

Obliczenia Naukowe - L1

Piotr Maciejończyk

17 października 2023

1 Rozpoznanie arytmetyki

1.1 Epsilon maszynowy

Należało znaleźć tzw. epsilon maszynowy - najmniejszą liczbę $macheps > 0$, taką że:

$$fl(1.0 + macheps) > 1.0 \text{ oraz } fl(1.0 + macheps) = 1 + macheps$$

Wartości epsilon maszynowego dla typów zmiennopozycyjnych w języku Julia otrzymałem na dwa sposoby:

- a) Korzystając z wbudowanej funkcji `eps()`.
- b) Metodą iteracyjną (**plik ex-1.1.jl**).

W języku C natomiast, znalazłem wartości dla odpowiednio `Float32`(float) oraz `Float64`(double) w pliku nagłówkowym **float.h**. Oto przedstawienie moich wyników:

<i>macheps</i>	<i>eps()</i>	iteracyjnie	float.h
Float16	0.000977	0.000977	————
Float32	$1.1920929 \cdot 10^{-7}$	$1.1920929 \cdot 10^{-7}$	$1.192093 \cdot 10^{-7}$
Float64	$2.220446049250313 \cdot 10^{-16}$	$2.220446049250313 \cdot 10^{-16}$	$2.220446 \cdot 10^{-16}$

Jak można zauważyć powyżej, metoda iteracyjna wyliczająca epsilon maszynowy dla typów zmiennopozycyjnych jest bardzo precyzyjna i zwraca odpowiednie wartości.

Natomiast, jaki dokładnie związek ma liczba *macheps* z precyzją arytmetyki ϵ ?

Precyzja arytmetyki (ϵ) jest górnym oszacowaniem błędu względnego zaokrąglenia liczby i w standardzie IEEE 754 przyjmuje formę:

$$\epsilon = 2^{-t} \geq \frac{|rd(x) - x|}{|x|} = |\delta|,$$

gdzie: t - liczba bitów mantysy (precyzja), x - liczba, której błąd względny chcemy ocenić.

Epsilon maszynowy (*macheps*) jest najmniejszą możliwą dodatnią liczbą większą od 1. Zgodnie ze standardem IEEE 754, na część ułamkową mantysy każdej liczby rzeczywistej jest przeznaczony $t - 1$ bitów. Zatem, aby otrzymać najmniejszą liczbę większą od 1.0, trzeba zwiększyć 1.0 o dokładnie $2^{-(t-1)}$, co jest szukaną przez nas wartością *macheps*. Można zatem zauważyć, że:

$$macheps = 2^{-(t-1)} = 2 \cdot 2^{-t} = 2 \cdot \epsilon$$

1.2 Maszynowe η

Zadaniem było iteracyjnie wyznaczyć dla wszystkich typów zmiennopozycyjnych liczbę maszynową η , czyli najmniejszą liczbę większą od 0 w danej arytmetyce fl .

Wartości epsilonu maszynowego dla typów zmiennopozycyjnych w języku Julia otrzymałem na dwa sposoby:

- a) Korzystając z wbudowanej funkcji $nextfloat(T(0.0))$ (T - typ zmiennopozycyjny).
- b) Metodą iteracyjną szukając najmniejszej dodatniej liczby, takiej że $fl(\eta + 0.0) > 0.0$ (plik **ex-1.2.jl**).

Oto tabela zawierająca moje wyniki:

η	$nextfloat(T(0.0))$	iteracyjnie
Float16	$6.0 \cdot 10^{-8}$	$6.0 \cdot 10^{-8}$
Float32	$1.0 \cdot 10^{-45}$	$1.0 \cdot 10^{-45}$
Float64	$5.0 \cdot 10^{-324}$	$5.0 \cdot 10^{-324}$

Jak można zauważyć powyżej, metoda iteracyjna wyliczająca liczbę maszynową η dla typów zmiennopozycyjnych jest bardzo precyzyjna i zwraca odpowiednie wartości.

Jaki związek ma liczba η z liczbą MIN_{sub} ?

Liczba MIN_{sub} to najmniejsza dodatnia liczba zmiennoprzecinkowa, która może być reprezentowana w danym systemie arytmetyki zmiennoprzecinkowej. Ta definicja dotyczy się również maszynowej liczby η , zatem oba te terminy są synonimiczne. Liczbę MIN_{sub} można zapisać jako:

$$MIN_{sub} = 2^{1-t} \cdot 2^{c_{min}},$$

gdzie: t - liczba bitów mantysy, c_{min} - najmniejsza wartość cechy.

1.3 Liczba (MAX)

Poleceniem było wyznaczyć iteracyjnie liczbę (MAX) dla wszystkich typów zmiennopozycyjnych w języku Julia oraz otrzymane wyniki porównać z wartościami zwracanymi przez funkcję $floatmax(T)$ i z wartościami zawartymi w pliku nagłówkowym **float.h** języka C.

Moja funkcja iteracyjnie szukała liczby MAX (czyli największej dodatniej wartości znormalizowanej dla danego typu zmiennoprzecinkowego), poprzez:

- a) Utworzenie mantysy dodając do siebie kolejno potęgi dwójki
- b) Iteracyjnie mnożąc wynik przez dwa, aż do uzyskania nieskończoności w danej arytmetyce
- c) Zwrócenie ostatniej liczby przed nieskończonością

Rozwiązanie znajduje się w pliku **ex-1.3.jl**. Znalazłem również wartości MAX dla typów *float* oraz *double* w pliku nagłówkowym języka C (są to wartości z dokładnością do 10 cyfry po przecinku). W tabeli poniżej umieściłem moje wyniki:

MAX	$floatmax(T)$	iteracyjnie	float.h
Float16	$6.55 \cdot 10^4$	$6.55 \cdot 10^4$	————
Float32	$3.4028235 \cdot 10^{38}$	$3.4028235 \cdot 10^{38}$	$3.4028234664 \cdot 10^{-38}$
Float64	$1.7976931348623157 \cdot 10^{308}$	$1.7976931348623157 \cdot 10^{308}$	$1.7976931349 \cdot 10^{308}$

Jak można zauważyć powyżej, metoda iteracyjna wyliczająca liczbę (MAX) dla typów zmiennopozycyjnych jest bardzo precyzyjna i zwraca odpowiednie wartości.

Co zwracają funkcje $floatmin(Float32)$ i $floatmin(Float64)$ i jaki jest związek zwracanych wartości z liczbą MIN_{nor} ?

Zgodnie z dokumentacją języka Julia, funkcja $floatmin(T)$ zwraca odpowiednio najmniejszą dodatnią liczbę znormalizowaną dla danego typu zmiennoprzecinkowego T . Zatem można powiedzieć, że wartością zwracaną przez tę funkcję jest $MIN_{nor} = 2^{c_{min}} = -2^{d-1} + 2$.

2 Metoda Kahana

Według Kahana, wartość *macheps* (czyli epsilon maszynowego) można uzyskać za pomocą równania:

$$macheps = fl(3 \cdot (\frac{4}{3} - 1) - 1)$$

Celem tego zadania było skorzystać z powyższej metody do obliczenia epsilon maszynowego oraz porównać otrzymane wartości z wartościami rzeczywistymi dla typów zmiennoprzecinkowych w języku Julia. Mój kod źródłowy zawarłem w pliku **ex-2.jl**, a tabela poniżej prezentuje moje wyniki:

<i>macheps</i>	<i>eps()</i>	Kahan
Float16	0.000977	-0.000977
Float32	$1.1920929 \cdot 10^{-7}$	$1.1920929 \cdot 10^{-7}$
Float64	$2.220446049250313 \cdot 10^{-16}$	$-2.220446049250313 \cdot 10^{-16}$

Jak można zauważyć powyżej, metoda Kahana wyliczająca epsilon maszynowy dla typów zmiennopozycyjnych jest bardzo precyzyjna i zwracająca poprawny, co do wartości bezwzględnej, wynik.

3 Rozmieszczenie liczb zmiennopozycyjnych

Zadanie polegało na eksperymentalnym sprawdzeniu równomiernego rozmieszczenia liczb zmiennopozycyjnych w zadanych przedziałach w arytmetyce **double**.

3.1 Przedział [1, 2]

Należało sprawdzić, czy w przedziale **[1, 2]** każda liczba zmiennopozycyjna x może być przedstawiona następująco:

$$x = 1 + k \cdot \delta,$$

gdzie:

$$k \in \{0, 1, 2, \dots, 2^{52}\}, \delta = 2^{-52}.$$

Do rozwiązania tego zadania skorzystałem z funkcji *bitstring()*, która m.in. zwraca bitową reprezentację otrzymanej liczby, aby sprawdzić, czy przy podanych parametrach liczby zmiennopozycyjne są równomiernie rozmieszczone w przedziale [1, 2]. Stworzyłem też funkcję *checkcorrectness()*, która na końcach przedziału sprawdza, czy dana wartość kroku jest dla niego prawidłowa. Zastosowałem również podejście iteracyjne i po kolei drukowałem kolejne liczby z zadanego przedziału:

[illegible]

W powyższej tabeli można zauważyć, że w każdej iteracji bitowa reprezentacja liczby zwiększa się o 1. Na podstawie tych wyników można wnioskować że liczby zmiennopozycyjne w przedziale **[1, 2]** są rozmieszczone równomiernie przy dobranych odpowiednio parametrach k oraz δ . Mój kod źródłowy znajduje się w pliku **ex-3.1.jl**.

3.2 Przedział $[\frac{1}{2}, 1]$

Jak dla przedziału $[1, 2]$, wykonujemy dokładnie takie same obliczenia dla przedziału $[\frac{1}{2}, 1]$, natomiast musimy zmienić parametr δ . Jako, że teraz działamy na przedziale dwukrotnie mniejszym, to $\delta = 2^{-53}$ (oraz oczywiście na początku $x = \frac{1}{2}$). Kod źródłowy dla tej części zadania znajduje się w pliku **ex-3.2.jl**. Tak samo jak wcześniej, iteracyjnie drukowałem bitową reprezentację kolejnych liczb:

Zatem można wywnioskować, że w przedziale $[\frac{1}{2}, 1]$ liczby zmiennoprzecinkowe również są równomiernie rozmieszczone.

W tym przypadku liczby zmiennoprzecinkowe także są równomiernie rozmieszczone w zadanym przedziale.

3.4 Dla rozpatrywanego przedziału

W arytmetyce *Float64* wartość kroku δ można obliczyć na podstawie wzoru:

$$\delta = 2^c \cdot 2^{1-t} = 2^{c_{KN} - (2^{d-1} - 1)} \cdot 2^{1-t} \stackrel{\text{Float64}}{=} 2^{c_{KN} - 1023} \cdot 2^{-52},$$

ponieważ dla *Float64*: $t = 53$, $d = 11$ (liczba bitów przeznaczona na eksponentę).

4 Błędy w arytmetyce

Celem tego zadania było znalezienie w arytmetyce *fl = Float64(double)* liczby zmiennopozycyjnej x w przedziale $1 < x < 2$, takiej że:

$$fl(x \cdot fl(\frac{1}{x})) \neq 1,$$

a także znalezienie najmniejszej takiej liczby x .

Mój program (**plik ex-4.jl**) od razu znajduje najmniejszą taką liczbę iteracyjnie. Do znalezienia rozwiązania posłużyłem się wartością *macheps*. W moim kodzie źródłowym pętla *while* wykonuje się do momentu, aż $fl(x \cdot fl(\frac{1}{x})) \neq 1$, zaczynając od $x = 1 + \text{macheps}$. Dla każdej kolejnej iteracji liczba x zwiększa się o *macheps*.

Liczba, którą podaje program, ma wartość: **1.0000000057228997**.

5 Obliczanie iloczynu skalarnego dwóch wektorów

W tym zadaniu należało stworzyć cztery algorytmy w języku Julia obliczające w arytmetyce *Float32* oraz *Float64* iloczyn skalarny dwóch wektorów o rozmiarach n :

- a) "w przód"
- b) "w tył"
- c) "od największego do najmniejszego"
- d) "od najmniejszego do największego"

Prawidłowość działania algorytmów należało przetestować dla dwóch wektorów x, y dla $n = 5$. Wektory x i y są postaci:

$$\begin{aligned} x &= [2.718281828, -3.141592654, 1.414213562, 0.5772156649, 0.3010299957] \\ y &= [1486.2497, 878366.9879, -22.37492, 4773714.647, 0.000185049] \end{aligned}$$

Wyniki trzeba było również porównać z rzeczywistą wartością iloczynu skalarnego wektorów x i y , która wynosi: $-1.00657107000000 \cdot 10^{-11}$. Wartość ta jest bardzo bliska zeru, co pozwala wnioskować, że "prawie" $x \perp y$ i wyniki algorytmów będą błędne. Kod źródłowy znajduje się w pliku **ex-5.jl**, a poniżej znajduje się tabela z wynikami zwróconymi przez mój program:

	algorytm a)	algorytm b)	algorytm c)	algorytm d)
Float32	-0.4999443	-0.4543457	-0.5	-0.5
Float64	$1.0251881368296672 \cdot 10^{-10}$	$-1.5643308870494366 \cdot 10^{-10}$	0.0	0.0

Jak można zauważyć powyżej, wyniki dla arytmetyki *Float32* znacząco odbiegają od rzeczywistej wartości iloczynu skalarnego wektorów x i y . To samo dotyczy arytmetyki *Float64*, w której jedynie wynik z wykorzystaniem algorytmu "w tył" jest jakkolwiek bliski dobrej odpowiedzi (lecz mimo wszystko podaje wynik około 15 razy mniejszy). Wniosek jaki można wyciągnąć, to taki, że kolejność wykonywania działań ma wpływ na precyzję obliczeń oraz że liczenie iloczynu skalarnego dla dwóch wektorów (prawie) prostopadłych jest zadaniem źle uwarunkowanym.

6 Błędy w arytmetyce ciąg dalszy

Zadanie dotyczyło policzenia w języku Julia w arytmetyce *Float64* wartości następujących funkcji

$$f(x) = \sqrt{x^2 + 1} - 1 \text{ oraz } g(x) = \frac{x^2}{\sqrt{x^2 + 1} + 1},$$

dla $x \in \{8^{-1}, 8^{-2}, 8^{-3}, \dots\}$.

Na podstawie uzyskanych rezultatów trzeba było również stwierdzić, która z tych funkcji zwraca bardziej wiarygodne wyniki. Na samym dole tej strony przedstawiłem otrzymane przeze mnie wyniki programu **ex-6.jl**, w którym zwracałem rezultaty funkcji f oraz g dla x ze zbioru podanego wyżej:

Co pierwsze rzuca się w oczy, to fakt, iż funkcja f dla wartości x mniejszych od 8^{-8} zwraca 0.0 jako wynik, gdzie funkcja g zachowuje się "normalnie" i dalej drukuje wiarygodne wyniki. Może to być spowodowane tym, iż należy unikać odejmowania od siebie liczb podobnej wielkości, gdyż taki proces może prowadzić do utraty liczb znaczących oraz precyzji.

W funkcji f , przy x^2 dążącym do zera, zmierzamy do sytuacji, w której odejmujemy jeden od "wartości coraz to bliższej jedynce". Natomiast w funkcji g działamy na dodawaniu i owy problem nie występuje, dlatego zwraca ona wiarygodniejsze wyniki.

x	f(x)	g(x)
8 ⁽⁻¹⁾	0.0077822185373186414	0.0077822185373187065
8 ⁽⁻²⁾	0.00012206286282867573	0.00012206286282875901
8 ⁽⁻³⁾	1.9073468138230965e-6	1.907346813826566e-6
8 ⁽⁻⁴⁾	2.9802321943606103e-8	2.9802321943606116e-8
8 ⁽⁻⁵⁾	4.656612873877393e-10	4.6566128719931904e-10
8 ⁽⁻⁶⁾	7.275957614183426e-12	7.275957614156956e-12
8 ⁽⁻⁷⁾	1.1368683772161603e-13	1.1368683772160957e-13
8 ⁽⁻⁸⁾	1.7763568394002505e-15	1.7763568394002489e-15
8 ⁽⁻⁹⁾	0.0	2.7755575615628914e-17
8 ⁽⁻¹⁰⁾	0.0	4.336808689942018e-19
8 ⁽⁻¹¹⁾	0.0	6.776263578034403e-21
8 ⁽⁻¹²⁾	0.0	1.0587911840678754e-22
8 ⁽⁻¹³⁾	0.0	1.6543612251060553e-24
8 ⁽⁻¹⁴⁾	0.0	2.5849394142282115e-26
8 ⁽⁻¹⁵⁾	0.0	4.0389678347315804e-28

Rysunek 1: Wyniki programu **ex-6.jl**

7 Przybliżanie pochodnej funkcji

Zadanie dotyczyło problemu przybliżania wartości pochodnej $f'(x)$ w punkcie x wzorem:

$$f'(x_0) \approx \tilde{f}'(x_0) = \frac{f(x_0+h) - f(x_0)}{h}.$$

Należało skorzystać z tego wzoru w języku Julia w arytmetyce *Float64* do przybliżenia wartości pochodnej funkcji:

$$f(x) = \sin x + \cos 3x,$$

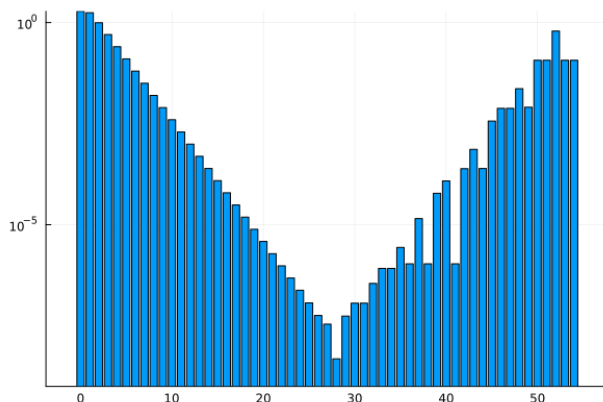
w punkcie $x_0 = 1$.

Poniżej znajdują się wyniki mojego programu **ex-7.jl**, w którym to dla $h = 2^{-n}$, gdzie $n \in \{0, 1, 2, \dots, 54\}$, obliczyłem przybliżoną wartość pochodnej funkcji f oraz błędy $|f'(x_0) - \tilde{f}'(x_0)|$.

Z wykresu na Rysunku 2. można wywnioskować, że najmniejszy błąd w aproksymacji pochodnej funkcji f występuje dla $n = 28$ (czyli dla $h = 2^{-28}$). Dla $n > 28$ błąd znowu staje się coraz większy.

Na Rysunku 3. w kolumnie " $1+h$ " można zauważyć, że wartość h stopniowo zbliża się do zera. Powinno się unikać odejmowania od siebie bardzo mocno przybliżonych wartości, co dzieje się w przypadku kiedy h przyjmuje wartości coraz to bliższe zeru. Wtedy $f(x_0+h) \approx f(x_0)$ dla $x_0 = 1$ i następuje wymieniony wyżej problem.

Dla wartości $h < 2^{-52}$ zmienna h przyjmuje nawet wartość 0. Zatem, do pewnego momentu precyzja arytmetyki osiąga swój maksymalny poziom, po czym (poprzez przykładowo odejmowanie od siebie dwóch bardzo mocno zbliżonych do siebie liczb), ta precyzja jest tracona przez utratę liczb znaczących w rachunkach komputera.



Rysunek 2: Wartość błędu aproksymacji pochodnej dla $x_0 = 1$.

h	1 + h	approximation	error
2 ⁰ (0)	2.0	2.0179892252685967	1.9010469435800585
2 ⁰ (-1)	1.5	1.8784413979316472	1.753490116243180
2 ⁰ (-2)	1.25	1.1077870952342974	0.9508448135457593
2 ⁰ (-3)	1.125	0.6232412792975817	0.5862989976690435
2 ⁰ (-4)	1.0625	0.3784080662835192	0.253457784514981
2 ⁰ (-5)	1.03125	0.24344307439754687	0.1265007927890807
2 ⁰ (-6)	1.015625	0.18009756330732785	0.0631552816187897
2 ⁰ (-7)	1.0078125	0.1484913953710958	0.03154911368255764
2 ⁰ (-8)	1.00390625	0.1327091142805159	0.01576683259197753
2 ⁰ (-9)	1.001953125	0.12402286929407085	0.007881412552170345
2 ⁰ (-10)	1.0009765625	0.12088247681186168	0.003940195122525265
2 ⁰ (-11)	1.00048828125	0.11891225046883847	0.001969968780308313
2 ⁰ (-12)	1.000244140625	0.117927223373901026	0.0009849520584721099
2 ⁰ (-13)	1.0001220703125	0.11743474981076572	0.0004924679222275685
2 ⁰ (-14)	1.00006103515625	0.11718851362093119	0.0002462319323930373
2 ⁰ (-15)	1.000030517578125	0.11706539714577957	0.00012311545724141837
2 ⁰ (-16)	1.0000152587890625	0.117003833928837255	0.155759983394246e-5
2 ⁰ (-17)	1.0000076293945312	0.11697306045971345	3.077077117529937e-5
2 ⁰ (-18)	1.0000038146972656	0.1169576106721178	1.5389378673624776e-5
2 ⁰ (-19)	1.0000019073486328	0.11694997636368498	7.694675146828866e-6
2 ⁰ (-20)	1.0000009536743164	0.11694612901192158	3.8473233834324105e-6

(a) Dla $0 \leq n \leq 20$

2 ⁰ (-21)	1.0000004768371582	0.1169442052487284	1.9235601962423127e-6
2 ⁰ (-22)	1.000000238418579	0.11694324295967817	9.612711400208096e-7
2 ⁰ (-23)	1.0000001192092896	0.11694276239722967	4.807886915192826e-7
2 ⁰ (-24)	1.0000000596046448	0.11694252118466285	2.394961446938737e-7
2 ⁰ (-25)	1.0000000298023224	0.116942398250103	1.1056156484463054e-7
2 ⁰ (-26)	1.0000000149011612	0.11694233864545822	5.6956920869239914e-8
2 ⁰ (-27)	1.0000000074505808	0.11694231629371643	3.468517827846843e-8
2 ⁰ (-28)	1.0000000037252903	0.11694228649139404	4.802855890773117e-9
2 ⁰ (-29)	1.0000000018626451	0.11694222688674927	5.480178888461751e-9
2 ⁰ (-30)	1.0000000009313226	0.11694216728210449	1.1440643366000813e-9
2 ⁰ (-31)	1.0000000004656613	0.11694216728210449	1.1440643366000813e-9
2 ⁰ (-32)	1.0000000002328306	0.11694192886352539	3.5282581276157003e-7
2 ⁰ (-33)	1.0000000001164153	0.11694145202636719	8.296621709646956e-7
2 ⁰ (-34)	1.0000000000582077	0.11694145202636719	8.296621709646956e-7
2 ⁰ (-35)	1.0000000000291838	0.11693954467773438	2.7370188037771956e-6
2 ⁰ (-36)	1.000000000014552	0.116943359375	1.0776864618478044e-6
2 ⁰ (-37)	1.000000000007276	0.1169281005859375	1.4181182600652196e-5
2 ⁰ (-38)	1.000000000003638	0.116943359375	1.0776864618478044e-6
2 ⁰ (-39)	1.000000000001819	0.11688232421875	5.9957469788152196e-5
2 ⁰ (-40)	1.0000000000009095	0.1168212890625	0.0001209926260381522
2 ⁰ (-41)	1.0000000000004547	0.116943359375	1.0776864618478044e-6
2 ⁰ (-42)	1.0000000000002274	0.1169921075	0.0002436629385381522

(b) Dla $21 \leq n \leq 42$

2 ⁰ (-43)	1.000000000001137	0.1162109375	0.0007313441885381522
2 ⁰ (-44)	1.000000000000568	0.1171875	0.0002452183114618478
2 ⁰ (-45)	1.000000000000284	0.11328125	0.003661031688538152
2 ⁰ (-46)	1.000000000000142	0.109375	0.007567281688538152
2 ⁰ (-47)	1.000000000000007	0.109375	0.007567281688538152
2 ⁰ (-48)	1.0000000000000036	0.09375	0.023192281688538152
2 ⁰ (-49)	1.0000000000000018	0.125	0.008057718311461848
2 ⁰ (-50)	1.0000000000000009	0.0	0.11694228168853815
2 ⁰ (-51)	1.0000000000000004	0.0	0.11694228168853815
2 ⁰ (-52)	1.0000000000000002	-0.5	0.6169422816885382
2 ⁰ (-53)	1.0	0.0	0.11694228168853815
2 ⁰ (-54)	1.0	0.0	0.11694228168853815

(c) Dla $43 \leq n \leq 54$ Rysunek 3: Wyniki programu **ex-7.jl** dla $x_0 = 1$.