Wojciech Pokój, 324526

Sprawozdanie z przedmiotu Analiza Numeryczna (M) Zadnie 1.4

Wrocław, 19 listopada 2021

Spis treści

1. Wprowadzenie

W 1734 roku słynny matematyk Leonhard Euler w swojej pracy zatytułowanej *"De Progressionibus harmonicis observationes"* po raz pierwszy zaproponował stałą, wyrażoną wzorem:

$$\gamma_n = (\sum_{k=1}^{n} \frac{1}{k} - \ln n) \tag{1}$$

$$\gamma = \lim_{n\to\infty} \gamma_n \approx 0.57721... \tag{2}$$

Niedługo później tą stałą nazwano *stałą Euler'a*. Jej pierwsze przybliżenie zostało obliczone z dokładnością do 5 cyfr znaczących, a na przestrzeni czasu zostały wyznaczone coraz dokładniejsze przybliżenia. Obecnie najprecyzyjniejsze składa się z 600 milionów cyfr. Do teraz jednak nie zostało pokazane czy ta stała jest niewymierna i czy jest transcendentalna i jest to jeden z ważniejszych problemów we współczesnej matematyce.

Jak się okazuje wyrazy ciągu wyznaczającego tą stałą dla odpowiednio dużego n można przybliżyć używając wzoru:

$$\gamma_n - \gamma \approx cn^{-d}, d > 0 \tag{3}$$

W tej pracy chciałbym doświadczalnie wyznaczyć takie wartości d oraz c żeby ciąg cn^{-d} najdokładniej możliwie wyznaczał wyrazy ciągu $\gamma_n - \gamma$. Obliczenia będę prowadził liczbach zmiennopozycyjnych o podwójnej precyzji (*double*).

2. Metoda obliczania

Korzystając z faktu, że logarytm dwójkowy jest funckją różnowartościową i monotonicznie rosnącą, wejściowe równanie jest równoważne:

Wojciech Pokój, 324526

Sprawozdanie z przedmiotu Analiza Numeryczna (M) Zadnie 1.4

Wrocław, 15 listopada 2021

Spis treści

1. Wprowadzenie

W 1734 roku słynny matematyk Leonhard Euler w swojej pracy zatytułowanej "De Progressionibus harmonicis observationes" po raz pierwszy zaproponował stałą, wyrażoną wzorem:

$$\gamma_n = (\sum_{k=1}^{n} \frac{1}{k} - \ln n) \tag{1}$$

$$\gamma = \lim_{n\to\infty} \gamma_n \approx 0.57721... \tag{2}$$

Niedługo później tą stałą nazwano stałą Euler'a. Jej pierwsze przybliżenie zostało obliczone z dokładnością do 5 cyfr znaczących, a na przestrzeni czasu zostały wyznaczone coraz dokładniejsze przybliżenia. Obecnie najprecyzyjniejsze składa się z 600 milionów cyfr. Do teraz jednak nie zostało pokazane czy ta stała jest niewymierna i czy jest transcendentalna i jest to jeden z ważniejszych problemów we współczesnej matematyce.

Jak się okazuje wyrazy ciągu wyznaczającego tą stałą dla odpowiednio dużego n można przybliżyć używając wzoru:

$$\gamma_n - \gamma \approx cn^{-d}, d > 0 \tag{3}$$

W tej pracy chciałbym doświadczalnie wyznaczyć takie wartości d oraz c żeby ciąg cn^{-d} najdokładniej możliwie wyznaczał wyrazy ciągu $\gamma_n - \gamma$. Obliczenia będę prowadził liczbach zmiennopozycyjnych o podwójnej precyzji (double).

2. Metoda obliczania

Korzystając z faktu, że logarytm dwójkowy jest funckją różnowartościową i monotonicznie rosnącą, wejściowe równanie jest równoważne:

$$\log(\gamma_n - \gamma) \approx \log cn^{-d} = \log c + -d \log n \tag{4}$$

Z kolei po wykonaniu podstawienia $[x/\log(n)]$:

$$\log(\gamma_{2^x} - \gamma) \approx \log c + -dx \tag{5}$$

Do przybliżenia wartości $\log c$ oraz $-d$ posłużę się liniowym równaniem regresowym dla wybranych wartości x. Równanie to służy do wyznaczenia współczynników funckji liniowej najlepiej odzwierciedlającej wykres dla wybranych punktów. W szczególności interesuje mnie znaleznienie a oraz b dla których funkcja (6) (gdzie x_i oraz y_i to wybrane punkty wykresu) będzie zwracała najmniejszą wartość.

$$f(a,b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2 \tag{6}$$

Wyznaczenia tych wartości można efektywnie wykonać poprzez wyprowadzenia pochodnych cząściowych funcji (6) względem zmiennych a oraz b oraz przyrównanie ich do zera. W szczególności do rozwiązania jest następujący układ równań:

$$f_a(a,b) = \sum_{i=1}^{n} (ax_i + b - y_i)x_i = 0 \tag{7}$$

$$f_b(a,b) = \sum_{i=1}^{n} (ax_i + b - y_i) = 0 \tag{8}$$

Czyli interesujące nas rozwiązanie to rowiązanie układu rówanań:

$$\begin{cases} a\sum_{i=0}^{n} x_i^2 + b\sum_{i=0}^{n} x_i = \sum_{i=0}^{n} x_i y_i \\ a\sum_{i=0}^{n} x_i + b\sum_{i=0}^{n} 1 = \sum_{i=0}^{n} y_i \end{cases} \tag{9}$$

Dalej ten układ równań jest równoważny równaniu:

$$\begin{bmatrix} \sum_{i=0}^{n} 1 & \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i y_i \end{bmatrix} \tag{10}$$

Niech X oraz Y będą następtującymi macierzami:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{11}$$

Zauważmy, że:

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{n} 1 & \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 \end{bmatrix} \tag{12}$$

$$X^T Y = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i y_i \end{bmatrix} \tag{13}$$

Zatem wyjściowe równanie dalej redukuje się do postaci:

$$\log(\gamma_n - \gamma) \approx \log cn^{-d} = \log c + -d \log n \tag{4}$$

Z kolei po wykonaniu podstawienia $[x/\log(n)]$:

$$\log(\gamma_{2^x} - \gamma) \approx \log c + -dx \tag{5}$$

Do przybliżenia wartości $\log c$ oraz $-d$ posłużę się liniowym równaniem regresowym dla wybranych wartości x. Równanie to służy do wyznaczenia współczynników funckji liniowej najlepiej odzwierciedlającej wykres dla wybranych punktów. W szczególności interesuje mnie znaleznienie a oraz b dla których funkcja (6) (gdzie x_i oraz y_i to wybrane punkty wykresu) będzie zwracała najmniejszą wartość.

$$f(a,b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2 \tag{6}$$

Wyznaczenia tych wartości można efektywnie wykonać poprzez wyprowadzenia pochodnych cząściowych funcji (6) względem zmiennych a oraz b oraz przyrównanie ich do zera. W szczególności do rozwiązania jest następujący układ równań:

$$f_a(a,b) = \sum_{i=1}^{n} (ax_i + b - y_i)x_i = 0 \tag{7}$$

$$f_b(a,b) = \sum_{i=1}^{n} (ax_i + b - y_i) = 0 \tag{8}$$

Czyli interesujące nas rozwiązanie to rowiązanie układu rówanań:

$$\begin{cases} a\sum_{i=0}^{n} x_i^2 + b\sum_{i=0}^{n} x_i = \sum_{i=0}^{n} x_i y_i \\ a\sum_{i=0}^{n} x_i + b\sum_{i=0}^{n} 1 = \sum_{i=0}^{n} y_i \end{cases} \tag{9}$$

Dalej ten układ równań jest równoważny równaniu:

$$\begin{bmatrix} \sum_{i=0}^{n} 1 & \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i y_i \end{bmatrix} \tag{10}$$

Niech X oraz Y będą następtującymi macierzami:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{11}$$

Zauważmy, że:

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{n} 1 & \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 \end{bmatrix} \tag{12}$$

$$X^T Y = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i y_i \end{bmatrix} \tag{13}$$

Zatem wyjściowe równanie dalej redukuje się do postaci:

2

$$X^T X \begin{bmatrix} b \\ a \end{bmatrix} = X^T Y \tag{14}$$

I w końcu do równania:

$$\begin{bmatrix} b \\ a \end{bmatrix} = (X^T X)^{-1} X^T Y \tag{15}$$

Tak wyprowadzone równanie posłuży mi do wyzanczenia doświadczalnie stałych c oraz d. Początkowy zbiór x-ów (argumentów) ograniczę do zbioru:

$$X = \{x \in N \mid \exists g, h \wedge 1 \leq g < 10 \wedge 3 \leq h < 5 \wedge n = g * 10^h\} \tag{16}$$

i będę go dopasowywać jeśli się okaże niewystarczający

3. Wyniki

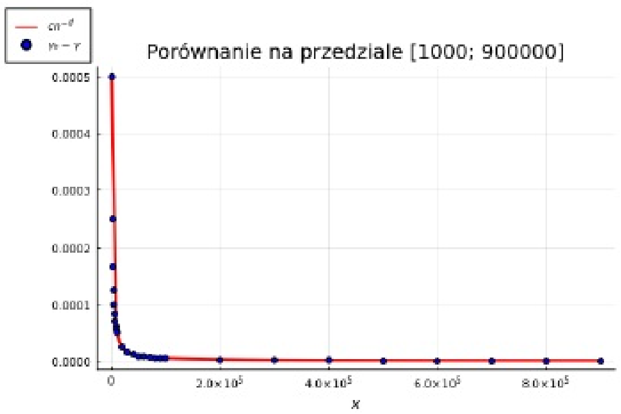
Dla tak ustalonego zbioru argumentów X obliczone c oraz d wynoszą:

$$\begin{aligned} c &= 0.499923057964 13217 \\ d &= 0.9999873486851487 \end{aligned} \tag{17}$$

Błąd względny i bezwględny przybliżenia zawierały się w przedziałach:

$$\begin{aligned} 0.0000019730 \leq |\gamma_n - \gamma - cn^{-d}| \leq 0.0001445246 \\ 0.0000001060 \leq |\frac{\gamma_n - \gamma - cn^{-d}}{\gamma_n - \gamma}| \leq 0.0000131793 \end{aligned} \tag{18}$$

Porównanie ciągu $\gamma_n - \gamma$ oraz funkcji cn^{-d} przedstawia wykres:



Jak widać na wykresie, ta funkcja kształtem bardzo dobrze oddaje ciąg $\gamma_n - \gamma$. Przetestuję tą samą metodę na kilku innych zbiorach i porównam wyniki z wynikami dla pierwotnego zbioru X.

$$X^T X \begin{bmatrix} b \\ a \end{bmatrix} = X^T Y \tag{14}$$

I w końcu do równania:

$$\begin{bmatrix} b \\ a \end{bmatrix} = (X^T X)^{-1} X^T Y \tag{15}$$

Tak wyprowadzone równanie posłuży mi do wyzanczenia doświadczalnie stałych c oraz d. Początkowy zbiór x-ów (argumentów) ograniczę do zbioru:

$$X = \{x \in N \mid \exists g, h \wedge 1 \leq g < 10 \wedge 3 \leq h < 5 \wedge n = g * 10^h\} \tag{16}$$

i będę go dopasowywać jeśli się okaże niewystarczający

3. Wyniki

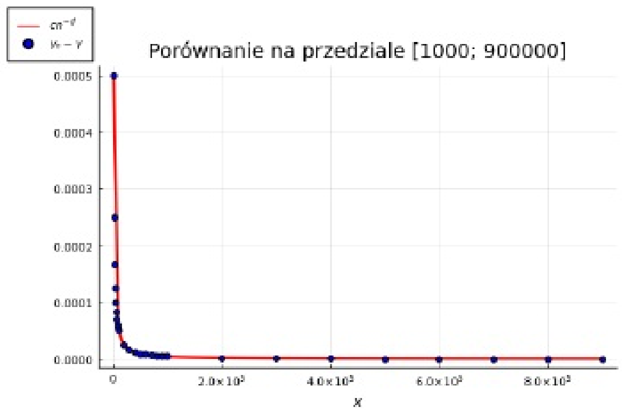
Dla tak ustalonego zbioru argumentów X obliczone c oraz d wynoszą:

$$\begin{aligned} c &= 0.499923057964 13217 \\ d &= 0.9999873486851487 \end{aligned} \tag{17}$$

Błąd względny i bezwględny przybliżenia zawierały się w przedziałach:

$$\begin{aligned} 0.0000019730 \leq |\gamma_n - \gamma - cn^{-d}| \leq 0.0001445246 \\ 0.0000001060 \leq |\frac{\gamma_n - \gamma - cn^{-d}}{\gamma_n - \gamma}| \leq 0.0000131793 \end{aligned} \tag{18}$$

Porównanie ciągu $\gamma_n - \gamma$ oraz funkcji cn^{-d} przedstawia wykres:



Jak widać na wykresie, ta funkcja kształtem bardzo dobrze oddaje ciąg $\gamma_n - \gamma$. Przetestuję tą samą metodę na kilku innych zbiorach i porównam wyniki z wynikami dla pierwotnego zbioru X.

Niech X_i będzie zbiorem:

$$X_i = \{n \in \mathbb{N} | \exists a \in \mathbb{N} \wedge a \leq 33 \wedge n = 3a + 10^i\} \tag{19}$$

3

Niech X_i będzie zbiorem:

$$X_i = \{n \in \mathbb{N} | \exists a \in \mathbb{N} \wedge a \leq 33 \wedge n = 3a + 10^i\} \tag{19}$$

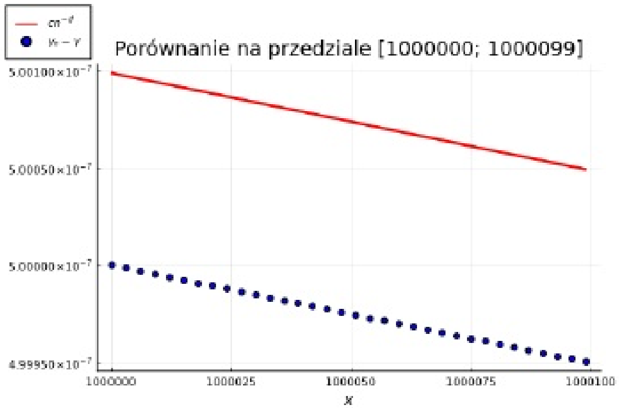
Wyniki metody regresji dla $X, X_2, X_3, X_4, X_5, X_6$ są następujące:

	C	D	min_{BB}	max_{BB}	min_{BW}	max_{BW}
X	0.4999230	0.9999873	1.060e-07	1.318e-05	1.973e-06	1.445e-04
X_2	0.4964835	0.9988178	1.281e-07	1.026e-05	1.083e-06	7.847e-05
X_3	0.4993679	0.9998409	1.843e-10	1.544e-08	2.041e-09	1.693e-07
X_4	0.4999153	0.9999834	1.599e-10	1.817e-10	2.286e-09	2.597e-09
X_5	0.4999910	0.9999985	5.014e-09	5.056e-09	8.830e-08	8.904e-08
X_6	0.5001478	1.0000071	1.358e-05	1.358e-05	2.842e-04	2.842e-04

Kolejne kolumny od lewej: Obliczone wartości c oraz d, minimalny i maksymalny błąd bezwzględny dla danego przedziału, minimlany i maksymalny błąd względny.

Z powyższych wyników można wyciągnąć kilka wniosków:

- Krótsze ale bardziej zagęszczone zbiory argumentów przybliżają lepiej wartości c oraz d w danych przydziałach.
- Dla wartości rzędu 10^4 wartości c i d są jakościowo najlepiej dobrane (najmniejsze błędy maksymalne).
- Dla wartości rzędu 10^6 powstają duże błędy względne i bezwzględne. Wynikać to może ze zjawiska utraty cyfr znaczących przy wykonywaniu odejmowania $\gamma_n - \gamma$. Dla zbioru X_6 obliczone wartości trzeba traktować ostrożnie. Problem dobitnie widać na wykresie dla przedziału X_6:



- Pomijając wątpliwej jakości wyniki dla zbioru X_6 możemy wywnioskować, że liczby c i d dążą do:

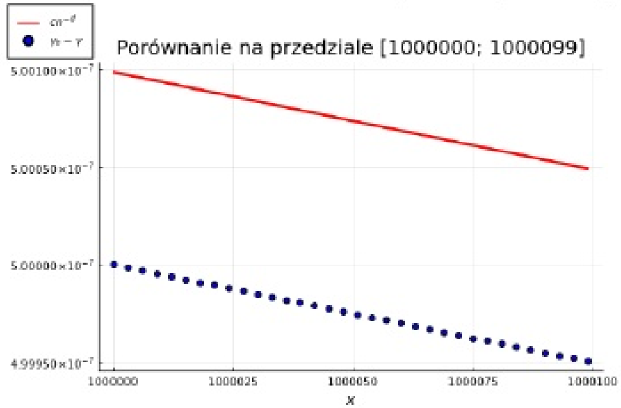
$$c \to \frac{1}{2}; \quad d \to 1 \tag{20}$$

Wyniki metody regresji dla $X, X_2, X_3, X_4, X_5, X_6$ są następujące:

	C	D	minBB	maxBB	minBW	maxBW
X	0.4999230	0.9999873	0.0000001060	0.0000131793	0.0000019730	0.0001445246
X_2	0.4964835	0.9988178	0.0000001281	0.0000102621	0.0000010830	0.0000784665
X_3	0.4993679	0.9998409	0.0000000002	0.0000000154	0.0000000020	0.0000001693
X_4	0.4999153	0.9999834	0.0000000002	0.0000000002	0.0000000023	0.0000000026
X_5	0.4999910	0.9999985	0.0000000050	0.0000000051	0.0000000883	0.0000000890
X_6	0.5001478	1.0000071	0.0000135779	0.0000135784	0.0002842072	0.0002842169

Z powyższych wyników można wyciągnąć kilka wniosków:

- Krótsze ale bardziej zagęszczone zbiory argumentów przybliżają lepiej wartości c oraz d w danych przydziałach.
- Dla wartości rzędu 10^4 wartości c i d są jakościowo najlepiej dobrane (najmniejsze błędy maksymalne).
- Dla wartości rzędu 10^6 powstają duże błędy względne i bezwzględne. Wynikać to może ze zjawiska utraty cyfr znaczących przy wykonywaniu odejmowania $\gamma_n - \gamma$. Dla zbioru X_6 obliczone wartości trzeba traktować ostrożnie. Problem dobitnie widać na wykresie dla przedziału X_6:



- Pomijając wątpliwej jakości wyniki dla zbioru X_6 możemy wywnioskować, że liczby c i d dążą do:

$$c \to \frac{1}{2}; \quad d \to 1 \tag{20}$$

4. Metoda druga

Sprawdzenie otrzymach wyników wykonam inną metodą. Zauważmy, że stałe c oraz d rozważanego ciąg cn^{-d} można potraktować jako stałą asymptotyczną zbieżności oraz wykładnik zbieżności ciągu γ_n o granicy γ.

Niech e_n będzie ciągiem błędów kolejnych przybliżeń ciągu γ_n, tj. $e_n = \gamma_n - \gamma$
Dla $n \to \infty$ mamy:

$$|e_{n+1}| \approx c|e_n|^d \tag{21}$$

$$|e_n| \approx c|e_{n-1}|^d \tag{22}$$

Stąd:

4. Metoda druga

Sprawdzenie otrzymach wyników wykonam inną metodą. Zauważmy, że stałe c oraz d rozważanego ciąg cn^{-d} można potraktować jako stałą asymptotyczną zbieżności oraz wykładnik zbieżności ciągu γ_n o granicy γ.

Niech e_n będzie ciągiem błędów kolejnych przybliżeń ciągu γ_n, tj. $e_n = \gamma_n - \gamma$
Dla $n \to \infty$ mamy:

$$|e_{n+1}| \approx c|e_n|^d \tag{21}$$

$$|e_n| \approx c|e_{n-1}|^d \tag{22}$$

Stąd:

$$\frac{|e_{n+1}|}{|e_n|} \approx \frac{c|e_n|^d}{c|e_{n-1}|} \approx |\frac{e_n}{e_{n-1}}|^d \tag{23}$$

Rozwiązując dla d:

$$d \approx \frac{log(e_{n+1}/e_n)}{log(e_n/e_{n-1})} \tag{24}$$

Ponieważ granica ciągu γ_n jest znana nie trzeba stosować dalszych przekształceń dla tej metody

5. Wyniki drugiej metody

Drugą metodę, podobnie jak pierwszą, przetestowałem na kilku przedziałach, zdefiniowanych następująco:

$$A_1 = \{10^3, 10^3 + 20, 10^3 + 40, \dots, 10^3 + 980\}$$
$$A_2 = \{10^4, 10^4 + 20, 10^4 + 40, \dots, 10^4 + 980\}$$
$$A_3 = \{10^6, 10^6 + 20, 10^6 + 40, \dots, 10^6 + 980\}$$

Wyniki dla przedziału A_1 zawiera poniższy wykres:

$$\frac{|e_{n+1}|}{|e_n|} \approx \frac{c|e_n|^d}{c|e_{n-1}|} \approx |\frac{e_n}{e_{n-1}}|^d \tag{23}$$

Rozwiązując dla d:

$$d \approx \frac{log(e_{n+1}/e_n)}{log(e_n/e_{n-1})} \tag{24}$$

Ponieważ granica ciągu γ_n jest znana nie trzeba stosować dalszych przekształceń dla tej metody

5. Usprawnienia do obliczeń

6. Wyniki z poprawioną precyzją

7. Wnioski