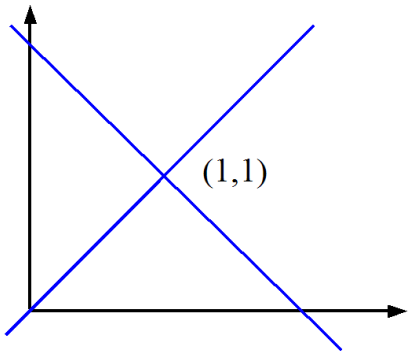
1.3 Občutljivost problema

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

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

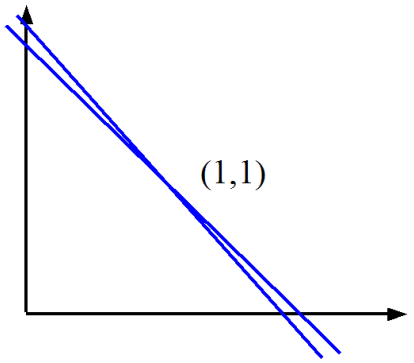


Zmotimo desno stran:

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

Ta sistem je neobčutljiv.

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



Zmotimo desno stran:

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

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

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

Čeprav so ničle enostavne in lepo separirane, majhna motnja povzroči velike spremembe.

Stopnja občutljivosti

Stopnjo občutljivosti merimo z razmerjem med velikostjo spremembe rezultata in velikostjo spremembe podatkov.

Zgled: Naj bo $f : \mathbb{R} \to \mathbb{R}$ zvezna in odvedljiva funkcija. Zanima nas razlika med $f(x)$ in $f(x + \delta x)$, kjer je δx majhna motnja.

Absolutna občutljivost: Iz ocene

$$|f(x + \delta x) - f(x)| \approx |f'(x)| \cdot |\delta x|,$$

sledi, da je $|f'(x)|$ *absolutna občutljivost* f v točki x.

Relativna občutljivost: Iz ocene

$$\frac{|f(x + \delta x) - f(x)|}{|f(x)|} \approx \frac{|f'(x)| \cdot |x|}{|f(x)|} \cdot \frac{|\delta x|}{|x|}$$

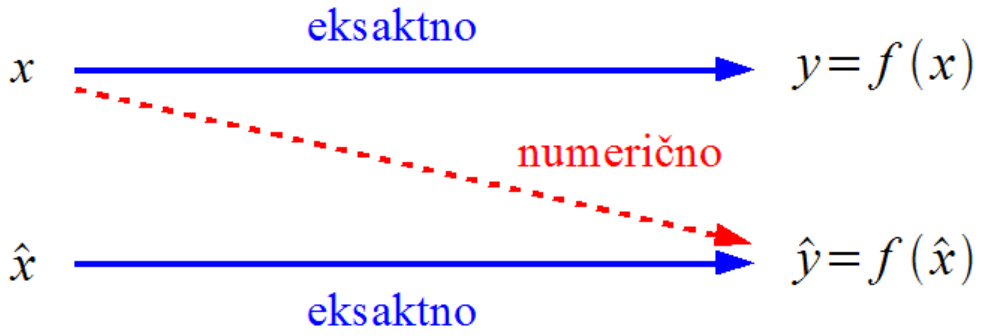
sledi, da je $\dfrac{|f'(x)| \cdot |x|}{|f(x)|}$ *relativna občutljivost* f v točki x.

1.4 Stabilnost metode

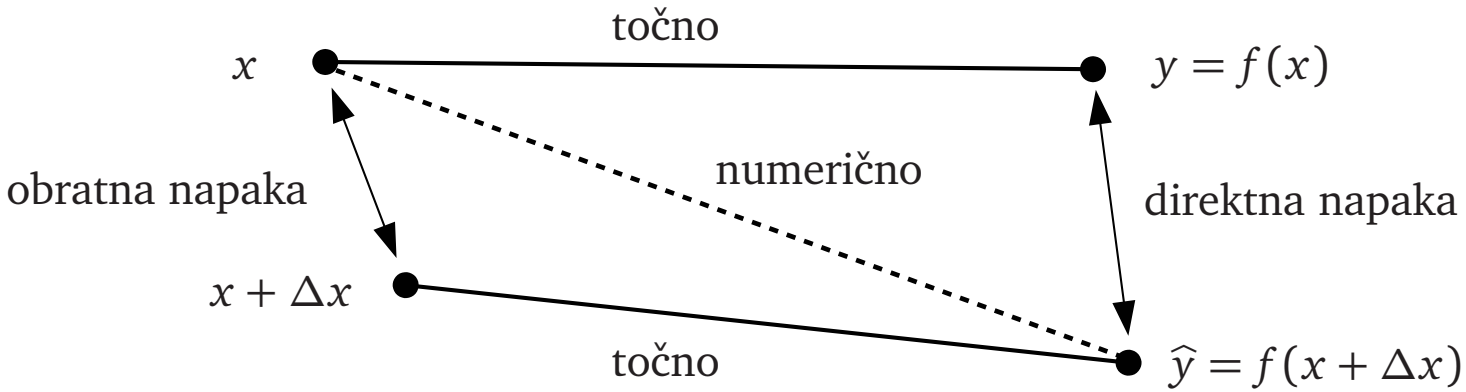
Pri računskem procesu pravimo, da je *stabilen* oz. *nestabilen*, ločimo pa *direktno* in *obratno stabilnost*. S tem se ukvarja *analiza zaokrožitvenih napak*.

- direktna analiza: Iz x namesto $y = f(x)$ izračunamo $\widehat{y}$. Če je za vsak x razlika med y in $\widehat{y}$ majhna (absolutno oz. relativno), je proces direktno stabilen (absolutno oz. relativno), sicer pa nestabilen.

- obratna analiza: Iz x namesto $y = f(x)$ izračunamo $\widehat{y}$. Sedaj se vprašamo, za koliko moramo spremeniti argument x v $\widehat{x}$, da bo $f(\widehat{x}) = \widehat{y}$. Če je za vsak x razlika med x in $\widehat{x}$ majhna (absolutno oz. relativno), je proces obratno stabilen (absolutno oz. relativno), sicer pa nestabilen.



Povezava med občutljivostjo, direktno in obratno napako



Direktna napaka: $\widehat{y} - y$

Obratna napaka: Δx

Groba ocena:

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

Občutljivost, stabilnost in natančnost

Algoritem je stabilen, če so rezultati, ki jih vrne, relativno neobčutljivi na motnje, ki se pojavijo zaradi zaokrožitvenih napak med samim računanjem.

Obratno stabilen algoritem tako vrne točno rešitev bližnjega problema.

Če je problem občutljiv, se točna rešitev bližnjega problema lahko zelo razlikuje od točne rešitve začetnega problema in izračunani rezultat je nenatančen.

Nenatančnost je tako lahko posledica:

- uporabe stabilnega algoritma na občutljivem problemu,
- uporabe nestabilnega algoritma na neobčutljivem problemu.

Natančnost je zagotovljena, kadar neobčutljiv problem rešimo s stabilno numerično metodo.

1.5 Tri vrste napak pri numeričnem računanju

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

Izvori napake so:

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

Celotno napako lahko razdelimo na tri dele

Neodstranljiva napaka: Namesto z x računamo s približkom $\overline{x}$ in namesto $y = f(x)$ izračunamo $\overline{y} = f(\overline{x})$. Neodstranljiva napaka je $D_n = y - \overline{y}$.

D_n je posledica napak začetnih podatkov.

Zgled: *Računanje* $\sin(\pi/10)$ *z osnovnimi operacijami v* $P(10, 4, -5, 5)$

Namesto z $x = \pi/10$ *računamo z* $\overline{x} = 0.3142 \cdot 10^0$

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

Napaka metode: Namesto f računamo vrednost funkcije g, ki jo lahko izračunamo s končnim številom operacij. Namesto $\overline{y} = f(\overline{x})$ tako izračunamo $\widetilde{y} = g(\overline{x})$. Napaka metode je $D_m = \overline{y} - \widetilde{y}$.

Pri sami numerični metodi pogosto neskončen proces nadomestimo s končnim (seštejemo le končno členov neskončne vrste, po končnem številu korakov prekinemo iterativno metodo).

Zgled: *Namesto* $\sin(\overline{x})$ *izračunamo* $g(\overline{x})$ *za* $g(x) = x - x^3/6$.

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

Celotna napaka

Zaokrožitvena napaka: Pri računanju $\widetilde{y} = g(\overline{x})$ se pri vsaki računski operaciji pojavi zaokrožitvena napaka, tako da namesto $\widetilde{y}$ izračunamo $\widehat{y}$. Sama vrednost $\widehat{y}$ je odvisna od vrstnega reda operacij in načina izračuna $g(\overline{x})$. Zaokrožitvena napaka je $D_z = \widetilde{y} - \widehat{y}$.

Zgled: *D_z je odvisna je od vrstnega reda in načina računanja $g(\overline{x})$. Primer:*

$$\begin{aligned} a_1 &= \mathrm{fl}(\overline{x} \cdot \overline{x}) = \mathrm{fl}(0.09872164) = 0.9872 \cdot 10^{-1} \\ a_2 &= \mathrm{fl}(a_1 \cdot \overline{x}) = \mathrm{fl}(0.031017824) = 0.3102 \cdot 10^{-1} \\ a_3 &= \mathrm{fl}(a_2/6) = \mathrm{fl}(0.00517) = 0.5170 \cdot 10^{-2} \\ \widehat{y} &= \mathrm{fl}(\overline{x} - a_3) = \mathrm{fl}(0.309032) = 0.3090 \cdot 10^{0} \end{aligned}$$

$$D_z = \widetilde{y} - g(\overline{x}) = 3.0 \cdot 10^{-5}$$

Celotna napaka: Končna napaka je $D = D_n + D_m + D_z$. Velja

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

Zgled: *Celotna napaka je* $D = D_n + D_m + D_z = 1.7 \cdot 10^{-5}$

1.6.1 Analiza zaokrožitvenih napak za produkt $n+1$ števil

Računamo produkt $p = x_0 x_1 \cdots x_n$ predstavljivih števil $x_0, x_1, \ldots, x_n$.

Eksaktni algoritem:

$p_0 = x_0$
$i = 1, \ldots, n$
$\quad p_i = p_{i-1} x_i$
$p = p_n$

Dejanski algoritem:

$\widehat{p}_0 = x_0$
$i = 1, \ldots, n$
$\quad \widehat{p}_i = \widehat{p}_{i-1} x_i (1+\delta_i), \quad |\delta_i| \leq u$
$\widehat{p} = \widehat{p}_n$

Dobimo

$$\widehat{p} = p(1+\gamma) = p(1+\delta_1)\cdots(1+\delta_n).$$

Velja

$$(1-u)^n \leq 1+\gamma \leq (1+u)^n.$$

Ocenimo

$$(1+u)^n = 1 + \binom{n}{1}u + \binom{n}{2}u^2 + \cdots = 1 + nu + \mathcal{O}(u^2),$$

$$(1-u)^n \geq 1 - nu. \quad \text{(indukcija)}$$

Če je $nu \ll 1$ ocenimo $|\gamma| \leq nu$. Računanje produkta $n+1$ števil je direktno in obratno stabilno.

1.6.2 Analiza zaokrožitvenih napak za skalarni produkt

Imamo dva vektorja predstavljivih števil $x = (x_1 \ \cdots \ x_n)^T$ in $y = (y_1 \ \cdots \ y_n)^T$, računamo pa $s = y^T x = \sum_{i=1}^{n} x_i y_i$.

Eksaktni algoritem:

$s_0 = 0$

$i = 1, \ldots, n$

$\quad p_i = x_i y_i$

$\quad s_i = s_{i-1} + p_i$

$s = s_n$

Dejanski algoritem:

$\widehat{s}_0 = 0$

$i = 1, \ldots, n$

$\quad \widehat{p}_i = x_i y_i (1 + \alpha_i), \ |\alpha_i| \leq u$

$\quad \widehat{s}_i = (\widehat{s}_{i-1} + \widehat{p}_i)(1 + \beta_i), \ |\beta_i| \leq u$

$\widehat{s} = \widehat{s}_n$

Obratna analiza vrne $\widehat{s} = \sum_{i=1}^{n} x_i y_i (1 + \gamma_i)$, kjer je

$$1 + \gamma_1 = (1 + \alpha_1)(1 + \beta_2) \cdots (1 + \beta_n)$$

in

$$1 + \gamma_i = (1 + \alpha_i)(1 + \beta_i) \cdots (1 + \beta_n), \ i = 2, \ldots, n.$$

Tako dobimo ocene $|\gamma_1| \leq nu$ in $|\gamma_i| \leq (n - i + 2)u$ za $i = 2, \ldots, n$. To pomeni, da je $\widehat{s}$ točni skalarni produkt relativno malo zmotenih vektorjev x in y. Računanje skalarnega produkta je obratno stabilno.

Direktna analiza za izračun skalarnega produkta

Pri direktni analizi najprej izračunamo absolutno napako:

$$\widehat{s} - s = \sum_{i=1}^{n} x_i y_i \gamma_i,$$

torej

$$|\widehat{s} - s| \leq \sum_{i=1}^{n} |x_i| \cdot |y_i| \cdot |\gamma_i| \leq nu \sum_{i=1}^{n} |x_i| \cdot |y_i| = nu|y|^T|x|.$$

Dobimo

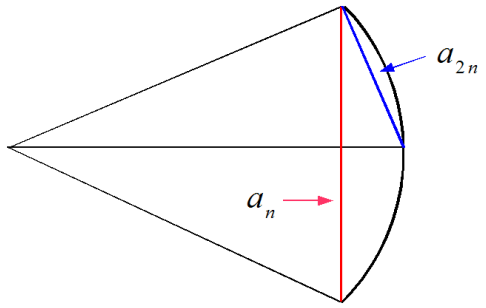
$$\left|\frac{\widehat{s} - s}{s}\right| \leq \frac{|y|^T|x|}{|y^T x|} nu.$$

Če so vsi $x_i y_i$ enakega predznaka, dobimo $\left|\frac{\widehat{s}-s}{s}\right| \leq nu$ in računanje je direktno stabilno, sicer pa imamo v primeru, ko sta vektorja skoraj pravokotna, lahko veliko relativno napako.

V splošnem torej računanje skalarnega produkta ni direktno stabilno.

1.7.1 Poučni primeri - računanje števila π

π je limita obsega S_n pravilnega mnogokotnika, včrtanega v krog s polmerom $r = \frac{1}{2}$. Naj bo a_n stranica pravilnega n-kotnika. Poiščimo zvezo med a_n in a_{2n}:



Velja

$$a_{2n} = \sqrt{\left(\frac{a_n}{2}\right)^2 + \left(\frac{1}{2} - \sqrt{\frac{1}{4} - \left(\frac{a_n}{2}\right)^2}\right)^2} = \sqrt{\frac{1 - \sqrt{1 - a_n^2}}{2}},$$

od tod pa iz $S_n = na_n$ sledi

$$S_{2n} = 2na_{2n} = 2n\sqrt{\frac{1 - \sqrt{1 - \left(\frac{S_n}{n}\right)^2}}{2}}.$$

Računanje števila π, 2.del

Začnemo pri $S_6 = 3$ in uporabljamo formulo $S_{2n} = 2n\sqrt{\frac{1-\sqrt{1-\left(\frac{S_n}{n}\right)^2}}{2}}$.

n	Sn	n	Sn
6	3.0000000	768	3.1430728
12	3.1058285	1536	3.1486604
24	3.1326292	3072	3.1374750
48	3.1393456	6144	3.1819806
96	3.1410186	12288	3.0000000
192	3.1414995	24576	4.2426405
384	3.1416743	49152	0.0000000

Formula odpove, saj pride do odštevanja skoraj enako velikih števil, napaka pa se množi z $2n$. V zgornji tabeli (enojna natančnost) so napačne decimalke rdeče.

Kadar imamo nestabilen postopek, nam ne pomaga niti računanje z večjo natančnostjo. Prava rešitev je preurediti postopek tako, da se med računanjem ne izgublja natančnost.

Računanje števila π, 3.del

Za stabilno računanje je potrebno formulo preurediti. Stabilna oblika je

$$S_{2n} = 2n \sqrt{\frac{\left(1 - \sqrt{1 - \left(\frac{S_n}{n}\right)^2}\right)\left(1 + \sqrt{1 - \left(\frac{S_n}{n}\right)^2}\right)}{2\left(1 + \sqrt{1 - \left(\frac{S_n}{n}\right)^2}\right)}} = S_n \sqrt{\frac{2}{1 + \sqrt{1 - \left(\frac{S_n}{n}\right)^2}}}\,.$$

Sedaj dobimo pravilne rezultate:

n	Sn	n	Sn
6	3.0000000	768	3.1415839
12	3.1058285	1536	3.1375904
24	3.1326287	3072	3.1415920
48	3.1393502	6144	3.1415925
96	3.1410320	12288	3.1415925
192	3.1414526	24576	3.1415925
384	3.1415577	49152	3.1415925

1.7.2 Računanje I_{10}

Integrale $I_n = \int_0^1 x^n e^{x-1} dx$, $n = 0, 1, \ldots$, lahko računamo rekurzivno preko formule

$$I_n = x^n e^{x-1}\Big|_0^1 - n \int_0^1 x^{n-1} e^{x-1} dx = 1 - nI_{n-1},$$

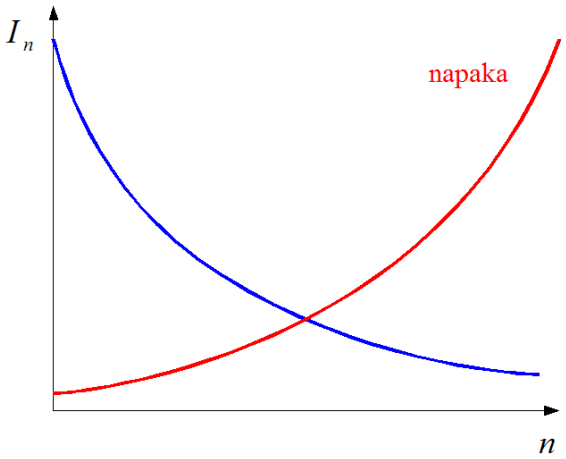
saj poznamo začetno vrednost $I_0 = 1 - e^{-1}$. V enojni natančnosti dobimo

n	I_n	n	I_n
0	0.6321205	7	0.1124296
1	0.3678795	8	0.1005630
2	0.2642411	9	0.0949326
3	0.2072767	10	0.0506744
4	0.1708932	11	0.4425812
5	0.1455340	12	−4.3109741
6	0.1267958	13	57.0426636

Razlog je v formuli $I_n = 1 - nI_{n-1}$. Napaka pri členu I_{n-1} se pomnoži z n in torej po absolutni vrednosti hitro narašča, točne vrednosti I_n pa padajo.

Računanje I_{10}, 2. del

Če računamo v obratni smeri: $I_{n-1} = \frac{1-I_n}{n}$, se napaka v vsakem koraku deli z n. Če začnemo pri nekem dovolj velikem členu, lahko z začetnim $I_n = 0$ izračunamo vse začetne člene dovolj natančno. Če začnemo z $I_{26} = 0$ tako dobimo (v enojni natančnosti) vse člene od I_{12} do I_0 na vse decimalke točno.



$$I_n = 1 - nI_{n-1}$$

n	I_n	n	I_n
0	0.6321205	8	0.1009320
1	0.3678795	9	0.0916123
2	0.2642411	10	0.0838771
3	0.2072766	11	0.0773522
4	0.1708934	12	0.0717733
5	0.1455329	13	0.0669477
6	0.1268024	⋮	⋮
7	0.1123835	26	0.0000000

1.7.3 Reševanje kvadratne enačbe

Rešitvi kvadratne enačbe $ax^2 + bx + c = 0$ sta podani s formulo

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

$a = 1.2345678$, $b = 76543210.5$, $c = 0.1122334455$, dvojna natančnost $\Longrightarrow$

$$x_1 = -6.2000005589000046 \cdot 10^7, \quad x_2 = -6.034970778375905 \cdot 10^{-9},$$

točni ničli pa sta

$$\widetilde{x}_1 = -6.2000005589000046 \cdot 10^7, \quad \widetilde{x}_2 = -1.466275647008561 \cdot 10^{-9}.$$

Težava se je pojavila pri odštevanje približno enako velikih števil pri računanju x_2. Rešitev je, da eno rešitev izračunamo po formuli (odvisno od predznaka b), drugo pa dobimo iz Vietove formule $x_1 x_2 = c/a$.

Tako iz $x_2 = c/(ax_1)$ dobimo pravilno $x_2 = -1.466275647008561 \cdot 10^{-9}$.