Uvod u numeričku matematiku

M. Klaričić Bakula

Ožujak, 2009.

1 Uvod

Numerička analiza je disciplina koja proučava i numerički rješava matematičke probleme koji se javljaju u znanosti, tehnici, gospodarstvu, itd. Iako se sama disciplina najčešće povezuje s numeričkim metodama, treba naglasiti da bez dubljeg poznavanja samog problema kojeg rješavamo nije moguće procijeniti je li neka metoda dobra u smislu da daje zadovoljavajuće točna rješenja u dovoljno kratkom vremenskom intervalu.

O problemu kojeg želimo riješiti treba znati barem sljedeće:

- postoji li njegovo rješenje i, ako postoji, da li je jedinstveno ili ako nije koliko rješenja ima;
- kako se rješenja (ili rješenje) ponašaju kada se polazni podaci malo promijene (teorija perturbacije).

Kada se konstruira neka metoda za rješavanje danog problema, otvara se nekoliko pitanja:

- **problem konvergencije** (konvergira li niz dobivenih aproksimacija prema rješenju);
- **brzina konvergencije** (ako niz konvergira kako to radi: linearno, kvadratno, kubno,...);
- **složenost metode** (broj računskih operacija, zauzeće memorije, prijenos podataka, dohvat operanada,...);
- **točnost metode** (koliko značajnih znamenaka izračunatog rješenja je točno);
- **stabilnost metode** (o čemu ovisi točnost dobivenog rješenja, analiza grešaka zaokruživanja);
- **adaptibilnost metode** za posebna (paralelna, vektorska) računala.

2 Pomoćni rezultati iz analize

TEOREM. (*Teorem o međuvrijednosti*) Neka je funkcija $f : [a, b] \to \mathbb{R}$ neprekidna na konačnom intervalu $[a, b] \subset \mathbb{R}$. Neka su

$$m = \inf_{a \le x \le b} f(x) \quad \text{i} \quad M = \sup_{a \le x \le b} f(x)$$

infimum i supremum funkcije f. Tada za svaki realni broj $\beta \in [m, M]$ postoji realni broj $\alpha \in [a, b]$ takav da je

$$f(\alpha) = \beta.$$

Posebno, postoje realni brojevi $\underline{x}$ i $\overline{x}$ takvi da vrijedi

$$m = f(\underline{x}) \quad \text{i} \quad M = f(\overline{x}).$$

TEOREM. (*Rolleov teorem*) Neka je funkcija f neprekidna na nekom intervalu $[a, b] \subseteq \mathbb{R}$. Ako je $f(a) = f(b) = 0$ i ako f' postoji na (a, b), onda je za neki $\xi \in (a, b)$ ispunjeno

$$f'(\xi) = 0.$$

TEOREM. (*Teorem o srednjoj vrijednosti*) Neka je funkcija $f : [a, b] \to \mathbb{R}$ neprekidna na konačnom intervalu $[a, b] \subset \mathbb{R}$ i derivabilna na otvorenom intervalu (a, b) . Tada postoji barem jedna točka $\xi \in (a, b)$ takva da je

$$f(b) - f(a) = f'(\xi)(b - a).$$

TEOREM. (*Teorem o integralnoj srednjoj vrijednosti*) Neka je funkcija $w : [a, b] \to \mathbb{R}$ nenegativna i integrabilna na konačnom intervalu $[a, b] \subset \mathbb{R}$, te neka je funkcija $f : [a, b] \to \mathbb{R}$ neprekidna na $[a, b]$. Tada postoji točka $\xi \in [a, b]$ takva da vrijedi

$$\int_a^b f(x) w(x) \, \mathrm{d}x = f(\xi) \int_a^b w(x) \, \mathrm{d}x.$$

Jedan od najvažnijih alata u numeričkoj matematici je **Taylorov teorem**. On nam, naime, daje jednostavnu metodu aproksimacije funkcije pomoću polinoma. No da bismo koristili takav pristup u aproksimaciji promatrana funkcija mora biti dovoljno glatka.

TEOREM. (*Taylorov teorem*) Neka funkcija f ima neprekidne derivacije do uključivo reda $n+1$, $n \in \mathbb{N}_0$, na intervalu $[a,b] \subseteq \mathbb{R}$. Ako su $x, x_0 \in [a,b]$, onda vrijedi

$$f(x) = p_n(x) + R_{n+1}(x),$$

gdje je

$$p_n(x) = f(x_0) + \sum_{k=1}^{n} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k,$$

$$R_{n+1}(x) = \frac{1}{n!} \int_{x_0}^{x} (x-t)^n f^{(n+1)}(t)\,\mathrm{d}t = \frac{(x-x_0)^{n+1}}{(n+1)!} f^{(n+1)}(\xi),$$

za neki ξ koji leži između točaka x i x_0.

Polinom p_n nazivamo razvojem u Taylorov red funkcije f u točki x_0. Ako je funkcija f beskonačno derivabilna polinom p_n prelazi u red potencija s općim članom $(x - x_0)^n$, a zove se Taylorov red. Tako vrijede sljedeći razvoji u Taylorov red:

$$
\begin{aligned}
e^x &= 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \frac{x^{n+1}}{(n+1)!} e^{\xi}, \\
\cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots + (-1)^n \frac{x^{2n}}{(2n)!} + (-1)^{n+1} \frac{x^{2n+2}}{(2n+2)!} \cos \xi, \\
\sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots + (-1)^{n-1} \frac{x^{2n-1}}{(2n-1)!} + (-1)^n \frac{x^{2n+1}}{(2n+1)!} \sin \xi, \\
(1+x)^{\alpha} &= 1 + \binom{\alpha}{1} x + \binom{\alpha}{2} x^2 + \cdots + \binom{\alpha}{n} x^n + \binom{\alpha}{n+1} \frac{x^{n+1}}{(1+\xi)^{n+1-\alpha}},
\end{aligned}
$$

pri čemu točka ξ leži između 0 i x.

Ponekad je problem izračunati n-tu derivaciju promatrane funkcije, i u nekim se od takvih slučajevima možemo koristiti već poznatim razvojima u Taylorov red nekih funkcija. Npr. da bismo razvili u Taylorov red oko nule funkciju zadanu sa

$$f(x) = e^{-x^2},$$

iskoristimo već poznati razvoj eksponencijalne funkcije u Taylorov red oko nule stavljajući $-x^2$ umjesto x. Dobijemo

$$e^{-x^2} = 1 - x^2 + \frac{x^4}{2!} - \cdots + (-1)^n \frac{x^{2n}}{n!} + (-1)^{n+1} \frac{x^{2n+2}}{(n+1)!} e^{\xi},$$

pri čemu je $\xi \in \left[-x^2, 0\right]$

Da bismo dobili razvoj funkcije $arctg$ oko nule možemo u red

$$(1+x)^{\alpha} = 1 + \binom{\alpha}{1}x + \binom{\alpha}{2}x^2 + \cdots + \binom{\alpha}{n}x^n + \binom{\alpha}{n+1}\frac{x^{n+1}}{(1+\xi)^{n+1-\alpha}}$$

uvrstiti $\alpha = -1$ i $x = t^2$, nakon čega dobijemo

$$\frac{1}{1+t^2} = 1 - t^2 + t^4 - \cdots + (-1)^n t^{2n} + (-1)^{n+1}\frac{t^{2n+2}}{1+t^2}.$$

Integriranjem ove jednakosti po t na $[0, x]$ dobijemo

$$\text{arctg}\,(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + (-1)^n \frac{x^{2n+1}}{2n+1} + (-1)^{n+1}\int_0^x \frac{t^{2n+2}}{1+t^2}\mathrm{d}t,$$

dok primjena Teorema o srednjoj vrijednosti daje

$$\int_0^x \frac{t^{2n+2}}{1+t^2}\mathrm{d}t = \frac{x^{2n+3}}{2n+3}\cdot\frac{1}{1+\xi^2},$$

za ξ između 0 i x.

Kada funkcija koju promatramo nije derivabilna često nam umjesto derivacija te funkcije mogu poslužiti njene podijeljene razlike.

DEFINICIJA. Neka su $x_0, \ldots, x_n$, $n \in \mathbb{N}_0$, različiti realni brojevi i f realna funkcija definirana na nekom intervalu koji ih sadrži. **Podijeljenu razliku** n-tog reda funkcije f u točkama $x_0, \ldots, x_n$ definiramo rekurzivno s

$$f[x_i] = f(x_i), \quad i \in \{0, 1, \ldots, n\},$$

$$f[x_0, x_1, \ldots, x_k] = \frac{f[x_1, x_2, \ldots, x_k] - f[x_0, x_1, \ldots, x_{k-1}]}{x_k - x_0}, \quad i \in \{1, \ldots, n\}.$$

Ako je funkcija f k puta derivabilna na odgovarajućem intervalu, može se pokazati da vrijedi:

$$f[x_0, x_1, \ldots, x_k] = \frac{f^{(k)}(\xi)}{k!},$$

pri čemu točka ξ leži između minimuma i maksimuma skupa $\{x_0, x_1, \ldots, x_k\}$.

3 Pomoćni rezultati iz algebre

Uvedimo najprije osnovne oznake.
S $\mathbb{R}^n$, $n \in \mathbb{N}$, označit ćemo skup svih jednostupčanih realnih matrica (koje još nazivamo i realnim vektorima)

$$\mathbb{R}^n = \left\{ \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \mid (\forall i \in \{1, \ldots, n\})\ x_i \in \mathbb{R} \right\}.$$

Elemente skupa $\mathbb{R}^n$ označavat ćemo malim podebljanim latiničnim slovima $(\boldsymbol{a},\ \boldsymbol{b},\ \boldsymbol{x},\ \boldsymbol{y},\ \ldots)$
Posebno istićemo vektor komu su sve komponente jednake nuli: zovemo ga nul-vektor i označavamo s $\mathbf{0}$.

Realne vektore ćemo zbrajati i množiti skalarima (realnim brojevima) na standardan način: ako su $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ i $\alpha \in \mathbb{R}$, onda se definira

$$\boldsymbol{x} + \boldsymbol{y} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix},$$

$$\alpha \boldsymbol{x} = \alpha \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{bmatrix}.$$

Na isti način se definira i $\mathbb{C}^n$ kao skup kompleksnih vektora.
Uz ovako definirane operacije zbrajanja vektora i množenja vektora skalarom skupovi $\mathbb{R}^n$ i $\mathbb{C}^n$ imaju čitav niz lijepih svojstava.
Prisjetimo se da smo za realne brojeve uveli pojam *apsolutne vrijednosti*. Točnije, funkcija $|\cdot| : \mathbb{R} \to \mathbb{R}_0^+$ definirana s

$$|x| = \begin{cases} -x, & x < 0 \\ x, & x \geq 0 \end{cases}$$

naziva se apsolutna vrijednost i ima svojstva:

1. $|a| = 0$ ako i samo ako je $a = 0$,
2. $|ab| = |a|\,|b|\,,$
3. $|a + b| \leq |a| + |b|\,,$

pri čemu su a, b proizvoljni realni brojevi.

Slična svojstva ima i funkcija *norma* koja nenegativne realne brojeve pridružuje vektorima. Funkcija $\|\cdot\| : \mathbb{R}^n$ (ili $\mathbb{C}^n$) $\to \mathbb{R}$ biti će norma ako ima sljedeća svojstva :

1. $\|\boldsymbol{a}\| \geq 0$ za sve vektore $\boldsymbol{a}$,
2. $\|\boldsymbol{a}\| = 0$ ako i samo ako je $\boldsymbol{a} = \boldsymbol{0}$,
3. $\|\alpha \boldsymbol{a}\| = |\alpha| \, \|\boldsymbol{a}\|$ za sve vektore $\boldsymbol{a}$ i sve skalare α,
4. $\|\boldsymbol{a} + \boldsymbol{b}\| \leq \|\boldsymbol{a}\| + \|\boldsymbol{b}\|$.

U uporabi su najčešće tri standardne norme:

1. $\|\boldsymbol{a}\|_2 = \sqrt{|a_1|^2 + \cdots + |a_n|^2}$, 2-norma ili *euklidska norma*,
2. $\|\boldsymbol{a}\|_\infty = \max\{|a_1|, \ldots, |a_n|\}$, *norma beskonačno*,
3. $\|\boldsymbol{a}\|_1 = |a_1| + \cdots + |a_n|$, *norma jedan*,

dok je nešto općenitija tzv. p-norma ($1 \leq p \leq \infty$) definirana izrazom

$$\|\boldsymbol{a}\|_p = \sqrt[p]{|a_1|^p + \cdots + |a_n|^p}.$$

Norme postoje i za matrice. Npr. ako je dana matrica

$$A = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix},$$

onda je

$$\|A\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}| .$$

No norme matrica obično zadovoljavaju i jedno dodatno svojstvo koje se tiče umnoška matrica. Za ovakvu normu beskonačno to svojstvo glasi

$$\|AB\|_\infty \le \|A\|_\infty \|B\|_\infty ,$$

gdje su A i B proizvoljne matrice reda n.

4 Tipovi grešaka

Da bismo mogli procijeniti da li neki algoritam implementiran na računalu izračunava rješenje promatranog problema s dovoljnom točnošću najprije se moramo upoznati s vrstama grešaka koje se pri tom javljaju.

4.1 Greške zbog polaznih aproksimacija

Ovaj tip grešaka se često javlja kod rješavanja praktičnih problema. Ovako nastale greške se mogu podijeliti u tri klase: greške *modela*, greške *metode* i greške u *polaznim podacima*.

- **Greške modela**
  Ove greške nastaju zamjenom složenih sustava jednostavnijima koje se mogu opisati matematičkim modelima odnosno zapisima (primjer bi bio zanemarivanje utjecaja otpora zraka na gibanje u zemaljskim uvjetima). Često se već postojeći dobri modeli zamjenjuju jednostavnijima da bi se mogle primijeniti numeričke metode (npr. linearizirane nelinearnih sustava parcijalnih diferencijalnih jednadžbi). Greške modela se mogu javiti prilikom rješavanja problema koji su granični slučajevi (npr. aproksimiranje vrijednosti $\sin x$ s x i kada vrijednost x nije bliska nuli). Ove greške su neuklonjive, a na karisniku je ocijeniti da li primjena daje dovoljno dobre rezultate.

- **Greške metode**
  Te greške nastaju kada se beskonačni procesi zamjenjuju konačnima. Također nastaju kod računanja veličina koje su definirane pomoću limesa (derivacije, integrali, granične vrijednosti nizova,...). Velik broj numeričkih metoda za aproksimiranje funkcija i rješavanje jednadžbi je upravo ovog oblika. Greške koje nastaju zbog zamjene beskonačnog nečim konačnim obično dijelimo u dvije kategorije: greške *diskretizacije* i greške *odbacivanja*.

(i) **Greške diskretizacije**
Ove greške nastaju zamjenom continuuma konačnim diskretnim skupom točaka ili kada se beskonačno mala veličina zamijeni nekim konkretnim malim brojem. One nastaju i kada se derivacija zamijeni podijeljenom razlikom, integral kvadraturnom formulom ili diferencijalna jednadžba diferencijskom jednadžbom. Možda je najjednostavniji primjer aproksimacija funkcije definirane na intervalu $[a, b]$ funkcijom definiranom na diskretnom skupu $\{x_1, \ldots, x_n\} \subset [a, b]$.

(ii) **Greške odbacivanja**
Greške odbacivanja nastaju kada se beskonačni niz, red, umnožak, suma i sl. zamijene konačnima (tj. kada odbacimo ostatak).

- **Greške u polaznim podacima**
  One imaju izvor u mjerenjima fizičkih veličina, smještanju podataka u računalo i prethodnim računanjima. No greške mjerenja i smještanja je puno jednostavnije ocijeniti od grešaka koje nastaju usljed brojnih zaokruživanja tijekom računanja.

Grubo rečeno: diskretizacija je vezana za continuum $(\mathbb{R}, \mathbb{C})$, a odbacivanje za diskretnu (prebrojivu) beskonačnost $(\mathbb{N}, \mathbb{Z})$. Objekti koji nedostaju zbog tih zamjena tvore tip grešaka koji se zovu greške metode.

4.2 Greške zaokruživanja

Greške zaokruživanja nastaju zbog toga što računala koriste konačnu aritmetiku, točnije binarnu **aritmetiku s pomičnom točkom**, kod koje je unaprijed rezerviran određeni broj binarnih mjesta za eksponent i mantisu. Usljed toga se svaka računska operacija u kojoj sudjeluju dva broja izračunava s nekom malom greškom (koja može biti i nula). Tu grešku, ako nije jednaka nula, može se precizno ocijeniti, a nazivamo je **greškom zaokruživanja.** Očito, što je neki algoritam složeniji to ima više računskih operacija, a kod gotovo svake će se javiti greška zaokruživanja. Stoga se postavlja pitanje s kojom ćemo greškom dobiti traženo rješenje?

Ovim problemom se bavi **teorija grešaka zaokruživanja**, a osjetljivošću rješenja problema kojeg rješavamo na pomake u polaznim podacima bavi se **teorija perturbacije**. Njihovom usklađenom uporabom često je moguće procijeniti točnost promatranog algoritma, a ako točnost izračunatih podataka ne odstupa znatno od točnosti ulaznih, onda govorimo o **stabilnom algoritmu**.

Ne samo iracionalni, već i mnogi racionalni brojevi umjerene veličine nemaju točnu reprezentaciju u računalu. Ako je x neki realni broj, onda njegovu *računalnu reprezentaciju* označavamo s

$$fl\,(x)\,.$$

Proučavanje pokazuje da se kod svake računske operacije u računalu javlja greška. To se zapisuje u obliku

$$fl\,(x \circ y) = (x \circ y)\,(1+\varepsilon)\,,\ \ |\varepsilon| \leq u,\ \ \circ \in \{+,-,*,/\}\,,$$

pričemu je u tzv. **preciznost računanja** ili **strojni** u. Greška ovisi o operandima x, y i operaciji $\circ$, dok u ovisi o računalu (IEEE standardu). Općenito, ako računalo koristi p binarnih znamenaka u mantisi, onda vrijedi $u = 2^{-p+1}$ ili $u = 2^{-p}$ ovisno o načinu zaokruživanja u računalu.
Glavna zadaća osobe koja se bavi numeričkom matematikom jest određivanje što bolje aproksimacije rješenja u što kraćem vremenu.

4.3 Apsolutna i relativna greška

Neka je $\widehat{x}$ neka aproksimacija realnog broja x. Najkorisnije mjere za točnost broja $\widehat{x}$ kao aproksimacije broja x su:

- **apsolutna greška**

$$G_{\text{aps}}(x) = |x - \widehat{x}|$$

- **relativna greška**

$$G_{\text{rel}}(x) = \frac{|x - \widehat{x}|}{|x|}$$

koja nije definirana za $x = 0$.

Ako je x poznat ili mu se zna red veličine, onda je apsolutna greška dobra mjera udaljenosti aproksimacije od točne vrijednosti. No u praksi x često varira od vrlo velikih do vrlo malih vrijednosti, pa je primjerenija mjera relativna greška. Ona ima dodatno lijepo svojstvo da je naovisna o skaliranju,

$$\frac{|x - \widehat{x}|}{|x|} = \frac{|\alpha x - \alpha\widehat{x}|}{|\alpha x|}, \quad \alpha \in \mathbb{R}.$$

Relativna greška povezana je s brojem točnih **značajnih znamenaka** neke aproksimacije. Značajne znamenke su prva netrivijalna znamenka i one koje slijede iza nje u zapisu. Npr. u broju 6.9990 imamo pet značajnih znamenaka, a u broju 0.0832 samo tri. Što znači broj točnih značajnih znamenaka vidjet ćemo kroz primjer:

$$\begin{aligned} x &= 1.00000,\ \widehat{x} = 1.00499, \quad G_{\mathsf{rel}}(x) = 4.99 \cdot 10^{-3}, \\ x &= 9.00000,\ \widehat{x} = 8.99899, \quad G_{\mathsf{rel}}(x) = 1.12 \cdot 10^{-4}. \end{aligned}$$

Evo jedne moguće definicije.
$\widehat{x}$ kao aproksimacija od x ima p točnih značajnih znamenaka ako se $\widehat{x}$ i x zaokružuju na isti broj od p značajnih znamenaka. Zaokružiti broj na p značajnih znamenaka znači zamijeniti ga s najbližim brojem koji ima p značajnih znamenaka. No prema ovoj definiciji brojevi $x = 0.9949$ i $\widehat{x} = 0.9951$ se ne slažu u dvije značajne znamenke, a slažu se u jednoj i u tri. Prema tome, definicija nije dobra.

Evo druge definicije.
$\widehat{x}$ kao aproksimacija od x ima p točnih značajnih znamenaka ako je $|x - \widehat{x}|$ manje od jedne polovine jedinice u p-toj značajnoj znamenci od x. Ova definicija implicira da se brojevi $x = 0.123$ i $\widehat{x} = 0.127$ slažu u dvije značajne znamenke, iako će mnogi misliti da se slažu u tri.

Kada se radi o vektorima greške se definiraju kao

$$G_{\text{aps}}(\boldsymbol{x}) = \|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|$$

i

$$G_{\text{rel}}(\boldsymbol{x}) = \frac{\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|},$$

a relacija

$$\frac{\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} \leq \frac{1}{2}10^{-p}$$

implicira da komponente x_i za koje vrijedi $|x_i| \approx \|\boldsymbol{x}\|$ imaju približno p točnih značajnih znamenaka.
Ako želimo sve komponente vektora staviti u prvi plan onda koristimo relativne greške po komponentama, a veličinu

$$\max_i \frac{|x_i - \widehat{x_i}|}{|x_i|}$$

nazivamo **maksimalna relativna greška po komponentama**.

Treba razlikovati pojam *preciznosti* od pojma *točnosti*.

- **Točnost** se odnosi na apsolutnu i relativnu grešku kojom se aproksimira tražena veličina.
- **Preciznost** je točnost kojom se izvršavaju osnovne računske operacije, a u aritmetici pomične točke mjerimo je pomoću u. Određena je brojem bitova u reprezentaciji mantise, pa se ista riječ koristi i za taj broj bitova.

Ipak, važno je znati da preciznost *ne limitira* točnost. Naime, uvijek se (uz povečanje potrošnje računalnog vremena) uz neku danu preciznost može simulirati i veća preciznost računanja.