ZASTOSOWANIA MATEMATYKI
VI (1963)

S. PASZKOWSKI (Varsovie)

APPROXIMATION UNIFORME DES FONCTIONS CONTINUES PAR LES FONCTIONS RATIONNELLES

1. Introduction. L'approximation uniforme consiste à remplacer une fonction continue ξ d'une variable réelle par une autre fonction plus simple ω, voisine de ξ. Le but de ce remplacement est, d'ordinaire, de simplifier le calcul approximatif des valeurs de ξ, par exemple à l'aide de la calculatrice automatique. C'est pourquoi on choisit d'habitude pour fonction ω un polynôme dont les valeurs se laissent calculer très simplement, surtout si son degré n'est pas élevé.

Supposons que la fonction ξ soit continue dans l'intervalle $I = \langle -1, 1\rangle$. Le nombre

$$\max_{t\in I} |\xi(t) - \omega(t)|$$

est dit erreur maximum (ou erreur au sens de l'approximation uniforme) de l'approximation de ξ par une autre fonction ω dans cet intervalle. Parmi les polynômes ω de degrés non supérieurs à un nombre naturel fixe n nous tâchons de choisir le polynôme auquel correspond la moindre valeur de l'erreur maximum. Souvent aussi ce dernier polynôme approche insuffisamment la fonction ξ, bien que le degré n soit choisi grand. Ce fait a lieu, par exemple, quand un pôle de cette fonction se trouve près de l'intervalle I. On pourrait supposer que les fonctions rationnelles qui peuvent avoir des singularités analogues, s'assortissent à une telle fonction ξ plus facilement que les polynômes. Par conséquent, l'application des fonctions rationnelles au calcul approximatif des valeurs d'une telle fonction ξ avec une précision fixée, permet d'abréger ce calcul et de réduire le nombre des coefficients employés en comparaison avec les formules polynomiales. Cela en vaut la peine, bien que le calcul qui mène à la fonction rationnele soit assez pénible (en comparaison avec les formules polynomiales). En effet, il suffit d'exécuter ce calcul une fois pour pouvoir appliquer la fonction autant de fois qu'il est nécessaire.

Pour la majorité des méthodes d'approximation rationnelle le point de départ est une série potentielle convergente vers la fonction ξ dans I. On transforme cette série en fonction rationnelle, par exemple par l'intermédiaire des fractions continues (cf. [8]). Récemment on a remarqué qu'on pourra obtenir de meilleurs résultats en appliquant le développement de ξ en série orthogonale de polynômes de Tchebycheff. A vrai dire cette série se rattache à l'approximation en métrique L^2, mais ses sommes partielles approchent la fonction ξ mieux que les sommes correspondantes de la série potentielle, aussi au sens de l'erreur maximum.

Stesin [6] a proposé une méthode de construction d'une fonction rationnelle, fondée sur la série orthogonale mentionnée. Hornecker [2] (voir aussi [3]) a récemment créé une autre méthode, plus générale et plus parfaite. Dans la présente note nous démontrerons qu'il est possible de réduire et de simplifier essentiellement le calcul des coefficients de la fonction rationnelle définie par Hornecker. Le sens de ces améliorations sera indiqué plus tard.

Dans le § 2 nous parlons du choix de la fonction rationnelle approximante, les §§ 3 et 4 contiennent des considérations théoriques. Aux §§ 5 et 6 nous décrivons les moyens de contrôle et l'ordre des calculs, enfin au § 7 nous illustrons la méthode proposée par un simple exemple.

2. Choix d'une fonction approximante. Supposons que la fonction approchée ξ soit développée en série orthogonale uniformément convergente de polynômes de Tchebycheff dans l'intervalle I:

$$\xi(t) = \frac{1}{2}a_0 + \sum_{j=1}^{\infty} a_j \tau_j(t), \tag{1}$$

$\tau_j(t) = \cos(j \arccos t)$ étant le polynôme de Tchebycheff d'ordre j,

$$a_j = \frac{2}{\pi} \int_{-1}^{1} \xi(t)\tau_j(t)(1-t^2)^{-1/2}\,dt \quad (j = 0, 1, \ldots).$$

Entre autres, toutes les fonctions ayant dans I une dérivée première bornée, satisfont à cette hypothèse ([4], p. 245). En fixant les deux nombres $m \geqslant -1$ entier et n naturel, on approche la fonction ξ par une fonction rationnelle de la forme

$$\omega(t) = \frac{1}{2}b_0 + \sum_{j=1}^{m} b_j \tau_j(t) + \sum_{k=1}^{n} \frac{c_k(1-p_k^2)}{2(1-2p_k t+p_k^2)} \tag{2}$$

avec des paramètres réels $b_0, b_1, \ldots, b_m$ et complexes $c_1, c_2, \ldots, c_n$, $p_1, p_2, \ldots, p_n$, où

$$|p_k| < 1 \quad (k = 1, 2, \ldots, n) \tag{3}$$

que nous préciserons plus loin. Par définition, si $m = -1$, la fonction ω se réduit à la dernière somme du second membre de (2) et si $m = 0$, cette fonction se réduit au terme $\frac{1}{2}b_0$ et à la somme mentionnée. On peut démontrer que toute fonction rationnelle dont le numérateur est un polynôme de degré tout au plus $m+n$ et le dénominateur est un polynôme de degré n, n'ayant pas de zéros dans l'intervalle I, s'exprime sous la forme (2). Cela résulte en particulier du fait que la fonction $z = \frac{1}{2}(p+p^{-1})$ transforme le cercle ouvert $|p| < 1$ sur le plan complexe sans I. D'autre part, pour tous les nombres complexes p tels que $|p| < 1$, on a l'identité

$$\frac{1-p^2}{2(1-2pt+p^2)} = \frac{1}{2} + \sum_{j=1}^{\infty} p^j \tau_j(t) \quad (t \in I) \tag{4}$$

([5], p. 54, formule 1.447.3). Elle permet d'écrire la fonction ω sous une forme analogue à celle de ξ:

$$\omega(t) = \frac{1}{2}\left(b_0 + \sum_{k=1}^{n} c_k\right) + \sum_{j=1}^{m}\left(b_j + \sum_{k=1}^{n} c_k p_k^j\right)\tau_j(t) + \tag{5}$$

$$+ \sum_{j=m+1}^{\infty}\left(\sum_{k=1}^{n} c_k p_k^j\right)\tau_j(t).$$

Les paramètres réels $b_0, b_1, \ldots, b_m$ et complexes $c_1, c_2, \ldots, c_n$, p_1, $p_2, \ldots, p_n$ ont été choisis de façon que dans le développement (1) de ξ et dans le développement (5) de ω les termes constants et les coefficients de $\tau_1, \tau_2, \ldots, \tau_{m+2n}$ soient égaux deux à deux:

$$b_j + \sum_{k=1}^{n} c_k p_k^j = a_j \quad (j = 0, 1, \ldots, m), \tag{6_1}$$

$$\sum_{k=1}^{n} c_k p_k^j = a_j \quad (j = m+1, m+2, \ldots, m+2n) \tag{6_2}$$

(pour $m = -1$ ce système d'équations se réduit, bien entendu, aux équations (6_2)). On peut espérer que la fonction ω définie de cette manière fournisse une bonne approximation de la fonction ξ.

La solution du système (6) n'existe pas toujours. Si elle existe, elle peut n'avoir aucun sens pratique lorsque les inégalités (3) assurant la

convergence de la série (5), n'ont pas lieu. Nous allons indiquer les conditions d'existence de la solution du système (6) et les moyens de vérifier les inégalités (3).

3. L'existence de la solution du système (6) et la forme de cette solution seront étudiées au moyen de la méthode appliquée dans des problèmes analogues (cf. [8]). Après avoir fixé ξ, on introduit la fonctionnelle Ξ définie pour tout polynôme de degré non supérieur à $2n-1$, avec les coefficients $g_0, g_1, \ldots, g_{2n-1}$:

$$\Xi\Big(\sum_{k=0}^{2n-1} g_k t^k\Big) = \sum_{k=0}^{2n-1} a_{k+m+1} g_k.$$

Nous définissons aussi par récurrence une suite de polynômes $v_0, v_1, \ldots, v_n$ (la théorème 1 donne les conditions assurant que cette définition soit correcte):

$$(7)\qquad r_k = -\frac{\Xi(tv_{k-1}^2)}{\Xi(v_{k-1}^2)} \quad (k=1,2,\ldots,n),$$
$$s_1 = 0, \quad s_k = -\frac{\Xi(v_{k-1}^2)}{\Xi(v_{k-2}^2)} \quad (k=2,3,\ldots,n),$$

$$(8)\qquad v_{-1}(t) = 0, \quad v_0(t) = 1,$$
$$v_k(t) = (r_k+t)v_{k-1}(t) + s_k v_{k-2}(t) \quad (k=1,2,\ldots,n).$$

THÉORÈME 1. *Si*

$$(9)\qquad \begin{vmatrix} a_{m+1} & a_{m+2} & \ldots & a_{m+k} \\ a_{m+2} & a_{m+3} & \ldots & a_{m+k+1} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m+k} & a_{m+k+1} & \ldots & a_{m+2k-1} \end{vmatrix} \neq 0 \quad (k=1,2,\ldots,n),$$

les polynômes $v_1, v_2, \ldots, v_n$ *existent,* v_k *est de degré* k *et le coefficient de* t^k *est* 1 $(k=1,2,\ldots,n)$.

Démonstration. On peut vérifier que pour tous les coefficients $a_{m+1}, a_{m+2}, \ldots, a_{m+2n}$ il existe une fonction ϱ dont les $2n$ moments initiaux sont égaux à ces coefficients:

$$(10)\qquad \int_{-1}^{1} t^k \varrho(t)\,dt = a_{k+m+1} \quad (k=0,1,\ldots,2n-1).$$

Cette fonction peut prendre des valeurs négatives. Néanmoins, on peut définir les polynômes $v_0, v_1, \ldots, v_n$ de degrés successifs, orthogonaux avec

le poids ϱ, dès que les déterminants de Gram de la fonction ϱ ne sont pas nuls ([4], p. 333). On déduit de là l'hypothèse (9). Les polynômes $v_0, v_1, \ldots, v_n$ satisfont à la relation de récurrence (8) où

$$r_k = \frac{\int_{-1}^{1} \varrho(t) t v_{k-1}^2(t)\,dt}{\int_{-1}^{1} \varrho(t) v_{k-1}^2(t)\,dt}, \qquad s_k = \frac{\int_{-1}^{1} \varrho(t) v_{k-1}^2(t)\,dt}{\int_{-1}^{1} \varrho(t) v_{k-2}^2(t)\,dt}$$

([4], p. 341). En vertu de la définition de la fonctionnelle Ξ et des moments de ϱ, on peut écrire ces égalités sous la forme (7). Il est tout à fait évident que le polynôme v_k est de degré k et le coefficient de t^k est 1.

THÉORÈME 2. *Si l'hypothèse* (9) *est satisfaite et, si en outre, pour* $m \geqslant 0$,

$$\text{(11)} \qquad \begin{vmatrix} a_{m+2} & a_{m+3} & \ldots & a_{m+n+1} \\ a_{m+3} & a_{m+4} & \ldots & a_{m+n+2} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m+n+1} & a_{m+n+2} & \ldots & a_{m+2n} \end{vmatrix} \neq 0,$$

il existe une solution du système (6) *telle que les nombres* $p_1, p_2, \ldots, p_n$ *soient tous des zéros du polynôme* v_n *et*

$$\text{(12)} \qquad c_k = \Xi\left(\prod_{i=1, i\neq k}^{n} \frac{t-p_i}{p_k-p_i}\right) p_k^{-m-1} \qquad (k = 1, 2, \ldots, n),$$

$$\text{(13)} \qquad b_j = a_j - \sum_{k=1}^{n} c_k p_k^j \qquad (j = 0, 1, \ldots, m).$$

Démonstration. Soient $p_1, p_2, \ldots, p_n$ des zéros du polynôme v_n et soit

$$c_k' = \int_{-1}^{1} \varrho(t) \prod_{i=1, i\neq k}^{n} \frac{t-p_i}{p_k-p_i}\,dt \qquad (k = 1, 2, \ldots, n).$$

Le théorème fondamental concernant les quadratures de Gauss, valable aussi pour la fonction ϱ définie dans ce paragraphe, nous montre que les égalités

$$\int_{-1}^{1} \varrho(t) t^j\,dt = \sum_{k=1}^{n} c_k' p_k^j \qquad (j = 0, 1, \ldots, 2n-1)$$

([4], p. 602) ont été vérifiées. Conformément à (10) on peut écrire ces égalités et la définition des nombres c_k' sous la forme suivante:

$$\sum_{k=1}^{n} c_k' p_k^j = a_{j+m+1} \quad (j = 0, 1, \ldots, 2n-1), \tag{14}$$

$$c_k' = \Xi\left(\prod_{i=1, i\neq k}^{n} \frac{t-p_i}{p_k-p_i}\right) \quad (k = 1, 2, \ldots, n).$$

Si $m = -1$, en prenant $c_k = c_k'$ $(k = 1, 2, \ldots, n)$, nous obtenons un système d'équations coïncidant avec le système (6_2). Supposons maintenant que $m \geqslant 0$. Comme

$$v_n(t) = \begin{vmatrix} 1 & t & \ldots & t^{n-1} & t^n \\ a_{m+1} & a_{m+2} & \ldots & a_{m+n} & a_{m+n+1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{m+n} & a_{m+n+1} & \ldots & a_{m+2n-1} & a_{m+2n} \end{vmatrix} : \begin{vmatrix} a_{m+1} & a_{m+2} & \ldots & a_{m+n} \\ a_{m+2} & a_{m+3} & \ldots & a_{m+n+1} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m+n} & a_{m+n+1} & \ldots & a_{m+2n-1} \end{vmatrix}$$

([4], p. 333), le terme constant du polynome v_n égale le quotient des deux déterminants. En vertu de l'hypothèse (11), le numérateur de ce quotient est non nul; par suite tous les zéros du polynôme v_n sont non nuls et les définitions

$$c_k = c_k' p_k^{-m-1} \quad (k = 1, 2, \ldots, n),$$

c'est-à-dire les définitions (12), sont correctes. Il en résulte l'identité des systèmes (14) et (6_2). En définissant les nombres b_k par les formules (13) pour m quelconque on satisfait aussi au système (6_1).

On sait déjà que la solution du système (6) existe sous l'hypothèse (9) et, pour $m \geqslant 0$, sous l'hypothèse (11). Toutefois il est inutile de calculer les déterminants de ces hypothèses. La première d'entre elles équivaut aux égalités $\Xi(v_k^2) \neq 0$ pour $k = 0, 1, \ldots, n-1$. Antérieurement nous avons constaté que la seconde hypothèse nous montre que le terme constant d_0 du polynôme

$$v_n(t) = d_0 + d_1 t + \ldots + d_{n-1} t^{n-1} + d_n t^n \quad (d_n = 1)$$

est non nul.

Après avoir calculé le polynôme v_n il faut encore vérifier si ses zéros satisfont à l'inégalité (3). Dans ce but on peut utiliser, par exemple, le théorème suivant de Westerfield [7]:

Soit la suite de nombres $q_0, q_1, \ldots, q_{n-1}$ résultant de l'arrangement des nombres

$$\sqrt[n]{|d_0|}, \quad \sqrt[n-1]{|d_1|}, \quad \ldots, \quad |d_{n-1}|$$

dans l'ordre décroissant. Alors tous les zéros du polynôme v_n sont situés dans le cercle

$$|z| \leqslant q_0 + 0{,}6180 q_1 + 0{,}3811 q_2$$

(et à plus forte raison, dans le cercle $|z| \leqslant q_0 + q_1$).

Notons encore que les expressions r_k et s_k des formules (8) se laissent calculer par l'algorithme q-d (cf. [8]). Alors le nombre des opérations arithmétiques serait un peu plus petit, mais l'hypothèse (9) ne serait plus suffisante.

4. Calcul d'une fonction approximante. Pour trouver la fonction ω on peut se baser directement sur la définition des polynômes $v_0, v_1, \ldots, v_n$ et le théorème 2, c'est-à-dire, après avoir trouvé v_n, calculer successivement les zéros (en général complexes) du polynôme v_n et les coefficients $c_1, c_2, \ldots, c_n$ et $b_0, b_1, \ldots, b_m$. C'est à peu près ainsi que procède Hornecker dans [2]. Nous présenterons en abrégé les étapes successives de sa méthode.

Hornecker propose de calculer les coefficients du polynôme v_n à partir du système d'équations

$$\sum_{j=0}^{n} a_{m+j+k} d_j = 0 \qquad (k = 1, 2, \ldots, n).$$

Cela réduit les hypothèses (9) à la dernière d'entre elles et permet de négliger les polynômes $v_0, v_1, \ldots, v_{n-1}$. Néanmoins, le nombre des opérations nécessaires pour résoudre le système ci-dessus est plus grand que dans le cas où l'on applique les formules (7) et (8) (cf. § 6). D'ailleurs, en général, nous ne fixons pas d'avance le paramètre n, mais nous le choisissons au cours du calcul de sorte que l'erreur d'approximation de ξ par ω soit suffisamment petite. Le méthode basée sur les formules (7) et (8) permet sans difficulté d'augmenter n si cela est nécessaire.

Ensuite, selon Hornecker, il faudrait trouver les zéros $p_1, p_2, \ldots, p_n$ du polynôme v_n et, à partir du système d'équations (6_2) à coefficients complexes p_k^j, trouver les nombres $c_1, c_2, \ldots, c_n$. Enfin on calculerait les coefficients de la fonction ω, en laquelle on réunirait en couples les fonctions rationnelles $c_k(1-p_k^2)/2(1-2p_k t+p_k^2)$ qui correspondent aux zéros complexes conjugués du polynôme v_n. Alors la fonction ω serait égale à la somme d'un polynôme et de fractions simples à coefficients réels, mais pour les trouver il faudrait encore exécuter des opérations supplémentaires sur des nombres complexes.

Dans ce paragraphe nous proposons une modification de la méthode de Hornecker. Elle donne la même fonction rationnelle ω, mais sous forme du quotient d'un polynôme de degré tout au plus $m+n$ par un polynôme de degré n. Ce quotient provient de la réduction au dénominateur commun

de tous les termes de ω. On verra que pour calculer les coefficients de ces deux polynômes il est inutile de trouver les zéros $p_1, p_2, \ldots, p_n$ (il suffit de constater que les inégalités (3) ont lieu) et les coefficients $c_1, c_2, \ldots, c_n$. Les opérations sur les nombres complexes sont également inutiles. Cette forme de la fonction rationnelle ω facilite les opérations soit lorsque ω constitue le but final du calcul, soit lorsque nous voulons transformer ω de nouveau, cette fois-ci en fraction continue ce qui est souvent pratiqué. Il est aussi facile ne réduire au dénominateur commun que la partie de ω qui se compose de fractions simples.

Nous donnons tout d'abord un lemme précisant le mode du calcul des valeurs auxiliaires

$$a_j^* = \sum_{k=1}^{n} c_k p_k^j \quad (j = 0, 1, \ldots),$$

dont le rôle sera expliqué.

THÉORÈME 3 ([2]). *On a*

$$\sum_{i=0}^{n} d_i a_{i+j}^* = 0 \quad (j = 0, 1, \ldots). \tag{15}$$

Ces relations de récurrence, dans lesquelles $d_0, d_1, \ldots, d_n = 1$ sont les coefficients du polynôme v_n, et les égalités

$$a_j^* = a_j \quad (j = m+1, m+2, \ldots, m+2n) \tag{16}$$

qui résultent de (6_2), permettent de calculer successivement a_{m+2n+1}^*, a_{m+2n+2}^*, ... et (dans l'hypothèse $d_0 \neq 0$) $a_m^*, a_{m-1}^*, \ldots, a_0^*$. Il est facile d'établir ces relations:

$$\sum_{i=0}^{n} d_i a_{i+j} = \sum_{i=0}^{n} d_i \sum_{k=1}^{n} c_k p_k^{i+j} = \sum_{k=1}^{n} c_k p_k^j \sum_{i=0}^{n} d_i p_k^i = \sum_{k=1}^{n} c_k p_k^j v_n(p_k) = 0.$$

THÉORÈME 4 ([2]). *L'erreur maximum d'approximation de la fonction ξ par la fonction rationnelle ω satisfait à l'inégalité*

$$\max_{t \in I} |\xi(t) - \omega(t)| \leqslant \sum_{j=m+2n+1}^{\infty} |a_j - a_j^*|. \tag{17}$$

Démonstration. Soit

$$\omega^*(t) = \sum_{k=1}^{n} \frac{c_k(1-p_k^2)}{2(1-2p_k t + p_k^2)}.$$

Conformément à l'identité (4), on a

$$\omega^* = \frac{1}{2}a_0^* + \sum_{j=1}^{\infty} a_j^* \tau_j \tag{18}$$

et de la définition (6) des paramètres b_j, c_k et p_k il résulte que

$$\xi - \omega = \sum_{j=m+2n+1}^{\infty} (a_j - a_j^*)\tau_j.$$

Ayant démontré de cette manière le théorème 4, nous passerons ensuite à la déduction des formules pour le numérateur λ et le dénominateur μ de la fonction rationnelle ω. On peut définir le polynôme μ comme il suit:

$$\mu(t) = \frac{1}{2}\prod_{k=1}^{n}(1 - 2p_k t + p_k^2).$$

THÉORÈME 5. *Le polynôme μ s'exprime par la formule*

$$\mu = \tfrac{1}{2}e_0 + e_1\tau_1 + \ldots + e_n\tau_n,$$

où

$$e_k = \sum_{i=0}^{n-k} d_i d_{i+k} \quad (k = 0, 1, \ldots, n). \tag{19}$$

Démonstration. Pour $t = \frac{1}{2}y + \frac{1}{2}y^{-1}$ on obtient

$$\mu(t) = \frac{1}{2}\prod_{k=1}^{n}(1 - p_k y - p_k y^{-1} + p_k^2)$$

$$= \frac{1}{2}\prod_{k=1}^{n}(y - p_k)(y^{-1} - p_k) = \frac{1}{2}\nu_n(y)\nu_n(y^{-1}).$$

Comme

$$\nu_n(y) = d_0 + d_1 y + \ldots + d_n y^n,$$

$$\nu_n(y^{-1}) = d_0 + d_1 y^{-1} + \ldots + d_n y^{-n},$$

on a

$$\nu_n(y)\nu_n(y^{-1}) = e_0 + e_1(y + y^{-1}) + \ldots + e_n(y^n + y^{-n}),$$

les coefficients e_k étant définis par la formule (19).
Si $t \notin I$, on a $y = t + \sqrt{t^2 - 1}$ et $y^{-1} = t - \sqrt{t^2 - 1}$, d'où

$$y^k + y^{-k} = (t + \sqrt{t^2 - 1})^k + (t - \sqrt{t^2 - 1})^k = 2\tau_k(t) \quad (k = 1, 2, \ldots)$$

([4], p. 72), ce qui établit la théorème 5. Il est évident que la formule pour le polynôme μ de ce théorème est valable pour tout t.

THÉORÈME 6. *On a*

$$\omega = \frac{\lambda}{\mu},$$

où

$$\lambda = \frac{1}{2}f_0 + \sum_{k=1}^{m+n} f_k \tau_k, \tag{20}$$

$$\frac{1}{2}f_0 = \frac{1}{2}\left(\frac{1}{2}a_0 e_0 + \sum_{l=1}^{n} a_l e_l\right),$$

$$f_k = \frac{1}{2}\left(a_k e_0 + \sum_{l=1}^{n} (a_{|k-l|} + a_{k+l}) e_l\right) \quad (k = 1, 2, \ldots, m+n). \tag{21}$$

Démonstration. Conformément à la définition du polynôme λ on a

$$\lambda = \omega\mu = \left(\frac{1}{2}a_0 + \sum_{k=1}^{m+2n} a_k \tau_k + \sum_{k=m+2n+1}^{\infty} a_k^* \tau_k\right)\left(\frac{1}{2}e_0 + \sum_{l=1}^{n} e_l \tau_l\right).$$

Le degré de ce polynôme n'est pas supérieur à $m+n$, ce qui prouve l'existence des nombres $f_0, f_1, \ldots, f_{m+n}$ satisfaisant à la formule (20). Appliquant les identités

$$\tau_k \tau_l = \tfrac{1}{2}(\tau_{|k-l|} + \tau_{k+l}) \quad (k, l = 1, 2, \ldots)$$

on obtient sans difficulté les formules (21). Il n'y a pas, dans ces formules, de coefficients a_k^* pour $k \geqslant m+2n+1$. Parfois il est préférable de ne réduire au dénominateur commun que la partie ω^* de la fonction rationnelle ω. On exprime alors la fonction ω par la somme d'un polynôme $\varkappa$ de degré tout au plus m et du quotient d'un polynôme λ^* de degré tout au plus $n-1$ par un polynôme μ.

THÉORÈME 7. *On a*

$$\omega = \varkappa + \frac{\lambda^*}{\mu},$$

où

$$\varkappa = \frac{1}{2}(a_0 - a_0^*) + \sum_{j=1}^{m} (a_j - a_j^*)\tau_j,$$

$$\lambda^* = \frac{1}{2}f_0^* + \sum_{k=1}^{n-1} f_k^* \tau_k,$$

$$\frac{1}{2}f_0^* = \frac{1}{2}\left(\frac{1}{2}a_0^* e_0 + \sum_{l=1}^{n} a_l^* e_l\right),$$

(22)

$$f_k^* = \frac{1}{2}\left(a_k^* e_0 + \sum_{l=1}^{n} (a_{|k-l|}^* + a_{k+l}^*) e_l\right) \quad (k = 1, 2, \ldots, n-1).$$

La forme du polynôme $\varkappa$ résulte des formules (13) et de la définition des coefficients a_j^*. Tenant compte de la formule (18), on obtient les coefficients $f_0^*, f_1^*, \ldots, f_{n-1}^*$ de même que $f_0, f_1, \ldots, f_{m+n}$ dans le théorème précédent.

Après avoir appliqué le théorème 6 ou 7 on doit encore trouver les coefficients des puissances successives de la variable t dans les polynômes $\lambda(t)$ et $\mu(t)$ ou $\varkappa(t)$, $\lambda^*(t)$ et $\mu(t)$. Nous utilisons ici la formule connue de récurrence pour les polynômes de Tchebycheff:

$$\tau_0(t) = 1, \quad \tau_1(t) = 1, \quad \tau_k(t) = 2t\tau_{k-1}(t) - \tau_{k-2}(t) \quad (k = 2, 3, \ldots),$$

ou directement les égalités qui en résultent:

$$\begin{aligned}
\tau_0(t) &= 1,\\
\tau_1(t) &= t,\\
\tau_2(t) &= 2t^2 - 1,\\
\tau_3(t) &= 4t^3 - 3t,\\
\tau_4(t) &= 8t^4 - 8t^2 + 1,\\
\tau_5(t) &= 16t^5 - 20t^3 + 5t,\\
\tau_6(t) &= 32t^6 - 48t^4 + 18t^2 - 1,\\
\tau_7(t) &= 64t^7 - 112t^5 + 56t^3 - 7t,\\
\tau_8(t) &= 128t^8 - 256t^6 + 160t^4 - 32t^2 + 1, \ldots
\end{aligned} \tag{23}$$

5. Contrôle du calcul. Le calcul qui mène à la fonction rationnelle ω n'est pas trop compliqué et sa vérification peut être limitée à deux opérations: 1° après avoir calculé le polynôme ν_n on vérifie la condition „d'orthogonalité" des polynômes $\nu_0 = 1$ et ν_n:

$$\Xi(\nu_n) = 0, \tag{24}$$

2° après avoir calculé les coefficients de ω on vérifie les inégalités

$$|\xi(\pm 1) - \omega(\pm 1)| \leqslant \sum_{j=m+2n+1}^{\infty} |a_j - a_j^*|, \tag{25}$$

qui résultent du théorème 4.

6. Schéma du calcul. Pour le calcul des coefficients de la fonction rationnelle ω on suivra les étapes que voici:

I. Calcul des coefficients $d_0, d_1, \ldots, d_n$ du polynôme ν_n à l'aide des formules de récurrence (8) et (7).

II. Contrôle de l'étape I (vérification de l'égalité (24)).

III. Vérification (à l'aide, par exemple, du théorème de Westerfield cité à la fin du § 3) si tous les zéros du polynôme ν_n sont situés dans le cercle $|z| < 1$. Sinon, il faut revenir à l'étape I et calculer le polynôme ν_{n+1} qui, peut-être, satisfera déjà à la condition demandée.

IV. Calcul des coefficients $a^*_{m+2n+1}, a^*_{m+2n+2}, \ldots$ des formules (15) et (16).

V. Examen de l'erreur maximum $\max\limits_{t\in I}|\xi(t)-\omega(t)|$ à l'aide de l'inégalité (17). Si son second membre est trop grand, il faut revenir à l'étape I et calculer le polynôme ν_{n+1}.

VI. Calcul des coefficients $e_0, e_1, \ldots, e_n$ du polynôme μ à partir des formules (19).

Les opérations restantes dépendent de la forme demandée de la fonction ω. Supposons d'abord que nous voulions l'obtenir sous forme du quotient $\omega = \lambda/\mu$ (cf. le théorème 6). Dans ce cas on doit exécuter les opérations suivantes:

VII. Calcul des coefficients $f_0, f_1, \ldots, f_{m+n}$ du polynôme λ à partir des formules (21).

VIII. Calcul des coefficients de $t^0, t^1, \ldots, t^n$ du polynôme μ et de $t^0, t^1, \ldots, t^{m+n}$ du polynôme λ à l'aide des formules (23).

IX. Contrôle des étapes III-VIII (vérification de l'inégalité (25)).

Supposons maintenant que nous voulions obtenir la fonction rationnelle ω sous forme de la somme $\omega = \varkappa + \lambda^*/\mu$. Le cas échéant, on doit exécuter les opérations suivantes:

VII*. Calcul des coefficients $a^*_0, a^*_1, \ldots, a^*_m$ des formules (15) et (16), et calcul des coefficients $\frac{1}{2}(a_0 - a^*_0), a_1 - a^*_1, \ldots, a_m - a^*_m$ du polynôme $\varkappa$.

VIII*. Calcul des coefficients $f^*_0, f^*_1, \ldots, f^*_{n-1}$ du polynôme λ^* des formules (22).

IX*. Calcul des coefficients de $t^0, t^1, \ldots, t^m$ du polynôme $\varkappa$, de $t^0, t^1, \ldots, t^n$ du polynôme μ et de $t^0, t^1, \ldots, t^{n-1}$ du polynôme λ^* à l'aide des formules (23).

X*. Contrôle des étapes III-VI et VII*-IX* (vérification de l'inégalité (25)).

Désignons par l le nombre des coefficients $a^*_{m+2n+1}, a^*_{m+2n+2}, \ldots$ calculés pour trouver le second membre de l'inégalité (17) aussi exactement

qu'on veut. Supposons en outre qu'on ait choisi convenablement le degré n, c'est-à-dire que les résultats des étapes III et V permettent de continuer le calcul. Alors on peut fixer le nombre des opérations arithmétiques qui seront à exécuter pour obtenir la fonction ω. Nous donnons ce nombre à part pour chacune des étapes, sauf l'étape III, où ce nombre est difficile à établir. Parmi les additions et les soustractions nous avons inséré la division d'un nombre par 2. On désigne par $[x]$ la partie entière de x.

Etapes	Additions et soustractions	Multiplications	Divisions
I	$\frac{1}{6}n(n-1)(2n+17)$	$\frac{1}{6}(n-1)(n^2\ldots 19n-6)$	$2n-1$
II	n	n	0
IV	$(n-1)l$	nl	0
V	$2l-1$	0	0
VI	$\frac{1}{2}n(n+1)+1$	$\frac{1}{2}n(n+1)$	0
VII	$2n^2+(m+1)(2n+1)$	$(n+1)(m+n+1)$	0
VIII	$[\frac{1}{4}(m+n)^2]+[\frac{1}{4}n^2]$	$[\frac{1}{4}(m+n+1)^2]+[\frac{1}{4}(n+1)^2]-2$	0
IX	$2m+4n$	0	2
VII*	$(m+1)(n-1)$	$(m+1)(n-1)$	$m+1$
VIII*	$2n^2+m+2$	$n(n+1)$	0
IX*	$[\frac{1}{4}m^2]+\frac{1}{2}n(n-1)$	$[\frac{1}{4}(m+1)^2]+\frac{1}{2}n(n+1)-3$	0
X*	$2m+4n$	0	2

7. **Exemple de calcul.** Voici le développement de la fonction e^t en série orthogonale de polynômes de Tchebycheff:

$$e^t = I_0\left(\frac{1}{2}\right)+2\sum_{j=1}^{\infty} I_j\left(\frac{1}{2}\right)\tau_j(t), \tag{26}$$

I_n étant la fonction de Bessel de première espèce, d'ordre n, d'une variable imaginaire. On obtient alors pour e^t

$$a_j = 2I_j(\tfrac{1}{2}) \quad (j = 0, 1, \ldots),$$

d'où il résulte (cf. ([1]) que

$$\begin{aligned}
a_0 &= 2{,}53213176, & a_1 &= 1{,}13031821, & a_2 &= 0{,}271495340,\\
a_3 &= 0{,}0443368498, & a_4 &= 0{,}0^2547424044, & a_5 &= 0{,}0^3542926312,\\
a_6 &= 0{,}0^4449773230, & a_7 &= 0{,}0^5319843646, & a_8 &= 0{,}0^6199212481,\\
a_9 &= 0{,}0^7110367717, & a_{10} &= 0{,}0^9550589608, & a_{11} &= 0{,}0^{10}249795662.
\end{aligned}$$

Nous supposons que $m = -1$ et $n = 3$. Par conséquent la fonction rationnelle cherchée ω sera le quotient d'un polynôme de degré au plus 2 par un polynôme de degré 3. La fonctionnelle Ξ est définie par la formule

$$\Xi\Big(\sum_{k=0}^{5} g_k t^k\Big) = \sum_{k=0}^{5} a_k g_k.$$

Etape I. Appliquant les formules (8) et (7) on calcule les polynômes ν_1, ν_2, ν_3:

$$\nu_{-1}(t) = 0, \quad \nu_0(t) = 1, \quad \nu_0^2(t) = 1,$$

$$\Xi(\nu_0^2) = 2{,}53213176, \quad \Xi(t\nu_0^2) = 1{,}13031821,$$

$$r_1 = -\frac{\Xi(t\nu_0^2)}{\Xi(\nu_0^2)} = -0{,}446389966, \quad s_1 = 0,$$

$$\nu_1(t) = (r_1+t)\nu_0(t) + s_1\nu_{-1}(t) = -0{,}446389966 + t,$$

$$\nu_1^2(t) = 0{,}199264002 - 0{,}892779932t + t^2,$$

$$\Xi(\nu_1^2) = -0{,}233067367, \quad \Xi(t\nu_1^2) = 0{,}027182989,$$

$$r_2 = \frac{\Xi(t\nu_1^2)}{\Xi(\nu_1^2)} = 0{,}116631467, \quad s_2 = -\frac{\Xi(\nu_1^2)}{\Xi(\nu_0^2)} = 0{,}0920439334,$$

$$\nu_2(t) = (r_2+t)\nu_1(t) + s_2\nu_0(t) = 0{,}0399808168 - 0{,}329758499t + t^2,$$

$$\nu_2^2(t) = 0{,}00159846571 - 0{,}0263680282t + 0{,}188702302t^2 - 0{,}659516998t^3 + t^4,$$

$$\Xi(\nu_2^2) = 0{,}0017083933, \quad \Xi(t\nu_2^2) = -0{,}00005298457,$$

$$r_3 = -\frac{\Xi(t\nu_2^2)}{\Xi(\nu_2^2)} = 0{,}0310142694, \quad s_3 = -\frac{\Xi(\nu_2^2)}{\Xi(\nu_1^2)} = 0{,}00733004076,$$

$$\nu_3(t) = (r_3+t)\nu_2(t) + s_3\nu_1(t)$$

$$= -0{,}00203208083 + 0{,}0370836387t - 0{,}298744230t^2 + t^3.$$

Par suite

$$d_0 = -0{,}00203208083, \; d_1 = 0{,}0370836387, \; d_2 = -0{,}298744230, \; d_3 = 1.$$

Etape II. Vérification de l'égalité (24):

$$\Xi(\nu_3) = a_0d_0 + a_1d_1 + a_2d_2 + a_3d_3 = -0{,}0000000008.$$

Etape III. Application du théorème de Westerfield au polynôme ν_3:

$$\sqrt[3]{|d_0|} \approx 0{,}127, \quad \sqrt[2]{|d_1|} \approx 0{,}193, \quad |d_2| \approx 0{,}299,$$

$$q_0 = 0{,}299, \quad q_1 = 0{,}193, \quad q_2 = 0{,}127,$$

les zéros du polynôme v_3 sont situés dans le cercle

$$|z| \leqslant q_0 + q_1 = 0{,}492 < 1.$$

Etape IV. Calcul des coefficients $a_6^*, a_7^*, \ldots$ à partir des formules (15) et (16):

$$\begin{aligned}
a_6^* &= -(d_0 a_3 + d_1 a_4 + d_2 a_5) = 0{,}00004928740,\\
a_7^* &= -(d_0 a_4 + d_1 a_5 + d_2 a_6^*) = 0{,}00000571474,\\
a_8^* &= -(d_0 a_5 + d_1 a_6^* + d_2 a_7^*) = 0{,}000000982759,\\
a_9^* &= -(d_0 a_6^* + d_1 a_7^* + d_2 a_8^*) = 0{,}0000001818262,\\
a_{10}^* &= 0{,}00000002948805,\\
a_{11}^* &= 0{,}000000004063653, \ldots
\end{aligned}$$

Etape V. Examen de l'erreur d'approximation $\max\limits_{t \in I} |e^t - \omega(t)|$ à l'aide de l'inégalité (17):

$$\begin{aligned}
a_6 - a_6^* &= -0{,}00000431008, & a_7 - a_7^* &= -0{,}00000251630,\\
a_8 - a_8^* &= -0{,}000000783547, & a_9 - a_9^* &= -0{,}0000001707894,\\
a_{10} - a_{10}^* &= -0{,}00000002893746, & a_{11} - a_{11}^* &= -0{,}000000004038673,
\end{aligned}$$

$$\max_{t \in I} |e^t - \omega(t)| \leqslant \sum_{j=6}^{\infty} |a_j - a_j^*| \leqslant 0{,}00000782. \tag{27}$$

A titre de comparaison il est utile de noter que la somme partielle

$$I_0\left(\frac{1}{2}\right) + 2\sum_{j=1}^{5} I_j\left(\frac{1}{2}\right)\tau_j$$

de la série (26) contenant autant de coefficients que la fonction ω approche la fonction e^t avec l'erreur 0,0000484.

Etape VI. Calcul des coefficients du polynôme μ à partir des formules (19):

$$\begin{aligned}
e_0 &= d_0^2 + d_1^2 + d_2^2 + d_3^2 = 1{,}090627440,\\
\tfrac{1}{2} e_0 &= 0{,}545313720,\\
e_1 &= d_0 d_1 + d_1 d_2 + d_2 d_3 = -0{,}309898110,\\
e_2 &= d_0 d_2 + d_1 d_3 = 0{,}0376907111,\\
e_3 &= d_0 d_3 = -0{,}00203208083.
\end{aligned}$$

Par suite,

$$\mu = 0{,}545313720 - 0{,}309898110\,\tau_1 + 0{,}0376907111\tau_2 - 0{,}00203208083\tau_3.$$

Etape VII. Calcul des coefficients du polynôme λ par les formules (21):

$$\begin{aligned}
\tfrac{1}{2}f_0 &= \tfrac{1}{2}(\tfrac{1}{2}a_0e_0+a_1e_1+a_2e_2+a_3e_3) = 0{,}520332735,\\
f_1 &= \tfrac{1}{2}\big(a_1e_0+(a_0+a_2)e_1+(a_1+a_3)e_2+(a_2+a_4)e_3\big) = 0{,}203814038,\\
f_2 &= \tfrac{1}{2}\big(a_2e_0+(a_1+a_3)e_1+(a_0+a_4)e_2+(a_1+a_5)e_3\big) = 0{,}012711529.
\end{aligned}$$

Par suite,

$$\lambda = 0{,}520332735+0{,}203814038\tau_1+0{,}012711529\tau_2.$$

Etape VIII. Calcul des coefficients de t^0, t^1, t^2, t^3 du polynôme μ et des coefficients de t^0, t^1, t^2 du polynôme λ à l'aide des formules (23):

$$\begin{aligned}
\mu(t) &= 0{,}507623009-0{,}303801868t+0{,}075381422t^2-0{,}008128323t^3,\\
\lambda(t) &= 0{,}507621206+0{,}203814038t+0{,}025423058t^2.
\end{aligned}$$

Etape IX. Vérification de l'inégalité (25):

$$\begin{aligned}
\lambda(-1) &= 0{,}329230226, & \mu(-1) &= 0{,}894934622,\\
\omega(-1) &= 0{,}36788187, & |e^{-1}-\omega(-1)| &= 0{,}00000243,\\
\lambda(1) &= 0{,}736858302, & \mu(1) &= 0{,}271074240,\\
\omega(1) &= 2{,}71828965, & |e-\omega(1)| &= 0{,}00000782,
\end{aligned}$$

ce qui s'accorde avec (27).

Finalement, nous obtenons la formule approchée

$$e^t \approx \frac{19{,}966961+8{,}016897t+t^2}{19{,}967032-11{,}949855t+2{,}965081t^2-0{,}319722t^3}$$

avec une erreur non supérieure à 0,00000782.

Travaux cités

[1] A. Gray and G. B. Mathew, *A treatise on Bessel functions*, London 1922.

[2] G. Hornecker, *Approximations rationnelles voisines de la meilleure approximation au sens de Tchebycheff*, Compt. Rend. Hebd. Acad. Sci. 249 (1959), p. 939-941.

[3] — *Détermination des meilleures approximations rationnelles (au sens de Tchebycheff des fonctions réelles d'une variable sur un segment fini et des bornes d'erreur correspondantes*, Compt. Rend. Hebd. Acad. Sci. 249 (1959), p. 2265-2267.

[4] I. P. Natanson (И. П. Натансон), *Конструктивная теория функций*, Москва 1949.

[5] I. M. Ryzhik et I. S. Gradstein (И. М. Рыжик и И. С. Градштейн), *Таблицы интегралов, сумм, рядов и произведений*, Москва 1951.

[6] I. M. Stesin (И. М. Стесин), *Обращение ортогональных разложений в последовательность подходящих дробей*, Вычислительная Математика, сборник 1, Москва 1957, p. 116-119.

[7] E. C. Westerfield, *A new bound for the zeros of polynomials*, Amer. Math. Monthly 40 (1933), p. 18-23.

[8] P. Wynn, *The rational approximation of functions which are formally defined by a power series expansion*, Math. Comput. 14 (1960), p. 147-186.

INSTYTUT MATEMATYCZNY POLSKIEJ AKADEMII NAUK

Praca wpłynęła 23. 11. 1961

S. PASZKOWSKI (Warszawa)

APROKSYMACJA JEDNOSTAJNA FUNKCJI CIĄGŁYCH ZA POMOCĄ FUNKCJI WYMIERNYCH

STRESZCZENIE

Na elektronowych maszynach cyfrowych wartości funkcji ciągłych oblicza się zwykle za pomocą wielomianów, dostatecznie dobrze aproksymujących te funkcje w sensie aproksymacji jednostajnej. Wiadomo jednak, że wielomiany są złym narzędziem aproksymacji funkcji mających biegun w pobliżu przedziału aproksymacji. W takich przypadkach lepiej posługiwać się funkcjami wymiernymi.

Hornecker w [2] i [3] zaproponował algorytm wyznaczania funkcji wymiernej, która — w określonej klasie podobnych funkcji — prawie najlepiej aproksymuje daną funkcję ciągłą. Metoda ta opiera się na rozwinięciu (1) funkcji przybliżanej w szereg ortogonalny względem wielomianów Czebyszewa i na analogicznym rozwinięciu (4) najprostszych funkcji wymiernych. Wadą metody Horneckera jest konieczność rozwiązywania układu równań liniowych i obliczania wszystkich zer wielomianu, a także wykonywania działań na liczbach zespolonych.

W niniejszej pracy zmodyfikowano metodę Horneckera, usuwając wszystkie wymienione wyżej wady jej pierwotnego wariantu. Schemat obliczeń i ich kontrolę szczegółowo opisano w § 6, w którym obliczono również ilość wykonywanych działań arytmetycznych. W § 7 podano przykład aproksymacji funkcji wykładniczej

---

С. ПАШКОВСКИЙ (Варшава)

РАВНОМЕРНАЯ АППРОКСИМАЦИЯ НЕПРЕРЫВНЫХ ФУНКЦИЙ РАЦИОНАЛЬНЫМИ ФУНКЦИЯМИ

РЕЗЮМЕ

Вычисление непрерывных функций на электронных цифровых машинах обычно осуществляется с помощью многочленов, достаточно хорошо приближающих эти функции в смысле равномерной аппроксимации. Известно, однако, что многочлены являются плохим инструментом приближения функций, которые имеют полюс вблизи сегмента аппроксимации. В таких случаях лучше пользоваться рациональными функциями.

Хорнекер в [2] и [3] предложил алгорифм вычисления рациональной функции, которая — в определенном классе таких же функций — почти наилучшим образом приближает данную непрерывную функцию. Этот метод основан на разложении (1) аппроксимируемой функции в ортогональный ряд по многочленам Чебышева и аналогичном разложении (4) простейших рациональных функций. Недостатки метода Хорнекера заключаются в необходимости решения системы линейных уравнений и нахождения всех корней многочлена, а также выполнения действий над комплексными числами.

В настоящей работе представлено видоизменение метода Хорнекера, устраняющее все перечисленные недостатки его первоначального варианта. Схема счета и его контроль подробно описаны в § 6. Там же подсчитано количество выполняемых арифметических действий. В § 7 дан пример аппроксимации экспоненциальной функции.

S. PASZKOWSKI (Warszawa)

UNIFORM APPROXIMATION OF CONTINUOUS FUNCTIONS BY MEANS OF RATIONAL FUNCTIONS

SUMMARY

The values of continuous functions are usually computed with the aid of electronic computers by means of polynomials, which approximate these functions sufficiently well in the sense of uniform approximation. It is known, however, that polynomials are not a good instrument of approximation of the functions whose pole lies near the interval of approximation. In such cases it is better to make use of rational functions.

Hornecker, in [2] and [3] proposed an algorithm for finding the rational function which approximates almost best a given continuous function, in a definite class of similar functions. This method consists in expanding (1) the function being approximated into an orthogonal series, with regard to Chebyshev polynomials, and in an analogous expansion (4) of the simplest rational functions. A drawback of Hornecker's method is the necessity of solving a system of linear equations and of computing all the zeros of the polynomial, as well as the necessity of performing operations on complex numbers.

In the present article Hornecker's method is modified: all drawbacks mentioned above are eliminated. The scheme of computations and their control are described in detail in § 6, which gives also the number of arithmetic operations to be performed. In § 7 an example of approximation of an exponential function is given.
