENGLISH (1972), but the extent and intensity of the reversed current does not appear sufficient for satisfaction of the condition (34). The possibility of current reversal was also noted by SHERCLIFF [15].

It must be concluded that, in spite of the conceptual simplicity of the idealised problem as posed by SHERCLIFF, the solutions that have so far been proposed have internal inconsistencies that have yet to be fully resolved. It should perhaps be noted, moreover, that even if the laminar problem were fully understood, the flow is very likely to be unstable when K exceeds some critical value, and a turbulent state is then the most likely outcome.

Acknowlegdment

I am grateful to Dr. W. R. HODGKINS of the Electricity Council Research Center, Capenhust, for drawing my attention to the induction furnace problem.

References

1 BRAUNBECK, W., Eine neue Methode elektrodenloser Leitfähigkeitmessung, Z. Phys. **73**, 312—334 (1932).
2 OZELTON, N. W., WILSON, J. R., A rotating field apparatus for determining resistivities of reactive liquid metals and alloys at high temperatures, J. Sci. Instrum. **43**, 359—363 (1966).
3 MOFFATT, H. K., On fluid flow induced by a rotating magnetic field, J. Fluid Mech. **22**, 521—528 (1965).
4 ALEMANY, A., The flow of conducting fluids in a circular duct under rotating magnetic fields with several dipoles, MHD-flows and Turbulence (Ed. H. BRANOVER), Israel Universities Press, Jerusalem, 17—31 (1976).
5 SMITH, P., The rotation of a conducting liquid in a uniform transverse field, ZAMM **44**, 495—502 (1964).
6 DAHLBERG, E., On the action of a rotating magnetic field on a conducting liquid, AB Atomenergi, Sweden Rep. AE-447 (1972).
7 RICHARDSON, A. T., On the stability of a magnetically driven rotation fluid flow, J. Fluid Mech. **63**, 593—605 (1974).
8 HAYES, D. J., BAUM, M. R., HOBDELL, M. R., The performance and applications of an electromagnetic rotary-flow device in liquid sodium, J. Br. Nucl. Energy Soc. **10**, 93—98 (1971).
9 KAPUSTA, A. B., Theory of centrifugal casting in a rotating magnetic field, Mag. Gidrod. **5**, 117—120 (1969).
10 ROBINSON, T., (with appendix by K. LARSSON), An experimental investigation of a magnetically driven rotating liquid-metal flow, J. Fluid Mech. **60**, 641—664 (1973).
11 TARAPORE, E. D., EVANS, J. W., Fluid velocities in induction melting furnaces: Part 1. Theory and laboratory experiments. Metal. Trans. 7B, 343—351 (1976).
12 HODGKINS, W. R., Mathematical calculations on electromagnetic stirring, Electricity Council Res. Center MM 12 (1972).
13 SNEYD, A., Generation of fluid motion in a circular cylinder by an unsteady applied magnetic field, J. Fluid Mech. **49**, 817—827 (1971).
14 ZHIGULEV, V. N., The phenomenon of ejection by an electrical discharge, Soviet Phys. Doklady, **5**, 36—39 (1960).
15 SHERCLIFF, J. A., Fluid motions due to an electric current source, J. Fluid Mech. **40**, 241—250 (1970).
16 SOZOU, C., On fluid motions induced by an electric current source, J. Fluid Mech. **46**, 25—32 (1971).
17 BATCHELOR, G. K., An introduction to fluid dynamics, Cambridge University Press (1967).
18 SOZOU, C., PICKERING, W. M., Magnetohydrodynamic flow due to the discharge of an electric current in a hemispherical container, J. Fluid Mech. **73**, 641—650 (1976).
19 SOZOU, C., ENGLISH, H., Fluid motions induced by an electric current discharge, Proc. R. Soc. Lond. A **329**, 71—81 (1972).

Address: Prof. H. K. MOFFATT, University of Bristol, School of Mathematics, University Walk, Bristol BS8 1TW, Great Britain

ZAMM 58, T 72 - T 85 (1978)

K. Nickel

Intervall-Mathematik

Eines der wesentlichen *Ziele* der Intervall-Mathematik ist es, allgemeine *Mengen* durch *Intervalle* einzuschränken. Die betrachteten *Mengen* sind dabei i. a. nicht explizit bekannt, es handelt sich meistens um *Lösungsmengen* von: Fixpunktgleichungen, Differentialgleichungen, Integralgleichungen, ... Es werden *Intervalle* verwendet, weil es sich dabei um eine besonders einfach beschreibbare, 2-parametrige Schar von speziellen Mengen handelt. Bei dieser Beschäftigung treten die verschiedensten Probleme auf aus: Arithmetik, Analysis, Topologie, Algebra, Numerik, etc.; ihre Gesamtheit ist in der Intervall-Mathematik enthalten. Der Vortrag berichtet über einige dieser Entwicklungen.

Die Intervall-Mathematik existiert seit über 10 Jahren. Bisher ist sie jedoch immer noch ein „Veilchen, das im Verborgenen blüht". Zum Beispiel gibt es bis heute nur zwei Bücher über dieses Gebiet, die Bücher von R. E. Moore (1966) und (1969) und Alefeld und Herzberger (1974). Die Methoden und Ergebnisse der Intervall-Mathematik sind bis jetzt nur in ganz wenige andere Bücher eingedrungen, man vgl. etwa Henrici (1974). Bis jetzt fanden — neben vielen lokalen Tagungen (meistens in Oberwolfach) — zwei internationale Kongresse statt, nämlich 1968 in Oxford (siehe die Proceedings von Hansen (1969)) und 1975 in Karlsruhe (siehe die Proceedings von Nickel (1975a)).

Ich bin dem Vorstand der GAMM und der örtlichen Tagungsleitung außerordentlich dankbar dafür, daß ich hier in Kopenhagen die Möglichkeit habe, erstmalig vor der GAMM einen zusammenfassenden Bericht über diese neue Disziplin zu geben. Selbstverständlich ist es nicht möglich, innerhalb von einer Stunde den Inhalt von über zehnjähriger Arbeit von vielen Mathematikern und von über 700 Publikationen vollständig darzustellen (eine Literaturübersicht wurde von Bierbaum (1976) erstellt). Ich werde theoretische Aspekte weitgehend vernachlässigen und mich — entsprechend der Zielsetzung der GAMM — hauptsächlich auf Aspekte der *Anwendungen und der Numerik* beschränken.

Vor fast genau 200 Jahren, am 30. April 1777, wurde Carl Friedrich Gauss geboren, einer der größten Mathematiker und Angewandten Mathematiker. Es scheint mir daher angemessen, die Methoden und Ergebnisse der Intervall-Mathematik auch an mathematischen Problemen zu erläutern, mit denen sich schon Gauss befaßte.

I. Grundlagen

1. Reelle Intervalle

Über Jahrhunderte hinweg galten bei Mathematikern die „imaginären" und die „komplexen" Zahlen nicht als wirkliche „Zahlen". Erst mit Gauss und der „Gaussschen Zahlenebene" wurden die komplexen Zahlen voll anerkannt. Ich möchte den Aufbau der reellen Intervalle und der reellen Intervall-Funktionen in Analogie zur Gaussschen Zahlenebene und zum Aufbau der Funktionentheorie bringen und damit (hoffentlich) verdeutlichen.

Komplexe Zahlen	Reelle Intervalle
$\mathbb{C}$	$I(\mathbb{R})$
Darstellung	
$z = x + iy$, $w = u + iv$.	$[x] = [\underline{x}, \bar{x}] := \{x \in \mathbb{R} \mid \underline{x} \leqq x \leqq \bar{x}\}$, $[y] = [\underline{y}, \bar{y}]$.
Basis	
$1, i$	$1, [-1, +1]$
Halbordnung(en)	
$w \leqq z \Leftrightarrow u \leqq x \wedge v \leqq y$ (komponentenweise).	$[x] < [y] \Leftrightarrow \bar{x} < \underline{y}$, $[x] \leqq [y] \Leftrightarrow \underline{x} \leqq \underline{y} \wedge \bar{x} \leqq \bar{y}$, $[x] \subseteqq [y] \Leftrightarrow \underline{y} \leqq \underline{x} \wedge \bar{x} \leqq \bar{y}$ (siehe Bild 1).
Arithmetik	
bekannt.	$[x] \overset{+}{/} [y] := \{x \overset{+}{/} y \mid x \in [x], y \in [y]\}$.

Die arithmetischen Operationen sind

Erweiterungen

aus dem Reellen. Sie werden durch endlich viele reelle Verknüpfungen erzeugt, sind also programmierbar.

$\mathbb{C}$ ist ein *Körper*.	$I(\mathbb{R})$ ist *kein Körper, keine* additive oder multiplikative Gruppe, es gilt *nicht* das *Distributivgesetz*.

Eine rationale Funktion

$F : \mathbb{C} \to \mathbb{C}$	$F : \mathbb{I}(\mathbb{R}) \to \mathbb{I}(\mathbb{R})$

ist unbeschränkt definierbar, solange

$0 \neq$ Nenner .	$0 \notin$ [Nenner] .

Sie ist eine *Erweiterung* der zugehörigen reellen rationalen Funktion $f: \mathbb{R} \to \mathbb{R}$, d. h.

$F(x + i \cdot 0) = f(x)$.	$F[x, x] = f(x)$.

Metrik

$\lvert w, z\rvert^2 := (u - x)^2 + (v - y)^2$	$\lvert\lvert[x], [y]\rvert\rvert := \max(\lvert\underline{x} - \underline{y}\rvert, \lvert\bar{x} - \bar{y}\rvert)$

damit lassen sich Stetigkeit, Konvergenz, etc. definieren, und es läßt sich Analysis treiben.

Eine rationale Funktion

$F : \mathbb{C} \to \mathbb{C}$	$F : \mathbb{I}(\mathbb{R}) \to \mathbb{I}(\mathbb{R})$

ist

holomorph, d. h. beliebig oft differenzierbar.	*inklusionsisoton*, d. h. $[x] \subseteq [y] \Rightarrow F[x] \subseteq F[y]$.

Man betrachtet daher die *Menge aller*

holomorphen Funktionen.	*inklusionsisotonen Funktionen.*

Die graphische Darstellung einer intervallwertigen Funktion F mit reellem Argument ($F: \mathbb{R} \to \mathbb{I}(\mathbb{R})$) ist ein Funktions-„Schlauch" (siehe Bild 2). Eine Funktion $F: \mathbb{I}(\mathbb{R}) \to \mathbb{I}(\mathbb{R})$ läßt sich (wie im Komplexen) nicht mehr so einfach graphisch interpretieren.

Von reellen Intervallen kann man übergehen zu Intervall-Vektoren und Intervall-Matrizen; sowie allgemeiner zu Intervallen über halbgeordneten Räumen. Beispiele dafür sind Funktionsintervalle (siehe Bild 3) und Intervalle von Operatoren. Die Intervallrechnung über Verbänden liefert ohne weitere Voraussetzungen allgemeine Fix-Intervall-Sätze, man vgl. etwa Nickel (1975b).

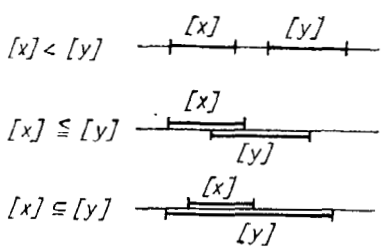


Bild 1. Ordnungsrelationen bei Intervallen

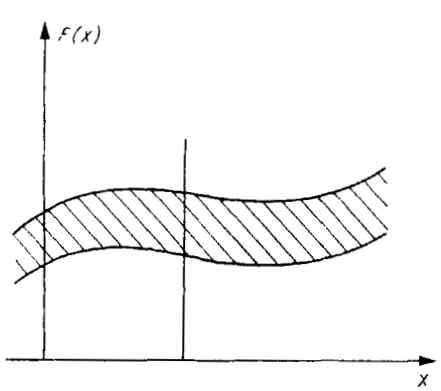


Bild 2. Darstellung einer Intervall-Funktion $F: \mathbb{R} \to \mathbb{I}(\mathbb{R})$

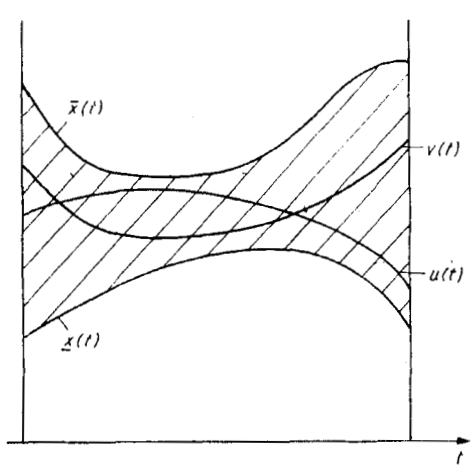


Bild 3. Funktionsintervall $[x] = [\underline{x}(t), \bar{x}(t)]$ mit 2 Repräsentanten $u, v \in [x]$

I. 2. Anwendungen

I.2.1. Schranken

Es sei F inklusionsisotone Intervallerweiterung zu f. Dann gilt für die Bildmenge der Funktion f auf dem Intervall $[x]$:

$$\{f(x) \mid x \in [x]\} \subseteq F[x] .$$

Durch bloßes „Ausrechnen" von $F[x]$ erhält man also *Schranken* für den Bildbereich. Ohne weitere Informationen (z. B. Lipschitz-Konstante o. ä.) läßt sich dieses Problem „rein reell" nicht lösen.

I.2.2. Numerische Quadratur

Zu einer gegebenen integrierbaren Funktion $f: [a, b] \to \mathbb{R}$ soll

$$x := \int_a^b f(t)\,\mathrm{d}t$$

einschließlich Schranken näherungsweise bestimmt werden.

Man zerlegt $[a, b]$ durch

$$a := t_0 < t_1 < \ldots < t_n := b\,, \qquad h_i := t_i - t_{i-1} \quad \text{für} \quad i = 1(1)\,n\,.$$

Ist dann F eine inklusionsisotone Intervall-Erweiterung zu f, dann gilt

$$x \in [x] := \sum_{i=1}^{n} h_i F[x_i]$$

(Approximation durch Riemannsche Unter- und Obersummen).

Ist F zusätzlich noch stetig, dann ergibt sich für $\max\limits_{i=1(1)n}(h_i) \to 0$ sogar Konvergenz; allerdings nur lineare Konvergenz. Die Bedeutung dieser Formel liegt trotz dieses Nachteils darin, daß keinerlei weiteren Informationen (wie etwa Lipschitz-Schranken, Ableitungs-Schranken, Holomorphie-Aussagen, ...) erforderlich sind. Selbstverständlich gibt es auch aufwendigere Verfahren von beliebig hoher Ordnung. In diesem Falle sind jedoch höhere Anforderungen an die Funktion $f(t)$ zu stellen.

I.2.3. Lipschitz-Bedingung

Gibt es zur Ableitung f' der Funktion $f: [a, b] \to \mathbb{R}$ eine inklusionsisotone Intervall-Erweiterung $F': \mathbb{I}[a, b] \to \mathbb{I}(\mathbb{R})$, dann genügt f der *Intervall-Lipschitz-Bedingung*

$$f(x) - f(y) \in [m]\,(x - y) \quad \text{für alle} \quad x, y \in [a, b]\,. \tag{1}$$

Das Lipschitz-Intervall $[m]$ ist dabei berechenbar durch $[m] := F'[a, b]$.

I.2.4. Intervall-Newton-Verfahren

Die Funktion $f \in C[a, b]$ besitze eine Nullstelle $\hat{x} \in [a, b]$ und genüge einer Intervall-Lipschitz-Bedingung (1) mit $[m] \ni 0$. Dann ist das Intervall-Newton-Verfahren

$$\left.\begin{aligned} &[x_0] := [a, b]\,,\\ &x_\nu \in [x_\nu] \quad \text{beliebig}\,,\\ &[x_{\nu+1}] := [x_\nu] \cap (x_\nu - f(x_\nu)/[m]) \end{aligned}\right\} \nu = 0, 1, \ldots$$

stets konvergent gegen $\hat{x}$, d. h. es gilt (im Sinne der Intervall-Metrik)

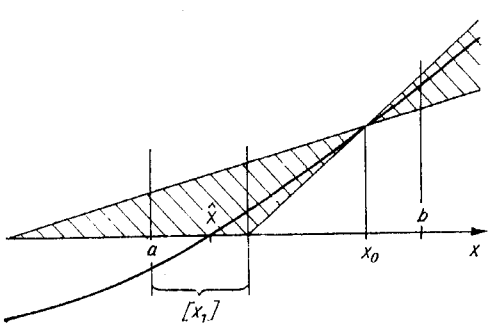
$$\hat{x} \in [x_\nu] \to \hat{x}\,, \qquad x_\nu \to \hat{x}\,.$$

Bild 4 veranschaulicht diese Formeln, die Funktion $f(x)$ und damit die Nullstelle $\hat{x}$ ist nach (1) „eingefangen" in dem schraffierten Bereich (der nur oberhalb der x-Achse gezeichnet ist).

Man beachte, daß in jedem Schritt gleichzeitig reelle Größen x_ν, $f(x_\nu)$ und Intervalle $[x_\nu]$, $[m]$ benutzt werden. Dies ist typisch für die meisten Intervall-Verfahren.

Ein entsprechender reeller Satz existiert ohne zusätzliche Annahmen bekanntlich nicht! Allein wegen dieses spektakulären Satzes „lohnt" sich die Erfindung der Intervall-Arithmetik und -Analysis.

Zur „Geschichte" dieses Verfahrens vergleiche man etwa R. E. Moore (1966) oder (1969) (wo die globale Konvergenz noch nicht bekannt war) und Nickel (1968), (1969). Man kann, wie üblich, unter passenden Voraussetzungen an f und/oder passender Abänderung des Verfahrens überlineare bzw. quadratische Konvergenz nach-



$$\left.\begin{aligned} &[x_0] := [a, b]\,,\\ &x_\nu \in [x_\nu] \quad \text{beliebig}\\ &[x_{\nu+1}] := [x_\nu] \cap (x_\nu - f(x_\nu)/[m]) \end{aligned}\right\} \nu = 0, 1, \ldots$$

Bild 4. Intervall-Newton-Verfahren

weisen. In den vergangenen Jahren wurden viele Variationen dazu angegeben, man vergleiche etwa KRAWCZYK (1969) sowie ALEFELD und HERZBERGER (1974). Weiter läßt sich — im Gegensatz zum Reellen — auch der Fall $0 \in [m]$ behandeln (siehe KRAWCZYK (1969) und andere) und damit sogar eine Methode zur Berechnung aller Nullstellen von f in dem Intervall $[a, b]$ erzeugen, siehe BARTH (1972a).

Bei der Übertragung auf (nichtlineare) Gleichungssysteme geht i. a. die Eigenschaft der globalen Konvergenz verloren (eine hinreichende Bedingung dafür wurde von ALEFELD und HERZBERGER (1970) angegeben). Jedoch bleiben Lösungseinschließung und lokale (überlineare und/oder quadratische) Konvergenz erhalten. Man vergleiche dazu NICKEL (1971), ALEFELD und HERZBERGER (1974).

I. 3. Gerundete reelle Intervalle

I.3.1. Zahlenlänge

Bekanntlich war GAUSS ein sehr eifriger, dabei aber ein sehr sorgfältiger Rechner. Von ihm stammt der Ausspruch, daß man die Güte eines menschlichen Rechners daran feststellen könne, wieviele überflüssige Schutzstellen er während der Rechnung mitführe. Die Verwendung der elektronischen Rechenautomaten hat die Frage nach der Genauigkeit einer Rechnung erheblich verändert. Wir sind gezwungen, mit fester Wortlänge, also fester Ziffernzahl unserer Ergebnisse zu rechnen und können während des Ablaufs der Rechnung die Anzahl der Schutzstellen nicht an die praktischen Gegebenheiten anpassen. Man versucht diese Schwierigkeit zu überwinden, indem man „hinreichend viele" Schutzstellen mitführt, aber was ist „hinreichend viele"? Die Benutzung der *gerundeten Intervall-Arithmetik* hilft uns bei dieser Schwierigkeit. Dabei werden reelle Zahlen zu (kleinen) Intervallen aus Maschinenzahlen vergröbert, diese Intervalle nach den Gesetzen der Intervall-Arithmetik miteinander verknüpft und die Ergebnisse wieder nach außen gerundet. Das Ergebnis ist „exakt", d. h. es enthält den wahren (im allgemeinen unbekannten) Wert. In ungünstigen Fällen, wenn eine Akkumulierung von Rundungsfehlern auftritt, können allerdings die erzielten Schranken recht pessimistisch sein.

Es lassen sich hinreichende Bedingungen dafür angeben, wann die Intervall-Abbildung optimal ist (dafür wurde der Terminus „schrankentreu" geprägt). Es kann jedoch nicht erwartet werden, daß diese Bedingungen für jede Funktion erfüllt wird. Beispiele für beide Verhaltensweisen, d. h. Schrankentreue und „Nicht-Optimalität" werden im folgenden wiederholt angegeben werden.

Es sei l die „Zahlenlänge", d. h. die Anzahl der Ziffern der Mantisse des Computers (gleichzeitig sei dafür gesorgt, daß die Schranken des Exponenten für $l \to \infty$ „passend" gegen ∞ gehen). Dann wird für das folgende angenommen, daß jeder reellen Zahl x eine „zugehörige" reell gerundete l-ziffrige Zahl $\tilde{x}(l)$ und ein gerundetes Intervall $[\tilde{x}(l)]$ zugeordnet wird. Weiter soll gelten, daß:

$$x, \tilde{x}(l) \in [\tilde{x}(l)], \qquad \lim_{l\to\infty} \tilde{x}(l) = x, \qquad \lim_{l\to\infty} [\tilde{x}(l)] = x \tag{2}$$

ist. Bei der arithmetischen Verknüpfung von Intervallen auf dem Computer werde stets „nach außen" gerundet und für die Resultate der Verknüpfung gelte wieder (2). Ist dann $f(x)$ eine rationale Funktion, $\tilde{f}_l(\tilde{x}(l))$ die zugehörige reelle Computerapproximation und $\tilde{F}_l[\tilde{x}(l)]$ eine entsprechende Computer-Intervall-Approximation, dann gilt durch Rekursion für alle x

$$f(x), \tilde{f}_l(\tilde{x}(l)) \in \tilde{F}_l[\tilde{x}(l)] \quad \text{und} \quad \lim_{l\to\infty} \tilde{F}_l[\tilde{x}(l)] = f(x).$$

Auf diese Weise lassen sich etwa die Verfahren der Nummer 2 auf den Computer übertragen, ohne daß die Schranken-Eigenschaft verloren geht. Außerdem werden die Ergebnisse für $l \to \infty$ „beliebig genau".

I.3.2. Intervall-Programmier-Sprachen

Schon sehr früh wurden die oben geschilderten Ideen einer gerundeten Intervall-Arithmetik durch die Definition einer allgemeinen Computersprache realisiert. Es ist dies die Erweiterung von ALGOL 60 zu Triplex-ALGOL 60, siehe WIPPERMANN et al. (1968). Es wurden mehrere Compiler erstellt, davon allein 3 in Karlsruhe (WIPPERMANN (1967), BROCKHAUS et al. (1969), ROTHMAIER (1974)). Eine Sprache Interval-FORTRAN ist in Vorbereitung (Madison/Wisc., USA) ein UNIVAC 1108-Compiler in Erprobung (BÖHMER und JACKSON (1977)).

Der große Vorteil dieser Sprachen ist, daß die Berücksichtigung der Rundungsfehler vollautomatisch durch den Computer vorgenommen wird und keinerlei Anstrengung durch den menschlichen Benutzer erfordert! Zwei einfache Beispiele werden in 3.5. und 6.1 angegeben.

I.3.3. Logik der Programmierung

Eine weitere wichtige Anwendung der Intervall-Mathematik finden wir in der Logik der Programmierung. Ein wesentlicher Bestandteil jedes Programms sind *Entscheidungen*, die zu Verzweigungen des Programms führen. Theoretisch genügt es dabei, sich auf eine Ja-Nein-Entscheidung zu beschränken. Dieses entspricht einer zweiwertigen Logik (tertium non datur). Bei der praktischen Berechnung jedoch sind Fehlentscheidungen zu befürchten. Sie kommen von den unvermeidlichen (Daten-, Rundungs-, ...) Fehlern. Um garantierte Entscheidungen treffen zu können, ist daher eine dreiwertige Logik erforderlich (wahr, falsch, unbestimmbar). Schreibt man Programme in dieser Art, so kann man garantieren, daß trotz der unvermeidlichen Ungenauigkeiten genau diejenigen Wege im Strukturdiagramm durchlaufen werden, die theoretisch gewünscht werden. Man vergleiche dazu Bild 5.

Ein krasses, wenn auch konstruiertes, Beispiel ist in Bild 6 wiedergegeben. Der dort definierte Funktionswert hat theoretisch den Wert 0. Bei endlicher Zahlenlänge innerhalb des Computers erhält man, falls bei der Division

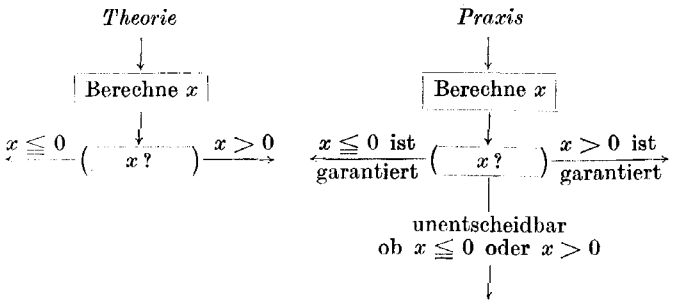


Bild 5. Logische Entscheidungen

Zu berechnen: $x := \begin{cases} 1 & \text{falls } 1 - (1/3) \cdot 3 > 0, \\ 0 & \text{sonst.} \end{cases}$

Theorie: $1 - (1/3) \cdot 3 = 0 \not> 0$

$\Rightarrow \boxed{x = 0}$

Gerundetes Rechnen (abgerundet):

$1 - (1/3) \cdot 3 = 1 - (0.33 \ldots 3) \cdot 3$
$= 1 - 0.99 \ldots 9$
$= 0.00 \ldots 01 > 0$

$\Rightarrow \boxed{\tilde{x} = 1}$

Intervall-Rechnung (gerundet):

$1 - (1/3) \cdot 3 = 1 - [0.3 \ldots 3,\ 0.3 \ldots 34] \cdot 3$
$= 1 - [0.9 \ldots 9,\ 1.0 \ldots 02]$
$= [-0.0 \ldots 02, +0.0 \ldots 01] \not> 0$

$\Rightarrow \boxed{[\tilde{x}] = 0}$

Bild 6. Numerische Konvergenz

abgerundet wird, stets den Wert 1, *und zwar für beliebige Ziffernanzahlen innerhalb der Maschine*! Verwendet man dagegen für die Zwischenergebnisse Intervall-Arithmetik (obwohl das Ergebnis eine ganze Zahl sein soll), so erhält man stets das richtige Ergebnis 0(!).

Es ist möglich, dieses Ergebnis zu einer Methode und zu einer Theorie zu verallgemeinern. In Untersuchungen von BIERBAUM (1975) wurde gezeigt, daß jedes numerische endliche Programm unter Benutzung der Intervall-Arithmetik derart umgeschrieben werden kann, daß es „numerisch konvergiert". Damit soll ausgedrückt werden, daß asymptotisch mit steigender Ziffernlänge das gesuchte Ergebnis schließlich beliebig genau approximiert wird gemäß folgender

Definition (numerische Konvergenz): Das theoretische Ergebnis eines numerischen Algorithmus sei $\hat{x}$. Der praktisch erzielte reelle bzw. Intervall-Wert sei $\tilde{x}(l)$ bzw. $[\tilde{x}(l)]$. Der Algorithmus heißt dann *numerisch konvergent*, wenn

$$\lim_{l \to \infty} \tilde{x}(l) = \hat{x} \quad \text{bzw.} \quad \lim_{l \to \infty} [\tilde{x}(l)] = \hat{x}\,.$$

Die Intervall-Version einer rationalen Funktion ist offenbar numerisch konvergent.

I.3.4. Abbrech-Kriterien

Die meisten (theoretischen) numerischen Verfahren führen auf eine unendliche Iteration. Es ist oft nicht einfach zu entscheiden, an welcher Stelle diese Iteration abzubrechen ist. Die Verwendung der gerundeten Intervall-Rechnung führt zu einem *universell gültigen* und *leicht programmierbaren* Abbrechkriterium:

Es sei $\hat{x} := \lim_{\nu \to \infty} x_\nu$ *zu berechnen.*

Man bestimmt (gerundete) *Intervalle* $[\tilde{x}_\nu(l)]$ *so, daß* $\hat{x} \in [\tilde{x}_\nu(l)]$, $\lim_{l \to \infty} [\tilde{x}_\nu(l)] = [x_\nu]$ mit $x_\nu \in [x_\nu]$, $[x_{\nu+1}] \subseteqq [x_\nu]$ und $\lim_{\nu \to \infty} [x_\nu] = \hat{x}$ *ist. Dann ist* $n = n(l)$ *mit*

$$[\tilde{x}_{\nu+1}(l)] \subseteqq [\tilde{x}_\nu(l)] \quad \text{für} \quad \nu = 0(1)\, n-1\,, \qquad [\tilde{x}_{n+1}(l)] \not\subseteqq [\tilde{x}_n(l)]$$

ein optimaler Abbrechindex, und es gilt numerische Konvergenz, d.h.

$$\lim_{l \to \infty} [\tilde{x}_{n(l)}(l)] = \hat{x}\,.$$

Man vergleiche dazu NICKEL (1975c).

Die entsprechenden Sätze für reelle Abbrechkriterien sind viel speziellerer Art und bei weitem nicht so universal.

I.3.5. Beispiel: Numerische Differentiation

Zu Recht wird die numerische Differentiation einer Funktion $f(t)$ als schwierig angesehen, die man am besten vermeidet. Mit Hilfe der gerundeten Intervall-Arithmetik, dem oben definierten Abbrechkriterium und der Theorie der numerischen Konvergenz lassen sich jedoch beliebig genaue, numerisch stabile Algorithmen angeben. Der unten in Triplex-ALGOL 60 geschriebene Algorithmus benutzt die einfachste Methode der vorwärts genommenen Differenzenquotienten zur Berechnung von $\hat{x} := f'(0)$. Ist $f(t)$ stabil bei $t = 0$ und wird die zugehörige Intervall-Näherung $\tilde{F}_l$: $[0, 1] \Rightarrow I(\mathbb{R})$ mit l Ziffern berechnet, dann wird danach die Intervall-Einschließung $\hat{x} \in [\tilde{x}(l)]$ mit $\cong l/2$ Ziffern bestimmt.

Numerische Differentiation

Gegeben: $f \in C^2[0, 1]$, $[u]$ mit $f''(t) \in 2[u]$ für $0 \leqq t \leqq 1$. Die Funktion $f(t)$ soll durch eine triplex procedure f abrufbar sein.

Ergebnis: Der Aufruf des Unterprogramms f prime (f, u) erzeugt ein optimales Intervall, das $f'(0)$ enthält.

Methode: Vorwärtsgenommener Differenzenquotient.

Programm:

```
triplex procedure f prime (f, u):
        value u;
        triplex u;
        triplex procedure f;
        begin real h;
                triplex th, x, xold, f0;
                h := 1;
                f0 := f(0);
                x := f(1) − f(0) − u;
        label: h := h/2;
                if h = 0 then goto fin;
                th := compose (h, h, h);
                xold := x;
                x := intsct (xold, (f(h) − f0)/th − th × u);
                if x ≠ xold then goto label;
        fin: f prime := x
        end;
```

Zu diesem Program wird ein Unterprogramm intsct (x, y) benötigt, das den Durchschnitt $x \cap y$ zweier Intervalle x und y mit gemeinsamen Punkten liefert. Eine mögliche Realisierung ist:

```
triplex procedure intsct (x, y);
        value x, y;
        triplex x, y;
        begin real a, b, c, d;
                c := inf (x);
                d := inf (y);
                a := if c > d then c else d;
                c := sup (x);
                d := sup (y);
                b := if c > d then d else c;
                intsct := compose (a, (a + b)/2, b)
end;
```

II. Anwendungen

II.4. Lineare Gleichungssysteme

II.4.1. Reelle Gleichungssysteme

Das System werde geschrieben in der Gestalt

$$Ax = b\,.$$

Der Einfachheit halber sei $A = (a_{ik})$ eine $n \times n$-Matrix mit $\det(A) \neq 0$, weiter sei $b = (b_k)$.

Die eindeutig bestimmte Lösung sei $\hat{x}$. Das Gausssche Eliminations-Verfahren ist eine stückweise (wegen der Pivotsuche) rationale Funktion in den Daten a_{ik} und b_k. Mit Hilfe der gerundeten Intervall-Arithmetik ist damit ein zugehöriges Intervall-Gauss-Verfahren sofort möglich. Auf einer l-ziffrigen Rechenmaschine laute das Resultat $[\tilde{x}(l)] = ([\underline{x}_k(l), \bar{x}_k(l)])$.

Eigenschaften *des Intervall-Gauß-Eliminations-Verfahrens*:

a) $\hat{x} \in [\tilde{x}(l)]$ (Fehler-Einschrankung), (3)

b) $\lim\limits_{l\to\infty} [\tilde{x}(l)] = \hat{x}$ (numerische Konvergenz). (4)

c) Die Suche nach einem größten Pivot (Spalten-, Zeilen-, totale Pivot-Suche) ist hier überflüssig, im Gegensatz zum Reellen. Es genügt, wenn die Null nicht in dem (Intervall-) Pivot enthalten ist! (Wongwises (1975))

d) Der Rechenaufwand (Anzahl der benötigten Operationen) ist genau gleich groß wie im Reellen.

Nachteile *des Intervall-Gauß-Eliminations-Verfahrens*:

a) Es sei $\tilde{x}(l) = (\tilde{x}_k(l))$ das reelle Computerergebnis des reellen Gauss-Verfahrens,

$$||\tilde{x}(l) - \hat{x}|| := \max_{k=1(1)\,n} |\tilde{x}_k(l) - \hat{x}_k|$$

und

$$\operatorname{span}[\tilde{x}(l)] := \max_{k=1(1)n} (\bar{x}_k(l) - \underline{x}_k(l))\,.$$

Dann zeigt die experimentelle Erfahrung nach Wongwises (1975) ein Verhalten wie

$$\operatorname{span}[\tilde{x}(l)]/||\tilde{x}(l) - \hat{x}|| \cong 10^{n/3}\,. \tag{5}$$

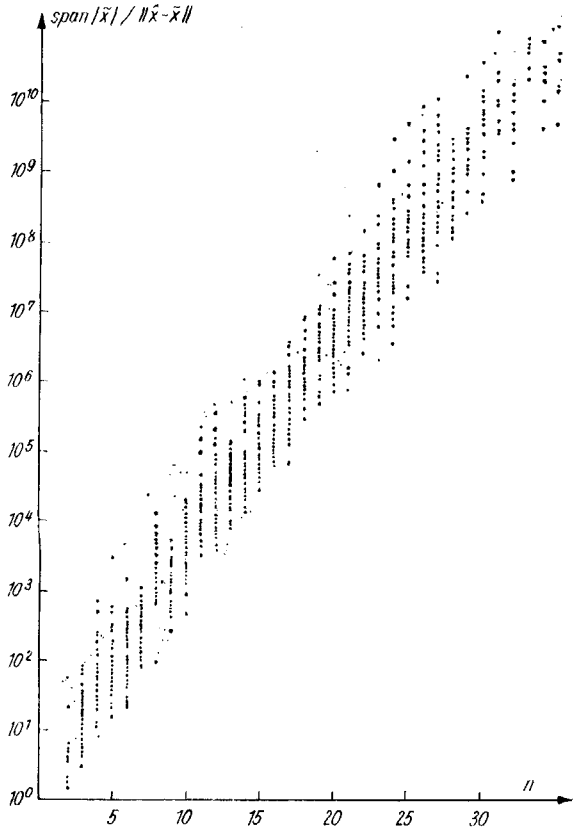


Bild 7. Experimentelle Ergebnisse des Intervall-Gauss-Eliminations-Verfahrens: Fehler-Überschätzung durch die Schranken, nach Wongwises (1975)

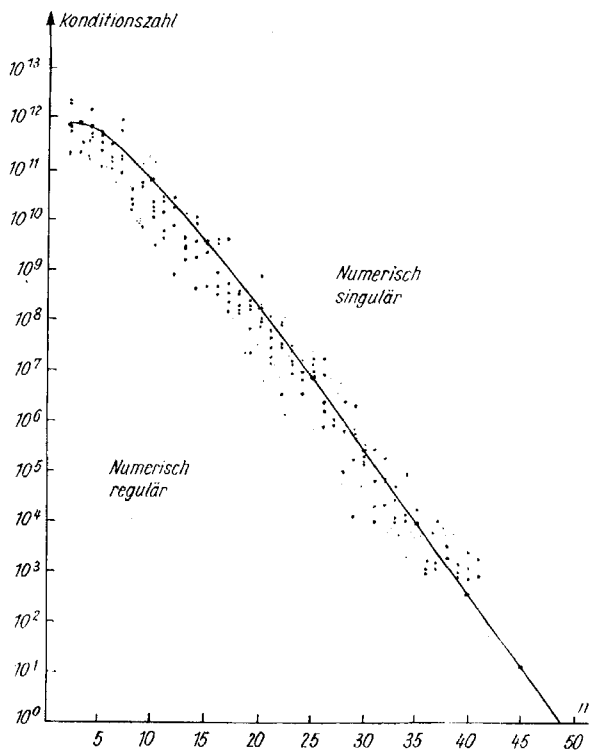


Bild 8. Experimentelle Ergebnisse des Intervall-Gauss-Eliminations-Verfahrens: Bereiche Numerischer Stabilität und Instabilität, nach Wongwises (1975) für Rechenanlage X8.

Dieses Verhalten ist unabhängig von der Konditionszahl cond (A) und von der Ziffernanzahl l (siehe Bild 7)! Die Ursache für dieses ungünstige Verhalten liegt darin, daß $\mathbb{I}(\mathbb{R})$ kein Körper ist. Das Verhalten (5) wurde von Wongwises auch theoretisch begründet, dabei wurde die Wilkinsonsche Theorie benützt.

b) Bei größeren Konditionszahlen cond (A) und größeren Gleichungssystemen ist i. a. kein Pivot-Element mehr auffindbar, das die 0 nicht enthält. Das Verfahren bricht dann zusammen („numerisch singulär" in Bild 8 nach Wongwises). Auch dieses unerfreuliche Verhalten läßt sich theoretisch vorhersagen und quantitativ bestätigen (siehe Wongwises (1975)).

Zur Angabe besserer Schranken wurden in den letzten zehn Jahren viele verschiedene Verfahren angegeben. Alle sind Iterationsverfahren und besitzen die gewünschten Eigenschaften (3) und (4), d. h. sie liefern numerisch konvergente Schranken. In allen Fällen ist eine Näherungsberechnung der Inversen erforderlich (N. B.: Im Reellen gilt dies als ein Kunstfehler!). Man vergleiche etwa Hansen (1965) und Hansen-Smith (1967). Das nach den numerischen Erfahrungen wohl erfolgreichste Verfahren ist das von Krawczyk (1969). Im nächsten Bild 9 sind typische Erfahrungen aus numerischen Experimenten von Wongwises aufgezeichnet. Das pessimistische Verhalten nach Gleichung (5) ist jetzt verschwunden. Im Durchschnitt sind etwa zwei Iterationsschritte erforderlich, das bedeutet etwa 6-mal mehr Rechenoperationen als im Reellen. Als Ausgleich für diesen größeren Rechenaufwand erhält

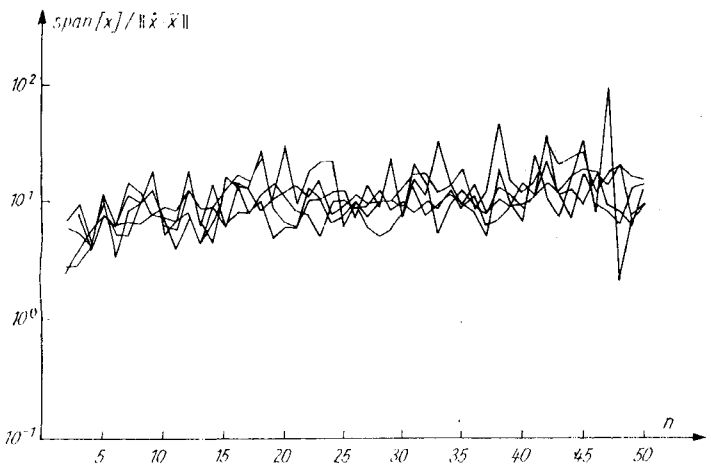


Bild 9. Experimentelle Ergebnisse des Krawczykschen Verfahrens: Fehler-Überschätzung durch die Schranken, nach Wongwises (1975)

man dafür jedoch das Ergebnis: *die oben beschriebenen Nachteile sind vollständig verschwunden.* Damit kann man das Problem der Lösungen eines reellen Gleichungssystems mit Hilfe intervall-arithmetischer Verfahren als vollständig gelöst betrachten.

II.4.2. Intervall-Gleichungssysteme

GAUSS hat viele Jahre seines Lebens der Geodäsie gewidmet, sowohl theoretisch als auch praktisch. Bei der Landmesserei kommt man ganz automatisch auf lineare Gleichungssysteme. Nun sind die Meßwerte jedoch niemals ganz exakt (zu GAUSS' Zeiten maß man Winkel, heute bevorzugt man Streckenmessungen). Wollen wir daher garantierte Ergebnisse, so müssen wir die bei den Messungen gemachten Fehler abschätzen und anschließend Intervall-Gleichungssysteme lösen. Das sind Gleichungssysteme, bei denen die Koeffizienten reelle Intervalle sind. Ich benutze die Schreibweise

$$[A]\,x = [b] \quad \text{mit} \quad [A] = ([\underline{a}_{ik}, \bar{a}_{ik}]), \qquad [b] = ([\underline{b}_k, \bar{b}_k]). \tag{6}$$

Die *Menge aller Lösungen* des Gleichungssystems (6) ist

$$\{\hat{x}\} := \{x \in \mathbb{R}^n \mid Ax = b, A \in [A], b \in [b]\}.$$

Erste Sätze über $\{\hat{x}\}$ stammen von OETTLI-PRAGER (1964), OETTLI (1965) und OETTLI-PRAGER-WILKINSON (1965).

Wie schon einfache Beispiele zeigen, ist diese Lösungsmenge $\{\hat{x}\}$ i. a. kein Intervall und nur recht kompliziert zu beschreiben. Man vergleiche dazu etwa das Beispiel von BARTH-NUDING (1974) vom nächsten Bild 10. Man ist daher an möglichst einfach zu beschreibenden Schranken interessiert, nämlich an der „Intervall-Hülle" oder „optimalen Intervall-Einschließung" $[\hat{x}] := [\inf\{\hat{x}\}, \sup\{\hat{x}\}]$, man vergleiche dazu das Bild von BARTH-NUDING. Schon einfache Beispiele zeigen, daß die Intervall-Version des GAUSSschen Eliminations-Verfahrens diese Intervall-Hülle $[\hat{x}]$ i. a. nicht liefert. Eine hinreichende Bedingung gibt der folgende

Satz (BARTH-BEECK-NUDING): *Jede Matrix $A \in [A]$ sei eine M-Matrix. Es gelte $[b] \geqq 0$ oder $[b] \leqq 0$ oder $0 \in [b]$. Dann liefert das Intervall-Gauß-Verfahren ohne Pivotisierung die Intervall-Hülle $[\hat{x}]$ des Systems* (6).

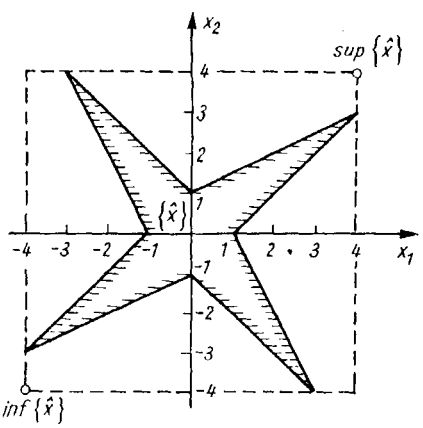


Bild 10. Lösungsmenge $\{\hat{x}\}$ und optimale Intervalleinschließung $[\hat{x}]$ für das System $\begin{pmatrix}[2,4] & [-2,1]\\ [-1,2] & [2,4]\end{pmatrix} x = \begin{pmatrix}[-2,2]\\ [-2,2]\end{pmatrix}$ nach BARTH-NUDING (1974)

(Man vergleiche BARTH-NUDING (1974), BEECK (1974)). M-Matrizen spielen eine große Rolle in den Anwendungen, z. B. bei der Diskretisierung von Randwertproblemen. Sehr viele praktisch vorkommende Intervall-Gleichungssysteme besitzen jedoch (leider) keine M-Intervall-Matrizen. In diesen Fällen kann jedoch immer noch das Iterationsverfahren angewendet werden. Es gelten die beiden folgenden Sätze über die Lösung eines linearen Intervall-Gleichungssystems in der iterativen Gestalt

$$x = [M]\,x + [r]. \tag{7}$$

Satz (O. MAYER): *Das Intervall-Iterationsverfahren*

$$[x_{\nu+1}] := [M]\,[x_\nu] + [r] \tag{8}$$

konvergiert für einen beliebigen Intervall-Anfangsvektor $[x_0]$ genau dann gegen die eindeutig bestimmte Lösung $[\hat{y}]$ des Systems

$$[y] = \langle M \rangle\,[y] + [r], \tag{9}$$

wenn der Spektralradius $\varrho(|\langle M \rangle|) < 1$ ist.

Dabei ist $|\langle M \rangle|$ die reelle Matrix mit dem Komponenten $\max(|\underline{m}_{ik}|, |\bar{m}_{ik}|)$, siehe O. MAYER (1968). Man beachte, daß (9) ein System ist, dessen Lösung ein Intervall-Vektor sein soll. Im Gegensatz dazu liefert der folgende Satz von BARTH (1972b) die Intervall-Hülle der Lösungsmenge für das System (7).

Satz (BARTH): *Es sei $\langle M \rangle = \langle \underline{M}, \bar{M} \rangle \geqq 0$ und $\varrho(\bar{M}) < 1$. Dann konvergiert das Intervall-Iterationsverfahren* (8) *gegen die Intervall-Hülle des Systems* (7).

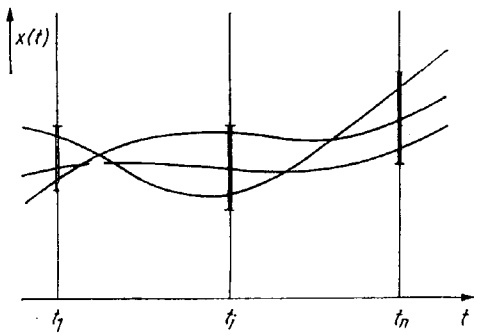
Das allgemeine Problem, die Intervall-Hülle des linearen Intervall-Gleichungssystems (6) oder (7) mit erträglichem Rechenaufwand zu bestimmen, wenn keine weiteren Informationen gegeben sind, ist bis heute noch ungelöst!

II. 5. Approximation

Die von GAUSS entwickelte Methode der kleinsten (Defekt-) Quadrate bestand ihre Feuertaufe im Jahre 1801 mit der spektakulären Wiederentdeckung des verloren gegangenen Asteroiden Ceres. Die Ceres war nur über einen kleinen Bereich ihrer Wanderung beobachtet worden und verschwand dann hinter der Sonne. Es galt, aus diesen wenigen Meßwerten die Konstanten ihrer KEPLER-Ellipse und damit ihre Bahn zu bestimmen. In den vergangenen fast zwei Jahrhunderten wurde die Approximationstheorie sehr stark weiterentwickelt und ist heute ein unentbehrliches Hilfsmittel der Mathematik.

Geht man statt von Meßwerten von Meßintervallen aus, (siehe Bild 11), dann kann man im Rahmen der Intervall-Mathematik eine Intervall-Approximationsaufgabe stellen:

Gegeben seien n Meßintervalle. Gegeben sei weiterhin eine Gesetzmäßigkeit für die gesuchte Funktion $x(t)$ (etwa die Schar aller KEPLER-Ellipsen, -Parabeln und -Hyperbeln). Gesucht sind dann entweder Schranken für die Parameter der Lösungsfamilie oder aber ein garantierter Lösungsstreifen (siehe Bild 12). Als durchaus (erwünschtes) Nebenergebnis kann dabei möglicherweise eine Verbesserung der Meßintervalle erzielt werden (siehe Bild 12)!



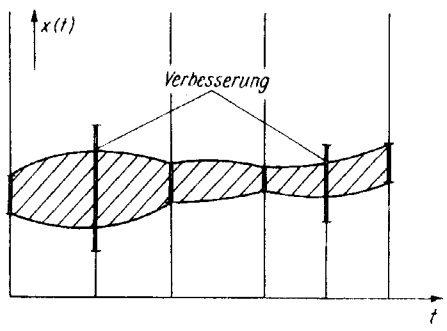


Bild 12. Garantiertes Funktionsintervall mit Verbesserung von Meßwerten

gegeben: Meßintervalle $[y_i]$ für $i = 1(1)n$,
Gesetz $x(t; \alpha, \beta, \ldots, \sigma)$.

gewünscht: $x(t; \alpha, \beta, \ldots, \sigma) \in [y_i]$ für $i = 1(1)n$.

gesucht: Entweder Schranken für $\alpha, \beta, \ldots, \sigma$,
oder Schranken für $x(t)$ für alle t.

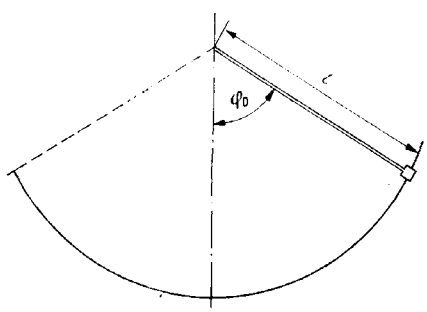
Bild 11. Intervall-Approximation

Die bisher aufgestellten reellen Approximationsmethoden sind i. a. offensichtlich nicht geeignet, dieses Problem der Intervall-Approximation zu lösen. In den letzten Jahren wurden spezielle Intervall-Approximations-Methoden entwickelt. Im Gegensatz zum „Reellen" sind wir jedoch noch weit von der vollständigen Lösung dieses Problems entfernt.

II. 6. Differentialgleichungen

II.6.1. Beispiel: Das mathematische Pendel

Ein großer Teil der Angewandten Mathematik besteht in der Lösung von Differentialgleichungen aus Mechanik, Physik, Chemie, Biologie, Volkswirtschaft, usf. Ich möchte die Anwendung der Intervall-Mathematik bei der Lösung von Differentialgleichungen an einigen Beispielen erläutern: Beim *mathematischen Pendel* ist die Lösung wohlbekannt und läßt sich durch elliptische Funktionen und Integrale darstellen (siehe Bild 13). Die darin auftre-



Schwingungsdauer T:

$$T = 4\sqrt{\frac{l}{g}} \int_0^{\pi/2} \frac{dt}{\sqrt{1 - \sin^2 \varphi_0/2 \sin^2 t}}$$

$$\varphi_0 \ll \pi/2 : T \approx 2\pi\sqrt{\frac{l}{g}}$$

Bild 13. Mathematisches Pendel

$$K(\sin\varphi_0/2) := \int_0^{\pi/2} \frac{dt}{\sqrt{1 - \sin^2\varphi_0/2 \sin^2 t}}.$$

$a_0 := \cos\varphi_0/2; \quad b_0 := 1;$

$a_{\nu+1} := \sqrt{a_\nu \cdot b_\nu}; \quad b_{\nu+1} := \frac{a_\nu + b_\nu}{2}$ für $\nu = 0, 1, \ldots$

Es gilt $a_\nu \nearrow c$, $b_\nu \searrow c$ und $K(\sin\varphi_0/2) = \frac{\pi}{2 \cdot c}$.

Bild 14. Gausssche Methode der geometrisch-arithmetischen Mittel

ν	a_ν	b_ν
0	0.250000000000 ...	1.000000000000 ...
1	0.500000000000 ...	0.625000000000 ...
2	0.559016994374 ...	0.562500000000 ...
3	0.560755792957 ...	0.560758497187 ...
4	0.560757145071 ...	0.560757145072 ...

$c = 0.560757145\,07\ldots$

$K(\sqrt{15}/4) = 2.801206082\,89\ldots$

Bild 15. Numerisches Beispiel für die Gausssche Methode der geometrisch-arithmetischen Mittel

tenden beiden Konstanten l (Länge des Pendelarms) und g (Gravitationsbeschleunigung) sind jedoch nicht exakt bekannt, sondern gemessen. Im günstigsten Fall lassen sie sich in Intervalle einschachteln. Die Auswirkung dieser Ungenauigkeit etwa auf die Schwingungsdauer läßt sich sofort durch elementare Intervallarithmetik berechnen. Man kann zeigen, daß man damit die optimalen Schranken erhält.

Das in der Schwingungsdauer auftretende elliptische Integral erster Gattung läßt sich nicht geschlossen durch elementare Funktionen darstellen, muß also durch ein passendes numerisches Verfahren berechnet werden. Die wohl günstigste Methode dafür ist die Gausssche Methode der geometrisch-arithmetischen Mittel (siehe Bild 14). Diese Methode ist global quadratisch konvergent und liefert simultan untere und obere Schranken für den Wert des elliptischen Integrals, man vergleiche das Zahlenbeispiel in Bild 15. Es ist jedoch zu beachten, daß diese Methode keine Kontraktion darstellt; durch Rundungsfehler kann daher das Ergebnis „abwandern". Eine Rechnung mit Hilfe der gerundeten Intervall-Arithmetik verhindert diese Abwanderung zuverlässig. Außerdem erhält man noch gratis die Information, wann die Iteration am zweckmäßigsten abzubrechen ist, nämlich dann, wenn infolge der Rundungsfehler keine Verbesserung mehr stattfindet. Ein mögliches Programm in der modernen Programmiersprache Triplex-ALGOL 60 ist im folgenden dargestellt.

Gausssches Verfahren der geometrisch-arithmetischen Mittel

Gegeben: $0 \leq a < b$ reell.

Gesucht: $G(a, b) := \int_0^{\pi/2} \frac{dt}{\sqrt{a^2 \sin^2 t + b^2 \cos^2 t}}$

Methode: Mit $a' := \sqrt{a \cdot b}$, $b' := (a + b)/2$ gilt $G(a, b) = G(a', b')$. Man iteriert, bis $a' = b' = c$ ist.

Programm:

```
triplex procedure gauss (a, b);
    value a, b; real a, b;
    begin triplex ta, tb, v, u, g, gneu;
        u := compose (a, a, a);
        v := compose (b, b, b);
        gneu := compose (a, a, b);
    marke: ta := u; tb := v;
        g := gneu;
        u := sqrt (ta × tb);
        v := (ta + tb)/2;
        gneu := compose (inf (u), (inf (u) + sup (v))/2, sup (v));
        if span (gneu) < span (g) then goto marke;
        gauss := 2 × arctan (1)/g
    end;
```

II.6.2. Numerische Lösung von Differentialgleichungen

Eines der einfachsten Probleme der Numerischen Mathematik ist die näherungsweise Lösung des Anfangswertproblems bei einem System von gewöhnlichen Differentialgleichungen (auf Sonderprobleme soll nicht eingegangen werden, wie etwa auf steife Systeme, Schrittweitensteuerung ...). Man wählt eine Folge von „passenden" Schrittweiten und eine numerische Methode, wie etwa das Runge-Kutta-Verfahren. Die einzige Voraussetzung an die rechte Seite der Differentialgleichung ist dann, daß diese Funktion numerisch für jeden Punkt des Lösungsraums berechenbar sein soll.

Wegen dieser (scheinbaren) Einfachheit wird oft übersehen, daß das zugehörige Problem der Fehlerschrankenbestimmung außerordentlich schwierig ist! Zwar sind die zugrundeliegenden theoretischen Sätze aus der Theorie der Differential-Un-Gleichungen schon seit langem bekannt (man vgl. etwa das Buch von W. Walter (1970)). Ihre praktische Realisierung ist jedoch so kompliziert, daß bis heute praktisch noch keine „reellen" Algorithmen zur Fehlererfassung bei Differentialgleichungen existieren. Seltsamerweise werden Intervall-Mathematiker sehr oft vorwurfsvoll gefragt, warum die Intervall-Mathematik nicht in der Lage wäre, „einfache" Algorithmen zur Lösung dieses Problems zur Verfügung zu stellen. Gelegentlich wird dies sogar als ein „Beweis" dafür angesehen, daß die Intervall-Numerik nicht in der Lage sei, die wesentlichen Grundprobleme zu lösen.

Dabei ist eine einfache Schrankenerfassung sehr leicht möglich: Man benutzt die Möglichkeit der numerischen Quadratur nach Kapitel 2.2. Damit läßt sich sofort das Verfahren von PICARD-LINDELÖF als Intervall-Verfahren formulieren. Es ist wichtig darauf hinzuweisen, daß die einfachste Version dieses Verfahrens garantierte Schranken liefert, ohne daß für die rechte Seite der Differentialgleichung mehr als die Existenz einer inklusionsisotonen Intervallerweiterung vorausgesetzt wird! Allerdings ist mit dieser Einfachheit nur lineare Konvergenz verknüpft. — Dasselbe gilt für ein entsprechendes Intervall-EULER-CAUCHY-Verfahren.

Mit höherem Aufwand lassen sich beliebig genaue Intervall-Algorithmen angeben. Ein besonders spektakuläres Verfahren ist das von MARCOWITZ (1975). Spätestens nach dem Erscheinen dieser Arbeit sind die oben erwähnten Vorwürfe gegenstandslos geworden. In seiner Arbeit hat Herr MARCOWITZ seinen Algorithmus auf das Problem des Wiedereintritts eines Raumflugkörpers in die Atmosphäre angewandt. Dieses Problem ist nicht „ausgesucht" derart, daß die Ergebnisse möglichst günstig werden. Fachleute sehen vielmehr an den Vorzeichen der Funktionalmatrizen, daß die rechte Seite der Differentialgleichung nicht quasi-isoton ist. Damit sind ungünstige Fehlerverhältnisse zu erwarten. Tatsächlich sind die berechneten Fehlerschranken jedoch außerordentlich günstig. Über weite Bereiche der Zeitvariablen hinweg ergeben sie sogar eine Verbesserung der ursprünglich berechneten reellen Näherungswerte. Auch ein Vergleich der Rechenzeiten zeigt, daß die Schrankenberechnung nur einen Bruchteil der Berechnung der Näherungslösung kostet. (Allerdings ist dieser Sachverhalt darauf zurückzuführen, daß es sich hier um ein Steuerungsproblem handelt).

Es muß jedoch ausdrücklich darauf hingewiesen werden, daß das von MARCOWITZ behandelte Problem in dem Sinne vereinfacht ist, daß die physikalischen Daten durch reelle Zahlen ersetzt wurden. Die berechneten Fehlerschranken geben also allein den Einfluß der Rundungsfehler wieder. Man ist damit sicher, daß bei einer nochmaligen Durchrechnung mit Datenintervallen jede Aufblähung der Fehlerschranken auf die Eingangsdaten zurückzuführen ist und nicht etwa auf eine unzweckmäßig gewählte Intervall-Methode. (Meines Wissens ist allerdings diese nochmalige Nachrechnung bisher noch nicht erfolgt.)

Damit kann das Problem der Erfassung der Rundungsfehler bei der Berechnung von Anfangswertaufgaben bei Systemen gewöhnlicher reeller Differentialgleichungen im Prinzip als erledigt gelten.

11.6.3. Intervall-Differential-Gleichungen

Ähnlich wie bei linearen Gleichungssystemen ist auch hier der Übergang von reellen Daten zu (großen) Daten-Intervallen nicht einfach und bringt spezielle Probleme. Im nächsten Bild 16 ist unter a) schematisch die Menge der Lösungen $\{\hat{x}\}$ eines Anfangswert-Problems skizziert. Diese Menge ist im allgemeinen so komplex, daß sie sich nicht auf einfache Weise beschreiben läßt. Die einfachste Einschränkung geschieht durch ein Intervall $[\hat{x}]$, die „Intervall-Hülle" oder „optimale-Intervall-Einschränkung". Im allgemeinen ist — genauso wie bei linearen Gleichungssystemen — auch diese Intervall-Hülle $[\hat{x}]$ nur schwer bestimmbar. Die Methode der Wahl ist dann ein Iterationsverfahren. In speziellen Fällen läßt sich jedoch die Intervall-Hülle $[\hat{x}]$ a priori angeben, nämlich dann, wenn die beiden Intervallschranken $\inf[\hat{x}]$ und $\sup[\hat{x}]$ Elemente der Lösungsmenge $\{\hat{x}\}$ sind und sich a priori angeben lassen, siehe die Skizze in Bild 16b). Dies gilt etwa unter den Voraussetzungen des folgenden allgemeinen Satzes:

Satz: *Zu lösen sei das System von Intervall-Anfangswertproblemen*

$$x'(t) \in [f(t, x(t))], \qquad x(0) \in [\alpha] .$$

Dabei sei $[\alpha]$ *ein Intervallvektor; in dem Funktions-Intervall-Vektor* $[f] = [\underline{f}, \bar{f}]$ *seien die beiden Schrankenfunktionen* $\underline{f}$ *und* $\bar{f}$ *stetig und quasiisoton. Man bestimmt* $\underline{x}(t)$ *als Minimalintegral von* $x' = \underline{f}(t, x)$, $x(0) = \underline{\alpha}$ *und* $\bar{x}(t)$ *als Maximalintegral von* $x' = \bar{f}(t, x)$, $x(0) = \bar{\alpha}$. *Dann gilt für die Lösungsmenge*

$$\underline{x}(t), \bar{x}(t) \in \{\hat{x}(t)\} \subseteq [\underline{x}(t), \bar{x}(t)] .$$

Es ist nun außerordentlich erfreulich, daß dieser günstige Sachverhalt sich auch bei sehr vielen anderen Problemen nachweisen läßt, nämlich bei elliptischen, parabolischen, hyperbolischen partiellen Differentialgleichungen etwa aus der Grenzschichttheorie, bei VOLTERRAschen Integralgleichungen der Biologie und so fort. Die Intervall-Hülle läßt sich, kurz gesagt, sicher dann a priori angeben, wenn die bekannten Voraussetzungen aus der Theorie der Differential- und Integral-Un-Gleichungen erfüllt sind derart, daß die beiden Schrankenprobleme „von monotoner Art" (d. h. invers-isoton) sind. — Sind diese Voraussetzungen jedoch nicht zu befriedigen, dann kann man bei einer sehr großen Klasse von Problemen immer noch mit dem Iterationsverfahren vermöge der eingangs geschilderten Fix-Intervall-Gleichungen trotzdem noch zur Intervall-Hülle $[\hat{x}]$ gelangen.

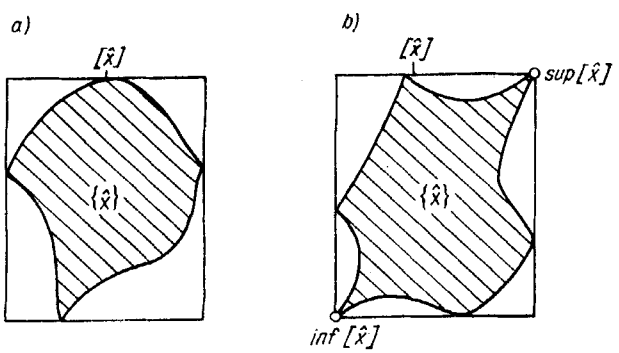


Bild 16. Lösungsmenge $\{\hat{x}\}$ und Intervallhülle $[\hat{x}]$

III. Ausblick

III. 7. Anwendbarkeit der Intervall-Mathematik

Zurückblickend auf die angegebenen Beispiele läßt sich jetzt die Frage beantworten: „In welchen Bereichen der Angewandten Mathematik helfen uns die Begriffe und Methoden der Intervall-Mathematik ?"

1. Die Intervall-Mathematik hilft uns, Fehler unter Kontrolle zu halten. „Fehler" sind dabei: Daten-Fehler, Darstellungs-Fehler, Rundungs-Fehler, Abbrech-Fehler, Verfahrens-Fehler, ... Alle diese verschiedenen Fehlertypen sind jetzt durch eine einheitliche Methode erfaßbar geworden.

2. Die Logik der Programmstruktur wird gesichert oder sogar überhaupt erst in Ordnung gebracht. Das geschieht mit Hilfe einer dreiwertigen Logik mit den logischen Werten: (garantiert) richtig, (garantiert) falsch, unentscheidbar ob richtig oder falsch.

3. Probleme der Angewandten Mathematik haben für gewöhnlich nicht nur eine einzige Lösung, sondern eine ganze Lösungsmenge (verursacht durch Datenungenauigkeiten, Mehrdeutigkeiten, Steuerungsparameter etc.). Die genaue Gestalt dieser Lösungsmenge ist meist (wenn überhaupt) nur außerordentlich schwierig zu bestimmen. Die Intervall-Mathematik liefert uns in vielen Fällen auf einfache Weise (oft optimale) *Schranken für die Lösungsmenge*.

4. Die Intervall-Mathematik liefert uns viele neue numerische Verfahren, die nicht unbedingt „aus dem Reellen" ableitbar sind. Bei bekannten Verfahren wird häufig das Verhalten einfacher und/oder durchsichtiger (Beispiele: Intervall-NEWTON-Verfahren, Intervall-Einzelschritt-Gesamtschritt-Verfahren).

III. 8. Vorurteile

Seit Bestehen der Intervall-Mathematik gibt es zwei Vorurteile. Sie wurden von Mathematikern erfunden, die einmal schlechte Erfahrungen mit den Anwendungen der Intervall-Mathematik gemacht haben, oder die glaubten, sie gemacht zu haben. In der Zwischenzeit wurden diese Vorurteile dann ohne jede Prüfung weitergegeben vom Institutsleiter an den Assistenten und die Studenten. Es handelt sich um die beiden folgenden angeblichen „Fakten":

1. *Die Fehlerschranken, die von der Intervall-Mathematik geliefert werden, sind immer viel zu pessimistisch!* Die numerische Erfahrung von Dutzenden von Mathematikern an hunderten von Problemen zeigte, daß diese Behauptung falsch ist. In der überwiegenden Zahl der Fälle liefert die Intervall-Mathematik sehr „günstige" Schranken, sehr häufig sind sie sogar optimal.

In einigen Fällen treten jedoch tatsächlich pessimistische Fehlerschranken auf, die Anwendung des GAUSSschen Eliminationsverfahrens auf lineare Gleichungssysteme bei nicht-speziellen Matrizen ist ein besonders eklatantes (und unangenehmes) Beispiel dafür. Sämtliche derartigen bisher bekannt gewordenen Fälle lassen sich jedoch als eine „naive" Anwendung der Intervall-Mathematik deuten. In allen diesen Fällen gibt es „angepaßte" Methoden, die zu „optimalen" Fehlerschranken führen. Selbstverständlich gibt es viele Intervall-Probleme, für die bis heute eine „angepaßte" Lösungsmethode noch fehlt, wie etwa das Problem der Lösung von linearen Gleichungssystemen mit Intervall-Matrizen. In all diesen Fällen gibt es jedoch auch keine einfache „reelle" Lösungsmethode.

2. *Alle gängigen Intervallmethoden sind viel zu zeitaufwendig.* Auch diese Behauptung wird nicht durch die Erfahrung bestätigt. Insbesondere ist sie falsch für Intervall-Iterationsmethoden, die als Erweiterung einer reellen Iterationsmethode entstanden sind. Der Rechenaufwand pro Iterationsschritt dürfte hierbei i. a. nicht größer sein als im Reellen. Da man erheblich mehr über den Fehler weiß, kann man im allgemeinen an einer früheren und günstigeren Stelle die Iteration abbrechen. Erfahrungsgemäß laufen deshalb Iterations-Intervall-Methoden im allgemeinen sogar schneller (!) als die entsprechenden reellen Iterationen.

Dieses zweite Vorurteil kommt vermutlich daher, daß bis heute bei fast allen Rechenmaschinen die arithmetischen Intervalloperationen noch nicht hardwaremäßig verwirklicht sind und daher durch Software simuliert werden müssen. Je nachdem, wie ungünstig die verdrahteten arithmetischen Operationen des Computers sind, laufen die (simulierten) Intervalloperationen bis zu mehrere Hundert mal langsamer als die entsprechenden reellen Operationen. Dieser Nachteil hat jedoch überhaupt nichts mit der Intervall-Mathematik zu tun, sondern nur mit der Tatsache, daß unsere heutigen Computer an die Intervall-Arithmetik denkbar schlecht angepaßt sind.

Erlauben Sie mir daher an dieser Stelle einen

Aufruf an die Computerhersteller:

Bitte sorgen Sie bei allen Neuentwicklungen von Computern dafür, daß die Intervalloperationen hardwaremäßig ohne Zeitverlust durchgeführt werden können. Die Kosten dafür liegen bei den heutigen Preisen für Rechenwerke unter 1% der Gesamtsumme eines Großrechners. Bitte stellen Sie dem Benutzer weiterhin eine passende Programmiersprache zur Verfügung, wie etwa Triplex-ALGOL 60 oder Interval-FORTRAN oder dergleichen. — Die Kosten der nachträglichen Umrüstung eines Computers sowohl hardwaremäßig als auch softwaremäßig sind — nach unseren Erfahrungen in Karlsruhe — außerordentlich hoch. Nachdem nunmehr die theoretischen Grundlagen vorhanden sind, bereitet es jedoch bei der Neuplanung eines Computers und des zugehörigen Systems nur geringfügige Mehrkosten, die Intervall-Arithmetik von Anfang an mit einzubeziehen.

Trotz meiner Beteuerungen über die Vorteile der Intervall-Mathematik und der Intervall-Analysis wird es sicherlich noch genug „ungläubige Thomasse" geben. Ihnen schlage ich einen Test vor. Meine Vermutung ist: Niemand von Ihnen kann die beiden oben angegebenen Programme (numerische Differentiation und elliptisches Integral) im „Reellen" einfacher programmieren, als dies geschehen ist. Voraussetzung ist natürlich, daß eine der gängigen Programmiersprachen benutzt wird (wie ALGOL, SIMULA, FORTRAN, ...), daß die gleiche Methode verwendet wird (vorwärtsgenommener Differenzenquotient bzw. Methode der geometrisch-arithmetischen Mittel) und daß als Ergebnis garantierte Fehlerschranken erzeugt werden.

III. 9. Erweiterungen

In meinem Vortrag habe ich an einigen speziellen ausgewählten Beispielen aus den Anwendungen gezeigt, wie uns die Intervall-Mathematik bei typischen Anwendungsproblemen helfen kann. Das ist nicht weiter verwunderlich, denn die Intervall-Arithmetik und die Intervall-Analysis wurden ursprünglich genau zu diesem Zweck konzipiert. Sehr viel überraschender jedoch ist, daß die Intervall-Mathematik sich heute immer mehr in Bereiche der „reinen" Mathematik hineinfrißt. In den letzten Jahren ist der abstrakte, nichtnumerische Teil der Intervall-Mathematik immer stärker in den Vordergrund getreten. Es handelt sich hier um den auch sonst in der Mathematik üblichen Prozeß der Vereinheitlichung und Abstrahierung.

Zwei Beispiele dazu:

1. Es gibt zwei Darstellungsmöglichkeiten für Intervalle, entweder in einem halbgeordneten Raum oder in einem metrischen Raum. Im Zweidimensionalen, also etwa auf der Gaussschen Zahlenebene (mit komponentenweiser Ordnungsrelation und der üblichen Metrik), werden diese beiden Intervall-Darstellungen zu Rechtecken bzw. Kreisen. Welche algebraischen Eigenschaften sind nun diesen beiden Darstellungsformen (und anderen) gemeinsam? Von Otto Mayer (1968) und W. Hahn (1971) wurde zur Beantwortung dieser Frage der *quasilineare Raum* eingeführt. Es handelt sich dabei um einen linearen Raum, in dem jedoch das Distributivgesetz der Multiplikation mit Konstanten aus dem Grundkörper abgeschwächt ist. Man kann zeigen, daß die Räume der beiden oben angegebenen Intervall-Darstellungen durch Rechtecke und Kreise solche quasilinearen Räume sind. Weiter gilt in beiden noch die Kürzungsregel der Addition.

Es scheint so zu sein, als ob die Struktur eines quasilinearen Raumes mit Kürzungsregel (zumindest) eine adäquate algebraische Darstellung der Menge von Intervallen ist. Aus diesem Grunde werden solche Räume nunmehr „Intervall-Räume" genannt. Einen ausgezeichneten Überblick über dieses und viele andere Ergebnisse der Intervall-Algebra findet man in dem Artikel von Ratschek (1975).

Insbesondere gilt die erfreuliche Tatsache, daß die (passend definierten) Intervalle über einem Intervall-Raum selbst wieder einen Intervall-Raum bilden. Man kann auf diese Weise rekursiv aufsteigen. Diese Schachtelung von Intervall-Räumen ist nicht nur theoretisch interessant, sondern — wie jedoch nicht gezeigt werden soll — auch wichtig für die Praxis.

2. In einer soeben fertiggestellten aber noch unveröffentlichten Arbeit hat Herr K.-U. Jahn (1977) die von Herrn Klaua entwickelten Ideen einer dreiwertigen Logik auf die Intervall-Analysis und Intervall-Topologie angewandt. Die sehr große Anzahl neuer Ergebnisse kann hier noch nicht einmal referiert werden. Es scheint jedoch so zu sein, daß die Benutzung der dreiwertigen Logik einen vereinheitlichenden Faktor darstellt. Durch die Anwendung der „üblichen", d. h. in der Analysis und Topologie gebräuchlichen Denkschemata wird man dann nämlich zwangsläufig auf den Begriff des Intervall-Raums und des quasilinearen Raums geführt. Die Begriffe der Stetigkeit, Differenzierbarkeit, Integrierbarkeit von Intervall-Funktionen lassen sich jetzt ebenso folgerichtig mit Hilfe einer dreiwertigen Logik einführen und viele andere Ergebnisse mehr. Ein kleiner „Leckerbissen" ist die Tatsache, daß die beiden oben angegebenen Darstellungen eines Intervalls mit Hilfe einer Halbordnung oder mit Hilfe einer Metrik (zumindest im eindimensionalen Reellen) nichts anderes als zwei zueinander duale Darstellungen desselben Sachverhaltes sind. Durch die Jahnschen Überlegungen wird die reelle Arithmetik und Analysis isomorph und isometrisch eingebettet in die Intervall-Arithmetik und -Analysis.

In den letzten Jahren hat es sich weiter gezeigt, daß die Verbandstheorie eine wesentliche Rolle für die Intervall-Analysis spielt. Allein unter der Voraussetzung der Inklusionsisotonie — sogar über viel allgemeineren Räumen, etwa über bedingt vollständigen Verbänden — lassen sich Fixpunktsätze in großer Allgemeinheit und mit großer Fruchtbarkeit formulieren.

Ein sehr allgemeiner (allerdings nichtkonstruktiver) Fixpunktsatz ist etwa der von Knaster-Tarski: *Eine inklusionsisotone Intervall-Funktion, die ein Intervall in sich abbildet, besitzt* (mindestens) *einen Fixpunkt.* Es lassen sich auch in sehr allgemeiner Weise konstruktive Fixintervall-Sätze angeben, die auf dem Iterationsverfahren basieren. Es gibt sehr allgemeine hinreichende Bedingungen, unter denen das Intervall-Iterationsverfahren nicht nur gegen ein Fixintervall, sondern sogar gegen einen Fixpunkt konvergiert, der dann der einzige Fixpunkt der gegebenen Funktion ist. Diese Fixintervallsätze sind zum Teil Realisierungen oder Erweiterungen von bekannten reellen Fixpunktsätzen, zum Teil sind sie jedoch unabhängig vom Reellen. Man vergleiche dazu etwa Nickel (1975b).

Ein besonders schönes Beispiel dafür, daß für Intervall-Funktionen manche Aussagen einfacher, klarer und durchsichtiger werden als im Reellen, ist ein Satz von Wisskirchen (1975): *Für ein beliebiges nichtlineares Gleichungssystem, das nur aus inklusions-isotonen Funktionen bestehen soll, ist* (bei gleichen Anfangswerten) *das Einzelschrittverfahren stets besser konvergent als das Gesamtschrittverfahren. Beide Verfahren konvergieren, selbst unter dem Einfluß von Rundungsfehlern, zu demselben Endwert.*

Bei der Intervall-Mathematik haben wir es nicht mit einem Gegensatz zur „üblichen" Mathematik zu tun, es handelt sich vielmehr um eine Erweiterung. Heute sieht es noch gelegentlich so aus, als ob die Intervall-Mathematik etwas „anderes wäre" als die „übliche" Mathematik. Es ist meine Vermutung, daß dieser Gegensatz in 10 oder 20 Jahren verschwunden sein wird und daß dann die Definitionen, Methoden und Ergebnisse der Intervall-Mathematik in die „üblichen" Bereiche der Mathematik integriert und von ihnen absorbiert sein werden, nämlich von: Funktionalanalysis, Logik, Numerischer Mathematik, sowie selbstverständlich auch von den Anwendungen der Mathematik.

In den letzten $2^1/_2$ Jahrtausenden haben wir uns daran gewöhnt, der Mathematik die folgenden Attribute zuzuordnen:

Sauberkeit der Methode,
Sicherheit des Urteils,
Exaktheit der logischen Schlüsse,
Garantie der Ergebnisse.

Es muß zugegeben werden, daß wir diese Attribute sowohl der heutigen Angewandten Mathematik als auch der heutigen Numerischen Mathematik oft nicht zuweisen können. Die Intervall-Mathematik gibt sie uns wieder zurück, zusammen mit einer durchaus erwünschten ästhetischen Komponente.

Literatur

Alefeld, G.; Herzberger, J., Über das Newton-Verfahren bei nichtlinearen Gleichungssystemen, ZAMM **50**, 773—774 (1970).
Alefeld, G.; Herzberger, J., Einführung in die Intervallrechnung, Herausg. K. H. Böhling, U. Kulisch, H. Maurer, Bibliographisches Institut Mannheim/Wien/Zürich 1974.
Barth, W., Ein Algorithmus zur Berechnung aller reellen Nullstellen in einem Intervall, Computing **9**, 327—333 (1972a).
Barth, W., Private Mitteilung (1972b).
Barth, W.; Nuding, E., Optimale Lösung von Intervallgleichungssystemen, Computing **12**, 117—125 (1974).
Beeck, H., Zur scharfen Außenabschätzung der Lösungsmenge bei linearen Intervallgleichungssystemen, ZAMM **54**, T 208—T 209 (1974).
Bierbaum, F., Einsatz der Intervallarithmetik bei der numerischen Konvergenz von ALGOL-60 Programmen, ‚Interval Mathematics', Ed. by K. Nickel, Lecture Notes in Computer Science 29, Springer Verlag 1975, 160—168.
Bierbaum, F., Intervall-Mathematik, Eine Literaturübersicht. 2. Auflage, Bericht Nr. 76/4 aus dem Institut für Praktische Mathematik der Universität Karlsruhe.
Böhmer, K.; Jackson, R. T., A FORTRAN-Triplex-Pre-Compiler based on the Augment Pre-Compiler, MRC Technical Summary Report 1732, University of Wisconsin, Madison 1977.
Brockhaus, M.; Rothmaier, B.; Schroth, P., Benutzeranleitung für Triplex-ALGOL im System Hydra 2, Interner Bericht des Inst. f. Informatik 69/11, Universität Karlsruhe 1969.
Hahn, W., Intervallarithmetik in normierten Räumen und Algebren, Dissertation, Bericht des Inst. f. Angew. Math. 3, Universität Graz 1971.
Hansen, E., Interval arithmetic in matrix computations. Part I, SIAM J. Numer. Analysis **2**, 308—320 (1965).
Hansen, E., Topics in Interval Analysis, Oxford University Press 1969.
Hansen, E.; Smith, R., Interval arithmetic in matrix computations. Part II, SIAM J. Numer. Analysis **4**, 1—9 (1967).
Henrici, P., Applied and Computational Complex Analysis, Wiley-Interscience, New York 1974.
Jahn, K.-U., Die Intervall-Arithmetik als Basis einer mehrwertigen Analysis, Dissertation zur Promotion B, Karl-Marx-Universität Leipzig 1977.
Krawczyk, R., Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken, Computing **4**, 187—201 (1969).
Marcowitz, U., Fehlerabschätzung bei Anfangswertaufgaben für Systeme von gewöhnlichen Differentialgleichungen mit Anwendung auf das Reentry-Problem, Numer. Math. **24**, 249—275 (1975).
Mayer, O., Über die in der Intervallrechnung auftretenden Räume und einige Anwendungen, Dissertation, Universität Karlsruhe 1968.
Moore, R. E., Interval Analysis, Prentice-Hall, Inc., Englewood Cliffs, N.J. 1966.
Moore, R. E., Intervallanalyse, Deutsche Übersetzung von: Moore, R. E., Interval Analysis, R. Oldenbourg Verlag, München, Wien 1969.
Nickel, K., Ein stets konvergenter Newton-Algorithmus mit Fehlerabschätzung (Notiz), ZAMM **48**, T 111 (1968).
Nickel, K., Triplex-ALGOL and applications, ‚Topics in Interval Analysis', Ed. by E. Hansen, Oxford University Press 1969, pp. 10—24.
Nickel, K., On the Newton method in interval analysis, MRC Technical Summary Report # 1136, University of Wisconsin, Madison 1971.
Nickel, K., Interval Mathematics, Lecture Notes in Computer Science 29, Springer Verlag 1975a.
Nickel, K., Verbandstheoretische Grundlagen der Intervall-Mathematik, ‚Interval Mathematics', Ed. by K. Nickel, Lecture Notes in Computer Science 29, Springer Verlag 1975b, pp. 251—262.
Nickel, K., Über die Stabilität und Konvergenz numerischer Algorithmen. Teil I und II, Computing **15**, 291—328 (1975c).
Oettli, W.; Prager, W., Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides, Numer. Math. **6**, 405—409 (1964).
Oettli, W., On the solution set of a linear system with inaccurate coefficients, SIAM J. Numer. Analysis **2**, 115—118 (1965).
Oettli, W.; Prager, W.; Wilkinson, J. H., Admissible solutions of linear systems with not sharply defined coefficients, SIAM J. Numer. Analysis **2**, 291—299 (1965).
Ratschek, H., Nichtnumerische Aspekte der Intervallarithmetik, ‚Interval Mathematics', Ed. by K. Nickel, Lecture Notes in Computer Science 29, Springer Verlag 1975, 48—74.
Rothmaier, B., Der Triplex-ALGOL Compiler der UNIVAC 1108, Interner Bericht des Inst. f. Prakt. Math. 74/1, Universität Karlsruhe 1974.
Walter, W., Differential and Integral Inequalities, Springer Verlag 1970.
Wippermann, H.-W., Manual für das System Triplex-ALGOL Karlsruhe, Institut für Angewandte Mathematik-Rechenzentrum, Universität Karlsruhe 1967.
Wippermann, H.-W. (Herausgeber); Apostolatos, N.; Krawczyk, R.; Kulisch, U.; Lortz, B.; Nickel, K.; Wippermann, H.-W., The Algorithmic Language Triplex-ALGOL 60, Numer. Math. **11**, 175—180 (1968).
Wisskirchen, P., Vergleich intervallarithmetischer Iterationsverfahren, Computing **14**, 45—49 (1975).
Wongwises, P., Experimentelle Untersuchungen zur numerischen Auflösung von linearen Gleichungssystemen mit Fehlererfassung, Dissertation, Interner Bericht des Inst. für Praktische Mathematik 75/1, Universität Karlsruhe 1975.

Anschrift: Professor Dr. K. Nickel, Institut für Angewandte Mathematik der Albert-Ludwigs-Universität, Hermann-Herder-Straße 10, D 7800 Freiburg i. Br., BRD

ZAMM 58, T 86 - T 95 (1978)

H. Werner

Neuere Entwicklungen auf dem Gebiete der nichtlinearen Splinefunktionen

Nach einer einleitenden Diskussion der Verallgemeinerung des Splinekonzepts und Definition des Begriffes regulärer Spline werden Anwendungen auf die Interpolation, Approximation, Anfangswertprobleme bei gewöhnlichen Differentialgleichungen und numerische Integration betrachtet. Neben Verfahren und theoretischen Überlegungen wird eine Fehleranalyse skizziert. Beispiele zeigen die Güte der Verfahren im Vergleich zu Standard-Methoden der praktischen Mathematik.

Einleitung

In diesem Vortrag soll der Begriff „Spline" allgemeiner gefaßt werden als es sonst in der Mehrzahl der Publikationen üblich ist. Als *Spline* werde eine Funktion bezeichnet, die stückweise aus mehrparametrischen (in der Regel durch geschlossene Darstellungen gegebenen) Funktionen zusammengesetzt ist. An den Übergangsstellen, auch *Knoten* genannt, sind die Bestandteile *glatt* von gegebener Ordnung k miteinander verheftet.

Meist dienen Splines dazu, eine anders schlecht zu handhabende Funktion gut anzunähern, etwa zum Zwecke der leichteren Berechnung in einem Computer. Dabei soll mit möglichst wenigen Operationen eine spezifizierte Genauigkeit erreicht werden.

Üblicherweise bestehen die geschlossenen Darstellungen in jedem einzelnen Teilintervall aus *linearen Kombinationen* gewisser Basisfunktionen, die ihrerseits Polynome, rationale Funktionen, Exponentialfunktionen oder ähnliches sein dürfen, wobei aber nur die linearen Faktoren als Parameter geändert werden.

Es sollen hier auch *nichtlineare Abhängigkeiten* zugelassen werden. Ziel dieses Vorgehens ist, unter geschickter *Ausnutzung bekannter Eigenschaften* der anzunähernden Funktion oder der sie definierenden Gleichungen, Differentialgleichungen oder Funktionalgleichungen mit *wenigen* Parametern zu *guten Approximationen* zu kommen.

Dabei darf die Ermittlung der Werte der Parameter und auch die Auswertung der Funktionen, die den Spline darstellen, natürlich nicht den Gewinn wettmachen, den man durch die Verringerung der Parameteranzahl erzielt hat.

Es sollen neben den zugrunde liegenden Konzepten und theoretischen Resultaten praktische Verfahren skizziert und Anwendungen beschrieben werden. Wir werden dementsprechend die Aufgaben der

Interpolation — Approximation — Numerischen Integration — Anfangswertprobleme

sowie die zugehörigen Fehlerabschätzungen und ihre asymptotischen Entwicklungen streifen.

Die einzelnen Bausteine findet man in den Arbeiten von Arndt, Baumeister, Braess, Meder, Micula, Runge, Schaback, Schomberg, Späth und dem Referenten.

Definitionen und Hilfssätze

Betrachtet werden sollen Funktionen in einem Intervall $I = [x_-, x_+]$. Gegeben seien ferner Zerlegungen dieses Intervalls, charakterisiert durch ihre Knoten:

$$x_- = x_0 < x_1 < \ldots < x_{m-1} < x_m = x_+ .$$

Das j-te Teilintervall $I_j = [x_{j-1}, x_j]$ habe die Länge

$$h_j = x_j - x_{j-1}\,, \quad \text{sei} \quad h = \max h_j .$$

Es werden Funktionenfamilien betrachtet, die in I_j wenigstens k-mal, in der Regel $k+2$-mal stetig differenzierbar sind,

$$\mathcal{F}_j \in C^k(I_j) .$$

Ein *nichtlinearer Spline* besteht dann aus einer Funktion $u(x) \in C^k(I)$, deren Restriktionen auf I_j für $j = 1, \ldots, m$ zu $\mathcal{F}_j$ gehören.

Typische Beispiele sind

„spezielle rationale Funktionen" $p(x) + \dfrac{c}{x-d}$ (Schaback, Arndt, Werner, Schomberg),

„spezielle Exponentialfunktionen" $p(x) + c \cdot e^{dx}$,
„reguläre Splinefunktionen" $p(x) + t(x, c, d)$ (i. w. Schaback).

Dabei ist $p(x)$ jeweils ein Polynom vom Grade $k-1$. Es werden allgemeiner auch Splines der Form

$$p(x) + a\left(1 + \frac{x}{b}\right)^{\gamma} \qquad \gamma \text{ fest oder variabel, also Parameter,}$$

betrachtet.

Die bei der Bearbeitung der ersteren Beispiele gemachten Erfahrungen führten zur Abstraktion gewisser Eigenschaften, die in der Definition des regulären Splines ihren Ausdruck gefunden haben.

Definition: Eine Klasse von Splines heißt *regulär und glatt*, wenn die Restriktionen auf jedes Teilintervall I_j die Form

$$p_j(x) + t_j(x, c, d) \quad \text{mit} \quad t_j \in \mathcal{T}_j$$

besitzen und die Funktionen von $\mathcal{T}_j$ in eindeutiger Weise durch die Ableitungen k-ter Ordnung,

$$c = t^{(k)}(x_{j-1}, c, d), \qquad d = t^{(k)}(x_j, c, d)$$

parametrisiert werden können.

In den meisten Fällen wird vorausgesetzt, daß die Funktionen t_j höhere Ableitungen nach x besitzen, und diese Ableitungen seien dann stetig differenzierbare Funktionen von c, d.

Allgemeiner werden später solche Funktionen betrachtet, bei denen die Funktionswerte und ihre Ableitungen in einem Knoten gerade als Parameter zur Repräsentation des Splinestückes der anschließenden Teilintervalle benutzt werden können.

Beschränkt man sich auf die Betrachtung regulärer Splinefunktionen, so zerfällt die Berechnung eines Splines, d. h. die Bestimmung der Parameter, in zwei Teilprobleme, nämlich erstens die Festlegung der k-ten Ableitung in den Knoten, also eine nichtlineare Aufgabe, und zweitens die Berechnung der polynomialen Anteile. Dies ist eine mit Methoden der linearen Algebra zu bewältigende lineare Aufgabe.

Bei der Aufstellung der Bestimmungsgleichungen für die Parameter leitet man zunächst mit Hilfe geeigneter Formalismen aus den Bestimmungsgleichungen solche Gleichungen ab, die nur noch die k-ten Ableitungen enthalten. Üblicherweise versucht man, mit Hilfe der Differenzenquotienten die polynomialen Bestandteile zu annullieren.

Ich möchte dies an dem Beispiel der Interpolation kurz erläutern. Dabei werden folgende Bezeichnungen und Hilfssätze verwendet.

Die *Differenzenquotienten* werden in der Form (Werner-Schaback [27]) geschrieben

$$\begin{aligned} &\Delta^1(x_j, x_{j+1})\, u := \frac{u(x_{j+1}) - u(x_j)}{x_{j+1} - x_j} \quad \text{und} \\ &\Delta^{n+1}(x_j, \ldots, x_{j+n+1})\, u := \Delta^1_t(x_j, x_{j+n+1})\, \Delta^n(t, x_{j+1}, \ldots, x_{j+n})\, u \quad \text{für} \quad n \geqq 1 . \end{aligned} \tag{1}$$

In der üblichen Weise erhält man „*konfluente*“ *Differenzenquotienten*, wenn Stützstellen x_j zusammenfallen. So ist beispielsweise

$$\Delta^2(x_j, x_j, x_{j+1})\, u = \frac{\Delta^1(x_j, x_{j+1})\, u - u'(x_j)}{x_{j+1} - x_j} \tag{2}$$

Eigenschaften der Differenzenquotienten sind an der zitierten Stelle zusammengetragen, vgl. auch Popoviciu [14]. Explizit erwähnt sei

Hilfssatz 1: *Sei $u(x) \in C^{k+2}$, und M_j bezeichne die k-te Ableitung $\mathrm{D}^k u(x_j)$. Dann gilt*

$$\Delta^k(\underbrace{x_0, x_0, \ldots, x_0}_{i}, \underbrace{x_1, \ldots, x_1}_{j})\, u = \frac{i \cdot M_0 + j \cdot M_1}{(k+1)!} + R \quad \text{mit} \quad R = O(|x_1 - x_0|^2), \quad i + j = k + 1 . \tag{3}$$

Der Beweis ergibt sich, indem man das Interpolationspolynom p betrachtet, welches in x_0 und x_1 mit $u(x)$ bis zur i-ten bzw. $(j-1)$-ten Ableitung übereinstimmt, so daß $\mathrm{D}^l(p(x) - u(x)) = O(|x_1 - x_0|^{k+2-l})$ gilt, $l = 0, \ldots, k$. Für $p(x)$ rechnet man die Formel elementar aus.

Unmittelbare Konsequenzen sind beispielsweise die Formeln

$$\lambda \Delta^2(x_{j-1}, x_j, x_j)\, u + \mu \Delta^2(x_j, x_j, x_{j+1})\, u = \Delta^2(x_{j-1}, x_j, x_{j+1})\, u \tag{4}$$

mit $\lambda = h_j/(h_j + h_{j+1})$, $\mu = h_{j+1}/(h_j + h_{j+1})$, sowie

$$\Delta^2(x_j, x_j, x_{j-1})\, u + \Delta^2(x_{j-1}, x_{j-1}, x_j)\, u = (u'(x_{j+1}) - u'(x_j))/h_j .$$

Die Formel (4) läßt sich auf höhere Differenzenquotienten übertragen und gestattet es, einen beliebigen k-ten Differenzenquotienten, über Knoten als Stützstellen genommen, als eine konvexe Kombination solcher konfluenter Differenzenquotienten zu schreiben, die nur je zwei benachbarte Knoten benutzen.

Hilfssatz 2: *Sei $u(x) \in C^k[x_-, x_+]$. Dann gilt*

$$\Delta^k(x_{j_0}, \ldots, x_{j_k})\, u(x) = \sum_{j=0}^{N-k} c_j \Delta^k(z_j, \ldots, z_{j+k})\, u \tag{5}$$

mit $c_j \geqq 0$, $\sum c_j = 1$ (eine konvexe Linearkombination) und

$$Z(X) := \{z_0, \ldots, z_N\} = \{\underbrace{x_{j_0}, \ldots, x_{j_0}}_{i\text{-mal}}, \underbrace{x_{j_0+1}, \ldots, x_{j_0+1}}_{k\text{-mal}}, \ldots, \underbrace{x_{j_k-1}, \ldots, x_{j_k-1}}_{k\text{-mal}}, \underbrace{x_{j_k}, \ldots, x_{j_k}}_{j\text{-mal}}\} ,$$

wenn x_{j_0} i-mal und x_{j_k} j-mal als konfluentes Argument auftritt.

Die Koeffizienten c_j lassen sich wieder mit Hilfe der h_j darstellen.

Der Beweis ist für $k = 1$ eine triviale Identität:

$$\Delta^1(x_j, x_n)\, u = \sum_{i=j}^{n-1} \frac{x_{i+1} - x_i}{x_n - x_j} \Delta^1(x_i, x_{i+1})\, u . \tag{6}$$

Für $k > 1$ wird ein Induktionsbeweis geführt.

Mit $X = (x_{j_0}, \dots, x_{j_k})$, dem Stützstellenvektor des Differenzenquotienten, assoziieren wir als $L(X)$ die Anzahl $N + 1$ der Elemente der Menge $Z(X)$. Die Zahl $L(X)$ wird Länge von X genannt.

Der Beweis wird durch Induktion nach L geführt. Ist $L(X) = k + 1$, so besteht die rechts stehende Summe nur aus dem gegebenen Differenzenquotienten selbst, es treten nur zwei benachbarte Stützstellen in Z auf.

Ist $L(X) > k + 1$, so müssen in Z zumindest drei verschiedene Stützstellen vorkommen. Für $L(X) < L_0$ sei die Behauptung bereits bewiesen.

Betrachte X mit $L(X) = L_0$. Dann kann man schreiben

$$\begin{aligned} \Delta^k(x_{j_0}, x_{j_1}, \dots, x_{j_k}) &= \Delta_t^2(x_{j_0}, x_{j_1}, x_{j_k})\, \Delta^{k-2}(t, x_{j_2}, \dots, x_{j_{k-1}}) \\ &= \lambda \cdot \Delta_t^2(x_{j_0}, x_{j_1}, x_{j_1}) \cdot \Delta^{k-2}(t, \dots) + \mu \cdot \Delta_t^2(x_{j_1}, x_{j_1}, x_{j_k}) \cdot \Delta^{k-2}(t, \dots) \end{aligned} \tag{7}$$

nach (4), sofern $x_{j_0} < x_{j_1} < x_{j_k}$ gilt, falls $x_{j_0} = x_{j_1}$:

$$\Delta^k(x_{j_0}, x_{j_1}, \dots, x_{j_k}) = \frac{1}{\mu} \Delta^2(x_{j_0}, z, x_{j_k}) \cdot \Delta^{k-2}(t, \dots) - \frac{\lambda}{\mu} \Delta^2(x_{j_0}, x_{j_0}, z) \cdot \Delta^{k-2}(t, \dots) . \tag{8}$$

Im ersten Fall ist dabei Formel (4) angewendet mit Identifikation der Punkte durch

$$x_{j-1} = x_{j_0}, \qquad x_j = x_{j_1}, \qquad x_{j+1} = x_{j_k},$$

im zweiten Fall mit

$$x_{j+1} = x_{j_k}, \qquad x_j = x_{j_0} = x_{j_1}, \qquad x_{j-1} = z,$$

ein zwischen x_{j_0} und x_{j_k} liegender Knoten, verwendet worden.

Im zweiten Fall erhält man für die Koeffizienten

$$\frac{1}{\mu} = \frac{x_{j_k} - z}{x_{j_k} - x_{j_0}}, \qquad -\frac{\lambda}{\mu} = \frac{z - x_{j_0}}{x_{j_k} - x_{j_0}},$$

beide sind wie im ersten Fall positiv und ihre Summe ist gleich 1.

In beiden Fällen ist also $\Delta^k(X)\, u$ als konvexe Summe zweier k-ter Differenzenquotienten ausgedrückt worden, deren Stützstellen die Länge $L_0 - 1$ zugeordnet bekommen, denn ihre Z-Mengen gehen aus $Z(X)$ durch Weglassen des ersten oder letzten Elementes hervor. Die Induktionsvoraussetzung ist also anwendbar. Der Beweis folgt aus der Tatsache, daß eine konvexe Summe konvexer Summen wieder eine konvexe Summe ist.

Interpolation mit regulären Spline-Funktionen

Das Interpolationsproblem (Arndt [1], Schaback [17, 18], Werner [22]) besteht darin, einen Spline $u(x)$ zu bestimmen, der gegebene Funktionswerte $f(x_i)$ an Knotenpunkten x_i $(i = 0, \dots, m)$ und, da noch weitere k Bedingungen frei sind, etwa außerdem noch am linken und rechten Intervallendpunkt x_0 und x_m vorgegebene Werte für die Ableitungen annimmt,

$$\begin{aligned} \mathrm{D}^i u(x_0) &= f^i(x_0) \qquad \text{für } i = 1, \dots, k_0 \\ \mathrm{D}^j u(x_m) &= f^j(x_m) \qquad \text{für } j = 1, \dots, k_m \end{aligned} \qquad \text{mit } k_0 + k_m = k .$$

Um das vorher skizzierte Schema anzuwenden, versucht man, durch Bildung von Differenzenquotienten, die nur die gegebenen Funktionswerte $f(x_j)$ und die gegebenen Ableitungen benutzen, die polynomialen Bestandteile zu eliminieren. Die Anwendung eines k-ten Differenzenquotienten führt aber nur dann zur Annullierung von $p_j(t)$, wenn die Stützstellen, über denen die Differenzenquotienten gebildet werden, alle zum Definitionsbereich dieses einen Polynoms gehören. Bei den vorliegenden Splines können jedoch die Polynome $p_j(t)$ von Teilintervall zu Teilintervall verschieden sein. In dieser Situation hilft der Hilfssatz 2. Die Anwendung dieses Hilfssatzes ergibt *also $m + 1$ Gleichungen für die k-ten Ableitungen der Funktion u in den Stützstellen x_i, für $i = 0, \dots, m$.* Auf der rechten Seite treten bekannte Werte auf.

Bedenkt man, daß für hinreichend kleine Abstände die k-ten Differenzenquotienten von f bis auf den Faktor $k!$ k-te Ableitungen sind, so ist es natürlich, zu verlangen, daß bei einer durch Interpolation zu approximierenden Funktion $f(x)$ die k-ten Ableitungen von $f(x)$ im Wertebereich der k-ten Ableitungen der den Spline erzeugenden Familien $\mathcal{F}_j$ liegen. Es ergibt dies eine für die Lösbarkeit des Interpolationsproblems notwendige Bedingung.

So bekommt man also die Interpolationsgleichungen

$$\sum c_{i,j_0}\, \Delta_x^k(z_{i,j_0}, \dots)\, t_j(x, M_{j-1}, M_j) = \Delta^k(x_{j_0}, \dots, x_{j_0+k})\, f , \tag{9}$$

wobei zwischen den z_{i,j_0}, den Koeffizienten c_{i,j_0} und den k-ten Ableitungen $M_j = \mathrm{D}^k u(x_j)$ die durch Hilfssatz 2 gegebenen Relationen bestehen.

Man kann die $\Delta^k t_j$ ihrerseits wieder entwickeln, indem man die Differenzenquotienten bis auf ein Restglied der Ordnung $O(h^2)$ durch lineare Kombinationen von Differentialquotienten ersetzt.

Diese Gleichungen bilden den Schlüssel zu einer Existenzaussage, die sich bei Arndt [1] findet. Für hinreichend kleine Intervalle ist danach die oben für reguläre Splines aufgestellte Interpolationsaufgabe lösbar, wenn dies für polynomiale Splines $(k + 1)$-ten Grades richtig ist. Man bekommt nach der Linearisierung von (9) auf der linken Seite eine invertierbare Matrix, rechts treten die k-ten Ableitungen mit kleinen Faktoren auf. Die Lösung kann iterativ gewonnen werden. Der Arbeitsaufwand ist vergleichbar mit dem für die Lösung mit einem polynomialen Spline bei der gleichen Aufgabe.

Nachdem die k-ten Ableitungen ermittelt sind, kann man die niedrigeren Ableitungen von u konstruieren. Im linken Randpunkte x_0 sind die Ableitungen bis zur Ordnung k_0 aufgrund der Vorgabe bekannt. Betrachtet man sukzessiv k-te Differenzenquotienten mit dem Punkt x_0 $k_0 + 2$, $k_0 + 3$, ... mal als Argument, so kann man die Ableitungen der Ordnung $k_0 + 1, \ldots, k - 1$ gewinnen, die k-te Ableitung im Punkte x_0 ist schon bekannt. So sind zunächst für das Intervall $[x_0, x_1]$ genügend Daten zur Festlegung von u vorhanden. Die Ableitungen in x_1 und der gegebene Wert $f(x_2)$ bestimmen u in $[x_1, x_2]$, usw. Der Fehler zwischen u und einer $k+2$-mal stetig differenzierbaren Funktion $f(x)$ ist von der Größenordnung h^{k+2}. Es ist durch diese Konstruktion gleichzeitig deutlich geworden, wie man die linearen Parameter, nämlich die Koeffizienten der Polynome $p_j(x)$, finden kann, nachdem die nichtlinear eingehenden k-ten Ableitungen bekannt sind.

Zur Demonstration werde für den Fall $k = 2$ und die speziellen rationalen Splines das System der Gleichungen explizit angegeben

$$\lambda_j(M_{j-1} \cdot M_j^2)^{1/3} + \mu_j(M_j^2 \cdot M_{j+1})^{1/3} = A^2(x_{j-1}, x_j, x_{j+1})\, f =: A_j\,. \tag{10}$$

Man sieht diesen Gleichungen nicht auf den ersten Blick an, daß sie sich bei Linearisierung nur um Glieder R_j der Größenordnung h^2 von den bekannten Relationen für die kubischen Splines

$$\lambda_j(\tfrac{1}{3} M_{j-1} + \tfrac{2}{3} M_j) + \mu_j(\tfrac{2}{3} M_j + \tfrac{1}{3} M_{j+1}) = A_j + R_j \tag{11}$$

unterscheiden.

Herr SCHABACK machte 1969 die Beobachtung, daß die Gleichungen (10) auch als EULERsche Gleichungen des Optimierungsproblems

$$E(M_0, \ldots, M_m) = \sum \left(\frac{h_j}{z_{j-1} \cdot z_j} + (h_j + h_{j+1}) \cdot A_j z_j \right) = \min \quad \text{mit} \quad z_i = (M_i)^{1/3} \tag{12}$$

aufgefaßt werden können. Damit kann man gleichzeitig einen Existenzbeweis im Großen führen.

Herr SCHABACK hat diesen Ansatz in einer weiteren Arbeit auf allgemeinere Klassen von regulären Splines ausgedehnt. Es gelingt ihm, sofern alle Funktionenfamilien $\mathcal{F}_j$ gleich sind, auch für diesen Fall, einen Existenzbeweis im Großen zu führen. Herr BAUMEISTER [5] hat in seiner Dissertation das zum SCHABACKschen Ansatz duale Problem behandelt und auf diese Weise ebenfalls einen Existenzbeweis für die Interpolation mit regulären Splines erhalten.

Approximation mit regulären Splines

Man kann die in der Approximationstheorie üblichen Techniken anwenden, um für reguläre Splines als Klasse der Approximationsfunktionen Existenz und Charakterisierung bester Approximationen im Sinne von TSCHEBYSCHEFF zu einer gegebenen im Intervall $I = [x_-, x_+]$ stetigen Funktion f zu behandeln. Hier sollen nur einige Bemerkungen zu diesem Thema gemacht werden, es sei auf die Arbeiten von BRAESS und WERNER [7], MEDER [12], SCHOMBERG [18], WERNER [20, 21], WERNER und LOEB [26] verwiesen.

Die Frage der Existenz ist aufs engste mit dem Problem der Abschließung der Familien regulärer Splines bezüglich gleichmäßiger Konvergenz in abgeschlossenen Teilintervallen von (x_-, x_+) verknüpft. Wieder abstrahiert man von den Erfahrungen mit der Approximation durch rationale Splines bzw. Exponentialsplines, um Eigenschaften zu formulieren, die in diesem Sinne eine Kompaktheit gleichmäßig beschränkter Funktionenfolgen garantieren. Um eine allgemeine Theorie herleiten zu können, kann man axiomatisch fordern, daß die k-ten Ableitungen der Approximationsfunktionen ihr Vorzeichen nicht wechseln. Das könnte als *Steifheit* bezeichnet werden. Man kann dann zeigen, daß bezüglich der punktweisen Konvergenz oder auch oben genannten gleichmäßigen Konvergenz diese Folgen kompakt sind und das genügt, um die Existenztheorie der TSCHEBYSCHEFF-Approximation aufzubauen.

Die Charakterisierung der Funktionen des Abschlusses und damit zusammenhängend auch die Charakterisierung der TSCHEBYSCHEFF-Approximierenden verlangt mehr Einblick in die Struktur der Funktionenklasse.

Von der Approximation mit rationalen Funktionen und Exponentialsummen weiß man, daß Entartungen auftreten können, wenn nicht alle Parameter der Approximationsfamilie ausgenutzt werden und dies findet in der Verringerung der zur Charakterisierung der besten Approximation notwendigen Alternantenpunkte seinen Niederschlag. Mit diesem Effekt hat man natürlich auch bei Splines zu rechnen.

Bei den rationalen Splines, beispielsweise, hat man jedoch ein Phänomen, das dieser Reduktion an Alternantenpunkten entgegenwirkt.

Entarten etwa im Grenzübergang die rechts und links von einem Knoten auftretenden speziellen rationalen Funktionen $a + bx + \dfrac{c}{x-d}$ zu linearen Funktionen, so kann dies bei geeigneter Kopplung der Parameter so vor sich gehen, daß die zweite Ableitung in dem Knoten gegen ∞ strebt, die ersten Ableitungen einen Sprung erleiden, die Funktion selbst stetig bleibt. Man kann leicht diskutieren, dies geschieht in den zitierten Arbeiten, welche Möglichkeiten für solche Entartungen auftreten können. Axiomatisch wird diesem Verhalten bei regulären Splinefamilien durch die „*Steilheit*" Rechnung getragen. Je nachdem, ob man mit festen oder variablen Knoten *arbeitet, können mehr oder weniger* starke Unstetigkeiten der k-ten und $(k-1)$-ten Ableitungen bei Grenzprozessen entstehen, dies ist in WERNER-LOEB [26] genauer ausgeführt. Bei der Ermittlung der zur Charakterisierung der besten Approximation notwendigen Anzahl von Alternantenpunkten sind die Knoten mit einem vom Verhalten der Grenzfunktion abhängigen Gewicht zu berücksichtigen.

Anfangswertprobleme gewöhnlicher Differentialgleichungen und ihre Behandlung mit regulären Splines

Betrachtet werde die klassische Fragestellung:

Gegeben sei eine Differentialgleichung

$$y' = f(x, y), \quad \text{mit} \quad f(x, y) \in C^{k+2}(G), \qquad G = I \times \mathbb{R}^n, \qquad I = |x_0, x_+|\,, \tag{13}$$

eine Anfangsbedingung

$$y(x_0) = y_0\,.$$

Gesucht ist eine Funktion $y(x)$, die die Differentialgleichung löst und durch den Punkt $(x_0, y_0) \in G$ geht.

Konstruiert werden soll ein Algorithmus, indem die Lösung des Problems ersetzt wird durch einen Spline, der dann natürlich nur näherungsweise die Differentialgleichung erfüllen wird. Der Ansatz dieser Art geht bereits zurück auf Loscalzo und Talbot [11], die mit kubischen Splines arbeiten. Andererseits haben nichtlineare Ansätze auch Lambert und Shaw [9, 10] verwendet, ohne jedoch glatte Lösungen zu konstruieren. Der hier vorgelegte Ansatz für reguläre Splines wurde für den Spezialfall rationaler Splines in der Dissertation von Runge behandelt, in der hier vorgetragenen allgemeinen Form bereits in Arbeiten des Autors und in einer Verallgemeinerung, die demnächst erscheinen wird, von Arndt [2].

Wir werden $x_j = h \cdot j + x_0$ als Knoten des zu konstruierenden Splines $u(x)$ ansetzen und haben nur noch einen Algorithmus anzugeben, der Stück um Stück diesen Spline zu berechnen gestattet. Durch den Anfangswert und die Differentialgleichungen sind Funktionswert $u(x_0)$ und erste Ableitung im Punkte x_0 gegeben. Man kann z. B. durch Ableitung der Differentialgleichung auch die zweite Ableitung von $u(x)$ im Punkte x_0 bestimmen. Wir wollen, und darauf muß man sich aus Stabilitätsgründen bei dieser einfachen Methode beschränken, welche wir beschreiben, annehmen, daß der Spline von der Ordnung $k = 2$ ist, also nur vier Parameter verwendet werden.

Durch die genannte Vorgabe im linken Endpunkt des Teilintervalls hat man also drei Bedingungen und kann den einen freien Parameter so wählen, daß im rechten Randpunkte des Intervalls der Spline die Differentialgleichung erfüllt.

Ist der Spline bis zu einem Knoten x_j ($j \geqq 0$) bereits konstruiert worden, so sind durch die Anfangswerte oder die bereits bis dorthin fortgesetzte Näherungslösung die 0-te, 1-te und 2-te Ableitung in diesem Punkte vorhanden, gegeben durch ihre linksseitigen Grenzwerte. Sie sollen aufgrund der Stetigkeitsforderungen mit den rechtsseitigen übereinstimmen.

$$u(x_j, h) = u(x_j - 0, h)\,, \qquad u'(x_j, h) = u'(x_j - 0, h)\,, \qquad u''(x_j, h) = u''(x_j - 0, h)\,. \tag{14}$$

In $|x_j, x_{j+1}|$ ist also die Funktion $u(x, h)$ so zu bestimmen, daß gilt

$$u'(x_{j+1}, h) = f(x_{j+1}, u(x_{j+1}, h))\,. \tag{15}$$

Es werde angenommen, daß diese Gleichung eine Lösung besitzt. In konkreten Fällen kann man dies genauer diskutieren.

Auf diese Weise wird der Spline bis zum Punkte x_{j+1} fortgesetzt und gleichzeitig werden die benötigten Anfangswerte für die Berechnung im nächsten Teilintervall ermittelt.

Es leuchtet ein, daß, da dieses Verfahren wie ein Einschrittverfahren aussieht, man insbesondere leicht die Schrittweite variieren kann.

Zwei Beobachtungen bezüglich der Ordnung dieses Verfahrens:

1. Unterscheidet man wie üblich zwischen lokalen und globalen Fehlern eines Verfahrens und untersucht man die Einschrittverfahren bezüglich ihrer Konvergenzordnung, so findet man bei Einschrittverfahren das Resultat, daß beim Übergang von lokalen zum globalen Fehler (je nachdem wie man die Definition gewählt hat) die Ordnung bezüglich der Schrittweite h um 1 erniedrigt wird.

   Vergleicht man beispielsweise beim Eulerschen Polygonverfahren u_1 mit dem Wert $y(x_1)$ der Lösung mit dem Anfangswert $y(x_0) = u_0$, so ist dieser lokale Fehler von der Ordnung $O(h^2)$.

   Vergleicht man hingegen u_j (mit $j \cdot h = x^* - x_0$) und $y(x^*)$ bei festem x^*, so verhält sich bekanntlich die Differenz wie $O(h)$, denn eine h-Potenz wird durch das Aufsummieren von j ($\sim 1/h$) Fehlertermen kompensiert.
2. Betrachtet man jetzt das auf die Splinefunktionen bauende Verfahren, so werden beim Start im Punkte x_0 Funktionswert und 1. und 2. Ableitung exakt sein, die 3. Ableitung wird durch die am anderen Intervallendpunkt x_1 gegebene Bedingung bestimmt, d. h. also, es ist sehr unwahrscheinlich, daß auch der Anfangswert der 3. Ableitung exakt ist. Man wird also erwarten, daß die Fehlerentwicklung mit dem Term $c \cdot h^3$ beginnt, also lokal zunächst scheinbar sogar nur die Ordnung 3 hätte, so daß man global ein Verfahren quadratischer Ordnung erwartet.

   Es ist jedoch natürlich, daß die zusätzliche Bedingung in x_1 eine zusätzliche Potenz von h zur Konvergenzordnung beiträgt, das gibt lokal $O(h^4)$. Es ist allerdings nicht so plausibel, daß durch dieses Verfahren sogar ein Verfahren 4. Ordnung entsteht.

Ohne auf die Einzelheiten des höchst technischen Beweises einzugehen, sei mitgeteilt, daß die entstehenden Verfahren von 4. Ordnung sind. Um ein Gefühl für die Konvergenzverhältnisse zu gewinnen, wenden wir uns zunächst dem lokalen Fehler zu, machen den Ansatz

$$u(x, h) = y(x) + w(x, h)$$

und versuchen, den Fehler w als eine Funktion von x in eine Potenzreihe zu entwickeln. Ihre Koeffizienten werden Funktionen des Parameters h sein und können ihrerseits nach h-Potenzen entwickelt werden.

Zur Vereinfachung der Schreibweise sei $x_0 = 0$, und es gilt

$$w(0, h) = w'(0, h) = w''(0, h) = 0 . \tag{16}$$

Setzt man die Spline mit den bekannten Daten an, so kann man schreiben

$$u(x, h) = u\big(x, y_0, y_0', y_0'', y_0''' + w'''(0, h)\big) = \frac{x^3}{3!} \cdot w'''(0, h) + R^3[w^{IV}](x, h) + y(x) \tag{17}$$

$$u'(x, h) = w'(x, h) + y'(x) = \frac{x^2}{2} \cdot w'''(0, h) + R^2[w^{IV}](x, h) + y'(x) , \tag{18}$$

das Fehlerglied läßt sich in Integralform schreiben

$$R^n[w](x, h) = \int_0^x \frac{(x - t)^n}{n!} \cdot w(t, h)\, dt .$$

Als Bestimmungsgleichung für $w'''(0, h)$ erhält man

$$y'(h) + w'(h, h) = f\big(h, y(h) + w(h, h)\big) . \tag{19}$$

Berücksichtigung der Differentialgleichung führt zu

$$w'(h, h) = f_y\big(h, y(h)\big) \cdot w(h, h) + \tfrac{1}{2} f_{yy} \cdot w^2 + \dots , \tag{20}$$

also gilt die Entwicklung

$$w'''(0, h) = w_0''' + h \cdot w_1''' + h^2 \cdot w_2''' + \dots \quad \text{mit} \quad w_0''' = 0, \quad w_1''' = -w^{IV}(0, y_0, y_0', y_0'', y_0''')/3 . \tag{21}$$

Lokal ist damit

$$w(x, h) = O(h^4) \tag{22}$$

gezeigt.

Der allgemeine Beweis kann nun so fortschreiten, daß man die Koeffizienten der TAYLOR-Entwicklung im nächsten Punkte ermittelt. Man kann zwischen den Koeffizienten benachbarter Knoten eine Rekursionsformel aufstellen und an Hand dieses Schemas für die Vektoren der Koeffizienten sehen, daß sie nicht anwachsen, sondern beschränkt bleiben. Lokale und globale Konvergenzordnung sind also gleich. Dies soll hier nicht ausgeführt werden. (Vgl. WERNER [22, 24]).

Man bekommt neben dem gewünschten Konvergenzresultat gleichzeitig einen guten Einblick in die Struktur des Fehlers. Dies soll hier nur für den Spezialfall weiter verfolgt werden, daß die rechte Seite der Differentialgleichung nicht von y abhängt, d. h. daß man es mit der numerischen Integration einer Funktion zu tun hat.

Quadratur

Die Berechnung von

$$y(x) = \int_{x_0}^{x} f(t)\, dt \tag{23}$$

wird zurückgeführt auf die Lösung der Differentialgleichung

$$y'(x) = f(x)$$

mit dem Anfangswert $y(x_0) = 0$.

Prinzipiell gelten dann die bisherigen Entwicklungen, sie vereinfachen sich nur. Diese Vereinfachung werde ausgenutzt, um in übersichtlicher Form die oben begonnenen Fehlerabschätzungen für diesen Spezialfall zu Ende zu führen. Zunächst betrachten wir nur

$$w'(x, h) = u'(x, h) - y(x) = u'(x, h) - f(x)$$

und bekommen dann den Fehler der Funktion selbst durch eine einfache Integration von $w'(x, h)$.

Nach Konstruktion ist $u(x, h)$ und damit auch $w(x, h)$ zweimal stetig differenzierbar, in den einzelnen Teilintervallen zwischen den Knoten jedoch sogar viermal, für die folgenden Abschätzungen voraussetzungsgemäß sogar fünfmal.

Die Differenz $w'(x, h)$ verschwindet in jedem Knoten. Der Mittelwertsatz der Differentialrechnung liefert deshalb in jedem (offenen) Teilintervall eine Nullstelle von w''. Außerdem verschwindet diese Ableitung aufgrund der Vorgabe in x_0. Dies läßt den Schluß zu, daß auch noch die nächste Ableitung w''' in (x_0, x_1) eine Nullstelle besitzt. Bei der Ableitung $w^{IV}(x, h)$ müssen wir davon ausgehen, daß sie bis auf eine Variation der Ordnung $O(h)$ konstant und durchaus von Null verschieden sein wird. Durch Integration folgt daraus, daß

w''' eine lineare Funktion bis auf $O(h^2)$, $\quad$ w' ein kubisches Polynom bis auf $O(h^4)$,
w'' eine Parabel bis auf $O(h^3)$, $\quad$ w ein Polynom 4. Grades bis auf $O(h^5)$

ist. Man kann also zwischen einem *Hauptterm* und *O-Term* des Fehlers unterscheiden. Aufgrund der doppelten Nullstelle von w_1 bei x_0 und der einfachen bei x_1 gilt

$$0 = \Delta_x^2(x_0, x_0, x_1)\, w'(x, h) = \frac{2w'''(x_0, h) + w'''(x_0, h)}{3!} + O(h^2) . \tag{24}$$

Daraus folgt, daß die Nullstelle von w''' bei $x_0 + \frac{1}{3} h + O(h^2)$ liegt, die Nullstelle von w'' bei $x_0 + \frac{2}{3} h + O(h^3)$. Der Fehlerverlauf ist auf der beigefügten Skizze zu sehen.

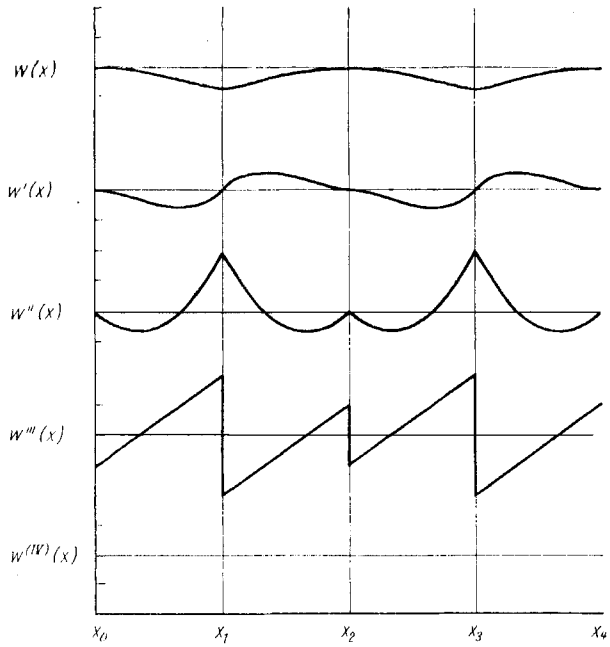


Fig. 1. Verhalten des Fehlers $w(x, h) = u(x, h) - y(x)$ und seiner Ableitungen

Mit den in x_1 erhaltenen Werten für w' und seinen Ableitungen und der Stetigkeitsforderung an w' und w'', sowie dem Verschwinden von w' in sämtlichen Knoten kann man nun die Konstruktion von w' für $[x_1, x_2]$ fortsetzen.

Ohne die Rechnungen im einzelnen wiederholen zu müssen, sieht man sofort, daß, bis auf die angegebenen O-Terme, diese Aufgabe gelöst wird, wenn man die obigen Polynome von

$$w,\ w'',\ w^{\mathrm{IV}} \text{ als } \textit{gerade},\quad w',\ w''' \text{ als } \textit{ungerade}$$

Funktion auf $[-h + x_0, x_0]$ und dann als periodische Funktionen der Periode $2h$ fortsetzt. Das Ergebnis für das Intervall $[x_0, x_4]$ ist skizziert. Bemerkenswert ist, daß w''' zwar in den Knoten unstetig ist, das Integral über eine Periode aber bis auf $O(h^2)$ verschwindet.

Der Hauptterm des Fehlers schaukelt sich also nicht auf, sondern bleibt beschränkt, eine Eigenschaft, wie man sie auch bei stabilen Zwei-Schritt-Verfahren beobachtet. Die O-Terme zeigen hingegen das von den Einschrittverfahren her bekannte Wachstum. All diese Überlegungen zusammenfassend sieht man also, daß man ein Verfahren vierter Ordnung vor sich hat.

Will man sich die Existenz asymptotischer Entwicklungen des Fehlers zunutze machen, um Extrapolationstechniken zur Verbesserung der Genauigkeit anzuwenden, so folgt aus den vorangegangenen Untersuchungen, daß man nur geradzahlige Punkte $x_{2j}^{(h)} = x_0 + 2jh$ verwenden sollte, denn dann ist gewährleistet, daß die vom Hauptterm und O-Term stammenden Fehler für $x_{2j}(h)$, $x_{4j}\left(\frac{h}{2}\right)$, ... „gleichgerichtet" sind.

Es sei nur erwähnt, daß bei der numerischen Quadratur und allgemeiner der Integration von Anfangswertproblemen die Ordnung erhöht werden kann, indem man etwa „Blockverfahren" anwendet, d. h. die Splines nicht nur in einem Intervall, sondern gleichzeitig für mehrere Intervalle definiert und dann auch auf höhere Ableitungen bei der Berechnung zurückgreift. Es ergeben sich die üblichen Stabilitätsprobleme und es gibt Verfahren, in denen man Stabilität bekommt und andere, in denen sie verlorengeht (Arndt [2]).

Zur numerischen Quadratur wird man in der Regel nicht ein so kompliziertes Verfahren verwenden, wenn man mit klassischen Verfahren, die auf polynomialen Ansätzen und Extrapolationsverfahren beruhen (Romberg-Integration), das Gleiche erreichen kann. Immerhin erscheint mir doch bemerkenswert, daß das Verfahren gegenüber diesen Methoden in gewissen Fällen überlegen ist. Beispiele zeigen die Überlegenheit bei solchen Funktionen, die in der Nähe des Integrationsintervalls singulär werden, vor allem dann, wenn die Singularität von hoher Ordnung ist. Die Splines gestatten es nämlich, einen Ansatz zu machen, der dieser Singularität Rechnung trägt oder, sofern die Ordnung der Singularität nicht leicht a priori zu bestimmen ist, deren Ordnung als zusätzlichen Parameter zu ermitteln.

Es bestehen für den Fall der numerischen Quadratur enge Beziehungen zu den von Wuytack vorgeschlagenen rationalen Ansätzen. In unserem Ansatz wird nur darür gesorgt, daß einmal die Ableitungen von $f(x, y)$ nur in (x_0, y_0), nicht aber beim Integrieren benötigt werden und andererseits die erhaltene Stammfunktion stückweise analytisch gegeben und sogar durchweg zweimal stetig differenzierbar ist. Beispiele, für die ein Vergleich der verschiedenen Verfahren durchgeführt wird, findet man in Werner und Wuytack [28].

Eines sei herausgegriffen. Zu berechnen sei

$$\int_0^1 \frac{x^{1{,}1}}{(1{,}1 - x)^3}\,dx = 100$$

Verglichen werden Trapezregel, SIMPSON-Verfahren und Spline-Verfahren mit festem $\alpha = -2$ und schließlich mit variablem α.

Zum Vergleich sind Zahlen untereinandergestellt, zu deren Berechnung die gleiche Anzahl k von Funktionswerten benutzt wurde.

k	9	17	33
Trapezregel	162,230	118,76507	105,050905
SIMPSON-Regel	124,377	104,27660	100,479518
Spline, $\alpha = -2$	99,908	99,98941	99,999136
Spline, α frei	99,959	99,99438	99,999468

Eine Anwendung

Es soll die Lage der Singularität der Lösungen von Anfangswertproblemen von Differentialgleichungen bestimmt werden, bei denen die rechte Seite $f(x, y)$ ein Polynom in x und y ist. Für den Spezialfall der RICCATIschen Differentialgleichungen ($f(x, y)$ quadratisch in y) weiß man, daß, von gewissen Ausnahmen abgesehen, mit Polen 1. Ordnung zu rechnen ist.

Bei allgemeineren Polynomen setzt man zweckmäßig mit algebraischen Singularitäten (oder allgemeiner als Exponentialfunktionen geschriebenen Singularitäten) versehene Splinefunktionen an.

Soll das vorherbeschriebene Verfahren verwendet werden, so arbeitet man die Anfangswerte u_0, u_0', u_0'' direkt in den Ansatz ein und kann etwa folgende *Funktionenklassen* benutzen:

1. Exponent α fest

$\alpha \neq 0, 1, 2$:

$$u(x) = u_j + u_j' \cdot z + u_j'' \left[\left(1 + \frac{z}{b_j}\right)^{\alpha} - 1 - \frac{\alpha}{b_j} z \right] \cdot \frac{b_j^2}{\alpha \cdot (\alpha - 1)};$$

$\alpha = 0$:

$$u(x) = u_j + u_j' z - b_j^2 \cdot u_j'' \left[\log\left(1 + \frac{z}{b_j}\right) - \frac{z}{b_j} \right]; \tag{25}$$

$\alpha = 1$:

$$u(x) = u_j + u_j' z + u_j'' \cdot b_j^2 \cdot \left[\left(1 + \frac{z}{b_j}\right) \log\left(1 + \frac{z}{b_j}\right) - \frac{z}{b_j} \right]$$

mit $z = x - x_j$ und $b_j = \dfrac{(\alpha - 2) \cdot u_j''}{u_j'''}$.

Der Wert $\alpha = 2$ führt zu quadratischen Polynomen, es sind nicht genug freie Parameter zur Erfüllung von vier Bedingungen vorhanden. Deshalb ist dieser Fall auszuschließen.

2. α variabel, also selbst Parameter

$\alpha \neq 0, 1$:

$$u(x) = u_j + (u_j')^2 \cdot \frac{\alpha_j - 1}{\alpha_j \cdot u_j''} \left[\left(1 + \frac{z \cdot u_j''}{(\alpha_j - 1) \cdot u_j'}\right)^{\alpha_j} - 1 \right]; \tag{26}$$

$\alpha = 0$:

$$u(x) = u_j + u_j' \cdot b_j \log\left(1 + \frac{z}{b_j}\right), \qquad \text{wobei } b_j = -\frac{u_j'}{u_j''}.$$

Für den Fall, daß die Nenner u_j''' bzw. u_j'' verschwinden, muß man im Programm des Algorithmus geeignete Vorkehrungen treffen. Man kann beispielsweise in einem Teilintervall auf kubische Splines umschalten.

Für diese Klassen von Splinefunktionen kann man die Bedingungen für die Lösbarkeit der Gleichung (15) zur Bestimmung des vierten Parameters vollständig übersehen. In WERNER und ZWICK [29] ist dies für die Quadratur im Einzelnen ausgeführt.

Einer der Fälle sei herausgegriffen. Für die Klasse mit variablen α ist der Exponent aus

$$u'(x_{j+1}, h) = f(x_{j+1})$$

zu bestimmen. Berechnet man die linksstehende Ableitung aus Formel (26), ersetzt z durch h, so findet man

$$\left(1 + \frac{1}{v}\right)^{v} = \left(\frac{f(x_{j+1})}{f(x_j)}\right)^{\frac{u_j'}{h \cdot u_j''}} =: A_j; \qquad \alpha_j = 1 + \frac{v \cdot h \cdot u_j''}{u_j'}. \tag{27}$$

Man verifiziert leicht, daß der Wertebereich der linksstehenden Funktion das Intervall $]1, \infty)$ ist. In der Regel liegt A_j in diesem Bereich. Nur wenn $f(x)$ verschwindet oder $u''(x, h)$ in $[x_j, x_{j+1}]$ *das Vorzeichen wechselt, kann* A_j kleiner als 1 ausfallen. Zu dieser direkt aus den bekannten Werten zu berechnenden Größe muß man v ermitteln.

Um den Definitionsbereich in ein zusammenhängendes Intervall zu verwandeln, setzt man

$$y = \frac{v}{1+v}$$

und kann als gute Näherungen für die Umkehrfunktion von

$$z = y^{\frac{y}{y-1}} = \left(1 + \frac{1}{v}\right)^v$$

die Ansätze

$$y = \begin{cases} \left(\dfrac{z-1}{e-1}\right)^c & \text{mit} \quad c = 2 \cdot \dfrac{e-1}{e} \quad \text{für} \quad 1 \leqq z \leqq 3{,}45 \\ z\left(1 - \dfrac{\log z}{z - 1{,}02} - \dfrac{0{,}13}{z}\right) & \text{für} \quad z > 3{,}45 \end{cases}$$

verwenden, die nach zwei, drei Iterationen auf zehnstellige Genauigkeit gebracht werden können. Entsprechende Ansätze erhält man auch für den Fall fester Exponenten α.

Mit den hier definierten Funktionenklassen sollen zwei Anfangswertprobleme behandelt werden. Die Nullstellen von $\left(1 + \dfrac{x - x_j}{b_j}\right)$ werden als Schätzungen für die Polstellen benutzt. Es ist nur eine Auswahl von Punkten aufgeführt.

1. $y' = 1 + y^2$, $y(0) = 1$.

Tabelle der Schätzwerte für die Lage des Poles
(Schrittweite $h = 0{,}125$)

x	$\alpha = -0.9$	-1.0	-1.1
0,000	0,735930	0,759675	0,783422
0,250	0,766229	0,782012	0,797797
0,500	0,778045	0,785257	0,792479
0,625	0,783749	0,785396	0,786985

Bei der Schrittweite $h = 0{,}03125$ bekommt man an der Stelle $x = 0{,}75$ die Schätzung 0,78539819, der exakte Wert wäre $\pi/4 = 0{,}78539816$.

Bemerkenswert ist, daß bei zu großem Betrag des Exponenten die Lage des Poles überschätzt und bei zu kleinem Exponenten unterschätzt wird, daß ein monotones Verhalten dieser Werte (insbesondere auch, wenn man noch mehr Punkte und kleinere Schrittweiten verwendet) unverkennbar ist, so daß man eine Einschließung des gesuchten Wertes bekommt, dies wird auch durch das folgende Beispiel demonstriert.

2. $y' = y^4 + x^2 \cdot y^6$, $y(0) = 1$.

Tabelle der Schätzwerte für die Lage des Poles
(Schrittweite $h = 0{,}03125$)

	α			
x	$-0{,}1$	$-0{,}2$	$-0{,}333333$	$-1{,}0$
0,000	0,280854	0,293697	0,310821	0,396460
0,0625	0,285900	0,295909	0,309254	0,375973
0,125	0,290847	0,298048	0,307650	0,355668
0,1875	0,295441	0,299837	0,305703	0,335078
0,250	0,299501	0,300960	0,302917	0,312849

Bei $\alpha = -0{,}2 = -\frac{1}{5}$(!) und $x = 0{,}2890625$ mit $h = 0{,}0078125$ wird der Pol zu 0,30104 geschätzt.

Schlußbemerkung

Die wesentlichen Vorteile der Splinefunktionen treten hervor, wenn Anfangswertprobleme bei Differentialgleichungen mit retardierten Argumenten zu lösen sind, wie sie beispielsweise in der Regeltechnik und den Biowissenschaften auftreten. Die Splinelösung kann während der Konstruktion der Lösung stückweise analytisch gespeichert und bei den weiteren Berechnungen mit beliebigem Argument aufgerufen werden.

Darüberhinaus haben diese Ausführungen versucht zu motivieren, daß es sich bei der Behandlung nichtlinearer Probleme lohnt, darüber nachzudenken, ob nicht durch Einsatz nichtlinearer Splines gute Lösungen mit ökonomischem Rechenaufwand erzielt werden können.

Literatur

1 Arndt, H., Interpolation mit regulären Spline-Funktionen, Dissertation, Münster 1974.
2 Arndt, H., Numerische Lösungen von gewöhnlichen Differentialgleichungen mit Block-Mehrschrittverfahren, eingereicht bei Numer. Math.

3 Barrar, B. R.; Loeb, H. L., Existence of best spline approximations with free knots, J. Math. Analysis Appl. **31**, 383–390 (1970).
4 Barrar, B. R.; Loeb, H. L., Spline functions with free knots as the limit of varisolvent families, J. Approximation Theory **12**, 70–77 (1974).
5 Baumeister, J., Extremaleigenschaften nichtlinearer Splines, Dissertation, München 1974.
6 Braess, D., Chebyshev approximation by spline functions with free knots, Numer. Math. **17**, 357–366 (1971).
7 Braess, D.; Werner, H., Tschebyscheff-Approximation mit einer Klasse rationaler Splinefunktionen. II, J. Approximation Theory **10**, 379–399 (1974).
8 Lambert, J. D., Computational Methods in Ordinary Differential for Equations, Wiley & Sons, New York-London 1973.
9 Lambert, J. D.; Shaw, B., On the numerical solution of $y' = f(x, y)$ by a class of formulae based on rational approximation, Math. Comput. **19**, 456–462 (1965).
10 Lambert, J. D.; Shaw, B., A method for the numerical solution of $y' = f(x, y)$ based on a self-adjusting non-polynomial interpolant, Math. Comput. **20**, 11–20 (1966).
11 Loscalzo, F. R.; Talbot, T. D., Spline function approximation for solutions of ordinary differential equations, SIAM J. Numer. Analysis **4**, 433–445 (1967).
12 Meder, G., Beiträge zur Formulierung und Konvergenz des Remes-Algorithmus für rationale Splinefunktionen, Dissertation, Münster 1975.
13 Micula, G., Bemerkungen zur numerischen Lösung von Anfangswertproblemen mit Hilfe nichtlinearer Splinefunktionen, In Lecture Notes in Mathematics **501**, 200–209, Spline Functions, Karlsruhe 1975, Springer-Verlag, Berlin-Heidelberg-New York 1976.
14 Popoviciu, T., Introduction â la théorie des différences divisées, Bull. Math. Soc. Roumaine Sci. **42**, 65–78 (1940).
15 Runge, R., Lösung von Anfangswertproblemen mit Hilfe nichtlinearer Klassen von Spline-Funktionen, Dissertation, Münster 1972.
16 Schaback, R., Spezielle rationale Splinefunktionen, J. Approximation Theory **7**, 281–292 (1973), Dissertation, Münster 1969.
17 Schaback, R., Interpolation mit nichtlinearen Klassen von Splinefunktionen, J. Approximation Theory **8**, 173–188 (1973).
18 Schomberg, H., Tschebyscheff-Approximation durch rationale Splinefunktionen mit freien Knoten, Dissertation, Münster 1973.
19 Späth, H., Vortrag auf der Tagung in Oberwolfach über Spline-Funktionen, April/May 1973.
20 Werner, H., Tschebyscheff-Approximation mit einer Klasse rationaler Splinefunktionen, J. Approximation Theory **10**, 74–92 (1974).
21 Werner, H., Tschebyscheff-Approximation nichtlinearer Splinefunktionen, in K. Böhmer – G. Meinardus – W. Schempp: Spline-Funktionen, BI-Verlag Mannheim-Wien-Zürich 1974, 303–313.
22 Werner, H., Interpolation and integration of initial value problems of ordinary differential equations by regular splines, SIAM J. Numer. Analysis **12**, 255–271 (1975).
23 Werner, H., An Introduction to regular splines and their application for initial value problems of ordinary differential equations, TR/53, Brunel University, Dept. of Mathematics 1975.
24 Werner, H., Numerische Behandlung gewöhnlicher Differentialgleichungen mit Hilfe von Splinefunktionen, ISNM **32**, 167–175 (1976).
25 Werner, H., Approximation by regular splines with free knots, Austin, Symposium on Approximation Theory 1976, S. 567–573.
26 Werner, H.; Loeb, H., Tschebyscheff-approximation by regular splines with free knots, in Approximation Theory, Bonn 1976, Lecture Notes in Math. **556**, 439–452, Springer-Verlag, Berlin-Heidelberg-New York 1976.
27 Werner, H.; Schaback, R., Praktische Mathematik II, Springer-Verlag, Berlin-Heidelberg-New York 1972.
28 Werner, H.; Wuytack, L., Nonlinear quadrature rules in the presence of a singularity, Universiteit Antwerpen, Dept. Wiskunde, 77–106 (1977).
29 Werner, H.; Zwick, D., Algorithms for numerical integration with regular splines, Rechenzentrum der Universität Münster Schriftenreihe Nr. **27** (1977).
30 Wuytack, L., Numerical integration by using nonlinear techniques, J. Comput. Appl. Math. **1**, 267–272 (1975).

Anschrift: Professor Dr. Helmut Werner, Institut für Numerische und instrumentelle Mathematik und Rechenzentrum der Universität Münster, Roxeler Str. 60/64, D-4400 Münster/Westf., BRD

Zusatz bei der Korrektur:

Einen Verbesserungsvorschlag von Herrn H. Arndt aufgreifend, kann man den Beweis von Hilfssatz 2 verkürzen, indem man die Zeilen 7 bis 19 auf Seite T 88 in folgender Weise abändert:

Betrachte X mit $L(X) = L_0$. Es sei z eine Stützstelle mit

$$x_{j_0} < z < x_{j_k} .$$

Wegen

$$\Delta^1(x_{j_0}, x_{j_k}) = \lambda \cdot \Delta^1(x_{j_0}, z) + \mu \cdot \Delta^1(z, x_{j_k}) \tag{7}$$

mit

$$\lambda = \frac{z - x_{j_0}}{x_{j_k} - x_{j_0}}, \qquad \mu = \frac{x_{j_k} - z}{x_{j_k} - x_{j_0}}$$

gilt

$$\begin{aligned} \Delta^k(x_{j_0}, \ldots, x_{j_k}) &= \Delta_t^1(x_{j_0}, x_{j_k})\, \Delta^{k-1}(t, x_{j_1}, \ldots, x_{j_{k-1}}) \\ &= \lambda \cdot \Delta^k(x_{j_0}, \ldots, x_{j_{k-1}}, z) + \mu \cdot \Delta^k(z, x_{j_1}, \ldots, x_{j_k}) . \end{aligned} \tag{8}$$

Die beiden Koeffizienten λ und μ sind positiv und ihre Summe ist gleich 1.