

## Wykład 2.

### *Analiza błędów w metodach numerycznych*

## Po co wprowadzamy liczby w formacie zmiennoprzecinkowym (floating point)?

- Przykład 1. W jaki sposób można zapisać liczbę 256.78 na 5-ciu miejscach?

2	5	6	.	7	8
---	---	---	---	---	---

Jak można zapisać najmniejszą liczbę w tym formacie?

0	0	0	.	0	0
---	---	---	---	---	---

Jak można zapisać największą liczbę w tym formacie?

9	9	9	.	9	9
---	---	---	---	---	---

## Po co wprowadzamy liczby w formacie zmiennoprzecinkowym (floating point)?

- Przykład 2. W jaki sposób można zapisać liczbę 256.786 na 5-ciu miejscach?

zaokrąglenie (rounded off)

2	5	6	.	7	9
---	---	---	---	---	---

urwanie (chopped)

2	5	6	.	7	8
---	---	---	---	---	---

Wniosek: Błąd jest mniejszy niż 0.01

## Jaki błąd popełniamy?

Błąd bezwzględny  $|x - x_o|$

  
wielkość dokładna lub rzeczywista  $x_o$

Błąd względny  $\frac{|x - x_o|}{x_o}$

Obliczenia:

$$\varepsilon_t = \frac{|x - x_o|}{x_o} \times 100\% = \frac{|256.79 - 256.786|}{256.786} \times 100\% = 0.001558\%$$

## Jaki błąd popełniamy?

**Względne błędy wielkości małych są duże.**

Porównajmy:

$$\varepsilon_t = \frac{|x - x_o|}{x_o} \times 100\% = \frac{|256.79 - 256.786|}{256.786} \times 100\% = 0.001558\%$$

$$\varepsilon_t = \frac{|x - x_o|}{x_o} \times 100\% = \frac{|3.55 - 3.546|}{3.546} \times 100\% = 0.11280\%$$

Błędy bezwzględne są jednakowe:

$$|x - x_o| = |256.786 - 256.79| = |3.546 - 3.55| = 0.004$$

## Jak utrzymać błędy względne na podobnym poziomie?

Można przedstawić liczbę w postaci:

$$\text{znak} \times \text{mantysa} \times 10^{\text{wykład}}$$

lub

$$\text{znak} \times \text{mantysa} \times 2^{\text{wykład}}$$

czyli

256.78 zapisujemy jako  $+ 2.5678 \times 10^2$

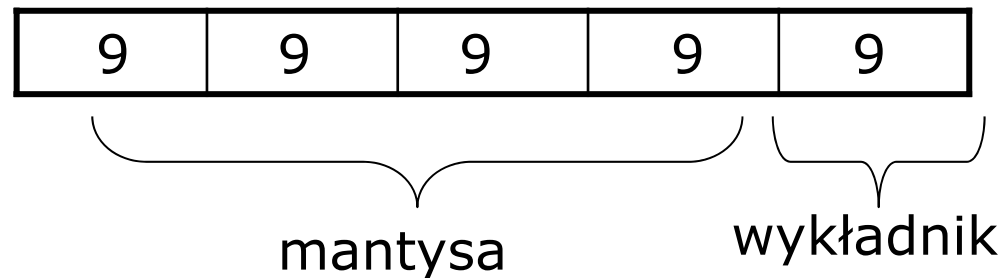
0.003678 zapisujemy jako  $+ 3.678 \times 10^{-3}$

$- 256.78$  zapisujemy jako  $- 2.5678 \times 10^2$

## Co zyskujemy stosując zapis zmiennoprzecinkowy?

### Zwiększa się zakres liczb, które możemy zapisać

Jeżeli użyjemy tylko 5 miejsc do zapisu liczby (dodatniej o dodatnim wykładniku) to najmniejsza liczba zapisana to 1 a największa  $9.999 \cdot 10^9$ .



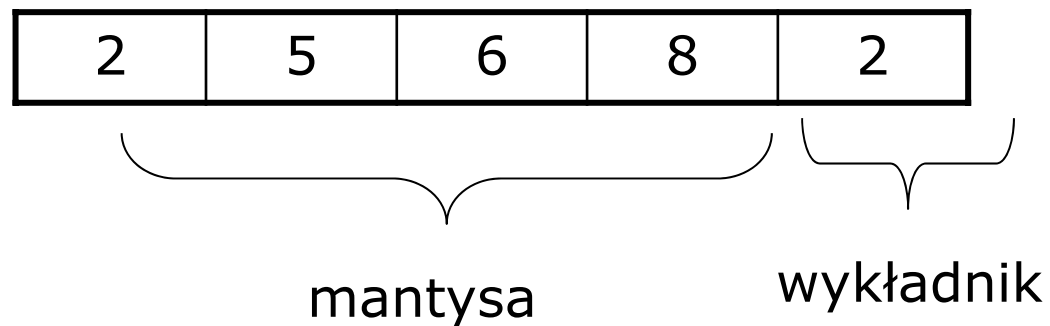
Zakres możliwych do zapisania liczb zwiększył się od 999.99 do  $9.999 \cdot 10^9$ .

## Co tracimy stosując zapis zmiennoprzecinkowy?

### Dokładność (precyzję).

Dlaczego?

Liczba 256.78 będzie przedstawiona jako  $2.5678 \cdot 10^2$  i na pięciu miejscach:

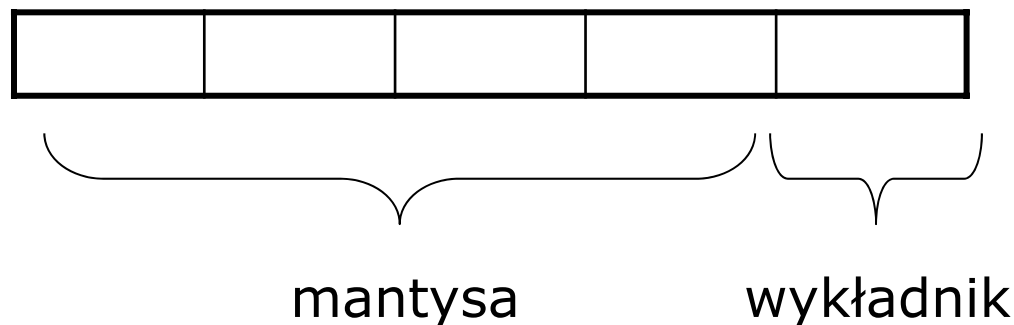


Wystąpi błąd zaokrąglenia.



## Przykład do samodzielnego rozwiązania

1. Proszę przedstawić liczbę 576329.78 na pięciu miejscach w podobny sposób jak w poprzednim przykładzie:

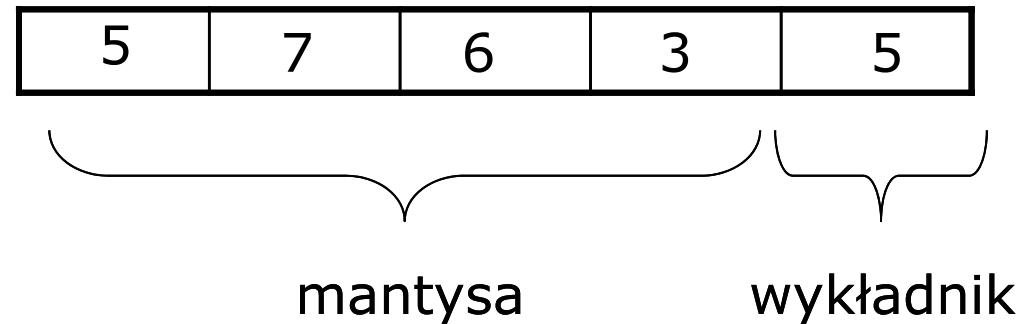


2. Proszę oszacować błąd bezwzględny i względny zaokrąglenia

3. Porównać z przykładem poprzednim (256.78) i wyciągnąć wnioski

## Rozwiązanie przykładu do samodzielnego rozwiązania

1. Liczba 576329.78 zapisana na pięciu miejscach:



2. Błąd bezwzględny przybliżenia wynosi 29.78 a względny 0.0051672%

3. Dla liczby 256.78 te błędy wynoszą odpowiednio: 0.02 (mniejszy) i 0.0077888% (porównywalny)

# Arytmetyka zmiennoprzecinkowa- system dziesiętny

Postać liczby

$$\sigma \times m \times 10^e$$

wykładnik będący  
liczbą całkowitą

znak liczby (-1 lub +1)

mantysa  $(1)_{10} \leq m < (10)_{10}$

Przykład

$$-2.5678 \times 10^2$$

$$\sigma = -1$$

$$m = 2.5678$$

$$e = 2$$



# Arytmetyka zmiennoprzecinkowa- systemem dwójkowy

Postać liczby

$$\sigma \times m \times 2^e$$

wykładnik będący  
liczbą całkowitą

znak liczby (0 dla  
dodatniej lub 1 dla  
ujemnej liczby)

mantysa  $(1)_2 \leq m < (2)_2$

Przykład

$$(1.1011011)_2 \times 2^{(101)_2}$$

$$\sigma = 0$$

$$m = 1011011$$

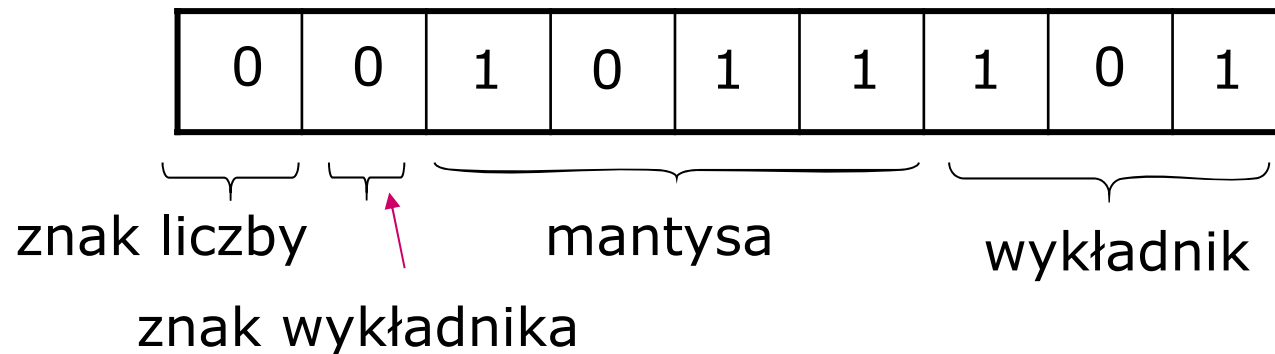
$$e = 101$$

1 nie jest zapisywane

## Przykład do samodzielnego rozwiązania

Mamy słowo 9-bitowe

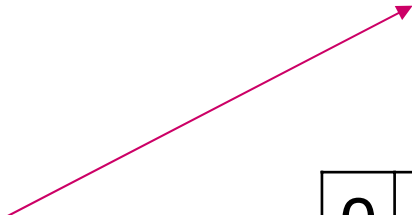
- pierwszy bit odpowiada znakowi liczby,
- drugi bit – znakowi wykładnika,
- następne cztery bity kodują mantysę,
- ostatnie trzy bity zapisują wykładnik



Znajdź liczbę (w postaci dziesiętnej), która jest przedstawiona w podany sposób.

## Odpowiedź

$$(54.75)_{10} = (110110.11)_2 = (1.1011011)_2 \times 2^5 \\ \cong (1.1011)_2 \times (101)_2$$



0	0	1	0	1	1	1	0	1
---	---	---	---	---	---	---	---	---

nie jest  
zapisywane

$(54)_{10}$

## Co to jest $\varepsilon$ maszyny cyfrowej?

Dla każdej maszyny cyfrowej definiuje się parametr epsilon  $\varepsilon$  określający dokładność obliczeń:

$$\varepsilon = N^{-t}$$

gdzie:  $N=2$  (w zapisie dwójkowym),  $N=10$  (w zapisie dziesiętnym),  $t$  jest liczbą bitów w mantysie liczby

$\varepsilon$  jest tym mniejsze im więcej bitów przeznaczono na reprezentowanie mantysy  $M$

Epsilon  $\varepsilon$  można traktować jako parametr charakteryzujący dokładność obliczeniową maszyny (im mniejsze  $\varepsilon$  tym większa dokładność).

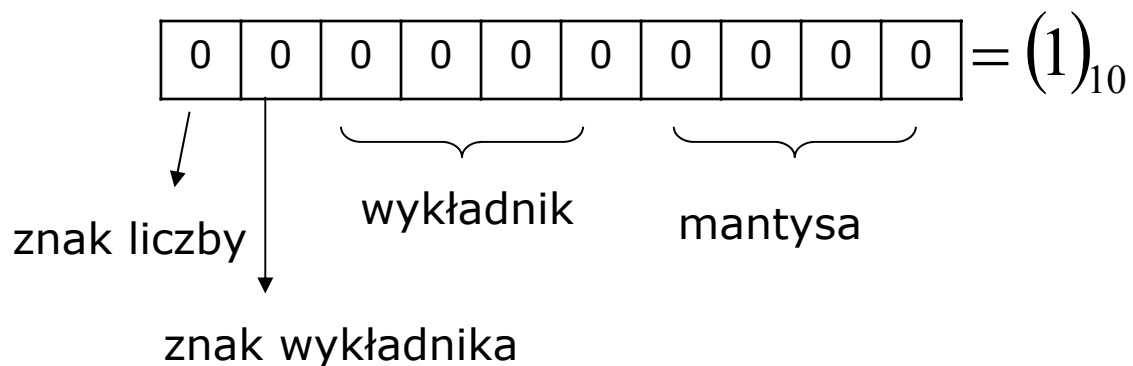
Podwójna precyzja (Fortran)  $\varepsilon_{DP} = \varepsilon^2$

## Co to jest $\epsilon$ maszyny cyfrowej?

Epsilon  $\epsilon$  jest to najmniejsza liczba, która po dodaniu do 1.000 produkuje liczbę, którą można przedstawić jako różną od 1.000.

Przykład: słowo dziesięciobitowe

$$x = M \times N^w$$



następna liczba

0	0	0	0	0	0	0	0	0	1
---	---	---	---	---	---	---	---	---	---

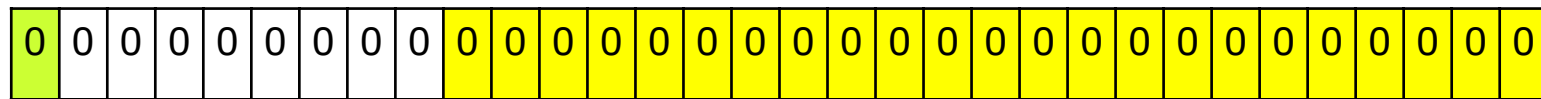
$= (1.0001)_2 = (1.0625)_{10}$

$$\epsilon_{mach} = 1.0625 - 1 = 2^{-4}$$



## Pojedyncza precyzja w formacie IEEE-754 (Institute of Electrical and Electronics Engineers)

32 bity dla pojedynczej precyzji



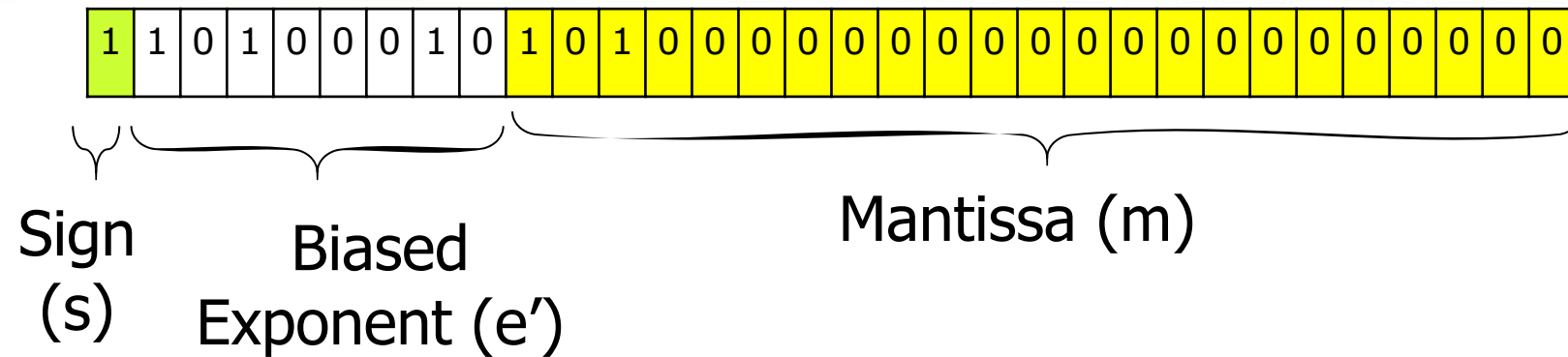
znak  
(s)

wykładnik  
(e')

mantysa (m)

$$Liczba = (-1)^s \times (1.m)_2 \times 2^{e'-127}$$

## Przykład



$$\begin{aligned}
 \text{Value} &= (-1)^s \times (1.m)_2 \times 2^{e'-127} \\
 &= (-1)^1 \times (1.10100000)_2 \times 2^{(10100010)_2 - 127} \\
 &= (-1) \times (1.625) \times 2^{162-127} \\
 &= (-1) \times (1.625) \times 2^{35} = -5.5834 \times 10^{10}
 \end{aligned}$$

8 bitów wykładnika oznacza  $0 \leq e' \leq 255$

Ustalone przesunięcie wykładnika wynosi 127 a  
zatem

$$-127 \leq e \leq 128$$

W istocie  $1 \leq e' \leq 254$

Liczby  $e' = 0$  i  $e' = 255$  są zarezerwowane dla  
przypadków specjalnych

Zakres wykładnika  $-126 \leq e \leq 127$

## Reprezentacja liczb specjalnych

$e' = 0$  — same zera

$e' = 255$  — same jedyнки

s	$e'$	m	Reprezentuje
0	same zera	same zera	0
1	same zera	same zera	-0
0	same jedyнки	same zera	$\infty$
1	same jedyнки	same zera	$-\infty$
0 lub 1	same jedyнки	różne od zera	NaN

## Format IEEE-754

Największa liczba

$$(1.1\dots\dots 1)_2 \times 2^{127} = 3.40 \times 10^{38}$$

Najmniejsza liczba

$$(1.00\dots\dots 0)_2 \times 2^{-126} = 2.18 \times 10^{-38}$$

Epsilon maszyny cyfrowej

$$\varepsilon_{mach} = 2^{-23} = 1.19 \times 10^{-7}$$

Jeżeli nie znamy wielkości dokładnej  $x_0$  możemy obliczać błąd bezwzględny przybliżenia (ang. approximate error) jako różnicę wartości uzyskanych w kolejnych przybliżeniach :

$$x_n - x_{n-1}$$

Błąd względny  $\varepsilon_a$  :

$$\varepsilon_a = \frac{x_n - x_{n-1}}{x_n}$$

## Przykład

Dla  $f(x) = 7e^{0.5x}$  w  $x = 2$  znajdź

a)  $f'(2)$  dla  $h = 0.3$

b)  $f'(2)$  dla  $h = 0.15$

c) błąd przybliżenia

## Rozwiązanie

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

a)  $h = 0.3$

$$f'(2) \approx \frac{f(2+0,3) - f(2)}{0,3} = \frac{7e^{0,5(2,3)} - 7e^{0,5(2)}}{0,3} = 10,265$$

## Przykład (cd)

b)  $h = 0.15$

$$f'(2) \approx \frac{f(2 + 0,15) - f(2)}{0,15} = \frac{7e^{0,5(2,15)} - 7e^{0,5(2)}}{0,15} = 9,880$$

c)  $\varepsilon_a = \frac{x_n - x_{n-1}}{x_n}$

$$\varepsilon_a = \frac{9,880 - 10,265}{9,8800} \approx -0,0389$$

Błąd procentowy 3,89%



## Błąd względny jako kryterium zakończenia procedury iteracyjnej

Jeżeli błąd względny jest mniejszy lub równy od pewnej określonej wcześniej liczby to dalsze iteracje nie są konieczne

$$|\varepsilon_a| \leq \varepsilon_s$$

Jeżeli wymagamy przynajmniej  $m$  cyfr znaczących w wyniku to

$$|\varepsilon_a| \leq 0.5 \times 10^{2-m}$$

## Podsumowanie przykładu

$$|\varepsilon_a| \leq 0.5 \times 10^{2-m}$$

$h$	$f'(2)$	$ \varepsilon_a $	$m$
0.3	10.265	N/A	0
0.15	9.8800	0.03894	1
0.10	9.7559	0.01271	1
0.01	9.5378	0.02286	1
0.001	9.5164	0.00225	2

Wartość dokładna 9.514

## Źródła błędów w obliczeniach numerycznych

1. Błędy wejściowe (błędy danych wejściowych)
2. Błędy obcięcia (ang. truncation error)
3. Błędy zaokrągleń (ang. round off error)

**Błędy wejściowe** występują wówczas gdy dane wejściowe wprowadzone do pamięci komputera odbiegają od dokładnych wartości tych danych.

**Błędy obcięcia** są to błędy wynikające z procedur numerycznych przy zmniejszaniu liczby działań.

**Błędy zaokrągleń** są to błędy, których na ogół nie da się uniknąć. Powstają w trakcie obliczeń i można je zmniejszać ustalając umiejętnie sposób i kolejność wykonywania zadań.

Źródła błędów wejściowych:

- dane wejściowe są wynikiem pomiarów wielkości fizycznych
- skończona długość słów binarnych i konieczność wstępnego zaokrąglania
- wstępne zaokrąglanie liczb niewymiernych

**Przybliżanie liczb**, których nie można wyrazić dokładnie dokonuje się poprzez:

- urywanie (ang. chopping)
- zaokrąglanie (ang. rounding)

## Przykład:

$$\pi \approx 3,14159265359$$

$$\pi \approx 3,1415$$

urywanie

$$\pi \approx 3,1416$$

zaokrąglanie

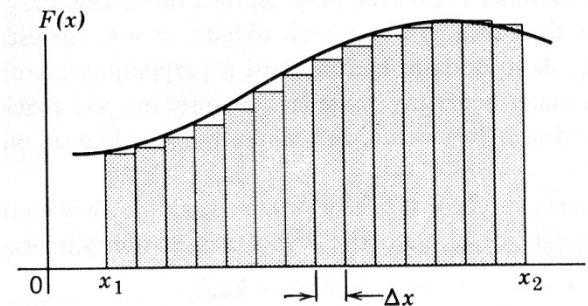
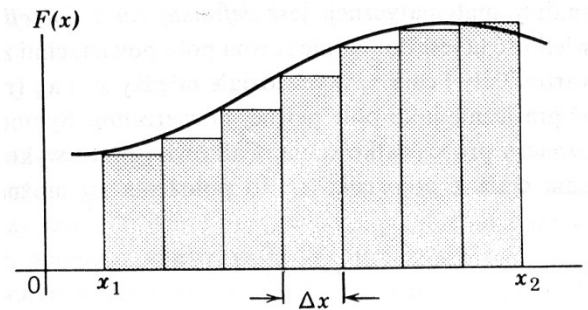
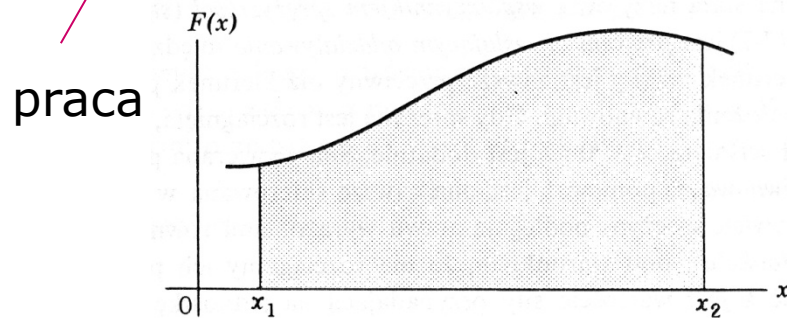
Zaokrąglanie prowadzi do mniejszego błędu niż urywanie.

## Błąd obcięcia

Spowodowany jest użyciem przybliżonej formuły zamiast pełnej operacji matematycznej:

- przy obliczaniu sum nieskończonych szeregów
- przy obliczaniu wielkości będących granicami (całka, pochodna)

$$W = \lim_{\Delta x \rightarrow 0} \sum_{x_1}^{x_2} F \Delta x = \int_{x_1}^{x_2} F dx$$



## Szereg Taylora

Jeżeli funkcja jest ciągła i wszystkie pochodne  $f'$ ,  $f''$ , ...,  $f^n$  istnieją w przedziale  $[x, x+h]$  to wartość funkcji w punkcie  $x+h$  można obliczyć jako:

$$f(x+h) = f(x) + f'(x)h + \frac{f''(x)}{2!}h^2 + \frac{f'''(x)}{3!}h^3 + \dots$$

Szereg Maclaurina jest to rozwinięcie wokół  $x=0$

$$f(0+h) = f(0) + f'(0)h + f''(0)\frac{h^2}{2!} + f'''(0)\frac{h^3}{3!} + \dots +$$

Typowe rozwinięcia w szereg wokół zera

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$



## Błąd obcięcia w szeregu Taylora

$$f(x+h) = f(x) + f'(x)h + f''(x)\frac{h^2}{2!} + \dots + f^{(n)}(x)\frac{h^n}{n!} + R_n(x)$$

reszta



$$R_n(x) = \frac{(x-h)^{n+1}}{(n+1)!} f^{(n+1)}(c)$$

$$x < c < x + h$$

## Przykład

Rozwinięcie w szereg  $e^x$  wokół  $x=0$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \dots$$

Im większa ilość wyrazów jest uwzględniana w rozwinięciu, tym błąd obcięcia jest mniejszy i możemy znaleźć tym dokładniejszą wartość wyrażenia

**Pytanie:** Ile należy uwzględnić wyrazów aby otrzymać przybliżoną wartość liczby  $e$  z błędem mniejszym niż  $10^{-6}$ ?

$$e^1 = 1 + 1 + \frac{1^2}{2!} + \frac{1^3}{3!} + \frac{1^4}{4!} + \frac{1^5}{5!} + \dots$$
$$\approx 2 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120}$$

$$x = 0, h = 1, f(x) = e^x \quad R_n(x) = \frac{h^{n+1}}{(n+1)!} f^{(n+1)}(c)$$

$$R_n(0) = \frac{1^{n+1}}{(n+1)!} f^{(n+1)}(c)$$

$$= \frac{(1)^{n+1}}{(n+1)!} e^c$$

ale  $x < c < x + h$

$$0 < c < 0 + 1$$

$$0 < c < 1$$

$$\frac{1}{(n+1)!} < |R_n(0)| < \frac{e}{(n+1)!}$$

## Rozwiązanie

$$\frac{e}{(n+1)!} < 10^{-6} \quad \text{założony poziom błędu}$$

$$(n+1)! > 10^6 e$$

$$(n+1)! > 10^6 \times 3$$

$$n \geq 9$$

Co najmniej 9 wyrazów musimy zastosować aby otrzymać wartość błędu na poziomie  $10^{-6}$

## Przykład tragicznego błędu zaokrąglenia

25 lutego 1991 w Dhahran, Arabia Saudyjska, zginęło 28 amerykańskich żołnierzy w wyniku ataku irackiej rakiety Scud. System obrony Patriot nie wykrył zagrożenia. Dlaczego?

System oblicza powierzchnię, którą powinien skanować na podstawie prędkości obiektu i czasu ostatniej detekcji. Zegar wewnętrzny był ustawiony na pomiar co  $1/10$  sekundy. Długość słowa 24 bity. Z powodu zaokrąglenia błąd bezwzględny wyniósł  $9.5 \cdot 10^{-8}$  s a po 100 godzinach:

$$9.5 \cdot 10^{-8} \times 10 \times 60 \times 60 \times 100 = 0.34 \text{ sec}$$

Przesunięcie obliczone na tej podstawie 687 m. Obiekt jest uznany poza zakresem gdy przesunięcie wynosi 137 m

## 1. Dodawanie i odejmowanie

Aby dodać lub odjąć dwie znormalizowane liczby w zapisie zmiennoprzecinkowym, wykładniki w powinny być zrównane poprzez odpowiednie przesunięcie mantysy.

**Przykład:** Dodać  $0,4546 \cdot 10^5$  do  $0,5433 \cdot 10^7$



przesuwamy

$$0,0045 \cdot 10^7 + 0,5433 \cdot 10^7 = 0,5478 \cdot 10^7$$

**Wniosek:** Tracimy pewne cyfry znaczące

## 2. Mnożenie

**Przykład:** Pomnożyć  $0,5543 \cdot 10^{12}$  przez  $0,4111 \cdot 10^{-15}$

Mnożymy mantysy i wykładniki  $w$  dodajemy.

$$0,5543 \cdot 10^{12} \cdot 0,4111 \cdot 10^{-15} = 0,2278273 \cdot 10^{-3} = 0,2278 \cdot 10^{-3}$$

## 3. Dzielenie

**Przykład:** Podzielić  $0,1000 \cdot 10^5$  przez  $0,9999 \cdot 10^3$

$$0,1000 \cdot 10^5 / 0,9999 \cdot 10^3 = 0,1000 \cdot 10^2$$

Za każdym razem tracimy pewne cyfry znaczące co jest źródłem błędu

## Kolejność działań

**$(a+b)-c \neq (a-c)+b$**  brak przemienności, łączności

**$a(b-c) \neq (ab-ac)$**  brak rozdzielności mnożenia  
względem dodawania

**Przykład:**  $a = 0,5665 \cdot 10^1$ ,  $b = 0,5556 \cdot 10^{-1}$ ,  
 $c = 0,5644 \cdot 10^1$

$$(a+b) = 0,5665 \cdot 10^1 + 0,5556 \cdot 10^{-1}$$
$$= 0,5665 \cdot 10^1 + 0,0055 \cdot 10^1 = 0,5720 \cdot 10^1$$

$$(a+b)-c = 0,5720 \cdot 10^1 - 0,5644 \cdot 10^1 = 0,7600 \cdot 10^{-1}$$

$$(a-c) = 0,5665 \cdot 10^1 - 0,5644 \cdot 10^1 = 0,0021 \cdot 10^1 = 0,2100 \cdot 10^{-1}$$

$$(a-c)+b = 0,2100 \cdot 10^{-1} + 0,5556 \cdot 10^{-1} = 0,7656 \cdot 10^{-1}$$



- W wielu przypadkach można uniknąć błędów wejściowych i błędów obcięcia.
- W trakcie obliczeń pojawiają się nowe błędy (błędy zaokrągleń), których nie da się uniknąć.
- Błędy zaokrągleń można zmniejszyć ustalając umiejętnie sposób i kolejność wykonywania działań.

## Zadanie do samodzielnego rozwiązania

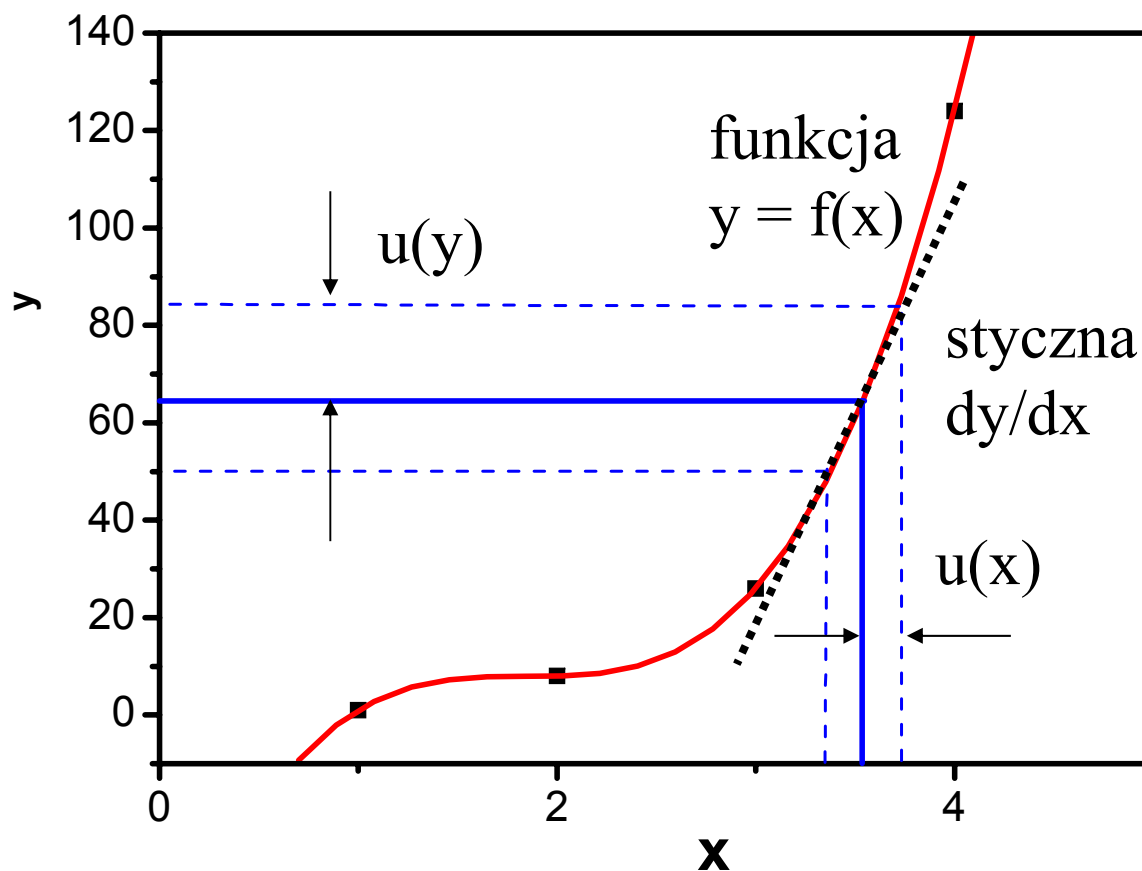
1. Szereg  $\sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} 4^n$

stanowi rozwinięcie Maclaurina funkcji:

(A)  $\cos(x)$ ; B)  $\cos(2x)$ ; C)  $\sin(x)$ ; D)  $\sin(2x)$

2. Liczba  $1/10$  jest zapisywana na 6 bitach w formacie stałoprzecinkowym (wszystkie bity są używane do zapisu części ułamkowej liczby). Różnica akumuluje się co  $1/10$  sekundy przez dobę. Jaka jest jej wartość?

# Propagacja błędów



$$u(y) = \frac{dy}{dx} u(x)$$

## Metoda różniczki zupełnej

Dla wielkości złożonej  $y=f(x_1, x_2, \dots, x_n)$  gdy niepewności maksymalne  $\Delta x_1, \Delta x_2, \dots, \Delta x_n$  są małe w porównaniu z wartościami zmiennych  $x_1, x_2, \dots, x_n$  niepewność maksymalną wielkości  $y$  wyliczamy z praw rachunku różniczkowego:

$$\Delta y = \left| \frac{\partial y}{\partial x_1} \right| |\Delta x_1| + \left| \frac{\partial y}{\partial x_2} \right| |\Delta x_2| + \dots + \left| \frac{\partial y}{\partial x_n} \right| |\Delta x_n|$$

## Przykład

Oszacować błąd pomiaru gęstości  $\rho$  kuli o masie  $m$  i promieniu  $R$

$$\rho(m, R) = \frac{m}{(4/3)\pi R^3}$$

błąd bezwzględny  $\Delta\rho = \left| \frac{\partial\rho}{\partial m} \right| |\Delta m| + \left| \frac{\partial\rho}{\partial R} \right| |\Delta R|$

ale  $\frac{\partial\rho}{\partial m} = \frac{1}{(4/3)\pi R^3}$        $\frac{\partial\rho}{\partial R} = \frac{-3}{(4/3)\pi R^4}$

błąd względny  $\varepsilon_\rho = \varepsilon_m + 3\varepsilon_R$

## Błędy działań arytmetycznych

### Błąd sumy

$$A = a \pm \Delta a$$

$$B = b \pm \Delta b$$

błędy bezwzględne składników sumy

$$A + B = a + b \pm \Delta a \pm \Delta b = a + b \pm \Delta(a + b)$$

błąd bezwzględny sumy

Zatem błąd bezwzględny sumy (różnicy) jest równy sumie błędów składników.

$$\Delta(a \pm b) = \Delta a + \Delta b$$

## Błędy działań arytmetycznych

Błąd względny sumy

$$\varepsilon_{a+b} = \frac{\Delta a + \Delta b}{a + b}$$

Błąd względny różnicy

$$\varepsilon_{a-b} = \frac{\Delta a + \Delta b}{a - b}$$

Błąd względny różnicy może być duży nawet gdy błędy względne odjemnej i odjemnika są małe. Należy unikać odejmowania prawie równych liczb przybliżonych!

Zjawisko zwane **redukcją** cyfr znaczących

Szczególnie istotne przy obliczeniach ilorazów różnicowych przybliżających pochodne funkcji, pierwiastków równania kwadratowego przy dominującym współczynniku przy pierwszej potędze, itp.

Tracimy dokładny sens liczby 0 jeśli dokonujemy obliczeń numerycznych

$$x^2 + 2x - 2 = 0$$

pierwiastkami są  $-1 \pm \sqrt{3}$

w przybliżeniu  
 $0,7320 \cdot 10^0$   
 $-0.2732 \cdot 10^1$

Sprawdzić, że po podstawieniu rozwiązań przybliżonych nie otrzymujemy dokładnie liczby zero

Powinno się zatem unikać odejmowania bliskich sobie liczb i warunek w pętli nie powinien być ustawiany „do zera”,

if  $a-b < \epsilon$

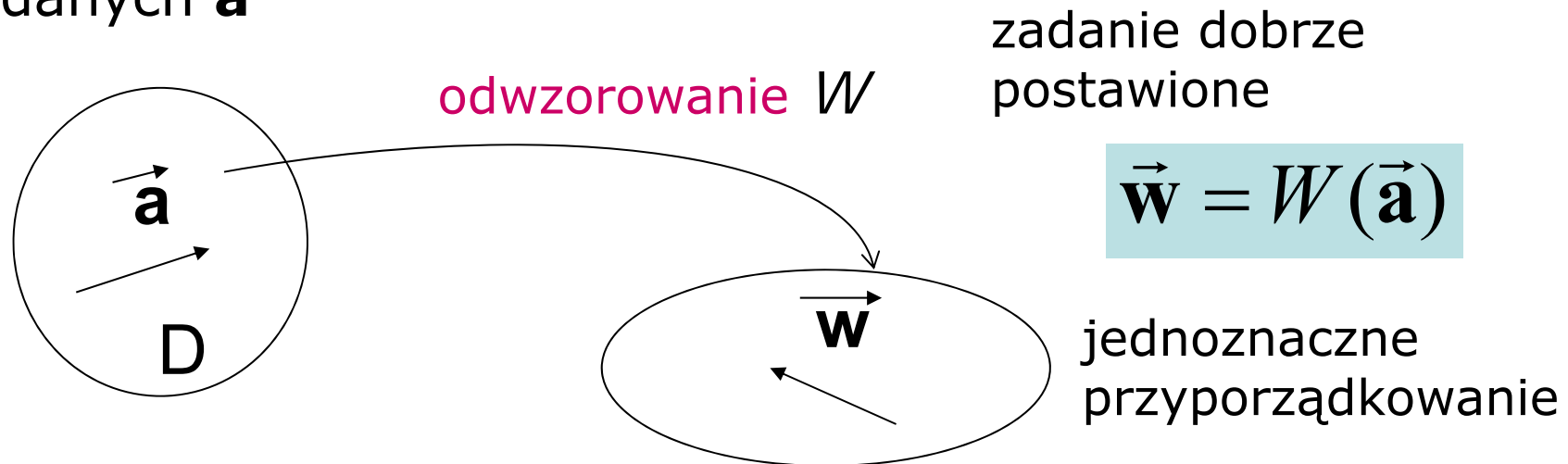


Przy obliczeniach numerycznych korzystne jest:

- ponowne rozwiązanie tego samego zagadnienia inną metodą lub taką samą metodą, ale z inną kolejnością operacji
- ponowne rozwiązanie zagadnienia przy nieznacznej zmianie danych wejściowych

## Zadania i algorytmy numeryczne

- **Zadanie numeryczne** wymaga jasnego i niedwuznacznego opisu powiązania funkcjonalnego między *danymi wejściowymi* czyli „zmiennymi niezależnymi” zadania i *danymi wyjściowymi*, tj. szukanymi wynikami.
- Zadanie numeryczne jest problemem polegającym na wyznaczeniu wektora wyników  $\vec{w}$  na podstawie wektora danych  $\vec{a}$



## Zadania i algorytmy numeryczne

- **Algorytm numeryczny** jest pełnym opisem poprawnie określonych *operacji* przekształcających wektor dopuszczalnych danych wejściowych (zbiór DN) na wektor danych wyjściowych.
- Algorytm jest poprawnie sformułowany gdy liczba niezbędnych działań będzie skończona

$$DN \cap D \neq \emptyset$$

$$\bar{\mathbf{w}} = WN(\mathbf{a}, \varepsilon)$$



wektor wyniku zależy od dokładności obliczeniowej  $\varepsilon$  maszyny cyfrowej

## Przykłady algorytmów

Dana jest liczba zespolona  $a = x + iy$ . Obliczyć  $1/a^2$

Algorytm I:

1.  $t = y / x$  (tangens fazy liczby  $a$ )

2.  $|a|^2 = x^2 + y^2$  (kwadrat modułu liczby  $a$ )

3.  $\operatorname{Re}\left(\frac{1}{a^2}\right) = \frac{1}{|a|^2} \frac{1-t^2}{1+t^2}$        $\operatorname{Im}\left(\frac{1}{a^2}\right) = \frac{1}{|a|^2} \frac{-2t^2}{1+t^2}$

Zadanie jest dobrze postawione, jeżeli:  $x^2 + y^2 \neq 0$

czyli:  $D = \mathbb{R}^2 - \{(0,0)\}$

Algorytm jest poprawnie sformułowany (11 niezbędnych działań)

## Przykłady algorytmów

Nie dla każdej pary danych  $(x,y) \neq 0$  można znaleźć rozwiązanie zadania stosując algorytm I.

1. Wystąpi nadmiar liczb zmiennopozycyjnych (dla  $x=0$  ale także z powodu zaokrąglenia do zera)
2. Nadmiar może nastąpić może już w pierwszym kroku gdy  $x=10^{-25}$  i  $y=10^{25}$  z powodu dzielenia  $y/x$
3. Dla  $x=0$ , istniejącego dla  $y \neq 0$  rozwiązania nie można wyznaczyć stosując ten algorytm. Wzrost dokładności obliczeń nie zmieni tego faktu.

Algorytm I nie jest numerycznie stabilny

## Przykłady algorytmów

Dana jest liczba zespolona  $a=x+iy$ . Obliczyć  $1/a^2$

Algorytm II:

$$1. \quad r = \operatorname{Re}\left(\frac{1}{a^2}\right) = \frac{x^2 - y^2}{x^2 + y^2}$$

$$2. \quad u = \operatorname{Im}\left(\frac{1}{a^2}\right) = \frac{-2xy}{x^2 + y^2}$$

Algorytm II jest poprawnie sformułowany (9 niezbędnych działań)

Algorytm II jest numerycznie stabilny co wynika z ciągłości wzorów dla  $x^2 + y^2 \neq 0$

## Schemat Hornera

Przykład **wzoru rekurencyjnego**

Aby obliczyć wartość wielomianu:

$$p(x) = x^n + a_1x^{n-1} + \dots + a_{n-1}x + a_n$$

w danym punkcie  $z$ , korzystamy ze schematu:

$$p_1 = z + a_1$$

$$p_2 = zp_1 + a_2$$

.....

$$p_n = zp_{n-1} + a_n$$

$$p(z) = p_n$$

co odpowiada obliczaniu wartości wyrażenia:

$$z\{z[z\dots(z + a_1) + a_2] + \dots + a_{n-1}\} + a_n$$

## Schemat Hornera

Schemat Hornera umożliwia znaczne zmniejszenie liczby działań arytmetycznych.

W schemacie Hornera wykonujemy  $n-1$  mnożeń i  $n$  dodawań.

Obliczając bezpośrednio:

$$\underbrace{z z \dots z}_{n \text{ razy}} + \underbrace{a_1 z \dots z}_{n-1 \text{ razy}} + \dots + a_{n-1} z + a_0$$

wykonujemy  $(n-1)(n+2)/2$  mnożeń i  $n$  dodawań.

Oszacowanie wielkości błędów zaokrągleń jest identyczne dla obu metod



## Schemat Hornera

Przykład: oblicz  $p(z) = a_0z^3 + a_1z^2 + a_2z + a_3$

w schemacie Hornera  $p(z) = ((a_0z + a_1)z + a_2)z + a_3$

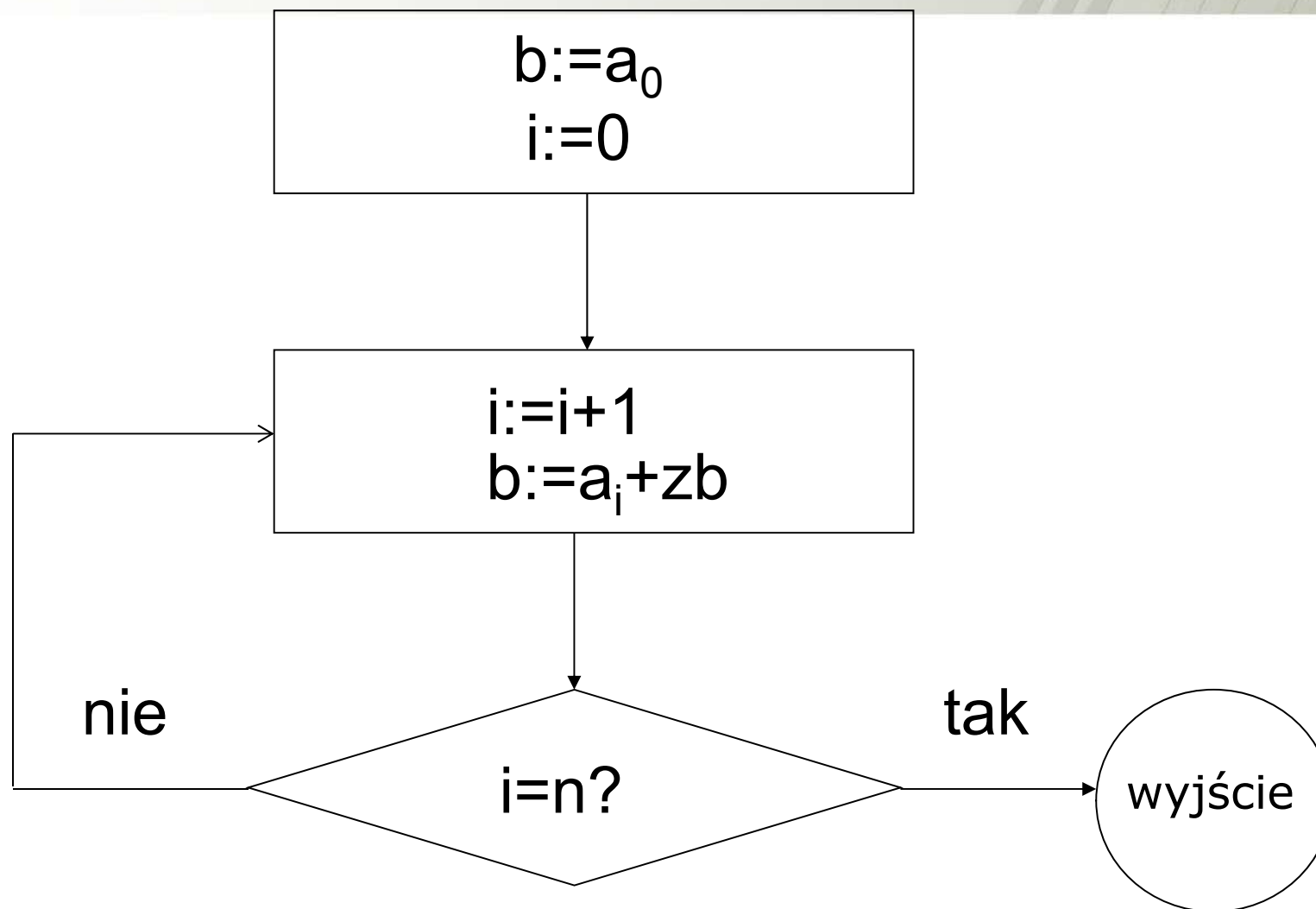
dla obliczeń ręcznych:

$$\begin{array}{cccc}
 a_0 & a_1 & a_2 & a_3 \\
 & zb_0 & zb_1 & zb_2 \\
 \hline
 b_0 & b_1 & b_2 & b_3 & p(z)=b_3
 \end{array}$$

Zadanie: Oblicz  $p(8)$  dla  $p(x)=2x^3+x+7$

$$\begin{array}{cccc}
 2 & 0 & 1 & 7 \\
 & 16 & 128 & 1032 \\
 \hline
 2 & 16 & 129 & 1039 & p(8)=1039
 \end{array}$$

## Schemat blokowy



## Uwarunkowanie zadania i stabilność algorytmów

Algorytm obliczeniowy jest **numerycznie stabilny**, jeżeli dla dowolnie wybranych danych

$$\mathbf{a}_0 \in D$$

istnieje taka dokładność obliczeń  $\varepsilon_0$ , że dla  $\varepsilon < \varepsilon_0$  mamy

$$\mathbf{a}_0 \in DN(\varepsilon)$$

oraz

$$\lim_{\varepsilon \rightarrow 0} WN(\mathbf{a}_0, \varepsilon) = W(\mathbf{a}_0)$$

Algorytm obliczeniowy jest **numerycznie stabilny** wtedy, gdy zwiększając dokładność obliczeń można wyznaczyć (z dowolną dokładnością) dowolne istniejące rozwiązanie zadania.

## Uwarunkowanie zadania i stabilność algorytmów

Uwarunkowaniem zadania nazywamy cechę, która mówi jak bardzo wynik dla zaburzonego wektora danych różni się od wyniku dla dokładnego wektora danych czyli:

$$W(\mathbf{a} + \delta\mathbf{a})$$

$$W(\mathbf{a})$$

**Wskaźnik uwarunkowania** zadania  $B(\mathbf{a})$  jest to liczba, dla której jest spełniony warunek:

$$\frac{\|\delta\mathbf{w}\|}{\|\mathbf{w}\|} \leq B(\mathbf{a}) \frac{\|\delta\mathbf{a}\|}{\|\mathbf{a}\|}$$

$$\delta\mathbf{w} = W(\mathbf{a} + \delta\mathbf{a}) - W(\mathbf{a})$$

## Wskaźnik uwarunkowania zadania

- Przyjmijmy względny błąd wielkości  $x$

$$\frac{x - \tilde{x}}{\tilde{x}}$$

- Względny błąd wielkości  $f(x)$

$$\frac{f(x) - f(\tilde{x})}{f(\tilde{x})} \approx \frac{f'(\tilde{x})(x - \tilde{x})}{f(\tilde{x})}$$

- Wskaźnik uwarunkowania:

$$\frac{\tilde{x}f'(\tilde{x})}{f(\tilde{x})}$$

## Wskaźnik uwarunkowania zadania

- Przykład

$$f(x) = \sqrt{x}$$

- Wskaźnik uwarunkowania:

$$\frac{\tilde{x}f'(\tilde{x})}{f(\tilde{x})} = \frac{x \frac{1}{2\sqrt{x}}}{\sqrt{x}} = \frac{1}{2}$$

zadanie dobrze uwarunkowane

## Wskaźnik uwarunkowania zadania

- Przykład

$$f(x) = \frac{10}{1-x^2}$$

- Wskaźnik uwarunkowania:

$$\frac{\tilde{x}f'(\tilde{x})}{f(\tilde{x})} = \frac{2x^2}{|1-x^2|}$$

zadanie źle uwarunkowane w pobliżu  $x=1$  i  $x=-1$