Predavanje 1

Uvod v numerične metode

Prvi sklop izročkov

Fakulteta za računalništvo in informatiko
Univerza v Ljubljani

5. oktober 2020

Viri in obveznosti

Viri:

- Bojan Orel, Osnove numerične matematike, Založba FE in FRI.
- Bor Plestenjak: Razširjen uvod v numerične metode, DMFA založništvo.

Obveznosti:

- Predavanja: 3 ure na teden
- Vaje: laboratorijske vaje vsak teden po dve uri.
- 3 domače naloge
- Pisni izpit iz teorije

Vsebina predmeta

1. Računanje in vloga napak pri numerični matematiki.
2. Sistemi linearnih enačb
   - Gausova eliminacija in LU razcep
   - Delno/kompletno pivotiranje - cena in občutljivost
   - Iterativne metode - Jacobijeva in Seidlova iteracija
3. Reševanje nelinearnih enačb
   - Bisekcija
   - Sekantna metoda in regula falsi
   - Tangentna metoda
   - Metoda fiksne točke
4. Aproksimacija in interpolacija
   - Lagrangeov in Newtonov interpolacijski polinom
   - Aproksimacija po metodi najmanjših kvadratov
5. Numerično odvajanje in integriranje
   - Trapezna metoda
   - Simpsonova metoda
   - Rombergova metoda
6. Diferencialne enačbe
   - Eulerjeva metoda
   - Runge-Kutta metode

Numerično in simbolno računanje

Numerično računanje:

- Takoj v formulo vstavljamo števila
- Pridemo do numeričnega rezultata - numerične rešitve

Simbolno računanje:

- simboli predstavljajo števila
- izraz preoblikujemo s simbolnim računanjem do novega simbolnega izraza - analitična rešitev

Primer

- *Numerično:*

$$\frac{(17.36)^2 - 1}{17.36 + 1} = 16.36; \qquad 0.25,\ 0.33333\ldots(?),\ 3.14159\ldots(?)$$

- *Simbolno:*

$$\frac{x^2 - 1}{x + 1} = x - 1; \qquad \frac{1}{4},\ \frac{1}{3},\ \pi,\ \tan 83$$

Metoda, algoritem in implementacija

Metoda. . . matematična konstrukcija, s katero rešujemo problem

Algoritem. . . koraki metode

Implementacija. . . zapis algoritma v izbranem jeziku

'Dober' v NM. . . majhna sprememba $\Rightarrow$ majhna napaka

Vprašanja:

- Ali je problem občutljiv?
- Ali je metoda 'dobra'?
- Ali je algoritem robusten - deluje na širokem spektru problemov?
- Ali je implementacija hitra - časovna in prostorska zahtevnost?

Občutljivost problema

Tipi problemov (glede na spremembo rezultata ob majhni spremembi začetnih podatkov):

- **občutljivi** oz. **slabo pogojeni** problemi.
- **neobčutljivi** oz. **dobro pogojeni** problemi.

Občutljivost je odvisna le od narave problema in ne od izbrane numerične metode.

Primer (Presečišča premic)

1. *Sistem in njegova perturbacija*

$$\begin{aligned} x + y &= 2 \quad \rightarrow \quad x + y = 1.9999 \\ x - y &= 0 \quad \rightarrow \quad x - y = 0.0002 \end{aligned}$$

ima rešitvi $x = y = 1$ *oz.* $x = 1.00005$ *in* $y = 0.99985$. *Začetni sistem je neobčutljiv, saj je šlo za spremembo za isti velikostni razred.*

2. *Sistem in njegova perturbacija*

$$\begin{aligned} x + 0.99y &= 1.99 \quad \rightarrow \quad x + 0.99y = 1.9899 \\ 0.99x + 0.98y &= 1.97 \quad \rightarrow \quad 0.99x + 0.98y = 1.9701 \end{aligned}$$

ima rešitvi $x = y = 1$ *oz.* $x = 2.97$ *in* $y = -0.99$. *Začetni sistem je občutljiv, saj je majhna sprememba začetnih podatkov povzročila veliko spremembo rezultata.*

Pristopi numerične matematike

- Neskončne procese nadomestimo s končnimi (Taylorjeva vrsta).
- Neskončno razsežne prostore nadomestimo s končno razsežnimi (funkcijske enačbe rešujemo v prostoru polinomov).
- Diferencialne enačbe nadomestimo z algebraičnimi.
- Nelinearne probleme nadomestimo z linearnimi.
- Matrike nadomestimo z enostavnejšimi.

Numerična matematika - zakaj?

- Znanost, ki temelji na matematičnih izračunih, je neposredno odvisna od NM.
- Nekatere katastrofe so se zgodile zaradi slabega numeričnega računanja: (`http://www-users.math.umn.edu/~arnold//disasters/`)
  - **Nesreča Misije Patriot, Dharan, Savdska Arabija, 25.2.1991**, 28 žrtev: slaba analiza zaokrožitvenih napak.
  - **Eksplozija rakete Ariana 5, Francoska Gvajana, 4.6.1996**: posledica prekoračitve obsega števil. `https://www.youtube.com/watch?v=PK_yguLapgA` `https://www.youtube.com/watch?v=W3YJeoYgozw` `https://www.arianespace.com/vehicle/ariane-5/`
  - **Potop ploščadi Sleipner A, Stavanger, Norveška, 12.8.1991**, miljarda dolarjev škode: nenatančna obdelava obremenitev pri reševanju PDE-jev. `https://www.youtube.com/watch?v=eGdiPs4THW8`

Predstavljiva števila

Števila shranjujemo v obliki

$$\boxed{x = \pm(0.d_1 d_2 d_3 \ldots d_m)_\beta \times \beta^e,}$$

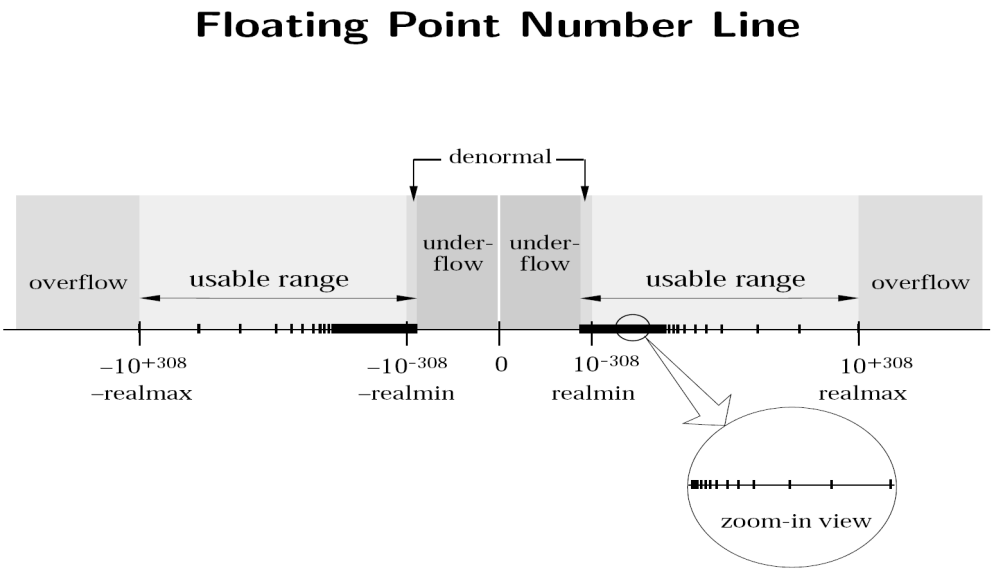
kjer je

- β naravno število,
- $d_1 d_2 d_3 \ldots d_m$ mantisa, e eksponent,
- e celo število,

Primer (baza 10)

- *1000.12345 zapišemo kot* $+(0.100012345)_{10} \times 10^4$.
- *0.000812345 zapišemo kot* $+(0.812345)_{10} \times 10^{-3}$.

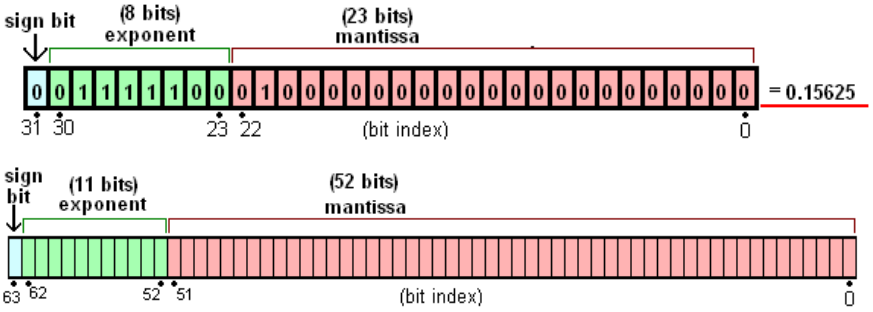
Prekoračitev in podkoračitev



- izračuni preblizu 0 lahko povzročijo **podkoračitev**
- preveliki izračuni lahko povzročijo **prekoračitev**
- prekoračitev je v splošnem hujši problem

IEEE standard

- IEEE Enojna natančnost: števila so predstavljena z 32 biti.
- IEEE Dvojna natančnost: števila so predstavljena z 64 biti.



Zaokrožitvene napake

- Vseh realnih števil ne moremo predstaviti v strojni aritmetiki $\Rightarrow$ zaokrožujemo in delamo zaokrožitvene napake.
- IEEE standard... zaokroži do najbližjega predstavljivega števila
- Naj bosta $x_- \leqslant x \leqslant x_+$ najbližji predstavljivi števili števila x:

$$\boxed{\mathrm{fl}(x) = \begin{cases} x_-, & \text{če je } x - x_- < x_+ - x, \\ x_+, & \text{če je } x - x_- \geqslant x_+ - x. \end{cases}}$$

- Kako velika je napaka? Recimo, da je x bližje x_-:

$$x = (0.1 b_2 b_3 \ldots b_{24} b_{25} b_{26})_2 \times 2^m$$

$$x_- = (0.1 b_2 b_3 \ldots b_{24})_2 \times 2^m, \qquad x_+ = \left((0.1 b_2 b_3 \ldots b_{24})_2 + 2^{-24}\right) \times 2^m$$

$$|x - x_-| \leqslant \frac{|x_+ - x_-|}{2} = 2^{m-25}, \qquad \left|\frac{x - x_-}{x}\right| \leqslant \frac{2^{m-25}}{1/2 \times 2^m} \leqslant 2^{-24} =: \frac{1}{2}\epsilon.$$

- $\boxed{\epsilon := 2^{1-24}}$ ... **strojni epsilon**
- $\boxed{u := \frac{1}{2}\epsilon}$ ... **osnovna zaokrožitvena napaka**

Strojni ϵ

Vzemimo $x = 1.0$ in dodajajmo $\frac{1}{2}, \frac{1}{4}, \ldots, 2^{-i}$:

Skrit bit ↓ ← 23 bitov →

1	1	0	0	0	0	0	0	0	0	0	0	e	e
1	0	1	0	0	0	0	0	0	0	0	0	e	e
1	0	0	1	0	0	0	0	0	0	0	0	e	e
						.....							
1	0	0	0	0	0	0	0	0	0	0	1	e	e
1	0	0	0	0	0	0	0	0	0	0	0	e	e

- $\mathrm{fl}(1 + 2^{-23}) \neq 1$, toda $\mathrm{fl}(1 + 2^{-24}) = 1$
- Enojna natančnost: 2^{-23}; Dvojna natančnost: 2^{-52}.

```
>> eps
ans = 2.2204e-16

>> eps('single')
ans = 1.1921e-07
```


Aritmetika v predstavljivih številih

- $\boxed{\mathrm{fl}(x) = x(1+\delta)}$, kjer je $|\delta| \leqslant u := \frac{1}{2}\beta^{1-m}$.
- Za predstavljivi števili x, y in $\odot \in \{+, -, *, /\}$ velja

$$\boxed{\mathrm{fl}(x \odot y) = (x \odot y)(1+\delta)}.$$

- $\boxed{(a+b)+c \neq a+(b+c)}$:

$$\begin{aligned}
(a+b)+c &= \mathrm{fl}(\mathrm{fl}(a+b)+c) = \mathrm{fl}((a+b)(1+\delta_1)+c) \\
&= [(a+b)(1+\delta_1)+c](1+\delta_2) \\
&= [(a+b+c)+(a+b)\delta_1)](1+\delta_2) \\
&= (a+b+c)\left[1+\frac{a+b}{a+b+c}\delta_1(1+\delta_2)+\delta_2\right]
\end{aligned}$$

Podobno

$$a+(b+c) = (a+b+c)\left[1+\frac{b+c}{a+b+c}\delta_3(1+\delta_4)+\delta_4\right].$$

Če pozabimo na člena $\delta_1\delta_2$ in $\delta_3\delta_4$, dobimo

$$(a+b)+c = (a+b+c)(1+\epsilon_3) \quad \text{kjer je} \quad \epsilon_3 \approx \frac{a+b}{a+b+c}\delta_1 + \delta_2,$$

$$a+(b+c) = (a+b+c)(1+\epsilon_4) \quad \text{kjer je} \quad \epsilon_4 \approx \frac{b+c}{a+b+c}\delta_3 + \delta_4.$$

Sklep: Ko seštevamo n števil, je najbolje začeti z najmanjšim in prištevati večje.

Napake pri numeričnem računanju

- **Neodstranljiva napaka** D_n . . . nenatančni začetni podatki.
- **Napaka metode** D_m . . . npr. neskončni proces aproksimiramo s končnim.
- **Zaokrožitvena napaka** D_z . . . računanje s približki in zaokroževanje.

Celotna napaka je

$$\boxed{D = D_n + D_m + D_z}.$$

Primer ($\sin \frac{\pi}{10}$ računamo v $\beta = 10$ in $m = 4$)

- D_n: $fl(\frac{\pi}{10}) = 0.3142 \cdot 10^0$.

$$|D_n| \approx \sin'(\frac{\pi}{10})|x - fl(x)| \leqslant \frac{1}{2} \cdot 10^{-4}.$$

- D_m : $\sin x \approx x - x^3/6$.

$$|D_m|| \leqslant x^5/120 = 2.6 \cdot 10^{-5}.$$

- D_z: $fl(x - fl(fl(fl(x \cdot x) \cdot x)/6))$.

$$|D_z| \leqslant 3.0 \cdot 10^{-5}.$$

Stabilnost metode

Stabilnost metode preverimo z **analizo zaokrožitvenih napak**.

Vrste napak (x naj bo točna vrednost, $\bar{x}$ pa približek zanjo):

- Prva delitev:
  - **Absolutna napaka** : $\boxed{\bar{x} - x}$.
  - **Relativna napaka** : $\boxed{\frac{\bar{x} - x}{x}}$.
- Druga delitev:
  - **Direktna napaka**: Absolutna in relativna napaka.
  - **Obratna napaka**: Koliko je potrebno spremeniti začetne podatke, da dobimo izračunan rezultat.

Glede na napake, nas lahko zanima:

- **direktna stabilnost** metode.
- **obratna stabilnost** metode.

Velja

$$\boxed{|\text{direktna napaka}| \approx \text{občutljivost} \times |\text{obratna napaka}|}.$$

Izračunana vrednost je blizu pravi, če rešujemo neobčutljiv problem z obratno stabilno metode.

Katastrofalno odštevanje

- Glavni problem zaokrožitvene aritmetike. . . **odštevanje/seštevanje dveh približno enakih/nasprotnih števil**

$$a = x.xxxx\ xxxx\ xxx1\ \overbrace{ssss\ldots}^{\text{izguba}}$$

$$b = x.xxxx\ xxxx\ xxx0\ \overbrace{tttt\ldots}^{\text{izguba}}$$

Potem

$$\begin{array}{cll} & \overbrace{x.xxx\ xxxx\ xxx1}^{\text{končna natančnost}} & \\ - & x.xxx\ xxxx\ xxx0 & \\ \hline = & 0.000\,0000\,0001 & \underbrace{????\,????}_{\text{izguba natančnosti}} \end{array}$$

- S ponavljanjem se napake seštevajo.

Katastrofalno odštevanje

Iščemo rešitve kvadratne enačbe

$$x^2 + 2a^2x - q = 0$$

Rešitev z manjšo absolutno vrednostjo je

$$x_2 = \frac{-2a^2 + \sqrt{4a^4 + 4q}}{2} = -a^2 + \sqrt{a^4 + q}.$$

1 $s := (a^2)^2$
2 $t := s + q$
3 $u := \sqrt{t}$
4 $x_2 := -a^2 + u$

Če je a^2 veliko večji od q, potem ima lahko korak 4 zgoraj veliko napako. Bolje:

$$x_2 = \left(-a^2 + \sqrt{a^4 + q}\right) \cdot \frac{a^2 + \sqrt{a^4 + q}}{a^2 + \sqrt{a^4 + q}} = \frac{q}{a^2 + \sqrt{a^4 + q}}.$$

1 $s := (a^2)^2$
2 $t := s + q$
3 $u := \sqrt{t}$
4 $v := a^2 + u$
5 $y := q/v$

Rešitve $x^2 + 2Ax - q = 0$

```
>> A = 10000;
>> q = 1;
>> y = -A+sqrt(A^2 + q)
y = 5.000000055588316e-05

>> y2 = q/(A+sqrt(A^2+q))
y2 = 4.999999987500000e-05

>> x = y;
>> x^2 + 2 * A * x -q
ans = 1.361766321927860e-08

>> x = y2;
>> x^2 + 2 * A * x -q
ans = -1.110223024625157e-16
```


Seštevanje in odštevanje nista direktno stabilni operaciji

$x, y \in \mathbb{R}$. Računamo približek $\overline{p}$ za $p = x \pm y$.

$$\begin{aligned} \overline{p} &= \mathrm{fl}(\mathrm{fl}(x) \pm \mathrm{fl}(y)) = \mathrm{fl}(x(1+\delta_1) \pm y(1+\delta_2)) \\ &= (x(1+\delta_1) \pm y(1+\delta_2))(1+\delta_3) \\ &= (x(1+\delta_1)(1+\delta_3) \pm y(1+\delta_2)(1+\delta_3)) \\ &= (x \pm y + x(\delta_1+\delta_3+\delta_1\delta_3) \pm y(\delta_2+\delta_3+\delta_2\delta_3)) \end{aligned}$$

kjer je $|\delta_i| \leqslant u$. Relativna napaka je

$$\boxed{\frac{|\overline{p}-p|}{|p|} \leqslant \frac{|x(\delta_1+\delta_3+\delta_1\delta_3) \pm y(\delta_2+\delta_3+\delta_2\delta_3)|}{|x+y|}}.$$

Torej:

$$x + y \approx 0 \quad \Rightarrow \quad \frac{|\overline{p}-p|}{|p|} \text{ veliko.}$$

Množenje je direktno stabilna operacija

$x, y \in \mathbb{R}$. Računamo približek $\overline{p}$ za $p = x \cdot y$.

$$\begin{aligned} \overline{p} &= \mathrm{fl}(\mathrm{fl}(x) \cdot \mathrm{fl}(y)) = \mathrm{fl}(x(1+\delta_1) \cdot y(1+\delta_2)) \\ &= x(1+\delta_1) \cdot y(1+\delta_2)(1+\delta_3) \\ &= xy(1+\delta_1+\delta_2+\delta_3 + \text{produkti več } \delta), \end{aligned}$$

kjer je $|\delta_i| \leqslant u$. Relativna napaka je

$$\boxed{\frac{|\overline{p}-p|}{|p|} \leqslant \frac{|xy||\delta_1+\delta_2+\delta_3+\mathcal{O}(u^2)|}{|xy|} = |\delta_1+\delta_2+\delta_3+\mathcal{O}(u^2)|}.$$

Torej:

Relativna napaka $\dfrac{|\overline{p}-p|}{|p|}$ ni odvisna od velikosti produkta xy.

Deljenje je direktno stabilna operacija

$x, y \in \mathbb{R}$. Računamo približek $\overline{p}$ za $p = x/y$.

$$\begin{aligned} \overline{p} &= \mathrm{fl}(\mathrm{fl}(x)/\mathrm{fl}(y)) = \mathrm{fl}(x(1+\delta_1)/y(1+\delta_2)) \\ &= x(1+\delta_1)/(y(1+\delta_2))(1+\delta_3) \\ &= x/y(1+\delta_1)(1-\delta_2+\delta_2^2-\cdots)(1+\delta_3), \\ &= x/y(1+\delta_1-\delta_2+\delta_3+\text{produkti več } \delta), \end{aligned}$$

kjer je $|\delta_i| \leqslant u$. Relativna napaka je

$$\boxed{\frac{|\overline{p}-p|}{|p|} \leqslant \frac{|x/y||\delta_1-\delta_2+\delta_3+\text{produkti več } \delta|}{|x/y|} = |\delta_1-\delta_2+\delta_3+\mathcal{O}(u^2)|}$$

Torej:

Relativna napaka $\dfrac{|\overline{p}-p|}{|p|}$ ni odvisna od velikosti kvocienta x/y.

Računanje skalarnega produkta - obratno, a ne direktno stabilno

Računamo skalarni produkt s vektorjev $x = (x_1, \ldots, x_n)$ in $y = (y_1, \ldots, y_n)$.

```
1 s̄ = x_1 · y_1
2 for i = 2 to n do
3   s̄ = s̄ + x_i · y_i
4 end
5 return s̄
```

Velja:

$$\bar{s} = \sum_{i=1}^{n} x_i \cdot y_i \cdot \underbrace{(1+3\delta_i) \cdot \prod_{j=i}^{n}(1+\epsilon_j)}_{1+\gamma_i}.$$

Ocenimo:

$$(1-3u)^{n-i+2} \leqslant 1+\gamma_i \leqslant (1+3u)^{n-i+2}.$$

Z indukcijo:

$$1-(n-i+2)3u \leqslant (1-3u)^{n-i+2}.$$

Z razvojem po binomski formuli:

$$(1+3u)^{n-i+2} \leqslant 1+(n-i+2)3u+\mathcal{O}(u^2).$$

Sledi:

$$|\gamma_i| \leqslant (n-i+2)3u \leqslant 3nu.$$

Računanje skalarnega produkta - obratno, a ne direktno stabilno

Relativna napaka je

$$\boxed{\frac{|s-\bar{s}|}{|s|} \leqslant \frac{3nu\sum_{i=1}^{n}|x_i|y_i|}{|x^T y|}}.$$

Če sta vektorja blizu pravokotnosti, je relativna direktna napaka lahko zelo velika. Je pa obratno stabilno, saj lahko vzamemo

$$\overline{x}_i = x_i, \quad \overline{y}_i = y_i(1+\gamma_i) \quad \text{za} \quad i = 1, \ldots, n.$$

Računanje vrednosti polinoma - obratno, a ne direktno stabilno

Računamo vrednost polinoma

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0$$

v neki točki $x = x_0$.

```
1 p̄ = a_n
2 for i = 1 : n
3   p̄ = p̄ · x + a_{n−i}
4 end
5 return p̄
```

Velja

$$\bar{p} = \sum_{i=0}^{n} a_{n-i} x^{n-i} \underbrace{\prod_{j=i+1}^{n} (1 + 3\delta_i) \prod_{k=i}^{n} (1 + \epsilon_i)}_{1+\gamma_i} .$$

Torej je

$$|\gamma_i| \leqslant 2(n - i + 1)3u \leqslant 6nu$$

in računanje vrednosti je obratno, ne pa direktno stabilno.