Numerična matematika

izročki predavanj

Matej Filip

Fakulteta za elektrotehniko
Univerza v Ljubljani

Viri

Viri v slovenščini:

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

Tuji viri - numerična linearna algebra:

- G.H. Golub, C.F. Van Loan: Matrix Computations, 3rd edition, Johns Hopkins Univ. Press, Baltimore, 1996.
- L.N. Trefethen, D. Bau: Numerical Linear Algebra, SIAM, Philadelphia, 1997.
- J.W. Demmel: Applied Numerical Linear Algebra, SIAM, 1997.

Tuji viri - numerična analiza:

- K. Atkinson, W. Han: Elementary Numerical Analysis, 3rd edition, John Wiley & Sons, Inc., New Jersey, 2003.
- R.L. Burden, J.D. Faires, A.M. Burden: Numerical Analysis, 10th edition, Cengage Learning, Boston, 2016.
- D.R. Kincaid, E.W. Cheney: Numerical Analysis, Mathematics of Scientific Computing, 3rd edition, Brooks/Cole, Pacific Grove, 2002.

Obveznosti

- Predavanja: 2 ure na teden
- Vaje: 2 uri na teden
- Pisni izpit: 50% ocene
- Izpit iz teorije: 50% ocene

Vsebina predmeta

1. Računanje in vloga napak pri numerični matematiki
2. Reševanje sistemov linearnih enačb
3. Reševanje nelinearnih enačb
4. Numerično odvajanje in integriranje
5. Numerično reševanje diferencialnih enačb

Prvo poglavje:

Uvod v numerično računanje

- Numerično računanje
- Predstavljiva števila
- Zaokrožitvene napake
- Katastrofalno seštevanje/odštevanje
- Primeri (ne)stabilnega računanja

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:*

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

- *Simbolno:*

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

Numerično in simbolno računanje

Primer

```
>> x=rand;(x^2-1)/(x+1)-(x-1)

ans=1.3877787807814446e-17
```

Analitično bi rezultat moral biti 0, vendar zaradi numeričnih napak dobimo majhno napako.

Kaj zanima numerično matematiko?

Metoda. . . matematična konstrukcija, s katero rešujemo problem

Algoritem. . . koraki metode

Implementacija. . . zapis algoritma v izbranem jeziku

Kaj pomeni ‘biti numerično dober’?

majhna sprememba podatkov $\Rightarrow$ majhna napaka rezultata

Tipična vprašanja numerične matematike:

- 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?

Občutljivih problemov NM ne more rešiti

Problem je občutljiv, če se ob majhni spremembi začetnih podatkov točen rezultat zelo spremeni.

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

Primer (presečišča premic)

Sistem in njegova perturbacija

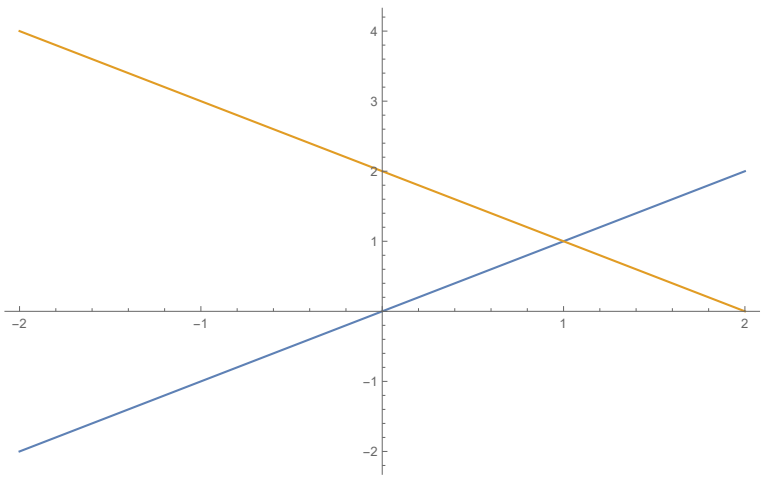
$$\begin{aligned} x + y = 2 \quad &\rightarrow \quad x + y = 1.9999 \\ x - y = 0 \quad &\rightarrow \quad x - y = 0.0002 \end{aligned}$$

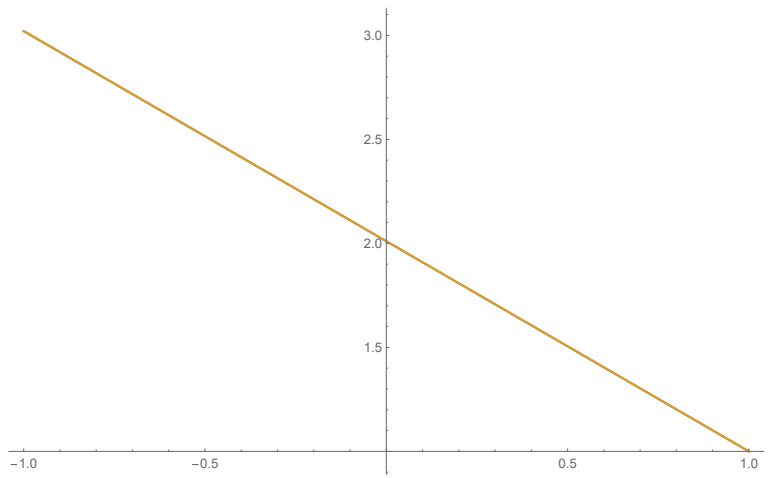
ima rešitvi $x = y = 1$ *oz.* $x = 1.00005$ *in* $y = 0.99985$. *Problem je neobčutljiv, saj je šlo za spremembo za isti velikostni razred.*

Sistem in njegova perturbacija

$$x + 0.99y = 1.99 \quad \rightarrow \quad x + 0.99y = 1.9899$$
$$0.99x + 0.98y = 1.97 \quad \rightarrow \quad 0.99x + 0.98y = 1.9701$$

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





Na čem temeljijo numerične metode?

- Matrike nadomestimo z enostavnejšimi (upoštevamo samo diagonalni ali zgornjetrikotni del).
- Nelinearne probleme nadomestimo z linearnimi (linearna aproksimacija v točki).
- Neskončne procese nadomestimo s končnimi (uporabimo Taylorjev polinom) .
- Neskončno razsežne prostore nadomestimo s končno razsežnimi (funkcije nadomestimo s polinomi).
- Diferencialne enačbe nadomestimo z algebraičnimi (znebimo se vseh parcialnih odvodov iz enačb).

Zakaj sploh potrebujemo numerično matematiko?

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, Zalivska vojna 1991, Savdska Arabija*, 28 žrtev: slaba analiza zaokrožitvenih napak.

  Čas zadetka iraške rakete, usmerjene na Savdsko Arabijo, je bil računan na vsako desetino sekunde v 24-bitnem sistemu. Ker velja

  $$\frac{1}{10} = 2^{-4} + 2^{-5} + 2^{-8} + 2^{-9} + 2^{-12} + 2^{-13} + 2^{-16} + 2^{-17} + 2^{-20} + 2^{-21} + \underbrace{+2^{-24} + 2^{-25} + 2^{-28} + \ldots}_{\text{zanemarimo}},$$

  je vsako desetinko sekunde napaka $9.5 \cdot 10^{-8}\ \mathrm{s}$. Po 100 urah računanja je bila napaka $9.5 \cdot 10^{-8}\ \mathrm{s} \cdot 100 \cdot 60 \cdot 60 \cdot 10 = 0.34\ \mathrm{s}$. Ker je hitrost rakete $1.676\ \mathrm{m/s}$, je bila pozicija rakete za več kot $500\ \mathrm{m}$ napačno predvidena in je ta ušla radarjem.

- *Eksplozija rakete Ariana 5, Francoska Gvajana, 1996*: posledica prekoračitve obsega števil.
  `https://www.youtube.com/watch?v=PK_yguLapgA`
  `https://www.youtube.com/watch?v=W3YJeoYgozw`

  Ob prenovi rakete so 'pozabili' nadgraditi uporabljen številski sistem, ki je horizontalno hitrost meril v 16-bitnem sistemu (1 bit porabimo za predznak). Največja hitrost v tem sistemu je

  $$2^0 + 2^1 + \ldots + 2^{13} + 2^{14} = \frac{2^{15} - 1}{2 - 1} = 32767.$$

  Ker je prenovljena raketa po 37 sekundah presegla to hitrost, je prišlo do zaustavitve motorjev...

- *Potop naftne ploščadi Sleipner A, Stavanger, Norveška, 1991*, miljarda dolarjev škode: nenatančna obdelava obremenitev pri reševanju PDE-jev.
  `https://www.youtube.com/watch?v=eGdiPs4THW8`

Ponovitev predstavljivih števil

Števila shranjujemo v obliki

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

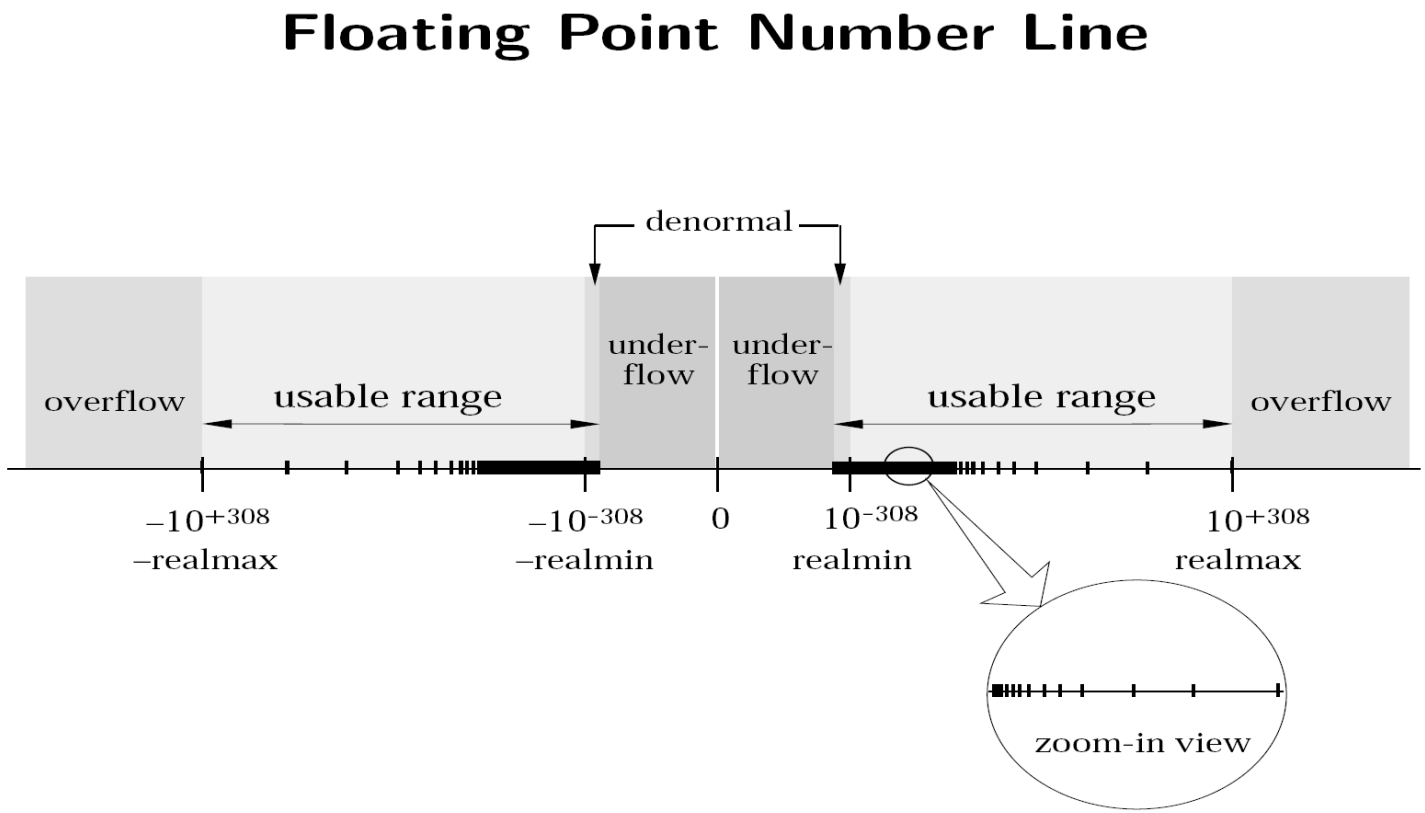
kjer je

- β naravno število (v računalništvu $\beta = 2$),
- $d_1 d_2 d_3 \ldots d_m$ mantisa, e eksponent.

Primer (baza 10)

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

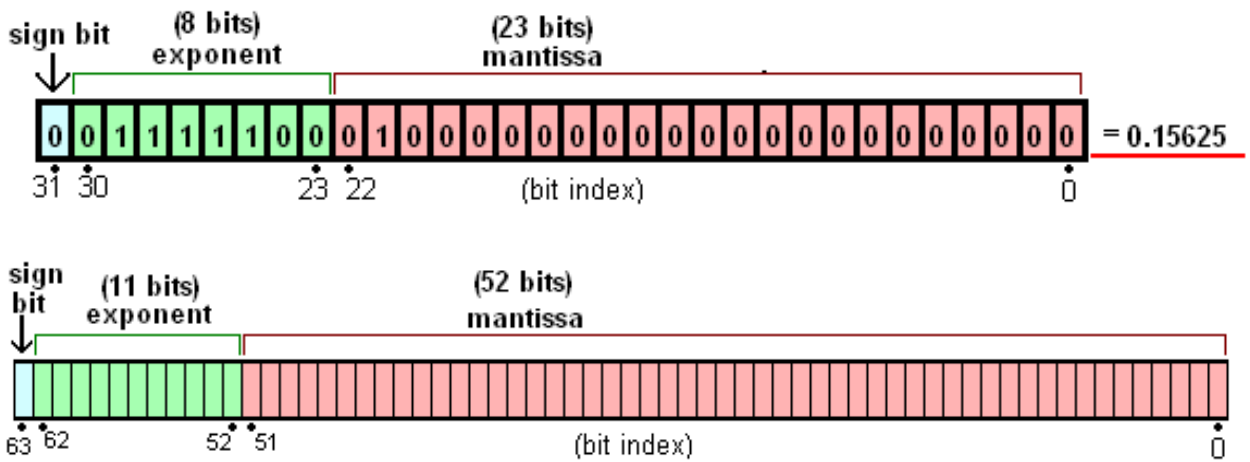
Prekoračitev in podkoračitev



- 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.



Kaj so zaokrožitvene napake?

- Večine realnih števil ne moremo predstaviti v strojni aritmetiki $\Rightarrow$ zaokrožujemo in delamo zaokrožitvene napake.

- IEEE standard... zaokroži x do najbližjega predstavljivega števila $\mathrm{fl}(x)$. Naj bosta

$$x_- \leqslant x \leqslant x_+$$

najbližji predstavljivi številili števila x. Potem je

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

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

$$\begin{aligned} x &= (0.b_1 b_2 b_3 \ldots b_m b_{m+1})_2 \times 2^e, \\ x_- &= (0.b_1 b_2 b_3 \ldots b_m)_2 \times 2^e, \\ x_+ &= \left((0.b_1 b_2 b_3 \ldots b_m)_2 + 2^{-m}\right) \times 2^e, \end{aligned}$$

$\mathrm{fl}(x) = x(1+\delta),\ |\delta| < 2^{-m}$

Absolutna napaka:

$$x - x_- \leqslant \frac{x_+ - x_-}{2} = 2^{e-m-1}.$$

Relativna napaka:

$$\frac{x - x_-}{x} \leqslant \frac{2^{e-m-1}}{1/2 \times 2^e} \leqslant \underbrace{2^{-m}}_{u} \ldots \text{osnovna zaokrožitvena napaka}$$

Torej je

$$x_- = x_- - x + x \geqslant -ux + x = x(1-u).$$

Podobno

$$x_+ \leqslant x(1+u).$$

Sledi

$$\boxed{\mathrm{fl}(x) = x(1+\delta)}, \quad \text{kjer je } |\delta| < u.$$

Kako računamo s predstavljivimi števili?

Za predstavljivi števili x, y in katerokoli od osnovnih operacij $\odot \in \{+, -, \cdot, :\}$ število $x \odot y$ ni nujno predstavljivo. Po zgornjem pa velja

$$\boxed{\mathrm{fl}(x \odot y) = (x \odot y)(1 + \delta)}, \quad \text{kjer je } |\delta| \leqslant u.$$

Seštevanje numerično ni asociativna operacija, tj.

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

Primer

```
>> a=rand;b=rand;c=rand;((a+b)+c)-(a+(b+c))

 ans=-2.220446049250313e-16
```

Seštevamo od manjših k večjim številom

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

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 $\delta_3\delta_4$ (Zakaj to lahko naredimo?), 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 števila, je za čim manjšo napako 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 D je

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

Primer ($\sin \frac{\pi}{10}$ računamo v desetiškem sistemu z $m = 4$)

- D_n: $fl(\frac{\pi}{10}) = 0.3142 \cdot 10^0$. *Ocenimo:* $|D_n| \approx \sin'(\frac{\pi}{10})|x - fl(x)| \leqslant \frac{1}{2} \cdot 10^{-4}$.
- D_m : $\sin x \approx x - x^3/6$. *Ocenimo:* $|D_m|| \leqslant x^5/120 = 2.6 \cdot 10^{-5}$.
- D_z: $fl(x - fl(fl(fl(x \cdot x) \cdot x)/6))$. *Ocenimo:* $|D_z| \leqslant 3.0 \cdot 10^{-5}$.

Stabilnost meri kakovost 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: $\boxed{\dfrac{\bar{x} - x}{x}}$.
- Druga delitev:
  - Direktna napaka: Numerična napaka rezultata.
  - Obratna napaka: Koliko je potrebno spremeniti začetne podatke, da dobimo izračunan rezultat.

Velja

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

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

Odštevanje in seštevanje sta lahko ‘katastrofalni’

odštevanje dveh približno enakih števil

seštevanje dveh približno nasprotnih števil

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

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

Potem

$$\begin{array}{lll} & \overbrace{x.xxx\ xxxx\ xxx1}^{\text{končna natančnost}} & \\ - & x.xxx\ xxxx\ xxx0 & \\ \hline = & 0.000\,0000\,0001 & ????\ ???? \\ = & 1.\ \underbrace{????\ ????}_{\text{izguba natančnosti}} \cdot \beta^{-m} & \end{array}$$

S ponavljanjem se napake seštevajo.

Primer katastrofalnega odštevanja

Iščemo rešitve kvadratne enačbe

$$x^2 + 2ax + b = 0, \quad \text{kjer je } a > 0 \text{ in } a^2 > b.$$

Rešitev z manjšo absolutno vrednostjo je

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

1 $k_1 := a^2$
2 $k_2 := k_1 - b$
3 $k_3 := \sqrt{k_2}$
4 $k_4 := -a + k_3$

Če je a^2 veliko večji od b, potem ima lahko korak 4 veliko napako. Možna rešitev:

$$x_2 = (-a + \sqrt{a^2 - b}) \cdot \frac{a + \sqrt{a^2 - b}}{a + \sqrt{a^2 - b}} = \frac{-b}{a + \sqrt{a^2 - b}}.$$

$$k_1 := a^2$$
$$k_2 := k_1 - b$$
$$k_3 := \sqrt{k_2}$$
$$k_4 := a + k_3$$
$$k_5 := \frac{-b}{k_4}$$

```
>> a = 10000;
>> b = -1;
>> x = -a+sqrt(a^2 - b)
x = 5.000000055588316e-05

>> x^2 + 2 * a * x +b
ans = 1.361766321927860e-08

>> x = -b/(a+sqrt(a^2-b))
x = 4.999999987500000e-05

>> x^2 + 2 * a * x +b
ans = -1.110223024625157e-16
```

Računanje s stabilnejšo obliko

- Izračun vrednosti funkcije

$$f(x) = x(\sqrt{x+1} - \sqrt{x})$$

ni stabilen za velike x, ker je $\sqrt{x+1} \approx \sqrt{x}$. Tej težavi se lahko izognemo:

$$f(x) = f(x) \cdot \frac{\sqrt{x+1} + \sqrt{x}}{\sqrt{x+1} + \sqrt{x}} = \frac{x}{\sqrt{x+1} + \sqrt{x}}.$$

- Vrsto

$$\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \ldots + \frac{1}{n(n+1)},$$

ki se sešteje v $\frac{n}{n+1}$ (dokaz: indukcija), je bolje numerično računati vzvratno kot

$$\frac{1}{n \cdot (n+1)} + \frac{1}{(n-1) \cdot n} + \ldots + \frac{1}{1 \cdot 2}.$$

Seštevanje in odštevanje v splošnem nista relativno direktno stabilni operaciji

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

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

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

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

Torej:

Če je $x + y$ blizu 0, potem je $\dfrac{|\overline{p} - p|}{|p|}$ veliko.

Množenje (in deljenje) je relativno 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.

Večina numeričnih metod ni relativno direktno stabilnih

Vse numerične metode, kjer sta vključeni

operaciji $+$ /$-$

in kot rezultat lahko dobimo npr. vrednost 0 ali nekje po poti kot vmesno vrednost skoraj singularno matriko, niso relativno direktno stabilne, tj. v rezultatu je lahko veliko relativna napaka.

Zato moramo vedno premisliti:

1. V katerih primerih so zgodi velika napaka?
2. Kako nestabilne primere preoblikovati v stabilne?

Primeri takih operacij:

- Računanje vrednosti polinoma.
- Računanje skalarnega produkta.
- Reševanje linearnega sistema.
- $\vdots$

Drugo poglavje:

Linearni sistemi

$Ax = b$

- Vektorske in matrične norme
- Pogojenostno število $\kappa(A)$
- Direktne metode za reševanje
  - *LU* razcep
  - Pivotna rast $\rho(A)$
  - Razcep Choleskega
- Predoločeni sistemi
  - *QR* razcep
  - Householderjeva zrcaljenja
  - *SVD* razcep

Vektorska norma je preslikava $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$, ki zadošča:

1. Pozitivna definitnost: $\|x\| \geqslant 0$ za vsak $x \in \mathbb{C}^n$ in $\|x\| = 0 \Leftrightarrow x = 0$.
2. Homogenost: $\|\alpha x\| = |\alpha| \, \|x\|$ za vsaka $\alpha \in \mathbb{C}$ in $x \in \mathbb{C}^n$
3. Trikotniška neenakost: $\|x + y\| \leqslant \|x\| + \|y\|$ za vsaka $x, y \in \mathbb{C}^n$.

Primer

Naj bo $x = (x_1, \ldots, x_n) \in \mathbb{C}^n$.

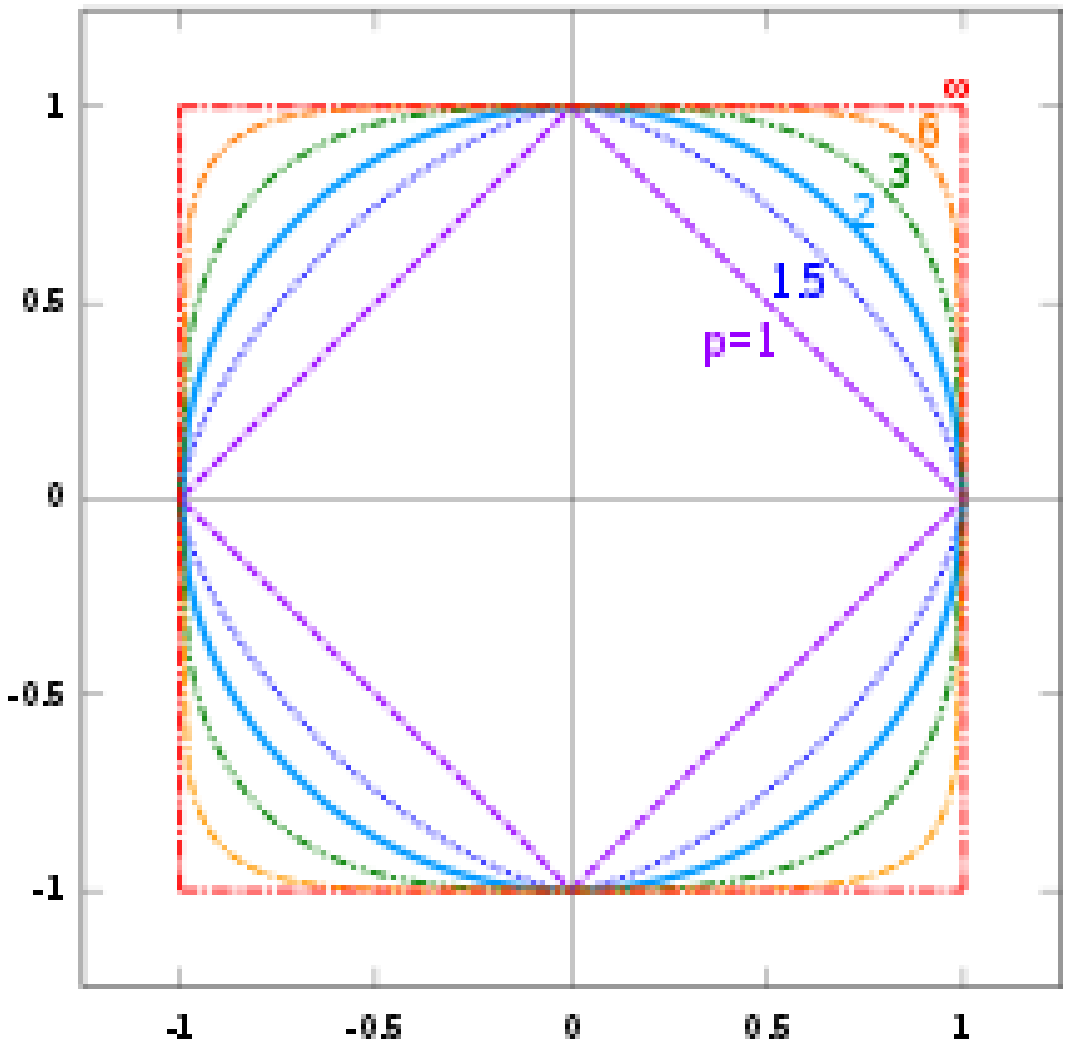
- *p–norma,* $p \in \mathbb{N}$*:*

$$\|x\|_p := \left(|x_1|^p + \ldots + |x_n|^p\right)^{1/p}.$$

- *Supremum norma:*

$$\|x\|_\infty = \max(|x_1|, \ldots, |x_n|).$$

Enotske krožnice v različnih normah



Matrična norma je preslikava $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$, ki zadošča:

1. Pozitivna definitnost: $\|A\| \geqslant 0$ za vsak $A \in \mathbb{C}^{n\times n}$ in $\|A\| = 0 \Leftrightarrow A = 0$.
2. Homogenost: $\|\alpha A\| = |\alpha| \, \|A\|$ za vsaka $\alpha \in \mathbb{C}$ in $A \in \mathbb{C}^{n\times n}$.
3. Trikotniška neenakost: $\|A + B\| \leqslant \|A\| + \|B\|$ za vsaka $A, B \in \mathbb{C}^{n\times n}$.
4. Submultiplikativnost: $\|AB\| \leqslant \|A\| \, \|B\|$ za vsaka $A, B \in \mathbb{C}^{n\times n}$.

Trditev

Naj bo $\|\cdot\|_*$ vektorska norma na $\mathbb{C}^n$. Potem predpis

$$\|A\|_* := \max_{\|x\|=1} \|Ax\|_* = \max_{x\neq 0} \frac{\|Ax\|_*}{\|x\|_*}.$$

določa matrično normo na $\mathbb{C}^{n\times n}$.

Dokaz

Naj bo $A = [a_{ij}]_{i,j=1,\ldots,n} \in \mathbb{C}^{n\times n}$ matrika. Nekaj matričnih norm:

1. 1–norma: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$ (1−norma),

$$\|A\|_1 = \max_{j=1,\ldots,n} \Big(\sum_{i=1}^{n} |a_{ij}| \Big). \quad \text{Dokaz}$$

2. Spektralna norma: Tu $\lambda_j(X)$ označuje j–to lastno vrednost matrike X.

$$\|A\|_2 = \sqrt{\max_{j=1,\ldots,n} \lambda_j(A^T A)}.$$

3. Frobeniusova norma:

$$\|A\|_F = \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2}.$$

4. Supremum norma:

$$\|A\|_\infty = \max_{i=1,\ldots,n} \Big(\sum_{j=1}^{n} |a_{ij}| \Big).$$

Zakaj imeti več matričnih norm?

Nekatere norme je bistveno zahtevneje izračunati od ostalih. Zahtevno je npr. določanje spektralne norme $\|\cdot\|_2$, saj je računanje lastnih vrednosti zahtevna naloga. Poceni pa je izračunati 1–normo, ∞–normo in F–normo. Iz različnih ocen, kot so

$$\begin{aligned}
\frac{1}{\sqrt{n}}\|A\|_F &\leqslant \|A\|_2 \leqslant \|A\|_F, \\
\frac{1}{\sqrt{n}}\|A\|_1 &\leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_1, \\
\frac{1}{\sqrt{n}}\|A\|_\infty &\leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_\infty,
\end{aligned}$$

pa lahko dobro ocenimo $\|A\|_2$.

Občutljivost sistema $Ax = b$

Zanima nas, kako na spremembo rešitve x vpliva napaka v začetnih podatkih, tj. napaka v A in b.

Zanima nas torej, kako velik je Δx v primeru majhnih perturbacij ΔA in Δb v rešitvi

$$(A + \Delta A)(x + \Delta x) = b + \Delta b. \tag{1}$$

Radi bi ocenili relativno napako $\frac{\|\Delta x\|}{\|x\|}$, kjer je $\|\cdot\|$ neka vektorska norma.

Izberimo vektorsko normo $\|\cdot\|$. Definirajmo občutljivost oz. pogojenostno število obrnljive matrike A v normi $\|\cdot\|$:

$$\kappa(A) = \|A\| \left\|A^{-1}\right\|.$$

$\kappa(A)$ meri občutljivost sistema $Ax = b$

Izrek

Naj bo A v (1) *obrnljiva matrika.*

1. *Privzemimo, da je $\Delta A = 0$. Potem velja:*

$$\frac{\|\Delta x\|}{\|x\|} \leqslant \kappa(A) \frac{\|\Delta b\|}{\|b\|}.$$

2. *Naj bo $\Delta A \neq 0$, naj za identično matriko I velja $\|I\| = 1$ in naj bo še $\left\|A^{-1}\right\| \|\Delta A\| < 1$. Potem velja:*

$$\frac{\|\Delta x\|}{\|x\|} \leqslant \frac{\kappa(A)}{1 - \kappa(A) \frac{\|\Delta A\|}{\|A\|}} \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|} \right).$$

Dokaz

Primer

1. *Če se spomnimo primera računanja prečišča dveh premic iz prvih predavanj, lahko vidimo, da je vprašanje občutljivosti glede na začetne podatke v resnici vprašanje občutljivosti matrik*

$$A_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad in \quad A_2 = \begin{pmatrix} 1.00 & 0.99 \\ 0.99 & 0.98 \end{pmatrix}.$$

Za njiju velja:

$$\kappa_1(A_1) = \kappa_F(A_1) = \kappa_\infty(A_1) = 2, \quad \kappa_2(A_1) = 1,$$
$$\kappa_1(A_2) = \kappa_\infty(A_2) = 3.96 \cdot 10^4, \quad \kappa_2(A_2) = 2, \quad \kappa_F(A_2) = 3.92 \cdot 10^4,$$

kjer κ_* označuje občutljivost v matrični normi $*$. Kot smo se geometrijsko prepričali, je drugi sistem res občutljiv, prvi pa ne.

2. *Primer zelo občutljive matrike je Hilbertova matrika*

$$H_n = \left[\frac{1}{i+j-1} \right]_{i,j} \in \mathbb{R}^{n \times n}.$$

Ta se pojavi pri iskanju polinoma, ki se v normi $\|f\| = \sqrt{\int_0^1 f^2 dx}$ najbolje prilega dani funkciji, saj je $\int_0^1 x^{i+j} dx = \frac{1}{i+j+1}$. Velja $\kappa_2(H_5) \approx 4.8 \cdot 10^5$, za naključno 5×5 matriko pa velja $\kappa_2 \approx 100$.

Primer

1. *Če z Matlabom z ukazom* $\backslash$ *rešimo sistem* $H_{15}x = v$*, kjer je*

$$v = H_{15} \cdot [1, \ldots, 1]^T,$$

bi morali dobiti za rezultat $x = [1, \ldots, 1]$*. Toda*

$$\|x - [1, \ldots, 1]^T\|_2 = 5.3 \cdot 10^{-3}.$$

Direktne metode

$$Ax = b$$

- Gaussova-eliminacija
- LU razcep
- Pivotiranje
- Pivotna rast
- Razcep Choleskega

Reševanje kvadratnih linearnih sistemov

Linearni sistem n enačb z n neznankami $x_1, \ldots, x_n$ je oblike

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2, \\ &\vdots \quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n, \end{aligned}$$

kjer so a_{ij}, b_j realna števila.

V matrični obliki ga zapišemo kot

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \ldots & \ldots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}}_{b}.$$

Geometrijski pomen sistema $Ax = b$

Naj bodo $a_{(1)}, a_{(2)}, \ldots, a_{(n)}$ stolpci matrike A, tj.,

$$a_{(i)} := \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{ni} \end{bmatrix} \in \mathbb{R}^n$$

Linearna kombinacija vektorjev $a_{(1)}, a_{(2)}, \ldots, a_{(n)}$ je vsak vektor oblike

$$x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{bmatrix} + \cdots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{nn} \end{bmatrix}, \qquad (2)$$

kjer so $x_i \in \mathbb{R}$ realna števila.

Zanima nas, ali obstaja linearna kombinacija (2), ki je enaka vektorju b.

Sistem $Ax = b$ z vidika numerične matematike

- Kako drago je reševanje sistema $Ax = b$?

  cena=število osnovnih računskih operacij $(+, -, \cdot, :)$.

- Kateri problemi in napake se pojavijo med reševanjem $Ax = b$?

  Ali obstajajo slabe matrike? Kako take matrike identificirati?

  Vemo že, da so slabe matrike z velikim $\kappa(A)$.

  Kaj pa, če $\kappa(A)$ *ni velik?*

- Za katere matrike se da enostavno in poceni rešiti tak sistem?

Ponovitev Gaussove eliminacije (GE)

Cilj je pretvoriti sistem v zgornjetrikotnega, nato pa ga rešiti z obratno substitucijo.

Primer

Rešujemo $Ax = b$, kjer sta

$$A = \begin{bmatrix} -3 & 2 & -1 \\ 6 & -6 & 7 \\ 3 & -4 & 4 \end{bmatrix}, \qquad b = \begin{bmatrix} -1 \\ -7 \\ -6 \end{bmatrix}.$$

Tvorimo razširjen sistem

$$\widetilde{A} = \left[\begin{array}{c|c} A & b \end{array}\right] = \left[\begin{array}{ccc|c} -3 & 2 & -1 & -1 \\ 6 & -6 & 7 & -7 \\ 3 & -4 & 4 & -6 \end{array}\right]$$

Prištejemo 2-kratnik prve vrstice drugi in 1-kratnik prve vrstice tretji.

$$\widetilde{A}_{(1)} = \left[\begin{array}{ccc|c} -3 & 2 & -1 & -1 \\ 0 & -2 & 5 & -9 \\ 0 & -2 & 3 & -7 \end{array}\right]$$

Primer

Odštejemo 1-kratnik druge vrstice od tretje

$$\widetilde{A}_{(2)} = \left[\begin{array}{rrr|r} -3 & 2 & -1 & -1 \\ 0 & -2 & 5 & -9 \\ 0 & 0 & -2 & 2 \end{array}\right]$$

Rešimo z obratno substitucijo

$$x_3 = \frac{2}{-2} = -1,$$
$$x_2 = \frac{1}{-2}\left(-9 - 5x_3\right) = 2,$$
$$x_1 = \frac{1}{-3}\left(-1 - 2x_2 + x_3\right) = 2.$$

V nadaljevanju bomo:

1. Prešteli število potrebnih računskih operacij za Gaussovo eliminacijo (GE).
2. GE bomo zapisali s pomočjo matričnih množenj.
3. Ukvarjali se bomo s stabilnostjo GE.

Algoritem GE in cena GE

```
1   -n × n matrika A = [a_ij]_ij in n × 1 vektor b = [b_i]_i
2   -preoblikujemo [A|b] v zgornjetrikotno z GE
3
4   for k = 1...n − 1
5     for i = k + 1...n
6       xmult = a_ik / a_kk
7       a_ik = 0
8       for j = k + 1...n
9         a_ij = a_ij − (xmult)a_kj
10      end
11      b_i = b_i − (xmult)b_k
12    end
13  end
```

Izrek

Število računskih operacij $(+, -, \cdot, :)$ *za prevedbo matrike* A *in razširjene matrike* $[A|b]$ *v zgornjetrikotno obliko je*

$$\frac{2}{3}n^3 + \mathcal{O}(n^2). \quad \textit{Dokaz}$$

Obratna substitucija in število operacij

```
1   -zgornjetrikotna n × n matrika U = [u_ij]_{i,j}, vektor
        c = [c_i]_i
2   -resimo sistem Ux = c
3
4   x_n = c_n / u_nn
5   for i = n − 1 ... 1
6      s = c_i
7      for j = i + 1 ... n
8         s = s − u_ij x_j
9      end
10     x_i = s / u_ii
11  end
```

Izrek

Število računskih operacij $(+, -, \cdot, :)$ *za rešitev sistem* $Ux = c$ *je*

$$n^2. \quad \textit{Dokaz}$$

Motivacija za zapis GE v matrični obliki

Videli smo, da je cena pretvorba matrike A oz. sistema $[A|b]$ v zgornjetrikotno obliko bistveno dražja kot pa obratna substitucija.

Če bomo v nekem postoku reševali sisteme $Ax = b$ pri fiksni matriki A, vektor b pa se bo spreminjal, bi bilo iz računskega vidika bistveno učinkoviteje preoblikovanje matrike A v zgornjetrikotno obliko narediti samo enkrat.

Ključno v tem procesu je ugotoviti, kako moramo preblikovati vektor b, ne da bi delali GE na razširjenem sistemu.

Eliminacijske matrike

Kako v matrični obliki zapišemo prištevanje večkratnika neke vrstice matrike k drugi?

Trditev

Prištejmo z $\alpha \in \mathbb{R}$ pomnoženo i–to vrstico matrike A njeni j–ti vrstici, kjer je $i \neq j$, in dobljeno matriko označimo z $\widetilde{A}$. Velja:

$$\widetilde{A} = A + \alpha \cdot E_{ij}A = (I_n + \alpha E_{ij})A,$$

kjer je E_{ij} matrika, ki ima v i–ti vrstici in j–tem stolpce 1, drugje pa 0, I_n pa identična matrika.

V prvem stolpcu pridelamo 0 pod diagonalo med GE na naslednji način:

$$A \mapsto (I_n - \frac{a_{21}}{a_{11}}E_{21})A \mapsto (I_n - \frac{a_{31}}{a_{11}}E_{31})(I_n - \frac{a_{21}}{a_{11}}E_{21})A \mapsto \cdots$$
$$\mapsto \underbrace{(I_n - \frac{a_{n1}}{a_{11}}E_{n1}) \cdots (I_n - \frac{a_{21}}{a_{11}}E_{21})}_{L_1} A =: A^{(1)}.$$

Eliminacijske matrike

Če odpravimo oklepaje v matriki L_1, in upoštevamo $E_{i1}E_{j1} = 0$, saj je $i, j > 1$, dobimo spodnjetrikotno matriko

$$L_1 = I_n - \frac{a_{21}}{a_{11}}E_{21} - \ldots - \frac{a_{n1}}{a_{11}}E_{n1}.$$

Z istim premislekom v drugem stolpcu nove matrike pridelamo 0 z množenjem $A^{(1)} = [a_{ij}^{(1)}]_{i,j}$ s spodnjetrikotno matriko

$$L_2 = I_n - \frac{a_{32}^{(1)}}{a_{22}^{(1)}}E_{32} - \ldots - \frac{a_{n2}^{(1)}}{a_{22}^{(1)}}E_{n2}.$$

Dobimo:

$$A^{(2)} = L_2A^{(1)} = L_2L_1A.$$

Na koncu GE dobimo s tem postopkom zgornjetrikotno matriko

$$U = L_{n-1}\cdots L_1A. \tag{3}$$

Eliminacijske matrike

Iz (3) bi radi izrazili A. Preprost račun pokaže, da inverze L_i^{-1} eliminacijskih matrik dobimo tako, da vse minuse v defincji L_i spremenimo v pluse:

$$L_1^{-1} = I_n + \frac{a_{21}}{a_{11}} E_{21} + \ldots + \frac{a_{n1}}{a_{11}} E_{n1},$$

$$L_2^{-1} = I_n + \frac{a_{32}^{(1)}}{a_{22}^{(1)}} E_{32} + \ldots + \frac{a_{n2}^{(1)}}{a_{22}^{(1)}} E_{n2},$$

$$\vdots$$

$$L_{n-1}^{-1} = I_n + \frac{a_{n,n-1}^{(n-2)}}{a_{n-1,n-1}^{(n-2)}} E_{n,n-1}$$

Iz (3) dobimo A na naslednji način:

$$A = \underbrace{L_1^{-1} L_2^{-1} \cdots L_{n-1}^{-1}}_{L} U.$$

Preprost račun pokaže (kjer upoštevamo $E_{ij}E_{k\ell} = 0$ za $j \neq k$), da velja

$$L = I_n + \frac{a_{21}}{a_{11}} E_{21} + \ldots + \frac{a_{n1}}{a_{11}} E_{n1} + \frac{a_{32}^{(1)}}{a_{22}^{(1)}} E_{32} + \ldots + \frac{a_{n,n-1}^{(n-2)}}{a_{n-1,n-1}^{(n-2)}} E_{n,n-1}$$

LU razcep matrike A

```
1   -Vhod: A = [a_ij]_{i,j} n × n matrika.
2   -Izhod: Spodnja trikotna matrika L in zgornja
        trikotna matrika U, da je A = LU
3   -ℓ_ik v spodnjem algoritmu so elementi pod
        diagonalo v L, na diagonali so same 1
4   -preostali elementi a_ij v zgornjem trikotniku so
        elementi matrike U
5
6   for k = 1, ..., n − 1
7     for i = k + 1, ..., n
8       ℓ_ik = a_ik / a_kk
9       for j = k + 1, ..., n
10        a_ij = a_ij − ℓ_ik a_kj
11      end
12    end
13  end
```

Izrek

Število računskih operacij $(+, -, \cdot, :)$ *za izračun LU razcepa matrike A je* $\frac{2}{3}n^3 + \mathcal{O}(n^2)$.

Prema substitucija in število operacij

```
1   -Vhod: spodnja trikotna n × n matrika L = [ℓ_ij]_{i,j} in
        vektor b = [b_i]_i
2   -Izhod: resitev y sistema Ly = b
3
4   y_1 = b_1/ℓ_11
5   for i = 2...n
6      s = b_i
7      for j = 1...i − 1
8         s = s − ℓ_ij y_j
9      end
10     y_i = s/ℓ_ii
11  end
```

Izrek

Število računskih operacij $(+, -, \cdot, :)$ *za rešitev sistem* $Ly = b$ *je*

$$n^2.$$

Reševanje sistema $Ax = b$ prek LU razcepa:

1. Izračunamo $A = LU$. Cena: $\frac{2}{3}n^3 + \mathcal{O}(n^2)$.
2. Rešimo $Ly = b$ s premo subsitucijo, tj. od y_1 proti y_n. Cena: $n^2 - n$.
3. Rešimo $Ux = y$ z obratno subsitucijo, t. od x_n proti x_1. Cena: n^2.

Cena preme substitucije je za n operacij manjša kot cena obratne substitucije, saj imamo ne diagonali L same enice in prihranimo v vsaki spremenljivki eno deljenje.

Reševanje sistema $Ax = b$ prek LU razcepa

Primer

$$A = \begin{pmatrix} 2 & 1 & 3 & -4 \\ -4 & -1 & -4 & 7 \\ 2 & 3 & 5 & -3 \\ -2 & -2 & -7 & 9 \end{pmatrix}, \quad b = \begin{pmatrix} 8 \\ -14 \\ 7 \\ -16 \end{pmatrix}.$$

1. $L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ -1 & -1 & 1 & 1 \end{pmatrix}, U = \begin{pmatrix} 2 & 1 & 3 & -4 \\ 0 & 1 & 2 & -1 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$
2. *Rešimo* $Ly = b$ *in dobimo* $y = \begin{pmatrix} 8 & 2 & -5 & -1 \end{pmatrix}^T$.
3. *Rešimo* $Ux = y$ *in dobimo* $x = \begin{pmatrix} 1 & -1 & 1 & -1 \end{pmatrix}^T$.

Obstoj LU razcep matrike

V nadaljevanju se bomo ukvarjali z obstojem in stabilnostjo *LU* razcepa.

Problematična sta npr. matriki

$$A = \begin{bmatrix} 0 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \qquad B = \begin{bmatrix} 10^{-17} & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix},$$

saj je 10^{-17} pod strojnim ϵ. Da pa se natančno povedati, kdaj *LU* razcep obstaja.

Podmatriki matrike $A \in \mathbb{R}^{n \times n}$, zožene na prvih k vrstic in stolpcev, pravimo k–ta glavna vodilna podmatrika.

Izrek (Obstoj LU razcepa)

Za $n \times n$ matriko A sta naslednji trditvi ekvivalentni:

1. *LU razcep matrike A obstaja in je enoličen.*
2. *k-ta glavna vodilna podmatrika matrike A je obrnljiva za vsak $k = 1, \ldots, n$.*

LU razcep z delnim pivotiranjem

Pri delnem pivotiranju pred eliminacijo v j-tem stolpcu primerjamo elemente

$$a_{jj}, a_{j+1,j}, \ldots, a_{nj},$$

nato pa zamenjamo j-to vrstico s tisto, ki vsebuje element z največjo absolutno vrednostjo.

Menjava j-te in k-te vrstice pa je množenje z leve s permutacijsko matriko P_{jk}, ki se od identitete razlikuje le v j-ti in k-ti vrstici, ki sta zamenjani:

$$P_{jk} = I_n - E_{jj} - E_{kk} + E_{jk} + E_{kj}.$$

Tu so E_{ij} standardne koordinatne matrike (1 v i-ti vrstici in j-tem stolpcu in 0 drugje).

LU razcep z delnim pivotiranjem

Izrek (LU razcep z delnim pivotiranjem)

Če izvedemo Gaussovo eliminacijo z delnim pivotiranjem matrike $A \in \mathbb{R}^{(n-1)\times(n-1)}$ *do zgornje trikotne matrike* U *z zaporedjem transformacij:*

$$\begin{aligned} A &\mapsto P_1 A \mapsto L_1 P_1 A_1 \mapsto P_2 L_1 P_1 A_1 \mapsto \cdots \\ &\mapsto L_{n-1} P_{n-1} \cdots L_1 P_1 A =: U, \end{aligned}$$

kjer so P_i *permutacijske,* L_i *pa eliminacijske matrike uporabljene na* i*–tem koraku postopka. Potem velja*

$$PA = LU$$

kjer so

$$P = P_{n-1} P_{n-2} \cdots P_1, \qquad L = \widehat{L}_1^{-1} \widehat{L}_2^{-1} \cdots \widehat{L}_{n-1}^{-1},$$
$$\widehat{L}_i^{-1} = P^{(i)} L_i^{-1} (P^{(i)})^T, \qquad P^{(i)} = P_{n-1} \cdots P_{i+1}.$$

Dokaz

LU razcep z delnim pivotiranjem - algoritem

```
1   -Vhod: A = [a_ij]_{i,j} n × n matrika
2   -Izhod: permutacijska matrika P, spodnja in
        zgornja trikotna matrika L in U, da je
        PA = LU
3
4   P in L identicni n × n matriki
5   for k = 1,...,n−1
6     poisci q-to in k-to vrstico, ki zadosca
          |a_qk| = max_{k≤p≤n} |a_pk|
7     zamenjaj q-to in k-to vrstico v matrikah A, P
          in strogem spodnjem trikotniku L
8     for i = k+1,...,n
9       ℓ_ik = a_ik / a_kk
10      for j = k+1,...,n
11        a_ij = a_ij − ℓ_ik a_kj
12        end
13    end
14   end
```

LU razcep z delnim pivotiranjem

Izrek (O računski zahtevnosti LU razcep z delnim pivotiranjem)

Število računskih operacij $(+, -, \cdot, :)$ *za izračun LU razcepa z delnim pivotiranjem je* $\frac{2}{3}n^3 + \mathcal{O}(n^2)$.

Dodatno delo pri LU razcepu z delnim pivotiranjem je $\mathcal{O}(n^2)$ primerjanj in menjav.

Reševanje $Ax = b$ prek LU razcepa z delnim pivotiranjem:

1. Izračunamo $PA = LU$. Cena: $\frac{2}{3}n^3 + \mathcal{O}(n^2)$.
2. Rešimo $Ly = Pb$ s premo subsitucijo. Cena: $n^2 - n$.
3. Rešimo $Ux = y$ z obratno subsitucijo. Cena: n^2.

Izrek (Obstoj LU razcepa z delnim pivotiranjem)

Za $n \times n$ *matriko* A *sta naslednji trditvi ekvivalentni:*

1. *LU razcep matrike* A *z delnim pivotiranjem obstaja.*
2. *Matrika* A *je obrnljiva.*

$Ax = b$ prek LU razcepa z delnim pivotiranjem

Primer.

$$A = \begin{pmatrix} 2 & 1 & 3 & -4 \\ -4 & -1 & -4 & 7 \\ 2 & 3 & 5 & -3 \\ -2 & -2 & -7 & 9 \end{pmatrix}, \quad b = \begin{pmatrix} 8 \\ -14 \\ 7 \\ -16 \end{pmatrix}.$$

1. $L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 & 0 \\ \frac{1}{2} & -\frac{3}{5} & 1 & 0 \\ -\frac{1}{2} & \frac{1}{5} & -\frac{1}{8} & 1 \end{pmatrix}, U = \begin{pmatrix} -4 & -1 & -4 & 7 \\ 0 & \frac{5}{2} & 3 & \frac{1}{2} \\ 0 & 0 & -\frac{16}{5} & \frac{58}{10} \\ 0 & 0 & 0 & \frac{1}{8} \end{pmatrix},$

   $P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$

2. Rešimo $Ly = Pb$ in dobimo $y = \begin{pmatrix} -14 & 0 & -9 & -\frac{1}{8} \end{pmatrix}^T$.

3. Rešimo $Ux = y$ in dobimo $x = \begin{pmatrix} 1 & -1 & 1 & -1 \end{pmatrix}^T$.

LU s kompletnim pivotiranjem

Pri kompletnem pivotiranju pred eliminacijo v j-tem stolpcu poiščemo element z največjo absolutno vrednostjo v podmatriki $A(j:n, j:n)$ in nato izvedemo ustrezni menjavi vrstic in stolpcev.

Dodatno delo pri LU razcepu s kompletnim pivotiranjem je $\mathcal{O}(n^3)$ primerjanj in menjav. Torej je skupna cena precej dražja od LU razcepa z delnim pivotiranjem. Ker bomo videli, da je LU razcep z delnim pivotiranjem statistično numerično stabilen, se v praksi kompletno pivotiranje redko uporablja.

Stabilnost LU razcepa matrike A

Sistem $Ax = b$ smo rešili prek LU razcepa in dobili približek $\widehat{x}$. Računali smo v treh korakih:

1. *Izračun LU razcepa:* $A + E = \widehat{L}\widehat{U}$.
2. *Prema substitucija:* $\widehat{L}\widehat{y} = b$.
3. *Obratna substitucija:* $\widehat{U}\widehat{x} = \widehat{y}$.

Izkaže se, da je (teoretično) nestabilen samo prvi korak.

Spomnimo se, da z u označujemo osnovno zaokrožitveno napako 2^{-m} kjer je m dolžina mantise. Z $|A| = [|a_{ij}|]_{i,j}$ označimo matriko absolutnih vrednosti vhodov matrike $A = [a_{ij}]_{i,j}$

Izrek (Ocena absolutne napake pri izračunu LU razcepa)

Naj bo $A \in \mathbb{R}^{n\times n}$ obrnljiva matrika, pri kateri se izvede LU razcep brez pivotiranja. Za izračunani matriki $\widehat{L}, \widehat{U}$ velja $A = \widehat{L}\widehat{U} + E$, kjer je

$$|E| \leqslant 3(n-1)u\left(|A| + |\widehat{L}||\widehat{U}|\right) + \mathcal{O}(u^2). \quad \textit{Dokaz}$$

Stabilnost LU razcepa matrike A

Iz zgornjega izreka sledi

$$\begin{aligned}\|E\|_\infty &\leqslant 3(n-1)u \cdot \left(\|A\|_\infty + (\|\widehat{L}\|\|\widehat{U}\|)_\infty\right) + \mathcal{O}(u^2)\\ &\leqslant 3(n-1)u \cdot \left(\|A\|_\infty + \|\widehat{L}\|_\infty\|\widehat{U}\|_\infty\right) + \mathcal{O}(u^2)\\ &\leqslant 3(n-1)u\|A\|_\infty + 3(n-1)n\|\widehat{U}\|_\infty + \mathcal{O}(u^2),\end{aligned}$$

kjer smo v drugi neenakosti upoštevali submultiplikativnost, v tretji neenakosti pa to, da so pri LU razcepu z delnim pivotiranjem vsi elementi matrike L navzgor omejeni z 1. Zato velja $\|L\|_\infty \leqslant n$. Torej je relativna napaka v supremum normi navzgor omejena z

$$\frac{\|E\|_\infty}{\|A\|_\infty} \leqslant 3(n-1)u + 3(n-1)nu \cdot \frac{\|\widehat{U}\|_\infty}{\|A\|_\infty} + \mathcal{O}(u^2).$$

Izrek (Ocena relativne napake pri izračunu LU razcepa)

Pri LU razcepu z delnim pivotiranjem velja ocena relativne napake:

$$\frac{\|E\|_\infty}{\|A\|_\infty} \leqslant 3(n-1)u + 3(n-1)nu \cdot \frac{\|\widehat{U}\|_\infty}{\|A\|_\infty} + \mathcal{O}(u^2).$$

Pivotna rast

Pivotna rast matrike A je definirana kot

$$\rho(A) := \frac{\max_{i,j} |\widehat{u}_{i,j}|}{\max_{i,j} |a_{i,j}|}.$$

Velja

$$\frac{\|\widehat{U}\|_\infty}{\|A\|_\infty} \leqslant n\rho(A).$$

Trditev

Pri delnem pivotiranju je pivotna rast omejena z 2^{n-1}.

Dokaz. Velja namreč $|\ell_{ij}| \leqslant 1$, a_{ij} pa na vsakem od največ $n-1$ korakov izračunamo kot

$$a_{ij} = a_{ij} - \ell_{ik} a_{kj}.$$

Torej se absolutna vrednost največjega elementa v matriki kvečjemu podvoji.

Pivotna rast pri delnem pivotiranju

Žal pa za vsak n obstajajo matrike s pivotno rastjo 2^{n-1}, tako da LU razcep z delnim pivotiranjem teoretično ni stabilen.

Primer

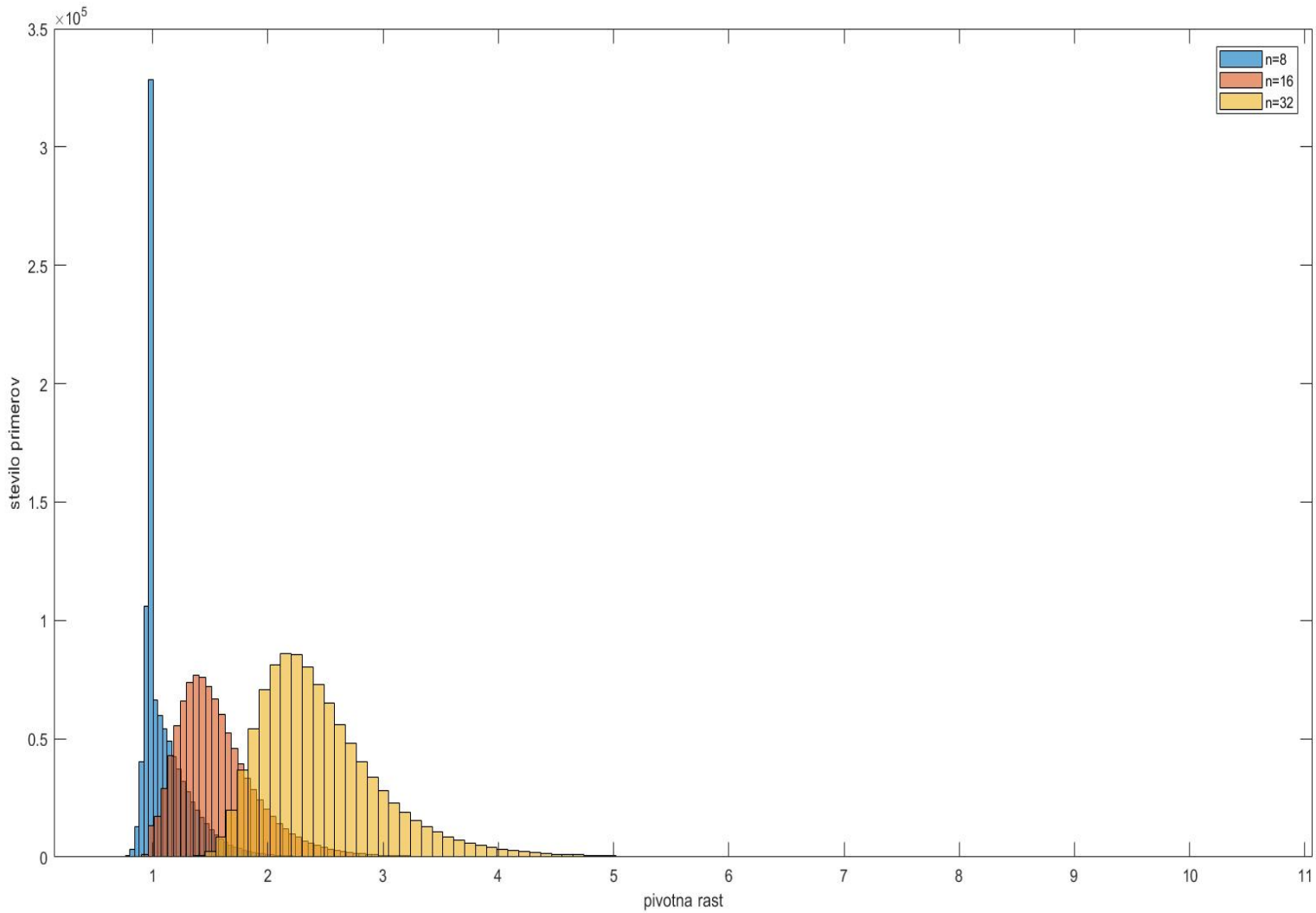
Matrika

$$A_n = \begin{pmatrix} 1 & 0 & \cdots & 0 & 1 \\ -1 & 1 & \ddots & \vdots & 1 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ -1 & \cdots & \cdots & -1 & 1 \end{pmatrix}$$

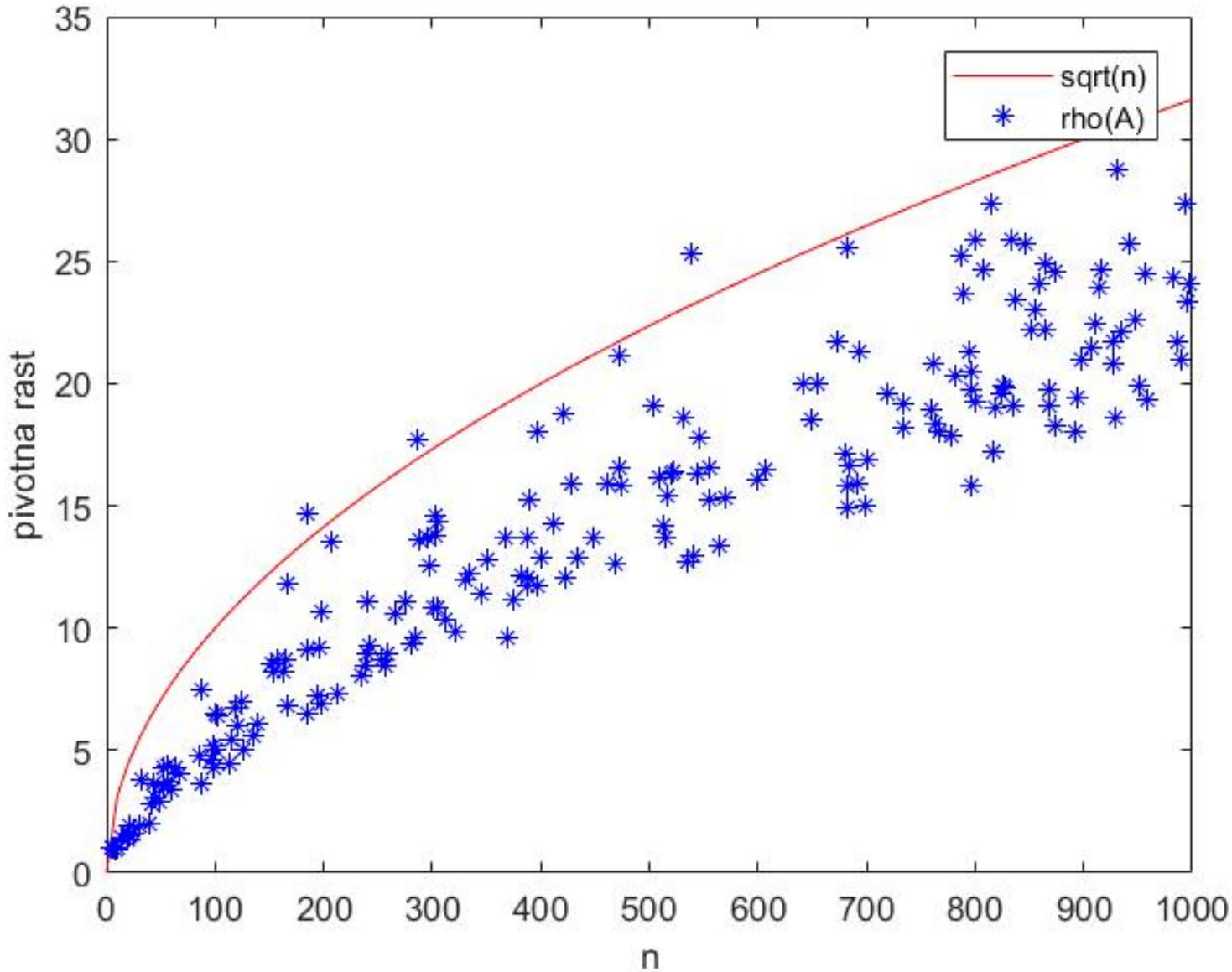
ima pivotno rast 2^{n-1}.

Statistično pa velja, da je pričakovana vrednost pivotne rasti $\mathcal{O}(n^{1/2})$, tako da LU razcep z delnim pivotiranjem v praksi je obratno stabilen.

Verjetnostne porazdelitve slučajne spremenljivke ρ, generirane z milijon naključnimi matrikami velikosti $n \times n$ (tj. vsak vhod naključen element iz enakomerne zvezne porazdelitve na intervalu $[0, 1]$):



Pivotna rast 200 naključnih matrik velikosti $n \times n$ (tj. vsak vhod naključen element iz enakomerne zvezne porazdelitve na intervalu $[0, 1]$):



LDL razcep simetrične matrike *A*

Trditev

Naj bo

$$A = A^T$$

$n \times n$ simetrična matrika in $A = LU$ njen LU razcep. Če je D diagonalna matrika, katere diagonala se ujema z diagonalo U-ja, potem je $U = DL^T$ in

$$A = LDL^T.$$

Dokaz. Velja

$$LU = A = A^T = (LU)^T = U^T L^T.$$

Z množenjem te verige enakosti z leve z L^{-1} in z desne z $(L^T)^{-1}$ dobimo

$$U(L^T)^{-1} = L^{-1}U^T =: D.$$

Ker je leva stran zgornja trikotna, desna pa spodnja trikotna, je D diagonalna matrika. Torej velja

$$A = LU = LDL^T.$$

Razcep Choleskega pozitivno definitne matrike A

Izrek (Razcep Choleskega)

Naj bo A simetrična in pozitivno definitna matrika, tj. za vsak $x \neq 0$ velja $x^T Ax > 0$. Potem obstaja spodnjetrikotna matrika V, da velja

$$A = VV^T.$$

Temu razcepu pravimo razcep Choleskega matrike A. Matrika V je enaka $V := LD^{1/2}$, kjer sta L in D matriki iz LDL razcepa matrike A.

Dokaz. Za obstoj je potrebno preveriti samo to, da $D^{1/2}$ res lahko izračunamo, tj. da so diagonalni elementi matrike D pozitivni. Da dobimo i-ti element na diagonali D, izračunamo $x^T Ax$ za vektor $x = (L^T)^{-1} e_i$, kjer je e_i i-ti stolpec identitete. Ker je A pozitivno definitna, je $x^T Ax > 0$.

Razcep Choleskega - algoritem

1 $A = [a_{ij}]_{i,j}$ je dana $n \times n$ matrika
2 ce se razcep v celoti izvede, je rezultat
spodnjetrikotna matrika V iz $A = VV^T$
3
4 for $k = 1, \ldots, n$
5 $v_{kk} = \sqrt{a_{kk} - \sum_{i=1}^{k-1} v_{ki}^2}$
6 for $j = k+1, \ldots, n$
7 for $i = 1, \ldots, k-1$
8 $a_{jk} = a_{jk} - v_{ji} v_{ki}$
9 end
10 $v_{jk} = a_{jk} / v_{kk}$
11 end
12 end

Razcep Choleskega

Izrek (Cena razcepa Choleskega)

Število računskih operacij $(+, -, \cdot, :)$ *za izračun razcepa Choleskega pozitivno definitne matrike* A *je* $\frac{n^3}{3} + \mathcal{O}(n^2)$.

Dokaz

Razcep Choleskega tako zahteva samo pol toliko operacij kot LU razcep in je najcenejši numerični način za ugotavljanje pozitivne definitnosti simetrične matrike.

Reševanje sistema $Ax = b$ prek razcepa Choleskega:

1. Izračunamo $A = VV^T$. Cena: $\frac{1}{3}n^3 + \mathcal{O}(n^2)$.
2. Rešimo $Vy = b$ s premo subsitucijo. Cena: $n^2 + \mathcal{O}(n)$.
3. Rešimo $V^T x = y$ z obratno subsitucijo. Cena: $n^2 + \mathcal{O}(n)$.

Izrek (Stabilnost računanja razcepa Choleskega)

Računanje razcepa Choleskega je numerično stabilna metoda.