### MOwNiT

# Rozwiązywanie układów równań liniowych metodami bezpośrednimi

Kacper Bieniasz

11 czerwca 2024

### 1 Dane techniczne sprzętu

Obliczenia zostały wykonane na komputerze o następującej specyfikacji:

- Procesor: AMD Ryzen 7 5800U

- Pamięć RAM: 16 GB DDR4 3200 MHz (2×8GB)

- System operacyjny: Windows 11 Home x64

### 2 Opis zagadnienia oraz sposób realizacji

Celem zadania było wykorzystanie algorytmu **Gaussa** [1] [2] oraz **Thomasa** [3] do rozwiązywania układów równań w postaci macierzowej  $\mathbf{A}\mathbf{x} = \mathbf{b}$ . Porównanie czasów wykonania, uwarunkowania układów oraz dokładności wyników w zależności od użytej precyzji. Eksperyment przeprowadzam dla układów (1), (2) stosując tylko algorytm **Gaussa**, a dla (4) również algorytm **Thomasa**.

W celu otrzymywania takich samych wyników niezależnie od czynników losowych ustawiam seed korzystając z funkcji np.random.seed(42). Jest ona konieczna, ponieważ na początku losuję wektor początkowy  $\mathbf{x}$  jako permutację  $\{-1,1\}$ , na jego podstawie obliczam wektor  $\mathbf{b}$ . Korzystam z jednego z algorytmów, w ten sposób otrzymuję wektor obliczony  $\mathbf{y}$ . Obliczenia powtarzam dla różnych wielkości układu oraz precyzji (float32 i float64).

# 3 Rozważane układy

Dla poniższych układów:

 $i, j \in \{1, \dots, n\}$ , gdzie n to rozmiar macierzy  $a_{i,j}$ — to element w i—tym wierszu oraz w j—tej kolumnie.

Układ 1

$$\begin{cases} a_{1,j} = 1 \\ a_{i,j} = \frac{1}{i+j-1} \text{ dla } i \neq 1 \end{cases}$$
 (1)

Układ 2

$$\begin{cases}
a_{i,j} = \frac{2i}{j} \text{ dla } j \ge i \\
a_{i,j} = a_{j,i} \text{ dla } j < i
\end{cases}$$
(2)

Układ 3

$$\begin{cases} a_{i,i} = k \\ a_{i,i+1} = \frac{1}{i+m} \\ a_{i,i-1} = \frac{k}{i+m+1} \\ a_{i,j} = 0 \end{cases} \quad \text{dla } i > 1$$

$$(3)$$

Gdzie:

k, m- parametry

W rozważanym układzie parametry przyjmują wartości k=4 i m=5, układ prezentuje się następująco:

$$\begin{cases}
 a_{i,i} = 4 \\
 a_{i,i+1} = \frac{1}{i+5} \\
 a_{i,i-1} = \frac{4}{i+6} \\
 a_{i,j} = 0
\end{cases}$$

$$\frac{d la i > 1}{d la j < i-1 \text{ oraz } j > i+1}$$
(4)

## 4 Kryterium pomiaru błędu

Zastosowane kryterium to norma maksimum. Sprawdza się ona najlepiej w tego typu zadaniach. Wyznacza ona maksymalną różnicę między współrzędnymi wektora danego i obliczonego. Wzór opisujący, normę maksimum ma się następująco:

$$\max_{i=1,\dots,n}\{|x_i-y_i|\}\tag{5}$$

gdzie:

 $x_i - i$ -ta współrzędna zadanego wektora **x**  $y_i - i$ -ta współrzędna obliczonego wektora **y** 

# 5 Współczynnik uwarunkowania macierzy [4]

Wskaźnik uwarunkowania macierzy pozwala na oszacowanie, w jakim stopniu błąd reprezentacji numerycznej danych wejściowych wpływa na błąd wyniku. Dokładność jest zależna od iloczynu epsilonu maszynowego i wskaźnika uwarunkowania. Współczynnik uwarunkowania oblicza się ze wzoru:

$$\kappa(A) = ||A^{-1}|| \cdot ||A|| \tag{6}$$

Norma jest określona poniższym wzorem:

$$||A|| = \max_{1 < i < n} \sum_{j=1}^{n} |a_{i,j}| \tag{7}$$

gdzie:

 $A{\rm -macierz}$ kwadratowa  $n{\rm -rozmiar}$ macierzy  $a_{i,j}{\rm -element}$ macierzy o współrzędnych i,j

# 6 Otrzymane wyniki

Ponieważ układ 3 (4) jest trójdiagonalnych, wyniki eksperymentu dla niego są przedstawione w osobnej sekcji niż dla dwóch pierwszych, które są porównywane między sobą.

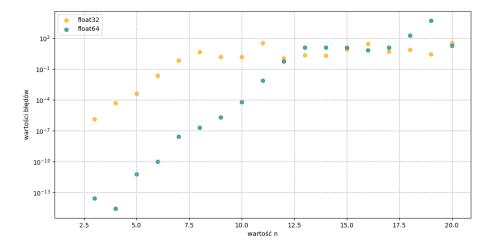
#### 6.1 Wyniki dla układu 1 oraz 2

#### Układ 1 (1)

Obliczenia ograniczyłem dla  $n \in \{3, \dots, 20\}$ . Tabela (1) zawiera zestawienie błędów w zależności od wybranej dokładności oraz rozmiaru układu. Rozważając dokładność float32, obserwujemy spadek dokładności do n=7, po którym następuje stabilizacja otrzymywanych błędów. Wartości wahają się między 1.13 do 34.69. Dla float64 błędy rosną do n=12. Dla większych n występuje podobne jak dla float32, wahanie się wyników jednak ich zakres jest znacznie większy. Użycie lepszej dokładności pozwala na dokładniejsze wyniki dla większego układu. Jednak uzyskuje akceptowalną dokładność tylko dla 4 więcej układów. Taki stan rzeczy spowodowany jest złym uwarunkowaniem macierzy. Dodatkowo wykres błędów w skali logarytmicznej (1) wizualnie przedstawia moment stabilizacji błędów.

	Dokładność		
n	float32	float64	
2	0.0000e+00	0.0000e+00	
3	1.3709e-06	2.7089e-14	
4	4.9570e-05	2.6645e-15	
5	4.0927e-04	6.1338e-12	
6	2.3369e-02	9.6068e-11	
7	7.2580e-01	2.7500e-08	
8	4.7162e+00	1.9665e-07	
9	1.5475e + 00	1.9872e-06	
10	1.5327e + 00	6.3692e-05	
11	$3.4511e{+01}$	7.3477e-03	
12	1.1266e+00	5.6456e-01	
13	2.3190e+00	1.2206e+01	
14	2.0446e+00	1.2799e+01	
15	$8.3503\mathrm{e}{+00}$	1.3026e+01	
16	$2.9455e{+01}$	6.6943e+00	
17	$5.0326 \mathrm{e}{+00}$	1.3044e+01	
18	7.5104 e + 00	1.8248e + 02	
19	2.9730e+00	5.2373e+03	
20	$3.4688e{+01}$	1.8461e + 01	

Tabela 1: Błędy dla układu 1 w zależności od dokładności i rozmiaru układu, kolorem czerwonym oznaczono punkty graniczne



Rysunek 1: Wykres błędów w skali logarytmicznej dla układu 1

#### **Układ 2** (2)

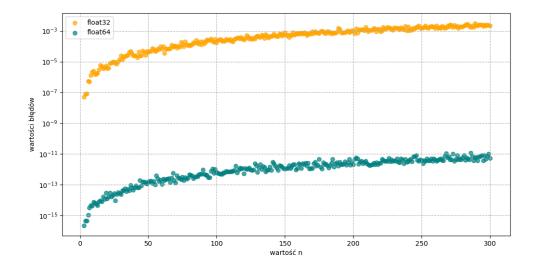
Obliczenia przeprowadziłem dla  $n \in \{3, \dots, 300\}$ . Dla  $2 \le n \le 20$  wyniki prezentuję, wykonując krok o 1, natomiast dla większych wartości n z krokiem 10. Wartości błędów znajdują się w tabelach (2, 3). Pierwsze co zauważamy to dużo lepsza dokładność dla float32 oraz float64 w porównaniu z wynikami dla układu 1 (1) jest to spowodowane dużo lepszym uwarunkowaniem układu (2). Rozważając float32, zauważamy szybki spadek dokładności, który wraz ze zwiększającym się rozmiarem układu zwalnia. Analogiczna sytuacja ma miejsce w przypadku użycia float64. Wykres błędów w skali logarytmicznej (2) pozwala łatwiej dostrzec spadek dokładności.

	Dokładność			
n	float32	float64		
2	0.0000e+00	0.0000e+00		
3	5.0664e-08	2.2204e-16		
4	7.9189e-08	4.4409e-16		
5	8.2850e-08	4.4409e-16		
6	5.3143e-07	1.1102e-15		
7	5.0194e-07	3.1086e-15		
8	1.3341e-06	4.4409e-15		
9	2.1807e-06	4.3299e-15		
10	2.4410e-06	7.5495e-15		
11	1.4380e-06	6.6613e-15		
12	1.6508e-06	5.7732e-15		
13	1.8719e-06	4.5519e-15		
14	3.0326e-06	8.1046e-15		
15	5.9677e-06	9.5479e-15		
16	5.0034e-06	8.4377e-15		
17	3.6367e-06	1.6764e-14		
18	3.9816e-06	1.1546e-14		
19	5.3337e-06	2.9754e-14		
20	3.8454e-06	1.2657e-14		

Tabela 2: Dokładność układu 2 dla  $2 \leq n \leq 51$ 

	Dokładność		
n	float32	float64	
30	1.0676e-05	4.3521e-14	
40	2.9145e-05	6.8834e-14	
50	5.1741e-05	2.4281e-13	
60	8.5000e-05	3.0165e-13	
70	6.0477e-05	2.6112e-13	
80	2.4061e-04	3.6326e-13	
90	1.6814e-04	4.3754e-13	
100	1.9357e-04	5.2669e-13	
110	2.1065e-04	7.2831e-13	
120	3.8785e-04	8.2656e-13	
130	3.1855e-04	1.1973e-12	
140	3.6399e-04	1.1915e-12	
150	4.8231e-04	1.2750e-12	
160	5.4983e-04	1.6511e-12	
170	5.6943e-04	1.9718e-12	
180	9.0764e-04	2.8568e-12	
190	9.2421e-04	1.8350e-12	
200	1.0114e-03	2.9092e-12	
210	1.2308e-03	2.5713e-12	
220	9.9002e-04	2.3277e-12	
230	1.4649e-03	3.5467e-12	
240	1.3729e-03	3.7881e-12	
250	1.3802e-03	4.1878e-12	
260	1.7794e-03	3.3167e-12	
270	2.3260e-03	5.9170e-12	
280	1.8544e-03	6.6123e-12	
290	2.5627e-03	8.0336e-12	
300	2.6331e-03	5.3975e-12	

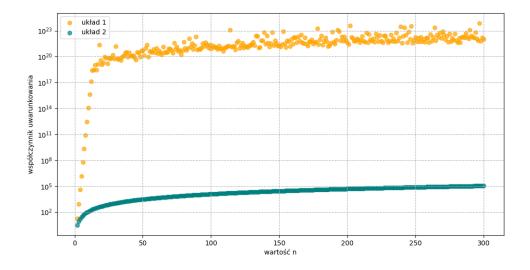
Tabela 3: Dokładność układu 2 dla  $53 \leq n \leq 101$ 



Rysunek 2: Wykres błędów w skali logarytmicznej dla układu 2

#### Współczynnik uwarunkowania obu macierzy

Analizując tabele (4, 5) oraz wykres (3) bardzo wyraźnie zauważalna jest różnica między uwarunkowaniem obu układów. Nawet dla małych rozmiarów macierzy ich uwarunkowania znacznie się od siebie różnią. W przypadku układu 1 zauważamy bardzo szybki wzrost współczynnika uwarunkowania, porównując do układu 2. Biorąc pod uwagę tabelę błędów dla układu 1 (1) oraz dla układu 2 (2, 3), zauważamy, że współczynnik uwarunkowania ma bardzo duży wpływ na dokładności wyników.



Rysunek 3: Wartości współczynników uwarunkowania w skali logarytmicznej

	Współczynnik		
	uwarunkowania		
n	układ 1	układ 2	
2	$1.8000\mathrm{e}{+01}$	3.0000e+00	
3	$8.6400 \mathrm{e}{+02}$	8.6667e+00	
4	3.7920e+04	1.6500e+01	
5	1.4420e + 06	2.6800e+01	
6	$5.6345 \mathrm{e}{+07}$	3.9667e+01	
7	2.2325e+09	$5.4571e{+01}$	
8	$8.1553e{+}10$	7.2417e + 01	
9	2.8432e+12	$9.2381e{+01}$	
10	$1.0896\mathrm{e}{+14}$	1.1473e + 02	
11	$3.9518\mathrm{e}{+15}$	1.3978e + 02	
12	$1.3645\mathrm{e}{+17}$	1.6682e+02	
13	$2.7254e{+}18$	1.9676e + 02	
14	$2.7451e{+}18$	2.2889e+02	
15	$1.0433e{+}19$	2.6329e+02	
16	$2.8954e{+}18$	3.0053e+02	
17	1.3547e + 19	3.3973e+02	
18	2.2139e+21	3.8174e + 02	
19	$3.5707\mathrm{e}{+19}$	4.2604e+02	
20	$1.2204\mathrm{e}{+19}$	4.7251e + 02	

Tabela 4: Współczynniki uwarunkowania dla 2  $\leq n \leq 25$ 

	Współczynnik		
	uwarunkowania		
n	układ 1 układ 2		
30	1.6223e + 21	1.0729e+03	
40	2.9472e + 20	1.9160e+03	
50	5.1337e + 20	3.0018e+03	
60	8.8226e + 20	4.3302e+03	
70	7.8330e + 20	5.9012e+03	
80	4.9367e + 21	7.7147e + 03	
90	6.6113e + 20	9.7709e+03	
100	8.4038e + 20	$1.2070\mathrm{e}{+04}$	
110	1.1057e + 21	1.4611e+04	
120	6.6618e + 21	1.7395e+04	
130	3.1385e+21	2.0422e+04	
140	3.0047e + 21	2.3691e+04	
150	2.0137e + 21	2.7202e+04	
160	$2.8956e{+21}$	3.0957e + 04	
170	2.7941e + 21	3.4954e + 04	
180	9.6809e + 21	3.9193e+04	
190	5.1792e + 21	4.3676e + 04	
200	4.4764e + 21	4.8401e+04	
210	8.4114e + 21	5.3368e + 04	
220	$5.0260 \mathrm{e}{+21}$	5.8578e + 04	
230	4.2158e + 21	6.4031e+04	
240	4.0346e + 21	6.9726e + 04	
250	1.7808e + 22	7.5664e + 04	
260	4.3854e + 21	8.1845e+04	
270	8.0901e+22	8.8268e + 04	
280	6.9784e + 21	9.4934e+04	
290	1.2795e + 23	$1.0184 \mathrm{e}{+05}$	
300	$1.0655\mathrm{e}{+22}$	1.0899e + 05	

Tabela 5: Współczynniki uwarunkowania dla 26  $\leq n \leq 49$ 

# 6.2 Wyniki dla układu 3 (4)

W tym przypadku rozważana macierz to macierz trójdiagonalna. Możemy zastosować dedykowany algorytm dla takiego typu macierzy, algorytm **Thomasa**. Zastosowanie dedykowanego sposobu rozwiązywania układów pozwala zmniejszyć złożoność czasową, która dla algorytmu **Gaussa** wynosi  $O(n^3)$  do O(n). Dodatkowo możemy zmniejszyć ilość potrzebnej pamięci, przechowując wartości nad przekątną, pod oraz samą przekątną.

#### Analiza błędów

Wybór metody nie ma żadnego wpływu na dokładność otrzymywanego wyniku. Oba algorytmy dostarczają wyniki o takim samym błędzie. W celu przejrzystości tabela (6) zawiera błędy z krokiem 1, natomiast w tabeli (7) z krokiem 50, zaczynając od wartości n=52.

	float32		float64	
n	Gauss	Thomas	Gauss	Thomas
2	0.0000e+00	0.0000e+00	0.0000e+00	0.0000e+00
3	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
4	1.1921e-07	1.1921e-07	0.0000e+00	0.0000e+00
5	1.1921e-07	1.1921e-07	1.1102e-16	1.1102e-16
6	1.1921e-07	1.1921e-07	1.1102e-16	1.1102e-16
7	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
8	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
9	1.1921e-07	1.1921e-07	1.1102e-16	1.1102e-16
10	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
11	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
12	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
13	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
14	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
15	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
16	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
17	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
18	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
19	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
20	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16

Tabela 6: Tabela błędów	metody Gaussa i Thomas	sa
dla $2 < n < 20$		

	float32		float64	
n	Gauss	Thomas	Gauss	Thomas
52	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
102	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
152	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
202	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
252	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
302	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
352	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
402	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
452	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
502	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
552	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
602	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
652	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
702	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
752	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
802	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
852	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
902	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16
952	1.1921e-07	1.1921e-07	2.2204e-16	2.2204e-16

Tabela 7: Tabela błędów metody Gaussa i Thomasa dla  $52 \leq n \leq 952$ 

#### Analiza czasów obliczeń

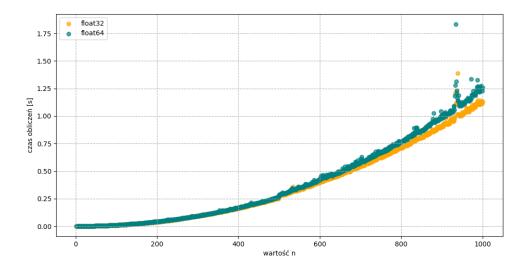
Główną zaletą stosowania algorytmu **Thomasa** jest oszczędność czasowa. Tabela (8) zawiera czasy obliczeń dla małych rozmiarów macierzy. Algorytm **Thomasa** nawet dla tak małych przypadków daje dużo lepsze wyniki. Działa  $\approx 7$  razy szybciej. Ponownie zachowując czytelność, tabela (9) zawiera czasy obu algorytmów dla  $52 \le n \le 952$  z krokiem 50. Pokazuje ona dużo wyraźniej, jak dużą przewagę ma algorytm **Thomasa** nad algorytmem **Gaussa**. Wykresy (4) i (5) pokazują krzywe czasów obliczeń dla odpowiedniej metody.

	float32		float64	
n	Gauss $[s]$	Thomas $[s]$	Gauss $[s]$	Thomas $[s]$
2	5.9700e-05	7.8000e-06	2.6700e-05	6.6000e-06
3	4.8000e-05	9.0000e-06	3.1700e-05	8.8000e-06
4	4.4300e-05	7.8000e-06	3.3400e-05	7.6000e-06
5	6.4500e-05	1.0800e-05	5.2700e-05	9.8000e-06
6	6.3500e-05	9.4000e-06	5.5800e-05	9.9000e-06
7	7.9200e-05	1.0500e-05	6.8100e-05	1.0000e-05
8	1.0130e-04	1.1500e-05	8.6500e-05	1.1500e-05
9	1.1860e-04	1.3800e-05	1.0230e-04	1.2000e-05
10	1.3420e-04	1.3200e-05	1.2120e-04	1.2700e-05
11	1.4970e-04	1.4400e-05	1.4180e-04	1.3600e-05
12	1.7210e-04	1.5800e-05	1.6490e-04	1.4600e-05
13	1.9580e-04	1.6000e-05	1.8870e-04	1.5300 e-05
14	2.3370e-04	1.7100e-05	2.1880e-04	1.6700 e-05
15	2.7130e-04	1.8900 e-05	2.4830e-04	1.7700e-05
16	2.8510e-04	1.8800e-05	2.7540e-04	1.8500e-05
17	3.1970e-04	2.1100e-05	3.1040e-04	1.9500e-05
18	3.5390e-04	2.0800e-05	3.4170e-04	2.0600e-05
19	3.8970e-04	2.3100e-05	3.8110e-04	2.1400e-05
20	4.3450e-04	2.3200e-05	4.5270e-04	2.9300e-05

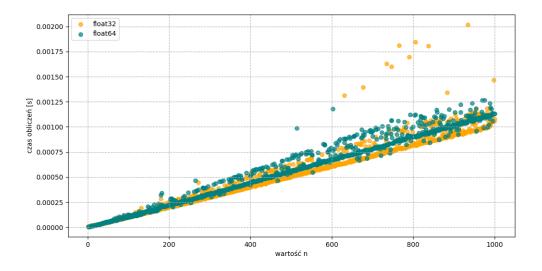
Tabela 8: Tabela czasów obliczeń metody Gaussa i Thomasa dla  $2 \leq n \leq 20$ 

	float32		float64	
n	Gauss $[s]$	Thomas $[s]$	Gauss $[s]$	Thomas $[s]$
52	2.6779e-03	7.2300e-05	2.7401e-03	5.8700e-05
102	1.0030e-02	1.1070e-04	1.0422e-02	1.0750e-04
152	2.2606e-02	1.8970e-04	2.4318e-02	1.5910e-04
202	3.9247e-02	2.1160e-04	4.2363e-02	2.0000e-04
252	6.1789e-02	2.5650 e-04	6.8413e-02	2.5470e-04
302	1.0694e-01	3.1010e-04	9.4389e-02	3.0120e-04
352	1.2432e-01	3.5590e-04	1.2955e-01	3.5020e-04
402	1.6278e-01	4.0960e-04	1.7196e-01	4.2530e-04
452	2.2129e-01	4.7260e-04	2.3604e-01	4.5150e-04
502	2.9323e-01	6.1430e-04	2.9098e-01	5.6970e-04
552	3.3633e-01	5.6150e-04	3.5082e-01	5.6010e-04
602	4.0072e-01	6.1800 e-04	4.2025e-01	6.0380e-04
652	4.8301e-01	6.6730 e-04	5.0359e-01	6.5580e-04
702	5.4935e-01	7.1260e-04	5.8661e-01	7.1600e-04
752	6.3196e-01	7.6490e-04	6.7903e-01	8.6830e-04
802	7.1457e-01	8.3980e-04	7.7072e-01	8.2190e-04
852	8.1659e-01	8.6570e-04	8.8536e-01	8.6170e-04
902	9.1227e-01	8.9570e-04	1.0224e+00	8.9860e-04
952	1.0255e+00	9.5870e-04	1.1144e+00	9.8530e-04

Tabela 9: Tabela czasów obliczeń metody Gaussa i Thomasa dla  $52 \leq n \leq 952$ 



Rysunek 4: Wykres czasów obliczeń dla metody Gaussa w zależności od dokładności



Rysunek 5: Wykres czasów obliczeń dla metody Thomasa w zależności od dokładności

### 7 Wnioski

Z analizy układów (1) i (2) zauważamy, jak duże znaczenie na dokładność obliczeń ma uwarunkowanie układu. Lepsza dokładność w przypadku układu 1 nie powoduje poprawności obliczeń dla znacznie większych rozmiarów układu. Dlatego uwarunkowanie macierzy ma kluczowe znaczenie w przypadku rozwiazywania układu.

Zastosowanie metody **Gaussa** do rozwiązywania układów równań jest stosunkowo proste i szybkie, jednak przy dużych rozmiarach jego złożoność czasowa  $O(n^3)$  może być nieakceptowalna. W takich sytuacjach, jeżeli nasza macierz jest trójdiagonalna, lepszym wyborem będzie zastosowanie algorytmu **Thomasa** przeznaczonego do takiego typu macierzy. Pozwala również zaoszczędzić pamięć potrzebną na przechowywanie macierzy.

### Literatura

```
[1] https://pre-epodreczniki.open.agh.edu.pl/tiki-index.php?page=Metoda+eliminacji+Gaussa
```

- [2] https://pl.wikipedia.org/wiki/Metoda\_eliminacji\_Gaussa
- [3] https://en.wikipedia.org/wiki/Tridiagonal\_matrix\_algorithm
- [4] https://pl.wikipedia.org/wiki/Wska%C5%BAnik\_uwarunkowania