BŁĘDY PRZETWARZANIA NUMERYCZNEGO

Maciej Patan

Instytut Sterowania i Systemów Informatycznych
Uniwersytet Zielonogórski

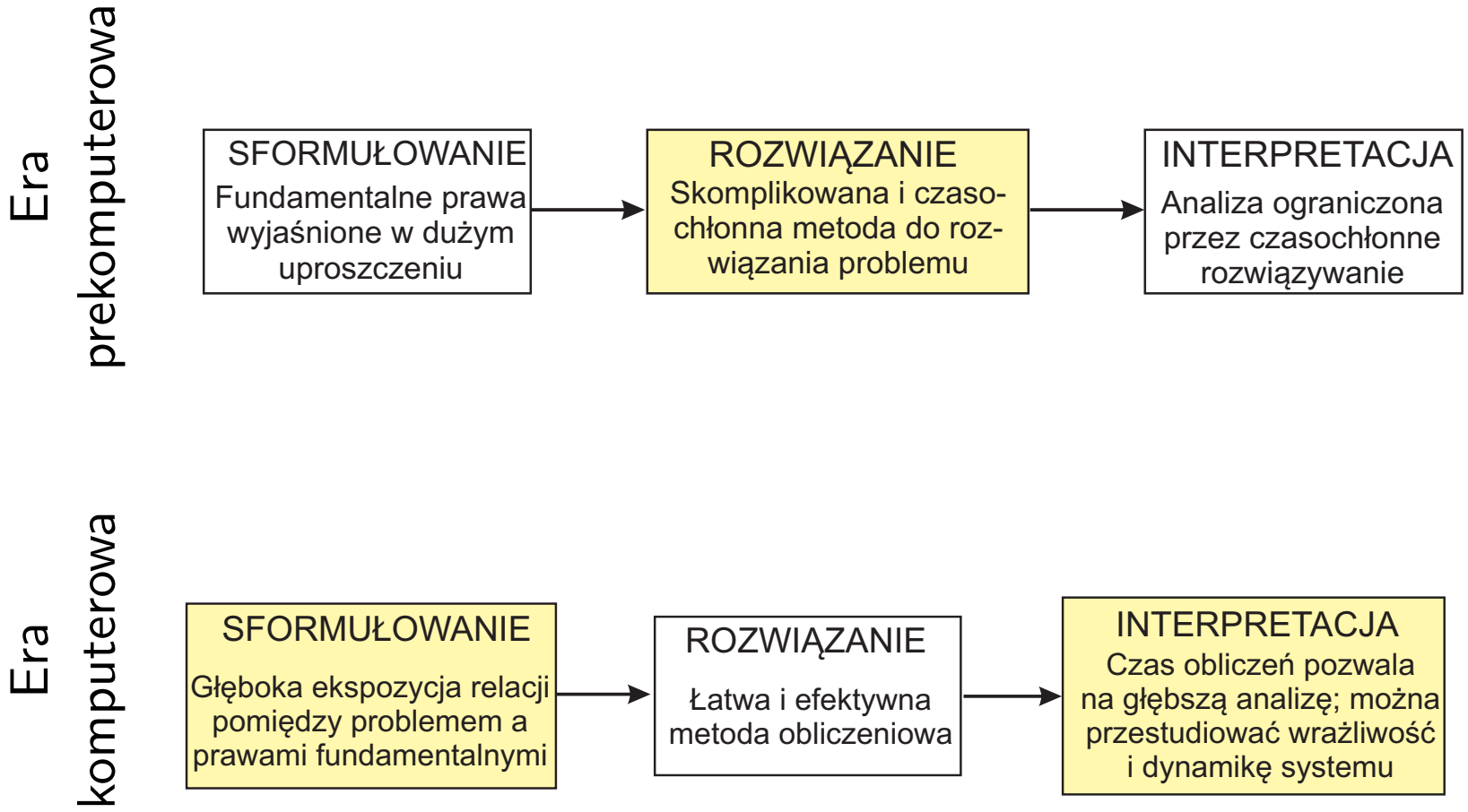
Dlaczego modelujemy ...

- systematyczne rozwiązywanie problemów,
- eksperymentalna eksploracja wielu rozwiązań,
- dostarczanie abstrakcyjnych metod zarządzania złożonością,
- redukcja czasu wdrożenia dla aplikacji biznesowych,
- zmniejszenie kosztów produkcji,
- zarządzanie ryzykiem błędu.

. . . numerycznie?

- szybkie i efektywne narzędzia rozwiązywania problemów,
- uniwersalność i szeroka użyteczność,
- często, jedyna alternatywa dla nieistniejących rozwiązań analitycznych,
- liczne istniejące efektywne programy i biblioteki,
- idealne do nauki obsługi maszyn cyfrowych,
- zmieniają rozumienie metod matematycznych (redukcja skomplikowanych technik do podstawowych operacji arytmetycznych)

Inżynierskie rozwiązywanie problemów



Podstawowe pojęcia

Metody numeryczne – rozwiązywanie zadań matematycznych w skończonej liczbie operacji arytmetyczno-logicznych,

Zadanie numeryczne – matematyczny opis relacji pomiędzy danymi wejściowymi i wyjściowymi,

Algorytm numeryczny – skończona sekwencja operacji przekształcających dane wejściowe w wyjściowe, przy czym operacja jest rozumiana jako funkcja arytmetyczna lub logiczna albo referencja do innych istniejących algorytmów.

Źródła błędów

1. Modelowanie matematyczne, np. najprostszy model przyrostu populacji
   $N(t) = N_0 e^{kt}$ jest prawidłowy tylko przy założeniu, że posiada nieograniczone zasoby,
2. Błędy grube i pomyłki, np. błędy programistyczne
3. Błędy pomiarowe, np. prędkość światła w próżni wynosi
   $c = (2.997925 + \varepsilon) \cdot 10^8 \text{m/sec}, \quad |\varepsilon| \leqslant 3 \cdot 10^{-6}$
4. Błędy maszynowe, np. błędy zaokrąglania lub odrzucania w arytmetyce zmiennoprzecinkowej
5. Błędy przybliżania matematycznego, np. obliczając całkę

$$I = \int_0^1 e^{x^2}\,\mathrm{d}x \approx \int_0^1 \left(1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!}\right)\mathrm{d}x$$

Błąd względny i bezwzględny

Błąd bezwzględny

$$E_{\mathrm{abs}} = \alpha - \hat{\alpha}$$

gdzie:

α – wartość dokładna

$\hat{\alpha}$ – wartość przybliżona (obliczona)

Błąd względny

$$E_{\mathrm{rel}} = \frac{\alpha - \hat{\alpha}}{\alpha} \quad \text{lub} \quad E_{\mathrm{rel}} = \frac{\alpha - \hat{\alpha}}{\alpha} \cdot 100\%$$

Wniosek: α jest najczęściej niedostępna *a priori*, stąd

$$E_{\mathrm{rel}} \cong \frac{\alpha - \hat{\alpha}}{\hat{\alpha}} \quad \text{lub nawet} \quad E_{\mathrm{rel}} \cong \frac{\tilde{\alpha} - \hat{\alpha}}{\tilde{\alpha}}$$

Przykład 1 Rozważmy zadanie pomiaru długości mostu oraz nita użytego do jego budowy. W wyniku pomiaru otrzymano odpowiednio 9999 i 9 cm. Jeżeli prawdziwe wartości to 10000 oraz 10 cm, policzyć (a) błąd bezwzględny i (b) procentowy błąd względny w każdym z przypadków.

a) Błąd bezwzględny dla długości mostu wynosi

$$E^m_{\text{abs}} = 10000 - 9999 = 1 \text{ cm}$$

oraz dla długości nita

$$E^n_{\text{abs}} = 10 - 9 = 1 \text{ cm}$$

b) Błąd względny dla długości mostu wynosi

$$E^m_{\text{rel}} = \frac{1}{10000} \cdot 100\% = 0.01\%$$

oraz dla nita

$$E^n_{\text{rel}} = \frac{1}{10} \cdot 100\% = 10\%$$

Cyfrowy zapis liczb całkowitych

baza 10	konwersja	baza 2	baza 8	baza 16
1	$1{=}2^0$	00000001	001	01
8	$8{=}2^3$	00001000	010	08
10	$8+2{=}2^3+2^1$	00001010	012	0C
27	$16+8+2+1{=}2^4+2^3+2^1$	00011010	032	1C
67	$64+2+1{=}2^6+2^1+2^0$	01000011	103	43
202	$128+64+8+2{=}2^8+2^7+2^3+2^1$	11001010	312	CA

Postać stałoprzecinkowa

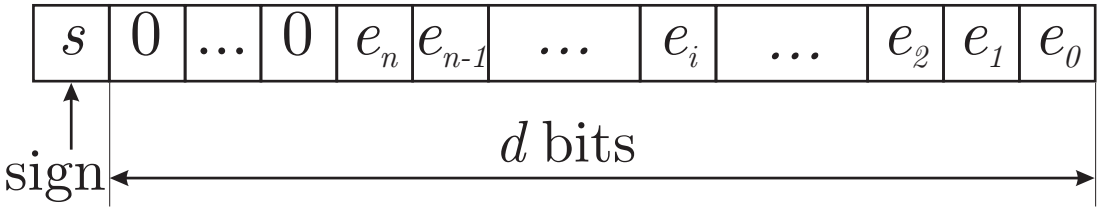
Liczba całkowita może być dokładnie reprezentowana w systemie dwójkowym jako

$$l = s \sum_{i=0}^{n} e_i 2^i$$

gdzie:

$s \in \{-1, 1\}$ – znak

$e_i \in \{0, 1\}$, $i = 0, \ldots, n-1$ oraz $e_n = 1$ – cyfry reprezentacji binarnej



Przykład 2 Określić zakres liczb całkowitych przy, które mogą być reprezentowane na 16-bitowym komputerze.

Pierwszy bit przechowuje znak. Pozostałe 15 bitów może służyć do przechowania liczb binarnych od 0 do 111111111111111. Stąd, górna granica wynosi

$$2^{14} + 2^{13} + \cdots + 2^1 + 2^0 = 2^{15} - 1 = 32767$$

W ten sposób zakres to [-32 767,32 767]. Ale istnieje nadmiarowość dla wartości zero, tj. 0000000000000000 oraz 1000000000000000, dlatego druga z kombinacji zazwyczaj jest używana do reprezentacji dodatkowej liczby ujemnej: -32768, stąd prawdziwy zakres jest od -32768 do 32767.

Wniosek: W ogólności, dla d cyfr zakres wynosi $[-2^d, 2^d - 1]$

Postać zmiennoprzecinkowa

Numeryczne wartości z niezerową częścią ułamkową są przechowywane jako liczby zmiennoprzecinkowe.

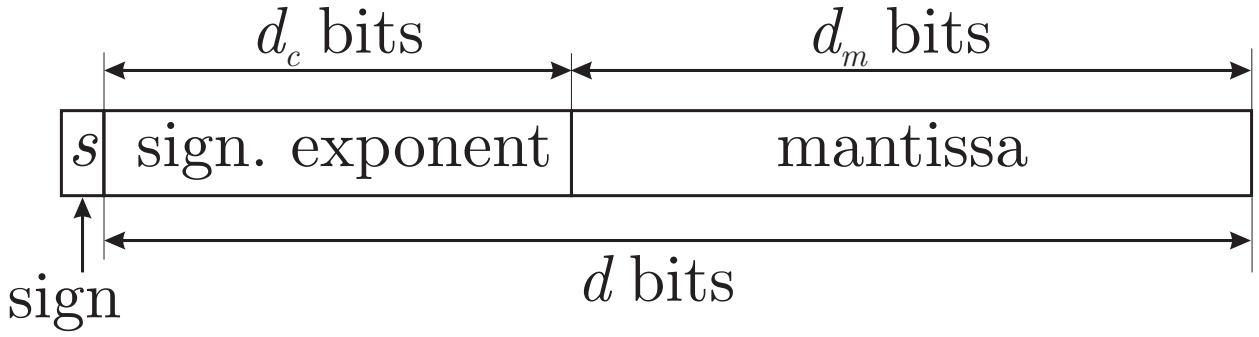
$$l = s \cdot 2^c \cdot m$$

gdzie:

$s \in \{-1, 1\}$ – znak

c – cecha (wykładnik) (d_c-bitowa liczba całkowita)

m – mantysa (d_m-bitowa liczba całkowita)



Cyfrowy zapis liczb zmiennoprzecinkowych

Standardy IEEE formatów liczb zmiennoprzecinkowych

Precyzja	Sign	Bity mantysy	Bity cechy	Suma
Single	1	23	8	32
Double	1	52	11	64
Quadruple	1	113	14	128

- d_m decyduje o precyzji reprezentacji,
- $d_c = d - d_m$ decyduje o zakresie reprezentacji,
- znormalizowana notacja naukowa, tzn. bez nieznaczących zer, np.

$$123.456 \longrightarrow \underbrace{0.123456}_{\text{mantissa}} \cdot 10^3,$$

czyli mantysa zawsze należy do przedziału $[\frac{1}{2}, 1)$.

Przykład 3 Utworzyć hipotetyczny zbiór liczb zmiennoprzecinkowych, które przechowują informację o 8 bitowych słowach (1 bit na znak, 3 na cechę ze znakiem i pozostałe 4 na mantysę).
Zakres cechy wynosi $[-2^2, 2^2 - 1]$. Zatem, najmniejsza wartość dodatnia to:
0100100=$2^{-1} \cdot 2^{-4} = 0.03125$, a największa wartość dodatnia to 0111111=$(2^{-1} + 2^{-2} + 2^{-3}) \cdot 2^3 = 7$. Czyli, zakres rozważanego systemu to

$$-7 \leqslant x \leqslant -0.03125 \text{ lub } x = 0 \text{ lub } 0.03125 \leqslant x \leqslant 7$$

Wniosek: W ogólności mamy:

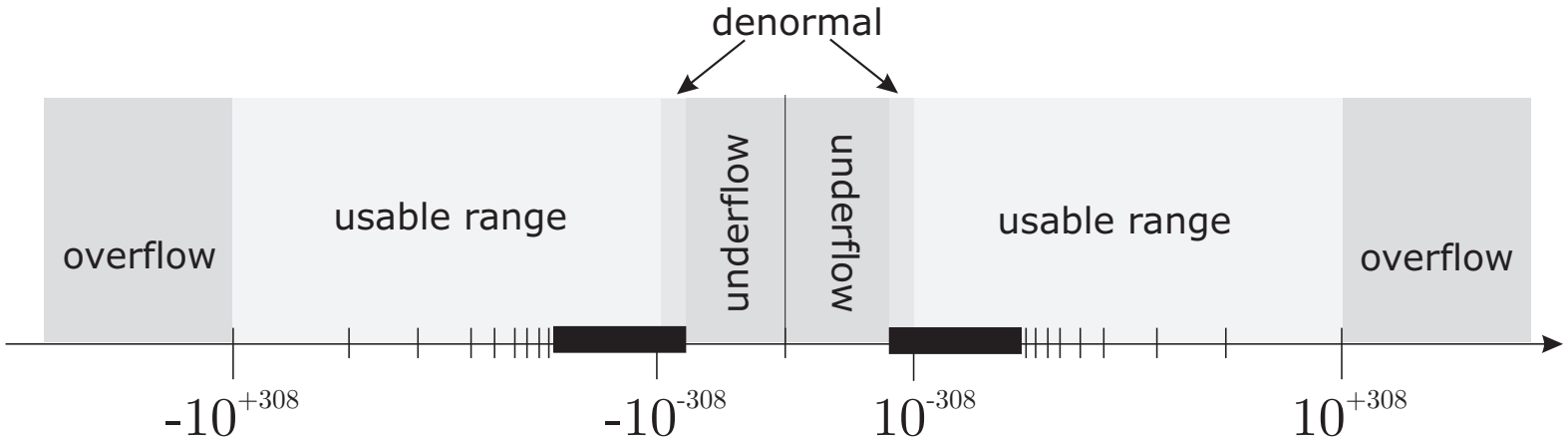
$$\boxed{\begin{array}{l} m_{\min} \cdot 2^{c_{\min}} \leqslant |x| \leqslant m_{\max} \cdot 2^{c_{\max}} \Rightarrow \\ \frac{1}{2} \cdot 2^{-2^{d_c - 1}} \leqslant |x| \leqslant (1 - 2^{-d_m}) \cdot 2^{2^{d_c - 1} - 1} \end{array}}$$

Zbiór liczb zmiennoprzecinkowych

	Rzeczywiste	Zmiennoprzecinkowe
Zakres	Nieskończony: istnieją dowolnie duże i dowolnie małe liczby rzeczywiste.	Skończony: liczba bitów cechy ogranicza amplitudę liczb zmiennoprzecinkowych.
Precyzja	Nieskończona: jest nieskończenie wiele liczb rzeczywistych pomiędzy dwoma dowolnymi l. rzeczywistymi.	Skończona: istnieje skończona liczba (czasami nawet równa zero) liczb zmiennoprzecinkowych pomiędzy dwoma dowolnymi l. zmiennoprzecinkowymi.

Wniosek: Linia liczb zmiennoprzecinkowych jest podzbiorem linii liczb rzeczywistych.

Linia liczb zmiennoprzecinkowych



Arytmetyka zmiennoprzecinkowa

Wyniki operacji arytmetycznych pomiędzy dwoma liczbami zmiennoprzecinkowymi:

- najczęściej nie może być reprezentowana przez inną wartość zmiennoprzecinkową,
- ma ograniczony zakres i precyzję.

Błędy zaokrąglenia w obliczeniach (1)

Przykład 4 Obliczyć $r = x^2 - y^2$, gdzie $x = 4.005$ oraz $y = 4.004$ z 4-cyfrową precyzją.

Przez bezpośrednie podstawienie mamy

$$r = x^2 - y^2 = 16.04(0025) - 16.03(2016) = 0.01$$

Prawdziwa wartość r to 0.008009. Zatem błąd względny wynosi

$$\boxed{E_{\text{rel}} = \frac{0.008009 - 0.01}{0.008009} \cdot 100\% \cong -24.859\%!!!}$$

Z drugiej strony, stosując znany wzór skróconego mnożenia $r = (x - y)(x + y)$ mamy:

$$r = (4.005 - 4.004)(4.004 + 4.005) = 0.001 \cdot 8.009 = 0.008009$$

Wynik jest dokładny i błąd względny

$$\boxed{E_{\text{rel}} = 0\%!!!}$$

Błędy zaokrąglenia w obliczeniach (2)

- Ograniczona precyzja prowadzi do zaokrągleń w pojedynczych kalkulacjach
- Efekty zaokrągleń akumulują się powoli
- Błędy zaokrągleń są nieuniknione, sposobem jest tworzenie lepszych algorytmów
- Odejmowanie niemal równych wartości prowadzi do poważnych strat precyzji

Błędy kasowania (1)

Dla dodawania: Błędy w

$$c = a + b \quad \text{and} \quad c = a - b$$

będą duże kiedy $a \ll b$ lub $a \gg b$.
Przykład 5 Rozważmy $c = a + b$ z
$a = \mathtt{x.xxx}\ldots \cdot 10^0, b = \mathtt{y.yyy}\ldots \cdot 10^{-8}$ oraz $z = x + y < 10$.

$$\begin{array}{cl}
 & \overbrace{\mathtt{x.xxx\ xxxx\ xxxx\ xxxx}}^{\text{osiągalna precyzja}} \\
+ & \mathtt{0.000\ 0000\ yyyy\ yyyy\ yyyy\ yyyy} \\
\hline
= & \mathtt{x.xxx\ xxxx\ zzzz\ zzzz\ \underbrace{yyyy\ yyyy}_{\text{stracone cyfry}}}
\end{array}$$

Błędy kasowania (2)

Dla odejmowania: Błąd w $c = a - b$ będzie znaczny dla $a \approx b$.

Przykład 6 Wyznaczyć $c = a - b$ w arytmetyce zmiennoprzecinkowej dla $a = \texttt{x.xxxxxxxxxxxxx1}$ oraz $b = \texttt{x.xxxxxxxxxxxxxx0}$.

$$
\begin{array}{rl}
 & \overbrace{\texttt{x.xxx xxxx xxxx xxx1}}^{\text{osiągalna precyzja}} \\
- & \texttt{x.xxx xxxx xxxx xxx0} \\
\hline
= & \texttt{0.000 0000 0000 0001 } \underbrace{\texttt{uuuu uuuu uuuu uuu}}_{\text{nieokreślone cyfry}} \\
= & \texttt{1.uuu uuuu uuuu uuuu } \cdot 10^{-15}
\end{array}
$$

Wynik posiada tylko jedną (!) cyfrę znaczącą.

Błędy kasowania (3)

Podsumowanie

- występują w dodawaniu: $a + b$ kiedy $a \gg b$ lub $a \ll b$,
- występują w odejmowaniu: $a - b$ kiedy $a \approx b$,
- ogromne błędy w pojedynczych operacjach, a nie powolna akumulacja,
- często mogą zostać zminimalizowane przez algebraiczne przekształcenie wrażliwej formuły.

Precyzja maszyny

Amplituda błędów zaokrąglania jest określona przez tzw. precyzję maszyny ε_m, tzn.

$$1 + \delta = 1, \quad \forall \delta \leqslant \varepsilon_m$$

Dla podwójnej precyzji (64 bity) $\varepsilon_m \cong 2.2204 \cdot 10^{-16}$

Wyznaczanie precyzji maszyny

```
eps=1
do
  if (eps+1 <= 1) exit
  eps = eps/2
end do
eps=2*eps
```

Porównywanie liczb zmiennoprzec.

```
if x==y % błędnie!!!

if abs(x-y) < eps
```

Błąd obcięcia

Rozważmy rozwinięcie w szereg $\sin(x)$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots$$

Dla małych x, tylko kilka wyrazów szeregu jest potrzebnych do dokładnego przybliżenia $\sin(x)$. Wyrazy wyższego rzędu są obcinane (ang. *truncated*)

$$f_{\text{true}} = f_{\text{sum}} + \text{błąd obcięcia}$$

Rozmiar błędu obcięcia zależy od x oraz liczby wyrazów włączonych w f_{sum}.

Szereg Taylora (1)

Dla odpowiednio ciągłej funkcji $f(x)$ określonej na przedziale $x \in [a, b]$ defniniujemy wyraz n-tego rzędu szeregu Taylora $P_n(x)$ jako

$$f(x) = f(x_0) + (x - x_0) \left.\frac{df}{dx}\right|_{x=x_0} + \frac{(x - x_0)^2}{2!} \left.\frac{d^2 f}{dx^2}\right|_{x=x_0} + \ldots + \frac{(x - x_0)^n}{n!} \left.\frac{d^n f}{dx^n}\right|_{x=x_0} + R_n(x)$$

gdzie istnieje wartość ξ spełniająca $x_0 \leqslant \xi \leqslant x$ taka, że

$$R_n(x) = \frac{(x - x_0)^{n+1}}{(n+1)!} \left.\frac{d^{(n)} f}{dx^{n+1}}\right|_{x=\xi}$$

Szereg Taylora(2)

Przykład 7 Użyć rozwinięcia w szereg Taylora $n = 4$ rzędu do przybliżenia $f(x) = \cos(x)$ w punkcie $x_1 = \pi/6$ na podstawie wartości $f(x)$ i jej pochodnych w $x_0 = 0$.

Przybliżenie 4-go rzędu

$$\cos(\pi/6) \cong 1 - \frac{(\pi/6)^2}{2} + \frac{(\pi/6)^4}{24} = 0.8660538$$

Wartość dokładna to $\sqrt{3}/2 = 0.8660254$.

Szereg Taylora(3)