Numerička analiza

M. Klaričić Bakula

Listopad, 2008.

1 Stabilnost numeričkog računanja

Sada ćemo se pozabaviti stabilnošću numeričkih algoritama, a s čime je usko povezana i pouzdanost dobivenih rješenja. Kroz primjere ćemo se upoznati s nekim nepoželjnim fenomenima koji se mogu pojaviti prilikom korištenja aritmetike računala.

1.1 Greške unazad i unaprijed

Neka je f realna funkcija jedne varijable. Pretpostavimo da je u aritmetici preciznosti u izračunata vrijednost $y = f(x)$ i da ona iznosi $\hat{y}$. Kako možemo mjeriti kvalitetu dobivenog $\hat{y}$ kao aproksimacije točnog y?

U većini slučajeva bit ćemo sretni ako postignemo neku malu relativnu grešku, no to neće uvijek biti moguće. Umjesto toga možemo se zapitati: Za koji stvari skup podataka smo zaista riješili problem? Dakle, za koji $\triangle x$ vrijedi

$$\hat{y} = f\left(x + \triangle x\right) \quad ?$$

Općenito može biti i više takvih $\triangle x$, a nas će zanimati najveći od njih, $\max|\triangle x|$ (ponekad se uzima i podijeljeno s $|x|$). Zove se **greška unatrag** ili **povratna greška**. S druge strane, apsolutna i relativna greška po funkcijskoj vrijednosti zovu se **greške unaprijed** ili jednostavno **greške**. Proces omeđivanja povratne greške izračunatog rješenja zove se **analiza povratne greške**, a motivacija za njega je dvostruka.

- **Omogućava interpretaciju grešaka zaokruživanja kao grešaka u ulaznim podacima**.
  Podaci često kriju netočnosti nastale usljed prethodnih računanja, nepreciznih rezultata mjerenja i spremanja podataka u računalo. Ako povratna greška nije veća od polaznih netočnosti, onda je dobiveno rješenje točno "do na ulaznu netočnost".

- **Reducira problem omeđivanja greške unaprijed na primjenu teorije perturbacije za dani problem**.
  Naime, teorija perturbacije je dobro poznata za večinu problema i važno je da ovisi o samom problemu, a ne o metodi koju koristimo. Kada imamo ocjenu greške unatrag rješenja promatrane metode, onda primjenom opće teorije perturbacije za dani problem lako dođemo do ocjene greške unaprijed.

Metoda za računanje vrijednosti $y = f(x)$ je **stabilna unazad** ili **povratno stabilna** ako za svaki x producira izračunati $\hat{y}$ s malom povratnom greškom, tj. ako vrijedi

$$\hat{y} = f(x + \triangle x) \tag{1}$$

za neki mali $\triangle x$. Pri tom značenje izraza "mala" povratna greška ovisi o kontekstu.
U načelu, za dani problem može postojati više metoda rješavanja od kojih će neke biti povratno stabilne, a neke neće.
Npr. sve osnovne računske operacije u računalu zadovoljavaju relaciju (1), pa daju rezultat koji je točan za malo pomaknute ulazne podatke:

$$x \to x(1+\delta) \ \text{ i } \ y \to y(1+\delta), \quad |\delta| \le u.$$

Međutim, većina metoda za računanje funkcije $\cos$ ne zadovoljava relaciju (1), već nešto slabiju

$$\hat{y} + \triangle y = \cos(x + \triangle x)$$

za neke male $\triangle x$ i $\triangle y$.

Općenito, greška u rezultatu koji se zapisuje kao

$$\hat{y} + \triangle y = f\,(x + \triangle x)\,, \qquad |\triangle x| \leq \xi\,|x|\,, \quad |\triangle y| \leq \eta\,|y| \tag{2}$$

naziva se **miješana naprijed-nazad greška**.

Može se reći ovako:
Izračunato rješenje $\hat{y}$ jedva se razlikuje od vrijednosti $\hat{y} + \triangle y$ koja se dobije egzaktnim računom na ulaznoj vrijednosti $x + \triangle x$ koja se jedva razlikuje od točne ulazne vrijednosti x.

Algoritam je **numerički stabilan** ako je stabilan u smislu relacije (2) s malim ξ i η.
Ova definicija se uglavnom odnosi na izračunavanja u kojima su greške zaokruživanja osnovnih aritmetičkih operacija dominantni oblici grešaka. U drugim područjima numeričke analize ovaj pojam ima različita značenja.

1.2 Uvjetovanost

Odnos između greške unazad i greške unaprijed za neki dani problem u velikoj je mjeri određen **uvjetovanošću problema**, tj. osjetljivošću rješenja problema na ulazne podatke.

Pretpostavimo da je dano približno rješenje $\hat{y}$ problema $y = f(x)$ koje zadovoljava relaciju

$$\hat{y} = f(x + \triangle x) .$$

Ako pretpostavimo da je funkcija f dvaput neprekidno derivabilna, onda razvoj u Taylorov red daje

$$\hat{y} - y = f(x + \triangle x) - f(x) = f'(x) \triangle x + \frac{f''(x + \Theta \triangle x)}{2!} (\triangle x)^2 , \quad \Theta \in (0, 1) ,$$

i možemo ocijeniti desnu stranu ove jednakosti.

Zbog

$$\frac{\hat{y}-y}{y}=\frac{f'(x)}{f(x)}\triangle x+\frac{f''(x+\Theta\triangle x)}{2f(x)}(\triangle x)^2$$

imamo

$$\frac{\hat{y}-y}{y}=\frac{xf'(x)}{f(x)}\frac{\triangle x}{x}+\mathcal{O}(\triangle x)^2,$$

pa veličina

$$\kappa(f)(x)=\left|\frac{xf'(x)}{f(x)}\right| \tag{3}$$

mjeri relativnu promjenu y za malu relativnu promjenu x.

Zato κ zovemo **uvjetovanost** funkcije f. Ako je f funkcija više varijabla, onda se u izrazu (3) umjesto apsolutne vrijednosti javlja norma.
Uvjetovanost služi za mjerenje najveće relativne promjene koja se dostiže za neku vrijednost broja x ili vektora $\boldsymbol{x}$.
Na primjer, ako je f logaritamska funkcija, onda je

$$\kappa(f)(x) = \frac{1}{|\ln(x)|},$$

pa je uvjetovanost jako velika za $x \approx 1$.

Kada se greške unatrag i unaprijed te uvjetovanost za neki problem definiraju na konzistentan način, vrijedi jednostavno pravilo:

$$\text{greška unaprijed} \lessapprox \text{uvjetovanost} \times \text{greška unazad}.$$

Dakle, izračunato rješenje loše uvjetovanog problema može imati veliku grešku unaprijed. Zato se uvodi sljedeća definicija.

DEFINICIJA. *Ako metoda daje rješenja s greškama unaprijed koja su sličnog reda veličine kao ona koja se dobiju primjenom povratno stabilne metode, onda se za metodu kaže da je stabilna unaprijed*.

Dakle, sama metoda ne treba biti povratno stabilna da bi bila stabilna unaprijed. Povratna stabilnost implicira stabilnost unaprijed, dok *obrat ne vrijedi*.

1.3 Akumulacija grešaka zaokruživanja

Rasprostranjeno je mišljenje da velike brzine modernih računala koja u svakoj sekundi izvršavaju i nekoliko milijardi računskih operacija imaju za posljedicu potencijalno velike greške u rezultatu. Na sreću, ta tvrdnja uglavnom nije istinita, a u rijetkim slučajevima kada dolazi do većih grešaka u rezultatu, kriva je jedna ili tek nekoliko podmuklih grešaka zaokruživanja.

PRIMJER. Ako se a novčanih jedinica investira na godinu dana po godišnjoj kamatnoj stopi x, uz n ukamačivanja godišnje, onda je vrijednost uloženog novca nakon jedne godine dana formulom

$$C_n(x,a) = aC_n(x), \quad C_n(x) = \left(1 + \frac{x}{n}\right)^n.$$

Ovo je formula tzv. složenog ukamaćivanja. Poznato je da ako broj ukamaćivanja n raste, onda i $C_n(x)$ monotono raste prema vrijednosti e^x. Kada $n \to \infty$ govorimo o neprekidnom ukamaćivanju, a ovakav se slučaju više sreće u biološkim i medicinskim područjima nego u bankarstvu.

Najprije ćemo odrediti uvjetovanost za funkciju C_n. Imamo

$$\kappa\left(C_n\right)(x) = \left|\frac{xC_n'(x)}{C_n(x)}\right| = \left|\frac{x}{1+\frac{x}{n}}\right| = \frac{|x|}{\left|1+\frac{x}{n}\right|},$$

pa uvjetovanost konvergira prema $|x|$ kada n raste (naravno, isto je i za uvjetovanost eksponencijalne funkcije).
Prilikom izvršavanja računskih operacija potrebnih za izračunavanje vrijednosti $C_n(x)$ u jednostrukoj preciznosti greška će se pojaviti već kod izračunavanja vrijednosti $1+x/n$, a potenciranjem na visoku potenciju n će se i dalje uvećavati. Naime, funkcija potenciranja

$$\mathsf{pot}_n(x) = x^n$$

nije dobro uvjetovana jer vrijedi

$$\kappa\left(\mathsf{pot}_n\right)(x) = \frac{\left|x \cdot nx^{n-1}\right|}{\left|x^n\right|} = n.$$

1.4 Kraćenje

Kraćenje nastaje kada se oduzimaju dva bliska broja. To najčešće, iako ne uvijek, ima za posljedicu netočan rezultat.

PRIMJER. Promotrimo funkciju f zadanu s

$$f(x) = \frac{1 - \cos(x)}{x^2}.$$

Za $x = 1.2 \times 10^{-5}$ vrijednost $\cos(x)$ zaokružena na 10 značajnih znamenki iznosi 0.9999999999, tako da je

$$1 - \cos(x) = 0.0000000001.$$

Dakle, aproksimacija za $f\left(1.2 \times 10^{-5}\right)$ je

$$\frac{10^{-10}}{1.44 \times 10^{-10}} \approx 0.6944,$$

što je očito jako loše jer je

$$(\forall x \neq 0) \quad 0 \leq f(x) \leq 0.5.$$

Iz ovoga se vidi da ni desetoroznamenkasta aproksimacija vrijednosti $\cos(x)$ nije dovoljno točna da bi izračunata vrijednost $f(x)$ imala barem jednu točnu značajnu znamenku.

Problem je u tomu što (iako je oduzimanje egzaktno) $1-\cos(x)$ ima samo jednu značajnu znamenku, pa je rezultat iste veličine kao i greška u $\cos(x)$. Drugim riječima, **oduzimanje podiže značaj prethodne greške**!

U ovom primjeru se f može napisati tako da se izbjegne kraćenje. Stavimo li

$$f(x) = \frac{1}{2}\left(\frac{\sin(x/2)}{x/2}\right)^2$$

izračunavanje će za isti $x = 1.2 \times 10^{-5}$ s desetoroznamnkastom aproksimacijom za $\sin\left(1.2 \times 10^{-5}\right)$ dati vrijednost 0.5 koja je točan rezultat na deset značajnih znamenki.

Da bismo dobili dublji uvid u fenomen kraćenja pogledajmo oduzimanje

$$\hat{x} = \hat{a} - \hat{b}$$

u egzaktnoj aritmetici, gdje su

$$\hat{a} = a\,(1 + \triangle_a)\,, \quad \hat{b} = b\,(1 + \triangle_b)\,.$$

Članovi $\triangle_a$ i $\triangle_b$ su relativne greške koje unosimo u račun. Izraz za relativnu grešku rezultata oduzimanja daje

$$\left|\frac{x - \hat{x}}{x}\right| = \left|\frac{-a\triangle_a + b\triangle_b}{a - b}\right| \le \max\left\{|\triangle_a|\,, |\triangle_b|\right\} \frac{|a| + |b|}{|a - b|}.$$

Očito je ograda za relativnu grešku visoka ako je

$$|a - b| \ll |a| + |b|\,,$$

a to je istina kada postoji bitno kraćenje u oduzimanju. Ova analiza pokazuje da se zbog kraćenja postojeće greške ili netočnosti u $\hat{a}$ i $\hat{b}$ povećavaju. Drugim riječima, **kraćenje dovodi prethodne greške na vidjelo**.

Ipak, postoje situacije kada kraćenje neče dovesti do loših pojava, a to su npr. sljedeće:

- ulazni podaci su točni,
- uvjetovanost je loša, pa je kraćenje nužnost,
- utjecaj kraćenja na daljnji račun nije loš (npr. kod $x + (y - z)$ ako je $x \gg y \approx z > 0$),
- kraćenje grešaka zaokruživanja.

1.5 Kako dizajnirati stabilne algoritme

Najprije naglasimo da je numerička stabilnost važnija od drugih karakteristika algoritma, kao što su npr. broj računskih operacija, paralelizacija, ušteda memorije i slično. Evo nekih općih uputa:

- izbjegavati oduzimanje bliskih brojeva koji nose greške
- minimizirati veličinu međurezultata u odnosu na konačni rezultat
- iskušavati razne formulacije istog problema
- koristiti jednostavne formule za ažuriranje tipa

  $$\text{nova vrijednost} = \text{stara vrijednost} + \text{mala korekcija}$$

  ako se korekcija može izračunati na dovoljan broj značajnih znamenki
- koristiti samo dobro uvjetovane transformacije
- poduzimati mjere opreza protiv prekoračenja i potkoračenja
- koristiti što manje cijepanje formula u više programskih linija uvođenjem pomoćnih varijabla jer CPU često koristi precizniju aritmetiku za operande u registrima, dok zaokruživanje nastupa tek prilikom spremanja u memoriju.

2 Osnove analize grešaka zaokruživanja

U ovom dijelu ćemo se pozabaviti osnovnim alatom za analiziranje stabilnosti numeričkih algoritama. Analiza grešaka zaokruživanja je, zajedno s perturbacijskom analizom problema koji se rješava, moćan alat za analiziranje, a samim time i za dizajniranje numeričkih algoritama.

Označimo s P skup negativnih potencija broja 2 koje ulaze u definiciju preciznosti IEEE standarda:

$$P = \left\{2^{-23}, 2^{-24}, 2^{-52}, 2^{-53}, 2^{-63}\right\}.$$

Ako je n prirodni broj koji igra ulogu dimenzije vektora ili matrice, stupnja polinoma, broja sumanada ili faktora i sl., onda ćemo pretpostaviti da vrijedi

$$u \in P \longrightarrow nu \leq 2^{-6}.$$

Posljedica ovoga je da ako radimo u jednostrukoj preciznosti i koristimo bilo koji način zaokruživanja, maksimalna vrijednost broja n bit će (zbog 23-6=17)

$$2^{17} = 131072,$$

a u standardnom načinu zaokruživanja prema najbližem n će biti najviše

$$2^{18} = 262144.$$

Kod dvostruke preciznosti, a pri standardnom načinu zaokruživanja prema najbližem, dobijemo (zbog 53-6=47)

$$n \leq 2^{47} \approx 1.40737488355328 \times 10^{14},$$

i tako dalje za ostale elemente skupa P.

U računanju će nam koristiti sljedeća tehnička lema.

LEMA. Neka je $u \in P$ i n takav da vrijedi $nu \leq 2^{-6}$. Ako je $|\varepsilon| \leq u$, onda vrijedi:

1. $(1+\varepsilon)^2 = 1 + \varepsilon_2, \quad |\varepsilon_2| \leq 2.00000012u,$
2. $(1+\varepsilon)^3 = 1 + \varepsilon_3, \quad |\varepsilon_3| \leq 3.00000036u,$
3. $(1+\varepsilon)^{-1} = 1 + \varepsilon_1', \quad |\varepsilon_1'| \leq 1.00000012u,$
4. $(1+\varepsilon)^{-2} = 1 + \varepsilon_2', \quad |\varepsilon_2'| \leq 2.00000036u,$
5. $(1+\varepsilon)^n = 1 + \varepsilon_n, \quad |\varepsilon_n| \leq 1.008nu,$
6. $(1+\varepsilon)^{-n} = 1 + \varepsilon_n', \quad |\varepsilon_n'| \leq 1.008nu,$
7. $(1+\varepsilon)^{\frac{1}{2}} = 1 + \varepsilon_{\sqrt{}}, \quad \left|\varepsilon_{\sqrt{}}\right| \leq 0.500000015u.$

DOKAZ. Samo kao primjer navodimo dokaz tvrdnje 2. Zbog $|\varepsilon| \leq u$, $u \leq 2^{-23}$ i

$$(1+\varepsilon)^3 = 1 + \varepsilon\left(3 + 3\varepsilon + \varepsilon^2\right)$$

imamo

$$\varepsilon_3 = \varepsilon\left(3 + 3\varepsilon + \varepsilon^2\right)$$

i

$$\begin{aligned} |\varepsilon_3| &\leq \left|\varepsilon\left(3 + 3\varepsilon + \varepsilon^2\right)\right| = |\varepsilon|\left|3 + 3\varepsilon + \varepsilon^2\right| \\ &\leq \left|3 + 3 \cdot 2^{-23} + 2^{-46}\right| u \\ &\leq 3.00000036u \end{aligned}$$

■

ZADATAK. Neka su x, y realni brojevi za koje vrijedi

$$x = fl\,(x)\,, \quad y = fl\,(y)\,.$$

S kolikom će relativnom greškom računalo koje koristi IEEE standard izračunati

$$z = \sqrt{x^2 + y^2} \qquad ?$$

RJEŠENJE.
Početne pretpostavke trebaju biti

$$\begin{aligned} fl\left(x^2\right) + fl\left(y^2\right) \ &< \ N_{\max}, \\ \min\left\{|x|\,, |y|\right\} \ &\geq \ \sqrt{N_{\min}}. \end{aligned}$$

Tada možemo pisati:

$$\begin{aligned} x_2 \ &= \ x \otimes x = fl\left(x^2\right) = x^2\,(1 + \varepsilon_1)\,, \quad |\varepsilon_1| \leq u \\ y_2 \ &= \ y \otimes y = fl\left(y^2\right) = y^2\,(1 + \varepsilon_2)\,, \quad |\varepsilon_2| \leq u. \end{aligned}$$

Također, umjesto

$$fl\left(fl\left(x^2\right) + fl\left(y^2\right)\right) = fl\left(x^2\right) \oplus fl\left(y^2\right)$$

kraće pišemo

$$fl\left(x^2 + y^2\right).$$

Vrijedi

$$\begin{aligned} z_2 &= fl\left(x^2 + y^2\right) = \left(x^2 + y^2\right)(1 + \varepsilon_3), \quad |\varepsilon_3| \leq u \\ z &= fl\left(\sqrt{z_2}\right) = \sqrt{z_2}\,(1 + \varepsilon_4), \quad |\varepsilon_4| \leq u. \end{aligned}$$

Povezivanjem svih jednadžbi dobijemo

$$
\begin{aligned}
z &= \sqrt{x^2+y^2}\,(1+\varepsilon_4)\,\sqrt{1+\varepsilon_3}\sqrt{1+\frac{\varepsilon_1 x^2+\varepsilon_2 y^2}{x^2+y^2}} \\
&= \sqrt{x^2+y^2}\,(1+\varepsilon_4)\,\sqrt{1+\varepsilon_3}\sqrt{1+\varepsilon_5} \\
&= \sqrt{x^2+y^2}\,(1+\varepsilon_z)\,,
\end{aligned}
$$

pri čemu je

$$
\varepsilon_5 = \frac{\varepsilon_1 x^2+\varepsilon_2 y^2}{x^2+y^2}, \quad 1+\varepsilon_z = (1+\varepsilon_4)\,\sqrt{(1+\varepsilon_3)\,(1+\varepsilon_5)}.
$$

Sada redom ocjenimo greške, i to najprije ε_5. Kako je (npr. za $\varepsilon_1 \le \varepsilon_2$)

$$\frac{\varepsilon_1 x^2 + \varepsilon_2 y^2}{x^2 + y^2} = \frac{x^2}{x^2 + y^2}\varepsilon_1 + \frac{y^2}{x^2 + y^2}\varepsilon_2 \in [\varepsilon_1, \varepsilon_2],$$

to je

$$|\varepsilon_5| \le \max\{|\varepsilon_1|, |\varepsilon_2|\} \le u.$$

Sada imamo,

$$\begin{aligned} |\varepsilon_z| &= \left|(1+\varepsilon_4)\sqrt{(1+\varepsilon_3)(1+\varepsilon_5)} - 1\right| \\ &\le \left|(1+u)\sqrt{(1+u)(1+u)} - 1\right| \\ &= |(1+u)(1+u) - 1| \\ &= \left|u^2 + 2u\right| = u(u+2) \\ &\le 2.00000012u \end{aligned}$$

■

2.1 Propagiranje grešaka zaokruživanja

Promotrimo sada kako izvođenje neke računske operacije na računalu povećava postojeće greške u podacima. Neka su

$$\hat{x} = x\,(1+\varepsilon_x)\,, \quad \hat{y} = y\,(1+\varepsilon_y)$$

podaci spremljeni u računalo koji aproksimiraju točne podatke x i y s pripadnim relativnim greškama ε_x i ε_y.

Pogledajmo redom što se događa kod izvođenja osnovnih aritmetičkih operacija u računalu s aproksimacijama brojeva x i y.

2.1.1 Množenje

$$\begin{aligned} fl\,(\hat{x}\cdot\hat{y}) &= (\hat{x}\cdot\hat{y})\,(1+\varepsilon_\times) = xy\,(1+\varepsilon_x)\,(1+\varepsilon_y)\,(1+\varepsilon_\times) \\ &= xy\,(1+\varepsilon_x+\varepsilon_y+\varepsilon_x\varepsilon_y+\varepsilon_\times+\alpha)\,, \end{aligned}$$

pri čemu je $|\varepsilon_\times| \leq u$ i

$$|\alpha| = |(\varepsilon_x+\varepsilon_y+\varepsilon_x\varepsilon_y)\,\varepsilon_\times| \approx |(\varepsilon_x+\varepsilon_y)\,\varepsilon_\times| \leq u\,(|\varepsilon_x|+|\varepsilon_y|) \leq 2u^2.$$

Dakle, za dovoljno male ε_x i ε_y članove $\varepsilon_x\varepsilon_y$ i α možemo odbaciti, pa se može staviti

$$fl\,(\hat{x}\cdot\hat{y}) \approx xy\,(1+\varepsilon_x+\varepsilon_y+\varepsilon_\times)\,.$$

Treba biti oprezan tek ako su ε_x i ε_y približno istih modula i suprotnih predznaka jer tada $\varepsilon_x\varepsilon_y$ utječe na ukupnu grešku.

2.1.2 Dijeljenje

$$\begin{aligned} fl\left(\frac{\hat{x}}{\hat{y}}\right) &= \frac{\hat{x}}{\hat{y}}\left(1+\varepsilon_/\right) = \frac{x\,(1+\varepsilon_x)}{y\,(1+\varepsilon_y)}\left(1+\varepsilon_/\right) \\ &= \frac{x}{y}\left(1+\varepsilon_x-\varepsilon_y-\varepsilon_x\varepsilon_y+\frac{\varepsilon_y^2}{1+\varepsilon_y}+\varepsilon_/+\beta\right), \end{aligned}$$

pri čemu je $\left|\varepsilon_/\right| \leq u$ i

$$\beta = \left(\varepsilon_x-\varepsilon_y-\varepsilon_x\varepsilon_y+(1+\varepsilon_x)\,\frac{\varepsilon_y^2}{1+\varepsilon_y}\right)\varepsilon_/ + \frac{\varepsilon_x\varepsilon_y^2}{1+\varepsilon_y}$$

broj takve veličine da se može zanemariti.

I opet, ako ε_x i ε_y nisu približno jednaki možemo zanemariti članove višega reda, pa dobijemo

$$fl\left(\frac{\hat{x}}{\hat{y}}\right) = \frac{x}{y}\left(1 + \varepsilon_x - \varepsilon_y + \varepsilon_/\right).$$

Kao i kod množenja, nova greška dijeljenja $\varepsilon_/$ ima utjecaj tek na zadnju decimalu binarnog prikaza kvocijenta.

2.1.3 Zbrajanje

$$
\begin{aligned}
fl\left(\hat{x}+\hat{y}\right) &= \left(\hat{x}+\hat{y}\right)\left(1+\varepsilon_+\right) = \left[x\left(1+\varepsilon_x\right)+y\left(1+\varepsilon_y\right)\right]\left(1+\varepsilon_+\right) \\
&= \left(x+y\right)\left[1+\frac{x}{x+y}\left(\varepsilon_x+\varepsilon_++\varepsilon_x\varepsilon_+\right)+\frac{y}{x+y}\left(\varepsilon_y+\varepsilon_++\varepsilon_y\varepsilon_+\right)\right].
\end{aligned}
$$

Označimo

$$
fl\left(\hat{x}+\hat{y}\right) = \left(x+y\right)\left(1+\varepsilon_s\right),
$$

gdje je

$$
\varepsilon_s = \frac{x}{x+y}\left(\varepsilon_x+\varepsilon_++\varepsilon_x\varepsilon_+\right)+\frac{y}{x+y}\left(\varepsilon_y+\varepsilon_++\varepsilon_y\varepsilon_+\right)
$$

Analizirajući ovaj izraz u ovisnosti o predznacima brojeva x i y dobijemo:

• ako je sign(x) =sign(y) , onda je

$$|\varepsilon_s| \leq \max\{|\varepsilon_1|, |\varepsilon_2|\} + u,$$

• ako je sign$(x) = -$sign(y) , onda problem nastupa u izrazima

$$\frac{x}{x+y} \quad \text{i} \quad \frac{y}{x+y}.$$

Naime, ako je $|x| \approx |y|$, onda će se greške $\varepsilon_x + \varepsilon_+ + \varepsilon_x\varepsilon_+$ i $\varepsilon_y + \varepsilon_+ + \varepsilon_y\varepsilon_+$ množiti s potencijalno vrlo velikim brojevima zbog $x + y \approx 0$. To je već poznati fenomen opasnog kraćenja.