Obliczenia inżynierskie

Projekt 1

Analiza dokładności obliczeń komputerowych

Autor

Jakub Strzelczyk **325325**

Oświadczam, że niniejsza praca, stanowiąca podstawę do uznania osiągnięcia efektów uczenia się z przedmiotu Obliczenia inżynierskie została wykonana przeze mnie samodzielnie.

Jakub Strzelczyk

Spis treści

Wydział Elektroniki
i Technik Informacyjnych

POLITECHNIKA WARSZAWSKA

1. Lista symboli matematycznych

— x, y, z, r, θ_2, ϕ_2, v - zmienne skalarne
— $\tilde{x}$, $\tilde{y}$, $\tilde{z}$, $\tilde{r}$, $\tilde{\theta}$, $\tilde{\phi}$, $\tilde{v}$ - zaburzone wartości zmiennych skalarnych
— ε, ε_1, ε_1 - błędy danych
— η_1, η_2 - błędy zaokrągleń operacji zmiennoprzecinkowych
— $\delta(\tilde{v})$, $\delta[\phi(\tilde{x}, y, z)$, $\delta\left[\tilde{y}\left(\tilde{r}, \tilde{\theta}, \tilde{\phi}\right)\right]$ - błędy względne $\tilde{v}$, $\tilde{\phi}$, $\tilde{y}$
— T(v), $T_{x\phi}(x, y, z)$, $T(z)$ - współczynniki charakteryzujące przenoszenie błędu względnego
— L - liczba cyfr znaczących mantysy
— eps - graniczny błąd względny

2. Wprowadzenie

Reprezentacja zmiennoprzecinkowa liczb jest oparta o notację wykładniczą, gdzie liczba przedstawiana jest za pomocą dwóch głównych komponentów: mantysy i cechy. Zakładając, że podstawa systemu liczbowego wynosi 10, ogólną formę takiej liczby x można wyrazić jako [1]:

$$x = \pm m \cdot 10^c \tag{1}$$

gdzie mantysa m jest postaci $m = 0.m_1 m_2 \dots m_{L-1} m_L m_{L+1} \dots$, przy czym $m_i \in \{0, \dots, 9\}$ i $m_1 \neq 0$,natomiast cecha $c \in C$.

Ze względu na ograniczoną precyzję z jaką mantysa i cecha są przechowywane w komputerze, przetrzymywane w pamięci liczby zmiennoprzecinkowe są tylko przybliżeniami liczby rzeczywistej, którą próbują przedstawić. Wartość zaburzoną (zaokrągloną) $\tilde{x}$ można obliczyć korzystając z poniższego wzoru:

$$\tilde{x} = \pm \tilde{m} \cdot 10^c, \quad \text{gdzie } \tilde{m} = \begin{cases} 0.m_1 m_2 \dots m_L & \text{dla } 0 \leqslant m_{L+1} < 5 \\ 0.m_1 m_2 \dots m_L + 10^{-L} & \text{dla } 5 \leqslant m_{L+1} < 9 \end{cases} \tag{2}$$

Korzystając z definicji błędu względnego skalara $\tilde{x}$ oznaczonego jako $\delta(\tilde{x})$, możemy oszacować jego maksymalną wartość:

$$|\delta(\tilde{x})| = \left|\frac{\tilde{x} - x}{x}\right| = \left|\frac{\tilde{m} \cdot 10^c - m \cdot 10^c}{m \cdot 10^c}\right| = \left|\frac{\tilde{m} - m}{m}\right| \tag{3}$$

Powyższy wzór przyjmie największą wartość, gdy licznik jest największy, natomiast mianownik najmniejszy z możliwych. Dla przyjętej reprezentacji licznik może maksymalnie przyjąć wartość $\frac{1}{2} \cdot 10^{-L}$ - dokładna wartość mantysy będzie znajdować się na osi liczbowej w połowie odległości między liczbą $0.m_1 m_2 \dots m_L$ a $0.m_1 m_2 \dots m_L + 10^{-L}$. Minimalna wartość mantysy w tym przypadku wynosi 0.1, zatem:

$$|\delta(\tilde{x})| \leqslant \frac{\frac{1}{2} \cdot 10^{-L}}{10^{-1}} = 5 \cdot 10^{-L} \tag{4}$$

Wówczas liczbę o powyższej wartości oznaczamy jako **eps** i błąd względny danej nie może przekroczyć tej wartości. Dzięki niej możemy szacować przenoszenie błędów względnych na określone dane. Powyższa wartość jest przyjęta dla określonych powyżej warunków. Do obliczania przenoszonych błędów względnych stosujemy metody rachunku epsilonów i różniczkowania analitycznego, które zostaną zastosowane w dalszej części projektu.

Wartość liczby **eps** zależy od precyzji, w jakiej komputer reprezentuje liczby. Najważniejszym standardem reprezentacji binarnej i operacji na liczbach zmiennoprzecinkowych jest **IEEE754** [2]. Standard ten definiuje dwa rodzaje liczb zmiennoprzecinkowych - pojedynczej i podwójnej precyzji. Przykładowo dla liczby pojedycznej precyzji w standardzie IEEE754 wartość eps wynosi $5,96 \cdot 10^{-8}$, natomiast dla liczby podwójnej precyzji wartość ta wynosi $1,11 \cdot 10^{-16}$.

Liczbę **pojedynczej precyzji** w tym formacie zapisujemy za pomocą 32 bitów - pierwszy bit jest bitem znaku, 8 bitów kodujących cechę (BIAS=127) oraz kolejne 23 bity to mantysa liczby.
Liczba **podwójnej precyzji** zapisana jest za pomoca 64 bitów - bit znaku, 11 bitów na cechę (BIAS=1023) i 52 bity mantysy.

Oprócz błędów danych wynikających z reprezentacji poszczególnych liczb, występują także błędy zaokrągleń wynikające z przybliżeń wykonywanych przez komputer po wykonywanych operacjach zmiennoprzecinkowych. Przykładowo, podczas dodawania dwóch liczb zmiennoprzecinkowych, komputer najpierw reprezentuje dwie liczby z określoną precyzją, a następnie po dodaniu zaburzonych wartości liczb zaokrągla sumę.

Kolejność i sposób wykonywania działań w algorytmach mają wpływ na wartość błędów zaokrągleń. Posługując się metodami numerycznymi, można oszacować błędy dla poszczególnych algorytmów i porównać je między sobą. Składowa błędu wyniku spowodowana przenoszeniem błędów danych nie zależy od algorytmu, natomiast składowa spowodowana przenoszeniem błędów zaokrągleń wyników operacji zależy od niego, zatem porównując dwa algorytmy prowadzące do wykonania tego samego zadania, możemy skupić się jedynie na błędach spowodowanych przybliżaniem wyników operacji zmiennoprzecinkowych.

W celu wykonania projektu skorzystałem z następujących wzorów dotyczących powiązania współrzędnych kartezjańskich $< x, y, z >$ i sferycznych $< r, \theta, \phi >$:

$$\begin{cases} r = \sqrt{x^2+y^2+z^2} \\ \theta = \arccos(\frac{z}{r}) \\ \varphi = sgn(y)\arccos(\frac{x}{\sqrt{x^2+y^2}}) \end{cases} \qquad \begin{cases} x = r\sin(\theta)\cos(\varphi) \\ y = r\sin(\theta)\sin(\varphi) \\ z = r\cos(\theta) \end{cases}$$

Celem projektu jest zbadanie i analiza wpływu dokładności reprezentacji numerycznej na przekształcenia między współrzędnymi kartezjańskimi i sferycznymi w przestrzeni trójwymiarowej.

3. Zadanie 1

3.1. Metodyka i wyniki doświadczeń

W przestrzeni trójwymiarowej współrzędna sferyczna ϕ jest powiązana ze współrzędnymi kartezjańskimi ⟨x, y, z⟩ w następujący sposób:

$$\phi = \operatorname{sgn}(y) \cdot \arccos\left(\frac{x}{\sqrt{x^2+y^2}}\right) \tag{5}$$

W celu wyznaczenia współczynnika charakteryzującego przenoszenie błędu względnego, jakim obarczona jest dana x na wartość współrzędnej ϕ, można skorzystać z metody różniczkowania analitycznego i metody rachunku epsilonów.

3.1.1. Metoda rachunku epsilonów

Najpierw obliczyłem przenoszenie błędu przez funkcję arccos :

$$\arccos(\tilde{v}) = \arccos(v(1+\delta(\tilde{v}))) = \arccos(v)\cdot(1+T(v)\delta(\tilde{v})) \tag{6}$$

$$T(v) = \frac{v}{\arccos(v)}\frac{d\arccos(v)}{dv} = \frac{v}{\arccos(v)}\frac{-1}{\sqrt{1-v^2}} = \frac{-v}{\sqrt{1-v^2}\arccos(v)} \tag{7}$$

$$\arccos(\tilde{v}) = \arccos(v)\cdot\left(1-\frac{v}{\sqrt{1-v^2}\arccos(v)}\delta(\tilde{v})\right) \tag{8}$$

Następnie wykonałem obliczenia pozwalające obliczyć współczynnik T(x) metodą rachunku epsilonów. Przyjąłem, że y jest stałą, która nie jest obarczona błędem danych, natomiast x jest reprezentowana w komputerze poprzez przybliżenie.

$$\tilde{x} = x(1+\varepsilon) \tag{9}$$

$$\tilde{\phi} = \operatorname{sgn}(y)\arccos\left(\frac{x(1+\varepsilon)}{\sqrt{x^2(1+\varepsilon)^2+y^2}}\right) \tag{10}$$

$$\tilde{\phi} \cong \operatorname{sgn}(y)\arccos\left(\frac{x(1+\varepsilon)}{\sqrt{x^2(1+2\varepsilon)+y^2}}\right) \tag{11}$$

$$\tilde{\phi} \cong \operatorname{sgn}(y)\arccos\left(\frac{x(1+\varepsilon)}{\sqrt{x^2+y^2}\sqrt{1+\frac{2x^2\varepsilon}{x^2+y^2}}}\right) \tag{12}$$

$$\tilde{\phi} \cong \text{sgn}(y) \arccos\left(\frac{x(1+\varepsilon)}{\sqrt{x^2+y^2}\left(1+\frac{x^2\varepsilon}{x^2+y^2}\right)}\right) \tag{13}$$

$$\tilde{\phi} \cong \text{sgn}(y) \arccos\left(\frac{x(1+\varepsilon)}{\sqrt{x^2+y^2}}\left(1-\frac{x^2\varepsilon}{x^2+y^2}\right)\right) \tag{14}$$

$$\tilde{\phi} \cong \text{sgn}(y) \arccos\left(\frac{x}{\sqrt{x^2+y^2}}\left[1+\varepsilon\left(1-\frac{x^2}{x^2+y^2}\right)\right]\right) \tag{15}$$

$$\tilde{\phi} \cong \text{sgn}(y) \arccos\left(\frac{x}{\sqrt{x^2+y^2}}\right) \cdot \left(1-\frac{\frac{x}{\sqrt{x^2+y^2}}}{\sqrt{1-\frac{x^2}{x^2+y^2}} \arccos\left(\frac{x}{\sqrt{x^2+y^2}}\right)}\left(1-\frac{x^2}{x^2+y^2}\right)\varepsilon\right) \tag{16}$$

$$\tilde{\phi} \cong \text{sgn}(y) \arccos\left(\frac{x}{\sqrt{x^2+y^2}}\right) \cdot \left(1-\frac{\frac{x}{\sqrt{x^2+y^2}}}{\frac{\sqrt{y^2}}{\sqrt{x^2+y^2}} \arccos\left(\frac{x}{\sqrt{x^2+y^2}}\right)}\frac{x^2-x^2-y^2}{x^2+y^2}\varepsilon\right) \tag{17}$$

$$\tilde{\phi} \cong \text{sgn}(y) \arccos\left(\frac{x}{\sqrt{x^2+y^2}}\right) \cdot \left(1+\frac{-xy^2}{|y| \arccos\left(\frac{x}{\sqrt{x^2+y^2}}\right) \cdot (x^2+y^2)}\varepsilon\right) \tag{18}$$

$$T_{x\phi}(x,y,z) = -\frac{xy^2}{|y| \arccos\left(\frac{x}{\sqrt{x^2+y^2}}\right) \cdot (x^2+y^2)} \tag{19}$$

3.1.2. Metoda różniczkowania analitycznego

$$T(x) = \frac{x}{\phi}\frac{d\phi}{dx} \tag{20}$$

$$T_{x\phi}(x,y,z) = \frac{x}{\text{sgn(y)} \arccos\left(\frac{x}{\sqrt{x^2+y^2}}\right)} \cdot \text{sgn(y)} \cdot \frac{-1}{\sqrt{1-\frac{x^2}{x^2+y^2}}} \cdot \frac{\sqrt{x^2+y^2}-x\cdot\frac{x}{\sqrt{x^2+y^2}}}{x^2+y^2} \tag{21}$$

$$T_{x\phi}(x,y,z) = \frac{-x}{\arccos\left(\frac{x}{\sqrt{x^2+y^2}}\right)} \cdot \frac{1}{\frac{\sqrt{y^2}}{\sqrt{x^2+y^2}}} \cdot \frac{\frac{y^2}{\sqrt{x^2+y^2}}}{x^2+y^2} \tag{22}$$

$$T_{x\phi}(x,y,z) = -\frac{xy^2}{|y| \arccos\left(\frac{x}{\sqrt{x^2+y^2}}\right) \cdot (x^2+y^2)} \tag{23}$$

3.2. Dyskusja wyników eksperymentów numerycznych

Zarówno w przypadku metody rachunku epsilonów, jak i metody różniczkowania analitycznego uzyskaliśmy takie same wyniki współczynnika charakteryzującego przenoszenie błędu względnego, jakim obarczona jest dana x, na wartość współrzędnej ϕ. Metoda rachunku epsilonów wymagała nieco więcej obliczeń oraz zastosowania podstawowych reguł rachunku epsilonów, upraszczających działania. Z kolei metoda różniczkowania analitycznego wymagała znajomości podstawowych pochodnych funkcji złożonych. Oba wyniki są identyczne, co potwierdza spójność wykonanych działań za pomocą obu metod.

3.3. Wnioski

Metody różniczkowania analitycznego oraz rachunku epsilonów umożliwiły obliczenie szukanego współczynnika, a otrzymane wyniki są takie same, co świadczy o możliwości zamienności stosowania obu metod. W celu obliczenia współczynnika możemy przeanalizować funkcję, a następnie wybrać metodę, która w prostszy sposób umożliwi estymację szukanego parametru. Kolejnym podejściem może być wykonanie obliczeń obiema metodami w celu uzyskania pewności poprawności wykonania działań - tak jak w przypadku tego zadania. Gdy wykonane przez nas obliczenia będą poprawne, otrzymamy identyczne wyniki dla obu przypadków.

4. Zadanie 2

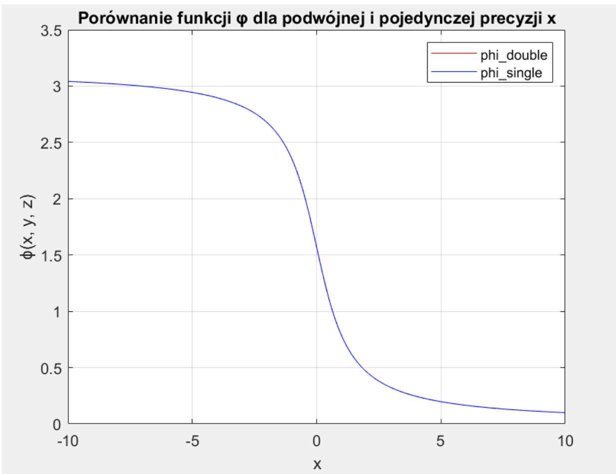
4.1. Metodyka i wyniki doświadczeń

W celu wykonania zadania przyjąłem następujące wartości danych: $x \in [-10, 10]$, y = 1, z = -0.1.

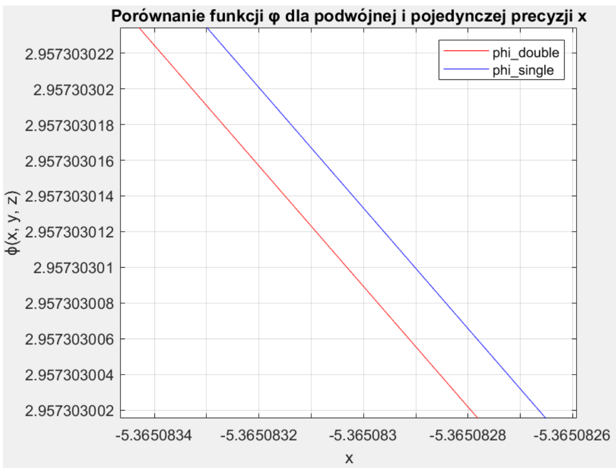
Korzystając z MATLABa wyznaczyłem wartość $\phi(x, y, z)$, używając reprezentacji zmiennopozycyjnej podwójnej precyzji, a następnie $\phi(\tilde{x}, y, z)$, gdzie $\tilde{x}$ jest wynikiem zapisu wartości x z pojedynczą precyzją.

Używając wbudowanej funkcji *plot* utworzyłem wykresy uzyskanych funkcji. Kod mojego programu został zorganizowany w logicznie uporządkowane funkcje, z których każda odpowiada za określoną funkcjonalność. W celu zwiększenia czytelności każda funkcja jest poprzedzona komentarzem, opisującym jej przeznaczenie.

Na rysunku 1. przedstawiono oryginalny wykres, prezentujący różnice pomiędzy $\phi(x, y, z)$ oraz $\phi(\tilde{x}, y, z)$ w zależności od wartości x; natomiast na rysunku 2. przedstawiono ten sam wykres po powiększeniu.



Rysunek 1. Wykres porównujący wartości $\phi(x, y, z)$ oraz $\phi(\tilde{x}, y, z)$

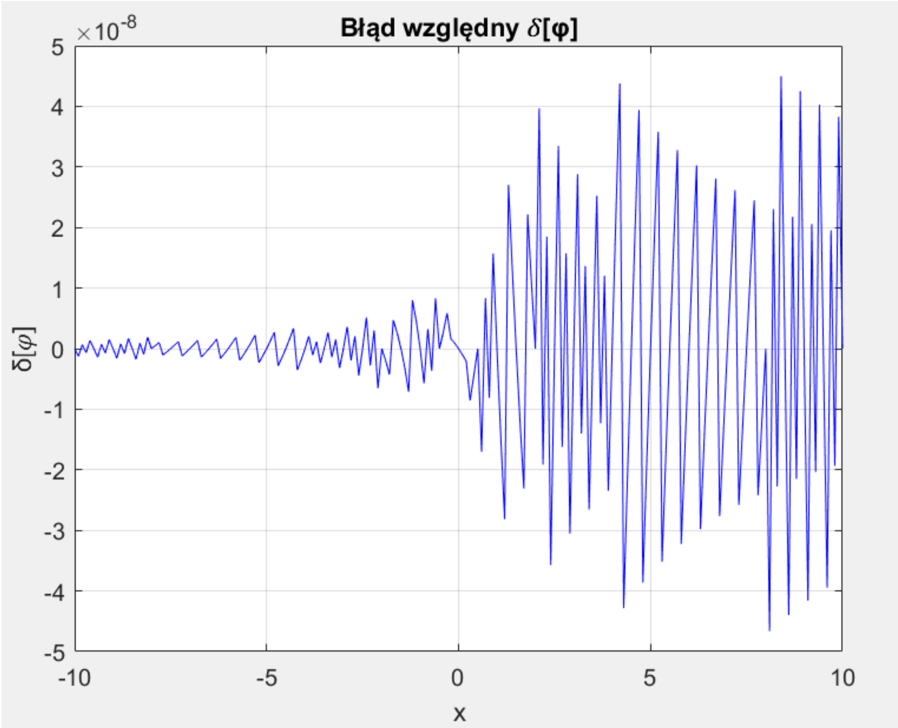


Rysunek 2. Wykres szczegółowo porównujący wartości$\phi(x, y, z)$ oraz $\phi(\tilde{x}, y, z)$

Następnie wyznaczyłem błąd względny korzystając z następującego wzoru:

$$\delta[\phi(\tilde{x}, y, z)] = \frac{\phi(\tilde{x}, y, z) - \phi(x, y, z)}{\phi(x, y, z)]} \tag{24}$$

Na rysunku 3. przedstawiono wykres prezentujący wartość błędu względnego $\delta[\phi(\tilde{x}, y, z)]$ w zależności od zmiany wartości x.



Rysunek 3. Wykres wartości błędu względnego $\delta[\phi(\tilde{x}, y, z)]$

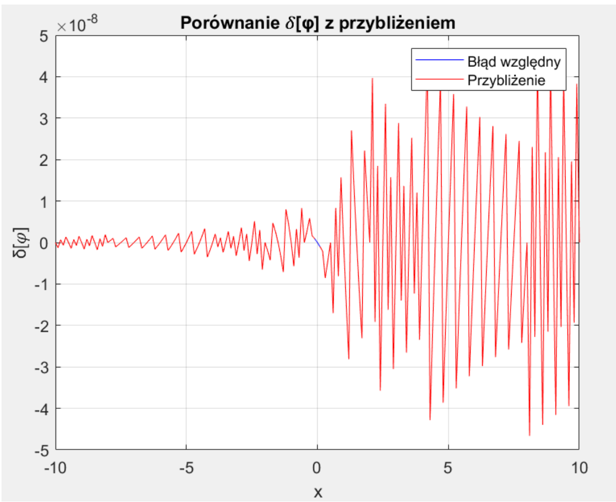
Następnie porównałem uzyskane wyniki z następującym przybliżeniem, uzyskanym dzięki współczynnikowi wyznaczonym w poprzednim zadaniu:

$$\delta[\phi(\tilde{x}, y, z)] \cong T_{x\phi}(x, y, z)\delta(\tilde{x}) \tag{25}$$

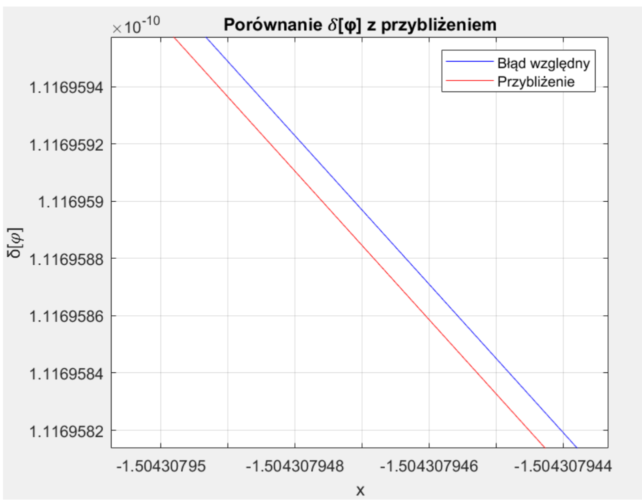
gdzie

$$\delta[\tilde{x}] = \frac{\tilde{x} - x}{x} \tag{26}$$

Na rysunku 4. przedstawiono wykres porównujący wartość błędu $\delta[\phi(\tilde{x}, y, z)]$ z jego przybliżeniem, natomiast na rysnku 5. przedstawiono ten sam wykres po zastosowaniu powiększenia.



Rysunek 4. Wykres porównujący wartość błędu $\delta[\phi(\tilde{x}, y, z)]$ z przybliżeniem



Rysunek 5. Wykres po powiększeniu, porównujący wartość błędu $\delta[\phi(\tilde{x}, y, z)]$ z jego przybliżeniem

4.2. Dyskusja wyników eksperymentów numerycznych

Wartości $\phi(x, y, z)$ oraz $\phi(\tilde{x}, y, z)$ zaznaczone na wykresie widocznym na rysunku 1. na pierwszy rzut oka nie różnią się, natomiast po powiększeniu wykresu jak na rysunku 2. widać różnice pomiędzy wartościami uzyskanymi dla pojedynczej i podwójnej precyzji x. Na podstawie kolejnego wykresu widocznego na rysunku 3. można odczytać wartość błędu względnego $\delta[\phi(\tilde{x}, y, z)]$ - nie przekracza ona $5 \cdot 10^{-8}$, a przenoszony błąd względny jest mniejszy dla wartości ujemnych x, niż dla wartości dodatnich, co może sugerować asymetrię w sposobie w jakim błędy są przenoszone w tej funkcji. Obliczony współczynnik przenoszenia wyznaczony w zadaniu 1. umożliwił mi przybliżenie uzyskanych wyników $\delta[\phi(\tilde{x}, y, z)]$. Wartość przybliżenia nieznacznie różni się od dokładnej wartości, a w celu ujrzenia różnic na wykresie widocznym na rysunku 4., musiałem go powiększyć.

4.3. Wnioski

Zaburzenie x poprzez zapis jego wartości w pojedynczej precyzji spowodowało także nieznaczne zmiany w wynikach funkcji zależnej od tej zmiennej. Jakikolwiek błąd danych argumentów funkcji powoduje także zmianę wartości funkcji, co ma praktyczne skutki w projektowaniu, programowaniu różnych urządzeń, systemów. W zależności od wartości argumentów danej funkcji, przenoszony błąd względny się różni; a przybliżenia, np. błędu $\delta[\phi(\tilde{x}, y, z)]$, wprowadzają kolejne błędy do wyniku końcowego. Gdy projektujemy aplikację mniej krytyczną i dopuszczalna jest w niej zmniejszona precyzja, wówczas możemy dokonywać przybliżeń. Dla zastosowań wymagających wyjątkowej dokładności, ważne jest stosowanie jak największej precyzji i unikanie przybliżeń, które wprowadzają kolejne błędy. Zatem, wybór stosowanych przez nas algorytmów czy upraszczania działań zależy od wymagań zadań projektowych.

5. Zadanie 3

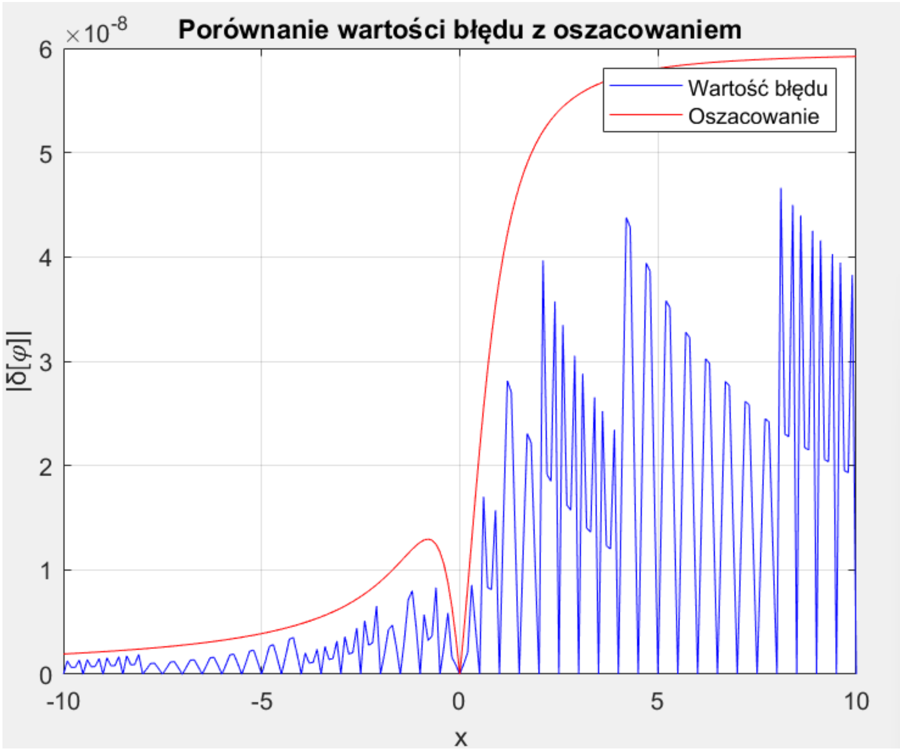
5.1. Metodyka i wyniki doświadczeń

Korzystając z poniższego wzoru:

$$|\delta[\phi(\tilde{x}, y, z)]| \leqslant |T_{x\phi}(x, y, z)| \, \mathrm{eps}_{\mathrm{single}} \tag{27}$$

gdzie eps_{single} oznacza wartość **eps** dla pojedynczej precyzji, dokonałem porównania wartości błędów, wyznaczonych w ramach zadania 2, z oszacowaniem.

Na rysunku 6. przedstawiono wykres porównujący moduł wartości błędów $\delta[\phi(\tilde{x}, y, z)]$ z ich oszacowaniem, obliczonym z wykorzystaniem wartości eps.



Rysunek 6. Wykres porównujący wartość błędu $\delta[\phi(\tilde{x}, y, z)]$ z oszacowaniem

5.2. Dyskusja wyników eksperymentów numerycznych

Na wykresie przedstawionym na rysunku 6. widoczne jest porównanie modułu wartości błędu $\delta[\phi(\tilde{x}, y, z)]$ z oszacowaniem. Oszacowane wartości są zawsze większe niż rzeczywista wartość, natomiast prezentują te same rzędy wielkości. Estymacja dla dodatnich wartości x nieco bardziej odbiega od rzeczywistych wartości błędu niż w przypadku wartości ujemnych.

5.3. Wnioski

Oszacowanie przenoszenia błędu z wykorzystaniem wartości eps pozwala dokładnie określić graniczny błąd względny określonej danej. Posiadając tę wiedzę, możemy ustalić górną granicę odchyleń, które mogą wystąpić, co umożliwia przewidywanie zachowania danego systemu lub algorytmu w najbardziej krytycznych scenariuszach.

6. Zadanie 4

6.1. Metodyka i wyniki doświadczeń

Za pomocą rachunku epsilonów wykonałem działania umożliwiające obliczenie błędu względnego jakim obarczona jest wartość współrzędnej y wyznaczona na podstawie współrzędnych sferycznych r, θ i ϕ zapisanych w reprezentacji zmiennopozycyjnej. Tym razem wziąłem pod uwagę zarówno błędy danych, jak i błędy zaokrągleń wyników zmiennopozycyjnych. Wyniki działań przedstawiam poniżej:

$$y = r\sin(\theta)\sin(\phi) \tag{28}$$

Błędy danych i błędy zaokrągleń:

$$\tilde{r} = r(1+\varepsilon_1) \tag{29}$$

$$\tilde{\theta} = \theta(1+\varepsilon_2) \tag{30}$$

$$\tilde{\phi} = \phi(1+\varepsilon_3) \tag{31}$$

$$v_1 = \sin(\theta) \tag{32}$$

$$\tilde{v_1} = v_1(1+\eta_1) \tag{33}$$

$$v_2 = \sin(\phi) \tag{34}$$

$$\tilde{v_2} = v_2(1+\eta_2) \tag{35}$$

$$v_3 = rv_1 \tag{36}$$

$$\tilde{v_3} = v_3(1+\eta_3) \tag{37}$$

$$v_4 = v_3v_2 \tag{38}$$

$$\tilde{v_4} = v_4(1+\eta_4) \tag{39}$$

Przenoszenie błędu przez funkcję sinus:

$$\sin(\tilde{z}) = \sin(z(1+\delta(\tilde{z}))) = \sin(z)\cdot(1+T(z)\delta(\tilde{z})) \tag{40}$$

$$T(z) = \frac{z}{\sin(z)}\frac{d\sin(z)}{dz} = \frac{z}{\sin(z)}\cos(z) = z\cot(z) \tag{41}$$

$$\sin(\tilde{z}) = \sin(z)\cdot(1+z\cot(z)) \tag{42}$$

Podstawiając zaburzone wartości danych, kontynuowałem działania:

$$\tilde{y} = \{[r(1+\varepsilon_1)\sin\left(\theta(1+\varepsilon_2)\right)(1+\eta_1)](1+\eta_3)\sin\left(\phi(1+\varepsilon_3)\right)(1+\eta_2)\}(1+\eta_4) \tag{43}$$

$$\tilde{y} = \{[r(1+\varepsilon_1)\sin(\theta)(1+\theta\cot(\theta)\varepsilon_2)(1+\eta_1)](1+\eta_3)\sin(\phi)(1+\phi\cot(\phi)\varepsilon_3)(1+\eta_2)\}(1+\eta_4) \tag{44}$$

$$\tilde{y} = r(1+\varepsilon_1)\sin(\theta)(1+\theta\cot(\theta)\varepsilon_2)(1+\eta_1)(1+\eta_3)\sin(\phi)(1+\phi\cot(\phi)\varepsilon_3)(1+\eta_2)(1+\eta_4) \tag{45}$$

$$\tilde{y} \cong r\sin(\theta)\sin(\phi)(1+\varepsilon_1+\theta\cot(\theta)\varepsilon_2+\phi\cot(\phi)\varepsilon_3+\eta_1+\eta_2+\eta_3+\eta_4) \tag{46}$$

$$\tilde{y} \cong y(1+\varepsilon_1+\theta\cot(\theta)\varepsilon_2+\phi\cot(\phi)\varepsilon_3+\eta_1+\eta_2+\eta_3+\eta_4) \tag{47}$$

Mając $\tilde{y}$ przedstawione w powyższej formie, dokonałem oszacowania błędu względnego $\delta(\tilde{y})$ jakim obarczona jest wartość y:

$$|\delta[\tilde{y}]| \leqslant |\varepsilon_1| + |\theta\cot(\theta)||\varepsilon_2| + |\phi\cot(\phi)||\varepsilon_3| + |\eta_1| + |\eta_2| + |\eta_3| + |\eta_4| \tag{48}$$

$$|\delta[\tilde{y}]| \leqslant |\theta\cot(\theta)|eps + |\phi\cot(\phi)|eps + 5eps \tag{49}$$

6.2. Dyskusja wyników eksperymentów numerycznych

Błąd względny współrzędnej y, wynikający z błędów danych oraz zaokrągleń wyników zmiennoprzecinkowych, dla dokładnych wartości θ, ϕ, co do modułu nie przekroczy wartości $|\theta\cot(\theta)|\varepsilon+|\phi\cot(\phi)|\varepsilon+5\varepsilon$. Działania wykonałem korzystając z rachunku epsilonów, przyjmując, że każda z wartości współrzędnych sferycznych jest zaburzona, a także wartości wyników zmiennopozycyjnych są zaokrąglane. Podstawowa zasada, która umożliwiła mi wykonanie powyższych obliczeń określona jest poniższym wzorem:

$$(1+\varepsilon_1)\cdot(1+\varepsilon_2) \cong 1+\varepsilon_1+\varepsilon_2 \tag{50}$$

6.3. Wnioski

Za pomocą prostych przekształceń matematycznych metoda epsilonów umożliwia sprawne oszacowanie błędów spowodowanych zaokrągleniami danych oraz zaokrągleniem wyników operacji zmiennopozycyjnych. Wymagana jest znajomość podstawowych reguł rachunku epsilonów a także pochodnych prostych funkcji.

7. Zadanie 5

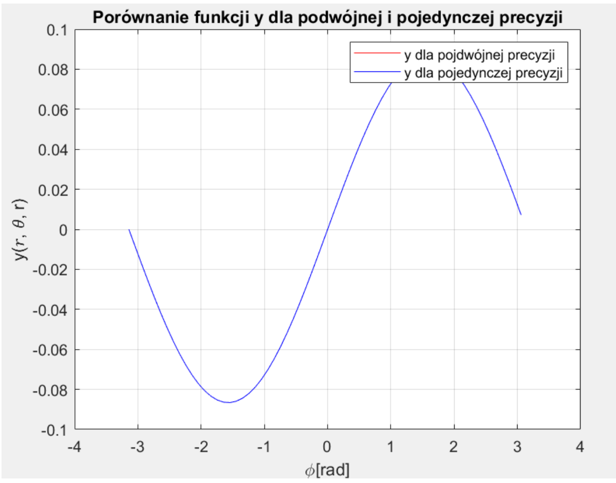
7.1. Metodyka i wyniki doświadczeń

W celu wykonania zadania przyjąłem następujące wartości danych: $r = 0.1$, $\theta = \frac{\pi}{3}$, $\phi \in [-\pi, \pi]$.

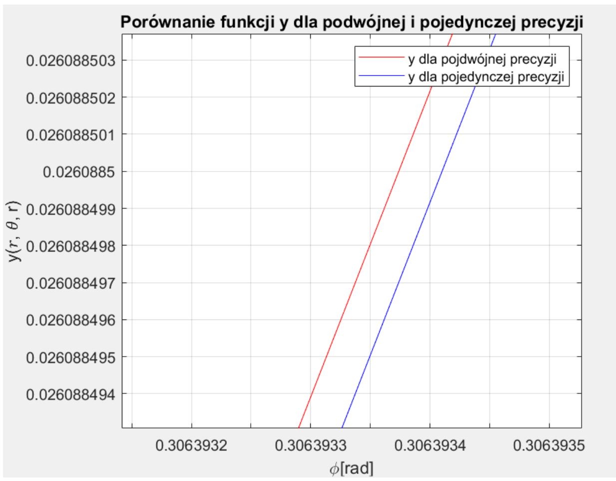
Korzystając z MATLABa wyznaczyłem wartość $y(r, \theta, \phi)$, używając reprezentacji zmiennopozycyjnej podwójnej precyzji, a następnie $\tilde{y}(\tilde{r}, \tilde{\theta}, \tilde{\phi})$, używając reprezentacji zmiennopozycyjnej pojedynczej precyzji.

Korzystając z wbudowanych funkcji plot oraz semilogy utworzyłem wykresy uzyskanych funkcji. Kod mojego programu został zorganizowany w logicznie uporządkowane funkcje, z których każda odpowiada za określoną funkcjonalność. W celu zwiększenia czytelności, każda funkcja jest poprzedzona komentarzem, opisującym jej przeznaczenie.

Na rysunku 6. przedstawiono oryginalny wykres oraz po przybliżeniu, prezentujący różnice pomiędzy $y(r, \theta, \phi)$ oraz $y(\tilde{r}, \tilde{\theta}, \tilde{\phi})$ w zależności od wartości x, natomiast na rysunku 7. przedstawiono ten sam wykres po przybliżeniu.



Rysunek 7. Wykres porównujący wartości $y(r, \theta, \phi)$ oraz $\tilde{y}(\tilde{r}, \tilde{\theta}, \tilde{\phi})$

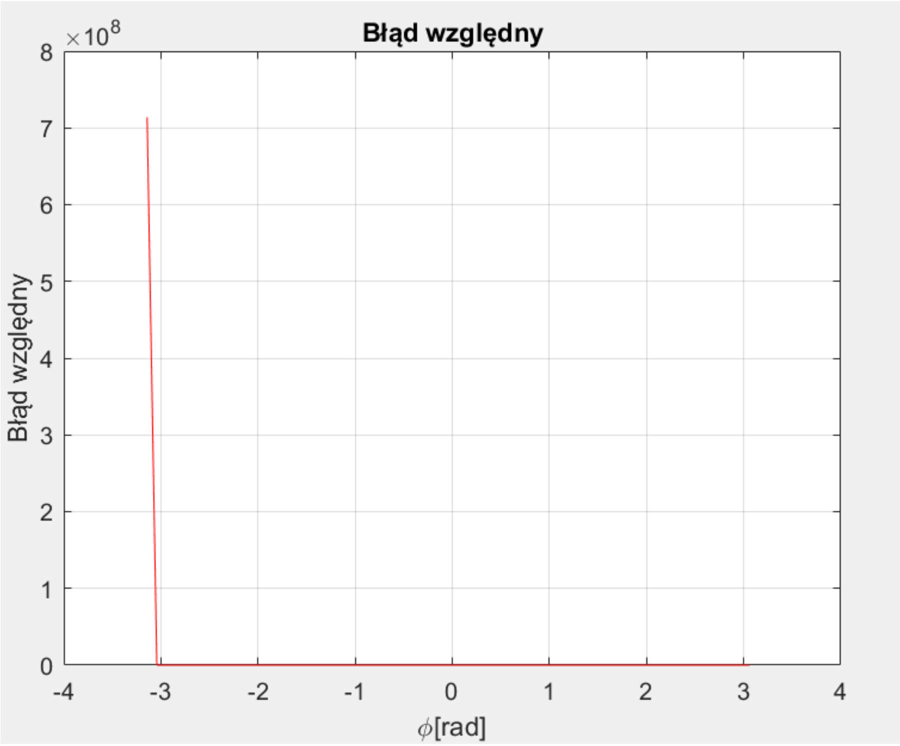


Rysunek 8. Wykres szczegółowo porównujący wartości $y(r, \theta, \phi)$ oraz $\tilde{y}(\tilde{r}, \tilde{\theta}, \tilde{\phi})$

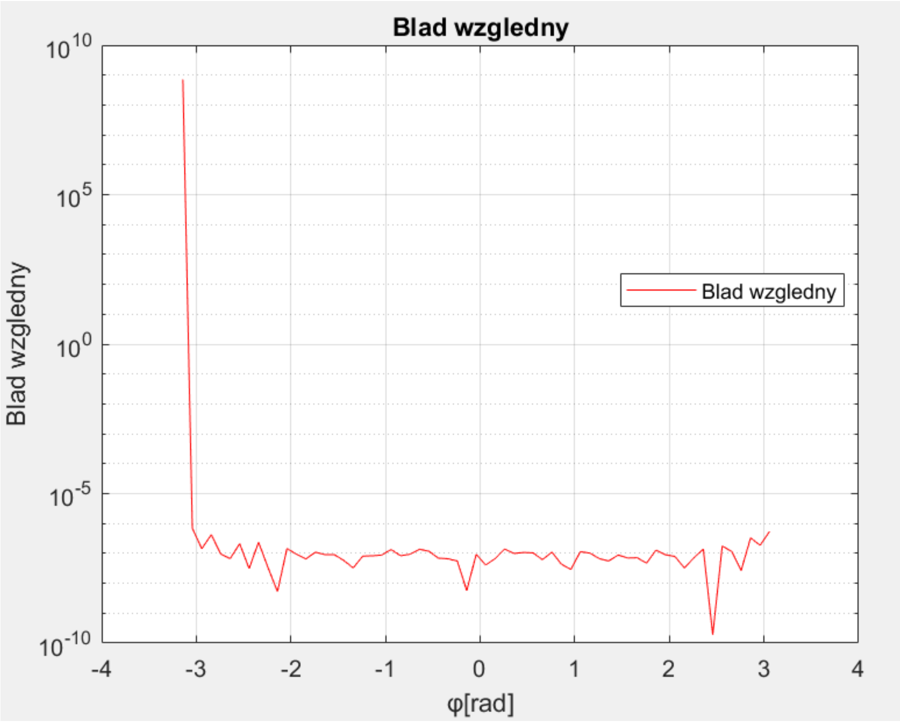
Następnie wyznaczyłem błąd względny korzystając z następującego wzoru:

$$\delta[\tilde{y}(\tilde{r}, \tilde{\theta}, \tilde{\phi})] = \frac{\tilde{y}(\tilde{r}, \tilde{\theta}, \tilde{\phi}) - y(r, \theta, \phi)}{y(r, \theta, \phi)} \tag{51}$$

Na rysunku 9. przedstawiono wykres prezentujący wartość błędu względnego $\delta[\tilde{y}(\tilde{r}, \tilde{\theta}, \tilde{\phi})]$ w zależności od zmiany wartości ϕ, natomiast na rysunku 10. zaprezentowano ten sam wykres po zastosowaniu skali logarytmicznej na osi pionowej.



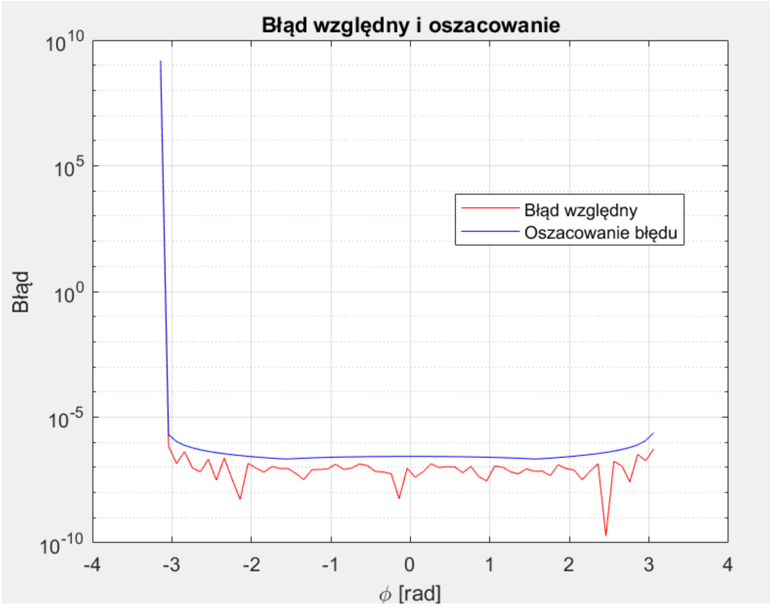
Rysunek 9. Wykres wartości błędu względnego $\delta[\tilde{y}(\tilde{r},\tilde{\theta},\tilde{\phi})]$



Rysunek 10. Wykres wartości błędu względnego $\delta[\tilde{y}(\tilde{r},\tilde{\theta},\tilde{\phi})]$ po zastosowaniu skali logarytmicznej

Następnie porównałem uzyskane wyniki z oszacowaniem uzyskanym w ramach zadania 4.

Na rysunku 11. przedstawiono wykres porównujący moduł wartości błędu $\delta[\tilde{y}(\tilde{r},\tilde{\theta},\tilde{\phi})]$ z oszacowaniem z wykorzystaniem wartości eps.



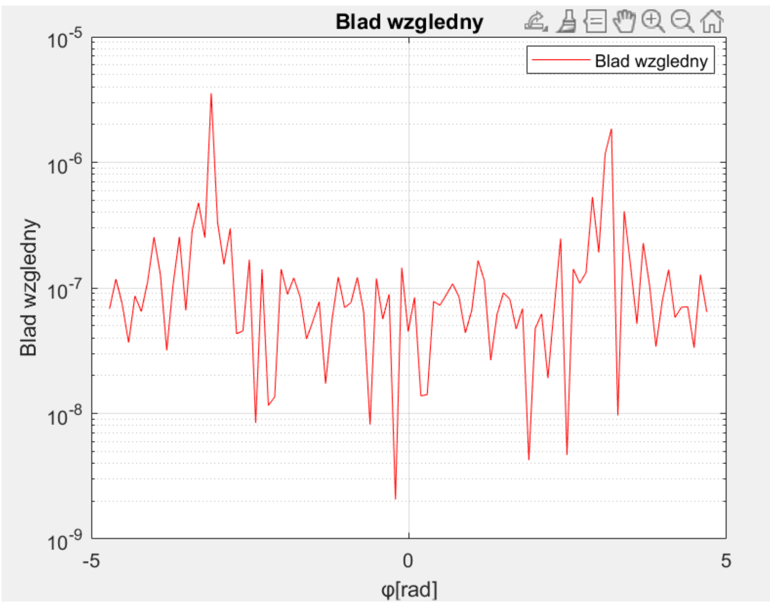
Rysunek 11. Wykres porównujący wartość błędu $\delta[\tilde{y}(\tilde{r}, \tilde{\theta}, \tilde{\phi})]$ z oszacowaniem

Na wykresie na osi pionowej została ustawiona skala logarytmiczna w celu uzyskania większej czytelności różnic pomiędzy oszacowaniem a wartością dokładną.

7.2. Dyskusja wyników eksperymentów numerycznych

W pierwszej analizie, wartości $y(r, \theta, \phi)$ oraz $\tilde{y}(\tilde{r}, \tilde{\theta}, \tilde{\phi})$, prezentowane na wykresie Rysunku 7. mogą na pierwszy rzut oka wydawać się identyczne. Jednakże, dokładniejsze przyjrzenie się danym, szczególnie po powiększeniu wykresu jak na Rysunku 8., ukazuje różnice między wartościami uzyskanymi przy użyciu pojedynczej i podwójnej precyzji.

Z kolejnych wykresów, widocznych na Rysunkach 9. oraz 10. wynika, że wartość błędu względnego $\delta[\tilde{y}(\tilde{r}, \tilde{\theta}, \tilde{\phi})]$ dla wartości $\phi = -\pi$ jest znacząco wysoka. Następnie, w miarę wzrostu wartości ϕ, błąd ten szybko maleje, aż stabilizuje się na poziomie bliskim 10^{-7}. Równie istotny jest fakt, iż wysoki błąd przy $\phi = -\pi$ może wynikać z błędów numerycznych w MATLAB-ie, spowodowanych przez wartość funkcji sinus, która wówczas wynosi zero i występuje w mianowniku. Rozszerzenie zakresu wartości ϕ do przedziału $[-\pi, \pi]$ umożliwiło uzyskanie bardziej szczegółowego wykresu wartości błędu $\delta[\tilde{y}(\tilde{r}, \tilde{\theta}, \tilde{\phi})]$, który przedstawiono na Rysunku 12.



Rysunek 12. Wykres porównujący wartość błędu $\delta[\tilde{y}(\tilde{r}, \tilde{\theta}, \tilde{\phi})]$ z oszacowaniem

Wyniki przedstawione na wykresie na Rysunku 10. mogą wynikać z metody obliczeń błędu w MATLAB-ie, szczególnie, gdy wartości funkcji sinus są bliskie zeru. W rzeczywistości, błąd dla tych wartości jest mniejszy, co można zauważyć po zmianie zakresu na osi poziomej.

Na wykresie na Rysunku 11. obserwujemy porównanie modułu wartości błędu $\delta[\tilde{y}(\tilde{r},\tilde{\theta},\tilde{\phi})]$ z jego oszacowaniem. Oszacowane wartości błędu są większe niż rzeczywiste, ale są tego samego rzędu wielkości. Estymata błędu jest zbliżona do rzeczywistych wartości. Obserwując wartości błądu względnego można zauważyć jego charakterystyczne wahania, natomiast oszacowany błąd prezentuje się bardziej równomiernie.

7.3. Wnioski

Proste przekształcenia matematyczne oraz metoda epsilonów umożliwiają skuteczne oszacowanie błędów wynikających z zaokrągleń danych oraz zaokrągleń wyników operacji zmiennopozycyjnych. Błędy zaokrąglenia operacji zmiennoprzecinkowych także wpływają na wyniki operacji, co należy brać pod uwagę np. przy wybieraniu konkretnego algorytmu rozwiązywania zadania numerycznego.

8. Zadanie 6

8.1. Metodyka i wyniki doświadczeń

W celu oszacowania minimalnej liczby cyfr znaczących mantysy reprezentacji zmiennoprzecinkowej o podstawie 2, dla której: $\left|\delta\left[\tilde{y}\left(\tilde{r},\tilde{\theta},\tilde{\phi}\right)\right]\right| \leqslant 10^{-9}$ gdy $r = 0.01$, $\theta = \frac{\pi}{8}$, $\phi = \frac{\pi}{4}$ skorzystałem z następującego oszacowania obliczonego w zadaniu 4:

$$|\delta[\tilde{y}]| \leqslant |\theta \cot(\theta)|eps + |\phi \cot(\phi)|eps + 5eps \tag{52}$$

Przeprowadzając podobne rozumowanie do zawartego we wprowadzeniu, można obliczyć wartość liczby eps dla reprezentacji zmiennopozycyjnej o podstawie 2 i liczbie cyfr znaczących mantysy L:

$$eps = \frac{sup|\tilde{m} - m|}{inf|m|} = \frac{\frac{1}{2}\cdot 2^{-L}}{1\cdot 2^{-1}} = 2^{-L} \tag{53}$$

Podstawiając wartość eps oraz wartości r, θ, ϕ, $\delta[\tilde{y}]$ do nierówności (45) otrzymałem:

$$|\delta[\tilde{y}]| \leqslant \left|\frac{\pi}{8}\cot(\frac{\pi}{8})\right| 2^{-L} + \left|\frac{\pi}{4}\cot(\frac{\pi}{4})\right| 2^{-L} + 5\cdot 2^{-L} \tag{54}$$

$$|\delta[\tilde{y}]| \leqslant \left|\frac{\pi}{8}\cot(\frac{\pi}{8})\right| 2^{-L} + \left|\frac{\pi}{4}\cot(\frac{\pi}{4})\right| 2^{-L} + 5\cdot 2^{-L} \tag{55}$$

$$|\delta[\tilde{y}]| \leqslant 2^{-L}\left(\frac{\pi}{8}\cdot(1+\sqrt{2}) + \frac{\pi}{4} + 5\right) \tag{56}$$

Błąd napewno będzie mniejszy niż 10^{-9}, gdy:

$$10^{-9} \geqslant 2^{-L}\left(\frac{\pi}{8}\cdot(1+\sqrt{2}) + \frac{\pi}{4} + 5\right) \tag{57}$$

$$10^{-9} \geqslant 2^{-L}\left(\frac{3+\sqrt{2}}{8}\pi + 5\right) \tag{58}$$

$$2^{-L} \leqslant \left(\frac{10^{-9}}{\frac{3+\sqrt{2}}{8}\pi + 5}\right) \tag{59}$$

$$-L \leqslant \log_2\left(\frac{10^{-9}}{\frac{3+\sqrt{2}}{8}\pi + 5}\right) \tag{60}$$

$$L \geqslant \log_2\left(\frac{\frac{3+\sqrt{2}}{8}\pi + 5}{10^{-9}}\right) \tag{61}$$

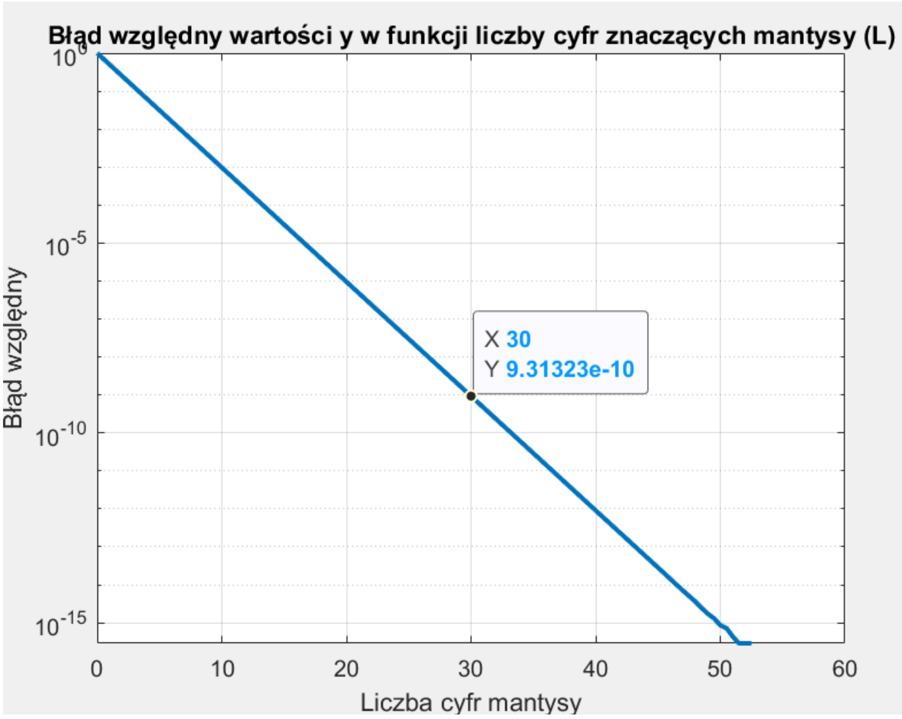
$$L \geqslant 32.99 \tag{62}$$

Powyższa metoda korzysta z oszacowania z Zadania 4. Zadanie można wykonać również poprzez obliczenie następującej wartości:

$$\delta[\tilde{y}(\tilde{r},\tilde{\theta},\tilde{\phi})] = \frac{\tilde{y}(\tilde{r},\tilde{\theta},\tilde{\phi}) - y(r,\theta,\phi)}{y(r,\theta,\phi)} \tag{63}$$

obliczając y na podstawie dostarczonych w treści zadania danych i podstawiając $\tilde{y} = y(1 + 2^{-L})$, gdzie L to liczba cyfr znaczących mantysy. Zatem przyjąłem, że liczba $\tilde{y}$ jest obarczona błędem względnym eps. Wykonałem powyższe działanie w Matlabie a następnie narysowałem wykres błędu względnego w zależności od liczby cyfr mantysy. Z wykresu odczytałem wartość L, dla której uzyskany błąd jest mniejszy niż 10^{-9}, ale jak najbliższy niemu.

Na rysunku 13. przedstawiono wykres oszacowanego błędu względnego w zależności od liczby cyfr znaczących mantysy.



Rysunek 13. Wykres oszacowanego błędu względnego w zależności od liczby cyfr znaczących mantysy

8.2. Dyskusja wyników eksperymentów numerycznych

Dla pierwszego sposobu nierówność zachodzi dla L $\geqslant$ 32.99, zatem minimalna liczba cyfr znaczących mantysy reprezentacji zmiennoprzecinkowej wynosi 33. Przeprowadzenie powyższych działań daje pewność, że błąd względny będzie mniejszy bądź równy szukanej wartości. Gwarantuje nam to użycie oszacowanej wartości błędu względnego, w której wykorzystana jest wartość eps do określenia granicznego błędu względnego dla y. Dla drugiego sposobu minimalna liczba cyfr mantysy, dla której spełnione jest szukane równanie wynosi 30. Oba wyniki różnią się o 3 cyfry, wynika to z różnych sposobów szacowania.

8.3. Wnioski

Pierwszy sposób polega na ręcznym szacowaniu wykorzystanym dzięki rachunkowi epsilonów. Wartość uzyskana za pomocą pierwszego sposobu jest większa, gdyż używając w tym przypadku rachunku epsilonów, zakładamy najgorszy możliwy scenariusz dla każdej operacji. W drugim sposobie za y podstawiamy dokładną wartość, a szacujemy tylko wartość błędu względnego y za pomocą wartości eps. Korzystając z rysunku 13. warto zauważyć, że po zastosowaniu skali logarytmicznej na osi pionowej wykres jest linią prosta. Linia prosta na wykresie logarytmicznym wskazuje na eksponencjalną zależność między błędem względnym a liczbą cyfr znaczących mantysy.

9. Listing programów

Listing 1. Zadanie2.m

```
% Inicjalizacja zmiennych x, y
x = -10:0.1:10;
y = 1;

% Funkcja φ dla podwojnej precyzji x
function out = phi_double(x, y)
    out = sign(y) .* acos(x ./ (sqrt(x .^ 2 + y .^ 2)));
end

% Funkcja φ dla pojedynczej precyzji x
function out = phi_single(x, y)
    xs = single(x);
    xd = double(xs);
    out = sign(y) .* acos(xd ./ (sqrt(xd .^ 2 + y .^ 2)));
end

%wykres - φ dla podwojnej i pojedynczej precyzji x
phi_1 = phi_double(x, y);
phi_2 = phi_single(x, y);

plot(x, phi_1, 'r')
hold on
plot(x, phi_2, 'b')
xlabel('x')
ylabel('φ(x, y, z)')
title('Porownanie funkcji φ dla podw jnej i pojedynczej precyzji x')
grid on
legend('phi\_double', 'phi\_single')


% Funkcja obliczajaca blad wzgledny δ[φ(∼x, y, z)]
function out = relative_err(phi_double, phi_single)
    out = (phi_single - phi_double) ./ phi_double;
end

% Obliczenie i wykres bledu wzglednego δ[φ(∼x, y, z)]
error = relative_err(phi_double(x, y), phi_single(x, y));
figure;
plot(x, error, 'b')
xlabel('x')
ylabel('δ[φ]')
title('Blad wzgledny δ[φ]')
grid on

% funkcja obliczaj ca δ[∼x]
function out = x_err(x)
    xs = single(x);
    xd = double(xs);
    out = (xd - x) ./ x;
end

% funkcja obliczajaca wspolczynnik przenoszenia bledu wzglednego
function out = T(x, y)
    out = - x.*y^2./ (abs(y).*acos(x./(sqrt(x.^2 + y.^2))).*(x.^2+y.^2));
end

% wykres porownujacy uzyskany wynik bledu wzglednego δ[φ(∼x, y, z)]
% z przyblizeniem:
delta_2 = T(x, y) .* x_err(x);
figure;
plot(x, error, 'b')
```

```
xlabel('x')
ylabel('δ[φ]')
title('Porownanie δ[φ] z przyblizeniem')
grid on
hold on
plot(x, delta_2, 'r')
legend('Blad wzgledny', 'Przyblizenie')
```

Listing 2. Zadanie3.m

```
%Inicjalizacja zmiennych x, y
x = -10:0.1:10;
y = 1;
eps_value = eps('single')/2;

%Funkcja φ dla x podwojnej precyzji
function out = phi_double(x, y)
    out = sign(y) .* acos(x ./ (sqrt(x .^ 2 + y .^ 2)));
end

%Funkcja φ dla x pojedynczej precyzji
function out = phi_single(x, y)
    xs = single(x);
    xd = double(xs);
    out = sign(y) .* acos(xd ./ (sqrt(xd .^ 2 + y .^ 2)));
end

%Funkcja obliczajaca blad wzgledny δ[φ(~x, y, z)]
function out = relative_err(phi_double, phi_single)
    out = (phi_single - phi_double) ./ phi_double;
end

%Funkcja obliczajaca wspolczynnik przenoszenia bledu wzglednego
function out = T(x, y)
    out = -x.*y^2./(abs(y).*acos(x./(sqrt(x.^2+y.^2))).*(x.^2+y.^2));
end

%Porownanie wartosci bledow, wyznaczone w ramach Zadania 2, z oszacowaniem
error_abs = abs(relative_err(phi_double(x, y), phi_single(x, y)));
plot(x, error_abs, 'b')
xlabel('x')
ylabel('|δ[φ]|')
title('Porownanie wartosci bledu z oszacowaniem')
grid on
hold on
plot(x, abs(T(x,y))*eps_value, 'r')
legend('Wartosc bledu', 'Oszacowanie')
```

Listing 3. Zadanie5.m

```
%inicjalizacja zmiennych
r = 0.1;
theta = pi/3;
phi = -1.5*pi:0.1:1.5*pi;

%funkcja do obliczenia y korzystajac z podwojnej precyzji
function out = calculate_y_double(r, theta, phi)
    out = r .* sin(theta) .* sin(phi);
end

%funkcja do obliczenia y korzystajac z pojedynczej precyzji
function out = calculate_y_single(r, theta, phi)
    r_s = single(r);
    theta_s = single(theta);
    phi_s = single(phi);
```

```
    out = r_s .* sin(theta_s) .* sin(phi_s);
end

%wykres - φ dla podwojnej i pojedynczej precyzji x
y_2 = calculate_y_double(r, theta, phi);
y_1 = calculate_y_single(r, theta, phi);

figure;
plot(phi, y_1, 'r')
hold on
plot(phi, y_2, 'b')
xlabel('φ[rad]')
ylabel('y(r, θ, φ)')
title('Porownanie funkcji y dla podwojnej i pojedynczej precyzji')
grid on
legend('y dla pojdwojnej precyzji', 'y dla pojedynczej precyzji')


%funkcja wyznaczajaca blad wzgledny δ[~y(~r, ~θ, ~φ)]
function out = relative_err(y_2, y_1)
    out = abs((y_1 - y_2) ./ y_2);
end

%wykres przedstawiajacy blad wzgledny δ[~y(~r, ~θ, ~φ)]
error = relative_err(y_2, y_1);
figure;
plot(phi, error, 'r')
xlabel('φ[rad]')
grid on
ylabel('Blad wzgledny')
title('Blad wzgledny')

%wykres prezentujacy blad wzgledny po zastosowaniu skali log na osi y
% Wykres ilustrujacy blad wzgledny oraz jego oszacowanie na skali logarytmicznej
figure
semilogy(phi, error, 'r')
xlabel('φ[rad]')
ylabel('Blad wzgledny')
title('Blad wzgledny')
legend('Blad wzgledny','Location', 'best')
grid on




estimated_delta = estimate_delta(theta, phi);

%funkcja zwracajaca oszacowanie bledu
function out = estimate_delta(theta, phi)
    eps_s = eps('single')/2;
    out = abs(theta .* cot(theta)).*eps_s + abs(phi .* cot(phi)).* eps_s + 5.*eps_s;
end

% Wykres ilustrujacy blad wzgledny oraz jego oszacowanie na skali logarytmicznej
figure
semilogy(phi, error, 'r', phi, estimated_delta, 'b')
xlabel('φ[rad]')
ylabel('Blad')
title('Blad wzgledny i oszacowanie')
legend('Blad wzgledny', 'Oszacowanie bledu', 'Location', 'best')
grid on
```

Listing 4. Zadanie6.m

```
%inicjalizacja zmiennych
```

```
r = 0.01;
theta = pi/8;
phi = pi/4;
l = 0:0.5:60;

%funkcja zwracajaca wartosc y
function out = y(r, theta, phi)
    out = r .* sin(theta) .* phi;
end

%funkcja zwracajaca y zaburzone bledem wzglednym 2^{-l}
function out = y_err(y, l)
    out = y .* (1 + 2 .^ (-l));
end

% Wywolanie funkcji dla obliczenia dokladnej wartosci y
y_exact = y(r, theta, phi);

% Obliczanie y z bledem wzglednym
y_dist = y_err(y_exact, l);

%funkcja zwracajaca blad wzgledny
function out = error(y_acc, y_err)
    out = (y_err - y_acc) ./ y_acc;
end

% Obliczenie bledu wzglednego
error_y = error(y_exact, y_dist);

% Tworzenie wykresu
figure;
semilogy(l, error_y, 'LineWidth', 2);
xlabel('Liczba cyfr mantysy');
ylabel('Blad wzgledny');
title('Blad wzgledny wartosci y w funkcji liczby cyfr znaczacych mantysy (L)');
grid on;
```

Literatura

[1] R. Z. Morawski, materiały do przedmiotu Obliczenia inżynierskie, Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, semestr zimowy 2023/24.

[2] https://pl.wikipedia.org/wiki/IEEE754 [12.02.2024]

Wydział Elektroniki i Technik Informacyjnych

POLITECHNIKA WARSZAWSKA