ANNALES
UNIVERSITATIS SCIENTIARUM BUDAPESTINENSIS DE ROLANDO EÖTVÖS NOMINATAE

SECTIO MATHEMATICA

TOMUS XLVI.

REDIGIT
Á. CSÁSZÁR

ADIUVANTIBUS
L. BABAI, A. BENCZÚR, K. BEZDEK., M. BOGNÁR, K. BÖRÖCZKY,
I. CSISZÁR, J. DEMETROVICS, A. FRANK, J. FRITZ, E. FRIED, A. HAJNAL,
G. HALÁSZ, A. IVÁNYI, I. KÁTAI, P. KOMJÁTH, M. LACZKOVICH,
L. LOVÁSZ, GY. MICHALETZKY, J. MOLNÁR, P. P. PÁLFY,
GY. PETRUSKA, A. PRÉKOPA, A. RECSKI, A. SÁRKÖZY, F. SCHIPP,
Z. SEBESTYÉN, L. SIMON, GY. SOÓS, J. SURÁNYI, G. STOYAN,
J. SZENTHE, G. SZÉKELY, L. VARGA

2003

ANNALES

Universitatis Scientiarum Budapestinensis de Rolando Eötvös nominatae

SECTIO BIOLOGICA
incepit anno MCMLVII
SECTIO CHIMICA
incepit anno MCMLIX
SECTIO CLASSICA
incepit anno MCMXXIV
SECTIO COMPUTATORICA
incepit anno MCMLXXVIII
SECTIO GEOGRAPHICA
incepit anno MCMLXVI
SECTIO GEOLOGICA
incepit anno MCMLVII
SECTIO GEOPHYSICA ET METEOROLOGICA
incepit anno MCMLXXV
SECTIO HISTORICA
incepit anno MCMLVII
SECTIO IURIDICA
incepit anno MCMLIX
SECTIO LINGUISTICA
incepit anno MCMLXX
SECTIO MATHEMATICA
incepit anno MCMLVIII
SECTIO PAEDAGOGICA ET PSYCHOLOGICA
incepit anno MCMLXX
SECTIO PHILOLOGICA
incepit anno MCMLVII
SECTIO PHILOLOGICA HUNGARICA
incepit anno MCMLXX
SECTIO PHILOLOGICA MODERNA
incepit anno MCMLXX
SECTIO PHILOSOPHICA ET SOCIOLOGICA
incepit anno MCMLXII

EINIGE EXTREMUMAUFGABEN FÜR D–V ZELLENSYSTEME VON DISKRETEN PUNKTSYSTEMEN

von

JENŐ HORVÁTH und ÁGOTA H. TEMESVÁRI

Herrn Professor J. Molnár seinem 80. Geburtstag gewidmet

(Juli 13, 2000)

1. Einleitung

1.1. Wir betrachten Punktsysteme in Ebenen konstanter Krümmung, die keinen Häufungspunkt haben. Im folgenden werden diese Punktsysteme diskret genannt.

Es sei $\{P_i\}$ ein diskretes Punktsystem und $P_k \in \{P_i\}$. Es bezeichne D_k die Menge derjenigen Punkte der Ebene, deren Abstand von P_k kleiner oder gleich dem Abstand von den anderen Punkten des Punktsystems ist. Die Menge D_k ist die zum Punkt P_k gehörige Dirichlet–Voronoische Zelle (kurz D–V Zelle). Die Menge der zum Punktsystem $\{P_i\}$ gehörigen D–V Zellen bilden eine normale Zerlegung der Ebene in konvexe Vielecken, wenn jede Zelle endlich ist.

Die Punkte P_j, $P_k \in \{P_i\}$ $j \neq k$ sind benachbart, wenn die entsprechenden D–V Zellen D_j und D_k eine gemeinsame Seite haben. Mit Q_{k1}, $Q_{k2}, \ldots$ $\ldots, Q_{ks}$ bezeichnen wir die Ecken von D_k.

In dieser Arbeit untersuchen wir die Extrema der Inhalte und Umfänge der D–V Zellen von diskreten Punktsystemen, wobei die Punktsysteme bestimmte Bedingungen erfüllen. Wir charakterisieren die Punktsysteme nach diesen Bedingungen.

J. Molnár [5] hat Kreissysteme unter angegebenen Bedingungen untersucht. Sein Problem kann man auch folgenderweise formulieren. Die Kreismittelpunkte werden als Punkte des Punktsystems $\{P_i\}$ betrachtet. Wir nehmen an, dass $P_jP_k \geq 2r$ für beliebige Punkte P_j, $P_k \in \{P_i\}$, $j \neq k$ und $P_kQ_{kq} \geq R$ $(q = 1, 2, \ldots, s)$ für jeden Punkt $P_k \in \{P_i\}$, wobei $0 < r < R$.

Die Grundaufgabe ist, den Inhalt (Umfang) der D–V Zelle D_k zu minimalisieren und die Zahlen r, R finden, für die jede D–V Zelle von $\{P_i\}$ den minimalen Inhalt (Umfang) hat.

J. Horváth und M. I. Stogrin [3] untersuchten solche diskrete Punktsysteme $\{P_i\}$, für die $P_jP_k \leq 2r$ für beliebige benachbarte Punkte P_j, $P_k \in \{P_i\}$ und $P_kQ_{kq} \leq R$ $(q = 1, 2, \dots, s)$ für jeden Punkt $P_k \in \{P_i\}$.

In diesem Fall ist die Frage, was das Maximum der Inhalte (Umfänge) der D–V Zellen ist, und für welche r und R Punktsysteme $\{P_i\}$ existieren, für die jede D–V Zelle den maximalen Inhalt (Umfang) hat.

1.2. In diesem Artikel beschäftigen wir uns mit den folgenden zwei Problemen.

1.2.1. Wir nehmen an, dass $P_jP_k \geq 2r$ für beliebige Punkte P_j, $P_k \in \in \{P_i\}$, $j \neq k$ und $P_kQ_{kq} \leq R$ $(q = 1, 2, \dots, s)$ für jeden Punkt $P_k \in \{P_i\}$ gelten, wobei $0 < r < R$.

1.2.2. Es gelten $P_jP_k \leq 2r$ für beliebige benachbarte Punkte P_j, $P_k \in \in \{P_i\}$ und $P_kQ_{kq} \geq R$ $(q = 1, 2, \dots, s)$ für jeden Punkt $P_k \in \{P_i\}$, wobei $0 < r < R$.

In beiden Fällen ist die Aufgabe, das Maximum bzw. Minimum der Inhalte (Umfänge) der D–V Zellen zu bestimmen und die Punktsysteme $\{P_i\}$ zu karakterisieren, für welche jede D–V Zelle den maximalen bzw. minimalen Inhalt (Umfang) hat.

2. Bezeichnungen, Hilfssätze

Es bezeichne $k(K, r)$ den Kreis vom Mittelpunkt K und Radius r, weiterhin $\hat{k}(K, r)$ seinen Rand. Es ist $\operatorname{int} k(K, r) = k(K, r) \setminus \hat{k}(K, r)$.

HILFSSATZ 1. *Es sei der Kreis $k(K, r)$ gegeben. Wir nehmen an, dass $\angle(AKC) > \angle(BKC)$ mit $A, B, C \in \hat{k}(K, r)$ gilt. Sind A und B fixe Punkte und nimmt $\angle(BKC)$ zu, dann nimmt der Inhalt und der Umfang des Dreiecks ABC streng monoton zu.*

Der Beweis des Hilfssatzes findet man in [4] für $\mathbf{E}^2$, in [2] für $\mathbf{H}^2$ und $\mathbf{S}^2$.

HILFSSATZ 2. *Das einem Kreis $k(K, r)$ einbeschriebene reguläre n-Eck besitzt unter allen in $k(K, r)$ enthaltenen konvexen n-Ecken den grösstmöglichen Inhalt und Umfang.*

Beweis in [4] für $\mathbf{E}^2$, in [2] für $\mathbf{H}^2$ und $\mathbf{S}^2$.

HILFSSATZ 3. *Es sei der Kreis $k(K, r)$ und die Punkte $A, B, C \in \hat{k}(K, r)$ gegeben. Es seien a, b, c die Tangenten an $k(K, r)$ mit den Berührunspunkten A, B, C. Es sei $A_1 = a \cap c$ und $B_1 = b \cap c$. Gilt $\angle(AKC) > \angle(BKC)$ und nimmt $\angle(BKC)$ zu, dann nimmt der Inhalt und der Umfang des Fünfecks KAA_1B_1B streng monoton ab.*

Beweis in [4] für $\mathbf{E}^2$, in [2] für $\mathbf{H}^2$ und $\mathbf{S}^2$.

HILFSSATZ 4. *Das einem Kreis $k(K, r)$ umbeschriebene reguläre n-Eck besitzt unter allen den Kreis $k(K, r)$ enthaltenden n-Ecken den kleinstmöglichen Inhalt und Umfang.*

Beweis in [4] für $\mathbf{E}^2$, in [2] für $\mathbf{H}^2$ und $\mathbf{S}^2$.

Es seien die konzentrischen Kreise $k(K, r)$ und $k(K, R)$ mit $r < R$ gegeben. Mit $H(r, R)$ bezeichnen wir das konvexe Vieleck, das dem Kreis $\hat{k}(K, R)$ einbeschrieben ist und dessen Seiten mit Ausnahme höchstens einer den Kreis $\hat{k}(K, r)$ berühren. Das Vieleck $H(r, R)$ wurde HAJÓS-Vieleck genannt.

HILFSSATZ 5 (HAJÓS LEMMA). *Es seien die konzentrischen Kreise $k(K, r)$ und $k(K, R)$ mit $r < R$ angegeben. Unter den konvexen Vielecken, die den Kreis $k(K, r)$ enthalten und deren Ecken keine innere Punkte von $k(K, R)$ sind, besitzt das Vieleck $H(r, R)$ den kleinstmöglichen Inhalt und Umfang.*

Beweis in [5].

Es seien die konzentrischen Kreise $k(K, r)$ und $k(K, R)$ mit $r < R$ gegeben. Mit $M(r, R)$ bezeichnen wir das in $k(K, R)$ enthaltene konvexe Vieleck, dessen Ecken mit Ausnahme höchstens einer auf der Kreislinie $\hat{k}(K, R)$ liegen und dessen Seiten den Kreis $\hat{k}(K, r)$ berühren.

HILFSSATZ 6. *Es seien die konzentrischen Kreise $k(K, r)$ und $k(K, R)$ mit $r < R$ angegeben. Unter den konvexen Vielecken, deren Seiten die Kreislinie $\hat{k}(K, r)$ schneiden oder berühren und deren Ecken im Kreis $k(K, R)$ liegen, besitzt das Vieleck $M(r, R)$ den grösstmöglichen Inhalt und Umfang.*

Beweis in [3].

Mit $R_0(r)$ bezeichnen wir den Umkreisradius des regulären Dreiecks von der Seitenlänge $2r$.

HILFSSATZ 7. *Nehmen wir an, dass die Ungleichung*

$$\min(AB, AC, BC) \geq 2r$$

für die Seiten des Dreiecks ABC gilt. Der Umkreisradius von ABC ist im Fall $AB = AC = BC = 2r$ minimal.

BEWEIS. Der Beweis ist sehr einfach.

Es sei $k(K, R)$ der Umkreis des Dreiecks ABC, wobei $AB = 2r$ und $\min(AC, BC) \geq 2r$ mit $r < R$ gelten. Es bezeichne φ_0 den Winkel $\angle(ACB)$, wenn $AC = BC$ in $\mathbf{H}^2$ und $AB = AC$ in $\mathbf{E}^2$ und $\mathbf{S}^2$ gilt. ∎

HILFSSATZ 8. *Es sei R der Umkreisradius des Dreiecks ABC, für das $\min(AB, AC, BC) \geq 2r$ gilt. Dann sind die Winkel von ABC nicht kleiner als φ_0. Der Winkel φ_0 tritt ein, wenn $AB = 2r$ und $AC = BC$ in $\mathbf{H}^2$, $AB = AC$ in $\mathbf{E}^2$ und $\mathbf{S}^2$ gilt.*

Beweis in [3].

Mit $R_1(r)$ bezeichnen wir den Umkreisradius des regulären Dreiecks $\Delta(r)$, dessen Inkreisradius r ist.

HILFSSATZ 9. *Es sei der Kreis $k(K, r)$ gegeben. Es gibt ein dem Kreis $k(K, r)$ umbeschriebenes reguläres Dreieck $\Delta(r)$ für alle $r \in R^+$ in der euklidischen Ebene. Der Umkreisradius dieses regulären Dreiecks ist $R_1(r) = 2r$.*

In der sphärischen Ebene existiert $\Delta(r)$ für $0 < r < \frac{\pi}{2}$ und sein Umkreisradius ist

$$R_1(r) = \arcsin \frac{2 \sin r}{\sqrt{1 + 3 \sin^2 r}}.$$

In der hyperbolischen Ebene existiert $\Delta(r)$ nur für $0 < r < \operatorname{arsh} \frac{1}{\sqrt{3}}$ und der Umkreisradius von $\Delta(r)$ ist

$$R_1(r) = \operatorname{arsh} \frac{2 \operatorname{sh} r}{\sqrt{1 - 3 \operatorname{sh}^2 r}}. \tag{1}$$

Im Fall $r = \operatorname{arsh} \frac{1}{\sqrt{3}}$ ist $\Delta(r)$ ein asymptotisches Dreieck.

Der Beweis beruht auf einfachen Rechnungen.

DEFINITION. Es sei $0 < r < R \leq R_1(r)$. Es seien die Kreise $k(K, r)$ und $k(K, R)$ gegeben. Das Dreieck ABC (falls es existiert) ist vom Typ $\Delta_1(r, R)$, wenn sein Inkreis $k(K, r)$ ist, $A, B \in \hat{k}(K, R)$ und die Ecke C kein innerer Punkt des Kreises $k(K, R)$ ist.

HILFSSATZ 10. *In der sphärischen Ebene existiert ein Dreieck vom Typ* $\Delta_1(r,R)$ *für alle* $0 < r < \frac{\pi}{2}$ *und* $0 < r < R \leq R_1(r)$. *In der euklidischen Ebene existiert ein Dreieck vom Typ* $\Delta_1(r,R)$, *wenn* $\sqrt{2}r < R \leq 2r$ *ist. Es gibt ein Dreieck vom Typ* $\Delta_1(r,R)$ *in der hyperbolischen Ebene, wenn* $0 < \mathrm{sh}r < \frac{1}{\sqrt{3}}$ *und* $R_3(r) < R \leq R_1(r)$ *sind, wobei*

$$R_3(r) = \mathrm{arsh}\sqrt{2}\,\mathrm{sh}r\sqrt{\frac{1-\mathrm{sh}^2r+\mathrm{sh}r\,\cosh r}{1-3\mathrm{sh}^2r}}\,. \tag{2}$$

Im Fall $R = R_3(r)$ *sind die Seiten* AC *und* BC *des Dreiecks* ABC *parallel.*

Wir legen den Beweis nicht ausführlich dar. Die Bedingungen für die Existenz des Dreiecks vom Typ $\Delta_1(r,R)$ ergeben sich daraus, dass sich die Tangenten an $k(K,r)$ durch A und B schneiden und der Schnittpunkt kein innerer Punkt des Kreises $k(K,R)$ ist.

HILFSSATZ 11. *Es sei* $R > R_1(r)$. *Es seien die Kreise* $k(K,r)$ *und* $k(K,R)$ *gegeben. Es gibt kein Dreieck* ABC, *das den Mittelpunkt* K *enthält, dessen Seiten sich den Kreis* $\hat{k}(K,r)$ *berühren oder schneiden und dessen Ecken keine innere Punkte von* $k(K,R)$ *sind.*

Die Behauptung des Hilfssatzes folgt daraus, dass $R_1(r)$ der Umkreisradius des regulären Dreiecks mit Inkreis $k(K,r)$ ist.

HILFSSATZ 12. *Es seien* $r \in R^+$ *und* $\sqrt{2}r < R \leq 2r$ *in* $\mathbf{E}^2$, $0 < r < \frac{\pi}{2}$ *und* $0 < r < R \leq R_1(r)$ *in* $\mathbf{S}^2$, $0 < \mathrm{sh}r < \frac{1}{\sqrt{3}}$ *und* $R_3(r) < R \leq R_1(r)$ *in* $\mathbf{H}^2$. *Es seien die Kreise* $k(K,r)$ *und* $k(K,R)$ *gegeben. Wir nehmen an, dass* K *ein innerer Punkt des Dreiecks* ABC *ist, die Seiten von* ABC *sich den Kreis* $\hat{k}(K,r)$ *berühren oder schneiden und die Ecken von* ABC *keine innere Punkte von* $k(K,R)$ *sind. Dann ist der Inhalt und der Umfang von* ABC *nicht grösser als der Inhalt und der Umfang von* $\Delta_1(r,R)$. *Gleichheit tritt dann und nur dann ein, wenn* $ABC \cong \Delta_1(r,R)$ *ist.*

BEWEIS.

1. Wir nehmen an, dass $k(K,r)$ der Inkreis des Dreiecks ABC ist und $AB \leq \min(AC,BC)$ gilt. Wir halten die Seitengeraden AC, AB fest und bewegen die Ecke B auf der fixen Seitengeraden AB gegen A bis der Lage $B \in \hat{k}(K,R)$. Nach Hilfssatz 3 nimmt der Inhalt und der Umfang des Dreiecks ABC streng monoton zu. Eine ähnliche Lageänderung

wird im Fall von A wiederholt, d.h., wir halten die Seitengeraden AB, BC fest und bewegen A gegen B bis Lage $A \in \hat{k}(K, R)$. Der Inhalt und der Umfang von ABC nimmt streng monoton zu. Endlich haben wir ein Dreieck vom Typ $\Delta_1(r, R)$. Die Existenz von $\Delta_1(r, R)$ ist nach Hilfssatz 10 garantiert.

2. Nehmen wir an, dass die Ecken des Dreiecks ABC keine äussere Punkte von $k(K, R_1(r))$ sind. Nach Hilfssatz 2 ist der Inhalt und der Umfang von ABC maximal, wenn ABC dem Kreis $\hat{k}(K, R_1(r))$ einbeschrieben ist. In diesem Fall berühren die Seiten des Dreiecks den Kreis $\hat{k}(K, r)$ (vgl. Fall 1).
3. Nach den Bedingungen unseres Hilfssatzes kann der Fall (vgl. Hilfssatz 11) nicht vorkommen, in dem A, B, $C \notin \operatorname{int} k(K, R_1(r))$ und mindestens eine der Ecken ein äusserer Punkt des Kreises $k(K, R_1(r))$ ist.
4. Es gelte $A \in \operatorname{int} k(K, R_1(r))$ und $B, C \notin \operatorname{int} k(K, R_1(r))$. Wir drehen die Gerade BA um B und die Gerade CA um C bis der Lage, in der die entspechende Seitengeraden den Kreis $\hat{k}(K, r)$ berühren. Der Schnittpunkt, der wieder mit A bezeichnet wird, kann kein äusserer Punkt von $k(K, R_1(r))$ (vgl. Fall 3) sein. Der Schnittpunkt existiert also und $A \notin k(K, R)$ gilt. Das neue Dreieck ABC besitzt einen grösseren Inhalt und Umfang als das ursprüngliche. Es ist leicht einzusehen, dass $BC > > \max(AC, AB)$ gilt. Es sei $AB > AC$. Dann drehen wir die Gerade BC um B. Während der Anwendung der Drehung erreichen wir entweder den Fall 1, d.h., BC berührt den Kreis $\hat{k}(K, r)$, oder den Fall $AC = AB$. Der Inhalt und der Umfang des Dreiecks ABC hat wieder zugenommen. Im letzteren Fall verschieben wir das Dreieck ABC entlang AK bis der Lage, in der BC den Kreis $\hat{k}(K, r)$ berührt oder $A \in \hat{k}(K, R)$ gilt. In beiden Fällen drehen wir die Gerade BA um B und die Gerade CA um C bis der Lage, in der die entspechende Seitengeraden den Kreis $\hat{k}(K, r)$ berühren. Wenn BC die Tangente von $\hat{k}(K, r)$ ist, dann haben wir den Fall 1 erreicht. Schneidet BC den Kreis $\hat{k}(K, r)$, dann wiederholen wir die Verschiebung und die entsprechenden Drehungen. Die Strecke AK nimmt streng monoton zu, deshalb erreichen wir mit der Wiederholung der Verschiebungen und der Drehungen den Fall 1. Der Inhalt und der Umfang von ABC hat inzwischen zugenommen.
5. Es sei $A, B \in \operatorname{int} k(K, R_1(r))$ und $C \notin \operatorname{int} k(K, R_1(r))$. Wir drehen die Geraden CA und CB um C und die Gerade AB bleibt fest. Wir erreichen eine der folgenden Fällen. Mindestens eine der neuen Ecken

A und B liegt auf $\hat{k}(K, R_1(r))$ (Fall 4), oder beide Seitengeraden CA und CB berühren $\hat{k}(K, r)$. Im letzteren Fall, wenn $CA < CB$ ist, dann drehen wir BA um B bis der Lage, in der BA die Tangente von $\hat{k}(K, r)$ ist, oder $AC = BC$ gilt. Es ist klar, dass der Inhalt und der Umfang des Dreiecks ABC während Anwendung der obigen Transformationen zugenommen haben. Endlich, im Fall $AC = BC$ verschieben wir die Gerade AB in Richtung KC bis zur Berührunglage an $\hat{k}(K, r)$. Nach Hilfssatz 11 kann nur $A, B \in \operatorname{int} k(K, R_1(r))$ vorkommen und wir haben Fall 1 erreicht.

Wir bemerken, dass auch ein einfacherer Beweis in $\mathbf{E}^2$ und $\mathbf{S}^2$ existiert. ▮

DEFINITION. Wir betrachten ein normales Mosaik in der Ebene konstanter Krümmung, dessen Flächen Vielecke sind. Das Mosaik wird *A-symmetrisch* genannt, wenn zwei beliebige benachbarte Flächen symmetrisch bezüglich der gemeinsamen Kantengeraden liegen.

Die Winkel des Dreiecks ABC vom Typ $\Delta_1(r, R)$ werden mit α, β und γ bezeichnet, wobei $\alpha = \beta \geq \gamma$.

HILFSSATZ 13. *In der Ebene konstanter Krümmung existieren A-symmetrische Mosaike, deren Flächen Dreiecke vom Typ $\Delta_1(r, R)$ sind. Die folgenden Fällen sind möglich:*

$\Delta_1(r, R)$ *ist regulär in* $\mathbf{E}^2$;

$\Delta_1(r, R)$ *ist regulär oder* $\alpha = \beta = \dfrac{\pi}{2}, \gamma = \dfrac{2\pi}{v}$ *mit* $v = 5, 6, 7, \ldots$ *in* $\mathbf{S}^2$*;*

$\Delta_1(r, R)$ *ist regulär oder* $\alpha = \beta = \dfrac{\pi}{u}, \gamma = \dfrac{2\pi}{v}$ *mit* $u = 3, 4, \ldots$ *und*

$v = 2u{+}1,\ 2u{+}2, \ldots$ *in* $\mathbf{H}^2$.

BEWEIS. Aus der Definition der A-Symmetrie folgt, dass gleiche Winkel herum eine Ecke des Mosaiks liegen. Die regulären Dreiecksmosaike sind vom Typ $\Delta_1(r, R)$. Im folgenden nehmen wir an, dass das Dreieck ABC vom Typ $\Delta_1(r, R)$ und nicht regulär ist. Dann gilt $AC = BC > AB$. Es ist klar, dass die Anzahl der Dreiecke herum A und herum B gerade ist. Es bezeichne $2u$ die Anzahl der Dreiecke herum A und herum B und v die Anzahl der Dreiecke herum C. Dann ist $\alpha = \beta = \frac{\pi}{u}, \gamma = \frac{2\pi}{v}$. Wir nehmen an, dass $u \leq 2$ und $v \leq 3$ sind.

Die Summe der Winkel von ABC ist $\Omega = \frac{2\pi}{u} + \frac{2\pi}{v}$. Wegen $AC = BC >$ $> AB$ gilt $\frac{\pi}{u} > \frac{2\pi}{v}$, woraus ergibt sich $v > 2u$.

In der euklidischen Ebene gilt $\Omega = \pi$, d.h. $(u-2)(v-2) = 4$. Die Gleichung hat nur die Lösung $u = 3$, $v = 6$ unter den Bedingungen $u \geq 2$, $v \geq 3$ und $v > 2u$, d.h. das Dreieck ABC kann nur regulär sein.

In der sphärischen Ebene gilt $\Omega > \pi$, d.h. $(u-2)(v-2) < 4$. Unter den obigen Bedingungen ergeben sich als Lösungen $u = 2$, $v = 5, 6, \dots$, d.h., es gelten $\alpha = \beta = \frac{\pi}{2}, \gamma = \frac{2\pi}{v}$ mit $v = 5, 6, 7, \dots$ für die Winkel von ABC.

In der hyperbolischen Ebene gilt die Ungleichung $(u-2)(v-2) > 4$. Es ist leicht zu zeigen, dass die Gleichung nur die im Hilfssatz angegebenen Lösungen hat. ∎

Bezeichnung. Wir betrachten ein Dreieck ABC vom Typ $\Delta_1(r, R)$, das zu den im Hifssatz 13 angegebenen Parametern u und v gehört. Es sei K der Inkreismittelpunkt von ABC, $r(u,v)$ sein Inkreisradius und $AK = BK = R(u,v)$.

Bemerkung. Ist $AC||BC$ und $\angle(CAB) = \angle(CBA) = \frac{\pi}{u}$, dann existiert ein A-symmetrisches Mosaik für $u \geq 3$, dessen Flächen zu ABC kongruente Dreiecke sind. Es bezeichne K den Inkreismittelpunkt des Dreiecks ABC, $r(u,\infty)$ sein Inkreisradius und $AK = BK = R(u,\infty)$.

Bezeichnung. Es sei $R_2(r)$ der Umkreisradius des gleichschenkligen Dreiecks ABC, wobei $R_2(r) = \frac{1}{2}AB$ ist, die Seiten AC und BC den Kreis $k(r)$ berühren und der Halbierungspunkt von AB der Mittelpunkt von $k(r)$ ist.

Durch einfache Rechnungen ergibt sich die Behauptung von

Hilfssatz 14. *In der euklidischen Ebene gilt* $R_2(r) = \sqrt{2}r$, *in der sphärischen Ebene ist* $R_2(r) = \arcsin\frac{\sqrt{2}\sin r}{\sqrt{1+\sin^2 r}}$. *In der hyperbolischen Ebene existiert* $R_2(r)$ *nur für* $0 < r < \operatorname{arsh} 1$ *und* $R_2(r) = \operatorname{arsh}\frac{\sqrt{2}\operatorname{sh} r}{\sqrt{1-\operatorname{sh}^2 r}}$. *Ist* $r = \operatorname{arsh} 1$, *dann sind zwei Seiten des Dreiecks* ABC *(vgl. Bezeichnung oben) parallel.*

Definition. Es sei $0 < r < R$ und $R_2(r) < R \leq R_1(r)$. Die Kreise $k(K,r)$ und $k(K,R)$ sind gegeben. Das Dreieck ABC ist vom Typ $\Delta_2(r,R)$, wenn sein Umkreis $k(K,R)$ ist, die Seiten AC und BC den Kreis $\hat{k}(K,r)$ berühren und die Seite AB den Kreis $\hat{k}(K,r)$ schneidet oder im Fall $R = R_1(r)$ berührt.

Das Dreieck vom Typ $\Delta_2(r,R)$ existiert unter den Bedingungen im Hilfssatz 14 und ist gleichschenklig.

DEFINITION. Es sei $0 < r < R \leq R_2(r)$. Die Kreise $k(K, r)$ und $k(K, R)$ sind gegeben. Das Dreieck ABC ist vom Typ $\Delta_3(r, R)$, wenn sein Umkreis $k(K, R)$, AB der Durchmesser dieses Kreises ist und mindestens eine der Seiten AC und AB (im Fall $R = R_2(r)$ beide Seiten) den Kreis $\hat{k}(K, r)$ berühren.

HILFSSATZ 15. *Es sei* $0 < r < R \leq R_1(r)$, *wobei* $r < \frac{\pi}{2}$ *in* $\mathbf{S}^2$ *und* $r < \operatorname{arsh}1$ *in* $\mathbf{H}^2$ *sind. Es seien die Kreise* $k(K, r)$ *und* $k(K, R)$ *gegeben. Wir nehmen an, dass* K *kein äusserer Punkt des Dreiecks* ABC *ist, die Seiten von* ABC *den Kreis* $\hat{k}(K, r)$ *schneiden oder berühren und die Ecken von* ABC *keine innere Punkte von* $k(K, R)$ *sind. Der Inhalt und der Umfang von* ABC *ist im Fall* $R_2(r) < R \leq R_1(r)$ *für ein Dreieck vom Typ* $\Delta_2(r, R)$, *im Fall* $r < R \leq R_2(r)$ *für ein Dreieck vom Typ* $\Delta_3(r, R)$ *minimal.*

BEWEIS. Sind die Ecken A, B, C äussere Punkte von $k(K, R)$, dann erhalten wir ein Dreieck vom kleineren Inhalt und Umfang, wenn wir die Schnittpunkte der Halbgeraden KA, KB, KC und der Kreislinie $\hat{k}(K, R)$ nehmen. Das neue Dreieck ABC ist dem Kreis $\hat{k}(K, R)$ einbeschrieben. Es sei $AB \geq \max(AC, BC)$. Wir halten die Ecken B und C fest und bewegen die Ecke A gegen C auf der Kreislinie $\hat{k}(K, R)$. Dann erreichen wir eine der folgenden Lagen. Die Seite CA berührt $\hat{k}(K, r)$, oder AB ist der Durchmesser von $\hat{k}(K, R)$. Im ersten Fall halten wir die Ecken A und C fest und bewegen die Ecke B gegen C auf $\hat{k}(K, R)$ bis der Lage, in der AB eine Tangente von $\hat{k}(K, r)$, oder AB der Durchmesser von $\hat{k}(K, R)$ ist. Nach Hilfssatz 1 nimmt der Inhalt und der Umfang von ABC während Anwendung der obigen Änderungen von ABC streng monoton ab.

Ist $R_2(r) < R \leq R_1(r)$, dann haben wir ein Dreieck vom Typ $\Delta_2(r, R)$ erhalten. Gilt aber $r < R \leq R_2(r)$, dann ist AB der Durchmesser von $\hat{k}(K, R)$. Es sei $BC \leq AC$. Jetzt bewegen wir C gegen B auf $\hat{k}(K, R)$. Dann erreichen wir die Lage, in der BC den Kreis $\hat{k}(K, r)$ berührt, d.h. das Dreieck ABC vom Typ $\Delta_3(r, R)$ ist und kleineren Inhalt und Umfang als das ursprüngliche hat. ∎

Die Winkel des Dreiecks ABC vom Typ $\Delta_2(r, R)$ werden mit α, β und γ bezeichnet, wobei $\alpha = \beta \leq \gamma$.

HILFSSATZ 16. *Es sei* $0 < r < R_2(r) < R \leq R_1(r)$. *(vgl. Hilfssatz 15). In der Ebene konstanter Krümmung existieren* A*-symmetrische Mosaike, deren Flächen Dreiecke vom Typ* $\Delta_2(r, R)$ *sind. Nur die folgenden Fällen sind möglich:*

$\Delta_2(r, R)$, *ist regulär in* $\mathbf{E}^2$;

$\Delta_2(r, R)$, *ist regulär oder* $\alpha = \beta = \dfrac{\pi}{u}, \gamma = \dfrac{2\pi}{v}$ *mit* $u = 2, v = 3$;

$u = 3, v = 4, 5$ *in* $\mathbf{S}^2$;

$\Delta_2(r, R)$, *ist regulär oder* $\alpha = \beta = \dfrac{\pi}{u}, \gamma = \dfrac{2\pi}{v}$ *mit* $u = 4, 5, \ldots$

und $v = u + 1, u + 2, \ldots, 2u - 1$ *in* $\mathbf{H}^2$.

Der Gedankengang des Beweises ist derselbe wie im Hilfssatz 13. Wir verwenden, dass der Umkreismittelpunkt des Dreiecks vom Typ $\Delta_2(r, R)$ ein innerer Punkt des Dreiecks ist. In diesem Fall gelten die Ungleichungen $\alpha \leq \gamma < 2\alpha$.

BEMERKUNG. In der sphärischen Ebene ist die Anzahl der Dreiecke vom Typ $\Delta_2(r, R)$ im entsprechenden A-symmetrischen Mosaik die folgende.

$u = 2, v = 3$:	6 Dreiecke;
$u = 3, v = 4$:	24 Dreiecke, die Ecken des A-symmetrischen Mosaiks sind die Ecken und Flächenmittelpunkte des Mosaiks $\{4, 3\}$;
$u = 3, v = 5$:	60 Dreiecke, die Ecken des A-symmetrischen Mosaik sind die Ecken und Flächenmittelpunkte des Mosaiks $\{5, 3\}$.

BEZEICHNUNG. Wir betrachten ein Dreieck ABC vom Typ $\Delta_2(r, R)$, das zu den im Hilfssatz 16 angegebenen Parametern u und v gehört. Es sei K der Umkreismittelpunkt und $R(u, v)$ der Umkreisradius von ABC. Es bezeichne $r(u, v)$ den Radius des Kreises vom Mittelpunkt K, der die Schenkel von ABC vom Typ $\Delta_2(r, R)$ berührt.

DEFINITION. Wir betrachten ein normales Mosaik in der Ebene konstanter Krümmung, dessen Flächen Dreiecke vom Typ $\Delta_2(r, R)$ sind. Das Mosaik wird C*-symmetrisch* genannt, wenn zwei beliebige benachbarte Dreiecke symmetrisch bezüglich des Mittelpunktes der gemeinsamen Seite liegen.

Wir nehmen an, dass die benachbarten Dreiecksflächen vom Typ $\Delta_2(r, R)$ des Mosaiks entweder symmetrisch bezüglich der gemeinsamen Kantengeraden, oder symmetrisch bezüglich des Mittelpunktes der gemeinsamen Kante liegen und beide Symmetrien vorkommen. Dann wird das Mosaik G*-symmetrisch* genannt.

Haben zwei benachbarte Dreiecke des Mosaiks beide Symmetrien, dann nehmen wir immer nur ihre A-Symmetrie in Betracht.

HILFSSATZ 17. *Es sei r eine positive reelle Zahl, zu der ein Dreieck ABC vom Typ $\Delta_2(r,R)$ gehört. Es sei weiterhin $R_2(r) < R \leq R_1(r)$ (vgl. Hilfssatz 14). In der Ebene konstanter Krümmung existieren G-symmetrische Mosaike mit Dreiecksflächen vom Typ $\Delta_2(r,R)$ nur in den folgenden Fällen.*

In der euklidischen Ebene gilt $\alpha = \beta < \gamma < \frac{\pi}{2}$ für die Winkel von $\Delta_2(r,R)$.

In der sphärischen Ebene existieren G-symmetrische Mosaike nicht.

In der hyperbolischen Ebene existieren zwei Type der Mosaike. Im ersten Typ liegen $2p$ Winkel α und q Winkel γ herum einer beliebigen Ecke des Mosaiks und es gilt $2p\alpha + q\gamma = 2\pi$, wobei $\frac{\pi}{p+q} < \alpha < \frac{2\pi}{2p+q}$. Für die möglichen Werte von p und q gelten die folgenden: $q = 1, p > 3$; $q = 2, p > > 2$; $q = 3,4, p \geq 2$; $q \geq 5, p \geq 1$.

Im zweiten Typ liegen entweder $2p_1$ Winkel α und q_1 Winkel γ herum einer Ecke des Mosaiks, oder $2p_2$ Winkel α und q_2 Winkel γ herum einer Ecke. Dann gelten

$$\alpha = \beta = \frac{q\pi}{p_1 q + q_1 p} \qquad und \qquad \gamma = \frac{2p\pi}{p_1 q + q_1 p},$$

wobei $p = p_1 - p_2 > 0,\ q = q_2 - q_1 > 0,\ 1 < p < q < 2p,\ q_1 \geq 1,\ p_1 \geq 1+p.$

BEWEIS.

1. In der euklidischen Ebene existieren G-symmetrische Mosaike, dessen Flächen nicht reguläre kongruente Dreiecke vom Typ $\Delta_2(r,R)$ sind. Betrachten wir nämlich ein beliebiges gleichschenkliges Dreieck ABC, wobei $\alpha = \beta < \gamma < \frac{\pi}{2}$ ist. Dann spiegeln wir das Dreieck an den Mittelpunkt einer der Schenkel. So entsteht ein Parallelogramm. Mit Verschiebungen um die linearen Kombinationen der nicht parrallelen Seitenvektoren des Parallelogramms können wir ein G-symmetrisches Mosaik bilden. Es ist klar, dass unendlich viele G-symmetrische Mosaike mit denselben Flächen existieren.
2. In der sphärischen Ebene ist die Anzahl der Winkel herum einer beliebigen Ecke wegen $\alpha = \beta$ gerade. Aus den Ungleichungen $2\alpha + \gamma > \pi$ und $\alpha < \gamma < 2\alpha$ folgt, dass die Anzahl der Winkel α 2 oder 4 ist.
   a) Liegen 4 Winkel α herum einer Ecke, dann gilt $4\alpha + \gamma = 2\pi$. Für jede Ecke kann diese Bedingung nicht gelten, sonst existiere eine

Ecke, herum der 2 Winkel γ sind. Dann gibt es mindestens eine Ecke, herum der 2 Winkel α liegen. Dann gilt $2\alpha + k\gamma = 2\pi$. Es ist leicht einzusehen, dass nur der Fall $k = 2$ vorkommen kann. Daraus folgt $\gamma = 2\alpha$, was nicht möglich ist.

b) Es gibt keine Ecke, herum der 4 Winkel α liegen. Es gilt $2\alpha + + k\gamma = 2\pi$. Aus den Ungleichungen $4\alpha + 2\gamma > 2\pi$ und $\alpha < \gamma$ folgt, dass nur $k = 2, 3$ möglich ist. In beiden Fällen, wenn wir das entsprechende Mosaik konstruiren möchten, ergibt sich Ecke, herum der 4 Winkel α liegen.

In der sphärischen Ebene existieren also G-symmetrische Mosaike nicht.

3. In der hyperbolischen Ebene betrachten wir ein Dreieck vom Typ $\Delta_2(r, R)$, für die $\alpha < \gamma < 2\alpha$ gilt. Dann spiegeln wir das Dreieck an seiner Basis, so entsteht ein Rhombus mit Winkeln 2α und γ. Die Existenz eines Mosaiks mit Dreieck $\Delta_2(r, R)$ ist mit der Existenz eines Mosaiks äquivalent, dessen Flächen zu diesem Rhombus kongruent sind.

a) Im ersten Fall gilt für das Dreiecksmosaik, dass dieselbe Anzahl vom Winkel α herum die Ecken des Mosaiks zusammentreffen. Es bezeichne $2p$ die Anzahl der Winkel α und q die Anzahl der Winkel γ. Dann gilt $2p\alpha + q\gamma = 2\pi$. Daraus ergibt sich $\gamma = \frac{2\pi - 2p\alpha}{q}$. Aus $\alpha < \gamma < 2\alpha$ folgt $\frac{\pi}{p+q} < \alpha < \frac{2\pi}{2p+q}$. Es gilt noch die Ungleichung $2\alpha + \gamma < \pi$, woraus erhält man $2\alpha(q - p) < \pi(q - 2)$. Durch einfache Rechnungen ergeben sich die möglichen Werte von p und q.

b) Im Mosaik vom zweiten Typ existieren Ecken, für die die Anzahl der Winkel α herum den Ecken nicht gleich ist. Es gelten $2p_1\alpha + + q_1\gamma = 2\pi$, $2p_2\alpha + q_2\gamma = 2\pi$, wobei $q_1 \neq q_2$ wegen $p_1 \neq p_2$ gilt. Es ist klar, dass ein anderes Typ von Ecken nicht existiert.

Es seien $p_1 > p_2$, $p = p_1 - p_2$, $q = q_2 - q_1$. Es gilt $p > 0$, $q > 0$ offenbar. Aus den obigen Gleichungssystem folgt $2p\alpha = q\gamma$. Aus $\alpha < \gamma < 2\alpha$ erhalten wir die Ungleichung $p < q < 2p$.

Durch einfache Rechnungen ergeben sich

$$\alpha = \beta = \frac{q\pi}{p_1 q + q_1 p} \qquad \text{und} \qquad \gamma = \frac{2p\pi}{p_1 q + q_1 p}.$$

Wegen der Ungleichung $2\alpha + \gamma < \pi$ muss auch

$$\frac{p + q}{p_1 q + q_1 p} < \frac{1}{2}$$

gelten. Aus $q_1 \geq 1$ und $p_1 = p + p_2 \geq 1 + p$ folgt, dass die obige Ungleichung für $p > 1$ gilt.

Wir bemerken, dass mehrere Herumlegung der Dreiecke für eine Ecke existiert, d.h. mehrere Mosaike extremal sind. ∎

3. Sätze

SATZ 1. *Es sei* $0 < r < R$. *Wir nehmen an, dass das Punktsystem* $\{P_i\}$ *für einen beliebigen Punkt* $P_k \in \{P_i\}$ *die folgenden Eigenschaften hat. Die D–V Zelle* D_k *enthält den Kreis* $\hat{k}(P_k, r)$ *und die Ecken von* D_k *liegen in* $k(P_k, R)$. *Der Inhalt und Umfang einer beliebigen D–V Zelle von* $\{P_i\}$ *ist für das Vieleck* $H(r, R_0(r))$ *minimal.*

Das Mosaik, deren Flächen zu $H(r, R_0(r))$ *kongruent sind, existiert nur im Fall, wenn das Vieleck* $H(r, R_0(r))$ *mit Winkel* $\frac{2\pi}{3}$ *regulär ist. In diesem extremalen Fall sind die Punkte von* $\{P_i\}$ *die Ecken eines regulären Dreiecksmosaiks von der Kantenlänge* $2r$.

BEWEIS. Es gilt $P_jP_k \geq 2r$ für beliebige $P_j, P_k \in \{P_i\}$, weil D_k den Kreis $\hat{k}(P_k, R)$ enthält. Aus Hilfssatz 7 folgt, dass die Ecken von D_k nur im Fall zum Kreis $k(P_k, R)$ gehören, wenn $R > R_0(r)$ ist, d.h., die Ecken von D_k ausser int $k(P_k, R_0(r))$ liegen. Nach Hilfssatz 5 ist der Inhalt und Umfang einer beliebigen D–V Zelle von $\{P_i\}$ für das Vieleck $H(r, R_0(r))$ minimal.

Das Minimum tritt für alle D–V Zellen ein, wenn sich ein Mosaik mit zu $H(r, R_0(r))$ kongruenten Flächen bilden lässt. Die Winkel des Vielecks $H(r, R_0(r))$ sind gleich $\frac{2\pi}{3}$ mit Ausnahme von höchstens zwei Winkeln. Diese zwei zurückgebliebende Winkel sind grösser als $\frac{2\pi}{3}$. Das Vieleck $H(r, R_0(r))$ also regulär sein.

Daraus folgt, dass die Punkte des Punktsystems $\{P_i\}$ sind die Ecken eines regulären Dreiecksmosaiks von der Kantenlänge $2r$. ∎

SATZ 2. *Es sei* $0 < r < R$. *Wir nehmen an, dass das Punktsystem* $\{P_i\}$ *für einen beliebigen Punkt* $P_k \in \{P_i\}$ *die folgenden Eigenschaften hat. Die D–V Zelle* D_k *enthält den Kreis* $\hat{k}(P_k, r)$ *und die Ecken von* D_k *liegen in* $k(P_k, R)$. *Der Inhalt und Umfang einer beliebigen D–V Zelle von* $\{P_i\}$ *ist für das reguläre* n_0*-Eck maximal, der einem Kreis vom Radius R einbeschrieben ist, wobei* $n_0 = \left[\frac{2\pi}{\varphi_0}\right]$. *(Die Definition von* φ_0 *findet man vor Hilfssatz 8.)*

Jede D–V Zelle ist nur im Fall vom maximalen Inhalt und Umfang, wenn $R = R_0(r)$ ist und ein Mosaik mit zu $H(r, R_0(r))$ kongruenten Flächen existiert. In diesem extremalen Fall sind die Punkte von $\{P_i\}$ die Ecken eines regulären Dreiecksmosaiks von der Kantenlänge $2r$.

BEWEIS. Aus Hilfssatz 7 folgt, dass ein Punktsystem unter den obigen Bedingungen nur im Fall existiert, wenn $R \geq R_0(r)$. Die Ecken von D_k liegen in $k(P_k, R)$, deshalb ist der Inhalt und Umfang von D_k für eine fixe Eckenzahl nicht grösser als der Umfang und Inhalt des dem Kreis $k(P_k, R)$ einbeschriebenen regulären n-Ecks.

Wir zeigen, dass $n \leq n_0 = \left[\frac{2\pi}{\varphi_0}\right]$. Aus der Definition der D–V Zelle folgt, dass eine ihrer Ecken der Mittelpunkt des Umkreises irgendeines Dreiecks $P_i P_j P_k$ ist. Ist der Radius dieses Umkreises nicht grösser als R, dann gilt $\angle(P_j P_k P_i) \geq \varphi_0$ nach Hilfssatz 8. Daraus ergibt sich $n \leq n_0$. Der Winkel φ_0 ist im Fall der grösste, wenn $R = R_0(r)$ ist, d.h., $P_i P_j P_k$ ein reguläres Dreieck von der Seitenlänge $2r$ ist. Daraus folgt schon die Behauptung des Satzes. ∎

SATZ 3. *Es sei $0 < r < R$, weiterhin im euklidischen Fall $\sqrt{2}r < R \leq 2r$, in der sphärischen Ebene $r < \frac{\pi}{2}$ und $r < R \leq R_1(r)$, im hyperbolischen Fall $r < \operatorname{arsh}\frac{1}{\sqrt{3}}$ und $R_3(r) < R \leq R_1(r)$. (Die Radien $R_1(r)$ und $R_3(r)$ findet man in (1) und (2).) Schneiden oder berühren die Seiten der D–V Zelle D_k für einen beliebigen Punkt $P_k \in \{P_i\}$ den Kreis $\hat{k}(P_k, r)$ und sind ihre Ecken keine innere Punkte von $k(P_k, R)$, dann ist der Inhalt und Umfang einer beliebigen D–V Zelle nicht grösser als der Inhalt und Umfang des Dreiecks vom Typ $\Delta_1(r, R)$.*

Jede D–V Zelle ist dann und nur dann vom maximalen Inhalt und Umfang, wenn ein Mosaik (vgl. Hilfssatz 13) existiert, deren Flächen zu $\Delta_1(r, R)$ kongruent sind, d.h. für die Fälle $r = r(u, v)$ und $R = R(u, v)$ (vgl. Bezeichnung nach Hilfssatz 13).

BEWEIS. Wir betrachten ein Punktsystem $\{P_i\}$, das den Bedingungen des Satzes entsprechend ist. Es sei $P_k \in \{P_i\}$ ein beliebiger Punkt. In welchem Fall ist der Inhalt und Umfang der entsprechenden D–V Zelle maximal? Aus $R_3(r) < R$ folgt, dass die Zelle D_k ein Dreieck ist. Nach Hilfssatz 12 ist der Inhalt und Umfang von D_k maximal, wenn das Dreieck vom Typ $\Delta_1(r, R)$ ist.

Das Mosaik, deren Flächen Dreiecke vom Typ $\Delta_1(r, R)$ sind, existiert nur für $r = r(u, v)$ und $R = R(u, v)$ nach Hilfssatz 13. Im extremalen Fall sind die Punkte von $\{P_i\}$ die Inkreismittelpunkte der Flächen vom Typ $\Delta_1(r, R)$. ∎

SATZ 4. *Es sei* $0 < r < R \leq R_1(r)$, *weiterhin in der sphärischen Ebene* $r < \frac{\pi}{2}$. *Schneiden oder berühren die Seiten der D–V Zelle* D_k *für einen beliebigen Punkt* $P_k \in \{P_i\}$ *den Kreis* $\hat{k}(P_k, r)$ *und sind ihre Ecken keine innere Punkte von* $k(P_k, R)$, *dann ist der Inhalt und Umfang einer beliebigen D–V Zelle nicht grösser als der Inhalt und Umfang des Dreiecks vom Typ* $\Delta_2(r, R)$ *für* $R_2(r) < R \leq R_1(r)$ *und ist grösser als der Inhalt und Umfang des Dreiecks vom Typ* $\Delta_3(r, R)$ *für* $r < R \leq R_2(r)$.

für $R_2(r) < R \leq R_1(r)$ *tritt der minimale Inhalt und Umfang für jede D–V Zelle ein, wenn A-symmetrische oder G-symmetrische Mosaike mit Flächen vom Typ* $\Delta_2(r, R)$ *existieren, d.h. für die im Hilfssatz 16 und 17 angegebene Werte von* r *und* R.

BEWEIS. Wir betrachten ein Punktsystem $\{P_i\}$, das den Bedingungen des Satzes entspricht. Es sei $P_k \in \{P_i\}$ ein beliebiger Punkt des Punktsystems und D_k die zum Punkt P_k gehörige D–V Zelle. Wir wählen das Dreieck ABC derart, dass seine Ecken zu D_k gehören und P_k der Punkt des Dreiecksbereichs ist. Es ist offenbar, dass der Inhalt und Umfang von D_k nicht kleiner als der Inhalt und Umfang von ABC ist und Gleichheit tritt nur im Fall $D_k = ABC$ ein. Nach Hilfssatz 15 ist der Inhalt und Umfang von ABC für $R_2(r) < R \leq R_1(r)$ nicht kleiner als der Inhalt und Umfang des Dreiecks vom Typ $\Delta_2(r, R)$ und Gleichheit tritt nur im Fall $ABC = \Delta_2(r, R)$ ein.

Jede D–V Zelle ist vom minimalen Inhalt und Umfang, wenn ein Mosaik mit Flächen vom Typ $\Delta_2(r, R)$ existiert. Ist dieses Mosaik A-symmetrisch bzw. G-symmetrisch, dann können nur die im Hilfssatz 16 bzw. 17 gegebene Kreisradien r und R eintreten.

Die Punkte des Punktsystems $\{P_i\}$ sind die Umkreismittelpunkte der Flächen vom Typ $\Delta_2(r, R)$. Wenn $r < R \leq R_2(r)$ ist, dann ist der Inhalt und Umfang einer beliebigen D–V Zelle grösser als der Inhalt und Umfang des Dreiecks vom Typ $\Delta_3(r, R)$. Gleichheit kann nie erreicht werden, weil ein Dreieck vom Typ $\Delta_3(r, R)$ keine D–V Zelle ist. Sein Umkreismittelpunkt ist nämlich kein innerer Punkt des Dreiecks. ∎

Literaturverzeichniss

[1] FEJES TÓTH, L.: *Lagerungen in der Ebene, auf der Kugel und im Raum,* Berlin–Göttingen–Heidelberg, (zweite Auflage, 1972).

[2] HORVÁTH, J.: Some extremal properities of regular polygons in Euclidean and hyperbolic plane, furthermore on the sphere, *Mat. Tanítása* **XXIX/4**, 106–112 (Hungarian).

[3] HORVÁTH, J. und STOGRIN, M. I.: Two extremal problems for Dirichlet–Voronoi domains, *Mat. Zametki* **41(2)**, 257–264 (Russian); Engl. Trans. *Math. Notes* **41**, 147–151.

[4] KÜRSCHÁK, J.: Über in dem Kreis ein- und umgeschriebene Vielecke, *Math. Ann.* **30**, 578–581.

[5] MOLNÁR, J.: Kreisanordnungen in der Ebene konstanter Krümmung, *III. Oszt. Közl.* **XII(3)**, 223–263 (ungarisch).

Jenő Horváth
Universität Sopron
Institut für Mathematik
9400 Sopron, Ungarn
jhorvath@emk.nyme.hu

Ágota H. Temesvári
Universität Sopron
Institut für Mathematik
9400 Sopron, Ungarn
hta@emk.nyme.hu

ON HYPERSURFACES WITH TYPE NUMBER TWO IN SPACE FORMS

By

RYSZARD DESZCZ and MARIAN HOTLOŚ

(Received March 21, 2003)

Dedicated to Professor Dr. Bang-Yen Chen on his sixtieth birthday

1. Introduction

Let (M, g), $n = \dim M \geq 4$, be a semi-Riemannian manifold satisfying at every point the following condition: the tensors $R \cdot C$ and $Q(S, C)$ are linearly dependent. This is equivalent on the set U consisting of all points of M at which $Q(S, C) \neq 0$ to

$$R \cdot C = L\, Q(S, C)\,, \tag{1}$$

where L is some function on U. For precise definition of the symbols used we refer to Sections **2** and **3** of this paper. We recall that the case: $S \neq 0$, $C \neq 0$ and $Q(S, C) = 0$ at a point $x \in M$ was considered in [10](Theorem 3.1). In this paper, without loss of generality, we restrict our investigations to the set $\mathcal{U}_L \subset U$ defined by $\mathcal{U}_L = \{x \in U \mid L \neq 0 \text{ at } x\}$. Manifolds fulfilling (1) were studied among others in: [6], [7], [10], [12], [13] and [14].

Let M be a hypersurface in a semi-Riemannian space of constant curvature $N_s^{n+1}(c)$, $n \geq 4$, with signature $(s, n + 1 - s)$. We denote by $\mathcal{U}_H$ the subset of M consisting of all points at which the tensor H^2 is not a linear combination of the tensor H and the metric tensor g induced on M from the metric of the ambient space.

The results of [6], [7], [13] and [14] are related to hypersurfaces in semi-Euclidean spaces $\mathbb{E}_s^{n+1}$, $n \geq 4$, fulfilling (1). For instance, in [13] (see Theorem 4.1 and Theorem 4.2) it was shown that if M is a hypersurface in

The authors are supported by the grant 2 P03A 006 17 from the Polish State Committee of Scientific Research (KBN).

$\mathbb{E}_s^{n+1}$, $n \geq 4$, satisfying (1) then on $\mathcal{U}_H \cap \mathcal{U}_L$ we have: $rank\big(S - \frac{\kappa}{n-1}\, g\big) = 1$ and

$$R \cdot C = Q(S, C)\,, \tag{2}$$

i.e. (1) with $L = 1$. An example of a hypersurface M in $\mathbb{E}_s^{n+1}$, $n \geq 4$, satisfying (2) was found in [6]. Some partial results on hypersurfaces in semi-Riemannian spaces of constant curvature satisfying (1) are contained in [13] (see also Section **4** for details). In this paper we investigate hypersurfaces M in a semi-Riemannian space of constant curvature $N_s^{n+1}(c)$, $c \neq 0$, $n \geq 4$, fulfilling (1). In Section **4** (see Proposition 4.3) we prove that if (2) is satisfied on $\mathcal{U}_H \neq \emptyset$ of a hypersurface M in $N_s^{n+1}(c)$, $n \geq 4$, then the ambient space must be semi-Euclidean. Therefore we investigate hypersurfaces M in $N_s^{n+1}(c)$ with nonzero sectional curvature c satisfying (1). Theorem 4.1 states that if M is a such hypersurface then on $\mathcal{U}_H \cap \mathcal{U}_L$ we have:

$$S - \frac{\kappa}{n}\, g = \beta\, w \otimes w\,, \quad \beta \in \mathbb{R}, \quad w \in T_x^* M\,, \tag{3}$$

$$R \cdot C = \frac{1}{n-1}\, Q(S, C)\,, \tag{4}$$

$$R \cdot R = \frac{\widetilde{\kappa}}{n(n+1)}\, Q(g, R)\,. \tag{5}$$

It is known that (5) is equivalent to the fact that at every point of $\mathcal{U}_H \cap \mathcal{U}_L$ the type number of M is equal to two ([5], Theorem 5.1). At the end of Section **4** we present also a corrected version of some results from [13]. In Section **5** (Example 5.1) we present an example of a warped product manifold satisfying (4), which can be locally realized as a hypersurface in $N_s^{n+1}(c)$, $c \neq 0$, $n \geq 4$. The type number of that hypersurface is equal to two, i.e. it is pseudosymmetric. On the other hand, there exist pseudosymmetric hypersurfaces in $N_s^{n+1}(c)$, $c \neq 0$, $n \geq 4$, with type number two, which do not satisfy (4). Namely, generalized Cartan hypersurfaces ([2], see also Example 5.2) have such properties. The hypersurface constructed in Example 5.1, resp., in Example 5.2, fulfils

$$(a)\ \ rank\,(H^2 - tr(H)\, H) = 1\,, \qquad (b)\ \ rank\,(H^2 - tr(H)\, H) = 2\,, \tag{6}$$

respectively. We prove (see Proposition 5.1(i)) that at every point $x \in \mathcal{U}_H$ of a pseudosymmetric hypersurface M in $N_s^{n+1}(c)$, $n \geq 4$, (6)(a) or (6)(b) must be satisfied. Moreover, Proposition 5.1(iii) shows that (6)(a) implies (4). On the other hand, Proposition 5.1(ii) states that hypersurfaces fulfilling (6)(b) cannot satisfy the condition (1), and consequently, (4). We mention that warped products as Riemannian submanifolds have been investigated by Professor Bang-Yen Chen, see [3].

The authors would like to express their thanks to Professor Leopold Verstraelen for his hints and comments related to this paper.

2. Basic notations

Throughout this paper all manifolds are assumed to be connected paracompact manifolds of class C^{∞}. Let (M,g) be an n-dimensional, $n \geq 3$, semi-Riemannian manifold. We denote by ∇, R, C, S and κ the Levi-Civita connection, the Riemann-Christoffel curvature tensor, the Weyl conformal curvature tensor, the Ricci tensor and the scalar curvature of (M,g), respectively. The Ricci operator $\mathcal{S}$ is defined by $g(\mathcal{S}X, Y) = S(X, Y)$, where $X, Y \in \Xi(M)$, $\Xi(M)$ being the Lie algebra of vector fields on M. We define the endomorphisms $X \wedge_A Y$, $\mathcal{R}(X, Y)$ and $\mathcal{C}(X, Y)$ of $\Xi(M)$ by $(X \wedge_A Y)Z = A(Y,Z)X - A(X,Z)Y$, $\mathcal{R}(X, Y)Z = [\nabla_X, \nabla_Y]Z - - \nabla_{[X,Y]}Z$ and $\mathcal{C}(X, Y) = \mathcal{R}(X, Y) - \frac{1}{n-2}(X \wedge_g \mathcal{S}Y + \mathcal{S}X \wedge_g Y - - \frac{\kappa}{n-1} X \wedge_g Y)$, respectively, where $X, Y, Z \in \Xi(M)$ and A is a symmetric $(0,2)$- tensor. Now the Riemann-Christoffel curvature tensor R, the Weyl conformal curvature tensor C and the $(0,4)$-tensor G of (M,g) are defined by $R(X_1,X_2,X_3,X_4) = g(\mathcal{R}(X_1,X_2)X_3,X_4)$, $C(X_1,X_2,X_3,X_4) = = g(\mathcal{C}(X_1,X_2)X_3,X_4)$ and $G(X_1,X_2,X_3,X_4) = g((X_1 \wedge_g X_2)X_3,X_4)$, respectively, where $X, Y, Z, X_1, X_2, \ldots \in \Xi(M)$. We define the following subsets of M: $\mathcal{U}_R = \{x \in M \mid R - \frac{\kappa}{(n-1)n} G \neq 0 \text{ at } x\}$, $\mathcal{U}_S = \{x \in M \mid S - - \frac{\kappa}{n} g \neq 0 \text{ at } x\}$, $\mathcal{U}_C = \{x \in M \mid C \neq 0 \text{ at } x\}$ and $\mathcal{U} = \mathcal{U}_S \cap \mathcal{U}_C$. We note that $\mathcal{U} \subset \mathcal{U}_R$. Let $\mathcal{B}(X, Y)$ be a skew-symmetric endomorphism of $\Xi(M)$ and let B be a $(0,4)$-tensor associated with $\mathcal{B}(X, Y)$ by

$$B(X_1,X_2,X_3,X_4) = g(\mathcal{B}(X_1,X_2)X_3,X_4)\,. \tag{7}$$

B is said to be a generalized curvature tensor if the following conditions are fulfilled

$$B(X_1,X_2,X_3,X_4) + B(X_2,X_3,X_1,X_4) + B(X_3,X_1,X_2,X_4) = 0\,,$$
$$B(X_1,X_2,X_3,X_4) = B(X_3,X_4,X_1,X_2)\,.$$

Clearly, the tensors R, C and G are generalized curvature tensors. For symmetric $(0,2)$-tensors E and F we define their Kulkarni-Nomizu product $E \wedge F$ by

$$\begin{aligned}(E \wedge F)(X_1,X_2,X_3,X_4) = E(X_1,X_4)F(X_2,X_3) + E(X_2,X_3)F(X_1,X_4)\\ - E(X_1,X_3)F(X_2,X_4) - E(X_2,X_4)F(X_1,X_3)\,.\end{aligned}$$

The tensor $E \wedge F$ is also a generalized curvature tensor. For a symmetric $(0,2)$-tensor E we define the $(0,4)$-tensor $\overline{E}$ by $\overline{E} = \frac{1}{2} E \wedge E$. In particular, we have $\overline{g} = G = \frac{1}{2}\, g \wedge g$. We note that the Weyl tensor C can be presented in the form

$$C = R - \frac{1}{n-2}\, g \wedge S + \frac{\kappa}{(n-2)(n-1)}\, G\,. \tag{8}$$

We have also (see e.g. [7], Section 3)

$$Q(E, E \wedge F) = -\,Q(F, \overline{E})\,. \tag{9}$$

Let $\mathcal{B}(X,Y)$ be a skew-symmetric endomorphism of $\Xi(M)$ and let B be the tensor defined by (7). We extend the endomorphism $\mathcal{B}(X,Y)$ to derivation $\mathcal{B}(X,Y)\cdot$ of the algebra of tensor fields on M, assuming that it commutes with contractions and $\mathcal{B}(X,Y)\cdot f = 0$ for any smooth function on M. Now for a $(0,k)$-tensor field T, $k \geq 1$, we can define the $(0,k+2)$-tensor $B \cdot T$ by

$$\begin{aligned} (B \cdot T)(X_1, \ldots, X_k; X, Y) &= (\mathcal{B}(X,Y) \cdot T)(X_1, \ldots, X_k; X, Y) \\ = -\,T(\mathcal{B}(X,Y)X_1, X_2, \ldots, X_k) &- \cdots - T(X_1, \ldots, X_{k-1}, \mathcal{B}(X,Y)X_k)\,. \end{aligned}$$

In addition, if A is a symmetric $(0,2)$-tensor then we define the $(0,k{+}2)$-tensor $Q(A,T)$ by

$$\begin{aligned} Q(A,T)(X_1, \ldots, X_k; X, Y) &= (X \wedge_A Y \cdot T)(X_1, \ldots, X_k; X, Y) \\ = -\,T((X \wedge_A Y)X_1, X_2, \ldots, X_k) &- \cdots - T(X_1, \ldots, X_{k-1}, (X \wedge_A Y)X_k)\,. \end{aligned}$$

In particular, in this manner, we obtain the $(0,6)$-tensors $B \cdot B$ and $Q(A,B)$. Setting in the above formulas $\mathcal{B} = \mathcal{R}$ or $\mathcal{B} = \mathcal{C}$, $T = R$ or $T = C$ or $T = S$, $A = g$ or $A = S$, we get the tensors $R \cdot R$, $R \cdot C$, $C \cdot R$, $R \cdot S$, $C \cdot S$, $Q(g,R)$, $Q(S,R)$, $Q(g,C)$ and $Q(g,S)$.

A semi-Riemannian manifold (M,g), $n \geq 3$, is said to be pseudosymmetric if at every point of M the tensors $R \cdot R$ and $Q(g,R)$ are linearly dependent. This is equivalent to

$$R \cdot R = L_R\, Q(g,R) \tag{10}$$

on $\mathcal{U}_R$, where L_R is some function on $\mathcal{U}_R$. The class of pseudosymmetric manifolds is an extension of the class of semisymmetric manifolds. A manifold (M,g), $n \geq 3$, is called semisymmetric ([16]) if on M we have $R \cdot R = 0$. Some geometrical considerations show that (10) is a more natural curvature condition than the condition of semisymmetry. For a presentation of facts related to this statement we refer to a recent review paper [1].

3. Some curvature identities

LEMMA 3.1 *(cf. [9], Lemma 3.4) Let (M, g), $n \geq 3$, be a semi-Riemannian manifold. Let at a point $x \in M$ be given a nonzero symmetric $(0,2)$-tensor E and a generalized curvature tensor B such that $Q(E, B) = 0$ at x. Moreover, let Y be a vector at x such that the scalar $\rho = a(Y)$ is nonzero, where a is a covector defined by $a(X) = E(X, Y)$, $X \in T_xM$. Then at x we have two possibilities: (i) the tensor E is of rank one (precisely, $E = \frac{1}{\rho}\, a \otimes a$) and*

$$\underset{X,Y,Z}{\mathcal{S}}\, a(X) B(Y, Z, X_1, X_2) = 0\,,$$

(ii) the tensor $E - \frac{1}{\rho}\, a \otimes a$ is nonzero and $B = \frac{\gamma}{2}\, E \wedge E$, $\gamma \in \mathbb{R}$.

LEMMA 3.2 *Let (M, g) be a semi-Riemannian manifold. If the curvature tensor R of M is of the form*

$$R = \phi\, \overline{S} + \mu\, g \wedge S + \eta\, G\,, \tag{11}$$

where ϕ, μ, η are some functions on M, then the above decomposition of R is unique on the set $\mathcal{U}$.

PROOF. Let $x \in \mathcal{U}$ and suppose that $R = \alpha_i\, \overline{S} + \beta_i\, g \wedge S + \gamma_i\, G$, $i = 1, 2$. If $\alpha_1 \neq \alpha_2$ then $\bar{S} = \alpha\, g \wedge S + \beta\, G$, which in view of Lemma 3.1 of [8], implies $S = \alpha g + \gamma u \otimes u$. This equality together with the above decomposition of R, in virtue of Remark 2.1 of [11], leads to $C = 0$ at x, a contradiction. Thus we have $\alpha_1 = \alpha_2$. Next, if we would have $\beta_1 \neq \beta_2$ then we would obtain $g \wedge S = \delta\, G$. But this immediately implies $S = \tilde{\delta} g$, a contradiction. Finally, if $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$ then obviously we also have $\gamma_1 = \gamma_2$.

From Theorem 4.1 of [12] we can deduce that if (M, g) is a semi-Riemannian manifold such that

$$\begin{aligned} (i)\quad & R \cdot R = \frac{\kappa}{n(n-1)}\, Q(g, R)\,, \\ (ii)\quad & R \cdot R = Q(S, R) - \frac{(n-2)\kappa}{n(n-1)}\, Q(g, C)\,, \\ (iii)\quad & R \cdot C = \frac{1}{n-1}\, Q(S, C)\,, \end{aligned} \tag{12}$$

then at every point $x \in \mathcal{U} \subset M$ we have

$$S = \frac{\kappa}{n}\, g + \beta\, w \otimes w\,, \quad w \in T_x^*M\,, \quad \beta \in \mathbb{R}, \qquad \underset{X,Y,Z}{\mathcal{S}}\, w(X) \mathscr{C}(Y, Z) = 0\,. \tag{13}$$

Thus taking into account also Theorem 4.2 of [12] we have the following equivalence:

PROPOSITION 3.1 *For every semi-Riemannian manifold* (M,g) *the conditions (12) and (13) are equivalent on* $\mathcal{U} \subset M$.

Finally, as a simple consequence of Theorem 4.2 of [9], we obtain:

LEMMA 3.3 *If the curvature tensor* R *of a semi-Riemannian manifold* (M,g) *satisfies (11), (12)(i) and (12)(ii) then*

$$R = \phi\, \overline{S} - \phi\, \frac{\kappa}{n}\, g \wedge S + \left(\frac{\kappa}{n(n-1)} + \phi\, \frac{\kappa^2}{n^2} \right) G . \tag{14}$$

4. Hypersurfaces satisfying $R \cdot C = L\, Q(S,C)$

Let M, $n \geq 3$, be a connected hypersurface isometrically immersed in a semi-Riemannian manifold (N, g^N). We denote by g the metric tensor of M induced from the metric tensor g^N. Further, we denote by ∇ and ∇^N the Levi-Civita connections corresponding to the metric tensors g and g^N, respectively. Let ξ be a local unit normal vector field on M in N and let $\varepsilon = g^N(\xi,\xi) = \pm 1$. We can present the Gauss formula and the Weingarten formula of (M,g) in (N,g^N) in the form: $\nabla^N_X Y = \nabla_X Y + \varepsilon\, H(X,Y)\xi$ and $\nabla_X \xi = -\mathcal{A}X$, respectively, where X, Y are vector fields tangent to M, H is the second fundamental tensor of (M,g) in (N,g^N), $\mathcal{A}$ is the shape operator and $H^k(X,Y) = g(\mathcal{A}^k X, Y)$, $k \geq 1$, $H^1 = H$ and $\mathcal{A}^1 = \mathcal{A}$. We denote by R and R^N the Riemann-Christoffel curvature tensors of (M,g) and (N,g^N), respectively. The *Gauss equation* of (M,g) in (N,g^N) has the form $R(X_1,\ldots,X_4) = R^N(X_1,\ldots,X_4) + \varepsilon\, \overline{H}(X_1,\ldots,X_4)$, where $\overline{H} = \frac{1}{2} H \wedge H$ and $X_1,\ldots,X_4$ are vector fields tangent to M. Let the equations $x^r = x^r(y^k)$ be the local parametric expression of (M,g) in (N,g^N), where y^k and x^r are the local coordinates of M and N, respectively, and $h,i,j,k \in \{1,2,\ldots,n\}$ and $p,r,t,u \in \{1,2,\ldots,n+1\}$. Let $\overline{H}_{hijk} = H_{hk}H_{ij} - H_{hj}H_{ik}$ denote the local components of the tensor $\overline{H}$. Now the Gauss equation turns into

$$R_{hijk} = R^N_{prtu} B_h^{\ p} B_i^{\ r} B_j^{\ t} B_k^{\ u} + \varepsilon\, \overline{H}_{hijk} , \qquad B_k^{\ r} = \frac{\partial x^r}{\partial y^k} , \tag{15}$$

where R^N_{rstu}, R_{hijk} and H_{hk} are the local components of the tensors R^N, R and H, respectively. We assume that the ambient space (N,g^N) is a conformally flat space. Using (8), (15) and the formulas (18) and (19) of [15] we get

$$C_{hijk} = \mu\, G_{hijk} + \varepsilon\, \overline{H}_{hijk} + \frac{\varepsilon}{n-2}\,(g \wedge (H^2 - tr(H)\,H))_{hijk}\,, \tag{16}$$

$$\mu = \frac{1}{(n-2)(n-1)}\,(\kappa - 2\widetilde{S}_{rt} B^r_h B^t_k g^{hk} + \widetilde{\kappa})\,, \tag{17}$$

where $\widetilde{S}_{rt}$ are the local components of the Ricci tensor $\widetilde{S}$ of the ambient space, G_{hijk} are the local components of the tensor G and $\widetilde{\kappa}$ and κ are the scalar curvatures of (N, g^N) and (M, g), respectively. From (16) we find

$$C \cdot H = \frac{\varepsilon}{n-2}(Q(g,H^3) + (n-3)Q(H,H^2) - tr(H)Q(g,H^2)) + \mu\, Q(g,H)\,, \tag{18}$$

$$C \cdot H^2 = \varepsilon(Q(H,H^3) + \frac{1}{n-2}(Q(g,H^4) - tr(H)Q(g,H^3)$$

$$- tr(H)Q(H,H^2))) + \mu\, Q(g,H^2)\,. \tag{19}$$

Let now M be a hypersurface in a semi-Riemannian space of constant curvature $N^{n+1}_s(c)$, $n \geq 4$. Clearly, (15) and (17) turn into

$$R_{hijk} = \varepsilon\, \overline{H}_{hijk} + \frac{\widetilde{\kappa}}{n(n+1)}\, G_{hijk}\,, \tag{20}$$

$$\mu = \frac{1}{n-2}\left(\frac{\kappa}{n-1} - \frac{\widetilde{\kappa}}{n+1}\right), \tag{21}$$

respectively, where $c = \frac{\widetilde{\kappa}}{n(n+1)}$. Contracting (20) with g^{ij} and g^{kh}, respectively, we obtain

$$S_{hk} = \varepsilon\,(tr(H)\,H_{hk} - H^2_{hk}) + \frac{(n-1)\widetilde{\kappa}}{n(n+1)}\, g_{hk}\,, \tag{22}$$

$$\kappa = \varepsilon\,((tr(H))^2 - tr(H^2)) + \frac{(n-1)\widetilde{\kappa}}{n+1}\,, \tag{23}$$

respectively, where $tr(H) = g^{hk}H_{hk}$, $tr(H^2) = g^{hk}H^2_{hk}$ and S_{hk} are the local components of the Ricci tensor S of M. Using (22) and Theorem 4.1 of [15] we can deduce that $\mathcal{U}_H \subset \mathcal{U}_S \cap \mathcal{U}_C \subset M$. It is known that at every point of a hypersurface M in $N^{n+1}_s(c)$, $n \geq 4$, the following condition of pseudosymmetry type is fulfilled ([15]): the tensors $R \cdot R - Q(S,R)$ and $Q(g,C)$ are linearly dependent. Precisely, on M we have

$$R \cdot R - Q(S,R) = -\frac{(n-2)\widetilde{\kappa}}{n(n+1)}\, Q(g,C)\,. \tag{24}$$

Evidently, if the ambient space is $\mathbb{E}^{n+1}_s$ then (24) reduces to $R \cdot R = Q(S,R)$. Similarly, in this case, (20) and (21) reduce to

$$(25) \qquad (a) \quad R_{hijk} = \varepsilon\, \overline{H}_{hijk}\,, \qquad (b) \quad \mu = \frac{\kappa}{(n-2)(n-1)}\,.$$

Let M be a hypersurface in semi-Riemannian space $N^{n+1}_s(c)$, $n \geq 3$. We recall that U_H is the set consisting of all points of M at which the tensor H^2 is not a linear combination of H and g. It is known that on the set $M - U_H$ the tensors $R \cdot R$ and $Q(g,R)$ are linearly dependent (cf. [13], Proposition 3.1(ii)). Thus we see that if for a hypersurface M in $N^{n+1}_s(c)$, $n \geq 3$, its set U_H is empty then M is pseudosymmetric. In particular, if at every point of a hypersurface M, in a Riemannian space form, there are at most two distinct principal curvatures then M is pseudosymmetric.

PROPOSITION 4.1 *Let M be a hypersurface in $N^{n+1}_s(c)$, $n \geq 4$, satisfying (1). Then on $\mathcal{U}_H \cap \mathcal{U}_L \subset M$ we have*

$$(26) \qquad H^3 = tr(H)\, H^2 + \lambda\, H\,, \quad \lambda = \frac{1}{n-1}(tr(H^2) - (tr(H))^2)\,,$$

$$(27) \qquad \varepsilon\lambda = \frac{\tilde{\kappa}}{n+1} - \frac{\kappa}{n-1}\,.$$

PROOF. We note that (1), in view of Corollary 4.1 of [8], implies (15) from that paper, so we have

$$(28) \qquad H^3 = tr(H)\, H^2 + \lambda\, H + \rho g\,.$$

In addition, on the set $\mathcal{U}_H \cap \mathcal{U}_L$ we have ([13], Proposition 3.9) the equality $C \cdot S = 0$, i.e., $tr(H)\, C \cdot H - C \cdot H^2 = 0$. Applying to this (18) and (19) we find

$$\begin{aligned} Q(g,H^4) = 2tr(H)\, Q(g,H^3) - (\varepsilon(n-2)\mu + (tr((H)^2)\, Q(g,H^2) \\ + \varepsilon(n-2)\mu tr(H) Q(g,H) + (n-2) Q(H, tr(H)H^2 - H^3)\,, \end{aligned}$$

and, in virtue of (28) also

$$Q(g, H^4 - 2tr(H)H^3 + (\varepsilon(n-2)\mu + (tr(H)^2)H^2 - (n-2)(\varepsilon\mu tr(H) + \rho)H) = 0\,.$$

This leads to

$$H^4 = 2tr(H)H^3 - (\varepsilon(n-2)\mu + (tr(H)^2)H^2 + (n-2)(\varepsilon\mu tr(H) + \rho)H + \overline{\lambda} g\,.$$

But (28) implies $H^4 = tr(H)H^3 = \lambda H^2 + \rho H$, so we have

$$tr(H)H^3 - (\varepsilon(n-2)\mu + \lambda + (tr(H)^2)H^2 + (\varepsilon(n-2)\mu tr(H) + (n-3)\rho)H + \overline{\lambda} g = 0.$$

Comparing this equality with (28), we obtain

$$-(\varepsilon(n-2)\mu+\lambda)H^2+(tr(H)(\varepsilon(n-2)\mu+\lambda)+(n-3)\rho)H+(\overline{\lambda}+tr(H)\rho)g=0\,.$$

Whence, in view of $x\in\mathcal{U}_H$, we immediately get

$$(a)\ \ \varepsilon(n-2)\mu+\lambda=0\,,\qquad (b)\ \ \rho=0 \tag{29}$$

and $\overline{\lambda}=0$. Thus (29)(b) and (28) lead to the first equality of (26). Finally, applying to (29)(a), (21) and (23) we obtain $\lambda=\frac{1}{n-1}\,((tr(H)^2-tr(H^2))$ and (27). This completes the proof. ∎

From Lemma 1.1 of [4] we have

PROPOSITION 4.2 *Let M be a pseudosymmetric hypersurface in $N^{n+1}_s(c)$, $n\geq 4$. Then on $\mathcal{U}_H\subset M$ we have*

$$H^3=tr(H)\,H^2+\lambda\,H\,,\qquad \lambda=\frac{1}{2}(tr(H^2)-(tr(H)^2)\,.$$

COROLLARY 4.1 *Let M be a pseudosymmetric hypersurface in $N^{n+1}_s(c)$ satisfying (1). Then on the set $\mathcal{U}_H\cap\mathcal{U}_L\subset M$ we have $\lambda=0$, i.e.*

$$\frac{\kappa}{n-1}=\frac{\widetilde{\kappa}}{n+1}\,. \tag{30}$$

We recall that (24) holds on every hypersurface M in $N^{n+1}_s(c)$. Using now (8) we have

$$R\cdot C=Q(S,R)-\frac{(n-2)\widetilde{\kappa}}{n(n+1)}\,Q(g,R)+\frac{(n-3)\widetilde{\kappa}}{(n-2)n(n+1)}\,Q(g,g\wedge S)\,.$$

Thus if M satisfies (1) then taking into account also (9) we get

$$\begin{aligned}(L-1)\,Q(S,R)=&-\frac{(n-2)\widetilde{\kappa}}{n(n+1)}\,Q(g,R)-\frac{L}{n-2}\,Q(g,\overline{S})\\ &+\frac{1}{n-2}\left(\frac{L\kappa}{n-1}+\frac{(n-3)\widetilde{\kappa}}{n(n+1)}\right)Q(g,g\wedge S)\,.\end{aligned} \tag{31}$$

PROPOSITION 4.3 *Let M be a hypersurface in $N^{n+1}_s(c)$ satisfying $R\cdot C=$ $=Q(S,C)$. If $\mathcal{U}_H\subset M$ is nonempty then the ambient space must be semi-Euclidean.*

PROOF. Under our assumption (31) turns into

$$\begin{aligned}&\frac{(n-2)\widetilde{\kappa}}{n(n+1)}\,Q(g,R)+\frac{1}{n-2}\,Q(g,\overline{S})-\\ &\qquad-\frac{1}{n-2}\left(\frac{\kappa}{n-1}+\frac{(n-3)\widetilde{\kappa}}{n(n+1)}\right)Q(g,g\wedge S)=0\,.\end{aligned} \tag{32}$$

Let $x \in \mathcal{U}_H$ and suppose that $\widetilde{\kappa} \neq 0$. Thus we can rewrite (32) in the form

$$Q(g, R + \frac{n(n+1)}{(n-2)^2\widetilde{\kappa}}\overline{S} - \frac{n(n+1)}{(n-2)^2\widetilde{\kappa}}\left(\frac{\kappa}{n-1} + \frac{(n-3)\widetilde{\kappa}}{n(n+1)}\right) g \wedge S) = 0$$

whence, we have

$$R = -\frac{n(n+1)}{(n-2)^2\widetilde{\kappa}}\overline{S} + \frac{n(n+1)}{(n-2)^2\widetilde{\kappa}}\left(\frac{\kappa}{n-1} + \frac{(n-3)\widetilde{\kappa}}{n(n+1)}\right) g \wedge S + \eta\, G$$

for some $\eta \in \mathbb{R}$. But such decomposition of R implies pseudosymmetry of M (see Theorem 4.2 of [9]). On the other hand, for every pseudosymmetric hypersurface M in $N_s^{n+1}(c)$ we have (5) on $\mathcal{U}_H$. Applying now (30) to (5) and (24) we get (12)(i) and (12)(ii). Next using Lemma 3.3 we obtain (14). Comparing the two obtained decompositions of R, in view of Lemma 3.2, we easily get a contradiction. This completes the proof. ∎

PROPOSITION 4.4 *Let M be a hypersurface in $N_s^{n+1}(c)$, $c \neq 0$, $n \geq 4$, satisfying (1). Then at every point $x \in \mathcal{U}_H \cap \mathcal{U}_L \subset M$ we have (3) and*

$$\underset{X,Y,Z}{\mathcal{S}}\ w(X)\mathcal{C}(Y,Z) = 0\,. \tag{33}$$

PROOF. According to Proposition 4.3, $c \neq 0$ implies $L \neq 1$ and we can rewrite (31) in the form

$$\begin{aligned} &Q(S,R) = \psi_1\, Q(g,R) + \psi_2\, Q(g,\overline{S}) + \psi_3\, Q(S,G)\,, \\ &\psi_1 = -\frac{(n-2)\widetilde{\kappa}}{n(n+1)(L-1)}\,, \quad \psi_2 = -\frac{L}{(n-2)(L-1)}\,, \\ &\psi_3 = -\frac{1}{(n-2)(L-1)}\left(\frac{L\kappa}{n-1} + \frac{(n-3)\widetilde{\kappa}}{n(n+1)}\right). \end{aligned} \tag{34}$$

We note that $\psi_1 \neq 0$ ($\widetilde{\kappa} = 0$, in view of Theorem 4.1 of [13], leads to $L = 1$) and (34) is equivalent to $Q\big(S - \psi_1 g, R - \psi_3 G + \frac{\psi_2}{\psi_1}\overline{S}\big) = 0$. According to Lemma 3.1 we have a priori two cases:

(a) $S - \psi_1 g = \beta\, w \otimes w$. In this case we have also

$$\underset{X,Y,Z}{\mathcal{S}}\ w(X)\mathcal{B}(Y,Z) = 0\,, \tag{35}$$

where $B = R - \psi_3 G + \frac{\psi_2}{\psi_1}\overline{S}$. But $\overline{S} = \psi_1^2 G + \psi_1\beta(g \wedge w \otimes w)$ and $B = R - (\psi_3 - - \psi_1\psi_2)G + \psi_2\beta(g \wedge w \otimes w)$. It is easy to see that the generalized curvature tensor P, defined by $P = g \wedge w \otimes w$ satisfies $\mathcal{S}_{X,Y,Z}\, w(X)\mathcal{P}(Y,Z) = 0$. Thus (35) implies that also $\mathcal{S}_{X,Y,Z}\, w(X)\mathcal{B}_1(Y,Z) = 0$, where $B_1 = R - (\psi_3 - - \psi_1\psi_2)G$. Applying now Theorem 4.1 of [9] we obtain our assertion.

(b) $R - \psi_3 G + \frac{\psi_2}{\psi_1} \overline{S} = \widetilde{\eta}(S - \psi_1 g) \wedge (S - \psi_1 g)$, $\widetilde{\eta} \in \mathbb{R}$. This equality can be written in the form $R = (\eta - \frac{\psi_2}{\psi_1})\overline{S} - \eta\psi_1\, g \wedge S + (\psi_3 + \eta\psi_1^2)G$, where $\eta = 2\widetilde{\eta}$. In the same manner as in the proof of Proposition 4.3 we can show that this leads to a contradiction. ∎

THEOREM 4.1 *Let M be a hypersurface in $N_s^{n+1}(c)$, $c \neq 0$, satisfying (1). Then at every point $x \in \mathcal{U}_H \cap \mathcal{U}_L \subset M$ we have (3), (30) and*

$$L = \frac{1}{n-1}, \quad R \cdot R = \frac{\kappa}{n(n-1)}\, Q(g,R), \quad R \cdot R = Q(S,R) - \frac{(n-2)\kappa}{n(n-1)}\, Q(g,C).$$

PROOF. Using Proposition 4.4 we get (3) and (33). Applying Proposition 3.1 we complete the proof. ∎

At the end of this section we present a corrected version of some results of [13]. The equality (17) is the corrected version of a formula from [15] defining the function μ. Unfortunately, in the proof of Proposition 3.10 of [6] was used that false formula. However, if we apply (17) in the proof of Proposition 3.10 of [13] then we do not obtain new results. Precisely, (45) and (47) of Proposition 3.10 of [13] are equivalent to (39) and (40) of Proposition 3.9 of [13], respectively. Thus we see that Proposition 3.10 is superfluous. Further, (56) of [13] must be removed and (61)(b) of [13] should be replaced by (27) from this paper. Therefore, Theorem 4.3 of [13] has the following form:

THEOREM 4.2 *If M is a hypersurface in $\mathbb{E}_s^{n+1}$, $n \geq 4$, fulfilling (1) then on $\mathcal{U}_H \cap \mathcal{U}_L \subset M$ we have:*

$$R \cdot S = 0, \quad C \cdot S = 0, \quad R \cdot C = Q(S,C),$$

$$C \cdot R = \frac{n-3}{n-2}\, Q(S,R), \quad \mathcal{A}^3 = tr(\mathcal{A})\mathcal{A}^2 - \frac{\varepsilon\,\kappa}{n-1}\mathcal{A}, \quad \varepsilon = \pm 1,$$

$$(36) \quad \mathcal{A}(W) = 0, \quad S = \frac{\kappa}{n-1}\, g + \beta\, w \otimes w, \quad w \in T_x^* M, \quad \beta \in \mathbb{R},$$

where the vector W is related to w by $g(W,X) = w(X)$, $X \in T_x M$.

Thus we see that the part of the assertion of Theorem 4.3 of [13], stating that if (1) is fulfilled on a hypersurface M in $N_s^{n+1}(c)$, $n \geq 4$, with nonempty set $\mathcal{U}_H \cap \mathcal{U}_L \subset M$, then the ambient space must be semi-Euclidean is false, in general. However, as we prove in Proposition 4.3, the above statement is true in the case when $R \cdot C = Q(S,C)$, i.e. $L = 1$, and in view of Theorem 4.1 of [13], only in this case. Examples of hypersurfaces M in $\mathbb{E}_s^{n+1}$, $n \geq 4$, satisfying (36), with nonempty set $\mathcal{U}_H \cap \mathcal{U}_L$, were found in [6]. Furthermore, as we present in Example 5.1, there are hypersurfaces in $N_s^{n+1}(c)$, $c \neq 0$, $n \geq 4$, fulfilling (1) with nonempty set $\mathcal{U}_H \cap \mathcal{U}_L \subset M$.

5. Examples

Let now M be a hypersurface in $N_s^{n+1}(c)$, $n \geq 4$, such that $\mathcal{U}_H \subset M$ is nonempty. From Theorem 5.1 of [5] it follows: (10) holds at a point $x \in \mathcal{U}_H$ if and only if at this point $rank\, H = 2$, i.e. the type number of M at this point is equal to 2. On any hypersurface in $N_s^{n+1}(c)$ we have the following identity (see e.g. [4], eq. (22))

$$R \cdot R - \frac{\widetilde{\kappa}}{n(n+1)}\, Q(g,R) = -Q(H^2, \overline{H})\,. \tag{37}$$

Further, it is known (see [4], Lemma 1.1) that if $rank\, H = 2$ at a point $x \in U_H$ then

$$Q(H^2, \overline{H}) = 0\,. \tag{38}$$

Therefore, if $rank\, H = 2$ at a point $x \in U_H$ then, by (38), (37) reduces to (5). It is easy to see that (5) implies

$$(a)\;\; R \cdot S = \frac{\widetilde{\kappa}}{n(n+1)}\, Q(g,S)\,, \qquad (b)\;\; R \cdot C = \frac{\widetilde{\kappa}}{n(n+1)}\, Q(g,C)\,. \tag{39}$$

We note also that in the particular case, if M is a hypersurface in a semi-Euclidean space $\mathbb{E}_s^{n+1}$, $n \geq 3$, then (5) reduces on $\mathcal{U}_H \subset M$ to $R \cdot R = 0$.

PROPOSITION 5.1 *Let M be a pseudosymmetric hypersurface in $N_s^{n+1}(c)$, $n \geq 4$.*
(i) (6)(a) or (6)(b) is satisfied at every point $x \in \mathcal{U}_H \subset M$.
(ii) If $c \neq 0$ and at $x \in \mathcal{U}_H \subset M$ we have (6)(b) then (4) cannot be satisfied at this point.
(iii) If $c \neq 0$ and at a point $x \in \mathcal{U}_H \subset M$ we have (6)(a), i.e.

$$H^2_{ij} - tr(H) H_{ij} = \frac{1}{\rho}\, a_i a_j\,, \quad \rho \in \mathbb{R} - \{0\}\,, \tag{40}$$

then the following relations are fulfilled at x: (4), (30) and

$$(a)\;\; a^k a_k = 0\,, \quad a^k = g^{jk} a_j\,, \quad (b)\;\; S_{ij} - \frac{\kappa}{n} g_{ij} = \frac{\varepsilon}{\rho}\, a_i a_j\,, \tag{41}$$

PROOF. (i) The relation (38), by the identity $Q(H, \overline{H}) = 0$, yields on $\mathcal{U}_H \subset M$

$$Q(H^2 - tr(H)\, H, \overline{H}) = 0\,. \tag{42}$$

Now, applying to (42) Lemma 3.1 and Lemma 1.1 of [4] we obtain easily our assertion.
(ii) Applying (23) into (22) we find

$$S - \frac{\kappa}{n} g = \varepsilon (tr(H)\, H - H^2) - \frac{\varepsilon}{n}((tr(H))^2 - tr(H^2))\, g\,. \tag{43}$$

We assume that (4) holds at x. Now, in view of Proposition 4.4, we have (3) at x. Thus (43) turns into

$$(44)\ \frac{\varepsilon}{n}((tr(H))^2 - tr(H^2))\, g = \varepsilon(tr(H)\, H - H^2) - \beta\, w \otimes w, \quad \beta \in \mathbb{R}, \quad w \in T^*_x M.$$

Applying to this (6)(b) we deduce that $(tr(H))^2 - tr(H^2) = 0$. Now (44) yields (6)(a), a contradiction with (6)(b).
(iii) The assumption that $x \in U_H$ implies $rank\, H = 2$ at x. This, in view of Lemma 1.1 of [4], gives

$$(45) \qquad H^3_{ij} = tr(H) H^2_{ij} + \lambda\, H_{ij}\, , \quad \lambda \in \mathbb{R}.$$

Further, (40) and (42), in view of Lemma 3.1(i), imply

$$(46) \qquad a_l\, \overline{H}_{hijk} + a_h\, \overline{H}_{iljk} + a_i\, \overline{H}_{lhjk} = 0\, .$$

On the other hand, transvecting (40) with $H_l{}^i$ we get

$$(47)(a)\ \ H^3_{lj} = tr(H) H^2_{lj} + \frac{1}{\rho}\, a^k H_{kl}\, a_j\, , \qquad (b)\ \ a^k H_{kl} = \phi\, a_l\, , \qquad \phi \in \mathbb{R}.$$

Now (47)(a), by (40) and (47)(b), turns into $H^3_{lj} = (tr(H) + \phi) H^2_{lj} - - \phi tr(H) H_{lj}$. This, together with (45), yields $\phi H^2_{lj} = (\lambda + \phi tr(H)) H_{lj}$. Since $x \in \mathcal{U}_H$, the last relation implies $\phi = 0$ and $\lambda = 0$. Thus (45) and (47)(b) reduce to

$$(48) \qquad (a)\ \ H^3_{ij} = tr(H)\, H^2_{ij}\, , \qquad (b)\ \ a^k H_{kl} = 0\, ,$$

respectively. Transvecting now (46) with a^l and using (48)(b) we get $a^l a_l \overline{H} = = 0$, whence we obtain (41)(a). Next, contracting (40) and using (41)(a), we get $(tr(H))^2 = tr(H^2)$. Substituting this into (23) we get (30). Applying (30) to (22) we obtain (41)(b). Further, using (20), (30), (41)(b) and (46), we can check that $a_l C_{hijk} + a_h C_{iljk} + a_i C_{lhjk} = 0$ at x. This fact, in view of Lemma 3.6 of [9], implies $Q(a \otimes a, C) = 0$, whence $Q(\frac{\varepsilon}{\rho}\, a \otimes a, C) = 0$ and by making use of (41)(b) we get $Q\big(S - \frac{\kappa}{n}\, g, C\big) = 0$. The last equality, by (30), turns into $Q(S, C) = \frac{(n-1)\widetilde{\kappa}}{n(n+1)}\, Q(g, C)$, which together with (39)(b), yields (4). This completes the proof. ∎

EXAMPLE 5.1. We denote by $(\widetilde{M}, \widetilde{g})$, $\dim \widetilde{M} = n - 1 \geq 3$, the warped product defined in Example 4.1 of [7]. We denote by $\widetilde{R}$, $\widetilde{S}$ and $\widetilde{\kappa}$ its curvature tensor, the Ricci tensor and the scalar curvature, respectively. In [7] it was shown that $(\widetilde{M}, \widetilde{g})$ is semisymmetric $(\widetilde{R} \cdot \widetilde{R} = 0)$ and

$$(49) \qquad rank\, (\widetilde{S}) = 1\, , \quad \widetilde{\kappa} = 0\, ,$$

on $\widetilde{M}$. In addition (see [7], Example 5.1), on $\widetilde{M}$ is defined the Codazzi tensor h such that

$$\widetilde{R} = \frac{\varepsilon}{2} h \wedge h\,, \quad \varepsilon = \pm 1\,, \tag{50}$$

on $\widetilde{M}$. This means that $(\widetilde{M}, \widetilde{g})$ can be locally realized as a hypersurface in $\mathbb{E}^{n+1}_s$. Further, h fulfils

$$rank\,(h) = 2\,, \quad h^3 = tr(h)\, h^2\,. \tag{51}$$

From (50), by a suitable contraction and making use of (49), we find

$$rank\,(h^2 - tr(h)\, h) = 1\,. \tag{52}$$

Let now $\overline{M}$ be an open nonempty interval of $\mathbb{R}$, $\overline{g}_{11} = 1$ the metric tensor on $\overline{M}$ and F the function on $\overline{M}$ defined by $F(x^1) = \exp(ax^1)$, $x^1 \in \overline{M}$, $a > 0$. We consider the warped product $\overline{M} \times {}_F\widetilde{M}$ of $(\overline{M}, \overline{g})$ and $(\widetilde{M}, \widetilde{g})$ with the warping function F. We have

$$\begin{aligned} &\Delta_1 F = \overline{g}^{11} F_1 F_1 = a^2 F^2, \\ &tr(T) = \overline{g}^{11} T_{11} = \overline{g}^{11} \left(\nabla_1 F_1 - \frac{1}{2F} F_1 F_1 \right) = \frac{a^2}{2} F, \end{aligned} \tag{53}$$

where $F_1 = \partial_1 F = \frac{\partial F}{\partial x^1}$. The scalar curvature κ of $\overline{M} \times {}_F\widetilde{M}$ is constant and

$$\kappa = -\frac{(n-1)na^2}{4}\,. \tag{54}$$

The last relation is an immediate consequence of (25) of [7] and (53). Using (23) and (24) of [7] and (54) we get on $\overline{M} \times \widetilde{M}$

$$rank\,(S - \frac{\kappa}{n}\, g) = 1\,. \tag{55}$$

We define on $\overline{M} \times \widetilde{M}$ the $(0,2)$-tensor H, with the local components

$$H_{11} = 0\,, \quad H_{1\alpha} = H_{\alpha 1} = 0\,, \quad H_{\alpha\beta} = \sqrt{F}\, h_{\alpha\beta}\,, \tag{56}$$

where $\alpha, \beta \in \{2, 3, \dots, n\}$. Using the fact that h is a Codazzi tensor, we can check that H is also a Codazzi tensor. Moreover, by making use of (51) and (56) we obtain $rank\,(H) = 2$ and (48)(a). Applying now in (22) of [7] the relations (53), (54) and (56) we get

$$R = \frac{\varepsilon}{2} H \wedge H + \frac{\widetilde{\kappa}}{n(n+1)}\, G\,, \tag{57}$$

where the constant $\widetilde{\kappa}$ is defined by (30). Thus we see that $\overline{M} \times {}_F\widetilde{M}$ can be locally realized as a hypersurface in a semi-Riemannian space of constant curvature $N_s^{n+1}(c)$, $c = \frac{\widetilde{\kappa}}{n(n+1)}$. From (57), by a suitable contraction and making use of (55), we get (6)(a). Finally, in view of Proposition 5.1, we see that $\overline{M} \times {}_F\widetilde{M}$ fulfils (4).

Example 5.2 ([2], Section 6) Let $N^2(c)$ be a minimal surface with nonzero constant curvature c in the unit $(n+1)$- sphere $S^{n+1}(1)$. We denote by M the tubular hypersurface $T_{\frac{\pi}{2}}(N^2(c))$ with radius $\frac{\pi}{2}$ about $N^2(c)$. This hypersurface is called a generalized Cartan hypersurface. Such hypersurface has at every point three principal curvatures: k, $-k$, $0, \dots, 0$, $k \neq 0$. It is easy to check that (6)(b) holds on M. Further, we have $(tr(H))^2 - tr(H^2) = $ $= -tr(H^2) = -2k^2$. Now from (23) it follows that $\frac{\kappa}{n-1} - \frac{\widetilde{\kappa}}{n+1}$ is nonzero on M. Thus we see that (30) is not satisfied on M. Therefore from Corollary 4.1 it follows that (1) cannot be satisfied on M.

References

[1] M. Belkhelfa, R. Deszcz, M. Głogowska, M. Hotloś, D. Kowalczyk and L. Verstraelen, On some type of curvature conditions, in: *Banach Center Publ.* **57**, Inst. Math., Polish Acad. Sci., 2002, 179–194.

[2] B. Y. Chen, A Riemannian invariant for submanifolds in space forms and its applications, in: *Geometry and Topology of Submanifolds*, **VI**, World Sci. Publishing, River Edge, NJ, 1996, 58–81.

[3] B. Y. Chen, Geometry of warped products as Riemannian submanifolds and related problems, *Soochow J. Math.* **28** (2002), 125–156.

[4] R. Deszcz, On certain classes of hypersurfaces in spaces of constant curvature, in: *Geometry and Topology of Submanifolds*, **VIII**, World Sci. Publishing, River Edge, NJ, 1996, 101–110.

[5] R. Deszcz, Pseudosymmetric hypersurfaces in spaces of constant curvature, *Tensor (N.S.)* **58** (1997), 253–269.

[6] R. Deszcz and M. Głogowska, Some examples of nonsemisymmetric Ricci-semisymmetric hypersurfaces, *Colloq. Math.* **94** (2002), 87–101.

[7] R. Deszcz, M. Głogowska, M. Hotloś and Z. Şentürk, On certain quasi-Einstein semisymmetric hypersurfaces, *Ann. Univ. Sci. Budapest, Eötvös Sect. Math.* **41** (1998), 151–164.

[8] R. Deszcz, M. Głogowska, M. Hotloś and L. Verstraelen, On some generalized Einstein metric conditions on hypersurfaces in semi-Riemannian space forms, *Colloq. Math.* **96** (2003), 149–166.

[9] R. Deszcz and M. Hotloś, On a certain subclass of pseudosymmetric manifolds, *Publ. Math. Debrecen* **53** (1998), 29–48.

[10] R. DESZCZ and M. HOTLOŚ, On a certain extension of the class of semisymmetric manifolds, *Publ. Inst. Math. (Beograd) (N.S.)* **63(77)** (1998), 115–130.

[11] R. DESZCZ and M. HOTLOŚ, On some pseudosymmetry type curvature condition, *Tsukuba J. Math.* **27** (2003), 13–30.

[12] R. DESZCZ, M. HOTLOŚ and Z. ŞENTÜRK, On the equivalence of the Ricci-pseudosymmetry and pseudosymmetry, *Colloq. Math.* **79** (1999), 211–227.

[13] R. DESZCZ, M. HOTLOŚ and Z. ŞENTÜRK, Quasi-Einstein hypersurfaces in semi-Riemannian space forms, *Colloq. Math.* **89** (2001), 81–97.

[14] R. DESZCZ, M. HOTLOŚ and Z. ŞENTÜRK, On curvature properties of quasi-Einstein hypersurfaces in semi-Euclidean spaces, *Soochow J. Math.* **27** (2001), 375–389.

[15] R. DESZCZ and L. VERSTRAELEN, Hypersurfaces of semi-Riemannian conformally flat manifolds, in: *Geometry and Topology of Submanifolds*, **III**, World Sci. Publishing, River Edge, NJ, 1991, 131–147.

[16] Z. I. SZABÓ, Structure theorems on Riemannian spaces satisfying $R(X, Y) \cdot R = $ $= 0$ I., The local version, *J. Differential Geom.* **17** (1982), 531–582.

Ryszard Deszcz
Department of Mathematics
Agricultural University of Wrocław
Grunwaldzka 53
50-357 Wrocław
Poland
rysz@ozi.ar.wroc.pl

Marian Hotloś
Institute of Mathematics
Wrocław University of Technology
Wybreże Wyspiańskiego 27
50-370 Wrocław
Poland
hotlos@im.pwr.wroc.pl

POWER INTEGRAL BASES IN ORDERS OF COMPOSITE FIELDS II

By

PÉTER OLAJOS

(Received May 8, 2003)

1. Introduction

Let K be an algebraic number field of degree n with ring of integers $\mathbb{Z}_K$. It is a classical problem in algebraic number theory to decide if there is an element α in K such that

$$\{1, \alpha, \alpha^2, \dots, \alpha^{n-1}\}$$

is an integral basis. Such an integral basis is called *power integral basis*. A further problem is to find all elements which generate power integral bases.

The index of a primitive algebraic integer α of K is defined as the module-index

$$I(\alpha) = (\mathbb{Z}^+{}_K : \mathbb{Z}^+[\alpha]).$$

Obviously α generates a power integral basis if and only if $I(\alpha) = 1$.

Note that

$$I(\alpha) = \frac{\left| \prod\limits_{1 \le j < k \le n} \left(\alpha^{(j)} - \alpha^{(k)} \right) \right|}{\sqrt{|D_K|}} \tag{1}$$

where $\alpha^{(i)}$ $(i = 1, \dots, n)$ are the conjugates of α and D_K is the discriminant of K.

Using Baker's method the first explicit bounds for the absolute values of the solutions of index form equations were given by K. Győry [7]. These

Research supported in part by Grant T-037367 from the Hungarian National Foundation for Scientific Research

upper bounds imply that up to equivalence there are only finitely many generators of power integral bases. For a detailed discussion of algorithmic results on calculating generators of power integral bases see the monograph I. Gaál [3].

Higher degree fields having subfields are very often given as composites of certain subfields. The problem of existence of power integral bases in such fields was investigated in [1], [4] and [5]. The purpose of this paper is to add some new results to this area and consider new applications, involving also infinite parametric families of fields, which are not covered by the former results.

2. New results

Let $f, g \in \mathbb{Z}[x]$ be distinct monic irreducible polynomials (over $\mathbb{Q}$) of degrees m and n, respectively. Let φ be a root of f and let ψ be a root of g. Set $L = \mathbb{Q}(\varphi)$, $M = \mathbb{Q}(\psi)$ and assume that the composite field $K = LM$ has degree mn. Denote by $d(f), d(g)$ the discriminants of f and g. Further we assume that there exists a square-free number q, such that f is a perfect power modulo q, that is

$$f(x) \equiv (x - t)^m \pmod{q} \tag{2}$$

with some $t \in \mathbb{Z}$.

Remark 1. If $\gcd(d(g), d(f)) = 1$, then the results of [1] are hardly applicable for higher degree number fields. If condition (2) is satisfied, we can draw conclusions on the existence of power integral bases even in higher degree fields. Further, the result of [4] are also not applicable, because g is an arbitrarily polynomial, so usually there are no square-free numbers p, q such that f and g are congruent to x^m and x^n modulo q and p, respectively.

Remark 2. The case $d = \gcd(d(f), d(g)) \neq 1$ have already been considered in [5]. In this case both f and g have a multiple linear factor modulo q, where q is a prime divisor of d. The result given below gives a condition when [5] is not applicable.

Consider the order $\mathcal{O}_f = \mathbb{Z}[\varphi]$ of the field L, the order $\mathcal{O}_g = \mathbb{Z}[\psi]$ of the field M and the composite order $\mathcal{O}_{fg} = \mathcal{O}_f \mathcal{O}_g = \mathbb{Z}[\varphi, \psi]$ in the composite field $K = ML$. Obviously $\{1, \varphi, \ldots, \varphi^{m-1}\}$, $\{1, \psi, \ldots, \psi^{n-1}\}$ and

$$\{1, \varphi, \ldots, \varphi^{m-1}, \psi, \varphi\psi, \ldots, \varphi^{m-1}\psi, \ldots, \psi^{n-1}, \varphi\psi^{n-1}, \ldots, \varphi^{m-1}\psi^{n-1}\},$$

are $\mathbb{Z}$-bases of $\mathcal{O}_f$, $\mathcal{O}_g$ and $\mathcal{O}_{fg}$, respectively.

Our main result is the following:

THEOREM 1. *If there exists a power integral basis in $\mathcal{O}_{fg}$, then the congruence*

$$(d(g))^{m(m-1)} \equiv \pm 1 \pmod{q} \tag{3}$$

is satisfied.

As a consequence we have:

THEOREM 2. *If (3) is not satisfied, then $\mathcal{O}_{fg}$ does not admit any power integral basis.*

3. Proof of Theorem 1

Denote by $\varphi^{(i)}$ $(1 \leq i \leq m)$ the conjugates of $\varphi \in L$ and by $\psi^{(j)}$ $(1 \leq j \leq n)$ the conjugates of $\psi \in M$. Denote by $\gamma^{(i,j)}$ the conjugate of any element $\gamma \in K$ under the automorphism mapping φ to $\varphi^{(i)}$ and ψ to $\psi^{(j)}$ $(1 \leq i \leq m,\ 1 \leq j \leq n)$. Denote by N the smallest normal extension of K and let $\mathfrak{q}_0$ be a prime ideal of N lying above a prime divisor q_0 of q.

Since $f(x) \equiv (x-t)^m \pmod{q}$, hence $f(x) = \prod_{j=1}^{m}(x - \varphi_j) \equiv$ $\equiv (x-t)^m \pmod{\mathfrak{q}_0}$, that is $\varphi_j \equiv t \pmod{\mathfrak{q}_0}$, where $1 \leq j \leq m$.

The discriminants of the polynomials f and g are

$$d(f) = \prod_{1 \leq i < j \leq m} \left(\varphi^{(i)} - \varphi^{(j)}\right)^2$$

$$d(g) = \prod_{1 \leq i < j \leq n} \left(\psi^{(i)} - \psi^{(j)}\right)^2. \tag{4}$$

These are also the discriminants of the bases $\{1, \varphi, \ldots, \varphi^{m-1}\}$ of the order $\mathcal{O}_f$ and $\{1, \psi, \ldots, \psi^{n-1}\}$ of the order $\mathcal{O}_g$, respectively. As it is known (cf. W. Narkiewicz [10]) the discriminant of the order $\mathcal{O}_{fg}$ is

$$D(\mathcal{O}_{fg}) = d(f)^n \cdot d(g)^m. \tag{5}$$

We can represent any element $\alpha \in \mathcal{O}_{fg}$ in the form

$$\alpha = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} x_{ij} \varphi^i \psi^j \tag{6}$$

with $x_{ij} \in \mathbb{Z}$. The index of α corresponding to the order $\mathcal{O}_{fg}$ (that is $(\mathcal{O}_{fg}^{+} : \mathcal{O}_{fg}^{+}[\alpha])$) is

$$I_{\mathcal{O}_{fg}}(\alpha) = \frac{1}{\sqrt{|D(\mathcal{O}_{fg})|}} \prod_{(i_1,j_1)<(i_2,j_2)} \left|\alpha^{(i_1,j_1)} - \alpha^{(i_2,j_2)}\right|$$

where the pairs of indices are ordered lexicographically. Now we rearrange the factors in the product above. Using (4) and (5) we have

$$I_{\mathcal{O}_{fg}}(\alpha) = I_1 \cdot I_2 \cdot I_3,$$

where

$$\begin{aligned} I_1 &= \prod_{i=1}^{m} \prod_{1\le j_1<j_2\le n} \left|\frac{\alpha^{(i,j_1)} - \alpha^{(i,j_2)}}{\psi^{(j_1)} - \psi^{(j_2)}}\right|, \\ I_2 &= \prod_{j=1}^{n} \prod_{1\le i_1<i_2\le m} \left|\frac{\alpha^{(i_1,j)} - \alpha^{(i_2,j)}}{\varphi^{(i_1)} - \varphi^{(i_2)}}\right|, \\ I_3 &= \prod_{\substack{(i_1,j_1)<(i_2,j_2) \\ i_1\neq i_2, j_1\neq j_2}} \left|\alpha^{(i_1,j_1)} - \alpha^{(i_2,j_2)}\right|. \end{aligned} \tag{7}$$

Obviously, the factors of I_1, I_2, I_3 appearing in (7) are algebraic integers. Further, using symmetric polynomials we can see that $I_1, I_2, I_3 \in \mathbb{Z}$. If α generates a power integral basis in $\mathcal{O}_{fg}$, then the index of α is 1, hence by $I_1 \cdot I_2 \cdot I_3 = \pm 1$ we also have $I_1, I_2, I_3 = \pm 1$ which implies that the factors of I_1, I_2, I_3 are in fact units.

For any $1 \le i_1 \neq i_2 \le m$ and $1 \le j_1 \neq j_2 \le n$ we have

$$\left(\alpha^{(i_1,j_1)} - \alpha^{(i_2,j_1)}\right) + \left(\alpha^{(i_2,j_1)} - \alpha^{(i_2,j_2)}\right) + \left(\alpha^{(i_2,j_2)} - \alpha^{(i_1,j_1)}\right) = 0$$

which implies the equation

$$\left(\varphi^{(i_1)} - \varphi^{(i_2)}\right) \varepsilon_{i_1 i_2 j_1 j_2} + \left(\psi^{(j_1)} - \psi^{(j_2)}\right) \eta_{i_1 i_2 j_1 j_2} + \rho_{i_1 i_2 j_1 j_2} = 0 \tag{8}$$

with

$$\varepsilon_{i_1 i_2 j_1 j_2} = \frac{\alpha^{(i_1,j_1)} - \alpha^{(i_2,j_1)}}{\varphi^{(i_1)} - \varphi^{(i_2)}}, \quad \eta_{i_1 i_2 j_1 j_2} = \frac{\alpha^{(i_2,j_1)} - \alpha^{(i_2,j_2)}}{\psi^{(j_1)} - \psi^{(j_2)}},$$

$$\rho_{i_1 i_2 j_1 j_2} = \alpha^{(i_2,j_2)} - \alpha^{(i_1,j_1)}.$$

By the above arguments these elements are units in

$$\mathcal{O} = \mathbb{Z}\left[\varphi^{(i_1)}, \varphi^{(i_2)}, \psi^{(j_1)}, \psi^{(j_2)}\right]$$

where $\left[\mathbb{Q}\left[\varphi^{(i_1)}, \varphi^{(i_2)}, \psi^{(j_1)}, \psi^{(j_2)}\right] : \mathbb{Q}\right] \leq m(m-1)n(n-1)$.

Consider equation (8) modulo $\mathfrak{q}_0$.

By our assumptions $\varphi^{(i_1)} - \varphi^{(i_2)} \equiv 0 \pmod{\mathfrak{q}_0}$, hence by equation (8) we get

$$\psi^{(j_1)} - \psi^{(j_2)} \equiv -\rho_{i_1 i_2 j_1 j_2} \cdot \eta_{i_1 i_2 j_1 j_2}^{-1} \pmod{\mathfrak{q}_o} \tag{9}$$

where $-\rho_{i_1 i_2 j_1 j_2} \cdot \eta_{i_1 i_2 j_1 j_2}^{-1}$ is also a unit in $\mathcal{O}$. This can be done for all $i_1 \neq i_2$, $j_1 \neq j_2$, so multiplying the left and right sides of equation (9) (for all $i_1 \neq i_2$ and $j_1 \neq j_2$) we become

$$(d(g))^{m(m-1)} \equiv \pm 1 \pmod{\mathfrak{q}_0} \tag{10}$$

since on the right side a power of the norm of a unit appears. This is a congruence with rational integers, hence as a consequence we also have

$$(d(g))^{m(m-1)} \equiv \pm 1 \pmod{q_0}. \tag{11}$$

We can prove (11) for all prime divisor q_0 of q, that is (3) must be satisfied.

4. Applications

EXAMPLE 1. *A parametric family of totally real cyclic* sextic fields

One of the most interesting application of Theorem 1 is the case when

$$f(x) = x^3 - (a+1)x^2 + (a+2)x + 1,$$

$$g(x) = x^2 - ax - 1$$

and $m = 3, n = 2$. This family was investigated by Odile Lecacheux [8] which has motivated our present result. We have

$$d(f) = (a^2 - a + 7)^2,$$

$$d(g) = a^2 + 4.$$

Let us consider the polynomial f. We get

$$f(x) - \left(x - \frac{a+1}{3}\right)^3 = \frac{1}{27} \cdot (a^2 - a + 7) \cdot (a + 4 - 9x).$$

Set $q = a^2 - a + 7$ and assume that q is square free. If $a \equiv 2 \pmod 3$, then $\gcd(q, 9) = 9$. Because of it we consider the family when $a \equiv 0, 1 \pmod 3$. Then we have $\gcd(q, 3) = 1$ which means

$$f(x) \equiv \left(x - \frac{a+1}{3}\right)^3 \pmod q.$$

Using congruence (3), by Theorem 1 if there exists a power integral basis in $\mathcal{O}_{fg}$ the following is satisfied:

$$(a^2+4)^6 \equiv (a-3)^6 \equiv \pm 1 \pmod q\ . \tag{3}$$

Using Maple for finding solutions we have the following:

if $a \notin [-840, 840]$, then (3) is not satisfied, so by Theorem 2 there exist no power integral basis in $\mathcal{O}_{fg}$.

Considering the values $|a| < 840$, (3) can only be satisfied for

$$a = -15, -2, 1, 4.$$

EXAMPLE 2. *Composite of a totally real cyclic quintic and a quadratic field*

Another application of Theorem 1 is the case when

$$f(x) = x^5 + a^2x^4 - (2a^3 + 6a^2 + 10a + 10)x^3 +$$

$$(a^4 + 5a^3 + 11a^2 + 15a + 5)x^2 + (a^3 + 4a^2 + 10a + 10)x + 1,$$

$$g(x) = x^2 - ax - 1$$

and $m = 5, n = 2$. The totally real cyclic quintic family generated by a root of f was investigated by E. Lehmer [9], see also cf. I. Gaál and M. Pohst [6]. We have

$$d(g) = a^2 + 4.$$

Let us consider the polynomial f. Set $q = a^4 + 5a^3 + 15a^2 + 25a + 25$ and assume that q is square free. Then we have

$$f(x) \equiv \left(x + \frac{a^2}{5}\right)^5 \pmod q.$$

Using congruence (3), by Theorem 1 if there exists a power integral basis in $\mathcal{O}_{fg}$, then

$$(a^2+4)^{20} \equiv \pm 1 \pmod q \tag{3}$$

is satisfied. Using Maple we have that if $a > 2.4 \cdot 10^{16}$, then (3) is not satisfied, so by Theorem 2 there exist no power integral basis in $\mathcal{O}_{fg}$.

References

[1] I. GAÁL: Power integral bases in composits of number fields, *Canad. Math. Bulletin* **41** (1998), 158–165.

[2] I. GAÁL: Solving index form equations in fields of degree nine with cubic subfields, *J. Symbolic Comput.* **30** (2000), 181–193.

[3] I. GAÁL: *Diophantine equations and power integral bases*, Birkhäuser, Boston, 2002.

[4] I. GAÁL and P. OLAJOS: Recent results on power integral bases of composite fields, *Acta Acad. Paed. Agriensis Sect. Math.*, to appear.

[5] I. GAÁL, P. OLAJOS and M. POHST: Power integral bases in orders of composit fields, *Experimental Math.* **11** (2002), 87–90.

[6] I. GAÁL and M. POHST: Power integral bases in a parametric family of totally real quintics, *Math. Comp.* **66** (1997), 1689–1696.

[7] K. GYŐRY: Sur les polynômes *à* coefficients entiers et de discriminant donné, *Publ. Math.* (Debrecen) **23** (1976), 141–165.

[8] O. LECACHEUX: Unités d´une famille de corps cycliques réels de degré 6 liés *à* la courbe modulaire $X_1(13)$, *J. of Number Theory* **31** (1989), 54–63.

[9] E. LEHMER: Connection between Gaussian periods and cyclic units, *Math. Comp.* **50** (1988), 535–541.

[10] W. NARKIEWICZ: *Elementary and Analytic Theory of Algebraic Numbers*, Springer, 1974.

Gyula Károlyi

University of Debrecen
Institute of Mathematics
H–4010 Debrecen Pf.12., Hungary
olaj@math.klte.hu

A SMALL CONVEX POLYTOPE WITH LONG EDGES, MANY VERTICES AND QUADRANGLE FACES ONLY

By

ZOLTÁN GYENES

(Received June 12, 2003)

It is known (see [1] or [2]) that there exists a 3-dimensional convex polytope of diameter 3 with arbitrary large number of vertices and edges of length at least 1. However, the constructions in these papers contain many triangle faces. We show a construction with quadrangle faces only by proving the following:

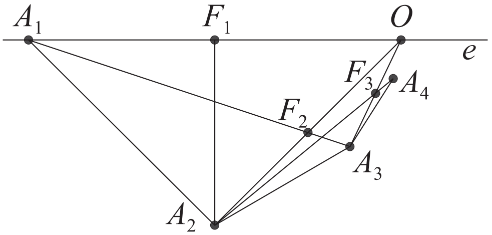
THEOREM 1. *There exists a 3-dimensional convex polytope in a sphere of diameter 3 with 2k+2 ($k \geq 4$) vertices, edge-lengths at least 1 and quadrangle faces only.*

To prove this we verify the following

LEMMA 1. *There exist a convex polygon with 2k ($k \geq 4$) vertices A_1, $A_2, \dots, A_{2k}$ and a point O in the interior of it, so that the intersection of OA_i and $A_{i-1}A_{i+1}$ is the midpoint of the former ($i = 1, 2, \dots, 2k$, $A_{2k+1} = A_1$).*

PROOF. Take an arbitrary "horizontal" line e and two points, A_1 and O on it, the former is on the "left". Let's take A_2 "below" e, so that $A_2A_1 = A_2O$. If we have already chosen A_i ($2 \leq i \leq 2k - 5$), then denoting the midpoint of OA_i by F_i, we choose A_{i+1} on the line of $A_{i-1}F_i$, so that $A_{i-1}A_iA_{i+1}O$ is a convex quadrangle. We choose the A_i's for $1 < i \leq 2k - 4$ in such a way that the orthogonal projection of A_i onto e is between A_1 and O. This is possible, since if this assumption holds for A_i, then it will hold for A_{i+1} as well, provided that A_{i+1} is close enough to F_i.

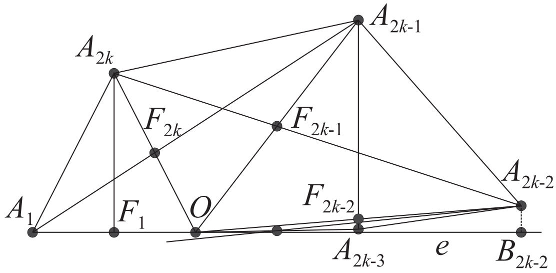
The author was supported by Hung. Nat. Sci. Found. (OTKA) grant no. T032042



Now we choose A_{2k-3} on the line of $A_{2k-5}F_{2k-4}$, so that it is above the line e, and the distance between O and the orthogonal projection of A_{2k-3} onto e is not equal to $\frac{A_1O}{2}$ (this is clearly possible).

We choose A_{2k-2} on the line of $A_{2k-4}F_{2k-3}$, so that $A_{2k-3}F_{2k-2}$ is orthogonal to e (again this is possible). Let us denote the distance of A_{2k-2} from e by a_{2k-2}, the orthogonal projection of A_{2k-2} to e by B_{2k-2} and the ratio $\frac{OB_{2k-2}}{A_1O}$ by λ (> 0). We have $\lambda \neq 1$ by the construction of A_{2k-3}.

Finally we choose A_{2k-1} on the line of $A_{2k-3}F_{2k-2}$ "above" e, so that its distance from e is $\frac{(\lambda+2)^2}{2(\lambda-1)^2}a_{2k-2}$ and A_{2k} on the line of A_2F_1 "above" e, so that its distance from e is $\frac{3(\lambda+2)}{2(\lambda-1)^2}a_{2k-2}$. It is easy to check, that this is a proper polygon with the point O. ∎



Now the construction of the polytope is the following. Contract the polygon in lemma 1 from O so that the distances OA_i become smaller than $\sqrt{2}$. Take a parallel plane lying at distance 1 from the plane of the polygon and translate the polygon and the point O into that plane by a shift orthogonal to the planes. Denote the corresponding points by $A'_1, \dots A'_k, O'$. Choose two more points: O'', O''' so that O''', O, O', O'' are on a line in this order and $O'''O = OO' = O'O'' = 1$. Finally, let the vertices of the polytope be: $A_1, A_3, \dots, A_{2k-1}, A'_2, A'_4, \dots, A'_{2k}, O'', O'''$.

Because of the construction of the polygon the faces of this polytope are the quadrangles $O'''A_{i-1}A_{i+1}A'_i$ and $A_iA'_{i-1}A'_{i+1}O''$ (we have to check only that these fourtuples are coplanar and this is true, since the midpoint of $O'''A'_i$ is on $A_{i-1}A_{i+1}$). Thus the edges are the segments $O'''A_i$, $A_iA'_{i+1}$, A'_iA_{i+1}, A'_iO'', all of which are longer than 1 because $O'''A_i$ is longer than $O'''O = 1$, $A_iA'_{i+1}$ is longer, than $OO' = 1$, etc. Finally, the polytope is in the Thales sphere of $O'''O''$. This proves our theorem.

References

[1] K. Böröczky, B. Csikós, *Small convex polytopes with long edges and many vertices* (to appear in Discrete Comput. Geom.)

[2] K. Bezdek, G. Blekherman, R. Connely and B. Csikós, The polyhedral Tammes problem, *Arch. Math.* **76** (2001), 314–320.

Zoltán Gyenes

ON RESTRICTED SET ADDITION IN ABELIAN GROUPS

By

GYULA KÁROLYI*

(Received July 11, 2003)

1. Introduction

Let $G \neq 0$ denote any Abelian group. Define $p(G)$ as the smallest positive integer p for which there exists a nonzero element g of G with $pg = 0$. If no such integer exists, we write $p(G) = \infty$. Thus, $p(G) = \infty$ if and only if G is torsion free, otherwise it is a prime number that equals the order of the smallest nontrivial subgroup of G. In particular, if G is finite, then $p(G)$ is the smallest prime divisor of $|G|$.

For nonempty subsets $A, B \subseteq G$ with $|A| = k$ and $|B| = \ell$, define

$$A + B = \{a + b \mid a \in A, b \in B\}$$

and

$$A \dotplus B = \{a + b \mid a \in A, b \in B, a \neq b\}.$$

If G is torsion free, that is, G is an ordered Abelian group, then the elements of A and B can be enumerated as $a_1 < a_2 < \ldots < a_k$ and $b_1 < < b_2 < \ldots < b_\ell$ such that

$$a_1 + b_1 < a_2 + b_1 < \ldots < a_k + b_1 < a_k + b_2 < \ldots < a_k + b_\ell.$$

Thus we can conclude that $|A + B| \geq k + \ell - 1$ and $|A \dotplus B| \geq k + \ell - 3$. In particular, $|A + A| \geq 2k - 1$ and $|A \dotplus A| \geq 2k - 3$.

According to the Cauchy–Davenport theorem [3], if p is a prime number and $p \geq k + \ell - 1$, then $|A + B| \geq k + \ell - 1$ holds for any $A, B \subseteq \mathbb{Z}/p\mathbb{Z}$ with $|A| = k, |B| = \ell$. This result has been generalized in several ways. In

* Visiting the Rényi Institute of the Hungarian Academy of Sciences. Research partially supported by Hungarian Scientific Research Grants OTKA T043623 and T043631.

particular, the following improvement can be obtained easily from Kneser's theorem [13, 16] or can be proved directly with a combinatorial argument, see [11].

THEOREM 1. *If A and B are nonempty subsets of an Abelian group G such that $p(G) \geq |A| + |B| - 1$, then $|A+B| \geq |A| + |B| - 1$.*

The case of restricted addition is apparently more difficult. In 1994 Dias da Silva and Hamidoune [4] proved the following analogue of the Cauchy–Davenport theorem, thus settling a problem of Erdős and Heilbronn (see [8]).

THEOREM 2. *If A is a k-element subset of the p-element group $\mathbb{Z}/p\mathbb{Z}$, p a prime, then*

$$|A \dotplus A| \geq \min\{p, 2k-3\}.$$

Later Alon, Nathanson and Ruzsa [1, 2] applying the so-called 'polynomial method' gave a simpler proof that also yields

$$|A \dotplus B| \geq \min\{p, |A| + |B| - 2\}$$

if $|A| \neq |B|$. Some lower estimates on the cardinality of $A \dotplus B$ in arbitrary Abelian groups were obtained recently by Lev [14, 15], and also by Hamidoune, Lladó and Serra [10] in the case $A = B$. Moreover, some more refined results in elementary Abelian groups have been proved by Eliahou and Kervaire, see [5, 6, 7].

In [12] we obtained the following extension of the Dias da Silva–Hamidoune theorem.

THEOREM 3. *If A is a k-element subset of an Abelian group G, then*

$$|A \dotplus A| \geq \min\{p(G), 2k-3\}.$$

The aim of the present note is to give a short alternative proof of this result.

2. The Key Idea of the Proof

The case $p(G) = 2$ being fairly obvious, we will assume that $p(G) \geq 3$ in order to avoid some minor technicalities. In this case Theorem 3 is clearly equivalent to the following.

THEOREM 4. *If A is a k-element subset of an Abelian group G with $p(G) \geq 2k - 3$, then*

$$|A \dotplus A| \geq 2k - 3.$$

Indeed, if $2k - 3 > p(G)$, then one can apply Theorem 4 for any subset A' of A with $p(G) = 2|A'| - 3$ to obtain the result; one only has to note that $A' \dotplus A' \subseteq A \dotplus A$.

Since A is contained in a finitely generated subgroup H of G, and obviously $p(H) \geq p(G)$, it is enough to prove Theorem 4 in the case when G is finitely generated. In this case we can write

$$G = G^1 \oplus G^2 \oplus \ldots \oplus G^m,$$

where each group G^i is isomorphic either to the infinite cyclic group $\mathbb{Z}$ or to a cyclic group $\mathbb{Z}/p^\alpha\mathbb{Z}$ with some prime number $p \geq p(G)$ and positive integer α. We have seen that the theorem is true if $G \cong \mathbb{Z}$, and Theorem 2 claims the same if $G \cong \mathbb{Z}/p\mathbb{Z}$ for some prime number p. In view of all this, to prove Theorem 4 it is enough to verify the following two statements.

STATEMENT 5. *If Theorem 4 is valid for the Abelian groups G^1 and G^2, then it also holds for their direct sum $G^1 \oplus G^2$.*

STATEMENT 6. *Let $p \geq 3$ be a prime number and let α be a positive integer. If Theorem 4 is valid for the group $\mathbb{Z}/p^\alpha\mathbb{Z}$, then it also holds for the group $\mathbb{Z}/p^{\alpha+1}\mathbb{Z}$.*

The key observation is that we can verify both statements using the same argument, based on the following notion. Let G^1 and G^2 be two Abelian groups for which we have already verified Theorem 4 for all possible values of k, and let $\varphi : G^1 \times G^1 \to G^2$ be any map. On the set of all ordered pairs (g^1, g^2) $(g^1 \in G^1, g^2 \in G^2)$, define an additive structure G_φ by introducing an operation $+_\varphi$ as follows:

$$\left(g^1, g^2\right) +_\varphi \left(h^1, h^2\right) =: \left(g^1 + h^1, g^2 + h^2 + \varphi\left(g^1, h^1\right)\right).$$

Note that if the map φ is symmetrical, then the operation $+_\varphi$ is commutative. Now Statements 5 and 6 can be easily derived from the following lemma.

LEMMA 7. *Let A be a k-element subset of G_φ such that*

$$2k-3 \le \min\{p(G^1), p(G^2)\}.$$

Then the set

$$A \dot{+} A =: \{a +_\varphi b \mid a, b \in A, a \neq b\}$$

has at least $2k-3$ different elements.

Indeed, letting $\varphi \equiv 0$ we get back the notion of direct sum: $G_\varphi \cong$ $\cong G^1 \oplus G^2$. Since $p(G^1 \oplus G^2) = \min\{p(G^1), p(G^2)\}$, Statement 5 follows immediately. On the other hand, if we choose $G^1 = \mathbb{Z}/p\mathbb{Z}$, $G^2 = \mathbb{Z}/p^\alpha\mathbb{Z}$, and we define

$$\varphi(x+p\mathbb{Z}, y+p\mathbb{Z}) = \begin{cases} 0 & \text{if } x+y<p \\ 1 & \text{otherwise} \end{cases}$$

for $x, y \in \{0, 1, \dots, p-1\}$, then $G_\varphi \cong \mathbb{Z}/p^{\alpha+1}\mathbb{Z}$. Since

$$p(\mathbb{Z}/p^{\alpha+1}\mathbb{Z}) = p(\mathbb{Z}/p^\alpha\mathbb{Z}) = p(\mathbb{Z}/p\mathbb{Z}) = p,$$

Lemma 7, coupled with Theorem 2 implies Statement 6 as well.

It only remains to prove Lemma 7.

3. Preliminary Lemmas

For a set $X \subseteq G_\varphi$ write

$$X^1 = \{g^1 \in G^1 \mid \text{there exists } g^2 \in G^2 \text{ with } (g^1, g^2) \in X\}.$$

We define X^2 in a similar way. For $A, B \subseteq G_\varphi$ we also introduce

$$A + B =: \{a +_\varphi b \mid a \in A, b \in B\}.$$

An immediate consequence of these definitions is the following statement.

PROPOSITION 8. *For arbitrary X, $Y \subseteq G_\varphi$ we have $(X \setminus Y)^1 \supseteq X^1 \setminus Y^1$ and $X^1 \dot{+} X^1 \subseteq (X \dot{+} X)^1 \subseteq X^1 + X^1$.*

The careful reader may observe that the second part of the statement does not remain valid in general if, instead of the projection to the first coordinate, one considers the projection to the second one.

We have to prove that $|A \dotplus A| \geq 2k - 3$ for the k-element set $A \subseteq G_\varphi$. Note that

$$2|A^i| - 3 \leq 2k - 3 \leq p\big(G^i\big)$$

for $i = 1, 2$. Write $A = A_0 \cup C$, where $C = C_1 \cup \ldots \cup C_t$,

$$A_0 = \{(a_i, b_i) \mid 1 \leq i \leq s\},\ C_i = \{(c_i, d_{ij}) \mid 1 \leq j \leq k_i\}$$

for $1 \leq i \leq t$ such that $2 \leq k_1 \leq k_2 \leq \ldots \leq k_t$, and $a_1, \ldots, a_s, c_1, \ldots, c_t$ are pairwise different elements of G^1. In particular, $k = s + k_1 + \ldots + k_t$ and $|A^1| = s + t$. The following easy lemma will be used frequently throughout the proof.

LEMMA 9. *For* $1 \leq \alpha, \beta \leq t, \alpha \neq \beta$ *we have*

$$|C_\alpha \dotplus C_\alpha| \geq 2k_\alpha - 3$$

and

$$|C_\alpha \dotplus C_\beta| \geq k_\alpha + k_\beta - 1.$$

PROOF. Adding $\varphi(c_\alpha, c_\alpha)$ to each element of $C_\alpha^2 \dotplus C_\alpha^2$, we obtain the set $(C_\alpha \dotplus C_\alpha)^2$. Consequently, $|C_\alpha \dotplus C_\alpha| = |(C_\alpha \dotplus C_\alpha)^2| = |C_\alpha^2 \dotplus C_\alpha^2|$. Since

$$2|C_\alpha^2| - 3 = 2k_\alpha - 3 \leq 2k - 3 \leq p\big(G^2\big),$$

the first estimate follows directly from our hypothesis on G^2. Similarly, $(C_\alpha \dotplus C_\beta)^2$ is obtained translating the set $C_\alpha^2 + C_\beta^2$ by $\varphi(c_\alpha, c_\beta)$. In this case we have

$$|C_\alpha^2| + |C_\beta^2| - 1 = k_\alpha + k_\beta - 1 \leq 2k - 5 < p\big(G^2\big),$$

and thus Theorem 1, applied to G^2, immediately implies

$$|C_\alpha \dotplus C_\beta| = |(C_\alpha \dotplus C_\beta)^2| = |C_\alpha^2 + C_\beta^2| \geq k_\alpha + k_\beta - 1.$$

4. Proof of Lemma 7

Assume first that $s = 0$, in which case $A = C = C_1 \cup \ldots \cup C_t$, $k = k_1 + \ldots + k_t$. The numbers $c_i + c_t$ $(1 \le i \le t)$ are t distinct elements of $C^1 + C^1$. It follows from Theorem 1 that $|C^1 \dotplus C^1| \ge 2t - 1$, and thus there is a set I of $t - 1$ pairs (γ, δ) such that the numbers

$$c_i + c_t \ (1 \le i \le t), \ c_\gamma + c_\delta \ ((\gamma, \delta) \in I)$$

are all different. Lemma 9 implies $|C_\gamma \dotplus C_\delta| \ge 1$ for these pairs (γ, δ). It follows that the sets

$$C_i \dotplus C_t \ (1 \le i \le t), \ C_\gamma \dotplus C_\delta \ ((\gamma, \delta) \in I)$$

are pairwise disjoint subsets of $A \dotplus A$. Based on Lemma 9 and the inequalities $k_i \le k_t$ for $1 \le i \le t$, we then indeed obtain

$$\begin{aligned} |A \dotplus A| &\ge \sum_{(\gamma,\delta)\in I} |C_\gamma \dotplus C_\delta| + \sum_{i=1}^{t-1} |C_i \dotplus C_t| + |C_t \dotplus C_t| \\ &\ge (t-1) + \sum_{i=1}^{t-1} (k_i + k_t - 1) + (2k_t - 3) \\ &\ge t - 1 + 2\sum_{i=1}^{t} k_i - (t-1) - 3 = 2k - 3. \end{aligned}$$

In the sequel we may assume that $s \ge 1$. If $t = 0$, that is, $|A_0^1| = s = k$, then we have

$$|A \dotplus A| \ge |A_0^1 \dotplus A_0^1| \ge 2k - 3$$

according to our assumption on the group G^1. Next, if $t = 1$ then we have $3 \le s + 2 \le (k+2) - 2$. Note that in this case $(A \setminus C) \dotplus C = A_0 \dotplus C$ and $C \dotplus C$ are disjoint, since $(g^1, g^2) \in C \dotplus C$ implies $g^1 = c_1 + c_1$, while $g^1 = a_i + c_1$ for some $1 \le i \le s$ if $(g^1, g^2) \in A_0 \dotplus C$. Moreover, the elements $(a_i + c_1, b_i + d_{1j})$ are pairwise different for $1 \le i \le s$, $1 \le j \le k_1$, thus we obtain the estimate

$$\begin{aligned} |A \dotplus A| \ge |A \dotplus C| &= |A_0 \dotplus C| + |C \dotplus C| \\ &\ge sk_1 + (2k_1 - 3) = s(k-s) + 2(k-s) - 3 \\ &= ((k+2) - (s+2))(s+2) - 3 \ge 2k - 3, \end{aligned}$$

as it was to be proved.

Finally, turning to the general case $s \geq 1, t \geq 2$, we can argue as follows. First we claim that there is an index $1 \leq j \leq t-1$ such that

$$a_1 + c_j \notin \{c_1 + c_t, c_2 + c_t, \ldots, c_{t-1} + c_t\}.$$

Indeed, were

$$\{a_1 + c_j \mid 1 \leq j \leq t-1\} = \{c_i + c_t \mid 1 \leq i \leq t-1\},$$

we would get $D + \{d\} = D$ with $D = \{c_1, c_2, \ldots, c_{t-1}\}$ and $d = a_1 - c_t \neq 0$. This would in turn imply that the pairwise different numbers

$$c_1, c_1 + d, c_1 + 2d, \ldots, c_1 + \big(p\big(G^1\big) - 1\big)d$$

all belong to D, which is absurd, since

$$|D| = t - 1 < 2t - 2 \leq k - 3 < 2k - 3 \leq p(G^1).$$

This way we specified $t+1$ pairwise different elements,

$$a_1 + c_j, a_1 + c_t, c_1 + c_t, c_2 + c_t, \ldots, c_{t-1} + c_t$$

of the set $A^1 \dot{+} A^1$ whose cardinality is at least $2(s+t) - 3$, based on our assumption on G^1. From Proposition 8 it follows that $A \dot{+} A$ contains at least

$$2(s+t) - 3 - (t+1) = 2s + t - 4$$

elements (g^1, g^2) such that

$$g^1 \notin \{a_1 + c_j, a_1 + c_t, c_1 + c_t, \ldots, c_{t-1} + c_t\}.$$

Denote the set of these elements by E. Introducing $F_\alpha = \{(a_1, b_1)\} \dot{+} C_\alpha$ for $\alpha = j, t$ we find that E, F_j, F_t and $C_i \dot{+} C_t$ $(1 \leq i \leq t-1)$ are pairwise disjoint subsets of $A \dot{+} A$. Obviously $|F_j| = k_j \geq k_1$ and $|F_t| = k_t$, hence Lemma 9 implies that

$$\begin{aligned}
|A \dot{+} A| &\geq |E| + |F_j| + |F_t| + \sum_{i=1}^{t-1} |C_i \dot{+} C_t| \\
&\geq (2s + t - 4) + k_1 + k_t + \sum_{i=1}^{t-1} (k_i + k_t - 1) \\
&\geq 2s + t - 4 + 2\sum_{i=1}^{t} k_i - (t-1) = 2k - 3.
\end{aligned}$$

This completes the proof of Lemma 7. ∎

References

[1] N. Alon, M.B. Nathanson, and I.Z. Ruzsa, Adding distinct congruence classes modulo a prime, *Amer. Math. Monthly* **102** (1995), 250–255.

[2] N. Alon, M.B. Nathanson, and I.Z. Ruzsa, The polynomial method and restricted sums of congruence classes, *J. Number Th.* **56** (1996), 404–417.

[3] H. Davenport, On the addition of residue classes, *J. London Math. Soc.* **10** (1935), 30–32.

[4] J.A. Dias da Silva and Y.O. Hamidoune, Cyclic spaces for Grassmann derivatives and additive theory, *Bull. London Math. Soc.* **26** (1994), 140–146.

[5] S. Eliahou and M. Kervaire, Sumsets in vector spaces over finite fields, *J. Number Th.* **71** (1998), 12–39.

[6] S. Eliahou and M. Kervaire, Restricted sums of sets of cardinality $1+p$ in a vector space over F_p, *Discrete Math.* **235** (2001), 199–213.

[7] S. Eliahou and M. Kervaire, Restricted sumsets in finite vector spaces: the case $p=3$, *Integers* **1** (2001), Research paper A2, 19 pages (electronic).

[8] P. Erdős and R.L. Graham, Old and New Problems and Results in Combinatorial Number Theory, *L'Enseignement Mathématique*, Geneva, 1980.

[9] G.A. Freiman, Foundations of a Structural Theory of Set Addition, *Translations of Mathematical Monographs* **37** AMS, 1973.

[10] Y.O. Hamidoune, A.S. Lladó, and O. Serra, On restricted sums, *Combin. Prob. Comput.* **9** (2000), 513–518.

[11] Gy. Károlyi, A compactness argument in the additive theory and the polynomial method, to appear in *Discrete Math.* (2004).

[12] Gy. Károlyi, The Erdős–Heilbronn problem in Abelian groups, to appear in *Israel J. Math.* (2004).

[13] M. Kneser, Abschätzungen der asymptotischen Dichte von Summenmengen, *Math. Z.* **58** (1953), 459–484.

[14] V.F. Lev, Restricted set addition in groups. I: The classical setting, *J. London. Math. Soc. (2)* **62** (2000), 27–40.

[15] V.F. Lev, Restricted set addition in groups. II: A generalization of the Erdős-Heilbronn conjecture, *Electron. J. Combin.* **7** (2000), Research paper R4, 10 pages (electronic).

[16] M.B. Nathanson, Additive Number Theory. Inverse Problems and the Geometry of Sumsets, *GTM* **165**, Springer, 1996.

Gyula Károlyi

Department of Algebra and Number Theory
Eötvös University
Pázmány P. sétány 1/C
Budapest, H–1117 Hungary
karolyi@cs.elte.hu

MODELING AND FORECASTING ATMOSPHERIC POLLUTION OVER REGION

By

VITALIY PRUSOV and ANATOLIY DOROSHENKO

(Received 27 November, 2003)

1. Introduction

Mathematical modelling of atmospheric pollution caused by industrial emissions is of great practical importance as it allows not only to estimate danger for humans but also to develop measures and means to reduce unhealthy consequences of atmospheric pollution and possible damages caused by it. These problems can be solved only on the base of complex ecological-meteorological modelling that, on one hand, has a required degree of detailed elaboration of a spectrum of atmospheric motions and, on the other hand, has potential to solve concrete physical tasks related to the transport and dispersion of pollutants in the air. These problems have been intensively discussed and investigated in recent years, especially in the context of future climate changes (see for example [1,2]). The major difficulties arise from the necessity of a fine grid resolution and carrying out long runs (covering many years), which lead to huge computational tasks. So the development of fast but sufficiently accurate numerical algorithms is of great importance, and techniques for speeding up parallel software implementations [3] must be intensively used.

The transport and diffusion of pollutants depend on the characteristics of the underlying surface as well as on atmospheric motions on various scales (wind and turbulence). The horizontal distribution of a pollutant, having been emitted from a source, is mainly determined by the wind field. Wind velocity affects both the distance of the substance spreading and its concentration in plumes by which the particles are transported. Temperature stratification is also of great importance as it determines the stability of the atmosphere, and the latter influences the intensity of turbulence and the thickness of the mixed layer on which the character of vertical dispersion of pollutants depends. Weather conditions influence also the processes of washing out the particles by atmospheric precipitation and absorption by the underlying surface.

The chemical changes of a substance in the atmosphere are connected to such meteorological characteristics as the quantity of water vapour or drops, the air temperature, the intensity of solar radiation and the presence of other atmospheric substances.

Hence, in any research of practical purpose, the task of the analysis and forecast of air pollution distributions cannot be studied within the framework of a hypothetical representation of atmospheric parameters. At the present stage, in the development of computer engineering, there is an opportunity of computing concentrations under real meteorological conditions on the base of complex atmospheric models [4].

2. A complex model of atmospheric state

The fundamental dynamical equations of a malleable medium are based on the universal physical laws of conservation of mass:

$$\frac{D\rho}{Dt} + \rho\,(\nabla \cdot V) = 0; \tag{1a}$$

conservation of momentum:

$$\frac{DV}{Dt} + 2\Omega \times V = -\rho^{-1}\nabla p - g + \nabla \cdot (\nu\,\Pi)\,; \tag{1b}$$

conservation of energy:

$$\rho c_p \frac{DT}{Dt} - \alpha\, T \frac{Dp}{Dt} = \nabla \left(k \nabla T - F^{\mathrm{rad}}\right) + Q_H; \tag{1c}$$

conservation of scalar entities $\Re = (q, q_L, q_w)$:

$$\frac{D\Re}{Dt} = \nabla\,(k\nabla\Re) + Q_q; \tag{1d}$$

and the ideal gas law:

$$p = \rho R T. \tag{1e}$$

In these equations (1a–e) and below the following notations are used:

$\frac{D}{Dt} = \frac{\partial}{\partial t} + V \cdot \nabla$; t is time, ρ is the density of the medium, V is its velocity, p is pressure, T is temperature, q is specific concentration of some ingredient, q_L is specific humidity, q_W is specific water content, Ω is the angular velocity of rotation of reference frame, g is the gravitational acceleration, ν is factor of turbulent viscosity, Π is tensor of tension, c_p is specific heat capacity at constant pressure, α is thermal expansion coefficient, k is the coefficient of turbulent heat conductivity (diffusion), F^{rad} is the density of radiation energy flux, Q_H is the intensity of allocation (absorption) of heat at the expense of phase transitions of moisture, Q_q is the source (sink) of scalars $\Re = (q_L, q_W)$ as a result of phase transitions.

When integrating the system of equations (1), the vector equation ($1b$) for $V = (v_1, v_2, v_3)$ should be split into three scalar equations, corresponding to the three coordinate directions x_1, x_2, x_3. Usually two axes of coordinates x_1, x_2 are chosen to be parallel to a hypothetical (smooth) surface of the Earth (for example, surface of the quiet sea). The first one, x_1 is directed to the East, the second one, x_2 to the North, and the third coordinate axis, x_3 upwards, orthogonal to this hypothetical surface of the Earth.

The existing numerical methods for solving the system (1) (finite difference, finite elements, spectral methods) are based on some discretization of the differential equations by means of projection into some finite dimensional space, a discrete set of values of points (grid). The more grid points we have, the better a function of continuous variable is approximated by the vector of discrete values of this function in nodes of the grid. Hence, on those grids which can be used for a realistic modelling of atmospheric circulation, these methods will actually give solutions only for long-wave processes. At the same time, equations (1a) and (1b) simulate disturbances as sound and gravitational waves which weakly influence meteorological phenomena, but sharply affect the stability of the numerical realization. The question arises naturally, there arises a question, whether it is possible to alter the original system of hydrodynamical equations (1) so that they do not include the solutions corresponding to sound, gravitational and other high-frequency waves, but describe the macroscale disturbances of the circulation.

In [5], maybe for the first time, on the base of an order estimation for the terms in equations (1a) and (1b), a method of ”filtering” of solutions, connected to high-frequency waves, was offered. Following the results of this work, we shall replace the third projection of the equation (1b) for velocity component v_3 by the equation of statics:

$$\frac{1}{\rho}\frac{\partial p}{\partial x_3} = -g\,. \tag{2}$$

As a consequence of assuming hydrostatic balance in the system of equations (1), there is no prognostic equation for the vertical movement.

As system (1) is closed, the vertical velocity should be defined from such a diagnostic equation in which the hydrostatic balance is supported. Meteorological quantities, which are measured directly in operative practice are pressure p, the absolute temperature T, horizontal wind components, v_1 and v_2 and also humidity q. Density ρ is easily determined from p and T with the help of equation (1e). On the other hand, the vertical component of velocity v_3, appearing in the equations of system (1), is not measured, and with the introduction of approximation (2) is not expressed explicitly through other quantities.

Let us express the vertical velocity at any time on distribution of v_1, v_2 and T by means of the equation

$$\text{(3)} \qquad \frac{\partial}{\partial x_3}(\nabla \cdot V) = \frac{g}{C_p T}\nabla \cdot V,$$

which can be obtained by a combination of the first law of thermodynamics for adiabatic processes:

$$\rho C_p \frac{DT}{Dt} - \frac{Dp}{Dt} = 0,$$

and equations (1e), (1a) and (2). To generalize the equation for the case of moist air, it is enough to replace absolute temperature by virtual one in equation (3):

$$T_v \equiv T(1 + 0,6078q - q_L).$$

The transition from the prognostic equation of conservation of mass (1a) to the diagnostic equation (3) violates the closeness of the complete system of equations in our circulation model because the equation of hydrostatics (2) is also diagnostic. Let us add to system (1)–(3) an equation for the tendency of pressure p. To this aim, we shall combine equations (1e), (1a) and (2) as follows:

$$\frac{\partial}{\partial x_3}\left(\frac{\partial p}{\partial t}\right) = g\nabla \cdot (\rho \mathbf{V}) \quad .$$

Integrating the obtained expression from a given height x_3 to the upper boundary of the model domain $x_3 = H$, we shall get:

$$\text{(4)} \quad \left(\frac{\partial p}{\partial t}\right)_{x_3=H} - \left(\frac{\partial p}{\partial t}\right)_{x_3} = (g\rho v_3)_{x_3=H} - (g\rho v_3)_{x_3} + g\int_{x_3}^{H} \bar{\nabla} \cdot (\rho \bar{\mathbf{V}})d\zeta\,,$$

where the overlines refer to the operator of horizontal divergence and the velocity vector. Further, using standard assumptions about the absence of turbulent flows and non-adiabatic conditions at height H above the surface of the earth, which at an order of magnitude exceeds the height of the atmospheric boundary layer, a condition for the upper boundary can be expressed as:

$$\left(\frac{\partial p}{\partial t}\right)_{x_3=H} = \left(\frac{\partial p}{\partial x_3}\frac{\partial x_3}{\partial t}\right)_{x_3=H} = -\,(\rho g v_3)_{x_3=H} \cdot \tag{5}$$

Taking into account the obtained condition in equality (4), this will give the final equation for the pressure tendency:

$$\frac{\partial p}{\partial t} = g\rho v_3 - 2\,(g\rho v_3)_H - g\int_{x_3}^{H} \bar{\nabla}\cdot\left(\rho\bar{\mathbf{V}}\right)d\zeta\,. \tag{6}$$

In order to close the obtained system of equations, it is necessary to define the physical characteristics of the medium, mode of its movement, to establish relationships between thermodynamic variables V, ρ, p, T, q, q_L, q_W and factors of transportation ν, k and, finally, to establish a method of parameterization of source (sink) members F^{rad}, Q_H, Q_q.

To express components $u = v_1$, $v = v_2$, $w = v_3$ of velocity $\mathbf{V}$ in the spherical system of coordinates λ (longitude), ϕ (latitude) and z (height), we can use the relations $x_1 = r\cos\varphi\ \cos\lambda$, $x_2 = r\cos\varphi\ \sin\lambda$, $z = x_3 = r\sin\varphi$. Let also be

$$\sigma = \frac{z - F\left(\lambda,\varphi\right)}{H - F\left(\lambda,\varphi\right)},\ \text{and}\ \overline{w} = \frac{1}{H-F}\left[w - (1-\sigma)\left(\frac{u}{r\,\cos\varphi}\frac{\partial F}{\partial\lambda} + \frac{v}{r}\frac{\partial F}{\partial\varphi}\right)\right],$$

where σ is a reduced value of vertical coordinate which is terrain-following; F is the height of the surface; H is the height above sea level of the upper boundary. The model in the limited area $\overline{G}$ can be presented by means of the following system of equations:

$$\begin{aligned}\frac{\partial u}{\partial t} = {} & -\frac{u}{r\cos\varphi}\frac{\partial u}{\partial\lambda} - \frac{v}{r}\frac{\partial u}{\partial\varphi} - \overline{w}\frac{\partial u}{\partial\sigma} + \left(2\Omega + \frac{u}{r\,\cos\varphi}\right)v\ \sin\varphi - \\ & - \frac{1}{r\cos\varphi}\left[\theta_v\frac{\partial\pi}{\partial\lambda} + (1-\sigma)\,g\frac{\partial F}{\partial\lambda}\right] + \frac{1}{r\cos\varphi}\frac{\partial}{\partial\lambda}\left(\mathrm{K}_G\frac{\partial u}{\partial\lambda}\right) + \\ & + \frac{1}{r}\frac{\partial}{\partial\varphi}\left(\mathrm{K}_G\frac{\partial u}{\partial\varphi}\right) + \frac{1}{(H-F)^2}\frac{\partial}{\partial\sigma}\left(\mathrm{K}_M\frac{\partial u}{\partial\sigma}\right),\end{aligned} \tag{7}$$

$$\frac{\partial v}{\partial t} = -\frac{u}{r\cos\varphi}\frac{\partial v}{\partial\lambda} - \frac{v}{r}\frac{\partial v}{\partial\varphi} - \overline{w}\frac{\partial v}{\partial\sigma} + \left(2\Omega + \frac{u}{r\ \cos\varphi}\right) u\ \sin\varphi -$$

$$-\frac{1}{r}\left[\theta_v\frac{\partial\pi}{\partial\varphi} + (1-\sigma)\,g\frac{\partial F}{\partial\varphi}\right] + \frac{1}{r\cos\varphi}\frac{\partial}{\partial\lambda}\left(\mathrm{K}_G\frac{\partial v}{\partial\lambda}\right) + \frac{1}{r}\frac{\partial}{\partial\varphi}\left(\mathrm{K}_G\frac{\partial v}{\partial\varphi}\right) +$$

$$(8)\qquad + \frac{1}{(H-F)^2}\frac{\partial}{\partial\sigma}\left(\mathrm{K}_M\frac{\partial v}{\partial\sigma}\right),$$

$$(9)\qquad \frac{1}{(H-F)^2}\frac{\partial^2 w}{\partial\sigma^2} + \frac{1}{H-F}\left(\frac{2}{r} - \frac{g}{\pi\theta_V}\right)\frac{\partial w}{\partial\sigma} - \frac{2}{r}\left(\frac{1}{r} + \frac{g}{\pi\theta_V}\right) w =$$

$$= \frac{1}{r\ \cos\varphi}\left\{\left(\frac{1}{r} + \frac{g}{\pi\theta_V}\right)\left[\left(\frac{\partial u}{\partial\lambda} + \frac{\partial v\ \cos\varphi}{\partial\varphi}\right) -\right.\right.$$

$$\left. - \frac{1}{H-F}\left(\frac{\partial F}{\partial\lambda}\frac{\partial u}{\partial\sigma} + \frac{\partial F}{\partial\varphi}\frac{\partial v\ \cos\varphi}{\partial\sigma}\right)\right] -$$

$$\left. - \frac{1}{H-F}\frac{\partial}{\partial\sigma}\left[\frac{\partial u}{\partial\lambda} + \frac{\partial v\ \cos\varphi}{\partial\varphi} - \frac{1-\sigma}{H-F}\left(\frac{\partial F}{\partial\lambda}\frac{\partial u}{\partial\sigma} + \frac{\partial F}{\partial\varphi}\frac{\partial v\ \cos\varphi}{\partial\sigma}\right)\right]\right\},$$

$$(10)\qquad \frac{\partial\theta}{\partial t} = -\frac{u}{r\cos\varphi}\frac{\partial\theta}{\partial\lambda} - \frac{v}{r}\frac{\partial\theta}{\partial\varphi} - \overline{w}\frac{\partial\theta}{\partial\sigma} +$$

$$+ \frac{1}{r\cos\varphi}\frac{\partial}{\partial\lambda}\left(\mathrm{K}_G\frac{\partial\theta}{\partial\lambda}\right) + \frac{1}{r}\frac{\partial}{\partial\varphi}\left(\mathrm{K}_G\frac{\partial\theta}{\partial\varphi}\right) + \frac{1}{(H-F)^2}\frac{\partial}{\partial\sigma}\left(\mathrm{K}_H\frac{\partial\theta}{\partial\sigma}\right) -$$

$$- \frac{L}{\pi}\left(\delta\frac{dq_H}{dt}\right) + Q_K - Q_\mathrm{I} + Q_R,$$

$$(11)\qquad \frac{\partial q}{\partial t} = -\frac{u}{r\cos\varphi}\frac{\partial q}{\partial\lambda} - \frac{v}{r}\frac{\partial q}{\partial\varphi} - \overline{w}\frac{\partial q}{\partial\sigma} + \frac{1}{r\cos\varphi}\frac{\partial}{\partial\lambda}\left(\mathrm{K}_G\frac{\partial q}{\partial\lambda}\right) +$$

$$+ \frac{1}{r}\frac{\partial}{\partial\varphi}\left(\mathrm{K}_G\frac{\partial q}{\partial\varphi}\right) + \frac{1}{(H-F)^2}\frac{\partial}{\partial\sigma}\left(\mathrm{K}_H\frac{\partial q}{\partial\sigma}\right) + \mathrm{M}_X,$$

$$(12)\qquad \frac{\partial q_S}{\partial t} = -\frac{u}{r\cos\varphi}\frac{\partial q_S}{\partial\lambda} - \frac{v}{r}\frac{\partial q_S}{\partial\varphi} - \overline{w}\frac{\partial q_S}{\partial\sigma} +$$

$$+ \frac{1}{r\cos\varphi}\frac{\partial}{\partial\lambda}\left(\mathrm{K}_G\frac{\partial q_S}{\partial\lambda}\right) + \frac{1}{r}\frac{\partial}{\partial\varphi}\left(\mathrm{K}_G\frac{\partial q_S}{\partial\varphi}\right) +$$

$$+ \frac{1}{(H-F)^2}\frac{\partial}{\partial\sigma}\left(\mathrm{K}_H\frac{\partial q_S}{\partial\sigma}\right) + \delta\frac{dq_H}{dt} + M_K - M_\mathrm{I},$$

$$(13)\qquad \frac{\partial q_L}{\partial t} = -\frac{u}{r\cos\varphi}\frac{\partial q_L}{\partial\lambda} - \frac{v}{r}\frac{\partial q_L}{\partial\varphi} - \overline{w}\frac{\partial q_L}{\partial\sigma} +$$

$$+\frac{1}{r\cos\varphi}\frac{\partial}{\partial\lambda}\left(\mathrm{K}_G\frac{\partial q_L}{\partial\lambda}\right)+\frac{1}{r}\frac{\partial}{\partial\varphi}\left(\mathrm{K}_G\frac{\partial q_L}{\partial\varphi}\right)+$$

$$+\frac{1}{(H-F)^2}\frac{\partial}{\partial\sigma}\left(\mathrm{K}_H\frac{\partial q_L}{\partial\sigma}\right)+\frac{1}{\rho\,(H-F)}\frac{\partial\rho\,V_o q_L}{\partial\sigma}-\delta\frac{dq_H}{dt}-M_{\mathrm{I}}$$

$$(14)\qquad \frac{\partial\pi}{\partial\sigma}=-\frac{g\,(H-F)}{\theta_v},$$

$$(15)\frac{\partial p}{\partial t}=g\rho\bar{w}-2\,(g\rho\bar{w})_{\sigma=1}-g\,(H-F)\int\limits_{\sigma}^{1}\frac{1}{r\,\cos\varphi}\left(\frac{\partial\rho u}{\partial\lambda}+\frac{\partial\rho v\cos\varphi}{\partial\varphi}\right)d\zeta\,,$$

$$(16)\frac{\partial k}{\partial t}=-\frac{u}{r\cos\varphi}\frac{\partial k}{\partial\lambda}-\frac{v}{r}\frac{\partial k}{\partial\varphi}-\overline{w}\frac{\partial k}{\partial\sigma}+\frac{K_M}{(H-F)^2}\left[\left(\frac{\partial u}{\partial\sigma}\right)^2+\left(\frac{\partial v}{\partial\sigma}\right)^2\right]-$$

$$-\frac{g}{\theta_v}\frac{K_H}{H-F}\frac{\partial\theta_v}{\partial\sigma}+\frac{2}{(H-F)^2}\frac{\partial}{\partial\sigma}\left(K_M\frac{\partial k}{\partial\sigma}\right)-\varepsilon,$$

$$(17)\qquad \frac{\partial\varepsilon}{\partial t}=-\frac{u}{r\cos\varphi}\frac{\partial\varepsilon}{\partial\lambda}-\frac{v}{r}\frac{\partial\varepsilon}{\partial\varphi}-\overline{w}\frac{\partial\varepsilon}{\partial\sigma}+$$

$$+C_2\frac{\varepsilon}{k}\left\{\frac{K_M}{(H-F)^2}\left[\left(\frac{\partial u}{\partial\sigma}\right)^2+\left(\frac{\partial v}{\partial\sigma}\right)^2\right]-\frac{g}{\theta_v}\frac{K_H}{H-F}\frac{\partial\theta_v}{\partial\sigma}\right\}-$$

$$-C_3\frac{\varepsilon^2}{k}+\frac{C_4}{(H-F)^2}\frac{\partial}{\partial\sigma}\left(K_M\frac{\partial\varepsilon}{\partial\sigma}\right)$$

$$(18)\qquad K_M=C_1k^2\Big/\varepsilon.$$

Here, in addition to some commonly known symbols and those already introduced, the following notations are used: $\pi = C_p\left(p/p_0\right)^{R/C_p}$ is the reduced pressure; θ is the potential temperature; δ is the attribute of presence of condensation of humidity (1: has a place, 0: is absent); V_o is the established speed of precipitation; k, ε-turbulence kinetic energy and its dissipation, respectively; K_G, K_M are the horizontal and vertical turbulent diffusion coefficients for momentum; K_H is the vertical turbulent diffusion coefficient for heat and humidity; $(C_1, C_2, C_3, C_4) = (0{,}09; 1{,}46; 1{,}83; 0{,}42)$ are constants of closure in the atmospheric boundary. Source-sink functions of subgrid scale are denoted as follows: Q_K is the intensity of latent heat release for vapour condensation; Q_{I} is the intensity of latent heat release for evaporation of water; Q_R is the intensity of radiation cooling or heating; M_X

is an additional ”source” of substance caused by chemical reactions; M_K is a source of humidity from condensation; M_{I} is a source of humidity from evaporation.

The mathematical model (7)–(18) differs from known models, widely used in operative practice, in using vertical coordinate σ and due to equations (9), (14) and (15).

3. Mathematical formulation of the problem and its numerical solution

Forecasting meteorological quantities and pollutant concentrations in the bounded atmospheric domain $\overline{G}$ by use of numerical methods is a rather complicated task. In the case of simpler models the splitting method proves to be efficient [6,7,8]. We have a more complex model, where the method of ”unilateral influence” can be used [9]. In other words, as boundary conditions for regional model (7)–(18), we will use results of analysis and forecast obtained by a macroscale (hemispheric or global) model.

Let the state of the atmosphere at point $r = (\lambda, \varphi, \sigma)$ of the macroscale area $G(r) \subset \overline{G}(r)$ be defined by a vector

$$\Re(r,t) = (u, v, w, \pi, T, q, q_S, q_L, k, \varepsilon)$$

of discrete values of the analysis and forecast $\Re\left(r, t^{m+1}\right) = \Re^{m+1}(r)$, received from a macroscale model at time $t = t^{m+1}$ $(m = 0, 1, \ldots, M)$ with a step $\tau = t^{m+1} - t^m$. To compute the atmospheric state in the bounded domain $\overline{G}$ for $\forall t \in \left[t^m, t^{m+1}\right]$, we will solve a task of the following form in vector representation:

$$\frac{\partial \Re}{\partial t} = D\Re, \quad \forall t \in \left[t^m, t^{m+1}\right], \forall r \in \overline{G},$$
$$\Re\left(r, t^{m+1}\right) = \Re^{m+1}(r), \quad m = 0, 1, \ldots, M. \tag{19}$$

Now replace continuum $\overline{G} = \overline{G}(r)$ by a finite set of points by breaking the region $\overline{G}$ into a set of $J-1$ elements of $\Delta\lambda_j$, $K-1$ elements of $\Delta\varphi_k$ and $L-1$ elements of $\Delta\sigma_l$. Let us construct a vector $\{r_{ijk} = \left(\lambda_j, \varphi_k, \sigma_l\right)$, $1 \le j \le J$, $1 \le k \le K$, $1 \le l \le L\}$, called grid. Then we will have:

$$\lambda_J = \lambda_1 + \sum_{\mu=2}^{J-1} \Delta\lambda_\mu, \quad \varphi_K = \varphi_1 + \sum_{\mu=2}^{K-1} \Delta\varphi_\mu, \quad \sigma_L = \sigma_1 + \sum_{\mu=2}^{L-1} \Delta\sigma_\mu .$$

In the domain of definition $\overline{G}$, instead of function $\Re(r,t)$, given on the macroscale grid, we will construct (see Section 4) a function of discrete arguments $\Re\left(r_{jkl}, t^m\right) = \Re^m_{jkl}$ on the regional grid in nodes $\left(\lambda_j, \varphi_k, \sigma_l, t^m\right) \in R$, $1 \le j \le J$, $1 \le k \le K$, $1 \le l \le L$, $1 \le l \le L$. Besides, we construct a grid operator Λ, corresponding to the differential operator D in (19) (see Section 4 for details).

After filling up function $\Re\left(t^{m+1}\right) = \Re^{m+1}$ in the nodes of the regional grid and computing values of the right-hand side functions $f\left(t^{m+1}\right) = f^{m+1} =$ $= \Lambda\Re^{m+1}$, $m = 1, 2, \ldots, M$ in all nodes of the grid $\{\left(\lambda_j, \varphi_k, \sigma_l\right)$, $1 \le j \le J$, $1 \le k \le K$, $1 \le l \le L\}$, we will search for a solution of the problem (19) for $\forall\, t \in \left[t^m, t^{m+1}\right]$ with the formula

$$\begin{aligned}
\Re(t) = \Re^m + \frac{t-t^m}{\tau}\Bigg[\tau f^m + \frac{t-t^m}{4\tau}\Bigg[4\left(\Re^{m+1} - 2\Re^m + \Re^{m-1}\right) - \\
- \tau\left(f^{m+1} - f^{m-1}\right) + \\
+ \frac{t-t^m}{4\tau}\left[5\left(\Re^{m+1} - \Re^{m-1}\right) - \tau\left(f^{m+1} + 8f^m + f^{m-1}\right) - \right. \\
- \frac{t-t^m}{4\tau}\left[2\left(\Re^{m+1} - 2\Re^m + \Re^{m-1}\right) - \tau\left(f^{m+1} - f^{m-1}\right) + \right. \\
+ \frac{t-t^m}{4\tau}\left[3\left(\Re^{m+1} - \Re^{m-1}\right) - \tau\left(f^{m+1} + 4f^m + f^{m-1}\right)\right]\Bigg]\Bigg]
\end{aligned} \tag{20}$$

for each node of the grid, $\left(\lambda_j, \varphi_k, \sigma_l\right)$, $1 \le j \le J$, $1 \le k \le K$, $1 \le l \le L$.

The scheme (20) is easily obtained by means of the Taylor expansion of function $\Re(r,t)$ for nodes $t = t^{m-1}$ and $t = t^{m+1}$ around the node $t = t^m$, taking into account the equation (19). The scheme has interpolation properties, i.e., at $t = t^m$ or $\left(\tau = t - t^m = 0\right)$ and $t = t^{m+1}$ or $\left(\tau = t^{m+1} - t = 0\right)$ the equalities $\Re\left(t^m\right) = \Re^m$ and $\Re\left(t^{m+1}\right) = \Re^{m+1}$ hold, respectively. Hence, by use of this method, the maximal error of the solution of problem (19) by means of (20) is inside the interval $t^m \le t \le t^{m+1}$ and is determined by the order of approximation, i.e., it is equal to $O\left[(\tau)^4\right]$.

From the above statements it is obvious, that:

- advance time t of a forecast depends on advance time t^{M+1} of the forecast gained from the macroscale model;
- a time step $\tau = t^{m+1} - t^{m}$ of defining macroscale information in view of a daily course of meteorological quantities can reach $\tau \leq 12$ hours;
- in contrast to classical numerical methods for solving equations of mathematical physics the offered method does not suffer from stability problems;
- the accuracy of the solution $\Re_{ijk}(t)$ depends on the accuracy of the applied interpolation method for filling up smoothly the given discrete function in the nodes of the regional grid and the method of approximation applied for the differential operator D in (19) by grid operator Λ.

4. Smooth filling up and approximation of differential operators by grid ones

Denote by η one of the coordinate axes in $r = \left(\lambda, \varphi, \sigma\right)$ and assume that the linear size of the domain of the macroscale model is in the interval $a \leq \eta \leq b$ along this coordinate axis. Let the arbitrary points $a < \eta_1 < \eta_2 <$ $< \ldots < \eta_{N-1} < b$ form a non-uniform macroscale grid $\varpi_h\,[a,b]$ with a grid step $h_{i-1} = \eta_i - \eta_{i-1}$. Let us renumber all nodes in some order $\eta_0, \eta_1, \eta_2, \ldots$ $\ldots, \eta_N$ and consider the values of the macroscale function $\Re\left(\eta_i, t^m\right)$ in the nodes of this grid as components of a vector $\Re = \left\{\Re_i\left(t^m\right),\ \ i = 0, 1, \ldots, N\right\}$.

Consider a cubic polynomial $P_{i-1}(\eta) = a_0 + a_1\eta + a_2\eta^2 + a_3\eta^3$ at the interval $\left[\eta_i, \eta_{i+1}\right]$ and choose it in such a way that in points $\eta_{i-1},\ \eta_i$ and η_{i+1} it has the same values as function $\Re$. Obviously, the polynomial P_{i-1} can be represented as a sum of quadratic polynomials which in points $\eta_{i-1},\ \eta_i$ and η_{i+1} satisfy the conditions

$$(21)\ P_{i-1}(\eta)\big|_{\eta=\eta_{i-1}} = \Re_{i-1}, \quad P_{i-1}(\eta)\big|_{\eta=\eta_i} = \Re_i, \quad P_{i-1}(\eta)\big|_{\eta=\eta_{i+1i}} = \Re_{i+1},$$

and a cubic polynomial which has zero values in these points. So it is easy to check that such a presentation is of the form:

$$(22a) P_{i-1}(\eta) = a_0^i + a_1^i(\eta-\eta_i) + a_2^i(\eta-\eta_i)^2 + a_3^i\left(\eta-\eta_{i-1}\right)(\eta-\eta_i)(\eta-\eta_{i+1}),$$

where

$$a_0^i = \Re_i, \quad a_1^i = \frac{1}{h_{i-1}+h_i}\left(\frac{\Re_{i+1}-\Re_i}{h_i}h_{i-1} + \frac{\Re_i-\Re_{i-1}}{h_{i-1}}h_i\right),$$

(22b) $$a_2^i = \frac{1}{h_{i-1}+h_i}\left(\frac{\Re_{i+1}-\Re_i}{h_i} - \frac{\Re_i - \Re_{i-1}}{h_{i-1}}\right).$$

Obviously it is possible to build on the interval $[\eta_i, \eta_{i+1}]$ a similar qubic polynomial

(23a) $$P_i(\eta) = b_0^i + b_1^i(\eta-\eta_{i+1}) + b_2^i(\eta-\eta_{i+1})^2 + b_3^i(\eta-\eta_i)(\eta-\eta_{i+1})(\eta-\eta_{i+2})$$

with the following coefficients:

$$b_0^i = \Re_{i+1}, \qquad b_1^i = \frac{1}{h_i+h_{i+1}}\left(\frac{\Re_{i+2}-\Re_{i+1}}{h_{i+1}}h_i + \frac{\Re_{i+1}-\Re_i}{h_i}h_{i+1}\right),$$

(23b) $$b_2^i = \frac{1}{h_i+h_{i+1}}\left(\frac{\Re_{i+2}-\Re_{i+1}}{h_{i+1}} - \frac{\Re_{i+1}-\Re_i}{h_i}\right),$$

Then in points η_i, η_{i+1} and P_i has η_{i+2} the same values as $\Re$.

In each of the polynomials P_{i-1}, P_i, parameters a_3^i and b_3^i are still arbitrary. Let us construct interpolation function $\Im(\eta)$ of the class of functions C^p $(p = 3)$. We will require that the interpolation function $\Im_i(\eta)$ in area $[\eta_{i-1}, \eta_{i+2}]$ satisfies conditions

(24) $$\left.\frac{d^k P_{i-1}}{d\eta^k}\right|_{\eta=\eta_i} = \left.\frac{d^k \Im_i}{d\eta^k}\right|_{\eta=\eta_i}, \qquad \left.\frac{d^k P_i}{d\eta^k}\right|_{\eta=\eta_{i+1}} = \left.\frac{d^k \Im_i}{d\eta^k}\right|_{\eta=\eta_{i+1}}, \qquad k = 0, 1, 2, 3.$$

Such interpolation functions $\Im_i(\eta)$ exist, they can be expressed as polynomials of degree $2p - 1$:

(25a) $$\Im_i(\eta) = c_0 + c_1(\eta-\eta_i) + c_2(\eta-\eta_i)^2 + c_3(\eta-\eta_i)^3 + c_4(\eta-\eta_i)^4 + c_5(\eta-\eta_i)^5,$$

and are uniquely defined by conditions (24). In the case of (25a) the coefficients can be given as

$$c_5^i = -\frac{1}{3h_i^5}\left[2\left(b_0^i - a_0^i\right) - h_i\left(b_1^i + a_1^i\right) + h_i^2\left(b_2^i - a_2^i\right)\right],$$

$$c_4^i = \frac{1}{3h_i^4}\left[5\left(b_0^i - a_0^i\right) - h_i\left(b_1^i + 4a_1^i\right) + h_i^2\left(b_2^i - 4a_2^i\right)\right],$$

(25b) $$c_3^i = \frac{1}{h_{i-1}+h_i+h_{i+1}}\left[\left(b_2^i - a_2^i\right) - 2h_i\left(h_i + 2h_{i+1}\right)c_4^i - 10h_i^2 h_{i+1} c_5^i\right],$$

$$c_2^i = a_2^i - \left(h_i - h_{i-1}\right)c_3^i, \qquad c_1^i = a_1^i - h_{i-1}h_i c_3^i, \qquad c_0^i = a_0^i.$$

So constructed, the polynomials $\Im_i(\eta)$ provide interpolation for a sufficiently wide class of functions inside the interval $[a,b]$ with third order of smoothness in points η_i, η_{i+1}, where the adjacent intervals $\left[\eta_{i-1}, \eta_{i+1}\right]$ and $\left[\eta_i, \eta_{i+2}\right]$ coincide.

Computation of grid values of partial derivatives of the first order $\psi_i = = \left(\partial \Re / \partial \eta\right)_i$ and partial derivative of the second order $\zeta_i = \left(\partial^2 \Re \middle/ \partial \eta^2\right)_i$ participating in $f^m_{jkl} = \Lambda \Re^m_{jkl}$, we shall carry out on the base of relations

$$\begin{gathered}
\psi_{i+1} + 2\left(1 + \frac{h_i}{h_{i-1}}\right)\psi_i + \frac{h_i}{h_{i-1}}\psi_{i-1} = \\
= \frac{3}{h_i}\left\{\Re_{i+1} - \left[1 - \left(\frac{h_i}{h_{i-1}}\right)^2\right]\Re_i - \left(\frac{h_i}{h_{i-1}}\right)^2 \Re_{i-1}\right\} - \\
- \frac{h_i h_{i-1}^2}{24}\left[1 - \left(\frac{h_i}{h_{i-1}}\right)^2\right]\frac{\partial^4 \Re}{\partial \eta^4},
\end{gathered} \tag{26}$$

$$\begin{gathered}
\frac{h_{i-1}}{h_i}\left[\frac{h_{i-1}}{h_i}\left(1 - \frac{h_{i-1}}{h_i}\right) + 1\right]\xi_{i+1} + \\
+ \left(1 + \frac{h_{i-1}}{h_i}\right)\left[\frac{h_{i-1}}{h_i}\left(3 + \frac{h_{i-1}}{h_i}\right) + 1\right]\xi_i + \\
+ \left[\frac{h_{i-1}}{h_i}\left(1 + \frac{h_{i-1}}{h_i}\right) - 1\right]\xi_{i-1} = \\
= \frac{12}{h_i^2}\left[\frac{h_{i-1}}{h_i}\Re_{i+1} - \left(1 + \frac{h_{i-1}}{h_i}\right)\Re_i + \Re_{i-1}\right] + \\
+ \frac{h_i^2 h_{i-1}}{360}\left[1 - \left(\frac{h_{i-1}}{h_i}\right)^2\right]\left\{5\frac{h_{i-1}}{h_i} + 2\left[1 - \left(\frac{h_{i-1}}{h_i}\right)^2\right]\right\}\frac{\partial^5 \Re}{\partial \eta^5},
\end{gathered} \tag{27}$$

which are obtained by decomposing function $\Re(r,t)$ in a Tailor series for the nodes $\eta = \eta_{i-1}$ and $\eta = \eta_{i+1}$ around $\eta = \eta_i$, in view of the equalities $\psi_i = \left(\partial \Re / \partial \eta\right)_i$ and $\zeta_i = \left(\partial^2 \Re \middle/ \partial \eta^2\right)_i$.

It is obvious that the relations (26), (27) have third order at $h_i \neq h_{i-1}$ and fourth order at $h_i = h_{i-1}$. Derivatives $\psi_i = \left(\partial \Re / \partial \eta\right)_i$ and $\xi_i = \left(\partial^2 \Re \middle/ \partial \eta^2\right)_i$ enter in (26), (27) implicitly. But as (26), (27) represent systems of algebraic equations with three-diagonal matrixes, their solution can be carried out rather effectively with the help of the sweep method [10] with boundary conditions:

$$-\frac{h_1}{6}\left(\xi_2-\xi_1\right)+\psi_1+\psi_2=2\frac{\Re_2-\Re_1}{h_1}, \tag{28a}$$

$$-\frac{h_{N-1}}{6}\left(\xi_N-\xi_{N-1}\right)+\psi_{N-1}+\psi_N=2\frac{\Re_N-\Re_{N-1}}{h_{N-1}}. \tag{28b}$$

Here it is necessary to note the main advantage of the offered method for approximating the derivatives in the (7)–(18). As the solution of the system of algebraic equations (26), (27) in each point i depends on values in other points, it depends on $\Re_i$ globally rather than locally. In other words, computed derivatives $\psi_i = \left(\partial\Re/\partial\eta\right)_i$ and $\xi_i = \left(\partial^2\Re\middle/\partial\eta^2\right)_i$ on interval $[a,b]$ will be more smoother, as they are not subjected to computational disturbances, inherent in the local three-point operators, approximating the derivatives.

5. Providing initial conditions for an ecological problem

Restricted resources of modern computers do not allow us to use spatial discretization of a general ecological-meteorological problem with sufficient resolution in order to compute processes of dispersion of a pollutant in the immediate proximity of single emission sources without involving a class of processes of “sub-grid” scale. On the other hand, use of a non-uniform grid with higher resolution in the vicinity of non-uniformly distributed sources is algorithmically difficult to solve.

As meteorological observations show, in the spectrum of pulsations of meteorological quantities in the real atmosphere there is a deep minimum in the area of perturbations with a period of $t_S \approx 60$ min. Therefore meteorological quantities obtained by measurement tools during 4-dimensional analysis are exposed to an hour averaging. The results of such averaging are equivalent to the data brought to synoptic maps. Hence, near the emission sources it is possible to neglect the dependence of meteorological quantities $v_1 = u$, $v_2 = v$, $v_3 = w$, and K_H, K_G on spatial coordinates. That is, the simplest and most accessible approaches of approximation in theoretical research: the assumptions of uniformity and isotropy of turbulent motions can be used.

Taking into account the small geometrical size of the area of abundant concentration near some source, we restrict ourselves in this area to the rectangular Cartesian system of coordinates $X = (x_1, x_2, x_3)$. Then the equation

(11) describing the transport and diffusion of substance $q\,(x_1,x_2,x_3,t)$ will take the form

$$(29)\frac{\partial q}{\partial t}+v_1\frac{\partial q}{\partial x_1}+v_2\frac{\partial q}{\partial x_2}+v_3\frac{\partial q}{\partial x_3}=\frac{\partial}{\partial x_1}\left(a\frac{\partial q}{\partial x_1}\right)+\frac{\partial}{\partial x_2}\left(a\frac{\partial q}{\partial x_2}\right)+\frac{\partial}{\partial x_3}\left(a\frac{\partial q}{\partial x_3}\right).$$

Equation (29) describes migration of concentration of aerosol substance q together with a flow of air with a speed $V=(v_1,v_2,v_3)$ and its diffusion by isotropic turbulence with turbulent diffusion coefficient $a=\mathrm{K}_G=\mathrm{K}_G=\mathrm{K}_H$. The fields v_j $(j=1,2,3)$ are considered homogeneous and defined on the base of a macroscale model at time $t=t^m\,(m=0,1,\ldots,M)$ with a step $\tau=t^{m+1}-t^m$.

Equation (29) with constant factors by the substitutions

$$(30)\qquad \eta_j=\frac{v_j}{2a},\quad \varpi_j=-\frac{v_j^2}{4a},\quad q=\varphi\prod_{j=1}^{3}e^{\varpi_j t+\eta_j x_j},\quad (j=1,2,3)$$

can be reduced to the relation

$$(31)\qquad \frac{\partial\varphi}{\partial t}=a\sum_{j=1}^{3}\frac{\partial^2\varphi}{\partial x_j^2}.$$

It is known [11] that the function φ, determined by the conditions

$$\left.\begin{aligned}\frac{\partial^2\varphi}{\partial r^2}&=\frac{1}{a}\frac{\partial\varphi}{\partial t},\\ \varphi\,(0,t)&=\frac{Q}{4\pi a}=\varphi_0,\\ \varphi\,(r,0)&=0,\end{aligned}\right\}$$

is expressed by the formula

$$\varphi\,(r,t)=\frac{2}{\sqrt{\pi}}\frac{M}{4\pi a}\int_{\frac{r}{2\sqrt{at}}}^{\infty}e^{-\alpha^2}d\alpha,$$

where

$$\alpha=\frac{\zeta-r}{2\sqrt{a\,(t-t_0)}}.$$

Hence, the solution of equation (31), describing the distribution of a pollutant emitted by a continuously working source of capacity Q, placed at the origin ($r = 0$), has the form

$$\varphi\,(r,t) = Q\Phi\,(r,t) = \frac{1}{r}\frac{2}{\sqrt{\pi}}\frac{Q}{4\pi a}\int\limits_{\frac{r}{2\sqrt{at}}}^{\infty} e^{-\alpha^2}d\alpha,$$

where $\Phi\,(r,t)$ is the concentration in case of an individual source ($Q = 1$).

To proceed to the case of an instant source, we will consider a source of capacity Q placed at point $\left(x_1^{ef}, x_2^{ef}, x_3^{ef}\right)$ and continuously working during a time interval of length τ. Such a source is equivalent to two sources of capacity $+Q$ and $-Q$, first of them is turned on at $t = 0$, and the second one at $t = \tau$. The distribution of concentration is thus expressed by the formula

$$\varphi_\tau\,(r,t) = Q\,[\Phi\,(r,t) - \Phi\,(r,t-\tau)]\,.$$

For a time interval of length τ, the amount of pollution $M = Q\tau$ is emitted, therefore

$$\varphi_\tau\,(r,t) = \frac{M}{\tau}\,[\Phi\,(r,t) - \Phi\,(r,t-\tau)]\,.$$

Passing to a limit at $\tau \to 0$ and considering M as constant, we will find that

$$\varphi_0\,(r,t) = \lim_{\tau\to 0}\varphi_\tau\,(r,t) = M\frac{\partial\Phi}{\partial t} = \frac{2}{\sqrt{\pi}}\frac{M}{4\pi ar}\frac{ra}{4a\sqrt{at^3}}e^{-\frac{r^2}{4at}},$$

or

$$\varphi_0\,(r,t) = MG\left(x_1, x_2, x_3, t, x_1^{ef}, x_2^{ef}, x_3^{ef}\right),$$

where

(32) $$G\left(x_1, x_2, x_3, t, x_1^{ef}, x_2^{ef}, x_3^{ef}\right) =$$

$$= \left(\frac{1}{2\sqrt{\pi at}}\right)^3 e^{-\frac{\left(x_1-x_1^{ef}\right)^2+\left(x_2-x_2^{ef}\right)^2+\left(x_3-x_3^{ef}\right)^2}{4at}}$$

represents concentration at a point (x_1, x_2, x_3) at time t as an effect of a point source of capacity M, placed at time $t = 0$ at the point $\left(x_1^{ef}, x_2^{ef}, x_3^{ef}\right)$.

If we carry out the obtained expression, returning back to the initial variable by the rule (30), then as a result we will have:

$$q\,(x_1, x_2, x_3, t) = M\left(\frac{1}{2\sqrt{\pi\, a\,(t - t_0)}}\right)^3 \cdot \tag{33}$$

$$\cdot\, e^{-\frac{\left[x_1 - x_1^{ef} - v_1(t - t_0)\right]^2 + \left[x_2 - x_2^{ef} - v_2(t - t_0)\right] + \left[x_3 - x_3^{ef} - v_3(t - t_0)\right]^2}{4a(t - t_0)}}.$$

Special estimations have been carried out in [8], which have shown that experimental data appear much below those computed on the base of the solution (33) for one hour $t_S \approx 1$ at constant actual meteorological quantities $v_j \quad (j = 1, 2, 3)$. This is explained by the presence of real large-scale fluctuations in the wind velocity, which contribute to an additional dispersion of the pollutants as well as the orientation of the axis of the plume, formed by the spreading substance. Such a result can be treated as a “mesoscale effect” of averaging the concentrations. Its estimation should be connected in view of features of fluctuations in the wind direction caused, in particular, by action of atmospheric whirlwinds of various scales. The change of concentration, depending on the time of averaging to which it concerns, was discussed in a number of works (for example in [12]) mainly on the base of qualitative reasons. To eliminate the differences between computed and measured data, arising as a result of this dependence, certain empirical factors are quite often used to decrease values of the calculated concentration by several times.

To receive the average concentration of a pollutant at any point $P\,(x, y, z)$ for an interval of time $t_S = 1$ hour, suppose that fluctuations of the wind in a given direction for a considered time interval occur randomly, and their probability is described by the Gaussian distribution function

$$f\,(\alpha) = \frac{1}{\sigma_\alpha\sqrt{2\pi}}\exp\left[-\frac{(\Delta\alpha)^2}{2\sigma_\alpha^2}\right]. \tag{34}$$

Here $\Delta\alpha$ is the deviation of wind direction from the mean value over an interval of length t_S; σ_α is the standard deviation of wind direction, determined as a result of processing input array of wind data under the formula

$$\sigma_\alpha^2 = \frac{1}{N - 1}\sum_{i=1}^{N}(\Delta\alpha_i)^2, \tag{35}$$

where N is the volume of the sample.

The data processing of measurements has allowed us to establish that σ_{α}^{2} depends on the external interval of averaging for cases of an unstable and a stable atmosphere. From the obtained results it follows that at the period of averaging $t_S \approx 1$ hour the value of σ_{α} grows by 26 % for unstable stratification and 41 % for stable one.

According to the probability theory, the average value of concentration for the period t_S is defined by the expression

$$q_t = \int\limits_{\bar{\alpha}-\frac{\pi}{2}}^{\bar{\alpha}+\frac{\pi}{2}} q_S(\alpha) f(\alpha)\, d\alpha, \tag{36}$$

where $q_S(\alpha)$ is a solution of the diffusion problem (33) for direction α.

By substituting (34) into (36) and by integrating, we will receive:

$$q_t = \frac{q_S(x - x_0, y - y_0, z - H)}{2\pi\,(x - x_0)\left(\bar{\sigma}^2 - \sigma_{\alpha}^2\right)} \exp\left(-\frac{\alpha_P^2}{2\sigma_{\alpha}^2}\right). \tag{37}$$

Here α_P is the angle between the average direction of wind $\bar{\alpha}$ and that at a considered point $P(x, y, z)$.

The concept of conditional division of area of dispersion of a pollutant emitted by a high point source in ”a near zone” and ”a distant zone” allows us to set concentration fields from the non-uniformly distributed sources at time $t = t^{m+1}$ $(m = 0, 1, \dots, M)$ with a step $\tau = t^{m+1} - t^m$ and to solve a task of the forecast of pollution under real weather conditions with the help of (20).

6. Conclusion

We have presented a new mathematical model and a non-standard numerical method for the effective solution of the complex problem of analysing and forecasting both meteorological quantities and distribution of atmospheric pollution over a region. The method offered replaces the Cauchy problem in the atmospheric model by a boundary-value problem and introduces a specific interpolation technique that have a number of advantages in the model and the method is computationally efficient. Firstly, advance time of a forecast and a time step of giving macroscale information in view of a daily course of meteorological quantities can be significantly increased and reach up to

12 hours. Secondly, in contrast to classical numerical methods for solving equations of mathematical physics, the offered method does not suffer from stability problems. And thirdly, the accuracy of the simulation depends on the accuracy of the interpolation method applied for filling up smoothly the given discrete function in nodes of the regional grid and the method for approximating the differential operator D in (20) by a grid operator Λ.

In overall, the approach undertaken in our method promises good computational efficiency. A work to implement the computational scheme for solving problems in meteorological and ecological forecasting of regions of Ukraine is in progress and results of its application will be available soon.

References

[1] J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell and C. A. Johnson, eds.: *Climate Change 2001: The Scientific Basis*, Cambridge University Press, 2001.

[2] Z. Zlatev: *Computer treatment of large air pollution models*, Environmental Science and Technology Library, Vol. 2., KLUWER Academic Publishers, Dordrecht–Boston–London, 1995.

[3] A. Doroshenko: Mathematical Models and Methods for High-Performance Parallel Computation, Kiev, *Naukova dumka* 2000, 177 p. (in Russian).

[4] Dovgyi S.A., Prusov V.A., Kopeika O.B.: Mathematical modelling man-caused pollution of environment, *Naukova Dumka* 2000, Kiev, 247 p. (in Russian).

[5] Charney J., Eliassen A.: A numerical method for predicting the perturbations of the middle latitude westerlies, *Tellus* 1949, 1, 38, pp. 1268–1274.

[6] W. Hundsdorfer, J. G. Verwer: *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer Series in Computational Mathematics, Vol. 33, Springer Verlag, 2003.

[7] Havasi, Á., Bartholy, J., Faragó, I.: Splitting method and its application in air pollution modelling, *Időjárás* **105** (2001), pp. 39–58.

[8] Dimov, I., Faragó, I., Havasi, Á., Zlatev, Z.: L-commutativity of the operators in splitting methods for air pollution models, *Annal. Univ. Sci. Sec. Math.* **44** (2002), pp. 127–148.

[9] Phillips N., Shukla J.: On the strategy of combining coarse and fine grid meshes in numerical weather prediction, *J. Appl. Meteor.* **12** (1973), pp. 763–770.

[10] Godunov S. K., Riabenkiy V. C.: Difference schemes (introduction to the theory), *Science*, Moscow (1973), 400 p (in Russian).

[11] TIKHONOV A. N., SAMARSKIY A. A.: The equations of mathematical physics, *Science*, Moscow (1977), 735 c. (in Russian).

[12] OKE T. R.: *Boundary layer climates*, London, Methuen & Co. Ltd., 1987, 359 p.

Vitaliy Prusov
Ukrainian Hydrometeorological Research Institute
Science prosp. 37
03650, Kiev-28
Ukraine
dor@isofts.kiev.ua

Anatoliy Doroshenko
Institute of Software Systems
National Academy of Sciences of Ukraine
Acad. Glushkov prosp., 40, block 5
03187 Kiev
Ukraine
dor@isofts.kiev.ua

ANNALES UNIV. SCI. BUDAPEST., **46** (2003), 75–80

PONZAG THEOREMS

By

ANDRÁS HRASKÓ

(Received November 29, 2003)

Introduction

The circumscribed circle k of any triangle ABC and the Feuerbach circle l circumscribed around the midpoints of its sides satisfy a simple relation. If the radii of these circles are r, R respectively, then

$$2R = r. \tag{1}$$

This theorem can be reversed to a less known statement: if (1) holds and the midpoint of the chord AB of k is on l, then AB can be completed to form a triangle ABC with circumscribed circle k and Feuerbach circle l.

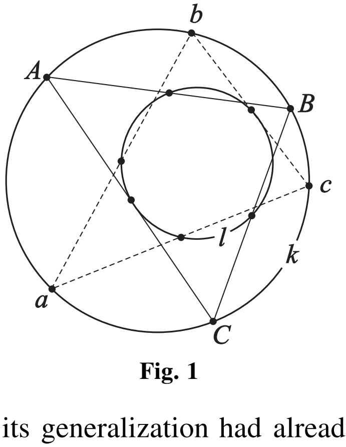


Fig. 1

This statement and its generalization had already been discussed in [2] where its equivalence to the theorem of Poncelet and to Zig-zag theorem was also proved. The generalization mentioned above can be shortly formulated as follows:

PONZAG THEOREM (SHORT VERSION)

If there is an n-gon whose vertices lie on the circle k and the midpoints of the sides lie on another circle l, then any chord of k with midpoint on l can be completed to form such an n-gon.

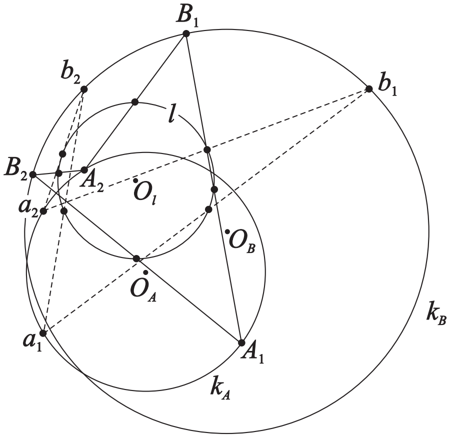


Fig. 2

As an exercise, the present author proved Ponzag theorem by the correspondence principle of Chasles (see [1]). The proof helped him to arrive at general statements of the same kind.

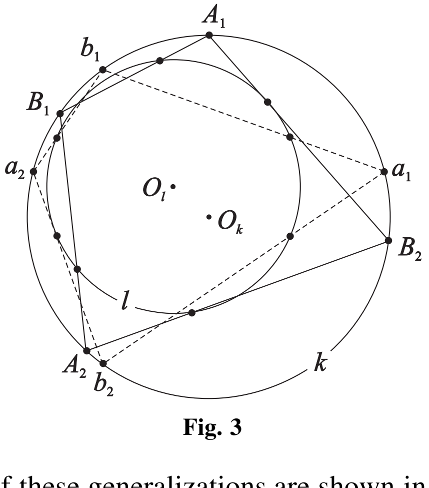


Fig. 3

Two examples of these generalizations are shown in advance in figures 2 and 3. In Fig. 2 circles k_A, k_B and l are on the plane such that every segment AB with endpoints A, B and midpoint F incident to k_A, k_B, l respectively can be completed to form a 4-gon with vertices alternately on k_A, k_B and midpoints all on l. $A_1B_1A_2B_2$, $a_1b_1a_2b_2$ are such 4-gons on the figure.

To understand the example in Fig. 3 a point H on the segment AB will be called ”thirdpoint” of AB if $AH = 2 \cdot HB$ or $BH = 2 \cdot HA$. The vertices of the 4-gons $A_1B_1A_2B_2$, $a_1b_1a_2b_2$ are on the same circle k, while certain thirdpoints of the sides are on l. Every chord AB of k with a thirdpoint on l can be completed to form such a 4-gon.

A ponzag-type general theorem

We will consider

A) three arbitrary circles k_A, k_B, l in the plane, or

B) two arbitrary circles k_A, k_B and any sphere l in 3-space,

and any

A) orientation preserving similarity $\mathscr{A}$ of the plane, different from translations[1],

B) dilatation $\mathscr{A}$ of the space.

This transformation will be denoted by $\mathscr{A}_P$ if its centre is translated to P, i.e.

$$\mathscr{A}_P(Q) = \mathscr{A}(Q - P) + P.$$

In cases A), B) there is always a similarity $\mathscr{B}$ of the plane or a dilatation of the space such that $\mathscr{A}_A(Q) = B$ holds iff $\mathscr{B}_B(Q) = A$. Indeed, if $\mathscr{A}$ is determined by the multiplication of some real or complex number λ then $\mathscr{B}$ can be defined as the multiplication by $\frac{\lambda}{\lambda - 1}$. Later we will also use the similarity $\mathscr{C}$ of the plane or the dilatation of the space such that $\mathscr{C}_Q(B) = A$ holds iff $\mathscr{B}_B(Q) = A$. The rate of enlargement of $\mathscr{C}$ is $|\frac{1}{\lambda - 1}|$.

If we exclude the existence of points $A \in k_A$ with $k_B \subset \mathscr{A}_A(l)$ and $B \in k_B$ with $k_A \subset \mathscr{B}_B(l)$, then Ponzag process constructing the sequence of points A_i, B_i in the theorem below will work unambiguosly. Apart from these conditions the following general theorem holds:

[1] such a transformation has a unique fixed point

General Ponzag theorem

Starting from any pair $A_1 \in k_A$, $B_1 \in k_B$ of points with

$$B_1 \in \mathcal{A}_{A_1}(l) \tag{2}$$

the sequence of points

$$A_1, \quad B_1, \quad A_2, \quad B_2, \quad A_3, \quad B_3, \quad \dots \tag{3}$$

($A_i \in k_A$, $B_i \in k_B$) can be uniquely determined by the following conditions:

I) $A_{i+1} \in \mathcal{B}_{B_i}(l)$, $\qquad B_i \in \mathcal{A}_{A_i}(l)$;

II) A_{i+1} differs from A_i, B_{i+1} differs from B_i if it is possible.

If sequence (3) *is $2n$-step periodic i.e. $A_{n+1} = A_1$ and $B_{n+1} = B_1$, then it is $2n$-step periodic starting from any pair of points $A_1 \in k_A$, $B_1 \in k_B$ satisfying condition* (2).

On the example in Fig. 2 $\mathcal{A}$ is the dilatation of ratio 2. In Fig. 3 $k_A = k_B = k$ and $\mathcal{A}$ is the dilatation of ratio 3; in Fig. 3 $\mathcal{A}$ is the composition of rotation 45° and dilatation of ratio $\sqrt{2}$ in Fig. 4. In this last figure there are infinitely many (twisted) quadrilaterals inscribed in k such that the vertices of the isosceles right angled triangles based on the sides of the quadrilateral are all incident to circle l. Traversing along the sides of the quadrilateral these isosceles triangles are oriented alternately.

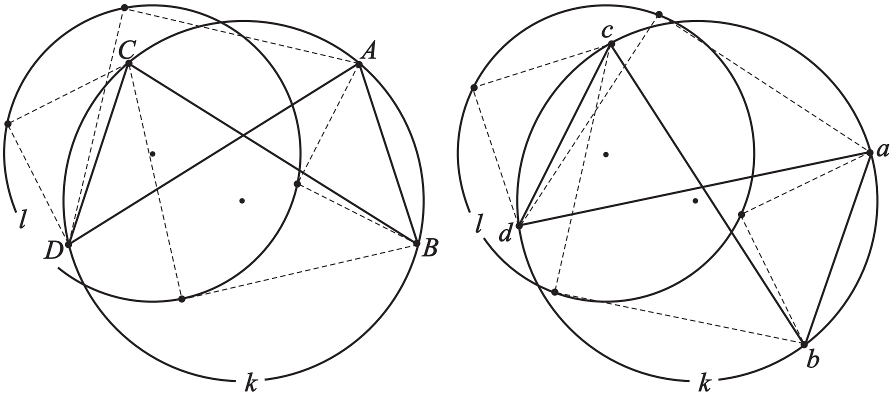


Fig. 4

Proof. We do not follow the way of the discovery of the theorem but trace it back to Zig-zag theorem [2] (refreshed below).

In Zig-zag theorem two arbitrary circles k_A, k_B of 3-space and a certain distance ρ is considered such that none of the spheres of radius ρ centered at k_A (k_B) contains k_B, (k_A).

ZIG-ZAG THEOREM

Starting from any pair $A_1 \in k_A$, $B_1 \in k_B$ *of points with*

$$A_1B_1 = \rho \tag{4}$$

the sequence of points

$$A_1, \quad B_1, \quad A_2, \quad B_2, \quad A_3, \quad B_3, \quad \dots \tag{5}$$

($A_i \in k_A$, $B_i \in k_B$*) can be uniquely determined by the following conditions:*

I) $B_iA_{i+1} = \rho, \qquad A_iB_i = \rho;$

II) A_{i+1} *differs from* A_i, B_{i+1} *differs from* B_i *if it is possible.*

If sequence (5) *is* $2n$*-step periodic i.e.* $A_{n+1} = A_1$ *and* $B_{n+1} = B_1$*, then it is* $2n$*-step periodic starting from any pair of points* $A_1 \in k_A$, $B_1 \in k_B$ *satisfying condition* (4).

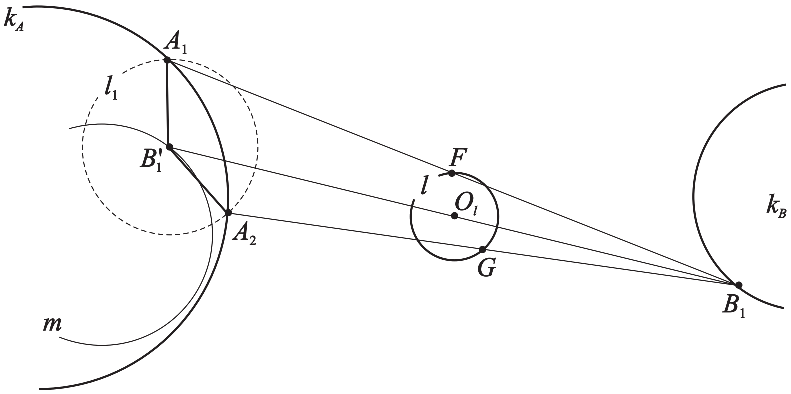


Fig. 5

In Fig. 5 a general ponzag configuration (k_A, k_B, l) can be seen. The transformation $\mathscr{A}$ is the enlargement of ratio 2, but this speciality of the figure will not be used in the following argument. The objects of the figure are defined in the following order:

$$B_1 \in k_B,$$

$$\mathscr{B}_{B_1}(l) = l_1, \qquad \mathscr{B}_{B_1}(O_l) = B_1',$$

$$\{A_1, A_2\} = k_A \cap l_1, \qquad \mathscr{B}_{B_1}(F) = A_1, \qquad \mathscr{B}_{B_1}(G) = A_2,$$

$$m = \{\mathscr{B}_B(O_l) \mid B \in K_B\},$$

i.e. $A_1B_1A_2$ is a piece of a general ponzag series. m is a circle, because $m = \mathscr{C}_{O_l}(k_B)$ and

$$|B_1'A_1| = |\mathscr{B}_{B_1}(O_lF)| = \frac{\lambda R}{\lambda - 1} = |\mathscr{B}_{B_1}(O_lG)| = |B_1'A_2|,$$

i.e. ponzag configuration (k_A, k_B, l) and zig-zag configuration $(k_A, m, \rho = = \frac{\lambda R}{\lambda - 1})$ are analogous: $A_1B_1A_2$ is a piece of a ponzag series iff $A_1\mathscr{C}_{O_l}(B_1)A_2$ is a piece of a zig-zag series.

COMMENTS.

1. Ponzag theorem is not included in the General Ponzag theorem, the latter for example does not say anything about the triangle and its midpoints.

2. The Feuerbach circle of an n-gon has a property analogous to that of the triangle. Namely if the vertices of an n-gon are all lying on circle k centered at O_k and the midpoints of the sides are lying on circle l centered at O_l and O_l is the midpoint of MO_k then the pedal points of the perpendiculars from M to the sides of the n-gon are also lying on l.

3. The following elementary problem is a challenge for the reader: *The radii of circles k_A, k_B, l on Fig. 2 are denoted by r_A, r_B, R respectively; the distance of the center O_l of l from the midpoint of O_AO_B the central of circles k_A, k_B is d. Find the algebraic condition satisfied by these four values.*

References

[1] J. DIEUDONNÉ, Historical development of algebraic geometry, *American Mathematical Monthly,* **8**, (1972) 827–866. Also available in book; the History of algebraic geometry had been published by Wadsworth Advanced Books and Software.

[2] ANDRÁS HRASKÓ, Poncelet-type problems, an elementary approach, *Elemente der Mathematik,* **55**, (2000) 45–62.

András Hraskó

Mihály Fazekas Secondary School
Horváth Mihály tér 8.
H-1082 Budapest
Hungary
andras@hrasko.com

THE CONJUGATE GRADIENT METHOD FOR 4TH ORDER NONLINEAR ELLIPTIC PROBLEMS

By

A. MÁRKUS

(Received Decemer 30, 2003)

1. Introduction

The conjugate gradient method (CGM) was first presented for linear systems in [8]. Later the method was extended to Hilbert space in the same form [2, 7] and also for nonlinear problems in Hilbert space [2, 3, 4].

In this paper we use the Hilbert space version of the CGM to develop a numerical algorithm for solving a class of 4th order elliptic problems. Such problems arise e.g. in the elasto-plastic bending of plates [12, 14]. In this paper we give the theoretical background and construct the method. We show that the convergence is linear.

In this paper the CG iteration is executed in FEM subspaces using Sobolev space background. The iteration involves a preconditioning matrix obtained as the discretized biharmonic operator. The idea of preconditioning operators is summarized in [6]; biharmonic preconditioning operators in other iterative methods arise in [10, 11]. As is well known, fast solvers have already been developed for the biharmonic problem [1, 5]: the use of such solvers make our algorithm an efficient method. Hence we do not present any concrete implementation or numerical result. Owing to the operator preconditioning, the convergence factor of our method is mesh independent.

2. The problem

In this section we formulate the 4th order boundary value problem and state that it has a unique weak solution. For this we introduce some notations:

$$H : V := \sum_{i,j=1}^{N} H_{ij} V_{ij} \qquad (H, V \in \mathbf{R}^{N\times N}),$$

$$\mathrm{div}^2 G := \sum_{i,j=1}^{N} \partial_i \partial_j G_{ij} \qquad (G \in C^2(\Omega, \mathbf{R}^{N\times N})).$$

We underline that in this paper

$$\nabla^2 v = \left(\partial_i \partial_j v\right)_{i,j=1}^{N}$$

denotes the Hessian of $v \in C^2(\Omega)$.

Now we consider 4th order nonlinear Dirichlet problems of the form

$$\left\{ \begin{aligned} T(u) \equiv \ & \mathrm{div}^2 A(x, \nabla^2 u) = g(x) \\ & u_{|\partial\Omega} = \frac{\partial u}{\partial \nu}_{|\partial\Omega} = 0, \end{aligned} \right. \tag{1}$$

satisfying the following conditions:

(i) $\Omega \subset \mathbf{R}^N$ is a bounded domain with piecewise smooth boundary.

(ii) The matrix-valued function $A : \Omega \times \mathbf{R}^{N\times N} \to \mathbf{R}^{N\times N}$ is measurable and bounded w.r. to the variable $x \in \Omega$ and C^2 in the other variables.

(iii) The Jacobian arrays

$$\frac{\partial A(x,\Theta)}{\partial \Theta} = \left\{ \frac{\partial A_{rs}(x,\Theta)}{\partial \Theta_{ik}} \right\}_{i,k,r,s=1}^{N} \in \mathbf{R}^{(N\times N)^2}$$

are symmetric (i.e., $\partial A_{rs}/\partial\Theta_{ik} = \partial A_{ik}/\partial\Theta_{rs}$) and their eigenvalues Λ satisfy

$$0 < m \leq \Lambda \leq M < \infty \tag{2}$$

with constants $M \geq m > 0$ independent of (x, Θ). Reformulating this assumption, the operators represented by the arrays are self-adjoint w.r.t. inner product

$$\langle H, K \rangle := H : K \qquad (H, K \in \mathbf{R}^{N\times N}) \tag{3}$$

and we have

$$m\,|H|_{\mathsf{F}}^2 \leq \frac{\partial A}{\partial \Theta}(x,\Theta)(H,H) \leq M\,|H|_{\mathsf{F}}^2 \quad (H \in \mathbf{R}^{N\times N}), \tag{4}$$

where

$$\frac{\partial A}{\partial \Theta}(x,\Theta)(H,H) := \left\langle \frac{\partial A}{\partial \Theta}(x,\Theta)H, H \right\rangle ,$$

and

$$|H|_{\mathsf{F}} := \langle H, H\rangle^{\frac{1}{2}} \tag{5}$$

is the Frobenius norm of the matrix H.

(iv) $g \in L^2(\Omega)$.

(v) The second derivatives satisfy

$$\left\| \frac{\partial^2 A}{\partial \Theta^2}(x,\Theta) \right\|_{\mathsf{F}} \le K \qquad ((x,\Theta) \in \Omega \times \mathbf{R}^{N\times N}))$$

with some $K > 0$ independent of (x,Θ), where

$$\left\| \frac{\partial^2 A}{\partial \Theta^2}(x,\Theta) \right\|_{\mathsf{F}} := \sup_{|H|_{\mathsf{F}}=|K|_{\mathsf{F}}=|L|_{\mathsf{F}}=1} \left| \frac{\partial^2 A}{\partial \Theta^2}(x,\Theta)(H,K,L) \right| .$$

PROPOSITION 1 ([6, p. 150]). *If $(i),\dots,(iv)$ hold then problem (1) has a unique weak solution $u^* \in H_0^2(\Omega)$. That is, u^* satisfies*

$$\int_\Omega A(x,\bigtriangledown^2 u^*) : \bigtriangledown^2 v = \int_\Omega g v \qquad (v \in H_0^2(\Omega)).$$

REMARK 1. One can also consider problem (1) with the boundary conditions

$$u_{|\partial\Omega} = A(x,\bigtriangledown^2 u)\nu \cdot \nu|_{\partial\Omega} = 0\,, \tag{6}$$

i.e., if the condition $\partial u/\partial \nu = 0$ is replaced by the second order conormal condition corresponding to the operator T. Then Proposition 1 holds in the space $H^2(\Omega) \cap H_0^1(\Omega)$.

3. The CGM in Hilbert space

In this section we state the main result for the CGM for nonlinear equations in Hilbert space [2, 6]. We give the construction of the approximating sequence and the rate of its convergence. Before constructing the approximating sequence, we first summarize the assumptions.

The CG assumptions ([6, p. 99]). Let H be a real Hilbert space, $F : H \to$ $\to H$ a continuous operator such that

(a) F is twice Gâteaux differentiable;

(b) the first Gâteaux derivative of F is bihemicontinuous, symmetric and satisfies

$$m\|h\|^2 \leq \langle F'(u)h, h\rangle \leq M\|h\|^2 \qquad (u, h \in H)$$

with constants $M \geq m > 0$ independent of u, h;

(c) there exist $u_0 \in H$ and constants $R, B > 0$ such that for any $u \in$ $\in B(u_0, R) := \{u \in H : \|u - u_0\| \leq R\}$ there holds $\|F''(u)\| \leq B$;

(d) let $b \in H$ and $\phi \colon H \to \mathbf{R}$ such that $\phi'(u) = F(u) - b$. (This ϕ exists by the previous assumptions.) We assume that $\{u \in H : \phi(u) \leq \phi(u_0)\} \subset$ $\subset B(u_0, R)$ holds for the level set corresponding to u_0.

(We note that the original paper [2] assumes Frêchet differentiability in (a), but the Gâteaux sense suffices if the bihemicontinuity of F' is assumed in (b) as above.)

Construction of the CG iteration. Let $u_0 \in H$ be as in assumption (c), $p_0 = r_0 = b - F(u_0)$. For $n \in \mathbf{N} = \{0, 1, \ldots\}$, successively, let $u_{n+1} := u_n +$ $+ c_n p_n$ where c_n is the smallest positive root of $\langle F(u_n + cp_n) - b, p_n\rangle = 0$; set $r_{n+1} := b - F(u_{n+1})$, $p_{n+1} := r_{n+1} + b_n p_n$, where $b_n := -\langle F'(u_{n+1})p_n, r_{n+1}\rangle \setminus$ $\langle F'(u_{n+1})p_n, p_n\rangle$.

We use the following further notations:

for all $n \in \mathbf{N}$ let $\varepsilon_n := \langle F'(u_n)^{-1} r_n, r_n\rangle^{\frac{1}{2}}$, further, let $d := \frac{B}{M^3}\left(3 + \frac{M}{2m}\right)$, $\eta_n := \frac{\sqrt{M}B}{2m^2}\varepsilon_n$, $\sigma_n := \frac{4mM}{(M+m)^2}\frac{\eta_n}{1+\eta_n} + d\varepsilon_n$, $q := \frac{M-m}{M+m}$, $q_n := (q^2 + \sigma_n)^{1/2}$, $R_n := \frac{\sqrt{M}}{m(1-q_n)}\varepsilon_n$.

Then there holds

THEOREM 1 [2]. *Under the above assumptions (a)–(d) the following hold:*

(1) *The equation $F(u) = b$ has a unique solution $u^* \in H$, and the sequence (u_n) of the CG iteration converges strongly to u^*.*

(2) *Let $N_0 \in \mathbf{N}$ be such that $R_{N_0} < R$ and $\sigma_{N_0} < 1 - q^2$. Then*

$$\|u_n - u^*\| \leq R_{N_0} \cdot q_{N_0} \cdot q_{N_0+1} \cdots q_{n-1} \quad (n > N_0).$$

(Note that $\lim q_n = q$*).*

(3) *Let N_0 be as in (2). Then for any $k > N_0$ there exists $N_k \in \mathbf{N}$ such that*

$$\varepsilon_{n+k} \leq \left(4 \left(\frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} \right)^{2k} + \delta_n \right) \varepsilon_n \quad (n > N_k)$$

where $\lim \delta_n = 0$. *(Note that ε_n is equivalent to $\|F(u_n) - b\|$ and $\|u_n - u^*\|$.)*

4. The CGM for the boundary value problem

The Sobolev space version of the nonlinear conjugate gradient method is based on the Hilbert space analogue of CG methods [2]. For second order problems, the construction and proof of this is found in [6, p. 208] for problems with nonlinear principal part and in [9] for semilinear problems. We present the extension of the algorithm in [6] for fourth order problems, i.e. the application of the Daniel iteration to problem (1). The presented method is based on the algorithm in [2], we note that computationally convenient modifications might be also applied using [3].

We consider problem (1) under the assumptions (i)–(v). The corresponding Sobolev space $H_0^2(\Omega)$ is endowed with the inner product

$$\langle u, v \rangle_{H_0^2(\Omega)} := \int_\Omega \nabla^2 u : \nabla^2 v \qquad (u, v \in H_0^2(\Omega)). \tag{7}$$

REMARK 2. The induced norm

$$\|v\|_{H_0^2(\Omega)} = \left\| \left| \nabla^2 v \right|_{\mathsf{F}} \right\|_{L^2(\Omega)} \tag{8}$$

(where the notation (5) was used) is equivalent to the usual norm

$$\|v\|^*_{H_0^2(\Omega)} := \left(\sum_{|\alpha|=2} \int_\Omega |\partial^\alpha v|^2 \right)^{\frac{1}{2}},$$

since

$$\sum_{|\alpha|=2} |\partial^\alpha v|^2 + \sum_{\substack{i,j=1,\\ i<j}}^{N} |\partial_{ij} v|^2 = \left|\nabla^2 v\right|_{\mathsf{F}}^2 = 2 \sum_{|\alpha|=2} |\partial^\alpha v|^2 - \sum_{i=1}^{N} |\partial_{ii} v|^2$$

which implies

$$\|v\|^*_{H_0^2(\Omega)} \le \|v\|_{H_0^2(\Omega)} \le 2\,\|v\|^*_{H_0^2(\Omega)}\,.$$

The Hilbert space version of the method is established in Theorem 1. In order to ensure the twice differentiability of the generalized differential operator simply by that of the nonlinearity A, we apply Theorem 1 in a finite-dimensional subspace $V \subset H_0^2(\Omega)$, endowed with the same inner product (7), and we also assume that $V \subset W^{2,\infty}(\Omega)$. We are maily interested in FEM subspaces, i.e., when V consists of piecewise polynomials u such that $u \in C^1(\overline{\Omega})$. (This, together with the boundary conditions of (1), ensures $V \subset W^{2,\infty}(\Omega) \cap H_0^2(\Omega)$.)

Let

$$\langle F(u), v\rangle_{H_0^2(\Omega)} := \int_\Omega A(x, \nabla^2 u) : \nabla^2 v \qquad (u, v \in V),$$

we will see in the proof of Theorem 2 that this expression defines an operator $F : V \to V$. Further, let $b \in V$ the element defined by

$$\langle b, v\rangle_{H_0^2(\Omega)} = \int_\Omega g v \qquad (v \in V).$$

Denote by $u^* \in V$ the unique solution of the problem

$$\langle F(u^*), v\rangle_{H_0^2(\Omega)} = \langle b, v\rangle_{H_0^2(\Omega)} \qquad (v \in V).$$

The CG iteration constructs a sequence $(u_n) \subset V$ together with $(p_n) \subset V$ and the residuals $r_n = b - F(u_n) \in V$ as follows.

Let $u_0 \in V$ be arbitrary. Then $r_0 \in V$ is the solution of the problem

$$\int_\Omega \nabla^2 r_0 : \nabla^2 v = -\int_\Omega (A(x, \nabla^2 u_0) : \nabla^2 v - gv) \qquad (v \in V)$$

and $p_0 = r_0$. If, for $n \in \mathbf{N}$, u_n and p_n are obtained, then

$$u_{n+1} := u_n + c_n p_n \,,$$

where c_n is the smallest positive root of equation $\langle F(u_n + cp_n) - b, p_n \rangle_{H_0^2(\Omega)} =$ $= 0$, further, $r_{n+1} \in V$ is the solution of the problem

$$\text{(9)} \quad \int_\Omega \nabla^2 r_{n+1} : \nabla^2 v = -\int_\Omega (A(x, \nabla^2 u_{n+1}) : \nabla^2 v - gv) \qquad (v \in V);$$

finally,

$$p_{n+1} := r_{n+1} + b_n p_n$$

with $b_n = -\alpha_n / \beta_n$ where

$$\alpha_n = \int_\Omega \frac{\partial A}{\partial \Theta}(x, \nabla^2 u_{n+1})\ \nabla^2 p_n : \nabla^2 r_{n+1}\,,$$

$$\beta_n = \int_\Omega \frac{\partial A}{\partial \Theta}(x, \nabla^2 u_{n+1})\ \nabla^2 p_n : \nabla^2 p_n\,.$$

The above algorithm is based on [2]. We note that this can be simplified by suitable modifications using an approximate calculation of c_n, b_n [3].

The convergence results are formulated using the following notations: for any $n \in \mathbf{N}$ let

$$\varepsilon_n := \langle F'(u_n)^{-1} r_n, r_n \rangle^{1/2}_{H_0^2(\Omega)}\,,$$

$d := \frac{B}{M^3}\left(3 + \frac{M}{2m}\right)$, $\eta_n := \frac{\sqrt{M}B}{2m^2}\varepsilon_n$, $\sigma_n := \frac{4mM}{(M+m)^2}\frac{\eta_n}{1+\eta_n} + d\varepsilon_n$, $q := \frac{M-m}{M+m}$, $q_n := (q^2 + \sigma_n)^{1/2}$, $R_n := \frac{\sqrt{M}}{m(1-q_n)}\varepsilon_n$.

THEOREM 2. *Let the assumptions (i)–(v) hold for problem (1). Denote by $u^* \in V$ the unique weak solution.*

Then for arbitrary $u_0 \in V$, the above constructed CG sequence satisfies the following convergence results:

(1) Let $N_0 \in \mathbf{N}$ *be such that* $R_{N_0} < R$ *and* $\sigma_{N_0} < 1 - q^2$*. Then*

$$\|u_n - u^*\|_{H_0^2(\Omega)} \leq R_{N_0} \cdot q_{N_0} \cdot q_{N_0+1} \cdots q_{n-1} \quad (n > N_0)\,.$$

(Note that $\lim q_n = q$*).*

(2) Let N_0 *be as in (1). Then for any* $k > N_0$ *there exists* $N_k \in \mathbf{N}$ *such that*

$$\varepsilon_{n+k} \leq \left(4 \left(\frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} \right)^{2k} + \delta_n \right) \varepsilon_n \quad (n > N_k) \tag{10}$$

where $\lim \delta_n = 0$*. (Note that* ε_n *is equivalent to* $\|F(u_n) - b\|_{H_0^2(\Omega)}$ *and* $\|u_n - u^*\|_{H_0^2(\Omega)}$*.)*

REMARK 3. The proof will use the following facts:

(i) For matrix valued functions $H, K \in L^2(\Omega, \mathbf{R}^{N\times N})$ the integral of (3) is an inner product, hence the Cauchy–Schwartz inequality implies that

$$\left| \int_\Omega H : K \right| \leq \|\, |H|_{\mathsf{F}} \|_{L^2(\Omega)} \, \|\, |K|_{\mathsf{F}} \|_{L^2(\Omega)} \,.$$

(ii) We have from (4)

$$\left\| \frac{\partial A}{\partial \Theta}(x, \Theta) \right\|_{\mathsf{F}} \leq M, \tag{11}$$

where

$$\left\| \frac{\partial A}{\partial \Theta}(x, \Theta) \right\|_{\mathsf{F}} := \sup_{|R|_{\mathsf{F}} = |T|_{\mathsf{F}} = 1} \left| \frac{\partial A}{\partial \Theta}(x, \Theta)(R, T) \right| .$$

PROOF OF THEOREM 2. The assumptions (a)–(d) of Theorem 1 have to be checked.

First, the Lagrange inequality yields that

$$\|A(x, \Theta) - A(x, 0)\|_{\mathsf{F}} \leq \sup_{\Xi \in \mathbf{R}^{N\times N}} \left\| \frac{\partial A}{\partial \Theta}(x, \Xi) \right\|_{\mathsf{F}} |\Theta|_{\mathsf{F}} \,,$$

hence

$$\|A(x, \Theta)\|_{\mathsf{F}} \leq \|A(x, 0)\|_{\mathsf{F}} + M \, |\Theta|_{\mathsf{F}} \qquad (x \in \Omega,\ \Theta \in \mathbf{R}^{N\times N}). \tag{12}$$

For any $u \in V$ let $F(u) \in V$ be defined by

$$\langle F(u), v\rangle_{H_0^2(\Omega)} := \int_\Omega A(x, \nabla^2 u) : \nabla^2 v \qquad (v \in V).$$

Here the existence of $F(u)$ is ensured by the Riesz theorem since Remark 3, (8), and (12) yield

$$\left|\langle F(u), v\rangle_{H_0^2(\Omega)}\right| \le \left\|\left|A(x, \nabla^2 u)\right|_{\mathsf{F}}\right\|_{L^2(\Omega)} \|v\|_{H_0^2(\Omega)} \le$$
$$\le \left(\| |A(x,0)|_{\mathsf{F}} \|_{L^2(\Omega)} + M \|u\|_{H_0^2(\Omega)}\right) \|v\|_{H_0^2(\Omega)} .$$

That is, F: $V \to V$ is a well-defined operator.

Further, we present that F is twice Gâteaux differenciable. For any $u \in$ $\in V$ let $S(u) \in L(V)$ be defined by

$$\langle S(u)v, w\rangle_{H_0^2(\Omega)} := \int_\Omega \frac{\partial A}{\partial \Theta}(x, \nabla^2 u)(\nabla^2 v, \nabla^2 w) \qquad (\forall v, w \in V)$$

The existence of $S(u)$ is provided again by the Riesz theorem since, using (11),

$$\left|\langle S(u)v, w\rangle_{H_0^2(\Omega)}\right| \le \left\|\left\|\frac{\partial A}{\partial \Theta}(x, \nabla^2 u)\right\|_{\mathsf{F}} \left|\nabla^2 v\right|_{\mathsf{F}} \left|\nabla^2 w\right|_{\mathsf{F}}\right\|_{L^2(\Omega)} \le$$
$$\le M \|v\|_{H_0^2(\Omega)} \|w\|_{H_0^2(\Omega)} .$$

We verify that $F' = S$ in Gâteaux sense:

$$(13) \qquad \left\|\frac{1}{t}(F(u + tv) - F(u)) - S(u)v\right\|_{H_0^2(\Omega)} =$$
$$= \sup_{\|w\|_{H_0^2(\Omega)}=1} \left\langle \frac{1}{t}(F(u + tv) - F(u)) - S(u)v, w \right\rangle_{H_0^2(\Omega)} =$$
$$= \sup_{\|w\|_{H_0^2}=1} \int_\Omega \left(\frac{A(x, \nabla^2 u + t\nabla^2 v) - A(x, \nabla^2 u)}{t} - \frac{\partial A}{\partial \Theta}(x, \nabla^2 u)\nabla^2 v \right) : \nabla^2 w =$$
$$= \sup_{\|w\|_{H_0^2}=1} \int_\Omega \left(\frac{\partial A}{\partial \Theta}(x, \nabla^2 u + \xi(x,t)\nabla^2 v) - \frac{\partial A}{\partial \Theta}(x, \nabla^2 u) \right) (\nabla^2 v, \nabla^2 w) \to 0$$

as $0 \le \xi(x,t) \le t \to 0$, using the Lebesgue theorem. Namely, we can check the conditions of this theorem as follows: $\frac{\partial A}{\partial \Theta}$ is a continuous mapping and $\nabla^2 u, \nabla^2 v \in L^2(\Omega)$ are fixed, hence implies that the integrand tends to 0 a.e. as $0 \le \xi(x,t) \le t \to 0$, further, (11) implies that

$$2M \left|\nabla^2 v\right|_{\mathsf{F}} \left|\nabla^2 w\right|_{\mathsf{F}}$$

is a major function of the integrand which belongs to $L^1(\Omega)$:

$$\int_\Omega 2M \left|\nabla^2 v\right|_{\mathsf{F}} \left|\nabla^2 w\right|_{\mathsf{F}} \le 2M \left\|v\right\|_{H_0^2(\Omega)} \left\|w\right\|_{H_0^2(\Omega)} < \infty.$$

In the next step we prove the existence of the second derivative. First we recall that $V \subset W^{2,\infty}(\Omega) \cap H_0^2(\Omega)$ is a finite dimensional subspace, hence the following holds:

$$\exists c(V) > 0: \qquad \left\|u\right\|_{W^{2,\infty}} := \left\| \left|\nabla^2 u\right|_{\mathsf{F}} \right\|_{L^\infty(\Omega)} \le c(V) \left\| \left|\nabla^2 u\right|_{\mathsf{F}} \right\|_{L^2(\Omega)} =$$

$$= c(V) \left\|u\right\|_{H_0^2(\Omega)} \quad (u \in V). \tag{14}$$

For any $u \in V$ let $P(u) \in L(V, L(V)) \equiv L_{(2)}(V^2, V)$ defined by

$$\left\langle P(u)(v,w), z \right\rangle_{H_0^2(\Omega)} := \int_\Omega \frac{\partial^2 A}{\partial \Theta^2}(x, \nabla^2 u)(\nabla^2 v, \nabla^2 w, \nabla^2 z) \quad (\forall v, w, z \in V).$$

Using again the Riesz theorem, we can see that $P(u)$ exists, since assumption (v) and (14) yield

$$\left| \left\langle P(u)(v,w), z \right\rangle_{H_0^2(\Omega)} \right| \le \int_\Omega K \left|\nabla^2 v\right|_{\mathsf{F}} \left|\nabla^2 w\right|_{\mathsf{F}} \left|\nabla^2 z\right|_{\mathsf{F}} \le$$

$$K \left\|v\right\|_{W^{2,\infty}} \left\|w\right\|_{H_0^2(\Omega)} \left\|z\right\|_{H_0^2(\Omega)} \le Kc(V) \left\|v\right\|_{H_0^2(\Omega)} \left\|w\right\|_{H_0^2(\Omega)} \left\|z\right\|_{H_0^2(\Omega)} . \tag{15}$$

We show that $F'' = P$ in Gâteaux sense:

$$\left\| \frac{1}{t}(F'(u+tv) - F'(u)) - P(u)v \right\|_{L(V)} =$$

$$= \sup_{\left\|w\right\|_{H_0^2(\Omega)} = \left\|z\right\|_{H_0^2(\Omega)} = 1} \left\langle \frac{1}{t}(F'(u+tv)w - F'(u)w) - P(u)(v,w), z \right\rangle_{H_0^2(\Omega)} =$$

$$= \int\limits_\Omega \left(\frac{\frac{\partial A}{\partial\Theta}(x,\nabla^2 u+t\nabla^2 v)-\frac{\partial A}{\partial\Theta}(x,\nabla^2 u)}{t} - \frac{\partial^2 A}{\partial\Theta^2}(x,\nabla^2 u)\nabla^2 v \right)(\nabla^2 w,\nabla^2 z)=$$

$$= \int\limits_\Omega \left(\frac{\partial^2 A}{\partial\Theta^2}(x,\nabla^2 u+\xi(x,t)\,\nabla^2\, v)-\frac{\partial^2 A}{\partial\Theta^2}(x,\nabla^2 u) \right)(\nabla^2 v,\nabla^2 w,\nabla^2 z)\to 0$$

as $0\le\xi(x,t)\le t\to 0$, using the Lebesgue theorem. Now we check the conditions of this theorem: $\frac{\partial^2 A}{\partial\Theta^2}$ is a continuous mapping and $\nabla^2 u,\nabla^2 v,\nabla^2 z\in$ $\in L^2(\Omega)$ are fixed, hence the integrand tends to 0 a.e. as $0\le\xi(x,t)\le t\to 0$, further, assumption (v) implies that

$$2K\left|\nabla^2 v\right|_{F}\left|\nabla^2 w\right|_{F}\left|\nabla^2 z\right|_{F}$$

is a major function of the integrand which belongs to $L^1(\Omega)$, since by (15),

$$\int\limits_\Omega 2K\left|\nabla^2 v\right|_{F}\left|\nabla^2 w\right|_{F}\left|\nabla^2 z\right|_{F}\le 2Kc(V)\,\|v\|_{H_0^2(\Omega)}\,\|w\|_{H_0^2(\Omega)}\,\|z\|_{H_0^2(\Omega)}\,.$$

The demonstration of the bihemicontinuity of F' is much similar:

$$\left\|(F'(u+tv+sw)-F'(u))h\right\|_{H_0^2(\Omega)}=$$

$$=\sup_{\|z\|_{H_0^2(\Omega)}=1}\left\langle(F'(u+tv+sw)-F'(u))h,z\right\rangle_{H_0^2(\Omega)}=$$

$$=\int\limits_\Omega\left(\frac{\partial A}{\partial\Theta}(x,\nabla^2 u+t\,\nabla^2\,v+s\,\nabla^2\,w)-\frac{\partial A}{\partial\Theta}(x,\nabla^2 u)\right)(\nabla^2 h,\nabla^2 z)\to 0$$

as $s,t,\to 0$ for the same reason as in (13).

The symmetry follows directly:

$$\left\langle F'(u)v,w\right\rangle_{H_0^2(\Omega)}=\int\limits_\Omega\frac{\partial A}{\partial\Theta}(x,\nabla^2 u)(\nabla^2 v,\nabla^2 w)=$$

$$=\sum_{i,k,r,s=1}^{N}\int\limits_\Omega\frac{\partial A_{ik}(x,\Theta)}{\partial\Theta_{rs}}\partial_i\partial_k v\;\partial_r\partial_s w=$$

$$=\sum_{i,k,r,s=1}^{N}\int\limits_\Omega\frac{\partial A_{rs}(x,\Theta)}{\partial\Theta_{ik}}\partial_r\partial_s w\;\partial_i\partial_k v=\left\langle v,F'(u)w\right\rangle_{H_0^2(\Omega)}\,.$$

It is easy to verify the spectral conditions using (4):

$$m\,\|h\|^2_{H_0^2(\Omega)} = \int\limits_\Omega m\left|\nabla^2 h\right|^2_{\mathsf{F}} \le \int\limits_\Omega \frac{\partial A}{\partial\Theta}(x,\nabla^2 u)(\nabla^2 h,\nabla^2 h) =$$

$$= \left\langle F'(u)h,h\right\rangle_{H_0^2(\Omega)} \le \int\limits_\Omega M\left|\nabla^2 h\right|^2_{\mathsf{F}} = M\,\|h\|^2_{H_0^2(\Omega)}.$$

F'' is bounded since, using (15):

$$\left\|F''(u)\right\|_{\mathsf{F}} = \sup_{\|v\|_{H_0^2(\Omega)}=\|w\|_{H_0^2(\Omega)}=\|z\|_{H_0^2(\Omega)}=1} \left|\left\langle F''(u)(v,w),z\right\rangle_{H_0^2(\Omega)}\right| \le$$

$$\le \sup_{\|v\|_{H_0^2(\Omega)}=\|w\|_{H_0^2(\Omega)}=\|z\|_{H_0^2(\Omega)}=1} Kc(V)\,\|v\|_{H_0^2(\Omega)}\,\|w\|_{H_0^2(\Omega)}\,\|z\|_{H_0^2(\Omega)} =$$

$$= Kc(V) \qquad (u\in V).$$

Since F has a symmmetric bihemicontinuous Gâteaux derivative, F also has a potential Ψ, therefore

$$\phi(u) = \Psi(u) - \langle u,b\rangle \qquad (u\in V)$$

is the potential mentioned in assumption (d) since

$$\phi'(u) = \Psi'(u) - b = F(u) - b.$$

Note that in this way $\phi\in C^2(V)$, therefore

$$\phi(u) = \phi(0) + \left\langle\phi'(0),u\right\rangle + \frac{1}{2}\left\langle\phi''(0)u,u\right\rangle \ge$$

$$\ge \phi(0) + \left(\frac{m}{2}\|u\|_{H_0^2(\Omega)} - \left\|\phi'(0)\right\|\right)\|u\|_{H_0^2(\Omega)}.$$

Hence

$$\lim_{\|u\|_{H_0^2(\Omega)}\to\infty} \phi(u) = \infty,$$

Therefore the level sets of ϕ are bounded, i.e for any $u_0\in V$ they are contained in some ball, which is suitable to satisfy the required assumptions (c) and (d) of Theorem 1. ∎

REMARK 4. The same construction of the CGM can be carried out and the same convergence result as in Theorem 2 holds if problem (1) is considered with the boundary conditions (6) mentioned in Remark 1. In this case we simply choose the finite-dimensional subspace V in $H_0^1(\Omega) \cap W^{2,\infty}(\Omega)$ instead of $H_0^2(\Omega) \cap W^{2,\infty}(\Omega)$.

REMARK 5. (Numerical aspects.) The iteration involves a preconditioning matrix obtained as the discretized biharmonic operator, that is, a discrete biharmonic problem (9) has to be solved stepwise. This is not costly since efficient fast solvers are available for the biharmonic problem [1, 5, 13]. (For this reason biharmonic preconditioning operators have also been used in other iterative methods [10, 11].) On the other hand, the operator preconditioning implies that the convergence factor of our method is mesh independent: as shown by (10), the factor $(\sqrt{M} - \sqrt{m})/(\sqrt{M} + \sqrt{m})$ only depends on the bounds of the coefficient in (2).

References

[1] BJORSTAD, P. E.: Fast numerical solution of the biharmonic Dirichlet problem on rectangles, *SIAM J. Numer. Anal.* **20** (1983), no. 1, 59–71.

[2] DANIEL, J. W.: The conjugate gradient method for linear and nonlinear operator equations, *SIAM J. Numer. Anal.* **4** (1967), 10–26.

[3] DANIEL, J. W.: Convergence of the conjugate gradient method with computationally convenient modifications, *Numer. Math.* **10** (1967), 125–131.

[4] DANIEL, J. W.: A correction concerning the convergence rate for the conjugate gradient method, *SIAM J. Numer. Anal.* **7** (1970), 277–280.

[5] EWING, R. E., MARGENOV, S. D., and VASSILEVSKI, P. S.: Preconditioning the biharmonic equation by multilevel iterations, *Math. Balkanica (N.S.)* **10** (1996), no. 1, 121–132.

[6] FARAGÓ, I. and KARÁTSON, J.: *Numerical solution of nonlinear elliptic problems via preconditioning operators: Theory and applications*, Nova Science Publisher, Inc., New York, 2002.

[7] HAYES, R. M.: Iterative methods of solving linear problems in Hilbert space, *Nat. Bur. Standards Appl. Math. Ser.* **39** (1954), 71–104.

[8] HESTENES, M. R. and STIEFEL, E.: Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Standards* Sect. B **49** (1952) No. 6., 409–436.

[9] KARÁTSON, J.: The conjugate gradient method for a class of non-differentiable operators, *Annales Univ. Sci. ELTE* **40** (1997), 121–130.

[10] KARÁTSON, J.: Sobolev space preconditioning of strongly nonlinear 4th order elliptic problems, in: *Numerical Analysis and Its Applications*, Sec. Int. Conf. NAA 2000 (Rousse, Bulgaria), eds. L. Vulkov, J. Wasniewski, P. Yalamov, pp. 459–466, *Lecture Notes Comp. Sci.* **1988**, Springer, 2001.

[11] KOSHELEV, A.: Regularity problem for quasilinear elliptic and parabolic systems, *Lecture Notes in Mathematics* **1614**, Springer-Verlag, Berlin, 1995.

[12] LANGENBACH, A.: *Monotone Potentialoperatoren in Theorie und Anwendung*, Springer, 1977.

[13] LANGER, U.: A fast iterative method for solving the first boundary value problem for the biharmonic equation (in Russian), *Zh. Vychisl. Mat. i Mat. Fiz.* **28** (1988), no. 2, 209–223, 302.

[14] MIKHLIN, S. G.: *The Numerical Performance of Variational Methods*, Walters–Noordhoff, 1971.

A. Márkus

Department of Applied Analysis
ELTE University
Pf. 120
H-1518 Budapest
Hungary
aurelius@cs.elte.hu

SPECTRAL APPROXIMANTS INVOLVING BALANCED AND CONVEX SETS

By

PHILIP MAHER

(February 6, 2004)

Let E denote a non-empty, closed set in $\mathbb{C}$ and let $\mathcal{S}(E)$ denote the set of all those (bounded, linear) normal operators X (on a fixed, separable Hilbert space H) each of whose spectrum $\sigma(X) \subseteq E$. A *spectral approximant* from $\mathcal{S}(E)$ of some operator A, with respect to some norm $|||\cdot|||$, is an operator, say, X_0 in $\mathcal{S}(E)$ that minimizes the quantity $|||A - X|||$ as X varies in $\mathcal{S}(E)$, subject to $|||A - X||| < \infty$, so that for all such X

$$|||A - X_0||| \leq |||A - X|||.$$

The subject of spectral approximants was initiated by Halmos in [6]. Halmos' work involves the concept of retraction. A (distance-minimizing) *retraction* for the non-empty, closed set E is a map $F\colon \mathbb{C} \to E$ such that

$$|\alpha - F(\alpha)| \leq |\alpha - \xi|$$

for all ξ in E where $\alpha \in \mathbb{C}$. Each non-empty, closed set E has a Borel measurable retraction and if E is convex there is a unique retraction; see [6] for more about retractions.

Halmos' main result [6, Theorem] says that if A normal and F is a Borel measurable retraction for the non-empty, closed set E then $F(A) \in \mathcal{S}(E)$ and

$$\text{(H)} \qquad ||A - F(A)|| \leq ||A - X||$$

for all X in $\mathcal{S}(E)$.

Bouldin [4] extended Halmos' work to the context of the von Neumann–Schatten classes $\mathcal{C}_p$ and norms $||\cdot||_p$ (For background on $\mathcal{C}_p$ and $||\cdot||_p$ see, for instance, [9, Chapter 2]). Bouldin's variant [4, Theorem 2] of (H) goes as follows: let E, F and A be as in the previous paragraph; then for all X

in $\mathcal{S}(E)$ such that $A - X \in \mathcal{C}_p$, provided that $2 \leq p < \infty$, it follows that $A - F(A) \in \mathcal{C}_p$ and

$$\text{(B)} \qquad ||A - F(A)||_p \leq ||A - X||_p.$$

Bouldin's result does does not necessarily hold for the $1 \leq p < 2$ case, as Bhattia showed by a very simple counter-example [3, p. 35]. In [3, Theorem 1] Bhattia extended (B) to the $1 \leq p < 2$ case for convex E. In Theorem 1(b) we give a slightly more direct proof of Bhattia's inequality for $|| \cdot ||_p$ (although Bhattia's proof applies to a wider class of unitarily invariant norms).

More significant is the corresponding result, Theorem 1(a), for balanced sets (A set E in $\mathbb{C}$ is *balanced* if $z \in E \Rightarrow e^{i\theta}z \in E$ for all θ). As an immediate consequence, viz. Corollary 3, of Theorem 1(a) we extend the result [8, Theorem 5.11(b)] on normal, partially isometric approximation of positive operators from $2 \leq p < \infty$ to $1 \leq p < \infty$ ([8] invokes Bouldin's $2 \leq p < \infty$ inequality (B)). The work of [8] is itself an extension of [1] which is relevant to quantum chemistry: see [2], [5].

THEOREM 1. *Let A be normal, F be a Borel measurable retraction for the non-empty, closed set E and let X vary such that $X \in \mathcal{S}(E)$ and $A - X \in$ $\in \mathcal{C}_p$ for $1 \leq p < \infty$. Then:*
(a) if E is balanced it follows that $A - F(A) \in \mathcal{C}_p$ and

$$||A - F(A)||_p \leq ||A - X||_p\,;$$

(b) if E is assumed convex, rather than balanced, the same conclusion as in (a) holds.

PROOF. (a) The proof is set up as in Bouldin [4, pp. 280–281]. Let $\{\phi_i\}$, where $1 \leq i \leq l(\leq \infty)$, be a maximal, orthonormal set of eigenvectors of A corresponding to the (countable set of) isolated eigenvalues $\{\alpha_i\}$ not contained in E (where the α_i are in decreasing order of magnitude and repeated according to (finite) multiplicity). From [4, p. 281] it follows that, for $1 \leq p < \infty$,

$$(1) \qquad ||A - F(A)||_p^p = \sum_{i=1}^{l} |\alpha_i - F(\alpha_i)|^p.$$

Let $\alpha = |\alpha|e^{i\theta}$. Since F maps $\mathbb{C}$ onto the balanced set E then $F(\alpha) =$ $= |F(\alpha)|e^{i\theta}$; and since F is a retraction on $\mathbb{R}^+$ onto $|E| \overset{\text{def}}{=} \{|f|: f \in E\}$ then $F(|\alpha|) = |F(\alpha)|$. Thus, for all α in $\mathbb{C}$

$$|a - F(a)| = |\,|\alpha| - |F(\alpha)|\,| = |\,|\alpha| - F(|\alpha|)\,| \leq |\,|\alpha| - |\xi|\,|$$

for all $|\xi|$ in $|E|$. Since A is normal then $|\alpha_i| = s_i(A)$ (the i^{th} singular value of A) for $1 \leq i \leq l$. Hence,

$$|\alpha_i - F(\alpha_i)| \leq |\, s_i(A) - |\xi||.$$

Since $\sigma(X) \subseteq E$ it follows that the point spectrum $\sigma_p(X) \subseteq E$ and hence, as X is normal, each $s_j(X) \in |E|$. Hence, the $|\xi|$ occurring above may vary over all the singular values $s_j(X)$ for $1 \leq j \leq \infty$ (the $s_j(X)$ being in decreasing order and repeated according to multiplicity). Hence, in particular,

$$|\alpha_i - F(\alpha_i)| \leq |s_i(A) - s_i(X)|$$

for $1 \leq i \leq l \leq \infty$. Therefore, from (1),

$$||A - F(A)||_p^p \leq \sum_{i=1}^{l} |s_i(A) - s_i(X)|^p. \tag{2}$$

From [10, (1.2.2)] it follows that if $\sum_{j=1}^{\infty} s_j(A - X)^p$ (= $||A - X||_p^p$) $< \infty$ then $\sum_{j=1}^{\infty} |s_j(A) - s_j(X)|^p \leq \sum_{j=1}^{\infty} s_j(A - X)^p$. Therefore, from (2),

$$||A - F(A)||_p^p \leq ||A - X||_p^p.$$

(b) Let $\{\alpha_i\}$ and $\{\phi_i\}$, where $1 \leq i \leq l \leq \infty$, be as in (a). Then, for $1 \leq p < \infty$,

$$\begin{aligned} ||A - F(A)||_p^p &= \sum_{i=1}^{l} |\alpha_i - F(\alpha_i)|^p \\ &\leq \sum_{i=1}^{l} |\alpha_i - \langle X\phi_i, \phi_i\rangle|^p \\ &= \sum_{i=1}^{l} |\langle (A - X)\phi_i, \phi_i\rangle|^p \\ &\leq ||A - X||_p^p, \end{aligned}$$

where the last inequality above follows from [9, Lemma 2.3.4] and the first inequality follows from the convexity of E: for, with W denoting numerical range [7, Problem 216],

$$\langle X\phi_i, \phi_i\rangle \in W(X) \subseteq \overline{W(X)} = \operatorname{conv} \sigma(X) \subseteq \operatorname{conv} E = E$$

whence $|\alpha_i - F(\alpha_i)| \leq |\alpha_i - \langle X\phi_i, \phi_i\rangle|$ for $1 \leq i \leq l$. ∎

The next two results concern approximation of a positive operator (An operator A is positive, denoted $A \geq 0$, if $\langle Af,f \rangle \geq 0$ for all f in H and strictly positive if $\langle Af,f \rangle > 0$ for all non-zero f in H). Corollary 2 below recaptures very simply a result of Aitken, Erdos and Goldstein [1, Corollary 3.6].

COROLLARY 2. *Let A be positive and X vary over those unitary operators such that $A - X \in \mathcal{C}_p$ for $1 \leq p < \infty$. Then $A - I \in \mathcal{C}_p$ and*

$$(1) \qquad ||A - I||_p \leq ||A - X||_p \ \ \textit{if} 1 \leq p < \infty\,.$$

PROOF. Let $E = \{z : |z| = 1\}$. If a normal operator has its spectrum in E then it is unitary and conversely. Let F be given by

$$F(z) = \begin{cases} \frac{z}{|z|} & \text{if } z \neq 0 \\ 1 & \text{if } z = 0\,. \end{cases}$$

Then F is a retraction onto E and, as $A \geq 0$, it follows that $F(A) = I$. Since E is balanced the inclusion $A - I \in \mathcal{C}_p$ and the inequality (1) follow from Theorem 1(a). ∎

COROLLARY 3. *Let A be positive and X vary over those normal partial isometries such that $A - X \in \mathcal{C}_p$ for $1 \leq p < \infty$. Then:*

(a) the map $X \to ||A - X||_p$ has a global minimizer;

(b) for $1 < p < \infty$, there exists a basis $\{\phi_n\}$ of H consisting of eigenvectors of A; and, with $E_{\frac{1}{2}}$ denoting the projection onto $\overline{S}\{\phi_n : A\phi_n = \alpha_n\phi_n$ and $\alpha_n \geq \frac{1}{2}\}$,

$$(1) \qquad ||A - E_{\frac{1}{2}}||_p \leq ||A - X||_p$$

with equality occurring in (1) *if, and for strictly positive A such that $\frac{1}{2} \notin \sigma_p(A)$ only if, $X = E_{\frac{1}{2}}$.*

PROOF. (a) Let $E = \{0\} \cup \{z : |z| = 1\}$. From [8, Theorem 5.10] it follows that $X \in \mathcal{S}(E)$ if and only if X is a normal partial isometry. Since E is balanced the existence of a global minimizer of $X \to ||A - X||_p$ is immediate from Theorem 1(a).

(b) This follows from (a) and from [8, Theorem 5.6] (The restriction that $p > 1$ is required since [8, Theorem 5.6] depends on the local theory developed in [8] which is valid only for $1 < p < \infty$). ∎

References

[1] J. G. AIKEN, J. A. ERDOS and J. A. GOLDSTEIN: Unitary approximation of positive operators, *Illinois J. Math.* **24** (1980), 61–72.

[2] J. G. AITKEN, J. A. ERDOS and J. A. GOLDSTEIN: On Lowdin orthogonalization, *Internat. J. Quantum Chem.* **18** (1980), 1101–1108.

[3] R. BHATTIA: Some inequalities for norm ideals, *Commun. Math. Phys.* **III** (1987), 33–39.

[4] R. BOULDIN: Best approximation of a normal operator in the Schatten p norm, *Proc. Amer. Math. Soc.* **80** (1980), 227–282.

[5] J. A. GOLDSTEIN and M. LEVY: Linear algebra and quantum chemistry, *Amer. Math. Monthly* **98** (1991), 710–718.

[6] P. R. HALMOS: Spectral approximants of normal operators, *Proc. Edinburgh Math Soc.* **19** (1974), 51–58.

[7] P. R. HALMOS: *A Hilbert space problem book*, 2nd ed., Springer-Verlag, New York, 1982.

[8] P. J. MAHER: Partially isometric approximation of positive operators, *Illinois J. Math.* **33** (1989), 227–243.

[9] J. R. RINGROSE: *Compact non-self-adjoint operators*, Van Nostrand Rheinhold, London, 1971.

[10] B. SIMON: *Trace ideals and their applications*, Cambridge University Press, Cambridge, 1979.

Philip Maher

Mathematics and Statistics Academic Group
Middlesex University, Tottenham Campus
White Hart Lane
London N17 8HR
England
p.maher@mdx.ac.uk

A COMMON FIXED POINT THEOREM FOR QUASI CONTRACTIVE TYPE MAPPINGS

By

VASILE BERINDE

(Received February 16, 2004)

1. Introduction

The well known Banach's fixed point theorem (also named contraction mapping principle) is one of the most useful results in fixed point theory. In a metric space setting it can be briefly stated as follows.

THEOREM B. *Let (X,d) be a complete metric space and $T : X \longrightarrow X$ a strict contraction, i.e., a map satisfying*

$$(1.1) \qquad d(Tx, Ty) \leq a\, d(x,y), \quad \textit{for all} \quad x,y \in X,$$

where $0 < a < 1$ is constant. Then T has a unique fixed point in X.

Theorem B, together with its direct generalizations and local variants, has many applications in solving nonlinear functional equations, but suffers from one drawback - the contractive condition (1.1) forces that T be continuous throughout X. In order to remove this drawback, in 1968 Kannan [9] obtained a fixed point theorem for mappings T that need not be continuous.

THEOREM K. *Let (X,d) be a complete metric space and $T : X \longrightarrow X$ a mapping for which there exists $a \in \left(0, \frac{1}{2}\right)$ such that*

$$(1.2) \qquad d(Tx, Ty) \leq a\left[d(x, Tx) + d(y, Ty)\right], \quad \textit{for all} \quad x,y \in X.$$

Then T has a unique fixed point in X.

EXAMPLE 1. Let $X = \mathbb{R}$ be the set of real numbers with the usual metric and $T : \mathbb{R} \longrightarrow \mathbb{R}$, given by $Tx = 0$, if $x \in (-\infty, 2]$ and $Tx = -\frac{1}{2}$, if $x \in (2, \infty)$.

Then T satisfies (1.2) with $a = \frac{1}{5}$, T is not continuous and $F_T = \{0\}$.

Following Kannan's theorem, a lot of papers were devoted to obtaining fixed point theorems for various classes of contractive type conditions that do not require the continuity of T, see for example Rus [13]. In this context, a very interesting theorem which extends both Banach's and Kannan's fixed point theorems, alongside many other similar results of this kind, was obtained in 1972 by Zamfirescu [14].

THEOREM Z. *Let (X,d) be a complete metric space and $T : X \longrightarrow X$ a mapping for which there exist the real numbers α, β and γ satisfying $0 < < \alpha < 1$, $0 < \beta < 1/2$ and $0 < \gamma < 1/2$ such that, for each $x, y \in X$, at least one of the following is true:*

(z_1) $\quad d(Tx, Ty) \leq \alpha\, d(x,y);$

(z_2) $\quad d(Tx, Ty) \leq \beta \left[d(x, Tx) + d(y, Ty)\right];$

(z_3) $\quad d(Tx, Ty) \leq \gamma \left[d(x, Ty) + d(y, Tx)\right].$

Then T has a unique fixed point in X.

One of the most general contraction conditions obtained in this way, for which the Picard iteration still converge to the unique fixed point, was given by Ciric [7] in 1974.

THEOREM C. *Let (X,d) be a complete metric space and $T : X \longrightarrow X$ a mapping that satisfies*

(1.3)

$$d(Tx, Ty) \leq h \cdot \max\{d(x,y), d(x, Tx), d(y, Ty), d(x, Ty), d(y, Tx)\},$$

for all $x, y \in X$ and some constant $0 < h < 1$.
Then T has a unique fixed point in X.

REMARK. It is easy to see that if T is an operator that satisfies the assumptions in any of the Theorems B, K and Z, then T also satisfies the assumptions of Theorem C.

The set $0_T(x) = \{x, Tx, T^2x, \ldots\}$ is called *the orbit* of T relative to x. It is shown in [15] that condition (1.3) does in fact assure that the orbits of T are bounded.

There exist many extensions and generalizations of these results. One of them was given in [1], for the class of the so called generalized φ-contractions, as a unifying fixed point theorem of many results of the same kind.

A mapping $T : X \longrightarrow X$ is said to be a *generalized* φ*-contraction* if there exists a function $\varphi : \mathbb{R}_+^5 \longrightarrow \mathbb{R}_+$ (called *comparison function* and satisfying certain appropriate conditions) such that for all $x, y \in X$

$$d(Tx, Ty) \leq \varphi\big(d(x,y),\, d(x, Tx),\, d(y, Ty),\, d(x, Ty),\, d(y, Tx)\big)\,. \tag{1.4}$$

EXAMPLE 2. The functions

1) $\varphi_1(t) = \alpha t_1$, for all $t = \big(t_1, t_2, t_3, t_4, t_5\big) \in \mathbb{R}_+^5$ $(0 \leq \alpha < 1)$;

2) $\varphi_2(t) = a(t_2 + t_3)$, for all $t = \big(t_1, t_2, t_3, t_4, t_5\big) \in \mathbb{R}_+^5$, $0 \leq a < \frac{1}{2}$;

3) $\varphi_3(t) \in \{\alpha t_1, \beta(t_2 + t_3), \gamma(t_4 + t_5)\}$, for all $t = \big(t_1, t_2, t_3, t_4, t_5\big) \in$ $\in \mathbb{R}_+^5$, $0 \leq \alpha < 1$; $0 \leq \beta < \frac{1}{2}$; $0 \leq \gamma < \frac{1}{2}$;

4) $\varphi_4(t) = h \cdot \max\{t_1, t_2, t_3, t_4, t_5\}$, for all $t = \big(t_1, t_2, t_3, t_4, t_5\big) \in$ $\in \mathbb{R}_+^5$, $0 < h < 1$

are all comparison functions. (Recall that a map satisfying (1.4) with $\varphi \equiv \varphi_4$ is usually called quasi contraction).

In a slightly corriged version, see Berinde [2], the main result in [1] can be briefly restated as follows.

THEOREM G. *Let* (X, d) *be a complete metric space and* $T : X \longrightarrow X$ *a generalized* φ*-contraction with* φ *such that* $\psi(t) = \varphi(t, t, t, t, t)$ *is a continuous comparison function and* $h(t) = t - \psi(t)$ *is an increasing bijection. Then*

(i) T *has a unique fixed point* p *in* X*;*

(ii) The Picard iteration $\{x_n\}_{n=0}^{\infty}$*, given by* $x_{n+1} = Tx_n$*,* $n \geq 0$ *and* $x_0 \in X$*, converges to* p*;*

(iii) $d(x_n, p) \leq \psi^n\big(h^{-1}(d(x_0, x_1))\big)$, $\quad n \geq 1$.

It is the main purpose of the present paper to extend Theorem G, and hence all fixed point theorems contained by it as particular cases, to a common fixed point theorem.

2. A common fixed point theorem

The important result given by Theorem C has been also extended in many directions: to nonself mappings, Ciric ([8], Theorem 2.1) by using Rothe's boundary condition, to generalized orbitally complete metric spaces with the metric satisfying a quadrilateral inequality instead of the usual triangle inequality, see Lahiri and Das [10], as well as to a common fixed point for nonself mappings, see Rakocevic [11] and Berinde [4], and also to orbitally complete metric spaces, see Ciric [6].

In this section we state and prove a general common fixed point theorem for self operators satisfying a generalized condition of quasi-contractive type.

To this end we need some appropriate notions and results related to mappings with contracting orbital diameters.

REMARKS.

1) A mapping satisfying a contractive condition of the form (1.4) is generally not continuous throughout X. However, as shown by Rhoades ([12], Theorem 2), a contractive mapping satisfying (1.3) is continuous *at the fixed point*. The argument is easily extendable to mappings satisfying (1.4) with φ an appropriate comparison function.

2) One of the first authors who considered conditions of the form (1.4) with $\varphi(t) \equiv \varphi(t_1)$, $t = (t_1, t_2, t_3, t_4, t_5) \in \mathbb{R}_+^5$, was Browder [5].
A scalar function $\varphi : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ involved in such a fixed point theorem is also called *comparison function* and is supposed to satisfy at least the following two conditions:

(i_φ) φ is monotonically increasing, i.e., $t_1 < t_2 \Rightarrow \varphi(t_1) \leq \varphi(t_2)$;

(ii_φ) The sequence $\{\varphi^n(t)\}_{n=0}^\infty$ converges to zero for each $t \in \mathbb{R}_+$, where φ^n stands for the n^{th} iterate of φ.

A prototype for the scalar comparison functions is $\varphi(t) = a \cdot t$, $t \in \mathbb{R}_+$, with $0 \leq a < 1$.
Considering $\varphi_1(t) = \frac{t}{1+t}$, $t \in \mathbb{R}_+$ and $\varphi_2(t) = \frac{1}{2}t$, if $0 \leq t < 1$ and $\varphi_2(t) = t - \frac{1}{3}$, if $t \geq 1$, it is easy to check that comparison functions need not be neither linear, nor continuous.

To prove our main result we shall use the following Lemma.

LEMMA 1. *Let* $\varphi : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ *satisfy* (i_φ) *and* (ii_φ) *and suppose*

$$t \leq \varphi(t)\,, \tag{2.1}$$

for a certain $t \in \mathbb{R}_+$. *Then* $t = 0$.

PROOF. Assume the contrary, i.e., there exists $t > 0$ such that (2.1) is satisfied. Then, by (i_φ) we inductively get

$$t \leq \varphi^n(t)\,, \quad n \geq 1$$

and so, in view of (ii_φ), this implies

$$0 \le t \le \lim_{n\to\infty} \varphi^n(t) = 0\,,$$

a contradiction. ∎

The main result of this paper is given by the next theorem.

THEOREM 1. *Let (X,d) be a complete metric space and S, $T : X \longrightarrow X$ two mappings with bounded orbits. Suppose T is continuous and*

$$d(Sx,Sy) \le \varphi\big(M(x,y)\big)\,, \quad \textit{for all} \quad x,y \in X, \tag{2.2}$$

where

(2.3)

$$M(x,y) = \max\{d(Tx,Ty),\, d(Tx,Sx),\, d(Ty,Sy),\, d(Tx,Sy), d(Ty,Sx)\}\,,$$

with $\varphi : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ a continuous scalar comparison function. Suppose

$$S(X) \subset T(X) \tag{2.4}$$

and also suppose T and S are weakly commutative, i.e.,

$$d(TSx,STx) \le d(Tx,Sx)\,, \quad \textit{for every} \quad x \in X\,. \tag{2.5}$$

Then T and S have a unique common fixed point.

PROOF. Let $x_0 \in X$ be arbitrary. Then by (2.4) $Sx_0 \in T(X)$, which shows that there exists $x_1 \in X$ such that

$$Tx_1 = Sx_0\,.$$

Consider now Sx_1. Since $Sx_1 \in T(X)$, there exists $x_2 \in X$ such that

$$Tx_2 = Sx_1\,.$$

By induction, we construct a sequence $\{x_n\}_{n=0}^{\infty}$ of points in X such that

$$Tx_{n+1} = Sx_n\,, \quad n = 0,1,2,\ldots\,.$$

We shall prove that $\{Tx_n\}_{n=1}^{\infty}$ is a Cauchy sequence.

To this end, consider

$$B(n,k) = \{Tx_j, Sx_j : n \le j \le n+k\}\,;\; b(n,k) = \operatorname{diam}\big(B(n,k)\big)\,;$$

$$B(n) = \{Tx_j, Sx_j : n \le j\}\,;\; b(n) = \operatorname{diam}\big(B(n)\big)\,.$$

It easy to see that $b(n,k) \uparrow b(n)$ as $k \to \infty$ and that $\{b(n)\}_{n=0}^{\infty}$ is a decreasing sequence of positive terms, hence

$$b = \lim_{n\to\infty} b(n)$$

does exist.

To prove that $\{Tx_n\}_{n=0}^{\infty}$ is a Cauchy sequence we must show that $b = 0$.

We claim that

$$b(n,k) \le \varphi\big(b(n-2,k-2)\big)\,, \quad n,k \ge 2\,, \tag{2.6}$$

and discuss the following three cases.

Case 1. $b(n,k) = d(Tx_i, Sx_j)$ with $n \le i, j \le n+k$:
Then $Tx_i = Sx_{i-1}$ and, by (2.2), we get

$$b(n,k) = d(Sx_{i-1}, Sx_j) \le \varphi\big(M(x_{i-1}, x_j)\big) \le \varphi\big(b(n-2,k+2)\big)\,,$$

since φ is monotonically increasing. The remaining cases:

Case 2. $b(n,k) = d(Sx_i, Sx_j)$ with $n \le i, j \le n+k$
and

Case 3. $b(n,k) = d(Tx_i, Tx_j)$ with $n \le i, j \le n+k$
can be easily reduced to Case 1.

Therefore (2.6) is true. Now, if we let $k \to \infty$ in (2.6) and use the continuity of φ we obtain

$$b(n) \le \varphi\big(b(n-2)\big)\,, \quad n \ge 1\,. \tag{2.7}$$

By (ii_φ) and continuity of φ, letting $n \to \infty$ in (2.7) we get

$$b \le \varphi(b)$$

which by Lemma 1 implies $b = 0$.

This shows that both $\{Tx_n\}_{n=1}^{\infty}$ and $\{Sx_n\}_{n=0}^{\infty}$ are Cauchy sequences. Since (X,d) is a complete metric space, we conclude that

$$\lim_{n\to\infty} Tx_n = p \in X\,,$$

and hence $\lim\limits_{n\to\infty} Sx_n = p$, too.

Since T is continuous, we obtain

$$\lim_{n\to\infty} T(Sx_n) = T\big(\lim_{n\to\infty} Sx_n\big) = Tp$$

which, in view of the weak commutativity condition (2.4), yields

$$\begin{aligned} &d(STx_n, Tp) \le d(STx_n, TSx_n) + d(TSx_n, Tp) \le \\ &\le d(Tx_n, Sx_n) + d(TSx_n, Tp) \longrightarrow 0, \quad \text{as} \quad n \to \infty\,. \end{aligned} \tag{2.8}$$

This shows that

$$\lim_{n\to\infty} (ST)(x_n) = Tp\,, \tag{2.9}$$

and therefore, by (2.8) and (2.9), we have

$$\begin{aligned}M(Tx_n,p) = \max\{&d(TTx_n,Tp), d(TTx_n,Sp), d(Tp,Sp),\, d(TTx_n,Sp),\\ &d(Tp,Sx_n)\} \longrightarrow \max\{d(Tp,Tp), d(Tp,Sp), d(Tp,Sp),\, d(Tp,Sp),\\ &d(Tp,Sp)\} = d(Tp,Sp), \quad \text{as} \quad n\to\infty\,.\end{aligned}$$

So by (2.3)

$$d(STx_n,Sp) \le \varphi\left(M(Tx_n,p)\right),$$

which by letting $n\to\infty$, yields

$$d(Tp,Sp) \le \varphi\left(d(Tp,Sp)\right)$$

and which by Lemma 1 implies $d(Tp,Sp)=0$, i.e.,

$$Tp = Sp\,. \tag{2.10}$$

To show that Sp is a common fixed point of S and T it suffices to show that Sp is a fixed point of S. Indeed, by (2.10) and (2.5) it results that

$$TSp = STp = SSp. \tag{2.11}$$

Now, by (2.2), (2.10) and (2.11), we have

$$d(SSp,Sp) \le \varphi\left(M(Sp,p)\right) = \varphi\left(d(SSp,Sp)\right),$$

which again by Lemma 1 implies $SSp=Sp$. From (2.11) it results that Sp is a fixed point of T, too. The uniqueness follows by (2.2). ∎

REMARKS.

1) For $T=1_X$, the identity map, by Theorem 1 we obtain a fixed point theorem similar to Theorem G;

2) For $\varphi(t)=h\cdot t$, $t\in\mathbb{R}_+$, $0<h<1$, from Theorem 1 we obtain a common fixed point theorem that contains Ciric's fixed point theorem as a particular case;

3) Note that if we denote for all $x,y\in X$,

$$D(x,y) = \left(d(x,y),\, d(x,Sx),\, d(y,Sy),\, d(x,Sy),\, d(y,Sx)\right),$$

then

$$\varphi_i\left(D(x,y)\right) \le \varphi\left(M(x,y)\right),$$

for all functions φ_1, φ_2 and φ_3 in Example 2.

This shows that, in the particular case $T=1_X$, Theorem 1 provides extensions of Banach's, Kannan's, Zamfirescu's and Ciric's fixed point theorems.

THEOREM 2. *Let (X,d) be a complete metric space and $T : X \longrightarrow X$ a generalized φ-contraction, i.e., a mapping satisfying*

$$d(Tx,Ty) \le \varphi\left(C(x,y)\right), \quad \textit{for all} \quad x,y \in X,$$

where $\varphi : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ is a continuous comparison function and

$$C(x,y) = \max\{d(x,y),\, d(x,Tx),\, d(y,Ty),\, d(x,Ty),\, d(y,Tx)\}.$$

If T has bounded orbits, then it has a unique fixed point.

PROOF. Take $T = 1_X$ and $S := T$ in Theorem 1. ∎

The continuity of T in Theorem 1 can be weakened to obtain a more general result, similar to Theorem 3 in Rakocevic [11] and Berinde [4]. Actually all the results given in Rakocevic [11] can be similarly adapted for self mappings, but we restrict to the result corresponding to Theorem 3 in [11].

THEOREM 3. *Let (X,d) be a complete metric space and $S,T : X \to X$ two mappings with bounded orbits. Suppose that T^m is continuous for some fixed positive integer m, that S and T satisfy* (2.2), (2.4) *and are commutative, that is,*

$$TSx = STx, \quad \textit{for each} \quad x \in K.$$

Then S and T have a unique common fixed point in K.

PROOF. Let $\{x_n\}$ be constructed as in the proof of Theorem 1.
Hence

$$\lim_{n\longrightarrow\infty} Sx_n = \lim_{n\longrightarrow\infty} Tx_n = p \in X.$$

For each $n \ge 1$,

$$\begin{aligned} d\big(T^m Sx_n, ST^{m-1}p\big) &= d\big(ST^m x_n, ST^{m-1}p\big) \le \\ &\le \varphi\left(M\big(T^m x_n, T^{m-1}p\big)\right) = \\ = \varphi\Big(\max\Big\{d\big(T^m Tx_n, T^m p\big),\, & d\big(T^m Tx_n, T^m Sx_n\big),\, d\big(T^m p, ST^{m-1}p\big), \\ &\big(T^m Tx_n, ST^{m-1}p\big),\, d\big(T^m p, T^m Sx_n\big)\Big\}\Big). \end{aligned}$$

Then by the continuity of T^m and letting $n \to \infty$ we get

$$d\big(T^m p, ST^{m-1}p\big) \le \varphi\left(d\big(T^m p, ST^{m-1}p\big)\right),$$

which by Lemma 1 shows that $T^m p = ST^{m-1}p$.
In order to prove that $T^m p$ is a fixed point of S, i.e.,

$$ST^m p = T^m p,$$

in view of $T^m p = ST^{m-1}p$, it suffices to show that

$$ST^m p = ST^{m-1}p. \tag{2.12}$$

Since

$$M\left(T^m p, T^{m-1}p\right) = \max\Big\{d\left(T^{m+1}p, T^m p\right), d\left(T^{m+1}p, ST^m p\right),$$
$$d\left(T^m p, ST^{m-1}p\right), d\left(T^{m+1}p, ST^{m-1}p\right), d\left(T^m p, ST^m p\right)\Big\},$$

in view of $T^m p = ST^{m-1}p$ and $T^{m+1}p = T\left(ST^{m-1}p\right) = ST^m p$, we obtain

$$M(T^m p, T^{m-1}p) =$$
$$\max\Big\{d\left(ST^m p, ST^{m-1}p\right), 0, 0, d\left(ST^m p, ST^{m-1}p\right), d\left(ST^m p, ST^{m-1}p\right)\Big\}.$$

Now by (2.3) we have

$$d\left(ST^m p, ST^{m-1}p\right) \leq \varphi\left(d\left(T^m p, T^{m-1}p\right)\right) = \varphi\left(d\left(ST^m p, ST^{m-1}p\right)\right)$$

which by Lemma 1 gives

$$d\left(ST^m p, ST^{m-1}p\right) = 0.$$

This proves (2.12) and hence $T^m p$ is a fixed point of S. Now

$$TT^m p = T^{m+1}p = ST^m p = T^m p,$$

which shows that $T^m p$ is a fixed point of T as well.

The uniqueness follows similarly, by the contraction condition (2.2). ∎

REMARKS.

1) Note that the results for nonself mappings in Rakocevic [11] and Berinde [4] are proven in a Banach space setting, while the results in this paper are obtained in the general setting of a complete metric space.

If we impose additional conditions on the comparison function φ, it is possible to obtain an error estimate for the method of successive approximations, like in Theorem G.

2) It is known, see Lemma 4.3.1 in [13] that if T is a generalized strict φ-contraction, i.e., T satisfies (1.4), with

$$t - \varphi(t,t,t,t,t) \to \infty, \text{as } t \to \infty,$$

then T has bounded orbits.

It is however an open question whether or not two mappings S and T satisfying (2.2) or the mapping T in Theorem 2, with φ an arbitrary comparison function, have bounded orbits.

References

[1] V. BERINDE: A fixed point theorem for mappings with contracting orbital diameters, *Bul. Stiint. Univ. Baia Mare* **10** (1994), 29–38.

[2] V. BERINDE: *Generalized Contractions and Applications* (Romanian), Editura Cub Press **22**, Baia Mare, 1997.

[3] V. BERINDE: *Iterative Approximation of Fixed Points*, Editura Efemeride, Baia Mare, 2002.

[4] V. BERINDE: A common fixed point theorem for discontinuous nonself mappings, *Miskolc Math. Notes* (to appear).

[5] F. E. BROWDER: On the convergence of successive approximations for nonlinear functional equations, *Indagat. Math.* **30** (1968), 27–35.

[6] LJ. B. CIRIC: On contraction type mappings, *Math. Balkanika* **1** (1971), 52–57.

[7] LJ. B. CIRIC: A generalization of Banach's contraction principle, *Proc. Am. Math. Soc.* **45** (1974), 267–273.

[8] LJ. B. CIRIC: Quasi contraction nonself mappings on Banach spaces, *Bull. Acad. Serbe Sci. Arts, Cl. Sci. Math. Natur. Sci. Math.* **23** (1998), 25–31.

[9] R. KANNAN: Some results on fixed points, *Bull. Calcutta Math. Soc.* **60** (1968), 71–76.

[10] B. K. LAHIRI and P. DAS: Fixed point of a Ljubomir Ciric's quasi-contraction mapping in a generalized metric space, *Publ. Math. Debrecen* **61(3–4)** (2002), 589–594.

[11] V. RAKOCEVIC: Quasi contraction nonself mappings on Banach spaces and common fixed point theorems, *Publ. Math. Debrecen* **58(3)** (2001), 451–460.

[12] B. E. RHOADES: Contractive definitions and continuity, *Contemporay Math.* **72** (1988), 233–245.

[13] I. A. RUS: *Generalized Contractions and Applications*, Cluj University Press, Cluj-Napoca, 2001.

[14] T. ZAMFIRESCU: Fix point theorems in metric spaces, *Arch. Math. (Basel)* **23** (1972), 292–298.

[15] W. WALTER: Remarks on a paper by F. Browder about contraction, *Nonlinear Anal. TMA* **5** (1981), 21–25.

Vasile Berinde

Department of Mathematics and Computer Science
North University of Baia Mare
Victoriel 76
430122 Baia Mare
Romania
vberinde@ubm.ro

ON THE ASYMPTOTIC BEHAVIOUR OF THE NON-AUTONOMOUS GURTIN–MACCAMY EQUATION

By

J. Z. FARKAS

(Received March 19, 2004)

1. Preliminaries

The Gurtin–MacCamy system, introduced in [5] and its generalizations, including vital rates depending on a finite number of weighted population size functions has been studied by many authors with different methods in different aspects [7], [1], [8], [9], [4]. It describes the dynamics of a single species population living in a closed territory, that is migration is excluded. The only way to leave the population is by death and the newborns of the individuals living in the population form the only after-growth. Thus, if these quantities are balanced the population can survive at a constant level. The measure of the balance is the so called inherent net reproduction number, the expected number of newborns for an individual in his lifetime.

The investigations of the stability of these constant level populations, i.e. stationary age-distributions, by linearization [3] lead to some results containing simple conditions for the net reproduction number [2].

In the present note we are going to investigate the asymptotic behaviour of solutions of the following (linear non-autonomus) model

$$p'_t(a,t)+p'_a(a,t)=-\mu(a,t)p(a,t),\quad 0\le a<m<\infty,\ t\ge 0$$

$$p(0,t)=\int\limits_0^m \beta(a,t)p(a,t)da,\quad t>0, \tag{1.1}$$

with the initial condition $p(a,0)=:p_0(a)$, which satisfies $p_0(0)=\int\limits_0^m \beta(a,0)\cdot p_0(a)da$. Here $p(a,t)$ denotes the density of members of age a at time $t\ge 0$.

This means that the quantity of members between age a and age $a + da$ is $p(a,t)da$ for small da. We assume finite life span denoted by m.

We believe that this linear but non-autonomous system is more useful modelling some population dynamical phenomena for example in the case of time periodic vital rate functions.

The dynamics of the system depends on the vital rates $\beta(a,t), \mu(a,t)$ for which we make the following general assumptions

$$\forall t \in [0,\infty), \forall a \in [0,m] \;\; 0 \leq \beta(a,t) \leq k < \infty, \; \mu(a,t) \geq 0, \tag{1.2}$$

$$\forall t \in [0,\infty) \int_0^m \mu(a,t)da = \infty, \;\; \forall t \in [0,\infty), \; a \in [0,m) \;\; \mu(a,t) < \infty. \tag{1.3}$$

Later we are to make other conditions on the vital rates.

Integrating along the characteristics the model (1.1) can be reduced to a pair of integral equations that corresponds to the cases $t \geq a$ and $a > t$. Since we are investigating here the asymptotic behaviour we consider only the case $t > m \geq a$.

The ODE system of characteristics is

$$\frac{da}{d\tau} = \frac{dt}{d\tau} = 1, \;\; \frac{dp}{d\tau} = -\mu(a,t)p(a,t). \tag{1.4}$$

From (1.4) we have the following formula for $p(a,t)$

$$p(a,t) = \varphi(t-a)e^{-\int_0^a \mu(s,t)ds}, \tag{1.5}$$

where φ is an arbitrary C^1 function which has to satisfy the following equation

$$p(0,t) = \int_0^m \beta(x,t)p(x,t)dx = \varphi(t), \tag{1.6}$$

and from (1.6) we obtain

$$p(a,t) = e^{-\int_0^a \mu(s,t)ds} \int_0^m \beta(x,t-a)p(x,t-a)dx, \tag{1.7}$$

thus

$$p(a,t) = p(0,t-a)\pi(a,t), \;\; with \;\; \pi(a,t) = e^{-\int_0^a \mu(s,t)ds}. \tag{1.8}$$

Here $\pi(a,t)$ denotes the probability for an individual to survive the age a at time t.

Finally recall the net reproduction function

$$R(t) = \int_0^m \beta(a,t) e^{-\int_0^a \mu(s,t)ds} = \int_0^m \beta(a,t)\pi(a,t)da, \tag{1.9}$$

which is the expected number of newborns of an individual at time t.

2. Extinction

In [6] Iannelli et al. studied the global boundedness of solutions of a generalized Gurtin–MacCamy system, where the vital rates depend on a weighted size of the population $S(t) = \int_0^m \gamma(a)p(a,t)da$. Under some natural condition they proved boundedness for the total population quantity $P(t) = \int_0^m p(a,t)da$.

They investigated two cases, first if the fertility function $\beta(a, S(t))$ is bounded by a non-increasing function $\phi(S)$ for which $lim_{S\to\infty}\phi(S) = 0$ holds.

Then they proved boundedness under conditions mainly for the mortality, namely $\beta(a,S) \leq C\gamma(a)$, $\mu(a,S) \geq \mu_0(a) + \omega(S)$, where γ is the weight function, C a positive constant and ω is a non-decreasing function of the weighted population size S, $lim_{S\to\infty}\omega(S) = \infty$.

In this section we are going to apply some of the idea of their proof for the non-autonomous system. That is first we show that under similar conditions for the fertility function the population goes to extinction. Then we consider the connection between the mortality and the fertility functions and establish a result in which a condition for the net reproduction number function $R(t)$ is given.

Consider the following assumptions on the fertility function $\beta(a,t)$

$$\beta(a,t) \leq \phi(t),\ \forall\, t \geq 0, \quad \exists T \geq m \ :\ \phi(T) \leq \frac{1}{2m}, \tag{2.1}$$

where $\phi(t)$ is a positive non-increasing function of $t \in [0,\infty)$.

THEOREM 1. *Let the conditions (2.1) be satisfied. For each non-negative initial age distribution* $p(.,0) \in L^1$ *we have* $\int_0^m p(a,t)da = P(t) \to 0$ *if* $t \to \infty$.

PROOF. From (1.7) we have

$$p(a,t) = p(0,t-a)\pi(a,t),$$

where $\pi(a,t) \leq 1$ for all $a \in [0,m]$, $t \in [m,\infty)$.

For the density of newborns at time t we have

$$p(0,t) = \int_0^m \beta(a,t)p(a,t)da \leq \phi(t)P(t). \tag{2.2}$$

That is we have

$$\int_0^m p(a,t)da = P(t) \leq \int_0^m p(0,t-a)da \leq \int_0^m \phi(t-a)P(t-a)da. \tag{2.3}$$

Now let $I_n := [(n-1)m, nm]$, $(n = 2,3,\ldots)$ and $P_n = max_{t\in I_n} P(t)$.

Then for $t \in I_{n+1}$ and $a \in [0,m]$ we have $(t-a) \in I_n \cup I_{n+1}$ thus, from (2.3) we obtain

$$P_{n+1} \leq max\{P_n, P_{n+1}\} * m * \phi((n-1)m).$$

Let n_* be sufficiently great to have $(n_* - 1)m \geq T$. Then we have

$$P_{n_*+1} \leq \frac{max\{P_{n*}, P_{n*+1}\}}{2}. \tag{2.4}$$

Then it follows that for $n \geq n_*$ we have $P_{n+1} \leq \frac{P_n}{2}$.

That is we have

$$\int_0^m p(a,t)da = P(t) \to 0, \; if \;\; t \to \infty.$$ ∎

As we have mentioned the net reproduction rate $R(t)$ is a key parameter to decide stability of stationary solutions of the autonomous model.

Now suppose that there exists a non-negative $\phi(.)$ function and some constant $\epsilon > 0$ such that

$$\beta(a,t) \leq \phi(t), \;\; \phi(t-a) \leq (1+\epsilon)\beta(a,t), \quad a \in [0,m], t > m. \tag{2.5}$$

Moreover suppose

$$\exists \; T \geq 0 \; s.t. \; R(T) \leq \frac{1}{1+\delta} \; for \; \delta > \epsilon, \tag{2.6}$$

and $R(t)$ is non-increasing.

THEOREM 2. *With the conditions (2.5)–(2.6) for each non-negative initial age distribution* $p(a,0) \in L^1$, $\int_0^m p(a,t)da = P(t) \to 0$ *if* $t \to \infty$.

PROOF. We have again

$$p(a,t) = p(0,t-a)\pi(a,t),\ t \in [m,\infty)$$

and in the same way as in the proof of Th.1 we obtain

$$P(t) \leq \int_0^m \phi(t-a)P(t-a)\pi(a,t)da$$

From the conditions in (2.5) we obtain

$$P(t) \leq \int_0^m (1+\epsilon)\beta(a,t)\pi(a,t)P(t-a)da, \tag{2.7}$$

and with the same $I_n := [(n-1)m, nm], (n = 2,3,\ldots)$ and $P_n := max_{t\in I_n} P(t)$, if $t \in I_{n+1},\ a \in [0,m], (t-a) \in I_n \cup I_{n+1}$ thus we obtain

$$P_{n+1} \leq max\{P_n, P_{n+1}\}(1+\epsilon)\int_0^m \beta(a,t)\pi(a,t)da, \tag{2.8}$$

and because $\int_0^m \beta(a,t)\pi(a,t)da = R(t) \leq \frac{1}{1+\delta}$ for $t \geq T$, for sufficiently large n_* we have for $n \geq n_*$

$$P_{n_*+1} \leq \frac{1+\epsilon}{1+\delta} max\{P_{n*}, P_{n_*+1}\}, \tag{2.9}$$

from where follows that $P_{n+1} \leq P_n \frac{1+\epsilon}{1+\delta} < P_n$, for $n \geq n_*$.

That is $P(t) \to 0$ if $t \to \infty$. ∎

REMARKS. The conditions in Th.2 for the fertility function is quite technical and the condition for $R(t)$ is the essential one. Roughly speaking it means that if there exists some finite $T \geq 0$ such that $R(t)$ is bounded by some $\frac{1}{1+\delta} \leq 1$ for $t \geq T$ then the population goes to extinction. In other words if the expected number of newborns at time t is less than 1 for $t \geq T$ then the total population quantity tends to zero, of course.

3. Sharper upper bound

In the previous section we determined conditions for the vital rates which guarantees the extinction of the population. One may expect that if there exists some finite T such that for $t \geq T$ the inherent net reproduction number $R(t)$ is lower than 1 in other words the number of per capita offspring is below 1 then the total population quantity decreases and the population goes to extinction.

In this section we are going to formulate some sharper "upper bound" for the total population quantity, which is also in close relation with the net reproduction rate R as we will see.

Integrating both sides of the equation in (1.1) from 0 to m we have

$$\dot{P}(t) = -\int_0^m \mu(a,t)p(a,t)da - \int_0^m p_a'(a,t)da = p(0,t) - \int_0^m \mu(a,t)p(a,t)da =$$

$$= \int_0^m \beta(a,t)p(a,t)da - \int_0^m \mu(a,t)p(a,t)da. \tag{3.1}$$

The solution of the ODE (3.1) obtained easily

$$P(t) = \int_0^t \int_0^m (p(a,s)\beta(a,s) - p(a,s)\mu(a,s))dads + P(0), \tag{3.3}$$

and we have

$$\lim_{t\to\infty} P(t) = \int_0^\infty \int_0^m (p(a,s)\beta(a,s) - p(a,s)\mu(a,s))dads + P(0). \tag{3.4}$$

Thus the question is when does the function

$$F(s) = \int_0^m (p(a,s)\beta(a,s) - p(a,s)\mu(a,s))da \tag{3.5}$$

belong to $L^1_{[0,\infty)}$.

From (1.4) we have $p(a,s) = p(0,s-a)\pi(a,s)$ for $s \geq a$, that is we have

$$F(s) = \int_0^m p(0,s-a)(\beta(a,s)\pi(a,s) - \mu(a,s)\pi(a,s))da \tag{3.6}$$

for $s \geq m$, and clearly $\int_0^m F(s)ds < \infty$ holds.

If the density of newborns $p(0,t)$ is finite for every t then there exists a function $C(s)$ which is also bounded, such that $p(0,s-a) < C(s)p(0,s)$ for every $a \in [0,m]$.

That is we have

$$F(s) \leq p(0,s)C(s)| \int_0^m \beta(a,s)\pi(a,s)da - \int_0^m \mu(a,s)\pi(a,s)da|. \tag{3.7}$$

Now observe that $\int_0^m \beta(a,s)\pi(a,s)da = R(s)$ by definition and $\int_0^m \mu(a,s) \cdot \pi(a,s)da = 1$ because $\mu(a,s)\pi(a,s)da$ is the probability for an individual to survive the age a and then die in $[a,a+da]$.

That is we have

$$\lim_{t\to\infty} P(t) \leq \int_0^\infty p(0,s)C(s)|R(s)-1|ds + P(0). \tag{3.8}$$

Note that if the net reproduction number $R(s) < M$ is bounded by some $M < \infty$ for every s, then the density of newborns $p(0,s)$ and the function $C(s)$ is bounded for every s, too. So if for example $(R(s)-1) \leq \frac{1}{s^{1+\alpha}}$ for some $\alpha > 0$, then the improper integral in (2.7) is convergent.

Example. Consider the following special vital rate functions with maximal life span $m = 100$

$$\beta(a,t) = b(a)f(t) = \frac{a^4}{C}(100-a)^2 1.11^{-a}(1+\frac{1}{t^2+1}), \quad \mu(a) = \frac{1}{100-a},$$

with $C = \int_0^{100} a^4(100-a)^2 1.11^{-a}\pi(a)da \sim 0,4045064485 * 10^{10}$.

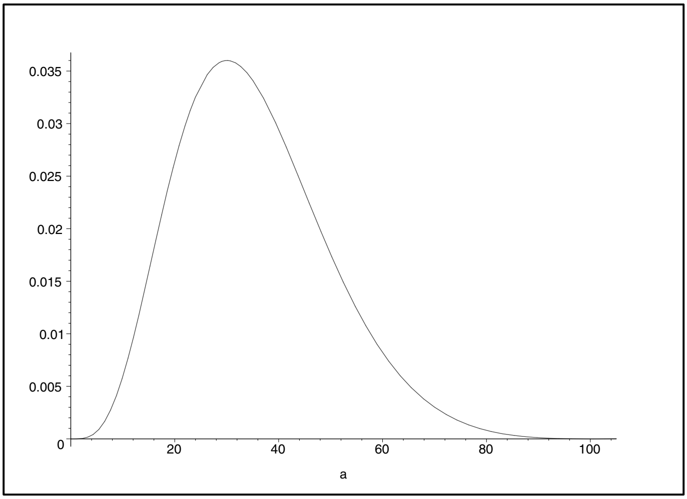


FIG. 1. $b(a) = \frac{a^4}{C}(100 - a)^2 1.11^{-a}$

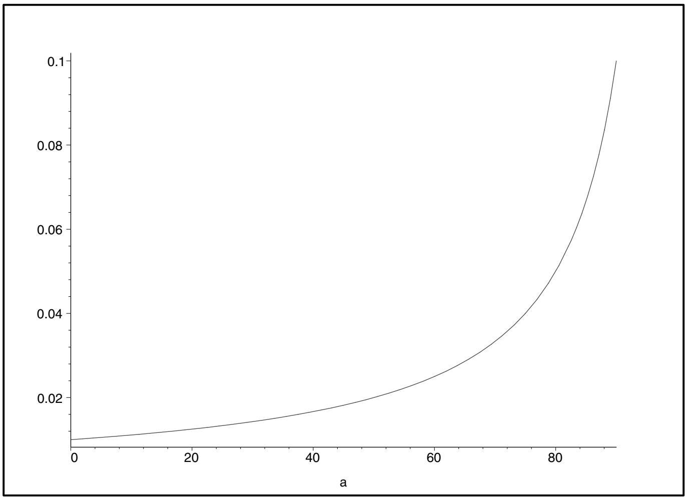


FIG. 2. $\mu(a) = \frac{1}{100-a}$

It is easy to show that these functions satisfy the conditions (1.2)–(1.3).

With $\pi(a) = 1 - \frac{a}{100}$

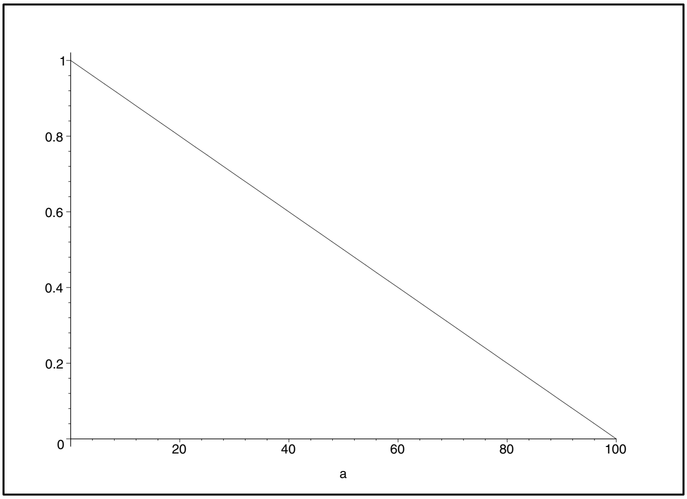


FIG. 3. $\pi(a) = 1 - \frac{a}{100}$

we arrive at

$$R(t) = \int_0^{100} \frac{a^4}{C}(100 - a)^2 1.11^{-a}(1 + \frac{1}{t^2+1})(1 - \frac{a}{100})da = 1 + \frac{1}{1+t^2}$$

Thus $R(t) \geq 1$ for $t \geq 0$ and $R(t) \to 1$ in a sufficient order.

From (3.8)

$$\lim_{t\to\infty} P(t) \leq \int_0^{\infty} \frac{p(0,s)C(s)}{1+s^2}ds + P(0), \tag{3.9}$$

that is for any given initial age distribution $p_0(a)$ which satisfies the compatibility condition

$$p_0(0) = \int_0^{100} 2p_0(a)\frac{a^4}{C}(100 - a)^2 1.11^{-a}da$$

the solution $p(a,t) \to p_*(a)$ if $t \to \infty$ with some non-trivial age distribution $p_*(a)$ in the following L^1 norm:

$$||p(.,t)|| := \int_0^{m} |p(a,t)|da. \tag{3.10}$$

REMARKS. The example above is a very special one but shows the essential role of the net reproduction function $R(t)$. Thus the general problem namely the formulation of necessary or sufficient conditions for the convergence to a non-trivial age distribution seems to be still open. As the probably much more interesting case of time periodic vital rates on which we are working.

ACKNOWLEDGEMENTS. The author is thankful to Prof. Miklós Farkas for reading the manuscript and making valuable suggestions. Research partially supported by the Hungarian National Foundation for Scientific Research.

References

[1] J. M. CUSHING: *An Introduction to Structured Population Dynamics*, SIAM, Philadelphia, PA, 1998.

[2] J. Z. FARKAS: Stability conditions for the non-linear McKendrick equations, to appear in *Applied Mathematics and Computations.*

[3] M. FARKAS: On the stability of stationary age distributions, *Applied Mathematics and Computation* **131(1)** (2002), 107–123.

[4] M. FARKAS: *Dynamical Models in Biology*, San Diego, Academic Press, 2001.

[5] M. E. GURTIN and R. C. MACCAMY: Non-linear age-dependent populations dynamics, *Arch. Rat. Mech. Anal.* **54** (1974), 281–300.

[6] M. IANNELLI, M.-Y. KIM, E.-J. PARK and A. PUGLIESE: Global boundedness of the solutions to a Gurtin–MacCamy system, *NoDEA* **9** (2002), 197–216.

[7] M. IANNELLI: *Mathematical Theory of Age-Structured Population Dynamics*, Applied Mathematics Monographs, Comitato Nazionale per le Scienze Matematiche, (C.N.R.), Vol. **7**, Giardini, Pisa, 1995.

[8] J. A. J. METZ and O. DIEKMANN: *The Dynamics of Physiologically Structured Populations*, Lecture Notes in Biomath. **68**, Springer-Verlag, Berlin, 1986.

[9] G. WEBB: *Theory of Nonlinear Age-Dependent Population Dynamics*, Marcel Dekker, New York, 1985.

J. Z. Farkas

Budapest University of Technology
Department of Differential Equations
Budapest, H-1521
Hungary
farkas@math.bme.hu

A UNIFIED THEORY OF $T_{\frac{1}{2}}$-SPACES

By

M. CALDAS, D. N. GEORGIOU, S. JAFARI and T. NOIRI

(Received May 7, 2004)

1. Introduction

In 1970, Levine [18] introduced the notion of $T_{\frac{1}{2}}$ spaces which properly lie between T_1-spaces and T_0-spaces. Dunham [12] obtained the following characterization of $T_{\frac{1}{2}}$-spaces: a topological space (X,τ) is $T_{\frac{1}{2}}$ if and only if each singleton of X is open or closed. Moreover, Arenas et al. [4] showed that a topological space (X,τ) is $T_{\frac{1}{2}}$ if and only if every subset of X is λ-closed. In 1987, semi-$T_{\frac{1}{2}}$ spaces are introduced by Bhattacharyya and Lahiri [6]. Sundaram et al. [28] showed that a topological space (X,τ) is semi-$T_{\frac{1}{2}}$ if and only if each singleton of X is semi-open or semi-closed. Recently, Caldas et al. ([7], [8], [9]) have defined and investigated the notions of (Λ,θ)-closed, (Λ,δ)-closed and (Λ,α)-closed sets in topological spaces. The characterization of Arenas et al. [4] motivated us to obtain unified characterizations of certain weak separation axioms containing $T_{\frac{1}{2}}$ and semi-$T_{\frac{1}{2}}$.

In this paper, we introduce the notions called m-structures which are weaker than topological structures. Using the m-structures, we investigate a unified theory of weak separation axioms containing $T_{\frac{1}{2}}$-spaces and semi-$T_{\frac{1}{2}}$-spaces.

2. Preliminaries

In what follows (X,τ) and (Y,σ) (or X and Y) denote topological spaces. Let A be a subset X. We denote the interior and the closure of a set A by Int(A) and Cl(A), respectively. A point $x \in X$ is called a θ-cluster point of A if $A \cap Cl(U) \neq \emptyset$ for every open set U of X containing x. The set of all θ-cluster points of A is called the θ-closure of A, denoted by $Cl_\theta(A)$. A subset A is called θ-closed [29] if $A = Cl_\theta(A)$. The complement of a θ-closed set is said to be θ-open. We denote the collection of all θ-open sets of (X,τ) by τ_θ.

A point $x \in X$ is called the δ-cluster point of A if $A \cap Int(Cl(U)) \neq \emptyset$ for every open set U of X containing x. The set of all δ-cluster points of A is called the δ-closure of A, denoted by $Cl_\delta(A)$. A subset A is called δ-closed [29] if $A = Cl_\delta(A)$. The complement of a δ-closed set is said to be δ-open. We denote the collection of all δ-open sets by τ_δ. The set $\{x \in X \mid x \in U \subset Int(Cl(U)) \subset A\}$ for some open set U of X is called the δ-interior of A and is denoted by $Int_\delta(A)$.

Definition 1. Let (X,τ) be a topological space. A subset A of X is said to be
(1) semi-open [17] if $A \subset Cl(Int(A))$,
(2) preopen [23] if $A \subset Int(Cl(A))$,
(3) α-open [25] if $A \subset Int(Cl(Int(A)))$,
(4) β-open [1] or semi-preopen [3] if $A \subset Cl(Int(Cl(A)))$.

The family of all semi-open (resp. preopen, α-open, β-open, semi-preopen) sets in X is denoted by $SO(X)$ (resp. $PO(X)$, $\alpha(X)$, $\beta(X)$, $SPO(X)$).

Definition 2. The complement of a semi-open (resp. preopen, α-open, β-open, semi-preopen) set is said to be semi-closed [10] (resp. preclosed [23], α-closed [24], β-closed [1], semi-preclosed [3]).

Definition 3. The intersection of all semi-closed (resp. preclosed, α-closed, β-closed) sets of X containing A is called the semi-closure [10] (resp. preclosure [13], α-closure [24], β-closure [2] or semi-preclosure [3]) of A and is denoted by $sCl(A)$ (resp. $pCl(A)$, $\alpha Cl(A)$, $\beta Cl(A)$ or $spCl(A)$).

Definition 4. The union of all semi-open (resp. preopen, α-open, β-open) sets of X contained in A is called the semi-interior (resp. preinterior,

α-interior, β-interior or semi-preinterior) of A and is denoted by $sInt(A)$ (resp. $pInt(A)$, $\alpha Int(A)$, $\beta Int(A)$ or $spInt(A)$).

DEFINITION 5. A subset A of a topological space (X,τ) is said to be
(1) g-closed [18] if $Cl(A) \subset U$ whenever $A \subset U$ and U is open in (X,τ),
(2) semi g-closed [6] if $sCl(A) \subset U$ whenever $A \subset U$ and U is semi-open in (X,τ).

DEFINITION 6. A topological space (X,τ) is said to be
(1) $T_{\frac{1}{2}}$-space [18] if every g-closed set is closed in (X,τ),
(2) semi-$T_{\frac{1}{2}}$ space [6] if every semi g-closed set is semi-closed in (X,τ).

LEMMA 2.1 (DUNHAM [12]; SUNDARAM ET AL. [28]). *Let (X,τ) be a topological space. Then*
(1) (X,τ) is $T_{\frac{1}{2}}$ if and only if every singleton of X is open or closed,
(2) (X,τ) is semi-$T_{\frac{1}{2}}$ if and only if every singleton of X is semi-open or semi-closed.

DEFINITION 7. A subset A of a topological space (X,τ) is called
(1) a Λ-set [4] if $A = \cap\{U \mid A \subset U, U \in \tau\}$,
(2) a semi-Λ-set [11] if $A = \cap\{U \mid A \subset U, U \in SO(X,\tau)\}$.

DEFINITION 8. A subset A of a topological space (X,τ) is said to be
(1) λ-closed [4] if $A = T \cap C$, where T is a Λ-set and C is closed,
(2) semi-λ-closed [11] if $A = T \cap C$, where T is a semi-Λ-set and C is semi-closed.

3. m-spaces

DEFINITION 9. A subfamily m of the power set $P(X)$ of a nonempty set X is called an m-structure on X if m satisfies the following:
(1) $\emptyset \in m$ and $X \in m$,
(2) $\cup_{\alpha\in\triangle} \in A_\alpha \in m$ whenever $A_\alpha \in m$ for each $\alpha \in \triangle$.

We call the pair (X,m) an m-space. Each member of m is said to be m-open and the complement of an m-open set is said to be m-closed.

REMARK 3.1. It should be noted that condition (2) in Definition 9 of m-structure is called property (B) by Maki in [22]. In this paper, we always assume the property (B) on m-structures.

REMARK 3.2. Let (X,τ) be a topological space. Then the families τ_θ, τ_δ, τ, $SO(X,\tau)$, $PO(X,\tau)$, $\alpha(X)$, $\beta(X)$ are all m-structures on X. It is well-known that τ_θ, τ_δ, $\alpha(X)$ are topologies for X.

DEFINITION 10. Let X be a nonempty set and m an m-structure on X. For a subset A of X, the m-closure of A and the m-interior of A are defined in [22] as follows:
(1) $m_X\text{-}Cl(A) = \cap\{F \mid A \subset F, X\backslash F \in m\}$,
(2) $m_X\text{-}Int(A) = \cup\{U \mid U \subset A, U \in m\}$.

In this paper, we denote $m_X\text{-}Cl(A)$ (resp. $m_X\text{-}Int(A)$) by $Cl_m(A)$ (resp. $Int_m(A)$).

REMARK 3.3. Let (X,τ) be a topological space and A a subset of X. If $m=\tau$ (resp. $SO(X)$, $PO(X)$, $\alpha(X)$, $\beta(X)$, τ_θ, τ_δ), then we have
(1) $Cl_m(A) = Cl(A)$ (resp. $sCl(A)$, $pCl(A)$, $\alpha Cl(A)$, $\beta Cl(A)$, $Cl_\theta(A)$, $Cl_\delta(A)$),
(2) $Int_m(A) = Int(A)$ (resp. $sInt(A)$, $pInt(A)$, $\alpha Int(A)$, $\beta Int(A)$, $Int_\theta(A)$, $Int_\delta(A)$).

LEMMA 3.4 (MAKI [22]). *Let m be an m-structure on a nonempty set X. For subsets A and B of X, the following properties hold:*
(1) $Cl_m(X\backslash A) = X\backslash Int_m(A)$ and $Int_m(X\backslash A) = X\backslash Cl_m(A)$,
(2) $Cl_m(\emptyset) = \emptyset$, $Cl_m(X) = X$, $Int_m(\emptyset) = \emptyset$ and $Int_m(X) = X$,
(3) If $A \subset B$, then $Cl_m(A) \subset Cl_m(B)$ and $Int_m(A) \subset Int_m(B)$,
(4) $A \subset Cl_m(A)$ and $Int_m(A) \subset A$,
(5) $Cl_m(Cl_m(A)) = Cl_m(A)$ and $Int_m(Int_m(A)) = Int_m(A)$.

LEMMA 3.5 (POPA AND NOIRI [27]). *Let m be an m-structure on a nonempty set X. Then $x \in Cl_m(A)$ if and only if $U \cap A \neq \emptyset$ for every $U \in m$ containing x.*

LEMMA 3.6. *Let m be an m-structure on a nonempty set X. Then for a subset A of X the following properties hold:*
(1) $A \in m$ if and only if $A = Int_m(A)$,
(2) A is m-closed if and only if $A = Cl_m(A)$,
(3) $Cl_m(A)$ is m-closed and $Int_m(A)$ is m-open.

PROOF. This is an immediate consequence of Lemma 3.4 and Lemma 3.5. ∎

DEFINITION 11. Let A be a subset of an m-space (X,m).
(1) A subset $\Lambda_m(A)$ is defined as follows: $\Lambda_m(A) = \cap\{O \in m \mid A \subset O\}$.
(2) The subset A is called a Λ_m-set if $A = \Lambda_m(A)$.
(3) The subset A is said to be (Λ,m)-closed if $A = T \cap C$, where T is a Λ_m-set and C is a m-closed set.

REMARK 3.7. Let (X,τ) be a topological space. If we set $m = \tau$ (resp. $SO(X)$, τ_θ, τ_δ, $\alpha(X)$), then the (Λ,m)-closed set is a λ-closed (resp. semi-λ-closed, (Λ,θ)-closed [8], (Λ,α)-closed [9], (Λ,δ)-closed [14]) set.

4. m-T_0 spaces

DEFINITION 12. An m-space (X,m) is said to be
(1) m-T_0 if for x, $y \in X$ such that $x \neq y$ there exists an m-open set U of X containing x but not y or an m-open set V of X containing y but not x,
(2) m-T_1 if for distinct points x, $y \in X$, there exist an m-open set containing x but not y and an m-open set containing y but not x,
(3) m-T_2 if for x, $y \in X$ such that $x \neq y$ there exist disjoint m-open sets U and V of X such that $x \in U$ and $y \in V$.

REMARK 4.1. Let (X,τ) be a topological space. Let us put $m = \tau$, $SO(X)$, $PO(X)$, τ_θ, τ_δ, $\alpha(X)$, $\beta(X)$, then we obtain the following table. In the table, each notion is defined in the literature shown in the square brackets.

m	τ	τ_θ	τ_δ	$\alpha(X)$	$SO(X)$	$PO(X)$	$\beta(X)$
m-T_2	T_2	θ-T_2 [5]	δ-T_2 [16]	α-T_2 [20]	semi-T_2 [19]	pre-T_2 [26]	β-T_2 [27]
m-T_1	T_1	θ-T_1 [15]	δ-T_1 [16]	α-T_1 [9]	semi-T_1 [19]	pre-T_1 [26]	β-T_1 [27]
m-T_0	T_0	θ-T_2 [7]	δ-T_0 [16]	α-T_0 [9]	semi-T_0 [19]	pre-T_0 [26]	β-T_0 [27]

LEMMA 4.2. *For an m-space (X,m), the following properties hold:*
(1) Every m-T_2 m-space is m-T_1 and every m-T_1 m-space is m-T_0,
(2) (X,m) is m-T_1 if and only if for each $x \in X$, the singleton $\{x\}$ is m-closed.

PROOF. The proof is obvious. ∎

THEOREM 4.3. *For an m-space (X,m), the following properties are equivalent:*
(1) (X,m) is m-T_0;
(2) For each pair of distinct points x and y of X, there exists a subset A of X such that $x \in A$, $y \notin A$ and A is m-open or m-closed;
(3) For each $x \in X$, the singleton $\{x\}$ is (Λ,m)-closed.

PROOF. (1) ⇒ (2): Let $x \neq y$. In case which there exists an m-open set U of X such that $x \in U$ and $y \notin U$, we put $A = U$. In case which there exists an m-open set V of X such that $x \notin V$ and $y \in V$, we put $A = X \setminus V$. Then A is the desired set.
(2) ⇒ (3): Let $x \in X$. By (2), for each point $y \neq x$ there exists a subset A_y of X such that $x \in A_y$, $y \notin A_y$ and A_y is m-open or m-closed. Let $T = \cap\{A_y \mid y \in X\setminus\{x\}, A_y$ is m-open$\}$ and $C = \cap\{A_y \mid y \in X\setminus\{x\}, A_y$ is m-closed$\}$. Then we obtain that T is a Λ_m-set, C is an m-closed set and $\{x\} = T \cap C$. Therefore, $\{x\}$ is (Λ,m)-closed.
(3) ⇒ (1): Let x and y be distinct points of X. By (3), $\{x\} = T \cap C$, where T is a Λ_m-set and C is m-closed. If C does not contain y, then $X\setminus C$ is an m-open set containing y but not x. If C contains y, then $y \notin T$. Since T is a Λ_m-set, there exists an m-open set U containing x such that $y \notin U$. Therefore, (X,m) is m-T_0. ∎

5. m-$T_{\frac{1}{2}}$ spaces

DEFINITION 13. An m-space (X,m) is said to be m-$T_{\frac{1}{2}}$ if every singleton of X is m-open or m-closed.

REMARK 5.1. Let (X,τ) be a topological space. Let $m = \tau$ (resp. $SO(X)$) then m-$T_{\frac{1}{2}} = T_{\frac{1}{2}}$ (resp. semi-$T_{\frac{1}{2}}$). By setting $m = \tau_\theta$, τ_δ or $\alpha(X)$, we can define θ-$T_{\frac{1}{2}}$, δ-$T_{\frac{1}{2}}$ or α-$T_{\frac{1}{2}}$ and obtain the characterizations by the following result.

THEOREM 5.2. *Let (X,m) be an m-space. Then the following properties are equivalent:*
(1) Every subset of X is (Λ,m)-closed;
(2) (X,m) is m-$T_{\frac{1}{2}}$.

PROOF. (1) $\Rightarrow$ (2): Let $x \in X$ and let us suppose that $\{x\}$ is not m-open. We prove that the singleton $\{x\}$ is m-closed. Let $A = X\backslash\{x\}$. Since $\{x\}$ is not m-open, the subset A is not m-closed. By assumption, the subset A is (Λ, m)-closed. Thus the subset A is a Λ_m-set. Since $A = X\backslash\{x\}$ and the set X is the only m-open set of the m-space (X, m) such that $A \subseteq X$, we have that A is m-open. Hence $\{x\}$ is m-closed and therefore the m-space (X, m) is an m-$T_{\frac{1}{2}}$ space.

(2) $\Rightarrow$ (1): Let A be any subset of the m-space (X, m). We prove that the subset A is (Λ, m)-closed, that is $A = T \cap C$, where T is a Λ_m-set and C is m-closed. Let $S = \{x \mid x \in X\backslash A \text{ and } \{x\} \text{ is } m\text{-open}\}$. Then, the set $C = \cap\{X\backslash\{x\} \mid x \in S\}$ is m-closed and $A \subseteq C$. Also, for the subset $T = \cap\{X\backslash\{x\} \mid x \in X\backslash(A \cup S)\}$ of X we have: $A \subseteq T$ and $\Lambda_m(T) = T$, that is T is a Λ_m-set. Also, it is clear that $T \cap C \subseteq A$. Thus $A = T \cap C$ and therefore the subset A is (Λ, m)-closed.

THEOREM 5.3. *For an m-space (X, m), the following properties hold:*
(1) (X, m) is m-T_1, then it is m-$T_{\frac{1}{2}}$,
(2) (X, m) is m-$T_{\frac{1}{2}}$, then it is m-T_0.

PROOF. (1) The proof is obvious from Lemma 4.2.
(2) Let x and y be two distinct elements of X. Since the m-space (X, m) is m-$T_{\frac{1}{2}}$, we have that $\{x\}$ is m-open or m-closed. Suppose that $\{x\}$ is m-open. Then the singleton $\{x\}$ is an m-open set such that $x \in \{x\}$ and $y \notin \{x\}$. Also, if $\{x\}$ is m-closed, then $X\backslash\{x\}$ is m-open such that $y \in X\backslash\{x\}$ and $x \notin X\backslash\{x\}$. Thus, in the above two cases, there exists an m-open set U of X such that $x \in U$ and $y \notin U$ or $x \notin U$ and $y \in U$. Thus the m-space (X, m) is m-T_0. ∎

By Theorem 5.3, we observe that the class of m-$T_{\frac{1}{2}}$ spaces is placed between the classes of m-T_0 and m-T_1 spaces.

DEFINITION 14. An m-space (X, m) is said to be
(1) m-R_0 if every m-open set contains the m-closure of each of its singletons,
(2) m-R_1 if for x, y in X with $Cl_m(\{x\}) \neq Cl_m(\{y\})$, there exist disjoint m-open sets U and V such that $Cl_m(\{x\}) \subset U$ and $Cl_m(\{y\}) \subset V$.

THEOREM 5.4. *For an m-R_0 m-space (X, m), the following properties are equivalent:*
(1) (X, τ) is m-T_0;

(2) (X,τ) *is* m-$T_{\frac{1}{2}}$*;*

(3) (X,τ) *is* m-T_1*.*

PROOF. It suffices to prove only (1) ⇒ (3): Let $x \neq y$ and since (X,m) is m-T_0, we may assume that $x \in U \subset X \backslash \{y\}$ for some $U \in m$. Then $x \in X \backslash Cl_m(\{y\})$ and $X \backslash Cl_m(\{y\})$ is m-open. Since (X,m) is m-R_0, we have $Cl_m(\{x\}) \subset X \backslash Cl_m(\{y\}) \subset X \backslash \{y\}$ and hence $y \notin Cl_m(\{x\})$. There exists $V \in m$ such that $y \in V \subset X \backslash \{x\}$ and (X,m) is an m-T_1 space. ▮

THEOREM 5.5. *Let* (X,m) *be an* m*-space. Then* (X,m) *is* m-$T_{\frac{1}{2}}$ *and* m-R_1 *if and only if it is* m-T_2*.*

PROOF. Necessity. Let x and y be two distinct points of X. Since (X,m) is m-$T_{\frac{1}{2}}$, we consider the following cases:

Case 1. $\{x\}$ and $\{y\}$ are m-closed: It follows from assumptions that there exist disjoint m-open sets U and V such that $\{x\} = Cl_m(\{x\}) \subset U$ and $\{y\} = Cl_m(\{y\}) \subset V$.

Case 2. $\{x\}$ is m-closed and $\{y\}$ is m-open: Let $U = \{y\}$. If $z \notin U$, then since $y \notin Cl_m(\{z\})$, $Cl_m(\{y\}) \neq Cl_m(\{z\})$. Since (X,m) is an m-R_1 space, there exists an m-open set V such that $Cl_m(\{z\}) \subset V$ and $y \notin V$, which implies $z \notin Cl_m(\{y\})$. Thus $Cl_m(\{y\}) \subset U = \{y\}$ and so $\{y\}$ is m-closed. Hence this case reduces to Case 1.

Case 3. $\{x\}$ is m-open and $\{y\}$ is m-closed: Also this is reduced to Case 1.

Case 4. $\{x\}$ and $\{y\}$ are m-open: Then m-open sets $\{x\}$, $\{y\}$ are required. Therefore, (X,m) is m-T_2.

Sufficiency. We recall that every m-T_2 m-space is m-T_1 and every m-T_1 m-space is m-$T_{\frac{1}{2}}$. Let x and y be points such that $Cl_m(\{x\}) \neq Cl_m(\{y\})$. Then, since (X,m) is m-T_2, there exist disjoint m-open sets U and V such that $Cl_m(\{x\}) = \{x\} \subset U$ and $Cl_m(\{y\}) = \{y\} \subset V$. Therefore (X,m) is m-R_1. ▮

6. m-$T_{\frac{1}{4}}$ spaces

DEFINITION 15. An m-space (X,m) is said to be m-$T_{\frac{1}{4}}$ if for every finite subset F of X and $x \notin F$ there exists a set F_x such that (1) $F \subseteq F_x$, (2) F_x is either m-open or m-closed, and (3) $F_x \cap \{x\} = \emptyset$.

THEOREM 6.1. *Every m-$T_{\frac{1}{2}}$ m-space (X, m) is m-$T_{\frac{1}{4}}$.*

PROOF. Let (X, m) be an m-$T_{\frac{1}{2}}$ m-space. We prove that the space (X, m) is m-$T_{\frac{1}{4}}$. Let F be a finite subset of X and $x \notin F$. Since the m-space (X, m) is m-$T_{\frac{1}{2}}$, we have that the singleton $\{x\}$ is m-open or m-closed. Let us suppose that $\{x\}$ is m-open. Setting $F_x = X \backslash \{x\}$ we have F_x is m-closed, $F \subseteq F_x$ and $F \cap \{x\} = \emptyset$. Similarly if $\{x\}$ is m-closed, then we have F_x is m-open, $F \subseteq F_x$ and $F \cap \{x\} = \emptyset$. Thus the m-space (X, m) is m-$T_{\frac{1}{4}}$. ∎

THEOREM 6.2. *Every m-$T_{\frac{1}{4}}$ m-space (X, m) is m-T_0.*

PROOF. This follows immediately from Theorem 4.3. ∎

THEOREM 6.3. *For an m-space (X, m) the following properties are equivalent:*
(1) Every finite subset of X is (Λ, m)-closed;
(2) (X, m) is m-$T_{\frac{1}{4}}$.

PROOF. (1) ⇒ (2): Let F be a finite subset of X and $x \notin F$. Since F is (Λ, m)-closed, we have that $F = T \cap C$, where T is a Λ_m-set and C is an m-closed subset of X. If $x \notin C$, then by setting $F_x = C$ we have: (1) $F \subseteq F_x$, (2) $x \notin F_x$ and (3) F_x is m-closed. Also, if $x \in C$, then $x \notin T$ and so for some m-open set U of X such that $F \subseteq U$ we have: (1) $F \subseteq U$, (2) $x \notin U$ and (3) U is m-open. Clearly, in the second case $F_x = U$. Thus the space (X, m) is m-$T_{\frac{1}{4}}$.

(2) ⇒ (1): Let F be a finite subset of the space X. We prove that the subset F is (Λ, m)-closed, that is $F = T \cap C$, where T is a Λ_m-set and C is m-closed. Since the m-space (X, m) is m-$T_{\frac{1}{4}}$, for every point $x \notin F$ there exists a subset F_x of X such that (1) $F \subseteq F_x$, (2) $F_x \cap \{x\} = \emptyset$ and (3) F_x is either m-open or m-closed. We set $T = \cap\{F_x \mid x \notin F \text{ and } F_x \text{ is } m\text{-open}\}$ and $C = \cap\{F_x \mid x \notin F \text{ and } F_x \text{ is } m\text{-closed}\}$. For the subsets T and C we have (1) C is m-closed, (2) T is a Λ_m-set and (3) $F \subseteq T \cap C$. Finally, we prove that $T \cap C \subseteq F$. Indeed, let $x \in T \cap C$ and $x \notin F$. Then, there exists an m-open or m-closed subset F_x of X such that $F \subseteq F_x$ and $F_x \cap \{x\} = \emptyset$. Let us suppose that F_x is m-open, then $x \notin T$ which is a contradiction. Similarly if F_x is m-closed, then $x \notin C$ which is also a contradiction. Thus $x \in F$ and therefore $T \cap C \subseteq F$. ∎

References

[1] M. E. Abd El-Monsef, S. N. El-Deeb and R. A. Mahmoud: β-open sets and β-continuous mappings, *Bull. Fac. Sci. Assiut Univ.* **12** (1983), 77–90.

[2] M. E. Abd El-Monsef, R. A. Mahmoud and E. R. Lashin: β-closure and β-interior, *J. Fac. Ed. Ain Shams Univ.* **10** (1986), 235–245.

[3] D. Andrijević: Semi-preopen sets, *Mat. Vesnik* **38** (1986), 24–32.

[4] F. G. Arenas, J. Dontchev and M. Ganster: On λ-sets and dual of generalized continuity, *Questions Answers Gen. Topology* **15** (1997), 3–13.

[5] S. Bandyopadhyayi: Some Problems Concerning Covering Properties and Function Spaces, Ph.D. Thesis, University of Calcutta, 1996.

[6] P. Bhattacharyya and B. K. Lahiri: Semi-generalized closed sets in topology, *Indian J. Math.* **29** (1987), 375–382.

[7] M. Caldas, S. Jafari and T. Noiri: Weak separation axioms via Veličko's θ-open set and θ-closure operator, *Atas Sem. Bras. Anal.* **56** (2002), 657–664.

[8] M. Caldas, D. N. Georgiou, S. Jafari and T. Noiri: On (Λ,θ)-closed sets (submitted).

[9] M. Caldas, D. N. Georgiou and S. Jafari: Study of (Λ,α)-closed sets and the related notions in topological spaces (submitted).

[10] S. G. Crossley and S. K. Hildbrand: Semi-closure, *Texas J. Sci.* **22** (1971), 99–112.

[11] J. Dontchev and H. Maki: On sg-closed sets and semi-λ-closed sets, *Questions Answers Gen. Topology* **15** (1997), 259–266.

[12] W. Dunham: $T_{\frac{1}{2}}$-spaces, *Kyungpook Math. J.* **17** (1977), 161–169.

[13] S. N. El-Deeb, I. A. Hasanein, A. S. Mashhour and T. Noiri: On p-regular spaces, *Bull. Math. Soc. Sci. Math. R. S. Roumanie* **27(75)** (1983), 311–315.

[14] D. N. Georgiou, S. Jafari and T. Noiri: Properties of (Λ,δ)-closed sets in topological spaces, Bolletino della unione matematica italiana (article in press).

[15] S. Jafari: Some properties of quasi θ-continuous functions, *Far East J. Math. Sci.* **6(5)** (1998), 689–696.

[16] R. C. Jain: The role of regularly open sets in general topological spaces, Ph.D. Thesis, Meerut Univ. Inst. Advance Stud. Meerut, India 1980.

[17] N. Levine: Semi-open sets and semi-continuity in topological spaces, *Amer. Math. Monthly* **70** (1963), 36–41.

[18] N. Levine: Generalized closed sets in topology, *Rend. Circ. Mat. Palermo (2)* **19** (1970), 89–96.

[19] S. N. Maheshwari and R. Prasad: Some new separation axioms, *Ann. Soc. Sci. Bruxelles* **89** (1975), 395–402.

[20] S. N. MAHESHWARI and S. S. THAKUR: On α-irresolute mappings, *Tamkang J. Math.* **11** (1980), 209–214.

[21] R. A. MAHMOUD and M. E. ABD EL-MONSEF: β-irresolute and β-topological invariant, *Proc. Pakistan Acad. Sci.* **27** (1990), 285–296.

[22] H. MAKI: On generalizing semi-open and preopen sets, Meeting on Topological Spaces Theory and its Applications, 24–25 August, 1996, Yatsushiro Coll. Tech., pp. 13–18.

[23] A. S. MASHHOUR, M. E. ABD EL-MONSEF and S. N. EL-DEEB: On precontinuous and weak precontinuous mappings, *Proc. Math. Phys. Soc. Egypt* **53** (1982), 47–53.

[24] A. S. MASHHOUR, I. A. HASANEIN and S. N. EL-DEEB: α-continuous and α-open mappings, *Acta Math. Hungar.* **41** (1983), 213–218.

[25] O. NJÅSTAD: On some classes of nearly open sets, *Pasific J. Math.* **15** (1965), 961–970.

[26] T. M. J. NOUR: Contributions to the Theory of Bitopological Spaces, Ph.D. Thesis, University of Delhi, 1989.

[27] V. POPA and T. NOIRI: On M-continuous functions, *Anal. Univ. "Dunarea Jos" – Galati, Ser. Mat. Fiz. Mec. Teor. Fasc. II* **18(23)** (2000), 31–41.

[28] P. SUNDARAM, H. MAKI and K. BALACHANDRAN: Semi-generalized continuous maps and semi-$T_{\frac{1}{2}}$ spaces, *Bull. Fukuoka Univ. Ed. Part III* **40** (1991), 33–40.

[29] N. V. VELIČKO: H-closed topological spaces, *Mat. Sb.* **70** (1966), 98–112; English transl. (2), in *Amer. Math. Soc. Transl.* **78** (1968), 102–118.

M. Caldas
Departamento de Matemática Aplicada
Universidade Federal Fluminense
Rua Mário Santos Braga, s/n
24020-140, Niterói, RJ-BRASIL
gmamccs@vm.uff.br

D. N. Georgiou
Department of Mathematics
University of Patras
26500 Patras, GREECE
georgiou@math.upatras.gr

S. Jafari
Department of Mathematics and Physics
Roskilde University
Postbox 260
4000 Roskilde, DENMARK
sjafari@ruc.dk

T. Noiri
Department of Mathematics
Yatsushiro College of Technology
Yatsushiro, Kumamoto
866-8501 JAPAN
noiri@as.yatsushiro-nct.ac.jp

ON A FUNDAMENTAL THEOREM OF REFLECTION GEOMETRY

By

ESZTER HORVÁTH

(Received May 14, 2004)

1. Introduction

Felix Klein, German mathematician, established a new aspect for classification of geometries in his famous inaugural address at the University of Erlangen in 1872. He describes a geometry from that point of view which geometric properties remain invariant while applying its certain possible transformations. These are the invariants of the geometry concerned. Such invariant is, e. g., the distance in the usual Euclidean geometry, the angle in its similarity- or in its circle-geometry, the cross ratio of any four collinear points in the projective geometry. Felix Klein gave a new direction of research that is referred to as Erlanger Program. An example for this aspect is reflection geometry. I shall discuss one of its problems without the claim of completeness.

Geometric transformations, as bijective mappings of a space onto itself, form a group with the successive application (composition) as the product operation. In the plane the reflection in a line, in the space the reflection in a plane has a prominent role in the outline and systematization of congruence transformations (isometries). Reflection geometry attempts to describe the geometries of a wide range on the basis of these transformations, with reference to the classical work of Bachmann [1] and the paper of Ahrens [2]. A possible, simplified discussion in plane and in space is given in the works of E. Molnár [3] and [4]. These papers called my attention to the topic and to these investigations.

Line-reflections afford possibility to systematize congruence transformations in plane. During this we replace two reflections in arbitrary two lines with reflections in other two lines, one of them going through any given point.

In space we can make such replacement for plane-reflections. The fundamental theorems on three reflections (due to Hjelmslev, see Theorem 1 and Theorem 2) present the possibility and the method, moreover, state uniqueness of construction. For convenience of the Reader this paper includes the proof of Theorem 2, where we can observe the effective and nice applications of axioms.

We can define line-reflection in space as well. The question arises whether the analogous unique construction of Theorem 1 is possible in space or not. The main theorem, Theorem 5, states that such construction exists. In the proof I show that the construction is not always unique in space.

In this paper a survey of primitive concepts, axioms and theorems will be given as well. To prove Theorem 5 mainly the Axiom 3, Axiom 4 and Theorem 4 are needed. But I give the proof of Theorem 3 and Theorem 4 too, because they offer the concrete construction of Theorem 5.

2. Primitive concepts and notations

Points: $A, B, C, \dots$

Lines: $a, b, c, \dots$

Planes: $\alpha, \beta, \gamma, \dots$

In the following α denotes the reflection in the plane α as well, the same holds for points and lines, respectively.

The point P is incident with the line e: $P \mathbf{I} e$.

The point P and the line e are incident with the plane α: $P \mathbf{I} \alpha$, $e \mathbf{I} \alpha$.

$\equiv$ denotes the equality of points, lines and planes.

$=$ denotes the equality of transformations.

Product ϕ of reflections in the planes $\alpha, \beta, \gamma, \dots$ (successive application in this order): $\alpha\beta\gamma \cdots =: \phi$.

The transformation ϕ maps the point P, the line e, and the plane α to their images: P^ϕ, e^ϕ, α^ϕ, respectively.

The inverse of transformation ϕ: ϕ^{-1}.

ϕ is involutive, if $\phi = \phi^{-1} \neq 1$ with the identity 1.

$\phi^{-1}\psi\phi$: ψ transformed (conjugated) by transformation ϕ.

If ϕ is involutive, then $\phi^{-1}\psi\phi = \phi\psi\phi$.

Especially $\varphi^{-1}\alpha\varphi =: \alpha^{\varphi}$ is the plane-reflection in the φ-image of α.

$\alpha \perp \beta$: The plane α is perpendicular to plane β, if $\alpha^{\beta} \equiv \alpha$ and $\alpha \not\equiv \beta$.

$a \perp \beta$: The line a is perpendicular to plane β, if $a^{\beta} \equiv a$ and $a \not\mathbf{I} \beta$.

Line-reflection: If $\alpha \perp \beta$ and $g \ \mathbf{I}\ \alpha,\beta$, then the transformation $\alpha\beta = \beta\alpha$ is called line-reflection g. The transformation $g = \alpha\beta$ is involutive, because

$$\alpha^{\beta} = \beta\alpha\beta = \alpha \iff \alpha\beta = \beta\alpha = (\alpha\beta)^{-1}.$$

Point-reflection: If $P \ \mathbf{I}\ a,\alpha$ and $a \perp \alpha$, then the transformation $\alpha a = a\alpha$ is called point-reflection P. The transformation $P = \alpha a$ is involutive, because

$$a^{\alpha} = \alpha a \alpha = a \iff \alpha a = a\alpha = (\alpha a)^{-1}.$$

We can discuss axioms and theorems on the basis of the same principle in Euclidean and hyperbolic geometry. The primitive concepts, axioms and theorems in elliptic geometry(where a quadrupel of pair-wise perpendicular planes exists) are defined analogously, but some questions need different, in general simpler discussions. Therefore, in this note I do not consider elliptic geometry.

3. Axioms and theorems

As Axiom 0., consider all the usual incidence and orthogonality statements in plane and in space, respectively. Then the reflection in a line of the plane, and the reflection in a plane of the space can be introduced [1]–[4] as usual. To these concepts refer the following axioms, in addition.

3.1. Axioms of tree line-reflections in plane

Axiom 1. *If $a,b,c \ \mathbf{I}\ P$, then there exists a line d, such that $abc = d$ and the consequence is $d \ \mathbf{I}\ P$.*

Axiom 2. *If $a,b,c \perp g$, then there exists a line d, such that $abc = d$ and the consequence is $d \perp g$.*

3.2. Axioms of tree plane-reflections in space

AXIOM 3. *If α,β,γ **I** g, then there exists a plane δ, such that $\alpha\beta\gamma = \delta$ and this implies δ **I** g.*

AXIOM 4. *If $g \perp \alpha,\beta,\gamma$, then there exists a plane δ, such that $\alpha\beta\gamma = \delta$ and this implies $g \perp \delta$.*

Of course, by Axiom 3 and Axiom 4 we can prove Axiom 1 and Axiom 2 in space by using the definition of line-reflection.

3.3. Theorems

THEOREM 1 (THE FUNDAMENTAL THEOREM IN PLANE). *Let a and b be two distinct lines in the plane and let X be a point not incident with both lines. Then there exists just one line g, such that g **I** X and $agb = h$ is a line-reflection.*

The consequence of the theorem is that there exists a unique line h_1 as well, such that $ab = gh_1$, where $h_1 = bhb = h^b$.

THEOREM 2 (THE FUNDAMENTAL THEOREM IN SPACE). *Let α and β be two distinct planes and let X be a point not incident with both planes. Then there exists exactly one plane φ, such that φ **I** X and $\alpha\varphi\beta = \psi$ is a reflection.*

The consequence of this theorem is that there exists a unique plane ψ_1 as well, such that $\alpha\beta = \varphi\psi_1$, where $\psi_1 = \beta\psi\beta = \psi^\beta$.

PROOF. (See Figure 1. and 2.)

Let us draw a perpendicular line a from X to the plane α. Let the intersection point of a and α be denoted by A. We similarly get the line b and the point B. Let g be the line connecting A and B. Let δ' be the plane, such that X **I** δ' and $\delta' \perp g$. The plane ξ is determined by X, a, b **I** ξ. Let α' and β' be planes, such that a **I** α', $\alpha' \perp \xi$, b **I** β', $\beta' \perp \xi$. X **I** α', β', therefore α' and β' meet in line m. δ' is incident with m as well. By Axiom 3 there exists a plane φ, such that $\alpha'\delta'\beta' = \varphi$ and φ **I** m, $\varphi \perp \xi$, thus $\varphi\xi = \xi\varphi$.

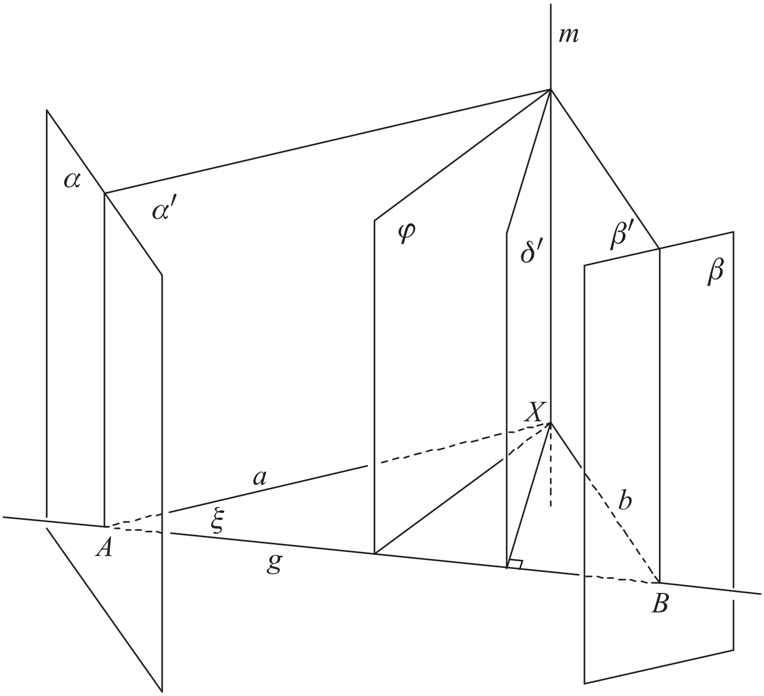


FIGURE 1.

Hence:

$$a\delta' b = (\xi\alpha')\delta'(\beta'\xi) = \xi(\alpha'\delta'\beta')\xi = \xi\varphi\xi = \varphi .$$

Therefore:

$$a\varphi b = \delta' . \tag{1}$$

According to the definition of point-reflection:

$$A = a\alpha = \alpha a \iff \alpha = aA = Aa ,$$
$$B = b\beta = \beta b \iff \beta = bB = Bb .$$

Using these relations:

$$\alpha\varphi\beta = (Aa)\varphi(bB) = A(a\varphi b)B .$$

By (1):

$$\alpha\varphi\beta = A\delta' B . \tag{2}$$

We produce the point-reflections A and B in another way:

$$\alpha'' \perp g ,\ \alpha'' \,\mathbf{I}\, A \Rightarrow A = g\alpha'' ;\ \beta'' \perp g ,\ \beta'' \,\mathbf{I}\, B \Rightarrow B = \beta'' g .$$

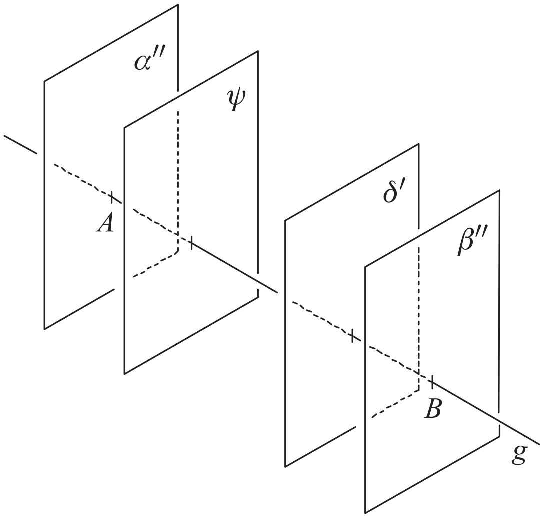


FIGURE 2.

By Axiom 4 there exists a plane $\psi \perp g$, such that

$$\alpha''\delta'\beta'' = \psi .$$

By (2):

$$\alpha\varphi\beta = A\delta' B = (g\alpha'')\delta'(\beta'' g) = g(\alpha''\delta'\beta'')g = g\psi g = \psi .$$

So we have proved our theorem. ∎

We get the consequence of Theorem 2 by the following:

$$\psi = \alpha\varphi\beta \Rightarrow \alpha\psi = \varphi\beta \Rightarrow \alpha = \varphi\beta\psi \Rightarrow \alpha\beta = \varphi\beta\psi\beta = \varphi\psi^{\beta}.$$

Uniqueness of φ (and ψ_1) follows by an indirect way [1], [2], [4], not detailed here.

THEOREM 3. *If $\Phi = \alpha\beta\gamma$ is a product of tree plane-reflections and X is an arbitrary point in the space, then there exist planes $\varepsilon, \varphi, \psi$, such that $X \mathbf{I} \varepsilon$ and $\varepsilon, \varphi \perp \psi$ and $\Phi = \varepsilon\psi\varphi$.*

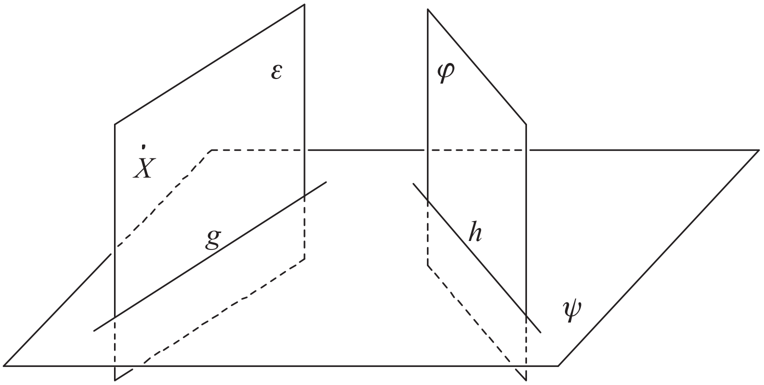


FIGURE 3.

If ε and ψ meet in line g, and φ and ψ meet in line h, then $\Phi = \varepsilon h = g\varphi$.

PROOF. (See the notations of Figure 4.)

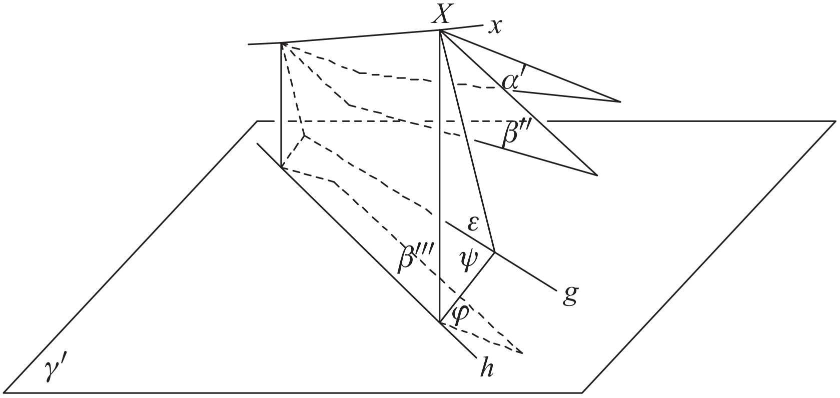


FIGURE 4.

By Theorem 2 there exist planes α' and β', such that $X \mathbf{I} \alpha'$ and $\alpha\beta = = \alpha'\beta'$. Applying again Theorem 2 there exist β'' and γ', such that $X \mathbf{I} \beta''$ and $\beta'\gamma = \beta''\gamma'$. Let x be the common line of α' and β''. We take the plane β''', such that $x \mathbf{I} \beta'''$ and $\beta''' \perp \gamma'$. (If $x \perp \gamma'$ then we can take an arbitrary plane β''' incident with x.) By Axiom 3 there exists the plane ε, such that $\varepsilon \mathbf{I} x$ and $\alpha'\beta'' = \varepsilon\beta'''$. Assume that β''' and γ' meet in the line h, and let ψ be the plane, such that $\psi \mathbf{I} h$ and $\psi \perp \varepsilon$. (If $h \perp \varepsilon$ than we can take an arbitrary plane ψ incident with h.) Now we use again Axiom 3, there exists a plane φ, such that $\beta'''\gamma' = \varphi\psi$ and $\varphi \mathbf{I} h$. By the assumption $\beta''' \perp \gamma'$ we have $\varphi \perp \psi$. Hence $\varphi\psi = \psi\varphi$.

Therefore

$$\alpha\beta\gamma = \alpha'\beta'\gamma = \alpha'\beta''\gamma' = \varepsilon\beta'''\gamma' = \varepsilon\varphi\psi = \varepsilon\psi\varphi \ .$$

ψ and φ meet in the line h, ε and ψ meet in the line g, ε, $\varphi \perp \psi$. We can use the definition of the line-reflection:

$$\alpha\beta\gamma = (\varepsilon\psi)\varphi = g\varphi$$

$$\alpha\beta\gamma = \varepsilon(\psi\varphi) = \varepsilon h \ .$$

We can choose planes β''' and ψ in some cases not uniquely (see above). Then the construction of the theorem is not necessarily unique. ∎

THEOREM 4. *If* $\Phi = \alpha\beta\gamma\delta$ *is a product of four plane-reflections and* X *is an arbitrary point in the space, then there exist lines* g *and* h*, such that* $\Phi = gh$ *and* $X \mathbf{I} g$.

PROOF. (See the notations of Figure 5.)

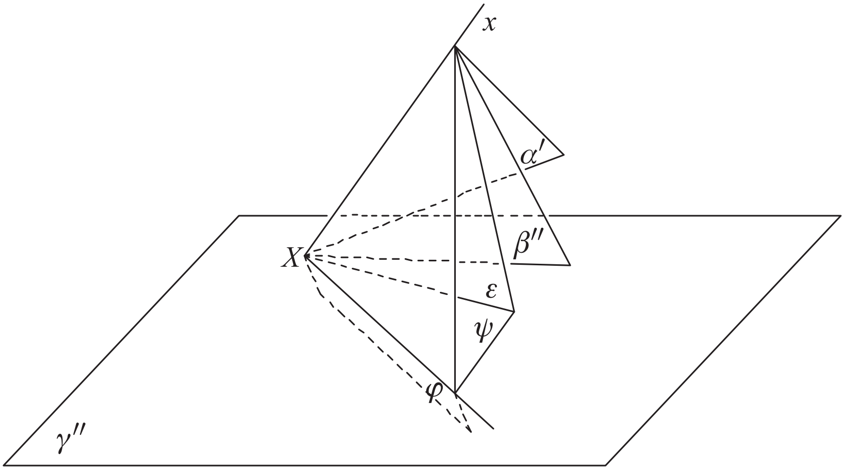


FIGURE 5.

By Theorem 2 there exist planes α' and β', such that $X \mathbf{I} \alpha'$ and $\alpha\beta = \alpha'\beta'$. Let us apply again Theorem 2, moreover, there exist β'' and γ' with $X \mathbf{I} \beta''$ and $\beta'\gamma = \beta''\gamma'$, and there exist γ'' and δ' with $X \mathbf{I} \gamma''$ and $\gamma'\delta = \gamma''\delta'$. By Theorem 3 we can take planes $\varepsilon, \varphi, \psi$, such that $\alpha'\beta''\gamma'' = \varepsilon\psi\varphi$. Now (see Figure 5) $\alpha', \beta'', \gamma'' \mathbf{I} X$ imply $X \mathbf{I} \varepsilon, \varphi, \psi$ and $\varepsilon, \varphi \perp \psi$. Hence $\alpha'\beta''\gamma'' = \psi\varepsilon\varphi$.

(See Figure 6.) Then ε and φ meet in a common line, say f, $f \mathbf{I} X$ and $f \perp \psi$. There exists the plane ω, such that $\omega \mathbf{I} f$ and $\omega \perp \delta'$. (If $f \perp \delta'$ then we can take an arbitrary plane incident with f.) ω and δ' meet in line h. By Axiom 3 there exists a plane π, such that $\varepsilon\varphi\omega = \pi$ and π **I** f. Then $\pi \perp \psi$. π and ψ meet in line g.

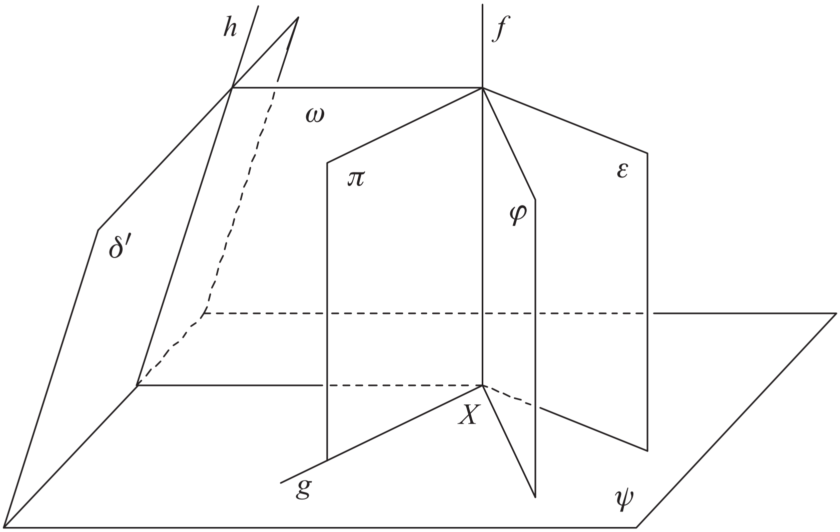


FIGURE 6.

Therefore,

$$\Phi = \alpha\beta\gamma\delta = \alpha'\beta''\gamma''\delta' = \psi\varepsilon\varphi\delta' = \psi(\varepsilon\varphi\omega)(\omega\delta') = (\psi\pi)h = gh$$

and X **I** g. ▮

In this proof we could see that the construction was not necessarily unique.

4. Line-reflections in Space

The question arises whether Theorem 1 has an analogon for line-reflections in space. The main purpose of this note is to prove:

THEOREM 5. *Let a and b be two distinct lines in the space and let X be an arbitrary point in the space not incident with both lines. Then there exists a line g, such that X is incident with g and agb can be replaced with a line-reflection h.*

PROOF. *If a and b has a common perpendicular line – which is always true in Euclidean space:*
(See the notations of Figure 7.)

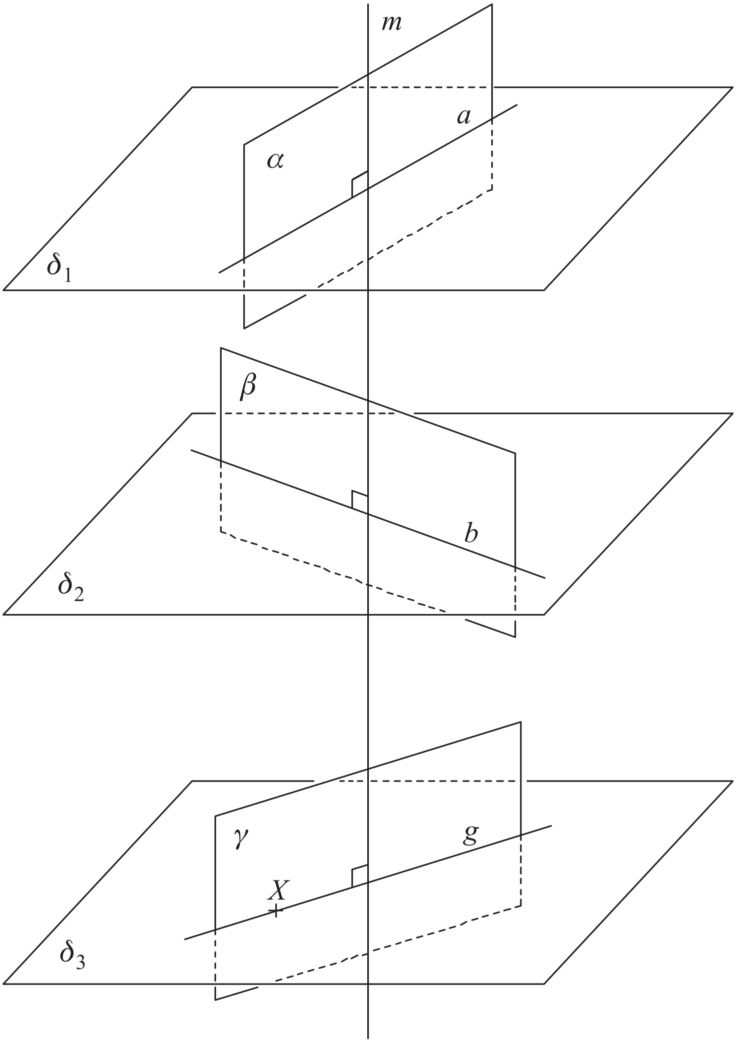


FIGURE 7.

Let m be the common perpendicular line of a and b. X and m determine a plane γ. (If $X\ \mathbf{I}\ m$ then there exist more such planes, we take any of them.) In this plane there exists a line g, such that $X\ \mathbf{I}\ g$ and $g \perp m$. The planes $\alpha \equiv [a,m]$, $\beta \equiv [b,m]$, $\gamma \equiv [g,m]$ are determined. We take the plane δ_1, such that $a\ \mathbf{I}\ \delta_1$ and $m \perp \delta_1$. The planes δ_2 and δ_3 are similarly introduced.

By the definition of line-reflection:

$$a = \alpha\delta_1\,, \qquad b = \beta\delta_2\,, \qquad g = \gamma\delta_3\,.$$

We would like to simplify the product $agb = (\alpha\delta_1)(\gamma\delta_3)(\beta\delta_2)$.
$\alpha,\beta,\gamma \perp \delta_i$ $(i = 1,2,3)$, because $m \perp \delta_i$ and $m \mathbf{I} \alpha,\beta,\gamma$.
Hence $\alpha\delta_i = \delta_i\alpha$; $\beta\delta_i = \delta_i\beta$; $\gamma\delta_i = \delta_i\gamma$.
Therefore:

$$agb = (\alpha\gamma\beta)(\delta_1\delta_3\delta_2) .$$

As $\alpha,\beta,\gamma \mathbf{I}\, m$, so by Axiom 3 there exists a plane φ, such that $\alpha\gamma\beta = \varphi$ and $\varphi \mathbf{I}\, m$. Since $m \perp \delta_1,\delta_2,\delta_3$, therefore by Axiom 4, there exists a plane ψ, such that $\delta_1\delta_3\delta_2 = \psi$ and $m \perp \psi$. Hence $agb = \varphi\psi$.

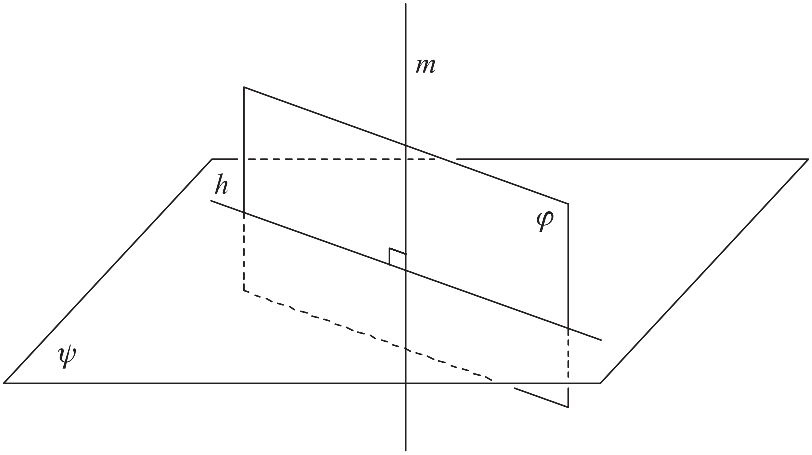


FIGURE 8.

We have $m \mathbf{I} \varphi$ and $m \perp \psi$, so $\varphi \perp \psi$. The planes φ and ψ meet in the line h and $\varphi\psi = h$. Hence we can replace the product agb by the line-reflection h, where m is perpendicular to h (Figure 8). From the proof we can see that if X is incident with the common perpendicular of a and b then the construction is not unique.

If we do not suppose the existence of a common perpendicular line for a and b: (see Figure 9) – it is possible in hyperbolic space.

Let α be the plane of the point X and of the line a. If X is not incident with the line a, then α is uniquely determined, otherwise we are done with $g = a$, $h = b$. We take the plane α_1 with the conditions $\alpha_1 \mathbf{I}\, a$ and $\alpha_1 \perp \alpha$.

Then

$$X \mathbf{I} \alpha \quad \text{and} \quad a = \alpha\alpha_1 = \alpha_1\alpha .$$

Planes β and β_1 are similarly introduced (X **I** b implies $g = b$, $h = a$):

$$X \mathbf{I} \beta \quad \text{and} \quad b = \beta\beta_1 = \beta_1\beta\,.$$

Using these notations:

$$ab = \alpha_1\alpha\beta\beta_1\,.$$

Applying Theorem 4 for these four reflections there exist line-reflections g and h_1, such that:

$$\alpha_1\alpha\beta\beta_1 = gh_1 \quad \text{and} \quad X \mathbf{I} g\,.$$

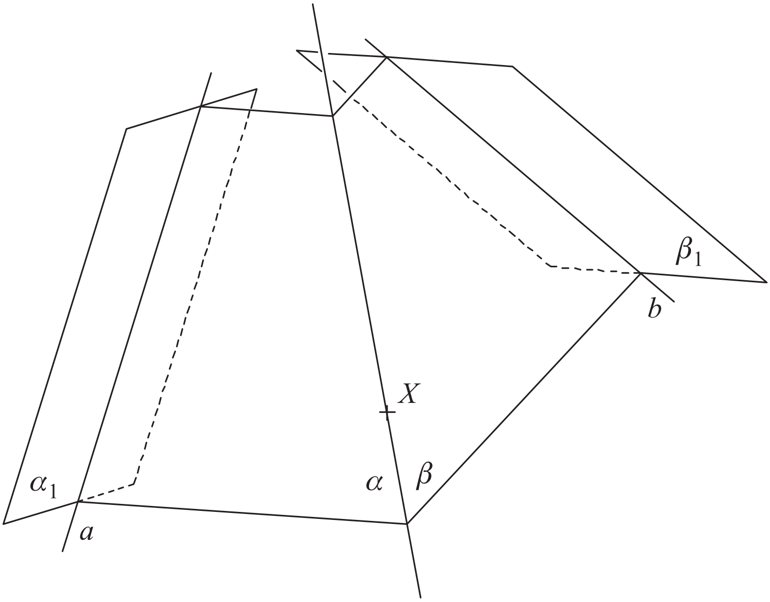


FIGURE 9.

Hence

$$\begin{aligned} ab &= gh_1 \\ gab &= h_1 \\ bga &= bh_1b \\ agb &= bh_1b = h_1^b. \end{aligned} \tag{3}$$

Since the line-reflection is involutive, bh_1b is just h_1 transformed by b, i.e., the reflection in the line $h \equiv h_1^b$. Hence the construction $agb = h$ exists, indeed. ∎

5. About uniqueness of the construction

The positions of the lines a and b and of the point X, respectively, may lead to different constructions for g and h in Theorem 5.

Now I examine, what are the criteria for that.

Let us suppose that there exist two lines $g_1, g_2 \mathbf{I} X$ and lines h_1, h_2, such that

$$ag_1b = bh_1b \quad \text{and} \quad ag_2b = bh_2b\,. \tag{4}$$

Then, applying equations (3) in opposite way, we get:

$$ab = g_1h_1 = g_2h_2$$
$$g_1g_2 = h_1h_2\,.$$

X is a fixed point of the transformation g_1g_2, so it is a fixed point of h_1h_2. If X is incident with h_1, then lines g_1 and h_1 have a common point, so they have a common perpendicular line, too.

Now let us suppose that X is not on the line h_1.

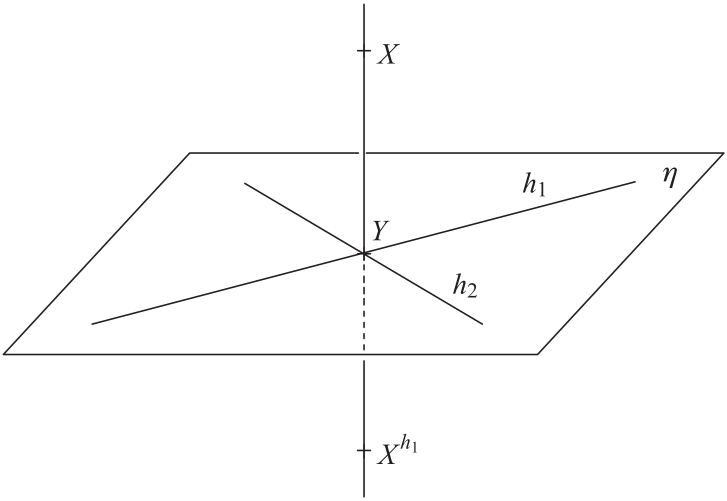


FIGURE 10.

By $X \equiv X^{h_1h_2}$ we have $X^{h_1} \equiv X^{h_2}$. According to the definition of line-reflection, h_1 and h_2 are incident with a plane η, such that is perpendicular to the line $x = XX^{h_1}$. The common point of the line x and the plane η is Y. h_1 and h_2 meet in the point Y, too. This point Y is a fixed point of transformations h_1h_2 and g_1g_2. In this case XY is an invariant (pointwise fixed) line of h_1h_2 and g_1g_2 and we have $XY \perp g_1, h_1$ as well. Hence the lines g_1 and h_1 have a common perpendicular line, such that is incident with the point X.

Therefore, if the construction of Theorem 5 according to (4) is not unique, then there exist lines g_1 and h_1 with $ab = g_1h_1$, such that g_1 and h_1 have a common perpendicular line and $g_1 \mathbf{I} X$.

I shall prove that if g_1 and h_1 have a common perpendicular line then we can construct many appropriate lines g_2 and h_2 being different from g_1 and h_1, with $g_1h_1 = g_2h_2$, $g_1, g_2 \mathbf{I} X$.

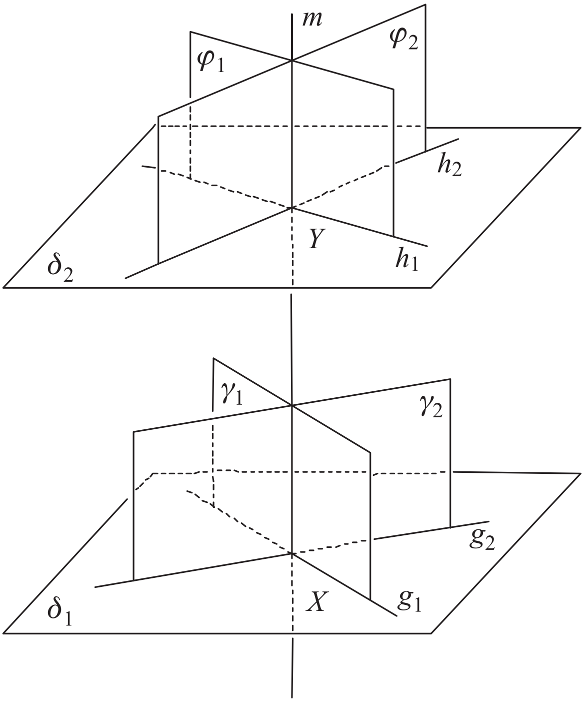


FIGURE 11.

The common perpendicular line of g_1 and h_1 is $m = [X, Y]$. Let δ_1 be the plane, such that $\delta_1 \mathbf{I} g_1$ and $\delta_1 \perp m$. Let δ_2 be the plane, such that $\delta_2 \mathbf{I} h_1$ and $\delta_2 \perp m$. Let g_2 be any line different from g_1, such that is incident with point X and plane δ_1.

The lines g_1, g_2, h_1 and m determine uniquely the planes $\gamma_1 \equiv [g_1, m]$, $\gamma_2 \equiv [g_2, m]$, $\varphi_1 \equiv [h_1, m]$.

By the definition of line-reflection:

$$\begin{aligned} g_1 &= \gamma_1\delta_1 = \delta_1\gamma_1 \\ g_2 &= \gamma_2\delta_1 = \delta_1\gamma_2 \\ h_1 &= \varphi_1\delta_2 = \delta_2\varphi_1 \,. \end{aligned}$$

Hence

$$h_1g_1g_2 = \delta_2\varphi_1\gamma_1\delta_1\delta_1\gamma_2 = \delta_2\varphi_1\gamma_1\gamma_2 \,.$$

$\varphi_1, \gamma_1, \gamma_2$ **I** m, by Axiom 3 there exists a plane φ_2 **I** m, such that $\varphi_1\gamma_1\gamma_2 = \varphi_2$. Whence

$$h_1g_1g_2 = \delta_2\varphi_2 \,.$$

Because of $\delta_2 \perp m$ and φ_2 **I** m we have $\varphi_2 \perp \delta_2$. φ_2 and δ_2 have a common point Y, therefore the planes meet in the line h_2. By the definition of line-reflection:

$$\begin{aligned} h_2 &= \delta_2\varphi_2 = \varphi_2\delta_2 \\ h_1g_1g_2 &= h_2 \\ g_1g_2 &= h_1h_2 \,. \end{aligned}$$

Now applying equation (3), we get two different constructions of the product $agb = h$, g **I** X.

What can be said about the starting lines a and b in the case above?

We have $ab = g_1h_1 = g_2h_2$, X **I** $g_1 \not\equiv g_2$, Y **I** $h_1 \not\equiv h_2$ and $m = XY$ is a common perpendicular to g_1, g_2, h_1, h_2. Let P be any point of line a. P and m determine a plane. There exists a line a_1 in this plane, such that P **I** a_1 and $a_1 \perp m$. Since $a_1, g_1, h_1 \perp m$, there exists a line $b_1 \perp m$ (according to the previous proof, see Figure 7), such that $a_1g_1h_1 = b_1 \Rightarrow a_1b_1 = g_1h_1$. Thus $g_1h_1 = ab = a_1b_1$ and P **I** a, a_1.

If $a \equiv a_1$, then $b \equiv b_1$, thus m is also the common perpendicular to a and b, and X is incident with their common perpendicular.

If $a \not\equiv a_1$, since $ab = a_1b_1$ and P **I** a, a_1, then a and b have a common perpendicular through the point P (as we proved it for g_1 and h_1 before). However, we can find by this method common perpendicular of a and b for any point of a. This can happen only in Euclidean geometry when a and b are parallel, i.e., they have more than one common perpendicular. In this case, ab is an Euclidean translation and trough any point X of space we have a line x parallel to the common perpendiculars of a and b. Then for any g with

X **I** $g \perp x$ there exists a unique h such that $ab = gh$, i.e., $agb = bhb = h^b$, as in Theorem 5.

SUMMARY. In Euclidean space: the construction in Theorem 5 is unique precisely if a and b are skew lines and X is not on their common perpendicular; in hyperbolic space: either if a and b have common perpendicular line and X is not incident with it, or if a and b do not have common perpendicular line.

This latter case occurs in the classical Bolyai–Lobachevskien hyperbolic space iff a and b are parallel in hyperbolic sense, i.e. they lie in a plane without common point and without common perpendicular line [1]. Then they determine a line bundle, with one line trough any point X of space. In the line bundle holds Euclidean plane geometry with bundle lines as "new points" and bundle planes as "new lines". Our Theorem 5 means in this line bundle: three "point-reflections" can be replaced with one "point-reflection" as usual in the Euclidean plane.

References

[1] F. BACHMANN: *Aufbau der Geometrie aus dem Spiegelunsbegriff*, Springer, 1959, 1973.

[2] J. AHRENS: Begründung der absoluten Geometrie des Raumes aus dem Spiegelungsbegriff, *Math. Zeitschrift* **71** (1959), 154–185.

[3] EMIL MOLNÁR: A tükrözésgeometriáról, *ELTE TTK Szakmódszertani Közleményei* **VII** (1974), 86–130. (On reflection geometry, in Hungarian with German summary; *Methodological Communications of Eötvös L. Univ., Fac. of Nat. Sci.*)

[4] EMIL MOLNÁR: Tükrözésgeometria a térben, *ELTE TTK Szakmódszertani Közleményei* **VIII** (1975), 76–107. (Reflection geometry in space, in Hungarian with German summary; *Methodological Communications of Eötvös L. Univ., Fac. of Nat. Sci.*)

Eszter Horváth

Szilágyi Erzsébet Gymnasium
Budapest
Hungary
h_esz@freestart.hu

ON THE JORDAN STRUCTURE OF TERNARY RINGS OF OPERATORS

By

JOSÉ M. ISIDRO[1] and LÁSZLÓ L. STACHÓ[2]

(Received May 14, 2004)

By a *ternary ring of operators (TRO)* we mean a norm-closed subspace in some $\mathscr{L}(H,K)$ (=$\{$bounded linear operators $H \to K\}$ with complex Hilbert spaces H, K) which is closed under the *ternary product* $[xyz] := xy^*z$. TRO's were introduced by Hestenes [9, 1962] who proved that, in the finite dimensional setting, TRO's can be faithfully represented as direct sums of spaces $\mathcal{M}_{m,n}(\mathbb{C})$ of $m \times n$ complex matrices. In infinite dimensions, Zettle [13, 1983] gave a characterization of TRO's among ternary Banach algebras, whence one could discover that Hilbert C*-modules are the same as TRO's. Henceforth many deep results have appeared studying TRO's and their applications, see [3, 2001], [11, 2002] and [6, 1999], among others showing that every TRO is isometrically isomorphic to a corner $pA(1-p)$ of a C*-algebra and that the ternary product is uniquely determined by the metric structure in a TRO. As a consequence, since the bidual of a C*-algebra is a W*-algebra, a TRO can be represented as a weak*-dense subTRO in $\bigoplus_{i\in I} p_i A_i(1-p_i)$, where $(A_i)_{i\in I}$ is the family of M-summands of A^{**}. The aim of this note is to show that this description can be refined somewhat to an infinite dimensional version of Hestenes' theorem. Namely we have the following

THEOREM 1.1. *Every TRO is isometrically isomorphic to a weak*-dense subTRO of the natural TRO of a direct sum* $\bigoplus_{i\in I}\mathscr{L}(H_i,K_i)$. *In particular,*

[1] Supported by Ministerio de Educación y Cultura of Spain, Research Project BFM 2002-1529.

[2] Supported by Spanish–Hungarian Project TéT E-50/02 and Hungarian Research Grant OTKA T34267.

up to isometric isomorphisms, TRO's with predual are ℓ_∞-direct sums of $\mathscr{L}(H,K)$-spaces and a reflexive TRO is a finite ℓ_∞-direct sum of copies of $\mathscr{L}(H,K)$ spaces with $\dim K < \infty$.

Our proofs rely upon the Jordan theory of Banach spaces with symmetric unit ball, the so called *JB*-triples*. According to a result of Harris [7, 1973], TRO's when equipped with the Jordan triple product $(*)$ $\{xyz\} :=$ $= (xy^*z + zy^*x)/2$ are JC*-triples and hence their unit ball is necessarily symmetric. Since the bidual of a C*-algebra is isometrically isomorphic to a weak*-closed subalgebra in some $\mathscr{L}(\widehat{H})$, the bidual of a TRO is a TRO again. Therefore, by Friedmann–Russo's Gelfand–Naimark type theorem for JB*-triples [4, 1985], it follows that any TRO E is isometrically isomorphic to a weak*-dense subTRO in the (ℓ^∞-direct) sum $\oplus_{j\in J} F_j$ of the minimal weak*-closed M-summands, the so called *Cartan factors,* of the bidual E^{**}, furthermore each Cartan factor F_j is a subTRO of E^{**}. From the theorem and its Jordan theoretical proof we obtain also the following characterization of TRO's among JB*-triples.

COROLLARY 1.2. *A JB*-triple E is the triple associated to a TRO if and only if in the canonical decomposition of the bidual $E^{**} = E_{\text{at}} \oplus E_n$, the atomic ideal E_{at} consists only of Cartan factors of type 1. A TRO admits no Jordan*-representation (JB*-homomorphism) with weak*-dense range into a Cartan factor that is not of type 1.*

REMARK 1.3. From a holomorphic view point, JC*-triples (norm-closed subspaces of some $\mathscr{L}(H)$ closed under the Jordan-triple product $\{xyz\} :=$ $= xy^*z/2 + zy^*x/2$) are known as (isometric) copies of Banach spaces with symmetric unit balls which admit only vanishing Jordan representations in exceptional Cartan factors. It would be tempting to conjecture that TRO's are copies of those Banach spaces with symmetric unit ball whose Jordan representations in Cartan factors not isomorphic to some $\mathscr{L}(H,K)$ vanish. However this is not the case. Namely the assumption of the weak*-density of the range in Corollary 1.2 is indispensable: *There is an isometric JB*-homomorphism of the TRO $\mathcal{M}_n(\mathbb{C})$ of complex n-square matrices into the space $\mathscr{S}_{2n}(\mathbb{C})$ of symmetric 2n-square matrices.*

2. Proofs

Before stating the proofs we recall some basic facts and notions involved. We know that given a surjective linear isometry $T\colon F_1 \to F_2$ between two TRO's, necessarily $T[xyz] = [(Tx)(Ty)(Tz)]$, $(x, y, z \in F_1)$. Furthermore if $F_i \subset \mathscr{L}(H_i, K_i)$, $(i \in I)$, are TRO's then their ℓ_∞-sum $\bigoplus_{i\in I} F_i$ is a TRO in the space $\mathscr{L}\big(\bigoplus_{i\in I}^2 H_i, \bigoplus_{i\in I}^2 K_i\big)$ with the ℓ_2-sums $\bigoplus_{i\in I}^2 H_i$ and $\bigoplus_{i\in I}^2 K_i$, and the natural pointwise operation $[(x_i)(y_i)(z_i)] := (x_i y_i z_i)$.

For later use, recall that JB*-triples can be equipped with a unique three variable operation $(x, y, z) \mapsto \{xyz\}$ which is symmetric linear in x, z and conjugate-linear in y satisfying among other axioms (for a complete list see [4]) the Jordan identity

$$\{ab\{xyz\}\} = \{\{abx\}yz\} - \{x\{bay\}z\} + \{xy\{abz\}\}$$

and the C*-axiom $\|\{xxx\}\| = \|x\|^3$. An element e in a JB*-triple is called a *tripotent* if $0 \neq e = \{eee\}$ in which case it has norm 1 and we write $Tri(E)$ for their family. Tripotents with respect to the Jordan triple product in a TRO are partial isometries. A tripotent e is said to be an *atom* in E if $\{eEe\} = \mathbb{C}e$ and we write $At(E)$ for the set of them. Recall that given $e, f \in Tri(E)$ we say that e *governs* f (written $e \vdash f$) if $e \in E_1(f) := \{x \in E\colon \{eex\} = x\}$ and $f \in E_{1/2}(e) := \{x \in E\colon \{eex\} = x/2\}$. We say that e,f are *collinear* (written $e \top f$) if $e \in E_{1/2}(f)$ and $f \in E_{1/2}(e)$.

In order to establish our main result we need some technical lemmas on JB*-triples.

LEMMA 2.1. *Let F be a TRO in $\mathscr{L}(H)$ and suppose e,$f \in Tri(F)$ are such that $\{eef\} = f/2$. Then the elements $x := ee^*f$ and $y := fe^*e$ are orthogonal tripotents in F that satisfy $f = x + y$.*

PROOF. By assumption $f = 2\{eef\} = ee^*f + fe^*e = x + y$. Hence $x = ee^*f = ee^*ee^*f + ee^*fe^*e = x + xe^*e$, that is $xe^*e = ee^*y = ee^*fe^*e = = 0$. It follows $xy^* = ee^*fe^*ef^* = 0$, $yx^* = (xy^*)^* = 0$. Similarly $x^*y = f^*ee^*fe^*e = 0$, $y^*x = (x^*y)^* = 0$. Therefore

$$x + y = f = ff^*f = (x + y)(x + y)^*(x + y) =$$
$$= xx^*x + yy^*y,$$
$$ee^*(x + y) = ee^*xx^*x + e$$

since $x = ee^*x$ and $ee^*y = 0$. This means that $x = xx^*x$ and $y = yy^*y$, thus $x, y \in Tri(F)$. On the other hand $2\{xxy\} = x(x^*y) + (yx^*)x = x0 + 0x = 0$, that is $x \perp y$. ∎

LEMMA 2.2. *Let F be a TRO in $\mathcal{L}(H)$, and suppose $0 \neq e, f \in At(F)$ with $e \top f$. Then for the projections $p := ee^*$, $q := ff^*$, $P := e^*e$, $Q := f^*f$ we have either $p = q$ and $PQ = QP = 0$ or $P = Q$ and $pq = qp = 0$.*

PROOF. By Lemma 2.1 and since atoms are indecomposable into sums of non-zero orthogonal tripotents, the tripotents

$$x := ee^*f \quad y := fe^*e \quad X := ff^*e \quad Y := ef^*f$$

satisfy the alternatives

$$\begin{array}{ll} 1)\ x = f,\ \ y = 0,\ \ X = e,\ \ Y = 0, & 2)\ x = f,\ \ y = 0,\ \ X = 0,\ \ Y = e, \\ 3)\ x = 0,\ \ y = f,\ \ X = e,\ \ Y = 0, & 4)\ x = 0,\ \ y = f,\ \ X = 0,\ \ Y = e. \end{array}$$

The alternative 2) implies $ee^*f = f$, $fe^*e = 0$, $ff^*e = 0$, $ef^*f = e$ and $ff^* = f * (ee^*f)^* = ff^*ee^* = (ff^*e)e^* = 0e^* = 0$ that is $f = 0$, contradicting the assumption $0 \neq f$.

3) implies $ee^*f = 0$, $fe^*e = f$, $ff^*e = e$, $ef^*f = 0$ and $ee^* = (ff^*e)e^* = = e(ee^*f)^* = e0^* = 0$ that is $e = 0$, contradicting the assumption $0 \neq e$.

1) means $ee^*f = f$, $fe^*e = 0$, $ff^*e = e$, $ef^*f = 0$. Hence $q = ff^* = = (ee^*f)f^* = (ee^*)(ff^*) = pq$ and also $q = ff^* = f(ee^*f)^* = (ff^*)(ee^*) = = qp$. Therefore $p = ee^* = (ff^*e)e^* = (ff^*)(ee^*) = (ee^*)(ff^*) = ff^* = q$. On the other hand $PQ = (e^*e)(f^*f) = e^*(ef^*f) = e^*0 = 0$, $QP = (f^*f)(e^*e) = = f^*(fe^*e) = f^*0 = 0$.

4) means $ee^*f = 0$, $fe^*e = f$, $ff^*e = 0$, $ef^*f = e$. Hence $P = e^*e = = e^*(ef^*f) = (ee^*)(f^*f) = PQ$ and also $P = e^*e = (ef^*f)^*e = (f^*f)(e^*e) = = QP$. Therefore $Q = f^*f = (fe^*e)^*f = e^*ef^*f = PQ = P$. On the other hand $qp = (ff^*)(ee^*) = (ff^*e)e^* = 0e^* = 0$ and $pq = (ee^*)(ff^*) = (ee^*f)f^* = = 0f^* = 0$. ∎

COROLLARY 2.3. *If F is a TRO in $\mathcal{L}(H)$ and $0 \neq e_1, \dots, e_N \in At(F)$ with $e_j \top e_k$ $(k \neq j)$ then either $p_1 = \cdots = p_N$ and $p'_k p'_j = 0$ $(k \neq j)$ or $p'_1 = \cdots = p'_N$ and $p_k p_j = 0$ $(k \neq j)$ for the projections $p_k := e_k e_k^*$, $p'_k := e_k^* e_k$ $(k = 1, \dots, N)$.*

PROOF. By Lemma 2.2 we have the alternatives: 1) $p_1 = p_2$ and $p'_1 p'_2 = 0$ or 2) $p'_1 = p'_2$ and $p_1 p_2 = 0$.

1) Suppose $p_j \neq p_1$. Then $p'_j = p'_1$, $p_1 p_j = p_j p_1 = 0$ and also (since $p_j \neq p_2 = p_1$) $p'_j = p'_2$, $p_2 p_j = p_j p_2 = 0$. In particular $p'_j = p'_1 = p'_2$. By our assumption 1), $p'_1 p'_2 = 0$. But then $p'_1 = p'_2 = p'_1 p'_2 = 0$ that is $e_1^* e_1 = p'_1 = 0$ and $e_1 = 0$ which is impossible.

2) Similarly we can exclude $p'_j \neq p'_1$ in this case. ∎

LEMMA 2.4. *Let F be a TRO in $\mathscr{L}(H)$, and suppose* $0 \neq e_1, e_2, e_3, e_4 \in \in At(F)$. *Then the situation* $e_3 \perp e_4$, $e_k \top e_\ell$ *(*$k < \ell$*,* $(k,\ell) \neq (3,4)$*) is impossible.*

PROOF. Let $p_k := e_k e_k^*$, $p_k' := e_k^* e_k$ $(k = 1, \dots, 4)$. We have the alternatives 1) $p_1 = p_2$ and $p_1'p_2' = p_2'p_1' = 0$ or 2) $p_1' = p_2'$ and $p_1 p_2 = p_2 p_1 = 0$.

Suppose 1). Since $e_1 \top e_2 \top e_3 \top e_1$, by Corollary 2.3 also $p_1 = p_3$. Since $e_1 \top e_2 \top e_4 \top e_1$, also $p_1 = p_4$. Thus 1) implies $p_1 = p_4$. However, the relationship $e_1 \perp e_4$ means (as it is well-known) that $0 = p_1 p_4 = p_4 p_1$ and $0 = p_1' p_4' = p_4' p_1'$. Therefore 1) is impossible. The case 2) can be treated analogously. ∎

We are now in the position to prove our maun result.

PROOF OF THEOREM 1.1. Let E be a TRO. We know that, without loss of generality, we may regard E^{**} as a weak* closed TRO in a space $\mathscr{L}(\widehat{H})$ with some Hilbert space $\widehat{H}$, moreover E is a weak* dense sub-TRO of E^{**} for the natural ternary product $[x,y,z] := xy^*z$, $(x,y,z \in \mathscr{L}(\widehat{H}))$. From a Jordan viewpoint, E^{**} is an ℓ^∞-direct sum of the form $E^{**} = E_{\mathrm{at}}^{**} \oplus E_n^{**}$ where $E_{\mathrm{at}}^{**} = \oplus_{j \in J} F_j$ and $\{F_j : j \in J\}$ is the family of all minimal atomic M-ideals of E^{**} with respect to the Jordan triple product $\{xyz\} := = \frac{1}{2}(xy^*z + zy^*x)$, $(x,y,z \in \mathscr{L}(\widehat{H}))$. Since the projection onto the atomic ideal $P_{\mathrm{at}} : E^{**} \to \oplus_{j \in J} F_j$ is an isometric JB*-homomorphism which is a bijection on E, it suffices to see that each factor F_j is a Cartan factor of type 1. Concerning Cartan factors, by the familiar classification, each F_j is isometrically isomorphic to some of the following classical JB*-triples:

$\mathscr{L}(H_j, K_j)$ [type 1],

$\mathscr{L}_\pm(H_j) := \{x \in \mathscr{L}(H_j) : x = \pm\overline{x}^*\}$ [types 2,3] with a conjugation $x \mapsto \overline{x}$,

$\mathrm{Spin}(H_j) := \big[H_j$ with $\{xyz\} := \langle x,y\rangle z + \langle z,y\rangle x - \langle x,\overline{z}\rangle\overline{y}\big]$ [type 4],

$\mathscr{M}_{1,2}(\mathbb{O})$ [type 5, of 16 dimensions], here $\mathbb{O}$ means the Cayley algebra of complex octonions.

$\mathscr{H}_3(\mathbb{O})$ [type 6, of 27 dimensions], the algebra of 3×3 hermitian matrices with entries in the octonions $\mathbb{O}$ equipped with the standard conjugation.

Our key observation is that, in all cases if F_j is not isomorphic to some $\mathscr{L}(H_j, K_j)$ then the standard covering atomic grid of F_j (see [12]) contains a couple of atoms e_1, e_2 with $e_1 \vdash e_2$ or it contains a family $\{e_1, e_2, e_3, e_4\}$ of atoms with $e_3 \perp e_4$, $e_k \top e_\ell$ $(k < \ell,\ (k,\ell) \neq (3,4))$. By the previous lemmas it immediately follows that this is impossible.

The statements concerning TRO's with predual are immediate. ∎

For the sake of completeness, we describe the mentioned systems $\{e_1, e_2\}$ respectively $\{e_1, \dots, e_4\}$ of atoms for the types 2–6.

To this aim, let H be a Hilbert space, let $x \mapsto \overline{x}$ be a conjugation on H, let $\{h_m : m \in M\}$ be a complete orthonormal system in H such that $h_m = \overline{h_m}$, $(m \in M)$, and let $e \otimes f$ denote the operator $x \mapsto \langle x, e\rangle f$ on H.

Case type 2. With $e_1 := h_1 \otimes h_1$, $e_2 := h_1 \otimes h_2 + h_2 \otimes h_1$ we have $e_1, e_2 \in At(\mathscr{L}^-{}_+(H))$ and $e_1 \vdash e_2$.

Case type 3, $\dim E > 3$. With $e_1 := h_1 \otimes h_2 - h_2 \otimes h_1$, $e_2 := h_2 \otimes h_3 - - h_3 \otimes h_2$, $e_3 := h_1 \otimes h_3 - h_3 \otimes h_1$, $e_4 := h_2 \otimes h_4 - h_4 \otimes h_2$ we have $e_1, e_2, e_3, e_4 \in At(\mathscr{L}(H))$ and $e_3 \perp e_4$, $e_k \top e_\ell$ $(k < \ell,\ (k,\ell) \neq (3,4))$.

Case type 4, $\dim E > 3$. With $e_k := 2^{-1/2}(h_k + ih_4)$, $(k = 1,2,3)$ and $e_4 := 2^{-1/2}(h_3 - ih_3)$ we have $e_1, e_2, e_3, e_4 \in At(\mathrm{Spin}(H))$ and $e_3 \perp e_4$, $e_k \top e_\ell$, $(k < \ell,\ (k,\ell) \neq (3,4))$.

In the cases of types 5–6 the standard grid of the unit matrices contains 8 atoms spanning a spin factor (type 4) of 8 dimensions. So as in the previous case, again there are atoms $e_1, \dots, e_4$ with $e_3 \perp e_4$, $e_k \top e_\ell$ $(k < \ell,\ (k,\ell) \neq = (3,4))$.

LEMMA 2.5 *If G is a Cartan factor then the atomic part of G^{**} is a copy of G.*

PROOF. We have $G^{**} = G_n^{**} \oplus \bigoplus_{j \in J} G_j$ where G_n^{**} is a non-atomic JBW*-triple and each G_j is a Cartan factor. Also there is an isometric JB*-homomorphism $U : G \to G^{**}$ onto some weak*-closed JB*-subtriple of G^{**}. Let π_j denote the canonical projection $G^{**} \to G_j$ and consider the representation $U_j := \pi_j U$ of G. The kernel K_j of U_j is a weak*-closed ideal in G. Since G is a factor, we have either $K_j = \{0\}$ or $K_j = G$. Since UG is weak*-dense in G^{**}, necessarily $U_j G \neq \{0\}$ and this excludes the possibility of $K_j = G$. Thus $K_j = \{0\}$, that is, the JB*-homomorphism U_j is injective. By a theorem of Horn–Dang–Neher on normal representations [10], injective JB*-homomorphisms are isometries. Thus $U_j G$ is a copy of G lying weak*-dense in the Cartan factor G_j. This is possible only if $U_j G = G_j$ and $U : G \leftrightarrow G_j$ is a JB*-isomorphism. By writing π for the canonical projection

$G^{**} \to \bigoplus_{j\in J} G_j$, it follows that πU is not weak*-dense in $\bigoplus_{j\in J} G_j$ unless the index set J is a singleton. ∎

PROOF OF COROLLARY 1.2. Let E be a TRO, G a Cartan factor and consider a JB*-homomorphism $T : E \to G$. It is well-known that the bidual operator $T^{**} : E^{**} \to G^{**}$ is also a JB*-homomorphism. We have $E^{**} = E_n^{**} \oplus \bigoplus_{i\in I} E_i$ where each term E_i is a Cartan factor and E_n^{**} is a non-atomic JBW*-triple. By the previous lemma, we may assume that $G^{**} = G_{\text{at}}^{**} \oplus G$ and, with the canonical projection $\pi : G^{**} \to G$, the operator πT^{**} is a JB*-homomorphism $E^{**} \to G$ which maps E onto a weak*-closed subtriple of G. Since πT^{**} is weak*-continuous, it follows that $\pi T^{**}E^{**} = = G$. The kernel K of the operator πT^{**} is a weak*-closed ideal of E^{**}. It is well known [1, 1985] that $E^{**} = K \oplus K^{\perp}$ where $K^{\perp} := \{x \in E^{**} : \{efx\} = = 0, \ \ e,f \in K\}$ is a weak*-closed ideal in E^{**}. Moreover, πT^{**} is an isometry on $K^{\perp}$ because injective JB*-homomorphisms are isometric [10]. Since $G = \pi T^{**}E = \pi T^{**}K^{\perp}$, the weak*-closed ideal $K^{\perp}$ must be a copy of the Cartan factor G. Hence $K^{\perp}$ is a minimal weak*-closed ideal in E^{**} and so $G \simeq K^{\perp} = E_i$ for some $i \in I$. By the theorem, each factor E_i is of type 1, hence so must be G. ∎

PROOF FOR REMARK 1.3. Let $e^{k\ell}$ denote the $n \times n$-matrix with 1 at the position (k,ℓ) and with 0 at other entries and let $s^{k\ell}$ be the symmetric $(2n) \times (2n)$-matrix with 1 at the positions $(2k-1, 2\ell)$ and $(2\ell, 2k-1)$ and 0 elsewhere. It is straightforward to verify that the linear extension T of the map $[e^{k\ell} \mapsto s^{k\ell} : \ 1 \le k,\ell \le n]$ satisfies the identity $T(xy^*z + zy^*x) = = (Tx)(Ty)^*(Tz) + (Tz)(Ty)^*(Tx)$ (by checking it for $n := 3$ and the unit matrices without loss of generality). ∎

References

[1] T. BARTON and R. TIMONEY: Weak*-continuity of Jordan triple products and applications, *Math. Scand.* **59** (1985), 177–191.

[2] T. DANG and Y. FRIEDMANN: Classification of atomic JBW*-triples and applications, *Math. Scan.* **61** (1987), 292–330.

[3] E. EFFROS, N. OZAWA and Z. RUAN: On injectivity and nuclearity for operator spaces, *Duke Math. J.* **110(3)** (2001), 489–521.

[4] Y. FRIEDMANN and B. RUSSO: Structure of the predual of a JBW*-triple, *J. Reine Angew. Math.* **356** (1985), 67–89.

[5] Y. FRIEDMANN and B. RUSSO: The Gelfand–Naimark theorem for JB*-triples, *Duke Math. J.* **53** (1986), 139–148.

[6] M. HAMANA: Triple envelops and Silov boundaries of operator spaces, *Math. J. Toyama Univ.* **22** (1999), 77–93.

[7] L. A. HARRIS: Bounded symmetric homogeneous domains in infinite dimensional spaces, in: *Proceedings on Infinite dimensional Holomorphy, Lecture Notes in Mathematics* **364** (1973), 13–40, Springer-Verlag, Berlin, 1973.

[8] L. A. HARRIS: A generalization of C*-algebras, *Proc. London Math. Soc.* **42(3)** (1981), 331–361.

[9] M. R. HESTENES: A ternary algebra with applications to matrices and linear transformations, *Arch. Rational Mech. Anal.* **11** (1962), 138–194.

[10] T. BARTON, T. DANG and G. HORN: Normal representations of Banach Jordan triple systems, *Proc. Amer. Math. Soc.* **102** (1987), 551–555.

[11] M. KAUR and Z. RUAN: Local properties of ternary rings of operators and their linking C*-algebras, *J. Funct. Anal.* **195** (2002), 262–305.

[12] E. NEHER: *Jordan triple systems by the grid approach*, Lecture Notes in Mathematics, Vol 1280, Springer-Verlag, Berlin, 1987.

[13] H. ZETTL: A characterization of ternary rings of operators, *Adv. in Math.* **48** (1983), 117–143.

José M. Isidro
Facultad de Matemáticas
Universidad de Santiago
Santiago de Compostela
Spain
jmisidro@zmat.usc.es

László L. Stachó
Bolyai Institute
Aradi Vértanúk tere 1
6720 Szeged
Hungary
stacho@math.u-szeged.hu

AN INEQUALITY BETWEEN THE MEASURES OF PSEUDORANDOMNESS

By

KATALIN GYARMATI

(May 14, 2004)

1. Introduction

In this paper I will improve on a generalization of an inequality of Mauduit and Sárközy [6]. They introduced the following measures of pseudorandomness in [5]:

For a binary sequence

$$E_N = \{e_1, \dots, e_N\} \in \{-1, +1\}^N,$$

write

$$U(E_N, t, a, b) = \sum_{j=1}^{t} e_{a+jb}$$

and, for $D = (d_1, \dots, d_k)$ with non-negative integers $0 \le d_1 < \dots < d_k$,

$$V(E_N, M, D) = \sum_{n=1}^{M} e_{n+d_1}, \dots, e_{n+d_k}.$$

Then the *well-distribution measure of E_N* is defined as

$$W(E_N) = \max_{a,b,t} |U(E_N, t, a, b)| = \max_{a,b,t} \left| \sum_{j=1}^{t} e_{a+jb} \right|,$$

where the maximum is taken over all a, b, t such that $a \in \mathbb{Z}$, $b, t \in \mathbb{N}$ and $1 \le a + b \le a + tb \le N$, while the *correlation measure of order k of E_N* is defined as

$$C_k(E_N) = \max_{M,D} |V(E_N, M, D)| = \max_{M,D} \left| \sum_{n=1}^{M} e_{n+d_1} \cdots e_{n+d_k} \right|,$$

where the maximum is taken over all $D = (d_1, \dots, d_k)$ and M such that $M + d_k \leq N$.

In [6] Mauduit and Sárközy proved that for all sequences $E_N \in \{-1,+1\}^N$ we have $W(E_N) \leq 3\sqrt{NC_2(E_N)}$. Later in [3] this inequality was generalized by me to correlation measure of any even order: If $3\ell^2 \leq N$ and $E_N \in \{-1,+ +1\}^N$ then $W(E_N) \leq 3\ell N^{1-1/(2\ell)} \left(C_{2\ell}(E_N)\right)^{1/(2\ell)}$. In the present paper I will improve on the factor 3ℓ showing that this inequality even holds with an absolute constant factor:

THEOREM 1. *If $\varepsilon > 0$, $N \geq 18\ell/\varepsilon^2$, then for all $E_N \in \{-1,+1\}^N$ we have*

$$W(E_N) \leq (\sqrt{2} + \varepsilon) N^{1-1/(2\ell)} C_{2\ell}(E_N)^{1/(2\ell)}.$$

Mauduit and Sárközy [6] also proved that their inequality is sharp by using probabilistic arguments. In [3] I presented an explicit construction for which the generalized inequality is sharp apart from a $\sqrt{\ell}$ factor. This construction was based on the notion of index (discrete logarithm): Denote ind$n$ the index of n modulo p, defined as the unique integer with

$$g^{\mathrm{ind}n} \equiv n \pmod{p},$$

and $1 \leq \mathrm{ind}n \leq p-1$, where g is a fixed primitive root modulo p. Let ind$^*n$ be the modulo m residue of ind n:

$$\mathrm{ind}^*n \equiv \mathrm{ind}n \pmod{m} \tag{1}$$

with $1 \leq \mathrm{ind}^*n \leq m$.

CONSTRUCTION 1. *Let $m \mid p-1$ and ind^*n be the function defined by (1). Then let the sequence $E_{p-1} = \{e_1, \dots, e_{p-1}\}$ be*

$$e_n = \begin{cases} +1 & \text{if } 1 \leq \mathrm{ind}^* f(n) \leq \frac{m}{2}, \\ -1 & \text{if } \frac{m}{2} < \mathrm{ind}^* f(n) \leq m \text{ or } p \mid f(n), \end{cases} \tag{2}$$

where $f(x) \in \mathbb{F}_p[x]$ is a polynomial with the degree k.

In Theorem 1 and 3 in [3] I gave estimates for the well-distribution measure and correlation measures of this sequence E_{p-1} if some, not too restrictive conditions hold on the polynomial $f(x)$. Then

$$W(E_{p-1}) \gg \frac{1}{\sqrt{\ell k^{\ell+1}}} p^{1-1/(2\ell)} \left(C_{2\ell}(E_{p-1})\right)^{1/(2\ell)} \tag{3}$$

follows from these theorems, where the implied constant factor is absolute.

This inspired me to consider the simplest polynomial $f(x) = x$ in Construction 1, hoping that inequality (3) holds with a factor larger than $\frac{1}{\sqrt{\ell}}$. Indeed we will study the following sequence:

CONSTRUCTION 2. *Let $m \mid p-1$ and $\operatorname{ind}^* n$ be the function defined by (1). Then let the sequence $E_{p-1} = \{e_1, \dots, e_{p-1}\}$ be*

$$e_n = \begin{cases} +1 & \text{if } 1 \le \operatorname{ind}^* n \le \frac{m}{2}, \\ -1 & \text{if } \frac{m}{2} < \operatorname{ind}^* n \le m. \end{cases} \tag{4}$$

For this sequence we have:

THEOREM 2. *If m is even then the sequence in Construction 2 satisfies*

$$W(E_{p-1}) \le 36 p^{1/2} \log p \log(m+1)$$

while for odd m we have

$$W(E_{p-1}) = \frac{p-1}{m} + O\big(p^{1/2} \log p \log(m+1)\big).$$

Indeed, this is Theorem 1 in [3] in the special case when k, the degree of the polynomial is 1.

In case of the correlation measure we will give slightly better upper bound than in Theorem 3 (in the special case $k = 1$) in [3]:

THEOREM 3. *If m is even then the sequence in Construction 2 satisfies:*

$$C_\ell(E_{p-1}) \le 9\ell 4^\ell p^{1/2} \log p \,(\log m)^\ell,$$

while for odd m we have

$$C_\ell(E_{p-1}) = \frac{p}{m^\ell} + O\big(5^\ell p^{1/2} \log p (\log m)^\ell\big).$$

It follows from Theorems 2 and 3:

COROLLARY 1. *For every $\varepsilon > 0$ there exist positive constants $p_0(\varepsilon)$ and $c_0(\varepsilon)$ such that if $p > p_0(\varepsilon)$ and m is an odd divisor of $p-1$ with*

$$m < c_0(\varepsilon)\frac{p^{1/(2\ell)}}{\ell(\log p)^{1+1/\ell}} \tag{5}$$

(so $\frac{p}{m^\ell} \gg 5^\ell p^{1/2}\log p\,(\log m)^\ell$), then

$$W(E_{p-1}) \geq (1-\varepsilon)p^{1-1/(2\ell)}\left(C_{2\ell}(E_{p-1})\right)^{1/(2\ell)}. \tag{6}$$

I remark that to make sure that condition (5) holds, first we fix an odd integer m, and after this we look for a prime number p with $m \mid p-1$ and (5). This is possible by Dirichlet's theorem on primes in arithmetic progressions.

So, indeed Theorem 1 is best possible apart from a constant factor. The interesting feature of this proof is that it is explicit, we give a sequence for which (6) holds. In the most cases there is only an existence proof for the sharpness of an inequality between pseudorandom measures.

2. Proofs of Theorem 1 and 3

PROOF OF THEOREM 1. It follows from the definition of $W(E_N)$ that there exist $a \in \mathbb{Z}$, $b, t \in \mathbb{N}$ with $1 \leq a+b < a+tb \leq N$ such that

$$W(E_N) = \left|\sum_{\substack{a+b\leq i\leq a+tb\\ i\equiv a+b \pmod b}} e_i\right|. \tag{7}$$

For $0 \leq h < b$ let

$$D_h \overset{\text{def}}{=} \Bigg(\sum_{\substack{a+b\leq i\leq a+tb\\ i\equiv h \pmod b}} e_i\Bigg)^{2\ell} - 2\ell! \sum_{\substack{a+b\leq i_1<\cdots<i_{2\ell}\leq a+tb\\ h\equiv i_1\equiv\cdots\equiv i_{2\ell} \pmod b}} e_{i_1}\cdots e_{i_{2\ell}}. \tag{8}$$

Using the multinomial theorem we get that D_h is a sum of products of the form $c\cdot e_{j_1}\cdots e_{j_r}$ where $c \geq 0$. Thus D_h takes his maximum when all e_i's are +1 (or all e_i's are -1). So:

$$D_h \leq \Bigg(\sum_{\substack{a+b\leq i\leq a+tb\\ i\equiv h \pmod b}} 1\Bigg)^{2\ell} - 2\ell! \sum_{\substack{a+b\leq i_1<\cdots<i_{2\ell}\leq a+tb\\ h\equiv i_1\equiv\cdots\equiv i_{2\ell} \pmod b}} 1$$

$$\leq t^{2\ell} - (t-1)(t-2)\cdots(t-2\ell) \leq t^{2\ell} - (t-2\ell)^{2\ell} \leq 4\ell^2 t^{2\ell-1}.$$

By this, (7) and (8) we have

$$
\begin{aligned}
(W(E_N))^{2\ell} &\le \sum_{h=0}^{b-1} \Big(\sum_{\substack{a+b\le i\le a+tb\\ i\equiv h \pmod b}} e_i \Big)^{2\ell}\\
&= \sum_{h=0}^{b-1} \Big(D_h + 2\ell! \sum_{\substack{a+b\le i_1<\cdots<i_{2\ell}\le a+tb\\ h\equiv i_1\equiv\cdots\equiv i_{2\ell} \pmod b}} e_{i_1}\cdots e_{i_{2\ell}} \Big)\\
&\le \sum_{h=0}^{b-1} \Big(4\ell^2 t^{2\ell-1} + 2\ell! \sum_{\substack{a+b\le i_1<\cdots<i_{2\ell}\le a+tb\\ h\equiv i_1\equiv\cdots\equiv i_{2\ell} \pmod b}} e_{i_1}\cdots e_{i_{2\ell}} \Big)\\
&= 4b\ell^2 t^{2\ell-1} + 2\ell! \sum_{\substack{a+b\le i_1<\cdots<i_{2\ell}\le a+tb\\ i_1\equiv\cdots\equiv i_{2\ell} \pmod b}} e_{i_1}\cdots e_{i_{2\ell}}.
\end{aligned}
$$

From this replacing i_2 by $i_1 + d_1$, i_3 by $i_1 + d_2$ and so on, finally $i_{2\ell}$ by $i_1 + d_{2\ell-1}$ we obtain

$$
\begin{aligned}
&(W(E_N))^{2\ell} \le 4b\ell^2 t^{2\ell-1} + 2\ell!\\
&\qquad \sum_{\substack{1\le d_1<\cdots<d_{2\ell-1}\le(t-1)b\\ d_1\equiv\cdots\equiv d_{2\ell-1}\equiv 0 \pmod b}} \ \sum_{i_1=a+b}^{a+tb-d_{2\ell-1}} e_{i_1}e_{i_1+d_1}\cdots e_{i_1+d_{2\ell-1}}. \qquad (9)
\end{aligned}
$$

By the definition of the correlation measure we have

$$
\left| \sum_{i_1=a+b}^{a+tb-d_{2\ell-1}} e_{i_1}e_{i_1+d_1}\cdots e_{i_1+d_{2\ell-1}} \right| \le C_{2\ell}E_N. \qquad (10)
$$

By $tb \le a + tb \le N$ we have $4b\ell^2 t^{2\ell-1} = 4\ell^2(tb)t^{2\ell-2} \le 4\ell^2 N^{2\ell-1}$, and so from (9) and (10) we obtain

$$
\begin{aligned}
(W(E_N))^{2\ell} &\le 4\ell^2 N^{2\ell-1} + 2\ell! \frac{N^{2\ell-1}}{(2\ell-1)!} C_{2\ell}(E_N)\\
&= 2\ell \left(1 + \frac{2\ell}{C_{2\ell}(E_N)} \right) N^{2\ell-1} C_{2\ell}(E_N).
\end{aligned}
$$

From this by the binomial theorem we get:

$$W(E_N) \le (2\ell)^{1/(2\ell)} \left(1 + \frac{1}{C_{2\ell}(E_N)}\right) N^{1-1/(2\ell)} \left(C_{2\ell}(E_N)\right)^{1/(2\ell)}.$$

Kohayakawa, Mauduit, Moreira and V. Rödl [4] proved that $C_{2\ell}(E_N) >$ $> \sqrt{\frac{N}{3(2\ell+1)}}$ holds for all $E_N \in \{-1,+1\}^N$ by this and since $(2\ell)^{1/(2\ell)} \le \sqrt{2}$ we get:

$$W(E_N) \le \sqrt{2} \left(1 + \sqrt{\frac{3(2\ell+1)}{N}}\right) N^{1-1/(2\ell)} \left(C_{2\ell}(E_N)\right)^{1/(2\ell)}.$$

If $N \ge 18\ell/\varepsilon^2 \ge 6(2\ell+1)/\varepsilon^2$ then this completes the proof of the theorem.

Proof of Theorem 3. The proof of the theorem is very similar to the proof of Theorem 1 in [2]. By the formula

$$\frac{1}{m} \sum_{\chi:\chi^m=1} \overline{\chi}^j(a)\chi(b) = \begin{cases} 1 & \text{if } m \mid \operatorname{ind} a - \operatorname{ind} b, \\ 0 & \text{if } m - \operatorname{ind} a - \operatorname{ind} b, \end{cases}$$

we obtain

$$e_n = 2 \sum_{\substack{1 \le i \le m/2 \\ i \equiv \operatorname{ind} n \pmod m}} 1 - 1 = \frac{2}{m} \sum_{1 \le i \le m/2} \sum_{\chi:\chi^m=1} \overline{\chi}(n)\chi(g^i) - 1.$$

Thus

$$e_n = \frac{2}{m} \left(\sum_{1 \le i \le m/2} \sum_{\chi \ne \chi_0:\chi^m=1} \overline{\chi}(n)\chi(g^j) + \frac{(-1)^m - 1}{4} \right). \tag{11}$$

To prove Theorem 3, consider any $\mathcal{D} = \{d_1, d_2, \dots, d_\ell\}$ with non-negative integers $d_1 < d_2 < \dots < d_\ell$ and positive integer M with $M + d_\ell \le p - 1$. Then arguing as in [7, p. 382] with m in place of $p-1$ from (11) we obtain:

$$V(E_N, M, D) = \frac{2^\ell}{m^\ell} \sum_{n=1}^{M} \prod_{j=1}^{\ell} \left(\sum_{1 \le i \le m/2} \sum_{\substack{\chi_j \ne \chi_0, \\ \chi_j^m = 1}} \overline{\chi_j}(n + d_j)\chi_j(g^i) + \frac{(-1)^m + 1}{4} \right)$$

$$= \frac{2^\ell}{m^\ell}\left(\sum_{k=0}^{\ell}\sum_{1\le j_1<\cdots<j_k\le \ell}\left(\frac{(-1)^m+1}{4}\right)^{\ell-k}\sum_{\substack{\chi_{j_1}\ne\chi_0,\\ \chi_{j_1}^m=1}}\cdots\sum_{\substack{\chi_{j_k}\ne\chi_0,\\ \chi_{j_k}^m=1}}\right.$$

$$\left.\sum_{n=1}^{M}\overline{\chi}_{j_1}(n+d_{j_1})\cdots\overline{\chi}_{j_k}(n+d_{j_k})\prod_{t=1}^{k}\left(\sum_{1\le \ell_t\le m/2}\chi_{j_i}(g^{\ell_t})\right)\right). \tag{12}$$

Let $S_0 = M$, $V_0 = \left(\frac{1}{2}\right)^\ell$ and for $1 \le k \le \ell$ let

$$S_k = \max_{\substack{\chi_1\ne\chi_0,\ldots,\chi_k\ne\chi_0\\ 1\le j_1<\cdots<j_k\le\ell}}\left|\sum_{n=1}^{M}\overline{\chi}_1\left(n+d_{j_1}\right)\cdots\overline{\chi}_k\left(n+d_{j_k}\right)\right| \tag{13}$$

and

$$V_k = \sum_{1\le j_1<\cdots<j_k\le\ell}\left(\frac{1}{2}\right)^{\ell-k}\sum_{\substack{\chi_{j_1}\ne\chi_0,\\ \chi_{j_1}^m=1}}\cdots\sum_{\substack{\chi_{j_k}\ne\chi_0,\\ \chi_{j_k}^m=1}}\prod_{t=1}^{k}\left|\sum_{1\le\ell_t\le m/2}\chi_{j_i}(g^{\ell_t})\right|. \tag{14}$$

Then by the triangle-inequality, the value of $\frac{(-1)^m+1}{4}$ and (12) we obtain that if m is even then

$$|V(E_N, M, D)| \le \frac{2^\ell}{m^\ell}S_\ell V_\ell \tag{15}$$

and

$$V(E_N, M, D) = \frac{2^\ell}{m^\ell}S_0V_0 + O\left(\frac{2^\ell}{m^\ell}\sum_{k=1}^{\ell}S_k V_k\right) \tag{16}$$

Next we give an upper bound for S_k. In order to do this we will use the following lemma:

LEMMA 1. *Suppose that p is a prime, χ is a non-principal character modulo p of order z, $f \in \mathbb{F}_p[x]$ has s distinct roots in $\overline{F_p}$, and it is not a constant multiple of a z-th power of a polynomial over $\mathbb{F}_p$. Let y be a real number with $0 < y \le p$. Then for any $x \in \mathbb{R}$:*

$$\left|\sum_{x<n\le x+y}\chi(f(n))\right| < 9sp^{1/2}\log p.$$

POOF OF LEMMA 1. This is a trivial consequence of Lemma 1 in [1]. Indeed, there this result is deduced from Weil's theorem, see [8].

Now let χ be a modulo p character of order m; for simplicity we will choose χ as the character uniquely defined by $\chi(g) = e\left(\frac{1}{m}\right)$.

Returning to the estimate of S_k, let $\overline{\chi_u} = \chi^{\delta_u}$ for $u = 1, 2, \dots, \ell$, whence by $\chi_1 \neq \chi_0, \dots, \chi_\ell \neq \chi_0$, we may take

$$1 \le \delta_u < m.$$

Thus in (13) we have

$$\begin{aligned}\left|\sum_{n=1}^{M} \overline{\chi}_1(n+d_{j_1}) \cdots \overline{\chi}_k(n+d_{j_k})\right| &= \left|\sum_{n=1}^{M} \chi^{\delta_1}(n+d_{j_1}) \cdots \chi^{\delta_\ell}(n+d_{j_k})\right| \\ &= \left|\sum_{n=1}^{M} \chi\left((n+d_{j_1})^{\delta_1} \cdots (n+d_{j_k})^{\delta_k}\right)\right|.\end{aligned}$$

Since $(n+d_{j_1})^{\delta_1} \cdots (n+d_{j_k})^{\delta_k}$ is not a perfect m-th power, this sum can be estimated by Lemma 1, whence

$$S_k \le 9kp^{1/2} \log p. \tag{17}$$

By (14) we have

$$V_k = \sum_{1 \le j_1 \le \cdots \le j_k \le \ell} \left(\frac{1}{2}\right)^{\ell-k} \left(\sum_{\substack{\chi \neq \chi_0, \\ \chi^m = 1}} \left|\sum_{j=1}^{[m/2]} \chi^j(g)\right|\right)^k.$$

LEMMA 2.

$$\sum_{\substack{\chi \neq \chi_0, \\ \chi^m = 1}} \left|\sum_{j=1}^{[m/2]} \chi^j(g)\right| \le \sum_{\substack{\chi \neq \chi_0, \\ \chi^m = 1}} \frac{2}{|1-\chi(g)|} < 2m \log(m+1).$$

PROOF OF LEMMA 2. This is Lemma 3 in [2] with m in place of d and $m/2$ in place of $(p-1)/2$, and it can be proved in the same way.

Using Lemma 2 we obtain

(18)
$$V_k \le \sum_{1\le j_1\le\cdots\le j_k\le \ell} \left(\frac{1}{2}\right)^{\ell-k} \left(2m\,(\log(m+1))^k\right) = \frac{4^k}{2^\ell}\binom{\ell}{k} m^k\,(\log(m+1))^k\,.$$

By (15), (16), (17) and (18) we obtain that if m is even then

$$|\,V(E_N, M, D)| \le 9\ell 4^\ell p^{1/2} \log p\,(\log(m+1))^\ell\,,$$

and if m is odd then

$$\begin{aligned} V(E_N, M, D) &= \frac{M}{m^\ell} + O\left(\frac{9p^{1/2}\log p}{m^\ell}\sum_{k=1}^{\ell} k\binom{\ell}{k} 4^k m^k\,(\log(m+1))^k\right) \\ &= \frac{M}{m^\ell} + O\left(\frac{9\ell p^{1/2}\log p}{m^\ell}\,(4m\log(m+1))^\ell\right) \\ &= \frac{M}{m^\ell} + O\left(5^\ell p^{1/2}\log p\,(\log(m+1))^\ell\right), \end{aligned}$$

which completes the proof of the theorem.

References

[1] R. Ahlswede, C. Mauduit and A. Sárközy: Large families of pseudorandom sequences of k symbols and their complexity, Part I, Part II.; *Proceedings on General Theory of Information Transfer and Combinatorics*, to appear.

[2] K. Gyarmati: On a family of pseudorandom binary sequences, *Periodica Math. Hungar.*, to appear.

[3] K. Gyarmati: On a fast version of a pseudorandom generator, *General Theory of Information Transfer and Combinatorics*, Conference Proceedings, to appear.

[4] Y. Kohayakawa, C. Mauduit, C. G. Moreira and V. Rödl: Measures of pseudorandomness for finite sequences: minimum and typical values, submitted to *J. London Math. Soc.*

[5] C. Mauduit and A. Sárközy: On finite pseudorandom binary sequences I: Measures of pseudorandomness, the Legendre symbol; *Acta Arithmetica* **82** (1997).

[6] C. Mauduit and A. Sárközy: On the measures of pseudorandomness of binary sequences, *Discrete Math.* **271** (2003), 195–207.

[7] A. Sárközy: A finite pseudorandom binary sequence, *Studia Scientiarum Mathematicarum Hungarica* **38** (2001), 377–384.

[8] A. WEIL: Sur les courbes algébriques et les variétés qui s'en déduisent, *Act. Sci. Ind.* **1041**, Hermann, Paris, 1948.

Katalin Gyarmati
ELTE TTK

gykati@cs.elte.hu

THE ω_1-LIMIT OF BAIRE-2 FUNCTIONS IS BAIRE-2

By

TAMÁS MÁTRAI*

(May 21, 2004)

1. Introduction

Almost a century ago, W. Sierpiński [6] observed that the pointwise limit of a sequence with length ω_1 of continuous real functions is necessarily continuous (Theorem 1 on page 133), which may seem to be quite paradox compared to the behavior of ordinary pointwise convergence. In the same paper, Sierpiński has also proved this class preserving property of ω_1-convergence for the Baire-1 functions (Theorem 2 on page 137); and he pointed out that by assuming the Continuum Hypothesis, every real function can be obtained as the ω_1-limit of Baire-2 functions (for more details and discussions, see [6], Section 5, page 139 and [2], specially Theorem 3 on page 499).

In view of these facts, T. Natkaniec [5] introduced a stronger notion of pointwise convergence. We recall the precise setting in the following definition.

Definition 1. Let (X,τ) be a Polish space, (Y,d) be a separable metric space and consider an ideal $\mathcal{I}$ on ω_1. We say that a sequence of functions $f_\alpha : X \to Y$ $(\alpha < \omega_1)$ *$\mathcal{I}$-converges* to the function $f: X \to Y$, in notation $f_\alpha \to_{\mathcal{I}} f$, if

$$\{\alpha < \omega_1 : f_\alpha(x) \neq f(x)\} \in \mathcal{I}$$

for every $x \in X$.

Similarly, we write $f_\alpha \underset{\mathcal{I}}{\to_d} f$ if for every $\varepsilon > 0$ and $x \in X$ we have

$$\{\alpha < \omega_1 : d(f(x), f_\alpha(x)) > \varepsilon\} \in \mathcal{I}.$$

* This research was supported by the OTKA grant F 043620.

In case of the ordinary ω_1 convergence, as used in [2] and [6], we have $\mathcal{I} = [\omega_1]^{\omega}$, that is the ideal of countable subsets of ω_1. However, our motivating theorem, giving partial answer to Problem 1 in [5] on page 490, is related to the particular case when the ideal contains the finite subsets of ω_1, that is $\mathcal{I}_< = [\omega_1]^{<\omega}$.

THEOREM 2. *Let (X, τ) be a Polish space, (Y, d) be a separable metric space, and consider a family $f_\alpha : X \to Y$ $(\alpha < \omega_1)$ of Baire-2 functions. If $f : X \to Y$ is such that $f_\alpha \to_{\substack{d \\ \mathcal{I}_<}} f$, then f is Baire-2.*

We note here that the original question asked by T. Natkaniec refered to $\mathcal{I}_<$-convergence. However, it is easy to see that $\mathcal{I}_<$-convergence implies $\substack{d \\ \mathcal{I}_<}$-convergence, so the result above is formally stronger than the required. The sufficiency of $\substack{d \\ \mathcal{I}_<}$-convergence was pointed out to the author by Petr Holický. We also note that using more sophisticated techniques, this result has already been generalized to every Baire-ξ class (see [4]).

In the following section, we present the characterization of $\Sigma^0_3(\tau)$ sets which is the key element of the proof of Theorem 2. The last section contains the proof of Theorem 2.

Our reference for the basic notions in descriptive set theory is [1]; in particular, $\Pi^0_\xi(\tau)$ ($\Sigma^0_\xi(\tau)$ resp.) stands for the ξ^{th} multiplicative (additive resp.) Borel class in (X, τ), starting with $\Pi^0_1(\tau)$ = closed sets, $\Sigma^0_1(\tau)$ = open sets.

2. $\Sigma^0_3(\tau)$ sets in the Borel hierarchy

Let (C, τ_C) denote the Polish space 2^ω with its usual product topology. To distinguish $\Sigma^0_3(\tau)$ sets from $\Pi^0_3(\tau)$ sets, we construct a $\Pi^0_3(\tau_C)$ set $\mathcal{P} \subseteq C$ such that every $\Sigma^0_3(\tau)$ subset of X containing a suitable copy $\varphi(\mathcal{P})$ of $\mathcal{P}$ is "much bigger" in sense of Baire category than $\varphi(\mathcal{P})$ (for the precise statement, see Lemma 3).

First we have to construct $\mathcal{P}$. The method had already been used by Lusin to build a proper $\Pi^0_3(\tau_C)$ set and was communicated to the author by Petra Šindelářová. Following [1], for two finite sequences $s, t \in \omega^{<\omega}$, we write $s < t$ if t is a (proper) extension of s. The length of s is denoted by $|s|$. If $s = s_1 s_2 \ldots s_n$ and $i \in \mathbb{N}$, then $s^\frown i$ stands for the sequence $s_1 s_2 \ldots s_n i$.

For every finite sequence $s \in \omega^{<\omega}$, fix a nonempty perfect set $P_s \subseteq C$ with the following properties:

$$P_\emptyset = C; \tag{1}$$

$$\text{for } t < s,\ P_s \subseteq P_t \text{ and } P_s \text{ is nowhere dense in } (P_t, \tau_C|_{P_t}); \tag{2}$$

$$\bigcup_{i \in \mathbb{N}} P_{s^\frown i} \text{ is dense in } (P_s, \tau_C|_{P_s}). \tag{3}$$

To have $P_{s^\frown i} \subseteq P_s$ $(i \in \mathbb{N})$, one simply has to take a countable dense subset $D_s = \{d_1, d_2, \dots\} \subseteq P_s$ and cover successively every d_i with a perfect set $P_{s^\frown i}$ which is nowhere dense in $(P_s, \tau_C|_{P_s})$. Then (1), (2) and (3) are obviously satisfied. Once this done, let

$$\mathcal{P} = \bigcap_{n=0}^{\infty} \bigcup_{\substack{s \in \omega^{<\omega} \\ |s|=n}} P_s. \tag{4}$$

Now we can formulate our characterization.

Lemma 3. *Let (X,τ) be a Polish space, $A \subseteq (X,\tau)$ be a Borel set.*

1. *If A is $\Sigma^0_3(\tau)$, then whenever for a continuous one-to-one map*
$$\varphi : (C, \tau_C) \to (X, \tau)$$
we have $\varphi(\mathcal{P}) \subseteq A$, then there is an $s \in \omega^{<\omega}$ for which $A \cap \varphi(P_s)$ is of the second category in $(\varphi(P_s), \tau|_{\varphi(P_s)})$.
2. *If A is not $\Sigma^0_3(\tau)$, then there is a continuous one-to-one map*
$$\varphi : (C, \tau_C) \to (X, \tau)$$
such that $\varphi(\mathcal{P}) \subseteq A$ and $A \cap \varphi(P_s)$ is meager in $(\varphi(P_s), \tau|_{\varphi(P_s)})$ for every $s \in \omega^{<\omega}$.

The statements involving Baire category do make sense since $\varphi(P_s)$, as a continuous image of the compact set P_s, is closed in the Polish space (X,τ), so itself is Polish with the restricted topology $\tau|_{\varphi(P_s)}$ (see e.g. [1], Proposition 3.3.(ii) on page 13). To prove Lemma 3, we will use the following result (see e.g. [3], page 433). In some sense, Lemma 3 is a quantitative analogue of this result in the special $\xi = 3$ case.

Theorem 4 (A. Louveau, J. Saint Raymond). *Let $3 \le \xi < \omega_1$ and (X,τ) be a Polish space. If $P_\xi \subseteq C$ is $\Pi^0_\xi(\tau_C)$ but not $\Sigma^0_\xi(\tau_C)$ and $A_0, A_1 \subseteq X$*

is any pair of disjoint Borel sets, then either A_0 can be separated from A_1 by a $\Sigma^0_\xi(\tau)$ set or there is a continuous one-to-one map $\varphi\colon(C,\tau_C)\to(X,\tau)$ with $\varphi(P_\xi)\subseteq A_0$ and $\varphi(C\setminus P_\xi)\subseteq A_1$.

Before giving the proof of Lemma 3, we make two easy observations.

LEMMA 5.

1. $\mathcal{P}\subseteq C$ is a $\Pi^0_3(\tau_C)$ set.

2. $\mathcal{P}\cap P_s$ is dense and meager in $(P_s,\tau_C|_{P_s})$ for every $s\in\omega^{<\omega}$.

PROOF. The first statement follows immediately from (4). To see that $\mathcal{P}\subseteq(P_{s_0},\tau_C|_{P_{s_0}})$ is dense for every $s_0\in\omega^{<\omega}$, take any nonempty closed ball $B_0\subseteq P_{s_0}$; we show that $B_0\cap\mathcal{P}\neq\emptyset$. We construct finite sequences $s_i\in\omega^{<\omega}$ $(i\in\mathbb{N})$ and a sequence of nonempty closed balls $B_i\subseteq(P_{s_i},\tau_C|_{P_{s_i}})$ $(i\in\mathbb{N})$ such that $s_i\leq s_j$ and $B_j\subseteq B_i$ for $0\leq i\leq j$. This proves the statement since such a $(B_i)_{i\in\mathbb{N}}$ is a nested sequence of nonempty closed sets in (C,τ_C), so

$$\bigcap_{i\in\mathbb{N}} B_i\subseteq B_0\cap\mathcal{P}$$

is nonempty.

Suppose that s_k and B_k have already been found. By (3), there is an $l\in\mathbb{N}$ such that for $s_{k+1}=s_k{}^\frown l$, $B_k\cap P_{s_{k+1}}\neq\emptyset$. Thus we can find a closed ball $B_{k+1}\subseteq(P_{s_{k+1}},\tau_C|_{P_{s_{k+1}}})$ contained in $B_k\cap P_{s_{k+1}}$, which completes the construction.

Finally, for every $s\in\omega^{<\omega}$, $\mathcal{P}\cap P_s$ is meager in $(P_s,\tau_C|_{P_s})$ since

$$\mathcal{P}\cap P_s\subseteq\left(\bigcup_{i\in\mathbb{N}}P_{s^\frown i}\right)\cap P_s \tag{5}$$

and by (2), the union on the right hand side of (5) is already meager in $(P_s,\tau_C|_{P_s})$. ∎

PROOF OF LEMMA 3. For the first statement, let $A\subseteq X$ be $\Sigma^0_3(\tau)$, $\varphi\colon(C,\tau_C)\to(X,\tau)$ be continuous, one-to-one, and suppose that $A\cap\varphi(P_s)$ is meager in $(\varphi(P_s),\tau|_{\varphi(P_s)})$ for every $s\in\omega^{<\omega}$. Then

$$A=\bigcup_{n\in\mathbb{N}}A_n$$

where the sets A_n $(n \in \mathbb{N})$ are $\Pi^0_2(\tau)$, and since in Polish spaces a $\Pi^0_2(\tau)$ set is meager if and only if it is nowhere dense, $A_n \cap \varphi(P_s)$ is nowhere dense in $(\varphi(P_s), \tau|_{\varphi(P_s)})$ for every $n \in \mathbb{N}$ and $s \in \omega^{<\omega}$. We define by induction finite sequences $s_n \in \omega^{<\omega}$ $(n \in \mathbb{N})$ and a corresponding sequence of closed balls $B_n \subseteq (X, \tau)$ $(n \in \mathbb{N})$ such that

(i) $|s_n| = n$ $(n \in \mathbb{N})$;

(ii) $s_n \leq s_m$ if $n \leq m$;

(iii) $B_m \subseteq B_n$ if $n \leq m$;

(iv) $B_n \cap \varphi(P_{s_{n+1}})$ $(n \in \mathbb{N})$ is nonempty and perfect;

(v) $B_n \cap \varphi(P_{s_n}) \cap A_n = \emptyset$ $(n \in \mathbb{N})$.

This completes the proof, since on one hand, by (2), (iii) and (iv), we have that $B_n \cap \varphi(P_{s_{n+1}})$ is a nested sequence of nonempty perfect sets, so

$$\mathcal{Q} = \bigcap_{n \in \mathbb{N}} B_n \cap \varphi(P_{s_{n+1}}) = \bigcap_{n \in \mathbb{N}} B_n \cap \varphi(P_{s_n})$$

is a nonempty subset of $\varphi(\mathcal{P})$, while on the other hand, by (v), $\mathcal{Q} \cap A = \emptyset$ since $\mathcal{Q} \cap A_n = \emptyset$ for every $n \in \mathbb{N}$, which contradicts $\varphi(\mathcal{P}) \subseteq A$.

Let $s_0 = \emptyset$ and suppose that s_i and B_{i-1} are found for $0 \leq i \leq n$ satisfying (i)–(v); we define B_n and s_{n+1}. Since $A_n \cap \varphi(P_{s_n})$ is nowhere dense in $(\varphi(P_{s_n}), \tau|_{\varphi(P_{s_n})})$, we can find a closed ball $B_n \subseteq B_{n-1}$ for which $B_n \cap \varphi(P_{s_n})$ is a nonempty perfect set and $B_n \cap \varphi(P_{s_n}) \cap A_n = \emptyset$; thus (iii) and (v) hold. By (3), we can find an $i \in \mathbb{N}$ such that $B_n \cap \varphi(P_{s_n{}^\frown i})$ is also nonempty and perfect. With $s_{n+1} = s_n{}^\frown i$, (i), (ii) and (iv) are satisfied, which completes the proof.

For the second statement, let $A \subseteq X$ be Borel but not $\Sigma^0_3(\tau)$. By the first part of the lemma for $(X, \tau) = (C, \tau_C)$ and $\varphi = \mathrm{Id}_C$, $\mathcal{P}$ is not $\Sigma^0_3(\tau_C)$ since by Lemma 5.2, it is meager in P_s for every $s \in \omega^{<\omega}$. Since $\mathcal{P}$ is $\Pi^0_3(\tau_C)$ by Lemma 5.1, we can apply Theorem 4 for $\xi = 3$, $P_3 = \mathcal{P}$, $A_0 = A$ and $A_1 = X \setminus A$.

The set A is not $\Sigma^0_3(\tau)$, so A_0 cannot be separated from A_1 by a $\Sigma^0_3(\tau)$ set. Thus we have a continuous one-to-one map

$$\varphi \colon (C, \tau_C) \to (X, \tau)$$

such that $\varphi(C) \cap A = \varphi(\mathcal{P})$; hence $\varphi(\mathcal{P}) \subseteq A$. We show that $A \cap \varphi(P_s)$ is meager in $(\varphi(P_s), \tau|_{\varphi(P_s)})$ for every $s \in \omega^{<\omega}$.

Take an $s \in \omega^{<\omega}$. Since φ is a continuous one-to-one map on the compact set P_s, it is a homeomorphism of $(P_s, \tau_C|_{P_s})$ and $(\varphi(P_s), \tau|_{\varphi(P_s)})$. We have

$$A \cap \varphi(P_s) = A \cap (\varphi(C) \cap \varphi(P_s)) = (A \cap \varphi(C)) \cap \varphi(P_s) = \varphi(\mathcal{P}) \cap \varphi(P_s).$$

Since homeomorphism preserves category, $A \cap \varphi(P_s) = \varphi(\mathcal{P}) \cap \varphi(P_s)$ is meager in $(\varphi(P_s), \tau|_{\varphi(P_s)})$ by Lemma 5.2. The proof is complete. ∎

3. $\mathcal{I}_<$-convergent functions

We will have to establish connection between function classes and sublevel sets. For this, we will use the following classical result (see e.g. [1], Chapter II, Theorem 24.3 on page 190).

THEOREM 6. *Let (X, τ) be a Polish space, (Y, d) be a separable metric space. Then for every $1 \leq \xi \leq \omega_1$, a function $f : X \to Y$ is Baire-ξ if and only if $f^{-1}(U) \subseteq X$ is $\Sigma^0_{\xi+1}(\tau)$ for every open set $U \subseteq Y$.*

In the metric space (Y, d), the open ball centered at $x \in Y$ with radius ρ is denoted by $B_d(x, \rho)$. Now we prove Theorem 2.

PROOF OF THEOREM 2. Let $f_\alpha \underset{\mathcal{I}_<}{\to_d} f$ for a family $f_\alpha : X \to Y$ $(\alpha < \omega_1)$ of Baire-2 functions and suppose that $f : X \to Y$ is not Baire-2. As the pointwise limit of the functions $\{f_\alpha : \alpha < \omega\}$, f is clearly Borel, so by Theorem 6, there is an open ball $B_d(x, \rho) \subseteq Y$ such that the $f^{-1}(B_d(x, \rho))$ is Borel but not $\Sigma^0_3(\tau)$. Set

$$H(\varepsilon) = f^{-1}(B_d(x, \rho - \varepsilon)), \ H_\alpha(\varepsilon) = f_\alpha^{-1}(B_d(x, \rho - \varepsilon))$$

for every $\alpha < \omega_1$ and $0 < \varepsilon < \rho$. Note that by Theorem 6, $H_\alpha(\varepsilon)$ is $\Sigma^0_3(\tau)$ for every $\alpha < \omega_1$ and $0 < \varepsilon < \rho$.

Since $H(0)$ is not $\Sigma^0_3(\tau)$, by Lemma 3.2 there is a continuous one-to-one map $\varphi : (C, \tau_C) \to (X, \tau)$ such that

(a) $\varphi(\mathcal{P}) \subseteq H(0)$, and

(b) $H(0) \cap \varphi(P_s)$ is meager in $(\varphi(P_s), \tau|_{\varphi(P_s)})$ for every $s \in \omega^{<\omega}$.

For $\varepsilon > 0$, let $J_1(\varepsilon)$ denote the set of those indices $\alpha < \omega_1$ for which $H_\alpha(\varepsilon)$ is of the second category in $(\varphi(P_s), \tau|_{\varphi(P_s)})$ for some $s \in \omega^{<\omega}$. We prove that $\omega_1 \setminus J_1(\varepsilon)$ is finite for every ε sufficiently small. Suppose that this is not true; take a positive sequence $(\varepsilon_i)_{i\in\mathbb{N}}$ with $\varepsilon_i \to 0$ and a countably infinite set $J'(\varepsilon_i) \subseteq \omega_1 \setminus J_1(\varepsilon_i)$ for every $i \in \mathbb{N}$. By the definition of $\frac{d}{\mathcal{I}_<}$ – convergence,

$$H(\varepsilon_i) \subseteq H'(\varepsilon_i) := \bigcup_{\alpha\in J'(\varepsilon_i)} H_\alpha(\varepsilon_i), \tag{6}$$

so by (a), we have that

$$\varphi(\mathcal{P}) \subseteq H(0) \subseteq \bigcup_{i\in\mathbb{N}} H(\varepsilon_i) \subseteq \bigcup_{i\in\mathbb{N}} H'(\varepsilon_i). \tag{7}$$

By (6), $H'(\varepsilon_i)$ $(i \in \mathbb{N})$ is $\Sigma^0_3(\tau)$, so by (7) we can apply Lemma 3.1 for $A = = \bigcup_{i\in\mathbb{N}} H'(\varepsilon_i)$. We obtain that A is of the second category in $(\varphi(P_s), \tau|_{\varphi(P_s)})$ for some $s \in \omega^{<\omega}$, which contradicts to the definition of $J_1(\varepsilon)$.

So there is an $\varepsilon_0 > 0$ such that $J_1(\varepsilon)$ is of cardinality ω_1 for every $\varepsilon < \varepsilon_0$. In particular, given that $\omega^{<\omega}$ is countable and $(\varphi(P_s), \tau|_{\varphi(P_s)})$ has countable base for every $s \in \omega^{<\omega}$, there is an $s \in \omega^{<\omega}$ and an open set $U \subseteq (\varphi(P_s), \tau|_{\varphi(P_s)})$ such that for a countably infinite set of indices $J'' \subseteq J_1(\varepsilon_0/2)$ we have that $H_\alpha(\varepsilon_0/2)) \cap \varphi(P_s)$ is comeager in U in the $\tau|_{\varphi(P_s)}$ topology whenever $\alpha \in J''$. Hence by the Baire Category Theorem for

$$H'' = \bigcap_{\alpha\in J''} H_\alpha(\varepsilon_0/2),$$

H'' is also comeager in U in the $\tau|_{\varphi(P_s)}$ topology, so by (b) we can find a point $x_0 \in H'' \setminus H(0)$. Thus f_α $(\alpha < \omega_1)$ is not $\frac{d}{\mathcal{I}_<}$-convergent since

$$J'' \subseteq \left\{\alpha < \omega_1 : d(f(x_0), f_\alpha(x_0)) > \frac{\varepsilon_0}{2}\right\}$$

is infinite; a contradiction. This completes the proof. ∎

As we have mentioned above, Theorem 2 is true for every Baire class (see [4]). The proof of the general theorem uses a characterization of $\Sigma^0_\xi(\tau)$ sets for every $\xi < \omega_1$ involving Baire category, similarly to Lemma 3. Finally we note that this approach makes also possible to treat the pointwise convergence of sequences of Borel functions with length λ where $\omega_1 < \lambda < 2^{\aleph_0}$ is a cardinal.

ACKNOWLEDGEMENT. The author would like to thank Márton Elekes for his helpful remarks.

References

[1] A. S. KECHRIS: Classical Descriptive Set Theory, *Graduate Texts in Mathematics* **156**, Springer-Verlag, (1994).

[2] P. KOMJÁTH: Limits of transfinite sequences of Baire-2 functions, *Real Anal. Exchange* **24(2)** (1998/99), 497–502.

[3] A. LOUVEAU and J. SAINT RAYMOND: Borel Classes and Closed Games: Wadge-type and Hurewicz-type Results, *Trans. Amer. Math. Soc.* **304(2)** (1987), 431–467.

[4] T. MÁTRAI: The ω_1-limit of Baire-ξ functions is Baire-ξ, *Fund. Math.* (2004), submitted for publication.

[5] T. NATKANIEC: The $\mathcal{J}$-almost constant convergence of sequences of real functions, *Real Anal. Exchange* **28(2)** (2002/03), 481–491.

[6] W. SIERPIŃSKI: Sur les suites transfinies convergentes de fonctions de Baire, *Fund. Math.* **1** (1920), 134–141.

Tamás Mátrai

CEU Central European University
Department of Mathematics and its Applications
1059 Budapest
Nádor utca 9.
Hungary
matrait@renyi.hu

ELEMENTARY RESULTS IN CONTROL OF ONE-DIMENSIONAL DISCRETE TIME DYNAMICAL SYSTEMS DEFINED BY A MULTIFUNCTION

By

GERGELY KOVÁCS and BÉLA VIZVÁRI

(May 30, 2004)

1. Introduction

This paper is devoted to the control of one-dimensional discrete time dynamical systems defined by a multifunction. The fact that the relation defining the system is an inclusion instead of an equation, reflects the uncertainties of the system and/or that our knowledge on the system is not complete.

In many applications of dynamical systems it is important to stabilize the system, i.e., the trajectory must be moved to and kept in a certain target region. To achieve this objective is more difficult in the case of the type of dynamical systems discussed in this paper than in the case of the traditional dynamical systems because of the higher degree freedom of the system.

The basic assumptions on the multifunction based dynamical system are as follows:

(A1) Let x_t denote the state of the system at time t. The next state, i.e., x_{t+1}, is an element of the set $G(x_t)$. It is assumed, that this set is a bounded interval for each x, i.e.,

$$G(x) = [a(x), b(x)],$$

where $a(x)$ and $b(x)$ are real valued functions with $a(x) \leq b(x)$, $\forall x \geq 0$.

(A2) It is assumed that $a(x)$ and $b(x)$ are linear functions, i.e.,

$$a(x) = \alpha_1 x + \alpha_0$$

and

$$b(x) = \beta_1 x + \beta_0,$$

i.e., the multifunction is

$$G(x_t) = [\alpha_1 x_t + \alpha_0, \beta_1 x_t + \beta_0].$$

(A3) The functions $a(x)$ and $b(x)$ are increasing, or equivalently α_1 and β_1 are positive real numbers with $\alpha_1 \leq \beta_1$.

(A4) The control is realized by an additive term. The control parameter denoted by q can be chosen from a bounded closed interval being symmetric to zero, i.e., a positive d exists such that

$$q \in [-d, d].$$

The control parameter can be changed in every step, too. Thus, the dynamics of the controlled system is described by the inclusion

$$x_{t+1} \in G(x_t) + q_{t+1}. \tag{1}$$

(A5) The control parameter q has a cost $c(q)$. It is assumed that this c function is continuous and differentiable (at $x = 0$ we consider side derivatives), symmetric to 0 with respect to the vertical axis and strictly increasing in the positive region. Hence if the derivative exists and $c'(q) = 0$, then $q = 0$ and if $q > 0$ then $c'(q) > 0$.

(A6) Our aim is to move the trajectory into a fixed interval $[A, B]$ in k steps from an initial state x_0 with minimal control cost. The number k is a fixed positive integer. It is assumed that

$$< x_0 < A < B.$$

According to (1) x_{t+1} is a shifted point of $G(x_i)$. This point of $G(x_t)$ is called *realization*. If we may choose the control parameter from the interval $[-d, d]$ after the realization becomes known, then the control is called *a posteori*. Otherwise, if we must choose the parameter from the interval without knowing the realization, then the control is called *a priori*.

The existence of an appropriate control of the system (A1)-(A6) has been discussed in [2].

(A7) If the system makes k iterations, then the total cost of the control is the sum of the costs of the controls in the individual iterations, i.e.,

$$\sum_{i=1}^{k} c(q_i).$$

In this paper a method is given for entering on the target interval in k steps at minimal cost.

2. The a priori case

The dynamics of the system is as follows. A new state x'_{t+1} is determined such that

$$x'_{t+1} \in [\alpha_1 x_t + \alpha_0, \beta_1 x_t + \beta_0] = \alpha_1 x_t + \alpha_0 + \epsilon_{t+1},$$

where $\epsilon_{t+1} \in [0, \beta_1 x_t + \beta_0 - \alpha_1 x_t - \alpha_0]$. Then it is transformed by the control into

$$x_{t+1} = x'_{t+1} + q_{t+1} = \alpha_1 x_t + \alpha_0 + \epsilon_{t+1} + q_{t+1}.$$

Thus, $x_1 = \alpha_1 x_0 + \alpha_0 + \epsilon_1 + q_1$, and in general

$$x_t = \alpha_1^t x_0 + \alpha_0 \sum_{i=0}^{t-1} \alpha_1^i + \sum_{i=1}^{t} \alpha_1^{t-i}(q_i + \epsilon_i). \tag{2}$$

Our aim is to reach the $[A, B]$ interval in exactly k steps, i.e., the relation $x_k \in [A, B]$ must hold, with a minimal cost. Three cases are investigated: (a) it can be assumed that all values of the control parameters are non-negative or (b) it can be assumed that the values of the control parameters are non-positive or (c) both positive and negative values of the control parameter can occur.

2.1. Nonnegative control

It is assumed, that B is great enough that for all realization and for all q the statement $x_k \leq B$ holds, i.e., for the maximal value of x_k, the inequality

$$\beta_1^t x_{k-t} + \beta_0 \sum_{i=0}^{t-1} \beta_1^i + \sum_{i=1}^{t} \beta_1^{t-i} d \leq B,$$

holds. It means that the only constraint, which must be satisfied is $x_k \geq A$. Hence $q_i \geq 0$ for all i.

The worst case is if for all i the equality $\epsilon_i = 0$ holds, because the system needs the highest amount of control in this case. On the other hand if this outcome of events is disregarded the system may miss the target region. Thus, at each iteration we must select the best possible value of the control variable, which still works even in the worst case of the future iterations. This can be modeled by the following nonlinear optimization problem if still t $(1 \leq t \leq k)$ iterations remained:

$$C_{k-t} + \sum_{i=k-t+1}^{k} \alpha_1^{k-i} q_i \geq A \tag{3}$$

$$\forall k-t+1 \le i \le k : \; 0 \le q_i \le d \tag{4}$$

$$\min \sum_{i=k-t+1}^{k} c(q_i), \tag{5}$$

where the q_i's are the decision variables according to (2) and

$$C_{k-t} = \alpha_1^t x_{k-t} + \alpha_0 \sum_{i=0}^{t-1} \alpha_1^i, \tag{6}$$

which is the point what the system can reach after the k-th iteration in the worst case, i.e., when $\epsilon_i = 0$, $i = k-t+1, \dots, k$.

For all i the value of q_i can be chosen nonnegative, because x must be increased.

When the decision is made on the value of q_{k-t} at the time $k-t$ the system is in a position determined by x_0, and q_i, ϵ_i $(i = 1, \dots, k-t-1)$. The sequence $\{q_{k-t+1}, \dots, q_k\}$ of the control values is denoted by the vector $\mathbf{q}$. All control vectors are evaluated under the hypothesis that all remaining ϵ_i's are equal to 0 unless it is stated else.

Theorem 2.1. *If* $\alpha_1 > 1$, *then there is an optimal solution* $\mathbf{q}^*$ *satisfying the inequalities* $q_i^* \ge q_{i+1}^*$ $(i = k-t+1, \dots, k-1)$.

Proof. If $q_{i+1}^* = a$, $q_i^* = b$ and $a < b$, then the control vector $\mathbf{q}$ with

$$q_l = \begin{cases} q_l^* & \text{if } t \neq i, i+1 \\ a & \text{if } l = i \\ b & \text{if } l = i+1 \end{cases}$$

has the same cost. Assuming that the sequence of ϵ_i's is the same, the state value x_k obtained by control vector $\mathbf{q}$ is greater than the state value x_k^* obtained by $\mathbf{q}^*$. Thus q_{i+1} can be decreased such that the control is still feasible and has less total cost. ▮

Theorem 2.2. *If* $\alpha_1 < 1$, *then there is an optimal solution* $\mathbf{q}^*$ *satisfying the inequalities* $q_i^* \le q_{i+1}^*$ $(i = k-t+1, \dots, k-1)$.

Proof. The proof is similar to the proof of the previous theorem. ▮

Theorem 2.3. *If* $\alpha_1 > 1$ *and* $c(q)$ *is linear or concave function on the interval* $[0, d]$, *then the components of the optimal solution* $\mathbf{q}^*$ *have the structure* $(d, d, \dots, d, a, 0, \dots, 0)$, *where* $0 \le a \le d$.

PROOF. Assume that at least two components of $\mathbf{q}^*$ are strictly between 0 and d. Then it follows from Theorem 2.1 that they are consecutive components with the same property, i.e., there is an index i such that $0 < < q^*_{i+1} \leq q^*_i < d$. Let $0 < \delta < \min\{d - q^*_i, q^*_{i+1}\}$. Then the total cost of control vector $\mathbf{q}$ with

$$q_l = \begin{cases} q^*_l & \text{if } l \neq i, i+1 \\ q^*_i + \delta & \text{if } l = i \\ q^*_{i+1} - \delta & \text{if } l = i+1 \end{cases}$$

is at most as high as the total cost of the control vector $\mathbf{q}^*$, but the state value of x_k obtained by control vector $\mathbf{q}$ is greater than the state value x^*_k obtained by $\mathbf{q}^*$. This is a contradiction. ∎

THEOREM 2.4. *If* $\alpha_1 < 1$ *and* $c(q)$ *is a linear or concave function on the interval* $[0, d]$*, then the components of the optimal solution* $\mathbf{q}^*$ *have the structure* $(0, 0, \ldots, 0, a, d, \ldots, d)$*, where* $0 \leq a \leq d$.

PROOF. The proof is similar to the proof of the previous theorem. ∎

DEFINITION 2.1. A vector $\mathbf{q}$ is admissible if

$$\sum_{i=k-t+1}^{k} \alpha_1^{k-i} q_i = A - C_{k-t}, 1 \leq t \leq k$$

where C_{k-t} is defined in (6).

THEOREM 2.5. *If* $\alpha_1 > 1$ *and* $c(q) = |q|^n$*, where* $n > 1$*, then the optimal solution of the problem is the admissible vector* $(d, \ldots, d, q_{s+1}, \ldots, q_k)$*, where* $d > q_i > 0$ $(i = s+1, \ldots, k)$, $\frac{q_i}{q_{i+1}} = \alpha_1^{\frac{1}{n-1}}$ *and*

$$s = \min\{l \mid \exists \mathbf{p} = (d, \ldots, d, p_{l+1}, \ldots, p_k), 0 < p_i < d, (i = l+1, \ldots, k), \mathbf{p} \textit{ is admissible}\}.$$

PROOF. Let λ_i, μ_i $(i = k - t + 1, \ldots, k)$, μ_0 be the Lagrange multipliers of the optimization problem (3)–(5) such that λ_i's belong the nonnegativity conditions, μ_i's belong the upper bounds in (4) and μ_0 belongs to (3). Then the Karush-Kuhn-Tucker necessary conditions (see [1], page 146) of optimality for problem (3)–(5) are:

$$\forall i : \lambda_i(-q_i) = 0 \tag{7}$$

$$\forall i : \mu_i(q_i - d) = 0 \tag{8}$$

$$\mu_0 \left(A - C_{k-t} - \sum_{i=1}^{t} \alpha_1^{t-i} q_i \right) = 0 \tag{9}$$

$$\forall i : c'(q_i) - \lambda_i + \mu_i - \mu_0 \alpha_1^{t-i} = 0 \tag{10}$$

The following cases may occur.

CASE I. $\mu_0 = 0$. Then $q_i = 0$ for all i. If $q_i \neq 0$, then equation (10) implies that $\lambda_i > 0$. Then equation (18) is not satisfied.

CASE II. $\mu_0 > 0$. If $0 < q_i < d$, then it follows that $\lambda_i = \mu_i = 0$ and $c'(q_i) = \mu_0 \alpha_1^{t-i}$. If $q_i = 0$, then $0 = \lambda_i + \mu_0 \alpha_1^{t-i}$ and the equation $\mu_0 = 0$ follows from the facts that λ_i is nonnegative and α_1 is positive. This is a contradiction. Finally if $q_i = d$ then $c'(d) = -\mu_i + \mu_0 \alpha_1^{t-i}$.

Thus it is proven that $\mu_0 > 0$ and $\forall i : q_i > 0$.

Let K_{s+2} the cost of the control belonging to the admissible vector

$$\mathbf{q} = (d, \dots, d, q_{s+2}, \dots, q_k)^T$$

and similarly K_{s+1} the cost of the control belonging to the admissible vector

$$\mathbf{q}' = (d, \dots, d, q'_{s+1}, q'_{s+2}, \dots, q'_k)^T,$$

where $0 < q_i < d$, $i = s+2, \dots, k$ and $0 < q'_i < d$, $i = s+1, \dots, k$. It is assumed, that both $\mathbf{q}$ and $\mathbf{q}$ are satisfying the necessary conditions (7)–(10) within appropriate multipliers. It will be shown that $K_{s+2} \leq K_{s+1}$ if the control effects of $\mathbf{q}$, and $\mathbf{q}'$ are equal, i.e., the equation

$$D = A - C_{s+1} = d\alpha_1^{k-s-1} + \sum_{i=2}^{k-s} q_{s+i} \alpha_1^{k-s-i} = \sum_{i=1}^{k-s} q'_{s+i} \alpha_1^{k-s-i}.$$

holds. It follows from $c'(q_{s+i}) = \mu_0 \alpha_1^{k-s-i}$ for all i that

$$\frac{c'(q_{s+i})}{c'(q_{s+j})} = \frac{q_{s+i}^{n-1}}{q_{s+j}^{n-1}} = \alpha_1^{j-i}.$$

Thus,

$$q_{s+i} = q_k \alpha_1^{\frac{k-s-i}{n-1}}$$

and similarly

$$q'_{s+i} = q'_k \alpha_1^{\frac{k-s-i}{n-1}}.$$

Therefore

$$q_k + q_k\alpha_1^{\frac{n}{n-1}} + \cdots + q_k\alpha_1^{\frac{(k-s-2)n}{n-1}} + d\alpha_1^{k-s-1} = D$$

and similarly

$$q_k' + q_k'\alpha_1^{\frac{n}{n-1}} + \cdots + q_k'\alpha_1^{\frac{(k-s-1)n}{n-1}} = D.$$

Thus,

$$q_k = \frac{D - d\alpha_1^{k-s-1}}{1 + \alpha_1^{\frac{n}{n-1}} + \cdots + \alpha_1^{\frac{(k-s-2)n}{n-1}}} = (D - d\alpha_1^{k-s-1})\frac{\alpha_1^{\frac{n}{n-1}} - 1}{\alpha_1^{\frac{(k-s-1)n}{n-1}} - 1}$$

and similarly

$$q_k' = \frac{\alpha_1^{\frac{n}{n-1}} - 1}{\alpha_1^{\frac{(k-s)n}{n-1}} - 1}D.$$

$$\begin{aligned} K_{s+2} &= d^n + q_{s+2}^n + \cdots + q_k^n = d^n + q_k^n\left(1 + \alpha_1^{\frac{n}{n-1}} + \cdots + \alpha_1^{\frac{(k-s-2)n}{n-1}}\right) \\ &= d^n + q_k^n\frac{\alpha_1^{\frac{(k-s-1)n}{n-1}} - 1}{\alpha_1^{\frac{n}{n-1}} - 1} = d^n + (D - \alpha_1^{k-s-1}d)^n\frac{(\alpha_1^{\frac{n}{n-1}} - 1)^{n-1}}{(\alpha_1^{\frac{(k-s-1)n}{n-1}} - 1)^{n-1}} \end{aligned}$$

and similarly

$$K_{s+1} = D^n\frac{(\alpha_1^{\frac{n}{n-1}} - 1)^{n-1}}{(\alpha_1^{\frac{(k-s)n}{n-1}} - 1)^{n-1}}.$$

Let $r = \frac{D}{d}$ and

$$S_{k-s-1} = \frac{\alpha_1^{\frac{(k-s-1)n}{n-1}} - 1}{\alpha_1^{\frac{n}{n-1}} - 1}.$$

Then

$$M_{s+1} := \frac{K_{s+2} - K_{s+1}}{d^n}$$

$$= 1 + \left(r - \alpha_1^{k-s-1}\right)^n \left(\frac{1}{S_{k-s-1}}\right)^{n-1} - r^n \left(\frac{1}{\alpha_1^{\frac{(k-s-1)n}{n-1}} + S_{k-s-1}}\right)^{n-1}.$$

If M_{s+1} is derivated by r, then the equation

$(M_{s+1})' =$

$$= n\left(r - \alpha_1^{k-s-1}\right)^{n-1} \left(\frac{1}{S_{k-s-1}}\right)^{n-1} - n r^{n-1} \left(\frac{1}{\alpha_1^{\frac{(k-s-1)n}{n-1}} + S_{k-s-1}}\right)^{n-1}$$

is obtained. The derivate is 0 only if

$$\left(r - \alpha_1^{k-s-1}\right) \left(\frac{1}{S_{k-s-1}}\right) = r \left(\frac{1}{\alpha_1^{\frac{(k-s-1)n}{n-1}} + S_{k-s-1}}\right),$$

i.e.,

$$r^* = \frac{\alpha_1^{k-s-1}\left(\alpha_1^{\frac{(k-s-1)n}{n-1}} + S_{k-s-1}\right)}{\alpha_1^{\frac{(k-s-1)n}{n-1}}}.$$

On the other hand r^* is the root of M_{s+1} too:

$$1+\left(\frac{\alpha_1^{k-s-1} S_{k-s-1}}{\alpha_1^{\frac{(k-s-1)n}{n-1}}}\right)^n \left(\frac{1}{S_{k-s-1}}\right)^{n-1} - \left(\frac{\alpha_1^{k-s-1} S_{k-s}}{\alpha_1^{\frac{(k-s-1)n}{n-1}}}\right)^n \left(\frac{1}{S_{k-s}}\right)^{n-1} =$$

$$= 1 + \left(\frac{\alpha_1^{k-s-1}}{\alpha_1^{\frac{(k-s-1)n}{n-1}}}\right)^n S_{k-s-1} - \left(\frac{\alpha_1^{k-s-1}}{\alpha_1^{\frac{(k-s-1)n}{n-1}}}\right)^n S_{k-s} =$$

$$= 1 - \frac{\left(\alpha_1^{k-s-1}\right)^n}{\left(\alpha_1^{\frac{(k-s-1)n}{n-1}}\right)^{n-1}} = 0.$$

The value of the second derivate of M_{s+1} at r^* is:

$$\begin{aligned} n(n-1)&\left(\frac{\alpha_1^{k-s-1}S_{k-s-1}}{\alpha_1^{\frac{(k-s-1)n}{n-1}}}\right)^{n-2}\left(\frac{1}{S_{k-s-1}}\right)^{n-1}\\ &\qquad - n(n-1)\left(\frac{\alpha_1^{k-s-1}S_{k-s}}{\alpha_1^{\frac{(k-s-1)n}{n-1}}}\right)^{n-2}\left(\frac{1}{S_{k-s}}\right)^{n-1}\\ &= n(n-1)\left(\frac{\alpha_1^{k-s-1}}{\alpha_1^{\frac{(k-s-1)n}{n-1}}}\right)^{n-2}\left(\frac{1}{S_{k-s-1}}-\frac{1}{S_{k-s}}\right), \end{aligned}$$

which is positive. Hence it follows, that M_{s+1} has a minimal value at r^* and M_{s+1} is nonnegative, i.e., $K_{s+2} > K_{s+1}$. ∎

THEOREM 2.6. *If $\alpha_1 < 1$ and $c(q) = |q|^n$, where $n > 1$, then the optimal solution of the problem is the admissible vector $(q_{k-t}, q_{k-t+1}, \dots, q_s, d, d, \dots$ $\dots, d)$, where $d > q_i > 0$ $(i = k - t, \dots, s)$, $\frac{q_i}{q_{i+1}} = \alpha_1^{\frac{1}{n-1}}$ and*

$$s = \max\{l \,|\, \exists \mathbf{p} = (p_{k-t}, p_{k-t+1}, \dots, p_l, d, \dots, d), 0 < p_i < d, (i = k-t, \dots, l),$$
$$\mathbf{p} \textit{ is admissible}\}.$$

PROOF. The proof is similar to the proof of the previous theorem. ∎

2.2. Non-positive control

It is enough to use only non-positive control values if for all realization and for all q the statement $x_k \geq A$ holds, i.e., for the minimal value of x_k:

$$\alpha_1^t x_{k-t} + \alpha_0 \sum_{i=0}^{t-1} \alpha_1^i + \sum_{i=1}^{t} \alpha_1^{t-i}(-d) \geq A$$

for all $t = 1, \dots, k$. It means that we need only $x_k \leq B$, i.e., $q_i \leq 0$ for all i.

The worst case is if for all i $\epsilon_i = \beta_1 x_{i-1} + \beta_0 - \alpha_1 x_{i-1} - \alpha_0$, because the system needs the highest amount of control in this case. On the other hand if this outcome of events is disregarded then the system may miss the target region. Thus, in each iteration the best possible value of the control variable must be selected, which still works even in the worst case. This can be modeled by the following nonlinear optimization problem if still t out of k, $(1 \leq t \leq k)$ iterations remained:

$$D_{k-t} + \sum_{i=k-t+1}^{k} \beta_1^{k-i} q_i \leq B$$

$$\forall i \in [k-t+1, k] : \; -d \leq q_i \leq 0$$

$$\min \sum_{i=k-t+1}^{k} c(q_i),$$

where the q_i's are the decision variables and

$$D_{k-t} = \beta_1^t x_{k-t} + \beta_0 \sum_{i=0}^{t-1} \beta_1^i. \tag{11}$$

For all i the value of q_i can be chosen non-positive, because the state x must be decreased.

When q_{k-t} is selected at the time $k-t$ the system is in a position determined by x_0, and q_i, ϵ_i $(i = 1, \ldots, k-t-1)$.

Of course, this case is very similar to the one when nonnegative control is possible. Therefore similar sequence of statement holds.

THEOREM 2.7. *If* $\beta_1 > 1$, *then there is an optimal solution* $\mathbf{q}^*$ *satisfying the inequalities* $q_i^* \leq q_{i+1}^*$, $(i = k-t+1, \ldots, k-1)$.

PROOF. The proof is similar to the proof of the Theorem 2.1. ∎

THEOREM 2.8. *If* $\beta_1 < 1$, *then there is an optimal solution* $\mathbf{q}^*$ *such that the values satisfy* $q*_i \geq q*_{i+1}$.

PROOF. The proof is similar to the proof of the Theorem 2.1. ∎

THEOREM 2.9. *If* $\beta_1 > 1$ *and* $c(q)$ *is a linear or concave function on the interval* $[-d, 0]$, *then the components of the optimal solution* $\mathbf{q}^*$ *have the structure* $(-d, -d, \ldots, -d, a, 0, \ldots, 0)$, *where* $-d \leq a \leq 0$.

PROOF. The proof is similar to the proof of Theorem 2.3. ∎

THEOREM 2.10. *If* $\beta_1 < 1$ *and* $c(q)$ *is a linear or concave function on the interval* $[-d, 0]$, *then the components of the optimal solution* $\mathbf{q}^*$ *have the structure* $(0, 0, \ldots, 0, a, -d, \ldots, -d)$, *where* $-d \leq a \leq 0$.

PROOF. The proof is similar to the proof of Theorem 2.3. ∎

DEFINITION 2.2. A vector $\mathbf{q}$ is admissible if

$$\sum_{i=k-t+1}^{k} \beta_1^{k-i} q_i = B - D_{k-t},$$

where D_{k-t} is given in (11).

THEOREM 2.11. *If* $\beta_1 > 1$ *and* $c(q) = |q|^n$, *where* $n > 1$, *then the optimal solution of the problem is the admissible vector* $(-d, \ldots, -d, q_{s+1}, \ldots, q_k)$, *where* $-d < q_i < 0$ $(i = s+1, \ldots, k)$, $\frac{q_i}{q_{i+1}} = \beta_1^{\frac{1}{n-1}}$ *and*

$$s = \min\{l \,|\, \exists \mathbf{p} = (-d, \ldots, -d, p_{l+1}, \ldots, p_k), -d < p_i < 0, (i = l+1, \ldots, k), \mathbf{p} \textit{ is admissible}\}.$$

PROOF. The proof is similar to the proof of Theorem 2.5. ∎

THEOREM 2.12. *If* $\beta_1 < 1$ *and* $c(q) = |q|^n$, *where* $n > 1$, *then the optimal solution of the problem is the admissible vector* $(q_{k-t}, q_{k-t+1}, \ldots, q_s, -d, -d, \ldots, -d)$, *where* $-d < q_i < 0$ $(i = k-t, \ldots, s)$, $\frac{q_i}{q_{i+1}} = \beta_1^{\frac{1}{n-1}}$ *and*

$$s = \max\{l \,|\, \exists \mathbf{p} = (p_{k-t}, p_{k-t+1}, \ldots, p_l, -d, \ldots, -d), -d < p_i < 0, (i = k-t, \ldots, l), \mathbf{p} \textit{ is admissible}\}.$$

PROOF. The proof is similar to the proof of Theorem 2.5. ∎

2.3. Control by positive and negative values

The system can be controlled only if the state of the highest realizations with the control parameter $-d$ is less then B, i.e.,

$$\beta_1^t x_{k-t} + \beta_0 \sum_{i=0}^{t-1} \beta_1^i + \sum_{i=1}^{t} \beta_1^{t-i}(-d) \leq B$$

and if the state of the lowest realizations with the control parameter d is more than A, i.e.,

$$\alpha_1^t x_{k-t} + \alpha_0 \sum_{i=0}^{t-1} \alpha_1^i + \sum_{i=1}^{t} \alpha_1^{t-i} d \geq A.$$

The optimal control problem can be solved by a generalization of the Bellman principle. In what follows it is elaborated in a backward manner.

I1. First the set X_{k-1} is calculated from the above conditions. The set contains the possible states of x_{k-1}, for which there exists a control parameter that all trajectories reach the target region $[A, B]$ with this parameter.

I2. Then the possible values of the control parameters are calculated for every point of the set X_{k-1} for which all controlled trajectories reach the interval $[A, B]$. These are non-empty intervals denoted by $[d^-(x_{k-1}), d^+(x_{k-1})]$.

I3. For every point x_{k-1} from the set X_{k-1} and for the sets $[d^-(x_{k-1}), d^+(x_{k-1})]$ the cheapest parameter from the set is chosen. It is the optimal control parameter for the state before the last iteration. In this way a function is determined for the last control parameter: $q_k(x_{k-1})$ and another for the cost of the control: $c_k(x_{k-1})$.

In general:

G1. Again from the above conditions the set X_{k-i} is calculated. The set contains the possible states of x_{k-i}, for which there exists a control parameter such that all trajectories reach the set X_{k-i+1} with this parameter.

G2. Then the possible values of the control parameters are calculated for every point of the set X_{k-i} for which all controlled trajectories reach the set X_{k-i+1}. These are non-empty intervals denoted by $[d^-(x_{k-i}), d^+(x_{k-i})]$. To every parameter from this set belongs a set too, the set of the possible next states using this parameter. Every point of this possible next states set has a cost $c_{k-i+2}(x_{k-i+1})$. The future cost of a possible parameter from $[d^-(x_{k-i}), d^+(x_{k-i})]$ is the maximal value of the above mentioned cost of the possible future states, denoted by $C_{k-i+1}(x_{k-i}, q_{k-i+1})$.

G3. For every point x_{k-i} from the set X_{k-i} and for the sets $[d^-(x_{k-i}), d^+(x_{k-i})]$ a parameter q_{k-i+1} is chosen for which the sum of the cost of this parameter $c(q_{k-i+1})$ and the future cost of this parameter $C_{k-i+1}(x_{k-i}, q_{k-i+1})$ is minimal. It is the optimal control parameter for this state. This way a new function is determined for the $k - i + 1$-th control parameter: $q_{k-i+1}(x_{k-i})$ and another for the cost of the control: $c_{k-i+1}(x_{k-i})$.

THEOREM 2.13. *If the function c is convex, then c_{k-i+1} is convex.*

PROOF. The proof is based on induction. Of course c_k is convex, because c is convex. If we assume that c_{k-i+2} is convex, then

$$C_{k-i}(x_{k-i,q}) = \max c_{k-i+2}(\alpha_1 x_{k-i} + \alpha_0 + q); c_{k-i+2}(\beta_1 x_{k-i} + \beta_0 + q).$$

The functions $c_{k-i+2}(\alpha_1 x_{k-i} + \alpha_0 + q)$ and $c_{k-i+2}\beta_1 x_{k-i} + \beta_0 + q$ are convex, because these are combination of a strictly increasing and convex and a convex function. The maximum of two convex functions is convex, too. Thus $C_{k-i+1}(x_{k-i,q})$ is convex. The sum of two convex function is convex, thus $c(q) + \max c_{k-i+2}(\alpha_1 x_{k-i} + \alpha_0 + q); c_{\beta_1 x_{k-i} + \beta_0 + q}$ is convex. The function

$$c_{k-i+1}(x_{k-i}) = \min_{q \in [d^-(x_{k-i}), d^+(x_{k-i})]} C_{k-i+1}(x_{k-i}, q),$$

i.e., the epigraph of this function is the projection of the epigraph of the above two-variables convex function. Thus the function c_{k-i+1} is convex. ▮

It follows, that

$$\begin{aligned} &c_{k-i+1}(x_{k-i}) = \\ &= \min_{q \in [d^-(x_{k-i}), d^+(x_{k-i})]} \max c_{k-i+2}(\alpha_1 x_{k-i} + \alpha_0 + q), c_{k-i+2}(\beta_1 x_{k-i} + \beta_0 + q). \end{aligned}$$

Finally we get the optimal control parameter for the actual step.

3. An example for control

Let $a(x) = x$ and $b(x) = 1.2x + 2$, $A = 45$, $B = 60$ and $d = 10$. Let $c(q) = q^2$.

I1. The set $X_{k-1} = [35, 56.667]$.

I2. If $35 \le x \le 55$, then $d^-{}_{k-1} = 45 - x$, and $55 \le x \le 56.667$, then $d^-{}_{k-1} = -10$. Moreover if $35 \le x \le 40$, then $d^+{}_{k-1} = 10$ and if $40 \le x \le 56.667]$, then $d^+{}_{k-1} = 1.2x - 58$.

I3. The cheapest parameters from the sets are the following:

a) $q_k(x_{k-1}) = 45 - x$ if $35 \le x \le 45$,

b) $q_k(x_{k-1}) = 0$ if $45 \le x \le 48.333$ and

c) $q_k(x_{k-1}) = 1.2x - 58$ if $48.333 \le x \le 56.667$.

Moreover

a) $c_k(x_{k-1}) = (45 - x)^2$ if $35 \le x \le 45$,

b) $c_k(x_{k-1}) = 0$ if $45 \leq x \leq 48.333$ and

c) $c_k(x_{k-1}) = (1.2x - 58)^2$ if $48.333 \leq x \leq 56.667$.

II1. The set $X_{k-2} = [25, 53.889]$

II2. If $25 \leq x \leq 45$, then $d^{-}{}_{k-2} = 35 - x$, and $45 \leq x \leq 53.889$, then $d^{-}{}_{k-2} = -10$. Moreover if $25 \leq x \leq 37.222$, then $d^{+}{}_{k-1} = 10$ and if $37.222 \leq x \leq 53.889$, then $d^{+}{}_{k-1} = 1.2x - 56.667$.

II3. The cheapest parameters from the sets are the following:

a) $q_{k-1}(x_{k-2}) = 22.5 - 0.5x$ if $25 \leq x \leq 38.134$,

b) $q_{k-1}(x_{k-2}) = 45.727 - 1.109x$ if $38.134 \leq x \leq 45.855$,

c) $q_{k-1}(x_{k-2}) = 27.34 - 0.708x$ if $45.855 \leq x \leq 52.746$ and

d) $q_{k-1}(x_{k-2} = -10$ if $52.746 \leq x \leq 53.889$.

4. The a posteori case

The a posteori case is similar to the a priori one, because of the worst cases are just the same. Only the iteration number is shifted (decreased) by one and the values of C_{k-t} and D_{k-t} are different:

$$C_{k-t} = \alpha_1^{t-1} x'_{k-t+1} + \alpha_0 \sum_{i=0}^{t-2} \alpha_1^i$$

and

$$D_{k-t} = \beta_1^{t-1} x'_{k-t+1} + \beta_0 \sum_{i=0}^{t-2} \beta_1^i,$$

where x'_{k-t+1} is the realization of after the state x_{k-t}.

References

[1] M. S. BAZARAA and C. M. SHETTY: *Nonlinear Programming*, John Wiley and Sons, New York, 1979.

[2] KOVÁCS, G., MURESAN, M. and VIZVÁRI, B.: *Viability results in control of one-dimensional discrete time dynamical systems defined by a multifunction*, Pural Mathematics, Siena–Budapest, 2003.

Gergely Kovács

College for Modern Business Studies
Tatabánya, Hungary
`kovacs.gergely@mutf.hu`

Béla Vizvári

Eötvös Loránd University
Budapest, Hungary
`vizvari@cs.elte.hu`

INDEX

ISSN 0524-9007

Address:
MATHEMATICAL INSTITUTE, EÖTVÖS LORÁND UNIVERSITY
ELTE, TTK, KARI KÖNYVTÁR, MATEMATIKAI TUDOMÁNYÁGI SZAKGYŰJTEMÉNY
PÁZMÁNY PÉTER SÉTÁNY 1/c, 1117 BUDAPEST, HUNGARY
Műszaki szerkesztő:
Fried Katalin PhD

A kiadásért felelős: az Eötvös Loránd Tudományegyetem rektora
A kézirat a nyomdába érkezett: 2004. július. Megjelent: 2004. július
Terjedelem: 17,16 A/4 ív. Példányszám: 500
Készült az EMTEX szedőprogram felhasználásával
az MSZ 5601–59 és 5602–55 szabványok szerint
Az elektronikus tipografálás Bori Tamás munkája
Nyomda: Haxel Bt.
Felelős vezető: Korándi József