Laboratorium 02

Arytmetyka komputerowa

Adam Biśta, 11.03.2023

1 Treść zadań

1. Napisać algorytm do obliczenia funkcji wykładniczej e^x przy pomocy nieskończonych szeregów

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

   (a) Wykonując sumowanie w naturalnej kolejności, jakie kryterium zakończenia obliczeń przyjmiesz ?

   (b) Proszę przetestować algorytm dla:

$$x = + - 1, + - 5, + - 10$$

   i porównać wyniki z wynikami wykonania standardowej funkcji exp(x)

   (c) Czy można posłużyć się szeregami w tej postaci do uzyskania dokładnych wyników dla x < 0 ?

   (d) Czy możesz zmienić wygląd szeregu lub w jakiś sposób przegrupować składowe żeby uzyskać dokładniejsze wyniki dla x < 0 ?

2. Które z dwóch matematycznie ekwiwalentnych wyrażeń x**2 - y**2 oraz (x - y)*(x + y) może być obliczone dokładniej w arytmetyce zmienno-przecinkowej ? Dlaczego ?

3. Dla jakich wartości x i y, względem siebie, istnieje wyraźna różnica w dokładności dwóch wyrażeń ?

   Zakładamy że rozwiązujemy równanie kwadratowe ax**2 + bx + c = 0, z a = 1.22, b = 3.34 i c = 2.28, wykorzystując znormalizowany system zmienno-przecinkowy z podstawa beta = 10 i dokładnością p = 3.

   (a) ile wyniesie obliczona wartość b**2 – 4ac ?

   (b) jaka jest dokładna wartość wyróżnika w rzeczywistej (dokładnej) arytmetyce ?

   (c) jaki jest względny błąd w obliczonej wartości wyróżnika ?

2 Rozwiązania zadań

1.

   (a) Kryterium zakończenia obliczeń można określić na przykład na podstawie porównania kolejnych składników szeregu z pewnym małym parametrem epsilon. Jeśli kolejny składnik jest mniejszy niż epsilon, to sumowanie szeregu zostaje zakończone.

   Poniżej zamieszczony został algorytm postępowania napisany w języku Python:

```python
import math

def exp(x, epsilon=1e-10):
    result = 1.0
    term = 1.0
    n = 1
    while abs(term) > epsilon:
        term *= x / n
        result += term
        n += 1
    return result
```

   W tym algorytmie używamy domyślnej tolerancji równej 1e-10. Następnie korzystamy z pętli, aby iterować przez kolejne wyrazy nieskończonego szeregu. Warunkiem zakończenia pętli jest osiągnięcie wartości bezwzględnej wyrazu mniejszej od tolerancji. W każdej iteracji mnożymy poprzedni wyraz przez $\frac{x}{n}$, dodajemy do wyniku i zwiększamy n o 1.

   (b) Poniżej przedstawione zostały wyniki dla przygotowanych danych z zadania 1b. math.exp() to funkcja załadowana z biblioteki natomiast exp() to funkcja zdefiniowana powyżej.

```python
test_values = [1, -1, 5, -5, 10, -10]
for x in test_values:
    result = exp(x)
    expec = math.exp(x)
    print(f"x={x}:\n exp(x):{result:.13f},\n math.exp(x):{expec:.13f}")
```

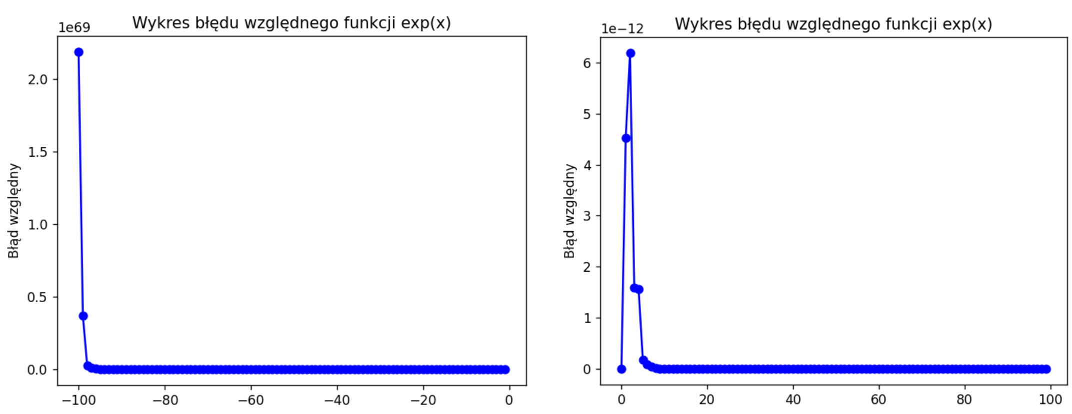
   Wyniki obliczeń:
   x=1:
   exp(x):2.7182818284**468**,
   math.exp(x):2.7182818284**590**
   x=-1:
   exp(x):0.3678794411**607**,
   math.exp(x):0.3678794411**714**
   x=5:

exp(x):148.4131591025**514**,
math.exp(x):148.4131591025**766**
x=-5:
exp(x):0.00673794**70171**,
math.exp(x):0.00673794**69991**
x=10:
exp(x):22026.4657948066**706**,
math.exp(x):22026.465794806r**7179**
x=-10:
exp(x):0.0000453998**989**,
math.exp(x):0.0000453999**298**

W powyższych wynikach na czarno zaznaczono, że od 10. miejsca po przecinku cyfry stały się niezgodne. Na czerwono oznaczono cyfry, które poniżej 10. miejsca zaczęły być niezgodne.
błąd względny: $|\frac{obliczona - dokladna}{dokladna}| = |\frac{exp(x) - math.exp(x)}{math.exp(x)}|$

```
blad wzgledny dla 1:   0.0000000000045
blad wzgledny dla -1:  0.0000000000292
blad wzgledny dla 5:   0.0000000000002
blad wzgledny dla -5:  0.0000000026780
blad wzgledny dla 10:  0.0000000000000
blad wzgledny dla -10: 0.0000006794807
```

Wartości ujemne mają większy błąd względny od wartości dodatnich



(a) Wykres 1
(b) Wykres 2

Rysunek 1: Porównywanie wykresów

(c) Nie możemy użyć szeregów w podanej wyżej postaci do uzyskania dokładnych wyników dla x < 0. Zauważamy, że dla wartości ujemnych wynik szeregu staje się mniej precyzyjny.

(d) Aby uzyskać dokładniejsze wyniki dla x < 0, można obliczyć najpierw wartość $e^{|x|}$, a następnie odwrócić wynik:

$$e^{-|x|} = \frac{1}{e^{|x|}}$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots, x >= 0$$

$$e^{-|x|} = \frac{1}{1 + |x| + \frac{|x|^2}{2!} + \frac{|x|^3}{3!} + \cdots}, x < 0$$

Podana wyżej operacja wykonywana dla liczb ujemnych zapobiega problemom związanym z odejmowaniem liczb bliskich sobie, co prowadzi do utraty dokładności wyniku i zwiększenia błędu wyniku w przypadku stosowania zwykłego szeregu dla x < 0. Zjawisko to nosi nazwę catastrophic cancellation.

2. $x^2 - y^2$ a $(x - y)(x + y)$

Wzór $x^2 - y^2$:

Zalety:

- Gdy $|x| >> |y|$, wyrażenie $x^2 - y^2$ może zostać obliczone z większą dokładnością, ponieważ błąd wynikający z ewaluacji funkcji $fl(y^2)$ nie wpłynie znacząco na wynik końcowy.

Wady:

- Istnieje ryzyko, że wartości x^2 lub y^2 będą zbyt duże, aby móc je reprezentować w pamięci, podczas gdy wyniki dodawania i odejmowania x i y będzie mieścić się w zakresie arytmetyki. W takim przypadku różnica między wynikami może być znaczna.

- Jeśli $x \approx y$, to wyrażenie $x^2 - y^2$ powoduje względnie duże błędy, ponieważ cyfry znaczące w wyniku zostaną zredukowane, a błąd pozostanie taki sam.

Wzór $(x+y)(x-y)$:

Zalety:

- Ten wzór zazwyczaj daje dokładniejszy wynik niż $x^2 - y^2$, gdyż ewaluacja błędu w obliczeniach jest mniejsza.

Wady:

- Wyrażenie $x - y$ również może prowadzić do dużych błędów względnych, ale w porównaniu do $x^2 - y^2$ jest znacznie mniejszy.

Podsumowując, wybór między tymi dwoma wzorami zależy od konkretnych argumentów i wartości. Oba wzory mają swoje zalety i wady, ale w przypadku, gdy obliczenia są wykonywane w niskiej precyzji arytmetyki, lepiej jest użyć wzoru $(x-y)(x+y)$.

3. Mamy dane równanie $1.22x^2 + 3.34x + 2.28 = 0$
   Zakładając, że obliczenia wykonywane są w systemie zmiennopozycyjnym o podstawie $\beta = 10$ i dokładności $p = 3$:
   (a) Wartość wyrażenia $b^2 - 4ac$:
   $fl(b \cdot b) = fl(3.34 \cdot 3.34) = fl(11.1556) = 11.2$
   $fl(4 \cdot a \cdot c) = fl(fl(4 \cdot a) \cdot c) = fl(fl(4 \cdot 1.22) \cdot 2.28) = fl(fl(4.88) \cdot 2.28) = fl(4.88 \cdot 2.28) = fl(11.1264) = 11.1$
   Wartość wyrażenia wynosi $fl(11.2 - 11.1) = 0.1$.
   (b) Dokładna wartość wyróżnika w rzeczywistej (dokładnej) arytmetyce:
   $b^2 - 4ac = 11.1556 - 11.1264 = 0.0292$
   (c) Względny błąd w obliczonej wartości wyróżnika:
   $blad\ wzgledny = |\frac{0.0292-0.1}{0.0292}| = \frac{0.0708}{0.0292} \approx 2.42466$

   Wnioski:
   Wadą rozwiązania w systemie zmiennopozycyjnym o podstawie $\beta = 10$ i dokładności p = 3 jest występowanie dużego błędu względnego przy obliczeniu wartości wyróżnika, który wynosi około 242%.

3 Bibliografia

- Katarzyna Rycerz: "Wykład z przedmiotu Metody Obliczeniowe w Nauce i Technice"
- https://en.wikipedia.org/wiki/Catastrophic-cancellation