KfK 5276
März 1994

KfK-Seminarreihe Aktuelle Forschungsgebiete in der Mathematik

Seminarbeiträge 1993

R. Seifert, T. Westermann (Hrsg.)
Institut für Angewandte Informatik

Kernforschungszentrum Karlsruhe

KERNFORSCHUNGSZENTRUM KARLSRUHE

Institut für Angewandte Informatik

KfK 5276

KfK-Seminarreihe
Aktuelle Forschungsgebiete in der Mathematik
Seminarbeiträge 1993

Herausgeber: R. Seifert, T. Westermann *

* Fachhochschule Karlsruhe

Kernforschungszentrum Karlsruhe GmbH, Karlsruhe

Kernforschungszentrum Karlsruhe GmbH
Postfach 3640, 76021 Karlsruhe

ISSN 0303-4003

KfK-Seminarreihe: Aktuelle Forschungsgebiete in der Mathematik

Seminarbeiträge 1993

Zusammenfassung

Am 25. Mai 1993 begann im KfK eine Seminarreihe über aktuelle Forschungsgebiete in der Mathematik, die im engen Zusammenhang mit praxisbezogenen Anwendungen stehen. Ziel dieser Seminarreihe ist, im KfK anwendungsorientierte, aktuelle Forschungsthemen aus der Mathematik zu präsentieren. Übersichtsvorträge sollen Einblicke in moderne Methoden und Verfahren der Mathematik ermöglichen. Organisiert wurde die Seminarreihe von Rolf Seifert (KfK, IAI) und Thomas Westermann (FH Karlsruhe).

Im vorliegenden Bericht sind die Seminarbeiträge in schriftlicher Form zusammengefaßt.

KfK-Seminar series on Selected Topics in Mathematics

Seminar reports 1993

Summary

In May 25, 1993 a series of seminars was held at KfK on selected topics in applied mathematics. The aim was to demonstrate the importance of applied mathematics and to present current research areas in mathematics. Survey lectures should give an insight in modern methods and methodologies. The seminars were organized by Rolf Seifert (KfK, IAI) and Thomas Westermann (FH Karlsruhe).

This report contains the collection of the seminar papers.

Inhaltsverzeichnis

Vorwort

Am 25. Mai 1993 begann die Seminarreihe "**Aktuelle Forschungsgebiete in der Mathematik - Praxisbezogene Anwendungen**", die von Mitarbeitern des IAI und HDI initiiert wurde.

Ziel dieser Seminarreihe ist, im KfK anwendungsorientierte, aktuelle Forschungsthemen aus der Mathematik zu präsentieren. Gedacht ist an Übersichtsvorträge, die Einblicke in moderne Methoden und Verfahren der Mathematik ermöglichen. Mit dieser Seminarreihe soll auch der Kontakt zu externen Forschungseinrichtungen erweitert und vertieft werden, um weitere kompetente Wissenschaftler als Gesprächspartner für das Kernforschungszentrum zu gewinnen. Darüber hinaus soll aber auch ein breiter Gedankenaustausch innerhalb des KfK ermöglicht werden.

Die Zielgruppe für das Auditorium sind somit mathematisch-interessierte Mitarbeiter, die sich neuen mathematischen Verfahren und Methoden aufgeschlossen zeigen bzw. die selbst an mathematischen Fragestellungen arbeiten.

Die Seminarreihe startete im Sommersemester mit zwei Vorträgen, die im Bereich der numerischen Mathematik bzw. der Intervall-Arithmetik ihre Anwendung finden. Am 25. Mai begann das Seminar mit einem Vortrag von Dr. Dobner (Universität Karlsruhe), der über das Thema "Was ist schlecht an schlecht gestellten Problemen?" vorgetragen hat. Die Reihe wurde am 29. Juni mit einem Übersichtsvortrag von Prof. Dr. Halter (Fachhochschule Karlsruhe) fortgesetzt, der über verschiedene mathematische Verfahren für technische Feldprobleme berichtete.

Im Wintersemester 1993/94 wurde die Seminarreihe durch drei weitere Vorträge fortgesetzt. Die beiden ersten Vorträge befaßten sich mit statistischen Verfahren in der Mathematik: Prof. Dr. Zamir (Hebrew University of Jerusalem, Israel) berichtete am 29. Oktober über "Material Accountancy: A Game Theoretical Analysis" und Prof. Dr. Yakovlev (St. Petersburg Technical University, Rußland) über "Statistical Methods for the Carcinogenic Risk Assessment". Die Reihe schloß DM Hugelmann (HDI) mit einem Vortrag über die "numerische Lösung partieller Differentialgleichungen mittels finiter Differenzen " ab.

Die Seminarreihe wird im Jahr 1994 fortgesetzt. Für Anregungen, Themen- und Vortragsvorschläge sind die Organisatoren (Rolf Seifert (IAI), Tel. 07247/ 824411 und Thomas Westermann (FH Karlsruhe), Tel. 0721/169 253) stets offen.

Was ist schlecht an schlecht gestellten Problemen?

H.-J. Dobner

Mathematisches Institut II, Universität Karlsruhe

Zusammenfassung

In diesem Artikel wird die Frage untersucht, was das Schlechte an schlecht gestellten Problemen ist. Dazu wird zunächst erklärt, was ein schlecht gestelltes Problem ist. An Hand mehrerer Beispiele wird sodann das charakteristische solcher Probleme aufgezeigt. Weiter werden Möglichkeiten zur Behandlung und Lösung dieser Fragestellungen diskutiert. Hierbei wird vor allem die Tikhonov Regularisierung eingehender untersucht.

1. Begriffsbildung

Schlecht oder inkorrekt gestellten Problemen haftet irgendwie der Makel des Falschen oder Unrichtigen an. Inwieweit dies gerechtfertigt ist, soll in dieser Arbeit abgeklärt werden.

Die Bezeichnungsweise stammt von Hadamard [6] und geht zurück auf das Jahr 1923. Hadamards Konzept kann durch nachfolgende Definition präzisiert werden.

Definition 1.1

Es seien X, Y normierte Räume und $A: X \to Y$. Das Problem

$$A(f) = g \tag{1.1}$$

heißt gut oder korrekt gestellt genau dann, wenn es folgende drei Eigenschaften hat

(1.2) Existenz: Für jede Inhomogenität $g \in Y$ existiert mindestens eine Lösung $f \in X$.

(1.3) Eindeutigkeit: Zu jedem $g \in Y$ existiert höchstens eine Lösung $f \in X$.

(1.4) Stabilität: Die nach (1.2) und (1.3) eindeutige Lösung f hängt stetig von der rechten Seite g ab, m.a.W. wird die Inhomogenität nur "wenig" geändert, so ändert sich auch die entsprechende Lösung nur "geringfügig".

Im anderen Fall, wenn eine dieser drei Forderungen verletzt ist, heißt (1.1) schlecht oder inkorrekt gestellt.

Hadamard war der Meinung, daß jede korrekte mathematische Formulierung einer sinnvollen physikalischen Fragestellung zu einer mathematischen Gleichung führen muß, welche den Bedingungen (1.2) – (1.4) genügt.

Entsprechend vorstehender Definition muß man verschiedene Arten der Schlechtgestelltheit unterscheiden und zwar

- Nichtexistenz
- Nichteindeutigkeit
- Instabilität

Vor allem dem letzten Punkt der extremen Verstärkung von Datenfehlern in der Lösung gilt das Hauptaugenmerk bei der Behandlung inkorrekt gestellter Fragestellungen, denn die Nichtexistenz läßt sich durch Ausweitung des Lösungsbegriffs und die Nichteindeutigkeit durch das Vorschreiben von Normierungsbedingungen meist erzwingen.

Bevor wir tiefer in die Behandlung inkorrekt gestellter Probleme einsteigen, werden wir uns deren charakterisierendes Verhalten an Hand einiger ausgewählter Beispiele verdeutlichen.

2. Beispiele

2.1 Lineares Gleichungssystem

(2.1) $$Ax = b$$

mit $A \in \mathbb{R}^{n\times n}$, $x, b \in \mathbb{R}^n$ und

(2.2) $$\det(A) = \varepsilon > 0 \ , \ \varepsilon \ll 1 \ .$$

Ist die rechte Seite lediglich näherungsweise bekannt, so kann die damit berechnete Lösung von der tatsächlichen sehr stark abweichen, auf Grund der fehlenden Stabilität haben

wir es mit einem schlecht gestellten Problem zu tun. So wird beispielsweise bei dem Gleichungssystem

$$\begin{pmatrix} \varepsilon & & 0 \\ & \ddots & \\ 0 & & \varepsilon \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \tag{2.3}$$

ein Fehler in b mit dem Faktor $\frac{1}{\varepsilon}$ verstärkt. Sind sowohl rechte Seite b als auch die Matrix A nur näherungsweise bekannt, d.h. liegt in Wirklichkeit nicht das System $Ax = b$, sondern das gestörte System $\tilde{A}x = \tilde{b}$ vor, kann es zusätzlich vorkommen, daß die Lösung nicht eindeutig ist, eventuell nicht einmal existiert, so daß wir es auch in dieser Situation mit einem inkorrekt gestellten Problem zu tun haben.

2.2 Anfangswertproblem für harmonische Funktionen

Dieses Problem wurde von Hadamard als Beispiel eines inkorrekt gestellten Problems angegeben:

$$\begin{aligned} u_{xx}(x,y) + u_{yy}(x,y) &= 0, \ (x,y) \in \mathbb{R} \times [0,\infty) \\ u(x,0) = 0, \ u_y^{[n]}(x,0) &= \frac{1}{n}\sin(nx), \ x \in \mathbb{R}, n \in \mathbb{N} \ . \end{aligned}$$

Die Anfangsdaten konvergieren gleichmäßig gegen Null

$$u_y^{[n]}(x,0) \Longrightarrow 0, \ n \to \infty,$$

wohingegen für die zugehörigen eindeutig bestimmten Lösungen

$$u^{[n]}(x,y) = \frac{1}{n^2}\sin(nx)\sinh(ny)$$

für kein $y > 0$ Konvergenz vorliegt. Wir erkennen für große n, daß obwohl sich die Anfangswerte nur wenig ändern, die Abweichung der entsprechenden Lösungsfunktionen exponentiell anwächst, also Forderung (1.4) unerfüllt ist.

Zur weiteren Verdeutlichung der bei schlecht gestellten Problemen auftretenden Phänomene diene das folgende Modellbeispiel (vgl. dazu auch Louis [9]).

2.3 Modellbeispiel

Eine Black Box transformiert das konstante Eingangssignal 1 in ein Ausgangssignal g. Nimmt man zur Identifizierung dieser Black Box an, daß sie linear und kausal ist, so läßt sie sich als Volterrasche Integralgleichung 1. Art modellieren:

$$A(f(x)) := \int_0^x 1 \cdot f(t)dt = g(x),\ 0 \leq x \leq a,\ a \in \mathbb{R}, \tag{2.4}$$

die Funktion f beschreibt also die Black Box. Ist $g \in C^1[0,a]$ und weiterhin die Lösbarkeitsbedingung

$$g(0) = 0 \tag{2.5}$$

erfüllt, so ist die eindeutig existierende Lösung von (2.4) gegeben als

$$\bigwedge_{x \in [0,a]} f(x) = g'(x). \tag{2.6}$$

Werden die Daten g nur ein wenig abgeändert, etwa:

$$g(0) = \gamma \neq 0,$$

so ist (2.5) und somit (1.2) verletzt. Ein weiterer Effekt tritt in Erscheinung, wenn lediglich Näherungswerte für die Ausgangsdaten vorliegen. Sei g gestört zu

$$g^\delta(x) = g(x) + \delta \sin(nx),\ x \in [0,a],$$

dann bleibt g^δ weiterhin differenzierbar und erfüllt die Lösbarkeitsforderung. Für den Datenfehler (in der Maximumnorm) gilt

$$\|g^\delta - g\| \leq \delta;$$

der Ergebnisfehler

$$\|f^\delta - f\| \leq n\delta,$$

ist abhängig von n und kann daher beliebig groß ausfallen, so daß bei dieser Konstellation zwar (1.2), (1.3) nicht aber (1.4) zutreffen.

Als nächstes wenden wir uns den inversen Problemen zu und erschließen auf diese Weise eine Klasse inkorrekt gestellter Fragestellungen.

3. Inverse Probleme

Ist die direkte Messung der Eigenschaften eines Objektes nicht möglich, sondern muß man von indirekten Beobachtungen auf diese Größe zurückschließen, so spricht man von einem inversen Problem. Eine inverse Aufgabe läßt sich auch durch die Kurzformel

"Wirkung beobachtet, Ursache gesucht"

beschreiben. Da es eine exakte Definition dieses Begriffs nicht gibt, soll der Unterschied zwischen direktem und inversem Problem an Hand der Wärmeleitungsgleichung

$$u_t(x,t) = u_{xx}(x,t),\ (x,t) \in [0,\pi] \times [0,T],\ T \in \mathbb{R}, \tag{3.1}$$

gegenüberstellend verdeutlicht werden (vgl. Kreß [7]).

direktes Problem	inverses Problem

die Randvorgaben sind in beiden Fällen die gleichen:

$$\bigwedge_{t\in[0,T]} (u(0,t) = 0 \ \wedge \ u(\pi,t) = 0).$$

Ausgehend von der bekannten Anfangstemperaturverteilung $f(x)$, $x \in [0,\pi]$, zum Zeitpunkt $t = 0$, soll die Endtemperatur $u(x,T) = g(x)$, $x \in [0,\pi]$ ermittelt werden.

Ausgangspunkt ist die gemessene Endtemperatur $g(x)$, $x \in [0,\pi]$, zum Zeitpunkt $t = T$; damit soll die Anfangstemperatur $u(x,0) = f(x)$ berechnet werden.

Durch einen Separationsansatz, wobei c_n bzw. d_n die Fourierkoeffizienten bedeuten, erhält man jeweils die eindeutig bestimmte Lösung:

$$u(x,t) = \sum_{n=1}^{\infty} c_n e^{-n^2 t} \sin(nx),$$

wobei

$$(x,t) \in [0,\pi] \times [0,T].$$

$$u(x,t) = \sum_{n=1}^{\infty} d_n e^{n^2(T-t)} \sin(nx),$$

wobei

$$(x,t) \in [0,\pi] \times [0,T].$$

Legen wir den Raum $L^2[0,T]$ zugrunde, so erhält man:

$$\|g\|_2^2 \le e^{-2T} \|f\|_2^2.$$

Hier ist die Endtemperatur stetig von der Anfangstemperatur abhängig, das Problem somit gut gestellt im Sinne von Definition 1.1.

$$\bigwedge_{C\in(0,\infty)} \|f\|_2^2 > C\|g\|_2^2,$$

Die Lösung des inversen Problems besteht in der Interpretation der Daten g, also der Konstruktion des Urbildes, dieses hängt nicht in stetiger Weise von g ab, das Problem folglich inkorrekt gestellt.

Charakteristisch für inverse Probleme ist, daß sie im Sinne von Hadamard schlecht gestellt sind.

Inverse Probleme spielen mittlerweile eine wichtige Rolle in der industriellen Praxis, was auch dazu geführt hat, sich intensiver mit der Lösung schlecht gestellter Aufgaben zu befassen. Bevor wir uns der Behandlung schlecht gestellter Probleme zuwenden, sollen zuerst noch einige praxisrelevante inverse (und damit inkorrekte) Probleme vorgestellt werden, um deren Bedeutung hervorzuheben.

4. Inverse Probleme in der Praxis

Es sollen hier einige Probleme beschrieben werden, die allesamt inverse Fragestellungen repräsentieren. Die Modellbildung, der Weg vom physikalischen Problem hin zur mathematischen Gleichung, würde die Zielsetzung dieser Abhandlung übersteigen. Wir verweisen daher auf die Literatur und begnügen uns mit einer kurzen Charakterisierung der einzelnen Aufgaben.

Bei der **Computer Tomographie** geht es darum, auf nichtinvasive Art und Weise Einblick in die Morphologie von Patienten zu gewinnen (s. Natterer [10]). Ein Gebiet wird von Röntgenstrahlen durchlaufen. Aus der Veränderung der Intensität der Strahlen wird die Gewebedichte in dem von Röntgenstrahlen durchlaufenen Gebiet bestimmt und so ein Bild von einem Schnitt durch den Patienten erzeugt. Mathematisch wird dies durch eine Integralgleichung 1. Art – die Radon Integralgleichung – dargestellt.

Auf ähnliche Probleme führt die **Laufzeitanalyse in der Seismik.** Ausgangssignal sind künstlich erzeugte seismische Wellen. Aus Laufzeitmessungen kann auf Formationen im Erdinnern (Ölfelder o.ä.) geschlossen werden. Auch dieser Prozeß wird durch eine Integralgleichung 1. Art beschrieben (vgl. Fawcett [4]).

In der Hydrologie geht man bei der **Bestimmung eines Diffusionskoeffizienten** von einem Gleichgewichtszustand aus und versucht daraus die Durchläßigkeit eines porösen Mediums zu berechnen. Dieses inverse Problem der Grundwasserströmung läßt sich als hyperbolische Differentialgleichung 1. Art auffassen (Richter [12]).

Aus der industriellen Praxis entstammt die Frage des **Reflektordesigns.** Gesucht ist der Entwurf eines Reflektors, so daß sich die Lichtverteilung in Ebene und Sphäre in vorgegebener Weise verhält. Mathematisch gesehen verbirgt sich hinter dieser Fragestellung eine nichtlineare partielle Differentialgleichung 2. Ordnung.

Das Auffinden von zylindrischen Armierungseisen im Stahlbeton läßt sich mathematisch als Integralgleichung 1. Art modellieren. Darauf aufbauend kann eine baustellentaugliche Methode zur **Lokalisierung von Armierungseisen** entwickelt werden.

In der Stahlindustrie hat man es häufig mit dem Problem zu tun, die **Wärmeverteilung innerhalb eines Stahlstrangs** zu ermitteln. Allerdings kann die Temperatur nur teilweise am Rand exakt gemessen werden. Die daraus resultierende nichtlineare parabolische Differentialgleichung ist schlecht gestellt (vgl. Engl [3]).

Auf eine Abelsche Integralgleichung 1. Art stößt man in der **Stereologie:** Eine Menge Kugeln mit unbekannter Radienverteilung ist in einem Volumen verteilt. Man legt nun endlich viele Schnitte durch dieses Volumen und erhält so Kreisscheiben. Aus der Verteilung dieser Radien (für endlich viele Schnitte) versucht man auf die Verteilung der Kugelradien zu schließen.

Eine umfangreiche Subklasse inverser Probleme bilden **inverse Streuprobleme.**
Aus Messungen von an Körpern gestreuten elektromagnetischen oder akustischen Wellen ist man bestrebt, auf Eigenschaften jener Objekte zurückzuschließen. Vor allem in der nichtzerstörenden Materialprüfung finden solche Verfahren ihre Anwendung. Auch hier können viele Probleme als Integralgleichungen 1. Art geschrieben werden.

5. Die Behandlung schlecht gestellter Probleme

Wie zuvor gesehen, führen zahlreiche korrekt mathematisch modellierte praxisrelevante Aufgaben auf schlecht gestellte Probleme. Wir werden uns jetzt damit auseinandersetzen, wie man solche Probleme behandelt. Um die methodischen Aspekte deutlich zu machen, treffen wir folgende Voraussetzungen:

X, Y Hilberträume

$A : X \to Y$ kompakt, linear, injektiv

Das Problem

$$A(f) = g, \; g \in Y, \tag{5.1}$$

sei schlecht gestellt.

Dabei ist die Forderung nach Injektivität keine prinzipielle Einschränkung der Allgemeinheit, da bei Homomorphismen die Eindeutigkeit durch entsprechende Modifikation des Lösungsraumes X gewährleistet werden kann. Ist X ein unendlichdimensionaler Raum, so bedeutet dies, daß die Umkehrabbildung $A^{-1}: A(X) \to X$ unbeschränkt ist und man somit vor die Aufgabe gestellt wird, eine Approximation für die unbeschränkte Inverse A^{-1} anzugeben. Dazu bedient man sich der Methode der Regularisierung.

Definition 5.1

Eine Regularisierung ist ein Verfahren zur Gewinnung einer stabilen Näherungslösung von (5.1), genauer: eine Regularisierung ist eine mit einer positiv reellen Zahl α parametrisierte Familie beschränkter linearer Operatoren R_α

(5.2) $$R_\alpha: Y \to X,\ \alpha > 0,$$

so daß

(5.3) $$\bigwedge_{\varphi \in X} \lim_{\alpha \to 0} R_\alpha(A(\varphi)) = \varphi.$$

Bemerkung 5.1

Legen wir $\dim X = \infty$ zugrunde, so kann wegen der Kompaktheit von A

- R_α nicht gleichmäßig beschränkt bezüglich α sein.
- $R_\alpha A$ nicht normkonvergent für $\alpha \to 0$ sein.

Wie in der Praxis häufig, wird die Lösung f von (5.1) ausgehend von einer gestörten rechten Seite g^δ mit bekanntem Fehlerniveau δ ermittelt:

(5.4) $$\|g^\delta - g\| \leq \delta.$$

Statt der ungestörten Lösung f erhält man jetzt die regularisierte Lösung $f_\alpha^\delta := R_\alpha(g^\delta)$. Ziel ist es, die Regularisierung derart zu konstruieren, daß f_α^δ stetig von g^δ abhängt. Für den Rekonstruktionsfehler $\|f_\alpha^\delta - f\|$ gilt:

(5.5) $$\|f_\alpha^\delta - f\| \leq \delta \|R_\alpha\| + \|R_\alpha(A(f)) - f\|,$$

er besteht demzufolge aus zwei Anteilen: Datenfehler $\delta\|R_\alpha\|$ und Regularisierungsfehler $\|R_\alpha(A(f)) - f\|$. Um das unterschiedliche Verhalten der beiden Fehler kennenzulernen, betrachten wir noch einmal Beispiel 2.3.

Beispiel 5.1

Gilt für die Inhomogenität in (2.4) $g \in C^3[0,a]$, so ist eine Regularisierung von (2.4) gegeben durch

(5.6) $$R_\alpha(g(x)) = \frac{1}{2\alpha}(g(x+\alpha) - g(x-\alpha)),\ \alpha \in (0,\infty).$$

Stehen nur gestörte Daten g^δ zur Verfügung, so erhält man für die Datenfehler die Abschätzung

(5.7) $$|R_\alpha((g^\alpha - g)(x))| \leq \frac{\delta}{\alpha},$$

und dieser Anteil wird mit größer werdendem α kleiner. Will man den Regularisierungsfehler

(5.8) $$\|R_\alpha(A(f)) - f\| \leq \frac{\alpha^2}{6} \cdot \|f''\|_\infty,$$

verkleinern, so muß man den Parameter α ebenfalls verkleinern. Aus (5.7) und (5.8) liest man ab, daß der Gesamtfehler nicht beliebig klein gemacht werden.

Tabellarisch kann das Verhalten dieser beiden Fehleranteile folgendermaßen erfaßt werden:

	Datenfehler	Regularisierungsfehler
$\alpha \to 0$	↗	↙
$\alpha \to \infty$	↙	↗

Damit läßt sich das für alle schlecht gestellten Probleme charakteristische Fehlerverhalten skizzieren.

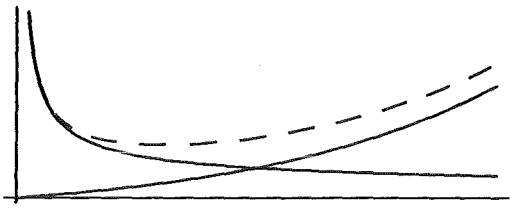


Abb. 1: Gesamtfehler (gestrichelt)

Somit erfordert jedes Regularisierungsverfahren eine eigene Strategie, um den Parameter α geeignet zu wählen.

6. Konstruktion von Regularisierungsverfahren

Wir behalten die Voraussetzungen des vorherigen Abschnittes bei.

Definition 6.1

Die singulären Werte μ_n von A sind die positiven Wurzeln aus den fallend angeordneten Eigenwerten des nichtnegativ, selbstadjungierten Operators A^*A, wobei diese entsprechend ihrer geometrischen Vielfachheit vorkommen.

Bemerkung 6.1

Für die Singulärwerte μ_n gilt

$$\lim_{n\to\infty} \mu_n = 0. \tag{6.1}$$

Definition 6.2

Ein singuläres System (μ_n, u_n, v_n) von A besteht aus den singulären Werten μ_n von A, sowie zwei Orthonormalfolgen $u_n \in X$, $v_n \in Y$, $n \in \mathbb{N}$, mit

$$Au_n = \mu_n v_n, \tag{6.2}$$

$$A^*v_n = \mu_n u_n. \tag{6.3}$$

Die Lösbarkeit von (1.1) läßt sich durch den Satz von Picard beantworten:

Satz 6.1

Es seien X, Y Hilberträume mit Innenprodukt $<,>$[1], $A: X \to Y$ kompakt, injektiv, linear. Die Gleichung

$$A(f) = g \tag{6.4}$$

ist genau dann lösbar, wenn

$$g \in N(A^*)^{\perp} \tag{6.5}$$

und

[1] Für das Innenprodukt in X bzw. Y verwenden wir das gleiche Symbol $<,>$. Aus dem Kontext ist ersichtlich, welches jeweils gemeint ist.

(6.6) $$\sum_{n=1}^{\infty} \frac{1}{\mu_n^2} | < g, v_n > |^2 < \infty$$

erfüllt sind. In diesem Falle ist die Lösung f von (6.4) darstellbar in der Form

(6.7) $$f = \sum_{n=1}^{\infty} \frac{1}{\mu_n} < g, v_n > u_n.$$

Beweis.

vgl. Kreß [7] □

Bemerkung 6.2

Die Glattheitsforderung (6.6) ist eine Bedingung an das Abklingverhalten der Fourierkoeffizienten von g bezüglich v_n.

Bemerkung 6.3

Der Darstellung (6.7) entnimmt man, daß Fehler in g mit dem Faktor $\frac{1}{\mu_n}$ verstärkt werden, so daß man daraus eine Klassifizierung der Schlechtgestelltheit ableiten kann:

> Ein Problem heißt stärker inkorrekt als ein anderes, wenn seine Singulärwerte schneller als die des anderen fallen.

Es bietet sich an, durch Abschwächen der großen Werte $\frac{1}{\mu_n}$, die Lösung (6.7) zu regularisieren, man erhält auf diese Weise

$$f_\alpha = \sum_{n=1}^{\infty} \frac{1}{\mu_n} F_\alpha(\mu_n) < g, v_n > u_n.$$

Die Funktion $F_\alpha(\mu_n)$ wird als Filter bezeichnet.

Satz 6.2

Genügt der Filter F_α, $\alpha \in (0,\infty)$, den Forderungen

$$\bigwedge_{\alpha>0} F_\alpha \quad \text{beschränkt,} \tag{6.8}$$

$$\lim_{\alpha\to 0} F_\alpha(\mu_n) = 1, \tag{6.9}$$

$$\bigwedge_{\alpha>0} \bigvee_{c(\alpha)>0} \bigwedge_{n\in\mathbb{N}} \left| F_\alpha(\mu_n)\frac{1}{\mu_n}\right| \le c(\alpha), \tag{6.10}$$

so ist die Familie der beschränkten, linearen Operatoren

$$R_\alpha(g) = f_\alpha = \sum_{n=1}^{\infty} \frac{1}{\mu_n} F_\alpha(\mu_n) < g, v_n > u_n, \quad g \in Y \tag{6.11}$$

ein Regularisierungsverfahren mit

$$\|R_\alpha\| \le c(\alpha). \tag{6.12}$$

Durch Wahl der Dämpfungsfunktion erhält man verschiedene Regularisierungsverfahren, die Daten der bedeutendsten Regularisierungen sind hier tabellarisch zusammengefaßt.

Regularisierung	$F_\alpha(\mu_n)$	$R_\alpha(g)$	Parameterbereich	$c(\alpha)$ aus (6.10)
Abgeschnittene Singulärwert-zerlegung	$\begin{cases} 1, & \mu_n \ge \alpha \\ 0, & \mu_n < \alpha \end{cases}$	$\sum_{\mu_n^2 \ge \alpha} \frac{1}{\mu_n} < g, v_n > u_n$	$\alpha \in (0,\infty)$	$\frac{1}{\sqrt{\alpha}}$
Landweber-Fridmann-Iteration	$1-(1-\omega\mu_n^2)^{\alpha+1}$, $0 \le \omega \le 1/\|\|A\|\|^2$	$\omega \sum_{k=0}^{\alpha} (I-\omega A^*A)^k A^*$	$\alpha \in \mathbb{N}$	$\sqrt{\omega(\alpha+1)}$
Tikhonov Regularisierung	$\frac{\mu_n^2}{\mu_n^2+\alpha}$	$(\alpha I + A^*A)^{-1}A^*g$	$\alpha \in (0,\infty)$	$\frac{1}{2\sqrt{\alpha}}$

Eine entscheidende Bedeutung bei der Anwendung all dieser Verfahren kommt der Wahl eines geeigneten Regularisierungsparameters zu, damit werden wir uns jetzt beschäftigen.

7. Wahl des Regularisierungsparameters

Der Regularisierungsparameter kann auch nach dem "trial and error" Prinzip ausgewählt werden. An Hand der bedeutsamsten Regularisierungsmethode – der Tikhonov Regularisierung – soll jedoch eine systematischere Vorgehensweise entwickelt werden. Dazu studieren wir die Tikhonov Regularisierung unter dem Gesichtspunkt einer Residuen Minimierung mit Strafterm. Es gilt

Satz 7.1

Es seien X, Y Hilberträume. $A: X \to Y$ kompakt, linear, injektiv und

$$A(f) = g \tag{7.1}$$

schlecht gestellt.

Dann gilt

$$\bigwedge_{\alpha>0} \bigwedge_{g\in Y} \overset{1}{\bigvee_{f_\alpha \in X}} \|A(f_\alpha) - g\|^2 + \alpha \|f_\alpha\|^2 \tag{7.2}$$

$$= \inf_{f\in X} \{\|A(f) - g\|^2 + \alpha \|f\|^2\}.$$

Dabei ist f_α gegeben als eindeutige Lösung von

$$(\alpha I + A^*A)(f_\alpha) = A^*(g), \tag{7.3}$$

und hängt dabei stetig von g ab.

Beweis.

s. Kreß [7]. □

Das quadratische Optimierungsproblem (7.2) läßt sich als penalty Methode eines der beiden Optimierungsprobleme mit Nebenbedingung interpretieren:

<u>Entweder</u>

Minimiere Residuum $\|A(f) - g\|$ durch Einschränken der Lösungsmannigfaltigkeit. Aus Gründen der Übersichtlichkeit beziehen wir uns hier darauf, daß die Norm von f beschränkt ist

$$\|f\| \leq \rho.$$

Man erhält eine sogenannte Quasi-Lösung f_q. Dieses Konzept geht auf Ivanov zurück

oder

Minimiere $\|f\|$ unter der Nebenbedingung, daß die Defektnorm beschränkt ist:

$$\|A(f) - g\| \leq \sigma,$$

diese auf Morozov zurückgehende Idee führt zur sogenannten Diskrepanz-Lösung f_d.

Die jeweilige Ermittlung des Regularisierungsparameters stellen wir für beide Prinzipien in Übersichtsform zusammen, wobei wir voraussetzen, daß $A(X)$ dicht in Y liegt.

8. Allgemeine Situation

Wir verzichten jetzt auf die Kompaktheit des Operators A, stattdessen betrachten wir (5.1) unter der Voraussetzung, daß, wenn X, Y Hilberträume bezeichnen, $A: X \to Y$ linear und stetig ist.

Man verallgemeinert den Lösungsbegriff auch für nicht aus dem Bildraum stammende Elemente, indem man den Defekt $\|A(f) - g\|$ minimiert und unter allen minimierenden Elementen dasjenige mit kleinster Norm auswählt. Das führt auf die verallgemeinerte Inverse.

Definition 8.1

Die verallgemeinerte Inverse oder Moore-Penrose Inverse $A^\dagger$ ist eine Abbildung definiert durch

$$(8.1) \quad \begin{cases} A^\dagger: D(A^\dagger) := R(A) \oplus R(A)^\perp \to X \\ \qquad\qquad g \mapsto A^\dagger(g) =: f, \\ \text{wobei} \\ \qquad f := \min\limits_{\|\varphi\|}\{\varphi \in M\}, M := \{\varphi \in X: \|A(\varphi) - g\| = \min\limits_{\psi \in X} \|A(\psi) - g\|\} \\ f \quad \text{heißt Moore-Penrose Lösung.} \end{cases}$$

Satz 8.1

Die Moore-Penrose Lösung $f = A^\dagger(g)$ ist die eindeutig bestimmte Lösung der Gleichung

Bestimmung des Regularisierungsparameters

	Quasi-Lösung f_q	Diskrepanz-Lösung f_d
Zu minimierende Größe	$\|A(f)-g\|$	$\|f\|$
Nebenbedingung	$\|f\| \leq \rho$, ρ a priori Information über Norm der Lösung	$\|A(f)-g\| \leq \sigma$, σ Schranke für Fehler in g
Grundidee	Stabilisierung durch Einschränkung der Lösungsmannigfaltigkeit	Es ist nicht sinnvoll, das Residuum kleiner als den Fehler in g zu machen. Stabilisierung wird durch Minimierung der Norm der Lösung erreicht.
Interpretation der Lösung	$A(f_q)$ Bestapproximation aus $A(K_\rho(0))$ an $g \in Y$. $K_\rho(0) := \{f \in X : \|f\| \leq \rho\}$	f_d Bestapproximation aus $\{f : \|A(f)-g\| \leq \sigma\}$ *an* $0 \in X$.
Bestimmung der Lösung	1) $g \in A(K_\rho(0))$: $f_q = 0$ 2) $g \notin A(K_\rho(0))$: Es gilt $\bigvee_{\alpha_q>0} G(\alpha_q) := \|f_{\alpha_q}\|^2 - \rho^2 = 0$ $\longrightarrow$ Newton-Verfahren zur Bestimmung von α_q: $\alpha_q^{(\nu+1)} = \alpha_q^{(\nu)} - G'(\alpha_q^{(\nu)})^{-1} G(\alpha_q^{(\nu)})$ Die dabei benötigten Größen $f\alpha_q^{(\nu)}$, $\frac{df\alpha_q^{(\nu)}}{d\alpha_q^{(\nu)}}$ bestimmt man mit Hilfe von (7.3).	1) $\|g\| \leq \sigma$: $f_d = 0$ 2) $\|g\| > \sigma$: Es gilt $\bigvee_{\alpha_d>0} H(\alpha_d) := \|A(f_{\alpha_d})-g\|^2 - \sigma^2 = 0$ $\longrightarrow$ Newton-Verfahren zur Bestimmung von α_d: $\alpha_d^{(\nu+1)} = \alpha_d^{(\nu)} - H'(\alpha_d^{(\nu)})^{-1} H(\alpha_d^{(\nu)})$ Die dabei benötigten Größen $f\alpha_d^{(\nu)}$, $\frac{df\alpha_d^{(\nu)}}{d\alpha_d^{(\nu)}}$ bestimmt man mit Hilfe von (7.3).
Stabilität	f_q hängt schwach stetig von g ab.	f_d hängt schwach stetig von g ab.

(8.2) $$A^*(A(f)) = A^*(g)$$

in $\overline{R(A^*)}$.

Beweis.

vgl. Louis [9] □

Bemerkung 8.1

Ist A ein linearer, kompakter Operator, so läßt sich f (und damit $A^\dagger$) mit Hilfe der Singulärwertzerlegung explizit in der Form (6.7) darstellen.

Die Moore-Penrose Inverse $A^\dagger$ ist aber i.a. unstetig, so daß noch gemäß Abschnitt 5, eine Regularisierung durchgeführt werden muß.

9. Einschließungsverfahren für schlecht gestellte Probleme

Unter dem Begriff E-Verfahren oder Einschließungsverfahren werden numerische Algorithmen subsummiert, welche die zusätzliche Qualitätseigenschaft aufweisen:

> Zusammen mit der numerisch berechneten Lösung werden mathematisch garantierte (enge) Fehlerschranken angegeben.

Die wichtigsten E-Verfahren zugrundeliegenden Konzepte werden nachfolgend stichpunktartig skizziert. Für eine weitergehende Darstellung sei auf den Sammelband von Kulisch [8] verwiesen.

Ultra-Arithmetik: Reihenentwicklung von Funktionen als arithmetische Methode.

Funktoid: Endliche Reihendarstellung von Funktionen mit Erfassung des Abschneidefehlers in Verbindung mit Ultra-Arithmetik.

Intervallfunktoid: Spezielles Funktoid, wobei die Objekte Funktionenmengen sind, deren Graph zwischen zwei Grenzfunktionen liegt.

Modifizierte Fixpunktsätze: Fixpunktsätze, die derart modifiziert sind, daß ihre Voraussetzungen von Rechnern überprüfbar sind.

In diesem Abschnitt sollen einige grundsätzliche Überlegungen angestellt werden, wie schlecht gestellte Aufgaben einschließungsmäßig behandelt werden können. Hierbei verstehen wir unter Einschließung der Lösung eines schlecht gestellten Problems stets die Einschließung der regularisierten Moore Penrose Lösung.

Cum grano salis wird durch eine Regularisierung eine Gleichung 1. Art in ein Problem 2. Art transformiert. Für zahlreiche Probleme 2. Art existieren jedoch effiziente Verifikationsalgorithmen. Die zur Einschließung nötigen Schritte sollen an Hand der Fredholmschen Integralgleichung 1. Art

$$A(f(s)) := \int_a^b k(s,t)f(t)dt = g(s) \quad , \tag{9.1}$$

basierend auf der Tikhonov Regularisierung mit Parameterwahl gemäß Abschnitt 7 beschrieben werden. Als Gleichung für die regularisierte Lösung erhält man die parameterabhängige Fredholmsche Integralgleichung 2. Art

$$\alpha f_\alpha(s) + \int_a^b \int_a^b k(\tau,s)k(\tau,t)f_\alpha(t)d\tau dt = \int_a^b k(s,t)g(t)dt, \quad a \le s \le b. \tag{9.2}$$

Will man Quasi-Lösungen behandeln, ist wie folgt zu verfahren (vgl. Abschnitt 7, der besseren Lesbarkeit wegen werden die Indizes q bzw. d weggelassen). Wähle $\alpha^{(0)}$, so daß

$$\alpha^{(0)} \cdot \rho \le \|A\|\delta. \tag{9.3}$$

Iteriere gemäß

$$\alpha^{(\nu+1)} = \alpha^{(\nu)} - (G'(\alpha^{(\nu)}))^{-1}G(\alpha^{(\nu)}), \quad \nu = 0,1,\ldots, \tag{9.4}$$

dabei ist

$$G(\alpha^{(\nu)}) = \|f_{\alpha^{(\nu)}}\|^2 - \rho^2, \tag{9.5}$$

$$G'(\alpha^{(\nu)}) = 2\mathrm{Re} < \frac{df_{\alpha^{(\nu)}}}{d\alpha^{(\nu)}}, \quad f_{\alpha^{(\nu)}} >, \tag{9.6}$$

und $\frac{df_{\alpha^{(\nu)}}}{d\alpha^{(\nu)}}$ genügt der Gleichung 2. Art

$$\alpha^{(\nu)}\frac{df_{\alpha^{(\nu)}}}{d\alpha^{(\nu)}} + \int_a^b \int_a^b k(\tau,s)k(\tau,t)\frac{df_{\alpha^{(\nu)}}}{d\alpha^{(\nu)}}d\tau dt = -f_{\alpha^{(\nu)}}, \tag{9.7}$$

$$f_{\alpha^{(\nu)}} \quad \text{aus (9.2).} \tag{9.8}$$

Bei Anwendung des Diskrepanz-Prinzips ergibt sich

Starte mit $\alpha^{(0)}$ gemäß

$$\alpha^{(0)}(\|g^\delta\| - \delta) \leq \|A\|^2 \cdot \delta. \tag{9.9}$$

Iteration nach

$$\alpha^{(\nu+1)} = \alpha^{(\nu)} - (H'(\alpha^{(\nu)}))^{-1} H(\alpha^{(\nu)}), \quad \nu = 0, 1, \ldots, \tag{9.10}$$

wobei

$$H(\alpha^{(\nu)}) = \|A(f_{\alpha^{(\nu)}}) - g\|^2 - \sigma^2 \tag{9.11}$$

$$H'(\alpha^{(\nu)}) = - < \frac{df_{\alpha^{(\nu)}}}{d\alpha^{(\nu)}}, \int_a^b k(t,s) f_{\alpha^{(\nu)}}(t) dt > \tag{9.12}$$

$$- \|f_{\alpha^{(\nu)}}\|^2 - 2\alpha^{(\nu)} \mathrm{Re} < \frac{df_{\alpha^{(\nu)}}}{d\alpha^{(\nu)}}, f_{\alpha^{(\nu)}} > \tag{9.13}$$

es gilt

$$\frac{df_{\alpha^{(\nu)}}}{d\alpha^{(\nu)}} \quad \text{erfüllt (9.7)} \tag{9.14}$$

und

$$f_{\alpha^{(\nu)}} \quad \text{aus (9.2).} \tag{9.15}$$

Die zur Verifikation nötigen Arbeiten können jetzt angegeben werden:

(A) Löse (9.4)/(9.10) mit dem Schumacher-Solver (vgl. Schumacher [13]); man erhält auf diese Weise ein Intervall $[\alpha]$, in dem mit Sicherheit die gesuchte Nullstelle liegt. Die Beziehungen (9.5),(9.6)/(9.11),(9.12) sind unter Einsatz des exakten Skalarprodukts auszuwerten. Die benötigten Größen $\frac{df_{\alpha^{(\nu)}}}{d\alpha^{(\nu)}}$ bzw. $f_{\alpha^{(\nu)}}$ können durch verifiziertes Lösen der Integralgleichungen (9.7) bzw. (9.2) eingeschlossen werden (siehe Dobner [2]).

(B) Mit dem so berechneten Parameterintervall $[\alpha]$ wird abschließend die Gleichung (9.2) einschließungsmäßig gelöst; man berechnet ein Funktoidelement $F_{[\alpha]}$, in dem mit Sicherheit die Moore Penrose-Lösung graphenmäßig enthalten ist.

(C) Durch Berechnung des Defekts

$$\int_a^b k(s,t)F_{[\alpha]}(t)dt - g(s), \quad a \leq s \leq b,$$

mit intervallanalytischen Methoden kann die erzielte Genauigkeit überprüft werden.

Numerische Studien für diesen Problemkreis sind gerade in Arbeit und werden in einem separaten Artikel veröffentlicht.

10. Was ist schlecht an schlecht gestellten Problemen?

Nach der eingehenden Untersuchung schlecht gestellter Fragestellungen können wir nun die eingangs aufgeworfene Frage wie folgt beantworten:

Schlecht an schlecht gestellten Problemen ist

- in erster Linie die Begriffsbildung, welche historisch bedingt ist. Wie gesehen, kann sehr wohl ein korrektes Modell eines sinnvollen technischen Problems durch eine mathematische Gleichung beschrieben werden, das eine der Bedingungen in Definition 1.1 verletzt. Nicht das Modell ist schlecht abgefaßt oder das Problem falsch formuliert, sondern die sie beschreibende Gleichung erfordert eine besonders sorgfältige (numerische) Behandlung.
- ihr sensibles Verhalten gegenüber kleinen Störungen, was ihre numerische Beherrschbarkeit einschränkt oder zumindest stark erschwert. Trotzdem ist es prinzipiell möglich, wie zuvor aufgezeigt, auch gesicherte Fehleraussagen zu berechnen.

Das Wort "schlecht" im Zusammenhang mit schlecht gestellten Problemen ist nicht wörtlich zu nehmen, sondern in diesem Sinne zu interpretieren.

Literatur

[1] Baumeister, J.: *Stable solution of inverse problems,* Vieweg, Braunschweig 1987.

[2] Dobner, H.-J.: *Verification Methods for Fredholm Integral Equations,* erscheint in Journal of Computer Mathematics Vol. 49.

[3] Engl, H.W.: *Inverse und inkorrekt gestellte Probleme,* Jahrbuch Überblicke Mathematik 1991, Vieweg, Braunschweig, pp 77 – 92, 1991.

[4] Fawcett, J.A.: *Inversion of n-dimensional spherical averages,* SIAM J. Appl. Math. 45, pp 336 – 341, 1985.

[5] Groetsch, C.W.: *Inverse Problems in the Mathematical Sciences,* Vieweg, Braunschweig, 1993.

[6] Hadamard, J.: *Lectures on the Cauchy problem in linear Partial Differential Equations,* Yale University Press, New Haven, 1923.

[7] Kreß, R.: *Linear Integral Equations,* Springer, Berlin/Heidelberg/New York, 1989.

[8] Kulisch, U. (Hrsg.): *Wissenschaftliches Rechnen mit Ergebnisverifikation,* Vieweg, Braunschweig, 1989.

[9] Louis, A.K.: *Inverse und schlecht gestellte Probleme,* Teubner, Stuttgart 1989.

[10] Natterer, F.: *The mathematics of computerized tomography,* Wiley & Teubner, Stuttgart 1986.

[11] Pucci, C.: *Sui problemo di Cauchy non "ben posti",* Atti Accad. Naz. Lincei. Rend. Cl. Sci. Fis. Mat. Nat. (8), 18, pp 473 – 477, 1955.

[12] Richter, G.R.: *An inverse problem for the steady state diffusion equation,* SIAM J. Appl. Math. 41, pp 210 – 221, 1981.

[13] Schumacher, G.: *Lösung nichtlinearer Gleichungen mit Verifikation des Ergebnisses in [8],* pp 137 – 154.

Mathematische Verfahren für technische Feldprobleme

Eberhard Halter
Fachbereich Feinwerktechnik, Fachhochschule Karlsruhe

Kurzfassung
In Spannungs-, Verformungs- und Schwingungsanalysen an festen Bauteilen, Strömungsberechnungen in Gasen und Flüssigkeiten, bei der Berechnung von elektromagnetischen Feldern und der Ermittlung von Temperaturverläufen und Wärmeflüssen geht es immer um die Bestimmung der räumlichen Verteilung und/oder des zeitlichen Verlaufs von physikalischen Größen aufgrund bekannter Gesetze. Typisch für diese technischen Feldprobleme sind komplizierte Geometrien und das Zusammentreffen mehrerer äußerer Einflüsse. Auf brauchbare Lösungen solcher Probleme führen einige bekannte numerische Verfahren wie die Finite-Elemente-Methode (FEM), die Randelementemethode (BEM) und Differenzenverfahren auf randangepaßten Koordinaten (BFC). Eine Skizze der Verfahrensideen und eine Diskussion ihrer Eigenschaften geben einen Überblick.

Technische Feldprobleme

Das räumliche Gebiet, auf dem eine physikalische Größe bestimmt werden soll, setzt sich in technischen Feldproblemen meist aus mehreren Teilen mit verschiedenen physikalischen Parametern zusammen. Es kann mehrfach zusammenhängend sein und sein Rand kann Ecken und Rundungen aufweisen. Am Beispiel Elastomechanik wird schnell die Vielfalt der Feldprobleme deutlich: die Geometrie der Bauteile, die Eigenschaften der verwendeten Materialien und die Lagerungen und Belastungen können vorgegeben werden. Durch Idealisierungen und Berücksichtigung von Symmetrien kann manches der Probleme in Bezug auf die geometrische Dimension und die Anzahl der abhängigen Variablen vereinfacht werden.

Mathematische Verfahren

Analytische Verfahren sind nur für einige einfachere Feldprobleme bekannt. Unter den numerischen Methoden für Feldprobleme sind die Finite-Elemente-Methode (FEM), die Randelementemethode (BEM) und Differenzenverfahren auf randangepaßten Koordi-

naten (BFC) die bekanntesten. Ihre Flexibilität zur Erfassung der Probleme, ihre Zuverlässigkeit und der Aufwand bei ihrem Einsatz soll erörtert werden. Eine Skizze der Verfahrensidee ist dabei hilfreich.

FEM

Die Geometrie des Berechnungsgebiets wird mit Hilfe von einfachen Teilen (Elemente) näherungsweise erfaßt. Es werden zunächst Punkte (Knoten) definiert. Benachbarte Knoten werden dann zu Elementen (Polygone oder Polyeder) verbunden. Die gesuchten physikalischen Größen werden als Überlagerungen von Formfunktionen mit den Werten der gesuchten Größe an den Knoten als Koeffizienten angesetzt. Im Ritzverfahren wird ein zum Feldproblem äquivalentes Variationsproblem gelöst. Das Einsetzen des Lösungsansatzes und partielles Ableiten nach den Knotenvariablen führt auf ein lineares Gleichungssystem. In der Elastomechanik ist das Feldproblem ein Randwertproblem und bedeutet ein Kräftegleichgewicht, das äquivalente Variationsproblem ist ein Extremalprinzip und bedeutet das Minimum der gesamten potentiellen Energie. Die Galerkinmethode ist eine Alternative zum Ritzverfahren. Beim Einsetzen des Lösungsansatzes in die Feldgleichung ergibt sich ein Residuum. Man versucht nun, die Knotenvariablen so zu wählen, daß das Residuum orthogonal zum linearen Erzeugnis der Formfunktionen ist. Dadurch wird die Lösung zur Bestapproximation im von den Formfunktionen aufgespannten linearen Unterraum. Es ergibt sich auch hier ein lineares Gleichungssystem für die Knotenvariablen. Die vom Ritzverfahren und der Galerkinmethode hergeleiteten linearen Gleichungssysteme sind gleichwertig, d. h. sie haben dieselbe Lösung, ihre Matrizen sind symmetrisch und positiv definit mit einer Bandstruktur. Zu ihrer Lösung gibt es mehrere robuste Verfahren, z. B. die Choleskymethode.

Die Vorteile der FEM sind der große Einsatzbereich sowohl in Bezug auf die Geometrievorgaben als auch hinsichtlich der Art der Feldprobleme. Es gibt ausgereifte Software von mehreren Anbietern, viele Anwender und eine stetige Weiterentwicklung. Unter den Nachteilen der FEM kann man den Bedienungsaufwand und bei hohem Anspruch an Genauigkeit den hohen Bedarf an Speicher und Rechenzeit sehen. Aber die Entwickler der Software haben gerade in neuerer Zeit erhebliche Verbesserungen erzielt, z. B. bequeme Bedienungsoberflächen, Kopplung mit CAD, iterative Löser für die Gleichungssysteme. Die Anforderungen an die Hardware wurden früher nur von Großrechnern erfüllt, jetzt gibt es ernsthafte FEM-Software auch für Kleinrechner.

BEM

In der Potentialtheorie werden Darstellungen von Feldgrößen mit Hilfe von Belegungen auf dem Rand des Gebiets hergeleitet. Damit können einige Feldprobleme als äquivalente Integralgleichungsprobleme formuliert werden, bei denen die Werte dieser Belegungen auf dem Rand des Gebiets gesucht sind. Durch eine geeignete Diskretisierung des Gebietsrandes und eine Approximation der gesuchten Belegung mit einem Ansatz aus Formfunktionen führt das Integralgleichungsproblem zu einem linearen Gleichungssystem für die Knotenvariablen. Diese Vorgehensweise wird als Randelementemethode (Boundary Element Method) bezeichnet.
Durch die Beschränkung auf den Gebietsrand wird das Problem um eine Dimension kleiner. Dies ist ein wesentlicher Vorteil der BEM, man hat im allgemeinen weniger Knoten und Gleichungen, die Modellerstellung und auch die CAD-Anbindung sind einfacher. Die BEM liefert bei Problemen mit Spannungskonzentrationen oder Rißbildungen eine gute Genauigkeit. Bei Außenraumproblemen kann sie eingesetzt werden, obwohl das Gebiet unbeschränkt ist (Schwierigkeit bei FEM). Die linearen Gleichungssysteme der BEM sind zwar kleiner, aber ihre Matrizen sind voll besetzt, unsymmetrisch und indefinit. Daher kommt es zu keiner wesentlichen Reduzierung des Rechenaufwandes. Der Anwendungsbereich in Bezug auf die Art der Feldprobleme ist kleiner als bei den FEM. Die Erfahrungen im praktischen Einsatz der BEM zeigen Vorteile gegenüber der FEM bei kompakten Bauteilen, jedoch Nachteile bei schlanken Körpern (bei großem Verhältnis von Umfang zu Fläche bzw. Oberfläche zu Volumen).

BFC

Im ursprünglichen Konzept der randangepaßten Koordinaten (Boundary Fitted Coordinates) wird das Feldproblem im physikalischen Raum ohne Typwechsel in ein äquivalentes Feldproblem auf einem Rechteck bzw. Quader transformiert. Die Gebietstransformation ist dabei Lösung eines elliptischen Randwertproblems. Ein reguläres Gitter im Rechteck bzw. Quader besitzt im ursprünglichen Gebiet ein randangepaßtes Gitter als Urbild. Das transformierte Feldproblem kann mit den bekannten Differenzenverfahren angegangen werden. Die Lösung wird auf die Knoten des randangepaßten Gitters im physikalischen Raum bezogen. Das durch die Anwendung der Differenzenverfahren entstehende lineare Gleichungssystem hat eine dünn besetzte Matrix mit Bandstruktur. Es ist nicht sehr speicherintensiv und kann iterativ gelöst werden. In Erweiterung dieses ursprünglichen Konzepts kann das Bild des Gebiets in ein Rechteck bzw. einen Quader lediglich eingebettet werden. Es ist dann möglich, Ausblendungen vorzunehmen (Inseln oder Halbinseln) und innere Ränder zu definieren. Ferner können Berechnungen auf

mehreren Gittern stattfinden, wobei ein Datenaustausch an Rändern erfolgt.
Die Diskretisierung eines Berechnungsgebiets mit Hilfe randangepaßter Koordinaten erfordert einen erheblichen Rechenaufwand und ist weniger flexibel als die freie Vernetzung mit rein algebraischen Methoden. Randangepaßte Gitter erfüllen aber leichter die Bedürfnisse der Numerik, die Verzerrung der Zellen und scharfe Wechsel in den Kantenlängen werden ausgeglichen. Vorteile haben BFC-Gitter auch in Bezug auf ihre Verwaltung. Durch die Mehrfachindizierung der Knoten (2 Indizes bei ebenen, 3 Indizes bei räumlichen Gittern) besteht eine natürliche Nachbarschaftsbeziehung und Zellendefinition.

Methodenvergleich

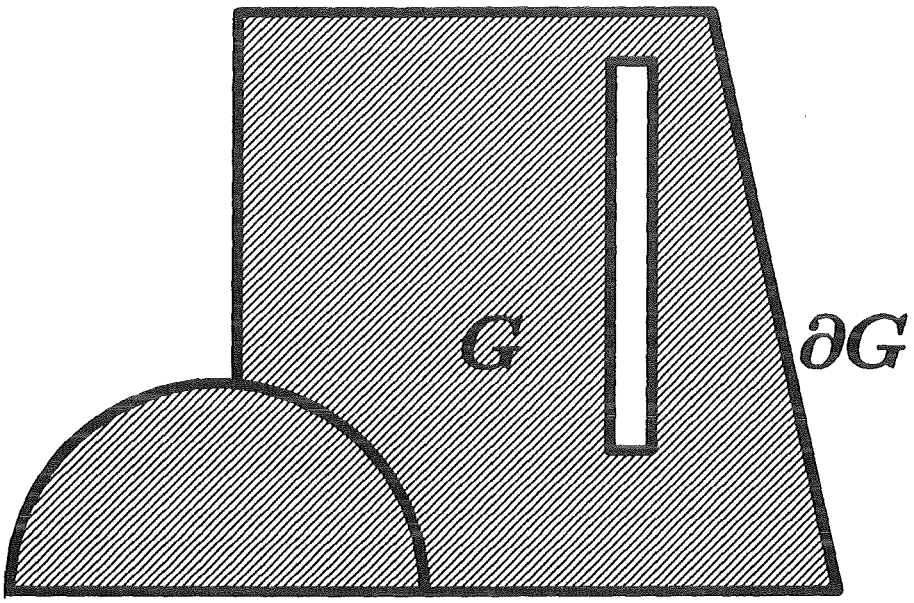
Bei einer Gegenüberstellung der Methoden zeigen sich die Stärken der FEM in der Breite des Einsatzbereichs und im Softwareangebot. Dies erklärt den überwiegenden Einsatz der FEM in der Industrie. BEM haben in einigen speziellen Anwendungen Vorteile gegenüber FEM, es werden einige BEM-Programme angeboten. Auf dem Softwaremarkt gibt es bisher kaum BFC-Angebote. Für Eigenentwicklungen sind BFC sicher geeignet und im Hinblick auf die Anwendungsbreite auch interessant.

Weiterentwicklungen

Auf dem FEM-Markt konkurieren viele Anbieter und kämpfen durch eigene Entwicklungen um Marktnischen. Nichtlinearitäten (große Verformungen, Kontaktprobleme, Materialgesetze) , Dynamik (Modalanalyse, Kriechen), Optimierungsstrategien und die Fluidmechanik sind Bereiche, in denen Neues erwartet werden kann. Die BEM-Software wird in Bezug auf die Kopplungen mit FEM und CAD verbessert. Neben einer wachsenden Zahl von Spezialanwendungen von BFC (PIC-Codes, Navier-Stokes-Gleichungen) laufen Arbeiten zu allgemeineren Gitterkonzepten, zum Einsatz von modernen numerischen Verfahren und zur Erzeugung von bequemen Benutzeroberflächen.

Anhang: Folien des Vortrags vom 29.06.1993, KfK/HDI.

Technische Feldprobleme



Vorgaben:

Gebiet G mit Rand ∂G

Partielle Differentialgleichung(en) in G

Randbedingungen auf ∂G

Typische Merkmale in der Technik:

Komplizierte Geometrien

Gemischte Randbedingungen

Unstetige Koeffizienten in DGLn

Beispiele

Elastomechanik

Vorgaben:

Geometrie, Materialeigenschaften, Belastungen

Bezeichnung:

Fachwerk, Balken, Scheibe, Platte, Schale, Körper

Von Interesse:

Spannungen, Verschiebungen, Schwingungen,
Bruchmechanik, Kontaktprobleme, ...

Weitere Gebiete

Elektromagnetische Felder

Wärmeleitung in Festkörpern

Strömungen

Mathematische Verfahren

Analytische Verfahren nur für einfache Fälle

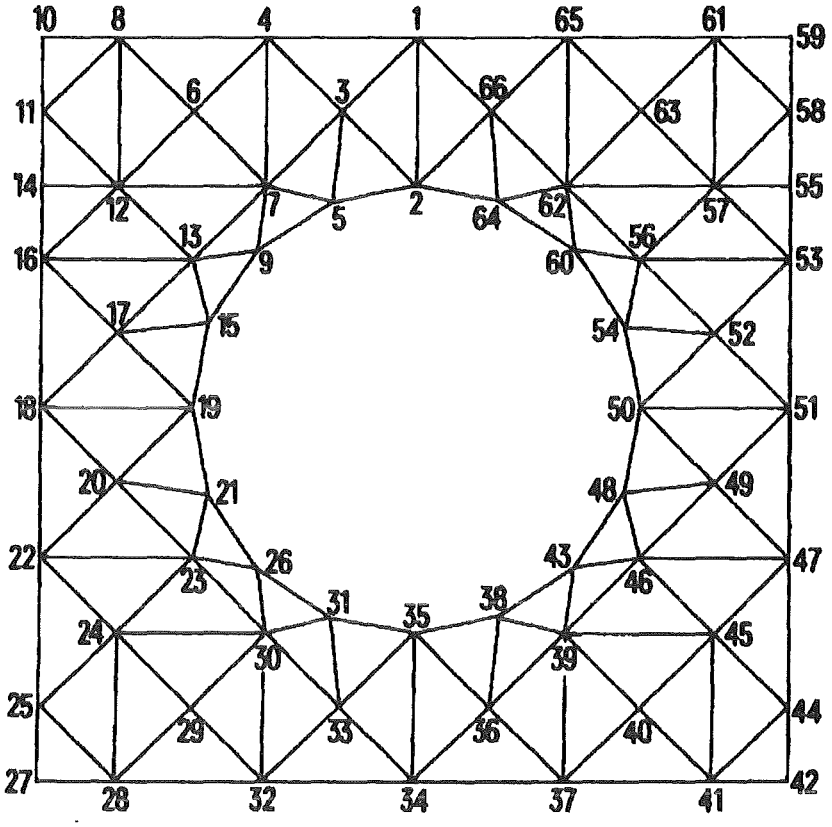
Numerische Verfahren auf Rechenanlagen

- FEM Finite Elemente Methoden
- BEM Randelementemethoden
- BFC Randangepaßte Koordinaten

Fragen:
Flexibilität (Geometrie, DGLn, RBen)
Zuverlässigkeit (Konvergenz, Genauigkeit)
Aufwand (Problemaufbereitung, Interpretation)
Software (Quelle, Benutzerqualifikation)

FEM

Diskretisierung



Ansatz

$$u(x,y) = \sum_{k=1}^{n} u_k N_k(x,y)$$

mit $N_k(x,y) = \begin{cases} 1 & \text{im Knoten k} \\ 0 & \text{in allen anderen Knoten} \end{cases}$

und u_k Knotenvariable als Koeffizient

Ritzverfahren

$$\begin{array}{ccc} \text{Randwertproblem} & \Longleftrightarrow & \text{Gleichgewicht} \\ \Updownarrow & & \Updownarrow \\ \text{Variationsproblem} & \Longleftrightarrow & \text{Extremalprinzip} \end{array}$$

Ansatz einsetzen und partiell ableiten nach u_k

$\Longrightarrow$ LGS für u_k

Galerkinmethode

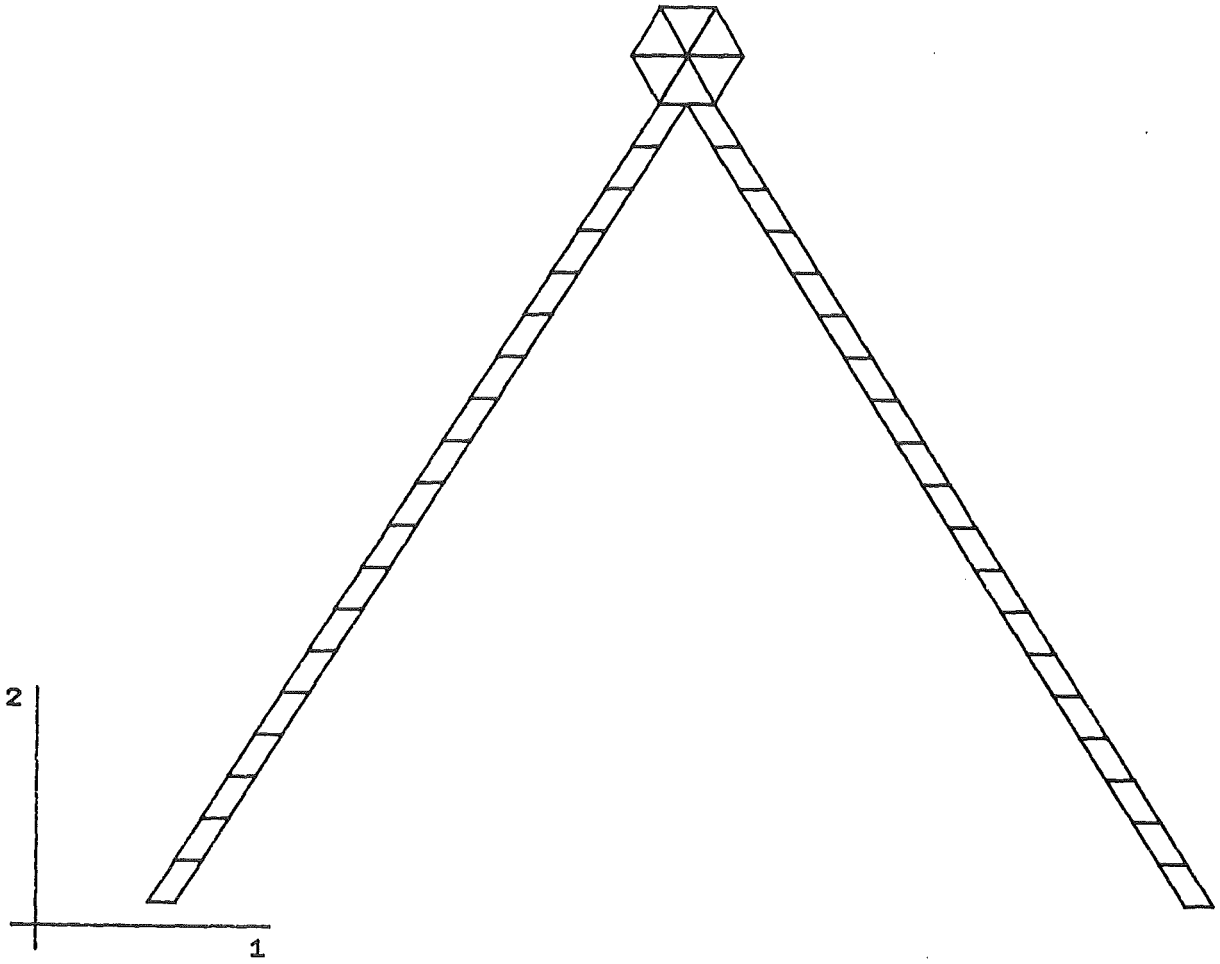
Ansatz einsetzen in DGL $\Longrightarrow$ Residuum

Wahl der u_k so, daß das Residuum orthogonal zu $[N_1, \ldots, N_k]$, d.h. beste Approximation dort

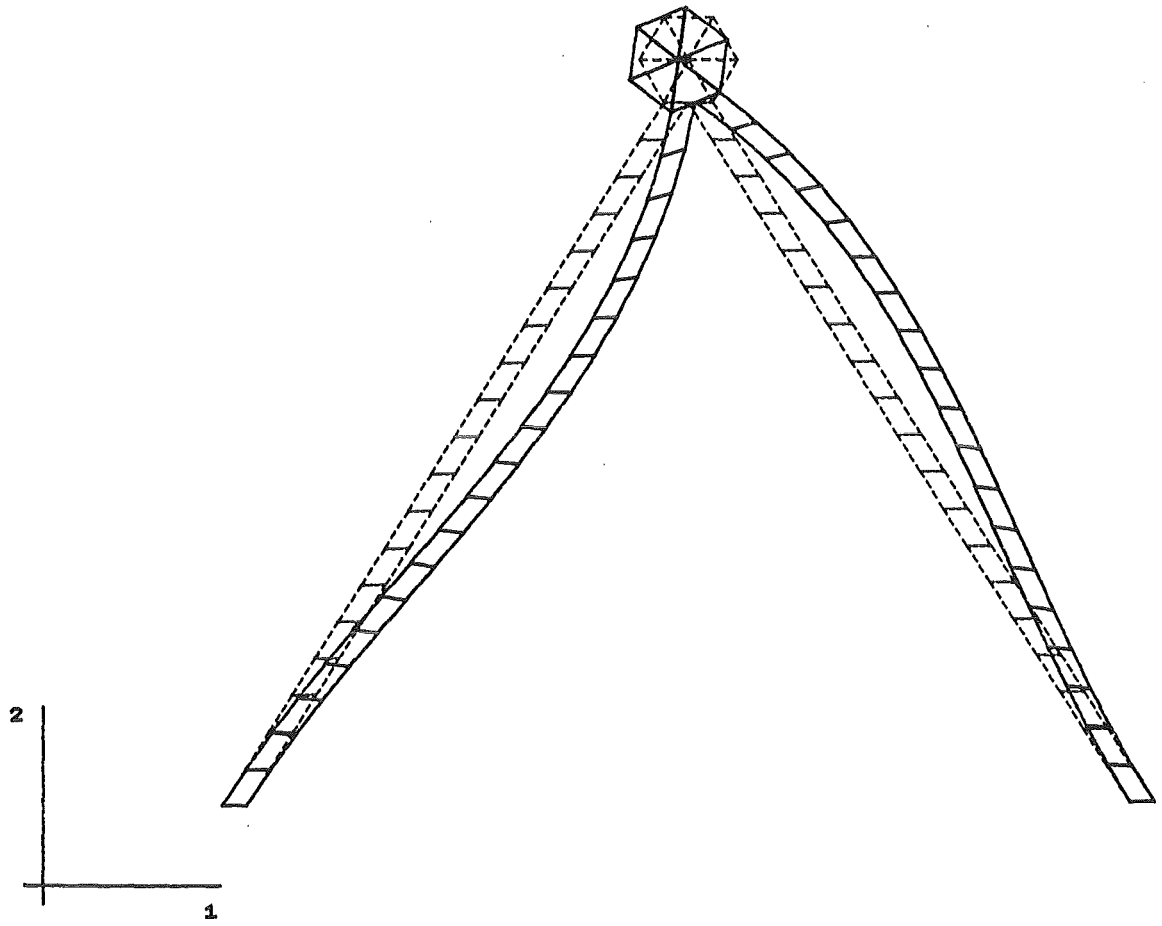
$\Longrightarrow$ LGS für u_k

Matrix des LGS

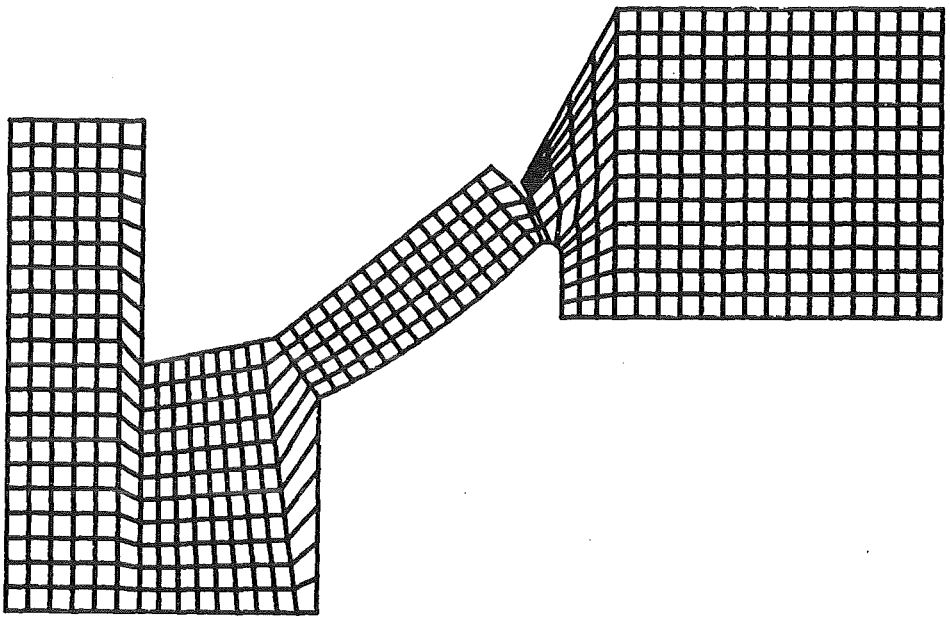
Positiv definit, Bandstruktur, dünn besetzt



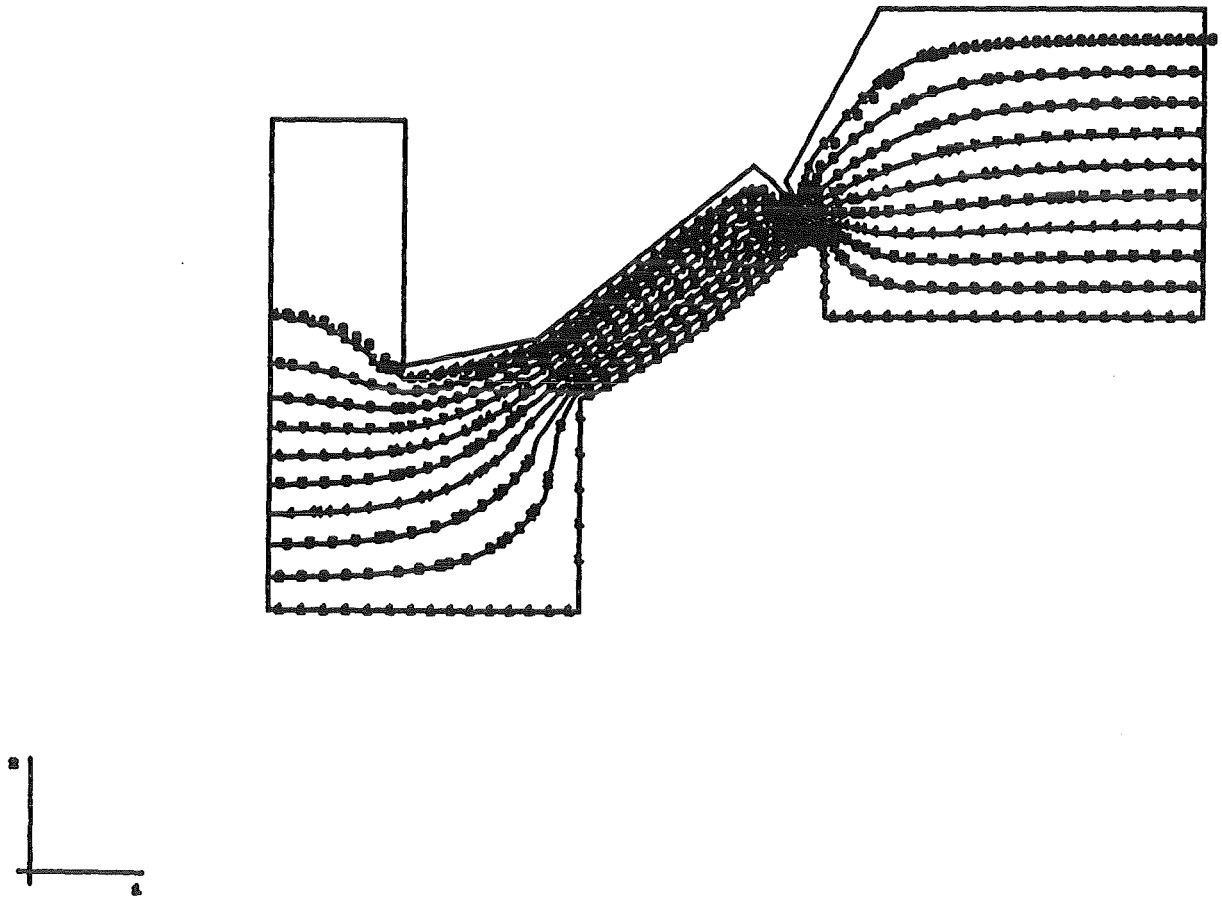
Das Bild zeigt eine einfache FEM-Diskretisierung einer Motorhalterung. Für das Gehäuse des Elektromotors wurden Dreieckzellen, für die Halterung Viereckzellen verwendet. Das Modell geht von ebenen Spannungen aus. Die unteren Enden der Halterung sind fest gelagert, d.h. die unteren vier Knoten sind nicht verschieblich.



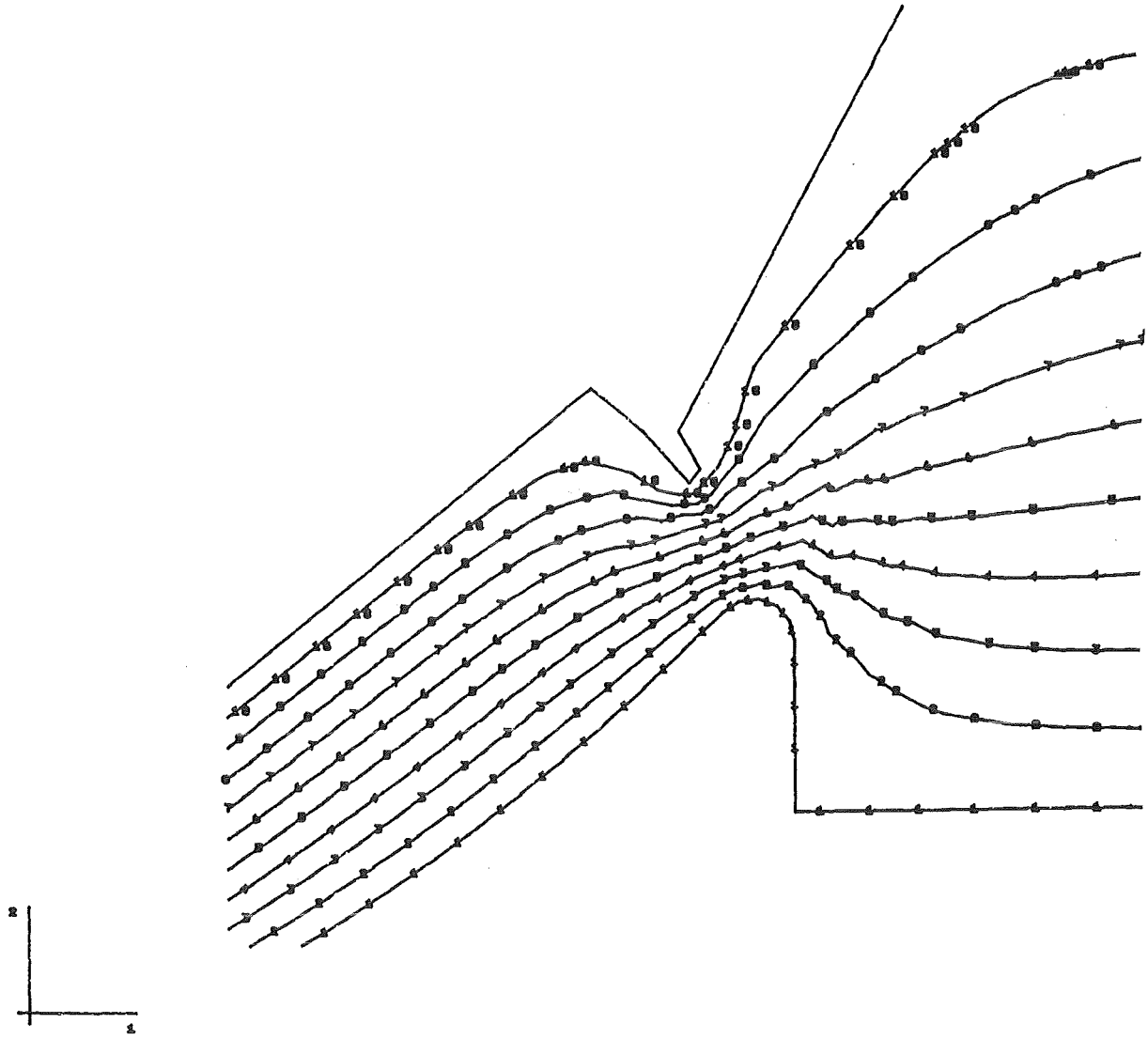
Die Verformung der Struktur unter dem Einfluß eines Drehmoments beim Anlaufen oder Abbremsen des Motors wird hier (überhöht) dargestellt. Die unbelastete Struktur ist gestrichelt unterlegt. Man erkennt die Bereiche der stärksten Verformung der Halterung.
Mit Hilfe eines Dehnungsmeßstreifens soll der Betriebszustand des Motors analysiert werden. Die Anbringung des Dehnungsmeßstreifens im Bereich der stärksten Verformung der Motorhalterung ergibt das beste Signal zur Analyse.



Das Bild zeigt den Querschnitt einer Diode mit Rotationssymmetrie, welcher mit Hilfe von Vierecken diskretisiert wurde. Es soll das elektrostatische Feld im Inneren der Diode berechnet werden, welches sich ergibt, wenn der Innenleiter auf das Potential 0 V und der Außenleiter auf das Potential 1 V gebracht werden.



Den Verlauf der Linien gleichen Potentials zeigt dieses Bild im Überblick. Der Potentialunterschied benachbarter Linien beträgt 0.1 V.



Durch Vergrößerung eines Teilbereichs der Diode wird die Lage der Linien gleichen Potentials deutlicher. Die Ursache für die Knickstellen der Linien im Inneren der Diode ist die grobe Diskretisierung.

Vorteile FEM

Sehr großer Einsatzbereich
- Geometrie (1D, 2D, 3D)
- Material (inhomogen, anisotrop, nichtlinear)
- Dynamik

Ausgereifte Software

Nachteile FEM

Bedienung (Geometrieerfassung, Diskretisierung)

Genauigkeit nur bei feiner Diskretisierung (große Datenmengen)

BEM

Randwertproblem auf G

$\Updownarrow$

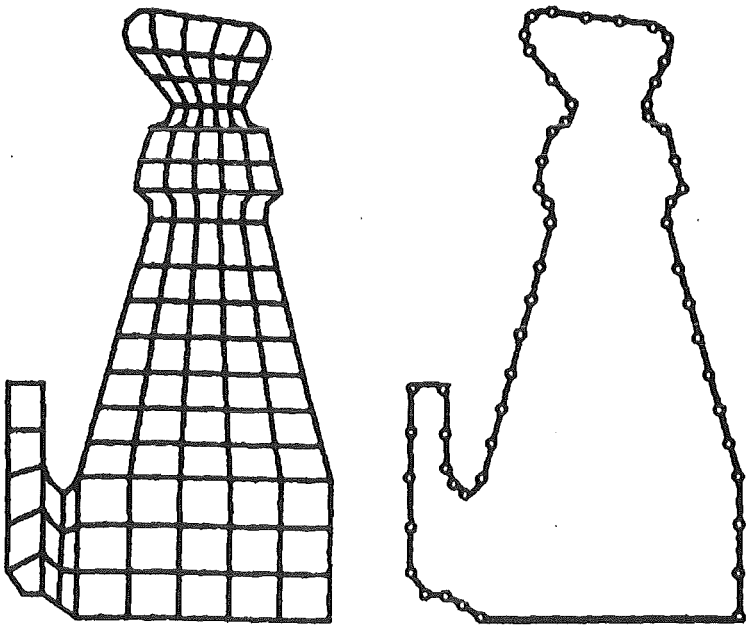
Integralgleichung auf ∂G

Diskretisierung des Randes ∂G

Approximation der Randfunktionen

$\Longrightarrow$ LGS für Knotenvariable auf ∂G

Beispiel:



FEM–Modell BEM–Modell

Vorteile BEM

Eine Dimension weniger

Weniger Gleichungen, Speicherbedarf

Einfachere Modellerstellung, CAD–Anbindung

Genauigkeit bei Spannungskonzentrationen und Rißproblemen

Außenraumprobleme

Nachteile BEM

Keine wesentliche Laufzeitreduzierung

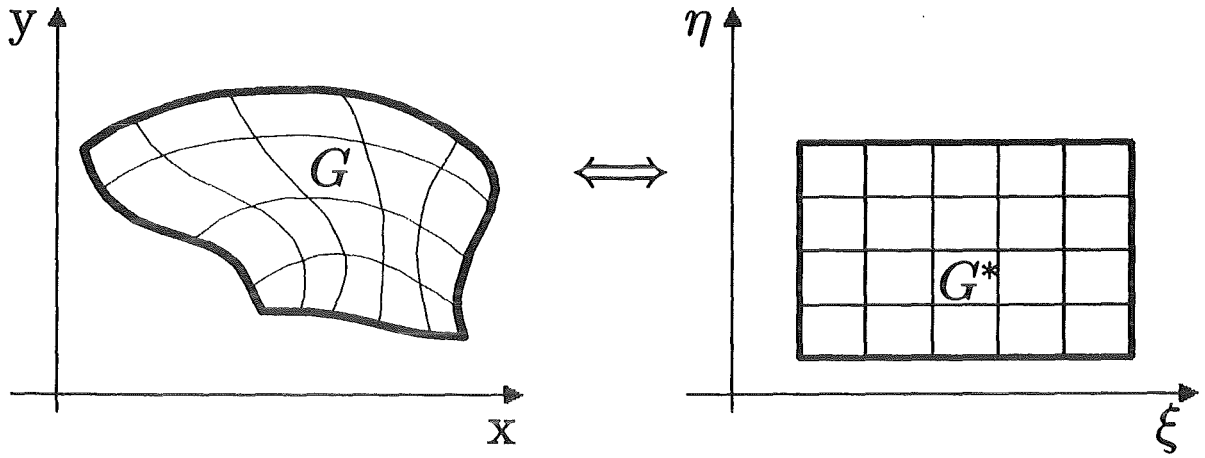
LGS voll besetzt, unsymmetrisch, indefinit

Anwendungsbereich kleiner als bei FEM

Vorteilhaft nur bei kompakten Bauteilen, nicht für Schalen, schlanke Körper, Stabwerke

Bisher keine Plastizität, Viskosität, Dynamik

Randangepaßte Koordinaten



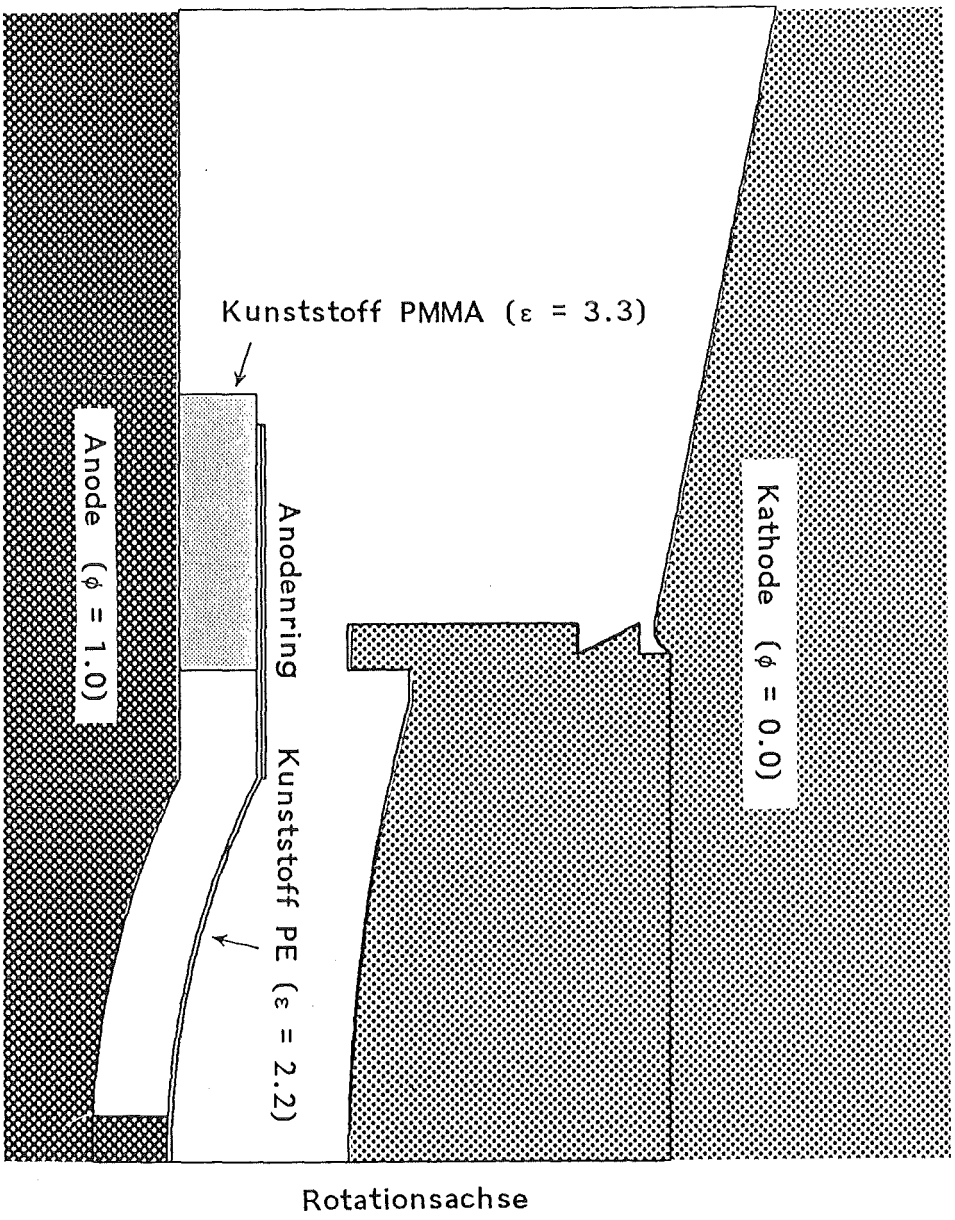
Transformation des Problems in G ohne Typwechsel auf ein äquivalentes Problem in G^*, Diskretisierung, Differenzenverfahren
$\Longrightarrow$ LGS, Bandstruktur, dünn besetzt

Konzepterweiterungen

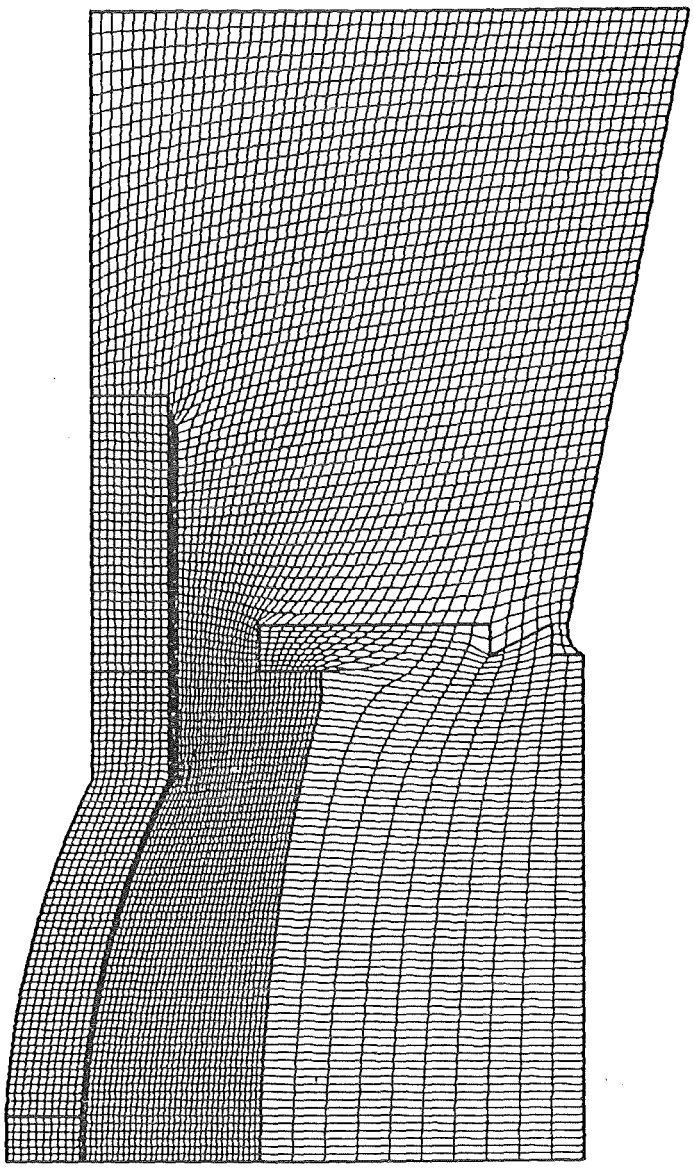
Ausblendungen und innere Ränder

Substrukturtechnik mit mehreren Gittern

PULSGENERATØR KALIF, DIØDE.



Das Bild zeigt einen Querschnitt der rotationssymmetrischen KALIF-Diode. Die Berechnung des elektrostatischen Feldes in der Diode, wenn die Anode auf das Potential 1 V und die Kathode auf das Potential 0 V gebracht werden, soll mit Hilfe eines Differenzenverfahrens auf randangepaßten Koordinaten erfolgen.



Das randangepaßte Gitter überdeckt den Bereich der Diode, in dem der Feldverlauf berechnet werden soll. Die Knoten des Gitters ergeben sich als numerische Lösung eines Randwertproblems, bei dem die Lage der Ränder vorgegeben wird. Teile der Kathode und der Anode werden ebenfalls vom Gitter überdeckt, hierauf finden jedoch keine Feldberechnungen statt, weil die Potentiale dort vorgegeben sind.

PULSGENERATØR KALIF, DIØDE.

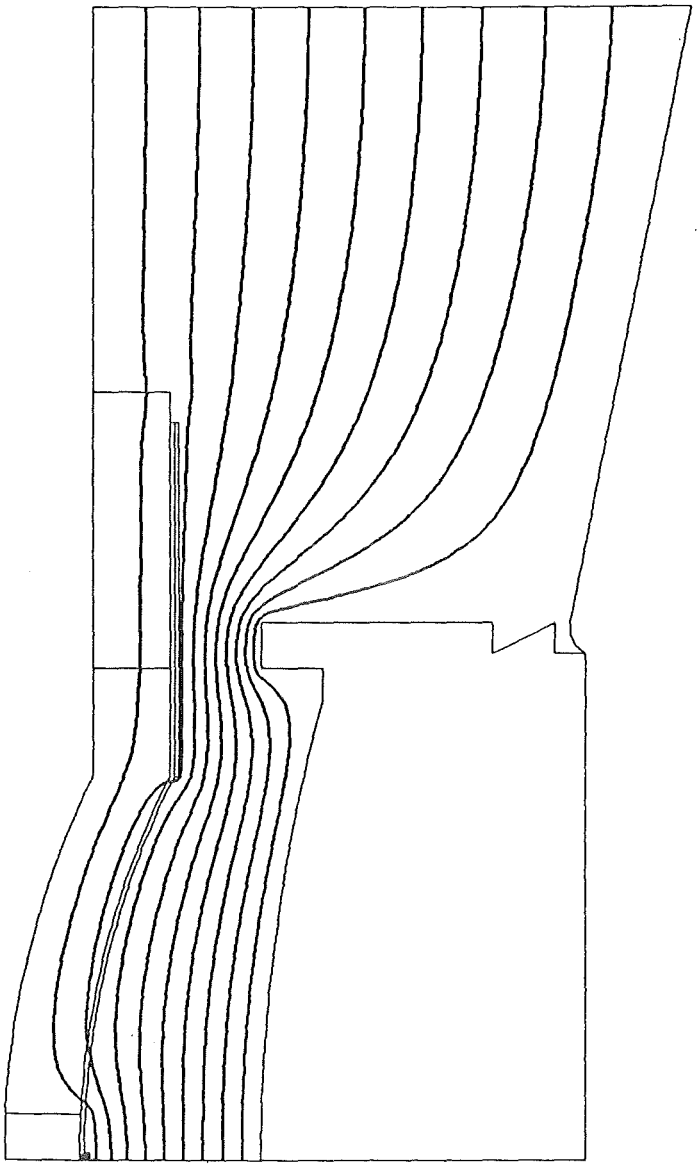
I1MX = 121, I2MX = 56.

LØGISCHES GITTER

SYMBOL	ATTRIBUT	BEDEUTUNG
▫	-3	AEUSSERER RANDPUNKT
◦	-2	AEUSSERER RANDPUNKT
▴	-1	AEUSSERER RANDPUNKT
·	0	FELDPUNKT
×	1	INNERER RANDPUNKT
•	2	INNERER RANDPUNKT
+	3	INNERER RANDPUNKT
⋆	4	INNERER RANDPUNKT
✕	5	INNERER RANDPUNKT
▪	6	INNERER RANDPUNKT
▪	7	INNERER RANDPUNKT

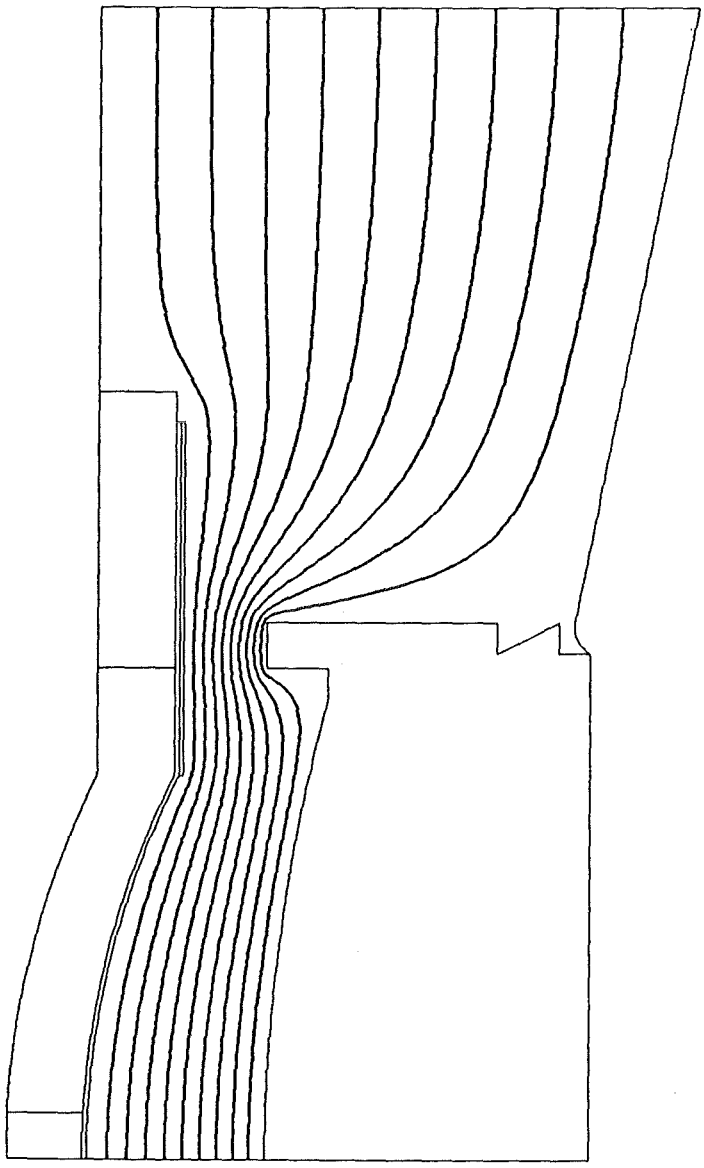
SYMBOL	ATTRIBUT	BEDEUTUNG
▫	8	INNERER RANDPUNKT
◦	9	INNERER RANDPUNKT
▴	10	INNERER RANDPUNKT
×	11	INNERER RANDPUNKT

Um die Feldberechnung und die Randbedingungen zu organisieren, trägt jeder Gitterknoten ein Attribut. Die Matrix dieser Attribute, das sogenannte logische Gitter, zeigt die Möglichkeiten für Berechnungen auf dem Diodengitter.



Das Bild zeigt den Verlauf der Linien gleichen Potentials in der Diode. Der Potentialunterschied benachbarter Linien beträgt 0.1 V. Weil die Kunststoffteile isolieren, bildet der Anodenring aus Aluminium eine sogenannte flotierende Elektrode, d. h. sein Potential ergibt sich aus einer Integralgleichung.

PULSGENERATOR KALIF, DIODE.



Die Kunststoffhaut und der Anodenring liegen bei dieser Berechnung auf Anodenpotential. Dadurch wird das Feld zwischen Anodenring und Kathode stärker. Das logische Gitter ist bei beiden Berechnungen dasselbe, jedoch die Zuordnung der Randbedingungen zu den Attributen ist verschieden.

Methodenvergleich

	FEM	BEM	BFC
Einsatzbereich	++	0	+
Verständnisaufwand	0	−	+
Softwareangebot	++	+	−
Einsatzfälle	~ 88%	~ 10%	~ 2%

Entwicklungen

FEM

Nichtlinearitäten (Materialgesetze, Geometrie), Fluidmechanik, Optimierungsstrategien

BEM

Kopplungen mit FEM und CAD

BFC

PIC, Substrukturtechnik, Software mit Industriestandard

Material Accountancy: A Game Theoretical Analysis

Shmuel Zamir
The Hebrew University of Jerusalem, ISRAEL.

Introduction

Material accountancy is a methodology designed to control materials with particular properties - rare, unpleasant, dangerous, precious - used by man in the course of his economic and social activities but their use requires the exercise of special care. In particular, material accountancy is practiced to a loss or diversion of materials for purposes unknown, but illegal according to some agreement, law or treaty.

Examples for this use of material accountancy are diverse environmental problems; here it is to be guaranteed that certain pollutants are released to the environment only within permitted standards. Other examples are arms control and disarmament agreements where the appropriate use of troops, equipment or special materials has to be controlled. In fact, it was the *Treaty for Non-Proliferation of Nuclear Weapons* which stimulated the most detailed work in this area.

A straightforward formulation might state that any material entering a well defined area - called *material balance area* - cannot simply disappear, i.e., either it is still there, or it has left it. So if at time t_0 there is an amount I_0 of the material (the beginning *real inventory*), if the net flow of material in the time interval $[t_0, t_1]$, is D, then the *book inventory* at time t_1 is $B = I_0 + D$. If the amount of material actually measured in the material balance area at time t_1 is I_1, then the difference is called *Material Unaccounted For* (*MUF*),

$$MUF = B - I_1 = I_0 + D - I_1.$$

This is not always zero due to measurements errors and possibly to illegal diversion of material.

Statistically, this problem is treated as a standard problem of hypothesis testing:

$$\begin{array}{lll} H_0 : & E(\mathit{MUF}) = 0 & \text{(legal behavior)} \\ H_1 : & E(\mathit{MUF}) = \mu > 0 & \text{(illegal behavior)}, \end{array}$$

and the corresponding is based on the observed value of MUF, by $\widehat{\mathit{MUF}}$:

Reject H_0 (and accept H_1) if and only if $\widehat{\mathit{MUF}} > s$.

The threshold s is determined by the *false alarm* (type I error) probability α.

In general, a sequence of n inventory periods $[t_0, t_1], \ldots, [t_{n-1}, t_n]$ is considered for each of which, a single material balance statistics is observed namely

$$\mathit{MUF}_i = I_{i-1} + D_i - I_i, \quad i = 1, \ldots, n.$$

Besides the technical difficulties involved in the statistical test for deciding whether or not an illegal action has taken place, we raise a more fundamental question regarding the use of statistical tests in this context. After all, this is not the standard scenario of statistical inference problem in which *one decision maker* - the statistician, observes a random sample generated randomly by 'nature'. In the situation considered here there is a *second decision maker*, namely the *operator*. This is the entity which decides whether to behave legally or illegally and what form of illegal behavior to choose.

One is thus driven from the standard statistical analysis to *game theoretical analysis*. This is the theory designed to provide mathematical models to situations of strategic conflicts involving several decision makers.

Game theoretical framework

We start by setting a general multistage inspection game. In this game there are two players: the *operator*, denoted as player **O** , and the *inspector*, denoted as player **I** . The game proceeds in n stages and it is played as follows:

- At stage 1, the operator **O** chooses an action m_1 not observable by the inspector **I** (this is the amount diverted in stage 1). An observation x_1 is drawn from a r.v. X_1 with density $f_1(\cdot \mid m_1)$ (this is MUF_1). This observation becomes common knowledge to both players.

- The inspector chooses either C (Clear), in which case the game continues to stage 2, or A (Alarm), in which case the game terminates with payoffs $(I_1(m_1), O_1(m_1))$.

- Inductively: At stage $i = 2, \ldots, n$, if the game has not been terminated before, player **O** chooses m_i (amount diverted at stage i), and an observation x_i (MUF_i) is made from the r.v. with density $f_i(\cdot \mid m_1, x_1, \ldots, m_{i-1}, x_{i-1}, m_i)$.

  Player **I** chooses either C, in which case the game continues to stage $i+1$, if $i < n$, and terminates with payoffs $(\tilde{I}_n(m_1, \ldots, m_n), \tilde{O}_n(m_1, \ldots, m_n))$ if $i = n$, or A, in which case the game terminates (by an alarm) with payoffs $(I_i(m_1, \ldots, m_i), O_i(m_1, \ldots, m_i))$.

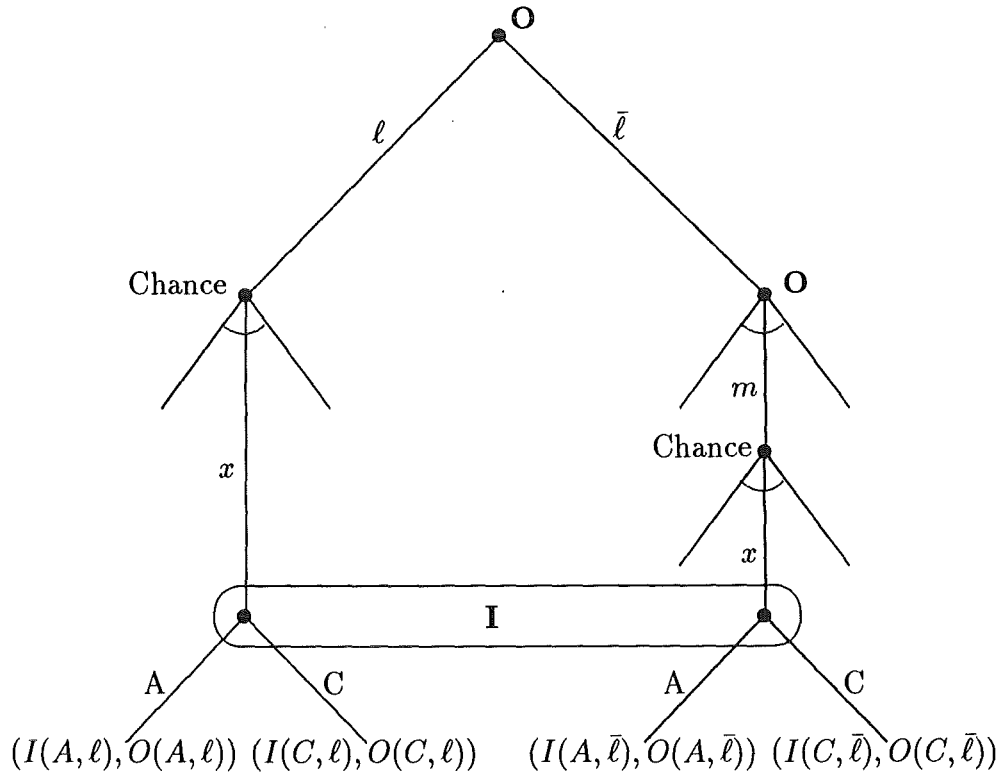
Restrictions on players' strategies: A static game.

We start our analysis of the model with a natural simple case which relates our model to the existing, mainly statistical, analysis which is basically *static*. The static model is obtained from the general model by imposing the following restrictions on the players' strategies:

Assumption 0.1

- *The operator's diversion strategy is completely determined at the beginning of the game, as he chooses either not to divert - ($m_i = 0$, $i = 1, \ldots, n$), or to divert according to a plan of the form $m = (m_1, \ldots, m_n)$ with $m_i \geq 0$ and $\sum_i m_i = M$, for a fixed constant $M > 0$.*
- *The inspector may call an alarm only at the end of the n-th stage, after having observed the whole vector $x = (x_1, \ldots, x_n)$.*

In view of Assumption 0.1, the resulting game, Γ_0, is basically a "one stage game" in which **O** moves first to choose ℓ (legal behavior) or $\bar{\ell}$ (illegal behavior) together with a diversion vector $m = (m_1, \ldots, m_n)$. Then a random vector x is observed and the inspector decides whether or not to call an alarm. This extensive form game, which we denote by Γ_0, is described in the following figure.



The restricted game Γ_0 in extensive form.

To analyze the game Γ_0 we first choose convenient payoff scales. Table 1 gives the payoffs for **I** and **O** at the end points of the game, i.e., for all combinations of ℓ or $\bar{\ell}$ with A or C.

		Operator	
		ℓ	$\bar{\ell}$
Inspector	C	$(0,0)$	$(-1,1)$
	A	$(-e,-h)$	$(-a,-b)$

Table 1. The payoffs for Γ_0.

The entries of Table 1 are ordered pairs (x,y) in which x is the payoff for **I** and y is the payoff for **O** .

A pure strategy of **I** is an alarm set $\mathbf{A} \subset \mathcal{R}^n$, with the interpretation that **I** calls an alarm, at the end of the n-th period, if and only if $(x_1,\ldots,x_n) \in \mathbf{A}$. A mixed strategy s is a probability distribution on pure strategies.

A pure strategy of **O** is a choice between ℓ and $\bar{\ell}$ and a diversion plan $m = (m_1,\ldots,m_n)$ satisfying $\sum_i m_i = 1$. A (*behavioral*) strategy of **O** is a pair $t = (q,p)$ where q is the probability of $\bar{\ell}$ and p is the probability distribution on diversion plans m if $\bar{\ell}$ is chosen.

The game Γ_0 can be described as a game in normal form: $\Gamma_0 = (S,T,I,O)$ where S is the set of mixed strategies for **I** , T is the set of behavioral strategies for **O** , I and O are the payoff functions, for **I** and **O** respectively, given by

$$I(s,t) = -q(a+(1-a)\beta(s,p)) - (1-q)e\alpha(s) \quad (1)$$
$$O(s,t) = -q(b-(1+b)\beta(s,p)) - (1-q)h\alpha(s) \quad (2)$$

where $\alpha(s)$ and $\beta(s,p)$ are, respectively, the type I and type II errors resulting from using the test s and the diversion plans distribution p.

Solution of the static game

The basic solution concept in non zero-sum games is the Nash Equilibrium (which we shall simply refer to as equilibrium), defined as follows:

Definition 0.2 *A pair of strategies (s^*,t^*) is called an equilibrium point if*

$$I(s^*,t^*) \geq I(s,t^*) \quad \forall s \in S$$
$$O(s^*,t^*) \geq I(s^*,t) \quad \forall t \in T$$

The corresponding $I(s^*,t^*)$ and $O(s^*,t^*)$ are then called equilibrium payoffs of the game.

For any strategy s of **I** we denote

$$\beta(s) := \sup_p \beta(s,p) = \sup_m \beta(s,m). \tag{3}$$

This is the highest type II error which can result from the test s. Let $\beta = \beta(\alpha)$ be the relation between α and the β attainable by the most powerful test, that is

$$\beta(\alpha) = \inf_s \{\beta(s) \mid \alpha(s) = \alpha\},$$

or, substituting $\beta(s)$ from (3),

$$\beta(\alpha) = \inf_{\{s|\alpha(s)=\alpha\}} \sup_m \beta(s,m). \tag{4}$$

We first solve an auxiliary zero-sum game G_α, in which

- Player **I** (the maximizer) is restricted to use tests with type I error probability not exceeding α.
- Player **O** has to divert, and his strategy is a diversion vectors $m = (m_1, \ldots, m_n)$.
- The payoff (from **O** to **I**) when (s,m) is played is the detection probability $1 - \beta(s,m)$.

After solving this game, the solution of game Γ_0 is given by the following:

Theorem 0.3 *The game Γ given in Figure 2 has a unique equilibrium $(s^*, (q^*, m^*))$ in which:*

- *The false alarm probability $\alpha^* = \alpha(s^*)$ is the solution of*

$$\beta(\alpha^*) = \frac{b}{b+1} - \frac{h}{b+1}\alpha^*.$$

- *The probability q^* of illegal behavior is given by the solution of*

$$\frac{e}{q^*} = e - (1-a)\beta'(\alpha^*).$$

- *The strategy s^* and the diversion plan m^* are optimal strategies in the game G_{α^*}.*

A sequential model

After solving the static game, we relax a few crucial assumptions to obtain the a *two period* model which captures the following features:

1. The significant amount of diverted material is 1 (by appropriate choice of units.)

2. The operator has no utility for less than 1 unit of mass diverted, and has no increment of utility for any additional mass beyond 1. This implies that (under any reasonable inspection policy) any second period diversion which is neither 0 nor completion to 1 is dominated.

3. The operator has the option to *retreat* in second the period from his diversion plan (i.e. not completing the diversion he started with) if it is "too risky".

4. Alarm as such has a negative utility for both players. However, the more material "caught" to be diverted the worse it is for the operator and (relatively) better it is for the inspector.

5. Time is valuable: Any amount of material diverted at first period, carries an *additional penalty* for the inspector to the second period.

6. The negative effect for the inspector of an undetected diversion, increases with the amount diverted (and may have discontinuous jump at the critical mass 1.)

The game is played as follows:

Period 1

- O chooses $m \in [0,1]$ (mass diverted in first period.)
- A random variable X_1 is observed by both players (denote this observation by x_1, this is MF_1).
- I chooses one of the two actions: C_1 (continue to second period) or A_1 (declare an alarm) which stops the game and assigns payoffs $(-a(m), -b(m))$ to I and O respectively.

Period 2

- O chooses one of the two actions: D (complete diversion to 1) or R (retreat.)
- A random variable X_2 is observed by both players (denote this observation by x_2, this is MF_2.)
- I chooses one of the two actions: C_2 (O.K.) or A_2 (declare an alarm.) In both cases the game ends and payoffs are made.

In the payoffs are chosen to have the following features:

- Undetected diversion of critical amount 1, which is the worst event for I and the best for O , corresponds to payoffs $(-1, 1)$ (bu appropriate choice of utility units.)

- The damage to the inspector from undetected diversion of amount m is $F(m)$ which we assume increasing in m, $F(0) = 0$ and $F(1) = 1$.
- Any alarm results in a damage to each player; $a(m)$ to **I** and $b(m)$ to **O** , where m is the amount which is actually diverted. We assume that $a(m)$ decreases in m, and $b(m)$ increases in m, with $a(1) = a < 1$ and $b(1) = b > 0$ corresponding to the damage of detection of a full diversion.
- Any amount m diverted in the first period and not detected at that period, results in *additional* damage (to the inspector) of $d(m)$. We assume that it is an increasing function of m and $d(0) = 0$.

The strategies in this game are:

- A strategy of the inspector is an ordered pair $s = (a_1, a_2)$, where a_1 is a transition probability from $\mathcal{R}$ to $\{A_1, C_1\}$ and a_2 is a transition probability from $\mathcal{R}^2$ to $\{A_2, C_2\}$, with the interpretation: if $X_1 = x_1$ then **I** chooses A_1 (i.e. calls an alarm) with probability $a_1(x_1)$. If the game reaches the second period with $X_1 = x_1$ and $X_2 = x_2$, then **I** calls an alarm according to the cumulative probability distribution $a_2(x_1, x_2)$.
- A strategy of the operator is an ordered pair $t = (q, p)$ where q is a measure on $[0, 1]$, the probability distribution of the first period diversion m, and p is a transition probability from $[0, 1] \times \mathcal{R}$ to $\{D, R\}$ that is, $p(m, x_1) = P(D \mid m, x_1)$ is the probability of completing the diversion at second period given that m was already diverted at first period, the observed X_1 (i.e. MUF_1) was x_1 and the **I** did not call an alarm at first period.

The structure of equilibrium

In studying the equilibrium (Definition 0.2) of the above described game, the major problem is the size and complexity of the strategy sets which makes it practically impossible even to write the equilibrium conditions in a workable form. We first prove that for finding equilibrium points, only much smaller sets of strategies may be considered. Only *threshold* strategies can be equilibrium strategies:

Theorem 0.4 *(Second period test for the inspector.) If s is an equilibrium strategy of* **I** *then:*

(i) *To any x_1 not followed by an alarm at first period, there is a critical value $c_2(x_1) \in (-\infty, \infty)$ such that:*

$$x_2 > c_2(x_1) \implies a_2(x_1, x_2) = 1 \quad \text{(i.e. call alarm surely.)}$$
$$x_2 < c_2(x_1) \implies a_2(x_1, x_2) = 0 \quad \text{(i.e. surely, alarm is not called.)}$$

(ii) *The function $c_2(x_1)$ is decreasing in x_1*

Theorem 0.5 *(Second period test for the operator.) If t is an equilibrium strategy of* **O** *then:*

(i) *for (almost) each x_1 there is $m^*(x_1)$ such that*

$$\begin{aligned} m > m^*(x_1) &\implies p(m, x_1) = 1 \\ m < m^*(x_1) &\implies p(m, x_1) = 0 \end{aligned}$$

(ii) *The function $m^*(x_1)$ is increasing in x_1*

Theorem 0.6 *(First period test for the inspector.) If s is an equilibrium strategy of* **I** *then the alarm set at the first period is determined by a critical value c_1, that is:*

$$\begin{aligned} x_1 > c_1 &\implies a_1(x_1) = 1 \quad \text{(i.e. call alarm surely.)} \\ x_1 < c_1 &\implies a_1(x_1) = 0 \quad \text{(i.e. surely, alarm is not called.)} \end{aligned}$$

The *CUMUF* test as an equilibrium strategy

At this stage in our analysis, we have an opportunity to 'calibrate' the game theoretical model: Since this is, in some sense, a generalization of the Statistical setup, it should be possible to see under which conditions we can derive known Statistical tests from the game. We do that for the *CUMUF*test; We prove, roughly, that under the assumptions underlying the statistical treatment, the *CUMUF*test emerges as the solution (i.e. equilibrium strategy) of the game. The assumptions needed on the payoffs are:

Assumption 0.7

$$\begin{aligned} (i) \quad a(m) &= a \quad \forall m \in \mathcal{R} \\ (ii) \quad b(m) &= b \quad \forall m \in \mathcal{R} \\ (iii) \quad d(m) &= 0 \quad \forall m \in \mathcal{R} \\ (iv) \quad F(m) &= \begin{cases} 0 & m < 1 \\ 1 & m \geq 1 \end{cases} \end{aligned}$$

for some constants $0 < a < 1$ and $b > 0$.

In addition, we need the following assumption to make the game static just like the statistical problem.

Assumption 0.8 *The operator is restricted to strategies $t = (q, m^*)$ in which $m^*(x_1)$ is constant in x_1.*

With these assumptions we prove the following:

Theorem 0.9 (i) *The game has an equilibrium* (s^*, t^*) *in which the strategies* $s^* = (c_1, c_2(\cdot))$ *and* $t^* = (q, m)$ *are determined as follows:*

- $c_1 = \infty$.
- $c_2 = C - x_1$, *where* C *is the unique solution of the equation:*

$$\frac{\phi\left(\frac{C-1}{\sqrt{\sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2}}\right)}{\phi\left(\frac{C}{\sqrt{\sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2}}\right)} = \frac{b}{1+b}. \tag{5}$$

- $m = \dfrac{1 + \rho\frac{\sigma_2}{\sigma_1}}{1 + 2\rho\frac{\sigma_2}{\sigma_1} + \frac{\sigma_2^2}{\sigma_1^2}}$.
- $q = \dfrac{K}{1+K}$ *where* K *is the following function of* C *and* m: *Denote*

$$d = 1 - m\left(1 + \rho\frac{\sigma_2}{\sigma_1}\right)$$

then

$$K = \frac{a}{1-a}\exp\left\{\frac{m^2}{2\sigma_1^2} + \frac{d^2 - 2dC}{2\sigma_2^2(1-\rho^2)}\right\}.$$

(ii) *Under some ('mild') assumption on the statistical distributions, this equilibrium is unique.*

Exclusion of a sure first period diversion

Coming back to our general two period model, we no longer make Assumption 0.8 that is, we reinstate the possibility for the operator to retreat from his diversion plan. One of the major difficulties in attempting to solve the equilibrium equations is the fact that one of the "unknowns" is a probability distribution q on $[0,1]$ which determines the first period diversion. In spite of richness of the set from which q can be chosen, we strongly conjecture that in equilibrium, q belongs to a much smaller (and hence more manageable) set of distributions namely, the support of q consists of only two values of m, one of which is $m = 0$. Our first result proves that this distribution is the simplest possible, that is there is no equilibrium in which q has a single point support, in other words:

Theorem 0.10 *There is no equilibrium in which* O *diverts (with probability* 1*) a fixed amount* m.

Positive probability for legal behavior

Assume that in the strategy of O the probability distribution q is of finite support. That is consider O 's strategies in which the first period diversion can take one of the values $(m_1, m_2, \ldots, m_n)$ with probabilities $(q_1, q_2, \ldots, q_n)$ respectively. Assume without loss of generality that $0 \leq m_1 < m_2 < \ldots < m_n \leq 1$. For this case we prove,

Theorem 0.11 (i) *In equilibrium, there is a positive probability for legal behavior of the operator (i.e., $m_1 = 0$ and $x_1^* = -\infty$).*

(ii) *The first period* I *threshold, c_1 is also the threshold for* O *'s completion if he has diverted the largest amount m_n.*

In particular this implies that in equilibria in which the support of q consists of only two points (which we conjecture is the only possible equilibrium), the equilibrium strategies must have the following structure:

- The operator behaves legally (in both periods) with probability q and with probability $1 - q$ he diverts m at the first period and completes the diversion at the second period if no alarm was called before.
- The inspector uses a two period threshold strategy $s = (c_1, c_2)$ which is to call an alarm at the first period if $x_1 > c_1$ and otherwise to call an alarm at the second period if $x_2 > c_2(x_1)$.

The interesting thing is that *in no event does the operator retreat* after starting a diversion plan. Although he has this option, he does not use it in equilibrium.

Now, this equilibrium becomes very similar to the two period model in which *we assumed* that there is no possibility for O to retreat. The conditions determining the equilibrium strategies $\tau = (q, m)$ and $s = (c_1, c_2)$ are similar except that now we will not be able to conclude that $c_1 = \infty$: Because of the "time value term" $d(m)$ we must find $c_1 < \infty$ that is, the inspector may well call an alarm at the first period.

Numerical solutions

The equilibrium equations for the sequential model cannot be solve analytically (with or without the 'time factor' $d(m)$). Therefore, we solved the equations numerically for linear time factor: $d(m) = d \cdot m$, for various values of d.

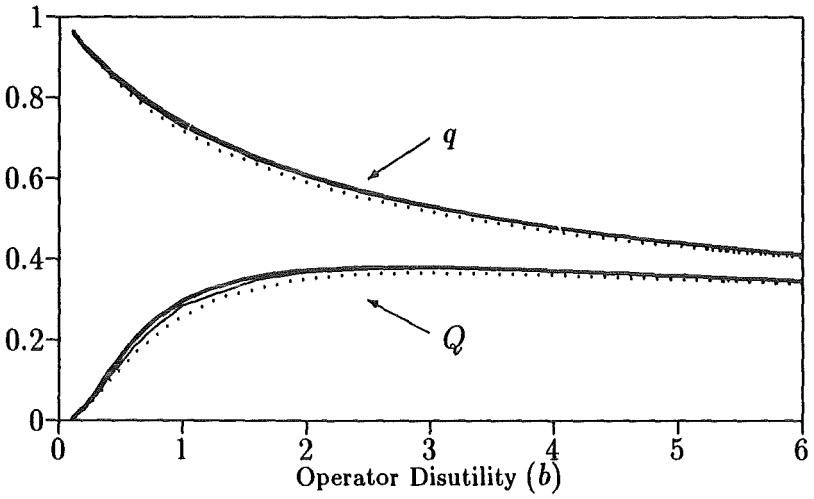
The input of the numerical procedure consists the statistical distribution parameters σ_1, σ_2, ρ, the payoff parameters b (alarm penalty) for the operator and a (alarm penalty), and finally d, the 'time factor'.

The output is an equilibrium point, which consists of: q - the diversion probability, m - the amount diverted at the first stage (in case of diversion), c_1 - the first stage threshold for alarm, and $c_2(x_1)$ - the second stage threshold (function). From these we computed several characteristics of the equilibrium, the most interesting of which are:

- *FAP* - The false alarm probability.

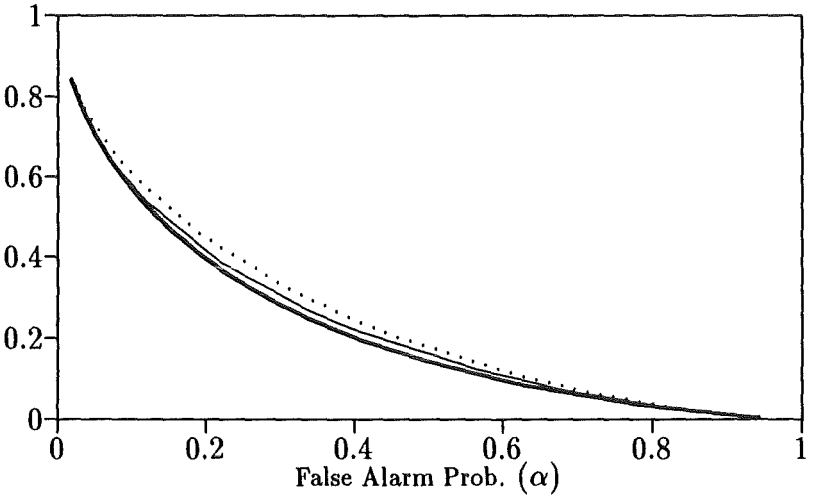
- β The probability of no detection, given that a diversion took place (type II error).
- $Q = q\beta$ - The probability that a diversion will take place and will not be detected.

The following is a sample of our numerical results:



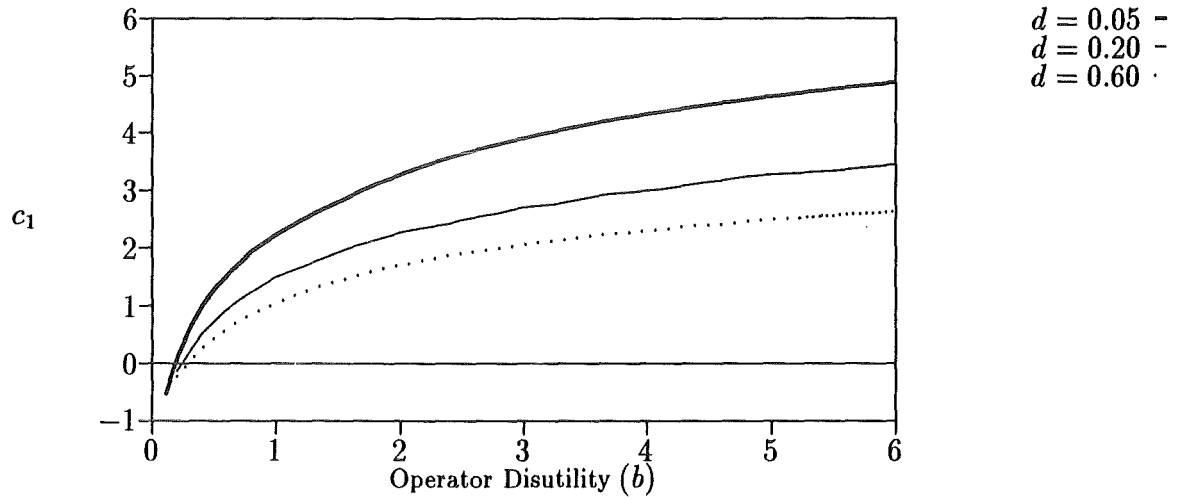
$d = 0.05$ –
$d = 0.20$ –
$d = 0.60$ ·

Here we see that although the diversion probability q decreases as the penalty b gets larger, the total probability of overlooked diversion increases in b for low values of b (but then decreases as expected).

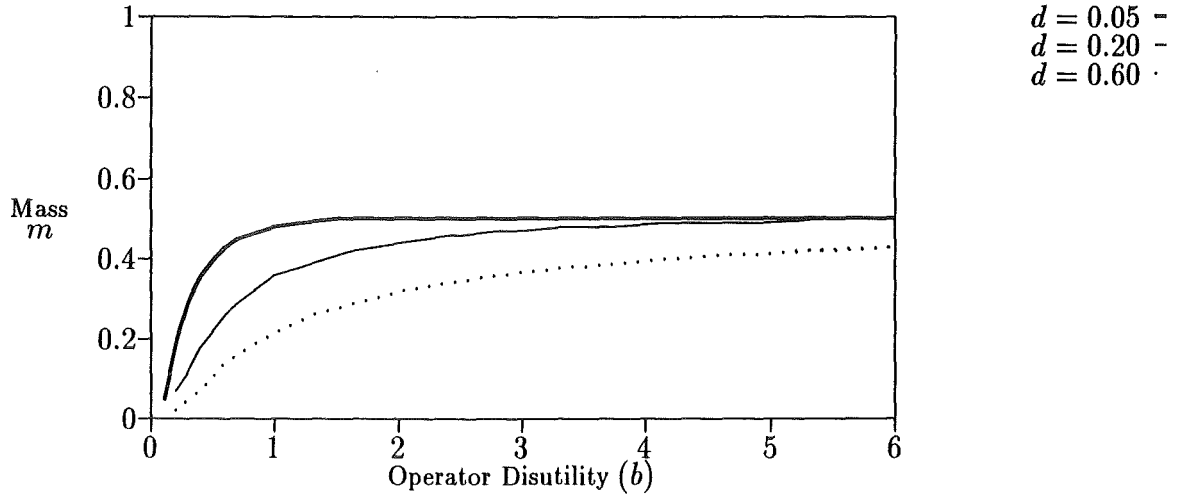


$d = 0.05$ —
$d = 0.20$ —
$d = 0.60$ ···

Here we see the classical dependence of the probabilities of first and second type errors. Note that each game corresponds to *a single* point on this line. To obtain the line we varied the game by varying the value of the penalty b.



As expected, the higher is the penalty b the less likely it is for the inspector to call an alarm in the first period.



The influence of the penalty b on the amount m diverted at the first period is non significant in a large range of b. Recall that with no time element ($d = 0$), the equilibrium value is constant in b (1/2 if $\sigma_1 = \sigma_2$).

As for the effect of the time factor d on the solution, we see that it has non significant effect on the 'performance' of the equilibrium as measured by FAP, β, q and Q. It does affect however both players' strategies namely c_1, and m.

STATISTICAL METHODS FOR THE CARCINOGENIC RISK ASSESSMENT

A.Yu. Yakovlev
Biostatistical Unit, Institut Curie, Paris, France
Department of Applied Mathematics, St. Petersburg Technical University, St. Petersburg, Russia

SUMMARY When applied to the statistical analysis of data on cancer incidence, a pertinent parametric method has the following distinct advantages:

(1) it allows a natural interpretation in terms of parameters bearing clear biological meaning,
(2) it provides a prediction of the carcinogenic risk forward in time beyond the follow–up period,
(3) it offers a means of estimating the proportion of unaffected individuals (*surviving fraction*) from time–to–tumor observations,
(4) it forms the basis for designing optimal strategies of cancer surveillance.

This paper discusses a new stochastic model of radiation–induced and spontaneous carcinogenesis based on the consideration of biological processes underlying the tumor latency within the random minima framework. A parametric family of distributions is obtained that is well suited for estimation purposes. The model can be a useful means in the statistical analysis of tumor recurrence data. When applying the model to clinical data on breast cancer, we estimate the expected number of clonogenic cells which give rise to early and late recurrences and their progression rate parameters. The prime object of our concern is discrimination between the true recurrence and a new cancer of the same histological type on the basis of the temporal characteristics of tumor latency. As evidenced by the data analysis, such a discrimination is feasible and allows to conclude that the contralateral breast cancer may be interpreted as a preexisting subclinical tumor at the time of treatment.

1. Introduction

The latent period of carcinogenesis (some authors call it the induction period) can be defined as the time from the start of exposure to a carcinogen to the detection (diagnosis) of an overt tumor. We will use the term "tumor onset" to mean the appearance of the first detectable tumor of the type being studied, but it will always be assumed that the tumor onset time is directly observable and can be treated in the same manner as lifetime, or failure time, in survival analysis. Even under controlled experimental conditions the value of a latent period is subject to considerable interindividual variations; therefore, it must be thought of as a random variable (r.v.).

Let U be a nonnegative r.v. representing the latent period of tumor development in an individual sampled from this homogeneous population, and let $G(u)$ be its cumulative distribution function (c.d.f.). It is obvious that the probability of tumor onset within the time interval $[0,t)$ is equal to $G(t)$, that is,

$$\Pr(\text{tumor onset before } t) = \Pr(U \leq t) = G(t). \tag{1}$$

One needs to make statistical inference (estimation, hypotheses testing, etc.) on the probability $G(t)$ from the time–to–tumor observations.

Despite considerable advances that have been made in nonparametric statistical methodology, parametric failure time models remain to be one of the most extensively studied areas in modern survival analysis (see Kalbfleisch and Prentice, 1980; Lawless, 1981; Miller, 1981; Cox and Oakes, 1983; Cohen and Whitten, 1988; for surveys). When applied to the statistical analysis of tumor incidence data, a pertinent parametric method has the following distinct advantages:
(1) it allows a natural interpretation in terms of parameters endowed with biological meaning,
(2) it provides a prediction of the time–varying cancer risk forward in time beyond the follow–up period,
(3) it offers a means of estimating a surviving fraction (probability of tumor cure in studies of treatment efficacy) from time–to–tumor observations,
(4) it forms the basis for designing optimal surveillance strategies,
(5) it allows to overcome a nonidentifiability aspect inherent to some statistical problems within the nonparametric framework.

The above listed strengths of parametric approach may well outweigh its weaknesses associated with strong distributional assumptions if a model is sufficiently realistic and produces reasonable results. It is obvious that a more reliable and substantive inference from real biomedical data is provided by using biologically–based models, rather than by selecting a suitable distribution for the time of tumor latency among standard parametric families.

2. Markovian Models of Carcinogenesis

Quite an ample literature exists concerning mathematical modeling of carcinogenesis irrespective of its origin. The models are usually aimed at risk assessment in a population affected by irradiation or chemical carcinogens but the basic line of reasoning applies to the phenomenon of tumor recurrence as well. The majority of the models in either way use the elements of the birth–and–death stochastic processes theory. A comprehensive analysis of this class of models is given by Tan (1991). A distinct group is formed by two–stage models because it is generally believed that only two mutation–like events in the course of neoplastic transformation are biologically meaningful (Knudson, 1990; Tan, 1991).

The most popular in current literature is the two–stage model developed by Moolgavkar and colleagues (Moolgavkar and Venzon, 1979; Moolgavkar and Knudson, 1981; Moolgavkar and Dewanji, 1988; Moolgavkar, Dewanji and Venzon, 1988; De-

wanji, Venzon and Moolgavkar, 1989; Moolgavkar et al., 1990; Moolgavkar, Luebeck and de Gunst, 1990; Luebeck and Moolgavkar, 1991).

The model is based on the following assumptions:

(1) Tumors arise from a single malignant progenitor cell.

(2) Each susceptible cell in a tissue becomes transformed independently of other cells. The authors believe that the target for carcinogen action is the population of stem cells, and genomic events that cause the transformation of a cell occur during cell division.

(3) In the small interval of time $(t, t+\Delta t)$ a normal susceptible cell divides into two cells of the same type with probability $\alpha_1(t)\Delta t + o(\Delta t)$; it dies (or differentiates) with probability $\beta_1(t)\Delta t + o(\Delta t)$; and it divides into one normal and one intermediate cell (a cell that has sustained the first genomic event) with probability $\mu_1(t)\Delta t + o(\Delta t)$; the usual independence hypotheses for the birth–and–death process are also to be accepted.

(4) Likewise, an intermediate cell divides into two intermediate cells or dies with rates $\alpha_2(t)$ and $\beta_2(t)$, respectively, and it divides into one intermediate and one malignant (second genetic event) cell with rate $\mu_2(t)$.

(5) Once a malignant cell is generated, its growth is deterministic, and it takes a nonrandom time for the tumor to become detectable.

Let ξ be the time until appearance of the first malignant cell in a tissue. To derive the incidence rate function

$$h(t) = \lim_{\Delta t \to 0+} \frac{1}{\Delta t} P\{t \le \xi < t + \Delta t \,|\, \xi \ge t\},$$

the three–dimensional birth–and–death stochastic process should be considered. Let $X(t), Y(t)$ and $Z(t)$ represent, respectively, the numbers of normal, intermediate, and malignant cells at time t. Let $\Psi(s_1, s_2, s_3; t)$ be the probability generating function of $\{X(t), Y(t), Z(t)\}$ given at $t = 0$ the intial condition: $X(0) = 1, Y(0) = 0, Z(0) = 0$. Then obviously,

$$h(t) = -\frac{\Psi'(1,1,0;t)}{\Psi(1,1,0;t)},$$

where $\Psi'(1,1,0;t)$ is the derivative of $\Psi(1,1,0;t)$ with respect to t. Partial differential equations are available for this or some auxiliary generating functions allowing solution in a few special cases (Tan, 1991; Tan and Chen, 1991; Tan and Chen, 1993).

Moolgavkar and Venzon (1979) proved that

$$h(t) = \mu_2 \, \mathbf{E}\{Y(t) \,|\, Z(t) = 0\}. \tag{2}$$

They noticed that if the probability of a malignant cell occurrence at time t is close to zero, then one may replace (2) by the approximation

$$h(t) \approx \mu_2(t) \, \mathbf{E}\{Y(t)\},$$

which allows to obtain the explicit expression for the incidence rate

$$h(t) \approx \mu_2(t) \int_0^t \mu_1(u) \, \mathbf{E}\{X(u)\} \, e^{\int_u^t [\alpha_2(x) - \beta_2(x)] dx} \, du, \tag{3}$$

where

$$\mathrm{E}\{X(t)\} = e^{\int_0^t [\alpha_1(x)-\beta_1(x)]dx}.$$

To be sure, this expression will retain its form if one assumes the function $X(t)$ to be nonrandom, such an assumption is familiar with modeling the development of adult tumors. Formula (3) is substantiated by assuming that μ_2 is negligible (Chen and Moini, 1990; Tan, 1991). Moolgavkar, Dewanji and Venzon (1988) have shown that expression (3) overestimates the actual incidence rate predicted by the model. Their preliminary calculations allowed to suggest that approximation (3) is quite good provided that no more than 50% – 60% of individuals develop tumors. A study of the approximation adequacy in the computationally feasible case of constant parameters has demonstrated that the approximation becomes worse with increasing α_2 at fixed $\alpha_2 - \beta_2$. It must be emphasized that it was formula (3) that was confidently used in many papers devoted to the carcinogenic risk assessment. Limitations of this simplified version of the model by Moolgavkar and colleagues are discussed at length by Chen (1993). More recent versions of the two–stage model do not involve an explicit description of normal cells; it suffices to specify the number of first–generation initiated (intermediate) cells by a nonhomogeneous Poisson process (Yang and Chen, 1991).

By prescribing the relationships between a carcinogen dose and parameter values, the dose–response versions of the model have been proposed (Chen, Kodell and Gaylor, 1988; Chen and Moini, 1990; Krewski and Murdoch, 1990). Potentially transforming damage repair can also be taken into consideration within the Markovian framework (Kopp–Schneider and Portier, 1991). This class of models remains very promising for analysis of experimental evidence, including initiation/promotion experiments (Portier and Edler, 1990; Kopp–Schneider and Portier, 1991; 1992 a; 1992 b) and dose–rate effects in the case of radiation–induced cancers (Moolgavkar et al., 1990). Elucidation of the limits of its application is far from complete. At the same time, however, it is not completely suitable for estimation purposes because of its rather complicated structure and the large number of unknown parameters involved.

Another common weak point in current Markovian models of carcinogenesis is that the description of tumor progression is not sufficiently advanced. In particular, the deterministic growth of a population of malignant cells assumed in the Moolgavkar–Venzon–Knudson model may be reasonably considered an oversimplification of reality. Clearly, the time to observing a tumor is not equal to the time at which the first malignant cell is generated. There is no relevant evidence allowing to neglect the time of tumor progression comparing with the duration of earlier stages of carcinogenesis. The only observable event, at least in carcinogenicity experiments without histopathological examination of premalignant lesions, is the detection of a tumor. One may proceed from the time to this event and its distribution in order to apply methods of modern survival analysis.

In the works by Chen and Farland (1991) and by Tan and Chen (1993), the time to tumor is thought of as the time to the occurrence of the first tumor cell (type Z) giving rise to such a birth–and–death process (specified by the parameters for Z cells) that will not become extinct, the extinction probability being equal to $1 - q(t)$. But the

time it takes for this Z cell to produce a detectable tumor (progression time) is not incorporated in the model. One way of doing this is to introduce a critical upper level (random or otherwise) of the number of tumor cells at detection. Then the dynamics of Z cells can be described as a birth–and–death process with two absorbing barriers and the first passage time with respect to the upper barrier will correspond to the time of tumor progression. An analytical form of the sought–for distribution is available for the linear homogeneous birth–and–death process with large values of the upper barrier (Saaty, 1961). The strictly Markovian property of the first passage time allows the convolution of the two distributions. This approach would make the probability $q(t)$ unnecessary. When employing the idea of an absorbing upper barrier, one assumes that, once the critical level has been attained, no further substantial reduction of tumor size is possible.

Another way of looking at this problem was proposed by Yang and Chen (1991). They assumed that each primary initiated cell gives rise to a birth–and–death process, $X(u,t)$, where $X(u,t)$ denotes the number of living initiated cells that are descendants of a primary initiated cell generated at time u. An X process produces malignant cells at a rate $cX(u,t), c > 0$. Let $Z(\eta,t)$ be the number of living malignant cells that are descendants of a first–generation malignant cell produced by an X process at some random instant η. $Z(\eta,t)$ is considered as a supercritical nonhomogeneous birth–and–death process. Denote by $Y(u,t)$ the number of malignant cells produced during $(0,t]$, then the total number of living malignant cells at time $t \geq u$, given $X(u,u) = 1$, is expressed by

$$K(u,t) = \begin{cases} \sum_{i=1}^{Y(u,t)} Z_i(\eta_i,t)\,, & Y(u,t) > 0, \\ 0\,, & Y(u,t) = 0, \end{cases}$$

where $Z_i(\eta_i,t)$ is the number of malignant cells at time t in the tumor that originates at time η_i.

The main concern in the paper of Yang and Chen (1991) was with animal bioassays that are commonly used in hepatocarcinogenesis studies. For this type of data, they assumed that at most one tumor can be identified in a focus by the examination of necropsied animals and that a tumor must be at least m cells in size to become detactable. The probability of observing a tumor at time t may be defined by introducing a threshold value, m, for the total number of tumor cells, $K(u,t)$, providing the time of initiation, u, is known. Yang and Chen (1991) gave an alternative definition,

$$\Pr\{Z_i(\eta_i,t) \geq m, \text{ for some } i = 1,...,Y(u,t) \,|\, X(u,u) = 1\}.$$

To compute this probability, Yang and Chen (1991) proposed a method implying numerical solution of a Riccati differential equation. In principle, serial sacrifice experiments provide the wanted information for estimating the model parameters, but the estimation procedure, even in computationally feasible cases, is quite tedious (Tan and Chen, 1991; 1993). As already noted above, this is true for the majority of

multistage models of carcinogenesis. Therefore, the search for new ways of modeling carcinogenesis, apart from those connected with Markov processes, seems to be quite reasonable. In the next section, such a possibility will be explored with regard to the process of tumor recurrence.

3. A Simple Stochastic Model of Tumor Recurrence

A useful model can be developed through considering the time of tumor latency within the random minima framework as was proposed by Hoang et al. (1993). The idea underlying this nonthreshold model is very simple. At the end of the treatment, the cells that will propagate into a newly detectable tumor – we call them clonogens – are surviving neoplastic cells capable of giving rise to tumor regeneration. Consider the case when tumors are exposed to large single doses of radiation or chemotherapy. In this case it is natural to assume that the number N of clonogens prior to irradiation is very large but the probability, p, of their survival after the treatment is very low. If N is nonrandom, one may confidently consider the number, ν, of surviving clonogens as a Poisson random variable (r.v.) The probability η of tumor cure (no surviving clonogens) is given by

$$\eta = \Pr(\nu = 0) = e^{-\theta} \tag{4}$$

where $\theta = Np$ is the mean number of clonogens surviving the treatment.

If N is random the situation is not simple except when N is also a Poisson variable – in which case formula (4) remains valid. Considering that cell proliferation might occur during the time intervals between successive fractions of radiation, in principle one can no longer expect the number of surviving clonogens to be Poisson. In a computer experiment, Tucker, Thames and Taylor (1990) showed that deviations from Poisson statistics might result in a bias of about 10% for the estimate of the probability of cure in most standard treatment regimens. However in our view this (small) bias has been overestimated due to the chosen probability of cell division between consecutive fractions. This probability – set by the authors to be 0.4 – is too high in view of

* the typical mitotic cycle duration in tumors,
* the asynchronous entry of cells into the prereplicative period after irradiation,
* the radiation-induced block of DNA synthesis and mitosis which frequently exceeds the one-day interval chosen by Turker, Thames and Taylor (see the discussion in Yakovlev and Zorin, 1988).

Therefore, we rejoin most authors in maintaining the assumption that the number of remaining clonogens is a Poisson variable. Yakovlev (1993) discussed the findings of Tucker, Thames and Taylor (1990) from the viewpoint of theory of branching stochastic processes.

Each surviving clonogenic cell possesses in the long run the capacity of giving rise to an overt tumor. Let X_i be the random time for the i-th clonogen to produce a detectable tumor. By analogy with the terminology accepted in carcinogenesis studies

we call X_i the progression time.

Remark. The notion of the progression time should not be taken too literally. It rather refers to *potential progression time*, serving to model the temporal organization of tumor latency in a relatively facile way.

Nonnegative r.v.'s $X_i, i = 1, 2, ...,$ are assumed to be independent and identically distributed with the common c.d.f. $F(x)$. This assumption is a forced one, otherwise no concise analytical form could be found for the latent time distribution function, hereafter denoted by G. The time to tumor recurrrence (latent period) can be defined as the random minimum

$$U = \min_{0 \le i \le \nu} X_i \ , \tag{5}$$

where $X_0 = +\infty$ with probability one.

If ν is a Poisson r.v. independent of the sequence $X_1, X_2, ...,$ the survivor function, $\bar{G}(t) = 1 - G(t)$, for the r.v. U can be obtained easily as follows

$$\bar{G}(t) = \Pr(U > t) = \sum_{k=0}^{\infty} \frac{\theta^k}{k!} e^{-\theta}(1 - F(t))^k = e^{-\theta F(t)} \ . \tag{6}$$

The key advantage of this model is to show explicitly the contribution of the two characteristics of tumor growth: the mean number of clonogens θ and the rate of their progression described by the function $F(t)$. Their estimation , if feasible, furnishes additional information on the biology of tumor recurrence, thereby offering a more refined interpretation of observational data. The survivor function $\bar{G}$ corresponds to a substochastic distribution and its limiting value $\bar{G}(+\infty) = e^{-\theta}$ represents the probability of tumor cure (compare with formula (4)). Note that this probability depends solely on the expected number of clonogens responsible for tumor growth.

The hazard function, $\lambda(t)$, defined with respect to $G(t)$ is

$$\lambda(t) = \theta f(t) \ , \tag{7}$$

where f is the density of the distribution F. If the progression time distribution F is unimodal, then the hazard function $\lambda(t)$ has a maximum. Note that the assumption on the exponentiality of F, $F(t) = 1 - e^{at}$, $t > 0$, should be rejected since that would correspond to the unrealistic case of a monotone decreasing hazard. To describe a possible heterogeneity of clonogens with respect to the progression time distribution, introduce k different types of tumor cells with distributions $F_j(t)$. Then the progression time distribution F is represented by a finite mixture

$$F(t) = \sum_{j=1}^{k} q_j F_j(t), \quad 0 < q_j < 1, \quad \sum_{j=1}^{k} q_j = 1 \ . \tag{8}$$

This mixture of distributions yields the independent competing risks model for the function $\bar{G}$, i. e.,

$$\bar{G}(t) = \prod_{j=1}^{k} \exp(-\theta_0 \, q_j F_j(t)), \tag{9}$$

where θ_0 is the expected total number of viable clonogens of various types existing in the treated tumor. Within the framework of this model the hazard functions λ_j are additive and

$$\lambda(t) = \theta_0 \sum_{j=1}^{k} q_j f_j(t) \ . \tag{10}$$

In view of the last formula, it is not surprising that the bimodal shape of the hazard function arises when tumor recurrences originate from two distinct subpopulations of progenitor cells as in the examples presented in Section 8.

Note that the function G, given by (6), is monotone with respect to the parameter θ. Therefore, θ induces a stochastic ordering of latent times and, introducing a prior distribution of θ, one may construct a randomized version of the model in order to describe inhomogeneity of a population of patients under study. Most convenient for this purpose are the gamma distributions and the class of $\frac{\alpha}{2}$ - stable distribution (Rachev and Yakovlev, 1993).

4. Computer Simulation of Tumor Recurrence

Parametric representation of the progression time distribution in formulas (6) and (9) is still an unsettled problem. Naturally, in specifying an analytical form of this distribution the principle of parsimony should be kept in mind. In the work of Hoang et al. (1993 a), preference was given to the two-parameter gamma distribution by virtue of its flexibility and the fact that this failure time model, very simple as it is, reflects a multistage structure of the process of tumor development. One more reason for such a choice is that the finite mixtures of gamma distributions are identifiable (Teicher, 1961; Yakowitz and Spragins, 1968), and so are competing risks models given by formula (9). The quantities X_i introduced in formula (5) are unobservable, therefore any distributional assumption on them can not be verified directly but only by fitting the function $\bar{G}(t)$. This will be accomplished in Section 8 concerned with real data analysis. When validating the model of tumor recurrence against clinical observations, use will be made of a special goodness of fit test developed for censored data and a hierarchical model for the progression time distribution.

There is another reasonable if not absolutely conclusive way to validate the progression time distribution selected for practical purposes. If we had a sufficiently realistic model of the processes underlying tumor promotion and progression it would be possible to test a specified form of the c.d.f. $F(t)$ by computer simulations. This possibility was explored with the aid of a simulation model that incorporates the description of proliferation, differentiation and death of tumor cells as well as growth control in neoplastic tissues (Ivankov et al., 1992). In like manner, one may validate the latent time distribution $G(t)$. Substantiation by this means of a given parametric family of distributions would add to one's confidence in putting it to practical use.

When constructing the model of clonal expansion (Ivankov et al., 1992), we proceeded from the following premises:

1. A proliferating cell, in its passage through the mitotic cycle, is delayed for this cycle duration which is assumed to be a gamma-distributed random variable with shape

parameter δ and scale parameter ρ . Thus the mean and the standard deviation of the mitotic cycle duration are equal to $\tau = \frac{\delta}{\rho}$ and $\sigma = \frac{\sqrt{\delta}}{\rho}$, respectively. No possibility is allowed for a cell to enter the resting phase before mitosis.

2. As a result of mitosis two daughter cells arise which either retain the capacity for further reproduction or become sterile and die the reproductive type of death. Three possible outcomes of the mitotic cycle, for irradiated tumor cells, are taken into account:

(i) both daughter cells retain the reproductive capacity;

(ii) both daughter cells are sterile;

(iii) one of the daughter cells is capable of proliferation, the other one is sterile.

Each of the above events occurs with the probabilities p_1, p_2 and p_3, $\sum_{i=1}^{3} p_i = 1$, respectively. The sterile cell is delayed for a random time obeying the exponential distribution with parameter λ. After a lapse of this time the sterile cell is eliminated from the clone.

3. Immediately after completion of the mitotic cycle every nonsterile cell goes to the resting phase and stays there until it is stimulated to either proliferation or terminal differentiation, the latter process resulting in the competence of a cell for specialized tissue functions and eventually in its death. We introduce three stages of reversible differentiation, their durations being exponentially distributed with parameter μ. A cell loses the capacity for proliferation after its passage through the third stage. The reverse process is modeled by the backward passage of a cell through the stages already passed (including the one it is staying at the moment) in the course of differentiation, and by its subsequent transition to the phase G_1 of the mitotic cycle. The temporal parameters of forward and backward passages are assumed identical. By dedifferentiation we mean transformation of a reversibly differentiated cell into a proliferating one. The fraction, d , of the resting cells set off to differentiation is assumed to be constant in time and independent of the total number of tumor cells.

4. To simulate the growth control mechanism operating in a neoplastic tissue we specify the fraction of cells entering the mitotic cycle by

$$r = \frac{1}{1 + aN^b}, \tag{12}$$

where a and b are constants, N is obtained by summing up the cells in all stages of their life cycle, i.e., proliferating, differentiating , resting and dying cells. If the value rN exceeds the current number of resting cells then some reversibly differentiated cells start the dedifferentiation process, the top priority being assigned to the cell whose differentiation process is the least advanced.

To simulate the effect of fractionated irradiation the above assumptions are supplemented with the following ones:

5. Let a sequence of fractional doses $D_1, D_2, ..., D_n$ represent the irradiation regimen. We begin with modeling the events occurring in a population of tumor cells after the

first irradiation. In doing so, we use a *multihit–one target* model of radiation cell survival (Turner, 1975) specified by the following survivor function

$$S(D) = \sum_{k=0}^{m} \frac{(xD)^k}{k!} e^{-xD}, \tag{13}$$

where D is the irradiation dose, x is the mean number of hits per unit dose, m is the critical number of hits a cell can bear without being killed. In applying expression (13) of the dose–effect relationship to simulation of irradiated cell kinetics, we proceed from a somewhat different interpretation of its parameters which will become evident subsequently. With probability $1 - S(D_1)$ every irradiated cell is classified as *damaged*, and with probability $S(D_1)$ it is considered as remaining *undamaged* after the first fractional dose. The parameter m value is taken to be the same for all phases of the cell cycle but the other parameter of radiosensitivity, x, is allowed to vary with the position of a cell in its life cycle. To specify such variations a baseline value, x_0, of the parameter x is chosen. This value is multiplied by a scale factor with values depending on the cell cycle phase. More specifically, we set this factor equal to 1.0, 1.5 and 1.0 for the phases G_1, S, and $G_2 + M$, respectively. For the differentiation stages, as well as for the G_0 – phase, the factor is assigned a value of 0.5.

The second irradiation is simulated similarly, except that the parameter m is set equal to 1 for all damaged cells and those of them found to be *dead* are eliminated from the model (interphase type of death). The simulation model is designed in such a way as to allow for a gradual increase of the parameter x_0 for undamaged cells with increasing the current total dose of irradiation. Both the undamaged and damaged cells enter the value of N in formula (12).

6. After every fraction of irradiation, each cell, no matter whether it is damaged or not, is delayed in its passage through the mitotic cycle. The radiation induced blocks $G_1 \to S, S \to G_2, G_2 \to M$ are introduced in the simulation model under discussion in much the same way as that was employed in the book by Yakovlev and Zorin (1988). The delay time, T, for every block is dependent on the fractional dose D_i, the dependence being specified by the following simple formula

$$T = T_0(1 - e^{-vD_i}), \ i = 1, ..., n,$$

given the values of T_0 and v vary depending on the mitotic cycle phase wherein a given cell is exposed to the dose D_i.

7. The processes of repair or reproductive death occur just prior to the mitotic division of a damaged cell. The enzymatic repair of radiation damage manifests as the transition of a damaged cell to the pool of undamaged cells. The probability, P, of this event is given by

$$P = P_{max} \frac{ht^2}{1 + ht^2}, \tag{14}$$

where h is a positive constant, and t is the time measured from the last irradiation. With probability η, every unrepaired cell is transfered to the pool of perishing cells

from which it is subsequently eliminated after an exponentially distributed delay with mean $1/\lambda$. With probability $1-\eta$, the unrepaired cell splits into two daughter cells entering the G_0 phase immediately afterwards. Undamaged cells die the reproductive type of death following the rules identical to those for unirradiated cells (Assumptions 1-4).

8. Each of a large number of tumors is initialized independently to contain a single progenitor cell. Irradiation is initiated at a prescribed tumor size. With the simulation of an irradiation regimen completed, the clonal growth of irradiated tumor cells is simulated until the size, N_c, of a detectable tumor is attained. Replicates of the simulation experiment yield an output sample consisting of times to tumor recurrence measured from the last irradiation.

In this study, a uniform regimen of fractionated irradiation was simulated, i. e. $D_1 = D_2 = ... = D_n = D$. The value of D was taken equal to 7 Gy. This number should be considered as arbitrary though it provides, in combination with other parameter values, a reasonably good description of reality. We are not striving to produce quantitative results as close to a particular dose–effect relationship as possible.

The following plausible values of the model parameters were prescribed: $\tau = 24, \sigma = 7.4$ (for the mitotic cycle phases: $\tau(G_1) = 12, \sigma(G_1) = 6; \tau(S) = 7, \sigma(S) = 3.5; \tau(G_2 + M) = 5, \sigma(G_2 + M) = 2.5), x_0 = 1.5, T_0(G_1 \to S) = 100, T_0(S \to G_2) = 140, T_0(G_2 \to M) = 160, v(G_1 \to S) = v(S \to G_2) = v(G_2 \to M) = 0.01, P_{max} = 0.2, h = 0.25, a = 1.3 \times 10^{-10}, b = 2, \lambda = 0.01, \mu = 0.02, d = 0.2, p_1 = 0.95, p_2 = 0.04, p_3 = 0.01, \eta = 0.5$.

The value of N_c was set equal to 10^6. The irradiation was initiated when the number of tumor cells attained a value of 0.8×10^6.

There is not a grain of prior evidence that the simple parametric model of tumor recurrence, expressed by (6), will be consistent with data generated by the comprehensive simulation model. But this is so indeed as evidenced by the results of computer simulations given below.

When the number of fractions was varied from 10 to 15, six samples were generated, each containing 950 values of the time to tumor recurrence. Each of the samples was individually centered with respect to the initial recurrence – free period. With these samples the parametric model of tumor recurrence was validated, for which purpose the c. d. f. $F(t)$ in formula (6) was specified by the generalized gamma distribution (Stacy, 1962) given by the following expression for its density

$$f(t) = \frac{\beta\varepsilon(\beta t)^{\varepsilon\alpha-1}\exp\{-(\beta t)^{\varepsilon}\}}{\Gamma(\alpha)} . \qquad (15)$$

Being a hierarchical family of distributions, expression (15) includes the two–parameter gamma distribution as a special case ($\varepsilon = 1$). The hypothesis : $\varepsilon = 1$, can be tested by the likelihood ratio test. Table 1 shows the results of testing the hypothesis for every sample resulted from computer simulations. As is seen from this table, in four cases out of six model (6) appears to be statistically consistent with simulations. This gives more grounds to use the gamma distribution for the function $F(t)$ in formula (6). By

way of illustration two (for $n = 12$ and $n = 15$) estimates, based on model (6), and the corresponding nonparametric estimates of the survivor function are presented in Figure 1.

Table 1. Testing the hypothesis: $\varepsilon = 1$.

Number of fractional doses	χ^2 – statistic	degrees of freedom	significance level
10	1.0	1	$\rho > 0.3$
11	4.4	1	$\rho < 0.05$
12	4.4	1	$\rho < 0.05$
13	1.2	1	$\rho > 0.2$
14	2.6	1	$\rho > 0.1$
15	1.0	1	$\rho > 0.3$

The dose–effect curve, depicted as a function of the number of dose fractions, is given in Figure 2. The most important result is shown in Figure 3. Referring to this figure, the application of model (6) provides a reasonable estimate of the *actual* mean number of surviving clonogens. The estimated parameter θ in (6) only slightly overestimates the number of undamaged cells in this simulation study. Considering the total number of irradiated tumor cells (damaged + undamaged), only some of them may be clonogenic. As of now, there is no way in which such an observation can be made except by conducting computer simulations.

5. A New Stochastic Model of Carcinogenesis

5.1. True Recurrence Versus Induced or Spontaneous Carcinogenesis

When considering the causes of local failures, one meets with three meaningful possibilities:

(i) The tumor detected after the treatment represents the true recurrence, i.e., it originates from the surviving primary tumor cells, including those from subclinical tumor foci preexisting at the time of treatment.
(ii) The observed tumor is induced by irradiation or/and chemotherapy (direct carcinogenic effect of a treatment).
(iii) The observed tumor is the new one, its appearance being due to an enhanced transformation rate and depression of the immune system in the organism treated by high doses of irradiation or/and chemotherapy (indirect carcinogenic effect of a treatment).

Case (i) seems most likely to be an explanation of cancer recurrence in the majority of clinical situations, but the other two cases can not a priori be excluded from theoretical consideration. At present there are no pathological or clinical criteria for discrimination between these possible causes of local failures. An appropriate solution to the problem may hopefully be found by studying the temporal characteristics of tumor latency, this issue being the prime concern of this section. In Section 8, we will present results by Hoang et al. (1993) for the contralateral breast cancer showing that discrimination between true recurrence and spontaneous carcinogenesis is feasible. In this section we consider theoretical aspects of the problem, taking advantage of a recently proposed model (Klebanov, Rachev and Yakovlev, 1993) which includes the description of radiation–induced and spontaneous carcinogenesis as special cases. A non–stationary generalization of this model was given by Yakovlev, Tsodikov and Bass (1993).

5.2. Assumptions and Notation

In constructing a model of carcinogenesis within the random minima framework, we proceed from the following assumptions: (A) The primary event in the process of carcinogenesis is the formation of an intracellular lesion which is potentially carcinogenic, i.e., it is capable of resulting in neoplastic transformation. One may see these precancerous lesions (located in different target cells) as possessing in the long run the capacity for producing a detectable tumor. Such primary events occur at random time instants and their sequence in time may be thought of as a point stochastic process. We specify this process by a Poisson one with intensity $h_0(t)$, so that the number of lesions $\nu_0(T)$ accumulated by time T is a Poisson r.v. with expectation $\int_0^T h_0(t)dt$. Considering radiation–induced cancers, it follows from the physical nature of ionizing radiation that the lesion occurences in the course of prolonged irradiation may be identified as the points of the Poisson process. In the case of an acute irradiation the number of radiation–induced lesions is also expected to be Poissonian in accordance with the "hit and target" principle (Turner, 1975). If this process on the time interval $(0, T]$ is superimposed on another one caused by other environmental factors, then we consider the superposition of two independent Poisson processes, their intensities being additive. The rationale of the Poisson character of the process of background lesion formation (spontaneous carcinogenesis) lies in the well–known asymptotic properties of the superposition of a large number of independent point processes (Cox and Isham, 1980, page 109).
(B) At present there is hardly a shadow of doubt that cells are endowed with a capacity to repair radiation and chemical injury, including injuries that result in cancer induction (Ainsworth, 1982; Raaphorst et al., 1990; Zhu and Hill, 1991). A potentially transforming damage repair is taken into consideration within the framework of a Markovian–type model of carcinogenesis developed by Kopp–Schneider and Portier (1991). It is natural to assume participation of the same repair mechanisms in the elimination of background lesions as well. All primary lesions are subject to repair processes but some of them remain unrecognized by the repair system and, consequently, unrepaired. Some of the lesions happen to be misrepaired due to errors in

the functioning of repair mechanisms (Tobias et al., 1980; Albright, 1989; Sachs et al., 1990). We do not distinguish between unrecognized and misrepaired lesions but consider them as a single pool of misrepaired lesions. The existing experimental evidence on the temporal characteristics of enzymatic repair of lesions (Tobias et al., 1980; Yakovlev and Zorin, 1988; Frankenberg–Schwager, 1989) in particular indicates that this process can be considered to be effectively instantaneous as compared with the typical life–lengths measured in carcinogenesis studies. Therefore, we assume that, unless there is exogeneous stimulation of repair systems, each lesion is repaired or misrepaired immediately after its origination. The repair effect is modeled as the specific thinning operation (see Cox and Isham, 1980, page 98) on the original Poisson process: with probability $1-p$ each point (lesion) is deleted independently of the others and of the whole point process. The probability p, in a general case, is allowed to be time dependent, i.e. $p = p(t)$. As a result we have a thinned Poisson process of intensity $h(t) = p(t)h_0(t)$ to represent the misrepaired lesion formation.

(C) The time from the i–th lesion formation to the observable effect, i.e. an overt tumor, eventually caused by this lesion is a r.v. X_i. We call X_i the progression time. The nonnegative r.v.'s $X_i, i = 1, 2, ...$, are assumed to be independent and identically distributed with the common cumulative distribution function $F(x)$. Denote by $\nu(t)$ the number of misrepaired lesions accumulated in the organism by the time t, and assume that the r.v. $\nu(t)$ is independent of the sequence $X_1, X_2, ...$ The latent period is defined as

$$U = \wedge_{i=0}^{\nu(t)}(E_i + X_i), \tag{16}$$

where E_i is the time of the i–th lesion formation given that this time is less than T, $\wedge$ is the minimum symbol, E_i and X_i are mutually independent and $E_0 + X_0 = +\infty$ (no lesion) with probability one. Introducing a promotion state which is independent of the unrepaired/misrepaired lesion formation and assuming that the promotion probability π is the same for all lesions, we retain the Poisson character of the r.v. $\nu(t)$. From the biological point of view this consideration presupposes that the processes of initiation and promotion are combined into a single promotion state (Kokoska, 1987). We will not try to attain a parametrization of the promotion probability π with respect to irradiation dose or dose rate. Even in dose–rate studies it seems more advantageous to regard π as a constant under given experimental conditions in order to learn more about possible trends. We usually need the dependence of p and π on characteristics of irradiation to be explicitly specified only when extrapolations are made to situations with insufficient statistical data on tumor incidence rate. It is very difficult to estimate p and π separately from the time–to–tumor observations but in the majority of applications it is not that necessary. In the sequel we will keep using the designation $\nu(t)$ to represent the number of promoted (among repaired/misrepaired) lesions accumulated by time t and the designation $h(t)$ for the corresponding intensity.

5.3. Spontaneous Carcinogenesis

The above assumptions enable us to derive the distribution function G for the random variable U given by (16). Indeed, the corresponding survivor function, $\bar{G} = 1 - G$, can be expressed by the formula of total probability as follows

$$\bar{G}(t) = \Pr\{U \geq t\} = \sum_{k=o}^{\infty} \bar{R}^k(t) \frac{(\int_0^t h(x)dx)^k}{k!} e^{-\int_0^t h(x)dx}, \tag{17}$$

where $\bar{R}$ is the conditional survivor function for the sum $E_i + X_i$ given $\nu(t) = k$. Now we can use the following property of the Poisson process (see, Cox and Isham, 1980, page 46): given that there are exactly k points in the interval $(0, t]$, these points are independent and identically distributed with density $h(x)/\int_0^t h(u)du$, $x \in (0, t]$. Then, for the distribution $R = 1 - \bar{R}$ we have

$$R(t) = \frac{\int_0^t F(t-x)h(x)dx}{\int_0^t h(x)dx}.$$

Substituting this expression for R in (17), we finally obtain

$$\bar{G}(t) = \exp\left\{ -\int_0^t h(x)F(t-x)dx \right\}. \tag{18}$$

In this expression, two substantive characteristics of carcinogenesis are confounded: the rate of formation of intracellular lesions, h, and the rate of their progression described by the function F. When $h(t)$ is constant in time, we have the following special case of (18)

$$\bar{G}(t) = \exp\left\{ -h\int_0^t F(x)dx \right\}, \tag{19}$$

which is best matched to model (2.6) as far as estimation purposes are concerned. If one selects, in accordance with the principle of parsimony, a two–parameter family of distributions to approximate the function F in (19), then there will be only 3 parameters to be estimated from the time–to–tumor observations.

Consider the hazard function, $\lambda(t)$, defined for the survivor function given by (19). It is easy to see that

$$\lambda(t) = hF(t) \leq h, \tag{20}$$

i.e., $\lambda(t)$ is a nondecreasing function bounded from above. Note, that for the existence of $\lambda(t)$ in this case the progression time distribution needs not to be absolutely continuous. Assume that $F(0) = 0$. Corresponding to (18) is the hazard function

$$\lambda(t) = \int_0^t h(t-x)dF(x),$$

which is also bounded if the rate $h(t)$ is bounded, the latter assumption being natural from the biological viewpoint.

5.4. Radiation–induced Carcinogenesis

Without any loss of generality one may confine the accumulation of lesions to the interval $(0,T]$, setting $h(t) = 0$ for $t > T$. When considering the case of prolonged irradiation, T represents the time of irradiation, assuming the background rate of lesion formation negligible. Now we have

$$\bar{G}(t) = \exp\left\{ - \int_0^{t\wedge T} F(t-x)h(x)dx \right\}, \tag{21}$$

or in the case of constant h

$$\bar{G}(t) = \exp\left\{ - h\int_0^{t\wedge T} F(t-x)dx \right\}, \tag{22}$$

where $\wedge$ is the minimum symbol.

Using standard probabilistic argument, it is easy to show that, depending on the interrelationship between t and T, the conditional survivor function $\bar{\Phi}(t,t_0)$ for the random variable U given $U > t_0, t_0 > 0$, can be expressed as follows:

$$\bar{\Phi}(t,t_0) = \begin{cases} \exp[-h\int_0^T F(t-x)dx + h\int_0^{t_0} F(t_0-x)dx] & \text{for } \mathrm{t > T > t_0}, \\ \exp[-h\int_0^{t_0} (F(t-x) - F(t_0-x))dx] & \text{for } \mathrm{t > T = t_0}, \\ \exp[-h\int_{t_0}^{t} F(x)dx] & \text{for } \mathrm{t_0 < t \leq T \leq +\infty}. \end{cases}$$

The last expression for $\bar{\Phi}(t,t_0)$ describes survival of the individuals selected at a prescribed age t_0.

Recall formula (22). Considering the particular case of prolonged irradiation at a constant dose rate, we can specify the parameter h by the formula

$$h = p\pi\Theta\frac{D}{T}, \tag{23}$$

where D is the total dose of irradiation, Θ is the expected number of lesions per unit dose, p and π were defined in 5.2 (Assumptions (B) and (C)). From (22) and (23) we obtain

$$\bar{G}(t) = \exp\left\{ - \frac{p\pi\Theta D}{T}\int_0^{t\wedge T} F(t-x)dx \right\}. \tag{24}$$

Obviously, $G = 1 - \bar{G}$ is the ordinary cumulative distribution function when $t \leq T$ and the ratio D/T is fixed. These conditions are met when tumor incidence is observed in the course of prolonged irradiation at a constant dose rate, but the values of D and T are not prescribed in advance. To single out this important case we write

$$\bar{G}(t) = \exp\left\{ - p\pi\gamma\int_0^t F(x)dx \right\}, \tag{25}$$

assuming γ to be proportional to the dose rate. This expression coincides with formula (19) but its interpretation is different.

The survivor function $\bar{G}$, given by (24), corresponds generally to a substochastic (improper) distribution, that is, $\lim_{t\to+\infty} G(t) = A < 1$, and this is explained by the fact that $\Pr\{\nu(T) = 0\} > 0$ for every bounded value of D.

Corresponding to the survivor function $\bar{G}$ given by (25) is the nondecreasing hazard function equal to

$$\lambda(t) = p\pi\gamma F(t). \tag{26}$$

The hazard function for $\bar{G}$, given by (24), coincides with expression (26) for $t \le T$ and $h = p\pi\gamma$, but for $t > T$

$$\lambda(t) = \frac{p\pi\Theta D}{T}[F(t) - F(t-T)].$$

has at least one maximum which is attained either at $t = T$ or at some point $t > T$.

If the background carcinogenesis can not be neglected we have to consider two competing risks with the marginal distributions given by (19) and (24). Assuming their independence, we may write

$$\bar{G}(t) = \exp\left\{ - h\int_0^t F(x)dx - \frac{p\pi\Theta D}{T}\int_0^{t\wedge T} F(t-x)dx \right\}.$$

It is beyond biological reason to introduce dissimilar forms of the progression time distributions for spontaneous and radiation–induced cancers. It is clear that the identification of this model parameters calls for the sample values of the tumor latency time exceeding the period of irradiation.

Formula (18) provides the basis for reproducing arbitrary regimen of irradiation. It is an easy task to derive formula for the fractionated irradiation consisting of $n+1$ fractions of the total dose D :

$$\bar{G}(t) = \exp\left\{ - \sum_{k=0}^{n} \frac{p_k\pi_k\Theta_k D_k}{\tau_k}\int_{t_k}^{t_k+\tau_k} F(t-x)dx \right\},\ t \ge t_n + \tau_n,$$

where the parameters p_k, π_k, Θ_k correspond to the k–th fractional dose D_k, $\sum_{k=0}^{n} D_k = D$, t_k is the time of the k–th irradiation, $k = 0, ..., n$, $t_0 = 0$, τ_k is the duration of the k–th irradiation.

In the case of short–term irradiation, it is reasonable to introduce a new parameter $\gamma_0 = p\pi\Theta$, the expected number of promoted lesions. This does not lead to loss of biological sense, because p and π, being constants at fixed dose value, cannot be separately estimated from the latent time observations but only as the product $p\pi\Theta$. Letting $T \to 0$ in expression (24), we get

$$\bar{G}(t) = \exp\{-\gamma_0 DF(t)\}. \tag{27}$$

The hazard function is given by

$$\lambda(t) = \gamma_0 Df(t), \tag{28}$$

where f is the density of F. If the progression time distribution is unimodal, then the hazard function has a maximum (compare with formula (7)).

The dose–effect dependence in the above considered models is determined by relation (23). Other parametrizations (e.g. the linear–quadratic dose–effect model) are also possible without violating the model structure as far as the temporal organization of tumor latency is concerned.

5.5. Discrimination between Spontaneous Carcinogenesis and Tumor Recurrence

The radiation–induced cancer risk will be our initial concern. Recalling the model of tumor recurrence given by (6), it is easy to see that its structure is similar to that of model (27), the only distinction is the biological meaning of the parameters involved in the description of these entirely different processes. Therefore, one may conclude that cases (i) and (ii), mentioned in Section 5.1, are formally similar in the parametric analysis for a fixed dose value. In other words, the time–to–tumor observations provide insufficient information to discriminate between the two models. Radiation-induced cancers are believed to have longer latencies than the true tumor recurrrence, but considering the randomness of the latent period this observation cannot be easily put into use in practice.

At first glance it would seem that the difficulty could be overcome with the help of dose-effect considerations. Actually, the survival of irradiated tumor cells is a decreasing function of dose, while the radiation-induced cancer incidence exhibits an extremum in the dose-response curve (Puri, 1982). Since at high doses the radiation-induced cell killing causes a reduction of cancer incidence, expression (27) should be modified in such a way as to take into account the survival of normal cells. Let $S(D)$ be the probability of cell survival at dose D, then (27) becomes

$$\bar{G}(t;D) = \exp\left(- \gamma_0 D S(D) F(t)\right) . \tag{29}$$

Assuming that $S(D)$ has a hazard rate $\lambda(D)$, it is easy to obtain the necessary condition for the minimum of $\bar{G}(t;D)$ with respect to D:

$$\lambda(D) = \frac{1}{D^*}$$

at some point $D^* \in (0, +\infty)$. In particular, the "multihit–one target" model (Turner, 1975) of radiation cell survival satisfies this condition. In clinical applications, the function $S(D)$ should be represented in such a form as to describe multifractional regimens of irradiation (Thames, 1985; Thames, 1987; Hanin, Rachev and Yakovlev, 1993; Hanin, Pavlova and Yakovlev, 1993) although the corresponding generalization of formula (29) appears to be quite combersome. To reveal the dissimilar behavior of the dose–effect relationship, regression analysis – with the irradiation dose as covariate – is faced with encompassing a sufficiently wide range of dose values. It seems doubtful, however, whether relevant data are available to provide such an analysis.

Besides, the radiation–induced cancer risk is expected to be fairly small as compared to the risk of tumor recurrences.

On the contrary, the model of spontaneous carcinogenesis possesses distinctive properties that can be used to discriminate between cases (i) and (iii), when analyzing the temporal aspect of tumor recurrence. As indicated earlier, the hazard rate for this model is a nondecreasing function bounded from above. Thus, this function exhibits a drastically different temporal pattern as compared with the hazard function in (7) which is usually of extremal type. The dissimilar behaviour of hazard rates in the two models can be a useful indicator for discriminating cases (i) and (iii) on the basis of time–to–tumor data. A preliminary distinction could be made with the aid of nonparametric estimators for the hazard function when the sample is sufficiently large and does not contain too many censored observations. But it is the parametric model that can provide a large part of explanation of an observed pattern. This idea will be used quite advantageously in the analysis of relapse–free time data for the contralateral breast cancer (Section 8).

6. Randomized Models and Associated Limiting Distributions

6.1. Randomization Procedure and Its Stability

Variations of sensitivity to treatment among individuals may be described by randomizing the parameter θ in formula (6). Let us slightly generalize this formula, assuming the parameter θ is random itself. If we postulate that θ is gamma distributed with shape parameter r and scale parameter β, then

$$\bar{G}(t) = \left(\frac{\beta}{\beta + F(t)}\right)^r, \quad r \geq 1. \tag{30}$$

In like manner, randomized counterparts of formulas (24) and (27) are readily obtained

$$\bar{G}(t) = \left(\frac{\beta}{\beta + \frac{p\pi D}{T}\int_0^{t\wedge T} F(t-x)dx}\right)^r, \tag{31}$$

$$\bar{G}(t) = \left(\frac{\beta}{\beta + DF(t)}\right)^r, \tag{32}.$$

Recalling that $r = 1/v^2$, where v is the variation coefficient for the r.v. Θ, one can infer from formulas (31) and (32) that *variability of individual response to irradiation reduces population risk.*

Various randomized versions of models (6), (24) and (27) can be constructed in a similar way. For instance, calculations are easy if Θ (or θ) is taken as a strictly positive $\alpha/2$ – stable r.v. (Weron, 1984). In contrast to the gamma distribution, the $\alpha/2$ – stable distribution possesses a heavy tail.

Following Klebanov, Rachev and Yakovlev (1993), consider stability of the randomization procedure. Let $M(x), M(0) = 0$ be an arbitrary cumulative distribution function for the parameter Θ and represent formulas (6), (24) and (27) in the following concise form

$$\bar{G}(t) = e^{-\Theta S(t)}.$$

By compounding $\bar{G}(t)$ with respect to $M(x)$, we get

$$\bar{G}_M(t) = \int_0^\infty e^{-xS(t)} dM(x) = S(t) \int_0^\infty M(x) e^{-xS(t)} dx.$$

Now we can see that the form of function $\bar{G}$ provides its stability to the parameter Θ randomization in the uniform metric. Indeed,

$$\sup_{t \geq 0} |G_{M_1}(t) - G_{M_2}(t)| \leq \sup_{x \geq 0} |M_1(x) - M_2(x)| S(t) \int_0^\infty e^{-xS(t)} dx$$

$$= \sup_{x \geq 0} |M_1(x) - M_2(x)|.$$

In practical applications one cannot always expect the given parameter to follow exactly the prior distribution $M(x)$. The stability property preserves the model even when there is a slight perturbation of the model structure.

6.2. Limiting Forms of the Latent Time Distribution

The first significant attempt at a stochastic discription of carcinogenesis on the basis of extreme value theory is attributed to Pike as early as 1966. His simplistic reasoning within nonrandom minima scheme leads, in particular, to the Weibull distribution of the time of tumor latency, which quite frequently provides a good fit to animal carcinogenesis data (see Durbin (1976) for a survey). As far as tumors induced by irradiation or chemical carcinogens are concerned, it is more realistic to assume that the initial number of altered (damaged) cells is random, i.e. to proceed from the random minima scheme. In Section 5 we presented a stochastic model of radiation carcinogenesis based on this assumption. This model appears to be much more flexible in applications to various experimental designs than the limiting distributions assotiated with the nonrandom minima scheme. It yields a parametric family of improper distributions for the time of tumor latency which provides a description both of the rate of tumor development and of the number of affected individuals. With this approach limiting forms of the latent time distribution are naturally expected to arise at high doses of irradiation, therefore one of the most reasonable areas of their application is the second cancer risk assessment for patients treated for cancer.

To find the form of $\bar{G}(t)$, given by (32), for large dose values, introduce the normalizing factor $N = D/\beta$. Assuming that in the neighborhood of $x = 0$ the asymptotic behaviour of the progression time distribution is polynomial, that is,

$$\lim_{x \to 0} x^{-\frac{1}{c}} F(x) = 0, \quad F(0+) = 0, \quad 0 < c < 1; \quad a > 0, \tag{33}$$

it is not difficult to show that the limiting cumulative distribution function for the r.v. $N^{c}U$ has the form of a modified log-logistic distribution,

$$\lim_{D\to+\infty} \bar{G}_{N^{c}U}(t) = \bar{K}(t) = \left(1 + at^{\frac{1}{c}}\right)^{-r}, \quad r \geq 1. \tag{34}$$

Conditions (33) define a fairly wide class of functions, so that the result (34) is quite general. The results of Gnedenko (1983) ensure the convergence for all $t \geq 0$. Similarly, the two-parameter Weibull distribution

$$\bar{W}(t) = e^{-at^{\frac{1}{c}}}$$

can be obtained as a limiting case of $\bar{G}$, given by formula (27).

The hazard rate for $\bar{K}(t)$

$$\lambda(t) = \frac{rat^{\frac{1}{c}} - 1}{c(1 + at^{\frac{1}{c}})}$$

has one maximum. On the other hand, setting $ar = \delta$, where δ is a positive constant, one can see that $K(t)$ tends to the Weibull distribution with paramrters δ and $\frac{1}{c}$ as $r \to \infty$. The rate of convergence is given by the inequality (Klebanov, Rachev and Yakovlev, 1993)

$$\sup_{t\geq 0} |\bar{K}(t) - e^{-\delta t^{\frac{1}{c}}}| \leq \frac{1}{er}.$$

Thus, distribution (34) can serve as an approximation for the Weibull distribution with monotone hazard function, the two distributions being nested in the sense of that convergence.

A more profound mathematical insight into the model properties can be given within the framework of the negative binomial random minima of independent and identically distributed random variables (Klebanov, Rachev and Yakovlev, 1993).

The limiting forms of the latent time distribution for radiation induced cancers may turn out to be contrary to reality owing to the fact that an intense cell death is expected to occur at high doses of irradiation. This may not but hinder the application of these distributions. In a region of high doses, the survival of cells is taken into account by formula (29). The dose values that call for using such a modification of the survivor function for the time of tumor latency may well be much higher than those for which the limiting forms of this function remain valid. More radiobiological experiments and clinical observations are necessary to reveal possible intersections of the two regions for tumors of differing localization and histological origin. Multidimensional generalizations of the limiting distributions considered above were discussed by Rachev and Yakovlev (1993)

7. Estimation of the Model Parameters

Within the scheme of right independent censoring the likelihood for a random sample of size n is of the form

$$L = \prod_j g(t_j) \prod_k \bar{G}_k(s_k) \,, \tag{35}$$

where $g(t) = -\bar{G}'(t)$ is the corresponding probability density function, t_j for $j = 1, ..., m$ (m random) represent the observed failure times and s_k for $k = 1, ..., n-m$ are the censored observations. If one selects a two-parameter family of distributions to approximate the function F in (6), then there will be only 3 parameters to be estimated from the time-to-recurrence observations, the estimation of which is feasible by the maximization of the likelihood function L . Because of its flexibility, and for other reasons discussed in Section 4, we choose F to be a gamma distribution with the density

$$f(t) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\beta t} \chi_{\mathbf{R}_+}(t), \quad \alpha > 1, \quad \beta > 0, \tag{36}$$

where α and β are shape and scale parameters, respectively.

In order to maximize the log-likelihood $\ell = \log L(\theta, \alpha, \beta)$ with respect to the unknown parameters θ, α, β we use the following 3–step procedure.

* Step 1: apply the random search algorithm (Zhigljavsky, 1992) that requires the specification of a domain A containing the overall maximum but not a starting point for the optimization.

* Step 2: apply the Davidon-Fletcher-Powell algorithm (Himmelblau, 1972) with the initial points provided by step 1. If the boundary of the set A is attained then go to step 3, otherwise step 2 gives the final solution.

* Step 3: apply the Zoutendijk algorithm (Himmelblau, 1972) allowing for constraints which specify A.

To simplify the computations, one may confine the search for the value of α that maximizes the log-likelihood ℓ to the set of positive integers, i.e., the problem $\max_{\theta,\alpha,\beta} \ell$ is replaced by

$$\max_{2 \le \alpha \le \alpha_{\max}} \tilde{\ell} \qquad \text{where} \quad \tilde{\ell} = \max_{\theta,\beta} \ell.$$

It suffices to take $\alpha_{max} = 25$ to cover all reasonable values of the variation coefficient $1/\sqrt{\alpha}$, the smallest being equal to 0.2 . The above-outlined numerical algorithm may readily be applied to the multicomponent model given by (9).

8. Application. Breast Cancer Recurrence

In recent years, conservative treatment of breast cancer by local surgery and/or radiotherapy has become a widely accepted alternative to mutilating mastectomy. The evaluation of such treatment techniques is often based on the risk of local recurrence. We apply the above methods to analyze data on breast cancer recurrence for 877 patients treated and followed at the Curie Institute from 1960 to 1988. Description

of the subcohort is given by Fourquet et al. (1989). The data include the localization of the recurrences in terms of their occurence in the ipsilateral (treated) breast and in the contralateral (opposite) one.

Table 2. Asymptotic likelihood inference on the ipsilateral breast cancer recurrence

Parameter	Maximum Likelihood Estimate	Asymptotic 0.95–Confidence Interval
θ_1	0.11	0.08 , 0.14
α_1	4.00	3.52 , 4.48
β_1	0.076	0.064 , 0.088
θ_2	1.07	0.00 , 3.19
α_2	5.00	1.33 , 8.67
β_2	0.012	0.00 , 0.027

Referring to formulas (6), (35) and (36), the following notation will be used:
θ – the mean number of surviving clonogens,
$\tau = \alpha/\beta$ – the mean progression time,
$\sigma = \sqrt{\alpha/\beta}$ – the standard deviation of the progression time.

First consider the contralateral breast. Plots of the parametric estimate, based on model (6), and the Kaplan–Meier estimate of the survivor function (disease–free curve) are shown in Figure 4A. The maximum likelihood estimates of the parameters θ, τ and σ are given in Table 3. Within the framework of this model the recurrence in the contralateral breast appears to originate from a small population of clonogens (mean number equalling 0.18). In other words, there seem to be preexisting subclinical tumor foci in the control breast. The χ^2 – like goodness of fit test developed by Hjort (1990) does not reject the null hypothesis even at a significance level of 0.2. This evidence against the hypothesis is very weak and we can confidently consider the model as consistent with observations. Another meanigful possibility, described as case (iii) in Section 1, is that the observed recurrence represents a new tumor caused by indirect carcinogenic effect of the treatment. To explore this possibility we turn to model (19). The estimated survivor function is depicted in Figure 4A (curve 2). Graphical analysis does not allow to discriminate between the two models. The same is true for the corresponding estimates of the hazard function as compared to the kernel estimate (Figure 4B) because it is not clear whether the true hazard rate is monotone or of extremal type.

On the other hand, there is no need whatever for testing the goodness of fit in order to decide whether or not the tumor in the contralateral breast is a new one. Model (19) gives the estimated value of $\tau = 4.8$ months which seems unrealistic in

view of the fact that even animal carcinogenesis studies (Klebanov et al., 1993) reveal longer progression times. The results of the parametric analysis favor the model of true recurrence to a greater extent.

Using the generalized gamma distribution, given by formula (15), one can validate the choice of the two–parameter gamma distribution for the function F in the model of tumor recurrence. For the data on contralateral breast cancer, the likelihood ratio test provides a rather weak evidence ($\chi^2 = 3.1$ on 1 degree of freedom) against the hypothesis $\varepsilon = 1$, thereby supporting our choice.

Table 3. Maximum likelihood estimates of the parameters for model of tumor recurrence

localization	# clonogens		time		std deviation	
	θ_1	θ_2	τ_1	τ_2	σ_1	σ_2
contralateral breast	0.18	–	140	–	99	–
ipsilateral breast	0.11	1.07	53	431	26	193
same quadrant	0.07	5.17	59	1048	34	468
other quadrant	0.05	0.34	50	315	29	157

clonogens: expected number of clonogens, time: mean progression time in months, std deviation: standard deviation of progression time

Consider next the ipsilateral (treated) breast as a whole. We proceed from the independent censoring of the data caused by recurrences in the contralateral breast because there are grounds (Asselain et al., 1993) to believe that cancers in the two breasts develop independently of one another after the treatment. Plots of the parametric estimate based on model (9) and the Kaplan–Meier disease-free curve are shown in Figure 5A. Figure 5B represents the parametric estimate and the kernel estimate (Belayev, 1987) of the corresponding hazard function. The model provides a good description of the data for $k = 2$, implying the existence of two competing populations of clonogens that give rise to the tumor recurrence. For $k = 1$, the goodness of fit test by Hjort (1990) rejects the null hypothesis at a significance level of less than 0.001. When we assume $k = 2$, the significance level is approximately 0.1, thereby indicating that the two competing risks model is consistent with the data.

Figure 1.

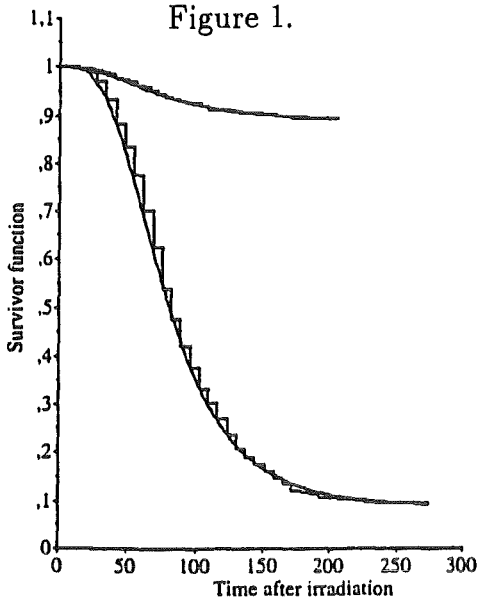


Figure 2.

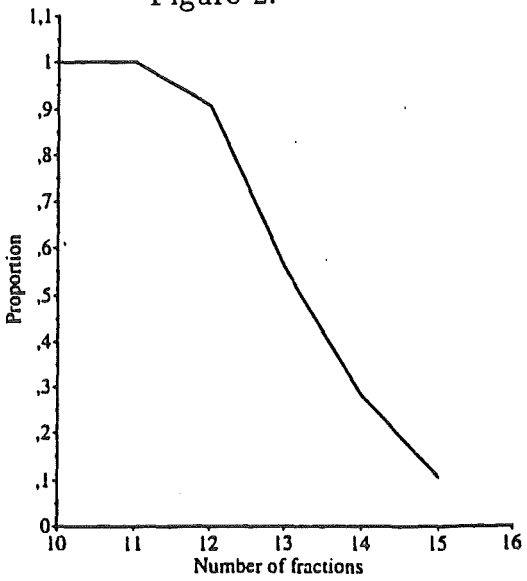


Figure 1. Parametric versus nonparametric estimation of the survivor function. Solid lines – parametric estimates, stepwise curves – the Kaplan–Meier estimates accomodated for grouped data. Upper curves correspond to the number of fractions $n = 12$, lower curves are given for $n = 15$.

Figure 2. The proportion of surviving tumors as a function of the number of fractional doses. Computer simulations.

Figure 3.

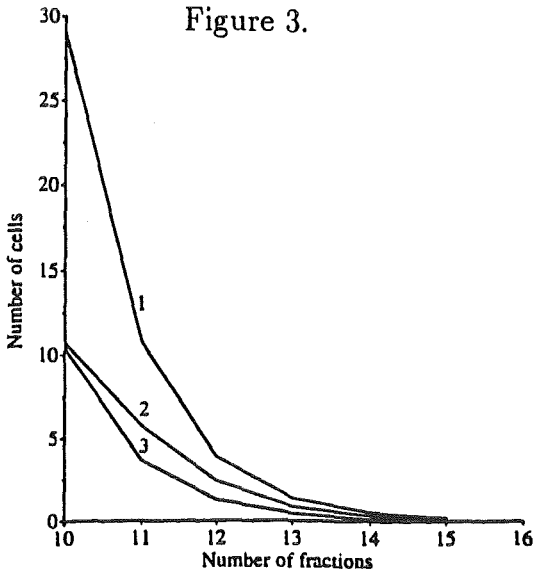


Figure 3. Estimation mean number of clonogens, 1 – the total number (damaged + undamaged) of surviving cells after a fractionated irradiation, 2 – the predicted number of surviving clonogens given by the estimated value of θ, 3 – the number of undamaged cells.

REFERENCES

Ainsworth, E.J. (1982). Radiation carcinogenesis–perspectives, in *Probability Models and Cancer*, L.Le Cam and L.Neyman, Eds, Amsterdam: North–Holland Publishing Company, pp. 99-169.

Albright, N. (1989). A Markov formulation of the repair–misrepair model of cell survival. *Radiation Research* **118**, 1-20.

Asselain, B., Fourquet, A., Hoang, T., Myasnikova, C., and Yakovlev, A.Yu. (1993). Testing the independence of competing risks: an application to the analysis of breast cancer recurrence. *Biometrical Journal*, to appear.

Belayev, Yu.K. (1987). *Statistical Methods of Incomplete Data Analysis in Reliability.* Znanie, Moscow, in Russian.

Chen, C.W. (1993). Armitage–Doll two–stage model: implications and extension. *Risk Anal.* to appear.

Chen, C.W. and Farland, W. (1991). Incorporating cell proliferation in quantitative cancer risk assessment: Approaches, issues and uncertainties, in *Chemically Induced Cell Proliferation: Implications for Risk Assessment*, Wiley–Liss Inc., pp. 481-499

Chen, J.J., Kodel, R.L., and Gaylor, D. (1988). Using the biological two–stage model to assess risk from short–term exposures. *Risk Anal.* **6**, 223-230.

Chen, C.W. and Moini, A. (1990). Cancer dose–response models incorporating clonal expansion, in *Scientific Issues in Quantitative Cancer risk Assessment*, S.H. Moolgavkar, Ed., Birkhauser, Boston, pp. 153-175.

Cohen, A.C. and Whitten, B.J. (1988). *Parameter Estimation in Reliability and Life Span Models.* Marcel Dekker Inc., New York.

Cox, D.R. and Isham, V. (1980). *Point Processes.* Chapman and Hall, London.

Cox, D.R. and Oakes, D. (1983). *Analysis of Survival Data.* Chapman and Hall, London.

Denwanji, A., Venzon, D.J., and Moolgavkar, S.H. (1989). A stochastic two–stage model for cancer risk assessment. II. The number and size of premalignant clones. *Risk Anal.* **9**, 179-187.

Durbin, N. (1976). *A Stochastic Model for Immunological Feedback in Carcinogenesis.* Lecture Notes in Biomathematics **9**, Springer–Verlag, Berlin.

Fisher, B., Anderson, S., Fisher, E. et al. (1991). Significance of ipsilateral breast tumour recurrence after lumpectomy. *The Lancet* **338**, 327-331.

Fourquet, A., Campana, F., Zafrani, B., Mosseri, V., Vielh, P., Durand, J.-C., and Vilcoq, J.R. (1989). Prognostic factors of breast recurrence in the conservative management of early breast cancer: A 25–year follow–up. *Int. J. Radiat. Oncol. Biol. Phys.* **17**, 719-725.

Frankenberg–Schwager, M. (1989). Review of repair kinetics for DNA damage induced in eukaryotic cells by in vitro by ionizing radiation. *Radiotherapy and Oncology* **14**,307-320.

Gaynor, J.J., Feuer, E.J., Tan, C.C., Wu, D.H., Little, C.R., Straus, D.J., Clarkson, B.D., and Brennan, M.F. (1993). On the use of cause–specific failure and conditional failure probabilities: examples from clinical oncology data. *J. Amer. Stat. Assoc.*

88, 400-409.
Gnedenko, B.V. (1983). On some stability theorems, in *Stability Problems for Stochastic Models.* Lecture Notes in Mathematics **982**, Springer–Verlag, Berlin, pp. 24-31.
Hanin, L.G., Pavlova, L.V., and Yakovlev, A.Yu. (1993). *Biomathematical Problems in Optimization of Cancer Radiotherapy.* CRC Press, Boca Raton, Florida.
Hanin, L.G., Rachev, S.T., Goot, R.E., and Yakovlev, A.Yu. (1989). Precise upper bounds for the functionals describing the tumor treatment efficiency. *Lecture Notes in Mathematics* **1432**, 50-67, Springer–Verlag, Berlin.
Hanin, L.G., Rachev, S.T., and Yakovlev, A.Yu. (1993). On the optimal control of cancer radiotherapy for non–homogeneous cell populations. *Advances in Applied Probability* **25**, 1-23.
Himmelblau, D.M. (1972). *Applied Nonlinear Programming.* McGraw–Hill Book Company, Austin, Texas.
Hjort, N. (1990). Goodness of fit tests in models for life history based on cumulative hazard rates. *Ann. Statist.* **18**, 1221-1258.
Hoang, T., Tsodikov, A., Yakovlev, A.Yu., and Asselain, B. (1993). Modeling breast cancer recurrence. Proceedings III International Conference *Mathematical population Dynamics*, Pau, France, to appear in *Biological Systems*, Wuerz Publications, Winnipeg, Manitoba, Canada.
Ivankov, A., Hoang, T., Loeffler, M., Asselain, B., Tsodikov, A., and Yakovlev, A.Yu. (1992). Distribution of clonogens progression time – A computer simulation study. In *Statistique des Processus en Milieu Médical*, B. Bru, C. Huber, B. Prum (eds), Université Paris V, Paris, France, pp. 287-294.
Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data.* John Wiley & Sons, New York.
Klebanov, L.B., Rachev, S.T. and Yakovlev, A.Yu. (1993). A stochastic model of radiation carcinogenesis: latent time distributions and their properties. *Mathematical Biosciences*, **113**, 51-75.
Knudson, A.G., Jr. (1990). Two–event carcinogenesis: roles of oncogenes and antioncogenes, in *Scientific Issues in Quantitative Cancer Risk Assessment*, S.H. Moolgavkar, Ed., Birkhauser, Boston, pp. 34-48.
Kokoska, S.M. (1987). The analysis of cancer chemoprevention experiments. *Biometrics* **43**, 525-534.
Kopp–Schneider, A. and Portier, C.J. (1991). The application of a multistage model that incorporates DNA damage and repair to the analysis of initiation/promotion experiments. *Math. Biosci.* **105**, 139-166.
Kopp–Schneider, A. and Portier, C.J. (1992 a). The role of clonal expansion in the use of multistage models for risk assessment. *Fundamental and Applied Toxicology* **14**, 601-613.
Kopp–Schneider, A. and Portier, C.J. (1992 b). Birth and death/differentiation rates of papillomas in mouse skin. *Carcinogenesis* **13**, 973-978.
Krewski, D. and Murdoch, D.J. (1990). Cancer modeling with intermittent exposures, in *Scientific Issues in Quantitative Cancer Risk Assessment*, S.H. Moolgavkar, Ed., Birkhauser, Boston, pp. 196-214.

Kurtz, J.M., Amalric, R., Brandone, H. et al. (1990). The prognostic significance of late local recurrence after conserving therapy. *Int. J. Radiat. Oncol. Biol. Phys.* **18**, 87-93.
Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data.* John Wiley, New York.
Luebeck, E.G. and Moolgavcar, S.H. (1991). Stochastic analysis of intermediate lesions in carcinogenesis experiments. *Risk Analysis* **11**, 149-157.
Miller, R.G. (1981). *Survival Analysis.* John Wiley, New York.
Moolgavkar, S.H., Cross, F.T., Luebeck, G., and Dagle, G. A two–mutation model for radon–induced lung tumors in rats. *Radiat. Res.* **121**, 28-37.
Moolgavkar, S.H. and Dewanji, A. (1988). Biologically based models for cancer risk assessment: A cautionary note. *Risk Analysis* **8**, 5-6.
Moolgavkar, S.H., Dewanji, A., and Venzon, D.J. (1988). A stochastic two–stage model for cancer risk assessment. I. The hazard function and the probability of tumor. *Risk Anal.* **8**, 383-392.
Moolgavkar, S.H. and Knudson, A.G. (1981). Mutation and cancer: a model for human carcinogenesis. *J. Natl. Cancer Inst.* **66**, 1037-1052.
Moolgavkar, S.H., Luebeck, G., and de Gunst, M. (1990). Two mutation model for carcinogenesis: relative roles of somatic mutations and cell prolifereation in determining risk, in *Scientific Issues in Quantitative Cancer Risk Assessment*, S.H. Moolgavkar, Ed., Birkhauser, Boston, pp. 136-152.
Moolgavkar, S.H. and Venzon, D.J. (1979). Two–event models for carcinogenesis: incidence curves for childhood and adult tumors. *Math. Biosci.* **47**, 55-77.
Pike, M.C. (1966). A method of analysis of a certain class of experiments in carcinogenesis. *Biometrics* **22**, 142-161.
Portier, C.J. and Edler, L. (1990). Two–stage models of carcinogenesis, classification of agents and design of experiments. *Fundamental and Applied Toxicology* **14**, 444-460.
Prentice, R.L., Kalbfleisch, J.D., Peterson, A.V.Jr., Flournoy, N., Farewell, V.T., and Breslow, N.E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, **34**, 541-554.
Puri, P.S. (1982). A hypothetical stochastic mechanism of radiation effects in single cells: some further throughts and results. *Probability Models and Cancer;* Le Cam, L. and Neyman, J., Eds, North–Holland Publishing Company, Amsterdam, pp. 171-187.
Raaphorst, G.P., Azzam, E.I., Feeley, M., and Sargent, M.D. (1990). Inhibition of repair of potentially lethal damage and DNA polymerases also influence the recovery of potentially neoplastic transforming damage in C3H1OT–$\frac{1}{2}$ cells. *Radiation Research* **123**, 49-54.
Rachev, S.T. and Yakovlev, A.Yu. (1993). Random minima scheme and carcinogenic risk estimation. *The Mathematical Scientist* **18**, 20-36.
Saaty, T.L. (1961). Some stochastic processes with absorbing barriers. *Journal of Royal Statistic Society, Ser. B* **23**, 319-334.
Sachs, R.K., Hlatky, L., Hahnfeldt, P., and Chen, P.L. (1990). Incorporating effects in Markov radiation cell–survival models. *Radiation Research* **124**, 216-226.

Stacy, E.W. (1962). A generalization of the gamma distribution. *Ann. Math. Stat.* **33**, 1187-1192.
Tan, W.Y. (1991). *Stochastic Models of Carcinogenesis.* Marcel Dekker, New York.
Tan, W.Y. and Chen, C.W. (1993). A nonhomogeneous stochastic model of carcinogenesis and its applications. Proceedings III International Conference *Mathematical Population Dynamics*, Pau, France, to appear in *Biological Systems*, Wuerz Publications, Winnipeg, Manitoba, Canada.
Teicher, H. (1961). Identifiability of finite mixtures. *Ann. Math. Statist.* **32**, 244-248.
Thames, H.D. (1985). An incomplete–repair model for survival after fractionated and continuous irradiations. *Int. J. Radiat. Biol.* **47**, 319-339.
Thames, H.D. (1987). Repair of irradiation injury and the time factor in radiotherapy, in *Cancer Modeling.* Statistics: Textbooks and Monographs Series **83**, Marcel Dekker Inc., New York, pp. 269-314.
Tobias, C.A., Blakely, E.A., Ngo, F.Q.H., and Yang, T.C.Y. (1980). The repair–misrepair model of cell survival, in *Radiation Biology in Cancer Research*, R.E. Meyn and H.R. Withers, Eds, Raven Press, New York, pp. 195-229.
Tucker, S.L., Thames, H.D., and Taylor, J.M.G. (1990). How well is the probability of tumor cure after fractionated irradiation described by Poisson statistic? *Radiat. Res.* **124**, 273-282.
Turner, M.M. (1975). Some classes of hit–target models. *Mathematical Biosciences* **23**, 219-235.
Weron, A. (1984). Stable processes and measures: a survey, in *Probability in Banach Spaces IV, Proceedings*, Lecture Notes in Mathematics **1080**, pp. 306-364.
Yakovlev, A.Yu. (1993). Comments on the distribution of clonogens in irradiated tumors. *Radiation Research* **134**, 117-120.
Yakovlev, A.Yu., and Zorin, A.V. (1988). *Computer Simulation in Cell Radiobilogy.* Lecture Notes in Biomathematics **74**, Springer Verlag, Berlin.
Yakowitz, S.J. and Spragins, J.D. (1968). On the idetifiability of finite mixtures. *Ann. Math. Statist.* **39**, 209-214.
Yang, G.L. and Chen, C.W. (1991). A stochastic two–stage carcinogenesis model: a new approach to computing the probability of observing tumor in animal bioassays. *Math. Biosci.* **104**, 247-258.
Zhigljavsky, A. (1992). *Theory of Global Random Search.* Kluwer, Dordrecht.
Zhu, L.X. and Hill, C.K. (1991). Repair of potentially mutagenic damage and radiation quality. *Radiation Research.* **127**, 184-189.

Numerische Lösung partieller Differentialgleichungen mittels finiter Differenzen

Ein kommentierter Rundgang
mit C.-P. Hugelmann
Kernforschungszentrum Karlsruhe HDI
Postfach 3640 D-76021 Karlsruhe

Einleitung

Das weite Feld der numerischen Behandlung partieller Differentialgleichungen in einem kurzen Überblick abzuhandeln, ist schlicht unmöglich - selbst wenn man sich, wie hier, auf die Methode der finiten Differenzen beschränkt. Mehr in der Art eines Streifzugs möchte ich Sie daher in das „Höhlen-Labyrinth" partieller Differentialgleichungen und ihrer Lösung mittels finiter Differenzen hinabführen und Ihnen von den vielen Kilometern solch eines gewaltigen Systems gerade mal ein paar gut ausgeleuchtete Meter kommentieren, aber auch an geeigneten Stellen auf die Bereiche verweisen, die wir hier nicht berühren können. Damit wir uns dabei nicht verirren, will ich Ihnen, wie seinerzeit die Ariadne dem Theseus, Leitfäden, gleich drei, an die Hand geben. Welche das sind, ergibt sich aus unserer Problemstellung, die man etwa so formulieren könnte :

Gesucht ist die Funktion f aus einem noch zu spezifizierenden Funktionenraum F, die eine partielle Differentialgleichung mit dem (Differential-) Operator D in einem Gebiet G erfüllt und darüber hinaus noch eine oder mehrere Bedingungen B auf einer Hyperfläche H von G oder der abgeschlossenen Hülle von G

$$D(f(x)) = g(x) \quad , \quad x \in G$$

$$B(f(x)) = r(x) \quad , \quad x \in H \subset \overline{G}$$

Unter der numerischen Lösung f_Δ solch eines Problems mit finiten Differenzen verstehen wir dann die Lösung eines Ersatzproblems mit einem diskretisierten Operator D_Δ in einem diskreten Gebiet G_Δ und entsprechenden modifizierten Randbedingungen :

$$D_\Delta(f_\Delta(x)) = g_\Delta(x) \quad , \quad x \in G_\Delta$$

$$B_\Delta(f_\Delta(x)) = r_\Delta(x) \quad , \quad x \in H_\Delta \subset \overline{G}_\Delta$$

Bitte erwarten Sie jetzt keine Aussagen darüber, unter welchen Voraussetzungen dies zum gewünschten Resultat führt, insbesondere wann die Gitterlösung „nahe" bei der Lösung des Ausgangsproblems liegt - diese Beschreibung hier ist ja auch viel zu allgemein. Fundamental für derartige theoretische Aussagen ist sicher die Festlegung des „richtigen" Funktionenraumes F, auf dem der Operator D wirkt. Meist wird ein „klassischer" Ansatz gewählt, die Funktion f sei eben genügend glatt (d.h. ausreichend

oft stetig differenzierbar) um eine Operatorgleichung zu erfüllen. Wie soll man dies aber sinnvoll in Bezug setzen zu der „diskreten" Lösung? Hier ergeben sich nun neue Ansätze von den finiten Element Methoden her, die ja in der „Praxis" grade mal höchstens stückweise stetige Funktionen ansetzen, und dafür einen soliden funktionalanalytischen Unterbau benötigen.

Ich möchte diesen sehr wichtigen Punkt nicht weiter vertiefen und einfach davon ausgehen, daß das gestellte Problem (sowie das Ersatzproblem) eine eindeutig bestimmte Lösung besitzt [Collatz, 1964]. Wir wollen vielmehr den wesentlichen Teilen der Beschreibung, dem Differentialoperator D, der Diskretisierung Δ und der Randbedingung B, ein wenig nachgehen, und ein paar (labyrinthartige) Verflechtungen dieser drei Leitlinien aufzeigen.

Soweit unsere „Übersicht", wir beginnen mit einigen Bemerkungen über Differentialoperatoren.

Zu den Eigenschaften von Differentialoperatoren

Beginnen wir mit dem einfachsten Operator für partielle Differentialgleichungen, dem quasilinearen oder linearen Differentialoperator erster Ordnung

$$D_1 f = a_1 \frac{\partial f}{\partial x_1} + a_2 \frac{\partial f}{\partial x_2} - a_0$$

mit

$$a_i = a_i(x_1, x_2, f) \quad \text{oder} \quad a_i = a_i(x_1, x_2) \quad , \quad i = 0, 1, 2 \quad ,$$

je nachdem ob die Koeffizientenfunktionen a_i von der Lösung f selbst abhängen oder nicht. Daran kann der für das Verständnis der Lösung partieller Differentialgleichungen wichtige Begriff der Charakteristiken am besten eingeführt werden. Sie bilden gewissermaßen die geologische Formation, aus denen die Lösungen partieller Differentialgleichungen aufgebaut sind.

Eine Charakteristik ist die Lösung folgenden Systems gewöhnlicher Differentialgleichungen :

$$\frac{dx_1}{ds} = a_1 \ , \quad x_1(0) = x_{1,0}$$

$$\frac{dx_2}{ds} = a_2 \ , \quad x_2(0) = x_{2,0}$$

$$\frac{dc}{ds} = a_0 \ , \quad c(0) = f_0$$

Dient der Punkt $f_0 = f(x_{1,0}, x_{2,0})$ einer Integralfläche f der partiellen Differentialgleichung (also einer Lösung f von Df = 0) als Anfangswert, so liegt nicht nur der Anfangspunkt, sondern die ganze Lösungskurve, eben die Charakteristik, $C:(x_1(s), x_2(s), c(s))$ in der Integralfläche. Grob gesprochen bedeutet dies, daß durch einen Punkt der Lösungsfläche bereits eine ganze Kurve der Fläche bestimmt ist. Dies bedeutet umgekehrt, daß längs der Kurve kein weiterer Punkt vorgegeben werden darf (entweder paßt er, und dann ist die zusätzliche Information redundant oder er paßt „nicht" und führt zu einem Widerspruch). Wenn man andererseits längs einer Kurve, die nicht mit irgend einer Charakteristik zusammenfällt, Punkt für Punkt die Charakteristiken berechnet und diese überdecken auf Grund der stetigen Abhängigkeit von den Anfangsdaten tatsächlich eine ganze Fläche, so ist dies eine Integralfläche, nämlich die Lösung des gerade formulierten Cauchy-Problems. Dies war auch bereits die Beweisidee von Kowalewski!

Die Menge aller Charakteristiken kann man auch quasi als das Analogon zum Steigungsfeld für eine gewöhnliche Differentialgleichung erster Ordnung auffassen.

Selbstverständlich ist dieser Begriff auch übertragbar auf partielle Differentialgleichungen mit mehr als zwei unabhängigen Variablen sowie auf Operatoren mit Ableitungen höherer Ordnung.

$$D_2 f = D_1 f + a_{1,1} \frac{\partial^2 f}{(\partial x_1)^2} + a_{1,2} \frac{\partial^2 f}{\partial x_1 \partial x_2} + a_{2,1} \frac{\partial^2 f}{\partial x_2 \partial x_1} + a_{2,2} \frac{\partial^2 f}{(\partial x_2)^2}$$

mit

$$a_{i,j} = a_{i,j}(x_1, x_2, f)$$

Auf Grund der Vertauschbarkeit der Differentiationsreihenfolge (unter gewissen Regularitätsvoraussetzungen) kann man zunächst fordern, daß $a_{i,j} = a_{j,i}$ gilt. Die Fallunterscheidung $\det(a_{i,j}) < , = , > 0$ führt dann auf die wohl allseits bekannten Klassifizierung der linearen partiellen Differentialgleichungen 2.Ordnung - hyperbolisch - parabolisch - elliptisch . Sind die

Koeffizientenfunktionen obendrein konstant, so überführt eine Koordinatentransformation den Operator je nachdem auf einen der klassischen Vertreter der Typen

d'Alembert-Operator (hyperbolisch)

$$[\,]\,f = \frac{\partial^2 f}{(\partial t)^2} - c^2 \frac{\partial^2 f}{(\partial x)^2}$$

Diffusions-Operator (parabolisch)

$$D\,f = \frac{\partial f}{\partial t} - d \frac{\partial^2 f}{(\partial x)^2}$$

Laplace-Operator (elliptisch)

$$\nabla^2 f = \frac{\partial^2 f}{(\partial x)^2} + \frac{\partial^2 f}{(\partial y)^2}$$

Das ist alles nicht neu und läuft, wie in der Geometrie, unter dem Stichwort Hauptachsentransformation. Der Typ liegt dann fest.

Der Typ kann aber auch von Teilgebiet zu Teilgebiet verschieden sein, wie etwa bei der Tricomi-Gleichung

$$x_2 \frac{\partial^2 f}{(\partial x_1)^2} + \frac{\partial^2 f}{(\partial x_2)^2} = 0$$

Sie ist in der Halbebene $x_2 > 0$ elliptisch, in der Halbebene $x_2 < 0$ hyperbolisch und für $x_2 = 0$ parabolisch. (parabolische Grenzlinie) [Collatz, 1990]. Ähnlichen „Typenmischmasch" findet man insbesondere im quasilinearen Fall.

Warum untersucht man das eigentlich, warum habe ich Sie hierher geführt? Nun jeder dieser Typen ist eine unerschöpfliche Quelle von Modellproblemen (durchaus physikalischer und nicht nur akademischer Natur) und birgt vor allem gewisse typische Eigenschaften in sich, die wesentlich für das Verständnis und für die analytische als auch numerische Behandlung sind.

Am besten untersucht ist die Laplace-, bzw. mit Inhomogenität die Poisson-Gleichung, die unter ihrem anderen Namen Potentialgleichung (die Lösungen heißen dann Potentiale) einen ganzen Zweig der Mathematik ihren Namen gab, eben der Potentialtheorie. Von hier wissen wir einiges über die Regularität der Lösungen, das Maximumprinzip, Symmetrien,

globalen Änderungen der Lösung bei lokalen Änderungen der Randvorgabe (man spricht hier von einer „unendlichen" Ausbreitung der „Störung" und es ist klar , daß sich das auch in der numerischen Lösung wiederspiegeln muß. Die Daten müssen irgenwie so „vernetzt" sein, daß jeder die Änderung des Datums eines anderen mitbekommt, wie etwa in einem Gleichungssystem. Ein deutlicher Hinweis auch für jene, die im Hinblick auf Parallelisierung von Algorithmen Domain Decomposition (eine Zerlegung also des Rechengebiets in Teilgebiete) im Auge haben. Dies ist letztlich alles nur die Folge der Tatsache, daß die elliptischen Differentialoperatoren nur ein paar komplexwertige Charakteristiken besitzen.

Dagegen besitzen die hyperbolischen Typen ein reelles Charakteristikenpaar , bzw. in höheren Dimensionen einen charakteristischen Kegel, der gleichzeitig den Einflußbereich darstellt (die Lösung der zugehörigen Differentialgleichung in einem Punkt ist durch die Wertes des Kegels und nur durch diese bestimmt). Da der Koeffizient der Zeitableitung (eine unabhängige Variable zeichnet sich von den anderen durch das Vorzeichen der Koeffizientenfunktion aus und wird als „Zeit" t interpretiert, was sich auch als physikalisch sinnvoll erweist) bei allen partiellen Differentialgleichungen von physikalischer Bedeutung konstant ist (meist = 1), sind die Charakteristiken sozusagen in die Zeitdimension hineingerichtet (die Zeit spielt die Rolle des Kurvenparameters). Mithin kann auf einer sogenannten Zeitscheibe, also für $t = t_0 = \text{const}$ keine Charakteristik liegen, diese bildet also eine ideale Vorgabefläche im Sinne eines Cauchyproblems. Das führt genau auf die Anfangswertprobleme der partieller Differentialgleichungen vom hyperbolischen Typ.

Der parabolische Typ nimmt wieder eine gewisse Zwitterstellung ein, die einzige vorhandene Charakteristik ist ebenfalls in den Zeitraum hineingerichtet, das Argument mit der Zeitebene sticht hier also auch. Hyperbolische und parabolische Operatoren führen so zu den sogenannten Evolutionsproblemen. Allerdings wirkt eine lokale Störung in einer Zeitscheibe bei einem parabolischen Problem, so wie beim elliptischen Typ, sich im vollen Raumgebiet aus. Ferner gilt, wenn auch abgeschwächt, ein Maximumsprinzip.

Bei der hyperbolischen Transport- oder Wellengleichung kann man dagegen durch ein geeignet vorgegebenes Geschwindigkeitsfeld erreichen, daß nicht nur Maximum und Minimum sondern das ganze transportierte Profil für alle Zeiten gleich bleibt. Eine „rückwärtslaufende" Zeit beschreibt dann einen Transportvorgang oder Wellenausbreitung mit umgekehrtem Geschwindigkeitsfeld. Eine Zeitumkehr beim Diffusionsprozeß widerspricht dagegen allen physikalischen Vorstellungen.

Viele weitere Eigenschaften der drei Typen ließen sich noch anführen (auf diese Einteilung bauen denn auch die meisten Lehrbücher auf, zumal dann auch via Faltung mit dem jeweiligen Greenschen Ansatz zumindest theoretisch analytische Lösungen angegeben werden können [Barton, 1989]. wichtig aber ist, diese Eigenschaften zu kennen und dafür zu sorgen, daß etwa solche Symmetrie-/ Antisymmetriebedingungen sich im Ersatzproblem widerspiegeln.

Wir aber wollen nun auf einen Sprung bei der Diskretisierung vorbeischauen.

Zur Diskretisierung des Gebietes und des Operators

Die Methode der finiten Differenzen beruht auf der Idee, den gewöhnlichen oder partiellen Differentialquotienten durch einen Differenzenquotienten zu ersetzen.

$$\frac{f(x+\Delta x)-f(x)}{\Delta x} \simeq f'(x) \quad , \quad x \in G$$

Dies ist einfach hingeschrieben (man muß in der Definition nur den Limes weglassen) und zur Rechtfertigung hat man ja auch noch die Taylorentwicklung, die dann die Quelle für höhere Ordnungen und kompliziertere Formel (etwa bei nichtäquidistanten Stützstellen) ist :

$$f_{i-1} = f(x_{i-1}) = f(x_i - \Delta x_1) = f(x_i) - \Delta x_1 f'(x_i) + \tfrac{1}{2}\Delta x_1^2 f''(x_i) \mp O(\Delta x_1^3)$$

$$f_i = f(x_i)$$

$$f_{i+1} = f(x_{i+1}) = f(x_i + \Delta x_2) = f(x_i) + \Delta x_2 f'(x_i) + \tfrac{1}{2}\Delta x_2^2 f''(x_i) + O(\Delta x_2^3)$$

$$\frac{\Delta x_1 f_{i+1} + (\Delta x_2 - \Delta x_1) f_i - \Delta x_2 f_{i-1}}{2\Delta x_1 \, \Delta x_2} = f'(x_i) + O(\tilde{\Delta}^2)$$

Für äquidistante Stützstellenverteilung $\Delta x_1 = \Delta x_2 = \Delta x$ vereinfacht sich dies zu

$$\frac{f_{i+1} - f_{i-1}}{2\Delta x} = f'(x_i) + O(\Delta x^3)$$

Hiermit ist unterschwellig eine sehr naive (oder besser: gar keine) Vorstellung über die Gitterfunktion f_Δ verbunden, die ja später durch die f_i repräsentiert werden soll. Man muß ja die Gesamtfunktion doch irgendwie im Griff haben, Bilanzen bilden etwa durch Integration über f . Dann wäre es schon zufällig, wenn dem gerade die Summe der Funktionswerte f_i entspräche. Aber man kann aus der Not eine Tugend machen und sich die f_i als Mittel über einem Intervallbereich vorstellen :

$$f_i = \frac{2}{\Delta x_1 + \Delta x_2} \int_{x_i - \Delta x_1/2}^{x_i + \Delta x_2/2} f(x)\, dx$$

Die intuitiven Formeln oben bleiben dabei in einem integralen Sinne richtig, sogar dann noch, wenn f gar nicht regulär genug für eine Taylorentwicklung ist. Wir kommen später noch auf das Verhältnis der Lösung f zu der Gitterlösung f_Δ zurück und wenden uns zunächst der Frage nach geeigneten Stützstellen im Gebiet G zu.

Unser Gebiet ist zunächst ein (Raum x Zeit)-Gebiet $G = \Omega \times [0, T]$ sofern die Zeit überhaupt involviert ist. In diesem Fall ist es üblich, gemäß den diesbezüglichen Äußerungen, in Zeitscheiben zu denken, das heißt zu verschiedenen Zeitpunkten liegt jeweils die gleiche Raumpunktverteilung vor. Dies ist wohlgemerkt eine vereinfachende Annahme, keine Notwendigkeit ! Nichtsdestotrotz wird beim Diskretisieren meist nur an die Raumverteilung der Stützstellen (in Ω) gedacht.

Hier gibt es eigentlich wieder drei verschiedene Kriterien, die unseren drei Fäden Operator-Raum-Rand zugeordnet werden können.

- Bezüglich des Operators haben wir am Differenzenquotient gesehen, daß eine äquidistante Einteilung der Stützstellen diesen besonders einfach gestaltet (was uns auch später bei der numerische Lösung wieder zu Gute kommt).

- Das gestellte Problem verlangt aber von uns manchmal eine ganz andere Verteilung der Stützstellen (da wo viel „los" ist, möchte man viele Punkte, also Lösungswerte, haben), man spricht hier von höherer Auflösung.

- Schließlich bricht auch die Beschreibung des Randes das äquidistante Raster auf, sofern nicht gerade ein Rechteckgebiet Ω vorliegt.

Diese Einteilung ist aber nicht ausschließlich, sondern man versucht alle drei Aspekte durch eine Kompromiß unter einen Hut zu bringen. Entscheidend für ein Differenzenverfahren ist die Stützstelle und ihre Verbindung zu den (unmittelbaren) Nachbarn, das Gitter. Hier nun der Versuch eine Einteilung der Gitter bezüglich der „Komplexität" dieser Verbindung vorzunehmen :

- statische Gitter
  - strukturierte Gitter
    - ein Gitterblock
      - Rechteck-Gitter
      - randangepaßte Gitter
    - mehrere Gitterblöcke
      - Gebietszerlegung
      - versetzte Gitter
      - lokale Verfeinerungen
  - unstrukturierte Gitter
    - Dreiecks-Gitter
    - polygonale Gitter
- dynamische Gitter
  - adaptive Gitter

- Free Lagrange Method

Rechteckgitter (äquidistant oder nicht) sind wohl am weitesten verbreitet in der Numerik und dienen dazu, den diskretisierten Operator (den Differenzenstern) möglichst einfach zu halten. Die Gitterpunkte (und die zugehörigen Funktionswerte) können durch eine Mehrfachindizierung leicht in Beziehung zu ihren unmittelbaren Nachbarn gesetzt werden (Erhöhen und Erniedrigen einzelner Indizes). Die Ausrichtung des Gitters entspricht dem Koordinatensystem.

Wo dies auf Grund der Gebietsform nicht möglich ist, versucht man durch Abbildung des Gebiets auf ein (logisches) Rechteck wenigstens die kanonische Nachbarstruktur zu erhalten. Das Bild eines Rechteckgitters in diesem logischen Gebiet bezüglich der Umkehrabbildung liefert im (physikalischen) Gebiet wieder ein entsprechend einfach strukturiertes Gitter, das sogenannte randangepaßte Gitter, aber für den Preis eines komplizierteren (diskreten) Operators.

Auch diese Methode droht dann zu versagen, wenn das Gebiet G nicht mehr topologisch äquivalent zu dem Rechteck ist. Man kann sich zwar meist durch „Aufschneiden" oder durch „Ausblenden" der „Löcher" helfen, aber oftmals ist es einfacher das Gebiets in topologisch einfachere Teilgebiete zu zerlegen und dafür die Verwaltung mehrerer Gitter in Kauf zu nehmen.

Nun besteht ja auch die Möglichkeit, die Teilgebiete so zu wählen, daß sie sich teilweise oder sogar vollständig überlappen. Dafür gibt es viele Anwendungsmöglichkeiten, etwa lokale Verfeinerungen zur besseren Auflösung, algorithmisch bedingter Datenaustausch (Multigrid-Verfahren, Schwarzsche Methode) oder um Funktionswerte verschiedener physikalischer Größen an verschiedenen Positionen (räumlich oder zeitlich) im Gebiet zu halten (versetzte Gitter, siehe Bild 1 des Anhangs).

Man kann natürlich auch ganz auf die durch das Rechteck (Quader, „Hypercube" - je nach Dimension) gegebene Struktur verzichten und zu Stützstellenverteilungen übergehen, deren Nachbar-Vernetzung zu Dreieckszellen oder gar Polygonmaschen führt (Delaunay-Triangulierung, Voronoi-mesh, siehe Bild 2 des Anhangs).

Die beschriebenen Gitter sind alle statisch, das heißt die (Raum-) Position der Stützstellen bleibt für alle Zeiten gleich. Es gibt aber eine Reihe von Problemen, bei denen ein dynamisches (also zeitlich veränderliches) Gitter von Vorteil ist. Als Beispiel seien solche Gebiete genannt, die sich selbst zeitlich verändern, (eine sich ausbreitende viskose Masse, Flüssigkeiten mit freien Oberflächen, ein expandierendes Gas) oder in denen sich etwas bewegt, (ein schwimmender Körper, eine Stoßfront, eine chemisch aktive Trennfläche zweier Fluids). Folgt man der „Aktion" mit den Gitterlinien, so führt das auf die adaptiven Gitter, die durch Lokalisierung (z.B. Shockfitting) und Interpolation einen zusätzlichen Aufwand erfahren. Schließlich sei noch die Free-Lagrange-Method (FLM) erwähnt, die durch eine Lagrange-

Betrachtungsweise der zu modellierenden Physik eine ständige Erzeugung oder Reorganisation des Gitters erfordert.

Diskretisierung zu treiben, ohne sich Gedanken zu machen darüber, was das diskrete Problem und seine Lösung mit dem ursprünglichen Problem zu tun hat, wäre sträflich.

Der Idealfall, wenn die diskrete (oder Gitter-) Lösung f_Δ der gesuchten Lösung f „beliebig" nahe kommt, wenn nur die Diskretisierung „fein" genug ist, wird als Konvergenz bezeichnet. Daß sich dabei auch die Operatoren D und D_Δ einander immer weiter annähern müssen, versteht sich von selbst, man spricht hier von Konsistenz. Es zeigt sich aber, daß die Konsistenz allein nicht für die Konvergenz ausreicht, vielmehr muß noch die Stabilität des diskreten Operators gefordert werden, der die Fortpflanzung von Fehlern beschränkt. Mathematisch exakt fassen läßt sich das natürlich mit einem Abstandsbegriff, der sich aus den Normen der zugrunde liegenden Funktionenräume ergibt (die Vergleichbarkeit muß dabei gewährleistet sein). In diesem Rahmen wollen wir, um die Begriffe wenigstens etwas zu erläutern, lediglich ein paar (warnende) Beispiele geben. Dazu betrachten wir, als Modellproblem, die Diffusionsgleichung mit einer Raumvariablen

$$\frac{\partial f}{\partial t} - d\,\frac{\partial^2 f}{(\partial x)^2} = 0$$

Im Jahre 1910 schlug Richardson dazu den, eigentlich plausiblen, Ansatz (Raum und Zeit sind mit zweiter Ordnung diskretisiert) vor :

$$\frac{f_j^{n+1} - f_j^{n-1}}{2\Delta t} = d\,\frac{f_{j+1}^n - 2f_j^n + f_{j-1}^n}{\Delta x^2}$$

(der obere Index gibt die zeitliche, der untere Index die räumliche Diskretisierung bezüglich äquidistanter Stützstellen wieder, also $f_j^n \simeq f(x_0 + j\Delta x\,,\, n\Delta t)$) Das Verfahren erweist sich aber, wegen des raschen Anwachsens von Fehlern, als völlig unbrauchbar - es ist absolut instabil. Wie man so etwas zeigen kann, möchte ich an dem etwas einfacheren Beispiel des FTCS-Verfahrens (Forward Time Centered Space) vorführen. Wie der Name sagt, wird die Zeitableitung durch eine vorwärtsgerichtete Differenz (also „nur" erster Ordnung), die Raumableitung durch zentrale Differenzen (wie oben) ersetzt :

$$\frac{f_j^{n+1} - f_j^n}{\Delta t} = d\,\frac{f_{j+1}^n - 2f_j^n + f_{j-1}^n}{\Delta x^2}$$

Bei von Neumanns Ansatz zur Stabilitätsanalyse wird nun für einen Funktionswert eine „Störung" angesetzt

$$f_j^n = A^n e^{ikj\Delta x}$$

und das zeitliche Anwachsen der Amplitude A (ausgedrückt durch den Amplifikationsfaktor G) beobachtet.

$$A^{n+1} = \left(1 - \frac{2d\Delta t}{\Delta x^2}(1 - \cos(k\Delta x))\right) A^n = GA^n \quad .$$

Wie man leicht sieht ist eine Dämpfung des Fehler ($|G| \leq 1$) für alle Wellenzahlen k nur dann gegeben, wenn

$$\frac{2d\Delta t}{\Delta x^2} \leq 1 \quad \text{oder} \quad \Delta t \leq \frac{\Delta x^2}{2d}$$

gilt. Das Verfahren ist also bedingt stabil, die sich ergebende Bedingung (das Stabilitätskriterium) ist typisch für viele explizite (hier läßt sich die Differenzengleichung einfach nach der neuen Zeitebene auflösen) Verfahren.

Der Wachstumsfaktor G des impliziten Verfahrens

$$\frac{f_j^{n+1} - f_j^n}{\Delta t} = d\,\frac{f_{j+1}^{n+1} - 2f_j^{n+1} + f_{j-1}^{n+1}}{\Delta x^2}$$

ist dagegen für alle Wellenzahlen

$$G = 1 \Big/ \left(1 + \frac{2d\Delta t}{\Delta x^2}(1 - \cos(k\Delta x))\right) < 1 \quad .$$

Das Verfahren ist damit unbedingt stabil. Eine Kombination („1/2 FCTS + 1/2 implizit") aus beiden, das Crank-Nicholson-Verfahren, schließlich ist ebenfalls unbedingt stabil , aber obendrein mit zweiter Ordnung in Raum und Zeit diskretisiert.

Zur Konsistenz betrachten wir nun das Rhombus-Schema (Dufort-Frankel)

$$\frac{f_j^{n+1} - f_j^n}{\Delta t} = d\,\frac{f_{j+1}^n - (f_j^{n+1} + f_j^{n-1}) + f_{j-1}^n}{\Delta x^2} \quad .$$

Durch Umordnung läßt sich dies auch schreiben als

$$\frac{f_j^{n+1} - f_j^n}{\Delta t} = d\,\frac{f_{j+1}^n - 2f_j^n + f_{j-1}^n}{\Delta x^2} - \frac{d\Delta t^2}{\Delta x^2}\,\frac{f_j^{n+1} - 2f_j^n + f_j^{n-1}}{\Delta t^2}$$

Wählt man den Grenzübergang $\Delta t, \Delta x \to 0$ so, daß das Verhältnis $\Delta t/\Delta x = c$ fest ist, wird nicht die Ausgangsgleichung, sondern die Wellengleichung

$$\frac{\partial f}{\partial t} = d\frac{\partial^2 f}{(\partial x)^2} - d\,c^2\frac{\partial^2 f}{\partial t^2}.$$

approximiert. Das Schema ist also nur bedingt konsistent (aber unbedingt stabil!).

Leider ist es nun so, daß die für eine partielle Differentialgleichung gewonnene Erkenntnis bezüglich Stabilität nicht einfach übertragbar ist auf eine andere, sie ist an den Operator gekoppelt (und nicht bloß an die Ableitung).

Wir wollen uns dies an einer anderen wichtigen Klasse von Evolutionsgleichungen, den Erhaltungsgleichungen in Flußform, klarmachen. Sie sind von der Bauart

$$\frac{\partial \rho}{\partial t} + \frac{\partial f(\rho)}{\partial x} = 0$$

und enthalten zum Beispiel die Wellengleichung -als System geschrieben-

$$\rho = \begin{pmatrix} u \\ v \end{pmatrix} \qquad u = c\frac{\partial g}{\partial x}$$

$$f(\rho) = \begin{pmatrix} 0 & -c \\ -c & 0 \end{pmatrix} \qquad v = \frac{\partial g}{\partial t}$$

$$\frac{\partial u}{\partial t} - c\frac{\partial v}{\partial x} = 0 \qquad \frac{\partial^2 g}{\partial t^2} - c^2\frac{\partial^2 g}{\partial x^2} = 0$$

$$\frac{\partial v}{\partial t} - c\frac{\partial u}{\partial x} = 0$$

Ihr einfachstes Exemplar ist durch den Fluß $f(\rho) = c\,\rho$ mit $c = \text{const}$ gegeben, die Lösung von

$$\frac{\partial \rho}{\partial t} + c\frac{\partial \rho}{\partial x} = 0$$

kann sofort zu $\rho(x, t) = \rho_0(x - c\,t)$ angegeben werden, wobei sich $\rho_0(x)$ unschwer als Anfangs(vorgabe)profil zur Zeit $t = 0$ ergibt.

Wenden wir hierauf das bei der Diffusionsgleichung erprobte FTCS-Verfahren an

$$\frac{\rho_j^{n+1} - \rho_j^n}{\Delta t} = -c\,\frac{\rho_{j+1}^n - \rho_{j-1}^n}{\Delta x},$$

so liefert die Stabilitätsanalyse dazu

$$|G| = \left|1 - i\frac{c\Delta t}{\Delta x}\sin k\Delta x\right| > 1 \quad ,$$

das Verfahren ist also hier total instabil !

Durch eine winzige Änderung (wir ersetzen ρ_j^n durch $\frac{1}{2}(\rho_{j+1}^n + \rho_{j-1}^n)$) erhält man jedoch das bedingt stabile Lax-Wendroff-Verfahren mit

$$|G| = \left|\cos k\Delta x - i\frac{c\Delta t}{\Delta x}\sin k\Delta x\right| \leq 1$$

und der sich daraus ergebenden Stabilitätsbedingung

$$\frac{|c|\Delta t}{\Delta x} \leq 1 \quad ,$$

die als Courant-Friedrichs-Levy-Kriterium (CFL) in der Literatur bekannt ist. Diese Bedingung kann geometrisch so interpretiert werden, daß der numerische Abhängigkeitsbereich (das sind die Diskretisierungspunkte in der Differenzengleichung) den physikalischen Abhängigkeitsbereich (man beachte, daß für konstantes c das Ausgangsprofil gerade um $c\,\Delta t < \Delta x$ verschoben wird) überdecken muß.

Warum funktioniert nun das Lax-Wendroff-Verfahren hier, FTCS aber nicht? Dazu ordnen wir das Schema etwas um

$$\frac{\rho_j^{n+1} - \rho_j^n}{\Delta t} = -c\,\frac{\rho_{j+1}^n - \rho_{j-1}^n}{2\Delta x} + \frac{1}{2}\,\frac{\rho_{j+1}^n - 2\rho_j^n + \rho_{j-1}^n}{\Delta t}$$

$$= -c\,\frac{\rho_{j+1}^n - \rho_{j-1}^n}{2\Delta x} + \frac{\Delta x^2}{2\Delta t}\,\frac{\rho_{j+1}^n - 2\rho_j^n + \rho_{j-1}^n}{\Delta x^2}$$

und erkennen, daß sich die beiden Schemata nur um den letzten Term rechts unterscheiden, der sich als Diffusionsterm entpuppt. Wir haben also Stabilität quasi durch numerische Diffusion erreicht.

Leider hat nun das Lax-Wendroff noch einen, in vielen Fällen, gravierenden Nachteil, es „produziert" in der Nähe steiler Gradienten des Anfangsprofils im Verlauf der Rechnung „Überschwinger" (Gibbs'sche Phänomene, siehe auch die nachlaufenden Wellen im Bild 5 des Anhangs), die zu fatalen (z.B. negative Werte von Konzentrationen) Fehlern führen. Die Ursache dafür ist letztlich in der nicht operatorgemäßen Diskretisierung der Raumableitung mittels zentraler Differenzen zu suchen. Die einzige Charakteristik der Erhaltungsgleichung ist entsprechend dem Vorzeichen der Koeffizientenfunktion $f'(\rho)$ nur entweder nach links oder nach rechts gerichtet, dürfte also auch nur von dort „Information" erhalten.

Eine adäquate Behandlung ist durch das sogenannte Charakteristiken-Verfahren (oder bekannter unter dem Namen Upwind-Verfahren) gegeben :

$$\frac{\rho_j^{n+1}-\rho_j^n}{\Delta t} = \begin{cases} +c\,\dfrac{\rho_{j+1}^n-\rho_j^n}{\Delta x} & c \le 0 \\ \text{für} & \\ -c\,\dfrac{\rho_j^n-\rho_{j-1}^n}{\Delta x} & c \le 0 \end{cases} .$$

Auch hier offenbart eine kleine Umordnung die Schwäche des Verfahrens

$$\frac{\rho_j^{n+1}-\rho_j^n}{\Delta t} = c\,\frac{\rho_{j+1}^n-\rho_{j-1}^n}{2\Delta x} - \frac{|c|\Delta x}{2}\,\frac{\rho_{j+1}^n-2\rho_j^n+\rho_{j-1}^n}{\Delta x^2} .$$

Es ist die starke numerische Dämpfung (siehe Bild 4 des Anhangs), durch den zweiten Term rechts gegeben. Diese numerische Dämpfung kann mit einer vorhandenen physikalischen Dämpfung „verrechnet" werden, für einen reinen Transportvorgang ist sie aber sehr lästig. Aber auch hier wurde ein Ausweg gefunden, es sind die Flußkorrekturverfahren, die sich grob als gewichtetes Mittel aus einem Verfahren niederer Ordnung und einem Verfahren höherer Ordnung beschreiben lassen. Bei einer konservativen Diskretisierung des Flusses f

$$\frac{\partial f(\rho)}{\partial x} \simeq \frac{f_{j+1/2}-f_{j-1/2}}{\Delta x}$$

setzt man also

$$\begin{aligned} f_{j+½} &= \Phi_{j+½} f^H_{j+½} + (1-\Phi_{j+½}) f^L_{j+½} \\ &= f^L_{j+½} + \Phi_{j+½}(f^H_{j+½} - f^L_{j+½}) \end{aligned}$$

Im eindimensionalen linearen Fall für $f(\rho) = c(x)\,\rho(x)$ kann etwa für $f^L_{j+½}$ der Upwind-Fluß

$$f^L_{j+½} = \frac{1}{2}(c_{j+½} + |c_{j+½}|)\rho_j + \frac{1}{2}(c_{j+½} - |c_{j+½}|)\rho_{j+1}$$

und für $f^H_{j+½}$ der Fluß des Lax-Wendroffverfahrens

$$f^H_{j+½} = \frac{1}{2}c_{j+½}(\rho_j+\rho_{j+1}) - \frac{1}{2}\frac{\Delta t}{\Delta x}c^2_{i+½}(\rho_{j+1}-\rho_j)$$

gesetzt werden. Die Idee dabei ist, den stark dämpfenden Charakter des Verfahrens niederer Ordnung durch den antidiffusiven Fluß $f^H_{j+½} - f^L_{j+½}$ zu korrigieren. Der Trick besteht darin, die Wichtungsfaktoren, die Limiter $\Phi_{j+1/2}$ lokal so zu bestimmmen, daß wiederum die unerwünschten Überschwinger vermieden werden können. Bei den FCT(Flux Corrected Transport)-Verfahren wurde der Limiter heuristisch (Verbot neuer Extrema) bestimmt. Die sehr aufwendigen Verfahren gestatten dabei lediglich $0 \le \Phi_{j+½} \le 1$. Ein besseres Maß dafür ist die Totalvariation

$$TV_{I}\rho_{\Delta}(\cdot,t) = \sum_{j}|\rho_{j+1}^{n} - \rho_{j}^{n}|$$

Die TVD(Total Variation Diminishing)-Verfahren basieren denn auch auf einer vollständig neuen Stabilitätstheorie bzgl. der L^{∞} – Norm und der Totalvariation (anstelle der üblichen Lax-Richtmyer-Theorie bzgl. der L^{2} – Norm. Damit wird nun zum einen eine größere Lösungsklasse erfaßt, zum anderen die unerwünschten Oszillationen vermieden, Stabilität kann formelmäßig erfaßt werden. Mit den Flußapproximationen $f_{i+\frac{1}{2}}^{L}$, $f_{i+\frac{1}{2}}^{H}$ von oben, genügt $\Phi_{i+\frac{1}{2}} = \Phi(r_{i+\frac{1}{2}})$ mit

$$r_{i+\frac{1}{2}} = \begin{cases} \dfrac{\rho_i - \rho_{i-1}}{\rho_{i+1} - \rho_i} & \text{für} \quad c_{i+\frac{1}{2}} \geq 0 \\ \dfrac{\rho_{i+2} - \rho_{i+1}}{\rho_{i+1} - \rho_i} & \text{für} \quad c_{i+\frac{1}{2}} < 0 \end{cases}$$

dem TVD-Kriterium, wenn nur

$$0 \leq \Phi_{i+\frac{1}{2}} \leq 2 \quad , \quad \Phi(r) \leq 2r \quad , \quad r \in \mathbb{R}$$

gilt. Die Wahl

$$\Phi(r) = \max(0,\min(2r,1),\min(r,2)) \quad , \quad r \in \mathbb{R}$$

erwies sich bei Testläufen als sehr günstig.

Ein Test sei hier abschließend auszugsweise vorgestellt. Es soll die lineare Advektionsgleichung

$$\frac{\partial \rho}{\partial t} + \frac{\partial (u\rho)}{\partial x} + \frac{\partial (v\rho)}{\partial y} = 0$$

mit den Geschwindigkeitskomponenten

$$\begin{aligned} u(x,y) &= -\omega \cdot (y - y_0) \\ v(x,y) &= \omega \cdot (x - x_0) \end{aligned}$$

gelöst werden. Die analytische Lösung ist eine Drehung der Anfangswerte um den Punkt (x_0,y_0) mit der Winkelgeschwindigkeit ω . Als Anfangswert wird auf einem 100x100 Gitter ein Kegel (mit der Höhe von 4 Raumeinheiten) vorgegeben (siehe Bild 3 im Anhang). Die weiteren numerischen Vorgaben waren

$$\omega = 0.1 \quad , \quad \Delta_x = \Delta_y = 1 \quad , \quad \Delta_t = 0.1 \quad .$$

Eine volle Umdrehung entspricht dann 628 Zeitschritten, gerechnet wurden 3768 Schritte (=6 Runden). Dieser zweidimensionale Fall wurde durch Splitting [Janenko, 1969] auf den oben beschriebenen eindimensionalen Fall zurückgeführt.

Im Anhang sind die Ergebnisse des Upwindverfahrens (Bild 4), des Lax-Wendroff-Verfahrens (Bild 5) und des TVD-Verfahrens mit dem „Superbee"-Limiter (Φ wie oben gewählt) vorgestellt (Bild 6).

Bemerkungen über künstliche Ränder

„Dirichlet"- (Randwerte vorgegeben), „Neumann"- (Ableitungen am Rand vorgegeben), „periodisch" sind die klassischen Randbedingungen. Will man den physikalischen Aspekt mehr betonen, so kann man auch von Haft- und Slip-Bedingungen, etwa für die Geschwindigkeit, oder von perfekt leitenden oder isolierenden Rändern für die Temperatur oder den Strom reden. Selbstverständlich gibt es auch alle nur erdenklichen Kombinationsmöglichkeiten und ich will hier nur auf eine besonders tückische Randbedingung eingehen, die sogenannten künstlichen Ränder. Ein künstlicher Rand entsteht dann, wenn man das zu berechnende Gebiet fern der eigentlichen physikalischen Rändern (unendlich ausgedehntes Gebiet etwa) dort abschneidet, wo die Lösung entweder hinreichend bekannt ist oder uninteressant zu werden beginnt. Zum ersteren gehören z.B. auch Einströmbereiche, die durch Meßwerte oder durch statistische Vorgaben ausreichend beschrieben werden können. Hierzu möchte ich auch jene Methoden zählen und beiseite lassen, die - durchaus mit numerischem Aufwand verbunden - auf analytische Weise das Fernfeld der Differentialgleichung, also die Lösung eines Außenraumproblems benutzen, um Randvorgaben zu machen (Dirichlet-to-Neumann-Verfahren, [Keller, 1989]). Ist die Lösung unbekannt, aber auch weitgehend uninteressant, so ist bei der Festlegung des künstliches Randes, der Schnittkante, und der dort aufgestellten Bedingung darauf zu achten, daß keine Rückwirkungen auf das eigentliche Rechengebiet zu befürchten sind - weder physikalisch, also durch die Differentialgleichung bedingt und daher gezwungenermaßen zu berücksichtigen (man spricht dann auch von Ausstrahlungsbedingungen), noch numerisch, wenn auf diese Weise ungewollt Reflexionen in das Rechengebiet zurücklaufen und die eigentliche Lösung stören.

Eine Methode basiert darauf, auslaufende Signale in einem größeren Bereich vor dem eigentlichen Rand „wegzudämpfen", sei es durch künstliche Erhöhung der physikalischen Viskosität, sei es gezielt durch numerische Dämpfung, oder „auszusieben" durch Fourieranalyse und Filterung. Diese Bereiche gehen dann meist aber für eine vernünftige Lösungsinterpretation verloren.

Eine weitere, teilweise bestechende, Methode möchte ich jetzt nur anritzen, an einem Beispiel erläutern. Sie geht von dem, ja immerhin über den künstlichen Rand hinweg geltenden, Differentialoperator aus. In unserem Modellfall sei es der Wellenoperator im Halbraum :

$$[\,]\ u(x,y,t) = (\partial_t^2 - \partial_x^2 - \partial_y^2)u = 0 \quad x,y \in \mathbb{R}\ ,\ y \geq 0$$

Mit der sehr komplexen Theorie der Pseudodifferentiale ist es möglich, oder besser gesagt erlaubt, den Wellenoperator hier formal so zu faktorisieren

$$[\,]\ u(x,y,t) = -\left(\partial_y - \sqrt{\partial_t^2 - \partial_x^2}\right)\left(\partial_y + \sqrt{\partial_t^2 - \partial_x^2}\right) u$$

und daraus eine Randbedingung

$$\left(\partial_y - \sqrt{\partial_t^2 - \partial_x^2}\right)u = 0 \quad \text{für} \quad y = 0$$

so auszuwählen, daß die vorgegebene ebene Welle

$$u(x,y,t) = e^{i(\tau t + \zeta x + \sqrt{\tau^2 - \zeta^2})}$$

$$\tau^2 > \zeta^2 \ , \quad \tau > 0 \ , \quad \zeta/\tau = \sin\varphi$$

absorbiert (wie das „funktioniert" wird eigentlich erst durch eine Fouriertransformation deutlich).

Wie setzt man das nun real um, schließlich kann man ja $\sqrt{\partial_t^2 - \partial_x^2}$ nicht programmieren? Man approximiert (Taylorentwicklung, Padé- Approximation, ...) diese „Gebilde" einfach -

Approximation von $\sqrt{1 - z^2}$	Randbedingung für $y = 0$
1	$(\partial_y - \partial_t)u = 0$
$1 - z^2/2$	$(\partial_y\partial_t - \partial_t^2 + \partial_x^2/2)u = 0$
$1 - z^2/2 - z^4/8$	$(\partial_y\partial_t^3 - \partial_t^4 + \partial_x^2\partial_t^2/2 + \partial_x^4/8)u = 0$
α	$(\partial_y - \alpha\partial_t)u = 0$

- und es funktioniert (nicht immer, die dritte Bedingung erweist sich als ill-posed im Sinne der Kreiss'schen Theorie!). Man kann auch andere Approximationen (im Sinne der Fehlerquadratmethode wäre etwa ein $\alpha = \pi/4$ besser als die 1 - das ist eine Frage der Norm) wählen und kommt dann sogar zu (im Durchschnitt) besseren Ergebnissen.

Zusammenspiel aller drei Komponenten

Wir haben nun alle drei Bereiche einmal gestreift und hie und da auch auf die Querverbindungen der einzelnen Komponenten hingewiesen. Gewis-

sermaßen in einem „Finale" kommt dann aber das Zusammenspiel aller drei Komponenten auf Sie als Numeriker zu, hat doch die Diskretisierung des Operators und der Randbedingungen zu einem Gleichungssystem geführt, das gelöst werden muß. Wenn die mühsam übertragenen Eigenschaften (wie Symmetrie, Positivität, etc.) des Operators auf sein diskretes Pendant nicht von den Randbedingungen wieder zerschlagen werden und man so glücklich zu linearen Band- oder Blockstrukturen gelangt oder auch nur die Dünnbesetztheit des diskreten Matrixoperators ausnutzen will, dann steht Ihnen eine breite Palette von Verfahren aus der Numerik zur Verfügung. Wählen Sie eine direkte Methode, ein FFT-Verfahren mit bestechender asymptotischer Aufwandsabschätzung (N log(N)) oder fahren Sie mit einer LU-Zerlegung besser, die man in jedem Zeitschritt wieder benutzen kann? Wählen Sie ein iteratives Verfahren, ein modernes cg-Verfahren oder einen klassischen SOR-Solver mit Tschebyscheff-Turbo und Red-black-Sortierung? Darf's ein Multigrid-Verfahren mit F- oder W-Zyklus sein ?

Der eigentliche Höhepunkt kommt aber dann, wenn man als Ergebnis schließlich doch das erhält, was auch immer durch die partiellen Differentialgleichungen beschrieben wird, was nur einfach der(en) Natur und ihrer schlichten Schönheit nahekommt. Dem haben sich alle „Simulanten" verschrieben.

Widmung

Vor gut einem Jahr ist Professor Dr. Winfried Schmidt durch einen tragischen Unfall ums Leben gekommen. Er hat sich mit seiner Gruppe immer wieder auf dieses „Labyrinth" der Numerik partieller Differentialgleichungen eingelassen und wäre auch sicher jetzt wieder gern bei dieser Tour dabei gewesen. Seinem Andenken möchte ich daher diesen Vortrag widmen.

Literatur

G.Barton
Elements of Green's Function and Propagation
Oxford Science Publications, 1989

Lothar Collatz
Funktionalanalysis und numerische Mathematik
Springer Verlag, Berlin, 1964

Lothar Collatz
Differentialgleichungen
B.G.Teubner, Stuttgart, 1990

R.Courant, E.Isaacson, M.Rees
On the Solution of nonlinear hyperbolic differential equations
Comm. Pure Appl. Math., Vol.5, 1952, pp243-255

R.Dautray, Jacques-Louis Lions
Mathematical Anaalysis and Numerical Methods for Science and Technology Vol. 1-6
Springer Verlag, Berlin Heidelberg New York Paris, (1984,1985)

B.Engquist, A.Majda
Absorbing boundary conditions for the numerical simulation of waves
Math. Comp., Vol.31, 1977, pp629-651

M.J.Fritts, W.P.Crowley, H.Trease
The Free-Lagrange Method
Lecture Notes in Physics Vol.238 Proceedings of the First International Conference Hilton Head Island, South Carolina, 1985 Springer Verlag, Berlin Heidelberg New York, Tokyo

W.Hackbusch
Theorie und Numerik elliptischer Differentialgleichungen
B.G.Teubner, Stuttgart, 1986

C.-P.Hugelmann
Differenzenverfahren zur Behandlung der Advektion
Wissenschaftliche Berichte des Instituts für Meteorologie und Klimaforschung der Universität Karlsruhe Nr.8 (1988)

N.N.Janenko
Die Zwischenschrittmethode zur Lösung mehrdimensionaler Probleme der mathematischen Physik
Springer Verlag, Berlin Heidelberg New York, 1969

Joseph B.Keller, Dan Givoli
Exact Non-reflecting Boundary Conditions
Journal of Computational Physics, Vol.82, (1989), pp172-192

P.D.Lax, B.Wendroff
Systems of conservation laws
Comm. Pure Appl. Math., Vol.13, (1960), pp217-237

E.L.Lindman
"Free Space" boundary conditions for the time dependent wave equation
J. Comp. Phys., Vol.18, (1975), pp66-78

R.E.Moore, I.O.Angell
Voronoi Polygons and Polyhedra
J. Comp. Phys. Vol.105, (1993), pp301-305

C.D.Munz
Monotone Differenzenverfahren zur Approximation von Schockwellen
Universität Karlsruhe, Fakultät für Mathematik, Bericht Nr. 26, 1984

R.D. Richtmyer, K.W.Morton
Difference Methods for Initial-Value Problems
John Wiley & Sons, New York London Sidney, 1967 (2nd Edition)

Patrick J.Roache
Computational Fluid Dynamics
Hermosa publishers, Albuquerque, 1982

W.I.Smirnow
Lehrgang der höheren Mathematik, Teil IV
Deutscher Verlag der Wissenschaften, 1982

J.Smoller
Shock waves and reaction-diffusion equations
Springer Verlag, New York Heidelberg Berlin, 1983

P.K.Sweby
High resolution schemes using flux limiter for hyperbolic conservation laws
SIAM J. Numer. Anal., Vol.21, 1984, pp217-235

F.Treves
Introduction to pseudodifferential and Fourier integral operators Vol.1,2
Plenum Press, New York London, 1980

L.Wagatha
Approximation of pseudodifferential operators in absorbing boundary conditions for hyperbolic equations
Numerische Mathematik, Vol.21, 1983, pp51-64

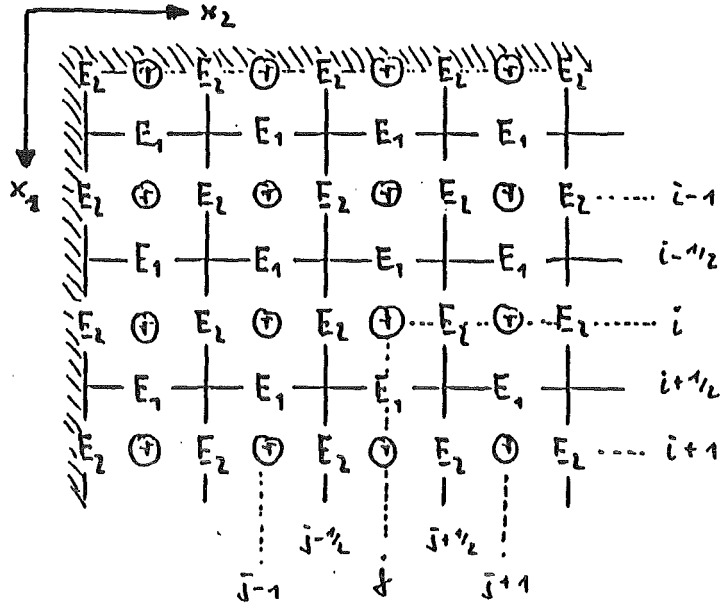
Anhang

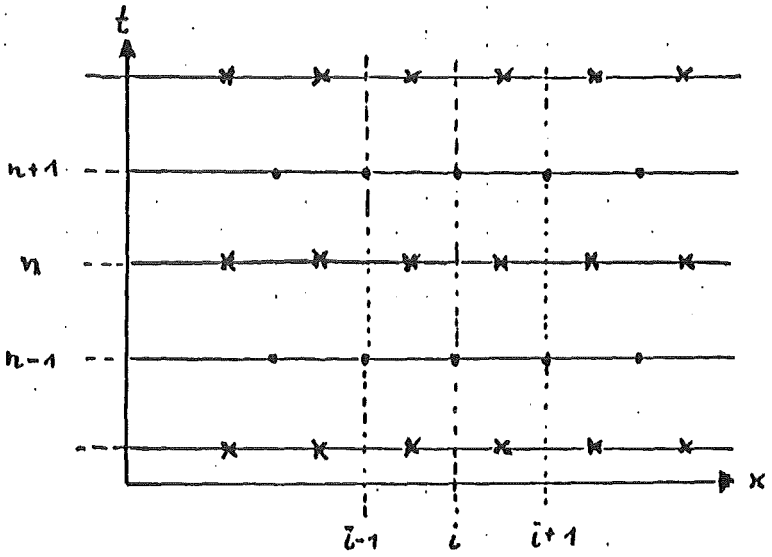
versetzte Gitter in Raum und Zeit

" staggered grid "

$$\nabla \cdot \vec{E} = \frac{\partial E_1}{\partial x_1} + \frac{\partial E_2}{\partial x_2} = r$$

$$r_{I,J} = \frac{E_{1,I+1/2,J} - E_{1,I-1/2,J}}{\Delta x_1} + \frac{E_{2,I,J+1/2} - E_{2,I,J-1/2}}{\Delta x_2}$$





" leap frog "
" midpoint rule "

$$\frac{\partial f}{\partial t} = -\frac{\partial f}{\partial x}$$

$$\frac{f_I^{n+1} - f_I^{n-1}}{\Delta t} = -\frac{f_{I+1/2}^{n} - f_{I-1/2}^{n}}{\Delta x} \qquad \frac{f_{I+1/2}^{n+2} - f_{I+1/2}^{n}}{\Delta t} = -\frac{f_{I+1}^{n+1} - f_{I}^{n+1}}{\Delta x}$$

versetzte Gitter

Bild 1

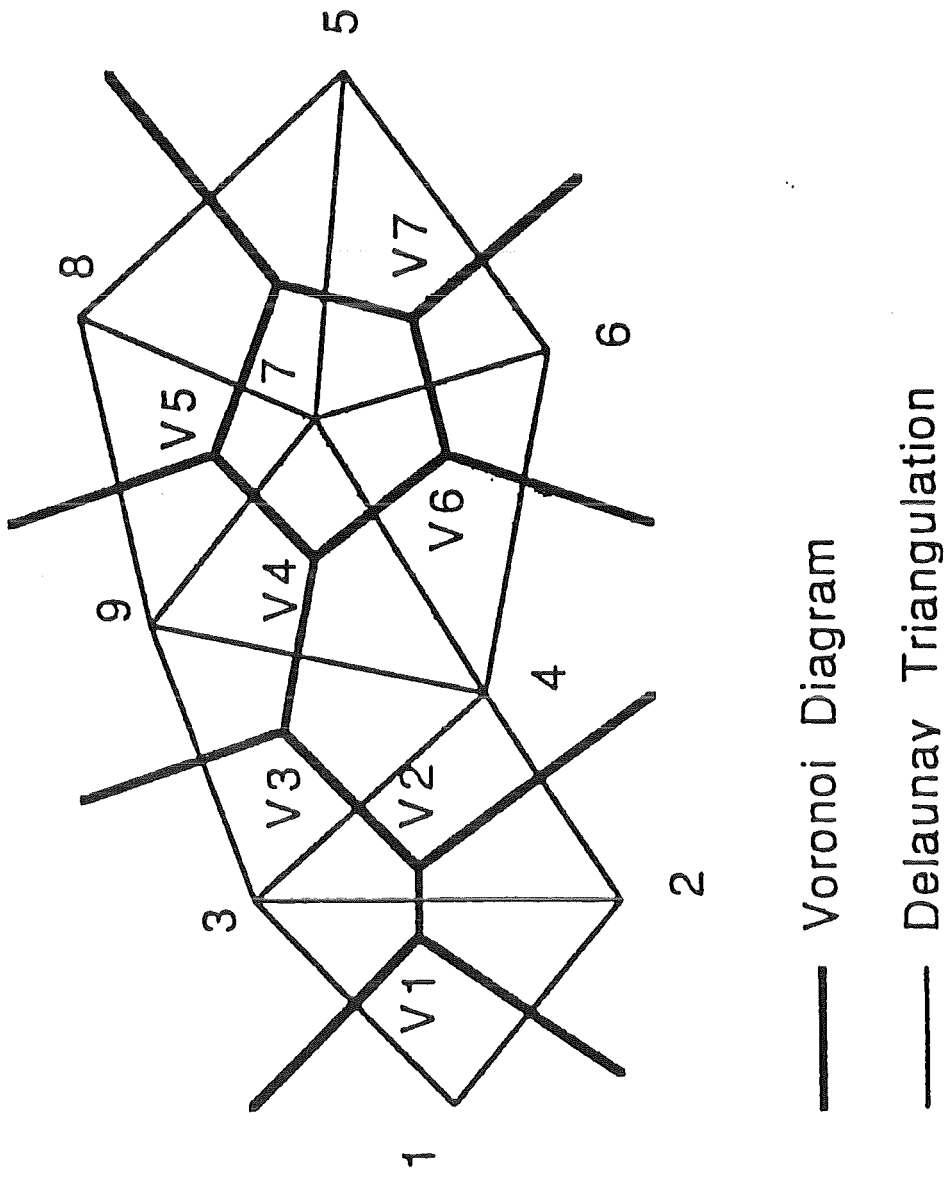
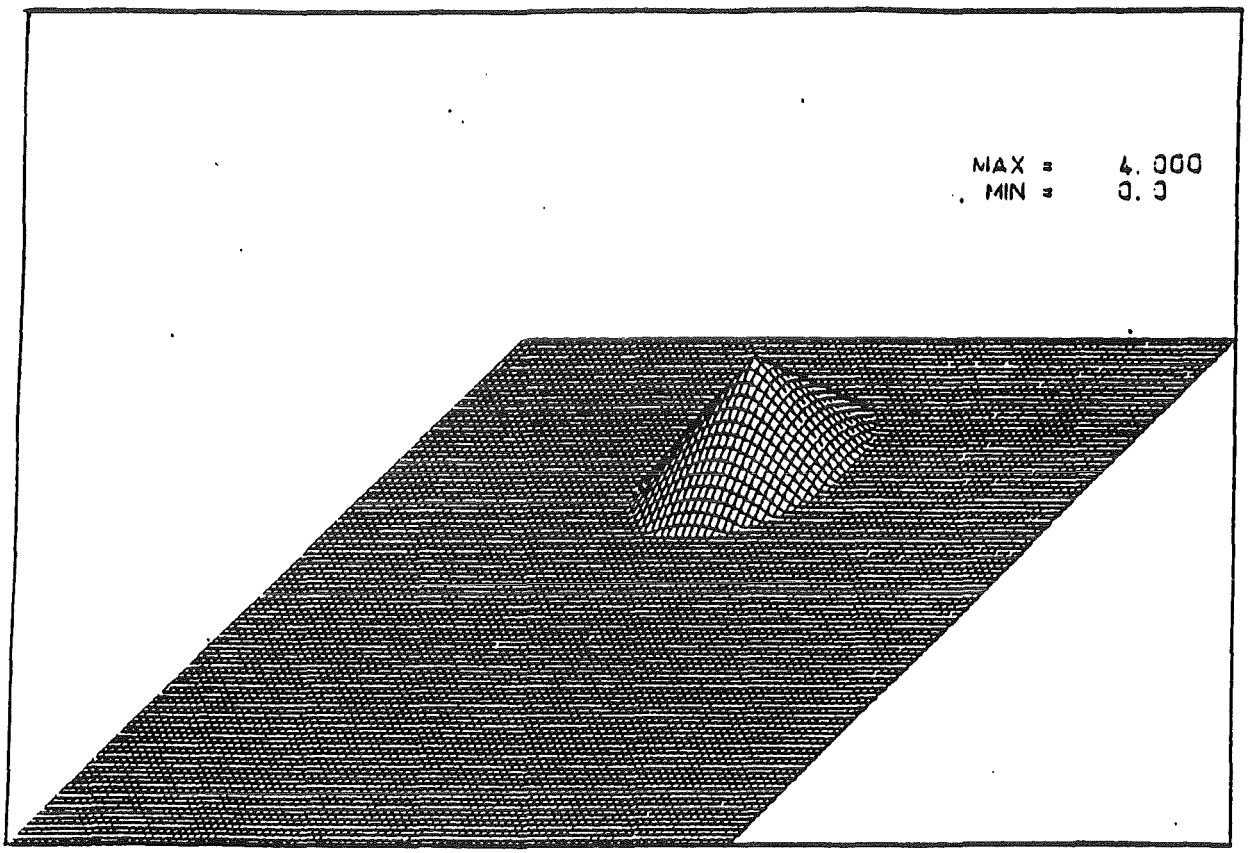
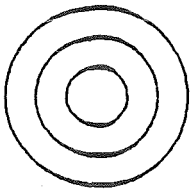

1
2
3
4
5
6
7
8
9
V1
V2
V3
V4
V5
V6
V7
Voronoi Diagram
Delaunay Triangulation

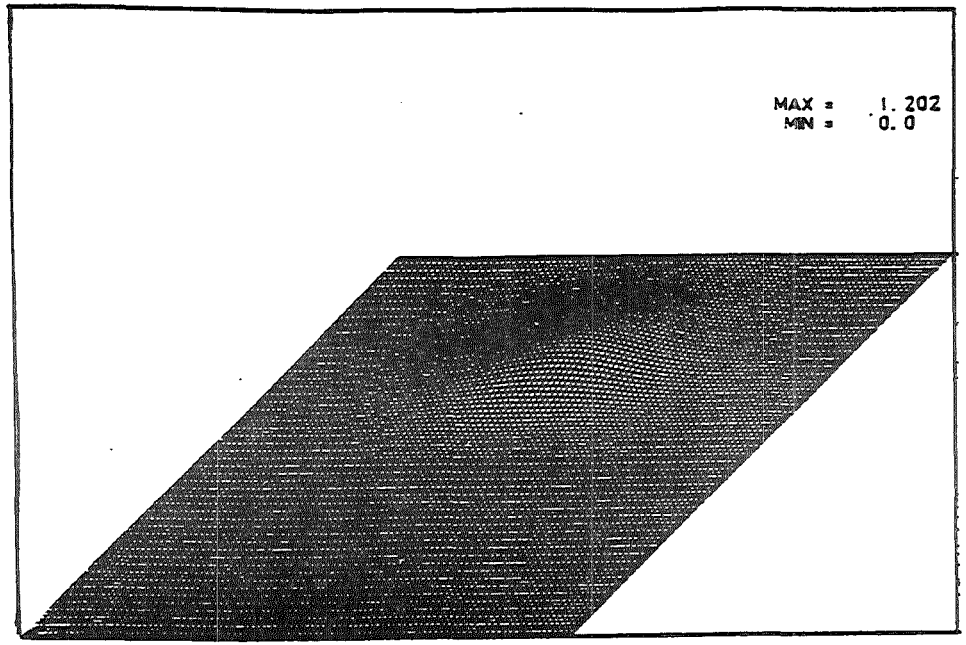
Bild 2

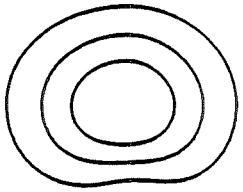




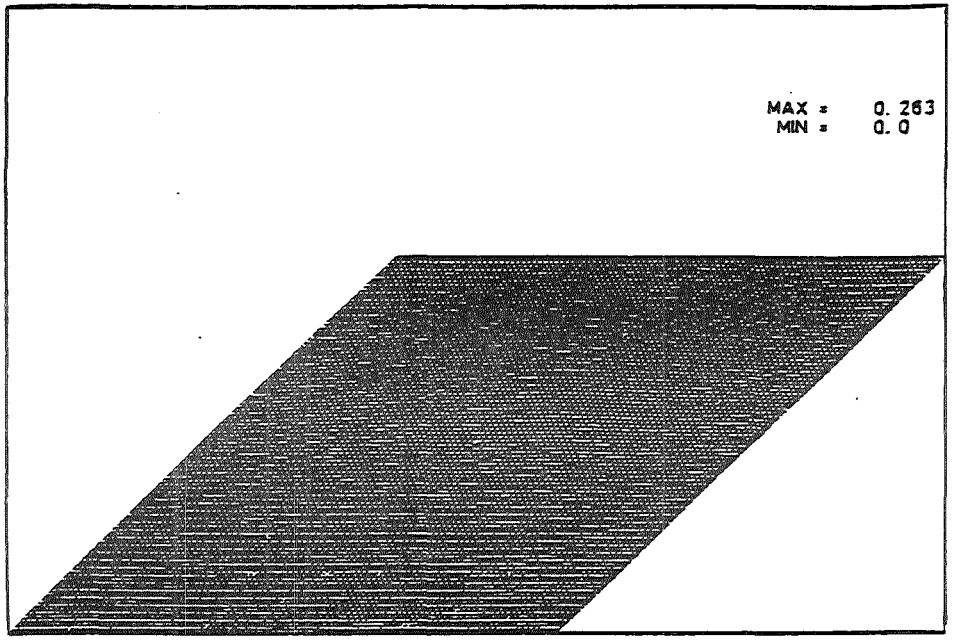
Anfangswert · Kegel

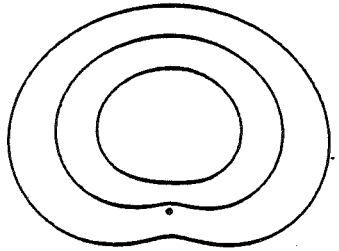
Bild 3





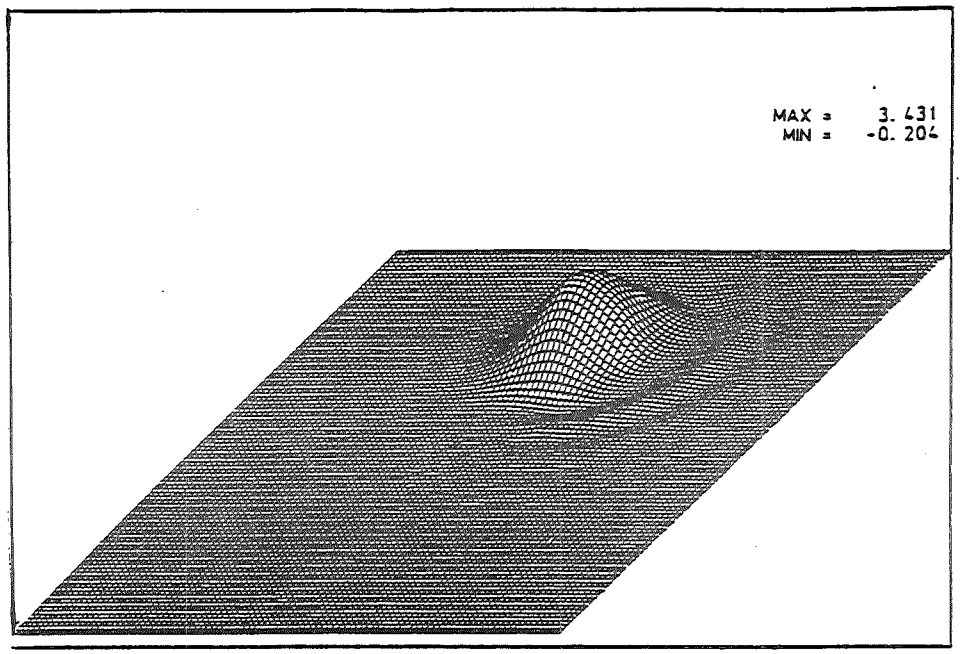
eine Umdrehung mit dem Upwind erster Ordnung Verfahren

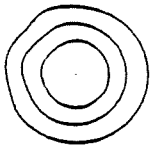


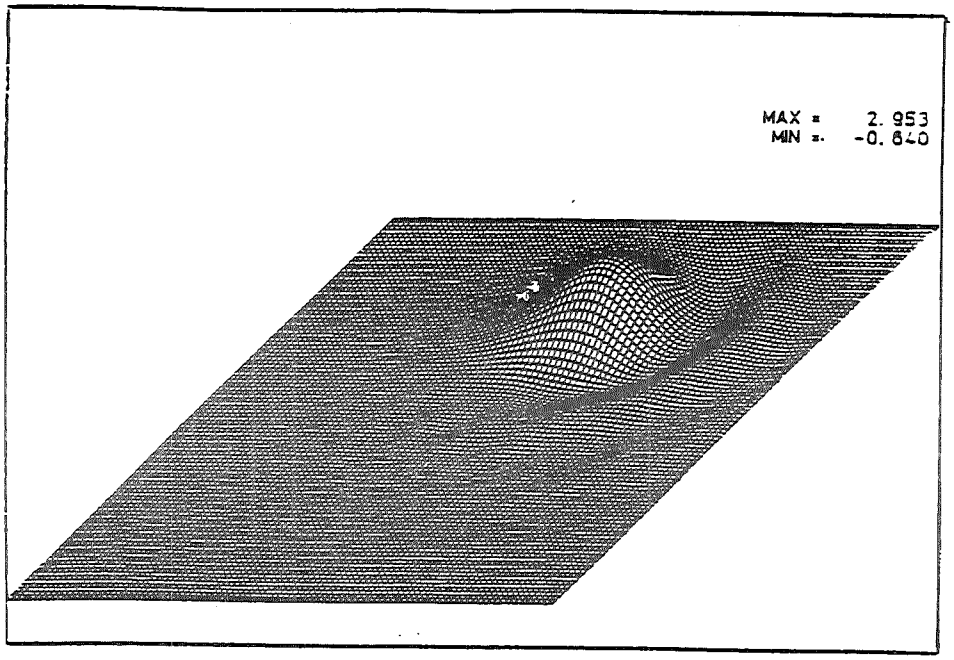


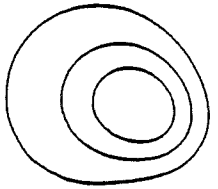
sechs Umdrehungen mit dem Upwind erster Ordnung Verfahren

Bild 4





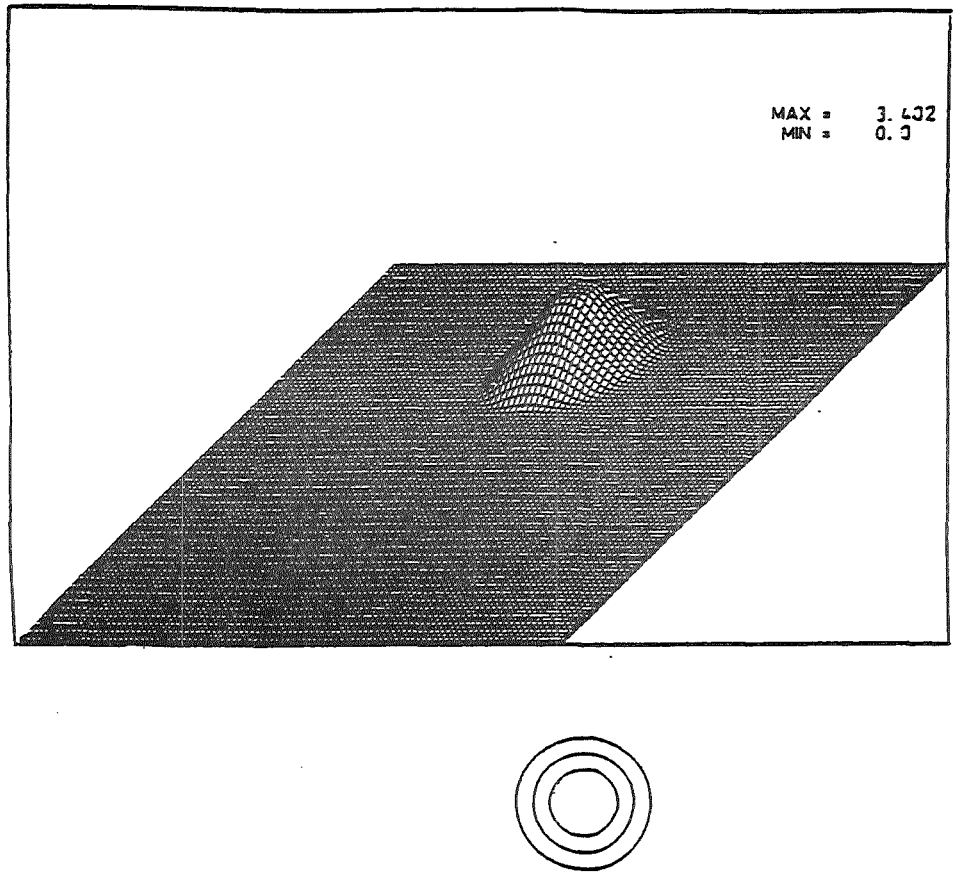




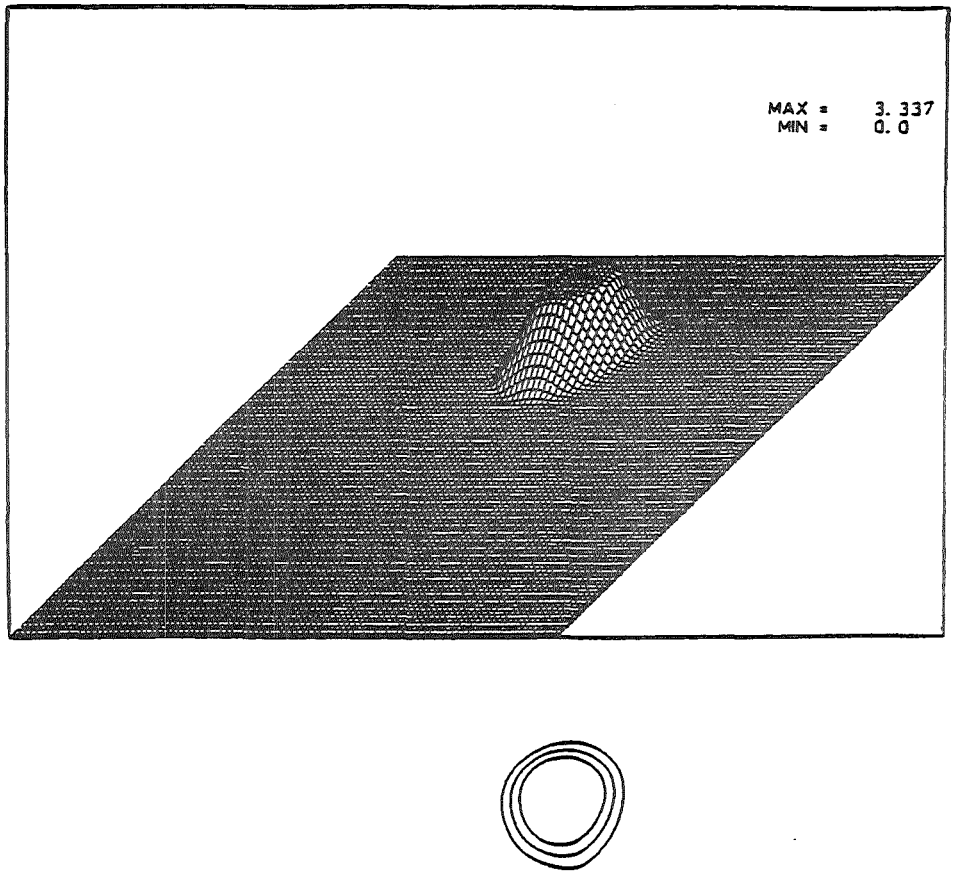
eine Umdrehung mit dem Verfahren von Lax-Wendroff

Bild 5

sechs Umdrehungen mit dem Verfahren von Lax-Wendroff



eine Umdrehung mit dem Flußkorrekturverfahren , Flußlimiter Φ_2



sechs Umdrehungen mit dem Flußkorrekturverfahren , Flußlimiter Φ_2

Bild 6