Maciej Borowiec
12.03.2025

Laboratorium 1
Analiza błędów

1. Wprowadzenie

Celem ćwiczenia było zrobienie czterech zadań służących do przeprowadzenia analizy błędów.

W pierwszym zadaniu należało obliczyć wartość pochodnej funkcji w punkcie dwoma metodami, zrobić analizę błędów obliczeniowych, błędów numerycznych i błędów metody w zależności od danych oraz porównać te metody obliczania pochodnej.

W drugim zadaniu należało porównać metody sumowania liczb zmiennoprzecinkowych pojedynczej precyzji i porównać błędy każdej z metod.

Zadanie trzecie polegało na przekształceniu danych wyrażeń, by uniknąć zjawiska kancelacji dla podanych argumentów.

W czwartym zadaniu trzeba było porównać dwie sprawności kolektora słonecznego i sprawdzić czy można stwierdzić, który jest bardziej sprawny po uwzględnieniu błędów.

2. Zadania

2.1. Zadanie 1.

W tym zadaniu należało wyznaczyć wartość pochodnej funkcji używając wzoru na różnice prawostronną:

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

oraz wzoru różnic centralnych:

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h}$$

Za rozważaną funkcję należało przyjąć $f(x) = tan(x)$ w punkcie $x = 1$, a za h wartości 10^{-k}, dla $k = 0, 1, \ldots, 16$.

Następnie należało wyznaczyć błędy metody, numeryczne i obliczeniowe w zależności od h dla obu metod i porównanie ich ze sobą.

Rzeczywista wartość pochodnej tangensa została wyznaczona przy użyciu tożsamości:

$$tan'(x) = 1 + tan^2(x)$$

2.1.1 Różnica prawostronna

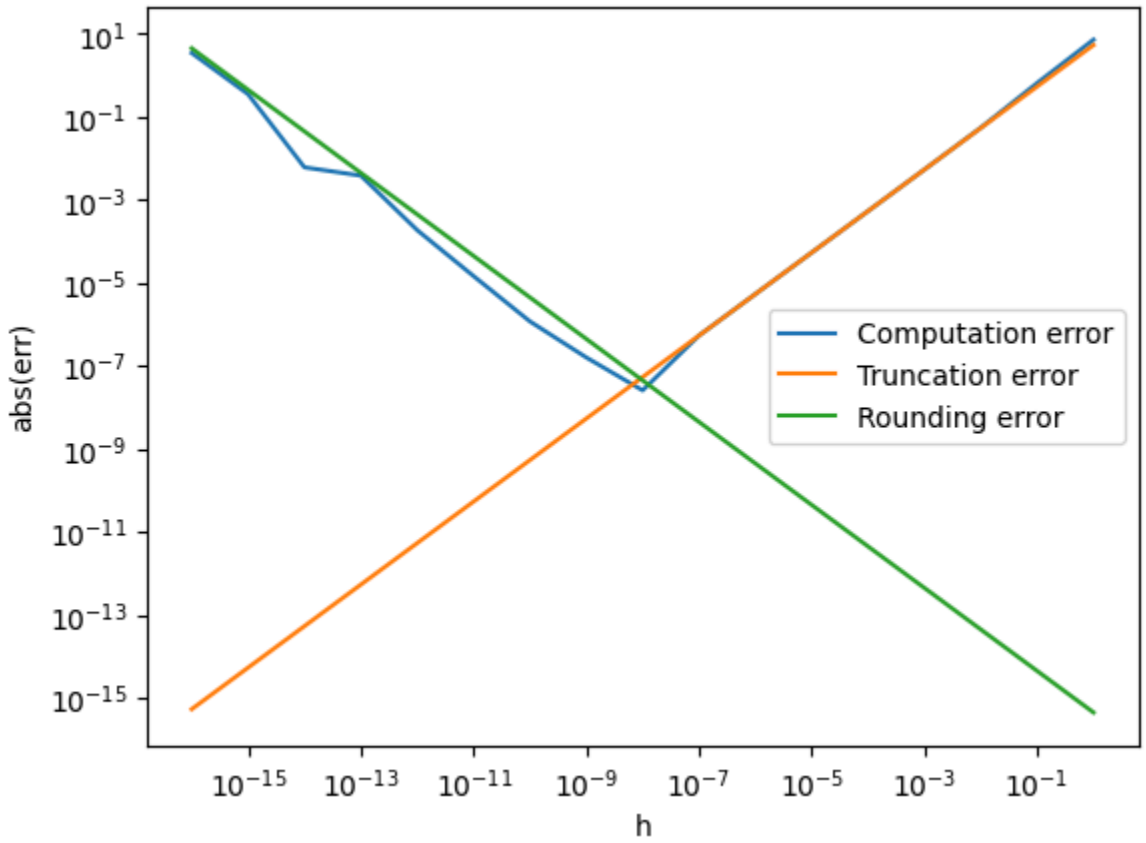
Błędy dla tej metody zostały wyznaczone następująco:

- błąd obliczeniowy $(E(h))$ to różnica wartości otrzymanej empirycznie, a rzeczywistej wartości wyliczonej z podanego wcześniej wzoru
- błąd metody można wyrazić wzorem $\frac{Mh}{2}$, gdzie $M \approx |f''(x)|$
- błąd numeryczny jest wyrażony wzorem $\frac{2\epsilon}{h}$, gdzie ϵ (tj. precyzję obliczeń) można łatwo wyznaczyć dzieląc liczbę 1.0 przez 2 aż do momentu, gdy warunek $1 + \epsilon > 1$ jest spełniony ostatni raz

Między błędami zachodzi następująca zależność:

$$E(h) \leq \frac{Mh}{2} + \frac{2\epsilon}{h}$$

Wartości bezwzględne tych błędów zostały przedstawione na wykresie poniżej. Użyto na nim skali logarytmicznej dla obu osi.



Rys. 1. Wykres wartości bezwzględnych błędów w zależności od wartości h dla różnicy prawostronnej

Na podanym wykresie (rys. 1) można zauważyć, że wartości błędów przyjmują wartości zgodne z oczekiwaniami opartymi na wzorach je opisujących. Można zauważyć, że błąd obliczeniowy osiąga minimalną wartość dla $h = 10^{-8}$. Możemy porównać wartość błędu dla tego h z wartością dla błędu dla h_{min} - teoretycznej wartości, dla której błąd powinien być najmniejszy. Wiemy, że:

$$h_{min} \approx 2\sqrt{\frac{\epsilon}{M}}$$

Przedstawmy porównanie w tabeli poniżej.

$h = 10^{-8}$	$E(h) \approx 2.5541 \cdot 10^{-8}$
$h_{min} \approx 9.1237 \cdot 10^{-9}$	$E(h) \approx 1.8288 \cdot 10^{-8}$

Tab. 1. Tabela porównująca wartości błędu obliczeniowego dla metody różnicy prawostronnej dla $h = 10^{-8}$ i dla h_{min}

Z tabeli (tab. 1) wynika, że błąd dla $h = 10^{-8}$ jest o $\sim 40\%$ większy od błędu dla h_{min}.

2.1.2 Różnice centralne

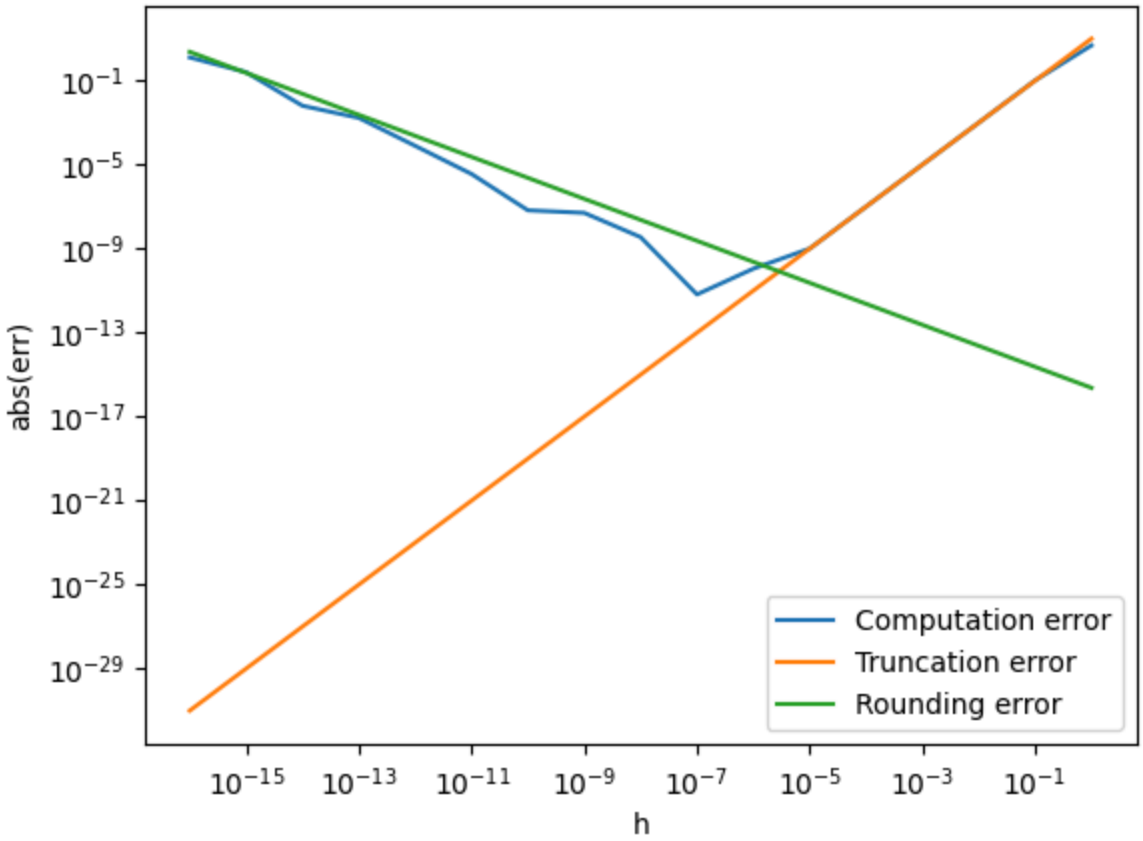
Błędy dla tej metody zostały wyznaczone następująco:

- błąd obliczeniowy $(E(h))$ to różnica wartości otrzymanej empirycznie, a rzeczywistej wartości wyliczonej z podanego wcześniej wzoru
- błąd metody można wyrazić wzorem $\frac{Mh^2}{6}$, gdzie $M \approx |f'''(x)|$
- błąd numeryczny jest wyrażony wzorem $\frac{\epsilon}{h}$, gdzie ϵ został wyznaczony jak poprzednio

Między błędami zachodzi analogiczna jak w poprzedniej metodzie zależność:

$$E(h) \leq \frac{Mh^2}{6} + \frac{\epsilon}{h}$$

Wartości bezwzględne tych błędów zostały przedstawione na wykresie poniżej. Użyto na nim skali logarytmicznej dla obu osi.



Rys. 2. Wykres wartości bezwzględnych błędów w zależności od wartości h dla różnic centralnych

Na podanym wykresie (rys. 2) można zauważyć, że wartości błędów przyjmują wartości zgodne z oczekiwaniami opartymi na wzorach je opisujących. Można zauważyć, że błąd obliczeniowy osiąga minimalną wartość dla $h = 10^{-7}$. Możemy, jak poprzednio, porównać wartość błędu dla tego h z wartością dla błędu dla h_{min}. Wiemy, że:

$$h_{min} \approx \sqrt[3]{\frac{3\epsilon}{M}}$$

Przedstawmy porównanie w tabeli poniżej.

$h = 10^{-7}$	$E(h) \approx 6.2230 \cdot 10^{-12}$
$h_{min} \approx 2.2733 \cdot 10^{-6}$	$E(h) \approx 8.6786 \cdot 10^{-11}$

Tab. 2. Tabela porównująca wartości błędu obliczeniowego dla metody różnic centralnych dla $h = 10^{-7}$ i dla h_{min}

O dziwo, w tym przypadku, z danej tabeli (tab. 2) można zauważyć, że błąd obliczeniowy dla h_{min} jest prawie 14 razy większy niż błąd dla $h = 10^{-7}$. Może to być spowodowane nieprecyzyjnym wzorem na h_{min} oraz jego komponenty lub niedokładnościami w obliczeniach komputerowych.

2.1.3 Porównanie metod

Możemy porównać te metody porównując ich błędy obliczeniowe dla h_{min}.

Metoda różnicy prawostronnej	$E(h_{min}) \approx 1.8288 \cdot 10^{-8}$
Metoda różnic centralnych	$E(h_{min}) \approx 8.6786 \cdot 10^{-11}$

Tab. 3. Tabela porównująca wartości błędów obliczeniowych dla h_{min} dla metody różnicy prawostronnej i metody różnic centralnych

Z danej tabeli (tab. 3), wynika że błąd obliczeniowy metody różnic centralnych jest ponad 210 razy mniejszy od błędy metody różnicy prawostronnej. Można zatem stwierdzić, że metoda różnic centralnych jest bardziej dokładna.

2.2. Zadanie 2.

W tym zadaniu należało zsumować listę n liczb zmiennoprzecinkowych pojedynczej precyzji na 5 różnych sposobów. Liczby zostały rozłożone losowo w przedziale $[0; 1]$ według rozkładu jednostajnego. Dla każdej z metod, za wyjątkiem metody a), należało użyć jedynie zmiennych o pojedynczej precyzji. Liczba n przyjmowała wartości 10^k, dla $k = 4, 5, \ldots, 8$.

Następnie należało porównać błędy względne tych metod. Jako wartość prawdziwą należało przyjąć wartość uzyskaną metodą **math.fsum(x)**.

2.2.1 Akumulator podwójnej precyzji

W tej metodzie należało zsumować liczby w kolejności jakiej zostały wygenerowane używając akumulatora podwójnej precyzji.

2.2.2 Akumulator pojedynczej precyzji

W tej metodzie należało zsumować liczby w kolejności jakiej zostały wygenerowane używając akumulatora pojedynczej precyzji.

2.2.3 Algorytm Kahana

W tej metodzie należało zsumować liczby w kolejności jakiej zostały wygenerowane używając algorytmu Kahana na sumowanie z kompensacją.

```python
# sum numbers using a Kahan algorithm with single precisison float accumulator
def sum_kahan(nums):
    acc = np.single(0.0)
    err = np.single(0.0)
    for num in nums:
        y = num - err
        temp = acc + y
        err = (temp-acc) - y
        acc = temp
    return acc
```

Rys. 3. Algorytm Kahana na sumowanie z kompensacją

2.2.4 Suma w kolejności rosnącej

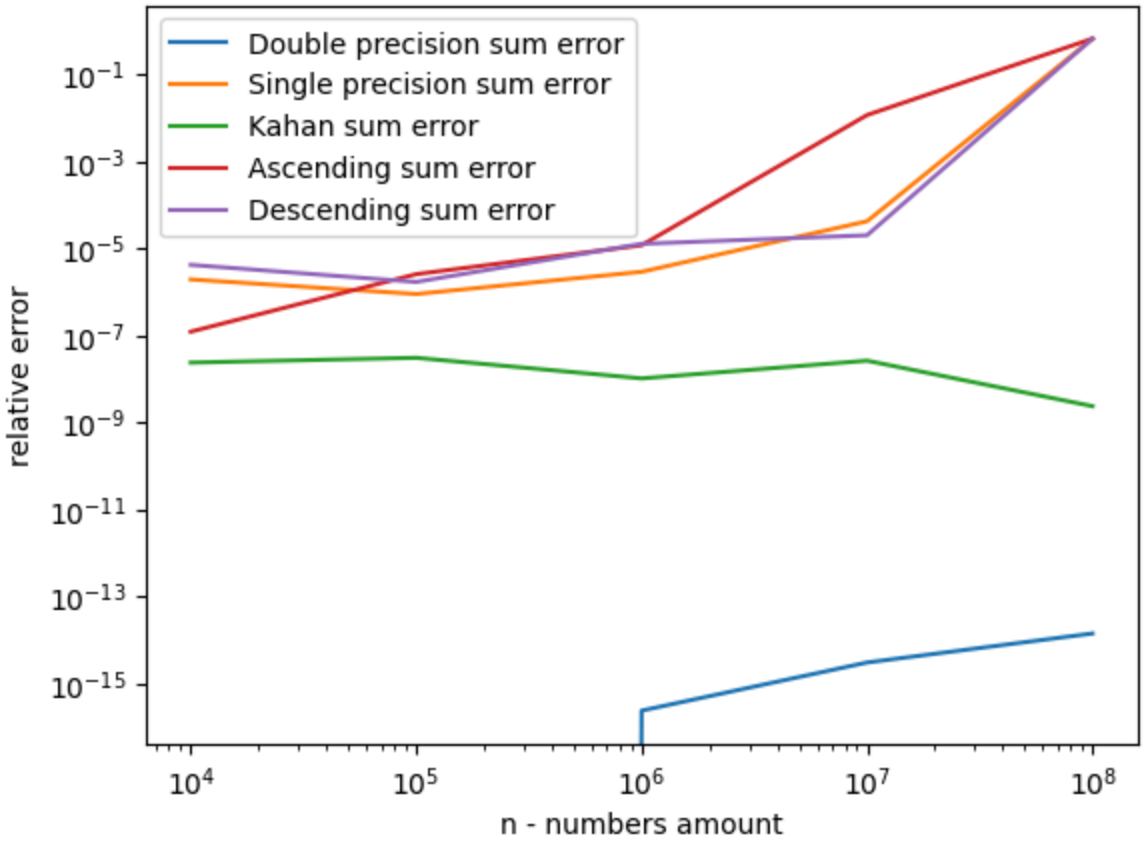
W tej metodzie należało zsumować liczby w kolejności rosnącej.

2.2.5 Suma w kolejności malejącej

W tej metodzie należało zsumować liczby w kolejności malejącej.

2.2.6 Porównanie metod

Po zsumowaniu wszystkich list o długościach przyjmujących wartości n przy użyciu różnych metod możemy porównać je ze sobą. Na wykresie poniżej przedstawiono zestawienie wartości bezwzględnych błędów względnych dla tych metod w zależności od długości listy n. Na wykresie użyto skali logarytmicznej dla obu osi.



Rys. 4. Wykres zestawiający wartości bezwzględne błędów obliczeniowych względnych dla sposobów sumowania listy w zależności od jej długości n

Na danym wykresie (rys. 4) można zauważyć jak nawet najmniejsza zmiana w sposobie sumowania liczb zmiennoprzecinkowych może zmienić wartość tej sumy. Sposoby 2, 4 i 5 jedynie zmieniają kolejność dodawania liczb, więc różnica w błędach nie jest duża, lecz nadal jest różna dla każdego z tych sposobów. Warto zauważyć, że w tym przypadku dodawanie liczb w kolejności rosnącej dla mniejszego zbioru liczb ma większą dokładność niż w kolejności malejącej, jednakże dla większego zbioru liczb zależność jest odwrotna. Algorytm Kahana radzi sobie znacząco lepiej w minimalizowaniu błędu obliczeniowego w sumowaniu. Tak jak można było się spodziewać, metoda, która użyła akumulatora o podwójnej precyzji popełniła najmniejszy błąd w obliczeniach. Ciekawe jest to, że dla mniejszych wartości n błąd jest tak mały, że nie zmieścił się na wykresie.

2.3. Zadanie 3.

W zadaniu trzecim należało przekształcić podane wyrażenia, tak, żeby uniknąć zjawiska kancelacji dla podanych argumentów. Porównałem również wartości wyrażeń wyjściowych i przekształconych z wartością prawdziwą, by zobaczyć czy faktycznie wersja przekształcona jest bardziej dokładna. Za wartość prawdziwą przyjąłem wynik podany przez serwis WolframAlpha.

2.3.1 $f_1(x) = \sqrt{x+1} - 1$ dla $x \approx 0$

Możemy przekształcić ten wzór mnożąc przez ułamek, którego licznik i mianownik jest sprzężeniem tego wyrażenia:

$$f_1(x) = (\sqrt{x+1} - 1) \cdot \frac{\sqrt{x+1}+1}{\sqrt{x+1}+1}$$

Po wymnożeniu i skróceniu:

$$f_1(x) = \frac{x}{\sqrt{x+1}+1}$$

Dzięki temu przekształceniu pozbyliśmy się odejmowania dwóch bardzo bliskich sobie liczb co powinno pomóc w pozbyciu się efektu kancelacji. Był to szczególny problem, gdyż pierwiastek zmniejszał różnicę odjemnej i odjemnika.

x	2^{-30}
Wartość prawdziwa	$4.6566128719931904 \cdot 10^{-10}$
Wartość przed przekształceniem	$4.6566128730773930 \cdot 10^{-10}$
Wartość po przekształceniu	$4.6566128719931904 \cdot 10^{-10}$

Tab. 4. Tabela porównująca wartości wyrażenia f_1 dla danego argumentu

Z tabeli (tab. 4) można zauważyć, że przekształcony wzór jest bardziej zgodny z wartością prawdziwą niż wzór przed przekształceniem. Wzór przed przekształceniem zaczyna się różnić już na 9. miejscu po przecinku, natomiast wzór przekształcony nie wykazuje różnicy dla pokazanych, pierwszych szesnastu liczb po przecinku.

2.3.2 $f_2(x) = x^2 - y^2$ dla $x \approx y$

Możemy przekształcić ten wzór używając wzoru skróconego mnożenia:

$$f_2(x) = (x+y)(x-y)$$

Dzięki temu przekształceniu pozbyliśmy się odejmowania kwadratów dwóch bardzo bliskich siebie liczb. Te kwadraty znacznie zmniejszały odstęp tych liczb od siebie. Co prawda nadal odejmujemy liczby nie podniesione do kwadratu, lecz jest to znacząco mniejszy problem niż odejmowanie tych właśnie kwadratów.

x	$1 + 2^{-29}$
y	$1 + 2^{-30}$
Wartość prawdziwa	$1.8626451518330422 \cdot 10^{-9}$
Wartość przed przekształceniem	$1.8626451492309570 \cdot 10^{-9}$
Wartość po przekształceniu	$1.8626451518330422 \cdot 10^{-9}$

Tab. 5. Tabela porównująca wartości wyrażenia f_2 dla danych argumentów

Z tabeli (tab. 5) można zauważyć, że przekształcony wzór jest bardziej zgodny z wartością prawdziwą niż wzór przed przekształceniem. Wzór przed przekształceniem zaczyna się różnić już na 8. miejscu po przecinku, natomiast wzór przekształcony nie wykazuje różnicy dla pokazanych, pierwszych szesnastu liczb po przecinku.

2.3.3 $f_3(x) = 1 - cosx$ dla $x \approx 0$

Możemy przekształcić ten wzór mnożąc przez ułamek, którego licznik i mianownik jest sprzężeniem tego wyrażenia:

$$f_3(x) = (1 - cosx) \cdot \frac{1 + cosx}{1 + cosx}$$

Po wymnożeniu i skróceniu:

$$f_3(x) = \frac{sin^2 x}{1 + cosx}$$

Dzięki temu przekształceniu pozbyliśmy się odejmowania dwóch bardzo bliskich sobie liczb co powinno pomóc w pozbyciu się efektu kancelacji.

x	2^{-30}
Wartość prawdziwa	$1.3210674732387130 \cdot 10^{-22}$
Wartość przed przekształceniem	0.0
Wartość po przekształceniu	$4.3368086899420180 \cdot 10^{-19}$

Tab. 6. Tabela porównująca wartości wyrażenia f_3 dla danego argumentu

Z tabeli (tab. 6) można zauważyć, że przekształcony wzór jest bardziej zgodny z wartością prawdziwą niż wzór przed przekształceniem. W tym przypadku jest on jednak dość mało precyzyjny, lecz na pewno bardziej precyzyjny niż przed przekształceniem, gdyż on się całkowicie wyzerował w tym przypadku.

2.3.4 $f_4(x) = cos^2 x - sin^2 x$ dla $x \approx \frac{\pi}{4}$

Możemy przekształcić ten wzór używając wzoru na cosinus kąta podwojonego:

$$f_4(x) = cos(2x)$$

Następnie podstawiając $x = \frac{\pi}{4} + \varepsilon$, gdzie ε to mała liczba:

$$f_4(x) = cos(\frac{\pi}{2} + 2\varepsilon)$$

Możemy użyć wzoru $cos(\frac{\pi}{2} + \theta) = -sin(\theta)$:

$$f_4(x) = -sin(2\varepsilon)$$

Wiemy, że dla małych θ zachodzi $sin(\theta) \approx \theta$, zatem:

$$f_4(x) \approx -2\varepsilon$$

Na koniec możemy wrócić z podstawieniem $\varepsilon = x - \frac{\pi}{4}$:

$$f_4(x) \approx -2x + \frac{\pi}{2}$$

Dzięki temu przekształceniu zastąpiliśmy odejmowanie dwóch, podniesionych do kwadratu, bardzo bliskich siebie liczb zwykłym odejmowaniem, co na pewno zredukuje efekt kancelacji.

x	$\frac{\pi}{4} + 2^{-25}$
Wartość prawdziwa	$-5.9604644775390580 \cdot 10^{-8}$
Wartość przed przekształceniem	$-5.9604644719879474 \cdot 10^{-8}$
Wartość po przekształceniu	$-5.9604644775390630 \cdot 10^{-8}$

Tab. 7. Tabela porównująca wartości wyrażenia f_4 dla danego argumentu

Z tabeli (tab. 7) można zauważyć, że przekształcony wzór jest bardziej zgodny z wartością prawdziwą niż wzór przed przekształceniem. Wzór przed przekształceniem zaczyna się różnić już na 9. miejscu po przecinku, natomiast wzór przekształcony dopiero ma pierwszą różnicę na miejscu czternastym.

2.3.5 $f_5(x) = lnx - 1$ dla $x \approx e$

Możemy przekształcić ten wzór używając zależności różnicy logarytmów i logarytmu ilorazu:

$$f_5(x) = lnx - lne = ln\frac{x}{e}$$

Możemy użyć podstawienia $x = e + \varepsilon$, gdzie ε jest małą liczbą:

$$f_5(x) = ln(\frac{e+\varepsilon}{e}) = ln(1 + \frac{\varepsilon}{e})$$

Wiemy, że dla małych x zachodzi $ln(1 + x) \approx x$, zatem:

$$f_5(x) \approx \frac{\varepsilon}{e} = \frac{x-e}{e}$$

Dzięki temu przekształceniu pozbyliśmy się operacji logarytmu, który może zmniejszyć precyzję obliczeń. Zamiast tego używamy liniowego odejmowania, które zmniejszy utratę precyzji w obliczeniach.

x	$e + 2^{-30}$
Wartość prawdziwa	$3.4261442824119880 \cdot 10^{-10}$
Wartość przed przekształceniem	$3.4261438131011346 \cdot 10^{-10}$
Wartość po przekształceniu	$3.4261442829989114 \cdot 10^{-10}$

Tab. 8. Tabela porównująca wartości wyrażenia f_5 dla danego argumentu

Z tabeli (tab. 8) można zauważyć, że przekształcony wzór jest bardziej zgodny z wartością prawdziwą niż wzór przed przekształceniem. Wzór przed przekształceniem zaczyna się różnić już na 6. miejscu po przecinku, natomiast wzór przekształcony dopiero na miejscu dziesiątym.

2.3.6 $f_6(x) = e^x - e^{-x}$ dla $x \approx 0$

Możemy przekształcić ten wzór używając szeregu Taylora:

$$e^x = 1 + x + \frac{x^2}{2} + O(x^3)$$

$$e^{-x} = 1 - x + \frac{x^2}{2} + O(x^3)$$

Dla małych x $O(x^3)$ jest pomijalnie małe, zatem:

$$f_6(x) = 1 + x + \frac{x^2}{2} - (1 - x + \frac{x^2}{2}) = 2x$$

To przekształcenie całkowicie pozbywa się funkcji wykładniczej i odejmowania podobnych liczb. Zastąpione zwykłym mnożeniem przez stałą powinno całkowicie zabezpieczyć wyrażenie przed efektem kancelacji.

x	2^{-60}
Wartość prawdziwa	$1.7347234759768070 \cdot 10^{-18}$
Wartość przed przekształceniem	0.0
Wartość po przekształceniu	$1.7347234759768070 \cdot 10^{-18}$

Tab. 9. Tabela porównująca wartości wyrażenia f_6 dla danego argumentu

Z tabeli (tab. 9) można zauważyć, że wartość przekształconego wzoru nie ma znaczącej różnicy od wartości prawdziwej. Wartość wyrażenia przed przekształceniem, natomiast, została zrównana z zerem.

2.4. Zadanie 4.

W tym zadaniu należało porównać dwie wartości sprawności kolektora słonecznego i stwierdzić czy po uwzględnieniu błędów możemy być pewni, że jeden kolektor jest lepszy od drugiego.

Efektywność η kolektora wyznaczamy wzorem:

$$\eta = K\frac{QT_d}{I}$$

gdzie:

- K - stała znana z dużą dokładnością
- Q - objętość przepływu
- T_d - różnica temperatur
- I - natężenie promieniowania

Na podstawie danego wzoru zmierzono sprawności $\eta_1 = 0.76$ i $\eta_2 = 0.70$ odpowiednio kolektorów $S1$ i $S2$. Wielkości Q, T_d oraz I zostały zmierzone z błędami podanymi w tabeli poniżej:

Kolektor	S1	S2
Q	1.5%	0.5%
T_d	1.0%	1.0%
I	3.6%	2.0%

Należy stwierdzić czy mamy pewność, że $S1$ jest bardziej sprawny niż $S2$.

2.4.1 Rozwiązanie

Sprawdzenie błędu dla kolektorów pozwoli odpowiedzieć na pytanie zadane w ćwiczeniu. Możemy to osiągnąć za pomocą prawa propagacji niepewności. Z racji takiej, że wzór na sprawność składa się tylko z iloczynów i ilorazów to błąd względny sprawności będzie równy pierwiastkowi sumy kwadratów błędów względnych wartości składających się na ten wzór. Zatem:

$$\frac{\Delta\eta_1}{\eta_1} = \sqrt{(1.5\%)^2 + (1.0\%)^2 + (3.6\%)^2} = 0.04026$$

$$\frac{\Delta\eta_2}{\eta_2} = \sqrt{(0.5\%)^2 + (1.0\%)^2 + (2.0\%)^2} = 0.02291$$

Możemy zatem obliczyć błąd bezwzględny:

$$\Delta\eta_1 = 0.04026 \cdot 0.76 = 0.030598$$

$$\Delta\eta_2 = 0.02291 \cdot 0.70 = 0.016037$$

Wyznaczmy zatem przedziały, w których te sprawności na pewno się zawierają:

$$\eta_1 \in [0.729402;\ 0.790598]$$
$$\eta_2 \in [0.683963;\ 0.716037]$$

Iloczyn tych przedziałów jest zbiorem pustym, zatem wiemy na pewno, że kolektor $S1$ jest bardziej sprawny niż kolektor $S2$.

To rozwiązanie zakłada pewną niezależność pomiędzy błędami pomiarów dla naszych danych. Ogólniejsze rozwiązanie dostaniemy, gdy użyjemy normy $||\cdot||_1$ zamiast normy $||\cdot||_2$. Wtedy otrzymamy:

$$\frac{\Delta\eta_1}{\eta_1} = |1.5\%| + |1.0\%| + |3.6\%| = 0.061$$
$$\frac{\Delta\eta_2}{\eta_2} = |0.5\%| + |1.0\%| + |2.0\%| = 0.035$$

Błąd bezwzględny:

$$\Delta\eta_1 = 0.061 \cdot 0.76 = 0.04636$$
$$\Delta\eta_2 = 0.035 \cdot 0.70 = 0.02450$$

Przedziały dla sprawności:

$$\eta_1 \in [0.71364;\ 0.80636]$$
$$\eta_2 \in [0.67550;\ 0.72450]$$

Dla ogólniejszego przypadku możemy zatem stwierdzić, że nie mamy pewności czy kolektor $S1$ bardziej sprawny niż kolektor $S2$.

5. Podsumowanie

Zadania, które należało wykonać pokazały, że trzeba być ostrożnym podczas pracy z liczbami rzeczywistymi na komputerach. Trzeba zwracać uwagę czy wybrane metody i narzędzia są najbardziej adekwatne do danego zadania, by jak najbardziej zminimalizować błędy wynikające z imperfekcji komputera.

6. Blbliografia

- Materiały zamieszczone wraz zadaniem
- Materiał z wikipedii nt. efektu kancelacji
- WolframAlpha do wyznaczenia “prawdziwych” wartości wyrażeń