Studia Scientiarum Mathematicarum Hungarica

AUXILIO
CONSILII INSTITUTI MATHEMATICI
ACADEMIAE SCIENTIARUM HUNGARICAE

REDIGIT
L. FEJES TÓTH

ADIUVANTIBUS
Á. CSÁSZÁR, I. CSISZÁR, A. HAJNAL, E. MAKAI,
P. RÉVÉSZ, O. STEINFELD, T. E. SCHMIDT,
J. SZABADOS, P. TURÁN, I. VINCZE

OMUS X.
ASC. 1–2.
975

AKADÉMIAI KIADÓ, BUDAPEST

Studia Scientiarum Mathematicarum Hungarica

A Magyar Tudományos Akadémia matematikai folyóirata

Szerkesztőség: 1061 Budapest V., Reáltanoda u. 13—15.

Technikai szerkesztő: Deák E.

Kiadja az Akadémiai Kiadó, 1053 Budapest V., Alkotmány u. 21.

A *Studia Scientiarum Mathematicarum Hungarica* angol, német, francia vagy orosz nyelven közöl eredeti értekezéseket a matematika tárgyköréből. Félévenként jelenik meg, évi egy kötetben.

Előfizetési ára belföldre 120,— Ft, külföldre 165,— Ft. Megrendelhető a belföld számára az Akadémiai Kiadónál, a külföld számára pedig a Kultúra Könyv és Hírlap Külkereskedelmi Vállalatnál (1011 Budapest II., Fő u. 32.).

Cserekapcsolatok felvétele ügyében kérjük a MTA Matematikai Kutató Intézete Könyvtárához (1053 Budapest V., Reáltanoda u. 13—15) fordulni.

Közlésre szánt dolgozatokat kérjük két példányban a szerkesztőség címére küldeni.

Studia Scientiarum Mathematicarum Hungarica is a journal of the Hungarian Academy of Sciences publishing original papers on mathematics, in English, German, French or Russian.

It is published semiannually, making up one volume per year.

Editorial Office: 1053 Budapest V., Reáltanoda u. 13—15, Hungary.
Technical Editor: E. Deák

Subseription rate: $ 16.00 per volume. Orders may be placed with *Kultúra* Trading Co. for Books and Newspapers, Budapest 62, P. O. B. 149 or with its representative sabroad.

For establishing exchange relations please write to the Library of the Mathematical Institute (1053 Budapest V., Reáltanoda u. 13—15.)

Papers intended for publication should be sent to the Editor in 2 copies.

Studia Scientiarum Mathematicarum Hungarica

Auxilio
Consilii Instituti Mathematici
Academiae Scientiarum Hungaricae

Redigit

L. Fejes Tóth

Adiuvantibus

Á. Császár, I. Csiszár, A. Hajnal, E. Makai, P. Révész, O. Steinfeld, T. E. Schmidt, J. Szabados, P. Turán, I. Vincze

Tomus X

Akadémiai Kiadó, Budapest
1975

STUDIA SCIENTIARUM MATHEMATICARUM HUNGARICA

Tomus X

INDEX

78-94 — Szegedi Nyomda — F. v.: Dobó József igazgató

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 1—18.

EIN VERFAHREN ZUR BERECHNUNG DER LÖSUNG MIT SINGULÄREM VERHALTEN BEI DIFFERENTIALGLEICHUNGEN ERSTER ORDNUNG

von

S. FRIVALDSZKY

I. TEIL

1. Einführung

Eine häufig benutzte Methode zur Lösung von Anfangswertaufgaben ist die abschnittsweise Annäherung der Lösung durch Polynome, [1]. In gewissen Fällen bekommen wir jedoch ein besseres Resultat, wenn wir eine rationale, gebrochene Funktion anstatt eines Polynoms verwenden, [2]. Diese Methode kann zu einem guten Ergebnis auch dann führen, wenn die Lösung auf dem untersuchten Abschnitt eine Funktion mit singulärem Verhalten wird.

Werden die Lösungfunktion oder eine der Ableitungen niedrigerer Ordnung von dieser auf dem untersuchten Abschnitt unbegrenzt, so kann die Lösungfunktion abschnittsweise mit dem Ansatz

$$y^*(x) = \sum_{p=0}^{L} a_p x^p + b|x+A|^N, \quad x \neq -A$$
$$N \notin \{0, 1, \ldots, L\}$$

oder

$$y^{**}(x) = \sum_{p=0}^{L} a_p x^p + b|x+A|^N \log|x+A|, \quad x \neq -A$$
$$N \in \{0, 1, \ldots, L\}$$

angenähert werden, wo $-A$ (die Stelle der Singularität) und N (der Grad der Singularität) bekannt sind. Die Konstanten a_p $(p=0, 1, \ldots, L)$, b werden auf dem betreffenden Abschnitt so gewählt, dass der Ansatz genau sein soll, wenn die Lösungfunktion ein einen gewissen Grad nicht übersteigendes Polynom oder der Ansatz selbst ist, [3].

Der Fehler in einem Schritt kann mit einem Integralausdruck aufgeschrieben werden. Haben die sogenannten Effektfunktionen unter dem Integralzeichen dauernd gleiche Vorzeichen, so hat dieser Fehler die Form

$$\bar{C}_{r+1} h^{r+1} \left\{ y^{(r+1)}(\xi_1) - (r-N) \frac{|A+\eta_1|^{r-N}}{|A+\xi_2|^{r-N+1}} y^{(r)}(\eta_2) \right\},$$

wo $(r+1)$ die Anzahl der Stützpunkte, $\bar{C}_{r+1}$ eine von der Lösungfunktion und der Schrittweite h unabhängige Konstante und

$$\begin{matrix} \xi_i \in [x_n - rh, x_n] \\ \eta_i \in [x_n - rh, x_n - h] \end{matrix} \quad (i = 1, 2)$$

sind.

Die Verfasser haben ihre Methode für den Fall verallgemeinert, dass der Ansatz die Form

$$y(x) = \sum_{p=0}^{L} a_p x^p + b\pi(x), \quad x \neq -A$$

hat, wo die Funktion $\pi(x)$ im Punkt $x=-A$ von singulärem Verhalten ist, [4]. Die Methode setzt voraus, dass die Funktion $\pi(x)$ bekannt sei.

Im Fall einer echten Singularität ($N<0$) und des genauen Ansatzes kann man bessere obere Schranken als $O(1/h^{-N})$ für diesen Fehlerausdruck nicht geben, vorausgesetzt, dass der Punkt x_n dem Punkt der Singularität $x=-A$ nahe liegt. Diese obere Schranke ist mit der Lösungfunktion von gleicher Grössenordnung, und so können diese Resultate nicht einmal entscheiden, ob der relative Fehler in einem Schritt genügend klein ist.

Der Artikel [3] gibt eine Annäherung in jedem Schritt der Berechnung für die Werte A und N derart, dass die zwei nicht verschwindenden Glieder niedrigsten Grades in der im betreffenden Schritt für den Fehler und der Schrittweite nach aufgeschriebenen Taylor—Reihe gleich Null gesetzt werden. Kann man voraussetzen, dass diese zwei, während der Berechnung für die Werte A und N bekommenen Zahlenreihen zu je einem Grenzwert konvergieren, so ist auch eine zweite Berechnung mit diesen Grenzwerten durchzuführen. Ähnlich verläuft das Verfahren, wenn eine der Zahlen A und N bekannt ist.

Die Anwendbarkeit dieser Methode ist nicht hinreichend begründet, da diese zwei Glieder in der Taylor—Reihe nicht unbedingt den überwiegenden Teil des Fehlers enthalten; die Koeffizienten dieser Reihe hängen ja von einer Ableitung höherer Ordnung der Lösungfunktion ab, demnach sind sie in der Nähe der Singularität nicht begrenzt.

In den in diesen Artikeln untersuchten numerischen Beispielen hat der singuläre Teil der Lösungfunktion die Form

$$b|x+A|^{-1}$$

d.h. er ist von der Art eines Pols erster Ordnung, und der Koeffizient des Glieds singulären Verhaltens ist konstant. In diesem Fall ist zu hoffen, dass diese Berechnung eine gute Annäherung für die Lösungfunktion gibt. In anderen Fällen ist es aber zweifelhalft, ob ein gutes Resultat erzielt werden kann.

Aus diesem Grund suchen wir einen anderen Weg.

2. Das Differenzenverfahren bei der Lösung mit singulärem Verhalten

Wir suchen die Lösung der Anfangswertaufgabe

$$y' = f(x, y), \quad y(x_0) = y_0$$

in Punkten

$$x_n = x_0 + nh, \quad n = 1, 2, \ldots, \quad h > 0$$

und es seien y_n der Näherungswert für den Wert $y(x_n)$ und

$$f_n = f(x_n, y_n).$$

An Stelle von Bedingungen für die Funktion $f(x, y)$ ist vorausgesetzt, dass die Lösung in der Form

$$y(x) = \frac{u(x)}{(s-x)^p}, \quad x_0 < x < s$$
$$p \notin \{0, -1, -2, \ldots\} \tag{2.1}$$

aufgeschrieben werden kann, wo $u(x) \in C_{r+1}[x_0, s]$ — $(r+1)$-mal stetig differenzierbar — und r eine positive ganze Zahl sind. Hier haben wir

$$f(x, y(x)) = \frac{u'(x)(s-x)+u(x)p}{(s-x)^{p+1}} = \frac{v(x)}{(s-x)^{p+1}} \tag{2.2}$$

und

$$y_{n+1} = y_{n-t} + \int_{x_{n-t}}^{x_{n+1}} f(x, y(x))\,dx = y_{n-t} + \int_{x_{n-t}}^{x_{n+1}} \frac{v(x)}{(s-x)^{p+1}}\,dx$$
$$t \in \{0, 1, \ldots, r\},$$

vorausgesetzt $y_n = y(x_n)$, wo $v(x) \in C_r[x_0, s]$ ist.

Auf dem Intervall des Integrals setzen wir ein interpolierendes Polynom mit den Stützstellen $x_n, x_{n-1}, \ldots, x_{n-r}$ an Stelle der Funktion $v(x)$, und im Integral führen wir die neue Veränderliche

$$x = x_n + hz$$

ein. Durch Anwendung des Zusammenhangs (2.2) bekommen wir die Formel

$$y_{n+1} = y_{n-t} + h \sum_{j=0}^{r} B_j^{(t)}(b) f_{n-j}, \tag{2.3}$$

wo die Koeffizienten

$$B_j^{(t)}(b) = \frac{1}{r!}(-1)^j \binom{r}{j} (b+j)^{p+1} \int_{-t}^{1} \frac{\prod_{\substack{i=0 \\ i \neq j}}^{r} (z+i)}{(b-z)^{p+1}}\,dz$$
$$b = \frac{s-x_n}{h} \tag{2.4}$$

von der Distanz von der Singularität abhängen und im Fall $b \to \infty$ in die Koeffizienten des gewöhnlichen Differenzverfahrens (Adams-Verfahren bei $t=0$) übergehen.

Zweckmässig ist, diese Koeffizienten bei $b=2, 3, 4, \ldots$ und bei einigen Werten von p und r auszurechnen. Aus diesen können die momentanen Koeffizienten während der Berechnung mit Hilfe linearer Interpolation bekommen werden. Im allgemeinen genügt es, sie bei etlichemal zehn Werten von b vorher zu bestimmen, fern von der Singularität kann ja irgendwelches gewöhnliche Differenzverfahren gebraucht werden.

Sind $p \leqq r-1$ eine positive ganze Zahl und $b > 1$, so gilt der folgende Zusammenhang für die Nachprüfung der richtigen Berechnung der Koeffizienten:

$$\sum_{j=0}^{r} B_j^{(t)}(b) = t+1,$$

was leicht zu beweisen ist. (Der Zusammenhang gilt für andere Werte von p im allgemeinen nicht.)

Im allgemeinen ist nicht zu hoffen, dass das Produkt $h(\partial f/\partial y)$ der Lösung entlang nahe der Singularität genügend klein sei. Deshalb sind die impliziten, der Formel (2.3) ähnlichen Formeln in diesem Fall nicht anzuwenden, weil ja die Iteration verbunden mit der Berechnung dieser Formeln nicht konvergiert. Ähnlich ist es nicht zweckmässig, sich hier mit dem Problem der Stabilität zu beschäftigen, [5]. Wäre dieses Produkt doch klein genug, so könnten wir die ähnlichen impliziten Formeln in der Form

$$y_{n+1} = y_{n-t} + \frac{h}{r!}\sum_{j=0}^{r}(-1)^j\binom{r}{j}(b-1+j)^{p+1} \tag{2.5}$$
$$\int_{-t}^{1}\frac{\prod_{\substack{i=0\\ i\neq j}}^{r}(z+i-1)}{(b-z)^{p+1}}\,dz\,f_{n+1-j}$$

anwenden. Man kann aber unmittelbar nicht überprüfen, ob diese Bedingung erfüllt ist.

Im zweiten Teil versuchen wir dennoch, die impliziten Formeln allgemein anzuwenden.

3. Der Fehler des Verfahrens in einem Schritt

Von der Formel (2.3) ausgehend schätzen wir den Fehler des Verfahrens in einem Schritt ab. Verwendet man die genauen Werte der Lösungfunktion unter (2.3), so ergibt die Differenz der linken und der rechten Seite unter (2.3) diesen Fehler. Zuerst ersetzen wir die Werte y_{n+1}, y_{n-t} bzw. f_{n-j} $(j=0, 1, \ldots, r)$ in dieser Differenz durch den Ausdruck (2.1) bzw. (2.2), dann die Werte u_{n-l} $(l=-1, 1, 2, \ldots, r)$ und u'_{n-l} $(l=1, 2, \ldots, r)$ durch ihre Taylor-Entwicklung nach den Potenzen der Schrittweite h. Diese Taylor-Formeln werden an der Stelle x_n entwickelt, mit dem Integralrestglied (der Veränderlichen w) benutzt, bis zur m-ten Potenz der Schrittweite h bei den Werten u_{n-l} und bis zur $(m-1)$-ten bei den u'_{n-l} genommen, $u(x)\in C_{m+1}[x_0, s]$, $(m\geqq 1)$ vorausgesetzt.

Jetzt werden der Zusammenhang

$$p\int_{-t}^{1}\frac{z^k}{(b-z)^{p+1}}\,dz = \frac{1}{(b-1)^p} - \frac{(-1)^k t^k}{(b+t)^p} - k\int_{-t}^{1}\frac{z^{k-1}}{(b-z)^p}\,dz$$
$$(k\geqq 1)$$

und das folgende Lemma verwendet, dessen Beweis wir dem Leser überlassen:

$$\delta_{0,k}\frac{1}{z} + \sum_{j=1}^{r}(-1)^r\binom{r}{j}\frac{j^k}{z+j} =$$
$$= (-1)^k r!\left(\frac{z^k}{\prod_{i=0}^{r}(z+i)} - \delta_{k,r+1}\right) \qquad \begin{array}{l} 0\leqq k\leqq r+1 \\ z\neq 0,-1,\ldots,-r. \end{array}$$

Ausserdem wird die Substitution

$$w = x_n + hz$$

in den Integralen durchgeführt. Wir erhalten so zwei Ausdrücke für den Fehler:

$$T_n(b) = u_n^{(r+1)} \frac{h^{r+1-p}}{(r+1)!} E_{r+1}(b) + I_{r+1}(b) \tag{3.1}$$

wenn $m = r+1$, $u(x) \in C_{r+2}[x_0, s]$ und

$$T_n(b) = I_r(b) \tag{3.2}$$

wenn $m = r$, $u(x) \in C_{r+1}[x_0, s]$ wo

$$E_{r+1}(b) = -(r+1-p) \int_{-t}^{1} \frac{\prod_{i=0}^{r} (z+i)}{(b-z)^{p+1}} dz$$

und

$$I_m(b) = \frac{h^{m+1-p}}{m!} \int_{-r}^{1} G_m(b, z) u^{(m+1)}(x_n + hz)\, dz\,.$$

Hier ist

$$G_m(b, z) = \frac{(\overline{1-z})^m}{(b-1)^p} + \frac{(\overline{-t-z})^m}{(b+t)^p} + \sum_{j=1}^{r} \frac{B_j^{(t)}(b)}{(b+j)^{p+1}} \cdot$$

$$\left[p(\overline{-j-z})^m + m(b+j)(\overline{-j-z})^{m-1}\right]$$

wo

$$(\overline{s-z}) = \begin{cases} s-z \begin{cases} s \leqq z \leqq 0 & \text{und} \quad s \leqq 0, \\ 1 \geqq z \geqq 0 & \text{und} \quad s > 0, \end{cases} \\ 0 \quad \text{sonst.} \end{cases}$$

Aus dem Ausdruck (3.2) ersieht man unmittelbar, dass

$$T_n(b) = O(h^{r+1-p})$$

— d.h. von der Grössenordnung $(r+1-p)$ ist, wenn $u(x) \in C_{r+1}[x_0, s]$.

Der Fehler in einem Schritt kann genügend klein werden, wenn $r+1 > p$ und die Schrittweite h genügend klein sind.

Hat die Effektfunktion $G_r(b, z)$ im Intervall (x_{n-r}, x_{n+1}) dauernd dasselbe Vorzeichen und ist $u(x) \in C_{r+1}[x_0, s]$, so ist

$$I_r(b) = \frac{h^{r+1-p}}{r!} u^{(r+1)}(\xi) \int_{-r}^{1} G_r(b, z)\, dz$$

aufzuschreiben und der Fehler erhält, von der Formel (3.2) ausgehend und das vorige Lemma verwendend, die neue Form

$$T_n(b) = u^{(r+1)}(\xi) \frac{h^{r+1-p}}{(r+1)!} E_{r+1}(b), \tag{3.3}$$

wo

$$\xi \in (x_{n-r}, x_{n+1}).$$

Wechselt die Effektfunktion ihr Vorzeichen auf (x_{n-r}, x_{n+1}) und ist $u(x) \in C_{r+2}[x_0, s]$, so verwenden wir die Formel (3.1) für die Fehlerabschätzung, während das Restglied weggelassen wird.

In beiden Fällen haben wir einen anwendbaren, der Formel (3.3) ähnlichen Zusammenhang für den Fehler, wenn es eine gute Abschätzung für den Wert $u^{(r+1)}(\xi)$ gibt.

Unabhängig von dieser ist der Fehler gut abzuschätzen, wenn sich die Funktion $u^{(r+1)}(x)$ in der Umgebung der Stelle x_n nur langsam ändert. Das Verfahren kann nämlich mit einer halben Schrittlänge zweimal nacheinander von der Stelle x_n ausgehend durchgeführt werden. Bekommen wir so den Wert $\bar{y}_{n+1}$, so ist näherungsweise

$$T_n(b) = \frac{\bar{y}_{n+1} - y_{n+1}}{1 - \dfrac{E_{r+1}(2b) + E_{r+1}(2b-1)}{2^{r+1-p} E_{r+1}(b)}}.$$

Unter ähnlichen Bedingungen ist die Fehlerabschätzung für die implizite Formel (2.5) zu gewinnen:

$$T_n^{imp}(b) = u^{(r+1)}(\xi) \frac{h^{r+1-p}}{(r+1)!} \left[-(r+1-p) \int_{-t}^{1} \frac{\prod_{i=0}^{r} (z+i-1)}{(b-z)^{p+1}} \, dz \right].$$

4. Verallgemeinerung des Verfahrens im Kapitel 2

Man kann die Formel (2.3) verallgemeinern. Wir drücken den Wert y_{n+1} in der linearen Form

$$y_{n+1} = \sum_{j=0}^{r} A_j^*(b) y_{n-j} + h \sum_{j=0}^{r} B_j^*(b) f_{n-j} \tag{4.1}$$

aus, und die Koeffizienten $A_j^*(b)$ und $B_j^*(b)$ $(j=0, 1, \ldots, r)$ werden so gewählt, dass die Differenz der beiden Seiten möglichst klein sei. Für den Fehler in einem Schritt wird die Differenz der linken und der rechten Seite gebildet. Ähnlich der Fehlerabschätzung im Kapitel 3. setzen wir die Formeln (2.1) und (2.2) in diesen Fehlerausdruck und auch jene, im Kapitel 3 erwähnten Taylor-Reihen anstelle der Werte der Funktionen $u(x)$ und $u'(x)$ ein. In dieser für den Fehler so bekommenen Potenzreihe nach der Schrittweite h versuchen wir einige der ersten Koeffizienten der Potenzen von h verschwinden zu lassen. Dies bedeutet lineare Bedingungen zu erfüllen. So bekommen wir für den Fehler, nach einer Ersetzung der Veränderlichen in den Integralen,

$$T_n^*(b) = u_n^{(m+1)} \frac{h^{m+1-p}}{(m+1)!} E_{m+1}^*(b) + I_{m+1}^*(b)$$

wenn $u(x) \in C_{m+2}[x_0, s]$, $m \geqq 1$.

Hier sind

$$E_k^*(b) = \frac{1}{(b-1)^p} - (-1)^k \sum_{j=1}^{r} A_j^*(b) \frac{j^k}{(b+j)^p} + (-1)^k$$

$$\sum_{j=1}^{r} B_j^*(b) \frac{j^{k-1}[k(b+j)-pj]}{(b+j)^{p+1}} - \frac{A_0^*(b)}{b^p} \delta_{0,k} - \frac{B_0^*(b)}{b^p} \left(\frac{p}{b} \delta_{0,k} + \delta_{1,k} \right)$$

$$(k = 0, 1, \ldots, m+1)$$

(δ_{ij} das Kronecker-Symbol) und

$$I_{m+1}^*(b) = \frac{h^{m+2-p}}{(m+1)!} \int_{-r}^{1} G_{m+1}^*(b, z) u^{(m+2)}(x_n + hz)\, dz,$$

wo

$$G_{m+1}^*(b, z) = \frac{\overline{(1-z)}^{m+1}}{(b-1)^p} + \sum_{j=1}^{r} A_j^*(b) \frac{\overline{(-j-z)}^{m+1}}{(b+j)^p} +$$

$$+ \sum_{j=1}^{r} \frac{B_j^*(b)}{(b+j)^{p+1}} \left[p \overline{(-j-z)}^{m+1} + (m+1)(b+j) \overline{(-j-z)}^{m} \right],$$

und das lineare Gleichungssystem

$$E_k^*(b) = 0 \quad (k = 0, 1, \ldots, m) \tag{4.2}$$

sollte durch diese Koeffizienten befriedigt werden, was bei $m \leqq 2r+1$ im allgemeinen erreichbar ist.

Das Ergebnis ist dem des Verfahrens im Kapitel 2 ähnlich, nur die Genauigkeit der Formel kann man steigern. Als Spezialfall bekommt man die Formel (2.3) und die Koeffizienten (2.4), wenn man nämlich die Gleichung (4.2) auflöst, wo $m=r$ und

$$A_j^*(b) = \delta_{j,t} \quad (j = 0, 1, \ldots, r)$$

sind. Dann wäre die besagte Lösung die einzelne des Gleichungssystems (4.2) bei $p \notin \{1, 2, \ldots, r\}$. Bei $p \in \{1, 2, \ldots, r\}$ ist die $(p+1)$-te Gleichung $(k=p)$ im Gleichungssystem (4.2) durch eine lineare Kombination der ersten p Gleichungen zu gewinnen. Nehmen wir die folgende Gleichung in diesem Fall zum Gleichungssystem (4.2) hinzu und ist das so erhaltene Gleichungssystem auflösbar, so sind die Koeffizienten eines genaueren Verfahrens (von der Genauigkeit $O(h^{r+2-p})$) als im Kapitel 2 zu bekommen. Ferner ist das Gleichungssystem (4.2) bei $m=r+1$ und

$$B_j^*(b) = 0 \quad (j = 1, 2, \ldots, r)$$

leicht auflösbar, aber das Resultat dürfte kaum interessant sein.

Bei der Formel (4.1) könnte man über die Stabilität auch dann nichts sagen, wenn das Produkt $h(\partial f/\partial y)$ der Lösung entlang klein genug wäre, weil die Koeffizienten von der Stelle abhängen. Trotzdem ist es zweckmässig, solche Formeln wie (4.1) anzuwenden, in der die Koeffizienten $A_j^*(b)$ die Gleichung der Stabilität im Grenzwert $b \to +\infty$ befriedigen, [5].

Später werden wir uns mit der Abschätzung der Werte (b, p) beschäftigen und ein numerisches Beispiel geben.

(Die Literaturhinweise beziehen sich auf das Literaturverzeichnis am Ende des II. Teils.)

II. TEIL

Einführung

Im ersten Teil haben wir uns mit einer Methode zur Berechnung der Lösung bei Differentialgleichungen erster Ordnung beschäftigt, welche Lösung auf dem untersuchten Abschnitt von singulärem Verhalten ist, was soviel bedeutet, dass sie oder eine ihrer Ableitungen niedrigerer Ordnung hier nicht begrenzt ist. Wir gingen von einem speziellen Ansatz aus, der zu einem Differenzenverfahren mit veränderlichen Koeffizienten führt. Die Fehlerabschätzung hat das Verfahren sehr gut anwendbar gezeigt.

Wir befassen uns in diesem Teil mit der Abschätzung der Daten (b, p) der Singularität, einem numerischen Beispiel, der Verallgemeinerung des Verfahrens für Differentialgleichungssysteme und der Anwendung der impliziten Formeln.

6. Die Abschätzung der Daten der Singularität

Für den relativen Abstand b von der Singularität und für den Grad p derselben kann man Zusammenhänge leicht bekommen. Durch Taylor-Reihen an einer Stelle x_m sind lineare Kombinationen aus den Werten u_{m-j}, u'_{m-j} $(j=-l', -l'+1, \ldots, \ldots, 0, 1, \ldots, l)$ so aufzuschreiben, dass sie die Null bei einer genügend kleinen Schrittweite h hinreichend annähern, vorausgesetzt, dass die Funktion $u(x)$ genügend vielmal stetig differenzierbar ist. Zwei von diesen Kombinationen geben, gleich Null gesetzt, ein Gleichungssystem für die Daten der Singularität, wenn wir den Ausdruck (2.2) des ersten Teils an Stelle der Werte u'_{m-j} und den unter (2.1) des ersten Teils an Stelle der übrigen Werte u_{m-j} einsetzen.

Einige dieser linearen Kombinationen kann man einfacher gewinnen. Angenommen $u(x) \in C_{l+1}[x_0, s]$, schreiben wir die Taylor-Reihen der Werte u_{m-j} $(j=1, 2, \ldots, l)$ an der Stelle x_m auf und bilden ihre lineare Kombination mit den Koeffizienten R_j $(j=1, 2, \ldots, l)$. Jetzt werden

$$R_j = (-1)^j \binom{l}{j} \quad (j = 1, 2, \ldots, l)$$

genommen, dann die übrig gebliebenen Werte u_{m-j} $(j=0, 1, \ldots, l)$ mit Hilfe der Formel (2.1) des ersten Teils durch die Werte y_{m-j} ersetzt. So bekommen wir

$$\begin{aligned} v_l(c, p) &= \sum_{j=0}^{l} (-1)^l \binom{l}{j} (1+cj)^p y_{m-j} = \\ &= u_m^{(l)} h^{l-p} c^p + O(h^{l+1-p}), \quad c = h/(s-x_m), \end{aligned} \tag{6.1}$$

nachdem das Lemma des ersten Teils bei $1 \leqq k \leqq l+1$ und bei dem Grenzwert $z \to 0$ verwendet worden ist.

Ähnlich werden

$$R_j = (-1)^j \binom{l}{j} \frac{1}{j} \quad (j = 1, 2, \ldots, l)$$

und $u(x)\in C_{l+2}[x_0, s]$ genommen, dann der Wert u'_m mit Hilfe von

$$u'_m = f_m \frac{h^p}{c^p} - u_m \frac{pc}{h}$$

nach der Formel (2.2) des ersten Teils, endlich die u_{m-j} ($j=0, 1, \ldots, l$) mit Hilfe von (2.1) durch die Werte y_{m-j} ersetzt:

$$\mu_l(c, p) = \sum_{j=1}^{l} (-1)^j \binom{l}{j} \frac{1}{j} (1+cj)^p y_{m-j} + y_m$$

(6.2)

$$\left(\sum_{i=1}^{l} \frac{1}{i} + pc\right) - hf_m = -\frac{1}{l+1} u_m^{(l+1)} h^{l+1-p} c^p + O(h^{l+2-p}),$$

nachdem das besagte Lemma bei $0\leqq k\leqq l+1$ und $z\to 0$ verwendet worden ist. (Bei $k=0$ haben wir das Glied $1/z$ zuerst auf die rechte Seite übergetragen.) Die zwei Gleichungen

$$\mu_l(c, p) = 0$$

(6.3)

$$\nu_l(c, p) = 0$$

bestimmen die Werte (c_m, p_m) an der Stelle x_m. Ist $u(x)\in C_{l+2}[x_0, s]$, so sind die weggelassenen Glieder in den Ausdrücken (6.1) bzw. (6.2) bei fixen Werten (c_m, p_m) von der Grössenordnung $O(h^{l-p})$ bzw. $O(h^{l+1-p})$.

Die beschriebene Methode kann man so anwenden, dass die Lösung zuerst z.B. durch die Predictor-Corrector Methode bekommen wird. Ist der Fehler in irgendwelchem Schritt zu gross, so halbieren wir die Schrittweite h. Ist sie schon zu klein, dann versuchen wir die Anwendung unserer Methode, wenn annehmbare Werte (s_m, p_m) an der betreffenden Stelle x_m aus den bisher bekommenen Werten der Lösung gewonnen werden, und die Stelle der Singularität nicht zu weit ist.

Die Berechnung mit dieser Methode fortgesetzt, bekommen wir in jedem Schritt der Berechnung der Lösung zwei Werte (s_m, p_m), wo $s_m = x_m + h/c_m$. Ist während der Berechnung anzunehmen, dass diese zwei Zahlenfolgen zu je einem Grenzwert

$$s_m \to s, \quad p_m \to p \quad (x_m < s)$$

konvergieren, so wird angenommen, dass es sich um eine Singularität von der Form (2.1) mit den Parametern (s, p) handelt. Deshalb führen wir eine zweite Berechnung mit diesen Parametern — genauer mit (b, p), wo $b=(s-x_n)/h$ — durch, wie es auch die Artikel [3], [4] empfehlen.

Wenn wir $l=r+l$ bzw. $l=r$ wählen, gibt die Formel (6.1) bzw. (6.2) eine gute Annäherung für den Wert $u_m^{(r+1)}$ im zweiten Berechnungsgang, das Glied höherer Grössenordnung weglassend. Einerseits kann man so überprüfen, ob die Werte $u_m^{(r+1)}$ als begrenzt angesehen werden können, anderseits ergibt diese Formel näheungsweise den Fehler in diesem Schritt, mit Hilfe der Formel (3.3).

Einer groben Untersuchung nach ist es ratsam

$$l \geqq r+1-p$$

zu wählen, um den durch die ungenauen Werte (b, p) bekommenen Fehler mit dem Fehler des Differenzenformel verglichen klein zu halten. Deshalb sei

$$l = \begin{cases} r & \text{wenn} \quad p > 1, \\ r+1 & \text{wenn} \quad 0 < p < 1. \end{cases}$$

7. Ein numerisches Beispiel

Die folgende Anfangswertaufgabe hat eine Singularität $p=0{,}5$ an der Stelle $s=0$. Hier sind $x_0=-1$, $h=0{,}025$.

Wir fingen von der Stelle $x_n=-0{,}300$ ausgehend an. In der angeführten Tabelle enthält die erste Spalte die Werte der Abszisse x_n, die anderen Spalten enthalten der Reihe nach die genauen Werte der Lösung, die Resultate des Verfahrens von Adams, von Lambert-Shaw bei $A=0$, $N=-0{,}5$ als zweiter Berechnungsgang, [3], [4]. Wir verwendeten unser Verfahren (2.3) bei $t=0$, $r=2$ in zwei Berechnungsgängen. Die letzten drei Spalten enthalten der Reihe nach die Abschätzung s_m für die Stelle der Singularität, den ersten und zweiten Berechnungsgang. Den Wert p haben wir als bekannt vorausgesetzt. Die Werte s_m wurden aus der Formel (6.2) bei $l=2$ berechnet weil $l \geqq r-p$ ausreicht, wenn wir nur die Formel (6.2) benutzen. Eine Fehlerabschätzung im zweiten Berechnungsgang wurde nicht durchgeführt.

Tabelle für die Lösung der Anfangswertaufgabe $y' = \dfrac{(y^2-1)y}{2(1-x)}$, $y(-1) = 1{,}4142136$

x_n	exakte Lösung	Adams-Verfahren	Lambert-Shaw-V.	s_m	eigenes V. mit s_m	eigenes V. mit s
−0,300	2,0816660	2 0808589	2,0816510	0,0017291	2,0816536	2,0816594
−0,275	2,1532217	2,1521284	2,1532025	0,0014030	2,1532069	2,1532141
−0,250	2,2360680	2,2345521	2,2360431	0,0011276	2,2360498	2,2360589
−0,225	2,3333333	2,3311715	2,3333003	0,0008222	2,3333103	2,3333218
−0,200	2,4494898	2,4462988	2,4494448	0,0007110	2,4494603	2,4494755
−0,175	2,5911939	2,5862775	2,5911305	0,0004937	2,5911552	2,5911755
−0 150	2,7688746	2,6608743	2,7687814	0,0003547	2,7688226	2,7688510
−0,125	3,0000000	2,9860131	2,9998547	0,0002357	2,9999256	2,9999670
−0,100	3,3166248	3,2896529	3,3163785	0,0001667	3,3165105	3,3165762
−0,075	3,7859389	3,7259949	3,7854658	0,0001133	3,7857400	3,7858586
−0,050	4,5825757	4,4156946	4,5814522	0,0000770	4,5821564	4,5824279
−0,025	6,4031244	5,6882611	6,3989711	0,0000436	6,4017218	6,4027248

Die Aufgabe wurde auf einer Rechenanlage MINSK-2 berechnet, zusammen mit anderen Beispielen.

Die Abschätzung des Wertes s ist schon an der Stelle $x_n=-0{,}300$ genügend genau, wo der Wert der Lösung nach dem Adams-Verfahren dem genauen Wert

nahe ist. Hätten wir ferner das Adams-Verfahren als den ersten Berechnungsgang angewandt und die Abschätzung für den Wert s aus diesen letzten drei, recht ungenauen Werten der Lösung durchgeführt, so hätten wir dennoch eine gute Annäherung

$$s_m = -0{,}00168$$

bekommen.

8. Das Verfahren für Differentialgleichungssysteme

Die Ergebnisse dieses Artikels können unmittelbar für Differentialgleichungssysteme erster Ordnung verallgemeinert werden, in denen gewisse Funktionen der Lösung von singulärem Verhalten, die übrigen aber etlichemal differenzierbar sind.

Wie gewöhnlich, schreibt man spezielle Differenzenformel nur für die Differentialgleichungen zweiter Ordnung auf, in denen die Ableitung erster Ordnung fehlt:

$$y'' = f(x, y), \qquad \begin{matrix} y(x_0) = y_0 \\ y'(x_0) = y_0' \end{matrix}$$

Wir setzen voraus, dass der Ansatz (2.1) gilt. Der Wert y_{n+1} wird jetzt in der Form

$$y_{n+1} = \sum_{j=0}^{r} A_j^{**}(b) y_{n-j} + h^2 \sum_{j=0}^{r} C_j^{**}(b) f_{n-j}$$

aufgeschrieben. Nach einer ähnlichen Ableitung für den Fehler bekommen wir seine Form:

$$T_n^{**}(b) = u_n^{(m+1)} \frac{h^{m+1-p}}{(m+1)!} E_{m+1}^{**}(b) + I_{m+1}^{**}(b)$$

wenn $u(x) \in C_{m+2}[x_0, s]$, $m \geqq 2$ ist und für die Koeffizienten das lineare Gleichungssystem

$$E_k^{**}(b) = 0 \quad (k = 0, 1, \ldots, m)$$

gilt. Hier sind

$$E_k^{**}(b) = \frac{1}{(b-1)^p} - (-1)^k \sum_{j=1}^{r} A_j^{**}(b) \frac{j^k}{(b+j)^p} - (-1)^k$$

$$\sum_{j=1}^{r} C_j^{**}(b) \frac{j^{k-2}}{(b+j)^{p+2}} [(b+j)^2 k(k-1) - 2pk(b+j)j + p(p+1)j^2] -$$

$$- \frac{A_0^{**}(b)}{b^p} \delta_{0,k} - \frac{C_0^{**}(b)}{b^p} \left[\frac{p(p+1)}{b^2} \delta_{0,k} + \frac{2p}{b} \delta_{1,k} + 2\delta_{2,k} \right], \quad (k = 0, 1, \ldots, m+1)$$

und

$$I_{m+1}^{**}(b) = \frac{h^{m+2-p}}{(m+1)!} \int_{-r}^{1} G_{m+1}^{**}(b, z) u^{(m+2)}(x_n + hz)\, dz,$$

wo

$$G_{m+1}^{**}(b,z)=\frac{\overline{(1-z)}^{m+1}}{(b-1)^p}+\sum_{j=1}^{r}A_j^{**}(b)\frac{\overline{(-j-z)}^{m+1}}{(b+j)^p}+$$

$$+\sum_{j=1}^{r}\frac{C_j^{**}(b)}{(b+j)^{p+2}}[p(p+1)\overline{(-j-z)}^{m+1}+2p(b+j)(m+1)$$

$$\overline{(-j-z)}^{m}+(b+j)^2m(m+1)\overline{(-j-z)}^{m-1}].$$

Unsere Methode für Anfangswertsaufgaben kann auch dann angewandt werden, wenn die Lösung zwar nicht von singulärem Verhalten ist, aber sie sich, oder sich ihre Ableitung niedrigerer Ordnung auf dem untersuchten Abschnitt recht schnell ändern.

9. Die Anwendung der impliziten Formeln

Wollen wir die implizite Formel (2.5) zur Lösung verwenden, so kann die Konvergenz der mit dieser Formel verbundenen Iteration in zwei Fällen gesichert werden.

1) Im allgemeinen ist die Gleichung

$$x=F(x)$$

durch Iteration auflösbar, wenn die Ableitung $F'(x)$ in der Umgebung der Wurzel $\bar{x}$ stetig und negativ ist, s. die Abbildung 1. Die Formel für diese Iteration lautet

$$x_{n+1}=\frac{x_nF(F(x_n))-F(x_n)^2}{F(F(x_n))-2F(x_n)+x_n} \tag{9.1}$$

wo der Nenner rechts von der Wurzel positiv, links von ihr negativ ist. Die Konvergenz ist gesichert, wenn die Ableitung zwischen den Abszissen $x=x_0$, $x=F(x_0)$ stetig und negativ ist. Diese Aussage folgt aus dem Beweis der Lösungsmethode für das Gleichungssystem (6.3) (vgl. den dritten Teil). Bezeichnen wir den Koeffizienten des Glieds f_{n+1} in der Formel (2.5) mit $\bar{B}_{-1}^{(t)}$, so lautet die Bedingung für die Konvergenz der Methode (9.1)

$$h\bar{B}_{-1}^{(t)}f_y(x_n,y_n)<0,$$

wenn die Ableitung f_y der Lösung entlang stetig ist. Bei $t=0$ reicht schon die Bedingung

$$f_y(x_n,y_n)<0$$

aus, weil $\bar{B}_{-1}^{(0)}>0$ ist.

2) Wenn zwei Gleichungen

$$x=F(x)$$

$$x=G(x)$$

die gemeinsame Wurzel $\bar{x}$ haben und es eine Abschätzung für den Quotienten der Werte der Ableitungen in diesem Punkt,

$$\vartheta\cong\frac{F'(\bar{x})}{G'(\bar{x})}\quad(\vartheta\neq 1)$$

gibt, so kann das gewöhnliche Iterationsverfahren für die Gleichung

$$x = H(x) = \frac{\vartheta G(x) - F(x)}{\vartheta - 1}$$

angewandt werden. Hier sind

$$\bar{x} = H(\bar{x}), \quad H'(\bar{x}) \cong 0.$$

Wir setzen voraus, dass die Formel (2.5) bei $t=1$ und $t=0$ denselben Wert y_{n+1} ergibt. Wenden wir die Formel (2.5) bei $t=1$ und $t=0$ zugleich an, so kann die gewünschte Abschätzung bei diesen zwei Iterationsformeln leicht gewonnen werden:

$$\vartheta = \frac{\bar{B}_{-1}^{(1)}(b)}{\bar{B}_{-1}^{(0)}(b)}.$$

Es ist leicht zu beweisen, dass

$$0 < \vartheta < 1,$$

wenn $p+1>0$ ist. Der Wert ϑ ist zu berechnen, aber er kann nahe zu 1 sein.

Im dritten Teil werden wir uns mit Methoden zur Auflösung des Gleichungssystems (6.3) beschätfigen.

LITERATUR

[1] LAMBERT, J. D.—MITCHELL, A. R.: On the solution of $y'=f(x, y)$ by a class of high accuracy difference formulae of low order. *Zeitschrift für Angewandte Mathematik und Physik* **13** (1962), 223—232.

[2] LAMBERT, J. D.—SHAW, B.: On the numerical solution of $y'=f(x, y)$ by a class of formulae based on rational approximation, *Mathematics of Computation* **19** (1965) 456—461.

[3] LAMBERT, J. D.—SHAW, B.: A method for the numerical solution of $y'=f(x, y)$ based on a self adjusting non-polynomial interpolant, *Mathematics of Computation* **20** (1966), 11—20.

[4] LAMBERT, J. D.—SHAW, B.: Generalisation of multistep methods for ordinary differential equations, *Numerische Mathematik* **8** (1966), 250—263.

[5] RALSTON, A.: *A first course in numerical analysis.* McGraw-Hill, Ing. 1965.

[6] FRIVALDSZKY, S. Numerische Methode für die gewöhnliche Differentialgleichung erster Ordnung mit der Lösung eines Pols, *MTA III. Osztályának Közleményei* **20** (1971) (ungarisch).

III. Teil

Einführung

Dieser Teil ist in gewissem Sinn von den ersten zwei Teilen unabhängig. Wir werden uns hier mit der Lösung des Gleichungssystems unter (6.3) des zweiten Teils beschäftigen, überdies geben wir eine Methode zur Lösung beliebiger Gleichungssysteme mit zwei Unbekannten.

10. Die Lösung des Gleichungssystems (6.3) mit der Davidenko-Methode

Eine naheliegende Methode zur Lösung des Gleichungssystems (6.3) ist die Davidenko-Methode, [1]. Wir gehen von den Funktionen

$$\mu(c, p, L) = L\{\mu_l(c, p) - \mu_1(c, p)\} + \mu_1(c, p)$$
$$\nu(c, p, L) = L\{\nu_l(c, p) - \nu_1(c, p)\} + \nu_1(c, p)$$

aus, die — bei

$$\mu(c, p, L) = 0,$$

$$\nu(c, p, L) = 0$$

— die Lösung $c(L)$, $p(L)$ bei einem beliebigen Wert L bestimmen. Für diese Lösung kann das Differentialgleichungssystem

$$\begin{aligned} \frac{\partial \mu}{\partial c} c'(L) + \frac{\partial \mu}{\partial p} p'(L) &= -\frac{\partial \mu}{\partial L} \\ \frac{\partial \nu}{\partial c} c'(L) + \frac{\partial \nu}{\partial p} p'(L) &= -\frac{\partial \nu}{\partial L} \end{aligned} \tag{10.1}$$

aufgeschrieben werden. Setzen wir voraus, dass uns die Werte $c(0)$, $p(0)$ bekannt sind und suchen wir die Werte $c(1)$, $p(1)$, die die Lösung von (6.3) darstellen.

Die Koeffizienten

$$\frac{\partial \mu}{\partial c}, \quad \frac{\partial \mu}{\partial p}, \quad \frac{\partial \mu}{\partial L}, \quad \frac{\partial \nu}{\partial c}, \quad \frac{\partial \nu}{\partial p}, \quad \frac{\partial \nu}{\partial L}$$

sind leicht zu bestimmen. Das Differentialgleichungssystem (10.1) kann z. B. mit der Hamming Predictor-Corrector Methode aufgelöst werden, [2].

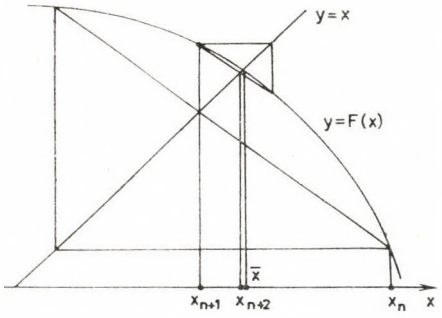


Fig. 1.

Das für die Anfangswerte $c(0)$, $p(0)$ aufgeschriebene Gleichungssystem

$$\mu_1(x, y) = 0$$

$$\nu_1(x, y) = 0$$

ist auf die Gleichungen

$$(1+x)^{\frac{1}{x}} = J_m = \left(\frac{y_m}{y_{m-1}}\right)^{\frac{y_m}{hf_m}} \tag{10.2}$$

$$y = \frac{1}{x}\frac{hf_m}{y_m}$$

zurückzuführen.

Die Gleichung (10.2) ist auflösbar, wenn der Wert J_m zwischen den Werten 2.25 und $e=2.718\ldots$ liegt. Weil

$$J_m \to (1+c)^{\frac{1}{c}}, \quad c = \frac{h}{s - x_m}$$

wenn $h \to 0$, $u(s) \neq 0$ und c ein fester Wert ist, so kann die Gleichung (10.2) bei einer gegebenen Schrittzahl vor der Singularität aufgelöst werden, wenn die Schrittweite h genügend klein ist. In diesem Fall ist die Gleichung

$$f(x) = (1+x)^{\frac{1}{x}} - J_m = 0$$

durch die umgestaltete Sehnen-Methode vom Intervall (0,001, 0,5) ausgehend aufzulösen und, vom Punkt $x_0=0{,}001$ ausgehend, auch durch die Newton-Methode, weil im Intervall (0,001, 0,5)

$$f'(x) < 0, \quad f''(x) > 0$$

sind.

11. Eine neue Methode zur Lösung der Gleichungssysteme mit zwei Unbekannten

Wir empfehlen die folgende iterative Methode zur Lösung des Gleichungssystems (6.3), die zur Lösung beliebiger Gleichungssysteme mit zwei Unbekannten geeignet ist. Es seien die Werte $(\bar{x}, \bar{y})$ die Lösung von (6.3), für die die Werte $(y^{(n)}, x^{(n)})$ eine gute Annäherung geben. Es seien beide partiellen Ableitungen der Funktionen $\mu \equiv \mu_l$, $\nu \equiv \nu_l$ in der Umgebung der Wurzel von (6.3) stetig und bei dieser Wurzel nicht verschwindend, und es gelte für diese Ableitungen die Ungleichung

$$\mu'_x \nu'_y \neq \mu'_y \nu'_x$$

in der Wurzel.

In diesem Fall sind die Werte (ξ_i, η_i), $(i=1, 2, 3)$ eindeutig zu bestimmen: (s. die Abbildung 2)

$$\begin{aligned}
\xi_1&: \mu(\xi_1, y^{(n)}) = 0,\\
\eta_1&: \nu(\xi_1, \eta_1) = 0,\\
\xi_2&: \mu(\xi_2, \eta_1) = 0,\\
\eta_2&: \nu(x^{(n)}, \eta_2) = 0,\\
\xi_3&: \mu(\xi_3, \eta_2) = 0,\\
\eta_3&: \nu(\xi_3, \eta_3) = 0.
\end{aligned}$$

Bezeichnen wir den gemeinsamen Punkt der Geraden $\{(\xi_1, y^{(n)}); (\xi_2, \eta_1)\}$, und $\{(x^{(n)}, \eta_2); (\xi_3, \eta_3)\}$ mit $(x^{(n+1)}, y^{(n+1)})$, wenn es diesen gibt. Er wird die folgende Annäherung abgeben. Hier gelten

$$x^{(n)} \to \bar{x}$$
$$y^{(n)} \to \bar{y}$$

wenn das Verfahren von einer der Wurzel nahen Anfangsannäherung $(x^{(0)}, y^{(0)})$ angefangen wird. Der Beweis wird abgekürzt beschrieben. Es sein an der Stelle der Wurzel

$$f' = -\frac{\mu'_x}{\mu'_y}, \quad g' = -\frac{\nu'_x}{\nu'_y}$$

Es gibt eine Zahl $0<\varepsilon_0<1$, für die

$$\frac{1}{\varepsilon_0} > |f'|, |g'| > \varepsilon_0, \quad |f'-g'| > \varepsilon_0$$

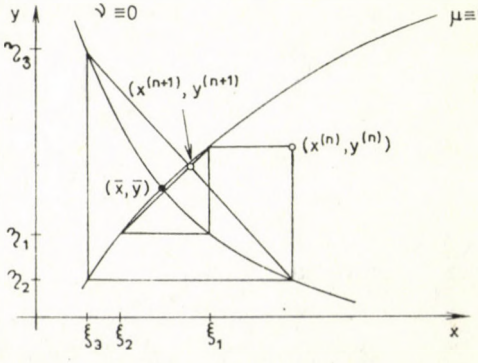


Fig. 2.

sind. Bei einer beliebigen, später zu bestimmenden Zahl $0<\varepsilon<\varepsilon_0/2$ gibt es ein Intervall $(\bar{x}-\tau, \bar{x}+\tau)$ in dem sich sowohl die Änderung $(-\mu'_x/\mu'_y)$ der Kurve $\mu=0$ entlang als auch die Änderung $(-\nu'_x/\nu'_y)$ der Kurve $\nu=0$ entlang höchstens auf den Wert ε belaufen. Ist der Punkt $(x^{(n)}, y^{(n)})$ dem Punkt $(\bar{x}, \bar{y})$ nahe genug, so liegen die Zahlen $\xi_i (i=1, 2, 3)$ in diesem Intervall. Man kann die Gleichungen

$$\xi_1 = \bar{x} + \frac{y^{(n)} - \bar{y}}{f' + \delta_1}$$

$$\eta_1 = \bar{y} + (\xi_1 - \bar{x})(g' + \delta_2)$$

$$\xi_2 = \bar{x} + \frac{\eta_1 - \bar{y}}{f' + \delta_3}$$

$$\eta_2 = \bar{y} + (x^{(n)} - \bar{x})(g' + \delta_4)$$

$$\xi_3 = \bar{x} + \frac{\eta_2 - \bar{y}}{f' + \delta_5}$$

$$\eta_3 = \bar{y} + (\xi_3 - \bar{x})(g' + \delta_6)$$

aufschreiben, wo $|\delta_i|<\varepsilon$ $(i=1, 2, \ldots, 6)$ sind. Die Bezeichnungen

$$\alpha^{(n)} = x^{(n)} - \bar{x}$$

$$\beta^{(n)} = y^{(n)} - \bar{y}$$

$$\gamma^{(n)} = \max(|\alpha^{(n)}|, |\beta^{(n)}|)$$

eingeführt ist das lineare Gleichungssystem

$$\alpha^{(n+1)} \left(\frac{g' + \delta_2}{f' + \delta_1} - 1 \right) - \beta^{(n+1)} \left(\frac{g' + \delta_2}{f' + \delta_3} - 1 \right) \frac{1}{f' + \delta_1} =$$

$$= \beta^{(n)} \frac{(g' + \delta_2)(\delta_3 - \delta_1)}{(f' + \delta_1)^2 (f' + \delta_3)}$$

$$\alpha^{(n+1)} \left(\frac{g' + \delta_6}{f' + \delta_5} - 1 \right) - \beta^{(n+1)} \left(\frac{g' + \delta_4}{f' + \delta_5} - 1 \right) \frac{1}{g' + \delta_4} =$$

$$= \alpha^{(n)} \frac{\delta_6 - \delta_4}{f' + \delta_5}$$

in Verbindung mit dem gemeinsamen Punkt der obigen Geraden aufzuschreiben. Wir haben vorausgesetzt, dass $\alpha^{(n)}, \beta^{(n)} \neq 0$ sind. Der Cramer-Regel nach sind

$$\alpha^{(n+1)} = \frac{D_\alpha}{D}, \quad \beta^{(n+1)} = \frac{D_\beta}{D}$$

wo

$$|D| > \frac{1}{2}\frac{|g'-f'|^3}{|f'|^3|g'|},$$

(11.1)

$$|D_\alpha| < 4\varepsilon\gamma^{(n)}\frac{|g'-f'|}{|f'|^3}\left(1+\frac{1}{|f'|}\right),$$

$$|D_\beta| < 4\varepsilon\gamma^{(n)}\frac{|g'-f'|}{|f'|^2}\left(1+\frac{|g'|}{|f'|^2}\right),$$

wenn die Zahl ε — die nur von den Werten f', g' abhängt — klein genug ist. Deshalb gilt

$$\gamma^{(n+1)} < \varepsilon K\gamma^{(n)} \quad \left(K < \frac{16}{\varepsilon_0^5}\right)$$

wo auch die Zahl K nur von den Werten f', g' abhängt. Gilt noch die Voraussetzung

$$\varepsilon < \frac{q}{K} \quad (0 < q < 1),$$

so ist

$$\gamma^{(n+1)} < q\gamma^{(n)},$$

und dann liegt auch die folgende Annäherung $(x^{(n+1)}, y^{(n+1)})$ in der obigen Umgebung der Wurzel und das Verfahren ist konvergent. (Der Zusammenhang (11.1) beweist, dass es sich um den Fall der zwei verschiedenen, nicht parallelen Geraden handelt.)

Das Verfahren is von einer guten Annäherung ausgehend konvergent. (Seine Konvergenz ist T. Frey nach quadratisch.)

Durch die Drehung des Koordinatensystems ist erreichbar, dass die genannten partiellen Ableitungen von Null verschieden seien. Zweckmässig ist, die Achsenrichtungen so zu wählen, dass

$$f'+g' = 0$$

sei.

Diese Methode wäre für die Lösung der Gleichungssysteme mit n Veränderlichen zu verallgemeinern, wobei zahlreiche Gleichungssysteme mit $(n-1)$ Veränderlichen in jedem Schritt der Iteration aufgelöst werden sollten. Aber das Verfahren würde selbst bei $n=3$ so viel Arbeit erfordern, dass es anzuwenden nicht zweckmässig wäre.

Diese Iteration ist für die Kurven $\mu=0$, $\nu=0$ nicht symmetrisch. Die Rolle der Kurven vertauschend können wir eine zweite, konvergierende Punktreihe bekommen. Man kann im voraus nicht wissen, welche Punktreihe schneller konvergieren wird, deshalb scheint es zweckmässig, beide zu berechnen. Die Kombination der beiden Punktreihen bietet mehrere Methoden für die Verbesserung der Konvergenz, auf welche wir aber hier nicht eingehen.

Die partiellen Ableitungen der Funktionen μ, ν existieren und sind in der Umgebung der Wurzel stetig, wenn $cl<1$ ist.

Es ist leicht zu beweisen, dass

$$\frac{f'}{p} < \frac{g'}{p}$$

gilt, wenn wir eine gegebene Anzahl von Schritten vor der Singularität sind, $u_m \neq 0$ ist, die zu bestimmenden Werte (c_m, p_m) den genauen Werten nahe und die Schritt-

weite h hinreichend klein sind. Für die erste Annäherung bei der Iteration kann man Werte aus der vorigen Iteration bekommen:

$$c_{m+1}^{(0)} = \frac{c_m}{1-c_m}, \quad p_{m+1}^{(0)} = p_m.$$

Gibt es keine vorige Iteration, so sind entweder die Davidenko-Methode für die gesuchten Werte (c_m, p_m), oder die folgende erste Annäherung für diese anzuwenden: es gelten näherungsweise

$$\frac{y_{m+1}y_{m-1}-y_m^2}{y_m^2} = c^2 p$$

$$\frac{y_m y_{m-2}-y_{m-1}^2}{y_{m-1}^2} = \frac{c^2 p}{(c+1)^2}$$

wenn $u_m \neq 0$ ist, von wo die Werte (c, p) für die erste Annäherung leicht auszudrükken sind.

12. Eine Methode beim Fall des einzigen unbekannten Parameters

Es kommt häufig vor dass einer der Parameter (c, p), besonders der Parameter p, bekannt ist. Wir wollen z. B. den Parameter c aus der Gleichung (6.2) des zweiten Teils bestimmen. Anstatt dieser ist die Gleichung

$$x = h(x)$$

wo

$$h(x) = x - \mu_l(x)\frac{\prod_{i=1}^{l}(1+xi)}{py_m l! x^l} \tag{12.1}$$

anzuwenden. Für diese gilt

$$h(c) \cong c$$

näherungsweise und auch

$$h'(c) \cong 0,$$

wenn die Schrittweite h bei einem fixen Wert c klein genug ist. Wir haben hier das Lemma des ersten Teiles bei $z = 1/c$ und $k = 0$ angewandt.

Die Gleichung (12.1) ist durch eine Iteration von einem guten Anfangswert ausgegangen aufzulösen, wenn die Annäherung c_m dem genauen Wert c nahe ist. Das Ergebnis wird die Lösung von (6.2) sein, weil

$$c_m \neq 0, -1/i \quad (i = 1, 2, \ldots, l)$$

sind.

LITERATUR

[1] FILIPPI S,—GLASMACHER, W.: Zum Verfahren von Davidenko, *Elektronische Datenverarbeitung*. 1967. 2.

[2] RALSTON, A.: *A first course in numerical analysis*, McGraw-Hill, Ing. 1965.

NIM IGÜSZI, 1363 Budapest, Pf. 33

(Eingegangen am 24. Juni 1971)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 19—22.

PANCONNECTED GRAPHS

by

YOUSEF ALAVI* and JAMES E. WILLIAMSON

A graph G is *connected* if for every pair u, v of distinct vertices of G, there exists a u—v path. If u and v are vertices of a connected graph G, then the *distance* $d_G(u, v) = d(u, v)$ between u and v is the length of a shortest $u-v$ path. Hence, if G has order p and l is the length of a $u-v$ path in G, then $d(u, v) \leqq l \leqq p-1$. We define a graph G to be *panconnected* if for each pair u, v of vertices in G, there exists a $u-v$ path of length l for each l such that $d(u, v) \leqq l \leqq p-1$.

J. A. BONDY has recently introduced the concept of edge-pancyclic graphs. A graph G is said to be *edge-pancyclic* if every edge of G is contained in a cycle of every length l, where $3 \leqq l \leqq p$. If a graph is panconnected, then it is edge-pancyclic; however, the concepts are not equivalent. The graph G shown in Fig. 1. is edgepancyclic but not panconnected, since G contains no $u-v$ paths of length l for $l=5$, 6, or 7, while $d(u, v)=2$ and $p-1=7$. In fact, the join of $\overline{K}_2$ with $2K_n$ for each integer $n \geqq 3$ is an infinite class of edge-pancyclic graphs which are not panconnected, where $\overline{K}_2$ denotes the graph with 2 vertices and no edges and $2K_n$ denotes the disconnected graph with two components each isomorphic with the complete graph of order n.

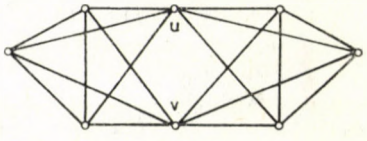


Figure 1.

The nth power G^n of a connected graph G is the graph whose vertex set is that of G and with the property that two distinct vertices are adjacent in G^n if and only if the distance between these vertices in G is at most n. It was shown independently by KARAGANIS [1] and SEKANINA [3] that if G is a connected graph of order p, then G^3 is hamiltonian-connected, i.e., every two distinct vertices are connected by a path of length $p-1$. The object of Theorem 1 is to present a strengthening of that result.

THEOREM 1. *If G is a connected graph, then G^3 is panconnected.*

PROOF. We proceed by induction on the order p of connected graphs. For small values of p, the result follows immediately. Assume for all connected graphs H of order less than p that H^3 is panconnected. Let G be a connected graphs of order p, and let u and v be any two distinct vertices of G. Let T be a spanning tree of G such that $d_T(u, v)=d_G(u, v)$. (The existence of such a tree is shown in [2, p. 103].) It is now sufficient to verify that there exists in T^3 a $u-v$ path of length l for each l such that $d_T(u, v) \leqq l \leqq p-1$.

We consider two cases depending on whether u and v are adjacent in T.

*Research partially sspported by a Western Michigan University Faculty Research Fellowship.

Case 1. Assume u and v are adjacent in T.

Here we show that there exists in T^3 a $u-v$ path of every positive length not exceeding $p-1$. Such a path of length one is obvious. Let T_1 and T_2 be the components containing u and v, respectively, in the graph $T-uv$, and let p_i, $i=1, 2$, be the order of T_i.

Suppose, first, that either $p_1=1$ or $p_2=1$, say the former. Let v_1 be a vertex of T_2 adjacent with v. Since $p_2=p-1$, it follows by the induction hypothesis that there exist $v-v_1$ paths in T_2^3 of length l_2 for all l_2 such that $1\leqq l_2\leqq p-2$. However, u and v_1 are adjacent in T^3; hence in T^3, there are $u-v$ paths of length l for all l such that $2\leqq l\leqq p-1$.

We now assume that $p_1\geqq 2$ and $p_2\geqq 2$. Let u_1 be adjacent with u in T_1, and let v_1 be adjacent with v in T_2. Thus $d_T(u_1, v_1)=3$, and u_1 and v_1 are adjacent in T^3.

By the induction hypothesis, there exists in T_1^3 a $u-u_1$ path of length l_1 for each l_1 such that $1\leqq l_1\leqq p_1-1$, and there exists in T_2^3 a $v-v_1$ path of length l_2 for each l_2 such that $1\leqq l_2\leqq p_2-1$. Hence by combining such a $u-u_1$ path, such a $v-v_1$ path, and the edge u_1v_1 of T^3, we obtain $u-v$ paths in T^3 of length l for each l such that $3\leqq l\leqq (p_1-1)+(p_2-1)+1=p_1+p_2-1=p-1$. Furthermore, u, v_1, v is a $u-v$ path of length two in T^3. Therefore, the result follows in this case.

Case 2. Assume u and v are not adjacent in T.

On the unique $u-v$ path P in T, select the vertex w on P adjacent with u. As in Case 1, denote by T_1 and T_2 the components of $T-uw$ containing u and v, respectively, and let p_i, $i=1, 2$, represent the order of T_i.

Assume that $p_1=1$ so that $p_2=p-1$. Since, by the induction hypothesis, T_2^3 is panconnected, T_2^3 and thus T^3 contains $w-v$ paths of length l_2 for all l_2 such that $d_{T^3}(w, v)\leqq l_2\leqq p-2$. Combining such a $w-v$ path with the edge uw, we obtain $u-v$ paths in T^3 of length l for each l such that $1+d_{T^3}(w, v)\leqq l\leqq p-1$. Observe that either $d_{T^3}(u, v)=d_{T^3}(w, v)$ or $d_{T^3}(u, v)=1+d_{T^3}(w, v)$. Hence there exist $u-v$ paths in T^3 of length l for each l such that $1+d_{T^3}(u, v)\leqq l\leqq p-1$. Since T^3 must by definition contain $u-v$ paths of length $d_{T^3}(u, v)$ we have the desired result.

Henceforth, we assume $p_1>1$, and let u' be a vertex of T_1 adjacent with u. By the induction hypothesis, T_1^3 is panconnected and thus contains $u-u'$ paths of length l_1 for all l_1 such that $1\leqq l_1\leqq p_1-1$. Also, T_2^3 is panconnected, from which it follows that there exist $w-v$ paths of length l_2 for all l_2 such that $d_{T^3}(w, v)\leqq l_2\leqq p_2-1$. Combining the above $u-u'$ paths with the $w-v$ paths and the edge $u'w$, we obtain $u-v$ paths in T^3 of length l for all l such that $2+d_{T^3}(w, v)\leqq l\leqq (p_1-1)+(p_2-1)+1=$ $=p-1$. If $d_{T^3}(u, v)=1+d_{T^3}(w, v)$, then this inequality gives the existence of $u-v$ paths of all lengths l such that $1+d_{T^3}(u, v)\leqq l\leqq p-1$, so once again the desired result has been obtained, since T^3 must contain $u-v$ paths of length $d_{T^3}(u, v)$. However, if $d_{T^3}(u, v)=d_{T^3}(w, v)$, then a $w-v$ path of length $d_{T^3}(w, v)$ in T_2^3 together with the edge uw produces a $u-v$ path in T^3 of length $1+d_{T^3}(u, v)$, so that this case and the proof is complete.

Since there exist graphs G whose square is not hamiltonian-connected, there exist graphs G whose square is not panconnected. However, it can be observed that if G is a hamiltonian graph, then its square is hamiltonian-connected. This result is strengthened by Theorem 2.

THEOREM 2. *If G is a hamiltonian graph, then G^2 is panconnected.*

PROOF. Let G be a hamiltonian graph and let u and v be a pair of vertices of G. Furthermore, let C be a hamiltonian cycle of G, and suppose that the vertices of G are labelled $v_1, v_2, \ldots, v_p$, consecutively along the cycle C. Without loss of generality we assume u has been labelled v_1 and v has been labelled v_k, for some k, $2 \leqq k < p$.

Let P be a $u-v$ path in G such that the length of P is $d_G(u, v)$. Then in G^2, the subgraph induced by the vertices of P contains a $u-v$ path P' of length m, where m is the least integer such that $2m \geqq d_G(u, v)$, which, with at most one exception, consists of edges of G^2 which are not edges of G. The path P' is necessarily a shortest $u-v$ path in G^2, so that $d_{G^2}(u, v) = m$.

We show that there exists a $u-v$ path of length l, for each l such that $d_{G^2}(u, v) \leqq \leqq l \leqq p-1$, by means of three cases, depending on the value of l.

Case 1. Suppose l is such that $d_{G^2}(u, v) \leqq l \leqq d_G(u, v)$.

Let H be the subgraph of G^2 defined by $V(H) = V(P)$ and $E(H) = E(P) \cup E(P')$. Since both P' and P are also subgraphs of H, the subgraph H contains $u-v$ paths of length $d_{G^2}(u, v)$ and $d_G(u, v)$. Now suppose H contains a $u-v$ path P'' of length l, where $d_{G^2}(u, v) \leqq l < d_G(u, v)$. Then P'' contains in its edge set an edge e which is an edge of P' that is not an edge of P. However, the edge e is adjacent with a pair of edges f_1 and f_2 which are mutually adjacent edges of P, such that the subgraph induced by e, f_1, and f_2 is a triangle in H. Then $P'' - e$, together with f_1, f_2, and their common vertex, is a $u-v$ path of length $l+1$ in H. Therefore H, and hence G^2, contains $u-v$ paths of all lengths l, where $d_{G^2}(u, v) \leqq l \leqq d_G(u, v)$.

Case 2. Suppose l is such that $d_G(u, v) \leqq l \leqq k-1$.

Recall that $u = v_1$ and $v = v_k$. There exists a $u-v$ path of length $k-1$ using only edges of C, so that indeed $d_G(u, v) \leqq k-1$. If $d_G(u, v) = k-1$, the result follows immediately in this case. Hence we assume $d_G(u, v) < k-1$.

Let v_i be the first vertex encountered on P such that $1 \leqq i \leqq k$, and v_j, the next vertex encountered on P, is not v_{i+1}. Furthermore, let m be the smallest integer such that $m > i$ and v_m lies on P. Necessarily, $m \leqq k$. Note that $m \neq i+1$, for otherwise P is not a shortest path in G. Now $v_i v_{i+1}$ and $v_i v_j$ are edges of G, so that $v_{i+1} v_j$ is an edge of G^2. Hence, if P is written

$$P\colon\ P_i, P_j$$

where P_i is the $v_1 - v_i$ subpath of P, and P_j is the $v_j - v_k$ subpath of P, then the path P^* written

$$P^*\colon\ P_i, v_{i+1}, P_j$$

is a $u-v$ path of length one greater than P.

For each t, $i < t < m-1$, $v_{t-1} v_t$ and $v_t v_{t+1}$ are edges of G, and hence $v_{t-1} v_{t+1}$ is an edge of G^2. Thus if $m \geqq i+2$, the edge $v_i v_{i+1}$ of P^* may be replaced by $v_i v_{i+2}$, v_{i+2}, and $v_{i+2} v_{i+1}$ to obtain a $u-v$ path of length one greater than P^*. Similarly, this may be repeated for each t, $i < t < m-1$, until a path Q is obtained which has length $m-i-1$ greater then P and uses all the vertices of G which lie between v_i and v_m on C. That is, Q may be written

$$Q\colon\ P_i, v_{i+2}, v_{i+4}, \ldots, v_{i+2n}, v_{i+2n+1}, v_{i+2n-1}, \ldots, v_{i+1}, P_j,$$

where $n=(m-i)/2-1$, and $m-i$ is even, or

$$Q\colon\ P_i, v_{i+2}, v_{i+4}, \ldots, v_{i+2n}, v_{i+2n-1}, v_{i+2n-3}, \ldots, v_{i+1}, P_j,$$

where $n=\dfrac{m-i-1}{2}$, and $m-i$ is odd.

Analogously, at each successive vertex v_s of P, where $1\leqq s\leqq k$, and the next vertex encountered on P is not v_{s+1}, the process produces a sequence of paths, each one greater in length than the previous, until all the vertices of G which lie between v_i and v_k have been used.

Case 3. *Suppose l is such that* $k\leqq l\leqq p-1$.

Let R denote the $u-v$ path of G such that $V(R)=\{v_1, v_2, \ldots, v_k\}$. Then R is a $u-v$ path of length $k-1$. Now $v_{k-1}v_k$ is an edge of G and of R. Also v_kv_{k+1} is an edge of G, so that $v_{k-1}v_{k+1}$ is an edge of G^2. Then the edge $v_{k-1}v_k$ of R can be replaced by $v_{k-1}v_{k+1}$, v_{k+1}, and $v_{k+1}v_k$ to obtain a path R' of length k in G^2. Similarly, the edge v_kv_{k+1} of R' may be replaced by the edge v_kv_{k+2}, v_{k+2}, and $v_{k+2}v_{k+1}$ to obtain a path R'' of length $k+1$ in G^2. Furthermore, for each s, $k<s<p$, $v_{s-1}v_s$ is an edge of G, v_sv_{s+1} is an edge of G, and $v_{s-1}v_{s+1}$ is an edge of G^2. Applying this fact recursively from R', the path R'' being the first such application, a sequence of paths of all lengths l, where $k\leqq l\leqq p-1$ is obtained. This completes the case and the proof, since u and v are arbitrary.

REFERENCES

[1] KARAGANIS, J. J. On the cube of a graph., *Canad. Math. Bull.*, **11** (1969), 295—296.
[2] ORE, O.: *Theory of Graphs*, Amer. Math. Soc. Colloq. Publ., 38, Providence (1962).
[3] SEKANINA, M.: On an ordering of the set of vertices of a connected graph. *Publ. Fac. Sci. Univ. Brno.*, **412** (1960), 137—142.

Western Michigan University
University of Dubuque

(Received April 9, 1973, revised April 23, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 23—31.

REMARKS ON FREE PRODUCTS IN REGULAR VARIETIES AND SINK-COMPLEMENTED SUBALGEBRAS

by

H. ANDRÉKA and I. NÉMETI

1. Introduction

First we give a natural construction for the absolutely free extension of a partial algebra. Then a general method is given to relativise an arbitrary algebra to a variety. Actually both a sintactical and an equivalent semantical method are given. Here an interlude follows, we turn our attention to sink-complemented subalgebras and list some important properties. Having done this, we are ready to prove a theorem on relatively free extensions of free algebras, which is a restatement of JONSSON's result on a more general level. In the same time the techniques outlined above result in considerable simplifications of JONSSON's proofs.

We use the following notations:

Algebras are denoted by german capitals $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{C}$, ... and their universes by the corresponding roman letters: A, B, C etc. The symbols $\subseteq$ and $\cong$ stand for "is a subalgebra of" and "is isomorphic to" respectively. The notation $\mathfrak{A} \overset{\sim}{\underset{h}{\rightarrow}} \mathfrak{B}$ means that h is a homomorphism from $\mathfrak{A}$ into $\mathfrak{B}$. Relations often are written transitively, e.g. $\mathfrak{A} \overset{\sim}{\underset{h}{\rightarrow}} \mathfrak{B} \in \mathscr{V}$ or $\mathfrak{A} \overset{\sim}{\underset{h}{\rightarrow}} \mathfrak{B} \overset{\sim}{\underset{h}{\rightarrow}} \mathfrak{C}$, the first of which stands for $\mathfrak{A} \overset{\sim}{\underset{h}{\rightarrow}} \mathfrak{B}$ and $\mathfrak{B} \in \mathscr{V}$. $Sg^{(\mathfrak{A})}X$, $Sg^{(\mathfrak{A})}X$ and $Cg^{(\mathfrak{A})}X$ denote respectively the subalgebra, subunivere and congruences generated by X in the algebra $\mathfrak{A}$. The superscript $(\mathfrak{A})$ is often omitted. $\mathfrak{B}/R$ denotes the factoralgebra of $\mathfrak{B}$ over the congrunce R.

By a regular variety we understand a variety without operations of rank zero. The formula $\mathscr{A} \Vdash e$ means that the equation e holds in the class $\mathscr{A}$ of algebras.

Functions are considered as (special) relations, that is, sets of ordered pairs. If R is a relation, then $R \restriction N$ denotes the restriction of R to N, that is the set of those ordered pairs in R the first elements of which belong to N. The identity relation over N is denoted by Id_N. The equivalence relation defined by a function f (on its domain) is denoted by f^0.

In formulating the theorems we use the metasymbols $\Rightarrow$, $\Leftrightarrow$, $\&$, $\exists$, $\forall$ in the usual sense; the letter d in the symbols $\overset{d}{=}$ and $\overset{d}{\Leftrightarrow}$ denotes, that the formula is a definition. The ends of proofs are marked by $\blacksquare$.

2. A construction of relatively free extensions

The following construction is a very natural way of extending — without specifying unnecessary order — a partial algebra to an algebra or, with other words, of extending it freely.

DEFINITION 1. Let $\mathfrak{N}$ be a partial algebra. First we define an operator $\mathfrak{Fr}$ on partial algebras. $\mathfrak{Fr}(\mathfrak{N})$ is the partial algebra obtained from $\mathfrak{N}$ by "one step extension": For every n-ary operation symbol f of $\mathfrak{N}$ and elements $a_1, \ldots, a_n$ of N if $f(a_1, \ldots, a_n)$ is not defined in $\mathfrak{N}$, then we put the element $\langle f, \langle a_1, \ldots, a_n\rangle\rangle$* into the universe of $\mathfrak{Fr}(\mathfrak{N})$, and define $f(a_1, \ldots, a_n)$ in $\mathfrak{Fr}(\mathfrak{N})$ as $f(a_1, \ldots, a_n) \stackrel{d}{=} \langle f, \langle a_1, \ldots, a_n\rangle\rangle$. Now, the absolute free extension of $\mathfrak{N}$ to an algebra is denoted by $\mathfrak{Fr}_{\mathfrak{N}}$, and $\mathfrak{Fr}_{\mathfrak{N}} \stackrel{d}{=} \bigcup_{n=1}^{\infty} \mathfrak{Fr}^n(\mathfrak{N})$, where $\mathfrak{Fr}^n$ denotes the n-iteration of $\mathfrak{Fr}$. The universe of $\mathfrak{Fr}_{\mathfrak{N}}$ is denoted by $\mathfrak{Fr}_{\mathfrak{N}}$. It is easily seen, that $\mathfrak{Fr}_{\mathfrak{N}}$ is an algebra indeed. We constructed this algebra without specifying unnecessary order in the following sense:

THEOREM 1. *Every homomorphism of $\mathfrak{N}$ can be extended to a homomorphism of $\mathfrak{Fr}_{\mathfrak{N}}$.*

In general instead of the absolute free extension we need only the free extension over a class $\mathscr{A}$ of algebras. To "relativise" an algebra $\mathfrak{B}$ to a class $\mathscr{A}$ of algebras we have the following tools:

DEFINITION 2. $\mathfrak{B}\mathscr{A} \stackrel{d}{=} \mathfrak{B}/Cr^{\mathfrak{B}}\mathscr{A}$, where $Cr^{\mathfrak{B}}\mathscr{A} \stackrel{d}{=} \bigcap \{h^{\circ}\colon \mathfrak{B} \underset{h}{\simeq} \mathfrak{A} \in \mathscr{A}\}$.

THEOREM 2. *Every homomorphism from $\mathfrak{B}$ to $\mathscr{A}$ can be taken through $\mathfrak{B}\mathscr{A}$, that is*

$$(\forall \mathfrak{B} \simeq \mathfrak{A} \in \mathscr{A})(\exists \mathfrak{B} \underset{g}{\simeq} \mathfrak{B}\mathscr{A} \underset{f}{\simeq} \mathfrak{A}) h = g|f,$$

where $g|f$ stands for the composition of g and f.

The proofs of Th. 1. and Th. 2. are essentially known.

In case $\mathscr{A}$ is a variety to relativise $\mathfrak{B}$ to $\mathscr{A}$ is just the same as to force $\mathfrak{B}$ to satisfy the equations of $\mathscr{A}$. More precisely:

Let α be an element of the t-type word algebra generated by a countable set. If $\mathfrak{A}$ is a t-type algebra, then $\alpha^{(\mathfrak{A})}$ denotes the polynomial in $\mathfrak{A}$ (of infinite arguments) corresponding to the word α. We call those arguments of $\alpha^{(\mathfrak{A})}$ which correspond to the variable symbols occurring in α the real arguments of $\alpha^{(\mathfrak{A})}$. When we write "α really depends: $\alpha^{(\mathfrak{A})}(a_1, \ldots, a_n)$" we mean that $a_1, \ldots, a_n$ are exactly the real arguments of $\alpha^{(\mathfrak{A})}$. On the other hand by just writing $\alpha^{(\mathfrak{A})}(a_1, \ldots, a_n)$ we mean that the real arguments are among $a_1, \ldots, a_n$.

DEFINITION 3. $E^{\mathfrak{B}}\mathscr{A} \stackrel{d}{=} \{\langle \alpha^{(\mathfrak{B})}(a_1, \ldots, a_n), \beta^{(\mathfrak{B})}(a_1, \ldots, a_n)\rangle : \mathscr{A} \Vdash \alpha = \beta,\ a_1, \ldots, a_n \in B\}$. (Note that here α and β need not really depend.)

---

* Of course, we suppose that $\langle f, \langle a_1, \ldots, a_n\rangle\rangle$ does not belong to N.

The relation $E^{\mathfrak{B}}\mathscr{A}$ is reflexive, symmetric but not necessarily transitive. An example follows:

Let $\mathfrak{B}$ be the algebra illustrated by figure 1. There are two fundamental operations in $\mathfrak{B}$, the $+$ and the $\cdot$, defined as:

$$+(a) \overset{d}{=} a^+ \overset{d}{=} c \overset{d}{=} b^+; \quad \cdot(a, b) \overset{d}{=} a \cdot b \overset{d}{=} f; \quad b \cdot d \overset{d}{=} g; \quad c \cdot d \overset{d}{=} h;$$

and in all the other cases the result is e.

Let $\mathscr{A}$ be the set of algebras (of type $\{\langle +, 1\rangle, \langle \cdot, 2\rangle\}$) in which the equation $x^+ \cdot y = x \cdot y$ holds.

It is easily seen that $\langle a \cdot d, c \cdot d\rangle$ and $\langle c \cdot d, b \cdot d\rangle$ belong to $E^{\mathfrak{B}}\mathscr{A}$. But $\langle a \cdot d, b \cdot d\rangle$ is not an element of $E^{\mathfrak{B}}\mathscr{A}$, because a, b and d cannot be represented as results of performing any operation, and so $\langle a \cdot d, b \cdot d\rangle$ could be obtained only from the equation $x \cdot y = z \cdot y$, but $\mathscr{A} \not\Vdash x \cdot y = z \cdot y$. So $E^{\mathfrak{B}}\mathscr{A}$ is not transitive.

THEOREM 3. *If $\mathscr{A}$ is a variety, then $Cr^{\mathfrak{B}}\mathscr{A} = Cg\, E^{\mathfrak{B}}\mathscr{A}$.*

PROOF: 1. $E^{\mathfrak{B}}\mathscr{A} \subseteq Cr^{\mathfrak{B}}\mathscr{A}$. Let $\mathfrak{B} \overset{\sim}{\underset{h}{\to}} \mathfrak{A} \in \mathscr{A}$ be arbitrary. Then $E^{\mathfrak{B}}\mathscr{A} \subseteq h^\circ$, because if $\mathscr{A} \Vdash \alpha = \beta$, then

$$h(\alpha(a_1, \ldots, a_n)) = \alpha(h(a_1), \ldots, h(a_n)) = \beta(h(a_1), \ldots, h(a_n)) = h(\beta(a_1, \ldots, a_n)).$$

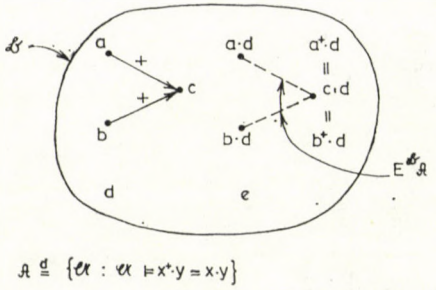


Figure 1.

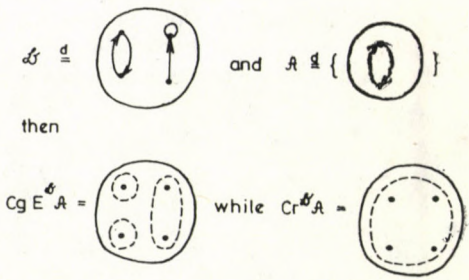


Figure 2.

2. $Cr^{\mathfrak{B}}\mathscr{A} \subseteq Cg\, E^{\mathfrak{B}}\mathscr{A}$. Because $\mathscr{A}$ is a variety and $\mathfrak{B}/Cg\, E^{\mathfrak{B}}\mathscr{A}$ satisfies the equations valid in $\mathscr{A}$, $\mathfrak{B}/Cg\, E^{\mathfrak{B}}\mathscr{A} \in \mathscr{A}$. So, the natural homomorphism f of the congruence $Cg\, E^{\mathfrak{B}}\mathscr{A}$ is a homomorphism from $\mathfrak{B}$ to $\mathscr{A}$, for which

$$Cg E^{\mathfrak{B}}\mathscr{A} = f^\circ \supseteq Cr^{\mathfrak{B}}\mathscr{A}. \qquad \blacksquare$$

Note: If $\mathscr{A}$ is not a variety, then we know only that $Cg\, E^{\mathfrak{B}}\mathscr{A} \subseteq Cr^{\mathfrak{B}}\mathscr{A}$. A counterexample is illustrated by fig. 2. Note, that $\mathfrak{Fr}_{\mathfrak{A}}\mathscr{A}$ is the relatively free extension of the partial algebra $\mathfrak{A}$ with respect to $\mathscr{A}$, in the usual sense.

3. Sink-complemented subalgebras

Before returning to the extensions of partial algebras we have to discuss a very important notion which plays a central role in our arguments: this is the notion of sink-complemented subalgebras (in short Sc-s) of an algebra. In this part of the article by an algebra we mean a partial algebra.

DEFINITION 4. $\mathfrak{B}$ is a sink-complemented subalgebra of $\mathfrak{A}$, in symbols $\mathfrak{B} \overline{\subseteq} \mathfrak{A}$:

$$\mathfrak{B} \overline{\subseteq} \mathfrak{A} \overset{d}{\Leftrightarrow} (\forall \alpha)[\alpha \text{ really depends} \Rightarrow (\alpha^{(\mathfrak{A})}(a_1, \ldots, a_n) \in B \text{ iff } a_1, \ldots, a_n \in B)].$$

The importance of Sc-s might be underlined by the fact that in semilattices the concept of ideal coincides with the notion of Sc.

Figure 3.

Picking out an Sc from an algebra means a nice "feedback — free" decomposition of the algebra into two parts: the Sc can influence its complement which however cannot influence the Sc back. (See fig. 3.) It is easily seen, that an algebra has a Sc if and only if it has the algebra of the form

(of the same type of course) as homomorphic image. Fig. 4. shows how the Sc-s define a partial ordering on the subalgebras of an algebra.

We do not discuss here some important properties of Sc-s because the purpose of our paper is different. A few theorems are listed below without proofs. (The proofs are straightforward).

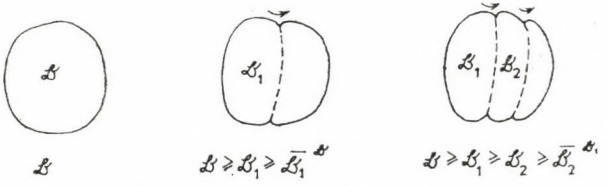


Figure 4.

THEOREM 4. *The relation* $\overline{\overline{\subseteq}}$ *is a partial ordering.*

THEOREM 5. $\mathfrak{C}, \mathfrak{B} \overline{\overline{\subseteq}} \mathfrak{A} \Rightarrow \mathfrak{C} \cap \mathfrak{B} \overline{\overline{\subseteq}} \mathfrak{B}$.

THEOREM 6. *The Sc-s of an algebra* $\mathfrak{A}$ *form an inductive closed-set system: moreover a subalgebra-representation for this i.c.s. system can be obtained by enriching* $\mathfrak{A}$ *with new operations constructed from the old ones.*

THEOREM 7. *Every finitely generated Sc can be generated by one element.*

THEOREM 8.

a) *A Sc of* $\mathfrak{A}$ *has common elements with every generator of* $\mathfrak{A}$.

b) $\mathfrak{B} \overline{\overline{\subseteq}} \mathfrak{A} \Rightarrow (Sg^{(\mathfrak{A})}X) \cap B = Sg^{(\mathfrak{B})}(X \cap B)$

Corollary:

c) $\mathfrak{B} \overline{\overline{\subseteq}} \mathfrak{A} \Rightarrow (Cg^{(\mathfrak{A})}X) \restriction B = Cg^{(\mathfrak{B})}(X \restriction B)$

THEOREM 9. $\mathfrak{B} \overline{\overline{\subseteq}} \mathfrak{A} \Rightarrow \mathfrak{Fr}_{\mathfrak{B}} \overline{\overline{\subseteq}} \mathfrak{Fr}_{\mathfrak{A}}$.

Now we return to the free extensions of partial algberas.

4. The main result

Our definitions for $\mathfrak{Fr}_{\mathfrak{N}}$ and $\mathfrak{Fr}_{\mathfrak{N}}\mathscr{V}$ coincide (up to isomorphism) with the categorial minded definitions (see e.g. GRÄTZER [3]). In the same time however $\mathfrak{Fr}_{\mathfrak{N}}$ and $\mathfrak{Fr}_{\mathfrak{N}}\mathscr{V}$ are concrete, explicit constructions (in the spirit of e.g. TARSKI [4]) as opposed to the categorial definitions which are not.

The constructions of $\mathfrak{Fr}_{\mathfrak{N}}$ and $\mathfrak{Fr}_{\mathfrak{N}}\mathscr{V}$ are completely analogous to the usual constructions of word algebras (and relatively free algebras as factor algebras of word algebras), generated by a set, moreover the latters are special cases of the firsts.

However there are also important differences in the behaviour of the new constructions. (These stem from the fact that $\mathfrak{N}$ is structured and so in essence $\mathfrak{Fr}_{\mathfrak{N}}$ is a free algebra, generated by N, under certain defining relations.) The following theorem states, for example, that an important property of ordinary free algebras holds also for free algebras generated by partial algebras. (This property is the equivalence of the categorial and constructive definitions.)

THEOREM 10. *Let* $\mathscr{V}$ *be a variety. Then* $\mathfrak{A} \cong \mathfrak{Fr}_{\mathfrak{N}}\mathscr{V}$ *iff*

$$\mathfrak{A} \in \mathscr{V} \text{ and } (\exists \mathfrak{N} \underset{h}{\simeq} \mathfrak{A})[Sg^{(\mathfrak{A})}\{h(x) : x \in N\} = \mathfrak{A} \,\&\, (\forall \mathfrak{N} \underset{\psi}{\simeq} \mathfrak{C} \in \mathscr{V})(\exists \mathfrak{A} \underset{\varphi}{\simeq} \mathfrak{C}) h|\varphi = \psi],$$

where $h|\varphi$ *denotes the composition of* h *and* φ.

On the other hand, e.g. it is not true for the relatively free algebra generated by $\mathfrak{N}$ that the factoring congruence can not relate the elements of N to each other, while it is true for the relatively free algebra generated by a set. Now we state a sufficient and necessary condition for this property of the factoring congruence to hold also for partial algebras. (See fig. 5.)

THEOREM 11. *Let $\mathscr{V}$ be a variety.*

$$(Cr^{\mathfrak{Fr}_{\mathfrak{N}}}\mathscr{V}) \restriction N = Id_N \Leftrightarrow [\mathscr{V} \Vdash \alpha = \beta \text{ and } \alpha^{(\mathfrak{N})}(a_1, \ldots, a_n) \text{ is defined} \Rightarrow$$

$$\Rightarrow \beta^{(\mathfrak{N})}(a_1, \ldots, a_n) \text{ is defined and } \alpha^{(\mathfrak{N})}(a_1, \ldots, a_n) = \beta^{(\mathfrak{N})}(a_1, \ldots, a_n)].$$

PROOF. The above condition is easily seen to be equivalent with the condition that $(E^{\mathfrak{Fr}_{\mathfrak{N}}}\mathscr{V}) \restriction N = Id_N$. Because $\mathfrak{N}$ is a Sc of $\mathfrak{Fr}_{\mathfrak{N}}$ and $\mathscr{V}$ is a variety, from Th. 3 and Th. 8c we have:

$$(Cr^{\mathfrak{Fr}_{\mathfrak{N}}}\mathscr{V}) \restriction N = (Cg^{(\mathfrak{Fr}_{\mathfrak{N}})} E^{\mathfrak{Fr}_{\mathfrak{N}}}\mathscr{V}) \restriction N = Cg^{(\mathfrak{N})}(E^{\mathfrak{Fr}_{\mathfrak{N}}}\mathscr{V} \restriction N) = Cg^{(\mathfrak{N})} Id_N = Id_N. \quad \blacksquare$$

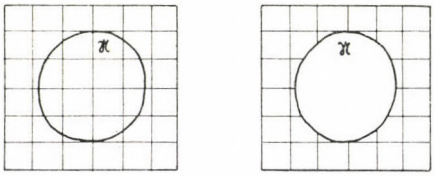


Figure 5.

Intiutively speaking, the above theorem states that if $\mathfrak{N}$ agrees (in the stronger sense) with the required equations then the extension of $\mathfrak{N}$ proceeds just in the style one always has dreamed of: we augment $\mathfrak{N}$ with the words for the missing values of the operations and then identify those of the new elements which are required. Thus starting from the partial algebra $\mathfrak{N}$ we have just naturally constructed the missing part of it.

THEOREM 12. (The principal theorem) *Let $\mathfrak{N}$ be a regular variety and* $(Cr^{\mathfrak{Fr}_{\mathfrak{N}}}) \restriction N = Id_N$

$$\mathfrak{C} \overline{\overline{\subseteq}} \mathfrak{Fr}_{\mathfrak{N}}\mathscr{V} \Rightarrow \mathfrak{C} = \mathfrak{Fr}_{\mathfrak{N} \cap \mathfrak{C}}\mathscr{V}.$$

PROOF. Let $\mathfrak{N}' \overset{d}{=} \mathfrak{N} \cap \mathfrak{C}$; $E' \overset{d}{=} E^{\mathfrak{Fr}_{\mathfrak{N}'}}\mathscr{V}$; $E \overset{d}{=} E^{\mathfrak{Fr}_{\mathfrak{N}}}\mathscr{V}$. Then

a) $\mathfrak{N}' \overline{\overline{\subseteq}} \mathfrak{N}$,

b) $\mathfrak{Fr}_{\mathfrak{N}'} \overline{\overline{\subseteq}} \mathfrak{Fr}_{\mathfrak{N}}$,

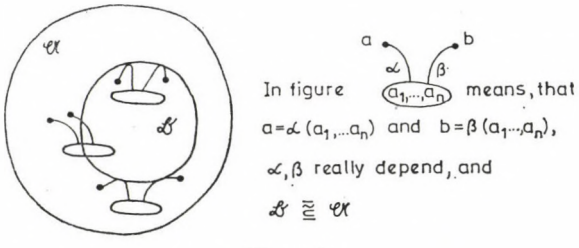


Figure 6.

c) $E' = E \restriction \mathfrak{Fr}_{\mathfrak{N}'}$, for (see fig. 6.)
$\langle a, b\rangle \in E'$ iff (because $\mathscr{V}$ is regular)

$$(\exists \mathscr{V} \Vdash \alpha = \beta;\ a_1, \ldots, a_n \in Fr_{\mathfrak{N}'};\ \alpha, \beta \text{ really depend})$$

$$a = \alpha^{(\mathfrak{Fr}_{\mathfrak{N}'})}(a_1, \ldots, a_n) \,\&\, b = \beta^{(\mathfrak{Fr}_{\mathfrak{N}'})}(a_1, \ldots, a_n) \text{ iff (because b)}$$

$$(\exists \mathscr{V} \Vdash \alpha = \beta;\ a_1, \ldots, a_n \in Fr_{\mathfrak{N}};\ \alpha, \beta \text{ really depend})$$

$$a \in Fr_{\mathfrak{N}'} \,\&\, a = \alpha^{(\mathfrak{Fr}_{\mathfrak{N}})}(a_1, \ldots, a_n) \,\&\, b = \beta^{(\mathfrak{Fr}_{\mathfrak{N}})}(a_1, \ldots, a_n) \text{ iff } \langle a, b\rangle \in E \,\&\, a \in Fr_{\mathfrak{N}'}.$$

d) $Cr^{\mathfrak{Fr}_{\mathfrak{N}'}}\mathscr{V} = (Cr^{\mathfrak{Fr}_{\mathfrak{N}}}\mathscr{V}) \restriction Fr_{\mathfrak{N}'}$, from c), Th. 8c and Th. 3.

e) $\mathfrak{Sg}^{(\mathfrak{Fr}_{\mathfrak{N}}\mathscr{V})}\mathfrak{N}' = \mathfrak{C}$.

(See fig. 7.) ∎

Now we discuss the application of this theorem to free products and especially the refinement result proved in JONSSON [1]. Free products are defined in [1] as follows: Given a variety $\mathscr{V}$ of algebras, an algebra $\mathfrak{A} \in \mathscr{V}$ is said to be a $\mathscr{V}$-free product of its subalgebras $\mathfrak{B}_i$, $i \in I$, provided $\mathfrak{A}$ is generated by the union of the sets B_i, and for any $\mathfrak{C} \in \mathscr{V}$ and system of homomorphisms $\mathfrak{B}_i \xrightarrow[f_i]{} \mathfrak{C}$ there exists $\mathfrak{A} \xrightarrow[g]{} \mathfrak{C}$ such that $f_i \subseteq g$ for all $i \in I$. A $\mathscr{V}$-free product of $\mathfrak{B}_i$, $i \in I$ is denoted by ${}^{\mathscr{V}}\prod_{i\in I}{}^{*}\mathfrak{B}_i$. It is easily seen that if ${}^{\mathscr{V}}\prod_{i\in I}{}^{*}\mathfrak{B}_i$ exists then $\mathfrak{Fr}_{\bigcup_{i\in I} B_i}\mathscr{V}$ is a free product of the system $\mathfrak{B}_i$, $i \in I$. It follows immediately from Th 11. that in regular varieties free products of algebras $\mathfrak{B}_i \in \mathscr{V}$, $i \in I$ exist if and only if the algebras $\mathfrak{B}_i$ are disjoint. From the construction of $\mathfrak{Fr}_{\bigcup_{i\in I} B_i}\mathscr{V}$ also follows that in a free product the factors are Sc-s. An advantage of the categorial minded definition of [1] is that it makes explicit the commutativity and complete associativity, while constructive definitions are advantageous e.g. in proving the refinement theorem: indeed the refinement theorem can be obtained by reformulating our principal theorem (Th 12): ${}^{\mathscr{V}}\prod_{i\in I}{}^{*}\mathfrak{B}_i = \mathfrak{C} * \mathfrak{A} \Rightarrow \mathfrak{C} = \prod^{*}(\mathfrak{B}_i \cap \mathfrak{C})$ (if $\mathscr{V}$ is a regular variety and $\mathfrak{B}_i \in \mathscr{V}$, $i \in I$.)

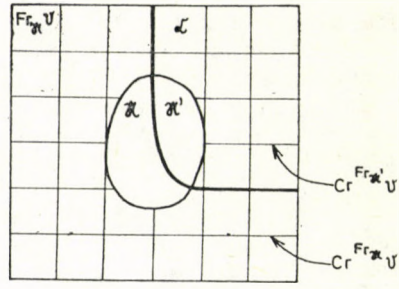


Figure 7.

5. Summing up

The chief aim of this article is to propose a simple, natural construction for ${}^{\mathscr{V}}\prod_{i\in I}{}^{*}\mathfrak{B}_i$. This construction is $\mathfrak{Fr}_{\bigcup_{i\in I} B_i}\mathscr{V}$.

In [1] a relatively free algebra with defining relations generaed by a set (in the notation of TARSKI [4]: $\mathfrak{Fr}^{(s)}_{\bigcup B_i}\mathscr{V}$, where $S = \bigcup\{Op_j^{(\mathfrak{B}_i)} : j \in In^{(\mathfrak{B}_i)}, i \in I\}$) plays the role of $\mathfrak{Fr}_{\bigcup \mathfrak{B}_i}\mathscr{V}$. That algebra $(\mathfrak{Fr}^{(s)}_{\bigcup B_i}\mathscr{V})$ however is only isomorphic to the free product,

since $\mathfrak{B}_i \cong \mathfrak{Fr}^{(s)}_{\cup B_i} \mathscr{V}$ does not hold. By using $\mathfrak{Fr}_{\cup \mathfrak{B}_i} \mathscr{V}$ we have got rid of the burden of dealing in terms of these isomorphisms since in $\mathfrak{Fr}_{\cup \mathfrak{B}_i} \mathscr{V}$ the $\mathfrak{B}_i$-s are preserved in their original form, well isolated from the other elements of the algebra. (See Th 11, and fig. 8.) The other, more important advantage of $\mathfrak{Fr}_{\cup \mathfrak{B}_i} \mathscr{V}$ to $\mathfrak{Fr}^{(s)}_{\cup B_i} \mathscr{V}$ is its more transparent structure. The source of this transparency is, that we did not start by

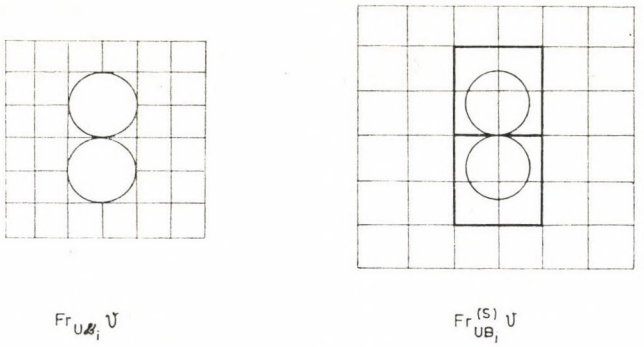


Figure 8.

abandonning the structure of $\cup \mathfrak{B}_i$ just to bring it back through the factoring congruence, which in the same time however has the job to provide for the structure of $\mathscr{V}$ also. The mixing of these two different requirements result in an opacity of the factoring congruence, which in the same time plays a central role in all the arguments. See fig. 9. However keeping in mind our goal to construct a free product this detour (with the structure of $\cup \mathfrak{B}_i$) is superflouous. As a result the factoring congruence needs only to provide for the structure of $\mathscr{V}$, and so the case distinctions in the proofs become superfluous. We got rid of the induction by using properties of the Sc-s. (Th 8. c.)

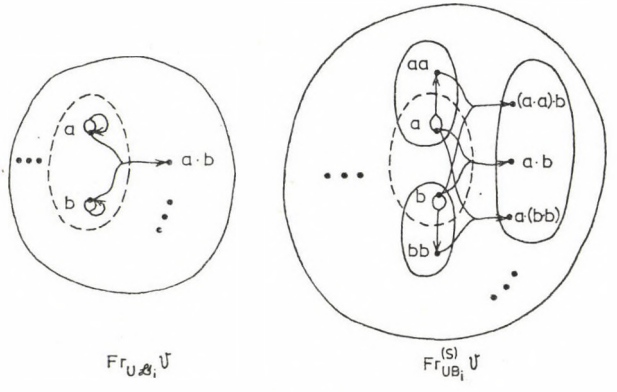


Figure 9.

REFERENCES

[1] JONSSON, B.: *Relatively free products in regular varieties,* Internal Report, Vanderbilt Univ., 1973.
[2] JONSSON, B.: Relatively free products in regular varieties, *Notices Amer. Math. Soc.* **20** (1973).
[3] GRÄTZER, G.: *Universal algebra,* Van Nostrand, 1968.
[4] HENKIN, L., MONK, J. D., TARSKI, A.: *Cylindric algebras,* North Holland, 1971.

MTA MAT. Kut. Int., 1053 Budapest, Reáltanoda u. 13—15, Hungary

(Received November 10, 1973)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 33—38.

CROSSINGS AND TOUCHINGS IN A RESTRICTED RANDOM WALK

by

K. G. ANEJA and KANWAR SEN

1. Introduction. In this paper, we derive some distributions concerning ENGELBERG's [3] restricted random walk (r.r.w.) where a particle located at the origin in the (t, y) plane at the epoch $t=0$ moves only at the epochs $t=1, 2, \ldots, a+b$ $(a>b)$; the movement at each epoch being a unit positive step or a unit negative step along the y-axis subject to restriction A, namely that the walk terminates at the point $(a+b, a-b)$. We consider a sequence of auxiliary random variables $\xi_1, \xi_2, \ldots, \xi_{a+b}$ such that ξ_i $(i=1, 2, \ldots, a+b)$ can assume only one of the values $+1$ or -1 according to whether the ith step in the r.r.w. is upward or downward. Let $S_i=\xi_1+\xi_2+\ldots+\xi_i$ and $S_0=0$; then restriction A implies that $S_{a+b}=a-b$. On representing a realization of the sequence $\{\xi_i\}$ in the (t, y) plane by a polygonal line whose ith side has slope ξ_i and whose ith vertex has ordinate S_i (or equivalently that the vertices of the polygonal line are (i, S_i); one obtains the graph of the r.r.w. which we call a path. It is assumed that all the $\binom{a+b}{a}$ different paths from $(0, 0)$ to $(a+b, a-b)$ are equally probable. However, no assumption is made with regard to the probabilities of the particle moving upward or downward at any individual step since our interest is limited to deriving some conditional distributions subject to restriction A. This is facilitated through the technique of generating functions (see ROSENSTOCK [7]) by which we derive some joint distributions concerning r.r.w. and extend the connected results due to ENGELBERG [3]. Besides we state and prove two new equivalence relations. The following definitions and notations are adopted:

Return. A return (to t-axis) for the path $\{S_i\}$ occurs at an even index i for which $S_i=0$.

Sojourn. The segment of path between the origin and the first return and that between any two consecutive returns is called a sojourn. Let ϱ_j denote the index of jth return and assume that $\varrho_0=0$; then the sojourn from ϱ_{j-1} to ϱ_j is positive or negative according as $S_i>0$ or $S_i<0$ for $\varrho_{j-1}<i<\varrho_j$ $(j\geqq 1)$.

Crossing. A crossing of the t-axis occurs at an even index i for which $S_i=0$ and $S_{i-1}\cdot S_{i+1}=-1$.

Wave. A segment of the path $\{S_i\}$ included between two consecutive crossings is called a wave [1]; the segments from the origin to the first crossing and from the last crossing to the last return also being regarded as waves. A wave is positive (negative) when all the sojourns which constitute it are positive (negative).

Touching. A touching of the t-axis occurs at an even index i for which $S_i=0$ and $S_{i-1}\cdot S_{i+1}=+1$; the touching being positive (or negative) when both S_{i-1} and S_{i+1} equal $+1$ (or -1).

With O as origin and Q as $(a+b, a-b)$, if P be the point of last return to t-axis in a path OQ; then the segment PQ can be regarded as a first passage through P as viewed from Q towards P.

The number of returns $Z_{a,b}$ in the path OQ (in fact in the segment OP) is also the number of sojourns in the segment OP. Let $Z^+_{a,b}$ and $Z^-_{a,b}$ denote respectively the number of positive and negative sojourns amongst $Z_{a,b}$.

Denoting the number of crossings in the path OQ and segment OP by $C_{a,b}$ and $C^*_{a,b}$ respectively we see that $C_{a,b}-C^*_{a,b}$ is either 0 or 1.

We denote the number of waves included in the segment OP by $W_{a,b}$ of which $W^+_{a,b}$ are positive and the remaining $T^-_{a,b}$ negative; then obviously $W_{a,b}=C^*_{a,b}+1$.

We denote the total number of touchings in the path OQ by $T_{a,b}$, of which $T^+_{a,b}$ are positive and the remaining $T^-_{a,b}$ negative. The corresponding total number of touchings in the segment OP is denoted by $\Gamma_{a,b}$ of which $\Gamma^+_{a,b}$ are positive and the remaining $\Gamma^-_{a,b}$ negative. Then obviously, $T_{a,b}-\Gamma_{a,b}\equiv T^+_{a,b}-\Gamma^+_{a,b}$, is either 0 or 1.

2. Generating Functions (GF). The GF [4] of the number of steps up to the first return to t-axis with $p=q=\frac{1}{2}$ is $F(s)=1-(1-s^2)^{1/2}$ and that for the first passage through r is $[F(s)/s]^r$; hence the GF of the length of the segment PQ is

$$[F(s)/s]^{a-b} = [1-(1-s^2)^{\frac{1}{2}}]^{a-b}\cdot s^{-a+b}. \tag{1}$$

Let $F^+(s)$ denote the GF of the first return time to t-axis through positive values or equivalently the GF of the length of a positive sojourn and let $F^-(s)$ denote the GF of the length of a negative sojourn, then

$$F^+(s) = F^-(s) = \frac{1}{2}F(s) \tag{2}$$

from which by elementary algebra, we get

$$[1-F^+(s)]^{-1} = 2s^{-2}F(s). \tag{3}$$

Regarding $a+b$ as a random variable but $a-b$ as fixed, er now determine a few generating functions.

(i) *Even Number of Crossings.* Let $G(s;\ T_{a,b}=u,\ T^+_{a,b}=v,\ C_{a,b}=2h)$ denote the GF of the number of steps in the paths entailing $2h$ crossings and u touchings of which v are positive. A path OQ with these characteristics is possible in either of the following two mutually exclusive cases for both of which $S_1=+1$.

Case (1). The last return to t-axis at P is a touching, so that

$$C_{a,b}=C^*_{a,b}=2h,\quad T_{a,b}=\Gamma^+_{a,b}=u \quad\text{and}\quad T^+_{a,b}=\Gamma^+_{a,b}+1=v. \qquad \text{(see fig. 1.)}$$

Case (2). The last return to t-xais at P is a crossing, so that

$$C_{a,b}=C^*_{a,b}+1=2h,\quad T_{a,b}=\Gamma_{a,b}=u \quad\text{and}\quad T^+_{a.b}=\Gamma^+_{a,b}=v. \qquad \text{(see fig. 2.)}$$

For case (1), we get

$$W_{a,b} = 2h+1, \quad W^{+}_{a,b} = h+1, \quad W^{-}_{a,b} = h, \quad Z^{+}_{a,b} = v+h \quad \text{and} \quad Z^{-}_{a,b} = u-v+h.$$

By a well-known result on occupancy problems FELLER [4], the number of ways of arranging the $v+h$ positive and $u-v+h$ negative sojourns amongst $h+1$ positive and h negative waves is $N \equiv \binom{v+h-1}{h}\binom{u-v+h-1}{h-1}$; hence the GF of the length of segment

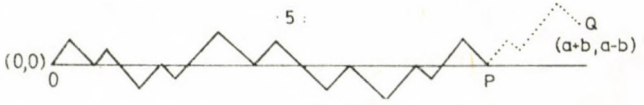


Fig. 1. Showing path OQ for which: $S_1 = +1$, the last return to t-axis at P is a touching, and $C^{*}_{a,b} = C_{a,b} = 2h = 4$, $T_{a,b} = \Gamma_{a,b} + 1 = u = 6$, $T^{+}_{a,b} = \Gamma^{+}_{a,b} + 1 = v = 3$

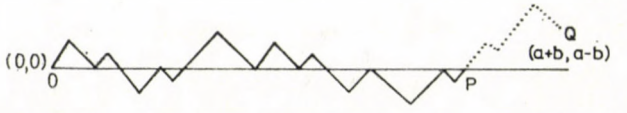


Fig. 2. Showing path OQ for which: $S_1 = +1$, the last return to t-axis at P is a crossing, and $C_{a,b} = C^{*}_{a,b} + 1 = 2h = 4$, $T_{a,b} = \Gamma_{a,b} = u = 6$, $T^{+}_{a,b} = \Gamma^{+}_{a,b} = v = 3$

OP by (2) is $N\left[\frac{1}{2}F(s)\right]^{(v+h)+(u-v+h)}$ which on multiplication by (1) gives the GF of the length of path OQ as

$$\binom{v+h-1}{h}\binom{u-v+h-1}{h-1}\left[\frac{1}{2}F(s)\right]^{u+2h}\left[\frac{F(s)}{s}\right]^{a-b}. \tag{4}$$

For case (2), we get

$$W_{a\,b} = 2h, \quad W^{+}_{a,b} = W^{-}_{a,b} = h, \quad Z^{+}_{a,b} = v+h \quad \text{and} \quad Z^{-}_{a,b} = u-v+h,$$

and proceeding similarly the GF of the length of path OQ is found to be

$$\binom{v+h-1}{h-1}\binom{u-v+h-1}{h-1}\left[\frac{1}{2}F(s)\right]^{u+2h}\cdot\left[\frac{F(s)}{s}\right]^{a-b}. \tag{5}$$

Adding (4), (5) and using (1) yields

$$G(s; T_{a,b} = u, T^{+}_{a,b} = v, C_{a,b} = 2h) = \tag{6}$$

$$= \binom{v+h}{h}\binom{u-v+h-1}{h-1}\left[1-(1-s^2)^{\frac{1}{2}}\right]^{u+2h+a-b}\cdot 2^{-(u+2h)}\cdot s^{-a+b}.$$

Summing (6) over u from v to ∞ by

$$\sum_{i=0}^{\infty}\binom{-1-\mu}{i}(-x)^i = (1-x)^{-1-\mu}$$

and using (3) we get

$$(7)\quad G(s; T^+_{a,b}=v, C_{a,b}=2h) = \binom{v+h}{h}\left[1-(1-s^2)^{\frac{1}{2}}\right]^{v+3h+a-b}\cdot 2^{-(v+h)}\cdot s^{-a+b-2h}$$

(ii) *Odd Number of Crossings.* A realization of the path OQ contributing to $G(s; T_{a,b}=u, T^+_{a,b}=v, C_{a,b}=2h+1)$ is possible in either of the following two mutually exclusive cases for both of which $S_1=-1$.

Case (1): The last return to t-axis at P is a touching, so that

$$C_{a,b} = C^*_{a,b} = 2h+1, \quad T_{a,b} = \Gamma_{a,b}+1 = u \quad \text{and} \quad T^+_{a,b} = \Gamma^+_{a,b}+1 = v. \text{ (see fig. 3.)}$$

Fig. 3. Showing path OQ for which: $S_1=-1$, the last return to t-axis at P is a touching, and $C_{a,b}=C^*_{a,b}=2h+1=5$, $T_{a,b}=\Gamma_{a,b}+1=u=6$, $T^+_{a,b}=\Gamma^+_{a,b}+1=v=3$

Case (2): The last return to t-axis at P is a crossing, so that

$$C_{a,b} = C^*_{a,b}+1 = 2h+1, \quad T_{a,b} = \Gamma_{a,b} = u \quad \text{and} \quad T^+_{a,b} = \Gamma^+_{a,b} = v. \quad \text{(see fig. 4.)}$$

For case (1), we get

$$W_{a,b} = 2h+2, \quad W^+_{a,b} = W^-_{a,b} = h+1, \quad Z^+_{a,b} = v+h \quad \text{and} \quad Z^-_{a.b} = u-v+h+1,$$

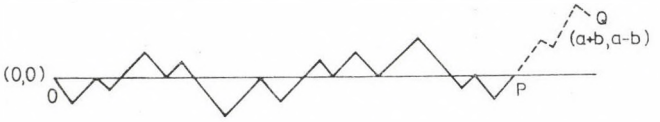


Fig. 4. Showing path OQ for which $S_1=-1$, the last return to t-axis at P is a crossing, and $C_{a,b}=C^*_{a,b}+1=2h+1=5$, $T_{a,b}=\Gamma_{a,b}=u=6$, $T^+_{a,b}=\Gamma_{a,b}=v=3$

so that the GF of the length of the path OQ is

$$(8)\qquad \binom{v+h-1}{h}\binom{u-v+h}{h}[F(s)]^{u+2h+1}\cdot\left[\frac{F(s)}{s}\right]^{a-b}.$$

For case (2),

$$W_{a,b}=2h+1, \quad W^+_{a,b} = h, \quad W^-_{a,b} = h+1, \quad Z^+_{a,b}= v+h \quad \text{and} \quad Z^-_{a,b} = u-v+h+1$$

and the GF of the length of the path OQ is

$$\binom{v+h-1}{h-1}\binom{u-v+h}{h}[F(s)]^{u+2h+1}\cdot\left[\frac{F(s)}{s}\right]^{a-b}. \tag{9}$$

Adding (9) and (10) we get

$$G(s;\, T_{a,b}=u,\, T^{+}_{a,b}=v,\, C_{a,b}=2h+1)= \tag{10}$$

$$=\binom{v+h}{h}\binom{u-v+h}{h}[1-(1-s^2)^{\frac{1}{2}}]^{u+2h+1+a-b}\cdot 2^{-(u+2h+1)}\cdot s^{-a+b}$$

from which on summation over u, we get

$$G(s;\, T_{a,b}=v,\, C_{a,b}=2h+1)= \tag{11}$$

$$=\binom{v+h}{h}[1-(1-s^2)^{\frac{1}{2}}]^{v+3h+2+a-b}\cdot 2^{-(v+h)}\cdot s^{-a+b-2h-2}.$$

3. Probability Distributions. For deriving the relevant probability distributions from the generating functions obtained in the previous section, we need the following results:

Power series expansion [2] valid for a positive integer λ

$$[1-(1-s^2)^{\frac{1}{2}}]^{\lambda}=2^{\lambda}\sum_{i=\lambda}^{\infty}\left(\frac{1}{2}s\right)^{2i}A_{i-\lambda}(\lambda,2) \tag{12}$$

where [5]: $$A_i(\lambda,k)=\frac{\lambda}{\lambda+ik}\binom{\lambda+ik}{i};$$

and combinatorial identities [5]:

$$\sum_{v=0}^{u}\binom{\alpha+u-v-1}{u-v}\binom{\beta+v-1}{v}=\binom{\alpha+\beta+u-1}{u}, \tag{13}$$

$$\sum_{i=0}^{m}(g+ik)\binom{x-1+i}{i}\binom{y-1-i}{y-m-1}=\frac{(x+y-m)g+mxk}{x+y}\binom{x+y}{m}. \tag{14}$$

The coefficient of s^{a+b} in the expansions of (6), (7) and (10), (11) by (12) when divided by the total number $\binom{a+b}{a}$ of possible paths from $(0,0)$ to $(a+b,\, a-b)$ gives the various conditional probabilities under restriction A viz. $S_{a+b}=a-b$. These on writing for convenience

$$\alpha=b-u-2h \quad \text{and} \quad \beta=b-v-2h$$

are

$$\mathrm{P}(T_{a,b}=u,\, T^{+}_{a,b}=v,\, c=2h)=\binom{v+h}{h}\binom{u-v+h-1}{h-1}A_{\alpha}(a-\alpha,2)/\binom{a+b}{a}; \tag{15}$$

(16)

$$\mathrm{P}(T_{a,b}=u,\, T^{+}_{a,b}=v,\, c=2h+1)=\binom{v+h}{h}\binom{u-v+h}{h}A_{\alpha-1}(a-\alpha+1,2)/\binom{a+b}{a};$$

$$\mathrm{P}(T^{+}_{a,b}=v,\, c_{a,b}=2h)=\binom{v+h}{h}A_{\beta}(a+h-\beta,2)/\binom{a+b}{a}; \tag{17}$$

and

$$\text{(18)} \qquad \mathsf{P}(T^{+}_{a,b} = v,\, c_{a,b} = 2h+1) = \binom{v+h}{h} A_{\beta-1}(a+h+2-\beta, 2)/\binom{a+b}{a}.$$

In the special case $h=0$, (17) has been obtained by ENGELBERG [3]. Summing respectively (15) and (16) over v from 0 to ∞ by means of (13), we get

$$\mathsf{P}(T_{a,b} = u,\, c_{a,b} = 2h) = \binom{u+2h}{u} A_{\alpha}(a-\alpha, 2)/\binom{a+b}{a}$$

and

$$\mathsf{P}(T_{a,b} = u,\, c_{a,b} = 2h+1) = \binom{u+2h+1}{u} A_{\alpha-1}(a-\alpha+1, 2)/\binom{a+b}{a}$$

which both combine to give

$$\text{(19)} \qquad \mathsf{P}(T_{a,b} = u,\, c_{a,b} = t) = \frac{a-b+u+t}{a+b-u-t}\binom{a+b-u-t}{b-u-t}\binom{u+t}{t}\Big/\binom{a+b}{a}.$$

This on summation over u by means of (14) verifies the known result ([3], [6])

$$\text{(20)} \qquad \mathsf{P}(c_{a,b} = t) = \frac{a-b+2t+1}{a+b+1}\binom{a+b+1}{a+t+1}\Big/\binom{a+b}{a}.$$

4. Two Equivalence Relations. Since (19) is invariant for interchange of u and t, we get the equivalence relation

$$\mathsf{P}(T_{a,b} = u,\, c_{a,b} = t) = \mathsf{P}(T_{a,b} = t,\, c_{a,b} = u)$$

which on summation over u gives

$$\mathsf{P}(c_{a,b} = t) = \mathsf{P}(T_{a,b} = t).$$

Acknowledgment. The authors are greatly indebted to Dr. H. C. GUPTA, Professor of Mathematical Statistics, University of Delhi for his guidance in these investigations.

REFERENCES

[1] CSÁKI, E. and VINCZE, I.: Two joint distribution laws in the theory of order statistics. *Mathematica* (Cluj), **5** (1963) 27—37.
[2] DWASS, MEYER: Simple Random walk and rank order statistics *Ann. Math. Statist.* **38** (1967) 1042—1053.
[3] ENGELBERG, O.: On some problems concerning a restricted random walk. *J. Appl. Prob.* **2** (1965) 396—404.
[4] FELLER, W.: *An Introduction to Probability Theory and Its Applications 1*, (3rd ed). John Wiley, New York, 1968.
[5] GOULD, H. W.: Some Generalizations of Vandermondes' convolution. *Amer. Math. Month.* **63** (1956) 84—91.
[6] KANWAR SEN (1964): On some combinatorial relations concerning the symmetric random walk. *Magyar Tud. Akad. Mat. Kutató. Int. Közl.* **9** (1964) 335—357.
[7] ROSENSTOCK, H. B. (1968): Level touchings in a random walk *SIAM J. Appl. Math.* **16** (1968) 1130—1131.

Institute of Agricultural Research Statistics
University of Delhi

(Received March 15, 1972)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 39—45.

ROBUSTNESS OF MANN-WHITNEY-WILCOXON TEST TO DEPENDENCE IN THE VARIABLES

by

Z. GOVINDARAJULU*

Abstract. Let (X, Y) have an unknown bivarite distribution function $H(x, y)$ having continuous marginals $F(x)$ and $G(y)$. The Mann-Whitney-Wilcoxon test statistic can be studentized so as to be asymptotically distribution-free for testing H_0: $F(x)=G(x)$, for all x against the alternative H_1: $F \geqq G$ (with strict inequality for some x). The test is consistent and its asymptotic efficiency relative to the t-test is evaluated and an explicit form for it is obtained when $H(x, y)$ is bivariate normal with correlation coefficient ϱ. The relative efficiency is $3/\pi$ when $\varrho=-1$ or 0, is increasing for $-1 \leqq \varrho < -.5$, decreasing for $-.5 < \varrho \leqq 1$ and is equal to $3/2=.866$ when $\varrho=1$.

1. Introduction. MANN-WHITNEY [4] have proposed a distribution-free test for H_0 against H_1 when X and Y are independent. It is of much interest to study the sensitivity of the test when X and Y are dependent having an unknown bivariate distribution function $H(x, y)$ with continuous marginals $F(x)$ and $G(y)$. Let (X_i, Y_i), $i=1, \ldots, n$ denote a random sample of size n from $H(x, y)$. Also, let $H_n(x, y)$, $F_n(x)$ and $G_n(Y)$ respectively denote the empirical distribution functions (e.d.f.'s) based on the samples (X_i, Y_i) $(i=1, \ldots, n)$, $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$. Let $Z_{ij}=1$ or 0 according as $X_i \leqq Y_j$ or $X_i > Y_j$ respectively for $1 \leqq i, j \leqq n$.

2. An Asymptotically Distribution-free Test. Define

$$U = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} Z_{ij} = \int_{-\infty}^{\infty} F_n(x)\, dG_n(x). \tag{1}$$

Then, we have the following result pertaining to the asymptotic normality of U.

THEOREM 2.1. *With the above notation, for all continuous F and G we have*

$$\lim_{n \to \infty} \mathrm{P}\{n^{1/2}(U-p)/\sigma \leqq z\} = \Phi(z), \tag{2}$$

where

$$\begin{aligned} \sigma^2 = {} & 2 \iint_{x<y} F(x)[1-F(y)]\, dG(x)\, dG(y) \\ & + 2 \iint_{x<y} G(x)[1-G(y)]\, dF(x)\, dF(y) \\ & - 2 \int_{-\infty}^{\infty} \int [H(x, y) - F(x)G(y)]\, dG(x)\, dF(y), \end{aligned} \tag{3}$$

$p=\int F dG$ and Φ denotes the standard normal distribution function.

* Part of this research was conducted while the author was a visiting professor at the University of Michigan. This research was in part supported by the Navy under the Office of Naval Research Contract No. N00014—73—A—0385—0001, Task Order NR042—295.

PROOF. We shall proceed as in GOVINDARAJULU [1] and write

$$n^{1/2}(U-p) = n^{1/2}\int (F_n - F)\, dG + n^{1/2}\int F d(G_n - G) + n^{1/2}\int (F_n - F)\, d(G_n - G).$$

Integrating by parts once in the second integral we obtain

$$(4)\quad n^{1/2}(U-p) = n^{1/2}\left[\int (F_n - F)\, dG - \int (G_n - G)\, dF\right] + n^{1/2}\int (F_n - F)\, d(G_n - G).$$

$$= B_n + C_n$$

where

$$B_n = n^{1/2}\left[\int (F_n - F)\, dG - \int (G_n - G)\, dF\right]$$

denotes the sum of n independent and identically distributed random variables. Thus B_n has an asymptotically normal distribution with variance σ^2 where

$$\sigma^2 = \operatorname{Var} B_n = \operatorname{Var}\left[\int (F_1 - F)\, dG - \int (G_1 - G)\, dF\right]$$

$$= \mathsf{E}\left[\int (F_1 - F)\, dG\right]^2 + \mathsf{E}\left[\int (G_1 - G)\, dF\right]^2$$

$$- 2\mathsf{E}\int\int [F_1(x) - F(x)][G_1(y) - G(y)]\, dG(x)\, dF(y)$$

and $F_1(x)$ and $G_1(y)$ are e.d.f.'s based on a single X and a single Y respectively. Noting that

$$\mathsf{E}\{F_1(x) - F(x)\}\{F_1(y) - F(y)\} = F(x)[1 - F(y)\} \quad \text{for} \quad x \leqq y$$

and

$$\mathsf{E}\{F_1(x) - F(x)\}\{G_1(y) - G(y)\} = H(x, y) - F(x)\, G(y),$$

we have,

$$\operatorname{Var} B_n = \sigma^2 \quad \text{where } \sigma^2 \text{ is given by (3).}$$

By integrating by parts once in the first term, one would get $n^{1/2}[\int F d(G_n - G) - \int G d(F_n - F)]$ for the first order random term. Consequently an alternative form for σ^2 is

(5)

$$\sigma^2 = \int F^2\, dG - p^2 + \int G^2\, dF - (1-p)^2 - 2\int_{-\infty}^{\infty}\int F(y)\, G(x)\, d[H(x, y) - G(y)\, F(x)]$$

$$= \int F^2\, dG + \int G^2\, dF - 2\iint H(x, y)\, dF(y)\, dG(x) - (1-2p)^2,$$

since

$$\iint F(y)[1 - G(x)]\, dH(x, y) = \mathsf{P}[X_1 < Y_2, X_2 < Y_3)$$

$$= \mathsf{P}(X_2 < Y_3) - \mathsf{P}(X_1 > Y_2, X_2 < Y_3)$$

$$= p - \iint H(x, y)\, dF(y)\, dG(x).$$

Thus, $n^{1/2}(U-p)/\sigma$ is asymptotically standard normal provided C_n tends to zero in probability. Using deep results in weak convergence*, one can show that C_n goes to

* I thank Professor M. B. WOODROOFE of the University of Michigan for pointing out this possibility.

zero in probability. However, in the following, we will assert the same using very elementary arguments. If $\mathsf{E}C_n^2$ tends to zero as $n \to \infty$, then one can assert via the Chebyshev's inequality that C_n tends to zero in probability. Towards this, one can rewrite C_n as

$$C_n = n^{-3/2} \sum_i^n \sum_j^n a_{ij} = n^{-3/2} \left[\sum_{i=1}^n a_{ii} + \sum_{i \neq j} \sum a_{ij} \right]$$

where $a_{ij} = \chi_i(Y_j) - F(Y_j) - \int \{\chi_i(y) - F(y)\}\, dG(y)$ and $\chi_i(x) = 1$ if $X_i \leqq x$ and zero otherwise. Then

$$\begin{aligned} \mathsf{E}C_n^2 &= n^{-3}\,\mathsf{E}\left[\left(\sum_{i=1}^n a_{ii}\right)^2 + \left(\sum_{i \neq j} \sum a_{ij}\right)^2 + 2\left(\sum_{i=1}^n a_{ii}\right)\left(\sum_{j \neq k} \sum a_{jk}\right)\right] \\ &= n^{-2}[\mathsf{E}(a_{11}^2) + (n-1)(\mathsf{E}a_{11})^2 + 2(n-1)\,\mathsf{E}(a_{12}^2) + (n-1)\,\mathsf{E}a_{11}a_{12} \\ &\qquad + (n-1)\,\mathsf{E}a_{11}a_{21}] \\ &\quad + (n-1)(n-2)n^{-2}[\mathsf{E}a_{12}a_{13} + \mathsf{E}a_{12}a_{31} + \mathsf{E}a_{12}a_{32} + \mathsf{E}a_{12}a_{23}]. \end{aligned}$$

Straightforward computations yield

$$\mathsf{E}a_{12}a_{13} = -\int (1-G)^2\, dF + 2\int F(1-G)\, dG = 0,$$

after integrating by parts once in the first integral. Also,

$$\mathsf{E}a_{12}a_{31} = p^2 - \int F^2\, dG,$$

$$\mathsf{E}a_{12}a_{32} = \int F^2\, dG - p^2,$$

and

$$\mathsf{E}a_{12}a_{23} = 0.$$

Thus

$$\mathsf{E}C_n^2 = n^{-1}[(\mathsf{E}a_{11})^2 + 2\mathsf{E}(a_{12}^2) + \mathsf{E}a_{11}a_{12} + \mathsf{E}a_{11}a_{21}] + O(n^{-2}) = O(n^{-1}) \tag{6}$$

since each of the expectations occurring in (6) is finite. This completes the proof of Theorem 2.1. Thus, in order to test H_0 against H_1, we reject H_0 when U exceeds some k_α $(1/2 < k_\alpha < 1)$ where k_α is determined by α. Now since $F \equiv G$, $E(F(Y) - G(Y)) = 0$ if and only if $F(Y) = G(Y)$ with probability one. Thus the test is consistent against H_1. However, since σ^2 under H_0 is not free of $H(x, y)$ the test is not distribution-free. A consistent estimator of σ^2 under H_0 is given by

$$\hat{\sigma}^2 = 2/3 - 2n^{-2} \sum_{i=1}^n \sum_{j=1}^n H_n(Y_i, X_j). \tag{7}$$

Thus, using $\hat{\sigma}^2$ in place of σ^2 in Theorem 2.1, one can construct an asymptotically distribution-free test of H_0 against H_1. That is,

$$k_\alpha \approx (1/2) - \hat{\sigma}\Phi^{-1}(\alpha)/n^{1/2}. \tag{8}$$

One can obtain from (5) a consistent estimator of σ^2 under H_1 given by

$$\hat{\hat{\sigma}}^2 = n^{-1} \sum_{i=1}^n [F^2(Y_i) + G^2(X_i)] - n^{-2}\left[2 \sum_i \sum_j H_n(Y_i, X_j) + \left\{n - 2 \sum_{i=1}^n F_n(Y_i)\right\}^2\right]. \tag{9}$$

Notice that under H_0, $p \cong \int G dG = 1/2$.

Certain Remarks: (i) Let $T=[n(n-1)]^{-1} \sum\sum_{i \neq j} Z_{ij}$. Consider

$$U-T = \{-1/n^2(n-1)\} \sum\sum_{i \neq j} Z_{ij} + n^{-2} \sum_{i=1}^{n} Z_{ii}.$$

Thus,

$$\sqrt{n}|U-T| \leqq 2/\sqrt{n} \to 0 \quad \text{as} \quad n \to \infty.$$

Consequently, T is asymptotically equivalent to U.

(ii) The assumption of continuity of F and G is somewhat strong. If $H(x, y)$ is discrete in either x or y or in both x and y it can be made continuous by the continuation process or a modificitation of it described in [2]. Moreover, U is well-defined even if X and Y have common discontinuities.

(iii) If n is random and there exists a positive integer N such that n/N converges to λ $(0<\lambda<\infty)$ in probability, then

$$\mathsf{P}[(\lambda N)^{1/2}(U-p)/\sigma \leqq z] \to \Phi(z) \quad \text{as} \quad N \to \infty. \tag{10}$$

(iv) If one wishes to test H_0 against $H_2: F \leqq G$, one should interchange the roles of F and G in the test procedure for H_0 against H_1.

3. Asymptotic Efficiency. In this section, we shall evaluate the Pitman efficiency of U with respect to location shift. Let $G(x)=F(x-\theta)$, $\theta=\xi/n^{\delta}$, ξ, $\delta>0$ and F be continuously differentiable. Then, since $\frac{\partial}{\partial\theta}\mathsf{E}(U) = \int_{-\infty}^{\infty} f(x+\theta)\,dF(x)$, the efficacy of U is given by

$$e(U) = \Big[\int_{-\infty}^{\infty} f^2(x)\,dx\Big]^2/A \tag{11}$$

where

$$A = (2/3) - 2\int_0^1 \int_0^1 H\big(F^{-1}(u), F^{-1}(v)\big)\,du\,dv. \tag{12}$$

Now, if $\mathsf{E}X=\theta_1$, $\mathsf{E}Y=\theta_2$, Var $X=$Var $Y=\tau^2$ and Cov$(X, Y)=\varrho\tau^2$ then for unknown τ and ϱ, the likelihood ratio test of $H_0: \theta_1=\theta_2$ for a bivariate normal sample would be the one-sample t-test based on the differences $d_i=Y_i-X_i$ $(i=1, \ldots, n)$. Consider the t-statistic, $t=\bar{d}/s_d$, where $\bar{d}=\bar{Y}-\bar{X}$ and $s_d^2=(n-1)^{-1}\sum_{i=1}^{n}(d_i-\bar{d})^2$. Then, t is asymptotically equivalent to

$$\tilde{t}=\bar{d}/\tau\{2(1-\varrho)\}^{1/2}. \tag{13}$$

Thus, the efficacy of t is

$$e(t) = 1/\{2(1-\varrho)\tau^2\}. \tag{14}$$

Notice that when ϱ is known the test criterion derived from likelihood ratio is asymptotically equivalent to $\tilde{t}$ given by (13). Hence, the efficiency of U relative to t is:

$$e(U,t) = \frac{2(1-\varrho)\tau^2}{A}\left[\int f^2(x)\,dx\right]^2. \tag{15}$$

It should be noted that the efficacy of the two-sample t-test is also equal to $e(t)$ given by (14). If $f(x)=\Phi(x/\tau)$ then $\int f^2(x)dx = 1/2\sqrt{\pi}\tau$. It is of interest to compute the value of A when X, Y have a bivariate normal distribution. For computing A, without loss of generality we can set $\tau=1$. Then

$$A = (2/3) - 2I(\varrho)$$

where

$$I(\varrho) = \int_0^1 \int_0^1 H(\Phi^{-1}(u), \Phi^{-1}(v))\,du\,dv.$$

Now, since $H(x,y) = \int_{-\infty}^{y} \varphi(s)\Phi\left(\frac{x-\varrho s}{\sqrt{1-\varrho^2}}\right)ds$ one can write

$$I(\varrho) = \int_{-\infty}^{\infty}\int \varphi(x)\varphi(y)\left[\int_{-\infty}^{y} \varphi(s)\Phi\left(\frac{x-\varrho s}{\sqrt{1-\varrho^2}}\right)ds\right]dx\,dy = \tag{16}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(s)[1-\Phi(s)]\Phi\left(\frac{x-\varrho s}{\sqrt{1-\varrho^2}}\right)\varphi(x)\,dx\,ds.$$

When $\varrho=0$, $e(U,t)=3/\pi$ as it should be in the independent case. When $\varrho=-1$, $I=\iint\limits_{x\gtrless -s} \varphi(s)\,[1-\Phi(s)]\,\varphi(x)\,dxds = 1/6$ and consequently,

$$e(U,t) = 3/\pi \quad \text{when} \quad \varrho = -1. \tag{17}$$

Also, when $\varrho=1$, $I=\iint\limits_{x\gtrless s} \varphi(s)\,[1-\Phi(s)]\varphi(x)dxds = 1/3$ and hence $A=0$. Thus $e(U,t)$ is of the form 0/0. Using L'Hospital's rule we obtain

$$\lim_{\varrho\to 1} e(U,t) = \left[4\pi \lim_{\varrho\to 1}\frac{\partial}{\partial\varrho} I\right]^{-1}. \tag{18}$$

Since differentiation underneath the integral sign is permissible

$$\frac{\partial}{\partial\varrho} I = (1-\varrho^2)^{-3/2}\int_{-\infty}^{\infty}\varphi(s)[1-\Phi(s)]\varphi\left(\frac{x-\varrho s}{(1-\varrho^2)^{1/2}}\right)\varphi(x)(\varrho x - s)\,dx\,ds.$$

Writing

$$\varphi\left(\frac{x-\varrho s}{(1-\varrho^2)^{1/2}}\right)\varphi(x) = (2\pi)^{-1/2}\exp\left[\frac{-(2-\varrho^2)}{2(1-\varrho^2)}\left(x-\frac{\varrho s}{2-\varrho^2}\right)^2 - \frac{\varrho^2 s^2}{2(2-\varrho^2)}\right]$$

and integrating with respect to x we obtain

$$\frac{\partial}{\partial \varrho} I = -2(2-\varrho^2)^{-3/2} \int_{-\infty}^{\infty} s\varphi(s)\varphi\left(\frac{\varrho s}{(2-\varrho^2)^{1/2}}\right)[1-\Phi(s)]\,ds \tag{19}$$

$$= -2(2-\varrho^2)^{-3/2} L \quad \text{(say)}.$$

Integrating by parts once we have

$$L = \int \varphi\left(\frac{\varrho v}{(2-\varrho^2)^{1/2}}\right)[1-\Phi(v)]\,d(-\varphi) =$$

$$= -\varrho^2(2-\varrho^2)^{-1} L - \int_{-\infty}^{\infty} \varphi\left(\frac{\varrho v}{(2-\varrho^2)^{1/2}}\right)\varphi^2(v)\,dv.$$

That is

$$L = -\left(\frac{2-\varrho^2}{2}\right)\int \varphi\left(\frac{\varrho v}{(2-\varrho^2)^{1/2}}\right)\varphi^2(v)\,dv = \tag{20}$$

$$= -(4\pi)^{-1}(2-\varrho^2)^{3/2}(4-\varrho^2)^{-1/2}.$$

Hence

$$\frac{\partial I}{\partial \varrho} = (2\pi)^{-1}(4-\varrho^2)^{-1/2}. \tag{21}$$

Thus, using (21) in (18) we have

$$\lim_{\varrho \to 1} e(U, t) = \sqrt{3}/2 = 0.866, \tag{22}$$

which is very slightly larger than 0.864, the lower bound for the asymptotic efficiency of the Mann-Whitney-Wilcoxon test procedure for shift alternatives, obtained by HODGES and LEHMANN [3]. Integrating both sides of (21) we have

$$I(\varrho) = (2\pi)^{-1} \operatorname{Arc\,sin}(\varrho/2) + c$$

and

$$I(0) = 1/4 \quad \text{implies that} \quad c = 1/4.$$

Thus

$$I(\varrho) = (2\pi)^{-1} \operatorname{Arc\,sin}(\varrho/2) + 1/4. \tag{23}$$

Notice that (23) yields $I(-1) = 1/6$ and $I(1) = 1/3$ as observed earlier. Now, using (23) in (16) for the normal case, we obtain

$$e(U, t) = (1-\varrho)\left\{\frac{\pi}{3} - 2 \operatorname{Arc\,sin}(\varrho/2)\right\}^{-1}. \tag{24}$$

From (24) we compute Table 3.1.

ϱ	$-.9$	$-.75$	$-.5$	$-.25$	0	.25	.5	.75	.9
$e(U,t)$	.959	.964	.966	.963	.955	.942	.922	.898	.874

TABLE 3.1: Asymptotic efficiency of U relative to t with normal alternatives

From Table 3.1 we surmise that $e(U,t)$ is increasing in ϱ for $-1 \leqq \varrho \leqq \varrho_0$ and is decreasing for $\varrho_0 \leqq \varrho \leqq 1$. It is of interest to evaluate ϱ_0. Setting $\partial/\partial\varrho\,\{e(U,t)\}=0$ we obtain

$$I(\varrho) = (1/3)-(1-e)\{2\pi(4-\varrho^2)^{1/2}\}^{-1}. \tag{25}$$

Trial and error method yields that $-.5 \leqq \varrho_0 \leqq -.45$ and for all practical purposes one can take $\varrho_0 = -.5$.

4. Concluding Remarks From Table 3.1 it is clear that the Mann-Whitney-Wilcoxon test procedure for location shift alternatives is robust for weakly to moderately correlated variables (that is, $-1 \leqq \varrho \leqq .5$). The asymptotic efficiency exceeds that of the independent case when $-1 \leqq \varrho \leqq -.5$. The test procedure proposed in section 2 can be construed as an asymptotically distribution-free test for equality of marginals in the bivariate distribution and its asymptotic efficiency relative to the likelihood ratio test procedure (based on the normality assymption) is given by (16).

REFERENCES

[1] GOVINDARAJULU, Z.: Distribution-free confidence bounds for $P(X < Y)$, *Ann. Inst. Statist. Math.* **20** (1968), 229—238.

[2] GOVINDARAJULU, Z., LECAM, L. and RAGHAVACHARI, M.: Generalizations of theorems of Chernoff-Savage on the asymptotic normality of test statistics, *Proc, Fifth Berkeley Symp. Math. Statist. and Prob.* **1** (1967), University of California Press, 609—638.

[3] HODGES, J. L., Jr. and LEHMANN, E. L.: The efficiency of some nonparametric competitors of the t-test, *Ann. Math. Statist.* **27** (1956), 324—335.

[4] MANN, H. B., and WHITNEY, D. R.: On a test of whether one of two random variables is stochastically larger than the other, *Ann. Math. Statist.* **18** (1947), 50—60.

University of Kentucky, Lexington Kentucky

(Received November 14, 1973)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 47—60.

LOCALLY MOST POWERFUL RANK ORDER TESTS FOR THE ONE-WAY RANDOM EFFECTS MODEL*

by

Z. GOVINDARAJULU

1. Introduction and Summary. The random effects model for the analysis of variance with the assumption of normality was considered by SCHEFFÉ (1959) (see Chapter 9). For one factor experiments the model is given by

$$X_{ij} = \mu + Y_i + \varepsilon_{ij}, \quad j = 1, \ldots, n_i \quad \text{and} \quad i = 1, \ldots, c, \tag{1.1}$$

where $\{Y_i\}$ and $\{\varepsilon_{ij}\}$ are mutually independent random variables. Assuming that the variance exists, the usual null hypothesis to be tested is

$$H_0\colon \mathsf{V}(X_{ij}) = \mathsf{V}(\varepsilon_{ij}) \quad \text{or equivalently} \quad \mathsf{V}(Y_i) = 0 \quad \text{for every } i$$

where V denotes variance, where it is further assumed that Y_i and ε_{ij} are normally distributed with zero means. The more general null hypothesis is given by

$$H_0\colon \mathsf{V}(Y_i) \leqq \theta\, \mathsf{V}(\varepsilon_{ij}) \quad \text{for all } i, j \text{ and } \theta(\theta > 0) \text{ is specified.}$$

NEYMANN (1967) has discussed the asymptotically 'optimal' $C(\alpha)$ test for one-sample problem (that is, when $n_i \equiv 1$ and $c = n$). KULKARNI (1969, 1970) has considered the problem of obtaining asymptotically 'optimal' $C(\alpha)$ tests for the c-sample problem when the underlying model is linear or nonlinear. However, the above tests do require the knowledge of the density of the underlying variables. GREENBERG (1964) has considered a more general model where the Y_i are normally distributed but the ε_{ij} have an arbitrary absolutely continuous distribution. Then, she develops partially distributionfree tests for the hypothesis quoted above. GOVINDARAJULU and DESHPANDÉ (1972) have formulated the nonparametric version of this problem and have derived locally most powerful tests foor the same under the assumption that $\mathsf{E}(Y_i)$ is not independent of i and study their asymptotic normality. In this paper we derive the locally most powerful tests when $\mathsf{E}(Y_i) = 0$ and study their asymptotic distributions under the null hypothesis.

2. Locally Most Powerful Tests. Let us consider the model given in (1.1) where, without loss of generality, we can set $\mu = 0$, and let F denote the common distribution of the ε_{ij} ($j = 1, \ldots, n_i$ and $i = 1, \ldots, c$) and $G(y)$ denote the common distribution

---

* Presented at the International Symposium on Statistical Design and Linear Models which was held at Colorado State University, Fort Collinns during March 19—23, 1973.

Research supported at the University of Kentucky in part by the Navy through ONR Contract No. N00014—73—A—0385—0001, Task Order NR 042—295.

function of the Y_i's, which are assumed to be mutually independent. Also, let $\mathsf{E}(Y_i)=0$ $(i=1, \ldots, c)$. Then, we are interested in testing

$$H_0;\ G(y) = \begin{cases} 0 & \text{for} \quad y < 0 \\ 1 & \text{for} \quad y \geqq 0 \end{cases}$$

against the alternative

$$H_1\colon\ G(y) \text{ is a non-trivial distribution function.}$$

In order to derive the locally most powerful (LMP) rank test, consider the alternative hypothesis:

$$H_\Delta\colon\ X_{ij} = \Delta Y_i + \varepsilon_{ij}, \quad \text{for small positive } \Delta.$$

Let $W_1 < \cdots < W_N$ $(N=n_1+\cdots+n_c)$ denote the combined ordered sample and let $Z=(Z_1, \ldots, Z_N)'$ denote the c-sample rank order, that is, $Z_k=i$ if $W_k=X_{il}$ for some $l=1, \ldots, n_i$. Also, let z be a possible realization of the $N!/\prod_{i=1}^{c} n_i!$ possible rank orders. Then, we are led to the following theorem.

THEOREM 2.1. *If (i) the density f has a derivative that is absolutely continuous on finite intervals,* (ii) $f'(x)$ *is continuous almost everywhere,* (iii) $\mathsf{E}Y_i^2 < \infty$ *and* (iv) $\mathrm{E}\left|\frac{\partial^2}{\partial x^2}\log f(X)\right| < \infty$, *then the locally most powerful test of H_0 against H_Δ is given by:*

Reject H_0 when

$$T = \sum_{j=1}^{c}\left\{\sum_{i\neq k}^{N}\sum^{N} E_0\left[\frac{f'(W_i)}{f(W_i)}\frac{f'(W_k)}{f(W_k)}\right]\delta_{j,z_i}\delta_{j,z_k} + \sum_{i=1}^{N} E_0\left[\frac{f''(W_i)}{f(W_i)}\right]\delta_{j,z_i}\right\} > K_\alpha \tag{2.1}$$

where K_α is determined by the level of significance α and $\delta_{i,j}$ denotes the Kronecker's delta (that is, $\delta_{i,j}=1$ if $i=j$ and zero otherwise).

PROOF. One may write

$$\mathsf{P}(Z = z|Y_i = y_i, i = 1, \ldots, c) = \left\{\sum_{i=1}^{c} n_i!\right\} \int_{-\infty<w_1<\cdots<w_N<\infty} \cdots \int \prod_{k=1}^{N} f\left(w_k - \sum_{i=1}^{c} y_i \delta_{i,z_k}\right) dw_k. \tag{2.2}$$

Notice that $\sum_{i=1}^{c} \delta_{i,z_k} = 1$. Hence

$$\mathsf{P}(Z = z|H_\Delta) - \mathsf{P}(Z = z|H_0) = \int_{-\infty}^{\infty} \cdots \int [\mathsf{P}(Z = z|H_\Delta, Y_i = y_i, i = 1, \ldots, c) - \mathsf{P}(Z = z|H_0)] \prod_{i=1}^{c} dG(y_i) \tag{2.3}$$

where

$$\mathsf{P}(Z = z|H_\Delta, Y_i = y_i, i = 1, \ldots, c) - \mathsf{P}(Z = z|H_0) = \left(\prod_{i=1}^{c} n_i!\right) \int_{-\infty<w_1<\cdots<w_N<\infty} \cdots \int \left[\prod_{k=1}^{N} f\left(w_k - \Delta \sum_{i=1}^{c} \delta_{i,z_k} y_i\right) - \prod_{k=1}^{N} f(w_k)\right] \prod_{1}^{N} dw_i. \tag{2.4}$$

Letting $\theta_k = \sum_{i=1}^{c} y_i \delta_{i, z_k}$ and expanding $h(\Delta) = \prod_{k=1}^{N} f(w_k - \Delta\theta_k)$ around $\Delta = 0$ we obtain

$$h(0) = \sum_{k=1}^{N} f(w_k), \quad h'(0) = h(0)\left[-\sum_{k=1}^{N} (f'/f)\theta_k\right] \tag{2.5}$$

$$h''(0) = h(0)\left[\sum_{k=1}^{N} (f'/f)\theta_k\right]^2 + h(0)\left[\sum_{k=1}^{N}\left\{\frac{f''}{f} - \left(\frac{f'}{f}\right)^2\right\}\theta_k^2\right] \tag{2.6}$$

$$h''(\Delta^*) = h(\Delta^*)\left[\sum_{k=1}^{N} (f'/f)\theta_k\right]^2 + h(\Delta^*)\left[\sum\left\{\frac{f''}{f} - \left(\frac{f'}{f}\right)^2\right\}\theta_k^2\right] \tag{2.7}$$

where the arguments of f, f' and f'' in (2.5) and (2.6) is w_k and in (2.7) it is $w_k - \Delta^*\theta_k$. Then

$$\lim_{\Delta\to 0} \Delta^{-2}[\mathsf{P}(Z = z|H_\Delta) - \mathsf{P}(Z = z|H_0)] =$$

$$= \lim_{\Delta\to 0} \int_{y_1=-\infty}^{\infty} \cdots \int_{y_c=-\infty}^{\infty} \left(\prod_{1}^{c} n_i!\right) \int_{-\infty<w_1<\cdots<w_N<\infty} \cdots \int \Delta^{-2} \times$$

$$\times \left\{\prod_{k=1}^{N} f(w_k - \Delta \sum_{i=1}^{c} \delta_{i, z_k} y_i) - \prod_{k=1}^{N} f(w_k)\right\} \prod_{k=1}^{N} dw_k \prod_{k=1}^{c} dG(y_i).$$

Using the computations in (2.5)–(2.7) we have

$$\text{L.H.S.} = (1/2) E_0\left[\sum_{i=1}^{c}\sum_{k=1}^{N} Y_i \delta_{i, Z_k} \frac{f'(W_k)}{f(W_k)}\right]^2 + (1/2)\sum_{k=1}^{N} E_0\left\{\frac{f''}{f} - \left(\frac{f'}{f}\right)^2\right\} \cdot \tag{2.8}$$

$$\cdot \left(\sum_{1}^{c} \delta_{i, Z_k} Y_i\right)^2 + (1/2)\lim_{\Delta\to 0} \int_{-\infty}^{\infty} \cdots \int \left(\prod_{1}^{c} n_i!\right) \int_{-\infty<w_1<\cdots<w_N<\infty} \cdots \int \{h''(\Delta^*) - h''(0)\} \prod_{k=1}^{N} dw_k \cdot$$

$$\cdot \prod_{1}^{c} dG(y_i) = T + (1/2)\lim_{\Delta\to 0} \int_{-\infty}^{\infty} \cdots \int \left(\prod_{1}^{c} n_i!\right) \int_{-\infty<w_1<\cdots<w_N<\infty} \cdots \int [h''(\Delta^*) - h''(0)] \prod_{1}^{c} dw_k \cdot$$

$$\cdot \prod_{i=1}^{c} dG(y_i)$$

where $h''(\Delta^*)$ is given by (2.7). Now if the limit on Δ can be taken beyond the integrations the second term vanishes. The interchange of limit on Δ and integration is justified via the convergence Theorem of HÁJEK and SIDÁK (1967 p. 154) or the Lebesgue dominated convergence theorem, because

$$\limsup_{\Delta\to 0} \int_{-\infty}^{\infty} \cdots \int \left(\prod_{1}^{c} n_i!\right) \int_{-\infty}^{\infty} \cdots \int |h''(\Delta)| \prod_{k=1}^{N} dw_k \prod_{i=1}^{c} dG(y_i)$$

$$\leqq (\mathsf{E}Y^2)\left\{\sum_{i=1}^{N} E_0\left|\frac{f''}{f} - \left(\frac{b'}{b}\right)^2\right| + \sum_{i=1}^{N}\sum_{k=1}^{N} E_0\left|\frac{f'(W_i)}{f(W_i)}\frac{f'(W_k)}{f(W_k)}\right|\right\} < \infty$$

by the hypothesis. This completes the proof of Theorem 2.1.

Remark 2.1. We can weaken the restriction of independence of Y's and assume that the $\{Y_i\}$ form an exchangeable sequence (that is, a symmetrically dependent sequence) of random variables. Then, if $G(y_1, \ldots, y_c)$ denotes the joint distribution function of $Y_1, \ldots, Y_c$, we will be interested in testing

$$H_0\colon G(y_1, \ldots, y_c) = \begin{cases} 0 & \text{if} \quad (y_1, \ldots, y_c) < (0, \ldots, 0) \\ 1 & \text{if} \quad (y_1, \ldots, y_c) \geqq (0, \ldots, 0) \end{cases}.$$

One can proceed and obtain the following locally most powerful test:

Reject H_0 if

$$(\mathbf{E}Y^2)\left[\sum_{i=1}^{c}\sum_{k=1}^{N} E_0\left(\frac{f''(W_k)}{f(W_k)}\right)\delta_{i,Z_k} + \sum_{i=1}^{c}\sum_{k\neq m}^{N}\sum^{N} \delta_{i,Z_k}\delta_{i,Z_m} E_0\left(\frac{f'(W_k)f'(W_m)}{f(W_k)f(W_m)}\right)\right] +$$

$$+ \mathsf{E}(Y_1 Y_2)\left[\sum_{i\neq j}^{c}\sum^{c}\sum_{k=1}^{N} E_0\left(\frac{f''(W_k)}{f(W_k)}\right)\delta_{i,Z_k}\delta_{j,Z_k} +\right.$$

$$\left. + \sum_{i\neq j}^{c}\sum^{c}\sum_{k\neq m}^{N}\sum^{N} E_0\left\{\frac{f'(W_k)}{f(W_k)}\,\frac{f'(W_m)}{f(W_m)}\right\}\delta_{i,Z_k}\delta_{j,Z_m}\right] > K_\alpha$$

where K_α is determined by the level of significance. Thus one needs the values of $\mathbf{E}Y^2$ and $\mathsf{E}(Y_1 Y_2)$ in order to carry out the test procedure, unless $\mathsf{E}(Y_1 Y_2) = \mathsf{E}(Y_1^2)$ in which case the Y_i are linearly related.

Remark 2.2: One can get the one-sample case by letting $n_1 = \ldots = n_c = 1$ and $c = n$.

Special case 2.1. Let $f(x) = (1/2)\exp(-|x|)$, $-\infty < x < \infty$. Then

$$T = \sum_{j=1}^{c}\left[\sum_{i=1}^{N}\sum_{k=1}^{N} E_0\{\operatorname{Sgn} W_i \operatorname{Sgn} W_k\}\delta_{j,Z_i}\delta_{j,Z_k}\right] = \tag{2.9}$$

$$= \sum_{j=1}^{c}\left[\sum_{i=1}^{N} E_0(\operatorname{Sgn} W_i)\,\delta_{j,Z_i}\right]^2 +$$

$$+ \sum_{j=1}^{c}\sum_{i=1}^{N}\sum_{k=1}^{N}[\operatorname{Cov}_0(\operatorname{Sgn} W_i, \operatorname{Sgn} W_k)]\,\delta_{j,Z_i}\delta_{j,Z_k}$$

where $\operatorname{Sgn} x$ denotes the sign function (namely $\operatorname{Sgn} x = 1, 0$ or -1 according as $x > 0$, $x = 0$ or $x < 0$ respectively).

Let $L = (\operatorname{Sgn} W + 1)/2$. Then L is a Bernoulli variable taking values 0 and 1 with probabilities 1/2 each. Also, let $L_i = (1 + \operatorname{Sgn} W_i)/2$, $i = 1, 2, \ldots, N$. Then, $L_1 \leqq L_2 \ldots \ldots \leqq L_N$ constitute order statistics in a sample of size N drawn from the symmetric Bernoulli population. Further,

$$\mathsf{P}(L_i = 1) = 2^{-N}\sum_{j=0}^{i-1}\binom{N}{j} \quad \text{and for} \quad i \leqq k, \mathsf{P}(L_i = L_k = 1) = \mathsf{P}(L_i = 1).$$

Hence, routine computations yield

$$\text{(2.10)} \qquad \mathsf{E}(\operatorname{Sgn} W_i) = -1+2^{-(N-1)}\sum_{j=0}^{i-1}\binom{N}{j} \approx -1+2\Phi\left(\frac{2i-1-N}{\sqrt{N}}\right)$$

and for $i \leqq k$

$$\text{(2.11)} \quad \operatorname{Cov}(\operatorname{Sgn} W_i, \operatorname{Sgn} W_k) = 4\operatorname{Cov}(L_i, L_k) = 4\cdot 2^{-2N}\left(\sum_{j=0}^{i-1}\binom{N}{j}\right)\left(\sum_{l=k}^{N}\binom{N}{l}\right) =$$

$$\approx 4\Phi\left(\frac{2i-1-N}{\sqrt{N}}\right)\Phi\left(\frac{N+1-2k}{\sqrt{N}}\right)$$

where Φ denotes the standard normal distribution function.

Special case 2.2. Let $f(x)=e^x(1+e^x)^{-2}$ or $F(x)=e^x(1+e^x)^{-1}$, $-\infty<x<\infty$. Then $f'(x)/f(x)=(1-2F)$ and $f''/f=(1-2F)^2-2F(1-F)$. Thus

$$\text{(2.12)} \qquad T=\sum_{k=1}^{c}\left[\sum_{i=1}^{N}\sum_{k=1}^{N}\mathsf{E}(1-2U_i)(1-2U_k)\delta_{j,z_i}\delta_{j,z_k}\right]-$$

$$-2\sum_{j=1}^{c}\sum_{i=1}^{N}\mathsf{E}\{U_i(1-U_i)\}\delta_{j,z_i}$$

where U_i denotes the ith smallest order statistic in a sample of size N drawn from the uniform distribution on [0, 1]. Thus,

$$\text{(2.13)} \qquad \frac{T}{N}+\frac{N}{3(N+2)} = \sum_{j=1}^{c}\left[N^{-1/2}\sum_{i=1}^{N}\mathsf{E}(1-2U_i)\delta_{j,z_i}\right]^2 +$$

$$+4N^{-1}\sum_{j=1}^{c}\sum_{i=1}^{N}\sum_{k=1}^{N}\operatorname{Cov}(U_i, U_k)\delta_{j,z_i}\delta_{j,z_k}$$

since $\sum_{j=1}^{c}\delta_{j,z_i}=1$ and $\sum_{i=1}^{N}\mathsf{E}\{U_i(1-U_i)\} = \sum_{i=1}^{N}\frac{i(N+1-i)}{(N+1)(N+2)} = \frac{N^2}{6(N+2)}$.

Special case 2.3. Let $f(x)=(2\pi)^{-1/2}\exp(-x^2/2)$, $-\infty<x<\infty$. Then

$$\text{(2.14)} \qquad T+N=\sum_{j=1}^{c}\left[\sum_{i=1}^{N}\sum_{k=1}^{N}L_{ik}\delta_{j,z_i}\delta_{j,z_k}\right]$$

where $L_{ik}=\mathsf{E}(V_iV_k)$ where V_i denotes the ith smallest normal order statistic in a sample of size N. Thus

$$\text{(2.15)} \qquad (T/N)+1 = \sum_{j=1}^{c}\left[N^{-1/2}\sum_{i=1}^{N}\mathsf{E}(V_i)\delta_{j,z_i}\right]^2 +$$

$$+N^{-1}\sum_{j=1}^{c}\sum_{i=1}^{N}\sum_{k=1}^{N}\operatorname{Cov}(V_i, V_k)\delta_{j,z_i}\delta_{j,z_k}.$$

3. Asymptotic Distribution of T under H_0. We shall define a class of statistics of which (2.9), (2.12) and (2.14) are special cases, and study its asymptotic distribution. Let $W_1 < W_2 < \cdots < W_N$ denote the order statistics in samples of sizes N drawn from an arbitrary distribution. Let g be a monotone Borel measurable function. Without loss of generality assume that g is monotonic nondecreasing. Then $g(W_1) \leqq \leqq g(W_2) \leqq \cdots \leqq g(W_N)$ constitute order statistics. Define the class of statistics:

$$T/N = N^{-1} \sum_{j=1}^{c} \sum_{i=1}^{N} \sum_{k=1}^{N} E_0\big(g(W_i)\, g(W_k)\big) \delta_{j, Z_i} \delta_{j, Z_k}. \tag{3.1}$$

In the double exponential case $g(x) = \operatorname{sgn} x$, in the logistic case $g(x) = 2F(x) - 1$ and in the normal case $g(x) = x$. One can rewrite

$$T/N = T^* + \tilde{T} \tag{3.2}$$

where

$$T^* = \sum_{j=1}^{c} \left[N^{-1/2} \sum_{i=1}^{N} E_0\big(g(W_i)\big)\, \delta_{j, Z_i} \right]^2, \tag{3.3}$$

$$\tilde{T} = \sum_{j=1}^{c} \tilde{T}_j \tag{3.4}$$

and

$$\tilde{T}_j = N^{-1} \sum_{i=1}^{N} \sum_{k=1}^{N} \operatorname{Cov}\big(g(W_i),\, g(W_k)\big) \delta_{j, Z_i} \delta_{j, Z_k}. \tag{3.5}$$

We need the following lemmas.

LEMMA 3.1. *Let* $c_{ik} = \operatorname{Cov}\big(g(W_i), g(W_k)\big)$, $1 \leqq i, k \leqq N$. *Then,*

(i) $c_{ik} \geqq 0$, *and*

(ii) $N^{-1} \sum_{i=1}^{N} \sum_{k=1}^{N} c_{ik} = \operatorname{Var}\{g(W)\} < \infty$.

PROOF. For (i) see, for instance BICKEL (1967), Lemma 2.1. For (ii) consider $\operatorname{Var}\left(\sum_{i=1}^{N} g(W_i) \right) = N \operatorname{Var}\{g(W)\}$. (See also Theorem 4.5 of GOVINDARAJULU (1963)).

LEMMA 3.2. (LEVER (1970)). *If* $\{b_i\}$ $i = 1, \ldots, n$, $n = 1, \ldots,$ *are real numbers such that* $0 \leqq b_i \leqq 1$, *and if* $n^{-1} \left(\sum_{i=1}^{n} b_i \right)^2 = 1$, *then* $n^{-1} \sum_{i=1}^{n} b_i^2 \to 0$ as $n \to \infty$.

PROOF. It suffices to show that $\max \left\{ n^{-1} \sum_{i=1}^{n} b_i^2 \right\}$ subject to the conditions of the hypothesis tends to zero as $n \to \infty$. Equivalently, we will maximize

$$\sum_{i=1}^{n} b_i^2 - n^{-1} \left(\sum_{i=1}^{n} b_i \right)^2 = \sum_{i=1}^{n} \left[b_i - n^{-1} \sum_{i=1}^{n} b_i \right]^2 = \sum_{i=1}^{n} [b_i - n^{-1/2}]^2.$$

The preceding expression will have its maximum value when the differences $b_i - n^{-1/2}$ are the maximum possible (in absolute value). Since $0 \leqq b_i \leqq 1$ and $\sum_{i=1}^{n} b_i = n^{1/2}$, let

$$b_i = \begin{cases} 1 & i = 1, \ldots, n^{1/2} \\ n^{1/2} - [n^{1/2}], & i = [n^{1/2}] + 1 \\ 0 & i = [n^{1/2}] + 2, \ldots, n, \end{cases}$$

where $[\cdot]$ denotes the largest integer contained in $(\cdot)$. Then $n^{-1} \sum_{i=1}^{n} b_i^2 < n^{-1}\{[n^{1/2}]+1\}$ since $n^{1/2} - [n^{1/2}] < 1$. This completes the proof of the lemma.

As a special case, let $b_i = c_{ik} =$ covariance of V_{iN} and V_{kN}. Then, $0 \leqq c_{ik} \leqq 1$ since $\sum_{i=1}^{n} c_{ik} = 1$ for each i, and from Lemma 3.1, $N^{-2}\left(\sum_{ik}\sum c_{ik}\right)^2 = N^{-2}(N)^2 = 1$. Thus for normal order statistics $N^{-2} \sum_{ik}\sum c_{ik}^2 \to 0$ as $N \to \infty$.

LEMMA 3.3. *If* $N^{-2} \sum_{i \neq k \neq l}^{N}\sum^{N}\sum^{N} c_{ik} c_{il} = o(1)$ *and* $N^{-2} \sum_{i \neq k}^{N}\sum^{N} c_{ik}^2 = o(1)$, *then* $(\tilde{T} - A) \to 0$ *in probability as* $N \to \infty$ *where*

$$A = \sum_{j=1}^{c} \{n_j(n_j - 1)/N(N-1)\} [\mathrm{Var}\{g(W)\}].$$

PROOF. It suffices to show that $\left[\tilde{T}_j - \frac{n_j(n_j-1)}{N(N-1)} \mathrm{Var}\{g(W)\}\right]$ tends to zero in probability for $j = 1, \ldots, c$ as $N \to \infty$. Now write

$$\tilde{T}_j = \tilde{T}_{j1} + \tilde{T}_{j2}$$

where

$$\tilde{T}_{j2} = N^{-1} \sum_{i=1}^{N} c_{ii} \delta_{j, Z_i}$$

and

$$\tilde{T}_{j1} = N^{-1} \sum_{i \neq k}\sum c_{ik} \delta_{j, Z_i} \delta_{j, Z_k}.$$

Then

$$\tilde{T}_{j2} \leqq N^{-1} \sum_{i=1}^{N} c_{ii} = N^{-1} \sum_{i=1}^{N} [\mathsf{E} g^2(W_i) - \{\mathsf{E} g(W_i)\}^2] \to \mathsf{E} g^2(W) - \mathsf{E} g^2(W) = 0,$$

by HOEFFDING's (1953) Theorem 2. Thus $\tilde{T}_{j2} \to 0$ for each j as $N \to \infty$. Next consider

$$E_0\{T_{j1} - E_0(T_{j1})\}^2 = N^{-2} \sum_{i \neq k}\sum \sum_{l \neq m}\sum c_{ik} c_{lm} E_0(\Delta_{ik}^{(j)} \Delta_{lm}^{(j)})$$

where $\Delta_{ik}^{(j)} = \delta_{j, Z_i} \delta_{j, Z_k} - n_j(n_j - 1)\{N(N-1)\}^{-1}$ and subscript 0 for E denotes expectation under H_0. Now using Lemma 3.1, the hypothesis of Lemma 3.3 and the computation:

$$N^{-2} E_0(\Delta_{ik}^{(j)} \Delta_{lm}^{(j)} \mid i \neq k \neq l \neq m) = \frac{-n_j(n_j-1)(N-n_j)(4Nn_j - 6N - 6n_k + 6)}{N^2(N-1)^2(N-2)(N-3)}$$

one can easily show that $E_0\{\tilde{T}_{j1}-E_0(\tilde{T}_{j1})\}^2=O(N^{-1})$ for $j=1, \ldots, c$. Now application of the Chebyshev's inequality completes the proof of Lemma 3.3. Then we are led to the following theorem.

THEOREM 3.1. *If the hypothesis of Lemma* 3.3 *is satisfied, then* T/N *is asymptotically equivalent to* T^* *where* T^* *is given by* (3.3) *and is the sum of squares of linear rank statistics.*

PROOF. Follows trivially from Lemma 3.3.

Now, variance of $g(W)$ is equal to unity when W is a normal or a double exponential random variable and is equal to $1/3$ when W is a logistic variable. In the case of the normal scores test (see special case 2.3) $W_i=V_i$ since $g(x)=x$. It is well-known (see, for instance, GOVINDARAJULU (1963), corollary 4.6.2) that $\sum_{k=1}^{N} c_{ik}=1$ for each i and from Lemma 3.2., a result of Mrs. Lever (1970) we have that $N^{-2}\sum_{i\neq k}^{N}\sum^{N} c_{ik}^2=o(1)$.

For the logisitc scores test (see Special case 2.2), $g(x)=2F(x)-1$ and $c_{ik}=$ $=i(N-k+1)/(N+1)^2(N+2)$ for $i\leqq k$. Consider

$$N^{-2}\sum_{i\neq k}\sum c_{ik}^2 = 2\sum_{i<k}\sum i^2(N-k+1)^2/N^2(N+1)^4(N+2)^2 = O(N^{-2}).$$

Next, let us turn to the exponential scores. We will show that

$$N^{-1}\sum_{i=1}^{N}\sum_{k=1}^{N} c_{ik}^2 = O(1).$$

Now,

$$N^{-1}\sum_{i=1}^{N} c_{ik}^2 \leqq N^{-1}\sum_{i=1}^{N} c_{ii} \to 0$$

as $N\to\infty$ due to HOEFFDING's (1953) theorem 2. Hence

$$N^{-1}\sum_{i}^{N}\sum_{k}^{N} c_{ik}^2 \simeq 2N^{-1}\sum_{i\leqq k}\sum c_{ik}^2 \approx 8\sum_{i\leqq k}\sum \Phi^2\left(\frac{2i-N}{\sqrt{N}}\right)\left\{1-\Phi\left(\frac{2k-N}{\sqrt{N}}\right)\right\}^2 \approx$$

$$\approx 2\iint\limits_{-\sqrt{N}\leqq x\leqq y\leqq\sqrt{N}} \Phi^2(x)[1-\Phi(y)]^2\,dx\,dy \to$$

$$\to 2\iint\limits_{-\infty<x\leqq y<\infty} \Phi^2(x)[1-\Phi(y)]^2\,dx\,dy.$$

Now, integrating by parts once in $\int_{-\infty}^{y}\Phi^2(x)\,dx$ and then performing the integration with respect to y one can show that the above integral reduces to

$$-4\iint\limits_{x\leqq y} x\Phi(x)[1-\Phi(y)]^2\,dy\,d\Phi(x).$$

Now integrating by parts once in $\int_y^\infty [1-\Phi(y)]^2 dy$ we see that the original integral is equal to

$$4\int_{-\infty}^{\infty} x^2\Phi(x)[1-\Phi(x)]^2\,d\Phi(x) - 8\iint_{x<y} xy\Phi(x)[1-\Phi(x)]\,d\Phi(x)\,d\Phi(y) =$$

$$= 4.\frac{1!2!}{4!}\mathsf{E}(W_{2,4}^2) - 8\cdot\frac{1!1!}{4!}\mathsf{E}(W_{2,4}W_{3,4}) =$$

$$= \frac{1}{3}\left[1-\frac{\sqrt{3}}{\pi}-\frac{2\sqrt{3}-3}{\pi}\right] = \frac{1}{3}\left[1-\frac{3(\sqrt{3}-1)}{\pi}\right]$$

where the W_{in} denote the standard normal order statistics and the exact-expected values are taken from JONES (1948). That is, $N^{-1}\sum_i^N\sum_k^N c_{ik}^2 = O(1)$. Next consider

$$N^{-3/2}\sum_{i\neq k\neq l}\sum\sum c_{ik}c_{il} \leqq 4N^{-3/2}\sum_{i\leqq k\leqq l}\sum\sum c_{ik}c_{il} \approx$$

$$\approx 16N^{-3/2}\sum_{1\leqq i\leqq k\leqq l\leqq N}\sum\sum\Phi^2\left(\frac{2i}{\sqrt{N}}-\sqrt{N}\right)\left\{1-\Phi\left(\frac{2k}{\sqrt{N}}-\sqrt{N}\right)\right\}\left\{1-\Phi\left(\frac{2l}{\sqrt{N}}-\sqrt{N}\right)\right\} =$$

$$= 16N^{-3/2}\sum_{\frac{2}{\sqrt{N}}-\sqrt{N}\leqq x\leqq y\leqq z\leqq\sqrt{N}}\sum\sum\Phi^2(x)[1-\Phi(y)][1-\Phi(z)] \leqq$$

$$\leqq 2\iiint_{-\infty<x\leqq y\leqq z<\infty}\Phi^2(x)[1-\Phi(y)][1-\Phi(z)\,dz\,dy\,dx =$$

$$= -2\iint_{x\leqq y} y\Phi^2(x)[1-\Phi(y)]^2\,dx\,dy + 2\iint_{x\leqq y}\Phi^2(x)[1-\Phi(y)]\,dx\,dy\Phi(y)$$

after integrating by parts once in the integral with respect to z. Now

$$2\iint_{x\leqq y}\Phi^2(x)[1-\Phi(y)]\,d\Phi(y)\,dx = \int_{-\infty}^{\infty}\Phi^2(x)[1-\Phi(x)]^2\,dx$$

which was earlier shown to be finite. Also integrating by parts with respect to x we have

$$2\iint_{x\leqq y} y\Phi^2(x)[1-\Phi(y)]^2\,dxdy = 2\int_{-\infty}^{\infty} y^2\Phi^2(y)[1-\Phi(y)]^2\,dy -$$

$$-4\int_0^{\infty} x\Phi(x)\,d\Phi(x)\left[\int_x^{\infty} y[1-\Phi(y)]^2\,dy\right[=$$

$$= 4\int_0^{\infty} y^2\Phi^2(y)[1-\Phi(y)]^2\,dy + 2\int_{-\infty}^{\infty} x^3\Phi(x)[1-\Phi(x)]^2\,d\Phi(x) -$$

$$-4\iint_{x\leqq y} xy^2\Phi(x)[1-\Phi(y)]\,d\Phi(x)\,d\Phi(y) \leqq$$

$$\leqq 4\int_0^{\infty} y^2[1-\Phi(y)]^2\,dy + 4\int_0^{\infty} x^3\,d\Phi(x) + 4\iint_{x\leqq y}|x|y^2\,d\Phi(x)\,d\Phi(y).$$

By elementary methods one can easily show that each of the integrals is finite. Hence, $N^{-3/2}\sum\sum\sum_{i\neq k\neq l} c_{ik}c_{il}=O(1)$. Thus the hypothesis of lemma 3.2 is satisfied by the statistics arising from special cases 2.1—2.3. Now let us analyze T^*. Towards this we can write

$$N^{-1/2}\sum_{i=1}^{N} E_0\big(g(W_i)\big)\delta_{j,Z_i} = (n_j/\sqrt{N})\int_{-\infty}^{\infty} J_N\left(\frac{N}{N+1}H_N(x)\right)dF_{n_j}^{(j)}(x) \tag{3.6}$$

where

$$H_N(x)=\sum_{j=1}^{c}\lambda_j F_{n_j}^{(j)}(x),\quad \lambda_j=n_j/N,\quad N=\sum_{1}^{c} n_j.$$

$J_N\big(i/(N+1)\big)=E_0\big(g(W_i)\big)$ and $F_{n_j}^{(i)}(x)$ denotes the empirical distribution based on the sample $X_{j1},\ldots,X_{j,n_j}$. Also, it is well-known that

$$J_N(u)\to J(u)\quad \text{for}\quad 0<u<1 \tag{3.7}$$

where $J=gF^{-1}$ and F^{-1} is the inverse of the distribution function of W. Also, by the asymptotic theory of GOVINDARAJULU, LECAM and RAGHAVACHARI (1967)

$$N^{-1/2}\sum_{i=1}^{N} E_0\big(g(W_i)\big)\delta_{j,Z_i}-\lambda_j N^{1/2}\int J\big(H(x)\big)\,dF^{(j)}(x)\sim\lambda_j B_{N,j} \tag{3.8}$$

where

$$B_{N,j}=N^{1/2}\Big[\int J(H)\,d(F_{n_j}^{(j)}(x)-F^{(j)})+\int (H_N-H)J'(H)\,dF^{(j)}\Big] \tag{3.9}$$

and $H(x)=\sum_{j=1}^{c}\lambda_j F^{(j)}(x)$. Integrating by parts in the first term of $B_{N,j}$ we obtain.

$$\begin{aligned}\sqrt{\lambda_j}B_{N,j} &= n_j^{1/2}\Big[\int (H_N-H)J'(H)\,dF^{(j)}-\int (F_{n_j}^{(j)}-F^{(j)})J'(H)\,dH\Big]\\ &= n_j^{1/2}\Big[\sum_{i\neq j}\lambda_i\int (F_{n_i}^{(i)}-F^{(i)})J'(H)\,dF^{(i)}-\\ &\quad-\sum_{i\neq j}\lambda_i\int (F_{n_j}^{(j)}-F^{(j)})J'(H)\,dF^{(i)}\Big]=\\ &= \sqrt{I}\Big[\sum_{i\neq j}\sqrt{\lambda_i\lambda_j}\tilde{Z}_{ij}-\sum_{i\neq j}\lambda_i\tilde{Z}_{ji}\Big]\end{aligned} \tag{3.10}$$

where

$$\tilde{Z}_i=\big\{n_i^{1/2}/\sqrt{I}\big\}\int\big(F_{n_i}^{(i)}(x)-F^{(i)}(x)\big)\,dF^{(j)},\quad i,j=1,\ldots,c, \tag{3.11}$$

and

$$I=2\iint_{0<x<y<1} x(1-y)J'(x)J'(y)\,dx\,dy=\int_0^1 J^2(x)\,dx-\left[\int_0^1 J(x)\,dx\right]^2 \tag{3.12}$$

Notice that $\tilde{Z}_{ij}$ is asymptotically standard normal when $F^{(1)}=\cdots=F^{(c)}$ $(i=1,\ldots,c)$. Then, we have the following theorem.

THEOREM 3.2. *If* $\lambda_i \equiv 1/c$ *and* $F^{(1)} = \cdots = F^{(c)}$ *then* cT^*/I *is asymptotically distributed as noncentral chi-square with* $c-1$ *degrees of freedom and noncentrality parameter*

$$\delta = (c-1)\mu^2, \quad \mu = \int_0^1 J(u)\,du. \tag{3.13}$$

PROOF. One can write

$$T^*/I = \sum_{j=1}^{c} \lambda_j \{\sqrt{\lambda_j}\, B_{N,j}\}^2 / I \tag{3.14}$$

where

$$\mathrm{Var}\,(\sqrt{\lambda_j}\, B_{N,j} | H_0) = (1-\lambda_j) I \tag{3.15}$$

and

$$\mathrm{Cov}\,(\sqrt{\lambda_j}\, B_{N,j}, \sqrt{\lambda_k}\, B_{N,k} | H_0) = -\sqrt{\lambda_j \lambda_k}\, I \tag{3.16}$$

with $\sum_{j=1}^{c} \lambda_j = 1$. Thus under H_0 when $\lambda_j \equiv 1/c$, cT^*/I is asymptotically equivalent to sum of c squares of correlated normal variables having the variance-covariance structure given by (3.15)–(3.16). Thus, by the Theorem of HÁJEK and SIDÁK (1967) (see p. 31)*, cT^*/I is asymptotically noncentral chi-square with $c-1$ degrees of freedom and noncentrality parameter δ.

Remark 3.2.1. For the normal scores statistic, $J = \Phi^{-1}$ and $\mu = 0$ and $I = 1$, for the logistic scores statistic, $J(u) = 2u - 1$, and $\mu = 0$, $I = 1/3$ and for the exponential scores, $I = \mathrm{Var}\, J(u) = \mathrm{Var}\, g(W) = 1$, $\mu = 0$.

COROLLARY 3.2.1. *For the exponential and normal scores,* cT^* *is asymptotically equivalent to a central chi-square with* $c-1$ *degrees of freedom and for the logistic scores,* $3cT^*$ *is asymptotically equivalent to a central chi-square with* $c-1$ *degrees of freedom.*

Next it is of interest to explore approximations to the distribution of T^* under H_0 when not all λ_i are equal to $1/c$. Towards this, let $\tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_c)'$, $B_N = (B_{N,1}, \ldots, B_{N,c})'$, $b = (\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_c})'$ and D_λ denote the diagonal matrix having $\lambda_1^2, \ldots, \lambda_c^2$ for the diagonal elements. Then using (3.10) one can easily write

$$T^* = I \tilde{Z}' A' D_\lambda A \tilde{Z} \tag{3.17}$$

where

$$A = \begin{pmatrix} -(1-\lambda_1)/\sqrt{\lambda_1}, & \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_c} \\ \sqrt{\lambda_1}, & -(1-\lambda_2)/\sqrt{\lambda_2}, \ldots, \sqrt{\lambda_c} \\ \sqrt{\lambda_1}, & \sqrt{\lambda_2}, \quad \ldots, -(1-\lambda_c)/\sqrt{\lambda_c} \end{pmatrix}$$

and

$$Ab = 0. \tag{3.19}$$

*Howewer, note that the statement of their theorem and the proof there are in error.

Further, there exists an orthogonal transformation which diagonalizes $A'D_\lambda A$. Hence, without loss of generality let us assume that

$$T^*/I \sim 2Q = \sum_{i=1}^{c-1} a_i \tilde{Z}_i^2 \tag{3.20}$$

where $a_1 < a_2 \ldots < a_{c-1}$ denote the characteristic roots of $A'D_\lambda A$ and $\tilde{Z}_i$ are independent standard normal variables.

Case 1. Let $c-1=2m$, m is a positive integer. Then let

$$2Q_1 = a_1(\tilde{Z}_1^2+\tilde{Z}_2^2)+\cdots+a_{c-2}(\tilde{Z}_{c-2}^2+\tilde{Z}_{c-1}^2) \tag{3.21}$$

$$2Q_2 = a_2(\tilde{Z}_1^2+\tilde{Z}_2^2)+\cdots+a_{c-1}(\tilde{Z}_{c-2}^2+\tilde{Z}_{c-1}^2). \tag{3.22}$$

$F_i(x)=\mathsf{P}(Q_i>x)$, $i=1, 2$. It is well known (see, for instance, Theorem 1 of SIDDIQUI (1965)) that $F_i(x)$ can be evaluated exactly as linear combinations of the distributions of exponential random variables. If $a_{2j-1} \neq a_{2j}$ for at least one j, then almost surely $Q_1 < Q < Q_2$, which implies, for all x

$$F_1(x) < F(x) < F_2(x) \tag{3.23}$$

Then, SIDDIQUI (1965) obtains the 'optimal' $\hat{F}(x)$ as

$$\hat{F}(x) = F_1(x)+\theta(F_2-F_1), \quad 0 < \theta < 1 \tag{3.24}$$

where

$$\theta = (F-F_1, F_2-F_1)\|F_2-F_1\|^{-2} \tag{3.25}$$

where $(p, q) = \int_0^\infty p(x)q(x)\,dx$, $\|p\|=(p, p)^{1/2}$.

Notice that the optimum θ minimizes $\|F-\hat{F}\|$. Siddiqui (1965) gives a systematic procedure for evaluating θ.

Case 2. let $c=2m$, then let

$$2Q_1 = a_1(\tilde{Z}_1^2+\tilde{Z}_2^2)+\cdots+a_{c-2}(\tilde{Z}_{c-3}^2+\tilde{Z}_{c-2}^2)$$

and

$$2Q_2 = a_1(\tilde{Z}_1^2+\tilde{Z}_c^2)+\cdots+a_{c-1}(\tilde{Z}_{c-2}^2+\tilde{Z}_{c-1}^2)$$

SIDDIQUI (1965) proposes a cruder choice for the Q_i namely $2Q_1=a_1\sum_1^{c-1}\tilde{Z}_i^2$ and $2Q_2=a_{c-1}\sum_{i=1}^{c-1}\tilde{Z}_i^2$ provided a_{b-1}/a_1 is near unity.

A somewhat simpler approximation can be obtained by using PATNAIK's (1949) approximation. This is given in Theorem 3.3.

THEOREM 3.3. *If all the n_i are not necessarily equal and if $\mu_{N,i}\equiv 0$ and $F^{(1)}=\cdots=F^{(c)}$, then, for suffcently large N,*

$$T^* \sim (a)\chi_b^2 \tag{3.26}$$

where $b=2(1-S_2)^2/\{S_4-14S_3-S_2^2+3S_2+c^2-c+11\}$

$$a=(1-S_2)I/b \tag{3.27}$$

and $S_i=\sum_{j=1}^{c}\lambda_j^i$, $i=1, \ldots, 4$.

PROOF. Here we resort to PATNAIK's (1949) approximation. That is, we write $T^*=(a)\chi_b^2$ and solve for a and b by equating the first two moments on both sides. So, when $\mu_{N,j}=0$ we have

$$T^*=\sum_{j=1}^{c}B_{N,j}^2=\sum_{j=1}^{c}\lambda_j(1-\lambda_j)I\tilde{B}_{N,j}^2 \tag{3.28}$$

where $\tilde{B}_{N,j}=B_{N,j}/\sqrt{\lambda_j(1-\lambda_j)I}$ which is approximately standard normal. The equations to be solved for, are

$$E_0(T^*)=ab \quad \text{and} \quad \operatorname{Var}(T^*|H_0)=2a^2b \tag{3.29}$$

where

$$E_0(T^*)=I\sum_{j=1}^{c}\lambda_j(1-\lambda_j) \text{ and}$$

$$\operatorname{Var}(T^*|H_0)=2I^2\sum_{j=1}^{c}\lambda_j^2(1-\lambda_j)^2+\sum_{j\neq k}\sum \mathsf{E}(B_{N,i}^2B_{N,k}^2|H_0)$$

$$-\sum_{i\neq k}\sum(\mathsf{E}B_{N,i}^2|H_0)(\mathsf{E}B_{N,k}^2|H_0).$$

Routine but tedious calculations yield

$$\mathsf{E}(B_{N,i}^2B_{N,k}^2|H_0, i\neq k)=I^2[12\lambda_i\lambda_k+1-(\lambda_i+\lambda_k)(6\lambda_i\lambda_k+1)]+O\left(\frac{1}{N}\right). \tag{3.30}$$

Using the significant term of the right side of (3.30) in (3.29) and solving for a and b we obtain (3.27).

*

Acknowledgement. I began working on this problem during the summer of 1969 when I was a visiting Professor at Florida State University. I am indebted to Professor RICHARD SAVAGE for some stimulating conversations I had with him and for his useful comments.

REFERENCES

[1] BICKEL, P. J.: Some contributions to the theory order statistics, *Proc. V Berkeley Symposium on Math. Statist. Prob.* **1** Univ. of Calif. Press. (1967), 575—592.

[2] GREENBERG, V. L.: *Robust Inference on Some Experimental Designs* A Dissertation submitted to the University of California for the Ph. D. degree. (1964).

[3] GOVINDARAJULU, Z.: On moments order statistics and quasi-ranges from normal populations, *Ann. Math. Statist.* **34** (1963) 633—651.

[4] GOVINDARAJULU, Z., LECAM, L. and RAGHAVACHARI, M.: Generalization of theorems of Chernoff and Savage on asymptotic normality of nonparametric test statistics, *Proc. V. Berkeley Symposium on Math. Statist. Prob.* **1**. Univ. of Calif. Press. 609—638. (1967).

[5] GOVINDARAJULU, Z. and DESPHANDÉ, J. V.: Random effects model: nonparametric case, *Ann. Inst. Statist. Math.* **24** (1972), 165—170.
[6] HÁJEK, J. and SIDÁK, Z.: *Theory of Rank Tests*, Academic Press, New York. (1967)
[7] HOEFFDING, W.: On the distribution of expected values of the order statistics, *Ann. Math. Statist.* **24** (1953), 93—100.
[8] JONES, H. L.: Exact lower moments of order statistics in small samples from a normal distribution, *Ann. Math. Staticst.* **19** (1948), 270—273.
[9] KULKARNI, S. R.: On the optimal asymptotic tests for the effects of cloud seeding on rainfall (2): the case of variable effects, The *Australian Journal of Statistics* **11** No. 1 (1969), 39—51.
[10] KULKARNI, S. R.: Locally asymptotically most powerful tests about the effects of K treatments, *Ann. Inst. Statist. Math.* **22** (1970), 145—158.
[11] LEVER, C.: Personal communication (1970).
[12] LOÈVE, M.: *Probability Theory*, Van Nostrand, Princeton (1960).
[13] NEYMAN, J.: Experimentation with weather control, *J. R. Statist. Soc. Series A.* **130** (1967), 285—326.
[14] PATNAIK, P. B.: The noncentral chi-square and F distributions and their applications, *Biometrika* **36** (1949), 202—232.
[15] SCHEFFÉ, H.: *The Analysis of Variance*, Wiley, New York (1959).
[16] SIDDIQUI, M. M.: Approximations to the distribution of quadratic forms, *Ann. Math. Statis.* **36** (1965), 677—682.

University of Kentucky, Lexington, Kentucky

(Received November 14, 1973)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 61—72.

BINOMIAL GROUP TESTING WITH TWO DIFFERENT SUCCESS PARAMETERS

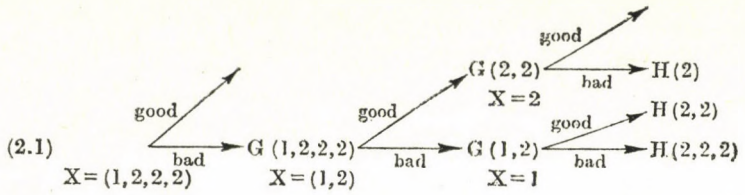
by

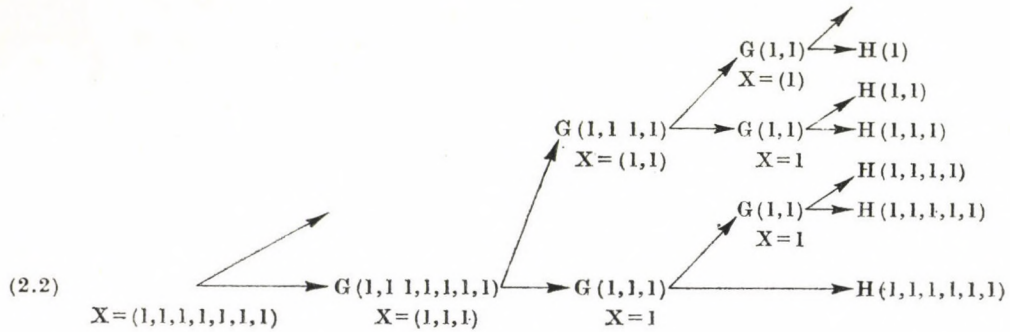
E. NEBENZAHL

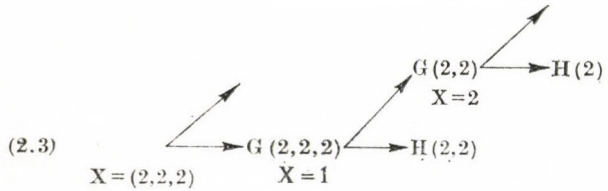
1. Introduction. n units identified by "x_1", "x_2", ... "x_n", are to be each classified as "good" or "bad", where the x_i^{th} unit has probability q_i of being good, independent of the classifications for the other units ($i=1, 2, \ldots, n$). We are not restricted to testing units individually (in order to classify them as good or bad) but may test any group of units $x=(x_{i_1}, \ldots, x_{i_n})$ simultaneously and arrive at one of two possible conclusions, that either 1) all the units tested are good or 2) at least one of them is bad ($1 \leqq i_1 < i_2 \ldots < i_n \leqq n$); A *procedure* represents a series of such tests for the purpose of classifying all n units. If we have no information about a set of units (as far as their being good or bad), we refer to it as a *binomial set;* if a set is known to have at least one defective, then it is referred to as a *defective set*. A procedure is said to be *non-mixing* if it never simultaneously tests a binomial and a defective set.

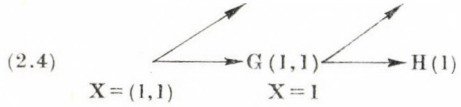
In the original "group-testing" problem of the above nature (see [1]), a test consisted of pooling the blood samples from a number of different people, for the purpose of detecting syphillis. In [3] and [4], the authors find the "optimal" non-mixing procedure, in the sense of minimizing the expected number of tests necessary to classify all the units; they only do so for the case that all the units have the same probability q of being good. Suppose, though, that there are M_1 units of type „1", having a probability q of being good and M_d units of type "d" (d, a positive integer $\geqq 2$), having a probability q^d of being good, i.e., we can consider a unit of type "d" (or a d-unit) to consist of d independent "components", all of which must be good in order for the unit to be good. One procedure that might be worked in this case is to use the one given in [3] and [4] separately for the two types of units. In this paper, we complete a discussion (begun in [2]) of a procedure (which is referred to later as H_{T^*}) that for the case when M_1 and M_d are large but finite, is a considerable improvement over the above one and which is conjectured to be the optimal (non-mixing) procedure. A modification of H_{T^*} (described in [2]) yields a useful prodecure for the case when $M_1, M_d \to \infty$ such that $M_1/M_d \to s < \infty$.

2. Preliminaries. A *tree* represents a series of tests, starting when all the remaining unclassified units form a binomial set (or are in the *H-state*) and continuing on until a first return to the H-state. For example, let $d=2$, i.e., we have only 1-units and 2-units, and define a tree for this case pictorially as (2.1) which means that one begins in the H-state by testing a single 1-unit and 3 2-units and then either 1) they are all good (the tree is ended) and we return to the H-state or 2) they form a defective set, called $G(1, 2, 2, 2)$; at the second step (the $G(1, 2, 2, 2)$ step), one tests a (1, 2) grouping and then either 1) they are both good and there remains a defective set of 2 2-units







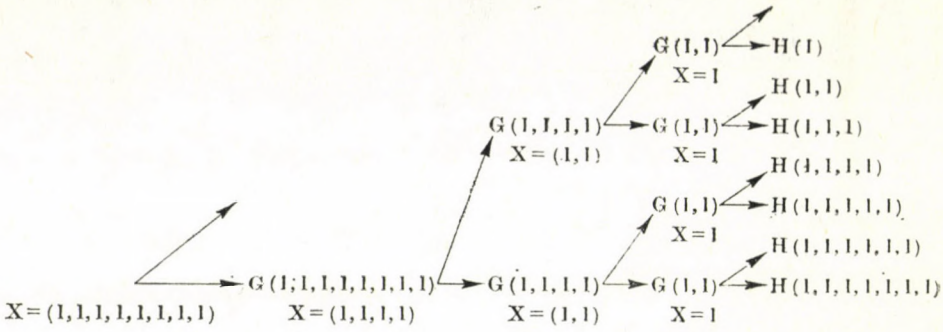


(called $G(2, 2)$) or 2) they form the $G(1, 2)$ defective set, while at the same time a (2, 2) grouping returns to the binomial state; at the $G(2, 2)$ third step, we test a 2-unit and then either 1) it is good (the tree is ended) and we return to the H-state or 2) it is bad and we return to the H-state, with a 2-unit from the initial (1, 2, 2, 2) grouping still unclassified, referred to by "$H(2)$"; we continue on in this fashion.

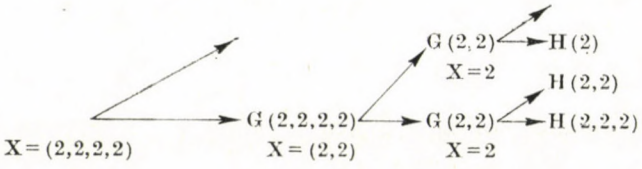
For the general (d, q) problem, where there are 1 and d-units having probabilities q and q^d, respectively, the optimal tree for classifying the 1-units (resp., d-units alone under the assumption that M_1 and M_d are large (see especially procedure R_{21} given after (18) of [4]) is to be denoted by T_1 (resp., T_d). Also for $j \geqq 1$, let $T(j, 1)$ denote the optimal tree for classifying the 1-units alone under the restriction that one begin testing at the initial (or H) step with j 1-units (also obtained from [3] and [4]); this tree is, of course, not as good as T_1. We illustrate some of the more general results of this paper with reference to two special cases, $(d, q)=(2, \cdot 9)$ and $(d, q)=$

$=(2, \cdot 92)$. For the first case, T_1, T_2 and $T(2, 1)$ are respectively given by (2.2)-(2.4), for the second case, T_1, T_2 and $T(2, 1)$ are given respectively by (2.4)-(2.6).

(2.5)



(2.6)



A procedure can be thought of as representing a sequence of trees. We focus in on a class of procedures that includes the conjectured optimal one. For any tree T, let H_T be the procedure which works T over and over again until either the M_1 1-units of the M_d d-units are classified; then it finishes by either 1) working the T_1 tree if 1-units still remain to be classified or 2) working the T_d tree if d-units still remain. For example, if the tree given in (2.1) is denoted by T^* and $(d, q)=(2, \cdot 9)$, then the expected number of 1 and 2-units classified per T^* tree is given by 1 and $q+q^3+q^5=2.2195$, respectively. This indicates that for $M_1=M_2$ (large but finite), the 1-unit will remain and we can write H_{T^*} as

$$H_{T^*} = \begin{Bmatrix} T^* \\ T_1 \end{Bmatrix}; \tag{2.7}$$

we note that

$$\begin{Bmatrix} T_1 \\ T_2 \end{Bmatrix} = H_{T_1} = H_{T_2} = \begin{Bmatrix} T_2 \\ T_1 \end{Bmatrix}, \tag{2.8}$$

For any arbitrary tree T_a, let $\mathsf{E}\{T|T_a\}$ be the expected number of tests per T_a tree and let $\mathsf{E}\{N(i)|T_a\}$ be the expected number of i-units classified per T_a tree $(i=1, \mathrm{d})$. Also, let $\mathsf{E}\{H_{T_a}\}$ be equal to the expected number of tests (necessary to classify all the units) for the H_{T_a} procedure. Thus for $(d, q)=(2, \cdot 9)$, $M_1=M_2$ (large but finite) and the H_{T^*} procedure given in (2.7),

$$\mathsf{E}\{H_{T^*}\} = \frac{M_2}{\mathsf{E}\{N(2)|T^*\}}\,\mathsf{E}\{T|T^*\} + \left\{\left[M_1 - \frac{M_2}{\mathsf{E}\{N(2)|T^*\}}\,\mathsf{E}\{N(1)|T^*\}\right] \cdot \left[\frac{\mathsf{E}\{T|T_1\}}{\mathsf{E}\{N(1)|T_1\}}\right]\right\} \approx M_1(1.1803), \tag{2.9}$$

where

$$\mathsf{E}\{T|T^*\} = 3-2q^7 \approx 2.0434, \quad \mathsf{E}\{N(1)|T^*\} = 1, \tag{2.10}$$

$$\mathsf{E}\{N(2)|T^*\} = q+q^3+q^5 \approx 2.2195, \quad \mathsf{E}\{T|T_1\} = 3+q-3q^7 \approx 2.4651,$$

$$\mathsf{E}\{N(1)|T_1\} = (1-q^7)/(1-q) \approx 5.2170.$$

A tree T_a is said to *agree* with another tree T_b if the same number of components is tested on every step of T_a as on the corresponding step of T_b. For the $(d, q)=(2, \cdot 9)$ case, the number of components tested at the initial step, $G(1, 2, 2, 2)$ step, ... of T^* is 7, 3, ... and is the same as on the initial step, $G(1, 1, 1, 1, 1, 1, 1)$ step, ... of T_1; T^*, therefore, agrees with T_1. We do not worry about the fact that T_1 extends longer than T^* and works with $G(1, 1)$ steps.

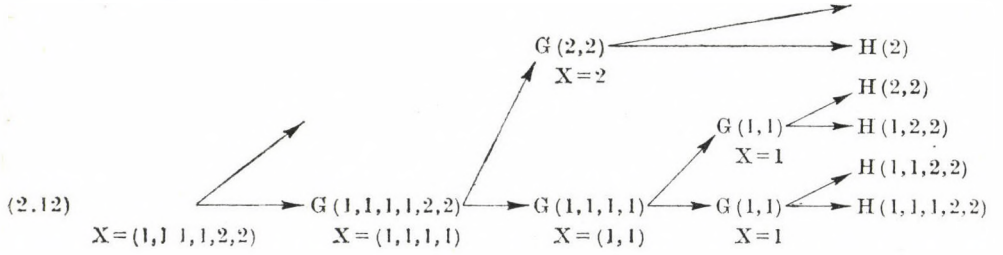
A group of units $(x_1, x_2, \ldots, x_i)$ will be called *coarser* than another group $(y_1, y_2, \ldots, y_j)$ if $i<j$ and there exists ordered unequal integers $l_1, l_2, \ldots, l_{i-1}$ such that

$$x_1 = \sum_{\alpha=1}^{l_1} y_\alpha, \; x_2 = \sum_{\alpha=l_1+1}^{l_2} y_\alpha, \quad \ldots, \; x_k = \sum_{\alpha=l_{i-1}+1}^{j} y_\alpha, \tag{2.11}$$

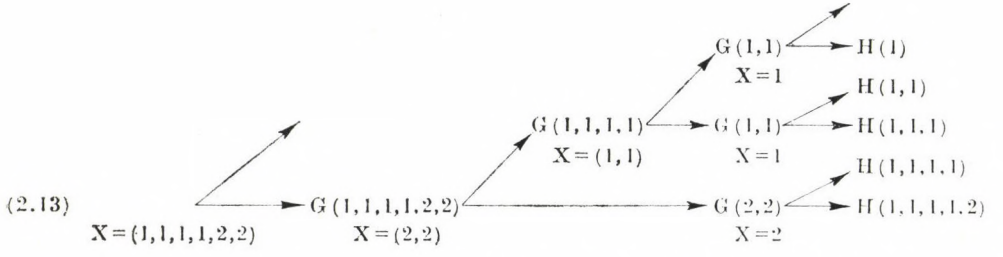
where at least one sum contains 2 or more elements; thus (1, 2, 2, 2) is coarser than (1, 1, 1, 1, 1, 1, 1).

Let $\mathcal{F}$ be a family of trees consisting only of trees that 1) agree with T_1 and 2) on a given G-step, always tests a d-unit in preference to d 1-units. For example, for the case $(d, q)=(2, \cdot 92)$ the trees

(2.12)



(2.13)



denoted by T' and T_f, respectively, both agree with T_1 (given in (2.5)) but only T_f is included in $\mathcal{F}$ because according to the T' tree, we test a (1, 1, 1, 1) grouping at $G(1, 1, 1, 1, 2, 2)$ and thus under T', preference is given to testing 1-units over 2-units.

For T_a and T_b, both included in $\mathscr{F}$, T_a is said to be *coarser* than T_b (or T_b is *finer* than T_a) if the group of units tested on the initial step (or H-step) of T_a is coarser than the corresponding group of T_b. Any two trees in $\mathscr{F}$ are comparable as to their coarseness and thus there exists a coarsest tree T^* in $\mathscr{F}$.

Let $\mathscr{F}'$ consist of trees that 1) agree with T_1 and 2) are identical on their H-step to some procedure in $\mathscr{F}$; by eliminating the preference condition, $\mathscr{F}'$ properl includes $\mathscr{F}$.

3. Main and Auxiliary Results. In [2], it is shown that under the condition that

$$\frac{\mathsf{E}\{T|T_d\}}{\mathsf{E}\{N(d)|T_d\}} \geqq (\mathsf{E}\{N(1)|T(d,1)\})\left(\frac{\mathsf{E}\{T|T_1\}}{\mathsf{E}\{N(1)|T_1\}}\right)-(\mathsf{E}\{T|T(d,1)\}-1), \tag{3.1}$$

$$\mathsf{E}H_{T^*} = \min_{T_f \in \mathscr{F}} \mathsf{E}H_{T_f}, \tag{3.2}$$

where T^* is the coarsest procedure in $\mathscr{F}$, i.e., $\mathsf{E}H_{T^*}$ minimizes $\mathsf{E}H_{T_f}$ over T_f included in $\mathscr{F}$. Let us note that T^*, T_1 and T_2 are functions of (d, q) and are given by (2.1), (2.2) and (2.3) when $(d, q)=(2, \cdot 9)$. Also, in this paper, we substitute "1" and "d" for the "e_1" and "e_2" in [2]. Thirdly, in [2] it is shown that condition (3.1) holds true for many pairs (d, q); it is also conjectured to be true for all (d, q).

We show in this paper that under condition (3.1), $\mathsf{E}H_{T^*}$ minimizes $\mathsf{E}H_{T'}$ over the much wider (than $\mathscr{F}$) class of procedures $\mathscr{F}'$, i.e.,

$$\mathsf{E}H_{T^*} = \min_{T' \in \mathscr{F}'} \mathsf{E}H_{T'}, \tag{3.3}$$

by showing that for any T' in $\mathscr{F}'$, there is a tree T_f in $\mathscr{F}$ such that

$$\mathsf{E}H_{T_f} \leqq \mathsf{E}H_{T'}. \tag{3.4}$$

Three auxiliary statements that must be proved in order to prove (3.4). First of all, for any $T' \in \mathscr{F}'$, there exists a $T_f \in \mathscr{F}$ such that

$$\begin{aligned} \mathsf{E}\{T|T'\} &= \mathsf{E}\{T|T_f\} + B_f, \\ \mathsf{E}\{N(1)|T'\} &= \mathsf{E}\{N(1)|T_f\} + A_f, \\ \mathsf{E}\{N(d)|T'\} &= \mathsf{E}\{N(d)|T_f\} - C_f, \end{aligned} \tag{3.5}$$

where

$$\begin{aligned} B_f &= W_f(\mathsf{E}\{T|T,(d,1)\}-1), \\ A_f &= W_f(\mathsf{E}\{N(1)|T(d,1\}), \\ C_f &= W_f, \end{aligned} \tag{3.6}$$

for some positive quantity, W_f.

Secondly, for any tree T, which begins (on its H-step) by testing a^* 1-units and b^* d-units

$$\mathsf{E}\{N(1)|T\} = \mathsf{E}\{N(1)|T(a^*+db^*,1)\} - \{(\mathsf{E}\{N(1)|T(d,1)\})(\mathsf{E}\{N(d)|T\})\}. \tag{3.7}$$

We note that for $T=T_f \in \mathscr{F}$

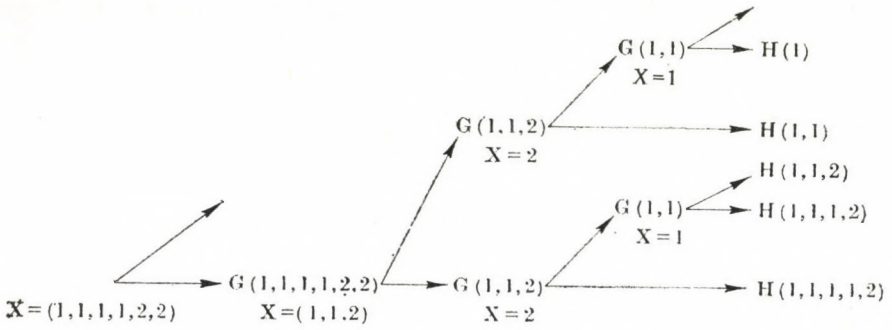
$$T(a^*+db^*,1) = T_1. \tag{3.8}$$

Thirdly, for any tree $T_f \in \mathscr{F}$, which begins by testing a^* 1-units and b^*d-units

$$\mathsf{E}\{T|T_f\} = \mathsf{E}\{T|T_1\} - (\mathsf{E}\{T|T(d,1)\}-1)\mathsf{E}\{N(d)|T_f\}. \tag{3.9}$$

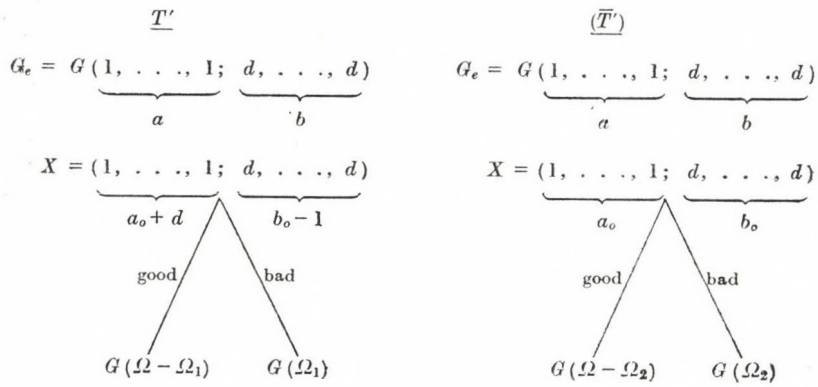
4. Proof of (3.5)—(3.6). Let T' be included in $\mathscr{F}'$ but not in $\mathscr{F}$, i.e., $T' \in \mathscr{F}' - \mathscr{F}$. Locate a G-step on T' in which a preference is given for d 1-units over a d-unit and where for all succeeding G-steps (if there are any), preference is always given to the d-unit; denote this G-step by G_e and let c denote the number of d-units that are bypassed (at G_e) in favor of 1-units ($c=1, 2, \ldots$). Let $(\overline{T'}) \in \mathscr{F}'$ be the tree which is identical to T' except for being amended so that at G_e only $(c-1)$ d-units are bypassed and for all succeeding G-steps (arising from G_e), it still gives preference to the d-unit. Using $(d, q)=(2, \cdot 9)$ as an example, we can let T' be given by (2.12) and $(\overline{T'})$ be given by

(4.1)



Here at $G_e=G(1, 1, 1, 1, 2, 2)$, 1-units are tested ahead of $c=2$ 2-units on T', whereas on $(\overline{T'})$, 1-units are tested ahead of only $(c-1)=1$ 2-unit. A general picture of T' and $(\overline{T'})$ around G_e would look like

(4.2)



where

(4.3)

$$\Omega = \{\underbrace{1, \ldots, 1}_{a}; \underbrace{d, \ldots, d}_{b}\}, \qquad \Omega_1 = \{\underbrace{1, \ldots, 1}_{a_o + d}; \underbrace{d, \ldots, d}_{b_o - 1}\}$$

$$\Omega_2 = \{\underbrace{1, \ldots, 1}_{a_o}; \underbrace{d, \ldots, d}_{b_o}\}$$

$(0 \leqq a_0 \leqq a-d,\ 1 \leqq b_0 \leqq b)$. Let EG (resp., EG_1) be the conditional expected number of tests (resp., 1-units classified) given the $G(\underbrace{1, \ldots, 1}_{d})$ step of $T(d, 1)$ as a starting point to complete (resp., in completing) the $T(d, 1)$ tree. We note that

$$EG = \frac{\mathsf{E}\{T|T(d,1)\}-1}{1-q^d}, \quad EG_1 = \frac{\mathsf{E}\{N(1)|T(d,1)\}-dq^d}{1-q^d}. \tag{4.4}$$

Let (1) $\hat{p}$ (resp., $\hat{\hat{p}}$) be equal to the probability on the T' tree of reaching the $G(\Omega-\Omega_1)$ step and then subsequently classifying all the d-units contained in the G_e defective set with the last one classified being bad (resp., good),

(2) $\hat{\hat{p}}$ (resp., $\breve{\breve{p}}$) be equal to the probability on the $(\overline{T'})$ tree of reaching the $G(\Omega_2)$ step and then subsequently classifying exactly b_0 of the d-units contained in the G_e defective set with the last one classified being bad (resp., good),

(3) l be defined by

$$l = q^{m_1+dm_d}(1-q)^{n_1}(1-q^d)^{n_d}, \tag{4.5}$$

where m_i (resp., n_i) is the number of i-units that have been classified as good (resp., bad) by the time G_e is reached $(i=1, d)$.

It is then clear from (4.2) that

$$\begin{aligned} \mathsf{E}\{T|T'\} &= \mathsf{E}\{T|(\overline{T'})\}+B, \\ \mathsf{E}\{N(1)|T'\} &= \mathsf{E}\{N(1)|(\overline{T'})\}+A, \\ \mathsf{E}\{N(d)|T'\} &= \mathsf{E}\{N(d)|(\overline{T'})\}-C, \end{aligned} \tag{4.6}$$

where

$$\begin{aligned} B &= (\hat{\hat{p}}-\hat{p})EG, \\ A &= \{dl[q^{a_0+db_0}-q^{a+db}]-\hat{p}EG_1-d\breve{p}\}+\{\hat{\hat{p}}EG_1+d\breve{\breve{p}}\}, \\ C &= \{l[q^{a_0+db_0}-q^{a+db}]-\hat{p}-\breve{p}\}+(\hat{\hat{p}}+\breve{\breve{p}}). \end{aligned} \tag{4.7}$$

Let (1) $\bar{p}$ (resp., $\bar{\bar{p}}$) be equal to the probability on the T' (resp., $(\overline{T'})$) tree of reaching the $G(\Omega-\Omega_1)$ (resp., $G(\Omega_2)$) step and subsequently classifying all (resp., b_0) of the d-units contained in G_e and let (2) $\hat{a}$ (resp., $\hat{\hat{a}}$) be equal to the number of 1-units contained in G_e that one must classify in order to classify all (resp., b_0) of the d-units contained in G_e, on the T' (resp., $(\overline{T'})$) tree. One can easily see that

$$\begin{aligned} \bar{p} &= l\{q^{\hat{a}}q^{d(b-1)}-q^{a+db}\}, \\ \hat{p} &= l\{q^{\hat{a}}q^{d(b-1)}(1-q^d)\} = l\left\{\left(\frac{\bar{p}}{l}+q^{a+db}\right)(1-q^d)\right\}, \\ \breve{p} &= l\{q^{\hat{a}}q^{db}-q^{a+db}\} = l\left\{q^d\frac{\bar{p}}{l}-q^{a+db}(1-q^d)\right\}, \\ \bar{\bar{p}} &= l\{q^{\hat{\hat{a}}}+d(b_0-1)-q^{a_0+db_0}\}, \\ \hat{\hat{p}} &= l\left\{\left(\frac{\bar{\bar{p}}}{l}+q^{a_0+db_0}\right)(1-q^d)\right\}, \\ \breve{\breve{p}} &= l\left\{q^d\left(\frac{\bar{\bar{p}}}{l}\right)-q^{a_0+db_0}(1-q^d)\right\}, \end{aligned} \tag{4.8}$$

where $a_0+d\leqq\hat{a}\leqq a$ and $0\leqq\hat{\hat{a}}\leqq a_0$. Substituting (4.4) and (4.8) into (4.7), one obtains the following simplified version of (4.7),

$$B = W(\mathsf{E}\{T|T(d,1)\}-1) \tag{4.9}$$
$$A = W(\mathsf{E}\{N(1)|T(d,1)\}),$$
$$C = W,$$

where

$$W = l\left\{\left(\frac{\bar{p}}{l}+q^{a_0+db_0}\right)-\left(\frac{\bar{p}}{l}+q^{a+db}\right)\right\}. \tag{4.10}$$

For $(d, q)=(2, .92)$, where T' and $(\overline{T'})$ are given by (2.12) and (4.1): $a=4$, $b=2$, $a_0=2$, $b_0=1$, $\hat{a}=4$, $\hat{\hat{a}}=0$, $l=1$. One can then easily verify that (4.6), (4.9) and (4.10) hold true for this case.

Having already "barred" T' to form $(\overline{T'})$, we continue by barring $(\overline{T'})$ to form $(\overline{\overline{T'}})=(\overline{(\overline{T'})})$, then barring $(\overline{\overline{T'}})$ to form $(\overline{\overline{\overline{T'}}})$ and so on until we obtain the finite sequence of trees

$$T', (\overline{T'}), (\overline{\overline{T'}}), \ldots, T_f; \tag{4.11}$$

it follows from the definitions of $\mathscr{F}$ and the barring operation that T_f is of necessity the last term in the sequence. In our $(d, q)=(2, .92)$ example, for T' given by (2.12), $T_f=(\overline{\overline{\overline{T'}}})$ is given by (2.13). Reapplying (4.6) and (4.9) to the pairs $((\overline{T'}),(\overline{\overline{T'}}))$, $((\overline{\overline{T'}}),(\overline{\overline{\overline{T'}}}))$, enables us to write
$\mathsf{E}\{T|T'\}$, $\mathsf{E}\{N(1)|T'\}$ and $\mathsf{E}\{N(d)|T'\}$ in terms of
$\mathsf{E}\{T|T_f\}$, $\mathsf{E}\{N(1)|T_f\}$ and $\mathsf{E}\{N(d)|T_f\}$, respectively,
in the form

$$\mathsf{E}\{T|T'\} = \mathsf{E}\{T|T_f\}+B_f, \tag{4.12}$$
$$\mathsf{E}\{N(1)|T'\} = \mathsf{E}\{N(1)|T_f\}+A_f,$$
$$\mathsf{E}\{N(d)|T') = \mathsf{E}\{N(d)|T_f\}-C_f,$$

where

$$B_f = W_f(\mathsf{E}\{T|T(d,1)\}-1), \tag{4.13}$$
$$A_f = W_f(\mathsf{E}\{N(1)|T(d,1)\}),$$
$$C_f = W_f,$$

for some appropriately defined W_f; this concludes the proof of (3.5)—(3.6).

5. Proof of (3.7). With reference to T (of (3.7)), let a_i be the number of additional 1-units that one must classify in order to classify i d-units over that needed to classify $(i-1)$ d-units ($i=1, 2, \ldots, b$), i.e., describing the above schematically, we have as the order of possible classification for the T tree

$$\underbrace{1,\ldots,1}_{a_1},\ \underset{1^{st}}{d},\ \underbrace{1,\ldots,1}_{a_2},\ \underset{2^{nd}}{d},\ \ldots,\ \underbrace{1,\ldots,1}_{a_{b^*}},\ \underset{(b^*)^{th}}{d},\ \underbrace{1,\ldots,1}_{a_{b^*+1}}; \tag{5.1}$$

where

$$a_{b^*+1} = a^* - \sum_{i=1}^{b^*} a_i. \tag{5.2}$$

It thus follows that

$$\begin{aligned} \mathsf{E}\{N(d)|T\} &= q^{a_1}(1-q^{a_2+d}) + 2q^{a_1+a_2+d}(1-q^{a_3+d}) + \ldots + \\ &+ (b^*-1)q^{\left(\sum_{i=1}^{b^*-1} a_i\right)+d(b^*-2)}(1-q^{a_{b^*}+d}) + b^* q^{\left(\sum_{i=1}^{b^*} a_i\right)+d(b^*-1)} = \\ &= q^{a_1} + q^{a_1+a_2+d} + \ldots + q^{\left(\sum_{i=1}^{b^*} a_i\right)+d(b^*-1)} \end{aligned} \tag{5.3}$$

and that

$$\begin{aligned} \mathsf{E}\{N(1)|T\} &= q^0 + q^{0+1} + \ldots + q^{0+(a-1)} + \\ &+ q^{d+a_1} + q^{d+(a_1+1)} + \ldots + q^{d+(a_1+a_2-1)} + \\ &\vdots \\ &+ q^{db^*+\sum_{i=1}^{b^*} a_i} + \ldots + q^{db^*+\left(\sum_{i=1}^{b^*+1} a_i\right)-1} \\ &= \{(1-q^{a_1}) + q^{(d+a_1)}(1-q^{a_2}) + \ldots + \\ &+ q^{\left(db^*+\sum_{i=1}^{b^*} a_i\right)}(1-q^{a_{b^*+1}})\}/\{1-q\} = \\ &= (1-q^{a^*+db^*})/(1-q) - [(1-q^d)/(1-q)][\mathsf{E}\{N(d)|T\}]. \end{aligned} \tag{5.4}$$

From [3], we see that

$$\mathsf{E}\{N(1)|T(j,1)\} = \frac{1-q^j}{1-q} \quad (j \geqq 1), \tag{5.5}$$

which completes the proof of (3.7).

6. Proof of (3.9). A pair of trees in $\mathscr{F}$, denoted by T_{co} and T_{FI}, are said to be adjacent if T_{FI}, the finer tree in the pair, tests exactly d more 1-units and one less d-unit on the H-step than the coarser tree T_{co}. Let (T_{co}, T_{FI}) be an adjacent pair in $\mathscr{F}$. It is shown in [2] that

$$\mathsf{E}\{T|T_{co}\} = \mathsf{E}\{T|T_{FI}\} - [\mathsf{E}\{T|T(d,1)\} - 1]Q_{co}, \tag{6.1}$$

where Q_{co} is the probability with reference to the T_{co} tree that all d-units tested on its initial step are classified before a return to the next H-state, i.e., if T_{co} begins by testing b^* d-units, Q_{co} is the probability that all b^* of the d-units are classified by the end of the tree. It is clear that beginning with some $T_f \in \mathscr{F}$, one can work one's way to T_1, by the sequence of adjacent pairs,

$$(T_f, (T_{FI})_1), \ldots, \quad ((T_{FI})_{b^*-1}, T_1), \tag{6.2}$$

where $(T_{FI})_i$ is the finer tree in the ith pair ($i=1, 2, \ldots, b^*$). Upon applying (6.1) to each of the b^* pairs, it follows that

$$\mathsf{E}\{T|T_f\} = \mathsf{E}\{T|T_1\} - [\mathsf{E}\{T|T(d,1)\} - 1]\left[\sum_{i=1}^{b^*} Q_i\right], \tag{6.3}$$

where Q_i is the probability that all the d-units are classified with the coarser tree of the ith pair. One immediately notices that Q_i can be defined equivalently as the probability that at least $[b^*-(i-1)]$ of the d-units on the T_f tree are classified and thus (by elementary probability)

$$\sum_{i=1}^{b^*} Q_i = \mathsf{E}\{N(d)|T_f\}, \tag{6.4}$$

which (from (6.3)) proves (3.9).

7. Proof of (3.4). Given a tree $T' \in \mathscr{F}'$, locate the corresponding tree $T_f \in \mathscr{F}$ by the barring operation given in section 4. Define r to be equal to M_1/M_d and for the purposes of this proof, we can assume without loss of generality that $M_d \equiv 1$. "1" could refer to 1,000,000 (say) and this certainly does not contradict the fact that we are assuming that M_1 and M_d are large but finite. Let us break up the proof into three parts, when

(1) $$r \leqq r^*,$$

(2) $$r^* < r \leqq r^{**},$$

(3) $$r^{**} < r,$$

where

$$r^* = \frac{\mathsf{E}\{N(1)|T_f\}}{\mathsf{E}\{N(d)|T_f\}}, \quad r^{**} = \frac{\mathsf{E}\{N(1)|T'\}}{\mathsf{E}\{N(d)|T'\}}. \tag{7.1}$$

Proof for case 1. In this case,

$$H_{T_f} = \begin{Bmatrix} T \\ T_d \end{Bmatrix}, \quad H_{T'} = \begin{Bmatrix} T' \\ T_d \end{Bmatrix}, \tag{7.2}$$

we note that

$$\mathsf{E}H_{T_f} = \frac{M_1}{\mathsf{E}\{N(1)|T_f\}} \mathsf{E}\{T|T_f\} + (1-M_1) \frac{\mathsf{E}\{N(d)|T_f\}}{\mathsf{E}\{N(1)|T_f\}} \frac{\mathsf{E}\{T|T_d\}}{\mathsf{E}\{N(d)|T_d\}} \tag{7.3}$$

and $\mathsf{E}H_{T'}$ has the same form as $\mathsf{E}H_{T_f}$ above with T_f replaced by T'. We first utilize (3.5) to rewrite $\mathsf{E}H_{T'}$ in terms of $\mathsf{E}\{T|T_f\}$, $\mathsf{E}\{N(1)|T_f\}$ and $\mathsf{E}\{N(d)|T_f\}$; after which, simple algebraic manipulation enables us to arrive at the result

$$[\mathsf{E}H_{T_f} \leqq \mathsf{E}H_{T'}] \Leftrightarrow \left[\frac{\mathsf{E}\{T|T_d\}}{\mathsf{E}\{N(d)|T_d\}} \geqq \frac{A_f \mathsf{E}\{T|T_f\} - B_f \mathsf{E}\{N(1)|T_f\}}{A_f \mathsf{E}\{N(d)|T_f\} + C_f \mathsf{E}\{N(1)|T_f\}}\right] \Leftrightarrow \tag{7.4}$$

$$\Leftrightarrow \left[\frac{\mathsf{E}\{T|T_d\}}{\mathsf{E}\{N(d)|T_d\}} \geqq \frac{(\mathsf{E}\{N(1)|T(d,1)\})\mathsf{E}\{T|T_f\} - (\mathsf{E}\{T|T(d,1)\}-1)\mathsf{E}\{N(1)|T_f\}}{\mathsf{E}\{N(1)|T(d,1)\}\mathsf{E}\{N(d)|T_f\} + \mathsf{E}\{N(1)|T_f\}}\right]$$

where "$\Leftrightarrow$" stands for "if and only if". In obtaining the last bracketed expression in (7.4), we make use of (3.6).

In [2], it was proved that the tree-function

$$H(T_0) = \frac{(\mathsf{E}\{N(1)|T(d,1)\})\mathsf{E}\{T|T_0\} - (\mathsf{E}\{T|T(d,1)\}-1)\mathsf{E}\{N(1)|T_0\}}{(\mathsf{E}\{N(1)|T(d,1)\})\mathsf{E}\{N(d)|T_0\} + \mathsf{E}\{N(1)|T_0\}} \tag{7.5}$$

is invariant over $T\in\mathscr{F}$. Utilizing (7.4) and the fact that

$$H(T_f) = H(T_1) \tag{7.6}$$

(remember that $T_1\in\mathscr{F}$) results in the proof of (3.4) for this case.

Proof for Case (2). In this case,

$$H_{T_f} = \begin{Bmatrix} T_f \\ T_1 \end{Bmatrix}, \quad H_{T'} = \begin{Bmatrix} T' \\ T_d \end{Bmatrix}. \tag{7.7}$$

Let $\mathsf{E}H_{T_f}(r_0)$ [resp., $\mathsf{E}H_{T'}(r_0)$] be the value of $\mathsf{E}H_{T_f}$[resp,, $\mathsf{E}H_{T'}$] at $r=r_0$. It is easy to see that for $r^*<r\leqq r^{**}$,

$$\begin{aligned} \mathsf{E}H_{T_f}(r) &= \mathsf{E}H_{T_f}(r^*)+(r-r^*)\Delta_f, \\ \mathsf{E}H_{T'}(r) &= \mathsf{E}H_{T'}(r^*)+(r-r^*)\Delta', \end{aligned} \tag{7.8}$$

where

$$\Delta_f = \frac{\mathsf{E}\{T|T_1\}}{\mathsf{E}\{N(1)|T_1\}}, \tag{7.9}$$

$$\Delta' = \frac{1}{\mathsf{E}\{N(1)|T_f\}+A_f}\left[\mathsf{E}\{T|T_f\}+B_f-(\mathsf{E}\{N(d)|T_f\}-C_f)\frac{\mathsf{E}\{T|T_d\}}{\mathsf{E}\{N(d)|T_d\}}\right].$$

We first show that under condition (3.1)

$$\Delta' \leqq \Delta_f. \tag{7.10}$$

Using (3.7), (3.8), (3.9), (3.6) to substitute in Δ' for $\mathsf{E}\{N(1)|T_f\}$, $\mathsf{E}\{T|T_f\}$, A_f, B_f and C_f, results in a verification of (7.10). It thus follows from (7.8) that proving (3.4) in this case is tantamount to proving that

$$\begin{aligned} \mathsf{E}H_{T_f}(r^{**}) &= r^*\frac{\mathsf{E}\{T|T_f\}}{\mathsf{E}\{N(1)|T_f\}}+(r^{**}-r^*)\frac{\mathsf{E}\{T|T_1\}}{\mathsf{E}\{N(1)|T_1\}} \\ &\leqq \mathsf{E}H_{T'}(r^{**}) = r^{**}\frac{\mathsf{E}\{T|T_f\}+B_f}{\mathsf{E}\{N(1)|T_f\}+A_f}. \end{aligned} \tag{7.11}$$

Upon substituting (from (3.7), (3.8), (3.9), (3.6), (7.1), and (3.5)) for $\mathsf{E}\{N(1)|T_f\}$, $\mathsf{E}\{T|T_f\}$, A_f, B_f, C_f, r^*, and r^{**}, one sees immediately that (7.11), is true, which proves (3.4) for this case.

Proof for Case (3). For $r>r^{**}$,

$$H_{T_f} = \begin{Bmatrix} T_f \\ T_1 \end{Bmatrix}, \quad H_{T'} = \begin{Bmatrix} T' \\ T_1 \end{Bmatrix}. \tag{7.12}$$

The proof of (3.4) for this case follows immediately from the result of the previous case by noting that for $r>r^{**}$,

$$\begin{aligned} \mathsf{E}H_{T_f}(r) &= \mathsf{E}H_{T_f}(r^{**})+(r-r^{**})\frac{\mathsf{E}\{T|T_1\}}{\mathsf{E}\{N(1)|T_1\}}, \\ \mathsf{E}H_{T'}(r) &= \mathsf{E}H_{T'}(r^{**})+(r-r^{**})\frac{\mathsf{E}\{T|T_1\}}{\mathsf{E}\{N(1)|T_1\}}; \end{aligned} \tag{7.13}$$

this completes the proof of (3.4).

8. Examples. For the $(d, q)=(2, .9)$ case, the conjectured optimal (non-mixing for M_1 and M_2 large, as usual) procedure is H_{T^*}, where T^* is given by (2.1). If $M_1=M_2$, then $\mathsf{E}H_{T^*}\approx M_1$ (1.1803), while classifying the 1 and 2-units separately results in a value of $\mathsf{E}H_{T_1}\approx M_1$ (1.1810); H_{T^*} thus shows savings over H_{T_1} in the expected number of tests necessary to classify all the units. One can derive an information lower bound (ILB) over all procedures on the expected number of tests for this case (see [2]) and get the value, $\mathrm{ILB}\approx M_1(.4690)+M_2(.7015)$; this lower bound is not always attainable and for $M_1=M_2$, $\mathrm{ILB}\approx M_1(1.1705)$. For the $(d, q)=(2, .92)$ case the conjectured optimal procedure is $H_{T_2}=H_{T_1}$, since $T^*=T_2$ (given in (2.6)) is the coarsest procedure in $\mathscr{F}$. If $M_1=M_2$, then $\mathsf{E}H_{T_1}\approx M_1(1.0270)$ and the ILB for this case is given by $\mathrm{ILB}\approx M_1(.4022)+M_2(.6188)=M_1(1.0210)$.

9. Extensions. A result analogous to (3.3) for the general (d, q) problem, where d is allowed to be any non-negative number, would be desirable. Other possible extensions would be in proving that H_{T^*} is the optimal non-mixing procedure or in finding the overall optimal procedure. We could also begin work on the case of more than two different types of units.

REFERENCES

[1] DORFMAN, R.: The detection of defective members of large populations. *Ann. Math. Statist.* **14** (1943), 436—440.

[2] NEBENZAHL, E. and SOBEL, M.: Finite and infinite models for generalized grouptesting with unequal probabilities of success for each item. A paper in *Discriminant Analysis and Applications*, ed. T. CACOULLOS. Academic Press Inc., New York (1973).

[3] SOBEL, M. and GROLL, P. A.: Group-testing to eliminate efficiently all defectives in a binomial sample. *Bell System Techn. Jour.* **38** (1959), 1179—1252.

[4] SOBEL, M.: Group-testing to classify all defectives in a binomial sample. A paper in *Information and Decision Processes*, ed. R. E. MACHOL. McGraw-Hill Book Co., New York, 127—161 (1960).

California State University, Hayward

(Received March 4, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 73—74.

AN EXTENSION OF AN INEQUALITY OF HOEFFDING

by

P. K. PATHAK

1. Statement of the result. The object of this note is to establish the following theorem.

THEOREM. *Let π be a finite population of N real numbers. Let $(X_1, \ldots, X_n)$ denote a simple random sample (with replacement) of size n from π and $\{Y_1, Y_2, \ldots, Y_n\}$, a simple random sample (without replacement) of size n from π. Then for every convex function g,*

$$\mathsf{E}[g(s_x^2)] \geqq \mathsf{E}[g((1-N^{-1})s_y^2)] \tag{1.1}$$

where s_x^2 and s_y^2 denote the sample variances based on $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ respectively.

An inequality analogous to (1.1) involving the sample total $\sum X_i$ and $\sum Y_i$ is originally due to HOEFFDING [1] (cf. ROSEN [3] in this connection). In this note we show that inequalities of this kind are consequences of the well-known technique of Rao-Blackwellization in sampling theory.

We turn now to the proof of the above theorem.

2. Proof of the theorem. The proof presented here involves the notion of sufficiency in sampling. The reader may refer to the paper [2] for necessary details. Now suppose units are drawn one-by-one with replacement from the population π until the sample first contains n distinct population units. Let $S=(X_1, \ldots, X_M)$ denote the sequence of units observed in the sample. (Note that $M \geqq n$ and repetitions may occur in S.) Let $T=\{Y_1, Y_2, \ldots, Y_n\}$ denote that set of distinct units in S. Then T is a sufficient statistic for this sampling scheme (cf. [2]). It is easily seen that the statistic T represents a simple random sample (without replacement) of size n from π and the first n observations in S, namely $X_1, \ldots, X_n$, represent a simple random sample (with replacement) of size n from π. Now based on $(X_1, \ldots, X_n)$, the sample variance $s_x^2=[1/(n-1)]\sum_{i=1}^{n}(X_1-\overline{X})^2$ is clearly an unbiased estimator of the population variance σ^2, say. Analogously based on the statistic $T=\{Y_1, \ldots, Y_n\}$ the sample variance $s_y^2=[1/(n-1)]\sum_{i=1}^{n}(Y_i-\overline{Y})^2$ is easily seen to be an unbiased estimator of a slightly different population variance $S^2=[N/(N-1)]\sigma^2$. By virtue of the sufficiency of T, it follows that $\mathsf{E}[s_x^2|T]$ is an unbiased estimator of σ^2 and for any convex loss function

g, $\mathsf{E}[g(s_x^2)] \geqq \mathsf{E}[g(\mathsf{E}(s_x^2|T))]$. To complete the proof of the theorem it will suffice to show that $\mathsf{E}[s_x^2|T] = (1-N^{-1})s_y^2$. Clearly

$$\text{(2.1)} \qquad \mathsf{E}[s_x^2|T] = \mathsf{E}\left[\frac{1}{2n(n-1)} \sum_{i \neq i'=1}^{n} (X_i - X_{i'})^2 | T\right] = \mathsf{E}\left[\frac{1}{2}(X_1 - X_2)^2 | T\right].$$

Owing to the symmetric nature of the sampling scheme, it follows that given $T = \{Y_1, Y_2, \ldots, Y_n\}$, the pair (X_1, X_2) assumes as its value any one of the pairs (Y_i, Y_i), $i \neq i' = 1, \ldots, n$, with equal probabilities. (Note than (X_1, X_2) assumes also as its value the pair (Y_i, Y_i), $i = 1, \ldots, n$, with a different probability, but this probability does not enter in our calculations.) Consequently we can write

$$\text{(2.2)} \qquad \mathsf{E}[s_x^2|T] = \frac{1}{2} \sum_{i=i'=1}^{n} (Y_i - Y_{i'})^2 \frac{k}{n(n-1)} = k s_y^2$$

where k is a suitable constant. Since $\sigma^2 = \mathsf{E}[s_x^2] = k\mathsf{E}[s_y^2] = [kN/(N-1)]\sigma^2$, $k = (1-N^{-1})$. This completes the proof of the theorem.

The following corollary follows easily from Theorem 1 and Chebycheff's inequality.

Corollary. Under simple random sampling without replacement of size n, the sample variance s_y^2 satisfies the following probability inequality

$$\text{(2.3)} \qquad \mathsf{P}[|s_y^2 - \mathsf{E}s_y^2| \geqq k] \leqq \frac{[N/(N-1)]^2}{k^2} \left[\frac{(\mu_4 - \sigma^4)}{n} + \frac{2\sigma^4}{n(n-1)}\right]$$

where μ_4 denotes the fourth central moment in the population.

REFERENCES

[1] HOEFFDING, W.: Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.* **58** (1963), 13—30.
[2] PATHAK, P. K.: Sufficiency in sampling theory, *Ann. Math. Statist.* **35** (1964), 795—808.
[3] ROSÉN, B.: On an inequality of Hoeffding, *Ann. Math. Statist.* **38** (1967), 382—392.

University of New Mexico

(Received April 30. 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 75—79.

SOME REMARKS ON SATURATED SETS OF POINTS

by

L. FEJES TÓTH

A set S of convex domains is said to be *saturated with respect to a convex domain* D, if any congruent replica of D has a point in common with a member of S. Similarly, we say that S is *saturated with respect to translates of* D, if any translate of D has a point in common with a domain of S.

These notions give rise to various problems [1, ..., 7]. Here we are especially interested in the following

PROBLEM. Find the thinnest set of translates of a closed convex domain A saturated with respect to translates of a closed convex domain B.

Here the term "thinnest" means "of minimal density", with the usual definition [1] of the area-density or number-density. The former may be interpreted as the total area of the domains divided by the total area of the plane. In this paper we prefer to use the number-density, in short the *density*, which gives the average number of the domains for the unit area. It is equal to the area-density divided by the area of A.

We start with the following

Remark 1. If the closed convex domains A and B are centro-symmetric, then a set of translates of A is saturated with respect to translates of B if and only if the translates of the Minkowski sum $A+B$ concentric with the respective translates of A cover the plane.

This is obvious because the Minkowski sum $A+B$ concentric with A consits of the centres of those translates of B which have a point in common with A.

E. MAKAI jr. observed that Remark 1 can be generalised as follows:

*Remark 1.** If A and B are closed convex domains, then a set of translates of A is saturated with respect to translates of B if and only if the repsective translates of the Minkowski sum $-B+A$ cover the plane.

Thus the above problem is equivalent with the problem of the thinnest covering of the plane with translates of $-B+A$. Let us consider the case that both A and B are centro-symmetric. Then we can refer to the fact [1, 8, 9] that the density of a covering of the plane with translates of a centro-symmetric convex domain D is never less than the density of the thinnest lattice-covering with translates of D. The latter is given by $1/H$, where H is the area of a hexagon of maximal area inscribed in D. According to a theorem of DOWKER [10], among the hexagons of maximal area inscribed in D, there is one which has central symmetry. Tessellating the plane with translates of this hexagon, and circumscribing about each hexagon a translate of D, we obtain a thinnest covering with translates of D. Since $A+B$ is convex and centro-

symmetric, our problem is reduced to the problem of finding a centro-symmetric hexagon of maximal area inscribed in $A+B$.

As an example, let A be a unit square and B the unit square arising from A by a rotation through 45°. Now $A+B$ is a regular octagon of sidelength 1. There are infinitely many centrosymmetric hexagons of maximal area inscribed in a regular octagon. Accordingly, the problem of the thinnest set of translates of A saturated with respect to translates of B has infinitely many solutions consisting of rows of equally spaced squares. One of the solutions is exhibited in Fig. 1. Within certain limits each row can be translated in its own direction, obtaining irregalur solutions of the problem.

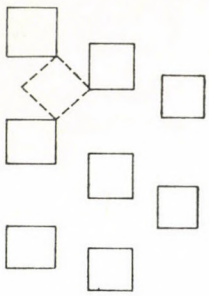


Figure 1.

Special attention is due to the limiting case of the above problem when A is a point: The density d of the thinnest set of points saturated with respect to translates of a centro-symmetric convex domain B is given by $d=1/H$, where H is the area of the hexagon of maximal area inscribed in B.

Let us recall the theorem of Sas [1, 11]: If T_n is the area of the n-gon of maximal area inscribed in a convex domain of area T, then

$$T_n \geqq T\frac{n}{2\pi}\sin\frac{2\pi}{n}.$$

Equality holds only for an ellipse. Thus if B has unit area, we have $\sqrt{27}/2\pi \leqq H \leqq 1$, showing the correctness of the following

Remark 2. If d is the density of the thinnest set of points saturated with respect to translates of a closed centro-symmetric convex domain of unit area then $1 \leqq d \leqq \leqq 2\pi/\sqrt{27}$. Equality holds on the left hand side only for quadrangles and hexagons, and on the right hand side only for ellipses.

It seems to be an interesting but difficult question, whether the thinnest set of points saturated with respect to translates of a convex domain without central symmetry constitutes a lattice, more precisely, whether there is a thinner saturated point-set than the thinnest saturated point-lattice. Since the answer to this question is not known yet, we restrict ourselves to point-lattices.

Remark 3. If d is the density of the thinnest lattice of points saturated with respect to translates of a closed convex domain of unit area, then $1 \leqq d \leqq 3/2$. Equality holds on the left hand side only for parallelograms and centro-symmetric hexagons, and on the right hand side only for triangles.

It is easily seen that if L is a thinnest point-lattice saturated with respect to translates of the convex domain B, then there is a translate of B containing on its boundary three vertices, say T, U and V, of a basic parallelogram $TUVU'$. We claim that this translate of B, let us call it simply B, contains a translate of the triangle TVU'. To see this we may suppose that, in its present position, B does not contain any of the triangles TVU', VUT' and UTV', where T' and V' are the mirror images of U' reflected in V and T, respectively (Fig. 2.). Again, we may suppose that the lines $T'U'$, $U'V'$ and $V'T'$ intersect B in segments shorter than TU, UV and VT, re-

spectively, because otherwise we could translate B in the direction of these lines so as to contain one of the triangles $VU'T$, $TV'U$ or $UT'V$. Finally, we may suppose that besides U, T and V B does not contain any other lattice-point. Indeed, since B is convex and it does not contain any of the lattice-points T', U' and V', it is enough to rule out the lattice-points T_V, T_U, V_U, V_T, U_T, U_V where $X_Y = Y + \overrightarrow{XY}$ denotes the image of the point X reflected in the point Y. Suppose that B contains one of these points, say U_T. Then B lies in the half-plane bounded by the line UT containing V. Since $U'T'$ intersects B in a segment shorter than UT we can translate B in the direction UT so as to separate it from all lattice points lying in $U'T'$. A subsequent small translation in the direction $\overrightarrow{TV}$ or $\overrightarrow{UV}$ would isolate B from all lattice-points. But this is impossible because the lattice was supposed to be saturated.

Consider the chords of B parallel to TV which intersect the triangle TUV. (By a chord we mean a segment joing two boundary points of B.) Among these chords there must be one, say T^*V^*, other than TV but congruent to it. For, the chords under consideration cannot be all smaller than TV, because otherwise B could be translated so as to get isolated from all lattice points. On the other hand, if among the chords there is one which is greater than TV, then there must be one chord which is equal to TV, because $T'V'$ intersects B in a segment smaller than TV. If the condition $\overline{T^*V^*} = \overline{TV}$ does not determine the chord T^*V^* uniquely, we choose the chord farthest from TV.

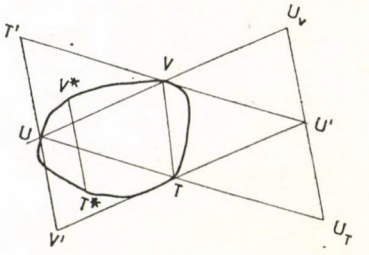


Figure 2.

T^* does not lie on the segment UT, since otherwise B would lie in the half-plane bounded by the line UT containing V, what we have seen to be impossible.

We claim that the translate of B, $B + \overrightarrow{T^*T}$, does not contain U_T. To see this consider a half-plane h containing B and bounded by a line containing T, and observe that the translation $\overrightarrow{T^*U}$ carries T in the interior of h. Therefore $B + \overrightarrow{T^*U}$ does not contain T, and consequently $B + \overrightarrow{T^*T} = B + \overrightarrow{UT} + \overrightarrow{T^*U}$ does not contain U_T. Similarly we see that $B + \overrightarrow{T^*T}$ does not contain the point U_V. It follows that it contains U', because otherwise the lattice would not be saturated.

Since B contains the triangle TUV, as well as an image of this triangle reflected in a point, B contains the convex hull C of these two triangles. C is a centro-symmetric hexagon or a parallelogram. Considering a parallelogram as the limiting case of a centro-symmetric hexagon, we can say that the lattice is generated by three non-adjacant vertices of C. The density of the lattice is equal to $1/H$, where H is the area of C.

Any lattice generated in this way by a centro-symmetric hexagon contained in B is saturated with respect to translates of B, because it is saturated with respect to translates of the hexagon. Thus the problem of finding the thinnest point-lattice saturated with respect to B is reduced to the problem of finding the centro-symmetric hexagon of maximal area H contained in B. If B is of unit area, then $H \leqq 1$. On the other hand, it is known [1, 10] that $H \geqq 2/3$, with equality only for triangles.

This completes the proof of Remark 3.

Referring to Remark 1*, it is easily seen that the inequalities $1 \leqq d \leqq 3/2$, continue to hold for an arbitrary set of points instead of a point-lattice. But the question whether now in $d \leqq 3/2$ equality can be attained, remains open.

We introduce a notion similar to the notion of a saturated point-lattice. We say that a point-lattice *blocks* a convex domain B (against translations) if there is a translate of B which does not contain any lattice point, but it is impossible to remove B by a set of translations arbitrarily far from this position so as to avoid, during the translations, all lattice points.

Remark 4. If d is the density of the thinnest lattice of points blocking a closed convex domain of unit area, then $1/2 \leqq d \leqq \pi/4$. Equality holds only for quadrangles and ellipses, respectively.

If L is a point-lattice blocking the closed convex domain B of unit area, then there is a translate of B containing two points T and U of L. Again, there is a translate of B containing two points V and W of L such that the segment VW is not parallel to TU. Since B contains a translate of each of the segments TU and VW, it also contains translates of these segments which have a point in common. Thus B contains a quadrangle (which can degenerate in a triangle) whose diagonals are translates of TU and VW. The lattice generated by the vectors $\overrightarrow{TU}$ and $\overrightarrow{VW}$ blocks B. On the other hand, any lattice generated by the diagonal vectors of a quadrangle contained in B is either saturated with respect to translates of B or blocks B. It has a basic parallelogram of twice as big area as the quadrangle. Thus the density of a thinnest lattice blocking B is equal to $1/2q$, where q is the area of the quadrangle of maximal area inscribed in B. Since $q \leqq 1$ and, by the theorem of SAS, $q \geqq 2/\pi$, we have the inequalities indicated in Remark 4.

To finish, let us mention some further problems.

Find the thinnest set of unit segments saturated with respect to segments of length s. For values of s less than a certain constant between $\sqrt{2}$ and 2 it seems to be convenient to arrange the segments along lines decomposing the plane into squares with diagonal of length s. For $s=2n$ with an integer n the thinnest set of segments is conjectured to form the edges of a tessellation with regular hexagons of edge-length n.

We say that a set of translates of a convex domain D is saturated, if it is saturated with respect to translates of D. If a set of translates of a centro-symmetric convex domain D is saturated, then concentric translates of $2D$ cover the plane. Since the area-density of the inflated domains is at least 1, the area-density of the original set is at least 1/4. What is the corresponding lower bound for general convex domains? Fig. 3 shows a saturated set of translates of a triangle with area-density 1/6.

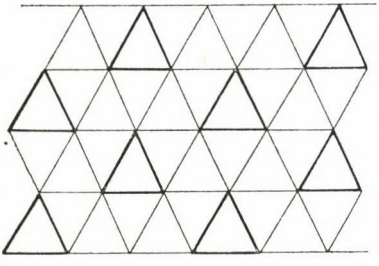


Figure 3.

The higher dimensional analogues of the problems considered in this paper are generally difficult. Nevertheless there are some simpler questions which are worth while to consider in higher space. For example: Find the thinnest point-lattice blocking a ball.

*

I thank Professor H. S. M. COXETER for assistance from his National Research Counsil of Canada grant and E. MAKAI jr. for several valuable comments.

REFERENCES

[1] FEJES TÓTH, L.: *Lagerungen in der Ebene, auf der Kugel und im Raum.* Zweite Auflage, Springer Verlag, Berlin—Heidelberg—New York, 1972.
[2] FEJES TÓTH, L.: *Packings and coverings in the plane,* Proc. Coll. Convexity (Copenhagen, 1965), Københavns Univ. Mat. Inst., København 1967, 78—87.
[3] EGGLESTON, H. G.: A minimal density plane covering problem, *Mathematika* **12** (1965), 226—234.
[4] BAMBAH, R. P. and WOODS, A. C.: On minimal density of maximal packings of the plane by convex bodies, *Acta Math. Acad. Sci. Hungar.* **19** (1968), 103—116.
[5] BAMBAH, R. P. and WOODS, A. C.: On minimal density of plane coverings by circles, *Acta Math. Acad. Sci. Hungar.* **19** (1968), 337—343.
[6] DUMIR, CH. and KHASSA, D. S.: Saturated systems of symmetric convex domains; results of Eggleston, Bambah and Woods, *Proc. Cambridge Phil. Soc.* **74** (1973), 107—116.
[7] DUMIR, CH. and KHASSA, D. S.: A conjecture of Fejes Tóth on saturated systems of circles, *Proc. Cambridge Phil. Soc.* **74** (1973), 453—460.
[8] FEJES TÓTH, L.: Some packing and covering theorems, *Acta Sci. Math. Szeged* **12**/A (1950), 62—67.
[9] BAMBAH, R. P. and ROGERS, C. A.: Covering the plane with convex sets, *J. London Math. Soc.* **27** (1952), 304—314.
[10] DOWKER, C. H.: On minimum circumscribed polygons, *Bull. Amer. Math. Soc.* **50** (1944), 120—122.
[11] SAS, E.: Über eine Extremumeigenschaft der Ellipsen, *Compositio Math.* **6** (1939), 468—470.
[12] FÁRY, I.: Sur la densité des réseaux de domaines convexes, *Bull. Soc. Math. France* **78** (1950), 152—161.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received August 27, 1974)

Remark during the galley-proof. If in n-space a set of translates of a convex body B of unit volume is saturated (with respect to translates of B) then, by Remark 1*, the respective translates of the vector body $-B+B$ cover the space. Thus the density of the vector bodies is $\geqq 1$. Recalling the theorem of C. A. Rogers and G. C. Shepard [The difference body of a convex body, *Arch. Math.* **8** (1957) 220-233] according to which the volume of $-B+B$ is $\leqq \binom{2n}{n}$ with equality only if B is a simplex, we see that the density of the original set is $\geqq 1/\binom{2n}{n}$. Equality can be attained only if the vector body of a simplex is a space-filler. This is the case for $n=2$. This answers the question raised in connection with Fig. 3.

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 81—83.

UNIFORM CONVERGENCE OF FOURIER SERIES IN HARMONIC ANALYSIS

by

R. KAUFMAN

A closed set E in $[-\pi, \pi]$ is an M_0-*set* if E carries a measure $\mu \neq 0$ whose Fourier–Stieltjes coefficients $\hat{\mu}(n)$ vanish at $n = \pm\infty$. It is known that an M_0-set is never a *Helson set:* Some continuous functions on E are not restrictions of absolutely convergent Fourier series, [3,5]. Nevertheless, certain M_0-sets possess an interpolation property comparable to the Helson property.

THEOREM 1. *There is an M_0-set F with this property: every continuous function on F is the restriction of a function on $[-\pi, \pi]$ whose Fourier series converges uniformly.*

THEOREM 2. *Every M_0-set E contains an M_0-set F.*

Theorem 1 is virtually a consequence of known results, contained in [1], [2], [4, p. 68]; an appendix to this note contains a proof of some inequalities cited from [1] and [2]. Theorem 2 seems to require a more elaborate (although technically simpler) procedure.

1. To prove theorem 1 we choose any g in $C^3[-\pi, \pi]$ with $g'' > 0$ everywhere. Then $(-\pi, \pi)$ contains a closed M_0-set F such that the functions e^{ing} $(n = 1, 2, 3, \ldots)$ are uniformly dense in the set of continuous functions on F to the complex unit circle [4, p. 68]. From the theory of Helson sets [5, p. 113], [3, p. 95], it follows that each continuous function h on F can be represented as a series $h = \sum a_n e^{ing}$ with $\sum |a_n|$ arbitrarily close to $\|h\|_\infty$.

Now F is closed and π, $-\pi$ belong to its complement, so there is a function $\varphi = 1$ on F, $\varphi = 0$ on a set $[\pi - \delta, \pi] \cup [-\pi, \delta - \pi]$ (for some $\delta > 0$). According to a theorem of HALÁSZ [2] (see also the Appendix) the function $\sum a_n e^{ing} \varphi$ has a uniformly convergent Fourier series, and of course represents h on F. (We suppose φ is C^1.)

2. In proving Theorem 2 we again use a function φ, $0 \leqq \varphi \leqq 1$, vanishing near π and $-\pi$, but allow the set $\varphi = 1$ to vary in the construction. Concerning g we specify that $1 < g' < 2$ on $[-\pi, \pi]$ and $g'' = 0$ on a neighborhood of the closed support of φ. Then, by a partition of the interval $(-\pi, \pi)$, we see that the partial sums of the Fourier series of the periodic functions $e^{iug} \varphi$ $(u > 0)$ are bounded by a constant independent of u. In fact these functions remain bounded in the (Wiener) A-norm: $\|h\|_A = \sum |b_n|$ when $h \sim \sum b_n e^{inx}$. It is also clear that $e^{ing} \varphi \to 0$ weakly in $L^1(-\pi, \pi)$.

The observations about $e^{ing}\varphi$ enable us to choose sequences $1<u_1<u_2<\ldots$ $\ldots<u_k<\ldots$ and $1<v_1<v_2<\ldots<v_k<\ldots$ so that

$$|S_m(e^{iu_k g}\varphi)| < 1+|\varphi| \leqq 2 \quad \text{for} \quad m \geqq v_k,$$

$$|S_m(e^{iu_k g}\varphi)| < 1 \quad \text{for} \quad m \leqq v_{k-1}, k \geqq 2.$$

We can also assume that $u_{k+1}>ku_k$.

From these inequalities we obtain $\left|\sum_{k=M}^{M+N} S_m(e^{iu_k g}\varphi)\right| < 2(N+1)+C(g,\varphi)$ for any numbers $M \geqq 1$, $N \geqq 1$, $m \geqq 0$.

3. Theorem 2 is proved by an inductive procedure employed in [4]; here we describe the typical step, but refer to [4] for the method of assembling the steps. Let λ be a probability measure whose closed support is contained in $(-\pi, \pi)$ and $\hat{\lambda}=0$ at $\pm\infty$. Let q be a real-valued function, continuous on $(-\pi, \pi)$ and η a positive number. We aim to construct a measure $\lambda_1 \geqq 0$ and a function G so that

a) λ_1 is absolutely continuous with respect to λ and $|\hat{\lambda}_1-\hat{\lambda}|<\eta$.

b) $|S_m(G)|<3$ for each $m \geqq 0$.

c) $|e^{iq}-G| \leqq \eta$ on the closed support of λ_1.

Following [4, p. 66] we choose a smooth function Γ of period 2π and mean-value 1, so that $\Gamma \geqq 0$ and $\Gamma(x)=0$ whenever $\eta \leqq |x| \leqq \pi$. Our candidate for λ_1 will be $\prod \cdot \lambda$, where $\prod = \prod_M^{M+N} \Gamma(u_k g-q)$, for an appropriate function g and corresponding numbers M, N. The associated function G will then be $(N+1)^{-1} \sum e^{iu_n g}\varphi$. As $\varphi=1$ on the closed support of λ, we obtain c) at once, and also the absolute continuity of λ_1.

In constructing the function g we can assume, by elementary measure theory, that the support S is totally disconnected, and cover S with disjoint closed intervals I_r, each of measure $\lambda(I_r)<\eta/2$. Then g is formed so that $g''=0$ on a neighborhood of each I_r, while the numbers $c_r=g'(I_r)$ belong to (1, 2) and are rationally independent. Next φ is chosen so that $\varphi=1$ on S and $\varphi=0$ outside UI_r, whence $\varphi=0$ near $-\pi, \pi$. This allows us to choose the sequence (u_k) and the number N so that $|S_m(G)|<3$ for all $m \geqq 0$ and $M \geqq 1$.

All that remains is estimation of the difference $(\prod \cdot \lambda\hat{}\,)-\lambda\hat{}\,$, and here we follow [4]. From the formula $\Gamma(x)=1+\sum' b_p e^{ipx}$, with $|b_p| \leqq 1$ and $\sum |b_p|<+\infty$ we obtain an absolutely convergent sum for $\prod \Gamma(u_k g-q)-1$ in which a typical term is $\prod_1^{N+1} b(p_j) \exp iu_{j+M-1} p_j g \exp -ip_j q$.

The term corresponding to $p_1=\ldots=p_{N+1}=0$ is absent, and if B is sufficiently large the sum outside the region R defined by inequalities $1 \leqq |p_1|+\ldots+|p_{N+1}| \leqq B$ is as small as we please.

On I_r we have $g'=c_r$ and this determines the derivative of $(p_1 u_M+\ldots$ $\ldots+p_{N+1}u_{M+N})g$. We consider all the derivatives obtained for different values of c_r and vectors $p_1, \ldots, p_{N+1}$ in the rectangle R. We claim that these numbers become dispersed as $M \to +\infty$, in the sense that the minimal distance between any two of them tends to $+\infty$. In fact, let us call l the leading index if $p_l \neq 0$, but $p_j=0$ for $j \geqq l+1$.

Then $(p_1u_M+\ldots+p_{N+1}u_{M+N})c_r \sim p_1u_{l+M-1}c_r$, so that two numbers in our set are far apart unless their leading indices are equal. But the numbers c_r are rationally independent, so the two numbers in question arise from the same interval I_r. From this it follows by a similar argument that the corresponding $(N+1)$-rows must be identical.

The estimation of $(\prod \cdot \lambda)\hat{} - \lambda\hat{}$ now proceeds as in [4, p. 67]. The inductive step is now complete and hence Theorem 2 is proved.

Appendix. The lemma stated below implies the theorem of ALPÁR [1], and also suffices for the proof of Theorem 1.

LEMMA. *Let $r>0$ and $f\in C^n[-r, r]$ for some $n\geqq 2$, and $\alpha\leqq f^{(n)}\leqq B\alpha$ everywhere on $[-r, r]$. Then* $\left|\int_{-r}^{r}(e^{if(x)}-1)x^{-1}dx\right| < A_n(B)$, *a constant depending only on n and B.*

PROOF. For any interval $I\subseteq(-r, r)$, we have $\left|\int_I e^{if}\right| \leqq C_n\alpha^{-1/n}$ by a variant of van der Corput's inequality proved in almost the same way as for $n=2$ [6, p. 197]. This allows us to assume $0<r<\leqq\alpha^{-1/n}$, for in the contrary case the two exterior integrals are controlled by partial integration.

Let M_k be the maximum of $|f^{(k)}|$ over $(-r, r)$, $2\leqq k\leqq n$, and observe that $|f(x)-f(0)-xf'(0)|\leqq\frac{1}{2}x^2M_2$. Thus, replacing $f(0)+xf'(0)$ on $(-r, r)$ entails an error r^2M_2 in the integral; consequently we can suppose $M_2>4^nBr^{-2}>4^nB\alpha^{2/n}$. This disposes of the case $n=2$ and allows us to use induction on n. Now $M_2>4^nB\alpha^{2/n}$ implies that for some $k=2, \ldots, n-1$, we have $M_{k+1}\leqq r^{-1}\alpha^{1/n}M_k$. Then $f^{(k)}$ varies at most $2\alpha^{-1/n}M_{k+1}\leqq M_k/2$, so that $M_k\geqq|f^{(k)}|\geqq M_k/2$. Thus the integral is bounded by $A_k(2)$ and the lemma is proved.

REFERENCES

[1] ALPÁR, L.: Sur une classe particulière de séries de Fourier ayant de sommes partielles bornées, *Studia Sci. Math. Hungarica* **1** (1966), 189—204.

[2] HALÁSZ, G.: On a theorem of L. Alpár concerning Fourier series of powers of certain functions, *Ibidem* **2** (1967), 67—72.

[3] KAHANE, J.-P.: *Séries de Fourier Absolument Convergentes*, Ergebnisse Math. Bd. 50 Springer-Verlag, Berlin, 1970.

[4] KAUFMAN, R.: Topics on Kronecker sets, *Annales Inst. Fourier* (Grenoble), **23** (1973), 65—74.

[5] RUDIN, W.: *Fourier Analysis on Groups*, Interscience, New York, 1960.

[6] ZYGMUND, A.: *Trigonometric Series*, I. Cambridge, 1959 and 1968.

University of Illinois, Urbana

(Received October 7, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 85—86.

GENERALIZED SPECTRAL OPERATORS AND NORMAL CONES

by

D. LUTZ

X always denotes a complex Banach space, $B(X)$ the Banach algebra of bounded linear operators on X, $\sigma(T)$ the spectrum of $T\in B(X)$. Let $\mathscr{A}$ be an admissible algebra of $\mathbf{C}$-valued functions on a set $\Omega\subset\mathbf{C}$ (for definitions see COLOJOARA and FOIAS [1] and DUNFORD—SCHWARTZ [2]), U an $\mathscr{A}$-spectral function, and

$$K=\{f\in\mathscr{A}, f(x)\geqq 0 \quad \text{for} \quad x\in\sigma(U_\lambda)\}.$$

Then

$$\mathscr{K}=\{U_f, f\in K\}$$

is a convex cone in $B(X)$.

If $T\in B(X)$ is a scalar operator with spectral measure E, the cone

$$\mathscr{K}=\{U_f, f\in C(\sigma(T)), f(x)\geqq 0 \quad \text{on} \quad \sigma(T)\},$$

generated by the $C(\sigma(T))$-spectral function

$$U_f=\int_{\sigma(T)} f(\lambda)\,dE(\lambda),$$

is normal in the strong operator topology on $B(X)$, i.e. there exists a norm $|||\cdot|||$ on X equivalent to the initial norm with

$$|||U_f x|||\leqq |||U_g x||| \quad \text{for all} \quad x\in X,$$

if $f, g\in C(\sigma(T))$, $0\leqq f\leqq g$ on $\sigma(T)$ (WALSH [4]).

In this case $|||U_f|||\leqq |||U_g|||$, i.e. $\mathscr{K}$ is also normal in the norm topology on $B(X)$. Hence $\mathscr{K}$ is proper.

In the present note we give a converse of the latter result.

THEOREM 1. *Let $\mathscr{A}$ be an admissible algebra of continuous functions on $\Omega\subset\mathbf{C}$ with* $\operatorname{Re}\mathscr{A}=\{\operatorname{Re} f, f\in\mathscr{A}\}\subset\mathscr{A}$. *Let $T\in B(X)$ be $\mathscr{A}$-scalar with continuous $\mathscr{A}$-spectral function U. Suppose $\mathscr{A}':=\{f|\sigma(T), f\in\mathscr{A}\}$ to be dense in $C(\sigma(T))$ with respect to the sup-norm topology on $C(\sigma(T))$. If the cone $\mathscr{K}=\{U_f, f\in\mathscr{A}, f\geqq 0$ on $\sigma(T)\}$ is normal in the norm topology on $B(X)$, then the following statements are valid:*

a) *T is $C(\sigma(T))$-scalar. There exists a continuous $C(\sigma(T))$-spectral function $\hat{U}$ of T with $\hat{U}|\mathscr{A}=U$.*

b) *If X is weakly sequentially complete, T is scalar.*

c) *$\hat{\mathscr{K}}=\{\hat{U}_f, f\in C(\sigma(T)), f(x)\geqq 0$ on $\sigma(T)\}$ is also normal in the norm topology on $B(X)$.*

PROOF. a), c) Let $|||\cdot|||$ be a norm on $B(X)$ equivalent to the initial norm satisfying $|||I|||=1$ and such that $f, g\in\mathscr{A}$, $0\leqq f\leqq g$ on $\sigma(T)$ imply $|||U_f|||\leqq |||U_g|||$.

For $f\in\mathscr{A}$, $|f|\leqq 1$ on $\sigma(T)$ one has

$$f_1 := \operatorname{Re} f\in\mathscr{A}, \quad f_2 := \operatorname{Im} f\in\mathscr{A}$$

and

$$0\leqq f_i(x)+1\leqq 2 \quad \text{on} \quad \sigma(T); \quad i=1,2.$$

So

$$\big|\,|||U_{f_i}|||-|||I|||\,\big| \leqq |||U_{f_i}+I||| \leqq 2,$$

$$|||U_{f_i}||| \leqq 3, \quad i=1,2,$$

and

$$|||U_f||| \leqq 6.$$

For arbitrary $f\in\mathscr{A}$ it follows that

$$|||U_f||| \leqq 6\cdot \max_{x\in\sigma(T)} |f(x)|,$$

i.e.

$$\|U_f\| = k\cdot \max_{x\in\sigma(T)} |f(x)|$$

with some $k\geqq 0$ for the initial norm $\|\cdot\|$.

Since $\mathscr{A}$ is dense in $C(\sigma(T))$, the existence of the continuous extension $\hat{U}$ on $C(\sigma(T))$ follows immediately; likewise it is clear that $\mathscr{U}$ is normal.

b) According to KANTOROVITZ [3], $T\in B(X)$ is scalar if and only if T is $C(\sigma(T))$-scalar with continuous $C(\sigma(T))$-spectral function.

COROLLARY 2. Modify the assumptions of Theorem 1 by requiring that $T\in B(X)$ be $\mathscr{A}$-spectral. Then T is $C(\sigma(T))$-spectral (and spectral, if X is weakly sequentially complete).

PROOF. Let $T=U_\lambda+N$, N quasinilpotent and commuting with U. Then, by continuity, N also commutes with $\hat{U}$.

We write

$$C^n(\Omega) = \{f|\Omega, f\in C^n(\mathbf{R}^2)\}, n\in\mathbf{N}\cup\{0\}\cup\{\infty\}, \Omega\subset\mathbf{C}.$$

COROLLARY 3. Let $T\in B(X)$. If for some $\Omega\subset\mathbf{C}$ and $n\in\mathbf{N}\cup\{\infty\}$ T is $C^n(\Omega)$-spectral, and if there exists a continuous $C^n(\Omega)$-spectral function U of T which generates a normal cone in the norm topology of $B(X)$, then T is $C(\sigma(T))$-spectral.

COROLLARY 4. Let X be weakly sequentially complete. $T\in B(X)$ is spectral if and only if T is generalized spectral, and if a spectral distribution U of T generates a normal cone in the norm topology on $B(X)$.

REFERENCES

[1] COLOJOARA, I.—FOIAS, C.: *Theory of generalized spectral operators*, New York, 1968.
[2] DUNFORD, N.—SCHWARTZ, J. T.: *Linear operators*, Part III. *Spectral operators*, New York, 1971.
[3] KANTOROVITZ, S.: Classification of operators by means of their operational calculus, *Trans. Amer. Math. Soc.* **115** (1965), 194—224.
[4] WALSH, B. J.: Structure of spectral measures on locally convex spaces, *Trans. Amer. Math. Soc.* **120** (1965), 295—326.

Fachbereich Mathematik, Universität Konstanz
D-775 Konstanz, Postfach 7733

(Received October 10, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 87—91.

CHARACTERIZATIONS OF UPPER RADICAL CLASSES OF SIMPLE RINGS

by

M. SADIQ ZIA*

1. Introduction. Generalizing the notion of almost nilpotency introduced by VAN LEEUWEN and HEYMAN [5], we define the notion of almost **P**-rings and dually that of nearby **Q**-rings. These notions are appropriate for characterizing upper radical classes of any class of simple rings. Particularly we obtain characterizations of the Brown—McCoy radical class (Corollary 1 and Theorems 2, 3) and of Jenkins' radical class (Corollary 2).

Throughout this paper a ring will always mean an associative ring. **A** and **O** will always denote the class of all rings and that of rings of one element, respectively. A class **R** of rings is called a *radical class*, if **R** satisfies the following requirement:

(R) $A \in \mathbf{R}$ iff every non-zero homomorphic image of A has a non-zero **R**-ideal.

Every ring A has a biggest **R**-ideal, the so called **R**-radical $\mathbf{R}(A)$ of A. A class $\mathbf{S} = \mathscr{S}\mathbf{R}$ of rings is said to be a *semisimple class* if $\mathbf{S} = \{A \in \mathbf{A} | \mathbf{R}(A) = 0\}$. For any homomorphically closed class **P** the *lower radical class* $\mathscr{L}\mathbf{P}$ is the smallest radical class containing the class **P**. The *upper radical class* $\mathscr{U}\mathbf{Q}$ of a hereditary class **Q** is the biggest radical class excluding the non-zero **Q**-rings. The semisimple class $\mathscr{S}\mathscr{U}\mathbf{Q}$ will sometimes be denoted by $\overline{\mathbf{Q}}$. As it is well known, the *Brown—McCoy radical class* is the upper radical class of the class of all simple rings with unity. We recall that *Jenkins' radical class* is the upper radical class of all non-zero simple prime rings (cf. JENKINS [2], VAN LEEUWEN [4] and VAN LEEUWEN—JENKINS [6]). An upper radical class $\mathscr{U}\mathbf{Q}$ is said to have the *intersection property*, if each $\mathscr{S}\mathscr{U}\mathbf{Q}$-ring (i.e. having 0 $\mathscr{U}\mathbf{Q}$-radical) is a subdirect sum of **Q**-rings (cf. LEAVITT [3]). For other details of radical theory we refer to DIVINSKY [1] and WIEGANDT [10].

Let **E** denote the class of all simple rings and consider a radical class **R**. The radical class **R** gives a partition (**L**; **M**) of the class **E** by $\mathbf{L} = \mathbf{R} \cap \mathbf{E}$ and $\mathbf{M} = \mathscr{S}\mathbf{R} \cap \mathbf{E}$ (the ring 0 is allowed to belong to both subclasses **L** and **M**). The class **L** and **M** is called the *lower class* and the *upper class* of the partition given by **R**, respectively.

Consider a class **R** and a class **P** of rings. We say that the *class* **R** *gives a proper partition of the class* **P** if $\mathbf{O} \neq \mathbf{R} \cap \mathbf{P} \neq \mathbf{P}$.

The author, with a deep sense of gratitude, thanks DR. RICHARD WIEGANDT for his guidance throughout the whole work.

2. General almost P-rings. Let **P** be any non-empty homomorphically closed class of rings.

---

* The author holds a post-graduate scholarship, granted by the Hungarian Institute of Cultural Relations.

Definition 1. A ring A is an *almost* **P**-*ring* if every proper homomorphic image of A is a **P**-ring. The class of almost **P**-rings will be denoted by $\mathscr{A}\mathbf{P}$.

Obviously $\mathbf{P} \subseteq \mathscr{A}\mathbf{P}$. For example, if **P** is the class of all nilpotent rings, then we get the definition of almost nilpotent rings (cf. VAN LEEUWEN [5] and WIEGANDT [8]) and obviously **P** is properly contained in $\mathscr{A}\mathbf{P}$.

In [5] idempotent simple rings are not considered as almost nilpotent rings, whereas, in accordance with [8], we suppose that every simple ring is in $\mathscr{A}\mathbf{P}$ for any homomorphically closed class **P**.

PROPOSITION 1. $\mathscr{A}\mathbf{P}$ *is homomorphically closed.*

PROOF. Consider a ring A in $\mathscr{A}\mathbf{P}$. Let $A/I=B$ be any homomorphic image of A. If $I=0$, then $A/I=A$ is in $\mathscr{A}\mathbf{P}$ and if $I=A$, then $B=0$ is in **P**. Whenever $I\neq A$, A/I is in **P** by definition 1.

It is easy to see that the lower radical class $\mathscr{L}\mathbf{P}$ has the following property:

Every $\mathscr{A}\mathbf{P}$*-ring is either in* $\mathscr{L}\mathbf{P}$ *or in* $\mathscr{S}\mathscr{L}\mathbf{P}$.

Let **R** be an arbitrary radical class and consider the following condition:

($A_{\mathbf{P}}$) *Every* $\mathscr{A}\mathbf{P}$*-ring is either in* **R** *or in* $\mathscr{S}\mathbf{R}$.

PROPOSITION 2. *Every radical class* $\mathbf{R} \supseteq \mathscr{L}\mathbf{P}$ *has property* ($A_{\mathbf{P}}$).

PROOF: Suppose $\mathbf{R} \supseteq \mathscr{L}\mathbf{P}$ and A is in $\mathscr{A}\mathbf{P}$. If $0 \neq \mathbf{R}(A)$, then $A \neq A/\mathbf{R}(A) \in \mathbf{P} \subseteq \mathscr{L}\mathbf{P} \subseteq \mathbf{R}$. On the other hand $A/\mathbf{R}(A) \in \mathscr{S}\mathbf{R} \subseteq \mathscr{S}\mathscr{L}\mathbf{P}$ because $\mathscr{L}\mathbf{P} \subseteq \mathbf{R}$. This implies that $A/\mathbf{R}(A)=0$ and thus $A=\mathbf{R}(A)$ is in **R**. If $\mathbf{R}(A)=0$, then A is in $\mathscr{S}\mathbf{R}$.

PROPOSITION 3. *If the radical class* **R** *gives a proper partition of the class of simple* **P**-*rings, then* **R** *does not have property* ($A_{\mathbf{P}}$).

PROOF: Take a non-zero simple ring $A \in \mathbf{R} \cap \mathbf{P}$ and a non-zero simple ring $B \in \mathscr{S}\mathbf{R} \cap \mathbf{P}$.

Consider the direct sum $C=A \oplus B$. By the isomorphism theorem,

$$(A+B)/A \cong B/A \cap B \cong B \quad \text{and} \quad (B+A)/B \cong A/B \cap A \cong A.$$

Thus $C/A \cong B$ and $C/B \cong A$ where A and B are both in **P**. Since C has no other proper non-zero homomorphic images, therefore C is in $\mathscr{A}\mathbf{P}$. But $0 \neq A \subseteq \mathbf{R}(C) \neq C$. This contradiction establishes the proposition.

Now we dualize the notion of an almost **P**-ring. Consider any non-empty hereditary class **Q** of rings.

Definition 2. A ring is a *nearby* **Q**-*ring* if every proper ideal of A is in **Q**. The class of nearby **Q**-rings will be denoted by $\mathscr{N}\mathbf{Q}$.

Obviously $\mathbf{Q} \subseteq \mathscr{N}\mathbf{Q}$. Consider the following condition:

($N_{\mathbf{Q}}$) *Every* $\mathscr{N}\mathbf{Q}$*-ring is either in* **R** *or in* $\mathscr{S}\mathbf{R}$.

PROPOSITION 4. *Every radical class* $\mathbf{R} \subseteq \mathscr{U}\mathbf{Q}$ *has property* ($N_{\mathbf{Q}}$).

PROOF: Consider a ring A in $\mathscr{N}\mathbf{Q}$. $\mathbf{R}(A)$ is an ideal of A. If $\mathbf{R}(A)=A$, then A is in **R**. If $\mathbf{R}(A) \neq A$, then, by definition 2, $\mathbf{R}(A)$ is in $\mathbf{Q} \cap \mathbf{R} \subseteq \mathbf{Q} \cap \mathscr{U}\mathbf{Q}=0$. This implies that $\mathbf{R}(A)=0$. Hence A is in $\mathscr{S}\mathbf{R}$.

PROPOSITION 5. *If the semisimple class* $\mathbf{S}$ *gives a proper partition of simple* $\mathbf{Q}$*-rings, then* $\mathscr{U}\mathbf{S}$ *does not have property* $(N_{\mathbf{Q}})$.

PROOF. Consider a non-zero simple ring $A \in \mathbf{S} \cap \mathbf{Q}$ and another non-zero simple ring $B \in \mathscr{U}\mathbf{S} \cap \mathbf{Q}$. Obviously the direct sum $c = A \oplus B$ in is $\mathscr{N}\mathbf{Q}$ and $0 \neq B \subseteq \mathscr{U}\mathbf{S}(c) \neq c$.

3. Applications. The Baer radical class as the lower radical of all zero-rings has already been characterized by almost $\mathbf{P}$-rings where $\mathbf{P}$ is the class of all nilpotent rings (cf. WIEGANDT [8]). We shall give here characterizations of upper radicals of simple rings, in particular of the Brown-McCoy radical and Jenkins' radical by virtue of almost $\mathbf{P}$-rings and nearby $\mathbf{Q}$-rings.

In some cases we can give more explicitly the class $\mathscr{A}\mathbf{P}$ and $\mathscr{N}\mathbf{Q}$, respectively.

Consider any class $\mathbf{M}$ of simple rings with unity. It is understood that $0 \in \mathbf{M}$. $\mathscr{N}\mathbf{M}$ is now a class of nearby simple rings with unity.

PROPOSITION 6. *Every* $\mathscr{N}\mathbf{M}$*-ring* A *is either a direct sum* $A = I \oplus K$ of two $\mathbf{M}$-rings *or* A *is simple.*

PROOF. Consider a ring A in $\mathscr{N}\mathbf{M}$. Suppose I is a non-zero proper ideal of A. So I is in $\mathbf{M}$. As it is well known that every ring I with unity embedded into A as an ideal is a direct summand of A, therefore $A = I \oplus K$. This implies that K is a proper ideal of A and so K is in $\mathbf{M}$. Otherwise A is simple. Similarly to Proposition 6 it is easy to check that for any class $\mathbf{L}$ of simple rings any $\mathscr{A}\mathbf{L}$-ring ($\mathscr{N}\mathbf{L}$-ring) A is either the direct sum of two $\mathbf{L}$-rings or A has exactly one ideal I with $A/I \in \mathbf{L}$ ($I \in \mathbf{L}$) or A is simple. Let $\mathbf{P}$ denote a homomorphically closed residual class (i.e. $\mathbf{P}$ is homomorphically closed and closed under taking subdirect sums.)

PROPOSITION 7. *The class* $\mathscr{A}\mathbf{P}$ *consists of* $\mathbf{P}$*-rings and of subdirectly irreducible* $\mathscr{A}\mathbf{P}$*-rings.*

PROOF. Take a ring $A \in \mathscr{A}\mathbf{P}$ and consider a subdirect decomposition $A = \text{subdirect} \sum A_\alpha$ of A into subdirectly irreducible rings A_α. Either each A_α is a proper homomorphic image of A or for an index α_0, $A \cong A_{\alpha_0}$.

In the first case, $A \in \mathscr{A}\mathbf{P}$ involves $A_\alpha \in \mathbf{P}$ and since $\mathbf{P}$ is residual we get $A \in \mathbf{P}$. In the second case A is subdirectly irreducible.

We say that a class $\mathbf{Q}$ is I-local if the sum $\mathbf{Q}(A) = \sum I_\alpha$ of all ideals $I_\alpha \in \mathbf{Q}$ of a ring A belongs to $\mathbf{Q}$. Examples for I-local classes are the radical classes. Assume $\mathbf{Q}$ is a hereditary I-local class.

PROPOSITION 8. *The class* $\mathscr{N}\mathbf{Q}$ *consists of* $\mathbf{Q}$*-rings and of local rings in* $\mathscr{N}\mathbf{Q}$. *Particularly if* $\mathbf{Q}$ *is a hereditary radical class, then* $\mathscr{N}\mathbf{Q} \setminus \mathbf{Q}$ *consists of all local rings* A *such that the unique maximal ideal of* A *is the* $\mathbf{Q}$*-radical of* A.

PROOF. Take a ring $A \in \mathscr{N}\mathbf{Q}$. Now either $A = \mathbf{Q}(A) \in \mathbf{Q}$ or $A \neq \mathbf{Q}(A)$. In the latter case, by $A \in \mathscr{N}\mathbf{Q}$ every proper ideal I of A is in $\mathbf{Q}$, and so $I \subseteq \mathbf{Q}(A)$. Hence $\mathbf{Q}(A)$ is the unique maximal ideal of A.

THEOREM 1. *Let* $\mathbf{M}$ *be any (non-empty) class of simple rings. A radical class* $\mathbf{R}$ *coincides with the upper radical* $\mathscr{U}\mathbf{M}$ *iff* $\mathbf{R}$ *is the biggest radical class such that*

(i) $\mathbf{R}$ *has property* $(N_{\mathbf{M}})$.

(ii) $\mathbf{M} \nsubseteq \mathbf{R}$.

PROOF. By Proposition 4, $\mathcal{U}\mathbf{M}$ has properties (i) and (ii). If $\mathbf{R}\subset\mathcal{U}\mathbf{M}$, then again by Proposition 4, $\mathbf{R}$ satisfies (i) and (ii). If $\mathbf{R}\not\subseteq\mathcal{U}\mathbf{M}$ and $\mathbf{R}$ satisfies conditions (i) and (ii), then there exists a ring $A\in\mathbf{R}\setminus\mathcal{U}\mathbf{M}$. Hence A can be mapped homomorphically onto a non-zero ring $B\in\mathbf{R}\cap\mathbf{M}$. On the other hand, condition (ii) implies the existence of a ring $C\in\mathbf{M}\cap\mathcal{S}\mathbf{R}$. Take the ring $D=B\oplus C$. Now $0\neq B\subseteq\mathbf{R}(D)\neq D$ and $(N_{\mathbf{M}})$ is not satisfied contradicting the hypothesis.

A straightforward application of Theorem 1 yields the following statements.

COROLLARY 1. *The Brown—McCoy radical* $\mathbf{G}$ *is the biggest radical class such that*
(i) *G has property* $(N_{\mathbf{M}_0})$
(ii) $\mathbf{M}_0\not\subseteq\mathbf{G}$ *where* $\mathbf{M}_0=\{$*all simple rings with unity*$\}$.

COROLLARY 2. *The Jenkins radical* $\mathbf{K}$ *is the biggest radical class such that*
(i) $\mathbf{K}$ *has property* $(N_{\mathbf{M}_j})$
(ii) $\mathbf{M}_j\not\subseteq\mathbf{K}$ *where* $\mathbf{M}_j=\{$*all simple prime rings*$\}$.

As mentioned in the introduction consider the class of all simple rings (allowing all rings A with $A^2=0$) and consider a partition $(\mathbf{L};\mathbf{M})$ of it into two classes $\mathbf{L}$ and $\mathbf{M}$ such that $\mathbf{L}\cap\mathbf{M}=0$ and $\mathbf{L}\cup\mathbf{M}=\mathbf{E}$.

PROPOSITION 9. *Let* $\mathbf{Q}$ *be a hereditary class of rings and denote the lower class of simple rings corresponding to* $\mathcal{U}\mathbf{Q}$ by $\mathbf{L}$. *If the upper radlcal* $\mathcal{U}\mathbf{Q}$ *is hereditary, then* $\mathbf{L}=\mathcal{U}\mathbf{Q}\cap\mathcal{N}\mathbf{Q}$.

PROOF. If suffices to show that any ring $A\in\mathcal{U}\mathbf{Q}\cap\mathcal{N}\mathbf{Q}$ is simple. If I is a proper ideal of A, then by $A\in\mathcal{N}\mathbf{Q}$, we have $I\in\mathbf{Q}$. On the other hand hereditariness of $\mathcal{U}\mathbf{Q}$ and $A\in\mathcal{U}\mathbf{Q}$ imply $I\in\mathcal{U}\mathbf{Q}$. But $\mathcal{U}\mathbf{Q}\cap\mathbf{Q}=0$. Hence $I=0$ and A is simple.

COROLLARY 3. *Using the same notations as in Proposition* 9, *suppose that the upper radical* $\mathcal{U}\mathbf{Q}$ *is hereditary. If A is an* $\mathcal{N}\mathbf{Q}$*-ring and* $A\notin\mathbf{L}$, *then A is in the semi-simple class* $\overline{\mathbf{Q}}$.

PROOF. If an $\mathcal{N}\mathbf{Q}$-ring A is not in $\mathbf{L}$, then by Proposition 9 A is not in $\mathcal{U}\mathbf{Q}$. Hence its radical $\mathcal{U}\mathbf{Q}(A)$ is not A. By $A\in\mathcal{N}\mathbf{Q}$ it follows $\mathcal{U}\mathbf{Q}(A)\in\mathbf{Q}\cap\mathcal{U}\mathbf{Q}=0$.

THEOREM 2. *Let* $\mathbf{M}$ *be a class of simple rings such that the upper radical* $\mathcal{U}\mathbf{M}$ *is hereditary. Then* $\mathcal{U}\mathbf{M}$ *coincides with the Brown-McCoy radical class* $\mathbf{G}$ *iff* $\mathcal{N}\mathbf{M}\cap\mathcal{U}\mathbf{M}$ *consists of all simple rings without unity.*

PROOF. By Proposition 9, $\mathbf{G}$ satisfies the required property; the converse statement is trivial.

PROPOSITION 10. *Let* $\mathbf{Q}$ *be a homomorphically closed hereditary class such that the upper radical* $\mathcal{U}\mathbf{Q}$ *gives the partition* $(\mathbf{L};\mathbf{M})$ *of simple rings. If the upper radical class* $\mathcal{U}\mathbf{Q}$ *has the intersection property, then* $\mathbf{M}=\mathcal{A}\mathcal{U}\mathbf{Q}\cap\overline{\mathbf{Q}}$.

PROOF. Clearly $\mathbf{M}\subseteq\mathcal{A}\mathcal{U}\mathbf{Q}\cap\overline{\mathbf{Q}}$. Suppose $0\neq A\in\mathcal{A}\mathcal{U}\mathbf{Q}\cap\overline{\mathbf{Q}}$. By the intersection property A can be given as a subdirect sum $A=\sum_{\text{subdirect}} A_\alpha$ of rings $A_\alpha\in\mathbf{Q}$. Now either $A_\alpha\in\mathcal{U}\mathbf{Q}$ for all indices α or there exists an index α_0 such that $A\cong A_{\alpha_0}$. In the first case we have $A_\alpha\in\mathcal{U}\mathbf{Q}\cap\mathbf{Q}=0$ for all α and consequently $A=0$. In the second case $A\in\mathcal{A}\mathcal{U}\mathbf{Q}\cap\mathbf{Q}$. Since $\mathbf{Q}$ is homomorphically closed and so for any non-zero ideal I of A we have $A/I\in\mathcal{U}\mathbf{Q}\cap\mathbf{Q}=0$. Hence A is simple and so $A\in\mathbf{M}$.

THEOREM 3. *Let* $(\mathbf{L}; \mathbf{M})$ *be a partition of simple rings such that the upper radical* $\mathscr{U}\mathbf{M}$ *has the intersection property.* $\mathscr{U}\mathbf{M}$ *coincides with the Brown—McCoy radical class* $\mathbf{G}$ *iff* $\mathscr{A}\mathscr{U}\mathbf{M} \cap \overline{\mathbf{M}}$ *consists of all simple rings with unity.*

PROOF. Apply Proposition 10 in the case $\mathbf{Q}=\mathbf{M}$ where $\mathbf{M}$ is the class of all simple rings with unity. The converse statement is trivial.

As it is well-known, each homomorphically closed semi-simple class $\mathbf{S}$ is determined by a strongly hereditary set $\mathbf{F}$ of finitely many finite fields and each $\mathbf{S}$-ring is a subdirect sum of $\mathbf{F}$-rings (cf. STEWART [7] and WIEGANDT [9], [10]). It is easy to see that $\mathbf{S}$ is a homomorphically closed semi-simple class iff $\mathbf{S}$ gives a partition $(\mathbf{L}; \mathbf{M})$ of simple rings such that $\mathbf{M}$ is a strongly hereditary set of finitely many finite fields. Moreover,

$$(*) \qquad \mathscr{A}\mathbf{L} \cap \mathbf{S} = \mathbf{M} \quad \text{and} \quad \mathscr{N}\mathbf{M} \cap \mathscr{U}\mathbf{S} = \mathbf{L}$$

hold. Observe that conditions $(*)$ are not characteristic to homomorphically closed semi-simple classes, e.g. the Brown—McCoy semisimple class is not homomorphically closed, but it satisfies conditions $(*)$.

REFERENCES

[1] DIVINSKY, N.: *Rings and radicals*, George Allen & Unwin Ltd., 1965.
[2] JENKINS, T. L.: A maximal ideal radical class, *J. Nat. Sci. Math. Lahore*, **7** (1967), 191—195.
[3] LEAVITT, W. G., The intersection property of the upper radical, *Archiv Math.*, **24** (1973), 486—492.
[4] VAN LEEUWEN, L. C. A.: On a generalization of Jenkins radical, *Archiv. Math.*, **22** (1971), 155—160.
[5] VAN LEEUWEN, L. C. A., HEYMAN, G. A. P.: A radical determined by a class of almost nilpotent rings, *Acta Math. Acad. Sci. Hung.* (to appear)
[6] VAN LEEUWEN, L. C. A., JENKINS, T. L.: Upper radicals and simple rings, *Periodica Math. Hung.* (to appear)
[7] STEWART, P. N.: Semi-simple radical classes, *Pacific J. Math.* Vol. **32**. No. 1. (1970), 249—254.
[8] WIEGANDT, R.: Characterizations of the Baer radical class by almost nilpotent rings, *Publ. Math. Debrecen* (to appear).
[9] WIEGANDT, Radical-Semisimple classes, *Periodica Math. Hung.* Vol. **3** (3—4), (1973), 243—245.
[10] WIEGANDT, R : *Radical and Semisimple classes of rings,* Queen's papers in pure and applied Mathematics No. 37, Kingston, Ontario, Canada, 1974.

Mathematical Institute of the Hungarian Academy of Sciences
1053 Budapest, Reáltanoda u. 13—15.

Department of Mathematics, Government Science College,
Lyallpur, Pakistan

(Received October 12, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 93—98.

A STATISTICAL PROPERTY OF THE WALSH FUNCTIONS

by

SHIGERU TAKAHASHI

§ 1. Introduction. It is known that the trigonometric orthogonal system has subsequences which share important properties of independent random variables (cf. [4][5]). In this note we study subsequences of the Walsh orthogonal system $\{w_n(x)\}$ from this point of view. Let

$$\varphi_k(x) = \operatorname{sign}(\sin 2^k \pi x), \quad 0 \leqq x \leqq 1, \tag{1.1}$$

the k-th Rademacher function and for an integer $n = \sum \varepsilon_k 2^k$ ($\varepsilon_k = 0$ or 1) let

$$w_n(x) = \prod \varphi_{k+1}^{\varepsilon_k}(x), \quad 0 \leqq x \leqq 1. \tag{1.2}$$

In [1] A. FÖLDES has proved the following

THEOREM A. *Let $\{n_k\}$ be a sequence of positive integers such that for every $k \geqq 1$*

$$n_{k+1}/n_k > 1 + k^{-\alpha}, \quad 0 < \alpha < 1/2.$$

Then we have

$$\overline{\lim_N} (2N \log\log N)^{-1/2} \sum_1^N w_{n_k}(x) = 1 \quad \text{a.e.}.$$

The purpose of the present note is to prove the following

THEOREM B. *Let $\{n_k\}$ be a sequence of positive integers such that for every $k \geqq 1$*

$$n_{k+1}/n_k > 1 + ck^{-\alpha}, \quad c > 0 \quad \text{and} \quad 0 < \alpha \leqq 1/2, \tag{1.3}$$

and $\{a_k\}$ be a sequence of real numbers satisfying the conditions:

$$A_N = \left(\sum_1^N a_k^2 \right)^{1/2} \to +\infty \quad \text{and} \quad a_N = O\left(A_N / N^{\alpha} (\log A_N)^{\frac{1+\varepsilon}{2}}\right) \tag{1.4}$$

as $N \to +\infty$, where ε is a positive constant. Then we have

$$\overline{\lim_N} (2A_N^2 \log\log A_N)^{-1/2} \sum_1^N a_k w_{n_k}(x) = 1 \quad \text{a.e.}.$$

For the proof we use the martingale properties of subsequences of partial sums of the Walsh series by which the asymptotic properties of the series are obtained slightly easily.* In fact under the condition (1.3), (1.4) is less restrictive than the

* More precisely using a theorem of martingale [3] we can approximate $\sum a_k w_{n_k}$ with Brownian motion in the same way.

conditions of log log law for trigonometric series (c f. [5, II]). Further as the theorem of log log law for martingales we use the following

THEOREM C. (W. F. SOUT [2]). *Let $\{\mathcal{J}_n\}$ be an increasing sequence of sub-σ fields of events and $\{\xi_n\}$ be a sequence of random variables such that:*

(i) *ξ_n is $\mathcal{J}_n$-measurable and $\mathsf{E}(\xi_n|\mathcal{J}_{n-1})=0$ a.e. $n\geqq 1$;*

(ii) $V_N=\sum_{n=1}^{N}\mathsf{E}(\xi_n^2|\mathcal{J}_{n-1})\to+\infty$ *as* $N\to+\infty$ *a.e.;*

(iii) $|\xi_n|\leqq\varepsilon_n(V_n/\log\log V_n)^{1/2}$ *a.e. where ε_n is $\mathcal{J}_{n-1}$-measurable random variable such that $\varepsilon_n\to 0$ a.e. as $n\to+\infty$.*

Then we have

$$\overline{\lim_N}\,(2V_N\log\log V_N)^{-1/2}\sum_1^N\xi_k=1 \quad a.e..$$

§ 2. Some Lemmas. In the following we consider the Walsh system as a sequence of random variables on a probability space $(\Omega, \mathcal{J}, \mathsf{P})$, where $\Omega=[0,1]$, $\mathcal{J}$ is the σ-field of all Borel-measurable sets on Ω and P is the Lebesgue measure on $\mathcal{J}$. Further we assume that $\{n_k\}$ and $\{a_k\}$ satisfy the conditions of Theorem B. Let us put

$$(2.1)\qquad \begin{cases} p(0)=0, \quad p(k)=\max\{m;\, n_m<2^k\}, \\ \Delta_k(x)=\sum_{m=p(k)+1}^{p(k+1)} a_m w_{n_m}(x) \quad \text{and} \quad B_k=A_{p(k+1)}. \end{cases}$$

If $p(k)+1<p(k+1)$, then (1.3) implies that

$$2>n_{p(k+1)}/n_{p(k)+1}>\sum_{m=p(k)+1}^{p(k+1)-1}(1+cm^{-\alpha})>1+c\{p(k+1)-p(k)-1\}p^{-\alpha}(k+1).$$

Therefore, we have

$$(2.2)\qquad p(k+1)-p(k)=O\big(p^{\alpha}(k)\big) \quad \text{as} \quad k\to+\infty.$$

Hence we have, by (1.4),

$$(2.3)\qquad \|\Delta_k\|_\infty<\sum_{m=p(k)+1}^{p(k+1)}|a_m|<\max_{p(k)<m\leqq p(k+1)}|a_m|\{p(k+1)-p(k)\}=$$

$$=O\big(B_k/(\log B_k)^{(1+\varepsilon)/2}\big) \quad \text{as} \quad k\to+\infty.$$

LEMMA 1. *For any positive integers k and j we define $\varphi(k,j)$ as an integer satisfying $w_k(x)w_j(x)=w_{\varphi(k,j)}(x)$. Then if $2^m\leqq j<k<2^{m+1}$, we have*

$$k-j\leqq\varphi(k,j)<2^m.$$

PROOF. The proof easily follows from (1.1) and (1.2).

LEMMA 2. *For any given integers k, j, q and h such that*

$$p(j)+1 < h \leqq p(j+1) < p(k)+1 < q \leqq p(k+1)$$

the number of pairs (n_r, n_i) satisfying the following relations

$$\begin{cases} \mathsf{E}(w_{n_q} w_{n_r} w_{n_h} w_{n_i}) = 1, \\ p(j) < i < h \quad \text{and} \quad p(k) < r < q, \end{cases} \tag{2.4}$$

is at most $C2^{j-k}p^{\alpha}(k)$ where C is a constant independent of k, j, q and h.

PROOF. Let m denote the smallest index r of pairs (n_r, n_i) satisfying the relation (2.4). Then we have by Lemma 1

$$n_q - n_m \leqq \varphi(n_q, n_m) = \varphi(n_h, n_i) < 2^j.$$

Thus we have

$$n_m > n_q - 2^j > n_q(1-2^{j-k}) \geqq n_q(1+2^{j-k+1})^{-1},$$

and by (1.3)

$$1+2^{j-k+1} > n_q/n_m > \prod_{j=m}^{q-1}(1+cj^{-\alpha}) > 1+c(q-m)p^{-\alpha}(k+1).$$

Since $p(k+1)/p(k) \to 1$ as $k \to +\infty$, $q-m < C2^{j-k}p^{\alpha}(k)$. Further for any given q, r and h, there exists at most one n_i satisfying (2.4). Therefore, we have

$$\{\text{the number of solutions } (n_r, n_i)\} \leqq q-m < C2^{j+k}p^{\alpha}(k).$$

LEMMA 3. *We have*

$$\mathsf{E}\left(\left|B_N^{-2}\sum_{k=1}^{N} \Delta_k^2 - 1\right|^2\right) = O\big((\log B_N)^{-1-\varepsilon}\big) \quad \text{as} \quad N \to +\infty.$$

PROOF. If we put $U_k = \Delta_k^2 - \mathsf{E}(\Delta_k^2)$, then we have

$$\begin{cases} U_k = 2\sum_{q=p(k)+2}^{p(k+1)} a_q \sum_{r=p(k)+1}^{q-1} a_r w_{n_q} w_{n_r}, \\ \mathsf{E}\left(\left|\sum_{1}^{N} \Delta_k^2 - B_N^2\right|^2\right) = \sum_{1}^{N} \mathsf{E}(U_k^2) + 2\sum_{k=1}^{N}\sum_{j=1}^{k-1} \mathsf{E}(U_k U_j). \end{cases} \tag{2.5}$$

By (2.5) and (2.3), it is seen by Minkowski inequality that

$$\mathsf{E}(U_k^2)^{1/2} \leqq 2\sum_{q=p(k)+2}^{p(k+1)} |a_q| \mathsf{E}\left(\left|\sum_{r=p(k)+1}^{q-1} a_r w_{n_r}\right|^2\right)^{1/2} =$$

$$= O\left(\mathsf{E}(\Delta_k^2)^{1/2} \sum_{q=p(k)+2}^{p(k+1)} |a_q|\right) = O\big(\mathsf{E}(\Delta_k^2)^{1/2} B_k/(\log B_k)^{(1+\varepsilon)/2}\big) \quad \text{as} \quad k \to +\infty.$$

Therefore we obtain

$$\sum_{1}^{N} \mathsf{E}(U_k^2) = O\big(B_N^4 (\log B_N)^{-(1+\varepsilon)}\big) \quad \text{as} \quad N \to +\infty.$$

For $k>j$ we have, by (2.5) and Lemma 2,

$$|\mathsf{E}(U_kU_j)| \leqq 8C2^{j-k}p^{\alpha}(k)\left(b_k\sum_{p(k)+1}^{p(k+1)}|a_q|\right)\left(b_j\sum_{p(j)+1}^{p(j+1)}|a_h|\right),$$

where $b_k=\max\{|a_j|, p(k)<j\leqq p(k+1)\}$. By (1.3) and (2.2) we have

$$b_k\sum_{p(k)+1}^{p(k+1)}|a_q| = O\big(B_k(\log B_k)^{-(1+\varepsilon)/2}p^{-\alpha}(k)\,\mathsf{E}(\Delta_k^2)^{1/2}\{p(k+1)-p(k)\}^{1/2}\big) =$$

$$= O\big(B_k(\log B_k)^{-(1+\varepsilon)/2}p^{-\alpha/2}(k)\,\mathsf{E}(\Delta_k^2)^{1/2}\big) \quad \text{as} \quad k\to+\infty,$$

and we obtain

$$\mathsf{E}(U_kU_j) = O\big(B_k^2(\log B_k)^{-1-\varepsilon}2^{j-k}p^{\alpha/2}(k)\,\mathsf{E}(\Delta_k^2)^{1/2}p^{-\alpha/2}(j)\,\mathsf{E}(\Delta_j^2)^{1/2}\big).$$

Therefore, for the proof of the Lemma it is enough to show that

$$\sum_{k=2}^{N}\sum_{j=1}^{k-1}2^{j-k}p^{\alpha/2}(k)\,\mathsf{E}(\Delta_k^2)^{1/2}p_{(j)}^{-\alpha/2}\,\mathsf{E}(\Delta_j^2)^{1/2} = O(B_N^2) \quad N\to+\infty.$$

Since $p(j+1)/p(j)\to 1$ as $j\to+\infty$, we have for every k

$$\sum_{j=1}^{k-1}2^{j-k}p^{-\alpha}(j)\leqq C_1^2p^{-\alpha}(k) \quad \text{for some} \quad C_1.$$

Thus we have by Cauchy inequality

$$\sum_{k-2}^{N}\sum_{j=1}^{k-1}2^{j-k}p^{\alpha/2}(k)\,\mathsf{E}(\Delta_k^2)^{1/2}p^{-\alpha/2}(j)\,\mathsf{E}(\Delta_j^2)^{1/2}\leqq$$

$$\leqq C_1\sum_{k=2}^{N}\mathsf{E}(\Delta_k^2)^{1/2}\left\{\sum_{j=1}^{k-1}2^{j-k}\mathsf{E}(\Delta_j^2)\right\}^{1/2}\leqq$$

$$\leqq C_1\left\{\sum_{k=2}^{N}\mathsf{E}(\Delta_k^2)\right\}^{1/2}\left\{\sum_{k=2}^{N}\sum_{j=1}^{k-1}2^{j-k}\mathsf{E}(\Delta_j^2)\right\}^{1/2}\leqq$$

$$\leqq C_1\sum_{k=1}^{N}\mathsf{E}(\Delta_k^2) = O(B_N^2) \quad \text{as} \quad N\to+\infty.$$

LEMMA 4. *We have* $\lim_{N\to\infty}B_N^{-2}\sum_{1}^{N}\Delta_k^2(x) = 1$ *a.e.*.

PROOF. Let us define a sequence $\{N(k)\}$ as follows:

$$N(k) = \min\{m; B_m^2\geqq e^{k^{\beta}}\},$$

where β is a constant such that

$$1>\beta>0 \quad \text{and} \quad (1+\varepsilon)\beta>1. \tag{2.6}$$

Since $B_m^2 - B_{m-1}^2 = \mathsf{E}(\Delta_m^2) = O\big(B_m^2/(\log B_m)^{1+\varepsilon}\big)$ as $m \to +\infty$, there exists an integer k_0 such that $k \geqq k_0$ implies that

$$e^{(k-1)^\beta} \leqq B_{N(k-1)}^2 < e^{k^\beta} \leqq B_{N(k)}^2. \tag{2.7}$$

On the other hand by Lemma 3 and (2.6) we have

$$\sum_{k=1}^{\infty} \mathsf{E}\left(\left| B_{N(k)}^{-2} \sum_{1}^{N(k)} \Delta_m^2 - 1 \right|^2\right) = O\left(\sum_{1}^{\infty} k^{-(1+\varepsilon)\beta}\right) = O(1),$$

and we obtain

$$\lim_{k \to \infty} B_{N(k)}^{-2} \sum_{1}^{N(k)} \Delta_m^2(x) = 1 \quad \text{a.e.}.$$

Since (2.7) implies that $B_{N(k)}^2 / B_{N(k-1)}^2 \to 1$ as $k \to +\infty$, $B_{N(k)}^{-2} \sum_{N(k-1)}^{N(k)} \Delta_m^2(x) \to 0$ a.e. as $k \to +\infty$ and we can complete the proof.

§ 3. Proof of Theorem B. For $n \geqq 0$ let us put

$$\xi_n(x) = \begin{cases} \Delta_n(x) + 2^{-n} w_{2^n}(x) & \text{if} \quad \mathsf{E}(\Delta_n w_{2^n}) = 0, \\ \Delta_n(x) & \text{otherwise.} \end{cases} \tag{3.1}$$

Then the σ-field $\mathscr{J}_n$ generated by $\{w_{2^k}(x),\ 0 \leqq k \leqq n\}$ is also the σ-field generated by $\{\xi_k(x),\ 0 \leqq k \leqq n\}$ and we apply Theorem C to the sequence $\{\xi_n, \mathscr{J}_n\}$, $n \geqq 1$. From (1.1) and (1.2) it is seen that

$$\mathsf{E}(\xi_n | \mathscr{J}_{n-1}) = 0 \quad \text{and} \quad \mathsf{E}(\xi_n^2 | \mathscr{J}_{n-1}) = \xi_n^2 \quad \text{a.e.}. \tag{3.2}$$

Further by Lemma 4 and (3.1) we have

$$B_N^{-2} \sum_{n=1}^{N} \xi_n^2(x) = B_N^{-2} \sum_{n=1}^{N} \Delta_n^2(x) + B_N^{-2} \sum_{n=1}^{N} \{\xi_n^2(x) - \Delta_n^2(x)\} = \tag{3.3}$$

$$= 1 + O\left(B_N^{-2} \left\{ \sum_{n=1}^{N} \Delta_n^2(x) \right\}^{1/2}\right) = 1 + o(1) \quad \text{a.e.} \quad \text{as} \quad N \to +\infty.$$

Putting $\varepsilon_N(x) = |\xi_N(x)| \left\{ \log\log \sum_{1}^{N} \xi_n^2(x) / \sum_{1}^{N} \xi_n^2(x) \right\}^{1/2}$,* then $\varepsilon_N(x)$ is $\mathscr{J}_{N-1}$-measurable

and by (3.3) and (2.3) $\varepsilon_N(x) \to 0$ a.e. as $N \to +\infty$. Therefore, the conditions of Theorem C are satisfied and we have

$$\overline{\lim_N} \left(2 \sum_{1}^{N} \xi_n^2 \log\log \sum_{1}^{N} \xi_n^2 \right)^{-1/2} \sum_{1}^{N} \Delta_m = 1 \quad \text{a.e.}.$$

By (3.3), (2.3) and the above relation we can prove the theorem.

---

* Without loss of generality we may assume that $\xi_1^2(x) \geqq 3$ a.e., so we can define $\varepsilon_N(x)$ a.e.

REFERENCES

[1] FÖLDES, A.: Further statistical properties of the Walsh functions, *Studia Sci. Math. Hung.* **7** (1972), 147—153

[2] STOUT, W. F.: A martingale analogue of Kolmogorov's law of the iterated logarithm, *Z. Wahr. Verw. Gebiete* **15** (1970), 279—290.

[3] STRASSEN, V.: Almost sure behavior of sums of independent random variables and martingales, Proc. 5th Berkeley Sympos. Math. Statist. Prob. (1965) vol II (Part I), 315—343.

[4] TAKAHASHI, S.: On lacunary trigonometric series, *Proc. Japan Acad.* **41** (1965), 503—506.

[5] —: On the law of the iterated logarithm for lacunary trigonometric series I and II, *Tôhoku Math. Journ.* **24** (1972), 319—329 (to appear).

Department of Mathematics, Kanazawa University, Kanazawa Japan

(Received October 29, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 99—101.

EINE ANWENDUNG EINES SATZES VON GASCHÜTZ AUF POLYNOMPERMUTATIONEN ÜBER ENDLICHEN MULTIOPERATORGRUPPEN

von

G. EIGENTHALER

In der folgenden Arbeit wird ein Satz über das Verhalten von gewissen Gruppen von Polynompermutationen über einer endlichen Multioperatorgruppe unter gewissen Kompositionsepimorphismen bewiesen. Dieser Satz stellt eine Verallgemetinerung von zwei Ergebnissen aus [2] dar, wo er für die Gruppe aller Polynompermutationen über einer endlichen Gruppe bzw. über einem endlichen Faktorring eines Dedekindschen Integritätsbereichs bewiesen wird ([2], ch. 5, § 3.3 bzw. ch. 4, § 5.21). Bezüglich der verwendeten Begriffe sei ebenfalls auf [2] verwiesen.

§ 1. Sei A eine universale Algebra im Sinne von [2] und $F_k(A)$ die volle k-stellige Funktionenalgebra über A, die von sämtlichen Funktionen von A^k in A mit den punktweise definierten Operationen gebildet wird, wobei k eine natürliche Zahl ist. Insbesondere liegen in $F_k(A)$ die konstanten Funktionen, die im folgenden mit den Elementen von A identifiziert werden, und die Projektionen $\xi_1, \ldots, \xi_k$, die definiert sind durch $\xi_i(a_1, \ldots, a_k)=a_i$, $(a_1, \ldots, a_k)\in A^k$, $i=1, \ldots, k$. Für $A_1\subseteq A$ sei $P_k(A, A_1)$ die von A_1 und $\xi_1, \ldots, \xi_k$ erzeugte Unteralgebra von $F_k(A)$. Die Elemente von $P_k(A, A_1)$ sind genau jene Funktionen aus $F_k(A)$, die sich mit Hilfe der Operationssymbole als Ausdrücke (im folgenden *Worte* genannt) in Elementen von A_1 und $\xi_1, \ldots, \xi_k$ darstellen lassen, und sollen als *k-stellige Polynomfunktionen über A mit Koeffizienten aus A_1* bezeichnet werden. Insbesondere erhält man für $A_1=A$ die Algebra $P_k(A, A)=P_k(A)$ aller *k-stelligen* Polynomfunktionen über A und für $A_1=\varnothing$ die Algebra $P_k(A, \varnothing)=G_k(A)$ der k-stelligen *Grätzerschen Polynomfunktionen* über A.

Im folgenden sei für eine Halbgruppe H mit Einselement $R(H)$ die Unterhalbgruppe der regulären Elemente von H (ein Element $a\in H$ heißt regulär, wenn sowohl aus $au=av$ als aus $ua=va$ stets $u=v$ folgt $(u, v\in H)$) und $E(H)$ die Gruppe der Einheiten von H. Es gilt $E(H)\subseteq R(H)$ und im Falle, daß H endlich ist, $E(H)=R(H)$. Das direkte Produkt $(F_k(A))^k$ von k Exemplaren von $F_k(A)$ bildet mit der Komposition $\circ$ von Funktionen, die definiert ist durch

$$(\varphi_1, \ldots, \varphi_k)\circ(\psi_1, \ldots, \psi_k) = (\varphi_1(\psi_1, \ldots, \psi_k), \ldots, \varphi_k(\psi_1, \ldots, \psi_k)),$$

$$(\varphi_1, \ldots, \varphi_k), (\psi_1, \ldots, \psi_k)\in(F_k(A))^k,$$

eine Halbgruppe mit Einselement, die isomorph ist zur symmetrischen Halbgruppe von A^k, und $(P_k(A, A_1))^k$ ist Unterhalbgruppe von $(F_k(A))^k$. Sei Sym A^k die symmetrische Gruppe von A^k und $U_k(A, A_1)=\text{Sym } A^k\cap(P_k(A, A_1))^k$. $U_k(A, A_1)$ ist eine Unterhalbgruppe von $(P_k(A, A_1))^k$, deren Elemente als *k-stellige Polynompermutationen über A mit Koeffizienten aus A_1* bezeichnet werden sollen, und es gilt

$$E((P_k(A, A_1))^k) \subseteq U_k(A, A_1) \subseteq R((P_k(A, A_1))^k).$$

Falls A endlich ist, gilt hier überall Gleichheit, und $U_k(A, A_1)$ bildet somit in diesem Falle mit der Komposition $\circ$ eine endliche Gruppe.

Sei η ein Epimorphismus der Algebra A auf die Algebra B, dann läßt sich η nach [2], ch. 3, § 3.31 eindeutig fortsetzen zu einem Epimorphismus von $P_k(A)$ auf $P_k(B)$, der für $i=1, \ldots, k$ die Projektion $\xi_i \in F_k(A)$ in die Projektion $\xi_i \in F_k(B)$ überführt. Bezeichnen wir diesen Epimorphismus mit $P_k(\eta)$ und ist $w(a_1, \ldots, a_n, \xi_1, \ldots, \xi_k)$ eine Darstellung eines Elementes von $P_k(A)$ als Wort in $a_1, \ldots, a_n \in A$ und $\xi_1, \ldots, \xi_k$, dann gilt

$$P_k(\eta) w(a_1, \ldots, a_n, \xi_1, \ldots, \xi_k) = w(\eta a_1, \ldots, \eta a_n, \xi_1, \ldots, \xi_k).$$

Die Abbildung $(P_k(\eta))^k$, die definiert ist durch

$$(P_k(\eta))^k(\varphi_1, \ldots, \varphi_k) = (P_k(\eta)\varphi_1, \ldots, P_k(\eta)\varphi_k), \quad (\varphi_1, \ldots, \varphi_k) \in (P_k(A))^k,$$

ist dann ein Epimorphismus von $(P_k(A))^k$ auf $(P_k(B))^k$, und zwar auch bezüglich der Komposition $\circ$, wie man leicht nachrechnet.

Eine Algebra A mit dem Operationensystem Ω heißt *Multioperatorgruppe*, falls es in Ω eine zweistellige Operation . , eine einstellige Operation $^{-1}$ und eine nullstellige Operation 1 gibt, so daß A mit . , $^{-1}$ und 1 eine Gruppe bildet und für alle Operationen $\omega \in \Omega$ mit einer Stelligkeit >0 gilt: $\omega\, 11 \ldots 1 = 1$. Insbesondere ist jede Gruppe, jeder Ring und jeder Ring mit Einselement eine Multioperatorgruppe. Ist η ein Homomorphismus von der Multioperatorgruppe A in die Multioperatorgruppe B, dann heißt die Menge aller $a \in A$ mit $\eta a = 1 \in B$ der *Kern* von η.

Im folgenden benötigen wir einen Satz von GASCHÜTZ ([1]) über Gruppen, der sich unmittelbar auf Multioperatorgruppen übertragen läßt und dessen Beweis auch in [2] (ch. 6, § 6.91) zu finden ist:

Sei k eine natürliche Zahl, A eine Multioperatorgruppe mit einem Erzeugendensystem aus k Elementen und η ein Epimorphismus von A auf eine Multioperatorgruppe B, dessen Kern endlich ist. Dann gibt es zu jedem Erzeugendensystem $\{f_1, \ldots, f_k\}$ von B ein Erzeugendensystem $\{e_1, \ldots, e_k\}$ von A mit $\eta e_i = f_i$, $i=1, \ldots, k$, d.h. man erhält sämtliche k-elementigen Erzeugendensysteme von B, indem man auf sämtliche k-elementigen Erzeugendensysteme von A den Epimorphismus η ausübt.

§ 2. Mit Hilfe dieses Satzes von GASCHÜTZ beweisen wir nun folgenden

SATZ. *Seien A, B endliche Multioperatorgruppen, η ein Epimorphismus von A auf B, $A_1 \subseteq A$ und $B_1 = \eta A_1$. Dann gilt: $(P_k(\eta))^k\, U_k(A, A_1) = U_k(B, B_1)$.*

BEWEIS. Da $(P_k(\eta))^k$ ein Epimorphismus bezüglich $\circ$ ist, gilt

$$(P_k(\eta))^k U_k(A, A_1) = (P_k(\eta))^k E((P_k(A, A_1))^k) \subseteq E((P_k(B, B_1))^k) = U_k(B, B_1).$$

Sei umgekehrt $(\varphi_1, \ldots, \varphi_k) \in U_k(B, B_1)$, so gibt es ein $(\psi_1, \ldots, \psi_k) \in U_k(B, B_1)$, so daß

$$(\psi_1, \ldots, \psi_k) \circ (\varphi_1, \ldots, \varphi_k) = (\xi_1, \ldots, \xi_k). \tag{1}$$

Hat ψ_i eine Darstellung $v_i(b_{i1}, \ldots, b_{im_i}, \xi_1, \ldots, \xi_k)$ als Wort in $b_{ij} \in B_1$ und $\xi_1, \ldots, \xi_k$, so gilt nach (1)

$$v_i(b_{i1}, \ldots, b_{im_i}, \varphi_1, \ldots, \varphi_k) = \xi_i, \quad i = 1, \ldots, k. \tag{2}$$

Faßt man die Elemente von A_1 bzw. B_1 als nullstellige Operationen auf, so bilden $P_k(A, A_1)$ und $P_k(B, B_1)$ mit dem Operationensystem $\Omega \cup A_1$ bzw. $\Omega \cup B_1$ endliche Multioperatorgruppen mit dem Erzeugendensystem $\{\xi_1, \ldots, \xi_k\}$ und $P_k(\eta)$ ist ein Epimorphismus von $P_k(A, A_1)$ auf $P_k(B, B_1)$, und zwar auch bezüglich der um A_1 bzw. B_1 erweiterten Operationensysteme, wenn man jedes Element von B_1 so oft als nullstellige Operation auffaßt, als es als Bild eines Elementes aus A_1 bei η auftritt. Nach (2) ist auch $\{\varphi_1, \ldots, \varphi_k\}$ ein Erzeugendensystem von $P_k(B, B_1)$ bezüglich $\Omega \cup B_1$. Nach obigem Satz von GASCHÜTZ gibt es daher ein Erzeugendensystem $\{\varrho_1, \ldots, \varrho_k\}$ von $P_k(A, A_1)$ bezüglich $\Omega \cup A_1$, so daß $P_k(\eta)\varrho_i = \varphi_i$, $i = 1, \ldots, k$. Da $\{\varrho_1, \ldots, \varrho_k\}$ ein Erzeugendensystem von $P_k(A, A_1)$ bezüglich $\Omega \cup A_1$ ist, gibt es Worte $w_i(a_{i1}, \ldots, a_{in_i}, \xi_1, \ldots, \xi_k)$, $i = 1, \ldots, k$, mit Operationssymbolen aus Ω und in Elementen $a_{ij} \in A_1$ und $\xi_1, \ldots, \xi_k$, so daß

$$w_i(a_{i1}, \ldots, a_{in_i}, \varrho_1, \ldots, \varrho_k) = \xi_i, \quad i = 1, \ldots, k. \tag{3}$$

Sei σ_i die durch das Wort $w_i(a_{i1}, \ldots, a_{in_i}, \xi_1, \ldots, \xi_k)$ dargestellte Funktion aus $P_k(A, A_1)$, so gilt nach (3)

$$(\sigma_1, \ldots, \sigma_k) \circ (\varrho_1, \ldots, \varrho_k) = (\xi_1, \ldots, \xi_k), \tag{4}$$

also $(\varrho_1, \ldots, \varrho_k) \in E\big((P_k(A, A_1))^k\big) = U_k(A, A_1)$. Wegen $\big(P_k(\eta)\big)^k(\varrho_1, \ldots, \varrho_k) = (\varphi_1, \ldots, \varphi_k)$ ist also $\big(P_k(\eta)\big)^k U_k(A, A_1) \supseteq U_k(B, B_1)$, womit der Satz bewiesen ist.

Für den Fall, daß $A_1 = A$ ist und A, B endliche Gruppen bzw. endliche Faktorringe von Dedekindschen Integritätsbereichen sind, erhält man die eingangs erwähnten Spezialfälle dieses Satzes, die bereits in [2] zu finden sind. Ein Beispiel dafür, daß der Satz nicht richtig bleibt, wenn die Algebra A unendlich ist, findet man ebenfalls in [2] (ch. 4, § 5.3), wo für A der Integritätsbereich Z der ganzen Zahlen genommen wird und für B ein Restklassenring $Z/(p)$, wobei $p > 3$ eine Primzahl ist.

LITERATUR

[1] GASCHÜTZ, W.: Zu einem von B. H. und H. Neumann gestellten Problem, *Math. Nachr.* **14** (1955), 249—252.

[2] LAUSCH, H. und W. NÖBAUER: *Algebra of Polynomials.* North-Holland Publishing Comp., Amsterdam—London 1973.

Institut für Algebra und Mathematische Strukturtheorie, Technische Universität Wien, Karlsplatz 13, A-1040 Wien

(Eingegangen: 12. November, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 103—109.

MAXIMAL CIRCUITS OF GRAPHS II

by

D. R. WOODALL

If $3 \leqq d+1 \leqq n$, let d_+ denote the number of vertices with valency $\geqq d$ in a graph G on n vertices. It is proved that, if $d_+ > (d+2)(n-1)/2d-1$ (d even) or $d_+ > d(n-2)/2(d-1)$ (d odd), then G contains a circuit of length at least $d+1$; this generalizes a result of Dirac's and is best possible for infinitely many values of n. Also, if $3 \leqq c \leqq d+1$ and $d_+ > (c+2)(n-1)/2d-1$, then G contains a circuit of length at least c. Stronger results can be conjectured if G is 2-connected.

1. Introduction. All graphs considered are finite, undirected, and without loops or multiple edges. Circuits are 'elementary', i.e., have no repeated vertices. $V(G)$ denotes the set of vertices of G. If $v \in V(G)$ and $X \subseteq V(G)$, $\Gamma(v)$ denotes the set of vertices joined to v and $\Gamma(X) := \bigcup_{v \in X} \Gamma(v)$*.

In his well known paper [1] of 1952, DIRAC proved that if G is a graph on n vertices in which every vertex v has valency $\varrho(v) \geqq k$, then

(D1) if $k \geqq \frac{1}{2}n$, then G is Hamiltonian;

(D2) if $k \geqq 2$, then G contains a circuit of length at least $k+1$; and

(D3) if G is 2-connected and $k \leqq \frac{1}{2}n$, then G contains a circuit of length at least $2k$.

In [2] and [3], PÓSA obtained theorems in which he showed that the same conclusions will follow even if there are some vertices in G with valency less than k, provided that there are not too many of them. His theorem generalizing (D1) in this way is best possible, but his generalizations of (D2) and (D3) are not, and examples suggest that Pósa-type conditions are not entirely appropriate to these cases. (As mentioned in [5], the Pósa-type conjecture on page 747 of [4] can be false if $n > 3k := {}^3/_2(n-r)$, although it may well be true if $2k \leqq n \leqq 3k$.) The purpose of this note is to prove the following theorem, which implies the truth of PÓSA's generalization of (D2) when $n \geqq 2d-1$.

THEOREM 1. *If $d \geqq 2$, and G is a graph on n ($\geqq 3$) vertices in which the number of vertices with valency $\geqq d$ is more than*

$$\begin{cases} (d+2)(n-1)/2d - 1 & (d \text{ even}), \\ d(n-2)/2(d-1) & (d \text{ odd}), \end{cases}$$

then G contains a circuit of length at least $d+1$.

---

* Throughout the paper the symbol $:=$ or $=:$ indicates that the equation in which it occurs acts as the definition of (some part of) the expression on the same side of the equality sign as the colon.

These bounds are best possible for infinitely many values of n, as shown by the following examples. If d is even, the examples consist of t copies of $K_{\frac{1}{2}d}+\frac{1}{2}(d+2)K_1$, disjoint except that each (except the last) has one of its $\frac{1}{2}(d+2)$ vertices in common with one of the $\frac{1}{2}(d+2)$ vertices of the next. If d is odd, we define the examples $O(d, t)$ recursively: Let $O(d, 1)$ be a graph on $2d$ vertices consisting of d vertices of valency d forming a K_d, each of which is joined to one of d vertices of valency 1; $O(d, t)$ is then anything that can be obtained from $O(d, 1)$ and any graph of the form $O(d, t-1)$ by identifying a terminal vertex of $O(d, t-1)$, and its adjacent vertex of valency d, with a similar pair of vertices taken from $O(d, 1)$, but in the reverse order. (So $O(d, t)$ has $t(2d-2)+2$ vertices, all of valency d or 1.)

Before proving Theorem 1, we shall prove:

THEOREM 2. *If $3 \leqq c \leqq d+1$, and G is a graph on n vertices in which the number of vertices with valency $\geqq d$ is more than $(c+2)(n-1)/2d - 1$ (and at least one), then G contains a circuit of length at least c.*

This result is not best possible, although examples similar to the above show that it is not far off. However, the restriction $c \leqq d+1$ is clearly necessary, unless we are to impose an extra condition such as 2-connectedness, in view of the graphs tK_{d+1} and K_1+tK_d. I have not been able to obtain a sensible generalization of (D3), but conjecture:

CONJECTURE. If $d \leqq \frac{1}{2}n$, and G is a 2-connected graph on n vertices in which the number of vertices with valency $\geqq d$ is at least $d+\frac{1}{2}n$, then G contains a circuit of length at least $2d$.

The figure $d+\frac{1}{2}n$ may not be best possible here; if n is large enough, it can probably be reduced to $\frac{1}{2}(n+3)$, the extremal examples then being those cited in [5] as disproving the conjecture of [4].

2. Proofs of the theorems. We prove Theorem 2 first. Theorem 1 is proved along similar lines, but a great deal of tedious calculation is needed in order to obtain this slight improvement, and I have therefore given the details only where they differ significantly from those in Theorem 2.

PROOF of Theorem 2. Let $D_+(G)$ and $D_-(G)$ denote the sets of vertices of G with valency $\geqq d$ and $< d$ respectively, with cardinalities $d_+(G)$ and $d_-(G)$, where

$$d_+(G) > (c+2)(n-1)/2d - 1 =: f(n, c). \tag{1}$$

We prove the result by induction on n. It is (vacuously) true if $n \leqq d$; so suppose $n \geqq d+1$, when $d_+(G) > 1$.

Suppose it is not true that all the vertices in $D_+(G)$ are contained in a single block of G. Then $G = G_1 \cup G_2$, where $|V(G_1) \cap V(G_2)| \leqq 1$, $d_+(G_1) \geqq 1$ and $d_+(G_2) \geqq 1$. Let $|V(G_i)| =: n_i$ $(i = 1, 2)$, so that $n_1+n_2 = n$ or $n+1$. If $d_+(G_i) \leqq f(n_i, c)$ for each i, then

$$d_+(G) \leqq f(n_1, c)+f(n_2, c)+1 \leqq f(n, c),$$

contrary to hypothesis. Thus either G_1 or G_2 satisfies the hypotheses of the theorem, and the result follows by induction. So from now on we may suppose that all the vertices in $D_+(G)$ are contained in a single block of G. Moreover, this block contains a circuit: for otherwise it would have to be K_2, with both vertices in $D_+(G)$, since $d_+(G)>1$, whence $n\geqq 2d$ and, by (1), $d_+(G)>2$, a contradiction.

Suppose that G does not contain a circuit of length $\geqq c$. Add edges to G to form a graph G^* with the maximum possible number of edges such that G^* contains no circuit of length $\geqq c$. From now on we write D_+, d_+, D_-, d_- for $D_+(G^*)$, $|D_+(G^*)|$, etc. There are two cases to consider.

Case 1. Each two vertices of D_+ are joined by an edge of G^*. In this case let $a_1, a_2, \ldots, a_m, a_1$ be the vertices in order round a circuit C of maximum length in G^* subject to the restriction that $D_+ \subseteq C$. Reducing suffices modulo m, let

$$X := \{a_i \in D_+ : a_{i-1} \notin D_+, a_{i+1} \notin D_+\},$$

$$Y_1 := \{a_i \in D_+ : a_{i-1} \in D_+, a_{i+1} \notin D_+\},$$

$$Y_2 := \{a_i \in D_+ : a_{i-1} \notin D_+, a_{i+1} \in D_+\},$$

$$Z := \{a_i \in D_+ : a_{i-1} \in D_+, a_{i+1} \in D_+\},$$

$$x := |X|, \quad y := |Y_1| = |Y_2|, \quad z := |Z|.$$

Note that

$$x+2y+z = d_+ \tag{2}$$

and

$$x+y \leqq m-d_+, \tag{3}$$

so that

$$y+z \geqq 2d_+ - m \tag{4}$$

and (subtracting (3) from (4))

$$z \geqq 3d_+ - 2m + x. \tag{5}$$

Moreover, since we are supposing that there is no circuit in G^* of length $\geqq c$, we have

$$d_+ \leqq m \leqq c-1 \leqq d \leqq n-1. \tag{6}$$

Let $\mathscr{P}$ denote the set of paths P such that P starts at a vertex in Y_1 and goes round C to the next vertex in Y_2. Then C can be broken up into the paths in $\mathscr{P}$ and the vertices in Z, and these units can be reassembled in any order to give a circuit with the same length as C. It is now easy to see, using the maximality of C, that if two vertices in $Y_1 \cup Y_2 \cup Z$ are joined to the same vertex outside C, then they must be the two end vertices of a path P in $\mathscr{P}$, the penultimate vertex of which is not joined to any other vertex of $Y_1 \cup Y_2 \cup Z$. So if there are k pairs of vertices in $Y_1 \cup Y_2 \cup Z$ with this property, forming two sets $Y_1' \subseteq Y_1$ and $Y_2' \subseteq Y_2$ with $|Y_1'| = |Y_2'| = k$ and

$$k \leqq y, \tag{7}$$

then each of the vertices of $Y_1 \cup Y_2 \cup Z$ apart from these $2k$ is joined to at most $m-1-k$ vertices of C, and each of these $2k$ vertices is joined to at most $m-k$ vertices of C. Thus we find

$$n \geqq |C| + |(Y_1 \cup Y_2 \cup Z) \setminus (Y_1' \cup Y_2')|[d-(m-1-k)] + |Y_1'|[d-(m-k)],$$

whence

$$n-m+k \geqq (2y+z-k)(d-m+1+k). \tag{8}$$

(4), (7) and (8) give

$$n-m+k \geqq (2d_+ - m)(d-m+1+k), \tag{9}$$

whence

$$\begin{aligned} n-m &\geqq (2d_+ - m)(d-m+1), \quad \text{since} \quad n-m \geqq d-m+1 \geqq 1, \quad \text{by (6)}, \\ &> [(c+2)(n-1)/d - 2 - m](d-m+1), \quad \text{by (1)}, \\ &\geqq [(m+3)(n-1)/d - 2 - m](d-m+1), \quad \text{since} \quad m \leqq c-1, \quad \text{by (6)}. \end{aligned}$$

Simplifying, this gives

$$(m-1)(n-1-d) > (n-1-d)(m+2)(d-m+1),$$

which is impossible since $n-1-d \geqq 0$ and $d-m+1 \geqq 1$, by (6). This completes the discussion of Case 1.

Case 2. There is at least one pair of vertices in D_+ that are not joined by an edge, Since the addition of an edge joining them creates a circuit of length at least c, they must be connected by a path of lenght at least c—1.

Let $a_1, a_2 \ldots a_m$ be the vertices in order along a longest path P in G^* subject to the restriction that a_1 and $a_m \in D_+$. Clearly $m \geqq c$. For $i = 1$ ur m, let

$$Q_i := \Gamma(a_i) \setminus V(P),$$

$$R_i := \Gamma(Q_i) \cap D_+ \cap V(P),$$

$$S_i := \Gamma(a_i) \cap V(P) \setminus R_i,$$

and, for any $X \subseteqq V(P)$, let

$$X^- := \{a_j \in V(P) : a_{j+1} \in X\}$$

and

$$X^+ := \{a_j \in V(P) : a_{j-1} \in X\}.$$

Let

$$q_i := |Q_i|, \quad r_i^- := |R_i^-|, \quad \text{etc.}$$

Since a_1 and $a_m \in D_+$, the result is obvious if Q_1 or $Q_m = \varnothing$; so suppose $Q_i \neq \varnothing$ $(i = 1, m)$. Then $a_i \in R_i$, and, since $a_i \notin \Gamma(a_i) \cap V(P)$,

$$q_i + r_i + s_i \geqq |\Gamma(a_i)| + 1 \geqq d+1. \tag{10}$$

Using the maximality of P, we see that $Q_i \subseteqq D_-$ and $\Gamma(Q_i) \cap D_+ \subseteqq V(P)$ $(i = 1, m)$. So we may suppose that

$$r_i = |\Gamma(Q_i) \cap D_+| > (c+2) q_i / 2d, \tag{11}$$

as otherwise the graph $G^* \setminus Q_i$ would satisfy the hypotheses of the theorem, and the result would follow by induction. (Notice that $G^* \setminus Q_i$ contains at least one vertex with valency $\geqq d$, namely a_{m+1-i} $(i=1, m)$.) Using the maximality of P again, we see that $R_1^- \cap D_+ = R_m^+ \cap D_+ = \varnothing$, so that

$$R_1^- \cap R_1 = R_m^+ \cap R_m = \varnothing.$$

Since $m \geqq c$, there is no circuit in G^* with at least as many vertices as P. Using this, and standard arguments for forming circuits out of paths, it is easy to see now that the only possible non-empty intersections among the sets

$$R_1, R_1^-, S_1, R_m^+, R_m^{++}, S_m^+$$

are

$$S_1 \cap R_1^- \quad \text{and} \quad S_m^+ \cap R_m^{++}.$$

Note that

$$r_1^- = r_1 - 1, \quad r_m^{++} = r_m^+ = r_m - 1 \quad \text{and} \quad s_m^+ = s_m,$$

and put

$$\delta_1 := |S_1 \setminus R_1^-|, \quad \delta_m := |S_m^+ \setminus R_m^{++}|. \tag{12}$$

Let

$$X := R_1 \cup R_1^- \cup S_1 \cup R_m^+ \cup R_m^{++} \cup S_m^+.$$

Then, by the above disjointness assertions,

$$|X| = 2r_1 + 2r_m - 3 + \delta_1 + \delta_m. \tag{13}$$

But (10), (11) and (12) give respectively

$$r_i \geqq d + 1 - s_i - q_i, \tag{14}$$

$$2dr_i/(c+2) > q_i, \tag{15}$$

and

$$r_i \geqq 1 + s_i - \delta_i, \tag{16}$$

so that, adding,

$$2r_i(c+d+2)/(c+2) > d + 2 - \delta_i. \tag{17}$$

Thus (13) gives

$$\begin{aligned} |X| &= 2(r_1 + r_m) + (\delta_1 + \delta_m)(c+2)/(c+d+2) - 3 + (\delta_1 + \delta_m)d/(c+d+2) \\ &> 2(d+2)(c+2)/(c+d+2) - 3 + (\delta_1 + \delta_m)d/(c+d+2), \text{ by (17)}, \\ &\geqq c \text{ if } c \leqq d \text{ or } \delta_1 + \delta_m \geqq 1. \text{ But if } c = d+1 \text{ and } \delta_i = 0, \text{ (17) gives} \end{aligned}$$

$$2r_i > (d+2)(d+3)/(2d+3) > \tfrac{1}{2}(d+3),$$

whence $r_i \geqq \frac{1}{4}(d+4)$ $(i=1, m)$. If r_1 or $r_m > \frac{1}{4}(d+4)$, then $|X| > c = d+1$, by (13). A study of (14), (15) and (16) now shows that $|X| \geqq c+1$ except when

$$\begin{cases} d \equiv 0 \pmod 4, \quad \delta_1 = \delta_m = 0, \\ q_1 = q_m = \frac{1}{2}d, \quad r_1 = r_m = \frac{1}{4}d + 1, \quad s_1 = s_m = \frac{1}{4}d. \end{cases} \tag{18}$$

We now prove the existence of a circuit in G^* of length at least $|X|-1$, coming back to the exceptional case (18) at the end. Since all the vertices if D_+ are contained in a single block of G^*, say B, the path P lies in this block. We construct a sequence of paths $P_1, P_2, \ldots$ iteratively. Let P_1 be a path outside P (except for its end vertices) connecting $a_{i_1} := a_1$ to a_{j_1}, where P_1 is chosen so that j_1 is as large as possible. Suppose now that $P_1, \ldots, P_s$ have been constructed, where P_s connects a_{i_s} to a_{j_s}. If $j_s = m$, stop; otherwise let P_{s+1} be a path outside P (except for its end vertices) connecting $a_{i_{s+1}}$ to $a_{j_{s+1}}$, where j_{s+1} is as large as possible subject to

$$i_{s+1} < j_s. \tag{19}$$

Note that $i_{s+1} \geqq j_{s-1}$ (if $s \geqq 2$), $j_{s+1} > j_s$ by the 2-connectedness of B, and P_{s+1} is disjoint from all of $P_1, \ldots, P_s$ except for the possibility that $i_{s+1} = j_{s-1}$. Having finished the construction with P_k, having $j_k = m$, *change* P_1 *and* P_k *if necessary* so that j_1 is as small as possible and i_k as large as possible without violating (19).

If $i \leqq j$, let us denote by $a_i P a_j$ and $a_j \overline{P} a_i$ the paths forwards along P from a_i to a_j and backwards from a_j to a_i, and similarly with P_s for P. Then if k is odd we have the circuit

$$a_{i_1} P_1 a_{j_1} P a_{i_3} P_3 a_{j_3} P \ldots P_k a_{j_k} \overline{P} a_{j_{k-1}} \overline{\overline{P_{k-1}}} a_{i_{k-1}} \overline{P} \ldots \overline{P_2} a_{i_2} \overline{P} a_{i_1};$$

and if k is even the circuit

$$a_{i_1} P_1 a_{j_1} P a_{i_3} P_3 a_{j_3} P \ldots P_{k-1} a_{j_{k-1}} P a_{j_k} \overline{P_k} a_{i_k} \overline{P} \ldots \overline{P_2} a_{i_2} \overline{P} a_{i_1}.$$

The only vertices of X that could be omitted from either of these circuits are a_{j_1-1}, a_{i_k+1} and a_{i_k+2}, which could be in R_1^-, S_m^+ or R_m^+, and R_m^{++}, respectively. However, if $a_{j_1-1} \in R_1^-$, then we can choose P_1 to have length at least two, and similarly with P_m if $a_{i_k+2} \in R_m^{++}$; so we can certainly find a circuit of length at least $|X|-1$.

This completes the proof except in the case (18), when $|X| = c = d+1$. In this case $S_1 \subseteq R_1^-$, since $\delta_1 = 0$, and so, since $a_2 \in S_1$, $a_3 \in R_1$. This means that there is a vertex a_2' in Q_1 that we can interchange with a_2. The effect of this is, for all practical purposes, to increase q_1 by one, and to give us an augmented set X' with $|X'| \geqq d+2$. For define

$$Q_1' := [\Gamma(a_1) \setminus V(P)] \cup \{a_2\},$$

and use this set to define (recursively) R_1', S_1', δ_1' and X' in a way exactly analogous to the way in which we defined R_1, S_1, δ_1 and X using Q_1. We obtain relations (13′)—(17′) involving $|X'|$, q_1', r_1', s_1' and δ_1' (in an obvious terminology) exactly analogous to the relations (13)—(17). But now $q_1' = \frac{1}{2}d+1$ and so, by (15′),

$$r_1' > (d+3)(d+2)/4d > \tfrac{1}{4}d+1.$$

Since $r_m = \frac{1}{4}d+1$, $|X'| \geqq d+2$, by (13). The longest circuit formed by the construction described above, possibly with a_2, in place of a_2, must have length at least $|X'|-1$, $\geqq d+1$ as required. This completes the proof of Theorem 2. ∎

PROOF of Theorem 1. The proof follows the same lines as the proof of Theorem 2. I indicate here the extra difficulties.

If d is even, we prove the result stated. If d is odd, we prove the stronger result that, if $d_+(G) > d(n-1)/2(d-1) - 1$, then either G contains a circuit of length at least $d+1$ or G is an $O(d, t)$ for some t. Since in the latter case $d_+(G) = d(n-2)/2(d-1)$, this will prove Theorem 1. (I cannot prove Theorem 1 without proving this stronger result.)

This causes difficulties in the second paragraph of the proof, because it is no longer enough for *one* of G_1 and G_2 to satisfy the hypotheses of the theorem. However, calculation shows that, if G_1 is an $O(d, t_1)$ and G_2 does not satisfy the hypotheses of the theorem, then the graph obtained by deleting all but one edge of G_1 from G does satisfy the hypotheses; and if this graph is an $O(d, t_2)$, then G is an $O(d, t_1 + t_2)$.

It still suffices to prove the result for G^*, since if G^* is an $O(d, t)$ then G^* minus an edge does not satisfy the hypotheses of the theorem, to that we must have $G = G^*$.

In Case 1 we choose C so that X is as small as possible (subject to C having maximum length and $D_+ \subseteqq C$). We then notice that, if $X \neq \varnothing$, each vertex of Z is joined to at most $m-3$ vertices of C. It is enough to prove that $n - m \leqq 1$ or (if d is odd) $n - m \leqq 2$, since, in the cases $n = d+1$ and (if d is odd) $n = d+2$ it is not difficult to prove the result directly. The details of the argument are repetitive and tedious, but the only difficulty arises in the case d odd, $k = 0$ and $X = \varnothing$. In this case the hypothesis of the theorem ensures that $d_+ \geqq \frac{1}{2}(n-1)$, and (2) and (8) yield $d - m + 1 = 1$ and $n \geqq 2d-1$, whence the hypothesis gives $d_+ \geqq d$. Thus $d_+ = d$, $m = d$, $x = y = 0$, and G^* consists of K_d with pendent edges added at each vertex. There is clearly no such graph if $n = 2d-1$. If $n = 2d$ we obtain $O(d, 1)$. If $n \geqq 2d+1$, the hypothesis gives $d_+ \geqq d+1$ and G^* contains the required circuit.

In Case 2, making heavy use of the integrity if the variables, we obtain the required result $|X| \geqq d+2$ except in case (18) and the following case:

$$d \equiv 1 \pmod 4, \quad \delta_1 + \delta_m \leqq 1, \quad r_1 = r_m = \tfrac{1}{4}(d+3),$$

$$q_i = \tfrac{1}{2}(d+1) \quad \text{and} \quad s_i = \tfrac{1}{4}(d-1) \quad \text{if} \quad \delta_i = 0 \quad (i = 1, m).$$

Relabelling P in the reverse direction if necessary, we may suppose that $\delta_1 = 0$, whence the result follows as in the proof of Theorem 2. This completes the proof of Theorem 1. ∎

REFERENCES

[1] DIRAC, G. A.: Some theorems on abstract graphs, *Proc. London Math. Soc.* (3) **2** (1952), 69—81.

[2] PÓSA, L.: A theorem concerning Hamilton lines, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **7** (1962), 225—226.

[3] PÓSA, L.: On the circuits of finite graphs, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **8** (1963), 355—361.

[4] WOODALL, D. R.: Sufficient conditions for circuits in graphs, Proc. London Math. Soc. (3) **24** (1972), 739—755.

[5] WOODALL, D. R.: Maximal circuits of graphs I. *Acta Math. Acad. Sci. Hungar.* (to appear).

Department of Mathematics, University of Nottingham
(Received December 27, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 111—115.

ON HADWIGER NUMBERS AND NEWTON NUMBERS OF A CONVEX BODY

by

L. FEJES TÓTH

§ 1. A set of open convex bodies is said to form a packing if each point of the space belongs to at most one body. The bodies having a boundary point in common with a body B are said to be neighbours or first neighbours of B. The neighbours of the first neighbours of B different from B and from its first neighbours are said to be second neighbours of B, and so on. Sometimes we will call the jth neighbours also neighbours of order j. The kth Hadwiger number H_k of B is defined [1, 12] as the maximum of the total number of its neighbours of order $\leqq k$ extended over all packings of translates of B. The kth Newton number N_k of B is defined [2, 12] in a similar way by replacing the word "translates" by "congruent replicas". Obviously, $N_k \geqq H_k$. The present paper contains some contributions to the various results [3, ..., 16] known about Hadwiger and Newton numbers.

§ 2. It is known [7, 8, 9] that in n-space we have, for any convex body,

$$H_k \leqq (2k+1)^n - 1$$

with equality only for parallelotopes. What is the exact lower bound of H_k? We shall see that this problem is almost hopeless, even for $n=2$.

In the plane we have the nice pair of inequalities

$$3k(k+1) \leqq H_k \leqq 4k(k+1).$$

The inequality on the left-hand side follows from the fact that in a lattice-packing of translates of a convex plate in which each plate has six neighbours, each plate has $6j$ neighbours of order j and thus all together $6 \cdot 1 + 6 \cdot 2 + \ldots + 6 \cdot k = 3k(k+1)$ neighbours up to the order k. The upper bound is attained for all k's if and only if the plate is a parallelogram. How accurate is the lower bound? We shall show that there is a constant $c > 3$ such that for any convex plate $\varliminf_{k \to \infty} H_k/k^2 \geqq c$. This strongly supports the conjecture that the lower bound is only for some small values of k exact.

It seems to be an interesting question what numbers can occur as the Hadwiger number H_k of a convex plate for small values of k. It is not difficult to show that H_1 is either 6 or 8. It is known [12] that for a circle $H_2 = 18$ (and it seems probable that the equality $H_2 = 18$ continues to hold for any strictly convex plate). For a centrosymmetric hexagon arising from a square by cutting off small triangles from opposite corners, we have $H_2 \geqq 20$ (Fig. 1). Thus we have at least one number between 18 and 24 which is the second Hadwiger number of a convex plate. It is not known whether there is another such number.

§ 3. Let B be a convex body, $B+v$ a translate of B and B^* the body arising from B by central symmetrisation [7]. Referring to the fact that B and $B+v$ have a point in common if and only if B^* and B^*+v do so, we see that B and B^* have the same Hadwiger numbers. Therefore we can restrict ourselves to centro-symmetric bodies. In §3 and §4 we mean by a plate always a centro-symmetric convex plate.

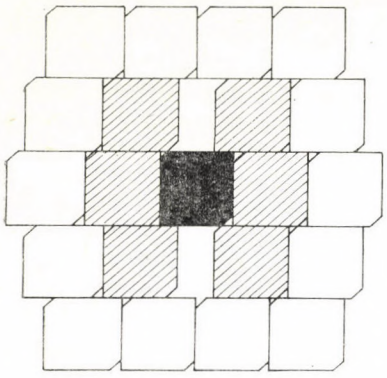


Figure 1.

Let O_1 and O_2 be the centers of two translates P_1 and P_2 of a plate P touching each other. We associate with the direction O_1O_2 the density Δ of the densest latticepacking of translates of P containing P_1 and P_2. If P_3 is a third translate of P touching both P_1 and P_2, then Δ is equal to the density of the plates P_1, P_2, P_3 in the triangle $O_1O_2O_3$, i.e. the total area of the parts of the plases lying in $O_1O_2O_3$ divided by the area of $O_1O_2O_3$.

Now we can phrase the following

THEOREM. *Let P be a centro-symmetric convex plate of area A. Let the areaelement dA of P be the area of the intersection of P with an infinitesimal angular region emanating from the center of P. Let Δ be the density associated with the direction of the bisector of this angular region. If H_k is the kth Hadwiger number of P, then*

$$(*) \qquad \varliminf_{k\to\infty} H_k/k^2 \geqq \frac{4}{A}\int^A \Delta dA.$$

The proof will be outlined in § 4. Here we make some remarks about the inequality (*).

Obviously, Δ is a continuous function of the direction. On the other hand, we have $\Delta \geqq 3/4$ for any plate and any direction. Equality holds only in the case of an affinely regular hexagon for the directions of its diagonals. Thus we have for any plate

$$\frac{1}{A}\int^A \Delta dA > \frac{3}{4}.$$

Observe that $\frac{1}{A}\int^A \Delta dA$ is a continuous, affine invariant functional of the plate. Any plate can be transformed by an affinity into a plate containing a unit circle and contained in a concentric circle of radius 4. The set of all such plates is compact. Therefore the last inequality implies the existence of a constant $c>3$ such that for any plate $\frac{1}{A}\int^A \Delta dA \geqq c/4$. Thus we have $\varliminf_{k\to\infty} H_k/k^2 \geqq c$, as claimed in § 2.

The greatest possible value of c is not known. As an orientation let us mention the interesting fact that for a regular hexagon we have

$$\frac{1}{A}\int^A \Delta dA = 6\int_0^{1/6} \frac{dx}{1+12x^2} = \sqrt{3}\arctan\frac{1}{\sqrt{3}} = \frac{\pi}{\sqrt{12}},$$

which is equal to the density of the densest lattice-packing of circles. Thus we have both for circles and for regular hexagons $\lim_{k\to\infty} H_k/k^2 \geqq 2\pi/\sqrt{3} \approx 3.62$.

Obviously, all neighbours of P of order not greater than k are contained in the domain $(2k+1)P$. Therefore the total number of these neighbours of P is less than $(2k+1)^2 d + o(k^2)$, where d is the density of the densest packing of translates of P. It is known [17, 18] that d cannot exceed the density $\bar{\Delta} = \max \Delta$ of the densest lattice-packing of translates of P. Thus we have

$$\overline{\lim_{k\to\infty}} H_k/k^2 \leqq 4\bar{\Delta}.$$

Comparing this inequality with (*) we see that for plates with direction-invariant Δ [19, 20, 21] we have

$$\lim_{k\to\infty} H_k/k^2 = 4\Delta.$$

It is very likely that for any plate

$$\lim_{k\to\infty} H_k/k^2 = \frac{4}{A}\int^{A} \Delta dA.$$

In n-space we have for any centro-symmetric convex body of volume V

$$\underline{\lim_{k\to\infty}} H_k/k^n \geqq \frac{2^n}{V}\int^{V} \Delta dV$$

with similar definition of Δ and the volume-element as in the plane. But here we cannot conclude for bodies with constant Δ that $\lim H_k/k^n = 2^n\Delta$, because for $n>2$ we do not know whether the density d of the densest packing of translates of the body is equal to Δ. For balls it is even conjectured [22] that in higher dimensions we have $d>\Delta$.

§ **4.** In order to prove the inequality (*) we must construct a packing of translates of P with a great total number of neighbours of P of order $\leqq k$. For this purpose we decompose the domain $(2k+1)P$ by half-lines emanating from its center O into a certain number of sectors. Let S be one of these sectors and let h be a half-line issuing from O and contained in S. Let $P=P_0$ be centered at O. Let $P_1, P_2, \ldots, P_k$ be the first, second, ..., kth neighbour of P with centers $O_1, O_2, \ldots, O_k$ lying on h. Let L be the densest lattice of translates of P containing P_0 and P_1. In L we consider those neighbours of P_0 of order at most k which are contained in S. We add to these plates those plates of $P_1, \ldots, P_k$ which are not contained in S. We call the plates obtained in this way sector-plates (Fig. 2).

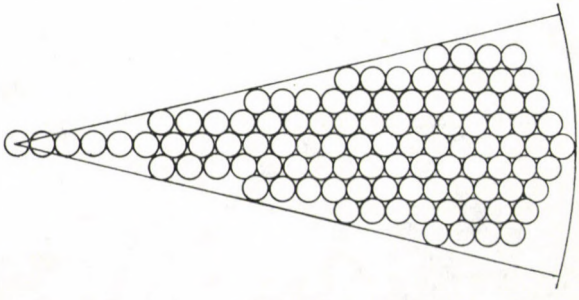


Figure 2.

If k is great, the sector-plates will leave a part of S, near to the boundary of $(2k+1)P$, empty, because among the sector-plates generally only P_k will reach the boundary of $(2k+1)P$. But if the angle of S is small then we can say that the sector-plates fill S with density Δ, where Δ is the density belonging to the direction of h. Thus the number of the sector-plates is approximately $\Delta s/A=(2k+1)^2\Delta a/A$, where s is the area of S and a is the area of the intersection of S and P. The sum of these numbers for all sectors is approxmately

$$\frac{4k^2}{A}\int^A \Delta dA.$$

Of course the sector-plates near to O belonging to different sectors will overlap. To avoid overlapping plates, we choose an integer m of order of magnitude $\sqrt{k}$ and we construct a "core" around P consisting of the first, ..., mth neighbours of P in a lattice of translates of P in which each plate has six neighborus. The centers $C_1, \ldots, C_{6m}$ of the mth neighbours of P lie, in this cyclic order, in the vertices and on the sides of an affinely regular hexagon $C_1C_{m+1}\ldots C_{5m+1}$. In the above construction of the sector-plates, let the number of sectors be $6m$, and choose the sectors S so that their "bisectors" h coincide with the halflines $OC_1, \ldots, OC_{6m}$. Translate the sector-plates belonging to the sector with the bisector OC_i through $\overrightarrow{OC_i}$, $i=1, \ldots, 6m$. The translated sector-plates along with the plates of the core will form a packing containing approximately

$$3m(m+1)+\frac{4k^2}{A}\int^A \Delta dA$$

plates, all of which are at most $(m+k)$th neighbours of P.

This implies the inequality $(*)$. A detailed proof was given, in the case of circles, in [13].

§ 5. In n-space let B be a convex body of volume V and diameter D. In any packing of congruent replicas of B the neighbours of B of order $\leqq k$ are contained in the parallel body of B of radius kD. Since the volume of this body is approximately $v_nk^nD^n$, where v_n is the volume of the unit ball, we have for the Newton number N_k of B the trivial inequality

$$\overline{\lim_{k\to\infty}} N_k/k^n \leqq v_nD^n/V.$$

The interesting thing in this inequality is that it is exact. Equality is attained for a special kind of space-fillers satisfying the condition that the space can be filled with linear chains consisting of congruent replicas of the body joined to each other along their diameters. This follows immediately from the idea of the proof of the inequality $(*)$.

Obviously, the parallelotopes satisfy the above condition. In the plane there are plates other than the parallelograms with the required property, namely the triangles and the quadrangles whose longest side is not longer than their longer diagonal. Are there bodies in more than two dimensions sharing the above property with the parallelotopes? As an answer to this question Professor H. S. M. COXETER called my attention to "certain simplexes such as affine variants of the orthoscheme $(0, 0, \ldots, 0)(1, 0, \ldots, 0)(1, 1, \ldots, 0)\ldots(1, 1, \ldots, 1)$". Another type of bodies with the required property was pointed out by E. MAKAI Jr.. Let A and B be opposite vertices

of an n-cube q. Decompose q into n equal pyramids based on the facets of q meeting at B and having A as common apex. Decomposing each cube in a cubical grid in this way, the pyramids will fit together in the desired way. The required property of the pyramids will be preserved by an affine elongation in the direction AB or by an affine contraction in the direction AB leaving the direction AB of the diameters of the pyramids unchanged.

The author thanks members of staff of the University of Calgary for assistance from their National Research Council of Canada grants.

REFERENCES

[1] FEJES TÓTH, L.: Über eine affininvariante Masszahl bei Eipolyedern, *Studia Sci. Math. Hung.* **5** (1970) 173—180.
[2] FEJES TÓTH, L.: Remarks on a theorem of R. M. Robinson, *Studia Sci. Math. Hung.* **4** (1969) 441—445.
[3] BENDER, C.: Bestimmung der grössten Anzahl gleich grosser Kugeln, welche sich auf eine Kugel von demselben Radius, wie die übrigen auflegen lassen, *Arch. Math. Phys.* **56** (1874) 302—312.
[4] GÜNTER, S.: Ein stereometrisches Problem, *Arch. Math. Phys.* **57** (1875) 209—215.
[5] SCHÜTTE, K. and VAN DER WAERDEN, B. L.: Das Problem der dreizehn Kugeln, *Math. Ann.* **125** (1953) 325—334.
[6] LEECH, J.: The problem of the thirteen spheres, *Math. Gaz.* **40** (1956) 22—23.
[7] HADWIGER, H , DEBRUNNER, H. and KLEE, V.: *Combinatorial Geometry in the Plane*, New York 1964. Theorem 43, p. 18.
[8] GRÜNBAUM, B.: On a conjecture of Hadwiger, *Pacific J. Math.* **11** (1964) 215—219.
[9] GROEMER, H.: Abschätzung für die Anzahl der konvexen Körper, die einen konvexen Körper berühren, *Monatsh. Math.* **65** (1961) 74—81.
[10] COXETER, H. S. M.: An upper bound for the number of equal non-overlapping spheres that can touch another of the same size, Proc. Symp. Pure Math. VII. Convexity. Amer. Math. Soc. 1963, 53—71.
[11] FEJES TÓTH, L.: On the number of equal discs that can touch another of the same kind, *Studia Sci. Math. Hung.* **2** (1967) 363—367.
[12] FEJES TÓTH, L. and HEPPES, A.: A variant of the problem of the thirteen spheres, *Can J. Math.* **19** (1967) 1092—1100.
[13] FEJES TÓTH, L.: Über die Nachbarschaft eines Kreises in einer Kreispackung, *Studia Sci. Math. Hung.* **4** (1969) 93—97.
[14] BÖRÖCZKY, K.: Über die Newtonsche Zahl regelmässiger Vielecke, *Periodica Math. Hung.* **1** (1971) 113—119.
[15] SCHOPP, J.: Über die Newtonsche Zahl von Bereichen konstanter Breite, *Studia Sci. Math. Hung.* **5** (1970)
[16] HORTOBÁGYI I.: Konvex lemezek Newton-számáról, *Mat. Lapok* **23** (1972) 313—317. (Über die Newtonsche Zahl konvexer Scheiben; ungarisch.)
[17] FEJES TÓTH, L.: Some packing and covering theorems, *Acta Sci. Math. (Szeged)* **12** A (1950) 62—67.
[18] ROGERS, C. A.: The closest packing of convex two-dimensional domains, *Acta Math.* **86** (1951) 309—321.
[19] FEJES TÓTH, L.: Über Scheiben mit richtungsinvarianter Packungsdichte, *Elem. Math.* **26** (1971) 58—59.
[20] MAKAI, E. Jr.: On centrosymmetric convex domains with a packing density independent of the direction, *Studia Sci. Math. Hungar.* **7** (1972) 423—424.
[21] CHAKERIAN, G. D.: On a certain affine invariant functional for convex bodies, *Studia Sci. Math. Hung.* **8** (1973) 91—93.
[22] ROGERS, C. A.: *Packing and Covering*, Cambridge 1964.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received December 2, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 117—121.

ON THE MAXIMAL DEVIATION BETWEEN TWO EMPIRICAL DISTRIBUTION FUNCTIONS

by

L. TAKÁCS

1. Introduction. Let $\xi_1, \xi_2, \ldots, \xi_m, \eta_1, \eta_2, \ldots, \eta_n$ be mutually independent random variables having a common distribution function $F(x)$. Denote by $F_m(x)$ and $G_n(x)$ the empirical distribution functions of the samples $(\xi_1, \xi_2, \ldots, \xi_m)$ and $(\eta_1, \eta_2, \ldots, \eta_n)$ respectively. The empirical distribution functions are defined to be continuous on the right. Denote by $\eta_1^*, \eta_2^*, \ldots, \eta_n^*$ the random variables $\eta_1, \eta_2, \ldots, \eta_n$ arranged in increasing order of magnitude.

In this paper we are concerned with the determination of the joint distribution of the magnitude and the rank order of the maximal deviation between the two empirical distribution functions.

The maximal deviation between $F_m(x)$ and $G_n(x)$ is defined either as

$$\delta^+(m,n) = \sup_{-\infty<x<\infty} [F_m(x) - G_n(x)] = \max_{1\leqq r\leqq n} [F_m(\eta_r^*) - G_n(\eta_r^* - 0)] \tag{1}$$

or as

$$\delta^-(m,n) = \sup_{-\infty<x<\infty} [G_n(x) - F_m(x)] = \max_{1\leqq r\leqq n} [G_n(\eta_r^*) - F_m(\eta_r^*)]. \tag{2}$$

Denote by $\varrho^+(m,n)$ the smallest $r=1,2,\ldots,n$ and by $\sigma^+(m,n)$ the largest $r=1,2,\ldots,n$ for which $F_m(\eta_r^*) - G_n(\eta_r^* - 0)$ attains its maximum. Denote by $\varrho^-(m,n)$ the smallest $r=1,2,\ldots,n$ and by $\sigma^-(m,n)$ the largest $r=1,2,\ldots,n$ for which $G_n(\eta_r^*) - F_m(\eta_r^*)$ attains its maximum. The variables $\varrho^+(m,n), \varrho^-(m,n), \sigma^+(m,n), \sigma^-(m,n)$ can be interpreted as the rank orders of the first and the last maximal deviation between the two empirical distribution functions relative to the sample $(\eta_1, \eta_2, \ldots, \eta_n)$.

Define

$$\tau^+(m,n) = [(m+n)(\varrho^+(m,n) - 1) + mn\delta^+(m,n)]/n, \tag{3}$$

$$\tau^-(m,n) = [(m+n)\varrho^-(m,n) - mn\delta^-(m,n)]/n, \tag{4}$$

$$\omega^+(m,n) = [(m+n)(\sigma^+(m,n) - 1) + mn\delta^+(m,n)]/n, \tag{5}$$

and

$$\omega^-(m,n) = [(m+n)\sigma^-(m,n) - mn\delta^-(m,n)]/n. \tag{6}$$

The random variables $\tau^+(m,n), \tau^-(m,n), \omega^+(m,n), \omega^-(m,n)$ can be interpreted as the rank orders of the first and the last maximal deviation between the two empirical distribution functions relative to the combined sample.

We can easily see that if $F(x)$ is a continuous distribution function, then $\delta^+(m, n)$, $\delta^-(m, n)$, $\varrho^+(m, n)$, $\varrho^-(m, n)$, $\sigma^+(m, n)$, $\sigma^-(m, n)$, $\tau^+(m, n)$, $\tau^-(m, n)$, $\omega^+(m, n)$, $\omega^-(m, n)$ are distribution-free statistics, and we have the following equivalence relations between the distributions of the indicated vector variables:

$$(\delta^+(m, n), \sigma^+(m, n)) \sim (\delta^-(m, n), n+1-\varrho^-(m, n))$$

and

$$(\delta^-(m, n), \sigma^-(m, n)) \sim (\delta^+(m, n), n+1-\varrho^+(m, n)).$$

These relations imply that

$$(\delta^+(m, n), \omega^+(m, n)) \sim (\delta^-(m, n), m+n-\tau^-(m, n))$$

and

$$(\delta^-(m, n), \omega^-(m, n)) \sim (\delta^+(m, n), m+n-\tau^+(m, n)).$$

The above relations show that in order to find the distributions of the indicated vector variables, it is sufficient to determine only the distributions of $(\delta^+(m, n), \varrho^+(m, n))$ and $(\delta^-(m, n), \varrho^-(m, n))$.

We note that in the particular case when $n=m$ we have also $(\delta^+(m, m), \varrho^+(m, m)) \sim (\delta^-(m, m), m+1-\varrho^-(m, m))$ and $(\delta^+(m, m), \tau^+(m, m)) \sim (\delta^-(m, m), 2m-\tau^-(m, m))$. These relations follow by symmetry.

In 1969 the author [3] determined the distribution of $(\delta^+(m, n), \varrho^+(m, n))$ for $n=mp$ where p is a positive integer. Now we shall determine the distribution of $(\delta^-(m, n), \varrho^-(m, n))$ for $n=mp$ where p is a positive integer. In the particular case when $p=1$ the distribution of $(\delta^+(m, n), \tau^+(m, n))$ was found in 1957 by I. VINCZE [4]. In 1973 G. P. STECK and G. J. SIMMONS [1] gave a general method of finding the distribution of $(\delta^+(m, n), \tau^+(m, n))$.

2. A combinatorial method. To find the distributions of the statistics introduced in this paper it is convenient to define some auxiliary variables. Let us define v_r $(r=1, 2, \ldots, n+1)$ as p times the number of variables $\xi_1, \xi_2, \ldots, \xi_m$ falling in the interval $(\eta^*_{r-1}, \eta^*_r]$ where $\eta^*_0=-\infty$, $\eta^*_{n+1}=\infty$ and p is a positive number. Write $N_r=v_1+\ldots+v_r$ for $r=1, 2, \ldots, n+1$. Clearly $N_{n+1}=mp$. By using this notation we can write that

$$F_m(\eta^*_r) = N_r/mp, \quad G_n(\eta^*_r) = r/n, \quad G_n(\eta^*_r-0) = (r-1)/n \tag{7}$$

for $r=1, 2, \ldots, n$. Thus we have

$$\delta^+(m, n) = \max_{1\leq r\leq n}\left[\frac{N_r}{mp}-\frac{r-1}{n}\right] \tag{8}$$

and

$$\delta^-(m, n) = \max_{1\leq r\leq n}\left[\frac{r}{n}-\frac{N_r}{mp}\right]. \tag{9}$$

Furthermore, $\varrho^+(m, n)$ is the smallest $r=1, 2, \ldots, n$ and $\sigma^+(m, n)$ is the largest $r=1, 2, \ldots, n$ for which $nN_r-(r-1)mp$ attains its maximum, and $\varrho^-(m, n)$ is the smallest $r=1, 2, \ldots, n$ and $\sigma^-(m, n)$ is the largest $r=1, 2, \ldots, n$ for which $rmp-nN_r$ attains its naximum.

Throughout the rest of the paper we assume that $F(x)$ is a continuous distribution function. If $F(x)$ is a continuous distribution function then $v_1, v_2, \ldots, v_{n+1}$

are interchangeable random variables whose joint distribution does not depend on $F(x)$. We have

$$\mathsf{P}\{N_i = sp\} = \frac{\binom{i+s-1}{s}\binom{m+n-i-s}{m-s}}{\binom{m+n}{m}} \tag{10}$$

for $i=1, 2, \ldots, n$ and $s=0, 1, \ldots, m$, and

$$\mathsf{P}\{N_i = sp,\ N_{i+j} - N_i = tp\} = \frac{\binom{i+s-1}{s}\binom{j+t-1}{t}\binom{m+n-i-j-s-t}{m-s-t}}{\binom{m+n}{m}} \tag{11}$$

for $1 \leqq i < i+j \leqq n$ and $0 \leqq s+t \leqq m$. Furthermore, $N_{n+1} = mp$.

Since the random variables $(v_1, v_2, \ldots, v_{n+1})$ have the same joint distribution as the variables $(v_{n+1}, v_n, \ldots, v_1)$ it follows that the equivalence relations mentioned in the Introduction are indeed true.

In what follows we shall consider the case when $n=mp$ and p is a positive integer. In this case $v_1, v_2, \ldots, v_{n+1}$ are interchangeable random variables taking on non-negative integers only. By the results of reference [2] for $1 \leqq s \leqq n+1$ we have

$$\mathsf{P}\{N_r < r \quad \text{for} \quad 1 \leqq r \leqq s \mid N_s = i\} = \begin{cases} (s-i)/s & \text{for} \quad 0 \leqq i \leqq s, \\ 0 & \text{for} \quad i \geqq s, \end{cases} \tag{12}$$

provided that $\mathsf{P}\{N_s = i\} > 0$. We can easily prove (12) by mathematical induction.

Define $\varrho(k)$ $(k=1, 2, \ldots, n+1)$ as the smallest $r=1, 2, \ldots, n+1$ for which $r - N_r = k$ if there is such an r at all. By (12) it follows immediately that

$$\mathsf{P}\{\varrho(k) = r\} = \frac{k}{r}\,\mathsf{P}\{r - N_r = k\} \tag{13}$$

for $1 \leqq k \leqq r \leqq n+1$.

3. The distribution of $(\delta^-(m, n), \varrho^-(m, n))$. By (12) and (13) we can easily determine the distribution of $(\delta^-(m, n), \varrho^-(m, n))$.

THEOREM. *If $F(x)$ is a continuous distribution function and $n=mp$ where p is a positive integer, then we have*

$$\binom{m+n}{m}\mathsf{P}\{n\delta^-(m, n) = k,\ \varrho^-(m, n) = k+sp\} = \frac{k}{k+sp}\binom{k+sp+s-1}{s}$$

$$\left[\binom{m+n-k-sp-s}{m-s} - \sum_{t=0}^{\left[\frac{n-k-1-sp}{p}\right]} \frac{1}{1+tp}\binom{tp+1}{t}\binom{m+n-k-1-sp-tp-s-t}{m-s-t}\right] \tag{14}$$

for $0 \leqq sp \leqq n-k$ and $1 \leqq k \leqq n$,

$$\binom{m+n}{m}\mathsf{P}\{n\delta^-(m, n) = 0,\ \varrho^-(m, n) = sp\} = \frac{\binom{sp+s-2}{s}\binom{m+n-sp-s}{m-s}}{(sp-1)(n+1-sp)} \tag{15}$$

for $1 < sp \leqq n$, *and*

$$\binom{m+n}{n} \mathsf{P}\{n\delta^-(m,n) = 0, \varrho^-(m,n) = 1\} = \frac{1}{2(2m-1)} \tag{16}$$

for $p = 1$.

PROOF. If $n = mp$, then by (9) we have

$$\mathsf{P}\{n\delta^-(m,n) = k, \varrho^-(m,n) = j\} = \tag{17}$$

$$= \mathsf{P}\{j - N_j = k, r - N_r < k \quad \text{for} \quad 1 \leqq r < j, r - N_r \leqq k \quad \text{for} \quad j \leqq r \leqq n\}$$

for $0 \leqq k \leqq j \leqq n$.

If $k \geqq 1$, then by (13) we can write that

$$\mathsf{P}\{n\delta^-(m,n) = k, \quad \varrho^-(m,n) = j\} =$$

$$= \mathsf{P}\{\varrho(k) = j\} - \sum_{i=1}^{n-j} \mathsf{P}\{\varrho(k) = j, \quad \varrho(k+1) - \varrho(k) = i\} = \tag{18}$$

$$= \frac{k}{j} \mathsf{P}\{N_j = j - k\} - \sum_{i=1}^{n-j} \frac{k}{ij} \mathsf{P}\{N_j = j - k \quad \text{and} \quad N_{i+j} - N_j = i - 1\}$$

and this proves (14).

If $j > 1$, then by (12) we get

(19)

$$\mathsf{P}\{n\delta^-(m,n) = 0, \varrho^-(m,n) = j\} = \sum_{r=1}^{j-1} \frac{r}{(j-1)(n+1-j)} \mathsf{P}\{N_1 = r+1, N_j = j\} =$$

$$= \frac{\mathsf{P}\{N_1 = 0, N_j = j\}}{(j-1)(n+1-j)}.$$

Here we used that $\mathsf{E}\{N_1 | N_j = j\} = 1$. This proves (15).

Finally, by (12) it follows that

$$\mathsf{P}\{n\delta^-(m,n) = 0, \varrho^-(m,n) = 1\} = \frac{\mathsf{P}\{N_1 = 1\}}{n} \tag{20}$$

which proves (16).

We note that if $n = mp$, then

$$\mathsf{P}\{\varrho^-(m,n) = j\} = \frac{1}{n+1} \tag{21}$$

for $j = 1, 2, \ldots, n$.

In the particular case when $n = m$, that is, $p = 1$ by (14), (15) and (16) we obtain that

(22)

$$\mathsf{P}\{m\delta^-(m,m) = k, \varrho^-(m,m) = j\} = \frac{k(k+1)}{(2j-k)(m-j+k+1)} \frac{\binom{2j-k}{j}\binom{2m-2j+k}{m-j}}{\binom{2m}{m}}$$

for $1 \leqq k \leqq j \leqq m$ and

$$(23) \quad \mathsf{P}\{m\delta^-(m,m) = 0, \varrho^-(m,m) = j\} = \frac{1}{(4j-2)(m-j+1)} \frac{\binom{2j}{j}\binom{2m-2j}{m-j}}{\binom{2m}{m}}$$

for $1 \leqq j \leqq m$. Formulas (22) and (23) are in agreement with the results of I. VINCZE [4].

REFERENCES

[1] STECK, G. P. and SIMMONS, G. J.: On the distributions of R^+_{mm} and (D^+_{mn}, R^+_{mn}), *Studia Sci. Math. Hung.* **8** (1973), 79—89.

[2] TAKÁCS, L.: The probability law of the busy period for two types of queuing processes, *Operations Research* **9** (1961), 402—407.

[3] TAKÁCS, L.: *Combinatorial methods in the theory of order statistics. Nonparametric Techniques in Statistical Inference*, Edited by M. L. Puri. Cambridge University Press, 1970, pp. 359—384.

[4] VINCZE, I.: Einige zweidimensionale Verteilungs- und Grenzverteilungssätze in der Theorie der geordneten Stichproben, *Publications of the Math. Inst. of the Hung. Acad. of Sciences* **2** (1957), 183—209.

Department of Mathematics and Statistics, Case Western Reserve University
Cleveland, Ohio 44106

(Received December 20, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 123—129.

EINE OPTIMALE FEHLERABSCHÄTZUNG ZUR TRIGONOMETRISCHEN INTERPOLATION

von

R. GÜNTTNER

§ 1. Einleitung und Ergebnisse

Es bezeichne $C_{2\pi}$ den Raum der reellwertigen, stetigen und 2π-periodischen Funktionen. In [1] wurde bewiesen

$$\|T_n[f]-f\| \leqq \frac{1}{2}(\|T_n\|+1)\cdot\omega\left(f, \frac{2\pi}{2n+1}\right), \quad f\in C_{2\pi} \tag{1.1}$$

wobei $T_n[f]$ das trigonometrische Interpolationspolynom von f an $2n+1$ äquidistanten Stützstellen bedeutet und $\|\cdot\|$ im Sinne der sup-Norm zu verstehen ist. Die Abschätzung (1.1) soll hier für gewisse Unterklassen von $C_{2\pi}$ weiter verschärft werden.

$H_\omega \subset C_{2\pi}$ sei die Menge aller Funktionen, deren Stetigkeitsmodul nicht größer als ein vorgegebener Stetigkeitsmodul ω ist. Speziell bezeichne $Lip_M\alpha$ die Klasse H_ω mit $\omega(\delta)=M\cdot\delta^\alpha$ $(0<\alpha\leqq 1)$.

Zum Vergleich seien noch zwei Fehlerabschätzungen aus der Approximationstheorie angegeben:

$$\|T_n[f]-f\| \leqq \frac{1}{2}(\|T_n\|+1)\cdot\omega\left(f, \frac{\pi}{n+1}\right), \quad f\in H_\omega \tag{1.2}$$

ω konkav; dies ist eine Folgerung aus den Ungleichungen

$$\|T_n[f]-f\| \leqq (\|T_n\|+1)\cdot E_n(f), \quad f\in C_{2\pi}$$

$$E_n(f) \leqq \frac{1}{2}\omega\left(f, \frac{\pi}{n+1}\right), \quad \omega \text{ konkav.}$$

Zu der letzten Abschätzung siehe etwa [4]. Speziell ergibt sich aus (1.2)

$$\|T_n[f]-f\| \leqq \frac{\pi\cdot M}{2(n+1)}(\|T_n\|+1), \quad f\in Lip_M 1. \tag{1.3}$$

Hier soll nun bewiesen werden:

SATZ 1. *Der Stetigkeitsmodul ω sei konkav. Dann ist*

$$\sup_{f\in H_\omega}\|T_n[f]-f\| = \frac{1}{2}(\|T_n\|+C_n)\cdot\omega\left(\frac{2\pi}{2n+1}\right) \tag{1.4}$$

mit

$$C_n = 2\omega\left(\frac{\pi}{2n+1}\right)/\omega\left(\frac{2\pi}{2n+1}\right)-1.$$

Ist ω streng monoton, so ist $C_n<1$ und damit (1.4) eine Verbesserung von (1.1) und (1.2). Als Spezialfall von Satz 1 sei hervorgehoben die

Folgerung 1. Für die Klasse $Lip_M\alpha$ ist $C_n=2^{1-\alpha}-1$, insbesondere gilt:

$$\sup_{f\in Lip_M 1} \|T_n[f]-f\| = \frac{\pi\cdot M}{2n+1}\cdot\|T_n\|.$$

Dies ist offensichtlich eine Verbesserung von (1.3).

Für die trigonometrische Interpolation bei gerader Knotenwahl wird ein analoger Satz bewiesen. Es bezeichne hierbei $T_n^*[f]$ das trigonometrische Interpolationspolynom an 2n äquidistanten Stützstellen (vgl. § 4):

SATZ 2. *Der Stetigkeitsmodul ω sei konkav. Dann ist*

$$\sup_{f\in H_\omega} \|T_n^*[f]-f\| = \frac{1}{2}(\|T_n^*\|+C_n^*)\cdot\omega\left(\frac{\pi}{n}\right)$$

mit

$$C_n^* = 2\omega\left(\frac{\pi}{2n}\right)/\omega\left(\frac{\pi}{n}\right)-1.$$

Folgerung 2. Für die Klasse $Lip_M\alpha$ ist $C_n^*=2^{1-\alpha}-1$, insbesondere gilt:

$$\sup_{f\in Lip_M 1} \|T_n^*[f]-f\| = \frac{\pi\cdot M}{2n}\cdot\|T_n^*\|.$$

Weitere Bemerkungen:
Für die Größen $\|T_n\|$ und $\|T_n^*\|$ ist bekannt:

$$C+\frac{2}{\pi}\ln(2n+1) \leqq \|T_n\| \leqq 1+\frac{2}{\pi}\ln(2n+1)$$

$$C+\frac{2}{\pi}\ln n \leqq \|T_n^*\| \leqq 1+\frac{2}{\pi}\ln n$$

mit

$$C = \frac{2}{\pi}\left(\gamma+\ln\frac{8}{\pi}\right) = 0{,}962\ldots$$

(γ Eulersche Konstante). Näheres hierzu siehe [2] und [3].

§ 2. Benötige Hilfssätze

Seien $y_k=f(\varphi_k)$, $(k=1, 2, \ldots, 2n+1)$, die Funktionswerte von f an den Interpolationsknoten. Das Interpolationspolynom läßt sich schreiben in der Form

$$T_n[f](\varphi) = \sum_{k=1}^{2n+1} y_k\cdot d_k(\varphi)$$

mit wohlbestimmten trigonometrischen Ausdrücken d_k. Es gilt

$$\sum_{k=1}^{2n+1} d_k(\varphi) \equiv 1 \tag{2.1}$$

$$\max_{\varphi} \sum_{k=1}^{2n+1} |d_k(\varphi)| = \|T_n\|. \tag{2.2}$$

Die Knoten seien nun speziell äquidistant vorgegeben

$$\varphi_k = a + \frac{2k\pi}{2n+1}, \quad (k = 1, 2, \ldots, 2n+1).$$

Zur Vereinfachung der Ausdrücke und ohne Beschränkung der Allgemeinheit werde im folgenden stets $a=0$ gewählt. Dann lassen sich die Grundpolynome vereinfachen zu

$$d_k(\varphi) = \frac{(-1)^k}{2n+1} \cdot \frac{\sin \frac{2n+1}{2} \varphi}{\sin \frac{1}{2}(\varphi - \varphi_k)}.$$

Für $\varphi \in]\varphi_n, \varphi_{n+1}[$ gilt demnach

$$\operatorname{sgn} d_k(\varphi) = \begin{cases} (-1)^{n+k} & \text{für} \quad 1 \leqq k \leqq n, \\ (-1)^{n+k+1} & \text{für} \quad n+1 \leqq k \leqq 2n+1. \end{cases} \tag{2.3}$$

Das in [1] angegebene Lemma muß hier dadurch verschärft werden, daß der n-te Summand genauer abgeschätzt wird:

LEMMA 1. *Sei*

$$\sum_{\mu=0}^{m} a_\mu = 0 \tag{2.4}$$

$$\operatorname{sgn} \sum_{\mu=0}^{\nu-1} a_\mu = -\operatorname{sgn} \sum_{\mu=0}^{\nu} a_\mu, \quad (\nu = 1, 2, \ldots, m-1) \tag{2.5}$$

und n, $1 \leqq n \leqq m-1$ *beliebig aber fest gewählt. Mit den Bezeichnungen*

$$A := \sum_{\mu=0}^{n-1} a_\mu, \quad B := \sum_{\mu=n+1}^{m} a_\mu = -\sum_{\mu=0}^{n} a_\mu$$

gilt $\operatorname{sgn} A = \operatorname{sgn} B = -\operatorname{sgn} a_n$, *und es ist*

$$\left| \sum_{\nu=0}^{m} a_\nu b_\nu \right| \leqq \left(\frac{1}{2} \sum_{\mu=0}^{m} |a_\mu| - |a_n| \right) \cdot \max_{1 \leqq k \leqq m} |b_{k-1} - b_k| +$$

$$+ |A| \cdot |b_{n-1} - b_n| + |B| \cdot |b_n - b_{n+1}|.$$

Der Beweis erfolgt am Ende dieses Paragraphen.

Zur Anwendung von Lemma 1 werde für $\varphi\in]\varphi_n, \varphi_{n+1}[$ gewählt:

$$a_\nu = \begin{cases} d_{k+1}(\varphi) & \nu = 0, 1, \ldots, n-1 \\ -1 & \nu = n \\ d_k(\varphi) & \nu = n+1, n+2, \ldots, 2n+1. \end{cases} \tag{2.6}$$

Wie in [1] gezeigt wurde, sind hierfür die Voraussetzungen (2.4) und (2.5) erfüllt. Es ist jetzt $A=A(\varphi)$, $B=B(\varphi)$ und $-\operatorname{sgn} a_n=1$, also

$$A(\varphi) = \sum_{k=1}^{n} d_k(\varphi) > 0, \quad B(\varphi) = \sum_{k=n+1}^{2n+1} d_k(\varphi) > 0.$$

Es gilt $A+B\equiv 1$. Für die Beweisführung von Satz 1 ist notwendig, daß auf mindestens einer Intervallhälfte von $]\varphi_n, \varphi_{n+1}[$ entweder $A\leqq\frac{1}{2}$ oder $B\leqq\frac{1}{2}$ erfüllt ist. Dazu

LEMMA 2. *Sei φ_M der Mittelpunkt des Intervalls $]\varphi_n, \varphi_{n+1}[$. Dann gilt*

(i) *entweder* $0<A(\varphi)\leqq\frac{1}{2}$ *für alle* $\varphi\in[\varphi_M, \varphi_{n+1}[$,

(ii) *oder* $0<B(\varphi)\leqq\frac{1}{2}$ *für alle* $\varphi\in]\varphi_n, \varphi_M]$.

Es kommen zwar beide im Lemma genannten Fälle vor (je nachdem n gerade oder ungerade), aber sie treten nicht gleichzeitig auf. Das kompliziert die Anwendung der Lemmata im nächsten Paragraphen.

BEWEIS von Lemma 1:
Partielle Summation und anschließendes Umordnen der Summanden liefert

$$\begin{aligned}
\left|\sum_{\nu=0}^{m} a_\nu b_\nu\right| &= \left|\sum_{\nu=1}^{m} (b_{\nu-1}-b_\nu)\cdot\sum_{\mu=0}^{\nu-1} a_\mu\right| = \\
&= \left|\sum_{\nu=1}^{n-1} + \sum_{\nu=n+2}^{m} + \sum_{\nu=n}^{n+1} (b_{\nu-1}-b_\nu)\cdot\sum_{\mu=0}^{\nu-1} a_\mu\right| \leqq \\
&\leqq \max_{1\leqq k\leqq m} |b_{k-1}-b_k|\cdot\left\{\sum_{\nu=1}^{m}\left|\sum_{\mu=0}^{\nu-1} a_\mu\right| - \sum_{\nu=n}^{n+1}\left|\sum_{\mu=0}^{\nu-1} a_\mu\right|\right\} + \\
&+ \sum_{\nu=n}^{n+1} |b_{\nu-1}-b_\nu)|\cdot\left|\sum_{\mu=0}^{\nu-1} a_\mu\right| = \\
&= \max_{1\leqq k\leqq m} |b_{k-1}-b_k|\cdot\left\{\sum_{\nu=1}^{m}\left|\sum_{\mu=0}^{\nu-1} a_\mu\right| - (|A|+|B|)\right\} + \\
&\quad + |b_{n-1}-b_n|\cdot|A| + |b_{n+1}-b_n|\cdot|B|.
\end{aligned}$$

Beim Beweis des Lemmas in [1] ist bereits gezeigt worden:

$$\sum_{\nu=1}^{m}\left|\sum_{\mu=0}^{\nu-1} a_\mu\right| \leqq \frac{1}{2}\sum_{\nu=0}^{m} |a_\nu|.$$

Weiter gilt unter Beachtung von (2.5)

$$\operatorname{sgn} A = \operatorname{sgn} B = \operatorname{sgn}(-a_n)$$

$$A+B = -a_n$$

und

$$|A|+|B| = |A+B| = |a_n|.$$

Damit ist Lemma 1 bewiesen.

BEWEIS von Lemma 2:

Auf Grund von $A(\varphi)+B(\varphi)\equiv 1$ und $A(\varphi)>0, B(\varphi)>0$ folgt zunächst $0<A(\varphi)<1$, $0<B(\varphi)<1$. Als nächstes soll gezeigt werden, daß A streng monoton fallend und B streng monoton steigend ist auf dem zu Grunde gelegten Intervall $]\varphi_n, \varphi_{n+1}[$. Es ist $A(\varphi_n)=1$, $A(\varphi_{n+1})=0$. Wäre das trigonometrische Polynom A vom Grade n nicht auf dem ganzen Intervall streng monoton fallend, so hätte die Ableitung dort wegen $0<A(\varphi)<1$ nicht weniger als 2 Nullstellen; zwischen φ_1 und φ_n hat sie nach dem Satz von Rolle auf Grund der Funktionswerte

$$A(\varphi_\nu) = \begin{cases} 1 & \text{für} \quad \nu = 1, 2, \ldots, n \\ 0 & \text{für} \quad \nu = n+1, n+2, \ldots, 2n+1 \end{cases}$$

mindestens $n-1$ Nullstellen, zwischen φ_{n+1} und φ_{2n+1} mindestens n, also zusammen nicht weniger als $2n+1$ Nullstellen, d.h. die Ableitung von A wäre identisch Null, Widerspruch. Also ist A monoton fallend im betrachteten Intervall und somit $B(\varphi)=$ $=1-A(\varphi)$ monoton steigend. Folglich gibt es wegen $A+B\equiv 1$ und der strengen Monotonie genau einen Wert $\bar{\varphi}\in]\varphi_n, \varphi_{n+1}[$ mit

$$A(\bar{\varphi}) = B(\bar{\varphi}) = \frac{1}{2},$$

und es gilt

$$0 < A(\varphi) \leqq \frac{1}{2} \quad \text{für alle} \quad \varphi\in[\bar{\varphi}, \varphi_{n+1}[,$$

$$0 < B(\varphi) \leqq \frac{1}{2} \quad \text{für alle} \quad \varphi\in]\varphi_n, \bar{\varphi}].$$

Es gilt somit Aussage (i) des Lemmas für den Fall $\bar{\varphi}\leqq\varphi_M$ und (ii) für den Fall $\bar{\varphi}\geqq\varphi_M$.

Damit ist Lemma 2 bewiesen.

§ 3. Beweis von Satz 1

Das Maximum von

$$R_n[f](\varphi) = |T_n[f](\varphi)-f(\varphi)|$$

werde in $]\varphi_n, \varphi_{n+1}[$ angenommen, sonst verschiebe man f und damit $T_n[f]$ um ein entsprechendes ganzzahliges Vielfaches des Knotenabstandes. Es sind nun die beiden Fälle von Lemma 2 zu unterscheiden. Angenommen, es gelte Teil (i). Dann kann auch angenommen werden, daß $R_n[f](\varphi)$ das Maximum für $\varphi\in[\varphi_M, \varphi_{n+1}[$ annimmt. (Ansonsten betrachte man im folgenden die Funktion $\tilde{f}$, die aus f durch Spiegelung an

φ_M hervorgeht, also $\tilde{f}(\varphi)=f(2\varphi_M-\varphi)$. Offenbar ist $\|R_n[f](\varphi)\|=\|R_n[\tilde{f}](\varphi)\|$, und nimmt $R_n[f]$ das Maximum in der einen Hälfte von $]\varphi_n, \varphi_{n+1}[$ an, so nimmt $R_n[\tilde{f}]$ dasselbe Maximum in der anderen Hälfte an.)

Zur Anwendung von Lemma 1 werde nun zusätzlich zu (2.6) noch gewählt

$$b_\nu = \begin{cases} f(\varphi_{\nu+1}) & \nu = 0, \ldots, n-1 \\ f(\varphi) & \nu = n \\ f(\varphi_\nu) & \nu = n+1, n+2, \ldots, 2n+1. \end{cases}$$

Damit ergibt sich unter Beachtung von (2.2)

$$R_n[f](\varphi) \leqq \left[\frac{1}{2}\|T_n\| - \frac{1}{2}\right] \cdot \omega\left(\frac{2\pi}{2n+1}\right) + A \cdot \omega(\varphi - \varphi_n) + B \cdot \omega(\varphi_{n+1} - \varphi).$$

Nun ist ω konkav, also

$$A \cdot \omega(\varphi - \varphi_n) + B \cdot \omega(\varphi_{n+1} - \varphi) \leqq \omega[A \cdot (\varphi - \varphi_n) + B \cdot (\varphi_{n+1} - \varphi)].$$

Aus

$$\varphi - \varphi_n = (\varphi_M - \varphi_n) + (\varphi - \varphi_M)$$
$$\varphi_{n+1} - \varphi = (\varphi_{n+1} - \varphi_M) - (\varphi - \varphi_M)$$

und $B = 1 - A$ folgt

$$A \cdot (\varphi - \varphi_n) + B \cdot (\varphi_{n+1} - \varphi) = (\varphi_M - \varphi_n) - (\varphi - \varphi_M) \cdot (1 - 2A) \leqq$$
$$\leqq (\varphi_M - \varphi_n) = \frac{\pi}{2n+1}.$$

§ 4. Beweis von Satz 2

Die Interpolationsknoten seien

$$\varphi_k^* = a + \frac{k\pi}{n} \quad (k = 1, 2, \ldots, 2n).$$

Als trigonometrisches Interpolationspolynom werde hier der folgende trigonometrische Ausdruck vom Grade n verstanden:

$$T_n^*[f](\varphi) = \sum_{k=1}^{2n} y_k \cdot d_k^*(\varphi)$$

mit

$$d_k^*(\varphi) = \frac{1}{2n} \sin n(\varphi - \varphi_k^*) \cdot \cot \frac{1}{2}(\varphi - \varphi_k^*).$$

Ohne Beschränkung der Allgemeinheit kann wieder zur Vereinfachung der Ausdrücke $a=0$ gewählt werden, so daß sich die Grundpolynome d_k^* vereinfachen lassen zu

$$d_k^*(\varphi) = \frac{(-1)^k}{2n} \sin n\varphi \cdot \cot \frac{1}{2}(\varphi - \varphi_k^*), \quad (k = 1, 2, \ldots, 2n).$$

Für $\varphi \in]\varphi_n^*, \varphi_{n+1}^*[$ gilt

$$\operatorname{sgn} d_k^*(\varphi) = \begin{cases} (-1)^{n+k} & 1 \leqq k \leqq n, \\ (-1)^{n+k+1} & n+1 \leqq k \leqq 2n. \end{cases}$$

Die Aussagen und Beweise der Paragraphen 2 und 3 gelten sinngemäß auch hier, wenn man den Index $2n+1$ stets fortläßt bzw. durch $2n$ ersetzt, und anstelle von $]\varphi_n, \varphi_{n+1}[$ bzw. d_ν jetzt $]\varphi_n^*, \varphi_{n+1}^*[$ bzw. d_ν^* schreibt usw.

Nur der Beweis von Lemma 2 muß modifiziert werden. Die dort angestellten Überlegungen liefern auf dem Intervall $]0, 2\pi]$ jetzt nur $2n$ Nullstellen für die Ableitung von A^*. Nun gilt aber

$$d_\nu^*(\varphi-\pi) = d_{n+\nu}^*(\varphi),$$

also $A^*(\varphi-\pi)=B^*(\varphi)$. Durch Addition von $A^*(\varphi)$ auf beiden Seiten ergibt sich auf Grund von $A^*+B^*\equiv 1$

$$A^*(\varphi-\pi) = 1-A^*(\varphi),$$

d.h. aber, die Anzahl der Extrema in $]\varphi_n^*, \varphi_{n+1}^*[= \left]\pi, \pi+\frac{\pi}{n}\right[$ wiederholt sich im Intervall $]0, \varphi_1^*[=]0, \frac{\pi}{n}[$, woraus die zum Beweis benötigte Anzahl von Nullstellen und demnach die Monotonie von A^* folgt. Es gilt sogar hier $\bar{\varphi}=\varphi_M$, denn aus

$$d_{n-\nu+1}^*(\varphi_M) = d_{n+\nu}^*(\varphi_M), \quad (\nu = 1, 2, \ldots, n)$$

folgt sofort

$$A^*(\varphi_M) = B^*(\varphi_M) = \frac{1}{2}.$$

Damit läßt sich Lemma 2 verschärfen zu

LEMMA 2^*. *Sei φ_M der Mittelpunkt des Intervalls $]\varphi_n^*, \varphi_{n+1}^*[$.* Dann gilt

(i) $0 < A^*(\varphi) \leqq \frac{1}{2}$ für alle $\varphi\in[\varphi_M, \varphi_{n+1}^*[$ und

(ii) $0 < B^*(\varphi) \leqq \frac{1}{2}$ für alle $\varphi\in]\varphi_n^*, \varphi_M]$.

Somit vereinfacht sich hier der in § 3 angegebene Beweis, denn nimmt $R_n[f](\varphi)$ das Maximum in $[\varphi_M, \varphi_{n+1}^*[$ bzw. $]\varphi_n^*, \varphi_M]$ an, so kann unmittelbar Lemma 1 in Verbindung mit Lemma 2^* (i) bzw. (ii) angewendet werden.

LITERATUR

[1] BRASS, H.—GÜNTTNER, R.: Eine Fehlerabschätzung zur Interpolation stetiger Funktionen. *Studia Sci. Math. Hungar.* **8** (1973) 363—367.

[2] EHLICH, H.—ZELLER, K.: Auswertung der Normen von Interpolationsoperatoren. *Math. Annalen* **164** (1966) 105—112.

[3] GÜNTTNER, R.: *Abschatzungen für Normen von Interpolationsoperatoren.* (Dissertation, Clausthal-Z., 1972) 79—80.

[4] KORNEICUK, N. P.: The best uniform Approximation on certain Classes of continous Functions. *Dokl. Akad. Nauk SSSR* **140** (1961) 748.= *Soviet Math. Dokl.* **2** (1961) 1254—1257.

Universität Osnabrück

(Eingegangen 28, November, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 131—140.

THE HAMMING-SPHERE HAS MINIMUM BOUNDARY

by

G. O. H. KATONA

Introduction

In their information-theoretical investigations [6] AHLSWEDE, GÁCS and KÖRNER needed the solution of the following problem. Let $\mathscr{A}$ be a subset of the space of 0—1 sequences of length n. The Hamming-distance $\varrho(a, b)$ of the sequences a and b is the number of places where they differ. $\delta(\mathscr{A})$ is the set of sequences which have Hamming-distance $\leqq 1$ at least from one element of $\mathscr{A}$. The question: what is the minimum of $|\delta(\mathscr{A})|$ ($|X|$ means the number of elements of X) if $|\mathscr{A}|$ is given. To determine the minimum of $|\delta(\mathscr{A})-\mathscr{A}|$ is an equivalent question. They have found an asymptotical solution in a paper of MARGULIS [5], but the problem of determining the exact minimum remained open*. The aim of this paper is to give the exact minimum.

If $|\mathscr{A}|$ allows, the optimal $\mathscr{A}$ is a Hamming-sphere. If $|\mathscr{A}|$ is different, then we have to choose some additional points in a suitable way.

The proof seems to be quite complicated, but it is very easy after knowing the technique of a similar combinatorial question described below (see also Theorem 1): Let $\mathscr{A}$ be a family of k-tuples of an n-element set (0—1 sequences with k 1's). Determine the minimal number of $(k-1)$-tuples which are contained at least in one k-tuple of $\mathscr{A}$ (that is, "lower" Hamming-boundary). This question was solved first by KRUSKAL [1], later (but independently) by the author [2]. The technique is used in the proofs of HANSEL [3] and ECKHOFF and WEGNER [4]. This last proof is the shortest variant of this type. (For other ways of proofs see [7] and [8].) We did not succeed in reducing our problem to this one, but we use the methods. We use heavily an inequality (see Lemma 2) which appears in different forms in [2], [3] and [4].

There is a natural correspondence between the 0—1 sequences of length n and the subsets of an n-element set. We use both terms alternately.

Summary of the used earlier results

LEMMA 1. *If m and k are given non-negative integers, then there is a unique representation*

$$m = \binom{a_k}{k} + \binom{a_{k-1}}{k-1} + \ldots + \binom{a_t}{t}, \tag{1}$$

where

$$a_k > a_{k-1} > \ldots > a_t \geqq t \geqq 1.$$

* In the paper of Margulis it is slightly differently formulated. $\partial(\mathscr{A})$ consists of the sequences x belonging to $\mathscr{A}$ and having a sequence $y \notin \mathscr{A}$ with Hamming distance $\partial(x, y)=1$. However, it is easy to see that $\partial(\mathscr{A})=\partial(\bar{\mathscr{A}})-\bar{\mathscr{A}}$.

The proof can be find in [1]–[2].

(1) is called the *k-canonical representation* of m. We define $(k, m>0)$

$$F(k, m) = \binom{a_k}{k-1} + \binom{a_{k-1}}{k-2} + \ldots + \binom{a_t}{t-1}, \tag{2}$$

$$F(k, 0) = 0.$$

LEMMA 2. *If* $k>0$, $m_1, m_2 \geqq 0$, *then*

$$F(k+1, m_1+m_2) \leqq \max\left(m_2, F(k+1, m_1)\right) + F(k, m_2). \tag{3}$$

This inequality appears in a modified form in [2], and [3]. This form and the shortest proof can be found in [4].

If $\mathscr{A}$ is a family of k-element subsets of a set of n elements, then $\delta_L(\mathscr{A})$ means the family of $k-1$-element subsets which are subsets of a k-element set $\in \mathscr{A}$.

THEOREM 1. *If* $\mathscr{A}$ *consists of different* k*-element subsets of an* n*-element set and* $0 \leqq m = |\mathscr{A}| \leqq \binom{n}{k}$, *then*

$$|\delta_L(\mathscr{A})| \geqq F(k, m)$$

and this is the best lower bound.

This theorem can be found in [1], [2], [3], [4], [7] and [8], and it is an easy consequence of Lemma 2. It is easy to see, that $|\delta_L(\mathscr{A})| = F(k, m)$ if we choose the first m 0—1 sequences with k 1's in the lexicographic order.

LEMMA 3. *If* $0<k$, $0 \leqq m_1, m_2$ *then*

$$F(k, m_1+m_2) \leqq F(k, m_1) + F(k, m_2), \tag{4}$$

PROOF. It can be found in [2]. However (4) is an easy consequence of Theorem 1, thus we give here the proof. Take two disjoint sets S_1 and S_2 of n_1 $\left(m_1 \leqq \binom{n_1}{k}\right)$ and n_2 elements $\left(m_2 \leqq \binom{n_2}{k}\right)$, respectively. Construct an optimal family of n_1 k-tuples on S_1 which contains $F(k, m_1)$ $(k-1)$-tuples. Take the same for S_2. It means, that the family on $S_1 \cup S_2$ contains exactly $F(k, m_1)+F(k, m_2)$ $(k-1)$-tuples. By theorem 1 this family must contain at least $F(k, m_1+m_2)$ $(k-1)$-tuples. This gives (4).

LEMMA 4. *If* $0<k$; $0 \leqq m_1 \leqq m_2 \leqq \binom{n}{k}$, $\binom{n}{k} \leqq m_1+m_2$, *then*

$$\binom{n}{k-1} + F\left(k, m_1+m_2-\binom{n}{k}\right) \leqq F(k, m_1) + F(k, m_2). \tag{5}$$

PROOF. If $m_1 = \binom{n}{k}$ or $m_2 = \binom{n}{k}$, equality holds in (5). Thus we may assume $m_1, m_2 < \binom{n}{k}$, that is,

$$F\left(k+1, \binom{n}{k+1} + m_1\right) = \binom{n}{k} + F(k, m_1) > m_2.$$

On the other hand,

$$F\left(k+1, \binom{n+1}{k+1}+m_1+m_2-\binom{n}{k}\right) = \binom{n+1}{k}+F\left(k, m_1+m_2-\binom{n}{k}\right),$$

using, that $m_1+m_2-\binom{n}{k}<\binom{n}{k}$. Now we shall use Lemma 2 with the numbers $\binom{n}{k+1}+m_1$ and m_2:

$$F\left(k+1, \binom{n}{k+1}+m_1+m_2\right) = F\left(k+1, \binom{n+1}{k+1}+m_1+m_2-\binom{n}{k}\right) =$$

$$= \binom{n+1}{k}+F\left(k, m_1+m_2-\binom{n}{k}\right) \leqq F\left(k+1, \binom{n}{k+1}+m_1\right)+F(k, m_2) =$$

$$= \binom{n}{k}+F(k, m_1)+F(k, m_2).$$

Thus we obtained an inequality which is equivalent to (5).

A consequence. Theorem 2 in [2] (which has a complicated proof in [2]) is an easy consequence of this lemma. The theorem says (in a slightly more general form), that if $\mathscr{A}$ is a family of different k-tuples on a set $S_1 \cup S_2$ ($S_1 \cap S_2 = \varnothing$), where $|S_2| \leqq |S_1| = n$, $\binom{n}{k} \leqq |\mathscr{A}| \leqq \binom{n}{k}+\binom{|S_2|}{k}$ and at most one of the relations $A \cap S_1 \neq \varnothing$ $A \cap S_2 \neq \varnothing$ ($A \in \mathscr{A}$) holds, then

$$|\delta_L(\mathscr{A})| \geqq \binom{n}{k-1}+F\left(k, |\mathscr{A}|-\binom{n}{k}\right).$$

That is, the best arrangement is, if we choose all the k-tuples from S_1 and the remainder from S_2.

Proof. Let m_1 and m_2 denote the number of the subsets ($\in \mathscr{A}$) contained by S_1 and S_2, respectively. The minimum of $(k-1)$-tuples "contained by $\mathscr{A}$" in S_1 is $F(k, m_1)$ by Theorem 1, and $F(k, m_2)$ in S_2. Thus Lemma 4 gives the result.

The results

We start with an analogue of lemma 1.

LEMMA 5. *If u and n are given non-negative integers ($u < 2^n$) then there is a unique representation (called n-bounded canonical representation)*

$$u = \binom{a_n}{n}+\binom{a_{n-1}}{n-1}+\ldots+\binom{a_t}{t}, \tag{6}$$

where $n = a_n = a_{n-1} = \ldots = a_{k+1} > a_k > a_{k-1} > \ldots > a_t \geqq t \leqq 1$ *for some* $k (t-1 \leqq k < n)$.

May be, it is better to write

$$u = \binom{n}{n} + \binom{n}{n-1} + \ldots + \binom{n}{k+1} + \binom{a_k}{k} + \ldots + \binom{a_t}{t} \tag{7}$$

$$(n > a_k > a_{k-1} > \ldots > a_t \geqq t \geqq 1),$$

with the remark that the part of a's can completely vanish.

Now we are able to introduce the following notation

$$G(n, u) = \binom{n}{n} + \binom{n}{n-1} + \ldots + \binom{n}{k+1} + \binom{n}{k} + \binom{a_k}{k-1} + \ldots + \binom{a_t}{t-1} \tag{8}$$

if $u > 0$, and $G(n, 0) = 0$.

LEMMA 6. *If* $0 \leqq u_1 \leqq u_2$, *then*

$$G(n, u_1 + u_2) \leqq \max\left(u_2, G(n-1, u_1)\right) + G(n-1, u_2). \tag{9}$$

If $\mathscr{A} = \{A_1, \ldots, A_u\}$ is a family of different subsets of an n-element set S, then $\delta(\mathscr{A})$ denotes the family of subsets B of S, the Hamming-distance $\varrho(B, A_i)$ of which is $\leqq 1$ at least for one member A_i of $\mathscr{A}$.

THEOREM 2. *If* $\mathscr{A}$ *is a system of different subsets of an n-element set S and* $|\mathscr{A}| = u < 2^n$, *then*

$$|\delta(\mathscr{A})| \geqq G(n, u)$$

and this is the best possible bound.

In general, $\delta_d(\mathscr{A})$ is defined in the following way:

$$\delta_d(\mathscr{A}) = \{B : \exists A \in \mathscr{A},\ \varrho(B, A) \leqq d\}.$$

Similarly, we need the generalization of (7):

$$G_d(n, u) =$$

$$= \binom{n}{n} + \binom{n}{n-1} + \ldots + \binom{n}{k+1} + \binom{n}{k} + \ldots + \binom{n}{k-d+1} + \binom{a_k}{k-d} + \ldots + \binom{a_t}{t-d}.$$

The following theorem is a more general form of Theorem 2:

THEOREM 3. *If* $|\mathscr{A}| = u < 2^n$ *then*

$$|\delta_d(\mathscr{A})| \geqq G_d(n, u).$$

PROOF of Lemma 5. (Warning: it is easier to prove than read!) First we prove there is a representation of form (6). Take the minimal k satisfying

$$u \geqq \binom{n}{n} + \ldots + \binom{n}{k+1} = v.$$

Then, applying lemma 1 for $u - v$ we obtain a representation of form (6). It remains only to prove that $n > a_k$. In the contrary case

$$u \geqq \binom{n}{n} + \ldots + \binom{n}{k+1} + \binom{a_k}{k} \geqq \binom{n}{n} + \ldots + \binom{n}{k+1} + \binom{n}{k}$$

holds, thus k was not the minimum, in contradiction with our suppositions.

We have to prove that (6) is unique. Suppose, the contrary case holds, there are two representations. If the k's are the same in both, then $u-v$ has two different representations of form (1) contradicting lemma 1. We can suppose that the k's are different ($k>k'$). Let the other representation be

$$u=\binom{n}{n}+\binom{n}{n-1}+\ldots+\binom{n}{k'+1}+\binom{b_{k'}}{k'}+\ldots+\binom{b_{t'}}{t'}. \tag{10}$$

Using a well-known formula

$$\binom{a_k}{k}+\ldots+\binom{a_t}{t}<\binom{n-1}{k}+\binom{n-2}{k-1}+\ldots+\binom{n-k+t-1}{t}+$$
$$+\binom{n-k+t-2}{t-1}+\ldots+\binom{n-k-1}{0}=\binom{n}{k}.$$

Thus,

$$u<\binom{n}{n}+\binom{n}{n-1}+\ldots+\binom{n}{k+1}+\binom{n}{k},$$

from (6), and

$$u\geqq\binom{n}{n}+\binom{n}{n-1}+\ldots+\binom{n}{k+1}+\binom{n}{k}$$

from (10). These two statements contradict each other. The lemma is proved.

PROOF of theorem 2. First we reduce the theorem to Lemma 6 which will be proved afterwards.

We use induction over n. If $n=1$, then $u=1=\binom{1}{1}$, $G(1,1)=\binom{1}{1}+\binom{1}{0}=2$, and $|\delta(\mathscr{A})|$ is always 2. Assume the theorem is proved for $n-1$, and prove it for n.

Fix an element x of S, and divide $\mathscr{A}$ into two families. $\mathscr{A}_1$ or $\mathscr{A}_2$ consists of the subsets which contain or do not contain x, respectively. The operation * on a family of subsets means that x is added to the members of the family which do not contain it and it is omitted from the members which do. Denote $|\mathscr{A}_1|$ and $|\mathscr{A}_2|$ by z_1 and z_2, respectively. Obviously,

$$z=|\mathscr{A}|=z_1+z_2.$$

We distinguish several cases.

Case 1. $z_1\leqq z_2\leqq G(n-1,z_1)$. We have, by the induction hypothesis,

$$|\delta(\mathscr{A}_2)|\geqq G(n-1,z_2) \tag{11}$$

and

$$|\delta(\mathscr{A}_1^*)|\geqq G(n-1,z_1) \tag{12}$$

as $|\mathscr{A}_1^*|=|\mathscr{A}_1|=z_1$. (Here δ is taken for the $(n-1)$-element set $S-\{x\}$). Similarly,

$$|\delta(\mathscr{A}_1^*)^*|\geqq G(n-1,z_1) \tag{13}$$

follows from (12). As $\delta(\mathscr{A}_2)\subset\delta(\mathscr{A})$, $\delta(\mathscr{A}_1^*)^*\subset\delta(\mathscr{A})$ and they are disjoint,

$$|\delta(\mathscr{A})|\geqq|\delta(\mathscr{A}_2)|+|\delta(\mathscr{A}_1^*)^*|\geqq$$
$$\geqq G(n-1,z_2)+G(n-1,z_1)$$

from (11) and (12). However, this is at least

$$G(n, z_1+z_2) = G(n, z)$$

by Lemma 6 and the suppositions of this case. Case 1 is settled.

Case 2. $z_1 \leqq z_2$, $G(n-1, z_1) \leqq z_2$. Now, we use $\delta(\mathscr{A}) \supset \delta(\mathscr{A}_2)$, $\delta(\mathscr{A}) \supset \mathscr{A}_2^*$ and $\delta(\mathscr{A}_2) \cap \mathscr{A}_2^* = \emptyset$. These fact result in

$$|\delta(\mathscr{A})| \geqq |\delta(\mathscr{A}_2)| + |\mathscr{A}_2^*|.$$

Here $|\mathscr{A}_2^*| = z_2$, and by the inductional hypothesis $|\delta(\mathscr{A}_2)| \geqq G(n-1, z_2)$, thus

$$|\delta(\mathscr{A})| \geqq z_2 + G(n-1, z_2).$$

The right hand side is at least $G(n, z_1+z_2) = G(n, z)$ by lemma 6 and the suppositions of this case. This case in settled, too.

Case 3. $z_2 \leqq z_1 \leqq G(n-1, z_2)$. We can repeat the proof of Case 1. The only difference, that in lemma 6 we have to write z_2 in place of u_1 and z_1 in place of u_2.

Case 4. $z_2 \leqq z_1$, $G(n-1, z_2) \leqq z_1$. Now, we use $\delta(\mathscr{A}) \supset \delta(\mathscr{A}_1^*)$, $\delta(\mathscr{A}) \supset \mathscr{A}_1$ and $\delta(\mathscr{A}_1^*) \cap \mathscr{A}_1 = \emptyset$. These facts result in

$$|\delta(\mathscr{A})| \geqq |\delta(\mathscr{A}_1^*)| + |\mathscr{A}_1|.$$

Here $|\mathscr{A}_1| = z_1$, and by the inductional hypothesis $|\delta(\mathscr{A}_1^*)| \geqq G(n-1, z_1)$, thus

$$|\delta(\mathscr{A})| \geqq z_1 + G(n-1, z_1).$$

The right-hand side is at least $G(n, z_1+z_2) = G(n, z)$ by lemma 6 and the suppositions of this case. The inequality of theorem 2 is proved.

We have to construct an $\mathscr{A}$ showing that the inequality is the best possible. Let $\mathscr{A}$ consist of all the subsets having at least $k+1$ elements and of the first $u - \binom{n}{n} - \ldots - \binom{n}{k+1}$ k-tuples in the lexicographic order. It is easy to see that $\delta(\mathscr{A})$ contains all the subsets having at least k elements and $\binom{a_k}{k-1} + \ldots + \binom{a_t}{t-1}$ $(k-1)$-tuples according to Theorem 1. The proof is completed.

PROOF of Lemma 6. *Case* 1. $G(n-1, u_1) \leqq u_2$. It is easy to see that $G(n, u)$ is monotonically increasing in u. We have to prove

$$G(n, u_1+u_2) \leqq u_2 + G(n-1, u_2). \tag{14}$$

By the monotonity it is enough to prove this inequality for the maximal possible u_1 satisfying $G(n-1, u_1) \leqq u_2$. Suppose, u_2 has the $(n-1)$-bounded representation

$$u_2 = \binom{n-1}{n-1} + \ldots + \binom{n-1}{\gamma+1} + \binom{c_\gamma}{\gamma} + \binom{c_{\gamma-1}}{\gamma-1} + \ldots + \binom{c_s}{s}, \tag{15}$$

and μ is the smallest index satisfying $c_\mu > \mu$. Then

$$U_1 = \binom{n-1}{n-1} + \ldots + \binom{n-1}{\gamma+2} + \binom{c_\gamma}{\gamma+1} + \ldots + \binom{c_\mu}{\mu+1} \tag{16}$$

satisfies $G(n-1, U_1) \leqq u_2$. But U_1+1 does not satisfy it. This is trivial if $\mu = s$.

If $\mu > s$ then U_1+1 has an additional term $\binom{c_{\mu-1}}{\mu} = \binom{\mu}{\mu} = 1$ and

$$G(n-1, U_1+1) =$$

$$= \binom{n-1}{n-1} + \ldots + \binom{n-1}{\gamma+1} + \binom{c_\gamma}{\gamma} + \ldots + \binom{c_\mu}{\mu} + \binom{\mu}{\mu-1} >$$

$$> \binom{n-1}{n-1} + \ldots + \binom{n-1}{\gamma+1} + \binom{c_\gamma}{\gamma} + \ldots + \binom{c_\mu}{\mu} + \mu - s.$$

Thus, we really can consider U_1 for u_1 in (14): Here, from (16)

$$U_1 + u_2 = \binom{n}{n} + \ldots + \binom{n}{\gamma+2} + \binom{c_\gamma+1}{\gamma+1} + \ldots + \binom{c_\mu+1}{\mu+1} + \binom{\mu}{\mu} + \ldots + \binom{s+1}{s+1},$$

and

$$G(n, U_1+u_2) =$$

$$= \binom{n}{n} + \ldots + \binom{n}{\gamma+1} + \binom{c_\gamma+1}{\gamma} + \ldots + \binom{c_\mu+1}{\mu} + \binom{\mu}{\mu-1} + \ldots + \binom{s+1}{s} =$$

$$= \binom{n-1}{n-1} + \ldots + \binom{n-1}{\gamma+1} + \binom{c_\gamma}{\gamma} + \ldots + \binom{c_\mu}{\mu} + \binom{\mu-1}{\mu-1} + \ldots + \binom{s}{s} +$$

$$+ \binom{n-1}{n-1} + \ldots + \binom{n-1}{\gamma} + \binom{c_\gamma}{\gamma-1} + \ldots + \binom{c_\mu}{\mu-1} + \binom{\mu-1}{\mu-2} + \ldots + \binom{s}{s-1} =$$

$$= u_2 + G(n-1, u_2).$$

The case is settled.

Case 2. $u_2 < G(n-1, u_1)$. Let u_1 have the form

$$u_1 = \binom{n-1}{n-1} + \ldots + \binom{n-1}{\beta+1} + \binom{b_\beta}{\beta} + \ldots + \binom{b_r}{r}. \tag{17}$$

From the inequality $u_1 \leqq u_2$ (see (15)) it follows $\beta \geqq \gamma$. On the other hand, from $u_2 < G(n-1, u_1)$ $\beta \leqq \gamma+1$ follows and if $\beta = \gamma+1$, then

$$v_2 = \binom{c_\gamma}{\gamma} + \ldots + \binom{c_s}{s} < \binom{b_{\gamma+1}}{\gamma} + \ldots + \binom{b_r}{r-1}. \tag{18}$$

Summarizing, β can be γ or $\gamma+1$. These two cases will be distinguished.

Case 2a. $\beta = \gamma$. Let us introduce the following notations

$$v_1 = \binom{b_\beta}{\beta} + \ldots + \binom{b_r}{r}$$

$$v_2 = \binom{c_\beta}{\beta} + \ldots + \binom{c_s}{s}.$$

Then

$$u_1 + u_2 = \binom{n}{n} + \ldots + \binom{n}{\gamma+2} + \left[\binom{n-1}{\gamma+2} + v_1 + v_2\right]. \tag{19}$$

Unfortunately, it is not a perfect form for taking $G(n, u_1+u_2)$. However, if

Case 2aa. $v_1+v_2<\binom{n-1}{\gamma}$, then $\binom{n-1}{\gamma+1}+v_1+v_2<\binom{n}{\gamma+1}$ that is, the bracket does not disturb the part $\binom{n}{n}+\ldots+\binom{n}{\gamma+2}$ in (19). Thus

$$G(n, u_1+u_2)=\binom{n}{n}+\ldots+\binom{n}{\gamma+1}+F\left(\gamma+1, \binom{n-1}{\gamma+1}+v_1+v_2\right)= \tag{20}$$

$$=\binom{n}{n}+\ldots+\binom{n}{\gamma+1}+\binom{n-1}{\gamma}+F(\gamma, v_1+v_2).$$

On the other hand

$$G(n-1, u_1)=\binom{n-1}{n-1}+\ldots+\binom{n-1}{\gamma}+F(\gamma, v_1) \tag{21}$$

and

$$G(n-1, u_2)=\binom{n-1}{n-1}+\ldots+\binom{n-1}{\gamma}+F(\gamma, v_2). \tag{22}$$

From (20), (21) and (22) it is easy to see that (9) is reduced to $F(\gamma, v_1+v_2)\leqq F(\gamma, v_1)+$ $+F(\gamma, v_2)$ which is lemma 3. We can turn to the next case.

Case 2ab. $v_1+v_2\geqq\binom{n-1}{\gamma}$. In this case we use a modified form of (19):

$$u_1+u_2=\binom{n}{n}+\ldots+\binom{n}{\gamma+2}+\binom{n}{\gamma+1}+\left[v_1+v_2-\binom{n-1}{\gamma}\right].$$

Here $v_1+v_2-\binom{n-1}{\gamma}<\binom{n-1}{\gamma}$ and the last term in bracket can not disturb the previous terms.

$$G(n, u_1+u_2)=\binom{n}{n}+\ldots+\binom{n}{\gamma}+F\left(\gamma, v_1+v_2-\binom{n-1}{\gamma}\right). \tag{23}$$

(23), (21) and (22) lead to

$$\binom{n-1}{\gamma-1}+F\left(\gamma, v_1+v_2-\binom{n-1}{\gamma}\right)\leqq F(\gamma, v_1)+F(\gamma, v_2).$$

This is true by lemma 4. Case 2a is proved.

Case 2b. $\beta=\gamma+1$. Now

$$u_1+u_2=\binom{n}{n}+\ldots+\binom{n}{\gamma+2}+(v_1+v_2)$$

holds, where $v_2<\binom{n-1}{\gamma}$, $v_1<\binom{n-1}{\gamma+1}$, thus $v_1+v_2<\binom{n}{\gamma+1}$. The term v_1+v_2 does not disturb the previous ones. Hence

$$G(n, u_1+u_2)=\binom{n}{n}+\ldots+\binom{n}{\gamma+1}+F(\gamma+1, v_1+v_2). \tag{24}$$

Furthermore the terms on the right-hand side of (9) have the forms

$$G(n-1, u_1) = \binom{n-1}{n-1} + \ldots + \binom{n-1}{\gamma+1} + F(\gamma+1, v_1) \tag{25}$$

$$G(n-1, u_2) = \binom{n-1}{n-1} + \ldots + \binom{n-1}{\gamma} + F(\gamma, v_2). \tag{26}$$

Comparing (24), (25) and (26), (9) reduces to

$$F(\gamma+1, v_1+v_2) \leqq F(\gamma+1, v_1) + F(\gamma, v_2).$$

This is true by lemma 2 if

$$v_2 \leqq F(\gamma+1, v_1) = \binom{b_{\gamma+1}}{\gamma} + \ldots + \binom{b_r}{r-1}.$$

But this is (18) which always holds when $\beta = \gamma+1$. The proof of the lemma is completed.

PROOF of theorem 3. We prove the theorem by induction over d. For $d=1$ it is theorem 2. Suppose $d>1$ and the statement is proved for smaller values. Observe that $\delta_d(\mathscr{A}) = \delta(\delta_{d-1}(\mathscr{A}))$ and hence by the inductional hypothesis

$$|\delta_d(\mathscr{A})| \geqq G(n, G_{d-1}(n, u)).$$

We have only to prove

$$G(n, G_{d-1}(n, u)) = G_d(n, u). \tag{27}$$

This is trivial, if

$$u = \binom{n}{n} + \ldots + \binom{n}{k+1} + \ldots + \binom{a_t}{t}$$

and $t \geqq d$ or $k+1 \leqq d$. Otherwise, if $t < d < k+1$, then

$$G_{d-1}(n, u) = \binom{n}{n} + \ldots + \binom{n}{k-d+2} + \binom{a_k}{k-d+1} + \ldots + \binom{a_t}{t-d+1} =$$

$$= \binom{n}{n} + \ldots + \binom{n}{k-d+2} + \binom{a_k}{k-d+1} + \ldots + \binom{a_{d-1}}{0}.$$

Let μ be minimal index such that $a_\mu < a_{\mu+1} - 1$ $(d-1 < \mu \leqq k)$. Then

$$G_{d-1}(n, u) = \binom{n}{n} + \ldots + \binom{n}{k-d+2} + \binom{a_k}{k-d+1} + \ldots + \binom{a_{\mu+1}}{\mu-d+2} + \binom{a_\mu+1}{\mu-d+1}$$

and

$$G(n, G_{d-1}(n, u)) =$$

$$= \binom{n}{n} + \ldots + \binom{n}{k-d+2} + \binom{n}{k-d+1} + \binom{a_k}{k-d} + \ldots + \binom{a_{\mu+1}}{\mu-d+1} + \binom{a_\mu+1}{\mu-d}.$$

On the other hand,

$$G_d(n,u) = \binom{n}{n} + \ldots + \binom{n}{k-d+1} + \binom{a_k}{k-d} + \ldots + \binom{a_d}{0} =$$

$$= \binom{n}{n} + \ldots + \binom{n}{k-d+1} + \binom{a_k}{k-d} + \ldots + \binom{a_{\mu+1}}{\mu-d+1} + \binom{a_\mu+1}{\mu-d}.$$

This shows (27) and the theorem. The proof is completed.

Remarks and an open problem

Theorem 2 gives a formula for the min $|\delta(\mathscr{A})|$. It is easy to derive formulas from it for min $|\delta(\mathscr{A})-\mathscr{A}|$ and min $|\partial(\mathscr{A})|$:

$$\min|\delta(\mathscr{A})-\mathscr{A}| = \min|\delta(\mathscr{A})| - |\mathscr{A}| =$$

$$= \binom{n}{k} + \binom{a_k}{k-1} + \ldots + \binom{a_t}{t-1} - \binom{a_k}{k} - \ldots - \binom{a_t}{t}.$$

On the other hand

$$\min|\partial(\mathscr{A})| = \min|\delta(\overline{\mathscr{A}})-\overline{\mathscr{A}}| = \min|\delta(\overline{\mathscr{A}})| - \overline{\mathscr{A}}.$$

Thus, we have to write $2^n-|\mathscr{A}|$ into the n-bounded canonical representation, and $G(n, 2^n-|\mathscr{A}|)$ gives the minimum.

An open question: What is the minimum of $|\delta(\mathscr{A})|$ if $|\mathscr{A}|$ is fixed and $|A|=k$ for $A\in\mathscr{A}$?

REFERENCES

[1] KRUSKAL, J. B.: The number of simplicies in a complex, *Mathematical Optimization Techniques*, University of Calif. Press, Berkeley and Los Angeles, 1963, pp. 251—278.
[2] KATONA, G.: A theorem of finite sets, Theory of Graphs, Proc. Coll. held at Tihany 1966, Akadémiai Kiadó, 1968, pp. 187—207.
[3] HANSEL, G.: Complexes et décompositions binomiales, *J. Combinatorial Th.* **7** (1969) 230—238.
[4] ECKHOFF, J. und WEGNER, G.: Über einen Satz von Kruskal (to appear in Periodica Math. Hung.).
[5] MARGULIS, A. A.: Probabilistic properties of graphs with large connectivity, *Probl. Peredachi i Informacii*, **10** (1974) (2) 101—108.
[6] AHLSWEDE, R., GÁCS P. and KÖRNER, J.: Bounds on conditional probabilities with applications in multi-user communication, *Zeitschrift f. Wahrscheinlichkeitsth. verw. Geb.* (To appear)
[7] CLEMENTS, G. F. and LINDSTRÖM, B.: A generalization of a combinatorial theorem of Macaulay, *J. Combinatorial Th.* **7** (1969) 230—238.
[8] DAYKIN, D. E.: A simple proof of the Kruskal-Katona theorem, *J. Combinatorial The. A.* **17** (1974) 252—253.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received December 10, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 141—146.

CENTRAL LIMIT THEOREMS FOR WEAKLY LACUNARY WALSH SERIES

by

ANTÓNIA FÖLDES

The statistical properties of lacunary Walsh series (like the trigonometric system) proved to be similar to that of the independent random variables (See [1], [2], [3]).

Let

$$R_k(x) = \operatorname{sign}(\sin 2^k \Pi x), \quad 0 \leqq x \leqq 1$$

the k-th Rademacher function, and for an integer $n = \sum \varepsilon_k 2^k$ ($\varepsilon_k = 0$, or $\varepsilon_k = 1$) let the Walsh functions $\{w_n(x)\}$ be defined as

$$w_n(x) = \prod_{\varepsilon_k = 1} R_{k+1}(x), \quad 0 \leqq x \leqq 1$$

In [2] the following two theorems are proved.

THEOREM 1. (*Central limit theorem.*) *If* $\{n_k\}$ *is a sequence of positive integers such that for every* $k \geqq 1$

$$\frac{n_{k+1}}{n_k} \geqq 1 + k^{-\alpha} \tag{1}$$

and $0 \leqq \alpha < \frac{1}{2}$, *then*

$$\lim_{N \to \infty} \mathrm{P}\left(N^{-\frac{1}{2}} \sum_{k=1}^{N} w_{n_k}(x) \leqq u \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{t^2}{2}} dt. \tag{2}$$

In turn for every $\alpha > \frac{1}{2}$ *there exists an* $\{n_k\}$ *satisfying* (1) *for which* (2) *is no longer valid.*

THEOREM 2. (*Law of the iterated logarithm.*) *If* $\{n_k\}$ *is a sequence of positive integers such that for every* $k \geqq 1$ (1) *holds and* $0 \leqq \alpha < \frac{1}{2}$ *then we have for a.e.* $x \in [0, 1]$

$$\overline{\lim_{N \to \infty}} (2N \log \log N)^{-\frac{1}{2}} \sum_{k=1}^{N} w_{n_k}(x) = 1. \tag{3}$$

In a recent paper S. TAKAHASHI [3] gave the following generalization of Theorem 2.

THEOREM 2A. *Let* $\{n_k\}$ *be a sequence of positive integers such that for every* $k \geqq 1$

$$\frac{n_{k+1}}{n_k} > 1 + ck^{-\alpha} \quad c > 0, \quad 0 \leqq \alpha \leqq \frac{1}{2} \tag{4}$$

and let further $\{a_k\}$ *be a sequence of real numbers satisfying the following conditions:*

$$A_N = \left(\sum_{k=1}^{N} a_k^2\right)^{\frac{1}{2}} \to +\infty$$

and

$$a_N = O\left(\frac{A_N}{N^{\alpha}(\log A_N)^{\frac{1+\varepsilon}{2}}}\right) \tag{5}$$

as $N \to \infty$, *where* ε *is a positive constant. Then we have for a.e.* $x \in [0, 1]$

$$\overline{\lim_{N\to\infty}} (2A_N^2 \log\log A_N^2)^{-\frac{1}{2}} \sum_{k=1}^{N} a_k w_{n_k}(x) = 1.$$

The aim of the present paper is to give a similar generalization of Theorem 1:

THEOREM 1A. *Let* $\{n_k\}$ *be a sequence of positive integers such that for every* $k \geqq 1$

$$\frac{n_{k+1}}{n_k} > 1 + ck^{-\alpha} \quad c > 0, \quad 0 \leqq \alpha \leqq \frac{1}{2}.$$

Let further $\{a_k\}$ *be a sequence of real numbers satisfying the following conditions:*

$$A_N = \left(\sum_{k=1}^{N} a_k^2\right)^{\frac{1}{2}} \to +\infty$$

and

$$a_N = o\left(\frac{A_N}{N^{\alpha}}\right) \quad \text{as} \quad N \to \infty. \tag{6}$$

Then we have

$$\lim_{N\to\infty} \mathsf{P}\left(\frac{\sum_{k+1}^{N} a_k w_{n_k}(x)}{A_N} < u\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{t^2}{2}} dt. \tag{7}$$

To prove this theorem we use some facts from TAKAHASHI [3] and a theorem of MCLEISH [4].

Let $\{X_{n,i};\ 0 \leqq i \leqq l_n\}$ be an array of random variables on the probability triple $(\Omega, \mathscr{F}, \mathsf{P})$ and denote $S_n = \sum_{i=0}^{l_n} X_{n,i}$. Let $\{\mathscr{F}_{n,i};\ 0 \leqq i \leqq l_n\}$ be any triangular array of sub-sigma fields of $\mathscr{F}$ such that $\mathscr{F}_{n,i-1} \subset \mathscr{F}_{n,i}$ for each n and $1 \leqq i \leqq l_n$. We shall call the array $\{X_{n,i}\}$ a martingale difference array with respect to $\{\mathscr{F}_{n,i}\}$, if $X_{n,i}$ is $\mathscr{F}_{n,i}$ measurable and $\mathsf{E}(X_{n,i}|\mathscr{F}_{n,i-1}) = 0$ a.s. for all n and $i \geqq 1$.

The theorem of MCLEISH [4] is the following:

THEOREM 3. *Let* $\{X_{n,i}\}$ *be a martingale difference array satisfying*

(a) $\max_{0\leqq i\leqq l_n} |X_{n,i}|$ *is uniformly bounded in* L_2 *norm,*

(b) $\max_{0\leqq i\leqq l_n} |X_{n,i}| \xrightarrow{p} 0$

(c) $\sum_{i=0}^{l_n} X_{n,i}^2 \xrightarrow{p} 1.$

Then $S_n \xrightarrow[w]{} N(0,1)$ (*where* $\xrightarrow[p]{}$ *and* $\xrightarrow[w]{}$ *denote convergence in probability and weak convergence* (*convergence in distribution*)).

In what follows we consider the Walsh system as a sequence of random variables on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ where $\Omega=[0,1]$, $\mathcal{F}$ if the σ-field of all Borel-measurable sets on Ω and P is the Lebesgue measure on $\mathcal{F}$. We assume further that $\{n_k\}$ and $\{a_k\}$ satisfy the conditions of Theorem 1A.

Following [3] let us put

$$\begin{cases} p(0)=0, \quad p(k)=\max\{m, n_m < 2^k\} & (k=1,2,\ldots), \\ \Delta_k(x)=\sum_{m=p(k)+1}^{p(k+1)} a_m w_{n_m}(x) \quad \text{and} \quad B_k = A_{p(k+1)} & (k=0,1,2,\ldots). \end{cases} \tag{8}$$

With an easy computation ([3] formula (2.2)) we have

$$p(k+1)-p(k)=O\big(p^{\alpha}(k)\big) \quad \text{as} \quad k\to\infty \tag{9}$$

and thus (6) implies

$$\sum_{m=p(k)+1}^{p(k+1)} |a_m| \leqq \max_{p(k)<m\leqq p(k+1)} |a_m|\big(p(k+1)-p(k)\big)= \tag{10}$$

$$= o\left(\frac{B_k}{p^{\alpha}(k)}\right) O\big(p^{\alpha}(k)\big) = o(B_k) \quad \text{as} \quad k\to\infty.$$

Further on we need

LEMMA 1. *Under the conditions of Theorem* 1*A we have*

$$\mathsf{E}\left|B_k^{-2}\sum_{i=0}^{k}\Delta_i^2(x)-1\right|^2 = o(1) \quad \textit{as} \quad k\to\infty. \tag{11}$$

The proof of this lemma is the same as that of Lemma 3 in [3].

PROOF of Theorem 1A. Suppose that

$$p(k) < N \leqq p(k+1). \tag{12}$$

Then

$$Z_N(x)=\frac{\sum_{m=1}^{N} a_m w_{n_m}(x)}{A_N}= \tag{13}$$

$$= \frac{B_{k-1}}{A_N}\,\frac{\sum_{i=0}^{k-1}\Delta_i(x)}{B_{k-1}} + \frac{\sum_{m=p(k)+1}^{N} a_m w_{n_m}(x)}{A_N}.$$

Put

$$X_{k,i}=\frac{\Delta_i(x)}{B_k} \qquad \begin{matrix} k=0,1,2\ldots \\ i=0,1,2,\ldots k. \end{matrix} \tag{14}$$

$$S_k=\sum_{i=0}^{k} X_{k,i}. \tag{15}$$

With these notations

$$Z_N(x) = \frac{B_{k-1}}{A_N} S_{k-1} + \frac{\sum_{m=p(k)+1}^{N} a_m w_{n_m}(x)}{A_N}. \tag{16}$$

Therefore to prove our theorem it is enough to show that

(i) $$S_k \xrightarrow{w} N(0,1),$$

(ii) $$\frac{B_{k-1}}{A_N} \to 1,$$

(iii) $$\frac{\sum_{m=p(k)+1}^{N} a_m w_{n_m}(x)}{A_N} \to 0 \quad \text{a.s.}$$

Let $\mathscr{F}_{k,i}$ be the sub-sigma field of $\mathscr{F}$ generated by the random variables $\{w_{2^n}(x): 0 \leqq n \leqq i\}$. In this case clearly $X_{k,i}$ is $\mathscr{F}_{k,i}$ measurable, $\mathscr{F}_{k,i-1} \subset \mathscr{F}_{k,i}$ and $\mathsf{E}(X_{k,i}|\mathscr{F}_{k,i-1}) = 0$ a.s. for all k and i, that is $\{X_{k,i}\}$ is a martingale difference array.

According to Theorem 3, for the proof of (i) we have to show that the above defined $\{X_{k,i}\}$ satisfy conditions (a), (b) and (c) of Theorem 3.

Using the orthogonality of the Walsh functions we get

$$\mathsf{E}\left(\max_{0 \leqq i \leqq k} |X_{k,i}|\right)^2 = \frac{\mathsf{E}\left(\max_{0 \leqq i \leqq k} |\Delta_i(x)|\right)^2}{B_k^2} \leqq$$

$$\leqq \frac{\mathsf{E}\left(\sum_{0 \leqq i \leqq k} \Delta_i^2(x)\right)}{B_k^2} = \frac{\sum_{0 \leqq i \leqq k} \mathsf{E}\left(\Delta_i^2(x)\right)}{B_k^2} = \tag{17}$$

$$= \frac{\sum_{0 \leqq i \leqq k} \sum_{m=p(i)+1}^{p(i+1)} a_m^2}{B_k^2} = \frac{B_k^2}{B_k^2} = 1$$

which means that condition (a) is satisfied.

To see (b) consider

$$\max_{0 \leqq i \leqq k} |X_{k,i}| = \max_{0 \leqq i \leqq k} \frac{|\Delta_i(x)|}{B_k} \leqq$$

$$\leqq \frac{\max_{0 \leqq i \leqq k} \sup_x |\Delta_i(x)|}{B_k} \leqq \frac{\max_{0 \leqq i \leqq k} \sum_{m=p(i)+1}^{p(i+1)} |a_m|}{B_k}.$$

Using (10) we get

$$\frac{\max_{0 \leqq i \leqq k} \sum_{m=p(i)+1}^{p(i+1)} |a_m|}{B_k} = \frac{\max_{0 \leqq i \leqq k} o(B_i)}{B_k} = o(1) \tag{18}$$

which proves (b).

Moreover from Lemma 1 trivially follows that the $\{X_{k,i}\}$ satisfy condition (c) of Theorem 3. Thus we proved (i).

To see (ii) it is sufficient to prove

$$\frac{B_{k-1}}{B_k} \to 1. \tag{19}$$

$$\left(\text{Clearly } 1 \geqq \frac{B_{k-1}}{A_N} \geqq \frac{B_{k-1}}{B_k} \geqq 0.\right)$$

We show that

$$1-\frac{B_{k-1}^2}{B_k^2} \to 0 \tag{20}$$

$$1-\frac{B_{k-1}^2}{B_k^2} = \frac{\sum_{m=p(k)+1}^{p(k+1)} a_m^2}{B_k^2} \leqq \frac{\max_{p(k)<m\leqq p(k+1)} |a_m|^2\{p(k+1)-p(k)\}}{B_k^2} =$$

$$= \frac{1}{B_k^2}\, o\left(\frac{B_k^2 p^\alpha(k)}{p^{2\alpha}(k)}\right) = o\left(\frac{1}{p^\alpha(k)}\right)$$

by (6) and (9) which proves (20) and (ii) follows.

Finally,

$$\frac{\left|\sum_{m=p(k)+1}^{N} a_m w_{n_m}(x)\right|}{A_N} \leqq \frac{\sum_{m=p(k)+1}^{p(k+1)} |a_m|}{B_{k-1}} =$$

$$= \frac{1}{B_{k-1}}\, o(B_k) = o(1) \quad \text{a.s.}$$

by (19) and (10) which proves (iii), and thus we proved our Theorem 1A.

Remark 1. To investigate the critical case $\alpha=\frac{1}{2}$ too, P. ERDŐS introduced (in the case of trigonometric system) the following gap condition

$$\frac{n_{k+1}}{n_k} > 1+\frac{c_k}{k^\alpha} \quad \text{for} \quad \alpha \leqq \frac{1}{2} \quad \text{where} \quad \lim_{k\to\infty} c_k = +\infty. \tag{21}$$

We remark that using this gap condition instead of (9) we have

$$p(k+1)-p(k) = o\big(p^\alpha(k)\big) \tag{22}$$

as

$$2 > \frac{n_{p(k+1)}}{n_{p(k)+1}} > \prod_{m=p(k)+1}^{p(k+1)-1} (1+c_m m^{-\alpha}) >$$

$$> 1+c_{p(k)}\big(p(k+1)-p(k)-1\big)\cdot p^{-\alpha}(k+1)$$

which implies (22).

With the method of the proof of Theorem 1A it is easy to prove using (22) the following

THEOREM 1B. *Under the gap condition (21), if*

$$A_N = \left(\sum_{k=1}^{N} a_k^2\right)^{1/2} \to \infty \quad \text{and} \quad a_N = O\left(\frac{A_N}{N^\alpha}\right) \quad \text{as} \quad N \to \infty \quad \text{for} \quad \alpha \leqq \frac{1}{2}$$

we have

$$\lim_{N\to\infty} \mathrm{P}\left(\frac{\sum_{k=1}^{N} a_k w_{n_k}(x)}{A_N} < u\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{t^2}{2}} dt.$$

Acknowledgement. I am indebted to I. BERKES for his valuable comments.

REFERENCES

[1] RÉVÉSZ, P. and WSCHEBOR, M.: On the statistical properties of the Walsh functions, *Publications of the Mathematical Institute of the Hung. Acad. Sci.* Vol. IX. Ser. A. Fasc. 3. (1964) 543—553.

[2] FÖLDES, A.: Further statistical properties of the Walsh functions, *Studia Sci. Math. Hung.* **7** (1972) 147—153.

[3] TAKAHASHI, S.: A statistical property of Walsh functions, *Studia Sci. Math. Hung.* **10** (1975) 93—98.

[4] MCLEISH, D. L.: Dependent central limit theorems and invariance principles, *The Annals of Probability*, **2**, No. 4, 1974. 620—628.

[5] P. ERDŐS: On trigonometric sums with gaps, *Magyar Tud. Akad. Mat. Kut. Int. Közl.* **7** (1962) 37—42.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received January 27, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 147—156.

MEAN OF OUTSTANDING ELEMENTS

by

A. NAGABHUSHANAM*, M. L. AGGARWAL**, H. C. GUPTA**

§ 1. Introduction

The problem of record values was motivated by the frequency with which the record weather conditions are reported in the newspapers. We define an observation to be an upper (lower) record if it is greater (lower) than all the preceding observations. CHANDLER [2] has obtained, for a semi-infinite series of independent observations from the same population, the probability distribution of the serial number of the rth lower record; he also gives the distribution of the record value itself. FOSTER and STUART [5] have obtained the joint distribution of the number of upper and lower records. DAVID and BARTON [4] have given the joint distribution of the epochs of occurrence of the records. Reference may also be made to a review paper by BARTON and MALLOWS [1] for various combinatorial aspects involving records, amalgamation and Simon Newcomb problems. For a mathematical formulation of the problem we introduce the following notations.

Let $x_1, x_2, \ldots$ be a sequence of independent observations from a continuous distribution (of the random variable X) with the density function $f(\cdot)$ and distribution function $F(\cdot)$. With $N_0 = N_0' = 1$, we recursively define

$$N_q = \text{Min}\{j | j > N_{q-1}, x_j > x_{N_{q-1}}\}$$

and

$$N_q' = \text{Min}\{j | j > N_{q-1}', x_j < x_{N_{q-1}'}\}.$$

$u_q \equiv x_{N_q}$ and $l_q \equiv x_{N_q'}$ are called the qth upper and lower records respectively occurring at epochs N_q, N_q' in the original sequence of observations.

Since $x_i < x_j \Rightarrow -\log[F(x_1)/F(x_i)] < -\log[F(x_1)/F(x_j)]$ and for $i < j$, $-\log[F(l_i)] < -\log[F(l_j)]$, we see that

$$z_j = \log[F(x_1)/F(l_j)], \tag{1}$$

is the jth upper record value in the sequence

$$\{\log F(x_1) - \log F(x_i)\}, \quad 1 \leqq i \leqq \infty.$$

Similarly

$$t_j = \log[1 - F(x_1)/1 - F(u_j)]$$

is the jth upper record value in the sequence

$$\{\log(1 - F(x_1)) - \log(1 - F(x_i))\}, \quad 1 \leqq i < \infty.$$

* Department of Mathematics, Indian Institute of Technology, Hauz Khas, New Delhi-29.

** Department of Mathematical Statistics, University of Delhi, Delhi-110007 (India).

The corresponding values z_j and t_j are called outstanding values in the sequel.

In this paper we derive the distribution of the mean of r outstanding values, namely the first r upper records in the sequence $\{\log F(x_1)-\log F(x_i)\}$ and that of the first s upper records in the sequence $\{\log(1-F(x_1))-\log(1-F(x_i))\}$. It is found that these variables are independent and we derive the distribution of their ratio. We observe that the limiting distribution of the ratio $r(r+1)(a_s+b_sT)/2z$ is standardized normal. Another problem treated in this paper is the distribution of the position of $\bar{z}$ amongst $z_1, \ldots, z_r$. Defining the random variable Y as the number of z's that exceed $\bar{z}$ we see that $\mathsf{P}\{Y \leqq r-i\} = \mathsf{P}\{\bar{Z} > Z_i\}$. It is interesting to note [1] that this is the distribution of "one less than the sum of $r-i$ independently distributed random variables, each having uniform distribution in (0,1)". Thus the moments of this distribution are readily available. This distribution also arises in the Simon-Newcomb problem in the theory of Non-parametric testing. A similar problem is the position amongst independent observations from a continuous population of their mean and is studied by many authors. In particular, when observations are taken from $N(\mu, \sigma)$ the probability that $\bar{x}$ lies between $x_{n-1:n}$ and $x_{n:n}$ known as Youden's Demon problem has been determined by KENDALL [6]. Another particular case namely when the observations are coming from an exponential population was taken up by SARKADI, VINCZE and SCHNELL [7].

§ 2. Joint distribution of X, Z, T

Using the vector notation, it is easily seen that if x be the first observation, then

$$dP(x, \mathbf{l}, \mathbf{u}) = dP(x, l_1, \ldots, l_r, u_1, \ldots, u_s) = \tag{2}$$

$$= f(x)\,dx \prod_{i=1}^{r} \prod_{j=1}^{s} \frac{dl_i du_j f(l_i) f(u_j)}{F(l_{i-1})\{1-F(u_{j-1})\}},$$

where $F(l_0) = F(x) = F(u_0)$.

We observe that $\mathbf{l}$, $\mathbf{u}$ are not independent; they are related through the initial observation x.

Now the Jacobian of the transformation from $(x, \mathbf{l}, \mathbf{u})$ to $(x, \mathbf{z}, \mathbf{t})$ given by (1) and (2) is

$$\frac{\partial(x, \mathbf{z}, \mathbf{t})}{\partial(x, \mathbf{l}, \mathbf{u})} = \prod_{i=1}^{r} \prod_{j=1}^{s} \frac{f(l_i) f(u_j)}{F(l_i)\{1-F(u_j)\}}$$

and

$$\prod_{i,j} dl_i\, du_j = \prod_{i,j} dz_i\, dt_j \left| \frac{\partial(x, \mathbf{z}, \mathbf{t})}{\partial(x, \mathbf{l}, \mathbf{u})} \right|.$$

The region for which $dP(x, \mathbf{l}, \mathbf{u})$ is non-zero is given by

$$\infty > u_r > u_{r-1} > \ldots > u_1 > x_1 > l_1 > \ldots > l_s > -\infty;$$

and thus the region for which $dP(x, \mathbf{z}, \mathbf{t})$ is non-zero is

$$-\infty < x < \infty, \quad 0 < z_1 < \ldots < z_r < \infty, \quad 0 < t_1 < \ldots < t_s < \infty.$$

Hence from (2)

$$dP(x, \mathbf{z}, \mathbf{t}) = dP(x, \mathbf{l}, \mathbf{u}) / \frac{\partial(x, \mathbf{z}, \mathbf{t})}{\partial(x, \mathbf{l}, \mathbf{u})} = \tag{3}$$

$$= f(x) \exp(-z_r - t_s)\, dx \prod_{i=1}^{r} dz_i \prod_{j=1}^{s} dt_j,$$

$$-\infty < x < \infty, \quad 0 < z_1 < \ldots < z_r < \infty, \quad 0 < t_1 < \ldots < t_s < \infty.$$

Since the joint density function of x, $\mathbf{z}$, $\mathbf{t}$ is the product of their marginal densities for the entire ranges (3) of the variables and these ranges are independent, it follows that X and the random vectors $\mathbf{Z}$, $\mathbf{T}$ are distributed independently of one another, and

$$dP(x) = f(x)\, dx$$

$$dP(\mathbf{z}) = \exp(-z_r)\, dz_1\, dz_2 \ldots dz_r, \quad 0 < z_1 < z_2 \ldots < z_r < \infty \tag{4}$$

$$dP(\mathbf{t}) = \exp(-t_s)\, dt_1\, dt_2 \ldots dt_s, \quad 0 < t_1 < t_2 \ldots < t_s < \infty \tag{5}$$

Distribution of $Z = Z_1 + \ldots + Z_r$

From (4),

$$C(\theta) = \mathsf{E}\{e^{i\theta z}\} \tag{6}$$

$$= \int_0^\infty \int_{z_1}^\infty \ldots \int_{z_{r-1}}^\infty \exp[i\theta(z_1 + \ldots z_r) - z_r]\, dz_1 \ldots dz_r$$

$$= [(1 - i\theta)(1 - 2i\theta)\ldots(1 - ri\theta)]^{-1}.$$

Hence by inversion

$$dP(z) = \frac{dz}{2\pi} \int C(\theta) \cdot e^{-i\theta z}\, d\theta = \tag{7}$$

$$= \sum_{j=1}^{r} \frac{(-1)^{r-j} e^{-z/j}\, j^{r-2}}{(j-1)!(r-j)!}\, dz,$$

expressing $C(\theta)$ by partial fractions and integrating. Similarly

$$dP(t) = \sum_{k=1}^{s} \frac{(-1)^{s-k} e^{-t/k} k^{s-2}}{(k-1)!(s-k)!}\, dt. \tag{8}$$

Distribution of T/Z

From (4) and (5), it is evident that Z and T are independent. Hence by (7) and (8), the joint probability density of Z, T is

$$\sum_{j=1}^{r} \sum_{k=1}^{s} \binom{r-1}{j-1} \binom{s-1}{k-1} \frac{j^{s-2} k^{s-2} (-1)^{r+s-j-k}}{\Gamma(r)\Gamma(s)} \exp\left(\frac{-z}{j} - \frac{t}{k}\right). \tag{9}$$

To find the distribution of T/Z we carry out the transformation

$$z = z, \; v = t/z.$$

The joint probability density of Z, V is z times (9) expressed in terms of z, v. This on integration over z from 0 to ∞ gives for the p.d.f. of V

$$(10)\qquad \frac{1}{(r-1)!(s-1)!}\sum_{j=1}^{r}\sum_{k=1}^{s}\binom{r-1}{j-1}\binom{s-1}{k-1}\frac{(-1)^{r+s-j-k}j^{r-2}k^{s-2}}{(1/j+v/k)^2},\quad 0<v<\infty.$$

Moments. From (4) we have

$$\mathsf{E}\left(\prod_{i=1}^{r}z_i^{\alpha_i}\right)=\int_0^\infty\int_{z_1}^\infty\dots\int_{z_{r-1}}^\infty e^{-z_r}z_1^{\alpha_1}\dots z_r^{\alpha_r}\,dz_1\dots dz_r,$$

the region of integration being indicated by

$$(11)\qquad \infty=z_{r+1}\geqq z_r\geqq\dots\geqq z_1\geqq z_0=0,$$

starting from z_1 lying between 0 and ∞. On reversing the order, this region is the same as that indicated by (11), starting from z_r lying between 0 and ∞. Hence

$$(12)\qquad \mathsf{E}\left(\prod_{i=1}^{r}z_i^{\alpha_i}\right)=\int_0^\infty\int_0^{z_r}\dots\int_0^{z_2}e^{-z_r}\left(\prod_{i=1}^{r}z_i^{\alpha_i}\right)dz_r\dots dz_1=$$

$$=\frac{\displaystyle\int_0^\infty e^{-z_r}z_r^{\alpha_1+\dots+\alpha_r+r-1}dz_r}{(\alpha_1+1)(\alpha_1+\alpha_2+2)\dots(\alpha_1+\dots+\alpha_{r-1}+r-1)}=$$

$$=\frac{\Gamma(\alpha_1+\alpha_2+\dots+\alpha_r+r)}{\displaystyle\prod_{i=1}^{r-1}(\alpha_1+\dots+\alpha_i+i)}.$$

Putting $\alpha_j=0$ for $j\neq i$ in (12) we get

$$\mathsf{E}(z_i^{\alpha_i})=\Gamma(r+\alpha_i)/(i-1)!(\alpha_i+i)\dots(\alpha_i+r-1)=$$
$$=\Gamma(i+\alpha_i)/(i-1)!$$

In particular, for a natural number α_i,

$$\mathsf{E}(z_i^{\alpha_i})=i(i+1)\dots(i+\alpha_i-1),\quad \alpha_i=1,2,\dots$$

so that

$$\mu_r(z_i)=\mathsf{E}(z_i-i)^r=\sum_{j=1}^{r}\binom{r}{j}(-i)^{r+j}\frac{\Gamma(i+j)}{\Gamma(i)}.$$

These give

$$\mathrm{Var}\,(Z_i)=i(i+1)-i^2=i,$$

$$\mu_3(Z_i)=i(i+1)(i+2)-3i^2(i+1)+2i^3=2i$$

$$\mu_4(Z_i)=i(i+1)(i+2)(i+3)-4i^2(i+1)(i+2)+6i^3(i+1)-3i^4=3i^2+6i.$$

Putting $\alpha_j=0$ for $j\neq k, m$ in (12) we get for $k<m$,

$$\mathsf{E}(Z_k^{\alpha_k} Z_m^{\alpha_m}) = \Gamma(\alpha_k+\alpha_m+r)/(k-1)! \prod_{j=k}^{m-1}(j+\alpha_k) \prod_{j=m}^{r-1}(j+\alpha_k+\alpha_m) =$$

$$= \Gamma(\alpha_k+\alpha_m+m)/(k-1)! \prod_{j=k}^{m-1}(j+\alpha_k).$$

In particular, for natural numbers p, q it is

$$\mathsf{E}(z_k^p z_m^q) = k(k+1)\dots(k+p-1)(m+p)\dots(m+p+q-1).$$

Thus

$$\mathrm{Cov}(z_k^p, z_m^q) = k(k+1)\dots(k+p-1)[(m+p)\dots(m+p+q-1) -$$

$$-m(m+1)\dots(m+q-1)],$$

$$\mathrm{Cov}(Z_k, Z_m^q) = k[(m+1)\dots(m+q-1)(m+q) - m(m+1)\dots(m+q-1)]$$

and

$$\mathrm{Cov}(Z_k, Z_m) = k, \quad \varrho(Z_k, Z_m) = \sqrt{k/m}.$$

Now

$$\mathsf{E}(Z) = \sum_i \mathsf{E}(Z_i) = \sum_i i = \frac{1}{2} r(r+1) = -a_r b_r^{-1}, \tag{13}$$

say

$$\mathrm{Var}(Z) = \sum \mathrm{Var}(Z_i) + 2 \sum_{j=2}^{r} \sum_{i=1}^{j-1} [\mathrm{Cov}(Z_i, Z_j)] = \tag{14}$$

$$= \sum_i i + 2 \sum \sum i = \sum i + 2 \cdot 2^{-1} \sum j(j-1)$$

$$= \sum_{i=1}^{r} i^2 = \frac{1}{6} r(r+1)(2r+1) = b_r^{-2} \text{ say.}$$

Also

$$\mathrm{Cov}(Z, Z_i) = \mathsf{E}(ZZ_i) - \mathsf{E}(Z)\mathsf{E}(Z_i) = \sum_{j=1}^{i} j + ir - \frac{1}{2} ir(r+1)$$

$$= \frac{1}{2} \cdot i(1+r-r^2+i).$$

Corresponding results for T, T_j are obtained on replacing i by j and r by s.

Asymptotic Distributions

From (6),

$$C_r(\theta) = \mathsf{E}[\exp\{i\theta(a_r+b_r Z)\}] =$$

$$= e^{i\theta a_r} \prod_{k=1}^{r} (1-ki\theta b_r)^{-1}$$

or

$$\log C_r(\theta) = i\theta \left\{ a_r + b_r \sum_{k=1}^{r} k \right\} - \frac{1}{2} b_r^2 \theta^2 \sum_{k=1}^{r} k^2 - \frac{1}{6} b_r^3 \theta^3 i(1+\varepsilon) \sum_{k=1}^{r} k^3,$$

where ε is bounded for small θ.

Hence from (13) and (14) for large r,

$$\log C_r(\theta) = -\frac{1}{2}\theta^2 + O(r^{-1/2}), \tag{15}$$

which shows that $a_r + b_r Z$ is asymptoticallay normal. The same result holds for $a_s + b_s T$.

We now make use of the following result [3]: If for finite a, finite non-zero b and finite positive h, k

$$x_n \sim N(a, h/\sqrt{n}), \quad Y_n \sim N(b, k/\sqrt{n}),$$

then for large n

$$b(x_n - a)/\frac{h}{\sqrt{n}}. \quad Y_n \sim N(0, 1),$$

irrespective of whether x_n and Y_n are independent or not.

$$\text{Take } X_n = \frac{a_s + b_s T}{\sqrt{n}}, \quad Y_n = \frac{b_r z}{a_r}.$$

These satisfy the above conditions with $a=0$, $h=1$, $b=-1$ and $k=\sqrt{C}$ if $n=s==cr$; and we conclude that

$$V_{r,s} = -\frac{a_s + b_s T}{\sqrt{n}} \Big/ \frac{1}{\sqrt{n}} \frac{b_r Z}{a_r} = \tag{16}$$

$$= \frac{1}{2} r(r+1)(a_s + b_s T)/Z$$

is asymptotically standardized normal.

§. 3 Distribution of position of $\bar{z}$

We shall now determine the probability $\mathsf{P}(\bar{z} > z_i)$ that the mean of r outstanding values exceeds a specified outstanding value and establish that this gives the position of z amongst z_j, $j=1, \ldots, r$.

The distribution of $\mathbf{z}=(z_1, \ldots, z_r)$ is given by (4) and we evaluate the integral of (4) over the region $\bar{z} \leqq z_i$. This on defining w_k by

$$z_1 + z_2 + \ldots + z_r + (r-k) w_k = r z_i \tag{17}$$

is

$$\mathsf{P}(\bar{z} \leqq z_i) = \int_0^{\infty} dz_i \int_0^{z_i} dz_1 \int_{z_1}^{z_i} dz_2 \ldots \int_{z_{i-2}}^{z_i} dz_{i-1} \tag{18}$$

$$\int_{z_i}^{w_i} dz_{i+1} \ldots \int_{z_{r-2}}^{w_{r-2}} dz_{r-1} \int_{z_{r-1}}^{w_{r-1}} \exp(-z_r)\, dz_r.$$

The upper limit w_k for z_{k+1} for $r>k\geqq i$ follows from the considerations that for z_{k+1} to be maximum

$$z_{k+i} = z_{k+1} \quad \text{for} \quad i = 2, 3, \ldots, r-k \quad \text{and} \quad \sum_{j=1}^{r} z_j \leqq r z_i.$$

The w_j's defined by (17) satisfy the recurrence relation

$$(j+1)w_{r-j-1} = jw_{r-j} + z_{r-j}, \quad j = 1, 2, \ldots, r-i-1. \tag{19}$$

Now writing $G(x)=e^{-x}$ and using (19)

$$I_{r-1} = \int_{z_{r-1}}^{w_{r-1}} e^{-z_r}\, dz_r = G(z_{r-1}) - G(2w_{r-2} - z_{r-1})$$

$$I_{r-2} = \int_{z_{r-2}}^{w_{r-2}} I_{r-1}\, dz_{r-1} =$$

$$= G(z_{r-2}) - G(w_{r-2}) + G(2w_{r-2} - z_{r-2}) - G(w_{r-2}) =$$

$$= G(z_{r-2}) - 2G\left(\frac{3}{2} w_{r-3} - \frac{1}{2} z_{r-2}\right) + G(3w_{r-3} - 2z_{r-2}).$$

In general for $i<j<r$ it can be shown by induction that

$$I_j = \int_{z_j}^{w_j} I_{j+1}\, dz_{j+1} = \tag{20}$$

$$= G(z_j) + \sum_{k=1}^{r-j} \frac{(-1)^{r-j-k-1}}{(r-j-1)!} \binom{r-j-1}{k-1} k^{r-j-1}$$

$$G[k^{-1}\{(r-j+1)w_{j-1} - (r-j-k+1)z_j\}].$$

In particular,

$$I_{i+1} = G(z_{i+1}) + \frac{1}{(r-i-2)!} \sum_{k=1}^{r-i-1} (-1)^{r-i-k} \binom{r-i-2}{k-1} k^{r-i-2} \times$$

$$\times G[k^{-1}\{(r-i)w_i - (r-i-k)z_{i+1}\}].$$

Now

$$I_i = \int_{z_i}^{w_i} I_{i+1}\, dz_{i+1} = \tag{21}$$

$$= G(z_i) - G(w_i) + \frac{1}{(r-i-1)!} \sum_{k=1}^{r-i-1} (-1)^{r-i-k} \binom{r-i-1}{k-1} k^{r-i-1} \times$$

$$\times \{G(w_i) - G[k^{-1}\{(k+i)z_i - z_1 . \{\ldots - z_i\}\}]$$

Let $x^d = \sum_{k=0}^{d} B_k^{(d)} x^{(k)}$, where $B_k^{(d)}$ are Stirling numbers of the second kind and in particular $B_d^{(d)} = 1$. Then

$$B_c^{(a)} \cdot c! = (E-1)^c 0^d = \sum_{k=1}^{c} \binom{c}{k} (-1)^{c-k} k^d,$$

so that on putting $c=d=r-i$,

$$1=\frac{1}{(r-i)!}\binom{r-i}{k}(-1)^{r-i-k}k^{r-i}$$

$$=\frac{1}{(r-i-1)!}\sum_{k=1}^{r-i-1}(-1)^{r-i-k}\binom{r-i-1}{k-1}k^{r-i-1}+\frac{(r-i)^{r-i-1}}{(r-i-1)!}.$$

Hence considering the sums over the two G-functions in (21) separately,

$$I_i=G(z_i)+$$

$$+\frac{1}{(r-i-1)!}\sum_{k=1}^{r-i}(-1)^{r-i-k+1}\binom{r-i-1}{k-1}k^{r-i-1}G\{k^{-1}((k+i)z_i-z_1-\ldots-z_i)\},$$

showing in virtue of (17) that (20) holds also for $j=i$. For $0<j<i$, let

$$I_{i-j}=\frac{k^{j-1}}{(j-1)!}\sum_{l=0}^{j-1}\binom{j-1}{l}(-1)^l G\{k^{-1}((i+k+l-j)z_i-z_1-\ldots z_{i-j}-lz_{i_j})\}.$$

Then for $0<j<i-1$,

$$\int_{z_{i-j-1}}^{z_i} I_{i-j}\,dz_{i-j}=\frac{k^j}{j!}\sum_{l=-1}^{j-1}\binom{j}{l+1}(-1)^l[G\{k^{-1}((i+k-j-1)z_i-$$

$$-z_1-\ldots-z_{i-j-1})\}-G\{k^{-1}(i+k+l-j)z_i-z_1-\ldots(l+1)z_{i-j-1})\}],$$

the additional value $l=-1$ only introducing a vanishing term. Further the argument of the first $G(\ )$ function is independent of l and its coefficients add up to $-(1-1)^j=0$.

Hence on writing $l-1$ for l,

$$\int_{z_{i-j-1}}^{z_i} I_{i-j}\,dz_{i-j}=$$

$$=\frac{k^j}{j!}\sum_{l=0}^{j}\binom{j}{l}(-1)^l G\{k^{-1}((i+k+l-j-1)z_i-z_1-\ldots-z_{i-j-1}-lz_{i-j-1})\}=$$

$$=I_{i-j-1},\quad 0<j<i-1.$$

In particular, we have

$$I_1=\frac{k^{i-2}}{(i-2)!}\sum_{m=0}^{i-2}\binom{i-2}{m}(-1)^m G\{k^{-1}[(k+m+1)z_i-(m+1)z_1]\},$$

$$\int_0^{z_i} I_1\,dz_1=\frac{k^{i-1}}{(i-2)!}\sum_{m=0}^{i-2}\binom{i-2}{m}\frac{(-1)^m}{(m+1)}\left[G(z_i)-G\left(\frac{k+m+1}{k}z_i\right)\right]=$$

$$=\frac{k^{i-1}}{(i-1)!}\sum_{m=0}^{i-1}\binom{i-1}{m}(-1)^m G\left(\frac{k+m}{k}z_i\right).$$

Lastly

$$\int_0^{\infty} dz_i\int_0^{z_i} I_1\,dz_1=\frac{k^i}{(i-1)!}\sum_{m=0}^{i-1}\binom{i-1}{m}\frac{(-1)^m}{k+m}. \tag{22}$$

Also

$$\int_0^{\infty}\int_0^{z_i}\int_0^{z_i}\dots\int_{z_{i-2}}^{z_i} G(z_i)\,dz_i\,dz_1\dots dz_{i-1} = \tag{23}$$

$$= \int_0^{\infty}\int_0^{z_i}\frac{(z_i-z_1)^{i-2}}{(i-2)!}e^{-z_i}\,dz_i\,dz_1 = \int_0^{\infty}\frac{z_i^{i-1}e^{-z_i}}{(i-1)!}\,dz_i = 1$$

Thus using (21) in (18) and simplifying in view of (23) and (22) we get

$$\mathsf{P}(\bar{Z} > z_i) = 1-\mathsf{P}(\bar{Z} \leqq z_i) =$$

$$= \frac{1}{(r-i-1)!}\sum_{k=1}^{r-i}\sum_{m=0}^{i-1}\frac{k^{r-1}}{(i-1)!}\binom{i-1}{m}\frac{(-1)^{r-i-k+m}}{k+m}\binom{r-i-1}{k-1} =$$

$$= \frac{(-1)^{r-i}}{(r-1-i)!(i-1)!}\sum_{k=1}^{r-i}\frac{(-1)^k k^{r-2}\binom{r-i-1}{k-1}}{\binom{i+k-1}{k}} =$$

$$= \frac{(-1)^{r-i}}{(r-1)!}\sum_{k=1}^{r-i}(-1)^k k^{r-1}\binom{r-1}{i+k-1}.$$

With $i=r-i-k$, we get

$$\mathsf{P}(\bar{Z} > z_i) = \frac{1}{(r-1)!}\sum_{j=0}^{r-i-1}(-1)^j\binom{r-1}{j}(r-i-j)^{r-1}. \tag{24}$$

Defining Y to be the number of z's that exceed $\bar{z}$ we observe that

$$\mathsf{P}(\bar{z} > z_i) = \mathsf{P}(Y \leqq r-i) = H(r-i), \quad \text{say,}$$

so that (24) gives the distribution function of the discrete valued random variable Y. We may write

$$\mathsf{P}(z_i < \bar{Z} < Z_{i+1}) = \mathsf{P}(y = r-i) = H(r-i)-H(r-i-1).$$

It is interesting to note that this distribution also arises in the Simon-Newcomb problem as the distribution of the number of strings that do not exceed $m=r-i-1$ in a given sequence $x_1, x_2, \dots, x_n$.

*

Acknowledgements. The authors are grateful to DR. KANWAR SEN and Prof. ENDRE CSÁKI for their valuable suggestions.

REFERENCES

[1] BARTON, D. E. and MALLOWS, C. L.: Some aspects of the random sequence, *Ann. Math. Stat.*, **36** (1965) 236—260.
[2] CHANDLER, K. N.: The distribution and frequency of record values, *J. R. Statist. Soc.*, B, **14** (1952), 220—228.
[3] CRAMER, H.: *Mathematical Methods of Statistics,* Princeton University Press, 259. 1946.
[4] DAVID, F. N. and BARTON, D. E.: *Combinatorial Chanel,* London: Griffin. 1962.
[5] FOSTER, F. G. and STUART, A.: Distribution free tests in time-series based on the breaking of records, *J. R. Statist. Soc.*, B. **16** (1954) 1—13.
[6] KENDALL, M. G.: Two problems in sets of measurements, *Biometrika* **41** (1954) 560—564.
[7] SARKADI, K., SCHNELL, E., VINCZE, I.: On the position of the sample mean among the ordered elements. *Publ. of the Math. Inst. of the Hung. Acad. of Sciences* **7** (1962) 239—254.

Dept. of Math. Stat., Univ. of Delhi, India
(Received January 10, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 157—170.

THE THINNEST THREE DIMENSIONAL POINT LATTICE TRAPPING A SPHERE*

by

M. N. BLEICHER

I. Introduction. The history of Discrete Geometry and the Geometry of Numbers are full of examples of the study of lattices which are extremal with respect to some covering or packing property. We give several examples:

Example 1. The densest lattice packing of unit spheres.

The problem is to find the lattice of greatest density (least volume of the fundamental domain) in which no two points are closer than 2. The two dimensional case was first solved by LAGRANGE [27] in 1773 in his work on the reduction of binary quadratic forms, but it waited for GAUSS [21] in 1831 to point out the geometric significance of LAGRANGE work. In the same paper GAUSS solved the 3-dimensional case. A number of workers, see ROGERS [29, p. 3] or BAMBAH [3] for more details, have attacked the problem in higher dimensions. The problem is now completely solved for all the dimensions through 8.

Example 2. The thinnest lattice covering by unit spheres.

The problem is to find the least dense lattice (maximal volume of the fundamental domain) such that no point in space is at a distance greater than 1 from some lattice point. The two dimensional case seems to have been solved first by KERSHNER [26] in 1939. The three dimensional case was first solved by BAMBAH in 1954. A number of people including BAMBAH, BARNES, BLEICHER, COXETER, DAVENPORT, DELONE, DICKSON, ERDŐS, FEW, GAMECKII, ROGERS, AND RYSKOFF [1—14, 17—21, 26, 29, 30], have simplified the proofs and worked on the problem in higher dimensions. The problem is now solved for dimensions 2, 3 and 4.

It is of interest to note that L. FEJES TÓTH [15] has shown that the ellipse is the least efficient symmetric planar convex body for lattice covering.

Example 3. The "Minimal Visibility" lattice packing of spheres.

In 1960—61 HEPPES [23] showed that in any lattice packing of spheres there are points from which one can look in certain directions without ever having one's view blocked. This work was extended by HORTOBÁGYI [24] when he proved that in any lattice packing of unit spheres one can pack an infinitely long cyclinder of radius $\frac{3\sqrt{2}}{4}-1=0.606\ldots$. HORVÁTH [25], has extended these results to higher dimensions.

* This research was initiated by a cooperative research project between American and Hungarian mathematicians sponsored by the National Science Foundation of the USA and the Institute for Cultural Relations in Hungary.

Example 4. The thinnest non-separable lattice of unit spheres.

A lattice of unit sphere (a lattice with unit spheres centered at each lattice point) is called separable if there is a plane which has no interior points of any sphere in it. MAKAI [28] has recently determined the density of the thinnest non-separable packing in three dimensions.

MAKAI solved the problem by showing how a solution to the packing problem leads to a solution of the separable lattice problem and vice versa; however, the extreme lattices are different. Thus this problem is solved for dimensions $d=2$, 3, 4, 5, 6, 7, 8.

Example 5. Trapping a sphere with a lattice.

L. FEJES TÓTH has recently asked the following question: What is the lattice of least density such that any closed unit sphere cannot move more than a finite distance without hitting a lattice point?

For the two dimensional case it is an easy exercise to prove that the lattice of points with even integral coordinates is the thinnest with density $^1/_4$. In this paper we confirm the conjecture of L. FEJES TÓTH that the least density in three dimensions is $\left(\frac{\sqrt{7142+1802\sqrt{17}}}{32}\right)^{-1}=0.265\ldots.$

The extreme lattice is the one generated by three vectors of length $\frac{1}{2}\sqrt{7+\sqrt{17}}$ any two of which make an angle of $\cos^{-1}\left(\frac{\sqrt{17}-1}{8}\right)=67.012\ldots^\circ$.

It is interesting to note that for $d=2$ Examples 1, 2 and 3 all have the same extreme lattice. For $d=3$ Examples 1 and 3 have the same solution.

In the next section we reinterpret the problem of Example 5 and reduce the problem to that of finding the lattice of maximal volume subject to certain restrictions on the generating vectors. In the third section we prove the main theorem by solving the reduced problem. In the concluding section we give some related conjectures and open problems.

II. The reduction of the problem. We reformulate the problem as follows:

Reformulation 1. Find the lattice of closed spheres of greatest determinant such that the complement of the spheres has only bounded components.

It is clear that this problem is equivalent to the first since a sphere can move freely without touching a lattice point if and only if its center remains in the complement of the sphere lattice.

Reformulation 2. Find the lattice of closed spheres of maximal determinant such that no point can move more than a bounded distance without touching the boundary of one of the spheres.

A lattice, of any determinant, which has the desired property is called a *blocking lattice*.

It is obvious (in any case it is a corollary of the work of HEPPES and HORTOBAGYI [23, 24] that at least two of the lattice spheres of a blocking lattice must have a com-

mon point. Thus there is a lattice vector A with $|A|<2$. We may suppose A is the shortest. If $\pm A$ are the only lattice vector of length less than 2, then it would be possible to move a point along a path close to the sphere centered at the origin, say O, until it is near the sphere centered at A, since no other spheres except those at $\pm A$ have an inner point in common with the sphere at O. We can continue to move the point without being in or on a lattice sphere until it is near the sphere centered at $2A$, then $3A$, etc.; thus the lattice is not a blocking lattice.

Thus there is a second lattice vector B not parallel to A with $|B|<2$. We suppose B is the shortest.

If all the vectors of length less than 2 are in the plane generated by A and B, then by a similar argument to the one above we may move a point as far as we like in a path close to the spheres centered in the plane of A and B. Thus there are vectors not in the plane of A and B which have length less than 2. Let C be the shortest. We summarize the above by

LEMMA 1. *In a blocking lattice there are three independent lattice vectors of length less than* 2.

In the sequel S denotes the surface of the sphere of radius 1 centered at O and S^* denotes the convex hull of S. $S(A)$ denotes the tranlation of S by A.

We next show that in a maximal blocking lattice there are three lattice planes in independent directions which are completely covered by the spheres at the lattice points in that plane where by a *lattice plane* we mean a plane through the origin generated by two independent lattice vectors. The planes are in *independent directions* if the normal vectors are linearly independent.

We begin with several lemmas.

LEMMA 2. *The lattice generated by three vectors each of length* $e=\frac{1}{2}\sqrt{7+\sqrt{17}}$ *and every pair meeting at an angle* $\alpha=2\cos^{-1}\frac{1}{4}\sqrt{7+\sqrt{17}}$ *is a blocking lattice. The determinant of this lattice is* $\delta_0=\frac{1}{32}\sqrt{7142+1802\sqrt{17}}=3.772\ldots.$

PROOF. It is a simple calculation to see that the circumcircle of the triangle with two edges of length e meeting at the angle α has radius 1. Thus the fundamental region determined by these three vectors has every face covered by spheres of the lattice. Thus a point is trapped in whatever fundamental region it lies. The lemma is proved.

LEMMA 3. *No point of space can be in or on more than* 4 *spheres of a maximal blocking lattice.*

PROOF. Suppose there is such a point. If the certers are all coplanar*, there are five lattice points in a circle of radius 1. Consider the convex hull of these points. It is a polygon, say P. P is either a triangle, quadrilateral, or pentagon. If one draws all

*If the five (or more centers of the spheres containing this point do not all lie in a plane, then the determinant of the lattice they generate is bounded above by twice the area of the eguilateral triangle inscribed in a unit circle; i.e., by $2\cdot 3\sqrt{3}/4=2.59\ldots<\delta_0=3.77\ldots$.

the edges between these points one divides P into at least 3 triangles the vertices of which are lattice points. The area of P is at most the area of a regular pentagon inscribed in a unit unit sphere. Thus the smallest triangles, say $\triangle EFG$, has area at most $\frac{1}{3}\left(\frac{5}{4}\sin 36°\right)$. If we take E as the origin, then $\overrightarrow{EF}$, $\overrightarrow{EG}$ and some other vector say A with $|A|<2$ generate the lattice. Thus the area of the fundamental domain is less than $\frac{5}{3}\sin 36° < \frac{5}{3} < \delta_0 = 3.77$, where δ_0 is the density of the blocking lattice of Lemma 2. Thus then lattice under consideration is not optimal Lemma 3 is proved.

LEMMA 4. *Let P, Q be points of the boundary of the fundamental domain of a blocking lattice such that $P-Q$ is a lattice point. If neither P nor Q are in or on the boundary of any lattice sphere, then there is no path joining P and Q which has no point in common with three spheres and which has no points interior to any sphere.*

PROOF. This is a standard compactness argument. Since no sphere except those which the path touches can be closer to it than ε for some $\varepsilon>0$, one can construct a path from P to Q which is always within ε of the original path and is just off the surface of any sphere upon which the original path lay. It follows that the lemma is true.

LEMMA 5. *In an optimal blocking lattice there is a point which is in at least* 3 *of the spheres.*

PROOF. Suppose that a sphere lattice has no point common to three lattice spheres. Let A and B be the vectors of Lemma 1. Let R' be that point on the intersection of the spheres centered at O and A such that the foot of the perpendicular from R' to the plane of A and B, $R(A, B)$, hits the plane at $\frac{A}{2}$. Let R be just past R' on the extension of the perpendicular through P', but close enough that P is on no lattice sphere.

To complete the proof we construct a path from R to $R+B$ consisting of the following parts in order:

1. The line segment joining R to R'.
2. The arc of the circle $S \cap S(A)$ joining R' to $R''=S \cap S(A) \cap P(A, B)$.
3. Either the circular arc of $P(A, B) \cap S$ joining P'' to $P(A, B) \cap S \cap S(B)$, if no point of this arc is interior ot any other lattice sphere. Or, if other lattice spheres say $S_1, S_2, \ldots, S_k$ intersect this arc they do so on disjoint pieces since we are supposing no point is common to three spheres, to that for each such intersection we can take a detour on the semicircle above $P(A, B)$ of the circle $S \cap S_i$.
4. We repeat the above type procedure, symmetrically in reverse order to construct the rest of the path from P'' to $R'+B$.

Lemma 4 says that this lattice is not a blocking lattice; thus Lemma 5 is established.

We note that if three spheres have a point in common, say the spheres S, $S(A)$ and $S(B)$ then the plane $P(A, B)$ is covered by the spheres centered at $S(mA+nB)$, $m, n=0, \pm 1, \pm 2, \ldots$.

LEMMA 6. *In an optimal blocking lattice, Λ there is a basis A, B, C such that $S\cap S(A)\cap S(B)\neq\emptyset$ and $S\cap S(A)\cap S(C)\neq\emptyset$.*

PROOF. Let A and B generate the lattice points in the blocking plane guaranteed by Lemma 5, so that $S\cap S(A)\cap S(B)\neq\emptyset$. We may suppose that $\sphericalangle AOB\leqq\frac{\pi}{2}$, by interchanging B and $-B$ if necessary.

We first suppose there is a point $C\in\Lambda\setminus P(A, B)$ such that $S(C)$ meets the intersection of two lattice spheres centered in $P(A, B)$. By using translations and symmetries we may suppose that $S\cap S(D)\cap S(C)\neq\emptyset$ where $D=kA+lB\neq 0, k\geqq 0\ |l|\geqq k$.

Table 1

D	A'	B'
B	B	B
$A+B$	$A+B$	A
$A-B$	$A-B$	$-B$
$A+2B$	$A+2B$	B
$A-2B$	$A-2B$	$-B$

Since $ld(\Lambda)\leqq|\det(A, D, C)|$ the determinant of these three points is at most $4\left(\frac{3\sqrt{3}}{4}\right)$ for it is at most twice the area of the parallelogram half of which is the inscribed triangle $\triangle OA'B'$, and $3\sqrt{3}<2\delta_0$, we deduce that $l=0, \pm 1$. There are 5 possibilities for D (see Table 1) each of which gives rise to three vectors A', B', C which are independent and satify the desired intersection condition.

Since $|\det(A', B' D)|\leqq 3\sqrt{3}<2\delta_0$ the vectors form a basis and the lemma holds if there is such a point C.

On the other hand if there is no such point C one can construct paths on the intersections of pairs of spheres centered in $P(A, B)$ which go arbitrarily far from the initial point (i.e. walking through the passes between the peaks). By raising this path normally to the plane $P(A, B)$, one obtains a path by which a point can be moved arbitrarily far without touching a sphere.

Lemma 6 is proved.

LEMMA 7. *Given an optimal lattice with a basis as in Lemma 6, then both $S\cap S(A+C)=\emptyset$ and $S\cap S(A+B)=\emptyset$.*

PROOF. It is sufficient to prove that $S\cap S(A+C)=\emptyset$, because of the symmetry in B and C. Suppose $S\cap S(A+C)\neq\emptyset$. It follows that $\frac{A+C}{2}\in S\cap S(A)\cap S(C)\cap\cap S(A+C)$. Thus the parallelogram $OA(A+C)C$ is contained in a unit circle. If $|A|=2a$ the area of the parallelogram is at most $4a\sqrt{1-a^2}$. Since $S\cap S(A)\cap\cap S(B)\neq\emptyset$, the height of B over the plane $P(A, C)$ is at most $1+\sqrt{1-a^2}$. Let $V=d(\Lambda)$ be the volume of a Fundamental region. It follows that $V\leqq 4a\sqrt{1-a^2}(1+\sqrt{1-a^2})$ but a short calculation shows that for $0\leqq a\leqq 1$, $V\leqq 4a\sqrt{1-a^2}(1+\sqrt{1-a^2})<\delta_0$. This contradiction to the optimality of the lattice proves the lemma.

This lemma implies that the angles $\sphericalangle AOC$ and $\sphericalangle AOB$ are both acute. We are now in a position to prove

LEMMA 8. *If Λ is a lattice such that the spheres centered at the points of Λ form an optimal blocking lattice then there is a basis, A, B, C of Λ such that $S\cap S(A)\cap S(B)\neq\emptyset$, $S\cap S(A)\cap S(C)\neq\emptyset$ and either $S\cap S(B)\cap S(C)\neq\emptyset$ or $S\cap S(B)\cap S(-C)\neq\emptyset$.*

PROOF. It is enough to exhibit three linearly independent lattice points, A, B and C which satisfy the intersection conditions, since if they were not a basis of Λ, the lattice they generate would be a blocking lattice with greater determinant.

Suppose we have an optimal blocking lattice. Let A, B and C be three independent lattice points, such that $S\cap S(A)\cap S(B)\neq\emptyset$ and $S\cap S(A)\cap S(C)\neq\emptyset$.

In each of $P(A, B)$, $P(A, C)$ there is a point where three spheres intersect; the centers of these spheres form a triangle of lattice points all of which are inside a circle of radius 1 centered at the intersection point. This triangle has area at most equal to the equalateral triangle inscribed in a unit circle. Thus the area of the triangle is at most $\frac{3\sqrt{3}}{4}$. Since the lattice must have determinant at least equal to $\delta_0=3.77\ldots$. It follows that the distance between parallel blocking planes is at least $\delta_0/\frac{3\sqrt{3}}{2}>1.4$.

We consider the infinite cyclinder, $\mathscr{C}$, bounded by the planes $P(A, B)$, $P(A, B)+C$, $P(A, C)$, and $P(A, C)+B$. Since the distances between planes parallel to $P(A, B)$ and $P(A, C)$ are greater than one only spheres centered in the boundary planes meet this cylinder. Even the planes parallel to $P(B, C)$ have a distance of at least $\delta_0/4>>.94\ldots$. Thus a given sphere on the cylinder can only intersect in the cylinder with the spheres centered in its own plane in the $P(A, B)$ and $P(A, C)$ directions and with spheres in its own or adjacent planes in the $P(B, C)$ direction. Consider those spheres in the cylinder centered in $P(A, B)$; they lie on the line λA or $B+\lambda A$, $-\infty<\lambda<+\infty$. Because the distance between $P(A, C)$ and $P(A, C)+B$ is greater than 1, no spheres on λA meet $B+\lambda A$ and vice versa. Let $\mathscr{S}=\mathscr{C}\cap\left(\cup\{S(L):L\in P(A, B)\cap\Lambda\}\right)$, then, since $P(A, B)$ is covered by the spheres centered in it, $\mathscr{S}$ is a surface in $\mathscr{C}$ above $P(A, B)$ and dividing the cylinder into two portion, the lower portion bounded below by $P(A, B)$ and an upper portion bounded above by $P(A, B)+C$. The part of the surface meeting $P(A, C)$ is from spheres centered on λA and the part meeting $P(A, C)+B$ is from spheres centered on $\lambda A+B$. Thus there is an infinity long path on $\mathscr{S}$ consisting entirely of intersection of spheres one of which is on λA and the other on $\lambda A+B$ which separates the part of $\mathscr{S}$ coming from λA from the part of $\mathscr{S}$ coming from $B+\lambda A$. If none of these intersection meet a sphere from $P(A, B)+C$ then this path can be raised slightly and the lattice would not be blocking. Thus some sphere centered in $P(A, B)+C$ must meet some intersection $S(k\mathrm{A})\cap S(lA+B)$. By translation we may suppose $k=0$. From the remarks on the spacing of parallel planes we see that $l=0, \pm1, \pm2$. But Lemma 7 eliminates $l=1$, which also eliminates $l=2$. Since $|A+B|\leqq|2A+B|$ because $\sphericalangle AOB$ is acute. If $l=-2$, then $S\cap S(B-2A)\neq\emptyset$ and $S(-2A)\cap S(B-2A)=S\cap S(B)-2A\neq\emptyset$; so that both

distances from $B-2A$ to 0 and $B-2A$ to $-2A$ are at most 2. It follows that

$$V \leqq \frac{1}{2}(4a)(\sqrt{4-4a^2})(1+\sqrt{1-a^2})$$

where the first three factors give half the area of the parallelogram 0, $-2A$, $B-2A$, B, and the last factor is an upper bound on the height of $P(A, B)+C$ since $S\cap S(A)\cap \cap S(C)\neq\varnothing$. But we saw in the proof of Lemma 7 that the above expression is always less than δ_0, so that in an optimal lattice $l=-2$ is impossible. Thus the only intersection with which we need be concerned are $S\cap S(B)$, and $S\cap S(B-A)$ and the translations by kA of these intersections. One of these intersections must meet a sphere centered at a point Q in $P(A, B)+C$, where Q has the form $Q=C+\lambda A$ or $Q=C+B+\lambda A$.

If $Q=C+\lambda A$, then by the above analysis with B and C interchanged $\lambda=0$ or $\lambda=-1$. If $Q=B+C+\lambda A$ then by translation through $-B$ we see that for $S\cap S(B)\cap \cap S(Q)\neq\varnothing$, $\lambda=0$ or -1 while for $S\cap S(B-A)\cap S(Q)\neq\varnothing$, $\lambda=-1$ or -2.

Table 2, below, gives the basis satisfying the lemma which arises from each of these cases. The "$\pm$" column indicate when $-C$ is used then rather than C is fulfilling the last intersection condition for Lemma 8 with the basis A', B', C'. While verifying the table it is helpful to recall that $S\cap S(B)\cap S(B+C)\neq\varnothing$ if and only if $S\cap S(B)\cap S(-C)\neq\varnothing$.

Table 2

Non-empty Intersection	A'	B'	C'	$\pm$
$S\cap S(B)\cap S(C)$	A	B	C	+
$S\cap S(B)\cap S(C-A)$	A	B	$A-C$	−
$S\cap S(B-A)\cap S(C)$	A	C	$A-B$	−
$S\cap S(B-A)\cap S(C-A)$	A	$A-B$	$A-C$	+
$S\cap S(B)\cap S(B+C)$	A	B	C	−
$S\cap S(B)\cap S(B+C-A)$	A	$A-C$	B	+
$S\cap S(B-A)\cap S(B+C-A)$	A	C	$B-A$	−
$S\cap S(B-A)\cap S(B+C-2A)$	$A-B$	A	$C-A$	−

Lemma 8 is established.

The converse of Lemma 8 is clearly true. Thus the maximal lattice trappings a sphere is the maximal lattice generated by three vectors A, B, C satisfying Lemma 8.

We now turn to the final lemma of the reduction.

LEMMA 9. *In an optimal blocking lattice there are three generating vectors, A, B, C such that each of the triangles $\triangle OAB$, $\triangle OAC$ and one of $\triangle OBC$ or $\triangle(-B)OC$ have circumradius exactly* 1. *The angles at O of the triangles of circumradius* 1 *are all acute.*

PROOF. Clearly at least one triangle must have circumradius 1 or we can enlarge the lattice by dilation and obtain a larger blocking lattice.

If none of the other triangles have circumradius 1 we may lengthen the basis vector involved in the smaller triangles without disturbing the blocking property of the lattice, thus the lattice is not optimal. We suppose therefore that two have circumradius exactly 1.

Finally suppose that exactly two of the three triangles with circumradius at most 1 have circumradius exactly 1. By choosing a different vertex of the fundamental region as the origin we may suppose that $\triangle OAB$ and $\triangle OAC$ have circumradius 1 and that either $\triangle OBC$ or $\triangle OB(B+C)$ has circumradius less than 1. We note that by translation $\triangle OB(B+C)$ is congruent to $\triangle(-B)OC$.

There are two cases to consider:

Case 1. The plane $P(A, B)$ is perpendicular to the plane $P(A, C)$. In this case we show that the lattice cannot be optimal.

Because of the perpendicularity we may suppose that $A=(2a, 0, 0)$ $B=(b_1, b_2, 0)$ and $C=(c_1, 0, c_3)$. Since either $|B-C|\leqq 2$ or $|B+C|\leqq 2$ we see that $b_2^2+c_3^2\leqq 4$. Also since $\triangle OAB$ and $\triangle OAC$ have circumradius 1 we see that $a\leqq 1$, $b_2\leqq 1+\sqrt{1-a^2}$ and $c_3\leqq 1+\sqrt{1-a^2}$. Thus

$$V=2ab_2c_3 \tag{1}$$

and

$$V\leqq 2a(1+\sqrt{1-a^2})^2. \tag{2}$$

From the restriction on $b_2^2+c_3^2$ we see that the product $b_2c_3\leqq 2$. Thus from (1) $\delta_0\leqq 4a$ whence $a\geqq\frac{\delta_0}{4}=.94\ldots$. Bur for $\frac{\delta_0}{4}\leqq a\leqq 1$ the right hand side (RHS) of (2) has its maximum at $a=\frac{\delta_0}{4}$ and we see that $V<\delta_0$ since for that value of a $1+\sqrt{1-a^2}<\sqrt{2}$. Therefore Case 1 cannot occur for an optimal lattice.

Case 2. The planes $P(A, B)$ and $P(A, C)$ are not perpendicular. In this case we can rotate the plane $P(A, C)$ about A a small amount toward the perpendicular. This does not change either $\triangle OAB$ or $\triangle OAC$ which still have circumradius 1. It may increase the circumradius of $\triangle OBC$ or $\triangle(-B)OC$, but if the rotation is small the circumradius remains less than 1. On the other hand $V=2\mathscr{A}\cdot h_c$ where $\mathscr{A}$ is the area of $\triangle OAB$ and h is the distance of C from the plane $P(A, B)$. Since $\mathscr{A}$ is unchanged by the rotation while h_c increases we see that the original lattice was not optimal.

Finally we turn to the angles. The fact that they are acute follows from Lemma 7.

Lemma 9 is proved.

III. The Main Theorem. In this section we prove the following:

THEOREM. *The unique lattice of least density of all those lattices which trap a unit sphere is that lattice generated by three vectors of length* $\frac{\sqrt{7+\sqrt{17}}}{2}=1.667\ldots$, *any two of which meet at an angle* $\cos^{-1}\left(\frac{\sqrt{17}-1}{8}\right)=67.0213\ldots^\circ$. *The density is* $\delta=\frac{1}{V}=32\Big/\sqrt{7142+1802\sqrt{17}}=.26\,508\ldots$, *where* V *denotes the volume of the fundamental domain.*

PROOF. In view of Lemma 9 it is sufficient to show that of all lattices generated by three vectors A, B, C such that the triangle $\triangle ABO$, $\triangle ACO$ and either $\triangle BCO$ or $\triangle(-B)OC$ all have circumradius 1 and also such that the angles $\sphericalangle AOB$ and $\sphericalangle AOC$ and either $\sphericalangle BOC$ or $\sphericalangle(-B)OC$ are acute the lattice described in the theorem is maximal.

Since the triangles have circumradius 1 it is clear that the lengths of the three generating vectors determine the lattice once it is known which is the third acute angle. We proceed in each of the two cases to find a formula for the volume in terms of the lengths of the generating vectors.

Let $|A|=2a$, $|B|=2b$ and $|C|=2c$. Let $\alpha=\sphericalangle BOC$, $\beta=\sphericalangle AOC$ and $\gamma=\sphericalangle AOB$. Let R_α be the circumcenter of $\triangle BOC$ or $\triangle(-B)OC$, R_β that of $\triangle AOC$ and R_γ that of $\triangle AOB$. Let $\varrho=\sphericalangle R_\beta OA=\sphericalangle R_\gamma OA=\cos^{-1}a$, $\sigma=\sphericalangle R_\alpha O(\pm B)=\sphericalangle R_\gamma OB=\cos^{-1}b$ and $\tau=\sphericalangle R_\alpha OC=\sphericalangle R_\beta OC=\cos^{-1}c$. We note that also each of α, β, γ is positive otherwise the lattice is degenerate.

We may choose coordinates so that $A=(2a, 0, 0)$, $B=2b(x_1, x_2, 0)$, and $C=2c(y_1, y_2, y_3)$, where $x_1^2+x_2^2=1$ and $y_1^2+y_2^2+y_3^2=1$.

The volume of the fundamental domain is then given by

$$V = 8abcx_2y_3. \tag{3}$$

From the fact that the three triangles have circumradius 1 we get

$$A\cdot B = |A|\cdot|B|\cos\gamma = 4ab\cos(\varrho+\sigma) \tag{4}$$

$$A\cdot C = |A|\cdot|C|\cos\beta = 4ac\cos(\varrho+\tau) \tag{5}$$

$$(\pm B)\cdot C = |B|\cdot|C|\cos\alpha = 4bc\cos(\sigma+\tau). \tag{6}$$

These equations reduce to

$$x_1 = \cos(\varrho+\sigma) = \cos\gamma \tag{7}$$

$$y_1 = \cos(\varrho+\tau) = \cos\beta \tag{8}$$

$$\pm(x_1y_1+x_2y_2) = \cos(\sigma+\tau) = \cos\alpha. \tag{9}$$

Since $(x_1, x_2, 0)$ and (y_1, y_2, y_3) are unit vectors we may use equations (7), (8) and (9) to solve for x_2, y_2 and y_3 and obtain

$$x_2 = \sin\gamma \tag{10}$$

$$y_2 = \frac{\pm\cos\alpha-\cos\gamma\cos\beta}{\sin\gamma} \tag{11}$$

$$y_3^2 = \frac{1-\cos^2\alpha-\cos^2\beta-\cos^2\gamma\pm2\cos\alpha\cos\beta\cos\gamma}{\sin^2\gamma}. \tag{12}$$

It follows from (3) that

$$V^2 = 64a^2b^2c^2\{1-\cos^2\alpha-\cos^2\beta-\cos^2\gamma\pm2\cos\alpha\cos\beta\cos\gamma\}. \tag{13}$$

Since the angles are all acute and nonzero it follows that the maximum of V occurs when the "$\pm$" is "$+$" in the term in braces in (13); this is the case when $\triangle OBC$ is the inscribed triangle.

We next find the maximum of V. We point out now that this will also yield the corollary on the maximal volume of a tetrahedron three faces of which have circumradius at most 1.

From the definitions of $\alpha, \beta, \gamma, \varrho, \sigma, \tau$ it is straight forward to show that

$$\cos\alpha = bc - \sqrt{1-b^2}\sqrt{1-c^2}, \tag{14}$$

$$\sin\alpha = b\sqrt{1-c^2} + c\sqrt{1-b^2},$$

$$\cos\beta = ac - \sqrt{1-a^2}\sqrt{1-c^2}, \tag{15}$$

$$\sin\beta = a\sqrt{1-c^2} + c\sqrt{1-a^2},$$

and

$$\cos\gamma = ab - \sqrt{1-a^2}\sqrt{1-b^2} \tag{16}$$

$$\sin\gamma = a\sqrt{1-b^2} + b\sqrt{1-c^2}.$$

By straight forward calculation one can verify that

$$\begin{aligned} V^2 = 256a^2b^2c^2\{&ab(1-c^2)\sqrt{1-b^2}\sqrt{1-a^2} + \\ &+ ac(1-b^2)\sqrt{1-a^2}\sqrt{1-c^2} + bc(1-a^2)\sqrt{1-b^2}\sqrt{1-c^2} - \\ &- (1-a^2)(1-b^2)(1-c^2)\} = \\ &= 256a^2b^2c^2\sqrt{1-b^2}\sqrt{1-c^2}\{a\sqrt{1-a^2}\sin\alpha + \\ &+ (1-a^2)\cos\alpha\}. \end{aligned} \tag{17}$$

We note that the first form of (17) is symmetric and a, b and c while the second form emphasizes the dependence on a since α depends only on b and c.

From (17) we can compute the partial derivative, of $W(a, b, c) = V^2/256$, to obtain

$$\frac{\partial W}{\partial a} = ab^2c^2\sqrt{1-b^2}\sqrt{1-c^2}\left\{\frac{a(3-4a^2)}{\sqrt{1-a^2}}\sin\alpha + 2(1-2a^2)\cos\alpha\right\}. \tag{18}$$

For $\frac{\partial W}{\partial a}$ or $\frac{\partial W}{\partial c}$ simply interchange the a and b or a and c respectively noting that α becomes γ or β respectively. Since α is a positive acute angle we see from (18) that $\frac{\partial W}{\partial a}$ is positive for $a^2 \leqq \frac{1}{2}$ and negative for $a^2 \geqq \frac{3}{4}$. By symmetry this leads to,

$$\frac{1}{2} < a^2, b^2, c^2 < \frac{3}{4}, \tag{19}$$

for a maximal lattice.

If we set the three partial derivatives equal to zero, after some simplification

we see that for a maximal lattice a, b and c must satisfy

$$\frac{a(3-4a^2)}{\sqrt{1-a^2}(2a^2-1)} = 2\cot\alpha \tag{20}$$

$$\frac{b(3-4b^2)}{\sqrt{1-b^2}(2b^2-1)} = 2\cot\beta \tag{21}$$

$$\frac{c(3-4c^2)}{\sqrt{1-c^2}(2c^2-1)} = 2\cot\gamma. \tag{22}$$

Let

$$f(x) = \frac{x(3-4c^2)}{\sqrt{1-x^2}(2x^2-1)}, \quad \frac{1}{\sqrt{2}} < x < \frac{\sqrt{3}}{2}$$

and

$$g(y,z) = \frac{2(zy - \sqrt{1-x^2}\sqrt{1-z^2})}{z\sqrt{1-y^2}+y\sqrt{1-z^2}}, \quad \frac{1}{\sqrt{2}} < y, z < \frac{\sqrt{3}}{2}.$$

The equations above can be summarized by

$$f(x) = g(y,z) \tag{23}$$

where x, y, z are replaced by the three cyclic permutations of a, b, c.

Setting $g(t)=g(t,t)$ we compute the following

$$f'(x) = \frac{1}{\sqrt{1-x^2}\,(2x^2-1)}\left(2+\frac{1}{2x^4-3x^2+1}\right) \tag{24}$$

$$g'(t) = \frac{1}{x^2(1-x^2)^{3/2}} \tag{25}$$

$$\frac{\partial g}{\partial y} = \frac{(z\sqrt{1-y^2}+y\sqrt{1-z^2})^2+(zy-\sqrt{1-z^2}\sqrt{1-y^2})^2}{\sqrt{1-y^2}(z\sqrt{1-y^2}+y\sqrt{1-z^2})^2} \tag{26}$$

Consider the polynomial $P(x)=2x^4-3x^2+1$ which occurs in (24). It has roots at $x^2=\frac{1}{2}$, 1 and is negative in the interval $\left(\frac{1}{2},1\right)$ with minimum at $x^2=\frac{3}{4}$. We note for future reference that

$$P\left(\sqrt{\frac{3}{4}}\right) = -\frac{1}{8}. \tag{27}$$

It is now clear that for $\frac{1}{2} < x^2, y^2, z^2, t^2 \leqq \frac{3}{4}$ we have

$$f'(x) < 0 \tag{28}$$

$$g'(t) > 0 \tag{29}$$

$$\frac{\partial g}{\partial y} > 0. \tag{30}$$

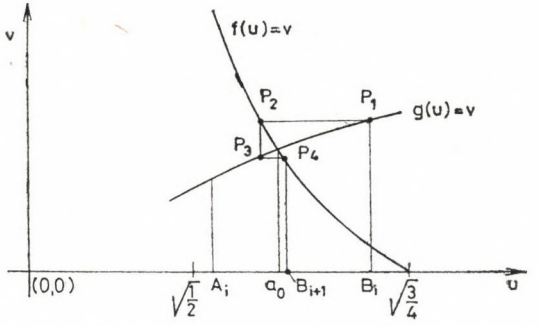
It is a straight forward calculation to verify that for $a=b=c=\frac{1}{4}\sqrt{7+\sqrt{17}}$ the three equations (20)—(22) hold. We wish to show that this is the only solution.

Suppose for a solution it is necessary that $a, b, c \in (A_i, B_i)$, where we know we may take $A_0 = \sqrt{\frac{1}{2}}$, $B_0 = \sqrt{\frac{3}{4}}$, then if (22) holds for $x=a$, $y=b$, $z=c$ we deduce from (30) that $f(a) < g(B_i, B_i)$. Setting $A_{i+1} = f^{-1}(g(B_i, B_i))$ we see that $A_{i+1} < a$. By symmetry $A_{i+1} < b, c$. It follows that $f(a) > g(A_{i+1}, A_{i+1})$ or that $a, b, c > B_{i+1} = f^{-1}(g(A_{i+1}, A_{i+1}))$. If $|B_{i+1} - A_{i+1}| \to 0$, then there can be only the one solution when $a_0 = b_0 = c_0 = \frac{1}{4}\sqrt{7+\sqrt{17}}$.

In order to show that the interval lengths go to zero it is necessary to observe that for all $x, y, z \in \left(\sqrt{\frac{1}{2}}, \sqrt{\frac{3}{4}}\right)$

$$|f'(x)| \geqq \left|f'\left(\sqrt{\frac{3}{4}}\right)\right| > g'\left(\sqrt{\frac{1}{2}}\right) \geqq g'(y),$$

since $\dfrac{\left|f'\left(\sqrt{\frac{3}{4}}\right)\right|}{\left|g'\left(\sqrt{\frac{1}{2}}\right)\right|} = 3\sqrt{2}$. Consider the following sequence of points (see Figure):



$$P_1 = (B_i, g(B_i)) = (B_i, g'(\theta_1)(B_i - a_0) + g(a_0))$$

$$P_2 = (f^{-1}(g_i(B_i)), g(B_i)) = \left(a_0 + \frac{g'(\theta_1)}{f'(\theta_2)}(B_i - a_0), g(B_i)\right)$$

$$P_3 = (f^{-1}(g(B_i), g(f^{-1}(g(B_i)))) =$$

$$= \left(f^{-1}(g(B_i)), \frac{g'(\theta_2)g(\theta_1)}{f'(\theta_2)}(B_i - a_0) + g(a_0)\right)$$

$$P_4 = (f^{-1}(g(f^{-1}(gB_i))), g(f^{-1}(g(B_i)))) =$$

$$= \left(\frac{g'(\theta_2)g'(\theta_1)}{f'(\theta_2)f'(\theta_1)}(B_i - a_0) + a_0, g(f^{-1}(g(B_i)))\right) = (B_{i+1}, g(f^{-1}(g(B_i))))$$

where $\theta_k \in (A_i, B_i)$, $k = 1, 2, 3, 4$.

Thus $|B_{i+1}-a_0|<\varrho^2(B_i-a_0)$. Similarly $|A_{i+1}-a_0|<\varrho^2(A_i-a_0)$. Since $\varrho<1$ we see there is precisely the one simultaneous solution of the three equation.*

Since there must exist a maximum and it cannot be at the boundary $a=b=c=a_0$ must be the maximum.

The theorem is proved.

Corollary. Of all tetrahedra three faces of which can be inscribed in circles of radius at most 1, the largest is the tetrahedra with base an equilateral triangle and 3 edges of length $\frac{1}{2}\sqrt{7+\sqrt{17}}$ meeting at the apex with equal face angles of $\cos^{-1}\left(\frac{17-1}{8}\right)=67.02\ldots°$.

IV. Related problems. *Problem 1.* Suppose instead of a sphere we wish to trap a convex body, K of more arbitrary shape. We then must specify if the permissible motions of K are restricted to translation or allow rotation.

In the restricted case the problem is affine invariant and the theorem provides a solution in the case of ellipsoids, but the problem is open for more general shapes. Also the problem of the thinnest lattice which can trap any set of contsant width 1 is of interest.

In the plane the restricted case again reduces to finding the maximal lattice with basis A, B such that $K\cap(K+A)\neq 0$ and $K\cap(K+B)\neq 0$ which is not too difficult for a given body K.

For the unrestricted case it is clear that a lattice which traps the in-sphere (in-circle in the plane) traps the body, however this does not seem in general to be the most efficient method. In the plane the square lattice with an edge length w, where w denotes the least distance between parallel support line of K, traps K, but it is not clear if it is always best possible.

Problem 2. For a given determinant which lattice of spheres covers the greatest volume of space?

The two dimensional case of this problem has been settled. The best lattice is the lattice generated by an equallateral triangle. In fact the best lattice is known to be the best configuration even if non lattice arrangements are allowed. See L. FEJES TÓTH [16].

In this problem, as in so many in discrete geometry, there seems to be great difficulty in going from two to three dimensions. Tools are needed by which two dimensional proofs can be changed to yield three (and higher) dimensional results.

---

* Professor H. GUNJI through an ingeneous set of trigonometric substitutions has shown that the equations (20)—(22) can be reduced to

$$\sin(\varrho-\sigma)[2\cos(2\varrho+2\sigma+\tau)\cos\tau] = 0$$

and the two similar equations obtained by permutation. Since $1/2\leqq a^2, b^2, c^2\leqq 3/4$ implies $45°\leqq \leqq\varrho, \sigma, \tau\leqq 60°$, $\cos(2\varrho+2\sigma+\tau)$ is negative and the only solutions are when $\varrho=\sigma=\tau$; i.e. $a=b=c$. It then follows that $a=b=c=\frac{1}{4}\sqrt{7+\sqrt{17}}$.

BIBLIOGRAPHY

[1] BAMBAH, R. P.: On lattice coverings, *Proc. Nat. Inst. Sci. India*, **19** (1953), 447—459.
[2] BAMBAH, R. P.: On lattice coverings by spheres, *Proc. Nat. Inst. Sci., India,* **20** (1954), 25—52.
[3] BAMBAH, R. P.: Packing and Covering, *Math. Student*, **38** (1970), 133—138.
[4] BAMBAH, R. P. and DAVENPORT, H.: The covering of n-dimensional space by spheres, *J. Lond. Math. Soc.* **27** (1952), 224—229.
[5] BARNES, E. S.: Covering of space by spheres, *Can. J. Math.*, **8** (1956), 293—304.
[6] BARNES, E. S. and DICKSON, T. J.: Extreme coverings of n-space by spheres, *J. Aust. Math. Soc.*, **7** (1967), 115—127. (Corrigendum *Ibid.* **8**, 638—640).
[7] BLEICHER, M. N.: Lattice Coverings of n-space by spheres, *Can. J. Math.*, **14** (1962), 632—650.
[8] COXETER, H. S. M., Few, L. and ROGERS, C. A.: Covering space with equal spheres, *Mathematika*, **6** (1959), 147—157.
[9] DAVENPORT, H.: The covering of space by spheres, *R. C. Circ. mat. Palermo,* (2), **1** (1952), 92—107.
[10] DELONE, B. N. and RYSKOV, S. S.: Solution of the problem of the least dense lattice covering of a 4-dimensional space by equal spheres, *Soviet Math. Dokl.*, **4** (1963), 1333.
[11] DICKSON, T. J.: An extreme covering of 4-space by spheres, *J. Aust. Math. Soc.*, **6** (1966), 179—192.
[12] DICKSON, T. J.: The extreme covering of 4-space by spheres, *J. Aust. Math. Soc.*, **7** (1967), 490—496.
[13] DICKSON, T. J.: A sufficient condition for an extreme covering of n-space by spheres, *J. Aust. Soc.*, **8** (1968), 56—62.
[14] ERDŐS, P. and ROGERS, C. A.: The covering of n-dimensional space by spheres, *J. Lond. Mat. Soc.*, **28** (1953), 287—293.
[15] FEJES TÓTH, L.: Eine Bemerkung über die Bedeckung der Ebene durch Eibereiche mit Mittelpunkt, *Acta Sci. Math. Szeged,* **11** (1946), 93—95.
[16] FEJES TÓTH, L.: *Lagerungen in der Ebene auf der Kugel und im Raum*, zweite Auflage, Springer-Verlag-Heidelberg, 1972.
[17] FEJES TÓTH, L. and MOLNÁR, J.: Unterdeckung und Überdeckung der Ebene durch Kreise, *Math. Nachr.*, **18** (1958), 235—243.
[18] FEW, L.: Covering space by spheres, *Mathematika*, **3** (1956), 136—139.
[19] GAMECKII, A. F.: The optimality of the principle lattice of Voronoi of first type among the lattices of first type of any numbers of dimensions, *Soviet Math. Dokl.*, **4** (1963), 1014—1016.
[20] GAMECKII, A. F.: On the theory of covering an n-dimensional Euclidean space with equal spheres, Soviet Math. Dokl., **3** (1962), 1410—1414.
[21] GAMECKII, A. F.: The geometrical meaning of the Zelling Parameter of n-dimension lattice of the first Voronoi type, (in Russian) *Research on Algebra and Mathematical Analysis,* Moldavian Academy of Science, 1956.
[22] GAUSS, C. F.: Untersuchungen über die Eigenschaften der positiven ternären quadratischen Formen von Ludwig August Seeber, *Göttingische gelehrte Anzeigen,* 1831 Juli 9, see Werke, Göttingen, 1836, II, 188—196, or *J. reine angew. Math.* **20** (1840), 312—320.
[23] HEPPES, A.: Ein Satz über gitterförmige Kugelpackungen, *Ann. Univ. Sci. Budapest. Eötvös Sect. Math.* **8—4** (1960—61), 89—90.
[24] HORTOBÁGYI, I.: Durchleuchtung gitterförmiger Kugelpackung mit Lichtbündeln, *Studia Sci. Math. Hung.*, **6** (1971)
[25] HORVÁTH, J.: Über die Durchsichtigkeit gitterförmiger Kugelpackungen, *Studia Sci. Math. Hung.*, **5** (1970), 421—426.
[26] KERSHNER, R.: The number of circles covering a set, *Amer. J. Math.*, **61** (1939), 665—671.
[27] LAGRANGE, J. L.: Recherches d'arithmétique, *Nouveaux Mémoires de l'Académie royal des Sciences at Belles-Lettres de Berlin,* pp. 265—312, *Oeuvres,* III, 693—758 (1773).
[28] MAKAI, E.: On the thinnest non-separable lattice of convex bodies, submitted to *Periodica Mathematica Hungarica.*
[29] ROGERS, C. A.: *Packing and covering,* Camb. Tracts. in Maths, and Math. Physics # 54, Camb. U.P., Cambridge, 1964.
[30] WATSON, G. L.: The covering of space by spheres, *R.C. Circ. mat. Palermo*, (2), **5** (1956), 93—100.

Dept. of Math., Univ. of Wis.. Madison

(Received October 7, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 171—173.

AN ISOPERIMETRIC PROBLEM FOR TESSELLATIONS

by

G. FEJES TÓTH

Let us recall the fact [1, pp. 84] that among the decompositions of the Euclidean plane into convex polygons of given equal area the average perimeter of the polygons attains its minimum for the regular hexagonal decomposition. In what follows we shall deal with the analogous problem if instead of polygons of equal area polygons of not too different areas are considered. Our result is contained in the following

THEOREM. *If a convex polygon with at most six sides is decomposed into n convex polygons such that the quotient of the areas of any two is at least*

$$q \geqq \left(\frac{\sqrt{6\tan\pi/6}-\sqrt{7\tan\pi/7}}{\sqrt{5\tan\pi/5}-\sqrt{6\tan\pi/6}}\right)^2 = 0.317\ldots$$

then the total perimeter of the polygons is at least

$$4\frac{\sqrt{nH}\sqrt[4]{12q}}{1+\sqrt{q}}.$$

It easily follows: *Decompose the plane into convex polygons of average area* 1 *such that the quotient of areas of any two is at least* $q \geqq 0.317\ldots$. *If the average perimeter p of the polygons exists then*

$$p \geqq \frac{4\sqrt[4]{12q}}{1+\sqrt{q}}.$$

As to the definition of the average of a functional see [1, pp. 57].

In this inequality equality is attained if "almost all" polygons are regular hexagons of area $\sqrt{q}$ and $1/\sqrt{q}$ such that the ratio of the numbers of the small and big hexagons is $1:\sqrt{q}$. We shall outline the construction of such a decomposition after the proof of the Theorem.

We shall denote a domain and its area with the same symbol.

Let H be a convex polygon with at most six sides. Decompose H into convex polygons $P_1, \ldots, P_n$ satisfying the condition of the Theorem. If p_i is the number of sides of P_i then, as a simple consequence of EULER's polyhedron theorem, we have $\sum_{i=1}^{n} p_i \leqq 6n$. Let T be the total perimeter of the polygons. Then in view of the isoperi-

metric property of the regular polygons we have

$$T \geqq \sum_{i=1}^{n} \sqrt{P_i} f(p_i),$$

where $f(p)=2\sqrt{p \tan \pi/p}$ is the perimeter of a regular p-gon of area 1. Observe that $f(3), f(4), \ldots$ is a decreasing convex sequence. Let us investigate the minimum of the sum $\sum_{i=1}^{n} \sqrt{P_i} f(p_i)$ under the conditions $\sum_{i=1}^{n} \sqrt{P_i}=H$, $\sum_{i=k}^{n} p_i \leqq 6n$ and

$$\frac{P_i}{P_j} \geqq q \geqq \left(\frac{f(6)-f(7)}{f(5)-f(6)}\right)^2, \quad i,j = 1, \ldots, n.$$

Assume that for some j we have $p_j>6$. Then for some i we have $p_i<6$. Thus by the convexity of the sequence $\{f(p)\}$ we have

$$\frac{\sqrt{P_i}}{\sqrt{P_j}} \geqq \sqrt{q} \geqq \frac{f(6)-f(7)}{f(5)-f(6)} \geqq \frac{f(p_j-1)-f(p_j)}{f(p_i)-f(p_i+1)},$$

i.e.

$$\sqrt{P_i} f(p_i) + \sqrt{P_j} f(p_j) \geqq \sqrt{P_i} f(p_i+1) + \sqrt{P_j} f(p_j-1).$$

Hence, replacing p_i by p_i+1 and p_j by p_j-1, the sum $\sum_{i=1}^{n} \sqrt{P_i}\, f(p_i)$ does not increase. Continuing this process until all p_i's become less than or equal to 6 and subsequently replacing the p_i's less than 6 by 6, we see that

$$\sum_{i=1}^{n} \sqrt{P_i} f(p_i) \geqq f(6) \sum_{i=1}^{n} \sqrt{P_i}.$$

We continue to give a lower bound for $\sum_{i=1}^{n} \sqrt{P_i}$ under the condition $\sum_{i=1}^{n} \sqrt{P_i}=H$ and $P_i/P_j \geqq q$. Since for $0 \leqq x-h<x \leqq y$ we have $\sqrt{x}+\sqrt{y}>\sqrt{x-h}+\sqrt{x+h}$, we may suppose that with the exception of at most one P_i all P_i's take only two values, $\underline{P}$ and $\overline{P}$ such that $\underline{P}/\overline{P}=q$. If the number of P_i's equal to $\underline{P}$ and $\overline{P}$ is $\underline{n}$ and $\overline{n}$, respectively and there is one $P_i=P$ such that $\underline{P}<P<\overline{P}$ then we have to scrutinize the sum $S=\underline{n}\sqrt{\underline{P}}+\sqrt{P}+\overline{n}\sqrt{\overline{P}}$. Since $\underline{n}+1+\overline{n}=n$, $\underline{n}\underline{P}+P+\overline{n}\overline{P}=H$ and $\underline{P}/\overline{P}=q$, we have

$$S = \underline{n}\sqrt{\frac{q(H-P)}{q\underline{n}+\overline{n}}} + \sqrt{P} + \underline{n}\sqrt{\frac{H-P}{q\underline{n}+\overline{n}}}.$$

This being a concave function of P, S attains its minimum if either $P=\underline{P}$ or $P=\overline{P}$. Thus our problem reduces to find the minimum of

$$S = \lambda n \sqrt{\underline{P}} + (1-\lambda)\sqrt{\overline{P}}$$

when $\lambda\overline{P}+(1-\lambda)\,\underline{P}=H/n$, $\underline{P}/\overline{P}=q$, $0 \leqq \lambda \leqq 1$ and λn is an integer. Now we have

$$S = \sqrt{nH}\frac{\lambda\sqrt{q+1-\lambda}}{\sqrt{\lambda q+1-\lambda}}.$$

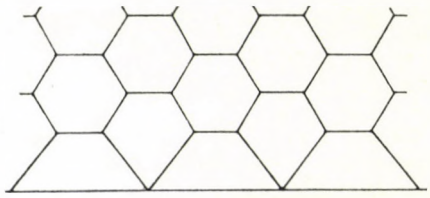
A lower bound for S will be obtained by dropping the condition that λn is an integer. It is easily seen that for $0 \leqq \lambda \leqq 1$ S is a convex function of λ which attains its minimum

$$S_{\min} = \sqrt{nH}\,\frac{2\sqrt[4]{q}}{1+\sqrt{q}} \text{ at } \lambda = \frac{1}{1+\sqrt{q}}.$$

This completes the proof of the Theorem.

To conclude we construct a tessellation which shows that in the above inequality for the average perimeter equality can be attained.

In a regular hexagonal tessellation let l be a line containing a side of a hexagon. Let l be in a horizontal position. The line l bisects some of the hexagons. Replace the upper part of these hexagons by isosceles triangles based on the horizontal diagonal of the respective hexagons so as to obtain pentagons having the same area as the hexagons. Drawing the line through the vertices of the pentagons lying above l we obtain isosceles trapezoids. It is easy to chek that these trapezoids also have the same area as the hexagons (Fig.).



It is obvious from this construction that we can decompose a parallel strip into equiareal convex polygons the bulk of which consists of regular hexagons. Decompose the plane into parallel strips which are decomposed in this manner alternatively into polygons of area $\sqrt{q}$ and $1/\sqrt{q}$. Let these strips be, in their natural order, $\ldots, s_{-2}, s_{-1}, s_0, s_1, s_2, \ldots$ so that the strips with even and odd indices contain small and big polygons, respectively. Choosing the width w_i of s_i so that $\lim_{n\to\pm\infty} \frac{w_{2n+1}}{w_{2n}} = \frac{1}{\sqrt{q}}$ and $w_n = O(\sqrt{n})$ we obtain the required decomposition of the plane. In fact, the condition that $w_n \to \infty$ implies that almost all polygons are regular hexagons of area $\sqrt{q}$ and $1/\sqrt{q}$. On the other hand, the additional condition that $w_n = 0(\sqrt{n})$ along with $\lim_{n\to\pm\infty} \frac{w_{2n+1}}{w_{2n}} = 1$ guarantees the existence of the average area and perimeter as well as the proper ratio of the numbers of the small and big hexagons.

It can be taken for granted that in the Theorem the constant 0.317... can be replaced by a smaller one. The Archimedean tessellation (4, 6, 12) shows that with the constant $\frac{2-\sqrt{3}}{3} = 0.0893\ldots$ the Theorem does not hold anymore.

It is conjectured that the Theorem continues to hold without the restriction of the convexity of the partial polygons but the proof seems to involve considerable difficulties.

REFERENCE

[1] FEJES TÓTH, L.: *Lagerungen in der Ebene auf der Kugel und im Raum* (zweite Auflage), Springer-Verlag Berlin—Heidelberg—New York, 1972.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received March 20, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 175—177.

ON AVERAGING INTERPOLATION OF HERMITE-FEJÉR TYPE

by

P. VÉRTESI

1. Introduction. Recently MOTZKIN, SHARMA and STRAUS have introduced the concept of averaging interpolation of Lagrange type ([1]). Later it was proved that for many cases this process has the same convergence properties than the ordinary Lagrange interpolation (see e.g. [2], [3]). Now we wish to show that supposing some natural conditions we also cannot achieve a better uniform convergence estimation for the averaging interpolation of Hermite—Fejér type defined in [4] than for the usual one.

2. Notations and preliminary results. Let, as usual,

$$-1 \leqq x_{n,n} < x_{n-1,n} < \ldots < x_{1,n} \leqq 1 \quad (n = 1, 2, 3, \ldots), \tag{2.1}$$

$$\omega_n(x) = c(x-x_{1,n})(x-x_{2,n})\ldots(x-x_{n,n}), \tag{2.2}$$

$$l_{k,n}(x) = \frac{\omega_n(x)}{\omega_n'(x_{k,n})(x-x_{k,n})} \quad (k = 1, 2, \ldots, n), \tag{2.3}$$

$$h_{k,n}(x) = \left[1 - \frac{\omega_n''(x_{k,n})}{\omega_n'(k_{k,n})}(x-x_{k,n})\right] l_{k,n}^2(x), \tag{2.4}$$

$$H_n(f; x) = \sum_{k=1}^{n} f(x_{k,n}) h_{k,n}(x). \tag{2.5}$$

As we know the degree of the Hermite—Fejér interpolatory polynomials H_n are not greater than $2n-1$, further

$$H_n(f; x_{k,n}) = f(x_{k,n}),\ H_n'(f; x_{k,n}) = 0 \quad (k = 1, 2, \ldots, n) \tag{2.6}$$

where $f(x)$ is defined on $[-1, 1]$. Let further

$$R_n = \sum_{k=1}^{n} (-1)^{k-1} \frac{\omega_n''(x_{k,n})}{[\omega_n'(x_{k,n})]^3}, \tag{2.7}$$

$$S_n(x) = \sum_{k=1}^{n} (-)^{k-1} h_{k,n}(x), \tag{2.8}$$

$$A_n(f; x) = H_n(f; x) - \frac{S_n(x)}{R_n} \sum_{k=1}^{n} \frac{\omega_n''(x_{k,n})}{[\omega'(x_{k,n})]^3} f(x_{k,n}) \tag{2.9}$$

(as (2.7)—(2.9), see [4]). In [4] R. B. SAXENA proved

THEOREM 2.1. *If $R_n \neq 0$ then $A_n(f; x)$ is the only polynomial of degree $\leq 2n-2$ for which*

$$A_n(f; x_{k,n}) - A_n(f; x_{k+1,n}) = f(x_{k,n}) + f(x_{k+1,n}) \quad (k = 1, 2, \ldots, n-1), \tag{2.8}$$

$$A_n'(f; x_{k,n}) = 0 \quad (k = 1, 2, \ldots, n). \tag{2.9}$$

The name of A_n is averaging interpolatory polynomial of Hermite—Fejér type.

He also proved that $\lim_{n\to\infty} \|A_n(f; x) - f(x)\|_{[-1,1]} = 0$ if $f \in C[-1, 1]$ ($=f$ is continuous on $[-1, 1]$) and $\omega_n(x) = \sin\vartheta \sin(n-1)\vartheta$ with $x = \cos\vartheta$ ($\|g(x)\|_{[-1,1]} = \max_{-1\leq x\leq 1} |g(x)|$).

3. New result. THEOREM 3.1. *Let us suppose that*

$$x_{k,n} = -x_{n-k+1,n} \quad (k = 1, 2, \ldots, n). \tag{3.1}$$

Then for any odd n

$$R_n = 0 \quad (n = 1, 3, 5, \ldots).$$

If n is even, $R_n \neq 0$ and for $f \in C[-1, 1]$

$$\|H_n(f; x) - f(x)\|_{[-1,1]} = O(1)\,|H_n(f; 0) - f(0)| \quad (n = 2n_1, 2n_2, 2n_3, \ldots) \tag{3.2}$$

then we have

$$\|H_n(f; x) - f(x)\|_{[-1,1]} = O(1)\,\|A_n(f(x) - f(x)\|_{[-1,1]} \quad (n = 2n_1, 2n_2, 2n_3, \ldots). \tag{3.3}$$

PROOF. Using the definition of $\omega_n(x)$ and (3.1) we have

$$\omega_n^{(i)}(x_{k,n}) = (-1)^{n+i}\omega_n^{(i)}(x_{n-k+1,n}) \quad (n = 1, 2, 3, \ldots; k = 1, 2, \ldots, n, i = 0, 1, 2). \tag{3.4}$$

So if $n = 2s+1$ we have (sometimes omitting the superfluous indices)

$$R_{2s+1} = \sum_{k=1}^{2s+1} (-1)^{k-1} \frac{\omega''(x_k)}{[\omega'(x_k)]^3} =$$

$$= \sum_{k=1}^{s} (-1)^{k-1} \left\{ \frac{\omega''(x_k)}{[\omega'(x_u)]^3} + \frac{\omega''(x_{2s+2-k})}{[\omega'(x_{2s+2-k})]^3} \right\} + (-1)^s \frac{\omega''(x_{s+1})}{[\omega'(x_{s+1})]^3} = 0 \quad (s = 1, 2, 3, \ldots)$$

because of $\omega''(x_{s+1}) = \omega''(0) = 0$ ($\omega_{2s+1}''(x)$ is an odd function).

If $n = 2s$ then by (2.8), (2.3) and (2.4)

$$S_{2s}(0) = \sum_{k=1}^{2s} (-1)^{k-1} h_k(0) = \sum_{k=1}^{2s} (-1)^{k-1} l_k^2(0) + \sum_{k=1}^{2s} (-1)^{k-1} \frac{\omega''(x_k)}{\omega'(x_k)} x_k l_k^2(0) =$$

$$= \omega^2(0) \sum_{k=1}^{s} (-1)^{k-1} \left\{ \frac{1}{[\omega'(x_k)x_k]^2} - \frac{1}{[\omega'(x_{2s+1-k})x_{2s+1-k}]^2} \right\} +$$

$$+ \omega^2(0) \sum_{k=1}^{s} (-1)^{k-1} \left\{ \frac{\omega''(x_k)}{[\omega'(x_k)]^3 x_k} - \frac{\omega''(x_{2s+1-k})}{[\omega'(x_{2s+1-k})]^3 x_{2s+1-k}} \right\} = 0.$$

So we have by (2.9) and (2.5) that

$$A_n(f; 0) = H_n(f; 0) \quad (n = 2, 4, 6, \ldots)$$

from where we have (3.3) by (3.2).

4. Remark. Let us consider the roots of the Jacobi polynomials $P_n^{(\alpha,\beta)}(x)$ $(\alpha, \beta > -1)$. If $\alpha = \beta$ then we have (3.1) (see [5], 4.1). Moreover, using [5], (14.5.1), (15.3.1) and (15.3.8) we have that $R_{2s}^{(\alpha,\alpha)} \neq 0$, (i.e. $A_{2s}^{(\alpha,\alpha)}(f; x)$ exist) if $-1 < \alpha \leqq -\frac{1}{2}$. (We denote by $R_n^{(\alpha,\alpha)}$ the numbers R_n corresponding to the roots of $P_n^{(\alpha,\alpha)}(x)$. Similar meaning have $A_n^{(\alpha,\alpha)}(f; x)$.) We can prove that $R_n^{(\alpha,\alpha)} \neq 0$ for other values of α but the proofs are not always so simple.

Let us see an example for (3.2). If we use the same roots and correspondence as above we have

$$\|H_n^{(\alpha,\alpha)}(|t|; x) - |x|\|_{[-1,1]} = O\left(\frac{\log n}{n}\right) \quad \left(-1 < \alpha \leqq -\frac{1}{2}, \quad n = 2, 3, 4, \ldots\right)$$

(see [6], (2.1)).

Further, using [5], (14.5.2), (8.21.10), (8.9.9) and (8.9.1) we get

$$H_n^{(\alpha,\alpha)}(|t|; 0) = \sum_{k=1}^{n} h_{k,n}^{(\alpha,\alpha)}(0)\,|x_{k,n}| \geqq \sum_{k=\left[\frac{k}{4}\right]}^{\left[\frac{n}{2}\right]} h_{k,n}^{(\alpha,\alpha)}(0)\,x_{k,n} \geqq$$

$$\geqq c \sum_{k=\left[\frac{n}{4}\right]}^{\left[\frac{n}{2}\right]} l_{k,n}^2(0)\,x_{k,n} = \frac{c_1}{n^2} \sum_{k=\left[\frac{n}{4}\right]}^{\left[\frac{n}{2}\right]} \frac{1}{x_{k,n}} = c_2 \frac{\log n}{n} \quad (n = 2, 4, 6, \ldots)$$

I.e., we have (3.2) for $|x|$ and $n = 2, 4, 6, \ldots$ $\left(1 < \alpha \leqq -\frac{1}{2}\right)$.

*

The author is grateful to the referee, Professor O. Kis, for his valuable remarks.

REFERENCES

[1] Motzkin, T. S., Sharma, A., Straus, E. G.: Averaging interpolation, *Spline Functions and Approximation Theory*, ISNM, **21** (1972), 191—233.
[2] Saxena, R. B., Sharma, A.: Convergence of averaging interpolation opreators, *Demonstratio Mathematica*.
[3] Botto, M. A., Vértesi, P.: On the convergence of the Bernstein—Grünwald averaging process for averaging interpolators, *Acta Math. Acad. Sci. Hungar.*, **26** (1975), 143—151.
[4] Saxena, R. B.: Averaging interpolation of Hermite—Fejér type, *Can. Math. Bull.* (to appear).
[5] Szegő, G.: *Orthogonal polynomials*, Amer. Math. Soc. New.York, 1959.
[6] Vértesi, P.: Notes on the Hermite—Fejér interpolation based on the Jacobi abscissas, *Acta Math. Acad. Sci. Hungar.*, **24** (1973), 233—239.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received May 9, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 179—183.

REMARKS ON A PAPER OF G. G. LORENTZ

by

G. PETRUSKA

1. Let $(X, \mathcal{A}, \mu)$ be a probability space, $\{f_n\}$ a sequence of $\mathcal{A}$-measurable functions satisfying

$$0 \leqq f_n(x) \leqq 1 \quad \mu-\text{a.e. } x \in X,\ n = 1, 2, \ldots . \tag{1}$$

Let $A = \{n_1, n_2, \ldots, n_k\}$ be a finite set of positive integers, $|A| = k$. We put

$$\prod_{n_j \in A} f_{n_j} = \prod_A f, \quad \int_X \left(\prod_A f\right) d\mu = p(A), \quad p(A)^{1/|A|} = m(A).$$

We are going to find a subsequence $N = \{n_1, n_2, \ldots\}$ of natural numbers such that $m(A)$ behaves regularly on the finite subsets $A \subset N$.

Results of such types were proved by VISSER [1], SUCHESTON [2] and in the most general form by LORENTZ [3]:

PROPOSITION. *If the sequence* $\{f_n\}$ *satisfies* (1) *and for a fixed integer* $r \geqq 1$

$$\underline{\lim} \int_X f_{n_1} f_{n_2} \ldots f_{n_r} d\mu \geqq \alpha^r \tag{2}$$

(where $n_j \to +\infty$, *independently for* $1 \leqq j \leqq r$*), then for every* $\varepsilon > 0$ *there exists a subsequence* $\{g_n\} \subset \{f_n\}$ *such that* $\int_X g_{n_1} \ldots g_{n_s} d\mu \geqq (1-\varepsilon)\alpha^s$ *for every* $n_1 < n_2 < \ldots < n_s$, $s = r, r+1, \ldots$.

This result is sharp in the sense that the factor $(1-\varepsilon)$ can not be deleted in general.

The proofs in [1]—[3] are based on the combinatorial Ramsey-theorem. The main tool in our argument is weak convergence which is inadequate if in an integral $\int_X f_{n_1} \ldots f_{n_2} d\mu$ many indices may move simultaneously to infinity.

2. Since the unit ball in L_∞ is weakly compact, by (1) there exists a weak-limit point φ, $(0 \leqq \varphi \leqq 1, \mu$ a.e. $x \in X)$ for the set $\{f_n\}$. Let

$$S = \{f : f = 1 \quad \text{or} \quad f = f_{n_1} f_{n_2} \ldots f_{n_k} \varphi^j,\ n_1 < n_2 < \ldots < n_k,\ j = 0, 1, \ldots\}.$$

S is a countable subfamily, therefore there exists a subsequence $\{g_n\} \subset \{f_n\}$, which is weakly convergent to φ with respect to S (i.e. for $f \in S$, $\int_X g_n f d\mu \to \int_X \varphi f d\mu$ holds). Thus (changing the symbols) we can assume that our sequence $\{f_n\}$ is weakly convergent with respect to the system S. In the sequel this will always be assumed.

3. LEMMA. *Let* $\varrho(x) \geqq 0$, $h(x) \geqq 0$ *be bounded measurable functions on* X, $p \geqq 0$. *Suppose*

(i) $\int_X \varrho \, d\mu = \sigma > 0;$

(ii) $\int_X h\varrho \, d\mu > \sigma^{\frac{p+2}{p+1}}.$

Then

$$\left(\int_X h^n \varrho \, d\mu\right)^{1/(n+p+1)} > \left(\int_X h^{n+1} \varrho \, d\mu\right)^{1/(n+p)} \tag{3}$$

holds for every $n = 1, 2, \ldots$.

(If in (ii) we assume $\geqq$, (3) holds with $\geqq$, too.)

PROOF. Put for brevity $I_n = \left(\int_X h^n \varrho d\mu\right)^{1/(n+p+1)}$, $n = 0, 1, \ldots$. For $n = 1$, assertion (3) holds by (ii). Suppose that $I_n > I_{n-1} > \ldots > I_0$ has already been proved. We recall the well known fact that, for any probability measure λ the mean

$$M_t(f) = \left(\int_X f^t \, d\lambda\right)^{1/t} \quad (t > 0, \quad f \geqq 0\}$$

is a non-decreasing function of t. Taking $d\lambda = \frac{1}{\sigma} \varrho d\mu$ we obtain

$$\frac{1}{\sigma} (I_{n+1})^{n+p+2} = [M_{n+1}(h)]^{n+1} \geqq [M_n(h)]^{n+1} = \left[\frac{1}{\sigma} (I_n)^{n+p+1}\right]^{\frac{n+1}{n}},$$

hence

$$(I_{n+1})^{n(n+p+2)} \geqq \frac{1}{\sigma} (I_n)^{(n+1)(n+p+1)} = (I_n)^{n(n+p+2)} \frac{(I_n)^{p+1}}{\sigma} > (I_n)^{n(n+p+2)},$$

where the latter inequality holds by the induction hypothesis $I_n^{p+1} > I_0^{p+1} = \sigma$. Thus $I_{n+1} > I_n$ follows.

4. THEOREM 1. *Suppose that the weak limit point* φ *is not constant* μ *a.e. Then there exists a subsequence* $N = \{n_1, \ldots, n_k, \ldots\}$ *such that* $m(A)$ *is monotone on* N *in the following sense:*

(i) *if* $A \subset B \subset N$ *and* $\max A < \min (B \setminus A)$ *then* $m(A) < m(B)$.

PROOF. The sequence N will be selected by induction. Since $\varphi \not\equiv c$, the strict inequality

$$\int_X \varphi^2 \, d\mu < \left(\int_X \varphi \, d\mu\right)^2 \tag{4}$$

holds. Hence for $n \geqq n_1$ we have

$$\int_X \varphi f_n \, d\mu > \left(\int_X f_n \, d\mu\right)^2 \tag{5}$$

as well. Let $N_1 = \{n_1\}$.

Induction hypothesis. Suppose that $N_k=\{n_1<n_2<\ldots<n_k\}$ has been chosen and it satisfies

$$m(A) < m(B) \tag{6}$$

if $A\subset B\subset N_k$ and $\max A<\min(B\setminus A)$ and

$$m(A)^{|A|+1} < \int \left(\prod_A f\right)\varphi\, d\mu \tag{7}$$

for every $A\subset N_k$.

If $k=1$, i.e. $N_1=\{n_1\}$ then (7) reduces to (5). In order to find n_{k+1} we apply our Lemma for every fixed $A\subset N_k$, $A\neq 0$ with $\varrho:=\prod_A f$, $h:=\varphi$, $p:=|A|-1$. Condition (7) justifies the application of the Lemma. Taking $n=2$ we have

$$\left[\int_X \left(\prod_A f\right)\varphi^2\, d\mu\right]^{|A|+1} > \left[\int_X \left(\prod_A f\right)\varphi\, d\mu\right]^{|A|+2}. \tag{8}$$

Hence we can find an integer $n_k(A)>\max A$ such that for every $n\geqq n_k(A)$

$$\left[\int_X \left(\prod_A f\right) f_n\,\varphi\, d\mu\right]^{|A|+1} > \left[\int_X \left(\prod_A f\right) f_n\, d\mu\right]^{|A|+2} \tag{9}$$

and

$$\left[\int_X \left(\prod_A f\right) f_n\, d\mu\right] > m(A)^{|A|+1} \tag{10}$$

hold as well. In fact, the terms in (8) are the limits of the corresponding terms in (9) and by the same reason (7) implies (10).

Now we put

$$n_{k+1} = \max\{n_k(A): A\subset N_k, \quad A\neq\varnothing\}$$

and

$$N_{k+1} = \{n_1, \ldots, n_k, n_{k+1}\}.$$

We have to verify properties (6) and (7) on N_{k+1}. Let $A\subset B\subset N_{k+1}$, $\max A<\min(B\setminus A)$. We can suppose $n_{k+1}\in B$. (10) implies $m(B\setminus\{n_{k+1}\})<m(B)$ and the general case follows by transitivity. To verify (7) we can again suppose $n_{k+1}\in B\subset N_{k+1}$. If $B\neq\{n_{k+1}\}$ then (9) implies $\int_X \prod_B f\varphi d\mu > m(B)^{|B|+1}$. If finally $B=\{n_{k+1}\}$, then by $n_{k+1}>n_1$, (7) follows from (5). Thus the induction process provides the required sequence $n_1<n_2<\ldots$.

Corollary 1. If for a fixed integer $r\geqq 1$ the sequence $f_1, \ldots, f_n, \ldots$ satisfies $m(A)\geqq c$ ($|A|=r$) and it has non-constant weak limit point, then $m(A)>c$ holds for every every $A\subset N$, $|A|>r$ for a suitable subsequence N. In particular, $\int_X f_n d\mu\geqq c$ ($n=1,2,\ldots$) implies $\int_X f_{n_1}\ldots f_{n_k} d\mu > c^k$ ($k=2, \ldots$) for a suitable subsequence.

Remark 1. Condition $\max(A)<\min(B\setminus A)$ can not be deleted from the monotonicity property. Let

$$f_n(x)=\begin{cases} x, & 0\leqq x<1-\dfrac{1}{2^n},\\ 0, & 1-\dfrac{1}{2^n}\leqq x\leqq 1,\end{cases}$$

$n=1, 2, \ldots$. Then $f_n\to x$ weakly on [0, 1] with respect to the Lebesgue measure. If $m(A)$ is fully monotone on a subsequence $N=\{n_1<n_2<\ldots\}$ then

$$\Big(\int_0^1 f_{n_1}f_{m_1}\ldots f_{m_k}\Big)^{1/k+1}\geqq\Big(\int_0^1 f_{m_1}\ldots f_{m_k}\Big)^{1/k},\quad n_1<m_1<\ldots<m_k,\ m_i\in N.$$

Taking $m_k\to+\infty, \ldots, m_n\to+\infty$, we obtain

$$\Big(\int_0^1 f_{n_1}\varphi^k\Big)^{1/k+1}\geqq\Big(\int_0^1\varphi^k\Big)^{1/k}\quad\text{for every } k=1, 2, \ldots.$$

For $k\to+\infty$, this gives $1-\dfrac{1}{2^{n_1}}\geqq 1$.

Remark 2. The example $f_n(x)=c\left(1+\dfrac{1}{n}\right)$ $(0<c<1)$ shows that $m(A)$ can be monotone decreasing $(m(A)>m(B)$ if $\varphi\equiv c$, and $\max A<\min(B\setminus A))$. If $\varphi\equiv c$, our method (with slight modification) would give the following assertion:

THEOREM 2. *If $\varphi\equiv c$, and a sequence $\varepsilon_1>\varepsilon_2\ldots$, $\varepsilon_k\to 0$ is given then there exists a subsequence $N=\{n_1<n_2<\ldots\}$ such that $A\subset B\subset N$, $\max A<\min(B\setminus A)$, $\min A\geqq n_k$ imply $m(B)>(1-\varepsilon_k)^t m(A)$, where $t=|A|^{-1}-|B|^{-1}$.*

We omit the proof.

Finally we show that a result like Cor. 1. still holds even if $\varphi\equiv c$.

THEOREM 3. *Suppose that, for a fixed $r\geqq 1$ $m(A)>d$ $(|A|=r)$ holds. Then there exists a subsequence N such that $m(A)>d$ for every $A\subset N$, $|A|\geqq r$.*

PROOF. If the sequence $f_1, \ldots, f_n, \ldots$ has a non-constant weak-limit point, the assertion holds by Corollary 1. Thus $\varphi\equiv c$ is supposed. From

$$\int_X f_{n_1}\ldots f_{n_r}\,d\mu>d^r$$

we obtain $c^r\geqq d^r$ if one after one $n_r\to+\infty, \ldots, n_1\to\infty$, i.e. $c\geqq d$. (In particular, we have $m(A)\geqq d$ for every $|A|\leqq r$.) We are going to use an induction argument like in Theorem 1. Let $N_1=\{1, 2, \ldots, r\}$ and suppose that $N_k=\{1, \ldots, r, n_{r+1}, \ldots, n_k\}$ has been selected with the property

$$m(A)>d\quad(A\subset N_k, |A|\geqq r). \tag{11}$$

Let $A\subset N_k$ be fixed, $|A|=r$. Then by

$$\lim_{n\to\infty}\int_X\Big(\prod_A f\Big)f_n\,d\mu=cp(A)\geqq dp(A)>d^{|A|+1}$$

we have for $n \geqq n_k(A)$

$$\int_X \left(\prod_A f\right) f_n \, d\mu > d^{|A|+1}.$$

Let $n_{k+1} = \max \{n_k(A): A \subset N_k, |A| \geqq r\}$ and $N_{k+1} = \{1, \ldots, n_k, n_{k+1}\}$. Then the sequence defined by induction obviously satisfies our requirements.

Remark 3. The induction argument above would easily give the following result: If for a fixed $r \geqq 1$, $m(A) \geqq d$ ($|A| = r$) holds and $\varepsilon > 0$ is given, then there exists a subsequence N such that for every $A \subset N$, $p(A) > (1-\varepsilon) d^{|A|}$.

REFERENCES

[1] VISSER, C.: On certain infinite sequences, *Nederl. Akad. Wetensch. Proc.*, **40** (1937) 358—367.
[2] SUCHESTON, L.: On sequences of events of which the probabilities admit a positive lower limit, *J. London Math. Soc.* **34** (1959) 386—394.
[3] LORENTZ, G. G.: Remark on a paper of Visser, *J. London Math. Soc.* **35** (1960) 205—208.

Eötvös Loránd University, Budapest

(Received May 4, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 185—189.

PACKING AND COVERING PROPERTIES OF SPLIT DISKS

by

H. GROEMER and A. HEPPES

In this paper a *disk* is defined to be a compact convex subset of the euclidean plane that is centrally symmetric and has non-empty interior. If D is a disk, then every line through the center of D splits the disk in an obvious way into two congruent closed convex sets. Either of these two sets will be called a *semidisk* obtained from D. The packing constant of a disk or semidisk X (maximal packing density of sets congruent to X) will be denoted by $\delta(X)$.

At a recent meeting on discrete geometry (Madison, Wisconsin, August 1974) L. FEJES TÓTH has raised the question whether one can always obtain from a given disk D a semidisk D' so that $\delta(D)<\delta(D')$ (for the case of circles see also [3]). In the present paper we shall answer this question and solve some related problems.

First, we note that there are obviously disks, for example parallelograms and centrally symmetric hexagons, with the property that all the related semidisks have the same packing constants as the original disks. On the other hand, it is also not difficult to find a disk D and a semidisk D' obtained from D so that $\delta(D)<\delta(D')$. For example, let D be a centrally symmetric octagon with one diagonal, say d, parallel to one of its sides. Then, it is clear (see figure 1) that the two semidisks D', D'' can be translated so as to form a centrally symmetric hexagon. Hence $\delta(D)<1$, but $\delta(D')=1$.

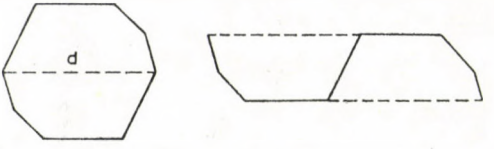


Fig. 1

Our further considerations are based on the well-known fact that for every disk D

$$\delta(D) = a(D)/a(S) \tag{1}$$

where S denotes a centrally symmetric hexagon (or parallelogram) that contains D and has minimal area. $a(X)$ signifies the area of X. We shall also discuss coverings of the plane by disks congruent to a given disk D. For this purpose we define the covering constant $\Delta(D)$ as the minimal density of coverings by disks congruent to D and subject to the condition that no two disks "cross" each other. Then, it can be shown that

$$\Delta(D) = a(T)/a(D) \tag{2}$$

where T is a centrally symmetric hexagon (or parallelogram) contained in D and of maximal area. For proofs and further details concerning the relations (1) and (2) see the books [1], [4] of L. FEJES TÓTH, and the pertinent literature cited there.

Since the relations (1) and (2) reduce our problem to the study of circumscribed and inscribed hexagons, and since results of this type are of independent interest we consider, more generally, centrally symmetric $2n$-gons ($n=2, 3, \ldots$) which contain or cover a disk or pair of semidisks. Our main result can be formulated in the following way.

THEOREM. *Let n be an integer that is greater than 1, and let D be a disk which is not a $2n$-gon. Furthermore, assume that P and Q are two centrally symmetric $2n$-gons with the property that $Q \subset D \subset P$ and that the centers of Q, D and P coincide. Then, the following statements are true.*

(i) *There exist a semidisk D_1 obtained from D and a centrally symmetric $2n$-gon $\tilde{P}$ such that $a(\tilde{P}) < a(P)$ and that $\tilde{P}$ contains two non-overlapping congruent copies of D_1.*

(ii) *There exist a semidisk D_2 obtained from D and a centrally symmetric $2n$-gon $\tilde{Q}$ such that $a(\tilde{Q}) > a(Q)$ and that $\tilde{Q}$ is contained in the union of two non-overlapping congruent copies of D_2.*

Before we turn to the proof of this theorem we formulate as a corollary an answer to the above question about packings and include also a solution of the corresponding problem for coverings. This corollary is an obvious consequence of (1), (2), and the specials case $n=2, 3$ of the Theorem.

COROLLARY. *If D is a disk that is not a centrally symmetric hexagon or a parallelogram there exist semidisks D_1, D_2 obtained from D so that*

$$\delta(D_1) > \delta(D)$$

and

$$\Delta(D_2) < \Delta(D).$$

PROOF OF THE THEOREM. First, we consider the case of a disk D that is contained in a $2n$-gon P. Let $p_1, p_2, \ldots, p_n$ and the images of these points in the center o of D be the vertices of P. Since D is not a $2n$-gon there exists a point p on the boundary of P that is not in D. There is no loss in generality by assuming that p is a point of the open segment (p_1, p_2) and that the closed segment $[p_1, p_2]$ is a side of P. Let now D' and D'' be the two semidisks obtained from D by splitting D along the line which passes through o and p. The same line splits P into two semidisks P', P'', where $D' \subset P'$, $D'' \subset P''$. It will be shown that D' can be taken as the D_1 of the theorem. Because of the convexity of D, and since $p \notin D$, either $[p, p_1]$ or $[p, p_2]$ contains no point of D. We may assume that the notation has been chosen so that $[p, p_1]$ is a side of the polygon P' and that this side contains no points of D (see figure 2 for these and the following considerations). Then, it is obviously possible to cut off from P', using a line parallel to $[p, p_1]$, a quadrangle so that the resulting polygon, say $\tilde{P}'$, has the same number of vertices as P', contains D', and has smaller area than P'. If the corresponding operation is performed on P'' it is clear that $\tilde{P}'$, and the resulting semidisk $\tilde{P}''$ can be translated so as to form a new $2n$-gon $\tilde{P}$ which has the desired properties with respect to the correspondingly translated disks D', D''. We remark that the previously excluded case $p=p_1$ could be settled by the same method.

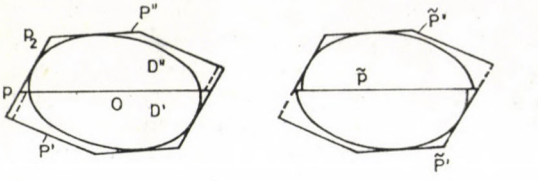


Fig. 2

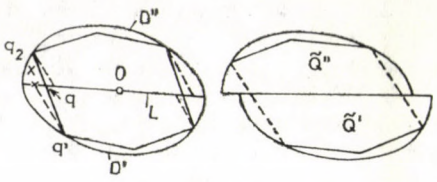


Fig. 3

To prove part (ii) of the Theorem we assume that $q_1, q_2, \ldots, q_n$ and their images in o are the vertices of Q. Since D is not a $2n$-gon we can infer that Q has a side, say $[q_1, q_2]$, so that the open segment (q_1, q_2) contains no point of D. Let q be a point on (q_1, q_2) that is closer to q_2 than to q_1, and let D be a line containing o and q. Splitting D and Q along L yields two semidisks D', D'' and two polygonal semidisks Q', Q'' so that $D' \subset Q'$, $D'' \subset Q''$. We may also assume that D' is the disk which contains q_1. Using Q' one can now construct a new semidisk $\tilde{Q}'$ by the following two modifications. First, we add to Q' a triangle with vertices at q_1, q, and at a point x on L that is inside D but outside Q. Then, we remove from Q' a similar triangle along the side parallel to $[q, q_1]$ (see figure 3). Since the area of the first triangle is larger then the area of the second one it follows that $a(\tilde{Q}') > a(Q')$. Furthermore, if x is chosen sufficiently close to q the semidisks Q' and $\tilde{Q}'$ have the same number of vertices. It is now clear that $\tilde{Q}'$ and an accordingly constructed semidisk $\tilde{Q}''$ can be translated so that there results a disk $\tilde{Q}$ which has the properties stated in the Theorem (D' can serve as D_2).

We conclude with some additional remarks concerning the packing properties of disks and semidisks.

1. Let D be a disk and let u be a given direction. We say that u is "exceptional" with respect to the packing (covering) problem if a splitting of D by a line of direction u does not improve the packing (covering) constant. Our proof of the Theorem shows that every strictly convex disk has at most three (six) exceptional directions with respect to the packing (covering) problem. But the directions singled out by our proof are not always exceptional. For example, a circle has obviously no exceptional direction, or the line used to split the octagon of figure 1 does not correspond to an

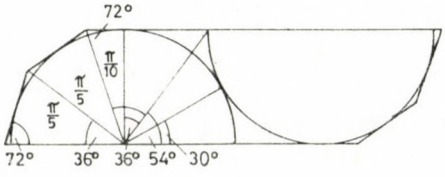


Fig. 4

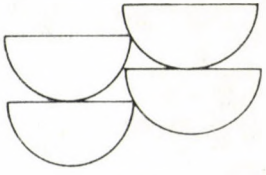


Fig. 5

exceptional direction with respect to the covering problem, although the method of proof of the Theorem would not be applicable for improving the covering constant in this case. In fact, it is not known to us whether there exist stricly convex disks, or any disks different from parallelograms and centrally symmetric hexagons, with exceptional directions.

2. The demonstration of the above theorem gives an actual procedure for improving the packing or covering constants by splitting the given disk. However, it does not provide any method for finding the actual packing or covering constants of a semidisk even if the packing and covering properties of the original disk are completely known. For example, the packing constant of a semicircle and the corresponding densest packings are apparently not known. A rather dense packing is obtained by placing two semicircles in a hexagon as in figure 4 and by tiling the plane with such hexagons. The density of this packing is found to be $\frac{\pi}{\sqrt{3}+5\tan\frac{\pi}{10}}=$ $=0{,}935\ldots$, as compared to the packing constant of circles which is $\frac{\pi}{\sqrt{12}}=0{,}906\ldots$. Another problem of this type would be to find those disks for which the increase in the packing constant (decrease in the covering constant) which can be achieved by splitting the disk is maximal.

3. FEJES TÓTH's original question can also be formulated for packings or coverings that are formed by the translates of a given disk or semidisk. However, it is to be expected that no improvement of the translative packing or covering constants can be achieved by splitting a given disk. As an example consider the case of a semicircle. The densest packing can be found by first centrally symmetrizing the semicircle, which yield the parallel domain of a line segment, and then by applying results of L. FEJES TÓTH [2] regarding packings of such domains. One obtains a packing as indicated in figure 5 of density $\frac{\pi}{2+\sqrt{3}}=0{,}841\ldots$. This constitutes a considerable reduction of the packing constant $\frac{\pi}{\sqrt{12}}=0{,}906\ldots$ of circles.

4. It is quite possible that for dimensions greater than two the packing and covering constants of most centrally symmetric convex bodies can also be improved by suitable splittings. For example, consider the case of sphere packings in 3-space. In the usual cartesian (x, y, z)-coordinate system let B be the sphere defined by $x^2+y^2+z^2\leqq 1$, and let B_1, B_2 be the semispheres consisting of those points of B with $x\geqq 0$ and $x\leqq 0$, respectively. Furthermore, let L be the lattice generated by the vectors $(\sqrt{3}, 0, 0)$, $(0, 2, 0)$, $(\sqrt{3}/2, 0, 3/2)$, and define a translate L' of L by $(0, 1, 0)+L$. Then it is easy to prove that $(B_1+L)\cup(B_2+L')$ is a packing of semispheres of density $\frac{4\pi}{9\sqrt{3}}=0{,}806\ldots$. Since is has been shown by ROGERS [5], [6] that the packing constant of B is not grenter than 0,78 it follows that the packing constant of semispheres is larger than the packing constant of spheres.

REFERENCES

[1] FEJES-TÓTH, L.: *Regular Figures*. Pergamon Press, 1964.
[2] FEJES-TÓTH, L.: On the arrangement of houses in a housing estate. *Studia Sci. Math. Hungar.* **2** (1967), 37—42.
[3] FEJES-TÓTH, L.: Lencsék legsűrűbb elhelyezése a síkon. *Math. Lapok* **22** (1971), 209—213.
[4] FEJES-TÓTH, L.: *Lagerungen in der Ebene auf der Kugel und im Raum* (Zweite Auflage). Springer-Verlag, 1972.
[5] ROGERS, C. A.: The packing of equal spheres. *Proc. Lond. Math. Soc.* **8** (1958), 309—620.
[6] ROGERS, C. A.: *Packing and covering*. Cambr. Univ. Press, 1964.

H. Groemer, Department of Mathematics, The University of Arizona, Tucson, Arizona 85721, U.S.A.

A. Heppes, Vércse u. 24/A Budapest, 1124. Hungary

(Received May 20, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 191—195.

ON ESTIMATIONS OF JACKSON AND TIMAN TYPE

by

P. VÉRTESI

1. Introduction and preliminary results

1.1. Our starting point is a problem raised by R. DEVORE. "Do there exist polynomials $q_0, q_1, \ldots, q_n$ of degree $\leqq n$ such that the linear operators

$$L_n(f;x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) q_k(x)$$

provide the Jackson estimates

$$\|f(x) - L_n(f;x)\| = O(1)\,\omega\left(f;\frac{1}{n}\right)$$

where $\|\cdot\|$ is the sup norm on [0, 1] ?" (See [1], p. 581.) In his paper [2] J. SZABADOS gave an affirmative answer for this question applying intermediate approximating rational functions and then using a standard polynomial approximation to these rational functions.

Now we prove a theorem from which we can easily get another positive answer for the above mentioned problem. Our solution is constructive and we do not need any intermediate auxiliary functions. Further, we shall obtain a certain connection between the degree of the pointwise approximation and the minimal distance of the nodes. We use the following result from [3]:

1.2. Let $|x| \leqq 1$, $x = \cos\vartheta$ further

$$t_{k,n} = \cos\vartheta_{k,n} \quad \text{with} \quad \vartheta_{k,n} = \frac{2k\pi}{2n+1}, \tag{1.1}$$

$$L_{k,n}(x) = \frac{\sin\frac{2n+1}{2}(\vartheta - \vartheta_{k,n})}{(2n+1)\sin\frac{\vartheta - \vartheta_{k,n}}{2}}. \tag{1.2}$$

The fundamental Lagrange polynomials corresponding to $x_{k,n}$ for $k = \overline{0, n}$ (which means $k = 0, 1, \ldots, n$) are

$$\begin{cases} l_{0,n}(x) = L_{0,n}(x), \\ l_{k,n}(x) = L_{k,n}(x) + L_{-k,n}(x) \quad (k = \overline{1, n}). \end{cases} \tag{1.3}$$

If

$$u_{k,n}(x) = 4L^3_{k,n}(x) - 3L^4_{k,n}(x) \quad (k = \overline{-n, n}),$$

$$\begin{cases} p_{0,n}(x) = u_{0,n}(x), \\ p_{k,n}(x) = u_{k,n}(x) + u_{-k,n}(x) \quad (k = \overline{1, n}), \end{cases}$$

then for the polynomials

$$I_n(f;x) \overset{\text{def}}{=} \sum_{k=0}^{n} f(t_{k,n}) p_{k,n}(x)$$

of degree $\leqq 4n$ we have for each $f \in C[-1, 1]$ ($=f$ is continuous on $[-1, 1]$)

THEOREM 1.1. (O. KIS—P. VÉRTESI). If $f \in C[-1, 1]$ *then*

$$I_n(f; t_{k,n}) = f(t_{k,n}) \quad (k = \overline{0, n};\ n = 1, 2, 3, \ldots) \tag{1.4}$$

further

$$|I_n(f;x) - f(x)| = O(1)\left[\omega\left(f; \frac{\sqrt{1-x^2}}{n}\right) + \omega\left(f; \frac{1}{n^2}\right)\right] \quad (n = 1, 2, 3, \ldots) \tag{1.5}$$

(see [3]).

2. New results

Let us define the sequence of real numbers φ_n as follows

$$1 \leqq \varphi_n \leqq n, \quad \varphi_n \leqq \varphi_{n+1} \quad (n = 1, 2, 3, \ldots). \tag{2.1}$$

We have the following

THEOREM 2.1. *For any fixed sequence* $\{\varphi_n\}$ *defined by* (2.1) *we have a node-system*

$$-1 < x_{n_1, n, \varphi_n} < x_{n_1 - 1, n, \varphi_n} < \ldots < x_{1, n, \varphi_n} < 1 \quad (n_1 \leqq n+1;\ n = 1, 2, 3, \ldots)$$

and polynomials $P_{k, n, \varphi_n}(x)$ $(k = \overline{1, n_1};\ n = 1, 2, 3, \ldots)$ *of degree* $\leqq 4n$ *such that for each* $f \in C[-1, 1]$ *and the operators*

$$L_{n, \varphi_n}(f; x) = \sum_{k=1}^{n_1} f(x_{k, n, \varphi_n}) P_{k, n, \varphi_n}(x)$$

$$|L_{n, \varphi_n}(f;x) - f(x)| = O(1)\left[\omega\left(f; \frac{\sqrt{1-x^2}}{n}\right) + \omega\left(f; \frac{\varphi_n}{n^2}\right)\right] \quad (n = 1, 2, 3, \ldots) \tag{2.2}$$

where

$$\min_{2 \leqq k \leqq n_1} (x_{k-1, n, \varphi_n} - x_{k, n, \varphi_n}) \sim \frac{\varphi_n}{n^2} \quad (n = 1, 2, 3, \ldots).$$

The special case of this theorem is a solution of the DEVORE's problem.

THEOREM 2.2. *We can choose the sequence* $\{\varphi_n\}$ *and the polynomials* $q_{k,n}(x)$ *of degree* $\leqq 4n$ *such that for each* $f \in C[-1, 1]$

$$|f(x) - \sum_{k=0}^{n} f\left(-1 + \frac{2k}{n}\right) q_{k,n}(x)| = O(1)\,\omega\left(f; \frac{1}{n}\right) \quad (n = 1, 2, 3, \ldots). * \tag{2.3}$$

---

* To have an exact answer we remark that by (2.3)

$$|f(x) - \sum_{j=0}^{4n} f\left(-1 + \frac{2j}{4n}\right) Q_{j, 4n}(x)| = O(1)\,\omega\left(f; \frac{1}{4n}\right)$$

if

$$Q_{j, 4n}(x) = q_{j,n}(x) \quad (j = 0, 4, 8, \ldots, 4n)$$

and equal to 0 otherwise.

Our theorems are the best possible in the following sense. Let $C(\omega_m)=\{f(x);\ \omega_m(f;t)\leqq \leqq a_m(f)\omega_m(t)\}$ where $\omega_m(0)=0$, $\omega_m(T)\geqq\omega_m(t)$ if $T\geqq t$, $\omega_m(t)$ is continuous, $\dfrac{t^m}{\omega_m(t)}$ is monotone increasing ($t\geqq 0$), $\omega_m(t)$ is the m-th modulus of smoothness of $f(x)$ on $[-1, 1]$, $m\geqq 1$ is a fixed integer, $\omega_1=\omega$. We have

THEOREM 2.3. *If* $\{\varphi_n\}$, $\{x_{k,n,\varphi_n}\}$ *and* L_{n,φ_n} *are as above then for any fixed* $\omega(t)$ *we have a funtion* $f_1\in C(\omega)$ *and a suitable subsequence* $\{r_i\}$ *such that*

$$|L_{n,\varphi_n}(f_1;1)-f_1(1)|>\omega\left(\frac{\varphi_n}{n^2}\right)\quad (n=r_1, r_2, r_3, \ldots). \tag{2.4}$$

By our theorems we can see that the pointwise convergence of the operator $\sum_{k=1}^{n_1} f(x_{k,n})\,p_{k,n}(x)$ generally strongly depends on the minimal distance of the nodes (see further 4.2).

3. Proofs

3.1. PROOF of Theorem 2.1.

Let

$$d_n(\varphi_n)=\frac{2\varphi_n}{n^2}\quad (n=1, 2, 3, \ldots) \tag{3.1}$$

and

$$s_{k,n,\varphi_n}=1-kd_n(\varphi_n)\quad (k=\overline{1,n})\quad \text{if}\quad \varphi_n\neq O(1). \tag{3.2}$$

Let further for $\varphi_n\neq O(1)$

$$\begin{cases} T_+ = \{t_{k,n};\ t_{k,n}\geqq 0,\ k\geqq \varphi_n\},\\ T_- = \{t_{k,n};\ -t_{k,n}\in T_+\},\\ S_+ = \{s_{k,n,\varphi_n};\ \sup(\{0\}\cup T_+)\leqq s_{k,n,\varphi_n}-d_n(\varphi_n)\},\\ S_- = \{s_{k,n,\varphi_n};\ -s_{k,n,\varphi_n}\in S_+\}\quad (\varphi_n\neq O(1)). \end{cases} \tag{3.3}$$

If $T\overset{\text{def}}{=}\{t_{k,n}\}_{k=0}^{n}$ we define our nodes as follows

$$\{x_{j,n,\varphi_n}\}_{j=1}^{n_1}=\begin{cases} T & \text{if}\quad \varphi_n=O(1),\\ T_+\cup T_-\cup S_+\cup S_- & \text{if}\quad \varphi=O(1). \end{cases} \tag{3.4}$$

Supposing that $x_{j,n,\varphi_n}>x_{i,n,\varphi_n}(1\leqq j<i\leqq n_1)$, we have

$$d_n(\varphi_n)\sim \min_k (x_{k,n,\varphi_n}-x_{k+1,n,\varphi_n}). \tag{3.5}$$

Indeed, by (3.1)—(3.4) we have to investigate only the case when x_k, $x_{k+1}\in T_+\cup T_-$, $\varphi_n\neq O(1)$. (Sometimes we omit the superfluous indices.)

By (1.1) and (2.4)

$$\min_{\varphi_n \leqq k \leqq \frac{n}{2}} (x_k - x_{k+1}) \sim \cos\frac{2\varphi_n\pi}{2n+1} - \cos\frac{2(\varphi_n+1)\pi}{2n+1} = \tag{3.6}$$

$$= 2\sin\frac{2\varphi_n+1}{2n+1}\pi \sin\frac{1}{2n+1}\pi \sim \frac{\varphi_n}{n^2}.$$

Let us define the polynomials $P_{j,n,\varphi_n}(x)$ as follows. $P_{j+1,n,\varphi_n}(x)=p_{j,n}(x)$ $(j=\overline{0,n})$ if $\varphi_n=O(1)$. On the other hand

$$P_{j,n,\varphi_n}(x) = \begin{cases} p_{k,n}(x) & \text{if} \quad x_j \in T_+ \cup T_- \quad \text{and} \quad x_{j,n} = t_{k,n}; \\ \sum\limits_{k\in K_j} p_{k,n}(x) & \text{if} \quad x_{j,n} \in S_+ \cup S_- \\ \text{where} \begin{cases} k\in K_j & \text{if} \quad x_{j+1,n} < t_{k,n} \leqq x_{j,n} \quad (j>1), \\ k\in K_1 & \text{if} \quad x_{2,n} < t_{k,n} < 1 \quad (\varphi_n \neq O(1)). \end{cases} \end{cases} \tag{3.7}$$

Now we can prove our theorem by the operator

$$L_{n,\varphi_n}(f;x) = \sum_{j=1}^{n_1} f(x_{j,n,\varphi_n}) P_{j,n,\varphi_n}(x). \tag{3.8}$$

Indeed, if $\varphi_n=O(1)$ then $L_{n,\varphi_n}\equiv I_n$ and we have our theorem by (1.5). If $\varphi_n\neq O(1)$ then

$$|f(x)-L_{n,\varphi_n}(f;x)| \leqq |f(x)-I_n(f;x)| + |I_n(f;x)-L_{n,\varphi_n}(f;x)| =$$

$$= O(1)\left[\omega\left(f;\frac{\sqrt{1-x^2}}{n}\right)+\omega\left(f;\frac{1}{n^2}\right)\right] + |I_n(f;x)-L_{n,\varphi_n}(f;x)|.$$

Here by (3.1) — (3.4)

$$|I_n(f;x)-L_{n,\varphi_n}(f;x)| = \left|\sum_{k=0}^{n} f(t_k)p_k(x) - \sum_{j=1}^{n_1} f(x_j)P_j(x)\right| =$$

$$= \left|\sum_{x_j\in S_+\cup S_-} [f(x_j)P_j(x) - \sum_{k\in K_j} f(t_k)p_k(x)]\right| = O(1)\omega(f;d_n)\sum_{k=0}^{n}|p_k(x)|$$

from where we get our statement considering that $\sum\limits_{k=0}^{n}|p_{k,n}(x)|=O(1)$ (see [3]).

3.2. PROOF of Theorem 2.2. We have to choose $\varphi_n=n$ and $P_0(x)=P_n(x)\equiv 0$.

3.3. PROOF of Theorem 2.3. By (3.7), (3.8) and (1.2)

$$\lambda_n(1) \overset{\text{def}}{=} \sum_{j=1}^{n_1}|P_{j,n,\varphi_n}(1)| = \sum_{k=0}^{n}|p_{k,n}(1)| = p_{0,n}(1) = 1.$$

Using that $1-x_{1,n,\varphi_n}\sim d_n(\varphi_n)\sim\varphi_n\cdot n^{-2}$ we have our statement by [4], Theorem 1 if $\lim\limits_{t=+0}(t/\omega(t))=0$. If $\omega(t)\sim t$ then we can apply [5], Lemma 2.1. choosing $g_n(x_2)=0$, $g_n(x_1)=1$ and $g_n((1+x_1)/2)=0$. Using [6], Theorem 2.1, we have the desired result. We omit the details.

4. Remarks

4.1. There are methods to prove the (1.5) Timan estimation using node-systems similar to (1.1). We can apply those to gain analogue theorems.

4.2. We can generalize our Theorem 2.3. Here is a possibility.

If $\{x_{k,n}\}_{k=1}^{n} \subset (-1, 1)$ is a point system for which $\min_k |x_{k+1,n} - x_{k,n}| = d_n$, $|1 - x_{1,n}| = d_n$, $L_n(f; x) = \sum_{k=1}^{n} f(x_{k,n}) l_{k,n}(x)$ with $l_{k,n}(x) \in C[-1, 1]$ and $\sum_{k=1}^{n} |l_{k,n}(1)| = O(1)$, then supposing that $\lim_{t=+0} (t/\omega(t)) = 0$ we have for a suitable $f_2 \in C(\omega)$ and $\{n_i\}$

$$L_n(f; 1) - f(1) \succ \omega(d_n) \quad (n = n_1, n_2, n_3, \ldots)$$

(see [4] and [5]).

Another generalization is settled in [7].

REFERENCES

[1] *Linear Operators and Approximation* II. Proceedings of the Conference held at Oberwolfach, 1974. Birkhäuser Verlag (Basel-Stuttgart, 1974).

[2] SZABADOS, J.: On a problem of R. DeVore, *Acta Math. Acad. Sci. Hungar.* **27** (1975).

[3] KIS, O., VÉRTESI, P.: On a new interpolatory process (Russian). *Annales Univ. Sci. Budapest Sectio Math.* **10** (1967), 117—128.

[4] VÉRTESI, P.: On certain linear operators. V. *Acta Math. Acad. Sci. Hungar.*, **23** (1972), 433—437.

[5] VÉRTESI, P.: On certain linear operators. VIII, *ibid.* **25** (1974), 171—187.

[6] VÉRTESI, P.: On certain linear operators. VII, *ibid* **25** (1974), 67—80.

[7] VÉRTESI, P.: On a problem of J. Szabados, *ibid*, **28** (1976).

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received June 3, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 197—202.

A PROOF OF THE SIEGEL LINEARIZATION THEOREM BY DILIBERTO BOUNDED DOMINANTS

by

BUCK WARE

In 1952 CARL SIEGEL proved his celebrated linearization theorem for analytic vectorfields: he showed that certain diophantine inequalities on the eigenvalues of the linear part are sufficient for a finite dimensional system of ordinary differential equations to admit an analytic linearization in the vicinity of a stationary point. In the 1970 volume of this journal IMRE BIHARI and Á. ELBERT generalized SIEGEL's result in the two dimensional case by allowing the possibility of relations among the eigenvalues. The problem of relations was also treated by ALEKSANDR BRJUNO in his long 1971 paper.

Recently a number of people have succeeded in proving infinite dimensional versions of various well known theorems. For instance, CHARLES PUGH [1969] and JACOB PALIS [1968] independently proved the C° linearization theorem of PHILIP HARTMAN and DAVID GROBMAN for any Banach space, and MICHAEL IRWIN extended the stable manifold theorem to Banach spaces in 1970.

The proof given here of the SIEGEL linearization theorem in a Banach space is an adaptation of a proof given by STEPHEN DILIBERTO in 1970 for the finite dimensional case. The method of DILIBERTO bounded dominants is different from the popular accelerated convergence technique developed by JOHN NASH, ANDREI KOLMOGOROV, VLADIMIR ARNOL'D, JÜRGEN MOSER, et al.

1. Notation. Let E, F be two complex Banach spaces. On the space $\mathfrak{L}_s^k(E, F)$ of all continuous symmetric k-linear maps

$$u\colon E^k \to F$$

we have the multilinear norm

$$|u| = \sup\{|u(x_1, \ldots, x_k)| : |x_j| = 1,\ x_j \in E,\ 1 \leqq j \leqq k\}$$

Restriction to the diagonal embeds $\mathfrak{L}_s^k(E; F)$ into the space of continuous functions from E to F by

$$u \mapsto (x \mapsto u(x^k)), \quad x \in E,$$

and the image is called the space of k-homogeneous polynomials on E with values in F, written $\mathfrak{P}_k(E; F)$. On this space we still use the multilinear norm. If $\mathfrak{p}$ is in $\mathfrak{P}_k(E, F)$, its polar is the unique symmetric k-linear map $\tilde{\mathfrak{p}}$ such that

$$\mathfrak{p}(x) = \tilde{\mathfrak{p}}(x^k)$$

for x in E, and its derivative at x in E is

$$(D\mathfrak{p})_x \cdot a = k\tilde{\mathfrak{p}}(a, x^{k-1})$$

for a in E.

By a formal powerseries on E with coefficients in F is meant a sequence $f=(\mathfrak{p}_k)_{k\in\mathbf{N}}$ with $\mathfrak{p}_k$ a k-homogeneous polynomial on E with values in F. Its strict majorant is the formal powerseries

$$\bar{f} = (|\mathfrak{p}_k|\varepsilon^k)_{k\in\mathbf{N}}$$

on $\mathbf{R}$ with coefficients in $\mathbf{R}$, where $\varepsilon^k: x\mapsto x^k$ for x in $\mathbf{R}$. (If r is the radius of convergence of f and $\bar{r}$ that of $\bar{f}$, then $r/e\leqq\bar{r}\leqq r$.) If $g=(\beta_k\varepsilon^k)_{k\in\mathbf{N}}$ is a formal powerseries on $\mathbf{R}$ with coefficients in $\mathbf{R}$, then the notation $f\prec g$ means that $|\mathfrak{p}_k|\leqq\beta_k$ for k in $\mathbf{N}$.

2. Formal Solution. Let E be a Banach space and let $f=u+f_2$ be a formal powerseries on E with coefficients in E, where u is a toplinear automorphism of E and the powerseries f_2 has order $\geqq 2$. For each integer $k\geqq 2$ define the linear endomorphism

$$\Phi_k(u)\colon \mathfrak{P}_k(E,E) \to \mathfrak{P}_k(E,E)$$

on k-homogeneous polynomials $\mathfrak{q}$ by

$$\Phi_k(u)\colon \mathfrak{q} \mapsto (D\mathfrak{q})\cdot u$$

where $(D\mathfrak{q})\cdot u$ is the polynomial map defined by

$$(D\mathfrak{q})\cdot u\colon x \mapsto (D\mathfrak{q})_x \cdot u(x)$$

for x in E.

Now consider the endomorphism $u_*-\Phi_k(u)$ defined by

$$u_*-\Phi_k(u)\colon \mathfrak{q} \mapsto u\cdot\mathfrak{q}-(D\mathfrak{q})\cdot u.$$

PROPOSITION. Suppose that there are no relations in the sense that $u_*-\Phi_k(u)$ is invertible for each $k\geqq 2$. Then there is a formal powerseries φ on E with coefficients in E which has no constant term, has linear term id_E, and which linearizes f in the sense that

$$(u+f_2)\circ\varphi = D\varphi\cdot u.$$

We define recursively, for each integer $k\geqq 2$, an homogeneous polynomial $\mathfrak{p}_k$ of degree k and a formal powerseries f_{k+1} of order $\geqq k+1$. Suppose that f_k has already been defined and define $\mathfrak{p}_k$ by

$$u\mathfrak{p}_k - D\mathfrak{p}_k\cdot u = (f_k)_k,$$

that is,

$$\mathfrak{p}_k = \big(u_*-\Phi_k(u)\big)^{\leftarrow}\cdot(f_k)_k,$$

where $(f_k)_k$ denotes the homogeneous component of degree k of f_k. Then define f_{k+1} by

$$f_{k+1} = (id-D\mathfrak{p}_k)^{\leftarrow}\cdot\big(f_k\circ(id-\mathfrak{p}_k)-(f_k)_k\big),$$

where $(id-D\mathfrak{p}_k)^{\leftarrow}$ stands for the formal powerseries

$$id_E + D\mathfrak{p}_k + (D\mathfrak{p}_k)\cdot(D\mathfrak{p}_k) + \ldots$$

on E with coefficients in End (E).

We also define the nonhomogeneous polynomial maps φ_k recursively by

$$\varphi_1 = id_E, \quad \varphi_k = \varphi_{k-1}\circ(id-\mathfrak{p}_k),$$

and denote by $\varphi=\lim \varphi_k$ the formal powerseries whose homogeneous component $(\varphi)_j$ of degree j is

$$(\varphi)_j = (\varphi_k)_j$$

for all $k \geqq j$.

We claim that φ linearizes the formal vectorfield $u+f_2$. Well, we've arranged our definitions so that we can prove by induction that

$$(u+f_2)\circ\varphi_k = (D\varphi_k)\cdot(u+f_{k+1}).$$

The equation being trivial for $k=1$, we assume it holds for $k-1$ and compute, using the definitions of φ_k, $\mathfrak{p}_k$, f_{k+1}, that

$$\begin{aligned}
(u+f_2)\circ\varphi_k &= \big((u+f_2)\circ\varphi_{k-1}\big)\circ(id-\mathfrak{p}_k) = \\
&= \big((D\varphi_{k-1})\cdot(u+f_k)\big)\circ(id-\mathfrak{p}_k) = \\
&= (D\varphi_{k-1})_{id-\mathfrak{p}_k}\cdot\big(u-u\mathfrak{p}_k+f_k\circ(id-\mathfrak{p}_k)\big) = \\
&= (D\varphi_{k-1})_{id-\mathfrak{p}_k}\cdot\big(u-\big((f_k)_k+(D\mathfrak{p}_k)\cdot u\big)+f_k\circ(id-\mathfrak{p}_k)\big) = \\
&= (D\varphi_{k-1})_{id-\mathfrak{p}_k}\cdot\big(u-(D\mathfrak{p}_k)\cdot u+(id-D\mathfrak{p}_k)\cdot f_{k+1}\big) \\
&= (D\varphi_{k-1})_{id-\mathfrak{p}_k}\cdot(id-D\mathfrak{p}_k)\cdot(u+f_{k+1}) = \\
&= D\varphi_k\cdot(u+f_{k+1}).
\end{aligned}$$

But the homogeneous component of $(u+f_2)\circ\varphi$ of degree $j\geqq 2$ is the same as the component of $(u+f_2)\circ\varphi_k$ for any $k\geqq j$, which by what we've just shown is the component of $(D\varphi_k)\cdot(u+f_{k+1})$ of degree j. Since f_{k+1} is of order $\geqq k+1$, it is even the same as the component of $D\varphi_k\cdot u$, which by the nature of φ is the j-th component of $D\varphi\cdot u$.

3. Convergence. We retain the notation of section 2 and prove the following

THEOREM. *Suppose that the least majorant $\bar{f}$ of f defines an analytic function with radius of convergence greater than $r_2>0$. Assume that the linear part $u=(Df)_0$ of f is non-Liouville in the sense that it satisfies the (additive) Siegel small divisor diophantine inequalities: there are constants $c>0$, $\nu>0$ such that*

$$|(u_*-\Phi_k(u))^{\leftarrow}| \leqq ck^{\nu}$$

for each $k\geqq 2$. Then the formal powerseries φ has radius of convergence $\geqq$

$$r_\infty = \exp(-\beta\pi^2/6),$$

where

$$\beta = \nu+4+\log\big(cr_2^{-1}\bar{f}_2(r_2)\big).$$

PROOF. Define recursively the radii r_k, $k\geqq 2$, by

$$r_{k+1} = r_k\cdot\exp\big(-\beta/(k-1)^2\big)$$

and observe that they converge monotonically down to r_∞. We claim that $\bar{f}_k$ has radius of convergence $>r_k$ and that we have the Diliberto bounded dominant relation

$$ck^{\nu+3}r_{k+1}^{k-1}r_k^{-k}\bar{f}_k(r_k) < 1.$$

This is satisfied for $k=2$ by our choice of β. Our task is to prove by induction that it holds for all k, so we assume it for $k\geqq 2$ and try to prove it for $k+1$.

Well, by the Siegel condition and the definition of $\mathfrak{p}_k$,

$$|\mathfrak{p}_k| = |(u_* - \Phi_k(u))^{\leftarrow} \cdot (f_k)_k| \leqq ck^{\nu}|(f_k)_k| \leqq$$

$$\leqq ck^{\nu} \sum_{j=0}^{\infty} |(f_k)_{k+j}| r_k^j = ck^{\nu} r_k^{-k} \overline{f_k}(r_k),$$

and so

$$|\mathfrak{p}_k| < k^{-3} r_{k+1}^{-k+1}$$

by the induction hypothesis. Hence

$$|(D\mathfrak{p}_k)|_x = k|\mathfrak{p}_k||x_k|^{k-1} \leqq$$

$$\leqq k(k^{-3} r_{k+1}^{-k+1}) r_{k+1}^{k-1} = k^{-2} < 1$$

for $|x|<r_{k+1}$, which implies in particular that $id_E-(D\mathfrak{p}_k)$ is invertible on the disk $r_{k+1}\mathbf{B}$ with inverse satisfying

$$\left|(id-(D\mathfrak{p}_k)_x^{\leftarrow}\right| \leqq (1-k^{-2})^{-1} < \exp\left((k^2-1)^{-1}\right)$$

for $|x| < r_{k+1}$.

We also observe that

$$|x-\mathfrak{p}_k(x)| \leqq |x|(1+|\mathfrak{p}_k||x|^{k-1}) \leqq$$

$$\leqq r_{k+1}(1+k^{-3}) < r_{k+1} e^{k^{-3}} =$$

$$= r_k \exp\left(k^{-3} - \beta/(k-1)^2\right) < r_k$$

(the exponent always being negative), so that $id-\mathfrak{p}_k$ sends the disk $r_{k+1}\mathbf{B}$ into $r_k\mathbf{B}$ and we can define the composition

$$f_k \circ (id-\mathfrak{p}_k).$$

These two facts already show that f_{k+1} is an analytic function defined on the disk $r_{k+1}\mathbf{B}$; it remains only to verify the DILIBERTO bounded dominance.

Write

$$f_k \circ (id-\mathfrak{p}_k) - (f_k)_k =$$

$$= \left((f_k)_k \circ (id-\mathfrak{p}_k) - (f_k)_k\right) + (f_k)_H \circ (id-\mathfrak{p}_k),$$

where $(f_k)_H = f_k - (f_k)_k$ denotes the sum of the higher order terms. For the first term we have, in view of Isaac Newton's binomial formula,

$$(f_k)_k(x-\mathfrak{p}_k(x)) - (f_k)_k(x) =$$

$$= \sum_{j=1}^{k} \binom{k}{j} (-1)^j (f_k)_k(x^{k-j}, \mathfrak{p}_k(x)^j),$$

and by taking multilinear norms we obtain the majorant relation

$$\overline{(f_k)_k \circ (id-\mathfrak{p}_k) - (f_k)_k} \prec$$

$$\prec \sum_{j=1}^{k} \binom{k}{j} |(f_k)_k| |\mathfrak{p}_k|^j \varepsilon^{j(k-1)+k} = |(f_k)_k|\left((\varepsilon + |\mathfrak{p}_k|\varepsilon^k)^k - \varepsilon^k\right).$$

Hence by the mean value theorem

$$\overline{(f_k)_k\circ(id-\mathfrak{p}_k)-(f_k)_k}(r_{k+1}) \leqq$$

$$\leqq |(f_k)_k|\big((r_{k+1}+|\mathfrak{p}_k|r_{k+1}^k)^k - r_{k+1}^k\big) \leqq |(f_k)_k|k(r_{k+1}+|\mathfrak{p}_k|r_{k+1}^k)^{k-1}\cdot|\mathfrak{p}_k|r_{k+1}^k \leqq$$

$$\leqq r_k^{-k}\overline{f_k}(r_k)kr_{k+1}^{2k-1}|\mathfrak{p}_k|(1+|\mathfrak{p}_k|r_{k+1}^{k-1})^{k-1} \leqq c^{-1}k^{-\nu-3}kk^{-3}r_{k+1}(1+k^{-3})^{k-1}.$$

For the other term we have

$$\overline{(f_k)_H\circ(id-\mathfrak{p}_k)} \prec \overline{(f_k)_H}\circ(\varepsilon+|\mathfrak{p}_k|\varepsilon^k)$$

and we use the Diliberto fact that (for $0\leqq t\leqq r_k$)

$$\overline{(f_k)_H}(t) \leqq (t^{k+1}/r_k^{k+1})\overline{(f_k)_H}(r_k) \leqq (t^{k+1}/r_k^{k+1})\overline{f_k}(r_k)$$

to obtain

$$\overline{(f_k)_H\circ(id-\mathfrak{p}_k)}(r_{k+1}) \leqq \overline{(f_k)_H}(r_{k+1}+|\mathfrak{p}_k|r_{k+1}^k) \leqq$$

$$\leqq \big((r_{k+1}^{k+1}(1+|\mathfrak{p}_k|r_{k+1}^{k-1})^{k+1})/r_k^{k+1}\big)\overline{f_k}(r_k) \leqq (r_{k+1}^2/r_k)c^{-1}k^{-\nu-3}(1+k^{-3})^{k+1}.$$

Combining these two inequalities yields

$$\overline{f_k\circ(id-\mathfrak{p}_k)-(f_k)_k}(r_{k+1}) \leqq$$

$$\leqq c^{-1}k^{-\nu-3}r_{k+1}\big((r_{k+1}/r_k)(1+k^{-3})^{k+1}+k^{-2}(1+k^{-3})^{k-1}\big) \leqq$$

$$\leqq c^{-1}k^{-\nu-3}r_{k+1}(1+k^{-3})^{k+1}(1+k^{-2}).$$

Now

$$f_{k+1} = (id-D_k)^{\leftarrow}\cdot\big(f_k\circ(id-\mathfrak{p}_k)-(f_k)_k\big)$$

so

$$\overline{f_{k+1}} \prec \overline{(id-D\mathfrak{p}_k)^{\leftarrow}}\cdot\overline{f_k\circ(id-\mathfrak{p}_k)-(f_k)_k}$$

and

$$\overline{f_{k+1}}(r_{k+1}) \leqq \overline{(id-D\mathfrak{p}_k)^{\leftarrow}}(r_{k+1})\overline{f_k\circ(id-\mathfrak{p}_k)-(f_k)_k}(r_{k+1}).$$

But

$$(id-D\mathfrak{p}_k)_x^{\leftarrow} = id+(D\mathfrak{p}_k)_x+(D\mathfrak{p}_k)_x^2+\ldots$$

so

$$\overline{(id-D\mathfrak{p}_k)^{\leftarrow}} = 1+k|\mathfrak{p}_k|\varepsilon^{k-1}+k^2|\mathfrak{p}_k|^2\varepsilon^{2k-2}+\ldots$$

and

$$\overline{(id-D\mathfrak{p}_k)^{\leftarrow}}(r_{k+1}) \leqq 1+k^{-2}+(k^{-2})^2+\ldots = (1+k^{-2})^{-1}.$$

Hence

$$\overline{f_{k+1}}(r_{k+1}) \leqq (1+k^{-2})^{-1}c^{-1}k^{-\nu-3}r_{k+1}(1+k^{-3})^{k+1}(1+k^{-2})$$

$$\leqq c^{-1}k^{-\nu-3}r_{k+1}\exp\big(k^{-2}(1+1/k)\big) \leqq c^{-1}k^{-\nu-3}r_{k+1}\exp(3/2k^2).$$

So now we can estimate that

$$c(k+1)^{\nu+3}r_{k+2}^k r_{k+1}^{-k-1}\overline{f_{k+1}}(r_{k+1}) \leqq$$

$$\leqq (1+k^{-1})^{\nu+3}(r_{k+2}/r_{k+1})^k\exp(3/2k^2) \prec$$

$$\prec \exp\big((\nu+3)/k-\beta/k+3/2k^2\big) \leqq \exp\big((1/k)(\nu+3+3/4-\beta)\big),$$

and our choice of β makes the exponent negative, completing the induction.

To complete the proof we need only verify that the sequence $(\varphi_k)_{k\geqq 1}$ converges to an analytic function. By the standard device of telescoping sums it is sufficient to show that the sequence $(\varphi_k-\varphi_{k-1})_{k\geqq 2}$ is absolutely summable. Well, by the mean value theorem we have, for $|x|\leqq r_{k+1}$,

$$|\varphi_k(x)-\varphi_{k-1}(x)| = \left|\varphi_{k-1}\left(x-\mathfrak{p}_k(x)\right)-\varphi_{k-1}(x)\right| = \|D\varphi_{k-1}\|_{[x,\,x-\mathfrak{p}_k(x)]}\cdot|\mathfrak{p}_k(x)| =$$

$$= \|id-D\mathfrak{p}_2\|_{r_3\mathbf{B}}\cdot\|id-D\mathfrak{p}_3\|_{r_4\mathbf{B}}\ldots\|id-D\mathfrak{p}_{k-1}\|_{r_k\mathbf{B}}\cdot|\mathfrak{p}_k|r_{k+1}^k \leqq$$

$$\leqq (1+2^{-2})(1+3^{-2})\ldots\left(1+(k-1)^{-2}\right)k^{-3}r_{k+1}^{-k-1}r_{k+1}^k <$$

$$< e^{-\pi^2/6}k^{-3}r_k < (r_2e^{-\pi^2/6})k^{-3},$$

which is all we need, by comparison.

So $(\varphi_k)_{k\geqq 1}$ converges uniformly on the disk $r_\infty\mathbf{B}$ to some function φ, and by the convergence theorem of Karl Weierstraß φ is analytic. But the definition of φ is that, for $j\geqq k$,

$$(D^k\varphi)_0 = (D^k\varphi_j)_0,$$

while on the other hand $(D^k\varphi)_0=\lim_{j\to\infty}(D^k\varphi_j)_0$, so the formal powerseries φ defines the analytic function φ and the proof is complete.

REFERENCES

BIHARI, I. and ELBERT, Á.: On the normalform of analytic differential equations in the neighborhood of a critical point (The case of the "saddle-point"). *Studia Scientiarum Mathematicarum Hungarica* **5** (1970) 337—351.

ВРЮНО АЛЕКСАНДР Д.: Аналитическая форма дифференциальных уравнений. *Труды Московского математического общества* **25** (1971) 119—262.

DILIBERTO, STEPHEN P. L.: A new method for establishing the existence of analytice functions I. The Siegel normal form theorem. Office of Naval Research Technical Report, Project NR 041 255 (1970).

IRWIN, MICHAEL C.: On the stable manifold theorem. *Bull. of the London Math. Soc.* **2** (1970) 196—198.

PALIS, JACOB JR.: On the local structure of hyperbolic points in Banach spaces. *Anais da Academia Brasileira de Ciências* **40** (1968) 263—266.

PUGH, CHARLES C.: On a theorem of P. Hartmann. *American Journal of Mathematics* **91** (1969) 363—367.

SIEGEL CARL L.: Über die Normalform analytischer Differentialgleichung in der Nähe einer Gleichgewichtslösung. *Nachrichten der Akademie der Wissenschaften in Göttingen, Mathematisch-physikalische Klasse*, Nr **5** (1952) 21—30.

University of California, Santa Cruz

(Received July 9, 1973)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 203—211.

COMPLEXITY OF LATTICE—CONFIGURATIONS

by

T. G. TARJÁN

Introduction

Let $Y=\{y_1, y_2, \ldots, y_r\}$ and $Z=\{z_1, z_2, \ldots, z_s\}$ be finite sets. Let us define the set $\mathscr{R}$ of the discrete rectangles:

$$\mathscr{R} = \{R: R = U\times V; U, V \neq \varnothing\ ; U\subset Y; V\subset Z\}$$

where $U\times V$ denotes the Cartesian product of U and V: $U\times V=\{(y, z): y\in U, z\in V\}$. For the power set $\mathscr{P}=2^{(Y\times Z)}=\{P: P\subset Y\times Z\}$ the relation $\mathscr{R}\subset\mathscr{P}$ holds. Let us denote the complexity of the rectangle $R=U\times V\in\mathscr{R}$ by $\pi(R)$, which is defined by $\pi(R)=$ $=|U|+|V|$.

Let us extend the complexity function π for $\mathscr{P}$ in the following way: $\pi(P)=$ $=\min\{\sum_{R\in\mathscr{Q}}\pi(R): \mathscr{Q}\subset\mathscr{R}, \bigcup_{R\in\mathscr{Q}} R=P\}$ if $P\in\mathscr{P}$. Further definitions: $\pi(r, s)=\max_{P\in\mathscr{P}}\pi(P)$, $\pi(r)=\pi(r, r)$.

These concepts arised in the switching theory (see e.g. [7]), but they have some interest in themselves, too.

The problem is to determine the values of $\pi(P)$ for certain sets $P\in\mathscr{P}$ and the value of $\pi(r, s)$.

Hencefort I shall try to determine the complexity for three types of sets P and I shall give an asymptotic lower bound for $\pi(r)$.

To the solution of the first problem I shall generalize a theorem of KATONA [2], in which he has proved a conjecture of EHRENFEUCHT and MYCIELSKI. Moreover, I draw two corollaries from this generalization, which do not follow from KATONA's theorem. In this case $\pi(P_1)\sim r\cdot\log_2 r$ holds.

In solving the second problem I use the idea of KATONA and SZEMERÉDI from the proof of Th. 1 in [3] and I get that $\pi(P_2)\sim r\cdot\log_2 r$.

The third problem is the complexity of the Hadamard-matrices [1]. This remained an open question. I shall give only the following lower bound: $\pi(P_3)\geqq$ $\geqq(r+1)\sqrt{r}$.

The forth problem is the computing of the value of $\pi(r)$. This is also an unsolved problem. Using LUPANOV's Th. 4 in [5] I shall give only the following asymptotic lower bound for $\pi(r)$ without proof: $\pi(r)\gtrsim\frac{r^2}{\log_2 r\cdot\log_2\log_2 r}$. Accordingly the complexity of P_1 and P_2 can not be maximal.

1. Complexity of diagonalless squares

Assume $Y=Z$ and consider the set $P_1=\{p:p=(y_i, y_j),\ i\neq j,\ 1\leqq i, j\leqq m\}$. What is the value of $\pi(P_1)$?

First I shall prove a lemma which is a generalization of KATONA's theorem in [2] and I shall draw two corollaries of it.

LEMMA 1. *Let* $A_1, A_2, \ldots, A_m$ *and* $B_1, B_2, \ldots, B_m$ *be subsets of* X $(|X|=n)$, *such that* $A_i\cap B_j\neq\emptyset$ *iff* $i\neq j$ $(1\leqq i, j\leqq m)$. *Then*

$$\sum_{i=1}^{m}\binom{|A_i|+|B_i|}{|A_i|}^{-1}\leqq 1 \quad \text{holds.}$$

PROOF. For arbitrary $i\neq j$, $1\leqq i$, $j\leqq m$ there exist distinct $x_k, x_l\in X$ for which $x_k\in A_i\cap B_j$ and $x_l\in A_j\cap B_i$ hold. Then there is no permutation of the elements $x_1, x_2, \ldots, x_n\in X$ for which all the elements of A_i could precede the elements of B_i and all the elements of A_j precede the elements of B_j.

There are $\binom{n}{|A_i|+|B_i|}|A_i|!|B_i|!(n-|A_i|-|B_i|)!$ permutations of the elements $x_1, x_2, \ldots, x_n\in X$ in which all the elements of A_i precede the elements of B_i. These permutations are different for different i's therefore

$$\sum_{i=1}^{m}\binom{n}{|A_i|+|B_i|}\cdot|A_i|!|B_i|!(n-|A_i|-|B_i|)!\leqq n!$$

holds which proves our lemma.

Take the particular case, when $B_i=A_i^c=X\setminus A_i$. In this case the assumptions of the lemma have the form $A_i\not\subset A_j$ $(i\neq j)$. That is, Lemma 1 gives the following

Corollary 1. (LUBELL—MESHALKIN—YAMAMOTO inequality [4], [6] and [8]) If $A_1, A_2, \ldots, A_m$ are subsets of an n-element set, and $A_i\not\subset A_j$ $(i\neq j)$ then

$$\sum_{i=1}^{m}\binom{n}{|A_i|}^{-1}\leqq 1.$$

Before the second corollary I shall prove a simple lemma.

LEMMA 2.

The function $f(x, y)=$

$$=\begin{cases}\dfrac{1-\langle x\rangle-\langle y\rangle}{\binom{[x]+[y]}{[x]}}+\dfrac{\langle y\rangle}{\binom{[x]+\{y\}}{[x]}}+\dfrac{\langle x\rangle}{\binom{\{x\}+[y]}{\{x\}}} & \text{if } \langle x\rangle+\langle y\rangle\leqq 1,\ \langle x\rangle\langle y\rangle>0\\ \dfrac{1-\prec x\succ-\prec y\succ}{\binom{\{x\}+\{y\}}{\{x\}}}+\dfrac{\prec x\succ}{\binom{[x]+\{y\}}{[x]}}+\dfrac{\prec y\succ}{\binom{\{x\}+[y]}{\{x\}}} & \text{if } 0\leqq\prec x\succ+\prec y\succ<1\end{cases}$$

is convex in the triangle $1\leqq x$, $1\leqq y$, $x+y\leqq n$, where

$$[z]=\max_{\substack{a\text{ integer}\\ a\leqq z}}a \qquad \{z\}=\min_{\substack{a\text{ integer}\\ a\geqq z}}a$$

$$\langle z\rangle=z-[z] \qquad \prec z\succ=\{z\}-z$$

Moreover in case of an arbitrary x_0, $1 \leqq x_0 \leqq n-1$ $f(x_0, y)$ *is a strictly monotonically decreasing function of* y *in* $[1, n-x_0]$. *Similarly,* $f(x, y_0)$ *is also a decreasing function of* x *in* $[1, n-y_0]$, *where* y_0 *is fixed in* $[1, n-1]$.

PROOF. Since $f(x, y)$ is a linear and continuous extension of $\binom{x+y}{x}^{-1}$, it is sufficient to show that

(i) for integer x_0, $1 \leqq x_0 \leqq n-1$, $f(x_0, y)$ is a convex function of y, $y \in [1, n-x_0]$
(ii) for integer y_0, $1 \leqq y_0 \leqq n-1$, $f(x, y_0)$ is a convex function of x, $x \in [1, n-y_0]$
(iii) for integer x_0, y_0, $2 \leqq x_0$, $2 \leqq y_0$, $x_0+y_0 \leqq n$,

$$f(x_0, y_0)-f(x_0, y_0-1)-f(x_0-1, y_0)+f(x_0-1, y_0-1) \geqq 0,$$

because for the plane

$$p(x, y) = f(x_0, y_0)+\frac{f(x_0+\varepsilon, y_0)-f(x_0, y_0)}{\varepsilon}(x-x_0)+$$

$$+\frac{f(x_0, y_0+\varepsilon)-f(x_0, y_0)}{\varepsilon}(y-y_0) \leqq$$

$$\leqq f(x_0, y_0)+f(x, y_0)-f(x_0, y_0)+f(x_0, y)-f(x_0, y_0)$$

holds ($\varepsilon = 1$ or -1) from (i) and (ii), and finally from (iii):

$$f(x, y)-p(x, y) \geqq f(x, y)-f(x_0, y)-f(x, y_0)+f(x_0, y_0) \geqq 0$$

is true and equality holds if (x, y) is equal to (x_0, y_0), $(x_0+\varepsilon, y_0)$ or $(x_0, y_0+\varepsilon)$.

(i) $$\frac{x_0!\,y!}{(x_0+y)!}-\frac{x_0!\,(y-1)!}{(x_0+y-1)!}=\frac{x_0!\,(y-1)!}{(x_0+y-1)!}\left(\frac{y}{x_0+y}-1\right)=-\frac{x_0!\,(y-1)!\,x_0}{(x_0+y)!},$$

$$-\frac{x_0!\,y!\,x_0}{(x_0+y+1)!}+\frac{x_0!\,(y-1)!\,x_0}{(x_0+y)!}=\frac{x_0!\,(y-1)!\,x_0}{(x_0+y)!}\times$$

$$\times\left(-\frac{y}{x_0+y+1}+1\right)=\frac{(x_0+1)!\,(y-1)!}{(x_0+y+1)!}x_0=x_0\binom{x_0+y+1}{x_0+1}^{-1}>0 \quad \text{if} \quad 2 \leqq y.$$

(ii) Similarly $$y_0\binom{x+y_0+1}{y_0+1}^{-1}>0 \quad \text{if} \quad 2 \leqq x.$$

(iii) $$\frac{x_0!\,y_0!}{(x_0+y_0)!}-\frac{(x_0-1)!\,y_0!}{(x_0+y_0-1)!}-\frac{x_0!\,(y_0-1)!}{(x_0+y_0-1)!}+\frac{(x_0-1)!)(y_0-1)!}{(x_0+y_0-2)!}=$$

$$=\frac{(x_0-1)!\,(y_0-1)!}{(x_0+y_0-2)!}\left(\frac{x_0y_0}{(x_0+y_0)(x_0+y_0-1)}-\frac{x_0}{x_0+y_0-1}-\frac{y_0}{x_0+y_0-1}+1\right)=$$

$$=\frac{x_0y_0-(x_0+y_0)^2+(x_0+y_0)^2-(x_0+y_0)}{\binom{x_0+y_0-2}{x_0-1}(x_0+y_0)(x_0+y_0-1)}=\frac{(x_0-1)(y_0-1)-1}{\binom{x_0+y_0-2}{x_0-1}(x_0+y_0)(x_0+y_0-1)} \geqq 0,$$

if $2 \leqq x_0$, $2 \leqq y_0$

From the proof of the convexity you can also see the monotonity of $f(x, y)$. The proof is completed.

Corollary 2. Let $A_1, A_2, \ldots, A_m$ and $B_1, B_2, \ldots, B_m$ be subsets of an n-element set, such that $A_i \cap B_j \neq \emptyset$ iff $i \neq j$. Then

$$m \leqq \left(f\left(\frac{\sum_{i=1}^{m} |A_i|}{m}, \frac{\sum_{i=1}^{m} |B_i|}{m} \right) \right)^{-1} \leqq \left(f\left(\max_{1 \leqq i \leqq m} |A_i|, \max_{1 \leqq i \leqq m} |B_i| \right) \right)^{-1}.$$

PROOF. Using Lemmas 2 and 1, we obtain

$$m \cdot f\left(\max_{1 \leqq i \leqq m} |A_i|, \max_{1 \leqq i \leqq m} |B_i| \right) \leqq mf\left(\frac{\sum_{i=1}^{m} |A_i|}{m}, \frac{\sum_{i=1}^{m} |B_i|}{m} \right) \leqq \sum_{i=1}^{m} f(|A_i|, |B_i|) \leqq 1,$$

which gives our statement.

LEMMA 3. *Let $U_1, U_2, \ldots, U_n$ and $V_1, V_2, \ldots, V_n$ be non-empty subsets of $Y = \{y_1, y_2, \ldots, y_m\}$ such that $P_1 = \bigcup_{k=1}^{n} (U_k \times V_k)$ holds, where*

$$P_1 = \{p : p = (y_i, y_j), i \neq j, 1 \leqq i, j \leqq m\}$$

Then for the unique λ, defined by $m = 1/f\left(\frac{\lambda}{2}, \frac{\lambda}{2}\right) : \sum_{k=1}^{n} (|U_k| + |V_k|) \geqq \lambda \cdot m$ holds, where $f(x, y)$ is defined in Lemma 2.

PROOF. The following two statements are equivalent

(i) $\bigcup_{k=1}^{n} (U_k \times V_k) = P_1 = \{p : p = (y_i, y_j), i \neq j, 1 \leqq i, j \leqq m\}$.

(ii) The sets $A_i = \{k : y_i \in U_k\}$ and $B_i = \{k : y_i \in V_k\}$ $(1 \leqq i \leqq m)$ have the property, that $A_i \cap B_j \neq \emptyset$ iff $i \neq j$.

(i → ii) If $i \neq j$, $1 \leqq i, j \leqq m$ then $p = (y_i, y_j)$ is an element of $P_1 = \bigcup_{k=1}^{n} (U_k \times V_k)$, that is, there exists a k, $1 \leqq k \leqq n$ for which $y_i \in U_k$ and $y_j \in V_k$. Thus $k \in A_i \cap B_j$ and so $A_i \cap B_j \neq \emptyset$. On the other hand, if $A_i \cap B_j \neq \emptyset$, then there exists a k, $1 \leqq k \leqq n$ for which $k \in A_i \cap B_j$ that is $p = (y_i, y_j)$ is an element of $(U_k \times V_k)$ which means that $p = (y_i, y_j) \in P_1$ and so $i \neq j$.

(ii → i) Assume $p = (y_i, y_j) \in P_1$, for some $1 \leqq i, j \leqq m$, $i \neq j$ follows. Thus, there exists a k, $1 \leqq k \leqq n$ for which $k \in A_i \cap B_j$. Then $p = (y_i, y_j)$ is an element of $(U_k \times V_k)$ which means that $p \in \bigcup_{k=1}^{n} (U_k \times V_k)$.

If $p = (y_i, y_j) \in \bigcup_{k=1}^{n} (U_k \times V_k)$ then there exists a k, $1 \leqq k \leqq n$ for which (y_i, y_j) is an element of $(U_k \times V_k)$. Thus $k \in A_i \cap B_j \neq \emptyset$ and so $i \neq j$, that is, $p \in P_1$.

Let us now suppose that $\sum_{k=1}^{n}(|U_k|+|V_k|)<\lambda m$ holds. Since $\sum_{k=1}^{n}(|U_k|+|V_k|)=$ $=\sum_{i=1}^{m}(|A_i|+|B_i|)$ is true, using the notation

$$\mu=\frac{\sum_{k=1}^{n}(|U_k|+|V_k|)}{m},$$

$$m=1/f\left(\frac{\lambda}{2},\frac{\lambda}{2}\right)>1/f\left(\frac{\mu}{2},\frac{\mu}{2}\right)\geqq\left(f\left(\frac{\sum_{i=1}^{m}|A_i|}{m},\frac{\sum_{i=1}^{m}|B_i|}{m}\right)\right)^{-1}\geqq m$$

hold because on the base of Lemma 2 $f(x, y)$ is decreasing in both x and y, just as $f(x, y)$ is convex and symmetric $(f(x, y)=f(y, x))$ and so $f\left(\frac{x+y}{2},\frac{x+y}{2}\right)\leqq f(x, y)$. Finally the last inequality follows from Corollary 2. This is a contradiction. The proof is completed.

THEOREM 1. *Put* $Y=Z$, $|Y|=|Z|=\left(\left[\frac{l}{2}\right]\right)=m$ *and* $P_1=\{p:p=(y_i, y_j), i\neq j,$ $1\leqq i, j\leqq m\}$. Then the complexity of $P_1:\pi(P_1)=l.\left(\left[\frac{l}{2}\right]\right)$.

PROOF. From Lemma 3 it follows $\pi(P_1)\geqq\lambda m$. This lower bound is the best possible if $\lambda=l$ is an integer. Namely let A_i's be all the $\left[\frac{l}{2}\right]$ element subsets of the set $X=\{1, 2, \ldots, l\}$ and $B_i=A_i^c=X\setminus A_i$. Thus, for A_i's and B_i's (ii) of Lemma 3 holds. In this way for

$$U_k=\{y_i:k\in A_i\},\quad V_k=\{y_i:k\in B_i\}$$

(i) $P_1=\bigcup_{k=1}^{l}(U_k\times V_k)$ also holds. Accordingly if $\lambda=l$ then

$$\lambda\cdot m=l\cdot\left(\left[\frac{l}{2}\right]\right)=l/f\left(\left[\frac{l}{2}\right],\left\{\frac{l}{2}\right\}\right)=l/f\left(\frac{l}{2},\frac{l}{2}\right)=\sum_{i=1}^{m}(|A_i|+|B_i|)=$$

$$=\sum_{k=1}^{l}(|U_k|+|V_k|)\geqq\pi(P_1)\geqq\lambda\cdot m,\quad\text{and so}\quad\pi(P_1)=\lambda\cdot m.$$

The proof is completeed

2. Complexity of isosceles right triangles

In case of $Y=Z$ and $P_2=\{p:p=(y_i, y_j),\ i<j,\ 1\leqq i, j\leqq m\}$ let us compute the value of $\pi(P_2)$, defined in our introduction.

First I shall prove a lemma which uses the idea of Th. 1. in [3] and I shall draw two conclusions of it.

LEMMA 4. *Let $A_1, A_2, \ldots, A_m$ and $B_1, B_2, \ldots, B_m$ be subsets of an n-element set X, such that $A_i \cap B_j \neq \varnothing$ iff $i<j$ $(1\leqq i, j\leqq m)$.*

Then $\sum_{i=1}^{m} 2^{-(|A_i|+|B_i|)} \leqq 1$ *holds.*

PROOF. For arbitrary pair $i<j$, $(1\leqq i, j\leqq m)$ $A_i \cap B_j \neq \varnothing$ holds. Then there is no $A \subset X$, for which both $A_i \subset A$, $A \cap B_i = \varnothing$ and $A_j \subset A$, $A \cap B_j = \varnothing$ hold, because otherwise $\varnothing \neq A_i \cap B_j \subset A \cap B_j = \varnothing$ would be true. Since there are $2^{(n-|A_i|-|B_i|)}$ sets A satisfying $A_i \subset A$ and $A \cap B_i = \varnothing$, thus $\sum_{i=1}^{m} 2^{(n-|A_i|-|B_i|)} \leqq 2^n$ holds which gives our lemma.

LEMMA 5. *Let $A_1, A_2, \ldots, A_m$ and $B_1, B_2, \ldots, B_m$ be subsets of an n-element set X such that $A_i \cap B_j \neq \varnothing$ iff $i<j$ $(1\leqq i, j\leqq m)$.*

Then $m \cdot \log_2 m \leqq \sum_{i=1}^{m} (|A_i| + |B_i|)$.

PROOF: Since 2^{-x} is a convex function and by Lemma 4

$$m \cdot 2^{-\frac{\sum_{i=1}^{m} (|A_i|+|B_i|)}{m}} \leqq \sum_{i=1}^{m} 2^{-(|A_i|+|B_i|)} \leqq 1 \text{ hold,}$$

which proves our lemma.

LEMMA 6. *Let $U_1, U_2, \ldots, U_n$ and $V_1, V_2, \ldots, V_n$ be non-empty subsets of $Y=\{y_1, y_2, \ldots, y_m\}$ such that*

$$\bigcup_{k=1}^{n} (U_k \times V_k) = P_2 = \{p:p=(y_i, y_j),\ i<j,\ 1\leqq i, j\leqq m\}.$$

Then $\sum_{k=1}^{m} (|U_k| + |V_k|) \geqq m \cdot \log_2 m$.

PROOF. The following two statements are equivalent:

(i) $\bigcup_{k=1}^{n} (U_k \times V_k) = P_2 = \{p:p=(y_i, y_j),\ i<j,\ 1\leqq i, j\leqq m\}$,

(ii) The sets $A_i=\{k: y_i \in U_k\}$ and $B_i=\{k:\ y_i \in V_k\}$ $(1\leqq i\leqq m)$ have the property, that $A_i \cap B_j \neq \varnothing$ iff $i<j$.

The proof of this equivalence is the same as it was in Lemma 3, the only difference is, that instead of $i \neq j$ you have to think of $i<j$.

Since $\sum_{k=1}^{n} (|U_k| + |V_k|) = \sum_{i=1}^{m} (|A_i| + |B_i|)$ is true, Lemma 5 gives our Lemma.

THEOREM 2. *Put* $Y=Z$, $|Y|=|Z|=2^l=m$ and $P_2=\{p:p=(y_i, y_j),\ i<j,\ 1\leqq i, j\leqq m\}$. *Then the complexity of* P_2: $\pi(P_2)=l\cdot 2^l$.

PROOF. From Lemma 6 it follows $\pi(P_2)\geqq m\cdot\log_2 m$. This lower bound is the best possible if $\log_2 m=l$ is an integer. Namely let $Y=\{0, 1\}^l$ and $E=\bigcup_{i=0}^{l-1}\{0, 1\}^i$ be. If $\varepsilon=(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)\in E$ let

$$W_{\varepsilon 0}=\left\{\delta\in\{0, 1\}^l: \delta=(\delta_1, \delta_2, \ldots, \delta_l),\ \delta_j=\varepsilon_j \quad (1\leqq j\leqq i),\ \delta_{i+1}=0\right\}$$

$$W_{\varepsilon 1}=\left\{\delta\in\{0, 1\}^l: \delta=(\delta_1, \delta_2, \ldots, \delta_l),\ \delta_j=\varepsilon_j \quad (1\leqq j\leqq i),\ \delta_{i+1}=1\right\} \text{ be.}$$

In this case $P_2=\bigcup_{\varepsilon\in E}(W_{\varepsilon 0}\times W_{\varepsilon 1})$ and $\sum_{\varepsilon\in E}|W_{\varepsilon 0}|+|W_{\varepsilon 1}|=l\cdot 2^l$ trivially hold. The proof is completed.

3. Complexity of Hadamard-matrices

Assume now, that $Y=Z=\{0, 1\}^l$ and $P_3=\{p:p=(y_i, y_j),\ y_i * y_j=0\}$ (where $y_i * y_j\equiv\sum_{k=1}^{l} y_{ik}\cdot y_{jk} \pmod 2$). Let us compute the value $\pi(P_3)$, defined in our introduction. Let us remark that P_3's are the Hadamard-matrices [1].

First I shall state a simple lemma and by means of it I shall give only a lower bound for $\pi(P_3)$. We introduce some notations:

If $U, V\subset Y$, $U * V=\{u * v:\ u\in U,\ v\in V\}$.

If $W\subset Y$, let $[W]$ be the subspace spanned by W, and finally $\dim W=\dim [W]$.

LEMMA 7. *Let* U, V *be subsets of the vector space* $Y=\{0, 1\}^l$ *satisfying* $U * V=0$. *Then* $\dim U+\dim V\leqq l$ holds.

PROOF. U has $\dim U$ linearly independent vectors $\{u_i\}_{i=1}^{\dim U}$. For arbitrary $x\in\{0, 1\}^l$ and $x * U=0$ the $x * u_i\equiv 0 \pmod 2$ $1\leqq i\leqq\dim U$ system of equations holds, which has $(l-\dim U)$ linearly independent solutions by the Cramer rule. The proof is completed.

THEOREM 3. *Put* $Y=Z=\{0, 1\}^l$ and $P_3=\{p:p=(y_i, y_j),\ y_i * y_j\equiv 0 \pmod 2\}$ (where $y_i * y_j\equiv\sum_{k=1}^{l} y_{ik}\cdot y_{jk} \pmod 2$). Then the complexity of P_3:

$$\pi(P_3)\geqq(2^l+1)2^{\frac{l}{2}}.$$

PROOF. Let $U_1, U_2, \ldots, U_n$ and $V_1, V_2, \ldots, V_n$ be non-empty subsets of Y, such that $P_3=\bigcup_{k=1}^{n}(U_k\times V_k)$. Then $\sum_{k=1}^{n}|U_k|\cdot|V_k|\geqq|P_3|$. Thus the following inequality holds:

$$(1)\qquad \sum_{k=1}^{n}(|U_k|+|V_k|)=\sum_{k=1}^{n}(|U_k|^{-1}+|V_k|^{-1})\cdot|U_k|\cdot|V_k|\geqq$$

$$\geqq\min_{1\leqq k\leqq n}(|U_k|^{-1}+|V_k|^{-1})\sum_{k=1}^{n}|U_k|\cdot|V_k|\geqq\min_{1\leqq k\leqq n}(|U_k|^{-1}+|V_k|^{-1})\cdot|P_3|.$$

On the other hand.

$$(2) \qquad |U_k|^{-1}+|V_k|^{-1} \geqq 2^{-\dim U_k}+2^{-\dim V_k} \geqq 2^{-\dim U_k}+2^{-l+\dim U_k} \geqq 2^{\left(1-\frac{l}{2}\right)}$$

follows from Lemma 7.

Let us prove now

$$(3) \qquad |P_3| = 2^{l-1}(2^l+1) \text{ by induction over } l.$$

If $l=1$, then $|P_3|=3$. Let us suppose that (3) is true for $l-1$. If y_{il} and y_{jl} are given and $y_{il}\cdot y_{jl}=0$ then $\sum_{k=1}^{l} y_{ik}\,y_{jk}\equiv 0 \pmod 2$ in $2^{l-2}(2^{l-1}+1)$ cases by our inductional assumption. If $y_{il}\cdot y_{jl}=1$ holds then $\sum_{k=1}^{l} y_{ik}\,y_{jk}\equiv 0 \pmod 2$ in $2^{2(l-1)}-2^{l-2}(2^{l-1}+1)$ cases. Since $y_{il}\,y_{jl}=0$ in 3 cases thus $|P_3|=3\cdot 2^{l-2}(2^{l-1}+1)+2^{2(l-1)}-2^{l-2}(2^{l-1}+1)=2^{l-1}(2^l+1)$ for l. From the formulas (1)–(3)

$$\sum_{k=1}^{n}(|U_k|+|V_k|) \geqq \min_{1\leqq k\leqq n}(|U_k|^{-1}+|V_k|^{-1})\,|P_3| \geqq 2^{1-\frac{l}{2}}2^{l-1}(2^l+1).$$

The proof is completed

4. Complexity of the most complex lattice-configurations

I shall formulate an asymptotic lower bound for $\pi(m)$ which is a simple consequence of LUPANOV's Th. 4 in [5].

THEOREM 4. $\pi(2^l) \gtrsim \dfrac{2^{2l}}{l\cdot\log_2 l}$, *where* $a_l \gtrsim b_l$ *iff* $\liminf\limits_l \dfrac{a_l}{b_l} \geqq 1$.

Open problems

The problems of Theorems 3 and 4, which are only partially solved, remained open questions:

1. Let

$$Y = \{y_1, y_2, \ldots, y_{2^l}\} = \{0,1\}^l$$

and

$$P_3 = \{p : p = (y_i, y_j),\ 1 \leqq i,j \leqq 2^l,\ y_i * y_j \equiv 0 \pmod 2\}$$

be where $y_i * y_j \equiv \sum_{k=1}^{l} y_{ik}\cdot y_{jk} \pmod 2$ and let us give a system $U_1, U_2, \ldots, U_n$; $V_1, V_2, \ldots, V_n$ for which $P=\bigcup_{k=1}^{n}(U_k\times V_k)$ and

$$\sum_{k=1}^{n}(|U_k|+|V_k|) = \text{minimal},$$

where $U_k\times V_k$ is the Cartesian product of U_k and V_k.

2. Let $Y=\{y_1, y_2, \ldots, y_m\}$ be and let us denote by $\pi(m)$ the minimal number that for every $P\subset(Y\times Y)$ there exists a system $U_1, U_2, \ldots, U_n$; $V_1, V_2, \ldots, V_n$ such that

$$P = \bigcup_{k=1}^{n} (U_k \times V_k) \text{ and}$$

$$\pi(m) \geqq \sum_{k=1}^{n} (|U_k| + |V_k|) \text{ hold,}$$

where $(U\times V)$ is the Cartesian product of U and V. What is the exact value of $\pi(m)$?

I am greatly indebted to G. O. H. KATONA for his helpful advices.

REFERENCES

[1] BERLEKAMP, E. R.: *Algebraic Coding Theory*, McGraw Hill, New York, 1969, 316—317.
[2] KATONA, G. O. H.: On a Conjecture of Ehrenfeucht and Mycielski, *J. Combinatorial Th.* (to appear).
[3] KATONA, G. and SZEMRÉDI, E.: On a Problem of Graph Theory, *Studia Sci. Math. Hungar.* **2** (1967), 23—28.
[4] LUBELL, D.: A Short Proof of Sperner's Lemma, *J. Combinatorial* Th. **1** (1966) 299.
[5] LUPANOV, O. B.: On Some Classes of Control Systems, "*Problemi Kybernetiki*" *M., Fizmat* **2** (1962) 63—97 (in Russian).
[6] MESHALKIN, L. D.: A Generalization of Sperner's Theorem on the Number of Subsets of a Finite Set, *Teor. Verojatnost. i Primenen.* **8** (1963) 219—220 (in Russian).
[7] TARJÁN, T. G.: On Complexity of Switching Circuits, *Problems of Control and Information Theory*, **3** (1974) 183—196.
[8] YAMAMOTO, K.: Logarithmic Order of Free Distributive Lattices, *J. Math. Soc. Japan*, **6** (1954) 343—353.

MTA Közgazdaságtud. Intézet (Institute of Economics, Hungarian Academy of Sciences), 1051 Budapest, Münnich F. u. 7.

(Received February 22, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 213—236.

EIN SIMPLEXARTIGER LÖSUNGSALGORITHMUS FÜR PSEUDOLINEARE OPTIMIERUNGSPROBLEME

von

HELGA HARTWIG

1. Einleitung

In der vorliegenden Arbeit soll ein Lösungsalgorithmus zu dem pseudolinearen Optimierungsproblem

$$(P) \qquad F(\mathbf{x}) = \text{Min!}, \quad \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geqq \mathbf{0}$$

entwickelt werden.

Der zulässige Bereich $B = \{\mathbf{x} \in \mathbf{R}^n | \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geqq \mathbf{0}\}$ von (P) sei eine nichtleere konvexe polyedrische Menge, über der die Zielfunktion $F(\mathbf{x})$ pseudolinear (d.h. zugleich pseudokonvex und pseudokonkav) ist. Die wichtigsten Spezialfälle der pseudolinearen Optimierung sind neben den linearen die hyperbolischen Optimierungsaufgaben.

Die Funktion $F(\mathbf{x})$ ist genau dann pseudolinear über B, wenn sie über einer B enthaltenden offenen Menge stetig differenzierbar ist und ferner für beliebige $\mathbf{x}^1, \mathbf{x}^2 \in B$ gilt

$$(1.1) \qquad (\mathbf{x}^1 - \mathbf{x}^2)^\top \operatorname{grad} F(\mathbf{x}^2) = 0 \Rightarrow F(\mathbf{x}^1) = F(\mathbf{x}^2).$$

In [2] wird gezeigt, daß bei einer über B pseudolinearen Funktion $F(\mathbf{x})$ für beliebige $\mathbf{x}^1, \mathbf{x}^2 \in B$ sogar erfüllt ist

$$(1.2) \qquad (\mathbf{x}^1 - \mathbf{x}^2)^\top \operatorname{grad} F(\mathbf{x}^2) \begin{Bmatrix} > \\ = \\ < \end{Bmatrix} 0 \Leftrightarrow F(\mathbf{x}^1) \begin{Bmatrix} > \\ = \\ < \end{Bmatrix} F(\mathbf{x}^2).$$

Eine pseudolineare Funktion ist stets auch streng quasikonvex und streng quasikonkav. Für das Problem (P) gilt deshalb nach MANGASARIAN [3]

1. Jedes lokale Minimum von $F(\mathbf{x})$ über B ist bereits ein globales Minimum.

Weiterhin haben wir nach MARTOS [5] bzw. als Folgerung aus einem allgemeineren Satz bei STOER/WITZGALL [6]

2. $F(x)$ nimmt sein Minimum über B (falls es existiert) in einer Ecke von B an.

Ist der zulässige Bereich B unbeschränkt, so ist auch die folgende, in [2] bewiesene Aussage interessant:

3. Nimmt $F(x)$ sein Infimum μ über B nicht (im Endlichen) auf B an, dann existiert eine unbeschränkte Kante von B, über der $F(x)$ den Grenzwert μ besitzt.

Der bekannteste Spezialfall einer pseudolinearen Funktion ist die gebrochen lineare Funktion

$$(1.3) \qquad F(\mathbf{x}) = \frac{\mathbf{z}^\top \mathbf{x} + z_0}{\mathbf{n}^\top \mathbf{x} + n_0}$$

mit $\mathbf{z}, \mathbf{n} \in \mathbf{R}^n$ und $z_0, n_0 \in \mathbf{R}^1$, die pseudolinear über jeder konvexen Menge $K \subset \subset \{\mathbf{x} \in \mathbf{R}^n | \mathbf{n}^\top \mathbf{x} + n_0 \neq 0\}$ ist. Es lassen sich aber auch andere Beispielklassen pseudolinearer Funktionen angeben. So ist, wie in [2] gezeigt wird,

$$F(\mathbf{x}) = \frac{l_2(\mathbf{x})}{l_3(\mathbf{x})} + \sqrt{\frac{l_2^2(\mathbf{x})}{l_3^2(\mathbf{x})} + \frac{l_1(\mathbf{x})}{l_3(\mathbf{x})}} \tag{1.4}$$

mit beliebigen affin linearen Funktionen $l_1(\mathbf{x})$, $l_2(\mathbf{x})$ und $l_3(\mathbf{x})$ pseudolinear über jeder konvexen Menge

$$K \subset \left\{ \mathbf{x} \in \mathbf{R}^n \,\middle|\, \begin{array}{l} l_3(\mathbf{x}) > 0,\ l_1(\mathbf{x}) \geqq 0 \text{ und} \\ (l_1(\mathbf{x}) > 0 \text{ oder } l_2(\mathbf{x}) > 0) \end{array} \right\}. \tag{1.5}$$

Die Eigenschaften 1 und 2 des pseudolinearen Optimierungsproblems lassen ein simplexartiges Lösungsverfahren für (P) als möglich erscheinen. Ein solcher Algorithmus ist tatsächlich anwendbar. Bei einem unbeschränkten zulässigen Bereich B kann man dabei aber — im Gegensatz zur linearen Optimierung — nicht in jedem Falle durch sukzessive Eckenübergänge mit jeweils fallender Zielfunktion zur Lösung gelangen. Dies wird am folgenden Beispiel ersichtlich:

Es sei $B \subset \mathbf{R}^2$ gegeben durch das System (vgl. Bild 1)

$$\begin{aligned} -x_1 + 2x_2 &\leqq 2 \\ x_1 - 3x_2 &\leqq 1 \\ x_1, x_2 &\geqq 0, \end{aligned} \tag{1.6}$$

und es sei

$$F(\mathbf{x}) = \frac{x_1 - x_2 + 1}{2x_1 + 1}. \tag{1.7}$$

Wegen $2x_1 + 1 > 0$ für alle $\mathbf{x} \in B$ ist $F(\mathbf{x})$ pseudolinear über B. In der Ecke $\mathbf{x}^3$ nimmt die Funktion ihr Minimum 0 über B an, $\mathbf{x}^3$ ist aber von der Ecke $\mathbf{x}^1$ aus wegen $F(\mathbf{x}^2) = 1 > \frac{2}{3} = F(\mathbf{x}^1)$ nicht durch Eckenübergänge mit fallender Zielfunktion erreichbar.

Das vorliegende Beispiel ist ein Gegenbeispiel zu der Aussage von MARTOS in [5], Corr. 5, wonach jedes lokale Eckpunktminimum einer pseudolinearen Funktion über einer (insbesondere auch unbeschränkten) konvexen polyedrischen Menge bereits das globale Minimum ist, sofern letzteres existiert. Hier wird nämlich, da $F(\mathbf{x})$ die Zielfunktionswerte der Nachbarecken von $\mathbf{x}^1$ nicht übersteigt, in $\mathbf{x}^1$ ein lokales Eckpunktminimum angenommen, welches nicht das globale Minimum ist.

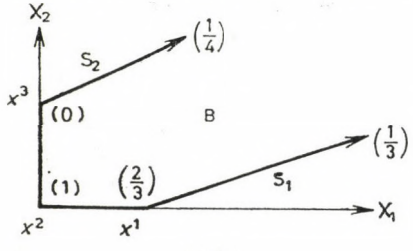


Bild 1

Das Beispiel demonstriert gleichzeitig das mögliche Versagen des Algorithmus von MARTOS [4] in der hyperbolischen Optimierung bei unbeschränkten zulässigen Berei-

chen. Dieser Algorithmus würde an der Ecke $\mathbf{x}^1$ abbrechen und fälschlicherweise die Nichtexistenz des Minimums von $F(\mathbf{x})$ über B konstatieren.

Untersuchen wir im Beispiel das Verhalten von $F(\mathbf{x})$ nicht nur in den Ecken, sondern auch über den beiden unbeschränkten Kanten S_1 und S_2 von B, so erhalten wir

(1.8)
$$F(\mathbf{x}^1) = \frac{2}{3} > \lim_{x_1 \to \infty} \{F(\mathbf{x}) | \mathbf{x} \in S_1\} = \frac{1}{3} > \lim_{x_1 \to \infty} \{F(\mathbf{x}) | \mathbf{x} \in S_2\} = \frac{1}{4} > F(\mathbf{x}^3) = 0.$$

Das heißt, die Niveaumenge $\{\mathbf{x} \in B | F(\mathbf{x}) \leqq F(\mathbf{x}^1)\}$ ist unbeschränkt, und es gibt einen "Weg" von $\mathbf{x}^1$ über S_1 ins Unendliche und von dort zurück über S_2 nach $\mathbf{x}^3$, auf dem die Eckpunktwerte bzw. Strahlgrenzwerte der Zielfunktion fallen. Diese Erkenntnis spielt für unseren Lösungsalgorithmus zu (P) eine entscheidende Rolle. Wir lassen darin zu, daß neben Ecken u.U. auch unbeschränkte Kanten von B durchlaufen werden.

2. Basisdarstellung der Kanten einer konvexen polyedrischen Menge

Um im Algorithmus mit unbeschränkten Kanten (im weiteren stets Kantenstrahlen genannt) ähnlich wie mit Ecken arbeiten zu können, geben wir in diesem Abschnitt eine Basisdarstellung für Kanten und die benötigten Vorschriften für Übergänge zwischen Ecken und Kantenstrahlen sowie Kantenstrahlen untereinander an. Dazu betrachten wir das System

(2.1) $$\mathbf{A}\mathbf{x} = \mathbf{b}$$

(2.2) $$\mathbf{x} \geqq \mathbf{0}$$

der Nebenbedingungen von (P). Es sei $\mathbf{A} = (\mathbf{a}_i)_{i=1,\ldots,n}$ mit $\mathbf{a}_i \in \mathbf{R}^m$ und Rang $\mathbf{A} = m$. Wir setzen $\mathcal{N} = \{1, \ldots, n\}$.

Definition 1. Das Paar $[\mathbf{x}^0, \mathcal{I}]$ mit $\mathbf{x}^0 \in \mathbf{R}^n$ und $\mathcal{I} \subset \mathcal{N}$ heißt Basislösung von (2.1), wenn gilt

(2.3) $$\mathbf{A}\mathbf{x}^0 = \mathbf{b}, \quad |\mathcal{I}| = m, \quad \text{Rang}\,(\mathbf{a}_i)_{i \in \mathcal{I}} = m,$$

$$x_i^0 = 0 \quad \text{für} \quad i \in \mathcal{I}' = \mathcal{N} \setminus \mathcal{I}.$$

Die x_i mit $i \in \mathcal{I}$ werden als Basisvariablen und die x_i mit $i \in \mathcal{I}'$ als Nichtbasisvariablen von $[\mathbf{x}^0, \mathcal{I}]$ bezeichnet. $\mathbf{x}^0$ ist der Basispunkt von $[\mathbf{x}^0, \mathcal{I}]$.

Die Basislösung $[\mathbf{x}^0, \mathcal{I}]$ von (2.1) heißt zulässige Basislösung oder kurz Eckbasis von (2.1), (2.2), wenn gilt $\mathbf{x}^0 \geqq \mathbf{0}$. Als kanonische Form von (2.1) bezüglich der Basislösung $[\mathbf{x}^0, \mathcal{I}]$ wird das System $\tilde{\mathbf{x}} + \mathbf{A}^0 \hat{\mathbf{x}} = \mathbf{b}^0$ bezeichnet, welches aus (2.1) durch Auflösung nach dem Vektor $\tilde{\mathbf{x}}$ der Basisvariablen hervorgeht ($\hat{\mathbf{x}}$ ist der Vektor der Nichtbasisvariablen). Es sei $\mathbf{A}^0 = (\mathbf{a}_i^0)_{i \in \mathcal{I}'}$.

Definition 2. Das Tripel $[\mathbf{x}^0, \mathscr{J}, \sigma]$ heißt Kantenbasis von (2.1), (2.2), wenn gilt
1. Es ist $[\mathbf{x}^0, \mathscr{J}]$ eine Basislösung von (2.1) und $\sigma \in \mathscr{J}'$.
2. Das Intervall

$$\Lambda = \{\lambda \in \mathbf{R}^1 \mid \lambda \geqq 0, \mathbf{b}^0 - \mathbf{a}_\sigma^0 \lambda \geqq \mathbf{0}\} \tag{2.4}$$

ist nicht leer und nicht einpunktig.

x_σ ist die markierte und jedes x_i mit $i \in \mathscr{J} = \mathscr{N} \setminus (\mathscr{J} \cup \sigma)$[1] eine gewöhnliche Nichtbasisvariable von $[\mathbf{x}^0, \mathscr{J}, \sigma]$. Λ heißt Zulässigkeitsintervall der Kantenbasis.

Ist $[\mathbf{x}^0, \mathscr{J}]$ eine Eckbasis von (2.1), (2.2), so ist $\mathbf{x}^0$ bekanntlich eine Ecke von B. Umgekehrt existiert zu jeder Ecke $\mathbf{x}^0$ von B ein $\mathscr{J} \subset \mathscr{N}$, so daß $[\mathbf{x}^0, \mathscr{J}]$ eine Eckbasis von (2.1), (2.2) ist. Eine analoge Aussage für Kantenbasen und Kanten liefert der folgende

SATZ 1: *Ist* $[\mathbf{x}^0, \mathscr{J}, \sigma]$ *eine Kantenbasis von* (2.1), (2.2), *so ist die Menge*

$$S = \{\mathbf{x} \in B \mid x_i = 0 \quad \textit{für} \quad i \in \mathscr{J}\} \quad \textit{mit} \quad \mathscr{J} = \mathscr{N} \setminus (\mathscr{J} \cup \sigma) \tag{2.5}$$

eine Kante von B. Umgekehrt existiert zu jeder Kante S von B eine Kantenbasis $[\mathbf{x}^0, \mathscr{J}, \sigma]$ *von* (2.1), (2.2), *die S gemäß* (2.5) *definiert.*

Der Beweis des Satzes (ausgeführt in [2]) basiert auf der Darstellung

$$\{\mathbf{x} \in B \mid x_i = 0 \quad \text{für} \quad i \in \mathscr{J}\} = \left\{ \mathbf{x} \in \mathbf{R}^n \;\middle|\; \begin{matrix} \tilde{\mathbf{x}} = \mathbf{b}^0 - \mathbf{a}_\sigma^0 \lambda, \; x_\sigma = \lambda, \\ x_i = 0 \quad \text{für} \quad i \in \mathscr{J}, \lambda \in \Lambda \end{matrix} \right\}. \tag{2.6}$$

Wir möchten bemerken, daß es — im Gegensatz zur Situation bei Ecken — zu einer gegebenen Kante von B auch im Falle der Nichtentartung i.a. mehrere verschiedene definierende Kantenbasen gibt.

Definition 3. Eine Kantenbasis von (2.1), (2.2) heißt (Kanten-) Strahlbasis, wenn ihr Zulässigkeitsintervall unbeschränkt ist.

SATZ 2. *Sei* $[\mathbf{x}^0, \mathscr{J}, \sigma]$ *eine Kantenbasis von* (2.1), (2.2). *Dann sind die folgenden drei Aussagen äquivalent:*
1. $[\mathbf{x}^0, \mathscr{J}, \sigma]$ *ist eine Strahlbasis von* (2.1), (2.2).
2. *Die Kante* $S = \{\mathbf{x} \in B \mid x_i = 0 \text{ für } i \in \mathscr{J}\}$ *ist ein Kantenstrahl.*
3. *Für alle* $i \in \mathscr{J}$ *gilt*

$$(-a_{i\sigma}^0, b_i^0) \geqq (0, 0)^2. \tag{2.7}$$

Für den Beweis (siehe [2]) werden folgende Überlegungen benutzt: Die Unbeschränktheit von Λ bzw. S ist gleichbedeutend damit, daß λ durch die Bedingung $\mathbf{b}^0 - \mathbf{a}_\sigma^0 \lambda \geqq \mathbf{0}$ nicht nach oben beschränkt wird. Letzteres ist genau mit (2.7) der Fall.

Da im Optimierungsverfahren zu (P) neben Ecken nur Kantenstrahlen von Interesse sind, schränken wir die folgenden Aussagen bereits auf Strahlbasen ein. Ähnlich wie das Benachbartsein zweier Eckbasen (die Indexmengen ihrer Basisvariablen unterscheiden sich in genau einem Element) definieren wir auch das Benachbartsein zweier Strahlbasen:

---

[1] Einpunktige Mengen sollen stets durch ihr Element symbolisiert werden.

[2] Das Symbol "$\geqq$" bezeichnet die bekannte reflexive lexikographische Ordnung von Vektoren.

Definition 4. Zwei Strahlbasen von (2.1), (2.2) heißen benachbart, wenn sich die Indexmengen ihrer Basisvariablen in höchstens einem und die Indexmengen ihrer gewöhnlichen Nichtbasisvariablen in genau einem Element unterscheiden.

Folgerung. Von einer gegebenen Strahlbasis $[\mathbf{x}^1, \mathscr{J}_1, \sigma_1]$ kann man auf genau vier Arten zu einer benachbarten Strahlbasis $[\mathbf{x}^2, \mathscr{J}_2, \sigma_2]$ gelangen:

A. Die markierte Nichtbasisvariable bleibt fest. Dann gilt (notwendig) mit einem $\alpha \in \mathscr{J}_1$ und einem $\beta \in \mathscr{J}_1$

$$\mathscr{J}_2 = (\mathscr{J}_1 \setminus \beta) \cup \alpha, \mathscr{J}_2 = (\mathscr{J}_1 \setminus \alpha) \cup \beta, \sigma_2 = \sigma_1. \tag{2.8}$$

B. Die markierte Nichtbasisvariable wird Basisvariable. Dann gilt mit einem $\alpha \in \mathscr{J}_1$ und einem $\beta \in \mathscr{J}_1$

$$\mathscr{J}_2 = (\mathscr{J}_1 \setminus \beta) \cup \sigma_1, \mathscr{J}_2 = (\mathscr{J}_1 \setminus \alpha) \cup \beta, \sigma_2 = \alpha. \tag{2.9}$$

C. Die markierte wird zur gewöhnlichen Nichtbasisvariablen.

C.1. Die Basisvariablen bleiben fest. Dann gilt mit einem $\alpha \in \mathscr{J}_1$

$$\mathscr{J}_2 = \mathscr{J}_1, \mathscr{J}_2 = (\mathscr{J}_1 \setminus \alpha) \cup \sigma_1, \sigma_2 = \alpha. \tag{2.10}$$

C.2. Die Basisvariablen ändern sich. Dann gilt mit einem $\alpha \in \mathscr{J}_1$ und einem $\beta \in \mathscr{J}_1$

$$\mathscr{J}_2 = (\mathscr{J}_1 \setminus \beta) \cup \alpha, \mathscr{J}_2 = (\mathscr{J}_1 \setminus \alpha) \cup \sigma_1, \sigma_2 = \beta. \tag{2.11}$$

Analog wie für Eckbasen und Ecken gilt nun

SATZ 3. *Sind zwei Strahlbasen von* (2.1), (2.2) *benachbart, so sind die beiden durch sie definierten Kantenstrahlen von B benachbart*[3] *oder — höchstens bei Entartung — identisch.*

Infolge der Verwendung beliebiger (also auch unzulässiger) Basislösungen für den Begriff der Strahlbasis kann man zu benachbarten Kantenstrahlen stets auch benachbarte definierende Strahlbasen finden.

Wir wollen zur Illustration der eingeführten Begriffe ein Beispiel betrachten: Es sei $B \subset \mathbf{R}^4$ gegeben durch

$$\begin{aligned} -x_1 + x_2 + x_3 &= 1 \\ 2x_1 + 4x_2 - x_4 &= 1 \\ x_1, x_2, x_3, x_4 &\geqq 0. \end{aligned} \tag{2.12}$$

Zur geometrischen Veranschaulichung transformieren wir B in den x_1, x_2-Raum (vgl. $\hat{B}$ in Bild 2).

B besitzt die beiden (benachbarten) Kantenstrahlen $S_1 = \{\mathbf{x} \in B | x_2 = 0\}$ und $S_2 = \{\mathbf{x} \in B | x_3 = 0\}$ (vgl. $\hat{S}_1$ und $\hat{S}_2$ in Bild 2). Sie lassen sich beide durch mehrere Strahlbasen definieren — S_1 etwa mit dem Basispunkt $\left(\frac{1}{2}, 0, \frac{3}{2}, 0\right)^\top$ (vgl. $\mathbf{y}^1$)

[3] Das heißt, sie liegen auf ein und derselben zweidimensionalen Randfläche von B.

und der markierten Nichtbasisvariablen x_4 oder mit dem Basispunkt $(-1, 0, 0, -3)^\top$ (vgl. $\mathbf{y}^3$) und der markierten Nichtbasisvariablen x_3. Als Basispunkt kommt jeder auf der Geraden durch den Kantenstrahl liegende (auch unzulässige) Schnittpunkt in Frage. Unter den Nichtbasisvariablen der zugehörigen Basislösung wird genau diejenige markiert, die über dem Strahl unbeschränkt ist.

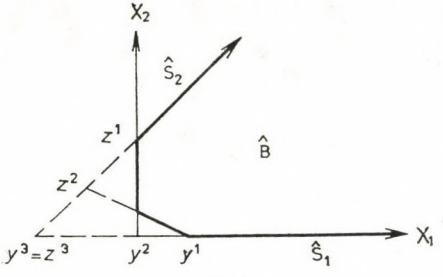


Bild 2

Um von einer Strahlbasis zu einer benachbarten zu gelangen, gehen wir (außer im Fall C.1, wo der Basispunkt beibehalten wird und sich nur die markierte Nichtbasisvariable ändert) von einem Basispunkt zu einem benachbarten über, wobei die markierte Nichtbasisvariable fest bleibt (Fall A) oder sich ändert (Fälle B und C.2). Im Beispiel sind etwa die Strahlbasen $\left[\left(\frac{1}{2}, 0, \frac{3}{2}, 0\right)^\top, \{1, 3\}, 4\right]$ und $\left[\left(-\frac{1}{2}, \frac{1}{2}, 0, 0\right)^\top, \{1, 2\}, 4\right]$ gemäß A bzw. die Strahlbasen $[(-1, 0, 0, -3)^\top, \{1, 4\}, 3]$ und $[(-1, 0, 0, -3)^\top, \{1, 4\}, 2]$ gemäß C.1 benachbart. Ihnen entsprechen in Bild 2 die Basispunkte $\mathbf{y}^1$ und $\mathbf{z}^2$ bzw. $\mathbf{y}^3$ und $\mathbf{z}^3 = \mathbf{y}^3$.

Das Beispiel zeigt die Notwendigkeit der Benutzung evtl. unzulässiger Basislösungen zur Darstellung der Kantenstrahlen. Ließen wir nur Eckbasen zu, so wären die die benachbarten Kantenstrahlen S_1 und S_2 definierenden Strahlbasen zwar eindeutig bestimmt, aber nicht benachbart — eine besitzt x_1 und x_3, die andere x_2 und x_4 als Basisvariablen (vgl. die Basispunkte $\mathbf{y}^1$ und $\mathbf{z}^1$ in Bild 2).

Definition 5. Eine Eckbasis und eine Strahlbasis von (2.1), (2.2) heißen inzident, wenn die Indexmenge der gewöhnlichen Nichtbasisvariablen der Strahlbasis in der Indexmenge der Nichtbasisvariablen der Eckbasis enthalten ist.

Folgerung 1. Von einer gegeben Eckbasis $[\mathbf{x}^1, \mathscr{I}_1]$ kann man auf genau zwei Arten zu einer inzidenten Strahlbasis $[\mathbf{x}^2, \mathscr{I}_2, \sigma]$ gelangen:

D.1. Eine Nichtbasisvariable wird zur markierten Nichtbasisvariablen. Dann gilt mit einem $\alpha \in \mathscr{I}_1'$

$$\mathscr{I}_2 = \mathscr{I}_1, \mathscr{J}_2 = \mathscr{J}_1' \setminus \alpha, \sigma = \alpha. \tag{2.13}$$

D.2. Eine Basisvariable wird zur markierten Nichtbasisvariablen. Dann gilt mit einem $\alpha \in \mathscr{I}_1'$ und einem $\beta \in \mathscr{I}_1$

$$\mathscr{I}_2 = (\mathscr{I}_1 \setminus \beta) \cup \alpha, \mathscr{J}_2 = \mathscr{J}_1' \setminus \alpha, \sigma = \beta. \tag{2.14}$$

Folgerung 2. Von einer gegebenen Strahlbasis $[\mathbf{x}^1, \mathscr{I}_1, \sigma]$ kann man auf genau zwei Arten zu einer inzidenten Eckbasis $[\mathbf{x}^2, \mathscr{I}_2]$ gelangen:

E.1. Die markierte Nichtbasisvariable bleibt Nichtbasisvariable. Dann gilt

$$\mathscr{I}_2 = \mathscr{I}_1, \mathscr{I}_1' = \mathscr{J}_1 \cup \sigma. \tag{2.15}$$

E.2. Die markierte Nichtbasisvariable wird Basisvariable. Dann gilt mit einem $\beta \in \mathscr{I}_1$

$$\mathscr{I}_2 = (\mathscr{I}_1 \setminus \beta) \cup \sigma, \ \mathscr{I}_2' = \mathscr{I}_1 \cup \beta. \tag{2.16}$$

SATZ 4. *Sind eine Eckbasis und eine Strahlbasis von* (2.1), (2.2) *inzident, so liegt die definierte Ecke auf dem definierten Kantenstrahl.*

Zu einer gegebenen Strahlbasis findet man stets eine (bei Nichtentartung eindeutig bestimmte) inzidente Eckbasis. Zu einer Eckbasis kann es mehrere (oder evtl. auch gar keine) inzidente Strahlbasen geben.

Wir kommen nun zu den Realisierungsbedingungen und Rechenregeln für die im Algorithmus benötigten Basisübergänge. Neben dem in bekannter Weise erfolgenden Übergang von einer Eckbasis zu einer benachbarten Eckbasis müssen im Algorithmus die folgenden Basisübergänge ausgeführt werden:

E-S: von einer Eckbasis zu einer inzidenten Strahlbasis,

S-S: von einer Strahlbasis zu einer benachbarten Strahlbasis,

S-E: von einer Strahlbasis zu einer inzidenten Eckbasis.

Die Vorschriften hierfür werden anhand von Tableaus angegeben.

Eine Eckbasis $[\mathbf{x}^0, \mathscr{I}]$ mit der kanonischen Form $\tilde{\mathbf{x}} + \mathbf{A}^0 \hat{\mathbf{x}} = \mathbf{b}^0$ von (2.1) wird wie üblich durch das Simplextableau

(2.17)

	x_{m+1}	$\cdots$	x_n	$\mathbf{b}^0$
x_1	$a^0_{1,m+1}$	$\cdots$	$a^0_{1,n}$	b^0_1
$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_m	$a^0_{m,m+1}$	$\cdots$	$a^0_{m,n}$	b^0_m

dargestellt. Dabei ist o.B.d.A. $\tilde{\mathbf{x}} = (x_1, \ldots, x_m)^\top$ und $\hat{\mathbf{x}} = (x_{m+1}, \ldots, x_n)^\top$. Im Tableau gilt $\mathbf{b}^0 \geqq \mathbf{0}$.

Eine Strahlbasis $[\mathbf{x}^0, \mathscr{I}, \sigma]$ wird durch das Tableau

(2.18)

			$*$			
	x_{m+1}	$\cdots$	$x_{6\cdot}$	$\cdots$	x_n	$\mathbf{b}^0$
x_1	$a^0_{1,m+1}$	$\cdots$	$a^0_{1,6}$	$\cdots$	$a^0_{1,n}$	b^0_1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_m	$a^0_{m,m+1}$	$\cdots$	$a^0_{m,6}$	$\cdots$	$a^0_{m,n}$	b^0_m

welches aus dem Simplextableau von $[\mathbf{x}^0, \mathscr{I}]$ durch Markierung von x_σ mit "$*$" hervorgeht, dargestellt. Entsprechend Satz 2 erfüllen die Elemente des Tableaus $(-a^0_{i\sigma}, b^0_i) \geqq (0, 0)$ für alle $i \in \mathscr{I}$.

Realisierung des Übergangs E-S

Eine Eckbasis $[\mathbf{x}^0, \mathscr{I}]$ sei durch ein Tableau der Gestalt (2.17) gegeben.

VORAUSSETZUNG. Es sei $\mathbf{a}^0_\alpha \leqq \mathbf{0}$ für ein $\alpha \in \mathscr{I}'$.

BEHAUPTUNG. Dann erhalten wir durch die Markierung von x_α in (2.17) eine zu $[\mathbf{x}^0, \mathscr{I}]$ gemäß D.1 inzidente Strahlbasis.

Realisierung der Übergänge S-S

Eine Strahlbasis $[\mathbf{x}^0, \mathscr{I}, \sigma]$ sei durch ein Tableau der Gestalt (2.18) gegeben Wir wählen ein beliebiges $\alpha \in \mathscr{J}$ aus und bezeichnen es als Übergangsindex.

Übergang S-S-1

VORAUSSETZUNG. Es sei $\mathbf{a}_\alpha^0 \not\equiv \mathbf{0}$, d.h. $a_{i\alpha}^0 > 0$ für gewisse $i \in \mathscr{I}$.

BEHAUPTUNG. Bilden wir

$$\left(-\frac{a_{\beta\sigma}^0}{a_{\beta\alpha}^0}, \frac{b_\beta^0}{a_{\beta\alpha}^0}\right) = \text{Lex. Min}\left\{\left(-\frac{a_{i\sigma}^0}{a_{i\alpha}^0}, \frac{b_0^i}{a_{i\alpha}^0}\right) \middle| i \in \mathscr{I}, a_{i\alpha}^0 > 0\right\}, \tag{2.19}$$

so erhalten wir durch Umrechnung von (2.18) mit dem Pivotelement $a_{\beta\alpha}^0$ eine zu $[\mathbf{x}^0, \mathscr{I}, \sigma]$ gemäß A benachbarte Strahlbasis.

Übergang S-S-2

VORAUSSETZUNG. Es sei $\mathbf{a}_\alpha^0 \leqq \mathbf{0}$ und $(-a_{i\alpha}^0, b_i^0) \prec (0, 0)$, d.h. $a_{i\alpha}^0 = 0$ und $b_i^0 < 0$ für gewisse $i \in \mathscr{I}$.

BEHAUPTUNG. Bilden wir

$$\frac{b_\beta^0}{a_{\beta\sigma}^0} = \text{Max}\left\{\frac{b_i^0}{a_{i\sigma}^0} \middle| i \in \mathscr{I}, a_{i\alpha}^0 = 0, b_i^0 < 0\right\}, \tag{2.20}$$

so erhalten wir durch Löschen der Marke bei x_σ, Markierung von x_α und Umrechnung von (2.18) mit dem (negativen) Pivotelement $a_{\beta\sigma}^0$ eine zu $[\mathbf{x}^0, \mathscr{I}, \sigma]$ gemäß *B* benachbarte Strahlbasis.

Übergang S-S-3

VORAUSSETZUNG. Es sei $\mathbf{a}_\alpha^0 \leqq \mathbf{0}$ und $(-a_{i\alpha}^0, b_i^0) \geqq (0, 0)$ für alle $i \in \mathscr{I}$.

BEHAUPTUNG. Durch Löschen der Marke bei x_σ und Markierung von x_α erhalten wir aus (2.18) eine zu $[\mathbf{x}^0, \mathscr{I}, \sigma]$ gemäß C.1 benachbarte Strahlbasis.

ANMERKUNG. Für jedes $\alpha \in \mathscr{J}$ ist genau eine der drei Voraussetzungen für die Übergänge S-S-1/2/3 erfüllt. Mit jedem beliebigen $\alpha \in \mathscr{J}$ als Übergangsindex läßt sich also ein Basisübergang S-S ausführen, und die Auswahl von α bestimmt bereits die Art dieses Übergangs.

Realisierung der Übergänge S-E

Eine Strahlbasis $[\mathbf{x}^0, \mathscr{I}, \sigma]$ sei durch ein Tableau der Gestalt (2.18) gegeben.

Übergang S-E-1

VORAUSSETZUNG. Es sei $\mathbf{b}^0 \geqq \mathbf{0}$.

BEHAUPTUNG. Durch Löschen der Marke bei x_σ erhalten wir aus (2.18) eine zu $[\mathbf{x}^0, \mathcal{J}, \sigma]$ gemäß E.1 inzidente Eckbasis.

Übergang S-E-2

VORAUSSETZUNG. Es sei $\mathbf{b}^0 \ngeqq \mathbf{0}$, d.h. $b_i^0 < 0$ für gewisse $i \in \mathcal{J}$.

BEHAUPTUNG. Bilden wir

$$\frac{b_\beta^0}{a_{\beta\sigma}^0} = \operatorname{Max}\left\{\frac{b_i^0}{a_{i\sigma}^0}\middle| i \in \mathcal{J},\ b_i^0 < 0\right\}, \tag{2.21}$$

so erhalten wir durch Löschen der Marke bei x_σ und Umrechnung von (2.18) mit dem (negativen) Pivotelement $a_{\beta\sigma}^0$ eine zu $[\mathbf{x}^0, \mathcal{J}, \sigma]$ gemäß E.2 inzidente Eckbasis.

Die Beweise dafür, daß unter den angegebenen Voraussetzungen und Rechenvorschriften tatsächlich stets die behaupteten Strahl- bzw. Eckbasen erreicht werden, sind in [2] geführt.

3. Allgemeines Prinzip des Algorithmus

Wir wollen nun darlegen, wie das pseudolineare Optimierungsproblem

$$F(\mathbf{x}) = \operatorname{Min}!, \quad \mathbf{Ax} = \mathbf{b},\ \mathbf{x} \geqq \mathbf{0} \tag{P}$$

mittels einer simplexartigen Abstiegsmethode prinzipiell gelöst werden kann. Dabei zeigen wir auch eine Möglichkeit zur rechnerischen Realisierung der auftretenden Tests unter einer gewissen (schwachen) Zusatzbedingung an die pseudolineare Zielfunktion auf. Wir setzen voraus, daß der zulässige Bereich $B = \{\mathbf{x} \in \mathbf{R}^n | \mathbf{Ax} = \mathbf{b},\ \mathbf{x} \geqq \mathbf{0}\}$ von (P) nicht entartet ist.

Sei $[-\mathbf{x}^0, \mathcal{J}]$ eine Basislösung von (2.1) mit der zugehörigen kanonischen Form $\tilde{\mathbf{x}} + \mathbf{A}^0\hat{\mathbf{x}} = \mathbf{b}^0$. Wir definieren die Funktion

$$f(\hat{\mathbf{x}}) = F(\mathbf{x})|_{\tilde{\mathbf{x}} = \mathbf{b}^0 - \mathbf{A}^0\hat{\mathbf{x}}}. \tag{3.1}$$

LEMMA 1. *Für beliebige* $\mathbf{x}^1, \mathbf{x}^2 \in B$ *gilt*

$$(\mathbf{x}^1 - \mathbf{x}^2)^\top \operatorname{grad} F(\mathbf{x}^2) = (\hat{\mathbf{x}}^1 - \hat{\mathbf{x}}^2)^\top \operatorname{grad} f(\hat{\mathbf{x}}^2). \tag{3.2}$$

3.1. *Beginn des Algorithmus, Optimaltest und Übergang zu Strahlbasen*

Wir gehen von einer Ecke $\mathbf{x}^0$ von B aus. Die zugehörige Eckbasis $[\mathbf{x}^0, \mathcal{J}]$ sei durch das Tableau (2.17) gegeben. Wir bilden entsprechend (3.1) $f(\hat{\mathbf{x}})$ und

$$g_i = \frac{\partial f}{\partial x_i}(\hat{\mathbf{x}})\bigg|_{\hat{\mathbf{x}}=\mathbf{0}}, \quad i \in \mathcal{J}', \quad \text{mit} \quad \mathbf{g} = (g_i)_{i \in \mathcal{J}'}. \tag{3.3}$$

Fall E.1 (Klassischer Optimaltest am Eckpunkt): $\mathbf{g} \geqq \mathbf{0}$.

BEHAUPTUNG. Dann nimmt $F(\mathbf{x})$ in $\mathbf{x}^0$ sein Minimum über B an.

BEWEIS. Sei $\mathbf{x}' \in B$ beliebig. Nach Lemma 1 gilt wegen $\hat{\mathbf{x}}' \geqq \mathbf{0}$

$$(\mathbf{x}' - \mathbf{x}^0)^\top \operatorname{grad} F(\mathbf{x}^0) = (\hat{\mathbf{x}}' - \hat{\mathbf{x}}^0)^\top \operatorname{grad} f(\hat{\mathbf{x}}^0) = \hat{\mathbf{x}}'^\top \mathbf{g} \geqq \mathbf{0}. \tag{3.4}$$

Da $F(\mathbf{x})$ pseudolinear ist, folgt nach (1.2) $F(\mathbf{x}') \geqq F(\mathbf{x}^0)$.

Fall E.2: $\mathbf{g} \ngeqq \mathbf{0}$, d.h. $g_i < 0$ für gewisse $i \in \mathscr{I}'$.

Dann wählen wir ein $\alpha \in \mathscr{I}'$ mit $g_\alpha < 0$ (zum Beispiel $g_\alpha = \operatorname{Min}\{g_i | i \in \mathscr{I}'\}$) und unterscheiden die Unterfälle

Fall E.2.1: $\mathbf{a}_\alpha^0 \ngeqq \mathbf{0}$, d.h. $a_{i\alpha}^0 > 0$ für gewisse $i \in \mathscr{I}$.

In diesem Falle bestimmen wir mit der Pivotspalte α nach der üblichen Regel ein Pivotelement und rechnen damit das Tableau (2.17) um.

BEHAUPTUNG. Für die erhaltene Nachbarecke $\mathbf{x}^1$ von $\mathbf{x}^0$ gilt $F(\mathbf{x}^1) < F(\mathbf{x}^0)$.

BEWEIS. Nach Lemma 1 gilt wegen der Nichtentartung von $\mathbf{x}^0$

$$(\mathbf{x}^1 - \mathbf{x}^0)^\top \operatorname{grad} F(\mathbf{x}^0) = x_\alpha^1 g_\alpha < 0, \tag{3.5}$$

nach (1.2) erhalten wir daraus $F(\mathbf{x}^1) < F(\mathbf{x}^0)$.

Fall E.2.2: $\mathbf{a}_\alpha^0 \leqq \mathbf{0}$.

Dann führen wir durch die Markierung von x_α den Übergang E-S zu einer inzidenten Strahlbasis aus.

BEHAUPTUNG. Für den erhaltenen Kantenstrahl S_0 gilt $\operatorname{Inf}\{F(\mathbf{x}) | \mathbf{x} \in S_0\} < F(\mathbf{x}^0)$.

BEWEIS. Sei $\mathbf{x}' \in S_0 \setminus \mathbf{x}^0$ beliebig. Dann gilt $x_\alpha' > 0$ und somit

$$(\mathbf{x}' - \mathbf{x}^0)^\top \operatorname{grad} F(\mathbf{x}^0) = x_\alpha' g_\alpha < 0. \tag{3.6}$$

Es folgt $F(\mathbf{x}') < F(\mathbf{x}^0)$ und daraus die Behauptung.

3.2. *Verallgemeinerter Optimaltest, Basisübergänge S-S und S-E*

Es sei im Algorithmus eine Strahlbasis $[\mathbf{x}^0, \mathscr{I}, \sigma]$ erreicht, die durch ein Tableau der Gestalt (2.18.) dargestellt wird. Analog zu (3.1) bilden wir $f(\hat{\mathbf{x}})$, außerdem definieren wir

$$g_i(\lambda) = \frac{\partial f}{\partial x_i}(\hat{\mathbf{x}})\bigg|_{x_i = 0 \text{ für } i \in \mathscr{J},\, x_\sigma = \lambda}, \quad i \in \mathscr{I}', \tag{3.7}$$

$$\lambda_0 = \operatorname{Min}\{\lambda \in \mathbf{R}^1 | \lambda \geqq 0,\ \mathbf{b}^0 - \mathbf{a}_\sigma^0 \lambda \geqq \mathbf{0}\}. \tag{3.8}$$

S_0 sei der durch $[\mathbf{x}^0, \mathscr{J}, \sigma]$ definierte Kantenstrahl. Mit dem durch

$$(3.9) \qquad s_i^0 = \begin{cases} -a_{i\sigma}^0 & \text{für} \quad i \in \mathscr{J} \\ 1 & \text{für} \quad i = \sigma \\ 0 & \text{für} \quad i \in \mathscr{J} \end{cases}$$

gegebenen Vektor $\mathbf{s}^0$ gilt $S_0 = \{\mathbf{x} \in \mathbf{R}^n | \mathbf{x} = \mathbf{x}^0 + \lambda \mathbf{s}^0, \ \lambda \geqq \lambda_0\}$.

Fall S.0: $\lim\limits_{\lambda \to \infty} F(\mathbf{x}^0 + \lambda \mathbf{s}^0) = -\infty$.

In diesem Falle wird das Verfahren wegen der Unbeschränktheit der Zielfunktion abgebrochen.

Wegen der Pseudolinearität von $F(\mathbf{x})$ ist $g_\sigma(\lambda)$ für $\lambda \geqq \lambda_0$ entweder stets negativ oder stets positiv oder stets gleich Null. Tritt der Fall S.0 nicht ein, so können wir daher die beiden Fälle S.1 und S.2 unterscheiden.

Fall S.1: $\lim\limits_{\lambda \to \infty} F(\mathbf{x}^0 + \lambda \mathbf{s}^0) > -\infty$ und $g_\sigma(\lambda) < 0$ für $\lambda \geqq \lambda_0$.

BEHAUPTUNG: Dann gilt

$$(3.10) \qquad \operatorname{Inf}\{F(\mathbf{x}) | \mathbf{x} \in S_1\} = \lim_{\lambda \to \infty} F(\mathbf{x}^0 + \lambda \mathbf{s}^0) < F(\mathbf{x}) \quad \text{für alle} \quad \mathbf{x} \in S_0.$$

BEWEIS. Wir betrachten beliebige $\mathbf{x}', \mathbf{x}'' \in S_0$ mit $\mathbf{x}' = \mathbf{x}^0 + \lambda' \mathbf{s}^0$ und $\mathbf{x}'' = \mathbf{x}^0 + \lambda'' \mathbf{s}^0$ für $\lambda'' > \lambda' \geqq \lambda_0$. Dann haben wir

$$(3.11) \qquad (\mathbf{x}'' - \mathbf{x}')^\top \operatorname{grad} F(\mathbf{x}') = (\lambda'' - \lambda') g_\sigma(\lambda') < 0,$$

also $F(\mathbf{x}'') < F(\mathbf{x}')$. Damit fällt $F(\mathbf{x})$ längs S_0 streng monoton, und es gilt (3.10).

Wir zerlegen nun die Indexmenge $\mathscr{J}$ in die Teilmengen

$$(3.12) \qquad \mathscr{J}^+ = \{i \in \mathscr{J} | \mathbf{a}_i^0 \not\leqq \mathbf{0}\} \quad \text{und} \quad \mathscr{J}^- = \{i \in \mathscr{J} | \mathbf{a}_i^0 \leqq \mathbf{0}\}$$

und führen den verallgemeinerten Optimaltest aus.

Verallgemeinerter Optimaltest: Es existiert eine Folge $\{\lambda^l\}_{l=1,2,\ldots}$ mit $\lambda^l \geqq \lambda_0$ und $\lim\limits_{l \to \infty} \lambda^l = \infty$, die für alle $i \in \mathscr{J}$ erfüllt

$$(3.13) \qquad g_i(\lambda^l) \geqq \begin{cases} 0 & \text{für} \quad i \in \mathscr{J}^+ \\ g_\sigma(\lambda^l) & \text{für} \quad i \in \mathscr{J}^-. \end{cases}$$

Fall S.1.1: Der verallgemeinerte Optimaltest sei erfüllt.

BEHAUPTUNG. Dann gilt $\lim\limits_{\lambda \to \infty} F(\mathbf{x}^0 + \lambda \mathbf{s}^0) < F(\mathbf{x})$ für alle $\mathbf{x} \in B$. $F(\mathbf{x})$ erreicht also im uneigentlichen Punkt $\mathbf{x}^0 + \infty \cdot \mathbf{s}^0$ sein endliches Infimum (verallgemeinertes Minimum) über B, ein Minimum im gewöhnlichen Sinne existiert nicht.

BEWEIS. Sei $\mathbf{x}' \in B$ beliebig. Wir wählen ein $\mathbf{x}'' \in S_0$ mit

$$x''_\sigma > x'_\sigma + \sum_{i \in \mathcal{J}^-} x'_i \quad \text{und} \quad x''_\sigma \in \{\lambda^l\}. \tag{3.14}$$

Wegen (3.13), (3.14) und $\mathbf{x}' \geqq \hat{\mathbf{x}}$ gilt dann

$$\begin{aligned}(\mathbf{x}' - \mathbf{x}'')^\top \operatorname{grad} F(\mathbf{x}'') &= (x'_\sigma - x''_\sigma) g_\sigma(x''_\sigma) + \sum_{i \in \mathcal{J}^+} x'_i g_i(x''_\sigma) + \sum_{i \in \mathcal{J}^-} x'_i g_i(x''_\sigma) \\ &\geqq (x'_\sigma - x''_\sigma) g_\sigma(x''_\sigma) + \sum_{i \in \mathcal{J}^-} x'_i g_i(x''_\sigma) \geqq \\ &\geqq (x'_\sigma - x''_\sigma) + \sum_{i \in \mathcal{J}^-} x'_i) g_\sigma(x''_\sigma) > 0.\end{aligned} \tag{3.15}$$

Daraus folgt $F(\mathbf{x}') > F(\mathbf{x}'')$ und somit nach (3.10) die Behauptung.

Fall S.1.2: Der verallgemeinerte Optimaltest sei nicht erfüllt.

HILFSBEHAUPTUNG. Dann existiert eine Folge $\{\mu^l\}_{l=1,2,\ldots}$ mit $\mu^l \geqq \lambda_0$ und $\lim_{l \to \infty} \mu^l = \infty$, die für ein $\alpha \in \mathcal{J}$ erfüllt

$$g_\alpha(\mu^l) < \begin{cases} 0 & \text{für} \quad \alpha \in \mathcal{J}^+ \\ g_\sigma(\mu^l) & \text{für} \quad \alpha \in \mathcal{J}^-. \end{cases} \tag{3.16}$$

BEWEIS. Ist der verallgemeinerte Optimaltest nicht erfüllt, dann gibt es ein $\mu' \geqq \lambda_0$, so daß zu jedem $\mu \geqq \mu'$ wenigstens ein $i(\mu) \in \mathcal{J}$ existiert mit

$$g_{i(\mu)}(\mu) < \begin{cases} 0 & \text{für} \quad i(\mu) \in \mathcal{J}^+ \\ g_\sigma(\mu) & \text{für} \quad i(\mu) \in \mathcal{J}^-. \end{cases} \tag{3.17}$$

Sei $\{\nu^l\}_{l=1,2,\ldots}$ eine beliebige Folge mit $\lim_{l \to \infty} \nu^l = \infty$. Wegen der Endlichkeit von $\mathcal{J}$ besitzt die Menge $\{i(\nu^l) \in \mathcal{J} \mid l \in \{1, 2, \ldots\}, \nu^l \geqq \mu'\}$ mindestens einen Häufungspunkt $\alpha \in \mathcal{J}$. Das heißt, es gibt eine Teilfolge $\{\mu^l\}_{l=1,2,\ldots}$ von $\{\nu^l\}$ mit

$$\mu^l \geqq \mu' \geqq \lambda_0, \ \lim_{l \to \infty} \mu^l = \infty \quad \text{und} \quad i(\mu^l) = \alpha, \tag{3.18}$$

woraus nach (3.17) die Behauptung folgt.

Wir wählen nun ein $\alpha \in \mathcal{J}$, das der Bedingung (3.16) genügt. Mit diesem Index α führen wir einen Übergang S-S von $[\mathbf{x}^0, \mathcal{J}, \sigma]$ zu einer benachbarten Strahlbasis aus. Die erreichte Strahlbasis möge den Basispunkt $\mathbf{x}^1$ besitzen und einen Kantenstrahl S_1 definieren. Die Größen λ_1 und $\mathbf{s}^1$ seien zu ihr analog wie λ_0 und $\mathbf{s}^0$ zur Ausgangs-Strahlbasis gebildet.

BEHAUPTUNG. Es gilt

$$\operatorname{Inf}\{F(\mathbf{x}) \mid \mathbf{x} \in S_1\} \leqq \operatorname{Inf}\{F(\mathbf{x}) \mid \mathbf{x} \in S_0\} = \lim_{\lambda \to \infty} F(\mathbf{x}^0 + \lambda \mathbf{s}^0). \tag{3.19}$$

BEWEIS. Sei $\mathbf{x}' \in S_0$ beliebig mit

$$x'_\sigma > \lambda_1 \quad \text{und} \quad x'_\sigma \in \{\mu^l\}. \tag{3.20}$$

a) $\alpha \in \mathscr{J}^+$: Dann existiert ein $\mathbf{x}'' \in S_1$ mit $x''_\sigma = x'_\sigma$ und, da S_0 nicht entartet ist, $x''_\alpha > 0$. Wegen (3.16) und (3.20) gilt also

$$(\mathbf{x}'' - \mathbf{x}')^\top \operatorname{grad} F(\mathbf{x}') = x''_\alpha g_\alpha(x'_\sigma) < 0. \tag{3.21}$$

Folglich ist $F(\mathbf{x}'') < F(\mathbf{x}')$, d.h. mit $\mathbf{x}' = \mathbf{x}^0 + x'_\sigma \mathbf{s}^0$ und $\mathbf{x}'' = \mathbf{x}^1 + x''_\sigma \mathbf{s}^1 = \mathbf{x}^1 + x'_\sigma \mathbf{s}^1$

$$F(\mathbf{x}^1 + x'_\sigma \mathbf{s}^1) < F(\mathbf{x}^0 + x'_\sigma \mathbf{s}^0). \tag{3.22}$$

b) $\alpha \in \mathscr{J}^-$: Dann existiert ein $\mathbf{x}'' \in S_1$ mit $x''_\alpha = x'_\sigma > 0$. Wegen (3.16), (3.20) und $g_\sigma(x'_\sigma) < 0$ haben wir folglich

$$\begin{aligned}(\mathbf{x}'' - \mathbf{x}')^\top \operatorname{grad} F(\mathbf{x}') &= (x''_\sigma - x'_\sigma) g_\sigma(x'_\sigma) + x'_\sigma g_\alpha(x'_\sigma) \leqq \\ &\leqq x'_\sigma [g_\alpha(x'_\sigma) - g_\sigma(x'_\sigma)] < 0.\end{aligned} \tag{3.23}$$

Also gilt wieder $F(\mathbf{x}'') < F(\mathbf{x}')$ und damit, da $\mathbf{x}'' = \mathbf{x}^1 + x'_\sigma \mathbf{s}^1$,

$$F(\mathbf{x}^1 + x'_\sigma \mathbf{s}^1) < F(\mathbf{x}^0 + x'_\sigma \mathbf{s}^0). \tag{3.24}$$

Aus (3.22) und (3.24) folgt, da x'_σ beliebig groß gewählt werden kann, mit (3.10) die Behauptung.

Fall S.2: $\lim\limits_{\lambda \to \infty} F(\mathbf{x}^0 + \lambda \mathbf{s}^0) > -\infty$ und $g_\sigma(\lambda) \geqq 0$ für $\lambda \geqq \lambda_0$.

Dann wird ein Übergang S-E von $[\mathbf{x}^0, \mathscr{J}, \sigma]$ zu einer inzidenten Eckbasis ausgeführt.

BEHAUPTUNG. Für die erreichte Ecke $\mathbf{x}^1$ gilt $F(\mathbf{x}^1) = \operatorname{Min} \{F(\mathbf{x}) | \mathbf{x} \in S_0\}$.

BEWEIS. Sei $\mathbf{x}' \in S_0$ beliebig. Wegen $x'_\sigma \geqq \lambda_0 = x^1_\sigma$ haben wir

$$(\mathbf{x}' - \mathbf{x}^1)^\top \operatorname{grad} F(\mathbf{x}^1) = (x'_\sigma - x^1_\sigma) g_\sigma(x^1_\sigma) \geqq 0, \tag{3.25}$$

also $F(\mathbf{x}') \geqq F(\mathbf{x}^1)$.

3.3. *Gesamtablauf des Verfahrens*

Im Optimierungsverfahren zu (P) werden, ausgehend von einer Eckbasis, nacheinander jeweils Eck- und Strahlbasen nach dem im Ablaufschema (Bild 3) angegebenen Prinzip durchlaufen. Dabei gilt der folgende

ENDLICHKEITSSATZ: Der angegebene Algorithmus ist endlich.

BEWEIS: Wir zeigen zunächst, daß im Algorithmus stets nur endlich viele Strahlbasen unmittelbar hintereinander durchlaufen werden. Dazu betrachten wir eine Kette direkt aufeinanderfolgender Strahlbasen

$$[\mathbf{x}^1, \mathscr{J}_1, \sigma_1], [\mathbf{x}^2, \mathscr{J}_2, \sigma^2], \ldots, [\mathbf{x}^\varkappa, \mathscr{J}_\varkappa, \sigma_\varkappa], \ldots \quad (\varkappa \geqq 2) \tag{3.26}$$

mit den zugehörigen Größen $\lambda_1, \lambda_2, \ldots, \lambda_\varkappa, \ldots$ und den die Hilfsbehauptung von Fall S.1.2 erfüllenden Folgen $\{\mu^l_1\}, \{\mu^l_2\}, \ldots, \{\mu^l_\varkappa\}, \ldots$ Nach dem Beweis der Hilfsbehauptung kann man diese Folgen sogar so bilden, daß gilt $\{\mu^l_1\} \supset \ldots \supset \{\mu^l_\varkappa\} \supset \ldots$.

Angenommen, es wäre $[\mathbf{x}^1, \mathscr{I}_1, \sigma_1]=[\mathbf{x}^\varkappa, \mathscr{I}_\varkappa, \sigma_\varkappa]$. Dann hätten wir $F(\mathbf{x}^1+\mu\mathbf{s}^1)=$ $=F(\mathbf{x}^\varkappa+\mu\mathbf{s}^\varkappa)$ für alle $\mu\geqq\lambda_1=\lambda_\varkappa$. Andererseits läßt sich aber ein μ^* wählen mit

$$u^* \geqq \operatorname{Max}\{\lambda_i | i\in\{2, \ldots, \varkappa\}\} \quad \text{und} \quad \mu^*\in\{\mu^l_{\varkappa-1}\}\subset\ldots\subset\{\mu^l_1\}, \tag{3.27}$$

für das nach (3.22) und (3.24) gilt

$$F(\mathbf{x}^1+\mu^*\mathbf{s}^1) > F(\mathbf{x}^2+\mu^*\mathbf{s}^2) > \ldots > F(\mathbf{x}^\varkappa+\mu^*\mathbf{s}^\varkappa). \tag{3.28}$$

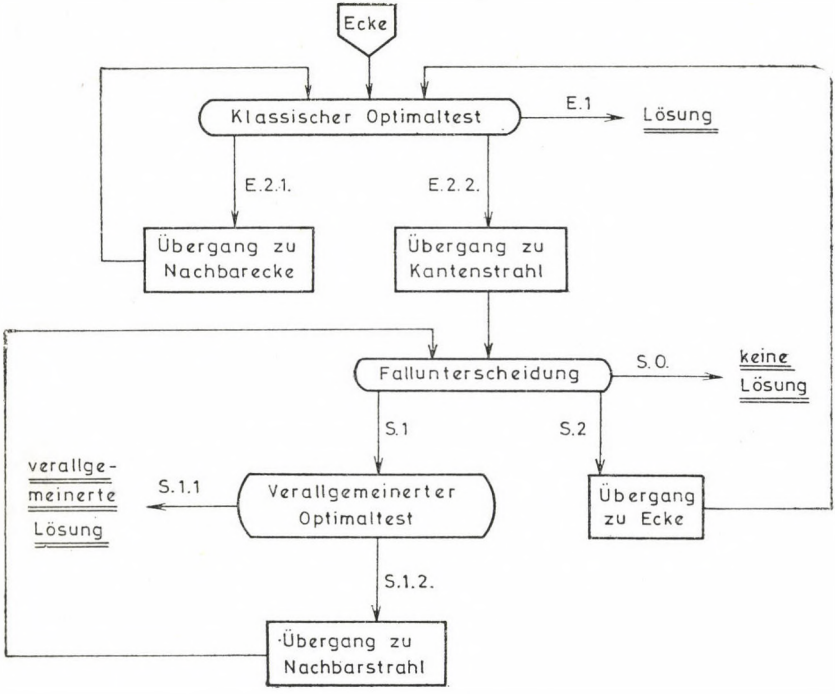


Bild 3 (Ablaufschema)

Das widerspricht der Annahme, folglich müssen die Strahlbasen (3.26) alle voneinander verschieden sein. Da es nur endlich viele verschiedene Strahlbasen von (2.1), (2.2) gibt, bricht die Kette (3.26) also nach endlich vielen Gliedern ab.

In der vom Algorithmus insgesamt erzeugten Kette von Ecken und Kantenstrahlen gilt für je zwei direkt aufeinanderfolgende Ecken $\mathbf{x}^1, \mathbf{x}^2$ stets $F(\mathbf{x}^2)<F(\mathbf{x}^1)$ und für je zwei durch Kantenstrahlen $S_1, \ldots, S_{k-1}$ $(k\geqq 2)$ verbundene Ecken $\mathbf{x}^1, \mathbf{x}^k$

$$F(\mathbf{x}^k) = \operatorname{Inf}\{F(\mathbf{x})|\mathbf{x}\in S_{k-1}\} \leqq \ldots \leqq \operatorname{Inf}(F(\mathbf{x})|\mathbf{x}\in S_1\} < F(\mathbf{x}^1), \tag{3.29}$$

also $F(\mathbf{x}^k)<F(\mathbf{x}^1)$. Das heißt, die Zielfunktionswerte fallen von Ecke zu Ecke streng. Folglich enthält die Kette nur endlich viele Ecken und ist somit, da zwischen je zwei Ecken auch nur endlich viele Kantenstrahlen durchlaufen werden, insgesamt endlich.

Der beschriebene Algorithmus liefert also nach endlich vielen Schritten die (gewöhnliche oder verallgemeinerte) Lösung von (P), sofern sie existiert. Eine bemerkenswerte Eigenschaft des Algorithmus ist die, daß trotz vorausgesetzter Nichtentartung das Erfülltsein des verallgemeinerten Optimaltests für das Vorliegen eines verallgemeinerten Minimums nur hinreichend, aber nicht notwendig ist. Eine Verschärfung des Kriteriums ist nicht ohne weiteres möglich, da bei einem Übergang S-S stets zugelassen werden muß, daß der Strahlgrenzwert von $F(\mathbf{x})$ nur schwach fällt — es gibt Beispiele, wo man andernfalls die Lösung nicht erreichen könnte.

3.4. *Realisierung des verallgemeinerten Optimaltests mittels asymptotischer Entwicklungen*

Der verallgemeinerte Optimaltest ist in der angegebenen Form nur sehr schwer zu realisieren. Schränken wir jedoch die Klasse der Zielfunktionen auf solche pseudolineare Funktionen ein, deren partielle Ableitungen $g_i(\lambda)$ gewisse asymptotische Entwicklungen besitzen, so können wir aus den zugehörigen Entwicklungskoeffizienten alle notwendigen Testgrößen gewinnen.

Ausführliche Darlegungen zu den folgenden Begriffen und Behauptungen sind z.B. bei BERG [1] zu finden.

Wir nennen eine über einem Intervall $[\lambda_0, \infty)$ definierte Funktionenfolge $\{\varphi_j(\lambda)\}_{j=0,1,\ldots}$ eine (asymptotische) Skala für $\lambda \to \infty$, wenn gilt

$$\varphi_j(\lambda) \not\equiv 0 \quad \text{und} \quad \varphi_{j+1}(\lambda) = o(\varphi_j(\lambda)) \quad (\lambda \to \infty) \quad \text{für} \quad j \in \{0, 1, \ldots\}. \tag{3.30}$$

Eine gegebene Funktion $h(\lambda)$ heißt für $\lambda \to \infty$ (asymptotisch) entwickelbar nach der Skala $\{\varphi_j(\lambda)\}$, wenn es Zahlen c_j ($j \in \{0, 1, \ldots\}$) gibt mit

$$h(\lambda) = \sum_{i=0}^{j} c_i \varphi_i(\lambda) + o(\varphi_j(\lambda)) \quad (\lambda \to \infty) \quad \text{für} \quad j \in \{0, 1, \ldots\}. \tag{3.31}$$

Wir schreiben dann

$$h(\lambda) \sim \sum_{j=0}^{\infty} c_j \varphi_j(\lambda) \quad (\lambda \to \infty). \tag{3.32}$$

Ist die Funktion $h(\lambda)$ für $\lambda \to \infty$ entwickelbar nach der Skala $\{\varphi_j(\lambda)\}$, so gilt für die Entwicklungskoeffizienten c_j

$$c_0 = \lim_{\lambda \to \infty} \frac{h(\lambda)}{\varphi_0(\lambda)}, \ldots, c_j = \lim_{\lambda \to \infty} \frac{h(\lambda) - \sum_{i=0}^{j-1} c_i \varphi_i(\lambda)}{\varphi_j(\lambda)}, \ldots . \tag{3.33}$$

LEMMA 2. *Die Funktion $h(\lambda)$ sei für $\lambda \to \infty$ entwickelbar nach einer Skala $\{\varphi_j(\lambda)\}$ mit $\varphi_j(\lambda) > 0$ für alle $j \in \{0, 1, \ldots\}$ und alle $\lambda \geqq \lambda_0$. Es existiere ein $j \in \{0, 1, \ldots\}$ mit $c_j \neq 0$ und $c_i = 0$ für alle $i < j$. Dann gilt für ein λ^**

$$c_j \left\{ \begin{matrix} > \\ < \end{matrix} \right\} 0 \Leftrightarrow h(\lambda) \left\{ \begin{matrix} > \\ < \end{matrix} \right\} 0 \quad \text{für} \quad \lambda \geqq \lambda^*. \tag{3.34}$$

Wir betrachten nun unser Problem (P). Im Falle der Unbeschränktheit des zulässigen Bereiches B möge $F(\mathbf{x})$ die folgende Zusatzvoraussetzung (Z) erfüllen:

(Z): Ist $[\mathbf{x}^0, \mathscr{J}, \sigma]$ eine Strahlbasis von (2.1), (2.2), so existiert zu jedem $i \in \mathscr{J}'$ eine Skala $\{\varphi_j^i(\lambda)\}$ mit $\varphi_j^i(\lambda) > 0$ für alle $j \in \{0, 1, \ldots\}$ und $\lambda \geqq \lambda_0$, nach der die Funktion

$$h_i(\lambda) = \begin{cases} g_i(\lambda) & \text{für} \quad i \in \mathscr{J}^+ \cup \sigma \\ g_i(\lambda) - g_\sigma(\lambda) & \text{für} \quad i \in \mathscr{J}^- \end{cases} \tag{3.35}$$

für $\lambda \to \infty$ entwickelbar ist. Ist $h_i(\lambda) \not\equiv 0$ für große λ, so soll die Funktion wenigstens einen nicht verschwindenden Entwicklungskoeffizienten besitzen.

Die gebrochen linearen Funktionen und speziellen Wurzelfunktionen (siehe (1.3) und (1.4)) genügen der Bedingung (Z). Hierfür läßt sich z.B. wählen $\varphi_j^i(\lambda) = \lambda^{-j}$ für alle $i \in \mathscr{J}'$.

Es sei im Algorithmus eine Strahlbasis $[\mathbf{x}^0, \mathscr{J}, \sigma]$ erreicht. Wegen (Z) gibt es Folgen $\{c_j^i\}$, so daß für obige $h_i(\lambda)$ gilt

$$h_i(\lambda) \sim \sum_{j=0}^{\infty} c_j^i \varphi_j^i(\lambda) \quad (\lambda \to \infty),\ i \in \mathscr{J}'. \tag{3.36}$$

Außerdem existiert zu jedem $i \in \mathscr{J}'$ mit $h_i(\lambda) \not\equiv 0$ für große λ ein (eindeutig bestimmtes) $j(i) \in \{0, 1, \ldots\}$ mit

$$c_{j(i)}^i \neq 0 \quad \text{und} \quad c_j^i = 0 \quad \text{für} \quad j < j(i). \tag{3.37}$$

Wir bilden die Größen

$$r_i = \begin{cases} c_{j(i)}^i, & \text{falls } j(i) \text{ existiert} \\ 0 & \text{sonst} \end{cases}, \quad i \in \mathscr{J}'. \tag{3.38}$$

Nach (Z) und Lemma 2 gibt es zu jedem $i \in \mathscr{J}'$ ein λ^i mit

$$r_i \begin{Bmatrix} > \\ = \\ < \end{Bmatrix} 0 \Leftrightarrow h_i(\lambda) \begin{Bmatrix} > \\ = \\ < \end{Bmatrix} 0 \quad \text{für} \quad \lambda \geqq \lambda^i. \tag{3.39}$$

Damit vereinfachen sich die Tests S.1 und S.2 wie folgt:

Fall S.1: $r_\sigma < 0$.

Fall S.1.1: $r_i \geqq 0$ für alle $i \in \mathscr{J}$.

Fall S.1.2: $r_\alpha < 0$ für ein $\alpha \in \mathscr{J}$.

Fall S.2: $r_\sigma \geqq 0$.

Die einzige Schwierigkeit des beschriebenen Vorgehens besteht offensichtlich darin, die Entwicklungskoeffizienten $c_{j(i)}^i$ zu bestimmen. Für die betrachteten speziellen Zielfunktionen kann dies jedoch nach einfachen Regeln direkt im Tableau realisiert werden.

4. Realisierung des Algorithmus für Beispielklassen pseudolinearer Zielfunktionen

Der Algorithmus zur Lösung von (P) soll nun noch für zwei spezielle Zielfunktionen konkret realisiert werden.

4.1. *Spezielle Wurzelfunktion als Zielfunktion*

Betrachtet wird das Problem (P) mit der Zielfunktion

$$F(\mathbf{x}) = \mathbf{d}^\top \mathbf{x} + d_0 + \sqrt{(\mathbf{d}^\top \mathbf{x} + d_0)^2 + \mathbf{e}^\top \mathbf{x} + e_0} \tag{4.1}$$

mit $\mathbf{d}, \mathbf{e} \in \mathbf{R}^n$ und $d_0, e_0 \in \mathbf{R}^1$, sowie mit $\mathbf{e}^\top \mathbf{x} + e_0 \geqq 0$ und ($\mathbf{e}^\top \mathbf{x} + e_0 > 0$ oder $\mathbf{d}^\top \mathbf{x} + d_0 > 0$) für alle $\mathbf{x} \in B$. Diese Funktion (vgl. (1.4) mit $l_3(\mathbf{x}) \equiv 1$) ist pseudolinear über B.

Wir gehen von einer Eckbasis $[\mathbf{x}^0, \mathscr{J}]$ aus und bilden zunächst analog zu (3.1) die auf die Nichtbasisvariablen $\hat{\mathbf{x}} = (x_{m+1}, \ldots, x_n)^\top$ umgerechnete Zielfunktion

$$f(\hat{\mathbf{x}}) = \mathbf{d}^{0\top} \hat{\mathbf{x}} + d_0^0 + \sqrt{(\mathbf{d}^{0\top} \hat{\mathbf{x}} + d_0^0)^2 + \mathbf{e}^{0\top} \hat{\mathbf{x}} + e_0^0}. \tag{4.2}$$

Es gilt

$$g_i = \frac{\partial f}{\partial x_i}(\hat{\mathbf{x}})\bigg|_{\hat{\mathbf{x}}=0} = d_i^0 + \frac{2 d_i^0 d_0^0 + e_i^0}{2\sqrt{(d_0^0)^2 + e_0^0}}, \quad i \in \mathscr{J}'. \tag{4.3}$$

Daraus folgt für jedes $i \in \mathscr{J}'$

$$g_i \begin{Bmatrix} > \\ = \\ < \end{Bmatrix} 0 \Leftrightarrow 2 d_i^0 \left(d_0^0 + \sqrt{(d_0^0)^2 + e_0^0}\right) + e_i^0 \begin{Bmatrix} > \\ = \\ < \end{Bmatrix} 0. \tag{4.4}$$

Wir stellen das Simplextableau zu $[\mathbf{x}^0, \mathscr{J}]$ auf und erweitern es um die drei Zeilen $(\mathbf{d}^0) = (d_{m+1}^0, \ldots, d_n^0, -d_0^0)$, $(\mathbf{e}^0) = (e_{m+1}^0, \ldots, e_n^0, -e_0^0)$ und $(\mathbf{p}^0) = (p_{m+1}^0, \ldots, p_n^0, p_0^0)$ mit

$$p_0^0 = d_0^0 + \sqrt{(d_0^0)^2 + e_0^0}, \quad p_i^0 = 2 d_i^0 p_0^0 + e_i^0 \quad \text{für} \quad i \in \mathscr{J}'. \tag{4.5}$$

Das erweiterte Tableau hat also die Gestalt

(4.6)

	x_{m+1}	. . .	x_n	$\mathbf{b}^0$
x_1	$a_{1,m+1}^0$	. . .	$a_{1,n}^0$	b_1^0
.	.		.	.
.	.		.	.
x_m	$a_{m,m+1}^0$	. . .	$a_{m,n}^0$	b_m^0
$\mathbf{d}^0$	d_{m+1}^0	. . .	d_n^0	$-d_0^0$
$\mathbf{e}^0$	e_{m+1}^0	. . .	e_n^0	$-e_0^0$
$\mathbf{p}^0$	p_{m+1}^0	. . .	p_n^0	p_0^0

Der Optimaltest wird nun wie folgt ausgeführt:

Fall E.1: $p_i^0 \geqq 0$ für alle $i \in \mathscr{I}'$.

Dann ist $\mathbf{x}^0$ für (P) optimal und $F(\mathbf{x}^0) = p_0^0$.

Fall E.2: $p_i^0 < 0$ für gewisse $i \in \mathscr{I}'$.

Dann wählen wir ein $\alpha \in \mathscr{I}'$ mit $p_\alpha^0 < 0$ und unterscheiden

Fall E.2.1: $\mathbf{a}_\alpha^0 \not\leqq \mathbf{0}$, d.h. $a_{i\alpha}^0 > 0$ für gewisse $i \in \mathscr{I}$.

Dann rechnen wir mit der Pivotspalte α das Tableau (4.6) einschließlich der Zeilen $(\mathbf{d}^0)$ und $(\mathbf{e}^0)$ wie üblich um und gelangen zu einer benachbarten Eckbasis. Die beiden umgerechneten Zusatzzeilen haben für diese dieselbe Bedeutung wie $(\mathbf{d}^0)$ und $(\mathbf{e}^0)$ nach (4.2) für $[\mathbf{x}^0, \mathscr{I}]$. Im erreichten Tableau wird die letzte Zeile $(\mathbf{p}^1)$ nun aualog wie $(\mathbf{p}^0)$ in (4.6) berechnet. Für die erhaltene Ecke $\mathbf{x}^1$ gilt $F(\mathbf{x}^1) = p_0^1 < p_0^0 = F(\mathbf{x}^0)$.

Fall E.2.2: $\mathbf{a}_\alpha^0 \leqq \mathbf{0}$.

Dann wird durch die Markierung von x_α zu einer inzidenten Strahlbasis übergegangen. Die Zeilen von Tableau (4.6) werden dabei außer der Zeile $(\mathbf{p}^0)$ unverändert übernommen. Für den definierten Kantenstrahl S_0 gilt $\operatorname{Inf}\{F(\mathbf{x}) | \mathbf{x} \in S_0\} < F(\mathbf{x}^0) = p_0^0$.

Im Algorithmus sei eine Strahlbasis $[\mathbf{x}^0, \mathscr{I}, \sigma]$ mit den zugehörigen (die Funktion $f(\hat{\mathbf{x}})$ analog wie in (4.2) bestimmenden) Zeilen $(\mathbf{d}^0)$ und $(\mathbf{e}^0)$ erreicht. Wir bilden nach (3.7)

$$g_i(\lambda) = d_i^0 + \frac{2d_i^0(d_\sigma^0\lambda + d_0^0) + e_i^0}{2\sqrt{(d_\sigma^0\lambda + d_0^0)^2 + e_\sigma^0\lambda + e_0^0}}, \quad i \in \mathscr{I}'. \tag{4.7}$$

Für den durch die Strahlbasis definierten Kantenstrahl $S_0 = \{\mathbf{x} \in \mathbf{R}^n | \mathbf{x} = \mathbf{x}^0 + \lambda\mathbf{s}^0, \lambda \geqq \lambda_0\}$ haben wir

$$F(\mathbf{x}^0 + \lambda\mathbf{s}^0) = d_\sigma^0\lambda + d_0^0 + \sqrt{(d_\sigma^0\lambda + d_0^0)^2 + e_\sigma^0\lambda + e_0^0}. \tag{4.8}$$

Da $\mathbf{x}^0 + \lambda\mathbf{s}^0$ für große λ in B liegt, muß für große λ gelten $e_\sigma^0\lambda + e_0^0 \geqq 0$. Das heißt, es ist $(e_\sigma^0, e_0^0) \geqq (0, 0)$. Wir machen nun eine vollständige Fallunterscheidung.

a) $(d_\sigma^0, e_\sigma^0) > (0, 0)$: Dann ist $\lim_{\lambda \to \infty} F(\mathbf{x}^0 + \lambda\mathbf{s}^0) = \infty$, dieser Fall kann nicht eintreten.

b) $(d_\sigma^0, e_\sigma^0) = (0, 0)$: Dann haben wir

$$F(\mathbf{x}^0 + \lambda\mathbf{s}^0) \equiv d_0^0 + \sqrt{(d_0^0)^2 + e_0^0}, \quad g_\sigma(\lambda) \equiv 0. \tag{4.9}$$

c) $(d_\sigma^0, e_\sigma^0) < (0, 0)$: Wegen $e_\sigma^0 \geqq 0$ ist dann $d_\sigma^0 < 0$ und

$$\lim_{\lambda \to \infty} F(\mathbf{x}^0 + \lambda\mathbf{s}^0) = -\frac{e_\sigma^0}{2d_\sigma^0}. \tag{4.10}$$

Die Funktionen $g_i(\lambda)$ lassen sich für $\lambda \to \infty$ nach der Skala $\{\lambda^{-j}\}$ entwickeln. Das heißt, es gibt Zahlen c_j^i mit

$$g_i(\lambda) \sim \sum_{j=0}^{\infty} c_j^i \lambda^{-j} \quad (\lambda \to \infty), \quad i \in \mathscr{I}'. \tag{4.11}$$

Die c_j^i erhalten wir nach (3.33) mittels

$$c_0^i = \lim_{\lambda\to\infty} g_i(\lambda) = 0, \quad c_1^i = \lim_{\lambda\to\infty} g_i(\lambda)\cdot\lambda = \frac{d_i^0 e_\sigma^0 - e_i^0 d_\sigma^0}{2(d_\sigma^0)^2}. \tag{4.12}$$

Ist $c_1^i=0$ und $d_i^0=0$, so folgt $e_i^0=0$ und somit $g_i(\lambda)\equiv 0$. Für alle $i\in\mathscr{I}'$ mit $c_1^i=0$ und $d_i^0\neq 0$ bestimmen wir

$$c_2^i = \lim_{\lambda\to\infty} g_i(\lambda)\cdot\lambda^2 = \frac{1}{2(d_\sigma^0)^2}\left[d_i^0 e_0^0 - e_i^0 d_0^0 - \frac{(e_i^0)^2}{4d_i^0}\right]. \tag{4.13}$$

Für jedes $i\in\mathscr{I}'$ mit $d_i^0\neq 0$ und $c_1^i=c_2^i=0$ gilt

$$g_i(\lambda) = d_i^0\left[1+\operatorname{sgn}\left(d_\sigma^0\lambda + d_0^0 + \frac{e_i^0}{2d_i^0}\right)\right], \tag{4.14}$$

wegen $d_\sigma^0<0$ folgt daraus $g_i(\lambda)\equiv 0$ für große λ.
Insgesamt gilt also

$$g_i(\lambda) = \frac{c_1^i}{\lambda} + \frac{c_2^i}{\lambda^2} + o\left(\frac{1}{\lambda^2}\right) \quad (\lambda\to\infty), \quad i\in\mathscr{I}', \tag{4.15}$$

mit $g_i(\lambda)\equiv 0$ für große λ, falls $c_1^i=d_i^0=0$ oder $c_1^i=c_2^i=0$.

Wir definieren für jedes $i\in\mathscr{I}'$

$$q_0^0 = \begin{cases} d_0^0 + \sqrt{(d_0^0)^2 + e_0^0} & \text{für} \quad d_\sigma^0 = 0 \\ -\dfrac{e_\sigma^0}{2d_\sigma^0} & \text{für} \quad d_\sigma^0 \neq 0, \end{cases} \tag{4.16}$$

$$q_i^0 = \begin{cases} u_i^0 = d_i^0 e_\sigma^0 - e_i^0 d_\sigma^0 & \text{für} \quad u_0^i \neq 0 \quad \text{oder} \quad d_i^0 = 0 \\ v_i^0 = d_i^0 e_0^0 - e_i^0 d_0^0 - \dfrac{(e_i^0)^2}{4d_i^0} & \text{für} \quad u_i^0 = 0 \quad \text{und} \quad d_i^0 \neq 0 \end{cases} \tag{4.17}$$

und bilden das Tableau

(4.18)

		*			
	x_{m+1}	$\cdots$	x_6	$\cdots$	x_n	b^0
x_1	$a^0_{1,m+1}$	$\cdots$	$a^0_{1,6}$	$\cdots$	$a^0_{1,n}$	b^0_1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_m	$a^0_{m,m+1}$	$\cdots$	$a^0_{m,6}$	$\cdots$	$a^0_{m,n}$	b^0_m
d^0	d^0_{m+1}	$\cdots$	d^0_6	$\cdots$	d^0_n	$-d^0_0$
e^0	e^0_{m+1}	$\cdots$	e^0_6	$\cdots$	e^0_n	$-e^0_0$
q^0	q^0_{m+1}	$\cdots$	q^0_6	$\cdots$	q^0_n	q^0_0

indem wir das vorliegende Strahltableau um die Zeile $(\mathbf{q}^0)=(q^0_{m+1},\ldots,q^0_n,q^0_0)$ erweitern und jede Spalte $i\in\mathscr{J}$ mit $q_i^0=v_i^0$ mit "0" markieren. Die Tests nehmen dann (vgl. die Ergebnisse von Abschnitt 3.4) folgende Gestalt an:

Fall S.1: $q_\sigma^0 < 0$.

Dann wird unterschieden

Fall S.1.1: Für alle $i \in \mathscr{J}$ gilt

$$q_i^0 \geqq \begin{cases} 0, & \text{falls } \mathbf{a}_i^0 \not\leqq \mathbf{0} \text{ oder } i \text{ nicht mit "0" markiert} \\ q_\sigma^0, & \text{falls } \mathbf{a}_i^0 \leqq \mathbf{0} \text{ und } i \text{ mit "0" markiert.} \end{cases} \tag{4.19}$$

In diesem Falle liefert der Kantenstrahl S_0 das verallgemeinerte Minimum $\lim_{\lambda \to \infty} F(\mathbf{x}^0 + \lambda \mathbf{s}^0) = q_0^0$.

Fall S.1.2: Für ein $\alpha \in \mathscr{J}$ gilt

$$q_\alpha^0 < \begin{cases} 0, & \text{falls } \mathbf{a}_\alpha^0 \not\leqq 0 \text{ oder } \alpha \text{ nicht mit "0" markiert} \\ q_\sigma^0, & \text{falls } \mathbf{a}_\alpha^0 \leqq 0 \text{ und } \alpha \text{ mit "0" markiert.} \end{cases} \tag{4.20}$$

Dann führen wir mit dem Index α einen Übergang S-S aus und behandeln dabei die Zeilen $(\mathbf{d}^0)$ und $(\mathbf{e}^0)$ ebenso wie die anderen. Sie haben danach für die neue Strahlbasis dieselbe Bedeutung wie vorher für $[\mathbf{x}^0, \mathscr{J}, \sigma]$. Im erreichten Tableau wird die q-Zeile wieder analog wie in (4.18) bestimmt.

Fall S.2: $q_\sigma^0 \geqq 0$.

Wir führen dann einen Übergang S-E aus, wobei die Zeilen $(\mathbf{d}^0)$ und $(\mathbf{e}^0)$ mit transformiert werden. Im neuen Tableau wird, da eine Eckbasis vorliegt, die p-Zeile analog wie $(\mathbf{p}^0)$ in (4.6) berechnet.

Damit ist das Verfahren zu (P) für den Fall der speziellen Zielfunktion (4.1) vollständig beschrieben. Alle darin benötigten Größen lassen sich nach einfachen Regeln direkt im Tableau bestimmen. Zur Illustration des Algorithmus geben wir noch ein Zahlenbeispiel an.

Beispiel. Es sei $B \subset \mathbf{R}^5$ gegeben durch

$$\begin{aligned} x_1 \quad + x_3 - x_4 \quad &= 1 \\ 3x_1 \quad - 4x_3 \quad - x_5 &= 6 \\ -x_1 + x_2 + x_3 + x_4 \quad &= 2 \\ x_1, x_2, x_3, x_4, x_5 &\geqq 0, \end{aligned} \tag{4.21}$$

und es sei

$$F(\mathbf{x}) = -\frac{5}{2}x_1 + x_2 + \sqrt{\left(-\frac{5}{2}x_1 + x_2\right)^2 - 3x_2 - 5x_3 + 5x_4 + 9}. \tag{4.22}$$

Es läßt sich zeigen, daß für alle $\mathbf{x} \in B$ gilt $-3x_2 - 5x_3 + 5x_4 + 9 \geqq 0$, wobei die Ungleichung für die $\mathbf{x} \in B$ mit $-\frac{5}{2}x_1 + x_2 \leqq 0$ sogar streng erfüllt ist. Also ist $F(\mathbf{x})$ pseudolinear über B.

Wir beginnen den Algorithmus mit der Anfangsecke $\mathbf{x}^0=(0, 0, 2, 1, 14)^\top$. Dann ist $\tilde{\mathbf{x}}=(x_3, x_4, x_5)^\top$, $\hat{\mathbf{x}}=(x_1, x_2)^\top$ und

$$f(\hat{\mathbf{x}}) = -\frac{5}{2}x_1+x_2+\sqrt{\left(-\frac{5}{2}x_1+x_2\right)^2+5x_1-3x_2+4}. \tag{4.23}$$

Das Anfangstableau hat damit die Gestalt

(4.24)

	x_1	x_2	$\mathbf{b}^0$
x_3	−1	1	2
x_4	−2	1	1
x_5	−1	4	14
d^0	$-\frac{5}{2}$	1	0
e^0	5	−3	−4
p^0	−5	1	2

Wegen $p_1^0=-5<0$ und $\mathbf{a}_1^0=(-1, -2, -1)^\top \leqq \mathbf{0}$ tritt der Fall E.2.2 ein, und wir gehen durch die Markierung von x_1 zu einer Strahlbasis über. Dabei erhalten wir das Tableau

(4.25)

	x_1	x_2	$\mathbf{b}^0$	Test
x_3	−1	1	2	(1,2)
x_4	−2	1	1	(2,1)
x_5	−1	[4]	14	$\left(\frac{1}{4}, \frac{7}{2}\right)$
d^0	$-\frac{5}{2}$	1	0	
e^0	5	−3	−4	
q^0	$-\frac{15}{2}$	$-\frac{5}{2}$	1	

Es ist $q_1^0=-\frac{15}{2}<0$ (Fall S.1) und $q_2^0=-\frac{5}{2}<0$ mit $\mathbf{a}_2^0=(1, 1, 4)^\top \not\leqq \mathbf{0}$. Daher gehen wir nach Fall S.1.2 vor und führen einen Übergang S-S-1 mit dem Index 2 aus. Hierzu bilden wir

$$\text{Lex. Min}\left\{\left(-\frac{a_{i1}^0}{a_{i2}^0}, \frac{b_i^0}{a_{i2}^0}\right)\middle| i\in\{3, 4, 5\}\right\} = \left(-\frac{a_{51}^0}{a_{52}^0}, \frac{b_5^0}{a_{52}^0}\right), \tag{4.26}$$

erhalten das Pivotelement a_{52}^0 und damit ein neues Tableau

(4.27)

	x_1	x_5	$\mathbf{b}^1$	Test
x_3	$\boxed{-\frac{3}{4}}$	$-\frac{1}{4}$	$-\frac{3}{2}$	2
x_4	$-\frac{7}{4}$	$-\frac{1}{4}$	$-\frac{5}{2}$	$-\frac{10}{7}$
x_2	$-\frac{1}{4}$	$\frac{1}{4}$	$\frac{7}{2}$	/
$\mathbf{d}^1$	$-\frac{9}{4}$	$-\frac{1}{4}$	$-\frac{7}{2}$	
$\mathbf{e}^1$	$\frac{17}{4}$	$\frac{3}{4}$	$\frac{13}{2}$	
$\mathbf{q}^1$	$\frac{253}{144}$	$-\frac{5}{2}$	$\frac{17}{18}$	

Wegen $q_1^1=\frac{253}{144}\geqq 0$ (Fall S.2) und $\mathbf{b}^1=\left(-\frac{3}{2}, -\frac{5}{2}, \frac{7}{2}\right)^\top \ngeqq \mathbf{0}$ erfolgt ein Übergang S-E-2. Mittels

$$\text{Max}\left\{\frac{b_i^1}{a_{i1}^1}\middle| i\in\{3,4\}\right\}=\frac{b_3^1}{a_{31}^1} \tag{4.28}$$

erhalten wir das negative Pivotelement a_{31}^1. Der Übergang liefert das Tableau

(4.29)

	x_3	x_5	$\mathbf{b}^2$
x_1	$-\frac{4}{3}$	$\frac{1}{3}$	2
x_4	$-\frac{7}{3}$	$\frac{1}{3}$	1
x_2	$-\frac{1}{3}$	$\frac{1}{3}$	4
$\mathbf{d}^2$	-3	$\frac{1}{2}$	1
$\mathbf{e}^2$	$\frac{17}{3}$	$-\frac{2}{3}$	-2
$\mathbf{p}^2$	$\frac{35}{3}-6\sqrt{3}$	$-\frac{5}{3}+\sqrt{3}$	$-1+\sqrt{3}$

Wegen $p_3^2\approx 1{,}29\geqq 0$ und $p_5^2\approx 0{,}06\geqq 0$ haben wir damit das Optimum erreicht. $F(\mathbf{x})$ nimmt also in der Ecke $(2, 4, 0, 1, 0)^\top$ sein Minimum $p_0^2=-1+\sqrt{3}$ über B an.

Wir veranschaulichen unser Vorgehen $\mathbf{x}^0\to S_0\to S_1\to\mathbf{x}^2$ in Bild 4 im Raum der Nichtbasisvariablen x_1 und x_2 der Anfangs-Eckbasis. Die (gerundeten) Eckpunktwerte bzw. Strahlgrenzwerte der Zielfunktion sind in Klammern angetragen.

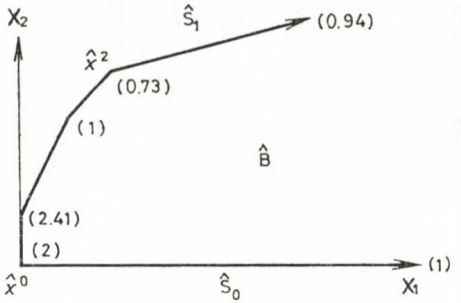


Bild 4

4.2. *Gebrochen lineare Zielfunktion (Hyperbolische Optimierung)*

Wir betrachten nun unser Problem (P) mit der gebrochen linearen Zielfunktion

$$F(\mathbf{x}) = \frac{\mathbf{z}^\top \mathbf{x} + z_0}{\mathbf{n}^\top \mathbf{x} + n_0} \tag{4.30}$$

mit $\mathbf{z}, \mathbf{n} \in \mathbf{R}^n$; $z_0, n_0 \in \mathbf{R}^1$ und $\mathbf{n}^\top \mathbf{x} + n_0 > 0$ für alle $\mathbf{x} \in B$. Für diese spezielle pseudolineare Zielfunktion nimmt der Algorithmus eine ähnlich einfache Gestalt wie im vorangegangenen Falle an.

Wir gehen von einer Eckbasis $[\mathbf{x}^0, \mathscr{J}]$ aus und bilden die auf die Nichtbasisvariablen $\hat{\mathbf{x}}$ umgerechnete Zielfunktion

$$f(\hat{\mathbf{x}}) = \frac{\mathbf{z}^{0\top} \hat{\mathbf{x}} + z_0^0}{\mathbf{n}^{0\top} \hat{\mathbf{x}} + n_0^0}. \tag{4.31}$$

Das Simplextableau zu $[\mathbf{x}^0, \mathscr{J}]$ erweitern wir um die drei Zeilen $(\mathbf{z}^0) = (z_{m+1}, \ldots, z_n^0, -z_0^0)$, $(\mathbf{n}^0) = (n_{m+1}^0, \ldots, n_n^0, -n_0^0)$ und $(\mathbf{p}^0) = (p_{m+1}^0, \ldots, p_n^0, p_0^0)$ mit

$$p_0^0 = \frac{z_0^0}{n_0^0}, \quad p_i^0 = z_i^0 n_0^0 - n_i^0 z_0^0 \quad \text{für} \quad i \in \mathscr{J}'. \tag{4.32}$$

Der Optimaltest wird nun ebenso wie unter 4.1. anhand der p_i^0 $(i \in \mathscr{J}')$ ausgeführt. Bei einem Basisübergang werden die Zeilen $(\mathbf{z}^0)$ und $(\mathbf{n}^0)$ wie die anderen im Tableau behandelt.

Haben wir eine Strahlbasis $[\mathbf{x}^0, \mathscr{J}, \sigma]$ mit den entsprechenden Zeilen $(\mathbf{z}^0)$ und $(\mathbf{n}^0)$ erreicht, so gelangen wir durch analoge Fallunterscheidungen wie für die speziellen Wurzelfunktionen und durch die Entwicklung der Funktionen

$$g_i(\lambda) = \frac{(z_i^0 n_\sigma^0 - n_i^0 z_\sigma^0)\lambda + z_i^0 n_0^0 - n_i^0 z_0^0}{(n_\sigma^0 \lambda + n_0^0)^2}, \quad i \in \mathscr{J}', \tag{4.33}$$

für $\lambda \to \infty$ nach der Skala $\{\lambda^{-j}\}$ zu den Größen

$$q_0^0 = \begin{cases} -\infty & \text{für} \quad n_\sigma^0 = 0 \quad \text{und} \quad z_\sigma^0 < 0 \\ \dfrac{z_0^0}{n_0^0} & \text{für} \quad n_\sigma^0 = 0 \quad \text{und} \quad z_\sigma^0 = 0 \\ \dfrac{z_\sigma^0}{n_\sigma^0} & \text{für} \quad n_\sigma^0 > 0 \end{cases} \tag{4.34}$$

$$q_i^0 = \begin{cases} u_i^0 = z_i^0 n_\sigma^0 - n_i^0 z_\sigma^0 & \text{für} \quad u_i^0 \neq 0 \\ v_i^0 = z_i^0 n_0^0 - n_i^0 z_0^0 & \text{für} \quad u_i^0 = 0 \end{cases}, \quad i \in \mathscr{I}'. \tag{4.35}$$

Die Zeile $(\mathbf{q}^0)$ fügen wir zum erreichten Tableau hinzu und markieren wieder jede Spalte $i \in \mathscr{J}$ mit $q_i^0 = v_i^0$ mit "0". Die Tests S.0, S.1 und S.2 werden nun ebenso wie unter 4.1. realisiert.

Damit haben wir für den Fall der gebrochen linearen Zielfunktion ein Lösungsverfahren zu (P) erhalten, das bezüglich der reinen Eckenübergänge rechnerisch mit dem Verfahren von MARTOS [4] zur hyperbolischen Optimierung übereinstimmt. Infolge des Einsatzes von Strahlbasen lassen sich damit aber auch hyperbolische Probleme mit unbeschränkten zulässigen Bereichen stets direkt und vollständig lösen.

Die beiden angegebenen konkreten Realisierungen des Algorithmus zum pseudolinearen Optimierungsproblem (P) besitzen eine starke Ähnlichkeit mit dem Simplexalgorithmus aus der linearen Optimierung. Sie erfordern in jedem der endlich vielen Schritte kaum mehr Rechenaufwand als dieser und können daher als sehr effektive und zuverlässige Lösungsmethoden angesehen werden.

LITERATUR

[1] BERG, L.: *Asymptotische Darstellungen und Entwicklungen,* Berlin, 1971.
[2] HARTWIG, H.: *Pseudolineare Optimierung* (Dissertation A), Leipzig, 1974.
[3] MANGASARIAN, O. L.: Pseudo-convex functions, *SIAM J. Control* 3 (1965), 284—289.
[4] MARTOS, B.: Hiperbolikus programozás, *Publ. Math. Inst. Hung. Ac. Sc.* **5** (1960), 383—406.
[5] MARTOS, B.: Quasi-convexity and quasi-monotonicity in nonlinear programming, *Stud. Sc. Math. Hung.* **2** (1967), 265—273.
[6] STOER, J.—WITZGALL, CH.: *Convexity and optimization in finite dimensions* I, Berlin—Heidelberg —New York, 1970.

701 Leipzig, Karl-Marx-Platz, Hauptgebäude der Karl-Marx-Universität,

Sektion Mathematik, DDR

(Eingegangen: 10 März, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 237—242.

BEMERKUNGEN ZUM BELEUCHTUNGSPROBLEM VON L. FEJES TÓTH

von

R. FISCHER

1. Wie muß man Lampen auf einer in beiden Richtungen unendlich langen Straße verteilen, damit die Helligkeit in dem am schlechtesten beleuchteten Punkt maximal ist, wenn die durchschnittliche Lampenanzahl pro 1 km vorgegeben ist? Dieses von L. FEJES TÓTH stammende Problem führt unter idealisierenden Annahmen auf die Frage, für welche doppelseitigen Folgen $(\ldots<x_{-2}<x_{-1}<x_0<x_1<x_2<\ldots)$ mit beschränkter oberer asymptotischer Dichte die Zahl

$$\inf_{x\in\mathbf{R}} \sum_{i\in\mathbf{Z}} f(|x-x_i|)$$

maximal ist ("optimale Anordnungen"). f ist dabei die von der Lampenart abhängige sogenannte "Beleuchtungsfunktion" d.h. $f(x)$ ist die von einer festen Lampe herrührende Helligkeit im Abstand x vom Fußpunkt der Lampe an (vgl. [1]). Wir verlagen: f ist stetig, nirgends zunehmend und $\int_0^\infty f(x)\,dx<\infty$. In [2] wurde eine hinreichende Bedingung angegeben, unter der die äquidistante Verteilung (=eindimensionales Punktegitter) stets eine optimale Anordnung ist. Hier sollen einige im Anschluß an [2] offen gebliebenen Fragen untersucht und ein Resultat für die physikalische Beleuchtung $\left(\text{mit der Beleuchtungsfunktion } f(x)=\frac{1}{(h^2+x^2)^{\frac{3}{2}}}\right)$ hergeleitet werden. Die Bekanntschaft mit den Resultaten und der Notation in [2] wird vorausgesetzt.

2. In [2] wurde bereits vermutet, daß man für den Satz 2 die Bedingung der strikten Konvexität für f durch die der Konvexität ersetzen kann. Dies konnte jedoch mit der dort verwendeten Methode, die wesentlich auf der Eindeutigkeit der optimalen Anordnung im Falle einer Straße endlicher Länge fußte, nicht gezeigt werden. Es soll hier mit einer anderen Methode das Resultat von Lemma 4 in [2], in dem eine Straße endlicher Länge betrachtet wird, ohne die Voraussetzung der Striktheit der Konvexität bewiesen werden, womit man dann auch Satz 2 von [2] ohne diese Voraussetzung erhält.

LEMMA 1. *Es sei $a_1, a_2, \ldots, a_n$ eine endliche Folge reeller Zahlen. Dann gibt es eine natürliche Zahl k mit $0\leqq k\leqq n$, sodaß gilt: Für $r=k+1, k+2, \ldots, n$ ist $a_{k+1}+\ldots+a_r\geqq 0$ und für $s=k, k-1, \ldots, 1$ ist $a_k+a_{k-1}+\ldots+a_s\leqq 0$.* (Leere Summe=0.)

BEWEIS. Vollständige Induktion nach n: Ist jede Partialsumme von $a_1+a_2+\ldots+a_n$ nicht negativ, dann ist mit $k=0$ alles gezeigt. Andernfalls gibt es $i=\min\{\nu|a_1+a_2+\ldots+a_\nu<0\}$. Auf die Zahlen $a_{i+1}, a_{i+2}, \ldots, a_n$ wenden wir die Induktionsvoraussetzung

an und erhalten: Es existiert k, $i \leqq k \leqq n$, sodaß jede Partialsumme von $a_{k+1}+\ldots+a_n$ nicht negativ und jede Partialsumme von $a_k+a_{k-1}+\ldots+a_{i+1}$ nicht positiv ist. Weiter ist wegen der Wahl von i jede Partialsumme von $a_k+a_{k-1}+\ldots+a_1$ nicht positiv.

LEMMA 2. *Im Falle einer konvexen Beleuchtungsfunktion gilt für Straßen endlicher Länge mit stetiger Grundbeleuchtung g und mit r Lampen: Ist* $(x_1, \ldots, x_r)$ *eine optimale Anordnung, so ist für jede andere Anordnung* $(x_1', x_2', \ldots, x_r')$ *die Ungleichung*

$$\max_{0 \leqq i \leqq r} m_i(x_1', \ldots, x_r') \geqq m(x_1, \ldots, x_r)$$

richtig.

BEWEIS. Es sei $\delta_i = x_i - x_i'$, $1 \leqq i \leqq r$. Nach Lemma 1 gibt es k, $0 \leqq k \leqq r$, sodaß jede Partialsumme von $\delta_{k+1}+\delta_{k+2}+\ldots+\delta_r$ nicht negativ und jede von $\delta_k+\delta_{k+1}+\ldots+\delta_1$ nicht positiv ist. Wir gehen von der Anordnung $(x_1', x_2', \ldots, x_r')$ aus und betrachten die Beleuchtung im Intervall $[x_k', x_{k+1}']$. Wir wollen nun die x_i' mit $i = k+1, \ldots, r$ so in die entsprechenden x_i schrittweise überführen, daß jede Bewegung eines x_i' nach links durch eine Bewegung eines x_j' mit $k+1 \leqq j < i$ nach rechts um mindestens den gleichen Betrag „ausgeglichen" wird. Mit Hilfe von Zerlegungen dieser Bewegungen in Teilbewegungen sieht man, daß dies sicher möglich ist, da alle Teilsummen $\delta_{k+1}+\delta_{k+2}+\ldots+\delta_v \geqq 0$ sind $(k+1 \leqq v \leqq r)$. Wegen der Konvexität von f gilt für $x \in [x_k', x_{k+1}']$:

$$f(|x-(x_j'+\delta)|)+f(|x-(x_i'-\delta)|) \leqq f(|x-x_j'|)+f(|x-x_i'|), \quad \delta > 0,$$

d.h. bei jeder Teilbewegung wird die Helligkeit in jedem Punkt von $[x_k', x_{k+1}']$ nicht größer, und damit auch bei der gesamten Verschiebung nicht. Durch eine analoge Betrachtung der Verschiebung der Punkte $x_1', x_2', \ldots, x_k'$ in $x_1, \ldots, x_k$ erhält man: Durch den Übergang $(x_1', \ldots, x_r') \to (x_1, \ldots, x_r)$ steigt die Beleuchtung in keinem Punkt des Intervalles $[x_\kappa', x_{k+1}']$. Wegen $x_k \leqq x_k' \leqq x_{\kappa+1}' \leqq x_{k+1}$ folgt daraus

$$\max_{0 \leqq i \leqq r} m_i(x_1', \ldots, x_r') \geqq m_k(x_1', \ldots, x_r') \geqq m_k(x_1, \ldots, x_r) \geqq m(x_1, \ldots, x_r),$$

womit Lemma 2 bewiesen ist.

Lemma 2 entspricht genau Lemma 4 in [2].

In [2] wurde die Bedingung der Beschränktheit der oberen asymptotischen Dichte der Folgen $(\ldots < x_{-2} < x_{-1} < x_0 < x_1 < x_2 < \ldots)$ so interpretiert:

$$\limsup_{n \to \pm\infty} \left| \frac{n}{x_n} \right| \leqq \frac{1}{d}$$

(d = „mittlerer Lampenabstand"), d.h. diese Bedingung wurde für jede Halbgerade verlangt. Die übliche Interpretation:

$$\limsup_{R \to \infty} \frac{N(R)}{2R} \leqq \frac{1}{d}, \quad N(R) = \operatorname{card}\{x_i \mid -R \leqq x_i \leqq R\}$$

stellt eine schwächere Voraussetzung dar. Wir wollen Optimalität ab nun bezüglich dieser Bedingung verstehen.

SATZ 1. *Für konvexe Beleuchtungsfunktionen ist die gitterförmige Anordnung* $(x_n)_{n\in\mathbf{Z}}$ *mit* $x_n=n\cdot d$ $(d>0)$ *optimal.*

Der folgende Beweis ist eine Modifikation des Beweises von Satz 2 in [2], von wo auch die Notation übernommen wird.

BEWEIS. Es sei (y_n) eine Anordnung mit $y_0=0$ und oberer asymptotischer Dichte $\leqq\frac{1}{d}$, sowie $\varepsilon>0$.

Wir werden zeigen:

$$\inf_{x\in\mathbf{R}}\sum_{i\in\mathbf{Z}}f(|x-y_i|)\leqq m(d)+5\varepsilon.$$

Wir wählen $\min\left\{1,\frac{d}{4}\right\}>\delta>0$, $h_1>0$ und $n_0\in\mathbf{N}$ so, daß gilt:

$$m(d-\delta-d\cdot\delta)\leqq m(d)+\varepsilon \tag{1}$$

$$R\left(x,\frac{d}{2}-\delta\right)\leqq\varepsilon\quad\text{für}\quad x\leqq -h_1 \tag{2}$$

$$\frac{|y_n|}{|n|+1}\geqq\frac{d}{2}-\delta\quad\text{für}\quad n\geqq n_0. \tag{3}$$

Für $T>0$ sei $l(T)=-\min\{i|y_i\in[-T,T]\}$ und $r(T)=\max\{i|y_i\in[-T,T]\}$. Nun wird T so groß gewählt, daß gilt:

$$l(T),r(T)\geqq n_0 \tag{4}$$

$$\frac{N(T)}{2T}=\frac{l(T)+r(T)+1}{2T}\leqq\frac{1}{d-\delta} \tag{5}$$

$$\frac{h_1}{T}<\delta \tag{6}$$

$$\frac{1}{T}\inf f^{-1}\left\{\frac{\varepsilon\cdot\left(\frac{d}{2}-d\right)}{T}\right\}<\delta. \tag{7}$$

Das letzte ist möglich, da $\lim_{x\to\infty}x\cdot f(x)=0$ ist, was aus der Monotonie von f und $\int_0^\infty f(x)\,dx<\infty$ folgt.

Wir setzen

$$h=\max\left\{h_1,\inf f^{-1}\left\{\frac{\varepsilon\left(\frac{d}{2}-\delta\right)}{T}\right\}\right\} \tag{8}$$

und betrachten $I=[-T+h,T-h]$.

Für die y_i, die außerhalb von $[-T, T]$ sind, gilt wegen (3) und (4): $|y_i| \geqq$ $\geqq |i| \cdot \left|\left(\frac{d}{2}-\delta\right)\right|$. Der von diesen y_i herrührende Anteil der Beleuchtung im Punkt $x \in I$ ist höchstens gleich

$$\frac{T}{\frac{d}{2}-\delta}(f(T+x)+f(T-x))+R\left(x-T, \frac{d}{2}-\delta\right)+R\left(-x-T, \frac{d}{2}-\delta\right).$$

Die Bedingungen (2) und (8) implizieren, daß diese Zahl kleiner als $4 \cdot \varepsilon$ ist.

Der mittlere Abstand d' bei jener äquidistanten Anordnung (x'_n), für die gerade die Punkte $x'_{-l(T)}, x'_{-l(T)+1}, \ldots, x'_{r(T)}$ in I liegen, ist

$$d' = \frac{2T\left(1-\frac{h}{T}\right)}{N(T)} \geqq (d-\delta)(1-\delta) \geqq d-\delta-d\delta.$$

Hier wurde (5), (6) und (7) verwendet. Das Minimum $m(d')$ zwischen aufeinanderfolgenden x'_n ist nach (1) nicht größer als $m(d)+\varepsilon$. Ersetzt man die x'_n mit $n>r(T)$ und $n<-l(T)$ durch die entsprechenden y_n, so ist das Minimum zwischen aufeinanderfolgenden x'_n in I nicht größer als $m(d)+5\varepsilon$. Werden nun die x'_n mit $-l(T) \leqq n \leqq r(T)$ durch die entsprechenden y_n ersetzt, so ist nach Lemma 2 das Minimum in I noch immer nicht größer als $m(d)+5\varepsilon$, womit der Satz bewiesen ist. Q.e.d.

FOLGERUNG. *Es sei f eine Beleuchtungsfunktion, die für $x \geqq \alpha$ konvex ist. Weiter sei $d>2\alpha$ und bei der Gitterverteilung (x_n), $x_n=n \cdot d$, werde das Minimum der Helligkeit im Punkt $\frac{d}{2}$ angenommen. Dann ist diese Verteilung optimal.*

BEWEIS. Es sei β so gewählt, daß $\alpha<\beta<\frac{d}{2}$ gilt. Die rechtsseitige Ableitung von f in β ist endlich, wir können daher f im Intervall $[0, \beta]$ so verändern, daß die neue Funktion $\bar{f}$, die mit f in $[\beta, \infty]$ übereinstimmt, eine konvexe Beleuchtungsfunktion ist und $\bar{f} \geqq f$ gilt. Es gilt

$$\inf_{x \in \mathbf{R}} \sum_{i \in \mathbf{Z}} \bar{f}(|x-x_i|) = \sum_{i \in \mathbf{Z}} \bar{f}\left(\left|\frac{d}{2}-x_i\right|\right) = \sum_{i \in \mathbf{Z}} f\left(\left|\frac{d}{2}-x_i\right|\right) = \inf_{x \in \mathbf{R}} \sum_{i \in \mathbf{Z}} f(|x-x_i|).$$

Da die Anordnung (x_n) für $\bar{f}$ optimal ist, ist sie es auch für f. Q.e.d.

3. Wir betrachten nun die „natürliche" Beleuchtungsfunktion $f(x)=\frac{1}{(h^2+x^2)^{\frac{\varepsilon}{2}}}$

Diese ist für $x \geqq \frac{h}{2}$ konvex.

SATZ 2. *Ist $d \geqq h$, so ist die gitterförmige Anordnung $(n \cdot d)_{n \in \mathbf{Z}}$ für die Beleuchtungsfunktion $f(x)=\frac{1}{(h^2+x^2)^{\frac{3}{2}}}$ optimal.*

BEWEIS. Ohne Beschränkung der Allgemeinheit nehmen wir $h=1$ an. Der Beweis der obigen Folgerung zeigt: Da man f im Intervall $\left[0, \frac{h}{2}\right]$ so vergrößern kann, daß insgesamt eine konvexe Funktion entsteht, genügt es zu zeigen: Für $d \geqq 1$ gilt:

$$S(x, d) = \sum_{i \in \mathbb{Z}} f(x - id) \geqq \sum_{i \in \mathbb{Z}} f\left(\frac{d}{2} - id\right) = S\left(\frac{d}{2}, d\right). \tag{11}$$

(Es ist ja $f(|x|)=f(x)$). Wegen Periodizität und Symmetrie $\left(\text{bezüglich } \frac{1}{2}\right)$ genügt es $0 \leqq x \leqq \frac{d}{2}$ zu betrachten und weil ferner S für festes d bezüglich x in $\left[\frac{1}{2}, d-\frac{1}{2}\right]$ konvex ist, reicht es hin, (11) für $0 \leqq x \leqq \frac{1}{2}$ zu zeigen. Für das Folgende ist es notwendig, das Monotonieverhalten der Funktionen f, f', f'' und f''' genau zu kennen.

Behauptung 1. Falls $|x-d| \geqq 0{,}8$ ist, so ist $\frac{\partial S}{\partial x}(x, d)$ für festes x eine in d monoton fallende Funktion. Dies ergibt sich, indem man $\frac{\partial S}{\partial x}(x, d)$ nach d differenziert und das Monotonieverhalten von f'' beachtet.

Behauptung 2. Für $0{,}2 \leqq x \leqq 0{,}5$ und $d \geqq 1$ gilt:

$$\frac{\partial S}{\partial x}(x, d) \leqq 0.$$

Wegen Behauptung 1 können wir $1 \leqq d \leqq 1{,}3$ voraussetzen. Es gilt

$$\begin{aligned} \frac{\partial S}{\partial x}(x, d) &= f'(x) - f'(d-x) + f'(d+x) - f'(2d-x) + \\ &+ \sum_{i=2}^{\infty} \left[f'(id+x) - f'\left((i+1) \cdot d - x\right)\right] \leqq \\ &\leqq f'(x) - f'(d-x) + f'(d+x) - f'(2d-x) =: h(x, d). \end{aligned}$$

Es genügt also

$$h(x, d) \leqq 0 \tag{12}$$

zu zeigen. Man überzeugt sich davon, daß

$$\frac{\partial h}{\partial d}(x, d) \leqq 0 \tag{13}$$

gilt, denn für $d - x \geqq \frac{2}{3}$ ist $\frac{\partial h}{\partial x}(x, d) \leqq f''(d-x) + f''(d+x) \leqq 0$ und im Fall $\frac{1}{2} \leqq d - x \leqq \frac{2}{3}$ zeigt man, daß $\frac{\partial^2 h}{\partial d^2}(x, d) \leqq -f'''\left(\frac{2}{3}\right) - 3f'''\left(\frac{3}{2}\right) \leqq 0$ gilt und muß dann nur noch $\frac{\partial h}{\partial x}(x, 1) \leqq f''\left(\frac{4}{3}\right) - 2f''\left(\frac{5}{3}\right) \leqq 0$ zeigen. Die Ungleichung (13) erlaubt jetzt die Be-

schränkung auf $d=1$ beim Beweis von (12). Mit $t(x)=h(x, 1)$ gilt für $0{,}3\leqq x\leqq 0{,}5$, $t'(x)\geqq f''(0{,}3)+f''(0{,}7)+2f''(1{,}5)\geqq 0$ und damit $t(x)\leqq t(0{,}5)=0$. Im Bereich $0{,}2\leqq x\leqq$ $\leqq 0{,}3$ gilt

$$t'(x)\leqq \max\left\{f''\left(\frac{\sqrt{3}}{2}\right)+f''(1,2)+f''(1{,}8),\ |f''(0,2)|-2f''\left(\frac{3}{2}\right)\right\}<1{,}8$$

und da $t(0{,}25)<-0{,}1$ gilt, folgt daraus $t(x)\leqq 0$ für $0{,}2\leqq x\leqq 0{,}3$. Damit ist die Behauptung 2 bewiesen.

Offenbar muß noch folgendes gezeigt werden:

Behauptung 3. Für $0\leqq x\leqq 0{,}2$ gilt

$$S(x, d)\geqq S\left(\frac{1}{2}, d\right)$$

für jedes $d\geqq 1$.

Da für $0\leqq x\leqq 0{,}2$ sicher $d-x\geqq 0{,}8$ gilt, gestattet Behauptung 1 die Annahme $d=1$ beim Beweis der obigen Behauptung.

Es gilt:

$$S(x, 1)-S\left(\frac{1}{2}, 1\right)=f(x)+f(1-x)+f(1+x)+f(2-x)+2\left(f\left(\frac{1}{2}\right)+f\left(\frac{3}{2}\right)\right)+$$

$$+\sum_{i=2}^{\infty}\left[f(i+x)-f\left(i+\frac{1}{2}\right)-\left(f\left((i+1)-\frac{1}{2}\right)-f(i+1-x)\right)\right]\geqq$$

$$\geqq f(x)+f(1-x)+f(1+x)+f(2-x)-2\left(f\left(\frac{1}{2}\right)+f\left(\frac{3}{2}\right)\right)=:r(x).$$

Es genügt also, zu zeigen, daß für $0\leqq x\leqq 0{,}2$ $r(x)\geqq 0$ gilt. Dies kann man tun, indem man die Intervalle $[0, 0.1]$, $[0.1, 0.15]$, $[0.15, 0.2]$ getrennt betrachtet und $r(x)\geqq$ $\geqq f(b)+f(1-a)+f(1+a)+f(2-a)-2\left(f\left(\frac{1}{2}\right)+f\left(\frac{3}{2}\right)\right)\geqq 0$ für $x\in[a, b]$ verwendet, wobei $[a, b]$ die angeführten Intervalle durchläuft.

LITERATUR

[1] Fejes-Tóth, L.: A problem of illumination, *Amer. Math. Monthly* **77** (1970), 868.
[2] Fischer, R.: Über die optimale Beleuchtung einer geraden Straße, *Monatsh. Math.* **79** (1975), 191—199.
[3] Heppes, A. Egy dimenziós probléma, *Mat. Lapok* **14** (1963), 124—127.
[4] Guy, R. u. Klee, V.: Monthly Research Problems 1969—1971, *Amer. Math. Monthly* **78** (1971), 1113—1122.
[5] Henry, B. R.: Solution of Fejes-Tóth's illumination problem, *Amer. Math. Monthly* **80** (1973).

Universität für Bildungswissenschaften Hochschulstrasse 67 A — 9010 Klagenfurt

(Eigegangen: 21. August 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 243—246.

ON THE L_4-NORM OF ORTHONORMAL LAGUERRE POLYNOMIALS

by

G. NÉMETH

1. Introduction. In this paper we shall deal with problem of monotonicity of L_4-norm of Laguerre polynomials. A similar question for Hermite polynomials was raised by Z. CIESIELSKI in [1]. Professor G. FREUD and the author have proved the total monotonicity of the L_4-norm of Hermite polynomials in [2]. Here we shall prove the L_4-norm of orthonormal Laguerre polynomials is a totally monotone decreasing sequence.

The author is indebted to Professor G. FREUD for several helpful discussions on this problem.

2. Monotonicity of the L_4-norm. Let us denote by $f_n(x)$ the orthonormal Laguerre functions:

$$f_n(x) = C_n \cdot x^{\alpha/2} \cdot e^{-x/2} \cdot L_n^{(\alpha)}(x), \tag{1}$$

where

$$L_n^{(\alpha)}(x) = \frac{1}{n!} x^{-\alpha} e^x \frac{d^n}{dx^n}(x^{n+\alpha} e^{-x}), \tag{2}$$

$$C_n = \sqrt{\frac{n!}{\Gamma(n+\alpha+1)}}, \quad \alpha > 1/2. \tag{3}$$

In this notation the fourth power of the L_4-norm is the integral:

$$\varrho_n^{(\alpha)} = \int_0^\infty [f_n(x)]^4 \, dx. \tag{4}$$

We prove the

THEOREM. *The sequence $\varrho_n^{(\alpha)}$ is totally monotone decreasing.*

The proof is by direct construction of a nonnegative function $s(t)$ for which

$$\varrho_n^{(\alpha)} = \int_0^1 t^n s(t) \, dt, \quad n = 0, 1, 2, \ldots \tag{5}$$

and then the total monotonicity is the consequence of Hausdorff's theorem [3].

We prove three lemmas.

LEMMA 1. Let $\alpha > -1/2$, then

$$\varrho_n^{(\alpha)} = \frac{1}{2\sqrt{\pi}} \frac{\Gamma(\alpha+1/2)}{\Gamma(\alpha+1)} \sum_{j=0}^{n} \frac{(1/2)_{n-j}^2}{(n-j)!^2} \frac{(1/2)_j}{j!} \frac{(\alpha+1/2)_j}{(\alpha+1)_j} \tag{6}$$

where $(a)_j$ is the Pochhammer's symbol:

$$(a)_0 = 1, \quad (a)_j = a(a+1)\dots(a+j-1), \quad j \geqq 1.$$

PROOF. We shall calculate a special generating function:

$$G(u_1, u_2) = \sum_{k=0}^{\infty} u_1^k \sum_{l=0}^{\infty} u_2^l \varrho_{k,l}^{(\alpha)}, \quad \varrho_{k,l}^{(\alpha)} = \int_0^{\infty} f_k^2(x) f_l^2(x)\, dx.$$

It is evident $\varrho_{n,n}^{(\alpha)} = \varrho_n^{(\alpha)}$.

Applying the generating function of HILLE—Hardy [4], we get:

$$G(u_1, u_2) = (1-u_1)^{-1-\alpha}(1-u_2)^{-1-\alpha} \frac{1}{\pi\Gamma^2(\alpha+1/2)} \int_{-1}^{+1} (1-t_1^2)^{\alpha-1/2} \int_{-1}^{+1} (1-t_2)^{\alpha-1/2} \times$$

$$\times \int_0^{\infty} x^{2\alpha} \exp\left\{-x\left[\frac{1+u_1}{1-u_1} + \frac{1+u_2}{1-u_2} + \frac{2u_1^{1/2}t_1}{1-u_1} + \frac{2u_2^{1/2}t_2}{1-u_2}\right]\right\} dx\, dt_1\, dt_2.$$

We can integrate by x and then by t_2. By elementary calculation we get the following result

$$G(u_1, u_2) = \frac{2^{2\alpha-1}}{\pi} (1-u_1)^{-1/2}(1-u_2)^{-1/2} \int_0^1 t_2^{\alpha-1/2}(1-t_1)^{\alpha-1/2} \times$$

$$\times \{(1-u_1^{1/2}u_2^{1/2}) + 4u_1^{1/2}u_2^{1/2}t_1\}^{-\alpha-1/2} dt_1.$$

This integral is a hypergeometric function of $u_1 \cdot u_2$:

$$G(u_1, u_2) = \frac{2^{2\alpha-1}}{\pi} \frac{\Gamma^2(\alpha+1/2)}{\Gamma(2\alpha+1)} (1-u_1)^{-1/2}(1-u_2)^{-1/2} {}_2F_1\left(\frac{1}{2}, \alpha+\frac{1}{2}; \alpha+1; u_1 \cdot u_2\right).$$

Expanding in power series of u_1 and u_2 we arrive at formula (6).

LEMMA 2.

$$\varrho_n^{(\alpha)} = \frac{1}{2\pi^2} \sum_{j=0}^{\infty} \frac{(1/2)_j^2}{j!^2} \cdot \frac{\Gamma(n+j+\alpha+1)\Gamma(1/2)}{\Gamma(n+j+\alpha+3/2)} \cdot \frac{\Gamma(n+j+1)\Gamma(1/2)}{\Gamma(n+j+3/2)} + \tag{7}$$

$$+ \frac{1}{2\pi^{5/2}} \cdot \frac{\Gamma(\alpha+1/2)}{\Gamma(\alpha+1)} \cdot \sum_{j=0}^{\infty} \frac{(1/2)_j}{j!} \frac{(\alpha+1/2)_j}{(\alpha+1)_j} \frac{\Gamma^2(n+j+\alpha+1)\Gamma^2(1/2)}{\Gamma^2(n+j+\alpha+3/2)}.$$

PROOF. From (6) with aid of Euler's beta integral we get:

$$\varrho_n^{(\alpha)} = \frac{1}{\pi^{3/2}} \frac{(1/2)_n}{n!} \frac{\Gamma(\alpha+1/2)}{\Gamma(\alpha+1)} \times$$

$$\int_0^1 s^{-1/2}(1-s)^{-1/2} \times \int_0^1 \omega^{n+\alpha-1/2}(1-\omega)^{-1/2} {}_2F_1\left(-n, \frac{1}{2}; \frac{1}{2}-n; S/\omega\right) d\omega\, ds.$$

In this integral we can expand the binomial terms in power series

$$\varrho_n^{(\alpha)} = \frac{1}{\pi^{3/2}} \frac{\Gamma(\alpha+1/2)}{\Gamma(\alpha+1)} \sum_{j=0}^{\infty} \frac{(1/2)_j}{j!} \sum_{l=0}^{\infty} \frac{(1/2)_l}{l!} \int_0^1 s^{j-1/2} \int_0^1 \omega^{n+l+\alpha-1/2} f(s/\omega)\, d\omega\, ds,$$

where

$$f(s) = \frac{(-1)^n}{(1/2)_n} s^{n+1/2} \frac{d^n}{ds^n} [s^{-1/2}(1-s)^n] = {}_2F_1\left(-n, \frac{1}{2}; \frac{1}{2}-n; s\right).$$

From the above double integral follows by partial integration

$$\int_0^1 \int_0^1 s^{j-1/2} \omega^{n+l+\alpha-1/2} f(s/\omega)\, d\omega\, ds = \frac{1}{n+j+l+\alpha+1} \times$$

$$\times \left[\int_0^1 s^{j-1/2} f(s)\, ds + \int_0^1 \omega^{n+l+\alpha-1/2} f\left(\frac{1}{\omega}\right) d\omega \right].$$

Finally, taking into account

$$\int_0^1 s^{j-1/2} f(s)\, ds = \frac{(j+1)_n}{(1/2)_n} \cdot \frac{\Gamma(j+1/2)\Gamma(1/2)}{\Gamma(n+j+3/2)},$$

$$\int_0^1 \omega^{n+l+\alpha-1/2} f\left(\frac{1}{\omega}\right) d\omega = \frac{(l+\alpha+1)_n}{(1/2)_n} \frac{\Gamma(l+\alpha+1/2)\Gamma(n+1)}{\Gamma(n+l+\alpha+3/2)},$$

$$\sum_{j=0}^{\infty} \frac{(1/2)_j}{j!} \frac{1}{n+l+j+\alpha+1} = \frac{\Gamma(n+l+\alpha+1)\Gamma(1/2)}{\Gamma(n+l+\alpha+3/2)},$$

we arrive at formula (7) by elementary calculations. From (7) we get easily

$$\varrho_n^{(\alpha)} = \frac{1}{2\pi^2} \int_0^1 \int_0^1 t_1^{n+\alpha}(1-t_1)^{-1/2} t_2^n (1-t_2)^{-1/2} {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1; t_1 \cdot t_2\right) dt_1 dt_2 +$$

$$+ \frac{1}{2\pi^{5/2}} \frac{\Gamma(\alpha+1/2)}{\Gamma(\alpha+1)} \int_0^1 \int_0^1 t_1^{n+\alpha}(1-t_1)^{-1/2} t_2^{n+\alpha}(1-t_2)^{-1/2} \times$$

$$\times {}_2F_1\left(\frac{1}{2}, \alpha+\frac{1}{2}; \alpha+1; t_1 \cdot t_2\right) dt_1\, dt_2.$$

LEMMA 3. For the function $g(s) = s^\beta {}_2F_1\left(\frac{1}{2}, \alpha+\frac{1}{2}; \alpha+1; s\right)$ if $\beta > -\frac{1}{2}, \alpha > -\frac{1}{2}$

$$\int_0^1 \int_0^1 t_1^\alpha (1-t_1)^{-1/2}(1-t_2)^{-1/2} g(t_1 \cdot t_2)\, dt_1\, dt_2 = \tag{8}$$

$$= \pi \int_0^1 t_1^\alpha g(t_1)\, {}_2F_1\left(\frac{1}{2}, \alpha+\frac{1}{2}; 1; 1-t_1\right) dt_1.$$

PROOF. Expanding the binomial terms in power series we can integrate by t_2 in the double integral. Proceeding than the summation we can arrive at formula (8).

Finally, when we apply the formula (8) to the above integrals, the proof of THEOREM is complete.

We get

$$\varrho_n^{(\alpha)} = \int_0^1 t^n s(t)\,dt,$$

$$(9) \qquad s(t) = \frac{t^\alpha}{2\pi}\left\{{}_2F_1\left(\frac{1}{2},\frac{1}{2};1;t\right){}_2F_1\left(\frac{1}{2},\alpha+\frac{1}{2};1-t\right)+\right.$$

$$\left.+\frac{\Gamma(\alpha+1/2)}{\Gamma(1/2)\Gamma(\alpha+1)}\,{}_2F_1\left(\frac{1}{2},\frac{1}{2};1;1-t\right){}_2F_1\left(\frac{1}{2},\alpha+\frac{1}{2};\alpha+1;t\right)\right\}.$$

In (9) the hypergeometric functions are series in positive terms therefore the function $s(t)$ is positive. Q.E.D.

3. An asymptotic formula. It is not difficult to derive an approximate formula for $\varrho_n^{(\alpha)}$ when $n\to\infty$.

It is well known from theory of hypergeometric functions

$${}_2F_1\left(\frac{1}{2},\frac{1}{2};1;t\right) \sim \frac{1}{\pi}\ln\frac{1}{1-t}, \quad t\to 1-0,$$

$${}_2F_1\left(\frac{1}{2},\alpha+\frac{1}{2};\alpha+1;t\right) \sim \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+1/2)\Gamma(1/2)}\ln\frac{1}{1-t}, \quad t\to 1-0,$$

and thus

$$\varrho_n^{(\alpha)} \sim \frac{1}{\pi^2}\int_0^1 t^n \ln\frac{1}{1-t}\,dt, \quad n\to\infty.$$

It follows by elementary calculations

$$\varrho_n^{(\alpha)} \sim \frac{1}{\pi^2}\frac{\ln n}{n}, \quad n\to\infty.$$

REFERENCES

[1] CIESIELSKI, Z.: Conference. Theory of Approximation, Problems. Poznan, 22—26. August 1972. p. 5.

[2] FREUD, G., NÉMETH, G.: On the L_p-norms of orthogonal Hermite functions, *Studia Sci. Math. Hung.* **8** (1973), 399—404.

[3] HAUSDORFF, F.: Momentprobleme für ein endliches Intervall, *Math. Zeitschr.* **16** (1923), 220—248.

[4] SZEGŐ, G.: *Orthogonal Polynomials.* New York. Amer. Math. Soc. Colloquium. Publ. 1939, Vol. 23.

Központi Fizikai Kutató Intézet, 1525 Budapest, Pf. 49

(Received April 20, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 247—253.

A PORTEMANTEAU THEOREM FOR VAGUE CONVERGENCE*

by

B. B. WINTER

Let Ω be a locally compact second countable space; then (see, e.g., Theorem 7.6.1 in [1]) Ω is a *locally compact Polish* space; i.e., Ω is locally compact and separable, and its topology can be metrized by a complete metric d. $\mathscr{B}$ is the σ-algebra of *Borel sets* of Ω, i.e. the smallest σ-algebra containing all the open sets; $\mathscr{M}$ is the set of all *Borel measures* on $\mathscr{B}$, i.e. all measures μ such that $\mu(K) < \infty$ for each compact K, and $\mathscr{M}_r = \{\mu \in \mathscr{M} : \mu(\Omega) < \infty\}$; $\mathscr{C}_c$ is the set of all continuous functions $\Omega \to \mathbf{R}$ having compact support, $\mathscr{C}_0$ is the set of all continuous functions $\Omega \to \mathbf{R}$ *vanishing at infinity* (i.e., for each $\varepsilon > 0$, $\{x : |f(x)| \geqq \varepsilon\}$ is compact), $\mathscr{C}_B$ is the set of all bounded continuous functions $\Omega \to \mathbf{R}$. If $(\mu_n)_0^\infty$ is a sequence in $\mathscr{M}$ then it converges *vaguely* to μ_0 $(\mu_n \xrightarrow{v} \mu_0)$ iff $\int f d\mu_n \to \int f d\mu_0$ for each $f \in \mathscr{C}_c$; if $(\mu_n)_0^\infty$ is a sequence in $\mathscr{M}_r$ then it converges *weakly* to μ_0 $(\mu_n \xrightarrow{w} \mu_0)$ iff $\int f d\mu_n \to \int f d\mu_0$ for each $f \in \mathscr{C}_B$. If $\mu_n \xrightarrow{v} \mu_0$ and $\mu_n \xrightarrow{v} \mu_0'$ then $\mu_0 = \mu_0'$, and likewise in the case of weak convergence. The support of f will be denoted $\operatorname{spt}(f)$; I_A is the indicator (or characteristic function) of the set A; the function $\varphi : \mathbf{R} \to \mathbf{R}$ is defined by

$$\varphi(x) = \begin{cases} 1 & \text{if } x \leqq 0 \\ 1 - x & \text{if } 0 \leqq x \leqq 1. \\ 0 & \text{if } x \geqq 1 \end{cases}$$

The well-known Portemanteau Theorem (Theorem 2.1 in [2]) states that if $(\mu_n)_0^\infty$ is a sequence of *probability measures* then $\mu_n \xrightarrow{w} \mu_0 \Leftrightarrow \overline{\lim}\, \mu_n(C) \leqq \mu_0(C)$ for each closed set $C \Leftrightarrow \underline{\lim}\, \mu_n(U) \geqq \mu_0(U)$ for each open set $U \Leftrightarrow \mu_n(A) \to \mu_0(A)$ whenever $\mu_0(\partial A) = 0$. It will be shown that something very similar to this is valid for vague convergence of sequences in $\mathscr{M}$.

(1) LEMMA. *For any $A \subset \Omega$ and $\varrho > 0$, let $A_\varrho = \{x \in \Omega : d(x, A) \leqq \varrho\}$. If $\mu_0 \in \mathscr{M}$ and K is a compact subset of Ω then $\exists r > 0$ such that the set*

$$\{\varrho : 0 < \varrho < r \;\&\; K_\varrho \text{ is compact} \;\&\; \mu_0(\partial K_\varrho) = 0\}$$

is dense in $(0, r)$.

PROOF. Consider $\mu_0 \in \mathscr{M}$ and K compact. It is easily shown that, due to local compactness, $\exists r > 0$ such that K_r is compact. Put $h(x) = d(x, K)$; then $\partial K_\varrho \subset \subset \{x \in \Omega : d(x, K) = \varrho\} = h^{-1}\{\varrho\}$. Put

$$A_n = \{\varrho : 0 < \varrho < r \;\&\; \mu_0(\partial K_\varrho) \geqq 1/n\}.$$

* Research partly supported by the National Research Council of Canada.

If A_n has distinct elements $\varrho_1, \varrho_2, \ldots$ then, since $\partial K_{\varrho_j} \subset h^{-1}\{\varrho_j\} \subset K_r$,

$$\infty > \mu_0(K_r) \geqq \mu_0(h^{-1}\{\varrho_1\} \cup h^{-1}\{\varrho_2\} \cup \ldots) = \mu_0 h^{-1}\{\varrho_1\} + \mu_0 h^{-1}\{\varrho_2\} + \ldots$$
$$\geqq 1/n + 1/n + \ldots;$$

this shows that A_n must be finite. Therefore $\bigcup_1^\infty A_n = \{\varrho : 0 < \varrho < r \ \& \ \mu_0(\partial K_\varrho) > 0\}$ is countable, hence $\{\varrho : 0 < \varrho < r \ \& \ \mu_0(\partial K_\varrho) = 0\}$ is dense in $(0, r)$. ∎

(2) LEMMA. *Suppose that $\mu_0 \in \mathscr{M}$. If $f \in \mathscr{C}_c$ then $\exists D \subset \mathbf{R}$ such that D is dense and, for each $y \in D$,*

(a) $\mu_0(f^{-1}\{y\}) = 0$ *and*

(b) $\mu_0(\partial\{x : f(x) \leqq y\}) = 0 = \mu_0(\partial\{x : f(x) > y\})$.

PROOF. Note that $y \neq 0 \Rightarrow f^{-1}\{y\} \subset \operatorname{spt}(f)$. Put

$$A_n = \left\{ y \in \mathbf{R} : y \neq 0 \quad \& \quad \mu_0(f^{-1}\{y\}) \geqq \frac{1}{n} \right\}.$$

If $y \in A_n$ then $1/n \leqq \mu_0(f^{-1}\{y\}) \leqq \mu_0(\operatorname{spt}(f)) < \infty$, hence A_n must be finite, and therefore $\{y \in \mathbf{R} : \mu_0(f^{-1}\{y\}) > 0\}$ is countable. Put $D = \mathbf{R} \setminus \bigcup_1^\infty A_n$; then D is dense in $\mathbf{R}$ and (a) is true for each $y \in D$.

If $A = f^{-1}(-\infty, y]$ then $A^c \subset f^{-1}[y, \infty)$. By continuity of f, both of these inverse images are closed; therefore

$$\partial A = \bar{A} \cap \bar{A^c} \subset f^{-1}(-\infty, y] \cap f^{-1}[y, \infty) = f^{-1}\{y\}.$$

Thus $y \in D \Rightarrow \mu_0 \partial f^{-1}(-\infty, y] = 0$ and, similarly,

$$y \in D \Rightarrow \mu_0 \partial f^{-1}(y, \infty) = 0.$$ ∎

(3) THEOREM. *If $(\mu_n)_0^\infty$ is a sequence in $\mathscr{M}$ then the following conditions are equivalent.*

(i) $\mu_n \xrightarrow{v} \mu_0$

(ii) (a) $\overline{\lim}\, \mu_n(K) \leqq \mu_0(K)$ *for any compact set* K

(b) $\underline{\lim}\, \mu_n(U) \geqq \mu_0(U)$ *for any open set* U

(ii)′ $\overline{\lim}\, \mu_n(K) \leqq \mu_0(K)$ *for any compact* K

$\underline{\lim}\, \mu_n(U) \geqq \mu_0(U)$ *for any relatively compact open* U

(iii) $\mu_n(A) \to \mu_0(A)$ *for any relatively compact* A *such that* $\mu_0(\partial A) = 0$.

PROOF. Suppose that (i) is true. Consider a compact K and put $f_k(x) = \varphi(k \cdot d(x, K))$, $k \in \mathbf{N}^+$. The support of f_k is the closure of $\{x : kd(x, K) < 1\}$, so $\operatorname{spt}(f_k) \subset \{x : d(x, K) \leqq 1/k\}$. For sufficiently large k, say $k \geqq k_0$, $\{x : d(x, K) \leqq 1/k\}$ is compact and therefore f_k is in $\mathscr{C}_c$. Clearly $k \geqq k_0 \Rightarrow$

$$I_K \leqq f_k \leqq f_{k_0} \leqq I_{\{x : d(x, K) \leqq 1/k_0\}};$$

therefore $\mu_n(K) \leqq \int f_k \, d\mu_n$ and

$$\overline{\lim}\, \mu_n(K) \leqq \overline{\lim_{n\to\infty}} \int f_k \, d\mu_n = \int f_k \, d\mu_0.$$

Since K is closed, $f_k \searrow I_K$; and $\int f_{k_0} d\mu_0 \leqq \mu_0\{x : d(x, K) \leqq 1/k_0\} < \infty$ because μ_0 is a Borel measure. By the Dominated Convergence Theorem, $\int f_k \, d\mu_0 \to \mu_0(K)$ which shows that (ii.a) is satisfied.

Now, still supposing that (i) is true, consider a relatively compact open set U. Let $C = \Omega \setminus U$ and put $g_k = 1 - \varphi(kd(x, C))$. Then $g_k \nearrow 1 - I_C = I_U$. Since $\overline{U}$ is compact, $\operatorname{spt}(g_k) \subset \operatorname{spt}(I_U) = \overline{U}$ shows that $g_k \in \mathscr{C}_c$. Therefore

$$\underline{\lim}\, \mu_n(U) \geqq \underline{\lim_{n\to\infty}} \int g_k \, d\mu_n = \int g_k \, d\mu_0.$$

By the Monotone Convergence Theorem, $\int g_k \, d\mu_0 \to \mu_0(U)$. Thus (ii.b) is satisfied when U is relatively compact and open. — Now let U be an arbitrary open set. Note that (as is easily shown) if a locally compact space has a countable base $\mathscr{U}$ then $\exists \mathscr{U}^* \subset \mathscr{U}$, such that $\mathscr{U}^*$ is a base consisting entirely of relatively compact sets. Therefore U can be written as $\bigcup_1^\infty U_j$, with each U_j open and relatively compact. Then, for any positive integer m, $\bigcup_1^m U_j$ is open and relatively compact; by what was just shown,

$$\mu_0\left(\bigcup_1^m U_j\right) \leqq \underline{\lim_{n\to\infty}}\, \mu_n\left(\bigcup_1^m U_j\right) \leqq \underline{\lim}\, \mu_n(U).$$

Letting $m \nearrow \infty$ shows that $\mu_0(U) \leqq \underline{\lim}\, \mu_n(U)$.

Obviously (ii)$\Rightarrow$(ii)′. Now suppose that (ii)′ is true and consider a relatively compact A such that $\mu_0(\partial A) = 0$. Since $\overline{A} = A^0 \cup \partial A$ and A^0 is relatively compact,

$$\mu_0(A) \leqq \mu_0(\overline{A}) = \mu_0(A^0) \leqq \underline{\lim}\, \mu_n(A^0) = \overline{\lim}\, \mu_n(\overline{A}) \leqq$$
$$\leqq \mu_0(\overline{A}) = \mu_0(A^0) \leqq \mu_0(A),$$

which shows that (iii) is satisfied.

Suppose that (iii) is true and consider $f \in \mathscr{C}_c$ and an arbitrary positive ε. Let $K = \operatorname{spt}(f)$; in view of (1), one can find $\varrho > 0$ such that $\operatorname{spt}(f) \subset K_\varrho$ and K_ϱ is compact and $\mu_0(\partial K_\varrho) = 0$. Let D be as in (2). Since D is dense, one can find $y_0 < y_1 < \ldots < y_k$ in D such that $y_0 < f(x) \leqq y_k$ for each $x \in \Omega$ and $y_j - y_{j-1} < \varepsilon$ for $j = 1, 2, \ldots, k$. Let $A_j = \{x : y_{j-1} < f(x) \leqq y_j\}$, $1 \leqq j \leqq k$, and

$$f_\varepsilon(x) = \left(\sum_{j=1}^k y_j I_{A_j}(x)\right) \cdot I_{K_\varrho}(x).$$

Note that $|f(x) - f_\varepsilon(x)| < \varepsilon$ for each $x \in \Omega$. Approximating f by f_ε,

$$\overline{\lim}\left|\int f \, d\mu_n - \int f \, d\mu_0\right| \leqq \overline{\lim}\left|\int f \, d\mu_n - \int f_\varepsilon \, d\mu_n\right| + \tag{4}$$
$$+ \overline{\lim}\left|\int f_\varepsilon \, d\mu_n - \int f_\varepsilon \, d\mu_0\right| + \overline{\lim}\left|\int f_\varepsilon \, d\mu_0 - \int f \, d\mu_0\right|.$$

The last term on the right of (4) is not more than

$$\int |f_\varepsilon - f|\, d\mu_0 = \int_{K_\varrho} |f_\varepsilon - f|\, d\mu_0 \leqq \varepsilon \mu_0(K_\varrho)$$

and the first term is not more than

$$\overline{\lim} \int |f_\varepsilon - f|\, d\mu_n = \overline{\lim} \int_{K_\varrho} |f_\varepsilon - f|\, d\mu_n \leqq \overline{\lim}\, \varepsilon \mu_n(K_\varrho) = \varepsilon \mu_0(K_\varrho).$$

As to the middle term, note that $\partial(A \cap B) \subset \partial A \cup \partial B$; then by (2), $\mu_0(\partial A_j) = 0$ and consequently $\mu_0 \partial(K_\varrho \cap A_j) \leqq \mu_0(\partial K_\varrho) + \mu_0(\partial A_j) = 0$. Therefore, as $n \to \infty$,

$$\int f_\varepsilon\, d\mu_n = \sum_{j=1}^{k} y_j \mu_n(K_\varrho \cap A_j) \to \sum_{j=1}^{k} y_j \mu_0(K_\varrho \cap A_j) = \int f_\varepsilon\, d\mu_0.$$

In view of (4), the above shows that $\overline{\lim} \left| \int f d\mu_n - \int f d\mu_0 \right| \leqq 2\varepsilon \mu_0(K_\varrho)$. Since ε is arbitrary and $\mu_0(K_\varrho)$ is finite, it follows that $\int f d\mu_n \to \int f d\mu_0$. ∎

In the case of weak convergence of probability measures, conditions (ii.a) and (ii.b) are equivalent because one can go from one to the other by complementation. But if the measures are not probabilities then $\mu_n(\Omega)$ can be "out of control" and these two conditions are therefore not equivalent. For instance, with $\Omega = \mathbf{R}$, let $\mu_1 = \mu_3 = \ldots$ be the uniform distribution on [0, 1] and $\mu_2 = \mu_4 = \ldots = 2\mu_1$. If $\mu_0 = \mu_1$ then (ii.b) is satisfied but (ii.a) fails with $K = [0, 1]$; if $\mu_0 = \mu_2$ then (ii.a) is satisfied but (ii.b) fails with $U = (0, 1)$.

Consider the special case $\Omega = \mathbf{R}^m$, with Borel sets $\mathscr{B}^m$. According to (3), if $\mu_n(A) \to \mu_0(A)$ for each bounded set A having $\mu_0(\partial A) = 0$ then $\mu_n \xrightarrow{v} \mu_0$. Actually, it is not necessary to verify that $\mu_n(A) \to \mu_0(A)$ for *all* bounded μ_0-null-boundary sets A; it suffices to look at sets of the form $(a, b]$. [If $x = (x_1, \ldots, x_m) \in \mathbf{R}^m$ then $(a, b]$ stands for $\{x \in \mathbf{R}^m : a_i < x_i \leqq b_i \text{ for } 1 \leqq i \leqq m\}$ and (a, b) is defined similarly; all the a_i's and b_i's are finite.] This is stated and proved for $\mathbf{R}^1$ on pp. 79—80 of [4] and pp. 180—181 of [5] but those proofs would become awkward if extended to $\mathbf{R}^m$. Instead, one can adapt the argument given on pp. 14—15 of [2]. The (adapted) argument is outlined below.

(5) THEOREM. *Let $(\mu_n)_0^\infty$ be a sequence of Borel measures on $\mathscr{B}^m$. If $\mu_n(a, b] \to \mu_0(a, b]$ whenever $\mu_0(\partial(a, b]) = 0$ then $\mu_n \xrightarrow{v} \mu_0$.*

PROOF. Let $\mathscr{U}$ be the class of all sets $(a, b] \subset \mathbf{R}^m$ which are such that $\mu_0(\partial(a, b]) = 0$ and let $S = \{x \in \mathbf{R}^m : d(\mathbf{0}, x) < 1\}$, where d is the usual metric for $\mathbf{R}^m$. For $\alpha \in \mathbf{R}$, let $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha) \in \mathbf{R}^m$. If $0 < \alpha < \beta$ then $\partial(-\boldsymbol{\alpha}, \boldsymbol{\alpha}]$ and $\partial(-\boldsymbol{\beta}, \boldsymbol{\beta}]$ are disjoint. It follows that $\{\alpha \in (0, 1) : \mu_0(\partial(-\boldsymbol{\alpha}, \boldsymbol{\alpha}]) > 0\}$ is countable, hence one can find $\alpha > 0$ such that $(-\boldsymbol{\alpha}, \boldsymbol{\alpha}] \in \mathscr{U}$ and $\mathbf{0} \in (-\boldsymbol{\alpha}, \boldsymbol{\alpha}) \subset (-\boldsymbol{\alpha}, \boldsymbol{\alpha}] \subset S$. Except for changes in notation, one shows similarly that

if x is in the open set G then one can find some $J_x \in \mathscr{U}$ such that $x \in J_x^0 \subset J_x \subset G$. Consider a relatively compact open set G and, for each $x \in G$, let J_x be as above. Then $\{J_x^0 : x \in G\}$ is an open cover of G, hence has a countable subcover; let $(J_k^0)_1^\infty$ be such a subcover, so that $G = \bigcup_1^\infty J_k^0$; since $J_k \subset G$, one has $G = \bigcup_1^\infty J_k$.

The next step is to notice that if $\mu\left(\bigcup_1^k A_i\right)<\infty$ then, as is easily proved by induction,

$$\mu\left(\bigcup_1^k A_i\right)=\sum_1^k \mu(A_i)-\sum_{1\leqq i_1<i_2\leqq k}\mu(A_{i_1}\cap A_{i_2})+$$

$$+\sum_{1\leqq i_1<i_2<i_3\leqq k}\mu(A_{i_1}\cap A_{i_2}\cap A_{i_3})-\ldots+(-1)^{k+1}\mu(A_1\cap\ldots\cap A_k).$$

If each A_i is in $\mathscr{U}$ then so is each of the above intersections, because $\partial(A\cap B)\subset$ $\subset\partial A\cup\partial B$ and the intersection of two sets of the type $(a, b]$ is again a set of that type. Since $\mu_n(G)<\infty$, it follows (as in [2]) that $\mu_0(G)\leqq\underline{\lim}\,\mu_n(G)$.

If K is a compact subset of $\mathbf{R}^m$ then one can find some $J\in\mathscr{U}$ such that $K\subset J^\circ$. Put $U=J^\circ\setminus K$; then $\mu_n(U)<\infty$ and $K=J^\circ\setminus U$, hence

$$\begin{aligned}\overline{\lim}\,\mu_n(K)&\leqq\overline{\lim}\,\mu_n(J^0)+\overline{\lim}\,(-\mu_n(U))\\&\leqq\lim\mu_n(J)-\underline{\lim}\,\mu_n(U)\\&\leqq\mu_0(J)-\mu_0(U)=\mu_0(J^0)-\mu_0(U)\\&=\mu_0(J^0\setminus U)=\mu_0(K).\end{aligned}$$

By (ii)′ in (3), $\mu_n\xrightarrow{v}\mu_0$. ∎

The next result is known but the simple proof given here may well be new. The result is stated and proved for $\mathbf{R}^1$ on pp. 79—80 of [4] and pp. 180—181 of [5] but the proofs given there would become extremely awkward if extended to $\mathbf{R}^m$. The result is proved in a more general setting (assuming only that Ω is locally compact and σ-compact) in Theorem 7.5.5 of [1] but that proof relies on a fair amount of topology and is "less transparent" than the proof given here for a locally compact Polish space.

(6) THEOREM. *If* $(\mu_n)_0^\infty$ *is a sequence in* $\mathscr{M}$, *such that* $\sup\{\mu_n(\Omega): n=1, 2, \ldots\ldots\}<\infty$, *then each of the conditions in* (3) *is equivalent to*

(iv) $$\int f\,d\mu_n\to\int f\,d\mu_0 \quad \text{for each} \quad f\in\mathscr{C}_0.$$

PROOF. Since $\mathscr{C}_c\subset\mathscr{C}_0$, (iv) implies $\mu_n\xrightarrow{v}\mu_0$. Conversely, suppose that $\mu_n\xrightarrow{v}\mu_0$ and consider $f\in\mathscr{C}_0$ and an arbitrary $\varepsilon>0$. Let $K=\{x: |f(x)|\geqq\varepsilon\}$; in view of (1), one can find $\varrho>0$ such that K_ϱ is compact and $\mu_0(\partial K_\varrho)=0$. Arguing as in the proof of (3), choose $y_0<y_1<\ldots<y_k$ in D such that $y_0<f(x)\leqq y_k$ for each each $x\in K$ and y_j- $-y_{j-1}<\varepsilon$ for $j=1, 2, \ldots, k$. Let f_ε be as in the proof of (3). Again $|f(x)-f_\varepsilon(x)|<\varepsilon$ for each $x\in\Omega$. Let $\beta=\sup\{\mu_n(\Omega)\}_1^\infty$ and again consider the inequality (4). If $n\neq0$ then

$$\overline{\lim}\int|f_\varepsilon-f|\,d\mu_n\leqq\varepsilon\,\overline{\lim}\,\mu_n(\Omega)\leqq\varepsilon\beta$$

and, by (ii.b) in (3),

$$\int|f_\varepsilon-f|\,d\mu_0\leqq\varepsilon\mu_0(\Omega)\leqq\varepsilon\,\underline{\lim}\,\mu_n(\Omega)\leqq\varepsilon\beta;$$

also, exactly as in the proof of (3), $\int f_\varepsilon\,d\mu_n\to\int f_\varepsilon\,d\mu_0$. In view of (4), since ε is arbitrary and β is finite, $\int f\,d\mu_n\to\int f\,d\mu_0$.

If one is dealing with a *bounded* family of measures then *vague* convergence on Ω can be studied by relating it to *weak* convergence on the Alexandrov one-point compactification of Ω. Put $\beta=\sup\{\mu_n(\Omega): n=0, 1, 2, \ldots\}$ and suppose that $\beta<\infty$; without loss of generality, assume that $\beta\leqq 1$. Let $\tilde{\Omega}=\Omega\cup\{*\}$ be the Alexandrov compactification of Ω, with Borel sets $\tilde{\mathscr{B}}$; it is easily seen that $\tilde{\mathscr{B}}=\mathscr{B}\cup\{B\cup\{*\}: B\in\mathscr{B}\}$. Extend μ_n to a *probability* measure $\tilde{\mu}_n$ on $\tilde{\mathscr{B}}$ by attaching the "missing mass" $1-\mu_n(\Omega)$ to the point $*$; i.e., for $A\in\tilde{\mathscr{B}}$, put

$$\tilde{\mu}_n(A)=\begin{cases}\mu_n(A) & \text{if} \quad *\notin A\\ \mu_n(A\setminus\{*\})+1-\mu_n(\Omega) & \text{if} \quad *\in A\end{cases}$$

In the sequel, Ω is again assumed to be a locally compact Polish space.

(7) LEMMA. *Let* $(\mu_n)_0^\infty$ *be a sequence on* $\mathscr{M}$, *such that* $\sup\{\mu_n(\Omega)\}_0^\infty\leqq 1$. *Then*

$$\mu_n\xrightarrow{v}\mu_0 \quad \text{iff} \quad \tilde{\mu}_n\xrightarrow{w}\tilde{\mu}_0.$$

PROOF. Suppose that $\tilde{\mu}_n\xrightarrow{w}\tilde{\mu}_0$ and consider $f\in\mathscr{C}_c(\Omega)$. Extend f to a continuous function $\tilde{f}$: $\tilde{\Omega}\to\mathbf{R}$, by putting $\tilde{f}(*)=0$ and $\tilde{f}(x)=f(x)$ if $x\in\Omega$. Then $\tilde{f}\in\mathscr{C}_B(\tilde{\Omega})$, hence

$$\int_\Omega f\,d\mu_n=\int_{\tilde{\Omega}}\tilde{f}\,d\tilde{\mu}_n\to\int_{\tilde{\Omega}}\tilde{f}\,d\tilde{\mu}_0=\int_\Omega f\,d\mu_0,$$

i.e. $\mu_n\xrightarrow{v}\mu_0$. ∎

Conversely, suppose that $\mu_n\xrightarrow{v}\mu_0$. Consider U, an open subset of Ω. By (3), $\tilde{\mu}_0(U)=\mu_0(U)\leqq\underline{\lim}\,\mu_n(U)=\underline{\lim}\,\tilde{\mu}_n(U)$. Next, consider U such that $U=\{*\}\cup(\Omega\setminus K)$, where K is a compact subset of Ω. Now $\tilde{\mu}_n(U)=1-\mu_n(\Omega)+\mu_n(\Omega\setminus K)=1-\mu_n(K)$ and therefore, by (3), $\underline{\lim}\,\tilde{\mu}_n(U)=1-\overline{\lim}\,\mu_n(K)\geqq 1-\mu_0(K)=\tilde{\mu}_0(U)$. Thus $\tilde{\mu}_0(U)\leqq\leqq\underline{\lim}\,\tilde{\mu}_n(U)$ whenever U is an open subset of $\tilde{\Omega}$; by the Portemanteau Theorem for weak convergence, $\tilde{\mu}_n\xrightarrow{w}\tilde{\mu}_0$.

For an application of (7) note that if $f\in\mathscr{C}_0(\Omega)$ and $\tilde{f}$ is defined as in the proof of (7) then $\tilde{f}\in\mathscr{C}_B(\tilde{\Omega})$; thus (6) follows immediately from (7). For another application of (7), let $(\mu_n)_1^\infty$ be a sequence in $\mathscr{M}$, such that $\sup\{\mu_n(\Omega)\}_1^\infty<\infty$. Since Ω is locally compact and second countable, so is $\tilde{\Omega}$; therefore $(\tilde{\mu}_n)_1^\infty$ has a weakly convergent subsequence, hence $(\mu_n)_1^\infty$ has a vaguely convergent subsequence. For a final application, let $\mathscr{M}_1(\Omega)$ and $\mathscr{M}_1(\tilde{\Omega})$ be the set of all probability measures on $\mathscr{B}$ and $\tilde{\mathscr{B}}$, respectively, and let $\tilde{\varrho}$ be a metric on $\mathscr{M}_1(\tilde{\Omega})$ such that $\tilde{\varrho}(\theta_n,\theta_0)\to 0$ iff $\theta_n\xrightarrow{w}\theta_0$. Such metrics exist; e.g., since $\tilde{\Omega}$ is again a Polish space, the PROKHOROV metric has this property (see [6]). The definition $\varrho(\mu_1,\mu_2)=\tilde{\varrho}(\tilde{\mu}_1,\tilde{\mu}_2)$ makes ϱ a metric on $\mathscr{M}_1(\Omega)$; in view of (7), $\mu_n\xrightarrow{v}\mu_0$ iff $\varrho(\mu_n,\mu_0)\to 0$. This gives a quick proof of the metrizability of vague convergence in $\mathscr{M}_1(\Omega)$; a more complicated proof (see Theorem 7.8.4 in [1]) establishes the metrizability of vague convergence in $\mathscr{M}(\Omega)$.

Acknowledgment. I am grateful to Peter Major for calling my attention to the result stated in (7) and its applications. He also pointed out that (3) and (6), which I originally stated for $\Omega=\mathbf{R}^m$, are valid whenever Ω is locally compact Polish space.

REFERENCES

[1] BAUER, H.: *Probability Theory and Elements of Measure Theory,* Holt, Rinehart and Winston, New York, 1972.
[2] BILLINGSLEY, P.: *Convergence of Probability Measures,* Wiley, New York, 1968.
[3] BREIMAN, L.: *Probability,* Addison-Wesley, Reading (Mass.), 1968.
[4] CHUNG, K. L.: *A Course in Probability Theory,* Harcourt, Brace & World, New York, 1968.
[5] LOEVE, M.: *Probability Theory,* 3rd ed., Van Nostrand, Princeton (N. J.), 1963.
[6] PROKHOROV, YU. V.: Convergence of random processes and limit theorems in probability theory, *Theor. Probability Appl.* **1** (1956), 157—214.

Mathematics Department, University of Ottawa

(Received May 30, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 255—271.

A THEOREM ON THE APPROXIMATION OF THE SPECTRUM OF A SELFADJOINT SYSTEM OF ORDINARY DIFFERENTIAL EQUATIONS BY THE LÁNCZOS PROCESS

by

K. MOSZYŃSKI

1. The selfadjoint eigenvalue problem

Consider the system of m linear ordinary differential equations

$$\frac{d}{dx}u(x) = [A(x)+\lambda B(x)]u(x) \tag{1}$$

with the boundary value conditions

$$Mu(a)+Nu(b) = 0, \tag{2}$$

where $u(x)$ is an m dimensional column vector, A, B, M, N are $m\times m$ real matrices and x ranges over the closed finite interval $[a, b]$ of the real axis. Assume the matrices A and B to be continuous on $[a, b]$ and the constant rectangular $m\times 2m$ matrix $[M, N]$ to be of the rank m.

Those values of the parameter λ, for which a non-zero solution of (1), (2) exists are the eigenvalues of the problem. The closure of the set of all eigenvalues is the spectrum of (1) (2). The non-zero solutions of (1) (2) are the eigenfunctions of the problem.

The problem (1) (2) is selfadjoint (see [3], [4]) if for any $x\in[a, b]$, there exists a non singular matrix $T(x)$ for which the following conditions hold.

$$\frac{d}{dx}T(x)+T(x)A(x)+A^T(x)T(x) = 0, \tag{3}$$

$$T(x)B(x)+B^T(x)T(x) = 0, \tag{4}$$

$$MT^{-1}(a)M^T-NT^{-1}(b)N^T = 0 \tag{5}$$

(6) for any $x\in[a, b]$, the matrix $S(x)=T(x)^T\cdot B(x)$ is symmetric and positive semidefinite.

We assume the spectrum of (1) (2) *not to be finite* and *not to contain zero.*

2. Some properties of the problem (1) (2)

G. A. Bliss proved in his papers [3] and [4] that:

1° All eigenvalues of the selfadjoint problem (1) (2) are real, and it is always possible to choose the corresponding eigenfunctions to be real. The order of any eigenvalue coincides with the number of real independent eigenfunctions corresponding to this eigenvalue.

2° The eigenvalues are zeros of some entire function, hence they form an at most countable set without finite accumulation points. Here we assume this set to be infinite.

3° It is possible to choose such eigenfunctions ψ_j to obtain an "orthonormal" system in the sense that:

$$\int_a^b \psi_i^T(x) S(x) \psi_j(x)\, dx = \delta_{ij}.$$

The matrix S plays the role of the weight function (though it can be singular!)

4° Let f be an integrable m dimensional vectorfuntion defined on $[a, b]$. Consider the generalized Fourier series

$$f(x) \sim \sum_{\nu=1}^{\infty} c_\nu \psi_\nu(x) \tag{7}$$

where

$$c_\nu = \int_a^b \psi_\nu^T(x) S(x) f(x)\, dx. \tag{8}$$

THEOREM 2.1. *If there exist two functions g and h such that the following equations are satisfied on* $[a, b]$:

$$\frac{d}{dx} f(x) - A(x) f(x) = B(x) g(x)$$

$$\frac{d}{dx} g(x) - A(x) g(x) = B(x) h(x)$$

and both the functions f and g satisfy the boundary conditions (2), *h being continuous on* $[a, b]$, *then*

$$f(x) = \sum_{\nu=1}^{\infty} c_\nu \psi_\nu(x),$$

and

$$\frac{d}{dx} f(x) = \sum_{\nu=1}^{\infty} c_\nu \frac{d}{dx} \psi_\nu(x).$$

Both series converge uniformely and absolutely on $[a, b]$. ∎

3. Orthogonal polynomials on a countable set of points

3.1. *Introduction.* Let $\mathfrak{A} = \{\mu_\nu, A_\nu\}_{\nu=1,2,3,\ldots}$ be a countable (infinite) set of pairs of real numbers. Put $\alpha = \inf_\nu \mu_\nu$, $\beta = \sup_\nu \mu_\nu$. Assume that the sequence $\mathfrak{C} = \{\mu_\nu\}_{\nu=1,2,3,\ldots}$ does not contain any infinite subsequence of equal numbers, that $A_\nu \neq 0$ for any ν, and that

$$\sum_{\nu=1}^{\infty} A_\nu^2 = 1, \quad 0 < |\alpha| + |\beta| < \infty.$$

Let f and g be arbitrary real functions defined and bounded on $\mathfrak{C}$, and put

$$(f, g) = \sum_{\nu=1}^{\infty} A_\nu^2 f(\mu_\nu) g(\mu_\nu),$$

$$\|f\| = \sqrt{(f,f)} \geqq 0.$$

Let $P_1(x), P_2(x), \ldots, P_{n+1}(x)$ be arbitrary polynomials such that P_k is of the degree $k-1$ (exactly), and $P_1(x) \equiv 1$. Then there exist such constants $\alpha_{n1}, \alpha_{n2}, \ldots, \alpha_{n,n}, \alpha_{n,n+1}$ that $\alpha_{n,n+1} \neq 0$ and

$$\alpha_{n,n+1} P_{n+1}(x) = (x - \alpha_{n,n}) P_n(x) - \sum_{\nu=1}^{n-1} \alpha_{n\nu} P_\nu(x). \tag{9}$$

Clearly, the formula (9) holds, because $xP_n(x)$ is of the degree n, $\{P_1, P_2, \ldots, P_{n+1}\}$ being the basis of the space of polynomials of the degree $\leqq n$.

3.2. *Definition of polynomials orthogonal on the set* $\mathfrak{A}$. Let $P_0(x) \equiv 0$, $P_1(x) \equiv 1$.

Assume the polynomials $P_0, P_1, \ldots, P_n$ to be already defined. P_{n+1} is a polynomial of the degree n (exactly), such that

(i) $(P_{n+1}, P_\nu) = \delta_{n+1,\nu}$ for $\nu = 0, 1, 2, \ldots, n+1$,

(ii) the coefficient of x^n in P_{n+1} is positive.

3.3. *Some properties of polynomials orthogonal on the set* $\mathfrak{A}$. Let us note that $(P_{k+1}, W) = 0$ for any polynomial W of the degree $m < k$, because $W(x) = \sum_{j=1}^{m+1} \beta_j P_j(x)$, where the β_j are constants, and $(P_{kn}, W) = \sum_{j=1}^{m+1} \beta_j (P_{k+1}, P_j) = 0$. Thus for the coefficients α_{nj} of the formula (9) we obtain

$$\alpha_{nj} = \sum_{\nu=1}^{\infty} A_\nu^2 \mu_\nu P_n(\mu_\nu) P_j(\mu_\nu), \quad j = 0, 1, 2, \ldots, n \tag{10}$$

$$\alpha_{n,n+1}^2 = \sum_{\nu=1}^{\infty} A_\nu^2 \mu_\nu^2 [p_n(\mu_\nu)]^2 - \alpha_{n,n}^2 - \alpha_{n,n-1}^2. \tag{11}$$

By symmetry it follows from the formula (10) that for $j = 1, 2, 3, \ldots, n$ $\alpha_{nj} = \alpha_{jn}$ and $\alpha_{nj} = 0$ for $j = 1, 2, 3, \ldots, n-2$.

In this case the formula (9) takes the following form

$$\alpha_{n,n+1} P_{n+1}(x) = (x - \alpha_{n,n}) P_n(x) - \alpha_{n,n-1} P_{n-1}(x)$$

where

$$\alpha_{n,n-1} = \alpha_{n-1,n} = \sum_{\nu=1}^{\infty} A_\nu^2 \mu_\nu P_n(\mu_\nu) P_{n-1}(\mu_\nu)$$

$$\alpha_{n,n} = \sum_{\nu=1}^{\infty} A_\nu^2 \mu_\nu [P_n(\mu_\nu)]^2 \tag{12}$$

$$\alpha_{n,n+1} = \sqrt{\sum_{\nu=1}^{\infty} A_\nu^2 \mu_\nu^2 [P_n(\mu_\nu)]^2 - \alpha_{n,n-1}^2 - \alpha_{n,n}^2} > 0.$$

Put $P^*_{n+1}(x)=(x-\alpha_{n,n})P_n(x)-\alpha_{n,n-1}P_{n-1}(x)$, then $\alpha_{n,n+1}=\|P^*_{n+1}\|\geqq 0$, hence $\alpha_{n,n+1}$ is always real.

The process defined by (12), determining the sequence of polynomials $\{P_k\}_{k=1,2,3,\ldots}$ fails when $\|P^*_{n+1}\|=0$ for some n. It can occur only when the set $\mathfrak{C}$ is finite, μ_ν being zeros of the polynomial P^*_{n+1}.

Let us observe that from the last formula of (12) we get

$$\alpha^2_{n,n+1}+\alpha^2_{n,n}+\alpha^2_{n,n-1}=\sum_{\nu=1}^{\infty}A^2_\nu\mu^2_\nu[P_n(\mu_\nu)]^2\leqq\max\{\alpha^2,\beta^2\}. \tag{13}$$

This inequality implies the uniform boundedness of the sequence of triads $\alpha_{n,n-1}$, $\alpha_{n,n}$, $\alpha_{n,n+1}$ under the condition that $\alpha^2,\beta^2<\infty$.

THEOREM 3.3.1 ([2]). *All zeros of P_{n+1}, $n=1, 2, \ldots$ are real, simple, and are contained in the open interval (α, β).*

PROOF. Since $\sum_{n=1}^{\infty}A^\nu_2P_{n+1}(\mu_\nu)=0$ and $n>0$, P_{n+1} changes sign in (α,β). Let $x_1, x_2, x_3, \ldots, x_l$ be the set of all different points of (α,β) in which a change of sign occurs. Then the polynomial $(x-x_1)\ldots(x-x_l)P_{n+1}(x)$ has a constant sign in (α,β), if not equal to zero. Since $l<n$ would imply $\sum_{\nu=1}^{\infty}A^2_\nu(\mu_\nu-x_1)\ldots(\mu_\nu-x_l)P_{n+1}(\mu_\nu)=0$ which is impossible, therefore $l=n$, and $x_1, x_2, \ldots, x_n$ are simple zeros of the polynomial P_{n+1} (which is of the degree n).

THEOREM 3.3.2 ([2]). *Let $x_1<x_2<\ldots<x_n$ be the zeros of the polynomial P_{n+1}. Put $x_0=\alpha$, $x_{n+1}=\beta$. Then any open interval $(x_\nu, x_{\nu+1})$ contains exactly one zero of P^*_{n+2}.*

PROOF. It is easy to see that

$$P_1(x)P_1(y)+\ldots+P_{n+1}(x)P_{n+1}(y)=\frac{P^*_{n+2}(x)P_{n+1}(y)-P_{n+1}(x)P^*_{n+2}(y)}{x-y},$$

hence, letting $y\to x$:

$$\sum_{\nu=1}^{n+1}P^2_\nu(x)=P^{*\prime}_{n+2}(x)P_{n+1}(x)-P^*_{n+2}(x)P'_{n+1}(x)$$

and

$$P^{*\prime}_{n+2}(x)P_{n+1}(x)-P^*_{n+2}(x)P'_{n+1}(x)>0$$

because $P_1(x)\equiv 1$. Let ξ and η be two consecutive zeros of P_{n+1}, and suppose $\xi<\eta$. Then

$$P'_{n+1}(\xi)P'_{n+1}(\eta)<0,$$

$$P^*_{n+2}(\xi)P'_{n+1}(\xi)<0,$$

$$P^*_{n+2}(\eta)P'_{n+1}(\eta)<0,$$

therefore

$$P^*_{n+2}(\xi)P^*_{n+2}(\eta)<0,$$

hence the interval (ξ, η) contains necessarily an odd number of zeros of P^*_{n+2}. Put $\xi=x_n$ and $\eta=\beta$, then $P'_{n+1}(\xi)>0$ and $P^*_{n+2}(\xi)<0$. But, simultanously, $P^*_{n+2}(\beta)>0$, hence the interval (ξ, η) contains again an odd number of zeros. The analoguous argument may be applied in case $\xi=\alpha$ and $\eta=x_1$. Since the number of intervals $(x_\nu, x_{\nu+1})$ is $n+1$ and the degree of P^*_{n+2} is $n+1$, eah of the intervals contains exactly one zero. ∎

THEOREM 3.3.3 ([2]). *Let $\{P_k\}_{k=0,1,2,\ldots}$ be a system of polynomials orthogonal on $\mathfrak{A}$, and $x_1, x_2, x_3, \ldots, x_n$ be all the zeros of P_{n+1}. Take any polynomial $\varrho(x)$ of the degree $\leqq 2n-1$. Then*

$$\sum_{\nu=1}^{\infty} A_\nu^2 \varrho(\mu_\nu) = \sum_{j=1}^{n} \varrho(x_j) \sum_{\nu=1}^{\infty} A_\nu^2 l_j(\mu_\nu) \tag{14}$$

where

$$l_j(x) = \frac{P_{n+1}(x)}{P'_{n+1}(x_j)(x-x_j)}. \tag{15}$$

PROOF. Let the polynomial $L(x)$ of the degree $n-1$ take the values $\varrho(x_j)$ at the points $x_j, j=1, 2, \ldots, n$. Then

$$L(x) = \sum_{j=1}^{n} \varrho(x_j) l_j(x)$$

and

$$\varrho(x) = L(x)+r(x),$$

further

$$\varrho(x_j) = L(x_j), \quad r(x_j) = 0, \quad j = 1, 2, \ldots, n$$

and $r(x)=\omega(x)P_{n+1}(x)$, where ω is a polynomial of the degree $\leqq n-1$. But

$$\sum_{\nu=1}^{\infty} A_\nu^2 \varrho(\mu_\nu) =$$

$$= \sum_{\nu=1}^{\infty} A_\nu^2 L(\mu_\nu) + \sum_{\nu=1}^{\infty} A_\nu^2 \omega(\mu_\nu) P_{n+1}(\mu_\nu) = \sum_{\nu=1}^{\infty} A_\nu^2 L(\mu_\nu)$$

because the second member vanishes, ω being of the degree $\leqq n-1$. In conclusion,

$$\sum_{\nu=1}^{\infty} A_\nu^2 \varrho(\mu_\nu) = \sum_{\nu=1}^{\infty} A_\nu^2 \sum_{j=1}^{\infty} \varrho(x_j) l_j(\mu_\nu) = \sum_{j=1}^{n} \varrho(x_j) \sum_{\nu=1}^{\infty} A_\nu^2 l_j(\mu_\nu).$$

THEOREM 3.3.4 ([2]).

$$B_i^2 = \sum_{\nu=1}^{\infty} A_\nu^2 l_i(\mu_\nu) = \sum_{\nu=1}^{\infty} A_\nu^2 l_i^2(\mu_\nu)$$

and

$$\sum_{j=1}^{n} B_j^2 = 1.$$

PROOF. Put $\varrho(x)=l_i(x)$ in the formula (14); then

$$0 < \sum_{\nu=1}^{\infty} A_\nu^2 l_i^2(\mu_\nu) = \sum_{j=1}^{n} \delta_{ij} \sum_{\nu=1}^{\infty} A_\nu^2 l_j(\mu_\nu) = \sum_{\nu=1}^{\infty} A_\nu^2 l_i(\mu_\nu)$$

and

$$\sum_{j=1}^{n} B_j^2 = \sum_{j=1}^{n} \sum_{\nu=1}^{\infty} A_\nu^2 l_j(\mu_\nu) = \sum_{\nu=1}^{\infty} A_\nu^2 \sum_{j=1}^{n} l_j(\mu_\nu).$$

The polynomial $Q(x)=\sum_{j=1}^{n} l_j(x)$ is of the degree $\leqq n-1$ and $Q(x_j)=1$ for $j=1,2,\ldots,n$. Hence $Q(x)-1$ vanishes at n different points thus $Q(x)\equiv 1$. Finally, $\sum_{j=1}^{n} B_j^2 = \sum_{\nu=1}^{\infty} A_\nu^2 = 1$. ∎

COROLLARY. The formula (14) for the polynomial $\varrho(x)$ of the degree $\leqq 2n-1$ can be now written as follows:

$$\sum_{\nu=1}^{\infty} A_\nu^2 \varrho(\mu_\nu) = \sum_{j=1}^{n} \varrho(x_j) B_j^2$$

where

$$\sum_{j=1}^{n} B_j^2 = 1, \quad B_j^2 > 0, \quad j = 1, 2, \ldots, n.$$

This means that for any polynomial of the degree $\leqq 2n-1$ the "nodes" μ_ν, in the sum on the left hand side of (14) can be replaced by the zeros of P_{n+1}, and the "weights" A_ν can be replaced by the new "weights" B_j.

THEOREM 3.3.5. ([2]). *Given an arbitrary point μ_k; in any neighbourhood of μ_k there are zeros of P_{n+1}, for n sufficiently great.*

PROOF. Let $\varepsilon>0$ be an arbitrary number but so small that there are no points $\mu_s\in\mathfrak{C}$, $\mu_s\neq\mu_k$ in the interval $[\mu_k-\varepsilon, \mu_k+\varepsilon]$. Put

$$f(x) = \begin{cases} 0 & \text{outside the closed interval } [\mu_k-\varepsilon, \mu_k+\varepsilon] \\ (x-\mu_k+\varepsilon)(\mu_k+\varepsilon-x) & \text{inside } [\mu_k-\varepsilon, \mu_k+\varepsilon]. \end{cases}$$

Choose the number $\delta>0$ (depending on ε) such that

$$\sum_{\nu=1}^{\infty} A_\nu^2 f(\mu_\nu) = \sum_{\mu_{k_i}=\mu_k} A_{k_i}^2 \varepsilon^2 > \delta > 0. \tag{16}$$

Since the function f is continuous, by Weierstrass Theorem there is a polynomial ϱ of the degree m such that for all $x\in[a, b]$

$$|f(x)-\varrho(x)| < \frac{\delta}{2}.$$

Take $n\geqq\frac{m+1}{2}$ and suppose that there are no zeros of P_{n+1} in the interval $[\mu_k-\varepsilon, \mu_k+\varepsilon]$. Obviously,

$$\left|\sum_{\nu=1}^{\infty} A_\nu^2 f(\mu_\nu) - \sum_{\nu=1}^{\infty} A_\nu^2 \varrho(\mu_\nu)\right| = \left|\sum_{\mu_{k_i}=\mu_k} A_{k_i}^2 \varepsilon^2 - \sum_{\nu=1}^{\infty} A_\nu^2 \varrho(\mu_\nu)\right| \leqq \tag{17}$$

$$\leqq \sum_{\nu=1}^{\infty} A_\nu^2 |f(\mu_\nu)-\varrho(\mu_\nu)| \leqq \frac{\delta}{2}.$$

By Theorems 3.3.3 and 3.3.4 $\sum_{\nu=1}^{\infty} A_\nu^2 \varrho(\mu_\nu) = \sum_{j=1}^{n} B_j^2 \varrho(x_j)$ where $x_1, x_2, \ldots, x_n$ are the zeros of P_{n+1}. Using the hypothesis: $x_j \notin [\mu_k - \varepsilon, \mu_k + \varepsilon]$ we obtain $|\varrho(x_j)| \leqq \frac{\delta}{2}$ for $j = 1, 2, \ldots, n$, and by the Corollary to Theorem 3.3.4.

$$\left| \sum_{\nu=1}^{\infty} A_\nu^2 \varrho(\mu_\nu) \right| \leqq \sum_{j=1}^{u} B_j^2 |\varrho(x_j)| \leqq \frac{\delta}{2}. \tag{18}$$

Finally by (16), (17), and (18)

$$\frac{\delta}{2} < \sum_{\mu_{k_i} = \mu_k} A_{k_i}^2 \varepsilon^2 - \frac{\delta}{2} \leqq \sum_{\mu_{k_i} = \mu_k} A_{k_i}^2 \varepsilon^2 - \left| \sum_{\nu=1}^{\infty} A_\nu^2 \varrho(\mu_\nu) \right| \leqq$$

$$\leqq \left| \sum_{\mu_{k_i} = \mu_k} A_{k_i}^2 \varepsilon^2 - \sum_{\nu=1}^{\infty} A_\nu^2 \varrho(\mu_\nu) \right| \leqq \frac{\delta}{2}.$$

This contradiction shows that the supposition that all the zeros are outside the interval $[\mu_k - \varepsilon, \mu_k + \varepsilon]$ is false. ∎

THEOREM 3.3.6. *Let ξ and η be two elements of the sequence $\mathfrak{C} = \{\mu_\nu\}_{\nu=1,2,\ldots}$ such that $\xi < \eta$ and there are no $\mu_s \in \mathfrak{C}$ satisfying the inequality $\xi < \mu_s < \eta$.*

Then the interval (ξ, η) contains at most one zero x_j of the polynomial P_{n+1}, $n = 1, 2, 3, \ldots$.

PROOF. Assume the zeros $x_{p+1}, x_{p+2}, \ldots, x_{p+s}$ of P_{n+1} to be in the interval (ξ, η). Take the polynomial

$$\varrho(x) = -(x - x_1)^2 (x - x_2)^2 \ldots (x - x_p)^2 (x - \xi)(x - \eta)(x - x_{p+s+1})^2 \ldots (x - x_n)^2.$$

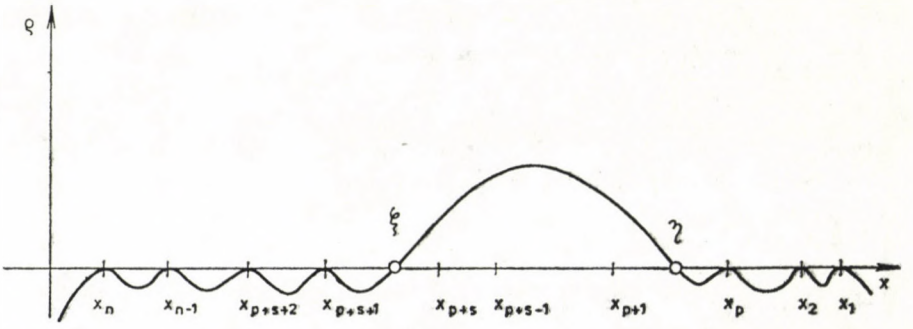


Fig. 1

Observe that

(i) ϱ is of the degree $2n - 2s + 2$,
(ii) $\varrho(x_j) = 0$ for $j = 1, 2, \ldots, p$ and $j = p+s+1, \ldots, n$,
(iii) $\varrho(x) > 0$ for $x \in (\xi, \eta)$,
(iv) $\varrho(x) \leqq 0$ for $x \notin (\xi, \eta)$,
(v) $2n - 2s + 2 \leqq 2n - 1$ when $s \geqq 2$.

Suppose that $s \geqq 2$; in this case we can use Theorem 3.3.3:

$$\sum_{\nu=1}^{\infty} A_\nu^2 \varrho(\mu_\nu) = \sum_{j=1}^{n} B_j^2 \varrho(x_j)$$

where $x_1, x_2, \ldots, x_n$ are the zeros of P_{n+1}. But $\sum_{\nu=1}^{\infty} A_\nu^2 \varrho(\mu_\nu) \leqq 0$, since $\varrho(\mu_2) \leqq 0$ for any $\mu_\nu \in \mathfrak{C}$. Moreover, by 3.3.4 and (i)–(v)

$$0 \geqq \sum_{\nu=1}^{\infty} A_\nu^2 \varrho(\mu_\nu) = \sum_{j=1}^{n} B_j^2 \varrho(x_j) = \sum_{j=p+1}^{p+s} B_j^2 \varrho(x_j) > 0$$

which is the contradiction. This proves that $s < 2$. ∎

Theorems 3.3.5 and 3.3.6 together imply the following

THEOREM 3.3.7. (The Qualitative Convergence Theorem.) *Assume μ to be the unique accumulation point of the sequence $\mathfrak{C}$. Take an arbitrary positive number δ and all elements of $\mathfrak{C}$ outside the closed interval $[\mu-\delta, \mu+\delta]$. Order all different elements of this new finite set in such a way that*

$$\mu_1 > \mu_2 > \mu_3 > \ldots > \mu_d.$$

Let ε satisfy the inequality

$$0 < \varepsilon < \min_{i \neq j} \min \left\{ \frac{\mu_i - \mu_j}{2}, \frac{\mu_i - \mu - \delta}{2}, \frac{\mu_i - \mu + \delta}{2} \right\}.$$

Then for any δ and ε satisfying all the above conditions, there is an integer N such that, for any $n > N$, each open interval $(\mu_i - \varepsilon, \mu_i + \varepsilon)$ $i = 1, 2, \ldots, d$ contains exactly one zero x_i of the polynomial P_{n+1}. Moreover,

$$0 < \mu_i - x_i < \varepsilon \quad \text{for} \quad \mu_i > \mu,$$

$$0 < x_i - \mu_i < \varepsilon \quad \text{for} \quad \mu_i < \mu,$$

$i = 1, 2, \ldots, d$. All remaining zeros x_j of P_{n+1} satisfy the condition

$$|\mu - x_j| < \delta.$$

PROOF. Theorems 3.3.5 and 3.3.6 imply the existence of an integer N such that, for any $n > N$, only one of the following three possibilities can occur (see the diagram for $d = 5$, Fig. 2).

The following theorem is of a quantitative character and gives some idea on the rate of convergence.

THEOREM 3.3.8. Let x_s be the zero of P_{n+1} *nearest to* $\mu_s \in \mathfrak{C}$. *For n great enough:*

$$|x_s - \mu_s| \leqq \frac{\sqrt{2}(\beta - \alpha)}{n \sqrt{\sum_{\mu_{s_i} = \mu_s} A_{s_i}^2}}, \tag{19}$$

and if $\mu_s = \alpha$ or $\mu_s = \beta$, then:

$$|x_s - \mu_s| \leqq \frac{\beta - \alpha}{2n^2 \sum_{\mu_{s_i} = \mu_s} A_{s_i}^2}. \tag{20}$$

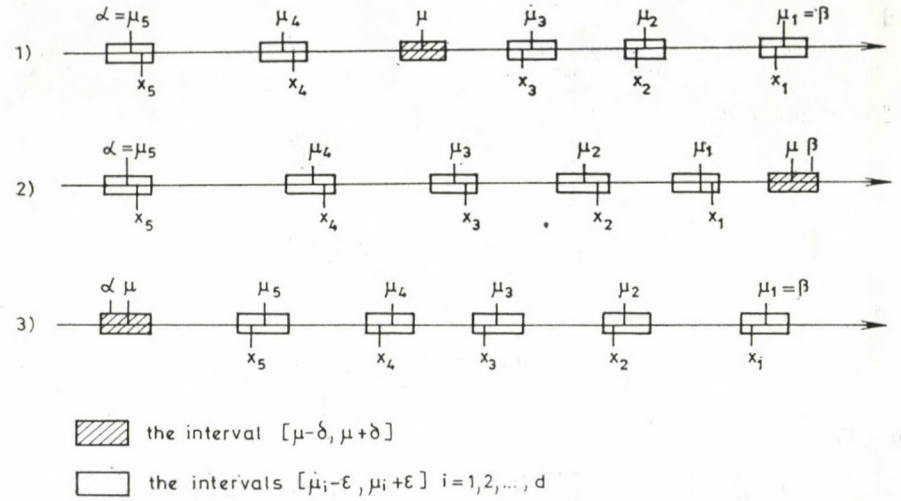


Fig. 2

PROOF. Let $x=\frac{\alpha+\beta}{2}+\frac{\alpha-\beta}{2}\cos\vartheta$. When $0\leqq\vartheta\leqq\pi$, then $\alpha\leqq x\leqq\beta$. Let $x_1<x_2<x_3<\ldots<x_n$ be the zeros of P_{n+1}. Define ϑ_j and γ_s as numbers from the interval $[0, \pi]$ such that:

$$x_j=\frac{\alpha+\beta}{2}+\frac{\alpha-\beta}{2}\cos\vartheta_j,$$

$$\mu_s=\frac{\alpha+\beta}{2}+\frac{\alpha-\beta}{2}\cos\gamma_s,$$

and take the following polynomial $\varrho(x)$ of the degree $2n-1$ on x:

$$\varrho(x)=\left[\frac{\sin n(\gamma_s-\vartheta)}{\sin\frac{\gamma_s-\vartheta}{2}}\right]^2+\left[\frac{\sin n(\gamma_s+\vartheta)}{\sin\frac{\gamma_s+\vartheta}{2}}\right]^2.$$

It is easy to see that

$$\varrho(\mu_s)=\left[\frac{\sin 2n\gamma_s}{\sin\gamma_s}\right]^2+4n^2\geqq 4n^2.$$

Using 3.3.3 we obtain

$$\sum_{\nu=1}^{\infty}A_\nu^2\varrho(\mu_\nu)=\sum_{j=1}^{n}B_j^2\varrho(x_j). \tag{21}$$

Since among $\vartheta_1, \vartheta_2, \ldots, \vartheta_{d_\Delta}$ the number ϑ_j is the nearest to γ_s, we have

$$0\leqq\frac{|\gamma_s-\vartheta_s|}{2}\leqq\frac{|\gamma_s-\vartheta_j|}{2}\leqq\frac{\pi}{2}$$

and

$$0 \leqq \frac{|\gamma_s - \vartheta_s|}{2} \leqq \frac{\gamma_s + \vartheta_j}{2} \leqq \pi - \frac{|\gamma_s - \vartheta_s|}{2},$$

and, consequently,

$$\left|\sin \frac{\gamma_s - \vartheta_s}{2}\right| \leqq \left|\sin \frac{\gamma_s - \vartheta_j}{2}\right|,$$

$$\left|\sin \frac{\gamma_s - \vartheta_s}{2}\right| \leqq \left|\sin \frac{\gamma_s + \vartheta_j}{2}\right|,$$

thus

$$\varrho(x_j) \leqq \frac{2}{\left[\sin \frac{\gamma_s - \vartheta_s}{2}\right]^2}.$$

By (21)

$$4 \sum_{\mu_{s_i} = \mu_s} A_{s_i}^2 n^2 \leqq \sum_{\mu_{s_i} = \mu_s} A_{s_i}^2 \varrho(\mu_s) \leqq \sum_{\nu=0}^{\infty} A_\nu^2 \varrho(\mu_\nu) = \sum_{j=1}^{n} B_j^2 \varrho(x_j) \leqq$$

$$\leqq \frac{2}{\left[\sin \frac{\gamma_s - \vartheta_s}{2}\right]^2}.$$

The last inequality implies

$$\left|\sin \frac{\gamma_s - \vartheta_s}{2}\right| \leqq \frac{1}{n\sqrt{2 \sum_{\mu_{s_i} = \mu_s} A_{s_i}^2}}. \tag{22}$$

Because of

$$\mu_s - x_s = (\beta - \alpha)\left[\sin \gamma_s \cos \frac{\gamma_s - \vartheta_s}{2} \sin \frac{\gamma_s - \vartheta_s}{2} - \cos \gamma_s \left(\sin \frac{\gamma_s - \vartheta_s}{2}\right)^2\right]$$

(see the definition of x_s and ϑ_s) it follows from the formula (22) that

$$|\mu_s - x_s| \leqq \frac{\sqrt{2}(\beta - \alpha)}{n\sqrt{\sum_{\mu_{s_i} = \mu_s} A_{s_i}^2}}.$$

When $\mu_s = \alpha$ or $\mu_s = \beta$ (then $\gamma_s = 0$ or $\gamma_s = \pi$), then only the last term of the expression for $\mu_s - x_s$ subsists, and

$$|\mu_s - x_s| = \frac{\beta - \alpha}{2n^2 \sum_{\mu_{s_i} = \mu_s} A_{s_i}^2}.$$

Note. Theorems similar to 3.3.1, 3.3.2, 3.3.3, 3.3.4, 3.3.5, but for a continuous weight function can be found in Szegö's book [2]. In the remaining theorems of this paragraph the assumption on the discrete character of the weights is essential.

4. The Lánczos Process

4.1. *Definition of the process.* Let us come back to the self adjoint eigenvalue problem (1) (2). Put for any m dimensional vectorfunction u of the class C^2 on $[a, b]$

$$L[u](x) = \frac{d}{dx}u(x) - A(x)u(x).$$

The following process, defining the sequence of m dimensional vectorfunctions $\{u_k\} = U$, is called the Lánczos Process:

(0) Let U_{-1} be any m dimensional vectorfunction continuous on $[a, b]$. Define u_0 as the solution of the following linear homogenous boundary value problem

$$L[u_0] = Bu_{-1}$$

$$Mu_0(a) + Nu_0(b) = 0.$$

(1) Define z_0 as the solution of the similar boundary value problem:

$$L[z_0] = Bu_0$$

$$Mz_0(a) + Nz_0(b) = 0$$

and $u_1 = \alpha z_0$ with α constant, such that

$$\int_a^b u_1^T(x) S(x) u_1(x)\, dx = 1.$$

(2) Let the function z_1 be the solution of the boundary value problem

$$L[z_1] = Bu_1$$

$$Mz_1(a) + Nz_1(b) = 0,$$

and define u_2 by the conditions

$$\beta_{12}u_2 + \beta_{11}u_1 = z_1$$

where

$$\beta_{11} = \int_a^b u_1^T(x) S(x) z_1(x)\, dx$$

$$\beta_{12} = \sqrt{\int_a^b z_1^T(x) S(x) z_1(x)\, dx - \beta_{11}^2}.$$

If $\beta_{12} > 0$, then u_2 is uniquely determined, if $\beta_{12} = 0$, then the process fails at $n = 1$.

Assume now the vectorfunctions $u_1, u_2, \ldots, u_k$ of the sequence U to be already defined in the first k steps of the process.

$(k+1)$ Let z_k be the solution of the boundary value problem

$$L[z_k] = Bu_k$$
$$Mz_k(a)+Nz_k(b) = 0,$$

we define u_{k+1} by the condition

$$\sum_{j=1}^{k+1} \beta_{k,j} u_j = z_k$$

where

$$\beta_{kj} = \int_a^b u_j^T(x) S(x) z_k(x)\, dx \quad \text{for} \quad j = 1, 2, \ldots, k;$$

$$\beta_{k,k+1} = \sqrt{\int_a^b z_k^T(x) S(x) z_k(x)\, dx - \sum_{j=1}^{k} \beta_{k,j}^2} \geqq 0.$$

If $\beta_{k,k+1}>0$, then u_{k+1} is uniquely determined, if $B_{k,k+1}=0$, then the process fails at $n=k$. ∎

4.2 *Properties of the sequence U.* Observe that the functions $u_1, u_2, \ldots, u_k, \ldots$ form an orthonormal system in the sense that

$$\int_a^b u_k^T(x) S(x) u_j(x)\, dx = \delta_{k,j}.$$

For $k=1$ the orthonormality condition is obviously satisfied. Assume now the orthonormality conditions to be satisfied for $u_1, u_2, \ldots, u_k$. Note that

$$0 \leqq \left\| z_k - \sum_{j=1}^{k} \beta_{kj} u_j \right\|^2 = \int_a^b \left(z_k^T(x) - \sum_{j=1}^{k} \beta_{k,j} u_j^T(x) \right) S(x) \left(z_k(x) - \sum_{j=1}^{k} \beta_{k,j} u_j(x) \right) dx =$$

$$= \int_a^b z_k^T(x) S(x) z_k(x)\, dx - 2 \sum_{j=1}^{k} \beta_{kj}^2 + \sum_{l,s=1}^{k} \beta_{kl} \beta_{k,s} \delta_{l,s} =$$

$$= \int_a^b z_k^T(x) S(x) z_k(x)\, dx - \sum_{j=1}^{k} \beta_{k,j}^2 = \|\beta_{k,k+1} u_{k+1}\|^2 = \beta_{k,k+1}^2.$$

Thus the coefficient $\beta_{k,k+1}$ is real (positive or zero), and

$$\|u_{k+1}\|^2 = \int_a^b u_{k+1}^T(x) S(x) u_{k+1}(x)\, dx = 1.$$

Moreover, for $j=1, 2, \ldots, k$,

$$\int_a^b u_{k+1}^T(x) S(x) u_j(x)\, dx = \int_a^b z_k^T(x) S(x) u_j(x)\, dx +$$

$$- \sum_{l=1}^{k} \beta_{k,l} \int_a^b u_l^T(x) S(x) u_j(x)\, dx = \beta_{k,j} - \sum_{l=1}^{k} \beta_{k,l} \delta_{l,j} = 0,$$

which proves the orthonormality of the system U.

THEOREM 4.2.1. *For* $u_k \in U$, $k=1, 2, 3, \ldots$,

$$u_k(x) = \sum_{\nu=1}^{\infty} c_\nu^{(k)} \psi_\nu(x)$$

(23)

$$\frac{d}{dx} u_k(x) = \sum_{\nu=1}^{\infty} c_\nu^{(k)} \frac{d}{dx} \psi_\nu(x)$$

where the $\psi_\nu(x)$ *are othonormal, real eigenfunctions of* (1,) (2),

$$c^{(k)} = \int_a^b \psi_\nu^T(x) S(x) u_k(x)\, dx, \tag{24}$$

and the series (23) *converge uniformely and absolutely on* $[a, b]$.

PROOF. We need only to show the existence of the continuous vectorfunctions g_k and h_k such that, for $k=1, 2, 3, \ldots$,

$$\begin{aligned} L[u_k] &= Bg_k \\ L[g_k] &= Bh_k \\ Mu_k(a) + Nu_k(b) &= 0 \\ Mg_k(a) + Ng_k(b) &= 0. \end{aligned} \tag{25}$$

(See paragraph 2).

For $k=1$ we can take $g_1(x)=u_0(x)$, $h_1(x)=u_{-1}(x)$. Assume now that the condition (25) is satisfied for $k=1, 2, 3, \ldots, i$. Since

$$u_{i+1} = \gamma_{i+1} z_i + \sum_{j=1}^{i} \gamma_j u_j$$

$$\gamma_{i+1} = \frac{1}{\beta_{i,i+1}}, \quad \gamma_j = -\frac{\beta_{i,j}}{\beta_{i,i+1}} \quad j = 1, 2, \ldots, i,$$

we have

$$L[u_{i+1}] = \gamma_{i+1} L[z_i] + \sum_{j=1}^{i} \gamma_j L[u_j] =$$

$$= \gamma_{i+1} Bu_i + \sum_{j=1}^{i} \gamma_j Bg_j = B\left[\gamma_{i+1} u_i + \sum_{j=1}^{i} \gamma_j g_j\right].$$

Moreover,

$$L\left[\gamma_{i+1} u_i + \sum_{j=1}^{i} \gamma_j g_j\right] = \gamma_{i+1} L[u_i] + \sum_{j=1}^{i} \gamma_j L[g_j] =$$

$$= \gamma_{i+1} Bg_i + \sum_{j=1}^{i} \gamma_j Bh_j = B\left[\gamma_{i+1} g_i + \sum_{j=1}^{i} \gamma_j h_j\right],$$

and one can put $g_{i+1}=\gamma_{i+1} u_i + \sum_{j=1}^{i} \gamma_j g_j$ and $h_{i+1}=\gamma_{i+1} g_i + \sum_{j=1}^{i} \gamma_j h_j$, the boundary conditions being obviously satisfied. ∎

Now by theorem 4.2.1, for any $u_k \in U$, $k=1, 2, 3, \ldots$,

$$L[z_k] = L\left[\sum_{j=1}^{k+1} \beta_{kj} u_j\right] = \sum_{j=1}^{k+1} \beta_{k,j} L[u_j] = Bu_k,$$

then

$$\sum_{j=1}^{k+1} \beta_{kj} \sum_{\nu=1}^{\infty} c_\nu^{(j)} L[\psi_\nu] = B \sum_{\nu=1}^{\infty} c_\nu^{(k)} \psi_\nu,$$

but, since $L[\psi_\nu] = \lambda_\nu B\psi_\nu$, we have

$$B \sum_{\nu=1}^{\infty} \psi_\nu \left[\sum_{j=1}^{k+1} \beta_{kj} c_\nu^{(j)} \lambda_\nu - c_\nu^{(k)}\right] = 0,$$

and using the orthnormality condition for φ_ν, $\nu=1, 2, 3, \ldots$ we obtain finally

$$\sum_{j=1}^{k+2} \beta_{k,j} c_\nu^{(j)} - \mu_\nu c_\nu^{(k)} = 0 \quad k = 1, 2, 3, \ldots \tag{26}$$

where $\mu_\nu = \dfrac{1}{\lambda_\nu}$ $(\lambda_\nu \neq 0)$.

We just proved the following

THEOREM 4.2.2. *The reciprocals μ_ν of the eigenvalues λ_ν of the problem* (1) (2) *are the eigenvalues of the infinite matrix X_∞*

$$X_\infty = \begin{pmatrix} \beta_{11}, \beta_{12}, & 0, & 0, & 0, \ldots \\ \beta_{21}, \beta_{22}, \beta_{23}, & 0, & 0, \ldots \\ \beta_{31}, \beta_{32}, \beta_{33}, \beta_{34}, & 0, \ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots \end{pmatrix}$$

the corresponding infinite eigenvectors being the vectors of the Fourier coefficients of the functions $u_1, u_2, \ldots$:

$$[c_\nu^{(1)}, c_\nu^{(2)}, c_\nu^{(3)}, \ldots \quad]^T. \qquad \blacksquare$$

Let us observe that

$$\int_a^b u_k^T(x) S(x) u_l(x)\, dx = \sum_{\alpha=1}^{\infty} \sum_{\beta=1}^{\infty} c_\alpha^{(k)} c_\beta^{(l)} \int_a^b \psi_\alpha^T(x) S(x) \psi_\beta(x)\, dx =$$

$$= \sum_{\alpha=1}^{\infty} \sum_{\beta=1}^{\infty} c_\alpha^{(k)} c_\beta^{(l)} \delta_{\alpha\beta} = \sum_{\nu=1}^{\infty} c_\nu^{(k)} c_{(l)}^{k} = \delta_{k,l}.$$

Thus the system of all the vectors $[c_1^{(k)}, c_2^{(k)}, c_3^{(k)}, \ldots]^T$ for $k=1, 2, \ldots$ is orthonormal. Assume now that the function u_{-1} of the sequence U has been chosen in such a way that $c_\nu^{(1)} \neq 0$ for all $\nu=1, 2, 3, \ldots$. Put:

$$A_\nu = c_\nu^{(1)}; \quad \sum_{\nu=1}^{\infty} A_\nu^2 = \sum_{\nu=1}^{\infty} c_\nu^{(1)2} = 1,$$

(27)

$$\mathfrak{A} = \{\mu_\nu, A_\nu\}_{\nu=1,2,3,\ldots}; \quad \mu_\nu = \frac{1}{\lambda_\nu} \quad (\lambda_\nu \neq 0),$$

and let $P_1, P_2, P_3, \ldots$ be a system of orthogonal polynomials on $\mathfrak{A}$. Observe that the sequence $\mathfrak{C}=\{\mu_\nu\}_{\nu=1,2,\ldots}$ has exactly one accumulation point $\mu=0$.

THEOREM 4.2.3. *Under the above definitions, for all* $\nu=1, 2, 3, \ldots; k=1, 2, 3, \ldots; j=1, 2, \ldots, k+1,$

$$c_\nu^{(k)} = P_k(\mu_\nu) c_\nu^{(1)} = P_k(\mu_\nu) A_\nu$$

(28)

$$\beta_{kj} = \alpha_{kj}$$

where the β_{kj} *and the* α_{kj} *are, respectively, the coefficients of the recurrence relations for the sequence* U *and for the polynomials* $P_1, P_2, P_3, \ldots$. *Hence, the matrix* X_∞ *from Theorem* 4.2.2 *is symmetric and tridiagonal:*

$$X_\infty = \begin{pmatrix} \alpha_{11}, \alpha_{12}, 0, 0, 0, \ldots \\ \alpha_{21}, \alpha_{22}, \alpha_{23}, 0, 0, \ldots \\ 0, \alpha_{32}, \alpha_{33}, \alpha_{34}, 0, \ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots \end{pmatrix}.$$

PROOF. Since for any $k=1, 2, 3, \ldots$

$$L[z_k] = Bu_k$$

and since the conditions of Theorem 2.1 are satisfied, for z_k,

$$z_k(x) = \sum_{\nu=1}^{\infty} d_\nu^{(k)} \psi_\nu(x),$$

$$L[z_k] = \sum_{\nu=1}^{\infty} d_\nu^{(k)} L[\psi_\nu] = \sum_{\nu=1}^{\infty} d_\nu^{(k)} \lambda_\nu B\psi_\nu = Bu_k = \sum_{\nu=1}^{\infty} c_\nu^{(k)} B\psi_\nu,$$

and, in conclusion,

(29)

$$d_\nu^{(k)} = c_\nu^{(k)} \mu_\nu.$$

Now we start the induction with respect to k with an arbitrary but constant v. Because of $P_1=1$, the first formula of (28) is obviously satisfied for $k=1$. We have

$$\beta_{11} = \int_b^a z_1^T(x) S(x) u_1(x)\, dx =$$

$$= \sum_{\alpha=1}^{\infty} \sum_{\beta=1}^{\infty} \mu_\alpha c_\alpha^{(1)} c_\beta^{(1)} \delta_{\alpha\beta} = \sum_{\nu=1}^{\infty} \mu_\nu (c_\nu^{(1)})^2 =$$

$$= \sum_{\nu=1}^{\infty} A_\nu^2 \mu_\nu [P_1(\mu_\nu)]^2 = \alpha_{11}$$

and

$$\beta_{12}^2 = \int_a^b z_1^T(x) S(x) z_1(x)\, dx - \beta_{11}^2 =$$

$$= \sum_{\nu=1}^{\infty} \mu_\nu^2 [c_\nu^{(1)}]^2 - \alpha_{11}^2 = \sum_{\nu=1}^{\infty} A_\nu^2 \mu_\nu^2 [P_1(\mu_\nu)]^2 - \alpha_{11}^2 = \alpha_{12}^2.$$

Assume now that the formulae (28) hold for $i \leqq k$. From (28) and (12), and by the inductive hypothesis

$$0 = \alpha_{i,i-1} c_\nu^{(i-1)} + (\alpha_{i,i} - \mu_\nu) c_\nu^{(i)} + \alpha_{i,i+1} c_\nu^{(i+1)} =$$

$$= \alpha_{i,i-1} P_{i-1}(\mu_\nu) c_\nu^{(1)} + (\alpha_{i,i} - \mu_\nu) P_i(\mu_\nu) c_\nu^{(1)} + \alpha_{i,i+1} c_\nu^{(i+1)} =$$

$$= -\alpha_{i,i+1} P_{i+1}(\mu_\nu) c_\nu^{(1)} + \alpha_{i,i+1} c_\nu^{(i+1)},$$

hence

$$c_\nu^{(i+1)} = P_{i+1}(\mu_\nu) c_\nu^{(1)} \quad \nu = 1, 2, \ldots.$$

Using (29) for $j \leqq i+1$ we obtain

$$\beta_{i+1,j} = \int_a^b z_{i+1}^T(x) S(x) u_j(x)\, dx =$$

$$= \sum_{\nu=1}^{\infty} \mu_\nu c_\nu^{(i+1)} c_\nu^{(j)} = \sum_{\nu=1}^{\infty} A_\nu^2 \mu_\nu P_{i+1}(\mu_\nu) P_j(\mu_\nu) = \alpha_{i+1,j}$$

and

$$\beta_{i+1,i+2}^2 = \int_a^b z_{i+1}^T(x) S(x) z_{i+1}(x)\, dx - \sum_{j=1}^{i+1} \alpha_{i+1,j}^2 =$$

$$= \sum_{\nu=1}^{\infty} \mu_\nu^2 [c_\nu^{(i+1)}]^2 - \sum_{j=1}^{i+1} \alpha_{i+1,j}^2 = \sum_{\nu=1}^{\infty} A_\nu^2 \mu_\nu P_{i+1}(\mu_\nu) P_i(\mu_\nu) - \sum_{j=1}^{i+1} \alpha_{i+1,j}^2 = \alpha_{i+1,i+2}^2,$$

which completes the proof. ∎

THEOREM 4.2.4. *Let X_n be the n dimensional tridiagonal symmetric matrix formed by the first n rows and columns of the infinite matrix X_∞ defined in the Lanczos Process* (see Theorems 4.2.2 and 4.2.3).

Take numbers $\delta>0$, $\varepsilon>0$ *and* $\mu_1>\mu_2>\mu_3>\ldots>\mu_d$ *satisfying the hypotheses of Theorem* 3.3.7. *Denote by* x_s *the eigenvalue of* X_n *nearest to* μ_s *for* $s=1, 2, \ldots, d$.

Then, there exists an integer N *such that for* $n>N$, $s=1, 2, \ldots, d$:

$$0 < \mu_s - x_s < \varepsilon \quad \text{for} \quad \mu_s > 0,$$

$$0 < x_s - \mu_s < \varepsilon \quad \text{for} \quad \mu_s < 0,$$

and $|x_i|<\delta$ *for all remaining eigenvalues* x_j *of* X_n *(we can assume* $n>d$*).*

Moreover,

$$|x_s - \mu_s| \leqq \frac{\sqrt{2}(\beta-\alpha)}{n\sqrt{R_s}} \quad s = 1, 2, \ldots, d$$

and

$$|x_s - \mu_s| \leqq \frac{\beta-\alpha}{2n^2 R_s} \quad \text{if } \mu_s = \alpha \text{ or } \mu_s = \beta,$$

where $\alpha=\inf_\nu \mu_\nu$, $\beta=\sup_\nu \mu_\nu$, $\mu_\nu=\dfrac{1}{\lambda_\nu}$ (λ_ν *are the eigenvalues of* (1) (2))

$$R_s = \sum_{\nu_{s_i}=\mu_s} A_{s_i}^2; \; A_\nu = c_\nu^{(1)} \text{ (see (27))}.$$

For the matrix X_∞ *the following estimation holds*

$$\alpha_{k,k-1}^2 + \alpha_{k,k}^2 + \alpha_{k,k+1}^2 \leqq \max[\alpha^2, \beta^2].$$

PROOF. The theorem is an obvious corollary of the previous theorems, because, as it is easy to see the eigenvalues of the n dimensional matrix X_n coincide with the zeros of the polynomial P_{n+1}. ∎

BIBLIOGRAPHY

[1] LÁNCZOS, C.: An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, *J. Res. National Bur. of Standards* **45** № 4 (1950), 256—288.

[2] SZEGŐ, G.: *Orthogonal polynomials*, Am. Math. Soc. (1939).

[3] BLISS, G. A.: A boundary value problem for a system of ordinary linear differential equations, *Transactions Amer. Math. Soc.* **28** (1926), 561—584.

[4] BLISS, G. A.: Definitely selfadjoint boundary value problems, *Transactions Amer. Math. Soc.* **44** (1938), 413—428.

Instytut Matematyczny PAN, ul. Śniadeckich 8, 00-950 Warszawa P.O.B. 137, Poland

(Received January 14, 1975)

INDEX

Printed in Hungary

A kiadásért felel az Akadémiai Kiadó igazgatója — Műszaki szerkesztő: Botyánszky Pál
A kézirat a nyomdába érkezett: 1975. IX. 19. — Terjedelem: 23,75 (A/5) ív, 48 ábra

75-4126 — Szegedi Nyomda — F. v.: Dobó József

Die *Studia Scientiarum Mathematicarum Hungarica* ist eine Halbjahrsschrift der Ungarischen Akademia der Wissenschaften. Sie veröffentlicht Originalbeiträge aus dem Bereich der Mathematik in deutscher, englischer, französischer oder russischer Sprache. Es erscheint jährlich ein Band.

Adresse der Redaktion: 1053 Budapest V., Reáltanoda u. 13—15, Ungarn.
Technischer Redaktor: E. Deák

Abonnementspreis pro Band (pro Jahr): $ 16.00. Bestellbar bei Buch- und Zeitungs-Aussenhandelsunternehmen *Kultúra* (Budapest 62, P. O. B. 149), oder bei den Vertretungen im Ausland.
Austauschabmachungen können mit der Bibliothek des Matematischen Instituts (1053 Budapest V., Reáltanoda u. 13—15) getroffen werden.
Die zur Veröffentlichung bestimmten Manuskripte sind in zwei Exemplaren an die Redaktion zu schicken.

Studia Scientiarum Mathematicarum Hungarica est une revue biannuelle de l'Académie Hongroise des Sciences publiant des essais originaux, en francais, anglais, allemand ou russe, du domaine des mathématiques.

Rédaction: 1053 Budapest V., Reáltanoda u. 13—15, Hongrie.
Rédacteur téchnique: E. Deák

Le prix de l'abonnement: $ 16.00 par an (volume). On s'abonne chez *Kultúra*, Société pour le Commerce de Livres et Journaux (Budapest 62, P. O. B. 149) ou chez ses représentants à l'étranger.
Pour établir des relations d'échange on est prié de s'adresser à la Bibliothèque de l'Institut de Mathématique (1053 Budapest V., Reáltanoda u. 13—15).
On est prié 'envoyer les articles destinés à la publication en deux exemplaires à l'adresse de la Rédaction

---

Studia Scientiarum Mathematicarum Hungarica — выходит два раза в год в Издании Академии Наук Венгрии. Журнал публикует оригинальные исследования в области математики на немецком, английском, французком и русском языках. Отдельные вышуски составляют ежегодно одни том.

Адрес редакции: 1053 Budapest V., Reáltanoda u. 13—15, Венгрия.
Технический редактор: E. Deák.

Подписная цена на год (за однии том): ? 16.00. Подписка на журнал принимается Внешнеторговым предприятием „Култура" (Budapest 62, P. O. B. 149) или его представительствами за границей.
По пиводу отношения обмена просим обрашаться к Библиотеке Института Математики (1953 Budapest V., Reáltanoda u. 13—15).
Работы, преднасначенные для опубликования в журнале следует направлять по адресу редакции в двух экземплярах.

All the reviews of the Hungarian Academy of Sciences may be obtained among others from the following bookshops:

ALBANIA

Ndermarja Shtetnore e Botimeve
Tirana

AUSTRALIA

A. Keesing
Box 4886, GPO
Sidney

AUSTRIA

Globus Buchvertrieb
Salzgries 16
Wien I.

BELGIUM

Office International de Librairie
30, Avenue Marnix
Bruxelles 5
Du Monde Entier
5, Place St. Jean
Bruxelles

BULGARIA

Raznoiznos
1 Tzar Assen
Sofia

CANADA

Pannonia Books
2 Spadina Road
Toronto 4, Ont.

CHINA

Waiwen Shudian
Peking
P.O.B. Nr. 88.

CHECHOSLOVAKIA

Artia A. G.
Ve Smeckách 30
Praha II.
Postova Novinova Sluzba
Dovoz tisku
Vinohradska 46
Praha 2
Postova Novinova Sluzba
Dovoz tlace
Leningradska 14
Bratislava

DENMARK

Ejnar Munksgaard
Nörregade 6
Kopenhagen

FINLAND

Akateeminen Kirjakauppa
Keskuskatu 2
Helsinki

FRANCE

Office International de Documentation et Libraire
48, rue Gay Lussac
Paris 5

GERMAN DEMOCRATIC REPUBLIC

Deutscher Buchexport und Import
Leninstraße 16.
Leipzig C. I.
Zeitungvertriebsamt
Clara Zetkin Straße 62.
Berlin N. W.

GERMAN FEDERAL REPUBLIC

Kunst und Wissen
Eich Bieber
Postfach 46.
7 Stuttgart S.

GREAT BRITAIN

Collet's' Subscription Dept.
44–45 Museum Street
London W. C. I.
Robert Maxwell and Co. Ltd.
Waynflete Bldg. The Plain
Oxford

HOLLAND

Swetz and Zeitlinger
Keizersgracht 471–487
Amsterdam C.
Martinus Nijhof
Lange Voorhout 9
The Hague

INDIA

Current Technical Literature
Co. Private Ltd.
Head Office:
India House OPP.
GPO Post Box 1374
Bombay I.

ITALY

Santo Vanasia
71 Via M. Macchi
Milano
Libreria Commissionaria Sanson
Via La Marmora 45
Firenze

JAPAN

Nauka Ltd.
2 Kanada-Zimbocho 2-ehome
Chiyoda-ku
Tokyo
Maruzen and Co. Ltd.
P.O. Box 605
Tokyo
Far Eastern Booksellers
Kanada P. O. Box 72
Tokyo

KOREA

Chulpanmul
Korejskoje Obschestvo po Exportu Importu Proizvedenij Pechati
Phenjan

NORWAY

Johan Grund Tanum
Karl Johansgatan 43
Oslo

POLAND

Export- und Import- Unternehmen RUCH
ul. Wilcza 46.
Warszawa

ROUMANIA

Cartimex
Str. Aristide Briand 14–18.
Bucuresti

SOVIET UNION

Mezhdunarodnaja Kniga
Moscow
G–200

SWEDEN

Almquist and Wiksell
Gamla Brogatan 26
Stockholm

USA

Stechert Hafner Inc.
31 East 10th Street
New York 3 N. Y
Walter J. Johnson
111 Fifth Avenue
New York 3 N. Y.

VIETNAM

Xunhasaba
Service d'Export et d'Import des Livres et Périodiques
19. Tran Quoc Toan
Hanoi

YUGOSLAVIA

Forum
Vojvode Misiva broj 1.
Novi Sad
Jugoslovenska Kniga
Terazije 27.
Beograd

10. VI. 1977. **Index: 26 748**

Scientiarum Mathematicarum Hungarica

TOMUS X.
FASC. 3–4.
1975

AKADÉMIAI KIADÓ, BUDAPEST

Studia Scientiarum Mathematicarum Hungarica

A Magyar Tudományos Akadémia matematikai folyóirata

Szerkesztőség: 1061 Budapest V., Reáltanoda u. 13—15.

Technikai szerkesztő: Deák E.

Kiadja az Akadémiai Kiadó, 1053 Budapest V., Alkotmány u. 21.

A *Studia Scientiarum Mathematicarum Hungarica* angol, német, francia vagy orosz nyelven közöl eredeti értekezéseket a matematika tárgyköréből. Félévenként jelenik meg, évi egy kötetben.

Előfizetési ára belföldre 120,— Ft, külföldre 165,— Ft. Megrendelhető a belföld számára az Akadémiai Kiadónál, a külföld számára pedig a Kultúra Könyv és Hírlap Külkereskedelmi Vállalatnál (1011 Budapest II., Fő u. 32.).

Cserekapcsolatok felvétele ügyében kérjük a MTA Matematikai Kutató Intézete Könyvtárához (1053 Budapest V., Reáltanoda u. 13—15) fordulni.

Közlésre szánt dolgozatokat kérjük két példányban a szerkesztőség címére küldeni.

Studia Scientiarum Mathematicarum Hungarica is a journal of the Hungarian Academy of Sciences publishing original papers on mathematics, in English, German, French or Russian.

It is published semiannually, making up one volume per year.

Editorial Office: 1053 Budapest V., Reáltanoda u. 13—15, Hungary.

Technical Editor: E. Deák

Subscription rate: $ 16.00 per volume. Orders may be placed with *Kultúra* Trading Co. for Books and Newspapers, Budapest 62, P. O. B. 149 or with its representatives abroad.

For establishing exchange relations please write to the Library of the Mathematical Institute (1053 Budapest V., Reáltanoda u. 13—15.)

Papers intended for publication should be sent to the Editor in 2 copies.

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 273—276.

NEAR-RINGS WHOSE GENERATOR IS A LIE IDEAL

by

K. K. SRIVASTAVA

1. Introduction

If the set of endomorphisms and anti-endomorphisms of a non-commutative group is considered then this set is a group under addition, a semigroup under multiplication and the left distributive law holds. This system has been named as "Near-Ring".

BLACKETT [2] and DESKINS [5] have considered the problem of semi-simplicity and the radical for a certain class of near-rings. FROHLICH [6] gave the ideal and representation theory for the class of distributively generated near-rings. Recently, the papers of CLAY [3], JACOBSON [8], LAXTON [10], MAXON [12], MALONE [11], and RAMAKOTAIAH [13] have appeared.

In the paper [14] we have extended the notion of annihilators to near-rings.

The main result of this paper is that if R is a simple distributively generated near-ring of characteristic $\neq 2$, then the generator of R which is a Lie ideal must either be R itself or is contained in the center of R.

2. Preliminaries

A distributively generated (d.g.) near-ring is a system with two binary operations, addition and multiplication, such that

(i) the elements of R form a group under one operation, say, addition,

(ii) the elements of R form a semigroup under the other operation, say, multiplication,

(iii) $x(y+z)=xy+xz$ for all $x, y, z \in R$,

(iv) $0x=0$ where 0 is the additive identity of R,

(v) R contains a multiplicative semigroup S whose elements generate R^+ and satisfy

$$(x+y)s = xs+ys \; \forall \; x, y \in R, \quad s \in S.$$

3.

(3.1) DEFINITION. We shall call a subset S of R a *Lie subnear-ring* of R if S is an additive subgroup such that for a, b in S, $ab-ba$ must also be in S. Similarly we define a subset S of R a Jordan subnear-ring of R if S is an additive subgroup such that for x, y in S, $xy+yx$ must also be in S.

(3.2) DEFINITION. Let S be a Lie subnear-ring of R. The additive subgroup $U \subset S$ is said to be a *Lie ideal* of S if whenever $u \in U$ and $s \in S$ then $[u, s] = us - su$ is in U.

We similarly define the concept of a Jordan subnear-ring of R.

4.

We begin with the following

(4.1) LEMMA. *If the generator U of R is a Jordan ideal of R then for all $a, b \in U$ and $x \in R$,*

$$(ab+ba)x - x(ab+ba) \in U.$$

PROOF. Since $a \in U$, for any $x \in R$

$$a(xb-bx)+(xb-bx)a \in U.$$

But

$$a(xb-bx)+(xb-bx)a =$$

$$= \{(ax-xa)b+b(ax-xa)\}+\{x(ab+ba)-(ab+ba)x\}.$$

The left hand side and the first term on the right hand side are in U, hence the lemma.

From this we now obtain:

(4.2) THEOREM. *Let R be a d.g. near-ring generated by U, in which $2x=0$ implies $x=0$ and suppose further that R has no non-zero nilpotent ideals. Then U contains a non-zero ideal of R.*

PROOF. Suppose $a, b \leftarrow U$. By the above lemma (4.1), for any $x \in R$, $xc - cx \in U$ where $c = ab+ba$. However since $c \in U$, $xc+cx \in U$. Adding, we get $2xc \in U$ for all x, hence for $y \in R$,

$$(2xc)y + y(2xc) \in U.$$

Since $2yxc \in U$ we obtain $2xcy \in U$, i.e. $2RcR \subset U$. Now $2RcR$ is an ideal of R, so we are done unless $2RcR = (0)$. If $2RcR = (0)$, by our assumption $RcR = (0)$ and so $(Rc)^2 = (0)$. Since R has no nilpotent ideals this forces $c = 0$, that is given $a, b \in U$ then $ab+ba = 0$.

Let $0 \neq a \in U$; then for $x \in R$, $b = ax+xa \in U$ hence $a(ax+xa)+(ax+xa)a = 0$. That is, $a^2x + xa^2 + 2axa = 0$.

Now for $a \in U$, $0 = aa+aa = 2a^2$ whence $a^2 = 0$. The top relation then reduces to $2axa = 0$ for all $x \in R$ and so $aRa = (0)$. But then $aR \neq (0)$ is a nilpotent right ideal of R, contrary to assumption. In other words, we have shown that U contains a non-zero ideal of R.

We now turn to the case of Lie ideals.

(4.3) LEMMA. *Let R has no non-zero nilpotent ideals in which $2x=0$ implies $x=0$. Suppose that the generator $U \neq (0)$ is both a Lie ideal and a subnear-ring of R. Then either $U \subset Z$, the center of R, or U contains a non-zero ideal of R.*

PROOF. Suppose U is not commutative. Then for some $x, y \in U$, $xy - yx = 0$. For any $r \in R$, $x(yr) - (yr)x \in U$, i.e.

$$(xy-yx)r + y(xr-rx) \in U.$$

The second member of this is in U since both y and $xr-rx\in U$ (since U is both a Lie ideal and a subnear-ring). The net result of all this is that

$$(xy-yx)R\subset U.$$

But then for $r, s\in R$,

$$\{(xy-yx)r\}s-s\{(xy-yx)r\}\in U$$

leading to $R(xy-yx)R\subset U$.

We now have shown that the ideal $R(xy-yx)R$ is in U. If $R(xy-yx)R=(0)$ then

$$\{R(xy-yx)\}^2=(0)$$

contrary to the assumption. We have shown that the result is correct if U as a subnear-ring of R is not commutative.

So, suppose that U is commutative; we want to show that it lies in the center of R. Given $a\in U$, $x\in R$ then $ax-xa\in U$ so commutes with a.

Now, for $x, y\in R$,

$$a(a(xy)-(xy)a)=(a(xy)-(xy)a)a.$$

Expanding $a(xy)-(xy)a$ as

$$(ax-xa)y+x(ay-ya)$$

and using that a commutes with this, with $ax-xa$ and with $ay-ya$ yields

$$2(ax-xa)(ay-ya)=0 \quad \forall x, y\in R.$$

Since $2r=0$ forces $r=0$ we obtain

$$(ax-xa)(ay-ya)=0.$$

In this we put $y=ax$, this result is

$$(ax-xa)R(ax-xa)=(0).$$

Since R has no nilpotent ideals we conclude that $ax-xa=0$ and so, a must be in the center of R.

(4.4) REMARK. In the proof of the above lemma we have also proved the following:

Let R be a near-ring (d.g.) having no non-zero nilpotent ideals in which $2x=0$ implies $x=0$. If $a\in R$ commutes with all $ax-xa$; $x\in R$ then a is in the center of R.

Lemma (4.3) immediately implies:

(4.5) THEOREM. *Let R be a simple d.g. near-ring of characteristic $\neq 2$. Then the generator U of R which is also a subnear-ring of R must either be R-itself or is contained in the center of R.*

(4.6) DEFINITION. Let $T(U)=\{x\in R: [x, R]\subset U\}$ where the generator U is a Lie ideal.

(4.7) LEMMA. *$T(U)$ is both a subnear-ring and a Lie ideal of R; moreover, $U\subset T(U)$.*

PROOF. Since U is a Lie ideal of R, $U\subset T(R)$; since $[T(U), R]\subset U\subset T(U)$; $T(U)$ must certainly be a Lie ideal of R.

Now suppose that $a, b \in T(U)$, $r \in R$. Then

$$(ab)r - r(ab) = \{a(br) - (br)a\} + \{b(ra) - (ra)b\},$$

so since $a, b \in T(U)$, the right side is in U. Therefore $[ab, R] \subset U$ that is $ab \in T(U)$.

We now prove the

(4.8) THEOREM. *Let R be a simple d.g. near-ring of characteristic $\neq 2$ and let the generator U be a Lie ideal of R. Then either $U \subset Z$, the center of R, or $U \supset [R, R]$.*

PROOF. By theorem (4.5) and lemma (4.7), since $T(U)$ is both a subnear-ring and a Lie ideal of R, either $T(U) \subset Z$ or $T(U) = R$. If $T(U) = R$ then by its very definition $[R, R] \subset U$; if $T(U)$ is contained in Z, since $U \subset T(U)$, we obtain $U \subset Z$.

COROLLARY. *If R is a non-commutative simple d.g. near-ring of characteristic $\neq 2$, then the subnear-ring generated by $[R, R]$ is R.*

PROOF. Any additive subgroup containing $[R, R]$ is, trivially, a Lie ideal of R. Hence the subnear-ring generated by $[R, R]$ is a Lie ideal thus, by theorem (4.5), it equals R or is in Z. If it is in Z then $[R, R] \subset Z$. Thus for $a \in R$ a commutes with all $ax - xa$, $x \in R$; by the sublemma we get that $a \in Z$, that is $R \subset Z$. Since we assumed R to be non-commutative, this is ruled out; hence the corollary.

Acknowledgement

I am thankful to Dr. H. M. Srivastava for his kind guidance during the preparation of this paper.

REFERENCES

[1] BAXTER, W.: Lie simplicity of a special class of associative rings, *Proc. Amer. Math. Soc.*, **7** (1958), 825—826.
[2] BLACKETT, D. W.: Simple and semi simple near-rings, *Proc. Amer. Math. Soc.* **4** (1953), 772—785.
[3] CLAY, J. R.: The near-rings on a finite cyclic group, *Amer. Math. Monthly* **71** (1964), 47—50.
[4] CLAY, J. R.: Imbedding of an arbitrary ring in a non trivial near-ring, *Amer. Math. Monthly* **74** (1967), 406—407.
[5] DESKINS, W. E.: A radical for near-rings, *Proc. Amer. Math. Soc.* **5** (1954), 825—827.
[6] FROHLICH, A.: Distributively generated near-rings (I) Ideal theory, *Proc. Lond. Math. Soc.* **8** (1958), 76—94, (II) Representation theory, *Proc. Lond. Math. Soc.*, **8** (1958), 95—108.
[7] HERSTEIN, I. N.: On the Lie and Jordan rings of a simple associative ring, *Amer. Jour. Math.* **77** (1955), 279—285.
[8] JACOBSON, N.: Structure of Rings, *Amer. Math. Soc. Coll. Publ. Vol. XXXVII*, 1956.
[9] JACOBSON, R. A.: The structure of a near-ring on a group of prime order, *Amer. Math. Monthly* **73** (1966), 59—61.
[10] LAXTON, R. R.: Prime ideal and the ideal radical of a distributively generated near-ring, *Math. Zeit* **83** (1964), 8—17.
[11] MALONE, J. J.: Near-rings with trivial multiplications, *Amer. Math. Monthly*, **74** (1967), 1111—1112.
[12] MAXON, C. J.: On finite near-rings with identity. *Amer. Math. Monthly*, **74** (1967), 1228—1230.
[13] RAMAKOTAIAH, D.: Radicals for near-rings, *Math. Zeit.*, **97** (1967) 45—56.
[14] SRIVASTAVA, K. K.: Annihilators in near-rings, Communicated for publications.

Department of Mathematics, University of Lucknow, Lucknow, India
(Received September 19, 1972)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 277—286.

FURTHER STABILITY CONDITIONS FOR CONTROLLABLY PERIODIC PERTURBED SOLUTIONS

by

H. M. EL OWAIDY

This paper is extension to the work of M. FARKAS (cf. [2], [3]) where a perturbed system containing a small parameter is considered. It was assumed that the unperturbed system is autonomous and has a unique periodic solution, the perturbation is nonautonomous and is controllably periodic. M. Farkas gave a criterion for the stability of its unique periodic solution. Here another condition will be given if that due to M. Farkas fails to hold. An effective form is given to this stability criterion by using Poincaré's method.

1. Notations

Let us consider the system:

$$dx/dt = f(x)+\mu g\left(\frac{t}{\tau}, x, \mu, \tau\right), \tag{1.1}$$

where x, f, g are n-dimensional vectors, and t, μ and τ are scalars. The function

$$F\left(\frac{t}{\tau}, x, \mu, \tau\right) = f(x)+\mu g\left(\frac{t}{\tau}x, \mu, \tau\right) \tag{1.2}$$

is analytic in the region $I_t\times\Omega\times I_\mu\times I_\tau$ where $I_t=\{t: -\infty<t<\infty\}$, Ω is an open and connected region in the n-dimensional space of x, $I_\mu=\{\mu: |\mu|<\alpha\}$ for some $\alpha>0$, and $I_\tau=\{\tau: |\tau-\tau_0|<\beta,\ 0<\beta<\tau_0\}$; τ_0 will be defined later.

We assume also that the function F is periodic in t with period τ.

It is assumed that the unperturbed system

$$dx/dt = f(x) \tag{1.3}$$

has a unique periodic solution $p(t)$ of period $\tau_0>0$. The first variational system of (1.3) corresponding to $p(t)$ is

$$dy/dt = f'_x(p(t))y. \tag{1.4}$$

The periodic function $\dot{p}(t)$ is a solution of (1.4). The fundamental matrix solution of (1.4) is denoted by $Y_0(t)$. It is assumed that one is a simple characteristic multiplier of the system (1.4). M. FARKAS (cf. [2]) proved that for each small enough value of $|\mu|$ and the parameter $|\vartheta|$ (defined there), the system (1.1) has a unique periodic solution $\varphi(t;\mu,\vartheta)$ with unique period $\tau(\mu,\vartheta)$ provided that $\tau(\mu,\vartheta)$ is substituted into (1.1) for τ.

The first variational system of (1.1) corresponding to the solution $\varphi(t; \mu, \vartheta)$ is

$$dy/dt = \left[f'_x(x)+\mu g'_x\left(\frac{t}{\tau}, x, \mu, \tau\right)\right]_{x=\varphi(t;\mu,\vartheta)} y. \tag{1.5}$$

The fundamental matrix solution of (1.5) is denoted by $Y(t; \mu, \vartheta)$ for which $Y(0; \mu, \vartheta)=U$, U the unit matrix, holds. It is clear that for $\mu=\vartheta=0$ we have $Y(t; 0, 0)=Y_0(t)$ which is the fundamental matrix solution of (1.4). Also for $\mu=0$ and all ϑ we have

$$\varphi\{t; 0, \vartheta) = p(t-\vartheta), \tag{1.6}$$

and $Y_0(t-\vartheta)$ is the fundamental matrix solution of

$$\dot{y} = f'_x(p(t-\vartheta))y. \tag{1.7}$$

Obviously

$$Y(t; 0, \vartheta) = Y_0(t-\vartheta)Y_0^{-1}(-\vartheta). \tag{1.8}$$

Let $\lambda(\mu, \vartheta)$ be the characteristic multiplier of the system (1.5) (i.e. eigenvalue of the matrix $C(\mu, \vartheta)=Y(\tau(\mu, \vartheta); \mu, \vartheta))$ for which $\lambda(0, 0)=1$ and $\lambda'_\mu(0, 0)$ the derivative of $\lambda(\mu, \vartheta)$ with respect to μ evaluated at $\mu=\vartheta=0$. Assuming that the remaining $n-1$ characteristic multipliers of system (1.4) are in modulus less than one, M. FARKAS (cf. [2]) proved that the solution $\varphi(t; \mu, \vartheta)$ is asymptotically stable if $|\mu|$ and $|\vartheta|$ are small enough and μ satisfies the inequality

$$\mu\lambda'_\mu(0, 0) < 0. \tag{1.9}$$

2. Stability theorem

If $\lambda'_\mu(0, 0)=0$ in M. FARKAS' theorem (cf. [2]) there exists no μ satisfying condition (1.9). Here we shall consider the periodic solution of the system (1.1) corresponding to the value $\vartheta=0$. If $\lambda'_\mu(0, 0)=0$ and $\lambda''_{\mu\mu}(0, 0)\neq 0$, we have asymptotical stability or unstability for all sufficiently small $|\mu|$ according as the point $\mu=0$ is a maximum or a minimum point of the function $\lambda(\mu, 0)$. The following theorem holds.

THEOREM 1. *Under the assumptions stated above, if the remaining* $n-1$ *characteristic multipliers of the system* (1.4) *are in modulus less than one,* $\lambda'_\mu(0, 0)=0$ *and*

$$\lambda''_{\mu\mu}(0, 0) < 0, \tag{2.1}$$

there exists $\varrho>0$ *such that for*

$$0 < |\mu'| < \varrho, \quad \vartheta = 0 \tag{2.2}$$

the periodic solution $\varphi(t; \mu, 0)$ *of the perturbed system* (1.1) *is asymptotically stable.*

PROOF. Since the remaining characteristic multipliers of the matrix are in modulus less than one, $C(\mu, 0)$ is an analytic function of μ, and if $\lambda'_\mu(0, 0)=0$ and $\lambda''_{\mu\mu}(0, 0)<0$ then $\mu=0$ is a maximum point of $\lambda(\mu, 0)$. Since $\lambda(\mu, 0)$ is an analytic function in a neighbourhood of $(0, 0)$ with $\lambda(0, 0)=1$, therefore

$$\lambda(\mu, 0) = 1+\frac{1}{2}\mu^2\lambda''_{\mu\mu}(0, 0)+o(\mu^2). \tag{2.3}$$

It follows that all the characteristic multipliers are in modulus less than one and thus the periodic solution $\varphi(t; \mu, 0)$ of the system (1.1) is asymptotically stable for all sufficiently small non-zero $|\mu|$, and by that the theorem is proved.

3. Expression for the stability criterion

Now we shall use Poincaré's method to obtain an effective form for the stability condition given in section 2.

Using the substitution

$$t = \vartheta+s\tau(\mu, \vartheta) \tag{3.1}$$

the system (1.1) and its periodic solution reduces to

$$dx/ds = \tau(\mu, \vartheta)\left[f(x)+\mu g\left(s+\frac{\vartheta}{\tau(\mu, \vartheta)}, x, \mu, \tau(\mu, \vartheta)\right)\right], \tag{3.2}$$

and

$$\psi(s; \mu, \vartheta) = \varphi(\vartheta+s\tau(\mu; \vartheta); \mu, \vartheta), \tag{3.3}$$

where $\psi(s; \mu, \vartheta)$ is obviously periodic in s with period one. Expanding the solution $\psi(s; \mu, \vartheta)$ of (3.2) and the function $\tau(\mu, \vartheta)$ for fixed ϑ by powers of μ, we obtain

$$\psi(s; \mu, \vartheta) = \sum_{k=0}^{\infty}\mu^k\psi^k(s; \vartheta), \tag{3.4}$$

$$\tau(\mu; \vartheta) = \sum_{k=0}^{\infty}\mu^k\tau_k(\vartheta), \tag{3.5}$$

and we have

$$\psi^0(s; \vartheta) = p(s\tau_0), \quad \tau_0(\vartheta) = \tau_0. \tag{3.6}$$

It is clear that the vectors $\psi^k(s; \vartheta)$, $k=1, 2, \ldots$ are periodic functions with period one. Substituting the expansions (3.4) and (3.5) into (3.2) we have the identity:

$$\sum_{k=0}^{\infty}\mu^k\frac{d\psi^k}{ds}(s, \vartheta) = \tag{3.7}$$

$$= \sum_{k=0}^{\infty}\mu^k\tau_k(\vartheta)\left[f(\psi(s; \mu, \vartheta))+\mu g\left(s+\frac{\vartheta}{\tau(\mu, \vartheta)}, \psi(s; \mu, \vartheta), \mu, \tau(\mu, \vartheta)\right)\right].$$

Equating the corresponding coefficients of μ on both sides we have:

$$\mu^0\colon \frac{d\psi^0}{ds}(s,\vartheta)=\tau_0 f\big(p(s\tau_0)\big), \tag{3.8}$$

$$\mu^1\colon \frac{d\psi^1}{ds}(s,\vartheta)=$$

$$=\tau_0 f'_x\big(p(s\tau_0)\big)\psi'(s,\vartheta)+\tau_0 g\left(s+\frac{\vartheta}{\tau_0},\, p(s\tau_0),\, 0,\, \tau_0\right)+\tau_1(\vartheta)f\big(p(s\tau_0)\big),$$

$$\mu^2\colon \frac{d\psi^2}{ds}(s,\vartheta)=\tau_0 f'_x\big(p(s\tau_0)\big)\psi^2(s,\vartheta)+$$

$$+\frac{\tau_0}{2}f''_{xx}\big(p(s\tau_0)\big)\psi'^2(s,\vartheta)+\tau_1(\vartheta)f'_x\big(p(s\tau_0)\big)\psi'(s,\vartheta)+$$

$$+\tau_2(\vartheta)f\big(p(s\tau_0)\big)+\tau_0 g'_x\left(s+\frac{\vartheta}{\tau_0},\, p(s\tau_0),\, 0,\, \tau_0\right)\psi'(s,\vartheta)+$$

$$+\tau_0 g'_\mu\left(s+\frac{\vartheta}{\tau_0},\, p(s\tau_0),\, 0,\, \tau_0\right)+\tau_0\tau_1(\vartheta)g'_\tau\left(s+\frac{\vartheta}{\tau_0},\, p(s\tau_0),\, 0,\, \tau_0\right)-$$

$$-\frac{\vartheta\tau_1}{\tau_0}g'_t\left(s+\frac{\vartheta}{\tau_0},\, p(s\tau_0),\, 0,\, \tau_0\right)+\tau_1 g\left(s+\frac{\vartheta}{\tau_0},\, p(s\tau_0),\, 0,\, \tau_0\right),$$

and in general

$$\mu^m\colon \frac{d\psi^m}{ds}(s,\vartheta)=\tau_0 f'_x\big(p(s\tau_0)\big)\psi^m(s,\vartheta)+\tau_m(\vartheta)f\big(p(s\tau_0)\big)+F^m, \tag{3.9}$$

where F^m depends on ψ^j, $j<m$ and τ_j, $j<m$.

Apart from these equations it is also the case that

$$\psi_1(0,\vartheta)=0 \quad \text{i.e. thus} \quad \psi_1^j(0,\vartheta)=0, \quad j=1,2,\dots .$$

Then the expressions for $\psi(s,\mu,\vartheta)$ and $\tau(\mu,\vartheta)$ given by (3.4) and (3.5) are solutions of the system above. It can be proved that the system of equations given above determines ψ^j and τ_j uniquely, subject to the conditions $\psi_1^j(0,\vartheta)=0$ and $\psi^j(s,\vartheta)$ periodic in s of period one.

In formula (3.8) f''_{xx} denotes the third order tensor with components $f''_{ix_kx_l}$ (the second partial derivatives of the components of the vector f), $i,l,k=1,2,\dots,n$, and $f''_{xx}\psi'^2$ stands for the vector whose i^{th} component is

$$\sum_{k,l=1}^{n} f''_{ix_kx_l}\psi_k^1\psi_l^1, \quad i=1,2,\dots,n \tag{3.10}$$

where ψ_k^1 and ψ_l^1 are the k^{th} and l^{th} components of the vector ψ^1 respectively. Similar meanings are attached to higher derivatives and this notation is used extensively

in the remainder of this paper. Since the characteristic matrix $C(\mu, \vartheta)$ of (1.5) corresponding to its fundamental matrix solution $Y(t; \mu, \vartheta)$ is

$$C(\mu, \vartheta) = Y(\tau(\mu, \vartheta)); \mu, \vartheta), \tag{3.11}$$

and since it is analytic, it can be expanded for fixed ϑ by powers of μ i.e.

$$C(\mu, \vartheta) = C_0(\vartheta) + \mu C_1(\vartheta) + \mu^2 C_2(\vartheta) + \mu^3 R(\mu, \vartheta) \tag{3.12}$$

where R is analytic, $C_0(0) = C(0, 0) = Y(\tau_0, 0, 0) = Y_0(\tau_0)$.

An expression for $C_1(\vartheta)$ was given by M. FARKAS (cf. [3]). Here we shall give an expression for $C_2(\vartheta)$ provided that the fundamental matrix of (1.4), $\psi^1(s, \vartheta)$, $\psi^2(s, \vartheta)$, $\tau_1(\vartheta)$ and $\tau_2(\vartheta)$ (in the expressions (3.4) and (3.5) respectively) are known. Differentiating the matrix C twice with respect to μ, and evaluating the result at $\mu = 0$ i.e.

$$\begin{aligned} C_2(\vartheta) = C''_{\mu\mu}(\mu, \vartheta) = \ddot{Y}[\tau(\mu, \vartheta), \mu, \vartheta]\tau'^2_\mu(\mu, \vartheta) + \\ + 2\dot{Y}'_\mu[\tau(\mu, \vartheta); \mu, \vartheta]\tau'_\mu(\mu, \vartheta) + \\ + \dot{Y}[\tau(\mu, \vartheta); \mu, \vartheta]\tau''_{\mu\mu}(\mu, \vartheta) + \ddot{Y}''_{\mu\mu}[\tau(\mu, \vartheta); \mu, \vartheta]|_{\mu=0} = \\ = \ddot{Y}(\tau_0, 0, \vartheta)\tau^2(\vartheta) + 2\dot{Y}'_\mu(\tau_0, 0, \vartheta)\tau_1(\vartheta) + \\ + \dot{Y}(\tau_0, 0, \vartheta)\tau_2(\vartheta) + Y''_{\mu\mu}(\tau_0, 0, \vartheta); \end{aligned}$$

for $\vartheta = 0$ we have

$$C_2(0) = \ddot{Y}(\tau_0, 0, 0)\tau_1^2(0) + 2\dot{Y}'_\mu(\tau_0, 0, 0) + \dot{Y}(\tau_0, 0, 0)\tau_2(0) + Y''_{\mu\mu}(\tau_0, 0, 0). \tag{3.13}$$

We shall give now an explicit form for each term in (3.13). Since (1.8) is a solution of (1.7) therefore

$$\dot{Y}(t; 0, \vartheta) = f'_x(p(t-\vartheta))Y_0(t-\vartheta)Y_0^{-1}(-\vartheta), \tag{3.14}$$

hence

$$\dot{Y}(\tau_0; 0, 0) = \dot{Y}_0(\tau_0) = f'_x(p(\tau_0)) \cdot Y_0(\tau_0). \tag{3.15}$$

Since the fundamental matrix solution $Y(t, \mu, \vartheta)$ satisfies the system (1.5) thus we can write

$$\dot{Y}(t; \mu, \vartheta) = \left[f'_x(x) + \mu g'_x\left(\frac{t}{\tau}, x, \mu, \tau\right)\right]_{x=\varphi(t;\mu,\vartheta)} Y(t; \mu, \vartheta). \tag{3.16}$$

The first term in (3.13) can be obtained by differentiating (3.16) with respect to t at $\mu = 0$ and $\vartheta = 0$; i.e.

$$\ddot{Y}(t; 0, 0) = [f''_{xx}(p(t))\dot{p}(t) + f'^2_x(p(t))]Y_0(t).$$

For $t = \tau_0$ we get

$$\ddot{Y}(\tau_0, 0, 0) = \ddot{Y}_0(\tau_0) = [f''_{xx}(p(\tau_0))\dot{p}(\tau_0) + f'^2_x(p(\tau_0))]Y_0(\tau_0). \tag{3.17}$$

Also the third term in (3.13) can be obtained by differentiating (3.16) with respect to μ at $\mu=0$, $\vartheta=0$ and $t=\tau_0$, thus we have

$$\dot{Y}'_\mu(\tau_0; 0, 0) =$$

$$= \left[f''_{xx}(p(\tau_0))\varphi'_\mu(\tau_0; 0, 0)+g'_x(1, p(\tau_0), 0, \tau_0)\right] Y_0(\tau_0)+f'_x(p(\tau_0))Y'_\mu(\tau_0; 0, 0) \tag{3.18}$$

where

$$Y'_\mu(\tau_0; 0, 0) = Y_0(\tau_0)\int_0^{\tau_0} Y_0^{-1}(t)\,B(t, 0)\,dt \tag{3.19}$$

and

$$B(t, 0) = \left[f''_{xx}(p(t))\varphi'_\mu(t; 0, 0)+g'_x\left(\frac{t}{\tau_0}, p(t), 0, \tau_0\right)\right] Y_0(t). \tag{3.20}$$

It is clear that

$$\varphi'_\mu(t; \mu, \vartheta) = \left\{-\dot{p}\left(\frac{t-\vartheta}{\tau(\mu, \vartheta)}\right)\cdot(\tau-\vartheta)(\tau_1+2\mu\tau_2+3\mu^2\tau_3+\ldots)/[\tau(\mu, \vartheta)]^2\right\}+ \tag{3.21}$$

$$+\psi'\left(\frac{\tau-\vartheta}{\tau(\mu, \vartheta)}, \vartheta\right)-\mu\left\{\dot{\psi}^1\left(\frac{\tau-\vartheta}{\tau(\mu, \vartheta)}, \vartheta\right)(t-\vartheta)(\tau_1+2\mu\tau_2+3\mu^2\tau_3+\ldots)/[\tau(\mu, \vartheta)]^2\right\}+$$

$$+2\mu\psi^2\left(\frac{t-\vartheta}{\tau(\mu, \vartheta)}, \vartheta\right)-\mu^2\left\{\dot{\psi}^2\left(\frac{t-\vartheta}{\tau(\mu, \vartheta)}, \vartheta\right)(t-\vartheta)(\tau_1+2\mu\tau_2+3\mu^2\tau_3+\ldots)/\tau^2\right\}$$

where $\tau=\tau(\mu, \vartheta)$, thus

$$\varphi'_\mu(t; 0, 0) = \psi'\left(\frac{t}{\tau_0}, 0\right)-\dot{p}(t)\frac{t}{\tau_0}\tau_1(0). \tag{3.22}$$

To obtain the last term in expansion (3.13), differentiating (3.16) with respect to μ we get

$$\dot{Y}'_\mu(t; \mu, \vartheta) =$$

$$= \left[f'_x\left(\varphi(t; \mu, \vartheta)\right)+\mu g'_x\left(\frac{t}{\tau}, \varphi(t; \mu, \vartheta); \mu, \tau\right)\right] Y'_\mu(t; \mu, \vartheta)+N(t; \mu, \vartheta) \tag{3.23}$$

where

$$N(t; \mu, \vartheta) = \tag{3.24}$$

$$= \left[f''_{xx}(\varphi(t; \mu, \vartheta))\varphi'_\mu(t; \mu, \vartheta)+g'_x\left(\frac{t}{\tau}, \varphi(t; \mu, \vartheta); \mu, \tau\right)+\right.$$

$$+\mu g''_{xx}\left(\frac{t}{\tau}, \varphi(t; \mu, \vartheta); \mu, \tau\right)\varphi'_\mu(t; \mu, \vartheta)+\mu g''_{x\mu}\left(\frac{t}{\tau}, \varphi(t; \mu, \vartheta); \mu, \tau\right)+$$

$$+\mu g''_{x\tau}\left(\frac{t}{\tau}, \varphi(t; \mu, \vartheta); \mu, \tau\right)\tau'_\mu(\mu, \vartheta)-$$

$$\left.-\mu\left[\frac{t\tau'_\mu}{\tau^2}\,g''_{x\tau}\left(\frac{t}{\tau}, \varphi(t; \mu, \vartheta); \mu, \tau\right)\right]\right] Y(t; \mu, \vartheta).$$

Thus $Y'_\mu(t; \mu, \vartheta)$ satisfies the nonhomogeneous matrix equation (3.23).

Since $Y(t; \mu, \vartheta)$ is the fundamental matrix solution of the homogeneous system corresponding to (3.23), we get by the method of the variation of constants that

$$Y'_\mu(t; \mu, \vartheta) = Y(t; \mu, \vartheta) \int_0^t Y^{-1}(s; \mu, \vartheta) N(s; \mu, \vartheta)\, ds.$$

Differentiating this formula with respect to μ we get:

$$Y''_{\mu\mu}(t; \mu, \vartheta) = Y(t; \mu, \vartheta) \int_0^t \frac{d}{d\mu}[Y^{-1}(s; \mu, \vartheta) N(s; \mu, \vartheta)]\, ds +$$

$$+ Y'_\mu(t; \mu, \vartheta) \int_0^t Y^{-1}(s; \mu, \vartheta) N(s; \mu, \vartheta)\, ds.$$

Thus

$$Y''_{\mu\mu}(\tau_0; 0, 0) = Y_0(\tau_0) \int_0^{\tau_0} [Y_\mu^{-1\prime}(t; 0, 0) B(t, 0) + Y^{-1}(t) N'_\mu(t; 0, 0)]\, dt +$$

(3.25)

$$+ Y_0(\tau_0) \int_0^{\tau_0} Y_0^{-1} B(t, 0)\, dt \int_0^{\tau_0} Y_0^{-1}(s)\, B(s, 0)\, ds$$

where

(3.26) $$N(t; 0, 0) = B(t, 0) \quad \text{which was given by (3.20),}$$

$$Y_\mu^{-1\prime}(t; 0, 0) = -Y^{-1}(t; 0, 0) Y'_\mu(t; 0, 0) Y^{-1}(t, 0, 0) =$$

$$= -Y_0^{-1}(\tau_0) Y'_\mu(t; 0, 0) Y_0^{-1}(t),$$

and $N'_\mu(t; 0, 0)$ is obtained by differentiating (3.24) with respect to μ and evaluating at $\mu=0$ and $\vartheta=0$. Thus we get

$$N'_\mu(t; 0, 0) = [f'''_{xxx}(p(t)) \varphi'^2_\mu(t; 0, 0) + f''_{xx}(p(t)) \varphi''_{\mu\mu}(t; 0, 0) +$$

$$+ 2g''_{xx}\left(\frac{t}{\tau_0}, p(t), 0, \tau_0\right) \varphi'_\mu(t; 0, 0) + 2g''_{x\mu}\left(\frac{t}{\tau_0}, p(t), 0, \tau_0\right) -$$

$$- \frac{\tau_1(0)}{\tau_0^2} t g''_{xt}\left(\frac{t}{\tau_0}, p(t), 0, \tau_0\right) + g''_{x\tau}\left(\frac{t}{\tau_0}, p(t), 0, \tau_0\right)\Big] Y_0(t) +$$

$$+ \left[f''_{xx}(p(t)) \cdot \varphi'_\mu(t; 0, 0) + g'_x\left(\frac{t}{\tau_0}, p(t), 0, \tau_0\right)\right] Y'_\mu(t; 0, 0)$$

where $f'''_{xxx}(p(t))\, \varphi'^2_\mu(t; 0, 0)$ denotes the $n \times n$ matrix with elements

$$\sum_{l, m=1}^{n} f'''_{ix_k x_l x_m} \varphi'_{l\mu} \varphi'_{m\mu}, \quad i, k = 1, 2, \ldots, n,$$

where $\varphi'_{l\mu}$ and $\varphi'_{m\mu}$ are the l^{th} and m^{th} components of the vector φ'_μ, and

$$Y'_\mu(t; 0, 0) = Y_0(\tau_0) \int_0^t Y^{-1}(s; 0, 0) B(s, 0)\, ds.$$

$\varphi''_{\mu\mu}$ can be obtained by differentiating (3.21) with respect to μ, and evaluating at $\mu=0$ and $\vartheta=0$; thus we get

$$\varphi''_{\mu\mu}(t;0,0)=\ddot{p}(t)\frac{t^2\tau_1^2(0)}{\tau_0^2}-2\dot{p}(t)t\left(\frac{\tau_0\tau_2(0)-\tau_1^2(0)}{\tau_0^2}\right)- \tag{3.27}$$

$$-2\dot{\psi}'\left(\frac{t}{\tau_0},0\right)t\frac{\tau_1(0)}{\tau_0^2}+2\psi^2\left(\frac{t}{\tau_0},0\right).$$

The second term in the expression (3.12) was given by M. FARKAS (cf. [3]) which is

$$C_1(0)=\tau_1(0)f'_x(p(\tau_0))Y_0(\tau_0)+Y_0(\tau_0)\int_0^{\tau_0}Y^{-1}(t)B(t,0)\,dt. \tag{3.28}$$

Now the first three terms in (3.12) were determined, so the second derivative of the eigenvalue $\lambda(\mu,0)$ (for which $\lambda(0,0)=1$ and $\lambda'_\mu(0,0)=0$) with respect to μ at $\mu=0$ will be given in an explicit form. Let the characteristic polynomial of $C(\mu;\vartheta)$ be denoted by

$$(-1)^n d(\lambda;\mu,\vartheta)=\det\left[C(\mu,\vartheta)-\lambda U\right],$$

$U=$the unit matrix. Now we can state and prove the following

THEOREM 2. *Under the conditions of theorem* 1:

$$\lambda''_{\mu\mu}(0,0)=\frac{(-1)^{n+1}\cdot 2}{d'_\lambda(1,0,0)}\left\{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\det\begin{vmatrix} c_1^0-u_1\\ c_2^0-u_2\\ \cdots\cdots\cdots\\ c_{i-1}^0-u_{i-1}\\ c_i^1\\ c_{i+1}^0-u_{i+1}\\ \cdots\cdots\cdots\\ \cdots\cdots\cdots\\ c_{j-1}^0-u_{j-1}\\ c_j^1\\ c_{j+1}^0-u_{j+1}\\ \cdots\cdots\cdots\\ c_n^0-u_n\end{vmatrix}+\sum_{i=1}^{n}\det\begin{vmatrix} c_1^0-u_1\\ c_2^0-u_2\\ \cdots\cdots\cdots\\ \cdots\cdots\cdots\\ c_{i-1}^0-u_{i-1}\\ c_i^2\\ c_{i+1}^0-u_{i+1}\\ \cdots\cdots\cdots\\ \cdots\cdots\cdots\\ c_n^0-u_n\end{vmatrix}\right\} \tag{3.29}$$

where c_i^0, c_i^1, c_i^2 and u_i are the i^{th} row vectors of the $n\times n$ matrices $C_0(0)$, $C_1(0)$, $C_2(0)$ and U, respectively.

PROOF. We have

$$(-1)^n d(\lambda;\mu,0)=\det\left|C(\mu,0)-\lambda U\right|=$$

$$=\det\begin{vmatrix} c_1^0+\mu c_1^1+\mu^2 c_1^2+\mu^3 r_1-\lambda u_1\\ c_2^0+\mu c_2^1+\mu^2 c_2^2+\mu^3 r_2-\lambda u_2\\ \cdots\cdots\cdots\cdots\cdots\cdots\\ \cdots\cdots\cdots\cdots\cdots\cdots\\ c_n^0+\mu c_n^1+\mu^2 c_n^2+\mu^3 r_n^3-\lambda u_n\end{vmatrix}.$$

Here r_i is the i^{th} row vector of R. Differentiating this determinant twice with respect to μ, evaluating the resulting determinant at $\lambda=1$, $\mu=0$, and summing up we get

$$(3.30)\quad (-1)^n d''_{\mu\mu}(1,0,0) = 2\left\{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \det \begin{vmatrix} & & & & & & \\ c_1^0-u_1 & & & & & & \\ c_2^0-u_2 & & & & & & \\ \cdots\cdots\cdots & & & & & & \\ \cdots\cdots\cdots & & & & & & \\ c_{i-1}^0-u_{i-1} & & & & & & \\ c_i^1 & & & & & & \\ c_{i+1}^0-u_{i+1} & & & & & & \\ \cdots\cdots\cdots & & & & & & \\ \cdots\cdots\cdots & & & & & & \\ c_{j-1}^0-u_{j-1} & & & & & & \\ c_j^1 & & & & & & \\ c_{j+1}^0-u_{j+1} & & & & & & \\ \cdots\cdots\cdots & & & & & & \\ c_n^0-u_n & & & & & & \end{vmatrix} + \sum_{i=1}^{n} \det \begin{vmatrix} & & & & & & \\ c_1^0-u_1 & & & & & & \\ c_2^0-u_2 & & & & & & \\ \cdots\cdots\cdots & & & & & & \\ c_{i-1}^0-u_{i-1} & & & & & & \\ c_i^2 & & & & & & \\ c_{i+1}^0-u_{i+1} & & & & & & \\ \cdots\cdots\cdots & & & & & & \\ \cdots\cdots\cdots & & & & & & \\ \cdots\cdots\cdots & & & & & & \\ c_n^0-u_n & & & & & & \end{vmatrix}\right\}.$$

Since 1 is a simple root of the polynomial $d(\lambda, 0, 0)$ i.e. $d(1,0,0)=0$ and $d'_\lambda(1,0,0)\neq 0$, hence $\lambda(\mu, 0)$ can be expressed from equation $d(\lambda, \mu, 0)=0$. Since $\lambda(\mu, 0)$ is an analytic function in a neighbourhood of $(0,0)$ such that $\lambda(0,0)=1$, therefore

$$\lambda(\mu,0) = 1+\mu\lambda'_\mu(0,0)+\frac{\mu^2}{2}\lambda''_{\mu\mu}(0,0)+o(\mu^2),$$

$\lambda'_\mu(\mu, 0)$ is obtained from the relation $d(\lambda, \mu, 0)=0$ by differentiation i.e.

$$\lambda'_\mu(\mu,0) = -d'_\mu(\lambda,\mu,0)/d'_\lambda(\lambda,\mu,0)$$

and its second derivative satisfies the relation

$$\lambda''_{\mu\mu}(\mu,0) = [-d''_{\mu\mu}(\lambda;\mu,0)d'^2_\lambda(\lambda;\mu,0)+$$

$$+2d''_{\mu\lambda}(\lambda;\mu,0)d'_\mu(\lambda;\mu,9)d'_\lambda(\lambda;\mu,0)-$$

$$-d''_{\lambda\lambda}(\lambda;\mu,0)d'^2_\mu(\lambda;\mu,0)]/d'^3_\lambda(\lambda;\mu,0).$$

If $\lambda'_\mu(0,0)=0$ i.e. $d'_\mu(\lambda,0)=0$, then

$$(3.31)\qquad \lambda''_{\mu\mu}(0,0) = -d''_{\mu\mu}(1,0,0)/d'_\lambda(1,0,0).$$

Substituting from (3.30) in this expression we get the result (3.29) which was to be proved.

Acknowledgement. The author wishes to thank Prof. M. Farkas for his continual encouragement, tolerance, constructive criticism and useful comments on the manuscript.

REFERENCES

[1] CODDINGTON, E. and LEVINSON, N.: *Theory of ordinary differential equations*, McGraw-Hill, New York, London, 1955.
[2] FARKAS, M.: Controllably periodic perturbations of autonomous systems, *Acta Math. Acad. Sci. Hungar.* **22** (1971), 337—348.
[3] FARKAS, M.: Determination of controllably periodic perturbed solutions by Poincaré's method, *Studia Sci. Math. Hungar.* **7** (1972), 257—266.
[4] HALE, J. K.: *Oscillations in nonlinear system*, McGraw-Hill, Advanced mathematics with applications, New York, 1963.
[5] LEFSCHETZ, S.: *Differential Equations: Geometric Theory*, Int. Pub. Inc., New York, 1957.

Math. Dept. Faculty of Science, Al Azhar University, Nasr City, Cairo, Egypt.

(Received March 2, 1973)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 287—296.

ON PERTURBATIONS OF LIÉNARD'S EQUATION

by

H. M. EL OWAIDY

We shall consider the periodic solution of Liénard's equation under small periodic perturbations. It is assumed that the perturbation is "controllably periodic" i.e. its period can be chosen appropriately. Also it is assumed that the perturbed Liénard's equation has a unique periodic solution. Poincaré's method is worked out for the determination of this solution up to the second approximation. Also a sufficient criterion will be given for the asymptotic stability of the perturbed periodic solution. The results are based on those of papers [3], [4] and [5], and are close to and generalizations of W. S. LOUD's result [8].

1. Notations

It is well known (cf. [7]) that Liénard's differential eq.

$$\ddot{u}+h(u)\dot{u}+k(u)=0 \quad (\cdot = d/dt) \tag{1.1}$$

has a unique periodic solution $u_0(t)$ with (least positive) period τ_0, (if $h(u)$ is an even continuous function, $k(u)$ is an odd differentiable function and the coefficients $h(u)$ and $k(u)$ of (1.1) are defined in $(-\infty, \infty)$, $H(u)=\int_0^u h(s)ds$, $K(u)=\int_0^u k(s)ds$, $H(u)$ has a single positive zero u_1 with $H(u)<0$ for $0<u<u_1$, $H(u)>0$ for $u>u_1$, $H(u)$ is monotonic increasing for $u>u_1$ and $H(u)\to\infty$, $K(u)\to\infty$ as $u\to+\infty$. Without loss of generality we select as origin of time a point for which

$$u_0(0)=0, \quad \dot{u}_0(0)=a>0. \tag{1.2}$$

Introducing the notations

$$x_1=u(t), \quad x_2=\dot{u}(t)+H(u). \tag{1.3}$$

Eq. (1.1) is equivalent to the system

$$\dot{x}=f(x) \tag{1.4}$$

where $x=\mathrm{col}\,[x_1, x_2]$ and

$$f(x)=\mathrm{col}\,[x_2-H(x_1), -k(x_1)]. \tag{1.5}$$

The periodic solution of (1.4) corresponding to $u_0(t)$ is

$$p(t)=\mathrm{col}\,[u_0(t), \dot{u}_0(t)+H(u_0(t))].$$

The first variational system of (1.4) is

$$\dot{y} = f'_x(p(t))\,y \tag{1.6}$$

where

$$f'_x(p(t)) = \begin{bmatrix} -h(u_0(t)) & 1 \\ -k'(u_0(t)) & 0 \end{bmatrix}. \tag{1.7}$$

The system (1.6) has the periodic solution $\dot{p}(t)$ of period τ_0 which is given by:

$$\dot{p}(t) = \operatorname{col}\left[\dot{u}_0(t),\ -k(u_0(t))\right]. \tag{1.8}$$

The second solution of the system (1.6) that is linearly independent from the first and satisfies

$$v_0(0) = 0, \quad \dot{v}_0(0) = \frac{1}{a} \tag{1.9}$$

is denoted by $v_0(t)$.

Let $Y_0(t)$ denote the fundamental matrix solution of the system (1.6) for which $Y_0(0) = U$, the unit matrix,

$$Y_0(t) = \begin{bmatrix} \dfrac{1}{a}\,\dot{u}_0(t) & a v_0(t) \\ -\dfrac{1}{a}\,k(u_0(t)) & a\dot{v}_0(t) \end{bmatrix} \tag{1.10}$$

and $Y_0(\tau_0)$ the characteristic matrix

$$Y_0(\tau_0) = \begin{bmatrix} 1 & a v_0(\tau_0) \\ 0 & a\dot{v}_0(\tau_0) \end{bmatrix}. \tag{1.11}$$

The Wronskian determinant $W(t)$ (for which $W(0)=1$) according to Liouville's formula is given by

$$W(t) = \dot{u}_0(t)\dot{v}_0(t) + k(u_0(t))v_0(t) = \exp\left[-\int_0^t h(u_0(s))\,ds\right]. \tag{1.12}$$

Thus the characteristic multipliers of (1.6) are 1 and

$$W(\tau_0) = a\dot{v}_0(\tau_0) = \exp\left[-\int_0^{\tau_0} h(u_0(t))\,dt\right]. \tag{1.13}$$

Also, we define another fundamental matrix $Y_1(t)$ such that its first column is $\dot{p}(t)$ and such that $Y_1^{-1}(0)Y_1(\tau_0)$ is diagonal. The relation between $Y_0(t)$ and $Y_1(t)$ is $Y_0(t) = Y_1(t)Y_1^{-1}(0)$.

Thus

$$Y(t) = \begin{bmatrix} \dot{u}_0(t) & v_0(t) - \dfrac{v_0(\tau_0)}{a[1-W(\tau_0)]}\,\dot{u}_0(t) \\ -k(u_0(t)) & \dot{v}_0(t) + \dfrac{v_0(\tau_0)}{a[1-W(\tau_0)]}\,k(u_0(t)) \end{bmatrix}$$

and

$$Y_1^{-1}(0)Y_1(\tau_0) = \begin{bmatrix} 1 & 0 \\ 0 & W(\tau_0) \end{bmatrix}.$$

Let us consider the system adjoint to $\dot{y}=f'_x(p(t))y$ which is

$$\dot{z} = -[f'_x(p(t))]^* z \tag{1.14}$$

where * denotes the transpose. A fundamental matrix of (1.14) is $Y_1^{*-1}(t)$ whose first column is $z(t)$, the only periodic solution of (1.14) of period τ_0, and thus the first row of $Y_1^{-1}(t)$ is $z^*(t)$, which is given by

$$z^*(t) = [z_1^*(t), z_2^*(t)] = \tag{1.15}$$

$$= \frac{1}{W(t)}\left[\dot{v}_0(t)+\frac{v_0(\tau_0)k(u_0(t))}{a[1-W(\tau_0)]}, -v_0(t)+\frac{v_0(\tau_0)\dot{u}_0(t)}{a[1-W(\tau_0)]}\right],$$

and is periodic in t with period τ_0.

Let us now consider the perturbed Liénard's equation

$$\ddot{u}+h(u)\dot{u}+k(u) = \mu\gamma\left(\frac{t}{c}, u, \dot{u}, \mu, \tau\right). \tag{1.16}$$

It is assumed that the perturbation is controllable, γ is periodic in t with period τ and $\gamma\in C^1$ for all its arguments. M. FARKAS and R. KARIM [6] proved that, if the mean value of the periodic function $h(u_0(t))$ (over a period) is different from zero, then to all sufficiently small $|\mu|$ and $|\vartheta|$ (defined there*), the perturbed equation (1.16) has a unique periodic solution $u_p(t;\mu,\vartheta)$ with unique period $\tau(\mu,\vartheta)$.

According to (1.3), equation (1.16) reduces to the perturbed system

$$\dot{x} = f(x)+\mu g\left(\frac{t}{\tau}, x, \mu, \tau\right) \tag{1.17}$$

where $f(x)$ is as defined before, and

$$g\left(\frac{t}{\tau}, x, \mu, \tau\right) = \operatorname{col}\left[0, \gamma\left(\frac{t}{\tau}, x, \mu, \tau\right)\right]. \tag{1.18}$$

Let the right hand side of (1.17) be analytic (less stringent conditions would be sufficient) in the region $I_t\times\Omega\times I_\mu\times I_\tau$ where $I_t=\{t:\ -\infty<t<\infty\}$, Ω is an open connected region in the u, $\dot{u}$ plane, $I_\mu=\{\mu:\ |\mu|<\alpha\}$ for some $\alpha>0$, $I_\tau=\{\tau:\ |\tau-\tau_0|<\beta\}$ for some β, $0<\beta<\tau_0$.

2. Determination of the periodic solution

Now Poincaré's method will be worked out for the approximate determination of the periodic solution of the perturbed Eq. (1.16) up to the second approximation.

Here the perturbed system (1.17) will be treated on basis of the unperturbed system corresponding to (1.17). It is assumed that we know

1) the unique periodic solution $p(t)$ corresponding to $u_0(t)$;
2) the fundamental matrix solution of the first variational system of (1.4) corresponding to $p(t)$.

*See also [2] or [3].

Let $\varphi(t;\mu,\vartheta)$ denote the periodic solution with period $\tau(\mu,\vartheta)$ of the system (1.17) corresponding to the periodic solution $u_p(t;\mu,\vartheta)$ of (1.16). Introducing the new variable s defined by

$$t=\vartheta+s\tau(\mu,\vartheta), \tag{2.1}$$

the system (1.17) and its periodic solution $\varphi(t;\mu,\vartheta)$ assume the forms

$$dx/ds=\tau(\mu,\vartheta)\left[f(x)+\mu g\left(s+\frac{\vartheta}{\tau(\mu,\vartheta)},\mu,\tau(\mu,\vartheta)\right)\right] \tag{2.2}$$

and

$$\psi(s;\mu,\vartheta)=\varphi(s+\vartheta\tau(\mu,\vartheta);\mu,\vartheta) \tag{2.3}$$

where $\psi(s;\mu,\vartheta)$ is periodic in s with period one. Expanding the solution $\psi(s;\mu,\vartheta)$ of (2.2) and the function $\tau(\mu,\vartheta)$ of fixed ϑ in powers of μ we get

$$\psi(s;\mu,\vartheta)=\sum_{k=0}^{\infty}\mu^k\psi^k(s,\vartheta), \tag{2.4}$$

$$\tau(\mu,\vartheta)=\sum_{k=0}^{\infty}\mu^k\tau_k(\vartheta). \tag{2.5}$$

From the definition of ϑ it follows that $\varphi(t;0,\vartheta)=p(t-\vartheta)$, $\tau_0(\vartheta)=\tau(0,\vartheta)=\tau_0$ and

$$\psi^0(s,\vartheta)=\psi(s,0,\vartheta)=\varphi(\vartheta+s\tau_0;0,\vartheta)=p(s\tau_0). \tag{2.6}$$

So if we determined the unknowns ψ^k and τ_k which are uniquely determined in the same manner as in [3], then $\psi(s;\mu,\vartheta)$ and $\tau(\mu,\vartheta)$ can be determined uniquely.

It is known [4] that the unique periodic solution $u_p(t;\mu,\vartheta)$ of (1.16) and its unique period $\tau(\mu,\vartheta)$ are analytic in a neighbourhood of $\mu=\vartheta=0$.

We shall assume throughout this paper that 1 is a simple characteristic multiplier of (1.6) (i.e. $\int_0^{\tau_0}h(u_0(t))dt\neq 0$).

THEOREM 1. *Under the assumptions stated before, the period and the periodic solution of perturbed Liénard's equation* (1.16), *corresponding to* μ *and* ϑ, *are given by:*

$$\tau(\mu,\vartheta)=\tau_0(\vartheta)+\mu\tau_1(\vartheta)+\mu^2\tau_2(\vartheta)+o(\mu^2) \tag{2.7}$$

and

$$u_p(t;\mu,\vartheta)=u_0\left(\frac{t-\vartheta}{\tau(\mu,\vartheta)}\right)+\mu\psi_1^1\left(\frac{t-\vartheta}{\tau(\mu,\vartheta)},\vartheta\right)+\mu^2\psi_1^2\left(\frac{t-\vartheta}{\tau(\mu,\vartheta)},\vartheta\right)+o(\mu^2) \tag{2.8}$$

with expressions for ψ_1^1, ψ_1^2, τ_1 *and* τ_2 *given by* (2.11), (2.14), (2.10) *and* (2.13), *respectively.*

PROOF. Let

$$\psi^1=\operatorname{col}[\psi_1^1,\psi_2^1] \quad\text{and}\quad \psi^2=\operatorname{col}[\psi_1^2,\psi_2^2].$$

Substituting from (2.4) and (2.5) in (2.2) and equating coefficients of μ on both sides, we find that $\psi'(s,\vartheta)$ satisfies the inhomogeneous equation

$$dz/ds=\tau_0 f_x'(p(s\tau_0))z+\tau_0 g\left(s+\frac{\vartheta}{\tau_0},p(s\tau_0),0,\tau_0\right)+\tau_1 f(p(s\tau_0)). \tag{2.9}$$

Since $Y_0(s\tau_0)$ is the fundamental matrix solution of the homogeneous equation corresponding to (2.9), for which $Y_0(0)=U$ where U is the unit matrix, thus the solution of this equation can be written in the form

$$\psi = Y_0(s\tau_0)c^1+Y_0(s\tau_0)\int_0^s Y_0^{-1}(q\tau_0)B_1(q\tau_0)\,dq,$$

where

$$C' = \operatorname{col}[C_1^1, C_2^1]$$

is a constant vector and will need to be determined too, and

$$B_1(q\tau_0) = \tau_1(\vartheta)\operatorname{col}[\dot u_0(q\tau_0), -k(u_0(q\tau_0))]+$$

$$+\tau_0\operatorname{col}\left[0, \gamma\left(q+\frac{\vartheta}{\tau_0}, u(q\tau_0), \dot u_0(q\tau_0), 0, \tau_0\right)\right].$$

Since $\psi_1^1(0,\vartheta)=0$ therefore $C_1^1=0$ and since the function $\psi'(s\vartheta)$ is periodic in s with period 1, therefore these two conditions determine $\psi^1(s,\vartheta)$ and $\tau_1(\vartheta)$ uniquely. After calculations we get

$$C_2^1 = \frac{W(\tau_0)}{a[1-W(\tau_0)]}\int_0^{\tau_0}\frac{\gamma(r)}{W(r)}\dot u_0(r)\,dr,$$

(2.10)

$$\tau_1(\vartheta) = -\int_0^{\tau_0} z_2^*(r)\gamma(r)\,dr,$$

$$\psi_1^1(s,\vartheta) = \frac{W(\tau_0)v_0(s\tau_0)}{1-W(\tau_0)}\int_0^{\tau_0}\frac{\gamma(r)}{W(r)}\dot u(r)\,dr+$$

(2.11)

$$+s\dot u(s\tau_0)\tau_1(\vartheta)-\dot u_0(s\tau_0)\int_0^{s\tau_0}\frac{\gamma(r)}{W(r)}v_0(r)\,dr+v_0(s\tau_0)\int_0^{s\tau_0}\frac{\gamma(r)}{W(r)}\dot u_0(r)\,dr,$$

where $\gamma(r)=\gamma\left(\frac{r+\vartheta}{\tau_0}, u_0(r), \dot u_0(r), 0, \tau_0\right)$ and $z_2^*(r)$ is defined by (1.15).

In the same way, equating the coefficients of μ^2 on both sides, we find that $\psi^2(s,\vartheta)$ satisfies the equation:

$$\frac{dz}{ds} = \tau_0 f_x'(p(s\tau_0))z+\tau_2(\vartheta)f(p(s\tau_0))+F. \tag{2.12}$$

Here

$$F = \frac{\tau_0}{2}f_{xx}''p(s\tau_0)\psi_1^2+\tau_1 f_x'(p(s\tau_0))\psi^1+$$

$$+\tau_0 g'\left(s+\frac{\vartheta}{\tau_0}, p(s\tau_0), 0, \tau_0\right)\psi^1+\tau_1 g\left(s+\frac{\vartheta}{\tau_0}, p(s\tau_0), 0, \tau_0\right)-$$

$$-\frac{\vartheta\tau_1}{\tau_0}g_t'\left(s+\frac{\vartheta}{\tau_0}, p(s\tau_0), 0, \tau_0\right)+$$

$$+\tau_0 g_\mu'\left(s+\frac{\vartheta}{\tau_0}, p(s\tau_0), 0, \tau_0\right)+\tau_0\tau_1 g_\tau'\left(s+\frac{\vartheta}{\tau_0}, p(s\tau_0), 0, \tau_0\right)$$

where f''_{xx} stands for the vector whose i^{th} component is

$$\sum_{k,l=1}^{2} f''_{ix_k x_l}\psi_k^1\psi_l^1, \quad i=1,2,$$

and $f'_x\psi'$ stands for the vector whose i^{th} component is

$$\sum_{k=1}^{2} f'_{ix_k}\psi_k^1.$$

The complete solution of (2.13) is

$$\psi^2(s,\vartheta) = Y_0(s\tau_0)C^2 + Y_0(s\tau_0)\int_0^s Y_0^{-1}(q\tau_0)B_2(q\tau_0)\,dq$$

where

$$B_2(q\tau_0) = \tau_2(\vartheta)f\big(p(q\tau_0)\big)+F$$

and

$$C^2 = \operatorname{col}[C_1^2, C_2^2]$$

is a constant vector. Since $\psi_1^2(0,\vartheta)=0$ therefore $C_1^2=0$. Since $\psi^2(s,\vartheta)$ is periodic in s with period one, after calculations we obtain:

$$\begin{aligned}(2.13)\qquad \tau_2(\vartheta) = {} & \frac{-av_0(\tau_0)}{W(\tau_0)}C_2^2 + \frac{1}{\tau_0}\int_0^{\tau_0}\frac{-1}{W(t)}\Big\{\big[\dot v_0(r)h'\big(u_0(r)\big)\psi_1^{12}(r,\vartheta)+\\ & + v_0(r)k''\big(u_0(r)\big)\psi_1^{12}(r,\vartheta)\big]+\tau_1\big[-\dot v_0(r)h\big(u_0(r)\big)\psi_1^1(r,\vartheta)+\\ & + \psi_2^1(r,\vartheta)+v_0(r)k'\big(u_0(r)\big)\psi_1^1(r,\vartheta)\big]-\Big[\tau_1\gamma(r)-\frac{\vartheta\tau_1}{\tau_0}\gamma'_t(r)+\\ & + \tau_0\gamma'_\mu(r)+\tau_0\tau_1\gamma'_\tau(r)+\tau_0\big(\gamma'_{x_1}(r)\psi_1^1+\gamma'_{x_2}(r)\psi_2^1\big)v_0(r)\Big\}\,dr\end{aligned}$$

and

$$\begin{aligned}(2.14)\qquad \psi_1^2(s,\vartheta) = {} & av_0(s\tau_0)C_2^2+\frac{1}{\tau_0}\int_0^{s\tau_0}\frac{1}{W(r)}\Big\{\big[\dot u_0(s\tau_0)\cdot\dot v_0(r)+\\ & + v_0(s\tau_0)k\big(u_0(r)\big)\big]\cdot\Big[-\frac{1}{2}h'\big(u_0(r)\psi_1^{12}(r,\vartheta)+\tau_1 h\big(u_0(r)\psi_1^1(r,\vartheta)+\\ & + \psi_2^1(r,\vartheta)\big]+\big[-\dot u_0(s\tau_0)v_0(r)+v_0(s\tau_0)\dot u_0(r)\big]\cdot\Big[-\frac{1}{2}k''\big(u_0(r)\big)\psi_1^{12}(r,\vartheta)-\\ & - \tau_1 k'\big(u_0(r)\big)\psi_1^1(r,\vartheta)+\tau_1\gamma(r)-\frac{\vartheta\tau_1}{\tau_0}\gamma'_t(r)+\tau_0\gamma'_{x_1}(r)\psi_1^1(r,\vartheta)+\\ & + \tau_0\gamma'_{x_2}(r)\psi_2^1(r,\vartheta)+\tau_0\gamma'_\mu(r)+\tau_0\tau_1\gamma'_\tau(r)\Big]\Big\}\,dr+\tau_0 s\dot u_0(s\tau_0).\end{aligned}$$

According to (1.3) and (1.5) the first component of the solution $\varphi(t;\mu,\vartheta)$ of (2.2) is $u_p(t;\mu,\vartheta)$, i.e.

$$u_p\big(\vartheta+s\tau(\mu,\vartheta);\mu,\vartheta\big) = \psi_1(s;\mu,\vartheta).$$

Thus according to (2.4), (2.5) and (2.6) we get (2.8).

3. Stability of the periodic solution

We shall give an explicit sufficient condition for the asymptotic stability of the periodic solution $u_p(t;\mu,\vartheta)$ of the perturbed Liénard's equation. This can be done by applying theorems 2 and 4 of M. FARKAS [4]. The nature of the transformation formula (1.3) which transforms the perturbed Liénard's differential Eq. (1.16) into system (1.17) ensures that the asymptotic stability of the solution $\varphi(t;\mu,\vartheta)$ of the system (1.7) implies the asymptotic stability of the corresponding solution $u_p(t;\mu,\vartheta)$ of (1.16). The first variational system of (1.17) with respect to the solution $\varphi(t;\mu,\vartheta)$ is

$$y' = \left[\frac{df(x)}{dx} + \mu \frac{\partial g\left(\frac{t}{\tau}, x, \mu, \tau\right)}{\partial x}\right]_{x=\varphi(t;\mu,\vartheta)} y, \tag{3.1}$$

which reduces for $\mu=0$, $\vartheta=0$ to the variational system (1.6). Let $\lambda(\mu,\vartheta)$ be the characteristic multiplier of the system (3.1) for which $\lambda(0,0)=1$ holds and $\lambda'_\mu(0,0)$ the partial derivative of $\lambda(\mu,\beta)$ with respect to μ at $\mu=\vartheta=0$.

THEOREM 2. *Under the assumptions stated in section* 2, *there exist* $\varrho_1>0$ *and* $\varrho_2>0$ *such that in the region*

$$|\mu| < \varrho_1, \quad |\vartheta| < \varrho_2, \tag{3.2}$$

$\lambda(\mu,\vartheta)$ *(for which* $\lambda(0,0)=1$*) is a real valued analytic function of its arguments* μ *and* ϑ, *and if the second characteristic multiplier of (1.6) is in modulus less than one, then the periodic solution* $\varphi(t;\mu,\vartheta)$ *of the perturbed system* (1.17) *with period* $\tau=\tau(\mu,\vartheta)$ *is asymptotically stable for all* (μ,ϑ) *satisfying* (3.2) *and the condition*

$$\mu\lambda'_\mu(0,0) < 0 \tag{3.3}$$

where

$$\lambda'_\mu(0,0) = -\int_0^{\tau_0} z_2^*(t)\gamma'_t\left(\frac{t}{\tau_0}, u_0(t), \dot{u}_0(t), 0, \tau_0\right) dt. \tag{3.4}$$

PROOF. The first part of the theorem is a consequence of M. FARKAS's theorem 3 (cf. [2]) and only the formula (3.4) is left to be proved.

According to theorem 4 (cf. [4]) we need to calculate the matrices $C^0(0)$, $C^1(0)$ and $d'_\lambda(1;0,0)$. We have

$$C^0(0) = \begin{bmatrix} 1 & av_0(\tau_0) \\ 0 & W(\tau_0) \end{bmatrix}$$

where $W(\tau_0)=\det[Y_0(\tau_0)]=a\dot{v}_0(\tau_0)$ and

$$d(\lambda;0,0) = \begin{vmatrix} 1-\lambda & av_0(\tau_0) \\ 0 & W(\tau_0)-\lambda \end{vmatrix},$$

thus

$$d'_\lambda(1;0,0) = 1 - W(\tau_0), \tag{3.5}$$

and so according to FARKAS's theorem 4 (cf. [4])

$$\lambda'_\mu 0,0) = C_{11}^1(0) + \frac{av_0(\tau_0)}{1-W(\tau_0)} C_{21}^1(0). \tag{3.6}$$

To evaluate $\lambda'_\mu(0, 0)$ we need only the first column of $C^1(0)$ which is given by

$$C^1(0) = \tau_1(0) f'_x(p(\tau_0)) Y_0(\tau_0) + Y_0(\tau_0) \int_0^{\tau_0} Y_0^{-1}(t) B(t, 0)\, dt \tag{3.7}$$

where

$$B(t, 0) = \left[f''_{xx}(p(t)) \varphi'_\mu(t; 0, 0) + g'_x\left(\frac{t}{\tau_0}, p(t), 0, \tau_0\right)\right] Y_0(t) \tag{3.8}$$

and

$$\varphi'_\mu(t; 0, 0) = \psi'\left(\frac{t}{\tau_0}, 0\right) - t\dot{p}(t)\frac{\tau_1(0)}{\tau_0}. \tag{3.9}$$

Thus

$$f''_{xx}(p(t)) \varphi'_\mu(t; 0, 0) = \begin{bmatrix} -h'(u_0(t)) \varphi'_{1\mu}(t; 0, 0) & 0 \\ -k''(u_0(t)) \varphi'_{1\mu} & 0 \end{bmatrix} \tag{3.10}$$

and

$$g'_x\left(\frac{t}{\tau_0}, u_0(t), \dot{u}_0(t), 0, \tau_0\right) = \begin{bmatrix} 0 & 0 \\ \gamma'_{x_1} & \gamma'_{x_2} \end{bmatrix} \tag{3.11}$$

where γ'_{x_1} and γ'_{x_2} are the partial derivatives of the function $\gamma\left(\frac{t}{\tau}, x_1, x_2, \mu, \tau\right)$ with respect to x_1 and x_2 evaluated at $\tau=\tau_0$, $\mu=0$ and $x=p(t)$.

After simple calculations and by using (1.5) we get the following expression:

$$\begin{aligned} \lambda'_\mu(0, 0) = &-\tau_1(0) h(0) + \frac{a\tau_1(0) v_0(\tau_0) k'(0)}{W(\tau_0) - 1} + \\ &+ \int_0^{\tau_0} z^*(t) \left[f''_{xx}(p(t)) \varphi'_\mu(t; 0, 0) + g'_x\left(\frac{t}{\tau_0}, p(t), 0, \tau_0\right)\right] \dot{p}(t)\, dt, \end{aligned} \tag{3.12}$$

where the integrand is a scalar product.

By using (1.14), (1.15) and taking into account the periodicity of $z^*(t)$, $f'_x(p(t))$ and $g\left(\frac{t}{\tau_0}, p(t), 0, \tau_0\right)$, it is easy to prove that

$$\int_0^{\tau_0} z^*(t) \left[f''_{xx}(p(t)) \varphi'_\mu(t; 0, 0) + g'_x\left(\frac{t}{\tau_0}, p(t), 0, \tau_0\right)\right] \dot{p}(t)\, dt =$$

$$= \tau_1(0) \left\{ h(0) + \frac{a v_0(\tau_0) k'(0)}{1 - W(\tau_0)} \right\} - \int_0^{\tau_0} z^*(t)\, g'_t\left(\frac{t}{\tau_0}, p(t), 0, \tau_0\right) dt.$$

Thus by substituting this expression in (3.12) we get:

$$\begin{aligned} \lambda'_\mu(0, 0) &= -\int_0^{\tau_0} z^*(t)\, g'_t\left(\frac{t}{\tau_0}, p(t), 0, \tau_0\right) dt = \\ &= -\int_0^{\tau_0} z_2^*(t) \gamma'_t\left(\frac{t}{\tau_0}, u_0(t), \dot{u}_0(t), 0, \tau_0\right) dt, \end{aligned} \tag{3.13}$$

and by this the theorem is proved.

NOTE. A special case arises from the above results if $\tau(\mu)=\tau_0$ i.e. the perturbation γ is of period τ_0. Then the condition (3.13) of the asymptotic stability is in accordance with that due to W. S. LOUD [8].

If $\lambda'_\mu(0,0)=0$, we restrict the consideration to solutions of system (1.17) corresponding to the case $\vartheta=0$. The value of $\lambda''_{\mu\mu}(0,0)$ can be calculated and if $\lambda''_{\mu\mu}(0,0)<0$ then the characteristic multiplier $\lambda(\mu,\tau)$ of the variational system (3.1) is less than one for all small non-zero μ. Thus if the latter condition holds and the second characteristic multiplier of (1.6) is in modulus less than one, then the periodic solution $u_p(t;\mu,0)$ of the perturbed Liénard's equation (1.16) with period $\tau(\mu,0)$ is asymptotically stable. An expression for $\lambda''_{\mu\mu}(0,0)$ in a general case was given by the author (cf. [2]), also an expression for $\lambda''_{\mu\mu}(0,0)$ was given for the perturbed equation

$$\ddot{u}+h(u)\dot{u}+u=\mu\sin\left(\frac{2\pi}{\tau}t\right)$$

(see the author's dissertation, Hungarian Academy of Sciences, Budapest, 1972).

Acknowledgements

The author wishes to express his appreciation to Prof. M. Farkas for several valuable discussions concerning these results.

REFERENCES

[1] CODDINGTON, E. and LEVINSON, N.: *Theory of Ordinary Differential Equations*, McGraw-Hill Book Company, New York, 1955.
[2] EL OWAIDY, H.: Further Stability Conditions for Controllably Periodic Perturbed Solutions, *Studia Sci. Math. Hung.* **10** (1975), 000—000.
[3] FARKAS, M.: Controllably Periodic Perturbations of Autonomous Systems, *Acta Math. Acad. Sci. Hungar.* **22** (1971), 337—348.
[4] FARKAS, M.: Determination of Controllably Periodic Perturbed Solutions by Poincaré's Method, *Studia Sci. Math. Hungar.* **7** (1972), 257—266.
[5] FARKAS, I. and FARKAS, M.: On Perturbations of Van der Pol's Equation, *Annales Univ. Sci. Budapest, Sectio Math.* **15** (1972) 155—164.
[6] FARKAS, M. and KARIM, R.: Periodic Solution of Perturbed Liénard's Equation, *Periodica Polytechnica, Electrical Eng. Hung.*, Vol. **16** No. 1. (1972), 41—45.
[7] LEVINSON, N. and SMITH, O. K.: A General Equation for Relaxation Oscillation, *Duke Math. Journal* **9** (1942), 382—403.
[8] LOUD, W. S.: Periodic Solutions of a Perturbed Autonomous System, *Ann Math.* **70** (1959), 496—529.

Math. Dept. Faculty of Science, Al Azhar University, Nasr City, Cairo, Egypt

(Received March 2, 1973)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 297—308.

A CLASS OF SQUARE INTEGRABLE IRREDUCIBLE UNITARY REPRESENTATIONS OF SOME LINEAR GROUPS OVER COMMUTATIVE p-FIELDS

by

R. G. LAHA

Introduction

Let k be a commutative p-field which is not necessarily of characteristic 0 (cf. WEIL [2]). We denote by $\mathcal{O}$, $\mathcal{P}$ and $\mathcal{O}^\times$ the unique maximal compact subring (that is, the ring of integers) of k, the unique maximal ideal contained in $\mathcal{O}$ and the group of invertible elements (that is, the unit group) of the ring $\mathcal{O}$ respectively. Let $\tilde{k}=\mathcal{O}/\mathcal{P}$ be the residue class field. Then $\tilde{k}$ is a finite field of characteristic $p>1$ (p being a prime number). Let $n>1$ be a fixed positive integer. Let G be the subgroup of $GL(n, k)$ consisting of those elements whose determinant belongs to O^*. Then $K = GL(n, \mathcal{O})$ is a maximal compact subgroup of G. Recently SHINTANI [1] constructed some square integrable irreducible unitary representations of G which are induced by suitable irreducible unitary representations of K, where these representations of K can be "parametrized" by certain characters of suitable compact Cartan subgroups of G which are contained in K. However, this construction is based on the assumption that $(n, p)=1$, that is, n and p are relatively prime. In the present article we show that an interesting subclass of these representations of G can be constructed without the assumption that $(n, p)=1$. We also give a more explicit description of the structure of these representations of G. In § 1 we present some preliminary results which are used in our investigation. In § 2 we construct some compact Cartan subgroups of G which are contained in K and which are in one-to-one correspondence with an unramified extension of k of degree n over k. In § 3 we construct some irreducible unitary representation of K which can be parametrized by certain characters of these compact Cartan subgroups of G and which induce square integrable irreducible unitary representations of G.

NOTATIONS. Let $\mathcal{R}$ be an arbitrary set of elements. We denote by $\mathcal{M}(n_1, n_2; \mathcal{R})$ the set of all $n_1 \times n_2$ matrices with elements belonging to $\mathcal{R}$. In particular, if $n_1 = n_2 = n$, we write $\mathcal{M}(n, n; \mathcal{R}) = \mathcal{M}(n, \mathcal{R})$. Suppose that $\mathcal{R}$ is a ring with identity. Then $\mathcal{M}(n, \mathcal{R})$ is a ring with identity which we denote by I_n. In this case, we denote by diag $(a_1, a_1, \ldots, a_n)$ $(a_j \in \mathcal{R}: 1 \leqq j \leqq n)$ the diagonal matrix belonging to $\mathcal{M}(n, \mathcal{R})$ whose diagonal elements are a_j $(1 \leqq j \leqq n)$. When $\mathcal{R}$ is a commutative ring, we denote by $\det x$, the determinant of an element $x \in \mathcal{M}(n, \mathcal{R})$.

This work was supported by the National Science Foundation through grant NSF-GP-35724X.

§ 1. Presentation of some preliminary results

This section is of an expository nature. Here we present some preliminary results (mostly without proofs) which will be used in our investigation. For the proofs of these results, we refer to SHINTANI [1].

1.1. Let G be a finite group and let H be a subgroup of G. Let σ be a representation of H in a finite dimensional vector space $\mathscr{V}$ over the field $\mathbf{C}$ of complex numbers. We denote by $\mathscr{V}_\sigma$ the vector space consisting of the mappings f of G into $\mathscr{V}$ which satisfy the condition

$$f(hg) = \sigma(h)f(g)$$

for every $h\in H$ and $g\in G$. We then define a representation π of G in the space $\mathscr{V}_\sigma$ by the formula

$$(\pi(g_0)f)(g) = f(gg_0) \quad (f\in\mathscr{V}_\sigma;\ g_1g_0\in G).$$

Then we say that π is the representation of G induced by the representation σ of H and we write $\pi = \underset{H\uparrow G}{\mathrm{Ind}}\,\sigma$.

In particular, suppose that H is a normal subgroup of G. We denote by $\hat{H}$ the set of all one-dimensional representations of H. Let $g\in G$ and $\chi\in\hat{H}$. We then define $g\cdot\chi\in\hat{H}$ by the formula

$$(g\cdot\chi)(h) = \chi(g^{-1}hg) \quad (h\in H).$$

We set

$$G_\chi = \{g\in G:\ g\cdot\chi = \chi\}.$$

Then G_χ is a subgroup of G containing H and is called the centralizer of χ in G.

LEMMA 1. (i) *Let $\chi\in\hat{H}$. Let σ_χ be an irreducible representation of G_χ in a finite dimensional vector space such that the restriction of σ_χ to H coincides with $\chi\cdot 1$. Then $\underset{G_\chi\uparrow G}{\mathrm{Ind}}\,\sigma_\chi$ is an irreducible representation of G.*

(ii) *Let σ be an arbitrary irreducible representation of G in a finite dimensional vector space $\mathscr{V}$. For every $\chi\in\hat{H}$, let $\mathscr{V}_\chi$ be the subspace of $\mathscr{V}$ defined by the formula*

$$\mathscr{V}_\chi = \{v\in\mathscr{V}:\ \sigma(h)v = \chi(h)v \text{ for every } h\in H\}.$$

Suppose that there exists a $\chi_0\in\hat{H}$ such that $\mathscr{V}_{\chi_0}\neq(0)$. Let $\mathcal{O}_{\chi_0} = \{g\cdot\chi_0\in\hat{H}:\ g\in G\}$ be the G-orbit in $\hat{H}$ containing χ_0. Then $\mathscr{V} = \sum_{\chi\in\mathcal{O}_{\chi_0}} \mathscr{V}_\chi$. Moreover for every $\chi\in\mathcal{O}_{\chi_0}$, $\mathscr{V}_\chi$ is a non-zero subspace of $\mathscr{V}$ which is invariant with respect to the subgroup G_χ of G. Let σ_χ be the representation of G_χ in the space $\mathscr{V}_\chi$ defined by the relation

$$\sigma_\chi(g) = \sigma(g)|\mathscr{V}_\chi \quad (g\in G_\chi).$$

Then σ_χ is an irreducible representation of G_χ in the space $\mathscr{V}_\chi$ and moreover $\underset{G_\chi\uparrow G}{\mathrm{Ind}}\,\sigma_\chi$ is equivalent to σ.

1.2. We now present some results from the theory of infinite dimensional representations.

Let G be a unimodular locally compact topological group with a countable base. Suppose that there exists a nontrivial compact open subgroup K of G. We normalize Haar measures dg and dk on G and K respectively such that $\int_K dg = \int_K dk = 1$.

Let σ be a continuous irreducible unitary representation of K in a Hilbert space $\mathscr{V}_\sigma$. Since K is compact, we note that $\mathscr{V}_\sigma$ is finite dimensional. Let $d_\sigma = \dim_{\mathbf{C}} \mathscr{V}_\sigma$. We denote by $\mathscr{H}_\sigma$ the set of measurable mappings f of G into $\mathscr{V}_\sigma$ which satisfy the following two conditions:

(i) $f(kg) = \sigma(k) f(g)$
for every $k \in K$ and $g \in G$;

(ii) $\int_G \langle f(g), f(g) \rangle_{\mathscr{V}_\sigma} dg < +\infty$ where $\langle\ ,\ \rangle_{\mathscr{V}_\sigma}$ denotes the inner product in the space $\mathscr{V}_\sigma$. Then $\mathscr{H}_\sigma$ is a Hilbert space with respect to the inner product

$$\langle f_1, f_2 \rangle = \int_G \langle f_1(g), f_2(g) \rangle_{\mathscr{V}_\sigma} dg \quad (f_1, f_2 \in \mathscr{H}_\sigma).$$

Let π_σ be the representation of G in the space $\mathscr{H}_\sigma$ defined by the relation

$$(\pi_\sigma(g_0) f)(g) = f(gg_0) \quad (f \in \mathscr{H}_\sigma;\ g, g_0 \in G).$$

Then π_σ is a continuous unitary representation of G in $\mathscr{H}_\sigma$ and is said to be induced by the representation σ of the subgroup K. We then write $\pi_\sigma = \operatorname{Ind}_{K \uparrow G} \sigma$.

We now denote by $\pi_\sigma | K$ the representation of K in the space $\mathscr{H}_\sigma$ which is obtained by restricting to the subgroup K. Let τ be an arbitrary continuous irreducible unitary representation of K. We denote $[\pi_\sigma | K\colon \tau]$ the multiplicity of τ in $\pi_\sigma | K$.

Let π be an arbitrary continuous irreducible unitary representation of G in a Hilbert space $\mathscr{H}$. Then π is said to be square integrable, if there exists a non zero vector $\xi_0 \in \mathscr{H}$ such that $\int_G |\langle \pi(g)\xi_0, \xi_0 \rangle|^2 dg < +\infty$.

In this case, it is known that there exists a positive number Δ_π (depending only on the equivalence class of π) such that the relation

$$\int_G \langle \pi(g)\xi_1, \eta_1 \rangle \overline{\langle \pi(g)\xi_2, \eta_2 \rangle}\, dg = \frac{1}{\Delta_\pi} \langle \xi_1, \xi_2 \rangle \overline{\langle \eta_1, \eta_2 \rangle}$$

holds for all $\xi_1, \xi_2, \eta_1, \eta_2 \in \mathscr{H}$. Here the number Δ_π is called the formal degree of the representation π.

LEMMA 2. (i) *With the same notations as above, let* $\pi_\sigma = \operatorname{Ind}_{K \uparrow G} \sigma$. *Suppose that* $[\pi_\sigma | K\colon \sigma] = 1$. *Then* π_σ *is a square-integrable irreducible unitary representation of* G *and moreover the formal degree of the representation* π_σ *coincides with* $d_\sigma = \dim_{\mathbf{C}} \mathscr{V}_\sigma$.

(ii) *Suppose that* G *is totally disconnected. Let* $\mathscr{C}_{\mathbf{C}}^{\infty}(G)$ *be the space of all complex-valued locally constant functions with compact support on* G.

Assume that $[\pi_\sigma | K\colon \sigma] = 1$ *and moreover for every irreducible unitary representation* τ *of* K, *the multiplicity* $[\pi_\sigma | K\colon \tau] < +\infty$.

Then for every $\theta \in \mathscr{C}_C^\infty(G)$, *the bounded linear operator* $\pi_\sigma(\theta)$ *in the space* $\mathscr{H}_\sigma$ *defined by the formula*

$$\pi_\sigma(\theta) = \int_G \theta(g)\pi_\sigma(g)\,dg$$

is of trace class and the trace is given by the formula

$$\operatorname{Tr} \pi_\sigma(\theta) = \int_G \Big(\int_K \theta(gkg^{-1}) \operatorname{Tr} \sigma(k)\,dk \Big)\,dg.$$

1.3. Let k be a commutative p-field. Let $\mathcal{O}$, $\mathscr{P}$, $\mathcal{O}^\times$ be the maximal compact subring of k, the maximal ideal in $\mathcal{O}$ and the group of units in $\mathcal{O}$ respectively. Let π be a prime element of k and let q be the module of k. Then $\pi \in \mathscr{P}$ and moreover $\mathscr{P} = \pi\mathcal{O}$. The residue class field $\tilde{k} = \mathcal{O}/\mathscr{P}$ is a finite field of characteristic p containing q elements. For every $v \in \mathbf{Z}$, we write $\mathscr{P}^v = \pi^v \mathcal{O}$ such that $\mathscr{P}^0 = \mathcal{O}$.

For every positive integer $m \geqq 1$, we set $\mathscr{R}_m = \mathcal{O}/\mathscr{P}^m$. Then $\mathscr{R}_m$ is a finite ring and moreover $\mathscr{R}_1 = \tilde{k}$. We denote by Φ_m the canonical homomorphism of the ring $\mathcal{O}$ onto the ring $\mathscr{R}_m$. Then Φ_m can be naturally extended to a homomorphism of the ring $\mathscr{M}(n, \mathcal{O})$ onto the ring $\mathscr{M}(n, \mathscr{R}_m)$. We denote this extension by Φ_m also.

Let $GL(n, k)$ be the group of all invertible elements belonging to the ring $\mathscr{M}(n, k)$. Let G be the subgroup of $GL(n, k)$ consisting of those elements whose determinant belongs to $\mathcal{O}^\times$. We set $K = GL(n, \mathcal{O})$. Then G is a unimodular locally compact topological group with a countable base and $K = GL(n, \mathcal{O})$ is a maximal compact open subgroup of G. Moreover the group G is totally disconnected.

We also note the following important double coset decomposition of G with respect to K:

Let π be a prime element of K. Let

$$m = (m_1, m_2, \ldots, m_n) \in \mathbf{Z}^n$$

and we set

$$\pi^m = \operatorname{diag}(\pi^{m_1}, \pi^{m_2}, \ldots, \pi^{m_n})$$

for every $m \in \mathbf{Z}^n$. We now write

$$\Lambda_0^+ = \Big\{ m \in \mathbf{Z}^n \colon m_1 \geqq m_2 \geqq \ldots \geqq m_n;\ \sum_{j=1}^n m_j = 0 \Big\}.$$

Then we have the decomposition

$$(*) \qquad G = \bigcup_{m \in \Lambda_0^+} K\pi^m K \quad \text{(disjoint union)}.$$

Let $m \geqq 1$ be a positive integer. We set

$$K_m = \{k \in K \colon k \equiv I_n \pmod{\mathscr{P}^m}\}.$$

Then K_m is a compact open normal subgroup of finite index in K and moreover the sequence $K_1 \supset K_2 \supset \ldots$ forms a fundamental system of neighbourhoods of I_n in K.

Let $m \geqq 2$ and $l = \left[\frac{m}{2}\right]$. Then the mapping $x \to \pi^{l-m}(x-1)$ defines an isomorphism of the multiplicative group K_{m-l}/K_m onto the additive group of the

ring $\mathcal{M}(n, \mathcal{O})/\mathcal{M}(n, \mathcal{P}^l) = \mathcal{M}(n, \mathcal{R}_l)$. Hence we conclude that K_{m-l}/K_m is a finite commutative group having the same number of elements as $\mathcal{M}(n, \mathcal{R}_l)$.

1.4. Let χ be a character of the additive group of k of order 0 (cf. [2]) that is, χ is trivial on $\mathcal{O}$, but not trivial on $\mathcal{P}^{-1} = \pi^{-1}\mathcal{O}$. Let $X \in \mathcal{M}(n, \mathcal{O})$. Let $r \geqq 2$ be a positive integer and let $s = \left[\frac{r}{2}\right]$. We now define the function χ_X^r on K_{r-s} by the formula

$$\chi_X^r(k) = \chi(\pi^{-r} \operatorname{Tr} X(k - I_n)) \quad (k \in K_{r-s}).$$

LEMMA 3. (i) *The function χ_X^r is a one-dimensional representation of K_{r-s} which is trivial on K_r.*

(ii) *The mapping $X \to \chi_X^r$ defines a homomorphism of the additive group of $\mathcal{M}(n, \mathcal{O})$ into the multiplicative group of all one dimensional representations of K_{r-s} which are trivial on K_r. Moreover the kernel of this homomorphism is $\mathcal{M}(n, \mathcal{P}^s)$.*

(iii) *For every $k \in K$, let $k \cdot \chi_X^r$ be the function on K_{r-s} defined by*

$$(k \cdot \chi_X^r)(h) = \chi_X^r(k^{-1}hk) \quad (h \in K_{r-s}).$$

Then $k \cdot \chi_X^r = \chi_{kXk^{-1}}^r$.

COROLLARY. *Let $x \in \mathcal{M}(n, \mathcal{R}_s)$. Let $X \in \mathcal{M}(n, \mathcal{O})$ be such that $x = \Phi_s(X)$. We set $\chi_x^r = \chi_X^r$. Then the mapping $x \to \chi_x^r$ $(x \in \mathcal{M}(n, \mathcal{R}_s))$ is an isomorphism of the additive group of $\mathcal{M}(n, \mathcal{R}_s)$ onto the multiplicative group of characters of the finite commutative group K_{r-s}/K_r.*

Let σ be a nontrivial continuous irreducible unitary representation of K in a Hilbert space $\mathcal{V}$. Then $\mathcal{V}$ is finite dimensional. Moreover since K is totally disconnected, there exists a positive integer $r = r(\sigma)$ such that σ is trivial on K_r, but not trivial on K_{r-1}. Hence σ can be identified with an irreducible representation of the finite group K/K_r in a finite dimensional vector space $\mathcal{V}$. Moreover we conclude from Lemma 3 and its corollary that for every $X \in \mathcal{M}(n, \mathcal{O})$ (respectively $X \in \mathcal{M}(n, \mathcal{R}_s)$) the function χ_X^r (respectively χ_x^r) is a one dimensional representation of the group K_{r-s}/K_r which is a commutative normal subgroup of the finite group K/K_r. We denote by K_X (respectively K_x) the centralizer of χ_X^r (respectively χ_x^r) in K. Then it follows from Lemma 3 that

$$K_X = \{k \in K: \Phi_s(kXk^{-1}) = \Phi_s(x)\},$$

$$K_x = \{k \in K: \Phi_s(k)x\Phi_s(k)^{-1} = x\}.$$

Let $x \in \mathcal{M}(n, \mathcal{R}_s)$. We define the subspace $\mathcal{V}_x$ of $\mathcal{V}$ by the formula

$$\mathcal{V}_x = \{v \in \mathcal{V}: \sigma(k)v = \chi_x^r(k)v \text{ for every } k \in K_{r-s}\}.$$

We set

$$O_\sigma = \{x \in \mathcal{M}(n, \mathcal{R}_s): \mathcal{V}_x \neq (0)\}.$$

Then the set O_σ is not empty and moreover for every $x_1, x_2 \in O_\sigma$, we have $\mathcal{V}_{x_1} \cap \mathcal{V}_{x_2} = (0)$.

We can now apply Lemma 1 to the above results and thus obtain

LEMMA. (i) *Let σ be a continuous irreducible unitary representation of K in a vector space $\mathscr{V}$ over $\mathbf{C}$. Let $r=r(\sigma)\geqq 2$ and $s=\left[\frac{r}{2}\right]$. Then the set O_σ is not empty and moreover the decomposition*

$$\mathscr{V}=\sum_{x\in O_\sigma}\mathscr{V}_x \quad \text{(direct sum)}$$

holds.

For every $x\in O_\sigma$, $\mathscr{V}_x$ is a non-zero subspace of $\mathscr{V}$ which is invariant with respect to the subgroup K_x of K. Let σ_x be the representation of K_x in the space $\mathscr{V}_x$ defined by the relation

$$\sigma_x(k)=\sigma(k)|\mathscr{V}_x \quad (k\in K_x).$$

Then σ_x is an irreducible representation of the group K_x and moreover $\operatorname{Ind}_{K_x\uparrow K}\sigma_x$ is equivalent with σ.

(ii) *Conversely, let $X\in\mathscr{M}(n,\mathbb{O})$ be such that $\Phi_1(X)\neq 0$. Let λ be an irreducible unitary representation of K_X such that the restriction of λ to the subgroup K_{r-s} coincides with $\chi_X^r\cdot 1$. Then $\sigma_\lambda=\operatorname{Ind}_{K_X\uparrow K}\lambda$ is an irreducible unitary representation of K such that $r=r(\sigma_\lambda)$ and $\Phi_s(X)\in O_{\sigma_\lambda}$.*

1.5. Next we consider some properties of the unitary representations of G which are induced by irreducible unitary representations of K.

Let σ be an irreducible unitary representation of K in the space $\mathscr{V}_\sigma$ (which is finite dimensional) and let $\pi_\sigma=\operatorname{Ind}_{K\uparrow G}\sigma$ be the corresponding induced representation of G in the Hilbert space $\mathscr{H}_\sigma$. We now consider the double coset decomposition $(*)$ of G with respect to K as given in 1.3.

Let $m\in\Lambda_0^+$. We set $K^m=K\cap(\pi^m)^{-1}K\pi^m$. Then we note that K^m is a subgroup of finite index in K. We denote by σ^m the representation of K^m in the space $\mathscr{V}_\sigma$ defined by the formula

$$\sigma^m(k)=\sigma\left(\pi^m k(\pi^m)^{-1}\right) \quad (k\in K^m).$$

Let $\mathscr{H}_\sigma^m$ be the subspace of the space $\mathscr{H}_\sigma$ consisting of those elements $f\in\mathscr{H}_\sigma$ which vanish outside the compact open subset $K\pi^m K$ of G. Then $\mathscr{H}_\sigma^m$ is a closed subspace of $\mathscr{H}_\sigma$ and moreover $\mathscr{H}_\sigma$ decomposes into the direct sum as follows:

$$\mathscr{H}_\sigma=\sum_{m\in\Lambda_0^+}\oplus\mathscr{H}_\sigma^m.$$

We note that the subspace $\mathscr{H}_\sigma^m$ is invariant with respect to the representation $\pi_\sigma|K$ of K. We denote by $\pi_\sigma^m|K$ the subrepresentation of K defined by the subspace $\mathscr{H}_\sigma^m$. Then the following result holds:

LEMMA. *Let τ be an arbitrary irreducible unitary representation of K. Then for every $m\in\Lambda_0^+$, the relation*

$$[\pi_\sigma^m|K:\ \tau]=[\tau|K^m:\ \sigma^m]$$

holds. Hence the relation

$$[\pi_\sigma|K:\ \tau]=\sum_{m\in\Lambda_0^+}[\tau|K^m:\ \sigma^m]$$

holds.

1.6. Let j be a positive integer such that $1 \leqq j \leqq n-1$. We define the subgroup N_j of G by the relation

$$N_j = \left\{ \begin{pmatrix} I_j & 0 \\ X & I_{n-j} \end{pmatrix} : X \in \mathcal{M}(n-j, j; k) \right\}.$$

Then the following result holds.

LEMMA. *Let τ be an arbitrary irreducible unitary representation of K. Let $m \in \Lambda_0^+$ be such that $[\tau | K^m : \sigma^m] > 0$. Moreover suppose that there exist some positive integers ν and j $(1 \leqq j \leqq n-1)$ such that $m_j - m_{j+1} \geqq r(\tau) - \nu$. Then the restriction of σ to the subgroup $N_j \cap K_\nu$ contains the identity representation of $N_j \cap K_\nu$.*

1.7. Let $r \geqq 2$ and $s = \left[\frac{r}{2}\right]$. Let $X \in \mathcal{M}(n, \mathcal{O})$ be such that $\Phi_1(X) \neq 0$. We define the one dimensional representation χ_X^r of the subgroup K_{r-s} which is trivial on K_r as described in 1.4. Let K_X be the centralizer of χ_X^r in K. Then $K_{r-s} \subset K_X$. Let λ be an irreducible unitary representation of K_x such that λ coincides with $\chi_X^r \cdot 1$ on K_{r-s}. Then we conclude from Lemma 4 that $\sigma_\lambda = \operatorname{Ind}_{K_X \uparrow K} \lambda$ is an irreducible unitary representation of K such that $r(\sigma_\lambda) = r$ and $\Phi_s(X) \in O_{\sigma_\lambda}$.

LEMMA. *With the same notations as above, suppose that the characteristic polynomial $\det(t \cdot 1_n - \Phi_1(X))$ of $\Phi_1(X)$ is irreducible over $\tilde{k}$. Let $m \in \Lambda_0^+$ be such that $\Delta(m) = \max_{1 \leqq j \leqq n-1} (m_j - m_{j+1}) \geqq 1$. Then $[\sigma_\lambda^m | K : \sigma_\lambda^m] = 0$. We now formulate the main result in this section which is essential for our later investigation.*

THEOREM. *With the same notations as above, suppose that there exists an $X \in \mathcal{M}(n, \mathcal{O})$ such that $\Phi_1(X) \neq 0$ and moreover the characteristic polynomial of $\Phi_1(X)$ is irreducible over $\tilde{k}$.*

Then $\pi_{\sigma_\lambda} = \operatorname{Ind}_{K \uparrow G} \sigma_\lambda$ is a squaire integrable irreducible unitary representation of G. Moreover the formal degree of π_{σ_λ} is equal to the dimension of the space $\mathcal{V}_{\sigma_\lambda}$.

For every $\theta \in \mathcal{C}_C^\infty(G)$, the operator $\pi_{\sigma_\lambda}(\theta)$ in the space $\mathcal{H}_{\sigma_\lambda}$ is of trace class.

PROOF. We conclude from Lemma 7 that

$$[\sigma_\lambda | K^m : \sigma_\lambda^m] = 0$$

for every such $m \in \Lambda_0^+$ for which $\Delta(m) = \max_{1 \leqq j \leqq n-1} (m_j - m_{j+1}) \geqq 1$. On the other hand we note that $\Delta(m) < 1$ that is, $\Delta(m) = 0$ if and only if $\pi^m = I_n$ so that we have $K^m = K$ and $\sigma_\lambda^m = \sigma_\lambda$.

Hence we conclude at once from Lemma 5 that

$$[\pi_{\sigma_\lambda} | K : \sigma_\lambda] = [\sigma_\lambda | K : \sigma_\lambda] = 1.$$

Then it follows at once from Lemma 2 that π_σ is a square integrable irreducible unitary representation of G and moreover the formal degree of π_{σ_λ} is equal to the dimension of the space $\mathcal{V}_{\sigma_\lambda}$.

Next let τ be an arbitrary irreducible unitary representation of K. Let $m \in \Lambda_0^+$ be such that

$$\Delta(m) = \max_{1 \leqq j \leqq n-1} (m_j - m_{j-1}) \geqq r(\tau) - r + 1.$$

We now proceed exactly in the same manner as in the proof of Lemma 7 by setting $v=r(\tau)-1$ and conclude that $[\tau|K^m\colon \sigma^m]=0$. This implies that there exists only a finite subset $P_0\subset\Lambda_0^+$ such that $[\tau|K^m\colon \sigma^m]=0$ for all $m\notin P_0$. Hence it follows at once from Lemma 5 that $[\pi_\sigma|K\colon \tau]<+\infty$. Then the proof is an immediate consequence of Lemma 2.

§ 2. Explicit construction of some compact Cartan subgroups of G

A subgroup A of G is said to be a Cartan subgroup of G, if A satisfies the following conditions:

(i) A is a maximal commutative subgroup of G.
(ii) Every element of A is semisimple.

We note that two subgroups H_1 and H_2 of G are said to be conjugate, if there exists an element $g\in G$ such that $gH_1g^{-1}=H_2$. On the other hand, two extensions k' and k'' of the field k are said to be conjugate, if there exists an isomorphism of k' onto k'' over k.

It is shown by SHINTANI [1] that there exists a bijection of the set of all conjugacy classes of compact Cartan subgroups of G onto the set of all conjugacy classes of extensions of the field k of degree n over k. Under this bijection, every compact Cartan subgroup of G is isomorphic with the group of units of the corresponding extension of k.

We can describe this bijection more precisely as follows:

Let $\mathscr{E}$ be a complete system of representatives of conjugacy classes of extensions of the field k of degree n over k. Then $\mathscr{E}$ is a finite set.

Let $k'\in\mathscr{E}$. We note that there exists an injective homomorphism τ of k' into $\mathscr{M}(n,k)$, when both k' and $\mathscr{M}(n,k)$ are considered as algebras over k. Let $\mathfrak{O}'^{\times}$ be the group of units of $\mathfrak{O}'$. Then the corresponding compact Cartan subgroup A of G is given by the relation $A=\tau(k')\cap G=\tau(\mathfrak{O}'^{\times})$. Moreover the set $\{A\colon k\in\mathscr{E}\}$ is a complete system of representative elements of conjugacy classes of compact Cartan subgroups of G.

We now carry out an explicit construction of a compact Cartan subgroup of G which corresponds to an unramified extension of k of degree n over k.

Let k' be an unramified extension of k of degree n over k. Let $\mathfrak{O}'$, $\mathscr{P}'$ and $\mathfrak{O}'^{\times}$ be the maximal compact subring of k', the maximal ideal in $\mathfrak{O}'$ and the group of units of $\mathfrak{O}'$ respectively. Then it is known (cf. [2]) that k' is generated over k by a primitive (q^n-1)-th root of unity. We set $q'=q^n$. Let ζ be a primitive $(q'-1)$-th root of unity. Then $\zeta\in\mathfrak{O}'^{\times}$ and moreover we have $\mathfrak{O}'=\mathfrak{O}[\zeta[$ and $k'=k[\zeta]$.

Let $G(k'/k)$ be the Galois group of k' over k. Then $G(k'/k)$ is finite cyclic group of order n and moreover a generator σ of this group which is called the Frobenius automorphism of k' over k is determined uniquely by the formula $\sigma\colon \zeta\to\zeta^q$. We also note that the minimal polynomial of ζ over k is given by

$$P(t)=(t-\zeta)(t-\sigma\zeta)\dots(t-\sigma^{n-1}\zeta)$$

so that P is irreducible over k.

For every positive integer $m \geqq 1$, we set $\mathfrak{R}'_m = \mathfrak{O}'/\mathfrak{P}'_m$. We denote by Φ'_m the canonical homomorphism of the ring $\mathfrak{O}'$ onto the ring $\mathfrak{R}'_m$. Then Φ'_m can be extended naturally to a homomorphism of $\mathfrak{M}(n, \mathfrak{O}')$ onto $\mathfrak{M}(n, \mathfrak{R}'_m)$. We denote this homomorphism also by Φ'_m. Since k' is unramified over k, we conclude that the restriction Φ'_m to $\mathfrak{O} \subset \mathfrak{O}'$ coincides with Φ_m. In particular, $\mathfrak{R}'_1 = \mathfrak{O}'/\mathfrak{P}' = \tilde{k}'$ is the residue class field of k' and is a cyclic extension of $\tilde{k}$ of degree n over k. Moreover the Galois group $G(\tilde{k}'/\tilde{k})$ of $\tilde{k}'$ over $\tilde{k}$ is a finite cyclic group of order n which is generated by the automorphism $\tilde{\sigma}$ of $\tilde{k}'$ over $\tilde{k}$, related with σ by the formula

$$\tilde{\sigma}(\Phi'_1(x)) = \Phi'_1(\sigma(x)), \quad x \in \mathfrak{O}'.$$

We set $\tilde{\zeta} = \Phi'_1(\zeta)$. Then we have $\tilde{k}' = \tilde{k}[\tilde{\zeta}]$ and the minimal polynomial of $\tilde{\zeta} \in \tilde{k}'$ over $\tilde{k}$ is given by

$$\tilde{P}(t) = (t - \tilde{\zeta})(t - \tilde{\sigma}\tilde{\zeta}) \ldots (t - \tilde{\sigma}^{n-1}\tilde{\zeta})$$

so that $\tilde{P}$ is irreducible over $\tilde{k}$.

Next we define an embedding τ of k' into $\mathfrak{M}(n, k)$ as follows:

We set

$$\mathfrak{Z} = \begin{pmatrix} 1 & \zeta & \ldots & \zeta^{n-1} \\ 1 & \sigma\zeta & \ldots & \sigma\zeta^{n-1} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & \sigma^{n-1}\zeta & \ldots & \sigma^{n-1}\zeta^{n-1} \end{pmatrix}$$

so that $\mathfrak{Z} \in GL(n, \mathfrak{O}')$.

Let $x \in k'$. We set

$$\iota(x) = \begin{pmatrix} x & & & \\ & \sigma(x) & & \\ & & \ddots & \\ & & & \sigma^{n-1}(x) \end{pmatrix} \in \mathfrak{M}(n, k')$$

and then

$$\tau(x) = \mathfrak{Z}^{-1}\iota(x)\mathfrak{Z}.$$

Then we can verify easily that τ is an injective homomorphism of k' into $\mathfrak{M}(n, k)$, when both k' and $\mathfrak{M}(n, k)$ are considered as algebras over k.

We then define the element $X \in \mathfrak{M}(n, \mathfrak{O})$ by the formula

$$X = \tau(\zeta) = \mathfrak{Z}^{-1}\iota(\zeta)\mathfrak{Z}.$$

Since $\mathfrak{O}' = \mathfrak{O}[\zeta]$, we conclude that $r(\mathfrak{O}') \subset \mathfrak{M}(n, \mathfrak{O})$. We now define the compact Cartan subgroup A by the formula $A = \tau(k') \cap G = \tau(\mathfrak{O}'^{\times})$. Then clearly $A \subset K$ and moreover $X \in A$.

LEMMA. (i) *The characteristic polynomial* $\det(t \cdot 1_n - \Phi_1(X))$ *of* $\Phi_1(X)$ *is irreducible over* $\tilde{k}$.

(ii) *The centralizer of* X *in* G *is* A.

PROOF. (i) We note that

$$\Phi_1(X) = \Phi'_1(X) = \Phi'_1(\mathfrak{Z})^{-1}\Phi'_1(l(\zeta))\Phi'_1(\mathfrak{Z})$$

so that we have

$$\begin{aligned}\det\left(t\cdot l_n-\Phi_1(X)\right) &= \det\left(t\cdot l_n-\Phi_1'(\iota(\zeta))\right)\\ &= \left(t-\Phi_1'(\zeta)\right)\left(t-\Phi_1'(\sigma\zeta)\right)\ldots\left(t-\Phi_1'(\sigma^{n-1}\zeta)\right)\\ &= (t-\tilde{\zeta})(t-\tilde{\sigma}\tilde{\zeta})\ldots(t-\tilde{\sigma}^{n-1}\tilde{\zeta})\\ &= \tilde{P}(t).\end{aligned}$$

Hence this polynomial is irreducible over $\tilde{k}$.

(ii) Let $Z_X=\{g\in G\colon gXg^{-1}=X\}$ be the centralizer of X in G.

Let $g\in Z_X$. Then $g\in G$. Since $X\in A$ and A is a maximal commutative subgroup of G, we conclude that $g\in A$ so that $Z_X\subset A$. Conversely suppose that $g\in A$. Since $\mathcal{O}'=\mathcal{O}[\zeta]$ and $X=\tau(\zeta)$, we conclude that g can be expressed as a polynomial in X with coefficients belonging to $\mathcal{O}$. Hence we have $gX=Xg$ so that $g\in Z_X$. Therefore $A=Z_X$.

COROLLARY. The minimal polynomial of $\Phi_1(X)$ over $\tilde{k}$ coincides with its characteristic polynomial and hence in particular it is of degree n.

This implies that the element $X\in\mathcal{M}(n,\mathcal{O})$ is quasiregular in the sense of SHINTANI [1].

§ 3. Construction of irreducible unitary representations of K

In this section we shall construct some irreducible unitary representations of K which can be parametrized by certain characters of the compact Cartan subgroup $A\subset K$ and moreover which induce square integrable irreducible unitary representations of G.

We now use the same notations as in § 2. For every positive integer $m\geqq 1$, we define the subgroup $\mathfrak{U}_m'$ of the group of units $\mathcal{O}'^{\times}$ of the field k' by the formula

$$\mathfrak{U}_m'=\{a\in\mathcal{O}'^{\times}\colon a-1\in\mathfrak{P}'^m\}.$$

Then $\mathfrak{U}_m'$ is a compact open subgroup of finite index in $\mathcal{O}'^{\times}$. Moreover since k' is totally disconnected, the subgroups $\mathfrak{U}_1'\supset\mathfrak{U}_2'\supset\ldots$ form a fundamental system of neighborhoods of the identity in $\mathcal{O}'^{\times}$.

Let ω be an arbitrary character of the multiplicative group $\mathcal{O}'^{\times}$ and let $\mathfrak{P}'^r$ be the conductor of ω (cf. [2]). We now assume that $r\geqq 2$ and we set $s=\left[\frac{r}{2}\right]$. This implies that ω is a character of the compact group $\mathcal{O}'^{\times}$ such that ω is trivial on $\mathfrak{U}_r'$, but is not trivial on $\mathfrak{U}_{r-1}'$. Let χ be a character of the additive group of k of order 0, that is, χ is trivial on $\mathcal{O}$, but not trivial on $\mathfrak{P}^{-1}=\pi^{-1}\mathcal{O}$.

Since k' is unramified over k, we conclude (cf. [2]) that the different of k' over k is $\mathcal{O}'$ so that the mapping

$$x\to\chi(\mathrm{Tr}_{k'/k}(x)),\quad x\in k'$$

defines a character of the additive group of k' of order 0 so that it is trivial on $\mathcal{O}'$, but not trivial on $\mathfrak{P}'^{-1}=\pi^{-1}\mathcal{O}'$. Here we note that since $\mathrm{ord}_{k'}(\pi)=1$, we take π, which is a prime element of k, also to be a prime element of k'. Since the mapping

$$a\to\pi^{s-r}(a-1),\quad a\in\mathfrak{U}_{r-s}'$$

defines an isomorphism of the multiplicative group $\mathfrak{P}'_{r-1}/\mathfrak{P}'_r$ onto the additive group $\mathfrak{O}'/\mathfrak{P}'^s$ we conclude from above that there exists an element $x_\omega \in \mathfrak{O}'$ such that the relation

$$\omega(a) = \chi\big(\mathrm{Tr}_{k'/k}(\pi^{-r} x_\omega (a-1))\big)$$

holds for all $a \in \mathfrak{P}'_{r-s}$.

We shall call the character ω of $\mathfrak{O}'^{\times}$ to be a primitive character, if the element $x_\omega \in \mathfrak{O}'$ is equal to $\zeta \in \mathfrak{O}'^{\times}$ which is the primitive $(q'-1)$th root of unity. Let ω be a primitive character of $\mathfrak{O}'^{\times}$ and let $\mathfrak{P}'^r$ be its conductor. Then the relation

$$\omega(a) = \chi\big(\pi^{-r}(\mathrm{Tr}_{k'/k}\zeta(a-1))\big)$$

holds for every $a \in \mathfrak{P}'_{r-s}$.

We now set $X = \tau(\zeta)$ and define the one-dimensional representation χ^r_X on K_{r-s} by the formula

$$\chi^r_X(k) = \chi\big(\pi^{-r}(\mathrm{Tr}\, X(k-1))\big), \quad k \in K_{r-s}.$$

We also note that $A = \tau(\mathfrak{O}'^{\times})$ is a compact Cartan subgroup of G such that $A \subset K$. We now define the character ω of the group A by the formula

$$\omega\big(\tau(a)\big) = \omega(a), \quad a \in \mathfrak{O}'^{\times}.$$

Then we conclude at once that χ^r_X defined above coincides with the character ω of A on $A \cap K_{r-s}$. We now assume r is even so that $r = 2s$ and $r - s = s$. Then we conclude that χ^{2s}_X coincides with ω on $A \cap K_s$. Since $X \in \mathcal{M}(n, \mathfrak{O})$ is quasiregular we conclude from a result of Shintani that $K_X = \mathcal{Z}_X \cdot K_s$ so that we conclude from Lemma 8 in § 2 that $K_X = A \cdot K_s$.

We now define the function λ_ω on $K_X = A \cdot K_s$ by the formula

$$\lambda_\omega(ak) = \omega(a)\chi^{2s}_X(k), \quad a \in A, k \in K_s.$$

Then we conclude that λ_ω is a one-dimensional representation of K_X which coincides with ω on A and with $\chi^{2s}_X \cdot 1$ on $K_s = K_{r-s}$. Hence it follows from Lemma 4 that $\sigma_\omega = \underset{K_X \uparrow K}{\mathrm{Ind}}\, \lambda_\omega$ is a continuous irreducible unitary representation of K. Then the following result holds.

THEOREM. *Let* $\pi_{\sigma_\omega} = \underset{K \uparrow G}{\mathrm{Ind}}\, \sigma_\omega$.

Then π_{σ_ω} *is a square integrable irreducible unitary representation of* G. *Moreover for every* $\theta \in \mathcal{C}^\infty_C(G)$, *the operator* $\pi_{\sigma_\omega}(\theta)$ *is of trace class.*

The proof follows immediately from Theorem 1, using the fact that the characteristic polynomial of $\Phi_1(X)$ is irreducible over k.

REFERENCES

[1] SHINTANI, T.: On Certain Square-integrable Irreducible Unitary Representations of Some P-adic Linear Groups, *Journal Math. Soc.* Japan **20** (1968), 522—565.

[2] WEIL, A.: *Basic Number Theory*, Springer-Verlag, New York, 1967.

Department of Mathematics, Bowling Green State University, Bowling Green, Ohio 43403, U.S.A.

(Received June 4. 1973)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 309—316.

A CONTRIBUTION TO THE PROBLEM OF WEIGHTED POLYNOMIAL APPROXIMATION OF THE DERIVATIVE OF A FUNCTION BY THE DERIVATIVE OF ITS APPROXIMATING POLYNOMIAL

by

NGUYEN-XUAN-KY

§ 1. Introduction

It was proved by E. STEIN [9] that, if $f(x)$ is 2π-periodic and k times continuously differentiable, $T_n(x)$ is a trigonometric polynomial of best approximation of order n to $f(x)$ then $T_n^{(k)}(x)$ tends uniformly to $f^{(k)}(x)$. At the same time G. FREUD [2] and J. CIPSZER—G. FREUD [1] discussed the problem of trigonometric approximation of the derivative of a function by the derivative of its approximating polynomial. Later G. FREUD [7] considered this problem for weighted polynomial approximation with the weight $W_0(x)=e^{-\frac{x^2}{2}}$ on the real axis. In the present paper the weights $W_\beta(x)=|x|^\beta e^{-\frac{x^2}{2}}$ $(-\infty<x<\infty,\ \beta\geqq 0)$ and $U_\alpha(x)=x^{\frac{\alpha}{2}}e^{-\frac{x^2}{2}}\left(0\leqq x<\infty,\ \alpha\geqq -\frac{1}{2}\right)$ will be analysed. We apply the ideas of G. FREUD [4], [5], [6] and [7].

Let $L_p(a, b)$ $(1\leqq p\leqq\infty)$ be the Banach-space of integrable functions with norm

$$\|f\|_p=\left\{\int_a^b |f(x)|^p\,dx\right\}^{\frac{1}{p}},\quad 1\leqq p<\infty,$$

and for $p=\infty$ the space of bounded measurable functions with norm

$$\|f\|_\infty=\operatorname*{vrai.\,sup}_{a<x<b}|f(x)|.$$

Let $\mathscr{P}_n$ be the set of polynomials of degree at most n. For every $\Pi\in\mathscr{P}_n$ we have $\Pi W_\beta\in L_p(-\infty,\infty)$ and $\Pi U_\alpha\in L_p[0,\infty)$, $(1\leqq p\leqq\infty)$. For $W_\beta f\in L_p(-\infty,\infty)$ and $U_\alpha g\in L_p[0,\infty)$ we define

$$E_n^{(p)}(W_\beta; f)=\inf_{\Pi\in\mathscr{P}_n}\|W_\beta(f-\Pi)\|_p \tag{1}$$

$$E_n^{(p)}(U_\alpha; g)=\inf_{\Pi\in\mathscr{P}_n}\|U_\alpha(g-\Pi)\|_p. \tag{2}$$

In this paper $c_k\langle x; y; z; \ldots\rangle$ denote constants, which depend only on $x, y, z, \ldots$ $(k=1, 2, \ldots)$.

We are going to prove the following two theorems.

THEOREM A. *Let* $2\leqq p\leqq\infty$, *and* $f(x)$ *an even function on the interval* $(-\infty,\infty)$ *satisfying the following conditions:*

(i) f *is absolutely continuous in every finite subinterval of* $(-\infty,\infty)$.

(ii) $W_\beta f\in L_p(-\infty,\infty)$.

(iii) $W_\beta f'\in L_p(-\infty,\infty)$.

Let further $\Pi_n \in \mathscr{P}_n$ *be even. If*

$$\|W_\beta(f-\Pi_n)\|_p \leqq \varepsilon_n \quad (n = 0, 1, \ldots)$$

then

$$\|W_\beta(f'-\Pi_n')\|_p \leqq c_1\langle p;\beta\rangle n^{1/2}\varepsilon_n + c_2\langle p;\beta\rangle E_{n-1}^{(p)}(W_\beta; f') \quad (n = 0, 1, \ldots).$$

THEOREM B. *Let* $\alpha \geqq -\frac{1}{p}$, $2 \leqq p \leqq \infty$, $\alpha + \frac{1}{p} \geqq 0$ *and let* $g(t)$ *be a function defined on* $[0, \infty)$ *satisfying the following conditions:*

(i) $g(t)$ *is absolutely continuous in every closed finite subinterval of* $[0, \infty)$,
(ii) $U_\alpha g \in L_p[0, \infty)$,
(iii) $U_{\alpha+1} g' \in L_p[0, \infty)$.

Let $\Pi_n \in \mathscr{P}_n$. *If*

$$\|U_\alpha(g-\Pi_n)\|_p \geqq \varrho_n \quad (n = 0, 1, \ldots)$$

then

$$\|U_{\alpha+1}(g'-\Pi_n')\|_p \leqq c_3\langle p;\alpha\rangle n^{1/2}\varrho_n + c_4\langle p;\alpha\rangle E_{n-1}^{(p)}(U_{\alpha+1}; g') \quad (n = 0, 1, \ldots).$$

In the case $\beta = 0$ and $1 \leqq p \leqq \infty$ Theorem A was proved by G. FREUD [7] for arbitrary (not necessarily even) f and Π_n. In the cases $\beta > 0$ and $p = \infty$ this theorem was proved by M. SALLAY [8] and she applied this result for a saturation problem.

§ 2. Lemmata on orthogonal polynomials

Let $\varrho(x)$ be a weight function on (a, b), $(-\infty \leqq a < b \leqq \infty)$. We denote by

$$p_n(\varrho; x) = \gamma_n(\varrho)x^n + \ldots \quad (\gamma_n(\varrho) > 0)$$

the n-th orthonormal polynomial with respect to the weight $\varrho(x)$. Let

$$u_\alpha(x) = \begin{cases} x^\alpha e^{-x} & \text{for} \quad x \geqq 0, \\ 0 & \text{for} \quad x < 0 \end{cases} \qquad (\alpha > -1).$$

LEMMA 2.1. *We have*

$$\begin{cases} p_{2n}(W_\beta^2; x) = p_n(u_{\beta-1/2}; x^2), \\ p_{2n+1}(W_\beta^2; x) = x p_n(u_{\beta+1/2}; x^2), \end{cases} \tag{3}$$

$$\begin{cases} \dfrac{\gamma_{2n}(W_\beta^2)}{\gamma_{2n+1}(W_\beta^2)} = (n+\beta+1/2)^{1/2}, \\ \dfrac{\gamma_{2n+1}(W_\beta^2)}{\gamma_{2(n+1)}(W_\beta^2)} = (n+1)^{1/2} \\ \qquad\qquad (n = 0, 1, 2, \ldots). \end{cases} \tag{4}$$

We have for every even $k \geqq 2$:

$$p_k'(W_\beta^2; x) = \frac{k\gamma_k(W_\beta^2)}{\gamma_{k-1}(W_\beta^2)} p_{k-1}(W_\beta^2; x), \tag{5}$$

$$W_\beta^2(x) p_k(W_\beta^2; x) = -\frac{\gamma_k(W_\beta^2)}{2\gamma_{k-1}(W_\beta^2)} [W_\beta^2(x) p_{k-1}(W_\beta^2; x)]'. \tag{6}$$

PROOF. (3): See G. FREUD [5]. (4): This is a consequence of (1) and the following relation (see G. SZEGŐ [10]):

$$\gamma_n(u_\alpha) = \frac{\Gamma^{\frac{1}{2}}(n+1)\Gamma^{-\frac{1}{2}}(n+\alpha+1)}{n!}.$$

(5), (6): See M. SALLAY [8].

REMARK. If k is even, then we have

$$\frac{k}{2}\left[\frac{\gamma_k(W_\beta^2)}{\gamma_{k-1}(W_\beta^2)}\right]^2 = 1. \tag{7}$$

Let $f(x)$ be a function on $(-\infty, \infty)$ and $W_\beta f \in L_p(-\infty, \infty)$ $(1 \leqq p \leqq \infty)$. Since $\Pi_n W_\beta \in L_q(-\infty, \infty)$ $\left(\frac{1}{p}+\frac{1}{q}=1\right)$ for every $\Pi_n \in \mathscr{P}_n$, we have $\Pi_n W_\beta^2 f \in L_1(-\infty, \infty)$. Let $S(W_\beta^2; f; x)$ be the generalized Fourier-series of f with respect to the system $\{p_n(W_\beta^2; x)\}$, i.e.

$$f(x) \sim S(W_\beta^2; f; x) = \sum_{k=0}^{\infty} c_k(W_\beta^2; f) p_k(W_\beta^2; x)$$

where

$$c_k(W_\beta^2; f) = \int_{-\infty}^{\infty} f(t) p_k(W_\beta^2; t) W_\beta^2(t)\, dt \quad (k = 0, 1, 2, \ldots).$$

Let

$$\begin{cases} S_n(W_\beta^2; f; x) = \sum_{k=0}^{n} c_k(W_\beta^2; f) p_k(W_\beta^2; x) \quad (n = 0, 1, 2, \ldots), \\ \sigma_n(W_\beta^2; f; x) = \frac{1}{n} \sum_{k=0}^{n-1} S_k(W_\beta^2; f; x) \quad (n = 1, 2, \ldots), \\ \vartheta_n(W_\beta^2; f; x) = 2\sigma_{2n}(W_\beta^2; f; x) - \sigma_n(W_\beta^2; f; x) \quad (n = 1, 2, \ldots). \end{cases} \tag{8}$$

LEMMA 2.2. *If f is the function in Theorem A then*

$$S_n'(W_\beta^2; f; x) = S_{n-1}(W_\beta^2; f'; x) \quad (n = 1, 2, \ldots). \tag{9}$$

PROOF. Since f is even, we have

$$S_n(W_\beta^2; f; x) = \sum_{k=0}^{n}{}^{*} c_k(W_\beta^2; f) p_k(W_\beta^2; x)$$

where the symbol Σ_* denotes that summation is taken only on even indices. Let

$$\max_{0\leqq x<\infty} W_\beta(x) = W_\beta(a).$$

In consequence of (iii) in Theorem A we have for $x\geqq a$, $k=0, 1, \ldots,$

$$|x^k W_\beta^2(x) f(x)| = \left|x^k W_\beta^2(x)\left[f(a)+\int_a^x f'(t)\,dt\right]\right| \leqq \tag{10}$$

$$\leqq |x^k W_\beta(x) f(a)| + x^k W_\beta(x)\int_a^x |W_\beta(t) f'(t)|\,dt \leqq$$

$$\leqq o(1) + x^k W_\beta(x)(x-a)^{1-1/p}\|W_\beta f'\|_p = o(1) \quad (x\to\infty).$$

We have analogously

$$x^k W_\beta^2(x) f(x) = o(1) \quad (x\to -\infty). \tag{11}$$

By (i) of Theorem A we can apply integration by parts, and we have by (10), (11) and (6)

$$c_k(W_\beta^2; f) = \int_{-\infty}^{\infty} f(t)\, p_k(W_\beta^2; t) W_\beta^2(t)\,dt = \left[\frac{-\gamma_k(W_\beta^2)}{2\gamma_{k-1}(W_\beta^2)} f(t)\, p_{k-r}(W_\beta^2; t) W_\beta^2(t)\right]_{-\infty}^{\infty} -$$

$$-\int_{-\infty}^{\infty} f'(t)\frac{\gamma_k(W_\beta^2)}{-2\gamma_{k-1}(W_\beta^2)}\, p_{k-1}(W_\beta^2; t) W_\beta^2(t)\,dt = \frac{\gamma_k(W_\beta^2)}{2\gamma_{k-1}(W_\beta^2)}\, c_{k-1}(W_\beta^2; f').$$

Thus from (5) and (7) we obtain

$$S_n'(W_\beta^2; f; x) = \sum_{k=0}^{n}{}_* c_k(W_\beta^2; f)\, p_k'(W_\beta^2; x) = \tag{12}$$

$$= \sum_{k=0}^{n}{}_* \frac{k}{2}\left[\frac{\gamma_k(W_\beta^2)}{\gamma_{k-1}(W_\beta^2)}\right]^2 c_{k-1}(W_\beta^2; f')\, p_{k-1}(W_\beta^2; x) = S_{n-1}(W_\beta^2; f'; x).$$

Q. e. d.

LEMMA 2.3. *If* $W_\beta f\in L_p(-\infty, \infty)$ $(1\leqq p\leqq\infty)$ *then*

$$\|W_\beta \sigma_n(W_\beta^2; f)\|_p \leqq c_5\langle p; \beta\rangle \|W_\beta f\|_p. \tag{13}$$

In the case when $p=\infty$, *we have more precisely*

$$W_\beta(x)\frac{1}{n}\sum_{\nu=0}^{n-1} |S_\nu(W_\beta^2; f; x)| \leqq c_5\langle\infty; \beta\rangle \|W_\beta f\|_\infty. \tag{14}$$

PROOF. In the case $\beta=0$ (13) and (14) were proved by G. FREUD [5]. In the more general case, that is when $\beta>0$, the proof is similar to the case $\beta=0$. We apply only the following inequalities:

(i) $\sum_{k=0}^{n-1} p_k^2(W_\beta^2; x) \leqq c_6\langle\beta\rangle n^{1/2} W_\beta^{-2}(x) \quad (\beta\geqq 0, n=1, 2, \ldots)$

(see formula (2.19) of G. FREUD [5]);

(ii) $$\frac{\gamma_{n-1}(W_\beta^2)}{\gamma_n(W_\beta^2)} \leqq c_7\langle\beta\rangle n^{1/2} \quad (n = 1, 2, \ldots)$$

(see (4)).

REMARK. We have by (8) and (13)

$$\|W_\beta \vartheta_n(W_\beta^2; f)\|_p \leqq c_8\langle p, \beta\rangle \|W_\beta f\|_p. \tag{15}$$

§ 3. A Markov-type inequality

THEOREM. *Let* $2 \leqq p \leqq \infty$. *For every even polynomial* $\Pi_n \in \mathscr{P}_n$, *we have*

$$\|W_\beta \Pi_n'\|_p \leqq c_9\langle p; \beta\rangle n^{1/2} \|W_\beta \Pi_n\|_p. \tag{16}$$

PROOF. a) In the case when $p = \infty$, (16) was proved by G. FREUD [6] for every polynomial (not only for even ones).

b) Let $p = 2$. Since

$$S_m(W_\beta^2; \Pi_n) \equiv \Pi_n, \quad m \geqq n \geqq 0,$$

we have by Parseval's formula and (12), (4)

$$\|W_\beta \Pi_n'\|_2 = \|W_\beta S_n(W_\beta^2; \Pi_n')\|_2 =$$

$$= \left\{\sum_{k=0}^{n} c_k^2(W_\beta^2; \Pi_n')\right\}^{1/2} \leqq c_9\langle 2; \beta\rangle n^{1/2} \left\{\sum_{k=0}^{n} c_k^2(W_\beta^2; \Pi_n)\right\}^{1/2} =$$

$$= c_9\langle 2; \beta\rangle n^{1/2} \|W_\beta \Pi_n\|_2.$$

c) Let $W_\beta f \in L_p(-\infty, \infty)$ and f be even. We define the linear operators T_n for $n = 0, 1, 2, \ldots$ as follows:

$$T_n(W_\beta f) = \vartheta_n'(W_\beta^2; f) W_\beta.$$

If $W_\beta f \in L_\infty(-\infty, \infty)$, we have by (a) and (15)

$$\|W_\beta \vartheta_n'(W_\beta^2; f)\|_\infty \leqq c_9\langle \infty; \beta\rangle n^{1/2} \|W_\beta \vartheta_n(W_\beta^2; f)\|_\infty \leqq$$

$$\leqq c_9\langle \infty; \beta\rangle c_8\langle \infty; \beta\rangle n^{1/2} \|W_\beta f\|_\infty.$$

If $W_\beta f \in L_2(-+\infty, \infty)$, we have by (b) and (15)

$$\|W_\beta \vartheta_n'(W_\beta^2; f)\|_2 \leqq c_9\langle 2; \beta\rangle n^{1/2} \|W_\beta \vartheta_n(W_\beta^2; f)\|_2 \leqq c_9\langle 2; \beta\rangle c_8\langle 2; \beta\rangle \|W_\beta f\|_2.$$

By restricting the even functions to $[0, \infty)$, T_n is a $(2, 2)$ and (∞, ∞) type operator (see Theorem 1.11 in [11]). From the Riesz—Thorin interpolation theorem it follows that

$$\|W_\beta \vartheta_n'(W_\beta^2; f)\|_p \leqq c_9\langle p; \beta\rangle n^{1/2} \|W_\beta f\|_p \quad (2 < p < \infty) \tag{17}$$

for every $W_\beta f \in L_p(-\infty, \infty)$, f even. Inserting an even polynomial $\Pi_n \in \mathscr{P}_n$ for f in (17) we obtain (16). Q. e. d.

§ 4. Proof of Theorems A and B

PROOF OF THEOREM A. We have by Lemma 2.2:

$$S_n'(W_\beta^2; f) = S_{n-1}(W_\beta^2; f') \quad (n = 1, 2, \ldots).$$

Thus

$$(18) \quad \vartheta_n'(W_\beta^2; f) = \frac{1}{n} \sum_{\nu=n-1}^{2n-1} S_\nu(W_\beta^2; f) = \frac{2n-1}{n} \sigma_{2n-1}(W_\beta^2; f') - \frac{n-1}{n} \sigma_{n-1}(W_\beta^2; f').$$

Since we have for every $\Pi_n \in \mathscr{P}_n$ $\vartheta_n(W_\beta^2; \Pi_n) \equiv \Pi_n$ we get

$$(19) \quad \|W_\beta[f - \vartheta_n(W_\beta^2; f)]\|_p \leqq \|W_\beta \vartheta_n(W_\beta^2; f - \Pi_n)\|_p + \|W_\beta(f - \Pi_n)\|_p \leqq$$

$$\leqq 2\|W_\beta \sigma_{2n}(W_\beta^2; f - \Pi_n)\|_p + \|W_\beta \sigma_n(W_\beta^2; f - \Pi_n)\|_p + \|W_\beta(f - \Pi_n)\|_p \leqq$$

$$\leqq c_{10}\langle p; \beta\rangle \|W_\beta(f - \Pi_n)\|_p.$$

Thus also

$$(20) \quad \|W_\beta[f - \vartheta_n(W_\beta^2; f)]\|_p \leqq c_{10}\langle p; \beta\rangle E_n^{(p)}(W_\beta; f).$$

Furthermore from $S_\nu(W_\beta^2; \Pi_{n-1}) \equiv \Pi_{n-1}$ $(\nu = n-1, n, n+1, \ldots)$ and (10) we have for an arbitrary $\Pi_{n-1} \in \mathscr{P}_{n-1}$:

$$\|W_\beta[f' - \vartheta_n'(W_\beta^2; f)]\|_p \leqq \|W_\beta(f - \Pi_{n-1})\|_p + \|W_\beta \vartheta_n'(W_\beta^2; f - \Pi_n)\|_p \leqq$$

$$\leqq \frac{2n-1}{n} \|\sigma_{2n-1}(W_\beta^2; f' - \Pi_{n-1}) W_\beta\|_p + \frac{n-1}{n} \|W_\beta \sigma_{n-1}(W_\beta^2; f' - \Pi_{n-1})\|_p +$$

$$+ \|W_\beta(f' - \Pi_{n-1})\|_p \leqq c_{11}\langle p; \beta\rangle \|W_\beta(f' - \Pi_{n-1})\|_p.$$

Thus also

$$(21) \quad \|W_\beta[\vartheta_n'(W_\beta^2; f) - f']\|_p \leqq c_{11}\langle p; \beta\rangle E_{n-1}^{(p)}(W_\beta; f').$$

It is a consequence of (21), (16) and (15) that

$$\|W_\beta(f' - \Pi_n')\|_p \leqq \|W_\beta[f' - \vartheta_n(W_\beta^2; f)]\|_p + \|W_\beta[\Pi_n' - \vartheta_n'(W_\beta^2; f)]\|_p \leqq$$

$$\leqq c_{11}\langle p; \beta\rangle E_{n-1}^{(p)}(W_\beta; f') + c_{12}\langle p; \beta\rangle n^{1/2} \|W_\beta[\Pi_n - \vartheta_n(W_\beta^2; f)]\|_p \leqq$$

$$\leqq c_{11}\langle p; \beta\rangle E_{n-1}^{(p)}(W_\beta; f') + c_{12}\langle p; \beta\rangle n^{1/2}\{\|W_\beta[f - \vartheta_n(W_\beta^2; f)]\|_p + \|W_\beta(f - \Pi_n)\|_p\} \leqq$$

$$\leqq c_{11}\langle p; \beta\rangle E_{n-1}^{(p)}(W_\beta; f') + c_{13}\langle p; \beta\rangle n^{1/2} \varepsilon_n.$$

This completes our proof.

PROOF OF THEOREM B. Let $g(t)$ be the function in Theorem B. We introduce the function:

$$g^*(x) = g(x^2), \quad -\infty < x < \infty.$$

It is a consequence of (i), (ii) and (iii) in Theorem B that $g^*(x)$ satisfies conditions (i), (ii) and (iii) in Theorem A with $\beta = \alpha + \frac{1}{p} \geqq 0$.

We get by the transformation $t=x^2$

$$\|U_\alpha(g-\Pi_n)\|_p = \|W_{\alpha+\frac{1}{p}}(g^*-\Pi_n^*)\|_p. \tag{22}$$

and

$$E_{2n-1}^{(p)}(W_{\alpha+\frac{1}{p}}; [g^*]') \leqq E_{n-1}^{(p)}(U_{\alpha+1}; g'). \tag{23}$$

Thus we have by Theorem A

$$\|U_{\alpha+1}(g'-\Pi_n')\|_p = \|W_{\alpha+1+\frac{1}{p}}[(g')^*-(\Pi_n')^*]\|_p = \|W_{\alpha+\frac{1}{p}}[(g^*)'-(\Pi_n^*)']\|_p \leqq$$

$$\leqq c_{14}\langle p;\alpha\rangle n^{\frac{1}{2}}\varrho_n + c_{15}\langle p;\alpha\rangle E_{2n-1}^{(p)}(W_{\alpha+\frac{1}{p}}; [g^*]') \leqq$$

$$\leqq c_{14}\langle p;\alpha\rangle n^{\frac{1}{2}}\varrho_n + c_{15}\langle p;\alpha\rangle E_{n-1}^{(p)}(U_{\alpha+1}; g'),$$

q. e. d.

REMARK. In Theorems A and B for one p it is not possible to replace $c_1\langle p;\beta\rangle$ and $c_2\langle p;\alpha\rangle$ by a sequence η_n for which $\eta_n\to 0$ as $n\to\infty$. It is suficient to prove that the Markov-type inequality (16) can not be strengthened. To prove this, assume the contrary for some $p\neq 4$. Then, interpolating between p and ∞ for $p<4$, and 2 and p if $p>4$, we obtain from the Riesz—Thorin theorem that (16) can be strengthened also in the case $p=4$. But this contradicts the fact that*

$$\|W_\beta[p_n^*(u_\alpha; x)]'\|_4 \sim n^{1/2}\|W_\beta p_n^*(u_\alpha; x)\|_4. \tag{24}$$

Indeed, (24) follows from (22) and the following relations:

$$\|U_\alpha p_n(u_\alpha; x)\|_4 \sim \sqrt[4]{\frac{1}{\pi^2}\frac{\ln n}{n}} \quad \left(\alpha > -\frac{1}{2}\right)$$

(see G. NÉMETH [12]),

$$p_n'(u_\alpha; x) = \frac{n\gamma_n(u_\alpha)}{\gamma_{n-1}(u_{\alpha+1})} p_{n-1}(u_{\alpha+1}; x).$$

(See formula (5.1.14) of G. SZEGŐ [10]), and

$$\frac{n\gamma_n(u_\alpha)}{\gamma_{n-1}(u_{\alpha+1})} \sim n^{\frac{1}{2}}$$

(see the relation in the proof of Lemma 2.1).

* The symbol $A_n\sim B_n$ denotes that

$$A < \frac{A_n}{B_n} < B$$

where A and B are independent of n.

REFERENCES

[1] CZIPSZER, J.—FREUD, G.: Sur l'approximation d'une fonction périodique et de ses dérivées successives par un polynome trigonométrique et par ses dérivées successives, *Acta Math. Uppsala* **99** (1958), 33—52.
[2] FREUD, G.: Über gleichzeitige Approximation einer Funktion und ihrer Derivierten, *Österreichische Mathem. Nachrichten* **47—48** (1957), 36—37.
[3] FREUD, G.: *Orthogonal polynomials*, Akadémiai Kiadó, Budapest 1971.
[4] FREUD, G.: On an inequality of Markov-type, *Soviet Math. Dokl.* **12** (1971), 570—573.
[5] FREUD, G.: A contribution to the problem of weighted polynomial approximation, ISNM. 20. Birkhäuser, Basel, 1972, 431—447.
[6] FREUD, G.: On two polynomial inequalities II, *Acta Math. Acad. Sci. Hung.* **23** (1972), 175—178.
[7] FREUD, G.: On weighted polynomial approximation of the derivative of a function by the derivative of its approximating polynomial, *Studia Sci. Math. Hung.* 1972.
[8] SALLAY, M.: Über ein Saturationsproblem, Lecture delivered on the conference: "Theory of Approximation" in Poznan, August 1972.
[9] STEIN, E.: Function of exponential type. *Ann. of. Math* **65** (1957), 582—592.
[10] SZEGŐ, G.: *Orthogonal polynomials*, American Math. Soc. 1939.
[11] ZYGMUND, A.: *Trigonometric series, Vol.* **2**, Cambridge University Press, 1968.
[12] NÉMETH, G.: On the L_4 norm of orthogonal Laguerre Polynomials, *Studia Sci. Math. Hung.* **10** (1975), 243—246.

Mathematical Institute of the Hungarian Academy of Sciences, H—1053 Budapest, Reáltanoda u. 13—15, Hungary

(Received April 20, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 317—322.

A NEW PROOF OF A THEOREM OF PÓLYA

by

P. K. PATHAK

Summary

This note furnishes a new proof of a theorem of Pólya on characteristic functions concerning a sufficient condition for a function on the real line to be the characteristic function of a probability density.

1. Introduction and Preliminaries

The object of this note is to furnish a new proof of the following theorem of Pólya (cf. [4], p. 83):

THEOREM 1. *Let φ be a function defined on R_1 and suppose that φ satisfies the following conditions: $\varphi(0)=1$, $\varphi(+\infty)=0$, for each t, $\varphi(t)=\varphi(-t)$, $\varphi(t)\geqq 0$, and φ is decreasing and convex in $[0, \infty]$. Then φ is the characteristic function of an absolutely continuous probability measure.*

In order to present a proof of this theorem we first present a few preliminary results.

LEMMA 1. *Let φ be continuous, decreasing and convex in $[0, \infty]$ with $\varphi(+\infty)=0$. Then* (a) $\lim_{t\to 0} t\varphi'(t)=0$, (b) $\lim_{t\to\infty} t\varphi'(t)=0$, *and* (c) $\int_0^\infty t\,d(\varphi'(t))=\varphi(0)$, *where φ' denotes the derivative of φ from the right.*

LEMMA 2. *Let φ satisfy the hypotheses of Theorem 1. Define, for each $x\neq 0$,*

$$f(x)=\frac{1}{\pi}\int_0^\infty \frac{(1-\cos tx)}{x^2}\,d(\varphi'(t)).$$

Then $f(x)\geqq 0$ and $\int f(x)dx=1$.

PROOF. Clearly $f(x)$ is well-defined and non-negative since $\varphi'(t)$ is increasing and right continuous. Also, for each $x\neq 0$, $f(x)=f(-x)$. So

$$\int f(x)\,dx = 2\int_0^\infty f(x)\,dx = \frac{2}{\pi}\int_0^\infty\int_0^\infty \frac{(1-\cos tx)}{x^2}\,d(\varphi'(t))\,dx.$$

Since the integrand in the last expression is non-negative, we can interchange the order of integration. We thus get

$$\int f(x)\,dx = \frac{2}{\pi}\int_0^\infty \int_0^\infty \frac{(1-\cos tx)}{x^2}\,dx\,d(\varphi'(t)) =$$

$$= \frac{2}{\pi}\int_0^\infty \int_0^\infty \frac{(1-\cos u)}{u^2}\,t\,du\,d(\varphi'(t)) =$$

$$= \left[\frac{2}{\pi}\int_0^\infty \frac{(1-\cos u)}{u^2}\,du\right]\left[\int_0^\infty td(\varphi'(t))\right] = \varphi(0) = 1. \quad \blacksquare$$

LEMMA 3. *Let $\varphi(t)$ be continuous with $\varphi(0)=1$ and $\int |\varphi(t)|dt<\infty$. Suppose that, for each x,*

$$f(x) = \frac{1}{2\pi}\int \exp(-itx)\varphi(t)\,dt \geqq 0.$$

Then $\int f(x)dx=1$ and $\varphi(t)$ is the characteristic function $f(x)$.

PROOF. Since $f(x)\exp(-x^2/2n^2)$ increases to $f(x)$ as n tends to infinity, we have, by the monotone convergence theorem,

$$\int f(x)\,dx = \lim_{n\to\infty}\int f(x)\exp(-x^2/2n^2)\,dx =$$

$$= \lim_{n\to\infty}\int \exp(-x^2/2n^2)\left[\frac{1}{2\pi}\int \exp(-itx)\varphi(t)\,dt\right]dx.$$

We can now interchange the order of integration since,

$$\iint \exp(-x^2/2n^2)|\varphi(t)|\,dt\,dx<\infty.$$

So

$$\int f(x)\,dx = \lim_{n\to\infty}\int \varphi(t)\left[\frac{1}{2\pi}\int \exp(-itx)\exp\left(-\frac{x^2}{2n^2}\right)dx\right]dt.$$

Since the quantity within brackets is the inversion formula for the normal density, we have

$$\int f(x)dx = \lim_{n\to\infty}\int \varphi(t)\left[\frac{n}{\sqrt{2\pi}}\exp\left(-\frac{t^2n^2}{2}\right)\right]dt = \lim_{n\to\infty}\frac{1}{\sqrt{2\pi}}\int \varphi\left(\frac{u}{n}\right)\exp\left(-\frac{u^2}{2}\right)du.$$

By the continuity of φ and since $\int |\varphi(t)|dt<\infty$, it follows that

$$\int f(x)\,dx = \varphi(0)\left[\frac{1}{\sqrt{2\pi}}\int \exp\left(-\frac{u^2}{2}\right)du\right] = \varphi(0) = 1.$$

It is now a simple matter to see that $\varphi(t)=\int \exp(itx)f(x)dx$. Actually this is a well-known result in Fourier integrals, namely if f is the Fourier transform of an L_1-function φ and if f is also L_1 then φ is the inverse Fourier transform of f

(cf. [2], p. 16). Nevertheless we present here an alternative simpler proof. Since $|\exp\left(-\frac{x^2}{2n^2}\right) f(x)| \leqq f(x)$ and $\int f(x)dx=1$, we have, by the dominated convergence theorem,

$$\int \exp(itx)f(x)\,dx = \lim_{n\to\infty} \int \exp(itx)f(x)\exp\left(-\frac{x^2}{2n^2}\right)dx =$$

$$= \lim_{n\to\infty} \int \exp\left(itx-\frac{x^2}{2n^2}\right)\left[\frac{1}{2\pi}\int \exp(-isx)\varphi(s)\,ds\right]dx.$$

We can now interchange the limits of integration since $\int \exp\left(-\frac{x^2}{2n^2}\right)|\varphi(x)|dx\,ds<\infty$. Therefore

$$\int \exp(itx)f(x)\,dx = \lim_{n\to\infty} \int \left[\frac{1}{2\pi}\int \exp\left(i(t-s)x-\frac{x^2}{2n^2}\right)dx\right]\varphi(s)\,ds =$$

$$= \lim_{n\to\infty} \int \varphi(s)\left[\frac{n}{\sqrt{2\pi}}\exp\left(-\frac{n^2(s-t)^2}{2}\right)\right]ds =$$

$$= \lim_{n\to\infty} \frac{1}{\sqrt{2\pi}}\int \varphi\left(t+\frac{u}{n}\right)\exp\left(-\frac{u^2}{2}\right)du.$$

Again, continuity of φ and $\int |\varphi(t)|dt<\infty$ imply that

$$\int \exp(itx)f(x)\,dx = \varphi(t)\left[\frac{1}{\sqrt{2\pi}}\int \exp\left(-\frac{u^2}{2}\right)du\right] = \varphi(t). \quad \blacksquare$$

With these lemmas we can now present a proof of Theorem 1.

2. Proof of Pólya's theorem

CASE 1. Suppose that $\varphi(t)$ of Theorem 1 satisfies the added condition $\int |\varphi(t)|dt<\infty$.

Then for each x, $x\neq 0$,

$$\int \exp(-itx)\varphi(t)\,dt = 2\int_0^\infty \cos tx\varphi(t)\,dt.$$

Since $\varphi(+\infty)=0$, we get on integrating by parts

$$\int \exp(-itx)\varphi(t)\,dt = 2\int_0^\infty \frac{\sin tx}{x}\left(-\varphi'(t)\right)dt.$$

Again since Lemma 1 implies $\lim_{t\to 0} t^2\varphi'(t)=0$ and $\lim_{t\to\infty}\varphi'(t)=0$, we have on integration by parts a second time

$$\int \exp(-itx)\varphi(t)\,dt = 2\int_0^\infty \frac{(1-\cos tx)}{x^2}\,d(\varphi'(t)).$$

Since the right hand side above is non-negative, it follows from Lemma 3 that $\varphi(t)$ is the characteristic function of the measure with density

$$f(x) = \frac{1}{\pi}\int_0^\infty \frac{(1-\cos tx)}{x^2}\, d(\varphi'(t)).$$

CASE 2. For the proof in the general case, let, for each $n \geqq 1$, $\psi_n(t) = \varphi(t)\exp(-|t|/n)$. Since for $t>0$, $-\psi_n'(t) = -\varphi'(t)\exp(-t/n)+(\varphi(t)/n)\exp(-t/n)$ is non-negative and decreasing, it follows that ψ_n satisfies the hypotheses of Theorem 1. In addition $\int |\psi_n(t)|\,dt < \infty$. Consequently by Case 1, $\psi_n(t)$ is the characteristic function of the density function

$$g_n(x) = \frac{1}{\pi}\int_0^\infty \frac{(1-\cos tx)}{x^2}\left[\exp\left(-\frac{t}{n}\right) d(\varphi'(t)) - \right.$$

$$\left. -\frac{2}{n}\exp\left(-\frac{t}{n}\right)\varphi'(t)\,dt + \frac{1}{n^2}\varphi(t)\exp\left(-\frac{t}{n}\right)dt\right] =$$

$$= T_{1n} - T_{2n} + T_{3n},$$

say. Now

$$\lim_{n\to\infty} T_{1n} = \lim_{n\to\infty}\frac{1}{\pi}\int_0^\infty \frac{(1-\cos tx)}{x^2}\exp\left(-\frac{t}{n}\right) d(\varphi'(t)) =$$

$$= \frac{1}{\pi}\int_0^\infty \frac{(1-\cos tx)}{x^2}\, d(\varphi'(t)) = f(x),$$

say, by the monotone convergence theorem.

Next

$$\lim_{n\to\infty} T_{2n} = \lim_{n\to\infty}\frac{2}{\pi}\int_0^\infty \frac{(1-\cos tx)}{nx^2}\exp\left(-\frac{t}{n}\right)\varphi'(t)\,dt.$$

Since the integrand here tends to zero as n tends to infinity, and $\int |\varphi'(t)|dt < \infty$, it follows, by the dominated convergence theorem, that $\lim_{n\to\infty} T_{2n} = 0$.

Finally

$$\lim_{n\to\infty} T_{3n} = \lim_{n\to\infty}\frac{1}{\pi}\int_0^\infty \frac{(1-\cos tx)}{n^2x^2}\exp\left(-\frac{t}{n}\right)\varphi(t)\,dt =$$

$$= \lim_{n\to\infty}\frac{1}{\pi}\int_0^\infty \frac{(1-\cos nux)}{nx^2}\,\varphi(nu)\exp(-u)\,du$$

where $u = t/n$. Since the integrand above is bounded by $(2/\pi x^2)\exp(-u)$ and $\int \exp(-u)du < \infty$,

$$\lim_{n\to\infty} T_{3n} = \frac{1}{\pi}\int_0^\infty \left[\lim_{n\to\infty}\frac{(1-\cos nux)}{nx^2}\,\varphi(nu)\right]\exp(-u)\,du.$$

Thus

$$\lim_{n\to\infty} g_n(x) = f(x) = \frac{1}{\pi}\int_0^\infty \frac{(1-\cos tx)}{x^2}\, d(\varphi'(t)).$$

It follows from Lemma 2 that $\int f(x)dx = 1$.

Also, from Case 1, we have $\int g_n(x)dx = 1$. Thus the sequence $\{g_n\colon n \geqq 1\}$ of density functions converge pointwise for $x \neq 0$ to a density function $f(x)$. Thus by SCHEFFÉ's theorem [5] and the Helly—Bray Lemma ([3])

$$\varphi(t) = \lim_{n\to\infty} \varphi(t)\exp(-|t|/n) = \lim_{n\to\infty} \int \exp(itx)\, g_n(x)\, dx = \int \exp(itx) f(x)\, dx.$$

Consequently $\varphi(t)$ is the characteristic function of the density $f(x)$.

3. Remark

An approach similar to the one used here can also be used to prove the following important theorem in Fourier series (cf. [3], Theorem 7.3.1, p. 113):

THEOREM 2. *Let* $\{a_n\colon n \geqq 1\}$ *be a given sequence of real numbers and suppose that*

(i) a_n *is decreasing;*

(ii) $\lim_{n\to\infty} n\Delta a_n = 0$, where $\Delta a_n = a_n - a_{n+1}$;

(iii) $\sum_{n=0}^{\infty} (n+1)|\Delta^2 a_n| < \infty$, *where* $\Delta^2 a_n = \Delta(\Delta a_n)$.

Then there exists a function $f(x)$, $-\pi < x \leqq \pi$, *such that* $f(x) = f(-x)$, $\int_{-\pi}^{\pi} |f(x)|dx < \infty$ *and* $a_n = \int_{-\pi}^{\pi} f(x)\cos nx\, dx$.

4. Acknowledgment

I am grateful to R. A. Wijsman for his helpful comments.

REFERENCES

[1] EDWARDS, R. E.: *Fourier Series*, Holt, Rinehart, Winston, (1967).
[2] GOLDBERG, R. R.: *Fourier Transforms*. Cambridge University Press, (1965).
[3] LOEVE, M.: *Probability Theory*, 3rd Edition, Van Nostrand, (1963).
[4] LUKACS, E.: *Characteristic Functions*, Hafner, (1970).
[5] SCHEFFÉ, H.: A useful theorem for probability distributions, *Ann. Math. Statist.* **18** (1947), 434—438.

The University of New Mexico

(Received April 30, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 323—328.

EXTREMAL NON-(p, q)-HAMILTONIAN GRAPHS

by

Z. SKUPIEŃ and A. P. WOJDA

There is proved a theorem announced in [9] on the structure of some extremal non-strongly-(p, q)-Hamiltonian graphs G. Those graphs G are of maximal size $|E(G)|$ for a given order $|V(G)|$ and a given lower bound for the minimal degree $\delta(G)$ of vertices of G, and are not strongly (p, q)-Hamiltonian. This theorem generalizes a result due to ORE, BONDY and CHVÁTAL, which concerns extremal non-Hamiltonian graphs.

1. TERMINOLOGY AND NOTATION. We shall use the terminology and notation of [9]. For the sake of completeness, we recall some definitions. Throughout the paper, graph G means ordinary graph $G=\langle V, E\rangle$ with the vertex set $V(G)=V$ and the edge set $E(G)=E$ where $E(G)\subseteq\{xy\colon x, y\in V \text{ and } x\neq y\}$. We assume that n is the order of the graph G, $n=|V(G)|$. The number of edges incident to a vertex x in G, denoted by $d(x, G)$, is the degree of x in G, and

$$\delta(G) = \min\{d(x, G)\colon x\in V(G)\}.$$

The degree sequence of the graph G is a non-decreasing sequence formed from degrees of all vertices of G.

Given graphs G and H, the symbol $H\subseteq G$ means that H is a subgraph of G and also that G is a supergraph of H. If at the same time $V(H)=V(G)$ then H is a factor (spanned subgraph) of G and G is a counter-factor (spanned supergraph) of H. Given a subset V_1 of vertices of G, the symbol $G\setminus V_1$ denotes the subgraph of G that is induced by $V(G)\setminus V_1$. The maximal counter-factor of G is clearly a complete graph and is denoted by $\langle G\rangle$. A complete graph with n vertices is as usually denoted by K_n (possibly with a distinguishing superscript). Then $\overline{K}_n$, the complementary graph of K_n, is a graph of order n and size 0. A path is a graph consisting of vertices and edges of a simple open chain. A graph whose all components are paths is called a path-system.

We write $G*H$ to denote the join of G and H if and only if the graphs G and H are disjoint. Then $G*H$ contains G, H, and all possible edges connecting vertices of G to those of H. Though the operation of join $*$ is associative we assume that

$$G*H*F = (G*H)\cup(H*F)$$

for any three mutually disjoint graphs G, H, and F, the symbol on the left-hand side being written only if G, H and F denote mutually disjoint graphs.

In our previous paper [9] there are introduced the notions of (p, q)-Hamiltonian, strongly (p, q)-Hamiltonian, and strongly q-edge Hamiltonian graphs. These are generalizations of known notions of p-Hamiltonian, q-edge Hamiltonian, and Ha-

miltonian-connected graphs as well as generalizations or specializations of different known concepts of highly Hamiltonian-connected graphs (cf. [9]). The notions of p-Hamiltonian, q-edge Hamiltonian, and Hamiltonian-connected graphs are due to CHARTRAND, KAPOOR, and LICK [3], KRONK [6], and ORE [7], respectively.

A graph G is said to be strongly (p, q)-Hamiltonian $(0 \leqq p \leqq n-3,\ 0 \leqq q \leqq n-1,$ and $p+q \leqq n-1)$ if, for any set $V_1 \subset V(G)$ such that $|V_1| \leqq p$ and for any path-system S, $S \subset \langle G \setminus V_1 \rangle$, of size $|E(S)| \leqq q$, there is in $G \cup S$ a Hamiltonian circuit containing S. Now, substituting here the symbol $S \subset G \setminus V_1$ for $S \subset \langle G \setminus V_1 \rangle$ one obtains the condition for G to be (p, q)-Hamiltonian. Any [strongly] $(0, q)$-Hamiltonian graph is said to be [strongly] q-edge Hamiltonian.

Obviously, $(p, 0)$-Hamiltonian graphs are also strongly $(p, 0)$-Hamiltonian, and are precisely also p-Hamiltonian. Notice also that Hamiltonian-connected graphs of order $\geqq 3$ are precisely strongly $(0, 1)$-Hamiltonian (i.e., strongly 1-edge Hamiltonian) graphs.

2. A GENERALIZATION OF CHVÁTAL'S THEOREMS. One of the most interesting conditions considered in Hamiltonian graph theory is the following condition $\mathscr{C}_{ns}$ being a slight modification of a condition due to CHVÁTAL [4] (see also BERGE [1]).

$\mathscr{C}_{ns}$: For any (or equivalently: There is) an arrangement of vertices of a graph G with n vertices such that

$$d(x_1, G) \leqq d(x_2, G) \leqq \ldots \leqq d(x_n, G)$$

and for any integer i such that $1 \leqq i < \frac{n-s}{2}$,

$$\text{either} \quad d(x_i, G) > i+s \quad \text{or} \quad d(x_{n-s-i}, G) \geqq n-i.$$

Theorems proved in [9] imply the following result.

THEOREM 1. *Given non-negative integers p, q, and n such that $0 \leqq p+q \leqq n-3$ $(n \geqq 3)$, the condition $\mathscr{C}_{n, p+q}$ is sufficient for a graph G to be strongly (p, q)-Hamiltonian.*

Moreover, the condition $\mathscr{C}_{ns}$ is best possible in the following sense.

THEOREM 2. *The degree sequence $(d(x_i, G))_{i=1}^{n}$ of a graph G which does not satisfy the condition $\mathscr{C}_{ns}$ where $0 \leqq s \leqq n-3$, is majorized by one of the following sequences*

$$(2.1) \qquad \underbrace{k+s, k+s, \ldots, k+s}_{k \text{ times}}, \quad \underbrace{n-k-1, n-k-1, \ldots, n-k-1}_{n-s-2k \text{ times}}, \quad \underbrace{n-1, \ldots, n-1}_{k+s \text{ times}}$$

where k is an integer such that $1 \leqq k < \frac{1}{2}(n-s)$. Furthermore, the sequence (2.1) *is the degree sequence of the uniquely determined graph*

$$\overline{K}_k^{(1)} * K_{k+s}^{(3)} * K_{n-s-2k}^{(2)}$$

which is not (p, q)-Hamiltonian (and therefore is not strongly (p, q)-Hamiltonian as well) for any two non-negative integers whose sum is equal to s.

PROOF. Let $d_1 \leqq d_2 \leqq \ldots \leqq d_n$ be the degree sequence of a graph G which does not satisfy the condition $\mathscr{C}_{ns}$ with $0 \leqq s \leqq n-3$. Therefore there is an integer k such

that $1 \leqq k < \frac{1}{2}(n-s)$, $d_k \leqq k+s$, and $d_{n-s-k} < n-k$. Thus, the degree sequence of G is majorized by the sequence (2.1). Now, if $H=H(k)$ is a graph on n vertices with the degree sequence (2.1) then H has $k+s$ vertices of degree $n-1$ and k vertices of minimal degree $\delta(H)=k+s$. Since the deletion of all k vertices of degree $\delta(H)$ results clearly in a complete subgraph (a graph on $n-k$ vertices, each of which being of degree $n-k-1$), the graph H exists and is unique, that is,

$$H \equiv H(k) \equiv H(k; n, s) = \overline{K}_k^{(1)} * K_{k+s}^{(3)} * K_{n-s-2k}^{(2)}.$$

Now, H is not (p, q)-Hamiltonian for any two non-negative integers p and q such that $p+q=s \leqq n-3$. To prove this, let H' be a subgraph of H obtained by deletion of any p vertices of degree $n-1$. Hence, since $s-p=q$, we have

$$H' = \overline{K}_k^{(1)} * K_{k+q}^{(4)} * K_{n-s-2k}^{(2)} \quad \text{with} \quad K_{k+q}^{(4)} \subseteq K_{k+s}^{(3)}.$$

It clearly suffices to prove that H' is not q-edge Hamiltonian. To do this, let S be a path of length q in the complete subgraph $K_{k+q}^{(4)}$ of H'. We shall show that there is no Hamiltonian circuit of H' passing through S. For the proof one can show that the graph H'_{S-} (defined similarly as the graph G_{S-} introduced for the similar purpose in [8]) is not Hamiltonian. We shall use another argument.

Let H'' be a subgraph of H' obtained by deletion of all vertices of S (together with all incident edges). Thus

$$H'' = \overline{K}_k^{(1)} * K_{k-1}^{(5)} * K_{n-s-2k}^{(2)} \quad \text{with} \quad K_{k-1}^{(5)} \subset K_{k+q}^{(4)}.$$

The subgraph H'' has no Hamiltonian path because the deletion of all $k-1$ vertices inducing $K_{k-1}^{(5)}$ $(k-1 \geqq 0)$ results in a graph in which the number of components (being $k+1$) exceeds by two the number of deleted vertices. Therefore H' contains no Hamiltonian circuit passing through S. So H is clearly non-(p, q)-Hamiltonian, as stated in the theorem.

REMARK 1. Theorems 1 and 2 generalize two theorems of CHVÁTAL [4] and a related theorem of BERGE [1], p. 204.

3. THE MAIN RESULT. Now we shall prove the following theorem.

THEOREM 3. *Let n, s, p, q, and d_1 be integers such that*

$$n \geqq 3, \quad s = p+q \leqq n-1, \quad p \geqq 0, \quad q \geqq 0, \quad d_1 < \frac{n-s}{2},$$

and let

$$\alpha = \alpha_{ns} = \begin{cases} \dfrac{n-s+4}{6} & \text{for even} \quad n-s, \\ \dfrac{n-s+1}{6} & \text{otherwise.} \end{cases} \tag{3.1}$$

Now, if G is a non-strongly-(p, q)-Hamiltonian graph on n vertices with $\delta(G) \geqq d_1+s$ and of maximal size then, for

$$d := \max\{1, d_1\} < \frac{n-s}{2} \quad (\text{so } n-s \geqq 2d+1 \geqq 3),$$

either

$$G = \overline{K}_d * K_{d+s} * K_{n-s-2d} \quad \text{if} \quad d \leqq \alpha_{ns},$$

or

$$G = \overline{K}_\tau * K_{s+\tau} * K_{n-s-2\tau} \quad \text{if} \quad d \geqq \alpha_{ns}$$

and

$$\tau = \tau_{ns} = \left[\frac{n-s-1}{2}\right]. \tag{3.2}$$

Moreover, for $1 \leqq n-s \leqq 2$, *any graph on n vertices, except* K_n *iff* $p \leqq n-3$, *is non-*(p, q)*-Hamiltonian. Therefore, if G is a non-strongly-*(p, q)*-Hamiltonian graph, which is maximal with respect to the inclusion, then*

$$G = \overline{K}_2 * K_{n-2} \quad \text{if} \quad 1 \leqq n-s \leqq 2, \quad p \leqq n-3,$$
$$G = K_n \quad \text{if} \quad p = n-2 \quad \text{or} \quad p = n-1.$$

PROOF. First assume that $n-s \geqq 3$. Let G be a non-strongly-(p, q)-Hamiltonian graph on n vertices with $\delta(G) \geqq d_1+s$ and of maximal size. Now, if $d_1 \geqq \frac{n-s}{2}$ then $\delta(G) \geqq \frac{n+s}{2}$, and therefore G would be strongly (p, q)-Hamiltonian by the condition of DIRAC—PÓSA (cf. [9]), what is impossible. So $d_1 < \frac{n-s}{2}$. Then the graph G does not satisfy the condition $\mathscr{C}_{ns}$ of Chvátal type. Hence, by the maximality of $|E(G)|$ and by Theorem 2, there is an integer k such that

$$1 \leqq d \leqq k \leqq \tau_{ns},$$
$$G = H = H(k; n, s) = K_k^{(1)} * K_{k+s}^{(3)} * K_{n-s-2k}^{(2)}, \tag{3.3}$$

and $|E(H)|$ is maximal. Put

$$f(k; n, s) := |E(H)|.$$

Hence, by (3.3), $f(k; n, s)$ is the following square trinomial in k

$$f(k; n, s) = \binom{n-k}{2} + (k+s)k$$

and therefore

$$\Phi(d; n, s) := |E(G)| = \max\{f(k; n, s): \ d \leqq k \leqq \tau_{ns}\} \tag{3.4}$$
$$= \max\{f(d; n, s), f(\tau_{ns}; n, s)\}.$$

To obtain Φ in an explicit form, we consider the difference

$$2f(d; n, s) - 2f(\tau; n, s) = (\tau - d)A$$

where

$$A = A(d, n, s) = 2n - 2s - 1 - 3d - 3\tau,$$

and $\tau = \tau_{ns}$.

First suppose that $n-s$ is even. Then

$$\tau = \frac{n-s-2}{2} \geqq 1 \quad \text{and} \quad A = \frac{1}{2}(n-s-(6d-4)).$$

Since, by our assumptions on n, s, d, there is $\tau-d\geqq 0$, therefor fore, even $n-s$ and $n-s>2d\geqq 2$, we have

$$\Phi(d;n,s)=\begin{cases} f(d;n,s) & \text{for } 6d-4<n-s,\\ f(d;n,s)=f(3d-3;n,s) & \text{for } 6d-4=n-s>2,\\ f\left(\dfrac{n-s-1}{2};n,s\right) & \text{for } n-s<6d-4<3(n-s)-4.\end{cases} \tag{3.5}$$

For odd $n-s$, and $2\leqq 2d<n-s$, we have

$$\tau=\frac{n-s-1}{2},\quad A=\frac{1}{2}(n-s-(6d-1)),$$

and

$$\Phi(d;n,s)=\begin{cases} f(d;n,s) & \text{for } 6d-1<n-s,\\ f(d;n,s)=f(3d-1;n,s) & \text{for } 6d-1=n-s,\\ f\left(\dfrac{n-s-1}{2};n,s\right) & \text{for } n-s<6d-1<3(n-s)-1.\end{cases} \tag{3.6}$$

We can join the formulas (3.5) and (3.6). To do this, we convert A into the form $A=3(\alpha-d)$ with $\alpha=\alpha_{ns}$ defined by (3.1).

Hence, by (3.5), (3.6) and (3.2), for $1\leqq d\leqq\tau_{ns}$, we have

$$\Phi(d;n,s)=\begin{cases} f(d;n,s) & \text{if } d<\alpha_{ns},\\ f(d;n,s)=f(\tau_{ns};n,s) & \text{if } d=\alpha_{ns},\\ f(\tau_{ns};n,s) & \text{if } \alpha_{ns}<d\leqq\tau_{ns},\end{cases}$$

and, moreover, there is no other value of k, $k\geqq d$, such that $\Phi(d;n,s)=|E(G)|=f(k;n,s)$. So, either

$$G=H(d;n,s) \quad \text{if} \quad d\leqq\alpha_{ns},$$

or

$$G=H(\tau_{ns};n,s) \quad \text{if} \quad d\geqq\alpha_{ns}$$

as stated in Theorem 3.

The case $n-s\leqq 2$ is obvious. This completes the proof.

REMARKS. 1. If $p\leqq n-3$ then, for $n-s=1, 2$, the graph G is the same as for $n-s=3$, that is G is a complete graph with one edge deleted.

2. The graph G is not uniquely determined iff $1\leqq d=\alpha_{ns}<\dfrac{n-s}{2}$. By the definition (3.1) of α_{ns}, it is the case iff $2<n-s\equiv i \pmod 6$ with $i=2$ or $i=5$.

ORE in [7] described non-Hamiltonian graphs with maximal size. BONDY [2] (see also CHVÁTAL [5]) proved the completness of Ore's list. This result is the following corollary to Theorem 3.

COROLLARY 1. *If G is a non-Hamiltonian graph on $n(\geqq 3)$ vertices of maximal size, then*

$$G=K_1*K_1*K_{n-2} \quad (n\geqq 3),$$

or additionally

$$G=\overline{K}_3*K_2 \quad \text{if} \quad n=5.$$

One can state two following other corollaries to Theorem 3.

COROLLARY 2. *If* $s=p+q\leqq n-3$, $p\geqq 0$, $q\geqq 0$ *and* G *is a non-strongly-*(p, q)*-Hamiltonian graph on* n *vertices and of maximal size, then*

$$G = K_1 * K_{1+s} * K_{n-s-2}$$

or additionally

$$G = \overline{K}_3 * K_{2+s} \quad \text{if} \quad n-s=5.$$

COROLLARY 3. *If* G *is a non-Hamiltonian-connected graph on* n *vertices and of maximal size, then*

$$G = K_1 * K_2 * K_{n-3}$$

or

$$G = \overline{K}_3 * K_3 \quad \text{if} \quad n=6.$$

The two extremal non-Hamiltonian-connected graphs on $n=6$ vertices are shown in Fig. 1.

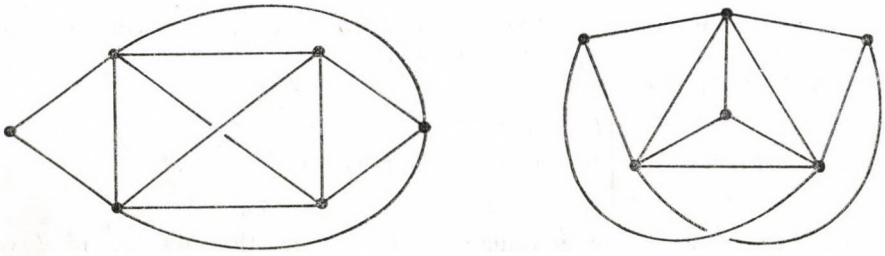


Fig. 1

REFERENCES

[1] BERGE, C.: *Graphs and hypergraphs,* North-Holland, Amsterdam and London, Amer. Elsevier, New York, 1973.
[2] BONDY, J. A.: Variations on the Hamiltonian theme, *Canad. Math. Bull.* **15** (1972), 57—62.
[3] CHARTRAND, G., KAPOOR, S. F. and LICK, D. R.: n-Hamiltonian graphs, *J. Comb. Theory* **9** (1970), 308—312.
[4] CHVÁTAL, V.: On Hamilton's ideals, *J. Comb. Theory,* **B 12** (1972), 163—168.
[5] CHVÁTAL, V.: New directions in Hamiltonian graph theory, in *New Directions in the Theory of Graphs, ed. by F. Harary,* Academic Press, New York and London, 1973, 65—95.
[6] KRONK, H. V.: Variations on a theorem of Pósa, in *The Many Facets of Graph Theory, ed. by G. Chartrand and S. F. Kapoor,* Lect. Notes Math. 110, Springer Verlag, 1969, 193—197.
[7] ORE, O.: Hamilton-connected graphs, *J. Math. Pures Appl.* **42** (1963), 21—27.
[8] SKUPIEŃ, Z. and WOJDA, A. P.: Sufficient conditions for λ-edge Hamiltonian graphs, *Bull. Acad. Polon. Sci. Sér. Math. Astronom. Phys.* **19** (1971), 391—396.
[9] SKUPIEŃ Z. and WOJDA, A. P.: On highly Hamiltonian graphs, *ibid.* **22** (5) (1974), 463—471.

Institute of Mathematics, Academy of Mining and Metallurgy, 30—059 Cracow

(Received April 30, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 329—336

ЧИСЛЕННОЕ ВЫДЕЛЕНИЕ ОГРАНИЧЕННЫХ РЕШЕНИЙ СИСТЕМ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

А. А. АБРАМОВ, Е. С. БИРГЕР, Н. Б. КОНЮХОВА, В. И. УЛЬЯНОВА

Summary. The systems of ordinary differential equations with singularities or on the infinite interval are discussed. The boundedness of solution is taken as a boundary condition at the singularity (or infinity). The method of numerical segregation of all sets of the solutions bounded in vicinity of the singularity discussed is proposed. Such segregation is performed by means of a power series expansion which is more simple and convenient for computation in respect to an expansion of the individual solutions. The obtained expansions are used in passing from the starting singular problem to a similar one considered at the finite interval without singularities.

1°. В различных областях математической физики встречаются задачи, сводящиеся к решению систем обыкновенных дифференциальных уравнений, которые имеют особенности в каких-либо точках или рассматриваются на бесконечном интервале. При этом роль граничного условия в особой точке или на бесконечности выполняет требование ограниченности решения. Типичные примеры дает квантовая механика, где ряд задач имеет следующий вид. Нужно найти собственные числа и собственные функции оператора Шредингера

$$H\psi = \lambda\psi; \tag{1}$$

здесь $\psi = \psi(\tau)$, $0 < \tau < \infty$, $H = -\frac{d^2}{d\tau^2} + U(\tau)$, $U(\tau) \to \infty$ при $\tau \to 0$ и имеет заданное поведение при $\tau \to \infty$. Нужно найти те λ, для которых существует нетривиальное решение $\psi(\tau)$, удовлетворяющее условиям: $|\psi(\tau)|$ ограничено при $\tau \to 0$ и при $\tau \to \infty$. При численном решении подобной задачи часто поступают следующим образом. Пользуясь разложением* $U(\tau)$ в окрестности $\tau = 0$, отыскивается разложение общего решения и выделяется то решение, которое является ограниченным. Тем самым получается разложение нужного решения в окрестности точки $\tau = 0$. Аналогично отыскивается разложение нужного решения при $\tau \to \infty$. Далее интервал $(0, \infty)$ заменяется конечным интервалом (τ_0, τ_∞), на котором и решается численно уравнение (1), а для граничных условий в τ_0 и τ_∞ используются упомянутые разложения, соответственно, в окрестности $\tau = 0$ и при $\tau \to \infty$.

*Здесь и всюду далее для простоты изложения не делаем различия между асимптотическими представлениями и сходящимися рядами. Для аккуратных формулировок нужно кроме учета такого различия еще зафиксировать непрерывность или определенную гладкость используемых функций.

В несложных случаях такой способ оказывается достаточным. Однако, он становится практически неприменимым, если необходимо решать громоздкую задачу.

Возьмем, например, линейную однородную систему

$$y' = \mathscr{A}(t)\,y, \quad y = \begin{Vmatrix} y_1 \\ \vdots \\ y_n \end{Vmatrix}, \quad 0 < t < T. \tag{2}$$

Пусть матрица $\mathscr{A}(t)$ имеет в окрестности точки $t=0$ разложение вида $\mathscr{A}(t)= =t^{-\tau}\ (\mathscr{A}_0+\mathscr{A}_1 t+\mathscr{A}_2 t^2+\ldots)$, $0<\tau$-целое. Пусть нужно найти решение уравнения (2), ограниченное при $t\to 0$ и удовлетворяющее каким-то заданным граничным условиям при $t\to T$. Попытка использовать прием, рассмотренный выше для уравнения (1), приведет к тому, что нужно будет найти разложения тех решений, которые ограничены при $t\to 0$. Это очень неприятная с вычислительной точки зрения задача. Во-первых, эти разложения отдельных решений очень громоздки и сложны по форме. Во-вторых, вид таких разложений зависит от значений некоторых параметров, которые в свою очередь должны бымь вычислены (например, при $\tau=1$ разложения, как праивло, имеют вид $y(t)= =t(y_0+ty_1+t^2y_2+\ldots)$, но могут появиться еще логарифмы, если среди собственных чисел матрицы $\mathscr{A}_0$ есть пары, отличающиеся на целое число). Ясно, что для величин вычисляемых на машине, часто невозможно сказать, являются ли они целыми или только близки к целым. Тем самым такой метод численно неустойчив. Для уравнения (2) поведение отдельных решений в окрестности особой точки оказывается очень капризным и плохо изучаемым численными методами. Разумеется, еще сложнее обстоит дело для систем нелинейных уравнений, где построение отдельных решений, ограниченных в окрестности особой точки, является еще более громоздкой и неустойчивой вычислительной задачей.

Поэтому для численного решения указанных задач желательно разработать методы, которые были бы лишены указанных выше недостатков.

В настоящем докладе излагаются некоторые результаты, полученные авторами и опубликованные в [1]—[20]*. Оказывается, для широкого класса систем имеет место следующая картина. Если не интересоваться поведением отдельных решений, а рассматривать все семейство решений, ограниченных в особой точке (или на бесконечности), то это семейство может быть задано соотношениями, которые имеют сравнительно простой вид и могут быть получены численно устойчивыми способами. Зафиксируем какую-либо точку, близкую к особой. Семейство решений, ограниченных в окрестности особой точки, может быть выделено некоторыми граничными условиями в упомянутой зафиксированной. Эти граничные условия получаются использованием разложений для семейства всех ограниченных решений. Таким образом, условие ограниченности решения в особой точке оказывается перенесенным в близкую точку. Если на другом конце интервала, на котором рассматривается система, также имеется особенность или интервал бесконечен, то тем же приемом

*Мы не рассматриваем сейчас других методов, развитых для более частных задач другими авторами (см., например, [21, 22]).

переносятся условия ограниченности решения в некоторую точку. Тем самым исходная задача сведется к решению соответствующей краевой задачи для исходной системы на конечном интервале без особенностей. Подчеркнем еще раз, что поведение отдельных решений внутри самого семейства ограниченных решений для поставленной задачи совершенно несущественно.

2°. Покажем, как именно осуществляется перенос условия ограниченности решения.

Пусть на интервале $T<t<\infty$ рассматривается система нелинейных уравнений

$$t^{-\tau}y' = f(t, y), \quad y = \begin{Vmatrix} y_1 \\ \vdots \\ y_n \end{Vmatrix}, \quad f = \begin{Vmatrix} f_1 \\ \vdots \\ f_n \end{Vmatrix}, \tag{3}$$

$0 \leqq \tau$-целое.

Пусть $f(t, y)$ имеет следующее поведение при больших t и малых $|y|$:

$$f(t, y) = \mathscr{A}y + g(t, y),$$

$$g(t, y) = \sum_{\substack{p_1 \geqq 0, \ldots, p_n \geqq 0,\ q \geqq 0 \\ p_1+p_2+\ldots+p_n+2q \geqq 2}} C_{q, p_1, \ldots, p_n} \frac{y_1^{p_1} \ldots y_n^{p_n}}{t^q}.$$

Пусть, наконец, матрица $\mathscr{A}$ не имеет собственных чисел на мнимой оси. Приведем матрицу $\mathscr{A}$ к квазидиагональному виду,

$$\begin{array}{c} \\ \uparrow \\ \varkappa \\ \downarrow \\ \uparrow \\ n-\varkappa \\ \downarrow \end{array} \left\| \begin{array}{c|c} \leftarrow \varkappa \rightarrow & \leftarrow n-\varkappa \rightarrow \\ \mathscr{A}_+ & 0 \\ \hline 0 & \mathscr{A}_- \end{array} \right\|,$$

где собственные числа $\mathscr{A}_+$ лежат в правой полуплоскости, а собственные числа $\mathscr{A}_-$ лежат в левой полуплоскости. В соответствующей системе координат обозначим

$$y = \left\| \begin{array}{c} y_+ \\ \hline y_- \end{array} \right\| \begin{array}{c} \uparrow \\ \varkappa \\ \downarrow \\ \hline \uparrow \\ n-\varkappa \\ \downarrow \end{array}, \quad f = \left\| \begin{array}{c} f_+ \\ \hline f_- \end{array} \right\| \begin{array}{c} \uparrow \\ \varkappa \\ \downarrow \\ \hline \uparrow \\ n-\varkappa \\ \downarrow \end{array}.$$

Тогда для достаточно больших t условие

$$|y| \to 0 \text{ при } t \to \infty$$

эквивалентно условию

$$y_+ = S(t, y_-), \quad S = \begin{Vmatrix} S_1 \\ \vdots \\ S_\varkappa \end{Vmatrix}.$$

Здесь функция $S(t, z)$ удовлетворяет уравнению в частных производных

$$t^{-\tau}\frac{\partial S}{\partial t} + \frac{\partial S}{\partial z} f_-(t, S(t, z), z) = f_+(t, S(t, z), z) \tag{4}$$

и условию

$$S(t, z) \to \beta(z) \quad \text{при} \quad t \to \infty,$$

где $\beta(z)$ есть решение уравнения $\frac{\partial \beta}{\partial z} f_-(\infty, \beta(z), z) = f_+(\infty, \beta(z), z)$, представимое рядом

$$\beta(z) = \sum_{\substack{p_1 \geqq 0, \ldots, p_{n-\varkappa} \geqq 0 \\ p_1 + p_2 + \ldots + p_{n-\varkappa} \geqq 2}} \beta_{p_1, \ldots, p_{n-\varkappa}} z_1^{p_1} \ldots z_{n-\varkappa}^{p_{n-\varkappa}}.$$

Эта функция $S(t, z)$ раскладывается в ряд по целым степеням $z, \frac{1}{t}$

$$S(t, z) = \sum_{\substack{p_1 \geqq 0, \ldots, p_{n-\varkappa} \geqq 0, \; q \geqq 0 \\ p_1 + p_2 + \ldots + p_{n-\varkappa} + 2q \geqq 2}} S_{q, p_1, \ldots, p_{n-\varkappa}} \frac{z_1^{p_1} \ldots z_{n-\varkappa}^{p_{n-\varkappa}}}{t^q}.$$

Коэффициенты $Sq, p_1, \ldots, p_{n-\varkappa}$ могут быть получены формальной подстановкой ряда для $S(t, z)$ в уравнение (4). При этом получается для этих коэффициентов последовательность систем линейных алгебраических уравнений вида

$$L\hat{S} = \varphi, \tag{5}$$

где $\hat{S}$-столбец из тех $S_{q, p_1, \ldots, p_{n-\varkappa}}$ при фиксированном q, для которых $\sum_{i=1}^{n-\varkappa} p_i$ одинакова φ-многочлен от предыду щих коэффициентов, а $\underset{L}{\lambda} = \underset{A_+}{\lambda} - \sum_{j=1}^{n-\varkappa} l_j \underset{A_-}{\lambda_j}$, $l_j \geqq 0$-целые, так что L-неособая матрица.

Отметим следующие удобства указанного разложения для практических вычислений.

1. В рассматриваемом случае граничное условие есть

$$|y| \to 0 \text{ при } t \to \infty.$$

Поэтому для достаточно больших t соотношение

$$y_+ = S(t, y_-)$$

нужно будет использовать лишь для малых $|y_-|$, а поэтому для практических вычислений можно и удобно пользоваться указанным разложением для $S(t, z)$.

2. Разложение имеет вид ряда по целым степеням независимо от таких частных факторов, как, например, Жорданова структура матрицы $\mathscr{A}$.

3. Коэффициенты разложения S ишутся из устойчиво решаемых систем линейных алгебраических уравнений. Для вычислений достаточно привести $\mathscr{A}$ к упомянутому квазидиагональному виду.

3°. Для того случая, когда система (3) линейна, формулы очень упрощаются. В частности, $S(t, z)$ принимает вид $S(t, z) = M(t)z + N(t)$, уравнение (4) заменяется системой обыкновенных дифференциальных уравнений для функций $M(t)$ и $N(t)$. Соответственно упрощаются формулы (5).

4°. В 2° было наложено ограничение: каждое собственное число матрицы $\mathscr{A}$ лежит или в правой или в левой полуплоскости. Это ограничение, грубо говоря, соответствует тому, что каждое отдельное решение при $t\to\infty$ или растет не медленнее некоторой экспоненты, или убывает не медленнее некоторой экспоненты. Промежуточная ситуация: $\mathscr{A}$ имеет собственные числа, лежащие на мнимой оси, создает большие дополнительные трудности для исследования. Удалось получить только некоторые результаты для линейных систем уравнений и, в частности, окончательный результат для того случая, когда $t=\infty$-правильная особая точка линейной системы.

5°. Укажем еще некоторые результаты, примыкающие к теме 2°—4°.

1. Пусть система

$$y''+p(t)\,y=0,\quad y=\begin{Vmatrix} y_1 \\ \vdots \\ y_n \end{Vmatrix}$$

рассматривается на интервале (α, β), где $\alpha<0<\beta$. Пусть в окрестности точки $t=0$ $p(t)$ имеет разложение вида

$$p(t)=\frac{1}{t}\,(p_0+tp_1+t^2p_2+\ldots).$$

Нужно выделить множество решений, допускающих аналитическое продолжение в нижнюю полуплоскость. Это выделение может быть осуществлено перенесением граничного условия из точки α в точку β по вещественной прямой. При этом в окрестности точки $t=0$ перенос осуществляется специальными разложениями, имеющими те же удобства, что и ряды, используемые в 2°.

2. Иногда в качестве ответов нужны не решения линейной системы (2), а однородные функционалы вида

$$\int_0^T y^*(t)\,Q(t)\,y(t)\,dt\Big/\int_0^T y^*(t)\,R(t)\,y(t)\,dt,$$

где $Q(t)$ и $R(t)$ — заданные квадратные матрицы. Можно дать способ вычисления таких функционалов, не требующий вычисления самой функции $y(t)$ и приводящий к решению некоторой вспомогательной системы дифференциальных уравнений, для которой решается задача Коши, устойчивая именно в том направлении, в котором и нужно решать эти уравнения.

3. В 2° встретилась алгебраическая задача, представляющая интерес в различных вычислительных задачах. Дана квадратная матрица $\mathscr{A}$, не имеющая собственных чисел на мнимой оси. Нужно привести ее к квазидиагональному виду

$$\begin{Vmatrix} \mathscr{A}_+ & 0 \\ 0 & \mathscr{A}_- \end{Vmatrix},$$

где все собственные числа матрицы $\mathscr{A}_+$ лежат в правой полуплоскости, в все собственные числа $\mathscr{A}_-$ — в левой.

Можно дать способы такого приведения, не требующие отыскания отдельных собственных чисел и собственных векторов матрицы, не зависящие от жордановой структуры $\mathscr{A}$ и не использующие комплексных чисел, если матрица $\mathscr{A}$ вещественна.

6°. Авторы применяли методы, изложенные выше, к решению многих задач квантовой механики, радиофизики, гидромеханики. Результаты получились хорошие.

ЛИТЕРАТУРА

[1] Абрамов, А. А.: О переносе условия ограниченности для некоторых систем обыкновенных линейных дифференциальных уравнений, *Журнал вычисл. мат. и мат. физ.* **1** (1961), № 4, 733—737.

[2] Абрамов, А. А.: О граничных условиях в особой точке для систем линейных обыкновенных дифференциальных уравнений, *Журнал вычисл. мат. и мат. физ.* **11** (1971), №1, 275—278.

[3] Абрамов, А. А., Тареев, Б. А., Ульянова, В. И.: Бароклинная неустойчивость в двуслойной фронтальной модели Кочина на бэта плоскости, *Известия АН СССР, физика атмосферы и океана* **8** (1972), № 2, 131—141.

[4] Абрамов, А. А., Тареев, Б. А., Ульянова, В. И.: Длинные волны и бароклинная неустойчивость на наклонн поверхности раздела, *Сб. трудов советско-французского симпозиума «Внутренние волны в океане», г. Новосибирск* (1972), 244—257.

[5] Абрамов, А. А., Тареев, Б. А., Ульянова, В. И.: Неустойчивость двуслойного геострофического течения с антисимметричным профилем скорости в верхнем слое, *Известия АН СССР, физика атмосферы и океана* **8** (1972), № 10, 1017—1028.

[6] Абрамов, А. А., Ульянова, В. И.: О решении уравнений для определения уровней энергии ионизированной молекулы водорода, *Журнал вычисл. мат. и мат. физ.* **1** (1961), № 2, 351—354.

[7] Абрамов, А. А., Ульянова, В. И.: О вычислении уровней энергии системы: два ядра- один электрон, *ТЭХ* **6** (1970), № 3, 384—386.

[8] Биргер, Е. С.: Об оценке погрешности замены условия ограниченности решения линейного дифференциального уравнения на бесконечном интервале. *Журнал вычисл. мат. и мат. физ.* **8** (1968), № 3, 674—678.

[9] Биргер, Е. С.: Об устойчивом вычислении некоторых функционалов от собственных функций задачи Штурма-Лиувилля на бесконечном интервале, *Журнал вычисл. мат. и мат. физ.* **8** (1968), № 5, 1126—1133.

[10] Биргер, Е. С.: О вычислении функционалов от собственных функций краевой задачи для системы линейных обыкновенных дифференциальнык уравнений, *Журнал вычисл. мат и мат. фис.* **13** (1973), № 1, 227—233.

[11] Биргер, Е. С., Ляликова, Н. Б.: О нахождении для некоторых систем обыкновенных дифференциальных уравнений решений с заданным условием на бесконечности, I, *Журнал вычисл. мат. и мат. физ.* **5** (1965), № 6, 979—990.

[12] Биргер, Е. С., Ляликова, Н. Б.: О нахождении для некоторых систем обыкновенных дифференциалных уравнений решений с заданным условием на бесконечности, II, *Журнал вычисл. мат. и мат. физ.* **6** (1966), № 3, 446—453.

[13] Биргер, Е. С., Конюхова, Н. Б.: К расчету молекул в одноэлектронном и одноцентровом приближении, *ТЭХ* **4** (1968), вып. 1, 29—36.

[14] Биргер, Е. С., Конюхова, Н. Б.: Численный расчет распространения радиоволн в вертикально-неоднородной тропосфере, *Радиотехника и электроника* **14** (1969), № 7, 1147—1156.

[15] Биргер, Е. С., Кербиков, Б. О., Конюхова, Н. Б., Шапиро, И. С.: О связанных квазиядерных состояниях системы $2N2\bar{N}$, *ЯФ* **17** (1973), вып. 1, 178—185.

[16] Конюхова, Н. Б.: О численном выделении стремящихся к нулю на бесконечности решений для некоторых двумерных нелинейных систем обыкновенных дифференциальных уравнений, *Журнал вычисл. мат. и мат. физ.* **10** (1970), № 1, 74—87.

[17] Конюхова, Н. Б.: О поведении решений внутри и вне устойчивого многообразия некоторых двумерных нелинейных систем обыкновенных дифференциальных уравнений, *Матем. заметки* **8** (1970), вып. 3, 285—295.
[18] Конюхова, Н. Б.: К решению краевых задач на бесконечном интервале для некоторых нелинейных систем обыкновенных дифференциальных уравнений с особенностью. *Журнал вычисл. мат. и мат. физ.* **10** (1970), № 5, 1150—1163.
[19] Конюхова, Н. Б.: О выделении устойчивых многообразий для некоторых нелинейных систем обыкновенных дифференциальных уравнений с особенностью, *Журнал вычисл. мат. и мат. физ.* **13** (1973), № 3, 609—626.
[20] Ульянова, В. И.: О перенесении граничных условий через некоторые особые точки, *Журнал вычисл. мат. и мат. физ.* **12** (1972), № 2, 528—532.
[21] Багмут, Г. И.: Разностные схемы высокого порядка точности для обыкновенного дифференциального уравнения с регулярной особенностью, *Журнал вычисл. мат. физ.* **9** (1969), № 1, 221—226.
[22] Russel, D. R.: Numerical solution of singular initial value problems, *SIAM J. Numer. Anal.* **7** (1970), No. 3, 399—417.

Вычисл Ценмр, ул Вавилова 40, Москва 117 333, СССР

(Посмупила 6-ого сенмядря 1974 г.)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 337—338.

A NOTE ON SEQUENTIAL WEAK COMPACTNESS

by

C. L. DE VITO

GROTHENDIECK, in his study of compactness in function spaces, constructs an example of a relatively sequentially compact set (defined below) whose closure is not even countably compact ([1], p. 170). An important setting for the study of sequential compactness is the weak topology of a locally convex topological vector space. Our purpose here is to show that, for the weak topology of a metrizable locally convex space, the closure of a relatively sequentially compact set is sequentially compact.

Let E denote a locally convex topological vector space over the field of real numbers. If we want to call attention to a specific, locally convex topology t on E we will write $E[t]$. The vector space of all continuous linear functionals on E will be denoted by E', the weak topology on E by $\sigma(E, E')$, and the weak* topology on E' by $\sigma(E', E)$. A sequence of points of E which converges for the topology $\sigma(E, E')$ will be called $\sigma(E, E')$-convergent. Similarly, we will speak of $\sigma(E, E')$-neighborhoods of a point, the $\sigma(E, E')$-closure of a set, etc.

DEFINITION. Let A be a subset of a locally convex space E. We say A is *sequentially weakly compact* if every sequence of points of A has a subsequence which is $\sigma(E, E')$-convergent to a point of A. We say A is *relatively sequentially weakly compact* if every sequence of points of A has a subsequence which is $\sigma(E, E')$-convergent to a point of E.

LEMMA. *Let $E[t]$ be a metrizable locally convex space and let M be a relatively sequentially weakly compact subset of E. If x_0 is a point in the $\sigma(E, E')$-closure of M, then there is a sequence of points in M which is $\sigma(E, E')$-convergent to x_0.*

PROOF. Since $E[t]$ is metrizable there is a countable fundamental system of neighborhoods of zero in this space. Let $\{V_n\}$ be such a system. Then each of the sets $V_n^0 = \{u \text{ in } E' \mid |ux| \leqq 1 \text{ for all } x \text{ in } V_n\}$ is $\sigma(E', E)$-compact [3; § 20, 9(4), p. 248] and clearly $E' = \bigcup_{n=1}^{\infty} V_n^0$. It follows that the point x_0 is in the $\sigma(E, E')$-closure of some countable subset N of M [3; § 24, 1(6), p. 312]. Let G be the closed linear subspace of E generated by the countable set $N \cup \{x_0\}$. Since $\sigma(G, G') = \sigma(E, E')|G$ [2; Corollary to Prop. 1, p. 262], the lemma will be proved if we can find a sequence of points in N which is $\sigma(G, G')$-convergent to x_0.

Now G with the topology $t|G$ is a separable metrizable locally convex space. Hence G' is $\sigma(G', G)$-separable [3; § 21, 3(5), p. 259]. Let $\{u_i\}$ be a countable $\sigma(G', G)$-dense subset of G'. For each positive integer n let x_n be a point of N such that

$|u_i(x_n-x_0)|<\frac{1}{n}$ for $1\leqq i\leqq n$. Since N is relatively sequentially weakly compact some subsequence $\{x_k\}$ of $\{x_n\}$ is $\sigma(G, G')$-convergent to a point y of G. Our inequality implies $\lim u_i x_k=u_i x_0$ for each fixed i and so $u_i x_0=u_i y$ for each i. But since $\{u_i\}$ is $\sigma(G', G)$-dense in G' we must have $x_0=y$.

THEOREM. *Let $E[t]$ be a metrizable locally convex space. Then the weak closure of a relatively sequentially weakly compact subset of E is sequentially weakly compact.*

PROOF. Let M be a relatively sequentially weakly compact subset of E and let $\{x_n\}$ be a sequence of points in the $\sigma(E, E')$-closure of M. For each n we can choose a sequence $\{y_{nm}\}$ of points of M which is $\sigma(E, E')$-convergent to x_n. Let H be the closed linear subspace of E generated by the countable set $\left[\bigcup_{n=1}^{\infty}\{y_{nm}|m=1, 2, \ldots\}\right]\cup$ $\cup\{x_n|n=1, 2, \ldots\}$. We can restrict our attention to the separable metrizable locally convex space $H[t|H]$ (see the last line in the first paragraph of the proof given above) and, furthermore, the space H' is $\sigma(H', H)$-separable (see the first two lines in the second paragraph of the proof given above). Let $\{u\ |i=1, 2, \ldots\}$ be $\sigma(H', H)$-dense in H'. For each fixed integer n the weak neighborhood $K_n=\{x \text{ in } H\,||u\ (x-x_n)|<\frac{1}{n}$ for $1\leqq i\leqq n\}$ of x_n contains a point of the sequence $\{y_{nm}\}$; let z_n denote such a point. In this way we obtain a sequence $\{z_n\}$ of points of M which, because of our hypothesis on M, has a subsequence $\{z_p\}$ which is weakly convergent to a point z of H. Clearly z is in the weak closure of M. We shall prove that the subsequence $\{x_p\}$ of $\{x_n\}$ has a subsequence $\{x_r\}$ which is weakly convergent to z. Let a circled, convex, weak neighborhood V of zero be given. We can choose a circled, convex, weak neighborhood U of zero such that $U+U\subset V$ [2; Theorem 1, p. 81]. For each integer n there is a point of the sequence $\{y_{nm}\}$ which is in $K_n\cap(U+x_n)$; let w_n be such a point. Now $\{w_p\}$ is in M and hence there is a subsequence $\{w_r\}$ of $\{w_p\}$ which has a weak limit w in H. Since $|u_i z_n-u_i w_n|\leqq|u_i z_n-u_i x_n|+|u_i x_n-u_i w_n|<\frac{2}{n}$ for $1\leqq i\leqq n$ it follows that, in particular, $u_i w=\lim u_i w_r=\lim u_i z_r=u_i z$ for every i. But since $\{u_i\}$ is $\sigma(H', H)$-dense in H' this says that $w=z$. Hence $z-x_r=(z-w_r)+(w_r-x_r)$ is in $U+U\subset V$ for all r sufficiently large.

The author would like to thank the refereee for his many helpful suggestions.

REFERENCES

[1] GROTHENDIECK, A.: Critères de Compacité dans les Espaces Fonctionnels Généraux, *Am. J. Math.* **74** (1952), 168—186.
[2] HORVÁTH, J.: *Topological Vector Spaces and Distributions, vol. I*, Addison—Wesley Publishing Company, Reading, Mass., 1966.
[3] KÖTHE, G.: *Topological Vector Spaces I*, Springer-Verlag, New York Inc., 1969.

Department of Mathematics, University of Arizona, Tucson, Arizona 85721, U.S.A.

(Received September 20, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 339—342.

A NONLINEAR PERIODIC BOUNDARY VALUE PROBLEM FOR A SYSTEM OF EQUATIONS OF THE SECOND ORDER

by

G. G. HAMEDANI and B. MEHRI

In this paper we prove existence and uniqueness theorems for the solutions of a nonlinear periodic boundary value problem for a system of second order equations.

Consider the vector boundary value problem

$$x''+f(t, x, x') = 0, \tag{1}$$

$$x(0)-x(\omega) = x'(0)-x'(\omega) = 0, \quad \omega\in[0, T], \tag{2}$$

where $x=(x_1, \ldots, x_n)$ is an n-dimensional vector;

$$f(t, x, y) = \big(f_1(t, x_1, \ldots, x_n, y_1, \ldots, y_n), \ldots, f_n(t, x_1, \ldots, x_n, y_1, \ldots, y_n)\big)$$

is a vector-valued function defined for $0\leqq t\leqq T$, $x, y\in R^n$.

Throughout this paper as norm of $x=(x_1, \ldots, x_n)$ and of $A=(a_{ik})$ it will be taken $\|x\|=\sum\limits_i |x_i|$ and $\|A\|=\sum\limits_{i,k} |a_{ik}|$, respectively.

In the sequel it is assumed that:

(A_1) $f(t, x, x')$ is a vector-valued, continuous function with domain $[0, T]\times R^{2n}$, $T>0$,

(A_2) there exist a matrix $A=(a_i\delta_{ik})_1^n$, $a_i>0$ for all i (δ_{ik} is the Kronecker delta), and a function $H(t, r)$ with the following properties:

1^0 $H(t, r)$ is piecewise continuous in $t\in[0, T]$, $r\geqq 0$, continuous in $t\in[0, T]$, and nondecreasing (for fixed t) with respect to $r\geqq 0$,

2^0 $\|Ax-f(t, x, x')\| \leqq H(t, \|x\|+\|x'\|)$, $t\in[0, T]$, $(x, x')\in R^{2n}$.

3^0 $\pi M_c\leqq 2\|A^{-1}\|^{-1}c$ for some constant $c>0$, where $M_c=\operatorname{Max}\limits_{t\in[0, T]} H(t, c)$. We have the following theorem.

THEOREM 1. *Under the above assumptions, there exists a positive real number* ω_0 $\left(0<\omega_0\leqq\dfrac{\pi}{a}, \text{ where } a=\operatorname{Max}\limits_i \sqrt{a_i}\right)$ *such that for each* $\omega\in\left[\omega_0, \dfrac{\pi}{a}\right]$ *the problem* (1), (2) *has at least one solution* $x(t)$ *continuous for* $0\leqq t\leqq\dfrac{\pi}{a}$ *satisfying* $\|x(t)\|\leqq c$, $0\leqq t\leqq\dfrac{\pi}{a}$.

PROOF. It is obvious that the problem (1), (2) is equivalent to

$$x''+Ax = Ax-f(t, x, x'), \tag{3}$$

$$x(0)-x(\omega) = x'(0)-x'(\omega) = 0. \tag{4}$$

Problem (3), (4) is equivalent to

$$x(t) = \int_0^{\omega} G(t, s)[Ax(s)-f(s, x(s), x'(s))]\, ds, \tag{5}$$

where $G(t, s)$ is Green's matrix for the problem (3), (4),

$$G(t, s) = \begin{cases} 2^{-1}(\sqrt{A})^{-1}\left[\sin\sqrt{A}\frac{\omega}{2}\right]^{-1}\cos\sqrt{A}\left(\frac{\omega}{2}+s-t\right) & \text{for } 0 \leqq s \leqq t \leqq \omega \\ 2^{-1}(\sqrt{A})^{-1}\left[\sin\sqrt{A}\frac{\omega}{2}\right]^{-1}\cos\sqrt{A}\left(\frac{\omega}{2}+t-s\right) & \text{for } 0 \leqq t \leqq s \leqq \omega \end{cases} \tag{6}$$

where $\omega \in \left(0, \frac{\pi}{a}\right]$, and the matrix functions $\sin\sqrt{A}\, t$ and $\cos\sqrt{A}\, t$ are defined by the matrix series (cf. [3], p. 118),

$$\sin\sqrt{A}t = \sum_{p=0}^{\infty}(-1)^p\frac{(\sqrt{A})^{2p+1}}{(2p+1)!}t^{2p+1}, \tag{7}$$

$$\cos\sqrt{A}t = \sum_{p=0}^{\infty}(-1)^p\frac{(\sqrt{A})^{2p}}{(2p)!}t^{2p}. \tag{8}$$

Using BIHARI's Theorem 1, [1], we shall prove that (5) has at least one solution with the desired properties. First we assume that $M_c>0$. It is easy to see that $G(t, s)$ is continuous and nonnegative in $0\leqq s\leqq t\leqq\omega$ and in $0\leqq t\leqq s\leqq\omega$. Since $\omega\in\left(0, \frac{\pi}{a}\right]$, it follows that $a.\frac{\omega}{2}\in\left(0, \frac{\pi}{2}\right]$. Since $a=\max_i \sqrt{a_i}$, it follows that $\sqrt{a_i}\ \frac{\omega}{2}\in\left(0, \frac{\pi}{2}\right]$ for each i, and hence $\sin\sqrt{a_i}\frac{\omega}{2}\geqq\frac{2}{\pi}\sqrt{a_i}\frac{\omega}{2}$ for each i involving

$$\|G(t, s)\| \leqq \sum_{i=1}^{n}\frac{1}{2\sqrt{a_i}}\cdot\frac{1}{\frac{2}{\pi}\sqrt{a_i}\frac{\omega}{2}} = \frac{\pi}{2\omega}\sum_{i=1}^{n}\frac{1}{a_i} = \frac{\pi\|A^{-1}\|}{2\omega}.$$

Now, we let

$$\omega_0 = \frac{\pi^2\|A^{-1}\|M_c}{2ac}.$$

From 3°, it follows that

$$0 < \omega_0 \leqq \frac{\pi}{a},$$

so that the interval $\left[\omega_0, \frac{\pi}{a}\right]$ actually exists. Moreover, it is easily seen that for any $\omega \in \left[\omega_0, \frac{\pi}{a}\right]$ we have

$$\|G(t, s)\| \leqq \frac{\pi \|A^{-1}\|}{2\omega_0} = \frac{a \cdot c}{\pi M_c}.$$

Thus $G(t, s)$ is continuous, bounded in $0 \leqq s \leqq t \leqq \omega$ and in $0 \leqq t \leqq s \leqq \omega$ for all $\omega \in \left[\omega_0, \frac{\pi}{a}\right]$, and hence the first hypothesis of Bihari's Theorem is satisfied.

From our assumptions it follows that the functions

$$\begin{matrix} F(t, x, x') = Ax - f(t, x, x') \\ H(t, \|x\| + \|x'\|) \end{matrix} \qquad t \in [0, T], \quad (x, x') \in R^{2n}$$

satisfy the second hypothesis of Bihari's Theorem. We also have

$$\frac{a \cdot c}{\pi M_c} \int_0^{\omega} H(s, c)\, ds \geqq \frac{a \cdot c}{\pi M_c} \cdot \frac{\pi}{a} M_c = c,$$

which shows that the third hypothesis of Bihari's Theorem is satisfied.

Finally, taking $z(t) \equiv 0$, we have shown that all the hypotheses of Bihari's Theorem are satisfied, and hence conclusion of his theorem tells us that at least one solution $x(t)$ of (5) (and hence of (1)) with the desired properties exists.

If $M_c = 0$, taking ω_0 to be any number in $\left(0, \frac{\pi}{a}\right]$, the proof trivially follows. Thus the proof of our Theorem 1 is completed.

COROLLARY 1. *Under the assumptions* 1^0, 2^0 *and the assumption*

$$4^\circ \quad \pi M_c \leqq a\left(2 - \frac{a}{b}\right) \left\|\sqrt{A^{-1}}\right\|^{-1} c \quad \textit{for some constant } c > 0,$$

where

$$M_c = \operatorname*{Max}_{t \in [0, T]} H(t, c)$$

and

$$b = \operatorname*{Min}_i \sqrt{a_i},$$

there exists a positive real number ω_0 $\left(\frac{\pi}{b} \leqq \omega_0 \leqq \frac{2\pi}{a}\right)$ *such that for each* $\omega \in \left[\frac{\pi}{b}, \omega_0\right]$ *there exists a solution* $x(t)$ *of* (1), (2) *satisfying* $\|x(t)\| \leqq c$, $t \in [0, T]$.

PROOF. Let $G(t, s)$ be as in Theorem 1. Since $|\sin u| \leqq -\frac{2}{\pi}u+2$ for $\frac{\pi}{2} \leqq u \leqq \pi$, we have for $\frac{\pi}{b} \leqq \omega \leqq \frac{2\pi}{a}$,

$$\|G(t, s)\| \leqq \sum_{i=1}^{n} \frac{1}{2\sqrt{a_i}} \cdot \frac{1}{-\frac{2}{\pi}\sqrt{a_i}\frac{\omega}{2}+2} \leqq \frac{\pi\left\|\sqrt{A^{-1}}\right\|}{2(-a\omega+2\pi)} = \Phi(\omega).$$

Now, we let

$$\omega_0 = \frac{2\pi}{a} - \frac{\pi^2\left\|\sqrt{A^{-1}}\right\| M_c}{a^2 c}.$$

From 4^0, it follows that

$$\frac{\pi}{b} \leqq \omega_0 \leqq \frac{2\pi}{a}.$$

Furthermore, for any $\omega \in \left[\frac{\pi}{b}, \omega_0\right]$ we have

$$\Phi(\omega) \leqq \Phi(\omega_0) = \frac{a \cdot c}{2\pi M_c}.$$

Now the remaining part of the proof is similar to that of Theorem 1.

COROLLARY 2. *If conditions of Theorem* 1 *or of Corollary* 1 *hold, and if the vector-valued function* $f(t, x, x')$ *is* ω*-periodic in* $t \in [0, T]$, *and satisfies the hypothesis of uniqueness with a given initial conditions, then* (1), (2) *has* ω*-periodic solution.*

REMARK. The results of this paper are generalizations of those of [2].

REFERENCES

[1] BIHARI, I.: Notes on a non-linear integral equation, *Studia Sci. Math. Hung.* **2** (1967), 1—6.

[2] HAMEDANI, G. G. and MEHRI, B.: Periodic boundary value problem for certain non-linear second order differential equation, *Studia Sci. Math. Hung.* **9** (1974,) 307—312.

[3] GANTMAKHER, F. R.: *The Theory of Matrices* (in Russian), Nauka, Moscow, 1966.

Arya-Mehr University of Technology, Tehran, Iran

(Received October 15, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 343—354.

ÜBER DIE STARKE CESÀRO-SUMMIERBARKEIT KONFORM-ÄQUIVALENTER REIHEN

von

R. WARLIMONT

§ 1. Einleitung

In der vorliegenden Arbeit werden die gemeinsam mit K.-H. INDLEKOFER begonnenen Untersuchungen ([5]) fortgesetzt.

Zunächst erläutern wir die im Titel genannten Begriffe.

1. Es sei $\varkappa \geqq 0$ und $0 \leqq \varrho \leqq 1$. Die Reihe

$$\sum_{m=0}^{\infty} a_m \quad (a_m \text{ komplex}) \tag{1}$$

heißt summierbar im Sinne $|C;\varkappa|^{\varrho}$, wenn ein komplexes a existiert derart, daß

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} \left| \binom{n+\varkappa}{n}^{-1} \sum_{m=0}^{n} \binom{n-m+\varkappa}{n-m} a_m - a \right|^{1/\varrho} = 0$$

gilt im Falle $0 < \varrho \leqq 1$ und wenn

$$\lim_{n\to\infty} \binom{n+\varkappa}{n}^{-1} \sum_{m=0}^{n} \binom{n-m+\varkappa}{n-m} a_m = a$$

gilt im Falle $\varrho = 0$.

Wir schreiben dann kurz

$$\sum_{m=0}^{\infty} a_m = a \quad \left(|C;\varkappa|^{\varrho}\right).$$

Offenbar ist also $|C;\varkappa|^0$ gerade die üblicherweise mit $(C;\varkappa)$ bezeichnete Cesàro-Summierbarkeit.

2. Es existiere

$$f(z) := \sum_{m=0}^{\infty} a_m z^m, \quad |z| < 1,$$

und es sei

$$z = \Phi(w), \quad \Phi(1) = 1$$

eine von der Identität verschiedene konforme Bijektivität der Einheitsscheibe auf sich.

Bekanntlich ist Φ von der Gestalt

$$\Phi(w) = \frac{1+\bar{\xi}}{1+\xi} \frac{w+\xi}{1+w\bar{\xi}} \quad \text{mit} \quad 0 < |\xi| < 1.$$

Die zur Reihe (1) bezüglich Φ konform-äquivalente Reihe

$$\sum_{m=0}^{\infty} a_m(\Phi) \tag{2}$$

ist dann erklärt durch die Beziehung

$$f(\Phi(w)) = \sum_{m=0}^{\infty} a_m(\Phi) w^m, \quad |w| < 1.$$

Nunmehr können wir unser Problem formulieren: Zu fest vorgegebenen $\varkappa_0 \geqq 0$, ϱ_0 $(0 \leqq \varrho_0 \leqq 1)$, Φ bestimme man die Menge

$$B(\varkappa_0, \varrho_0; \Phi)$$

all jener Tupel $(\varkappa, \varrho)$, $\varkappa \geqq 0$, $0 \leqq \varrho \leqq 1$, für welche die Aussage

„(1) summierbar im Sinne $|C; \varkappa_0|^{\varrho_0} \Rightarrow$ (2) summierbar im Sinne $|C; \varkappa|^{\varrho}$“ richtig ist.

TURÁN ([6]) erschloß diesen Problemkreis, als er (in unserer Terminologie)

$$(0, 0) \notin B(0, 0, \Phi)$$

nachwies.

Der Fall $\varrho_0 = 0$ wurde in [5] erledigt. Nunmehr lösen wir den Fall

$$0 \leqq \varrho_0 \leqq 1/2.$$

Der ausstehende Fall

$$1/2 < \varrho_0 \leqq 1$$

scheint schwierig zu sein.

Unser Resultat lautet: *Bedeutet $B(\varkappa_0)$ jenen abgeschlossenen Teilbereich des Halbstreifens*

$$\{(\varkappa, \varrho) | \varkappa \geqq \varkappa_0, \ 0 \leqq \varrho \leqq 1\},$$

der aus diesem durch Abtrennen des Dreiecks mit den Eckpunkten

$$(\varkappa_0, 0), \quad (\varkappa_0 + 1/2, 0), \quad (\varkappa_0, 1/2)$$

hervorgeht, so ist

$$(*) \qquad B(\varkappa_0, \varrho_0, \Phi) = \begin{cases} B(\varkappa_0) & \text{für} \quad 0 \leqq \varrho_0 < 1/2, \\ B(\varkappa_0) - (\varkappa_0 + 1/2, 0) & \text{für} \quad \varrho_0 = 1/2. \end{cases}$$

Die Unabhängigkeit des Bereiches von Φ darf nicht zu einer Fehldeutung des Resultates verleiten:

Schreiben wir kurz $B(\varkappa_0, \varrho_0)$ statt $B(\varkappa_0, \varrho_0, \Phi)$ und sind die Größen $\varkappa_0$, $\varkappa$, ϱ_0, ϱ mit

$$(\varkappa, \varrho) \notin B(\varkappa_0, \varrho_0)$$

vorgegeben, dann existiert zu jedem Φ eine im Sinne $|C; \varkappa_0|^{\varrho_0}$ summierbare Reihe (1), deren bezüglich Φ konformes Äquivalent (2) nicht summierbar ist im Sinne $|C; \varkappa|^{\varrho}$.

Ob es Reihen gibt, die sich gleichzeitig für alle Φ so verhalten, oder bei nur vorgegebenen $\varkappa_0$, ϱ_0 gleichzeitig für alle Φ und alle $(\varkappa, \varrho) \notin B(\varkappa_0, \varrho_0)$ ist ein Problem, zu dem wir hier nichts beitragen (man vgl. aber [4]!).

Beim Beweis von $(*)$ beschränken wir uns bequemlichkeitshalber auf den Fall $\varkappa_0 = 0$ und Φ mit zugehörigem reellen ξ.

Es gelten die folgenden Sätze.

SATZ 1. *Aus*

$$\sum_{m=0}^{\infty} a_m = a \quad \left(|C; 0|^{1/2}\right)$$

folgt

$$\sum_{m=0}^{\infty} a_m(\Phi) = a \quad \left(|C; 0|^{1/2}\right).$$

SATZ 2. *Zu jedem Φ existiert eine Reihe* (1) *mit den Eigenschaften*:

(1) *ist summierbar im Sinne* $|C; 0|^{1/2}$,

(2) *ist nicht summierbar im Sinne* $(C; 1/2)$.

SATZ 3. *Aus*

$$\sum_{m=0}^{\infty} a_m = a \quad \left(|C; 0|^{\varrho}\right) \quad \text{mit} \quad 0 \leqq \varrho < 1/2$$

folgt

$$\sum_{m=0}^{\infty} a_m(\Phi) = a \quad (C; 1/2).$$

Unser Satz 3 verschärft eines der beiden Hauptergebnisse ALPÁRS, nämlich [1], Théorème 1, der den Fall $\varrho = 0$ unseres Satzes wiedergibt. Obendrein ist Satz 3 wegen Satz 2 auch optimal.

Fügen wir unseren Sätzen das zweite Hauptergebnis ALPÁRS ([1], Théorème 2), hinzu, also

„Zu jedem $\varkappa$, $0 \leqq \varkappa < 1/2$, und Φ existiert eine konvergente Reihe (1), deren bezüglich Φ konformes Äquivalent (2) nicht summierbar ist im Sinne $(C; \varkappa)$“, so folgt schließlich $(*)$ für $\varkappa_0 = 0$ aus alledem, kombiniert mit den Konsistenzsätzen für das $|C; \varkappa|^{\varrho}$-Verfahren ([2], [3]; diese Details, die schon in [5] vorkamen, sind dort eingehender dargestellt).

§ 2. Bezeichnungen und Hilfssätze

Es sei also

$$\Phi(w) = \frac{w + \xi}{1 + w\xi} \quad \text{mit } \xi \text{ reell}, \quad 0 < |\xi| < 1.$$

Wir setzen noch

$$\eta := |\xi|, \quad q := \frac{1-\eta}{1+\eta}, \quad Q := q^{-1}, \quad \Psi(z) := \frac{1 - z\xi}{z - \xi},$$

$$V(r) := \max_{|w|=r} |\Phi(w)|, \quad v(r) := \min_{|w|=r} |\Phi(w)|, \quad M(r) := \max_{|z|=r} |\Psi(z)|.$$

Ohne Beweis notieren wir: Für $0<r\leqq 1$ ist

$$V(r)=\frac{r+\eta}{1+r\eta},\quad v(r)=\frac{|r-\eta|}{1-r\eta}. \tag{1}$$

Es ist

$$M(r)=\begin{cases}\dfrac{1+r\eta}{r+\eta} & \text{für} \quad r\geqq 1\\[2ex] \dfrac{1-r\eta}{|r-\eta|} & \text{für} \quad 0<r\leqq 1.\end{cases} \tag{2}$$

Mit einer höchstens von ξ abhängenden O-Konstanten gilt

$$r^{q}M(r)=1+O\big((r-1)^3\big) \quad \text{für} \quad r\to 1+ \tag{3}$$

$$r^{Q}M(r)=1+O\big((1-r)^3\big) \quad \text{für} \quad r\to 1-. \tag{4}$$

Setzt man abkürzend

$$s:=\operatorname{sign}(r-1)\xi,$$

so existiert eine höchstens von ξ abhängende Konstante $K>0$ derart, daß

$$|\Psi(re^{i\varphi})|\leqq M(r)\big(1-K|1-r|(1+s\cos\varphi)\big) \tag{5}$$

gilt für $1/2\leqq r\leqq 3/2$.

LEMMA 1. *Es sei $g(z)$ holomorph für $|z|<1$, es sei $p>0$ und*

$$\eta\leqq r<1.$$

Dann existieren höchstens von ξ abhängende Konstante $K_j>0$ $(j=1,2,3)$ und ein R, $0<R<1$ derart, daß

$$K_1\leqq\frac{1-R}{1-r}\leqq K_2$$

und

$$\int_0^{2\pi}\big|g\big(\Phi(re^{i\varphi})\big)\big|^p\,d\varphi\leqq K_3\int_0^{2\pi}|g(Re^{i\varphi})|^p\,d\varphi.$$

Nur der Fall $p=2$ wird benötigt, welcher schon in [5] erledigt worden ist. Weil das Lemma vielleicht an sich interessiert, nutzen wir die Gelegenheit, dieses in voller Allgemeinheit vorzutragen.

Unser BEWEIS beruht auf folgendem Sachverhalt ([7], VII, Theorem 7.12): Ist $h(z)$ holomorph für $|z|<1$ und ist $p>0$, so wächst

$$\int_0^{2\pi}|h(re^{i\varphi})|^p\,d\varphi$$

monoton in $0<r<1$.

Es sei $\eta\leqq r<\tilde{r}<1$. Man setze

$$G(t):=\int_0^{2\pi}\big|g\big(\Phi(te^{i\varphi})\big)\big|^p\,d\varphi.$$

Dann ist

$$G(r) \leqq (\tilde{r}-r)^{-1} \int_r^{\tilde{r}} G(t)\,dt \leqq \big(\eta(\tilde{r}-r)\big)^{-1} \int_r^{\tilde{r}} G(t)\,t\,dt =$$

$$= \big(\eta(\tilde{r}-r)\big)^{-1} \int_A \big|g\big(\Phi(w)\big)\big|^p\,du\,dv,$$

wobei

$$w = u+iv \quad \text{und} \quad A := \{w \mid r \leqq |w| \leqq \tilde{r}\}$$

gesetzt wurde.

Die Funktionaldeterminante der Substitution

$$w = \Phi^{-1}(z), \quad z = x+iy,$$

ist gleich

$$\left|\frac{d}{dz}\,\Phi^{-1}(z)\right|^2 \leqq Q^2,$$

so daß

$$G(r) \leqq \big(\eta(\tilde{r}-r)\big)^{-1} Q^2 \int_{\Phi(A)} |g(z)|^p\,dx\,dy.$$

Wegen

$$\Phi(A) \subseteq B := \{z \mid v(r) \leqq |z| \leqq V(\tilde{r})\}$$

ist

$$G(r) \leqq \big(\eta(\tilde{r}-r)\big)^{-1} Q^2 \int_B |g(z)|^p\,dx\,dy =$$

$$= \big(\eta(\tilde{r}-r)\big)^{-1} Q^2 \int_{v(r)}^{V(\tilde{r})} t\,dt \int_0^{2\pi} |g(te^{i\varphi})|^p\,d\varphi \leqq$$

$$\leqq Q^2 \eta^{-1} \frac{V(\tilde{r})-V(r)}{\tilde{r}-r'} \int_0^{2\pi} \big|g\big(V(\tilde{r})e^{i\varphi}\big)\big|^p\,d\varphi.$$

Wählt man etwa

$$\tilde{r} := \frac{1}{2}(1+r) \quad \text{und} \quad R := V(\tilde{r})$$

und beachtet noch (1), so folgt die Behauptung. □

Nun setze

$$s_n := \sum_{m=0}^{n} a_m,$$

$$t_n = t_n(\Phi) := \binom{n+1/2}{n}^{-1} \sum_{m=0}^{n} \binom{n-m+1/2}{n-m} a_m(\Phi)$$

und sei C ein Kreis mit Zentrum 0 und Radius r, $0<r<1$. Dann ist

$$\binom{n+1/2}{n} t_n = \frac{1}{2\pi i}\int_C \left(\sum_{m=0}^{\infty} a_m(\Phi) w^m\right)(1-w)^{-3/2} w^{-n-1}\,dw =$$

$$= \frac{1}{2\pi i}\int_C \left(\sum_{m=0}^{\infty} a_m(\Phi(w))^m\right)(1-w)^{-3/2} w^{-n-1}\,dw =$$

$$= \frac{1}{2\pi i}\int_C \left(\sum_{m=0}^{\infty} s_m(\Phi(w))^m\right)(1-\Phi(w))(1-w)^{-3/2} w^{-n-1}\,dw.$$

Setzen wir

$$J_{nm} = J_{nm}(\Phi) := \frac{1}{2\pi i}\int_C (\Phi(w))^m (1-\Phi(w))(1-w)^{-3/2} w^{-n-1}\,dw, \tag{6}$$

so gilt

$$t_n - a = \binom{n+1/2}{n}^{-1} \sum_{m=0}^{\infty} J_{nm}(s_m - a) \tag{7}$$

für beliebiges komplexes a. In (6) substitutieren wir $w=\Phi^{-1}(z)$ und bekommen

$$J_{nm} = (1-\xi)(1+\xi)^{-1/2}\frac{1}{2\pi i}\int_C z^m (\Psi(z))^n (1-z\xi)^{1/2}(1-z)^{-1/2}(z-\xi)^{-1}\,dz, \tag{8}$$

wobei C ein Kreis ist mit Zentrum 0 und Radius r, $\eta<r<1$.

Es folgen einige Abschätzungen von J_{nm}.

Es gilt ([1], § 5)

$$\sum_{m=0}^{\infty} |J_{nm}|^2 \ll \log n. \tag{9}$$

Es sei $A\geqq 1$ und

$$n_2 := [(q+\log^{-A} n)n], \quad n_3 := [(Q-\log^{-A} n)n].$$

Dann gilt ([1], § 9)

$$J_{nm} \ll \left(n\left(Q-\left(\frac{m}{n}-q\right)\right)\right) \quad \text{für} \quad n_2 \leqq m \leqq n_3. \tag{10}$$

Die O-Konstanten hängen höchstens von ξ und A ab.

Die Abschätzung (10) steht bei Alpár leider nicht explizit, sondern setzt sich aus in [1] § 9 verstreuten Einzelresultaten zusammen. Obendrein hat Alpár nur $A=1$; man überzeugt sich von der Richtigkeit von (10) für jedes $A\geqq 1$.

LEMMA 2. *Es sei*

$$n_4 := [Qn] \quad \textit{und} \quad 1 \leqq p < 2.$$

Dann gilt

$$\sum_{m=0}^{n_4} |J_{nm}|^p \ll n^{1-p/2}$$

mit einer höchstens von ξ und p abhängenden O-Konstanten.

BEWEIS. Mit

$$n_1 := [qn]$$

wird die abzuschätzende Summe

$$\leqq \left(\sum_{m=0}^{n_1} + \sum_{m=n_1}^{n_2} + \sum_{m=n_2}^{n_3} + \sum_{m=n_3}^{n_4} \right) |J_{nm}|^p := S_1 + S_2 + S_3 + S_4 .$$

Nach (9) ist

$$S_2 + S_4 \leqq \left(\sum_{m=0}^{\infty} |J_{nm}|^2 \right)^{p/2} \qquad (n \log^{-A} n)^{1-p/2} \ll n^{1-p/2},$$

wenn wir

$$A := p(2-p)^{-1}$$

wählen.

Nach (10) ist

$$S_3 \ll n^{1-p/2} \sum_{m=n_2}^{n_3} \frac{1}{n} \left(\left(Q - \frac{m}{n} \right) \left(\frac{m}{n} - q \right) \right)^{-p/2} \ll$$

$$\ll n^{1-p/2} \int_q^Q ((Q-x)(x-q))^{-p/2} \, dx \ll n^{1-p/2}.$$

Bei der *Abschätzung von* S_1 verwenden wir ALPÁRS Methode (man vgl. [1] § 6, insbesondere Fig. 1 auf p. 116).

Es sei C_1 der Kreis mit Zentrum 1 und Radius $r_1 := n^{-1}$ und es sei C_2 der Kreis mit Zentrum 0 und Radius $r_2 := 1 + n^{-1/3}$. Wir denken uns die z-Ebene längs der positiven reellen Achse aufgeschnitten. Unter C_3 werde verstanden das Intervall $[1+r_1, r_2]$, einmal aufgefaßt als Teil des „oberen Ufers", ein andermal als ein Teil des „unteren Ufers".

Aus (8) folgt bei passender Orientierung der C_j

$$J_{nm} = \int_{C_1} + \int_{C_2} + \int_{C_3} =: J_{nm1} + J_{nm2} + J_{nm3}.$$

Mit

$$S_{1j} := \sum_{m=0}^{n_1} |J_{nmj}|^p \quad (j = 1, 2, 3)$$

ist dann

$$S_1 \ll S_{11} + S_{12} + S_{13}.$$

Abschätzung von S_{11}. Es ist

$$J_{nm1} \ll r_1 r_1^{-1/2} (1+r_1)^m \left(\max_{z \in C_1} |\Psi(z)| \right)^n.$$

Wegen $\Psi(1) = 1$ ist

$$\max_{z \in C_1} |\Psi(z)| = 1 + O(r_1),$$

so daß

$$J_{nm1} \ll r_1^{1/2} (1+r_1)^m$$

und folglich

$$S_{11} \ll r_1^{p/2} \sum_{m=0}^{n_1} (1+r_1)^{mp} \ll r_1^{p/2-1} (1+r_1)^{n_1 p} \ll n^{1-p/2}.$$

Abschätzung von S_{12}. Es ist

$$J_{nm2} \ll r_2^m \int_0^{2\pi} |\Psi(r_2 e^{i\varphi})|^n |\varphi|^{-1/2} d\varphi.$$

Da die hieraus folgende triviale Abschätzung

$$J_{nm2} \ll r_2^m (M(r_2))^n$$

nur für $1 \leqq p \leqq 4/3$ zum Ziel führt, müssen wir uns mehr bemühen. Es werde $M(r_2)$ in α angenommen. Dann ist $\alpha = 0$ oder $= \pi$. Es sei $0 < \varepsilon \leqq \pi/2$. Nach (5) ist

$$J_{nm2} \ll r_2^m (M(r_2))^n \Big(\int_{\alpha-\varepsilon}^{\alpha+\varepsilon} |\varphi|^{-1/2} d\varphi + (1 - K(r_2-1)\varepsilon^2)^n \int_{\substack{\alpha-\pi \\ |\varphi-\alpha| \geqq \varepsilon}}^{\alpha+\pi} |\varphi|^{-1/2} d\varphi \Big).$$

Daher ist stets

$$J_{nm2} \ll r_2^m (M(r_2))^n (\varepsilon^{1/2} + \exp(-Kn(r_2-1)\varepsilon^2)).$$

Mit

$$\varepsilon := n^{-\delta/3}, \quad \text{wobei} \quad 0 < \delta < 1$$

wird

$$J_{nm2} \ll r_2^m (M(r_2))^n \varepsilon^{1/2} \tag{11}$$

und folglich ist

$$S_{12} \ll (M(r_2))^{np} \varepsilon^{p/2} \sum_{m=0}^{n_1} r_2^{mp} \ll (M(r_2))^{np} \varepsilon^{p/2} r_2^{n_1 p} (r_2-1)^{-1} \ll n^{-(\delta p)/6 + 1/3} (r_2^q M(r_2))^{np}$$

und wegen (3) ist dies $\ll n^{-(\delta p)/6+1/3}$. Wählen wir

$$\max(0, 3-4/p) < \delta < 1,$$

so ergibt sich

$$S_{12} \ll n^{1-p/2}.$$

Abschätzung von S_{13}. Es ist

$$J_{nm3} \ll \int_{1+r_1}^{r_2} r^m (M(r))^n (r-1)^{-1/2} dr.$$

Sind a, b derart, daß

$$a > 0, \quad b > (p-1)/p, \quad a+b = 1/2,$$

so folgt mit der Hölderschen Ungleichung

$$|J_{nm3}|^p \ll r_1^{p-1-bp} \int_{1+r_1}^{r_2} r^{mp} (M(r))^{np} (r-1)^{-ap} dr,$$

so daß

$$S_{13} \ll r_1^{p-1-bp} \int_{1+r_1}^{r_2} (r^q M(r))^{np} (r-1)^{-ap-1} dr.$$

Nach (3) ist

$$(r^q M(r))^{np} = (1 + O(r_2-1)^3)^{np} = O(1),$$

und folglich

$$S_{13} \ll r_1^{p-1-bp} \int_{1+r_1}^{\infty} (r-1)^{-ap-1}\, dr \ll n^{1-p/2}.$$

LEMMA 3. *Es sei*

$$r_3 := 1-n^{-1/3} \quad \textit{und} \quad \varepsilon := n^{-\delta/3} \quad \textit{mit} \quad 0<\delta<1.$$

Dann ist

$$J_{nm} \ll r_3^m \big(M(r_3)\big)^n \varepsilon^{1/2}$$

mit einer höchstens von ξ und δ abhängenden O-Konstanten.

BEWEIS. Ganz analog zu den Abschätzungen für J_{nm2}, die (11) lieferten.

§ 3. Die Beweise der Sätze 1, 2, 3

BEWEIS VON SATZ 1. Es sei

$$s_n := \sum_{m=0}^{n} a_m \quad \text{und} \quad s_n(\Phi) := \sum_{m=0}^{n} a_m(\Phi).$$

Man setze

$$f(z) := \sum_{m=0}^{\infty} a_m z^m, \quad F(z) := \sum_{n=0}^{\infty} (s_n - a) z^n, \quad G(w) := \sum_{n=0}^{\infty} \big(s_n(\Phi)-a\big) w^n.$$

Dann ist

$$G(w) = \big(f(\Phi(w))-a\big)(1-w)^{-1} =$$

$$= F\big(\Phi(w)\big)\big(1-\Phi(w)\big)(1-w)^{-1} = F\big(\Phi(w)\big)(1-\xi)(1+w\xi)^{-1}.$$

Folglich ist

$$|G(w)| \leqq Q\big|F\big(\Phi(w)\big)\big| \quad \text{für} \quad |w| \leqq 1$$

und somit

$$\sum_{n=0}^{\infty} |s_n(\Phi)-a|^2 r^{2n} = \frac{1}{2\pi} \int_0^{2\pi} |G(re^{i\varphi})|^2\, d\varphi \ll \int_0^{2\pi} \big|F\big(\Phi(re^{i\varphi})\big)\big|^2\, d\varphi$$

und also, nach Lemma 1,

$$\ll \int_0^{2\pi} |F(Re^{i\varphi})|^2\, d\varphi = 2\pi \sum_{n=0}^{\infty} |s_n - a|^2 R^{2n}.$$

Aus der vorausgesetzten Beziehung

$$\sum_{n=0}^{N} |s_n - a|^2 = o(N),$$

zusammen mit

$$K_1 \leqq \frac{1-R}{1-r} \leqq K_2,$$

folgt nunmehr

$$\sum_{n=0}^{N} |s_n(\Phi)-a|^2 = o(N).$$

BEWEIS VON SATZ 2. Nach [3] p. 129 existiert eine unendliche Reihe

$$\sum_{m=0}^{\infty} b_m$$

mit folgenden Eigenschaften:

(i) sie ist summierbar im Sinne $|C;0|^{1/2}$,

(ii) sie ist nicht summierbar im Sinne $(C;1/2)$.

Nun setze

$$a_m := b_m(\Phi^{-1}). \tag{1}$$

Dann ist

$$a_m(\Phi) = b_m. \tag{2}$$

Nach (i), (1) und Satz 1 ist

$$\sum_{m=0}^{\infty} a_m \quad \text{summierbar im Sinne} \quad |C;0|^{1/2}.$$

Nach (ii), (2) ist

$$\sum_{m=0}^{\infty} a_m(\Phi) \quad \text{nicht summierbar im Sinne} \quad (C;1/2).$$

BEWEIS VON SATZ 3. Es sei

$$\sum_{m=0}^{n} |s_m - a|^{1/\varrho} = o(n) \quad \text{wobei} \quad 0 < \varrho < 1/2.$$

Zu zeigen ist

$$\lim_{n\to\infty} t_n = a.$$

Nach (7) ist

$$|t_n - a| \ll n^{-1/2} \sum_{m=0}^{\infty} |J_{nm}|\,|s_m - a| \leqq n^{-1/2}\left(\sum_{m=0}^{n_4} + \sum_{m=n_4}^{\infty}\right)|J_{nm}|\,|s_m - a| =: n^{-1/2}(S_1 + S_2).$$

Nach Lemma 2 ist

$$S_1 \leqq \left(\sum_{m=0}^{n_4} |s_m - a|^{1/\varrho}\right)^{\varrho}\left(\sum_{m=0}^{n_4} |J_{nm}|^{1/(1-\varrho)}\right)^{1-\varrho} = o(n^{1/2}).$$

Nach Lemma 3 ist

$$S_2 \ll \big(M(r_3)\big)^n \varepsilon^{1/2} \sum_{m=n_4}^{\infty} |s_m - a|\, r_3^m \ll$$

$$\ll \big(M(r_3)\big)^n \varepsilon^{1/2} r_3^{n_4(1-\varrho)}(1-r_3)^{\varrho-1}\left(\sum_{m=n_4}^{\infty} |s_m - a|^{1/\varrho} r_3^m\right)^{\varrho}.$$

Mit

$$S(n) := \sup_{m \geqq n_4} (m+1)^{-1} \sum_{k=0}^{m} |s_k - a|^{1/\varrho}$$

wird die Summe rechts

$$\leqq (1-r_3)\sum_{m=n_4}^{\infty}\left(\sum_{k=0}^{\infty}|s_k-a|^{1/\varrho}\right)r_3^m \leqq$$

$$\leqq (1-r_3)S(n)\sum_{m=n_4}^{\infty}(m+1)r_3^m =$$

$$= (1-r_3)S(n)r_3^{n_4}\sum_{k=0}^{\infty}(n_4+k+1)r_3^k =$$

$$= (1-r_3)S(n)r_3^{n_4}\left(n_4(1-r_3)^{-1}+(1-r_3)^{-2}\right) \ll r_3^{n_4}nS(n),$$

so daß

$$S_2 \ll \left(r_3^{\varrho}M(r_3)\right)^n n^{-\delta/6+(1+2\varrho)/3}\left(S(n)\right)^{\varrho},$$

nach (4) also

$$S_2 \ll n^{-\delta/6+(1+2\varrho)/3}\left(S(n)\right)^{\varrho}.$$

Wählen wir

$$\max(0, 4\varrho-1) < \delta < 1,$$

so folgt

$$S_2 = o(n^{1/2}).$$

LITERATUR

[1] ALPÁR, L. Remarque sur la sommabilité des séries de Taylor sur leurs cercles de convergence, III, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5** (1960), 97—152.

[2] BOYD, A. V., HYSLOP, J. M.: A definition of strong Rieszian summability and its relationship to strong Cesáro summability, *Proc. Glasgow Math. Assoc.* **1** (1952—53), 94—99.

[3] GLATFELD, M.: On strong Rieszian summability, *Proc. Glasgow Math. Assoc.* **3** (1956—58), 123—131.

[4] HALÁSZ, G.: On the behaviour of Taylor series under conformal mappings of the circle of convergence, *Studia Scientiarum Math. Hung.* **1** (1966), 389—401.

[5] INDLEKOFER, K.-H., WARLIMONT, R.: Über die starke Cesáro-Summierbarkeit von Potenzreihen auf dem Rande des Konvergenzkreises, *Math. Nachr.* **63** (1974), 393—399.

[6] TURÁN, P.: A remark concerning the behaviour of a power series on the periphery of its convergence circle, *Publ. Inst. Math. Beograd* **12** (1958), 19—26.

[7] ZYGMUND, A.: *Trigonometric series, Vol. 1,* Cambridge, 1968.

Universität Regensburg, Fachbereich Mathematik, D—8400 Regensburg, BRD

(Eingegangen am 15. November, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 355—364.

ON OSCILLATION AND ASYMPTOTIC NONOSCILLATION OF FUNCTIONAL RETARDED EQUATIONS

by

B. SINGH

1. Introduction

Certain physical mechanisms whose performance is related to their past performance, have lead to the study of differential difference equations [1]. Systems representing such mechanisms are mathematically represented by differential equations carrying delay terms. One example of such a system is the pulse code modulation system in telephone network — SMITH [11].

The oscillatory behavior of these equations is an interesting phenomenon in regard to certain high speed mechanisms. A sudden burst of oscillations caused by the delay term may cause instability in them — see [8, P. 518].

The literature is rich with many oscillation criteria about differential-difference equations of various types—see [6, 9, 10, 12]. The goal in discovering these criteria so far has been to unify these results with similar results about ordinary differential equations when the delay term is set equal to zero. For example the equations

$$y^{(2n)}(t)+p(t)\,y(t)=0 \tag{1}$$

and

$$y^{(2n)}(t)+p(t)\,y\big(t-1(t)\big)=0 \tag{1a}$$

have all their bounded solutions oscillatory if

$$\int^{\infty} t^{2n-1}p(t)\,dt=\infty$$

provided $0\leqq 1(t)\leqq M$. See ONOSE [9] and SINGH [10]. When the delay term is bounded, then the two criteria are usually the same with the delay term playing no role. But the equation

$$y''(t)-y(t-\pi)=0 \tag{2}$$

due to BRADLEY [3] tells a different story. Its solutions Sin t and Cos t impart it a different behavior than the equation

$$y''(t)-y(t)=0 \tag{2a}$$

which is nonoscillatory. Obviously this variant behavior is being imposed upon equation (2) by the delay term.

Recently LADAS and LAKSHMIKANTHAN [7] showed that if

$$p(t)>0,\quad p'(t)\leqq 0\quad\text{and}\quad \tau^2 p(t)\geqq 2,$$

then the bounded solutions of the equation

$$y''(t)-p(t)y(t-\tau)=0 \tag{3}$$

are oscillatory. The corresponding equation obtained by taking $\tau=0$ and $p(t)\equiv 1$ iso bviously non-oscillatory.

The goal of theorem (2.1) in the second section of this paper is to improve and generalize this result of Ladas and Lakshmikanthan to a more general even order retarded equation

$$y^{(2n)}(t)-p(t)h(y_\tau(t))=0 \tag{4}$$

where $n\geqq 1$ is an integer and $y_\tau(t)\equiv y(t-\tau)$, $\tau>0$ is a positive constant. An example at the end of this section highlights the use of this theorem.

The third section of this paper deals with the non-oscillation of equation (4). Four theorems in this section actually characterize the non-oscillatory solutions of equation (4). Theorem (3.2) establishes a relationship between the oscillation of equation

$$y^{(2n)}(t)+p(t)h(y_\tau(t))=0 \tag{5}$$

and the nonoscillation of equation (4).

We now give definitions and assumptions that will hold in the remainder of this paper.

In what follows the term "solution" is going to apply to such functions as are solutions of equations under consideration and which are continuously extendable over some positive half line $[T, \infty]$, $T>0$.

We call a function on $C[T, \infty]$ oscillatory if it has sequence of zeros converging to ∞. Otherwise we call it non-oscillatory.

$p(t)$ is assumed to be positive and continuous on the whole real line R. τ will remain a positive constant. The notation for differentiation will be as

$$y^{(i)}(t)\equiv\frac{d^i}{dt^i}[y(t)],\quad i=1, 2, \ldots, 2n.$$

2. On oscillation

THEOREM 2.1. *Suppose that* $h: R\to R$ *is continuous and*

(6) (A) $\operatorname{Sgn} h(x)=\operatorname{Sgn} x,\quad h(-x)=-h(x)$,

(7) (B) *There exists a positive number* β *such that* $\dfrac{h(x)}{x}\geqq\beta$ *for all* x,

(8) (C) $\limsup\limits_{t\to\infty}\int\limits_{t-\tau}^{t}\dfrac{(t-s)^{2n-1}p(s)\,ds}{(2n-1)!}>\dfrac{1}{\beta}$;

then all bounded solutions of equation (4) are oscillatory.

REMARK. If we take $n=1$, $p'(t)\leqq 0$ and $h(x)\equiv x$, then we obtain equation (3) and related result of LADAS and LAKSHMIKANTHAN (7). In fact for $n=1$ from (C) we obtain $\beta=1$ and

$$\int_{t-\tau}^{t}(t-s)p(s)\,ds \geqq p(t)\int_{t-\tau}^{t}(t-s)\,ds = \frac{1}{2}\tau^2 p(t) > 1.$$

Thus this result of [7] is not only generalized but also improved in that differentiability of $p(t)$ and $p'(t)\leqq 0$ are not required.

PROOF OF THEOREM 2.1. Let $y(t)$ be a nonoscillatory solution of equation (4). Without any loss we can assume that there exists a $T_1>T$ such that for $t\geqq T_1$, $y(t)>0$ and $y_\tau(t)>0$. From equation (4) and condition (A) of this theorem,

$$y^{(2n)}(t) > 0, \quad t \geqq T_1. \tag{9}$$

Now suppose $y(t)$ is bounded. Since $y^{(2n-1)}(t)$ is increasing, $y^{(2n-1)}(t)$ cannot be eventually positive because this will cause $y(t)$ to be unbounded. The following conclusion is now obvious.

$$(-1)^i y_\tau^{(i)}(t) \geqq 0, \quad i = 1, 2, \ldots, 2n-1, \tag{10}$$

$$y_\tau(t) > 0 \quad \text{and} \quad y^{(2n)}(t) > 0 \quad \text{for some} \quad t \geqq T_2 > T_1.$$

By generalized mean value theorem and the fact that $y^{(2n)}(t)>0$ for $t\geqq T_2$ we have for $t\geqq s\geqq T_2$

$$y(s-\tau) \geqq y(t-\tau)+(s-t)y'(t-\tau)+\frac{(s-t)^2}{2!}y''(t-\tau)+\ldots+\frac{(s-t)^{2n-1}}{(2n-1)!}y^{(2n-1)}(t-\tau).$$

From where we have

$$y_\tau(s) \geqq y_\tau(t)+(s-t)y_\tau'(t)+\frac{(s-t)^2}{2!}y_\tau''(t)+\ldots+\frac{(s-t)^{2n-1}}{(2n-1)!}y_\tau^{(n-1)}(t) \tag{11}$$

from where we have

$$\frac{y_\tau(s)}{h(y_\tau(s))}h(y_\tau(s)) \geqq y_\tau(t)+(s-t)y_\tau'(t)+\frac{(s-t)^2}{2!}y_\tau''(t)+\ldots+\frac{(s-t)^{2n-1}}{(2n-1)!}y_\tau^{(2n-1)}(t). \tag{12}$$

From condition (B) and (12) we obtain

$$\frac{1}{\beta}h(y_\tau(s)) \geqq y_\tau(t)+(s-t)y_\tau(t)+\frac{(s-t)^2}{2!}y''(t)+\ldots+\frac{(s-t)^{2n-1}}{(2n-1)!}y_\tau^{(2n-1)}(t). \tag{13}$$

Multiplying (13) by $p(s)$ and integrating with respect to s between $t-\tau$ and t we have

$$\frac{1}{\beta}\int_{t-\tau}^{t}y^{(2n)}(s)\,ds = \frac{1}{\beta}\int_{t-\tau}^{t}p(s)h(y_\tau(s))\,ds$$

from where

$$\frac{1}{\beta}\int_{t-\tau}^{t} y^{(2n)}(s)\,ds \geqq y_\tau(t)\int_{t-\tau}^{t} p(s)\,ds + y'_\tau(t)\int_{t-\tau}^{t}(s-t)p(s)\,ds + \tag{14}$$
$$+ y''_\tau(t)\int_{t-\tau}^{t}\frac{(s-t)^2}{2!}p(s)\,ds + \ldots + y_\tau^{(2n-1)}(t)\int_{t-\tau}^{t}\frac{(s-t)^{2n-1}}{(2n-1)!}p(s)\,ds,$$

$$y^{(2n-1)}(t) - y_\tau^{(2n-1)}(t) > \beta y_\tau^{(2n-1)}(t)\cdot\frac{1}{(2n-1)!}\int_{t-\tau}^{t}(s-t)^{2n-1}p(s)\,ds, \tag{15}$$

since all the terms on the right hand side of (14) are non-negative due to conclusion (10) and the fact that $s \leqq t$.

From (15) we have

$$y^{(2n-1)}(t) > -\beta y_\tau^{(2n-1)}(t)\left[\int_{t-\tau}^{t}\frac{(t-s)^{2n-1}}{(2n-1)!}p(s) - \frac{1}{\beta}\right]. \tag{16}$$

But the left hand side of (16) is negative while due to condition (C) and conclusion (10), the right hand side is positive. This contradiction proves the result.

To justify the use of this theorem consider the following

EXAMPLE 1. For the equation

$$y^{(iv)}(t) - y(t-2\pi) = 0.$$

Sin t and Cos t are bounded oscillatory solutions. The conditions of the theorem are satisfied since

$$\int_{t-2\pi}^{t}\frac{(t-s)^{2n-1}}{(2n-1)!}p(s)\,ds = \int_{t-2\pi}^{t}\frac{(t-s)^3}{6}\,ds = \frac{1}{24}(2\pi)^4 > 1.$$

3. On non-oscillation

The next two theorems throw light on the asymptotic nature of the nonoscillatory solutions of equation (4).

THEOREM 3.1. *Suppose for every function $F(t)$ that becomes negative eventually and every $c>0$, there exists points $t_c > c$ such that*

$$F^{(2n)}(t_c) + p(t_c)h\big(F_\tau(t_c)\big) \leqq 0.$$

Further suppose that conditions (A) and (B) of theorem (2.1) hold and that $h'>0$. Let $y(t)$ be a non-oscillatory solution of (4). Then either $|y(t)|\to\infty$ or $|y(t)|\to 0$.

PROOF. As in the proof of theorem (2.1) we can assume that $y(t)$ is eventually positive and that there exists T_2 (same as before in theorem (2.1)) such that if $y(t)$ is bounded then for $t \geqq T_3$, (10) holds. Since all the derivatives are monotonic

$y(t)\to\alpha$ as $t\to\infty$ where α is finite or infinite. We only need to show that if α is finite then $\alpha=0$. Suppose to the contrary that $\alpha>0$. Consider the function

$$F(t) = y(t)-2\alpha. \tag{17}$$

Since $y(t)$ decreases to α, there exists a large $T_3>T_2$ such that for $t\geqq T_3$, $y(t)\leqq \leqq\frac{3}{2}\alpha$.

Hence for $t\geqq T_3$, $F(t)$ in (17) is negative. Now from (17)

$$F^{(2n)}(t)+p(t)h\big(F_\tau(t)\big) = y^{(2n)}(t)+p(t)h\big(y_\tau(t)-2\alpha\big). \tag{18}$$

Making use of equation (4) in (18) we have

$$\begin{aligned} F^{(2n)}(t)+p(t)h\big(F_\tau(t)\big) &= p(t)\big[h\big(y_\tau(t)-2\alpha\big)+h\big(y_\tau(t)\big)\big] = \\ &= 2p(t)\left[\frac{h\big(y_\tau(t)\big)+h\big(y_\tau(t)-2\alpha\big)}{2y_\tau(t)-2\alpha}\right]\big(y_\tau(t)-\alpha\big) = \\ &= 2p(t)\left[\frac{h\big(y_\tau(t)\big)-h\big(2\alpha-y_\tau(t)\big)}{y_\tau(t)-\big(2\alpha-y_\tau(t)\big)}\right]\big(y_\tau(t)-\alpha\big)\geqq 0, \end{aligned}$$

since $y_\tau(t)$ decreases to α and the expression in square bracket being positive by mean value theorem and the fact $h'>0$.

But for $T_3=c$, this is a contradiction to the hypothesis of the theorem and the proof is complete.

REMARK. The above theorem suggests the following theorem.

THEOREM 3.2. *Suppose conditions (A) and (B) of theorem (2,1) hold. Let*

$$\int^{\infty} t^{2n-1}p(t)\,dt=\infty \quad \textit{and} \quad h'>0.$$

Let $y(t)$ be a non-oscillatory solution of equation (4). Then either $|y(t)|\to\infty$ or $|y(t)|\to 0$ as $t\to\infty$.

PROOF. We proceed as in the proof of theorem 2.1. Let T_2 be large enough so that $y(t)$ and $y_\tau(t)$ are positive (without any loss of generality) for $t\geqq T_2$. By equation (4), $y^{(2n)}(t)>0$. Thus all the derivatives are monotonic and hence $y(t)$ approaches a limit finite or infinite. We only need to show that if $y(t)\to\alpha$, $0\leqq\alpha<\infty$, then $\alpha=0$. Suppose that $\alpha>0$. Since $y(t)$ is bounded, conclusion (10) of the proof of theorem (2.1) holds. Thus $y'(t)<0$ and $y(t)$ decreases to α. As in the proof of theorem (3.1) we define

$$F(t) = y(t)-2\alpha$$

and arrive at

$$F^{(2n)}(t)+p(t)h\big(F_\tau(t)\big)\geqq 0 \tag{19}$$

for $t\geqq T_3$. Since $F(t)$ is bounded and negative for $t\geqq T_3$, (19) implies

$$F(t)<0,\ F'(t)\leqq 0,\ F''(t)\geqq 0,\ F'''(t)\leqq 0,\ \ldots,\ F^{(2n-1)}(t)\leqq 0,\ F^{(2n)}(t)>0. \tag{20}$$

Multiplying (19) by $(t-\tau)^{2n-1}$ and dividing by $F_\tau(t)$ we get

$$\frac{F^{(2n)}(t)(t-\tau)^{2n-1}}{F(t-\tau)}+(t-\tau)^{2n-1}p(t)\cdot\frac{h(F_\tau(t))}{F_\tau(t)}\leqq 0. \tag{20}$$

Integrating (20) between T_3 and t and making use of

$$\frac{h(F(t))}{F_\tau(t)}\geqq\beta$$

we get

$$\frac{F^{(2n-1)}(t)(t-\tau)^{2n-1}}{F(t-\tau)}-\frac{F^{(2n-1)}(T_3)(T_3-\tau)^{2n-1}}{F(T_3-\tau)}- \tag{21}$$

$$-\int_{T_3}^{t}\frac{F^{(2n-1)}(s)(2n-1)(s-\tau)^{2n-2}}{F(s-\tau)}\,ds+$$

$$+\int_{T_3}^{t}\frac{F^{(2n-1)}(s)(s-\tau)^{2n-1}F'(s-\tau)}{F^2(s-\tau)}\,ds+\beta\int_{T_3}^{t}(s-\tau)^{2n-1}p(s)\,ds\leqq 0.$$

Now in (21), the first term is non-negative, the second is finite, fourth is positive due to (20) and last term tends to ∞ as $t\to\infty$, it follows from inequality (21) that

$$\lim_{t\to\infty}\int_{T_3}^{t}\frac{F^{(2n-1)}(s)(s-\tau)^{2n-2}}{F(s-\tau)}\,ds=\infty. \tag{22}$$

Now

$$\int_{T_3}^{t}\frac{F^{(2n-1)}(s)(s-\tau)^{2n-2}}{F(s-\tau)}\,ds=\frac{F^{(2n-2)}(t)(t-\tau)^{2n-2}}{F(t-\tau)}-$$

$$-\frac{F^{(2n-2)}(T_3)(T_3-\tau)^{2n-2}}{F(T_3-\tau)}-\int_{T_3}^{t}\frac{F^{(2n-2)}(s)(2n-2)(s-\tau)^{2n-3}}{F(s-\tau)}\,ds+$$

$$+\int_{T_3}^{t}\frac{F^{(2n-2)}(s)(s-\tau)^{2n-2}F'(s-\tau)}{F^2(s-\tau)}\,ds$$

in which the first term is negative, second is finite, the fourth is negative. Therefore for (22) to hold

$$\lim_{t\to\infty}\int_{T_3}^{t}\frac{F^{(2n-2)}(s)(s-\tau)^{2n-3}\,ds}{F(s-\tau)}=+\infty. \tag{23}$$

Proceeding in this manner we find, that for inequality (21) to hold we must have

$$\lim_{t\to\infty}\int_{T_3}^{t}\frac{F'(s)}{F(s-\tau)}\,ds=-\infty. \tag{24}$$

Now by (20)

$$\int_{T_3}^{t} \frac{F'(s)}{F(s-\tau)}\,ds \geqq \int_{T_3}^{t} \frac{F'(s-\tau)}{F(s-\tau)}\,ds = \ln|F(t-\tau)| - \ln|F(T_3-\tau)| < \infty$$

a contradiction since $|F(t-\tau)|$ is bounded and increasing. The proof is now complete.

To justify this theorem, consider the following examples.

EXAMPLE 2. The equation

$$y^{(iv)}(t) - e^{-\pi} y(t-\pi) = 0 \tag{25}$$

has the solution $y=e^{-t}$ which goes to zero. Now from our theorem

$$e^{-\pi} \int_{T_3}^{\infty} (s-\pi)^3\,ds = \infty$$

satisfying the required condition.

EXAMPLE 3. The equation

$$y^{(iv)}(t) - e^{\pi} y(t-\pi) = 0 \tag{26}$$

has $y=e^{t}$ as solution that goes to ∞. Again the conditions of the theorem are satisfied.

COROLLARY 3.1. *Suppose all bounded solutions of equation*

$$y^{(2n)}(t) + p(t)h\big(y_\tau(t)\big) = 0 \tag{27}$$

are oscillatory. Further suppose that conditions of theorem (3.2) are satisfied. Then the conclusion of (3.2) is true.

PROOF. By theorems (1) and (2) of SINGH [10], it follows that

$$\int^{\infty} t^{2n-1} p(t)\,dt = \infty. \tag{28}$$

The proof of theorem (3.2) remains true when

$$\int^{\infty} (t-\tau)^{2n-1} p(t)\,dt = \infty$$

is replaced by (28).

REMARK. The last two theorems of this section assumed that h be differentiable and $h'>0$. In the next theorem this restriction is dropped and a stronger result is proved.

THEOREM 3.3. *Suppose conditions (A) and (B) of theorem (2.1) hold. Let*

$$\int^{\infty} t\big(\beta p(t) - 1\big)\,dt = \infty. \tag{29}$$

Suppose $y(t)$ is a non-oscillatory solution of equation (4). Then either $|y(t)| \to 0$ or $|y(t)| \to \infty$ as $t \to \infty$.

PROOF. Without any loss, we can assume that $y(t)$ is eventually positive. Since from equation (4), $y^{(2n)}(t)>0$, all the preceeding derivatives are monotonic. This means that $y(t)$ approaches a limit as $t\to\infty$. The only thing we need to show that if the limit is finite, then it must be zero. Let then $\lim_{t\to\infty} y(t)=\alpha>0$.

Due to boundedness of $y(t)$, conclusion (10) holds and $y^{(2n-1)}(t)<0$ eventually. Let T_3 be the same as in proof of theorem (3.2) so that for $t\geqq T_3$, conclusion (4) is true.

Define

$$z(t)=\frac{y^{(2n-1)}(t)\,t}{y(t-\tau)},\quad t\geqq T_3. \tag{30}$$

Then $z(t)\leqq 0$. Also observe that $y^{(2n-1)}(t)\to 0$ due to (10) and non-negativeness of $y(t)$. If $\alpha>0$ then

$$\lim_{t\to\infty}\frac{y^{(2n-1)}(t)}{y(t-\tau)}=0. \tag{31}$$

Differentiating (30) we obtain

$$z'(t)=\frac{y^{(2n)}(t)\,t}{y(t-\tau)}-\frac{y^{(2n-1)}(t)\,y'(t-\tau)\,t}{y^2(t-\tau)}+\frac{y^{(2n-1)}(t)}{y(t-\tau)}=$$

$$=\frac{tp(t)h\big(y(t-\tau)\big)}{y(t-\tau)}-\frac{y^{(2n-1)}(t)\,y'(t-\tau)\,t}{y^2(t-\tau)}+\frac{y^{(2n-1)}(t)}{y(t-\tau)}\geqq$$

$$\geqq \beta p(t)\,t-\frac{y'(t-\tau)\,y^{(2n-1)}(t)\,t}{y^2(t-\tau)}+\frac{y^{(2n-1)}(t)}{y(t)}.$$

Now

$$0<\frac{y'(t-\tau)\,y^{(2n-1)}(t)}{y^2(t-\tau)}\leqq 1$$

in view of (31) for $t\geqq T_4>T_3$. Thus for $t\geqq T_4$

$$z'(t)-\frac{y^{(2n-1)}(t)}{y(t)}\geqq t\big(\beta p(t)-1\big). \tag{32}$$

Integrating (32) between T_4 and t we get

$$z(t)-z(T_4)-\frac{1}{\alpha}\left[y^{(2n-2)}(t)-y^{(2n-2)}(T_4)\right]\geqq\int_{T_4}^{t} s\big(\beta p(s)-1\big)\,ds. \tag{33}$$

The left hand side of (33) is either negative or bounded since $\alpha>0$. Since the right hand side tends to $+\infty$ as $t\to\infty$, this is the required contradiction that proves the theorem.

The proof of theorem (3.3) suggests the following

THEOREM 3.4. *Let* $\varphi(t)$ *be a positive differentiable function which has a positive bounded derivative on some positive halfline* $[T_4, \infty]$. *Suppose*

$$\int_{T_4}^{\infty} \varphi(t)(\beta p(t)-1)\,dt = \infty.$$

Then the conclusion of theorem (3.3) is true.

PROOF. We set

$$z(t) = \frac{y^{(2n-1)}(t)}{y(t-\tau)}\,\varphi(t)$$

and proceed as in theorem (3.3). The conclusion follows.

EXAMPLE 4. Consider example 3 as the following equation.

$$y^{(VI)}(t) - e^{\pi} y(t-\pi) = 0$$

which has e^t as a non-oscillatory solution.

Here let

$$\varphi(t) = t, \quad \beta = 1;$$

then

$$\int_{T_4}^{\infty} t(\beta p(t)-1)\,dt = \int_{T_4}^{\infty} t(e^{\pi}-1)\,dt = \infty$$

thus satisfying conditions of theorem (3.3) or (3.4).

REFERENCES

[1] BELLMAN, R., and COOKE, K.: *Differential-Difference Equations*, Academic Press, New York, 1963.

[2] BHATIA, NAM P.: Some oscillation theorems for second order differential equations, *J. Math. Anal. Appl.* **15** (1966), 442—446.

[3] BRADLEY, JOHN S.: Oscillation theorems for second order delay equations, *J. Differential Equations* **8** (1970), 397—403.

[4] BURTON, T. and GRIMMER, R.: On the asymptotic behavior of solutions of $x''(t)+a(t)x'(t)=$ $=e(t)$, *Pac. J. Math.* Vol. **41**, No. 1 (1972), 77—88.

[5] ELIASON, STANLEY B.: A Lyapunov inequality for a certain second order nonlinear differential equation, *J. London Math. Soc.* (2), **2** (1970), 467—472.

[6] HAMMET, MICHAEL E.: Non-oscillation properties of a non-linear differential equation, *Proc. Amer. Math. Soc.* Vol. **30**, No. 1 (1971).

[7] LADAS, G., and LAKSHMIKANTHAN, S.: Oscillations Caused by Retarded actions, *J. Applicable Analysis*. (To appear.)

[8] MINORSKY, N.: *Non-linear Oscillations*, Van Nostrand, Princeton, NJ, 1962.

[9] ONOSE, H.: Oscillatory properties of ordinary differential equations of arbitrary order, *J. Differential Equations* **7** (1970), 454—458.

[10] SINGH, B.: A necessary and sufficient condition for the oscillation of even order non-linear delay differential equation, *Canad. J. Math.* (To appear.)

[11] SMITH, R. A.: Solution estimates for certain linear-differential systems, *J. London Math. Soc.* (2), **2** (1970), 581—588.

[12] TRAVIS, C. C.: Oscillation theorems for second order differential equations with functional arguments, *Proc. Amer. Math. Soc.* Vol. **31**, No. 1, 1972.

Dept. of Mathematics, University of Wisconsin, Manitowoc County, Manitowoc, WI 54220

(Received November 30, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 365—380.

THE ALGEBRAIC DERIVATIVE AND INTEGRAL IN THE DISCRETE OPERATIONAL CALCULUS, II

by

T. FÉNYES and P. KOSIK

Introduction

In paper [1] we defined the concept of the algebraic integral on the discrete Mikusiński operator field M as the inverse of the algebraic derivative and have given the necessary and sufficient condition of the existence of it. Moreover, the linear, first order algebraic differential equation has been discussed in detail.

This paper consists of three parts. The brief summary of the elements of the discrete operational calculus is given in the first chapter. The alternative theorem related to the operational solutions of the linear, first order differential equation will be proved in the second chapter. This chapter contains the detailed operational discussion of the algebraic Bernoulli equation of the form

$$Dx+ax+bx^m=0, \quad x\in M,$$

too. D denotes the symbol of the algebraic derivative, $a, b\in M$, m is an arbitrary integer $(m\neq 0, 1)$.

The third chapter contains the application of the discrete operational calculus to the solution of difference equations with polynomial coefficients being the discrete analogue of the discussion of GESZTELYI [2].

The operational notations and symbols of BUTZER—SCHULTE [3] will be generally used.

1. The brief summary of the elements of the discrete operational calculus

The symbol $a=\{a_n\}$ denotes an arbitrary finite and real valued function defined in the discrete points $n=0, 1, 2, \ldots$, of the positive half line. The symbol a_n denotes the values of the function $\{a_n\}$ for $n=0, 1, 2, \ldots$. The set of the above functions is denoted by E. Two operations are defined in the set E.

Addition:

$$a+b=\{a_n\}+\{b_n\}=\{a_n+b_n\};$$

multiplication:

$$ab=\{a_n\}\{b_n\}=\left\{\sum_{k=0}^{n} a_k b_{n-k}\right\}.$$

The set E is a commutative ring with respect to addition and multiplication defined above. It has no divisors of zero and can be extended to a quotient field M. The

elements of M are called discrete convolution quotients or operators. They are of the form

$$\frac{\{a_n\}}{\{b_n\}}, \quad \{a_n\}, \{b_n\} \in E, \quad \{b_n\} \neq 0.$$

The field of the real numbers is denoted by K.

E and K can be embedded isomorphic in the field M. The unit element of E is $\{\delta_{0,n}\}$ where $\delta_{0,n}$ denotes the Kronecker symbol. The unit element of E, M and K can be identified algebraically. It is denoted by 1. Similarly the zero-element of E, M, K will be denoted by 0 since it can be identified too.

Every number is also a function in the discrete operational calculus, the value of the null-th component of which equals the value of the given number, the other components are zero.

Special operators

THE SUMMING OPERATOR. The function $h=\{1\}$ having the value 1 for every $n=0, 1, \ldots,$ defines the summing operator, since $h\{a_n\}=\{1\}\{a_n\}=\left\{\sum_{k=0}^{n} a_k\right\}$ for $\{a_n\} \in E$.

THE DIFFERENCE OPERATOR. The operator

$$q = \frac{1}{h-1} \tag{1.1}$$

defines the difference operator. Obviously $q \notin E$. Its fundamental property is

$$q\{a_n\} = \{\Delta a_n\} + (1+q)a_0$$

where

$$\{\Delta a_n\} = \{a_{n+1} - a_n\}.$$

More generally for the i-th difference the formula

$$q^i\{a_n\} = \{\Delta^i a_n\} + (1+q)\sum_{\nu=0}^{i-1} q^{i-1-\nu}\Delta^\nu a_0, \quad i \geqq 1 \tag{1.2}$$

holds.

THE TRANSLATION OPERATOR. The translation operator is defined by $v=\frac{1}{1+q}$.

It holds that

$$\frac{1}{(1+q)^m} = \{\delta_{m,n}\}, \quad m = 0, 1, 2, \ldots,$$

and

$$v^m\{a_n\} = \{b_n\}, \quad b_n = \begin{cases} a_{n-m}, & \text{for} \quad n \geqq m \\ 0, & \text{for} \quad 0 \leqq n < m. \end{cases} \tag{1.3}$$

Moreover $v^i v^j = v^{i+j}$ for every integer i, j.

In the following the symbol $a^{(\nu)}=\{a_{\nu,n}\}\in E$, $\nu=0,1,\dots$, denotes a sequence of discrete functions. [3] defines a convergence in E as follows:

$$\sum_{\nu=0}^{\infty} a^{(\nu)} = \sum_{\nu=0}^{\infty} \{a_{\nu,n}\} = \left\{\sum_{\nu=0}^{\infty} a_{\nu,n}\right\}.$$

The operational form of an arbitrary $b=\{b_n\}\in E$ is

$$b = \{b_n\} = \sum_{\nu=0}^{\infty} \frac{b_\nu}{(1+q)^\nu}, \tag{1.4}$$

in the sense of convergence defined above. We can easily deduce that every operator $d\in M$ can be written as

$$d = \sum_{\nu=\varkappa}^{-1} \frac{a_\nu}{(1+q)^\nu} + \sum_{\nu=0}^{\infty} \frac{a_\nu}{(1+q)^\nu} = \sum_{\nu=\varkappa}^{\infty} \frac{a_\nu}{(1+q)^\nu}, \tag{1.5}$$

$$(a_\nu\in K,\ \varkappa \text{ is a negative integer}).$$

There are only a finite number of terms of (1.5) not contained in E (see also BERG [4]). An equivalent statement is the following: Every $b\neq 0$ operator can be written as

$$b = (1+q)^N\{\varrho_n\} \quad (\{\varrho_n\}\in E,\ \varrho_0\neq 0,\ N \text{ is an integer}). \tag{1.6}$$

If $N\leqq 0$, then $b\in E$.

We have introduced in the ring E an operation by the definition

$$Da = D\{a_n\} = \{-na_n\} \tag{1.7}$$

(see [1]). It can be easily seen that

$$D\{a_n+b_n\} = D\{a_n\}+D\{b_n\},\quad D[\{a_n\}\{b_n\}] = \{a_n\}D\{b_n\}+\{b_n\}D\{a_n\}. \tag{1.8}$$

The operation D will be termed the algebraic derivative (see also [5]). The definition of the algebraic derivative can be extended to M by

$$D\frac{a}{b} = \frac{bDa-aDb}{b^2},\quad (a, b\in E,\ b\neq 0). \tag{1.9}$$

It is easy to verify that D retains properties (1.8) in M. If $\alpha\in K$, then $D\alpha=0$. Moreover, D is linear in M.

The algebraic derivative of the difference operator q is

$$Dq = 1+q, \tag{1.10}$$

(see [1]) or

$$D(1+q) = 1+q. \tag{1.11}$$

It is easy to verify by induction that for every integer ν and $x\in M$

$$D(x^\nu) = \nu x^{\nu-1}D(x),$$

holds.

Choosing $x=1+q$ we obtain

$$D[(1+q)^\nu] = \nu(1+q)^\nu. \tag{1.12}$$

From (1.5) and (1.12) we see that

$$Dx = \sum_{\nu=\varkappa}^{\infty} \frac{-\nu a_\nu}{(1+q)^\nu}. \tag{1.13}$$

It can be seen that if $Dx=0$, for $x\in M$, then $x\in K$.

Let us consider now the exponential function defined in the ring E (see [1]). This has a role in the theory of algebraic differential equations.

$$e^a = e^{\{a_n\}} = \sum_{\nu=0}^{\infty} \frac{\{a_n\}^\nu}{\nu!} \quad (a\in E).$$

It can be easily seen that

$$e^a e^b = e^{a+b} \quad \text{for} \quad a, b\in E.$$

Moreover, the exponential function

$$\{e_n\} = e^{\{a_n\}} \tag{1.14}$$

has the following properties:

$$e_0 = e^{a_0}, \quad \text{so} \quad e^{\{a_n\}} \neq 0,$$

$$De^a = (Da)e^a,$$

$$e^{\{a_n\}}\in K, \quad \text{iff} \quad \{a_n\}\in K$$

(see [1]).

The concept of the algebraic integral is defined as the inverse of D. If for an arbitrary $x\in M$ there exists an $y\in M$ such that $Dy=x$, then y is called the algebraic integral of x and is denoted by

$$y = \int x. \tag{1.15}$$

The algebraic integral is linear in M, i.e.

$$\int(\alpha x+\beta z) = \alpha\int x+\beta\int z, \quad (x, z\in M,\ \alpha, \beta\in K) \tag{1.16}$$

holds, provided that x, z are algebraic integrable. Two integrals of an operator — if they exist — differ from each other in an arbitrary number.

We have given a simple necessary and sufficient condition of the algebraic integrability (see [1]).

Let $x\in M$ be an arbitrary operator given by

$$x = \sum_{\nu=\varkappa}^{\infty} \frac{a_\nu}{(1+q)^\nu}. \tag{1.17}$$

x is algebraic integrable, if and only if, $a_0=0$. If so, the formula

$$\int x = \sum_{\nu=\varkappa}^{\infty} \frac{b_\nu}{(1+q)^\nu} \tag{1.18}$$

holds where

$$b_\nu = -\frac{a_\nu}{\nu}, \quad \text{for} \quad \nu\neq 0 \quad (b_0 \text{ arbitrary}). \tag{1.19}$$

2. Algebraic differential equations

Let us consider now the linear, first order, algebraic differential equation

$$Dx - wx = y \tag{2.1}$$

where $w, y \in M$ are given. We have shown that the homogeneous equation

$$Dx - wx = 0 \tag{2.2}$$

has a solution $x_0 \neq 0$, if and only if, $w = \{f_n\} \in E$ and f_0 is an arbitrary integer (see [1]). The general operational solution of (2.2) is

$$x = C(1+q)^{f_0} \exp\left[\int (\{f_n\} - f_0)\right], \quad (C \in K). \tag{2.3}$$

Moreover we have shown, that if $x_0 \neq 0$ exists, then (2.1) is solvable in M, if and only if $\int \frac{y}{x_0}$ exists and a particular solution is of the form

$$x = x_0 \int \frac{y}{x_0}.$$

By introducing the denotations

$$y = (1+q)^N \{\varrho_n\} \quad (\varrho_0 \neq 0),$$

$$\{H_n\} = \{\varrho_n\} \exp\left[-\int (\{f_n\} - f_0)\right]$$

and defining the function G by the definition

$$G_n = \frac{H_n}{N - f_0 - n}, \quad \text{for} \quad n \neq N - f_0$$

$$G_n \text{ is arbitrary}, \quad \text{for} \quad n = N - f_0$$

we obtained that the differential equation

$$Dx - fx = y \quad (f \in E, \, f_0 \text{ is an integer}) \tag{2.4}$$

is solvable in M, if and only if,

$$N < f_0, \quad \text{or} \quad N > f_0 \quad \text{and} \quad H_{N-f_0} = 0,$$

and a particular solution of (2.4) is of the form

$$x = (1+q)^N \exp\left[\int (\{f_n\} - f_0)\right] \{G_n\}. \tag{2.5}$$

The general solution of (2.4) is

$$x = [C(1+q)^{f_0} + (1+q)^N \{G_n\}] \exp\left[\int (\{f_n\} - f_0)\right].$$

provided that it exists.

In [1] we have not discussed the case of non-integer f_0. In this case the homogeneous equation has the only solution $x_0=0$ and the inhomogeneous equation cannot be solved by the method of variation of parameters since the solution formula

$$x = x_0 \int \frac{y}{x_0}$$

looses its meaning.

Nevertheless the operator (2.5) exists also for non integer f_0, and it can be easily shown that it is the only solution of

$$Dx - fx = y.$$

So it holds the following operational

ALTERNATIVE THEOREM. *If the homogeneous algebraic differential equation*

$$Dx - fx = 0 \quad (f \in E)$$

has only the trivial solution (f_0 is not an integer), then the corresponding inhomogeneous equation

$$Dx - fx = y \quad (y \in M)$$

always has one and only one solution in M. If the homogeneous equation has some non-trivial solutions (f_0 is an integer), then the inhomogeneous differential equation has either no solution or an infinity of solutions, depending on the given operator $y \in M$.

REMARK. What can be said on the operational solutions of

$$Dx - wx = y \tag{2.6}$$

if w is not a function? Since the corresponding homogeneous equation has the only solution, $x_0=0$ we see that (2.6) can have only one solution in M.

However we are not able to prove the existence of such a solution and cannot give the explicit operational form of it.

In the sequel we shall deal with the discrete operational Bernoulli equation of the form

$$Dx + ax + bx^m = 0. \tag{2.7}$$

Here $a \in E$, $b \in M$ are given, m is an arbitrary integer ($m \neq 0, 1$). If m is odd and $x_0 \in M$ is a solution of (2.7), then $-x_0$ is also a solution of the above equation. In the sequel we shall not distinguish between solutions differing from each other only by their signs.

By the formal application of the substitution

$$z = x^{1-m} \tag{2.8}$$

we can reduce (2.7) to the linear equation of the form

$$Dz - (m-1)az = (m-1)b. \tag{2.9}$$

If (2.7) has a non-trivial solution in M, it can be seen from (2.8) that (2.9) is also solvable in M. The converse statement does not hold. If (2.9) has a solution $z \in M$, we obtain

$$x = z^{\frac{1}{1-m}} \tag{2.10}$$

as a formal solution of (2.7). However (2.10) does not exist necessarily, since the operator field is not closed algebraically.

The formal solutions of the Bernoulli equation can be given explicitly and it holds the following

LEMMA. *Let us consider the discrete operational Bernoulli equation*

$$Dx + ax + bx^m = 0 \quad (a \in E,\ b \in M,\ m \text{ is an integer},\ m \neq 0, 1) \tag{2.11}$$

where

$$b = (1+q)^N \{\varrho_n\} \quad (\varrho_0 \neq 0).$$

Moreover let

$$h = (m-1)\{\varrho_n\} \exp\left[-\int (\{(m-1)a_n\} - (m-1)a_0)\right],$$

$$g = \{g_n\} = \left\{\frac{h_n}{N-(m-1)a_0-n}\right\} \quad \text{for} \quad n \neq N-(m-1)a_0,$$

$$g_n \text{ is arbitrary, for } n = N-(m-1)a_0.$$

If a_0 is not an integer, then (2.11) *has the only formal solution*

$$x = [(1+q)^N \{g_n\}]^{-\frac{1}{m-1}} \exp\left[-\int (\{a_n\} - a_0)\right]. \tag{2.12}$$

If a_0 is an integer, then (2.11) *has formal solutions if and only if*

$$N < (m-1)a_0$$

or

$$N > (m-1)a_0$$

and

$$h_{N-(m-1)a_0} = 0$$

holds. The formal solutions can be written as follows

$$x = [C(1+q)^{a_0(m-1)} + (1+q)^N \{g_n\}]^{-\frac{1}{m-1}} \exp\left[-\int (\{a_n\} - a_0)\right]. \tag{2.13}$$

PROOF. This lemma is a simple consequence of the operational theory of the linear algebraic differential equation. It arises now the following question. Under what conditions do the formal solutions represent proper solutions of the Bernoulli equation? In the following we are going to answer this question and give the necessary and sufficient condition guaranteeing the existence of the proper solution of (2.7).

It is well-known that the operator field is not closed algebraically. The algebraic equation

$$x^m = A \tag{2.14}$$

can have no solution in the discrete operator field. Let A be given as

$$A = (1+q)^L\{\Theta_n\} \quad (\Theta_0 \neq 0)$$

and we look for a solution of (2.14) in the form

$$x = (1+q)^M\{\vartheta_n\} \quad (\vartheta_0 \neq 0).$$

Substituting this to (2.14) we have

$$(1+q)^{Mm}\{\vartheta_n\}^m = (1+q)^L\{\Theta_n\}$$

or

$$(1+q)^{Mm-L} = \frac{\{\Theta_n\}}{\{\vartheta_n\}^m} = \{\tau_n\}, \quad (\tau_0 \neq 0).$$

Since $\Theta_0 \neq 0$, $\vartheta_0 \neq 0$ it can be seen that the operator τ is a function the null-th component of which is not zero. So

$$Mm - L = 0, \quad M = \frac{L}{m} \tag{2.15}$$

must hold. If so,

$$\{\vartheta_n\}^m = \{\Theta_n\}$$

will be obtained and finally we have

$$\{\vartheta_n\} = \sqrt[m]{\{\Theta_n\}}.$$

For odd m $\{\vartheta_n\}$ exists and is unique, for even m $\{\vartheta_n\}$ exists, if and only if, $\Theta_0 > 0$ ($\vartheta_0 = \sqrt[m]{\Theta_0}$). Both signs of the root must be taken into account.

Consequently (2.14) *is solvable in M, if and only if* $\frac{L}{m}$ *is an integer and for even* m $\Theta_0 > 0$ *holds.*

Let us consider the case where a_0 is not an integer. The formal solution (2.12) is a proper solution of the differential equation (2.11), if and only if, the operator

$$[(1+q)^N\{g_n\}]^{-\frac{1}{m-1}} \tag{2.16}$$

exists. From the definition of $\{g_n\}$ we have

$$g_0 = \frac{(m-1)\varrho_0 P}{N-(m-1)a_0} = \frac{\varrho_0 P}{\frac{N}{m-1} - a_0} \tag{2.17}$$

where P is some positive number. From the above discussion follows that (2.16), (2.12) exist if and only if $k = \frac{N}{m-1}$ is an integer and for odd m

$$\frac{\varrho_0}{k - a_0} > 0$$

holds.

If so, (2.12) can be written as

$$x = (1+q)^{-\frac{N}{m-1}} \{g_n\}^{-\frac{1}{m-1}} \exp\left[-\int (\{a_n\}-a_0)\right],$$

and from the property of the translation operator we have that the proper solution so obtained represents a function, if and only if $k=\dfrac{N}{m-1}\geqq 0$.

Now we discuss the case of integer a_0.

We take into account naturally the cases, where the formal solutions (2.13) exist. For $C=0$ (2.13) equals to (2.12). Let $C\neq 0$. We have two cases, $N>(m-1)a_0$ and $N<(m-1)a_0$.

I. If $N>(m-1)a_0$, then (2.13) can be written as

$$x = \left[(1+q)^N \left(C(1+q)^{a_0(m-1)-N}+\{g_n\}\right)\right]^{-\frac{1}{m-1}} \exp\left[-\int (\{a_n\}-a_0)\right]. \tag{2.18}$$

Since $C(1+q)^{a_0(m-1)-N}\in E$ is a function, the null-th component of which is zero, from (2.17) we obtain that (2.13) represents operators, if and only, if $k=\dfrac{N}{m-1}$ is an integer and for odd m, $\dfrac{\varrho_0}{k-a_0}>0$ holds.

If so, the formal solution (2.13) is the proper general solution of (2.11) and can be written as

$$x = (1+q)^{-\frac{N}{m-1}} \left[C(1+q)^{a_0(m-1)-N}+\{g_n\}\right]^{-\frac{1}{m-1}} \exp\left[-\int (\{a_n\}-a_0)\right]. \tag{2.19}$$

The well-known property of the translation operator shows that (2.19) is a function for every C if and only if $k=\dfrac{N}{m-1}\geqq 0$.

II. If $N<(m-1)a_0$, (2.13) is of the form

$$x = \left[(1+q)^{a_0(m-1)}\left(C+(1+q)^{N-(m-1)a_0}\{g_n\}\right)\right]^{-\frac{1}{m-1}} \exp\left[-\int (\{a_n\}-a_0)\right]. \tag{2.20}$$

Since

$$C+(1+q)^{N-(m-1)a_0}\{g_n\}\in E$$

where the null-th component is $C\neq 0$, we obtain for even m that (2.20), (2.13) represent operators for every C, for odd m, if and only if $C>0$. If so, the proper general operational solution of (2.11) reads as

$$x = (1+q)^{-a_0}\left[C+(1+q)^{N-(m-1)a_0}\{g_n\}\right]^{-\frac{1}{m-1}} \exp\left[-\int (\{a_n\}-a_0)\right]. \tag{2.21}$$

From the property of the translation operator we have that (2.21) is a function for every allowable value of C if and only if $a_0\geqq 0$. It holds the following

THEOREM 1. *Let us consider the algebraic Bernoulli equation*

$$Dx+ax+bx^m = 0, \quad \begin{pmatrix} a\in E \\ b=(1+q)^N\{\varrho_n\},\ \varrho_0\neq 0 \end{pmatrix} \tag{2.22}$$

and let m be an arbitrary integer $(m \neq 0, 1)$. *If* a_0 *is not an integer, then* (2.22) *is non-trivially solvable in M, if and only if,*

$$k = \frac{N}{m-1} \quad \text{is an integer} \tag{2.23}$$

and for odd m $\frac{\varrho_0}{k-a_0} > 0$ *holds. If so,* (2.22) *has the only non-trivial solution*

$$x = (1+q)^{-k}\{g_n\}^{-\frac{1}{m-1}} \exp\left[-\int(\{a_n\}-a_0)\right]. \tag{2.24}$$

(2.24) *is a function, if and only if* $k \geqq 0$.

If a_0 *is an integer and* $N > (m-1)a_0$, (2.22) *is non-trivially solvable in M, if and only if the following conditions are satisfied:*

$$h_{N-(m-1)a_0} = 0, \quad k \text{ is an integer} \tag{2.25}$$

and for odd m

$$\frac{\varrho_0}{k-a_0} > 0.$$

If so, the general solution of (2.22) *is*

$$X = [C(1+q)^{a_0(m-1)} + (1+q)^N\{g_n\}]^{-\frac{1}{m-1}} \exp\left[-\int(\{a_n\}-a_0)\right] \tag{2.26}$$

being a function for every real C, if and only if

$$k = \frac{N}{m-1} \geqq 0.$$

If a_0 *is an integer and* $N < (m-1)a_0$ *the formal general nontrivial solution of* (2.22)

$$x = [C(1+q)^{a_0(m-1)} + (1+q)^N\{g_n\}]^{-\frac{1}{m-1}} \exp\left[-\int(\{a_n\}-a_0)\right] \tag{2.27}$$

represents a proper solution of (2.22) *for every* $C > 0$. (2.27) *is a proper solution of* (2.22) *for every* $C < 0$, *if and only if, m is even.*

The operators so obtained are functions, if and only if $a_0 \geqq 0$.

Finally, for $C = 0$ (2.27) *is a proper solution of* (2.22) *if and only if k is an integer, and for odd m,* $\frac{\varrho_0}{k-a_0} > 0$ *holds. The solution represents a function, if and only if* $k \geqq 0$.

In the case of $N = (m-1)a_0$ (2.22) *has the only solution* $x = 0$.

REMARK. From practical point of view, the discrete algebraic Bernoulli equation has some importance at the solution of non-linear recursions of the following type:

$$\{A_n\} \cdot \{nf_n\} + \{B_n\} \cdot \{f_n\} + \{C_n\} \cdot \{f_n\} \cdot \{f_n\} \cdot \overset{(m)}{\cdots} \cdot \{f_n\} = 0 \tag{2.28}$$

where the functions $\{A_n\}$, $\{B_n\}$, $\{C_n\}$ are given and $\{f_n\}$ is unknown. (2.28) has the operational form

$$-Df + \frac{B}{A}f + \frac{C}{A}f^m = 0.$$

It can be seen from the above discussion that the operational method is useful if $\frac{\{B\}}{\{A\}}$ is a function.

3. The application of the operational method to the solution of difference equations with polynomial coefficients

In this chapter we shall solve difference equations with polynomial coefficients by the application of the discrete operational method. Let us consider the difference equation

$$\sum_{i=0}^{m}(a_i-b_i n)\Delta^i x_n = \varphi_n \tag{3.1}$$

where a_i and b_i are given numbers, φ_n is a given function and x_n is the unknown function. The initial conditions $x_0, x_1, \ldots, x_{m-1}$ are also given.

Operationally (3.1) can be written as

$$\sum_{i=0}^{m}[a_i\Delta^i x + b_i D(\Delta^i x)] = \varphi. \tag{3.2}$$

Taking into account (1.2) and $D(q^i)=iq^{i+1}(1+q)$ we have

$$D(\Delta^i x) =$$

$$= q^i Dx + iq^{i-1}(1+q)x - (1+q)\left[\sum_{v=0}^{i-2}(i-1-v)\,q^{i-2-v}\Delta^v x_0 + \sum_{v=0}^{i-1}(i-v)\,q^{i-v-1}\Delta^v x_0\right].$$

Substituting this in (3.2), the following linear equation will be obtained:

$$Dx + \frac{\sum_{i=0}^{m}[(a_i+ib_i)q^i + ib_i q^{i-1}]}{\sum_{i=0}^{m} b_i q^i}\,x = \tag{3.3}$$

$$= \frac{\varphi + (1+q)\sum_{i=0}^{m}\left[\sum_{v=0}^{i-1}\left(a_i+b_i(i-v)\right)q^{i-1-v}\Delta^v x_0 + \sum_{v=0}^{i-2} b_i(i-1-v)\,q^{i-2-v}\Delta^v x_0\right]}{\sum_{i=0}^{m} b_i q^i}$$

where $\sum_{v=0}^{i-1}$ is zero for $i<1$, and $\sum_{v=0}^{i-2}$ is zero for $i<2$. We can solve (3.3) only in the case when the operator

$$f = -\frac{\sum_{i=0}^{m}[(a_i+ib_i)q^i + ib_i q^{i-1}]}{\sum_{i=0}^{m} b_i q^i} \tag{3.4}$$

is a function.

We need the following statement.
The operator

$$z = \frac{\sum_{i=0}^{m} \gamma_i q^i}{\sum_{i=0}^{n} \delta_i q^i} \quad (\gamma_i, \delta_i \in K, \ \gamma_m, \delta_n \neq 0)$$

is a function, if and only if $m \leqq n$. This may be easily seen since z can be written as

$$z = \frac{\sum_{i=0}^{m} \gamma_i'(1+q)^i}{\sum_{i=0}^{n} \delta_i'(1+q)^i} \quad (\gamma_i', \delta_i' \in K)$$

where $\gamma_m = \gamma_m'$, $\delta_n' = \delta_n$.
We have

$$z = \frac{(1+q)^m \sum_{i=0}^{m} \gamma_i' \frac{1}{(1+q)^{m-i}}}{(1+q)^n \sum_{i=0}^{n} \delta_i' \frac{1}{(1+q)^{n-i}}}$$

and

$$z = (1+q)^{m-n}\{\tau_n\} \quad (\{\tau_n\} \in E, \ \tau_0 \neq 0). \tag{3.5}$$

We see from (3.5) that the above statement is true. The numbers a_m, b_m may not be zero simultaneously, because then the order of the difference equation would be lower than m.

If $b_m = 0$, $a_m \neq 0$, the polynomial occuring in the numerator of (3.9) is of higher degree than that of the denominator, consequently (3.4) is not a function.

If $b_m \neq 0$, the polynomial occuring in the denominator of (3.4) is of m-th degree, the polynomial in the numerator is at most of m-th degree, so (3.4) is a function.

In the sequel we assume that $b_m \neq 0$. The right-hand side of (3.3) is a function. This follows from the fact that the polynomial occuring in the numerator of the right-hand side of (3.3) is at most of m-th degree. From (3.4) we have

$$f_0 = -\frac{a_m + m b_m}{b_m} = -\left(\frac{a_m}{b_m} + m\right). \tag{3.6}$$

The solutions of (3.3) can be written explicitly by the aid of the formulas given in [1] and the preceding paragraph. Occasionally, it is more convenient to express the right hand side of (3.3) as a function of the operator q. We shall illustrate this on two simple examples.

We give now an operational proof of the following well-known fact.

THEOREM 2. *Let us consider the difference equation* (3.1) *with the initial conditions* $x_0, x_1, \ldots, x_{m-1}$. *Let* $b_m \neq 0$. *If* $\frac{a_m}{b_m}$ *is not a non-negative integer, then* (3.1)

has exactly one solution satisfying the given initial conditions. (Regular case.) If $\frac{a_m}{b_m}$ *is a non-negative integer, then* (3.1) *has either no solution or an infinity of solutions satisfying the given initial conditions. (Singular case.)*

PROOF. Leet $\{x'_n\}$ be an arbitrary function satisfying the initial conditions $x_0, x_1, \ldots, x_{m-1}$. By introducing the substitution

$$\{u_n\} = \{x_n\} - \{x'_n\}$$

in (3.1) we obtain the difference equation

$$\sum_{i=0}^{m} (a_i - b_i n) \Delta^i u_n = \psi_n \tag{3.7}$$

where the initial conditions of $\{u_n\}$ are zero, $\{\psi_n\}$ is some function.

From (3.3) it follows that (3.7) can be written operationally as

$$Du + \frac{\sum_{i=0}^{m} [(a_i + ib_i) q^i + ib_i q^{i-1}]}{\sum_{i=0}^{m} b_i q^i} u = \frac{\psi}{\sum_{i=0}^{m} b_i q^i}. \tag{3.8}$$

First let $\frac{a_m}{b_m}$ be an integer. We obtain by (3.4), (3.6) that the general solution of the corresponding homogeneous equation is of the form

$$u_h = C(1+q)^{-\left(\frac{a_m}{b_m}+m\right)} \exp\left[\int (\{f_n\} - f_0)\right]. \tag{3.9}$$

We know that (3.8) is solvable in M, if and only if,

$$\int \frac{\psi}{\sum_{i=0}^{m} b_i q^i} (1+q)^{\frac{a_m}{b_m}+m} \exp\left[-\int (\{f_n\} - f_0)\right] \tag{3.10}$$

exists. Since $\frac{(1+q)^m}{\sum_{i=0}^{m} b_i q^i} \in E$, it may be easily seen that the above algebraic integral reads as

$$\int (1+q)^{\frac{a_m}{b_m}} \{\chi_n\} \quad (\chi \in E). \tag{3.11}$$

If $\frac{a_m}{b_m} < 0$, the integrand is a function the null-th component of which is zero, so (3.11) exists. If $\frac{a_m}{b_m} \geqq 0$, (3.11) does not exist necessarily. We have from (3.5) that $\frac{1}{\sum_{i=0}^{m} b_i q^i}$ is a function, the values of the first m components are zero. This

is also true for the right hand side of (3.8) and it can be written that

$$\frac{\psi}{\sum_{i=0}^{m} b_i q^i} = (1+q)^N \{\mu_n\} \quad (\mu_0 \neq 0,\ N \leqq -m).$$

In the case of existence of (3.11) a particular solution of (3.8) may be derived from (2.5) and is of the form

$$u_p = (1+q)^N \exp\left[\int (\{f_n\} - f_0)\right] \{G_n\}. \tag{3.12}$$

Since $N \leqq -m$, it can be seen that the values of the first m components of the solution so obtained are zero and (3.12) satisfies (3.7) and the vanishing initial conditions.

For $\frac{a_m}{b_m} < 0$, (3.9) cannot be a function satisfying zero initial conditions.

Namely, for $\frac{a_m}{b_m} < 0$, $\frac{a_m}{b_m} + m < 0$ (3.9) is not a function. On the other hand if $\frac{a_m}{b_m} < 0$, $\frac{a_m}{b_m} + m \geqq 0$ (3.9) is a function for every value of C. However the values of the first m components of (3.9) are not all zero ($C \neq 0$).

Consequently, if $\frac{a_m}{b_m} < 0$ (3.8) and (3.7) have only one solution being a function with m vanishing initial values.

If $\frac{a_m}{b_m} \geqq 0$, (3.9) is a function for every value of C, the first m components of which are zero.

So, if (3.8) has a solution, then it has an infinity of solutions satisfying the difference equation (3.7) with the vanishing initial conditions. The general solution of (3.7) is obtained as

$$u = u_p + u_h.$$

Now let us consider the case where $\frac{a_m}{b_m}$ is not integer. Then (3.12) is the only solution of (3.8), consequently also of (3.7) satisfying the zero initial conditions.

So we have proved the Theorem in the case if the initial conditions $x_0, x_1, \ldots, x_{m-1}$ are zero.

If we rewrite the function $\{x_n\}$ by

$$\{x_n\} = \{u_n\} + \{x'_n\}$$

we have that $\{x_n\}$ satisfies the difference equation (3.1) and the given initial conditions. The Theorem is proved.

We end this paragraph whit some examples.

FIRST EXAMPLE.

$$(n+1)\Delta x_n + n x_n = n, \quad x_0 = 0. \tag{3.13}$$

Since $\{n\}=\dfrac{q+1}{q^2}$, the operational form of (3.13) is

$$Dx+\frac{1}{1+q}x=-\frac{1}{q^2}. \tag{3.14}$$

The null-th component of $\dfrac{1}{1+q}$ is zero, the general solution of the homogeneous equation is of the form

$$x_h = Ce^{-\int \frac{1}{1+q}} = Ce^{\frac{1}{1+q}}.$$

We look for a particular solution of (3.14) as

$$x_p = C^* e^{\frac{1}{1+q}}.$$

Substituting this in (3.14) we have

$$C^* = -\int \frac{1}{q^2} e^{-\frac{1}{1+q}}. \tag{3.15}$$

It can be seen that

$$C^* = \frac{1+q}{q} e^{-\frac{1}{1+q}}$$

and

$$x_p = C^* e^{\frac{1}{1+q}} = \frac{1+q}{q} = \{1\}.$$

Since

$$x_h = Ce^{\frac{1}{1+q}} = C\sum_{\nu=0}^{\infty} \frac{1}{\nu!\,(1+q)^\nu} = C\left\{\frac{1}{n!}\right\},$$

the general solution of (3.14), (3.13) reads

$$x = \{x_n\} = \left\{1+\frac{C}{n!}\right\}.$$

From $x_0=0$ we obtain $C=-1$ and

$$\{x_n\} = \left\{1-\frac{1}{n!}\right\}. \tag{3.16}$$

(3.16) is the only solution satisfying (3.13) with the given initial condition.

SECOND EXAMPLE.

$$(n-1)\Delta x_n + nx_n = n, \quad x_0 = 1. \tag{3.17}$$

Obviously

$$x_p = \{1\}$$

is a particular solution of (3.17). The general solution of (3.17) is the sum of x_p and the general solution of the corresponding homogeneous equation with zero initial condition.

The corresponding homogeneous equation with zero initial condition is of the following operational form

$$Dx+\frac{1+2q}{1+q}\,x=0. \tag{3.18}$$

Since

$$f=-\frac{1+2q}{1+q}=-1-\frac{q}{1+q}\in E$$

we have $f_0=-2$.

It can be seen by (2.3) that the general solution of (3.18) is

$$x_h=C(1+q)^{-2}\exp\left[\int\left(-\frac{1+2q}{1+q}+2\right)\right]= \tag{3.19}$$

$$=C(1+q)^{-2}\exp\left[\int\frac{-1-2q+2+2q}{1+q}\right]=$$

$$=C(1+q)^{-2}\exp\left[\int\frac{1}{1+q}\right]=C(1+q)^{-2}\exp\left[-\frac{1}{1+q}\right]=$$

$$=C(1+q)^{-2}\sum_{v=0}^{\infty}\frac{(-1)^v}{v!\,(1+q)^v}=C(1+q)^{-2}\left\{\frac{(-1)^n}{n!}\right\}.$$

There are an infinite number of solutions satisfying the zero initial condition. Consequently (3.17) has also infinitely many solutions with the condition $x_0=1$. They are of the form

$$x=\{x_n\},$$

$$x_n=\begin{cases}1, & \text{for} \quad n=0,1\\ 1+C\,\dfrac{(-1)^{n-2}}{(n-2)!}, & \text{for} \quad n\geqq 2\end{cases}\qquad (C\in K).$$

REFERENCES

[1] FÉNYES, T.—KOSIK, P.: The algebraic derivative and integral in the discrete operational calculus, *Studia Sci. Math. Hung.* **7** (1972), 117—130.

[2] GESZTELYI, E.: Anwendung der Operatorenrechnung auf lineare Differentialgleichungen mit Polynom-Koeffizienten, *Publ. Math. Debrecen* **10** (1963), 215—243.

[3] BUTZER, L.—SCHULTE, H.: *Ein Operatorenkalkül zur Lösung gewöhnlicher und partieller Differenzengleichungssysteme von Funktionen diskreter Veränderlichen und seine Anwendungen,* Forschungsberichte des Landes Nordrhein—Westfalen. Nr. 1537, Westdeutscher Verlag, Cologne, 1965.

[4] BERG, L.: *Einführung in die Operatorenrechnung,* VEB Deutscher Verlag der Wissenschaften, Berlin, 1965.

[5] MOORE, D. H.: Convolution products and quotients and algebraic derivatives of sequences, *The American Mathematical Monthly* **69** (1962), 132—138.

Mathematical Institute of the Hungarian Academy of Sciences, H—1053 Budapest, Reáltanoda u. 13—15. Hungary

(Received March 24, 1974)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 381—388.

NEARBY ALTERNATING CHEBYSHEV APPROXIMATION*

by

CHARLES B. DUNHAM

Abstract. The best Chebyshev approximation from an alternating family depends on the function being approximated, on the domain, and on the weight function. This dependence is continuous in a neighbourhood of a point where the best approximation is non-degenerate. A local existence result for varisolvent families is proven. It is shown that non-existence or discontinuity occurs when the best approximation is degenerate and the number of alternations is minimal.

Let (F, P) be an alternating family of approximations on a closed interval $[\alpha, \beta]$. All functions considered will be continuous on $[\alpha, \beta]$ and for such functions we define the norm on a non-empty closed subset Y of $[\alpha, \beta]$ to be

$$\|g\|_Y = \max\{|g(x) : x \in Y\}.$$

(When Y is omitted, it is understood that $Y=[\alpha, \beta]$). The Chebyshev problem on Y is for a given continuous function f and given continuous non-negative weight function w to find an element $F(A^*, \cdot)$, $A^* \in P$, for which $\|w(f-F(A, \cdot))\|_Y$ attains its infimum $\varrho(f, Y, w)$ over $A \in P$. Such an element $F(A^*, \cdot)$ is called a best Chebyshev approximation to f on Y with respect to w. It is unique if it exists. Denote it by $T(f, Y, w)$, defining the generalized Chebyshev mapping. In this paper is studied the dependence of T on its arguments.

Some restrictions of this problem have already been studied. The classical Chebyshev operator is $T(\cdot, [\alpha, \beta], 1)$. A study of it for varisolvent F appears in [1] and for general F appears in [3]. The subset mapping for fixed f is $T(f, \cdot, 1)$. A study of it for general F and subsets filling out $[\alpha, \beta]$ appears in [4] for varisolvent F and in [3] for general F. MAEHLY and WITZGALL [5] considered varying f and w jointly in approximation by ordinary rational functions on an interval $[\alpha, \beta]$. The behaviour of T for approximation by finite dimensional linear families is considered in [8].

A hitherto unstudied case in which continuity of T is of interest, is where X is an alternant of the error curve of the best approximation on a larger set and the subsets Y are trial alternants used in the Remez algorithm. Continuity of T implies that if perturbations in f and w are small enough and the trial alternant is close enough, the trial solution will be close to the best approximation on the alternant, which is the best approximation on the larger set.

To avoid trivial cases, we will assume that in all cases studied card $\{y : y \in Y, w(y) > 0\}$ is not smaller than the degree of F.

* This work was assisted by a grant from the National Research Council of Canada.

The following definition is given in [3] and agrees with the accepted definitions of degeneracy for polynomial rational families $R_m^n[\alpha, \beta]$ and exponential sums.

DEFINITION. F is *degenerate at* A $(F(A, \cdot)$ is degenerate) if every neighbourhood of $F(A, \cdot)$ contains elements of higher degree.

Clearly elements of maximum degree are non-degenerate. In most approximating families of interest, including those of the two classes mentioned above, all elements not of maximum degree are degenerate. An example where this is not the case precedes Theorem 4 of [3].

DEFINITION. Let $\{g, g_1, \ldots\} \subset C[\alpha, \beta]$. $\{g_k\} \to g$ means $\|g - g_k\| \to 0$.

DEFINITION. For X, Y non-empty closed subsets of $[\alpha, \beta]$, define

$$\operatorname{dist}(X, Y) = \sup\{\inf\{|x-y| : y \in Y\} : x \in X\},$$

$$d(X, Y) = \max\{\operatorname{dist}(X, Y), \operatorname{dist}(Y, X)\}.$$

DEFINITION. Let $X, X_1, \ldots$ be non-empty closed subsets of $[\alpha, \beta]$. We say $\{X_k\} \to X$ if $d(X, X_k) \to 0$.

LEMMA 1. *Let* $F(A, \cdot)$ *be the best approximation to* f *on* X *with respect to* w. *Let* $\{x_0, \ldots, x_n\}$ *be an alternant of* $w(f - F(A, \cdot))$ *on* X. *Let* $\|v - w\| \max\{\|f - F(A, \cdot)\|_X, 1\} \leqq \leqq \varepsilon$. *Let* $\|v\| \|f - g\| < \varepsilon$. *Let* $v(x_i) w(x_i) > 0$, $i = 0, \ldots, n$.

There exists $\delta > 0$ *such that if*

$$|x_i' - x_i| < \delta, \quad i = 0, \ldots, n,$$

and

$$\max\{v(x_i')|g(x_i') - F(B, x_i')| : i = 0, \ldots, n\} \leqq \left\|w(f - F(A, .))\right\|_X + 3\varepsilon \tag{1}$$

then

$$\operatorname{sgn}[f(x_0) - F(A, x_0)](-1)^i[F(B, x_i') - F(A, x_i')] \geqq -6\varepsilon / \min\{v(x_i') : i = 0, \ldots, n\}. \tag{2}$$

PROOF. Assume without loss of generality that $f(x_0) - F(A, x_0) \geqq 0$, and let E denote $\|w(f - F(A, \cdot))\|_X$. We have

$$w(x_i)[f(x_i) - F(A, x_i)] = (-1)^i E.$$

As $w(f - F(A, \cdot))$ is uniformly continuous on $[\alpha, \beta]$, there exists $\delta > 0$ such that if $|x_i - x_i'| < \delta$,

$$(-1)^i w(x_i')(f(x_i') - F(A, x_i')) \geqq E - \varepsilon, \quad i = 0, \ldots, n.$$

By choice of v,

$$(-1)^i v(x_i')(f(x_i') - F(A, x_i')) \geqq E - 2\varepsilon, \quad i = 0, \ldots, n.$$

By choice of g,

$$(-1)^i v(x_i')(g(x_i') - F(A, x_i')) \geqq E - 3\varepsilon, \quad i = 0, \ldots, n. \tag{3}$$

Inequality (1) can be rewritten as

$$(-1)^i v(x_i')(g(x_i') - F(B, x_i')) \geqq E + 3\varepsilon, \quad i = 0, \ldots, n. \tag{4}$$

Subtracting (4) from (3) gives

$$(-1)^i v(x_i')(F(B, x_i') - F(A, x_i')) \geqq -6\varepsilon, \quad i = 0, \ldots, n.$$

LEMMA. *Let F be non-degenerate at A. Then for given $\varepsilon>0$ there exists $\eta(\varepsilon)$ such that $\|F(A,\cdot)-F(B,\cdot)\|<\eta(\varepsilon)$ if (2) holds and $\eta(\varepsilon)\to 0$ as $\varepsilon\to 0$.*

This is Lemma 2 of [3].

LEMMA. *Let F be of degree m at A_k, $k=0,1,\ldots$ and let $\{F(A_k,\cdot)\}$ converge pointwise to $F(A_0,\cdot)$ on m distinct points. Then $\{F(A_k,\cdot)\}$ converges uniformly to $F(A_0,\cdot)$.*

This is Theorem 3 of [3].

Best approximations may fail to exist. Further, numerical procedures may only give approximations which are near them. We therefore introduce nearly best approximations, which do exist.

DEFINITION. $F(A,\cdot)$ is an *ε-nearly best approximation* to f on Y with respect to w if $\|w(f-F(A,\cdot))\|_Y \leqq \varrho(f,Y,w)+\varepsilon$.

Best approximations are 0-nearly best approximations.

THEOREM 1. *Let $T(f,X,w)=F(A,\cdot)$ and F be non-degenerate at A. Let $\{f_k\}\to f$, $d(X,X_k)\to 0$, $\{w_k\}\to w$, and $\{\varepsilon_k\}\to 0$. Let $F(A_k,\cdot)$ be ε_k-nearly best to f_k with respect to w_k on X_k. Then $\{F(A_k,\cdot)\}$ converges uniformly to $F(A,\cdot)$.*

This is a consequence of Lemmas 1 and 2.

In case F is varisolvent, a local existence result holds. A definition of varisolvence and theory for approximation on an interval is given in [6]. We assume that the difficulty of [2] does not occur.

THEOREM 2. *Let F be unisolvent of variable degree. Let $T(f,X,w)=F(A,\cdot)$ and F be non-degenerate at A. There exists μ, ν such that if $d(X,Y)<\mu$, $\|v-w\|\leqq\nu$, and $\|v\|\,\|f-g\|<\nu$, then g has a (non-degenerate) best approximation on Y with respect to v. If $\{f_k\}\to f$, $\{X_k\}\to X$, $\{w_k\}\to w$, then $T(f_k,X_k,w_k)\to T(f,X,w)$.*

PROOF. Let F have degree n at A. By definition of degeneracy and the second corollary to Theorem 2 of [3], there exists λ such that $\|F(A,\cdot)-F(B,\cdot)\|<\lambda$ implies that F is of degree n at B. Let $x_0,\ldots,x_n$ be as in Lemma 1. By definition of solvency of degree n at A there exists $\gamma>0$ such that if $|y_k-F(A,x_k)|<\gamma$, $k=1,\ldots,n$, then there exists a parameter B satisfying

$$F(B,x_k)=y_k \quad k=1,\ldots,n, \quad \|F(A,.)-F(B,.)\|<\lambda. \tag{5}$$

By choice of λ, F is unisolvent of degree n at any such B, and hence B is completely determined by (5). Choose ε such that $\eta(\varepsilon)<\gamma/2$, then by Lemma 1 and 2, if $|x_i'-x_i|<\delta$ for $i=0,\ldots,n$ and

$$\max\{v(x_i')|g(x_i')-F(B,x_i')|: i=0,\ldots,n\}\leqq\|f-F(A,.)\|_X+3\varepsilon$$

then

$$\|F(A,.)-F(B,.)\|<\gamma/2.$$

Now choose Y with $d(X,Y)<\delta$. Let $\{B_k\}$ be a sequence of parameters such that

$$\{\|v(g-F(B_k,.))\|_Y\}\to\inf\{\|v(g-F(C,.))\|_Y: C\in P\}:=\varrho(g,Y,v).$$

If $\varrho(g,Y,v)=\|v(g-F(A,\cdot))\|$ then $F(A,\cdot)$ is best to f on Y with respect to v. Otherwise choose $x_i'\in Y$ such that $|x_i-x_i'|<\delta$, $i=0,\ldots,n$. We can assume that

$$\max\{v(x_i')|g(x_i')-F(B_k,x_i')|: i=0,\ldots,n\}\leqq\|w(f-F(A,.))\|_X+3\varepsilon.$$

It follows that $\|F(A, \cdot)-F(B_k, \cdot)\|<\gamma/2$ and the n-tuples of values of $F(B_k, \cdot)$ at the points $x_1, \ldots, x_n$ form a bounded sequence with subsequence converging to an accumulation point $(y_1, \ldots, y_n)$, which determines a parameter B at which F is unisolvent of degree n. Using Lemma 3 we can show that for all $x \in Y$, $|v(x)(g(x)-F(B, x))| \leqq \varrho(g, Y, v)$ and so $F(B, \cdot)$ is a best approximation to g on Y with respect to v. The first part of the theorem is proven. Now let $\{f_k\} \to f$, $\{X_k\} \to X$, and $\{w_k\} \to w$. Then for all k sufficiently large, there is a best approximation $F(A_k, \cdot)$ to f_k on X_k with respect to w_k. From Lemmas 1 and 2 it follows immediately that $\|F(A, \cdot)-F(A_k, \cdot)\| \to 0$.

If $T(f, X, w)$ is degenerate, we may not have nearby existence. This is shown in [4] for the subset mapping. We now show that it may not hold for even the classical Chebyshev operator.

EXAMPLE. Let $n>1$ and X be a closed subset of $[0, 1]$ containing at least $2n+1$ points. Let

$$F(A, x) = \sum_{k=1}^{n} a_k \exp(a_{n+k} x).$$

F is a varisolvent approximating function. Let $f(x)=0$ and $f_k(x)=x/k$. f_k is a uniform limit of approximations but is not itself an approximation, hence it has no best approximation on X.

Let us now consider the case where $T(f, X, w)$ is degenerate. A special case, which we consider at length, is that in which f is an approximant $F(A, \cdot)$. It is well-known that the classical Chebishev operator $T(\cdot, [\alpha, \beta], 1)$ is continuous in this case. It is also obvious that the mapping $T(f, \cdot, \cdot)$ is constant (hence continuous) in this case. However, as will be shown in the next example, T may not be continuous. The most general result appears to be

THEOREM 3. *Let f be an approximant $F(A, \cdot)$. Let $\{f_k\} \to f$, $\varepsilon_k \to 0$ and $\{w_k\} \to w>0$. Let $F(A_k, \cdot)$ be ε_k-nearly best to f_k on X with respect to w_k, then $\|F(A_k, \cdot)-F(A, \cdot)\|_X \to 0$.*

The theorem is easily proven by standard arguments. The theorem is, however, not true if we also vary the domain.

EXAMPLE. Let us approximate by the family $R_1^0[0, 1]$ of rational functions. Let $w=1$. Let $X_k=[1/k, 1]$ and $X=[0, 1]$. Let $f=0$ and

$$f_k(x) = 1/(1+k^2 x) \quad (x \geqq 1/k)$$
$$= 1/(1+k) \quad (0 \leqq x < 1/k.)$$

We have $T(f_k, X_k, w)=1/(1+k^2 x) \nrightarrow T(f, X, w)=0$.

Let us now consider the case where $T(f, X, w)$ is degenerate and not equal to f. There do exist sequences $\{f_k\} \to f$, $\{X_k\} \to X$, $\{w_k\} \to w$ such that $T(f_k, X_k, w_k) \to T(f, X, w)$. In particular these sequences can be chosen such that $T(f_k, X_k, w_k) = T(f, X, w)$. However, the theory to date suggests that if $T(f, X, w)$ is degenerate and not equal to f, then suitably varying any one of the three arguments of T is likely to cause failure of existence or failure of continuity of T. Let $d(A)$ denote the degree of F at A. In most cases of interest, $w(f-F(A, \cdot))$ alternates exactly $d(A)$ times when A is best.

THEOREM 4. *Let F be degenerate at A and $w(f-F(A,\cdot))$ alternate exactly $d(A)$ times on X. There is a sequence $\{f_k\}\to f$ such that if $T(f_k, X, w)$ exists, $T(f_k, X, w)\dashrightarrow T(f, X, w)$.*

To prove this, we use arguments similar to those of Lemma 6 and the second Theorem 8 of [3].

Define $M(A)=\{x: w(x)|f(x)-F(A,x)|=\|w(f-F(A,\cdot))\|_X\}$. If $M(A)$ is "sufficiently large", A best in X implies that A is best on all sets near X. For example, let $[\alpha,\beta]=[-1,1]$, $F(A,x)=a_1\exp(a_2x)$, $w=1$, and

$$\begin{aligned} f(x) &= -1 && -1\leqq x\leqq -1/2\\ &= 2x && -1/2 < x < 1/2\\ &= 1 && 1/2\leqq x\leqq 1. \end{aligned}$$

0 is best on any subset of $[-1,1]$ containing a point $[-1,-1/2]$ and a point in $[1/2,1]$, since two such points form an alternant. However, in a typical case of approximation, $M(A)$ is nowhere dense.

THEOREM 5. *Let $w(f-F(A,\cdot))$ alternate exactly $d(A)$ times on $[\alpha,\beta]$, F be degenerate at A and $M(A)$ be nowhere dense. There exists a sequence $\{X_k\}$ of closed subsets such that $\{X_k\}\to[\alpha,\beta]$ and if $\{T(f,X_k,w)\}$ exists, $T(f,X_k,w)\dashrightarrow T(f,[\alpha,\beta],w)$.*

This is proven by arguments similar to those of Theorem 7 of [3]. A similar result can be given for finite X.

LEMMA 4. *Let F be degenerate at A and $w(f-F(A,\cdot))$ alternate exactly $d(A)$ times on X. Then there is a sequence $\{w_k\}\dashrightarrow w$ such that either $T(f,X,w_k)$ does not exist or if $T(f,X,w_k)=F(C_k,\cdot)$, $d(C_k)>d(A)$.*

PROOF. By definition, there is $A_k\in P$ such that $d(A_k)>d(A)$ and $\|F(A,\cdot)-F(A_k,\cdot)\|<1/k$. Define $E(x)=w(x)(f(x)-F(A,x))$ and $Y=\{x:|E(x)|\geqq\|E\|_X/2\}$. It can be seen that $(f(x)-F(A,x))/(f(x)-F(A_k,x))\to 1$ uniformly on Y. Define

$$w_k(x)=w(x)(f(x)-F(A,x))/(f(x)-F(A_k,x)),\quad x\in Y.$$

By the Tietze extension theorem as given in the text of Dugundji [12, 149] there is a continuous extension of w_k to $[\alpha,\beta]$ such that $\|w_k-w\|<\|w_k-w\|_Y+1/k\to 0$. We have

$$w_k(x)(f(x)-F(A_k,x))=w(x)(f(x)-F(A,x)),\quad x\in Y$$

and it is readily seen that for all k sufficiently large, $\|w_k(f-F(A_k,\cdot))\|_X=\|w(f-F(A,\cdot))\|_X$. It follows that $w_k(f-F(A_k,\cdot))$ alternates exactly $d(A)$ times on X. Since $d(A_k)>d(A)$, $F(A_k,\cdot)$ is not best to f on X with respect to w_k. Let $\|w_k(f-F(B,\cdot))\|_X<\|w_k(f-F(A_k,\cdot))\|_X$, then $F(B,\cdot)-F(A_k,\cdot)$ alternates in sign on an alternant of $w(f-F(A,\cdot))$, hence $F(B,\cdot)-F(A_k,\cdot)$ has at least $d(A)$ zeros. It follows that if B_k is best to f on X with respect to w_k, $d(B_k)>d(A)$.

THEOREM 6. *Let F be degenerate at A and $w(f-F(A,\cdot))$ alternate exactly $d(A)$ times on X. There is a sequence of weights $\{w_k\}\to w$ such that if $\{T(f,X,w_k)\}$ exists, $T(f,X,w_k)\dashrightarrow T(f,X,w)$.*

PROOF. Suppose $T(f, X, w_k)=F(B_k, \cdot)$ for all k, then by the previous lemma, $w_k(f-F(B_k, \cdot))$ alternates at least $d(A)+1$ times on X. Thus $\{w_k(f-F(B_k, \cdot))\}$ cannot converge uniformly to $w(f-F(A, \cdot))$, which alternates exactly $d(A)$ times.

In case a best approximation always exists in approximation with respect to a positive weight function, the preceding theorem guarantees discontinuity of T by varying w. Such a case is that of approximation on an interval $[\alpha, \beta]$ by ordinary rational functions $R_m^n[\alpha, \beta]$ or by $F(A, x)=a_1 \exp(a_2 x)$.

A simple example of discontinuity of $T(f, [\alpha, \beta], \cdot)$ follows.

EXAMPLE. Let $[\alpha, \beta]=[-1, 1]$, $f(x)=x$, and $F(A, x)=a_1 \exp(a_2 x)$. Let $w=1$ and $w_k(x)=1+\frac{x}{k}$. As wf alternates once, 0 is the best approximation to f with respect to w. As $w_k f$ does not alternate, 0 is not a best approximation to f with respect to w_k and hence the best approximation $F(A_k, \cdot)$ to f with respect to w_k must be of degree 2, hence $w_k(f-F(A_k, \cdot))$ alternates at least twice. If we had $\{T(f, [\alpha, \beta], w_k)\} \to T(f, [\alpha, \beta], w)$, we would have $\{w_k(f-F(A_k, \cdot))\} \to wf$. But the sequence is composed of elements alternating at least twice, whereas wf is monotone.

Example the same thing happens in approximation by the family of ordinary rational functions $R_1^0[-1, 1]$, the families of ratios of constants p to first degree polynomials q, $q>0$ on $[-1, 1]$.

It appears (see [3, examples on 104—105 and 106—107; 10, Section 5]) that special properties of F must be used to obtain stronger discontinuity and local non-existence results. The property of irregularity has proven useful [9; 10; 11]. It would be of interest if this property could be proven useful in other cases, in particular with the weight variable (a general discontinuity result comparable to [11] would be desirable).

The continuity of ϱ has not been seriously studied, even for some special cases. It is well known that $\varrho(\cdot, X, w)$ must be continuous. However $\varrho(f, \cdot, w)$ need not be continuous [3, 100]. The behavior of $\varrho(f, X, \cdot)$ is completely unstudied. ϱ is continuous where T is continuous, but may be continuous where T is not continuous.

It is expected that comparable results hold for alternation with a fixed point [13]. The property corresponding to varisolvence is a special case of S-varisolvence, studied by KAUFMAN and BELFORD [14].

In the case that convergence does occur, the rate of convergence of T and ϱ is of interest. To get any results, it may be necessary to assume that (F, P) satisfy the hypotheses of Meinardus and Schwedt, given by BURKE [16], or of BARRAR and LOEB [15]. Studies have been made of rates with f varying [15] and of rates with discretization of an interval [16]. In the case of approximation by ordinary rational functions on an interval, a rate for f and w variable has been obtained by MAEHLY and WITZGALL [5, 289—300]. However no study has been made of rates when varying X always finite or always an interval, of rates when varying w, or of rates when varying all three arguments. A linear (first order) rate of convergence is expected under favorable circumstances. That this is best possible in general follows from consideration of approximation by constants and varying only f. It should be noted that in the special case of an interval, a quadratic (second order) rate occurs under circumstances generally encountered in practical problems.

REFERENCES

[1] DUNHAM, C. B.: Continuity of the Varisolvent Chebyshev Operator, *Bull. Amer. Math. Soc.* **74** (1968), 606—608.
[2] DUNHAM, C. B.: Necessity of Alternation, *Can. Math. Bull.* **10** (1968), 743—744.
[3] DUNHAM, C. B.: Alternating Chebyshev Approximation, *Trans. Amer. Math. Soc.* **178** (1973), 95—109.
[4] DUNHAM, C. B.: Varisolvent Chebyshev Approximation on Subsets, in *G. G. Lorentz (ed.), Approximation Theory*, Academic Press, New York, 1973, 337—340.
[5] MAEHLY, H., and WITZGALL, C.: Tschebyscheff-Approximationen in kleinen Intervallen II, *Numer. Math.* **2** (1960), 293—307.
[6] RICE, J. R.: Tchebycheff Approximations by Functions Unisolvent of Variable Degree, *Trans. Amer. Math. Soc.* **99** (1961), 298—302.
[7] DUNHAM, C. B.: Chebyshev approximation with respect to a vanishing weight function, *J. Approximation Theory* **12** (1974), 305—306.
[8] DUNHAM, C. B.: Nearby linear Chebyshev approximation, *Aequationes Mathematicae*, to appear.
[9] DUNHAM, C. B.: Approximation by alternating families on subsets, *Computing* **9** (1972), 261—265.
[10] DUNHAM, C. B.: Nonexistence of best alternating approximations on subsets, *Proc. Amer. Math. Soc.* **60** (1976), 203—206.
[11] SCHMIDT, E.: Discontinuity of the alternating Chebyshev operator (introduction by C. DUNHAM), *Rocky Mtn. J. Math.* (to appear).
[12] DUGUNDJI, J.: *Topology*, Allyn and Bacon, Boston, 1966.
[13] DUNHAM, C. B.: Alternation with a null point, *J. Approximation Theory* **15** (1975), 157—160.
[14] KAUFMAN, E. H., JR. and BELFORD, G. G.: A generalization of the varisolvency and unisolvency properties, *J. Approximation Theory* **7** (19 3), 21—35.
[15] BARRAR, R., and LOEB, H.: On the continuity of the nonlinear Tschebyscheff operator, *Pacific J. Math.* **32** (1970), 593—601.
[16] BURKE, M.: Nonlinear best approximations on discrete sets, *J. Approximation Theory* **16** (1976), 133—141.

Computer Science Department, University of Western Ontario,
London, Ontario N6A 5B9, Canada

(Received April 15, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 389—396.

COSTRICT RADICAL CLASSES OF ASSOCIATIVE RINGS

by

B. J. GARDNER and E. AHMED

This paper introduces a property of radical classes of rings which is at the same time a dualization and a generalization of strictness. One way of characterizing strict radical classes is in terms of their semi-simple rings: a radical class is strict if and only if subrings of semi-simple rings are semi-simple or equivalently, whenever B is semi-simple and there exists a ring monomorphism $A \to B$, A is semi-simple. A radical class will be called *costrict* if for every radical ring A and ring epimorphism $A \to B$ (the category-theoretical language is that of [4]), B is also radical. It turns out that strict radical classes are costrict, though the converse is false. Some characterizations of costrict radicals and a construction of the lower costrict radical class (in the obvious sense) are given and some examples of costrict and non-costrict radical classes presented.

The basic references for radical theory are [3] and [14]. All rings considered are associative.

1. Epimorphisms

A ring homomorphism f is an *epimorphism* if it satisfies the condition

$$gf = hf \Rightarrow g = h.$$

Any epimorphism f has a factorization $f = ip$ where i is injective, p is surjective and both i and p are epimorphisms. Thus we need only consider surjective and injective epimorphisms, and in the latter case we shall be mainly concerned with inclusion maps. If A is a subring of a ring B such that the inclusion $A \to B$ is an epimorphism, we call B an *epimorphic extension* of A and represent this by the notation $A \subseteq\!\!\!\rightarrow B$. Most of the literature on ring epimorphisms deals with the category $\hat{\mathbf{R}}$ of rings with identity elements and identity-preserving homomorphisms. The relevant connections between that category and $\mathbf{R}$, the category of rings and homomorphisms, in which we shall work, can be summarized briefly as follows. For any ring A, let A^1 denote the ring obtained by adjoining an identity to A in the standard way.

(1) If $f\colon A \to B$ is a homomorphism, we define $f^1\colon A^1 \to B^1$ by $f^1(a, n) = (f(a), n)$. Then f is an epimorphism in $\mathbf{R}$ if and only if f^1 is an epimorphism in $\hat{\mathbf{R}}$. (2) If $g\colon A \to B$ is an epimorphism in $\mathbf{R}$ and A has an identity, then B has an identity and g is an epimorphism in $\hat{\mathbf{R}}$. (3) Any epimorphism $h\colon R \to S$ in $\hat{\mathbf{R}}$ is an epimorphism in $\mathbf{R}$.

(Proof of (3): Let k, l be homomorphisms in $\mathbf{R}$ such that $kh=lh$. Then k and l induce maps k', l' to $k(e)Ak(e)=l(e)Al(e)$ where A is the range of k and l and e is the identity of S.)

If A and B are rings, we denote by dom (A, B) (the *dominion* of A in B) the subring of B consisting of those elements b of B for which $f(b)=g(b)$ whenever two homomorphisms $f, g\colon B\to R$ (any R) agree on A. dom (A, B) consists of all elements of B which have the form

$$(*) \qquad a+XPY$$

where $a\in A$, X is a $1\times m$ matrix over B, P is an $m\times n$ matrix over A^1 and Y is an $n\times 1$ matrix over B such that XP and PY are matrices over A. When A has an identity, dom (A, B) consists of elements of the form

$$XPY$$

where P is a matrix over A and otherwise the symbols have the same meaning as before. Clearly $A\subseteq\!\!\!\!\to B$ if and only if dom $(A, B)=B$. For more details see [7] and [8].

We conclude this section with a few results on preservation of properties by epimorphisms.

PROPOSITION 1.1. *Let A be a commutative ring, $f\colon A\to B$ an epimorphism. Then B is commutative.*

PROOF. Clearly only the case where f is an inclusion needs to be treated. In this case we have $A^1\subseteq\!\!\!\!\to B^1$, so by Proposition 1.3 of [12], B^1 is commutative. Hence B is too. □

For a ring A, we denote by $A^{(n)}$, $A[x]$ the ring of $n\times n$ matrices and the ring of polynomials over A. In $A^{(n)}$, $[a]_{ij}$ represents the matrix with a in the (i, j) position and 0 in all the others, $\Delta(a)$ the matrix $\sum_{i=1}^{n}[a]_{ii}$.

PROPOSITION 1.2. *If $A\subseteq\!\!\!\!\to B$ then $A^{(n)}\subseteq\!\!\!\!\to B^{(n)}$, for all n.*

PROOF. By $(*)$, every element b of B has the form

$$b = a+\sum_{i,j} b_i a_{ij} b'_j$$

where $a\in A$, $b_i, b'_j\in B$, $a_{ij}\in A^1$, $\sum_i b_i a_{ij}$, $\sum_i a_{ij}b'_j\in A$. It follows that for any k, l, we have

$$[b]_{kl} = [a]_{kl}+\sum_{i,j}[b_i]_{ks}\Delta(a_{ij})[b'_j]_{sl} \quad \text{(any } s\text{)}$$

where $[a]_{kl}\in A^{(n)}$, $[b_i]_{ks}, [b'_j]_{sl}\in B^{(n)}$, $\sum_i [b_i]_{ks}\Delta(a_{ij})=\left[\sum_i b_i a_{ij}\right]_{ks}\in A^{(n)}$ and $\sum_j \Delta(a_{ij})[b'_j]_{sl}=$ $=\left[\sum_j a_{ij}b'_j\right]_{sl}\in A^{(n)}$. Also (identifying $(A^{(n)})^1$ with the subring of $(A^1)^{(n)}$ generated by $A^{(n)}$ and the matrices $\Delta(n)$, $n\in Z$) $\Delta(a_{ij})\in(A^{(n)})^1$, for all i, j. Hence $[b]_{kl}\in$ dom $(A^{(n)}, B^{(n)})$ for all $b\in B$ and for all k, l. It follows that dom $(A^{(n)}, B^{(n)})=B^{(n)}$. □

PROPOSITION 1.3. *If $A\subseteq\!\!\!\!\to B$, then $A[x]\subseteq\!\!\!\!\to B[x]$.*

PROOF. By a similar argument to the one just given, one can show that b and bx^n belong to dom $(A[x], B[x])$ for all $b\in B$ and for all n. □

2. Costrict radical classes

Recall that a radical class $\mathscr{R}$ is *strict* if for every ring A, $\mathscr{R}(A)$ contains all $\mathscr{R}$-subrings of A. An equivalent characterization, more relevant to our present discussion, is the following: a radical class is strict if and only if the corresponding semi-simple class is closed under formation of subrings. (Further information concerning strict radical classes can be obtained from [11].) Since the monomorphisms of rings are precisely the injections, the following property of radical classes is, in a sense, dual to strictness. A radical class $\mathscr{R}$ will be called *costrict* if it satisfies:

$$\text{If } A\in\mathscr{R} \text{ and } f\colon A\to B \text{ is an epimorphism, then } B\in\mathscr{R}.$$

Since radical classes are homomorphically closed, we have immediately

PROPOSITION 2.1. *A radical class is costrict if and only if it is closed under formation of epimorphic extensions.* □

From another point of view, costrictness is a generalization of strictness:

PROPOSITION 2.2. *Strict radical classes are costrict.*

PROOF. Let A be a ring in a strict radical class $\mathscr{R}$. If $A\underset{\to}{\subseteq}B$ then $\mathscr{R}$ contains A^*, the ideal of B generated by A [11]. But the natural map and the zero map from B to B/A^* agree on A, so $A^*=B$. □

Further characterizations of costrict radical classes are given by the following theorem.

THEOREM 2.3. *Let $\mathscr{R}$ be a radical class, $\mathscr{S}$ the corresponding semi-simple class. The following conditions are equivalent:*

(i) *$\mathscr{R}$ is costrict;*

(ii) *If $A\underset{\to}{\subseteq}B\in\mathscr{S}$ and $A\in\mathscr{R}$, then $A=0$;*

(iii) *If A is a ring such that, for every non-zero homomorphic image A'', there exists a chain*

$$0\neq R\underset{\to}{\subseteq}I\lhd A'',$$

with $R\in\mathscr{R}$, then $A\in\mathscr{R}$.

PROOF. (i)⇒(ii): This is clear.

(ii)⇒(iii): Suppose A satisfies the condition but $A\notin\mathscr{R}$. Then $A/\mathscr{R}(A)\neq 0$, so there is a chain

$$0\neq R\underset{\to}{\subseteq}I\lhd A/\mathscr{R}(A),$$

with $R\in\mathscr{R}$. But I is in $\mathscr{S}$, so this can't happen.

(iii)⇒(i): Let A be in $\mathscr{R}$, with $A\underset{\to}{\subseteq}B$. If $B/I\neq 0$, then $A\nsubseteq I$ (cf. Proposition 2.2) so we have a commutative diagram

$$\begin{array}{ccc} A & \longrightarrow & B \\ \downarrow & & \downarrow \\ 0\neq A/A\cap I & \cong & (A+I)/I\to B/I \end{array}$$

where $A/A\cap I$ is in $\mathscr{R}$ and all maps are epimorphisms. By (iii), B belongs to $\mathscr{R}$. □

3. The lower costrict radical construction

In this section we define a construction, analogous to the lower radical construction, which yields the smallest costrict radical class containing a given class.

Let $\mathscr{M}$ be a non-empty class of rings. We define the classes $\mathscr{M}_{(\alpha)}$ inductively as follows:

$\mathscr{M}_{(1)} = \{B | \exists$ an epimorphism $A \to B$ with $A \in \mathscr{M}\}$.

$\mathscr{M}_{(\alpha)} = \{B |$For every non-zero homomorphic image B'' of B, $\exists$ a chain $0 \neq R \subseteq\!\!\!\!\rightarrow I \lhd B''$ with $R \in \mathscr{M}_{(\alpha)}$ for some $\alpha < \beta\}$.

THEOREM 3.1. *With the notation used above,* $\bigcup_{\alpha} \mathscr{M}_{(\alpha)}$ *is the smallest costrict radical class containing* $\mathscr{M}$.

PROOF. It is straightforward to show that $\mathscr{M}_{(\alpha)} \subseteq \mathscr{M}_{(\beta)}$ whenever $\alpha \leqq \beta$ and that each $\mathscr{M}_{(\alpha)}$ is homomorphically closed, whence $\bigcup_{\alpha} \mathscr{M}_{(\alpha)}$ is also.

Let B be a ring such that every non-zero homomorphic image B'' has a chain

$$0 \neq R \subseteq\!\!\!\!\rightarrow I \lhd B''$$

with $R \in \bigcup_{\alpha} \mathscr{M}_{(\alpha)}$. Let $\{J_s | s \in S\}$ be the set of all non-zero ideals of B, and for each s let

$$\alpha_s = \min \{\alpha | \exists \text{ a chain } 0 \neq R \subseteq\!\!\!\!\rightarrow I \lhd B/J_s \text{ with } R \in \mathscr{M}_{(\alpha)}\}.$$

Then for any β which is greater than all the α_s, we have $B \in \mathscr{M}_{(\beta)} \subseteq \bigcup_{\alpha} \mathscr{M}_{(\alpha)}$. As a special case, if every $B'' \neq 0$ has a non-zero ideal in $\bigcup_{\alpha} \mathscr{M}_{(\alpha)}$, then B is in $\bigcup_{\alpha} \mathscr{M}_{(\alpha)}$. This, together with the fact that $\bigcup_{\alpha} \mathscr{M}_{(\alpha)}$ is homomorphically closed, means that $\bigcup_{\alpha} \mathscr{M}_{(\alpha)}$ is a radical class, while from the general case and Theorem 2.3, we see that $\bigcup_{\alpha} \mathscr{M}_{(\alpha)}$ is costrict.

Finally, if $\mathscr{U}$ is a costrict radical class containing $\mathscr{M}$, a transfinite induction argument making use of Theorem 2.3 (iii) establishes that $\bigcup_{\alpha} \mathscr{M}_{(\alpha)} \subseteq \mathscr{U}$. □

We shall call the radical class $\bigcup_{\alpha} \mathscr{M}_{(\alpha)}$ the *lower costrict radical class* defined by $\mathscr{M}$ and represent it by the notation $L_C(\mathscr{M})$.

4. Examples

As was shown in [7], every epimorphism $A \to B$, where A is (von Neumann) regular, is surjective. This provides some examples of costrict radical classes which are not strict.

EXAMPLE 4.1. The (radical) class $\mathscr{V}$ [2] of all regular rings, and all its radical subclasses, which include the semi-simple radical classes [10], are costrict. None

of these is strict, however. For any radical class $\mathscr{R}$, the class $\mathscr{R}^* = \{A \mid A[x] \in \mathscr{R}\}$ is also radical. If $\mathscr{R}$ is strict, then $\mathscr{R}^* = \mathscr{R}$ ([11], Proposition 3.1) while if $\mathscr{R}$ is hereditary and consists of idempotent rings, $\mathscr{R}^* = \{0\}$ ([5], Theorem 10). Thus there are no hereditary strict radical classes $\neq \{0\}$ consisting of idempotent rings. A similar argument shows that no non-trivial radical subclass $\mathscr{R}$ of $\mathscr{V}$ is strict, as we would then have $\mathscr{R} = \mathscr{R}^* \subseteqq \mathscr{V}^* = \{0\}$.

Some further examples are provided by the next result.

PROPOSITION 4.2. *Let $\mathscr{M}$ be a non-empty class of rings such that*

(i) *$\mathscr{M}$ is homomorphically closed,*

(ii) *if $I \lhd J \lhd A$ and $I \in \mathscr{M}$, then the ideal I^* of A generated by I is in $\mathscr{M}$, and*

(iii) *if $0 \neq A \lhd B \underset{\rightarrow}{\subseteq} C$ and $A \in \mathscr{M}$, then C has a non-zero ideal in $\mathscr{M}$.*

Then $L(\mathscr{M})$, the lower radical class defined by $\mathscr{M}$, is costrict.

PROOF. Let A be a ring such that for every non-zero homomorphic image A'', there is a chain

$$0 \neq R \underset{\rightarrow}{\subseteq} I \lhd A''$$

with $R \in L(\mathscr{M})$. Then this chain can be augmented by a finite number of terms to give a chain

$$0 \neq I_1 \lhd \ldots \lhd I_n \lhd R \underset{\rightarrow}{\subseteq} I \lhd A''$$

where $I_1 \in \mathscr{M}$. By (ii), R has a non-zero ideal $J \in \mathscr{M}$, so by (iii) I has a non-zero ideal $K \in \mathscr{M}$. We have $K \lhd I \lhd A''$, so by (ii) again, A'' has a non-zero ideal in $\mathscr{M}$. But A'' is an arbitrary non-zero homomorphic image of A, so A is in $L(\mathscr{M})$. By Theorem 2.3, $L(\mathscr{M})$ is costrict. □

COROLLARY 4.3. *Let $\mathscr{M}$ be a non-empty epimorphically closed class of rings with identity elements. Then $L(\mathscr{M})$ is costrict.*

PROOF. Being idempotent, the rings in $\mathscr{M}$ satisfy condition (ii) of Proposition 4.2. If A is in $\mathscr{M}$ and $A \lhd B \underset{\rightarrow}{\subseteq} C$, then adjoining identities to B and C we get $A \lhd B^1 \underset{\rightarrow}{\subseteq} C^1$. Since A has an identity, we can write $B^1 = A \oplus R$ (ring direct sum). Let e be the identity of A; then by Corollary 2.2 of [12], eC^1e is an epimorphic extension of A and an ideal of C^1. But $eC^1e \subseteqq C$, so $eC^1e \lhd C$. This establishes condition (iii) of Proposition 4.2. □

A class can be closed under epimorphisms and yet define a lower radical class which is not costrict.

EXAMPLE 4.4. The class of commutative rings is closed under epimorphisms (Proposition 1.1). Let $\mathscr{R}$ be its lower radical class. Then for any (commutative) field K, $\mathscr{R}$ contains $\begin{bmatrix} 0 & K \\ 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} K & 0 \\ 0 & K \end{bmatrix} \cong K \oplus K$, so $\mathscr{R}$ contains $\begin{bmatrix} K & K \\ 0 & K \end{bmatrix}$. The latter has the full matrix ring $K^{(2)}$ as an epimorphic extension ([8], p. 268) but $K^{(2)}$ clearly is not in $\mathscr{R}$.

PROPOSITION 4.5. *Let $\mathscr{R}$ be a costrict radical class,*

$$\mathscr{R}^* = \{A \mid A[x] \in \mathscr{R}\} \text{ and } \mathscr{R}_n = \{A \mid A^{(n)} \in \mathscr{R}\}.$$

Then $\mathscr{R}^*$ and $\mathscr{R}^{(n)}$ are costrict radical classes.

PROOF. $\mathscr{R}^*$ is a radical class ([5], Theorem 1) and so is each $\mathscr{R}^{(n)}$ ([9], p. 61). The rest follows from propositions 1.2 and 1.3 □

Finally we give some examples of radical classes which are not costrict. In what follows, $\mathscr{J}$ is the Jacobson radical class, $\mathscr{K}$ the corpoidal radical class [13].

PROPOSITION 4.6. *Let $\mathscr{R}$ be a radical class such that*

$$\mathscr{J} \subseteqq \mathscr{R} \subseteqq \mathscr{K}.$$

Then $\mathscr{R}$ is not costrict.

PROOF. Let A be the ring of rational numbers of the form $\frac{2m}{2n+1}$. Then $A \in \mathscr{J}$ but the field Q of rational numbers is an epimorphic extension of A. If $f, g: Q \to R$ are non-zero homomorphisms which agree on A, then in particular $2f(1)=f(2)=$ $=g(2)=2g(1)$. The additive group of $f(A)+g(A)$ is torsion-free and contains $f(1)$ and $g(1)$. Since $2(f(1)-g(1))=0$, we conclude that $f(1)=g(1)$. Call this common value e. Then f and g have the same image — call it $R=eRe$ — and may be regarded as isomorphisms from Q to R. For any integers m, n with $n \neq 0$ we have

$$f\left(\frac{m}{n}\right)f(n) = f(m) = me = g(m) = g\left(\frac{m}{n}\right)g(n) = g\left(\frac{m}{n}\right)f(n),$$

so $f\left(\frac{m}{n}\right)=g\left(\frac{m}{n}\right)$. On the other hand, $\mathscr{K}(Q)=0$. □

5. The lower costrict radical class defined by the class of zerorings

Let $\mathscr{Z}$ denote the class of zerorings. Clearly we have $\mathscr{B} \subseteqq L_C(\mathscr{Z}) \subseteqq \mathscr{N}_g$, where $\mathscr{B}$ is the prime, $\mathscr{N}_g$ the generalized nil, radical class, since the latter is strict and hence (by Proposition 2.2) costrict. In this section we shall obtain a little more information about the position of $L_C(\mathscr{Z})$ among other radical classes.

PROPOSITION 5.1. $L_C(\mathscr{Z}) \neq \mathscr{N}_g$.

PROOF. The class $\mathscr{Z}$ is closed under epimorphisms — nilpotent rings have no proper epimorphic extensions ([8], p. 267) — so we look at the classes $\mathscr{Z}_{(\alpha)}$ of the construction in § 3. Suppose $L_C(\mathscr{Z})$ contains a full matrix ring over a finite field; let β be the least ordinal for which the corresponding $\mathscr{Z}_{(\beta)}$ contains such a ring and let $K^{(n)}$ be in $\mathscr{Z}_{(\beta)}$ for a finite field K. Since $K^{(n)}$ is simple, there is a ring $R \neq 0$ in some $\mathscr{Z}_{(\alpha)}$, with $\alpha < \beta$, such that $R \subseteq K^{(n)}$. Now R is not nilpotent, by the result quoted at the beginning of this proof, so, being finite, it has a semi-simple artinian image, necessarily in $\mathscr{Z}_{(\alpha)}$. This violates the minimality condition on β, so $L_C(\mathscr{Z})$ can contain no $K^{(n)}$. On the other hand, $\mathscr{N}_g$ contains them all. □

COROLLARY 5.2. *$L_C(\mathscr{Z})$ is not strict.* □

It is not known whether or not $L_C(\mathscr{Z})=\mathscr{B}$. The next two results show, however, that certain subclasses of $\mathscr{B}$ are closed under epimorphisms. Let $\mathscr{C}$ denote the class of commutative rings.

PROPOSITION 5.3. *$\mathscr{B} \cap \mathscr{C}$ is closed under epimorphisms and $\mathscr{B} \cap \mathscr{C} = L_C(\mathscr{Z}) \cap \mathscr{C}$.*

PROOF. $\mathscr{B}\cap\mathscr{C}\subseteq L_C(\mathscr{Z})\cap\mathscr{C}\subseteq\mathscr{N}_g\cap\mathscr{C}=\mathscr{B}\cap\mathscr{C}$, while $\mathscr{C}$ is closed under epimorphisms (Proposition 1.1) so $L_C(\mathscr{Z})\cap\mathscr{C}$ is too. □

PROPOSITION 5.4. *If A is a (left or right) T-nilpotent ring, then so is any epimorphic extension of A.*

PROOF. Suppose A is left T-nilpotent. This is equivalent to every non-zero homomorphic image of A having a non-zero left annihilator ([6], Theorem 1.2). If $A\underset{\to}{\subseteq}B$ and B'' is a non-zero homomorphic image of B, then as in the proof of Theorem 2.3, we have $A''\underset{\to}{\subseteq}B''$, where A'' is a non-zero homomorphic image of A. Let $(0\colon A'')$ denote the left annihilator of A''. As shown in § 1, every element in B'' has the form $a+\Sigma a_i b_i$, where a and the a_i belong to A'' and the b_i to B''. Clearly $(0\colon A'')$ annihilates B''. Hence B is left T-nilpotent. □

Using Proposition 4.5, it's easy to show that if a ring R belongs to $L_C(\mathscr{Z})$, then so do $R[x]$ and $R^{(n)}$ for each n. If $L_C(\mathscr{Z})\subseteq\mathscr{J}$, then for every $R\in L_C(\mathscr{Z})$ we have $R[x]\in\mathscr{J}$, so by Lemmas 2J and 3J of [1], R is nil.

REFERENCES

[1] AMITSUR, S. A.: Radicals of polynomial rings, *Canad. J. Math.* **8** (1956), 355—361.
[2] BROWN, B., MCCOY, N. H.: The maximal regular ideal of a ring, *Proc. Amer. Math. Soc.* **1** (1950), 165—171.
[3] DIVINSKY, N. J.: *Rings and radicals,* Allen and Unwin, London, 1965.
[4] FREYD, P.: *Abelian categories,* Harper and Row, New York, Evanston and London, 1964.
[5] GARDNER, B. J.: A note on radicals and polynomial rings, *Math. Scand.* **31** (1972), 83—88.
[6] GARDNER, B. J.: Some aspects of T-nilpotence, *Pacific J. Math.* **53** (1974), 117—130.
[7] GARDNER, B. J.: Epimorphisms of regular rings, *Comment. Math. Univ. Carolinae* **16** (1975), 151—160.
[8] ISBELL, J. R.: Epimorphisms and dominions. IV, *J. London Math. Soc.* (2) **1** (1969), 265—273.
[9] KREMPA, J.: Radicals of semi-group rings, *Fund. Math.* **85** (1974), 57—71.
[10] STEWART, P. N.: Semi-simple radical classes, *Pacific J. Math.* **32** (1970), 249—254.
[11] STEWART, P. N.: Strict radical classes of associative rings, *Proc. Amer. Math. Soc.* **39** (1973), 273—278.
[12] STORRER, H. H.: Epimorphic extensions of non-commutative rings, *Comment. Math. Helv.* **48** (1973), 72—86.
[13] THIERRIN, G.: Sur le radical corpoïdal d'un anneau, *Canad. J. Math.* **12** (1960), 101—106.
[14] WIEGANDT, R.: *Radical and semisimple classes of rings,* Queen's Papers in Pure and Applied Mathematics, No. 37, Kingston, Ontario, 1974.

University of Tasmania, Hobart, Australia

(Received May 1, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 397—412.

LARGE α-CRITICAL GRAPHS WITH SMALL DEFICIENCY (ON LINE-CRITICAL GRAPHS II)

by

L. SURÁNYI

0. Introduction

If G is a graph, $V(G)$, $E(G)$ and $\alpha(G)$ denotes its vertex-, edge-set and the maximum number of independent vertices in G, resp.; G is *α-critical* if $\alpha(G)=\alpha$, but $\alpha(G-e)>\alpha(G)$ for all $e\in E(G)$.

$(m, n, k)\to n+l$ ($(m, n, k)\not\to n+l$) means, that the following statement is true (false): For any graph G with m vertices and $\alpha(G)<n+l$, there exists a subset X of $V(G)$, satisfying $|X|\leqq k$ and $\alpha(G-X)<n$.

This symbol was defined and first investigated by P. ERDŐS in [2]. (The equivalence of his definition and the one given here is proved in [7], lemma 5.3.) This symbol was studied in [2], [4], [6], [7], [8]; [5] studies the analogous question for t-graphs. Let $f(n, k, l)=\max\{m\colon (m, n, k)\to n+l\}$. (Clearly $1\leqq l\leqq k$, $1\leqq n$ may be assumed.) It was proved by Erdős, that $f(n, k, l)=2n+k-2$ for $\left[\frac{k+1}{2}\right]\leqq l\leqq k$. However the determination of $f(n, 3, 1)$ seems to be more difficult. It was proved in [6] and [7], that $2n+3\leqq f(n, 3, 1)\leqq\frac{5}{2}n+9$. In Theorem 1.11. of this paper we prove $f(n, 3, 1)-2n=o(n)$, in fact $f(n, 3, 1)-2n<10\left[\sqrt{n}\right]$ for all $n\geqq 1$. We do not know whether $\lim_{n\to\infty}(f(n, 3, 1)-2n)=\infty$ or $f(n, 3, 1)-2n=o(\sqrt{n})$ holds. In Theorem 4 we give a general non-trivial upper bound for $f(n, k, l)$.

If one wants to verify $(m, n, k)\to n+l$ for some m, n, k, l, one clearly may restrict oneself to the $(n+l-1)$-critical graphs having m vertices. On the other hand whenever X contains a vertex y and all its neighbours in G, one always has $\alpha(G-X)<\alpha(G)$. That is why the investigation of $f(n, k, l)$ is related to the following (unsolved) problem:

Let δ be any fixed number. Does there exist a number $\alpha_0(\delta)$ with the following property: Whenever G is a connected α-critical graph, $\alpha>\alpha_0(\delta)$, and $\delta=|V(G)|-2\alpha(G)$, we always find a vertex in G with degree $\leqq 2$.

A theorem of HAJNAL ([3], Theorem 1) shows that $\delta\geqq 0$ and $\alpha_0(0)=\alpha_0(1)=1$, ANDRÁSFAI in [1] proved, that $\alpha_0(2)=1$. $\alpha_0(3)\leqq 12$ was proved in [7], Theorem 5. It was also shown there, that an affirmative answer to this question implies $|E(G)|-|V(G)|\leqq f(\delta)$ (with some function f depending only on $\delta=|V(G)|-2\alpha(G)$, for all connected α-critical graphs G. This means, that if $\alpha\to\infty$, G is connected α-critical and $\delta=|V(G)|-2\alpha(G)$ is fixed, then "almost all' vertices of G have degree 2. (See [7] 4.2). There are several constructions known proving $\alpha_0(\delta)>c\cdot\delta$ with different numerical constants c. (See e.g. [9], [10], [11], [6].) In Theorems 2.6 and 2.8 of this paper we prove $\alpha_0(\delta)\geqq\frac{\delta^2}{4}$ for any even δ (i.e. $\alpha_0(\delta)\geqq\left[\frac{\delta}{2}\right]^2$ in general).

We use the following non-standard notations:

$$E(G, A, B) = \{yz : y \in A,\ z \in B,\ yz \in E(G)\}, \quad \varrho(G, x)$$

is the degree of the vertex x in G, $\mathscr{F}(\beta, G) = \{F \subset V(G)$: F is independent in G, $|F| = \beta\}$, $\mathscr{F}(G) = \bigcup_{\beta} \mathscr{F}(\beta, G)$. $C(q)$ denotes the set of residue classes modulo q; we sometimes identify k with the residue class of $C(q)$ containing it, but this will cause no confusion. Also if there is no danger of confusion, we identify the set $A \subset V(G)$ with the subgraph of G spanned by A. We put $\eta(r) = \frac{1}{2}\left(r+1+\left[\frac{r+1}{2}\right]\right)$ for any integer r.

1. The construction

1.1. First we define the graph $\mathscr{B}_k$ for any odd $k \geqq 3$.

$$V(\mathscr{B}_k) = \{\langle i, p, q\rangle : i \in C(k+1),\ p \in C(k),\ q \in C(2)\};$$

$$E(\mathscr{B}_k) = \left\{\langle i, p, 0\rangle \langle i + \frac{k-1}{2}, p, 1\rangle : i \in C(k+1),\ p \in C(k)\right\} \cup$$

$$\left\{\langle i, p, 0\rangle \langle i + \frac{k+1}{2}, p, 1\rangle : i \in C(k+1),\ p \in C(k)\right\} \cup$$

$$\left\{\langle i, p, q\rangle \langle i + \frac{k+1}{2}, p+1, q\rangle : i \in C(k+1),\ p \in C(k),\ q \in C(2)\right\}.$$

This is the graph that will prove 1.11 (i). This graph is "almost" a Descartes-product of a $2k+2$-circuit and a k-circuit. (See in 2.7 the definition of $\mathscr{C}_k$.) Note, that this graph is *not* symmetric, only anti-symmetric in q in the following sense: $\Phi_0(\langle i, p, q\rangle) = \langle i, p, 1-q\rangle$ is *not* an isomorphism on $\mathscr{B}_k$, only $\Phi_1(\langle i, p, q\rangle) = \langle -i, p, 1-q\rangle$ *is* one. To avoid the complications in the technical details of the proof this lack of symmetry would cause (see 1.9), we construct a graph G_k, symmetric and anti-symmetric in q, that has following properties:

$(*)$ $\alpha(G_k) = 2\alpha(\mathscr{B}_k) = 2k^2$.

$(**)$ $V(G_k) = \mathscr{B}^k(0) \cup \mathscr{B}^k(1)$ *where both* $\mathscr{B}^k(0)$ *and* $\mathscr{B}^k(1)$ *span a graph isomorphic to* $\mathscr{B}_k$ *in* G_k; $\mathscr{B}^k(0)$ *and* $\mathscr{B}^k(1)$ *are (vertex-) disjoint.*

$(***)$ *Deleting any 2 vertices* x *and* y *from* G_k, *the graph* $G_k - x - y$ *contains 2 disjoint independent sets with* $\alpha(G_k)$ *vertices, each.*

Now it is easy to see that in view of $(*)$ and $(**)$, the corresponding statement of $(***)$ holds for $\mathscr{B}_k$, as well. This, however, yields 1.11 (i).

So we define G_k. It will be symmetric and anti-symmetric in q, but will look very much like $\mathscr{B}_k$. Put for any p and q[1], and any integer l:

$$A_l(p, q) = \{\langle i, p, q\rangle : i \in C(l)\} \quad \text{and} \quad A_l(p) = A_l(p, 0) \cup A_l(p, 1).$$

$V(\mathscr{B}_k)$ is the union of the sets $A_{k+1}(p)$ for all p's.

[1] In this chapter p and q — with or without indices — will always denote members of $C(k)$ resp. $C(2)$.

Similarly to this we define

$$V(G_k) = \bigcup_{p \in C(k)} A_{2k+2}(p).$$

We want to prove $(***)$ in the following way: for any x, y we find an F and an F_1 in $\mathscr{F}(2k^2, G_k - x - y)$ such that $F \cap F_1 = \varnothing$ and $|(F \cup F_1) \cap A_{2k+2}(p, q)| = 2k$. Now the fact that we have 2 more vertices in every $A_{2k+2}(p, q)$ — instead of 1, which would be the case if we considered $\mathscr{B}_k$ — makes this possible in all cases, and this is the other point (besides the above mentioned symmetry) why we prove $(***)$ for G_k instead of $\mathscr{B}_k$. So we define

$$E(G_k) =$$

$$= \{\langle i_0, p, q\rangle\langle i_1, p, q\rangle : i_0, i_1 \in C(2k+2),\ p \in C(k),\ q \in C(2),\ i_0 - i_1 = k,\ k+1 \text{ or } k+2\} \cup$$

$$\cup \{\langle i, p, q\rangle\langle i+k+1, p+1, q\rangle : i \in C(2k+2),\ p \in C(k),\ q \in C(2)\}.$$

Now we prove $(*)$ and $(**)$. Put

$$\mathscr{B}^k(q) = \bigcup_{p \in C(k)} (\{\langle 2j+1, p, q\rangle : 0 \leqq j \leqq k\} \cup \{\langle 2j, p, 1-q\rangle : 0 \leqq j \leqq k\}).$$

Then clearly $\mathscr{B}^k(0)$, $\mathscr{B}^k(1)$ and $\mathscr{B}_k$ are isomorphic. This proves $(**)$. Next we show, that

1.2. $\alpha(G_k) \leqq 2k^2$.

Since $\alpha(G_k) \leqq \alpha(\mathscr{B}^k(0)) + \alpha(\mathscr{B}^k(1)) = 2\alpha(\mathscr{B}_k)$, so we only have to prove

1.3. $\alpha(\mathscr{B}_k) \leqq k^2$.

As we mentioned in 1.1, $A_{k+1}(p)$ spans a circuit of length $2k+2$ in $\mathscr{B}_k$ so $\mathscr{F}(k+1, A_{k+1}(p)) = \{A_{k+1}(p, 0), A_{k+1}(p, 1)\}$. Let F be an independent set in $\mathscr{B}_k$. If $|F \cap A_{k+1}(p)| \leqq k$ for all p, then $|F| \leqq k^2$. So next assume $|F \cap A_{k+1}(p)| = k+1$. But then $F \cap A_{k+1}(p) = A_{k+1}(p, q)$ for some q. So we may assume because of obvious symmetries, that, $F \cap A_{k+1}(0) = A_{k+1}(0, 0)$. Now $A_{k+1}(p+1, q) \cup A_{k+1}(p, q)$ has $2k+2$ elements, spans a 1-factor, so $|F \cap (A_{k+1}(p, q) \cup A_{k+1}(p+1, q))| \leqq k+1$. So $F \supset A_{k+1}(0, 0)$ implies $F \cap A_{k+1}(1, 0) = F \cap A_{k+1}(-1, 0) = \varnothing$ and so

$$|F| = |F \cap A_{k+1}(0)| + \sum_{1 \leqq j \leqq \frac{k-3}{2}} |F \cap (A_{k+1}(2j, 0) \cup A_{k+1}(2j+1, 0))| +$$

$$+ \sum_{1 \leqq j \leqq \frac{k-1}{2}} |F \cap (A_{k+1}(2j-1, 1) \cup A_{k+1}(2j, 1))| \leqq$$

$$\leqq |A_{k+1}(0, 0)| + \frac{k-3}{2}(k+1) + \frac{k-1}{2}(k+1) = (k-1)(k+1) < k^2.$$

This proves 1.3 and so 1.2.

1.4. Our next aim is to define certain elements of $\mathscr{F}(2k^2, G_k)$. This will also yield $\alpha(G_k) = 2k^2$ and $\alpha(\mathscr{B}_k) = k^2$.

We need some notations. We put for each $i\in C(2k+2)$, p, q, $0\leqq r\leqq 2k+2$

$$I(p, q, i, r) = \left\{\langle i+t, p, q\rangle : -\frac{r-1}{2} \leqq t \leqq \frac{r-1}{2}\right\} \quad \text{if } r \text{ is odd},$$

$$I\left(p, q, i+\frac{1}{2}, r\right) = \left\{\langle i+t, p, q\rangle : -\frac{r}{2}+1 \leqq t \leqq \frac{r}{2}\right\} \quad \text{if } r \text{ is even}.$$

So $I(p, q, u, r)$ is the arc of $A_{2k+2}(p, q)$ with length r and centre u.

The following observations are trivial consequences of the definitions:

1.5. *Let* $0\leqq r\leqq 2k+2$ *and* u *such that* $u-\eta(r)\in C(2k+2)$. *Then*

(i) $I(p, q, u, r)\cup I(p, 1-q, u, t)\in \mathscr{F}(A_{2k+2}(p))$ *for all* p, q, $0\leqq t\leqq 2k-r$ *if* $t\equiv r\ (2)$.

(ii) $I(p, q, u, r)\cup I(p+1, q, u, t)\in \mathscr{F}(A_{2k+2}(p, q)\cup A_{2k+2}(p+1, q))$ *for all* p, q, $0\leqq t\leqq 2k+2-r$ *if* $t\equiv r\ (2)$.

For all p, q, $0\leqq r\leqq 2k$ and for all u with $u-\eta(r)\in C(2k+2)$ we put

$$F_p(q, u, r) = I(p, q, u, r)\cup I(p, 1-q, u, 2k-r).$$

If $0\leqq t\leqq k$, then t^* denotes 1 or 0 corresponding to whether t is odd or even. Now for all $q=0, 1$; p; $0\leqq t_0\leqq \frac{k-1}{2}$; $\beta=\pm 1$ we define

(iii) $$F^0(p, q, u) = \bigcup_{t=0}^{k-1} F_{p+t}((t+q)^*, u, 2t) = \bigcup_{t=0}^{k-1} F_{p-t}((t+q)^*, u, 2t) =$$

$$= \bigcup_{\substack{t=0\\ t \text{ even}}}^{k-1} F_{p+t}(q, u, 2t)\cup \bigcup_{\substack{t=0\\ t \text{ odd}}}^{k-1} F_{p+t}(q, u, 2k-2t) \quad \text{if } u-\frac{1}{2}\in C(2k+2),$$

(iv) $$F^1(p, q, u) = \bigcup_{t=0}^{k-1} F_{p+t}((t+q)^*, u, 2t+1) = \bigcup_{t=0}^{k-1} F_{p-1-t}(1-(t+q)^*, u, 2t+1) =$$

$$= \bigcup_{\substack{t=0\\ t \text{ even}}}^{k-1} F_{p+t}(q, u, 2t+1)\cup \bigcup_{\substack{t=0\\ t \text{ odd}}}^{k-1} F_{p+t}(q, u, 2k-2t-1), \quad \text{if } u\in C(2k+2),$$

(v) $$F^k(p, u, t_0, \beta) = \bigcup_{|t|<t_0} F_{p+t}(0, u+\beta|t|, k)\cup \bigcup_{t_0\leqq|t|\leqq\frac{k-1}{2}} F_{p+t}(0, u+\beta t_0, k),$$

if $u\in C(2k+2)$.

Now we define the adiungant of any set F, and we denote it by F_1:

$$F_1 = \{\langle i, p, q\rangle : \langle i+k+1, p, 1-q\rangle\in F\}.$$

1.5. and trivial calculations yield, that

1.6. (i) *The sets defined in* 1.5. (iii)—(v) *belong to* $\mathscr{F}(2k^2, G_k)$.

(ii) $\alpha(G_k)=2k^2$, $\alpha(\mathscr{B}_k)=k^2$.

(iii) *The following equations hold, whenever the left-hand sides are defined:*

$$F_1^j(p, q, u) = F^j(p, 1-q, u+k+1), \quad F_1^k(p, u, t_0, \beta) = F^k(p, u, +k+1, t_0, \beta).$$

(iv) *If* F *is a set defined by* 1.5 (iii)—(v), *then* $F\cap F_1=\varnothing$ *and for some* u_0, u_1

$$F\cap A_{2k+2}(p, q) = \{\langle i, p, q\rangle : u_0 \leqq i \leqq u_1\},$$

$$F_1\cap A_{2k+2}(p, q) = \{\langle i, p, q\rangle : u_1+2 \leqq i \leqq 2k+u_0\}.$$

1.7. DEFINITION: We say a graph G has *property* $\mathbf{u}_l^j$ for some $j, l\geqq 0$ if $\mathscr{F}(\alpha(G), G-H)$ has j disjoint elements whenever $H\subset V(G)$ and $|H|=l$.

1.8.

(i) *If* G *has property* $\mathbf{u}_l^j$ *for some* $l\geqq 0$, $j\geqq 1$ *then* G *has property* $\mathbf{u}_{l'}^{j'}$ *for all* $0\leqq l'\leqq l$, $1\leqq j'\leqq j$.

(ii) *If* G *has property* $\mathbf{u}_l^j$ *for some* $l\geqq 0$, $j\geqq 1$, *then* G *has property* $\mathbf{u}_{l+j'}^{j-j'}$ *for all* $0\leqq j'\leqq j-1$.

(iii) *Suppose that the spanned subgraphs* G_i, $1\leqq i\leqq m$ *satisfy* $V(G_i)\cap V(G_{i'})=\varnothing$ *for* $i\neq i'$, *and let* $V(G)=\bigcup_1^m V(G_i)$, $\alpha(G)=\sum_1^m \alpha(G_i)$. *Then* G *has property* $\mathbf{u}_l^j$ *iff all* G_i's *have it.*

These assertions are trivial. Now we want to prove, that G_k has property $\mathbf{u}_3^1$. This will yield Theorem 1.11. However we prove more; we prove (* * *):

1.9. G_k *has property* $\mathbf{u}_2^2$ *for all odd* $k>1$.

PROOF. Let $x_0=\langle i_0, p, q\rangle$, $x_1=\langle i_1, p', q'\rangle$ be 2 vertices of G_k. Without loss of generality we may assume $0\leqq i_1-i_0\leqq k+1$. Applying the isomorphism $\Phi_0(\langle i, p, q\rangle)=\langle i-i_0, p-p_0, q-q_0\rangle$ we may assume $x_0=\langle 0, 0, 0\rangle$, $x_1=\langle i_1, p_1, q_1\rangle$ with $0\leqq i_1\leqq k+1$, $0\leqq p_1\leqq k-1$. If $\frac{k-1}{2}<p_1<k$, then using the isomorphism $\Phi_1(\langle i, p, q\rangle)=\langle i, -p, q\rangle$, too, we can get $0\leqq p_1\leqq\frac{k-1}{2}$. So we may assume $x_0=\langle 0, 0, 0\rangle$, $x_1=\langle i_1, p_1, q_1\rangle$, $0\leqq i_1\leqq k+1$, $0\leqq p_1\leqq\frac{k-1}{2}$. We distinguish 4 cases:

Case a. $0\leqq i_1\leqq p_1\leqq\frac{k-1}{2}$, q_1 arbitrary;

Case b. $0\leqq p_1<k-p_1\leqq k+1$, q_1 arbitrary;

Case c. $0\leqq p_1<i_1\leqq k-p_1$, $(p_1)^*=q_1$;

Case d. $0\leqq p_1<i_1\leqq k-p_1$, $(p_1)^*=1-q_1$.

In all this cases we will find an F defined by one of the equalities (iii)—(v) in 1.5 such, that $x_0, x_1\notin F\cup F_1$. (F_1 is the adiungant of F.) 1.6 (i) and (iii) gives $F, F_1\in\mathscr{F}(2k^2, G_k)$ and 1.6 (iv) gives $F\cap F_1=\varnothing$, so this will prove 1.9.

Case a. $F=F^k\left(0, \frac{k+1}{2}, i_1, 1\right)$. F is defined. $F\cup F_1\cap\{x_0, x_1\}=\varnothing$, because (see 1.6 (iv))

$$(F\cup F_1)\cap A_{2k+2}(0,0)=\{\langle i,0,0\rangle : 1\leqq i\leqq k \text{ or } k+2\leqq i\leqq 2k+1\}$$

and

$$(F\cup F_1)\cap A_{2k+2}(p_1,q_1)=$$

$$=\{\langle i,p_1,q_1\rangle : i_1+1\leqq i\leqq i_1+k \text{ or } i_1+k+2\leqq i\leqq i_1+2k+1\}.$$

$\left(\text{Note that } i_1\leqq p_1\leqq\frac{k-1}{2}.\right)$

Case b. Now $F=F^k\left(0,\frac{k+1}{2},k-i_1+1,-1\right)$. F is defined, since $0\leqq k-i_1+1\leqq p_1\leqq\frac{k-1}{2}$. $(F\cup F_1)\cap\{x_0,x_1\}=\emptyset$, since we have

$$(F\cup F_1)\cap A_{2k+2}(0,0)=\{\langle i,0,0\rangle : 1\leqq i\leqq k, \text{ or } k+2\leqq i\leqq 2k+1\}$$

and

$$(F\cap F_1)\cap A_{2k+2}(p_1,q_1)=$$

$$=\{\langle i,p_1,q_1\rangle : -k+i_1\leqq i\leqq i_1-1 \text{ or } i_1+1\leqq i\leqq i_1+k\}.$$

$\left(\text{We used 1.6 (iv) and } 0\leqq k-i_1+1\leqq p_1\leqq\frac{k-1}{2}.\right)$

Case c. Put $v=i_1-p_1-1$ and $F=F^{1-2\eta(v)}\left(-\left[\frac{v}{2}\right];\left[\frac{v}{2}\right]^*,\frac{v+1}{2}\right)$. F is always defined. Let $\left[\frac{v}{2}\right]^*=q$. Then $\left(q+\left[\frac{v}{2}\right]\right)^*=0$, and

$$F\cap A_{2k+2}(0,0)=F_{-\left[\frac{v}{2}\right]+\left[\frac{v}{2}\right]}\left(\left(\left(q+\left[\frac{v}{2}\right]\right)^*,\frac{v+1}{2},2\left[\frac{v}{2}\right]+v-2\left[\frac{v}{2}\right]\right)\cap A_{2k+2}(0,0)=\right.$$

$$=F_0\left(0,\frac{v+1}{2},v\right)\cap A_{2k+2}(0,0)=\{\langle i,0,0\rangle : 1\leqq i\leqq v\}.$$

(We used $0\leqq v\leqq k-1$.) 1.6 (iv) gives $F_1\cap A_{2k+2}(0,0)=\{\langle i,0,0\rangle : v+2\leqq i\leqq 2k+1\}$ and this proves $x_0\notin F\cup F_1$. On the other hand

$$F\cap A_{2k+2}(p_1,q_1)=$$

$$=F_{-\left[\frac{v}{2}\right]+\left[\frac{v}{2}\right]+p_1}\left(\left(q+\left[\frac{v}{2}\right]+p_1\right)^*,\frac{v+1}{2},2\left[\frac{v}{2}\right]+2p_1+v-2\left[\frac{v}{2}\right]\right)\cap A_k(p_1,q_1)=$$

$$=F_{p_1}\left(q_1,\frac{v+1}{2},v+2p_1\right)\cap A_{2k+2}(p_1,q_1)=$$

$$=\{\langle i,p_1,q_1\rangle : -p_1+1\leqq i\leqq v+p_1=i_1-1\}.$$

Here we used $\left(q+\left[\frac{v}{2}\right]+p_1\right)^*=\left(\left[\frac{v}{2}\right]^*+\left[\frac{v}{2}\right]+p_1\right)^*=\left(2\left[\frac{v}{2}\right]^*+p_1^*\right)^*=p_1^*=q_1$, and this holds, since $0\leqq\left[\frac{v}{2}\right]\leqq 2\left[\frac{v}{2}\right]+p_1\leqq v+p_1=i_1-1\leqq k-1$. 1.6 (iv) gives now $F_1\cap A_{2k+2}(p_1,q_1)=\{\langle i,p_1,q_1\rangle : i_1+1\leqq i\leqq 2k-p_1+1\}$ proving $x_1\notin F\cup F_1$.

Case d. Put $v=k-i_1-p_1$, $\left[\frac{v}{2}\right]^*=q$ and $F=F^{1-2\eta(v)}\left(-\left[\frac{v}{2}\right], q, -\frac{v+1}{2}\right)$. Since $0\leqq v\leqq k-1$ so we have

$$F\cap A_{2k+2}(0,0)=F_{-\left[\frac{v}{2}\right]+\left[\frac{v}{2}\right]}\left(\left[\frac{v}{2}\right]^*, -\frac{v+1}{2}, v\right)\cap A_{2k+2}(0,0)=$$

$$=\{\langle i,0,0\rangle : -1\geqq i\geqq -v\}.$$

So 1.6 (iv) gives $F_1\cap A_{2k+2}(0,0)=\{\langle i,0,0\rangle\colon 1\leqq i\leqq 2k-v\}$ and $x_0\notin F\cup F_1$. Using $\left[\frac{v}{2}\right]+p_1\leqq k-1$ we have

$$F\cap A_{2k+2}(p_1,q_1)=F_{p_1}\left(\left(q+\left[\frac{v}{2}\right]+p_1\right)^*, -\frac{v+1}{2}, 2p_1+v\right)\cap A_{2k+2}(p_1,q_1).$$

Now we have $\left(q+\left[\frac{v}{2}\right]+p_1\right)^*=\left(q^*+\left[\frac{v}{2}\right]^*+p_1^*\right)^*=2\left[\frac{v}{2}\right]^*+p_1^*=p_1^*=1-q_1$, since $0\leqq\left[\frac{v}{2}\right]\leqq 2\left[\frac{v}{2}\right]+p_1\leqq v+p_1=k-i_1\leqq k-1$. Thus we get

$$F\cap A_{2k+2}(p_1,q_1)=F_{p_1}\left(1-q_1, -\frac{v+1}{2}, v+2p_1\right)\cap A_{2k+2}(p_1,q_1)=$$

$$=I\left(p_1,q_1,-\frac{v+1}{2},2k-v-2p_1\right)=$$

$$=\{\langle i,p_1,q_1\rangle : -k+p_1\leqq i\leqq i_1-1\}.$$

So —according to 1.6 (iv) — we have $F_1\cap A_{2k+2}(p_1,q_1)=\{\langle i,p_1,q_1\rangle\colon i_1+1\leqq i\leqq k+p_1\}$, and $x_1\notin F\cup F_1$. This proves Case d., too. So we proved $(*)$—$(***)$. In view of 1.8 (iii) this gives 1.10.

1.10. *$\mathscr{B}_k$ has property* $\mathbf{u}_2^2$.

1.8 (ii) and 1.10 imply the first part of the

1.11. THEOREM. (i) $(2k^2+2k, k^2, 3)\nrightarrow k^2+1$ *for all odd* $k\geqq 1$.

(ii) $(2n+10\lceil\sqrt{n}\rceil+8, n, 3)\nrightarrow n+1$ *for all integer* $n\geqq 1$.

PROOF. (i) means the existence of a graph G with $2k^2+2k$ vertices having $\alpha(G)=k^2$ and property $\mathbf{u}_3^1$. But $\mathscr{B}_k$ is such a graph for $k>1$ and K_4 for $k=1$.

To prove (ii) let $n\geqq 1$ be any integer, $k=k(n)$ the unique odd number satisfying $k^2\leqq n<(k+2)^2$ and put $l(n)=n-k^2(n)$. Then $l(n)\leqq(k+2)^2-k^2=4(k+1)\leqq$ $\leqq 4\lceil\sqrt{n}\rceil+4$. Let X_i be a K_4 for all $2\leqq i\leqq l(n)$, X_1 be a K_s with $s=10\lceil\sqrt{n}\rceil+12-$ $-2k-2l(n)\geqq 4$, $X_0=\mathscr{B}_k$ if $k>1$, and $X_0=K_4$ otherwise. ($X_i\cap X_j=\varnothing$ if $i\neq j$.) Let the graph G^n be $\bigcup_{i=0}^{l(n)} X_i$.

Clearly $\alpha(G^n)=\sum_{i=0}^{l(n)}\alpha(X_i)=k^2+l(n)=n$, $|V(G^n)|=|V(\mathscr{B}_k)|+s+4l(n)-4=2k^2+$ $+2k+10\lceil\sqrt{n}\rceil+12-2k-2l(n)+4l(n)-4=2(k^2+l(n))+10\lceil\sqrt{n}\rceil+8=2n+10\lceil\sqrt{n}\rceil+8$. All the X_i-s have property $\mathbf{u}_3^1$, so G^n must have it, too (see 1.8 (iii)). This proves the Theorem.

REMARKS. 1. (i) gives an affirmative answer to a special case of the conjecture stated in [6]. The general case is settled in Theorem 3.1.

2. If we put $K_5=X_i$ for $2\leqq i\leqq l(n)$ and $K_s=X_1$ with $s=14[\sqrt{n}]+17-2k-3l(n)\geqq 5$ then $H^n=\bigcup_0^{l(n)} X_i$ has property $\mathbf{u}_2^2$ (since $\mathscr{B}_k$, K_5 and $K_s\supset K_5$ have it.) This yields

1.12. *For all integer values of n there exists a graph H^n of $2n+14[\sqrt{n}]+12$ vertices such that $\alpha(H)=n$ and H^n has property $\mathbf{u}_2^2$.*

2. Consequences for line-critical graphs

2.1. DEFINITIONS (i) If $E\subset E(G)$ and $\alpha(G-E)>\alpha(G)$ then we call E *a critical set in* G.

(ii) $e\in E(G)$ *is critical in* G iff $\{e\}$ is a critical set in G.

(iii) $\delta(G)=|V(G)|-2\alpha(G)$ for any graph G.

2.2. In this chapter we investigate the $2k^2$-critical spanning subgraphs Γ_k of G_k. Of course our main interest lies in the special type of graphs $\Gamma_k=\Gamma_k^0\cup\Gamma_k^1$, where Γ_k^q is a spanning k^2-critical subgraph of $\mathscr{B}^k(q)$. In 2.3 we prove that the edges of $\mathscr{B}_p^k(q)$ are critical in G_k, and that Γ_k must contain almost all edges of G_k connecting different $A_{2k+2}(p)$ sets. The main consequences of this facts are summarized in Theorem 2.6. In 2.7—2.11 we give a construction of another type that gives a result, analogous to Theorem 2.6 (see Theorem 2.8). That $k\geqq 3$ is odd is assumed throughout 2.3—2.6.

2.3. (i) *For all $i\in C(2k+2)$, $p\in C(k)$, $q\in C(2)$, the edge $e=\langle i,p,q\rangle\langle i+k,p,q+1\rangle$ is critical in G_k.*

(ii) *Let i, p, q be as in* (i), *and $1\leqq t_0\leqq k$, $e_0=\langle i,p,q\rangle\langle i+k+1,p-1,q\rangle$ $e_1=\langle i+2t_0,p,q\rangle\langle i+2t_0+1+k,p-1,q\rangle$. Then the set $\{e_0,e_1\}$ is a critical set in G_k.*

PROOF: (i) We may assume $e=\langle 0,0,0\rangle\langle k,0,1\rangle$. Put

$$T=F^k\left(0,\frac{k-1}{2},0,1\right)\cup\{\langle k,0,1\rangle\}.$$

$F^k\left(0,\frac{k-1}{2},0,1\right)$ is independent and contains neither $\langle 2k+1,1;1\rangle$ nor $\langle 2k+1,-1,1\rangle$, $\langle 2k,0,0\rangle$, $\langle 2k+1,0,0\rangle$ but it contains $\langle 2k+2,0,0\rangle=\langle 0,0,0\rangle$. Since these are the 5 vertices connected to $\langle k,0,1\rangle$ in G_k, so we have $E(G_k,T,T)=\{e\}$, i.e. $T\in\mathscr{F}(2k^2+1,G_k-e)$ proving (i).

(ii) Again we assume $e=\langle 0,0,0\rangle\langle k+1,-1,0\rangle$ and $e_1=\langle 2t_0,0,0\rangle\langle 2t_0+k+1,-1,0\rangle$, $1\leqq t_0\leqq\frac{k+1}{2}$. We put

$$T=\Big(\bigcup_{-(t_0-1)\leqq t\leqq -1}F_t((-t)^*,t_0,2(t+t_0)-1)\cup\Big(\bigcup_{0\leqq t\leqq k-(t_0+1)}F_t((t)^*,t_0,2(t+t_0)+1)\Big)\cup$$
$$\cup I(k-t_0,(k-t_0)^*,t_0,2k+1)$$

Clearly $|T|=2k^2+1$. Our claim is $E(G_k, T, T)=\{e_0, e_1\}$. Once we proved this, we have $\alpha(G_k-\{e_0, e_1\})\geqq|T|>\alpha(G_k)$, proving (ii).

It is clear that

$$\bigcup_{-(t_0-1)\leqq t\leqq -1} F_t((-t)^*, t_0, 2(t+t_0)-1) = \bigcup_{-(t_0-1)\leqq t\leqq -1} A_{2k+2}(p)\cap F^1((1-t_0), (t_0-1)^*, t_0)$$

and

$$\bigcup_{0\leqq t\leqq k-t_0-1} F_t((t)^*, t_0, 2(t+t_0)+1) = \Big(\bigcup_{0\leqq p\leqq k-(t_0+1)} A_{2k+2}(p)\Big)\cap F^1(-t_0, (t_0)^*, t_0),$$

so these sets are independent, and so is $I(k-t_0, (k-t_0)^*, t_0, 2k+1)$. So all we have to prove, is that $I(k-t_0, (k-t_0)^*, t_0, 2k+1)$ is not joined to $T\cap A_{2k+2}(k-t_0+1, (k-t_0)^*)$ and to $T\cap A_{2k+2}(k-t_0-1, (k-t_0)^*)$; and, that $E(G_k, T\cap A_{2k+2}(0), T\cap A_{2k+2}(-1))=\{e_0, e_1\}$. The second one is trivial from 1.6 (i); and 1.6 (ii) with

$$T\cap A_{2k+2}(k-t_0+1, (k-t_0)^*) = T\cap A_{2k+2}(-(t_0-1), (t_0-1)^*) =$$
$$= F_{-(t_0-1)}((t_0-1)^*, t_0, 1)\cap A_{2k+2}(-(t_0-1), (t_0-1)^*) =$$
$$= I(-(t_0-1), (t_0-1)^*, t_0, 1) = I(k-t_0+1, (k-t_0)^*, t_0, 1)$$

and with

$$T\cap A_{2k+2}(k-(t_0+1), (k-t_0)^*) =$$
$$= F_{k-t_0-1}((k-t_0-1)^*, t_0, 2k-1)\cap A_{2k+2}(k-t_0-1, 1-(k-t_0-1)^*) =$$
$$= I(k-(t_0+1), (k-t_0)^*, t_0, 1)$$

gives the first part of the proposition. This proves our claim and (ii).

The following observation is trivial.

2.4. *If $E\subset E(\mathscr{B}^k(q))$ for some $q\in C(2)$ and E is critical in G_k, then E is also critical in $\mathscr{B}^k(q)$.*

This is obvious: E is critical in G_k, and this means, that an F in $\mathscr{F}(2k^2+1, G_k-E)$ exists with $|F\cap\mathscr{B}^k(1-q)|\leqq k^2$ since $\mathscr{B}^k(1-q)$ spans the same subgraph in G_k as in G_k-E, (we used $E\cap E(\mathscr{B}^k(1-q))=\varnothing$ and $\alpha(\mathscr{B}^k(1-q))=$ $=\alpha(\mathscr{B}_k)=k^2$ (see 1.3).) So $|F\cap\mathscr{B}^k(q)|\geqq k^2+1$ proving $\alpha(\mathscr{B}^k(q)-E)\geqq k^2+1$. We note, that given any 2 edges e_0, e_1 connecting $\mathscr{B}^k(q_0)\cap A_{2k+2}(p, q)$ and $\mathscr{B}^k(q_0)\cap$ $\cap A_{2k+2}(p-1, q)$ they can always be written in the form of 2.3 (ii). So 2.4 and 2.3 gives

2.5. (i) *Any edge of $\mathscr{B}^k_p(q)$ is critical in $\mathscr{B}^k(q)$.* ($p\in C(k)$, $q\in C(2)$.)

(ii) *If e_0, e_1 are 2 different edges connecting $A_{2k+2}(p, q)$ and $A_{2k+2}(p-1, q)$ in $\mathscr{B}^k(q_0)$ for some q, $q_0\in C(2)$, $p\in C(k)$, then $\{e_0, e_1\}$ is a critical set in $\mathscr{B}^k(q_0)$.*

Let now Γ_k be any k^2-critical spanning subgraph of $\mathscr{B}^k(0)$. Since by 2.5, $E(\mathscr{B}^k(0))-E(\Gamma_k)$ contains no edge of $\mathscr{B}^k_p(0)$ and at most one edge connecting $A_{2k+2}(p, q)$ to $A_{2k+2}(p-1, q)$, ($p\in C(k)$, $q\in C(2)$), we have

$$|E(\mathscr{B}^k(0))-E(\Gamma_k)|\leqq 2k \quad \text{i.e.} \quad |E(\Gamma_k)|\geqq 4(k+1)k-2k = 4k^2+2k.$$

Finally we note that $\varrho(\Gamma_k, 3)\geqq 3$ for all vertices of Γ_k: $\mathscr{F}(k^2, \Gamma_k-(\{x\}\cup E(\Gamma_k, x)))=\varnothing$ is trivial and $\mathscr{B}^k(0)$ has property $\mathbf{u}^2_2$, and so property $\mathbf{u}^1_3$ according to 1.10. However this means $|\{x\}\cup E(\Gamma_k, x)|\geqq 4$ and $\varrho(\Gamma_k, x)\geqq 3$, as stated.

So we get the following theorem putting $\delta=2k$ (k odd!) and $H_\delta=\Gamma_k$ if $\delta\geqq 6$, $H_2=K_4$:

2.6. THEOREM. *For any positive* $\delta \equiv 2 \bmod 4$ *there exists a graph* H_δ *with the following properties:*

(i) H_δ *is connected and is* $\delta^2/4$*-critical;*

(ii) $|V(H_\delta)| = \frac{\delta^2}{2} + \delta$, i.e. $\delta(H_\delta) = \delta$;

(iii) $|E(H_\delta)| \geqq \delta^2 + \delta = |V(H_\delta)| + \frac{\delta^2}{2}$;

(iv) H_δ *has no vertex of valency* $\leqq 2$*; and moreover*

(v) H_δ *has property* $\mathbf{u}_2^2$ *and so property* $\mathbf{u}_3^1$.

REMARKS: 1. The main interest of this theorem are (i), (ii), and (iv) together (see introduction).

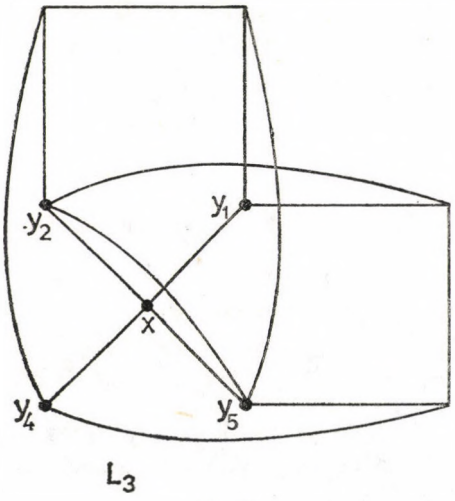


Fig. 1.

Let $\delta \geqq 3$, let L_δ be the graph we get by substituting the vertex x of the graph on Fig. 1 with a $K_{\delta-2}$ and joining all the vertices of this $K_{\delta-2}$ to all y_i $1 \leqq i \leqq 4$. Then L_δ is 3-critical, $\delta(L_\delta) = \delta$ and

$$|E(L_\delta)| - |V(L_\delta)| = \frac{\delta^2 + \delta}{2}.$$

So $f(\delta) \geqq \binom{\delta+1}{2}$, showing that (iii) gives no good information on $f(\delta)$. We don't think $f(\delta) \geqq \binom{\delta+1}{2}$ to be a sharp result, but as for now, do not know any better.

2. We did not derive Γ_k explicitely from $\mathscr{B}_k$. Our next aim is to sketch a totally explicit construction of a H_δ satisfying (i)—(iv) of theorem 2 for any $\delta \equiv 0 \bmod 4$.:

2.7. Let $\delta \equiv 0 \bmod 4$, $\delta > 0$ and put $\delta = 2k$. Then $k \geqq 2$ is even. The graph $\mathscr{D}_k$ we are going to define is similar to $\mathscr{B}_k$. (Its definition is motivated by the following graph $\mathscr{C}_k$, which is isomorphic to $\mathscr{B}_k$ for odd $k \geqq 3$:

$$V(\mathscr{C}_k) = \{\langle i, p\rangle : i \in C(2k+2),\ p \in C(k)\},$$

$$E(\mathscr{C}_k) = \{\langle i, p\rangle\langle i+1, p\rangle : i \in C(2k+2),\ p \in C(k)\} \cup$$

$$\{\langle i, p\rangle\langle i, p+1\rangle : i \in C(2k+2),\ p \in C(k),\ p \neq k\} \cup$$

$$\{\langle i, k\rangle\langle i+k+1, 1\rangle : i \in C(2k+2)\}.$$

Let for any $p \in C(k+1)$, $q \in C(2)$,

$$U_k(p) = \{\langle i, p\rangle : i \in C\{2k)\}.$$

We define $\mathscr{D}_k$ as follows:

$$V(\mathscr{D}_k) = \bigcup_{p \in C(k+1)} U_k(p),$$

$$E(\mathscr{D}_k) = \{\langle 2j-1, 0\rangle\langle 2j, 0\rangle : 1 \leqq j \leqq k\} \cup \{\langle 2j, 1\rangle\langle 2j+1, 1\rangle : 1 \leqq j \leqq k\} \cup$$
$$\{\langle i, p\rangle\langle i+1, p\rangle : i \in C(2k),\ 2 \leqq p \leqq k\} \cup$$
$$\{\langle i, 0\rangle\langle i+k, 1\rangle : i \in C(2k)\} \cup \{\langle i, p\rangle\langle i, p+1\rangle : i \in C(2k),\ 1 \leqq p \leqq k\}.$$

We put $U_k(p, q) = \left\{\langle 2j-q, p\rangle : 1 \leqq j \leqq \frac{k}{2}\right\}$ and $V_k(p, q) = U_k(p, q) \cup U_k(p+1, q)$. Clearly $U_k(p) = U_k(p, 0) \cup U_k(p, 1)$. $U_k(p)$ spans an alternating $U_k(p, 0)\, U_k(p, 1)$ circuit of length $2k$ for $2 \leqq p \leqq k$ in $\mathscr{D}_k$, $U_k(0)$ and $U_k(1)$ spans a matching between $U_k(0, 0)$ and $U_k(0, 1)$ resp. $U_k(1, 0)$ and $U_k(1, 1)$ in $\mathscr{D}_k$, and $V_k(p, q)$ spans a matching between $U_k(p, q)$ and $U_k(p+1, q)$. Finally we mention, that $U_k(0) \cup U_k(1)$ spans a circuit of length $4k$ in $\mathscr{D}_k$. We claim that $\mathscr{D}_k$ is k^2-critical. Suppose we have proved this. Since $\delta = 2k$, this proves the following theorem with $H_\delta = \mathscr{D}_k$.

2.8. THEOREM. *For any positive integer δ with $\delta \equiv 0 \bmod 4$ there exists a graph H_δ which satisfies conditions* (i)—(iv) *of Theorem* 2.6, *with equality in* (iii).

$\mathscr{D}_k$ is connected, $|V(\mathscr{D}_k)| = 2k^2 + 2k = \frac{\delta^2}{2} + \delta$, $|E(\mathscr{D}_k)| = 4k^2 + 2k = \delta^2 + \delta$, and for all x, $\varrho(\mathscr{D}_k, x) \geqq 3$. So nothing else, but our claim is left to be proved.

2.9. First we prove $\alpha(\mathscr{D}_k) \leqq k^2$. We use the same idea as in 1.3. Let $F \in \mathscr{F}(\mathscr{D}_k)$.

a) $|F \cap U_k(p)| \geqq k$ for some $2 \leqq p \leqq k$. Then $|F \cap U_k(p)| = k$ and $F \cap U_k(p) = U_k(p, q)$ for some $q \in C(2)$, say, for $q = 0$, since $U_k(p)$ spans a circuit of length $2k$ in $\mathscr{D}_k$. Clearly

$$F = \big(F \cap U_k(p)\big) \cup \big(F \cap \big(U_k(p-1, 0)\big) \cup \big(F \cap U_k(p+1, 0)\big) \cup$$
$$\cup \Big(\bigcup_{1 \leqq t \leqq \frac{k}{2}-1} \big(F \cap V_k(p+2t, 0)\big)\Big) \cup \Big(\bigcup_{1 \leqq t \leqq \frac{k}{2}} \big(F \cap V_k(p+2t-1, 1)\big)\Big).$$

But $V_k(p+j, q)$ spans a 1-factor in $\mathscr{D}_k$, so $|F \cap V_k(p+j, q)| \leqq \frac{|V_k(p+j, q)|}{2} \leqq k$. $U_k(p, 0) \subset F$ implies $F \cap U_k(p-1, 0) = F \cap U_k(p+1, 0) = \varnothing$. All these together yield

$$|F| \leqq k + \left(\frac{k}{2} - 1\right) k + \frac{k}{2} k = k^2.$$

b) $|F| \cap U_k(p)| \leqq k-1$ for all $2 \leqq p \leqq k$.
If $\big|F \cap \big(U_k(0) \cup U_k(1)\big)\big| \leqq 2k-1$, then we have

$$|F| = \sum_{2 \leqq p \leqq k} |F \cap U_k(p)| + \big|F \cap \big(U_k(0) \cup U_k(1)\big)\big| \leqq (k-1)^2 + 2k - 1 = k^2.$$

So we only have to consider the case, when $\big|F \cap \big(U_k(0) \cup U_k(1)\big) \geqq 2k$. In this case $\big|F \cap \big(U_k(0) \cup U_k(1)\big)\big| = 2k$ and $F \cap \big(U_k(0) \cup U_k(1)\big) = U_k(0, q) \cup U_k(1, 1-q)$ must hold with some $q \in C(2)$, say $q = 0$, because $U_k(0) \cup U_k(1)$ spans a circuit of length

$4k$. So we have $F\cap U_k(0,1)=F\cap U_k(1,0)=F\cap U_k(2,1)=F\cap U_k(k,0)=\varnothing$ and

$$|F| = |U_k(0,0)|+|U_k(1,1)|+\sum_{1\leqq t\leqq \frac{k}{2}-1}|F\cap V_k(2t,0)|+\sum_{1\leqq t\leqq \frac{k}{2}-1}|F\cap V_k(2t+1,1)|\leqq$$

$$\leqq 2k+2\left(\frac{k}{2}-1\right)k = k^2.$$

2.10. So only $\mathscr{F}(k^2+1, \mathscr{D}_k-e)\neq 0$ for all $e\in E(\mathscr{D}_k)$ is left to be proved. To simplify the proof of this, let Φ be a function on $V(\mathscr{D}_k)$ defined by $\Phi(\langle i,p\rangle)=\langle i+1, 1-p\rangle$. Φ is clearly an isomorphism on $\mathscr{D}_k$, and so is Φ^j. But $\Phi^j(\langle i,p\rangle)=\langle i+j, 1-p\rangle$ if j is odd and $=\langle i+j,p\rangle$ if j is even. So to prove $\mathscr{F}(k^2+1,\mathscr{D}_k-e)\neq\varnothing$ for all e of the type $\langle j,p\rangle\langle j+1,p\rangle$, we may restrict ourselves to the type $e=\langle 0,p\rangle\langle 1,p\rangle$. ($1-p$ is ranging over $C(k+1)$ if p is so.) So let $e=\langle 0,p_0\rangle\langle 1,p_0\rangle\in E(\mathscr{D}_k)$ i.e. $1\leqq p_0\leqq k$.

a. p_0 is odd. We define F by

$$F\cap U_k(0)=\left\{\langle 2j-1,0\rangle : 1\leqq j\leqq\frac{k}{2}\right\}\cup\left\{\langle 2j,0\rangle : \frac{k}{2}+1\leqq j\leqq k\right\};$$

$$F\cap U_k(p)=\left\{\langle 2j-1,p\rangle : 1\leqq j\leqq\frac{k}{2}\right\}\cup\left\{\langle 2j,p\rangle : \frac{k}{2}+1\leqq j\leqq k-1\right\},$$

$$\text{if}\quad 1\leqq p\leqq k-1,\ p \text{ odd};$$

$$F\cap U_k(p)=\left\{\langle 2j,p\rangle : 1\leqq j\leqq\frac{k}{2}\right\}\cup\left\{\langle 2j+1,p\rangle : \frac{k}{2}+1\leqq j\leqq k-1\right\},$$

$$\text{if}\quad 1\leqq p\leqq k,\ p \text{ even}.$$

One can easily check that $F\in\mathscr{F}(k^2,\mathscr{D}_k)$. If p_0 is odd, $1\leqq p_0\leqq k$, then the vertices connected to $\langle 2k,p_0\rangle=\langle 0,p_0\rangle$ in $\mathscr{D}_k$ are contained in the set

$$T=\{\langle k,0\rangle,\langle 1,p_0\rangle\}\cup\{\langle 0,p\rangle : 1\leqq p\leqq k\}\cup$$

$$\{\langle 2k-1,p\rangle : 1\leqq p\leqq k-1,\ p \text{ odd}\}.$$

$T\cap F=\{\langle 1,p_0\rangle\}$ so $E(\mathscr{D}_k, F\cup\{\langle 0,p_0\rangle\}, F\cup\{\langle 0,p_0\rangle\})=\{e_0\}$ proving $F\cup\{\langle 0,p_0\rangle\}\in\mathscr{F}(k^2+1,\mathscr{D}_k-e_0)$.

b. p_0 is even. Then $2\leqq p_0\leqq k$. We define F' by

$$F'\cap U_k(0)=\left\{\langle 2j,0\rangle : 1\leqq j\leqq\frac{k}{2}\right\}\cup\left\{\langle 2j-1,0\rangle : \frac{k}{2}+1\leqq j\leqq k\right\};$$

$$F'\cap U_k(p)=\left\{\langle 2j,p\rangle : 1\leqq j\leqq\frac{k}{2}-1\right\}\cup\left\{\langle 2j-1,p\rangle : \frac{k}{2}+1\leqq j\leqq k\right\}$$

$$\text{if}\quad 1\leqq p\leqq k-1,\ p \text{ odd};$$

$$F'\cap U_k(p)=\left\{\langle 2j-1,p\rangle : 1\leqq j\leqq\frac{k}{2}\right\}\cup\left\{\langle 2j,p\rangle : \frac{k}{2}+1\leqq j\leqq k-1\right\}$$

$$\text{if}\quad 1\leqq p\leqq k \text{ even}.$$

One can easily check again that $F' \in \mathscr{F}(k^2, \mathscr{D}_k)$. If p_0 is even, $2 \leqq p_0 \leqq k$, then the vertices connected to $\langle 2k, p_0\rangle = \langle 0, p_0\rangle$ in $\mathscr{D}_k$ are contained in the set

$$T' = \{\langle 1, p_0\rangle\} \cup \{\langle 2k-1, p\rangle : 2 \leqq p \leqq k,\ p \text{ even}\} \cup \{\langle 0, p\rangle : 0 \leqq p \leqq k\}.$$

Since $T' \cap F' = \{\langle 1, p_0\rangle\}$, so again we have $F' \cup \{\langle 0, p_0\rangle\} \in \mathscr{F}(k^2+1, \mathscr{D}_k - e_0)$.

2.11. To complete our proof we have to prove $\mathscr{F}(k^2+1, \mathscr{D}_k - e) \neq \varnothing$ for any $e \in E(V_k(p, q))$ if $p \in C(k+1)$, $q \in C(2)$. We define $F_p(q)$ by

$$F_p(q) = \bigcup_{0 \leqq t \leqq k-1} U_k\left(p+t+1,\ q+t-2\left[\frac{t}{2}\right]\right) =$$

$$= U_k(p+1, q) \cup U_k(p+2, 1-q) \cup U_k(p+3, q) \cup \ldots \cup U_k(p+k, 1-q).$$

$F_p(q) \in \mathscr{F}(k^2, \mathscr{D}_k)$ is immediate. If $e \in E(V_k(p, q))$, then e has the form $e = \langle i, p\rangle\langle j, p+1\rangle$. $\langle i, p\rangle$ is only connected to some vertices of $U_k(p, 1-q)$, $U_k(p-1, q) = U_k(p+k, q)$, and to this $\langle j, p+1\rangle$. ($j=i$ if $p \neq 0$, $j=i+k$ if $p=0$.) But $F_p(q) \cap \cap (U_k(p) \cup U_k(p+k, q)) = \varnothing$, yielding $F_p(q) \cup \{\langle i, p\rangle\} \in \mathscr{F}(k^2+1, \mathscr{D}_k - e)$. This completes the proof of Theorem 2.8.

3. The general case of the arrow symbol

3.1. THEOREM. (i) $(4lk^2+4lk, 2k^2-t+1, 2lt+1) \nrightarrow 2k^2+1$ *for any* t, $l \geqq 1$, $k \geqq 1$ *odd.*

(ii) $(m(n, k, l), n, k) \nrightarrow n+l$ *for every* $k>1$, $k \equiv 1 \bmod 2l$, $l \geqq 1$, *where* $m(n, k, l) = \dfrac{k-1}{l}(n+l-1) + 4(k-1)\left[\sqrt{2(n+l-1)}\right] + \dfrac{k-1}{l}\left[\sqrt{2(n+l-1)}\right]$. *So in this case we have* $f(n, k, l) < m(n, k, l)$.

COROLLARIES: (iii) $(2kn+10k\left[\sqrt{2n}\right], n, 2k+1 \nrightarrow n+1)$ for $n \geqq 1$, $k \geqq 1$.

(iv) $\left(2(n+l-1)+8l+2\right)\left[\sqrt{2(n+l-1)}\right], n, 2l+1) \nrightarrow n+l$ for $n, l \geqq 1$.

REMARKS. 1. It was proved in [7] Theorem 6, that $(2n+2l+1, n, 2l+1) \nrightarrow n+l$ for $n > n_0(l)$. So we have $3 \leqq f(n, 2l+1, l) - 2(n+l-1) \leqq (8l+2)\left[\sqrt{2(n+l-1)}\right]$.

(iv) is the first proof of $f(n, 2l+1, l) - 2n = o(n)$ (l fixed). We do not know, whether $\lim\limits_{\substack{n \to \infty \\ l \text{ fixed}}} f(n, 2l+1, l) = \infty$ holds. (This would be a consequence of the existence of a finite $\alpha_0(\delta)$ for all δ in the problem of the Introduction. See [7] 5.5.)

(iv) was conjectured in [6].

2. (iii) gives the best known upper bound for $f(n, 2k+1, 1)$ if $k>1$. (For $k=1$ see Theorem 1.11.) The only general lower bound on $f(n, 2k+1, 1)$ can be found in [4]: $f(n, 2k+1, 1) \geqq 2n+2k-1+\Delta(k)$ with $\Delta(k) = \max\left\{q : \binom{q}{2} \leqq k+1\right\}$.

PROOF: (ii)$\Rightarrow$(i) can be proved in the same way as in Theorem 1.

To prove (i) we generalize the construction of G_k. Let k be odd, $l \geqq 1$, $k \geqq 3$. We define a graph $G_{k,l}$ as follows:

$$V(G_{k,l}) = \bigcup_{p \in C(k)} A_{2l(k+1)}(p)$$

$$E(G_{k,l}) = \{\langle i, p, q\rangle\langle j, p, q\rangle : i+2k+2 \leqq j \leqq i+2(l-1)(k+1)\} \cup$$
$$\{\langle i, p, q\rangle\langle j, p, 1-q\rangle : i+k \leqq j \leqq i+2l+(2l-1)k\} \cup$$
$$\{\langle i, p, q\rangle\langle j, p+1, q\rangle : i+k+1 \leqq j \leqq i+(2l-1)(k+1)\}.$$

Clearly $G_{k,1}=G_k$.

$G_{k,l}$ will prove (i).

First we show that

3.2. $\alpha(G_{k,l}) \leqq 2k^2$.

We claim that

(i) $$\mathscr{F}(2k+1, A_{2l(k+1)}(p)) = \bigcup_{q=0}^{1} \mathscr{F}(2k+1, A_{2l(k+1)}(p, q))$$

and

(ii) $$\mathscr{F}(2k+3, A_{2l(k+1)}(p, q) \cup A_{2l(k+1)}(p+1, q)) =$$
$$\mathscr{F}(2k+3, A_{2l(k+1)}(p, q)) \cup \mathscr{F}(2k+3, A_{2l(k+1)}(p+1, q)) = \emptyset.$$

Suppose we proved this. Then we prove 3.2 as follows: Let $F \in \mathscr{F}(G_{k,l})$. If $|F \cap A_{2l(k+1)}(p)| \leqq 2k$ for all $p \in C(k)$, then $|F| \leqq 2k^2$. So suppose $|F \cap A_{2l(k+1)}(p)| \geqq 2k+1$ for some p. Then according to (i)

$$F \cap A_{2l(k+1)}(p) = F \cap A_{2l(k+1)}(p, q)$$

for some q. Because of obvious symmetries we may assume $p=0$, $q=0$. Put for all p, q $F \cap A_{2l(k+1)}(p, q) = F(p, q)$. We have

$$|F(0, 0)| \geqq 2k+1, \quad |F(0, 1)| = 0,$$

and — according to (ii) —

$$|F(1, 0)|, \ |F(-1, 0)| \leqq 2k+2-|F(0, 0)|,$$

$$|F(2j, 0) \cup F(2j+1, 0)| \leqq 2k+2 \quad \text{for} \quad j = 1, \ldots, \frac{k-3}{2},$$

$$|F(2j-1, 1) \cup F(2j, 1)| \leqq 2k+2 \quad \text{for} \quad j = 1, \ldots, \frac{k-1}{2}.$$

So we have

$$|F| = |F(0, 0)| + |F(0, 1)| + |F(-1, 0)| + |F(1, 0)| +$$
$$+ \sum_{j=1}^{\frac{k-3}{2}} |F(2j, 0) \cup F(2j+1, 0)| + \sum_{j=1}^{\frac{k-1}{2}} |F(2j-1, 1) \cup F(2j, 1)| \leqq$$
$$\leqq 2k+2+1+(k-2)(2k+2) = 2k^2-1 < 2k^2.$$

This proves 3.2.

So we only have to prove (i) and (ii). To do this we define the following bipartite graph $\mathscr{P}_t^v$:

$$V(\mathscr{P}_t^v) = \{\langle i, q\rangle : i \in C(t),\ q \in C(2)\},$$

$$E(\mathscr{P}_t^v) = \{\langle i, 0\rangle\langle j, 1\rangle : 0 \leqq i-j \leqq v;\ i, j \in C(t)\}.$$

We denote by $P_t(q)=\{\langle i, q\rangle : i \in C(t)\}$. Our claim is that the following lemma holds:

3.3. *Let* $F \in \mathscr{F}(\mathscr{P}_t^v)$ *and suppose* $F \cap P_t(q) \neq \varnothing$ *for* $q=0, 1$. *Then* $|F| \leqq t-v$.

The edges between $A_{2l(k+1)}(p, 0)$ and $A_{2l(k+1)}(p, 1)$ span a bipartite graph isomorphic to $\mathscr{P}_{2l(k+1)}^{2l+2(l-1)k}$, so 3.3 gives 3.2 (i). The edges between $A_{2l(k+1)}(p, q)$ and $A_{2l(k+1)}(p+1, q)$ span a bipartite graph isomorphic to $\mathscr{P}_{2l(k+1)}^{2(l-1)k}$. This proves 3.2 (ii). So 3.2 is proved if we prove 3.3. As a matter of fact 3.3 is the same as the Lemma in [6]. We give a short proof of it. (Note that this lemma is trivial for $v=1, 2$.) Let F be as in the lemma; we want to prove this lemma by induction on the number l of the elements in $F \cap P_t(0)=F_0$. $l \geqq 1$ and if $l=1$ the lemma is trivial. So suppose that $l \geqq 2$ and the lemma holds for $l-1$. We may suppose that

$$(*) \qquad F_1 = F \cap P_t(1) = P_t(1) - \{x : \exists y \in F_0 (xy \in E(\mathscr{P}_t^v))\}.$$

So $F_1 \neq \varnothing$ and $F_0 \neq \varnothing$ give $P_t(1) - F_1 \neq \varnothing$, too. So there exists some vertex x such that $x=\langle i, 1\rangle \in F$ and $x'=\langle i-1, 1\rangle \notin F_1$. $\langle i, 1\rangle \in F$ implies $\langle i+u, 0\rangle \notin F_0$ for $0 \leqq u \leqq v$. So if $y=\langle i-1, 0\rangle \notin F_0$, then $x'=\langle i-1, 1\rangle$ would not be connected to F_0 in $\mathscr{P}_t^v$ and then $x' \in F_0$ would follow by $(*)$. So we have $y \in F_0$, and the only edge connecting x' to F_0 in $\mathscr{P}_t^v$ is $x'y$. This means that

$$F' = (F_1 \cup \{x'\}) \cup (F_0 - \{y\}) \in \mathscr{F}(\mathscr{P}_t^v).$$

F' and F have the same cardinality, F' has verices both in $P_t(0)$ and $P_t(1)$, it has $l-1$ vertices in $P_t(0)$ so the induction hypothesis applies to it. So we get $|F| = |F'| \leqq t-v$.

3.4. Now, in what follows, we only sketch the proof of 3.1 (i), since it goes on the same way as the proof of 1.10 (i) in **1.** If in 1.4—1.5 we put $C(2l(k+1))$, $A_{2l(k+1)}(p, q)$, $A_{2l(k+1)}(p)$ instead of $C(2k+2)$ $A_{2k+2}(p, q)$ $A_{2k+2}(p)$ resp., everything remains valid. Now we may define the j'th adiungant F_j of any set F by $(F_{j-1})=F_j$ for $j=2, \ldots, 2l$ or otherwise

$$F_j = \{\langle j(k+1)+i, p, (q+j)^*\rangle : \langle i, p, q\rangle \in F\}.$$

Clearly $F_{2l+j}=F_j$, $F_0=F$. $F_1, \ldots, F_{2l}$ are pairwise disjoint. $F_j \in \mathscr{F}(2k^2, G_{k,l})$ and the analogue of (1.6) (i)—(iv) can be easily found and proved.

Let $|X| \geqq 2l$. Then we clearly have 2 elements $x_0=\langle i_0, p_0, q_0\rangle$ and $x_1=\langle i_1, p_1, q_1\rangle$ in X such that $0 \leqq i_0 - i_1 \leqq k+1$. Now the argument of 1.9 gives without any change an F with its $2l-1$ pairwise disjoint adiungants all contained in $\mathscr{F}(2k^2, G_{k,l})$. (This proves that $G_{k,l}$ has property U_{2l}^2.) So if $|X| \geqq 2l$ and $\alpha(G_{k,l}-X) \leqq 2k^2-t$, then $|(X-\{x_0, x_1\}) \cap F_i| \geqq t$ must hold for $i=1, \ldots, 2l$. The sets F_i are pairwise disjoint, and none of them contains x_0 or x_1, so $|X| \geqq 2lt+2$ must also hold. This completes the proof of 3.1 (i).

ADDED IN PROOF. LOVÁSZ recently proved that $\alpha_0(\delta) \leqq 2^{c\delta^2}$.

REFERENCES

[1] ANDRÁSFAI, B.: On critical graphs, in: *Theory of Graphs, International symposium held at Rome, 1966* (Dunod, Paris; Gordon and Breach, New York, 1967), 9—19.
[2] ERDŐS, P.: On a lemma of Hajnal—Folkmann, in: *Combinatorial theory and its applications, Coll. Math. Soc. J. Bolyai,* **4** (1970), 311—316.
[3] HAJNAL, A.: On k-saturated graphs, *Canad. Journ. of Math.* 1965.
[4] MILNER, E.—SAUER, N.: Generalizations of a lemma of Folkmann, in: *Annals of the New York Acad. of Sci. Vol. 175,* Art. 1 (1970), 295—307.
[5] PETRUSKA, G.—SZEMERÉDI, E.: On a combinatorial problem I, *Studia Sci. Math. Hung.* **7** (1972), 363—374.
[6] SURÁNYI, L.: On a problem of P. Erdős and A. Hajnal, in: *Comb. Theory and its applications, Coll. Math. Soc. J. Bolyai* 4 (1970), 1029—1041.
[7] SURÁNYI, L.: On linecritical graphs, *Coll. Math. Soc. J. Bolyai, 10, Infinite and Finite Sets, Keszthely, 1973,* 1411—1444.
[8] SZEMERÉDI, E.: On a problem of P. Erdős, in: *Comb: theory and its applications, Coll. Math. Soc. J. Bolyai 4* (1970), 1051—2.
[9] WESSEL, W.: Kantenkritische Graphen mit der Zusammenhangszahl *2, Manuscripta Math.* **2** (1970), 309—334.
[10] WESSEL, W.: On the problem of determining whether a given graph is edge-critical or not, in: *Comb. theory and its applications,* Coll. Math. Soc. J. Bolyai, **4** (1970), 1123—1139.
[11] WESSEL, W.: A first family of edge-critical wheels. *To appear.*

Mathematical Institute of the Hungarian Academy of Sciences,
H—1053 Budapest, Reáltanoda u. 13—15.

(Received May 2, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 413—418.

STEADY STATE HEAT FLOW IN A SHELL ENCLOSED BETWEEN TWO PROLATE SPHEROIDS

by

S. N. PANDEY

1.

In the study of the temperature distribution within various types of shells the expansions in terms of orthogonal polynomials have been widely used. It is well known that the study of the diffusion of heat, produced by any means in the solids, has several important applications (see [2], pp. 12—13) in Mathematical Physics and Engineering sciences.

The temperature distribution in a spherical shell, for example, is already known [6], p. 329. An analogous problem for the prolate spheroidal shell has been discussed by BHONSLE [1].

It is known that (see [5], p. 213) prolate spheroidal coordinates α, β, φ are related to the rectangular coordinates x, y, z by the relations:

$$\begin{aligned} x &= c \sinh\alpha \sin\beta \cos\varphi, \\ y &= c \sinh\alpha \sin\beta \sin\varphi, \\ z &= c \cosh\alpha \cos\beta, \end{aligned} \tag{1.1}$$

where $0 \leqq \alpha < \infty$, $0 \leqq \beta \leqq \pi$, $-\pi < \varphi < \pi$ and $c > 0$ is a scalar factor.

It may be mentioned here that Bhonsle (loc. cit.) in his study of this problem, assumed that

$$L_n \equiv \int_0^{\pi} f(\beta) P_n(\cos\beta) \sin\beta dB$$

is bounded and monotonically decreasing, where $f(\beta)$ is the temperature distribution function and $P_n(\cos\beta)$ is the nth Legendre polynomial. The proof, in fact, depends on the inequality

$$P_n(\cos\beta) \leqq 1,$$

which holds uniformly in the range $0 \leqq \beta \leqq \pi$. It is interesting to note that in the interior of the interval $(0, \pi)$, we may use more convenient order estimates (see [7], p. 167) for $P_n(\cos\beta)$, $0 < \beta < \pi$.

The object of this paper is to study the problem of temperature distribution within a prolate spheroidal shell from a different angle, making use of the convolution structure formula defined by GASPER [4]. Instead of imposing any condition on L_n, we start from the temperature distribution function $f(\beta)$, satisfying certain conditions, and discuss the solution of the problem in a straightforward way.

2. Statement of the problem

We consider the temperature distribution for the set of points $I(\alpha_1<\alpha<\alpha_2)$, when $B(\alpha=\alpha_2)$ is kept at the temperature $u(\alpha_2, \beta)=0$ and $A(\alpha=\alpha_1)$, is kept at the temperature $u(\alpha_1, \beta)=\varphi(\cos\beta)\equiv f(\beta)$ say, such that

$$f(\beta)\in \mathrm{lip}^* \frac{1}{2}. \tag{1.2}$$

We shall use the following lemmas in the solution of the problem:

LEMMA 1. *The n^{th} partial sum $S_n(x)$ of the series expansion*

$$\sum_{n=0}^{\infty} a_n P_n(x) \tag{2.1}$$

is given by

$$S_n(x)-\varphi(x) = \int_{-1}^{1} \psi(z) K_n(z)\, dz, \tag{2.2}$$

where

$$a_n = \left(n+\frac{1}{2}\right) \int_{-1}^{1} \varphi(y) P_n(y)\, dy,$$

$$\psi(z) = \psi(x, z) = \int_{-1}^{1} [\varphi(y)-\varphi(x)] K(x, y, z)\, dy,$$

$$K(x, y, z) = (1-x^2-y^2-z^2+2xyz)^{-\frac{1}{2}}$$

and

$$K_n(z) = \frac{1}{2}\{P_n'(z)+P_{n+1}'(z)\}.$$

PROOF. We have

$$S_n(x)-\varphi(x) = \sum_{k=0}^{\infty} a_k P_k(x)$$

$$= \int_{-1}^{1} [\varphi(y)-\varphi(x)] \left\{\sum_{k=0}^{n} \left(k+\frac{1}{2}\right) P_k(x)\, P_k(y)\right\} dy$$

$$= \int_{-1}^{1}\int_{-1}^{1} [\varphi(y)-\varphi(x)]\cdot K(x, y, z) \left\{\sum_{k=9}^{n} \left(k+\frac{1}{2}\right) P_k(z)\right\} dy\, dz$$

(see [4], Theorem 1, for $\alpha=\beta=0$)

$$= \int_{-1}^{1}\int_{-1}^{1} [\varphi(y)-\varphi(x)]\cdot K(x, y, z)\, dy \cdot \frac{1}{2}[P_n'(z)+P_{n+1}'(z)]\, dz$$

$$= \int_{-1}^{1} \psi(z) K_n(z)\, dz.$$

LEMMA 2.

$$f(\beta)\in \mathrm{lip}^*\frac{1}{2} \quad \textit{implies} \quad \psi(z)\in \mathrm{lip}^*\frac{1}{2}.$$

PROOF. We have

$$\psi(z+\tau)-\psi(z)=\int_{-1}^{1}[\varphi(y)-\varphi(x)]\{K(x,y,z+\tau)-K(x,y,z)\}\,dy$$

$$=\int_{-1-x}^{1-x}[\varphi(x+t)-\varphi(x)]\{K(x,x+t,z+\tau)-K(x,x+t,z)\}\,dt.$$

Substituting $x=\cos\beta$, $x+t=\cos\theta$ and $\theta=\beta-\gamma$ we find that

$$|\psi(z+\tau)-\psi(z)|=\Big|\int_{\beta-\pi}^{\beta}[\varphi\{\cos(\beta-\gamma)\}-\varphi(\cos\beta)]\cdot\{K[\cos\beta,\cos(\beta-\gamma)z+\tau]-$$

$$-K[\cos\beta,\cos(\beta-\gamma),z]\}\,d\gamma\Big|$$

$$\geqq \sup|\varphi\{\cos(\beta-\gamma)\}-\varphi(\cos\beta)|\int_{\beta-\pi}^{\beta}|K\{\cos\beta,\cos(\beta-\gamma),z+\tau\}-$$

$$-K\{\cos\beta,\cos(\beta-\gamma),z|\,d\gamma$$

$$=A\sup\{|f(\beta-\gamma)-f(\beta)|\},$$

where A is a positive constant, not necessarily the same at each occurrence.

Thus we infer that, for $|\tau|=|\gamma|$

$$\psi(z)\in \mathrm{lip}^*\frac{1}{2}.$$

3. Solution of the problem

In view of the azimuthal symmetry the equation satisfied by the temperature function $u(\alpha,\beta)$, may be written as

$$\frac{1}{\sinh\alpha}\frac{\partial}{\partial\alpha}\left(\sinh\alpha\frac{\partial\mu}{\partial\alpha}\right)+\frac{1}{\sin\beta}\cdot\frac{\partial}{\partial\beta}\left(\sin\beta\frac{\partial\mu}{\partial\beta}\right)=0. \tag{3.1}$$

The solution of this equation is given by (see [5], p. 216)

$$u(\alpha,\beta)=\sum_{n=0}^{\infty}[M_nP_n(\cosh\alpha)+N_nQ_n(\cosh\alpha)]\cdot P_n(\cos\beta). \tag{3.2}$$

Using the boundary condition at B, we get

$$N_n=-M_n\cdot\frac{P_n(\cosh\alpha_2)}{Q_n(\cosh\alpha_2)}.$$

Substituting the expression for N_n in (3.2), we have

$$(3.3)\qquad u(\alpha,\beta)=\sum_{n=0}^{\infty} M_n P_n(\cos\beta)\cdot\frac{P_n(\cosh\alpha)Q_n(\cosh\alpha_2)-Q_n(\cosh\alpha)P_n(\cosh\alpha_2)}{Q_n(\cosh\alpha_2)}.$$

Now on account of the boundary condition at A, we find that

$$(3.4)\qquad f(\beta)=\sum_{n=0}^{\infty} M_n\cdot P_n(\cos\beta)\cdot\frac{P_n(\cosh\alpha_1)Q_n(\cosh\alpha_2)-Q_n(\cosh\alpha_1)P_n(\cosh\alpha_2)}{Q_n(\cosh\alpha_2)}.$$

Finally, using the orthogonal property of the Legendre polynomials, we have

$$M_n=\frac{\left(n+\frac{1}{2}\right)Q_n(\cosh\alpha_2)\int_0^{\pi} f(\beta)P_n(\cos\beta)\sin\beta\,d\beta}{P_n(\cosh\alpha_1)Q_n(\cosh\alpha_2)-Q_n(\cosh\alpha_1)P_n(\cosh\alpha_2)}.$$

Thus the series (3.3) may be rewritten in the form

$$(3.5)\qquad u(\alpha,\beta)=\sum_{n=0}^{\infty}\frac{f_n[P_n(\cosh\alpha)Q_n(\cosh\alpha_2)-Q_n(\cosh\alpha)P_n(\cosh\alpha_2)]}{[P_n(\cosh\alpha_1)Q_n(\cosh\alpha_2)-Q_n(\cosh\alpha_1)P_n(\cosh\alpha_2)]}\cdot P_n(\cos\beta)$$

where

$$f_n=\left(n+\frac{1}{2}\right)\int_0^{\pi} f(\beta)P_n(\cos\beta)\sin\beta\,d\beta=\left(n+\frac{1}{2}\right)\int_{-1}^{1}\varphi(y)P_n(y)\,dy.$$

Convergence of the Series in (3.5).
We have

$$(3.6)\qquad \begin{aligned}\sum_{n=0}^{\infty} f_n(P_n(x))&=\sum_{n=0}^{\infty}\left(n+\frac{1}{2}\right)\int_{-1}^{1}\varphi(y)P_n(x)P_n(y)\,dy\\&=\sum_{n=0}^{\infty}\left(n+\frac{1}{2}\right)\int_{-1}^{1}\int_{-1}^{1}\varphi(y)K(x,y,z)\cdot P_n(z)\,dy\,dz.\end{aligned}$$

(See [4], theorem 1, for $\alpha=\beta=0$.)
By Lemma 1, the nth partial sum of the series (3.6) is given by

$$S_n(x)-\varphi(x)=\int_{-1}^{1}\psi(z)\cdot K_n(z)\,dz$$

i.e.

$$S_n(x)-[\varphi(x)+\psi(1)]=\int_{-1}^{\pi}[\psi(z)-\psi(1)]\cdot K_n(z)\,dz$$

that is

$$(3.7)\qquad S_n(\cos\beta)-[f(\beta)+\psi(1)]=\int_0^{\pi}\chi(\omega)K_n(\cos\omega)\sin\omega\,d\omega$$

where

$$\chi(\omega)=\psi(\cos\omega)-\psi(1).$$

By Lemma 2, it can be easily seen that

$$\chi(\omega) \in \operatorname{lip}^* \frac{1}{2}. \tag{3.8}$$

DU PLESSIS [3], has proved that under the condition (3.8) the expression on the right hand side of (3.7) is $O(1)$ as $n \to \infty$ thus, it follows that the sequence $\{s_n(\cos \beta)\}$ is uniformly bounded.

Also, on the lines of Bhonsle (loc. cit.) it can be easily seen that if $n > n_0$, where n_0 is a suitable integer, then

$$\frac{P_n(\cosh \alpha) Q_n(\cosh \alpha_2) - Q_n(\cosh \alpha) P_n(\cosh \alpha_2)}{P_n(\cosh \alpha_1) Q_n(\cosh \alpha_2) - Q_n(\cosh \alpha_1) P_n(\cosh \alpha_2)}$$

is a positive function of n and tends steadily to zero as $n \to \infty$.

Hence, by virtue of Dirichlet's criterion (see [8], p. 4) for uniform convergence, the series (3.5) converges uniformly for $\alpha_1 < \alpha < \alpha_2$, and $0 \leqq \beta \leqq \pi$ and we infer that the temperature distribution function $u(\alpha, \beta)$ is a continuous function of α and β.

Finally it is interesting to note that Du Plesis (loc. cit.) by means of an example, has shown that for the convergence of the sequence $\{S_n(\cos \beta)\}$ the condition (3.8) is the best possible. This implies that for the solution of our problem any improvement in the condition (1.2) is not possible.

I wish to express my very deep gratitude to professor D. P. Gupta for his valuable suggestions in the preparation of this paper.

REFERENCES

[1] BHONSLE, B. R.: Steady State heat flow in a shell enclosed between two prolate spheroids, *Mathematica Japonicae* Vol. **12**, No. 1 (1967), 101—105.
[2] CARSLAW, H. S., JAEGER, J. C.: *Conduction of heat in solids*, Oxford University Press, 1959.
[3] DU PLESSIS, N.: The Cesaro summability of Laplace series, *Journ. London Math. Soc.* Vol. **27** (1952), 337—352.
[4] GASPER, G.: Positivity and the convolution structure for Jacobi series, *Annals of Maths.* Vol. **93**, No. 1 (1971), 112—118.
[5] LEBEDEV, N. N.: *Special functions and their applications*, Prentice-Hall, Inc. 1965.
[6] REDDICK, H. W. and MILLER, F. H.: *Advanced Mathematics for Engineers*, Asia Publishing House, Bombay, 1962.
[7] SZEGŐ, G.: *Orthogonal Polynomials*, American Math. Soc., Colloquium publications, Vol. XXIII, New York, 1959.
[8] TITCHMARSH, E. C.: *The theory of functions*, Oxford University Press, 1952.

Department of Mathematics, Saifia College, Bhopal, India

(Received May 6, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 419—422.

DELAYED AVERAGES OF A STATIONARY GAUSSIAN SEQUENCE

by

C. M. DEO

T. L. LAI (1974) has proved an interesting limit theorem for the so-called delayed averages of independent random variables. It is apparent from his proof that his theorem would apply to a stationary but dependent sequence of random variables only if one would impose fairly stringent conditions on the moments and dependence structure of the sequence. In this note we deal with a stationary Gaussian sequence and show that such a sequence obeys Lai's theorem under a simple condition on the correlation sequence. It was shown, in DEO (1974) that under the same sufficient condition a stationary Gaussian sequence obeys Strassen's law of iterated logarithm. Also Example 3 in DEO (1974) indicates that this condition cannot be significantly weakened even for Lai's theorem.

Let $\{\xi_n: 0 \leqq n < \infty\}$ be a stationary Gaussian sequence with $E(\xi_0^2)=1$, $E(\xi_0 \xi_j)= r_j$, $j \geqq 0$. We will assume that

$$\lim_{n \to \infty} n^{1+\beta} r_n = 0 \quad \text{for some} \quad \beta > 0. \tag{1}$$

Let $0<\alpha<1$. Let S_n be the partial sum $\sum_{j=1}^{n} \xi_j$, and let T_n denote the "delayed" partial sum $\sum_{j=n}^{n+[n^\alpha]-1} \xi_j$. Here, as usual, $[\cdot]$ is the greatest integer function. Under condition (1) the series $\sum_{n=1}^{\infty} r_n$ converges absolutely and we write $\sigma^2 = 1 + 2 \sum_{n=1}^{\infty} r_n$. We will assume that $\sigma > 0$; although by a straightforward adaptation of the proof given below it is easy to see that the theorem in this note remains true even when $\sigma = 0$.

THEOREM. *Assume* (1). *Then for each* α, $0<\alpha<1$, *relations* (2) *and* (3) *below hold with probability one.*

$$\limsup_{n \to \infty} \{2(1-\alpha) n^\alpha \log n\}^{-1/2} T_n = \sigma. \tag{2}$$

$$\liminf_{n \to \infty} \{2(1-\alpha) n^\alpha \log n\}^{-1/2} T_n = -\sigma. \tag{3}$$

PROOF. It suffices to prove (2), since (3) follows from (2) by considering the sequence $\{-\xi_n\}$.

Note that variance of T_n is asymptotically equal to $n^\alpha \sigma^2$. We first show that the limsup in (2) is at least σ. Let $n_k = \left[\lambda k^{\frac{1}{1-\alpha}}\right]$ where we choose $\lambda > 0$ large enough

so that for all large k, $n_{k+1}-(n_k+n_k^\alpha)>\lambda k^{\frac{\alpha}{1-\alpha}}$. For such λ it is easy to see that we can find an integer k_0 such that $n_{k+p}-(n_k+n_k^\alpha)>p$ for all $k\geqq k_0$ and all $p\geqq 1$. Let now $\delta_n=\sup_{k\geqq n}|r_k|$. Then, by (1), we can find a finite positive number A such that $n^{1+\beta}\delta_n<A$ and $\sum_{j=n}^{\infty}\delta_j<An^{-\beta}$ for all n. Let us write $Y_k=n_k^{-\alpha/2}\sigma^{-1}T_{n_k}$.

Note that $E(Y_k)=0$, $E(Y_k^2)\sim 1$. We next show that $|E(Y_kY_{k+p})|<Bp^{-\beta}$, $\forall k\geqq k_0$ and $\forall p\geqq 1$ where the constant B is independent of k and p. We have

$$|E(Y_kY_{k+p})| = \left|(\sigma^4 n_k^\alpha n_{k+p}^\alpha)^{-1/2}\sum_{i=n_k}^{n_k+[n_k^\alpha]-1}\sum_{j=n_{k+p}}^{n_{k+p}+[n_{k+p}^\alpha]-1}E(\xi_i\xi_j)\right| \leqq \tag{4}$$

$$\leqq (\sigma^4 n_k^\alpha n_{k+p}^\alpha)^{-1/2}\left\{\sum_{j=n_{k+p}-(n_k+[n_k^\alpha])}^{n_{k+p}-n_k-1} j\delta_j + n_k^\alpha\sum_{j=n_{k+p}-n_k}^{\infty}\delta_j\right\}.$$

Now the first sum in the brackets has $[n_k^\alpha]$ terms each of which is less than $Ap^{-\beta}$. This follows from the fact that $n_{k+p}-(n_k+[n_k^\alpha])>p$ and $j\delta_j<Aj^{-\beta}$. Also the second sum above is less than $Ap^{-\beta}$. Combining this with $n_{k+p}>n_k$ we see that the entire expression above is dominated by $2A\sigma^{-2}p^{-\beta}$. Thus $|E(Y_kY_{k+p})|<Bp^{-\beta}$, $\forall k\geqq k_0$ and $\forall p\geqq 1$ where B can be taken to be $2A\sigma^{-2}$. We can now apply Lemma 1 in DEO (1973) to conclude that, for each $\delta>0$, Y_k exceeds $(1-\delta)(2\log k)^{1/2}$ infinitely often with probability one. This is equivalent to asserting that T_{n_k} exceeds $\sigma(1-\delta)(2n_k^\alpha(1-\alpha)\log n_k)^{1/2}$ for infinitely many k with probability one. Thus the limsup in (2) is at least σ.

We now show that the limsup in (2) is no greater than σ. Again, as in LAI (1974), let $m_k=[(k/\log k)^{\frac{1}{1-\alpha}}$. For a standard normal variable Z,

$$P[Z>x]\sim(2\pi x^2)^{-1/2}\exp\left(-\frac{1}{2}x^2\right) \quad \text{as} \quad x\to\infty. \tag{5}$$

Let now $\varepsilon>0$. Using (5) it is easy to see that the probability $P[T_{m_k}>(1+\varepsilon)\sigma\{2m_k(1-\alpha)\log m_k\}^{1/2}]$ is, for large k, dominated by $k^{-\gamma}$ for some $\gamma>1$. Hence by Borel—Cantelli lemma T_{m_k} exceeds $(1+\varepsilon)\sigma\{2m_k(1-\alpha)\log m_k\}^{1/2}$ only finitely often with probability one.

Note that $\{2m_k^\alpha(1-\alpha)\log m_k\}\sim\{2m_{k+1}^\alpha(1-\alpha)\log m_{k+1}\}$. Hence to complete the proof it suffices to show that, for each $\delta>0$, $\max_{m_k\leqq n\leqq m_{k+1}}|T_n-T_{m_k}|$ exceeds $\sigma\delta\{2(1-\alpha)m_k^\alpha\log m_k\}^{1/2}$ only finitely often a.s. This in turn is implied by

$$\sum_{k=1}^{\infty}\sum_{n=m_k}^{m_{k+1}}P[|T_n-T_{m_k}|>\delta\sigma\{2(1-\alpha)m_k\log m_k\}^{1/2}]<\infty. \tag{6}$$

It remains to prove (6). Let us note that $m_{k+1}-m_k\sim k^{\frac{\alpha}{1-\alpha}}(\log k)^{-\frac{1}{1-\alpha}}$. Hence $m_k+[m_k^\alpha]>m_{k+1}$ for large k. Now $T_n-T_{m_k}$ is a sum of ξ_j' some occuring with a positive sign and some with negative sign. The total number of ξ_j's involved in this sum is $(n-m_k)+(n+[n^\alpha]-m_k-[m_k^\alpha])$ which is easily seen to be less than $3(m_{k+1}-m_k)$.

Thus the variance of $T_n - T_{m_k}$ is asymptotically at most $3\sigma^2(m_{k+1}-m_k)$. Hence, by (5) again, for large k,

$$P[|T_n - T_{m_k}| > \delta\sigma\{2(1-\alpha)m_k^\alpha \log m_k\}^{1/2}] > \exp\left[-\frac{(1-\alpha)m_k^\alpha \log m_k}{3(m_{k+1}-m_k)} \cdot \delta'^2\right]$$

where $0 < \delta' < \delta$. Again use the fact that

$$m_{k+1} - m_k \sim k^{\frac{\alpha}{1-\alpha}} (\log k)^{-\frac{1}{1-\alpha}}$$

and

$$m_k^\alpha (1-\alpha) \log m_k \sim k^{\frac{\alpha}{1-\alpha}} (\log k)^{\frac{1-2\alpha}{1-\alpha}}$$

and conclude that

$$\sum_{n=m_k}^{m_{k+1}} P[|T_n - T_{m_k}| > \delta\sigma\{2(1-\alpha)m_k \log m_k\}^{1/2}] \leqq$$

$$\leqq k^{\frac{\alpha}{1-\alpha}} (\log k)^{-\frac{1}{1-\alpha}} \exp\left[-\frac{\delta'^2 (\log k)^2}{3}\right].$$

This proves (6) and hence the theorem.

In conclusion it would be interesting to find out if the hypotheses of the theorem are sufficient for a Strassen-type functional form of this theorem.

REFERENCES

[1] DEO, C. M.: An iterated logarithm law for maxima of nonstationary Gaussian processes, *J. Appl. Probl.* **10** (1973), 402—408.
[2] DEO, C. M.: A note on stationary Ganssian sequences. *Ann. Probability* **2** (1976), 954—957.
[3] LAI, T. L.: Limit theorems for delayed sums, *Ann. Probability* **2** (1974), 432—441.

Mathematics Department, University of Ottawa, Ottawa, Ontario, K1N 6N5, Canada

(Received June 20, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 423—426.

SOME REGULARITY PROPERTIES OF THE L^1 AND L^2 METRICS ON PROBABILITY MEASURES

by

M. KANTER

1. Introduction

In this paper we study monotonicity and continuity properties of two equivalent metrics on probability measures for a class of Markov processes which include the "semi-stable" processes of Lamperti as well as stable stochastic processes.

2. General Result

Letting $\mathscr{B}$ represent a σ-field of subsets of an abstract set Ω, we consider two probability measures μ and ν defined on $\mathscr{B}$ and define $H_k(\mu, \nu)$ as follows:

$$H_k(\mu, \nu) = \int_\Omega \left| \left(\frac{d\mu}{d\lambda}\right)^{k-1} - \left(\frac{d\nu}{d\lambda}\right)^{k-1} \right|^k d\lambda \tag{2.1}$$

where λ is any non-negative measure on B which dominates both μ and ν.

We will consider $H_k(\mu, \nu)$ only for $k=1$ or $k=2$. When $k=1$, $H_k(\mu, \nu)$ is the usual total variation metric on probability measures, while for $k=2$, $H_k^{1/2}(\mu, \nu)$ is a metric on probability measures which is equivalent to the total variation metric as shown in [6]. We call these two metrics respectively the L^1 and L^2 metrics.

We now make a simple but useful observation. We define

$$f_k(\mu, \nu) = 1-(1/2)H_k(\mu, \nu) \tag{2.2}$$

and we notice that for $k=1$ we can rewrite (2.2) as

$$f_1(\mu, \nu) = \int_\Omega \min\left\{\frac{d\mu}{d\lambda}, \frac{d\nu}{d\lambda}\right\} d\lambda \tag{2.3}$$

while for $k=2$ we can rewrite (2.2) as

$$f_2(\mu, \nu) = \int_\Omega \left(\frac{d\mu}{d\lambda} \cdot \frac{d\nu}{d\lambda}\right)^{1/2} d\lambda. \tag{2.4}$$

Remembering that both the functions $\sqrt{a \cdot b}$ and $\min\{a, b\}$ are non-negative definite on $(0, \infty)\times(0, \infty)$ we conclude the following lemma.

LEMMA 2.1. *If $\{\mu_t | t \in T\}$ is any set of probability measures on (Ω, B) then the functions $f_k(\mu_t, \mu_s)$ are non-negative definite on $T\times T$ for $k=1$ or 2.*

We shall need the following basic lemma in the proof of our main result.

LEMMA 2.2. *Let $P(A, w)$ be a Markov kernel on $(\Omega, \mathscr{B})$. Let ν_1 and ν_2 be probability measures. Then for $k=1$ or $k=2$ we have*

$$H_k(P\circ\nu_1, P\circ\nu_2) \leqq H_k(\nu_1, \nu_2)$$

where, for any measure μ, we let $P\circ\mu$ denote the measure with values

$$P\circ\mu(A) = \int_\Omega P(A, w)\, d\mu(w)$$

for $A\in\mathscr{B}$.

PROOF. Let $\varphi_1(x)=|1-x|$ and let $\varphi_2(x)=2-2\sqrt{x}$. Then we have $H_k(\nu_1, \nu_2)=$ $=I_{\varphi_k}(\nu_1, \nu_2)$ where for any function $\varphi(x)$: $[0, \infty)\to R$ we define

$$\Phi_\varphi(\nu_1, \nu_2) = \int_\Omega p_1(w)\varphi\left(\frac{p_2(w)}{p_1(w)}\right) d\lambda(w)$$

where λ is any measure that dominates ν_1 and ν_2, and where $p_i=\dfrac{d\nu_i}{d\lambda}$.

Now the main theorem of [3] states that if $\varphi(x)$ is a convex function on $[0, \infty]$ then $I_\varphi(\nu_1, \nu_2)\geqq I_\varphi(P\circ\nu_1, P\circ\nu_2)$ for any Markov kernel $P_x(A, w)$. To get the result of the lemma, we just apply this fact to the particular convex functions φ_1 and φ_2. Q. e. d.

We shall call a set of measures $\{\mu_t | t\in(0, \infty)\}$ a screw process of order k if $H_k(\mu_{ts}, \mu_{t's})=H_k(\mu_t, \mu_{t'})$ for all t, t', s in $(0, \infty)$. If, furthermore, the measures μ_t are the marginal distributions of some Markov process $(X(t)|t\in(0, \infty))$ with stationary transition probability kernel and with state space $(\Omega, \mathscr{B})$ then we shall call the process $\{\mu_t | t\in(0, \infty)\}$ a screw Markov process of order k. The following is our main result.

THEOREM 2.1. *Let $\{\mu_t | >0\}$ be a non-constant screw Markov process. Then, if $k=1$ or 2, the function $H_k(\mu_1, \mu_t)$ is a unimodal function on $(0, \infty)$ with unique minimum at $t=1$. Furthermore $H_k(\mu_1, \mu_t)$ is continuous on $(0, 1)\cup(1, \infty)$ and symmetric in the sense that $H_k(\mu_1, \mu_t)=H_k(\mu_1, \mu_t-1)$.*

REMARK. Before we start the proof of Theorem 2.1 we remember that a real-valued function f defined on a sub-interval J of R, the real line, is said to be unimodal if there exists some t in J such that f is monoton on $(-\infty, t)\cap J$ and on $(t, \infty)\cap J$.

PROOF. Let $k=1$ or 2. The fact that $H_k(\mu_1, \mu_t)=H_k(\mu_1, \mu_t-1)$ is obvious using the fact that $H_k(\mu, \nu)=H_k(\nu, \mu)$ and the screw property of $\{\mu_t | t>0\}$.

We now prove unimodality. By symmetry we need only show that $H_k(\mu_1, \mu_t)$ is increasing with t on $(1, \infty)$.

Let $x>0$. We use the fact that μ_t are the marginals of a Markov process with stationary transition probability kernel and Lemma 2.2 to conclude that $H_k(\mu_1, \mu_t)\geqq$ $\geqq H_k(\mu_{1+x}, \mu_{t+x})$. By our assumptions we know $H_k(\mu_{1+x}, \mu_{t+x})=H_k(\mu_1, \mu_{(t+x)(1+x)^{-1}})$. Now if $1<t'<t$, then setting $x'=(t-t')(t'-1)^{-1}$ we have $(t+x')(1+x')^{-1}=t'$ hence we can conclude that $H_k(\mu_1, \mu_t)\geqq H_k(\mu_1, \mu_{t'})$. This finishes the proof of unimodality.

To prove continuity note that $f_k(\mu_1, \mu_{ts-1})$ is a non-negative definite function of s and t on $(0, \infty)\times(0, \infty)$. From [8] it follows that $f_k(\mu_1, \mu_{e^x})$ is a.e. equal to some characteristic function $f(x)=\int_R e^{ixy}dF(y)$ where F is a probability distribution on R. We claim that $f_k(\mu_1, \mu_{e^x})$ is continuous in any open interval (a, b) which does not contain 0. To prove this claim we note that $f_k(\mu_1, \mu_{e^x})$ is monotone on (a, b) by unimodality. If for some $x\in(a, b)$ we have $\lim_{y\uparrow x} f_k(\mu_1, \mu_{e^y})\neq\lim f_k(\mu_1, \mu_{e^y})$ then it also follows that $\lim_{y\uparrow x} f(y)\neq\lim_{y\downarrow x} f(y)$, which contradicts the fact that f is continuous. By monotonicity we conclude that $\lim_{y\to x} f_k(\mu_1, \mu_{e^y})=f_k(\mu_1, \mu_{e^x})$ and our claim is proved. We conclude that $f_k(\mu_1, \mu_{e^x})$ is continuous except possibly to $x=0$.

All that is left to prove is that the point $t=1$ is the unique minimum value of $H_k(\mu_1, \mu_t)$. If not then $f_k(\mu_1, \mu_{e^x})$ is equal to 1 in a non-empty open interval about $x=0$ which implies $f_k(\mu_1, \mu_{e^x})=1$ for all x. It then follows that $\mu_t=\mu_1$ for all t, which contradicts the assumption that $\{\mu_t|t>0\}$ is not constant. Q. e. d.

3. Applications and Examples

Suppose $(\Omega, \mathcal{B})$ is a measurable linear space, by which we mean that the operation of addition $(x, y)\to x+y$ from $\Omega\times\Omega$ into Ω is measurable with respect to $\mathcal{B}\times\mathcal{B}$, while the operation of scalar multiplication $(a, x)\to ax$ from $R\times\Omega$ into Ω is measurable with respect to $\mathcal{B}_R\times\mathcal{B}$, where $\mathcal{B}_R$ denotes the Borel subsets of R.

DEFINITION 3.1. A random process $(X(t)|t\in(0, \infty))$ with state space $(\Omega, \mathcal{B})$ is said to be *semi-stable of index* α if for all $a>0$ the process $(X(at)|t\in(0, \infty))$ has the same joint distributions as the process $(a^{1/\alpha}X(t)|t\in(0, \infty))$. LAMPERTI has studied semi-stable processes in [5], in the special case when Ω equals R. A semi-stable process with stationary and independent increments is called a "stable convolution semi-group".

LEMMA 3.1. *If* $\{\mu_t|t>0\}$ *are the marginal distributions of a semi-stable Markov process of index* α, *then* $\{\mu_t|t>0\}$ *is a screw Markov process of order* k, *where* $k=1$ *or* 2.

PROOF. Let S_a: $\Omega\to\Omega$ stand for the 1-1, bimeasurable map sending x into $a^{1/\alpha}x$, where $a\in(0, \infty)$ and $x\in\Omega$. We can express μ_{at} as $\mu_t\circ S_a^{-1}$ by the hypotheses of theorem. We wish to show that $H_k(\mu_t, \mu_s)=H_k(\mu_t\circ S_a^{-1}, \mu_s\circ S_a^{-1})$. However if λ dominates μ_t and μ_s then $\lambda\circ S_a^{-1}$ dominates $\mu_t\circ S_a^{-1}$ and $\mu_s\circ S_a^{-1}$. Furthermore a simple computation shows that the function $\left(\frac{d\mu}{d\lambda}\right)\circ S_a^{-1}$ is a version of the Radon—Nikodym derivative $\frac{d\mu\circ S_a^{-1}}{d\lambda\circ S_a^{-1}}$. The asserted equality now follows trivially by the general formula

$$\int_\Omega f\circ S_a^{-1}(y)\, d\lambda\circ S_a^{-1}(y) = \int_\Omega f(y)\, d\lambda(y) \tag{3.1}$$

which is valid for any non-negative and measurable function f. Q. e. d.

To construct a screw Markov process which is not semi-stable, let $(X_1(t)|t>0)$ with state space $(\Omega_1, \mathscr{B}_1)$ and $(X_2(t)|t>0)$ with state space $(\Omega_2, \mathscr{B}_2)$ be two semi-stable and mutually independent stochastic processes of index α_1 and α_2 respectively with $\alpha_1 \neq \alpha_2$. Consider the new process $(X(t)|t>0)=((X_1(t), X_2(t))|t>0))$ with state space $\Omega=\Omega_1\times\Omega_2$ and $\mathscr{B}=\mathscr{B}_1\times\mathscr{B}_2$. The process $(X(t)|t>0)$ is clearly not semi-stable, however it is a screw Markov process of order k where $k=1$ or 2. This follows by the proof of Lemma 3.1 where S_a: $\Omega_1\times\Omega_2 \to \Omega_1\times\Omega_2$ is

$$S_a(x, y) = a^{1/\alpha_1}x, a^{1/\alpha_2}y) \tag{3.2}$$

for $(x, y)\in\Omega_1\times\Omega_2$.

We end this paper with some remarks on stable distributions. A measure μ on R^n is called stable of index α in $(0, 2]$ if for all $a, b>0$ we have

$$\mu\circ S_a^{-1} * \mu\circ S_b^{-1} = \mu\circ S_c^{-1}$$

for $c=(a^\alpha+b^\alpha)^{1/\alpha}$, where $*$ denotes convolution. It is easy to see that a stable measure μ on R^n of index α can be imbedded into a stable convolution semi-group $\{\mu_t|t>0\}$ of index α so that $\mu=\mu_1$. We conclude that for any stable measure μ on $(R^n, \mathscr{B}_{R^n})$ the function $f(a)=H_k(\mu, \mu\circ S_a^{-1})$ is a unimodal function of a. (In [4] this fact was established when μ was either (1) a symmetric stable measure on R or (2) a stable measure on $(0, \infty)$.)

REFERENCES

[1] BERMAN, S.: A New Characterization of Characteristic Functions of Absolutely Continuous Distributions, *Pacific Journal of Math.* (to appear).

[2] CSISZÁR, I.: Über Topologische und Metrische Eigenschaften der Relativen Information der Ordnung α, *Trans. of the 3rd Prague Conference on Information Theory* (1964), 63—75.

[3] CSISZÁR, I.: Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten, *Publ. Math. Inst. Hung. Acad. Sci.* **A8** (1963), 85—108.

[4] KANTER, M.: Stable Densities Under Change of Scale and Total Variation Inequalities, *Annals of Probability* (1975), 697—707.

[5] LAMPERTI, L.: Semi-stable Markov Processes I, *Zeitschr. f. Wahrscheinlichkeitstheorie u. Verw. Geb.* **22** (1972), 205—225.

[6] LECAM, L.: On the Assumptions Used to Prove Asymptotic Normality of Maximum Likelihood Estimates, *Ann. Math. Stat.* (1970), 802—828.

[7] LOÈVE, M.: *Probability Theory*, 3rd ed., Van Nostrand, New York, 1963.

[8] RIESZ, F.: Über Sätze von Stone und Bochner, *Acta Sci. Math Szeged* **6** 1933, 184—198.

Sir George Williams University and University of New South Wales

(Received October 15, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 427—434.

A STRONG APPROXIMATION OF THE MULTIVARIATE EMPIRICAL PROCESS

by

M. CSÖRGŐ[1] and P. RÉVÉSZ

§ 1. Introduction

In order to study the properties of the one-dimensional empirical distribution function (e.d.f.) the following approximation theorem is fundamental:

THEOREM A. ([1]) *Let $X_1, X_2, \ldots$ be a sequence of independent, uniform — [0, 1] r.v.'s defined on a rich enough probability space. Further, let $E_n(x)$ $(0 \leqq x \leqq 1)$ be the e.d.f. based on the sample $X_1, X_2, \ldots, X_n$ and let $\alpha_n(x) = \sqrt{n}\,(E_n(x) - x)$ be the Empirical Process (E. P.). Then one can define a sequence $\{B_n(x)\}$ of Brownian Bridges (B. B.'s) and a Kiefer Process (K.P.) $K(x; y)$ $(0 \leqq x \leqq 1, 0 \leqq y < \infty)$ such that*

$$\sup_{0 \leqq x \leqq 1} |\alpha_n(x) - B_n(x)| \overset{\text{a.s.}}{=} O(n^{-\frac{1}{2}} \log n)$$

and

$$\sup_{0 \leqq x \leqq 1} |\sqrt{n}\alpha_n(x) - K(x; n)| \overset{\text{a.s.}}{=} O(\log^2 n).$$

The precise meaning of "rich enough" will not be formulated from time to time; it will, however, be enough to assume all the time that an independent sequence of Wiener Processes (W.P.'s) $\{W_n(x)\}$, which is independent of the originally given i.i.d. sequence $\{X_n\}$, can be constructed on the assumed probability space. From now on it will be assumed that the underlying probability space is rich enough in this sense.

It is natural to ask how these results can be extended to the multidimensional case. An answer to this question is the following:

THEOREM B. ([2]) *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ be a sequence of independent r.v.'s uniformly distributed over the unit cube I^d of the d-dimensional Euclidean space. Then one can define a sequence $\{B_n(\mathbf{x})\}$ of B.B.'s and a K.P. $K(\mathbf{x}; y)$ $(\mathbf{x} \in I^d, 0 \leqq y < \infty)$ such that*

$$\sup_{\mathbf{x} \in I^d} |\alpha_n(\mathbf{x}) - B_n(\mathbf{x})| \overset{\text{a.s.}}{=} O\left(n^{-\frac{1}{2(d+1)}} (\log n)^{\frac{3}{2}}\right),$$

$$\sup_{\mathbf{x} \in I^d} |n^{\frac{1}{2}} \alpha_n(\mathbf{x}) - K(\mathbf{x}; n)| \overset{\text{a.s.}}{=} O\left(n^{\frac{d+1}{2(d+2)}} \log^2 n\right)$$

where the E.P. $\alpha_n(\mathbf{x}) = \alpha_n(x_1, x_2, \ldots, x_d) = n^{\frac{1}{2}}(E_n(\mathbf{x}) - x_1 x_2 \ldots x_d)$ and $E_n(\mathbf{x})$ is the e.d.f. based on the sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$.

[1] Research partially supported by a Canadian N.R.C. Grant.

(The definition of $B_n(\mathbf{x})$ and $K(\mathbf{x}; n)$ is given in paragraph 2.)

Up to now we investigated only the case when the sample is coming from the uniform law. In the one dimensional case Theorem A immediately implies:

THEOREM C. *Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d.r.v.'s having a continuous distribution function $F(x)$. Then one can define a sequence $\{B_n(x)\}$ of B.B.'s and a K.P. $K(x; y)$ such that*

$$\sup_{-\infty<x<\infty} |\beta_n(x) - B_n(F(x))| \overset{\text{a.s.}}{=} O(n^{-\frac{1}{2}} \log n),$$

$$\sup_{-\infty<x<\infty} |n^{\frac{1}{2}} \beta_n(x) - K(F(x); n)| \overset{\text{a.s.}}{=} O(\log^2 n)$$

where the E.P. $\beta_n(x) = n^{\frac{1}{2}}(F_n(x) - F(x))$ and $F_n(x)$ is the e.d.f. based on the sample $Y_1, Y_2, \ldots, Y_n$.

In order to prove Theorem C, one only notes that $F(Y_1), F(Y_2), \ldots$ are independent, uniform $-[0, 1]$ r.v.'s.

A two-dimensional analogue of Theorem C is given in [3], where the more general problem of strong approximation of the two-dimensional empirical *measure* over sets with smooth boundaries is studied. Applying then the result on the approximation of the empirical measure, in [3] there is given a strong approximation of the two-dimensional E.P. with a somewhat weaker rate of convergence than in our Theorem 1. It is not immediate how the method of [3] could be generalized to more than two dimensions, which would be of interest in itself.

The aim of our present paper is to study the problem of strong approximation of the multivariate empirical process directly for arbitrary dimensions and for arbitrary continuous distribution functions and to retain the rate of convergence of Theorem B while doing so.

The multivariate weak approximation problem of the E.P. was first studied by DUDLEY ([4]).

§ 2. The result and notations

First some notations and definitions:

D. 1. Wiener Process (W.P.): A separable Gaussian Process (G.P.) $W(\mathbf{x}) = \{W(x_1, x_2, \ldots, x_d);\ 0 \leqq x_i < \infty;\ i = 1, 2, \ldots, d\}$ with $EW(\mathbf{x}) = 0$ and

$$R(\mathbf{x}_1, \mathbf{x}_2) = EW(\mathbf{x}_1)W(\mathbf{x}_2) = \prod_{i=1}^{d} \min(x_{1i}, x_{2i}),$$

$$\mathbf{x}_1 = (x_{11}, x_{12}, \ldots, x_{1d}), \quad \mathbf{x}_2 = (x_{21}, x_{22}, \ldots, x_{2d}).$$

D. 2. B.B.:

$$B(\mathbf{x}) = B(x_1, x_2, \ldots, x_d) = W(\mathbf{x}) - x_1 x_2, \ldots, x_d W(1, 1, \ldots, 1)$$

$$(0 \leqq x_i \leqq 1;\ i = 1, 2, \ldots, d)$$

where $W(\mathbf{x})$ is a W.P.

D. 3. K.P.:

$$K(\mathbf{x}; y) = K(x_1, x_2, \ldots, x_d; y) = W(x_1, x_2, \ldots, x_d, y) - x_1 x_2 \ldots x_d W(1, 1, \ldots, 1, y)$$

where $W(x_1, x_2, \ldots, x_d, y)$ is a W.P. of $(d+1)$-dimensions.

D. 4. Set

$$I^{(r)}_{i_1, i_2, \ldots, i_d} = \left\{(x_1, x_2, \ldots, x_d) : \frac{i_k}{r} \leqq x_k < \frac{i_k+1}{r},\ k = 1, 2, \ldots, d\right\}$$

where $i_k = 0, 1, 2, \ldots, r-1$; $k = 1, 2, \ldots, d$.

D. 5. For any set $A \subset I^d = [0, 1] \times [0, 1] \times \ldots \times [0, 1]$ and integer r, let:

(i) $$A_r(i) = \sum^{(i)} I^{(r)}_{i_1, i_2, \ldots, i_d}$$

where union $\Sigma^{(i)}$ is taken over all $I^{(r)}_{i_1, i_2, \ldots, i_d} \subset A$,

(ii) $$A_r(0) = \sum^{(0)} I^{(r)}_{i_1, i_2, \ldots, i_d}$$

where union $\sum^{(0)}$ is taken over all $I^{(r)}_{i_1, i_2, \ldots, i_d} \subset \bar{A}$,

(iii) $$A_r(b) = \overline{A_r(i)} \cap \overline{A_r(0)}$$

and $\bar{A} = I^d - A$.

D. 6. Any set $A \subset I^d$ is said to be an element of the class of sets $S^d(C)$ $(C>0)$ if $\lambda(A_r(b)) \leqq Cr^{-1}$ for any integer r, where $\lambda(\cdot)$ is the Lebesgue measure.

D. 7. For any finite set S, let $N(S)$ be the number of elements of S.

D. 8. For any r $(r = 1, 2, \ldots)$, the Wiener measure $W(I^{(r)}_{i_1, i_2, \ldots, i_d})$ is defined the usual (inclusion-exclusion) way, whence $W(A_r(i))$ can be also defined by additivity. For any set $A \in S^d(C)$, we define

$$W(A) = \lim_{n \to \infty} W(A_n(i)).$$

It is quite easy to prove that this limit exists a.s.

D. 9. For any $A \in S^d(C)$, let

$$B(A) = W(A) - \lambda(A) W(1, 1, \ldots, 1).$$

D. 10. For any $A \in S^d(C)$, let $K(A; y) = W(A, y) - \lambda(A)\ W(1, 1, \ldots, 1, y)$, where $W(A, y)$ is defined via D. 8., using the fact that for every fixed y, $y^{-\frac{1}{2}} W(x_1, x_2, \ldots, x_d, y)$ is a d-dimensional W.P.

D. 11. For any $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, set

$$D_{\mathbf{x}} = \{(a_1, a_2, \ldots, a_d) : a_i \leqq x_i\ (i = 1, 2, \ldots, d)\}.$$

D. 12. Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_d)$ be a r.v. with a continuous distribution function

$$F(\mathbf{y}) = F(y_1, y_2, \ldots, y_d) = P(Y_1 < y_1, Y_2 < y_2, \ldots, Y_d < y_d)$$

and set

$$F_1(y_1) = P(Y_1 < y_1),$$

$$F_i(y_i | y_1, y_2, \ldots, y_{i-1}) = P(Y_i < y_i | Y_1 = y_1, Y_2 = y_2, \ldots, Y_{i-1} = y_{i-1})$$

$$(i = 2, 3, \ldots, d).$$

D. 13. Let $F(\mathbf{y})=F(y_1, y_2, \ldots, y_d)$ be a distribution function and define the transformations

$$T_i(F) = T_i = T_i(y_1, y_2, \ldots, y_i) =$$
$$= (F_1(y_1), F_2(y_2|y_1), \ldots, F_i(y_i|y_1, y_2, \ldots, y_{i-1})) \quad (i = 1, 2, \ldots, d).$$

Now we can formulate our main result:

THEOREM 1. *Let* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ *be d-dimensional i.i.d.r.v.'s with a continuous distribution function* $F(\mathbf{y})$. *Assume that* F_i $(i=1, 2, \ldots, d)$ *are continuous, strictly monotone distribution functions for any fixed* $y_1, y_2, \ldots, y_{i-1}$ *(see D.*12.*) and satisfy the following conditions:*

the functions

$$F_i(y_i|T_{i-1}^{-1}(x_1, x_2, \ldots, x_{i-1})) \quad (i = 1, 2, \ldots, d)$$

are differentiable with respect to $x_1, x_2, \ldots, x_{i-1}$ *over the interior of* I^{i-1} *and*

$$\sum_{j=1}^{i-1} \int_{I^{i-1}} \left| \frac{\partial}{\partial x_j} F_i(y_i|T_{i-1}^{-1}(x_1, x_2, \ldots, x_{i-1})) \right| dx_1\, dx_2 \ldots dx_{i-1} \leqq K \tag{1}$$

where K *is a positive constant (for* T_i *see* D. 13.*). Let* $\beta_n(\mathbf{y})=n^{\frac{1}{2}}(F_n(\mathbf{y}) - F(\mathbf{y}))$ *where* $F_n(\mathbf{y})$ *is the e.d.f. based on the sample* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_d$.

Then we can define a sequence $\{B_n(\mathbf{x})\}$ $(\mathbf{x}\in I^d)$ *of B.B.'s and a K.P.* $K(\mathbf{x}; y)$ $(\mathbf{x}\in I^d;\ 0\leqq y<\infty)$ *such that*

$$\sup_{\mathbf{y}\in R^d} |\beta_n(\mathbf{y}) - B_n(T_d D_{\mathbf{y}})| \overset{\text{a.s.}}{=} O(n^{-\frac{1}{2(d+1)}} (\log n)^{\frac{3}{2}}) \tag{2}$$

and

$$\sup_{\mathbf{y}\in R^d} |n^{\frac{1}{2}}\beta_n(\mathbf{y}) - K(T_d D_{\mathbf{y}}; n)| \overset{\text{a.s.}}{=} O(n^{\frac{d+1}{2(d+2)}} \log^2 n). \tag{3}$$

REMARK 1. As a consequence of condition (1), $T_d D_{\mathbf{y}} \in S^d(d(K+2))$, whence $B_n(T_d D_{\mathbf{y}})$ and $K(T_d D_{\mathbf{y}}; n)$ are defined.

This remark is implied by the following elementary

PROPOSITION. *Let* $f(\mathbf{x})$ $(\mathbf{x}\in I^{d-1})$ *be a differentiable function in the interior of* I^{d-1}, *and assume that*

$$\sum_{i=1}^{d-1} \int_{I^{d-1}} \left| \frac{\partial f}{\partial x_i} \right| dx_1\, dx_2 \ldots dx_{d-1} \leqq K$$

for some $K>0$. *Then the graph of* f *intersects at most* $(K+2)r^{d-1}$ *of the* r^d *cubes* $I^{(r)}_{i_1, i_2, \ldots, i_d}$ *of D.* 4.

REMARK 2. It is of interest to formulate Theorem 1 for R^2; then (1) simplifies to

$$\int_0^1 \left| \frac{\partial}{\partial x_1} F_2(y_2|F_1^{-1}(x_1)) \right| dx_2 \leqq K.$$

REMARK 3. Condition (1) can be easily checked for any given distribution function $F(\mathbf{x})$ $(\mathbf{x}\in R^d)$. For example it is satisfied when F is the multi-dimensional normal distribution function with any covariance matrix.

The proof of our Theorem 1 is quite similar to those of Lemmas 11 and 12 of [3]. Therefore we take the freedom of omitting some details of the present proof. In fact, in [3] a fundamental tool is the simple Lemma 1, which says that, in a sense, there are not too many polygons in the unit-square. The analogous lemma in our present paper is the elementary Lemma 2, which says in the same sense that there are not too many sets in the class $\{T_d D_{\mathbf{x}}\colon \mathbf{x}\in R^d\}$. Roughly speaking, the proof of our Theorem 1 can be obtained if we repeat the proofs of Lemmas 11 and 12 of [3], but we apply our present Lemma 2 instead of Lemma 1 of [3].

§ 3. Proof of theorem 1

First, we list a number of results which are going to be used in the sequel.

THEOREM D. ([5]) *For each d and $\varepsilon>0$, there is a constant $c(\varepsilon, d)$ such that, for all d-dimensional distribution functions $F(\cdot)$ and $z\geqq 0$, we have*

$$P\Big\{\sup_{\mathbf{x}\in R^d}|\beta_n(\mathbf{x})|\geqq z\Big\}\leqq c(\varepsilon, d)e^{-(2-\varepsilon)z^2},$$

where $\beta_n(\mathbf{x})$ is the E.P. based on a sample coming from F.

THEOREM E. ([2]) *Let r be an integer and let N_{ij} $(i,j=1,2,\ldots,r)$ be a double array of independent standard normal r.v.'s. Then there exists a W.P.*

$$\{W(x_1, x_2);\ 0\leqq x_1, x_2\leqq 1\}$$

such that

$$W(i/r, j/r)=\frac{1}{r}\sum_{\substack{k\leqq i\\ h\leqq j}} N_{kh}.$$

This is Lemma 6 of [2], and a d-dimensional version of this result is immediate.

THEOREM F. ([2]) *Let $0<y_1<y_2<\ldots$ be a sequence of real numbers and let $\{B_i(\mathbf{x})\}$ $(\mathbf{x}\in I^2)$ be a sequence of independent B.B.'s. Then there exists a K.P. $K(\mathbf{x}; y)$ $(\mathbf{x}\in I^2;\ 0\leqq y<\infty)$ such that*

$$K(\mathbf{x}; y_i)=\sqrt{y_1}B_1(\mathbf{x})+\sqrt{y_2-y_1}B_2(\mathbf{x})+\ldots+\sqrt{y_i-y_{i-1}}B_i(\mathbf{x}).$$

This is Lemma 7 of [2] and a d-dimensional version of this result is immediate.

THEOREM G. ([1]) *Let $p_1, p_2, \ldots$ be independent Poisson r.v.'s with mean 1. Put $\Pi_n=p_1+p_2+\ldots+p_n$. Then there exists a W.P. $\{W(t);\ t\geqq 0\}$ and positive constants C, K and λ such that for every z*

$$P\Big\{\sup_{1\leqq k\leqq n}|(\Pi_k-k)-W(k)|>C\log n+z\Big\}<Ke^{-\lambda z}.$$

Now we give some lemmas:

LEMMA 1. *Let* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ *be as in Theorem* 1 *and let* $p_1, p_2, \ldots$ *be independent Poisson r.v.'s with mean* 1, *which are independent from* $\{\mathbf{Y}_i\}$. *Then for* $c>2$

$$P\left\{\sup_{\mathbf{x}\in R^d} |\beta_n(\mathbf{x})-\beta_{\Pi_n}(\mathbf{x})| \geqq \frac{c(\log n)^{3/4}}{n^{1/4}}\right\} = O\left(n^{-\frac{c-2}{4}}\right) \tag{5}$$

where $\Pi_n=p_1+p_2+\ldots+p_n$.

PROOF. Let $A(n)$ be the event of the probability statement of (5) and put $a_n= =n-c(n\log n)^{1/2}$, $b_n=n+c(n\log n)^{1/2}$. Then

$$\begin{aligned} P(A(n)) &\leqq P\{|\Pi_n-n| > c(n\log n)^{1/2}\}+P\{A(n), |\Pi_n-n| \leqq c(n\log n)^{1/2}\} \\ &\leqq O(n^{-c^2/2})+P\left\{\sup_{a_n\leqq k\leqq b_n}\sup_{\mathbf{x}\in R^d} |\beta_n(\mathbf{x})-\beta_k(\mathbf{x})| > \frac{c(\log n)^{3/4}}{n^{1/4}}\right\} \\ &\leqq O(n^{-c^2/2})+2c(n\log n)^{1/2}P\left\{\sup_{\mathbf{x}\in R^d} |\beta_n(\mathbf{x})-\beta_{b_n}(\mathbf{x})| > \frac{c(\log n)^{3/4}}{n^{1/4}}\right\}. \end{aligned} \tag{6}$$

Since

$$\begin{aligned} \beta_{b_n}(\mathbf{x})-\beta_n(\mathbf{x}) &= c^{1/2}\left(\frac{\log n}{n}\right)^{1/4}(b_n-n)^{1/2}(F_{n,b_n}(\mathbf{x})-F(\mathbf{x}))+ \\ &+O\left(\left(\frac{\log n}{n}\right)^{1/2}\right) b_n^{1/2}(F_{b_n}(\mathbf{x})-F(\mathbf{x})), \end{aligned} \tag{7}$$

where $F_{n,b_n}(\mathbf{x})$ stands for the e.d.f. based on the random sample $\mathbf{Y}_{n+1}, \mathbf{Y}_{n+2}, \ldots, \mathbf{Y}_{b_n}$, and $O(\cdot)$ of (7) is independent of $\mathbf{x}$ and by Theorem D

$$P\left\{c^{1/2}\left(\frac{\log n}{n}\right)^{1/4}(b_n-n)^{1/2}\sup_{\mathbf{x}}|F_{n,b_n}(\mathbf{x})-F(\mathbf{x})| > \frac{c}{2}\frac{(\log n)^{3/4}}{n^{1/4}}\right\} = O\left(n^{-\frac{c}{4}(2-\varepsilon)}\right), \tag{8}$$

$$P\left\{O\left(\left(\frac{\log n}{n}\right)^{1/2}\right) b_n^{1/2}\sup_{\mathbf{x}}|F_{b_n}(\mathbf{x})-F(\mathbf{x})| > \frac{c}{2}\frac{(\log n)^{3/4}}{n^{1/4}}\right\} = O\left(n^{-\frac{c}{4}(2-\varepsilon)}\right). \tag{9}$$

Lemma 1 follows from (6), (7), (8) and (9).

The next lemma follows from a simple combinatorial argument, and its proof will be omitted.

LEMMA 2. *The number of the elements of the class of the sets* $(T_d D_{\mathbf{x}})_r(i)$ $(\mathbf{x}\in R^d)$ *is less than or equal to* $r^{\frac{d(d+1)}{2}}$, *i.e.*

$$N\{T_d D_{\mathbf{x}})_r(i) : \mathbf{x}\in R^d\} \leqq r^{\frac{d(d+1)}{2}}$$

for any fixed distribution function $F(\mathbf{x})$ $(\mathbf{x}\in R^d)$.

Using the notations of Lemma 1 and Theorem G, we get from Theorem G immediately:

LEMMA 3. *Let* $r=r_n=o(n^{1/d})$ *be a sequence of integers. Then for every* n $(n=1, 2, \ldots)$ *one can construct a set of independent standard normal r.v.'s* $N^{(n)}_{i_1, i_2, \ldots, i_d}$ $(i_j=0, 1, 2, \ldots, r-1;\ j=1, 2, \ldots, d)$ *such that*

$$P\left\{|r^{d/2}\tilde{\beta}_{\Pi_n}(T_d^{-1}I^{(r)}_{i_1, i_2, \ldots, i_d})-N^{(n)}_{i_1, i_2, \ldots, i_d}| \geqq r^{d/2}\frac{c\log n+z}{\sqrt{n}}\right\} \leqq Ke^{-\lambda z}$$

where

$$\tilde{\beta}_{\Pi_n}(A)=\sqrt{n}\left(\frac{H_{\Pi_n}(A)}{n}-F(A)\right)$$

(A is a Borel set of R^d*) and* $H_{\Pi_n}(A)$ *is the number of the elements of the sample* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{\Pi_n}$ *lying in the set* $A\subset R^d$.

Let $W_n(\mathbf{x})=W_n(x_1, x_2, \ldots, x_d)$ be a W.P. for which

$$W_n\left(\frac{i_1}{r}, \frac{i_2}{r}, \ldots, \frac{i_d}{r}\right)=r^{-d/2}\sum_{j_l\leqq i_l} N^{(n)}_{j_1, j_2, \ldots, j_d} \quad (l=1, 2, \ldots, d) \tag{10}$$

(by Theorem E there exists such a W.P.).

Let $T=T_d(F)$, and for any $\mathbf{x}\in R^d$ consider the following sets $D_{\mathbf{x}}$, $TD_{\mathbf{x}}$, $(TD_{\mathbf{x}})_r(i)$, $T^{-1}((TD_{\mathbf{x}})_r(i))$. Studying the differences between the measures of these sets one can say:

LEMMA 4. *For any* $\mu>0$ *there exists a* $C>0$ *depending only on* $F(\cdot)$ *such that*

$$P\left\{\sup_{\mathbf{x}}\left|W_n((TD_{\mathbf{x}})_r(i))-\tilde{\beta}_{\Pi_n}\left(T^{-1}((TD_{\mathbf{x}})_r(i))\right)\right| \geqq \frac{C(\log n)^{3/2}r^{d/2}}{\sqrt{n}}\right\} \leqq n^{-\mu}, \tag{11}$$

$$P\left\{\sup_{\mathbf{x}}\left|\tilde{\beta}_{\Pi_n}(D_{\mathbf{x}})-\tilde{\beta}_{\Pi_n}\left(T^{-1}((TD_{\mathbf{x}})_r(i))\right)\right| \geqq Cr^{-1/2}\log n\right\} \leqq n^{-\mu}, \tag{12}$$

$$P\left\{\sup_{\mathbf{x}}\left|W_n(TD_{\mathbf{x}})-W_n((TD_{\mathbf{x}})_r(i))\right| \geqq Cr^{-1/2}\log n\right\} \leqq n^{-\mu}. \tag{13}$$

PROOF. (11) follows from Lemmas 2 and 3. (For details see the method of proof of Lemma 8 of [2]). Our statements (12) and (13) follow from condition (1) and Lemma 2. In fact condition (1) implies that $TD_{\mathbf{x}}\in S^d(d(K+2))$, i.e. $\lambda(TD_{\mathbf{x}}-(TD_{\mathbf{x}})_r(i))=O(r^{-1})$. (For details see the method of proof of Lemma 2 of [3]).

LEMMA 5. *For any* $\mu>0$, *there exists a* $C>0$ *depending only on* $F(\cdot)$ *such that*

$$P\left\{\sup_{\mathbf{y}\in R^d}|\beta_n(\mathbf{x})-B_n(TD_{\mathbf{y}})| \geqq C(\log n)^{3/2}n^{\frac{-1}{2(d+1)}}\right\} \leqq n^{-\mu} \tag{14}$$

where

$$B_n(\mathbf{x})=W_n(\mathbf{x})-x_1x_2\ldots x_dW_n(1, 1, \ldots, 1), \quad \mathbf{x}\in I^d,$$

and $W_n(\cdot)$ *is defined by* (10).

PROOF. Applying (11) for $D_{\mathbf{y}}=R^d$, we get

$$P\left\{|W_n(I^d)-\tilde{\beta}_{\Pi_n}(R^d)| \geqq \frac{C(\log n)^{3/2}r^{d/2}}{\sqrt{n}}\right\} \leqq n^{-\mu}, \tag{15}$$

also we clearly have

$$\left(\tilde{\beta}_{\Pi_n}(A)-\frac{\Pi_n-n}{\sqrt{n}}F(A)\right)-\beta_{\Pi_n}(A)=\beta_{\Pi_n}(A)\frac{\sqrt{\Pi_n}-\sqrt{n}}{\sqrt{n}}. \tag{16}$$

Now choosing $r=n^{\frac{1}{d+1}}$, we get (14) as a consequence of Lemmas 1, 4 and (15) and (16).

The inequality (14) clearly implies statement (2) of our Theorem 1.

PROOF OF (3). Let $n_k=k^{d+2}$ and $m_k=n_k-n_{k-1}$. Denote by $\hat{\beta}_k(\mathbf{y})$ the E.P. based on the sample $\mathbf{Y}_{n_{k-1}+1}, \mathbf{Y}_{n_{k-1}+2}, \ldots, \mathbf{Y}_{n_k}$. Further let $\hat{B}_k(\mathbf{x})$ be a B.B. for which

$$P\left\{\sup_{\mathbf{y}}|\hat{\beta}_k(\mathbf{y})-\hat{B}_k(TD_{\mathbf{y}})|\geqq C(\log n)^{3/2}n^{\frac{-1}{2(d+1)}}\right\}\leqq n^{-\mu} \tag{17}$$

(by Lemma 5 there exists such a B.B.).

By Theorem F there exists a K.P. such that

$$K(\mathbf{x};n_k)=\sqrt{n_1}\hat{B}_1(\mathbf{x})+\sqrt{m_2}\hat{B}_2(\mathbf{x})+\ldots+\sqrt{m_k}\hat{B}_k(\mathbf{x}).$$

Since $\beta_{n_k}(\mathbf{y})=\sqrt{n_1}\hat{\beta}_1(\mathbf{y})+\sqrt{m_2}\hat{\beta}_2(\mathbf{y})+\ldots+\sqrt{m_k}\hat{\beta}_k(\mathbf{y})$, (17) easily implies (3). (For details see the proof of Lemma 12 of [3]).

REFERENCES

[1] KOMLÓS, J., MAJOR, P., TUSNÁDY, G.: An Approximation of Partial Sums of independent R. V.'s and the Sample D. F. I., *Zeitschrift für Wahrscheinlichkeitstheorie* **32** (1975), 111—131.

[2] CSÖRGŐ, M., RÉVÉSZ, P.: A New Method to Prove Strassen Type Laws of Invariance Principle II, *Zeitschrift für Wahrscheinlichkeitstheorie* **31** (1975), 261—269.

[3] RÉVÉSZ, P.: On Strong Approximation of the Multidimensional Empirical Process, *The Annals of Probability* **4** (1976), 729—743.

[4] DUDLEY, R. M.: Weak Convergence of Probabilities on Nonseparable Metric Spaces and Empirical Measures on Euclidean Spaces, *Illinois Journal of Mathematics* **10** (1966), 109—126.

[5] KIEFER, J.: On large deviations of the empiric d. f. of vector chance variables and a law of the iterated logarithm, *Pacific Journal of Mathematics* **11** (1961), 649—660.

Mathematical Institute of the Hungarian Academy of Sciences, 1053 Budapest, Reáltanoda u. 13—15, Hungary

(Received November 21, 1975)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 435—458.

UNTERSUCHUNGEN ÜBER RICHTUNGSSTRUKTUREN, I. WEITERE BEZIEHUNGEN DER RICHTUNGSDIMENSION ZU DEN KLASSISCHEN DIMENSIONEN* FÜR GEWISSE KLASSEN TOPOLOGISCHER RÄUME

von

E. DEÁK

Einleitung

Die „Richtungsdimension" Dim X eines Raumes X, um die es sich in diesem Artikel handelt, ist ein in [3] und [5] eingeführter neuer topologischer Dimensionsbegriff. (Die bequemste Übersicht (und auch weitere Literaturangaben) über einen — auch die Theorie der Richtungsdimension umfassenden — Teil der seit damals entwickelten Theorie der sog. Richtungsstrukturen bietet der Bericht ohne Beweise [8].) Das Hauptergebnis von [5] lautet so: *ein separabler metrisierbarer Raum X ist dann und nur dann in den n-dimensionalen euklidischen Raum topologisch einbettbar, wenn* $\operatorname{Dim} X \leqq n$ *gilt.* (Ein anderer Beweis dieses Satzes ist — bisher nur in ungarischer Sprache — in [7] erschienen.)

Es folgt daraus (man beachte, daß im Bereich der separablen metrisierbaren Räume die drei klassischen Dimensionen miteinander äquivalent sind), daß für jeden separablen metrisierbaren Raum X, dessen Dimension endlich (im Sinne irgendeiner der hier behandelten vier Dimensionsbegriffe) ist,

$$\left.\begin{matrix}\operatorname{ind} X\\ \operatorname{Ind} X\\ \dim X\end{matrix}\right\} \leqq \operatorname{Dim} X,$$

und es erhob sich ganz natürlich die Frage, ob noch weitere (d. h. sich auf andere Klassen von Räumen beziehende) ähnliche Abschätzungen nach oben einer klassi-

---

* *Bezeichnungen und Wortgebrauch*

Unter „Raum" wird durchwegs ein nichtleerer allgemeiner topologischer Raum verstanden. Die „klassischen Dimensionen" eines Raumes X sind: die (Menger—Urysohnsche) kleine induktive Dimension $\operatorname{ind} X$, die (Čechsche) große induktive Dimension $\operatorname{Ind} X$ und die (Lebesguesche) Überdeckungsdimension $\dim X$. $|A|$ ist die Mächtigkeit einer Menge A. Für eine Teilmenge A eines Raumes B bedeutet $A^{-(B)}$ (evtl. einfach A^-) bzw. $\operatorname{Gr}_B A$ (evtl. einfach $\operatorname{Gr} A$) die abgeschlossene Hülle bzw. die Begrenzung von A in B. Für zwei Überdeckungen $\mathcal{U}$ und $\mathcal{V}$ bedeutet $\mathcal{U} \prec \mathcal{V}$, daß $\mathcal{U}$ eine Verfeinerung von $\mathcal{V}$ ist. Eine Basis bzw. Subbasis eines Raumes bedeutet eine offene Basis bzw. Subbasis. „Kompakt" wird im Sinne „bikompakt" gebraucht; „regulär", „normal", „kompakt", parakompakt" u. ä. m. sind ohne Voraussetzung auch nur des T_0-Axioms gemeint. Das Zeichen $\subset$ wird im Sinne „echte Teilmenge" gebraucht.

Was die Literaturhinweise (Zahlen in eckigen Klammern) betrifft: ist die Zahl aus Antiqua gesetzt, so verweist das auf irgendeine leicht zugängliche Quelle; kursivierte Zahlen deuten die Urquelle (wenn uns diese überhaupt bekannt ist) an.

schen Dimension eines Raumes durch die Richtungsdimension desselben zu ermitteln sind.

Nach einem in [4] (wo übrigens auch einige wesentliche Berichtigungen zu [3] gegeben wurden) nur andeutungsweise bewiesenen Satz dieser Art haben wir dann in [6] tatsächlich nicht nur den Beweis dieses Satzes nachgeholt, sondern eine ganze Reihe solcher Sätze bewiesen. (Einige von ihnen werden wir im § 0 dieser Arbeit anführen.)

Mit der vorliegenden Arbeit haben wir das Ziel, diese Untersuchungen weiterzuführen. Unsere hiesigen Ergebnisse sind sämtlich Verallgemeinerungen oder Verschärfungen (oder beides auf einmal) gewisser Sätze aus [6] (darauf soll auch der Titel dieses Artikels hinweisen).

Unsere Hauptergebnisse finden sich im § 2; die zu ihrem Beweis nötigen Hilfssätze hat § 1 zum Gegenstand; im § 3 werden einige Spezialfälle und Konsequenzen der Fundamentalsätze aus § 2 — zum Teil mit eigenen Beweisen — hervorgehoben.

Zunächst seien — im § 0 — die anzuwendenden Begriffe, Sätze und andere Hilfsmittel aus der Theorie der Richtungsstrukturen (ohne Beweise) angeführt.

§ 0. Zusammenstellung der nötigen Hilfsmittel aus der Theorie der Richtungsstrukturen*

(0.1) DEFINITION. Es sei X eine nichtleere Menge.

(a) Eine *Richtung der Menge X* ist ein System $\mathscr{R}$ von geordneten Paaren (G, F) mit $G \subseteq F \subseteq X$ und mit

(I) $(\varnothing, \varnothing), (X, X) \in \mathscr{R}$,

(II) $(G_1, F_1), (G_2, F_2) \in \mathscr{R}, \quad (G_1, F_1) \neq (G_2, F_2 \Rightarrow F_1 \subseteq G_2 \vee F_2 \subseteq G_1$

und weiter

(III) $\bigcup \{G: G \in \mathscr{G}^*\} \in \mathscr{G}(\mathscr{R}) \quad (\varnothing \neq \mathscr{G}^* \subseteq \mathscr{G}(\mathscr{R}))$,
(IV) $\bigcap \{F: F \in \mathscr{F}^*\} \in \mathscr{F}\{\mathscr{R}) \quad (\varnothing \neq \mathscr{F}^* \subseteq \mathscr{F}(\mathscr{R}))$,

wo $\mathscr{G}(\mathscr{R})$ bzw. $\mathscr{F}(\mathscr{R})$ die Familie der ersten bzw. zweiten Komponenten der Elemente von $\mathscr{R}$ bedeutet.

(b) Die Elemente von $\mathscr{G}(\mathscr{R})$ und die Komplemente der Mengen aus $\mathscr{F}(\mathscr{R})$ bzw. die Elemente von $\mathscr{F}(\mathscr{R})$ sowie die Komplemente der Mengen aus $\mathscr{G}(\mathscr{R})$ werden die *$\mathscr{R}$-offenen* bzw. *$\mathscr{R}$-abgeschlossenen Halbmengen* von X genannt.

(0.2) BEZEICHNUNGEN. (a) Die Beziehung $(G, F) \in \mathscr{R}$ wird — bei festgesetztem $\mathscr{R}$ — auch durch die zwar nicht immer eindeutigen (weil es nämlich verschiedene Paare $(G_1, F_1), (G_2, F_2) \in \mathscr{R}$ mit $G_1 = G_2$ oder $F_1 = F_2$ geben kann) aber immer höchstens zweideutigen Symbole $G = G(F)$ oder $F = F(G)$ ausgedrückt.

(b) Mit $\underline{G}(F)$ bzw. $\overline{G}(F)$ wird die kleinere bzw. größere Menge $G(F)$ und mit $\underline{F}(G)$ bzw. $\overline{F}(G)$ die kleinere bzw. größere Menge $F(G)$ bezeichnet.

---

* Die hiesige Terminologie stimmt nicht ganz mit jener von [5] überein (wohl aber mit jener des Artikels [8], der allerdings englisch abgefaßt ist); das hat aber bloß bezeichnungs- und beweistechnische Gründe und ist sonst belanglos.

(c) Soll auch auf die Richtung $\mathscr{R}$ hingewiesen werden (vgl. die Bemerkung weiter unten (0.4),(a)), so benutzen wir die entsprechenden Symbole

$$\underline{G}(\mathscr{R}; F), \quad \overline{G}(\mathscr{R}; F), \quad \underline{F}(\mathscr{R}; G), \quad \overline{F}(\mathscr{R}; G).$$

(0.3) DEFINITION. Ist eine Menge X mit einer Richtung $\mathscr{R}$ versehen, so werden die Mengen

$$F \diagdown G \quad ((G, F) \in \mathscr{R},\ G \neq F)$$

die *$\mathscr{R}$-Ebenen* von X genannt; ihre Familie wird mit $\mathscr{S}(\mathscr{R})$ bezeichnet.

(0.4) BEMERKUNGEN. (a) Verschiedene Richtungen ein und derselben Menge können gemeinsame Elemente haben; d. h. wenn $\mathscr{R}_1$ und $\mathscr{R}_2$ beide Richtungen einer Menge X sind und $\mathscr{R}_1 \neq \mathscr{R}_2$ ist, so kann es doch Teilmengen G und F von X mit $(G, F) \in \mathscr{R}_1$ und gleichzeitig $(G, F) \in \mathscr{R}_2$ geben.

(b) Ist $\mathscr{R}$ eine Richtung einer Menge X, so gilt

$$S_1, S_2 \in \mathscr{S}(\mathscr{R}), \quad S_1 \neq S_2 \Rightarrow S_1 \cap S_2 = \varnothing.$$

(0.5) DEFINITION. (a) Eine *Richtungsstruktur* (RS) einer Menge X ist eine nichtleere Familie $\mathfrak{R}$ von Richtungen $\mathscr{R}$ dieser Menge.

(b) Wir gebrauchen die Bezeichnungen

$$\mathscr{G}(\mathfrak{R}) = \bigcup \{\mathscr{G}(\mathscr{R}) \colon \mathscr{R} \in \mathfrak{R}\}, \quad \mathscr{F}(\mathfrak{R}) = \bigcup \{\mathscr{F}(\mathscr{R}) \colon \mathscr{R} \in \mathfrak{R}\}.$$

(c) Für eine RS $\mathfrak{R}$ einer Menge X nennen wir alle $\mathscr{R}$-Ebenen ($\mathscr{R} \in \mathfrak{R}$) auch $\mathfrak{R}$-Ebenen von X; es wird das Symbol

$$\mathscr{S}(\mathfrak{R}) = \bigcup \{\mathscr{S}(\mathscr{R}) \colon \mathscr{R} \in \mathfrak{R}\}$$

gebraucht.

(d) Eine *$\mathfrak{R}$-offene* bzw. *$\mathfrak{R}$-abgeschlossene Halbmenge einer Menge X*, $\mathfrak{R}$ eine RS von X, ist eine $\mathscr{R}$-offene bzw. $\mathscr{R}$-abgeschlossene Halbmenge dieser Menge ($\mathscr{R} \in \mathfrak{R}$).

(0.6) Die *Spur* $\mathscr{R}|X^*$ bzw. $\mathfrak{R}|X^*$ *einer Richtung* $\mathscr{R}$ bzw. *einer RS* $\mathfrak{R}$ einer Menge X *auf einer Teilmenge* X^* ist das System

$$\{(G \cap X^*, F \cap X^*) \colon (G, F) \in \mathscr{R}\}$$

bzw.

$$\{\mathscr{R}|X^* \colon \mathscr{R} \in \mathfrak{R}\}.$$

(0.7) BEMERKUNGEN. (a) Die Definition unter (0.5),(a) schließt den Fall nicht aus, daß eine RS einer Menge mehrere — etwa durch irgendwelche Parameter oder Indizes voneinander abweichende, also nur in der Bezeichnung verschiedene — Exemplare ein und derselben Richtung enthält.

(b) Andererseits kann es vorkommen, daß die Spuren $\mathscr{R}_1|X^*$, $\mathscr{R}_2|X^*$ (wo $\mathscr{R}_1$ und $\mathscr{R}_2$ Richtungen einer Menge X sind) für ein X^* mit $\varnothing \neq X^* \subsetneqq X$ zusammenfallen, obwohl $\mathscr{R}_1 \neq \mathscr{R}_2$ ist.

(0.8) DEFINITION. *Die trivialen Richtungen* einer beliebigen Menge X sind die beiden Systeme von Paaren

$$\{(\varnothing, \varnothing), (X, X)\}, \quad \{(\varnothing, \varnothing), (\varnothing, X), (X, X)\}.$$

(0.9) DEFINITION. *Eine Richtungsstruktur* (RS) $\mathfrak{R}$ *(eine Richtung* $\mathscr{R}$*) eines Raumes* X ist eine RS (eine Richtung) der Trägermenge X des Raumes, für welche jede $\mathfrak{R}$-offene bzw. $\mathfrak{R}$-abgeschlossene ($\mathscr{R}$-offene bzw. $\mathscr{R}$-abgeschlossene) Halbmenge offen bzw. abgeschlossen im Sinne der Topologie von X ist — und übrigens in diesem Zusammenhang auch $\mathfrak{R}$-*offener* bzw. $\mathfrak{R}$-*abgeschlossener* ($\mathscr{R}$-*offener* bzw. $\mathscr{R}$-*abgeschlossener*) *Halbraum des Raumes* X genannt wird.

(0.10) DEFINITIONEN. (a) Die durch eine *Richtung* $\mathscr{R}$ bzw. *RS* $\mathfrak{R}$ einer Menge X auf dieser Menge *induzierte Topologie* $\mathscr{T}(\mathscr{R})$ bzw. $\mathscr{T}(\mathfrak{R})$ ist diejenige, für welche die $\mathscr{R}$-offenen bzw. $\mathfrak{R}$-offenen Halbmengen eine Subbasis bilden.

(b) Eine RS $\mathfrak{R}$ eines Raumes X wird *eine kompatible RS* (KRS) *des Raumes* genannt, wenn $\mathscr{T}(\mathfrak{R})$ mit der Topologie des Raumes übereinstimmt.

(Nach (0.9) ist $\mathscr{T}(\mathfrak{R})$ immer gröber als die Topologie von X.)

(0.11) DEFINITIONEN. (a) Die — mit $\operatorname{Dim} X$ bezeichnete — *Richtungsdimension* (RD) eines Raumes X ist das Minimum der Mächtigkeiten seiner KRS-en, triviale Richtungen nicht miteingerechnet.

(Die letztere Bestimmung hat natürlich nur bei endlicher RD Bedeutung.)

(b) Eine *minimale KRS* (MKRS) *eines Raumes* X ist eine KRS $\mathfrak{R}$ des letzteren mit $|\mathfrak{R}| = \operatorname{Dim} X$.

(c) Wir vereinbaren noch, daß $\operatorname{Dim} \varnothing = 0$ sein soll.

(0.12) BEMERKUNGEN. (a) Jeder Raum X hat KRS-en und somit auch eine RD $\operatorname{Dim} X$; das ist eine — dem Raum eindeutig zugeordnete — Kardinalzahl mit

$$0 \leqq \operatorname{Dim} X \leqq w(X),$$

wo $w(X)$ das topologische Gewicht des Raumes X bedeutet.

(Nebenbei bemerkt: zu jeder beliebigen Kardinalzahl $\mathfrak{n}$ gibt es topologische Räume X mit $\operatorname{Dim} X = \mathfrak{n}$.

(b) Da eine KRS (ja überhaupt jede RS) eines trivialen (indiskreten) Raumes nur aus trivialen Richtungen bestehen kann (und, nebenbei: solche Räume lassen überhaupt keine anderen als kompatible RS-en zu), gilt nach (0.11),(a) für jeden trivialen Raum X notwendig $\operatorname{Dim} X = 0$.

(b′) Ebenso selbstverständlich ist für jeden nichttrivialen Raum X $\operatorname{Dim} X \geqq 1$.

(c) Die Spur einer KRS eines Raumes X auf einer Menge $X^* \subset X$ ist jeweils eine KRS des Teilraumes X^*.

(d) Wegen (c) ist die RD monoton, d. h.

$$\operatorname{Dim} X^* \leqq \operatorname{Dim} X \quad (X^* \subset X).$$

(e) Für jede KRS $\mathfrak{R}$ eines Raumes X gibt es einen Teil $\mathfrak{R}^* \subseteq \mathfrak{R}$ mit paarweise verschiedenen Richtungen (vgl. (0.7), (a)), der gleichfalls eine KRS dieses Raumes ist.

(Diese Bemerkung ist sogar für MKRS-en nicht belanglos; eine *endliche* MKRS kann allerdings selbstverständlich nur aus paarweise verschiedenen Richtungen bestehen.)

(0.13) DEFINITIONEN. (a) Für eine RS $\mathfrak{R}$ einer Menge bzw. eines Raumes X und für $\varnothing \neq \mathfrak{R}^* \subseteq \mathfrak{R}$ mit $|\mathfrak{R}^*| < \aleph_0$ wird jede Menge

$$I = \bigcap \{M \setminus N\colon\ M, N \in \mathscr{G}\{\mathscr{R}) \cup \mathscr{F}\{\mathscr{R}),\ \mathscr{R} \in \mathfrak{R}^*\}$$

(eigentlich sollte $M(\mathscr{R})\setminus N(\mathscr{R})$ mit $M(\mathscr{R})$, $N(\mathscr{R})\in\mathscr{G}(\mathscr{R})\cup\mathscr{F}(\mathscr{R})$ $(\mathscr{R}\in\mathfrak{R}^*)$ geschrieben werden, es ist auch so gemeint) ein $\mathfrak{R}$-*Intervall* von X genannt.

(b) Im Falle

$$M\in\mathscr{G}(\mathscr{R}),\quad N\in\mathscr{F}(\mathscr{R})\quad(\mathscr{R}\in\mathfrak{R}^*)$$

bzw.

$$M\in\mathscr{F}(\mathscr{R}),\quad N\in\mathscr{G}\{\mathscr{R})\quad(\mathscr{R}\in\mathfrak{R}^*)$$

heißt I ein $\mathfrak{R}$-*offenes* bzw. $\mathfrak{R}$-*abgeschlossenes Intervall* von X.

(c) Die Familie aller $\mathfrak{R}$-offener Intervalle wird mit $\mathscr{I}(\mathfrak{R})$ bezeichnet.

(0.14) BEMERKUNGEN. (a) Wenn es sich um eine RS $\mathfrak{R}$ eines *Raumes* X handelt, so ist die Menge I unter (0.13),(a), im Falle unter (0.13),(b) auch bezüglich der Topologie von X eine offene bzw. abgeschlossene Menge.

(b) Jede $\mathfrak{R}$-offene bzw. $\mathfrak{R}$-abgeschlossene Halbmenge, $\mathfrak{R}$ eine RS einer Menge, X, ist ein $\mathfrak{R}$-offenes bzw. $\mathfrak{R}$-abgeschlossenes Intervall.

(c) Jede $\mathfrak{R}$-Ebene ist ein $\mathfrak{R}$-abgeschlossenes Intervall.

(d) Die Kompatibilität einer RS $\mathfrak{R}$ eines Raumes X (vgl. (0.10), (b)) kann auch so definiert werden: $\mathscr{I}(\mathfrak{R})$ ist eine Basis für die Topologie von X.

(0.15) BEMERKUNG. Für ein $\mathfrak{R}$-Intervall I wie unter (0.13) (im topologischen Fall) und mit den Bezeichnungen

$$\mathfrak{R}_1^* = \{\mathscr{R}\in\mathfrak{R}^*:\ M\in\mathscr{G}(\mathscr{R})\},$$
$$\mathfrak{R}_2^* = \{\mathscr{R}\in\mathfrak{R}^*:\ M\in\mathscr{F}(\mathscr{R})\},$$
$$\mathfrak{R}_3^* = \{\mathscr{R}\in\mathfrak{R}^*:\ N\in\mathscr{G}(\mathscr{R})\},$$
$$\mathfrak{R}_4^* = \{\mathscr{R}\in\mathfrak{R}^*:\ N\in\mathscr{F}(\mathscr{R})\}$$

gilt allgemein

$$\begin{aligned}\mathrm{Gr}\,I\subseteq\ & \bigcup\{\underline{F}(\mathscr{R};M)\setminus M:\ \mathscr{R}\in\mathfrak{R}_1^*\}\cup\\ & \cup\bigcup\{M\setminus\bar{G}(\mathscr{R};M):\ \mathscr{R}\in\mathfrak{R}_2^*\}\cup\\ & \cup\bigcup\{\underline{F}(\mathscr{R};N)\setminus N:\ \mathscr{R}\in\mathfrak{R}_3^*\}\cup\\ & \cup\bigcup\{N\setminus\bar{G}(\mathscr{R};N):\ \mathscr{R}\in\mathfrak{R}_4^*\};\end{aligned}$$

im Falle

$$M\in\mathscr{G}(\mathscr{R}),\quad F\in\mathscr{F}(\mathscr{R})\quad(\mathscr{R}\in\mathfrak{R}^*)$$

hat man einfach

$$\mathrm{Gr}\,I\subseteq\bigcup\{(\underline{F}(\mathscr{R};M)\setminus M)\cup(N\setminus\bar{G}(\mathscr{R};N)):\ \mathscr{R}\in\mathfrak{R}^*\}.$$

(0.16) SATZ. *Es seien* $\mathfrak{R}$ *eine KRS eines Raumes* X *mit paarweise verschiedenen Richtungen,* $\emptyset\neq\mathfrak{R}^*\subseteq\mathfrak{R}$ *und*

(0.16.1) $$S(\mathscr{R})\in\mathscr{S}(\mathscr{R})\quad(\mathscr{R}\in\mathfrak{R}^*).$$

(a) *Es gilt dann*

$$\mathrm{Dim}\bigcap\{S(\mathscr{R}):\ \mathscr{R}\in\mathfrak{R}^*\}\leqq|\mathfrak{R}|-|\mathfrak{R}^*|.$$

([3] 36; berichtigt: [4] 303.)

(b) *Mit der Bezeichnung unter* (0.16.1) *ist* $\bigcap\{S(\mathscr{R}):\ \mathscr{R}\in\mathfrak{R}\}$ *als Teilraum von* X *in jedem Fall indiskret.*

(Man stelle das (a) und (0.12),(b) gegenüber.)

(0.17) SUMMENSATZ. *Es seien B eine nichtleere Indexmenge und*

$$X = \sum\{X_\beta\colon \beta \in B\}$$

die topologische Summe von Räumen X_β.

(a) *Ist jedes X_β ein trivialer Raum, so gilt*

$$\operatorname{Dim} X \leqq 1.$$

(b) *Gibt es ein $\beta_0 \in B$ so, daß X_{β_0} nicht-trivial ist, so haben wir*

$$\operatorname{Dim} X = \sup\{\operatorname{Dim} X_\beta\colon \beta \in B\}.$$

(0.18) BEMERKUNG. Es sind nur ganz schwache Sätze über die RD von Vereinigungen von Teilräumen bekannt, und das hat wohl den Grund, daß diese Sache tatsächlich schlecht steht: als besonders krasses (zugleich aber sehr einfaches) Beispiel sei erwähnt, daß die RD eines Zahlenintervalls (und überhaupt der Zahlengerade) gleich 1, die RD der Kreislinie (die ja topologisch aufgefaßt die Vereinigung von zwei abgeschlossenen Zahlen-intervallen ist) jedoch 2 beträgt (s. den in der Einleitung hervorgehobenen Einbettungssatz).

Es folgt eine Auswahl aus den bisher gewonnenen Ergebnissen, die sich auf das Verhalten der RD zu irgendeiner klassischen Dimension beziehen. Es werden hauptsächlich solche Sätze angeführt, mit denen die Ergebnisse der vorliegenden Arbeit in einen leicht erkennbaren Zusammenhang gebracht werden können.

(0.19) SATZ. *Für einen Raum X mit* $\operatorname{Dim} X = 1$ *gilt* $\operatorname{ind} X \leqq 1$ ([6] 251).

(0.20) SATZ. *Für einen Raum X mit* $\operatorname{Dim} X = 1$ *gilt* $\dim X \leqq 1$ ([6] 256).

(0.21) SATZ. *Für einen Raum X mit* $\operatorname{Dim} X = 1$ *gilt* $\operatorname{Ind} X \leqq 1$ ([6] 256).

(0.22) BEMERKUNG. Dem Fall $\operatorname{Dim} X = 1$ kommt — abgesehen von den vorhergehenden drei Sätzen — eine besondere Bedeutung zu, weil er bei je einer vollständigen Charakterisierung der Klasse der zusammenhängenden ordnungsfähigen (ordnungstopologischen) und der unterordnungsfähigen (d. h. in einen ordnungsfähigen Raum topologisch einbettbaren) Räume die Hauptrolle spielt:

(a) *Ein T_1-Raum X mit $|X| > 1$ ist genau dann unterordnungsfähig, wenn* $\operatorname{Dim} X = 1$ gilt.

(b) *Ein zusammenhängender T_1-Raum X mit $|X| > 1$ ist genau dann ordnungsfähig, wenn* $\operatorname{Dim} X = 1$ *gilt.*

(Für weiteres darüber s. [8] 188—192.)

(0.23) BEZEICHNUNG. Es seien $0^* = 0$ und

$$n^* = n[(n-1)^* + 1] \quad (n = 1, 2, \ldots).$$

($\{n^*\colon 1, 2, \ldots\}$ ist also eine (streng) monoton wachsende Folge natürlicher Zahlen.)

(0.24) SATZ. *Für einen vollständig normalen T_2-Raum X endlicher RD gilt*

$$\operatorname{ind} X \leqq (\operatorname{Dim} X)^*.$$

(0.25) SATZ. *Für einen vollständig normalen, kompakten T_2-Raum X endlicher RD gilt*

$$\operatorname{Ind} X \leqq (\operatorname{Dim} X)^*.$$

(0.26) SATZ. *Für einen lokalkompakten metrisierbaren Raum X endlicher RD gilt*

(0.26.1) $$\dim X \leqq \operatorname{Dim} X.$$

(0.27) SATZ. *Für einen vollkommen normalen Lindelöfschen T_2-Raum endlicher RD gilt*

$$\dim X \leqq \operatorname{Dim} X.$$

(0.28) BEMERKUNG. Man könnte die Sätze (0.26) und (0.27) — die in dieser Fassung in [6] bewiesen wurden — anstelle der Voraussetzung „Dim X ist endlich" auch mit der Voraussetzung „dim X ist endlich" formulieren.

Ist nämlich, bei $\dim X < \infty$, $\operatorname{Dim} X < \aleph_0$, so gilt unter den Bedingungen des Satzes (0.26) bzw. (0.27) die Ungleichung (0.26.1), die für $\operatorname{Dim} X \geqq \aleph_0$ zu einer Trivialität wird.

§ 1. Hilfssätze

(1.1) DEFINITION. Ein Raum wird *total-parakompakt* genannt, wenn sich aus jeder Basis des Raumes eine lokal-endliche Überdeckung desselben auswählen läßt.

(1.2) BEMERKUNGEN. (a) R. M. FORD bewies in [11] den folgenden Satz: *Für jeden total-parakompakten metrisierbaren Raum X mit* $\operatorname{ind} X < \infty$ *gilt*

$$\operatorname{Ind} X = \operatorname{ind} X.$$

(Das war der Anlaß zur Einführung des Begriffs der Total-Parakompaktheit.)

(b) Mit der unter (1.1) angeführten Begriffsbildung im Jahre 1963 — eine eigentlich sehr naheliegende Verallgemeinerung der Kompaktheit und zugleich Spezialisierung der Parakompaktheit — nahmen die auf das Verhältnis der beiden klassischen induktiven Dimensionen gerichteten Forschungen einen neuen, bedeutenden Aufschwung. Dem Satz unter (a) folgten ziemlich viele weitere Sätze, die für gewisse Klassen von Räumen die Äquivalenz dieser beiden Dimensionen aussagen; der Begriff der Total-Parakompaktheit selbst hat seither viele Verallgemeinerungen und Abänderungen erfahren.

(c) Es ist hier nicht der Ort, über all dies zu berichten. Wir bemerken aber, daß jeder Satz vom Typ jenes unter (a) auch für den Gegenstand der vorliegenden Arbeit von Bedeutung ist, weil er nämlich mit unserem weiter unten folgenden Satz (2.1) kombiniert werden kann.

(1.3) HILFSSATZ. *Es seien X ein total-parakompakter Raum, $\emptyset \neq H \subseteqq X$ eine abgeschlossene Menge, $\mathscr{W}$ ein System offener Mengen mit*

$$H \subseteqq \bigcup \{W : W \in \mathscr{W}\}$$

und $\mathscr{B}$ eine beliebige Basis der Topologie des Raumes X.

Dann gibt es ein lokal-endliches System

$$\mathscr{B}^* \subseteq \mathscr{B} \tag{1.3.1}$$

mit

$$\mathscr{B}^* < \mathscr{W} \tag{1.3.2}$$

und

$$H \subseteq \bigcup\{B\colon B \in \mathscr{B}^*\}. \tag{1.3.3}$$

BEWEIS. Mit der Bezeichnung

$$\mathscr{W}' = \mathscr{W} \cup \{X \setminus H\}$$

ist auch der Teil

$$\mathscr{B}_* = \{B \in \mathscr{B}\colon \exists W \in \mathscr{W}',\ B \subseteq W\}$$

von $\mathscr{B}$ eine Basis für den Raum X. Auf Grund der Total-Parakompaktheit des letzteren gibt es einen lokal-endlichen Teil

$$\mathscr{B}_{**} \subseteq \mathscr{B}_*$$

dieser Basis mit

$$X = \bigcup\{B\colon B \in \mathscr{B}_{**}\},$$

und das um so mehr lokal-endliche System

$$\mathscr{B}^* = \{B \in \mathscr{B}_{**}\colon B \cap H \neq \varnothing\}$$

entspricht den Forderungen (1.3.1), (1.3.2) und (1.3.3). □

(1.4) DEFINITION. Es seien X eine nichtleere Menge, $M \subseteq X$, $\mathfrak{R}$ eine RS der Menge X und $\mathfrak{m}$ eine Kardinalzahl mit $\mathfrak{m} \leqq |\mathfrak{R}|$.

Die Menge M soll eine $(\mathfrak{m}; \mathfrak{R})$-*kleine Teilmenge von* X heißen, wenn es ein System $\mathfrak{R}^* \subseteq \mathfrak{R}$ mit $|\mathfrak{R}^*| \geqq \mathfrak{m}$ und je eine $\mathscr{R}$-Ebene

$$S(\mathscr{R}) \in \mathscr{S}(\mathscr{R}) \quad (\mathscr{R} \in \mathfrak{R}^*)$$

mit

$$M \subseteq \bigcap\{S(\mathscr{R})\colon \mathscr{R} \in \mathfrak{R}^*\} \tag{1.4.1}$$

gibt.

(1.5) BEMERKUNGEN. (a) Für zwei Kardinalzahlen $\mathfrak{m}_1, \mathfrak{m}_2$ mit $\mathfrak{m}_1 < \mathfrak{m}_2$ ist — unter den Bedingungen der vorigen Definition — jede $(\mathfrak{m}_2; \mathfrak{R})$-kleine Menge M auch $(\mathfrak{m}_1; \mathfrak{R})$-klein.

(b) Jede Menge M ist unter allen Umständen $(0; \mathfrak{R})$-klein; man nimmt eben $\mathfrak{R}^* = \varnothing$ unter (1.4.1) (um das zu ermöglichen wurde ja auch für $\mathfrak{R}^*$ unter (1.4) nicht das Wort „Richtungsstruktur" gebraucht — vgl. die Definition (0.5),(a)).

(c) Die Menge $\varnothing$ ist in jedem Fall $(|\mathfrak{R}|; \mathfrak{R})$-klein.

(d) Wir werden diesen „Kleinheitsgrad-Begriff" in der vorliegenden Arbeit zwar nur auf endliche RS-en anwenden, doch hätte es wegen anderwärtiger Anwendungen keinen Sinn, die Definition derart einzuschränken.

(1.6) DEFINITION. Unter den Bedingungen der Definition (1.4) wird die Kardinalzahl

$$\mathfrak{m} = \sup\{\mathfrak{m}'\colon M \text{ ist } (\mathfrak{m}'; \mathfrak{R})\text{-klein}\}$$

als der $\mathfrak{R}$-*Kleinheitsgrad der Menge* M (in der Grundmenge X) bezeichnet. Ist die Menge M für dieses $\mathfrak{m}$ $(\mathfrak{m}; \mathfrak{R})$-klein (was insbesondere für $\mathfrak{m} < \aleph_0$ immer der Fall ist), so kann man sie auch *genau* $(\mathfrak{m}; \mathfrak{R})$-*klein* nennen.

(1.7) BEMERKUNGEN. (a) Der $\mathfrak{R}$-Kleinheitsgrad ist eine monotone Mächtigkeitsfunktion: der $\mathfrak{R}$-Kleinheitsgrad einer Menge $M^* \subset M$ kann nicht kleiner sein als jener von M. (Daß jeder Menge ein $\mathfrak{R}$-Kleinheitsgrad zukommt, folgt aus (1.5), (b).)

(b) Die Menge $\varnothing$ hat unter allen Umständen den $\mathfrak{R}$-Kleinheitsgrad $|\mathfrak{R}|$.

(c) Als $\mathfrak{R}$-Kleinheitsgrad einer nichtleeren Menge $M \subseteqq X$ kann jede der Kardinalzahlen $0, 1, 2, \ldots, |\mathfrak{R}|$ auftreten (es kommt auf die Menge M und die RS $\mathfrak{R}$ an).

(1.8) BEMERKUNG. *Die Begrenzung eines jeden* $\mathfrak{R}$-*Intervalls,* $\mathfrak{R}$ *eine RS eines Raumes* X*, ist die endliche Vereinigung von abgeschlossenen* $(1; \mathfrak{R})$-*kleinen Mengen.*

Das folgt unmittelbar aus (0.15), (0.16),(a) und (1.7),(b). □

(1.9) HILFSSATZ. *Für eine beliebige KRS* $\mathfrak{R}$ *und für*

(a) *jede einelementige Menge* $H = \{x\}$ $(x \in X)$ *eines beliebigen Raumes und*

(b) *jede nichtleere abgeschlossene Menge* H *eines total-parakompakten Raumes* X *besitzt das Umgebungssystem* $\mathfrak{U}$ *von* H *eine solche Basis* $\mathfrak{U}^*$*, daß die Begrenzung einer jeden Menge aus* $\mathfrak{U}^*$ *die Vereinigung einer lokal-endlichen (im Fall* (a) *sogar endlichen) Familie abgeschlossener* $(1; \mathfrak{R})$-*kleiner Mengen ist.*

BEWEIS. 1° Die Behauptung für den Fall (a) folgt aus (1.8) und (0.14), (d).

2° Was den Fall (b) betrifft: Es seien $\mathscr{I}$ eine Basis des Raumes X bestehend aus $\mathfrak{R}$-offenen Intervallen aus $\mathscr{I}(\mathfrak{R})$ (etwa $\mathscr{I}(\mathfrak{R})$ selbst, vgl. (0.14),(b)),

$$\mathscr{I}_U \subseteqq \mathscr{I} \quad (U \in \mathfrak{U})$$

je ein lokal-endliches System und

$$H \subseteqq J_U \subseteqq U \quad (U \in \mathfrak{U}),$$

mit der Bezeichnung

$$J_U = \bigcup \{I\colon I \in \mathscr{I}_U\} \quad (U \in \mathfrak{U})$$

(die Existenz der $\mathscr{I}_U$ ist durch (1.3) gesichert).

Nun haben wir wegen der Lokal-Endlichkeit der Systeme $\mathscr{I}_U$

$$\operatorname{Gr} J_U \subseteqq \bigcup \{\operatorname{Gr} I\colon I \in \mathscr{I}_U\}$$

woraus nach (1.8) die Behauptung folgt:

$$\mathfrak{U}^* = \{J_U\colon U \in \mathfrak{U}\}$$

ist eine Umgebungsbasis von H in X mit den gewünschten Eigenschaften. □

(1.10) HILFSSATZ. *Es seien* X *ein Raum,* $\{Y_\alpha : \alpha \in A\}$ *ein lokal-endliches Mengensystem im Raume* X *und*

$$\{Z_\alpha^\beta\colon \beta \in B_\alpha\} \quad (\alpha \in A)$$

je ein lokal-endliches Mengensystem in X *mit*

$$Z_\alpha^\beta \subseteqq Y_\alpha \quad (\beta \in B_\alpha, \alpha \in A).$$

Dann ist auch das Mengensystem

$$\{Z_\alpha^\beta\colon \beta\in B_\alpha, \alpha\in A\}$$

lokal-endlich. □

Wir übergehen den Beweis dieser elementaren Tatsache.

§ 2. Die Fundamentalsätze

Es handelt sich hier eigentlich um zwei verschiedene Sätze ((2.1),(a) und (2.1),(b)) die aber im wesentlichen mit ein und demselben Verfahren bewiesen werden können. Wir werden also unter (2.1) einen einheitlichen Beweis geben, der nur an einigen Zwischenstellen einer Zweiteilung bedarf.

(2.1) SATZ. *Es sei X ein Raum mit* $\operatorname{Dim} X<\aleph_0$.

(a) *Dann gilt*

$$\operatorname{ind} X\leqq \operatorname{Dim} X. \tag{2.1.1}$$

(b) *Ist X obendrein total-parakompakt, so gilt auch noch*

$$\operatorname{Ind} X\leqq \operatorname{Dim} X. \tag{2.1.2}$$

BEWEIS. 1° Teil (a) bzw. (b) des Satzes kann auf den Teil (a′) bzw. (b′) des folgenden Satzes zurückgeführt werden:

Es seien n eine natürliche Zahl, $\mathfrak{R}=\{\mathscr{R}_1, \mathscr{R}_2, \ldots, \mathscr{R}_n\}$ *eine KRS eines Raumes X (mit notwendigerweise* $\operatorname{Dim} X\leqq n$*) und K die Vereinigung einer lokal-endlichen* Familie $\mathscr{K}$ abgeschlossener* $(1;\mathfrak{R})$*-kleiner Mengen in X.*

(a′) *Es gilt dann*

$$\operatorname{ind} K\leqq n-1.$$

(b′) *Im Falle, daß X total-parakompakt ist, gilt auch noch*

$$\operatorname{Ind} K\leqq n-1.$$

Tatsächlich: nimmt man $n=\operatorname{Dim} X$, so folgt (a) bzw. (b) auf Grund des Hilfssatzes (1.9) aus (a′) bzw. (b′).

2° Nun ist aber auch (a′) bzw. (b′) bloß ein Spezialfall des folgenden Satzes (a″) bzw. (b″) (und man kann diese Verallgemeinerung — bei einem induktiven Beweisverfahren — auch gar nicht umgehen):

Es seien

$$n,\quad 1\leqq k\leqq n \tag{2.1.3}$$

natürliche Zahlen,

$$\mathfrak{R}=\{\mathscr{R}_1, \mathscr{R}_2, \ldots, \mathscr{R}_n\}$$

* Im Falle (*a′*) könnte man anstelle von „$\mathscr{K}$ lokal-endlich" auch „$\mathscr{K}$ endlich" setzen; wegen der lokalen Beschaffenheit der kleinen induktiven Dimension würde das keine Einschränkung des Satzes bewirken. Wir werden uns dennoch — lediglich der Einheitlichkeit wegen — an die obige Formulierung halten. (Dieselbe Bemerkung bezieht sich auch auf die Formulierung des weiter unten folgenden Satzes (*a″*).)

eine KRS eines Raumes X (*mit notwendigerweise* $\operatorname{Dim} X \leqq n$),

$$\mathscr{K} = \{K_\mu : \mu \in M\}$$

eine lokal-endliche Familie von abgeschlossenen, $(k; \mathfrak{R})$*-kleinen Mengen* K_μ *in* X *und*

$$K = \bigcup \{K_\mu : \mu \in M\}.$$

(a″) *Es gilt dann*

$$\operatorname{ind} K \leqq n - k.$$

(Zur Erläuterung dieses Satzes sei zweierlei bemerkt:

(1) Nach (0.4), (a) sind die k $\mathfrak{R}$-Ebenen, deren Durchschnitt ein K_μ laut Definition (1.4) als Teilmenge enthält, nicht notwendig paarweise verschiedene Mengen im Raume X. Der $\mathfrak{R}$-Kleinheitsgrad einer solchen Menge K_μ kann also kleiner als k sein. Wir erwähnen diese Tatsache übrigens nur zur Klärung der Sachlage; für die folgenden Ausführungen ist sie belanglos.

(2) Auf Grund von (1.5),3° gilt

$$\operatorname{Dim} K_\mu \leqq n - k \quad (\mu \in M);$$

dennoch kann

$$\operatorname{Dim} K > n - k$$

sein, obwohl sämtliche K_μ abgeschlossene Mengen des Teilraumes K von X sind (man vergleiche die Bemerkung (0.18)). Das Wesentliche des Inhalts von (a″) ist also die Tatsache, daß sich die kleine induktive Dimension — im Gegensatz zur RD — bei Vereinigungen der besagten Art nicht vergrößert.)

(b″) *Ist unter denselben Bedingungen wie diejenigen von* (a″) X *obendrein ein total-parakompakter Raum, so gilt auch noch*

$$\operatorname{Ind} K \leqq n - k.$$

Tatsächlich geht (a″) bzw. (b″) für $k = 1$ in (a′) bzw. (b′) über. Es kommt also nunmehr darauf an, (a″) und (b″) zu beweisen. Aus beweistechnischen Gründen setzen wir dabei

$$K_{\mu_1} \neq K_{\mu_2} \quad (\mu_1, \mu_2 \in M,\ \mu_1 \neq \mu_2) \tag{2.1.4}$$

voraus.

3° Es ist ein leichtes die Gültigkeit von (a″) und (b″) im Spezialfall $k = n$ nachzuweisen. Nach (0.16),(b) ist nämlich in diesem Fall jedes K_μ $(\mu \in M)$ — wenn es nicht die leere Menge ist — als Teilraum von X und daher auch als Teilraum von K indiskret, und es gelten somit

$$\operatorname{ind} K_\mu \leqq 0, \quad \operatorname{Ind} K_\mu \leqq 0 \quad (\mu \in M). \tag{2.1.5}$$

Wegen

$$K_{\mu_1} \cap K_{\mu_2} = \varnothing \quad (\mu_1, \mu_2 \in M,\ \mu_1 \neq \mu_2)$$

(was unter diesen Umständen aus (2.1.4) folgt) und daher

$$K_\mu = K \setminus \bigcup \{K_{\mu'} : \mu' \in M,\ \mu' \neq \mu\} \quad (\mu \in M)$$

ist nun aber jedes K_μ $(\mu\in M)$ nicht nur eine abgeschlossene, sondern — auf Grund der Lokal-Endlichkeit von $\mathscr{K}$ — auch eine offene Teilmenge des Raumes K, und K ist somit die topologische Summe der Mengen aus $\mathscr{K}$. Wir haben deshalb, auf Grund von (2.1.5),

$$\operatorname{ind} K \leqq 0 \ (= n-n) \quad \text{bzw.} \quad \operatorname{Ind} K \leqq 0 \ (= n-n)$$

w. z. b. w.

4° der soeben erledigte Fall $k=n$ impliziert nach (2.1.3) auch den Fall $n=1$*. Wir nehmen darum

$$n \geqq 2$$

an und führen den Beweis mittels Induktion „von oben nach unten" bezüglich k weiter; d. h. wir beweisen, daß die Gültigkeit des Satzes (a″) bzw. (b″) für ein k mit $2\leqq k\leqq n$ dieselbe für $k-1$ nach sich zieht.

5° Es seien also (a″) und (b″) für ein k mit

$$2 \leqq k \leqq n$$

als schon bewiesen angenommen. Demgemäß seien sämtliche abgeschlossenen Mengen K_μ $(\mu\in M)$ $(k-1;\mathfrak{R})$-klein, d. h.

$$K_\mu \subseteq D_\mu \quad (\mu\in M), \tag{2.1.6}$$

wobei

$$D_\mu = \bigcap \{S_{\varkappa,\mu}^{\nu(\varkappa,\mu)}\colon \varkappa = 1, 2, \ldots, k-1\} \quad (\mu\in M) \tag{2.1.7}$$

ist mit

$$\nu(\varkappa,\mu)\in\{1, 2, \ldots, n\}, \quad S_{\varkappa,\mu}^{\nu(\varkappa,\mu)}\in\mathscr{S}(\mathscr{R}_{\nu(\varkappa,\mu)})$$

$$(\varkappa = 1, 2, \ldots, k-1, \ \mu\in M)$$

und

$$\nu(\varkappa_1,\mu) \neq \nu\{\varkappa_2,\mu)$$

$$(\varkappa_1, \varkappa_2\in\{1, 2, \ldots, k-1\}, \ \varkappa_1\neq\varkappa_2, \ \mu\in M).$$

Es kann noch

$$\left\{\left(\nu(\varkappa,\mu_1), \ S_{\varkappa,\mu_1}^{\nu(\varkappa,\mu_1)}\right)\colon \varkappa = 1, 2, \ldots k-1\right\} \neq \tag{2.1.8}$$

$$\neq \left\{\left(\nu(\varkappa,\mu_2), \ S_{\varkappa,\mu_2}^{\nu(\varkappa,\mu_2)}\right)\colon \varkappa = 1, 2, \ldots, k-1\right\}$$

$$(\mu_1,\mu_2\in M, \ \mu_1\neq\mu_2)$$

angenommen werden. (Wäre nämlich (2.1.8) nicht allgemein der Fall, so könnte man auf Grund der Relation

$$R = \{(\mu_1,\mu_2)\colon \mu_1,\mu_2\in M, \ (2.1.8) \text{ gilt nicht}\},$$

die offensichtlich eine Äquivalenzrelation ist,

$$\tilde{M} = \{R[\mu]\colon \mu\in M\}$$

* Für $\dim X=1$ haben wir (2.1.1) und (2.1.2) ohnehin schon bewiesen, u. zw. auch (2.1.2) ohne jedwelche zusätzliche Voraussetzungen über den Raum X s. ((0.19) und (0.21)).

anstelle von M und dementsprechend

$$\mathscr{K}^{\sim} = \{K_{\tilde{\mu}}\colon\ \tilde{\mu}\in\tilde{M}\}$$

— mit

$$K_{\tilde{\mu}} = \bigcup\{K_{\mu'}\colon\ \mu'\in\tilde{\mu}\}\quad (\tilde{\mu}\in\tilde{M})\ -$$

anstelle von $\mathscr{K}$ einführen. Es wird ja auch $\mathscr{K}$ eine lokal-endliche Familie von abgeschlossenen, $(k-1;\mathfrak{R})$-kleinen Mengen mit

$$K = \bigcup\{K_{\tilde{\mu}}\colon\ \tilde{\mu}\in\tilde{M}\}$$

sein, wo aber obendrein noch die (2.1.8) entsprechende Bedingung

$$\{(\nu(\varkappa, \tilde{\mu}_1),\ S^{\nu(\varkappa,\tilde{\mu}_1)}_{\varkappa,\tilde{\mu}_1})\colon\ \varkappa = 1, 2, \ldots, k-1\} \neq$$

$$\neq \{(\nu(\varkappa, \tilde{\mu}_2),\ S^{\nu(\varkappa,\tilde{\mu}_2)}_{\varkappa,\tilde{\mu}_2})\colon\ \varkappa = 1, 2, \ldots, k-1\}$$

$$(\tilde{\mu}_1, \tilde{\mu}_2\in\tilde{M},\ \tilde{\mu}_1 \neq \tilde{\mu}_2)$$

sichlerlich erfüllt ist.)

Weiter seien H eine beliebige einelementige (für den Fall (a″)) bzw. nichtleere abgeschlossene (für den Fall (b″)) Menge im Teilraum K von X und U eine Umgebung von H in K d. h.

$$U = K\cap U',$$

wo U' eine Umgebung von H im Raume X ist.

6° Jedem Punkt $x\in H$ kann eine Umgebung V_x von x in X mit

$$V_x \subseteqq U'\quad (x\in H)$$

derart zugeordnet werden, daß V_x mit hochstens endlich vielen Elementen der Familie $\mathscr{K}$ Punkte gemeinsam hat. Dann ist auch

$$W_x = V_x\setminus\bigcup\{K_\mu\colon\ \mu\in M,\ x\notin K_\mu\} \tag{2.1.9}$$

eine Umgebung von x in X $(x\in H)$ — weil ja der Subtrahend eine abgeschlossene Menge ist — und wir haben

$$W_x\cap K \subseteqq U\quad (x\in H). \tag{2.1.10}$$

Da das System $\mathscr{K}$ — wegen seiner Lokal-Endlichkeit — punktal-endlich ist, muß jedes

$$M(x) = \{\mu\in M\colon\ x\in K_\mu\}\quad (x\in H) \tag{2.1.11}$$

eine *endliche* Telmenge von M sein, und es gilt nach (2.1.9)

$$M(x) = \{\mu\in M\colon\ W_x\cap K_\mu \neq \varnothing\}\quad (x\in H). \tag{2.1.12}$$

7° Es seien nun $x\in H$ festgesetzt und

$$I\in\mathscr{I}(\mathfrak{R})$$

mit

$$I \subseteqq W_x, \tag{2.1.13}$$

sonst beliebig. (Es kommt im folgenden tatsächlich nur auf (2.1.13), an es braucht also nicht etwa $x\in I$ zu sein.)

Dann haben wir nach (2.1.11), (2.1.12) und (2.1.13)

$$M(x) \supseteq \{\mu \in M\colon I \cap K_\mu \neq \varnothing\}. \tag{2.1.14}$$

Weiter gilt dann

$$\begin{aligned} \mathrm{Gr}_K(I \cap K) &= \\ &= \mathrm{Gr}_K\left(\bigcup\{I \cap K_\mu\colon \mu \in N(x)\}\right) \subseteq \\ &\subseteq \bigcup\{\mathrm{Gr}_K(I \cap K_\mu)\colon \mu \in N(x)\} \quad (x \in H), \end{aligned} \tag{2.1.15}$$

mit der Bezeichnung

$$N(x) = \{\mu \in M\colon I \cap K_\mu \neq \varnothing\},$$

wobei es sich um endliche Vereinigungen handelt (weil auch die $N(x)$ endliche Mengen sind), und es müssen also zunächst die Mengen $\mathrm{Gr}_K(I \cap K_\mu)$ unter (2.1.15) einzeln untersucht werden.

Zu diesem Zweck führen wir noch die folgenden Bezeichnungen ein:

$$P_\mu = \bigcup\{K_{\mu'}\colon \mu' \in M,\ \mu' \neq \mu\} \quad (\mu \in M), \tag{2.1.16}$$

$$Q_\mu = P_\mu \cap K_\mu \quad (\mu \in M). \tag{2.1.17}$$

8° Es sei also jetzt

$$\mu \in N(x)$$

(beliebig, aber) festgesetzt, und wir behaupten:

$$\mathrm{Gr}_K(I \cap K_\mu) \subseteq (K_\mu \cap \mathrm{Gr}_X I) \cup Q_\mu.^* \tag{2.1.18}$$

Um dies zu beweisen setzen wir

$$y \in \mathrm{Gr}_K(I \cap K_\mu) \setminus Q_\mu \tag{2.1.19}$$

(sonst beliebig), und zeigen, daß

$$y \in K_\mu \cap \mathrm{Gr}_X I \tag{2.1.20}$$

ist, was — da offensichtlich $y \in K_\mu$ und $y \in I^{-(K)}$ — darauf hinausläuft,

$$y \notin I \tag{2.1.21}$$

zu beweisen.

Nun ist aber $y \in K_\mu \setminus Q_\mu$ und daher

$$y \notin P_\mu.$$

Wäre also $y \in I$, so müßte, da P_μ eine abgeschlossene Menge ist,

$$G = (I \cap K) \setminus P_\mu$$

---

* Hierzu zwei BEMERKUNGEN:

(1) Daß man auf der rechten Seite der Inklusion (2.1.18) zur Menge $K_\mu \cap \mathrm{Gr}_X I$ überhaupt irgend etwas hinzunehmen muß, erhellt daraus, daß $I \cap K_\mu$ keine offene Menge in der Topologie des Teilraumes K zu sein braucht.

(2) Unter Voraussetzung der Regularität des Raumes X wäre der Beweis etwas einfacher zu führen, weil man sich dann nämlich auf „so kleine" I beschränken könnte, daß sogar die $I^{-(K)}$ keine der Mengen $K_{\mu'}$ $(\mu' \in M \setminus N(x)$ schneiden. Da nun aber unser Satz dieser Einschränkung nicht bedarf, wollen wir auf diese Variante nicht eingehen.

eine offene Menge des Raumes K sein, für die

$$y \in G \subseteq I \cap K_\mu$$

gilt; das würde bedeuten, daß y ein innerer Punkt der Menge $I \cap K_\mu$ bezüglich der Topologie von K ist, was aber der Annahme unter (2.1.19) widerspricht.

Damit ist — über (2.1.21) und (2.1.20) — (2.1.18) bewiesen.*

9° Untersuchen wir zunächst das zweite Glied der rechten Seite der Inklusion unter (2.1.18).

Wir behaupten, daß die Menge Q_μ als Vereinigung einer lokal-endlichen Familie von abgeschlossenen, $(k; \mathfrak{R})$-kleinen Mengen darstellbar ist.

Tatsächlich gibt es ja nach (2.1.8) für jedes Paar

$$\mu \in M, \quad \mu' \in M \setminus \{\mu\}$$

einen Index

$$\varkappa[\mu', \mu] \in \{1, 2, \ldots, k-1\}$$

mit entweder

$$\nu(\varkappa[\mu', \mu], \mu') \neq \nu(\varkappa, \mu) \quad (\varkappa = 1, 2, \ldots, k-1)$$

oder — im Falle, daß

$$\nu(\varkappa, \mu') = \nu(\varkappa, \mu) \quad (\varkappa = 1, 2, \ldots, k-1) \; -$$

einen solchen mit

$$S^{\nu(\varkappa[\mu', \mu], \mu')}_{\varkappa[\mu', \mu], \mu'} \neq S^{\nu(\varkappa[\mu', \mu], \mu)}_{\varkappa[\mu', \mu], \mu}$$

und daher (s. (0.4),(b))

$$S^{\nu(\varkappa[\mu', \mu], \mu')}_{\varkappa[\mu', \mu], \mu'} \cap S^{\nu(\varkappa[\mu', \mu], \mu)}_{\varkappa[\mu', \mu], \mu} = \varnothing .$$

Somit ist — auch (1.5),(c) beachtend — diese Behauptung bewiesen.

10° Jetzt fassen wir das erste Glied der rechten Seite unter (2.1.18) ins Auge. Wir haben

(2.1.22)
$$K_\mu \cap \mathrm{Gr}_\chi I \subseteq$$
$$\subseteq \bigcup \{K_{\nu, j, \mu}[x]: \; \nu = 1, 2, \ldots, n; \; j = 1, 2\}$$

mit

(2.1.23)
$$K_{\nu, j, \mu}[x] =$$
$$= S^\nu_j[x] \cap \left(\bigcap \{S^{\nu(\varkappa, \mu)}_{\varkappa, \mu}: \; \varkappa = 1, 2, \ldots, k-1\}\right)$$
$$(\nu = 1, 2, \ldots, n; \; j = 1, 2),$$

---

* Die Beziehung unter (2.1.18) kann, etwas formaler, etwa auch so hergeleitet werden: es gilt — mit der Bezeichnung unter (2.1.16) —

$$K \setminus (I \cap K_\mu) \subseteq (K_\mu \setminus I) \cup P_\mu ,$$

wobei hervorgehoben sei, daß auf der rechten Seite eine abgeschlossene Menge des Raumes K steht, und wir haben daher (wegen der Lokal-Endlichkeit von $\mathscr{K}$ in K)

$$\mathrm{Gr}_K (I \cap K_\mu) = \mathrm{Gr}_K [K \setminus (I \cap K_\mu)] \subseteq (K_\mu \setminus I) \cup P_\mu ,$$

woraus wegen

$$\mathrm{Gr}_K (I \cap K_\mu) \subseteq I^{-(K)} \cap K_\mu$$

tatsächlich (2.1.18) folgt.

und zwar so, daß — sofern $S_j^\nu[x] \neq \emptyset$ —

$$S_j^\nu[x] \in \mathscr{S}(\mathscr{R}_\nu) \quad (\nu = 1, 2, \ldots, n;\ j = 1, 2)$$

und — selbstverständlich —

$$I \cap S_j^\nu[x] = \emptyset \quad (\nu = 1, 2, \ldots, n;\ j = 1, 2). \tag{2.1.24}$$

Da nun aber nach (2.1.6), (2.1.7) und (2.1.14)

$$I \cap S_{\varkappa,\mu}^{\nu(\varkappa,\mu)} \neq \emptyset \quad (\varkappa = 1, 2, \ldots, k-1)$$

und daher, auf Grund von (2.1.24)

$$S_j^\nu[x] \neq S_{\varkappa,\mu}^{\nu(\varkappa,\mu)}$$

$$(\nu = 1, 2, \ldots, n;\ j = 1, 2;\ \varkappa = 1, 2, \ldots, k-1)$$

ist, ergibt sich, wegen (0.4), (b), für jedes zulässige Paar $\nu, \varkappa$ unter (2.1.23) mit

$$\nu = \nu(\varkappa, \mu)$$

notwendigerweise

$$K_{\nu, j, \mu}[x] = \emptyset \quad (j = 1, 2).$$

Auf der rechten Seite der Inklusion unter (2.1.22) haben wir also die Vereinigung von *endlich vielen* Mengen, deren jede — wenn nicht die leere Menge — der Durchschnitt von k $\mathfrak{R}$-Ebenen ist, die aus k Ebenenscharen $\mathscr{S}(\mathscr{R}_\nu)$ mit paarweise verschiedenen Indizes ν stammen, und das bedeutet, daß die Menge $K_\mu \cap \mathrm{Gr}_X I$ als Vereinigung von endlich vielen, abgeschlossenen, $(k; \mathfrak{R})$-kleinen Mengen dargestellt werden kann.

11° Die Ergebnisse unter 9° und 10° zusammenfassend finden wir also auf Grund von (2.1.18), daß die Menge $\mathrm{Gr}_K(I \cap K_\mu)$ eine Darstellung als Vereinigung einer lokal-endlichen Familie abgeschlossener Mengen, deren jede $(k; \mathfrak{R})$-klein ist, zuläßt.

Da es sich weiter unter (2.1.15) um eine endliche Vereinigung handelt, gilt — die Festsetzung von $\mu \in M(x)$ aufgehoben — dieselbe Aussage auch auf die Menge $\mathrm{Gr}_K(I \cap K)$ bezogen.

Bisher waren aber auch noch $x \in H$ und $I \subseteqq W_x$ (s. 6°) festgesetzt. Jetzt soll auch das aufgehoben werden.

12° Durch Anwendung des Hilfssatzes (1.3) auf X und H (was nur im Fall (b″) möglich, im Fall (a″) aber auch unnötig ist), mit

$$\mathscr{W} = \{W_x\colon\ x \in H\}$$

(s. 6°) und

$$\mathscr{B} = \mathscr{I}(\mathfrak{R}),$$

ergibt sich die Existenz eines lokal-endlichen Systems

$$\mathscr{J} \subseteqq \mathscr{I}(\mathfrak{R}) \tag{2.1.25}$$

mit

$$\mathscr{J} < \mathscr{W}, \tag{2.1.26}$$

weiter — nach (1.3) und (2.1.10) —

$$H \subseteqq J \cap K \subseteqq U,$$

wo

$$J = \bigcup \{I\colon I \in \mathscr{J}\}$$

bedeutet, und wobei — nach (2.1.25), (2.1.26) und 7° — für jedes $I \in \mathscr{J}$ die Schlußfolgerung unter 11° gilt. (Im Falle (a″) nimmt man einfach ein einziges $I \in \mathscr{I}(\mathfrak{R})$ mit $H \subseteqq I \subseteqq W_x$ und setzt $\mathscr{J} = \{I\}$.)

Die Mengensysteme

$$\{I \cap K\colon I \in \mathscr{J}\}, \quad \{\mathrm{Gr}_K(I \cap K)\colon I \in \mathscr{J}\}$$

sind dann ebenfalls lokal-endlich.

13° Nunmehr haben wir

$$\mathrm{Gr}_K(J \cap K) = \mathrm{Gr}_K\left(\bigcup \{I \cap K\colon I \in \mathscr{J}\}\right) \subseteqq$$

$$\subseteqq \bigcup \{\mathrm{Gr}_K(I \cap K)\colon I \in \mathscr{J}\}.$$

Nach Hilfssatz (1.10) und dem Ergebnis unter 11° ist also auch die Menge $\mathrm{Gr}_K(I \cap K)$ als Vereinigung einer lokal-endlichen Familie von abgeschlossenen, $(k; \mathfrak{R})$-kleinen Mengen darstellbar.

Nach der Induktionsvoraussetzung unter 5° gilt daher

$$\operatorname{ind} \mathrm{Gr}_K(J \cap K) \leqq n-k$$

dzw.

$$\operatorname{Ind} \mathrm{Gr}_K(J \cap K) \leqq n-k$$

(für den Fall (a″) bzw. (b″)), woraus endlich — nach (2.1.10) und weil H und U unter 5° beliebig gewählt worden waren —

$$\operatorname{ind} K \leqq (n-k)+1 \quad (= n-(k-1))$$

dzw.

$$\operatorname{Ind} K \leqq (n-k)+1 \quad (= n-(k-1))$$

(also (a″) bzw. (b″)) folgt, und damit ist der Satz bewiesen. □

(2.2) BEMERKUNG. Die Bemerkung unter (0.28) kann sinngemäß auch auf die Sätze (2.1),(a) und (2.1),(b) sowie auf die weiteren Sätze dieser Art im § 3 bezogen werden; bei diesen Sätzen kommt es darauf an, daß man die induktiven Dimensionen, falls sie endlich sind (es gibt ja auch eine transfinite Erweiterung ihrer Definition, s. [14] 50), als Kardinalzahlen betrachtet, obwohl sie eigentlich als Ordinalzahlen definiert werden.

(2.3) Es sei noch bemerkt, daß beide Sätze (2.1),(a) und (2.1),(b) Verschärfungen (indem Dim X an Stelle von $(\mathrm{Dim}\, X)^*$ tretet) und zugleich Verallgemeinerungen (indem die Voraussetzungen bedeutend abgeschwächt wurden) der entsprechenden Sätze (0.24) und (0.25) sind; mit (2.1),(a) haben wir sogar die weitgehendste überhaupt mögliche Verallgemeinerung von (0.24) erreicht.

§ 3. Einige nennenswerte Spezialfälle der Fundamentalsätze

Es werden in diesem Paragraphen einige Sätze angeführt, die sämtlich entweder unmittelbare Spezialfälle der Sätze unter (2.1) sind oder sich durch Kombination der letzteren mit bekannten dimensions-theoretischen Sätzen leicht beweisen lassen. „Nennenswert" sind diese Sätze, weil sie sich auf wohlbekannte und vielbehandelte Klassen von Räumen beziehen; darüber hinaus können wir für einige von ihnen auch selbständige (d. h. die ziemlich schwierig bewiesenen Sätze unter (2.1) umgehende) Beweise geben.

(3.1) SATZ. *Für einen total-parakompakten T_2-Raum X endlicher RD gilt*

$$\dim X \leqq \operatorname{Dim} X.$$

BEWEIS. Nach einem Satz von N. B. WEDENISOFF ([*22*]; [17] 56) gilt für jeden normalen T_2-Raum $\dim X \leqq \operatorname{Ind} X$. Da jeder total-parakompakte T_2-Raum (eigentlich sogar — wie bekannt — jeder parakompakte T_2-Raum) normal ist, kann nun der erwähnte Satz mit unserem Satz (2.1),(b) kombiniert werden, wodurch sich der Beweis der vorliegenden Satzes ergibt. □

(3.2) SATZ. *Für einen kompakten Raum X endlicher RD gilt*

$$\operatorname{Ind} X \leqq \operatorname{Dim} X.$$

BEWEIS. Es ist ein Spezialfall von (2.1.),(b). □

(3.3) SATZ. *Für einen kompakten T_2-Raum X endlicher RD gilt*

$$\dim X \leqq \operatorname{Dim} X.$$

1. BEWEIS. Es handelt sich um einen Spezialfall von (3.1). □

2. BEWEIS. Man verbindet den Satz von P. S. ALEXANDROFF ([*1*]), nach dem für jeden kompakten T_2-Raum X $\dim X \leqq \operatorname{ind} X$ gilt, mit unserem Satz (3.2) oder unmittelbar mit (2.1),(a). □

(3.4) HILFSSATZ. *Es seien $\mathfrak{R}$ eine endliche KRS eines T_1-Raumes Y (mit notwendigerweise endlicher RD) und*

$$\dim Y > 0. \tag{3.4.1}$$

Dann gibt es ein $\mathscr{R} \in \mathfrak{R}$ und ein $(G, F) \in \mathscr{R}$ mit $G \neq \varnothing$ und $Y \setminus F \neq \varnothing$.

BEWEIS. Wären

$$G = \varnothing \vee F = Y$$

$$((G, F) \in \mathscr{R},\ \mathscr{R} \in \mathfrak{R}),$$

so müßte

$$|\mathscr{R}| < \aleph_0 \quad (R \in \mathfrak{R})$$

sein (man sieht nämlich sofort, daß dann — von trivialen Richtungen abgesehen — überhaupt nur Richtungen der Art

$$\{(\varnothing, \varnothing),\ (G, Y),\ (Y, Y)\} \quad (G \subsetneqq Y),$$

$$\{(\varnothing, \varnothing), (\varnothing, F), (Y, Y)\} \quad (\varnothing \neq F \subsetneqq Y)$$

oder

$$\{(\varnothing, \varnothing), (\varnothing, F), (G, Y), (Y, Y)\}$$

$$(\varnothing \neq F \subsetneqq G \subsetneqq Y)$$

als Elemente von $\mathfrak{R}$ auftreten können), und wegen $|\mathfrak{R}|<\aleph_0$ würde das $w(Y)<\aleph_0$ nach sich ziehen; da Y ein T_1-Raum ist, müßte er daher indiskret sein, was aber (3.4.1) widerspricht. □

(3.5) SATZ. *Für jeden kompakten T_2-Raum X mit* $\dim X<\infty$ *gilt*

(3.5.1) $$\dim X \leqq \operatorname{Dim} X.$$

BEWEIS. 1° In den Trivialfällen

$$\dim X = 0, \quad \dim X = 1, \quad \operatorname{Dim} X = 0$$

gilt (3.5.1) einfach wegen (0.12),(a), (0.12),(b) bzw. (0.12),(b),(b′).*

Der Beweis unseres Satzes erfolgt nunmehr induktiv, u. zw. ganz unabhängig vom Fundamentalsatz (2.1).

2° Es sei

$$k \geqq 2$$

eine natürliche Zahl, und nehmen wir — auf Grund von 1° — als Induktionsvoraussetzung an, daß die Behauptung (3.5.1) für

$$\dim X = 0, 1, \ldots, k-1$$

und für

$$\operatorname{Dim} X = 0, \ldots, k-2$$

schon bewiesen ist; wir haben zu zeigen, daß dann (3.5.1) auch für die Fälle

$$\dim X = k, \quad \operatorname{Dim} X = k-1$$

gültig ist.

3° Ein Satz von W. HUREWICZ und K. MENGER ([*13*], [16] 217; mit einer leichten Abänderung — wie K. NAGAMI in [17] bemerkte — kann auch der Beweis in [14], S. 94, benutzt werden, obwohl der Satz dort auf metrisierbare (kompakte) Räume bezogen ist) besagt, daß jeder kompakte T_2-Raum X mit $\dim X=n$ $(n\geqq 1)$ eine Cantorsche n-Mannigfaltigkeit C als Teilraum enthält.

(Der von P. S. URYSOHN in [*20*], S. 124, geprägte Begriff einer *Cantorschen n-Mannigfaltigkeit* hat seine traditionelle Benennung his heute behalten, obwohl die Sache kaum etwas mit den üblichen Bedeutungs-varianten des Fachwortes „Mannigfaltigkeit" zu tun hat; es handelt sich nämlich um diejenigen kompakten

* Eigentlich ist auch der Fall $\operatorname{Dim} X=1$ — nämlich durch unseren Satz (0.20) — schon erledigt; das folgende Beweisverfahren ist aber derart aufgebaut, daß es unnötig ist, sich auf den — nicht gerade leicht bewiesenen und für den hiesigen Zweck überdies auch viel zu allgemeinen — Satz (0.20) zu berufen.

Es wäre auch nicht stilgerecht, den Satz (3.3) heranzuziehen, weil in dessen Beweis der Satz (2.1) benutzt wird, den wir jetzt eben umgehen wollen.

T_2-Räume C mit $\dim C=n$, $n\geqq 1$, deren jeder Teilraum $C^*\subseteq C$ zusammenhängend ist, wenn nur $\dim (C\setminus C^*)\leqq n-2$ gilt.)

Auch der besagte Hurewicz—Mengersche Satz wurde durch ähnliche Sätze ([20]) und eine Fragestellung ([21] 285) Urysohns angeregt.)

4° Haben wir nun einen kompakten T_2-Raum X mit

$$\dim X = k, \tag{3.5.2}$$

so gibt es unter den Teilräumen von X eine Cantorsche k-Mannigfaltigkeit C.

Wäre — im Gegensatz zu unserer Behauptung — $\operatorname{Dim} X\leqq k-1$, so hätten wir nach (0.12),(d) auch $\operatorname{Dim} C\leqq k-1$, und es gäbe — nach dem Hilfssatz (3.4), der hier wegen (3.5.2) mit $Y=C$ anwendbar ist — für jede MKRS $\mathfrak{R}$ von C ein $\mathscr{R}\in\mathfrak{R}$ und ein $(G, F)\in\mathscr{R}$ mit $G\neq\varnothing$ und gleichzeitig $C\setminus F\neq\varnothing$, so daß also der Teilraum $C\setminus(F\setminus G)$ von C wegen $G\subseteq F$ (s. (0.1)) nicht zusammenhängend ist, obwohl laut Induktionsvoraussetzung (die auf den Teilraum $F\setminus G$ von C, weil er als abgeschlossene Teilmenge eines kompakten Raumes kompakt ist und weil nach (0.16),(a) die RD dieses Teilraumes höchstens $k-2$ beträgt, angewendet werden kann)

$$\dim (F\setminus G)\leqq \operatorname{Dim}(F\setminus G)\leqq k-2$$

wäre. Mit diesem Widerspruch ist die Behauptung für den Fall (3.5.2) bewiesen.

5° Ist aber X ein kompakter T_2-Raum mit

$$\operatorname{Dim} X = k-1 \tag{3.5.3}$$

und wäre — im Gegensatz zur Behauptung (3.5.1) —

$$\dim X = j \geqq k+1 \tag{3.5.4}$$

(dim X wurde als endlich vorausgesetzt, und $\dim X=k$ ist durch das Ergebnis von 4° schon als unvereinbar mit (3.5.3) erkannt worden), so gäbe es eine Cantorsche j-Mannigfaltigkeit C mit $\operatorname{Dim} C\leqq k-1$.

Wieder folgt (genau wie unter 4°) aus dem Hilfssatz (3.4) für eine beliebige MKRS $\mathfrak{R}$ von C, die Existenz eines $\mathscr{R}\in\mathfrak{R}$ und eines $(G, F)\in\mathscr{R}$ mit

$$\operatorname{Dim}(F\setminus G)\leqq k-2, \tag{3.5.5}$$

so daß der Raum $C\setminus(F\setminus G)$ in die zwei disjunkte nichtleere offene Teilmengen G und $C\setminus F$ zerlegt werden könnte und daher

$$\dim (F\setminus G)\geqq j-1 \tag{3.5.6}$$

sein müßte, obwohl aus (3.5.5) laut Induktionsvoraussetzung

$$\dim (F\setminus G)\leqq j-3 \tag{3.5.7}$$

folgt.

Mit dem Widerspruch zwischen (3.5.6) und (3.5.7) ist nun die Behauptung auch im Falle (3.5.3) bewiesen. □

(3.6) BEMERKUNG. Auf die Bemerkungen (0.27) und (2.2) zurückverweisend muß geklärt werden, daß, für irgendeine Klassse $\mathscr{X}$ von Räumen,

$$[\dim X < \infty \Rightarrow \dim X \leqq \operatorname{Dim} X \ (X \in \mathscr{X})] \not\Rightarrow$$

$$\not\Rightarrow [\operatorname{Dim} X < \aleph_0 \Rightarrow \dim X \leqq \operatorname{Dim} X \ (X \in \mathscr{X})]$$

(obwohl

$$[\operatorname{Dim} X < \aleph_0 \Rightarrow \dim X \leqq \operatorname{Dim} X \ (X \in \mathscr{X})] \Rightarrow$$

$$\Rightarrow [\dim X < \infty \Rightarrow \dim X \leqq \operatorname{Dim} X \ (X \in \mathscr{X})]).$$

In der Formulierung des Satzes (3.5) kann also die Voraussetzung „dim X ist endlich" nicht ohne weiteres durch die Voraussetzung „Dim X ist endlich" ersetzt werden. M. a. W.: (3.3) impliziert (3.5), aber (3.5) impliziert nicht (jedenfalls nicht so unmittelbar) den Satz (3.3).

(3.7) SATZ. *Für jeden lokalkompakten, parakompakten Raum X endlicher RD gilt*

$$\operatorname{Ind} X \leqq \operatorname{Dim} X.$$

BEWEIS. Nach R. TELGÁRSKY [*19*] ist jeder lokalkompakte, parakompakte Raum total-parakompakt (Telgársky beweist eigentlich eine noch schärfere Aussage); es handelt sich also um einen Spezialfall von (2.1),(b). □

Der folgende Satz ist, da für den — normalen — T_2-Raum X $\dim X \leqq \operatorname{Ind} X$ gilt (wir haben uns darauf schon unter (3.1) berufen), eigentlich ein Spezialfall von (3.7) (man beachte auch (0.27)), dessen eigene Formulierung damit gerechtfertigt ist, daß wir auch einen eigenen — d. h. unseren Fundamentalsatz (2.1),(b) und überhaupt den Begriff der Total-Parakompaktheit vermeidenden — Beweis geben.

(3.8) SATZ. *Für jeden lokalkompakten, parakompakten T_2-Raum X endlicher RD gilt*

(3.8.1) $$\dim X \leqq \operatorname{Dim} X.$$

BEWEIS. 1° Bekanntlich (s. etwa [19] 286) ist ein lokalkompakter T_2-Raum genau dann parakompakt, wenn er die topologische Summe von σ-kompakten Räumen ist. Wir haben also

(3.8.2) $$X = \sum_{\alpha \in A} X_\alpha, \quad X_\alpha = \bigcup_{i=1}^{\infty} Y_\alpha^i \quad (\alpha \in A),$$

wo Y_α^i für jedes Paar $\alpha \in A$, $i = 1, 2, \ldots$ kompakt ist.

2° Wegen der Monotonie der RD haben wir

$$\operatorname{Dim} Y_\alpha^i \leqq \operatorname{Dim} X \quad (\alpha \in A, \ i = 1, 2, \ldots),$$

woraus nach Satz (3.5) (dessen Beweis unabhängig von (2.1) war!)

(3.8.3) $$\dim Y_\alpha^i \leqq \operatorname{Dim} X \quad (\alpha \in A, \ i = 1, 2, \ldots)$$

folgt.

3° Ein Satz von E. ČECH ([2], [17] 53) besagt, daß die Überdeckungsdimension eines normalen T_2-Raumes, der als Vereinigung abzählbar vieler abgeschlossener Teilräume dargestellt werden kann, gleich dem Supremum der Überdeckungsdimensionen dieser Teilräume ist.

Aus (3.8.3) folgt demnach

$$\dim X_\alpha \leqq \operatorname{Dim} X \quad (\alpha \in A),$$

und weiter, da es sich unter (3.8.2) um eine topologische Summe handelt, deren Überdeckungsdimension trivialerweise gleich dem Supremum der Überdeckungsdimensionen der Summanden ist, die Behauptung (3.8.1) des Satzes. □

(3.9) BEMERKUNGEN. 1° Was das Verhältnis der bisherigen Ergebnisse dieses Paragraphen zu jenen in [6] betrifft: man sieht sofort, daß mit den Sätzen (3.2) und (3.7) Fortschritte gegenüber (0.25) und mit den Sätzen (3.1) (hier beachte man wieder den schon unter (3.7) angeführten Satz von Telgársky) und (3.8) Fortschritte gegenüber (0.26) gemacht wurden.

2° Satz (3.5) stellt ein völlig neues Ergebnis gegenüber jenen in [6] dar.

3° Was endlich (0.27) betrifft: dieser Satz ist durch unseren Fundamentalsatz (2.1),(a) überhaupt belanglos geworden, da nach dem Satz von K. NAGAMI ([*18*] 289), *für σ-total-parakompakte*, total-normale** T_2-Räume* X, $\operatorname{Ind} X = \operatorname{ind} X$ *gilt* (und da offensichtlich jeder Lindelöfsche Raum σ-total-parakompakt und jeder vollkommen normale Raum total-normal ist). Damit haben wir nicht nur eine Verallgemeinerung sondern gleichzeitig auch eine Verschärfung (wegen $\operatorname{Ind} X \geqq \dim X$ für normale Räume X) des Satzes (0.27) gewonnen. Es gilt sogar die folgende gemeinsame Verallgemeinerung der Sätze (0,27) und (3.3):

(3.10) SATZ. *Für einen Lindelöfschen, regulären T_2-Raum X endlicher RD gilt*

$$\dim X \leqq \operatorname{Dim} X.$$

BEWEIS. Man verbindet Satz (2.1,(a) mit dem Satz von K. MORITA [15], nach dem für Lindelöfsche, reguläre T_2-Räume X $\dim X \leqq \operatorname{ind} X$ gilt. □

LITERATUR

[1] ALEXANDROFF, P. S.: Der endliche dimensionstheoretische Summensatz für bikompakte Räume (russisch), *Soobsč. Akad. Nauk Grusin. SSR* **2** (1941), 1—6.

[2] ČECH, E.: Contribution à la théorie de la dimension, *Čas. Math. Fys.* **62** (1933), 277—290.

[3] DEÁK E.: Ein neuer topologischer Dimensionsbegriff, *Revue Roum. Math. Pures Appl.* **10** (1965), 31—42.

[4] DEÁK E.: Bemerkung zu meiner vorangehender Arbeit „Ein neuer topologischer Dimensionsbegriff", *Revue Roum. Math. Pures Appl.* **13** (1968), 303—305.

[5] DEÁK E.: Eine vollständige Charakterisierung der Teilräume eines euklidischen Raumes mittels der Richtungsdimension, *Publ. Math. Inst. Hung. Acad. Sci.* **9** (1965), 437—465.

* Ein Raum heißt *σ-total-parakompakt*, wenn aus jeder Basis des Raumes eine σ-lokalendliche Überdeckung desselben ausgewählt werden kann ([*18*] 289).

** Ein normaler Raum X heißt *total-normal*, wenn jede offene Menge G von X als Vereinigung eines Mengensystems dargestellt werden kann, welches lokal-endlich in G ist und dessen Elemente offene F_σ-Mengen bezüglich X sind ([9] 273, [17] 42).

[6] DEÁK E.: Einige Beziehungen der Richtungsdimension zu den klassischen Dimensionsbegriffen der allgemeinen Topologie, *Math. Nachr.* **37** (1968), 247—266.
[7] DEÁK E.: Dimension und Konvexität II (ungarisch), *MTA III. Oszt. Közl.* **17** (1967), 312—329.
[8] DEÁK, E.: Theory and Applications of Directional Structures, *Coll. Math. Soc. János Bolyai* **8** *(Topics in Topology, Keszthely, Hungary, 1972)*, Amsterdam—London, 1974.
[9] DOWKER, C. H.: Inductive dimension of completely normal spaces, *Quart. J. Math. Oxford, second series* **4** (1953), 267—281.
[10] FITZPATRICK, B. JR.—FORD, R. M.: On the Equivalence of Small and Large Inductive Dimension in Certain Metric Spaces, *Duke Math. J.* **34** (1967), 33—37.
[11] FORD, R. M.: *Basis Properties in Dimension Theory*, Doctoral dissertation, Auburn University, Auburn, Ala., 1963.
[12] FRENCH, J. A.: Coincidence of Small and Large Inductive Dimension, *Lecture Notes in Math.* **378** (1974) *(Second Pittsburgh Internat. Conf. on General Topology and its Applications, December 18—22, 1972)*, 132—139.
[13] HUREWICZ, W.—MENGER, K.: Dimension und Zusammenhangsstufe, *Math. Ann.* **100** (1928), 618—633.
[14] HUREWICZ, W.—WALLMAN, H.: *Dimension Theory*, Princeton 1948.
[15] MORITA, K.: On the dimension of normal spaces, I, *Jap. Journ. Math.* **20** (1950), 5—36.
[16] MENGER, K.: *Dimensionstheorie*, Leipzig und Berlin, 1928.
[17] NAGAMI, K.: *Dimension Theory*, Academic Press, New York and London, 1970.
[18] NAGAMI, K.: A note on the large inductive dimension of totally normal spaces, *J. Math. Soc. Japan* **21** (1969), 282—290.
[19] RINOW, W.: *Lehrbuch der Topologie*, VEB Deutscher Verlag der Wiss., Berlin, 1975.
[20] URYSOHN, P. S.: Mémoire sur les multiplicités cantoriennes I, *Fund. Math.* **7** (1925), 30—139.
[21] URYSOHN, P. S.: Mémoire sur les multiplicités cantoriennes II, *Fund. Math.* **8** (1926), 225—359.
[22] VEDENISOFF, N. B.: Sur la dimension au sens de E. Čech (russisch; französische Zusammenfassung), *Isv. Akad. Nauk SSSR, ser. math.* **5** (1941), 211—216.

Mathematisches Forschungsinstitut der Ungarischen Akademie der Wissenschaften
H—1053 Budapest, Reáltanoda u. 13—15, Ungarn

(Eingegangen am 10. Januar, 1976)

Studia Scientiarum Mathematicarum Hungarica 10 (1975) 459—462.

DIMENSION AND METRIZATION OF UNIFORM SPACES

by

SVETLANA BÚZÁSI

In this paper Theorem 2. gives a characterization of the "large covering" dimension ΔdX of metrizable uniform space X by the property of its (uniformity preserving) metric. Theorem 1. states that X admits the uniformity preserving metric of property Nag^n, if $\Delta dX \leqq n$. These results correspond to I. NAGATA's and P. OSTRAND's theorems where topological spaces are characterized in such way (see [1], p. 138 and [2]). As a corollary we have that the covering dimension of metrizable topological space X ($\dim X$) is the minimum of $\Delta d(X, \mu)$ for all topology preserving metrizable uniformities μ on X (Theorem 15 in [3], p. 153).

By ΔdX we denote the large covering dimension of a uniform space X (see [3]), by $\dim X$ we mean the topological covering dimension of X considered with uniform topology.

For $A \subset X$ and a covering $\mathcal{U}$ of X let

$$S^t(A, \mathcal{U}) = \begin{cases} A, & \text{if} \quad t = 0, \\ \cup \{U | U \in \mathcal{U},\ U \cap A \neq \emptyset\} & \text{if} \quad t = 1, \\ S(S^{t-1}(A, \mathcal{U}), \mathcal{U}) & \text{if} \quad t > 1, \end{cases}$$

$$[\mathcal{U}]^t = \{S^t(A, \mathcal{U}) | A \in \mathcal{U}\}, \quad [\mathcal{U}]^1 = [\mathcal{U}].$$

We put $S_\varepsilon = \{S_\varepsilon(x) | x \in X\}$ where $\varepsilon > 0$ is a real number and $S_\varepsilon(x)$ an open ε-ball, in a metric space X.

LEMMA 1. *Let X be a metric uniform space and $\Delta dX \leqq n$. For each positive integer j and an arbitrary fixed positive integer t there exist $n+1$ uniformly discrete families of sets $\mathcal{U}_j^1, \ldots, \mathcal{U}_j^{n+1}$ such that for each j:*

(1) $\mathcal{U}_j = \bigcup_{i=1}^{n+1} \mathcal{U}_j^i$ *covers* X;

(2) $\operatorname{mesh} \mathcal{U}_j < (2t)^{-j}$;

(3) $[\mathcal{U}_{j+1}]^{t-1}$ *refines* $\mathcal{U}_j$;

(4) *if* $j < k$ *and* $1 \leqq i \leqq n+1$, *each member of* $[\mathcal{U}_k]^{t-1}$ *meets at most one member of* $\mathcal{U}_j^i$.

PROOF. Since each uniform covering of X can be refined by an uniform covering consisting of $n+1$ uniformly discrete families (see [3], p. 67) we can construct the sequence $\mathcal{U}_j$ ($j = 1, 2, \ldots$) by induction. Begin with $\mathcal{S}_{\frac{1}{4t}} > \mathcal{U}_1$ and if we already have $\mathcal{U}_1, \ldots, \mathcal{U}_j$ which satisfy the (1)—(4) of lemma 1, let $\mathcal{U}_j > \mathcal{S}_{\varepsilon_j} > \mathcal{S}_{\frac{\varepsilon_j}{4t}} > \mathcal{U}_{j+1}$ where $S_{\varepsilon_j}(x)$ meets at most one member of $\mathcal{U}_j^i$ for all $x \in X$.

THEOREM 1. *If* $\Delta dX \leqq n$ *for a metrizable uniform space* X, *then* X *admits a metric* d *compatible with the uniformity which satisfies the condition* Nag^n: *For any* $n+3$ *points* $x, y_1, \ldots, y_{n+2}$ *of* X *there exist distinct indices* i, j *such that* $d(x, y_j) \geqq d(y_i, y_j)$.

PROOF. We shall make use of OSTRAND's construction of [2], working with uniform coverings; in this way the new metric d induces the uniformity of X originally generated by a metric ϱ. We get a sequence of uniform coverings $\mathscr{U}_j$ $(j=1, 2, \ldots)$ in (X, ϱ) satisfying conditions (1)—(4) of [2] applying Lemma 1. with $t=32$. Then for each dyadic rational $m=\sum_{k=1}^{t} 2^{-m_k}$ where $1 \leqq m_1 < \ldots < m_t$, the covering $\mathscr{S}_m$ can be constructed from the sequence $\mathscr{U}_j$ $(j=1, 2, \ldots)$ and then a metric d of the property Nag^n can be defined in the same way as in [2]. To show that d is compatible with the uniformity of (X, ϱ) we need the following

LEMMA 2. *For each* $m=\sum_{k=1}^{t} 2^{-m_k}$ *where* $1 < m_1$ *we have* $\mathscr{U}_{m_1} < \mathscr{S}_m < \mathscr{U}_{m_1-1}$.

Proof of Lemma 2. The relation $\mathscr{U}_{m_1} < \mathscr{S}_m$ is clear from the definition of $\mathscr{S}_m$ in [2]. For any $A \subset X$ and each $\mathscr{U}_j$ we have

$$S^{31}(A, \mathscr{U}_{j+1}) \subset S(A, [\mathscr{U}_{j+1}]^{30}) \subset S\{A, \mathscr{U}_j\},$$

so the coverings ${}^*\mathscr{U}_j$ from [2] satisfy

$$\mathscr{U}_j < {}^*\mathscr{U}_j < [\mathscr{U}_j].$$

Hence for any $A \subset X$ we get

$$S(A, {}^*\mathscr{U}_j) \subset S^3(A, \mathscr{U}_j), \quad S^3(A, {}^*\mathscr{U}_j) \subset S^9(A, \mathscr{U}_j).$$

Let now consider the set $T(A, i, j)$ defined in [2] which can be also written as follows:

$$T(A, i, j) = \ldots = S\big(S(A, {}^*\mathscr{U}_j^i), {}^*\mathscr{U}_{j+1}^i\big)\ldots,$$

that is we form stars of A relative to families ${}^*\mathscr{U}_j^i, {}^*\mathscr{U}_{j+1}^i, \ldots$. Finishing the star-forming after $k+1$ steps we get $T^k(A, i, j)$. Note that for families with the property $[\mathscr{U}_{j+1}]^m < \mathscr{U}_j$ and for any $A \subset X$ we have

$$S^{k_r}\big(\ldots S^{k_2}\big(S^{k_1}(A, \mathscr{U}_1), \mathscr{U}_2\big)\ldots, \mathscr{U}_r\big) \subset S^{k_1+1}(A, \mathscr{U}_1) \tag{6}$$

if $k_2, \ldots, k_r \leqq m$ (easy to prove by induction on r). Hence for each $k>0$

$$T^k(A, i, j) = S\big(\ldots S\big(S(A, {}^*\mathscr{U}_j^i), \mathscr{U}_{j+1}^i\big)\ldots\big), {}^*\mathscr{U}_{j+k}^i\big) \subset$$

$$\subset S\big(\ldots S\big(S(A, {}^*\mathscr{U}_j), {}^*\mathscr{U}_{j+1}\big)\ldots\big), {}^*\mathscr{U}_{j+k}\big) \subset$$

$$\subset S^3\big(\ldots S^3\big(S^3(A, \mathscr{U}_j), \mathscr{U}_{j+1}\big)\ldots\big), \mathscr{U}_{j+k}\big) \subset S^4(A, \mathscr{U}_j),$$

and so $T(A, i, j) \subset S^4(A, \mathcal{U}_j)$. Let now $m = \sum_{k=1}^{t} 2^{-m_k}$, $1 < m_1$, $i_U m \in \mathcal{S}_m^i$, $U \in \mathcal{U}_{m_1}^i$ (see [2]). Then

$$\begin{aligned} i_U m = T\big(S^3(\dots T\big(S^3\big(T(U, i, m_1+1),\ {}^*\mathcal{U}_{m_2}\big), i, m_2\big)\dots,\ {}^*\mathcal{U}_{m_t}\big), i, m_t\big) \subset \\ \subset S^4\big(S^9\big(\dots S^4\big(S^9\big(S^4(U, \mathcal{U}_{m_1+1}),\ \mathcal{U}_{m_2}\big)\dots,\ \mathcal{U}_{m_t}\big),\ \mathcal{U}_{m_t}\big) \subset \\ \subset S^{13}\big(\dots S^{13}\big(S^4(U, \mathcal{U}_{m_1+1}),\ \mathcal{U}_{m_2}\big)\dots,\ \mathcal{U}_{m_t}\big) \subset \\ \subset S^{14}\big(S^4(U, \mathcal{U}_{m_1+1}),\ \mathcal{U}_{m_2}\big) \subset S^{18}(U, \mathcal{U}_{m_1+1}). \end{aligned}$$

So we get for each i, $1 \leqq i \leqq n+1$,

$$S_m^i < {}^*\mathcal{U}_{m_1}^i < [\mathcal{U}_{m_1}] < \mathcal{U}_{m_1-1}$$

which completes the proof of Lemma 2.

Now we can easily show that ϱ and d generate the same uniformity in X. Let $d(x, y) < m$ where $m = \sum_{k=1}^{t} 2^{-m_k}$. By the definition of d for some dyadic rational p, $0 < p < m$, $y \in S(x, S_p)$, but $\mathcal{S}_p < \mathcal{S}_m < \mathcal{U}_{m_1-1}$ so $y \in S(x, \mathcal{U}_{m_1-1})$ and hence $\varrho(x, y) < (2 \cdot 32)^{-m_1+1} = 2^{6(1-m_1)}$. In the proof of Lemma 1. we got the sequence ε_j $(j=1, 2, \dots)$ where $\varepsilon_1, \varepsilon_2, \dots \to 0$ and $\mathcal{U}_j > \mathcal{S}_{\varepsilon_j}$. Let now $\varrho(x, y) < \varepsilon_j$ for some $1 < j$. Then $y \in S(x, \mathcal{U}_j) \subset S(x, S_m)$ is true for each m with $m_1 = j$. Hence we have $d(x, y) \leqq \frac{1}{2^j} < \frac{1}{2^{j-1}}$, which completes the proof of Theorem 1.

THEOREM 2. *A metrizable uniform space X is of $\Delta dX \leqq n$ iff X admits a metric ϱ compatible with the uniformity such that for every $\varepsilon > 0$ and for every point $x \in X$*

$$(*) \qquad \varrho\big(S_{\frac{\varepsilon}{2}}(x), y_i\big) < \varepsilon, \quad i = 1, \dots, n+2$$

imply $\varrho(y_i, y_j) < \varepsilon$ for some i, j with $i \neq j$.

PROOF. Let (X, d) be a metric uniform space with $\Delta dX \leqq n$. Using Lemma 1 with $t=7$ we get a sequence $\mathcal{U}_j$ $(j=1, 2, \dots)$ of uniform coverings of X such that for each j:

(2) mesh $\mathcal{U}_j < 2^{-4j}$,

(3) $[\mathcal{U}_{j+1}]^6$ refines $\mathcal{U}_j$,

(4) each member of $[\mathcal{U}_{j+1}]^6$ meets at most $n+1$ members of $\mathcal{U}_j$.

From (2) it follows that $\{S(x, \mathcal{U}_j) \mid j=1, 2, \dots\}$ is a neighborhood base at each $x \in X$ in the topology induced by d in X. We shall show that for each j and $x \in X$ the set $S^2(x, [\mathcal{U}_{j+1}])$ meets at most $n+1$ members of $\mathcal{U}_j$. Let $x \in U \in \mathcal{U}_{j+1}$ then $S(x, [\mathcal{U}_{j+1}]) \subset S^3(U, \mathcal{U}_{j+1})$, so $S^2(x, [\mathcal{U}_{j+1}]) \subset S^6(U, \mathcal{U}_{j+1}) \in [\mathcal{U}_{j+1}]^6$ and from (3) and (4) we get our statement. Now it is clear that the sequence $\mathcal{U}_j$ $(j=1, 2, \dots)$ is suitable for NAGATA's construction of Theorem V. 3 (see [1], page 138). For integers $1 \leqq m_1 < m_2 < \dots < m_p$ and $U \in \mathcal{U}_{m_1}$ the set $S_{m_1 m_2 \dots m_p}(U)$ is defined as follows:

$$S_{m_1 m_2 \dots m_p}(U) = \begin{cases} U & \text{if} \quad p = 1 \\ S^2\big(\dots S^2\big(S^2(U, \mathcal{U}_{m_2}), \mathcal{U}_{m_3}\big)\dots, \mathcal{U}_{m_p}\big) & \text{if} \quad p > 1. \end{cases}$$

From the definition and (6) follows

$$S_{m_1 m_2 \dots m_p}(U) \subset S^3(U, \mathcal{U}_{m_2}) \subset S(U, [\mathcal{U}_{m_2}]) \subset S(U, \mathcal{U}_{m_1}).$$

So we have for $1 < m_1 < m_2 < \dots < m_p$

$$\mathcal{U}_{m_1} < \sigma_{m_1 \dots m_p} < [\mathcal{U}_{m_1}] < \mathcal{U}_{m_1 - 1} \tag{7}$$

where

$$\sigma_{m_1 \dots m_p} = \{S_{m_1 \dots m_p}(U) \mid U \in \mathcal{U}_{m_1}\}.$$

The metric ϱ defined in Theorem V. 3. by means of coverings $\sigma_{m_1 \dots m_p}$ is compatible with the uniformity of (X, d). This can be proved with the help of (5) and (7) in the same manner as in Theorem 1.

Conversely, let the uniformity of X be generated by a metric ϱ with the property described in Theorem 2. Let $\mathcal{U}$ be a uniform covering of X and $\delta > 0$ a real number such that $\mathcal{S}_\delta < \mathcal{U}$. Put $\delta = \frac{3\varepsilon}{2}$ and let M be a maximal set in X with the property: $x, y \in M$, $x \neq y \Rightarrow \varrho(x, y) \geqq \varepsilon$. The covering $\mathcal{S}_\delta^M = \{S_\delta(x) \mid x \in M\}$ is a subcovering of $\mathcal{S}_\delta$ and since $\mathcal{S}_{\frac{\varepsilon}{2}} < \mathcal{S}_\delta^M$ it is a uniform covering of X. The properties of ϱ and M garantee that ord $S_\delta^M \leqq n+1$ so we have $\Delta dX \leqq n$ which completes the proof of Theorem 2.

COROLLARY. *If* dim X *of a metrizable topological space* X *is finite,* dim X *is the minimum of* $\Delta d(X, \mu)$ *for all topology preserving metrizable uniformities* μ *on* X.

PROOF. From NAGATA's Theorem V. 3. ([1], p. 138) and Theorem 1. above it follows that $\dim X \leqq \Delta d(X, \mu)$. If $\dim X \leqq n$ we can introduce a topology preserving metric ϱ which has the property $(*)$, but the second part of the proof of Theorem 2. shows that the metric of such property generates a uniformity μ on X with $\Delta d(X, \mu) \leqq n$.

REFERENCES

[1] NAGATA, I.: *Modern dimension theory,* Amsterdam, 1965.
[2] OSTRAND, P.: A conjecture of J. Nagata on dimension and metrization, *Bull. Amer. Math. Soc.* **71** (1965), 623—625.
[3] ISBELL, J. R.: *Uniform spaces,* Providence, R. I, 1964.

Kossuth Lajos University, Debrecen, Hungary

(Received May 13, 1976)

Studia Scientiarum Mathematicarum Hungarica 10 (1975), 463—472.

DISCRETE AND EQUAL CONVERGENCE

by

Á. CSÁSZÁR and M. LACZKOVICH

1.

Let X be a non-empty set and let f_n, f be real valued functions defined on X. We say that f is the *discrete limit* of the sequence (f_n) if, for every $x \in X$, there exists $n_0 = n_0(x)$ such that $f(x) = f_n(x)$ for $n \geqq n_0$. The terminology is motivated by the fact that this condition means precisely the convergence of $(f_n(x))$ to $f(x)$ with respect to the discrete topology of the real line.

f is said to be the *equal limit* of the sequence (f_n) if there is a sequence of positive numbers $\varepsilon_n \to 0$ such that, for every $x \in X$, there exists $n_0 = n_0(x)$ with $|f(x) - f_n(x)| < \varepsilon_n$ for $n \geqq n_0$.

Let Φ be an arbitrary class of functions defined on X. We denote by Φ^λ, Φ^d, and Φ^e the classes of all functions defined on X which are pointwise limits, discrete limits, and equal limits of sequences of functions belonging to Φ, respectively.

In this paper we are going to examine the classes Φ^d and Φ^e for certain classes Φ. We give a description of Φ^d if $\Phi = C(X)$ where X is a topological space fulfilling some restrictions (Corollary 14, Theorem 15). In 2. we define the "Baire classes" with respect to the pointwise, discrete, and equal convergence. We describe the "discrete Baire functions" (Theorem 7) and applying some results of the theory of Baire functions we examine the relations among these classes.

Finally we characterize those classes Φ which are used as natural starting points in the above mentioned Baire classification (Theorem 16).

It is obvious that $\Phi^d \subset \Phi^e \subset \Phi^\lambda$ for every Φ. First we give a sufficient condition to have $\Phi^d = \Phi^e$. We remark that if $f \in \Phi^e$, i.e. if, for a suitable sequence (ε_n) satisfying $\varepsilon_n > 0$, $\varepsilon_n \to 0$, there are $f_n \in \Phi$ such that, for every $x \in X$, we have $|f(x) - f_n(x)| < \varepsilon_n$ if n is sufficiently large, then the same condition holds for an *arbitrary* sequence (η_n) satisfying $\eta_n > 0$, $\eta_n \to 0$ and a suitable subsequence (f_{k_n}) of (f_n) (in fact, one has to take k_n such that $\varepsilon_{k_n} < \eta_n$).

THEOREM 1. *Suppose that Φ is a subtractive lattice* (that is Φ contains the constants and, for every $f, g \in \Phi$ we have $f - g$, $\max(f, g) \in \Phi$). *In addition suppose Φ to be complete* (that is if $f_n \in \Phi$ and (f_n) converges to f uniformly then $f \in \Phi$). *Then* $\Phi^d = \Phi^e$.

PROOF. It is enough to prove $\Phi^e \subset \Phi^d$. Let $f \in \Phi^e$, then there are $f_n \in \Phi$ such that, for every $x \in X$, there exists $n_0 = n_0(x)$ with $|f(x) - f_n(x)| \leqq 2^{-n}$ if $n \geqq n_0$.

We put

$$A_k = \{x \in X: |f(x) - f_n(x)| \leqq 2^{-n} \text{ for } n \geqq k\},$$

then $A_1 \subset A_2 \subset \ldots$ and $\bigcup_{k=1}^{\infty} A_k = X$. We show that, for every k, there exists $\varphi_k \in \Phi$

such that $\varphi_k|A_k=f|A_k$. Then f is the discrete limit of the sequence (φ_k) so that $f\in\Phi^d$ will be proved.

Let k be fixed and let

$$g_0=f_1,\quad g_n=f_{n+1}-f_n\quad (n=1,2,\ldots).$$

Then $f=\sum_{n=0}^{\infty} g_n$ and for every $x\in A_k$ and $n\geqq k$ we have

$$|g_n(x)|\leqq|f_{n+1}(x)-f(x)|+|f_n(x)-f(x)|\leqq 2^{-n-1}+2^{-n}<2^{-n+1}.$$

Put

$$h_n=g_n\quad\text{for}\quad n<k$$

and

$$h_n=\operatorname{med}(g_n,2^{-n+1},-2^{-n+1})=\max(\min(g_n,2^{-n+1}),-2^{-n+1})\quad\text{for}\quad n\geqq k.$$

Since Φ is a subtractive lattice, $h_n\in\Phi$ for every n. The series $\sum_{n=0}^{\infty} h_n$ converges uniformly so that the completeness of Φ implies that $\varphi_k=\sum_{n=0}^{\infty} h_n\in\Phi$.

If $x\in A_k$ then $h_n(x)=g_n(x)$ for every n, hence $\varphi_k(x)=\sum_{n=0}^{\infty} g_n(x)=f(x)$ that is $\varphi_k|A_k=f|A_k$ which proves the theorem.

COROLLARY 2. *If X is a topological space and $\Phi=C(X)$ then $\Phi^d=\Phi^e$.*

We say that Φ is an *ordinary system* if it is a subtractive lattice and $f,g\in\Phi$ implies $fg\in\Phi$ and, in the case $g(x)\neq 0$ for $x\in X$, also $f/g\in\Phi$.

PROPOSITION 3. *If Φ is an ordinary system then so are Φ^d and Φ^e.*

PROOF. It is obvious that Φ^d and Φ^e are subtractive lattices and that $f,g\in\Phi^d$ implies $fg\in\Phi^d$. Let $f\in\Phi^d$, $f(x)\neq 0$ for $x\in X$, then f is the discrete limit of the sequence (f_n), where $f_n\in\Phi$. We put $g_n=\max\left(f_n^2,\frac{1}{n}\right)$, then $g_n\in\Phi$, $g_n>0$ and f^{-2} is the discrete limit of $g_n^{-1}\in\Phi$. Hence $f^{-2}\in\Phi^d$ so that $f^{-1}=f\cdot f^{-2}\in\Phi^d$.

Suppose $f,g\in\Phi^e$, we prove $fg\in\Phi^e$. There are functions $f_n,g_n\in\Phi$ such that $|f(x)-f_n(x)|<n^{-2}$ and $|g(x)-g_n(x)|<n^{-2}$ for every $x\in X$ and $n\geqq n_0(x)$. Hence

$$|f(x)g(x)-f_n(x)g_n(x)|\leqq|f(x)-f_n(x)|\cdot|g_n(x)|+$$
$$+|f(x)|\cdot|g(x)-g_n(x)|<n^{-2}(|g(x)|+1)+|f(x)|\cdot n^{-2}<n^{-1}$$

if $n>\max(n_0(x),2|f(x)|,2|g(x)|+2)$. That is fg is the equal limit of (f_ng_n) with $\varepsilon_n=n^{-1}$; $fg\in\Phi^e$.

Let $f\in\Phi^e$, $f(x)\neq 0$ for $x\in X$, then $f^2\in\Phi^e$ hence there are $f_n\in\Phi$ such that $|f_n(x)-f(x)^2|<n^{-3}$ if $n\geqq n_0(x)$. If $g_n=\max\left(f_n,\frac{1}{n}\right)$ then $g_n\in\Phi$, $g_n\geqq\frac{1}{n}$ and $|g_n(x)-f(x)^2|<n^{-3}$ whenever $n\geqq\max(n_0(x),2\cdot f(x)^{-2})=n_1(x)$. Thus for $h_n=g_n^{-1}$ we have $h_n\in\Phi$ and

$$|h_n(x)-f(x)^{-2}|=|g_n(x)-f(x)^2|\cdot g_n(x)^{-1}\cdot f(x)^{-2}<n^{-3}\cdot nn=n^{-1}$$

if $n \geqq n_1(x)$ which gives $f^{-2} \in \Phi^e$ whence

$$f^{-1} = f \cdot f^{-2} \in \Phi^e,$$

q. e. d.

REMARK 4. It is well-known that Φ^λ is a complete ordinary system whenever Φ is an ordinary system ([1], 5.6.3.3). We show that Φ may be a complete ordinary system without $\Phi^d = \Phi^e$ being complete.

For instance let $\Phi = C([0, 1])$ and let the set $D = \{x_n : n = 1, 2, \ldots\}$ be everywhere dense in $[0, 1]$. Then the function

$$f(x) = \begin{cases} \dfrac{1}{k} & \text{if} \quad x = x_k, \\ 0 & \text{if} \quad x \in [0, 1] - D \end{cases}$$

is the uniform limit of the functions

$$f_n(x) = \begin{cases} \dfrac{1}{k} & \text{if} \quad x = x_k, \ k \leqq n, \\ 0 & \text{if} \quad x \in [0, 1] - \{x_1, \ldots, x_n\}. \end{cases}$$

Now it is obvious that $f_n \in \Phi^d$ but $f \notin \Phi^d$ (see Theorem 13).

In addition in this case $(\Phi^d)^d \neq (\Phi^d)^e$ (see Corollary 12).

2.

During this section we suppose that Φ is a complete ordinary system.

On the analogy of the Baire classification we define the classes Φ_α, $\Phi_\alpha^{(d)}$, and $\Phi_\alpha^{(e)}$, for every countable ordinal α as follows.

Let $\Phi_0 = \Phi$, $\Phi_{\beta+1} = \Phi_\beta^\lambda$ and for a limit ordinal α let Φ_α be the smallest complete ordinary system containing the class $\bigcup_{\beta<\alpha} \Phi_\beta$. (That is let Φ_α be the intersection of all complete ordinary systems containing $\bigcup_{\beta<\alpha} \Phi_\beta$. See [1], 5.6.6.)

The classes Φ_α are complete ordinary systems for every $\alpha < \omega_1$ and for the class $\Phi_{\omega_1} = \bigcup_{\alpha<\omega_1} \Phi_\alpha$ we have $\Phi_{\omega_1}^\lambda = \Phi_{\omega_1}$.

The relation between the classes Φ_α and the classes of Baire given by the other usual (and perhaps more natural) definition $\left(\Phi_0^* = \Phi, \ \Phi_\alpha^* = \left(\bigcup_{\beta<\alpha} \Phi_\beta^*\right)^\lambda\right)$ is the following:

$$\Phi_\alpha = \Phi_\alpha^* \quad \text{if} \quad \alpha \text{ is finite},$$

$$\Phi_{\alpha+1} = \Phi_\alpha^* \quad \text{if} \quad \alpha \geqq \omega, \tag{2.1}$$

$$\Phi_\alpha \subset \Phi_\alpha^* \quad \text{if} \quad \alpha \text{ is a limit ordinal.}$$

Hence $\Phi_{\omega_1} = \bigcup_{\alpha<\omega_1} \Phi_\alpha^*$.

In order to show (2.1) we shortly summarize the basic properties of the classes Φ_α.

Denote by $\{f>c\}$ and $\{f\geqq c\}$ the sets

$$\{x\in X\colon f(x)>c\} \quad \text{and} \quad \{x\in X\colon f(x)\geqq c\}$$

respectively. For an arbitrary class Ψ of real valued functions let

$$\mathfrak{M}_\Psi=\big\{\{f>c\}\colon f\in\Psi,\ c\in\mathbf{R}\big\}$$

and

$$\mathfrak{N}_\Psi=\big\{\{f\geqq c\}\colon f\in\Psi,\ c\in\mathbf{R}\big\}.$$

Conversely, if $\mathfrak{M}$ and $\mathfrak{N}$ are arbitrary classes of subsets of X then let $[\mathfrak{M},\mathfrak{N}]$ denote the class of all functions defined on X and satisfying $\{f>c\}\in\mathfrak{M}$, $\{f\geqq c\}\in\mathfrak{N}$ for every $c\in\mathbf{R}$.

It is known that, for every complete ordinary system Ψ, we have $\Psi=[\mathfrak{M}_\Psi,\mathfrak{N}_\Psi]$ ([1], 5.6.5.1). The classes Φ_α are complete ordinary systems so that

$$\Phi_\alpha=[\mathfrak{M}_\alpha,\mathfrak{N}_\alpha], \tag{2.2}$$

where

$$\mathfrak{M}_\alpha=\mathfrak{M}_{\Phi_\alpha},\quad \mathfrak{N}_\alpha=\mathfrak{N}_{\Phi_\alpha}.$$

On the other hand, if Ψ is an ordinary system, then

$$\Psi^\lambda=[(\mathfrak{N}_\Psi^\delta)^\sigma,(\mathfrak{M}_\Psi^\sigma)^\delta]$$

where $\mathfrak{S}^\sigma$ and $\mathfrak{S}^\delta$ denote, respectively, the classes of all sets of the form $\bigcup_{i=1}^{\infty} S_i$ and $\bigcap_{i=1}^{\infty} S_i$ $(S_i\in\mathfrak{S})$ (see [2], p. 241, IX.).

It follows that

$$\mathfrak{M}_\alpha=\Big(\bigcup_{\beta<\alpha}\mathfrak{N}_\beta\Big)^\sigma \quad\text{and}\quad \mathfrak{N}_\alpha=\Big(\bigcup_{\beta<\alpha}\mathfrak{M}_\beta\Big)^\delta \tag{2.3}$$

for $1\leqq\alpha<\omega_1$. In fact, if $\alpha=\beta+1$, then we have $\mathfrak{M}_\alpha=(\mathfrak{N}_\beta^\delta)^\sigma=(\mathfrak{N}_\beta)^\sigma$ since $(\mathfrak{N}_\beta)^\delta=\mathfrak{N}_\beta$ for the complete ordinary system Φ_β ([1], 5.6.5). Similarly $\mathfrak{N}_\alpha=(\mathfrak{M}_\beta)^\delta$ and because the classes $\mathfrak{M}_\alpha$, $\mathfrak{N}_\alpha$ are increasing, we have (2.3). If α is a limit ordinal, then the definition of Φ_α and [1], 5.6.5 give (2.3), taking into account that now $\bigcup_{\beta<\alpha}\mathfrak{M}_\beta=\bigcup_{\beta<\alpha}\mathfrak{N}_\beta$ as a consequence of the inclusions

$$\mathfrak{M}_\beta\subset\mathfrak{M}_\beta^\delta=\mathfrak{N}_{\beta+1},\quad \mathfrak{N}_\beta\subset\mathfrak{N}_\beta^\sigma=\mathfrak{M}_{\beta+1}.$$

Now we prove (2.1) by transfinite induction. The cases $\alpha=0,1,2,\ldots$ are obvious. If (2.1) holds for every $\beta<\alpha$ and α is a limit ordinal, then $\Phi_\alpha\subset\Phi_\alpha^*$ follows from the fact that Φ_α^* is a complete ordinary system and

$$\Phi_\alpha^*\supset\bigcup_{\beta<\alpha}\Phi_\beta^*=\bigcup_{\beta<\alpha}\Phi_\beta.$$

Let $\Psi=\bigcup_{\beta<\alpha}\Phi_\beta$, then by [1], 5.6.5 we have

$$\Phi_\alpha=[\mathfrak{M}_\Psi^\sigma,\mathfrak{N}_\Psi^\delta]$$

and hence

$$\Phi_{\alpha+1}=\Phi_\alpha^\lambda=\big[\big((\mathfrak{N}_\Psi^\delta)^\delta\big)^\sigma,\big((\mathfrak{M}_\Psi^\sigma)^\sigma\big)^\delta\big]=[(\mathfrak{N}_\Psi^\delta)^\sigma,(\mathfrak{M}_\Psi^\sigma)^\delta]=\Psi^\lambda=\Phi_\alpha^*$$

which is the second statement of (2.1). If $\alpha=\beta+1$, then $\Phi_{\alpha+1}=(\Phi_{\beta+1})^\lambda=(\Phi_\beta^*)^\lambda=\Phi_\alpha^*$, q. e. d.

The validity of the formulas (2.3) motivates the somewhat artificial definition of the classes Φ_α.

Now we define the classes $\Phi_\alpha^{(d)}$: let $\Phi_0^{(d)}=\Phi$ and $\Phi_\alpha^{(d)}=(\bigcup_{\beta<\alpha}\Phi_\beta^{(d)})^d$ for every $\alpha<\omega_1$. For the class $\Phi_{\omega_1}^{(d)}=\bigcup_{\alpha<\omega_1}\Phi_\alpha^{(d)}$ we have $(\Phi_{\omega_1}^{(d)})^d=\Phi_{\omega_1}^{(d)}$.

The class $\Phi_{\omega_1}^{(d)}$ is that of the "discrete Baire functions".

The definition of the classes $\Phi_\alpha^{(e)}$ is similar: $\Phi_0^{(e)}=\Phi$ and $\Phi_\alpha^{(e)}=(\bigcup_{\beta<\alpha}\Phi_\beta^{(e)})^e$ for every $\alpha<\omega_1$. It follows from Proposition 3 that the classes $\Phi_\alpha^{(d)}$ and $\Phi_\alpha^{(e)}$ are ordinary systems for every α.

In order to describe the connection between the classes $\Phi_\alpha^{(d)}$ and Φ_α we need the following lemmas.

LEMMA 5. *Let $\alpha>0$. Then the class $\mathfrak{M}_\alpha$ has the reduction property, i.e. for every $A_1, \ldots, A_n\in\mathfrak{M}_\alpha$ there exist pairwise disjoint sets $C_1, \ldots, C_n\in\mathfrak{M}_\alpha$ such that $C_1\subset A_1, \ldots, C_n\subset A_n$ and $\bigcup_{i=1}^{n} C_i=\bigcup_{i=1}^{n} A_i$.*

PROOF. We prove it by induction. The case $n=1$ is obvious. Suppose the assertion is true for $n-1$ and let $A_1, \ldots, A_n\in\mathfrak{M}_\alpha$. By the hypothesis there are disjoint sets $C_1, \ldots, C_{n-1}\in\mathfrak{M}_\alpha$ with $C_i\subset A_i$ $(i=1, \ldots, n-1)$ and

$$\bigcup_{i=1}^{n-1} C_i=\bigcup_{i=1}^{n-1} A_i=B.$$

By (2.3)

$$A_n=\bigcup_{k=1}^{\infty} D_k \quad \text{and} \quad B=\bigcup_{k=1}^{\infty} B_k$$

where $D_k, B_k\in\bigcup_{\beta<\alpha}\mathfrak{N}_\beta$. Let

$$C_n=\bigcup_{k=1}^{\infty}\Big(D_k-\bigcup_{i=1}^{k} B_i\Big), \quad B'=\bigcup_{k=1}^{\infty}\Big(B_k-\bigcup_{i=1}^{k-1} D_i\Big).$$

Since $\mathfrak{N}_\beta$ and $\mathfrak{M}_\beta$ are lattices for every β ([1], 5.6.3.1), the elements of $\mathfrak{N}_\beta$ are the complements of the elements of $\mathfrak{M}_\beta$, and $\mathfrak{N}_\beta\subset\mathfrak{M}_\alpha$, $\mathfrak{M}_\beta\subset\mathfrak{N}_\alpha$ $(\beta<\alpha)$, we have $C_n, B'\in\mathfrak{M}_\alpha$, $C_n\subset A_n$, $B'\subset B$, $C_n\cap B'=\emptyset$, $C_n\cup B'=A_n\cup B$. Thus the sets $C_i'=C_i\cap B'$ $(i=1, \ldots, n-1)$, C_n satisfy the requirements of the lemma.

LEMMA 6. *If $A\in\mathfrak{M}_\alpha\cap\mathfrak{N}_\alpha$ then the characteristic function $k_A\in\Phi_\alpha^{(d)}$.*

PROOF. First we prove that for $\alpha>0$ and for $A, B\in\mathfrak{N}_\alpha$, $A\cap B=\emptyset$, there is a $D\in\mathfrak{M}_\alpha\cap\mathfrak{N}_\alpha$ such that $A\subset D$ and $D\cap B=\emptyset$. By [1], 5.6.3.1 we have $X-A$, $X-B\in\mathfrak{M}_\alpha$, hence by Lemma 5 there exist $C\subset X-A$, $D\subset X-B$, $C\cap D=\emptyset$, $C\cup D=(X-A)\cup(X-B)=X$ and $C, D\in\mathfrak{M}_\alpha$. Thus $A\subset D$, $B\cap D=\emptyset$ and $D=X-C\in\mathfrak{M}_\alpha\cap\mathfrak{N}_\alpha$.

Now we prove $k_A\in\Phi_\alpha^{(d)}$ by transfinite induction. If $\alpha=0$, then $A\in\mathfrak{M}_0\cap\mathfrak{N}_0$ and (2.2) imply $k_A\in\Phi=\Phi_0^{(d)}$. If $\alpha=1$, then by (2.3)

$$A=\bigcup_{n=1}^{\infty} A_n, \quad X-A=\bigcup_{n=1}^{\infty} B_n \tag{2.4}$$

where $A_n, B_n\in\mathfrak{N}_0$. Since $\mathfrak{N}_0$ is a lattice, we can suppose $A_1\subset A_2\subset\ldots$, $B_1\subset B_2\subset\ldots$. For every n there are $f_n, g_n\in\Phi$ satisfying $f_n(x)=0$ for $x\in A_n$, $f_n(x)>0$ for $x\in X-A_n$,

$g_n(x)=0$ for $x\in B_n$, $g_n(x)>0$ for $x\in X-B_n$. Then $h_n=g_n/(f_n+g_n)\in\Phi$ and k_A is the discrete limit of (h_n).

Assume now $\alpha>1$, and suppose that the lemma is proved for $\beta<\alpha$. We have again (2.4) with $A_n, B_n\in\bigcup_{\beta<\alpha}\mathfrak{N}_\beta$, and the latter system being a lattice, we can suppose as above $A_1\subset A_2\subset\ldots$, $B_1\subset B_2\subset\ldots$. For every n we have a set $D_n\in\bigcup_{\beta<\alpha}(\mathfrak{M}_\beta\cap\mathfrak{N}_\beta)$ for which $A_n\subset D_n$, $D_n\cap B_n=\emptyset$. By the induction hypothesis $k_{D_n}\in\bigcup_{\beta<\alpha}\Phi_\beta^{(d)}$. Since k_A is the discrete limit of k_{D_n} we have $k_A\in\Phi_\alpha^{(d)}$, q. e. d.

THEOREM 7. *$f\in\Phi_{\omega_1}^{(d)}$ if and only if $f\in\Phi_{\omega_1}$ and there are a decomposition $X=\bigcup_{n=1}^{\infty}A_n$ and functions $f_n\in\Phi$ such that $f|A_n=f_n|A_n$ $(n=1, 2, \ldots)$.*

PROOF. We prove the part "only if" by transfinite induction. If $f\in\Phi_0^{(d)}=\Phi$ then trivially $f\in\Phi_{\omega_1}$ and the decomposition $X=A$ will do. If the assertion holds for every $\beta<\alpha$ and $f\in\Phi_\alpha^{(d)}$ then there are functions $f_n\in\bigcup_{\beta<\alpha}\Phi_\beta^{(d)}$ such that f is the discrete limit of f_n. Hence $X=\bigcup_{n=1}^{\infty}B_n$ where $B_n=\{x\in X: f(x)=f_n(x)\}$. By the induction hypothesis $f_n\in\Phi_{\omega_1}$ and thus $f\in\Phi_{\omega_1}^d=\Phi_{\omega_1}$, and there exist decompositions $X=\bigcup_{i=1}^{\infty}A_{ni}$ and functions $g_{ni}\in\Phi$ such that $f_n|A_{ni}=g_{ni}|A_{ni}$ $(n, i=1, 2, \ldots)$. Consequently for the decomposition $X=\bigcup_{n,i=1}^{\infty}(B_n\cap A_{ni})$ we have $f|B_n\cap A_{ni}=g_{ni}|B_n\cap A_{ni}$, q. e. d.

The part "if": suppose $X=\bigcup_{n=1}^{\infty}A_n$, $f_n\in\Phi$, $f\in\Phi_{\omega_1}$ and $f|A_n=f_n|A_n$. Since $f, f_n\in\Phi_{\omega_1}$, the set

$$B_n=\{x\in X: f(x)=f_n(x)\}$$

belongs to $\bigcup_{\alpha<\omega_1}\mathfrak{N}_\alpha$. Let $C_n=B_n-\bigcup_{i=1}^{n-1}B_i$, then $X=\bigcup_{n=1}^{\infty}C_n$ is a disjoint decomposition for which $C_n\in\bigcup_{\alpha<\omega_1}\mathfrak{N}_\alpha$ and $f|C_n=f_n|C_n$. Consequently f is the discrete limit of the functions $g_n=\sum_{i=1}^{n}f_i\cdot k_{C_i}$. If $C_i\in\mathfrak{N}_\alpha$ then $C_i\in\mathfrak{M}_{\alpha+1}\cap\mathfrak{N}_{\alpha+1}$ so that by Lemma 6 $k_{C_i}\in\Phi_{\alpha+1}^{(d)}$. Since $\Phi_{\alpha+1}^{(d)}$ is an ordinary system we have $g_n\in\Phi_{\alpha+1}^{(d)}$ and hence $f\in\Phi_{\omega_1}^{(d)}$, q. e. d.

COROLLARY 8. *If X is a topological space and $\Phi=C(X)$ then $f\in\Phi_{\omega_1}^{(d)}$ if and only if f is Baire measurable and if the graph of f can be covered by countable many graphs of functions belonging to $C(X)$.*

COROLLARY 9. *If $\Phi=C([0, 1])$ then there are Baire 1 functions not belonging to $\Phi_{\omega_1}^{(d)}$.*

PROOF. Let f be monotone on $[0, 1]$ such that the discontinuities of f are everywhere dense in $[0, 1]$; f is a Baire 1 function. If $g(x)$ is an arbitrary continuous function on $[0, 1]$ then the set $\{x\in[0, 1]: f(x)=g(x)\}$ is nowhere dense on $[0, 1]$. Hence for every sequence of continuous functions f_n the sets $\{x\in[0, 1]: f(x)=f_n(x)\}$ cannot cover $[0, 1]$ by the Baire category theorem. Thus by Theorem 7 $f\notin\Phi_{\omega_1}^{(d)}$.

THEOREM 10. *For every* $\alpha<\omega_1$ *we have* $\Phi_\alpha\subset(\Phi_\alpha^{(d)})^e$.

PROOF. Let $f\in\Phi_\alpha$, n be a natural number. The sets

$$A=\{x\in X: f\{x\}<-n \quad \text{or} \quad f(x)>n\},$$

$$A_i=\left\{x\in X: \frac{i-1}{n}<f(x)<\frac{i+1}{n}\right\}$$

$(-n^2\leqq i\leqq n^2)$ belong to $\mathfrak{M}_\alpha$, hence by Lemma 5 there are pairwise disjoint sets $C, C_i\in\mathfrak{M}_\alpha$ such that $C\subset A$, $C_i\subset A_i$ $(-n^2\leqq i\leqq n^2)$ and

$$X=\bigcup_{i=-n^2}^{n^2}A_i\cup A=\bigcup_{i=-n^2}^{n^2}C_i\cup C.$$

Thus $C, C_i\in\mathfrak{M}_\alpha\cap\mathfrak{N}_\alpha$.

Let $f_n=\sum_{i=-n^2}^{n^2}\frac{i}{n}\cdot k_{C_i}$, then Lemma 6 implies $f_n\in\Phi_\alpha^{(d)}$. It is easy to see that f is the equal limit of f_n, thus $f\in(\Phi_\alpha^{(d)})^e$, q. e. d.

COROLLARY 11. *For every* $\alpha<\omega_1$, *we have* $\Phi_\alpha^{(e)}\subset\Phi_{\alpha+1}\subset\Phi_{\alpha+2}^{(e)}$; *hence* $\Phi_{\omega_1}^{(e)}=\Phi_{\omega_1}$.

PROOF. $\Phi_\alpha^{(d)}\subset\Phi_\alpha^{(e)}\subset\Phi_\alpha^*$ can be easily proved by transfinite induction. Thus by Theorem 10 $\Phi_\alpha^{(e)}\subset\Phi_\alpha^*\subset\Phi_{\alpha+1}\subset(\Phi_{\alpha+1}^{(d)})^e\subset(\Phi_{\alpha+1}^{(e)})^e=\Phi_{\alpha+2}^{(e)}$.

COROLLARY 12. *If* $\Phi=C([0,1])$ *then* $(\Phi^d)^d\neq(\Phi^d)^e$.

PROOF. For the class Φ_1 (the class of Baire 1 functions) $\Phi_1\subset(\Phi^d)^e$ by Theorem 10, but $\Phi_1\not\subset(\Phi^d)^d$ by Corollary 9.

3.

It is well-known that the functions of the first Baire class $(C(X))^\lambda=C^\lambda(X)$ can be characterized, under suitable hypotheses on the topological space X, by the fact that $f|A$ has at least one point of continuity for each closed set $\emptyset\neq A\subset X$. Our purpose is to prove a similar characterization theorem for the first discrete Baire class $(C(X))^d=C^d(X)$.

THEOREM 13. *If* X *is a Baire space and* $f\in C^d(X)$, *then the points of discontinuity of* f *constitute a rare (nowhere dense) subset of* X.

PROOF. Assume $f_n\in C(X)$, $X=\bigcup_{n=1}^{\infty}A_n$, $A_1\subset A_2\subset\ldots$, $f_n|A_n=f|A_n$. For an arbitrary open set $G\neq\emptyset$, there is an open set $\emptyset\neq G'\subset G$ such that $G'\subset\overline{A}_n$ for a suitable integer n. Since $f_k|A_n=f_n|A_n$ for $k\geqq n$, clearly $f_k|G'=f_n|G'$ $(k\geqq n)$ by the continuity of f_k and f_n, and then $f|G'=f_n|G'$. Hence f has no point of discontinuity in G', q. e. d.

COROLLARY 14. *If* X *is a Hausdorff space and* $f\in C^d(X)$, *then, for any compact subspace* $K\subset X$, *the points of discontinuity of the restriction* $f|K$ *constitute a rare subset of the subspace* K.

Now this statement admits the following converse:

THEOREM 15. *Let X be a σ-compact, perfectly normal Hausdorff space and f a real valued function on X such that, for any compact set $K\subset X$, the points of discontinuity of $f|K$ constitute a rare set in the subspace K. Then $f\in C^d(X)$.*

PROOF. Let $K\subset X$ be compact. Define $K_0=K$, and if the compact sets K_β are defined for $\beta<\alpha$ and $K_{\beta_1}\supset K_{\beta_2}$ for $\beta_1<\beta_2<\alpha$, define K_α to be the closure of the set of the points of discontinuity of the restriction

$$f\Big|\bigcap_{\beta<\alpha} K_\beta.$$

Then clearly $K_\alpha\subset K_\beta$ for $\beta<\alpha$ and K_α is compact. Moreover, $K_\alpha\neq K_\beta$ for $\beta<\alpha$ if $\beta<\alpha$ implies $K_\beta\neq\emptyset$ since, by hypothesis, K_α is a rare subset of the (non-empty) compact set $\bigcap_{\beta<\alpha} K_\beta$.

Therefore the sets K_α must be empty from a certain ordinal on; let $\gamma=\gamma(f,K)$ be the smallest ordinal such that $K_\gamma=\emptyset$.

We show $f|K\in C^d(K)$ for any compact $K\subset X$ by means of transfinite induction with respect to $\gamma(f,K)$. This is true if $\gamma(f,K)=0$ (since then $K=\emptyset$) and for $\gamma(f,K)=1$ (since then $f|K\in C(K)$). Assume that the assertion holds for $\gamma(f,K)<\alpha$ and suppose $\gamma(f,K)=\alpha$. Then $f|F$ is continuous where

$$F=\bigcap_{\beta<\alpha} K_\beta\neq\emptyset.$$

If $K'\subset K-F$ is compact, then $K'_\beta=(K')_\beta\subset K_\beta\cap K'$ is proved by an easy transfinite induction for each ordinal β, hence

$$\bigcap_{\beta<\alpha} K'_\beta\subset\bigcap_{\beta<\alpha} K_\beta\cap K'=F\cap K'=\emptyset$$

which implies $K'_\beta=\emptyset$ for a $\beta<\alpha$ so that $\gamma(f,K')<\alpha$ and $f|K'\in C^d(K')$ by the induction hypothesis.

By the perfect normality of X we have

$$K-F=\bigcup_{n=1}^{\infty} C_n$$

where C_n is compact and $C_1\subset C_2\subset\ldots$, moreover

$$C_n=\bigcap_{i=1}^{\infty} G_{n,i} \tag{3.1}$$

where $G_{n,i}$ is open and $G_{n,1}\supset G_{n,2}\supset\ldots$. Since $f|C_n\in C^d(C_n)$, there are functions $f_{ni}\in C(C_n)$ such that $f|C_n$ is the discrete limit of f_{ni} for $i\to\infty$. By the Tietze—Urysohn theorem there is $g_i\in C(K)$ satisfying

$$g_i|C_1=f_{1i},\quad g_i|C_k-G_{k-1,i}=f_{ki}|C_k-G_{k-1,i} \tag{3.2}$$

for $k=2,\ldots,i$, and

$$g_i|F=f|F.$$

Clearly $f|K$ is the discrete limit of the sequence (g_i), whence $f\in C^d(K)$.

Now by the σ-compactness of X we have

$$X=\bigcup_{n=1}^{\infty} C_n$$

where C_n is compact and $C_1 \subset C_2 \subset \ldots$, and (3.1) is again valid with open sets $G_{n,i}$ such that $G_{n,1} \supset G_{n,2} \supset \ldots$. A similar construction as above furnishes functions $g_i \in C(X)$ satisfying (3.2) for $k=2, \ldots, i$; then f is the discrete limit of (g_i), hence $f \in C^d(X)$. Q. e. d.

The hypotheses of Theorem 15 are fulfilled e.g. if X is a σ-compact metrizable space. It would be interesting to know whether a similar statement holds for larger classes of topological spaces, possibly by replacing compact subspaces by other ones (closed, Baire, etc.).

In this direction, the following example may be useful. Let X be an uncountable set, $p \in X$, let the points $x \in X - \{p\}$ be isolated, and the neighbourhoods of p be the sets V such that $p \in V$ and $X - V$ is countable. Then X is a Lindelöf T_3–space. Moreover, if V_n $(n=1, 2, \ldots)$ is a neighbourhood of p, then $\bigcap_{n=1}^{\infty} V_n$ is a neighbourhood of p as well, and p is an accumulation point of X. Hence $f \in C(X)$ iff the real valued function f is constant on a neighbourhood of p. Consequently

$$C^{\lambda}(X) = C^d(X) = C(X).$$

However, each real valued function defined on X fulfils the condition that the points of discontinuity of $f|A$ constitute, for *every* subspace $A \subset X$, a rare set in A.

4.

In the theory of Baire classes and discrete Baire classes, a fundamental role was played by complete ordinary systems. As we mentioned above, we know from [1] that these systems admit a characterization by means of the sets $\{x \in X : f(x) > c\}$ and $\{x \in X : f(x) \geqq c\}$. Now we will add to this further characterizations by means of the notion of a composition-closed class, introduced by the first of the authors. For our purposes, it suffices to recall the following definitions (see [3]): a class Φ of real valued functions defined on X is said to be *countably strongly composition-closed* if the following is true: whenever $f_i \in \Phi$ $(i=1, 2, \ldots)$, $h: X \to \mathbf{R}^{\mathbf{N}}$ is defined by $h(x)=(f_i(x))$, and $k: h(X) \to \mathbf{R}$ is continuous, then $k \circ h \in \Phi$. Similarly, Φ is said to be *finitely strongly composition-closed* if the same is true for $n \in N$, $f_i \in \Phi$ $(i=1, \ldots, n)$, $h: X \to \mathbf{R}^n$ defined again by $h(x)=(f_i(x))$, and $k \in C(h(X))$. Finally Φ is said to be *finitely composition-closed* if $n \in \mathbf{N}$, $f_i \in \Phi$ $(i=1, \ldots, n)$, $h: X \to \mathbf{R}^n$, $h(x)=(f_i(x))$, $k \in C(\mathbf{R}^n)$ implies $k \circ h \in \Phi$.

Now we can prove the following

THEOREM 16. *Let Φ be a class of real valued functions defined on X. The following statements are equivalent:*

(a) *Φ is a complete ordinary system,*

(b) *there is a σ-lattice $\mathfrak{M}$ in X such that $f \in \Phi$ iff*

$$\{x \in X : f(x) > c\} \in \mathfrak{M}, \quad \{x \in X : f(x) < c\} \in \mathfrak{M}$$

for each $c \in \mathbf{R}$,

(c) *Φ is countably strongly composition-closed,*

(d) *Φ is finitely strongly composition-closed and complete,*

(e) *Φ is finitely composition-closed, complete, and $f \in \Phi$, $f(x) \neq 0$ for $x \in X$ implies $1/f \in \Phi$,*

(f) *Φ is complete, it contains all constants, $f, g\in\Phi$ implies $f+g, fg\in\Phi$, and $f\in\Phi$, $f(x)\neq 0$ for $x\in X$ implies $1/f\in\Phi$.*

REMARK. (a)$\Leftrightarrow$(b) was proved in [1], 5.6.3.1 and 5.6.5, and (e)$\Leftrightarrow$(f) in [4], 1.14.

PROOF. (a)$\Rightarrow$(b): [1], 5.6.3.1 and 5.6.5.

(b)$\Rightarrow$(c): Let $f_i\in\Phi$ $(i\in\mathbf{N})$, $h:X\to\mathbf{R}^{\mathbf{N}}$, $k:h(X)\to\mathbf{R}$ be as in the definition. We show

$$\{x\in X:\ k(h(x))>c\}\in\mathfrak{M},\quad \{x\in X:\ k(h\{x\})<c\}\in\mathfrak{M}$$

for $c\in\mathbf{R}$. Clearly these sets have the form

$$h^{-1}(G\cap h(X))=h^{-1}(G)$$

where

$$G\cap h(X)=\{y\in h(X):\ k(y)>c\}$$

or

$$G\cap h(X)=\{y\in h(X):\ k(y)<c\}$$

respectively, and G is open in $\mathbf{R}^{\mathbf{N}}$. Now

$$G=\bigcup_{n=1}^{\infty}G_n,\quad G_n=\bigtimes_{i=1}^{\infty}I_{ni}$$

where $I_{ni}=(a_{ni},b_{ni})$ for $i\in F$, $I_{ni}=\mathbf{R}$ for $i\in\mathbf{N}-F$, and F is a finite subset of $\mathbf{N}$. Hence

$$h^{-1}(G_n)=\bigcap_{i\in F}(A_{ni}\cap B_{ni}),$$

$$A_{ni}=\{x\in X:\ f_i(x)>a_{ni}\}\in\mathfrak{M},$$

$$B_{ni}=\{x\in X:\ f_i(x)<b_{ni}\}\in\mathfrak{M},$$

so that $h^{-1}(G_n)\in\mathfrak{M}$ and $h^{-1}(G)=\bigcup_{n=1}^{\infty}h^{-1}(G_n)\in\mathfrak{M}$.

(c)$\Rightarrow$(d): Countable strong composition-closedness clearly implies finite strong composition-closedness, and also the completeness by [3], (2.7) and (2.11) (in fact, (2.11) holds for countably weakly composition-closed classes and a fortiori for countably strongly composition-closed ones).

(d)$\Rightarrow$(e)$\Rightarrow$(f): Obvious.

(f)$\Rightarrow$(a): It suffices to show that Φ is a lattice under the hypotheses of (f) or, more precisely, that $f\in\Phi$ implies $|f|\in\Phi$. Now if f is bounded, $|f|\leqq c$, then the function $|u|$ is the limit of a uniformly convergent sequence of polynomials $p_n(u)$ on $[-c,c]$; thus $p_n\circ f\in\Phi$ for $f\in\Phi$ and by the completeness $|f|\in\Phi$. In the general case we take $g=f/(1+f^2)\in\Phi$, then $|g|=|f|/(1+f^2)\in\Phi$ and $|f|=(1+f^2)|g|\in\Phi$. Q. e. d.

REFERENCES

[1] AUMANN, G.: *Reelle Funktionen*, Berlin—Göttingen—Heidelberg, 1954.
[2] HAUSDORFF, F.: *Mengenlehre*, Berlin—Leipzig, 1927.
[3] CSÁSZÁR, Á.: Function classes, compactifications, real-compactifications, *Ann. Univ. Budapest, Sect. Math.* **17** (1974), 139—156.
[4] ISBELL, J. R.: Algebras of uniformly continuous functions, *Ann. of Math.* **68** (1958), 96—125

Eötvös Loránd University Budapest, Hungary

(Received July 5, 1976)

Studia Scientiarum Mathematicarum Hungarica 10 (1975)

INDEX

Printed in Hungary

A kiadásért felel az Akadémiai Kiadó igazgatója — Műszaki szerkesztő: Botyánszky Pál
A kézirat a nyomdába érkezett: 1978. I. 5. — Terjedelem: 17,75 (A/5) ív, 2 ábra

78-94 — Szegedi Nyomda — F. v.: Dobó József igazgató

Die *Studia Scientiarum Mathematicarum Hungarica* ist eine Halbjahrsschrift der Ungarischen Akademie der Wissenschaften. Sie veröffentlicht Originalbeiträge aus dem Bereich der Mathematik in deutscher, englischer, französischer oder russischer Sprache. Es erscheint jährlich ein Band.

Adresse der Redaktion: 1053 Budapest V., Reáltanoda u. 13—15, Ungarn.
Technischer Redaktor: E. Deák

Abonnementspreis pro Band (pro Jahr): $ 16.00. Bestellbar bei Buch- und Zeitungs-Aussenhandelsunternehmen *Kultúra* (Budapest 62, P. O. B. 149), oder bei den Vertretungen im Ausland.
Austauschabmachungen können mit der Bibliothek des Matematischen Instituts (1053 Budapest V., Reáltanoda u. 13—15) getroffen werden.
Die zur Veröffentlichung bestimmten Manuskripte sind in zwei Exemplaren an die Redaktion zu schicken.

Studia Scientiarum Mathematicarum Hungarica est une revue biannuelle de l'Académie Hongroise des Sciences publiant des essais originaux, en francais, anglais, allemand ou russe, du domaine des mathématiques.

Rédaction: 1053 Budapest V., Reáltanoda u. 13—15, Hongrie.
Rédacteur téchnique: E. Deák

Le prix de l'abonnement: $ 16.00 par an (volume). On s'abonne chez *Kultúra*, Société pour le Commerce de Livres et Journaux (Budapest 62, P. O. B. 149) ou chez ses représentants à l'étranger.
Pour établir des relations d'échange on est prié de s'adresser à la Bibliothèque de l'Institut de Mathématique (1053 Budapest V., Reáltanoda u. 13—15).
On est prié 'envoyer les articles destinés à la publication en deux exemplaires à l'adresse de la Rédaction

---

Studia Scientiarum Mathematicarum Hungarica — выходит два раза в год в Издании Академии Наук Венгрии. Журнал публикует оригинальные исследования в области математики на немецком, английском, французком и русском языках. Отдельвные вышуски составляют ежегодно одни том.

Адрес редакции: 1053 Budapest V., Reáltanoda u. 13—15, Венгрия.
Технический редактор: E. Deák.

Подписная цена на год (за одии том): ? 16.00. Подписка н журнал принимается Внешнеторговым предприятием „Култура" (Budapest 62, P. O. B. 149) или его представительствами за границей.
По пиводу отношения обмена просим обрашаться к Библиотеке Института Математики (1953 Budapest V., Reáltanoda u. 13—15).
Работы, преднасначенные для опубликования в журнале следует направлять по адресу редакции в двух экземплярах.

All the reviews of the Hungarian Academy of Sciences may be obtained among others from the following bookshops:

ALBANIA

Ndermarja Shtetnore e Botimeve
Tirana

AUSTRALIA

A. Keesing
Box 4886, GPO
Sidney

AUSTRIA

Globus Buchvertrieb
Salzgries 16
Wien I.

BELGIUM

Office International de Librairie
30, Avenue Marnix
Bruxelles 5
Du Monde Entier
5, Place St. Jean
Bruxelles

BULGARIA

Raznoiznos
1 Tzar Assen
Sofia

CANADA

Pannonia Books
2 Spadina Road
Toronto 4, Ont.

CHINA

Waiwen Shudian
Peking
P.O.B. Nr. 88.

CHECHOSLOVAKIA

Artia A. G.
Ve Smeckách 30
Praha II.
Postova Novinova Sluzba
Dovoz tisku
Vinohradska 46
Praha 2
Postova Novinova Sluzba
Dovoz tlace
Leningradska 14
Bratislava

DENMARK

Ejnar Munksgaard
Nörregade 6
Kopenhagen

FINLAND

Akateeminen Kirjakauppa
Keskuskatu 2
Helsinki

FRANCE

Office International de Documentation et Libraire
48, rue Gay Lussac
Paris 5

GERMAN DEMOCRATIC REPUBLIC

Deutscher Buchexport und Import
Leninstraße 16.
Leipzig C. I.
Zeitungvertriebsamt
Clara Zetkin Straße 62.
Berlin N. W.

GERMAN FEDERAL REPUBLIC

Kunst und Wissen
Eich Bieber
Postfach 46.
7 Stuttgart S.

GREAT BRITAIN

Collet's' Subscription Dept.
44–45 Museum Street
London W. C. I.
Robert Maxwell and Co. Ltd.
Waynflete Bldg. The Plain
Oxford

HOLLAND

Swetz and Zeitlinger
Keizersgracht 471–487
Amsterdam C.
Martinus Nijhof
Lange Voorhout 9
The Hague

INDIA

Current Technical Literature
Co. Private Ltd.
Head Office:
India House OPP.
GPO Post Box 1374
Bombay I.

ITALY

Santo Vanasia
71 Via M. Macchi
Milano
Libreria Commissionaria Sanson
Via La Marmora 45
Firenze

JAPAN

Nauka Ltd.
2 Kanada-Zimbocho 2-ehome
Chiyoda-ku
Tokyo
Maruzen and Co. Ltd.
P.O. Box 605
Tokyo
Far Eastern Booksellers
Kanada P. O. Box 72
Tokyo

KOREA

Chulpanmul
Korejskoje Obschestvo po Exportu Importu Proizvedenij Pechati
Phenjan

NORWAY

Johan Grund Tanum
Karl Johansgatan 43
Oslo

POLAND

Export- und Import- Unternehmen RUCH
ul. Wilcza 46.
Warszawa

ROUMANIA

Cartimex
Str. Aristide Briand 14–18.
Bucuresti

SOVIET UNION

Mezhdunarodnaja Kniga
Moscow
G–200

SWEDEN

Almquist and Wiksell
Gamla Brogatan 26
Stockholm

USA

Stechert Hafner Inc.
31 East 10th Street
New York 3 N. Y
Walter J. Johnson
111 Fifth Avenue
New York 3 N. Y.

VIETNAM

Xunhasaba
Service d'Export et d'Import des Livres et Périodiques
19. Tran Quoc Toan
Hanoi

YUGOSLAVIA

Forum
Vojvode Misiva broj 1.
Novi Sad
Jugoslovenska Kniga
Terazije 27.
Beograd

31. I. 1979. **Index: 26 748**