Numerične metode - NTF

2013/2014

- Gašper Jaklič
- pisarna: FMF, Jadranska 21, soba 5.15
- govorilna ura: po dogovoru
- e-pošta: gasper.jaklic@fmf.uni-lj.si
- domača stran predmeta: spletna učilnica
- asistentka: Selena Praprotnik
- vaje ta teden odpadejo

Režim

- 2 kolokvija
- 3 pisni izpiti
- domače naloge
- ustni izpit

Pisni izpit lahko opravite s kolokviji. Domače naloge prinesejo 20% pisne ocene, pisni izpit ali kolokviji pa 80% ocene.

Pogosta vprašanja

- na kolokviju ali izpitu imate lahko priročnik, kalkulator in 2 A4 lista
- pisno oceno lahko enkrat popravljate (velja boljša ocena), naslednjič velja zadnja ocena
- pisna ocena velja do konca študijskega leta

Numerična matematika

Numerična matematika se ukvarja z razvojem in analizo algoritmov za numerično reševanje matematičnih problemov.

Ukvarjali se bomo z naslednjimi problemi:

- linearni sistemi: Reši:

$$\begin{aligned} 3x_1 + 2x_2 &= 7, \\ -4x_1 + 5x_2 &= 3. \end{aligned}$$

- nelinearne enačbe in sistemi: Reši:

$$\log x + x = 0.$$

- linearni problem najmanjših kvadratov: Reši:

$$\begin{aligned} 3x_1 + 2x_2 &= 7, \\ -4x_1 + 5x_2 &= 3, \\ 3x_1 + 2x_2 &= -2. \end{aligned}$$

  - Poišči premico, ki se najbolje prilega danim točkam.

- interpolacija: poišči polinom, ki gre skozi točke $(1,3),(2,5),(4,-1),(5,2)$.
- Bézierove krivulje: zapis polinomske krivulje, posebej primeren za grafično oblikovanje
- integriranje: izračunaj integral

$$\int_1^2 \frac{e^t}{t}\, dt.$$

Numerično reševanje

Numerično reševanje pomeni, da rešitev problema iščemo v numerični obliki, in ne eksaktno. Naprimer namesto $\sqrt{3}$ računamo $1.7320\ldots$.

Numerična metoda je postopek, s katerim iz vhodnih numeričnih podatkov s končnim zaporedjem elementarnih operacij izračunamo numerični približek za rezultat določenega problema.

Elementarne operacije so odvisne od okolja, mi bomo pod to šteli +, -, *, / in $\sqrt{}$.

Uporabljali bomo program Octave.

Numerično računanje

Numerično računanje se razlikuje od eksaktnega, ker običajno na računalniku računamo s števili, predstavljenimi s premično piko (plavajočo vejico).

Primer: $100 * (100/3 - 33) - 100/3$
Nepričakovano dobimo (Octave) $2.34479102800833e - 13$.

Za osnovne računske operacije naj bi veljala asociativnost, toda npr. za
$x = 0.1234567890$
$y = 0.0987654321$
$z = 0.9911991199$
dobimo

$$x * y * z - z * y * x = 1.73472347597681e - 18$$

Uporaba numeričnih metod

Ko drugih možnosti ne poznamo, npr.:

- iskanje ničel polinoma pete stopnje $x^5 + 3x - 1 = 0$,
- reševanje enačbe: $x + \ln x = 0$,
- računanje določenega integrala: $\int_0^1 e^{x^2}\, dx$,
- pri večini nelinearnih enačb, diferencialnih enačb,...

Kadar so bolj preproste od analitičnih rešitev, npr.:

- Cardanove formule za ničle kubičnega polinoma.

Prevedba na lažji problem

Glavni princip numeričnega reševanja problema je, da namesto podanega težkega problema rešujemo lažji problem, ki ima ali enako rešitev ali pa se rešitvi ne razlikujeta dosti. Npr.:

- neskončne procese nadomestimo s končnimi,
- nelinearni problem nadomestimo z linearnim,
- zapletene funkcije aproksimiramo z enostavnejšimi, npr. s polinomi.

Dobra numerična metoda

Glavne zahteve za dobro numerično metodo:

- zanesljivost: na enostavnih problemih vedno deluje pravilno.
- robustnost: običajno deluje na težjih problemih, kadar pa ne deluje, vrne informacijo o tem.
- natančnost: izračuna rešitev tako natančno, kot je to možno glede na natančnost podanih začetnih podatkov.
- ekonomičnost: časovna (število operacij) in prostorska (poraba spomina).
- uporabnost: lahko jo uporabimo na širokem spektru problemov.
- prijaznost do uporabnika: je dobro dokumentirana in ima enostaven uporabniški vmesnik.

Razvoj

Numerične metode se stalno razvijajo. Razlog:

- novi problemi,
- novi pristopi in novi algoritmi,
- razvoj računalnikov,
- razvoj paralelnih računalnikov.

Absolutna in relativna napaka

Pri numeričnem računanju izračunamo numerični približek za točno rešitev problema. Razlika med približkom in točno vrednostjo je napaka približka. Ločimo absolutno in relativno napako:

$$\text{absolutna napaka} = \text{približek - točna vrednost},$$

$$\text{relativna napaka} = \frac{\text{absolutna napaka}}{\text{točna vrednost}}.$$

Naj bo x točna vrednost, $\widehat{x}$ pa približek za x.

- Če je $\widehat{x} = x + d_a$, potem je $d_a = \widehat{x} - x$ absolutna napaka.
- Če je $\widehat{x} = x(1 + d_r)$ potem je $d_r = \frac{\widehat{x}-x}{x}$ relativna napaka.

Občutljivost problema

Če se rezultat pri majhni spremembi podatkov (motnji oz. perturbaciji) ne spremeni veliko, je problem neobčutljiv, sicer pa je občutljiv. Npr.

$$\begin{aligned} x + y &= 2 \\ x - y &= 0 \end{aligned} \Longrightarrow x = y = 1.$$

Zmotimo desno stran:

$$\begin{aligned} x + y &= 1.9999 \\ x - y &= 0.0002 \end{aligned} \Longrightarrow x = 1.00005, y = 0.99985.$$

Ta sistem je neobčutljiv.

$$\begin{aligned} x + 0.99y &= 1.99 \\ 0.99x + 0.98y &= 1.97 \end{aligned} \Longrightarrow x = y = 1.$$

Zmotimo desno stran:

$$\begin{aligned} x + 0.99y &= 1.9899 \\ 0.99x + 0.98y &= 1.9701 \end{aligned} \Longrightarrow x = 2.97, y = -0.99.$$

Ta sistem je zelo občutljiv.

Wilkinsonov zgled

Polinom

$$p(x) = (x-1)(x-2)\cdots(x-20) = x^{20} - 210x^{19} + \cdots + 20!$$

ima ničle $1, 2, \ldots, 20$, polinom

$$g(x) = p(x) - 2^{-23}x^{19}$$

pa ima ničle

$$\begin{aligned} x_9 &= 8.91752 \\ x_{10,11} &= 10.0953 \pm 0.64310i \\ x_{16,17} &= 16.7307 \pm 2.81263i \\ x_{18,19} &= 19.5024 \pm 1.94033i \\ x_{20} &= 20.8469 \end{aligned}$$

Majhna motnja povzroči velike spremembe.

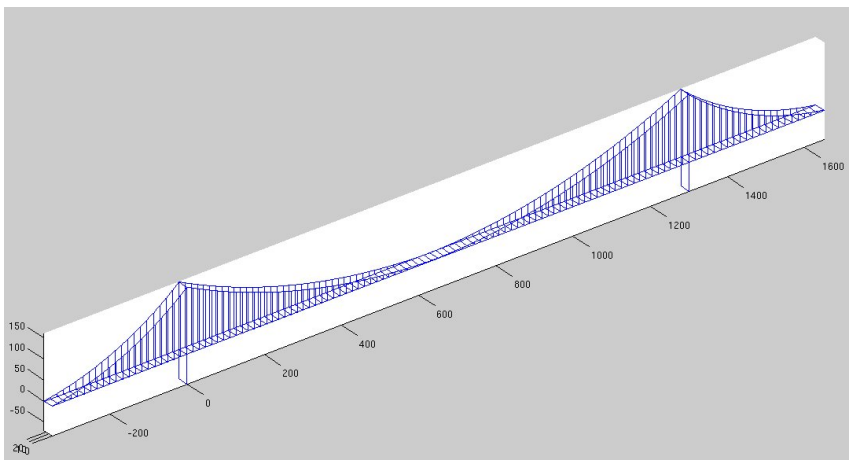
Računalniške simulacije

Številne fizikalne, tehnološke, in druge procese lahko simuliramo na računalniku. Postopek je sestavljen iz naslednjih korakov:

1. razvoj matematičnega modela,
2. razvoj numeričnih metod za numerično reševanje modela,
3. implementacija numeričnih metod,
4. simuliranje procesov z računalnikom,
5. predstavitev dobljenih numeričnih rezultatov v pregledni obliki,
6. analiza dobljenih rezultatov, po potrebi se vrnemo na enega izmed prejšnjih korakov.

Zgled - model mosta Golden Gate

Modelirati je potrebno most Golden Gate v San Franciscu za realne podatke. Most (200000*ton*) nosijo jeklene vrvi premera 92*cm*, napete na 230*m* stebre, nanje so v enakomernih razmikih pritrjene vertikalne vrvi, ki nosijo most.



Reševati je potrebno sisteme nelinearnih enačb.

Zgled - tračnica

Imamo $150m$ dolgo tračnico, ki je na obeh koncih trdno vpeta. Zaradi velike vročine se tračnica raztegne za $1cm$ in se dvigne v obliki krožnega loka. Na 8 decimalk točno izračunaj, kolikšna je maksimalna oddaljenost od tal.

Odgovor: $75.00074999cm$.

Rešiti je potrebno nelinearno enačbo.

Dinamični zgled

- modeliranje vrste na banki
- modeliranje prometa skozi križišče
- avtomatski pilot v letalu
- vodenje robota
- ...

Zapis števil v premični piki

V računalniku so števila zapisana v premični piki kot

$$x = \pm m \cdot b^e,$$

kjer je $m = 0.c_1 c_2 \ldots c_t$ mantisa in

- b: baza (2, lahko tudi 10 ali 16),
- t: dolžina mantise,
- e: eksponent v mejah $L \leq e \leq U$,
- c_i: števke v mejah od 0 do $b - 1$.

Če je $c_1 \neq 0$, potem je število normalizirano, sicer pa subnormalizirano.

Zapis označimo s $P(b, t, L, U)$.

Npr. $0.1101 \cdot 2^2 = 3.25$.

Zgled

Vsa normalizirana pozitivna predstavljiva števila iz množice $P(2, 3, -1, 1)$ so:

$$0.100_2 \cdot 2^{-1} = 0.2500 \quad 0.100_2 \cdot 2^{0} = 0.500 \quad 0.100_2 \cdot 2^{1} = 1.00$$
$$0.101_2 \cdot 2^{-1} = 0.3125 \quad 0.101_2 \cdot 2^{0} = 0.625 \quad 0.101_2 \cdot 2^{1} = 1.25$$
$$0.110_2 \cdot 2^{-1} = 0.3750 \quad 0.110_2 \cdot 2^{0} = 0.750 \quad 0.110_2 \cdot 2^{1} = 1.50$$
$$0.111_2 \cdot 2^{-1} = 0.4375 \quad 0.111_2 \cdot 2^{0} = 0.875 \quad 0.111_2 \cdot 2^{1} = 1.75$$

Subnormalizirana števila (možna le pri najmanjšem eksponentu) so:

$$0.011_2 \cdot 2^{-1} = 0.01875$$
$$0.010_2 \cdot 2^{-1} = 0.01250$$
$$0.001_2 \cdot 2^{-1} = 0.00625$$

Standard IEEE

- single: $P(2, 24, -125, 128)$, število je shranjeno v 32 bitih, predznak:1, eksponent: 8, mantisa: 23 $(c_2 c_3 \dots c_{24}\,(c_1 = 1))$
- double: $P(2, 53, -1021, 1024)$, število je shranjeno v 64 bitih, predznak: 1, eksponent: 11, mantisa: 52 $(c_2 c_3 \dots c_{53}\,(c_1 = 1))$
- obstajajo posebna števila $0, \infty, -\infty$ in NaN.

Zapis predstavljivih števil (enojna dolžina)

Naj bo s predznak, $0 \leq e \leq 255$ eksponent, in $0 \leq f < 1$ število $0.c_2c_3 \ldots c_{24}$. Velja:

$0 < e < 255$	poljuben f	$x = (-1)^s(1+f) \cdot 2^{e-127}$
$e = 255$	$f = 0$	$x = (-1)^s \cdot \infty$
$e = 255$	$f \neq 0$	$x = \mathrm{NaN}$
$e = 0$	$f = 0$	$x = (-1)^s \cdot 0$
$e = 0$	$f \neq 0$	$x = (-1)^s \cdot (0+f) \cdot 2^{-126}$

Primeri pozitivnih števil:

e	f	število
10000010	01100000000000000000000	$x = (1 + 2^{-2} + 2^{-3}) \cdot 2^{130-127} = 11$
11111111	00000000000000000000000	$x = \infty$
11111111	01011010100000000000000	$x = \mathrm{NaN}$
00000000	00000000000000000000000	$x = 0$
00000000	00000100000000000000000	$x = 2^{-6} \cdot 2^{-126} = 2^{-132}$

Natančnost IEEE zapisa v dvojni dolžini

- po molekularni fiziki je v metru zraka (daljica) približno $3 \cdot 10^8$ molekul
- obseg Zemlje je približno $4 \cdot 10^7$ metrov, torej je v krogu okrog Zemlje okrog 10^{16} molekul.
- med 1 in 2 je 2^{52} števil v premični piki (približno 10^{16}).
- če postavimo krog okrog Zemlje s koordinatnim sistemom med 1 in 2, bodo števila v premični piki razporejena približno tako gosto kot molekule zraka.

Osnovna zaokrožitvena napaka

Števila, ki niso predstavljiva, predstavimo s približki, ki jih dobimo z zaokrožanjem. Naj bo x število in $fl(x)$ najbližje predstavljivo število. Velja

$$fl(x) = x(1+\delta), \text{ in } |\delta| \leq u,$$

kjer je

$$u = \frac{1}{2} b^{1-t}$$

osnovna zaokrožitvena napaka:

- single: $u = 2^{-24} = 6 \cdot 10^{-8}$,
- double: $u = 2^{-53} = 1 \cdot 10^{-16}$.

Izrek

Če število x leži znotraj intervala predstavljivih števil, velja

$$\frac{|fl(x) - x|}{|x|} \leq \frac{u}{1+u} \approx u.$$

Računanje v IEEE aritmetiki

Standard IEEE zagotavlja, da velja:

- $fl(x \circ y) = (x \circ y)(1+\delta), |\delta| \leq u$ za $\circ = +, -, *, /$,
- $fl(\sqrt{x}) = \sqrt{x}(1+\delta), |\delta| \leq u$.

Izjema je, če pride do prekoračitve (overflow) ali podkoračitve (underflow) obsega predstavljivih števil. V tem primeru dobimo po IEEE:

- prekoračitev: $\pm\infty$,
- podkoračitev: 0.

Nesreča rakete Ariane

4. junija 1996 je pri prvem poletu rakete Ariane 5 prišlo do nesreče: raketa je po 40 sekundah zavila s poti in eksplodirala. Do nesreče je prišlo zaradi prekoračitve obsega. Ker program ni imel testiranja prekoračitve, se je sesul, s tem pa tudi celoten polet.

Program, ki je povzročil napako, je prišel iz programa za starejšo Ariane 4. Tokrat je močnejša raketa povzročila, da so bile izmerjene količine prevelike in prišlo je do prekoračitve obsega.

Več na: http://www.fmf.uni-lj.si/∼jaklicg/arianne.htm

Napaka numerične metode

Računamo vrednost funkcije $f : X \to Y$ pri danem x.
Numerična metoda vrne približek $\widehat{y}$ za $y = f(x)$, razlika $D = y - \widehat{y}$ je celotna napaka približka.

Izvori napake so:

- nenatančnost začetnih podatkov,
- napaka numerične metode,
- zaokrožitvene napake med računanjem.

Celotna napaka je iz 3 delov

1. Neodstranljiva napaka: namesto z x računamo s približkom $\bar{x}$ in namesto $y = f(x)$ izračunamo $\bar{y} = f(\bar{x})$. Neodstranljiva napaka je $D_n = y - \bar{y}$ in je posledica napak podatkov.
2. Napaka metode: namesto f z numerično metodo računamo vrednost funkcije g. Namesto $\bar{y} = f(\bar{x})$ izračunamo $\widetilde{y} = g(\bar{x})$. Napaka metode je $D_m = \bar{y} - \widetilde{y}$. Običajno je posledica dejstva, da neskončno iteracijo zaključimo v končno korakih.
3. Zaokrožitvena napaka: pri izračunu $\widetilde{y} = g(\bar{x})$ se pri vsaki računski operaciji pojavi zaokrožitvena napaka, namesto $\widetilde{y}$ izračunamo $\widehat{y}$. Vrednost $\widehat{y}$ je odvisna od načina izračuna in vrstnega reda operacij. Zaokrožitvena napaka: $D_z = \widetilde{y} - \widehat{y}$.

Celotna napaka: $D = D_n + D_m + D_z$ in velja

$$|D| \leq |D_n| + |D_m| + |D_z|.$$

Računanje $\sin(\pi/10)$ v $P(10, 4, -5, 5)$

1. D_n: namesto z $x = \pi/10$ računamo z $\bar{x} = 0.3142 \cdot 10^0$

$$D_n = y - \bar{y} = \sin(\pi/10) - \sin(0.3142) = -3.9 \cdot 10^{-5}$$

2. D_m: namesto $\sin(\bar{x})$ izračunamo $g(\bar{x})$ za $g(x) = x - x^3/6$.

$$D_m = \bar{y} - \widetilde{y} = 2.5 \cdot 10^{-5}$$

3. D_z: odvisna je od vrstnega reda in načina računanja $g(\bar{x})$.

$$\begin{aligned} a_1 &= fl(\bar{x} \cdot \bar{x}) = fl(0.09872164) = 0.9872 \cdot 10^{-1} \\ a_2 &= fl(a_1 \cdot \bar{x}) = fl(0.03101154) = 0.3101 \cdot 10^{-1} \\ a_3 &= fl(a_2/6) = fl(0.0051683\ldots) = 0.5168 \cdot 10^{-2} \\ \widehat{y} &= fl(\bar{x} - a_3) = fl(0.309032) = 0.3090 \cdot 10^0 \end{aligned}$$

$$D_z = \widetilde{y} - \widehat{y} = 3.0 \cdot 10^{-5}$$

Celotna napaka je $D = D_n + D_m + D_z = 1.6 \cdot 10^{-5}$.