Numerička analiza

2. predavanje

Saša Singer

singer@math.hr

web.math.hr/~singer

PMF – Matematički odjel, Zagreb

Sadržaj predavanja

- Uvodna priča o greškama:
  - Problemi numeričke matematike (zašto ona postoji).
  - Pojam greške, apsolutna i relativna greška.
  - Izvori grešaka — model, ulazni podaci (mjerenje), metoda, zaokruživanje.
    - Ilustracija grešaka na modelnim primjerima.

Sadržaj predavanja

- Prikaz brojeva u računalu i greške zaokruživanja:
  - Prikaz cijelih brojeva i tipične greške.
  - Prikaz realnih brojeva. IEEE standard.
  - Jedinična greška zaokruživanja.
  - Greške zaokruživanja osnovnih aritmetičkih operacija.
    - Opasno ili "katastrofalno" kraćenje.
  - "Širenje" grešaka (zaokruživanja), stabilni i nestabilni algoritmi.
  - Mjerenje grešaka — razne norme.
  - Uvjetovanost problema.
  - Stabilnost algoritma.
    - Primjeri stabilnih i nestabilnih algoritama.
  - Primjeri grešaka zaokruživanja.

Numerička matematika

Problemi numeričke matematike

U matematici postoji niz problema koje

- ne znamo ili ne možemo egzaktno riješiti,

tj. prisiljeni smo tražiti približno rješenje.

Neki klasični "zadaci" u numeričkom računanju su:

- rješavanje sustava linearnih i nelinearnih jednadžbi,
- računanje integrala,
- računanje aproksimacije neke zadane funkcije (zamjena podataka nekom funkcijom),
- minimizacija (maksimizacija) zadane funkcije, uz eventualna ograničenja (obično, u domeni),
- rješavanje diferencijalnih i integralnih jednadžbi ...

Problemi numeričke matematike (nastavak)

Neke probleme čak znamo egzaktno riješiti (bar u principu),

- poput sustava linearnih jednadžbi (ponoviti LA1),

no to predugo traje, pa koristimo računala.

Međutim, tada imamo dodatni problem, jer

- računala ne računaju egzaktno, već približno!

Oprez, tada ni osnovne aritmetičke operacije nisu egzaktne.

Dakle, ključni pojam u numerici je

- približna vrijednost, odnosno, greška.

Ciljevi numeričke matematike

U skladu s tim, osnovni zadatak numeričke matematike je naći (dati) odgovore na sljedeća pitanja:

- kako riješiti neki problem — metoda,
- koliko je "dobro" izračunato rješenje — točnost, ocjena greške.

Malo preciznije, za svaku od navedenih klasa problema, treba proučiti sljedeće "teme" — potprobleme:

1. Uvjetovanost problema — osjetljivost problema na greške, prvenstveno u početnim podacima (tzv. teorija pertubacije ili smetnje — vezana uz sam problem).
2. Konstrukcija standardnih numeričkih metoda za rješavanje danog problema.

Ciljevi numeričke matematike (nastavak)

Kad jednom "stignemo" do numeričkih metoda, treba još proučiti sljedeće "teme" — potprobleme:

3. Stabilnost numeričkih metoda — njihova osjetljivost na "smetnje" problema.
4. Efikasnost pojedine numeričke metode — orijentirano prema implementaciji na računalu:
   - broj računskih operacija i potreban memorijski prostor za rješavanje problema (= Složenost).
5. Točnost numeričkih metoda, u smislu neke "garancije" točnosti izračunatog rješenja.

Ilustracija ovih "potproblema" na primjerima — malo kasnije.

Greške

Greške

Pri numeričkom rješavanju nekog problema javljaju se različiti tipovi grešaka:

- greške modela — svođenje realnog problema na neki "matematički" problem,
- greške u ulaznim podacima (mjerenja i sl.),
- greške numeričkih metoda za rješavanje "matematičkog" problema,
- greške "približnog" računanja — obično su to
  - greške zaokruživanja u aritmetici računala.

Greške modela su "izvan" dosega numeričke matematike.

- Spadaju u fiziku, kemiju, biologiju, tehniku, ekonomiju, ...

Mjere za grešku

Oznake:

- prava vrijednost — x,
- izračunata ili približna vrijednost — $\hat{x}$.

Standardni naziv: $\hat{x}$ je aproksimacija za x.

Trenutno, nije bitno odakle (iz kojeg skupa) su x i $\hat{x}$.

- Zamislite da su to "obični" realni brojevi — $x, \hat{x} \in \mathbb{R}$.

Mjere za grešku (nastavak)

Apsolutna greška:

- mjeri udaljenost izračunate vrijednosti $\hat{x}$ obzirom na pravu vrijednost x.

Ako imamo vektorski prostor i normu, onda je

- udaljenost = norma razlike.

Dakle, apsolutna greška je definirana ovako:

$$E_{\mathrm{abs}}(x, \hat{x}) := |\hat{x} - x|.$$

Često se koristi i oznaka $\Delta x = \hat{x} - x$ (na pr. u analizi), pa je $E_{\mathrm{abs}}(x, \hat{x}) = |\Delta x|$.

Mjere za grešku (nastavak)

Primjer. Dojam o "veličini" greške:

- ako smo umjesto 1 izračunali 2, to nam se čini lošije nego
- ako smo umjesto 100 izračunali 101.

Relativna greška:

- mjeri relativnu točnost aproksimacije $\hat{x}$ obzirom na veličinu broja x,
- na pr. koliko se vodećih znamenki brojeva x i $\hat{x}$ podudara.

Relativna greška definirana je za $x \neq 0$,

$$E_{\mathrm{rel}}(x, \hat{x}) := \frac{|\hat{x} - x|}{|x|}.$$

Često se koristi i oznaka δ_x. Katkad se u nazivniku javlja $|\hat{x}|$.

Mjere za grešku (nastavak)

Ideja relativne greške: ako $\hat{x}$ napišemo kao $\hat{x} = x(1+\rho)$, onda je njegova relativna greška

$$E_{\mathrm{rel}}(x, \hat{x}) := |\rho|.$$

Dakle, relativna greška mjeri

- koliko se faktor $(1+\rho)$ apsolutno razlikuje od 1.

Sad možemo detaljnije opisati one četiri vrste grešaka:

- greške modela,
- greške u ulaznim podacima (mjerenjima),
- greške metoda za rješavanje modela,
- greške aritmetike računala.

Greške modela

Greške modela mogu nastati:

- zbog zanemarivanja utjecaja nekih sila,
  - na primjer, zanemarivanje utjecaja otpora zraka ili trenja (v. primjer),
- zbog zamjene kompliciranog modela jednostavnijim,
  - na primjer, sustavi nelinearnih običnih ili parcijalnih diferencijalnih jednadžbi se lineariziraju, da bi se dobilo barem približno rješenje,
- zbog upotrebe modela u graničnim slučajevima,
  - na primjer, kod matematičkog njihala se $\sin x$ aproksimira s x, što vrijedi samo za male kutove.

Modelni primjer — Problem gađanja

Primjer. Imamo top (ili haubicu) u nekoj točki — recimo, ishodištu.

- Treba pogoditi cilj koji se nalazi u nekoj drugoj točki.

Najjednostavniji model za ovaj problem je poznati kosi hitac. Projektil ispaljujemo prema cilju,

- nekom početnom brzinom v_0 (vektor),
- pod nekim kutem α, obzirom na horizontalnu ravninu.

Slikica!

Cijela stvar se odvija pod utjecajem gravitacije (prema dolje). Ako zanemarimo otpor zraka, dobijemo "obični" kosi hitac.

Modelni primjer — Jednadžba

Osnovna jednadžba je

$$F = ma,$$

gdje je m masa projektila (neće nam trebati na početku), a

- a je akceleracija — vektor u okomitoj (x, y)-ravnini,
- F je sila gravitacije, prema dolje, tj. $F_x = 0$ i $F_y = -mg$.

Gornja jednadžba je diferencijalna jednadžba drugog reda u vremenu. Ako je $(x(t), y(t))$ položaj projektila u danom trenutku, jednadžba ima oblik po komponentama:

$$m \, \frac{d^2x}{dt^2} = F_x, \quad m \, \frac{d^2y}{dt^2} = F_y.$$

Akceleracija je druga derivacija položaja.

Modelni primjer — Rješenje jednadžbe

Neka je projektil ispaljen u trenutku $t_0 = 0$.

Nakon integracije, za brzinu $v =$ prva derivacija položaja, imamo jednadžbu

$$mv = F \cdot t + mv_0,$$

ili, po komponentama (masa se skrati)

$$v_x = \frac{dx}{dt} = v_0 \cos\alpha, \qquad v_y = \frac{dy}{dt} = v_0 \sin\alpha - gt.$$

Još jednom integriramo (početni položaj je $x_0 = 0$, $y_0 = 0$). Za položaj projektila u trenutku t dobivamo:

$$x(t) = v_0 t \cos\alpha, \qquad y(t) = v_0 t \sin\alpha - \frac{1}{2} g t^2.$$

Reklo bi se — znamo sve!

Modelni primjer — Još neke relacije

Jednadžba "putanje" projektila u (x, y)-ravnini je

$$y = x \operatorname{tg} \alpha - \frac{g}{2v_0^2 \cos^2 \alpha} \, x^2.$$

To je parabola, s otvorom nadolje, koja prolazi kroz ishodište.

Najveća visina projektila je

$$y_{\max} = \frac{(v_0 \sin \alpha)^2}{2g},$$

a maksimalni domet na horizontalnoj x-osi je

$$x_{\max} = \frac{v_0^2 \sin 2\alpha}{g}.$$

Modelni primjer — Stvarnost

Nažalost, s ovim modelom nećemo ništa pogoditi.

- Fali otpor zraka, tlak pada s visinom, vjetrovi i sl.

Praksa:

- Koeficijent za otpor ovisi o obliku projektilu — mjeri se.
- Izračunate tablice se eksperimentalno “upucavaju” i korigiraju.
- Primjena u praksi ide obratno — znam daljinu, tražim kut.

Greške modela (nastavak)

Primjer. Među prvim primjenama jednog od prvih brzih paralelnih računala na svijetu (ASCI Blue Pacific) bilo je

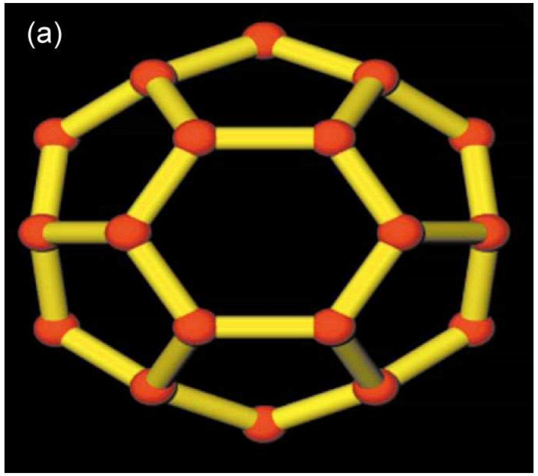
- određivanje trodimenzionalne strukture i elektronskog stanja ugljik-36 fulerena.

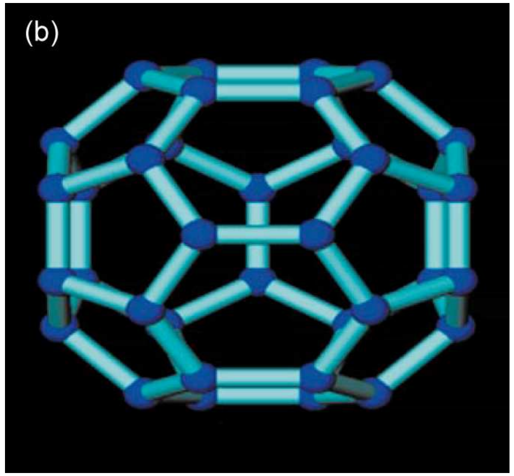
Primjena spoja je višestruka:

- supravodljivost na visokim temperaturama,
- precizno doziranje lijekova u stanice raka.

Greške modela (nastavak)

Prijašnja istraživanja kvantnih kemičara dala su dvije moguće strukture tog spoja.





Te dvije strukture imaju različita kemijska svojstva.

Greške modela (nastavak)

Stanje stvari:

- eksperimentalna mjerenja pokazivala su da je struktura (a) stabilnija,
- teoretičari su tvrdili da je stabilnija struktura (b).

Prijašnja računanja,

- zbog pojednostavljivanja i interpolacije,

kao odgovor davala su prednost “teoretskoj” strukturi.

Definitivan odgovor,

- proveden računanjem bez pojednostavljivanja,

pokazao je da je struktura (a) stabilnija.

Greške u ulaznim podacima

Greške u ulaznim podacima javljaju se zbog

- nemogućnosti ili besmislenosti točnog mjerenja (Heisenbergove relacije neodređenosti).
- Primjer, tjelesna temperatura se obično mjeri na desetinku stupnja Celziusa točno. Pacijent je podjednako loše ako ima tjelesnu temperaturu 39.5° ili 39.513462°.

Bitno praktično pitanje:

- Mogu li male greške u ulaznim podacima bitno povećati grešku rezultata?

Nažalost MOGU!

- Takvi problemi zovu se loše uvjetovani problemi.

Greške u ulaznim podacima (nastavak)

Primjer.
Zadana su dva sustava linearnih jednadžbi — recimo, umjesto ispravnih (prvih) koeficijanata, izmjerili smo druge:

$$\begin{aligned} 2x &+ 6y &= 8 \\ 2x &+ 6.0001y &= 8.0001, \end{aligned}$$

i

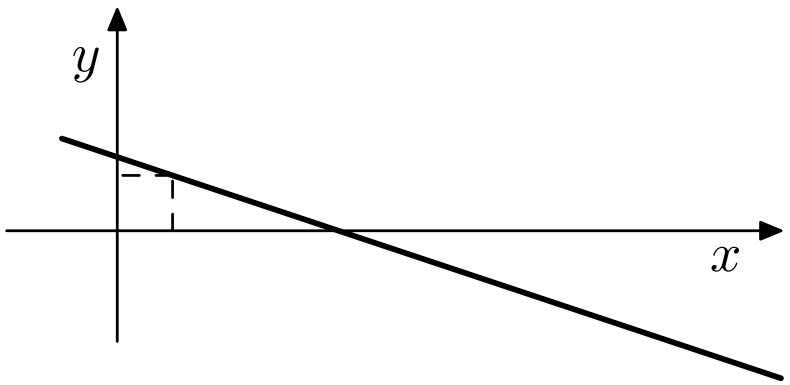
$$\begin{aligned} 2x &+ 6y &= 8 \\ 2x &+ 5.99999y &= 8.00002. \end{aligned}$$

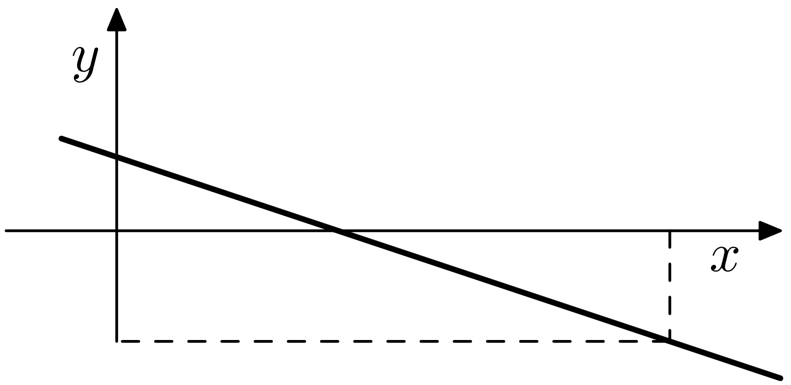
Perturbacije koeficijenata: reda veličine 10^{-4}. Je li se rezultat također promijenio za red veličine 10^{-4}?

Greške u ulaznim podacima (nastavak)

- Rješenje prvog problema: $x = 1$, $y = 1$.
- Rješenje drugog problema: $x = 10$, $y = -2$.

Grafovi presjecišta dva pravca za prvi i drugi sustav:





Greške metoda za rješavanje problema

Najčešće nastaju kad se nešto beskonačno zamjenjuje nečim konačnim. Razlikujemo dvije kategorije:

- greške diskretizacije koje nastaju zamjenom kontinuuma konačnim diskretnim skupom točaka, ili "beskonačno" malu veličinu h ili $\varepsilon \to 0$ zamijenjujemo nekim "konačno" malim brojem;
- greške odbacivanja koje nastaju "rezanjem" beskonačnog niza ili reda na konačni niz ili sumu, tj. odbacujemo ostatak niza ili reda.

Greške metoda za rješavanje problema (nast.)

Tipični primjeri greške diskretizacije:

- aproksimacija funkcije f na $[a, b]$, vrijednostima te funkcije na konačnom skupu točaka (tzv. mreži) $\{x_1, \ldots, x_n\} \subset [a, b]$,
- aproksimacija derivacije funkcije f u nekoj točki x. Po definiciji je

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h},$$

a za približnu vrijednost uzmemo dovoljno mali $h \neq 0$ i

$$f'(x) \approx \frac{\Delta f}{\Delta x} = \frac{f(x+h) - f(x)}{h}.$$

Greške metoda za rješavanje problema (nast.)

Tipični primjeri greške odbacivanja:

- zaustavljanje iterativnih procesa nakon dovoljno velikog broja n iteracija (recimo kod računanja nultočaka funkcije);
- zamjena beskonačne sume konačnom kad je greška dovoljno mala (recimo kod sumiranja Taylorovih redova).

Taylorov red, Taylorov polinom, ...

Za dovoljno glatku funkciju f, Taylorov red oko točke x_0

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k$$

možemo aproksimirati Taylorovim polinomom p

$$f(x) = p(x) + R_{n+1}(x), \qquad p(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k,$$

pri čemu je $R_{n+1}(x) = \dfrac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_0)^{n+1}$ greška odbacivanja, a ξ neki broj između x_0 i x. $R_{n+1}(x)$ obično ocjenjujemo po apsolutnoj vrijednosti.

Taylorov red, Taylorov polinom, ... (nastavak)

Primjer.

- Funkcije e^x i $\sin x$ imaju Taylorove redove oko točke 0 koji konvergiraju za proizvoljan $x \in \mathbb{R}$.
- Zbrajanjem dovoljno mnogo članova tih redova, možemo, barem u principu, dobro aproksimirati vrijednosti funkcija e^x i $\sin x$.
- Traženi Taylorovi polinomi s istim brojem članova (ali ne istog stupnja) su

$$e^x \approx \sum_{k=0}^{n} \frac{x^k}{k!}, \qquad \sin x \approx \sum_{k=0}^{n} \frac{(-1)^k x^{2k+1}}{(2k+1)!}.$$

Taylorov red, Taylorov polinom, ... (nastavak)

Za grešku odbacivanja trebaju nam derivacije:

$$(e^x)^{(n)} = e^x, \qquad (\sin x)^{(n)} = \sin\left(x + \frac{n\pi}{2}\right),$$

pa su pripadne greške odbacivanja

$$R_{n+1}(x) = \frac{e^{\xi} x^{n+1}}{(n+1)!}, \qquad R_{2n+3}(x) = \frac{\sin(\xi + \frac{2n+3}{2}\pi) x^{2n+3}}{(2n+3)!},$$

Pretpostavimo sada da je $x > 0$. Iz $\xi \leq x$ dobivamo

$$|R_{n+1}(x)| \leq \frac{e^x x^{n+1}}{(n+1)!}, \qquad |R_{2n+3}(x)| \leq \frac{x^{2n+3}}{(2n+3)!}.$$

Taylorov red, Taylorov polinom, ... (nastavak)

Zbrojimo li članove reda sve dok apsolutna vrijednost prvog odbačenog člana ne padne ispod zadane točnosti $\varepsilon > 0$, napravili smo grešku odbacivanja manju ili jednaku

$$\begin{cases} e^x \varepsilon, & \text{za } e^x, \\ \varepsilon, & \text{za } \sin x. \end{cases}$$

U prvom slučaju očekujemo

- malu relativnu grešku,

a u drugom slučaju očekujemo

- malu apsolutnu grešku.

Provjerimo to eksperimentalno — u aritmetici računala!

Greške (nastavak)

Izvori i vrste grešaka (ponavljanje)

Pri numeričkom rješavanju nekog problema javljaju se četiri vrste grešaka:

- greške modela,
- greške u ulaznim podacima (mjerenjima),
- greške metoda za rješavanje modela,
- greške aritmetike računala.

Za početak, pogledajmo detaljnije zadnju vrstu grešaka koja nastaje zbog "približnog" računanja. To su greške zaokruživanja u

- prikazu brojeva u računalu i
- aritmetici računala.

Prikaz brojeva u računalu

Tipovi brojeva u računalu

U računalu postoje dva bitno različita tipa brojeva:

- cijeli brojevi
- realni brojevi.

Oba skupa su konačni podskupovi odgovarajućih skupova $\mathbb{Z}$ i $\mathbb{R}$ u matematici.

Kao baza za prikaz oba tipa koristi se baza 2.

Cijeli brojevi

Cijeli se brojevi prikazaju korištenjem n bitova — binarnih znamenki, od kojih jedna služi za predznak, a ostalih $n-1$ za znamenke broja.

Matematički gledano,

- aritmetika cijelih brojeva u računalu je modularna aritmetika u prstenu ostataka modulo 2^n, samo je sustav ostataka simetričan oko 0, tj.

$$-2^{n-1}, \ldots, -1, 0, 1, \ldots, 2^{n-1}-1.$$

- Računalo ne zna izravno operirati s brojevima izvan tog raspona.

Realni brojevi

Realni brojevi r prikazuju se korištenjem mantise m (ili češće, signifikanda) i eksponenta e u obliku

$$r = \pm m \cdot 2^e,$$

pri čemu je e cijeli broj u određenom rasponu, a m racionalni broj za koji vrijedi $1 \leq m < 2$ (tj. mantisa započinje s $1. \ldots$).

- Vodeća jedinica se često ne pamti, pa je mantisa "dulja" za 1 bit tzv. "skriveni bit" (engl. hidden bit).
- Eksponent se prikazuje kao s-bitni cijeli broj, a za mantisu pamti se prvih t znamenki iza binarne točke.
- Po standardu, eksponentu se dodaje "pomak" (engl. bias), da bi eksponent bio nenegativan. Ovo je nebitno za ponašanje aritmetike.

Realni brojevi

Skup svih realnih brojeva prikazivih u računalu je omeđen, a parametriziramo ga duljinom mantise i eksponenta i označavamo s $\mathbb{R}(t, s)$.

mantisa						eksponent				
$\pm$	m_{-1}	m_{-2}	$\cdots$	m_{-t}		e_{s-1}	e_{s-2}	$\cdots$	e_1	e_0

Ne može se svaki realni broj egzaktno spremiti u računalo.

Ako je broj $x \in \mathbb{R}$ unutar prikazivog raspona i

$$x = \pm \left(\sum_{k=1}^{\infty} b_{-k} 2^{-k} \right) 2^e$$

i mantisa broja ima više od t znamenki, ...

Realni brojevi

... bit će spremljena aproksimacija tog broja $f\ell(x) \in \mathbb{R}(t, s)$ koja se može prikazati kao

$$f\ell(x) = \pm \left(\sum_{k=1}^{t} b_{-k}^{*} 2^{-k} \right) 2^{e^*}.$$

Slično kao kod decimalne aritmetike

- ako je prva odbačena znamenka 1, broj zaokružujemo nagore,
- a ako je 0, nadolje.

Time smo napravili apsolutnu grešku manju ili jednaku od "pola zadnjeg prikazivog bita", tj. 2^{-t-1+e}.

Relativna greška zaokruživanja

Gledajući relativno, greška je manja ili jednaka

$$\left|\frac{x - f\ell(x)}{x}\right| \leq \frac{2^{-t-1+e}}{2^{-1} \cdot 2^{e}} = 2^{-t},$$

tj. imamo vrlo malu relativnu grešku.

Veličinu 2^{-t} zovemo jedinična greška zaokruživanja (engl. unit roundoff) i uobičajeno označavamo s u.

Za $x \in \mathbb{R}$ unutar prikazivog raspona, umjesto x sprema se zaokruženi broj $f\ell(x) \in \mathbb{R}(t, s)$ i vrijedi

$$f\ell(x) = (1 + \varepsilon)x, \quad |\varepsilon| \leq u,$$

gdje je ε relativna greška napravljena tim zaokruživanjem.

IEEE standard za prikaz brojeva

Prikaz realnih brojeva u računalu zove se prikaz s pomičnim zarezom/točkom (engl. floating point representation), a aritmetika je aritmetika pomičnog zareza/točke (engl. floating point arithmetic).

Veličine s i t prema novom IEEE standardu:

`format`	`32-bitni`	`64-bitni`	`128-bitni`
duljina mantise	23 bita	52 bita	112 bita
duljina eksponenta	8 bitova	11 bitova	15 bitova
jedinična gr. zaokr.	2^{-24}	2^{-53}	2^{-113}
$u \approx$	$5.96 \cdot 10^{-8}$	$1.11 \cdot 10^{-16}$	$9.63 \cdot 10^{-35}$
raspon brojeva $\approx$	$10^{\pm 38}$	$10^{\pm 308}$	$10^{\pm 4932}$

IEEE standard za prikaz brojeva (nastavak)

Većina PC računala (procesora) još ne podržava 128-bitni prikaz i aritmetiku.

Umjesto toga, FPU (Floating–point unit) stvarno koristi

- tzv. tip `extended` iz starog IEEE standarda.

Dio primjera koje ćete vidjeti napravljen je baš u tom tipu!

`format`	`80-bitni`
duljina mantise	64 bita
duljina eksponenta	15 bitova
jedinična gr. zaokr.	2^{-64}
$u \approx$	$5.42 \cdot 10^{-20}$
raspon brojeva $\approx$	$10^{\pm 4932}$

IEEE standard za aritmetiku računala

IEEE standard propisuje i svojstva aritmetike.
Pretpostavka standarda — za osnovne aritmetičke operacije ($\circ$ označava $+$, $-$, $*$, $/$) nad $x, y \in \mathbb{R}(t, s)$ vrijedi

$$f\ell(x \circ y) = (1 + \varepsilon)\,(x \circ y), \quad |\varepsilon| \leq u,$$

za sve $x, y \in \mathbb{R}(t, s)$ za koje je $x \circ y$ u dozvoljenom rasponu.
Dobiveni rezultat je tada prikaziv, tj. vrijedi $f\ell(x \circ y) \in \mathbb{R}(t, s)$.
Postoje rezervirani eksponenti koji označavaju "posebno stanje":

- overflow,
- underflow,
- dijeljenje s 0,
- nedozvoljenu operaciju kao što su $0/0$, $\sqrt{-1}$.