FORSCHUNG - AUSBILDUNG - WEITERBILDUNG

Bericht Nr. 51

PUNKTE, DIE EINEN MEHRDIMENSIONALEN WÜRFEL GLEICHMÄSSIG AUSFÜLLEN

I.M. Sobol *)

*) Interner Bericht
aus dem Russischen übersetzt von
Frau Kerstin Rjasanowa
März 1991

UNIVERSITÄT KAISERSLAUTERN
Fachbereich Mathematik
Arbeitsgruppe Technomathematik
Postfach 3049

6750 Kaiserslautern

Sobol, I.M.

PUNKTE, DIE EINEN MEHRDIMENSIONALEN WÜRFEL GLEICHMÄSSIG AUSFÜLLEN

Moskau, Wissen 1985
(Neues im Leben der Wissenschaft und Technik.
Serie "Mathematik, Kybernetik" Nr. 2)

Einführung

Wenn man jemanden bittet, auf einer Strecke vier Punkte gleichmäßig zu verteilen, so ruft das keine Schwierigkeiten hervor, obwohl vielleicht der eine die Verteilung wie in Abb. 1a und der andere die wie in Abb. 1b bevorzugt. Stellt man nun die schwierigere Frage nach der gleichmäßigen Verteilung von vier Punkten im Quadrat, so wird die Mehrheit sicherlich die Verteilung wie in Abbildung 2a bevorzugen, aber in Anbetracht der Konfigurationen 2b und 2c zu zweifeln beginnen und sich fragen: Was heißt denn "gleichmäßig"?

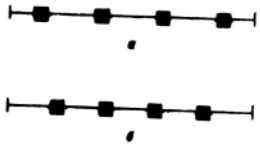

Abb. 1

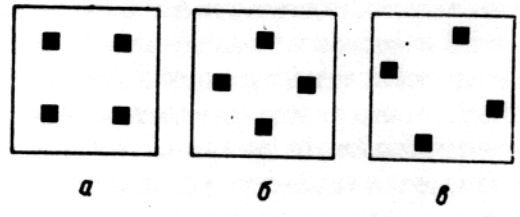

Abb. 2

In Wirklichkeit ist bereits im eindimensionalen Falle nicht ganz klar, was "gleichmäßig" bedeutet; das kann man daran erkennen, daß eben verschiedene Leute die unterschiedlichen Verteilungen a bzw. b aus Abb. 1 bevorzugen. Die Frage wird noch komplizierter, wenn man die Punkte in einem Würfel oder einem n-dimensionalen Würfel (Hyperwürfel) zu verteilen hat, wobei die Anzahl der

Punkte $N \neq 2^n$ ist und man N zu vergrößern hat ohne Beeinträchtigung der Gleichmäßigkeit... .

Solche Fragen hat die Theorie der gleichmäßig verteilten Folgen zum Gegenstand. Sie erforscht unendliche Folgen von Punkten $P_0, P_1, \ldots, P_i, \ldots$, die die Eigenschaft besitzen, daß eine Gruppe von Punkten $P_0, P_1, \ldots, P_{N-1}$ für jedes N in irgendeinem Sinne gleichmäßig im Würfel verteilt ist. Bei der Vergrößerung von N wächst die "Dichte der Verteilung", und ihre Gleichmäßigkeit wird beibehalten.

Die Theorie der Gleichverteilung wurde von H. Weyl im Jahre 1916 begründet. Sie erschien als Berührungspunkt einiger mathematischer Disziplinen (reelle und komplexe Analysis, Zahlentheorie, Wahrscheinlichkeitstheorie u.a.), und lange Zeit begrenzte sich ihre Anwendung auf verschiedene Fragen der "reinen" Mathematik und Mechanik. Die numerische Mathematik begann sich in den 50er Jahren für die gleichverteilten Folgen zu interessieren, nämlich nach der Entstehung der Monte-Carlo-Methode, als sich erwies, daß die Punkte solcher Folgen in einigen Fällen die Rolle von quasizufälligen Punkten spielen können. Die klassische Richtung der Theorie der Gleichverteilung ist verbunden mit dem bekannten Kriterium von Weyl (siehe unten), den Abschätzungen von trigonometrischen Summen und der Betrachtung von gebrochenen Anteilen verschiedener Funktionen. Sie ist recht vollständig in den Monographien von Kuipers und Niederreiter [2] dargelegt. Allerdings wurden die am besten gleichverteilten Folgen nicht auf diesem Wege konstruiert. Die vorliegende Broschüre ist einer nicht klassischen Richtung gewidmet, die es gestattet, solche Folgen zu konstruieren. Viel Aufmerksamkeit wurde den Anwendungen der gleichverteilten Folgen in der numerischen Mathematik gewidmet (Kapitel 3). Leider ist es aufgrund der Beschränkung der Seitenanzahl nicht möglich gewesen, in den Text eine Reihe schöner Resultate der klassischen Richtung sowie Fragen der unmittelbaren technischen Anwendung aufzunehmen. (Z.B. ist bekannt, daß, wenn man ein Fernsehraster mit einer Abbildung nicht zeilenweise, sondern quasizufällig ausfüllt (Abb. 3, Raster 16×16), man große Objekte und ihre Bewegung bereits ausmachen kann, ohne daß man abwarten muß, bis die Abbildung alle Zellen ausgefüllt hat unabhängig davon, wo sich die Abbildung befindet.)

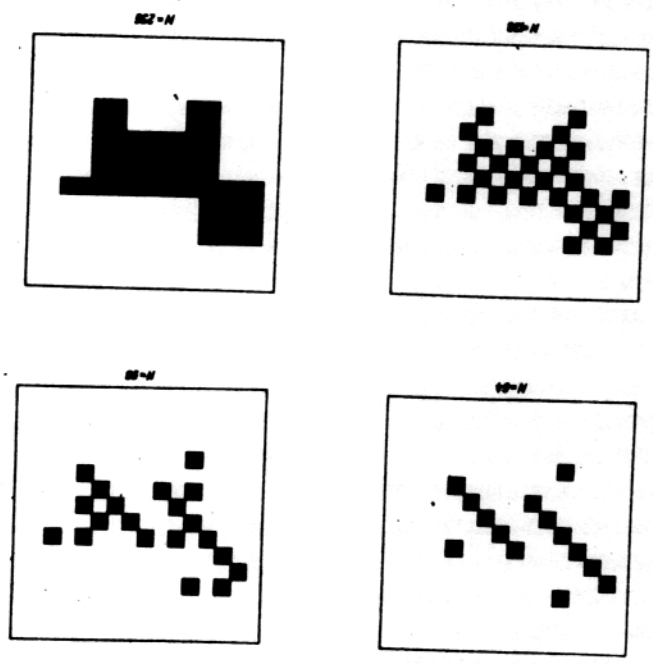

Abb. 3

Kapitel 1. Eindimensionale Aufgaben

§ 1 Gleichverteilte Folgen von Punkten

Binäre Strecken. Wir bezeichnen mit dem Buchstaben ℓ eine beliebige Strecke, die dem Intervall $[0,1]$ angehört, z.B. $\ell=(a,b)$, wobei $0\leq a<b\leq 1$ gilt. Der Eindeutigkeit wegen sei im weiteren festgelegt, daß eine Strecke links abgeschlossen und rechts offen sei mit der einzigen Ausnahme, wenn das rechte Ende $b=1$ ist; in diesem Falle sei ℓ auch rechts abgeschlossen. Die Länge von ℓ sei mit $|\ell|=b-a$ bezeichnet.
Strecken, die man durch Teilung des Intervalls $[0,1]$ in 2^m gleiche Teile ($m=0,1,2,\ldots$) erhalten kann, nennen wir <u>binäre Strecken</u>. Für solche Strecken wird im folgenden die Bezeichnung

$$\ell_{mj} = [(j-1)2^{-m}, j2^{-m}) \ , \quad 1\leq j\leq 2^m \ ,$$

verwendet. Im Falle $j=2^m$ ist gemäß unserer Vereinbarung ℓ_{mj} beidseitig abgeschlossen. Offensichtlich gilt

$$\ell_{m1} + \ell_{m2} + \ldots + \ell_{m,2^m} = [0,1] \ .$$

Die linke bzw. rechte Hälfte von ℓ_{mj} bezeichnen wir mit ℓ^{+}_{mj}; das sind ebenfalls binäre Strecken, wobei $|l^{-}_{mj}| = |\ell^{+}_{mj}| = |\ell_{mj}|/2$ gilt.
Neben der doppelten Numerierung werden wir außerdem die einfache Numerierung verwenden, indem $\ell_{mj}=\ell_k$ mit $k=2^m+j-1$ gesetzt wird.

Gleichverteilte Folgen. Wir betrachten eine beliebige Folge von Punkten $x_0,x_1,\ldots,x_i,\ldots$, die dem Intervall $[0,1]$ angehören. Sei ℓ eine beliebige Strecke, $\ell \subset [0,1]$. Wir wählen den Abschnitt $x_0,x_1,\ldots,x_{N-1}$ der Folge aus und bezeichnen mit $S_N(\ell)$ die Anzahl der Punkte dieses Abschnittes, die ℓ angehören.

Die Folge $x_0,x_1,\ldots,x_i,\ldots$ heißt <u>gleichverteilt</u> auf dem Intervall $[0,1]$, wenn für beliebiges ℓ gilt:

$$\lim_{N\to\infty} S_N(\ell)/N = |\ell| \ . \qquad (1)$$

Der Kürze halber werden wir für "Die Folge $x_0,x_1,\ldots,x_i,\ldots$ ist gleichverteilt auf dem Intervall $[0,1]$" schreiben: "$\{x_i\}$ ist g.v.".

Der geometrische Sinn dieser Definition ist hinreichend klar: Wenn $\{x_i\}$ g.v. ist, so ist für große N die Anzahl der Punkte $S_N(\ell) \sim N|\ell|$, d.h. $S_N(\ell)$ ist der Länge $|\ell|$ proportional. Die Eigenschaften der Gleichverteilung sind asymptotisch. Man kann in $\{x_i\}$ eine beliebige endliche Anzahl von Punkten austauschen, hinzufügen oder hinwegnehmen, und an der Gleichverteilung verändert sich dabei nichts. So kann sich, wenn für $i > N_0$ alle Punkte x_i erhalten bleiben, für alle hinreichend großen N der neue Wert $S'_N(\ell)$ vom alten $S_N(\ell)$ nicht um mehr als N_0 unterscheiden. Auf den Grenzwert in (1) hat das keinen Einfluß, da $N_0/N \to 0$ ist.

Lemma. Dafür, daß $\{x_i\}$ g.v. ist, ist notwendig und hinreichend, daß (1) für alle binären Strecken erfüllt ist. Wir bemerken weiterhin, daß man in der Definition (1) die halboffenen Stecken ℓ durch offene oder geschlossene ersetzen kann; daher kann man sich auf Strecken ℓ der Gestalt $\ell = [0,x)$ beschränken.

Weylsches Theorem. Die Verbindung zwischen der Definition (1) mit Aufgaben der Funktionentheorie und der numerischen Mathematik ist in gewissem Maße aus dem folgenden Theorem ersichtlich.

Weylsches Theorem. Dafür, daß $\{x_i\}$ g.v. ist, ist notwendig und hinreichend, daß für eine beliebige im Riemannschen Sinne integrierbare Funktion f(x) folgende Gleichung erfüllt ist:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} f(x_i) = \int_0^1 f(x)dx \qquad (2)$$

Das Riemannsche Integral ist das gewöhnliche bestimmte Integral, das in der mathematischen Analysis gelehrt wird. Wir erinnern daran, daß die Forderung der Existenz des Riemannschen Integrals die Forderung nach der Beschränktheit der Funktion f(x) einschließt (für unbeschränkte Funktionen werden uneigentliche Integrale eingeführt). Interessant ist, daß für Funktionen, die nach Lebesgue integrierbar sind, die Gleichung (2) sogar dann nicht erfüllt sein kann, wenn diese beschränkt sind.
Auf den ersten Blick scheinen die Forderungen (1) und (2) völlig voneinander verschieden. Um die Verbindung zwischen ihnen zu zeigen, wählen wir eine beliebige Strecke ℓ und betrachten die Funktion $f_\ell(x)$, die wir gewöhnlich Indikator dieser Strecke nennen:

$f_\ell(x) = 1$, wenn $x \in \ell$; $\quad f_\ell(x) = 0$, wenn $x \notin \ell$.

Da gilt

$$\sum_{i=0}^{N-1} f_\ell(x_i) = S_N(\ell), \qquad \int_0^1 f_\ell(x)dx = |\ell| ,$$

so fallen für $f=f_\ell(x)$ die Gleichungen (1) und (2) zusammen. In der Formulierung des Weylschen Theorems heben wir die direkte und die umgekehrte Behauptung hervor. Die direkte Behauptung lautet: Wenn $\{x_i\}$ g.v. ist, so ist für eine beliebige im Riemannschen Sinne integrierbare Funktion f(x) (2) erfüllt. Die umgekehrte Behauptung lautet: Wenn (2) für eine beliebige im Riemannschen Sinne integrierbare Funktion f(x) erfüllt ist, so ist $\{x_i\}$ g.v.

Es stellt sich heraus, daß man die umgekehrte Behauptung wesentlich verschärfen kann. Zum Beispiel ist, wenn (2) für eine beliebige stetige Funktion f(x) erfüllt ist, $\{x_i\}$ g.v. Noch schärfer ist das bekannte Weylsche Kriterium, das eine große Rolle bei der Entwicklung der "klassischen" Richtung der Theorie der gleichmäßigen Verteilung spielte (aber von uns nicht verwendet wird): Wenn (2) für alle trigonometrischen Funktionen $f=\cos 2\pi kx$ und $f=\sin 2\pi kx$ mit ganzem k erfüllt ist, so ist $\{x_i\}$ g.v.

Wir setzen nun voraus, daß f(x) nicht der Forderung des Weylschen Theorems genügt: Sie sei nicht im Riemannschen Sinne integrierbar. Dann kann man eine solche g.v. $\{x_i\}$ finden, für die die Gleichung (2) nicht erfüllt ist (für gegebenes f(x) und $\{x_i\}$). Dieses Resultat, das N.G. de Brujin, K.A. Post (1968) und C. Binder (1970) erhielten, gestattet, noch eine "umgekehrte" Behauptung zu formulieren: Wenn für eine gegebene endliche Funktion f(x) und eine beliebige g.v. $\{x_i\}$ (2) erfüllt ist, so ist f(x) im Riemannschen Sinne integrierbar.

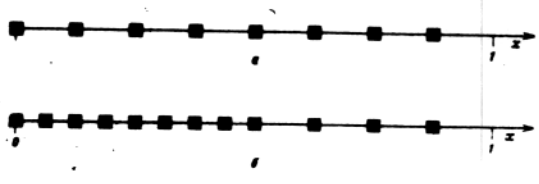

Abb. 4

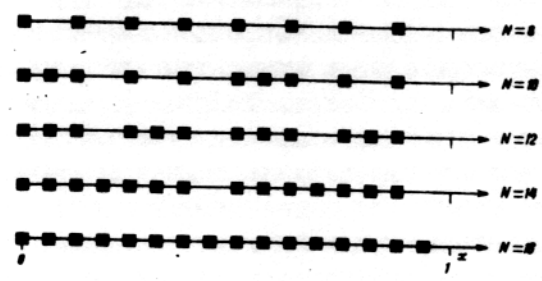

Abb. 5

B e i s p i e l 1. Die Folge der binär-rationalen Brüche in natürlicher Reihenfolge: 0, 1/2, 1/4, 3/4, 1/8, 3/8, 5/8, 7/8, 1/16,... .

Unabhängig davon, daß die Anfangsabschnitte dieser Folge, die $N=2^k$ Punkte enthalten, sehr gleichmäßig liegen (Abb. 4a), ist diese Folge nicht g.v. Um das zu beweisen, betrachten wir Anfangsabschnitte, die $N=N_k=2^k+2^{k-1}$ Punkte beinhalten (Abb. 4b). Es sei $\ell=[0,1/2)$. Offenbar ist für $N=N_k$ die Anzahl der Punkte $S_N(\ell)=2\cdot 2^{k-1}=(2/3)N$. Folglich gilt

$$\lim_{k\to\infty} S_{N_k}(\ell)/N_k = \frac{2}{3} \pm |\ell| ,$$

und die Gleichung (1) ist für $\ell=[0,\frac{1}{2}]$ nicht erfüllt. Wenn man aber Abschnitte betrachtet, deren Länge $N=N_k=2^k$ beträgt, so gilt $\lim_{k\to\infty} S_{N_k}(\ell)/N_k = \frac{1}{2} = |\ell|$.

B e i s p i e l 2. Die Folge von van der Corput (1985).

Aus obigen binär-rationalen Brüchen konstruieren wir eine g.v. $\{x_i\}$, wobei $x_i=p(i)$ gilt. Die Zahlen p(i) kann man durch folgende Formeln definieren: Wenn im Dualsystem $i=e_m e_{m-1}\ldots e_2 e_1$ ist, so ist ebenfalls im Dualsystem $p(i)=0,e_1e_2\ldots e_{m-1}e_m$.

Hierbei sind alle e_i binäre Ziffern, d.h. entweder 0 oder 1. Im Dezimalsystem heißen diese Formeln

$$i = e_1+2^1e_2 + \ldots + 2^{m-1}e_m ,$$

$$p(i) = e_1 2^{-1} + e_2 2^{-2} + \ldots + e_m 2^{-m} .$$

Einige Zahlenwerte für p(i) sind in der folgenden Tabelle angegeben. Für diese Folge sind die Anfangsabschnitte für $N=2^k$ dieselben wie im vorherigen Beispiel. Allerdings ist die Reihenfolge der Verteilung jeder folgenden Gruppe von 2^k Punkten viel komplizierter (Abb. 5). Der Beweis dafür, daß $\{p(i)\}$ g.v. ist, wird in § 3 geführt.

i	Дроби		p(i)	i	Дроби		p(i)
	i	p(i)			i	p(i)	
0	0	0	0	8	1000	0,0001	1/16
1	1	0,1	1/2	9	1001	0,1001	9/16
2	10	0,01	1/4	10	1010	0,0101	5/16
3	11	0,11	3/4	11	1011	0,1101	13/16
4	100	0,001	1/8	12	1100	0,0011	3/16
5	101	0,101	5/8	13	1101	0,1011	11/16
6	110	0,011	3/8	14	1110	0,0111	7/16
7	111	0,111	7/8	15	1111	0,1111	15/16

§ 2 Quantitative Charakteristiken der Gleichmäßigkeit

Die Gleichungen (1) und (2) und andere Eigenschaften, an die in §1 erinnert wurde, gestatten es, die Gleichverteilung einer gegebenen Folge festzustellen, geben aber keine Antwort auf die Frage, welche von zwei g.v. Folgen "gleichmäßiger" verteilt ist. Hier betrachten wir Kriterien, die nicht nur die Gleichverteilung feststellen, sondern auch eine quantitative Einschätzung der Gleichmäßigkeit geben.

Diskrepanz. Das ist die meistverbreitete Charakteristik der Gleichmäßigkeit, deren Untersuchung in den 30er Jahren begann. Wir fixieren die Punkte $x_0,x_1,\ldots,x_{N-1}$ aus dem Intervall $[0,1]$, die wir der Kürze halber Gitter nennen. Mit $S_N(x)$ bezeichnen wir die Anzahl der Punkte, die der Strecke $[0,x)$ angehören. Anders ausgedrückt ist $S_N(x)=S_N(\ell)$, wobei $\ell=[0,x)$ gilt. Diskrepanz des Gitters $x_0,x_1,\ldots,x_{N-1}$ heißt die Zahl

$$D(x_0,\ldots,x_{N-1}) = \sup_{0\le x\le 1} |S_N(x)-Nx| \qquad (3)$$

Mitunter wird der Kürze halber anstelle von $D(x_0,x_1,\ldots,x_{N-1})$ einfach D verwendet. Notwendig ist die Bemerkung, daß in der Literatur verschiedene nicht identische Definitionen der Diskrepanz anzutreffen sind. Oft wird als Diskrepanz das Verhältnis D/N bezeichnet, manchmal das Supremum von $|S_N(\ell)-N|\ell||$ über alle $\ell\le[0,1]$.

Der geometrische Sinn der Definition (3) ist offensichtlich: Nx ist die Anzahl der Gitterpunkte, die auf die Strecke $[0,x)$ bei idealer (proportionaler) Verteilung entfallen, und $S_N(x)$ ist die Anzahl der Punkte, die faktisch auf $[0,x)$ entfallen. So schätzt D in gewissem Sinne die maximale Abweichung der faktischen Punkteverteilung von der idealen gleichmäßigen Verteilung ab. Da $S_N(x)\le N$ und $Nx\le N$ gilt, ist immer $D\le N$.

Theorem 1. Dafür, daß $\{x_i\}$ g.v. ist, ist notwendig und hinreichend, daß für $N\to\infty$ gilt

$$D(x_0,\ldots,x_{N-1})/N \to 0 \ . \qquad (4)$$

Berechnung der Diskrepanz. Es ist leicht zu zeigen, daß die obere Grenze in der Formulierung (3) stets für einen Gitterpunkt realisiert wird, nämlich für $x=x_i-0$ oder für $x=x_i+0$ (Abb. 6). Folglich kann man D nach der Formel

$$D = \max_{0\leq i\leq N-1} \left(|S_N(x_i+0)-Nx_i|, |S_N(x_i-0)-Nx_i| \right) \qquad (5)$$

berechnen, und anstelle des Auffindens der oberen Grenze in (3) ist es hinreichend, die größte unter den 2N Zahlen (5) auszuwählen.

Der Algorithmus der Berechnung der Abweichung kann weiter vereinfacht werden. Wir betrachten die Formel (5) unter der Annahme, daß die Gitterpunkte schon geordnet sind: $x_{(1)}\leq x_{(2)}\leq \ldots \leq x_{(N)}$, und (der Kürze wegen) beschränken wir uns auf den Fall, daß alle diese Punkte voneinander verschieden sind. Da $S_N(x_{(i)}-0) = i-1$ und $S_N(x_{(i)}+0) = i$ gilt, kann man anstelle von (5) schreiben

$$D = \max_{1\leq i\leq N} \left(|i-Nx_{(i)}|, |i-1-Nx_{(i)}| \right) .$$

Hierbei ist $i-Nx_{(i)} = (i-1/2-Nx_{(i)}) + 1/2$ und $i-1-N_{(i)} = (i-1/2-Nx_{(i)}) - 1/2$. Daher ist leicht zu überprüfen (Abb. 7), daß

$$\max(|i-Nx_{(i)}|, |i-1-Nx_{(i)}|) = |i-1/2-Nx_{(i)}| + 1/2 .$$

Im Resultat erhalten wir die Formel von H. Niederreiter (1972):

$$D = \frac{1}{2} + \max_{1\leq i\leq N} |i-\frac{1}{2}-Nx_{(i)}| . \qquad (6)$$

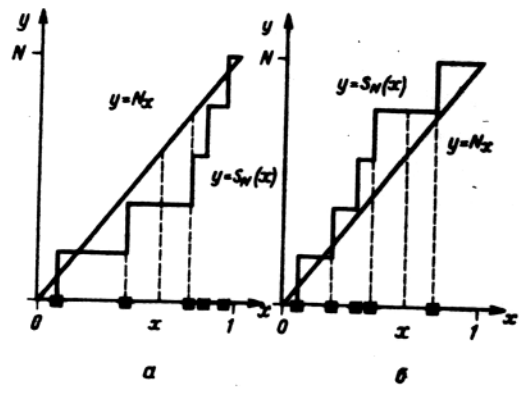

Abb. 6

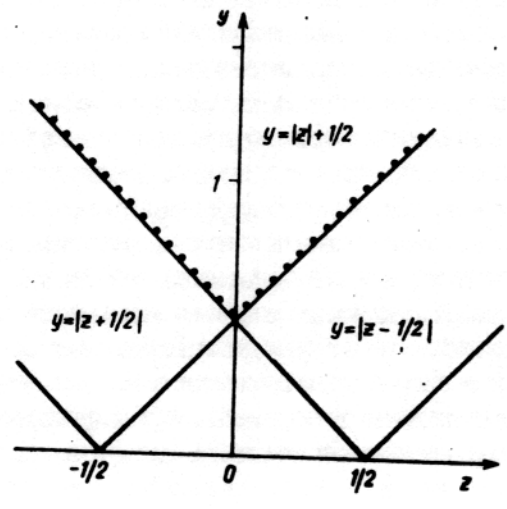

Abb. 7

Für die Berechnung nach Formel (6) ist es hinreichend, die größte unter N Zahlen zu finden. Sie ist auch in dem Falle gültig, wenn einige der Punkte x_i zusammenfallen. Wir erinnern daran, daß in (6) im Unterschied zu (5) geordnete Werte x_i vorausgesetzt wurden.

Optimierung der Diskrepanz. Aus der Formel (6) folgt sofort, daß das Minimum von D gleich 1/2 ist und nur dann angenommen wird, wenn das Gitter aus den Punkten $x_{(i)} = (i-1/2)N$ besteht, wobei $i=1,2,\ldots,N$ ist; ein solches Gitter ist in Abb. 1a zu sehen. Allerdings ist es unmöglich, eine unendliche Folge $\{x_i\}$ so zu wählen, daß beliebige Anfangsabschnitte $x_0,x_1,\ldots,x_{N-1}$ optimal sind: Beim Übergang von N Punkten zum (n+1)ten müssen alle Punkte verändert werden. Nicht genug, van der Corput fand heraus, daß es nicht gelingt, eine Folge $\{x_i\}$ so zu konstruieren, daß für alle N die Diskrepanzen $D(x_0,\ldots,x_{N-1})$ beschränkt sind. Seine Hypothese wurde von T. van Aardenne-Ehrenfest (1945) streng nachgewiesen: In einer beliebigen Folge $\{x_i\}$ ist der obere Grenzwert der Diskrepanz

$$\lim_{N\to\infty} \sup D(x_0,\ldots,x_{N-1}) = \infty .$$

Das letzte Resultat bedeutet, daß in einer beliebigen Folge $\{x_i\}$ beliebig lange "schlechte" Abschnitte vorkommen können, deren Diskrepanzen unbeschränkt wachsen. Weiter bewies K.F. Roth (1954), daß für solche "schlechten" Abschnitte $D \geq c_1\sqrt{\ln N}$ gilt, wobei c_1 eine absolute Konstante ist, die nicht von den Parametern der betrachteten Folge abhängt. Schließlich verbesserte W.M. Schmidt (1972) die Aussage für "schlechte" Abschnitte auf $D \geq c_2 \ln N$, wobei c_2 eine andere absolute Konstante ist. Da für die Folgen $\{p_i\}$ aus Beispiel 2 die Abschätzung

$$D(x_0,\ldots,x_{N-1}) \leq (1/3)\log_2 N + O(1)$$

gültig ist, die von S. Haber (1966) erhalten wurde (hierbei ist bewiesen, daß der Wert der Konstanten nicht kleiner als 1/3 sein kann), so kann die Ordnung von ln N in der Abschätzung von Schmidt nicht verbessert werden. Die Folge $\{p_i\}$ ist aber der Wachstumsordnung der Diskrepanzen nach optimal.

Neben der Diskrepanz (3) werden ebenfalls verschiedene mittlere Diskrepanzen

$$D_{(p)} = \left\{\int_0^1 |S_N(x)-Nx|^p dx\right\}^{1/p}$$

untersucht, wobei $1<p<\infty$ ist. Wir werden diese nicht betrachten. Wir bemerken nur, daß das Vierpunktegitter aus Abb. 2c aus der Bedingung $D_{(2)} = \min$ auf einer gewissen Klasse zweidimensionaler Gitter erhalten wurde (I.W. Wilenkin, 1973).

Ungleichmäßigkeit (nonuniformity). Diese "nichtklassische" Charakteristik der Gleichmäßigkeit wurde im Jahre 1957 in Verbindung mit der Verwendung der für den gegebenen Bereich neuen Methode der Fourier-Haare-Reihen [1] entwickelt. Die Struktur dieser Charakteristik und ihre Eigenschaften zeigten Wege der Konstruktion neuer Klassen gleichverteilter Folgen auf, die die besten Charakteristiken der Gleichmäßigkeit aufweisen.
Wir betrachten wieder ein Gitter, das aus N Punkten $x_0,x_1,\ldots,x_{N-1}$ aus dem Intervall [0,1] besteht. Wir wählen eine beliebige binäre Strecke ℓ_k. Bei "idealer" Gleichverteilung der Punkte des Netzes müssen auf die linke und rechte Hälfte von ℓ_k jeweils die gleiche Anzahl dieser Punkte entfallen. Daher charakterisiert die Größe $|S_N(\ell_k^-)-S_N(\ell_k^+)|$ in gewissem Maße die Ungleichmäßigkeit der Verteilung der Punkte des Netzes auf ℓ_k.
<u>Ungleichmäßigkeit</u> des Gitters $x_0,x_1,\ldots,x_{N-1}$ heißt die ganze Zahl

$$\Phi_\infty(x_0,\ldots,x_{N-1}) = \sup_k |S_N(\ell_k^-)-S_N(\ell_k^+)| , \qquad (7)$$

wobei die obere Grenze über alle binären Strecken genommen wird.

Es ist leicht nachzuweisen, daß für ein beliebiges Gitter $x_0,\ldots,x_{N-1}$ gilt

$$1 \leq \Phi_\infty(x_0,\ldots,x_{N-1}) \leq N .$$

Die rechte Ungleichung folgt aus der Tatsache, daß jede der Größen $S_N(\ell_k^-)$ und $S_N(\ell_k^+)$ jeweils N nicht übersteigt. Für den Beweis der linken Ungleichung nehmen wir zuerst an, daß alle Punkte des Gitters voneinander verschieden sind. Dann kann man ein solch großes M finden, daß jede der Strecken ℓ_{mj}, $1\leq j\leq 2^m$ entweder leer ist oder einen Punkt enthält. Im letzten Falle ist offensichtlich $|S_N(\ell_{mj}^-)-S_N(\ell_{mj}^+)| = 1$.

Setzt man nicht voraus, daß alle Punkte des Gitters voneinander verschieden sind, so erhält man, wenn man eine solch feine Unterteilung wählt und alle geometrisch voneinander verschiedenen Punkte isoliert, die Ungleichung $|S_N(\ell_k^-)-S_N(\ell_k^+)| = s$, wobei s die maximale Anzahl von zusammenfallenden Punkten des Gitters ist. Diese Überlegungen zeigen, daß die obere Grenze in der Formel (7) in Wirklichkeit über eine endliche Menge von Strecken ℓ_k berechnet wird; sowie die Unterteilung des Intervalls [0,1] in ℓ_{mj} so fein ist, daß alle geometrisch voneinander verschiedenen Punkte isoliert sind, kann man aufhören, da die Betrachtung noch kleinerer binärer Strecken bereits nichts mehr ändert. Folglich kann die Größe Φ_∞ für ein gegebenes Gitter unmittelbar nach Formel (7) berechnet werden.
Es ist nicht schwer, eine Abschätzung für Φ_∞ durch D zu erhalten:

$$\Phi_\infty(x_0,\ldots,x_{N-1}) \leq 4D(x_0,\ldots,x_{N-1}) \ . \qquad (8)$$

Theorem 2. Dafür, daß $\{x_i\}$ g.v. ist, ist notwendig und hinreichend, daß für $N\to\infty$ gilt

$$\Phi_\infty(x_0,\ldots,x_{N-1})/N \to 0 \ . \qquad (9)$$

Optimierung der Ungleichmäßigkeit. Die Größe Φ_∞ ist ungenauer als D: Sie kann nur ganzzahlige Werte annehmen. Diese Ungenauigkeit gestattet aber eine große Freiheit für die Optimierung. Zum Beispiel sahen wir, daß ein einziges Gitter $x_0,\ldots,x_{N-1}$ existiert, das das Minimum von D realisiert. Andererseits kann man unendlich viele Gitter $x_0,\ldots,x_{N-1}$ angeben, die das Minimum von Φ_∞ realisieren.

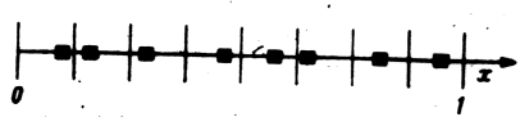

Abb. 8

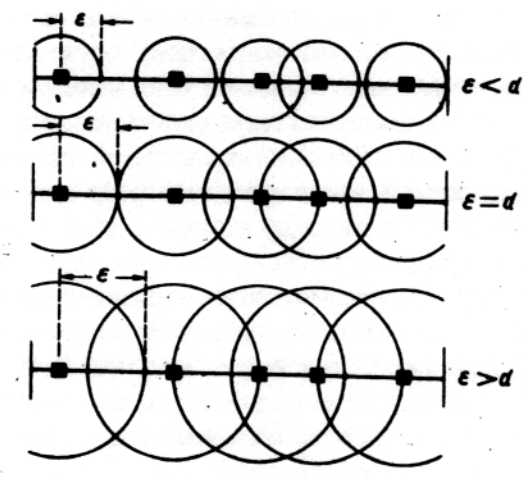

Abb. 9

B e i s p i e l . Sei $N=2^{\nu}$, wobei ν ganz ist. Wir unterteilen das Intervall [0,1] in N gleiche Strecken $\ell_{\nu j}$ und wählen in jeder einen beliebigen Punkt aus (Abb. 8). Wir zeigen, daß für dieses Gitter $\phi_{\infty}(x_0,\ldots,x_{N-1}) = 1$ gilt.
Erstens besteht für $m<\nu$ jede Strecke $\ell^{\pm}_{mj}$ aus der gleichen Anzahl von Strecken $\ell_{\nu j}$ und enthält deswegen die gleiche Anzahl von Gitterpunkten. Folglich gilt

$$S_N(\ell^{-}_{mj}) - S_N(\ell^{+}_{mj}) = 0 .$$

Zweitens enthält für $m\geq\nu$ jede Strecke ℓ_{mj} nicht mehr als einen Punkt, und alle $|S_N(\ell^{-}_{mj})-S_N(\ell^{+}_{mj})|$ sind entweder gleich 0 oder 1. In beiden Fällen erweist sich

$$|S_N(\ell^{-}_{mj}) - S_N(\ell^{+}_{mj})| \leq 1 ,$$

woraus folgt, daß $\phi_{\infty}=1$ ist.

Wir sahen, daß in einer beliebigen Folge $\{x_i\}$ "schlechte" Abschnitte existieren, auf denen die Diskrepanzen $D(x_0,\ldots,x_{N-1})$ unbeschränkt wachsen. Aus Sicht des Kriteriums ϕ_{∞} ist es viel besser: Man kann eine unendliche Menge von Folgen $\{x_i\}$ angeben, für die $\phi_{\infty}(x_0,\ldots,x_{N-1})=1$ für jedes N ist. Solchen Folgen ist §3 gewidmet. Eine davon ist die Folge $\{p(i)\}$, die in §1 betrachtet wurde.

Streuung (dispersion of points). Wir betrachten noch eine quantitative Charakteristik der Verteilung einer Gruppe von Punkten $x_0,x_1,\ldots,x_{N-1}$, die oft in Arbeiten zu numerischen Methoden anzutreffen sind.
Streuung der Punkte $x_0,x_1,\ldots,x_{N-1}$ heißt die Größe

$$d(x_0,\ldots,x_{N-1}) = \sup_{0\leq x\leq 1} \min_{0\leq i\leq N-1} |x-x_i| . \qquad (10)$$

Der geometrische Sinn der Formel (10): Wir fixieren einen beliebigen Punkt x und finden den zu ihm nächstgelegenen Punkt des Gitters (der Abstand zu ihm ist $\min_i |x-x_i|$), und danach wählt man x auf "ungünstigste" Art und Weise.
Die Größe d ist mit dem in theoretischen Untersuchungen oft verwendeten Begriff des ε-Netzes verbunden. (Eine endliche Gruppe von Punkten heißt ε-Netz, wenn der Abstand von einem beliebigen Punkt x bis zum nächstgelegenen Punkt der Gruppe ε nicht übersteigt.) In unserem Fall ist die Gruppe der Punkte $x_0,\ldots,x_{N-1}$

fixiert; bei beliebigem $\varepsilon \geq d$ ist sie ε-Netz und für beliebiges $\varepsilon < d$ nicht (Abb. 9). Damit ist d der kleinste Wert für ε, bei dem die Punkte $x_0,\ldots,x_{N-1}$ ein ε-Netz bilden.
Da N Kreise des Durchmessers d das Intervall [0,1] bedecken, ist $2dN \geq 1$, woraus eine Abschätzung für d von unten folgt:

$$(2N)^{-1} \leq d(x_0,\ldots,x_{N-1}) \ . \qquad (11)$$

Für das Gitter, das aus den Punkten $x_i=(i-1/2)/N$ mit $i=1,2,\ldots,N$ besteht, ist der Wert für d minimal und beträgt $(2N)^{-1}$.
Es ist nicht schwer, eine Abschätzung für d durch D zu finden:

$$d(x_0,\ldots,x_{N-1}) \leq 2D(x_0,\ldots,x_{N-1})/N \ . \qquad (12)$$

Theorem 3. Dafür, daß $\{x_i\}$ g.v. ist, ist notwendig, daß für $N\to\infty$ gilt

$$d(x_0,\ldots,x_{N-1}) \to 0 \ . \qquad (13)$$

Allerdings ist die Bedingung (13) nicht hinreichend.

Der Beweis der Notwendigkeit der Bedingung (13) folgt sofort aus Theorem 1 und der Ungleichung (12). Für den Beweis der zweiten Behauptung des Theorems betrachten wir die Folge aus Beispiel 1. Wenn $N=2^k$ ist (Abb. 4a), so ist offenbar $d=1/2^k$; der "schlechteste" Punkt ist in diesem Falle x=1. Solange $2^k \leq N < 2^{k+1}$ gilt, ändert sich der Wert d nicht (vgl. Abb. 4b). Erst bei $N=2^{k+1}$ stellt sich $d=1/2^{k+1}$ heraus. Aus diesem Grunde gilt für die betrachtete Folge

$$d(x_0,\ldots,x_{N-1}) = 2^{-[\log_2 N]} \ ,$$

wobei [z] den ganzen Teil der Zahl z bezeichnet. Es ist klar, daß $d(x_0,\ldots,x_{N-1}) \to 0$ für $N\to\infty$ gilt, obwohl diese Folge nicht g.v. ist. Theorem 3 ist vollständig bewiesen.
Aus dem letzten Theorem folgt, daß die Größe d als Kriterium der Gleichmäßigkeit nicht benutzt werden sollte.

§ 3 LP_0-Folgen

Definitionen. Ein Gitter, das aus $N=2^\nu$ Punkten besteht, wobei ν ganz ist, heißt P_0-Gitter, wenn jeder binären Strecke mit der Länge 1/N ein Gitterpunkt angehört.
Solche Gitter wurden bereits in §2 betrachtet (Beispiel), wo be-

wiesen wurde, daß für ein beliebiges P_o-Gitter $\Phi_\infty=1$ gilt. Mit Hilfe der Formel (6) ist leicht zu zeigen, daß für ein beliebiges P_o-Gitter $D\leq 1$ gilt.

Tatsächlich ist in Abb. 8 zu sehen, daß der i-te Punkt des P_o-Gitters $x(i)$ der Strecke $(i-1)/N\leq x<i/N$ angehört. Dabei ist die Größe $|i-1/2-Nx(i)|$, die in Formel (6) vorkommt, nicht größer als 1/2. Das heißt, es ist $D\leq 1$.

P_o-Gitter sind sehr gute Gitter (der Gleichmäßigkeit der Verteilung der Punkte nach), allerdings ist dieser Begriff so elementar, daß er keiner speziellen Bezeichnung bedürfte, wenn nicht die Verallgemeinerung auf den mehrdimensionalen Fall bevorstünde. Demgegenüber ist der Begriff der LP_o-Folge, zu deren Definition wir übergehen, sogar im eindimensionalen Falle nichttrivial.

Binärer Abschnitt der Folge $x_o,x_1,\ldots,x_i,\ldots$ heißt die Menge der Glieder x_i mit den Nummern i, die einer Ungleichung der Gestalt $k2^s\leq i\leq(k+1)2^s$ mit $k=0,1,2,\ldots$; $s=1,2,\ldots$ genügen. Zum Beispiel ist der Abschnitt $16\leq i<24$ binär ($k=2$, $s=3$), aber der Abschnitt $4\leq i\leq 16$ nicht. Ein anderes Beispiel: Wenn man die Menge der Glieder von $\{x_i\}$ in folgende Abschnitte aufteilt:

$$[x_o,\ldots,x_{h-1}],\ [x_h,\ldots,x_{2h-1}],\ [x_{2h},\ldots,x_{3h-1}],\ldots,$$

wobei $h=2^s$ ist, so erhalten wir alle binären Abchnitte der Länge 2^s. Die Folge $\{x_i\}$ heißt LP$_o$-Folge, wenn ein beliebiger binärer Abschnitt von ihr ein P_o-Netz ist.

Die Bezeichnung LP_o entstand im Resultat der Abkürzung des Satzes "Ein beliebiger binärer Abschnitt ist ein P_o-Gitter.".

Abschätzungen der Gleichmäßigkeit. Es ist nicht schwer zu zeigen, daß für einen beliebigen Anfangsabschnitt einer beliebigen LP_o-Folge gilt

$$\Phi_\infty(x_o,\ldots,x_{N-1}) = 1,\quad D(x_o,\ldots,x_{N-1}) \leq t\ , \qquad (14)$$

wobei t die Anzahl der Einsen in der binären Darstellung der Zahl N ist.

Um die Wachstumsordnung von D festzustellen, tauschen wir die Ungleichung (14) durch eine ungenauere aus. Wenn $2^{s-1}\leq N\leq 2^{s+1}-1$, so ist die Anzahl t der Einsen in der binären Darstellung von N nicht größer als s. Folglich ist $t\leq s=[\log_2(N+1)]$, wobei [z] der ganze Teil der Zahl z ist. Auf diese Weise gilt für einen be-

liebigen Anfangsabschnitt einer beliebigen LP_0-Folge

$$D(x_0,\ldots,x_{N-1}) \leq [\log_2(N+1)] \ . \qquad (15)$$

Die erhaltenen Abschätzungen zeigen, daß alle LP_0-Folgen gleichmäßig verteilt sind. In bezug auf das Kriterium D haben alle LP_0-Folgen eine optimale Wachstumsordnung D=O(ln N), und in bezug auf das Kriterium Φ_0 sind alle LP_0-Folgen optimal: $\Phi_\infty \equiv 1$.

Folgen binär rationalen Typs. Wir betrachten eine unendliche Matrix

$$(v_{sj}) = \begin{pmatrix} v_{11} & v_{12} & \cdots \\ v_{21} & v_{22} & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix} , \qquad (16)$$

deren Elemente Nullen und Einsen sind. Wir nehmen an, daß in jeder Zeile die Anzahl der Einsen endlich und ungleich Null ist. Dann entspricht jeder Zeile eine binär-rationale Zahl, deren binäre Darstellung lautet:

$$V_s = 0,v_{s1}v_{s2} \cdots v_{sj} \cdots \ . \qquad (17)$$

Die Matrix (v_{sj}) heißt <u>Richtungsmatrix</u>, und die Zahlen $V_1,V_2,\ldots,V_s,\ldots$ <u>Richtungszahlen.</u>

Eine Folge binär-rationalen Typs (abgekürzt: BR-Folge) ist eine Folge $\{r(i)\}$, die nach drei Regeln konstruiert wird:

1° $r(0) = 0$.

2° Wenn $i=2^s$, so ist $r(i) = V_{s+1}$.

3° Wenn $2^s<i<2^{s+1}$, so ist $r(i) = r(2^s)*r(i-2^s)$, wobei die Operation $*$ die stellenweise Addition nach Modul 2 im Dualsystem bedeutet.

Unter den Befehlen eines beliebigen Computers gibt es eine logische Operation, die die Operation $*$ ausführt. Sie heißt "ausschließendes ODER", obwohl die Numeriker die Bezeichnung "Vergleich der binären Darstellung" bevorzugen, weil $a*b=0$ dann und nur dann gilt, wenn alle Stellen der Darstellungen von a und b zusammenfallen, d.h. a=b ist. Wir erklären diese Operation an Zahlenbeispielen.

$$5/16 * 7/8 = 0,0101 * 0,1110 = 0,1011 = 11/16 ,$$

$$13/16 * 19/32 = 0,11010 * 0,10011 = 0,01001 = 9/32 .$$

Es ist nicht schwer zu sehen, daß die Regeln 1^o-3^o folgender Regel äquivalent sind: Wenn im Dualsystem

$$i = e_m \dots e_2 e_1$$

gilt, so ist ebenfalls im Dualsystem

$$r(i) = e_1 V_1 * e_2 V_2 * \dots * e_m V_m \ . \qquad (18)$$

Wir bemerken, daß man in der Formel (18) keine Multiplikation auszuführen hat: Wenn $e_j=1$ ist, so ist der "Summand" V_j in der Formel dabei, ist hingegen $e_j=0$, so wird er ausgelassen. So ist zum Beispiel für $i=25$ im Zweiersystem $i=11001$; folglich $r(25)=V_1*V_4*V_5$.
Im folgenden Theorem sind einfache hinreichende Bedingungen dafür formuliert, daß bei ihrer Erfüllung eine BR-Folge eine LP_o-Folge ist.

Theorem. Wenn in der Richtungsmatrix auf der Hauptdiagonalen Einsen stehen und oberhalb Nullen, so ist die ihr entsprechende BR-Folge eine LP_o-Folge.

Die Folge $\{p(i)\}$.
Die in §1 konstruierte Folge $\{p(i)\}$ ist die einfachste BR-Folge. Ihre Richtungsmatrix ist die unendliche Einheitsmatrix. Die Richtungszahlen $V_s=2^{-s}$ enthalten Einsen an verschiedenen Stellen der Darstellung im Dualsystem, was die Benutzung des Zeichens "+" anstelle von "$*$" in (18) gestattet.
Aus dem vorherigen Theorem folgt, daß $\{p(i)\}$ eine LP_o-Folge ist und für sie die Ungleichungen (14) gelten. Interessant ist, daß die grobe Abschätzung $D(x_o,\dots,x_{N-1})\leq t$ für einige Werte für N besser ist als die in §2 angeführte Abschätzung

$$D(x_o,\dots,x_{N-1}) \leq (1/3)\log_2 N + 0(1) \ ,$$

wo die Konstante 1/3 nicht verkleinert werden kann.
Wir bemerken noch zwei Eigenschaften von $\{p(i)\}$, die bei weitem nicht alle LP_o-Folgen vom binär-rationalen Typ haben.

Eigenschaft 1. Ein beliebiger Abschnitt der Folge $\{p(i)\}$, der 2^m Punkte enthält ($m=1,2,\dots$), ist ein P_o-Gitter.

Eigenschaft 2. Für einen beliebigen Abschnitt $i'\leq i\leq i''$ der Folge $\{p(i)\}$ ist die Ungleichmäßigkeit gleich Eins:

$$\Phi_\infty(p(i),\dots,p(i'')) = 1 \ .$$

Die Folge $\{q(i)\}$

Pascal-Matrix nennen wir die unendliche Matrix

$$C = (c_{sj}) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \dots \\ 1 & 1 & 0 & 0 & 0 & \dots \\ 1 & 2 & 1 & 0 & 0 & \dots \\ 1 & 3 & 3 & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix} . \qquad (19)$$

Die Elemente c_{sj}, die oberhalb der Hauptdiagonalen stehen, sind gleich Null. Die anderen Elemente $c_{sj} = C_{s-1}^{j-1}$ sind die bekannten Binomialkoeffizienten. So ist der Anteil der Matrix C, der keine Nullen enthält, das bekannte Pascaldreieck.

Folge $\{q(i)\}$ heißt die BR-Folge mit der Richtungsmatrix (v_{sj}), wobei $v_{sj}=c_{sj}$ (mod2) gilt. Die letzte Darstellung bedeutet, daß in der Pascal-Matrix alle ungeraden Elemente durch Einsen und alle geraden durch Nullen zu ersetzen sind:

$$(v_{sj}) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & \dots \\ 1 & 1 & 0 & 0 & 0 & 0 & \dots \\ 1 & 0 & 1 & 0 & 0 & 0 & \dots \\ 1 & 1 & 1 & 1 & 0 & 0 & \dots \\ 1 & 0 & 0 & 0 & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix} \qquad (20)$$

Da die Matrix (20) den Bedingungen des letzten Theorems genügt, so ist $\{q(i)\}$ eine LP_0-Folge, und für sie gelten die Beziehungen (14).

Lemma.
Wir bezeichnen mit ω_{sj} die Elemente der Matrix C^2, so daß $\omega_{sj} = \sum_{\alpha=1}^{\infty} c_{s\alpha}c_{\alpha j}$ gilt. Dann ist $\omega_{sj}=0$ für $j>s$; $\omega_{ss}=1$,

$\omega_{sj}=0$ (mod2) für $j<s$.
Mit Hilfe des Lemmas kann man beweisen, daß $\{p(i)\}$ und $\{q(i)\}$ symmetrisch in folgendem Sinne sind: Wenn $p(i)=q(k)$ ist, so ist $p(k)=q(i)$.

In §3 des Kapitels 2 ist bewiesen, daß $\{p(i)\}$ und $\{q(i)\}$ in einem gewissen Sinne unabhängig sind: Die Folge der Punkte mit den kartesischen Koordinaten $(p(i),q(i))$ ist eine zweidimensionale LP_0-Folge.

§ 4 Über die näherungsweise Berechnung von Integralen

Für die näherungsweise Berechnung von Integralen werden gewöhnliche Quadraturformeln der Gestalt

$$\int_0^1 f(x)dx \approx \sum_{i=0}^{N-1} C_i f(x_i) \tag{21}$$

verwendet; die Punkte x_i heißen Stützstellen und die Koeffizienten C_i Gewichte der Formel (21).
Wenn die Funktion f(x) hinreichend glatt ist (der Sinn dieser Worte wird im folgenden klar), so werden anstelle von (21) Formeln mit gleichen Gewichten verwendet:

$$\int_0^1 f(x)dx \approx \frac{1}{N} \sum_{i=0}^{N-1} f(x_i) \tag{22}$$

Das Weylsche Theorem behauptet, daß, wenn als Stützstellen in der Formel (21) die Punkte einer g.v. Folge $\{x_i\}$ gewählt werden, die Konvergenz für eine beliebige im Riemannschen Sinne integrierbare Funktion f(x) (vgl. (2)) gewährleistet werden kann. Natürlich entsteht die Frage nach der Abschätzung der Konvergenzgeschwindigkeit.
Es stellt sich heraus, daß solche Klassen von Funktionen f(x) existieren, für die die Konvergenz von $D(x_0,\ldots,x_{N-1})$ oder $\Phi_\infty(x_0,\ldots,x_{N-1})$ bestimmt wird. Wir beschränken uns auf die Betrachtung einer solchen Klasse.

Hauptlemma. Wir bezeichnen mit $\delta(f)$ den Fehler der Formel (2)

$$\delta(f) = \frac{1}{N} \sum_{i=0}^{N-1} f(x_i) - \int_0^1 f(x)dx \ . \tag{23}$$

Wir setzen voraus, daß die Funktion f(x) stetig ist und eine stückweise stetige Ableitung f'(x) besitzt.

Lemma. Für beliebige Punkte $x_0,\ldots,x_{N-1}$ gilt

$$\delta(f) = \frac{1}{N} \int_0^1 (Nx - S_N(x))f'(x)dx \ . \tag{24}$$

Den Beweis beginnen wir mit der offensichtlichen Beziehung

$$f(x) = f(1) - \int_x^1 f'(t)dt \ .$$

Mit Hilfe der Heavysidefunktion e(x) (die durch e(x)=0 für $x \leq 0$, e(x)=1 für $x > 0$ definiert ist) kann man diese Beziehung in der

Gestalt

$$f(x) = f(1) - \int_0^1 f'(t)e(t-x)dt$$

schreiben. Da der Ausdruck $\delta(f)$ linear von f abhängig ist, gilt

$$\delta(f(x)) = - \int_0^1 f'(t)\delta(e(t-x))dt \ .$$

Es ist noch die Größe $\delta(e(t-x))$ zu berechnen, die nach (23) gleich

$$\frac{1}{N} \sum_{i=0}^{N-1} e(t-x_i) - \int_0^1 e(t-x)dx = \frac{1}{N} S_N(t) - t$$

ist. Setzt man diesen Ausdruck in das vorherige Integral ein, so erhält man (24).
Aus der Formel (24) folgt die Abschätzung für den Fehler $\delta(f)$ durch die Diskrepanz

$$|\delta(f)| \leq \frac{D(x_o,\ldots,x_{N-1})}{N} \int_0^1 |f'(x)|dx \ . \qquad (25)$$

Die hier eingehende Konstante $\int_0^1 |f'(x)|dx$ ist gleich der Variation von f(x) auf dem Intervall [0,1]. Die Verallgemeinerung der Abschätzung (25) auf beliebige Funktionen f(x) mit beschränkter Variation erhielt J.F. Koksma (1942).

Abschätzungen für Klassen von Funktionen.
In der Theorie der Quadraturformeln wird der Konstruktion der besten Quadraturformel für eine gegebene Klasse von Funktionen viel Aufmerksamkeit geschenkt. Die Formel (24) gestattet, ähnliche Aufgaben zu lösen. Wir beschränken uns auf ein Beispiel.
Wir betrachten die Klasse der Funktionen $W_1^1(L)$, die alle stetigen Funktionen f(x) mit stückweise stetiger Ableitung f'(x) enthält, die der Ungleichung

$$\int_0^1 |f'(x)|dx \leq L$$

genügen. Uns interessiert die Größe $R = \sup|\delta(f)|$, wobei das Supremum über alle Funktionen f(x) aus $W_1^1(L)$ genommen wird. Man kann sagen, daß R der Fehler ist, der bei der Wahl der "schlechtesten" Funktionen der Klasse entsteht.

Mit Hilfe des Hauptlemmas läßt sich beweisen, daß bei beliebiger Wahl der Punkte $x_0, x_1, \ldots, x_{N-1}$ aus $[0,1]$ gilt:

$$R = LD(x_0, \ldots, x_{N-1})/N \ . \qquad (26)$$

Aus der Formel (26) folgt sofort, daß der minimale Wert für R, der gleich $L/(2N)$ ist, bei $x_i=(i+1/N)/N$, $0 \leq i \leq N-1$, angenommen wird. Anders ausgedrückt erweist sich als beste Quadraturformel mit gleichen Gewichten in der Klasse $W_1^1(L)$ die gut bekannte Rechtecksregel. In Wirklichkeit bleibt die Rechtecksregel auch optimal in der Klasse $W_1^1(L)$, wenn in die Betrachtungen Quadraturformeln der Gestalt (21) mit Gewichten einbezogen werden (S.M. Nikotskij, 1950).

Die Berechnung von uneigentlichen Integralen. Wir setzen voraus, daß die Funktion f(x) stetig ist und auf $0 < x \leq 1$ eine stückweise stetige Ableitung f'(x) hat; für $x \to 0$ ist diese Funktion unbeschränkt, es existiert aber das uneigentliche Integral $\int_0^1 f(x)dx$. Wie in §1 bemerkt wurde, kann man für eine beliebige solche Funktion eine g.v. Folge $\{x_i\}$ so ungünstig finden, daß die Gleichung (2) nicht erfüllt ist. Folglich muß eine Bedingung, die die Erfüllung von (2) garantiert, die Eigenschaften von $\{x_i\}$ mit denen von f(x) verbinden.

Wir betrachten eine g.v. $\{x_i\}$ mit $x_i > 0$. Wir bezeichnen $a_N = \min\limits_{0 \leq i \leq N-1} x_i$. Offensichtlich ist $a_N \to 0$, wenn $N \to \infty$.

Das folgende Resultat (I.M. Sobol), 1973) stellt das Analogon zu (25) dar.

Wenn für die betrachteten f(x) und $\{x_i\}$ für $N \to \infty$ die Bedingung

$$\frac{D(x_0, \ldots, x_{N-1})}{N} \int_{a_N}^1 |f'(x)|dx \to 0 \qquad (27)$$

efüllt ist, so ist die Beziehung (2) gültig.

Als Beispiel der Verwendung der Bedingung (27) betrachten wir die Folge p(1),p(2),... (ohne Nullpunkte), für die $1/2 < (N+1)a_N < 2$ gilt. Wir nehmen an, daß die Funktion f(x) bei Null eine einfache Besonderheit der Gestalt $f=x^{-\lambda}$ oder $f=x^{-1}(\ln x)^{-\lambda}$ hat. In diesem Fall ist die Bedingung (27) dann und nur dann erfüllt, wenn das entsprechende uneigentliche Integral konvergiert.

§ 5 Analogien aus der Wahrscheinlichkeitstheorie

Gesetz der großen Zahlen. Wir bezeichnen mit γ eine Zufallsgröße, die im Intervall $(0,1)$ gleichverteilt ist. Die Wahrscheinlichkeitsdichte einer solchen Größe ist $p(x)\equiv 1$ und die Verteilungsfunktion $F(x)=x$ mit $0\leq x\leq 1$.

Wir betrachten eine Folge von unabhängigen Werten $\gamma_1,\gamma_2,\ldots,\gamma_i,\ldots$. Bezeichnet man wie vorher die Anzahl der Werte γ_i mit den Nummern $1\leq i\leq N$, die auf das Intervall ℓ entfallen, mit $S_N(\ell)$, so ist das Verhältnis $S_N(\ell)/N$ gleich der Frequenz des Eintretens des zufälligen Ereignisses $\{\gamma \in \ell\}$ in N unabhängigen Versuchen. Da die Wahrscheinlichkeit dieses Ereignisses

$$P\{\gamma \in \ell\} = \int_{\ell} p(x)dx = |\ell|$$

ist, bedeutet die bekannte Tatsache der Konvergenz der Häufigkeiten, die gewöhnlich Bernoulli-Theorem genannt wird, daß für $N\to\infty$ gilt:

$$S_N(\ell)/N \xrightarrow{P} |\ell| \; ; \qquad (1')$$

Mit $\xrightarrow{P}$ wird die Konvergenz in der Wahrscheinlichkeit bezeichnet. Offensichtlich kann man die Formel (1) als deterministisches Analogon zur Formel (1') betrachten.

Wir betrachten nun eine absolut integrierbare Funktion $f(x)$. Die mathematische Erwartung der Zufallsgröße $f(\gamma)$ existiert und ist gleich

$$Mf(\gamma) = \int_0^1 f(x)p(x)dx = \int_0^1 f(x)dx \; .$$

Nach dem bekannten Theorem von A.J. Chintschin (1928) genügt die Folge der Werte $\{f(\gamma_i)\}$ dem Gesetz der großen Zahlen. Für $N\to\infty$ ist

$$\frac{1}{N}\sum_{i=1}^{N} f(\gamma_i) \xrightarrow{P} \int_0^1 f(x)dx \; . \qquad (2')$$

Offensichtlich stellt die Formel (2) ein deterministisches Analogon zur Formel (2') dar. Allerdings können die Konvergenzgeschwindigkeiten in (2) bzw. (2') sehr unterschiedlich sein.

Wir nehmen an, daß das Quadrat der Funktion $f(x)$ ebenfalls integrierbar ist. Dann ist die Dispersion endlich:

$$Df(\gamma) = Mf^2(\gamma) - (Mf(\gamma))^2 = \int_0^1 f^2(x)dx - (\int_0^1 fdx)^2 ,$$

und die Folge der Werte $\{f(\gamma_i)\}$ genügt dem zentralen Grenzwertsatz. Mit diesem Satz ist leicht zu zeigen, daß die Konvergenzordnung in Formel (2') mit einer beliebig großen Wahrscheinlichkeit gleich $1/\sqrt{N}$ ist.
In der deterministischen Formel (2) kann die Konvergenzordnung beliebig schlecht sein. Wenn man aber eine "gute" g.v. Folge $\{x_i\}$ wählt, z.B. irgendeine LP_σ-Folge und eine "nicht allzu schlechte" Funktion $f(x)$, z.B. $f(x) \in W_1^1(L)$, so ist, wie im vorangegangenen Paragraphen gezeigt wurde, die Konvergenzordnung in (2) nicht schlechter als $N^{-1}\ln N$ und für $N=2^m$ sogar $1/N$.

Die Statistik von Kolmogorow. Wir betrachten wieder eine Folge unabhängiger Werte $\{\gamma_i\}$, und sei $S_N(x)$ die Anzahl der $\gamma_i < x$ mit $1 \leq i \leq N$. Die Funktion

$$F_N(x) = S_N(x)/N$$

heißt <u>ausgewählte</u> oder <u>empirische Verteilungsfunktion</u> der Stichprobe $\gamma_1, \ldots, \gamma_N$. Die Behauptung über die Konvergenz von $F_N(x)$ in der Wahrscheinlichkeit gegen die Verteilungsfunktion $F(x)=x$ der Größe γ ist der Behauptung (1') äquivalent. Die Abweichung $F_N(x)$ von $F(x)$ wird oft mit der Statistik von Kolmogorow abgeschätzt:

$$\kappa_N = \sqrt{N} \sup_{0 \leq x \leq 1} |F_N(x) - F(x)| , \qquad (3')$$

wofür A.N. Kolmogorow (1933) einen Grenzwertsatz zeigte, der gewöhnlich Theorem von Kolmogorow genannt wird:

$$\lim_{N\to\infty} P(\kappa_N < x) = K(x) ,$$

wobei

$$K(x) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2x^2}$$

die <u>Kolmogorow-Funktion</u> ist.

Da im von uns betrachteten Falle $F(x)=x$ ist, kann man bei Einsetzen von (3) in (3') leicht sehen, daß gilt:

$$\kappa_N = D/\sqrt{N} .$$

Auf diese Weise erweist sich die Abweichung D als deterministisches Analogon der Statistik von Kolmogorow (mit einer Genauigkeit bis auf den Faktor $\sqrt{N}$, der weder von x noch von $F_N(x)$ abhängt).

Deterministische Interpretation der Vereinbarkeitskriterien.
Wir nehmen an, daß eine Gruppe von Zahlen $x_1,\ldots,x_N$ aus dem Intervall (0,1) gegeben ist. Kann man annehmen, daß diese Zahlen unabhängige Werte der Zufallsgröße γ sind? Für die Antwort auf eine solche Frage werden verschiedene Vereinbarkeitskriterien verwendet, darunter das Kolmogorow-Kriterium, das auf der Statistik κ_N beruht. Wir erklären das Kriterium an einem Beispiel.

Wir nehmen an, daß N hinreichend groß ist und daß $P\{\kappa_N<x\} \approx K(x)$ gilt. In der Praxis nimmt man diese Gleichheit hinreichend genau für $N\geq 40$ an. Mit den gegebenen Werten $x_1,\ldots,x_N$ berechnen wir κ_N. Wenn $\kappa_N\geq 1{,}95$ gilt, so wird die Hypothese darüber, daß die gegebenen Zahlen unabhängige Werte von γ sind, abgelehnt, da $K(1{,}95)=0{,}999$ und die Wahrscheinlichkeit $P\{\kappa_N\geq 1{,}95\} \approx 0{,}001$ ist: Das Eintreten eines so unwahrscheinlichen Ereignisses wird als ein Widerspruch eingeschätzt.

Offensichtlich kann ein beliebiges Vereinbarkeitskriterium die Hypothese ablehnen, sie aber nicht beweisen.
Wir engen nun die Fragestellung ein: Kann man die Zahlen $x_1,\ldots,x_N$ anstelle der unabhängigen Werte γ in der Formel

$$Mf(\gamma) \approx \frac{1}{N} \sum_{i=1}^{N} f(\gamma_i)$$

für die näherungsweise Berechnung der mathematischen Erwartung verwenden? Schreibt man Formel (26) um in die Gestalt

$$\sup_{f\in W_1^1(L)} \left|\frac{1}{N} \sum_{i=1}^{N} f(x_i) - Mf(\gamma)\right| = \frac{L\kappa_N}{\sqrt{N}} ,$$

so sehen wir, daß, wenn man sich auf Funktionen der Klasse $W_1^1(L)$ beschränkt, ein geringes $\varkappa_N$ einen kleinen Fehler bei solchen Rechnungen garantiert. Auf diese Weise gestattet die deterministische Interpretation, verschiedenen Vereinbarkeitskriterien verschiedene Aufgabenklassen gegenüberzustellen, für die diese Kriterien in gewissem Sinne sowohl notwendig als auch hinreichend sind. Natürlich garantiert die Vereinbarkeit mit der Hypothese nach einem der Kriterien nicht die Möglichkeit der Verwendung der überprüften Zahlen in Aufgaben anderer Klassen. So garantiert zum Beispiel ein kleines $\varkappa_N$ nicht die Möglichkeit der Berechnung zweifacher Integrale mit den Zahlen $x_1,\ldots,x_N$.

Kapitel 2. Mehrdimensionale Aufgaben

§ 1 Gleichverteilte Folgen im mehrdimensionalen Würfel

Bezeichnungen. Wir bezeichnen mit K^n den Einheitshyperwürfel im n-dimensionalen Raum: K^n besteht aus allen Punkten P mit den kartesischen Koordinaten $P=(x_1,\ldots,x_n)$, die den Ungleichungen $0 \leq x_j \leq 1$ $(j=1,2,\ldots,n)$ genügen. Der Kürze wegen schreiben wir anstelle des Wortes "Hyperwürfel" "Würfel". Ebenso werden wir im folgenden anstelle "Hyperebene" "Ebene" und anstelle "Hyperoktant" "Oktant" schreiben.

Wir benötigen n-dimensionale Parallelepipede π mit Kanten, die parallel zu den Koordinatenachsen verlaufen (Abb. 10). Die Projektion eines solchen Parallelepipedes auf die Achse $0x_j$ stellt eine Strecke dar, die wir ℓ_j nennen. Offensichtlich besteht π aus allen Punkten $P=(x_1,\ldots,x_n)$, deren Koordinaten den Bedingungen $x_j \in \ell_j$ mit $1 \leq j \leq n$ genügen. Das Volumen (n-dimensional) eines solchen Parallelepipedes ist gleich dem Produkt $V_\pi = |\ell_1| \ldots |\ell_n|$. Wir erinnern daran, daß wegen unserer Vereinbarung darüber, daß alle betrachteten Strecken links abgechlossen und rechts offen sind (mit Ausnahme des Falls, wenn das rechte Ende gleich 1 ist), bei allen π die linken und unteren Kanten (bezüglich jeder Koordinate) zu π gehören, die rechten und oberen aber nicht (mit der offensichtlichen Ausnahme, wenn irgendeine dieser Kanten mit der Begrenzung des Würfels K^n zusammenfällt).

Ein Parallelepiped π_k heißt <u>binär</u>, wenn alle ihn definierenden Strecken $\ell_{k_1},\ldots,\ell_{k_n}$ binär sind. In Abb. 11 sind alle binären Rechtecke (n=2) mit der Fläche 1/8 abgebildet. Wegen der getroffenen Vereinbarungen bildet die Summe aller π_k desselben Typs den ganzen Würfel K^n.

Wir betrachten eine Folge von Punkten $P_0, P_1, \ldots, P_i, \ldots$, die K^n

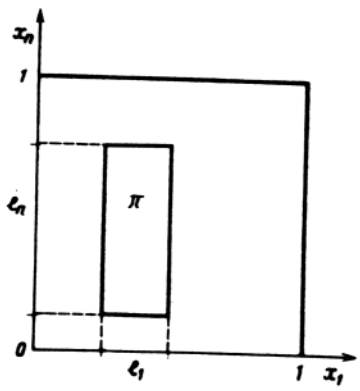

Abb. 10

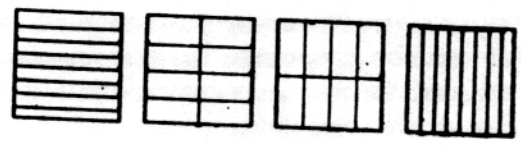

Abb. 11

angehören, und bezeichnen mit $S_N(G)$ die Anzahl der Punkte P_i mit den Nummern $0 \leq i \leq N-1$, die der Menge G angehören. Die Folge der Punkte $P_0, P_1, \ldots, P_i, \ldots$ heißt gleichverteilt in K^n, (kurz g.v.), wenn für ein beliebiges π gilt:

$$\lim_{N\to\infty} S_N(\pi)/N = V_\pi \ . \tag{28}$$

Man kann zeigen, daß, wenn G ein beliebiges Gebiet aus K^n mit dem Volumen V_G ist, aus (28) folgt:

$$\lim_{N\to\infty} S_N(G)/N = V_G \ .$$

Auf diese Weise ist die Anzahl der Punkte einer g.v. Folge, die einem beliebigen Gebiet G angehören, bei großen N proportional dem Volumen G.

Lemma. Dafür, daß $\{P_i\}$ g.v. ist, ist notwendig und hinreichend, daß (28) für alle binären π_k erfüllt ist.

Theorem von Weyl. Dafür, daß $\{P_i\}$ g.v. ist, ist notwendig und hinreichend, daß für eine beliebige im Riemannschen Sinne integrierbare Funktion f(P) die folgende Beziehung erfüllt ist:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} f(P_i) = \int_{K^n} f(P)dP \ . \tag{29}$$

Das in (29) rechts stehende Integral heißt:

$$\int_{K^n} f(P)dP = \int_0^1 \ldots \int_0^1 f(x_1, \ldots, x_n)dx_1 \ldots dx_n \ .$$

Diskrepanz. Wir betrachten in K^n ein Gitter, das aus N beliebigen Punkten $P_0, P_1, \ldots, P_{N-1}$ besteht. Jedem Punkt P aus K^n wird ein Parallelepiped π_p mit der Diagonale OP (Abb. 12) zugeordnet. Das Volumen V_p dieses Parallelepipedes ist gleich dem Produkt der Koordinaten $x_1, \ldots, x_n$ des Punktes P.
Diskrepanz des Gitters $P_0, P_1, \ldots, P_{N-1}$ heißt die Zahl

$$D(P_0, \ldots, P_{N-1}) = \sup_P |S_N(\pi_p) - NV_p| \ , \tag{30}$$

wobei die obere Grenze über alle $P \in K^n$ genommen wird. Wie auch im eindimensionalen Falle gilt $D \leq N$.

Theorem 1. Dafür, daß $\{P_i\}$ g.v. ist, ist notwendig und hinreichend, daß für $N\to\infty$ gilt

$$D(P_o,\ldots,P_{N-1})/N \to 0 \ . \tag{31}$$

Bis zu dieser Stelle ließen sich alle eindimensionalen Resultate einfach auf die n-dimensionalen Aufgaben übertragen. Allerdings entsteht bei dem Versuch, den Wert $D(P_o,\ldots,P_{N-1})$ auszurechnen, die erste Schwierigkeit: Um alle Punkte zu finden, für die die obere Grenze in (30) realisiert wird, muß man durch jeden Gitterpunkt n Ebenen legen, die parallel zu den Koordinatenebenen sind, und alle N^n Schnittpunkte betrachten. Der Grund dafür ist leicht zu verstehen, wenn man Abb. 13 aufmerksam betrachtet, auf der das zweidimensionale Gitter (n=2), bestehend aus N=4 Punkten, abgebildet ist und die Werte der Funktion $S_N(x_1,x_2)$ angeführt sind, die innerhalb jedes unterteilenden Rechtecks konstant sind: Die Funktion $S_N(P)$ hat nicht nur in den Punkten des Gitters Sprünge, sondern auch in anderen Ecken der unterteilenden Rechtecke.

Überhaupt stellt sich die Charakteristik $D(P_o,\ldots,P_{N-1})$ als recht schwierig untersuchbar dar. Ihre untere Grenze ist bis jetzt unbekannt (mit Ausnahme des Falles n=1, wenn inf D = 1/2). Die Spezialisten sind sich einig in der Meinung [1,2]*), daß die beste Möglichkeit der Abschätzung von D für ein n-dimensionales Gitter, das aus N Punkten besteht, gleich

$$D = O(\ln^{n-1}N) \tag{32}$$

und für eine n-dimensionale Folge bei beliebigen N

$$D = O(\ln^{n}N) \tag{33}$$

ist.

Für den speziellen Fall N=1 sind diese Behauptungen bewiesen; die Optimalität von (32) ist sogar für n=2 gezeigt. Für ein beliebiges n sind Gitter und Folgen, die den Abschätzungen (32) und (33) genügen, konstruiert, die Optimalität dieser Abschätzungen ist aber nicht bewiesen.

---

*) Siehe ebenfalls Übersichtsartikel: Niederreiter M. Quasi-Monte-Carlo methods and pseudorandom numbers. Bull. Amer. Math. Soc., 1978, 84, No. 6, p. 957-1041

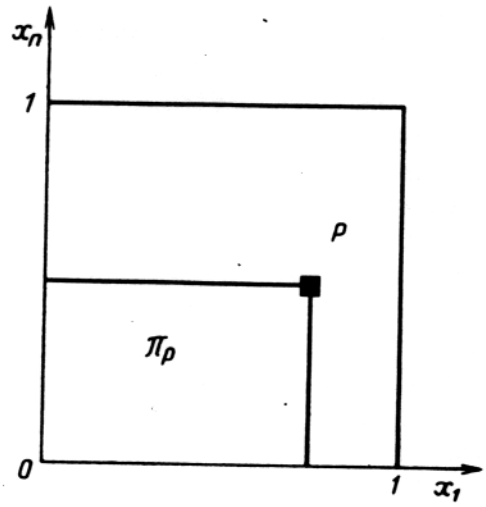

Abb. 12

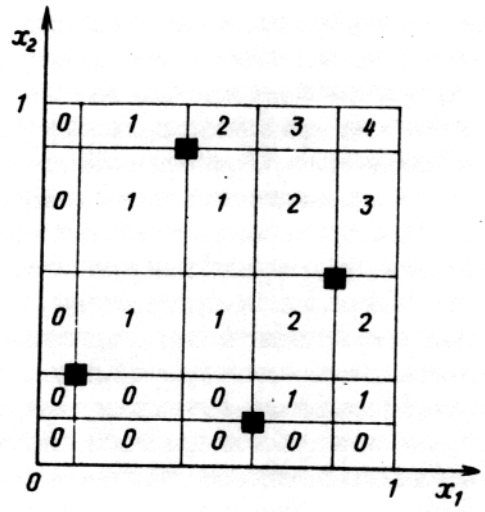

Abb. 13

Ungleichmäßigkeit. Wir betrachten in K^n wieder ein beliebiges Gitter, das aus N Punkten $P_0, P_1, \ldots, P_{N-1}$ besteht. Wir wählen ein beliebiges binäres Parallelepiped π_k. Die Koordinatenebenen teilen π_k in 2^n gleichgroße Oktanten (genauer: Hyperoktanten, Abb. 14). Wir bezeichnen mit V_k^+ die Menge aller positiven und mit V_k^- die Menge aller negativen Oktanten.
Die Größe $|(S_N(V_k^-)-S_N(V_k^+)|$ charakterisiert die Lage der Gitterpunkte zu π_k. Die obere Grenze dieser Größe über alle möglichen π_k

$$\sup_k |S_N(V_k^-)-S_N(V_k^+)| \qquad (34)$$

ist eine ganze Zahl, die irgendwie die Lage der Gitterpunkte in K^n charakterisiert. Leider kann diese Größe nicht als n-dimensionales Analog für Φ_∞ angenommen werden: Gute Werte (34) garantieren nicht die Gleichverteiltheit der Gitterpunkte in K^n. Um eine solche Garantie zu erhalten, muß man die Projektionen der Punkte $P_0, \ldots, P_{N-1}$ auf verschiedene Seitenflächen des Würfels K^n betrachten und für jede davon die entsprechende Größe der Gestalt (34) abschätzen.

Wir bemerken, daß bei der Definition von D im mehrdimensionalen Fall eine solche Schwierigkeit nicht entsteht, da die Diskrepanz der Projektionen niemals die Diskrepanz des Gitters selbst übersteigt. Im Falle eines zweidimensionalen Gitters ist z.B.

$$D = \sup_{0 \le x,y \le 1} |S_N(x,y)-Nxy| \ ;$$

für die Projektionen dieses Gitters auf die Achse 0x

$$D^1 = \sup_{0 \leq x \leq 1} |S_N^1(x) - Nx| ,$$

da aber hier $S_N^1(x) - Nx = S_N(x,1) - Nx \cdot 1$ gilt, ist offenbar $D^1 \leq D$.

Wir bezeichnen mit K_β eine s-dimensionale Seitenfläche des Würfels K^n, wobei $1 \leq s \leq n$ gilt; bei s=n wird der Würfel K^n selbst betrachtet. Die Projektionen der Punkte $P_0, \ldots, P_{N-1}$ auf K_β bilden ein s-dimensionales Gitter in K_β. Dafür kann die Größe (34) berechnet werden.

Ungleichmäßigkeit des Gitters $P_0, P_1, \ldots, P_{N-1}$ heißt die größte obere Grenze

$$\phi_\infty(P_0, \ldots, P_{N-1}) = \max_\beta \sup_k |S_N(V_k^-) - S_N(V_k^+)| . \qquad (35)$$

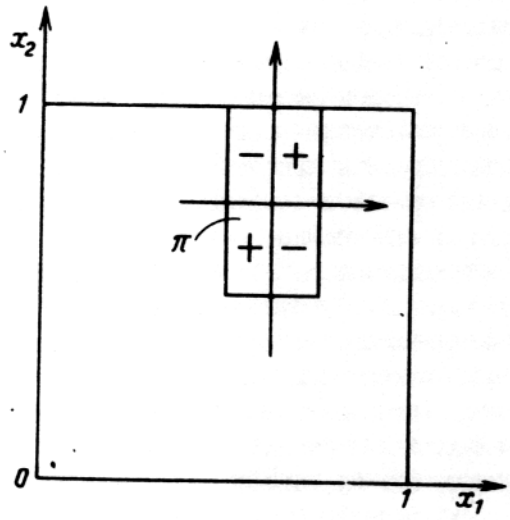

Abb. 14

Genauso wie im eindimensionalen Falle läßt sich zeigen, daß ϕ_∞ eine ganze Zahl mit $1 \leq \phi_\infty \leq N$ ist. Allerdings ist im n-dimensionalen Fall die untere Schranke schon nicht genau. Für n=2 ist bekannt, daß, wenn $N \geq 2$ gilt, $\phi_\infty(P_0, \ldots, P_{N-1}) \geq 2$ ist. Für $n > 2$ ist aber die genaue untere Schranke ϕ_∞ nicht bekannt. Die Abschätzung von ϕ_∞ durch D lautet:

$$\phi_\infty(P_0, \ldots, P_{N-1}) \leq 4^n D(P_0, \ldots, P_{N-1}) . \qquad (36)$$

Theorem 2. Dafür, daß $\{P_i\}$ g.v. ist, ist notwendig und hinreichend, daß für $N\to\infty$ gilt

$$\phi_\infty(P_0,\ldots,P_{N-1})/N \to 0 \ . \tag{37}$$

Im mehrdimensionalen Fall gibt die Ungenauigkeit der Charakteristik ϕ_∞ gewisse Vorteile vom Blickpunkt der Optimierung aus. Die genaue untere Grenze von ϕ_∞ in K^n ist nicht bekannt, bekannt ist aber die bestmögliche Wachstumsordnung: $\phi_\infty(P_0,\ldots,P_{N-1}) = O(1)$. Folgen mit beschränkten Ungleichmäßigkeiten sind weiter unten in §3 konstruiert.

Streuung. Diese Größe wird im n-dimensionalen Falle genauso definiert wie im eindimensionalen. Wenn die Punkte $P_0,P_1,\ldots,P_{N-1}$ gegeben sind, so ist ihre Streuung $d=d(P_0,\ldots,P_{N-1})$ gleich

$$d = \sup_P \min_{0\le i\le N-1} \rho(P,P_i) \ , \tag{38}$$

wobei die obere Grenze über alle möglichen Lagen der Punkte P im K^n genommen wird und $\rho(P,P')$ der euklidische Abstand zwischen den Punkten $P=(x_1,\ldots,x_n)$ und $P'=(x_1',\ldots,x_n')$ ist.
Wir merken zwei äußerst wichtige Eigenschaften der Größe d im n-dimensionalen Falle an. Erstens gelten die Abschätzungen (Niederreiter, 1977)

$$(\omega_n N)^{-1/n} \le d(P_0,\ldots,P_{N-1}) \le 2\sqrt{n}(D/N)^{1/n} \ , \tag{39}$$

wobei ω_n eine Konstante zahlenmäßig gleich dem Volumen der Einheitskugel im n-dimensionalen Raum ist. Zweitens gilt genauso wie im eindimensionalen Falle: Dafür, daß $\{P_i\}$ g.v. ist, ist notwendig, daß $d(P_0,\ldots,P_{N-1}) \to 0$ für $N\to\infty$ gilt. Diese Bedingung ist aber nicht hinreichend.

Kubische Gitter. Wir betrachten ein Gitter, das aus $N=M^n$ Punkten mit den Koordinaten

$$\left(\frac{i_1-1/2}{M}, \frac{i_2-1/2}{M}, \ \ldots\ , \frac{i_n-1/2}{M}\right) \tag{40}$$

besteht, wobei $i_1,i_2,\ldots,i_n$ unabhängig voneinander die Werte $1,2,\ldots,M$ durchlaufen. Auf Abb. 15 ist ein kubisches Gitter für $n=2$, $M=4$ abgebildet. Die Charakteristiken der Gleichmäßigkeit für ein solches Gitter sind leicht zu berechnen.

Es ist nicht schwer zu überprüfen, daß der Wert $|S_N(\pi_p)-NV_p|$ maximal wird z.B. im Punkt $P' = (1/2M,1,1,\ldots,1)$, wenn gilt $S_N(\pi_{p'}) = 0$, $NV_{p'} = N/2M = M^{n-1}/2$. Folglich ist

$$D = (1/2)N^{1-1/n} \quad . \tag{41}$$

Ebenso leicht ist es, sich davon zu überzeugen, daß unter allen Größen (34), die in der Formel (35) beteiligt sind, die maximalen diejenigen sind, die den eindimensionalen Projektionen des Gitters entsprechen. Tatsächlich stellen die Projektionen der Punkte (40) auf eine beliebige Koordinatenachse M geometrisch unterschiedliche Punkte dar, von denen jeder M^{n-1}-mal wiederholt wurde. Bei hinreichend feiner Unterteilung des Intervalls [0,1] in binäre Strecken gehören alle geometrisch voneinander verschiedenen Punkte verschiedenen ℓ_k an, so daß gilt:

$$\sup|S_N(\ell_k^-)-S_N(\ell_k^+)| = M^{n-1} \quad .$$

Hieraus folgt

$$\Phi_\infty = N^{1-1/n} \quad . \tag{42}$$

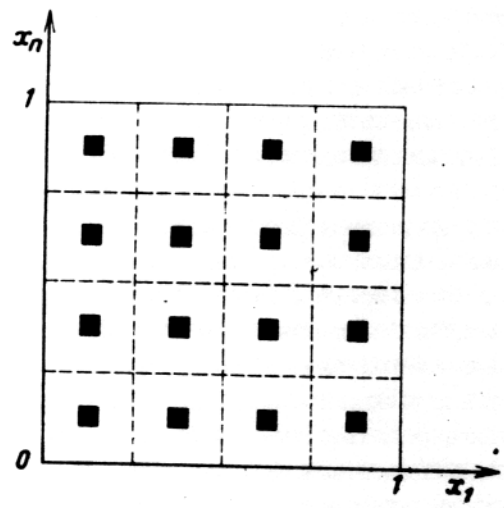

Abb. 15

Die Formeln (41) und (42) sind außerordentlich interessant: Aus ihnen folgt, daß bei n=1 die "kubischen Gitter" optimal sind (s. Abb. 1a): D=1/2, Φ_∞=1. Allerdings verschlechtert sich die Gleichmäßigkeit der Gitter (40) mit Vergrößerung von n, und die Ordnungen in den Formeln (41) und (42) nähern sich den schlechteren, die gleich N sind.

Notwendig ist die Bemerkung, daß die überwiegende Mehrheit der Leute intuitiv annimmt, daß im mehrdimensionalen Fall kubische Gitter die beste Gleichverteiltheit von Punkten im K^n realisieren. Viele Mißerfolge, die mit Versuchen, die kubischen Gitter für $n>3$ zu verwenden, zusammenhängen, werden oft als "verdammte Dimensionen" interpretiert. In Wirklichkeit sind, wie oben bereits bemerkt wurde, die besten Ordnungen für D und ϕ_∞ für die bekannten n-dimensionalen Gitter für alle Dimensionen $n\geq 2$ entsprechend gleich $O(\ln^{n-1}N)$ und $O(1)$, was besser ist als (41) und (42). Schon bei $n=2$ sind die Ordnungen von (41) und (42) gleich $\sqrt{N}$ - diese Ordnung entspricht ebenfalls den zufälligen Gittern, die aus N unabhängigen zufälligen Punkten bestehen, welche in K^2 gleichverteilt sind. Das heißt, für $n\geq 3$ sind die Gitter (40) asymptotisch (d.h. für $N\to\infty$) schlechter als die zufälligen.

Die Größe der Streuung d für die Gitter (40) berechnet sich ebenfalls einfach, da der zu einem beliebigen Punkt P nächstgelegene Gitterpunkt der Mittelpunkt eines entsprechenden kleinen Würfels ist (s. Abb. 15). Die schlechteste Lage eines Punktes P ist die Ecke eines kleinen Würfels. Da die Diagonale eines kleinen Würfels (nach dem n-dimensionalen Satz von Pythagoras) gleich $\sqrt{n(1/M)^2}$ ist, so gilt

$$d = (\sqrt{n}/2)N^{-1/n} . \qquad (43)$$

Vergleicht man (43) mit der linken Schranke von (39), so gelangt man zu dem (nach den vorherigen Überlegungen) unerwarteten Schluß, daß die Streuung kubischer Gitter für $N\to\infty$ von optimaler Ordnung ist.

Wir kehren zur Betrachtung dieses Paradoxes in §3 des Kapitels 3 zurück. Hier bemerken wir nur, daß, wie wir wissen, die Größe d keine vollwertige Charakteristik der Gleichmäßigkeit eines Gitters ist, und darum sollte man auf Grundlage der Abschätzung (43) nicht schließen, daß kubische Gitter sehr gut sind.

§ 2 LP_τ-Folgen

P_o-Gitter. In Analogie zu dem eindimensionalen Fall nennen wir P_o-Gitter ein Gitter, das aus $N=2^\nu$ Punkten des Würfels K^n besteht, wenn jedem binären π_k mit dem Volumen $1/N$ ein Gitterpunkt angehört. Im eindimensionalen Falle war die Konstruktion der P_o-Gitter eine triviale Sache, weil nur N binäre Strecken der Länge $1/N$ existieren. Mit der Vergrößerung von n wächst die Anzahl der binären Parallelepipede mit dem Volumen $1/N$ schnell, und N Punkte so zu verteilen, daß sie ein P_o-Gitter bilden, ist bei weitem nicht einfach. Zum Beispiel sind in Abb. 11 alle 32 binären Rechtecke des Flächeninhaltes 1/8 konstruiert. Man kann einfach überprüfen, daß die 8 Punkte, die in Abb. 16 dargestellt sind, ein P_o-Gitter im K^2 bilden: Jedem der 32 oben angemerkten Rechtecke gehört ein Gitterpunkt an.

Zweidimensionale P_o-Gitter bilden die Punkte mit den Koordinaten $(i/N,p(i))$ für $0\leq i\leq N-1$, dreidimensionale P_o-Gitter die Punkte mit den Koordinaten $(i/N,p(i),q(i))$ für $0\leq i\leq N-1$; hier ist $N=2^\nu$, und $\{p(i)\}$ und $\{q(i)\}$ sind Folgen, die in Kapitel 1 definiert wurden. Die Gültigkeit dieser Behauptungen folgt aus dem unten angeführten Theorem über die Konstruktion von P_τ-Gittern im K^{n+1} mit Hilfe der LP_π-Folgen aus K^n.

Allerdings erwies sich der Übergang zum vierdimensionalen Fall als unerwartet schwierig. Vielfältige Versuche, P_o-Gitter im K^4 zu konstruieren, waren erfolglos und endeten mit dem Beweis der Gegenbehauptung: Im K^4 ist es nicht möglich, ein P_o-Gitter zu konstruieren, das $N\geq 4$ Punkte enthält.

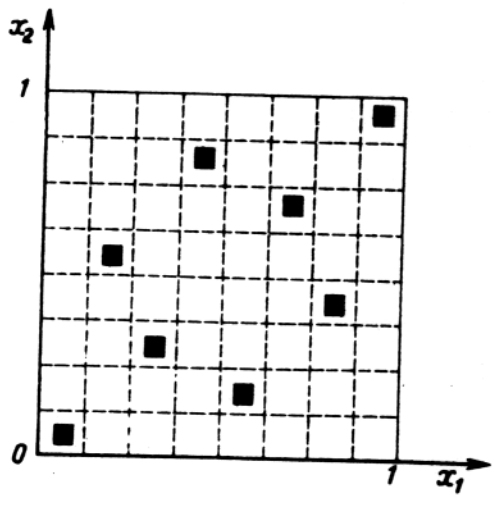

Abb. 16

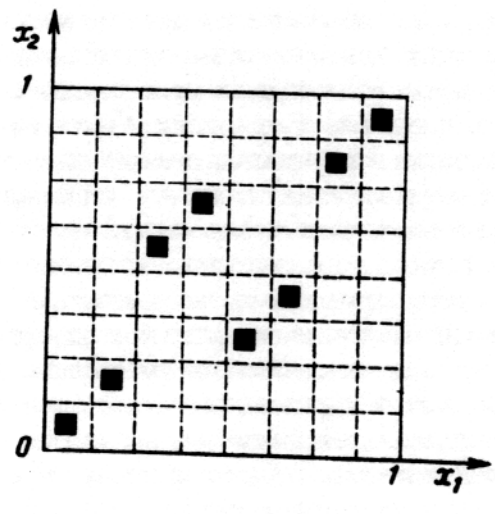

Abb. 17

Daher mußten die Forderungen, die an die Verteilung der Gitterpunkte an die binären Parallelepipede gestellt wurden, abgeschwächt werden, und es mußten allgemeinere Definitionen eingeführt werden.

P_τ-Gitter. Ein Gitter, das aus $N=2^\nu$ Punkten des Würfels K^n besteht, heißt P_τ-Gitter, wenn jedem binären π_k mit dem Volumen $2^\tau/N$ 2^τ Gitterpunkte angehören (es wird $\nu \geq \tau$ vorausgesetzt).

Zum Beispiel ist das ebene Gitter, das in Abb. 17 dargestellt ist und N=8 Punkte enthält, ein P_1-Gitter (jedem π_k mit $|\pi_k|=1/4$ gehören 2 Punkte an), ist aber kein P_0-Gitter.

P_τ-Gitter, die eine beliebig große Anzahl von $N=2^\nu$ Punkten enthalten, existieren in Räumen beliebiger Dimension n, aber den Wert τ muß man mit dem Wachstum von n ebenfalls vergrößern. Wir bezeichnen mit $\tau(n)$ den kleinsten Wert τ so, daß im K^n P_τ-Gitter existieren, die eine beliebig große Anzahl $N=2^\nu$ von Punkten enthalten. Die genauen Werte dieser Konstanten sind nur für die unteren Dimensionen bekannt:

$$\tau(1) = \tau(2) = \tau(3) = 0 , \quad \tau(4) = 1 .$$

Für beliebige Dimensionen ist in [1] gezeigt, daß $\tau(n)=O(n\ln n)$ ist, obwohl ebenfalls möglich wäre, daß in Wirklichkeit $\tau(n) \sim n$ ist.

Für beliebige P_τ-Gitter im K^n sind die Abschätzungen

$$\Phi_\infty = O(1), \quad D = O(\ln^{n-1} N)$$

gültig, deren Wachstumsordnung (nach N) unabhängig von τ ist. Die erste dieser Abschätzungen ist der Ordnung nach optimal, die zweite offenbar auch (vgl. (32)). Von τ hängen nur die Konstanten in diesen Abschätzungen ab, so daß die Frage über die Größe von $\tau(n)$ eine Frage nach den besten Konstanten in diesen Abschätzungen ist.
Wir merken die folgenden drei Eigenschaften von P_τ-Gittern an.

1° Die Projektionen der Punkte des P_τ-Gitters auf irgendeine s-dimensionale Seitenfläche K_β des Würfels K^n, wobei $1 \leq s \leq n-1$ gilt, bilden ein s-dimensionales P_τ-Gitter. Für dieses s-dimensionale Gitter übersteigt der Wert τ nicht den Wert τ für das Ausgangsgitter und kann streng kleiner sein.

2° Im K^n ist für ein beliebiges P_τ-Gitter folgende Abschätzung gültig:

$$\Phi_\infty \leq 2^{n-1+\tau} . \qquad (44)$$

3° Für $n=1,2,3$ und $N \geq 2^{n-1}$ wird für ein beliebiges P_0-Gitter im K^n die Ungleichung (44) zur Gleichung: $\Phi_\infty = 2^{n-1}$.

LP_τ-Folgen. Eine Folge der Punkte $P_0, P_1, \ldots, P_i, \ldots$ im K^n heißt LP_τ-Folge, wenn ein beliebiger binärer Abschnitt von ihr, der nicht weniger als $2^{\tau+1}$ Punkte enthält, ein P_τ-Gitter darstellt.
Aus Eigenschaft 1° der P_τ-Gitter folgt, daß die Projektion der Punkte einer LP_τ-Folge auf eine beliebige s-dimensionale Seitenfläche K_β eine s-dimensionale LP_τ-Folge (mit demselben oder einem kleineren Wert für τ) darstellt.
Für einen beliebigen Anfangsabschnitt einer beliebigen LP_τ-Folge im K^n gilt die Abschätzung

$$\Phi_\infty(P_0, \ldots, P_{N-1}) \leq 2^{n-1+\tau} . \qquad (45)$$

Aus Formel (45) und Theorem 2 folgt, daß beliebige LP_τ-Folgen gleichverteilt im K^n sind. Mehr noch, sie sind alle der Ordnung nach optimal, weil $\Phi_\infty(P_0, \ldots, P_{N-1}) = O(1)$ gilt.
Für einen beliebigen Anfangsabschnitt einer beliebigen LP_τ-Folge ist die Abschätzung (33) der Diskrepanz gültig. Da die Anfangsabschnitte der Länge $N=2^m$ für alle hinreichend großen m ein P_τ-Gitter darstellen, gilt für solche N auch die stärkere Abschätzung (32).

Konstruktion von P_τ-Gittern aus LP_τ-Folgen. Das folgende Theorem stellt eine Verallgemeinerung der Konstruktion von K.F. Roth (1954) dar, der mit Hilfe eindimensionaler gleichverteilter Folgen gute Gitter in Quadraten erzielte.

Theorem. Wenn die Punkte $P_0,P_1,\ldots,P_i,\ldots$ mit den Koordinaten $P_i=(x_{i1},\ldots,x_{in})$ eine LP_τ-Folge im K^n darstellen, so ist das Gitter, das aus $N=2^\nu$ Punkten mit den Koordinaten $(x_{i1},\ldots,x_{in},i/N)$, $0\leq i\leq N-1$, besteht, ein P_τ-Gitter im K^{n+1}.

Aus diesem Theorem folgt unter anderem, daß es nicht möglich ist, im K^3 eine LP_0-Folge zu konstruieren, da es im entgegengesetzten Fall möglich wäre, ein P_0-Gitter im K^4 zu konstruieren, was aber, wie wir wissen, unmöglich ist.

Gut gleichverteilte Folgen. Die Folge der Punkte $P_0,P_1,\ldots,P_i,\ldots$ heißt gut gleichverteilt (well distributed), wenn

$$D(P_k,P_{k+1},\ldots,P_{k+N-1})/N \to 0 \qquad (31')$$

gleichmäßig bezüglich $k=0,1,2,\ldots$ gilt. Dieser Begriff wurde von E. Hlawka (1955) und G.M. Petersen (1956) eingeführt. Offensichtlich ist jede gut gleichverteilte (kurz: w.v.) Folge ebenfalls gleichverteilt. Die Forderung (31) ist aber viel schärfer als die bisher existierende Forderung (31).
Man kann beweisen, daß für eine Folge $r_0,r_1,\ldots,r_i$ von unabhängigen zufälligen Punkten, die in K^n gleichverteilt im wahrscheinlichkeitstheoretischen Sinne sind, die Forderung (31) mit der Wahrscheinlichkeit 1 erfüllt ist, die Wahrscheinlichkeit der Gültigkeit von (31') aber gleich Null ist.
Die Forderung (31') kann durch eine äquivalente Forderung ersetzt werden: Gleichmäßig bezüglich $k=0,1,2\ldots$ gilt:

$$\Phi_\infty(P_k,P_{k+1},\ldots,P_{k+N-1})/N \to 0$$

Mit Hilfe dieses Kriteriums kann man die hinreichende Bedingung beweisen: Wenn $\Phi_\infty(P_0,P_1,\ldots,P_{N-1}) \leq C$ für alle N gilt, so ist $\{P_i\}$ w.v.. Hieraus folgt, daß alle LP_0-Folgen gut gleichverteilt sind.

§ 3 Die Konstruktion von LP_τ-Folgen

Natürlicherweise versucht man, LP_τ-Folgen im K^n so zu konstruieren, daß jede Koordinate eine eindimensionale LP_0-Folge darstellt. Solche LP_0-Folgen zu konstruieren, ist nicht schwer. Sie

aber so zu konstruieren, daß sie in gewissem Sinne unabhängig sind, ist weitaus schwieriger. Für dieses Ziel wurden lineare Differenzenoperatoren im Körper GF(2), der aus zwei Elementen besteht, der 0 und der 1, verwendet. Die Multiplikationsregeln im Körper sind die gewöhnlichen, die Additionsregeln entsprechen der Operation *, d.h. 0+1=1+0=1. 0+0=1+1=0.

Monozyklische Operatoren im Körper GF(2). Wir betrachten eine lineare Differenzengleichung der Ordnung m mit konstanten Koeffizienten

$$Lu_i = 0 , \qquad (46)$$

wobei der Operator L durch den Ausdruck

$$Lu_i = u_{i+m} + a_{m-1}u_{i+m-1} + \ldots + a_1 u_{i+1} + u_i$$

definiert ist; hierbei sind alle a_i und u_i Elemente des Körpers, d.h. Nullen oder Einsen.

<u>Lösung</u> der Gleichung (46) heißt die unendliche Folge

$$\ldots,u_{-2},u_{-1},u_o,u_1,u_2,\ldots,$$

die für alle $-\infty<i<\infty$ definiert ist und der Gleichung (46) für jedes i genügt. Jede Lösung ist eindeutig durch das Vorgeben der Gruppe $(u_1,\ldots,u_m)$ definiert, weil alle Werte $u_{m+1},u_{m+2},\ldots$ und alle Werte $u_o,u_{-1},u_{-2},\ldots$ nacheinander mit Hilfe von (46) ausgerechnet werden:

$$u_{i+m} = a_{m-1}u_{i+m-1} + \ldots + a_1 u_{i+1} + u_i , \quad i=1,2,\ldots;$$

$$u_i = u_{i+m} + a_{m-1}u_{i+m-1} + \ldots + a_1 u_{i+1} , \quad i=0,-1,-2,\ldots .$$

Da insgesamt 2^m verschiedene Gruppen $(u_1,\ldots,u_m)$ existieren, die aus Nullen und Einsen bestehen, so existieren insgesamt 2^m Lösungen, darunter eine triviale: $u_i \equiv 0$.
Wir betrachten die Gruppen $(u_1,\ldots,u_m),(u_2,\ldots,u_{m+1}),(u_3,\ldots,u_{m+2}),\ldots$. Darunter befindet sich auf jeden Fall eine Gruppe, die mit einer der schon betrachteten zusammenfällt. Folglich ist jede beliebige Lösung der Gleichung (46) periodisch, wobei die Periode 2^m-1 nicht übersteigt.
Der Operator L heißt <u>monozyklisch</u>, wenn die Gleichung (46) eine Lösung mit der größten Periode 2^m-1 hat.

Es ist nicht schwer zu überprüfen, daß sich beliebige nichttriviale Lösungen der monozyklischen Gleichung (46) nur in der Verschiebung der Numerierung der Elemente unterscheiden. In der Literatur werden solche Lösungen manchmal M-Folgen (Folgen mit maximaler Periode) genannt. Sie werden in der Kodierungstheorie und in Schemata der Generierung von pseudo-zufälligen Nullen und Einsen verwendet. Es gibt auch Tabellen von monozyklischen Operatoren.
Die ersten vier monozyklischen Operatoren lauten:

$$u_{i+1}+u_i\ , \qquad u_{i+2}+u_{i+1}+u_i\ ,$$

$$u_{i+3}+u_{i+1}+u_i\ , \qquad u_{i+3}+u_{i+2}+u_i\ .$$

Der Operator $u_{i+3}+u_{i+2}+u_i$ ist monozyklisch, und jede nichttriviale Lösung der Gleichung $u_{i+3}+u_{i+1}+u_i=0$ hat die Periode $2^3-1=7$. Zum Beispiel ist folgendes Lösung:

$$\ldots,1,0,0,1,0,1,1,1,0,0,1,0,1,1,\ldots\ .$$

Der Operator $u_{i+3}+u_{i+2}+u_{i+1}+u_i=0$ ist nicht monozyklisch, und die Gleichung $u_{i+3}+u_{i+2}+u_{i+1}+u_i=0$ hat Lösungen mit verschiedenen Perioden, z.B.

$$\ldots,1,0,0,1,1,0,0,1,\ldots \text{ oder } \ldots\ 1,0,1,0,\ldots$$
$$\text{oder } \ldots\ 1,1,1,1,\ldots\ .$$

Die Lösungen verschiedener monozyklischer Gleichungen sind in gewissem Sinne unabhängig. Gerade die Unabhängigkeit gestattete die Konstruktion von LP_τ-Folgen. Wir einigen uns auf die Sprechweise, daß die Richtungsmatrix (v_{sj}) zum Operator L der Ordnung m gehört, wenn drei Bedingungen erfüllt sind:

a) Jede der ersten m Spalten der Matrix ist Lösung der homogenen Gleichung $Lu_i=0$, so daß $Lv_{ij}=0$ für fixiertes $j=1,2,\ldots,m$ ist.
b) Jede der folgenden Spalten der Matrix ist Lösung der inhomogenen Gleichung $Lv_{ij}=v_{i,j-m}$ für fixiertes $j=m+1,m+2,\ldots$.
c) Auf der Hauptdiagonalen stehen Einsen und oberhalb von ihr Nullen.

Es ist nicht schwer zu beweisen, daß, wenn man die Anfangswerte der ersten m Spalten so wählt, daß in der linken oberen Ecke der Matrix (v_{sj}) auf der Hauptdiagonalen Einsen und oberhalb von ihr Nullen stehen, die Bedingung c) automatisch aus (b) folgt.

Aus dem Theorem, das in §3 des Kapitels 1 formuliert wurde, folgt, daß die BR-Folge, die einer solchen Richtungsmatrix entspricht, eine LP_o-Folge ist. Über eine solche Folge werden wir sagen, daß sie zum Operator L gehört.

Beispiel. Wir beweisen, daß die Matix (v_{sj}), nach der $\{q(i)\}$ im §3 des Kapitels 1 konstruiert wurde, zum monozyklischen Operator $u_{i+1}+u_i$ gehört.
Tatsächlich ist die Ordnung dieses Operators gleich 1, und die einzige nichttriviale Lösung der Gleichung $u_{i+1}+u_i=0$ besteht aus Einsen: $u_i \equiv 1$. Das ist die erste Spalte der Matrix. Aus der Bedingung (b) folgt die Gleichung zur Definition der restlichen Elemente der Matrix $v_{i+1,j}+v_{i,j}=v_{i,j-1}(\mathrm{mod}2)$. Es ist leicht zu sehen, daß das die nach Modul 2 betrachtete rekurrente Formel für die Berechnung der Binomialkoeffizienten, die in der Pascal-Matrix stehen, ist: $c_{i+1,j}=c_{ij}+c_{i,j-1}$.

Theorem. Seien $L_1,\ldots,L_{n-1}$ verschiedene monozyklische Operatoren, deren Ordnungen gleich $m_1,\ldots,m_{n-1}$ sind. Sei $\{p_k(i)\}$ irgendeine BR-Folge, die zum Operator L_k gehört. Die Folge der Punkte $Q_o,Q_1,\ldots,Q_i,\ldots$ mit den Koordinaten

$$Q_i = (p_1(i),p_2(i),\ldots,p_{n-1}(c))$$

ist eine LP_τ-Folge im K^{n-1} mit dem Wert

$$\tau = \sum_{k=1}^{n-1} (m_k-1) \ . \qquad (47)$$

Das vorläufige Theorem kann man verschärfen, wenn man als eine Koordinate $\{p(i)\}$ verwendet. Der Sinn der Verschärfung liegt darin, daß die Folge im K^n konstruiert wird, der Wert τ aber derselbe bleibt wie im vorherigen Theorem.

Theorem. Alle Bedingungen des vorherigen Theorems nehmen wir als erfüllt an. Die Folge der Punkte $Q_o,Q_1,\ldots,Q_i,\ldots$ mit den Koordinaten

$$Q_i = (p(i),p_1(i),\ldots,p_{n-1}(i)) \qquad (48)$$

ist eine LP_τ-Folge im K^n mit dem Wert τ, der in der Formel (47) definiert wurde.

§ 4 LP_τ-Folgen in der Anwendung

Die Folge der Punkte $Q_0, Q_1, \ldots, Q_i$, die für die Anwendung in der numerischen Praxis vorgesehen ist, muß drei Forderungen genügen:

1° Die Gleichmäßigkeit muß asymptotisch optimal sein.
2° Die Gleichmäßigkeit der Lage der Punkte muß nicht nur bei $N \to \infty$, sondern schon bei kleinen N gewährleistet sein.
3° Der Algorithmus der Berechnung der Punkte Q_i muß hinreichend einfach sein.

Der ersten Forderung genügt eine beliebige LP_τ-Folge, weil $\Phi_\infty(Q_0 \ldots, Q_{N-1}) = 0(1)$ eine optimale Abschätzung der Wachstumsordnung von Φ_∞ bezüglich N ist.
Günstig ist es, die Anfangsabschnitte der Länge $2^m, 2^{m+1}, 2^{m+2}, \ldots$ zu verwenden, da für sie $D = 0(\ln^{n-1} N)$ gilt.

Über zusätzliche Eigenschaften der Gleichmäßigkeit. Um in einem gewissen Maße der Forderung 2 zu genügen, wurden LP_τ-Folgen mit zusätzlichen Eigenschaften der Gleichmäßigkeit konstruiert. Wir erläutern die einfachste dieser Eigenschaften, die Eigenschaft A genannt wird.
Es sei $P_0, P_1, \ldots, P_i, \ldots$ eine Folge von Punkten im K^n. Wir unterteilen sie in Abschnitte der Länge $h = 2^n$:

$$[P_0, \ldots, P_{h-1}], [P_h, \ldots, P_{2h-1}], [P_{2h}, \ldots, P_{3h-1}], \ldots .$$

Man sagt, daß eine Folge $\{P_i\}$ die Eigenschaft A besitzt, wenn zu jedem der 2^n Oktanten des Würfels K^n ein Punkt aus jedem solchen Abschnitt gehört.
Das Vorhandensein der Eigenschaft A sichert in gewissem Maße die Gleichmäßigkeit der Lage der Anfangspunkte der Folge sogar für $N < 2^n$, weil jeder folgende Punkt in einen der "leeren" Oktanten gelangt, solange nicht alle Oktanten gefüllt sind.
In Entsprechung mit der Formel (47) ist es für die Verkleinerung von Wert τ wünschenswert, monozyklische Operatoren geringerer Ordnungen zu verwenden. I.B. Matusow (1980) bewies, daß man, wenn man alle monozyklischen Operatoren nach dem Wachstum der Ordnungen verteilt, die zu ihnen gehörenden Richtungsmatrizen so wählen kann, daß die Eigenschaft A für alle n erfüllt ist.

Über Algorithmen der Berechnung. Ein sehr einfacher Algorithmus für die Berechnung der Punkte (48) setzt das Vorhandensein einer Tabelle von Richtungspunkten $V^{(1)},V^{(2)},\ldots$ voraus, wobei $V^{(s)}=Q_2^{s-1}$ gilt. So seien $(V_1^{(s)},\ldots,V_n^{(s)})$ die Koordinaten des Punktes $V^{(s)}$, (so, daß $V_j^{(s)}$ die Richtungszahlen für $\{p_j(i)\}$ sind). Wenn im Dualsysem $i=e_m\ldots e_2e_1$ gilt, berechnen sich die Koordinaten $(q_{i1},\ldots,q_{in})$ des Punktes Q_i nach der Formel

$$q_{ij} = (e_1V_j^{(1)} * e_2V_j^{(2)} * \ldots * e_mV_j^{(m)}\ ,\ 1\leq j\leq n\ . \qquad (49)$$

Da alle $V_j^{(s)}$ binär rationale Zahlen der Gestalt $V_j^{(s)}=r_j^{(s)}2^{-s}$ sind, ist es günstiger, die Tabelle der Zähler $r_j^{(s)}$ zu speichern. Solche Tabellen sind erstellt für $1\leq s\leq 20$, $1\leq j\leq 51$ und gestattten eine einfache Berechnung der Punkte Q_i mit der Dimension $n\leq 51$ und einer Anzahl von $N<2^{21}$.

Schnelle Programme zur Berechnung der Punkte Q_i, die die Formel (49) mit Hilfe logischer Befehle realisieren, sind recht einfach für beliebige Computer zu erstellen. Ein sehr bequemes, wenn auch "langsames" Programm in FORTRAN ist in [3] und [4] enthalten.

Zur Beschleunigung der Geschwindigkeit der Berechnung der Punkte Q_i schlagen I.A. Antonow und W.M. Salejew (1979) vor, die Ordnung der Punkte Q_i so zu verändern (die binären Abschnitte aber beizuhalten), daß man die Koordinaten jedes nachfolgenden Punktes aus den entsprechenden Koordinaten des vorherigen Punktes mit Hilfe der einen Operation * erhält. In [3] ist ein "superschnelles" Programm angeführt, das diesen Algorithmus realisiert. Es ist in der Sprache FORTRAN erstellt, benötigt aber den Compiler FOREX.

§ 5 LP_τ-Folgen zur Basis r

Definition. Wir fixieren eine ganze Zahl $r\geq 2$. Wir nennen solche Strecken, die durch Teilung des Intervalles [0,1] in r^m gleiche Teile (m=0,1,2,...) erhalten werden können, Strecken zur Basis r. Ein Parallelepiped heißt zur Basis r, wenn alle ihn definierenden Strecken Strecken zur Basis r sind. In Abb. 18 sind alle Rechtecke (n=2) zur Basis 3 der Fläche 1/9 abgebildet.

Ein Gitter, das aus $N=r^\nu$ Punkten des Würfels K^n besteht, heißt P_τ-Gitter zur Basis r, wenn jedem Parallelepiped π_k mit dem Volumen 1/N ein Punkt des Gitters zugehört. Es ist nicht schwer

zu verstehen, daß sich bei ungefähr gleichen N die Anzahl der Parallelepipede zur Basis r des Volumens 1/N mit der Vergrößerung von r verkleinert.

Beispiel. Im vierdimensionalen Würfel K^4 betrachten wir alle möglichen π_k zur Basis r des Volumens 1/N, wobei $N=r^\nu$ ist. Wenn die Länge der j-ten Kante mit r^{m_j} bezeichnet wird, so ist das Volumen von π_k gleich $r^{-m_1}\ldots r^{-m_4}=r^{-\nu}$. Folglich genügen die Zahlen $m_1,\ldots,m_4$, die den Typ des Parallelepipedes π_k definieren, der Gleichung

$$m_1+m_2+m_3+m_4 = \nu \ . \qquad (50)$$

Die Anzahl der nichtnegativen ganzzahligen Lösungen dieser Gleichung ist $(\nu+1)(\nu+2)(\nu+3)/6$.
Wenn $N=2^7=128$ ist, so ist die Anzahl der Typen gleich 120 und die Anzahl der verschiedenen binären π_k gleich $120\cdot128=15360$.
Wenn $N=5^3=125$ ist, so ist die Anzahl der Typen gleich 20 und die Anzahl der verschiedenen π_k zur Basis 5 gleich $20\cdot125=2500$.
Wenn $N=11^2=121$ ist, so ist die Anzahl der Typen gleich 10 und die Anzahl der verschiedenen π_k zur Basis 11 nur $10\cdot121=1210$.

Daraus, daß sich mit dem Wachstum von r die Anzahl der Parallelepipede zur Basis r verringert, folgt, daß die Bedingungen für die Existenz der P_o-Gitter zur Basis r mit der Vergrößerung von r besser werden. Das heißt, in den Fällen, in denen es nicht möglich ist, im K^n binäre P_o-Gitter zu konstruieren, kann man versuchen, P_o-Gitter zur Basis r mit hinreichend großem r zu konstruieren.
Die Folge der Punkte $P_o,P_1,\ldots,P_i,\ldots$ im K^n heißt LP_o-Folge zur Basis r, wenn ein beliebiger Abschnitt zur Basis r von ihr ein P_o-Gitter zur Basis r darstellt.

Konstruktion. Zum ersten Mal werden P_o-Gitter und LP_o-Folgen zur Basis r in einer Arbeit von H. Faure (1982) betrachtet, der bewies, daß, wenn r eine Primzahl mit $r\geq n$ ist, im K^n LP_o-Folgen zur Basis r existieren.

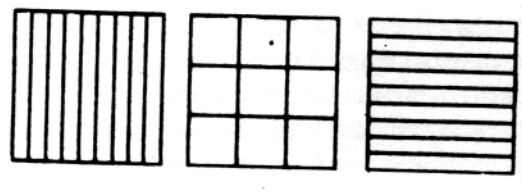

Abb. 18

Die von Faure konstruierten FLP_o-Folgen zur Basis r stellen eine direkte Verallgemeinerung der binären LP_o-Folgen im K^2 dar, deren Punkte $Q_i=(p(i),q(i))$ sind. Tatsächlich stellen, wie wir bereits sahen, die Richtungsmatrizen, nach denen $\{p(i)\}$ und $\{q(i)\}$ konstruiert worden sind, die unendliche Einheitsmatrix E und die Pascal-Matrix C(mod2) dar. Aus dem Lemma, das am Ende des §3 von Kapitel 1 angeführt ist, folgt, daß $C^2=E(mod2)$ ist. Das letzte Resultat kann leicht auf ein beliebiges ganzes r verallgemeinert werden: $C^r=E(modr)$.
Als Richtungsmatrizen für die Konstruktion einer LP_o-Folge zur Basis r im K^n werden beliebige n Matrizen aus der Gruppe

$$E,C(modr),C^2(modr),\ldots,C^{r-1}(modr)$$

vorgeschlagen. Die Koordinaten der Punkte einer solchen Folge sind Folgen des "rationalen Typs zur Basis r"; diese Definition ist völlig analog zur Definition der Folgen vom binär-rationalen Typ in Kapitel 1.
Für die Anfangsabschitte einer beliebigen LP_o-Folge zur Basis r ist folgende Abschätzung gültig: $D(P_o,\ldots,P_{N-1})=0(\ln^n N)$, und für $N=r^m$ die stärkere Abschätzung $D(P_o,\ldots,P_{N-1})=0(\ln^{n-1} N)$.
Es ist vorerst unklar, welche Folgen - die binären LP_τ-Folgen oder die LP_o-Folgen zur Basis r - bei praktischen Anwendungen bevorzugt werden sollten. Zugunsten der ersteren sprechen die zusätzlichen Eigenschaften der Gleichmäßigkeit, zugunsten der letzteren - die asymptotischen Abschätzungen der Konstanten in den Formeln $D=0(\ln^n N)$ (s. unten). Es ist nicht ausgeschlossen, daß sich bei nicht sehr großen n die ersteren, bei großen n hingegen die letzteren als besser erweisen.

Über Folgen mit der besten Asymptotik der Abweichungen. Im Jahre 1960 schluß J.M. Hammersley vor, für die Konstruktion von mehrdimensionalen Gittern die Verallgemeinerung der Folgen $\{p(i)\}$ zur Basis r zu verwenden: Wenn im Zahlensystem zur Basis r $i=e_m,\ldots,e_2e_1$ ist, so ist ebenfalls im Zahlensystem zur Basis r $\phi_r(i)=0,e_1,e_2,\ldots,e_m$; hier sind alle e_j Ziffern des Zahlensystems zur Basis r, d.h. $0,1,2,\ldots,r-1$. Offensichtlich gilt $\phi_2(i)\equiv p(i)$.
Seien $r_1,\ldots,r_n$ Primzahlen. Die Folgen der Punkte $P_o,P_1,\ldots,P_i,$ deren Koordinaten

$$P_i = (\phi_{r_1}(i), \phi_{r_2}(i), \ldots, \phi_{r_n}(i))$$

sind, untersuchte J.H. Halton (1960) und bewies, daß für diese gilt

$$D(P_o, \ldots, P_{N-1}) = O(\ln^n N) \ . \qquad (51)$$

Der Chronologie nach war das die erste Klasse von Folgen mit einer so guten Asymptotik. Wie bekannt ist, ist eine ebensolche Abschätzung wahr für alle LP_τ-Folgen und für LP_o-Folgen zur Basis r.

Wenn man in (51) das Hauptglied der Abschätzung heraushebt, so wird es in der Gestalt $B(n)\ln^n N$ geschrieben. Wie verhält sich B(n) für $n \to \infty$? Es stellt sich heraus, daß für Haltonfolgen (für die günstigste Wahl von $r_1, \ldots, r_n$) die Größe $\ln B(n) = O(n \ln n)$, für LP_τ-Folgen (bei Wahl des minimalen τ) $\ln B(n) = O(n \ln \ln n)$ und für LP_o-Folgen zur Basis r (bei Wahl des minimalen r) die Größe $B(n) \to 0$ ist.

Man sollte die Resultate dieses Vergleiches nicht überbewerten. Erstens sind diese Abschätzungen richtig für Klassen von Folgen; zweitens ist die Abschätzung von B(n) für LP_τ-Folgen unter der Voraussetzung hergeleitet worden, daß das beste $\tau \sim n \log_2 n$ ist; das war die Ordnung der Abschätzung von τ aus [1]. Diese Abschätzung ist aber nicht genau. Und wenn sich erweist, daß $\lim_{n \to \infty} \sup(\tau / n \log_2 n) < 1$ gilt, zieht das die Beziehung $B(n) \to 0$ nach sich.

Die Bevorzugung von LP_τ-Folgen gegenüber den Halton-Folgen erfolgt gewöhnlich nicht der Aymptotik von B(n) wegen, sondern der zusätzlichen Eigenschaften der Gleichmäßigkeit und dem Vorhandensein von Abschnitten wegen, die eine verbesserte Abschätzung der Abweichungen $D = O(\ln^{n-1} N)$ zulassen.

Kapitel 3. Anwendungen der gleichverteilten Folgen in der Numerischen Mathematik

§ 1 Näherungsweise Berechnung mehrdimensionaler Integrale

Die näherungsweise Abschätzung eindimensionaler Integrale ruft gewöhnlich keine Schwierigkeiten hervor. Deshalb sind die Resultate von §4 des Kapitels 1 in der Hauptsache von theoretischem Interesse. Andererseits findet die Verallgemeinerung dieser Resultate auf mehrdimensionale Integrale nicht wenige praktische Anwendungen.

Bezeichnungen. Wir bezeichnen mit β die Menge der natürlichen Zahlen $\beta=(i_1,\ldots,i_s)$, die den Ungleichungen

$$1\leq i_1<i_2< \ldots <i_2\leq n\ , \qquad 1\leq s\leq n\ , \tag{52}$$

genügen. Jedem β entspricht eine s-dimensionale Seitenfläche des Würfels K^n, auf der sich die Koordinaten $x_{i_1},\ldots,x_{i_s}$ von 0 bis 1 verändern und alle anderen Koordinaten gleich 1 sind. Wir bezeichnen diese Seitenfläche mit K_β. Wenn s=n ist, so ist $\beta=(1,2,\ldots,n)$, und die Rolle von K_β spielt der Würfel K^n selbst (Abb. 19, n=3).

Die Projektion eines beliebigen Punktes $T=(t_1,\ldots,t_n)$ aus dem K^n auf K_β bezeichnen wir mit T_β, so daß die Koordinaten von T_β gleich $(1,\ldots,1,t_{i_1},1,\ldots,1,t_{i_2},1,\ldots)$ sind. Sei der Kürze halber $dt_{i_1},\ldots,dt_{i_s} = dT_\beta$.

Neben dem gegebenen Gitter $P_0,\ldots,P_{N-1}$ im K^n betrachten wir die Projektion der Punkte dieses Gitters auf K_β. Sie bilden ein s-dimensionales Gitter, deren Abweichung wir mit

$$D^\beta \sup_{P_\beta} \left|S_N^\beta(\pi_{P_\beta}) - NV_{P_\beta}\right|$$

bezeichnen; hierbei ist $S_N^\beta(\pi_{P_\beta})$ die Anzahl der Projektionen, die in den s-dimensionalen Parallelepiped π_{P_β}, die Projektion von π_P, entfallen und V_{P_β} sein Volumen gleich dem Produkt $x_{c_1}\ldots x_{c_s}$ (Abb. 20). Wie schon auf Seite 16 bemerkt wurde, sind alle $D^\beta\leq D$.

Wir setzen voraus, daß die Funktion f(P) stetig im K^n und alle ihre partiellen Ableitungen, die nicht mehr als eine Differentiation nach jeder Koordinate enthalten, stückweise stetig im K^n sind. Für die Bezeichnung dieser Ableitung ist es ebenfalls bequem, den Index β zu verwenden:

$$f^{(\beta)}(P) = \partial^s f(P) \,/\, \partial x_{i_1} \dots \partial x_{i_s} , \qquad (53)$$

wobei $i_1,\dots,i_s$ den Bedingungen (52) genügt.

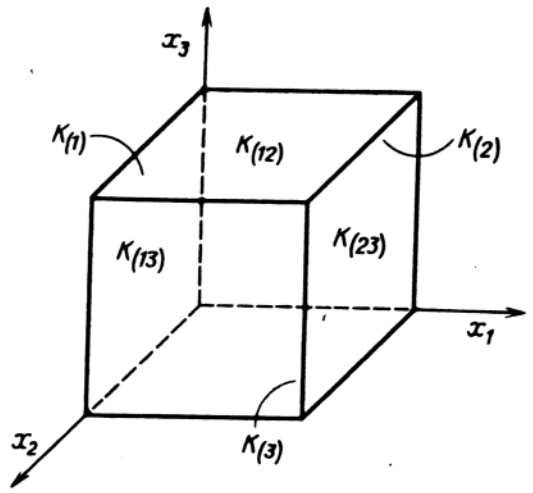

Abb. 19

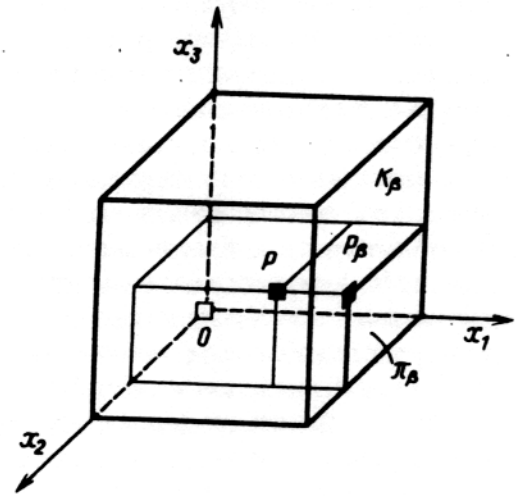

Abb. 20

Die höchste unter den Ableitungen (53) entspricht $\beta=(1,2,\dots,n)$ und ist gleich $\partial^n f/\partial x_1,\dots,\partial x_n$.
Die Summation über alle β, die (52) genügen, wollen wir mit einem Zeichen schreiben, so daß z.B. gilt

$$\sum_\beta F_\beta = \sum_{i_1=1}^{n} F_{i_1} + \sum_{i \le i_1 < i_2 \le n} F_{i_1 i_2} + \dots + F_{12\dots n} ,$$

das sind insgesamt 2^n-1 Summanden.
Sei $P=(x_1,\dots,x_n)$, $T=(t_1,\dots,t_n)$, $A=(1,1,\dots,1)$. Dann gilt

$$f(P) = f(A) + \sum_\beta (-1)^s \int_{x_{i_1}}^{1} \dots \int_{x_{i_s}}^{1} f^{(\beta)}(T_\beta)\,dT_\beta . \qquad (54)$$

Zur Erklärung dieser Formel schreiben wir sie für den Fall n=2 auf:

$$f(x,y) = f(1,1) - \int_x^1 f'_x(t,1)\,dt - \int_y^1 f'_y(1,v)\,dv + \int_x^1 \int_y^1 f''_{xy}\,dtdv .$$

Hier gehen nicht alle Werte der Ableitungen f'_x und f'_y ein, sondern nur die, die für die Definition der Werte von $f(x,y)$ benötigt werden.

Hauptlemma. Wir bezeichnen mit $\delta(f)$ den Fehler der Näherung

$$\delta(f) = \frac{1}{N} \sum_{i=0}^{N-1} f(P_i) - \int_{K^n} f(P)dP \ . \qquad (55)$$

Lemma. Wie man die Punkte $P_0, P_1, \ldots, P_{N-1}$ auch immer wählt, es gilt

$$\delta(f) = \frac{1}{N} \sum_{\beta} (-1)^s \int_{K_\beta} (NV_{T_\beta} - S_N^\beta(\pi_{T_\beta})) \cdot f^{(\beta)}(T_\beta)dT_\beta \ . \qquad (56)$$

Ausgehend von der Darstellung (54) wird das Lemma genauso bewiesen wie im eindimensionalen Fall.

Abschätzungen für Klassen von Funktionen. Wir betrachten eine Zusammenstellung von Konstanten L_β, die sich aus 2^n-1 nicht negativen Konstanten $L_{i_1 \ldots i_s}$ zusammensetzt. Mit $W_1^1(L_\beta)$ bezeichnen wir die Menge der stetigen Funktionen $f(P)$, deren partielle Ableitungen (53) stückweise stetig sind und für die gilt

$$\int_{K_\beta} |f^{(\beta)}(T_\beta)|dT_\beta \leq L_\beta \ .$$

Genauso wie im eindimensionalen Fall interessiert uns die Größe $R=\sup|\delta(f)|$, wobei die obere Grenze über alle möglichen Funktionen $f(P) \in W_1^1(L_\beta)$ genommen wird. Mit Hilfe des Hauptlemmas kann man beweisen, daß, wie man auch immer die Punkte $P_0, \ldots, P_{N-1}$ aus K^n nimmt, stets

$$R = \frac{1}{N} \sum_{\beta} L_\beta D^\beta$$

gilt. Wenn man als Stützstellen für die Integration in (55) die Punkte einer LP_τ-Folge wählt, so erhält man bei beliebigen N, daß $R=O(N^{-1}\ln^n N)$ gilt und für $N=2^m$ die schärfere Abschätzung

$$R = O(N^{-1}\ln^{n-1} N)$$

gilt. N.S. Bachwalow (1972) bewies, daß die letzte Abschätzung der Ordnung nach nicht verbessert werden kann.

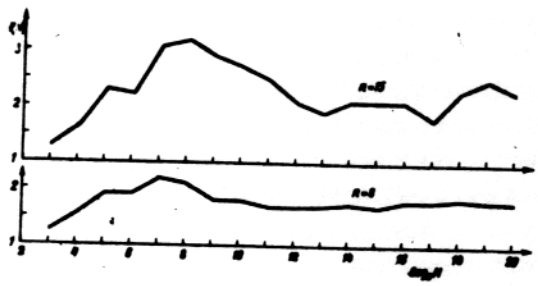

Abb. 21

Mehrdimensionale Analoge der Abschätzung (25) wurden unabhängig voneinander in den Arbeiten von I.M. Sobol (1961) und E. Hlavka (1961) erhalten. In der ersten dieser Arbeiten wurde die Hauptaufmerksamkeit auf die Konstruktion von Klassen solcher Funktionen gelegt, für die genaue Fehlerabschätzungen erhalten werden können; in der zweiten Arbeit wird die Möglichkeit der Verallgemeinerung der Abschätzungen auf einige Klassen von Funktionen mit Unstetigkeiten hervorgehoben.

Beispiel (I.M. Sobol, J.L. Lewitan 1978):
Wir betrachten die Funktionen

$$g_n(P) = \prod_{j=1}^{n} \frac{j+2x_j}{j+1} ,$$

für die die Integrale über K^n gleich 1 sind. Für n=8 und n=15 können diese Integrale mit Hilfe der Punkte einer LP_τ-Folge abgeschätzt werden.
In der Zeichnung 21 sind die Werte $H=N(\delta(g_n f))$ als Funktion von $\log_2 N$ angegeben. Für n=8 gelangen die Werte von H(N) bereits in die Asymptotik, so daß offensichtlich $\delta(g_8) \sim 1{,}8/N$ gilt. Bei n=15 gelangt man noch nicht in den asymptotischen Bereich, bei allen hinreichend großen N ist jedoch $2<H(N)<3$, so daß auch in diesem Falle $\delta(g_{15}) = O(1/N)$ gilt.

§ 2 Quasizufällige Punkte

In der Mehrheit der Aufgaben, die mit der Monte-Carlo-Methode gelöst werden, wird die gesuchte Größe in der Gestalt der mathematischen Erwartung einer gewissen Zufallsgröße dargestellt: $a=M\eta$. Dann kann man, indem man N unabhängige Werte dieser Größe $\eta_1,\ldots,\eta_N$ erhält, a näherungsweise abschätzen:

$$a \approx \frac{1}{N} \sum_{i=1}^{N} \eta_i \ . \qquad (57)$$

Nach dem Gesetz der großen Zahlen konvergiert die rechte Seite von (57) in der Wahrscheinlichkeit gegen a, wenn $N\to\infty$ ist.

Die Formel (57) ist aber noch kein Algorithmus zur Berechnung: Zu ihr gehörten noch eine Formel zur Modellierung von η mit Hilfe von Standardzufallszahlen $\gamma_1,\gamma_2,\ldots$, die gleichverteilt im Intervall (0,1) sind.
Wir schreiben die Formel zur Modellierung von η in der Gestalt

$$\eta = \phi(\gamma_1,\gamma_2,\ldots) \ . \qquad (58)$$

Beide Formeln (57) und (58) definieren den <u>Monte-Carlo-Algorithmus</u> zur Berechnung von a.
Wenn die Funktion ϕ in (58) von n Argumenten abhängt, $\phi=\phi(\gamma_1,\ldots,\gamma_n)$, so sagt man, daß die konstruktive Dimension des Algorithmus (57)-(58) gleich n ist (kurz: k.D.=n).
Alle Monte-Carlo-Algorithmen mit k.D.=n lassen eine gemeinsame Interpretation zu. Tatsächlich ist

$$a = M\eta = M\phi(\gamma_1,\ldots,\gamma_n) = \int_{K^n} \phi(P)dP \ , \qquad (59)$$

weil der Zufallspunkt $\Gamma=(\gamma_1,\ldots,\gamma_n)$ gleichverteilt im Würfel K^n ist.
(Wir erinnern daran, daß ein Zufallspunkt gleichverteilt heißt, wenn seine Wahrscheinlichkeitsdichte $p(x_1,\ldots,x_n)\equiv 1$ in K^n ist.)

Weiter, wenn $\gamma_{1i},\ldots,\gamma_{ni}$ Zufallszahlen sind, nach denen η_i modelliert wird, so kann man sie als kartesische Koordinaten des Zufallspunktes $\Gamma_i=(\gamma_{1i},\ldots,\gamma_{ni})$ betrachten, der im K^n gleichverteilt ist. Folglich gilt $\eta_i=\phi(\Gamma_i)$, und die Formel (57) kann man in der Gestalt

$$a \approx \frac{1}{N} \qquad (60)$$

schreiben. Die Beziehungen (59) und (60) zeigen, daß man einen beliebigen Monte-Carlo-Algorithmus der Gestalt (57)-(58) mit k.D.=n als Algorithmus zur Berechnung eines n-dimensionalen Integrals über K^n auf dem Wege der Mittlung der zu integrierenden Funktion in den unabhängigen Zufallspunkten $r_1,\ldots,r_N$ interpretieren kann.

Andererseits wissen wir, daß man solche Integrale mit Hilfe der Punkte einer n-dimensionalen LP_τ-Folge $Q_1,Q_2,\ldots$ berechnen kann:

$$a = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \Phi(Q_i) \ . \qquad (61)$$

Der Vergleich der Formeln (60) und (61) zeigt, daß man, wie auch immer der Monte-Carlo-Algorithmus (57)-(58) mit k.D.=n geartet ist, für seine Realisierung anstelle der Zufallspunkte r_i die nicht zufälligen Punkte Q_i verwenden kann. Das bedeutet, daß anstelle der Standardzufallszahlen für die Berechnung der i-ten Realisierung $\eta=\eta_i$ die Koordinaten des i-ten Punktes Q_i verwendet werden sollten. In allem anderen bleibt der Monte-Carlo-Algorithmus ohne Veränderungen.

In einigen Algorithmen beschleunigt der Übergang zu den Punkten Q_i die Konvergenz wesentlich: anstelle des Fehlers der Ordnung $N^{-1/2}$, der für die Näherung (60) charakteristisch ist, hat man für (61) einen Fehler der Ordnung $N^{-1}\ln^n N$, d.h. besser als $N^{-1+\varepsilon}$ mit beliebigem $\varepsilon>0$. Besonders interessant ist, daß, während die Fehlerordnung von (60), die gleich $N^{-1/2}$ ist, von der k.D. des Algorithmus nicht abhängt, die Fehlerordnung von (61) mit der Vergrößerung der k.D. schlechter wird, was zur Bevorzugung des Algorithmus mit kleinerer k.D. nötigt (falls man die Auswahl hat).

Die Punkte, die man in den Monte-Carlo-Algorithmen anstelle der zufälligen verwenden kann und bei denen die Konvergenz des Algorithmus erhalten bleibt, nennt man mitunter <u>quasizufällig</u>. So kann man sagen, daß die Punkte einer n-dimensionalen LP_τ-Folge quasizufällige Punkte für die Monte-Carlo-Algorithmen mit k.D.=n darstellen.

Beispiel. (I.M. Sobol, S.G. Rosin, L.Ch. Chomskij, 1976)
Es wurden verschiedene Monte-Carlo-Algorithmen zur Abschätzung der mathematischen Erwartung $M\phi$ untersucht, wobei $\phi=\exp(\xi\eta^2)$ und der Zufallspunkt (ξ,η) im K^2 mit der Wahrscheinlichkeitsdichte $p(x,y)=2y$ definiert ist.
Einer der betrachteten Algorithmen hat die k.D.=3:

$$\varepsilon = \exp\left\{\gamma_1[\max(\gamma_2,\gamma_3)]^2\right\}$$

Die Resultate der Berechnungen sind in Abb. 2 vorgestellt, wobei auf der Abszissenachse die Werte $\log_2 N$ und auf der Ordinatenachse die Größen

$$H(N) = \sqrt{N}\ \left|M\phi-\frac{1}{N}\sum_{i=1}^{N}\phi_i\right|$$

abgetragen sind. Die Linien 1 und 2 sind die Berechnungen mit Hilfe eines Gebers von Pseudozufallszahlen; die Resultate vereinbaren sich gut mit der theoretischen Abschätzung des wahrscheinlichen Fehlers $\delta_{wahr} = 0{,}22/\sqrt{N}$, der $H_{wahr} = \sqrt{N}\delta_{wahr} = 0{,}22$ entspricht. Die Linie 3 ist das Resultat der Berechnung mit dreidimensionalen Punkte Q_i; in dieser Berechnung erwies sich, daß $H \sim 0{,}8/\sqrt{N}$ ist, so daß für den Fehler gilt $\delta \sim 0{,}8/N$.

Der Fall K.D.=∞. Da die Anzahl der monozyklischen Operatoren, mit denen in §3 des Kapitels 2 LP_τ-Folgen konstruiert wurden, unendlich ist, kann man eine unendlichdimensionale LP_τ-Folge $x_0,x_1,\ldots,x_i,\ldots$ konstruieren, wobei jeder Punkt von ihr die Gestalt $x_i=(p(i),p_1(i),p_2(i),\ldots,p_n(i),\ldots)$ hat. Solche Punkte können die Rolle von quasizufälligen Punkten für die Algorithmen (57)-(58) mit beliebigen k.D. spielen. Allerdings gibt es hinreichend effektive Möglichkeiten zur Berechnung solcher Punkte (für beliebig große n) bisher nicht.

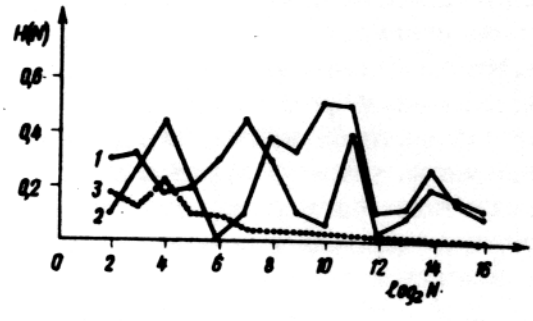

Abb. 22

§ 3 LP-Suche

Die einfachste Suche. Wir setzen voraus, daß im Würfel K^n eine stetige Funktin $f(P)$ gegeben ist und daß ihr minimaler Funktionswert $f^*=\inf f(P)$ gesucht ist, wobei $P \in K^n$ ist. Die einfachste Suche von f^* besteht in folgendem. Im K^n werden die Versuchspunkte $P_1,\ldots,P_N$ ausgewählt; in jedem von ihnen wird der Funktionswert $f(P_i)$ berechnet, und als Näherung zu f^* wird $f_N^*=\min f(P_i)$ bei $1\leq i\leq N$ genommen. Es entsteht die Frage: Wie sind die Versuchspunkte zu wählen?

Die einfachste Suche heißt zufällig, wenn in der Eigenschaft der Versuchspunkte unabhängige Zufallspunkte r_i genommen werden, die im K^n gleichverteilt sind. Die Konvergenz einer solchen Suche ist leicht nachzuweisen. Sei P^* ein Punkt, in dem das Minimum realisiert wird: $f(P^*)=f^*$. Sei U eine beliebig kleine Umgebung des Punkte P^*, und sei das Volumen dieser Umgebung gleich $\varepsilon>0$. Die Wahrscheinlichkeit dafür, daß mindestens einer der Punkte $r_1,\ldots,r_N$ auf U entfällt, ist gleich $1-(1-\varepsilon)^N$ und strebt für $N\to\infty$ gegen 1.

Es ist völlig klar, daß für den Erfolg der einfachsten Suche die Zufälligkeit der Versuchspunkte nicht benötigt wird: Wichtig ist die Gleichmäßigkeit ihrer Aufteilung. Darum ist es natürlich, für dieses Ziel die Punkte von am besten gleichverteilten Folgen zu verwenden. In Abb. 23 sind 32 zufällige Punkte und 32 Punkte mit den Koordinaten $(p(i),i/32)$, $0\leq i\leq 31$, dargestellt. Sogar in diesem einfachen Beispiel ist der Vorteil der nicht zufälligen Punkte in der Gleichmäßigkeit der Aufteilung offensichtlich.

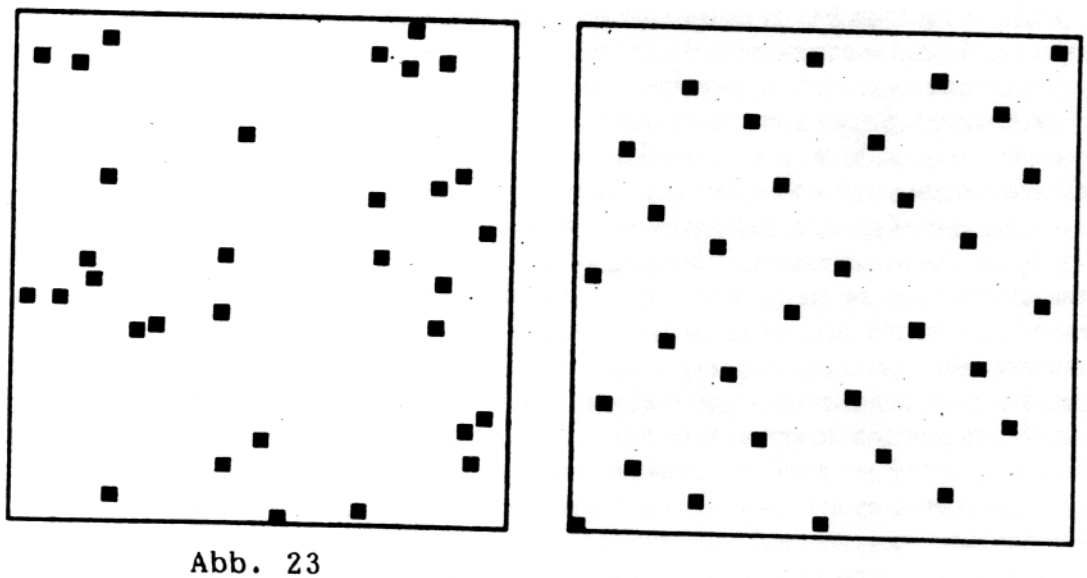

Abb. 23

zufälliges Gitter π_o-Gitter

Die einfachste Suche heißt LP-Suche, wenn in der Eigenschaft der Versuchspunkte die Punkte einer LP_τ-Folge $\{Q_i\}$ gewählt werden. In diesem Falle läßt sich die Konvergenz der Suche sehr leicht nachweisen: Da $S_N(U)/N \to \varepsilon$ gilt, ist $S_N(U) \to \infty$, und die Anzahl der Punkte Q_i, die auf U entfallen, wird beliebig groß, wenn N hinreichend groß ist.

Auf den ersten Blick scheint es, daß die einfachste Suche nicht in der Lage ist, mit den gerichteten Suchmethoden zu konkurrieren (Gradientenmethode, konjugierte Richtungen u.a.), in denen die Lage des aktuellen Versuchspunktes P_i nicht von vornherein festgelegt ist, sondern unter Berücksichtigung der schon gefundenen Werte $f(P_1),\ldots,f(P_{i-1})$ gewählt wird. Das ist wirklich so, wenn die Funktion f(P) insgesamt ein Minimum hat (und ihr Relief hinreichend gut ist). Demgegenüber ändert sich die Situation, wenn f(P) einige Minima hat: Eine beliebige lokale Suchmethode führt zu einem der Punkte, in denen das Minimum erreicht wird. Dafür, daß man das absolute (globale) Minimum nicht ausläßt, ist es günstig, wenn man vorher mit der einfachsten Suche den gesamten Würfel sondiert und erst dann zumn gerichteten Suchen übergeht, wobei man als Anfangspunkte einige der besten Punkte wählt, die während der einfachsten Suche gefunden wurden.

Die zweite Aufgabenklasse, für die die einfachste Suche mit gerichteten Suchmethoden konkurrieren kann, sind Aufgaben, in denen gefordert ist, die Maxima und (oder) Minima einiger Funktionen abzuschätzen. Die einfachste Suche gestattet es, alle diese Größen mit ein und denselben Versuchspunkten abzuschätzen, indem in jedem Punkt P_i alle uns interessierenden Funktionen berechnet werden. Eine solche Situation ist typisch für Aufgaben mit mehreren Kriterien, die im nächsten Paragraphen betrachtet werden.

Abschätzung des Fehlers der einfachsten Suche. Wir lassen zu, daß die Funktion f(P) folgender Bedingung genügt: Für beliebige zwei Punkte P und P' aus K^n gilt

$$|f(P)-f(P')| \leq L\rho(P,P') , \qquad (62)$$

wobei $\rho(P,P') = ((x_1-x_n')^2+\ldots+(x_n-x_n')^2)^{1/2}$ der Abstand zwischen diesen Punkten darstellt und L eine Konstante ist. Dann kann man für die Abschätzung des Fehlers $|f_N^*-f^*)$ die Streuung der

Versuchspunkte $d=d(P_1,\ldots,P_N)$ verwenden. Indem man den zu P^* nächstgelegenen Versuchtspunkt P_k wählt, kann man schreiben, daß $f(P_k)-f(P^*)<Ld$ und umso mehr

$$f_N^*-f^* \leq Ld \tag{63}$$

gilt. Hieraus drängen sich sofort zwei seltsame Schlüsse auf. Erstens: Da die bestmögliche Ordnung von d gleich $N^{-1/n}$ (s. (39)) ist, stellt sich die Konvergenzgeschwindigkeit der einfachsten Suche als außerordentlich schlecht heraus. Zweitens, da für kubische Gitter die Ordnung von d gleich $N^{-1/n}$ (s. (43)) ist, erweisen sich kubische Gitter (der Ordnung nach) als -optimal. Das ist nun ganz und gar ein paradoxer Schluß, weil wir wissen, daß der Gleichmäßigkeit der Lage der Punkte nach kubische Gitter ganz klar den Anfangsabschnitten von LP_τ-Folgen unterlegen sind.
Die Sache ist darin begründet, daß d zu groß und seine Ordnung "die beste" sowohl für gute als auch für schlechte Gitter ist. Es kann kein Kriterium für die Qualität des Gitters sein. Aber wie kann man dann die Fehler (63) abschätzen?

Vektorcharakteristik der Streuung. Wir betrachten den Fall, wenn $f(P)=f(x_1,\ldots,x_n)$ in Wirklichkeit nur von den Veränderlichen $x_{i_1},\ldots,x_{i_s}$ abhängt. Ohne Einschränkung der Allgemeinheit können wir annehmen, daß $i_1,\ldots,i_s$ den Bedingungen (52) genügen, so daß $f(P)\equiv f(P_\beta)$ ist. In diesem Falle ist $f(P)-f(P')=f(P_\beta)-f(P'_\beta)$, und die Einschränkung (62) zieht eine noch schärfere Einschränkung nach sich:

$$|f(P)-f(P')| \leq L\rho(P_\beta,P'_\beta)\ . \tag{64}$$

Sie ist daher schärfer, weil der Abstand zwischen den Projektionen niemals den Abstand zwischen den Punkten selbst übersteigt. Es ist klar, daß in der betrachteten Situation der Fehler $f_N^*-f^*$ nicht von der Streuung der Punkte $P_1,\ldots,P_N$ an sich, sondern von der Streuung der Projektionen dieser Punkte auf K_β abhängt. Bezeichnet man diese Streuung mit d_β, so erhält man aus (64) eine noch genauere Abschätzung als (63):

$$f_N^*-f^* \leq Ld_\beta\ . \tag{65}$$

In der Eigenschaft der Kriterien für die Qualität eines beliebigen Gitters *) wird vorgeschlagen, die Gesamtheit aller $d\beta$ zu nehmen, zu ihrer Zahl gehört auch $d=d(1,2,\ldots,n)$. Wenn man sich der Ungleichungen (39) im s-dimensionalen Falle bedient, kann man schreiben:

$$(\omega_s N)^{-1/s} \leq d_\beta \leq 2\sqrt{s}\ (D^\beta/N)^{1/s} \ . \qquad (66)$$

Der Übergang zur Vektorcharakteristik der Streuung gestattet es, die seltsamen Schlüsse zu erklären, die aus (63) erhalten wurden. Beginnen wir mit dem zweiten. Im oben angemerkten Artikel ist bewiesen, daß, wenn die Punkte $P_0,P_1,\ldots,P_{N-1}$ ein π_τ-Gitter in K^n bilden, bei beliebigen $i_1,\ldots,i_s$, die (52) genügen, gilt:

$$d_\beta \leq b(\tau,s)N^{-1/s} \ .$$

Folglich sind alle $d\beta$ optimal der Ordnung nach. Wenn man für die Suche die Punkte einer LP_τ-Folge $Q_0,Q_1,\ldots,Q_i,\ldots$ und, beginnend mit $N=2^m$, die Anzahl der Versuchspunkte verdoppelt, so daß $N=2^m,2^{m+1},2^{m+2},\ldots$ ist, so wird das Gitter jedesmal ein P_τ-Gitter, und alle $d\beta$ werden ständig optimal der Ordnung nach sein.

Auf der anderen Seite bilden die Projektionen der Punkte eines kubischen Gitters, das aus $N=M^n$ Punkten besteht, auf K_β wieder ein kubisches Gitter, das allerdings aus nur M^s Punkten besteht. Da der Durchmesser jeder der M^s kleinen s-dimensionalen Würfel $\sqrt{s}/M$ ist, gilt

$$d_\beta = (\sqrt{s}/2)N^{-1/n} \ . \qquad (67)$$

Die Ordnung der Formel (67) hängt nicht von s ab und wird bei kleinen s wesentlich schlechter sein als die beste Ordnung $N^{-1/s}$.

Jetzt wenden wir uns dem Schluß zu, daß die bestmögliche Konvergenzordnung in (63) so schlecht ist, daß es schiene, daß die einfachste Suche bereits bei $n\sim6$ keine Chance mehr hat. In

---

*) Siehe: I.M. Sobol. Über die Abschätzung der Genauigkeit der einfachsten mehrdimensionalen Suche, Vortr. AdW der UdSSR, 1982, 266, No. 3, 569-572.

Wirklichkeit wird aber dieser pessimistische Schluß von der langjährigen Praxis der Nutzung der LP-Suche widerlegt.
Hier ist eine der dafür möglichen Erklärungen. Funktionen f von n Argumenten, die in der Praxis angetroffen werden, hängen zwar von allen Argumenten ab, aber nicht von allen in gleichem Maße. Umgekehrt gibt es meistens eine Gruppe von "führenden" Veränderlichen, von denen die Funktion viel stärker als von allen anderen Veränderlichen abhängt. In einem solchen Falle wird die Genauigkeit der Suche von df* nicht so sehr von der Größe d als vielmehr von der Größe des entsprechenden $d\beta$ bestimmt, dessen Ordnung $N^{-1/s}$ sich als wesentlich besser erweist als $N^{-1/n}$. So ist es zum Beispiel in dem Falle, wenn die zu optimierende Funktion $f(x_1,\ldots,x_n)$ eine Darstellung in der Gestalt der Summe

$$f = g(x_1,\ldots,x_n) + h(x_{i_1},\ldots,x_{i_s})$$

mit $s<n$, aber $h \gg g$ zuläßt. Leider ist es meistens nicht möglich, von vornherein die Möglichkeit einer solchen Darstellung zu überprüfen.
Alle diese Überlegungen beziehen sich nicht auf speziell ausgedachte Testfunktionen, deren Abhängigkeit von $x_1,\ldots,x_n$ völlig beliebig sein kann.

§ 4 Über die Optimierung nach vielen Kriterien

Aufgaben der Projektierung von Maschinen. Jedem ist klar, daß die Aufgaben der Projektierung von Maschinen viele Kriterien umfassen: Es ist wünschenswert, die Selbstkosten zu verringern, die Haltbarkeit zu erhöhen, den Metallverbrauch zu senken, den Wirkungsgrad zu vergrößern, den Energieverbrauch zu senken usw.
Als es noch keine Computer gab, wählten die Konstrukteure ein Projekt für die Realisierung mit Berücksichtigung aller Kriterien, wobei sie ihr Wissen, ihre Erfahrungen, Intuition und die Ratschläge von Experten einbrachten. Mit dem Erscheinen von Computern entstand natürlich der Gedanke an ihre Nutzung, und es erschien der Wunsch, die "beste" Variante der zu projektierenden Maschine auszuwählen. Und hier, so sagt man, haben die Mathema-

---

*) Es ist am Platze, sich an die lustige Aussage von A. Einstein zu erinnern: "Der Herrgott ist raffiniert, aber nicht böswillig gesinnt."

tiker, ohne es zu wollen, die Konstrukteure irritiert: Sie haben herausgefunden, daß (im allgemeinen Fall) die Aufgabe über das Finden der optimalen Variante nur dann eine eindeutige Lösung hat, wenn ein Kriterium optimiert wird; im entgegengesetzen Fall kann man nur über die sogenannte Menge von Pareto sprechen - die Menge der Varianten, die keine Verbeserung nach allen Kriterien gleichzeitig zulassen. Es begannen die Versuche, die Aufgaben mit mehreren Kriterien auf solche mit einem Kriterium zurückzuführen: Eine ganze Wissenschaft entstand ... .
Ich versichere nicht, daß alles genauso geschehen ist. Aber in der überwiegenden Mehrzahl der Arbeiten, die der Optimierung von Maschinen und Konstruktionen gewidmet sind, werden diese Aufgaben als solche mit einem Kriterium formuliert; die auf diesem Wege erhaltenen Lösungen befriedigen in der Regel die Konstrukteure nicht. Vor vergleichsweise kurzer Zeit versuchte FH. Ashley (1982) zu analysieren, was die Optimierung der Luftfahrtindustrie der USA brachte und gelangt dabei zu einem für viele unerwarteten Schluß: Fast immer wurden nicht diejenigen Projekte realisiert, die in Optimierungsdrechnungen erhalten wurden. Offensichtlich war das der Preis für die ungenügend begründete Aufgabenstellung der Optimierungsaufgaben.
Die Anzahl der Arbeiten, in denen die Autoren bewußt die a priori Definition eines einzigen (lösenden, globalen) Kriteriums ablehnten und alle oben angemerkten Mengen von Pareto ausrechnen, ist recht klein. Dabei wird die endgültige Wahl der "besten" Variante gewöhnlich den Konstrukteuren überlassen.

Als einfachste Methode der näherungsweisen Konstruktion der Menge von Pareto kann man die LP-Suche ansehen. Tatsächlich, wenn man das Gebeit der möglichen Lösungen im mehrdimensionalen Parameterraum nicht gleichmäßig aufgeteilten Versuchspunkten ausfüllt, in jedem Punkt das projektierte System und die Werte aller Kriterien berechnet und danach alle nichteffektiven Punkte ausschließt (ein Punkt heißt nichteffektiv, wenn ein anderer Punkt existiert, in dem die Werte aller Kriterien nicht schlechter als im betrachteten sind und mindestens ein Kriterium streng besser ist), so dienen die verbleibenden Punkte als Näherung zur Menge von Pareto, die aus allen effektiven Punkten besteht.

Die weitere Entwicklung dieser Methodik führte zur Schaffung der interaktiven Methode [3], die man Methode der Erforschung des Parameterraumes nennt (viele nennen sie übrigens ebenfalls LP-Suche). Sie fand eine weite Verbreitung unter den Maschinenkonstrukteuren; dem Autor sind nicht weniger als 30 Arbeiten bekannt, die mit Hilfe dieser Methode ausgeführt wurden. Da ihr eine spezielle Ausgabe der gegenwärtigen Serie [4] gewidmet ist, so beschränken wir uns nur auf eine schematische Beschreibung.

Die Untersuchung des Parameterraumes. Wir setzen voraus, daß ein hinreichend stabiles mathematisches Modell vorhanden ist, das es gestattet, alle die Konstrukteure interessierenden Charakteristiken der zu projektierenden Maschine zu berechnen. Das Modell hängt von n Parametern ab, die auf bestmögliche Art und Weise auszuwählen sind. Diese Parameter können sein: die Mase, die Steifigkeit, die Abmessungen der Einzelteile, die Spielräume, u.a.. Wir bezeichnen alle Parameter mit $\alpha_1,\ldots,\alpha_n$ und werden sie als einen Punkt $A(\alpha_1,\ldots,\alpha_n)$ im n-dimensionalen Parameterraum ansehen.

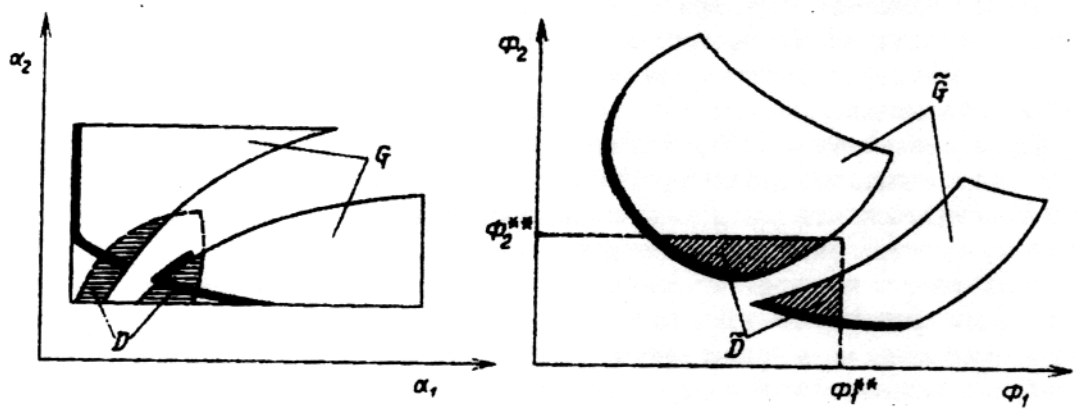

Abb. 24

In der Regel ist die Menge der möglichen Lösungen G direkt oder indirekt bestimmt, und es ist eine Menge von Kriterien $\Phi_1(A),\ldots,\Phi_k(A)$ gegeben, die für $A \in G$ definiert sind. Der Einfachheit halber werden wir annehmen, daß es wünschenswert ist, jedes dieser Kriterien zu verringern.

Die Punkte $Q_1,Q_2,\ldots$ werden in Versuchspunkte $A_1,A_2,\ldots,A_N$ transformiert, die in G gleichverteilt sind. In jedem Punkt A_i

wird das Modell der Maschine und alle $\phi_\nu(A_i)$ berechnet. Für jedes Kriterium wird eine Versuchstabelle erstellt, das Analogon einer Variationsreihe in der Statistik, mitderen Hilfe der Konstrukteur die Kriterienbeschränkung ϕ_ν^{**} auswählt - die schlechtesten, aber noch annehmbaren Werte von ϕ_ν. Damit ist die Menge der zulässigen Lösungen D definiert, die aus allen solchen A besteht, für die $A \in G$ und $\phi_\nu(A) \preceq \phi_\nu^{**}$, $1 \leq \nu \leq k$ gilt.
In Abb. 24 ist der Fall n=2, k=2 abgebildet; die Menge von Pareto ist die fettgedruckte Linie, die Menge D ist gestrichelt.

Es ist notwendig zu unterstreichen, daß D keine einfache Einschränkung der Menge G ist: Alle ausgesonderten Punkte sind für den Konstrukteur unannehmbar, selbst wenn sie effektiv sind. Gewöhnlich wählt der Konstrukteur selbst die "beste" Variante, indem er die endliche Menge der zugelassenen effektiven Punkte analysiert.
Ein hinreichend ausführliches Durchsehen der Menge G gestattet ebenfalls auf die Frage zu antworten, ob die Forderungen, die an die zu projektierende Maschine gestellt werden, vereinbar sind.

Optimierung nach mehreren Kriterien in numerischen Experimenten.
Das numerische Experiment ist eine vergleichsweise neue Methode der Untersuchung komplizierter Aufgaben der Physik und Technik [5]. Die Perspektiven seiner Entwicklung und seine Möglichkeiten sind unabsehbar, da auch jede Verallgemeinerung der Computer diese Möglichkeiten erweitert.
Auf einer Etappe des numerischen Experimentes muß die Aufgabe über die Einstellung (oder Kalibrierung) des mathematischen Modells gelöst werden. Zum Beispiel enthält ein Gleichungssystem, das den zu untersuchenden Prozeß modelliert, oft unbestimmte Parameter, die man so auswählen muß, daß die Werte $y_1,\ldots,y_k$, die nach dem mathematischen Modell berechnet werden, gut mit den vorhandenen experimentellen Daten $y_1^o,\ldots,y_k^o$ übereinstimmen. Diese Aufgabe führt gewöhnlich zur Auffindung des Minimums einer Funktion der Art

$$\phi = \sum_{\nu=1}^{k} a_\nu (y_\nu - y_\nu^o)^2 ,$$

wobei a_ν positive Gewichte sind.

Wenn eine solche Variante von Parametern existieren würde, so daß für alle $y_\nu=y_\nu^0$ gilt, so wäre eine ähnliche Aufgabenstellung völlig begründet. Aber in der Mehrheit der realen Aufgaben gibt es eine solche Variante von Parametern nicht, und das absolute Minimum von ϕ ist positiv. In diesem Falle ist es viel interessanter, die Aufgbe als eine mit mehreren Kriterien anzusehen, in der es gefordert ist, k Kriterien zu minimieren:

$$\phi_\nu = |y_\nu - y_\nu^0| , \quad 1 \leq \nu \leq l .$$

(Möglich sind dazwischenliegende Varianten der Aufgabenstellung, wenn Kriterien eingeführt werden, die gleich einige y_ν umfassen.) Die Herangehensweise mit mehreren Kriterien gestattet es, das Anwendungsgebiet des Modells zu klären, auf seine schwachen Stellen hinzuweisen und Wege zu seiner Vervollkommnung anzugeben.

Vorerst wird die Methode der Untersuchung des Parameterraumes (LP-Suche) bei der Lösung solcher Aufgaben selten verwendet. Das dafür interessanteste Beispiel ist die Arbeit von N.W. Lukaschewa, N.S. Milewskaya und A.M. Jeljaschewitsch (1981), in der ein effektiver Algorithmus zur Dekodierung der Struktur polymerer Ketten geschaffen wurde.

Allerdings beschränkt sich die Nutzung der Optimierung nach mehreren Kriterien im numerischen Experiment nicht nur auf die Einstellung des Modells, sie kann auch zur Auswahl des Modells genutzt werden.

Wie bekannt, kann für jede reale (physikalische oder technische) Aufgabe eine Hierarchie komplizierter werdender Modelle konstruiert werden: vom einfachsten, was gewöhnlich eine analytische Formel gibt und näherungsweise die wichtigsten Charakteristiken der Aufgabe verbindet, bis zum kompliziertesten, das eine Menge von Faktoren berücksichtigt, aber so kompliziert ist, daß seine Berechnung für die modernen Computer unmöglich ist (und darum eine weitere Verkomplizierung unsinnig ist). Welches der Modelle sollte man für die Lösung der konkreten Aufgabe wählen? Man kann mit der Betrachtung eines der einfachsten Modelle beginnen: Wenn die Untersuchung des Parameterraumes die Unmöglichkeit einer befriedigenden Einstellung des Modells zeigt, so sollte man zu einem komplizierteren System übergehen. Dabei ist klar, daß man ohne ausführliche Betrachtung des Parameterraumes diese Unmöglichkeit der Einstellung nicht feststellen kann.

Literatur

1. **Sobol, I.M.**, Mehrdimensionale Quadraturformeln und die Funktionen von Haare. M., Nauka, 1969 (russisch)

2. **Kuipers, L, Niederreiter, M.**, Uniform distribution of sequences, New York, Wiley, 1974

3. **Sobol, I.M., Statnikow, R.B.**, Auswahl optimaler Parameter in Aufgaben mit mehreren Kriterien. M., Nauka, 1981 (russisch)

4. **Samarskij, A.A.**, Numerisches Experiment in Aufgaben der Technologie - Westnik AN SSSR, 1984, № 3, 77-86 (russisch)