JUNE / JUIN 2001

IN THIS ISSUE / DANS CE NUMÉRO

23
No 2

C. R. Math. Rep. Acad. Sci. Canada Vol. **23** (2), 2001 pp. 33–38

PERTURBATION D'UNE INCLUSION DIFFÉRENTIELLE NON CONVEXE AVEC VIABILITÉ

SAÏD SAJID

Présenté par Vlastimil Dlab, FRSC

ABSTRACT. We establish the existence of viable solutions of a class of multivalued differential equation with Carathéodory nonconvex right-hand side.

RÉSUMÉ. On établit l'existence de solutions d'une classe d'équations différentielles multivoques avec contrainte sur l'état dont le second membre est de Carathéodory et non convexe.

1. Introduction. Dans la littérature, l'existence de solutions d'équations différentielles du premier ordre avec contrainte sur l'état est obtenue grâce à des conditions tangentielles telles le champs de vecteurs est contenu ou rencontre le cône contingent. Dans cette note, on présente un résultat d'existence de solution viable d'une inclusion différentielle dont le second membre est de Caratheodory et non convexe.

Dans tout ce qui suit, H est un espace de Hilbert séparable, $\mathcal{B}(H)$ l'ensemble des parties convexes faiblement compactes d'intérieur non vide de H muni de la distance de Hausdorff h, pour tout intervalle J, $AC(J,H)$ et $L^1(J)$ désignent l'espace des applications absolument continues de J dans H muni de la topologie de la convergence uniforme et l'espace des applications réelles intégrables sur J. Pour toute partie non vide A d'un espace de Banach E, on note par ∂A, $\operatorname{int} A$, $\operatorname{ext} A$, χ_A, $\overline{A}$, $d(.,A)$ et $\pi_A(.)$ la frontière, l'intérieur, l'ensemble des points extremaux, la fonction caractéristique, l'adhérence de A, la distance à A et la projection sur A respectivement. Pour tout $r > 0$ et $x \in E$, on note par $B(x,r)$ la boule de centre x et de rayon r.

2. Hypothèses et résultat principal. Soient K un convexe fermé non vide de H, $x_0 \in K$, $f\colon [0,T] \times K \to H$ et $F\colon [0,T] \times K \to \mathcal{B}(H)$. On considère les hypothèses suivantes :

(H_1) $\forall x \in K$, $t \to f(t,x)$ est mesurable et $f(t,x) \in T_K(x) \forall (t,x) \in [0,T] \times K$.

(H_2) $\exists k \in L^1([0,T])$, $\forall t \in [0,T]$, $x \to f(t,x)$ est $k(t)$-lipschitzienne.

(H_3) $\exists g \in L^1([0,T])$, $\forall (t,x) \in [0,T] \times K \|f(t,x)\| \leq g(t)$.

Reçu par les éditeurs le 1 août 2000.
Classification (AMS) par sujet : 34A60.
© Société mathématique du Canada 2001.

(H_4) F est continue pour la distance de Hausdorff.

(H_5) Il existe un convexe compact non vide D de H tel que $\forall (t,x) \in [0,T] \times K$, on a $\begin{cases} (\operatorname{int} F(t,x)) \cap T_K(x) \cap \mathrm{D} \neq \emptyset \\ \overline{\operatorname{co}}\big[(\operatorname{ext} F(t,x)) \cap T_K(x) \cap \mathrm{D}\big] = F(t,x) \cap T_K(x) \cap \mathrm{D}. \end{cases}$

REMARQUE. Si K est un convexe compact non vide et non réduit à un singleton de H. Alors (H_5) est satisfaite dans le cas suivant :

$$\mathrm{D} = \{x - y : (x,y) \in K \times K\} \quad \text{et} \quad F(t,x) = \pi_{\mathrm{D}}(x) + d(0, \partial ri \mathrm{D}) \overline{B}(0,1)$$

où ∂riD est la frontière relative de D.

En outre, si H est de dimension finie, alors (H_5) est satisfaite dans le cas où F est uniformément borné et $F(t,x) \subset T_K(x)$ $\forall (t,x) \in [0,T] \times K$.

Sous les hypothèses (H_1)–(H_5) on a :

THÉORÈME 2.1. *Il existe $T_0 \in]0,T]$ et $x \in AC([0,T_0],H)$ tels que*

$$\begin{cases} \dot{x}(t) \in f(t,x(t)) + \operatorname{ext} F(t,x(t)) & p.p \text{ sur } [0,T_0], \\ x(0) = x_0, x(t) \in K & \forall t \in [0,T_0]. \end{cases}$$

REMARQUE. Si $K = H$ alors on a l'existence de solutions seulement sous (H_1), (H_2), (H_3) et les conditions suivantes :

(C_1) F est continue pour la distance de Hausdorff.

(C_2) $\exists l \in L^1([0,T])$, $\forall (t,x) \in [0,T] \times K$, $h\big(F(t,x), \{0\}\big) \leq l(t)$.

COROLLAIRE 2.2. *Soit C une multifonction mesurable de $[0,T]$ à valeurs convexes fermées non vides dans H telle que l'application $t \to h\big(C(t), \{0\}\big)$ soit intégrable. Si $K = H$ et si F vérifie (C_1) et (C_2), alors il existe $T_1 \in]0,T]$ et $y \in AC([0,T_1],H)$ tels que :*

$$\dot{y}(t) \in C(t) + \operatorname{ext} F\big(t, y(t)\big) \quad p.p \text{ sur } [0,T_0], y(0) = x_0.$$

Dans tout ce qui suit et pour des raisons techniques on pose $I = [0,T]$ et on définit sur $I \times H$ la multifonction $G(t,x) = F\big(t, \pi_K(x)\big)$. Il est claire que G hérite toutes les propriétés de F.

3. Fonctionnelles semi-continues supérieurement. Les techniques de démonstrations reposent sur le théorème de Baire appliqué à des ensembles engendrés par des fonctions semi-continues supérieurement. À cet effet, on introduit la fonction de Choquet.

PROPOSITION 3.1. *Il existe une fonctionnelle $d_G : I \times I \times H \to [-\infty, +\infty[$:*

(i) $0 \leq d_G(t,x,y) \leq R^2$ $\forall (t,x,y) \in \operatorname{gr}(G)$ *et* $\|y\| \leq R$, $R > 0$. $\operatorname{gr}(G)$ *désigne le graphe de* G.

(ii) $d_G(t,x,y) = 0$ *si et seulement si* $y \in \operatorname{ext} G(t,x)$.

Pour la preuve, voir [2] et [3].

PROPOSITION 3.2. *La fonction* $(x,y) \to d(y,T_K(x))$ *définie sur* $K \times H$ *est semi-continue supérieurement.*

PREUVE. Voir [1, Th. 1, p. 220 et Cor. 1, p. 52].

LEMME 3.3. *Pour toute fonction absolument continue* $f: I \to H$, *on a :*

$$\frac{d}{dt}[d(f(t),K)] \leq d\Big(\dot{f}(t), T_K\Big(\pi_K(f(t))\Big)\Big) \quad p.p \text{ sur } I.$$

PREUVE. Voir [1, Cor. 1, p. 179].

4. Preuve du résultat principal.

PROPOSITION 4.1. *Soit* $Q: I \times H \to \mathcal{B}(H)$ *une multifonction continue. Alors pour tout* $y \in H$, $\{(t,x) \in I \times H : y \in \operatorname{int} Q(t,x)\}$ *est un ouvert de* $I \times H$.

La démonstration repose sur l'égalité: $\forall C, D \in \mathcal{B}(H)$, $h(C,D) = h(\partial C, \partial D)$ et la continuité de l'application $(t,x,y) \to d(y, \partial Q(t,x))$.

LEMME 4.2. *Il existe* $T_0 \in]0,T]$ *et* $x_1(.) \in AC([0,T_0],H)$ *tels que*

(i) $\dot{x}_1(.) - f\Big(.,\pi_K(x_1(.))\Big)$ *soit une constante sur* $[0,T_0]$ *et* $x_1(0) = x_0$.

(ii) $\forall t \in [0,T_0]$, $\dot{x}_1(t) - f\Big(t,\pi_K(x_1(t))\Big) \in \Big(\operatorname{int} G(t,x_1(t))\Big) \cap T_K(x_0) \cap D$.

IDÉE DE LA DÉMONSTRATION. On choisit v_0 dans $(\operatorname{int} G(0,x_0)) \cap T_K(x_0) \cap D$ qui est non vide par hypothèse et on considère une solution du problème de Cauchy $\dot{x}(t) = f\Big(t,\pi_K(x(t))\Big) + v_0$ notée $x_1(.)$ qui, grâce à la proposition 4.1, vérifie (i) et (ii) sur un intervalle $[0,T_0]$. On peut choisir T_0 de sorte que $\int_0^{T_0} k(t)\,dt < 1$.

Soit S l'ensemble de solutions sur $I_0 = [0,T_0]$ du problème

$$\dot{x}(t) = f\Big(t,\pi_K(x(t))\Big) + G(t,x(t)) \quad \text{p.p sur } I_0, x(0) = x_0$$

et S^* le sous-ensemble de S tel que pour tout $x \in S^*$, on a :

a) $\dot{x}(.) - f\Big(.,\pi_K(x(.))\Big)$ est une constante sur chaque $\operatorname{int} J_n$, où $(J_n)_{n\in\mathbb{N}}$ est une suite d'intervalles vérifiant $I_0 = \bigcup_{n\in\mathbb{N}} \overline{J}_n$ et $\sup J_n = \inf J_{n+1}$ $\forall n \in \mathbb{N}$.

b) $x(0) = x_0$, $\dot{x}(t) - f\Big(t,\pi_K(x(t))\Big) \in \Big[\Big(\operatorname{int} G(t,x(t))\Big) \cap D\Big]$ p.p sur I_0.

Il est claire que $x_1(.) \in S^*$. Pour tout $\alpha > 0$ et pour tout $n \in \mathbb{N}^*$, on définit

$$S_G^\alpha = \left\{x \in \overline{S^*} : \int_{I_0} d_G\Big(t,x(t),\dot{x}(t) - f\Big(t,\pi_K(x(t))\Big)\Big)\,dt < \alpha\right\}$$

$$S_d^\alpha = \left\{x \in \overline{S^*} : \int_{I_0} d\Big(\dot{x}(t) - f\Big(t,\pi_K(x(t))\Big), T_K\Big(\pi_K(x(t))\Big)\Big)\,dt < \alpha\right\}$$

$$S^\alpha = S_G^\alpha \cap S_d^\alpha$$

$$R^n = S^{\frac{1}{n}} \quad \text{et} \quad R = \bigcap_{n\in\mathbb{N}^*} R^n.$$

En vertu de la proposition 3.1 et du lemme 3.3, tout élément de R est solution du problème (1.1). Il suffit de prouver que R^n est un ouvert dense dans $\overline{S^*}$.

LEMME 4.3. *Pour tout $\alpha > 0$, S^α est une partie ouverte de $\overline{S^*}$.*

DÉMONSTRATION. Soit $\{x_n\}_{n\in\mathbb{N}} \subset \overline{S^*} \setminus S^\alpha$ telle que $x_n(.) \to x(.)$ dans $\overline{S^*}$. Alors

$$\int_{I_0} d_G\Big(t, x_n(t), \dot{x}_n(t) - f\Big(t, \pi_K\big(x_n(t)\big)\Big)\Big)\, dt \geq \alpha$$

et

$$\int_{I_0} d\Big(\dot{x}_n(t) - f\Big(t, \pi_K\big(x_n(t)\big)\Big), T_K\Big(\pi_K\big(x_n(t)\big)\Big)\Big)\, dt \geq \alpha$$

vue la construction de S^* on peut supposer que

$$\Big\{\dot{x}_n(t) - f\Big(t, \pi_K\big(x_n(t)\big)\Big) : t \in I_0, n \in \mathbb{N}\Big\}$$

est une partie dénombrable du compact D; donc quitte à extraire un sous-suite, on suppose que $\Big(\dot{x}_n(.) - f\Big(t, \pi_K\big(x_n(.)\big)\Big)\Big)_{n\in\mathbb{N}}$ converge; notons y sa limite. Comme $\dot{x}_n(.) \to \dot{x}(.)$ faiblement, on a $y = \dot{x}(.) - f\Big(., \pi_K\big(x(.)\big)\Big)$. On achève la démonstration en utilisant la proposition 3.1 et la proposition 3.2.

LEMME 4.4. *Soient $x \in S^*$, $\alpha > 0$ et J_0 un intervalle tels que $\dot{x}(.) - f\Big(., \pi_K\big(x(.)\big)\Big)$ soit une constante sur $\operatorname{int} J_0 =]0, t_1[$ $(t_1 > 0)$. Il existe alors une suite (s_n) dans $]0, t_1[$, une suite $(y_n)_{n\in\mathbb{N}}$ dans $AC(J_0, H)$ et une famille $(P_n)_{n\in\mathbb{N}}$ de suite d'intervalles $(J_q^n)_{q\in\mathbb{N}}$ vérifiant : pour tout $n, q \in \mathbb{N}$ $\sup J_q^n = \inf J_{q+1}^n$, tels que*

(i) $y_n(0) = x_0$, $\dot{y}_n(t) \in \Big[\Big(\operatorname{int} G\big(t, y_n(t)\big)\Big) \cap D\Big]$, $\forall t \in [0, s_n[$.

(ii) $\dot{y}_n(.) - f\Big(., \pi_K\big(y_n(.)\big)\Big)$ *est une constante sur chaque* $\operatorname{int} J_q^n$, $\forall q \in \mathbb{N}$.

(iii) $\int_0^{s_n} d_G\Big(t, y_n(t), \dot{y}_n(t) - f\Big(t, \pi_K\big(y_n(t)\big)\Big)\Big)\, dt \leq \frac{\alpha s_n}{2T_0}$.

(iv) $\int_0^{s_n} d\Big(\dot{y}_n(t) - f\Big(t, \pi_K\big(y_n(t)\big)\Big), T_K\Big(\pi_K\big(y_n(t)\big)\Big)\Big)\, dt \leq \frac{\alpha s_n}{2T_0}$.

(v) $\sup\{\|y_n(t) - x(t)\| : t \in J_0\} \to 0$ *si* $n \to +\infty$.

DÉMONSTRATION. Notons par a la constante $\dot{x}(.) - f\Big(., \pi_K\big(x(.)\big)\Big)$ sur $\operatorname{int} J_0 =]0, t_1[$. Pour tout $n \in \mathbb{N}^*$ et $i \in \{0, \ldots, n\}$ on pose $t_i^n = \frac{it_1}{n}$. D'après l'hypothèse (H_5) il existe $\lambda_j^n > 0$, $b_j^n \in \big[\big(\operatorname{ext} G(0, x_0)\big) \cap T_K(x_0) \cap D\big]$ avec $j = 1, \ldots, m_n$, $m_n \in \mathbb{N}$, tels que $\sum_{j=1}^{m_n} \lambda_j^n = 1$ et

$$\Big\|a - \sum_{j=1}^{m_n} \lambda_j^n b_j^n\Big\| < \frac{1}{2^n}. \tag{1}$$

En vertu de la proposition 3.1, du lemme 3.2 et de la proposition 4.1, il existe $\gamma_0 \in]0, 1[$ et $\zeta_n > 0$ tels que pour tout $(t, x) \in B\big((t, x_0), \zeta_n\big) \cap I_0 \times H$, on a

$$c_j^n(\gamma_0) \in [\operatorname{int} G(0, x_0) \cap T_K(x_0) \cap D] \tag{2}$$

$$\max\Big\{d_G\big(t, x, c_j^n(\gamma_0)\big), d\Big(c_j^n(\gamma_0), T_K\big(\pi_K(x)\big)\Big)\Big\} < \frac{\alpha}{2T_0} \tag{3}$$

où $c_j^n(\gamma_0) = \gamma_0 a + (1-\gamma_0)b_j^n$. Quitte à choisir γ_0 assez petit, (1) entraîne

$$\left\| a - \sum_{j=1}^{m_n} \lambda_j^n c_j^n(\gamma_0) \right\| < \frac{1}{2^n}. \tag{4}$$

Pour $i \in \{0,\dots,n\}$ et $j \in \{1,\dots,m_n\}$ on définit

$$\tau_{i,0}^n = t_i^n, \quad \tau_{i,j}^n = \tau_{i,j-1}^n + \lambda_j^n \frac{t_1}{n} \quad \text{et} \quad \Delta_{i,j}^n = [\tau_{i,j-1}^n, \tau_{i,j}^n].$$

Observons que pour tout $i \in \{0,\dots,n-1\}$, $\bigcup_{j=1}^{m_n} \Delta_{i,j}^n = [t_i^n, t_{i+1}^n]$. Soit $y_{0,1}^n(.)$ une solution sur $\Delta_{0,1}^n$ du problème $\dot{x}(t) = f\Big(t, \pi_K\big(x(t)\big)\Big) + c_1^n(\gamma_0)$, $x(0) = x_0$. Par récurrence, on considère $y_{0,j}^n(.)$ une solution sur $\Delta_{0,j}^n$ du problème $\dot{x}(t) = f\Big(t, \pi_K\big(x(t)\big)\Big) + c_j^n(\gamma_0)$, $x(\tau_{0,j}^n) = y_{0,j-1}^n(\tau_{0,j}^n)$, $y_{i,j}^n(.)$ une solution sur $\Delta_{i,j}^n$ du problème $\dot{x}(t) = f\Big(t, \pi_K\big(x(t)\big)\Big) + c_j^n(\gamma_0)$, $x(\tau_{i,j}^n) = y_{i-1,j}^n(\tau_{i,j}^n)$. Finalement, on définit sur $[0,t_1]$ la fonction $y_n(t) = \sum_{i=0}^{n-1} \chi_{[t_i^n, t_{i+1}^n]}(t) y_i^n(t)$ où pour tout $t \in [t_i^n, t_{i+1}^n]$, $y_i^n(t) = \sum_{j=1}^{m_n-1} \chi_{\Delta_{i,j}^n}(t) y_{i,j}^n(t)$.

Pour tout $n \in \mathbb{N}$, choisissons un réel s_n tel que $0 < s_n < \min\{\zeta_n, \frac{\zeta_n}{2M}\}$ et $\frac{1}{s_n}\int_0^{s_n} g(t)\,dt \le \frac{M}{2}$ où $M = \sup\{\|x\|, x \in D\}$. Alors pour tout $t \in [0, s_n]$, on a

$$\|y_n(t) - x_0\| = \int_0^t \|\dot{y}_n(s)\|\,ds \le M s_n.$$

Tenant compte de (2), il vient alors

$$\Big(\dot{y}_n(t) - f\Big(t, \pi_K\big(y_n(t)\big)\Big)\Big) \in \big[\operatorname{int} G(t, y_n(t)) \cap T_K(x_0) \cap D\big]$$

$$d_G\Big(t, y_n(t), \dot{y}_n(t) - f\Big(t, \pi_K\big(y_n(t)\big)\Big)\Big) \le \frac{\alpha}{2T_0}$$

$$d\Big(\dot{y}_n(t) - f\Big(t, \pi_K\big(y_n(t)\big)\Big), T_K\Big(\pi_K\big(y_n(t)\big)\Big)\Big) \le \frac{\alpha}{2T_0}.$$

Pour achever la démonstration, il suffit d'établir la propriété (v). En effet,

$$\|y_n(t_i^n) - x(t_i^n)\| \le \frac{t_i}{2^n} + \sup_{t\in[0,t_1]} \|y_n(t) - x(t)\| \int_0^{t_i} k(t)\,dt, \quad \forall n, \forall i = 0,\dots,n.$$

Ainsi, pour tout $t \in J_0$, si nous notons i_n l'indice tel que $t \in [t_{i_n}^n, t_{i_n+1}^n]$, alors

$$\begin{aligned}\|y_n(t) - x(t)\| &\le \|y_n(t) - y_n(t_{i_n}^n)\| + \|y_n(t_{i_n}^n) - x(t_{i_n}^n)\| + \|x(t_{i_n}^n) - x(t)\| \\ &\le 2\frac{M t_1}{n} + \frac{t_1}{2^n} + \sup_{t\in[0,t_1]} \|y_n(t) - x(t)\| \int_0^{t_1} k(t)\,dt.\end{aligned}$$

Ceci achève la démonstration du lemme 4.4.

LEMME 4.5. *S^α est dense dans $\overline{S^*}$.*

IDÉE DE LA DÉMONSTRATION. Soit $x \in S^*$. D'aprés le lemme 4.4, il existe $\big(s,(s_n),(y_n)\big)$ dans $]0,T_0]\times]0,s]\times AC([0,s],H)$ vérifiant les propriétés (i)–(v) du lemme. Grâce au lemme de Zorn, on démontre l'existence d'un élément maximal (pour un ordre bien précis) noté $\big(s^m,(s_n^m)_n,(y_n^m)_n\big)$. On démontre que pour tout $n \in \mathbb{N}$, $s_n = T_0$. Ceci achève la preuve du théorème 2.1.

Pour démontrer le corollaire, il suffit de poser $f(t,x) = \pi_{C(t)}(x)$ et d'appliquer le théorème 2.1.

REFERENCES

1. J. P. Aubin et A. Cellina, *Differential Inclusions*. Springer-Verlag, 1984.
2. C. Castaing et M. Valadier, *Convex Analysis and Measurable Multifunctions*. Lecture Notes in Math. **580**, Springer-Verlag, 1977.
3. G. Choquet, *Lectures on Analysis*. W. A. Benjamin, 1969.
4. G. Haddad, *Monotone Trajectories of Differential Inclusions and Functional Differential Inclusions with Memory*. Israel J. Math. **39**(1981), 83–100.
5. S. Sajid, *Solution Viable d'une Inclusion Différentielle non Convexe*. C. R. Acad. Sci. Paris Sér. I **324**(1997), 143–148.

Département de Mathématiques
FSTM
BP. 146
Mohammadia
Maroc
courriel: sajid@uh2m.ac.ma

C. R. Math. Rep. Acad. Sci. Canada Vol. **23** (2), 2001 pp. 39–45

ON A CONJECTURE OF KEMNITZ

R. THANGADURAI

Presented by M. Ram Murty, FRSC

ABSTRACT. In this note, we extend the work of W. D. Gao [6] on a conjecture of Kemnitz.

RÉSUMÉ. Dans cette Note, nous étendons le travail de W. D. Gao [6] sur une conjecturé de Kemnitz.

1. Introduction. For $d, n \in \mathbb{N}$, let $f(n,d)$ denote the smallest positive integer such that every sequence of f numbers of, not necessarily distinct, integer lattice points in $\mathbb{Z}^d$ contains a subsequence of size n whose sum is divisible by n.

The main problem is to find the exact values of $f(n,d)$ for all $n, d \in \mathbb{N}$.

Note that given a sequence $x_1, x_2, \ldots, x_{f(n,d)}$ of integer lattice points in $\mathbb{Z}^d$ where $x_i = (r_{i1}, r_{i2}, \ldots, r_{id})$, we write $r_{ij} = s_{ij} + nm_{ij} \forall i = 1, 2, \ldots, f(n,d)$ and $j = 1, 2, \ldots, d$ where m_{ij} is an integer and $0 \leq s_{ij} \leq n-1$. Let $y_i = (s_{i1}, s_{i2}, \ldots, s_{id}) \in \mathbb{Z}^d$ for all $i = 1, 2, \ldots, f(n,d)$. If we can find an n-element subset of $\{y_1, y_2, \ldots, y_{f(n,d)}\}$ whose sum is zero modulo n, then by translation, we arrive at an n-element subset of $\{x_1, x_2, \ldots, x_{f(n,d)}\}$ whose sum is also zero modulo n and vice versa. Thus, henceforth, we always assume that the given integer lattice points are elements of $(\mathbb{Z}/n\mathbb{Z})^d$.

The existence of $f(n,d)$ is clear from the following inequalities which are obtained using the simple, yet powerful Dirichlet's Pigeon Hole principle. Clearly,

$$1 + 2^d(n-1) \leq f(n,d) \leq 1 + n^d(n-1). \tag{1}$$

To get the lower bound, we have to take all the d-tuples (2^d in number) with coordinates 0 or 1, each with multiplicity $(n-1)$. The upper bound follows from the fact that any sequence of $1 + n^d(n-1)$ elements of $(\mathbb{Z}/n\mathbb{Z})^d$ must contain the same vector n times. Thus, $f(2,d) = 2^d + 1$.

The upper bound in equation (1) is too weak. With sophisticated techniques from Combinatorial Number Theory, Alon and Dubiner [2] proved that

$$f(n,d) \leq c(d)n$$

where $c(d)$ is an absolute constant which depends only on d.

Received by the editors February 2, 2000.
AMS subject classification: Primary: 11P21; secondary: 11B50.
© Royal Society of Canada 2001.

When the values of f are known for m and n, Harborth [8] gave a formula for an upper bound of $f(nm,d)$ as follows.

LEMMA 1.A (HARBORTH [8]). *For each integers $n,m \geq 1$, we have*

$$f(nm,d) \leq \min\Big(f(n,d)+n\big(f(m,d)-1\big), f(m,d)+m\big(f(n,d)-1\big)\Big).$$

When $d=1$, from equation (1), we get $2n-1 \leq f(n,1) \leq n^2-n+1$. In 1961, Erdős, Ginzburg and Ziv [5] (much before the general result [2]) completely solved this case by proving the formula $f(n,1)=2n-1$.

Thus the one-dimensional case was completely solved. Now let us pass onto the case of dimension two of this lattice point problem.

The two-dimensional case was first considered by Harborth [8] and he proved that $f(3,2)=9$. Kemnitz [9], in 1983, proved that $f(p,2)=4p-3$ for $p=5,7$. These results together with the above Lemma 1.A implies that $f(n,2)=4n-3$ for all $n=2^a3^b5^c7^r$ where $a,b,c,r \in \mathbb{N}$. Moreover, in the same paper, Kemnitz [9] proved the following theorem.

THEOREM 1.B (KEMNITZ [9]). *If $a_1,a_2,\ldots,a_{4p-3}$ is a sequence of $4p-3$ elements in $(\mathbb{Z}/p\mathbb{Z})^2$ such that all a_i's are distinct modulo p, then we can find a p-element subsequence whose sum is zero modulo p. Here p is a prime number.*

REMARK 1.1. Theorem 1.B suggests that we can henceforth consider $4p-3$ lattice points with one of the lattice points repeated at least twice.

Also, Alon and Dubiner [2] proved that for any given $3p$ lattice points in $\mathbb{Z}^2$ such that all $3p$ lattice points add up to zero modulo p, we can find p-element subset whose sum is zero modulo p.

CONJECTURE. $f(n,2)=4n-3$ for all $n \in \mathbb{N}$.

This was first conjectured by Kemnitz and was suggested, independently, by N. Zimmerman and Y. Peres (see for instance [3]).

REMARK 1.2. If we prove $f(p,2)=4p-3$ for all primes, then from Lemma 1.A we get $f(n,2) \leq 4n-3$ for all positive integer n. But the lower bound of the equation (1) is just $4n-3$. Putting together we get $f(n,2)=4n-3$. Thus, it is enough to prove the above conjecture for the primes alone.

Alon and Dubiner [2] proved that $f(n,2) \leq 6n-5$ for all n. This was improved later by Weidong Gao [7] for all primes p by proving $f(p,2) \leq 5p-1$. The present author has been recently informed that Lajos Rónyai [10] has almost proved Kemnitz's conjecture; he proved that $f(p,2) \leq 4p-2$ for all primes p, using the methods in a recent paper of Alon [1].

REMARK 1.3. Since $f(p,2) \leq 4p-2 \leq 5p-4$, from [10] and using Lemma 1.A, we can get $f(n,2) \leq 5n-4$ for all positive integer n.

Related to the expected bound, in 1996, Weidong Gao [6], proved that if $f(n,2)=4n-3$ and $n \geq \big((3m-4)(m-1)m^2+3\big)/4m$ with $m \geq 2$, then

the result is true for nm. To prove this result he has used the following crucial observation. If $4n-3$ integer lattice points are given in the plane with one of the lattice points is repeated at least $n-1$ times, then there is an n-element subset whose sum is zero modulo n.

One should mention two expository articles ([4] and [12]) around this topic. Also, it is known that the lower bound in equation (1) is not tight for dimension $d > 2$. Harborth proved that $f(3,3) = 19$ in [8] which is greater than the lower bound 17.

Indeed our main theorem, here, is an improvement of Gao's result as well as his crucial lemma. In contrast to Gao's idea, we shall not use the crucial lemma for proving the main theorem. Rather, we independently prove these along the same line.

In our method we often use the following result of van Emde Boas [13]. This result was first proved for primes by Olson [11]. Then it was extended to all natural numbers by van Emde Boas [13] by an easy induction.

LEMMA 1.C (VAN EMDE BOAS [13]). *If $a_1, a_2, \ldots, a_{3n-2}$ is a sequence of $3n-2$ elements in $(\mathbb{Z}/n\mathbb{Z})^2$, then we can find an integer t, $1 \leq t \leq n$ such that there is a t-element subsequence whose sum is zero modulo n.*

Our main interest is to prove the following theorems.

THEOREM 2.1. *If a sequence of $4n-3$ integer lattice points in $\mathbb{Z}^2$ such that one of the given lattice points is repeated at least $\lfloor n/2 \rfloor$ times, then we can find an n-element subset whose sum is zero modulo n. Here $\lfloor . \rfloor$ denotes the floor function.*

THEOREM 3.1. *If $f(n,2) = 4n-3$ and $n > (2m^3 - 3m^2 + 3)/4m$, for some integer $m \geq 2$, then, $f(nm,2) = 4nm - 3$.*

2. Proof of Theorem 2.1 Let

$$S := \{a_1, a_2, \ldots, a_{4n-3}\}$$

be the given set of integer lattice points in the plane. It suffices to consider the a_i's as elements in the group $(\mathbb{Z}/n\mathbb{Z})^2$. Let $a \in (\mathbb{Z}/n\mathbb{Z})^2$ be repeated at least $\lfloor n/2 \rfloor$ times. Also we can assume that a is the element repeated maximum number of times. Otherwise, we could have chosen that element which is repeated more than a. Also from the result of W. D. Gao [6], we can assume that a is repeated at most $n-2$ times.

By rearranging the indices, we assume that $a_1 = a_2 = \cdots = a_s = a$ where $\lfloor n/2 \rfloor \leq s \leq n-2$. Translate the given $4n-3$ integer lattice points by a. We get, $\underbrace{0, 0, \ldots, 0}_{s \text{ times}}$ repeated s times and $S^* := \{b_1, b_2, \ldots, b_{4n-3-s}\}$ where b_i's are the non-zero elements of the translated set S. Since s is at most $n-2$, we have $4n-3-s > 3n-2$.

We know from Theorem 1.C that if T consists of $3n-2$ number of integer lattice points in $(\mathbb{Z}/n\mathbb{Z})^2$, then it contains a t-element subset whose sum is zero modulo n with $1 \leq t \leq n$. We pair such a subset T with an integer t. If we write (T,t) we mean that $|T| = 3n-2$ and the integer t attached to T in the above manner.

From S^*, we collect all possible subsets T_i's with $|T_i| = 3n-2$. So corresponding to each subset T_i, we have an integer t_i. The number of such pairs (T_i, t_i) in S^* is equal to

$$\binom{4n-3-s}{3n-2} = \binom{4n-3-s}{n-1-s} \geq 4n-3-s.$$

Let T be one of the above subsets T_i's of S^* such that the corresponding t (in the pairing (T,t)) is the maximum of all such t_i's. We choose one such pair and we denote it by (T,t).

We can assume that $t < n-s$. For, if $s+t \geq n$, then adding some zeros to T we get an n-element subset whose sum is zero modulo n.

CLAIM. $\lfloor n/2 \rfloor + 1 \leq t \leq n$.

Assume the contrary. That is, if t had been at most $\lfloor n/2 \rfloor$, then $4n-3-s-\lfloor n/2 \rfloor \geq 3n-2$, as $s+t < n$. Therefore, we can get the next maximal pair (T_1, t_1) with $t_1 \leq t$. So, $1 \leq t_1 \leq \lfloor n/2 \rfloor$. This implies $t+t_1 \leq n$. In that case, considering $T \cup T_1$ we would have chosen $t+t_1$ as our t in the first step. Hence t has to be strictly more than $\lfloor n/2 \rfloor$.

Now, observe that we have s zeros outside S^* and a t-element set from S^* whose sum is zero modulo n. Since $s \geq \lfloor n/2 \rfloor$ and $\lfloor n/2 \rfloor + 1 \leq t \leq n$, we thus get an n-element set whose sum is zero modulo n, by adding to the t-element set the appropriate number of zeros from the s zeros outside S^*. ∎

REMARK 2.2. In Theorem 2.1, if we assume n is odd and one of the lattice points is repeated at least $\lfloor n/2 \rfloor - 1 = (n-3)/2$ times, then in this method, we cannot prove the assertion. For, since $4n-3-(n-3)/2 = 3n-2+(n+1)/2$, we can choose a maximal pair (T,t) such that $(n+1)/2 \leq t \leq n$. Suppose $t = (n+1)/2$. Then since we have $3n-2+(n+1)/2-(n+1)/2 = 3n-2$ lattice points left with, we can choose a pair (T_1, t_1). We must have $t+t_1 > n$ implies $t_1 \geq (n+1)/2$ which forces $t_1 = (n+1)/2$. Hence in this method we cannot get an n element subset whose sum is zero modulo n.

As a corollary to the above Theorem, we arrive at a new proof of some known results as follows.

COROLLARY 2.3. *$f(p,2) = 4p-3$ for $p = 3, 5$.*

PROOF. Let $\{a_1, a_2, \ldots, a_{4p-3}\}$ be a sequence of $4p-3$ integer lattice points in $\mathbb{Z}^2$. It is enough to consider all the a_i's in $(\mathbb{Z}/p\mathbb{Z})^2$. By Lemma 1.B, it is enough to assume that at least one of the given lattice points is repeated twice. Without

loss of generality it is enough to assume that the zero element is repeated at least twice. When $p = 3, 5$, by Theorem 2.1, the above observation gives the desired result. ∎

3. **Proof of Theorem 3.1** It is enough to prove that $f(nm, 2) \leq 4nm - 3$. Consider

$$S = \{a_1, a_2, \ldots, a_{4nm-3}\} \subset \mathbb{Z}^2$$

possibly with repetitions. Without loss of generality we can assume $S \subset [0, nm-1] \times [0, nm-1]$.

Let S_m be the set of all elements of S modulo m. Then, we see that there exists an element $x \in S_m$ which is repeated maximum number of times. We can assume x to be the zero element of S_m, if necessary by translating the elements of S. Note that in S_m at least $[(4nm-3)/m^2] + 1$ zeros are available.

Let S_m^* be the set of all non-zero elements of S_m. From S_m^* take out all possible $k \geq 0$ disjoint non-empty subsets $R_1, R_2, \ldots, R_k$ with $|R_i| = m$ such that $\sum_{r \in R_i} r \equiv 0 \pmod{m}$ $\forall\, i = 1, 2, \ldots, k$. Hence, $W := S_m^* \setminus (\bigcup_{i=1}^{k} R_i)$ contains no m-element subset whose sum is zero modulo m. Hence by Remark 1.3, we have $|W| \leq 5m - 5$.

If $|W| \geq 3m - 2$, then by Theorem 1.C, we can find a natural number t, $2 \leq t \leq m-1$ such that we find such a t-element subset of W sums to zero. Let B_1 be a maximal subset of W such that $|B_1| = t_1$ with $2 \leq t_1 \leq m-1$ and its sum is zero modulo m. Then we can take A_1 which contains $m - t_1$ zeros and together with B_1 we get an m-element subset whose sum is zero modulo m.

If $|W \setminus B_1| \geq 3m - 2$, we can find B_2 which is the maximal subset of $W \setminus B_1$ with $|B_2| = t_2$ with $2 \leq t_2 \leq m-1$ whose sum is zero modulo m. Note that $t_1 \geq t_2$ and $|B_1 \cup B_2| > m$. If not, we would have chosen $B_1 \cup B_2$ in the first step and it would have contradicted the maximality. Once we have chosen B_2, take A_2, a subset of all zeros disjoint from A_1 and having cardinality $m - |B_2|$. Then, $A_2 \cup B_2$ produces an m-element subset of S_m whose sum is zero modulo m.

Continue this process, until we arrive at $|W \setminus (\bigcup_{i=1}^{\ell} B_i)| \leq 3m - 3$ where ℓ is a non-negative integer.

Therefore, we would have used at most $2(m-1)$ zeros which we have out side S_m^* to bring down the cardinality of W from $5m-5$ to $3m-3$. Hence in order to make sure that there are at least $2(m-1)$ zeros in S_m, we need the following condition,

$$|S_m \setminus S_m^*| \geq \left[\frac{4nm-3}{m^2}\right] + 1 \geq 2(m-1).$$

This implies, $n > (2m^3 - 3m^2 + 3)/(4m)$.

Note that we are not using Theorem 2.1 to produce a zero-sum subsets of cardinality m in the above process.

If $|(S_m \setminus S_m^*) \setminus (\bigcup_{i=1}^{\ell} A_i)| \geq m$, remove all possible disjoint m-element subsets and call the remaining set be A. Clearly $1 \leq |A| \leq m-1$.

Let us count all the disjoint m-element subsets of S_m whose sum is zero modulo m. Let the number be t. Then,

$$tm = 4nm - 3 - \left|W \setminus \left(\bigcup_{i=1}^{\ell} B_i\right)\right| - |A|.$$

Hence,

$$t \geq \frac{1}{m}\left(4nm - 3 - (3m-3) - (m-1)\right) = 4n - 4 + 1/m.$$

Since t is integer, $t \geq 4n-3$. Hence we have $I_1, I_2, \ldots, I_{4n-3}$ disjoint m-element subsets of S such that $\sum_{b\in I_j} b \equiv 0 \pmod{m}$ for every $j = 1, 2, \ldots, 4n-3$. Write $c_j = 1/m \sum_{b\in I_j} b$. Since $f(n,2) = 4n-3$ and we have $4n-3$ number of $c_1, c_2, \ldots, c_{4n-3}$ integer lattice points, there exist n element subsequence $c_{i_1}, c_{i_2}, \ldots, c_{i_n}$ such that its sum is zero modulo n. Thus we get

$$\sum_{j=1}^{n} c_{i_j} \equiv 0 \pmod{n} \Longrightarrow \sum_{j=1}^{n} \sum_{b\in I_{i_j}} b \equiv 0 \pmod{mn}.$$

Hence, we get nm-elements from the set S such that their sum is zero modulo nm. ∎

COROLLARY 3.2. *Let $n = 2^a 3^b 5^c 7^d$ with $a, b, c, d \geq 0$, and let $m_1 \geq m_2 \geq \cdots \geq m_k$. Suppose that $n > (2m_1^3 - 3m_1^2 + 3)/(4m)$. Then,*

$$f(nm_1m_2\cdots m_k, 2) = 4nm_1m_2\cdots m_k - 3.$$

PROOF. Using Theorem 1.A and the known results, we get $f(n,2) = 4n-3$ for $n = 2^a 3^b 5^c 7^d$. Hence by Theorem 3.1, we have $f(nm_1, 2) = 4nm_1 - 3$. The result follows by induction on k. ∎

REMARK 3.3. Theorem 3.1 can be marginally improved if we take $m = p$ a prime number. By the recent result of Lajos Rónyai [10] in which he proved that $f(p,2) \leq 4p-2$ for all primes p. Using this result, if we proceed as in Theorem 3.1, we arrive at the following conditions on n. That is, we have to make sure that there are at least p zeros in $S_p \setminus S_p^*$. Thus we arrive at

$$\frac{4np-3}{p^2} + 1 \geq p \Longrightarrow n \geq (p^2(p-1)+3)/(4p).$$

In this way, we can prove the following theorem.

THEOREM 3.4. *If $f(n,2) = 4n-3$ and $n > (p^2(p-1)+3)/(4p)$ for some prime number p, then, $f(np,2) = 4np-3$.*

ACKNOWLEDGEMENTS. This work is a part of my Ph.D. thesis done at Harish Chandra Research Institute (formerly, Mehta Research Institute), Allahabad. I would like to thank Professor R. Balasubramanian for valuable discussions. I also thank Dr. S. D. Adhikari, my thesis advisor, for his constant encouragements and for his help to improve the presentation of this paper.

REFERENCES

1. N. Alon, *Combinatorial Nullstellensatz.* Combin. Probab. Comput. **8**(1999), 7–29.
2. N. Alon and M. Dubiner, *A lattice point problem and additive number theory.* Combinatorica (3) **15**(1995), 301–309.
3. ———, *Zero-sum sets of prescribed size.* In: Combinatorics, Paul Erdös is Eighty, Vol. **1**, Jànos Bolyai Math. Soc., Budapest, 1993, 33–50.
4. Y. Caro, *Zero-sum problems—A survey.* Discrete Math. **152**(1996), 93–113.
5. P. Erdös, A. Ginzburg and A. Ziv, *Theorem in the additive number theory.* Bull. Res. Council Israel **10 F**(1961), 41–43.
6. W. D. Gao, *On zero-sum subsequences of restricted size.* J. Number Theory **61**(1996), 97–102.
7. ———, *Addition Theorems and Group Rings.* J. Combin. Theory Ser. A **77** (1997), 98–109.
8. H. Harborth, *Ein Extremalproblem Für Gitterpunkte.* J. Reine Angew. Math. **262/263** (1973), 356–360.
9. A. Kemnitz, *On a lattice point problem.* Ars Combin. **16b**(1983), 151–160.
10. Lajos Rónyai, *On a conjecture of Kemnitz.* To appear.
11. J. E. Olson, *On a combinatorial problem of Erdös, Ginzburg and Ziv.* J. Number Theory **8**(1976), 52–57.
12. R. Thangadurai, *Some Direct and Inverse problems in additive number theory.* Bull. Allahabad Math. Soc. **12-13**(1997/98), 37–55.
13. P. van Emde Boas, *A combinatorial problem on finite abelian groups, II.* Math. Centre Amsterdam **1969**, ZW-1969-007.

The Institute of Mathematical Sciences
C. I. T. Campus, Taramani
Chennai 600 113
India
email: thanga@imsc.ernet.in

C. R. Math. Rep. Acad. Sci. Canada Vol. **23** (2), 2001 pp. 46–52

A CONDITIONAL PROOF OF ARTIN'S CONJECTURE FOR PRIMITIVE ROOTS

AMORA NONGKYNRIH

Presented by M. Ram Murty, FRSC

ABSTRACT. Assuming a hypothesis which is weaker than the generalized Riemann hypothesis, this paper gives a proof of Artin's conjecture that any integer $a \neq \pm 1$ or a perfect square is a primitive root (mod p) for infinitely many primes p.

RÉSUMÉ. En supposant une hypothèse qui est plus faible que l'hypothèse de Riemann généralisée, cet article donne une démonstration de la conjecture d'Artin qui dit que tout entier $a \neq \pm 1$ ou un carré parfait est racine primitive (modulo p) pour une infinité de premiers p.

A conjecture of E. Artin formulated in 1927, states that any integer $a \neq \pm 1$ or a perfect square is a primitive root (mod p) for infinitely many primes p. Moreover, if $N_a(x)$ denotes the number of such primes up to x, he conjectured an asymptotic formula of the form $N_a(x) \sim A(a)\frac{x}{\log x}$ as $x \to \infty$, where $A(a)$ is a constant depending on a.

In 1967, Hooley [1] proved Artin's conjecture and an asymptotic formula for $N_a(x)$ subject to the assumption of the generalized Riemann hypothesis for Dedekind zeta functions of certain number fields. Recall that Hooley proved the following theorem [1]:

If it is assumed that the generalized Riemann hypothesis holds for the Dedekind zeta function over Kummer fields of the type $\mathbf{Q}(\sqrt[k]{2}, \sqrt[k]{1})$, *where* k *is square-free, then we have*

(a) Let $N_2(x)$ *be the number of primes* p *not exceeding* x *for which 2 is a primitive root modulo* p. *Then*

$$N_2(x) = \frac{Ax}{\log x} + O\left(\frac{x \log\log x}{\log^2 x}\right) \text{ where } A = \prod_q \left(1 - \frac{1}{q(q-1)}\right).$$

(b) There are infinitely many primes p *for which 2 is a primitive root modulo* p.

He also remarked that the second part of the theorem is still true if no zero of the zeta functions over $\mathbf{Q}(\sqrt[k]{2}, \sqrt[k]{1})$ has real part exceeding $1 - \frac{1}{2}e^{-1} - \delta$.

In this paper, we show that it is possible to prove Artin's conjecture on a hypothesis weaker than the generalized Riemann hypothesis. However, our proof

Received by the editors February 10, 2000.
AMS subject classification: Primary: 11A07; secondary: 11M26, 11N36.
© Royal Society of Canada 2001.

rests on a hypothesis which is slightly weaker than the one remarked by Hooley. We shall consider the particular case when $a = 2$. The general case differs from this only in that there are some extra points of detail that require discussion. The precise result we prove is as follows.

THEOREM 1. *Let* $K = \mathbf{Q}(\sqrt[k]{2}, \sqrt[k]{1})$, k *a square-free integer. Assume that the Dedekind zeta function* $\zeta_K(s)$ *is zero-free in* $\mathrm{Re}(s) > 1 - \frac{1}{2}e^{-12A/5} - \delta$, *where* $A = \prod_q(1 - \frac{1}{q(q-1)})$, q *prime. Then 2 is a primitive root for a positive proportion of primes, i.e.,*

$$N_2(x) \gg \frac{x}{\log x}.$$

REMARK. Notice that $A = \prod_q(1 - \frac{1}{q(q-1)}) < \frac{1}{2} \cdot \frac{5}{6} = \frac{5}{12}$ which implies that $\frac{12A}{5} < 1$ so that our result falls short of the remark made by Hooley.

A part of Hooley's proof does not assume any hypothesis. This will also hold in our case with a few minor changes in the choice of parameters. We shall then apply a variant of the Bombieri-Vinogradov theorem proved in [2] for the part of the proof which requires the assumption of a zero-free region for the Dedekind zeta function of the form stated in Theorem 1.

The principle underlying Hooley's treatment of Artin's conjecture was that of the *simple asymptotic sieve* [1]. In order to formulate the problem in sieve theory notation, we make use of the following observation: 2 is a primitive root modulo p if and only if $p \neq 2$ and there is no prime divisor q of $p-1$ for which 2 is a q-th power residue mod p.

Let $p > 2$. For any such prime p and for any prime q, let $R(q,p)$ denote the simultaneous conditions "2 is a q-th power residue mod p, $q \mid p-1$"; and for any square-free integer k, let the generalized symbol $R(k,p)$ indicate the conjunction of $R(q,p)$ for all prime divisors q of k.

We set up the following notation: $S(x,\eta)$ counts the number of primes p up to x that do not satisfy $R(q,p)$ for any q not exceeding η. Then the criterion for primitive roots implies that $N_2(x) = S(x, x-1)$. For any square-free integer k, $P(x,k)$ counts the number of primes up to x for which $R(k,p)$ hold (no condition being implied if $k = 1$). Finally, $M(x,\eta_1,\eta_2)$ counts the number of primes p up to x for which $R(q,p)$ hold for at least one prime q satisfying $\eta_1 < q \leq \eta_2$.

We shall need the following theorem proved in a paper of Ram Murty and Kumar Murty [2]:

NOTATION. Let K be a Galois extension of $\mathbf{Q}$, $G = \mathrm{Gal}(K/\mathbf{Q})$, C a conjugacy class in G. Let l, q be positive integers with $1 \leq l \leq q$, $(l,q) = 1$. Denote by $\pi_C(x,q,l)$ the number of primes $p \leq x$ which are ramified in K, which satisfy $(p, K/\mathbf{Q}) = C$ and $p \equiv l \pmod q$. Here $(p, K/\mathbf{Q})$ is the Artin symbol of p in G.

Let H be the largest abelian subgroup of G, $H \cap C \neq \phi$, $d = [G : H]$, and let

$$\eta = \begin{cases} d-2 & \text{if } d \geq 4 \\ 2 & \text{if } d \leq 4. \end{cases}$$

THEOREM 2. *Let* $Q = x^{\frac{1}{\eta}-\epsilon}$. *Then for any* $A > 0$,

$$\sum_{q<Q}{}' \max_{(l,q)=1} \max_{y\le x} \left|\pi_C(y,q,l) - \frac{|C|}{|G|}\frac{\pi(y)}{\phi(q)}\right| \ll \frac{x}{(\log x)^A}$$

where the prime on the summation indicates that we range only over those q *satisfying* $\mathbf{Q}(\zeta_q)\cap K = \mathbf{Q}$.

PROOF OF THEOREM 1. The simple asymptotic sieve applied to the problem under consideration gives

$$\begin{aligned} N_2(x) &= S(x,\xi_1) + O\big(M(x,\xi_1,\xi_2)\big) + O\big(M(x,\xi_2,\xi_3)\big) \\ &\quad + O\big(M(x,\xi_3,\xi_4)\big) + O\big(M(x,\xi_4,x-1)\big) \end{aligned} \tag{1}$$

where $\xi_1 = \alpha\log x$ with $0<\alpha<1$, the value of α will be chosen in due course; $\xi_2 = x^\theta/\log^2 x$ with $\theta < 1/2$; $\xi_3 = x^{\frac{1}{2}-\epsilon}$; $\xi_4 = x^{1/2}\log x$.

The last two terms on the right of (1) can be estimated without any hypothesis. As in [1], using the Brun-Titchmarsh theorem we obtain

$$\begin{aligned} M(x,\xi_3,\xi_4) &\le \sum_{\xi_3<q\le\xi_4} P(x,q) \\ &\ll \frac{\epsilon x}{\log x}, \end{aligned}$$

with ξ_3 and ξ_4 as chosen above. To estimate the final term we observe that the condition $R(q,p)$ implies that

$$2^{\frac{p-1}{q}} \equiv 1 (\mathrm{mod}\ p).$$

Therefore, since $q > x^{1/2}\log x$ and $p \le x$, any primes p that the sum $M(x,\xi_4,x-1)$ counts must divide the product

$$T = \prod_{m<x^{1/2}(\log x)^{-1}} (2^m - 1).$$

The number of prime divisors of T is $\ll \frac{x}{\log^2 x}$. Therefore (1) reduces to

$$N_2(x) = S(x,\xi_1) + O\big(M(x,\xi_1,\xi_2)\big) + O\left(\frac{\epsilon x}{\log x}\right). \tag{2}$$

We express $S(x,\xi_1)$ in terms of $P(x,k)$ as follows:

$$S(x,\xi_1) = \sum_{l'} \mu(l')P(x,l') \tag{3}$$

where l' indicates either 1 or positive square-free numbers composed entirely of prime factors $q \le \xi_1$. $S(x,\xi_1)$ can be estimated in the same way as in [1]. The

details can be found in [1], so we shall omit them and just briefly recall the method used.

The primes counted in the sum $P(x,k)$ can be characterized in terms of conditions formulated in the language of algebraic number theory.

The primes contributing to $P(x,k)$ are just those primes p for which the simultaneous conditions

$$v^q \equiv 2 (\mathrm{mod}\ p) \text{ soluble}, \quad p \equiv 1 (\mathrm{mod}\ q)$$

hold for every prime divisor q of k, which are equivalent to the simultaneous conditions

$$v^k \equiv 2 (\mathrm{mod}\ p) \text{ soluble}, \quad p \equiv 1 (\mathrm{mod}\ k). \tag{4}$$

But (4) are together equivalent to the requirement that $v^k \equiv 2 (\mathrm{mod}\ p)$ have exactly k incongruent roots. By a principle due to Dedekind we deduce that (4) is equivalent to the condition that $p \nmid k$ and that p splits completely in the Kummerian field $\mathbf{Q}(\sqrt[k]{2}, \sqrt[k]{1})$.

Let $n(k)$ denote the degree of $\mathbf{Q}(\sqrt[k]{2}, \sqrt[k]{1})$ over $\mathbf{Q}$, and let $\pi(x,k)$ be the number of prime ideals $\wp$ in $\mathbf{Q}(\sqrt[k]{2}, \sqrt[k]{1})$ such that $N\wp \leq x$. In this case, $n(k) = k\phi(k)$ and we obtain

$$P(x,k) = \frac{\pi(x,k)}{k\phi(k)} + O\big(\nu(k)\big) + O(x^{1/2}). \tag{5}$$

In order to estimate $\pi(x,k)$ and hence $P(x,k)$, we apply the theory of Dedekind's zeta function and assume the following hypothesis:

The real part β of every complex zero $\rho = \beta + i\gamma$ of the Dedekind zeta function $\zeta_K(s)$ is less than or equal to $1-\theta$ for every Kummer field of type $K = \mathbf{Q}(\sqrt[k]{2}, \sqrt[k]{1})$ where $\theta > (1/2)e^{-12A/5}$.

We then obtain the following expression:

$$\pi(x,k) = \mathrm{li}\, x + O\big(k\phi(k)x^{1-\theta}\log kx\big). \tag{6}$$

From (5) and (6) we get

$$\begin{aligned} P(x,k) &= \frac{\mathrm{li}\, x}{k\phi(k)} + O(x^{1-\theta}\log kx) + O\big(\nu(k)\big) + O(x^{1/2}) \\ &= \frac{\mathrm{li}\, x}{k\phi(k)} + O(x^{1-\theta}\log kx). \end{aligned} \tag{7}$$

Thus, from (3) and (7) we get

$$\begin{aligned} S(x,\xi_1) &= \sum_{l'} \mu(l') \left\{ \frac{\mathrm{li}\, x}{l'\phi(l')} \right\} + O(x^{1-\theta}\log l'x) \\ &= \mathrm{li}\, x \sum_{l'} \frac{\mu(l')}{l'\phi(l')} + O\Big(x^{1-\theta}\log x \sum_{l'} 1\Big). \end{aligned}$$

Notice that

$$l' \leq \prod_{q \leq \xi_1} q \leq e^{2\xi_1} = x^{2\alpha}.$$

Choose $\alpha = \theta/3$. So,

$$\begin{aligned} S(x, \xi_1) &= \operatorname{li} x \sum_{l'} \frac{\mu(l')}{l'\phi(l')} + O\big(x^{1-\theta}(\log x)x^{\alpha}\big) \\ &= A \operatorname{li} x + O\left(\frac{x}{\xi_1 \log x}\right) + O\left(\frac{x}{\log^2 x}\right) \\ &= A \operatorname{li} x + O\left(\frac{x}{\log^2 x}\right) \end{aligned} \tag{8}$$

where

$$A = \prod_{q} \left(1 - \frac{1}{q(q-1)}\right).$$

The proof given so far is essentially Hooley's treatment of Artin's conjecture. However, with the weaker hypothesis we have assumed here, this method breaks down in estimating the second term $M(x, \xi_1, \xi_2)$. In order to estimate this term, we observe that

$$M\left(x, \frac{x^{\theta}}{\log^2 x}, x^{\frac{1}{2}-\epsilon}\right) \leq S_{2,3}(x) \tag{9}$$

where

$$\begin{aligned} S_{2,3}(x) = \#\Big\{ & p \leq x : p \text{ does not split completely in } L_2 \text{ and in } L_3, \\ & p \equiv 1 (\operatorname{mod} q), \frac{x^{\theta}}{\log^2 x} < q \leq x^{\frac{1}{2}-\epsilon} \Big\}. \end{aligned}$$

At this point, we appeal to Theorem 2 stated earlier. We shall apply this theorem taking $K = L_2 L_3$ where $L_2 = \mathbf{Q}(\sqrt{2})$ and $L_3 = \mathbf{Q}(\sqrt[3]{1}, \sqrt[3]{2})$.

We know that if K_1 and K_2 are finite field extensions of a field F in some algebraic closure $\overline{F}$, which are linearly disjoint over F, then $K_1 K_2 \simeq K_1 \otimes_F K_2$. If K_1 and K_2 are Galois extensions of F, it is easy to check that

$$\operatorname{Gal}(K_1 \otimes_F K_2 / F) \simeq \operatorname{Gal}(K_1/F) \times \operatorname{Gal}(K_2/F).$$

Therefore, with our choice of K as above, it follows that

$$\begin{aligned} \operatorname{Gal}(K/\mathbf{Q}) &\simeq \operatorname{Gal}(L_2/\mathbf{Q}) \times \operatorname{Gal}(L_3/\mathbf{Q}) \\ &= (\mathbf{Z}/2) \times S_3. \end{aligned}$$

Every $g \in \operatorname{Gal}(K/\mathbf{Q})$ can be written as $g = (g_1, g_2)$, and every conjugacy class C in $\operatorname{Gal}(K/\mathbf{Q})$ is of the form (C_1, C_2) where C_1 is a conjugacy class in $\operatorname{Gal}(L_2/\mathbf{Q})$ and C_2 is a conjugacy class in $\operatorname{Gal}(L_3/\mathbf{Q})$.

If p does not split completely in L_2 and in L_3, then $(p, K/\mathbf{Q}) = C$ where

(i) C is not of the form $\{\mathrm{Id}\} \times C_2$, and
(ii) C is not of the form $C_1 \times \{\mathrm{Id}\}$.

There are two conjugacy classes in $\mathrm{Gal}(K/\mathbf{Q})$ which satisfy (i) and (ii), *viz.*, $C = \{1\} \times \{\sigma, \tau\}$ and $C' = \{1\} \times \{\alpha, \beta, \gamma\}$ where 1 denotes the non-trivial element in $\mathbf{Z}/2$; σ, τ are elements of order 3 in S_3, and α, β, γ are elements of order 2 in S_3.

Then

$$\pi_C(x, q, 1) = \frac{2}{12} \frac{\mathrm{li}\, x}{\phi(q)} + E$$

and

$$\pi_{C'}(x, q, 1) = \frac{3}{12} \frac{\mathrm{li}\, x}{\phi(q)} + E'$$

where E and E' denote the error terms.

For these conjugacy classes C and C', we verify below that $d \leq 4$.

CASE I. Let $H = (\mathbf{Z}/2) \times \langle \sigma \rangle$. Then $(1, \sigma) \in H \cap C \neq \phi$, and H is an abelian subgroup of G whose order is 6.

CASE II. Let $H = (\mathbf{Z}/2) \times \langle \alpha \rangle$. Then $(1, \alpha) \in H \cap C' \neq \phi$, and H is an abelian subgroup of G whose order is 4.

Therefore, in both cases, $d \leq 4$ giving $\eta = 2$, and we now apply Theorem 3 to estimate $S_{2,3}(x)$.

$$S_{2,3}(x) \leq \left| \sum_{\frac{x^\theta}{\log^2 x} < q < x^{(1/2)-\epsilon}} \left(\frac{2}{12} \frac{\mathrm{li}\, x}{\phi(q)} + \frac{3}{12} \frac{\mathrm{li}\, x}{\phi(q)} + E + E' \right) \right|$$

$$\leq \frac{5}{12} \frac{x}{\log x} \log \frac{(1/2) - \epsilon}{\theta} + \frac{2x}{\log^2 x}. \tag{10}$$

The estimate of the error term follows from Theorem 2.

Combining (2), (8), (9) and (10) we get

$$N_2(x) \geq A \frac{x}{\log x} - \frac{\epsilon x}{\log x} - \frac{5}{12} \left(\log \frac{1 - 2\epsilon}{2\theta} \right) \frac{x}{\log x}$$

which implies that

$$N_2(x) \gg \frac{x}{\log x} \text{ whenever } A > \epsilon + \frac{5}{12} \log \frac{1 - 2\epsilon}{2\theta}$$

that is, whenever

$$\theta > \frac{1}{2} e^{-12A/5} + \delta, \quad \text{for } \delta > 0.$$

This completes the proof of the theorem.

ACKNOWLEDGEMENTS. This paper is the result of a suggestion of Professor M. Ram Murty; I thank him for his encouragement and support. I thank the Mehta Research Institute, Allahabad, for computing and library facilities during preparation of the paper.

After the paper was submitted for publication, research has subsequently been funded by the Council of Scientific and Industrial Research, India.

REFERENCES

1. C. Hooley, *Applications of sieve methods to the theory of numbers.* Cambridge Tracts in Math. **70**, Cambridge University Press, 1976.
2. M. Ram Murty and V. Kumar Murty, *A variant of the Bombieri-Vinogradov theorem.* CMS Conference Proceedings **7**(1987), 243–272.

Mehta Research Institute
Chhatnag Road
Jhusi, Allahabad 211 019
India
email: amora@mri.ernet.in

C. R. Math. Rep. Acad. Sci. Canada Vol. **23** (2), 2001 pp. 53–59

AVERAGING OF AN ELLIPTIC SPECTRAL PROBLEM IN A VARYING DOMAIN

MAMADOU SANGO

Presented by Vlastimil Dlab, FRSC

ABSTRACT. We consider the spectral problem for a higher-order elliptic equation in a sequence of perforated domains. Using a variational method, we establish the convergence of the eigensolutions of the problem to the eigensolutions of a limit problem containing an additional term of capacity type.

RÉSUMÉ. Nous considérons un problème spectral pour une équation élliptique d'ordre arbitraire dans une suite de domaines perforés. Utilisant une méthode variationnelle, nous établissont la convergence des valeurs propres et functions propres du problème vers les valeurs propres et vecteurs propres correspondants d'un problème limite qui contient un terme complementaire de type capacitaire.

1. Introduction. Let Ω be a bounded open set in the n-dimensional Euclidean space $\mathbf{R}^n$, with a sufficiently smooth boundary Γ, and let there be defined a finite number of closed sets $F_i^{(s)}$, $i = 1, \ldots, I(s)$ lying inside Ω and pairwise disjoint, *i.e.*, $F_i^{(s)} \cap F_j^{(s)} = \emptyset$ for $i \neq j$. In the domain $\Omega^{(s)} = \Omega \setminus \bigcup_{i=1}^{I(s)} F_i^{(s)}$, we consider the spectral boundary value problem

$$\sum_{|\alpha|,|\beta|\leq m} D^\alpha \big(a_{\alpha\beta}(x) D^\beta u(x)\big) = \lambda u(x) \text{ in } \Omega^{(s)}, \tag{1}$$

$$D^\alpha u(x) = 0, \quad |\alpha| \leq m-1 \text{ on } \partial\Omega^{(s)}. \tag{2}$$

We use the following notations: $\partial\bullet$ denotes the boundary of the set $\bullet$, $\bar{\bullet}$ denotes the closure of the set $\bullet$, $\alpha = (\alpha_1, \ldots, \alpha_n)$ is a multi-index with non negative integer components, $|\alpha| = \alpha_1 + \cdots + \alpha_n$, $D^\alpha u(x) = (\frac{\partial}{\partial x_1})^{\alpha_1} \cdots (\frac{\partial}{\partial x_n})^{\alpha_n} u(x)$, $D^k u(x) = \{D^\alpha u(x) : |\alpha| = k\}$. By $W_p^l(\bullet)$ ($p \in (1, \infty)$, l is a non negative integer), $\overset{0}{W}{}_p^l(\bullet)$ and $L_p(\bullet)$ we denote the usual Sobolev and Lebesgue spaces on $\bullet$.

Under appropriate geometric conditions on the closed sets $F_i^{(s)}$, we aim to prove that any eigensolution (λ_s, u_s) of problem (1)–(2) converges in suitable

Received by the editors April 13, 2000.
Research supported by the National Research Foundation of South Africa.
AMS subject classification: 35Jxx, 35Pxx.
© Royal Society of Canada 2001.

topologies to a corresponding eigensolution (λ, u) of the limit problem

$$\sum_{|\alpha|,|\beta|\le m} (-1)^{|\alpha|} D^{\alpha}\big(a_{\alpha\beta}(x) D^{\beta} u(x)\big) + c(x)u(x) = \lambda u(x) \text{ in } \Omega, \tag{3}$$

$$D^{\alpha} u(x) = 0, \quad |\alpha| \le m-1 \text{ on } \Gamma, \tag{4}$$

where c is a function expressed in terms of specific characteristic of the sets $F_i^{(s)}$.

The investigation of problem (1)–(2) in the stationary case (non spectral case) goes back to the works of Marchenko and Khruslov (see [3, Chap. 2] or [2]). Oleinik, Yocifian and Shamaev have studied in [4] the spectral problem for some classes of elliptic problems with rapidly oscillating coefficients in domains with periodic structure; earlier works in this direction were undertaken by Kesavan [1] and Vanninathan [5]. We refer to the bibliography of [4] for other works on this subject. The sequence of domains that we consider (the same as in [3]) need not have a periodic structure. For such domains, analogous investigations of the spectral problem seem not to have been done for higher-order elliptic problems. In the above mentioned papers the convergence results for the spectral problem were derived through reduction of the problem to operator formulation and subsequent application of abstract results of functional analysis combined with the convergence results obtained in the stationary case. Here we propose another approach based on analytic tools using some special trial test functions in the min-max formulation of the eigenvalues of problem (1)–(2).

2. **Hypotheses and results.** We assume that the functions $a_{\alpha\beta}(x)$ ($|\alpha|$, $|\beta| \le m$) in the equation (1) satisfy the following conditions:

(A1) $a_{\alpha\beta}(x)$ are real valued functions, defined and m-times continuously differentiable in $\bar{\Omega}$ and $a_{\alpha\beta}(x) = a_{\beta\alpha}(x)$.

(A2) For all $x \in \bar{\Omega}$, $\xi = (\xi_1, \dots, \xi_n) \in \mathbf{R}^n$, and any function $u(x) \in \overset{0}{W}{}_2^m(\Omega)$, the following inequalities hold:

$$\sum_{|\alpha|=|\beta|=m} a_{\alpha\beta}(x)\xi^{\alpha}\xi^{\beta} \ge \mu_1 \Big(\sum_{i=1}^{n} \xi_i^2\Big)^m, \tag{5}$$

$$\int_{\Omega} \sum_{|\alpha|,|\beta|\le m} a_{\alpha\beta}(x) D^{\alpha}u(x) D^{\beta}u(x) \ge \mu_2 \int_{\Omega} \sum_{|\alpha|\le m} |D^{\alpha}u(x)|^2 \, dx, \tag{6}$$

where μ_1 and μ_2 are some positive constants independent of x and u, and $\xi^{\alpha} = \xi_1^{\alpha_1} \cdots \xi_n^{\alpha_n}$. The inequality (5) is known as the condition of strong ellipticity.

Throughout the work, we restrict ourselves for simplicity to the case when $n > 2m$.

Let us proceed now to the formulation of the conditions on the sets $F_i^{(s)}$. Let $B(x, \rho)$ be a ball of radius ρ centered at the point x. We set $d_i^{(s)} = \min_{x \in \mathbf{R}^n} \{\rho :$

$F_i^{(s)} \subset B(x,\rho)\}$ and let $x_i^{(s)}$ be the center of the ball of radius $d_i^{(s)}$ such that $F_i^{(s)} \subset \overline{B(x_i^{(s)}, d_i^{(s)})}$. By $r_i^{(s)}$ we denote the distance from $B(x_i^{(s)}, d_i^{(s)})$ to $\bigcup_{i\neq j} B(x_j^{(s)}, d_j^{(s)}) \cup \Gamma$. Let $B(x_i^{(s)}, a)$ be the concentric ball to $B(x_i^{(s)}, d_i^{(s)})$ with $a > d_i^{(s)}$.

We introduce the auxiliary functions $v_i^{(s)}(x)$, solutions of the model boundary value problem

$$\sum_{|\alpha|=|\beta|=m} (-1)^m a_{\alpha\beta}(x_i^{(s)}) D^{\alpha+\beta} v_i^{(s)}(x) = 0 \quad \text{in } B(x_i^{(s)}, a) \setminus F_i^{(s)}, \tag{7}$$

$$D^\alpha \big(v_i^{(s)}(x) - 1\big) = 0, \quad |\alpha| \le m-1, \quad x \in \partial F_i^{(s)}, \tag{8}$$

$$D^\alpha v_i^{(s)}(x) = 0, \quad |\alpha| \le m-1, \quad x \in \partial B(x_i^{(s)}, a). \tag{9}$$

We extend the functions $v_i^{(s)}(x)$ to Ω by setting $v_i^{(s)}(x) = 1$ on $F_i^{(s)}$ and $v_i^{(s)}(x) = 0$ in $\Omega \setminus B(x_i^{(s)}, a)$. They will play a central role in the investigation of the problem (1)–(2).

For each set $F_i^{(s)}$, $s = 1, 2, \ldots$, $i = 1, \ldots, I(s)$, we define the number

$$C(F_i^{(s)}) = \int_{B(x_i^{(s)}, a)} \sum_{|\alpha|=|\beta|=m} a_{\alpha\beta}(x_i^{(s)}) D^\alpha v_i^{(s)}(x) D^\beta v_i^{(s)}(x)\, dx, \tag{10}$$

where the functions $v_i^{(s)}(x)$ are solutions of problem (7)–(9). The numbers $C(F_i^{(s)})$ represent the local energetic characteristics of the sets $F_i^{(s)}$.

We shall require the following conditions:

(H1) $d_i^{(s)} \le C_2 r_i^{(s)}$, $\lim_{s\to\infty} \max_{1\le i\le I(s)} \{r_i^{(s)}\} = 0$, where C_2 is a constant independent of i and s.

(H2)

$$\sum_{i=1}^{I(s)} \frac{(d_i^{(s)})^{2(n-2m)}}{(r_i^{(s)})^n} \le C_3.$$

(H3) There exists a bounded function $c(x)$ such that for any region $G \subset \Omega$,

$$\lim_{s\to\infty} \sum_{i\in I(s,G)} C(F_i^{(s)}) = \int_G c(x)\, dx,$$

where $I(s,G)$ denotes the set of numbers $i \in \{1, \ldots, I(s)\}$ for which $F_i^{(s)} \subset G$.

Throughout we understand a solution of the boundary value problem (1)–(2) or (3)–(4) in the weak sense.

The conditions (A1) and (A2) imply the selfadjointness of problem (1)–(2), the existence of a sequence of eigenvalues $0 < \lambda_1^{(s)} \le \lambda_2^{(s)} \le \cdots \le \lambda_k^{(s)} \le \cdots$ of (1)–(2) in $\mathbf{R}$, arranged in increasing order and a sequence $u_j^{(s)}(x)$, $j = 1, 2, \ldots$ of

eigenfunctions of (1)–(2) corresponding to the $\lambda_j^{(s)}$, $j = 1, 2, \ldots$ and forming an orthonormal basis in $L_2(\Omega^{(s)})$. The problem (3)–(4) is selfadjoint as well.

Let us introduce further notations:

$$(u,v)_s = \int_{\Omega^{(s)}} uv\,dx, (u,v)_0 = \int_{\Omega} uv\,dx,$$

$$\langle u,v\rangle_s = \int_{\Omega^{(s)}} \sum_{|\alpha|,|\beta|\le m} a_{\alpha\beta}(x) D^\alpha u D^\beta v\,dx,$$

$$\langle u,v\rangle_0 = \int_{\Omega} \Big(\sum_{|\alpha|,|\beta|\le m} a_{\alpha\beta}(x) D^\alpha u D^\beta v + c(x)uv \Big)\,dx.$$

We denote by $\mathcal{L}(u_1^{(s)}, \ldots, u_{k-1}^{(s)})$ (resp. $\mathcal{L}(u_1, \ldots, u_{k-1})$) the subspace generated in $L_2(\Omega^{(s)})$ (resp. $L_2(\Omega)$) by the functions $u_j^{(s)}$, $j = 1, \ldots, k-1$ (resp. u_j, $j = 1, \ldots, k-1$) introduced above.

Let

$$W_s(u_1^{(s)}, \ldots, u_{k-1}^{(s)}) = \{w \in \overset{0}{W}{}_2^m(\Omega^{(s)}) : \|w\|_{L_2(\Omega^{(s)})} = 1, (w, u_j^{(s)})_s = 0, j = 1, \ldots, k-1\},$$

$$W_0(u_1, \ldots, u_{k-1}) = \{w \in \overset{0}{W}{}_2^m(\Omega) : \|w\|_{L_2(\Omega)} = 1, (w, u_j)_0 = 0, j = 1, \ldots, k-1\},$$

and let us denote by δ_{ij} the symbol of Kronecker. By the min-max principle (see *e.g.* [4, Chap. 3]), it is known that the k-th eigenvalue $\lambda_k^{(s)}$ (resp. λ_k) of problem (1)–(2) (resp. (3)–(4)) is defined as follows,

$$\lambda_k^{(s)} = \inf_{w \in W_s(u_1^{(s)}, \ldots, u_{k-1}^{(s)})} \langle w, w\rangle_s \tag{11}$$

(resp.

$$\lambda_k = \inf_{w \in W_0(u_1, \ldots, u_{k-1})} \langle w, w\rangle_0). \tag{12}$$

The inf in (11) (resp. (12)) is attained if w is an eigenfunction of (1)–(2) (resp. (3)–(4)).

In the sequel by the k-th eigensolution $(\lambda_k^{(s)}, u_k^{(s)})$ (resp. (λ_k, u_k)) we shall mean that $\lambda_k^{(s)}$ (resp. λ_k) is the k-th eigenvalue of (1)–(2) (resp. (3)–(4)) and $u_k^{(s)}$ (resp. u_k) the eigenfunction corresponding to it. The symbol $A \to B$ will mean A converges to B.

The main result of this work is

THEOREM 1. *Assume that the hypotheses (A1), (A2), (H1), (H2) and (H3) are satisfied. Let* $0 < \lambda_1^{(s)} \le \lambda_2^{(s)} \le \cdots \le \lambda_k^{(s)} \le \cdots$ *be the sequence of eigenvalues of the problem (1)–(2) arranged in increasing order and let* $u_1^{(s)}, u_2^{(s)}, \ldots,$

$u_k^{(s)}, \ldots \in \overset{0}{W}{}_2^m(\Omega^{(s)})$ be the sequence of the corresponding eigenfunctions of (1)–(2) extended to Ω by setting $u_j^{(s)}(x) = 0$ in $\Omega \setminus \Omega^{(s)}$, and such that $(u_i^{(s)}, u_j^{(s)})_s = \delta_{ij}$, $i, j = 1, 2, \ldots$. Then there exists a sequence of real numbers $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k \leq \cdots$, arranged in increasing order, and a sequence of functions $u_1, \ldots, u_k, \ldots \in \overset{0}{W}{}_2^m(\Omega)$ such that

$$\lim_{s \to \infty} \lambda_k^{(s)} = \lambda_k,$$

and the sequence $u_k^{(s)}(x)$ converges weakly in $\overset{0}{W}{}_2^m(\Omega)$ and strongly in $\overset{0}{W}{}_p^m(\Omega)$ (for all $p \in (1, 2)$) to the function $u_k(x)$ as $s \to \infty$. Furthermore (λ_k, u_k) is the $k-th$ eigensolution of problem (3)–(4).

3. Sketch of the proof of Theorem 1.

STEP 1. Let $(\lambda_1^{(s)}, u_1^{(s)})$, $u_1^{(s)} \in \overset{0}{W}{}_2^m(\Omega^{(s)})$ be the first eigensolution of problem (1)–(2) such that $\|u_1^{(s)}\|_{L_2(\Omega^{(s)})} = 1$. The sequence $\{\lambda_1^{(s)}\}_{s=1,2,\ldots}$ is bounded; thus there exists a $\lambda_1^* > 0$ such that $\lambda_1^{(s)} \to \lambda_1^*$ modulo the extraction of a subsequence. Also the norm of $\{u_1^{(s)}\}_{s=1,2,\ldots}$ is bounded in $\overset{0}{W}{}_2^m(\Omega)$, hence modulo the extraction of a subsequence, $u_1^{(s)}$ converges weakly in $\overset{0}{W}{}_2^m(\Omega)$ to a function $u_1 \in \overset{0}{W}{}_2^m(\Omega)$ and $\|u_1\|_{L_2(\Omega)} = 1$. We prove that (λ_1^*, u_1) is an eigensolution of (3)–(4), furthermore $u_1^{(s)} \to u_1$ strongly in $\overset{0}{W}{}_p^m(\Omega)$ with $p \in (1, 2)$. A key point is to show that λ_1^* is indeed the first eigenvalue of (3)–(4), *i.e.*,

$$\lambda_1^* = \lambda_1 = \inf_{w \in \overset{0}{W}{}_2^m(\Omega), \|w\|_{L_2(\Omega)} = 1} \langle w, w \rangle_0.$$

We proceed as follows. Clearly $\lambda_1^* \geq \lambda_1$. Hence we must make sure that $\lambda_1^* \leq \lambda_1$. In order to prove the latest inequality we insert in the functional defining $\lambda_1^{(s)}$;

$$\lambda_1^{(s)} = \inf_{w \in \overset{0}{W}{}_2^m(\Omega), \|w\|_{L_2(\Omega^{(s)})}} \langle w, w \rangle_s$$

trial test functions of the form

$$\tilde{u}_{1s}^{(k)}(x) = \frac{u_{1s}^{(k)}(x)}{\|u_{1s}^{(k)}\|_{L_2(\Omega^{(s)})}},$$

where

$$u_{1s}^{(k)}(x) = u_1^{(k)}(x) - \sum_{i \in I_s} v_i^{(s)}(x) u_1^{(k)}(x) \psi_i^{(s)}(x);$$

$v_i^{(s)}(x)$ is a solution of problem (7)–(9) in $B(x_i^{(s)}, a) \setminus F_i^{(s)}$ with $v_i^{(s)}(x) = 1$ on $F_i^{(s)}$ and $v_i^{(s)}(x) = 0$ in $\Omega \setminus B(x_i^{(s)}, a)$, $\psi_i^{(s)}(x)$ are some appropriate test functions,

$\{u_1^{(k)}(x)\}_{k=1,2,\ldots} \in C_o^\infty(\Omega)$ and $u_1^{(k)} \to u_1$ strongly in $\overset{0}{W}{}_2^m(\Omega)$. As a result of appropriate calculations and using some sharp pointwise and integral estimates of the functions $v_i^{(s)}$, we get that

$$\lambda_1^* \leftarrow \lambda_1^{(s)} \leq \langle \tilde{u}_{1s}^{(k)}, \tilde{u}_{1s}^{(k)} \rangle_s \to \langle u_1, u_1 \rangle_0 = \lambda_1,$$

as $s, k \to \infty$. Thus $\lambda_1^* \leq \lambda_1$ and the equality $\lambda_1^* = \lambda_1$ follows. Therefore Theorem 1 holds for the first eigensolution of (1)–(2).

STEP 2. Let $(\lambda_1^{(s)}, u_1^{(s)}), \ldots, (\lambda_k^{(s)}, u_k^{(s)})$ $(u_j^{(s)} \in \overset{0}{W}{}_2^m(\Omega^{(s)}))$ be the first k eigensolutions of (1)–(2) such that $(u_i^{(s)}, u_j^{(s)})_s = \delta_{ij}$, $i, j = 1, \ldots, k$. Let $(\lambda_1, u_1), \ldots, (u_{k-1}, \lambda_{k-1})$ be the first $(k-1)$ eigensolutions of (3)–(4) such that $\lambda_j^{(s)} \to \lambda_j$ and $u_j^{(s)}$ converges to u_j weakly in $\overset{0}{W}{}_2^m(\Omega)$ and strongly in $\overset{0}{W}{}_p^m(\Omega)$ (for all $p \in (1,2)$), for $j = 1, \ldots, k-1$. We show that there exists a $\lambda_k > 0$ and a function $u_k \in \overset{0}{W}{}_2^m(\Omega)$ such that $\lambda_k^{(s)} \to \lambda_k$, $u_k^{(s)}$ converges to u_k weakly in $\overset{0}{W}{}_2^m(\Omega)$ and strongly in $\overset{0}{W}{}_p^m(\Omega)$ $(p \in (1,2))$; furthermore (λ_k, u_k) is the k-th eigensolution of (3)–(4).

The proof is in spirit the same as in Step 1, but it is technically more involved. The role of the trial test functions in (11) are played by the functions

$$\tilde{u}_{sk}^{(r)}(x) = \frac{W_{sk}^{(r)}(x)}{\|W_{sk}^{(r)}\|_{L_2(\Omega^{(s)})}},$$

where

$$W_{sk}^{(r)}(x) = u_{sk}^{(r)}(x) - \sum_{j=0}^{k-1} (u_{sk}^{(r)}, u_j^{(s)})_s u_j^{(s)}(x),$$

$$u_{sk}^{(r)}(x) = u_k^{(r)}(x) - \sum_{i \in I_s} v_i^{(s)}(x) u_k^{(r)}(x) \psi_i^{(s)}(x);$$

the functions $v_i^{(s)}(x)$, $\psi_i^{(s)}(x)$ are the same as in Step 1, and $\{u_k^{(r)}(x)\}_{r=1,2,\ldots}$ is a sequence of functions in $C_o^\infty(\Omega)$ which converges to u_k strongly in $\overset{0}{W}{}_2^m(\Omega)$. Thus Theorem 1 holds for the k-th eigensolution of (1)–(2). The validity of Theorem 1 for all eigenvalues and their corresponding eigenfunctions then follows by induction.

REFERENCES

1. S. Kesavan, *Homogenization of elliptic eigenvalue problems.* Appl. Math. Optim. **5**(1979), 153–167, 197–216.
2. E. Ya. Khruslov, *The first boundary value problems in domains with a complicated boundary for higher-order equations.* Math. USSR Sb. **32**(1977), 535–549.

3. V. A. Marchenko and E. Ya. Khruslov, *Boundary value problems in domains with fine-grained boundaries.* Naukova Dumka, Kiev, 1974 (Russian).
4. O. A. Oleinik, A. S. Shamaev and G. A. Yocifian, *Mathematical problems in the theory of strongly nonhomogeneous elastic media.* Moscow University Press, Moscow, 1990 (Russian).
5. M. Vanninathan, *Homogénéisation des valeurs propres dans les milieux perforés.* C. R. Acad. Sci. Paris, **287**(1978), 403–406.

Department of Mathematics
Vista University
Private Bag X1311
Silverton 0127
Pretoria
South Africa
email: sango-m@marlin.vista.ac.za

C. R. Math. Rep. Acad. Sci. Canada Vol. **23** (2), 2001 pp. 60–64

PROOF OF SOME CONJECTURES BY KAPLANSKY

R. A. MOLLIN

Presented by Vlastimil Dlab, FRSC

ABSTRACT. In correspondence with this author, Professor Irving Kaplansky posed several conjectures, largely prompted by his being inspired by the proof of [2, Theorem 6.5.9, p. 348]. Although some of these conjectures may be "known" in the folklore, they certainly are not *well* known. Moreover, the proofs discovered by this author of three of these conjectures link the solutions of quadratic Diophantine equations with the theory of continued fractions, thereby continuing work done by this author and others in [4]–[6].

RÉSUMÉ. Lors de quelques échanges de correspondence entre l'auteur et le professeur Irving Kaplansky, ce dernier a énoncé plusieurs conjectures, sous l'influence, pour la plupart d'entre elles, de la preuve du [2, Théorème 6.5.9, p. 348]. Même si certaines de ces conjectures font partie du folklore, elles ne sont probablement pas toutes *bien* connues. De plus, les preuves obtenues par l'auteur de trois de ces conjectures établissent un lien entre les solutions d'équations quadratiques diophantiennes et les fractions continues, et sont dans la ligne des travaux de l'auteur et autres chercheurs [4]–[6].

1. **Notation and preliminaries.** We assume that the reader is familiar with basic algebraic number theoretic concepts such as those contained in [2]–[3]. We denote simple continued fraction expansions by

$$\langle q_0 \, ; \, q_1, q_2, \ldots, q_l, \ldots \rangle.$$

These partial quotients are linked to the following recursive sequences, which we will need in the next section. For $D \in \mathbb{N}$ not a perfect square, and $(P + \sqrt{D})/Q$ a quadratic irrational, define

$$P_0 = P, \quad Q_0 = Q, \quad \text{and recursively for } j \geq 0,$$

$$q_j = \left\lfloor \frac{P_j + \sqrt{D}}{Q_j} \right\rfloor, \tag{1.1}$$

$$P_{j+1} = q_j Q_j - P_j, \tag{1.2}$$

and

$$D = P_{j+1}^2 + Q_j Q_{j+1}. \tag{1.3}$$

Received by the editors October 27, 2000.
AMS subject classification: 11A55, 11R11, 11D09, 11A51.
Key words and phrases: continued fractions, diophantine equations, ideals.
© Royal Society of Canada 2001.

It follows that we have the simple continued fraction expansion of the quadratic irrational:

$$\alpha = \frac{P + \sqrt{D}}{Q} = \frac{P_0 + \sqrt{D}}{Q_0} = \langle q_0 \,;\, q_1, \ldots, q_j, \ldots \rangle,$$

and for such an α, we define two sequences of integers $\{A_j\}$ and $\{B_j\}$ inductively by:

$$A_{-2} = 0, \quad A_{-1} = 1, \quad A_j = q_j A_{j-1} + A_{j-2} \quad (\text{for } j \geq 0), \tag{1.4}$$

$$B_{-2} = 1, \quad B_{-1} = 0, \quad B_j = q_j B_{j-1} + B_{j-2} \quad (\text{for } j \geq 0). \tag{1.5}$$

If $\alpha = \sqrt{D}$, then by [2, Theorem 5.3.4, p. 246],

$$A_{j-1}^2 - B_{j-1}^2 D = (-1)^j Q_j \quad (\text{for } j \geq 1), \tag{1.6}$$

Finally, we need the following result.

THEOREM 1.1. *Suppose that $D > 0$ is a squarefree radicand, and $\ell(\sqrt{D}) = \ell$ is the period length of the simple continued fraction expansion of $\sqrt{D}$, with the Q_j defined in that expansion from equations (1.1)–(1.3). Then $Q_j | 2D$ with $Q_j > 1$ if and only if $j = \ell/2$. Furthermore, if D is even, then $Q_j | D$ with $Q_j > 1$ if and only if $j = \ell/2$. In either case, $q_{\ell/2} = 2P_{\ell/2}/Q_{\ell/2}$.*

PROOF. See [1, Theorem 6.1.4, p. 193]. ∎

2. Three conjectures. In correspondence with this author over the past couple of years, the following were posed by Irving Kaplansky.

CONJECTURE 2.1. *Let $p \equiv 3 \pmod 4$ be a prime, and let*

$$\langle q_0 \,;\, q_1, \ldots, q_{\ell/2}, \ldots, q_\ell \rangle$$

be the simple continued fraction expansion of $\sqrt{p}$ of period length ℓ. Then either $q_{\ell/2} = \lfloor \sqrt{p} \rfloor = q_0$ or $q_{\ell/2} = \lfloor \sqrt{p} \rfloor - 1 = q_0 - 1$, whichever is odd.

CONJECTURE 2.2. *Let $p \equiv 1 \pmod 8$ and $q \equiv 3 \pmod 4$ be primes, $T + U\sqrt{pq}$ the minimal solution of $x^2 - pqy^2 = 1$, and $\left(\frac{p}{q}\right) = 1$, where $\left(\frac{*}{*}\right)$ is the Legendre symbol. Then if*

$$q^{(p-1)/4} \equiv -1 \pmod p, \tag{2.7}$$

T is even.[1]

CONJECTURE 2.3. *If p is a prime such that $p = a^2 + b^2$ for some integers a, b, then there exist integers x, y such that $a = x^2 - py^2$.*

[1] Note that (2.7) is equivalent to $\left(\frac{q}{p}\right)_4 = -1$ where $\left(\frac{*}{*}\right)_4$ denotes the quartic residue symbol.

PROOF OF CONJECTURE 2.1. First, since $p \equiv 3 \pmod 4$, then ℓ is even (by equation (1.6) with $j = \ell$). By Theorem 1.1, $q_{\ell/2} = 2P_{\ell/2}/Q_{\ell/2}$, and $Q_{\ell/2} \mid 2p$. Since,

$$p = P_{\ell/2}^2 + Q_{\ell/2}Q_{\ell/2-1},$$

by equation (1.3), then $Q_{\ell/2} \neq p, 2p$, given that $Q_j \geq 1$ for all integers $j \geq 0$. Since $Q_{\ell/2} > 1$, then $Q_{\ell/2} = 2$, so

$$q_{\ell/2} = P_{\ell/2}.$$

However, by equation (1.1),

$$q_{\ell/2} = \left\lfloor \frac{P_{\ell/2} + \sqrt{p}}{2} \right\rfloor.$$

Thus,

$$\frac{P_{\ell/2} + \sqrt{p}}{2} > P_{\ell/2} > \frac{P_{\ell/2} + \sqrt{p}}{2} - 1,$$

which forces

$$\sqrt{p} > P_{\ell/2} > \sqrt{p} - 2.$$

Since the only integers in that range are $\lfloor \sqrt{p} \rfloor$ and $\lfloor \sqrt{p} \rfloor - 1$, then $P_{\ell/2}$ must be one of them. Also, by equation (1.3) again,

$$p = P_{\ell/2}^2 + 2Q_{\ell/2},$$

so $P_{\ell/2}$ cannot be even. Hence, $q_{\ell/2}$ is one of $\lfloor \sqrt{p} \rfloor$ or $\lfloor \sqrt{p} \rfloor - 1$, whichever is odd. This completes the proof of Conjecture 2.1.

PROOF OF CONJECTURE 2.2. Suppose that (2.7) holds and T is odd.

CLAIM 2.1. *There exist $a, b \in \mathbb{Z}$ such that $pb^2 - qa^2 = 1$.*

Since $T^2 - pqU^2 = 1$, then

$$(T+1)(T-1) = T^2 - 1 = pqU^2.$$

However, since $g = \gcd(T+1, T-1) \mid 2$, then $g = 2$ given that T is odd. There are four possibilities.

CASE 2.1. *$T - 1 = 2a^2$ and $T + 1 = 2pqb^2$ where $2ab = U$.*

By subtracting, we get that $1 = pqb^2 - a^2$, so $a^2 \equiv -1 \pmod q$, a contradiction since $q \equiv 3 \pmod 4$.

CASE 2.2. *$T - 1 = 2pa^2$ and $T + 1 = 2qb^2$ where $2ab = U$.*

By subtracting, we get that $1 = qb^2 - pa^2$. Thus, $(pa)^2 \equiv -p \pmod q$, so

$$1 = \left(\frac{pa}{q}\right)^2 = \left(\frac{(pa)^2}{q}\right) = \left(\frac{-p}{q}\right) = \left(\frac{-1}{q}\right)\left(\frac{p}{q}\right) = -\left(\frac{p}{q}\right),$$

so $\left(\frac{p}{q}\right) = -1$, a contradiction.

CASE 2.3. $T-1=2pqa^2$ *and* $T+1=2b^2$ *where* $2ab=U$.

By subtracting, we get $1=b^2-pqa^2$, but $b<T$, contradicting the minimality of T.

CASE 2.4. $T-1=2qa^2$ *and* $T+1=2pb^2$ *where* $2ab=U$.

By subtracting, we get $1=pb^2-qa^2$, which is Claim 2.1.

By Claim 2.1, $-qa^2\equiv 1 \pmod{p}$. Therefore,

$$1=\left(\frac{-qa^2}{p}\right)_4=\left(\frac{-q}{p}\right)_4\left(\frac{a}{p}\right)_4^2=\left(\frac{-q}{p}\right)_4\left(\frac{a}{p}\right)=\left(\frac{q}{p}\right)_4\left(\frac{a}{p}\right)=-\left(\frac{a}{p}\right).$$

We have shown that

$$\left(\frac{a}{p}\right)=-1. \tag{2.8}$$

Let $a=2^jy$ where y is odd and j is a nonnegative integer. Since $pb^2-qa^2=1$, then $pb^2\equiv 1 \pmod{y}$. Therefore, by the Quadratic Reciprocity Law, we have the following Jacobi symbol equation:

$$1=\left(\frac{pb^2}{y}\right)=\left(\frac{p}{y}\right)=\left(\frac{y}{p}\right)=\left(\frac{2^jy}{p}\right)=\left(\frac{a}{p}\right),$$

where the penultimate equality comes from the fact that $p\equiv 1 \pmod{8}$. We have contradicted (2.8). This proves Conjecture 2.2.

For the following proof, the reader should be familiar with the basics of quadratic orders as found in [1].

PROOF OF CONJECTURE 2.3. Let $p=a^2+b^2$, and consider the non-maximal quadratic order $\mathcal{O}_{4p}=\mathbb{Z}[\sqrt{p}]$ of discriminant $4p$. Then $I=(a,b+\sqrt{p})$ is an ideal in $\mathcal{O}_{4p}$ of norm a. By multiplication formulae for ideal given in [2, (3.5.2)–(3.5.7), pp. 178–179] (see also [1, pp. 10–11]), $I^2=(a^2,b+\sqrt{p})$. By [3, Exercise 3.73, p. 158], $I^2\sim 1$ in the class group of $\mathcal{O}_{4p}$. By [3, Theorem 3.70, p. 162] and the class number formula for quadratic orders given [2, p. 345] (see also the development given in [1, pp. 25–26]), the class number of $\mathcal{O}_{4p}$ is odd. Hence, by [2, Exercise 6.5.33, p. 357], $I\sim 1$ in the class group of $\mathcal{O}_{4p}$. Thus, there are $x,y\in\mathbb{Z}$ such that $I=(x+y\sqrt{p})$, where the norm of I is given via [3, Corollary 3.44, p. 150]:

$$N(I)=a=N(x+y\sqrt{p})=x^2-py^2,$$

as required.

ACKNOWLEDGEMENTS. This author is indebted to not only lively and inspired correspondence with Irving Kaplansky, but also for input from Franz Lemmermeyer in the proof of Conjecture 2.2 and from Andrew Granville in the solution of Conjecture 2.3. Also, thanks go to Claude Levesque for translating the abstract into French.

NOTE. Upon acceptance of this paper for publication, the editors indicated that the (anonymous) referee informed them that a (different) proof of Conjecture 2.3 has been submitted by Feit to the American Mathematical Monthly.

REFERENCES

1. R. A. Mollin, *Quadratics*. CRC Press, Boca Raton-New York-London-Tokyo, 1996.
2. ———, *Fundamental Number Theory with Applications*. CRC Press, Boca Raton-New York-London-Tokyo, 1998.
3. ———, *Algebraic Number Theory*. Chapman and Hall/CRC Press, Boca Raton-New York-London-Tokyo, 1999.
4. ———, *Jacobi symbols, ambiguous ideals, and continued fractions*. Acta Arith. **85**(1998), 331–349.
5. R. A. Mollin and A. J. van der Poorten, *Continued fractions, Jacobi symbols, and quadratic Diophantine equations*. Canad. Math. Bull. **43**(2000), 218–225.
6. A. J. van der Poorten and P. G. Walsh, *A note on Jacobi symbols and continued fractions*. Amer. Math. Monthly **106**(1999), 52–56.

Department of Mathematics and Statistics
University of Calgary
Calgary, Alberta
T2N 1N4
website: http://www.math.ucalgary.ca/~ramollin/
email: ramollin@math.ucalgary.ca