REVISTA DE ANALIZĂ NUMERICĂ ȘI TEORIA APROXIMAȚIEI
Volumul 2, Fascicola 1, 1973, pp. 51–59

UN ALGORITM PENTRU DETERMINAREA ELEMENTULUI DE CEA MAI BUNA APROXIMARE CU RESTRICȚII

de

GH. CIMOCA

(Cluj)

1. Introducere

În această notă se prezintă un algoritm de tip REMEZ [1] pentru determinarea elementului de cea mai bună aproximare într-o problemă de aproximare cu restricții, al cărei studiu a fost publicat în [2]. G. D. TAYLOR și M. J. WINTER [6] au dat un astfel de algoritm pentru o problemă de aproximare cu restricții, care este un caz particular al problemei studiate în [2]. Mai amintim că P. J. LAURENT [4], R. C. JONES și L. A. KARLOVITZ [3] au dat algoritmi de același tip pentru probleme de apxroximare unilaterală.

2. Notații, definiții, rezultate

Fie T, U, L submulțimi compacte, nu neapărat disjuncte, ale intervalului $[a, b]$ Presupunem că $S = T \cup U \cup L$ conține cel puțin $n + 1$ puncte. Se notează prin $C[S]$ spațiul liniar al funcțiilor continue cu valori reale definite pe S.

Fie $M = [\varphi_1, \ldots, \varphi_n]$ un subspațiu Haar n-dimensional al spațiului $C[S]$ pentru care funcțiile $\varphi_1, \ldots, \varphi_n$ formează o bază ($\varphi_1, \ldots, \varphi_n$ este un sistem Cebîsev)

În continuare se fixează un cuplu (u, l) de funcții cu valori reale extinse :

$$u : U \to \bar{R} \text{ și } l : L \to \bar{R},$$

supuse la următoarele restricţii:

(i) $\{s \in U : u(s) = -\infty\} = \emptyset$, dar
$\{s \in U : u(s) = +\infty\} = U_{+\infty}$ poate fi nevidă.

(ii) $\{s \in L : l(s) = +\infty\} = \emptyset$, dar
$\{s \in L : l(s) = -\infty\} = L_{-\infty}$ poate fi nevidă.

(iii) $U_{+\infty}$ este o submulţime deschisă a lui U, iar
$L_{-\infty}$ este o submulţime deschisă a lui L.

(iv) u este continuă pe $U^* = U - U_{+\infty}$, iar
l este continuă pe $L^* = L - L_{-\infty}$.

(v) Dacă $s \in U \cap L$ atunci $u(s) > l(s)$.

Mulţimea funcţiilor cu care se face aproximarea este:

$$\overline{M} = \{p \in M : l(s) \leq p(s),\ s \in L \ \text{ şi } \ p(s) \leq u(s),\ s \in U\}$$

presupunînd în continuare că $\overline{M} \neq \emptyset$.

Fie:

$$\|f\| = \max_{s \in T} |f(s)| \text{ pentru } f \in C[S].$$

D e f i n i ţ i e. *O funcţie $p_0 \in \overline{M}$ se numeşte element de cea mai bună aproximare din $\overline{M}$ pentru funcţia $f \in C[S]$ dacă:*

$$e_0 = \|f - p_0\| = \inf_{p \in \overline{M}} \|f - p\|.$$

O b s e r v a ţ i e. *Dacă $T = U = L$ se obţine problema de aproximare cu restricţii studiată în* [5], *problemă pentru care se cunoaşte un algoritm de calcul al elementului de cea mai bună aproximare* [6].

În vederea caracterizării elementului de cea mai bună aproximare pentru o funcţie dată $f \in C[S]$ se definesc, pentru fiecare element $p_0 \in \overline{M}$, următoarele mulţimi de puncte critice:

$$\gamma_0^+ = \{s \in T : f(s) - p_0(s) = \|f - p_0\|\},$$

$$\gamma_0^- = \{s \in T : f(s) - p_0(s) = -\|f - p_0\|\},$$

$$\gamma_0^u = \{s \in U : p_0(s) = u(s)\},$$

$$\gamma_0^l = \{s \in L : p_0(s) - l(s)\}$$

$$\Gamma_0^+ = \gamma_0^+ \cup \gamma_0^l,\ \ \Gamma_0^- = \gamma^- \cup \gamma_0^u,\ \ \Gamma_0 = \Gamma_0^+ \cup \Gamma_0^-.$$

Are loc următoarea teoremă de caracterizare:

TEOREMA 1. [2, *Teorema* 4.6] *Fie $f \in C[S]$ astfel încît: $l(s) \leq f(s) \leq u(s)$ pentru $s \in U \cap L$. Elementul $p_0 \in \overline{M}$ este un element de cea mai bună aproximaţre pentru f dacă şi numai dacă există $n + 1$ puncte $s_1 < s_2 < \ldots < s_{n+1}$ în Γ_0 astfel încît:*

$$\xi(s_i) = (-1)^{i+1}\xi(s_1),\ i = 2, 3, \ldots, n+1$$

unde:

$$\xi(s) = \begin{cases} +1 & \text{dacă } s \in \Gamma_0^+, \\ -1 & \text{dacă } s \in \Gamma_0^- \end{cases}$$

Înainte de a trece la descrierea algoritmului vom defini pentru fiecare $\in \overline{M}$ şi $f \in C[S]$ următoarea funcţie:

$$\delta(f(s) - p(s)) = \begin{cases} +1 \text{ dacă } p(s) = f(s) = l(s), & s \in L \\ -1 \text{ dacă } p(s) = f(s) = u(s), & s \in U \\ \operatorname{sgn}\{f(s) - p(s)\} \text{ în celelalte cazuri} \end{cases}$$

3. Algoritmul

Presupunem dat cuplul (u, l) care verifică restricţiile (i) — (v). Fie $f \in C[S] - \overline{M}$ fixată.

Se consideră pentru început sistemul arbitrar de puncte din S:

$$S_1 = \{s_{11}, \ldots, s_{n+1,1}\} \text{ astfel încît } s_{11} < \ldots < s_{n+1,1}$$

şi sistemul de ecuaţii:

$$\sum_{i=1}^{n} \alpha_i \varphi_i(s_{j1}) + (-1)^j \alpha_{n+1} = f(s_{j1}),\ j = 1, 2, \ldots, n+1. \tag{1}$$

Deoarece $\varphi_1, \ldots, \varphi_n$ este un sistem Cebîşev, sistemul (1) are o soluţie unică pe care o vom nota cu:

$$A_1 = \{\alpha_{11}, \alpha_{21}, \ldots, \alpha_{n+1,1}\}.$$

Fie $p_1 = \sum_{i=1}^{n} \alpha_{i1}\varphi_i$ şi $e_1 = |\alpha_{n+1,1}|$.

Dacă $\|f - p_1\| = e_1$, $p_1(s) \leq u(s)$, $s \in U$ şi $p_1(s) \geq l(s)$, $s \in L$ atunci p_1 este elementul de cea mai bună aproximare pentru f şi iteraţia se opreşte.

Dacă p_1 nu este un element de cea mai bună aproximare pentru f se trece la al doilea pas al iterației procedîndu-se în modul următor:

Fie

$$E_1 = \|f - p_1\| - e_1,$$

$$M_1 = \max \{p_1(s) - u(s),\ s \in U\},$$

$$m_1 = \max \{l(s) - p_1(s),\ s \in L\},$$

$$\eta_1 = \max \{E_1,\ M_1,\ m_1\}.$$

O b s e r v a ț i e. *În cazul egalității unor numere din tripletul de mai sus, se alege η_1 ca fiind primul element maxim din triplet.*

Pasul al doilea al iterației se începe prin alegerea unui punct din S relativ la următoarele cazuri:

— dacă $\eta_1 = E_1$ se alege $\sigma \in T$ încît $\|f - p_1\| = |f(\sigma) - p_1(\sigma)|$
— dacă $\eta_1 = M_1$ se alege $\sigma \in U$ încît $\eta_1 = p_1(\sigma) - u(\sigma)$
— dacă $\eta_1 = m_1$ se alege $\sigma \in L$ încît $\eta_1 = l(\sigma) - p_1(\sigma)$.

Se înlocuiește sistemul S_1 prin $S_2 = \{s_{12}, \ldots, s_{n+1,2}\}$ care are proprietatea că $s_{i2} = s_{i1}$ pentru toți indicii i cu excepția unui singur indice i_0 pentru care $s_{i_0,2} = \sigma$ și îndeplinește condițiile:

$$\delta(f(s_{i2}) - p_1(s_{i2})) = (-1)^{i+1}\delta(f(s_{12}) - p_1(s_{12})),\quad i = 1, \ldots, n+1.$$

Se împarte apoi mulțimea de indici $\{1, 2, \ldots, n+1\}$ în trei submulțimi două cîte două disjuncte, în felul următor:

— dacă i nu este indicele corespunzător punctului extras $s_{i_0,1}$ atunci $i \in T_2$
— dacă i coincide cu indicele punctului extras atunci:

$$i_0 \in T_2 \quad \text{dacă} \quad \eta_1 = E_1,$$
$$i_0 \in U_2 \quad \text{dacă} \quad \eta_1 = M_1,$$
$$i_0 \in L_2 \quad \text{dacă} \quad \eta_1 = m_1.$$

Continuarea pasului secund al interației constă în rezolvarea sistemului:

$$\sum_{j=1}^{n} \alpha_{j2}\varphi_j(s_{12}) + (-1)^i\alpha_{n+1,2} = f(s_{i2}),\quad i \in \mathcal{T}_2,$$

$$(2)\qquad \sum_{j=1}^{n} \alpha_{j2}\varphi_j(s_{i2}) = u(s_{i2}),\quad i \in \mathcal{U}_2,$$

$$\sum_{j=1}^{n} \alpha_{j2}\varphi_j(s_{i2}) = l(_{i2}s),\quad i \in \mathcal{L}_2,$$

cu necunoscutele $\alpha_{12}, \ldots, \alpha_{n+1,2}$, unde ultimele ecuații se iau în considerare numai dacă $U_2 \neq \emptyset$ sau $L_2 \neq \emptyset$

Notînd cu $A_2 = \{\alpha_{12}, \ldots, \alpha_{n+1,2}\}$ soluția sistemului (2) fie:

$$p_2 = \sum_{i=1}^{n} \alpha_{i2}\varphi_i \quad \text{și} \quad e_2 = |\alpha_{n+1,2}|.$$

Dacă $\|f - p_2\| = e_2$, $p_2(s) \leq u(s)$, $s \in U$ și $p_2(s) \geq l(s)$, $s \in L$ atunci p_2 este elementul de cea mai bună aproximare.

Dacă p_2 nu este elementul de cea mai bună aproximație, iterația se continuă prin inducție astfel:

La pasul k se cunosc:

1° Un sistem de puncte din S: $S_k = \{s_{1k}, \ldots, s_{n+1,k}\}$ astfel încît: $s_{1k} < s_{2k} < \ldots < s_{n+1,k}$.

2° Trei mulțimi $\mathcal{T}_k$, $\mathcal{U}_k$, $\mathcal{L}_k$ două cîte două disjuncte a căror reuniune este $\{1, 2, \ldots, n+1\}$.

3° Un element $p_k = \sum_{j=1}^{n} \alpha_{jk}\varphi_j$

4. Un număr real $\alpha_{n+1,k}$ astfel încît este satisfăcut sistemul:

$$\sum_{j=1}^{n} \alpha_{jk}\varphi_j(s_{ik}) + (-1)^i\alpha_{n+1,k} = f(s_{ik}),\quad i \in \mathcal{T}_k$$

$$\sum_{j=1}^{n} \alpha_{jk}\varphi_j(s_{ik}) = u(s_{ik}),\quad i \in \mathcal{U}_k$$

$$\sum_{j=1}^{n} \alpha_{jk}\varphi_j(s_{ik}) = l(s_{ik}),\quad i \in \mathcal{L}_k$$

și în plus au loc relațiile:

$$\delta(f(s_{ik}) - p_k(s_{ik})) = (-1)^{i+k}\delta(f(s_{1k}) - p_k(s_{1k})),\quad i = 1, \ldots, n+1.$$

Dacă $\|f - p_k\| = e_k = |\alpha_{n+1,k}|$ și $p_k(s) \leq u(s)$, $s \in U$ și $p_k(s) \geq l(s)$, $s \in L$ atunci p este elementul de cea mai bună aproximare și iterația se oprește.

Dacă p_k nu este elementul de cea mai bună aproximare se trece la pasul $k+1$ în modul următor:

Fie

$$E_k = \|f - p_k\| - e_k,$$

$$M_k = \max \{p_k(s) - u(s),\ s \in U\},$$

$$m_k = \max \{l(s) - p_k(s),\ s \in L\}$$

și $\eta_k = \max \{E_k, M_k, m_k\}$ ținîndu-se cont de observația făcută la al doilea pas.

Urmează acum alegerea unui punct din S conform criteriilor:

— dacă $\eta_k = E_k$ se alege $\sigma \in T$ încît $\|f - p_k\| = (f(\sigma) - p_k(\sigma)|$.

— dacă $\eta_k = M_k$ se alege $\sigma \in U$ încît $\eta_k = p_k(\sigma) - u(\sigma)$.

— dacă $\eta_k = m_k$ se alege $\sigma \in L$ încît $\eta_k = l(\sigma) - p_k(\sigma)$.

Se înlocuiește sistemul S_k prin sistemul $S_{k+1} = \{s_{1,k+1}, \ldots, s_{n+1,k+1}\}$ unde $s_{i,k+1} = s_{ik}$ pentru toți indicii i cu excepția unui indice i_0 pentru care $s_{i_0,k+1} = \sigma$ astfel încît să fie satisfăcute relațiile:

$$\delta(f(s_{i,k+1}) - p_k(s_{i,k+1})) = (-1)^{i+1}\delta(f(s_{1,k+1}) - p_k(s_{1,k+1}))$$

pentru $i = 1, 2, \ldots, n + 1$.

În continuare se definesc mulțimile $\mathcal{T}_{k+1}$, $\mathcal{U}_{k+1}$, $\mathcal{L}_{k+1}$ două cîte două disjuncte a căror reuniune este $\{1, \ldots, n + 1\}$ definire ce se face în raport cu noul punct $s_{i_0,k+1}$ conform cu faptul că $\eta_k = E_k$, M_k respectiv m_k.

Se consideră acum sistemul:

$$\sum_{j=1}^{n} \alpha_{j,k+1}\varphi_j(s_{i,k+1}) + (-1)^i\alpha_{n+1,k+1} = f(s_{i,k+1}), \quad i \in \mathcal{T}_{k+1}$$

$$\sum_{j=1}^{n} \alpha_{j,k+1}\varphi_j(s_{i,k+1}) = u(s_{i,k+1}), \quad i \in \mathcal{U}_{k+1}$$

$$\sum_{j=1}^{n} \alpha_{j,k+1}\varphi_j(s_{i,k+1}) = l(s_{i,k+1}), \quad i \in \mathcal{L}_{k+1}$$

cu soluția $A_{k+1} = \{\alpha_{1,k+1}, \ldots, \alpha_{n+1,k+1}\}$ și fie:

$$p_{k+1} = \sum_{i=1}^{n} \alpha_{i,k+1}\varphi_i \quad \text{și} \quad e_{k+1} = |\alpha_{n+1,k+1}|.$$

Vom demonstra următoarea teoremă:

TEOREMA 2. *Presupunem că iterația nu se sfîrșește după un număr finit de pași. Atunci șirul de elemente $(p_k)_{k=1}^{\infty}$ construit prin algoritmul descris, converge uniform către elementul de ceea mai bună aproximare $p_0 \in \overline{M}$ pentru f și $e_k \nearrow e_0 = \|f - p_0\|$.*

Demonstrația acestei teoreme se face cu ajutorul a două leme din [6], valabile și în cazul problemei noastre.

L e m a 1. *Dacă iterația nu se termină după k pași, adică $p_0 \neq p_k$ atunci sînt adevărate următoarele afirmații:*

(3) $$\delta(f(s_{i,k+1}) - p_k(s_{i,k+1})) = \delta(f(s_{i,k+1}) - p_{k+1}(s_{i,k+1})),$$
$$i = 1, 2, \ldots, n.$$

(4) $$\delta(f(s_{i,k+1}) - p_{k+1}(s_{i,k+1})) = (-1)^{i+1}\delta(f(s_{1,k+1}) - p_{k+1}(s_{1,k+1})),$$
$$i = 1, 2, \ldots, n.$$

(5) $$e_{k+1} > e_k$$

(6) $$e_{k+1} \geqq \max\ \{f(s_{i,k+1}) - \varphi(s_{i,k+1}),\ i \in \mathcal{L}_{k+1}\}$$

(7) $$e_{k+1} \geqq \max\ \{u(s_{i,k+1}) - f(s_{i,k+1}),\ i \in \mathcal{U}\}$$

(8) *p_{k+1} este elementul de cea mai bună aproximare pentru f, pe $T \cap S_{k+1}$ din mulțimea:*

$$\overline{M}_{k+1} = \{(p \in M : l_{k+1}(s) \leqq p(s) \in L \text{ și } p(s), s \leqq u_{k+1}(s),\ s \in U\}$$

unde:

$$l_{k+1}(s) = \begin{cases} l(s) \ \textit{dacă} \ s = s_{i,k+1} \ \textit{cu} \ i \in \mathcal{L}_{k+1} \\ -\infty \ \textit{în caz contrar} \end{cases}$$

iar

$$u_{k+1}(s) = \begin{cases} u(s) \ \textit{dacă} \ s = s_{i,k+1} \ \textit{cu} \ i \in \mathcal{U}_{k+1} \\ +\infty \ \textit{în caz contrar} \end{cases}$$

L e m a 2. *Pentru orice $k \geqq 2$ există sistemul de numere reale $B_k = \{\beta_{1k}, \ldots, \beta_{n+1,k}\}$ astfel încît:*

(9) $$e_k = \sum_{i \in \mathcal{T}_k} \beta_{ik} f(s_{ik}) + \sum_{i \in \mathcal{L}_k} \beta_{ik}\varphi(s_{ik}) + \sum_{i \in \mathcal{U}_k} \beta_{ik} u(s_{ik})$$

(10) $$\sum_{i \in \mathcal{T}_k} |\beta_{ik}| = 1$$

(11) $$\sum_{=1}^{n+1} \beta_{k}\varphi_j(s_{k}) = 0, \quad j = 1, 2, \ldots, n.$$

(12) $$\beta_{ik}\delta(f(s_{ik}) - p_{k-1}(s_{ik})) > 0, \quad i = 1, 2, \ldots, n + 1.$$

(13) $$\sum_{i=1}^{n+1} |\beta_{ik}| \leqq A < \infty,\ A \ \textit{o constantă independentă de} \ k.$$

(14) $$e_k - e_{k+1} = \sum_{i \in \mathcal{T}_k} |\beta_{ik}|\,||f(s_{ik}) - p_{k-1}(s_{ik})| - e_{k-1}| +$$
$$+ \sum_{i \in \mathcal{L}_k} |\beta_{ik}|\,|l(s_{ik}) - p_{k-1}(s_{ik})| + \sum_{i \in \mathcal{U}_k} |\beta_{ik}|\,|u(s_{ik}) - p_{k-1}(s_{ik})|$$

(15) $$|\beta_{ik}| \geqq \beta > 0,\ \beta \ \textit{o constantă independentă de} \ i \ \textit{și} \ k.$$

Demonstrația teoremei 2. Afirmația (8) ne arată că p_k este un element de cea mai bună aproximare pentru f pe mulțimea $T \cap S_k$ în raport cu funcțiile l_k și u_k. Pe baza unicității elementului de cea mai bună aproximare [2, Teorema 5.1] și pe baza inegalității (5) obținem $e_n < e_0$ pentru orice k deoarece $p_0 \in \overline{M}_k$, adică $e_k \nearrow e \leq e_0$.

Folosind afirmațiile (6) și (7) se deduce existența unei constante $B < 0$ astfel încît $\|p_k\| \leq B$ pentru orice k și deci există $\bar{p} \in \overline{M}$ cu proprietatea că un subșir (p_m) al șirului (p_k) converge uniform către $\bar{p}$.

Avînd în vedere Lema 2 obținem inegalitatea:

$$e_k - e_{k-1} \geq \beta \max \{E_{k-1}, M_{k-1}, m_{k-1}\}$$

care implică:

$$\lim_{k\to\infty} E_k = 0, \quad \limsup_{k\to\infty} M_k \leq 0, \quad \limsup_{k\to\infty} m_k \leq 0$$

adică $\bar{p} \in \overline{M}$ și ținînd seama de unicitatea elementului de cea mai bună aproximare $\bar{p} = p_0$.

4. Exemple. În încheiere se dau cîteva exemple pentru:

$$T = [0,1], \ M = [1, s, s^2, \ldots, s^{n-1}] \ \text{și} \ f(s) = \sqrt{s}, \ s \in S,$$

cuprinse în următorul tabel:

S	n	Restricții	p_0	Observații
[0, 1]	1	—	$s + 0{,}125$	Dat greșit în [6]
[0, 1]	1	$l(0{,}5) = 0{,}7$ $u(0{,}5) = 0{,}71$	$s + 0{,}2$	
$[0,1] \cup \{2\}$	1	$l(0{,}5) = 0{,}7$ $u(0{,}5) = 0{,}71$ $u(2) = 1{,}3$	$0{,}4\ s + 0{,}5$	
[0, 1]	3	$l(0{,}5) = 0{,}7$ $u(0{,}5) = 0{,}71$	$2{,}171525 s^3 -$ $-3{,}90341\ s^2 +$ $+2{,}7319\ s +$ $+0{,}04847$	Din [6]

UN ALGORITHME POUR LA DÉTERMINATION DE L'ÉLÉMENT DE LA MEILLEURE APPROXIMATION AVEC CONTRAINTES

RÉSUMÉ

Dans cette note on présent un algorithme de type Remes [1] pour la détermination de l'élément de la meilleure approximation dans un problème avec contraintes [2].

BIBLIOGRAFIE

[1] Cheney E. W., *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.
[2] Cimoca Gh., *On uniform approximation by functions having restricted ranges*, Mathematica (Cluj) **12** (35), 2, 237–251 (1970).
[3] Jones R. C., Karlovitz, L. A., *Iterative construction of contrained Chebyshev approximation of continuous functions*, SIAM J. Numer. Anal. **5**, 574–585 (1968).
[4] Laurent P. J., *Approximation uniforme de fonctions continues sur un compact avec constraintes de type inégalité*, RIRO, **1**, 5, 81–95 (1967).
[5] Taylor D. G., *Approximation by Functions Having Restricted Ranges III*, J. Math. Anal. Appl., **27**, 241–248 (1969).
[6] Taylor G. D., Winter M. J., *Calculation of best restricted approximations*, SIAM J. Numer. Anal. **7**, 248–255 (1970).

Institutul de calcul din Cluj
al Academiei Republicii Socialiste România

Primit la 2. IX. 1972.