MOwNiT - Laboratorium 1: Arytmetyka komputerowa

Wojciech Dąbek

5 marca 2024

1 Treści zadań

1. Znaleźć *"maszynowe epsilon"*, czyli najmniejszą liczbę a, taką że $a+1 > 1$.
2. Rozważamy problem ewaluacji funkcji $sin(x)$, m.in. propagację błędu danych wejściowych, tj. błąd wartości funkcji ze względu na zakłócenie h w argumencie x:
   - Ocenić błąd bezwzględny przy ewaluacji $sin(x)$.
   - Ocenić błąd względny przy ewaluacji $sin(x)$
   - Ocenić uwarunkowanie dla tego problemu
   - Dla jakich wartości argumentu x problem jest bardzo czuły?
3. Funkcja sinus zadana jest nieskończonym ciągiem:

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

   - Jakie są błędy progresywny i wsteczny jeśli przybliżamy funkcję sinus biorąc tylko pierwszy człon rozwinięcia,
     tj. $sin(x) \approx x$, dla $x = 0.1, 0.5$ i 1.0?
   - Jakie są błędy progresywny i wsteczny jeśli przybliżamy funkcję sinus biorąc pierwsze dwa człony rozwinięcia,
     tj. $sin(x) \approx x - \frac{x^3}{6}$, dla $x = 0.1, 0.5$ i 1.0?
4. Zakładamy że mamy znormalizowany system zmiennoprzecinkowy z $\beta = 10, p = 3, L = -98$.
   - Jaka jest wartość poziomu UFL (underflow) dla takiego systemu?
   - Jeśli $x = 6.87 \cdot 10^{-97}$ i $y = 6.81 \cdot 10^{-97}$, jaki jest wynik operacji $x - y$?

2 Rozwiązania

2.1

Przyjmując definicję *maszynowego epsilonu* z polecenia będzie on równy wartości "jednostki ostatniego miejsca" względem 1, czyli wartości przyjmowanej, gdy mantysa jest równa 1 (najmniejsza), a wykładnik taki jak dla 1.
Zatem:

$$\varepsilon = b^{1-p}$$

gdzie b jest podstawą systemu liczbowego, a p precyzją.

Przykładowo dla wartości typu *double* w języku C++ stosowana jest podstawa 2 i precyzja 53, więc $\varepsilon = 2^{-52} \approx 2,22 \cdot 10^{-16}$.

2.2

- Błąd bezwzględny:

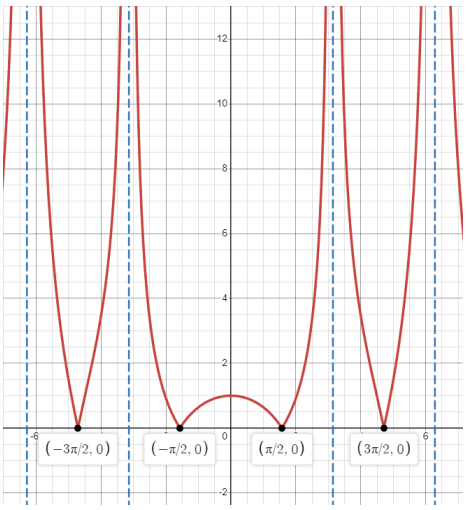
$$\Delta \sin x = |\sin x - \sin(x(1+\varepsilon))|$$

- Błąd względny:

$$\frac{\Delta \sin x}{\sin x} = \frac{|\sin x - \sin(x(1+\varepsilon))|}{\sin x}$$

- Uwarunkowanie:

$$\text{cond}(\sin x) = \left|\frac{x \sin' x}{\sin x}\right| = \left|\frac{x \cos x}{\sin x}\right| = |x \operatorname{ctg} x|$$



Rysunek 1: $y = \text{cond}(\sin x) = |x \operatorname{ctg} x|$

- Jak widać na wykresie funkcji uwarunkowania, problem jest bardzo czuły w parzystych wielokrotnościach π oprócz zera, gdyż funkcja ucieka tam do $+\infty$.

Wnioski: Sinus jest funkcją najgorzej uwarunkowaną w otoczeniu swoich miejsc zerowych, a najlepiej w otoczeniu swoich ekstremów lokalnych. Wynika to z "przeplatania się" miejsc zerowych i ekstremów funkcji $\sin x$ i jej pochodnej $\cos x$.

2.3

Dla funkcji $y = f(x)$ błąd progresywny określamy jako $|\hat{y} - y|$, a błąd wsteczny jako $|\hat{x} - x|$. Tutaj rozważamy funkcję $sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$

- Tylko pierwszy człon rozwinięcia

x	$y = \sin x$	$\hat{y} \approx x$	$\lvert\hat{y} - y\rvert$	$\hat{x} = \arcsin \hat{y}$	$\lvert\hat{x} - x\rvert$
0,1	0,09983341664	0,1	0,00016658336	0,100167421	0,000167421
0,5	0,4794255386	0,5	0,0205744614	0,523598776	0,023598776
1	0,8414709848	1	0,1585290152	1,57079633	0,57079633

Tabela 1: Wartości dla przybliżenia $\sin x \approx x$

- Dwa pierwsze człony rozwinięcia

x	$y = \sin x$	$\hat{y} \approx x - \frac{x^3}{6}$	$\lvert\hat{y} - y\rvert$	$\hat{x} = \arcsin \hat{y}$	$\lvert\hat{x} - x\rvert$
0,1	0,09983341664	0,0998(3)	0,0000000833	0,0999999163	0,0000000837
0,5	0,4794255386	0,4791(6)	0,0002588719	0.499705041	0.000294959
1	0,8414709848	0,8(3)	0.00813765147	0,985110783	0.014889217

Tabela 2: Wartości dla przybliżenia $\sin x \approx x - \frac{x^3}{6}$

Wnioski: Można zauważyć ogromne zmniejszenie obu rodzajów błędów przy zastosowaniu większej ilości wyrazów, nawet dodając tylko jeden. Wzrost błędów wraz z argumentem funkcji zgadza się z wnioskami z poprzedniego zadania ze względu na uwarunkowanie sinusa.

2.4

- Wartość poziomu UFL (underflow) jest najmniejszą dodatnią liczbą możliwą do zapisania w danym systemie, a więc z mantysą 1 i możliwie najmniejszym wykładnikiem.
  Dla systemu z polecenia otrzymujemy więc:

$$UFL = \beta^L = 10^{-98}$$

- Licząc matematycznie $x - y = 6,87 \cdot 10^{-97} - 6,81 \cdot 10^{-97} = 6 \cdot 10^{-99}$.
  Jest to liczba mniejsza od UFL, przez co w tym systemie wynikiem tej operacji będzie 0.

Wnioski: Skoro UFL stanowi miarę dokładności (co widać również z powyższej analizy), to system służący do bardzo dokładnych obliczeń powinien mieć jak najmniejszy parametr L (dolną granicę zakresu wykładnika).

3 Bibliografia

Wykład prof. Heatha *Scientific Computing*
https://en.wikipedia.org/wiki/Machine_epsilon
https://en.wikipedia.org/wiki/Condition_number