Aproximarea funcțiilor - metoda celor mai mici pătrate

Aproximare în L^2

Radu Trîmbițaș

Universitatea "Babeș-Bolyai"

25 martie 2010

Introducere

- Funcțiile de aproximat pot să fie definite pe:
  - un continuu (de regulă un interval) – funcții speciale pe care dorim să le evaluăm ca parte a unei subrutine
  - pe o mulțime finită de puncte – situație întâlnită în științele fizice sau inginerie, când măsurătorile fizice se fac în funcție de alte cantități (cum ar fi timpul)
- Deoarece o astfel de evaluare trebuie să se reducă la un număr finit de operații aritmetice, trebuie în ultimă instanță să aproximăm funcțiile prin intermediul polinoamelor sau funcțiilor raționale.
- Dorim să aproximăm o funcție dată, cât mai bine posibil în termeni de funcții mai simple.

- În general o schemă de aproximare poate fi descrisă prin:
  1. o funcție $f \in X$ ce urmează a fi aproximată;
  2. o clasă Φ de aproximante;
  3. o normă $\|\cdot\|$ ce măsoară mărimea funcțiilor.

- Căutăm o aproximare $\widehat{\varphi} \in \Phi$ a lui f astfel încât
$$\|f - \widehat{\varphi}\| \leq \|f - \varphi\| \text{ pentru orice } \varphi \in \Phi. \tag{1}$$
- Această problemă se numește problemă de cea mai bună aproximare a lui f cu elemente din Φ, iar funcția $\widehat{\varphi}$ se numește cea mai bună aproximare a lui f relativ la norma $\|\cdot\|$.
- Cunoscându-se o bază $\{\pi_j\}_{j=1}^{n}$ a lui Φ putem scrie
$$\Phi = \Phi_n = \left\{ \varphi : \ \varphi(t) = \sum_{j=1}^{n} c_j \pi_j(t), \ c_j \in \mathbb{R} \right\}. \tag{2}$$
- Φ este un spațiu liniar finit dimensional sau o submulțime a acestuia.

Exemple de clase de aproximante I

Exemplu

$\Phi = \mathbb{P}_m$ - *mulțimea polinoamelor de grad cel mult m. O bază a sa este* $e_j(t) = t^j$, $j = 0, 1, \ldots, m$. *Deci* $\dim \mathbb{P}_m = m + 1$. *Polinoamele sunt cele mai utilizate aproximante pentru funcții pe domenii mărginite (intervale sau mulțimi finite de funcții). Motivul – teorema lui Weierstrass – orice funcție din* $C[a, b]$ *poate fi aproximată oricât de bine printr-un polinom de grad suficient de mare.*

Exemple de clase de aproximante II

Exemplu

$\Phi = \mathbb{S}_m^k(\Delta)$ *spațiul funcțiilor spline polinomiale și cu clasa de netezime k pe subdiviziunea*

$$\Delta: \ a = t_1 < t_2 < t_3 < \cdots < t_{N-1} < t_N = b$$

a intervalului $[a, b]$. *Acestea sunt funcții polinomiale pe porțiuni de grad* $\leq m$, *legate în* $t_1, \ldots, t_{N-1}$, *astfel încât toate derivatele până la ordinul k să fie continue. Presupunem* $0 \leq k < m$. *Pentru* $k = m$ *se obține* $\mathbb{P}_m$. *Dacă* $k = -1$ *permitem discontinuități în punctele de joncțiune.*

Exemplu

$\Phi = \mathbb{T}_m[0, 2\pi]$ *spațiul polinoamelor trigonometrice de grad cel mult m pe* $[0, 2\pi]$. *Acestea sunt combinații liniare ale funcțiilor*

$$\begin{aligned} \pi_k(t) &= \cos(k-1)t, \quad k = \overline{1, m+1}, \\ \pi_{m+1-k}(t) &= \sin kt, \quad k = \overline{1, m}. \end{aligned}$$

Dimensiunea spațiului este $n = 2m + 1$. *Astfel de aproximante sunt alegeri naturale dacă funcția de aproximat este periodică de perioadă* 2π. *(Dacă* f *are perioada* p *se face schimbarea de variabilă* $t \to tp/2\pi$.*)*

Exemple de norme și tipuri de aproximare I

Câteva alegeri posibile ale normei, atât pentru aproximări continue, cât și pentru cele discrete apar în tabelul de mai jos

normă continuă	tip	normă discretă
$\lVert u\rVert_\infty = \max\limits_{a\le t\le b} \lvert u(t)\rvert$	L_∞	$\lVert u\rVert_\infty = \max\limits_{1\le i\le N} \lvert u(t_i)\rvert$
$\lVert u\rVert_{1,w} = \int_a^b \lvert u(t)\rvert w(t)dt$	L_w^1	$\lVert u\rVert_{1,w} = \sum_{i=1}^{n} w_i \lvert u(t_i)\rvert$
$\lVert u\rVert_{2,w} = \left(\int_a^b \lvert u(t)\rvert^2 w(t)dt\right)^{1/2}$	L_w^2	$\lVert u\rVert_{2,w} = \left(\sum_{i=1}^{N} w_i \lvert u(t_i)\rvert^2\right)^{1/2}$

- Cazul continuu presupune un interval $[a,b]$ și o funcție pondere $w(t)$ (posibil și $w(t)\equiv 1$) definită pe intervalul $[a,b]$ și pozitivă, exceptând zerourile izolate. Intervalul $[a,b]$ poate fi nemărginit, dacă funcția pondere w este astfel încât integrala pe $[a,b]$ care definește norma să aibă sens. Funcția dată f și funcția φ din clasa Φ trebuie definite pe $[a,b]$ și norma $\lVert f-\varphi\rVert$ să aibă sens.

Exemple de norme și tipuri de aproximare II

- Cazul discret presupune o mulțime de N puncte distincte $t_1, t_2, \dots, t_N$ împreună cu ponderile $w_1, w_2, \dots, w_N$ (posibil $w_i = 1$, $i = \overline{1, N}$). f și φ trebuie definite în punctele t_i în cazul discret.
- Deci combinând normele din tabelă cu spațiile liniare din exemple se obține o problemă de cea mai bună aproximare (1) cu sens.
- Dacă cea mai bună aproximantă $\widehat{\varphi}$ în cazul discret este astfel încât $\|f - \widehat{\varphi}\| = 0$, atunci $\widehat{\varphi}(t_i) = f(t_i)$, pentru $i = 1, 2, \dots, N$, spunem că $\widehat{\varphi}$ interpolează f în punctele t_i și numim această problemă de aproximare problemă de interpolare.
- Cele mai simple probleme de aproximare sunt problema celor mai mici pătrate și problema de interpolare.

Instrument notațional I

- Definim în cazul continuu

$$\lambda(t) = \begin{cases} 0, & \text{dacă} \quad t < a \ (\text{când} \ -\infty < a), \\ \int_a^t w(\tau)d\tau, & \text{dacă} \quad a \leq t \leq b, \\ \int_a^b w(\tau)d\tau, & \text{dacă} \quad t > b \ (\text{când } b < \infty). \end{cases} \tag{3}$$

- putem scrie

$$\int_{\mathbb{R}} u(t)d\lambda(t) = \int_a^b u(t)w(t)dt, \qquad \forall u \in C[a,b]$$

deoarece

$$d\lambda(t) = \begin{cases} w(t)dt, & t \in (a,b); \\ 0, & t \notin (a,b). \end{cases}$$

Instrument notațional II

- $d\lambda$ se numește măsură (pozitivă) continuă
- măsura discretă (numită și ,,măsura Dirac“) asociată mulțimii de puncte $\{t_1, t_2, \dots, t_N\}$ este o măsură $d\lambda$ care este nenulă numai în punctele t_i și are aici valoarea w_i. Astfel în acest caz

$$\int_{\mathbb{R}} u(t) d\lambda(t) = \sum_{i=1}^{N} w_i u(t_i). \tag{4}$$

(definim $\lambda(t)$ ca fiind o funcție în scară cu saltul în t_i egal cu w_i)

- unificare cu integrală Stieltjes: definim norma lui L_2 prin

$$\|u\|_{2,d_\lambda} = \left(\int_{\mathbb{R}} |u(t)|^2 d\lambda(t) \right)^{\frac{1}{2}} \tag{5}$$

și obținem norma continuă sau discretă după cum λ este ca în (3) sau o funcție în scară ca în (4).

Instrument notațional III

- Vom numi suportul lui $d\lambda$ – notat cu $\operatorname{supp} d\lambda$ – intervalul $[a, b]$ în cazul continuu (presupunem că w este pozitivă pe $[a, b]$ exceptând zerourile izolate) și mulțimea $\{t_1, t_2, \ldots, t_N\}$ în cazul discret.
- Spunem că mulțimea de funcții π_j din (2) este liniar independentă pe $\operatorname{supp} d\lambda$ dacă

$$\forall\ t\ \in \operatorname{supp} d\lambda \quad \sum_{j=1}^{n} c_j \pi_j(t) \equiv 0 \Rightarrow c_1 = c_2 = \cdots = c_k = 0 \tag{6}$$

Vom specializa problema (1) luând ca normă norma din L_2

$$\|u\|_{2,d_\lambda} = \left(\int_{\mathbb{R}} |u(t)|^2 d\lambda(t) \right)^{\frac{1}{2}}, \tag{7}$$

unde $d\lambda$ este fie o măsură continuă (conform (3)) sau discretă (conform (4)) și utilizând aproximanta φ dintr-un spațiu liniar n-dimensional

$$\Phi = \Phi_n = \left\{ \varphi : \ \varphi(t) = \sum_{j=1}^{n} c_j \pi_j(t), \ c_j \in \mathbb{R} \right\}. \tag{8}$$

π_j liniar independente pe $\operatorname{supp} d\lambda$; integrala din (7) are sens pentru $u = \pi_j$, $j = 1, \dots, n$ și $u = f$. Problema astfel obținută se numește problemă de aproximare în sensul celor mai mici pătrate sau problemă de aproximare în medie pătratică.

Produse scalare

- Definim produsul scalar prin

$$(u, v) = \int_{\mathbb{R}} u(t)v(t)d\lambda(t). \tag{9}$$

- Definiție corectă, conform inegalității Cauchy-Buniakovski-Schwarz

$$\|(u, v)\| \leq \|u\|_{2,d_\lambda} \|v\|_{2,d_\lambda}$$

- Proprietăți
  (i) simetria $(u, v) = (v, u)$;
  (ii) liniaritatea $(\alpha_1 u_1 + \alpha_2 u_2, v) = \alpha_1 (u_1, v) + \alpha_2 (u_2, v)$;
  (iii) pozitiv definirea $(u, u) \geq 0$ și $(u, u) = 0 \Leftrightarrow u \equiv 0$ pe $\mathrm{supp} d\lambda$.
- De asemenea

$$\|u\|^2_{2,d_\lambda} = (u, u). \tag{10}$$

Ortogonalitate

- Spunem că u și v sunt ortogonale dacă

$$(u, v) = 0. \tag{11}$$

- Mai general, putem considera sisteme ortogonale $\{u_k\}_{k=1}^n$:

$$(u_i, u_j) = 0 \text{ dacă } i \neq j,\ \ u_k \neq 0 \text{ pe } \mathrm{supp}\, d\lambda;\ \ i, j = \overline{1, n},\ \ k = \overline{1, n}. \tag{12}$$

- Pentru un astfel de sistem are loc teorema generalizată a lui Pitagora

$$\left\| \sum_{k=1}^n \alpha_k u_k \right\|^2 = \sum_{k=1}^n |\alpha_k|^2 \|u_k\|^2. \tag{13}$$

- (13) $\Rightarrow$ orice sistem ortogonal este liniar independent pe $\mathrm{supp}\, d\lambda$. Într-adevăr, dacă membrul stâng al lui (13) se anulează, atunci și membrul drept se anulează și deoarece $\|u_k\|^2 > 0$, din ipoteză rezultă $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$.

- Putem scrie pătratul erorii din L_2 sub forma

$$E^2[\varphi] := \|\varphi - f\|^2 = (\varphi - f, \varphi - f) = (\varphi, \varphi) - 2(\varphi, f) + (f, f).$$

- Înlocuind pe φ cu expresia sa se obține

$$E^2[\varphi] = \int_{\mathbb{R}} \left(\sum_{j=1}^{n} c_j \pi_j(t) \right)^2 d\lambda(t) - 2 \int_{\mathbb{R}} \left(\sum_{j=1}^{n} c_j \pi_j(t) \right) f(t) d\lambda(t) + \tag{14}$$

$$+ \int_{\mathbb{R}} f^2(t) d\lambda(t).$$

- Pătratul erorii din L_2 este o funcție cuadratică de coeficienții $c_1, \ldots,$ c_n ai lui φ. Problema celei mai bune aproximații în L_2 revine la a minimiza această funcție pătratică; ea se rezolvă anulând derivatele parțiale.

Ecuațiile normale II

- Se obține

$$\frac{\partial}{\partial c_i} E^2[\varphi] = 2 \int_{\mathbb{R}} \left(\sum_{j=1}^{n} c_j \pi_j(t) \right) \pi_i(t) d\lambda(t) - 2 \int_{\mathbb{R}} \pi_i(t) f(t) d\lambda(t) = 0$$

  adică

$$\sum_{j=1}^{n} (\pi_i, \pi_j) c_j = (\pi_i, f), \quad i = 1, 2, \dots, n. \tag{15}$$

- Aceste ecuații se numesc ecuații normale pentru problema celor mai mici pătrate.
- Ele formează un sistem de forma

$$Ac = b \tag{16}$$

  unde matricea A și vectorul b au elementele

$$A = [a_{ij}], \quad a_{ij} = (\pi_i, \pi_j), \quad b = [b_i], \quad b_i = (\pi_i, f). \tag{17}$$

Existența și unicitatea soluției I

- Datorită simetriei produsului scalar, A este o matrice simetrică. Mai mult, A este pozitiv definită, adică

$$x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j > 0 \text{ dacă } x \neq [0, 0, \ldots, 0]^T. \tag{18}$$

- Funcția (18) se numește formă pătratică (deoarece este omogenă de grad 2). Pozitiv definirea lui A ne spune că forma pătratică ai cărei coeficienți sunt elementele lui A este întotdeauna nenegativă și zero numai dacă variabilele x_i se anulează.
- Pentru a demonstra (18) să inserăm definiția lui a_{ij} și să utilizăm proprietățile (i)-(iii) ale produsului scalar

$$x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j (\pi_i, \pi_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i \pi_i, x_j \pi_j) = \left\| \sum_{i=1}^{n} x_i \pi_i \right\|^2 .$$

Existența și unicitatea soluției II

- Aceasta este evident nenegativă. Ea este zero numai dacă $\sum_{i=1}^{n} x_i \pi_i \equiv 0$ pe $\operatorname{supp} d\lambda$, care pe baza liniar independenței lui π_i implică $x_1 = x_2 = \cdots = x_n = 0$.
- Este un rezultat cunoscut din algebra liniară că *o matrice A simetrică pozitiv definită este nesingulară*. Într-adevăr, determinantul său, precum și minorii principali sunt strict pozitivi. Rezultă că sistemul de ecuații normale (15) are soluție unică.
- Corespunde această soluție minimului lui $E[\varphi]$ în (14)? Matricea hessiană $H = \left[\partial^2 E^2 / \partial c_i \partial c_j\right]$ trebuie să fie *pozitiv definită*. Dar $H = 2A$, deoarece E^2 este o funcție cuadratică. De aceea, H, ca și A, este într-adevăr pozitiv definită și soluția ecuațiilor normale ne dă minimul dorit.

- Problema de aproximare în sensul celor mai mici pătrate are o soluție unică, dată de

$$\widehat{\varphi}(t) = \sum_{j=1}^{n} \widehat{c}_j \pi_j(t) \tag{19}$$

unde $\widehat{c} = \left[\widehat{c}_1, \widehat{c}_2, \ldots, \widehat{c}_n\right]^T$ este vectorul soluție al ecuațiilor normale (15).

Exemplu I

Exemplu

Dându-se punctele

$$(0,-4),(1,0),(2,4),(3,-2),$$

determinați polinomul de gradul I corespunzător acestor date prin metoda celor mai mici pătrate.

Soluție. Aproximanta căutată are forma

$$\varphi(x)=c_0+c_1x.$$

Sistemul de ecuații normale se determină din condițiile $f-\varphi\perp 1$ și $f-\varphi\perp x$. Se obține

$$\begin{cases} c_0(1,1)+c_1(x,1)=(f,1) \\ c_0(1,x)+c_1(x,x)=(f,x) \end{cases}$$

Exemplu II

Dar, $(1,1)=\sum_{i=1}^{4} 1\cdot 1=3$,
$(1,x)=(x,1)=\sum_{i=1}^{4} 1\cdot x_i = 1\cdot 0+1\cdot 1+1\cdot 2+1\cdot 3=6$,
$(x,x)=\sum_{i=1}^{4} x_i^2=14$. Pentru membrul drept avem $(f,1)=(y,1)=-2$ și $(f,x)=(y,x)=2$. Am obținut sistemul

$$\begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} -2 \\ 2 \end{bmatrix}$$

cu soluția $c_0=-2$, $c_1=1$. Deci $\varphi(x)=x-2$. ■

Neajunsuri ale MCMMP I

- Ecuațiile normale rezolvă problema de aproximare în sensul celor mai mici pătrate complet în teorie. Dar în practică?
- Referitor la o mulțime generală de funcții de bază liniar independente, pot apărea următoarele dificultăți:
  1. Sistemul de ecuații normale (15) poate fi prost condiționat. Un exemplu simplu este următorul: $\operatorname{supp} d\lambda = [0,1]$, $d\lambda(t) = dt$ pe $[0,1]$ și $\pi_j(t) = t^{j-1}$, $j = 1, 2, \ldots, n$. Atunci

     $$(\pi_i, \pi_j) = \int_0^1 t^{i+j-2} dt = \frac{1}{i+j-1}, \quad i, j = 1, 2, \ldots, n,$$

     adică matricea A este matricea Hilbert. Prost condiționarea ecuațiilor normale se datorează alegerii neinspirate a funcțiilor de bază. Acestea devin aproape liniar dependente când exponentul crește. O altă sursă de degradare provine din elementele membrului drept $b_j = \int_0^1 \pi_j(t) f(t) dt$. Când j este mare $\pi_j(t) = t^{j-1}$ se comportă pe

$[0,1]$ ca o funcție discontinuă. Un polinom care oscilează mai rapid pe $[0,1]$ ar fi de preferat, căci ar angaja mai viguros funcția f.

2. Al doilea dezavantaj este faptul că toți coeficienții $\widehat{c}_j$ din (19) depind de n, adică $\widehat{c}_j = \widehat{c}_j^{(n)}$, $j = 1, 2, \ldots, n$. Mărirea lui n ne dă un nou sistem de ecuații mai mare și cu o soluție complet diferită. Acest fenomen se numește nepermanența coeficienților $\widehat{c}_j$.

- Amândouă neajunsurile (1) și (2) pot fi eliminate (sau măcar atenuate) alegând ca funcții de bază un sistem ortogonal,

$$(\pi_i, \pi_j) = 0 \text{ dacă } i \neq j \quad (\pi_j, \pi_j) = \|\pi_j\|^2 > 0 \tag{20}$$

- Atunci sistemul de ecuații normale devine diagonal și poate fi rezolvat imediat cu formula

$$\widehat{c}_j = \frac{(\pi_j, f)}{(\pi_j, \pi_j)}, \quad j = 1, 2, \ldots, n. \tag{21}$$

Evident, acești coeficienți $\widehat{c}_j$ sunt independenți de n și odată calculați rămân la fel pentru orice n mai mare. Avem acum proprietatea de permanență a coeficienților. De asemenea nu trebuie să rezolvăm sistemul de ecuații normale, ci putem aplica direct (21).

Neajunsuri ale MCMMP IV

- Orice sistem $\{\widehat{\pi}_j\}$ care este liniar independent pe $\operatorname{supp} d\lambda$ poate fi ortogonalizat (în raport cu măsura $d\lambda$) prin procedeul Gram-Schmidt. Se ia

$$\pi_1 = \widehat{\pi}_1$$

și apoi, pentru $j = 2, 3, \dots$ se calculează recursiv

$$\pi_j = \widehat{\pi}_j - \sum_{k=1}^{j-1} c_k \pi_k, \quad c_k = \frac{(\widehat{\pi}_j, \pi_k)}{(\pi_k, \pi_k)}, \quad k = \overline{1, j-1}.$$

Atunci fiecare π_j astfel determinat este ortogonal pe toate funcțiile precedente.

Eroarea în MCMMP I

- Am văzut că dacă $\Phi = \Phi_n$ constă din n funcții π_j, $j = 1, 2, \ldots, n$ care sunt liniar independente pe $\mathrm{supp} d\lambda$, atunci problema de aproximare în sensul celor mai mici pătrate pentru această măsură

$$\min_{\varphi \in \phi_n} \|f - \varphi\|_{2,d_\lambda} = \|f - \widehat{\varphi}\|_{2,d_\lambda} \tag{22}$$

are o soluție unică $\widehat{\varphi} = \widehat{\varphi}_n$, dată de (19).

- Există multe moduri de a selecta baza $\{\pi_j\}$ a lui Φ_n și de aceea mai multe moduri de a reprezenta soluția, care conduc totuși la aceeași funcție. Eroarea în sensul celor mai mici pătrate – cantitatea din dreapta relației (22) – este independentă de alegerea funcțiilor de bază (deși calculul soluției, așa cum s-a menționat anterior, nu este).
- În studiul acestor erori, putem presupune fără a restrânge generalitatea că baza π_j este un sistem ortogonal (fiecare sistem liniar independent poate fi ortogonalizat prin procedeul Gram-Schmidt).

- Avem conform (21)

$$\widehat{\varphi}_n(t) = \sum_{j=1}^{n} \widehat{c}_j \pi_j(t), \quad \widehat{c}_j = \frac{(\pi_j, f)}{(\pi_j, \pi_j)}. \tag{23}$$

- Observăm întâi că eroarea $f - \varphi_n$ este ortogonală pe Φ_n, adică

$$(f - \widehat{\varphi_n}, \varphi) = 0, \ \forall \ \varphi \in \Phi_n \tag{24}$$

  unde produsul scalar este cel din (9).
- Deoarece φ este o combinație liniară de π_k, este suficient să arătăm (24) pentru fiecare $\varphi = \pi_k$, $k = 1, 2, \ldots, n$.

Eroarea în MCMMP III

- Înlocuind φ_n cu expresia sa din (23) în (24), găsim

$$(f - \widehat{\varphi}_n, \pi_k) = \left(f - \sum_{j=1}^{n} \widehat{c}_j \pi_k, \pi_k \right) = (f, \pi_k) - \widehat{c}_k (\pi_k, \pi_k) = 0,$$

ultima ecuație rezultând din formula pentru $\widehat{c}_k$ din (23).

- Rezultatul din (24) are o interpretare geometrică simplă. Dacă reprezentăm funcțiile ca vectori și spațiul Φ_n ca un plan, atunci pentru orice funcție f care înțeapă planul Φ_n, aproximanta în sensul celor mai mici pătrate $\widehat{\varphi}_n$ este proiecția ortogonală a lui f pe Φ_n, vezi figura 1.

Eroarea în MCMMP IV

- În particular, alegând $\varphi = \widehat{\varphi}_n$ în (24) obținem

$$(f - \widehat{\varphi}_n, \widehat{\varphi}_n) = 0$$

și de aceea, deoarece $f = (f - \widehat{\varphi}) + \widehat{\varphi}$, conform teoremei lui Pitagora și generalizării sale (13)

$$\begin{aligned}
\|f\|^2 &= \|f - \widehat{\varphi}\|^2 + \|\widehat{\varphi}\|^2 \\
&= \|f - \widehat{\varphi}_n\|^2 + \left\| \sum_{j=1}^{n} \widehat{c}_j \pi_j \right\|^2 \\
&= \|f - \widehat{\varphi}_n\|^2 + \sum_{j=1}^{n} |\widehat{c}_j|^2 \|\pi_j\|^2 .
\end{aligned}$$

Eroarea în MCMMP V

- Exprimând primul termen din dreapta obținem

$$\|f-\widehat{\varphi}_n\| = \left\{ \|f\|^2 - \sum_{j=1}^{n} |\widehat{c}_j| \|\pi_j\|^2 \right\}^{1/2}, \quad \widehat{c}_j = \frac{(\pi_j, f)}{(\pi_j, \pi_j)}. \tag{25}$$

- De notat că expresia dintre acolade trebuie să fie nenegativă.

Eroarea în MCMMP VI

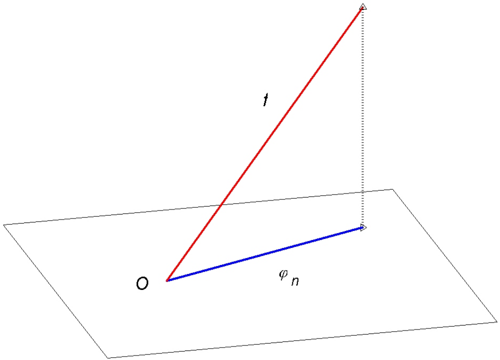


Figure: Aproximația în sensul celor mai mici pătrate ca proiecție ortogonală

- Dacă se dă acum o secvență de spații liniare Φ_n, $n = 1, 2, 3, \ldots$, avem evident

$$\|f - \widehat{\varphi}_1\| \geq \|f - \widehat{\varphi}_2\| \geq \|f - \widehat{\varphi}_3\| \geq \ldots,$$

  care rezultă nu numai din (25), dar mai direct din faptul că

$$\Phi_1 \subset \Phi_2 \subset \Phi_3 \subset \ldots.$$

- Deoarece există o infinitate de astfel de spații, atunci secvența de erori din L_2, fiind monoton descrescătoare, trebuie să conveargă la o limită. Este limita 0?
- Dacă este așa, spunem că aproximarea prin metoda celor mai mici pătrate converge în medie pătratică când $n \to \infty$.

Convergența II

- Este evident din (25) că o condiție necesară și suficientă pentru aceasta este
$$\sum_{j=1}^{\infty} |\widehat{c}_j|^2 \|\pi_j\|^2 = \|f\|^2. \tag{26}$$
- Un mod echivalent de a formula convergența este următorul: dându-se f cu $\|f\| < \infty$, adică $\forall\ f \in L_{2,d\lambda}$ și dându-se un $\varepsilon > 0$ arbitrar de mic, există un întreg $n = n_\varepsilon$ și o funcție $\varphi^* \in \Phi_n$ astfel încât $\|f - \varphi^*\| \leq \varepsilon$.
- O clasă de funcții având această proprietate se numește completă în raport cu norma $\|\cdot\| = \|\cdot\|_{2,d\lambda}$. Vom numi relația (26) relația de completitudine sau relația Parseval-Liapunov.

Carl Friedrich Gauss 1777-1855
Matematică, astronomie, geodezie, magnetism
(Ca adolescent în Braunschweig a descoperit teorema binomială, reciprocitatea pătratică, media aritmetico-geometrică. . .)
1807-1855: Universitatea din Göttingen

Adrien Marie Legendre (1752-1833) matematician francez, analiză (integrale eliptice), teoria numerelor, geometrie. Descoperitor, alături de Gauss, (în 1805) al metodei celor mai mici pătrate, deși Gauss a utilizat metoda încă din 1794, dar a publicat-o doar în 1809.

Exemple de sisteme ortogonale

1. Sistemul trigonometric – cunoscut din analiza Fourier.
2. Polinoame ortogonale

- *Sistemul trigonometric* este format din funcțiile:

$$1, \cos t, \cos 2t, \cos 3t, \dots, \sin t, \sin 2t, \sin 3t, \dots$$

- El este ortogonal pe $[0, 2\pi]$ în raport cu măsura

$$d\lambda(t) = \begin{cases} dt & \text{pe } [0, 2\pi] \\ 0 & \text{în rest} \end{cases}$$

$$\int_0^{2\pi} \sin kt \sin \ell t dt = \begin{cases} 0, & \text{pentru} \quad k \neq \ell \\ \pi, & \text{pentru} \quad k = \ell \end{cases} \qquad k, \ell = 1, 2, 3, \dots$$

$$\int_0^{2\pi} \cos kt \cos \ell t dt = \begin{cases} 0, & k \neq \ell \\ 2\pi, & k = \ell = 0 \\ \pi, & k = \ell > 0 \end{cases} \qquad k, \ell = 0, 1, 2$$

$$\int_0^{2\pi} \sin kt \cos \ell t dt = 0, \quad k = 1, 2, 3, \dots, \quad \ell = 0, 1, 2, \dots$$

- Aproximarea are forma

$$f(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kt + b_k \sin kt). \tag{27}$$

- Utilizând (21) obținem

$$\begin{aligned} a_k &= \frac{1}{\pi} \int_0^{2\pi} f(t) \cos kt dt, \quad k = 1, 2, \ldots \\ b_k &= \frac{1}{\pi} \int_0^{2\pi} f(t) \sin kt dt, \quad k = 1, 2, \ldots \end{aligned} \tag{28}$$

  numiți coeficienți Fourier ai lui f. Ei sunt coeficienții (21) pentru sistemul trigonometric.
- Prin extensie coeficienții (21) pentru orice sistem ortogonal (π_j) se vor numi coeficienții Fourier ai lui f relativ la acest sistem.

Sistemul trigonometric III

- În particular, recunoaștem în seria Fourier trunchiată pentru $k = n$ aproximarea lui f în clasa polinoamelor trigonometrice de grad $\leq n$ relativ la norma

$$\|u\|_2 = \left(\int_0^{2\pi} |u(t)|^2 dt \right)^{1/2}$$

Polinoame ortogonale I

- Dându-se o măsură $d\lambda$, știm că orice număr finit de puteri $1, t, t^2, \ldots$ sunt liniar independente pe $[a, b]$, dacă $\mathrm{supp} d\lambda = [a, b]$, iar $1, t, \ldots, t^{N-1}$ liniar independente pe $\mathrm{supp} d\lambda = \{t_1, t_2, \ldots, t_N\}$.
- Deoarece o mulțime de vectori liniar independenți a unui spațiu liniar poate fi ortogonalizată prin procedeul Gram-Schmidt, orice măsură $d\lambda$ de tipul considerat generează o mulțime unică de polinoame ortogonale monice $\pi_j(t, d\lambda)$, $j = 0, 1, 2, \ldots$ ce satisfac

$$\mathrm{grad} \pi_j = j, \quad j = 0, 1, 2, \ldots$$

$$\int_{\mathbb{R}} \pi_k(t)\pi_\ell(t) d\lambda(t) = 0, \text{ dacă } k \neq \ell \tag{29}$$

- Aceste polinoame se numesc polinoame ortogonale relativ la măsura $d\lambda$.

Polinoame ortogonale II

- Vom permite indicilor să meargă de la 0. Mulțimea π_j este infinită dacă $\operatorname{supp} d\lambda = [a, b]$ și constă din exact N polinoame $\pi_0, \pi_1, \dots, \pi_{N-1}$ dacă $\operatorname{supp} d\lambda = \{t_1, \dots, t_N\}$. În ultimul caz polinoamele se numesc polinoame ortogonale discrete.
- Între trei polinoame ortogonale monice (un polinom se numește monic dacă coeficientul său dominant este 1) consecutive există o relație liniară. Mai exact, există constantele reale $\alpha_k = \alpha_k(d\lambda)$ și $\beta_k = \beta_k(d\lambda) > 0$ (depinzând de măsura $d\lambda$) astfel încât

$$\pi_{k+1}(t) = (t - \alpha_k)\pi_k(t) - \beta_k \pi_{k-1}(t), \quad k = 0, 1, 2, \dots \tag{30}$$

$$\pi_{-1}(t) = 0, \quad \pi_0(t) = 1.$$

(Se subînțelege că (30) are loc pentru orice $k \in \mathbb{N}$ dacă $\operatorname{supp} d\lambda = [a, b]$ și numai pentru $k = \overline{0, N-2}$ dacă $\operatorname{supp} d\lambda = \{t_1, t_2, \dots, t_N\}$).

- Pentru a demonstra (30) și a obține expresiile coeficienților să observăm că $\pi_{k+1}(t) - t\pi_k(t)$ este un polinom de grad $\leq k$, și deci poate fi exprimat ca o combinație liniară a lui $\pi_0, \pi_1, \ldots, \pi_k$. Scriem această combinație sub forma

$$\pi_{k+1} - t\pi_k(t) = -\alpha_k \pi_k(t) - \beta_k \pi_{k-1}(t) + \sum_{j=0}^{k-2} \gamma_{k,j} \pi_j(t) \qquad (31)$$

(sumele vide se consideră nule).

- Înmulțim scalar ambii membri ai relației anterioare cu π_k și obținem

$$(-t\pi_k, \pi_k) = -\alpha_k(\pi_k, \pi_k)$$

adică

$$\alpha_k = \frac{(t\pi_k, \pi_k)}{(\pi_k, \pi_k)}, \quad k = 0, 1, 2, \ldots \qquad (32)$$

- La fel, înmulțind scalar cu π_{k-1} obținem

$$(-t\pi_k, \pi_{k-1}) = -\beta_k(\pi_{k-1}, \pi_{k-1}).$$

Deoarece $(t\pi_k, \pi_{k-1}) = (\pi_k, t\pi_{k-1})$ și $t\pi_{k-1}$ diferă de π_k printr-un polinom de grad $< k$ se obține prin ortogonalitate $(t\pi_k, \pi_{k-1}) = (\pi_k, \pi_k)$, deci

$$\beta_k = \frac{(\pi_k, \pi_k)}{(\pi_{k-1}, \pi_{k-1})}, \quad k = 1, 2, \ldots \tag{33}$$

- Înmulțind (31) cu π_ℓ, $\ell < k-1$, se obține

$$\gamma_{k,\ell} = 0, \quad \ell = 0, 1, \ldots, k-1 \tag{34}$$

- Formula de recurență (30) ne dă o modalitate practică de determinare a polinoamelor ortogonale. Deoarece $\pi_0 = 1$, putem calcula α_0 cu (32) pentru $k = 0$ și apoi π_1, etc. Procedeul – numit procedura lui Stieltjes – este foarte potrivit pentru polinoame ortogonale discrete, căci în acest caz produsul scalar se exprimă prin sume finite.
- În cazul continuu, calculul produsului scalar necesită calcul de integrale, ceea ce complică lucrurile. Din fericire, pentru multe măsuri speciale importante, coeficienții se cunosc explicit.
- Cazul special când măsura este simetrică (adică $d\lambda(t) = w(t)$ cu $w(-t) = w(t)$ și $\operatorname{supp} d\lambda$ simetrică față de origine) merită o atenție specială, deoarece în acest caz $\alpha_k = 0, \ \forall\ k \in \mathbb{N}$, conform lui (27) căci

$$(t\pi_k, \pi_k) = \int_{\mathbb{R}} w(t) t\pi_k^2(t) dt = \int_a^b w(t) t\pi_k^2(t) dt = 0,$$

deoarece avem o integrală dintr-o funcție impară pe un domeniu simetric.

Figure: Thomas Ioannes Stieltjes (1856-1894)

- Se definesc prin așa-numita formulă a lui Rodrigues

$$\pi_k(t) = \frac{k!}{(2k)!}\frac{d^k}{dt^k}(t^2-1)^k. \tag{35}$$

- Verificăm întâi ortogonalitatea pe $[-1,1]$ în raport cu măsura $d\lambda(t) = dt$. Pentru orice $0 \le \ell < k$, prin integrare repetată prin părți se obține:

$$\int_{-1}^{1}\frac{d^k}{dt^k}(t^2-1)^k$$

$$= \sum_{m=0}^{\ell}\ell(\ell-1)\dots(\ell-m+1)t^{\ell-m}\frac{d^{k-m-1}}{dt^{k-m-1}}(t^2-1)^k\bigg|_{-1}^{1} = 0,$$

ultima relație având loc deoarece $0 \le k-m-1 < k$.

Polinoamele lui Legendre II

- Deci,
$$(\pi_k, p) = 0, \qquad \forall p \in \mathbb{P}_{k-1},$$
demonstrându-se astfel ortogonalitatea.
- **Relația de recurență**
- Datorită simetriei, putem scrie
$$\pi_k(t) = t^k + \mu_k t^{k-2} + \ldots, \quad k \geq 2$$
și observând (din nou datorită simetriei) că relația de recurență are forma
$$\pi_{k+1}(t) = t\pi_k(t) - \beta_k \pi_{k-1}(t),$$
obținem
$$\beta_k = \frac{t\pi_k(t) - \pi_{k+1}(t)}{\pi_{k-1}(t)},$$
care este valabilă pentru orice t.

- Făcând $t \to \infty$,

$$\beta_k = \lim_{t\to\infty} \frac{t\pi_k(t) - \pi_{k+1}(t)}{\pi_{k-1}(t)} = \lim_{t\to\infty} \frac{(\mu_k - \mu_{k+1})t^{k-1} + \dots}{t^{k-1} + \dots} = \mu_k - \mu_{k+1}.$$

(Dacă $k = 1$, punem $\mu_1 = 0$.)

- Din formula lui Rodrigues rezultă

$$\pi_k(t) = \frac{k!}{(2k)!}\frac{d^k}{dt^k}\left(t^{2k} - kt^{2k-2} + \dots\right)$$

$$= \frac{k!}{(2k)!}(2k(2k-1)\dots(k+1)t^k -$$

$$k(2k-2)(2k-3)\dots(k-1)t^{k-1} + \dots)$$

$$= t^k - \frac{k(k-1)}{2(2k-1)}t^{k-2} + \dots,$$

așa că

$$\mu_k = \frac{k(k-1)}{2(2k-1)}, \quad k \geq 2.$$

Deci,

$$\beta_k = \mu_k - \mu_{k+1} = \frac{k^2}{(2k-1)(2k+1)}$$

și deoarece $\mu_1 = 0$,

$$\beta_k = \frac{1}{4 - k^{-2}}, \quad k \geq 1. \tag{36}$$

- Ele se pot defini prin relația

$$T_n(x) = \cos(n \arccos x), \quad n \in \mathbb{N}. \tag{37}$$

- Din identitatea trigonometrică

$$\cos(k+1)\theta + \cos(k-1)\theta = 2\cos\theta\cos k\theta$$

și din (37), punând $\theta = \arccos x$ se obține

$$\begin{aligned} T_{k+1}(x) &= 2xT_k(x) - T_{k-1}(x) \quad k = 1, 2, 3, \ldots \\ T_0(x) &= 1, \quad T_1(x) = x. \end{aligned} \tag{38}$$

- De exemplu,

$$T_2(x) = 2x^2 - 1,$$
$$T_3(x) = 4x^3 - 3x,$$
$$T_4(x) = 8x^4 - 8x^2 + 1$$

  ș.a.m.d.

- Din relația (38) se obține pentru coeficientul dominant al lui T_n valoarea 2^{n-1} (dacă $n \geq 1$), deci polinomul Cebîșev de speța I monic este

$$\overset{\circ}{T}_n(x) = \frac{1}{2^{n-1}} T_n(x), \quad n \geq 0, \quad \overset{\circ}{T}_0 = T_0. \tag{39}$$

- Din (37) se pot obține rădăcinile lui T_n

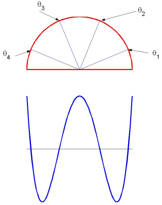
$$x_k^{(n)} = \cos \theta_k^{(n)}, \quad \theta_k^{(n)} = \frac{2k-1}{2n}\pi, \quad k = \overline{1,n}. \tag{40}$$

Polinoamele Cebîșev de speța I III

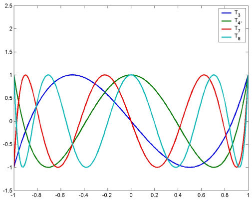
- Ele sunt proiecțiile pe axa reală ale punctelor de pe cercul unitate de argument $\theta_k^{(n)}$.
- Pe intervalul $[-1, 1]$ T_n oscilează de la $+1$ la -1, atingând aceste valori extreme în punctele

$$y_k^{(n)} = \cos \eta_k^{(n)}, \quad \eta_k^{(n)} = \frac{k\pi}{n}, \quad k = \overline{0, n}.$$

T_4 și rădăcinile sale



T_3, T_4, T_7, T_8 pe [-1,1]



- Polinoamele Cebîșev de speța I sunt ortogonale în raport cu măsura

$$d\lambda(x) = \frac{dx}{\sqrt{1-x^2}}, \quad \text{pe } [-1,1].$$

- Se verifică ușor din (37) că

$$\begin{aligned}
&\int_{-1}^{1} T_k(x) T_\ell(x) \frac{dx}{\sqrt{1-x^2}} = \int_0^\pi T_k(\cos\theta) T_\ell(\cos\theta) d\theta \\
&= \int_0^\pi \cos k\theta \cos \ell\theta d\theta = \begin{cases} 0 & \text{dacă} \quad k \neq \ell \\ \pi & \text{dacă} \quad k = \ell = 0 \\ \pi/2 & \text{dacă} \quad k = \ell \neq 0 \end{cases}
\end{aligned} \tag{41}$$

- Dezvoltarea în serie Fourier de polinoame Cebîșev este dată de

$$f(x) = \sum_{j=0}^{\infty}{}' c_j T_j(x) = \frac{1}{2}c_0 + \sum_{j=1}^{\infty} c_j T_j(x), \tag{42}$$

unde

$$c_j = \frac{2}{\pi}\int_{-1}^{1} f(x) T_j(x) \frac{dx}{\sqrt{1-x^2}}, \quad j \in \mathbb{N}.$$

- Păstrând în (42) numai termenii de grad cel mult n se obține o aproximare polinomială utilă de grad n

$$\tau_n(x) = \sum_{j=0}^{n}{}' c_j T_j(x), \tag{43}$$

având eroarea

$$f(x) - \tau_n(x) = \sum_{j=n+1}^{\infty} c_j T_j(x) \approx c_{n+1} T_{n+1}(x). \tag{44}$$

- Aproximanta din (43) este cu atât mai bună cu cât coeficienții din extremitatea dreaptă tind mai repede către zero. Eroarea (44) oscilează în esență între $+c_{n+1}$ și $-c_{n+1}$ și este deci de mărime ,,uniformă“. Acest lucru contrastează puternic cu dezvoltarea Taylor în jurul lui $x = 0$, unde polinomul de grad n are eroarea proporțională cu x^{n+1} pe $[-1, 1]$.

Teoremă

Pentru orice polinom monic $\overset{\circ}{p}_n$ de grad n are loc

$$\max_{-1\leq x\leq 1}\left|\overset{\circ}{p}_n(x)\right| \geq \max_{-1\leq x\leq 1}\left|\overset{\circ}{T}_n(x)\right| = \frac{1}{2^{n-1}}, \quad n\geq 1, \tag{45}$$

unde $\overset{\circ}{T}_n(x)$ este dat de (39).

Demonstrație. Se face prin reducere la absurd. Presupunem că

$$\max_{-1\leq x\leq 1}\left|\overset{\circ}{p}_n(x)\right| < \frac{1}{2^{n-1}}. \tag{46}$$

Atunci polinomul $d_n(x) = \overset{\circ}{T}_n(x) - \overset{\circ}{p}_n(x)$ (de grad $\leq n-1$) satisface

$$d_n\left(y_0^{(n)}\right) > 0,\ d_n\left(y_1^{(n)}\right) < 0,\ d_n\left(y_2^{(n)}\right) > 0, \ldots, (-1)^n d_n\left(y_n^{(n)}\right) > 0. \tag{47}$$

Deoarece d_n are n schimbări de semn, el este identic nul; aceasta contrazice (47) și astfel (46) nu poate fi adevărată. ■

Rezultatul (45) se poate interpreta în modul următor: cea mai bună aproximare uniformă din $\mathbb{P}_{n-1}$ pe $[-1,1]$ a lui $f(x) = x^n$ este dată de $x^n - \overset{\circ}{T}_n(x)$, adică, de agregarea termenilor până la gradul $n-1$ din $\overset{\circ}{T}_n$ luați cu semnul minus. Din teoria aproximațiilor uniforme se știe că cea mai bună aproximare polinomială uniformă este unică. Deci, egalitatea în (45) poate avea loc numai dacă $\overset{\circ}{p}_n(x) = \overset{\circ}{T}_n(x)$.

Polinoamele Cebîșev de speța a II-a

- Se definesc prin

$$Q_n(t) = \frac{\sin[(n+1)\arccos t]}{\sqrt{1-t^2}}, \quad t \in [-1,1]$$

- Ele sunt ortogonale pe $[-1,1]$ în raport cu măsura $d\lambda(t) = w(t)dt$, $w(t) = \sqrt{1-t^2}$.
- Relația de recurență este

$$Q_{n+1}(t) = 2tQ_n(t) - Q_{n-1}(t), \quad Q_0(t) = 1, \quad Q_1(t) = 2t.$$

Figure: Pafnuti Levovici Cebîșev (1821-1894)

Polinoamele lui Laguerre

- Sunt ortogonale pe $[0, \infty)$ în raport cu ponderea $w(t) = t^\alpha e^{-t}$.
- Se definesc prin

$$\ell_n^\alpha(t) = \frac{e^t t^{-\alpha}}{n!} \frac{d^n}{dt^n}(t^{n+\alpha} e^{-t}) \text{ pentru } \alpha > 1$$

- Relația de recurență pentru polinoamele monice este

$$\ell_{k+1}^\alpha(t) = (t - 2k - \alpha - 1)\ell_k^\alpha(t) - \beta_k \ell_{k-1}^\alpha(t),$$

unde

$$\beta_k = \begin{cases} \Gamma(1+\alpha), & \text{pentru } k = 0; \\ k(k+\alpha), & \text{pentru } k > 0. \end{cases}$$

Figure: Edmond Laguerre (1834-1886)

Polinoamele lui Hermite I

- Se definesc prin

$$H_n(t) = (-1)^n e^{t^2} \frac{d^n}{dt^n}(e^{-t^2}).$$

- Ele sunt ortogonale pe $(-\infty, \infty)$ în raport cu ponderea $w(t) = e^{-t^2}$ și verifică relația de recurență

$$H_{k+1}(t) = tH_k(t) - \beta_k H_{n-1}(t)$$

unde

$$\beta_k = \begin{cases} \sqrt{\pi}, & \text{pentru } k = 0; \\ \frac{k}{2}, & \text{pentru } k > 0. \end{cases}$$

Figure: Charles Hermite (1822-1901)

Polinoamele lui Jacobi

Sunt ortogonale pe $[-1, 1]$ în raport cu ponderea

$$w(t) = (1-t)^{\alpha}(1+t)^{\beta}.$$

Coeficienții din relația de recurență sunt

$$\alpha_k = \frac{\beta^2 - \alpha^2}{(2k+\alpha+\beta)(2k+\alpha+\beta+2)}$$

și

$$\begin{aligned}
\beta_0 =& 2^{\alpha+\beta+1} B(\alpha+1, \beta+1), \\
\beta_k =& \frac{4k(k+\alpha)(k+\alpha+\beta)(k+\beta)}{(2k+\alpha+\beta-1)(2k+\alpha+\beta)^2(2k+\alpha+\beta+1)}, \quad k > 0.
\end{aligned}$$

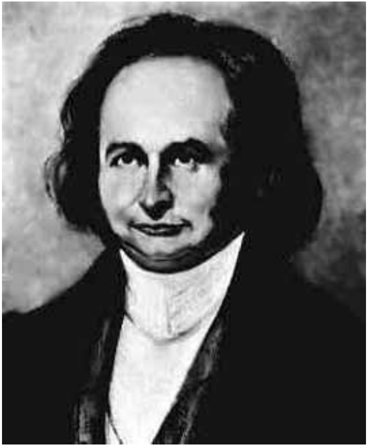


Figure: Carl Gustav Jacob Jacobi (1804-1851)

Exemplu

Pentru funcția $f(t) = \arccos t,\ t \in [-1,1]$, *obțineți aproximanta în sensul celor mai mici pătrate,* $\widehat{\varphi} \in P_n$ *a lui* f *relativ la funcția pondere* $w(t) = (1-t^2)^{-\frac{1}{2}} = \frac{1}{\sqrt{1-t^2}}$ *adică, găsiți soluția* $\varphi = \widehat{\varphi}$ *a problemei*

$$\min\left\{\int_{-1}^{1}[f(t)-\varphi(t)]^2\frac{dt}{\sqrt{1-t^2}}\ :\ \varphi \in P_n\right\}.$$

Exprimați φ *cu ajutorul polinoamelor Cebîșev* $\pi_j(t) = T_j(t)$.

Soluție. $\widehat{\varphi}(t) = \dfrac{c_0}{2} + c_1T_1(x) + \cdots + c_nT_n(x)$

$$c_k = \frac{(f, T_k)}{(T_k, T_k)} = \frac{2}{\pi}(f, T_k) = \frac{2}{\pi}\int_{-1}^{1}\frac{\arccos t}{\sqrt{1-t^2}}\cos(k\arccos t)dt$$

$$= \frac{2}{\pi}\int_0^{\pi} u\cos ku du = \frac{2}{\pi}\left[\left.\frac{u\sin ku}{k}\right|_0^{\pi} - \frac{1}{k}\int_0^{\pi}\sin ku du\right]$$

$$= \frac{2}{\pi}\left[\frac{1}{k}\frac{\cos ku}{k}\bigg|_0^{\pi}\right] = -\frac{2}{\pi k^2}\left[(-1)^k - 1\right]$$

k par $c_k = 0$

k impar $c_k = -\dfrac{2}{\pi k^2}(-2) = \dfrac{4}{\pi k^2}$ ■

Bibliografie I

- Å. Björk, *Numerical Methods for Least Squares Problem*, SIAM, Philadelphia, 1996.
- E. Blum, *Numerical Computing: Theory and Practice*, Addison-Wesley, 1972.
- P. G. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson, Paris, Milan, Barcelone, Mexico, 1990.
- Gheorghe Coman, *Analiză numerică*, Editura Libris, Cluj-Napoca, 1995.
- W. Gautschi, *Numerical Analysis. An Introduction*, Birkhäuser, Basel, 1997.

Bibliografie II

- W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes in C*, Cambridge University Press, Cambridge, New York, Port Chester, Melbourne, Sidney, 1996, disponibila prin www, `http://www.nr.com/`.
- D. D. Stancu, *Analiză numerică – Curs și culegere de probleme*, Lito UBB, Cluj-Napoca, 1977.
- J. Stoer, R. Burlisch, *Introduction to Numerical Analysis*, 2nd ed., Springer Verlag, 1992.