MASZYNOWA REPREZENTACJA LICZB

Reprezentacja zmiennoprzecinkowa $x_{t,r}$ liczby x:

$$x_{t,r} = m_t \cdot P^{c_r}$$

m_t - mantysa, $|m_t| < 1$, t - liczba znaków mantysy,

c_r - wykładnik (cecha), r - liczba znaków wykładnika,

P - podstawa (ograniczymy się do $P = 2$).

Naturalna reprezentacja **znormalizowana:**

$$0.5 \leq |m_t| < 1,$$

wówczas t pozycjom mantysy odpowiada t pierwszych pozycji po przecinku mantysy jako liczby,

np.: $m_7 = 1011011 \sim$ 0,1011011, $m_4 = 1011 \sim$ 0,1011.

(jedynka jest zawsze na pierwszej pozycji każdej mantysy znormalizowanej).

MASZYNOWA REPREZENTACJA LICZB 2

Np. dla $t = 4$ i $r = 2$ oraz dwóch bitów znaków przykładowa reprezentacja zmiennoprzecinkowa binarna:

$$0\,10|0\,1011,$$

(niebieskie bity znaków, zero = plus: $(-1)^0 = 1$).

Podana liczba w systemie dziesiętnym:

$$x = 2^{(-1)^0(1\cdot 2^1+0\cdot 2^0)} \cdot (-1)^0(1\cdot 2^{-1} + 0\cdot 2^{-2} + 1\cdot 2^{-3} + 1\cdot 2^{-4}) = 2.75,$$

kolejne potęgi dwójki rosną płynnie od końca mantysy do początku wykładnika.

Zbiór liczb maszynowych dla reprezentacji znormalizowanej, uzupełniony o 0:

$$\{-x_{max}, \ldots, -x_{min}\} \cup \{0\} \cup \{x_{min}, \ldots, x_{max}\},$$

$$\text{gdzie}: \quad x_{min} = m_{min}2^{r_{min}}, \quad x_{max} = m_{max}2^{r_{max}}$$

MASZYNOWA REPREZENTACJA LICZB 3

W standardzie IEEE 754 przyjęto konwencję normalizacji:

$$1 \leq |m_t| < 2,$$

np. $m_7 = 1011011 \sim$ 1,011011, $m_4 = 1011 \sim$ 1,011

Liczba 2.75 zapisana w tej konwencji jest w postaci :

$$x = 2^{(-1)^0(0\cdot 2^1+1\cdot 2^0)} \cdot (-1)^0(1\cdot 2^0 + 0\cdot 2^{-1} + 1\cdot 2^{-2} + 1\cdot 2^{-3}) = 2.75$$

(mantysa pomnożona przez 2, wykładnik zmniejszony o 1).

Wg **standardu IEEE 754**, format 32 bitowy (*single precision*):

- mantysa 24 bitowa, z normalizacją: $1 \leq |m_{24}| < 2$,
- wykładnik 8 bitowy kodowany z przesunięciem o $2^7 - 1 = 127$

co daje zakres liczb ok. od $\pm 1.8 \times 10^{-38}$ do $\pm 3.4 \times 10^{38}$.

MASZYNOWA REPREZENTACJA LICZB 4

bit znaku (1 bit)	wykładnik bez znaku przesunięty o 127 (8 bitów)	znormalizowana mantysa 24 pozycyjna (pierwszy bit zawsze równy 1 - domyślny) (23 bity)

Reprezentacja liczby dziesiętnej w tym standardzie:

$$x_{IEEE754,32} = (-1)^s \cdot m_{24} \cdot 2^{c_8 - 127}, \quad \text{gdzie}$$

s – bit znaku,
m_{24} – liczba dziesiętna odpowiadająca mantysie 24 bitowej,
$c_8 - 127$ – liczba z zakresu -126 ($c_8 \sim 00000001$) do $+127$($c_8 \sim 11111110$),
np. wykładnik zero (dla mantysy z zakresu [1,2)) jest reprezentowany przez $c_8 \sim 01111111$ $(0 \cdot 2^7 + 2^6 + 2^5 + 2^4 + 2^3 + 2^2 + 2^1 + 2^0 = 127)$.

Reprezentacja liczby dziesiętnej $x = 2.75$:

$$0|10000000|01100000000000000000000000000$$

$$\left(x = 2^{(-1)^0(0\cdot 2^1 + 1\cdot 2^0)} \cdot (-1)^0(1 \cdot 2^0 + 0 \cdot 2^{-1} + 1 \cdot 2^{-2} + 1 \cdot 2^{-3}) = 2.75 \right)$$

BŁĘDY REPREZENTACJI

Zbiór liczb maszynowych

$M = \{-x_{max}, \ldots, -x_{min}\} \cup \{0\} \cup \{x_{min}, \ldots, x_{max}\} \subset R$ jest **skończony**;

– im większe t, tym M jest "gęściejszy",

– im większe r, tym M pokrywa większy zakres liczb.

Zakładając $c_r = c$, oznaczmy najlepsze **maszynowe przybliżenie** liczby x przez rd(x) (od *rounding*):

$$\mathrm{rd}\,(x) = m_t \cdot 2^c$$

Przybliżenie jest najlepsze, jeśli:

$$|\mathrm{rd}\,(x) - x| \leq \min_{g \in M} |g - x|$$

Postulat ten jest spełniony przez **typowe zaokrąglenie**:

$$m_t = \sum_{i=1}^{t} e_{-i} \cdot 2^{-i} + e_{-(t+1)} \cdot 2^{-t},$$

gdzie $e_{-1} = 1, \quad e_{-i} = 0$ lub 1 dla $i = 2, 3, ..., t+1$ (wzór dla $\frac{1}{2} \leq m_t < 1$).

BŁĘDY REPREZENTACJI 2

Oszacowanie błędu zaokrąglenia (*roundoff error*):

$$|m - m_t| \leq 2^{-(t+1)}$$

Uzasadnienie:

- na pierwszej odrzucanej pozycji mantysy jest 0, to odrzucana część (z pierwszą jedynką na pozycji $t+2$ lub dalej) jest z zakresu $[0 \;\; 2^{-(t+1)})$
  tzn. mantysa m_t jest z niedomiarem $< \;\; 2^{-(t+1)}$;
- na pierwszej odrzucanej pozycji mantysy jest 1, tzn. odrzucana część jest z zakresu $[2^{-(t+1)} \;\; 2^{-t}) = [\frac{1}{2}2^{-t} \;\; 2^{-t})$,
  ale wtedy zaokrąglamy dodając 2^{-t} do mantysy – ponieważ odrzucana część jest z zakresu $[\frac{1}{2}2^{-t} \;\; 2^{-t})$, to błąd jest dodatni z zakresu

$$2^{-t} - [\frac{1}{2}2^{-t} \;\; 2^{-t}) = (0 \;\; \frac{1}{2}2^{-t}] = (0 \;\; 2^{-(t+1)}],$$

  tzn. mantysa m_t jest z nadmiarem $\leq \;\; 2^{-(t+1)}$.

BŁĘDY REPREZENTACJI 3

Błąd względny reprezentacji:

$$\frac{\mathrm{rd}\,(x)-x}{x}=\frac{m_t\cdot 2^c-m\cdot 2^c}{m\cdot 2^c}=\frac{m_t-m}{m},$$

skąd **oszacowanie błędu względnego zaokrąglenia**:

$$\left|\frac{\mathrm{rd}\,(x)-x}{x}\right|=\frac{|m_t-m|}{|m|}\leq\frac{2^{-(t+1)}}{2^{-1}}=2^{-t},\quad \text{gdy}\quad |m|\geq 2^{-1}\quad (\frac{1}{2}\leq m_t<1)$$

Zależność

$$\left|\frac{\mathrm{rd}\,(x)-x}{x}\right|=\frac{|m_t-m|}{|m|}\leq 2^{-t}$$

jest uniwersalna i nie zależy od przyjętego wzorca normalizacji mantysy.

Stąd, *maksymalny błąd względny reprezentacji zmiennoprzecinkowej zależy jedynie od liczby bitów mantysy* i nazywany jest **dokładnością maszynową**, czy **precyzją maszynową (machine precision)** i oznaczany jest przez eps.

Dla mantysy t-bitowej i zaokrąglania (dla każdej normalizacji) $eps=2^{-t}$.

BŁĘDY REPREZENTACJI 4

Mamy:

$$|\mathrm{rd(x)} - x| \leq eps \cdot |x|$$

stąd błąd reprezentacji można zapisać w równoważnej postaci:

$$\mathrm{rd(x)} - x = \varepsilon \cdot x, \qquad |\varepsilon| \leq eps,$$

gdzie ε to liczba reprezentująca błąd zaokrąglenia liczby x (dodatnia, zero lub ujemna).

Ostatnią równość zapisuje się powszechnie w postaci

$$\mathrm{rd}\,(x) = x\,(1 + \varepsilon)\,, \qquad |\varepsilon| \leq eps$$

BŁĘDY REPREZENTACJI 5

Aproksymacja maszynowa liczby przez **obcięcie**:

$$m_t = \sum_{i=1}^{t} e_{-i} \cdot 2^{-i} \quad (\text{wzór dla } \frac{1}{2} \leq m_t < 1),$$

wprowadza tzw. błąd obcięcia (*truncation error, chopping error*).

Odrzucana część jest z zakresu $[2^{-(t+1)}, 2^{-t})$,
stąd m_t jest z niedomiarem $\leq 2^{-t}$,

$$|m - m_t| \leq 2^{-t}.$$

Przy zaokrąglaniu było $|m - m_t| \leq 2^{-(t+1)}$,
stąd teraz $eps = 2^{-t+1}$ (zamiast 2^{-t}),

$$\mathrm{rd}\,(x) = x\,(1+\varepsilon)\,, \quad |\varepsilon| \leq 2^{-t+1} = eps.$$

BŁĘDY REPREZENTACJI 6

W przypadku stosowania słowa o podwójnej długości mantysy:

$$eps_{mdl} = (eps)^2 \,,$$

gdzie indeks dolny "mdl" oznacza *mantissa double length*.

Wg standardu IEEE 754 liczba w podwójnej precyzji ("double precision"):

bit znaku (1 bit)	wykładnik przesunięty (11 bitów)	znormalizowana mantysa 53 pozycyjna (pierwszy bit zawsze równy 1 - domyślny) (52 bity)

co daje zakres liczb ok. od $\pm 2.2 \times 10^{-308}$ do $\pm 1.8 \times 10^{308}$.

ARYTMETYKA ZMIENNOPOZYCYJNA

Wynikiem działań na liczbach maszynowych nie są na ogół liczby maszynowe.

Oznaczając przez fl (*floating point*) wynik działania zmiennopozycyjnego, zapisujemy **wyniki działań elementarnych** w postaci:

$$fl\,(x \pm y) = (x \pm y) \cdot (1 + \varepsilon)\,,$$
$$fl\,(x \cdot y) = (x \cdot y)\,(1 + \varepsilon)\,,$$
$$fl\,(x/y) = (x/y)\,(1 + \varepsilon)\,, \quad \text{gdzie } |\varepsilon| \leq eps.$$

Przyczyna: Standard IEEE 754 wymaga organizacji jednostki arytmetycznej komputera zapewniającej osiąganie takiej dokładności.

Z reguły funkcje elementarne również są realizowane w taki sposób, np.:

$$fl\left(\sqrt{x}\right) = \left(\sqrt{x}\right) \cdot (1 + \varepsilon)\,, \quad |\varepsilon| \leq eps.$$

Uwaga: dokładność maszynową eps można również określić jako najmniejszą dodatnią liczbę maszynową g taką, że zachodzi relacja: $fl\,(1 + g) > 1$, tzn.

$$eps = \min\{g \epsilon M : fl(1 + g) > 1;\ g > 0\}.$$

PROPAGACJA BŁĘDÓW

W programach mamy:

- dane wejściowe obarczone błędami reprezentacji maszynowej,

- ciągi elementarnych operacji arytmetycznych (i obliczania funkcji elementarnych), przy wykonywaniu generujących błędy numeryczne.

Błędy te ulegają propagacji – wymnażaniu, sumowaniu, itp., co może prowadzić do poważnej kumulacji błędów i stąd
całkowity błąd względny wyniku zadania obliczanego numerycznie jest z reguły znacznie większy od dokładności maszynowej.

Dwa sposoby analizy błędy wyniku:

– metodą probablistyczną zakładając, że poszczególne błędy jednostkowe to niezależne, nieskorelowane zmienne losowe i wyznaczając np. błąd średni,

– metodą najgorszego przypadku, wyznaczając oszacowanie maksymalnego możliwego modułu błędu,

– metodami analitycznymi, w ograniczonym zakresie prostszych przypadków.

PROPAGACJA BŁĘDÓW – PRZYKŁAD

Przykład. Oszacowanie błędu maksymalnego zadania dodawania dwóch liczb:

$$y = a + b,$$

Realizując algorytm w arytmetyce zmiennopozycyjnej mamy

$$y = [a(1+\varepsilon_a) + b(1+\varepsilon_b)](1+\varepsilon_1)$$

gdzie $|\varepsilon_a|, |\varepsilon_b|, |\varepsilon_1| \leq eps$.

Przekształcając powyższe wyrażenie dostajemy

$$\begin{aligned}
y &= (a + a\varepsilon_a + b + b\varepsilon_b)(1+\varepsilon_1) \\
&= a + a\varepsilon_a + b + b\varepsilon_b + a\varepsilon_1 + a\varepsilon_a\varepsilon_1 + b\varepsilon_1 + b\varepsilon_b\varepsilon_1 \\
&\underset{1}{=} a + a\varepsilon_a + b + b\varepsilon_b + a\varepsilon_1 + b\varepsilon_1 \\
&= a + b + a(\varepsilon_a + \varepsilon_1) + b(\varepsilon_b + \varepsilon_1) \\
&= (a+b)\left[1 + \frac{a(\varepsilon_a+\varepsilon_1)}{a+b} + \frac{b(\varepsilon_b+\varepsilon_1)}{a+b}\right] = (a+b)\,[1+\delta]
\end{aligned}$$

PROPAGACJA BŁĘDÓW – PRZYKŁAD (2)

Błąd względny:

$$\delta = \frac{a(\varepsilon_a + \varepsilon_1)}{a+b} + \frac{b(\varepsilon_b + \varepsilon_1)}{a+b}$$

Szacujemy maksymalny moduł błędu:

$$\begin{aligned}
|\delta| &= \left| \frac{a(\varepsilon_a + \varepsilon_1)}{a+b} + \frac{b(\varepsilon_b + \varepsilon_1)}{a+b} \right| \\
&\leq \left| \frac{a(\varepsilon_a + \varepsilon_1)}{a+b} \right| + \left| \frac{b(\varepsilon_b + \varepsilon_1)}{a+b} \right| \\
&= \frac{|a|\,|(\varepsilon_a + \varepsilon_1)|}{|a+b|} + \frac{|b|\,|(\varepsilon_b + \varepsilon_1)|}{|a+b|} \\
&\leq \frac{|a|2eps}{|a+b|} + \frac{|b|2eps}{|a+b|} = \frac{|a|+|b|}{|a+b|} 2eps
\end{aligned}$$

$$\left(\text{Dla algorytmu } y = (a+b)+c \;\; \text{otrzymamy}: \quad \delta \leq \frac{(|a|+|b|+|c|)}{|a+b+c|} 3eps \right)$$

UWARUNKOWANIE ZADANIA

Matematyczne **zadanie obliczeniowe** – odwzorowanie $\phi : R^n \longmapsto R^m$,

$$\mathbf{w} = \phi(\mathbf{d}),$$

gdzie: $\mathbf{d} = [d_1\ d_2 \cdots d_n]^T$ - wektor danych,
$\mathbf{w} = [w_1\ w_2 \cdots w_m]^T$ - wektor wyników.

W praktyce mamy jedynie reprezentacje maszynowe danych, tj.

$$rd(d_i) = d_i(1 + \varepsilon_i), \quad \text{gdzie } |\varepsilon_i| \leq eps,\ i = 1, ..., n.$$

Definicja. Zadanie obliczeniowe nazywamy **źle uwarunkowanym**, jeśli niewielkie (względne) zaburzenia danych zadania powodują duże (względne) zmiany jego rozwiązania.

Wskaźnikiem uwarunkowania nazwiemy wielkość charakteryzującą ilościowo zmiany wartości wyniku w stosunku do zmian (zaburzeń) wartości danych.

UWARUNKOWANIE ZADANIA Prosty przykład

Przykład. Policzmy iloczyn skalarny

$$w = \phi(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{n} a_i b_i,$$

dla:

a) $\mathbf{a} = [1 \;\; 2 \;\; 3]$, $\mathbf{b} = [2 \;\; 6 \;\; -5]$, wynik: $w = -1$,

b) $\mathbf{a} = [1.02 \;\; 2.04 \;\; 2.96]$, $\mathbf{b} = [2 \;\; 6 \;\; -5]$, wynik: $w = -0.52$,

a następnie:

c) $\mathbf{a} = [1 \;\; 2 \;\; 3]$, $\mathbf{b} = [2 \;\; 6 \;\; +5]$, wynik: $w = 29$,

d) $\mathbf{a} = [1.02 \;\; 2.04 \;\; 2.96]$, $\mathbf{b} = [2 \;\; 6 \;\; +5]$, wynik: $w = 29.8$ – zaburzenie wyniku na poziomie zaburzenia danych !

Wniosek: **Uwarunkowanie zadania zależy od konkretnych wartości danych**, dla których to zadanie liczymy.

UWARUNKOWANIE ZADANIA 2

Błędy danych:

bezwzględny: $\Delta\mathbf{d} = [\Delta d_1\ \Delta d_2 \cdots \Delta d_n]^{\mathrm{T}}$, względny: $\frac{\|\Delta\mathbf{d}\|}{\|\mathbf{d}\|}$.

Błąd bezwzględny wyniku: $\Delta\mathbf{w} = [\Delta w_1\ \Delta w_1 \cdots \Delta w_m]^{\mathrm{T}}$,

błąd względny wyniku:

$$\frac{\|\Delta\mathbf{w}\|}{\|\mathbf{w}\|} = \frac{\|\phi(\mathbf{d}+\Delta\mathbf{d}) - \phi(\mathbf{d})\|}{\|\phi(\mathbf{d})\|}.$$

Wskaźnik uwarunkowania:

$$\mathrm{cond}_\phi(\mathbf{d}) = \lim_{\delta\to 0} \sup_{\|\Delta\mathbf{d}\|\le\delta} \frac{\frac{\|\phi(\mathbf{d}+\Delta\mathbf{d})-\phi(\mathbf{d})\|}{\|\phi(\mathbf{d})\|}}{\frac{\|\Delta\mathbf{d}\|}{\|\mathbf{d}\|}},$$

tzn. granica kresu górnego ilorazu błędów względnych wyniku i danych, względem wszystkich malejących $\Delta\mathbf{d}$, dla których zadanie jest dobrze postawione.

UWARUNKOWANIE ZADANIA 3

Niech

$$\phi(\mathbf{d}) = f(\mathbf{d}),\ \ f : \mathbb{R}^n \to \mathbb{R},$$

f różniczkowalna, ograniczamy się do **normy euklidesowej wektorów**.

Ponieważ dla dostatecznie małych $\triangle\mathbf{d}$:

$$|f(\mathbf{d}+\Delta\mathbf{d}) - f(\mathbf{d})| \approx \left| f'(\mathbf{d}) \cdot \triangle\mathbf{d} \right|$$

(gdzie $f'(\mathbf{d}) = \left[\frac{\partial f(\mathbf{d})}{\partial d_1} \ \ \frac{\partial f(\mathbf{d})}{\partial d_2} \ \ \cdots \ \ \frac{\partial f(\mathbf{d})}{\partial d_n} \right]$ – wektor wierszowy)

oraz dla normy euklidesowej i każdego $\delta > 0$

$$\max_{\|\triangle\mathbf{d}\|=\delta} \left| f'(\mathbf{d}) \cdot \triangle\mathbf{d} \right| = \left\| f'(\mathbf{d}) \right\| \|\triangle\mathbf{d}\| = \left\| f'(\mathbf{d}) \right\| \delta$$

(równość osiągana dla współliniowych wektorów),

to można sformułować wskaźnik uwarunkowania $\mathrm{cond}_f(\mathbf{d})$ w postaci

$$\mathrm{cond}_f(\mathbf{d}) = \lim_{\delta\to 0} \left(\max_{\|\Delta\mathbf{d}\|=\delta} \frac{\frac{|f(\mathbf{d}+\Delta\mathbf{d})-f(\mathbf{d})|}{|f(\mathbf{d})|}}{\frac{\|\Delta\mathbf{d}\|}{\|\mathbf{d}\|}} \right) = \frac{\frac{\left\| f'(\mathbf{d}) \right\| \delta}{|f(\mathbf{d})|}}{\frac{\delta}{\|\mathbf{d}\|}} = \frac{\left\| f'(\mathbf{d}) \right\| \|\mathbf{d}\|}{|f(\mathbf{d})|}$$

WSKAŹNIK UWARUNKOWANIA – Przykład 1

$$\mathrm{cond}_f(\mathbf{d}) = \frac{\left\| f'(\mathbf{d}) \right\| \|\mathbf{d}\|}{|f(\mathbf{d})|}$$

Obliczymy powyższy wskaźnik uwarunkowania dla $\phi(\mathbf{d}) = f(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{n} a_i b_i$.

Wektor danych zdefiniujemy w postaci

$$\mathbf{d} = [a_1 \ a_2 \cdots a_n \ b_1 \ b_2 \cdots b_n]^{\mathrm{T}}$$

Mamy

$$\mathrm{cond}_f(\mathbf{d}) = \frac{\|[b_1 \ b_2 \cdots b_n \ a_1 \ a_2 \cdots a_n]\|_2 \, \|\mathbf{d}\|_2}{|\sum_{i=1}^{n} a_i b_i|} = \frac{\sum_{i=1}^{n} (a_i^2 + b_i^2)}{|\sum_{i=1}^{n} a_i b_i|}$$

Dla danych w poprzednim przykładzie mamy:

1) $\mathbf{a} = [1 \ \ 2 \ \ 3]$, $\mathbf{b} = [2 \ \ 6 \ \ -5]$, stąd $\mathrm{cond}_f(\mathbf{d}) = 79$

2) $\mathbf{a} = [1 \ \ 2 \ \ 3]$, $\mathbf{b} = [2 \ \ 6 \ \ +5]$, stąd $\mathrm{cond}_f(\mathbf{d}) = \frac{79}{29} \cong 2.7$.

WSKAŹNIK UWARUNKOWANIA – Przykład 2

Aproksymacja pochodnej funkcji $g(x)$ ilorazem różnicowym wstecznym, tzn.

$$f(\mathbf{d}) = \frac{g(x) - g(x-h)}{h}$$

gdzie $\mathbf{d} = [g(x)\ g(x-h)]^{\mathrm{T}}$ – wektor danych z błędami względnymi $\varepsilon_1, \varepsilon_2$, tzn.

$$\begin{aligned} fl(g(x)) &= g(x)(1+\varepsilon_1), \quad \varepsilon_1 \le Eps, \\ fl(g(x-h)) &= g(x-h)(1+\varepsilon_2), \quad \varepsilon_2 \le Eps, \quad Eps \ge eps. \end{aligned}$$

Jeśli krok $h > 0$ jest potęgą liczby 2 (dana bez błędu), to

$$\begin{aligned} \mathrm{cond}_f(\mathbf{d}) &= \frac{\|[1/h\ \ -1/h]\|_2\, \|[g(x)\ g(x-h)]\|_2}{|g(x) - g(x-h)|/h} \\ &= \frac{\sqrt{2}\sqrt{g(x)^2 + g(x-h)^2}}{|g(x) - g(x-h)|} \cong \frac{2|g(x)|}{|g(x) - g(x-h)|} \end{aligned}$$

Wniosek: malenie h zmniejsza $|g(x) - g(x-h)|$ – uwarunkowanie pogarsza się,
ale: im mniejszy krok h tym mniejszy bląd aproksymacji pochodnej – efekt przeciwny.

Stąd: istnieje optymalny h minimalizujący sumę błędów aproksymacji i numerycznego.

WSKAŹNIK UWARUNKOWANIA – Przykład 3

Rozważymy uwarunkowanie funkcji wyznaczających różne pierwiastki wielomianu kwadratowego $w(x) = ax^2 + bx + c$, $a > 0$, wg wzorów:

$$x_1(a,b,c) = \frac{-b + \sqrt{b^2 - 4ac}}{2a},$$

$$x_2(a,b,c) = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

Wektor danych to wektor współczynników wielomianu, $\mathbf{d} = [a\ b\ c]^{\mathrm{T}}$.

Mamy(dla $b^2 - 4ac \neq 0$):

$$x_1'(a,b,c) = \left[\frac{2ac - b^2 + b\sqrt{b^2 - 4ac}}{2a^2\sqrt{b^2 - 4ac}} \quad \frac{b - \sqrt{b^2 - 4ac}}{2a\sqrt{b^2 - 4ac}} \quad \frac{-1}{\sqrt{b^2 - 4ac}} \right],$$

$$x_2'(a,b,c) = \left[\frac{-2ac + b^2 + b\sqrt{b^2 - 4ac}}{2a^2\sqrt{b^2 - 4ac}} \quad \frac{-b - \sqrt{b^2 - 4ac}}{2a\sqrt{b^2 - 4ac}} \quad \frac{1}{\sqrt{b^2 - 4ac}} \right]$$

WSKAŹNIK UWARUNKOWANIA – Przykład 3 cd.

$$\mathrm{cond}_{x_1}(\mathbf{d}) = \frac{2a\left\|x_1'(\mathbf{d})\right\|\sqrt{a^2+b^2+c^2}}{|-b+\sqrt{b^2-4ac}|} \qquad \left(x_1(a,b,c) = \frac{-b+\sqrt{b^2-4ac}}{2a}\right)$$

$$\mathrm{cond}_{x_2}(\mathbf{d}) = \frac{2a\left\|x_2'(\mathbf{d})\right\|\sqrt{a^2+b^2+c^2}}{|-b-\sqrt{b^2-4ac}|} \qquad \left(x_1(a,b,c) = \frac{-b-\sqrt{b^2-4ac}}{2a}\right)$$

Wskaźnik uwarunkowania funkcji definiującej pierwiastek o mniejszym module bardzo wzrasta, gdy $b^2 \gg |4ac|$ – możliwy znacznie większy błąd względny.

Aby tego uniknąć, wykorzystujemy tylko wzór dla pierwiastka o większym module (z większym modułem licznika we wzorze na pierwiastek), drugi pierwiastek wyznaczamy z wzoru Viéte'a $x_1 + x_2 = -b/a$.

Jeśli potrzebny jest tylko pierwiastek o mniejszym module, stosujemy wzory

$$x_1(a,b,c) = \frac{2c}{-b-\sqrt{b^2-4ac}}, \qquad x_2(a,b,c) = \frac{2c}{-b+\sqrt{b^2-4ac}},$$

(wykorzystując jedynie ten wzór, w którym mianownik ma większy moduł).

WSKAŹNIK UWARUNKOWANIA (SZACOWANIE)

Jeśli $\Delta d_i = d_i\varepsilon_i$, $\mathrm{rd}(d_i) = d_i(1+\varepsilon_i)$, $|\varepsilon_i| \leq eps$, $i = 1, ..., n$,

to dla normy $\|\cdot\|_2$ (też $\|\cdot\|_1$, $\|\cdot\|_\infty$) mamy $\frac{\|\Delta\mathbf{d}\|}{\|\mathbf{d}\|} \leq eps$, ponieważ

$$\frac{\|\Delta\mathbf{d}\|}{\|\ \mathbf{d}\ \|} = \frac{\sqrt{(d_1\varepsilon_1)^2 + (d_2\varepsilon_2)^2 + \cdots + (d_n\varepsilon_n)^2}}{\|\ \mathbf{d}\ \|} \leq \frac{\|\ \mathbf{d}\ \|eps}{\|\ \mathbf{d}\ \|} = eps.$$

Z naszej definicji wskaźnika uwarunkowania wynika bezpośrednio

$$\frac{\|\Delta\mathbf{w}\|}{\|\mathbf{w}\|} \approx \mathrm{cond}_\phi(\mathbf{d}) \cdot \frac{\|\Delta\mathbf{d}\|}{\|\ \mathbf{d}\ \|}$$

Nie znając postaci funkcyjnej zadania, można póbować szacować wskaźnik jako najmniejszą możliwą do wyznaczenia wartość $\mathrm{cond}^s_\phi(\mathbf{d})$, dla której zachodzi (dla wszystkich $\triangle\mathbf{d}$)

$$\frac{\|\Delta\mathbf{w}\|}{\|\mathbf{w}\|} \leq \mathrm{cond}^s_\phi(\mathbf{d}) \cdot \frac{\|\Delta\mathbf{d}\|}{\|\ \mathbf{d}\ \|}.$$

Tak prosta postać często jest trudna do uzyskania, ogólniej szukamy możliwie dokładnego oszacowania θ

$$\frac{\|\Delta\mathbf{w}\|}{\|\mathbf{w}\|} \leq \theta\left(\frac{\|\Delta\mathbf{d}_1\|}{\|\mathbf{d}_1\|}, \ldots, \frac{\|\Delta\mathbf{d}_m\|}{\|\mathbf{d}_m\|}\right)$$

gdzie argumentami są błędy względne danych/grup danych (*będą przykłady*).

ALGORYTM I JEGO NUMERYCZNE REALIZACJE

Trzy *podstawowe i różne* pojęcia:

- **Zadanie obliczeniowe** (matematyczne): $\mathbf{w} = \phi(\mathbf{d})$,
- **Algorytm** $A(\mathbf{d})$ obliczenia wyniku zadania $\phi(\mathbf{d})$, tj sposób wyznaczenia wyniku zgodnie z jednoznacznie określoną kolejnością wykonywania elementarnych działań arytmetycznych,
- **Numeryczna realizacja** $fl(A(\mathbf{d}))$ **algorytmu** $A(\mathbf{d})$, polegająca na:
  a) zastąpieniu wielkości liczbowych występujących w $A(\mathbf{d})$ ich reprezentacjami zmiennopozycyjnymi,
  b) wykonaniu operacji arytmetycznych (działań elementarnych, obliczaniu wartości funkcji standardowych) w arytmetyce zmiennopozycyjnej „fl", tj. w sposób przybliżony.

Przykład: zadanie: $\phi(a,b) = a^2 - b^2$,

algorytm $A1(a,b)$: $a \cdot a - b \cdot b$,

algorytm $A2(a,b)$: $(a+b) \cdot (a-b)$

ALGORYTM I JEGO NUMERYCZNE REALIZACJE 2

Realizacja numeryczna algorytmu $A1(a,b)$:

$$\begin{aligned}
fl(A1(a,b) &= fl(a\cdot a - b\cdot b)\\
&= fl[fl(a\cdot a) - fl(b\cdot b)]\\
&= [a^2(1+\varepsilon_1) - b^2(1+\varepsilon_2)](1+\varepsilon_3) \qquad (|\varepsilon_i| \le eps\ )\\
&= [a^2 - b^2 + a^2\varepsilon_1 - b^2\varepsilon_2](1+\varepsilon_3)\\
&= a^2 - b^2 + a^2\varepsilon_1 - b^2\varepsilon_2 + (a^2-b^2)\varepsilon_3 + a^2\varepsilon_1\varepsilon_3 - b^2\varepsilon_2\varepsilon_3\\
&\underset{1}{=} a^2 - b^2 + a^2\varepsilon_1 - b^2\varepsilon_2 + (a^2-b^2)\varepsilon_3\\
&= (a^2-b^2)(1 + \tfrac{a^2\varepsilon_1 - b^2\varepsilon_2}{a^2-b^2} + \varepsilon_3)\\
&= (a^2-b^2)(1+\delta_1),
\end{aligned}$$

gdzie

$$\delta_1 = \frac{a^2\varepsilon_1 - b^2\varepsilon_2}{a^2-b^2} + \varepsilon_3$$

$$|\delta_1| = \left|\frac{a^2\varepsilon_1 - b^2\varepsilon_2}{a^2-b^2} + \varepsilon_3\right| \le \frac{a^2+b^2}{|a^2-b^2|}eps + eps.$$

ALGORYTM I JEGO NUMERYCZNE REALIZACJE 3

Realizacja numeryczna algorytmu $A2(a,b)$:

$$\begin{aligned} fl(A2(a,b) &= fl[(a+b)\cdot(a-b)] \\ &= fl[fl(a+b)\cdot fl(a-b)] \\ &= [(a+b)(1+\varepsilon_1)\cdot(a-b)(1+\varepsilon_2)](1+\varepsilon_3) \\ &= (a^2-b^2)(1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3) \\ &\underset{1}{=} (a^2-b^2)(1+\varepsilon_1+\varepsilon_2+\varepsilon_3) \\ &= (a^2-b^2)(1+\delta_2), \end{aligned}$$

gdzie

$$\delta_2 = \varepsilon_1+\varepsilon_2+\varepsilon_3$$

$$|\delta_2| = |\varepsilon_1+\varepsilon_2+\varepsilon_3| \leq 3eps.$$

(Dla algorytmu A1(a,b) mieliśmy:

$$|\delta_1| \leq \frac{a^2+b^2}{|a^2-b^2|}eps + eps. \ )$$

Przyjmując względny błąd danych: $\frac{\|\triangle \mathbf{d}\|}{\|\mathbf{d}\|} \leq eps$

mamy
$$\frac{\| \phi(\mathbf{d}+\Delta\mathbf{d}) - \phi(\mathbf{d}) \|}{\|\phi(\mathbf{d})\|} \leq \mathrm{cond}_\phi(\mathbf{d}) \cdot eps$$

Definicja: Algorytm $A(\mathbf{d})$ realizujący zadanie $\phi(\mathbf{d})$ nazywamy **numerycznie stabilnym**, jeśli istnieje taka stała dodatnia K_s, że dla każdych danych $\mathbf{d} \in D$ i dostatecznie małego eps (tj. dostatecznie silnej arytmetyki) zachodzi nierówność:
$$\frac{\|fl(A(\mathbf{d})) - \phi(\mathbf{d})\|}{\|\phi(\mathbf{d})\|} \leq K_s \cdot \mathrm{cond}_\phi(\mathbf{d}) \cdot eps$$

gdzie stałą K_s nazywamy **wskaźnikiem stabilności**.

Z definicji powyższej wynika bezpośrednio
$$\lim_{eps \to 0} \frac{\| fl(A(\mathbf{d})) - \phi(\mathbf{d}) \|}{\|\phi(\mathbf{d})\|} = 0$$

$$\frac{\|fl\left(A\left(\mathbf{d}\right)\right)-\phi\left(\mathbf{d}\right)\|}{\|\phi\left(\mathbf{d}\right)\|}\leq K_s\cdot\mathrm{cond}_\phi\left(\mathbf{d}\right)\cdot eps$$

Całkowity błąd względny wyniku uzyskiwanego algorytmem numerycznie stabilnym zależy od:

– uwarunkowania zadania ($\mathrm{cond}_\phi(\mathbf{d})$),

– stosowanej arytmetyki zmiennopozycyjnej (eps),

– wskaźnika stabilności numerycznej algorytmu (K_s), tj. jakości algorytmu.

Metoda pozornych równoważnych zaburzeń: polega na wykazaniu, że $fl(A(\mathbf{d}))$ jest zaburzonym dokładnym rozwiązaniem zadania ϕ o zaburzonych danych, tj.

$$fl\left(A\left(\mathbf{d}\right)\right)=\phi\left(\mathbf{d}+\mathbf{\Delta d}\right)\cdot\left(1+\eta\right),$$

gdzie błędy względne danych i wyniku spełniają:

$$\frac{|\Delta d_i|}{|d_i|}\leq k_i\cdot eps,\quad |\eta_j|\leq k_j\cdot eps,$$

zaś k_i i k_j to stałe (o niewielkich wartościach).