NUMERIČKE METODE (FIZIKA) / NUMERIČKA ANALIZA (C SMJER)

1. INTERPOLACIJA

U numeričkim metodama razmatraju se matematičke formule i računski procesi koji mogu da budu efektivno sprovedeni. To znači da mogu da budu sprovedeni za konačno mnogo izvedenih računskih radnji i da mogu da budu dovedeni do poznatih konkretnih brojnih vrijednosti, najčešće do približnih vrijednosti.

U glavi 1. govorimo o zadatku interpolacije. Dato je nekoliko tačaka u ravni, a treba odrediti kruvu liniju da prolazi kroz sve te tačke.

1.1. LAGRANŽOV INTERPOLACIONI POLINOM

Razmotrimo funkciju f: $R \to R$ ili f: $[a,b] \to R$, Razmotrimo $n \geq 1$ međusobno različitih tačaka $x_1, \ldots, x_n$ na realnoj osi. Neka su date vrijednosti $f(x_i)$ za $i = \overline{1,n}$. Želimo da odredimo funkciju L_n: $R \to R$ ili L_n: $[a,b] \to R$ koja zadovoljava uslov $L_n(x_i) = f(x_i)$ za $i = \overline{1,n}$. Za funkciju L_n se kaže da je interpolaciona funkcija. Kažemo da želimo da riješimo zadatak o interpolaciji za podatke $(x_i, f(x_i))$, sa $i = \overline{1,n}$. Treba definisati klasu funkcija $\{L_n\}$ u kojoj se traži funkcija L_n. Uzmimo da je to klasa svih mogućih polinoma $R \to R$ čiji je stepen $< n$ (čiji je stepen $\leq n-1$). Dakle, polazimo od predstavljanja $L_n(x) = a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$, gdje su $a_j \in R$ ($j = \overline{0, n-1}$) zasad neodređene veličine. Iz uslova interpolacije $L_n(x_i) = f(x_i)$, $i = \overline{1,n}$ imamo $a_{n-1}x_i^{n-1} + \ldots + a_1 x_i + a_0 = f(x_i)$, $i = \overline{1,n}$ ili

$$\begin{bmatrix} x_1^{n-1} & \cdots & x_1 & 1 \\ \cdots & & & \\ x_n^{n-1} & \cdots & x_n & 1 \end{bmatrix} \cdot \begin{bmatrix} a_{n-1} \\ \vdots \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} f(x_1) \\ \cdots \\ f(x_n) \end{bmatrix}$$

Imamo linearni sistem od n jednačina sa n nepoznatih. Njegova determinanta D je poznata Vandermondova determinanta i znamo da je $|D| = \prod_{1 \leq i < j \leq n}(x_j - x_i)$. Po uslovu je $x_i \neq x_j$ za $i \neq j$. Tako da je $D \neq 0$. Zato sistem ima jedinstveno rješenje. Dakle, postavljeni zadatak o interpolaciji ima rješenje u razmatranoj klasi $\{L_n\}$, i to jedinstveno. Na redu je konstrukcija interpolacionog polinoma $L_n = L_n(x)$, odnosno dobijanje njegovog eksplicitnog izraza. Mogao bi da bude riješen linearni sistem, a mi ćemo izvršiti konstrukciju drugim putem.

Razmotrimo proizvod $\omega_n(x) = \prod_{i=1}^{n}(x - x_i)$. Funkcija ω_n je polinom stepena tačno n i zadovoljava $\omega_n(x_i) = 0$ za $i = \overline{1,n}$. U izrazu za $\omega_n(x)$ izostavimo jedan faktor, recimo prvi faktor $x - x_1$. Vidimo da je funkcija $y(x) = \prod_{i=2}^{n}(x - x_i)$ polinom stepena tačno $n-1$ i da zadovoljava $y(x_i) = 0$ za $i = \overline{2,n}$. A $y(x_1) = \prod_{i=2}^{n}(x_1 - x_i) \neq 0$. Prema tome, lako formiramo funkciju $\Phi_i = \Phi_i(x)$ da zadovoljava: $\Phi_i(x_j) = 0$ za $j \neq i$ i $\Phi_i(x_i) = 1$. Upravo

$$\Phi_i(x) = \prod_{\substack{j=1 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \quad \text{ili se piše samo} \quad \Phi_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$

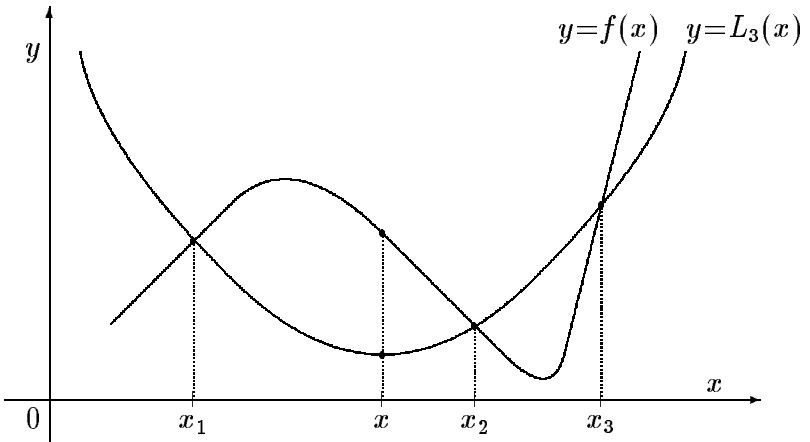
Na kraju, funkcija $L_n(x) = \Phi_1(x)f(x_1) + \ldots + \Phi_n(x)f(x_n)$ predstavlja rješenje postavljenog zadatka, za L_n se kaže da predstavlja Lagranžov interpolacioni polinom, pišemo

$$L_n(x) = \sum_{i=1}^{n} f(x_i) \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$

U numeričkoj praksi, veličina $f(x)$ nije poznata, a veličina $L_n(x)$ je poznata odnosno može da bude izračunata, tako da je $L_n(x)$ približna zamjena za $f(x)$. Ovdje je x jedna tačka na realnoj osi koja nije čvor tj. koja ne pripada mreži čvorova tj. koja se ne poklapa ni sa kojim x_i, $i = \overline{1,n}$. Ponekad se upotrebljava sljedeća terminologija. Ako tačka x za koju se interpolacija vrši pripada odsječku $[a,b]$ obrazovanom od strane čvorova onda se kaže da se vrši interpolacija u užem smislu. A u suprotnom slučaju se kaže da se vrši ekstrapolacija. Ovdje je $a = \min_{i=1,\ldots,n} x_i$ i $b = \max_{i=1,\ldots,n} x_i$.

Tabela Ulazni podaci za zadatak o interpolaciji

i	x_i	$f(x_i)$
1	x_1	$f(x_1)$
2	x_2	$f(x_2)$
$\ldots$	$\ldots$	$\ldots$
n	x_n	$f(x_n)$



Slika za L. i. p. u slučaju $n = 3$

Donekle sličan zadatku o interpolaciji je **zadatak o najboljoj aproksimaciji**, engl. best fit. Skup ulaznih podataka $(x_i, f(x_i))$, $i = \overline{1,n}$ je jedan te isti kod oba zadatka. Kod aproksimacije se ne traži više da pojedinačno odstupanje $r_i = f(x_i) - L_n(x_i)$ bude $= 0$, već se sada traži da $r_i = f(x_i) - g(x_i)$ bude što je moguće manje. Ovdje je g funkcija koja predstavlja rješenje zadatka o aproksimaciji. Traži se ustvari da vrijednost izraza $r^2 = \sum_{i=1}^n |r_i|^2$ bude što je moguće manja, pa se zato kaže i – metoda najmanjih kvadrata. Iz koje klase funkcija $\{g\}$ treba izabrati g? U najprostijem slučaju g je linearna funkcija. Ili je g polinom stepena $\le k$, gdje je $k \le n-2$. Ili je g oblika $g(x) = a + \frac{b}{x}$ ili $g(x) = ae^{bx}$ ili itd.

Najbolja aproksimacija sa $g(x) = ax + b$ za $(x_i, f(x_i)) = (x_i, f_i)$, gdje je $1 \le i \le n$ (linearna regresija):

$$\begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{bmatrix} \cdot \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i f_i \\ \sum_{i=1}^n f_i \end{bmatrix}$$

Navedena formula (za određivanje a i b) lako se izvodi iz uslova za stacionarnu tačku funkcije $F = F(a,b)$, odnosno iz uslova za njenu tačku minimuma. Znamo da uslov za stacionarnu tačku glasi: $\frac{\partial F(a,b)}{\partial a} = 0$, $\frac{\partial F(a,b)}{\partial b} = 0$. Uvedena je oznaka $F(a,b) = \sum_{i=1}^n (ax_i + b - f_i)^2$. Vidimo da je $r^2 = F(a,b)$.

Uzmimo da treba naći zavisnost oblika $y = ae^{bx}$, na osnovu podataka $\{(x_i, y_i),\ 1 \le i \le n\}$. Svođenje na prethodni slučaj pomoću: $y = ae^{bx} \Rightarrow \ln y = \ln a + bx$. Treba primijeniti model prave linije na podatke $\{(x_i, \ln y_i),\ 1 \le i \le n\}$. Kao rezultat, imaćemo b i $\ln a$.

Primjer 1: Naći najbolju aproksimaciju oblika $y = ax + b$ po metodi najmanjih kvadrata za podatke

x	1	2	3	4
y	1	4	6	8

Izrada: $\sum x_i^2 = 30$ $\sum x_i = 10$ $\sum x_i y_i = 59$ $\sum y_i = 19$

$$\begin{cases} 30a + 10b = 59 \\ 10a + 4b = 19 \end{cases} \quad a = 2{,}3 \quad b = -1. \text{ Odgovor: } y = 2{,}3x - 1.$$

Primjer 2: Po metodi najmanjih kvadrata, naći aproksimaciju oblika $y = ae^{bx}$ (svođenjem na linearni slučaj) za podatke

x	1	2	3	4
y	1	4	10	20

Izrada: $\ln y = Y$

x	1	2	3	4
Y	0	1,39	2,30	2,99

$Y = Ax + B$ $\begin{cases} 30A + 10B = 21{,}64 \\ 10A + 4B = 6{,}68 \end{cases}$ $A = 0{,}99$ $B = -0{,}80$ $e^{-0{,}80} = 0{,}45$. Odgovor: $y = 0{,}45e^{0{,}99x}$.

1.2. OCJENA GREŠKE ZA LAGRANŽOV INTERPOLACIONI POLINOM

Ova sekcija predstavlja nastavak prethodne, tako da se sve oznake preuzimaju. Neka $x \in [a, b]$ ili $x \in R$, s tim da x nije čvor. Razmotrimo numerički odgovor $f(x) \approx L_n(x)$ na pitanje: čemu je jednako $f(x)$. Greškom ili greškom numeričkog odgovora nazivamo razliku $r(x) = f(x) - L_n(x)$. U ovoj sekciji treba da dobijemo izraz za $r(x)$, odnosno treba da ocijenimo sa gornje strane $|r(x)|$. Pretpostavlja se da $f \in C^n(R)$ ili $f \in C^n[a, b]$. Uvedimo funkciju φ relacijom

$$\varphi(t) = f(t) - L_n(t) - K\omega_n(t), \tag{1}$$

gdje je K zasad neodređena konstanta. Izaberimo K tako da bude $\varphi(x) = 0$. Dakle, neka je

$$K = \frac{f(x) - L_n(x)}{\omega_n(x)}. \tag{2}$$

Sada je vrijednost K određena. Pogledajmo rješenja jednačine $\varphi(t) = 0$. Vidimo da φ ima bar $n + 1$ nulu. Upravo, njene nule su sigurno $t = x$, $t = x_1$, ..., $t = x_n$. Naime, $\varphi(x_i) = f(x_i) - L_n(x_i) - K\omega_n(x_i) = f(x_i) - f(x_i) - K \cdot 0 = 0$. Tačke $x, x_1, \ldots, x_n$ obrazuju na realnoj osi odsječak $[a, b]$. Dakle, neka bude $a = \min(x, x_1, \ldots, x_n)$ i $b = \max(x, x_1, \ldots, x_n)$. Nabrojane tačke obrazuju na realnoj osi n malih odsječaka. (Rolova teorema: ako je f neprekidna na $[a, b]$, diferencijabilna u (a, b) i $f(a) = f(b)$ onda postoji broj $\xi \in (a, b)$ takav da je $f'(\xi) = 0$.) Rolova teorema govori da između dvije nule funkcije postoji bar jedna nula njenog izvoda. Tako da funkcija φ' ima bar n nula unutar $[a, b]$. Slično, φ'' ima bar $n - 1$ nulu. Itd. Na kraju, $\varphi^{(n)}$ ima bar jednu nulu: postoji $\xi \in (a, b)$ takav da je $\varphi^{(n)}(\xi) = 0$. S druge strane, jasno je da je $L_n^{(n)}(t) \equiv 0$ i da je $\omega_n^{(n)}(t) \equiv n!$ Ako se (1) diferencira n puta po t onda imamo $\varphi^{(n)}(t) = f^{(n)}(t) - Kn!$ Uvrstimo $t = \xi$: $\varphi^{(n)}(\xi) = f^{(n)}(\xi) - Kn!$ ili $0 = f^{(n)}(\xi) - Kn!$ ili

$$K = \frac{f^{(n)}(\xi)}{n!}. \tag{3}$$

Upoređivanjem (2) i (3) dobijamo rješenje tj. traženi izraz za grešku:

$$r(x) = f(x) - L_n(x) = \frac{1}{n!} f^{(n)}(\xi)\omega_n(x), \tag{4}$$

$\xi = \xi(x) \in (a, b)$. Često se (4) piše u obliku:

$$|f(x) - L_n(x)| \le \frac{1}{n!} \cdot M_n \cdot |\omega_n(x)|, \text{ gdje je } M_n = \max_{t \in [a,b]} |f^{(n)}(t)|.$$

greška = ... ili |greška| ≤ ...

1.3. PODIJELJENE RAZLIKE I NJIHOVA SVOJSTVA

U ovoj sekciji uvodi se pojam podijeljene razlike i utvrđuju se neka njihova jednostavna svojstva. Kaže se konačna razlika ili razlika ili engl. difference za izraz $f(x_2)-f(x_1)$. A kaže se podijeljena razlika za izraz $\dfrac{f(x_2)-f(x_1)}{x_2-x_1}$, a za taj izraz koriste se razne oznake, kao $[x_1,x_2]$ ili (f,x_1,x_2) ili $f(x_1;x_2)$. Dakle, neka su o funkciji $f\colon R\to R$ dati na raspolaganje oni isti podaci kao u prethodnim sekcijama: $(x_i,f(x_i))$, $i=\overline{1,n}$. Podijeljena razlika prvog reda funkcije f označava se sa $f(x_i;x_j)$ i jednaka je

$$f(x_i;x_j)=\frac{f(x_j)-f(x_i)}{x_j-x_i}. \tag{1}$$

Slično, drugog reda

$$[x_i,x_j,x_k]\ \text{ ili }\ f(x_i;x_j;x_k)=\frac{f(x_j;x_k)-f(x_i;x_j)}{x_k-x_i}.$$

Uopšte, razlika reda k definiše se pomoću razlika reda $k-1$, i to

$$f(x_1;\ldots;x_{k+1})=\frac{f(x_2;\ldots;x_{k+1})-f(x_1;\ldots;x_k)}{x_{k+1}-x_1}.$$

Podijeljena razlika nultog reda označava se sa $f(x_i)$ i jednaka je $f(x_i)$.

Sljedeća lema utvrđuje jedno svojstvo podijeljenih razlika. Lako se vidi da je podijeljena razlika jednaka jednoj linearnoj kombinaciji vrijednosti funkcije u obuhvaćenim čvorovima. Lema daje izraz za koeficijente linearne kombinacije.

Lema. Za svako $k\geq 1$ važi jednakost

$$f(x_1;\ldots;x_k)=\sum_{j=1}^{k} f(x_j)\cdot\frac{1}{\prod_{i\neq j}(x_j-x_i)}. \tag{2}$$

$\prod_{i=1,\ldots,j-1,j+1,\ldots,k}$

Lemu ćemo dokazati primjenom principa matematičke indukcije po k. Za $k=1$ razmatrana jednakost svodi se na $f(x_1)=f(x_1)$ i očito je istinita. Za $k=2$ ona se svodi na definicionu relaciju (1). Uzmimo da je jednakost (2) tačna za $k=\ell$ i dokažimo da je ona tada tačna i za $k=\ell+1$. Prva naredna transformacija oslanja se na definiciju, a druga na indukcijsku pretpostavku:

$$f(x_1;\ldots;x_{\ell+1})=\frac{1}{x_{\ell+1}-x_1}\Big(f(x_2;\ldots;x_{\ell+1})-f(x_1;\ldots;x_\ell)\Big)=$$

$$\frac{1}{x_{\ell+1}-x_1}\left(\sum_{j=2}^{\ell+1} f(x_j)\cdot\frac{1}{\prod_{\substack{2\leq i\leq \ell+1\\ i\neq j}}(x_j-x_i)}-\sum_{j=1}^{\ell} f(x_j)\cdot\frac{1}{\prod_{\substack{1\leq i\leq \ell\\ i\neq j}}(x_j-x_i)}\right). \tag{3}$$

Čemu je jednak koeficijent c_j uz $f(x_j)$ u posljednjem izrazu? Vidimo da je $1\leq j\leq \ell+1$. Uzmimo prvo da je $1<j<\ell+1$. Tada je

$$c_j=\frac{1}{x_{\ell+1}-x_1}\left(\frac{1}{\prod_{\substack{2\leq i\leq \ell+1\\ i\neq j}}(x_j-x_i)}-\frac{1}{\prod_{\substack{1\leq i\leq \ell\\ i\neq j}}(x_j-x_i)}\right)=$$

$$\frac{1}{x_{\ell+1}-x_1}\cdot\frac{1}{\prod_{\substack{1\le i\le \ell+1\\ i\ne j}}(x_j-x_i)}\cdot\Big((x_j-x_1)-(x_j-x_{\ell+1})\Big)=\frac{1}{\prod_{\substack{1\le i\le \ell+1\\ i\ne j}}(x_j-x_i)}.$$

Vidimo da c_j ima predviđeni oblik. Uzmimo sada da je $j=1$. Veličina $f(x_1)$ pojavljuje se samo u drugom sabirku desne strane formule (3) i očito je

$$c_1=\frac{1}{x_{\ell+1}-x_1}\Big(\prod\nolimits_{\substack{1\le i\le \ell\\ i\ne j}}(x_j-x_i)\Big)^{-1}=\Big(\prod\nolimits_{\substack{1\le i\le \ell+1\\ i\ne j}}(x_j-x_i)\Big)^{-1},$$

ima predviđeni oblik. Slično kada je $j=\ell+1$ tj. za $c_{\ell+1}$. Lema je dokazana.

Iz leme neposredno slijede sljedeća dva svojstva podijeljenih razlika.

1. Podijeljena razlika je linearni operator od funkcije f tj. važi $(\alpha_1 f_1+\alpha_2 f_2)(x_1;\ldots;x_k)=\alpha_1 f_1(x_1;\ldots;x_k)+\alpha_2 f_2(x_1;\ldots;x_k)$.

2. Podijeljena razlika $f(x_1;\ldots;x_k)$ je simetrična funkcija svojih argumenata $x_1,\ldots,x_k$ tj. ne mijenja se pri bilo kakvoj njihovoj permutaciji. Recimo, $f(x_1;x_2)=f(x_2;x_1)$ i $f(x_1;x_2;x_3)=f(x_3;x_1;x_2)$ i slično.

Na kraju, uobičajeno je da se podijeljene razlike prikazuju u obliku tabele. U nastavku je data tabela u slučaju kada su poznate vrijednosti funkcije f u tačkama $x_1,\ldots,x_n$.

$$\begin{array}{lllll}
f(x_1) & f(x_1;x_2) & f(x_1;x_2;x_3) & \ldots & f(x_1;x_2;\ldots;x_n)\\
f(x_2) & f(x_2;x_3) & \ldots & & \\
f(x_3) & \ldots & & & \\
\ldots & & & & \\
 & f(x_{n-1};x_n) & & & \\
f(x_n) & & & &
\end{array}$$

1.4. NJUTNOVA INTERPOLACIONA FORMULA SA PODIJELJENIM RAZLIKAMA

Neka su opet poznate vrijednosti $f(x_1),\ldots,f(x_n)$ funkcije $f\colon R\to R$ u n međusobno različitih tačaka $x_1,\ldots,x_n$. U ovoj sekciji će: a) L.i.p. biti prikazan u drugom obliku. Naime, u 1.1. je bila dokazana jedinstvenost interpolacionog polinoma, tako da Nj.i.p. koji slijedi predstavlja samo drugi oblik jednog te istog polinoma (drugi prikaz, drugi zapis). Još će: b) biti izvedena formula o vezi između podijeljene razlike n–tog reda neke funkcije i njenog n–tog izvoda.

Kao prvo, dokažimo jedan identitet. Vidi se da je

$$f(x)-L_n(x)=f(x)-\sum_{i=1}^{n}f(x_i)\prod\nolimits_{j=1\ \ j\ne i}^{n}\frac{x-x_j}{x_i-x_j}=$$

$$\prod\nolimits_{i=1}^{n}(x-x_i)\cdot\left(\frac{f(x)}{\prod_{i=1}^{n}(x-x_i)}+\sum_{i=1}^{n}\frac{f(x_i)}{(x_i-x)\prod_{j=1\ \ j\ne i}^{n}(x_i-x_j)}\right)=\omega_n(x)\cdot f(x;x_1;\ldots;x_n),$$

uz primjenu leme iz prethodne sekcije. Ranija oznaka $\omega_n(x)=\prod_{i=1}^{n}(x-x_i)$.

$$\text{Identitet}\quad f(x)-L_n(x)=\omega_n(x)\cdot f(x;x_1;\ldots;x_n). \tag{1}$$

S druge strane, prepišimo formulu (4) iz sekcije 1.2.

$$r(x)=f(x)-L_n(x)=\frac{1}{n!}f^{(n)}(\xi)\omega_n(x),\ \ \xi=\xi(x)\in[a,b],$$

izraz za grešku L.i.p. Uporedimo formulu (1) i prepisanu formulu. Tako dobijamo

$$f(x; x_1; \ldots; x_n) = \frac{1}{n!} f^{(n)}(\xi) \tag{2}$$

za neko $\xi \in [a, b]$, gdje je $a = \min(x, x_1 \ldots, x_n)$, $b = \max(x, x_1, \ldots, x_n)$, $f \in C^n[a, b]$, $n \geq 1$.

Na lijevoj strani formule (2) napisana je podijeljena razlika n–tog reda funkcije f.

Uradili smo b) i na redu je a).

Prelazimo na izvođenje Nj.i.f. Uvedimo oznaku $\omega_k(x) = \prod_{i=1}^{k}(x - x_i)$. Neka $L_k = L_k(x)$ označava L.i.p. funkcije f po mreži čvorova $\{x_1, \ldots, x_k\}$. Tako da je naravno $L_k(x_i) = f(x_i)$ za $i = 1, \ldots, k$. Funkcija $L_m(x) - L_{m-1}(x)$ predstavlja očito polinom stepena manjeg od m, a za $x = x_1$, ..., $x = x_{m-1}$ očito je

$$L_m(x) - L_{m-1}(x) = f(x) - f(x) = 0.$$

Primjera radi, kvadratni polinom $p = p(x)$ čije su nule $x = 3$ i $x = 4$ prikazuje se kao $p(x) = A_2(x - 3)(x - 4)$. Tako da se polinom $L_m(x) - L_{m-1}(x)$ prikazuje kao

$$L_m(x) - L_{m-1}(x) = A_{m-1}\omega_{m-1}(x),$$

a treba odrediti A_{m-1}. Supstitucija $x = x_m$:

$$L_m(x_m) - L_{m-1}(x_m) = A_{m-1}\omega_{m-1}(x_m),$$

$$f(x_m) - L_{m-1}(x_m) = A_{m-1}\omega_{m-1}(x_m). \tag{3}$$

S druge strane, upotrebimo (1) kada je $n = m - 1$:

$$f(x) - L_{m-1}(x) = \omega_{m-1}(x) f(x; x_1; \ldots; x_{m-1}),$$

$$\text{supstitucija } x = x_m: \quad f(x_m) - L_{m-1}(x_m) = \omega_{m-1}(x_m) f(x_m; x_1; \ldots; x_{m-1}),$$

$$f(x_m) - L_{m-1}(x_m) = \omega_{m-1}(x_m) f(x_1; \ldots; x_m), \tag{4}$$

jer je podijeljena razlika simetrična funkcija svojih argumenata.

Uporedimo (3) i (4). Tako $A_{m-1} = f(x_1; \ldots; x_{m-1})$ i zato

$$L_m(x) - L_{m-1}(x) = f(x_1; \ldots; x_m)\omega_{m-1}(x).$$

Slijedi

$$L_n(x) = L_1(x) + (L_2(x) - L_1(x)) + (L_3(x) - L_2(x)) + \ldots + (L_n(x) - L_{n-1}(x)),$$

$$L_n(x) = L_1(x) + f(x_1; x_2)\omega_1(x) + f(x_1; x_2; x_3)\omega_2(x) + \ldots + f(x_1; x_2; \ldots; x_n)\omega_{n-1}(x),$$

$$L_n(x) = f(x_1) + f(x_1; x_2)(x - x_1) + f(x_1; x_2; x_3)(x - x_1)(x - x_2) + \ldots +$$

$$f(x_1; x_2; \ldots; x_n)(x - x_1) \ldots (x - x_{n-1}). \tag{5}$$

Interpolacioni polinom zapisan u obliku (5) naziva se Njutnovim interpolacionim polinomom sa podijeljenim razlikama.

Na kraju, želimo da zapišemo zajedno tj. u jednoj formuli i izraz za interpolacioni polinom i izraz za njegovu grešku, želimo da napišemo formulu oblika tačna vrijednost = približna vrijednost + greška tj. formulu $f(x) = L_n(x) + r(x)$. Na osnovu (1) i (5) imamo:

$$f(x) = f(x_1) + f(x_1; x_2)\omega_1(x) + f(x_1; x_2; x_3)\omega_2(x) + \ldots +$$
$$f(x_1; \ldots; x_n)\omega_{n-1}(x) + f(x; x_1; \ldots; x_n)\omega_n(x).$$

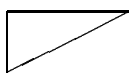
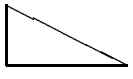
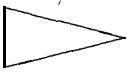
Najzad, zapazimo jedno dobro svojstvo koje posjeduje Njutnova interpolaciona formula sa podijeljenim razlikama, neka vrsta "aditivnosti". Zamislimo ovakvu situaciju. Mi procjenjujemo $f(x)$ pomoću $L_4(x)$ tj. koristimo čvorove $x_1, \ldots, x_4$. U okviru toga, ocijenimo odgovarajuću grešku $f(x) - L_4(x)$, pri čemu se ispostavi da je greška isuviše velika, ispostavi se da nismo zadovoljni dobijenom preciznošću. Tada, radi dobijanja bolje aproksimacije, uključimo u račun još jedan čvor x_5. Zapaziti da prilikom računanja zbira $L_5(x)$ ne moramo da ponovo računamo sve njegove sabirke, već da je dovoljno da se dosad izračunatoj približnoj vrijednosti $L_4(x)$ doda još jedan sabirak; uporedi izraze za $L_4(x)$ i $L_5(x)$.

1.5. KONAČNE RAZLIKE

Konačne razlike čine osnovni aparat u numeričkim metodama, a definišu se samo u slučaju ravnomjerne (ekvidistantne) mreže čvorova. U ovoj sekciji daju se definicije i svojstva. Ekvidistantna mreža čvorova definiše se pomoću početnog čvora $x_0 \in R$ i svog koraka $h > 0$. Čvorovi su $x_i = x_0 + ih$, obično za i važi $i = 0, 1, \ldots, n$, gdje je $n \geq 1$. Sljedeća tablica predstavlja primjer. Mreža je definisana sa $x_0 = 0 \quad h = 0{,}1 \quad n = 5$. Tablica se odnosi na funkciju $y(x) = e^{2x}$. Prikazane su njene vrijednosti u čvorovima y_i i odgovarajuće konačne razlike prvog reda Δy_i i drugog reda $\Delta^2 y_i$:

i	x_i	y_i	Δy_i	$\Delta^2 y_i$
0	0	1	0,22140	0,04902
1	0,1	1,22140	0,27042	0,05988
2	0,2	1,49182	0,33030	0,07312
3	0,3	1,82212	0,40342	0,08932
4	0,4	2,22554	0,49274	
5	0,5	2.71828		

Razmotrimo funkciju $f\colon R \to R$ ili $f\colon [a, b] \to R$. Označimo sa f_i njene vrijednosti u čvorovima mreže tj. $f_i = f(x_i)$ za $i = 0, 1, \ldots, n$. Za jednu te istu konačnu razliku koriste se tri različita naziva. Izraz $f_{i+1} - f_i$ naziva se konačnom razlikom prvog reda. Za taj izraz koriste se sljedeće tri različite oznake: $\Delta f_i = f_{i+1} - f_i$ konačna razlika unaprijed, $\nabla f_{i+1} = f_{i+1} - f_i$ konačna razlika unazad i $\delta f_{i+1/2} = f_{i+1} - f_i$ ili svejedno $f^1_{i+1/2} = f_{i+1} - f_i$ centralna konačna razlika. Razlike višeg reda definišu se na osnovu razlika nižeg reda: $\Delta^m f_i = \Delta(\Delta^{m-1} f_i) = \Delta^{m-1} f_{i+1} - \Delta^{m-1} f_i$, $\nabla^m f_i = \nabla(\nabla^{m-1} f_i) = \nabla^{m-1} f_i - \nabla^{m-1} f_{i-1}$ i $\delta^m f_i = \delta(\delta^{m-1} f_i) = \delta^{m-1} f_{i+1/2} - \delta^{m-1} f_{i-1/2}$ ili svejedno $f^m_i = f^{m-1}_{i+1/2} - f^{m-1}_{i-1/2}$. Vidimo da se za centralne razlike koriste dvije različite oznake $\delta^m f_i$ i f^m_i. U slučaju centralne razlike $\delta^m f_i = f^m_i$, ako je m paran onda je i cio, a ako je m neparan onda je i polucio.

Tabela razlika unaprijed Δ^m ima oblik  , tabela razlika unazad ∇^m ima oblik  , a tabela centralnih razlika δ^m ima oblik  . U nastavku je prikazana tablica centralnih razlika:

x	f	$\delta f{=}f^1$	$\delta^2 f{=}f^2$	$\delta^3 f{=}f^3$	
x_0	f_0				$\vdots$
		$f^1_{1/2}$			
x_1	f_1		f^2_1		$\vdots$
		$f^1_{3/2}$		$f^3_{3/2}$	
x_2	f_2		f^2_2		$\vdots$
		$f^1_{5/2}$		$f^3_{5/2}$	
x_3	f_3		f^2_3		
		$f^1_{7/2}$		$\vdots$	
x_4	f_4		$\vdots$		
		$\vdots$			
$\vdots$	$\vdots$				

Na primjer: $\Delta f_i = f_{i+1} - f_i$, $\Delta^2 f_i = f_{i+2} - 2f_{i+1} + f_i$, $\Delta^3 f_i = f_{i+3} - 3f_{i+2} + 3f_{i+1} - f_i$, $\Delta^4 f_i = f_{i+4} - 4f_{i+3} + 6f_{i+2} - 4f_{i+1} + f_i$, itd. Vidimo da je konačna razlika jedna linearna kombinacija vrijednosti funkcije u čvorovima koji su obuhvaćeni. Sljedeća lema utvrđuje čemu su jednaki koeficijenti linearne kombinacije.

Lema 1. Za svako $m \geq 1$ važi jednakost

$$\Delta^m f_i = \sum_{j=0}^{m} (-1)^j \binom{m}{j} f_{i+m-j}.$$

Lemu ćemo dokazati indukcijom. Za $m = 1$ jednakost se svodi na $\Delta f_i = f_{i+1} - f_i$ i očito je tačna. Uzmimo da je jednakost tačna za $m = l$ i dokažimo da je tada jednakost tačna i za $m = l + 1$:

$$\Delta^{l+1} f_i = \Delta^l f_{i+1} - \Delta^l f_i = \sum_{j=0}^{l} (-1)^j \binom{l}{j} f_{i+1+l-j} - \sum_{j=0}^{l} (-1)^j \binom{l}{j} f_{i+l-j} =$$

$$f_{i+1+l} + \sum_{j=1}^{l} (-1)^j \binom{l}{j} f_{i+1+l-j} - \sum_{j=0}^{l-1} (-1)^j \binom{l}{j} f_{i+l-j} - (-1)^l f_i =$$

$$f_{i+1+l} + \sum_{j=1}^{l} (-1)^j \binom{l}{j} f_{i+1+l-j} - \sum_{j=1}^{l} (-1)^{j-1} \binom{l}{j-1} f_{i+l-j+1} + (-1)^{l+1} f_i =$$

$$\text{pomoću} \quad \binom{l}{j} + \binom{l}{j-1} = \binom{l+1}{j} \quad \text{imamo}$$

$$f_{i+1+l} + \sum_{j=1}^{l} (-1)^j \binom{l+1}{j} f_{i+1+l-j} + (-1)^{l+1} f_i =$$

$$\sum_{j=0}^{l+1} (-1)^j \binom{l+1}{j} f_{i+l+1-j},$$

što je i trebalo. Lema je dokazana.

Lema pokazuje da je operator uzimanja konačne razlike od neke funkcije linearan po toj funkciji tj. da važi jednakost $\Delta^m(\alpha f + \beta g)_i = \alpha \Delta^m f_i + \beta \Delta^m g_i$.

Podijeljena razlika $f(x_i;\ldots;x_{i+m})$ reda m obrazuje se po vrijednostima funkcije u $m+1$ čvorova $x_i,\ldots,x_{i+m}$ koji su raspoređeni bilo ravnomjerno bilo neravnomjerno. Konačna razlika $\Delta^m f_i$ reda m obrazuje se po tim istim čvorovima, jedino kada oni čine ravnomjernu mrežu.

Lema 2. Važi jednakost

$$\Delta^m f_i = h^m \cdot m! \cdot f(x_i;\ldots;x_{i+m}).$$

Dokazuje se indukcijom. Za $m=1$ imamo $\Delta f_i = hf(x_i;x_{i+1})$ ili $\Delta f_i = h\frac{f_{i+1}-f_i}{x_{i+1}-x_i}$ ili $\Delta f_i = f_{i+1}-f_i$. Indukcijski korak:

$$\Delta^{l+1} f_i = \Delta^l f_{i+1} - \Delta^l f_i =$$

$$\Big(h^l \cdot l! \cdot f(x_{i+1};\ldots;x_{i+l+1}) - h^l \cdot l! \cdot f(x_i;\ldots;x_{i+l})\Big) \cdot \frac{h\cdot(l+1)}{x_{i+l+1}-x_i} =$$

$$h^{l+1}\cdot(l+1)!\cdot f(x_i;\ldots;x_{i+l+1}), \quad \text{dokazano.}$$

Bila je formula (2) u sekciji 1.4. $f(x_i;\ldots;f_{i+m}) = \frac{1}{m!}f^{(m)}(\xi)$.

Formula (1.4.2) izražava vezu između podijeljene razlike određenog reda i izvoda funkcije tog istog reda. Jednostavnim njenim kombinovanjem sa prethodnom lemom, dobija se veza konačnih razlika i izvoda, kako slijedi.

Lema 3. Ako $f \in C^m[x_i, x_{i+m}]$ onda važi

$$\Delta^m f_i = h^m \cdot f^{(m)}(\xi), \quad \text{za neko } \xi \in [x_i, x_{i+m}].$$

Naravno da se ista jednakost može zapisati i u drugim oznakama: $\Delta^m f_i = \nabla^m f_{i+m} = \delta^m f_{i+m/2} = f^m_{i+m/2} = h^m \cdot f^{(m)}(\xi)$, za neko $\xi \in [x_i, x_{i+m}] = [x_0+ih, x_0+(i+m)h]$.

1.6. NJUTNOVE INTERPOLACIONE FORMULE SA KONAČNIM RAZLIKAMA

U ovoj sekciji izvode se dvije najpoznatije interpolacione formule. I one predstavljaju samo drugi zapis L. i. p. Račun je jednostavan, jer je sve pripremljeno u prethodnim sekcijama.

Razmotrimo ravnomjernu mrežu čvorova čiji je osnovni čvor $x = x_0 \in R$ i čiji je korak $h > 0$. Sami čvorovi su $x_i = x_0 + ih$ za razne cijele i. Pored nezavisno promjenljive x koristi se i druga nezavisno promjenljiva t, kako je to uobičajeno kada se radi o ravnomjernoj mreži. Dvije promjenljive povezane su sljedećom relacijom o smjeni promjenljive: $x = x_0 + ht$ ili svejedno $t = \frac{x-x_0}{h}$.

Razmotrimo funkciju f i označimo sa f_i njene vrijednosti u čvorovima x_i.

Neka je $L_n = L_n(x)$ L. i. p. za funkciju f za mrežu čvorova $x_0, x_1, \ldots, x_{n-1}$. Prepišimo formulu (1.4.5) iz sekcije 1.4.

$$L_n(x) = f(x_0) + f(x_0;x_1)(x-x_0) + \ldots + f(x_0;\ldots;x_{n-1})(x-x_0)\ldots(x-x_{n-2})$$

sa podijeljenih razlika prelazimo na konačne razlike po lemi iz prethodne sekcije

$$L_n(x) = f(x_0) + \frac{1}{h}\Delta f_0 \cdot (x-x_0) + \ldots + \frac{1}{(n-1)!}\cdot\frac{1}{h^{n-1}}\cdot\Delta^{n-1}f_0\cdot(x-x_0)\ldots(x-x_{n-2})$$

sa x prelazimo na t

$$x - x_0 = ht,\ x - x_1 = h(t-1),\ x - x_2 = h(t-2),\ \ldots$$

$$L_n(x_0+ht) = f_0 + \Delta f_0 \cdot t + \frac{1}{2!}\cdot\Delta^2 f_0 \cdot t(t-1) + \ldots + \frac{1}{(n-1)!}\cdot\Delta^{n-1} f_0 \cdot t(t-1)\ldots(t-n+2),\ (I)$$

ovo je I Nj. i. f. ili Nj. i. f. za interpolaciju unaprijed.

Moglo bi se naravno umjesto $\Delta f_0, \Delta^2 f_0, \ldots, \Delta^{n-1} f_0$ pisati i $f^1_{1/2}, f^2_1, \ldots, f^{n-1}_{(n-1)/2}$.

Što se tiče izraza za grešku formule (I), kako se radi samo o drugom prikazu jednog te istog polinoma, to možemo naravno da iskoristimo raniju formulu za grešku (1.2.4), samo što ćemo toj formuli dati nešto drugi oblik, kako slijedi:

$$f(x) - L_n(x) = f(x_0+ht) - L_n(x_0+ht) = \frac{1}{n!}\cdot f^{(n)}(\xi)\cdot\prod_{i=0}^{n-1}(x-x_i),$$

$$f(x) - L_n(x) = f(x_0+ht) - L_n(x_0+ht) = \frac{1}{n!}\cdot f^{(n)}(\xi)\cdot h^n\cdot t(t-1)\ldots(t-n+1),$$

gdje je ξ neki broj takav da je $\min(x_0, x) \le \xi \le \max(x, x_{n-1})$.

Razmotrimo sada n čvorova $x_{-(n-1)} < \ldots < x_{-1} < x_0$. Neka je sada $L_n = L_n(x)$ L. i. p. za f po toj mreži čvorova. I dalje je $x = x_0 + ht$. Sličnim izvođenjem dobija se II Nj. i. f. ili Nj. i. f. za interpolaciju unazad:

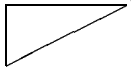
$$L_n(x) = f(x_0) + f(x_0; x_{-1})\cdot(x-x_0) + \ldots + f(x_0;\ldots;x_{-(n-1)})\cdot(x-x_0)\ldots(x-x_{-(n-2)}),$$

$$L_n(x_0+ht) = f_0 + \nabla f_0\cdot t + \frac{1}{2!}\cdot\nabla^2 f_0\cdot t(t+1) + \ldots + \frac{1}{(n-1)!}\cdot\nabla^{n-1} f_0\cdot t(t+1)\ldots(t+n-2),\ (II)$$

a za grešku formule (II) važi sljedeći izraz:

$$r(x) = f(x) - L_n(x) = f(x_0+ht) - L_n(x_0+ht) = \frac{1}{n!}\cdot f^{(n)}(\xi)\cdot h^n\cdot t(t+1)\ldots(t+n-1),$$

$\min(x_{-(n-1)}, x) \le \xi \le \max(x, x_0)$. Ili $x_{-(n-1)} \le \xi \le x_0$, kada se radi o interpolaciji u užem smislu.

Pogledajmo još jednom I Nj. i. f. Uobičajeno je da se ta formula zapisuje preko konačnih razlika unaprijed. Tako da se prethodno formira odgovarajuća tablica konačnih razlika unaprijed, ta tablica ima gornji trougaoni oblik  Vidi se da sve potrebne brojne vrijednosti čitamo samo iz gornjeg reda tablice.

Ako koristimo mrežu $\{x_0, x_1, x_2\}$ onda je

$$L_3(x_0+ht) = f_0 + t\cdot\Delta f_0 + \frac{1}{2}\cdot t(t-1)\cdot\Delta^2 f_0.$$

Dodajmo još jedan čvor, da bismo sa većom preciznošću saznali $f(x_0+ht)$. Ako koristimo mrežu $\{x_0, x_1, x_2, x_3\}$ onda je

$$L_4(x_0+ht) = f_0 + t\cdot\Delta f_0 + \frac{1}{2}\cdot t(t-1)\cdot\Delta^2 f_0 + \frac{1}{6}\cdot t(t-1)(t-2)\cdot\Delta^3 f_0.$$

Vidimo da se doda jedan sabirak, kada se sa L_3 prelazi na L_4.

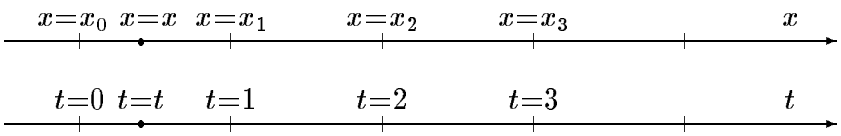
Zapišimo L_4 zajedno sa izrazom za grešku:

$$f(x_0 + ht) = L_4(x_0 + ht) + \frac{1}{24} \cdot t(t-1)(t-2)(t-3) \cdot f^{IV}(\xi) \cdot h^4.$$

Kaže se da je I Nj. i. f. slična Tejlorovoj formuli.

Pogledajmo mali primjer. Neka je $h = 0{,}1 \quad x_0 = 0 \quad x_1 = 0{,}1 \quad x_2 = 0{,}2 \quad x_3 = 0{,}3$. Poznata je veza $x = x_0 + ht$. Recimo, ako je $x = 0{,}04$ tada je $t = 0{,}4$.

Sljedeća slika prikazuje vezu dvije promjenljive x i t:



Pogledajmo mali primjer za II Nj. i. f. Neka bude $h = 0{,}1 \quad x_{-3} = 9{,}7 \quad x_{-2} = 9{,}8 \quad x_{-1} = 9{,}9$ $x_0 = 10$. Stalno je $x = x_0 + ht$. Recimo, ako je $x = 9{,}87$ tada je $t = -1{,}3$.

Iskustvo pokazuje da je najbolje da se interpolacioni polinom obrazuje pomoću vrijednosti funkcije u malom broju tačaka, obično manje ili jednako od pet ili šest tačaka. Pogledajmo tri moguća slučaja. Neka imamo veliku tablicu vrijednosti funkcije f, po ravnomjernoj mreži. Recimo od $x = 0$ do $x = 10$ sa korakom $h = 0{,}1$. Na osnovu datih informacija treba, za dato x koje nije čvor, primjenom aparata interpolacije, naći dobru približnu vrijednost za broj $f(x)$. Tačna vrijednost $f(x)$ očito je van našeg domašaja. Prvi slučaj: tačka x je blizu početka mreže. Tada treba primijeniti I Nj. i. f. Drugi slučaj: tačka x je blizu kraja mreže, blizu krajnjeg desnog čvora. Tada je pogodno da se primijeni II Nj. i. f. I još treći mogući slučaj: za dati broj x može da se kaže da se nalazi u sredini tablice. Ako se opredijelimo da koristimo šest čvorova (oni su uzastopni) onda uzimamo tri lijevo od x i tri desno od x. Zato što se tako minimizuje greška, jer se minimizuje proizvod $\omega_n(x)$. A i intuitivno je jasno da je tako povoljno. A za sprovođenje samog računanja treba upotrebiti bilo koji od raznih, a međusobno ekvivalentnih, zapisa (prikaza) interpolacionog polinoma. Recimo L. i. p. Ili Nj. i. f. sa podijeljenim razlikama. Ili I Nj. i. f. gdje sa x_0 označimo krajnji lijevi od tri lijeva čvora.

1.7. INTERPOLACIJA SA VIŠESTRUKIM ČVOROVIMA

U slučaju L. i. p. pretpostavljalo se da su u čvorovima date samo vrijednosti funkcije. Razmotrimo sada jedan opštiji zadatak interpolacije. V. sliku. Neka mrežu čine čvorovi $x_1, \ldots, x_n \in [a,b]$, među kojima nema poklapanja. Ovdje je $a = \min(x, x_1, \ldots, x_n)$ i $b = \max(x, x_1, \ldots, x_n)$. Neka su u čvorovima date vrijednosti funkcije $f(x_i)$ i njenih izvoda $f^{(j)}(x_i)$ do reda $m_i - 1$ uključeno, $j = 1, \ldots, m_i - 1$, za $i = 1, \ldots, n$. Prema tome, u svakoj tački x_i poznate su brojne vrijednosti $f(x_i), f'(x_i), \ldots, f^{(m_i-1)}(x_i)$, $i = 1, \ldots, n$. Tako da ulaznih veličina ukupno ima $m_1 + \ldots + m_n = s$. Želimo da odredimo interpolacionu funkciju koja odgovara ovim ulaznim podacima. Opredijelimo se da interpolacionu funkciju tražimo u klasi svih polinoma stepena manjeg od s. Za H_s se kaže da je Hermitov interpolacioni polinom ili da je interpolacioni polinom sa višestrukim čvorovima. Za $m_i \geq 1$ se kaže da predstavlja višestrukost čvora x_i, $i = 1, \ldots, n$.

Mi treba da razmotrimo sljedeća četiri pitanja: 1) postojanje i. p. 2) jedinstvenost i. p. 3) konstrukcija i. p. tj. dobijanje eksplicitnog izraza za i. p. i 4) ocjena greške: kolika se greška čini kada se kaže da je $f(x)$ približno jednako $H_s(x)$, za određeno $x \in [a,b]$.

Dakle, $H_s = H_s(x)$ treba da bude polinom stepena $\leq s-1$ koji zadovoljava uslove: $H_s^{(j)}(x_i) = f^{(j)}(x_i)$, $j = 0, \ldots, m_i - 1$, $i = 1, \ldots, n$. Ima s uslova. Opšti izraz za polinom stepena $\leq s-1$ glasi: $H_s(x) = a_{s-1}x^{s-1} + \ldots + a_1 x + a_0$. Ima s koeficijenata tj. s stepeni slobode.

Oko jedinstvenosti. Iz algebre je poznato sljedeće tvrđenje: jedini polinom $p = p(x)$ stepena $\leq s-1$ koji ima s nula, pri čemu se pojedina nula broji svojom višestrukošću, jeste $p(x) \equiv 0$. Dopustimo da postoje dva polinoma $H_s^{(1)}(s)$ i $H_s^{(2)}(x)$ stepena $\leq s-1$ da oba zadovoljavaju sve uslove (svih s ograničenja). Neka bude $H_s^{(3)}(x) = H_s^{(1)}(x) - H_s^{(2)}(x)$. Imamo da je $H_s^{(3)(j)}(x) = 0$, $j = 0, \ldots, m_i - 1$, $i = 1, \ldots, n$. $H_s^{(3)}$ je polinom stepena $\leq s-1$ i ima barem s nula (brojano sa višestrukostima). Zato je $H_s^{(3)}(x) \equiv 0$ tj. $H_s^{(1)}(x) \equiv H_s^{(2)}(x)$. Ovim je jedinstvenost dokazana.

Oko postojanja. Dovedimo u vezu s ograničenja i s stepeni slobode. Dakle, treba da budu zadovoljene sljedeće jednakosti (ima ih s):

$$H_s^{(j)}(x_i) = f^{(j)}(x_i), \quad j = 0, \ldots, m_i - 1, \quad i = 1, \ldots, n. \quad (1)$$

Mi smo upravo napisali sistem od s linearnih jednačina sa s nepoznatih $a_{s-1}, \ldots, a_0$. Razmotrimo i odgovarajući homogeni sistem:

$$H_s^{(j)}(x_i) = 0, \quad j = 0, \ldots, m_i - 1, \quad i = 1, \ldots, n. \quad (2)$$

Ima li homogeni sistem drugog rješenja osim trivijalnog rješenja $a_{s-1} = \ldots = a_0 = 0$? Kao što je maločas navedeno prilikom analize jedinstvenosti, iz (2) slijedi da je $H_s(x) \equiv 0$; slijedi da je $a_{s-1} = \ldots = a_0 = 0$. Homogeni sistem (2) ima samo trivijalno rješenje. Zato i sistem (1) ima jedinstveno rješenje, ma kakve da su brojne vrijednosti $\{f^{(j)}(x_i)\}$ koje čine njegovu desnu stranu (njegove slobodne članove). Ovim je postojanje dokazano; v. i šablon kasnije.

Sada o konstrukciji. Brojne vrijednosti $\{f^{(j)}(x_i)\}$ ne učestvuju u matrici sistema (1) već jedino učestvuju na desnoj strani sistema. Zato će pojedina nepoznata a_k da bude linearna kombinacija tih brojnih vrijednosti. Broj a_k je koeficijent polinoma. Kako je $H_s(x) = a_{s-1}x^{s-1} + \ldots + a_1 x + a_0$ to polinom $H_s(x)$ može da bude prikazan u obliku (kada se linearne kombinacije uvrste u $H_s(x) = \ldots$)

$$H_s(x) = \sum_{i=1}^{n} \sum_{j=0}^{m_i-1} c_{ij}(x) f^{(j)}(x_i),$$

gdje su $c_{ij}(x)$ polinomi stepena $\leq s-1$. Nećemo izvoditi eksplicitne izraze za $c_{ij}(x)$, zato što su ti izrazi glomazni, već ćemo samo navesti jedan primjer kasnije.

Još o ocjeni greške. Neka $x \in [a,b]$ i neka x nije čvor. Treba ocijeniti razliku $f(x) - H_s(x)$. Neka $f \in C^{(s)}[a,b]$. Uvedimo proizvod $\omega_s(t) = \prod_{i=1}^{n}(t-x_i)^{m_i}$. Tako da je ω_s polinom stepena tačno s čiji je najstariji koeficijent $= 1$. Pa će biti $\omega_s^{(s)}(t) = s!$ Uvedimo u razmatranje i funkciju φ relacijom

$$\varphi(t) = f(t) - H_s(t) - K\omega_s(t), \quad (3)$$

gdje je K zasad neodređena brojna vrijednost. Vrijednost K biramo tako da u tački interpolacije $t = x$ bude zadovoljen uslov $\varphi(x) = 0$. Dakle, mi stavljamo $K = \dfrac{f(x) - H_s(x)}{\omega_s(x)}$. Rekli smo da je tačka nterpolacije $t = x$ fiksirana. Koliko nula ima funkcija $\varphi = \varphi(t)$? Tačka $t = x_i$ je nula reda m_i, za svako $i = 1, \ldots, n$. Plus $t = x$. Ukupno ima nula, uzimajući u obzir njihove višestrukosti, najmanje $m_1 + \ldots + m_n + 1 = s + 1$. Odavde, po Rolovoj teoremi, prvi izvod $\varphi' = \varphi'(t)$ ima bar s nula na odsječku $[a,b]$, uračunavajući višestrukosti. Slično, φ'' ima bar $s-1$ nula. Itd. Funkcija $\varphi^{(s)}$ ima bar jednu nulu. Označimo tu nulu sa ξ: $\varphi^{(s)}(\xi) = 0$, $\xi \in [a,b]$. Diferencirajmo relaciju (3) s puta. Tako $\varphi^{(s)}(t) = f^{(s)}(t) - Ks!$ Uvrstimo ovdje $t = \xi$. Tako $0 = f^{(s)}(\xi) - Ks!$ Odranije imamo da je $K = \dfrac{f(x) - H_s(x)}{\omega_s(x)}$. Prema tome

$$f(x) - H_s(x) = \frac{1}{s!} f^{(s)}(\xi)\omega_s(x).$$

Ovdje je $\xi = \xi(x)$. Dobili smo izraz za grešku.

Razmotrimo jedan primjer. Neka imamo $n = 3$ čvora $x = -1$, $x = 0$ i $x = 1$. Neka su čvorovi $x = -1$ i $x = 1$ jednostruki (prosti), a $x = 0$ neka je dvostruk, tako da je $s = 4$. Dakle, date su četiri brojne vrijednosti, označavaju se kao $f(-1)$, $f(0)$, $f'(0)$, $f(1)$ ili kao f_{-1}, f_0, f_0', f_1. Treba odrediti polinom trećeg stepena $H_4 = H_4(x)$ da zadovoljava:

$$H_4(-1) = f_{-1}, \quad H_4(0) = f_0, \quad H_4'(0) = f_0', \quad H_4(1) = f_1.$$

U okviru primjera, sastavićemo eksplicitni izraz za H_4 i napisaćemo izraz za ocjenu greške. V. sliku za H_4. Za eksplicitni izraz, treba da se pođe od predstavljanja $H_4(x) = a_3x^3 + a_2x^2 + a_1x + a_0$. Dobiće se:

$$H_4(x) = -\frac{1}{2}x^2(x-1)f_{-1} - (x+1)(x-1)f_0 + \frac{1}{2}(x+1)x^2 f_1 - (x+1)x(x-1)f_0'.$$

Izraz za grešku:

$$f(x) - H_4(x) = \frac{1}{24} f^{IV}(\xi)(x+1)x^2(x-1), \quad \xi \in [-1,1] \quad (\text{ako } x \in [-1,1]).$$

Šablon: $H_s(x) = a_{s-1}x^{s-1} + \ldots + a_1x + a_0$

$H_s'(x) = (s-1)a_{s-1}x^{s-2} + \ldots + a_1$, itd.

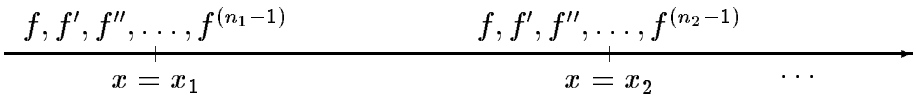
$H_s(x_1) = f(x_1) \Rightarrow a_{s-1}x_1^{s-1} + \ldots + a_1x_1 + a_0 = f(x_1)$

$$H_s'(x_1) = f'(x_1) \Rightarrow (s-1)a_{s-1}x^{s-2} + \ldots + a_1 = f'(x_1), \quad \text{itd.}$$

$$H_s(x_2) = f(x_2) \Rightarrow a_{s-1}x_2^{s-1} + \ldots + a_1x_2 + a_0 = f(x_2), \quad \text{itd.} \quad \text{itd.}$$

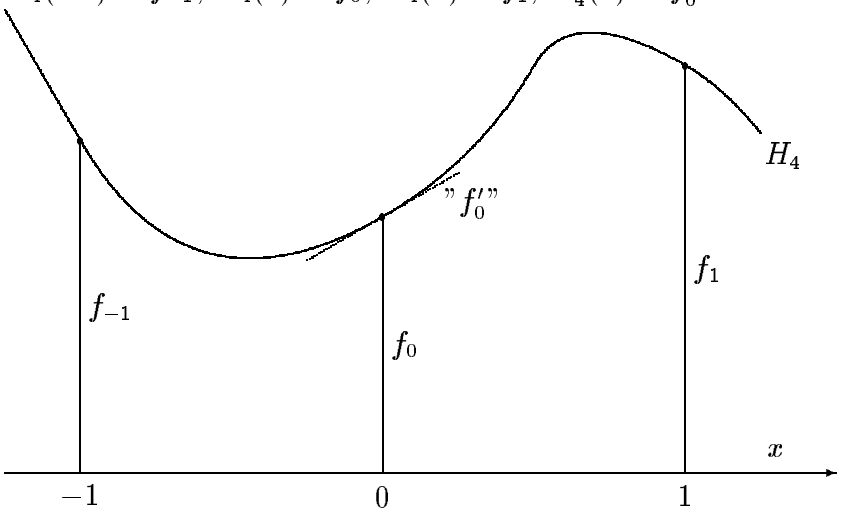
$$\begin{bmatrix} x_1^{s-1} & \cdots & x_1 & 1 \\ (s-1)x_1^{s-2} & \cdots & 1 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ x_2^{s-1} & \cdots & x_2 & 1 \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \cdot \begin{bmatrix} a_{s-1} \\ \cdots \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f'(x_1) \\ \cdots \\ f(x_2) \\ \cdots \end{bmatrix}$$

ako su na desnoj strani sve nule onda postoji samo trivijalno rješenje $\Rightarrow \det M \neq 0$



$H_4(x) = a_3x^3 + a_2x^2 + a_1x + a_0$

$H_4(-1) = f_{-1}$, $H_4(0) = f_0$, $H_4(1) = f_1$, $H_4'(0) = f_0'$



1.8. INTERPOLACIJA POMOĆU SPLAJNA

Razmotrimo opet zadatak o interpolaciji funkcije. Neka n bude prirodan broj i neka $[a, b]$ bude odsječak na realnoj osi. Neka je po odsječku $[a, b]$ postavljeno $n + 1$ čvorova $a = x_0 < x_1 < \ldots < x_n = b$ i neka je u svakom čvoru poznata brojna vrijednost $f(x_i) = f_i \in R$. Na osnovu tih podataka o funkciji f, treba naći približnu vrijednost za $f(x)$, gdje je x data tačka iz odsječka $[a, b]$. Iskustvo pokazuje da interpolacija pomoću polinoma visokog stepena

obično ne daje zadovoljavajuće aproksimacije. Zato razmotrimo drugu mogućnost za rješavanje postavljenog zadatka. Neka interpolaciona funkcija $s = s(x)$ na svakom malom odsječku $[x_0, x_1], \ldots, [x_{n-1}, x_n]$ bude polinom čiji stepen nije visok. Za $s = s(x)$ se kaže da je splajn, a kaže se da se vrši interpolacija pomoću splajna ili pomoću dio–po–dio polinoma; engl. spline – krivuljar. Ako je s polinom trećeg stepena na $[x_{i-1}, x_i]$ onda se kaže da je s kubni splajn (ovo ćemo raditi). Kada dobijemo splajn, onda ćemo mi, kada je data tačka x, da izračunamo brojnu vrijednost $s(x)$. Broj $s(x)$ i predstavlja naš odgovor tj. kazaćemo da je $f(x) \approx s(x)$. Treba da se dobije i ocjena za grešku $r(x) = f(x) - s(x)$. Radi jednostavnosti računanja, razmotrićemo samo slučaj ekvidistantno raspoređenih čvorova $x_i = a + ih$, za $i = \overline{0,n}$, gdje je $h = \dfrac{b-a}{n}$.

Definicija. Za podatke n, a, b i $\{f_i\}_{i=0}^n$, kubnim splajnom naziva se funkcija s: $[a,b] \to R$ koja zadovoljava sljedeća tri uslova: a) na svakom malom odsječku $[x_{i-1}, x_i]$, $s(x)$ je polinom trećeg stepena, b) $s \in C^2[a,b]$ i c) $s(x_i) = f_i$ za $i = \overline{0,n}$ (uslov interpolacije).

Jasno je da je

$$s(x) = s_i(x) \quad \text{za} \quad x_{i-1} \le x \le x_i,$$

gdje je $s_i(x)$ polinom trećeg stepena. Napišimo predstavljanje polinoma trećeg stepena u nešto prilagođenom obliku:

$$s_i(x) = a_i + b_i(x - x_i) + \frac{c_i}{2}(x - x_i)^2 + \frac{d_i}{6}(x - x_i)^3, \quad i = \overline{1,n}.$$

Imamo $4n$ slobodnih veličina $\{a_i, b_i, c_i, d_i\}_{i=1}^n$. Očito je

$$s_i'(x) = b_i + c_i(x - x_i) + \frac{d_i}{2}(x - x_i)^2, \quad s_i''(x) = c_i + d_i(x - x_i).$$

Neprekidnost splajna i njegovog prvog i drugog izvoda je pod znakom pitanja samo u tačkama dodira $x_1, \ldots, x_{n-1}$ dva susjedna mala odsječka. Uslov b) možemo da rastavimo na uslove: b_1) $s(x_i - 0) = s(x_i + 0)$ tj. $s_i(x_i) = s_{i+1}(x_i)$, b_2) $s'(x_i - 0) = s'(x_i + 0)$ tj. $s_i'(x_i) = s_{i+1}'(x_i)$ i b_3) $s''(x_i - 0) = s''(x_i + 0)$ tj. $s_i''(x_i) = s_{i+1}''(x_i)$, $i = \overline{1,n-1}$. Tako da b) daje $3(n-1)$ uslova, a uslov interpolacije c) daje još $n+1$ uslova; ukupno $4n-2$ uslova. Broj stepeni slobode je $4n - (4n-2) = 2$.

Iskoristimo prvo b_1) i c) u obliku $s_i(x_{i-1}) = f_{i-1}$ i $s_i(x_i) = f_i$ za $i = \overline{1,n}$:

$$s_i(x_{i-1}) = f_{i-1} \qquad a_i + b_i(x_{i-1} - x_i) + \frac{c_i}{2}(x_{i-1} - x_i)^2 + \frac{d_i}{6}(x_{i-1} - x_i)^3 = f_{i-1}$$

$$a_i - b_i h + \frac{c_i}{2}h^2 - \frac{d_i}{6}h^3 = f_{i-1} \qquad i = 1, \ldots, n$$

$$s_i(x_i) = f_i \qquad a_i = f_i \qquad i = 1, \ldots, n \qquad \text{(svi } a_i \text{ su određeni i izlaze iz računa)}$$

$$b_i h - \frac{c_i}{2}h^2 + \frac{d_i}{6}h^3 = f_i - f_{i-1} \qquad i = 1, \ldots, n \qquad (1)$$

b_2): $\quad s_i'(x_i) = s_{i+1}'(x_i) \qquad b_i + c_i(x_i - x_i) + \dfrac{d_i}{2}(x_i - x_i)^2 = b_{i+1} + c_{i+1}(x_i - x_{i+1}) + \dfrac{d_{i+1}}{2}(x_i - x_{i+1})^2$

$$b_i = b_{i+1} - c_{i+1}h + \frac{d_i}{2}h^2 \qquad i = 1, \ldots, n-1 \qquad (2)$$

b$_3$): $s_i''(x_i) = s_{i+1}''(x_i)$ $\quad c_i + d_i(x_i - x_i) = c_{i+1} + d_{i+1}(x_i - x_{i+1})$

$$c_i = c_{i+1} - d_{i+1}h \qquad i = 1, \ldots, n-1 \qquad (3)$$

Na račun dva stepena slobode koji su ostali, treba dodati neka dva nova uslova. Postoji nekoliko običnih načina da se to izvede, a mi ćemo primijeniti jedan od načina. Možemo smatrati da funkcija f zadovoljava $f''(a) = 0$ i $f''(b) = 0$ (ponekad se kaže da su ovo dva granična uslova). Tada možemo da tražimo da bude $s''(a) = 0$ i $s''(b) = 0$. Dakle, dopunimo definiciju splajna sljedećim uslovom: d) $s''(a) = 0$ i $s''(b) = 0$. Izrazimo sada ova dva nova uslova preko naših oznaka. $s''(a) = 0$ znači $s_1''(x_0) = 0$ tj. $c_1 + d_1(x_0 - x_1) = 0$ ili

$$c_1 - d_1 h = 0. \qquad (4)$$

$s''(b) = 0$ znači $s_n''(x_n) = 0$ tj. $c_n + d_n(x_n - x_n) = 0$ ili

$$c_n = 0. \qquad (5)$$

Ima $3n$ jednačina (1)–(5) i $3n$ nepoznatih $\{b_i, c_i, d_i\}_{i=1}^n$. Zapišimo (3) i (4) zajedno kao

$$c_i = c_{i+1} - d_{i+1}h \qquad i = 0, \ldots, n-1, \qquad (6)$$

stavljajući da je

$$c_0 = 0, \qquad (7)$$

gdje je uvedena pomoćna promjenljiva c_0.

Razmotrimo sada (1), (2) i (6); kao i (7) i (5): $c_0 = 0$, $c_n = 0$. Želimo da dobijemo sistem u kome se pojavljuju samo $\{c_i\}_{i=0}^n$:

prepišimo (1) $\quad b_i h - \frac{c_i}{2}h^2 + \frac{d_i}{6}h^3 = f_i - f_{i-1} \qquad i = 1, \ldots, n$

pomjerimo indeks za jedan naviše $\quad b_{i+1}h - \frac{c_{i+1}}{2}h^2 + \frac{d_{i+1}}{6}h^3 = f_{i+1} - f_i \qquad i = 0, \ldots, n-1$

oduzmimo dvije relacije:

$$(b_{i+1} - b_i)h - \frac{c_{i+1} - c_i}{2}h^2 + \frac{d_{i+1} - d_i}{6}h^3 = f_{i+1} - 2f_i + f_{i-1} \qquad i = 1, \ldots, n-1 \qquad \Big/ : h$$

ovo se uvrsti u (2) u obliku $b_{i+1} - b_i = \ldots$ i isto $b_{i+1} - b_i = \ldots$

$$c_{i+1}h - \frac{d_{i+1}}{2}h^2 = \frac{c_{i+1} - c_i}{2}h - \frac{d_{i+1} - d_i}{6}h^2 + \frac{f_{i+1} - 2f_i + f_{i-1}}{h} \qquad i = 1, \ldots, n-1 \qquad (*_1)$$

(b_j su eliminisani)

S druge strane, (6) govori da je $\quad d_{i+1}h = c_{i+1} - c_i \qquad i = 0, \ldots, n-1 \qquad (*_2)$

isto $\quad d_i h = c_i - c_{i-1} \qquad i = 1, \ldots, n \qquad (*_3)$

Relacije $(*_2)$ i $(*_3)$ uvedu se u $(*_1)$:

$$c_{i+1}h - \frac{h}{2}(c_{i+1}-c_i) = \frac{c_{i+1}-c_i}{2}h - \frac{c_{i+1}-c_i-c_i+c_{i-1}}{6}h + \frac{f_{i+1}-2f_i+f_{i-1}}{h} \qquad \Big/ \cdot \frac{6}{h}$$

(i d_j su eliminisani)

$$6c_{i+1} - 3(c_{i+1}-c_i) = 3(c_{i+1}-c_i) - (c_{i+1}-2c_i+c_{i-1}) + \frac{6}{h^2}(f_{i+1}-2f_i+f_{i-1}) \qquad i = 1,\dots,n-1$$

$$\left.\begin{array}{l} c_{i-1}+4c_i+c_{i+1} = \dfrac{6}{h^2}(f_{i+1}-2f_i+f_{i-1}) \qquad i=1,\dots,n-1 \\ c_0 = 0, \quad c_n = 0 \end{array}\right\} \qquad (8)$$

Sistem linearnih jednačina (8) ima $n+1$ jednačina i ima $n+1$ nepoznatih $\{c_i\}_{i=0}^n$. Taj sistem ima jedinstveno rješenje, kako će kasnije biti pokazano (uzimamo obavezu).

Kada su $\{c_i\}_{i=0}^n$ određeni onda se na osnovu (6) neposredno izračunaju svi $\{d_i\}_{i=1}^n$. Zatim se na osnovu (1) dobiju svi $\{b_i\}_{i=1}^n$. Sada su svi polinomi $\{s_i(x)\}_{i=1}^n$ određeni. Mi smo riješili zadatak o interpolaciji pomoću kubnog splajna koji se opisuje uslovima a)–d). Bliže rečeno, dokazano je postojanje i jedinstvenost rješenja $s = s(x)$ i dat je postupak za njegovu konstrukciju: formirati sistem (8), riješiti taj sistem, itd.

Samo napominjemo da se ponekad umjesto para uslova $f''(a) = 0$, $f''(b) = 0$ (koji vode do tzv. prirodnog kubnog splajna) uzima par uslova $f'(a) = 0$, $f'(b) = 0$ ili $f(a) = f(b)$, $f'(a) = f'(b)$ ili nešto slično.

Kao primjer, prikazaćemo sistem (8) u slučaju $n = 6$. Može se smatrati da su nepoznate $\{c_i\}_{i=0}^n$ ili se može smatrati da su nepoznate $\{c_i\}_{i=1}^{n-1}$. Tako da je matrica sistema oblika 7×7 ili 5×5:

$$\begin{array}{l} c_0 = 0 \\ c_0+4c_1+c_2 = \dots \\ c_1+4c_2+c_3 = \dots \\ c_2+4c_3+c_4 = \dots \\ c_3+4c_4+c_5 = \dots \\ c_4+4c_5+c_6 = \dots \\ c_6 = 0 \end{array} \qquad \text{ili} \qquad \begin{array}{l} 4c_1+c_2 = \dots \\ c_1+4c_2+c_3 = \dots \\ c_2+4c_3+c_4 = \dots \\ c_3+4c_4+c_5 = \dots \\ c_4+4c_5 = \dots \end{array}$$

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{ili} \quad M = \begin{bmatrix} 4 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 4 & 1 \end{bmatrix}$$

Sada ćemo da razdužimo obavezu.

Definicija. Za kvadratnu matricu $A = [a_{ij}] \in R^{n\times n}$ kaže se da je dijagonalno dominantna (po redovima) ako za svako $i = 1,\dots,n$ važi nejednakost $|a_{ii}| > \sum_{j=1 \quad j\neq i}^{n} |a_{ij}|$.

Vidimo da je matrica M sistema (8) dijagonalno dominantna, jer je $4 > 1+1$.

Teorema. Ako je A dijagonalno dominantna onda je A regularna ($\det A \neq 0$).

Dokaz. Razmotrimo homogeni sistem linearnih jednačina $Ax = 0$, gdje je $x = (x_1, \ldots, x_n)$, tj. razmotrimo

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n = 0 \quad \text{za} \quad i = 1, \ldots, n.$$

Dopustimo da razmatrani sistem ima netrivijalno rješenje $(x_1, x_2, \ldots, x_n) \neq (0, 0, \ldots, 0)$. Među brojevima $|x_1|, |x_2|, \ldots, |x_n|$ uočimo najveći, odnosno ako ima više jednakih najvećih uočimo bilo koji od njih. Neka uočenom broju odgovara indeks ℓ odnosno neka je $|x_\ell|$ najveći; $|x_j| \leq |x_\ell|$ za $j = 1, \ldots, n$. Imamo da je $|x_\ell| > 0$ tj. da je $x_\ell \neq 0$. Dopušteno rješenje naravno zadovoljava svih n jednačina sistema, pa posebno zadovoljava i ℓ–tu jednačnu (jednačinu $i = \ell$), tako da je

$$a_{\ell 1}x_1 + \ldots + a_{\ell,\ell-1}x_{\ell-1} + a_{\ell\ell}x_\ell + a_{\ell.\ell+1}x_{\ell+1} + \ldots + a_{\ell n}x_n = 0$$

$$a_{\ell\ell}x_\ell = -a_{\ell 1}x_1 - \ldots - a_{\ell,\ell-1}x_{\ell-1} - a_{\ell.\ell+1}x_{\ell+1} - \ldots - a_{\ell n}x_n \qquad |a+b| \leq |a| + |b|$$

$$|a_{\ell\ell}| \cdot |x_\ell| \leq |a_{\ell 1}| \cdot |x_1| + \ldots + |a_{\ell,\ell-1}| \cdot |x_{\ell-1}| + |a_{\ell.\ell+1}| \cdot |x_{\ell+1}| + \ldots + |a_{\ell n}| \cdot |x_n|$$

$$|a_{\ell\ell}| \cdot |x_\ell| \leq |a_{\ell 1}| \cdot |x_\ell| + \ldots + |a_{\ell,\ell-1}| \cdot |x_\ell| + |a_{\ell.\ell+1}| \cdot |x_\ell| + \ldots + |a_{\ell n}| \cdot |x_\ell| \qquad / : |x_\ell|$$

$$|a_{\ell\ell}| \leq |a_{\ell 1}| + \ldots + |a_{\ell,\ell-1}| + |a_{\ell.\ell+1}| + \ldots + |a_{\ell n}|$$

A uslov dijagonalne dominantnosti za ℓ–ti red (za red $i = \ell$) govori upravo suprotno da je $|a_{\ell\ell}| > |a_{\ell 1}| + \ldots + |a_{\ell,\ell-1}| + |a_{\ell.\ell+1}| + \ldots + |a_{\ell n}|$. Dobili smo kontradikciju. Ne može da postoji netrivijalno (nenulto) rješenje. Homogeni sistem ima dakle samo nulto rješenje. Pokazali smo da je $\det A \neq 0$. Teorema je dokazana.

Zapaziti da je matrica sistema (8) trodijagonalna; to je jedno značajno dobro svojstvo razmatrane numeričke metode.

Definicija. Za kvadratnu matricu $A = [a_{ij}] \in R^{n\times n}$ kaže se da je trodijagonalna ako je ispunjen sljedeći uslov: $a_{ij} \neq 0 \Rightarrow j = i - 1$ ili $j = i$ ili $j = i + 1$.

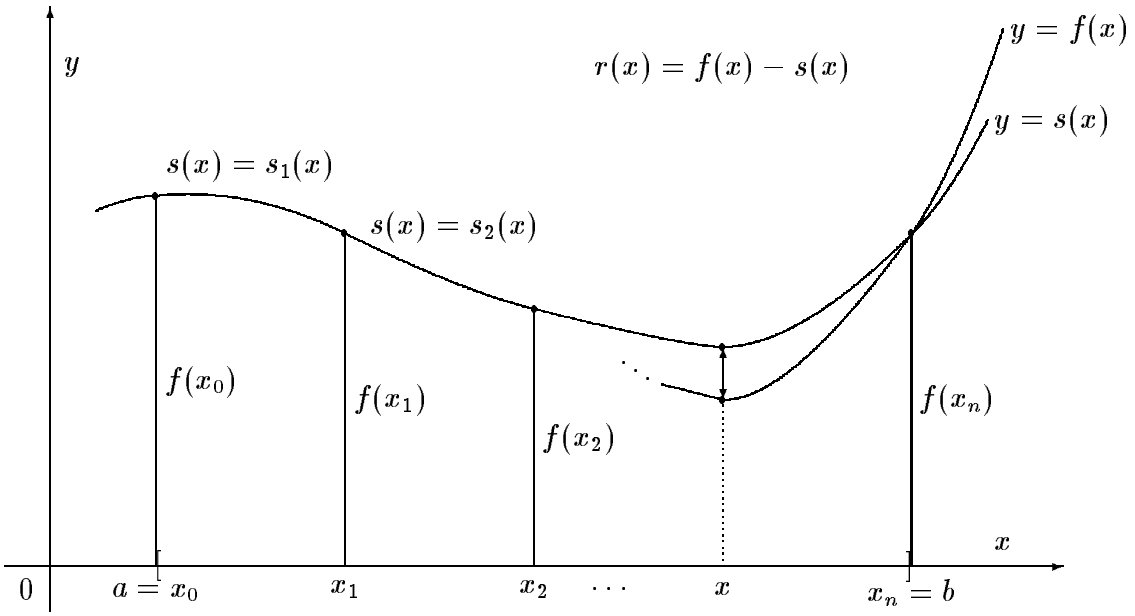
Ako je matrica sistema linearnih jednačina trodijagonalna onda to omogućava da se sistem znatno lakše (brže) riješi postupkom uzastopne eliminacije nepoznatih. Poznato je sljedeće o vremenskom trošku rješavanja sistema linearnih jednačina oblika $n \times n$; trošak se mjeri potrebnim brojem izvršenih aritmetičkih operacija. Za opšti ili puni sistem trošak iznosi $O(n^3)$, a za trodijagonalni iznosi svega $O(n)$.

Sljedeću teoremu o ocjeni greške navodimo bez dokaza.

Teorema. Neka $f \in C^4[a, b]$. Neka je $M_4 = \max\limits_{a\leq x\leq b} |f^{IV}(x)|$. Tada važe sljedeće nejednakosti:

$$\max_{a\leq x\leq b} |f(x) - s(x)| \leq M_4 h^4, \quad \max_{a\leq x\leq b} |f'(x) - s'(x)| \leq M_4 h^3 \quad \text{i} \quad \max_{a\leq x\leq b} |f''(x) - s''(x)| \leq M_4 h^2.$$

Vidimo da za svako fiksirano $x \in [a, b]$ važi relacija $s(x) \to f(x)$ kad $h \to 0$ odnosno kad $n \to \infty$. Zato se kaže da razmatrana numerička metoda konvergira; kad korak mreže $h \to 0$ onda greška metode $r(x) = f(x) - s(x) \to 0$. Kaže se da metoda ima četvrti red ili stepen konvergencije, budući da je $r(x) = O(h^4)$ kad $h \to 0$. Značajno je i što $s'(x) \to f'(x)$, a isto tako i što $s''(x) \to f''(x)$. Drugim riječima, izvod splajna može dobro da posluži za aproksimaciju izvoda funkcije, a drugi izvod splajna za aproksimaciju drugog izvoda funkcije.


y
$r(x) = f(x) - s(x)$
$y = f(x)$
$y = s(x)$
$s(x) = s_1(x)$
$s(x) = s_2(x)$
$f(x_0)$
$f(x_1)$
$f(x_2)$
$f(x_n)$
x
0
$a = x_0$
x_1
x_2
$\cdots$
x
$x_n = b$

1.9. NUMERIČKO DIFERENCIRANJE

Na osnovu date tablice vrijednosti funkcije treba procijeniti vrijednost izvoda funkcije u nekoj tački. U ovoj sekciji biće izvedene formule za numeričko diferenciranje, biće dobijen odgovarajući izraz za grešku i biće navedeni primjeri formula za numeričko diferenciranje. Neka bude $n \geq 1$ i razmotrimo na realnoj osi n međusobno različitih tačaka $x_1, \ldots, x_n$. Neka su poznate vrijednosti funkcije f u čvorovima tj. neka su date brojne vrijednosti $f(x_i) = f_i \in R$. Neka $x \in R$ i neka je $k \geq 1$. Treba procijeniti $f'(x)$ ili uopšte $f^{(k)}(x)$. Stavimo $a = \min(x, x_1, \ldots, x_n)$ i $b = \max(x, x_1, \ldots, x_n)$. Pretpostavlja se da $f \in C^{n+k}[a, b]$.

Neka je $L_n = L_n(x)$ L.i.p. za f po mreži $\{x_i\}_{i=1}^n$. Mi ćemo uzeti da je $f'(x) \approx L_n'(x)$ i uopšte da je $f^{(k)}(x) \approx L_n^{(k)}(x)$. Nema teškoća da se izvodi polinoma $L_n(x)$ izračunaju tačno. Kaže se da se razmatraju formule za numeričko diferenciranje koje su interpolacionog tipa. Kasnije će biti dati primjeri konkretnih realizacija takvih formula, u zavisnosti od mreže, od x i od k (u zavisnosti od podataka).

Prelazimo na ocjenu greške formule $f^{(k)}(x) \approx L_n^{(k)}(x)$ tj. na dobijanje izraza za grešku $r(x) = f^{(k)}(x) - L_n^{(k)}(x)$. Odranije iz sekcije 1.4. znamo formule (1) i (2), kako su tamo bile numerisane:

$$f(x) - L_n(x) = f(x; x_1; \ldots; x_n)\omega_n(x) \quad \text{i} \quad f(x; x_1; \ldots; x_n) = \frac{f^{(n)}(\xi)}{n!}, \quad \xi \in [a, b];$$

oznaka $\omega_n(x) = \prod_{i=1}^n (x - x_i)$. Primijenimo na $f(x) - L_n(x) = f(x; x_1; \ldots; x_n)\omega_n(x)$ Lajbnicovu formulu za k–ti izvod proizvoda $(u(x)v(x))^{(k)} = \sum_{j=0}^k \binom{k}{j} u^{(j)}(x)v^{(k-j)}(x)$:

$$r(x) = f^{(k)}(x) - L_n^{(k)}(x) = \sum_{j=0}^{k} \binom{k}{j} (f(x; x_1; \ldots; x_n))^{(j)} \omega_n^{(k-j)}(x).$$

Razmotrimo izraz

$$A = f(x; x + \varepsilon; \ldots; x + j\varepsilon; x_1; \ldots; x_n).$$

S jedne strane, A predstavlja podijeljenu razliku reda j od funkcije $f(t; x_1; \ldots; x_n)$ po mreži čvorova $t = x$, $t = x + \varepsilon$, ..., $t = x + j\varepsilon$. Formula oblika formule (2) iz sekcije 1.4. govori o vezi podijeljenih razlika i izvoda funkcije i daje nam $A = \frac{1}{j!}(f(\xi(\varepsilon); x_1; \ldots; x_n))^{(j)}$, gdje je $x < \xi(\varepsilon) < x + j\varepsilon$. Kad $\varepsilon \to 0$ onda je očito $\lim_{\varepsilon \to 0} \xi(\varepsilon) = x$ tako da postoji i $\lim_{\varepsilon \to 0} A$;

$$\lim_{\varepsilon \to 0} A = \frac{1}{j!}(f(x; x_1; \ldots; x_n))^{(j)}.$$

S druge strane, A predstavlja podijeljenu razliku reda $n + j$ funkcije $f = f(t)$ po naznačenih $n + j + 1$ čvorova $t = x$, $t = x + \varepsilon$, ..., $t = x_n$. Opet upotrebimo poznatu formulu o vezi podijeljenih razlika i izvoda funkcije: $A = \frac{1}{(n+j)!} f^{(n+j)}(\xi_j(\varepsilon))$, gdje je $a \leq \xi_j(\varepsilon) \leq b$. Uvedimo oznake $m_1 = \min_{t \in [a,b]} f^{(n+j)}(t)$ i $m_2 = \max_{t \in [a,b]} f^{(n+j)}(t)$. Tako da je $m_1 \leq f^{(n+j)}(\xi_j(\varepsilon)) \leq m_2$. Maločas smo pokazali da postoji $\lim_{\varepsilon \to 0} A$, tako da znamo da postoji i $\lim_{\varepsilon \to 0} f^{(n+j)}(\xi_j(\varepsilon))$. Vidimo da je $m_1 \leq \lim_{\varepsilon \to 0} f^{(n+j)}(\xi_j(\varepsilon)) \leq m_2$. Funkcija $f^{(n+j)}$ je neprekidna na odsječku

$[a,b]$. Poznata je teorema da neprekidna na odsječku funkcija uzima sve svoje međuvrijednosti $m \in [m_1, m_2]$. Zato postoji tačka $\xi_j \in [a,b]$ takva da je

$$\lim_{\varepsilon \to 0} A = \frac{1}{(n+j)!} f^{(n+j)}(\xi_j).$$

Uporediti dva prikazivanja za $\lim_{\varepsilon \to 0} A$. Prema tome

$$r(x) = f^{(k)}(x) - L_n^{(k)}(x) = \sum_{j=0}^{k} \binom{k}{j} \frac{j!}{(n+j)!} f^{(n+j)}(\xi_j) \omega_n^{(k-j)}(x), \tag{1}$$

$\xi_j \in [a,b]$ za $j = 0, \ldots, k$.

Formula (1) važi za svako $x \in [a,b]$ tj. ona važi i kada se tačka x poklapa sa nekim čvorom (kada je $x = x_i$ za neko i), s tim da prethodno izvođenje treba da bude malo prilagođeno ako je x čvor. Upravo, ako je x čvor onda treba staviti $A = f(x_i + \varepsilon; x_i + 2\varepsilon; \ldots; x_i + (j+1)\varepsilon; x_1; \ldots; x_n)$.

$$f^{(k)}(x) - L_n^{(k)}(x) = \sum_{j=0}^{k} \frac{k!}{(k-j)!(n+j)!} f^{(n+j)}(\xi_j) \omega_n^{(k-j)}(x), \quad \text{jer je } \binom{k}{j} = \frac{k!}{j!(k-j)!}$$

$$\left| f^{(k)}(x) - L_n^{(k)}(x) \right| \le \sum_{j=0}^{k} \frac{k!}{(k-j)!(n+j)!} M_{n+j} \left| \omega_n^{(k-j)}(x) \right|, \quad \text{gdje je } M_{n+j} = \max_{t \in [a,b]} \left| f^{(n+j)}(t) \right|$$

Prelazimo na primjere. Biće navedena tri primjera, sva tri se odnose na slučaj ekvidistantne mreže čvorova.

1. Formula za prvi izvod u čvoru, jednostrana formula. Definišimo podatke koji odgovaraju ovom specijalnom slučaju. Imamo $n+1$ čvor $x_i = x_0 + ih$, gdje je $i = 0, \ldots, n$, a treba da se procijeni $f'(x_0)$. Treba napisati izraz za odgovarajući L.i.p. Zatim treba diferencirati taj izraz i onda naravno treba uvrstiti $x = x_0$. Onda još samo treba konkretizovati formulu za grešku (1).

Tokom rada, prikazaćemo L.i.p. $L_{n+1} = L_{n+1}(x)$ u obliku I Nj.i.f, uz upotrebu obične smjene $x = x_0 + ht$:

$$L_{n+1}(x) = L_{n+1}(x_0 + ht) =$$

$$f_0 + \Delta f_0 \cdot t + \frac{1}{2!} \cdot \Delta^2 f_0 \cdot t(t-1) + \ldots + \frac{1}{n!} \cdot \Delta^n f_0 \cdot t(t-1)\ldots(t-n+1);$$

$$\frac{dy}{dt} = h \frac{dy}{dx} \qquad L'_{n+1}(x) = L'_{n+1}(x_0 + ht) =$$

$$\frac{1}{h} \left(\Delta f_0 \cdot \frac{d}{dt} t + \frac{1}{2!} \cdot \Delta^2 f_0 \cdot \frac{d}{dt} t(t-1) + \ldots + \frac{1}{n!} \cdot \Delta^n f_0 \cdot \frac{d}{dt} t(t-1)\ldots(t-n+1) \right);$$

$$t = 0 \qquad L'_{n+1}(x_0) = \frac{1}{h} \left(\Delta f_0 - \frac{1}{2} \Delta^2 f_0 + \ldots + \frac{(-1)^{n+1}}{n} \Delta^n f_0 \right) \tag{2}$$

Za $k = 1$ (1) glasi

$$f'(x) - L'_n(x) = \frac{1}{n!} f^{(n)}(\xi_0) \omega'_n(x) + \frac{1}{(n+1)!} f^{(n+1)}(\xi_1) \omega_n(x), \quad \xi_0, \xi_1 \in [a,b].$$

Da se ocijeni greška za formulu (2), u formuli (1) se stavi $k=1$ i $x=x_0$, a n se zamijeni sa $n+1$ i takođe $\omega_n(x)$ se zamijeni sa $\omega_{n+1}(x)=\prod_{i=0}^{n}(x-x_i)$. Vidi se da je $\omega_{n+1}(x_0)=0$. Izračunati $\omega'_{n+1}(x)$ i $\omega'_{n+1}(x_0)$. Kada se sprovede jednostavni račun onda se dobije

$$f'(x_0)-L'_{n+1}(x_0)=\frac{(-1)^n}{n+1}f^{(n+1)}(\xi)h^n,\quad \text{za neko } \xi\in[x_0,x_n]. \tag{3}$$

Tako da (2) predstavlja formulu za numeričko diferenciranje, u smislu $f'(x_0)\approx L'_{n+1}(x_0)$, a (3) izražava njenu grešku. Navedimo neke konkretne slučajeve formule (2):

$$n=1\qquad f'(x_0)\approx\frac{1}{h}\Delta f_0=\frac{f(x_1)-f(x_0)}{h} \tag{4}$$

ili $\quad f'(x_i)=\frac{1}{h}\big(f(x_{i+1})-f(x_i)\big)-\frac{h}{2}f''(\xi),\ \ x_i\le\xi\le x_{i+1},\ \ \text{ako } f\in C^2[x_i,x_{i+1}]$

$$n=2\qquad f'(x_0)\approx\frac{1}{h}\left(\Delta f_0-\frac{1}{2}\Delta^2 f_0\right)=\frac{-f(x_2)+4f(x_1)-3f(x_0)}{2h}$$

2. Formula za prvi izvod u tački koja se nalazi na sredini između dva čvora, simetrična formula. Razmotrimo mrežu čvorova $x_i=x_0+ih$, gdje je $i=\overline{-(\ell-1),\ell}$ i razmotrimo tačku $\overline{x}=x_{1/2}=x_0+\frac{h}{2}$. Recimo. kada je $\ell=2$ onda mrežu čine čvorovi $\{x_{-1},x_0,x_1,x_2\}$. Izostavlja se odgovarajući račun. Važi:

$$f'(\overline{x})=f'\left(x_0+\frac{h}{2}\right)=\frac{1}{h}\sum_{j=0}^{\ell-1}\frac{(-1)^j}{(2j+1)!}\left(\frac{1}{2}\cdot\frac{3}{2}\cdot\ldots\cdot\left(j-\frac{1}{2}\right)\right)^2\delta^{2j+1}f_{1/2}+$$

$$\frac{(-1)^\ell}{(2\ell+1)!}\left(\frac{1}{2}\cdot\frac{3}{2}\cdot\ldots\cdot\left(\ell-\frac{1}{2}\right)\right)^2 f^{(2\ell+1)}(\xi)h^{2\ell}$$

u smislu tačno = približno + greška.
Za grešku: $\omega_{2\ell}(x)=\prod^{\ell}$ $\overline{x})=0$. Neki slučajevi:

$$i=-(\ell-1)(x-x_i);\ \omega'_{2\ell}($$

$$\ell=1\qquad f'\left(x_0+\frac{h}{2}\right)\approx\frac{1}{h}\delta f_{1/2}=\frac{1}{h}\big(f(x_1)-f(x_0)\big) \tag{5}$$

ili (umjesto $\frac{h}{2}$ piši h) $\quad f'(x_i)=\frac{1}{2h}\big(f(x_{i+1})-f(x_{i-1})\big)-\frac{h^2}{6}f'''(\xi)$

$$\ell=2\qquad f'\left(x_0+\frac{h}{2}\right)\approx\frac{1}{h}\left(\delta f_{1/2}-\frac{1}{24}\delta^3 f_{1/2}\right)=$$

$$\frac{1}{24h}\big(-f(x_2)+27f(x_1)-27f(x_0)+f(x_{-1})\big)$$

3. Drugi izvod u čvoru, simetrična formula. $f''(x_0)$ aproksimira se preko vrijednosti funkcije u čvorovima $x_{-\ell}$, $x_{-(\ell-1)}$, …, x_ℓ (ima ih $2\ell+1$), gdje je $x_i=x_0+ih$ za $i=-\ell,-(\ell-1),\ldots,\ell$

$$f''(x_0)=\frac{1}{h^2}\sum_{j=1}^{\ell}\frac{2(-1)^{j-1}}{(2j)!}((j-1)!)^2\delta^{2j}f_0+\frac{2(-1)^\ell}{(2\ell+2)!}(\ell!)^2 f^{(2\ell+2)}(\xi)h^{2\ell}$$

tačno = približno + greška. Za grešku:

$$\omega_{2\ell+1}(x) = \prod_{i=-\ell}^{\ell}(x - x_i), \quad \omega_{2\ell+1}(x_0) = 0,$$

$$\omega'_{2\ell+1}(x_0) = (-1)^\ell(\ell!)^2 h^{2\ell}, \quad \omega''_{2\ell+1}(x_0) = 0$$

$$\ell = 1 \qquad f''(x_0) \approx \frac{1}{h^2}\delta^2 f_0 = \frac{1}{h^2}\Big(f(x_1) - 2f(x_0) + f(x_{-1})\Big) \tag{6}$$

ili $\quad f''(x_i) = \dfrac{1}{h^2}\Big(f(x_{i+1}) - 2f(x_i) + f(x_{i-1})\Big) - \dfrac{h^2}{12}f^{IV}(\xi)$

$$\ell = 2 \quad f''(x_0) \approx \frac{1}{h^2}\left(\delta^2 f_0 - \frac{1}{12}\delta^4 f_0\right) =$$

$$\frac{1}{12h^2}\Big(-f(x_2) + 16f(x_1) - 30f(x_0) + 16f(x_{-1}) - f(x_{-2})\Big)$$

U zaključku, ako se po ekvidistantnoj mreži koja se sastoji od n čvorova procjenjuje k–ti izvod funkcije u nekoj tački x onda za grešku važi relacija $r(x) = O(h^{n-k})$, za male vrijednosti koraka mreže h. Međutim, ako je tačka x postavljena simetrično u odnosu na čvorove onda se okolnosti poboljšavaju (greška se smanjuje); u tom slučaju važi $r(x) = O(h^{n-k+1})$, red aproksimacije se poveća za jedan. Takvo svojstvo imaju formule iz drugog i trećeg primjera, za $f'(\overline{x})$ i za $f''(x_0)$.

U zaključku, za formule (4)–(6) se kaže da predstavljaju osnovne formule za numeričko diferenciranje. Postoji lak i neposredan način da se one dokažu. Izvršiti razvoj funkcije f po Tejlorovoj formuli do odgovarajućeg izvoda u okolini tačke x. U okviru toga, i odgovarajući izrazi za grešku mogu da budu provjereni (izvedeni).

1.10. NESTABILNOST NUMERIČKOG DIFERENCIRANJA. TRI VRSTE GREŠKE U NUMERIČKIM METODAMA

Mi ćemo sada razmotriti pojam nestabilnosti formula za numeričko diferenciranje (nestabilnosti u odnosu na približnost ulaznih veličina) na jednom jednostavnom primjeru formule za numeričko diferenciranje (na jednostavnom modelu). Tako da izlaganje neće biti tehnički opterećeno. A sve karakteristike razmatrane pojave (nestabilnosti) lijepo se vide.

Neka imamo dva čvora x_0 i $x_1 = x_0 + h$ i dvije odgovarajuće vrijednosti funkcije $f(x_0) = f_0$ i $f(x_1) = f_1$ i neka treba da se procijeni $f'(x_0)$. Kako je rađeno:

$$f'(x_0) \approx \frac{1}{h}\Big(f_1 - f_0\Big) \tag{1}$$

$$f'(x_0) = \frac{1}{h}\Big(f_1 - f_0\Big) + r_1 \qquad r_1 = -\frac{1}{2}f''(\xi)h \qquad x_0 < \xi < x_1$$

$$|r_1| \le \frac{1}{2}M_2 h \qquad M_2 = \max_{t\in[a,b]}|f''(t)| \qquad x_0, x_1 \in [a,b] \qquad f \in C^2[a,b]$$

Za r_1 se kaže da predstavlja grešku ili grešku metode.

Neka su f_0 i f_1 ulazne veličine. Uzmimo sada da umjesto sa tačnim vrijednostima ulaznih veličina f_0 i f_1 raspolažemo samo sa odgovarajućim približnim vrijednostima f_0^* i f_1^*. Uvedimo oznake za dvije odgovarajuće greške. Neka bude $\varepsilon_0 = f_0 - f_0^*$ i $\varepsilon_1 = f_1 - f_1^*$. Neka je poznata granica greške ulaznih veličina u oznaci E i to $|\varepsilon_0| = |f_0 - f_0^*| \le E$ i $|\varepsilon_1| = |f_1 - f_1^*| \le E$. Za E se kaže da predstavlja mjeru greške ulaznih veličina (ulaznih podataka).

U datim okolnostima, mi možemo da efektivno izračunamo jedino broj $t^{**} = \frac{1}{h}\big(f_1^* - f_0^*\big)$, gdje u računu ušestvuju približne vrijednosti f_0^* i f_1^*. Tačan broj $t = f'(x_0)$ je nedostižan. Nedostižan je i približni broj $t^* = \frac{1}{h}\big(f_1 - f_0\big)$. Njegovu ulogu preuzima približni broj t^{**}, koji predstavlja numerički odgovor (predstavlja rezultat), budući da numerički odgovor glasi $t \approx t^{**}$.

Neka bude $r_2 = t^* - t^{**}$. Za r_2 se kaže da predstavlja grešku izazvanu približnošću ulaznih veličina. Ponekad se za r_2 kaže da predstavlja grešku zaokruživanja; jer se r_2 stvara kada su ulazne veličine date zaokružene, date samo približno. Ponekad se za r_2 kaže da predstavlja neotklonjivu grešku, imajući u vidu da nismo u stanju da otklonimo grešku ulaznih veličina (nismo u stanju saznati njihove tačne vrijednosti), tako da će ta greška "proći" kroz računski proces i (znači) odraziće se na krajnji numerički odgovor.

Na redu je procjena veličine r_2:

$$r_2 = t^* - t^{**} = \frac{1}{h}\big(f_1 - f_0\big) - \frac{1}{h}\big(f_1^* - f_0^*\big) = \frac{1}{h}\big(f_1 - f_1^*\big) - \frac{1}{h}\big(f_0 - f_0^*\big) = \frac{1}{h}\varepsilon_1 - \frac{1}{h}\varepsilon_0,$$

$$|r_2| \le \frac{1}{h}|\varepsilon_1| + \frac{1}{h}|\varepsilon_0| \le \frac{1}{h}E + \frac{1}{h}E = \frac{2E}{h}.$$

Ako je $E = 0$ onda je $r_2 = 0$. Što je E veće to je i r_2 veće.

Na redu je procjena greške numeričkog odgovora, procjena veličine $r = t - t^{**}$:

$$r = t - t^{**}, \qquad r = t - t^* + t^* - t^{**}, \qquad r = r_1 + r_2,$$

$$|r| \le |r_1| + |r_2|, \qquad |r| \le \frac{1}{2}M_2 h + \frac{2E}{h}.$$

Na r, r_1 i r_2 gledamo kao na funkcije od h.

Nije ispunjen uslov da $r_2 = r_2(h) \to 0$ kad $h \to 0$. Zato se kaže da je formula (1) nestabilna u odnosu na grešku ulaznih veličina.

Nacrtati grafik funkcije $g = g(x) = \frac{1}{2}M_2 x + \frac{2E}{x}$ za $x > 0$. Funkcija dostiže minimum za $x = 2\sqrt{E/M_2}$ a sama vrijednost minimuma iznosi $g(2\sqrt{E/M_2}) = 2\sqrt{M_2 E}$. Ako se kao korak h uzme $h = h_0 = 2\sqrt{E/M_2}$ onda se to može smatrati najpovoljnijim izborom koraka. Budući da je $|r(x)| \le g(x)$ to je $|r(h_0)| \le 2\sqrt{M_2 E}$.

Pogledajmo odnos između greške ulaznih veličina E i ukupne greške $r(h_0)$. Ako je $E = 10^{-2n}$ (ulazne veličine znamo sa $2n$ tačnih decimala) onda je greška jednaka 10^{-n}, grubo govoreći; $\sqrt{10^{-2n}} = 10^{-n}$. Dakle, od polaznog broja tačnih decimala, u rezultatu je ostalo (sačuvalo se) samo pola decimala tačnih, a pola se izgubilo. Drugim riječima, greška rezultata je znatno veća od greške ulaznih veličina; nepovoljna okolnost.

Prelazimo na opšti slučaj formule za numeričko diferenciranje. Greška formule za numeričko diferenciranje obično ima oblik

$$r = r_1 + r_2 \ \text{ sa } \ r_1 \sim C_1 h^{n-k} \ \text{ i } \ r_2 \sim C_2 h^{-k},$$

gdje se računa približna vrijednost k–tog izvoda u nekoj tački po ekvidistantnoj mreži čvorova; h je korak mreže, a čvorova ima ukupno n na broju. Uporedi sa primjerima iz prethodne sekcije. Tako da negativno svojstvo nestabilnosti imamo kod svake (kod praktično svake) formule za numeričko diferenciranje.

Pogledajmo jedan mali primjer. Neka bude $y = y(x) = e^x$ i $x_0 = 0$, $x_1 = 0{,}1$, $x_2 = 0{,}2$, $y_0 = y(x_0)$, $y_1 = y(x_1)$, $y_2 = y(x_2)$. Tada je $y_0 = 1$, $y_1 = 1{,}10517$, $y_2 = 1{,}22140$; prikazati u obliku tabele. Neka na osnovu nabrojanih podataka treba da bude procijenjen $y''(x_1)$. Primjenom formule iz prethodne sekcije dobijamo sljedeću procjenu:

$$\frac{y_2 - 2y_1 + y_0}{h^2} = \frac{1{,}22140 - 2 \cdot 1{,}10517 + 1}{0{,}1^2} = 1{,}106.$$

Sada imamo jednostavne opservacije. Ulazni podaci y_0, y_1 i y_2 imaju po šest značajnih cifara, a rezultat 1,106 ima svega četiri značajne cifre; došlo je do gubitka dvije značajne cifre. Ulazni podaci imaju grešku reda 10^{-5}, a rezultat ima grešku reda 10^{-3}. Dakle, rezultat ima znatno veću i apsolutnu i relativnu grešku od ulaznih podataka.

Prelazimo na dio: tri vrste greške u numeričkim metodama. Za bilo koju numeričku metodu, greška ili ukupna greška r računa se kao $r = r_1 + r_2 + r_3$, gdje se za treću komponentu r_3 kaže da predstavlja grešku računanja ili grešku operacija. Iz samog naziva je jasno kako se formira r_3. Naime, tokom računanja po datom obrascu redom se izvode naznačene aritmetičke operacije. Rezultat pojedine aritmetičke operacije je približan broj (ponekad se kaže zaokružen broj), makar da su argumenti operacije tačni brojevi. Tako da ukupna računanja po datom obrascu unose dodatnu grešku r_3.

Znamo da računar izvodi aritmetičke operacije (i druge operacije) nad realnim brojevima samo približno tačno.

Od metode do metode, treba voditi računa o sve tri komponente greške: greška metode + greška izazvana približnošću ulaznih veličina + greška računanja.

U zaključku, ako je metoda nestabilna u odnosu na grešku ulaznih veličina onda se to po pravilu posebno naglasi. Ako već ne možemo da izbjegnemo upotrebu takve metode onda treba dobro pratiti (kontrolisati) ponašanje greške r_2. S druge strane, za stabilnu metodu lako može biti da je r_2 zanemarljivo mala u odnosu na grešku metode r_1.

U zaključku, često je realno smatrati da je greška računanja r_3 zanemarljivo mala u odnosu na grešku metode r_1.

Na primjer, kod L.i.p. smo radili kao da je $r_2 = 0$ i $r_3 = 0$, a umjesto "greška metode" govorili smo jednostavno "greška".

1.11. POJAM PRIBLIŽNOG BROJA

Neka je a tačna vrijednost neke veličine i neka je a^* poznata približna vrijednost te veličine. Greškom ili apsolutnom greškom približnog broja a^* naziva se veličina $A(a^*) = a - a^*$. Relativnom greškom približnog broja a^* naziva se veličina $R(a^*) = \frac{A(a^*)}{|a^*|} = \frac{a-a^*}{|a^*|}$. Ponekad se u literaturi ove dvije veličine uzimaju po modulu, pa bi se stavilo $A(a^*) = |a-a^*|$ i $R(a^*) = \frac{|a-a^*|}{|a^*|}$. Kada piše recimo $a = 7{,}2 \pm 0{,}1$ onda to ima smisao $a^* = 7{,}2$ i $|A(a^*)| \leq 0{,}1$, drukčije rečeno $7{,}1 \leq a \leq 7{,}3$. Relativna greška $R(a^*)$ se često izražava u procentima. Znamo da realni brojevi zapisani u memoriji računara imaju granicu relativne greške 10^{-7}.

U stvarnosti, mi ne znamo $A(a^*)$ niti $R(a^*)$, već znamo samo neke njihove ocjene, a bolje je da su te dvije ocjene što bliže veličinama $A(a^*)$ i $R(a^*)$. Tako se definišu i granica apsolutne greške $\Delta(a^*)$ i granica relativne greške $\delta(a^*)$ približnog broja a^* kao ma koji brojevi koji ispunjavaju $|a - a^*| \leq \Delta(a^*)$ i $\frac{|a-a^*|}{|a^*|} \leq \delta(a^*)$.

Značajnom cifrom broja naziva se svaka cifra njegovog zapisa počev od prve nenulte cifre slijeva. Recimo, broj $a^* = 0{,}0\underline{3045}$ ima četiri značajne cifre, one su podvučene, a broj $a^* = 0{,}0\underline{3045000}$ ima sedam značajnih cifara. Za značajnu cifru se kaže da je sigurna ako apsolutna greška tog broja ne prevazilazi vrijednost pozicije (težinu dekadnog mjesta) koja odgovara toj cifri. Na primjer $a^* = 0{,}0\underline{3045}$ $\Delta(a^*) = 3 \cdot 10^{-6}$ broj a^* ima četiri sigurne cifre, one su podvučene. Na primjer $a^* = 0.0\underline{3045}00$ $\Delta(a^*) = 7 \cdot 10^{-7}$ broj a^* ima pet sigurnih cifara. Ponekad se značajna cifra naziva sigurnom (u užem smislu) samo ako apsolutna greška ne prevazilazi polovinu (polovinu jedinice) vrijednosti odgovarajuće pozicije.

Lako se vidi da je relativna greška približnog broja blisko povezana sa brojem njegovih sigurnih cifara: ako broj ima k sigurnih cifara onda njegova relativna greška iznosi 10^{-k}, grubo govoreći.

Ako je približan broj prosto napisan sa recimo 3 decimale onda se podrazumijeva da je njegova granica greške 10^{-3} ili $\frac{1}{2} \cdot 10^{-3}$ (konvencija).

1.12. GREŠKA FUNKCIJE

Ako je $a = 4 \pm 0{,}1$ kolika se greška čini kada se kaže da je $\sqrt{a} \approx 2$? Ako je $x_1 = x_1^* \pm \Delta(x_1^*)$ i $x_2 = x_2^* \pm \Delta(x_2^*)$ kolika se greška čini kada se uzme da približan broj $f(x_1^*, x_2^*)$ zamjenjuje tačnu vrijednost $f(x_1, x_2)$, recimo da je $f(x, y) = x^2 y$. U nastavku će biti izvedena formula koja daje izraz za odgovarajuću grešku. Kaže se da je to formula za grešku funkcije, da je to formula za grešku u slučaju kada su argumenti funkcije približni brojevi.

Pogledajmo prvo slučaj funkcije od jedne promjenljive f: $[a, b] \to R$. Neka je $x \in [a, b]$ tačna vrijednost argumenta, neka je $x^* \in [a, b]$ raspoloživa približna vrijednost i neka je $\Delta(x^*)$ granica greške tj. neka je $|x - x^*| \leq \Delta(x^*)$. Tada je $f(x)$ tačna vrijednost funkcije, a $f(x^*)$ je približna vrijednost. Neka bude

$$A = f(x) - f(x^*).$$

Za A se kaže da je greška funkcije. Naš zadatak je da ocijenimo A. Po Lagranžovoj teoremi (po formuli o konačnim priraštajima) imamo

$$A = f'(\xi) \cdot (x - x^*), \quad \xi = x^* + \theta \cdot (x - x^*), \quad 0 < \theta < 1$$

$$|A| = |f'(\xi)| \cdot |x - x^*|$$

$$|A| \leq M_1 \cdot |x - x^*|, \quad M_1 = \sup_{t \in G} |f'(t)|, \quad G = [x^* - \Delta(x^*), x^* + \Delta(x^*)]$$

Smatramo da je $G \subset [a, b]$ tako da se može svejedno uzeti da je $M_1 = \sup_{t \in [a,b]} |f'(t)|$:

$$|A| \leq M_1 \cdot \Delta(x^*). \tag{1}$$

Vidimo da formula (1) rješava postavljeni zadatak. Pretpostavili smo da $f \in C^1[a, b]$.

Dalje, razmotrimo veličinu

$$L = |f'(x^*)| \cdot \Delta(x^*).$$

Lako se vidi da važi nejednakost

$$|A| \leq L + o(\Delta(x^*)) \text{ kad } \Delta(x^*) \to 0.$$

Tako da L može da posluži kao zadovoljavajuća zamjena za grešku $|A|$, kada je greška argumenta mala. L se lako računa i zato se, u praktičnom radu, veličina L često uzima kao zamjena za grešku. Za L se kaže da predstavlja linearnu ocjenu za grešku funkcije.

Primjer $f(x) = \sqrt{x}$ $x^* = 4$ $\Delta(x^*) = 0{,}1$. Tada je $|A| \leq 0{,}0252$ i $L = 0.025$.

Pogledajmo slučaj funkcije od više promjenljivih. Vidjećemo da važe slične okolnosti. Neka su x_i^* približne vrijednosti za x_i sa granicama greške $\Delta(x_i^*)$, tako da je $|x_i - x_i^*| \leq \Delta(x_i^*)$ za $i = 1, \ldots, n$. Uvedimo oznaku $G = [x_1^* - \Delta(x_1^*), x_1^* + \Delta(x_1^*)] \times \ldots \times [x_n^* - \Delta(x_n^*), x_n^* + \Delta(x_n^*)]$. Neka je funkcija f definisana na skupu G. Pretpostavimo da $f \in C^1(G)$. Naš zadatak je da ocijenimo grešku $A = f(x_1, \ldots, x_n) - f(x_1^*, \ldots, x_n^*)$. U tom cilju, uvedimo parametar θ za odsječak od tačke $(x_1^*, \ldots, x_n^*)$ do tačke $(x_1, \ldots, x_n)$, tako da $\theta = 0$ i $\theta = 1$ odgovaraju početku odnosno kraju odsječka. Posmatrajmo funkciju f samo na tom odsječku. Primijenimo Lagranžovu teoremu. Kada se izvod u smjeru izrazi preko parcijalnih izvoda onda imamo

$$A = \sum_{i=1}^{n} \frac{\partial f(\xi_1, \ldots, \xi_n)}{\partial t_i} \cdot (x_i - x_i^*), \quad \xi_i = x_i^* + \theta(x_i - x_i^*), \quad i = 1, \ldots, n, \quad 0 < \theta < 1$$

$$|A| \leq \sum_{i=1}^{n} \left| \frac{\partial f(\xi_1, \ldots, \xi_n)}{\partial t_i} \right| \cdot |x_i - x_i^*|$$

$$|A| \leq \sum_{i=1}^{n} B_i \cdot \Delta(x_i^*), \quad B_i = \sup_{(t_1, \ldots, t_n) \in G} \left| \frac{\partial f(t_1, \ldots, t_n)}{\partial t_i} \right| \tag{2}$$

Vidimo da formula (2) rješava postavljeni zadatak.

Dalje, kao praktična ocjena za grešku funkcije uzima se veličina

$$L = \sum_{i=1}^{n} \left| \frac{\partial f(x_1^*, \ldots, x_n^*)}{\partial t_i} \right| \cdot \Delta(x_i^*),$$

ovo je tzv. linearna ocjena za grešku funkcije.

Razmotrimo dva važna specijalna slučaja.

1. $f(t_1, \ldots, t_n) = t_1 + \ldots + t_n$. Greška zbira jednaka je zbiru grešaka sabiraka.

Ako se sabira 10 približnih brojeva čije su granice greške po 10^{-4} onda će zbir imati granicu greške 10^{-3}.

2. $f(t_1,\ldots,t_n)=t_1\cdot\ldots\cdot t_n$. Relativna greška proizvoda približno je jednaka zbiru relativnih grešaka činilaca. Drukčije rečeno, linearna ocjena za relativnu grešku proizvoda jednaka je zbiru relativnih greški činilaca. Za dokaz, dovoljno je izračunati

$$L=\frac{\partial f(x_1,\ldots,x_n)}{\partial x_i}\cdot\Delta(x_i),\quad\text{odnosno}\quad\frac{L}{f}=\frac{L}{x_1\cdot\ldots\cdot x_n}.$$

Dokaz na drugi način (u slučaju $n=3$):

$$f=f(x_1,x_2,x_3)=x_1x_2x_3\quad f^*=f(x_1^*,x_2^*,x_3^*)=x_1^*x_2^*x_3^*\quad A=f-f^*$$

$$A=x_1x_2x_3\mp x_1^*x_2x_3\mp x_1^*x_2^*x_3-x_1^*x_2^*x_3^*=$$

$$(x_1-x_1^*)x_2x_3+(x_2-x_2^*)x_1^*x_3+(x_3-x_3^*)x_1^*x_2^*$$

$$\frac{A}{f^*}=\frac{x_1-x_1^*}{x_1^*}\cdot\frac{x_2x_3}{x_2^*x_3^*}+\frac{x_2-x_2^*}{x_2^*}\cdot\frac{x_3}{x_3^*}+\frac{x_3-x_3^*}{x_3^*}$$

$$\frac{A}{f^*}\approx\frac{x_1-x_1^*}{x_1^*}+\frac{x_2-x_2^*}{x_2^*}+\frac{x_3-x_3^*}{x_3^*}\quad\text{tj.}\quad R(f^*)\approx R(x_1^*)+R(x_2^*)+R(x_3^*)$$

Ako x_1^* ima 6 sigurnih cifara x_2^* 8 x_3^* 10 onda će proizvod $f^*=x_1^*x_2^*x_3^*$ imati 6 sigurnih cifara.

A razlike $=A$ umanjenika $+A$ umanjioca

R količnika $\approx R$ djeljenika $+R$ djelioca

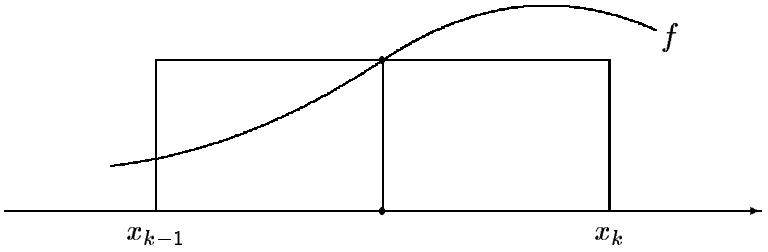
Obrnuti problem greške. Razmotrimo funkciju $f=f(x_1,\ldots,x_n)$ i neka je unaprijed data najveća dozvoljena granica greške funkcije Δf, npr. $\Delta f=0{,}5\cdot10^{-2}$. Treba odrediti granice grešaka argumenata $\Delta x_1,\ldots,\Delta x_n$ tako da greška funkcije bude manja ili jednaka od date gornje granice. Kako da se riješi? Postavljeni zadatak se lako rješava ako se upotrebi linearna ocjena za grešku funkcije L, iako nije savršena. Dakle, treba da bude ispunjeno: $(\forall i)\ |x_i-x_i^*|\le\Delta x_i\Rightarrow|f(x_1,\ldots,x_n)-f(x_1^*,\ldots,x_n^*)|\le\Delta f$. Prema tome, dovoljno je da bude $L\le\Delta f$ ili svejedno $\sum_{i=1}^n|\partial f(x_1^*,\ldots,x_n^*)/\partial x_i|\cdot\Delta x_i\le\Delta f$ ili $\sum_{i=1}^n|\partial f(x_1,\ldots,x_n)/\partial x_i|\cdot\Delta x_i\le\Delta f$. Drukčije zapisano: $\sum_{i=1}^n B_i\Delta x_i\le\Delta f$, gdje je uvedena oznaka $B_i=\sup|\partial f/\partial x_i|$ ili $B_i\ge\sup|\partial f/\partial x_i|$. Kako izabrati $\Delta x_1\ge0,\ldots,\Delta x_n\ge0$? Postoje razni pristupi, a popularna su sljedeća tri pristupa. Princip jednakih apsolutnih grešaka: $\Delta x_1=\ldots=\Delta x_n$. Princip jednakih relativnih grešaka: $\Delta x_1/|x_1|=\ldots=\Delta x_n/|x_n|$. Princip jednakih doprinosa: $B_1\Delta x_1=\ldots=B_n\Delta x_n$, tako da je u razmatranom slučaju očito $(\forall i)\ B_i\Delta x_i=\Delta f/n$ ili $B_i\Delta x_i\le\Delta f/n$. Odgovor treba dati u obliku $\Delta x_1=\ldots,\ldots,\Delta x_n=\ldots$. Pretpostavlja se da unaprijed raspolažemo sa grubim aproksimacijama za argumente $x_1,\ldots,x_n$, odnosno sa njihovim intervalima. Npr. $7{,}4<x_1<7{,}5$ i slično ostali x_i. U sva tri pristupa iskoristi se naravno nejednakost $B_1\Delta x_1+\ldots+B_n\Delta x_n\le\Delta f$ ili $=\Delta f$.

Praktično, uzima se $B_i=|\partial f(x_1,\ldots,x_n)/\partial x_i|$ ($x_1,\ldots,x_n$ iz grube aproksimacije).

2. NUMERIČKA INTEGRACIJA

2.1. TRI FORMULE

Neka je $[a,b]$ odsječak realne ose i neka je f funkcija $f\colon [a,b] \to R$. Razmotrimo integral $I = I(f) = \int_a^b f(x)dx$. Kao približna vrijednost služi zbir $S = S(f) = \sum_{k=1}^n c_k f(x_k)$. Za $c_k \in R$ kaže se da je koeficijent kvadraturne sume, za $x_k \in R$ kaže se da je čvor kvadraturne sume ($x_i \neq x_j$ za $i \neq j$). Ovdje je $n \geq 1$. Za zbir $\sum_{k=1}^n c_k f(x_k)$ ili svejedno $\sum_{k=1}^n c_k f_k$ kaže se da predstavlja kvadraturnu sumu, a za formulu oblika $I(f) \approx S(f)$ kaže se da predstavlja kvadraturnu formulu ili formulu za numeričku integraciju. Razmatra se i greška $R = R(f) = I(f) - S(f)$. Razmotrićemo tri formule koje imaju jasno geometrijsko tumačenje. Prelazimo na slučaj ravnomjerne mreže čiji je korak $h > 0$ i malo prilagođavamo oznake. Neka bude $x_0 = a$, $x_n = b$, $nh = b - a$ i $x_k = x_0 + kh$ za $k = 0, 1, \ldots, n$. Pisaćemo i $x_{k+1/2} = x_0 + (k + \frac{1}{2})h$.



a) Na malom odsječku $[x_{k-1}, x_k]$, površinu ispod grafika funkcije $y = f(x)$ zamijenimo površinom pravougaonika čija je osnovica $[x_{k-1}, x_k]$ i čija je visina $f(x_{k-1/2}) = f_{k-1/2}$:

$$I_k = \int_{x_{k-1}}^{x_k} f(x)dx \quad S_k = \int_{x_{k-1}}^{x_k} f(x_{k-1/2})dx = hf(x_{k-1/2}) \quad I_k \approx S_k \tag{1}$$

$$R_k = I_k - S_k = \int_{x_{k-1}}^{x_k} (f(x) - f(x_{k-1/2}))dx$$

$f \in C^2[a,b]$ Tejlorova formula:

$$f(x) = f(x_{k-1/2}) + f'(x_{k-1/2})(x - x_{k-1/2}) + \frac{1}{2!}f''(\xi_k(x))(x - x_{k-1/2})^2 \quad \xi_k(x) \in [x_{k-1}, x_k]$$

$$R_k = f'(x_{k-1/2}) \underbrace{\int_{x_{k-1}}^{x_k} (x - x_{k-1/2})dx}_{=0} + \frac{1}{2!}\int_{x_{k-1}}^{x_k} f''(\xi_k(x)) \underbrace{(x - x_{k-1/2})^2}_{\geq 0} dx$$

teorema o srednjoj vrijednosti za integrale: f neprekidna, g integrabilna, g stalnog znaka $\Rightarrow (\exists \xi)\ a < \xi < b$, $\int_a^b f(x)g(x)dx = f(\xi)\int_a^b g(x)dx$

$$R_k = \frac{1}{2!}f''(\xi_k) \underbrace{\int_{x_{k-1}}^{x_k} (x - x_{k-1/2})^2 dx}_{=h^3/12} \quad \xi_k \in [x_{k-1}, x_k] \quad R_k = \frac{1}{24}h^3 f''(\xi_k) \tag{2}$$

$$I = \int_a^b f(x)dx = \sum_{k=1}^n \int_{x_{k-1}}^{x_k} f(x)dx = \sum_{k=1}^n I_k \quad S = \sum_{k=1}^n S_k = h\sum_{k=1}^n f(x_{k-1/2}) \quad I \approx S \tag{3}$$

$$R = I - S = \sum_{k=1}^{n} R_k = \frac{1}{24}h^3 \sum_{k=1}^{n} f''(\xi_k) = \frac{1}{24}h^3 n \cdot \frac{1}{n}\sum_{k=1}^{n} f''(\xi_k)$$

$$m_1 = \min_{x\in[a,b]} f''(x) \quad m_2 = \max_{x\in[a,b]} f''(x) \quad m_1 \le f''(\xi_k) \le m_2$$

$$n \cdot m_1 \le \sum_{k=1}^{n} f''(\xi_k) \le n \cdot m_2 \quad m_1 \le \frac{1}{n}\sum_{k=1}^{n} f''(\xi_k) \le m_2$$

funkcija neprekidna na odsječku dostiže sve svoje međuvrijednosti (teorema o međuvrijednostima za neprekidnu funkciju)

$$\exists \xi \in [a,b] \quad \frac{1}{n}\sum_{k=1}^{n} f''(\xi_k) = f''(\xi)$$

$$R = \frac{1}{24}h^3 n f''(\xi) \quad R = \frac{1}{24}(b-a)h^2 f''(\xi) \tag{4}$$

$$|R| \le \frac{1}{24}(b-)h^2 M_2 \quad M_2 = \sup_{x\in[a,b]} |f''(x)|$$

Za (1) se kaže da je osnovna formula pravougaonika, za (2) se kaže da izražava ocjenu greške. Za (3) se kaže da predstavlja (sastavljenu) formulu pravougaonika, a za (4) se kaže da izražava njenu grešku. Iz (4): $R = O(h^2)$ kad $h \to 0$, pa se kaže da je stepen ili red tačnosti ili preciznosti formule pravougaonika jednak $N = 2$. Ako se umjesto $h = 0{,}1$ stavi $h = 0{,}05$ onda se h^2 svede na četvrtinu, a $f''(\xi)$ se promijeni na nepoznat način; po (4) se kaže da se polovljenjem koraka greška svede (grubo govoreći) na četvrtinu prethodne greške. F $\square$: $S = h\sum_{k=1}^{n} f_{k-1/2}$.

b) Kao približna vrijednost za I_k neka sada služi površina trapeza čija su tjemena $(x_{k-1}, 0)$, $(x_k, 0)$, (x_{k-1}, f_{k-1}) i (x_k, f_k). Stranici trapeza čija su tjemena (x_{k-1}, f_{k-1}) i (x_k, f_k) odgovara jednačina prave linije $y = L_2(x)$, gdje je $L_2(x)$ Lagranžov i. p. za f po mreži čvorova $\{x_{k-1}, x_k\}$. Tako da je očito $S_k = \int_{x_{k-1}}^{x_k} L_2(x)dx$. Poznata je formula za grešku interpolacije:

$$f(x) - L_2(x) = \frac{1}{2!}\omega_2(x) f''(\xi(x)), \quad \xi(x) \in [x_{k-1}, x_k], \quad x \in [x_{k-1}, x_k],$$

$$\omega_2(x) = (x - x_{k-1})(x - x_k), \quad f \in C^2[x_{k-1}, x_k] \text{ ili } f \in C^2[a,b]$$

$$I_k = \int_{x_{k-1}}^{x_k} f(x)dx \quad S_k = \frac{h}{2}\Big(f(x_{k-1}) + f(x_k)\Big) \text{ (površina trapeza)}$$

$$S_k = \frac{h}{2}\Big(f_{k-1} + f_k\Big) \quad I_k \approx S_k \tag{5}$$

$$R_k = I_k - S_k = \int_{x_{k-1}}^{x_k} f(x)dx - \int_{x_{k-1}}^{x_k} L_2(x)dx = \frac{1}{2!}\int_{x_{k-1}}^{x_k} f''(\xi(x)) \underbrace{(x - x_{k-1})(x - x_k)}_{\le 0} dx =$$

$$\frac{1}{2!}f''(\xi_k)\int_{x_{k-1}}^{x_k} (x - x_{k-1})(x - x_k)dx \quad \xi_k \in [x_{k-1}, x_k] \quad R_k = -\frac{1}{12}h^3 f''(\xi_k) \tag{6}$$

$$I = \int_a^b f(x)dx = \sum_{k=1}^{n}\int_{x_{k-1}}^{x_k} f(x)dx = \sum_{k=1}^{n} I_k$$

$$S = \sum_{k=1}^{n} S_k = h\left(\frac{1}{2}f_0 + \sum_{k=1}^{n-1} f_k + \frac{1}{2}f_n\right) \quad I \approx S \tag{7}$$

$$R = I - S = \sum_{k=1}^{n} R_k = -\frac{1}{12}h^3 \sum_{k=1}^{n} f''(\xi_k) = -\frac{1}{12}h^3 n f''(\xi), \quad a \le \xi \le b$$

$$R = -\frac{1}{12}(b-a)h^2 f''(\xi) \tag{8}$$

Trapezna formula (7) ima grešku reda veličine h^2 odnosno ima red tačnosti $N = 2$, v. (8). Ako se korak prepolovi onda čvorovi iz prvobitne mreže ostaju i u novoj mreži (vrijednosti funkcije se za njih ne računaju ponovo). Rezime o trapeznoj: $I \approx S = h(\frac{1}{2}f_0 + \sum_{k=1}^{n-1} f_k + \frac{1}{2}f_n)$. $R(\frac{h}{2}) \approx \frac{1}{4}R(h)$ (grubo govoreći).

Iz formule (8): $|R| \le \frac{1}{12}(b-a)M_2h^2$, gdje je $M_2 = \max_{a\le x\le b}|f''(x)|$. Ponekad nemamo izraz za $f''(x)$, tako da ne raspolažemo sa M_2. Tada se, radi praktične ocjene greške, može uzeti da veličina $M_2^* = h^{-2}\max_{0\le k\le n-2}|\Delta^2 f_k|$ posluži kao približna zamjena za M_2.

c) U slučaju Simpsonove formule, za obrazovanje približne vrijednosti S_k za mali odsječak $[x_{k-1}, x_k]$ iskoriste se tri podatka, iskoriste se vrijednosti funkcije u tačkama x_{k-1}, $x_{k-1/2}$ i x_k. Neka $L_3 = L_3(x)$ bude L. i. p. za funkciju f po mreži čvorova $\{x_{k-1}, x_{k-1/2}, x_k\}$. Površina ispod grafika $y = f(x)$ na odsječku $x_{k-1} \le x \le x_k$ približno je jednaka površini ispod parabole na tom istom odsječku. Sastaviti izraz $L_3(x) = \ldots$ (parabola). Izračunati integral od polinoma $\int_{x_{k-1}}^{x_k} L_3(x)dx$. Dobija se $\int_{x_{k-1}}^{x_k} L_3(x)dx = \frac{1}{6}(f_{k-1}+4f_{k-1/2}+f_k)$. U cilju ocjene greške, uvedimo u razmatranje i i. p. sa višestrukim čvorovima (Hermitov i. p.) $H_4 = H_4(x)$. Neka je $H_4 = H_4(x)$ H. i. p. za funkciju f definisan uslovima:

$$H_4(x_{k-1}) = f(x_{k-1}), \quad H_4(x_{k-1/2}) = f(x_{k-1/2}), \quad H_4(x_k) = f(x_k) \ \text{ i } \ H_4'(x_{k-1/2}) = f'(x_{k-1/2}).$$

Sastaviti izraz $H_4(x) = \ldots$. Izračunati $\int_{x_{k-1}}^{x_k} H_4(x)dx$. Dobiće se isti rezultat kao maločas. Dobiće se da je $\int_{x_{k-1}}^{x_k} H_4(x)dx = \frac{1}{6}(f_{k-1} + 4f_{k-1/2} + f_k)$.

Račun se za Simpsonovu formulu odvija po sličnom šablonu kao maločas u slučajevima a) i b): mali odsječak, greška, veliki odsječak, greška. Formula (9) će biti osnovna Simpsonova formula, a (11) će biti (sastavljena) Simpsonova formula. Pretpostavlja se da $f \in C^4[a,b]$:

$$I_k = \int_{x_{k-1}}^{x_k} f(x)dx \quad S_k = \frac{h}{6}\Big(f_{k-1} + 4f_{k-1/2} + f_k\Big) \quad I_k \approx S_k \tag{9}$$

$$R_k = I_k - S_k = \int_{x_{k-1}}^{x_k} f(x)dx - \int_{x_{k-1}}^{x_k} L_3(x)dx =$$

$$\int_{x_{k-1}}^{x_k} f(x)dx - \int_{x_{k-1}}^{x_k} H_4(x)dx = \int_{x_{k-1}}^{x_k} (f(x) - H_4(x))dx$$

poznati izraz za grešku za H. i. p. $f(x) - H_4(x) = \dfrac{1}{4!}f^{IV}(\xi(x))\omega_4(x)$

$$x_{k-1} \le \xi(x) \le x_k \quad x_{k-1} \le x \le x_k \quad \omega_4(x) = (x - x_{k-1})(x - x_{k-1/2})^2(x - x_k) \quad f \in C^4[x_{k-1}, x_k]$$

$$R_k = \frac{1}{4!}\int_{x_{k-1}}^{x_k} f^{IV}(\xi(x))\underbrace{\omega_4(x)}_{\le 0}dx = \frac{1}{4!}f^{IV}(\xi_k)\underbrace{\int_{x_{k-1}}^{x_k}\omega_4(x)dx}_{\text{izračunati}} \quad \xi_k \in [x_{k-1}, x_k]$$

$$R_k = -\frac{1}{2880}h^5 f^{IV}(\xi_k) \tag{10}$$

$$I = \int_a^b f(x)dx = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} f(x)dx = \sum_{k=1}^{n} I_k$$

$$S = \sum_{k=1}^{n} S_k = \frac{h}{6}\Big(f_0 + f_n + 4\sum_{k=1}^{n} f_{k-1/2} + 2\sum_{k=1}^{n-1} f_k\Big) \quad I \approx S \tag{11}$$

$$R = I - S = \sum_{k=1}^{n} R_k = -\frac{1}{2880}h^5 \sum_{k=1}^{n} f^{IV}(\xi_k) = -\frac{1}{2880}h^4(b-a)\cdot\frac{1}{n}\sum_{k=1}^{n} f^{IV}(\xi_k)$$

$$R = -\frac{1}{2880}(b-a)h^4 f^{IV}(\xi) \text{ za neko } \xi \in [a,b] \tag{12}$$

Vidimo da je $R = O(h^4)$ pa se kaže da je red tačnosti $N = 4$. Ako se korak prepolovi (radi dobijanja manje greške) onda se stare vrijednosti funkcije iskoriste. Polovljenje koraka svodi grešku na otprilike $\frac{1}{16}$ njene ranije vrijednosti.

Kvadraturne formule b) i c) se koriste više od a), jer su pogodne kod polovljenja koraka.

Kvadratura c) se koristi više od b), jer ima upadljivo veći (bolji) red tačnosti.

Mala dopuna oko c). Obično se izbjegava da u indeksu pišu polovine, obično se sa h označi rastojanje između dvije susjedne tačke. Drugim riječima, izvršićemo malu izmjenu u oznakama. Kada se to izvrši onda se dobije kako slijedi. Simpsonova formula i njen izraz za grešku glase kako slijedi:

$$I = \int_a^b f(x)dx \quad S = \frac{h}{3}\sum_{k=0}^{n} c_k f_k \quad c_0 = c_n = 1,\ c_k = 4 \text{ za } k \text{ neparno},\ c_k = 2 \text{ za } k \text{ parno}$$

$$h = \frac{b-a}{n} \quad x_k = a + kh,\ f_k = f(x_k),\ k = \overline{0,n} \quad \text{broj } n \text{ je obavezno paran} \quad I \approx S$$

$$R = I - S = -\frac{1}{180}(b-a)h^4 f^{IV}(\xi) \text{ za neko } \xi \in [a,b]$$

Na primjer, $\int_0^1 f(x)dx \approx \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + 2f_4 + \ldots + 4f_9 + f_{10})$.

2.2. RUNGEOVO PRAVILO ZA PRAKTIČNU OCJENU GREŠKE

Prvo ćemo izvesti drugi izraz za grešku trapezne formule. Vidjećemo da je taj izraz bolji od izraza iz prethodnog naslova. Iskoristićemo taj izraz da izvedemo tzv. Rungeovo pravilo ili Rungeovu formulu za procjenu greške trapezne formule. Za procjenu greške na Rungeov način ili po Rungeovom principu karakteristično je da se izvrše dva proračuna tj. da se dobiju dvije približne vrijednosti I_1 i I_2 za jedan te isti tačan broj I. I_1 je dobijeno sa korakom h a I_2 je dobijeno sa korakom $\frac{h}{2}$. I_2 je dobijeno po mreži koja ima više čvorova tako da I_2 ima manju grešku od I_1. Kako je približan broj I_2 bolji od I_1 to se I_2 usvaja kao numerički odgovor. Gruba približna vrijednost I_1 ima pomoćnu ulogu. Ona služi da se procijeni greška numeričkog odgovora tj. da se procijeni $R_2 = I - I_2$. Rungeov princip za dobijanje procjene greške može da bude primijenjen na bilo koju numeričku metodu čiji izraz za grešku ima oblik $\sim Ch^k$.

Uvedimo potrebne oznake. Razmotrimo funkciju f definisanu na odsječku $[a,b]$ i pretpostavimo da $f \in C^4[a,b]$. Neka bude $I = \int_a^b f(x)dx$. Izaberimo $n \geq 1$ i stavimo $h = \frac{b-a}{n}$.

Neka bude $I_1 = I(h) = h(\frac{1}{2}f(a) + \sum_{i=1}^{n-1} f(a+ih) + \frac{1}{2}f(b))$ trapezna formula sa korakom h i neka bude $I_2 = I(\frac{h}{2}) = \frac{h}{2}(\frac{1}{2}f(a) + \sum_{i=1}^{2n-1} f(a+i\cdot\frac{h}{2}) + \frac{1}{2}f(b))$ trapezna formula sa korakom $\frac{h}{2}$. Zapaziti da ranije izračunati broj I_1 može da posluži prilikom računanja I_2 jer čvorovi grube mreže pripadaju i detaljnijoj mreži; $I_2 = \frac{1}{2}I_1 + \frac{h}{2}\sum_{i=1}^{n} f(a+(2i-1)\cdot\frac{h}{2})$. Uvedimo i oznake za dvije greške: $R_1 = R(h) = I - I_1 = I - I(h)$, $R_2 = R(\frac{h}{2}) = I - I_2 = I - I(\frac{h}{2})$. Odgovor glasi $I \approx I_2$, samo treba R_2 da se ocijeni ($|R_2| \le \ldots$ ili $R_2 \approx \ldots$).

Razvijamo funkciju f po Maklorenovoj formuli do četvrtog izvoda:

$$f(x) = f(0) + xf'(0) + \frac{1}{2!}x^2f''(0) + \frac{1}{3!}x^3f'''(0) + \frac{1}{4!}x^4f^{IV}(\xi(x)).$$

Za grešku trapezne formule po malom odsječku imamo:

$$r = \int_{-h/2}^{h/2} f(x)dx - \frac{h}{2}\left(f\left(-\frac{h}{2}\right) + f\left(\frac{h}{2}\right)\right) =$$

$$\int_{-h/2}^{h/2} f(0)dx + \frac{1}{2!}f''(0)\int_{-h/2}^{h/2} x^2dx + \frac{1}{4!}\underbrace{\int_{-h/2}^{h/2} f^{IV}(\xi(x))x^4dx}_{\text{t o srednjoj v}} -$$

$$\frac{h}{2}\left(2f(0) + 2\cdot\frac{1}{2!}\left(\frac{h}{2}\right)^2 f''(0) + \frac{1}{4!}\left(\frac{h}{2}\right)^4 f^{IV}(\xi_1) + \frac{1}{4!}\left(\frac{h}{2}\right)^4 f^{IV}(\xi_2)\right)$$

$$\int_{-h/2}^{h/2} xdx = 0,\ \int_{-h/2}^{h/2} x^3dx = 0,\ \xi_1 = \xi\left(-\frac{h}{2}\right),\ \xi_2 = \xi\left(\frac{h}{2}\right)$$

$$r = hf(0) + f''(0)\cdot\frac{1}{3}\left(\frac{h}{2}\right)^3 + \frac{1}{4!}f^{IV}(\overline{\xi})\int_{-h/2}^{h/2} x^4dx -$$

$$hf(0) - \left(\frac{h}{2}\right)^3 f''(0) - \frac{1}{4!}\left(\frac{h}{2}\right)^5 f^{IV}(\xi_1) - \frac{1}{4!}\left(\frac{h}{2}\right)^5 f^{IV}(\xi_2) =$$

$$-\frac{1}{12}f''(0)h^3 + \rho,\ \ |\rho| \le cM_4h^5,\ \ M_4 = \max_{x\in[a,b]}|f^{IV}(x)|$$

Maločas smo izvršili translaciju po x–osi, samo radi lakšeg pisanja. Neka sada ulogu odsječka $[-\frac{h}{2}, \frac{h}{2}]$ preuzme odsječak $[x_{i-1}, x_i] = [a+(i-1)h, a+ih]$:

$$r_i = \int_{x_{i-1}}^{x_i} f(x)dx - \frac{h}{2}\Big(f(x_{i-1}) + f(x_i)\Big) = -\frac{1}{12}h^3f''(x_{i-1/2}) + \rho_i,$$

$$|\rho_i| \le cM_4h^5, \quad \rho_i = O(h^5), \quad \sum_{i=1}^{n}\rho_i = O(h^4)$$

Sabiranjem po $i = 1, \ldots, n$:

$$R(h) = I - I(h) = \sum_{i=1}^{n} r_i = -\frac{1}{12}h^2\cdot\underbrace{\sum_{i=1}^{n} hf''(x_{i-1/2})}_{\text{f }\square\text{ za } f''} + \sum_{i=1}^{n}\rho_i =$$

$$-\frac{1}{12}h^2\left(\int_a^b f''(x)dx-\frac{1}{24}(b-a)h^2f^{IV}(\xi)\right)+O(h^4)=-\frac{1}{12}h^2\int_a^b f''(x)dx+O(h^4)+O(h^4)$$

(v. formulu pravougaonika i izraz za grešku)

$$R(h)=Ch^2+O(h^4),\quad C=-\frac{1}{12}\int_a^b f''(x)dx,\quad C \text{ ne zavisi od } h.$$

Mi smo dobili:

$$R(h)=R_1=I-I(h)=Ch^2+O(h^4),\quad h\to 0,\quad C=-\frac{1}{12}\Big(f'(b)-f'(a)\Big). \tag{1}$$

Isto tako:

$$R\left(\frac{h}{2}\right)=R_2=I-I\left(\frac{h}{2}\right)=C\left(\frac{h}{2}\right)^2+O\left(\left(\frac{h}{2}\right)^4\right)=\frac{1}{4}Ch^2+O(h^4).$$

Zapaziti da su brojevi I_1 i I_2 efektivno poznati, što se ne može reći za I i C.
Mi imamo:

$$R_1=I-I_1=Ch^2+O(h^4)$$

$$R_2=I-I_2=\frac{1}{4}Ch^2+O(h^4)$$

oduzimanjem: $\quad I_2-I_1=\frac{3}{4}Ch^2+O(h^4).$

Ako je $C\neq 0$ onda je

$$\lim_{h\to 0}\frac{R_2}{I_2-I_1}=\lim_{h\to 0}\frac{\frac{1}{4}Ch^2+O(h^4)}{\frac{3}{4}Ch^2+O(h^4)}=\frac{1}{3}\ \text{ ili }\ R_2\sim\frac{1}{3}\Big(I_2-I_1\Big)\ \text{kad } h\to 0\ \text{ ili }\ R_2\approx\frac{1}{3}\Big(I_2-I_1\Big). \tag{2}$$

Ovo je završna–glavna formula. Broj na desnoj strani formule (2) je efektivno poznat. Za Rungeovu ocjenu (2) kaže se da je praktična zato što je ona efektivno ostvarljiva (njena dobra strana) i zato što je ona samo približno tačna (njena loša strana).

U slučaju $C=0$ može se pokazati da važi $|R_2|\leq\frac{1}{3}|I_2-I_1|$, za male vrijednosti h.

Pogledajmo mali primjer. Razmotrimo $\int_0^1 x^4dx$ i neka bude $n=1$ tj. $h=1$. Tada je $I_1=0{,}5000$, $I_2=0{,}2812$ i $\frac{1}{3}(I_2-I_1)=-0{,}0729$. A $R_2=-0{,}0812$.

Slijede razne dopune

1. Ako $I(\frac{h}{2})$ ne zadovoljava unaprijed traženu preciznost ($R(\frac{h}{2})$ prelazi unaprijed datu dozvoljenu grešku ε, tj. $\frac{1}{3}|I_2-I_1|>\varepsilon$) onda izračunati sa korakom $\frac{h}{4}$ novu približnu vrijednost $I(\frac{h}{4})$ i procijeniti njenu grešku na Rungeov način. Itd.

2. Neka $I(h)$ znači približnu vrijednost dobijenu po Simpsonovoj formuli. Tada važi $R(h)=I-I(h)=Ch^4+O(h^6)$, $h\to 0$, C ne zavisi od h. Isto tako važi $R(\frac{h}{2})\approx\frac{1}{15}(I(\frac{h}{2})-I(h))$ (procjena greške na Rungeov način). Pretpostavlja se da $f\in C^6[a,b]$.

3. Uopšte, ako je $R(h)\sim Ch^k$ onda važi $R(\frac{h}{2})\approx\frac{1}{2^k-1}(I(\frac{h}{2})-I(h))$.

2.3. ROMBERGOVA FORMULA

Mali primjer iz prethodnog naslova: $I(\frac{h}{2}) + \frac{1}{3}(I(\frac{h}{2}) - I(h)) = 0{,}2083$.

U prethodnom naslovu smo vidjeli da broj $b = \frac{1}{3}(I(\frac{h}{2}) - I(h))$ može da posluži kao približna vrijednost za $R(\frac{h}{2}) = I - I(\frac{h}{2})$. Ako je već $R(\frac{h}{2}) \approx b$ onda je očito $I \approx I(\frac{h}{2}) + b$. Dakle, neka broj b služi za popravku približnog broja $I(\frac{h}{2})$, neka sada numerički odgovor bude $I(\frac{h}{2}) + b$. Zanimljivo je da se broj $I(\frac{h}{2}) + b$ poklapa sa približnom vrijednošću za I izračunatom po Simpsonovoj formuli sa korakom $\frac{h}{2}$ tj. po mreži čvorova $a, a+\frac{h}{2}, a+h, \ldots, b$; uvjeriti se neposrednim računom. Nije bitno što odgovara baš Simpsonovoj formuli već je bitno što odgovara formuli čiji je stepen tačnosti (četvrti) veći od stepena tačnosti polazne trapezne formule (drugi).

Jedno sredstvo za dobijanje preciznije približne vrijednosti određenog integrala $I = \int_a^b f(x)\,dx$ jeste da se umjesto koraka h primijeni korak $\frac{h}{2}$, zatim $\frac{h}{4}$, itd. Drugo sredstvo bilo bi da se umjesto kvadraturne formule drugog stepena tačnosti primijeni formula četvrtog reda tačnosti, zatim šestog reda, itd. Kod Rombergove formule usaglašemo se koriste jedno i drugo sredstvo, s tim da se prelazak na formulu višeg reda tačnosti vrši pomoću popravke ranije izračunate približne vrijednosti.

U prvom dijelu izlaganja biće izveden izraz za grešku trapezne formule $R(h)$. Upravo, biće pokazano da se greška $R(h)$ razlaže po parnim stepenima h. U drugom dijelu izlaganja biće izvedena formula za popravku.

Prvi dio izlaganja. Neka bude $nh = b - a$ i $x_i = a + ih$. Imamo:

$$R(h) = I - I(h) = \int_a^b f(x)dx - h\left(\frac{1}{2}f(x_0) + \sum_{i=1}^{n-1} f(x_i) + \frac{1}{2}f(x_n)\right) =$$

$$\sum_{i=1}^{n}\left[\int_{x_{i-1}}^{x_i} f(x)dx - h\left(\frac{1}{2}f(x_{i-1}) + \frac{1}{2}f(x_i)\right)\right]$$

razviti funkciju f po Tejlorovoj formuli oko tačke $x = x_{i-1/2}$ do šestog izvoda

$$R(h) = \sum_{i=1}^{n}\Bigg[\frac{1}{24}h^3 f''(x_{i-1/2}) + \frac{1}{1920}h^5 f^{IV}(x_{i-1/2}) + \frac{1}{6!}f^{VI}(\overline{\xi_i})\int_{-h/2}^{h/2} x^6 dx -$$

$$h\left(\frac{1}{2!}\left(\frac{h}{2}\right)^2 f''(x_{i-1/2}) + \frac{1}{4!}\left(\frac{h}{2}\right)^4 f^{IV}(x_{i-1/2}) +\right.$$

$$\left.\frac{1}{2}\cdot\frac{1}{6!}\left(\frac{h}{2}\right)^6 f^{VI}(\xi_{1i}) + \frac{1}{2}\cdot\frac{1}{6!}\left(\frac{h}{2}\right)^6 f^{VI}(\xi_{2i})\right)\Bigg] =$$

$$-\frac{1}{12}h^2\underbrace{\sum_{i=1}^{n} hf''(x_{i-1/2})}_{\text{f } \square \text{ za } f''} - \frac{1}{480}h^4\underbrace{\sum_{i=1}^{n} hf^{IV}(x_{i-1/2})}_{\text{f } \square \text{ za } f^{IV}} + O(h^6)$$

po sličnosti sa prethodnim naslovom pokazati da se greška sastavljene formule pravougaonika prikazuje u obliku $c_2h^2 + O(h^4)$, gdje c_2 ne zavisi od h, ali naravno zavisi od podintegralne funkcije

$$R(h) = -\frac{1}{12}h^2\left(\int_a^b f''(x)dx - c_2(f'')\cdot h^2 + O(h^4)\right) -$$

$$\frac{1}{480}h^4\left(\int_a^b f^{IV}(x)dx - c_2(f^{IV})\cdot h^2 + O(h^4)\right) + O(h^6) = C_2h^2 + C_4h^4 + O(h^6).$$

Ako se nastavi sa primjenom i dogradnjom postupka od maločas onda će se za grešku sastavljene trapezne formule za funkciju f po odsječku $[a,b]$ sa korakom h dobiti sljedeća formula:

$$R(h) = I - I(h) = C_2h^2 + C_4h^4 + \ldots + C_{2m}h^{2m} + O(h^{2m+2}),\ \ h \to 0, \tag{1}$$

gdje veličine $C_2, C_4, \ldots, C_{2m}$ ne zavise od h, a pretpostavlja se da $f \in C^{2m+2}[a,b]$.

Drugi dio izlaganja. Popravljanjem, sa koeficijentom popravke $\frac{1}{3}$, približnih vrijednosti čija greška ima oblik $C_2h^2+C_4h^4+C_6h^6+\ldots$ dobijaju se nove–bolje približne vrijednosti čija greška počinje sa h^4 (čija greška je jednaka $C_4'h^4+C_6'h^6+\ldots$). U nastavku je dat račun koji potkrepljuje ovo tvrđenje. Popravljanjem maločas dobijenih popravljenih vrijednosti, ovog puta uzimajući $\frac{1}{15}$ kao koeficijent popravke, dobijaju se još bolje približne vrijednosti, njihova greška počinje sa h^6. U nastavku je dat račun koji potkrepljuje ovo tvrđenje. Sljedeći koeficijent popravke biće $\frac{1}{63}$. Itd.

Zaista, iz pretpostavke da je $I - I(h) = C_2h^2 + C_4h^4 + C_6h^6 + \ldots$, ako se uvede oznaka $J\left(\frac{h}{2}\right) = I\left(\frac{h}{2}\right) + \frac{1}{3}\left(I\left(\frac{h}{2}\right) - I(h)\right)$, slijedi da je

$$I - J\left(\frac{h}{2}\right) = -\frac{1}{3}\left(I - I(h)\right) + \frac{4}{3}\left(I - I\left(\frac{h}{2}\right)\right) =$$

$$-\frac{1}{3}\left(C_2h^2 + C_4h^4 + C_6h^6 + \ldots\right) + \frac{4}{3}\left(C_2\left(\frac{h}{2}\right)^2 + C_4\left(\frac{h}{2}\right)^4 + C_6\left(\frac{h}{2}\right)^6 + \ldots\right) =$$

$$-\frac{1}{4}C_4h^4 - \frac{5}{16}C_6h^6 + \ldots = C_4'h^4 + C_6'h^6 + \ldots.$$

Zaista, iz pretpostavke da je $I - J(h) = \text{koeficijent}\cdot h^4 + O(h^6)$, ako se uvede oznaka $K\left(\frac{h}{2}\right) = J\left(\frac{h}{2}\right) + \frac{1}{15}\left(J\left(\frac{h}{2}\right) - J(h)\right)$, slijedi da je $I - K(\frac{h}{2}) = \ldots = O(h^6)$.

Da se dobije jedna nova popravljena približna vrijednost dovoljne su svega tri–četiri aritmetičke operacije. Time je izvođenje Rombergove formule urađeno. Samo se treba uvjeriti da su formule koje slijede saglasne sa dosad rečenim. Rombergova formula glasi:

$$\begin{cases} S_{0j} = h_j\left(\frac{1}{2}f(a) + \sum_{k=1}^{M-1} f(a+kh_j) + \frac{1}{2}f(b)\right), & j \geq 0 \\ S_{ij} = S_{i-1,j} + \frac{1}{4^i-1}\left(S_{i-1,j} - S_{i-1,j-1}\right), & i > 0,\ \ j \geq i \end{cases} \tag{2}$$

gdje je $h_j = 2^{-j}h_0$, $M = 2^jn$ i $nh_0 = b - a$.

Formula (2) služi za dobijanje približne vrijednosti za $I = \int_a^b f(x)dx$. Njen parametar je početni korak h_0. Dokaz formule (2) sadržan je u dosadašnjim razmatranjima. Zapaziti da za realizaciju formule (2) nije potrebno da znamo čemu su jednake vrijednosti $C_2, C_4, \ldots, C_{2m}$ iz (1).

Vrijednosti S_{0j} iz prvog reda tabele 1 odgovaraju trapeznoj formul: S_{00} sa korakom h_0, S_{01} sa korakom $h_1 = 2^{-1}h_0$, S_{02} sa korakom $h_2 = 2^{-2}h_0$, itd. Vrijednosti S_{1j} iz drugog reda

odgovaraju Simpsonovoj formuli. Vrijednosti S_{2j} iz trećeg reda odgovaraju jednoj formuli čiji je stepen tačnosti šest. Itd. Prilikom računanja vrijednosti S_{0j} treba iskoristiti već ranije izračunati broj $S_{0,j-1}$ koji sadrži polovinu sada potrebnih vrijednosti funkcije u čvorovima, zato smo nacrtali horizontalne strelice. Za računanje približne vrijednosti S_{ij} (kada je $i \geq 1$) dovoljne su svega tri–četiri aritmetičke operacije, vertikalne i kose strelice pokazuju na osnovu kojih ranije izračunatih vrijednosti se izračuna S_{ij}. Tabela 2 govori da precizan redosljed računanja jeste upravo ①, ②, ③, ….

Poslije ③, ⑥, ⑩, … računanja se prekidaju za trenutak da bi se provjerilo da li je dostignuta željena tačnost ε (unaprijed određena). Ako je ispunjen izlazni kriterijum $|S_{j-1,j} - S_{jj}| < \varepsilon$ onda se S_{jj} usvaja kao numerički odgovor. Program za računar obično se zaustavlja kada u nizu $S_{00}, S_{11}, S_{22}, \ldots$ dođe do poklapanja neka dva njegova susjedna člana. Kaže se da se računa sa najvećom mogućom tačnošću ili do mašinske tačnosti. Dalje popravke ne bi koristile tj. računar prosto ne bi mogao da ih registruje.

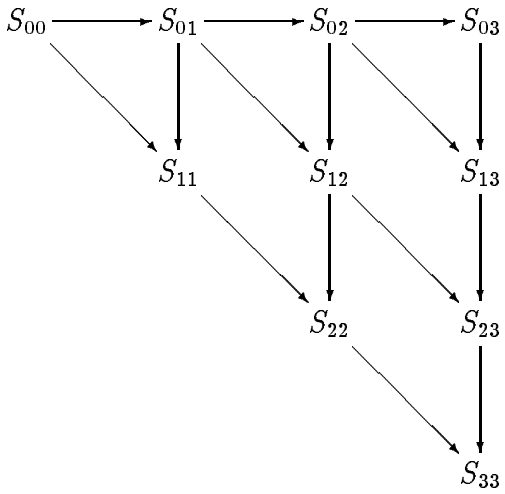


Tabela 1 $S_{11} = S_{01} + \frac{1}{3}\big(S_{01} - S_{00}\big)$

$I(h)$	$I(\frac{h}{2})$	$I(\frac{h}{4})$	$I(\frac{h}{8})$	$\cdots$
	$J(\frac{h}{2})$	$J(\frac{h}{4})$	$J(\frac{h}{8})$	$\cdots$
		$K(\frac{h}{4})$	$K(\frac{h}{8})$	$\cdots$
			$\cdots$	$\cdots$

①	②	④	⑦
	③	⑤	⑧
		⑥	⑨
			⑩

Tabela 2

2.4. KVADRATURNE FORMULE U SLUČAJU PRISUSTVA TEŽINSKE FUNKCIJE

Neka je $p = p(x)$ fiksirana funkcija $p\colon [a,b] \to R$ za koju se pretpostavlja da je integrabilna na odsječku $[a,b]$. Za funkciju p se kaže da je težinska funkcija ako je ispunjen uslov $p(x) \geq 0$ za svako $x \in [a,b]$. Razmotrimo određeni integral $I = I(f) = \int_a^b f(x)p(x)dx$. Neka je na odsječku $[a,b]$ postavljena mreža čvorova $\{x_i\}_{i=1}^n$ i neka su poznate vrijednosti funkcije u čvorovima $f(x_i) = f_i$. Konstruisaćemo kvadraturnu formulu (interpolacionog tipa) oblika $I(f) \approx S(f)$, gdje je $S(f) = \sum_{i=1}^n c_i f(x_i)$. Funkcija f biće zamijenjena svojim Lagranžovim i. p. po razmatranoj mreži čvorova. Odgovarajući izraz za grešku $R = R(f) = I - S = I(f) - S(f)$ takođe se lako dobija, polazeći od poznatog izraza za grešku interpolacije.

Iz naslova 1.1. i 1.2. poznato je sljedeće:

$$L_n(x) = \sum_{i=1}^n \Phi_i(x) f(x_i), \quad \Phi_i(x) = \prod_{j=1\ j\neq i}^n \frac{x - x_j}{x_i - x_j}, \quad f(x) = L_n(x) + \frac{1}{n!}\omega_n(x) f^{(n)}(\xi(x)),$$

$$a \leq \xi(x) \leq b \ \ (\text{jer } a \leq x \leq b), \quad f \in C^n[a,b], \quad \omega_n(x) = \prod_{i=1}^n (x - x_i).$$

Izvođenje:

$$f(x) = \sum_{i=1}^n \Phi_i(x) f(x_i) + \frac{1}{n!}\omega_n(x) f^{(n)}(\xi(x)) \quad \Big/ \cdot p(x)$$

$$f(x)p(x) = \sum_{i=1}^n \Phi_i(x) p(x) f(x_i) + \frac{1}{n!}\omega_n(x) f^{(n)}(\xi(x)) p(x) \quad \Big/ \int_a^b \ldots dx$$

$$\underbrace{\int_a^b f(x)p(x)dx}_{I(f)} = \sum_{i=1}^n \underbrace{\Big(\int_a^b \Phi_i(x)p(x)dx\Big)}_{c_i} f(x_i) + \underbrace{\frac{1}{n!}\int_a^b \omega_n(x) f^{(n)}(\xi(x))p(x)dx}_{R(f)}$$

$$I(f) = \sum_{i=1}^n c_i f(x_i) + R(f)$$

$$I(f) \approx S(f) = \sum_{i=1}^n c_i f(x_i) \tag{1}$$

Za grešku formule (1) imamo:

$$|R(f)| \leq \frac{1}{n!} M_n \int_a^b |\omega_n(x)| p(x)dx, \tag{2}$$

$$M_n = \max_{x\in[a,b]} |f^{(n)}(x)|.$$

Neka je mreža ravnomjerna i neka tačke $x = a$ i $x = b$ pripadaju mreži. Drugim riječima, neka čvorovi budu $x_i = a + ih$ za $0 \leq i \leq n$, gdje je $nh = b - a$. Tada se za kvadraturnu formulu oblika (1) kaže da je Njutn–Kotesova kvadraturna formula. Primjera radi, navedimo

Njutn–Kotesove formule kada je $n = 3$ i kada je $n = 4$, u slučaju da je težinska funkcija $p(x) \equiv 1$:

$$\int_a^b f(x)dx \approx \frac{b-a}{8}\Big(f_0 + 3f_1 + 3f_2 + f_3\Big) \quad \text{(tri–osminska formula)}$$

$$\int_a^b f(x)dx \approx \frac{b-a}{90}\Big(7f_0 + 32f_1 + 12f_2 + 32f_3 + 7f_4\Big)$$

Zadatak. Neka za kvadraturnu formulu (1), pored njene greške metode $R = R_1$ koja je izražena relacijom (2), treba proučiti i njenu grešku izazvanu približnošću ulaznih podataka R_2. Ulaznim podacima smatraju se brojevi $\{f_i\}_{i=1}^n$. Uzimamo da ulazni podaci nisu poznati sasvim tačno tj. uzimamo da raspolažemo samo sa približnim vrijednostima $\{f_i^*\}_{i=1}^n$. Neka je data mjera greške ulaznih podataka δ: $|f_i - f_i^*| \le \delta$ za $1 \le i \le n$. Ulogu numeričkog odgovora $S = \sum_{i=1}^n c_i f_i$ sada naravno preuzima broj $S^* = \sum_{i=1}^n c_i f_i^*$, tako da je $R_2 = S - S^*$. A naravno da je u datim okolnostima ukupna greška $I - S^*$ jednaka $I - S^* = R_1 + R_2$.

Treba riješiti postavljeni zadatak tj. treba ocijeniti R_2:

$$R_2 = S - S^* = \sum_{i=1}^n c_i f_i - \sum_{i=1}^n c_i f_i^* = \sum_{i=1}^n c_i(f_i - f_i^*)$$

$$|R_2| = \Big|\sum_{i=1}^n c_i(f_i - f_i^*)\Big| \le \sum_{i=1}^n |c_i(f_i - f_i^*)| = \sum_{i=1}^n |c_i| \cdot |f_i - f_i^*| \le \sum_{i=1}^n |c_i| \cdot \delta$$

$$|R_2| \le C \cdot \delta, \text{ gdje je } C = \sum_{i=1}^n |c_i|$$

Na redu je analiza posljednje relacije. Što je veličina C manja to je R_2 manje, odnosno to se greška ulaznih podataka slabije odražava–prenosi na rezultat (na numerički odgovor). Drugim riječima, što je veličina C manja to je formula (1) stabilnija u odnosu na grešku svojih ulaznih podataka, odnosno to je situacija povoljnija. Ili iz suprotnog ugla: ako je broj C velik onda je formula (1) numerički nedovoljno stabilna u odnosu na grešku ulaznih podataka. Zadatak je riješen.

Posmatrajmo relaciju (1) kada je $f(x) \equiv 1$. Tada je $L_n(x) \equiv f(x)$ i zato $R(f) = 0$ odnosno $I(f) = S(f)$. $I(f) = \int_a^b p(x)dx$, $S(f) = \sum_{i=1}^n c_i$, tako da je $\sum_{i=1}^n c_i = \int_a^b p(x)dx = const$.

Ako su svi koeficijenti $\{c_i\}_{i=1}^n$ pozitivni onda je očito $\sum_{i=1}^n c_i = \sum_{i=1}^n |c_i|$. Međutim, ako među koeficijentima $\{c_i\}_{i=1}^n$ ima i pozitivnih i negativnih onda je očito $\sum_{i=1}^n |c_i| > \sum_{i=1}^n c_i$. Za Njutn–Kotesove formule sa velikim n (recimo sa $n \ge 10$) karakteristično je da su neki koeficijenti pozitivni a neki negativni, što umanjuje njihovu numeričku stabilnost u odnosu na grešku ulaznih podataka, pa se zato te formule u praksi rijetko koriste.

2.5. GAUSOVA KVADRATURNA FORMULA

Priprema o ortogonalnim polinomima

U realnom Hilbertovom prostoru H posmatrajmo n linearno nezavisnih elemenata φ_1, ..., φ_n. Primjenom Gram–Šmitovog postupka ortogonalizacije, sistem $\{\varphi_1, \ldots, \varphi_n\}$ može da se ortogonalizuje, tj. može da bude dobijen sistem $\{\psi_1, \ldots, \psi_n\}$ za koji važi: a) ψ_k je linearna

kombinacija od $\varphi_1, \ldots, \varphi_k$ i b) $\psi_i \perp \psi_j$ tj. $(\psi_i, \psi_j) = 0$ za $i \neq j$; $(\ ,\ )$ je oznaka za skalarni proizvod. Elementi $\psi_1, \ldots, \psi_n$ određeni su jednoznačno do na multiplikativne konstante.

Razmotrimo konkretni Hilbertov prostor $H = L_{2,p(x)}[a,b]$. Neka je funkcija $p = p(x)$ integrabilna po $[a,b]$ i neka je $p(x) > 0$ skoro svuda na $[a,b]$. Funkcija f pripada H ako i samo ako je $\int_a^b |f(x)|^2 p(x)dx$ konačan. Ako se dvije funkcije razlikuju samo na skupu mjere nula onda se one identifikuju. U H se skalarni proizvod definiše kao $(f,g) = \int_a^b f(x)g(x)p(x)dx$. Ovo je Lebegov prostor. Na primjer $L_2[a,b]$.

U našem Hilbertovom prostoru H razmotrimo njegovih n elemenata $\varphi_i(x) = x^{i-1}$, za $i = 1, \ldots, n$. Sistem $\{1, x, x^2, \ldots, x^{n-1}\}$ je linearno nezavisan. Njegovom ortogonalizacijom nastaje sistem $\{\psi_1, \ldots, \psi_n\}$. Kako je φ_i polinom stepena tačno $i-1$ to je i ψ_i polinom stepena tačno $i-1$. Po n može da se napreduje, tako da može da se posmatra sistem $\{\psi_1, \psi_2, \ldots\}$. Izvršimo malo prilagođavanje oznaka: umjesto $\psi_1, \psi_2, \ldots$ pisaćemo $\psi_0, \psi_1, \ldots$. Tako da je odsad ψ_i polinom stepena tačno i. Za sistem $\{\psi_0, \psi_1, \ldots\}$ kaže se da je sistem ili da je niz ortogonalnih polinoma. Navedimo dva primjera.

Primjer 1. Neka bude $[a,b] = [-1,1]$ i $p(x) \equiv 1$. Sistem ortogonalnih polinoma u prostoru $L_2[-1,1]$ je sistem tzv. Ležandrovih polinoma $\{L_n\}_{n=0}^{\infty}$, za koje važi relacija: $L_n(x) = \dfrac{1}{2^n n!} \dfrac{d^n}{dx^n} (x^2-1)^n$. Recimo, $L_0(x) = 1$, $L_1(x) = x$, $L_2(x) = \dfrac{3}{2}x^2 - \dfrac{1}{2}$, $L_3(x) = \dfrac{5}{2}x^3 - \dfrac{3}{2}x$.

Prinjer 2. Neka bude $[a,b] = [-1,1]$ i $p(x) = \dfrac{1}{\sqrt{1-x^2}}$. U ovom prostoru $L_{2,p(x)}[-1,1]$, ortogonalni polinomi su tzv. Čebiševljevi polinomi $T_n(x)$; $n = 0, 1, \ldots$; za koje važi relacija: $T_n(x) = \cos(n \arccos x)$. Recimo, $T_0(x) = 1$, $T_1(x) = x$, $T_2(x) = 2x^2 - 1$, $T_3(x) = 4x^3 - 3x$.

Za dokaz prvog primjera, treba se uvjeriti da se ortogonalizacijom sistema $\{1, x, x^2, \ldots\}$ dobije sistem $\{L_0, L_1, L_2, \ldots\}$. Ili, pošto je već napisana gotova relacija $L_n(x) = \ldots$, dovoljno je uvjeriti se: a) da je L_n polinom stepena tačno n i b) da je $L_i \perp L_j$ tj. da je $(L_i, L_j) = \int_{-1}^{1} L_i(x)L_j(x)dx = 0$ za $i \neq j$. Slično važi naravno i za drugi primjer. Dokaze za jedan i drugi primjer izostavljamo.

U sljedećoj lemi govori se o jednom svojstvu polinoma iz sistema ortogonalnih polinoma. Prije toga, o faktorizaciji polinoma u realnom području. Polinom $P = P(x)$ na jednoznačan način može da se prikaže u obliku

$$P(x) = c_0 \cdot \prod_i (x - \alpha_i)^{m_i} \cdot \prod_j (x^2 + \beta_j x + \gamma_j)^{n_j},$$

gdje jednačina $x^2 + \beta_j x + \gamma_j = 0$ nema realnog rješenja; za izraz $x^2 + \beta_j x + \gamma_j$ kaže se da je nesvodljiv trinom. Broj $x = \alpha_i$ je nula polinoma višestrukosti m_i. Polinom P je očito stepena $\sum_i m_i + 2\sum_j n_j$. Recimo

$$P(x) = 2(x-1)^4(x-2)(x-6)^2(x^2+1)^2(x^2+4x+5).$$

Dalje, o broju nula polinoma poznato je sljedeće pravilo: polinom stepena n ima najviše n realnih nula.

Lema. Neka je $\{\psi_0, \psi_1, \ldots\}$ sistem ortogonalnih polinoma u prostoru $L_{2,p(x)}[a,b]$; ovdje je $\psi_n = \psi_n(x)$ polinom stepena tačno n. Polinom ψ_n ima u otvorenom intervalu (a,b) tačno n međusobno različitih nula.

Dokaz leme. Uočimo sve nule polinoma ψ_n koje pripadaju intervalu (a,b) i čija je višestrukost neparna; neka su to $x_1, \ldots, x_\ell$. Dopustimo da je $\ell < n$. Neka bude

$$f(x) = \prod_{i=1}^{\ell} (x - x_i) \quad \text{i} \quad g(x) = \psi_n(x) f(x).$$

Razmotrimo veličinu

$$A = \int_a^b g(x)p(x)dx = \int_a^b \psi_n(x) \cdot \prod_{i=1}^{\ell}(x - x_i) \cdot p(x)dx = \int_a^b \psi_n(x)f(x)p(x)dx = (\psi_n, f).$$

Pođimo od faktorizacije za ψ_n i pogledajmo faktorizaciju za g.

Nesvodljivi trinomi u faktorizaciji za ψ_n ne mijenjaju znak uopšte na čitavoj realnoj osi. Faktori $(x - \alpha_i)^{m_i}$, gdje je m_i paran, takođe ne mijenjaju znak. Faktori $(x - \alpha_i)^{m_i}$, gdje je m_i neparan a $\alpha_i \notin (a, b)$, ne mijenjaju znak u (a, b).

Ako je eksponent m_i neparan i $\alpha_i \in (a, b)$ onda se u faktorizaciji za g pojavljuje paran eksponent, jer je izvršeno množenje sa $x - \alpha_i$, tj. sa $f(x)$. Tako da je $g(x) \geq 0$ za $x \in [a, b]$, ako je $c_0 > 0$, odnosno $g(x) \leq 0$ za $x \in [a, b]$, ako je $c_0 < 0$. Funkcija g je polinom, g ima samo konačno mnogo nula, tako da je $A = \int_a^b g(x)p(x)dx > 0$, ili < 0. Dakle, $A \neq 0$.

S druge strane, $A = (\psi_n, f)$. Pretpostavili smo da je $\ell < n$. Zato je f polinom stepena nižeg od n. Polinom f može se prikazati kao linearna kombinacija od $\psi_0, \ldots, \psi_\ell$; $f = \sum_{i=0}^{\ell} c_i\psi_i$. Važi $\psi_i \perp \psi_n$ tj. $(\psi_i, \psi_n) = 0$ za $i = 0, \ldots, \ell \Rightarrow$

$$A = (\psi_n, f) = (\psi_n, c_0\psi_0 + \ldots + c_\ell\psi_\ell) = c_0(\psi_n, \psi_0) + \ldots + c_\ell(\psi_n, \psi_\ell) = c_0 \cdot 0 + \ldots + c_\ell \cdot 0 = 0.$$

$A = 0$. Dobili smp kontradikciju. Ne može biti $\ell < n$. Mora biti $\ell = n$. Lema je dokazana. Sve nule su jednostruke (proste).

Svi članovi niza ortogonalnih polinoma $\psi_0, \psi_1, \ldots$ određeni su jednoznačno do na multiplikativne konstante. Neka su te konstante izabrane tako da najstariji koeficijent svakog polinoma iz niza bude jednak 1, tj. da bude $\psi_k(x) = x^k + \ldots$.

Prelazimo na Gausovu kvadraturnu formulu. Neka je odsječak $[a, b]$ fiksiran i neka je težinska funkcija $p = p(x)$ fiksirana. Za težinsku funkciju se pretpostavlja da je integrabilna i da je skoro svuda pozitivna

Neka bude $n \geq 1$. Razmotrimo kvadraturnu formulu sa n čvorova $x_1, \ldots, x_n$:

$$I(f) \approx S(f), \qquad I(f) = \int_a^b f(x)p(x)dx, \qquad S(f) = \sum_{i=1}^{n} c_i f(x_i) \tag{1}$$

i uvedimo sljedeću oznaku za njenu grešku: $R(f) = I(f) - S(f)$.

Definišimo algebarski stepen tačnosti a kvadraturne formule (1). Za (1) se kaže da ima algebarski stepen tačnosti a ako je $R(f) = 0$ (greška je jednaka nuli, kvadraturna formula je tačna) kada je f – bilo koji polinom stepena $\leq$a. A postoji bar jedan polinom f čiji je stepen jednak a+1 takav da je za njega $R(f) \neq 0$.

Gaus je razmatrao i riješio sljedeći ekstremalni zadatak: konstruisati kvadraturnu formulu oblika (1) sa najvećim mogućim algebarskim stepenom tačnosti. Vidjećemo da je rješenje zadatka a=$2n - 1$. Ovo nije neočekivano. Zaista, s jedne strane, vidimo da u (1) ima $2n$ stepeni slobode $c_1, \ldots, c_n, x_1, \ldots, x_n$ (pogodno ih izaberi). S druge strane, ima $2n$ uslova ako se traži: neka bude $R(f) = I(f) - S(f) = 0$ kada je $f(x) = x^k$ za $k = 0, \ldots, 2n - 1$.

Dokažimo dvije leme.

Lema 1. Pretpostavimo da kvadraturna formula (1) ima svojstvo da je tačna za sve polinome f čiji je stepen $\leq 2n - 1$. Neka bude $\omega_n(x) = \prod_{i=1}^{n}(x - x_i)$. Neka je $P_{n-1} = P_{n-1}(x)$ proizvoljni polinom stepena $\leq n - 1$. Tada je $\int_a^b \omega_n(x)P_{n-1}(x)p(x)dx = 0$.

Zapaziti da lema 1. ima oblik inplikacije. Ona polazi od toga da postoji kvadraturna formula oblika (1) sa svojstvom "tačna je za sve polinome f čiji je stepen $\leq 2n - 1$". Ona ne tvrdi da takva formula postoji. Ona ne dokazuje da takva formula postoji.

Dokaz leme 1. Stavimo $f(x) = \omega_n(x)P_{n-1}(x)$. Proizvod ω_n je polinom stepena tačno n, P_{n-1} je polinom stepena $\leq n-1$. f je polinom stepena $\leq 2n-1$, zato je $R(f) = I(f) - S(f) = 0$, po pretpostavci. Imamo redom:

$$0 = R(f) = I(f) - S(f) = \int_a^b f(x)p(x)dx - \sum_{i=1}^{n} c_i f(x_i) =$$

$$\int_a^b \omega_n(x)P_{n-1}(x)p(x)dx - \sum_{i=1}^{n} c_i \left(\omega_n(x)P_{n-1}(x)\right)|_{x=x_i} =$$

$$\int_a^b \omega_n(x)P_{n-1}(x)p(x)dx - \sum_{i=1}^{n} c_i \cdot 0 \cdot P_{n-1}(x_i) = \int_a^b \omega_n(x)P_{n-1}(x)p(x)dx - 0.$$

Lema 1. je dokazana.

Umjesto $\int_a^b \omega_n(x)P_{n-1}(x)p(x)dx = 0$ možemo da pišemo $(\omega_n, P_{n-1}) = 0$ tj. $\omega_n \perp P_{n-1}$. Polinom ω_n je ortogonalan na sve polinome čiji je stepen $\leq n-1$. Dakle, mora da bude $\omega_n = \psi_n$. Ovo određuje koje tačke su čvorovi kvadraturne formule. To su nule polinoma n–tog stepena ψ_n iz niza ortogonalnih polinoma $\{\psi_0, \psi_1, \ldots\}$.

Lema 2. Neka su $x_1, \ldots, x_n$ nule polinoma ψ_n. Neka je kvadraturna formula oblika (1) tačna za sve polinome f čiji je stepen $\leq n-1$. Tada je (1) tačna za sve polinome čiji je stepen $\leq 2n-1$.

Zapaziti da i lema 2. samo hipotetički govori o formuli oblika (1) koja bi sada imala svojstvo "tačna je (greška je nula) za sve polinome čiji je stepen $\leq n-1$".

Dokaz leme 2. Uzmimo ma koji polinom P_{2n-1} čiji je stepen $\leq 2n-1$. Podijelimo ga sa polinomom ψ_n. Ostatak dijeljenja je naravno nižeg stepena od ψ_n. Označimo količnik sa g_{n-1} a ostatak sa r_{n-1}: $P_{2n-1}(x) = g_{n-1}(x)\psi_n(x) + r_{n-1}(x)$, gdje su i g_{n-1} i r_{n-1} izvjesni polinomi stepena $\leq n-1$. Obavezni smo da pokažemo da je $R(P_{2n-1}) = 0$. Imamo redom:

$R(P_{2n-1}) = R(g_{n-1}\psi_n + r_{n-1}) =$ (integral je aditivan, formula (1) je aditivna)

$R(g_{n-1}\psi_n) + R(r_{n-1}) =$ (pretpostavka leme)

$$R(g_{n-1}\psi_n) + 0 = I(g_{n-1}\psi_n) - S(g_{n-1}\psi_n) =$$

$$\int_a^b g_{n-1}(x)\psi_n(x)p(x)dx - \sum_{i=1}^{n} c_i \left(g_{n-1}(x)\psi_n(x)\right)|_{x=x_i} =$$

$$(g_{n-1}, \psi_n) - \sum_{i=1}^{n} c_i \cdot g_{n-1}(x_i) \cdot 0 = 0,$$

jer $\psi_n \perp \psi_0, \ldots, \psi_{n-1} \Rightarrow \psi_n \perp$ linearna kombinacija od $\psi_0, \ldots, \psi_{n-1} \Rightarrow \psi_n \perp g_{n-1}$.

Lema 2. je dokazana.

Uz pomoć dvije leme, postavljeni zadatak sveo se na sljedeće. Čvorovi kvadraturne formule su definisani. Još nisu definisani koeficijenti $c_1, \ldots, c_n$. Traži se da formula bude tačna za proizvoljnu funkciju f koja je polinom čiji je stepen $\leq n-1$, pa će onda automatski biti tačna i za sve polinome do stepena $2n-1$ uključeno, prema lemi 2. Lako se konstruiše formula koja

je tačna (čija je greška $R(f)$ jednaka nuli) za polinome stepena $\leq n-1$. To je urađeno u prethodnom naslovu 2.4.

$$\int_a^b f(x)p(x)dx \approx \int_a^b L_n(x)p(x)dx \quad \leftarrow \quad \text{nađi tačan izraz za ovo;}$$

L_n je L. i. p.

Drugim riječima, formula (1) je interpolacionog tipa.

U ovom trenutku je izvedena Gausova kvadraturna formula. Čvorovi su nule od ψ_n. A koeficijenti se izračunaju po šablonu za kvadraturnu formulu u slučaju prisustva težinske funkcije.

Slijede dvije napomene.

Napomena 1. Ne postoji kvadraturna formula sa n čvorova oblika (1) čiji bi algebarski stepen tačnosti a bio jednak $2n$. Zaista, zamislimo da takva formula postoji; čvorovi su označeni kao $x_1, \ldots, x_n$; neka bude $\omega_n(x) = \prod_{i=1}^n (x - x_i)$. Posmatrajmo kada je podintegralna funkcija $f(x) = \omega_n^2(x) = \prod_{i=1}^n (x - x_i)^2 \geq 0$. Zapazimo unaprijed da su sve vrijednosti f u čvorovima jednake nuli. Sada, s jedne strane, $I(f) = \int_a^b f(x)p(x)dx > 0$, a s druge strane $S(f) = \sum_{i=1}^n c_i f(x_i) = \sum_{i=1}^n c_i \cdot 0 = 0$. Tako da je $R(f) = I(f) - S(f) \neq 0$. Vidimo da je f polinom stepena tačno $2n$, tako da je dokaz završen. Tek sada je ekstremalni zadatak u potpunosti riješen.

Napomena 2. Svi koeficijenti Gausove kvadraturne formule su pozitivni tj. za $i = 1, \ldots, n$ važi $c_i > 0$. Zaista, neka podintegralna funkcija bude $f(x) = \prod_{j=1 \ j\neq i}^n (x - x_j)^2$; f je polinom stepena tačno $2n-2$; $f(x) \geq 0$. Tada je

$$I(f) = \int_a^b f(x)p(x)dx > 0 \quad \text{i} \quad S(f) = \sum_{j=1}^n c_j f(x_j) = c_i f(x_i) \quad \text{i} \quad R(f) = 0 \quad \text{tj.} \quad I(f) = S(f).$$

Slijedi $c_i f(x_i) > 0$. Slijedi $c_i > 0$, jer je $f(x_i) > 0$. Time je dokaz završen.

Ako se uvrsti $f(x) = 1$ onda $\int_a^b p(x)dx = \sum_{i=1}^n c_i = C$. Veličina C ne zavisi od n.

Svojstvo $c_i > 0$ za $i = 1, \ldots, n$ daje numeričku stabilnost Gausove kvadraturne formule u odnosu na grešku ulaznih podataka $f(x_1), \ldots, f(x_n)$; v. analizu u 2.4.

Ostaje da se ocijeni greška kvadraturne formule (1). Neka $f \in C^{2n}[a,b]$. Pridružimo funkciji f njen L. i. p. po tačkama $x_1, \ldots, x_n$; $L_n = L_n(x)$; stepena $\leq n-1$. Pridružimo i Hermitov i. p. $H_{2n} = H_{2n}(x)$, stepena $\leq 2n-1$, definisan sa $2n$ uslova:

$$H_{2n}(x_i) = f(x_i), \quad H'_{2n}(x_i) = f'(x_i), \quad i = 1, \ldots, n$$

(ima n čvorova, svi čvorovi su dvostruki). Dalje imamo:

$$f(x) - H_{2n}(x) = \frac{1}{(2n)!} f^{(2n)}(\xi(x))\psi_n^2(x) \qquad \text{(v. u Interpolacija sa višestrukim čvorovima),}$$

$$\text{(ponovimo da je)} \qquad \psi_n(x) = \prod_{i=1}^n (x - x_i),$$

$$R(f) = I(f) - S(f) = \qquad \text{(formula (1) izvedena je po šablonu)}$$

$$I(f) - I(L_n) = \qquad (\text{stepen}(L_n) \leq \text{a} = 2n-1)$$

$$I(f) - S(L_n) = \qquad (L_n(x_i) = H_{2n}(x_i) = f(x_i),\ \sum_{i=1}^{n} c_i L_n(x_i) = \sum_{i=1}^{n} c_i H_{2n}(x_i))$$

$$I(f) - S(H_{2n}) = \qquad (\text{stepen}(H_{2n}) \leq \text{a})$$

$$I(f) - I(H_{2n}) = \int_a^b f(x)p(x)dx - \int_a^b H_{2n}(x)p(x)dx =$$

$$\int_a^b \frac{1}{(2n)!} f^{(2n)}(\xi(x))\psi_n^2(x)p(x)dx = \qquad (\psi_n^2(x)p(x) \geq 0)$$

$$\frac{1}{(2n)!} f^{(2n)}(\xi) \int_a^b \psi_n^2(x)p(x)dx, \quad a < \xi < b$$

(posljednji integral je konstanta, za konkretnu kvadraturnu formulu). Imamo konačno:

$$R(f) = \frac{1}{(2n)!} f^{(2n)}(\xi) \int_a^b \psi_n^2(x)p(x)dx.$$

Na kraju, navedimo dva specijalna slučaja Gausove kvadraturne formule. Odgovaraju redom prvom i drugom primjeru iz pripreme. Izvođenje (račun) za dva specijalna slučaja se izostavlja. Dokažite sami za vježbu. Sastaviti i izraze za grešku.

Primjer 1. Neka bude $[a, b] = [-1, 1]$ i $p(x) \equiv 1$

$$n = 1: \quad \int_{-1}^{1} f(x)dx \approx 2f(0), \qquad n = 2: \quad \int_{-1}^{1} f(x)dx \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

$$n = 3: \quad \int_{-1}^{1} f(x)dx \approx \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right)$$

Primjer 2. Neka bude $[a, b] = [-1, 1]$ i $p(x) = \dfrac{1}{\sqrt{1 - x^2}}$

$$\int_{-1}^{1} \frac{f(x)dx}{\sqrt{1 - x^2}} \approx \frac{\pi}{n} \sum_{i=1}^{n} f(x_i), \ \text{ gdje je } \ x_i = \cos\frac{(2i - 1)\pi}{2n} \quad \text{(Hermitova formula)}$$

3. NUMERIČKE METODE ALGEBRE

Numeričkim metodama algebre pripada numeričko rješavanje sljedećih zadataka: naći rješenje sistema linearnih jednačina, izvršiti inverziju matrice, izračunati vrijednost determinante, naći svojstvene / sopstvene vrijednosti i vektore matrice, odrediti nule polinoma.

Razmotrimo zadatak o rješavanju sistema od n linearnih jednačina sa n nepoznatih. Postoje tri vrste numeričkih metoda: a) tačne ili direktne, b) iterativne i c) vjerovatnosne ili stohastičke. Ako je $n < 10^3$ onda treba primijeniti neku metodu oblika a), ako je $10^3 < n < 10^6$ onda metodu oblika b), ako je $n > 10^6$ onda metodu oblika c), po pravilu.

Za numeričku metodu se kaže da je tačna metoda ako je njena greška metode jednaka nuli. Drugim riječima, za numeričku metodu se kaže da je tačna ako ona daje tačan rezultat (greška je jednaka nuli), nakon izvođenja konačno mnogo aritmetičkih i logičkih operacija, pod pretpostavkom da nema greške računanja i greške ulaznih podataka.

3.1. GAUSOVA METODA ELIMINACIJE

Ili Gausova metoda uzastopne eliminacije nepoznatih. To je jedna tačna metoda, služi za rješavanje sistema linearnih jednačina. Neka je A linearni operator $R^n \to R^n$ ili neka je $A = [a_{ij}]_{i,j=1}^n$ realna kvadratna matrica oblika $n \times n$. Razmotrimo sistem linearnih jednačina $Ax = b$ ili

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = \overline{1,n},$$

$x = (x_1, \ldots, x_n) \in R^n$, $b = (b_1, \ldots, b_n) \in R^n$, $n \geq 1$; A – matrica sistema, b – vektor slobodnih članova, x – nepoznata.

Možemo pisati $Ax = b$ gdje je $x = [x_1 \ \ldots \ x_n]^T$, $b = [b_1 \ \ldots \ b_n]^T$, x i b su matrice oblika $n \times 1$ (x i b su vektori–kolone); matrični zapis ili svejedno $Ax = b$ gdje je $x = (x_1, \ldots, x_n)$, $b = (b_1, \ldots, b_n)$, x i b pripadaju R^n; operatorski zapis.

Formalno gledano, sistem $Ax = b$ može da bude riješen primjenom Kramerovog pravila (izračuna se $n + 1$ determinanta). Međutim, za vremensku složenost ili za potreban broj aritmetičkih operacija t_n tada važi $t_n > n!$ što čini da je taj postupak praktično neupotrebljiv već za recimo $n = 20$. Dodatno, što je broj aritmetičkih operacija veći to je i greška računanja veća. Gausova metoda eliminacije svodi t_n na $t_n = O(n^3)$, a ako se pod t_n podrazumijeva samo broj izvršenih operacija množenja i dijeljenja onda imamo $t_n \sim \frac{1}{3}n^3$.

Izložimo algoritam. Uzmimo da je $a_{11} \neq 0$. Veličina x_1 eliminiše se iz jednačina $i = 2, \ldots, i = n$. Upravo, prva jednačina $i = 1$ podijeli se sa a_{11}, pa se onda nova prva jednačina, pomnožena sa $-a_{i1}$, doda i–toj jednačini, za $i = 2, \ldots, i = n$. Sistem sada ima oblik:

$$x_1 + \sum_{j=2}^{n} a_{1j}^{(1)} x_j = b_1^{(1)}$$
$$\sum_{j=2}^{n} a_{ij}^{(1)} x_j = b_i^{(1)}, \quad i = \overline{2,n}$$

Uzmimo da je $a_{22}^{(1)} \neq 0$. Veličina x_2 eliminiše se iz jednačina $i = 3, \ldots, i = n$, na sličan način kao maločas. Itd. Time se polazni sistem svodi na gornji trougaoni oblik:

$$x_i + \sum_{j=i+1}^{n} a_{ij}^{(i)} x_j = b_i^{(i)}, \quad i = \overline{1,n}$$

ili

$$\begin{bmatrix} 1 & a_{12}^{(1)} & a_{13}^{(1)} & \dots & a_{1n}^{(1)} \\ & 1 & a_{23}^{(2)} & \dots & a_{2n}^{(2)} \\ & & \dots & \dots & \dots \\ 0 & & & & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(n)} \end{bmatrix}$$

Ako je $a_{11} = 0$ onda se izvrši pogodna zamjena mjesta jednačina tako da time postane $a_{11} \neq 0$. Slično ako se desi da je $a_{22}^{(1)} = 0$, itd. Zapaziti da postoje dvije mogućnosti: prva: zamjenom mjesta se postigne da su svi eliminatorni elementi $a_{11}, a_{22}^{(1)}, \dots$ različiti od nule i druga: to ne može da bude postignuto. Lako se vidi da prva mogućnost odgovara slučaju da je $\det A \neq 0$ (matrica A je regularna, sistem $Ax = b$ ima jedinstveno rješenje), dok druga mogućnost nastaje kada je $\det A = 0$ (matrica A je singularna, nije tačno da sistem $Ax = b$ ima jedinstveno rješenje). Dakle, tokom realizacije algoritma pokazaće se da li je $\det A \neq 0$ ili je pak suprotno $\det A = 0$.

Prema tome, pretpostavimo da je $\det A \neq 0$.

Gausov algoritam eliminacije sastoji se iz dva dijela: a) direktni hod i b) inverzni hod. Direktni hod algoritma obuhvata računanja kojima se polazni sistem $Ax = b$ svede na gornji trougaoni oblik. U inverznom hodu se pređe put od trougaonog oblika do nalaženja rješenja sistema.

Ostaje da se izloži b). Nastali trougaoni sistem se trivijalno rješava. Iz jednačine $i = n$ odredi se x_n, zatim se iz jednačine $i = n - 1$ odredi x_{n-1}, itd.

3.2. GAUSOVA METODA ELIMINACIJE SA IZBOROM GLAVNOG ELEMENTA

Riječ je opet o jednoj tačnoj metodi za rješavanje sistema linearnih jednačina, a predstavlja jednu modifikaciju ili poboljšanje metode izložene u prethodnom naslovu. U čemu se sastoji modifikacija? Izbor eliminatornog ili glavnog elementa više se ne prepušta slučaju. Glavni element se bira po kriterijumu: neka bude što je moguće veći po apsolutnoj vrijednosti. Dakle, vrši se upravljanje tokom računskih operacija, vrši se vođenje ili pivotiranje. Šta se time dobija? Prethodno, praksa ili iskustvo pokazuje da algoritam iz prethodnog naslova (bez vođenja) ima jednu slabu stranu: uticaj greške računanja često sasvim pokvari numerički rezultat. Takođe se i eventualno prisutna greška ulaznih podataka često veoma negativno odrazi na grešku numeričkog rezultata. Ako se vrši vođenje onda se rezultujuća greška radikalno smanji. Prva aksioma numeričkih metoda – prilikom rješavanja sistema linearnih jednačina direktnom metodom obavezno vršiti izbor glavnog elementa. U prvom dijelu izlaganja biće izložen sami algoritam. U drugom dijelu izlaganja biće dato obrazloženje za – korisno je vršiti vođenje.

Razmotrimo sistem linearnih jednačina $Ax = b$. Modifikacija se odnosi na direktni hod algoritma. U prvom koraku se za glavni element izabere najveći po apsolutnoj vrijednosti član čitave matrice A, odnosno najveći po apsolutnoj vrijednosti među brojevima a_{ij}, $1 \leq i \leq n$, $1 \leq j \leq n$. Neka je to a_{kl}. Broj a_{kl} dovodi se na mjesto a_{11}. Jednačine $i = 1$ i $i = k$ zamijene mjesta. Pored toga, neka stupci $j = 1$ i $j = l$ zamijene mjesta, čime promjenljive x_1 i x_l zamijene uloge; voditi računa o ovoj permutaciji. U drugom koraku se za glavni element izabere najveći po apsolutnoj vrijednosti među brojevima a_{ij}, $2 \leq i \leq n$, $2 \leq j \leq n$. Slično u trećem koraku, itd. u n–tom koraku. Kada se trougaoni sistem riješi, onda razdužiti permutacije, u obrnutom redosljedu.

Napišimo grubi algoritam. Program ima dvije osnovne promjenljive: matrica A oblika $n \times n$ i niz b dužine n. Prvo se učitaju ulazni podaci n, A i b.

Za svako i od 1 do n ponovi
{ 1. Među brojevima a_{kl}, $i \le k \le n$, $i \le l \le n$, pronađi najveći po apsolutnoj vrijednosti, neka je to a_{kl}
2. Ako je $a_{kl} = 0$ onda se štampa "$\det A = 0$" i program se zaustavlja
3. Neka i–ti i k–ti redak matrice A zamijene mjesta i neka b_i i b_k zamijene mjesta
4. Neka i–ti i l–ti stubac matrice A zamijene mjesta i neka se zapamti koja je permutacija stubaca izvršena
5. Učini da postane $a_{ii} = 1$ tj. i–tu jednačinu podijeli sa a_{ii}; konkretno:
$a_{ij} \leftarrow 0$ za $1 \le j \le i-1$, $a_{ij} \leftarrow a_{ij}/a_{ii}$ za $i+1 \le j \le n$, $b_i \leftarrow b_i/a_{ii}$, $a_{ii} \leftarrow 1$
6. Za svako m od $i+1$ do n ponovi { Učini da postane $a_{mi} = 0$ tj. i–tu jednačinu, pomnoženu sa $-a_{mi}$, dodaj m–toj jednačini; konkretno:
$a_{mj} \leftarrow 0$ za $1 \le j \le i-1$, $a_{mj} \leftarrow a_{mj} - a_{mi}a_{ij}$ za $i+1 \le j \le n$, $b_m \leftarrow b_m - a_{mi}b_i$, $a_{mi} \leftarrow 0$ } }
7. Sada je A gornja trougaona matrica. Riješiti sistem $Ax = b$. Iz jednačine $i = n$ odrediti x_n, zatim iz jednačine $i = n-1$ odrediti x_{n-1}, itd.
8. Razduži permutacije i štampaj rezultat $x_1, \ldots, x_n$

Zapaziti da pronalaženje glavnog elementa samo malo povećava vremensku složenost t_n.

Može se vršiti samo djelimično vođenje: glavni element traži se samo u stupcu tj. među brojevima a_{ki}, $i \le k \le n$. Tada nema permutacija.

Za obrazloženje, pogledajmo po kom zakonu se mijenja opšti član matrice a_{mj}, u i–tom koraku:

$$\begin{cases} a_{11}x_1 + \ldots + a_{1n}x_n = b_1 \\ \ldots \\ a_{ii}x_i + a_{i,i+1}x_{i+1} + \ldots + a_{ij}x_j + \ldots + a_{in}x_n = b_i \\ \ldots \\ a_{mi}x_i + a_{m,i+1}x_{i+1} + \ldots + a_{mj}x_j + \ldots + a_{mn}x_n = b_m \\ \ldots \end{cases}$$

$$a_{mj} \leftarrow a_{mj} - \frac{a_{mi}a_{ij}}{a_{ii}}$$

Sve članove matrice smatramo približnim brojevima, zbog greške računanja koja se do datog trenutka akumulirala. Vrše se nova računanja nad članovima matrice. Pogledajmo grešku izraza na desnoj strani u posljednjoj relaciji, u zavisnosti od veličine eliminatornog elementa a_{ii}. Pogledajmo uže grešku izraza $\frac{a_{mi}a_{ij}}{a_{ii}}$. Uvedimo oznake $x = a_{mi}a_{ij}$, $y = a_{ii}$ i $f(x,y) = \frac{x}{y}$. Neka x ima grešku $\Delta(x)$ i neka y ima grešku $\Delta(y)$. Kako se to odražava na grešku broja $f(x,y)$? Naslov Greška funkcije, formula za grešku funkcije od više promjenljivih:

$$x = x^* + \Delta(x), \;\; y = y^* + \Delta(y), \;\; f(x,y) = f(x^*,y^*) + \Delta(f)$$

$$\frac{\partial f}{\partial x} = \frac{1}{y}, \;\; \frac{\partial f}{\partial y} = -\frac{x}{y^2}$$

$$\Delta(f) \approx \frac{\partial f(x,y)}{\partial x} \cdot \Delta(x) + \frac{\partial f(x,y)}{\partial y} \cdot \Delta(y) = \frac{1}{y} \cdot \Delta(x) - \frac{x}{y^2} \cdot \Delta(y)$$

Što je broj $|y| = |a_{ii}|$ veći to je greška $|\Delta(f)|$ manja. Ako se eliminatorni element udvostruči onda se greška približno prepolovi (koeficijent koristi približno je jednak $c = 2$), i slično.

Pogodno biranje eliminatornog elementa a_{ii} doprinosi iz koraka u korak smanjivanju greške, da se opiše ukupno smanjenje ili ukupna korist treba da se pomnože pojedinačni koeficijenti koristi c.

3.3. MJERA USLOVLJENOSTI MATRICE

Kvadratnoj matrici A biće pridružen broj u oznaci $\mathrm{cond}(A)$ (č. kondicija) ili $M(A)$, a naziva se njenom mjerom uslovljenosti ili kondicijom ili kondicionim brojem. Kod sistema linearnih jednačina $Ax = b$, greška ulaznih podataka odražava se na grešku numeričkog rezultata u stepenu koji zavisi od mjere uslovljenosti matrice A: što je mjera uslovljenosti veća to se više odražava. Može se pokazati da slične okolnosti važe i kada je riječ o greški računanja.

Definicija mjere uslovljenosti matrice. U skupu R^n, norma vektora x u oznaci $\|x\|$ može da bude uvedena na razne načine; neka $\|x\|$ označava jednu moguću normu za koju smo se opredijelili. Znamo da važi $\|x + y\| \leq \|x\| + \|y\|$, nejednakost trougla. Neka je $A \in R^{n\times n}$ kvadratna matrica oblika $n \times n$ ili neka je A: $R^n \to R^n$ linearni operator u prostoru R^n. Neka $\|A\|$ označava normu linearnog operatora A koja je saglasna sa uvedenom normom vektora (koja odgovara uvedenoj normi vektora), a definiše se relacijom $\|A\| = \sup_{x\neq 0} \|Ax\|/\|x\|$; tzv. indukovana norma. Vidimo da važi nejednakost $\|Ax\| \leq \|A\|\cdot\|x\|$. Označimo jediničnu matricu kao $E \in R^{n\times n}$; važi $\|E\| = 1$. Ako $A \in R^{n\times n}$ i $B \in R^{n\times n}$ onda važi $\|A + B\| \leq \|A\| + \|B\|$ i $\|AB\| \leq \|A\| \cdot \|B\|$.

Definicija. Za $A \in R^{n\times n}$, $\det A \neq 0$, stavlja se da je

$$\mathrm{cond}(A) = \|A\| \cdot \|A^{-1}\|.$$

Ako je $\det A = 0$ onda $\mathrm{cond}(A)$ nije definisano. Neki uzimaju da je tada $\mathrm{cond}(A) = +\infty$.

Koliko najmanje može da bude $\mathrm{cond}(A)$? Iz $A \cdot A^{-1} = E$ imamo $\|A \cdot A^{-1}\| = \|E\| = 1$ i $\|A\| \cdot \|A^{-1}\| \geq 1$ tj. $\mathrm{cond}(A) \geq 1$.

Ako je broj $\mathrm{cond}(A)$ mali tj. ako je broj $\mathrm{cond}(A)$ blizu donje granice 1 onda se za matricu A kaže da je ona dobro uslovljena. Tada je odgovarajući sistem linearnih jednačina $Ax = b$ manje osjetljiv na grešku ulaznih podataka i na grešku računanja, kao što ćemo vidjeti. Ako je broj $\mathrm{cond}(A)$ veliki tj. ako je broj $\mathrm{cond}(A)$ blizu gornje granice $+\infty$ onda se za matricu A kaže da je slabo uslovljena. Tada odgovarajući sistem linearnih jednačina $Ax = b$ ima slaba svojstva numeričke stabilnosti tj. sistem je nepogodan za numeričko rješavanje, kao što ćemo vidjeti.

Navedimo jedno elementarno svojstvo karakteristike $\mathrm{cond}(A)$. Prethodno, ako je $Ax = \lambda x$, gdje je $x \neq 0$, onda se za broj λ kaže da je svojstvena vrijednost matrice A, a za x se kaže da je odgovarajući svojstveni vektor, kao što je poznato iz linearne algebre. Lako se vidi da je $|\lambda| \leq \|A\|$, riječima: norma matrice je veća ili jednaka od apsolutne vrijednosti bilo koje njene svojstvene vrijednosti. Ima li kakve veze između svojstvenih vrijednosti dvije matrice A i A^{-1}?

$$Ax = \lambda x, \;\; A^{-1}Ax = A^{-1}\lambda x, \;\; x = \lambda A^{-1}x, \;\; A^{-1}x = \frac{1}{\lambda}x,$$

riječima: svojstvene vrijednosti inverzne matrice jednake su recipročnim vrijednostima svojstvenih vrijednosti same matrice, sa jednim te istim odgovarajućim svojstvenim vektorom. Važi: $\det A = 0 \Leftrightarrow \lambda = 0$ je svojstvena vrijednost matrice A. Svojstvo $\mathrm{cond}(A)$:

$$\mathrm{cond}(A) \geq \frac{|\lambda_2|}{|\lambda_1|},$$

gdje su $|\lambda_2|$ i $|\lambda_1|$ redom najveći odnosno najmanji među svim brojevima $|\lambda|$, λ – svojstvena vrijednost matrice A. Zaista,

$$\|A\| \geq |\lambda_2|, \;\; \|A^{-1}\| \geq \frac{1}{|\lambda_1|} \;\; \Rightarrow \;\; \|A\| \cdot \|A^{-1}\| \geq \frac{|\lambda_2|}{|\lambda_1|}.$$

Provjeri navedeno svojstvo $\mathrm{cond}(A)$ u slučaju $\|x\| = \|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$, A – dijagonalna matrica.

Razmotrimo zadatak o rješavanju sistema linearnih jednačina $Ax = b$ čiji su ulazni podaci A i b samo približno poznate veličine. I rješenje x ćemo saznati samo približno, greška ulaznih podataka očito utiče na grešku rješenja. Uvedimo potrebne oznake. Razmotrimo sistem $Ax = b$, gdje se pretpostavlja da je $\det A \neq 0$. Neka je A^* raspoloživa približna vrijednost matrice sistema. Uvedimo oznaku δA za odgovarajuće odstupanje tj. neka bude $A = A^* + \delta A$. Neka je b^* raspoloživa približna vrijednost vektora slobodnih članova. Uvedimo oznaku δb za odgovarajuće odstupanje tj. neka bude $b = b^* + \delta b$. Pretpostavimo da je i $\det A^* \neq 0$. Mi riješimo sistem $A^* x^* = b^*$ (sa x^* smo označili njegovo rješenje) i saopštimo naš numerički odgovor $x \approx x^*$. Kolika je greška? Označimo odgovarajuće odstupanje kao δx tj. neka bude $x = x^* + \delta x$; $\delta x \in R^n$ je apsolutna greška numeričkog odgovora. Dok su $\delta A \in R^{n \times n}$ i $\delta b \in R^n$ očito apsolutne greške matrice sistema i vektora slobodnih članova, redom. Možemo razmatrati i relativne greške $\|\delta x\|/\|x\|$, $\|\delta A\|/\|A\|$ i $\|\delta b\|/\|b\|$. Sljedeća teorema izražava relativnu grešku rješenja $\|\delta x\|/\|x\|$ preko relativne greške ulaznih podataka i veličine $\mathrm{cond}(A)$. Prije teoreme dolazi jedna lema.

Banahova lema. Neka je C kvadratna matrica koja zadovoljava $\|C\| < 1$. Tada postoji matrica $(E - C)^{-1}$ i važi ocjena

$$\|(E - C)^{-1}\| \leq \frac{1}{1 - \|C\|}.$$

Dokaz leme. Za bilo koji vektor x imamo

$$\|(E - C)x\| = \|x - Cx\| \geq \|x\| - \|Cx\| \geq \|x\| - \|C\| \cdot \|x\| = \delta \cdot \|x\|,$$

gdje je $\delta = 1 - \|C\| > 0$. Ako je $(E - C)x = 0$ onda je $0 \geq \delta \cdot \|x\| \Rightarrow \|x\| = 0$, $x = 0$. Vidimo da homogeni sistem $(E - C)x = 0$ ima samo trivijalno rješenje. Tako da je matrica $E - C$ regularna (invertibilna). Zato u nejednakosti $\|(E - C)x\| \geq \delta \cdot \|x\|$ možemo da uvedemo oznake $(E - C)x = y$ i $x = (E - C)^{-1} y$, pa ta nejednakost dobija oblik

$$\|y\| \geq \delta \cdot \|(E - C)^{-1} y\| \;\; \Rightarrow \;\; \|(E - C)^{-1} y\| \leq \frac{1}{\delta} \|y\|.$$

Budući da vektor y prolazi kroz čitav skup R^n to slijedi

$$\|(E - C)^{-1}\| \leq \frac{1}{\delta} = \frac{1}{1 - \|C\|}.$$

Lema je dokazana. E – jedinična matrica.

Teorema. Neka matrica A^{-1} postoji i neka je $\|\delta A\| \cdot \|A^{-1}\| < 1$. Tada postoji i matrica $(A - \delta A)^{-1}$ i važi nejednakost

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{1}{1 - \mathrm{cond}(A) \cdot \frac{\|\delta A\|}{\|A\|}} \cdot \mathrm{cond}(A) \cdot \left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right), \tag{1}$$

gdje je $Ax = b$ i $(A - \delta A)(x - \delta x) = b - \delta b$.

Dokaz teoreme. $Ax = b \quad A^*x^* = b^* \quad A = A^* + \delta A \quad b = b^* + \delta b \quad x = x^* + \delta x$

$$(A - \delta A) \cdot (x - \delta x) = b - \delta b$$

$$Ax - \delta A \cdot x - A \cdot \delta x + \delta A \cdot \delta x = b - \delta b \quad / \cdot (-1)$$

$$(A - \delta A) \cdot \delta x = \delta b - \delta A \cdot x \quad / \cdot A^{-1}$$

$$(E - A^{-1}\delta A) \cdot \delta x = A^{-1}\delta b - A^{-1}\delta A \cdot x \qquad (E - A^{-1}\delta A \text{ je invertibilna; lema})$$

$$\delta x = (E - A^{-1}\delta A)^{-1}(A^{-1}\delta b - A^{-1}\delta A \cdot x)$$

$$\|\delta x\| \le \|(E - A^{-1}\delta A)^{-1}\| \cdot (\|A^{-1}\| \cdot \|\delta b\| + \|A^{-1}\| \cdot \|\delta A\| \cdot \|x\|) \qquad \left(\frac{\|Ax\|}{\|b\|} = 1 \quad \Rightarrow \quad \frac{\|A\| \cdot \|x\|}{\|b\|} \ge 1\right)$$

$$\|\delta x\| \le \|(E - A^{-1}\delta A)^{-1}\| \cdot \left(\|A^{-1}\| \cdot \|\delta b\| \cdot \frac{\|A\| \cdot \|x\|}{\|b\|} + \|A^{-1}\| \cdot \|\delta A\| \cdot \|x\| \cdot \frac{\|A\|}{\|A\|}\right) \qquad \text{(lema)}$$

$$\|\delta x\| \le \frac{1}{1 - \|A^{-1}\delta A\|} \cdot \left(\text{cond}(A) \cdot \frac{\|\delta b\|}{\|b\|} \cdot \|x\| + \text{cond}(A) \cdot \frac{\|\delta A\|}{\|A\|} \cdot \|x\|\right) \quad / : \|x\|$$

$$\frac{\|\delta x\|}{\|x\|} \le \frac{1}{1 - \|A^{-1}\| \cdot \|\delta A\|} \cdot \text{cond}(A) \cdot \left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right)$$

Teorema je dokazana, budući da je

$$\|A^{-1}\| \cdot \|\delta A\| = \|A^{-1}\| \cdot \|\delta A\| \cdot \frac{\|A\|}{\|A\|} = \text{cond}(A) \cdot \frac{\|\delta A\|}{\|A\|}$$

Pogledajmo nejednakost (1) kada je $\delta A \approx 0$ i $1 - \text{cond}(A) \cdot \frac{\|\delta A\|}{\|A\|} \approx 1$. Vidimo da se tada relativna greška rješenja $\frac{\|\delta x\|}{\|x\|} = \frac{\|x - x^*\|}{\|x\|}$ procjenjuje mjerom uslovljenosti matrice sistema $\text{cond}(A)$ puta zbir relativnih grešaka matrice sistema i vektora slobodnih članova $\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}$.

Specijalan slučaj teoreme dobijamo kada stavimo da je $\|\delta A\| = 0$, matrica sistema je data tačno. Vidimo da se tada relativna greška desne strane sistema (vektora slobodnih članova) prenosi na relativnu grešku rješenja sistema sa koeficijentom uveličavanja koji je jednak $\text{cond}(A)$. Formulom, ako je $Ax = b$, $Ax^* = b^*$, $b = b^* + \delta b$ i $x = x^* + \delta x$ onda je

$$\frac{\|\delta x\|}{\|x\|} \le \text{cond}(A) \cdot \frac{\|\delta b\|}{\|b\|}. \tag{2}$$

Drugi specijalan slučaj teoreme dobijamo ako stavimo da je $\|\delta b\| = 0$. Sada je desna strana sistema data tačno, dok je matrica sistema poznata samo približno tačno. Matrica sistema poznata je sa izvjesnom greškom, ona je poznata sa relativnom greškom $\|\delta A\|/\|A\|$.

Ta relativna greška se prenosi (odražava) na relativnu grešku rješenja sistema sa koeficijentom uveličavanja koji je jednak otprilike $\mathrm{cond}(A)$.

Primjedba. Prilikom rješavanja sistema $Ax = b$ direktnom metodom, računar ne saopštava tačno rješenje x, već će računar saopštiti približno rješenje x^*, zbog prisustva greške računanja. Kako da se ocijeni greška $x - x^*$? Uvrstiti x^* u sistem tj. izračunati vektor $b - Ax^* = b - b^* = \delta b$. Za vektor δb kaže se da predstavlja tzv. ekvivalentnu smetnju. Ono što je računar uradio ekvivalentno je sa time da je tačno (bez greške računanja) riješen sistem čiji je vektor slobodnih članova dat približno. Dovoljno je da se primijeni formula (2).

3.4. ITERATIVNE METODE ZA RJEŠAVANJE SISTEMA LINEARNIH JEDNAČINA

Biće riječi o metodi proste iteracije i o jednom njenom posebnom slučaju (o tzv. Jakobijevoj metodi), a prethodno o normi u konačno–dimenzionom prostoru.

Neka bude $n \geq 1$ i razmotrimo skup R^n. Postoje razne funkcije $\| \ \|: R^n \to R$ koje ispunjavaju aksiome norme. Skup R^n zajedno sa jednom takvom normom postaje normirani prostor.

Za dvije norme $\| \ \|_a$ i $\| \ \|_b$ koje su definisane u jednom te istom skupu X kaže se da su ekvivalentne ako postoje konstante $\alpha > 0$ i $\beta > 0$ takve da važi $\alpha\|x\|_a \leq \|x\|_b \leq \beta\|x\|_a$, za svako $x \in X$. Poznato je sljedeće tvrđenje: bilo koje dvije norme u skupu R^n su ekvivalentne. Zato pitanje konvergencije niza $\{x^{(k)}\}_{k=1}^{\infty}$ čiji članovi pripadaju R^n ne zavisi od izbora norme i svodi se na pitanje konvergencije po koordinatama (konvergencije n brojnih nizova). Dakle, ako niz konvergira po jednoj normi onda taj niz konvergira i po bilo kojoj drugoj normi, tako da se tada prosto kaže da niz konvergira.

Tri norme u prostoru R^n koje se često koriste označavaju se kao $\| \ \|_1$, $\| \ \|_2$ i $\| \ \|_\infty$ a definišu se sa

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \quad \|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} \quad \text{i} \quad \|x\|_\infty = \max_{1 \leq i \leq n} |x_i|,$$

gdje je $x = (x_1, \ldots, x_n) \in R^n$. Znamo da su prva i druga norma specijalni slučajevi norme $\|x\|_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$, gdje je $p \geq 1$. Isto tako znamo da važi $\lim_{p \to +\infty} \|x\|_p = \|x\|_\infty$. Slično je $\| \ \|_p$ norma i u skupu C^n, gdje je $1 \leq p \leq +\infty$.

Razmotrimo matricu $A = [a_{ij}]_{i,j=1}^{n} \in R^{n \times n}$. Relacija $\|A\| = \sup_{x \neq 0} \|Ax\|/\|x\|$ definiše normu matrice A koja je indukovana normom vektora $x \in R^n$ u istoj oznaci $\|x\|$. Za tri uobičajene norme u R^n imamo sljedeće eksplicitne izraze za odgovarajuće norme matrice:

$$\|A\|_1 = \max_{1 \leq j \leq n} \left(\sum_{i=1}^{n} |a_{ij}| \right), \quad \|A\|_2 = \sqrt{\max_{1 \leq i \leq n} \lambda_i(A^T A)} \quad \text{i} \quad \|A\|_\infty = \max_{1 \leq i \leq n} \left(\sum_{j=1}^{n} |a_{ij}| \right);$$

trebalo bi ovo dokazati; A^T je transponovana matrica matrice A, a $\lambda_i(A^T A)$ su svojstvene vrijednosti matrice $A^T A$. Ako je matrica A simetrična ($A = A^T$) onda važi $\|A\|_2 = \max_{1 \leq i \leq n} |\lambda_i(A)|$, gdje su sa $\lambda_i(A)$ označene svojstvene vrijednosti matrice A.

Brojevi λ^2 su svojstvene vrijednosti matrice A^2.

U slučaju $A \in C^{n \times n}$ ne mijenja se ništa kod $\|A\|_1$ i $\|A\|_\infty$ a mijenja se samo malo kod $\|A\|_2$ kako slijedi: $\|A\|_2 = \sqrt{\max_{1 \leq i \leq n} \lambda_i(A^* A)}$, gdje je A^* konjugovana matrica. Ako je A samokonjugovana ($A = A^*$) onda važi $\|A\|_2 = \max_{1 \leq i \leq n} |\lambda_i(A)|$.

Za iterativnu metodu se kaže da je ona metoda proste iteracije ako se ona zasniva na Banahovoj teoremi o nepokretnoj tački.

Neka su dati matrica $A \in R^{n\times n}$ i vektor $b \in R^n$. Razmotrimo sistem linearnih jednačina

$$Ax = b. \tag{1}$$

Sistem (1) transformisati u njemu ekvivalentni sistem oblika

$$x = Bx + c, \tag{2}$$

$B \in R^{n\times n}$, $c \in R^n$; sistemi (1) i (2) su ekvivalentni – oni imaju jedna te ista rješenja $x \in R^n$.

Uzmimo da sistem ili jednačina (1) odnosno (2) ima jedinstveno rješenje x. Za iterativne metode (za metode uzastopnih ili sukcesivnih aproksimacija) karakteristično je da se konstruiše niz vektora (iterativni niz) $\{x^{(k)}\}_{k=0}^{\infty}$ čiji članovi pripadaju R^n koji bi trebalo da konvergira ka rješenju tj. trebalo bi da bude $\lim_{k\to\infty} x^{(k)} = x$ ili svejedno $\lim_{k\to\infty} \|x^{(k)} - x\| = 0$.

Kada se kaže da se rješava jednačina $x = Bx + c$ i da se primjenjuje metoda proste iteracije tada se ustvari već podrazumijeva da su članovi iterativnog niza definisani relacijom

$$x^{(k+1)} = Bx^{(k)} + c \ \text{ za } k = 0, 1, 2, \ldots. \tag{3}$$

Niz će biti definisan kada se još odabere i početna aproksimacija $x^{(0)} \in R^n$.

Sljedeća teorema predstavlja ustvari Banahovu teoremu o mepokretnoj tački u slučaju preslikavanja φ koje djeluje u prostoru R^n i koje je afino: $\varphi(x) = Bx + c$. Fiksirajmo u R^n jednu normu $\|\ \|$ i sa $\|B\|$ označimo naravno saglasnu normu matrice B.

Teorema o dovoljnim uslovima za konvergenciju metode proste iteracije. Ako je $\|B\| < 1$ onda: a) sistem (2) ima jedinstveno rješenje x, b) iterativni niz (3) konvergira ka x za bilo koju početnu aproksimaciju $x^{(0)} \in R^n$ i c) postoje konstante $\gamma > 0$ i $0 < q < 1$ takve da je $\|x^{(k)} - x\| \le \gamma q^k$ za svako $k \ge 0$.

Dokaz teoreme. Po uslovu $\|B\| < 1$ i po Banahovoj lemi iz prethodnog naslova slijedi da je matrica $E - B$ invertibilna. Znači da sistem $(E - B)x = c$ ima jedinstveno rješenje, čime je a) dokazano. Uvedimo oznaku $r^{(k)}$ za grešku k–te aproksimacije tj. stavimo $r^{(k)} = x - x^{(k)}$. Imamo redom:

$$x = Bx + c \ \text{ i } \ x^{(k+1)} = Bx^{(k)} + c \ \ \Rightarrow \ \ x - x^{(k+1)} = Bx - Bx^{(k)}, \ \ r^{(k+1)} = Br^{(k)}$$

$$r^{(1)} = Br^{(0)}, \ \ r^{(2)} = Br^{(1)}, \ \ \ldots \ \ \Rightarrow \ \ r^{(k)} = B^k r^{(0)}$$

$$\|r^{(k)}\| = \|B^k r^{(0)}\| \le \|B^k\| \cdot \|r^{(0)}\| \le \|B\|^k \cdot \|r^{(0)}\|, \quad \lim_{k\to\infty} r^{(k)} = 0$$

Dakle, dobili smo da je $\lim_{k\to\infty} x^{(k)} = x$, čime je dokazano b). Vidimo da je i c) već dokazano, sa $\gamma = \|r^{(0)}\| = \|x - x^{(0)}\|$ i $q = \|B\| < 1$. Teorema je dokazana.

Ako je $\|B\| \le q$ i $\varphi(x) = Bx + c$ onda očito $\|\varphi(y) - \varphi(x)\| = \|By + c - Bx - c\| = \|B(y - x)\| \le \|B\| \cdot \|y - x\| \le q\|y - x\|$. Znači, φ je kontrakcija ako je $q < 1$.

Kolika je greška? Mi računamo: $x - x^{(0)} = x - x^{(1)} + x^{(1)} - x^{(0)}$

$$\|x - x^{(0)}\| \le \|x - x^{(1)}\| + \|x^{(1)} - x^{(0)}\| \le q\|x - x^{(0)}\| + \|x^{(1)} - x^{(0)}\|$$

$$(1 - q)\|x - x^{(0)}\| \le \|x^{(1)} - x^{(0)}\|$$

(ranije smo pokazali da je $\gamma = \|x - x^{(0)}\|$)

$$\|x^{(k)} - x\| \leq \frac{q^k}{1-q}\|x^{(1)} - x^{(0)}\| \tag{4}$$

Mi računamo: $x - x^{(k)} = x - x^{(k+1)} + x^{(k+1)} - x^{(k)}$ (po aksiomi trougla $\Rightarrow$)

$$\|x - x^{(k)}\| \leq \|x - x^{(k+1)}\| + \|x^{(k+1)} - x^{(k)}\|$$

(imamo da je $\varphi(x) = x$, $\varphi(x^{(k-1)}) = x^{(k)}$ i $\varphi(x^{(k)}) = x^{(k+1)}$)

$$\|x - x^{(k)}\| \leq q\|x - x^{(k)}\| + q\|x^{(k)} - x^{(k-1)}\|$$

(jer φ ima koeficijent kontrakcije q)

$$\|x - x^{(k)}\| \leq \frac{q}{1-q}\|x^{(k)} - x^{(k-1)}\| \tag{5}$$

Formule (4) i (5) služe za ocjenu greške k–te aproksimacije $x^{(k)}$, za bilo koje $k \geq 1$. Iz koraka u korak, greška se množi sa q. Tako da je tempo ili brzima konvergencije prvog reda ili linearna.

Analiza teoreme. Može se desiti da je po nekoj normi $\|B\| < 1$ a po nekoj drugoj normi da nije $\|B\| < 1$. Dovoljno je da po jednoj normi bude $\|B\| < 1$, da bi iterativni niz konvergirao ka rješenju, u vezi međusobne ekvivalentnosti svih vektorskih normi nad R^n, o čemu je bilo riječi na početku. Nema protivrječnosti u tome što dovoljan uslov $\|B\| < 1$ može da bude ispunjen u jednoj normi a neispunjen u nekoj drugoj normi.

Kako da se datom sistemu (1) pridruži ekvivalentan sistem oblika (2)? Za ovo postoji više načina, postoji beskonačno mnogo načina. Navedimo jedan način. Neka je $D \in R^{n\times n}$ bilo koja regularna matrica. Jednačina $Ax - b = 0$ očito je ekvivalentna jednačini $x = x + D(Ax - b)$. Odatle:

$$x = (E + DA)x - Db = Bx + c; \quad B = E + DA, \quad c = -Db$$

Primjer: Jakobijeva metoda. Neka polazni sistem (1) zadovoljava $a_{ii} \neq 0$ za $1 \leq i \leq n$. Tada se i–ta jednačina može podijeliti sa a_{ii} i onda iz i–te jednačne izraziti x_i. Vidimo da smo (1) sveli na oblik $x = Bx + c$ tj. da smo definisali iterativni niz $\{x^{(k)}\}_{k=0}^{\infty}$ čiji članovi pripadaju R^n; izabrati $x^{(0)} \in R^n$. Realizacija:

$$Ax = b \quad \text{ili} \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2 \\ \quad \ldots \end{cases}$$

$$x_i = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}}x_j - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}}x_j + \frac{b_i}{a_{ii}} \text{ za } 1 \leq i \leq n \quad \text{ili} \quad x = Bx + c \quad \text{ili}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} & \dots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} & \dots & -\frac{a_{2n}}{a_{22}} \\ \dots & \dots & & & \end{bmatrix}}_{B} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} + \underbrace{\begin{bmatrix} \frac{b_1}{a_{11}} \\ \frac{b_2}{a_{22}} \\ \vdots \end{bmatrix}}_{c}$$

$$x^{(k+1)} = Bx^{(k)} + c \quad \text{ili} \quad \begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ \vdots \end{bmatrix} = B \cdot \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \\ \vdots \\ x_n^{(k)} \end{bmatrix} + c \ \text{ za } \ k \geq 0 \tag{6}$$

Da li Jakobijev iterativni proces (6) konvergira? Pokušajmo da primijenimo teoremu od maločas. Da li je $\|B\| < 1$? Izaberimo $\| \ \| = \| \ \|_\infty$. Imamo redom:

$$\|B\|_\infty = \max_{1 \leq i \leq n} \left\{ \sum_{j=1}^{n} |b_{ij}| \right\} =$$

$$\max \left\{ \left| -\frac{a_{12}}{a_{11}} \right| + \left| -\frac{a_{13}}{a_{11}} \right| + \ldots + \left| -\frac{a_{1n}}{a_{11}} \right|, \left| -\frac{a_{21}}{a_{22}} \right| + \left| -\frac{a_{23}}{a_{22}} \right| + \ldots + \left| -\frac{a_{2n}}{a_{22}} \right|, \ldots \right\} =$$

$$\max \left\{ \frac{|a_{12}| + |a_{13}| + \ldots + |a_{1n}|}{|a_{11}|}, \frac{|a_{21}| + |a_{23}| + \ldots + |a_{2n}|}{|a_{22}|}, \ldots \right\}$$

Dovoljno je da matrica A bude dijagonalno dominantna. Mi smo dokazali tvrđenje: ako je matrica A dijagonalno dominantna ($\sum_{j=1 \ j \neq i}^{n} |a_{ij}| < |a_{ii}|$ za $1 \leq i \leq n$) onda Jakobijev iterativni proces konvergira. Završen primjer.

U nastavku želimo da nađemo tačne tj. neophodne i dovoljne uslove da iterativni niz (3) konvergira, bez obzira na izbor $x^{(0)} \in R^n$ tj. za ma kakvu početnu aproksimaciju.

Iz linearne algebre je poznato sljedeće: za svaku matricu $P \in R^{n \times n}$ postoji matrica $Q \in R^{n \times n}$ pomoću koje se P prevodi u njenu Žordanovu kanonsku formu $Q^{-1}PQ$ u smislu da važi jednakost

$$Q^{-1}PQ = \begin{bmatrix} \lambda_1(P) & \alpha_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \lambda_2(P) & \alpha_2 & 0 & \dots & 0 & 0 \\ \dots & \dots & & & & & \\ 0 & 0 & 0 & 0 & \dots & \lambda_{n-1}(P) & \alpha_{n-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & \lambda_n(P) \end{bmatrix}$$

gdje su $\lambda_i(P) \in C$ svojstvene vrijednosti matrice P, a $\alpha_i \in \{0, 1\}$; za matrice P i $Q^{-1}PQ$ kaže se da su slične.

Lema. Neka za matricu $B \in R^{n\times n}$ važi $\max_{1\le i\le n} |\lambda_i(B)| < q$. Tada postoji matrica $D \in R^{n\times n}$ takva da je $\|D^{-1}BD\|_\infty \le q$.

Dokaz leme. Stavimo $\eta = q - \max_{1\le i\le n} |\lambda_i(B)| > 0$. Želimo da odredimo Žordanovu kanonsku formu matrice $\eta^{-1}B$. Dakle, postoji matrica D takva da važi:

$$D^{-1}(\eta^{-1}B)D = \begin{bmatrix} \lambda_1(\eta^{-1}B) & \alpha_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \lambda_2(\eta^{-1}B) & \alpha_2 & 0 & \dots & 0 & 0 \\ \dots & \dots & & & & & \\ 0 & 0 & 0 & 0 & \dots & \lambda_{n-1}(\eta^{-1}B) & \alpha_{n-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & \lambda_n(\eta^{-1}B) \end{bmatrix}$$

$$= \begin{bmatrix} \eta^{-1}\lambda_1(B) & \alpha_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \eta^{-1}\lambda_2(B) & \alpha_2 & 0 & \dots & 0 & 0 \\ \dots & \dots & & & & & \\ 0 & 0 & 0 & 0 & \dots & \eta^{-1}\lambda_{n-1}(B) & \alpha_{n-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & \eta^{-1}\lambda_n(B) \end{bmatrix} \Big/ \cdot \eta$$

$$D^{-1}BD = \begin{bmatrix} \lambda_1(B) & \eta\alpha_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \lambda_2(B) & \eta\alpha_2 & 0 & \dots & 0 & 0 \\ \dots & \dots & & & & & \\ 0 & 0 & 0 & 0 & \dots & \lambda_{n-1}(B) & \eta\alpha_{n-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & \lambda_n(B) \end{bmatrix}$$

Ovdje $\alpha_i \in \{0, 1\}$.

Znamo sljedeće: ako je λ svojstvena vrijednost matrice A onda je $c\lambda$ svojstvena vrijednost matrice cA, što je maločas upotrebljeno. Zaista, $Ax = \lambda x \Rightarrow (cA)x = (c\lambda)x$.

Dalje:

$$\|D^{-1}BD\|_\infty = \max\Big\{|\lambda_1(B)| + \eta\alpha_1, |\lambda_2(B)| + \eta\alpha_2, \dots, |\lambda_{n-1}(B)| + \eta\alpha_{n-1}, |\lambda_n(B)|\Big\} \le$$

$$\max\Big\{|\lambda_1(B)| + \eta, |\lambda_2(B)| + \eta, \dots, |\lambda_{n-1}(B)| + \eta, |\lambda_n(B)|\Big\} \le \max_{1\le i\le n} |\lambda_i(B)| + \eta = q$$

Lema je dokazana.

Teorema o neophodnim i dovoljnim uslovima za konvergenciju metode proste iteracije. Neka sistem (2) ima jedinstveno rješenje x. Iterativni proces (3) konvergira ka x (za ma kakvu početnu aproksimaciju $x^{(0)}$) ako i samo ako su sve svojstvene vrijednosti matrice B po apsolutnoj vrijednosti < 1.

Dokaz teoreme. Uslov je dovoljan. Dato je $|\lambda_i(B)| < 1$. Uzmimo proizvoljno q u granicama $\max_{1\le i\le n} |\lambda_i(B)| < q < 1$. Kako su uslovi prethodne leme ispunjeni za matricu B i broj q to postoji matrica D takva da važi $\|D^{-1}BD\|_\infty \le q$ ili svejedno $\|\Lambda\|_\infty \le q$, gdje je uvedena oznaka $\Lambda = D^{-1}BD$. Imamo redom:

$$\Lambda = D^{-1}BD \quad \Rightarrow \quad B = D\Lambda D^{-1}$$

$$B^2 = B \cdot B = D\Lambda D^{-1} \cdot D\Lambda D^{-1} = D\Lambda^2 D^{-1}$$

$$B^3 = D\Lambda^3 D^{-1}, \ \ldots, \ B^k = D\Lambda^k D^{-1}, \ \ldots$$

$$\|B^k\|_\infty = \|D\Lambda^k D^{-1}\|_\infty \le \|D\|_\infty \cdot \|\Lambda^k\|_\infty \cdot \|D^{-1}\|_\infty \le$$

$$\|D\|_\infty \cdot \|\Lambda\|_\infty^k \cdot \|D^{-1}\|_\infty \le \|D\|_\infty \cdot q^k \cdot \|D^{-1}\|_\infty$$

Odaberimo $x^{(0)} \in R^n$, čime je niz $x^{(k)} = Bx^{(k-1)} + c$ definisan. U prethodnoj teoremi je pokazano da važi $r^{(k)} = B^k r^{(0)}$, gdje je uvedena oznaka $r^{(k)} = x - x^{(k)}$. Slijedi

$$\|r^{(k)}\|_\infty = \|B^k r^{(0)}\|_\infty \le \|B^k\|_\infty \cdot \|r^{(0)}\|_\infty \le \|D\|_\infty \cdot q^k \cdot \|D^{-1}\|_\infty \cdot \|r^{(0)}\|_\infty$$

$$\lim_{k\to\infty} \|r^{(k)}\|_\infty = 0$$

Dokazali smo da je $\lim_{k\to\infty} \|x - x^{(k)}\|_\infty = 0$ tj. da je $\lim_{k\to\infty} x^{(k)} = x$ po normi $\| \ \|_\infty$, $\forall x^{(0)}$. Ustvari smo dokazali da je $\lim_{k\to\infty} \|x - x^{(k)}\| = 0$ gdje je $\| \ \|$ bilo koja norma tj. da je $\lim_{k\to\infty} x^{(k)} = x$ po bilo kojoj normi, $\forall x_0$, u vezi ekvivalentnosti svih normi u konačno–dimenzionom prostoru.

Uslov je neophodan. Dopustimo da matrica B ima svojstvenu vrijednost $\lambda \in C$ takvu da je $|\lambda| \ge 1$. Označimo sa $v \in C^n$ odgovarajući svojstveni vektor: $Bv = \lambda v$, $v \ne 0$. Treba pokazati da se može naći bar jedna početna aproksimacija $x^{(0)}$ takva da odgovarajući iterativni niz $x^{(k)} = Bx^{(k-1)} + c$ ne konvergira ka x kad $k \to \infty$. Odaberimo $x^{(0)} = x - v$. Imamo

$$x^{(1)} = Bx^{(0)} + c = B(x - v) + c = Bx - Bv + c = x - Bv = x - \lambda v$$

$$x^{(2)} = Bx^{(1)} + c = B(x - \lambda v) + c = Bx - \lambda Bv + c = x - \lambda Bv = x - \lambda^2 v, \ \ldots$$

$$x^{(k)} = x - \lambda^k v, \ \ldots$$

$$r^{(k)} = x - x^{(k)} = x - x + \lambda^k v = \lambda^k v$$

$$\|r^{(k)}\| = \|\lambda^k v\| = |\lambda|^k \cdot \|v\| \ge \|v\| > 0, \quad \text{nije} \ \lim_{k\to\infty} \|r^{(k)}\| = 0$$

$$\text{nije} \ \lim_{k\to\infty} \|x - x^{(k)}\| = 0 \quad \left(\text{nije} \ \lim_{k\to\infty} x^{(k)} = x\right)$$

Teorema je dokazana.

3.5. ZAJDELOVA METODA

Razmatra se sistem oblika $Ax = b$ tj. $a_{11}x_1 + \ldots + a_{1n}x_n = b_1, \ldots, a_{n1}x_1 + \ldots + a_{nn}x_n = b_n$. U slučaju Jakobijeve metode:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(-\sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} + b_i \right), \ \ 1 \le i \le n, \ \ k \ge 0. \tag{1}$$

A u slučaju Zajdelove metode:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(-\sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} + b_i \right), \ \ 1 \le i \le n, \ \ k \ge 0. \tag{2}$$

Zajdelova metoda je primjer iterativne metode za rješavanje sistema linearnih jednačina $Ax = b$; $A = [a_{ij}]_{i,j=1}^{n} \in R^{n\times n}$, $b = [b_i]_{i=1}^{n} \in R^n$, $n \geq 1$. Zajdelova metoda predstavlja malu modifikaciju ili malo poboljšanje Jakobijeve metode. Za jednu i drugu metodu pretpostavlja se da je $a_{ii} \neq 0$ za $1 \leq i \leq n$. Znamo da "matrica A je dijagonalno dominantna" predstavlja dovoljan uslov za konvergenciju Jakobijeve metode, a vidjećemo da je to i dovoljan uslov za konvergenciju Zajdelove metode. Ako je matrica $A \in R^{n\times n}$ dijagonalno dominantna onda je $\det A \neq 0$; ovo je dokazano ranije, kod kubnog splajna (Interpolacija pomoću splajna).

Neka je matrica A sistema linearnih jednačina $Ax = b$ koji se rješava dijagonalno dominantna:

$$\sum_{j=1\ j\neq i}^{n} |a_{ij}| < |a_{ii}| \quad \text{za} \quad 1 \leq i \leq n. \tag{3}$$

Tada sistem $Ax = b$ ima jedinstveno rješenje $x = (x_1, \ldots, x_n) \in R^n$. Za Jakobijevu metodu važi formula (1). Kod Zajdelove metode računanja se sprovode po formuli (2). Iterativni niz $\{x^{(k)}\}_{k=0}^{\infty}$, gdje $x^{(k)} \in R^n$, biće definisan kada se izabere $x^{(0)} = (x_1^{(0)}, \ldots, x_n^{(0)}) \in R^n$. Može se uzeti $x^{(0)} = (0, \ldots, 0)$.

Dakle, u formuli (2), neka je u toku računanje $(k+1)$–ve iteracije $x^{(k+1)} = (x_1^{(k+1)}, \ldots, x_n^{(k+1)})$ i neka se konkretno u datom trenutku računa komponenta $x_i^{(k+1)}$. Već su poznate komponente $x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}$ vektora $x^{(k+1)}$. Upotrebiti ih u datom trenutku, umjesto $x_1^{(k)}, \ldots, x_{i-1}^{(k)}$. Iskustvo pokazuje da Zajdelova metoda daje bolje rezultate od Jakobijeve metode (greška je manja, konvergencija je brža). Vidimo da se Zajdelova metoda lakše programira od Jakobijeve metode: čim je izračunato $x_i^{(k+1)}$ više nam ne treba $x_i^{(k)}$. Zato se u praksi koristi Zajdelova metoda, a ne Jakobijeva metoda.

Na osnovu (2) je

$$\begin{cases} a_{11}x_1^{(k+1)} + a_{12}x_2^{(k)} + a_{13}x_3^{(k)} + \ldots + a_{1n}x_n^{(k)} = b_1 \\ a_{21}x_1^{(k+1)} + a_{22}x_2^{(k+1)} + a_{23}x_3^{(k)} + \ldots + a_{2n}x_n^{(k)} = b_2 \\ \ldots \\ a_{n1}x_1^{(k+1)} + a_{n2}x_2^{(k+1)} + a_{n3}x_3^{(k+1)} + \ldots + a_{nn}x_n^{(k+1)} = b_n \end{cases}$$

ili $Bx^{(k+1)} + Cx^{(k)} = b$ za $k \geq 0$, gdje su uvedene oznake $A = B + C$ i

$$B = \begin{bmatrix} a_{11} & 0 & 0 & \ldots & 0 \\ a_{21} & a_{22} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & & \\ a_{n1} & a_{n2} & a_{n3} & \ldots & a_{nn} \end{bmatrix} \quad \text{i} \quad C = \begin{bmatrix} 0 & a_{12} & a_{13} & \ldots & a_{1n} \\ 0 & 0 & a_{23} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & & \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix}$$

Uvedimo oznaku $r^{(k)} = x - x^{(k)}$ za grešku k–te aproksimacije $x^{(k)}$.

Teorema. Pretpostavimo da je

$$\sum_{j=1\ j\neq i}^{n} |a_{ij}| \leq q|a_{ii}| \quad \text{za} \quad 1 \leq i \leq n, \tag{4}$$

gdje je $q < 1$. Tada važi nejednakost

$$\|r^{(k+1)}\|_\infty \leq q \cdot \|r^{(k)}\|_\infty \quad \text{za} \quad k \geq 0. \tag{5}$$

Dokaz teoreme. Imamo $Ax = b$ ili svejedno $Bx + Cx = b$ i imamo $Bx^{(k+1)} + Cx^{(k)} = b$. Oduzimanjem: $Bx - Bx^{(k+1)} + Cx - Cx^{(k)} = b - b$ ili

$$Br^{(k+1)} + Cr^{(k)} = 0, \tag{6}$$

čime smo dobili sistem koji se odnosi na grešku $r^{(k)}$. Posmatrajmo brojeve $|r_1^{(k+1)}|, \ldots, |r_n^{(k+1)}|$ i uočimo najveći među njima. Neka je najveći $|r_l^{(k+1)}|$, tako da je

$$\|r^{(k+1)}\|_\infty = \|(r_1^{(k+1)}, \ldots, r_n^{(k+1)})\|_\infty = \max_{1\le i\le n} |r_i^{(k+1)}| = |r_l^{(k+1)}|.$$

Napišimo l–tu jednačinu sistema (6):

$$\sum_{j=1}^{l-1} a_{lj} r_j^{(k+1)} + a_{ll} r_l^{(k+1)} + \sum_{j=l+1}^{n} a_{lj} r_j^{(k)} = 0 \quad / : a_{ll}$$

$$r_l^{(k+1)} = -\sum_{j=1}^{l-1} \frac{a_{lj}}{a_{ll}} r_j^{(k+1)} - \sum_{j=l+1}^{n} \frac{a_{lj}}{a_{ll}} r_j^{(k)}$$

$$\left|r_l^{(k+1)}\right| \le \sum_{j=1}^{l-1} \left|\frac{a_{lj}}{a_{ll}}\right| \cdot \left|r_j^{(k+1)}\right| + \sum_{j=l+1}^{n} \left|\frac{a_{lj}}{a_{ll}}\right| \cdot \left|r_j^{(k)}\right|$$

$$\left|r_l^{(k+1)}\right| \le \sum_{j=1}^{l-1} \left|\frac{a_{lj}}{a_{ll}}\right| \cdot \max_{1\le i\le n} |r_i^{(k+1)}| + \sum_{j=l+1}^{n} \left|\frac{a_{lj}}{a_{ll}}\right| \cdot \max_{1\le i\le n} |r_i^{(k)}|$$

$$|r_l^{(k+1)}| \le \alpha \cdot \|r^{(k+1)}\|_\infty + \beta \cdot \|r^{(k)}\|_\infty$$

gdje su uvedene oznake $\alpha = \sum_{j=1}^{l-1} \left|\frac{a_{lj}}{a_{ll}}\right|$ i $\beta = \sum_{j=l+1}^{n} \left|\frac{a_{lj}}{a_{ll}}\right|$

$$\|r^{(k+1)}\|_\infty \le \alpha \cdot \|r^{(k+1)}\|_\infty + \beta \cdot \|r^{(k)}\|_\infty \quad / : (1-\alpha)$$

$$\|r^{(k+1)}\|_\infty \le \frac{\beta}{1-\alpha} \|r^{(k)}\|_\infty$$

po uslovu teoreme je $\alpha + \beta \le q$ (u (4) stavi $i = l$) i $q < 1 \Rightarrow \frac{\beta}{1-\alpha} \le q$

$$\|r^{(k+1)}\|_\infty \le q \cdot \|r^{(k)}\|_\infty$$

Teorema je dokazana.

Uzastopnom primjenom nejednakosti (5) nalazimo da je $\|r^{(k)}\|_\infty \le q^k \cdot \|r^{(0)}\|_\infty$, tako da je $\lim_{k\to\infty} \|r^{(k)}\|_\infty = 0$. Pokazali smo da je $\lim_{k\to\infty} r^{(k)} = 0$ odnosno pokazali smo da je $\lim_{k\to\infty} x^{(k)} = x$. Dakle, ako je ispunjen uslov (4) onda Zajdelov iterativni proces konvergira ka rješenju.

Vidi se da je (3) $\Leftrightarrow$ (4).

3.6. PRIMJER ITERATIVNE METODE (ZA RJEŠAVANJE SISTEMA LINEARNIH JEDNAČINA) VARIJACIONOG TIPA

Za numeričku metodu se kaže da je varijaciona ako ona uključuje minimiziranje nekog funkcionala. Za preslikavanje φ se kaže da je funkcional ako su njegove vrijednosti realni brojevi. Postoje razne varijacione metode za rješavanje sistema linearnih jednačina. U ovom naslovu govori se o jednoj takvoj iterativnoj metodi. U ovom naslovu u prostoru R^n koristi se norma $\|\ \|_2$; $\|(x_1,\ldots,x_n)\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}$; biće označavana prosto kao $\|\ \|$. Znamo da je ta norma povezana sa skalarnim proizvodom $(x,y) = \sum_{i=1}^n x_i y_i$ relacijom $\|x\| = \sqrt{(x,x)}$. Iz linearne algebre je poznato sljedeće. Neka je matrica A simetrična ($A = A^T$). Tada A ima n realnih svojstvenih vrijednosti tj. sve njene svojstvene vrijednosti su realni brojevi; višestrukost svojstvenih vrijednosti je naravno uzeta u obzir. Uredimo ih po veličini: $\lambda_{\min}(A) = \lambda_1 \leq \ldots \leq \lambda_n = \lambda_{\max}(A)$. Neka je dodatno matrica A i pozitivno definitna, u oznaci $A > 0$, što znači da $x \neq 0 \Rightarrow (Ax,x) > 0$. Tada su sve njene svojstvene vrijednosti pozitivni brojevi ($0 < \lambda_1$) i njena norma $\|\ \|_2$ jednaka je najvećoj njenoj svojstvenoj vrijednosti; $\|A\| = \|A\|_2 = \lambda_n$. Još je i $\det A \neq 0$. Prelazimo na izlaganje metode minimalne nepovezanosti. (Ako treba riješiti $f(x) = b$ i ako odgovor glasi $x \approx x^*$ onda se kaže da $x - x^*$ predstavlja grešku i još se kaže da $f(x^*) - b$ predstavlja nepovezanost.) Neka $A \in R^{n\times n}$ i $b \in R^n$ i razmotrimo sistem

$$Ax = b.$$

Sistem $Ax = b$ ekvivalentan je sistemu

$$x = x + \tau(Ax - b),$$

za $\tau \neq 0$. Znači da uzastopne aproksimacije mogu da se računaju po formuli

$$x^{(k+1)} = x^{(k)} + \tau(Ax^{(k)} - b),\ \ k \geq 0.$$

Mi možemo da τ podešavamo od koraka do koraka. Mi ćemo uzastopne aproksimacije računati po formuli

$$x^{(k+1)} = x^{(k)} + \tau_{k+1}(Ax^{(k)} - b),\ \ k \geq 0. \qquad (1)$$

(Slično kao gradijentne metode za minimizaciju funkcije f: $R^n \to R$: od tačke $x^{(k)} \in R^n$ preći određeni put po polupravoj čiji je smjer suprotan smjeru gradijenta $\mathrm{grad}(f(x^{(k)}))$; na kraju tog puta je $x^{(k+1)} \in R^n$. Smjer antigradijenta je smjer najbržeg opadanja funkcije.) Niz $\{x^{(k)}\}_{k=0}^{\infty}$ je definisan ako se izabere $x^{(0)} \in R^n$. Neka x označava rješenje sistema $Ax = b$. Možemo sa $r^{(k)} = x - x^{(k)}$ da označimo grešku k–te aproksimacije $x^{(k)}$. Bolje je što je greška manja. Ako se u izrazu $Ax - b$ uvrsti $x = x^{(k)}$ da li će se dobiti nula, šta će se dobiti, bolje je da se dobije što manje. Uvedimo oznaku

$$z^{(k)} = Ax^{(k)} - b,\ \ k \geq 0. \qquad (2)$$

Za $z^{(k)} \in R^n$ kaže se da je nepovezanost (lijeve i desne strane sistema). Bolje je što je norma nepovezanosti manja. Niz $\{x^{(k)}\}_{k=0}^{\infty}$, relacija (1), biće definisan ustvari tek kada se odrede i brojevi $\{\tau_{k+1}\}_{k=0}^{\infty}$. Uvedimo oznaku

$$\varphi(\tau_{k+1}) = \|z^{(k+1)}\| = \|Ax^{(k)} - b\| \geq 0.$$

Želimo da $\varphi(\tau_{k+1})$ ima što je moguće manju vrijednost. To je kriterijum za izbor veličine τ_{k+1}. Prelazimo na dobijanje eksplicitnog izraza za τ_{k+1}. Za račun je pogodnije da se gleda najmanja moguća vrijednost kvadrata $\varphi^2(\tau_{k+1}) = \|z^{(k+1)}\|^2 = (z^{(k+1)}, z^{(k+1)})$:

$$x^{(k+1)} = x^{(k)} + \tau_{k+1}(Ax^{(k)} - b) = x^{(k)} + \tau_{k+1} z^{(k)}$$

$$/ \cdot A$$

$$Ax^{(k+1)} = Ax^{(k)} + \tau_{k+1} Az^{(k)}$$

$$Ax^{(k+1)} - b = Ax^{(k)} - b + \tau_{k+1} Az^{(k)}$$

$$z^{(k+1)} = z^{(k)} + \tau_{k+1} Az^{(k)} \qquad (*)$$

$$\varphi^2(\tau_{k+1}) = (z^{(k+1)}, z^{(k+1)}) = (z^{(k)} + \tau_{k+1} Az^{(k)}, z^{(k)} + \tau_{k+1} Az^{(k)}) =$$

$$(z^{(k)}, z^{(k)}) + \tau_{k+1}(z^{(k)}, Az^{(k)}) + \tau_{k+1}(Az^{(k)}, z^{(k)}) + \tau_{k+1}^2 (Az^{(k)}, Az^{(k)}) =$$

$$(z^{(k)}, z^{(k)}) + 2\tau_{k+1}(z^{(k)}, Az^{(k)}) + \tau_{k+1}^2 (Az^{(k)}, Az^{(k)}),$$

jer je $A = A^T$. Parabola $y = ax^2 + bx + c$ ima najmanju moguću vrijednost kada je $x = \dfrac{-b}{2a}$, $a > 0$. Prema tome

$$\tau_{k+1} = \frac{-(z^{(k)}, Az^{(k)})}{(Az^{(k)}, Az^{(k)})} = \frac{-(z^{(k)}, Az^{(k)})}{\|Az^{(k)}\|^2}, \quad k \geq 0. \qquad (3)$$

Riješen je zadatak o minimumu funkcionala. Uzastopne aproksimacije su sada definisane. Redosljed računanja je: $x^{(0)}$, $z^{(0)}$, τ_1, $x^{(1)}$, itd.

Teorema o dovoljnim uslovima za konvergenciju metode minimalne nepovezanosti. Neka je $A = A^T$ i $A > 0$. Tada sistem $Ax = b$ ima jedinstveno rješenje x. Neka je $x^{(0)} \in R^n$ bilo koji vektor. Tada niz $\{x^{(k)}\}_{k=0}^{\infty}$, relacije (1)–(3), konvergira ka x. Pored toga, važi sljedeća nejednakost:

$$\|A(x^{(k)} - x)\| \leq \rho_0^k \cdot \|A(x^{(0)} - x)\| \text{ za } k \geq 0. \qquad (4)$$

Svojstvene vrijednosti matrice A su $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ $(0 < \lambda_1)$. Stavljeno je $\rho_0 = \dfrac{\lambda_n - \lambda_1}{\lambda_1 + \lambda_n} \geq 0$.

Dokaz teoreme. Po konstrukciji je $\varphi(\tau_{k+1}) = \min\limits_{\tau \in R} \varphi(\tau)$. Zato je $\varphi(\tau_{k+1}) \leq \varphi(\tau')$ za bilo koje $\tau' \in R$. Stavimo da je $\tau' = \dfrac{-2}{\lambda_1 + \lambda_n}$. Imamo da je $\varphi(\tau_{k+1}) = \|z^{(k+1)}\| = \|(E + \tau_{k+1}A)z^{(k)}\|$, v. (*). Imamo da je $\varphi(\tau') = \|(E + \tau' A)z^{(k)}\|$. Dakle

$$\|(E + \tau_{k+1}A)z^{(k)}\| \leq \|(E + \tau' A)z^{(k)}\|. \qquad (**)$$

Odredićemo $\|E + \tau' A\|$. Matrica A ima svojstvene vrijednosti $0 < \lambda_1 \leq \ldots \leq \lambda_n$. $\tau' A$ ima svojstvene vrijednosti $\dfrac{-2\lambda_n}{\lambda_1 + \lambda_n} \leq \ldots \leq \dfrac{-2\lambda_1}{\lambda_1 + \lambda_n} < 0$. Zapaziti da je i matrica $\tau' A$ simetrična.

Zapaziti da je i matrica $E+\tau' A$ simetrična. Znamo da je tada $\|E+\tau' A\| = \max\limits_{1\le i\le n} |\lambda_i(E+\tau' A)|$. Iz linearne algebre znamo sljedeće: svojstvene vrijednosti matrice $A+cE$ su veće od svojstvenih vrijednosti matrice A za c. Matrica $E+\tau' A$ ima svojstvene vrijednosti po veličini od $1-\dfrac{2\lambda_n}{\lambda_1+\lambda_n}$ do $1-\dfrac{2\lambda_1}{\lambda_1+\lambda_n}$ tj. od $-\dfrac{\lambda_n-\lambda_1}{\lambda_1+\lambda_n}$ do $\dfrac{\lambda_n-\lambda_1}{\lambda_1+\lambda_n}$, od $-\rho_0$ do ρ_0. Dakle, $\|E+\tau' A\| = \rho_0 < 1$.

Nastavljamo od (**):

$$\|(E+\tau_{k+1}A)z^{(k)}\| \le \|(E+\tau' A)z^{(k)}\|$$

$$\|z^{(k+1)}\| \le \|(E+\tau' A)z^{(k)}\|$$

$$\|z^{(k+1)}\| \le \|E+\tau' A\| \cdot \|z^{(k)}\|$$

$$\|z^{(k+1)}\| \le \rho_0 \cdot \|z^{(k)}\|$$

uzastopnom primjenom, $\qquad \|z^{(k)}\| \le \rho_0^k \cdot \|z^{(0)}\|$

$$\|Ax^{(k)} - b\| \le \rho_0^k \cdot \|Ax^{(0)} - b\|$$

$$\|Ax^{(k)} - Ax\| \le \rho_0^k \cdot \|Ax^{(0)} - Ax\|$$

$$\|A(x^{(k)} - x)\| \le \rho_0^k \cdot \|A(x^{(0)} - x)\|$$

Dokazali smo (4). Iz (4) slijedi da $\|x^{(k)} - x\| \to 0$ kad $k \to \infty$. Teorema je dokazana.

Znamo da je $\lambda_1 \cdot \|x^{(k)} - x\| \le \|A(x^{(k)} - x)\| \le \lambda_n \cdot \|x^{(k)} - x\|$.

Formula (4) služi za ocjenu greške aproksimacije $x^{(k)}$.

3.7. METODA SKALARNOG PROIZVODA

Riješiti tzv. potpuni problem svojstvenih vrijednosti za matricu $A \in R^{n\times n}$ znači odrediti sve njene svojstvene vrijednosti i odrediti sve odgovarajuće svojstvene vektore. Riješiti djelimični problem znači odrediti neke svojstvene vrijednosti ili odrediti jednu svojstvenu vrijednost (dominantnu). Za male vrijednosti n, svojstvene vrijednosti λ mogu da budu određene iz uslova $\det(A-\lambda E) = 0$. Metoda skalarnog proizvoda (metoda stepena) predstavlja jednu numeričku metodu za rješavanje djelimičnog problema svojstvenih vrijednosti. Razmotrimo jednostavni slučaj kada je matrica A simetrična, mada se ta metoda može primijeniti i na nesimetrične matrice. Neka je $A \in R^{n\times n}$ simetrična matrica:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \qquad (a_{ji} = a_{ij})$$

Označimo njene svojstvene vrijednosti kao $\lambda_i \in R$, $i = \overline{1,n}$, svaka svojstvena vrijednost broji se sa svojom višestrukošću. Neka je numeracija izvršena tako da bude

$$|\lambda_1| \ge |\lambda_2| \ge \ldots \ge |\lambda_n|.$$

Označimo odgovarajuće svojstvene vektore kao e_i, $i = \overline{1,n}$; važi $Ae_i = \lambda_i e_i$. Iz linearne algebre znamo da je $e_i \perp e_j$ tj. da je $(e_i, e_j) = 0$ za $i \neq j$; (x, y) označava skalarni proizvod. Izaberimo svojstvene vektore tako da bude $(e_i, e_i) = 1$ tj. $\|e_i\| = 1$ za $i = \overline{1,n}$. Sistem vektora $\{e_i\}_{i=1}^n$ čini ortonormiranu bazu prostora R^n. Biće konstruisan niz brojeva $\{\mu_k\}_{k=0}^\infty$, $\mu_k \in R$, takav da $\mu_k \to \lambda_1$ kad $k \to \infty$. Izaberimo na proizvoljan način vektor $x^{(0)} \in R^n$. Vektor $x^{(0)}$ razložimo po bazi $(e_1, \ldots, e_n)$; imamo da je $x^{(0)} = \sum_{i=1}^n c_i e_i$, gdje $c_i \in R$; znamo da je $c_i = (x^{(0)}, e_i)$. Stavimo da je

$$x^{(1)} = Ax^{(0)}, \quad x^{(2)} = Ax^{(1)}, \quad \ldots, \quad x^{(k)} = Ax^{(k-1)}, \quad \ldots$$

Imamo da je

$$x^{(1)} = Ax^{(0)} = A\sum_{i=1}^n c_i e_i = \sum_{i=1}^n c_i Ae_i = \sum_{i=1}^n c_i \lambda_i e_i$$

Slično, imamo da je

$$x^{(2)} = \sum_{i=1}^n c_i \lambda_i^2 e_i, \quad \ldots, \quad x^{(k)} = \sum_{i=1}^n c_i \lambda_i^k e_i, \quad \ldots$$

Izračunajmo skalarne proizvode $(x^{(k)}, x^{(k)})$ i $(x^{(k+1)}, x^{(k)})$:

$$(x^{(k)}, x^{(k)}) = \left(\sum_{i=1}^n c_i \lambda_i^k e_i, \sum_{j=1}^n c_j \lambda_j^k e_j\right) = \sum_{i,j=1}^n c_i c_j \lambda_i^k \lambda_j^k (e_i, e_j) =$$

$$\sum_{i=1}^n c_i c_i \lambda_i^k \lambda_i^k (e_i, e_i) = \sum_{i=1}^n c_i^2 \lambda_i^{2k}$$

$$(x^{(k+1)}, x^{(k)}) = \left(\sum_{i=1}^n c_i \lambda_i^{k+1} e_i, \sum_{j=1}^n c_j \lambda_j^k e_j\right) = \sum_{i,j=1}^n c_i c_j \lambda_i^{k+1} \lambda_j^k (e_i, e_j) =$$

$$\sum_{i=1}^n c_i c_i \lambda_i^{k+1} \lambda_i^k (e_i, e_i) = \sum_{i=1}^n c_i^2 \lambda_i^{2k+1}$$

Stavimo da je

$$\mu_k = \frac{(x^{(k+1)}, x^{(k)})}{(x^{(k)}, x^{(k)})}, \quad k \geq 0.$$

Sada je aproksimacioni niz $\{\mu_k\}_{k=0}^\infty$ definisan i proces računanja ili algoritam je definisan.

Teorema. Ako je $|\lambda_1| > |\lambda_2|$ i ako je $c_1 = (x^{(0)}, e_1) \neq 0$ onda $\mu_k \to \lambda_1$ kad $k \to \infty$ i

$$|\mu_k - \lambda_1| = O(q^k) \text{ kad } k \to \infty, \qquad (1) \qquad \text{gdje je } q = \left|\frac{\lambda_2}{\lambda_1}\right|^2 < 1.$$

Dokaz teoreme:

$$\mu_k = \frac{(x^{(k+1)}, x^{(k)})}{(x^{(k)}, x^{(k)})} = \frac{c_1^2 \lambda_1^{2k+1} + c_2^2 \lambda_2^{2k+1} + c_3^2 \lambda_3^{2k+1} + \ldots + c_n^2 \lambda_n^{2k+1}}{c_1^2 \lambda_1^{2k} + c_2^2 \lambda_2^{2k} + c_3^2 \lambda_3^{2k} + \ldots + c_n^2 \lambda_n^{2k}} \qquad (2)$$

Vidimo da $\mu_k \to \lambda_1$, jer je $|\lambda_2| < |\lambda_1|, \ldots, |\lambda_n| < |\lambda_1|$; kao $\lim\limits_{n\to\infty} \dfrac{5\cdot 3^{n+1} + 7\cdot 2^{n+1}}{5\cdot 3^n + 7\cdot 2^n} = 3$

$$\mu_k - \lambda_1 = \frac{c_2^2(\lambda_2-\lambda_1)\lambda_2^{2k} + c_3^2(\lambda_3-\lambda_1)\lambda_3^{2k} + \ldots + c_n^2(\lambda_n-\lambda_1)\lambda_n^{2k}}{c_1^2\lambda_1^{2k} + c_2^2\lambda_2^{2k} + c_3^2\lambda_3^{2k} + \ldots + c_n^2\lambda_n^{2k}}$$

$$|\mu_k - \lambda_1| \le \frac{1}{c_1^2\lambda_1^{2k}}\left(c_2^2|\lambda_2-\lambda_1|\lambda_2^{2k} + c_3^2|\lambda_3-\lambda_1|\lambda_3^{2k} + \ldots + c_n^2|\lambda_n-\lambda_1|\lambda_n^{2k}\right) \le$$

$$\frac{1}{c_1^2\lambda_1^{2k}}\left(c_2^2\cdot 2|\lambda_1|\cdot\lambda_2^{2k} + c_3^2\cdot 2|\lambda_1|\cdot\lambda_2^{2k} + \ldots + c_n^2\cdot 2|\lambda_1|\cdot\lambda_2^{2k}\right) =$$

$$\left(\frac{\lambda_2}{\lambda_1}\right)^{2k}\cdot\frac{1}{c_1^2}\cdot 2|\lambda_1|\left(c_2^2 + c_3^2 + \ldots + c_n^2\right) = O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right)$$

Teorema je dokazana.

Formula (1) služi za ocjenu greške k–te aproksimacije i govori da je brzina konvergencije (tempo konvergencije) metode skalarnog proizvoda prvog reda ili prvog stepena.

a) Uslov teoreme $|\lambda_1| > |\lambda_2|$ govori da A ima svojstvenu vrijednost koja je strogo dominantna po apsolutnoj vrijednosti. Šta će se desiti ako uslov nije ispunjen (nepovoljna okolnost). Ima više dominantnih po apsolutnoj vrijednosti. Sve one su istog znaka ili među njima ima i pozitivnih i negativnih.

Ako je $\lambda_1 = \lambda_2$, $|\lambda_2| > |\lambda_3|$ tada također $\mu_k \to \lambda_1$ kad $k \to \infty$ pod uslovom da je $c_1^2 + c_2^2 \neq 0$. Ako je $\lambda_1 = \ldots = \lambda_p$, $|\lambda_p| > |\lambda_{p+1}|$ tada također $\mu_k \to \lambda_1$ kad $k \to \infty$ pod uslovom da je $c_1^2 + \ldots + c_p^2 \neq 0$. Izračunati $\lim\limits_{k\to\infty} \mu_k$ (formula (2)) i uvjeriti se.

Ako je $\lambda_1 = \ldots = \lambda_p = -\lambda_{p+1} = \ldots = -\lambda_q$, $|\lambda_q| > |\lambda_{q+1}|$ onda nije istina da niz $\{\mu_k\}_{k=0}^{\infty}$ konvergira ka λ_1; niz $\{\mu_k\}_{k=0}^{\infty}$ ne koristi; taj niz "lažno" konvergira. Uvjeriti se posmatranjem dva posebna slučaja kako slijedi. Ako je $\lambda_1 = -\lambda_2$, $|\lambda_2| > |\lambda_3|$ onda je $\lim\limits_{k\to\infty} \mu_k = \dfrac{c_1^2 - c_2^2}{c_1^2 + c_2^2}\cdot\lambda_1$ pod uslovom da je $c_1^2 + c_2^2 \neq 0$. Pogledati i drugi poseban slučaj $\lambda_1 = \lambda_2 = -\lambda_3$, $|\lambda_3| > |\lambda_4|$.

Ako je već nastupila nepovoljna okolnost koja stvara teškoće, kako da teškoće prevaziđemo? Prvi savjet: primijeniti metodu skalarnog proizvoda na matricu $A + cE$ čije su svojstvene vrijednosti $\lambda_i + c$. Drugi savjet: neka niz $\nu_k = \dfrac{(x^{(k+2)}, x^{(k)})}{(x^{(k)}, x^{(k)})}$ posluži kao aproksimacioni niz. Ako je recimo $\lambda_1 = -\lambda_2$, $|\lambda_2| > |\lambda_3|$ onda je $\lim\limits_{k\to\infty} \nu_k = \lambda_1^2$ pod uslovom da je $c_1^2 + c_2^2 \neq 0$.

b) Početna aproksimacija $x^{(0)}$ bira se na slučajan način, ne koristeći bilo kakve informacije o matrici A odnosno o njenim svojstvenim vrijednostima i svojstvenim vektorima. Mi uostalom po pravilu i ne raspolažemo sa takvim informacijama. Zato se može desiti da bude $c_1 = 0$ (nepovoljna okolnost). Kada izaberemo $x^{(0)}$, mi tada ne znamo da li je $c_1 \neq 0$ ili je pak suprotno $c_1 = 0$. Zato se može pokušati sa dva–tri početna vektora $x^{(0)}$.

Slijede razne dopune

a) Posmatrajmo veličinu

$$\|x^{(k)}\| = \sqrt{(x^{(k)}, x^{(k)})} = \sqrt{\sum_{i=1}^{n} c_i^2\lambda_i^{2k}} = \sqrt{c_1^2\lambda_1^{2k} + c_2^2\lambda_2^{2k} + \ldots + c_n^2\lambda_n^{2k}}.$$

Ako je $|\lambda_1| > 1$ onda je $\lim_{k\to\infty} \|x^{(k)}\| = +\infty$; $c_1 \neq 0$. Kod računara će lako doći do prekoračenja. Vršiti prilagođavanje veličine $\|x^{(k)}\|$ s vremena na vrijeme ili vršiti njeno prilagođavanje na svakom koraku. Ulogu vektora $x^{(k)}$ neka preuzme vektor koji je kolinearan sa $x^{(k)}$ a čija je dužina = 1.

b) Kada k neograničeno raste onda vektor

$$x^{(k)} = \sum_{i=1}^{n} c_i \lambda_i^k e_i = c_1\lambda_1^k e_1 + c_2\lambda_2^k e_2 + \ldots + c_n\lambda_n^k e_n$$

postaje skoro kolinearan sa vektorom e_1; pod uslovima $|\lambda_1| > |\lambda_2|$ i $c_1 \neq 0$. Imamo način da približno odredimo svojstveni vektor e_1 koji odgovara dominantnoj po apsolutnoj vrijednosti svojstvenoj vrijednosti λ_1.

c) Ako su spektralni podaci simetrične matrice $A = Ax$

$$\lambda_1, \lambda_2, \ldots, \lambda_n \text{ i } e_1, e_2, \ldots, e_n$$

onda su spektralni podaci matrice $B = Bx = Ax - \lambda_1(x, e_1)e_1$

$$0, \lambda_2, \ldots, \lambda_n \text{ i } e_1, e_2, \ldots, e_n.$$

Kada smo odredili makar i samo približno λ_1 i e_1 ($\|e_1\| = 1$), mi izračunamo matricu B. Primjenom metode skalarnog proizvoda na matricu B sada može da bude približno određeno λ_2.

S. l. j. $\begin{cases} a_{11}x_1 + a_{12}x_2 = b_1 \\ a_{21}x_1 + a_{22}x_2 = b_2 \end{cases}$ ili $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ (matrični zapis)

Princip kontrakcije ili teorema o nepokretnoj tački ili Bamahova teorema o fiksnoj tački. Neka je X kompletan metrički prostor. Razmotrimo preslikavanje φ: $X \to X$. Ako je φ kontrakcija onda φ ima jedinstvenu (jednu jedinu, tačno jednu) nepokretnu tačku.

Definicija. Razmotrimo metrički prostor (X, d), d – distanca, rastojanje i razmotrimo preslikavanje φ: $X \to X$. Za φ se kaže da je kontrakcija ako postoji broj $q < 1$ takav da važi $d(\varphi(x), \varphi(y)) \leq qd(x, y)$, $\forall\, x, y \in X$. Tada se za broj q kaže da je koeficijent kontrakcije.

Definicija. Za $\xi \in X$ kaže se da je nepokretna tačka preslikavanja φ ako važi $\varphi(\xi) = \xi$.

Nastavak teoreme. Izaberimo proizvoljno $x_0 \in X$. Stavimo $x_{n+1} = \varphi(x_n)$ za $n \geq 0$. Tada važi $\lim_{n\to\infty} d(x_n, \xi) = 0$, tj. $\lim_{n\to\infty} x_n = \xi$.

Nastavak. Važe nejednakosti $d(x_n, \xi) \leq \frac{q^n}{1-q} d(x_1, x_0)$ i $d(x_n, \xi) \leq \frac{q}{1-q} d(x_n, x_{n-1})$. Još $d(x_n, \xi) \leq qd(x_{n-1}, \xi)$ ili $|x_n - \xi| \leq q|x_{n-1} - \xi|$ ($X = R$), $\|x_n - \xi\| \leq q\|x_{n-1} - \xi\|$ ($X = R^n$).

Npr. $d(x, y) = |x - y|$ u slučaju prostora R, npr. $d(x, y) = \|x - y\|$ u slučaju prostora R^n.

Dopuna o $\mathbf{x} = (x_1, \ldots, x_n)$ i $\mathbf{y} = (y_1, \ldots, y_n)$: iz skalarnog proizvoda $\langle \mathbf{x}, \mathbf{y} \rangle = x_1y_1 + \ldots + x_ny_n$ proizilazi norma vektora $\|\mathbf{x}\| = \sqrt{x_1^2 + \ldots + x_n^2}$, saglasno relaciji $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.

4. RJEŠAVANJE SISTEMA NELINEARNIH JEDNAČINA

4.1. METODA POLOVLJENJA

Neka je postavljen sljedeći zadatak: naći približno rješenje jednačine $f(x) = 0$. Ovdje se pretpostavlja da je funkcija f neprekidna i da f: $R \to R$ ili f: $[a,b] \to R$. Za rješavanje ovog zadatka postoji više metoda, a metoda polovljenja je najjednostavnija među njima. U vezi velike brzine savremenih kompjutera, razmatrani zadatak se skoro uvijek rješava ovom metodom, ovim algoritmom.

Pripremni korak kod rješavanja nelinearne jednačine $f(x) = 0$ po bilo kojoj numeričkoj metodi jeste lokalizacija nula funkcije f tj. određivanje početnog odsječka $[a_0, b_0]$, pri čemu važi $f(a_0) \cdot f(b_0) < 0$, u krajnjim tačkama odsječka vrijednosti funkcije su različitog znaka. Budući da je funkcija f po uslovu neprekidna to slijedi, poznato je iz matematičke analize, da f ima bar jednu nulu unutar odsječka $[a_0, b_0]$. Metoda polovljenja omogućuje nam da jednostavnim algoritmom i brzo dobijemo približnu vrijednost jedne nule sa neogramičeno dobrom preciznošću (sa po želji malom greškom).

Pogledajmo sliku 1. Prikazano je da je u tački a_0 vrijednost funkcije negativna a da je u b_0 samim tim pozitivna. Pogledajmo kakvog je znaka vrijednost funkcije u srednjoj tački odsječka. Uvedimo oznaku $c = (a_0 + b_0)/2$. Izračunajmo vrijednost $f(c)$. Ova vrijednost je ili jednaka nuli ili je različita od nule. Ako je $f(c) = 0$ onda smo našli tačno rješenje razmatrane jednačine, pa se izvršavanje algoritma očito prekida. Mnogo je realnije očekivati da će biti $f(c) \neq 0$. Ako je $f(c) \neq 0$ onda ćemo uzeti da je $[a_1, b_1] = [a_0, c]$ u slučaju da je $f(a_0) \cdot f(c) < 0$ ili ćemo uzeti $[a_1, b_1] = [c, b_0]$ u slučaju da je $f(c) \cdot f(b_0) < 0$. Dakle, naša dalja pažnja odnosiće se na odsječak $[a_1, b_1]$, budući da na lijevom i desnom kraju tog odsječka funkcija f ima vrijednosti različitog znaka, pa unutar tog odsječka sigurno postoji (bar jedno) rješenje jednačine koja nas interesuje. Odsječak $[a_1, b_1]$ očito je lijeva ili desna polovina početnog odsjeka $[a_0, b_0]$.

Pogledajmo kakvog je znaka vrijednost funkcije f u srednjoj tački odsječka $[a_1, b_1]$. U zavisnosti od rasporeda znakova brojeva $f(a_1)$, $f((a_1 + b_1)/2)$ i $f(b_1)$, definišemo novi odsječak $[a_2, b_2]$ koji predstavlja lijevu ili desnu polovinu od $[a_1, b_1]$. Bilo da predstavlja lijevu ili desnu polovinu, ima svojstvo da sadrži nulu funkcije f. Očito je da poslije drugog koraka mi imamo sljedeću informaciju: nula pripada odsječku $[a_2, b_2]$. Mi sada možemo da za približnu vrijednost rješenja postavljenog zadatka proglasimo bilo koju tačku iz ovog odsječka (ako ne želimo da dalje računamo). Takođe je očito koliko je u tom slučaju rastojanje između tačnog i približnog rješenja tj. kolika je greška našeg numeričkog odgovora. To rastojanje je manje od dužine samog odsječka $[a_2, b_2]$. Dužina ovog novog odsječka iznosi jednu četvrtinu dužine polaznog odsječka, tj. $b_2 - a_2 = (b_0 - a_0)/4$.

Slično se naravno vrši treći korak odnosno određuje novi odsječak $[a_3, b_3]$. Jasno je da je dužina svakog novog odsječka jednaka polovini dužine prethodnog odsječka. To je mjera brzine kojom se korak po korak približavamo ka tačnom rješenju jednačine. Na svakom koraku može da se desi da (slučajno) pogodimo tačno rješenje, označimo tačno rješenje sa ξ, da nađemo odgovor sa greškom nula, mada je ovo zaista malo vjerovatno.

Itd. Nastavljajući ovako, naša udaljenost od nule funkcije f očito teži ka nuli.

Dokle ćemo mi ovako da računamo ili dokle će kompjuter ovako da produži? Obično je unaprijed definisana preciznost tj. dozvoljena greška $\varepsilon > 0$ sa kojom treba riješiti postavljenu jednačinu. Drugim riječima, naš odgovor odnosno približno rješenje ξ' treba da ispunjava uslov $|\xi - \xi'| < \varepsilon$, izlazni kriterijum.

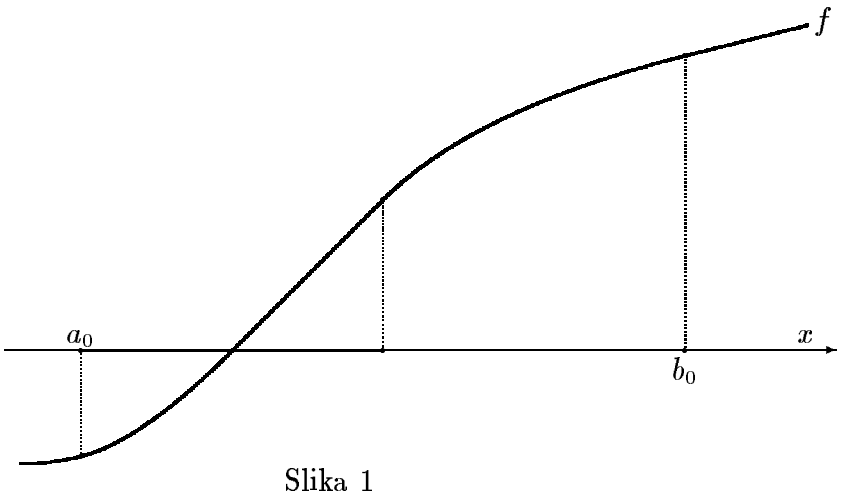
Izvršimo ocjenu greške poslije k izvedenih koraka. Dakle, došli smo do odsječka $[a_k, b_k]$. Najprirodnije je da u ovoj situaciji upravo srednju tačku tog odsječka smatramo približnim

rješenjem, tj. da stavimo $\xi \approx \xi' = (a_k + b_k)/2$. Sada tvrdimo da je $|\xi - \xi'| \leq (b_0 - a_0)/2^{k+1}$. Ovu formulu možemo zvati formulom za ocjenu greške metode polovljenja. I iz ove formule je očito da važi $\xi' \to \xi$ kad $k \to \infty$.

Ako je unaprijed specifikovana dozvoljena greška ε onda se pomoću formule za ocjenu greške može utvrditi da li je učinjenih k koraka dovoljno ili treba produžiti računanje. Ili prosto utvrditi pomoću računanja dužine $b_k - a_k$ odsječka $[a_k, b_k]$. Takođe se može, ako to nečemu koristi, unaprijed determinisati broj koraka nakon kojih se tačnost ε ostvari.

Vidimo da se od koraka do koraka greška pomnoži sa 1/2. Drukčije rečeno, greška opada tempom geometrijske progresije (čiji je količnik 1/2). Zato se kaže da metoda polovljenja ima prvi ili linearni red (brzinu) konvergencije. Sadašnja greška $\leq$ konstanta puta prethodna greška.

Vidimo da nas izloženi algoritam vodi ka jednom rješenju postavljene jednačine. A postavljena jednačina može da ima više od jednog rješenja.



Slika 1

Opisivanje metode je uglavnom završeno. Dosad rečeno može da se sakupi i formuliše u obliku teoreme, ovo se prepušta čitaocu da uradi. Slijede dopune–komentari.

Zanimljivo je i jednostavno eksplicitno izraziti tempo kojim greška opada, izraziti ga u terminima decimalnih mjesta koja su "osvojena". Mi kažemo da su (recimo) dvije decimale osvojene ako je greška manja od 10^{-2}. Izvođenje koje slijedi oslanja se jedino na okolnost da se jednim korakom greška svede na polovinu pređašnje. Za dva koraka greška će biti pomnožena sa faktorom 1/4. A za otprilike 3,3 koraka biće pomnožena sa faktorom 0,1, budući da je $\log_{10} 2 = 0{,}3$. Dakle, jedna decimala se osvoji za 3.3 koraka. Uzmimo da je dužina početnog odsječka $[a_0, b_0]$ jednaka 1. Neka želimo da nađemo približnu vrijednost sa pet decimala. Koliko koraka će trebati? 17 koraka.

Na početku je rečeno da upotreba neke složenije metode za numeričko rješavanje nelinearne jednačine sa jednom nepoznatom izgleda danas vrlo slabo opravdana. Imamo u vidu brzinu kojom današnji kompjuteri mogu da računaju. Vidimo da se za realizaciju jednog koraka po metodi polovljenja segmenta najviše vremena potroši za računanje (jedne samo) vrijednosti funkcije f, za računanje vrijednosti funkcije u jednoj tački.

Kod upotrebe metode polovljenja treba obratiti pažnju na sljedeću okolnost. Vrijednosti koje računar saopštava su približni brojevi. Ako je saopštena vrijednost blizu nule onda nismo baš sigurni kada tvrdimo da je pozitivna ili negativna. Kako da se ovo prevaziđe? Uvedimo potrebne oznake. Neka treba da bude izračunata vrijednost funkcije f u nekoj tački c. Dakle, $f(c)$ naravno označava (tačnu) vrijednost te funkcije u toj tački. Označimo sa $f^*(c)$ odgovarajuću vrijednost koju nam računar saopštava. Veličinom $f^*(c)$ raspolažemo, veličinu $f(c)$

možemo samo da "zamišljamo". Interesuje nas – kojeg je znaka veličina $f(c)$. Kada možemo da po znaku broja $f^*(c)$ sudimo o znaku broja $f(c)$? Da na ovo odgovorimo, treba da uvedemo u razmatranje još jedan pokazatelj – $\varepsilon > 0$ – greška sa kojom kompjuter saopštava realne brojeve (da li se ima u vidu prosječna ili najveća moguća?). Tek ako je $f^*(c) > \varepsilon$ ili $f^*(c) > 2\varepsilon$ možemo s pravom da presudimo da je i broj $f(c)$ pozitivan.

U okviru ovoga, i slučaj "$f(c) = 0$", za koji je prije rečeno da može da nastupi iako malo vjerovatno – treba kritički pogledati. Ovaj slučaj je najpovoljniji sa stanovišta matematičke analize, "uspjeli smo da dobijemo ne približno nego čak tačno rješenje". Sa stanovišta numeričkih metoda, budući da i greška računanja treba da se uzima u obzur, ovaj slučaj nije toliko poželjan. Ako nam je računar saopštio da je vrijednost funkcije u toj nekoj tački jednaka nuli onda mi ustvari ne znamo kog znaka je vrijednost. Ista priča važi naravno za $|f^*(c)| < \varepsilon$. Kako prilagoditi metodu polovljenja u ovakvoj situaciji? Tačku c treba za nekoliko pomjeriti tako da nam za to izmijenjeno c računar saopšti da je $|f^*(c)| > \varepsilon$ ili saopšti da je $|f^*(c)| > 2\varepsilon$. Ovo pomijeranje tačke c, ova izmjena algoritma se odražava na ocjenu greške, dolazi do izvjesnog usporavanja konvergencije.

Koliko je ε? Znamo da kod većine programskih jezika, kada se radi sa običnom preciznošću, relativna greška, kod samog upisivanja realnog broja u memoriju, iznosi prosječno 10^{-7}. U slučaju dvostruke tačnosti (engl. double precision), relativna greška iznosi nekih 10^{-16}.

4.2. METODA PROSTE ITERACIJE

Razmotrimo zadatak o numeričkom rješavanju sistema od n nelinearnih jednačina sa n nepoznatih. Ovakav sistem (ovakva jednačina) definiše se pomoću jednog preslikavanja f. Oblast definisanosti preslikavanja jeste prostor R^n ili neki njegov podskup. A vrijednosti preslikavanja su očito takođe iz R^n. Dakle, dat je sistem (data je jednačina) $f(x) = 0$, pri čemu je f: $R^n \to R^n$ ili f: $A \to R^n$, gdje je $A \subset R^n$. Očito 0 – nula u R^n. Označimo (tačno) rješenje ovog sistema sa x.

Većina metoda za približno (za numeričko) rješavanje sistema jednačina je iterativnog tipa. Računaju se redom tzv. uzastopne (sukcesivne) aproksimacije, koje bi trebalo da teže ka tačnom rješenju. Za iterativne metode koje se baziraju na Banahovoj teoremi o nepokretnoj tački kaže se da su – metode proste iteracije.

Pripremni korak u rješavanju jeste – ispitivanje da li sistem uopšte ima rješenja, koliko rješenja ima, lokalizacija pojedinih rješenja (određivanje oblasti koja sadrži rješenje). Tako nalazimo početnu (ili nultu ili grubu) aproksimaciju $x^{(0)}$.

Prethodni korak kod primjene metode proste iteracije jeste da se dati sistem $f(x) = 0$ transformiše u neki ekvivalentni sistem oblika $x = g(x)$. Postoji više načina da se ova transformacija izvrši, postoji beskonačno mnogo sistema oblika $x = g(x)$ koji su ekvivalentni sa datim–polaznim sistemom. Neki od tih oblika su, pokazaće se, pogodni za primjenu metode proste iteracije. To znači da će iteracije računate po takvom obliku da konvergiraju. A neki nisu pogodni. Važno je pitanje – kako uraditi transformaciju da se ispostavi da su uslovi Banahove teoreme ispunjeni.

Znamo da se Banahova teorema odnosi na jednačinu oblika upravo $x = g(x)$. Zato se i vrši transformacija datog sistema u takav oblik.

Dakle, već imamo početnu aproksimaciju $x^{(0)}$, gdje je $x^{(0)}$ jedan vektor dužine n, a slično naravno i $x^{(1)}$, itd. Mi računamo, kompjuter računa sljedeće: $x^{(1)} = g(x^{(0)})$, $x^{(2)} = g(x^{(1)})$, ... Da li niz $\{x^{(k)}\}$ konvergira, ako konvergira – da li konvergira ka rješenju sistema x? Pod određenim uslovima važi da $x^{(k)} \to x$ (da $x^{(k)} \to$ rješenju) kad $k \to \infty$. Dovoljne uslove za ovo daje Banahova teorema.

Kada se kaže da se jednačina (sistem) $x = g(x)$ rješava metodom proste iteracije onda je samim tim implicitno rečeno i po kojoj formuli se računaju uzastopne aproksimacije. Po formuli $x^{(k)} = g(x^{(k-1)})$ za $k \in N = \{1, 2, \ldots\}$.

Navedimo kako glasi Banahova teorema o nepokretnoj tački.

T. Neka je X (realan) kompletan metrički prostor i neka je njegova metrika označena sa ρ. Neka preslikavanje g: $X \to X$ zadovoljava uslov kontrakcije: $\rho(g(x), g(y)) \leq q\rho(x, y)$ za ma koje x i y, gdje je konstanta $q < 1$; za samo g se kaže da je kontrakcija. Neka je $x^{(k)} = g(x^{(k-1)})$ za $k \in N$. Tada: (1) jednačina $x = g(x)$ ima jedinstveno rješenje u prostoru X (označimo to rješenje sa ξ), (2) $x^{(k)} \to \xi$, bez obzira kako je izabrano $x^{(0)} \in X$ i (3) važi sljedeća formula (koja nam služi za ocjenjivanje greške): $\rho(x^{(k)}, \xi) \leq \dfrac{q^k}{1-q}\rho(x^{(1)}, x^{(0)})$.

Dokaz ove teoreme je poznat iz matematičke analize, iz funkcionalne analize, pa ga ovdje nećemo ponavljati. Za ξ se kaže da je nepokretna ili fiksna tačka funkcije g. Znamo također da ulogu prostora X može da preuzme neki (bilo koji) njegov zatvoreni podskup.

Napominjemo da je bolje što je koeficijent kontrakcije q manji, što je bliži nuli. Jer tada iterativni niz brže konvergira, greška se od koraka do koraka brže smanjuje. A ako q pređe 1 onda g prestaje da bude kontrakcija.

Uzmimo da je A pravi podskup od R^n. Veći su izgledi da će preslikavanje g biti kontrakcija ako A ima ulogu prostora X nego ako čitav R^n ima tu ulogu. Tj. ukoliko je A manji utoliko je lakše da ualov kontrakcije bude ispunjen. Međutim, ne treba previdjeti sljedeću okolnost. U teoremi se pojavljuje i uslov: vrijednosti preslikavanja g pripadaju skupu X. Ako je $X = R^n$ onda je ovaj uslov očito ispunjen sam po sebi. A ako je $X = A$ onda ne mora da bude. U numeričkoj praksi – ustanoviti da li je ovaj uslov zaista ispunjen – nije trivijalno. Upravljanje ovim uslovom zahtijeva jednako truda koliko i upravljanje uslovom kontrakcije. I jedan i drugi uslov zavise od načina transformacije polaznog sistema $f(x) = 0$ u oblik pripremljen za vršenje iteracija $x = g(x)$.

Znamo da je R^n jedan kompletan metrički prostor, treba reći koja metrika (norma) se ima u vidu. U nastavku, mi ćemo uglavnom nastojati da konkretizujemo razne elemente Banahove teoreme u slučaju $X = R^n$.

Slučajevi $n = 2$ i $n > 2$ se vrlo malo razlikuju. Samo radi lakšeg pisanja, uzmimo odsad da je $n = 2$. A čitalac neka sam zaključi da se $n > 2$ tretira analogno.

Ako je već $n = 2$ onda $x = g(x)$ zapisuje jedan sistem od dvije jednačine sa dvije nepoznate. Vektorski zapis. Recimo, ovdje $x \in R^2$. Pogodnije nam je odsad da te dvije jednačine zapišemo pojedinačno. Da upotrebljavamo skalarni zapis. Umjesto $x = g(x)$ odsad ćemo pisati

$$\begin{cases} x = f(x, y) \\ y = g(x, y) \end{cases}$$

Nove oznake se ukrštaju sa dosadašnjim, obratiti pažnju da se izbjegne zbrka. Sada je očito $x \in R$ i $y \in R$. Isto tako, sada je f: $R^2 \to R$ i g: $R^2 \to R$. Počinju skalarne oznake.

Pretpostavlja se da je A – zatvoreni konveksni podskup od R^2. Zatvoreni – da bi metrički prostor bio kompletan. Konveksni – u vezi primjene Tejlorove formule, v. niže, ako krajnje tačke jedne duži pripadaju skupu A onda i bilo koja tačka te duži pripada tom skupu. Također se pretpostavlja da $f \in C^1(A)$ kao i $g \in C^1(A)$.

Želimo da analiziramo uslove pod kojima važi da je preslikavanje (f, g) kontrakcija. Ispostaviće se da to zavisi od ponašanja prvih parcijalnih izvoda funkcija f i g. Bolje je što su ti izvodi (po modulu) manji, što su bliži nuli. Razmatra se supremum, uzet po skupu A.

Uvedimo sljedeće oznake: neka su m_{ij} $(i,j=1,2)$ ma koji brojevi koji zadovoljavaju sljedeće nejednakosti:

$$\sup_{(x,y)\in A}\left|\frac{\partial f(x,y)}{\partial x}\right|\le m_{11},\quad \sup_{A}\left|\frac{\partial f(x,y)}{\partial y}\right|\le m_{12},$$

$$\sup_{A}\left|\frac{\partial g(x,y)}{\partial x}\right|\le m_{21},\quad \sup_{A}\left|\frac{\partial g(x,y)}{\partial y}\right|\le m_{22},\quad M=\begin{bmatrix} m_{11} & m_{12}\\ m_{21} & m_{22}\end{bmatrix}$$

Ovaj način definisanja brojeva m_{ij} odgovara numeričkoj praksi. Bolje nego da smo u prethodne četiri formule u kojima se definišu ova četiri broja umjesto znaka manje ili jednako pisali znak jednako. Zato što u praksi po pravilu mi nismo u stanju da odredimo tačno pomenute supremume. Nego smo u stanju da odredimo samo izvjesne njihove ocjene sa gornje strane. A bolje je ukoliko su te ocjene preciznije, ukoliko su bliže supremumima samim. Pa upotrebljavamo očito te ocjene sa kojima raspolažemo, a ne upotrebljavamo u računanju supremume koje jedino "zamišljamo".

Pomenuta četiri pokazatelja možemo da smatramo elementima jedne matrice M, ima oblik 2×2, kao što je već napisano u prethodnoj formuli.

Određenim kombinovanjem pokazatelja m_{ij} dobićemo sada nove pokazatelje q_1 i q_∞. Pokazatelje m_{ij} smo maločas zapisali kao matricu, jer će se u izvođenju pojaviti norma te matrice M. I to $\|\ \|_1$ i $\|\ \|_\infty$, ove norme su već ranije upotrebljavane, kada se govorilo o iterativnom rješavanju linearnih sistema.

Podsjećamo da je u prostoru R^2, $\|(x,y)\|_1=|x|+|y|$ i $\|(x,y)\|_\infty=\max\{|x|,|y|\}$.

Takođe podsjećamo, ranije je rađeno, na formule za indukovanu normu matrice tj. linearnog operatora, $\|M\|_1=\max\{m_{11}+m_{21},m_{12}+m_{22}\}$ kao i $\|M\|_\infty=\max\{m_{11}+m_{12},m_{21}+m_{22}\}$. U ovim formulama bi umjesto m_{ij} trebalo da piše $|m_{ij}|$, ali je ustvari svejedno jer znamo da su sva četiri broja nenegativna, $m_{ij}\ge 0$.

Neka je q_1 ma koji broj koji zadovoljava $m_{11}+m_{21}\le q_1$ i $m_{12}+m_{22}\le q_1$. Slijedi da važi $\|M\|_1\le q_1$. Slično, neka je q_∞ ma koji broj koji zadovoljava $m_{11}+m_{12}\le q_\infty$ i $m_{21}+m_{22}\le q_\infty$. Tada možemo pisati da je $\|M\|_\infty\le q_\infty$.

Mogli smo broj q_1 prostije da definišemo kao maksimalni od dva zbira $m_{11}+m_{21}$ i $m_{12}+m_{22}$. Tada bi norma matrice M sa indeksom jedan bila jednaka q_1. Slično u slučaju norme sa indeksom beskonačno.

Mi posmatramo Banahovu teoremu u slučaju da je $X=R^2$ ili $X=A\subset R^2$. A metrika ρ koja se u toj teoremi spominje jeste ona koja proističe iz norme $\|\ \|_1$ ili $\|\ \|_\infty$. Znamo da se uzima $\rho(\alpha,\beta)=\|\alpha-\beta\|$. Mi ćemo sada da dokažemo dvije teoreme, jedna će se odnositi na normu sa indeksom jedan a jedna na onu sa indeksom beskonačno. Izvođenje jedne i druge teoreme će teći paralelno. Slijedi mali računski dio tog izvođenja.

Posmatrajmo razliku vrijednosti funkcije f u dvije tačke (x_1,y_1) i (x_2,y_2). Vidi sliku 2. Izraz za tu razliku dobićemo uz pomoć Tejlorove formule. Bolje reći, uz pomoć Lagranžove teoreme (formule o konačnim priraštajima). U izrazu se pojavljuje jedna tačka (x_3,y_3) sa odsječka čiji su krajevi prve dvije pomenute tačke. Uočiti da se, ako se funkcija f posmatra samo na toj duži, na tu funkciju može gledati kao na funkciju od jedne promjenljive. Parametrizovati duž. Tako da možemo reći da se primjenjuje Lagranžova teorema na funkciju od jedne promjenljive. Izraziti izvod po parametru preko parcijalnih izvoda od f.

Po Tejlorovoj formuli imamo:

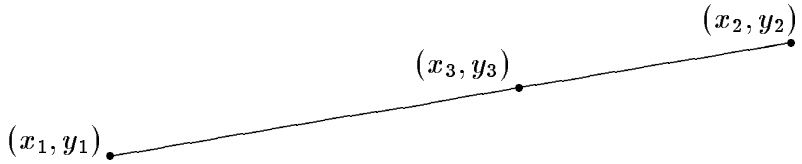
$$f(x_2,y_2)-f(x_1,y_1)=\frac{\partial f(x_3,y_3)}{\partial x}(x_2-x_1)+\frac{\partial f(x_3,y_3)}{\partial y}(y_2-y_1),$$

gdje je (x_3, y_3) neka tačka sa duži od (x_1, y_1) do (x_2, y_2), detaljnije se piše $x_3 = x_1 + \theta(x_2 - x_1)$ i $y_3 = y_1 + \theta(y_2 - y_1)$, za neko θ, $0 < \theta < 1$.

Na isti način,

$$g(x_2, y_2) - g(x_1, y_1) = \frac{\partial g(x_4, y_4)}{\partial x}(x_2 - x_1) + \frac{\partial g(x_4, y_4)}{\partial y}(y_2 - y_1),$$

gdje je (x_4, y_4) neka (druga) tačka sa iste one duži.



Slika 2

Odavde i iz ranijih jednakosti imamo:

$$|f_2 - f_1| \le m_{11}|x_2 - x_1| + m_{12}|y_2 - y_1|, \quad |g_2 - g_1| \le m_{21}|x_2 - x_1| + m_{22}|y_2 - y_1|, \qquad (*)$$

gdje su uvedene skraćenice $f_i = f(x_i, y_i)$, $g_i = g(x_i, y_i)$, $i = 1, 2$.

Uvedimo nove skraćenice $\vec{a} = (x_1, y_1)$, $\vec{b} = (x_2, y_2)$, $\vec{c} = (f_1, g_1)$, $\vec{d} = (f_2, g_2)$. Tako da preslikavanje (f, g) prevodi tačku $\vec{a}$ u tačku $\vec{c}$, a prevodi $\vec{b}$ u $\vec{d}$. Ako želimo da to preslikavanje bude kontrakcija onda norma od $\vec{b} - \vec{a}$ treba da prevazilazi normu od $\vec{d} - \vec{c}$.

Saberimo dvije nejednakosti $(*)$:

$$|f_2 - f_1| + |g_2 - g_1| \le (m_{11} + m_{21})|x_2 - x_1| + (m_{12} + m_{22})|y_2 - y_1| \le q_1(|x_2 - x_1| + |y_2 - y_1|),$$

odnosno $\|\vec{d} - \vec{c}\|_1 \le q_1 \|\vec{b} - \vec{a}\|_1$.

Slično iz $(*)$ imamo, kada se umjesto $|x_2 - x_1|$ odnosno $|y_2 - y_1|$ napiše veći od ta dva broja,

$$|f_2 - f_1| \le q_\infty \cdot \max\{|x_2 - x_1|, |y_2 - y_1|\} \text{ i } |g_2 - g_1| \le q_\infty \cdot \max\{|x_2 - x_1|, |y_2 - y_1|\}.$$

Slijedi da je $\|\vec{d} - \vec{c}\|_\infty \le q_\infty \cdot \|\vec{b} - \vec{a}\|_\infty$. Jer $\alpha \le \gamma$ i $\beta \le \gamma \Rightarrow \max\{\alpha, \beta\} \le \gamma$.

Time smo dokazali sljedeće dvije teoreme.

Teorema. Ako (i) (f, g): $A \to A$ i (ii) $q_1 < 1$ onda važi $x_k \to \xi$, $y_k \to \eta$ (kad $k \to \infty$), za bilo koje $(x_0, y_0) \in A$. Tačka (ξ, η) je jedinstvena. Ponovimo uslove: prvo: A zatvoren konveksan i drugo: $f \in C^1(A)$ i $g \in C^1(A)$. Ovdje su upotrebljene oznake: $x_{k+1} = f(x_k, y_k)$, $y_{k+1} = g(x_k, y_k)$. Ovdje (ξ, η) označava rješenje sistema $x = f(x, y)$, $y = g(x, y)$.

Teorema. ... (ii) $q_\infty < 1$ onda ... .

Slijede razne dopune.

O ocjenjivanju greške

Mi smo našli prilično eksplicitno izražene dovoljne uslove za konvergenciju metode proste iteracije. Treba izvršiti dobru lokalizaciju tačnog rješenja, time se postigne da skup A bude "mali", time se postigne da su dobri izgledi da se taj skup preslikava u samog sebe. Takođe treba da se polazni sistem na pogodan način transformiše u oblik po kome se onda vrše iteracije. Time se postigne da uslov kontrakcije bude ispinjen. Konkretno, treba izračunati pokazatelje m_{ij}. Dovoljno je da važi jedan od dva uslova $q_1 < 1$, $q_\infty < 1$. U toj situaciji, mi smo sigurni da ćemo primjenom metode proste iteracije postići željeni cilj. Računajući sve više i više članova niza $\{(x_k, y_k)\}_{k=1}^{\infty}$ mi ćemo se neograničeno dobro približiti ka traženom rješenju (ξ, η). Na

ovom mjestu, u numeričkim metodama, uvijek se samo po sebi postavlja jedno te isto pitanje. Ako je specifična iteracija izračunata, koliko je ona udaljena od tačnog rješenja (ξ, η). Da li da računamo još iteracija. Ako je unaprijed propisana tačnost koju treba postići, da li posljednje izračunato zadovoljava. O ocjeni greške govorićemo generalno, nezavisno od toga koja norma se ima u vidu, tako da se ovo što slijedi odnosi i na jednu i na drugu prethodnu teoremu.

Kraj skalarnih oznaka. Vraćamo se dakle na vektorske oznake, jer nam je tako za ubuduće pogodnije.

Jednu formulu za ocjenu greške već imamo. Iz Banahove teoreme, njen dio (3), tamo piše:

$$\rho(x^{(k)}, \xi) \leq \frac{q^k}{1-q}\rho(x^{(1)}, x^{(0)}) \qquad \left(\text{ili } \|x^{(k)} - \xi\| \leq \frac{q^k}{1-q}\|x^{(1)} - x^{(0)}\|\right). \tag{1}$$

Ponovimo osnovni koncept iz numeričke. Rastojanje između približnog $x^{(k)}$ i tačnog ξ jeste upravo greška približne vrijednosti. Ako poslije tog rastojanja piše $\leq$ onda je to upravo formula za ocjenu greške. Važno je da broj koji dolazi poslije znaka $\leq$ može da bude izračunat efektivno.

Da bi se ova formula upotrebljavala, samo je treba konkretizovati. U njoj se pojavljuje "opšta" metrika ρ. A mi radimo po normi sa indeksom 1 ili po normi sa indeksom ∞. Ako je norma sa indeksom 1 onda ρ ima svoj sljedeći konkretni izraz, itd.

Pogledajmo formulu (1). Kojim tempom opada greška kada se ide sve dalje u aproksimacionom nizu $\{x^{(k)}\}$? Kolika je brzina konvergencije metode proste iteracije? Veličina $\rho(x^{(k)}, \xi)$ tj. greška$(x^{(k)})$ ocjenjuje se brojem q^k puta jedna konstanta koja ne zavisi od k. A greška$(x^{(k+1)})$ se ocjenjuje sa q^{k+1} puta ona ista konstanta. Prilikom prelaska od jedne iteracije na sljedeću, došlo je do množenja ocjene za grešku sa brojem $q < 1$. Možemo reći da se greška na svakom koraku množi sa q. U vezi toga, kaže se da metoda proste iteracije ima prvi (linearni) stepen ili red ili brzinu konvergencije.

Postoji još jedna formula za ocjenu greške. Upravo, važi sljedeća nejednakost:

$$\rho(x^{(k)}, \xi) \leq \frac{q}{1-q}\rho(x^{(k)}, x^{(k-1)}). \tag{2}$$

Dokažimo ovu nejednakost. Po aksiomi trougla imamo da je

$$\rho(x^{(k)}, \xi) \leq \rho(x^{(k)}, x^{(k+1)}) + \rho(x^{(k+1)}, \xi) =$$

(s obzirom da je $g(x^{(k-1)}) = x^{(k)}$, $g(x^{(k)}) = x^{(k+1)}$ i $g(\xi) = \xi$)

$$\rho(g(x^{(k-1)}), g(x^{(k)})) + \rho(g(x^{(k)}), g(\xi)) \leq q\rho(x^{(k-1)}, x^{(k)}) + q\rho(x^{(k)}, \xi).$$

Spajajući početak $\rho(x^{(k)}, \xi)$ i kraj $q\rho(x^{(k-1)}, x^{(k)}) + q\rho(x^{(k)}, \xi)$, odmah dolazimo do formule (2).

U praktičnom radu, za ocjenjivanje greške, bolje je da se koristi formula (2) nego formula (1). Formula (2) daje precizniju ocjenu za grešku. Budući da se oslanja na novije iteracije $x^{(k-1)}$ i $x^{(k)}$, a ne na polazne $x^{(0)}$ i $x^{(1)}$.

Iskoristimo još formulu (2). Da li razlika između dvije posljednje izračunate iteracije može da posluži kao ocjena greške posljednje iteracije? Može ako je $q \leq 1/2$. Zaista, $q \leq 1/2 \Rightarrow \rho(x^{(k)}, \xi) \leq \rho(x^{(k)}, x^{(k-1)})$. Preko norme: ako je $q \leq 1/2$ onda je $\|x^{(k)} - \xi\| \leq \|x^{(k)} - x^{(k-1)}\|$.

O slučaju $n = 1$

Korisno je da se posebno pogleda jedno–dimenzioni slučaj, tj. slučaj jedne nelinearne jednačine $x = \varphi(x)$ sa jednom nepoznatom x. Treba se uvjeriti da φ preslikava izvjesni odsječak $[a, b]$ u taj isti odsječak. I da važi $|\varphi'(x)| \leq q < 1$ za $x \in [a, b]$. Za ocjenu greške koristiti formulu (2): $|x^{(k)} - \xi| \leq \frac{q}{1-q}|x^{(k)} - x^{(k-1)}|$. Očito je $\rho(\alpha, \beta) = |\alpha - \beta|$.

REZIME o metodi proste iteracije u jedno–dimenzionom slučaju. Želimo da nađemo rješenje jednačine oblika $x = \varphi(x)$. U nastavku se ponavlja teorema koja je već dokazana ranije, samo što su oznake malo prilagođene.

Teorema. Razmotrimo funkciju $\varphi \in C^1[a,b]$. Neka su ispunjeni uslovi: (1) ako je $a \le x \le b$ onda je $a \le \varphi(x) \le b$ i (2) postoji q takav da je $|\varphi'(x)| \le q < 1$ za $a \le x \le b$. Definišimo niz brojeva $\{x_n\}_{n=0}^{\infty}$ sa $x_{n+1} = \varphi(x_n)$ za $n \ge 0$, gdje $x_0 \in [a,b]$. Tada važi: (1) jednačina $x = \varphi(x)$ ima jedinstveno rješenje na odsječku $[a,b[$ (označimo ga sa ξ) i (2) $\lim_{n\to\infty} x_n = \xi$.

Dokaz teoreme. Po Lagranžovoj $\varphi(\alpha_2) - \varphi(\alpha_1) = \varphi'(\beta)(\alpha_2 - \alpha_1)$ gdje je $\alpha_1 < \beta < \alpha_2 \Rightarrow |\varphi(\alpha_2) - \varphi(\alpha_1)| = |\varphi'(\beta)| \cdot |\alpha_2 - \alpha_1| \le q|\alpha_2 - \alpha_1|$.

Kako φ: $[a,b] \to [a,b]$ i $x_0 \in [a,b]$ to $x_n \in [a,b]$ za svako n.

Dokažimo da je $\{x_n\}$ Košijev niz. Imamo $|x_{n+1} - x_n| = |\varphi(x_n) - \varphi(x_{n-1})| \le q|x_n - x_{n-1}|$. Slično $|x_{n+2} - x_{n+1}| = |\varphi(x_{n+1}) - \varphi(x_n)| \le q|x_{n+1} - x_n| \le q^2|x_n - x_{n-1}|$. Itd. Isto tako $|x_{n+p} - x_{n+p-1}| \le q^p|x_n - x_{n-1}|$. Na isti način se dokazuje i $|x_n - x_{n-1}| \le q^{n-1}|x_1 - x_0|$. Ukupno

$$|x_{n+p} - x_n| = |x_{n+p} - x_{n+p-1} + \ldots + x_{n+2} - x_{n+1} + x_{n+1} - x_n| \le$$

$$|x_{n+p} - x_{n+p-1}| + \ldots + |x_{n+2} - x_{n+1}| + |x_{n+1} - x_n| \le (q^p + \ldots + q^2 + q)|x_n - x_{n-1}| \le$$

$$(q + q^2 + \ldots)|x_n - x_{n-1}| = \frac{q}{1-q}|x_n - x_{n-1}| \le \frac{q^n}{1-q}|x_1 - x_0| \to 0 \ \text{ kad } \ n \to \infty$$

bez obzira na $p \ge 1$.

Budući da je metrički prostor kompletan, to je niz $\{x_n\}$ i konvergentan i odmah uvodimo oznaku $X = \lim_{n\to\infty} x_n$. Iz $a \le x_n \le b$ za svako $n \Rightarrow a \le X \le b$. Iz $x_{n+1} = \varphi(x_n)$ slijedi $X = \lim_{n\to\infty} \varphi(x_n)$ i dalje slijedi (budući da je φ neprekidna funkcija) $X = \varphi(X)$. Znači $X = \xi$. Ne mogu postojati dva rješenja ξ_1 i ξ_2 jer bi tada bilo $\varphi(\xi_1) = \xi_1$, $\varphi(\xi_2) = \xi_2$ i (za neko β) $|\xi_2 - \xi_1| = |\varphi(\xi_2) - \varphi(\xi_1)| = |\varphi'(\beta)| \cdot |\xi_2 - \xi_1| \le q|\xi_2 - \xi_1|$ a znamo da je $q < 1$. Dokaz je završen.

Važi nejednakost (za ocjenu greške) $|x_n - \xi| \le \frac{q^n}{1-q}|x_1 - x_0|$ za svako n. Isto tako, $|x_n - \xi| \le \frac{q}{1-q}|x_n - x_{n-1}|$ za svako n.

Dokaz druge nejednakosti:

$$|x_n - \xi| \le |x_n - x_{n+1}| + |x_{n+1} - \xi| = |\varphi(x_{n-1}) - \varphi(x_n)| + |\varphi(x_n) - \varphi(\xi)| \le$$

$$q|x_{n-1} - x_n| + q|x_n - \xi| \ \ \Rightarrow \ \ (1-q)|x_n - \xi| \le q|x_{n-1} - x_n| \ \ \big/ : (1-q)$$

Dokaz prve nejednakosti:

$$|x_n - \xi| \le \frac{q}{1-q}|x_n - x_{n-1}| \le \frac{q}{1-q}|\varphi(x_{n-1}) - \varphi(x_{n-2})| \le \frac{q^2}{1-q}|x_{n-1} - x_{n-2}| =$$

$$\frac{q^2}{1-q}|\varphi(x_{n-2}) - \varphi(x_{n-3})| \le \frac{q^3}{1-q}|x_{n-2} - x_{n-3}| \le \ldots \le \frac{q^n}{1-q}|x_1 - x_0|$$

Jasno, po Lagranžovoj teoremi imamo $\varphi(\alpha_2) - \varphi(\alpha_1) = \varphi'(\beta)(\alpha_2 - \alpha_1)$ gdje je $\beta = \alpha_1 + \theta(\alpha_2 - \alpha_1)$ za neko $0 < \theta < 1 \Rightarrow |\varphi(\alpha_2) - \varphi(\alpha_1)| \le q|\alpha_2 - \alpha_1|$ za bilo koje $\alpha_1, \alpha_2 \in [a,b]$.

Primjer. Razmotrimo jednačinu $x = \sqrt{1+x}$ na odsječku $1 \le x \le 2$. Nacrtati odgovarajuću sliku, nacrtati grafik funkcije $y = \sqrt{1+x}$ na dijelu $1 \le x \le 2$, kao i pravu $y = x$. Stavimo $\varphi(x) = \sqrt{1+x}$. Data jednačina ima na tom odsječku jedinstveno rješenje, rješenje može da bude nađeno po metodi proste iteracije (sa proizvoljnom preciznošću). Zato što su ispunjeni uslovi teoreme od maločas. Zaista, $1 \le x \le 2 \Rightarrow \sqrt{2} \le \varphi(x) \le \sqrt{3}$, odnosno funkcija φ prevodi odsječak $[1,2]$ u samog sebe. Pored toga, $\varphi'(x) = \frac{1}{2\sqrt{1+x}}$, pa je $|\varphi'(x)| \le \frac{1}{2}$ za $x \in [1,2]$. Imamo

da je $q = \frac{1}{2}$. Radeći preciznije, možemo pisati da je $q = \frac{1}{2\sqrt{2}}$. Kao x_0 uzme se bilo koja tačka odsječka $[1,2]$, recimo stavimo da je $x_0 = 1{,}5$. Zatim redom računamo po formuli $x_{n+1} = \sqrt{1+x_n}$. Proces će da konvergira dosta dobrim tempom zato što je koeficijent kontrakcije q manji od jedne polovine. Imamo da je $\lim_{n\to\infty} x_n = \xi$, gdje je $\xi = (1+\sqrt{5})/2 = 1{,}61803$.

Kontrakcija: mali primjer: $\alpha_2 - \alpha_1 = 2 - 1 = 1$, $\varphi(\alpha_2) - \varphi(\alpha_1) = 1{,}73 - 1{,}41 = 0{,}32$, $q = 1/2\sqrt{2} = 0{,}35$.

Uopšte, kada treba riješiti jednačinu npr. $x = \sqrt{1+x}$, mi prvo nacrtamo dva grafika $y = x$ i $y = \sqrt{1+x}$ i vidimo gdje se krive sijeku. Tako odredimo otprilike gdje ima rješenja, to je tzv. lokalizacija rješenja. Odredimo $[a,b]$. Jednačina $x = \varphi(x) \Rightarrow$ dva grafika $y = x$ i $y = \varphi(x)$.

Kako postići da uslov kontrakcije bude ispunjen?

Budući da se radi o numeričkoj, dosadašnja priča ne vrijedi puno ako se ne da neko uputstvo o pogodnom načinu transformacije iz polaznog oblika $f(x) = 0$ u oblik $x = g(x)$. O izboru g. Uputstvo koje slijedi ne garantuje da se uvijek može uspješno primijeniti metoda proste iteracije, ali u mnogim slučajevima omogućuje.

Pogledajmo prvo na primjeru, ovdje je $n = 1$. Neka se traži rješenje jednačine $x = \varphi(x)$ na nekom odsječku $[a,b]$. Neka smo za prvi izvod funkcije φ na tom odsječku ustanovili da važi $4 \le \varphi'(x) \le 6$. Ovo vrlo slabo izgleda sa stanovišta primjene metode proste iteracije. Međutim, polazna jednačina je ekvivalentna sa $x = \frac{x+\varphi(x)}{2}$. Izvod funkcije sa desne strane $\frac{x+\varphi(x)}{2}$ kreće se očito između 2,5 i 3,5. Izgleda nam da se nekako može podesiti da se izvod desne strane kreće između $-q$ i q, sa $q < 1$.

Uopšte, jednačina $x = \varphi(x)$ očito je ekvivalentna sa jednačinom $x + \lambda x = \varphi(x) + \lambda x$ tj. $x = \frac{\varphi(x)+\lambda x}{1+\lambda}$. Ovdje je λ bilo koja konstanta, samo da je $\neq -1$. Veza izvoda funkcije $\varphi(x)$ i funkcije $\frac{\varphi(x)+\lambda x}{1+\lambda}$ je očita. Koliko je najbolje da se izabere λ u primjeru od maloprije? Uspješnost primjene ovog tzv. λ–postupka zavisi od toga – koliko jasnom informacijom o ponašanju izvoda φ' raspolažemo. Tako da će se sukcesivne aproksimacije računati po formuli $x_{k+1} = \frac{\varphi(x_k)+\lambda x_k}{1+\lambda}$. Drukčije se može reći: neka funkciju $\frac{\varphi(x)+\lambda x}{1+\lambda}$ označavamo odsad kao $\varphi(x)$.

Izložimo ovo uputstvo u slučaju $n = 2$. Neka polazni sistem glasi

$$a(x,y) = 0,\ b(x,y) = 0;$$

opet skalarne oznake. Ovaj sistem može očito da se zapiše u drugom obliku kao:

$$x = x + \alpha a(x,y) + \beta b(x,y),\ y = y + \gamma a(x,y) + \delta b(x,y).$$

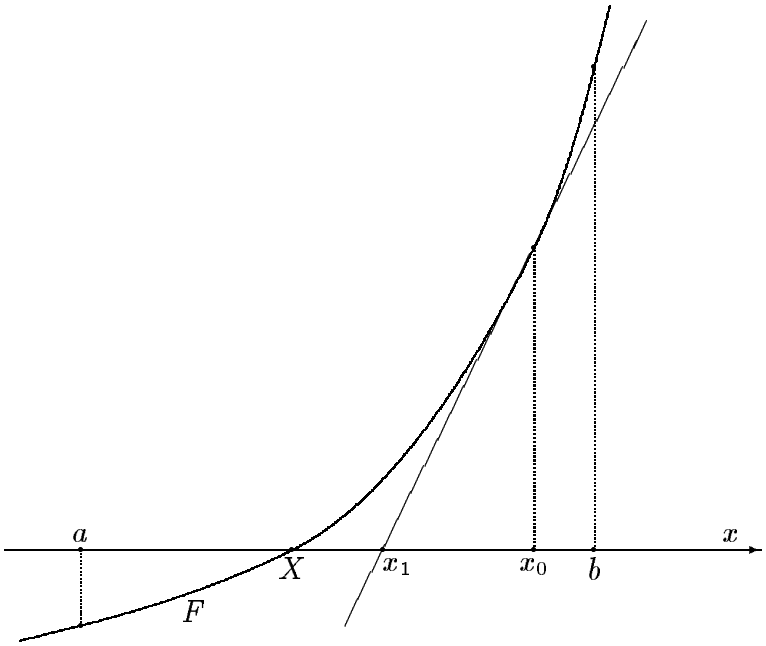
Samo treba da bude $\alpha\delta - \beta\gamma \neq 0$. Tako da će drugi oblik poslužiti za formiranje iterativnog niza. Konstante α, β, γ i δ biraju se pogodno tj. biraju se tako da prvi parcijalni izvodi funkcije $x + \alpha a(x,y) + \beta b(x,y)$ i funkcije $y + \gamma a(x,y) + \delta b(x,y)$ budu bliski nuli.

4.3. NJUTNOVA METODA

Uvod. Neka je F nelinearno preslikavanje i neka se razmatra jednačina $F(x) = 0$, da li ima rješenja, kako naći to rješenje. Preslikavanje F može da dejstvuje u R (to znači da F: $R \to R$) ili da dejstvuje u R^n ili da dejstvuje u opštem konačno– ili beskonačno–dimenzionom normiranom prostoru (Banahovom prostoru) X. U prethodnom naslovu smo vidjeli da za ovaj zadatak može da posluži metoda proste iteracije, zasnovana na Banahovoj teoremi o nepokretnoj tački. U ovom naslovu ćemo vidjeti da za rješavanje jednačine $F(x) = 0$ može da posluži druga jedna metoda – Njutnova metoda. Ova se metoda još naziva i metoda tangente, vezano za njenu ideju i za njeno geometrijsko tumačenje, kao što ćemo vidjeti. Možemo reći da je metoda tangente

(Njutnova metoda) vezana za sami pojam prvog izvoda nekog preslikavanja (u nekoj tački). U smislu da je izvodno preslikavanje – ono linearno preslikavanje koje, među svim mogućim linearnim preslikavanjima, ostvaruje najbolju aproksimaciju polaznog (nelinearnog) preslikavanja (u određenoj tački). O značaju ove dvije metode, metode proste iteracije i Njutnove metode, može se suditi već po veličini dvojice matematičara Banaha i Njutna. Ovo su dvije glavne metode za numeričko rješavanje jednačine $F(x) = 0$, jednačina – u Banahovom prostoru. Isto tako, ovo su dvije glavne metode za bilo kakvo (teorijsko) ispitivanje jednačine $F(x) = 0$, u funkcionalnoj analizi, u teoriji operatora. Što se tiče Njutnove metode, Njutn je ovu metodu razvio za slučaj $F\colon R \to R$. Kasnije je ova metoda uopštena za slučaj $F\colon R^n \to R^n$. Kasnije je ovu metodu dalje uopštio ruski matematičar L. Kantorovič na slučaj $F\colon X \to Y$, X i Y – Banahovi prostori.

Pogledajmo prvo slučaj $F\colon R \to R$. Dakle, neka je F (nelinearna) realna funkcija realne promjenljive, tj. $F\colon R \to R$ ili $F\colon [a,b] \to R$. Treba odrediti približnu vrijednost (jednog) rješenja jednačine $F(x) = 0$. Ideja Njutnove metode i njene glavne osobine lijepo se vide na ovom jednostavnom slučaju jedne nelinearne jednačine sa jednom nepoznatom. Pripremni korak za primjenu Njutnove metode jeste lokalizacija korijena jednačine, nule funkcije F. Sprovođenje ovog pripremnog koraka nije vezano za samu Njutnovu metodu, vrši se na isti način kao u slučaju metode proste iteracije. Naravno da je bolje što je polazni odsječak $[a,b]$ za koji smo sigurni da sadrži traženi korijen – što kraći. Do tog polaznog odsječka dolazi se tako što se vrijednosti funkcije F u nekoliko tačaka izračunaju, pa se vidi gdje dolazi do promjene znaka funkcije. Možemo neku tačku x_0 polaznog odsječka $[a,b]$ smatrati početnom aproksimacijom. Tačno rješenje razmatrane jednačine označavaćemo kao X; dakle, važi $F(X) = 0$.



Slika 3

Vidi sliku 3. Prikazan je grafik jedne neprekidne funkcije $y = F(x)$ na dijelu $a \le x \le b$. Prikazano je da funkcija F u krajnjim tačkama odsječka ima vrijednosti koje se razlikuju po znaku, $F(a) \cdot F(b) < 0$. U lijevom da je recimo negativna a u desnom kraju da je samim tim pozitivna. Zato jednačina $F(x) = 0$ ima bar jedno rješenje X. Prikazano je da je F strogo monotona funkcija na odsječku $[a,b]$, da je rastuća, njen prvi izvod je pozitivan. Zato je rješenje

X jedinstveno. Još je prikazano da funkcija raste sve brže i brže kako se krećemo po x–osi; ovo znači da je $F''(x) > 0$ na pomenutom odsječku. Na slici je prikazana i početna aproksimacija x_0. Zapaziti da je tačka x_0 desno od tačke X. Zato je $F(x_0) > 0$. Mogli smo da uzmemo i $x_0 = b$. Kako naći neku aproksimaciju za X bolju od x_0? Nacrtajmo i grafik tangente $y = G(x)$ na krivu $y = F(x)$ u tački $x = x_0$. U maloj okolini tačke $x = x_0$, odgovarajuće vrijednosti $F(x)$ i $G(x)$ su bliske jedna drugoj. U blizini tačke $x = x_0$ jedan i drugi grafik se skoro poklapaju. Govoreći drukčije, neka je $F(x)$ predstavljena po Tejlorovoj formuli. I $G(x)$ takođe. Nulti i prvi sabirak jednog i drugog predstavljanja, razvoja se poklapaju. Tek na drugom sabirku, koji ima $(x - x_0)^2$, pojavljuje se razlika. Drugim riječima, važi $F(x_0) = G(x_0)$ i važi $F'(x_0) = G'(x_0)$. Na primjer, jednačina tangente na krivu $y = 5 + 4x + 3x^2$ u tački $x = 0$ glasi $y = 5 + 4x$. Kriva $y = F(x)$ siječe x–osu u tački X. Kriva (ustvari prava) $y = G(x)$ siječe x–osu u jednoj tački; označimo tu tačku sa x_1. Upravo x_1 uzimamo za bolju aproksimaciju korijena X. Da je tangenta bila postavljena u tački $x_0 < X$ onda bismo se radeći ovako udaljili od X a ne približili X. Ako $y = F(x)$ nije monotona onda se grafik funkcije i grafik tangente raziđu. Slično ako F' nije monotona funkcija. Kako dobiti još bolju aproksimaciju x_2? Istim postupkom. U tački $x = x_1$ postaviti tangentu na $y = F(x)$ i odrediti presječnu tačku te tangente i x–ose; ta presječna tačka jeste upravo x_2. Ako crtež dopunimo, ako na njemu prikažemo, izvedemo još i tačke x_2 i x_3, stičemo utisak da je x_2 još puno bliže tački X, a da se x_3 skoro poklapa sa X. Ispostaviće se da je ovaj utisak ispravan, vidi kasnije formule. Upravo, tempo konvergencije Njutnove metode od iteracije do iteracije u nizu iteracija nije stalan, već se povećava kako se ide dalje u tom nizu, dolazi do ubrzanja. Na redu je mali i jednostavni računski dio izvođenja, konstrukcije Njutnove metode (za slučaj F: $R \to R$). Treba naći izraz za x_1 preko x_0. Istog oblika će naravno biti i izraz za x_{k+1} u zavisnosti od x_k.

Jednačina tangente $y = G(x)$ glasi $y - F(x_0) = F'(x_0)(x - x_0)$ (jednačina prave kroz jednu tačku) ili $y = F(x_0) + F'(x_0)(x - x_0)$. Rečeno je da ćemo rješenje jednačine $G(x) = 0$ tj. $F(x_0) + F'(x_0)(x - x_0) = 0$ da označimo sa x_1. Nalazimo da je $x_1 = x_0 - F(x_0)/F'(x_0)$. Zapaziti da razvoj funkcije $y = F(x)$ po Tejlorovoj formuli u okolini tačke $x = x_0$ sa ostatkom koji sadrži drugi izvod glasi $F(x) = F(x_0) + F'(x_0)(x - x_0) + O((x - x_0)^2)$. Već znamo da se svaki novi član aproksimacionog niza dobija na osnovu prethodnog po istom šablonu, u x_k se postavi tangenta, gleda se njen presjek sa x–osom. Dakle, $x_{k+1} = x_k - F(x_k)/F'(x_k)$.

Teorema (o dovoljnim uslovima za konvergenciju Njutnove metode). Neka su ispunjeni sljedeći uslovi: $F \in C^2[a, b]$; $F(a) \cdot F(b) < 0$; $F'(x)$ je konstantnog znaka na $[a, b]$ (ovo znači sljedeće: važi da je funkcija F' pozitivna na čitavom odsječku $[a, b]$ ili važi da je $F'(x) < 0$ za svako $x \in [a, b]$); funkcija F'' ima stalni znak na cijelom odsječku $[a, b]$; tačka x_0 iz $[a, b]$ izabrana je tako da bude $F(x_0) \cdot F''(x_0) > 0$. Neka je niz brojeva $\{x_k\}$ definisan sa:

$$x_{k+1} = x_k - \frac{F(x_k)}{F'(x_k)} \tag{1}$$

za $k \geq 0$. Tada jednačina $F(x) = 0$ ima jedinstveno rješenje na $[a, b]$ (označimo ga sa X). I važi da $x_k \to X$ kad $k \to \infty$.

Dokaz. Iz $F(a) \cdot F(b) < 0$ i F' je stalnog znaka slijedi postojanje i jedinstvenost rješenja X. U zavisnosti od toga kakvog su znaka $F'(x)$ i $F''(x)$ moguća su četiri slučaja, i to: 1) $F' > 0$, $F'' > 0$, 2) $F' > 0$, $F'' < 0$, 3) $F' < 0$, $F'' > 0$ i 4) $F' < 0$, $F'' < 0$. Mi ćemo sprovesti dokaz za prvi slučaj. Za ostala tri slučaja, dokaz je sličan. Dakle, imamo okolnosti kao na slici. Prvo. Važi da je $x_k > X$ za svako $k \geq 0$ (zapaziti odmah da je ovaj uslov ekvivalentan sa uslovom $F(x_k) > 0$). Dokazuje se indukcijom. Imamo da je $x_0 > X$ jer je $F''(x) > 0$ i $F(x_0) \cdot F''(x_0) > 0$. Ako je $x_k > X$ onda je i $x_{k+1} > X$. Zaista, geometrijski, gledajući od tačke $x = x_0$ unazad, tangenta opada brže od funkcije, pa će tangenta presjeći x–osu prije nego

funkcija. A analitički se pokaže ako se napiše razvoj po Tejlorovoj formuli čiji ostatak je izražen preko drugog izvoda. Drugo. Niz $\{x_k\}_{k=0}^{\infty}$ je monotono opadajući, tj. važi da je $x_{k+1} < x_k$ za $k \geq 0$. Zaista, $x_{k+1} - x_k = -F(x_k)/F'(x_k)$, pri čemu je $F(x_k) > 0$ (maločas je pokazano), a $F'(x_k)$ takođe > 0 (definicioni uslov prvog slučaja). Treće. Taj niz je konvergentan, budući da je monoton i da je ograničen (svi njegovi elementi su veći od X). Označimo sa ξ graničnu vrijednost ovog niza $\{x_k\}$. I četvrto. Na relaciju (1) primijenimo operaciju $\lim_{k\to\infty}$. Niz $\{x_{k+1}\}$ je podniz od $\{x_k\}$, pa i on teži ka ξ. Funkcija F je neprekidna pa važi $\lim_{k\to\infty} F(x_k) = F(\lim_{k\to\infty} x_k)$. Tako da važi sljedeće: $\xi = \xi - F(\xi)/F'(\xi)$. Kako $\xi \in [a,b]$ to je $F'(\xi) \neq 0$. Dakle, $F(\xi) = 0$. Znači da je $\xi = X$. Dokaz je završen.

Teorema (o ocjeni greške). Važi sljedeća nejednakost koja nam služi kao formula za ocjenjivanje greške: greška(x_k) =

$$|X - x_k| \leq \frac{M_2}{2m_1}(x_k - x_{k-1})^2$$

za svako $k \geq 1$. Ovdje je m_1 ma koji broj za koji važi $m_1 \leq |F'(x)|$ za svako $x \in [a,b]$. A M_2 je bilo koji broj koji zadovoljava uslov $|F''(x)| \leq M_2$ za $x \in [a,b]$.

Ova teorema predstavlja nastavak prethodne teoreme, pa od nje preuzima oznake i uslove. Naravno da se može reći da je m_1 (tačni) infimum (a ne bilo koja njegova ocjena sa donje strane) od $|F'(x)|$. Broj $m_1 > 0$ sa ovim svojstvom sigurno postoji, jer je $F'(x) \neq 0$ i F' neprekidna i $[a,b]$ kompaktan. Slično se može reći da je M_2 supremum modula drugog izvoda.

Dokaz. Razvijmo funkciju $y = F(x)$ po Tejlorovoj formuli u okolini tačke $x = x_{k-1}$ do drugog izvoda:

$$F(x) = F(x_{k-1}) + F'(x_{k-1})(x - x_{k-1}) + F''(\alpha)(x - x_{k-1})^2/2,$$

α zavisi od x. Iskoristimo ovaj razvoj za $x = x_k$: $F(x_k) = F(x_{k-1}) + F'(x_{k-1})(x_k - x_{k-1}) + F''(\alpha)(x_k - x_{k-1})^2/2$, α je neka tačka između x_{k-1} i x_k, $|F''(\alpha)| \leq M_2$. Izraz $F(x_{k-1}) + F'(x_{k-1})(x_k - x_{k-1})$, linearni dio Tejlorovog razvoja, jednak je u ovoj situaciji nuli; geometrijska definicija prvog izvoda, definicija metode tangente. Tako da se ranija formula svodi na $F(x_k) = F''(\alpha)(x_k - x_{k-1})^2/2$. Slijedi da je

$$|F(x_k)| \leq \frac{M_2}{2}(x_k - x_{k-1})^2. \tag{2}$$

S druge strane imamo: $F(x_k) = F(x_k) - F(X) = F'(\beta)(x_k - X)$, Lagranžova teorema. Slijedi da je $|F(x_k)| = |F'(\beta)| \cdot |x_k - X|$, $|F(x_k)| \geq m_1|x_k - X|$ ili svejedno

$$|x_k - X| \leq \frac{|F(x_k)|}{m_1}. \tag{3}$$

Zapaziti usput da se i posljednja relacija (3) može smatrati jednom formulom za ocjenu greške, za ocjenu udaljenosti nekog broja x_k od nule X funkcije F, kao i da izvođenje te relacije nije zavisno od Njutnove metode, pa se ta relacija može koristiti za Njutnovu metodu a i za druge metode. Kombinovanjem (2) i (3) imamo $|x_k - X| \leq \frac{1}{m_1} \cdot \frac{M_2}{2}(x_k - x_{k-1})^2$. Dokaz je završen.

Nabrojimo sada četiri glavne karakteristike 1)–4) Njutnove metode. 1) Red ili brzina konvergencije Njutnove metode je drugi (kvadratni). Ovo je glavni kvalitet Njutnove metode! Znamo da metoda proste iteracije ima (samo) prvi red konvergencije. Pogledajmo, da se uvjerimo, formule koje izražavaju grešku(x_k) preko razlike između dvije susjedne aproksimacije, u slučaju jedne i druge metode. Kod metode proste iteracije bilo je $\|\xi - x_k\| \leq const \cdot \|x_k - x_{k-1}\|$. A kod Njutnove metode je $\|\xi - x_k\| \leq const \cdot \|x_k - x_{k-1}\|^2$; v. posljednju teoremu, ovdje je

$const = \frac{M_2}{2m_1}$. 2) Za uspješnu primjenu Njutnove metode potrebno je da raspolažemo dovoljno dobrom početnom aproksimacijom x_0. Možemo reći da ovaj iskaz važi i kada je riječ o metodi proste iteracije. Ipak, za Njutnovu metodu, ovaj zahtjev je nekako izraženiji. 3) Da bi Njutnova metoda konvergirala treba da $|F'|$ bude odvojen od nule. Ili svejedno, treba da $|F'|^{-1}$ bude ograničena veličina. Zaista, v. pokazatelj m_1 iz prethodne teoreme. I 4) da bi Njutnova metoda konvergirala, treba da $|F''|$ bude odvojen od $+\infty$ (tj. da bude ograničen). Zaista, v. M_2.

Za Njutnovu metodu u slučaju jedne dimenzije drukčije se kaže da je metoda tangente.

Prelazimo na slučaj $n = 2$, sistem od dvije nelinearne jednačine sa dvije nepoznate, $F\colon R^2 \to R^2$, $F(x) = 0$. Neka bude $F = (F_1, F_2)$ gdje $F_1\colon R^2 \to R$ i $F_2\colon R^2 \to R$ i neka bude $x = (x_1, x_2) \in R^2$. Mi tražimo približno rješenje sistema $F_1(x_1, x_2) = 0$, $F_2(x_1, x_2) = 0$. Dopustimo da već raspolažemo početnom aproksimacijom $x_0 = (x_{10}, x_{20})$. Kako odrediti iduću aproksimaciju (x_{11}, x_{21})? Mi funkcije F_1 i F_2 razvijemo u Tejlorov red u okolini tačke $(x_1, x_2) = (x_{10}, x_{20})$. Zatim od ta dva Tejlorova reda zadržimo samo nulti i prvi sabirak, a sve ostale sabirke odbacimo. Drugim riječima, izvršimo linearizaciju funkcija $F_1(x_1, x_2)$ i $F_2(x_1, x_2)$. Umjesto da rješavamo dati sistem $F_1(x_1, x_2) = 0$, $F_2(x_1, x_2) = 0$, čijim bismo rješavanjem našli tačni korijen $X = (X_1, X_2)$, mi rješavamo približni linearizovani sistem, čije rješenje i jeste iduća aproksimacija $k = 1$. Na sličan način se od k–te aproksimacije $x_k = (x_{1k}, x_{2k})$ prelazi na $(k+1)$–vu $(x_{1,k+1}, x_{2,k+1})$. Napišimo odgovarajuće formule. Nulti sabirak Tejlorovog reda ima vrijednost funkcije (u tački u kojoj se razvoj vrši), recimo $F_1(x_{1k}, x_{2k})$. A prvi sabirak ima, u slučaju funkcije od dvije promjenljive, prvi parcijalni izvod po prvom argumentu (sračunat u tački u kojoj se razvoj vrši) puta priraštaj prvog argumenta plus prvi parcijalni izvod po drugom argumentu puta priraštaj drugog argumenta. "Tačni" nelinearni sistem, čije je rješenje $(x_1, x_2) = (X_1, X_2)$, glasi:

$$\begin{cases} F_1(x_1, x_2) = 0 \\ F_2(x_1, x_2) = 0 \end{cases}$$

tj.

$$\begin{cases} F_1(x_{1k}, x_{2k}) + \dfrac{\partial F_1(x_{1k}, x_{2k})}{\partial x_1}(x_1 - x_{1k}) + \dfrac{\partial F_1(x_{1k}, x_{2k})}{\partial x_2}(x_2 - x_{2k}) + \ldots = 0 \\ F_2(x_{1k}, x_{2k}) + \dfrac{\partial F_2(x_{1k}, x_{2k})}{\partial x_1}(x_1 - x_{1k}) + \dfrac{\partial F_2(x_{1k}, x_{2k})}{\partial x_2}(x_2 - x_{2k}) + \ldots = 0 \end{cases}$$

A odgovarajući približni, linearni sistem, čije je rješenje $(x_1, x_2) = (x_{1,k+1}, x_{2,k+1})$, glasi:

$$\begin{cases} F_1(x_{1k}, x_{2k}) + \dfrac{\partial F_1(x_{1k}, x_{2k})}{\partial x_1}(x_1 - x_{1k}) + \dfrac{\partial F_1(x_{1k}, x_{2k})}{\partial x_2}(x_2 - x_{2k}) = 0 \\ F_2(x_{1k}, x_{2k}) + \dfrac{\partial F_2(x_{1k}, x_{2k})}{\partial x_1}(x_1 - x_{1k}) + \dfrac{\partial F_2(x_{1k}, x_{2k})}{\partial x_2}(x_2 - x_{2k}) = 0 \end{cases}$$

ili

$$\begin{bmatrix} F_1(x_{1k}, x_{2k}) \\ F_2(x_{1k}, x_{2k}) \end{bmatrix} + \begin{bmatrix} \dfrac{\partial F_1(x_{1k}, x_{2k})}{\partial x_1} & \dfrac{\partial F_1(x_{1k}, x_{2k})}{\partial x_2} \\ \dfrac{\partial F_2(x_{1k}, x_{2k})}{\partial x_1} & \dfrac{\partial F_2(x_{1k}, x_{2k})}{\partial x_2} \end{bmatrix} \cdot \begin{bmatrix} x_1 - x_{1k} \\ x_2 - x_{2k} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Posljednje se može napisati u obliku

$$F(x_k) + F'(x_k)(x - x_k) = 0,$$

gdje se pojavljuje (Jakobijeva) matrica $F'(x_k)$ prvih izvoda preslikavanja F. Već je rečeno da rješenje jednačine $F(x_k) + F'(x_k)(x - x_k) = 0$ jeste iduća aproksimacija $x = x_{k+1}$. Možemo

pisati da važi $F(x_k) + F'(x_k)(x_{k+1} - x_k) = 0$. Da se ovo riješi, pomnožiti sa lijeve strane sa $(F'(x_k))^{-1}$. Linearni sistem se lako riješi, vektor x_{k+1} može efektivno da se izračuna. Vidimo da u računanju treba izvršiti inverziju matrice $F'(x_k)$, to je jedna matrica oblika 2×2, a pojavljuje se uslov $\Delta_k = \det F'(x_k) \neq 0$. Napišimo najzad formule koje određuju iduću aproksimaciju $x_{k+1} = (x_{1,k+1}, x_{2,k+1})$:

$$\Delta_k = \begin{vmatrix} \dfrac{\partial F_1(x_{1k}, x_{2k})}{\partial x_1} & \dfrac{\partial F_1(x_{1k}, x_{2k})}{\partial x_2} \\ \dfrac{\partial F_2(x_{1k}, x_{2k})}{\partial x_1} & \dfrac{\partial F_2(x_{1k}, x_{2k})}{\partial x_2} \end{vmatrix}, \quad \Delta_{1k} = \begin{vmatrix} F_1(x_{1k}, x_{2k}) & \dfrac{\partial F_1(x_{1k}, x_{2k})}{\partial x_2} \\ F_2(x_{1k}, x_{2k}) & \dfrac{\partial F_2(x_{1k}, x_{2k})}{\partial x_2} \end{vmatrix},$$

$$\Delta_{2k} = \begin{vmatrix} \dfrac{\partial F_1(x_{1k}, x_{2k})}{\partial x_1} & F_1(x_{1k}, x_{2k}) \\ \dfrac{\partial F_2(x_{1k}, x_{2k})}{\partial x_1} & F_2(x_{1k}, x_{2k}) \end{vmatrix}, \quad \begin{cases} x_{1,k+1} = x_{1k} - \Delta_{1k}/\Delta_k \\ x_{2,k+1} = x_{2k} - \Delta_{2k}/\Delta_k \end{cases}$$

Ovdje je $k = 0, 1, \ldots$. Time je čitav niz aproksimacija $\{(x_{1k}, x_{2k})\}_{k=0}^{\infty}$ definisan. Da li ovaj niz konvergira ka korijenu $X = (X_1, X_2)$? Dovoljni uslovi za konvergenciju i formula za ocjenu greške biće dati kasnije, u opštem slučaju, za Banahov prostor.

Sada nekoliko riječi o slučaju bilo kakvog $n > 1$. Ovaj slučaj se skoro ne razlikuje od slučaja $n = 2$. Sada je F: $R^n \to R^n$, jednačina (sistem od n nelinearnih jednačina sa n nepoznatih) glasi $F(x) = 0$, ovdje očito $x \in R^n$, $0 \in R^n$, tačni korijen sistema označićemo kao X, očito $X \in R^n$. Za komponente se uvode obične oznake, $F = (F_1, \ldots, F_n)$, $x = (x_1, \ldots, x_n)$. Za preslikavanje F se pretpostavlja da je definisano u čitavom prostoru R^n ili da je definisano u nekom podskupu A tog prostora R^n. Takođe se pretpostavlja da je to preslikavanje neprekidno–diferencijabilno. Izvod $F'(x)$ ovog preslikavanja će se pojaviti u računanju. Znamo da je $F'(x)$ jedna matrica oblika $n \times n$, njeni elementi izražavaju se preko prvih parcijalnih izvoda funkcija F_i (jasno da je F_i: $R^n \to R$). Znamo da je $F'(x) = \left[\dfrac{\partial F_i(x)}{\partial x_j}\right]_{i,j=1}^{n}$ Ako je tačka x fiksirana onda je matrica $F'(x)$ konstanta. Prelazimo na izvođenje formula koje definišu Njutnovu metodu (treći put ih izvodimo). Neka u datom trenutku raspolažemo k–tom aproksimacijom x_k tačnog rješenja X. Jednačina koja se rješava glasi $F(x) = 0$. Izračunajmo $F(x_k)$ i $F'(x_k)$. Pomoću ove dvije veličine može dobro da se procijeni $F(x_k + \eta)$ u okolini tačke x_k, tj. kada je η mala veličina, kada je $\|\eta\|$ mali broj, ima se u vidu neka norma u prostoru R^n, recimo $\| \ \|_2$. Greška kod ovakve procjene je reda $o(\|\eta\|)$. Ako je F dvaput neprekidno–diferencijabilno preslikavanje onda greška iznosi $O(\|\eta\|^2)$. Tako, važi da je $F(x_k + \eta) = F(x_k) + F'(x_k)\eta + o(\|\eta\|)$. Napišimo sljedeću jednačinu: $F(x_k) + F'(x_k)\eta = 0$ tj.

$$F(x_k) + F'(x_k)(x_{k+1} - x_k) = 0. \tag{4}$$

Dovoljan uslov da jednačina (4) po nepoznatoj x_{k+1} ima rješenja jeste da je linearni operator $F'(x_k)$ invertibilan tj. da je matrica $F'(x_k)$ regularna ($\det F'(x_k) \neq 0$). Tada rješenje glasi:

$$x_{k+1} = x_k - (F'(x_k))^{-1} \cdot F(x_k), \ k = 0, 1, \ldots. \tag{5}$$

Time je računski algoritam za Njutnovu metodu u slučaju R^n opisan. Vidimo da na svakom koraku treba da se izvrši inverzija jedne (konkretne, za taj korak) matrice oblika $n \times n$. Dovoljni uslovi za konvergenciju i ocjenu greške biće dati u nastavku, za opšti slučaj Banahovog prostora X.

Postoji i tzv. modifikovana Njutnova metoda. S jedne strane, tokom numeričke realizacije Njutnove metode najviše (računarskog) vremena se troši na računanje inverznih matrica $(F'(x_k))^{-1}$. S druge strane, obično se te matrice dobro stabilizuju poslije već malog broja iteracija. Modifikacija se sastoji u tome da se počev od neke iteracije pa nadalje, recimo poslije pete iteracije, više ne računaju inverzne matrice $(F'(x_6))^{-1}$, $(F'(x_7))^{-1}$, ..., nego da se one u računskom procesu zamijene sa (za njih približnom vrijednošću) $(F'(x_5))^{-1}$. Ipak, modifikovana Njutnova metoda ima slabiji stepen konvergencije (prvi) od Njutnove metode (drugi stepen).

Prelazimo na izlaganje Njutnove metode u opštem slučaju. Neka su X i Y dva Banahova prostora (realna), u njima se norma označava redom kao $\| \ \|_X$ i $\| \ \|_Y$, moguće je da se X i Y poklapaju. Vidimo da X označava i prostor originala i traženi korijen jednačine, ali se te dvije stvari ipak upadljivo razlikuju. Neka je F (nelinearni) operator, $F\colon X \to Y$. Razmotrimo jednačinu $F(x) = 0$. Prethodno uvedimo pojam izvoda (nelinearnog) preslikavanja F (u određenoj tački). Preslikavanje F djeluje iz jednog Banahovog prostora u drugi. Ovaj pojam predstavlja uopštenje pojma izvoda preslikavanja $F\colon R^n \to R^n$. Ovdje se uvodi tzv. pojam jakog izvoda ili izvoda po Frešeu. Za linearni operator $P\colon X \to Y$ (ako takav operator P postoji) kažemo da predstavlja izvod operatora F u tački x ako važi sljedeća jednakost:

$$\|F(x+\eta) - F(x) - P\eta\|_Y = o(\|\eta\|_X) \text{ kad } \|\eta\|_X \to 0;$$

odsad ćemo umjesto P pisati $F'(x)$. Dalje, konstrukcija Njutnove metode u ovom slučaju poklapa se sa maloprije urađenim izvođenjem te metode u slučaju $X = Y = R^n$, formule (4) i (5). Očito se pretpostavlja da $F'(x)$ postoji i da $(F'(x_k))^{-1}$ postoji.

Neka su za neke konstante $a > 0$, $a_1 < +\infty$ i $a_2 < +\infty$ ispunjena sljedeća dva uslova:

$$\|(F'(x))^{-1}\| \le a_1 \text{ za } x \in \Omega_a = \{x\colon \|x - X\|_X < a\} \tag{6}$$

(na lijevoj strani znaka $\le$ je norma operatora) i

$$\|F(U_1) - F(U_2) - F'(U_2)(U_1 - U_2)\|_Y \le a_2\|U_1 - U_2\|_X^2 \text{ za } U_1, U_2 \in \Omega_a. \tag{7}$$

Uvedimo oznake $c = a_1 a_2$ i $b = \min\left\{a, \dfrac{1}{c}\right\}$.

Teorema. Ako su ispunjeni uslovi (6) i (7) i ako $x_0 \in \Omega_b$ onda Njutnov iterativni proces (5) konvergira i važi sljedeća formula (za ocjenu greške):

$$\|x_k - X\|_X \le \frac{1}{c}\Big(c\|x_0 - X\|_X\Big)^{2^k}. \tag{8}$$

Dokaz. Vidi sliku 4. Prvo. Svi x_k pripadaju $\Omega_b = \{x\colon \|x - X\|_X < b\}$, ovo ćemo dokazati indukcijom. Za $k = 0$ ovo je ispunjeno po uslovu teoreme. Treba izvršiti indukcijski korak, prelazak sa k na $k+1$. Uzmimo da $x_k \in \Omega_b$. Kada u (7) uvrstimo $U_1 = X$, $U_2 = x_k$ tada imamo da je

$$\|F(X) - F(x_k) - F'(x_k)(X - x_k)\|_Y \le a_2\|x_k - X\|_X^2$$

slijedi, budući da je $F(X) = 0$, a za $F(x_k)$ v. formulu (4),

$$\|F'(x_k)(x_{k+1} - X)\|_Y \le a_2\|x_k - X\|_X^2$$

slijedi, pomoću (6),

$$\|x_{k+1} - X\|_X \le c\|x_k - X\|_X^2 \tag{9}$$

i dalje $< cb^2 = (cb)b \leq b$. Dobili smo da je $\|x_{k+1} - X\|_X < b$ odnosno da $x_{k+1} \in \Omega_b$. Dokazano je da svi elementi aproksimativnog niza $\{x_k\}$ pripadaju toj otvorenoj lopti. Ova okolnost je prikazana na slici 5. Drugo. Dokažimo formulu (8). Uvedimo oznaku $q_k = c\|x_k - X\|_X$. Iz (9) imamo da je $q_{k+1} \leq q_k^2$ za svako k. Recimo, $q_1 \leq q_0^2$, $q_2 \leq q_1^2$, $q_3 \leq q_2^2$ pa na primjer slijedi da je $q_3 \leq q_0^8$. Dakle, $q_k \leq q_0^{2^k}$, indukcijom. Dobili smo da je $c\|x_k - X\|_X \leq (c\|x_0 - X\|_X)^{2^k}$, time je (8) dokazano. I treće, posljednji dio dokaza teoreme. Posmatrajmo nejednakost (8). Budući da je $c\|x_0 - X\|_X < 1$ to desna strana te nejednakosti $\to 0$ kad $k \to \infty$. Zato i njena lijeva strana $\to 0$. Dobili smo da $\|x_k - X\|_X \to 0$ odnosno da $x_k \to X$. Dokaz je završen.

Slijede razne dopune.

1. Važno je istaći da se u prethodnjoj teoremi pretpostavlja da rješenje X jednačine $F(x) = 0$ postoji.

Lako se vidi da je rješenje jednačine $F(x) = 0$ u lopti Ω_b jedinstveno. Zaista, dopustimo da pored X postoji i neko drugo rješenje Y (važi $F(Y) = 0$). Uzmimo tada da je $x_0 = Y$. Po (5) izlazi onda da je i $x_1 = Y$, $x_2 = Y$, ... A u teoremi je dokazano da mora biti $\lim\limits_{k\to\infty} x_k = X$.

2. Uslov (6) govori da je prvi izvod odvojen od nule.

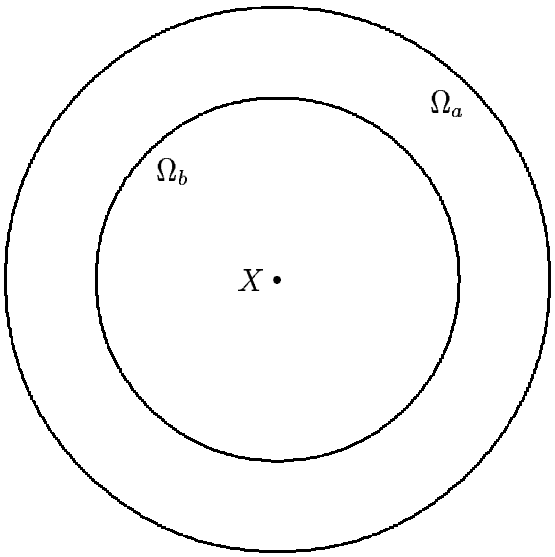
Uslov (7) odgovara uslovu: drugi izvod je odvojen od $+\infty$. Na račun ovoga, znamo da umjesto $F(U_1) - F(U_2) - F'(U_2)(U_1 - U_2)$, u slučaju $F\colon R \to R$, možemo pisati $F''(\alpha)(U_1 - U_2)^2/2$.

3. U slučaju da je $F\colon R^n \to R^n$, ako funkcije $F_1, \ldots, F_n$ (to su komponente od F) imaju ograničene druge parcijalne izvode onda je uslov (7) ispunjen, jer po Tejlorovoj formuli važi:

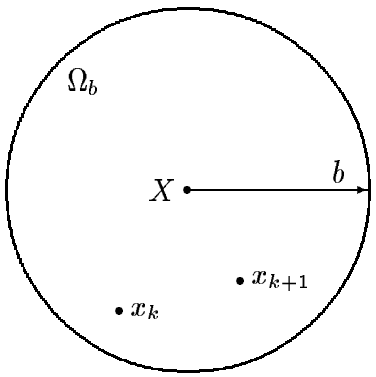
$$F_i(y) = F_i(x) + \sum_{j=1}^{n} \frac{\partial F_i(x_1, \ldots, x_n)}{\partial x_j}(y_j - x_j) + O(\|y - x\|^2),$$

$i = 1, \ldots, n$, $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$.

U slučaju da je $F\colon R^n \to R^n$, konstanta a_2 može da bude izražena preko $\dfrac{\partial^2 F_i}{\partial x_j \partial x_k}$. Prethodno se treba opredijeliti za određenu normu. Preciznije rečeno, korisno je napisati relacije koje efektivno izražavaju vezu između a_2 i drugih parcijalnih izvoda funkcija F_i. Kao i vezu između a_1 i $(F'(x))^{-1}$ odnosno $\dfrac{\partial F_i}{\partial x_j}$.



Prostor X, Slika 4,
X – tačno rješenje



Slika 5

5. NUMERIČKE METODE ZA RJEŠAVANJE KOŠIJEVOG ZADATKA ZA OBIČNE DIFERENCIJALNE JEDNAČINE

5.1. UVOD O KOŠIJEVOM ZADATKU I LEMA O DVA RJEŠENJA DIFERENCIJALNE JEDNAČINE

Da se opiše značaj numeričkih metoda za rješavanje običnih diferencijalnih jednačina dovoljne su sljedeće dvije rečenice. Većina procesa koji se proučavaju u fizici i uopšte prirodnim naukama opisuje se parcijalnim diferencijalnim jednačinama, a u jednostavnijim slučajevima – običnim diferencijalnim jednačinama. Veoma su rijetki slučajevi kada može da bude određeno tačno ili egzaktno (ili analitičko) rješenje diferencijalne jednačine.

U ovoj glavi razmatraju se numeričke metode za rješavanje početnog (Košijevog) zadatka za obične diferencijalne jednačine, pa formulišimo zadatak. Dati su diferencijalna jednačina $y' = f(x, y)$ i početni uslov $y(x_0) = y_0$, nepoznata funkcija ili rješenje zadatka $y = y(x)$ razmatra se na odsječku x–ose $[a, b] = [x_0, x_0 + X]$. Prilikom numeričkog rješavanja, pretpostavlja se da rješenje (da tačno rješenje) postoji, da je ono jedinstveno i da je ono jedna dovoljno glatka funkcija. Ako $f \in C^p$ onda $y \in C^{p+1}$.

Kroz svaku tačku oblasti prolazi samo po jedna integralna kriva diferencijalne jednačine.

Prilikom numeričkog rješavanja postavljenog zadatka, mi postavimo po x–osi mrežu čvorova $a = x_0 < x_1 < x_2 < \ldots < x_n = x_0 + X = b$. Označimo sa y_i približnu vrijednost za broj $y(x_i)$ tj. za vrijednost tačnog rješenja $y = y(x)$ u čvoru $x = x_i$. Brojne vrijednosti $\{y_i\}_{i=0}^n$ čine naš numerički odgovor, a razmatra se naravno i greška $R_i = y(x_i) - y_i$.

Numeričke metode za Košijev zadatak za o. d. j. dijele se u dvije klase i to: a) jednokoračne i b) višekoračne (k–koračne) ili diferencne. U slučaju a), približna vrijednost y_i određuje se na osnovu približne vrijednosti y_{i-1} koja je u datom trenutku već određena. U slučaju b), y_i se određuje po nekoliko ranijih približnih vrijednosti, recimo da se određuje po y_{i-4}, y_{i-3}, y_{i-2} i y_{i-1} (tada je $k = 4$). Primjer numeričke metode oblika a) jeste metoda Runge–Kuta. Primjer numeričke metode oblika b) jeste Adamsova metoda. Obično je mreža ekvidistantna tj. $x_i = x_0 + ih$, $nh = b - a = X$.

Od numeričke metode se očekuje da njena greška teži ka nuli kada mreža čvorova na $[a, b]$ postaje sve gušća, odnosno kada broj čvorova neograničeno raste. Takođe se očekuje da greška teži ka nuli određenim tempom, odnosno teži ka nuli što brže, kada $n \to \infty$, odnosno kada $h \to 0$. Razmotrimo jednu fiksiranu numeričku metodu. Neka je mreža čvorova ekvidistantna sa korakom $h > 0$. Definicija 1. Neka $x \in [a, b] = [x_0, x_0 + X]$. Neka $z_h(x)$ označava približnu vrijednost za $y(x)$ dobijenu sa korakom h i neka bude $R_h(x) = y(x) - z_h(x)$. Smatramo da je $nh = b - a$ za neki cio broj $n \geq 1$. Za numeričku metodu se kaže da konvergira u tački x ako važi $\lim_{h \to 0} R_h(x) = 0$. Definicija 2. Za numeričku metodu se kaže da konvergira na odsječku $[a, b]$ ako ona konvergira u svakoj tački tog odsječka. Definicija 3. Za numeričku metodu se kaže da ima red konvergencije s u tački x ako važi $|R_h(x)| = O(h^s)$ kad $h \to 0$. Definicija 4. Za numeričku metodu se kaže da ima red konvergencije s na odsječku $[a, b]$ ako ona ima red konvergencije s ravnomjerno u svim tačkama tog odsječka.

Završavajući uvod, metode koje ćemo raditi uopštavaju se neposredno ili vrlo lako na slučaj implicitno date jednačine, na slučaj sistema jednačina prvog reda, kao i na slučaj jednačine višeg reda.

Na redu je jedna lema iz teorije običnih diferencijalnih jednačina.

Lema. Neka je $f(x, y)$ neprekidna funkcija i neka je neprekidno diferencijabilna po promjenljivoj y. Neka su $Y_1(x)$ i $Y_2(x)$ dva rješenja diferencijalne jednačine $y' = f(x, y)$. Tada

važi

$$Y_2(\beta) - Y_1(\beta) = (Y_2(\alpha) - Y_1(\alpha)) \exp\left\{\int_\alpha^\beta f'_y(x, \overline{y}(x))dx\right\}, \qquad (*)$$

gdje je $\overline{y}(x)$ neki broj između $Y_1(x)$ i $Y_2(x)$.

Dokaz. Imamo $Y_1' = f(x, Y_1)$ i $Y_2' = f(x, Y_2)$, oduzimanjem $(Y_2 - Y_1)' = f(x, Y_2) - f(x, Y_1)$. Na $f(x, Y_2) - f(x, Y_1)$ primijenimo Lagranžovu teoremu (teoremu o konačnim priraštajima), pa izlazi

$$f(x, Y_2) - f(x, Y_1) = f'_y(x, \overline{y}) \cdot (Y_2 - Y_1),$$

gdje je $\overline{y}$ neki broj između $Y_1(x)$ i $Y_2(x)$. Uradimo ovo za razne x, pa tako imamo

$$(Y_2 - Y_1)' = f'_y(x, \overline{y}(x)) \cdot (Y_2 - Y_1).$$

Posljednje predstavlja linearnu diferencijalnu jednačinu oblika $y' = P(x)y$ po nepoznatoj funkciji $Y_2 - Y_1$. Rješavanjem te jednačine dobija se $(*)$.

Zapaziti da mi ne možemo ništa reći o neprekidnosti ili eventualnoj glatkosti funkcije $\overline{y}(x)$. To nam nije ni potrebno, jer je funkcija

$$f'_y(x, \overline{y}(x)) = \frac{f(x, Y_2(x)) - f(x, Y_1(x))}{Y_2(x) - Y_1(x)}$$

neprekidna. Zaista, u brojiocu i imeniocu su neprekidne funkcije. A još, imenilac je različit od nule, jer se dva rješenja $Y_1(x)$ i $Y_2(x)$ jedne te iste jednačine ne sijeku (jedinstvenost rješenja Košijevog zadatka). Zato su sve gore sprovedene transformacije – ispravne. Lema je dokazana.

Vodeći primjer za ovu lemu: razmotrimo diferencijalnu jednačinu $y' = ay$. Ovdje je očito $f(x, y) = ay$ i $f'_y(x, y) = a$. Opšte rješenje glasi $y(x) = Ce^{ax}$. Razmotrimo dva partikularna rješenja $Y_1(x) = C_1 e^{ax}$ i $Y_2(x) = C_2 e^{ax}$. Mi upoređujemo rastojanje po visini između ova dva rješenja u dvije tačke na x–osi. Rastojanje u tački $x = \beta$ može da bude znatno veće od rastojanja u ranijoj tački $x = \alpha$:

$$Y_2(\beta) - Y_1(\beta) = (Y_2(\alpha) - Y_1(\alpha))e^{a(\beta-\alpha)}.$$

Dakle, prilikom napredovanja po x–osi dolazi do dodatnog razilaženja jednog i drugog rješenja. Zaključak: na odnos između dva rastojanja po visini suštinski utiče veličina $f'_y(x, y)$.

Mi ćemo ubuduće formulu $(*)$ koristiti obično u nešto obrađenom obliku, kako slijedi. Treba pretpostaviti da funkcija $f = f(x, y)$ zadovoljava Lipšicov uslov po drugoj promjenljivoj y, sa konstantom L, tj. da važi $|f(x, y_2) - f(x, y_1)| \leq L \cdot |y_2 - y_1|$. Mi ustvari pretpostavljamo da važi $|\frac{\partial f}{\partial y}| \leq L$, što predstavlja nešto jači uslov, jer je ovako jednostavnije pisanje. Tada

$$\exp\left\{\int_\alpha^\beta f'_y(x, \overline{y}(x))dx\right\} \leq \exp\left|\int_\alpha^\beta f'_y(x, \overline{y}(x))dx\right| \leq$$

$$\exp\left\{\int_\alpha^\beta |f'_y(x, \overline{y}(x))|dx\right\} \leq \exp\left\{\int_\alpha^\beta L dx\right\} = \exp\{L(\beta - \alpha)\}$$

i

$$|Y_2(\beta) - Y_1(\beta)| \leq |Y_2(\alpha) - Y_1(\alpha)| \cdot e^{L(\beta-\alpha)} \qquad (**)$$

5.2. OJLEROVA METODA I DRUGI PRIMJERI

Ojlerova metoda predstavlja najjednostavniji primjer jednokoračne metode.

Navedimo na početku jednu poznatu činjenicu iz teorije običnih diferencijalnih jednačina. Razmotrimo diferencijalnu jednačinu $y' = f(x,y)$. U dijelu ravni R^2 gdje je f definisana nacrtajmo tzv. polje pravaca (polje smjerova). Drugim riječima, u svakoj tački (x,y) tog dijela ravni nacrtajmo pravu koja prolazi kroz tu tačku a čiji je koeficijent pravca jednak $f(x,y)$. S druge strane, razmotrimo funkciju $y = y(x)$ koja je rješenje jednačine $y' = f(x,y)$. Grafik funkcije $y = y(x)$ dodiruje polje smjerova, tj. u zajedničkoj tački prave i grafika – prava je tangenta grafika. Dakle, diferencijalna jednačina određuje tangentu grafika. Iz ove činjenice se već može konstruisati najjednostavnija numerička metoda za Košijev zadatak – Ojlerova metoda, što slijedi u nastavku. Ilustrujmo na konkretnom primjeru Košijevog zadatka: $y' = x^2 + y^2$, $y(1) = 2$. Imamo da je $f(x_0, y_0) = x_0^2 + y_0^2 = 1 + 4 = 5$. Odavde je $y(x_0 + h) = y(1 + h) \approx y_0 + 5h = 2 + 5h$. Recimo, $y(1{,}01) \approx 2{,}05$.

Postavka zadatka. Razmotrimo Košijev (početni) zadatak $y' = f(x,y)$, $y(x_0) = y_0$ na odsječku $x_0 \le x \le x_0 + X$. Označimo njegovo tačno rješenje kao $y = y(x)$. Treba konstruisati numeričku metodu za dobijanje približnih vrijednosti tačnog rješenja $y = y(x)$. Postavimo po odsječku $[x_0, x_0+X]$ ravnomjernu mrežu čvorova čiji je korak h, gdje je $nh = X$. Dakle, stavimo $x_k = x_0 + kh$ za $k = \overline{0,n}$. Neka y_k označava približnu vrijednost za broj $y(x_k)$ ili svejedno neka y_k bude aproksimacija za vrijednost tačnog rješenja postavljenog Košijevog zadatka u čvoru $x = x_k = x_0 + kh$, gdje je $k = \overline{0,n}$. Veličine $\{y_k\}_{k=0}^n$ biće efektivno određene i one će i predstavljati numerički odgovor. Analiziraćemo i grešku $R_k = y(x_k) - y_k$, za $k = \overline{0,n}$.

Kao prvi primjer numeričke metode za K. z. za o. d. j. navodimo Ojlerovu metodu. Ova metoda ima slab stepen konvergencije $O(h)$, pa se zato u praksi ne koristi. Međutim, ova metoda je tehnički jednostavna, pa ćemo lako vidjeti neka svojstva koja su zajednička za sve metode iz klase metoda Runge–Kuta.

Kako procijeniti $y(x_1)$? Napišimo Tejlorovu formulu za tačno rješenje $y = y(x)$ vršeći razvoj oko tačke $x = x_0$. Tako

$$y(x_0 + h) = y(x_0) + hy'(x_0) + \frac{1}{2}h^2 y''(\alpha),$$

gdje je $x_0 < \alpha < x_0 + h$. Odavde je

$$y(x_0 + h) \approx y(x_0) + hy'(x_0) \text{ tj.}$$

$$y(x_1) \approx y(x_0) + hy'(x_0) \text{ ili svejedno } y_1 = y_0 + hy'(x_0).$$

U vezi date d. j.

$$y_1 = y_0 + hf(x_0, y_0).$$

Kako procijeniti $y(x_2)$? Na sličan način, mi stavljamo da je

$$y_2 = y_1 + hf(x_1, y_1).$$

Zapaziti da je broj y_1 približan po jednom osnovu: odbačen je sabirak $\frac{1}{2}h^2y''(\alpha)$. Dok je broj y_2 približan još po jednom osnovu: za njegovo računanje oslanjamo se na približan broj y_1 kojim raspolažemo, a ne oslanjamo se na tačan broj $y(x_1)$ kojim ne raspolažemo. Slično su i brojevi $y_3, y_4, \ldots$ približni po dva osnova.

Kako procijeniti $y(x_3), y(x_4), \ldots$? Postupa se na isti način. Dakle, šema za računanje koja i definiše Ojlerovu metodu izražava se sljedećom formulom:

$$y_k = y_{k-1} + hf(x_{k-1}, y_{k-1}) \text{ za } k = \overline{1,n}.$$

Znamo da šema za računanje obuhvata formule pomoću kojih se onda može sastaviti program za računar (za saznavanje približnih vrijednosti y_k).

Mali primjer za Ojlerovu metodu: $y' = x + y$, $y(0) = 1$. Izaberimo $h = 0{,}1$ i n=4, tako da je $x_1 = 0{,}1$ $x_2 = 0{,}2$ $x_3 = 0{,}3$ $x_4 = 0{,}4$ Tada se dobija numerički rezultat $y_1 = 1{,}1$ $y_2 = 1{,}22$ $y_3 = 1{,}362$ $y_4 = 1{,}5282$ S druge strane, odgovarajuće tačne vrijednosti glase $y(x_1) = 1{,}11034$ $y(x_2) = 1{,}24281$ $y(x_3) = 1{,}39972$ $y(x_4) = 1{,}58365$ budući da je analitičko rješenje $y(x) = 2e^x - x - 1$. Tako da pojedinačne greške $R_k = y(x_k) - y_k$ iznose redom: $R_1 = 0{,}01034$ $R_2 = 0{,}02281$ $R_3 = 0{,}03772$ $R_4 = 0.05545$

Prelazimo na ocjenu greške. Prvo uvodimo pojam lokalne greške i vršimo njenu ocjenu. Lokalna greška ili greška na koraku definiše se u slučaju Ojlerove metode relacijom:

$$\rho = y(x+h) - [y(x) + hy'(x)] \text{ tj. } \rho = y(x+h) - [y(x) + hf(x, y(x))].$$

Na primjer, greška metode na prvom koraku jednaka je

$$\rho_1 = y(x_1) - y_1 = y(x_1) - [y_0 + hy'(x_0)] = y(x_1) - [y_0 + hf(x_0, y_0)].$$

Zapaziti da lokalna greška obuhvata samo prvi osnov po kome je broj y_k ($k \geq 2$) približan, od dva osnova koji postoje.

Imamo:

$$y(x+h) = y(x) + hy'(x) + \frac{1}{2}h^2 y''(\beta), \text{ gdje je } x < \beta < x+h \Rightarrow \rho = \frac{1}{2}h^2 y''(\beta).$$

Uzmimo da je drugi izvod rješenja K. z. ograničen na odsječku $[x_0, x_0+X]$ tj. da je $|\frac{1}{2}y''(\beta)| \leq C$. Tada je očito $|\rho| \leq Ch^2$. Lokalna greška Ojlerove metode je reda h^2.

Prelazimo na ocjenu greške (globalne greške). Uvedimo potrebne oznake. Kao i dosad, y_k su približne vrijednosti za čvorove $x_k = x_0 + kh$ koje je računar saopštio. Tako da greška u pojedinom čvoru iznosi naravno $R_k = y(x_k) - y_k$ za $k = \overline{0,n}$. Kao i dosad, $y = y(x)$ je tačno rješenje postavljenog K. z. Ocijenićemo grešku $R_n = y(x_n) - y_n$ u posljednjem čvoru $x_n = x_0 + nh = x_0 + X$, a slično izvođenje može da bude sprovedeno za bilo koji čvor.

Uvedimo u razmatranje pomoćne funkcije. Neka funkcija $y_k = y_k(x)$ zadovoljava postavljenu d. j. $y' = f(x,y)$ i neka zadovoljava jednakost $y_k(x_k) = y_k$ kao svoj početni uslov. Ovdje je $k = \overline{0,n}$. Vidi se da je $y(x) = y_0(x)$. Kao što je maločas rađeno, lokalne greške definisane su relacijom $\rho_k = y_{k-1}(x_k) - y_k = y_{k-1}(x_k) - y_k(x_k)$. Imamo:

$$R_n = y(x_n) - y_n(x_n) = y_0(x_n) - y_n(x_n) = \sum_{k=1}^{n} (y_{k-1}(x_n) - y_k(x_n))$$

$$|R_n| \leq \sum_{k=1}^{n} |y_{k-1}(x_n) - y_k(x_n)| \leq \text{ po } (**)$$

$$\sum_{k=1}^{n} |y_{k-1}(x_k) - y_k(x_k)| e^{L(x_n - x_k)} = \sum_{k=1}^{n} |\rho_k| e^{L(x_n - x_k)}$$

Sada se mijenja definicija konstante C, ona se sada opštije definiše (i zato se ona povećava). Uzmimo da je $|\frac{1}{2}y_k''(x)| \leq C$, $\forall x \in [x_0, x_0 + X]$ i $\forall k = \overline{0,n}$. Više od toga (još opštije), stavimo da je C broj takav da važi $|\frac{1}{2}y''(x)| \leq C$, $\forall x \in [x_0, x_0 + X]$, gdje je $y = y(x)$ bilo koje rješenje d. j. $y' = f(x, y)$. Nastavljamo:

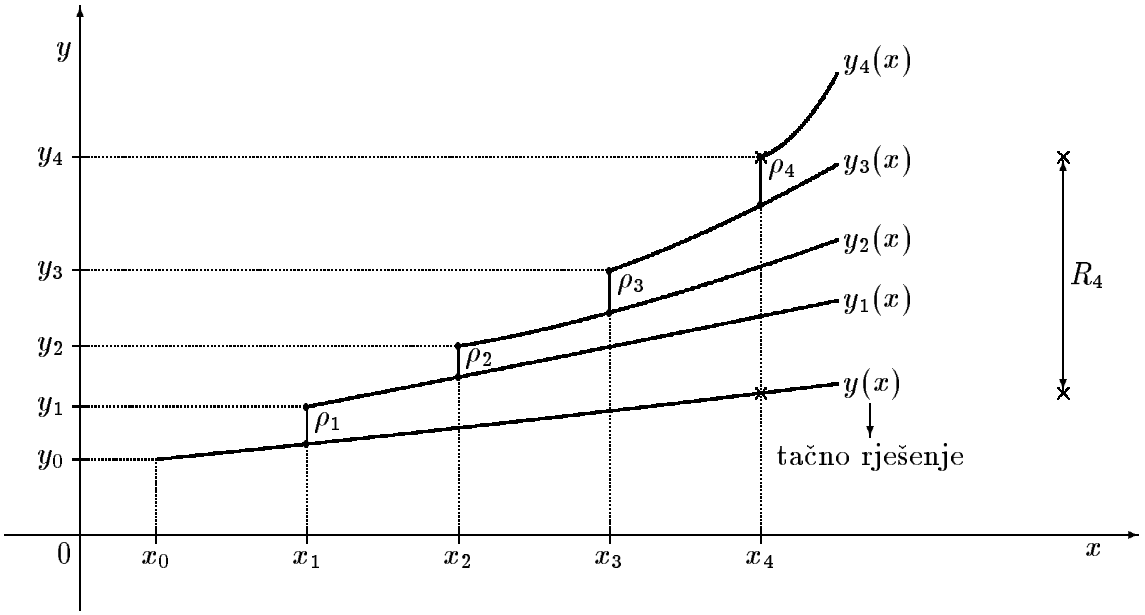
$$|R_n| \leq \sum_{k=1}^{n} Ch^2 e^{L(x_n - x_k)}$$

$$|R_n| \leq \sum_{k=1}^{n} Ch^2 e^{LX} = nCh^2 e^{LX} = CXhe^{LX} = const \cdot h$$

Ili $R_n = O(h)$. Dakle, greška Ojlerove metode je reda h. Ponovimo, pod uslovom da je $|\frac{\partial f}{\partial y}| \leq L$ za neko $L > 0$ i da rješenja $y(x)$ d. j. $y' = f(x, y)$ imaju ograničene druge izvode $y''(x)$. Ili nestrogo zapisano: pod uslovom da je 1 $L < \infty$ i 2 $C < \infty$.

Slika se odnosi na ocjenu greške (globalne greške), $R_4 = y(x_4) - y_4(x_4)$.

Završeno o Ojlerovoj metodi.



Napomena. Ne važi nejednakost $|R_n| \leq |\rho_1| + |\rho_2| + \ldots + |\rho_n|$ u opštem slučaju. Drugim riječima, R_n je uopšte uzev znatno veće od napisanog zbira. Znamo da u slučaju određenog integrala zbir lokalnih greški jeste jednak globalnoj greški (ukupnoj greški).

Napomena. Ako se data jednačina $y' = f(x, y)$ diferencira po x onda

$$y'' = f_x'(x, y) + f_y'(x, y)y' = f_x'(x, y) + f_y'(x, y)f(x, y),$$

gdje na desnoj strani učestvuju sve poznate veličine, učestvuju samo date veličine i one koje se po datima mogu izračunati. Tako da će uslovi $|f_y'| < \infty$ i $|y''| < \infty$ tj. $L < \infty$ i $C < \infty$ biti ispunjeni ako su tri funkcije f, f_x' i f_y' ograničene na skupu $\Omega = [x_0, x_0 + X] \times R$.

Napomena. Za male h, dovoljno je da bude $|\frac{1}{2}z''(x)| \leq C$, gdje je $z = z(x)$ bilo koje rješenje postavljene d. j. koje je "blisko" tačnom rješenju razmatranog K. z.

Na redu su primjeri jednokoračnih metoda (šema)

Primjer 1: jedna implicitna šema

Napišimo identitet

$$y(x+h) = y(x) + \int_0^h y'(x+t)dt. \tag{1}$$

Ako za računanje integrala u (1) upotrebimo osnovnu trapeznu formulu $\int_0^h F(t)dt = \frac{h}{2}(F(0) + F(h)) + O(h^3)$ onda

$$y(x+h) = y(x) + \frac{h}{2}\Big(y'(x) + y'(x+h)\Big) + O(h^3) \text{ ili}$$

$$y(x+h) = y(x) + \frac{h}{2}\Big(f(x,y(x)) + f(x+h,y(x+h))\Big) + O(h^3). \tag{2}$$

Tako se dobija numerička formula ili šema za računanje

$$y_{j+1} = y_j + \frac{h}{2}\Big(f(x_j,y_j) + f(x_{j+1},y_{j+1})\Big) \text{ za } j \geq 0. \tag{3}$$

Jednakost (3) ustvari predstavlja jednu (nelinearnu) jednačinu po nepoznatoj y_{j+1}. Zato se za ovu šemu kaže da je implicitna (da je u neriješenom obliku). Ako se približna vrijednost koja se računa (oblika y_{j+1}) može saznati neposrednim računom onda se za šemu ili metodu kaže da je eksplicitna ili da je u riješenom obliku.

Može se tvrditi da postoji jedinstveno rješenje jednačine (3) jedino ako je h dovoljno malo (jedino za $h \leq h_0$).

Lokalna greška navedene metode iznosi $\rho = O(h^3)$, a njena greška je jednaka $O(h^2)$.

Primjer 2: poboljšana ili modifikovana Ojlerova metoda

Pokušajmo da u prethodnoj metodi nešto promijenimo, da bi se implicitnost otklonila. Ako se na desnoj strani formule (2) broj $y(x+h)$ zamijeni nekim brojem y^* koji od $y(x+h)$ odstupa za $O(h^2)$ onda se ta desna strana izmijeni za $O(h^3)$. Kako je ostatak formule (2) bio $O(h^3)$ i kako se ovom zamjenom uvećava za $O(h^3)$, to će i lokalna greška metode sa y^* biti $O(h^3)$.

Kao y^* može da posluži približna vrijednost za $y(x+h)$ koju daje Ojlerova metoda.

Odavde, šema za računanje modifikovane Ojlerove metode glasi:

$$\begin{cases} y^*_{j+1} = y_j + hf(x_j,y_j) \\ y_{j+1} = y_j + \dfrac{h}{2}\left(f(x_j,y_j) + f(x_{j+1},y^*_{j+1})\right), \quad j = \overline{0,n-1} \end{cases}$$

Lokalna greška modifikovane Ojlerove metode iznosi $\rho = O(h^3)$, a njena greška je jednaka $O(h^2)$.

Primjer 3

Za računanje integrala u (1) upotrebimo formulu pravougaonika, čime se dobija da je $y(x+h) \approx y(x) + hf(x+\frac{h}{2}, y(x+\frac{h}{2}))$. Kao približna vrijednost za $y(x+\frac{h}{2})$ može da posluži vrijednost izračunata po Ojlerovoj formuli sa korakom $\frac{h}{2}$. Tako imamo metodu čija je šema za računanje definisana sljedećim parom jednakosti:

$$\begin{cases} y_{j+1/2} = y_j + \dfrac{h}{2}f(x_j,y_j) \\ y_{j+1} = y_j + hf(x_{j+1/2},y_{j+1/2}) \end{cases}$$

Lako se vidi da lokalna greška ove metode iznosi $\rho = O(h^3)$ i da je njena greška jednaka $O(h^2)$.

5.3. OPŠTI SLUČAJ EKSPLICITNE METODE TIPA RUNGE–KUTA

U ovoj sekciji se definiše opšti oblik metode tipa Runge–Kuta, eksplicitne. Zatim se izlaže opšti postupak za izvođenje numeričke metode tipa Runge–Kuta, za taj postupak se kaže da predstavlja tzv. metodu neodređenih koeficijenata. Na kraju se analizira ρ – lokalna greška metode.

U prethodnom naslovu smo vidjeli da postoje razne ideje kako da se iskombinuje priraštaj rješenja kada x napreduje za h. Kombinuje se tako da lokalna greška ρ bude manja. Možda je bolje da se umjesto ideja (dosjetki) razvije aparat opšteg tipa, za kombinacije. Tako radimo u ovom naslovu.

Metoda Runge–Kuta služi za rješavanje Košijevog zadatka $y' = f(x, y)$, $y(x) = y$. Radi se o jednokoračnoj metodi. Dovoljno je definisati kako se vrši jedan korak, tj. definisati kako se iz tačke x prelazi u tačku $x + h$. Dakle, postavlja se pitanje: kako naći približnu vrijednost u oznaci $z(h)$ za tačnu vrijednost $y(x + h)$. Naravno da se lokalna greška metode ili greška metode na koraku definiše kao $\rho = \rho(h) = y(x + h) - z(h)$. Šablon oznaka:

po x–osi	tačno	približno	greška
x	$y = y(x)$	$y = y(x)$	0
$x + h$	$y(x + h)$	$z(h)$	$\rho(h)$

Vidimo da je Košijev uslov označen kao $y(x) = y$, umjesto uobičajenog zapisivanja $y(x_0) = y_0$.

Evo opšte definicije metode ove klase. U procesu računanja fiksirani su izvjesni brojevi $\alpha_2, \ldots, \alpha_q, p_1, \ldots, p_q, \beta_{ij}$ $(1 \le j \le i-1, 2 \le i \le q)$. Redom se računaju veličine $k_1(h), \ldots, k_q(h)$, Δy i $z(h)$:

$$\begin{cases} k_1(h) = hf(x, y), \;\; k_2(h) = hf(x + \alpha_2 h, y + \beta_{21} k_1(h)), \;\; \ldots, \\ k_q(h) = hf(x + \alpha_q h, y + \beta_{q1} k_1(h) + \ldots + \beta_{q,q-1} k_{q-1}(h)), \\ \Delta y = \sum_{i=1}^{q} p_i k_i(h), \;\; z(h) = y + \Delta y \end{cases}$$

i zatim se stavlja da je $y(x + h) \approx z(h)$. Lako se zapaža da sve veličine $k_i(h)$ imaju smisao priraštaja. Zapaža se da mora da bude $\sum_{i=1}^{q} p_i = 1$, ako se želi postići makar kakva aproksimacija. Kao što je već rečeno, greška metode na koraku definiše se kao $\rho(h) = y(x+h) - z(h) = y(x + h) - [y(x) + \Delta y]$. Vidimo da $\rho(h)$ zavisi od tačke (x, y), a zavisi naravno i od funkcije f.

Postavlja se pitanje izbora zasad neodređenih (slobodnih) brojnih vrijednosti α_i, p_i i β_{ij}. Te brojne vrijednosti biće izabrane tako da postane $\rho(0) = \rho'(0) = \ldots = \rho^{(s)}(0) = 0$. Jasno je da je bolje što je broj s veći. To i jeste kriterijum za određivanje parametara α_i, p_i i β_{ij}. Uzima se da ovdje broj s ne može da bude povećan, tj. da za neku glatku funkciju $f(x, y) = f_0(x, y)$ važi $\rho^{(s+1)}(0) \neq 0$. Za s se kaže da je red greške metode.

Možemo napisati Tejlorovu formulu:

$$\rho(h) = \sum_{i=0}^{s} \frac{\rho^{(i)}(0)}{i!} h^i + \frac{\rho^{(s+1)}(\theta h)}{(s+1)!} h^{s+1} \;\text{ ili }\; \rho(h) = \frac{\rho^{(s+1)}(\theta h)}{(s+1)!} h^{s+1},$$

gdje je $0 < \theta < 1$.

Ispitajmo za prvih nekoliko q.

$\underline{q=1}$ $k_1(h)=hf(x,y)$, $\Delta y=p_1k_1(h)$, $z(h)=y+\Delta y$, $y(x+h)\approx z(h)$

$$\rho(h)=y(x+h)-y(x)-p_1hf(x,y)$$

$$\rho'(h)=y'(x+h)-p_1f(x,y)$$

$$\rho'(0)=y'(x)-p_1f(x,y)=f(x,y)-p_1f(x,y)=(1-p_1)f(x,y)$$

Ako želimo da postignemo da bude $\rho'(0)=0$ onda mora biti $p_1=1$. Ovim je jedna konkretna metoda tipa Runge–Kuta definisana. Vidimo da smo ustvari dobili Ojlerovu metodu.

$\underline{q=2}$ Opet ćemo izračunati $\rho'(h),\rho''(h),\ldots$ i opet ćemo nastojati da postignemo $\rho'(0)=0$, $\rho''(0)=0$, ...

Po opštem šablonu: $k_1(h)=hf(x,y)$, $k_2(h)=hf(x+\alpha_2h,y+\beta_{21}k_1(h))$, $\Delta y=p_1k_1(h)+p_2k_2(h)$, $z(h)=y+\Delta y$, $y(x+h)\approx z(h)$

$$\rho(h)=y(x+h)-y(x)-p_1k_1(h)-p_2k_2(h)=y(x+h)-y(x)-p_1hf(x,y)-p_2hf(\overline{x},\overline{y})$$

skraćenice: $x+\alpha_2h=\overline{x}$ i $y+\beta_{21}k_1(h)=\overline{y}$

$$\rho'(h)=y'(x+h)-p_1f(x,y)-p_2f(\overline{x},\overline{y})-p_2h\Big(\alpha_2f'_x(\overline{x},\overline{y})+\beta_{21}f'_y(\overline{x},\overline{y})f(x,y)\Big)$$

$$\rho'(0)=(1-p_1-p_2)f(x,y)$$

$$\begin{cases} y'=f(x,y)\\ y''=f'_x+f'_y\cdot y'=f'_x+f'_yf\\ y'''=f''_{xx}+2f''_{xy}f+f''_{yy}f^2+f'_yy''= \text{ izraziti } y'' \text{ po prethodnoj formuli}\\ y^{IV}=\ldots \text{ itd.}\end{cases}$$

$$\rho''(h)=y''(x+h)-2p_2\Big[\alpha_2f'_x(\overline{x},\overline{y})+\beta_{21}f'_y(\overline{x},\overline{y})f(x,y)\Big]-p_2h\Big[\alpha_2^2f''_{xx}(\overline{x},\overline{y})+$$

$$2\alpha_2\beta_{21}f''_{xy}(\overline{x},\overline{y})f(x,y)+\beta_{21}^2f''_{yy}(\overline{x},\overline{y})(f(x,y))^2\Big]$$

$$\rho''(0)=(1-2p_2\alpha_2)f'_x(x,y)+(1-2p_2\beta_{21})f'_y(x,y)f(x,y)$$

$$\rho'''(h)=\ldots,\quad \rho'''(0)=\ldots$$

Ako želimo da bude $\rho'(0)=0$ onda mora biti $\underline{1-p_1-p_2=0}$. Ako želimo da bude $\rho''(0)=0$ onda mora biti $\underline{1-2p_2\alpha_2=0}$ i $\underline{1-2p_2\beta_{21}=0}$. Vidimo da imamo tri uslova, a četiri parametra α_2, p_1, p_2 i β_{21}. Jedan parametar se bira proizvoljno. Navedimo dvije mogućnosti.

(1) Izaberimo $p_1=\frac{1}{2}$. Onda izlazi $p_2=\frac{1}{2}$, $\alpha_2=1$, $\beta_{21}=1$. Vidimo da je ovo modifikovana Ojlerova metoda.

Mi pišemo: $k_1=hf(x_j,y_j)$, $k_2=hf(x_j+h,y_j+k_1)$, $y_{j+1}=y_j+\frac{1}{2}(k_1+k_2)$ (RK2 metoda).

(2) Ako se izabere $p_1=0$ onda izlazi $p_2=1$, $\alpha_2=\frac{1}{2}$, $\beta_{21}=\frac{1}{2}$. Vidimo da je ovo posljednja metoda iz prethodne sekcije 5.2. gdje piše Primjer 3.

Kada kažemo $\rho'(0)=\rho''(0)=0$ onda mislimo: $\rho'(0)=\rho''(0)=0$ za svaku d. j. (za svaku funkciju f), a takođe i za bilo koju tačku (x,y) (ima ulogu tačke (x_0,y_0)).

Razmotrimo diferencijalnu jednačinu $y' = y$. Neposredni račun pokazuje da je $\rho'''(0) = y$, nezavisno od izbora parametara α_2, p_1, p_2 i β_{21}. To znači da se u slučaju $q = 2$ ne može pronaći formula Runge–Kuta čiji bi stepen tačnosti bio $s = 3$.

$\underline{q = 3}$ Razmatra se slično dosadašnjem. Može se dokazati da se najviše može postići da bude $s = 3$. Drugim riječima, ne postoji šema za koju bi bilo $s = 4$ (kada je $q = 3$). Evo jedne formule čiji je red tačnosti $s = 3$:

$$\begin{cases} k_1 = hf(x,y), \;\; k_2 = hf\left(x + \dfrac{h}{2}, y + \dfrac{k_1}{2}\right), \;\; k_3 = hf(x+h, y - k_1 + 2k_2), \\ \Delta y = \dfrac{1}{6}\Big(k_1 + 4k_2 + k_3\Big), \;\; z(h) = y + \Delta y \end{cases}$$

stavlja se $y(x + h) \approx z(h)$.

$\underline{q = 4}$ Najviše se može postići da bude $s = 4$. Evo jedne šeme čiji je red tačnosti $s = 4$:

$$\begin{cases} k_1 = hf(x,y), \;\; k_2 = hf\left(x + \dfrac{h}{2}, y + \dfrac{k_1}{2}\right), \;\; k_3 = hf\left(x + \dfrac{h}{2}, y + \dfrac{k_2}{2}\right), \\ k_4 = hf(x+h, y+k_3), \;\; \Delta y = \dfrac{1}{6}\Big(k_1 + 2k_2 + 2k_3 + k_4\Big), \;\; z(h) = y + \Delta y \end{cases}$$

Predložena šema se često koristi u praksi. Ponekad se za nju prosto kaže da je "metoda Runge–Kuta IV reda". Lako se mogu napisati kompletne formule, odnosno formule pomoću kojih se onda sastavi program za računar, kako slijedi. Razmotrimo Košijev zadatak $y' = f(x,y)$, $x_0 \le x \le x_0 + X$, $y(x_0) = y_0$. Neka je $n \ge 1$ i $h = X/n$. Koriste se obične oznake za čvorove $x_j = x_0 + jh$ i približne vrijednosti u čvorovima y_j: RK4:

$$\begin{aligned} &k_1 = hf(x_j, y_j), \;\; k_2 = hf\left(x_j + \frac{h}{2}, y_j + \frac{k_1}{2}\right), \;\; k_3 = hf\left(x_j + \frac{h}{2}, y_j + \frac{k_2}{2}\right), \\ &k_4 = hf(x_j + h, y_j + k_3), \;\; y_{j+1} = y_j + \frac{1}{6}\Big(k_1 + 2k_2 + 2k_3 + k_4\Big), \;\; j = \overline{0, n-1} \end{aligned}$$

Završeno je ispitivanje za razne q. Napominje se da se u praksi koriste samo metode u kojima je $q \le 6$.

Sada ćemo navesti uslove pod kojima je funkcija greške $\rho = \rho(h)$ dovoljno glatka i ograničena na skupu $\Omega = [x_0, x_0 + X] \times R \subset R^2$. Pretpostavimo da $f \in C^s(\Omega)$ i $f \in B^s(\Omega)$; c – continuous – neprekidan, b – bounded – ograničen. Zapis $f \in B^s(\Omega)$ znači da su funkcija f i svi njeni parcijalni izvodi do reda s uključeno ograničeni na skupu Ω. Glatkost funkcije f prenosi se na funkcije $k_1(h), \ldots, k_q(h)$, pa se prenosi i na funkciju $\rho = \rho(h)$. Neposredno se vidi da je funkcija $\rho = \rho(h)$ neprekidno diferencijabilna $s + 1$ puta. Ograničenost funkcije f i njenih parcijalnih izvoda prenosi se na funkciju $\rho = \rho(h)$. Tako da je $\rho(h) = O(h^{s+1})$, odnosno $|\rho(h)| \le Ch^{s+1}$, gdje je C konstanta. Na primjer: Ojlerova metoda: $\rho(h) = \frac{1}{2!}y''(x + \theta h)h^2$, gdje je $y''(x) = f'_x(x,y) + f'_y(x,y)f(x,y)$.

Posljednje, sada ćemo izdvojiti glavni sabirak greške $\rho(h)$. Pretpostavimo da $f \in C^{s+1}(\Omega)$ i $f \in B^{s+1}(\Omega)$. Tada važi Tejlorov razvoj:

$$\rho(h) = \frac{\rho^{(s+1)}(0)}{(s+1)!}h^{s+1} + \frac{\rho^{(s+2)}(\theta h)}{(s+2)!}h^{s+2},$$

gdje je $0 < \theta < 1$ (različito od ranijeg θ). Slijedi

$$\rho(h) = c(x,y)h^{s+1} + O(h^{s+2}),$$

gdje $c(x,y)$ ne zavisi od h. Glavni sabirak greške je $c(x,y)h^{s+1}$. Zapaža se da $c = c(x,y)$ neprekidno zavisi od (x,y). Primjer: Ojlerova metoda: $\rho(h) = \frac{1}{2!}y''(x)h^2 + \frac{1}{3!}y'''(x+\theta h)h^3$.

5.4. OCJENA GREŠKE ZA METODU RUNGE–KUTA

Ocjena greške metode jedne bilo koje fiksirane metode tipa Runge–Kuta doslovno se poklapa sa ocjenom greške Ojlerove metode sprovedenom u naslovu 5.2. Samo što je kod Ojlerove lokalna greška ρ bila reda h^2 a sada će biti reda h^{s+1}, tako da je tamo greška metode R_n u čvoru $x = x_n = x_0 + nh = x_0 + X$ izašla reda h a sada će izaći red h^s. Međutim, sada će biti izvršena potpuna analiza greške, odnosno biće uzete u obzir sve tri komponente greške R_n. Drugim riječima, pored greške metode biće uzeta u obzir i greška računanja, kao i greška izazvana približnošću (greškom) ulaznog podatka y_0.

Uvedimo oznake i pretpostavke. Dopustimo da mreža čvorova nije ekvidistantna; neka bude $x_0 < x_1 < \ldots < x_n = x_0+X$ i $x_{i+1}-x_i = h_i$. Neka se približna vrijednost y_{j+1} računa na osnovu prethodne y_j po formuli oblika $y_{j+1} = \Phi(f, x_j, h_j, y_j)$. Uzmimo da veličinu Φ nismo u stanju da izračunamo potpuno tačno, već da je u stvarnosti računamo sa izvjesnim zaokruživanjem. Zato prestaje da važi $y_{j+1} = \Phi(f, x_j, h_j, y_j)$, već uzimamo da važi $y_{j+1} = \Phi(f, x_j, h_j, y_j) - \delta_{j+1}$. Za δ_{j+1} se kaže da je greška računanja na koraku.

$y(x_0) = \pi$, $y_0(x_0) = y_0 = 3{,}14$, $R_0 = \pi - 3{,}14$, $|R_0| \leq 0{,}002$.

Mi numerički rješavamo Košijev zadatak

$$y' = f(x,y), \quad x_0 \leq x \leq x_0 + X, \quad y(x_0) = y(x_0),$$

čije tačno rješenje je funkcija $y = y(x)$. Može se desiti da je ulazni podatak ili ulazna veličina (to je jedna brojna vrijednost) $y(x_0)$ poznata samo približno. Uzmimo da imamo takve okolnosti i neka je y_0 odgovarajući približni broj. Neka je $R_0 = y(x_0) - y_0 = y(x_0) - y_0(x_0)$. Za R_0 se kaže da predstavlja grešku ulaznih podataka ili ulaznog podatka. I R_0 će se naravno odraziti na rezultujuću ukupnu grešku. Već je upotrebljena oznaka za funkciju $y_0 = y_0(x)$; ona po definiciji zadovoljava $y_0' = f(x, y_0)$, $y_0(x_0) = y_0$; zadovoljava d.j. $y' = f(x,y)$ i zadovoljava približni početni uslov. Oko oznaka zapažamo: u ovoj sekciji y_0 znači približni početni uslov (u ranijim sekcijama značilo je tačni početni uslov), a tačni početni uslov u ovoj sekciji označen je kao $y(x_0)$.

Uvedimo u razmatranje još n rješenja naše d. j. $y' = f(x,y)$. Za $j = \overline{1,n}$, neka je $y_j = y_j(x)$ funkcija koja zadovoljava razmatranu d.j. i zadovoljava početni uslov u tački $x = x_j$ po približnom broju y_j:

$$y_j' = f(x, y_j), \quad y_j(x_j) = y_j.$$

$y_j'(x) = f(x, y_j(x))$.

Sa ρ_j ćemo označavati grešku metode na koraku: $\rho_j = y_{j-1}(x_j) - \Phi(f, x_{j-1}, h_{j-1}, y_{j-1})$. Pretpostavićemo da je greška metode na koraku (greška metode na koraku ove fiksirane metode tipa Runge–Kuta čija se ukupna greška razmatra) reda h^{s+1}, odnosno da je $|\rho_j| \leq Ch_{j-1}^{s+1}$ za svako j.

Za funkciju $\dfrac{\partial f}{\partial y}$ ili svejedno $f_y'(x,y)$ pretpostavićemo da je neprekidna na skupu $\Omega = [x_0, x_0+X] \times (-\infty, +\infty)$. I da je ona ograničena na tom skupu, da je za neko $L > 0$ ispunjeno $\left|\dfrac{\partial f}{\partial y}\right| \leq L$ za $(x,y) \in \Omega$. Ovo posljednje je u vezi primjene leme iz naslova 5.1.

Nas interesuje greška $R_n = y(x_n) - y_n = y(x_n) - y_n(x_n)$ u tački $x_n = x_0 + X$ (u krajnjoj desnoj tački razmatranog odsječka).

Teorema. Neka je $|\rho_j| \le Ch_{j-1}^{s+1}$ i neka je $\left|\dfrac{\partial f}{\partial y}\right| \le L$ na skupu Ω. Uvedimo oznake $H = \max_{1\le j\le n} h_{j-1}$ i $\delta = \max_{1\le j\le n} |\delta_j|$. Tada važi nejednakost

$$|R_n| \le e^{LX}\Big(CXH^s + n\delta + |R_0|\Big). \tag{1}$$

Dokaz:

$$R_n = y(x_n) - y_n = y(x_n) - y_n(x_n) = y(x_n) - y_0(x_n) + y_0(x_n) - y_n(x_n) =$$

$$y(x_n) - y_0(x_n) + \sum_{j=1}^{n}\Big(y_{j-1}(x_n) - y_j(x_n)\Big)$$

$$|R_n| \le \underbrace{|y(x_n) - y_0(x_n)|}_{\text{prvi izraz}} + \underbrace{\sum_{j=1}^{n} |y_{j-1}(x_n) - y_j(x_n)|}_{\text{drugi izraz}}$$

Za pojedini sabirak u drugom izrazu primijenimo lemu, sa $\alpha = x_j$, $\beta = x_n$, $Y_1(x) = y_{j-1}(x)$, $Y_2(x) = y_j(x)$:

$$y_{j-1}(x_n) - y_j(x_n) = (y_{j-1}(x_j) - y_j(x_j))\exp\Big\{\int_{x_j}^{x_n} f'_y(x, \overline{y}_j(x))dx\Big\} =$$

$$(y_{j-1}(x_j) - \Phi(f, x_{j-1}, h_{j-1}, y_{j-1}) + \Phi(f, x_{j-1}, h_{j-1}, y_{j-1}) - y_j(x_j))\exp\Big\{\dots \text{ isto } \dots\Big\} =$$

$$(\rho_j + \delta_j)\exp\Big\{\dots \text{ isto } \dots\Big\}$$

$$|y_{j-1}(x_n) - y_j(x_n)| \le (|\rho_j| + |\delta_j|)\exp\Big\{\int_{x_j}^{x_n} |f'_y(x, \overline{y}_j(x))|dx\Big\} \le$$

$$(Ch_{j-1}^{s+1} + \delta)\exp\Big\{\int_{x_j}^{x_n} Ldx\Big\} \le (Ch_{j-1}^{s+1} + \delta)\exp\Big\{LX\Big\}$$

Sabiranjem:

$$\sum_{j-1}^{n} |y_{j-1}(x_n) - y_j(x_n)| \le \Big(\sum_{j=1}^{n} Ch_{j-1}^{s+1} + n\delta\Big)\exp\Big\{LX\Big\} \le \Big(CXH^s + n\delta\Big)\exp\Big\{LX\Big\},$$

jer je

$$h_0^{s+1} + h_1^{s+1} + \ldots + h_{n-1}^{s+1} = h_0^s \cdot h_0 + h_1^s \cdot h_1 + \ldots + h_{n-1}^s \cdot h_{n-1} \le$$

$$H^s \cdot h_0 + H^s \cdot h_1 + \ldots + H^s \cdot h_{n-1} = H^s X$$

Slično za prvi izraz, također pomoću leme:

$$|y(x_n) - y_0(x_n)| \le |y(x_0) - y_0(x_0)|\exp\Big\{\int_{x_0}^{x_n} |f'_y(x, \overline{y}_0(x))|dx\Big\} \le |R_0|\exp\Big\{LX\Big\}$$

Na kraju, prvi izraz plus drugi izraz:

$$|R_n| \le \Big(CXH^s + n\delta + |R_0|\Big) \exp\Big\{LX\Big\}$$

Dokaz je završen.

Šablon oznaka: $y_k(x_k) = y_k$ i

	približno	tačno	greška
x_0	y_0	$y(x_0)$	$R_0 = y(x_0) - y_0$
x_1	y_1	$y(x_1)$	$R_1 = y(x_1) - y_1$
x_2	y_2	$y(x_2)$	$R_2 = y(x_2) - y_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_n = x_0 + X$	y_n	$y(x_n)$	$R_n = y(x_n) - y_n$
	(računar)		

Slika kada je $n = 4$; $R_4 = y(x_4) - y_4 = y(x_4) - y_4(x_4)$:

Slika je slična slici za Ojlerovu metodu, samo što su tamo visinske razlike u tačkama $x = x_i$ $(i = 1, \ldots, 4)$ bile ρ_i a sada su $\rho_i + \delta_i$ i samo što je tamo $y = y(x)$ bilo isto što i $y_0 = y_0(x)$ a sada su to dvije različite funkcije, njihova visinska razlika u tački $x = x_0$ na ovoj slici iznosi R_0.

Pogledajmo završnu formulu (1) koja izražava totalnu ili sveukupnu grešku R_n u krajnjoj desnoj tački odsječka od x_0 do $x_0 + X$. Vidimo da greška $R_n \to 0$ ako $H \to 0$ uz istovremeno $n\delta \to 0$ i $R_0 \to 0$; numerička metoda konvergira. Veličina $n\delta$ predstavlja zbirnu grešku računanja, a R_0 je greška ulaznih podataka. Ako se razmatra samo greška metode onda važi nejednakost $|R_n| \le e^{LX} \cdot CXH^s$.

5.5. ALGORITAM ZASNOVAN NA METODI RUNGE–KUTA

Samo ćemo naznačiti elemente koji čine algoritam. Neka smo se opredijelili za jednu određenu metodu. Označimo sa s red greške metode; obično je $s = 4$. Uzimamo da je prisutna jedino greška metode. Koristićemo formulu za grešku metode:

$$R_h(x_0 + X) = \sum_{j=1}^{n} \rho_j \exp\Big\{\int_{x_j}^{x_n} f'_y(x, \overline{y}_j(x))dx\Big\}, \tag{1}$$

gdje je $X = nh$, v. u 5.4. Koristićemo i formulu u kojoj je izdvojen glavni sabirak lokalne greške:

$$\rho(h) = ch^{s+1} + O(h^{s+2}), \tag{2}$$

gdje je $c = c(x, y)$ neprekidna funkcija, v. na kraju 5.3.

1. Jedan postupak da se ocijeni greška

Neka su $z_h(x_0+X)$ i $z_{2h}(x_0+X)$ dvije približne vrijednosti za $y(x_0+X)$ dobijene redom sa korakom h odnosno $2h$. Neka su $R_h(x_0+X)$ i $R_{2h}(x_0+X)$ dvije odgovarajuće greške, tako da je očito

$$y_(x_0+X) = z_h(x_0+X) + R_h(x_0+X) \;\text{ i }\; y(x_0+X) = z_{2h}(x_0+X) + R_{2h}(x_0+X).$$

Iz (1) i (2) lako se može izvesti sljedeća formula koja izražava vezu između jedne i druge greške:

$$\lim_{h\to 0} R_{2h}(x_0+X)/R_h(x_0+X) = 2^s. \tag{3}$$

Ostavlja se čitaocu da izvede formulu (3), kao jedan zadatak za vježbu iz Analize. Zapaziti da je $\lim_{h\to 0}\overline{y}_j(x) = y(x)$, gdje je $y=y(x)$ tačno rješenje K.z. Isto tako, $\rho(2h)\sim 2^{s+1}\rho(h)$ kad $h\to 0$.

Iz (3) se vrlo lako izvodi sljedeća formula:

$$R_h(x_0+X) \sim \Big(z_h(x_0+X) - z_{2h}(x_0+X)\Big)\Big/\Big(2^s-1\Big) \;\text{ kad }\; h\to 0. \tag{4}$$

Ona se izvodi na običan Rungeov način. U numeričkoj praksi, formula (4) koristi se u obliku

$$R_h(x_0+X) \approx \Big(z_h(x_0+X) - z_{2h}(x_0+X)\Big)\Big/\Big(2^s-1\Big).$$

Očito je da posljednja napisana formula služi da se dobije procjena za grešku približnog broja $z_h(x_0+X)$. Bitno je što je procjena efektivna (mi smo u mogućnosti da izračunamo izraz na desnoj strani).

2. Drugi postupak da se ocijeni greška (početak)

Znamo da formula (1) važi i za ravnomjernu i za neravnomjernu mrežu. Obično se $f(x,\overline{y}_j(x))$ zamijeni sa $L=\sup f'_y$ čime se dobija jedna prilično uveličana ocjena greške. Formula (1) postaće praktično upotrebljiva kada se nađu efektivne procjene za ρ_j. U nastavku će i biti izvedene takve procjene za ρ_j.

3. Drugi postupak da se ocijeni greška (središnji dio i kraj)

Neka je $h>0$ i neka je $X=nh$. Prvo se izvrši glavni proračun. Glavni proračun izvrši se čitav sa korakom h. Dakle, neka y_k označava približnu vrijednost koja se odnosi na čvor $x_k = x_0+kh$, za $k=\overline{1,n}$. Ponovimo: računar je saopštio brojeve $\{y_k\}_{k=1}^n$, oni predstavljaju numerički odgovor i pitamo se kolika je njihova greška. Može se reći da je broj y_{k+1} dobijen na osnovu (x_k,y_k) sa korakom h. Može se reći da je broj y_{k+2} dobijen na osnovu (x_{k+1},y_{k+1}) sa korakom h. V. sliku. U cilju dobijanja efektivne ocjene greške numeričkog odgovora y_n, sada se izvrši i pomoćni proračun, kako slijedi. Na osnovu (x_k,y_k) sa korakom $2h$ izračunati i približnu vrijednost Y_{k+2} koja se (znači) odnosi na čvor x_{k+2}; izračunati za svako parno k. Približna vrijednost Y_{k+2} ima pomoćnu ulogu, ona služi da se procijene lokalne greške ρ_{k+1} i ρ_{k+2} koje se tiču glavnog proračuna, one se odnose na y_{k+1} i y_{k+2}. Lako se vidi da je $\rho_{k+2}\sim\rho_{k+1}$ kad $h\to 0$ (tj. da je $\lim_{h\to 0}\rho_{k+2}/\rho_{k+1}=1$), što se u numeričkoj praksi koristi u obliku $\rho_{k+2}\approx\rho_{k+1}$, za male h. Upotrebom temeljnih formula (1) i (2) dosta lako se može izvesti sljedeća relacija:

$$\rho_{k+1}+\rho_{k+2} \approx \frac{1}{2^s-1}\Big(y_{k+2}-Y_{k+2}\Big). \tag{5}$$

4. Podešavanje koraka

Kvalitetni programi (moćan softver) za rješavanje Košijevog zadatka nastaju ako se tokom rješavanja (tokom napredovanja po x–osi) podešava korak integracije h. Tekuća vrijednost koraka se povećava ili smanjuje, u zavisnosti od okolnosti. Bolje rečeno, u zavisnosti od tekućeg iznosa lokalne greške. Kaže se da je u programu korak promjenljiv ili se kaže da program automatski vrši podešavanje koraka. Pogledajmo malo detaljnije o čemu se radi.

Tokom izvršavanja programa mi postepeno dobijamo nove i nove približne vrijednosti. Jedan parametar programa jeste broj $\varepsilon > 0$ čiji je smisao – najveći dopušteni iznos lokalne greške; vidjećemo da ε utiče na tekuću dužinu koraka. U algoritmu, kada se naprave dva poteza odnosno kada se saznaju dvije približne vrijednosti onda se učini i jedan pomoćni potez sa dvokorakom, čime se sazna broj oblika Y_{k+2}. Po formuli (5) izračunati procjenu za lokalnu grešku $\rho_{k+1} = e$. Uporediti ε i $|e|$, pri čemu postoje tri mogućnosti, kako slijedi:

1. Ako je $|e| > \varepsilon$ onda treba poništiti dvije posljednje izračunate približne vrijednosti oblika y_{k+1} i y_{k+2} (vratiti se u tačku $x = x_k$) i zatim nastaviti sa prepolovljenim korakom $h/2$.

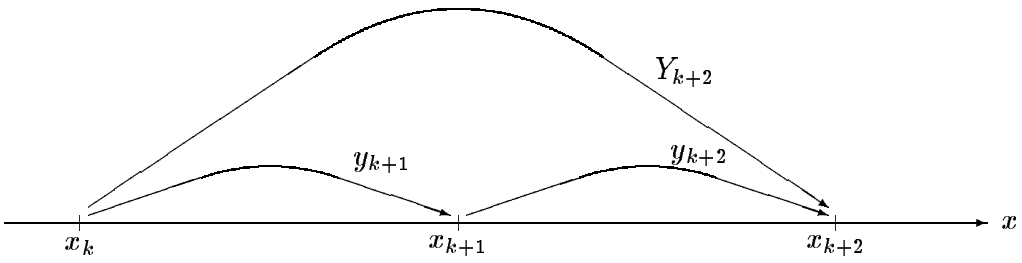
2. Ako je $|e| < \dfrac{1}{64}\varepsilon$ onda treba udvostručiti korak.

3. Ako je $\dfrac{1}{64}\varepsilon < |e| < \varepsilon$ onda treba nastaviti sa tekućim korakom.

Pisali smo za slučaj kada je $s = 4$. Formula (5) govori da je $2e \approx \dfrac{1}{15}\Big(y_{k+2} - Y_{k+2}\Big)$. Udvostručiti korak – pomnožiti lokalnu grešku sa 32.

5. Najmanji i najveći korak

Program obično ima još dva parametra $h_{\min}$ i $h_{\max}$ čiji je smisao jasan iz oznake: to su redom najmanja odnosno najveća dopuštena vrijednost koraka. Ako bi tekući korak pao ispod donje granice $h_{\min}$ onda bi se moglo desiti da greška računanja postane neprihvatljivo velika.



5.6. DIFERENCNE METODE

1. Pojam Adamsove metode i njene lokalne greške

Neka je $p_k(x)$ L.i.p. funkcije $F(x) = y'(x)$ po mreži čvorova $x_{n-j} = x_n - jh$, gdje je $j = \overline{1,k}$. Napiši eksplicitni izraz za $p_k(x)$ preko $y'(x_{n-k})$, ..., $y'(x_{n-1})$. Po izrazu za grešku interpolacije:

$$F(x) - p_k(x) = y'(x) - p_k(x) = r_k(x) =$$

$$\frac{1}{k!}\omega_k(x)F^{(k)}(\xi(x)) = \frac{1}{k!}\omega_k(x)y^{(k+1)}(\xi(x)), \quad \omega_k(x) = \prod_{j=1}^{k}(x - x_{n-j}),$$

ako $y \in C^{k+1}[x_0, x_0 + X]$. Na jednakost $y'(x) = p_k(x) + r_k(x)$ primijeni $\int_{x_{n-1}}^{x_n} dx$:

$$y(x_n) - y(x_{n-1}) = \int_{x_{n-1}}^{x_n} p_k(x)dx + \int_{x_{n-1}}^{x_n} r_k(x)dx,$$

$$y(x_n) = y(x_{n-1}) + \int_{x_{n-1}}^{x_n} p_k(x)dx + \rho_n \quad \text{ili} \quad y(x_n) \approx y(x_{n-1}) + \int_{x_{n-1}}^{x_n} p_k(x)dx.$$

Za ρ_n se kaže da je lokalna greška metode ili da je greška metode na koraku. Izračunaj $\int_{x_{n-1}}^{x_n} p_k(x)dx$. Račun pokazuje da je $\int_{x_{n-1}}^{x_n} p_k(x)dx$ jedna linearna kombinacija vrijednosti $y'(x_{n-k})$, ..., $y'(x_{n-1})$:

$$\int_{x_{n-1}}^{x_n} p_k(x)dx = h\Big(b_{-k}y'(x_{n-k}) + \ldots + b_{-1}y'(x_{n-1})\Big),$$

gdje su b_{-k}, ..., b_{-1} konstante. Sprovesti račun u slučajevima $k = 1$, $k = 2$, $k = 3$ i $k = 4$.

Neka $y = y(x)$ označava rješenje d.j. $y' = f(x, y)$. Tada je $y'(x_{n-j}) = f(x_{n-j}, y(x_{n-j}))$.

Razmotrimo K.z.

$$y' = f(x, y), \quad x_0 \le x \le x_0 + X, \quad y(x_0) = y_0$$

i razmotrimo mogućnost njegovog numeričkog rješavanja po ravnomjernoj mreži čvorova čiji je korak $h = X/N$ tj. po mreži čvorova $x_n = x_0 + nh$, gdje je $n = \overline{0,N}$. Označimo sa y_n približne vrijednosti u čvorovima, za $n = \overline{0,N}$. Iz prethodnog izvođenja vidimo da smo ustvari konstruisali jednu numeričku metodu za rješavanje postavljenog K.z. Ta metoda izražava se sljedećom šemom za računanje koja ide u računar:

$$y_n = y_{n-1} + h\Big(b_{-k}f(x_{n-k}, y_{n-k}) + \ldots + b_{-1}f(x_{n-1}, y_{n-1})\Big), \quad n = \overline{k,N};$$

ovo je eksplicitna (e.) ili ekstrapolaciona Adamsova formula. Vidi se da veličine b_{-k}, ..., b_{-1} zavise od k.

U nastavku su date eksplicitne formule u slučajevima $k = 1$, $k = 2$, $k = 3$ i $k = 4$. Koristi se obična skraćenica $f_i = f(x_i, y_i)$:

$$y_n = y_{n-1} + hf_{n-1}, \quad y_n = y_{n-1} + \frac{h}{2}\Big(3f_{n-1} - f_{n-2}\Big),$$

$$y_n = y_{n-1} + \frac{h}{12}\Big(23f_{n-1} - 16f_{n-2} + 5f_{n-3}\Big),$$

$$y_n = y_{n-1} + \frac{h}{24}\Big(55f_{n-1} - 59f_{n-2} + 37f_{n-3} - 9f_{n-4}\Big)$$

Vidimo da je $k = 1$ Ojlerova metoda. Za $k = 4$ kaže se da je Adamsova e. metoda IV reda.

Zapaža se da metoda nije u stanju sama da startuje odnosno da je za startovanje potrebna i druga pomoćna metoda. Upravo, približne vrijednosti y_1, ..., y_{k-1} izračunaju se po nekoj metodi tipa Runge–Kuta.

Prilikom buduće ocjene greške $R_n = y(x_n) - y_n$ treba uzeti u obzir faktore koji će sada biti nabrojani. a) Lokalnu grešku ρ_n, jer sabirak ρ_n ne ide u računar. b) Prilikom računanja y_n mi koristimo približne brojeve y_{n-j} sa kojima raspolažemo; ne koristimo tačne vrijednosti $y(x_{n-j})$, jer te vrijednosti ne znamo. c) Grešku približnih vrijednosti y_1, ..., y_{k-1} koje je saopštila pomoćna metoda tipa Runge–Kuta.

Sada ćemo izvršiti ocjenu lokalne greške ρ_n:

$$\omega_k(x) = (x - x_{n-k}) \cdot \ldots \cdot (x - x_{n-1})$$

$$r_k(x) = \frac{1}{k!}\omega_k(x)y^{(k+1)}(\xi(x))$$

$$\min(x_{n-k}, \ldots, x_{n-1}, x) < \xi(x) < \max(x_{n-k}, \ldots, x_{n-1}, x)$$

$$\min(x_{n-k}, x) < \xi(x) < \max(x_{n-1}, x)$$

$$x_{n-k} < \xi(x) < x_n, \text{ jer } x \in [x_{n-1}, x_n]$$

$$\rho_n = \int_{x_{n-1}}^{x_n} r_k(x)dx = \frac{1}{k!}\int_{x_{n-1}}^{x_n} \omega_k(x)y^{(k+1)}(\xi(x))dx =$$

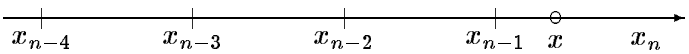
t. o srednjoj vrijednosti, jer je $\omega_k(x)$ stalnog znaka za $x \in [x_{n-1}, x_n]$

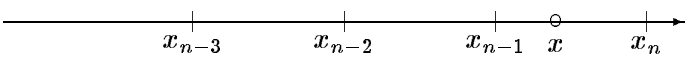
$$\frac{1}{k!}y^{(k+1)}(\xi_n)\int_{x_{n-1}}^{x_n} \omega_k(x)dx =$$

izračunaj integral, pomoću smjene $x = x_n + ht$

$$\frac{1}{k!}y^{(k+1)}(\xi_n) \cdot C_k h^{k+1}, \quad \text{ovdje } \xi_n \in (x_{n-k}, x_n)$$

Na primjer, u slučaju $k = 4$ dobija se $C_4 = \dfrac{251}{30}$ tako da je $\rho_n = \dfrac{251}{720}h^5 y^V(\xi_n)$.

e. $k = 4$ 

i. $k = 4$ 

Prelazimo na implicitnu (i.) ili interpolacionu Adamsovu formulu. Neka sada $p_k(x)$ bude interpolacioni polinom funkcije $F(x) = y'(x)$ po mreži čvorova $x_{n-j} = x_n - jh$, gdje je $j = \overline{0, k-1}$. Sada je $\omega_k(x) = \prod_{j=0}^{k-1}(x - x_{n-j})$. Izvođenje je slično onome od maločas. Račun pokazuje da je:

$$\int_{x_{n-1}}^{x_n} p_k(x)dx = h\Big(b_{-k+1}y'(x_{n-k+1}) + \ldots + b_0 y'(x_n)\Big).$$

I. Adamsova formula:

$$y_n = y_{n-1} + h\Big(b_{-k+1}f(x_{n-k+1}, y_{n-k+1}) + \ldots + b_0 f(x_n, y_n)\Big), \quad n = \overline{k-1, N}$$

(b_j e. $\neq b_j$ i.)

Ustvari smo napisali jednu (nelinearnu) jednačinu po nepoznatoj y_n, jer y_n figuriše i na lijevoj i na desnoj strani posljednje jednakosti. U praksi će ta jednačina biti rješavana samo približno tačno. Tako da prilikom buduće ocjene greške u slučaju implicitne šeme treba uzeti u obzir još jedan faktor, kako slijedi. d) Greška sa kojom je riješena nelinearna jednačina.

U nastavku su date implicitne formule u slučajevima $k = 1$, $k = 2$, $k = 3$ i $k = 4$:

$$y_n = y_{n-1} + hf_n, \quad y_n = y_{n-1} + \frac{h}{2}\Big(f_n + f_{n-1}\Big), \quad y_n = y_{n-1} + \frac{h}{12}\Big(5f_n + 8f_{n-1} - f_{n-2}\Big),$$

$$y_n = y_{n-1} + \frac{h}{24}\Big(9f_n + 19f_{n-1} - 5f_{n-2} + f_{n-3}\Big)$$

Za posljednju formulu kaže se da je Adamsova i. metoda IV reda.

Naravno da se lokalna greška ρ_n definiše kao $\rho_n = \int_{x_{n-1}}^{x_n} r_k(x)dx$, gdje je $r_k(x)$ greška interpolacije. Kada se sprovede račun koji je po svemu sličan onome od maločas onda će se dobiti da je:

$$\rho_n = \frac{1}{k!}y^{(k+1)}(\xi_n) \cdot C_k h^{k+1}, \quad \text{ovdje } \xi_n \in (x_{n-k+1}, x_n)$$

(C_k e. $\neq C_k$ i.)

Recimo, $k = 4$: $C_4 = -\dfrac{19}{30}$ i $\rho_n = -\dfrac{19}{720}h^5 y^V(\xi_n)$.

2. Pojam prediktor–korektor metode

Kako da se (nelinearna) jednačina riješi brzo i sa dobrom preciznošću? Zapaža se da je pogodno tu jednačinu rješavati metodom proste iteracije. Stavimo

$$y_n^{(\ell)} = y_{n-1} + h\Big(b_{-k+1}f(x_{n-k+1}, y_{n-k+1}) + \ldots + b_{-1}f(x_{n-1}, y_{n-1}) + b_0 f(x_n, y_n^{(\ell-1)})\Big)$$

ili $\quad y_n^{(\ell)} = \varphi(y_n^{(\ell-1)})$ za $\ell \geq 1$,

gdje je $\quad \varphi(t) = y_{n-1} + h\Big(b_{-k+1}f(x_{n-k+1}, y_{n-k+1}) + \ldots + b_{-1}f(x_{n-1}, y_{n-1}) + b_0 f(x_n, t)\Big)$.

Tako da je $\quad \varphi'(t) = hb_0 f'_y(x_n, t)$.

Želimo da bude $y_n^{(\ell)} \to y_n$ kad $\ell \to \infty$, a treba da izaberemo početnu aproksimaciju $y_n^{(0)}$. Po čemu je pogodno? S jedne strane, za dovoljno male h biće $|\varphi'(t)| \leq q < 1$ odnosno uslov kontrakcije biće ispunjen, ako je $\dfrac{\partial f}{\partial y}$ ograničena funkcija. S druge strane, dovoljno dobru početnu aproksimaciju može da nam pruži neka eksplicitna formula.

Iskustvo pokazuje da je dovoljno da se izvrši samo jedna iteracija. Drugim riječima, ne ide se dalje od $\ell = 1$; stavlja se da je $y_n \approx y_n^{(1)}$. Opisali smo postupak rješavanja (nelinearne) jednačine.

Pogledajmo još jednom taj postupak. Računaju se dva broja $y_n^{(0)} = y_n^{\text{pred}}$ i $y_n^{(1)} = y_n^{\text{kor}}$; y_n^{kor} se uzima za y_n. Vidimo da se kombinuju jedna e. šema i jedna i. šema. Jedna i druga šema primijene se po jednom. Tako nastaje jedna nova (treća) šema, za koju se kaže da je prediktor–korektor tipa (skraćeno p.–k. tipa).

Ponovimo: broj y_n^{pred} računa se po nekoj eksplicitnoj šemi, a broj $y_n = y_n^{\text{kor}}$ računa se po nekoj implicitnoj šemi.

Dvije šeme koje se kombinuju uzimaju se sa jednim te istim k po pravilu.

Također se i pomoćna startna metoda tipa Runge–Kuta uzima po pravilu da ima isti red lokalne greške kao i dvije metode koje se kombinuju.

Broj y_n^{pred} predskazuje, a $y_n^{\text{kor}} = y_n$ je popravljena vrijednost.

Iskustvo ili praksa pokazuje da upotreba metoda oblika p.–k. daje najbolje aproksimacije.

Navedimo jedan primjer šeme oblika p.–k. Adamsova p.–k. šema IV reda data je sa:

$$y_n^* = y_{n-1} + \frac{h}{24}\Big(55f_{n-1} - 59f_{n-2} + 37f_{n-3} - 9f_{n-4}\Big), \quad f_n^* = f(x_n, y_n^*),$$

$$y_n = y_{n-1} + \frac{h}{24}\Big(9f_n^* + 19f_{n-1} - 5f_{n-2} + f_{n-3}\Big), \quad n = \overline{4, N}.$$

3. Primjeri diferencnih šema

Budući da je $y' = f(x, y)$ to je

$$\int_{x_n - ph}^{x_n} y'(x)dx = \int_{x_n - ph}^{x_n} f(x, y(x))dx \text{ ili } y(x_n) - y(x_n - ph) = \int_{x_n - ph}^{x_n} f(x, y(x))dx.$$

Sada se na posljednji integral primijeni neka kvadraturna formula.

Neka bude $p = 1$ i neka se primijeni trapezna formula:

$$y(x_n) - y(x_{n-1}) \approx \frac{h}{2}\Big(f(x_{n-1}, y(x_{n-1})) + f(x_n, y(x_n))\Big).$$

Odavde imamo šemu:

$$y_n = y_{n-1} + \frac{h}{2}\Big(f(x_{n-1}, y_{n-1}) + f(x_n, y_n)\Big) \text{ ili svejedno } y_n = y_{n-1} + \frac{h}{2}\Big(f_{n-1} + f_n\Big).$$

Koristi se obična oznaka $f_i = f(x_i, y_i)$. Vidimo da je ovo jedna i. šema.

Neka bude $p = 2$ i neka se primijeni Simpsonova formula:

$$y(x_n) - y(x_{n-2}) \approx \frac{h}{3}\Big(f(x_{n-2}, y(x_{n-2})) + 4f(x_{n-1}, y(x_{n-1})) + f(x_n, y(x_n))\Big).$$

Odavde imamo šemu:

$$y_n = y_{n-2} + \frac{h}{3}\Big(f_{n-2} + 4f_{n-1} + f_n\Big).$$

Ponekad se zapisuje kao $y_n = y_{n-2} + \frac{h}{3}(y'_{n-2} + 4y'_{n-1} + y'_n)$. Vidimo da je ovo jedna i. šema.

Neka bude $p = 2$ i neka se primijeni formula pravougaonika:

$$y(x) - y(x - 2h) \approx 2hf(x - h, y(x - h)) \text{ ili } y(x_n) - y(x_{n-2}) \approx 2hf(x_{n-1}, y(x_{n-1})).$$

Odavde imamo šemu za računanje:

$$y_n = y_{n-2} + 2hf_{n-1}.$$

Vidimo da je ovo jedna e. šema.

4. Opšti oblik diferencne šeme

Razmotrimo ekvidistantnu mrežu čvorova $x_0, x_1, x_2, \ldots$ čiji je korak $h > 0$; $x_i = x_0 + ih$ za $i \geq 0$. Sljedeća jednakost definiše opšti oblik diferencne (ili višekoračne ili k–koračne) metode:

$$\sum_{i=0}^{k} a_{-i} y_{n-i} - h \sum_{i=0}^{k} b_{-i} f(x_{n-i}, y_{n-i}) = 0. \tag{1}$$

Formula (1) služi za računanje približnog broja y_n, a smatra se da su već poznati približni brojevi y_j, gdje je $j < n$.

Za startovanje šeme treba da su već određeni $y_1, \ldots, y_{k-1}$. Za njihovo određivanje koristi se neka jednokoračna metoda.

U praksi se upotrebljavaju jedino šeme sa $a_0 \neq 0$. Ako bi bilo $a_0 = 0$ onda bi se y_n nalazilo u sabirku $-hb_0 f(x_n, y_n)$. Razmatraćemo samo šeme u kojima je $a_0 \neq 0$.

Diferencne šeme dijele se na eksplicitne i implicitne. Ako je $b_0 = 0$ onda se za šemu kaže da je eksplicitna: y_n se neposredno računa. Ako je $b_0 \neq 0$ onda se za šemu kaže da je implicitna: y_n učestvuje i u sabirku $-hb_0 f(x_n, y_n)$, tako da (1) predstavlja jednu (nelinearnu) jednačinu po nepoznatoj y_n.

Diferencne šeme dijelimo na one koje su Adamsovog tipa i one koje nisu Adamsovog tipa. Pogledajmo koeficijente $a_0, a_{-1}, \ldots, a_{-k}$. Ako je $a_{-2} = \ldots = a_{-k} = 0$ onda se za šemu kaže da je Adamsova ili da je Adamsovog tipa; tada je očito $a_0 = 1$ i $a_{-1} = -1$.

Na kraju, lokalna greška ρ_n opšte diferencne šeme izražene formulom (1) definiše se slično kao lokalna greška za ranije navedene posebne slučajeve. Neka bude $a_0 = 1$. Formula (1) definiše sljedeću šemu za računanje koja ide u računar:

$$y_n = -\sum_{i=1}^{k} a_{-i} y_{n-i} + h \sum_{i=0}^{k} b_{-i} y(x_{n-i}, y_{n-i}).$$

Ta šema za računanje nastala je od neke formule oblika

$$y(x_n) \approx -\sum_{i=1}^{k} a_{-i} y(x_{n-i}) + h \sum_{i=0}^{k} b_{-i} f(x_{n-i}, y(x_{n-i})).$$

Prema tome

$$\rho_n = y(x_n) + \sum_{i=1}^{k} a_{-i} y(x_{n-i}) - h \sum_{i=0}^{k} b_{-i} f(x_{n-i}, y(x_{n-i})).$$

5.7. METODA NEODREĐENIH KOEFICIJENATA

Napišimo opet opšti oblik diferencne šeme:

$$\sum_{i=0}^{k} a_{-i} y_{n-i} - h \sum_{i=0}^{k} b_{-i} f_{n-i} = 0 \text{ ili svejedno } \sum_{i=0}^{k} a_{-i} y_{n-i} - h \sum_{i=0}^{k} b_{-i} f(x_{n-i}, y_{n-i}) = 0,$$

gdje su a_{-i} i b_{-i} konstante, tj. a_{-i} i b_{-i} ne zavise od h. Posmatrajmo sljedeći izraz:

$$L = \sum_{i=0}^{k} a_{-i} y(x_{n-i}) - h \sum_{i=0}^{k} b_{-i} f(x_{n-i}, y(x_{n-i})) \text{ ili } L = \sum_{i=0}^{k} a_{-i} y(x_{n-i}) - h \sum_{i=0}^{k} b_{-i} y'(x_{n-i}),$$

gdje je očito uzeto da je funkcija $y = y(x)$ tačno rješenje d. j. $y' = f(x, y)$. Metoda neodređenih koeficijenata predstavlja jedan mogući način da se izaberu zasad neodređene veličine a_{-i} i b_{-i} tj. da se pogodno odrede koeficijenti a_{-i} i b_{-i}. Postupa se kako slijedi. Napišimo Tejlorov razvoj funkcije $y = y(x)$ u okolini tačke $x = x_n$:

$$y(x_n + \alpha) = y(x_n) + y'(x_n) \cdot \alpha + \frac{1}{2!} y''(x_n) \cdot \alpha^2 + \dots$$

(do nekog stepena α). Diferencirajmo ili svejedno napišimo Tejlorov razvoj funkcije $y' = y'(x)$ u okolini tačke $x = x_n$:

$$y'(x_n + \alpha) = y'(x_n) + y''(x_n) \cdot \alpha + \frac{1}{2!} y'''(x_n) \cdot \alpha^2 + \dots.$$

Napišimo jedan i drugi razvoj kada je $\alpha = -kh$, ..., $\alpha = -h$ i $\alpha = 0$ redom. Sve to uvrstimo u L. Na taj način, kada se izvrši sređivanje:

$$L = E_0 y(x_n) + E_1 h y'(x_n) + E_2 h^2 y''(x_n) + \dots + E_{q-1} h^{q-1} y^{(q-1)}(x_n) + O(h^q).$$

Sada želimo da što više početnih koeficijenata E_p bude jednako nuli. Recimo, želimo da bude $E_0 = E_1 = E_2 = E_3 = 0$. Na osnovu te želje određuju se koeficijenti a_{-i} i b_{-i}.

Pretpostavlja se da je tačno rješenje $y = y(x)$ dovoljno glatka funkcija.

Nećemo dalje konkretizovati ovaj postupak.

5.8. OCJENA GREŠKE DIFERENCNE METODE

1. Slučaj eksplicitne Adamsove metode

Razmotrimo eksplicitnu Adamsovu šemu:

$$y_n = y_{n-1} + h \sum_{i=1}^{k} b_{-i} f(x_{n-i}, y_{n-i}), \quad n = \overline{k, N}.$$

Kao i dosad, $y = y(x)$ označava tačno rješenje razmatranog K. z. a y_n su približne vrijednosti. Isto tako, $x_n = x_0 + nh = x_0 + nX/N$ (mreža čvorova je ravnomjerna).

Brojeve $y_1, \dots, y_{k-1}$ dala je neka pomoćna metoda. Za grešku tih brojeva pretpostavlja se sljedeće: $|R_j| = |y(x_j) - y_j| \le C_0 h^{k+1}$ za $j = 1, \dots, k-1$.

Iz 5.6. za lokalnu grešku ρ_n znamo:

$$\rho_n = y(x_n) - y_n^* = C_k h^{k+1} y^{(k+1)}(\xi), \quad \text{gdje je } \; y_n^* = y(x_{n-1}) + h\sum_{i=1}^{k} b_{-i} f(x_{n-i}, y(x_{n-i})).$$

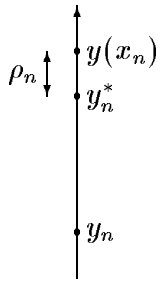
Slijedi da je

$$|\rho_n| \leq C_1 h^{k+1},$$

ako se pretpostavi sljedeće: $(k+1)$–vi izvod $y^{(k+1)}(x)$ funkcije $y = y(x)$ je ograničen na odsječku $[x_0, x_0 + X]$, gdje $y = y(x)$ označava bilo koje rješenje d. j. $y' = f(x, y)$. Ako $f \in C^k(\Omega) = C^k([x_0, x_0 + X] \times R)$ onda važi ta ograničenost, ako su još naravno svi parcijalni izvodi funkcije $f = f(x, y)$ do reda k uključeno ograničeni na skupu Ω.

Mi treba da ocijenimo grešku metode $R_n = y(x_n) - y_n$ za razne n.

Mi ćemo u izvođenju da izrazimo R_n preko $R_{n-k}, \ldots, R_{n-1}$. Mi ćemo napredovati po x–osi od tačke $x = x_0$ prema tački $x = x_N = x_0 + X$.



Slika za 1. Eksplicitna Adamsova
$y(x_n)$ je tačan broj, y_n je približan broj (računar saopštava y_n)
$R_n = y(x_n) - y_n$ je greška, ρ_n je lokalna greška

Vidjećemo da je potrebna još jedna pretpostavka (treća), pa se sada i ona uvodi, kako slijedi. Neka funkcija od dvije promjenljive $f = f(x, y)$ zadovoljava Lipšicov uslov po svojoj drugoj promjenljivoj y, sa Lipšicovom konstantom L, tj. neka za ma koje dvije tačke $(x, y_1) \in \Omega$ i $(x, y_2) \in \Omega$ važi nejednakost

$$|f(x, y_1) - f(x, y_2)| \leq L \cdot |y_1 - y_2|.$$

Za treću pretpostavku je dovoljno da je funkcija $f'_y(x, y)$ ograničena na skupu Ω tj. da je $|f'_y(x, y)| \leq L$ za svako $(x, y) \in \Omega$.

Imamo redom:

$$R_n = y(x_n) - y_n = y(x_n) - y_n^* + y_n^* - y_n = \rho_n + y_n^* - y_n =$$

$$\rho_n + y(x_{n-1}) - y_{n-1} + h\sum_{i=1}^{k} b_{-i}[f(x_{n-i}, y(x_{n-i})) - f(x_{n-i}, y_{n-i})] =$$

$$\rho_n + R_{n-1} + h\sum_{i=1}^{k} b_{-i} \cdot f'_y(x_{n-i}, \widetilde{y}_{n-i}) \cdot (y(x_{n-i}) - y_{n-i}) =$$

$$\rho_n + R_{n-1} + h\sum_{i=1}^{k} b_{-i} \cdot f'_y(x_{n-i}, \widetilde{y}_{n-i}) \cdot R_{n-i}$$

$$|R_n| \leq |\rho_n| + |R_{n-1}| + h\sum_{i=1}^{k} |b_{-i}| \cdot |f'_y(x_{n-i}, \widetilde{y}_{n-i})| \cdot |R_{n-i}| \leq$$

$$|\rho_n| + |R_{n-1}| + h\sum_{i=1}^{k} |b_{-i}| \cdot L \cdot |R_{n-i}|$$

Uvedimo u razmatranje niz brojeva $\{g_n\}_{n=0}^{N}$ sljedećom jednakošću:

$$g_n = \max(|R_0|, |R_1|, \ldots, |R_n|), \quad n \geq k.$$

Jasno, $|R_n| \leq g_n$. Jasno, uvedeni niz brojeva je neopadajući. Možemo reći da uvedeni brojevi odražavaju "rastući" oblik greške i da majoriraju grešku. Dalje imamo:

$$|R_n| \leq |\rho_n| + g_{n-1} + h\sum_{i=1}^{k} |b_{-i}| \cdot L \cdot g_{n-1}$$

$$|R_n| \leq |\rho_n| + g_{n-1} + h \cdot B \cdot L \cdot g_{n-1}$$

$$g_n \leq |\rho_n| + g_{n-1}(1 + hBL) \text{ za } n = k, k+1, \ldots, N. \tag{1}$$

Uvedena je oznaka $B = \sum_{i=1}^{k} |b_{-i}|$.

Odranije raspolažemo sa sljedećim relacijama:

$$g_0 = 0,\ g_1 = \ldots = g_{k-1} = C_0 h^{k+1} \ \text{ i } \ |\rho_n| \leq C_1 h^{k+1} \text{ za } n = k, k+1, \ldots, N. \tag{2}$$

Rekurentne formule (1) i (2) primjenjuju se uzastopno u nastavku.

Pisaćemo kao da je $C_1 \geq C_0$. Imamo redom:

$$g_k \leq |\rho_k| + (1 + hBL)g_{k-1} \leq C_1 h^{k+1} + (1 + hBL)C_0 h^{k+1} \leq$$

$$C_1 h^{k+1} + (1 + hBL)C_1 h^{k+1} = C_1 h^{k+1}[1 + (1 + hBL)]$$

$$g_{k+1} \leq |\rho_{k+1}| + (1 + hBL)g_k \leq C_1 h^{k+1} + (1 + hBL) \cdot C_1 h^{k+1}[1 + (1 + hBL)] =$$

$$C_1 h^{k+1}[1 + (1 + hBL) + (1 + hBL)^2]$$

slično $g_{k+2} \leq C_1 h^{k+1}[1 + (1 + hBL) + (1 + hBL)^2 + (1 + hBL)^3]$ itd.

$$g_N \leq C_1 h^{k+1}[1 + (1 + hBL) + (1 + hBL)^2 + \ldots + (1 + hBL)^{N-k+1}]$$

$$g_N \leq C_1 h^{k+1}[1 + (1 + hBL) + \ldots + (1 + hBL)^{N-1}] = \text{ zbir geometrijske progresije}$$

$$C_1 h^{k+1} \cdot \frac{1}{(1 + hBL) - 1} \cdot [(1 + hBL)^N - 1] \leq C_1 h^{k+1} \cdot \frac{1}{hBL} \cdot (1 + hBL)^N$$

$$\text{poznato je } \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = e, \ \ \forall n \ \left(1 + \frac{1}{n}\right)^n < e$$

$$\text{poznato je } \lim_{n\to\infty} \left(1 + \frac{a}{n}\right)^n = e^a, \ \ \forall n \ \left(1 + \frac{a}{n}\right)^n < e^a, \ \ a > 0$$

Mi imamo $g_N \leq C_1 h^{k+1} \cdot \frac{1}{hBL} \cdot e^{XBL} = Ch^k$

$$\text{jer je } (1+hBL)^N = \left(1+\frac{XBL}{N}\right)^N \le e^{XBL}$$

Mi imamo definitivno $|R_N| \le Ch^k$

Formulišimo teoremu koju smo upravo dokazali.

Teorema. Neka su ispunjeni sljedeći uslovi. a) Neka $y = y(x)$ označava bilo koje rješenje d. j. $y' = f(x,y)$. Za svaku funkciju $y = y(x)$ važi $y \in C^{k+1}[x_0, x_0+X]$. Sve funkcije $y^{(k+1)} = y^{(k+1)}(x)$ su ravnomjerno ograničene na odsječku $[x_0, x_0+X]$. b) Funkcija $f'_y(x,y)$ je neprekidna i ograničena na skupu $\Omega = [x_0, x_0+X]\times(-\infty,+\infty)$. c) Startne približne vrijednosti $y_1,\ldots,y_{k-1}$ dobijene su po nekoj šemi tipa Runge–Kuta čiji red lokalne greške iznosi h^{k+1}. Tada važi nejednakost $|R_N| \le Ch^k$. Ovdje je C neka konstanta, tj. C ne zavisi od h.

Važe nejednakosti $|R_n| \le Ch^k$, gdje je $0 \le n \le N$.

2. Slučaj implicitne Adamsove metode

Razmotrimo implicitnu Adamsovu šemu:

$$y_n = y_{n-1} + h\sum_{i=0}^{k-1} b_{-i} f(x_{n-i}, y_{n-i}), \quad n = \overline{k-1, N}. \tag{3}$$

Biće opet: znamo da je ovdje lokalna greška reda h^{k+1}, a vidjećemo da je ovdje greška reda h^k. Izvođenje je slično i iskaz teoreme je sličan, pa ćemo iznijeti bez dokaza. Ipak, pojavljuje se jedna razlika, kako slijedi. Kod eksplicitne šeme u iskazu teoreme je bilo: neka je N ma kakav prirodan broj i neka je onda $h = X/N$. Sada u iskazu teoreme imamo: neka je N dovoljno veliki prirodan broj i neka je onda $h = X/N$. Ili: neka je broj $h > 0$ dovoljno mali. Ovim se postiže sljedeća stvar: (nelinearna) jednačina (3) ima jedinstveno rješenje, za svako n.

$(\exists h_0 > 0) \quad (\forall h \le h_0) \quad$ postoje jedinstveni brojevi $\{y_n\} \quad \& \quad$ važi $|R_n| \le Ch^k$

Ima još jedna okolnost karakteristična za implicitnu šemu. (Nelinearna) jednačina (3) po nepoznatoj y_n u praksi bude riješena samo približno tačno, po pravilu. Ako se ona svaki put tj. za svako n riješi sa greškom do C_2h^{k+1}, gdje je C_2 neka konstanta, onda iskaz teoreme ostaje da važi.

3. Opšti slučaj diferencne metode ili slučaj diferencne metode koja ne mora da bude Adamsovog tipa

Razmotrimo šemu:

$$\sum_{i=0}^{k} a_{-i} y_{n-i} - h\sum_{i=0}^{k} b_{-i} f(x_{n-i}, y_{n-i}) = 0, \quad a_0 \ne 0, \quad n = \overline{k, N}. \tag{4}$$

Teorema. a) Neka lokalna greška ima red veličine h^{s+1}. b) $|f'_y(x,y)| \le L$ za $(x,y) \in \Omega$. c) Red veličine greške brojeva $y_1,\ldots,y_{k-1}$ jednak je h^{s+1}. d) I neka je ispunjen uslov α, v, niže. Tada je red veličine greške šeme (4) jednak h^s. $(\forall h \le h_0)$

Ovu teoremu navodimo bez dokaza, a samo ćemo razjasniti šta je to "uslov α".

Razmotrimo jednačinu po nepoznatoj $\alpha \in C$:

$$a_0\alpha^k + a_{-1}\alpha^{k-1} + \ldots + a_{-k} = 0,$$

ovo je tzv. karakteristična jednačina. Sagledajmo sva njena rješenja (sve njene korijene) u kompleksnoj α–ravni. Uslov α ili uslov korijena glasi kako slijedi: svi korijeni su po modulu ≤ 1 i svi korijeni koji su po modulu $= 1$ su prosti (jednostruki).

Primjer. Šema

$$y_n = -4y_{n-1} + 5y_{n-2} + h(4f_{n-1} + 2f_{n-2})$$

ima lokalnu grešku $O(h^4)$. Njena karakteristična jednačina glasi $\alpha^2 + 4\alpha - 5 = 0$, korijeni jednačine su $\alpha_1 = 1$ i $\alpha_2 = -5$, tako da uslov α nije ispunjen. U nastavku ćemo pokazati da tu šemu ne treba upotrebljavati u praksi. Mi ćemo analizirati prostiranje greške ulaznih podataka. Izaćiće na vidjelo nestabilnost šeme.

Razmotrimo d. j. $y' = 0$. Sada je $f = 0$ pa $y_n = -4y_{n-1} + 5y_{n-2}$. Razmotrimo rješenje te d. j. $y = y(x) = 0$. Smatramo da su y_0 i y_1 ulazni podaci. Uzmimo da je $y_0 = \delta$ i $y_1 = -5\delta$, gdje je $\delta > 0$. Iz $y_n = -4y_{n-1} + 5y_{n-2}$ imamo redom $y_2 = (-5)^2\delta$, $y_3 = (-5)^3\delta$, ... Vidimo da greška neograničeno raste. Dakle, greška ulaznih podataka se nadlinearno prenosi ili odražava na grešku približnih vrijednosti. Drugim riječima, šema nije stabilna u odnosu na približnost ulaznih podataka.

Uz ovaj primjer. Razmotrimo homogenu diferencnu jednačinu drugog reda sa konstantnim koeficijentima

$$ax_n + bx_{n-1} + cx_{n-2} = 0.$$

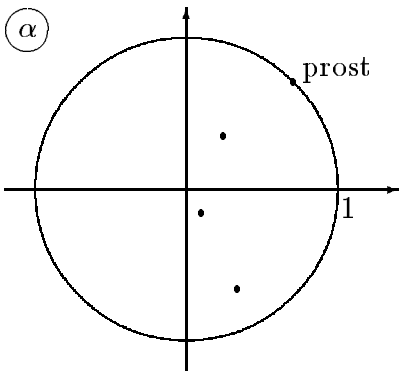
Njeno opšte rješenje je

$$x_n = C_1 \cdot \lambda_1^n + C_2 \cdot \lambda_2^n,$$

gdje su λ_1 i λ_2 korijeni tzv. karakteristične jednačine $a\lambda^2 + b\lambda + c = 0$; ovo ako je $\lambda_1 \neq \lambda_2$. A ako je $\lambda_1 = \lambda_2$ onda je opšte rješenje

$$x_n = C_1 \cdot \lambda_1^n + C_2 \cdot n \cdot \lambda_1^n.$$

Ovdje su C_1 i C_2 proizvoljne konstante. One mogu da budu određene, odnosno partikularno rješenje može da bude izdvojeno ako je definisan par početnih uslova: $x_1 = x_{10}$, $x_2 = x_{20}$, gdje su $x_{10} \in R$ i $x_{20} \in R$ dati brojevi.



Slika za 3. Opšta diferencna

Napomena za buduće gradivo. Kasnije ćemo raditi Milnovu p.–k. metodu, tako da se izražava preko dvije formule:

$$\text{prva formula } \; y_n = y_{n-4} + h(\ \dots\ )$$

znači $\alpha^4 - 1 = 0$, slijedi $\alpha_{1,2} = \pm 1$, $\alpha_{3,4} = \pm i$, tako da je uslov α ispunjen

$$\text{i druga formula } \; y_n = y_{n-2} + h(\,\dots\,)$$

znači $\alpha^2 - 1 = 0$, slijedi $\alpha_{1,2} = \pm 1$, tako da je uslov α ispunjen.

5.9. ADAMSOVA METODA ČETVRTOG REDA

Adamsova p.–k. šema IV reda izražava se pomoću sljedeće dvije formule:

$$\begin{cases} y_n^{\text{pred}} = y_{n-1} + \dfrac{h}{24}\Big(55f(x_{n-1}, y_{n-1}) - 59f(x_{n-2}, y_{n-2}) + 37f(x_{n-3}, y_{n-3}) - 9f(x_{n-4}, y_{n-4})\Big) \\ y_n = y_{n-1} + \dfrac{h}{24}\Big(9f(x_n, y_n^{\text{pred}}) + 19f(x_{n-1}, y_{n-1}) - 5f(x_{n-2}, y_{n-2}) + f(x_{n-3}, y_{n-3})\Big) \end{cases}$$

za $n = 4, 5, \dots, N$. Njena lokalna greška ρ_n je reda h^5, a njena greška $R_n = y(x_n) - y_n$ je reda h^4. Za lokalnu grešku ρ_n' Adamsove e. šeme IV reda važi:

$$\rho_n' = \frac{1}{4!} C' \cdot y^V(\xi') \cdot h^5 = \frac{1}{4!} \cdot \frac{251}{30} y^V(\xi') \cdot h^5, \;\; x_{n-4} < \xi' < x_n. \tag{1}$$

A za lokalnu grešku ρ_n razmatrane p.–k. šeme važi:

$$\rho_n = \frac{1}{4!} C \cdot y^V(\xi) \cdot h^5 = \frac{1}{4!}\Big(-\frac{19}{30}\Big) y^V(\xi) \cdot h^5, \;\; x_{n-3} < \xi < x_n. \tag{2}$$

Dosad rečeno rađeno je ranije u naslovima 5.6. i 5.8. a sada će biti korišćeno. Zadržavamo ranije oznake. Primjera radi: $h = X/N$ i $y_n = y_n^{\text{kor}}$. U ovom naslovu samo će biti izveden efektivni izraz za ρ_n.

Dva broja y_n^{pred} i y_n izračunati su oslanjanjem na jednu te istu prethodnu približnu vrijednost y_{n-1}. Ta dva broja se razlikuju zato što se razlikuju njihove lokalne greške ρ_n' i ρ_n. Može se pokazati da su $y_n^{\text{pred}} - y_n$ i $\rho_n - \rho_n'$ ekvivalentne beskonačno male kad $h \to 0$. Mi pišemo

$$y_n^{\text{pred}} - y_n \approx \rho_n - \rho_n'. \tag{3}$$

Vidimo da $\xi' \to \xi$ kad $h \to 0$, a mi ćemo to koristiti u obliku $y^V(\xi') \approx y^V(\xi)$.

U nastavku se još samo primijene sredstva aritmetike na formule (1)–(3):

$$y_n^{\text{pred}} - y_n \approx \rho_n - \rho_n' \approx -\frac{19}{720} y^V(\xi) h^5 - \frac{251}{720} y^V(\xi) h^5 = -\frac{270}{720} y^V(\xi) h^5 = \frac{270}{19} \rho_n$$

$$\text{digitron pokazuje da je } \;\; \frac{270}{19} = 14{,}21 \approx 14$$

Konačno,

$$\rho_n \approx \frac{1}{14}\Big(y_n^{\text{pred}} - y_n\Big).$$

5.10. ALGORITAM ZASNOVAN NA DIFERENCNOJ METODI

Razmatra se početni zadatak $y' = f(x, y)$, $x_0 \le x \le x_0 + X$, $y(x_0) = y_0$.

Samo ćemo naznačiti elemente koji čine algoritam. Zadatak je definisan sljedećim parametrima: $f = f(x,y)$, x_0, y_0 i X. Neka smo se opredijelili za jednu određenu diferencnu metodu. Biće izložena dva načina za praktičnu ocjenu greške.

Prvi način. Sprovedu se dva odvojena proračuna. Neka su $z_h(x_0+X)$ i $z_{2h}(x_0+X)$ dvije približne vrijednosti za $y(x_0+X)$ dobijene redom sa korakom h odnosno $2h$. Označimo sa $R_h(x_0+X)$ i $R_{2h}(x_0+X)$ odgovarajuće greške:

$$y(x_0+X) = z_h(x_0+X) + R_h(x_0+X) \text{ i } y(x_0+X) = z_{2h}(x_0+X) + R_{2h}(x_0+X).$$

Uzmimo da je red greške diferencne metode koja se primjenjuje jednak h^4.

Između jedne i druge greške postoji sljedeća veza:

$$\lim_{h\to 0} \frac{R_{2h}(x_0+X)}{R_h(x_0+X)} = 16 \text{ ili } R_{2h}(x_0+X) \approx 16 R_h(x_0+X).$$

Na običan Rungeov način, lako se dobija sljedeća procjena greške približnog broja $z_h(x_0+X)$:

$$R_h(x_0+X) \approx \frac{1}{15}\Big(z_h(x_0+X) - z_{2h}(x_0+X)\Big).$$

Drugi način. Izvrši se jedan proračun, a greška se procjenjuje preko lokalnih greški. Koristimo obične oznake $Nh = X$ i $x_n = x_0 + nh$. Za grešku važi:

$$R_n = \sum_{k=1}^{n} \rho_k \exp\{(x_n - x_k) f'_y(\xi_k, \eta_k)\} \text{ ili } |R_n| \le \sum_{k=1}^{n} |\rho_k| \exp\{(x_n - x_k)L\}, \text{ za razne } n.$$

5.11. MILNOVA METODA

Neka bude $Nh = X$ i $x_n = x_0 + nh$. Milnova metoda definisana je sljedećom šemom za računanje:

$$\begin{cases} y_n^{\text{pred}} = y_{n-4} + \dfrac{4h}{3}\Big(2f(x_{n-3}, y_{n-3}) - f(x_{n-2}, y_{n-2}) + 2f(x_{n-1}, y_{n-1})\Big) \\ y_n = y_{n-2} + \dfrac{h}{3}\Big(f(x_{n-2}, y_{n-2}) + 4f(x_{n-1}, y_{n-1}) + f(x_n, y_n^{\text{pred}})\Big), \quad n = 4, 5, \ldots, N \end{cases}$$

Prva formula može da se dobije kada se u identitetu $\int_{x_{n-4}}^{x_n} y'(x)dx = \int_{x_{n-4}}^{x_n} f(x, y(x))dx$ funkcija $F(x) = f(x, y(x))$ zamijeni svojim interpolacionim polinomom po tačkama $x = x_{n-4}$, $x = x_{n-3}$, $x = x_{n-2}$ i $x = x_{n-1}$. Kada se izračuna integral od polinoma onda će se kao jedan sabirak dobiti $0 \cdot F(x_{n-4})$. Druga formula može da se dobije kada se napiše $\int_{x_{n-2}}^{x_n} y'(x)dx = \int_{x_{n-2}}^{x_n} f(x, y(x))dx$ i onda se $f(x, y(x))$ zamijeni svojim interpolacionim polinomom koji se odnosi na tačke x_{n-2}, x_{n-1}, x_n i x_{n+1}. Kada integral od polinoma onda će se kao jedan sabirak dobiti $0 \cdot f(x_{n+1}, y(x_{n+1}))$.

Lokalna greška Milnove metode je reda h^5, a njena greška R_n je reda h^4.

Za lokalnu grešku ρ_n važi sljedeća praktična procjena:

$$\rho_n \approx \frac{1}{29}\Big(y_n^{\text{pred}} - y_n\Big).$$

6. NUMERIČKE METODE ZA RJEŠAVANJE GRANIČNOG ZADATKA ZA OBIČNE DIFERENCIJALNE JEDNAČINE

Razmatraćemo dvije vrste metoda. a) Diferencne metode, drukčije se kaže – metoda konačnih razlika. b) Varijacione metode, za neke od njih kaže se – metoda konačnih elemenata. Jednu i drugu vrstu opisaćemo u slučaju diferencijalne jednačine drugog reda.

6.1. METODA KONAČNIH RAZLIKA

Biće konstruisana numerička metoda za dobijanje približnog rješenja graničnog zadatka drugog reda. Biće dokazano da numerička metoda ima tzv. red aproksimacije h^2 i da je ona stabilna u odnosu na svoju desnu stranu i u odnosu na svoja dva granična uslova. Na kraju će biti dokazano da ona konvergira, s tim da je red konvergencije (red tačnosti) takođe jednak h^2.

Prvo se daje postavka zadatka koji treba da bude numerički riješen. Razmotrimo jednačinu

$$Ly = f(x) \ \text{ tj. } \ y'' - p(x)y(x) = f(x), \ \text{ za } \ 0 \le x \le X, \tag{1}$$

i granične uslove

$$y(0) = a, \ \ y(X) = b. \tag{2}$$

Iz teorije običnih d. j. poznata su sljedeća svojstva graničnog zadatka (1)–(2). Neka $p \in C^2[0,X]$, neka $f \in C^2[0,X]$ i neka je $p(x) \ge 0$ za $x \in [0,X]$. Tada zadatak (1)–(2) ima jedinstveno rješenje $y = y(x)$ i $y \in C^4[0,X]$. Napominje se da je linearni diferencijalni operator $-L$ pozitivan u prostoru $L_2[0,X]$. To znači da važi $(-Ly, y) = \int_0^X [-y'' + p(x)y(x)]y(x)dx > 0$ za bilo koju funkciju $y = y(x)$ koja nije $\equiv 0$ i koja zadovoljava granične uslove $y(0) = 0$, $y(X) = 0$.

Neka je N prirodan broj i neka je $h = X/N$. Uvedimo ravnomjernu mrežu čvorova $x_n = nh$ za $n = \overline{0,N}$. Neka je $y = y(x)$ tačno rješenje razmatranog zadatka (1)–(2), tako da je $y(x_n)$ vrijednost tačnog rješenja u čvoru $x = x_n$. Neka su y_n odgovarajuće približne vrijednosti, biće dobijene kasnije. Uvedimo i oznaku za grešku: $R_n = y_n - y(x_n)$, za $n = \overline{0,N}$.

Neka bude

$$\ell_0(y(x_n)) = \frac{1}{h^2}\Big(y(x_{n+1}) - 2y(x_n) + y(x_{n-1})\Big)$$

i

$$r_n = \ell_0(y(x_n)) - y''(x_n) = \frac{1}{h^2}\Big(y(x_{n+1}) - 2y(x_n) + y(x_{n-1})\Big) - y''(x_n).$$

Za r_n se kaže da je greška aproksimacije. Poznata je sljedeća formula:

$$r_n = \frac{1}{12}y^{IV}(\xi_n)h^2, \ \text{ gdje je } \ x_{n-1} \le \xi_n \le x_{n+1}.$$

Tako da je greška aproksimacije reda h^2; red konzistencije je h^2. Ili $|r_n| \le \frac{1}{12}M_4h^2$, gdje je $M_4 = \max_{x\in[0,X]}|y^{IV}(x)|$, gdje je $y = y(x)$ tačno rješenje za (1)–(2). Formula $r_n = \frac{1}{12}y^{IV}(\xi_n)h^2$ rađena je ranije u okviru svojstava podijeljenih i konačnih razlika, a takođe i u okviru numeričkog diferenciranja. Uostalom, ona može da bude provjerena razvojem funkcije $y = y(x)$ po Tejlorovoj formuli. ℓ_0 aproksimira y''. Ako je $y = y(x)$ linearna funkcija onda je $\ell_0 = 0$. Ako je $y(x) = ax^2 + bx + c$ onda je $\ell_0(y(x_n)) = 2a$.

Neka bude $p_n = p(x_n)$ i $f_n = f(x_n)$. Napišimo jednačinu (1) kada je $x = x_n$:

$$y''(x_n) - p_n y(x_n) = f_n.$$

Uvedimo oznaku

$$\ell(y(x_n)) = \frac{1}{h^2}\Big(y(x_{n+1}) - 2y(x_n) + y(x_{n-1})\Big) - p_n y(x_n).$$

Imamo da je

$$\ell(y(x_n)) = f_n + r_n.$$

ℓ aproksimira L ili svejedno $Ly(x_n) \approx \ell(y(x_n))$, sa greškom r_n.

U numeričkoj metodi, drugi izvod $y''(x)$ u čvoru $x = x_n$ biva zamijenjen podijeljenom razlikom koja je sastavljena od efektivno poznatih brojeva $\{y_n\}_{n=0}^N$ (biće dobijeni kasnije) tj. od y_{n-1}, y_n i y_{n+1}, a ne od brojeva $y(x_{n-1})$, $y(x_n)$ i $y(x_{n+1})$. Mi ćemo saznati brojeve $\{y_n\}_{n=0}^N$ ako riješimo sljedeći sistem linearnih jednačina:

$$\frac{1}{h^2}\Big(y_{n+1} - 2y_n + y_{n-1}\Big) - p_n y_n = f_n \ \text{ ili } \ \ell(y_n) = f_n, \ \ n = \overline{1, N-1}, \tag{3}$$

$$y_0 = a, \ \ y_N = b. \tag{4}$$

Mali primjer za metodu konačnih razlika: $y'' - \sqrt{x}y = e^x$, $y(0) = 1$, $y(1) = 2$. Izaberimo $N = 10$. Tada sistem linearnih jednačina glasi $y_0 = 1$, $(y_2 - 2y_1 + y_0)/h^2 - \sqrt{0{,}1}y_1 = e^{0{,}1}$, $(y_3 - 2y_2 + y_1)/h^2 - \sqrt{0{,}2}y_2 = e^{0{,}2}$, ..., $y_{10} = 2$. Kada riješimo sistem, onda i dobijamo numerički rezultat $\{y_n\}_{n=0}^{10}$.

Imamo da je $\ell(y(x_n)) = f_n + r_n$ i $\ell(y_n) = f_n$. Oduzimanjem, uzimajući u obzir linearnost diferencnog operatora ℓ,

$$\ell(y_n) - \ell(y(x_n)) = f_n - (f_n + r_n) \ \text{ ili } \ \ell(y_n - y(x_n)) = -r_n \ \text{ ili } \ \ell(R_n) = -r_n, \ \ n = \overline{1, N-1},$$

veza greške metode R_n i greške aproksimacije r_n.

Biće dokazana sljedeća teorema.

Teorema. Sistem linearnih jednačina (3)–(4) ima jedinstveno rješenje $\{y_n\}_{n=0}^N$ i važi sljedeća formula (za ocjenu greške)

$$\max_{0 \le n \le N} |R_n| \le \frac{1}{96} X^2 M_4 h^2.$$

Mi ćemo ustvari dokazati nešto opštiju teoremu koja se odnosi na nešto opštiju situaciju. Prethodna teorema analizira samo grešku metode. Sljedeća teorema uzima u obzir i grešku računanja i grešku izazvanu približnošću ulaznih podataka; ulazni podaci su a i b. Neka veličine $\{y_n\}_{n=0}^N$ više ne zadovoljavaju sistem (3)–(4) nego odsad uzimamo da one zadovoljavaju sljedeći sistem:

$$\frac{1}{h^2}\Big(y_{n+1} - 2y_n + y_{n-1}\Big) - p_n y_n = f_n + \delta_n \ \text{ ili } \ \ell(y_n) = f_n + \delta_n, \ \ n = \overline{1, N-1}, \tag{5}$$

$$y_0 = a + R_0, \ \ y_N = b + R_N. \tag{6}$$

(5)–(6) bi se očito svelo na (3)–(4) da je $\delta_n = 0$ i $R_0 = R_N = 0$. Zašto su uvedeni δ_n? Kada se sistem linearnih jednačina riješi onda se njegovo rješenje radi provjere uvrsti u sami taj sistem. Lijeva i desna strana se ne poklope već se razlikuju za δ_n. Razlikuju se zato što se tokom rješavanja sistema nakupila greška računanja. Još, brojevi δ_n odgovaraju i slučaju kada je desna strana jednačine $f = f(x)$ poznata samo približno, poznata sa nekom greškom. Zašto

su potrebni R_0 i R_N? Moguće je da su brojevi a i b koji definišu par graničnih uslova samo približno poznate veličine.

Za mjeru greške računanja uzećemo $\max_{0<n<N}|\delta_n|$. Za mjeru greške ulaznih podataka uzećemo $\max(|R_0|,|R_N|)$. I dalje je $R_n = y_n - y(x_n)$. Sada R_n odražava sve tri komponente greške (metode, računanja i od ulaznih podataka).

Ranija veza $\ell(R_n) = -r_n$ sada u novoj situaciji očito postaje

$$\ell(R_n) = -r_n + \delta_n, \quad n = \overline{1, N-1}.$$

Sada u novoj situaciji važi sljedeća teorema (koja će biti dokazana).

Teorema. Sistem linearnih jednačina (5)–(6) ima jedinstveno rješenje $\{y_n\}_{n=0}^N$ i važi sljedeća formula (za ocjenu greške)

$$\max_{0\le n\le N}|R_n| \le \frac{1}{96}X^2M_4h^2 + \frac{1}{8}X^2\max_{0<n<N}|\delta_n| + \max(|R_0|,|R_N|).$$

Dokaz teoreme. Sistem linearnih jednačina (5)–(6) sastoji se od $N+1$ jednačina i ima $N+1$ nepoznatih $\{y_n\}_{n=0}^N$. Možemo pomnožiti sa h^2 jednačine od $n=1$ do $n=N-1$. Označimo sa M matricu sistema, ona je oblika $(N+1)\times(N+1)$. Vidimo da je matrica M trodijagonalna. Prvo pitanje: pokazaćemo da je matrica M regularna tj. da je $\det M \neq 0$; koristićemo sljedeće: $p(x)\ge 0$ za svako $x\in[0,X] \Rightarrow p_n \ge 0$ za svako $n\in\{1,\dots,N-1\}$. Zato sistem ima jedinstveno rješenje. Napišimo sistem (5)–(6) i napišimo matricu M:

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & \dots & & & \\ 1 & -2-p_1h^2 & 1 & 0 & \dots & & & \\ 0 & 1 & -2-p_2h^2 & 1 & & & & \\ \dots & \dots & \dots & & & & & \\ 0 & 0 & 0 & 0 & \dots & 1 & -2-p_{N-1}h^2 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{bmatrix}}_{=M} \cdot \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} a+R_0 \\ h^2(f_1+\delta_1) \\ h^2(f_2+\delta_2) \\ \dots \\ h^2(f_{N-1}+\delta_{N-1}) \\ b+R_N \end{bmatrix}$$

Posmatrajmo matricu $N = [n_{ij}]_{i,j=1}^{N-1}$, gdje je $n_{ii} = -2$ i $n_{i,i-1} = n_{i,i+1} = 1$, a inače je $n_{ij} = 0$. Posmatrajmo proces svođenja matrice N na gornji trougaoni oblik. Tako vidimo da je $\det N \neq 0$ (ustvari je $\det N = (-1)^{N+1}\cdot N$). Vidimo da je tokom istog svođenja situacija sa M jednaka ili povoljnija od N. Dakle, $\det M \neq 0$. Ili: iz kasnijih eksplicitnih formula (10) za rješavanje trodijagonalnog sistema vidi se da je $\det M \neq 0$. Napominje se sljedeće: ako bi bilo $p_n > 0$ za svako $n\in\{1,\dots,N-1\}$ onda bi matrica M bila dijagonalno dominantna; znamo da iz dijagonalne dominantnosti matrice slijedi njena regularnost. Prvo pitanje je završeno.

Mi raspolažemo sa

$$\ell(R_n) = -r_n + \delta_n, \quad n = \overline{1, N-1}. \tag{7}$$

Bilo bi bolje da raspolažemo sa $R_n = \dots$. Kao da bi se na relaciju $\ell(R_n) = -r_n + \delta_n$ primijenio operator ℓ^{-1}. O tome je ustvari riječ u nastavku.

Mi raspolažemo i sa

$$|r_n| \le \frac{1}{12}M_4h^2, \quad n = \overline{1, N-1}. \tag{8}$$

Dokažimo dvije leme.

Lema 1. Neka bude $p_n \geq 0$ za $n = \overline{1, N-1}$. Neka bude $\ell(z_n) = \frac{1}{h^2}\Big(z_{n+1} - 2z_n + z_{n-1}\Big) - p_n z_n$. Neka $\{z_n\}_{n=0}^N$ bude ma kakav (konačan) niz brojeva. Ako je $\ell(z_n) \leq 0$ za $n = \overline{1, N-1}$ i $z_0 \geq 0$, $z_N \geq 0$ onda je $z_n \geq 0$ za $n = \overline{1, N-1}$.

(Uporediti sa: $y(0) = 0$, $y(1) = 0$, $y'' \leq 0 \Rightarrow y(x) \geq 0$)

Dokaz. Uvedimo oznaku $d = \min_{0 \leq n \leq N} z_n$ i dopustimo da je $d < 0$. Za koje n važi $z_n = d$? Ne može biti $z_0 = d$ niti $z_N = d$. Neka je q najmanji cio broj za koji je $z_q = d$. Imamo $z_{q-1} > d$ i $z_{q+1} \geq d$. Tako da je $z_{q+1} - 2z_q + z_{q-1} > 0$. Još, $p_q \geq 0$ i $z_q = d < 0 \Rightarrow -p_q z_q \geq 0$. Sabiranjem

$$\ell(z_q) = \frac{1}{h^2}\Big(z_{q+1} - 2z_q + z_{q-1}\Big) - p_q z_q > 0.$$

Po uslovu leme je $\ell(z_q) \leq 0$, tako da smo dobili kontradikciju. Dakle, ne može biti $d < 0$, nego mora da bude $d \geq 0$. Lema je dokazana.

Lema 2. Neka bude $p_n \geq 0$ za $n = \overline{1, N-1}$. Neka bude $\ell(z_n) = \frac{1}{h^2}\Big(z_{n+1} - 2z_n + z_{n-1}\Big) - p_n z_n$. Neka $\{z_n\}_{n=0}^N$ bude ma kakav (konačan) niz brojeva. Važi nejednakost

$$\max_{0 \leq n \leq N} |z_n| \leq \max(|z_0|, |z_N|) + \frac{1}{8} X^2 Z, \quad \text{gdje je} \quad Z = \max_{0 < n < N} |\ell(z_n)|.$$

(Uporediti sa: $y(0) = 0$, $y(1) = 0 \Rightarrow |y(x)| \leq \frac{1}{8} \max_{0 \leq x \leq 1} |y''(x)|$. Znak jednakosti se dostiže za $y(x) = x(1-x)$)

Dokaz. Uvedimo u razmatranje niz brojeva

$$\omega_n = |z_0| \frac{X - nh}{X} + |z_N| \frac{nh}{X} + \frac{1}{2} Z (X - nh) nh.$$

Iz eksplicitnog izraza za ω_n je $\omega_n \geq 0$. Neposrednim računom nalazimo da je $\frac{1}{h^2}\Big(\omega_{n+1} - 2\omega_n + \omega_{n-1}\Big) = -Z$, tako da je $\ell(\omega_n) = -Z - p_n \omega_n \leq -Z$. Dalje, za $n = \overline{1, N-1}$ imamo $\ell(\omega_n \pm z_n) = \ell(\omega_n) \pm \ell(z_n) \leq -Z \pm \ell(z_n) \leq 0$. Pored toga, $\omega_0 \pm z_0 = |z_0| \pm z_0 \geq 0$ i $\omega_N \pm z_N = |z_N| \pm z_N \geq 0$. Prema tome, niz brojeva $\{\omega_n \pm z_n\}_{n=0}^N$ zadovoljava sve uslove prethodne leme. Zato imamo $\omega_n \pm z_n \geq 0$ za $n = \overline{0, N}$. Napišimo odvojeno: $\omega_n + z_n \geq 0$ i $\omega_n - z_n \geq 0$. Znači da je $|z_n| \leq \omega_n$. Slijedi da je $|z_n| \leq \max_{0 \leq n \leq N} \omega_n$.

Ostaje samo da se izračuna $\max_{0 \leq n \leq N} \omega_n$. Imamo:

$$|z_0| \frac{X - nh}{X} + |z_N| \frac{nh}{X} \leq \max(|z_0|, |z_N|) \frac{X - nh}{X} + \max(|z_0|, |z_N|) \frac{nh}{X} = \max(|z_0|, |z_N|)$$

$$\frac{1}{2} Z (X - nh) nh \leq \frac{1}{2} Z \cdot \frac{1}{4} X^2, \quad \text{jer je} \quad x(1-x) \leq \frac{1}{4} \quad \text{za} \quad x \in [0, 1]$$

$$\text{sabiranjem,} \quad \omega_n \leq \max(|z_0|, |z_N|) + \frac{1}{2} Z \cdot \frac{1}{4} X^2$$

Znači da je $\max_{0 \leq n \leq N} \omega_n \leq \max(|z_0|, |z_N|) + \frac{1}{2} Z \cdot \frac{1}{4} X^2$. Lema je dokazana, jer $\forall n \ |z_n| \leq$ izraz $\Rightarrow \max_{0 \leq n \leq N} |z_n| \leq$ izraz.

Lema 2 govori sljedeće. Neka $\{z_n\}_{n=0}^N$ predstavlja rješenje diferencnog graničnog zadatka $\ell(z_n) = f_n$ za $0 < n < N$, $z_0 = a$, $z_N = b$. Tada je ispunjen tzv. uslov stabilnosti tj. tada važi nejednakost $\|z\| \leq C_1 \|f\| + C_2 \|g\|$. Ovdje je $z = (z_0, z_1, \ldots, z_N)$, $f = (f_1, f_2, \ldots, f_{N-1})$ i $g = (a, b)$. Ima se u vidu max–norma.

Slijedi završni dio dokaza teoreme. Primijenimo lemu 2 na niz brojeva $\{R_n\}_{n=0}^N$:

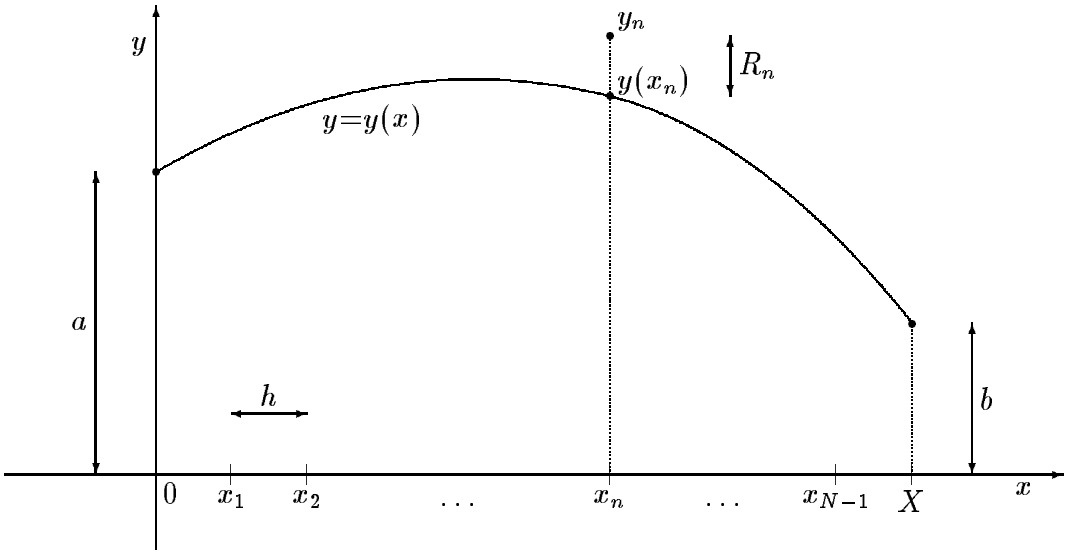
$$\max_{0\le n\le N}|R_n| \le \max(|R_0|,|R_N|) + \frac{1}{8}X^2 \max_{0<n<N}|\ell(R_n)| = \quad \text{po (7)}$$

$$\max(|R_0|,|R_N|) + \frac{1}{8}X^2 \max_{0<n<N}|-r_n+\delta_n| \le$$

$$\max(|R_0|,|R_N|) + \frac{1}{8}X^2 \max_{0<n<N}|r_n| + \frac{1}{8}X^2 \max_{0<n<N}|\delta_n| \le \quad \text{po (8)}$$

$$\max(|R_0|,|R_N|) + \frac{1}{8}X^2\cdot\frac{1}{12}M_4h^2 + \frac{1}{8}X^2 \max_{0<n<N}|\delta_n|$$

Teorema je dokazana.



U nastavku – dopune.

Izvršimo sada dogradnju konstruisane numeričke metode, da dobijemo potpuno i moćno sredstvo za rješavanje graničnog zadatka.

1. Posmatrajmo samo grešku metode, tj. neka je $\delta_n = 0$ i $R_0 = R_N = 0$. Red greške je h^2, greška $\approx Ch^2$. Na običan Rungeov način dobija se praktični izraz za grešku. Dakle, neka je x tačka sa odsječka $[0, X]$ koja je čvor i po mreži sa korakom h i po mreži sa korakom $2h$. Neka su $z_h(x)$ i $z_{2h}(x)$ dvije odgovarajuće približne vrijednosti. Tada je

$$\text{greška}(z_h(x)) = R_h(x) = y(x) - z_h(x) \approx \frac{1}{3}\Big(z_h(x) - z_{2h}(x)\Big). \tag{9}$$

2. Navedimo i gotove formule za rješavanje trodijagonalnog sistema. Formule i oznake su preuzete iz knjige Tihonov–Samarski. Te formule izražavaju tzv. metodu progona. Napominje se da je vremenski trošak za rješavanje punog sistema linearnih jednačina jednak $O(N^3)$, a trodijagonalnog sistema jednak $O(N)$; N – dimenzija sistema.

$$A_iy_{i-1} - C_iy_i + B_iy_{i+1} = -F_i,\ 0<i<N;\ \ y_0 = \chi_1y_1 + \nu_1,\ y_N = \chi_2y_{N-1} + \nu_2$$

$$\xi_N = \chi_2,\ \eta_N = \nu_2;\ \ \xi_i = \frac{A_i}{C_i - B_i\xi_{i+1}},\ \eta_i = \frac{B_i\eta_{i+1} + F_i}{C_i - B_i\xi_{i+1}},\ i = N-1,\dots,1$$

$$y_0 = \frac{\nu_1 + \chi_1\eta_1}{1 - \chi_1\xi_1};\ \ y_{i+1} = \xi_{i+1}y_i + \eta_{i+1},\ i = 0, \ldots, N-1 \tag{10}$$

Ipak, treba reći da su navedene formule prilično nepraktične i da su nepotrebne–suvišne.

3. Algoritam. Formirati matricu M i riješiti trodijagonalni sistem po formulama (10) ili Gausovom metodom eliminacije, čime se dobijaju približne vrijednosti, sa izabranim korakom. Izvršiti i pomoćni grubi proračun sa dvostrukim korakom, primijeniti (9), tj. dobiti ocjenu greške.

Primjer.

Ako je $N = 6$ onda sistem linearnih jednačina može da glasi recimo:

$$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 1 & -2 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{bmatrix}$$

Nisu potrebne formule (10) da bi se sistem riješio. Sistem se lako rješava.

U nastavku – naknadne dopune.

1. Ako se ukine uslov $p(x) \geq 0$ onda u iskazu teoreme treba dodati: za dovoljno male $h > 0$.

2. Uslovima (2) odgovaraju jednačine (4). Ako bi se (2) zamijenilo sa $y'(0) = a$, $y'(X) = b$ onda bi (4) trebalo zamijeniti sa $\dfrac{y_1 - y_0}{h} = a$, $\dfrac{y_N - y_{N-1}}{h} = b$. Slično, ako u postavci zadatka umjesto (2) figurišu granični uslovi $\alpha_0 y(0) + \alpha_1 y'(0) = a$, $\beta_0 y(X) + \beta_1 y'(X) = b$ onda u numeričkoj metodi par jednačina (4) treba zamijeniti parom jednačina

$$\alpha_0 y_0 + \alpha_1 \frac{y_1 - y_0}{h} = a,\ \beta_0 y_N + \beta_1 \frac{y_N - y_{N-1}}{h} = b.$$

3. Metoda konačnih razlika primjenjuje se za numeričko rješavanje raznih graničnih zadataka, na način koji je sličan dosad izloženom načinu za granični zadatak (1)–(2). Upravo, data diferencijalna jednačina napiše se za $x = x_n$ (x_n – čvor), a onda se $y'(x_n)$, $y''(x_n)$ i slično zamijene nekom podijeljenom razlikom. Za y' se može koristiti formula:

$$y'(x_n) = \frac{1}{2h}\Big(y(x_{n+1}) - y(x_{n-1})\Big) + O(h^2).$$

Za y'' se može koristiti naša formula

$$y''(x_n) = \frac{1}{h^2}\Big(y(x_{n+1}) - 2y(x_n) + y(x_{n-1})\Big) + O(h^2)$$

ili eventualno formula

$$y''(x_n) = \frac{1}{12h^2}\Big(-y(x_{n+2}) + 16y(x_{n+1}) - 30y(x_n) + 16y(x_{n-1}) - y(x_{n-2})\Big) + O(h^4).$$

U slučaju primjene metode konačnih elemenata, kao približno rješenje uzima se funkcija $z = z(x)$ koja predstavlja dio–po–dio polinom prvog stepena, tj. čiji je grafik izlomljena linija, s tim da zadovoljava $z(0) = a$ i $z(X) = b$.

Uzima se ona funkcija $z = z(x)$ navedenog oblika za koju se ostvaruje najmanja moguća vrijednost nelinearnog funkcionala $I(z) = \int_0^X [(z'(x))^2 + p(x)z^2(x) + 2f(x)z(x)]dx$.

Numeričke metode (Fizika) / Numerička analiza (C smjer)

Sadržaj predavanja:

Fajlovi (gsview): numerame.tex 11 pages numerb.tex 8 pages numerc.tex 9 pages numerd.tex 16 pages numere.tex 14 pages numerf.tex 6 pages numerg.tex 16 pages numerh.tex 14 pages numeri.tex 12 pages numerj.tex 6 pages. Σ = 10 files & 112 pages