Przegląd metod całkowania numerycznego

Kacper Kingsford

Analiza Numeryczna (M) - P2 - Zadanie P1.9
28 listopada 2020

Streszczenie

Całkowanie numeryczne to metoda numeryczna polegająca na przybliżonym obliczaniu całek oznaczonych. Proste metody polegają na przybliżeniu całki za pomocą odpowiedniej sumy ważonej wartości całkowanej funkcji w kilku punktach. Aby uzyskać dokładniejsze przybliżenie dzieli się przedział całkowania na niewielkie fragmenty. Ostateczny wynik jest sumą oszacowań całek w poszczególnych podprzedziałach. Najczęściej przedział dzieli się na równe podprzedziały, ale bardziej wyszukane algorytmy potrafią dostosowywać krok do szybkości zmienności funkcji.

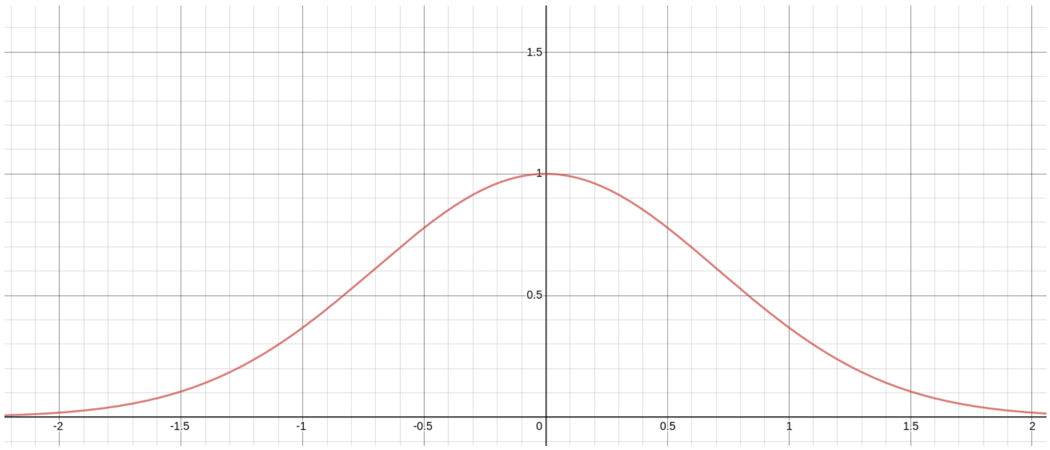
Ze względu na częstotliwość występowania całek oznaczonych w matematyce, istotne jest to, żeby móc je dokładnie aproksymować. Niniejsze sprawozdanie jest podsumowaniem eksperymentu numerycznego polegającego na obliczaniu całek metodą trapezów oraz Romberga.

Spis treści

1 Metody całkowania numerycznego

Całki oznaczone mogą być nieskończone. Spójrzmy choćby na funkcje

$$t(x) = e^{-x^2}$$



Aby wyeliminować ten problem będziemy przybliżać funkcje które całkujemy innymi, których całki da się łatwo policzyć:

$$\int_a^b f(x)\,dx \approx \int_a^b g(x)\,dx$$

Będziemy stosować interpolacje wielomianową. Weźmy:

$$g(x) = \sum_{i=1}^{n} f(x_i) \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x - x_i}$$

wtedy

$$\int_a^b f(x)\,dx \approx \int_a^b g(x)\,dx = \sum_{i=1}^{n} f(x_i) \prod_{j=0, j\neq i}^{n} \int_a^b \frac{x - x_j}{x - x_i} dx.$$

Niech

$$A_i = \int_a^b \frac{x - x_j}{x - x_i} dx$$

zatem

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{n} A_i f(x_i).$$

Wzór ten nazywamy kwadraturą.

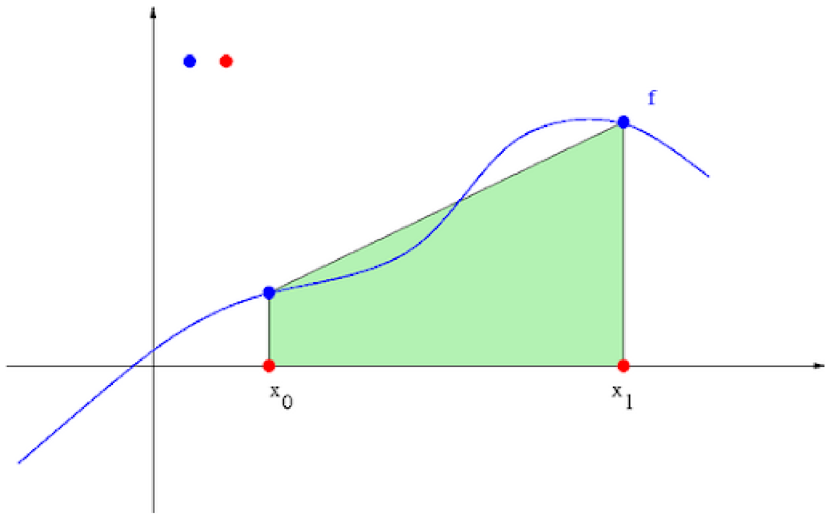
1.1 Metoda złożonych trapezów

Spójrzmy na wzór trapezów. Weźmy $n = 1 \implies x_0 = a, \;\; x_1 = b$. Wtedy

$$A_0 = \int_a^b \frac{b - x}{b - a} dx = \frac{1}{2}(b - a), A_1 = \int_a^b \frac{x - a}{b - a} dx = \frac{1}{2}(b - a)$$

$$\int_a^b f(x)\,dx \approx \frac{1}{2}(b - a)[f(a) + f(b)].$$

Łatwo sprawdzić, że wzór ten jest dokładny, tylko dla $f \in \pi_1$.



Uogólnijmy tą metode. Podzielmy przedział $[a, b]$ punktami

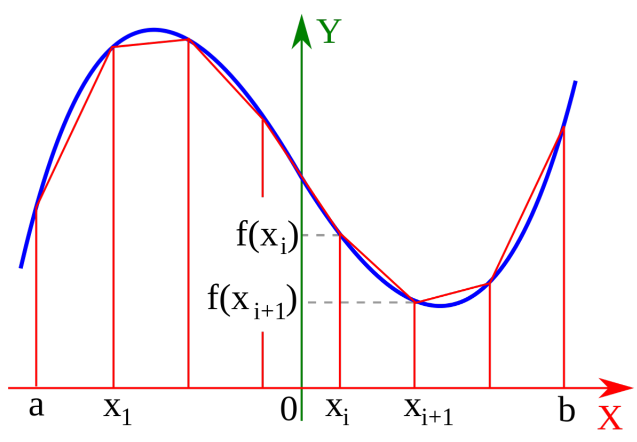
$$a = x_0 < x_1 < ... < x_{n-1} < x_n = b$$

na podprzedziały i zastosujmy wzór trapezów:

$$\int_a^b f(x)\,dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)dx \approx \frac{1}{2}\sum_{i=1}^{n}(x_i - x_{i-1})[f(x_{i-1}) + f(x_i)].$$

Wzór ten jest dokładny gdy wykres funkcji jest łamaną, której wierzchołki mają pierwszą współrzędną x_i. W dodatku, jeśli odległości między punktami są równe $x_i = a + ih, \quad h = \frac{b-a}{n}, \quad h = 0, 1, ..., n$ wzór przyjmuje postać:

$$\int_a^b f(x)dx \approx \frac{1}{2}h[f(a) + 2\sum_{i=1}^{n-1} f(a + ih) + f(b)]$$



1.2 Metoda Romberga

Oznaczmy jako $T(n, 1)$ złożony wzór trapezów. Wtedy: Dla $n = 1, 2$:

$$T(1, 1) = \frac{b - a}{2}[f(a) + f(b)]$$

$$T(2, 1) = \frac{b - a}{4}[f(a) + 2f(\frac{a + b}{b}) + f(b)]$$

Załóżmy, że f ma ciągłe pochodne wszystkich rzędów na $[a, b]$. Wtedy:

$$\int_a^b f(x)dx = \frac{1}{2}h[f(a) + 2\sum_{i=1}^{n-1} f(a + ih) + f(b)] + \sum_{i=1}^{\infty} K_i h^{2i}$$

gdzie $h = \frac{b-a}{n}$, a stałe K_i zależą od pochodnych funkcji f. Stąd wynika, że możemy użyć ekstrapolacji Richardsona, aby obliczyć przybliżenie z wyższą dokładnością. Oznaczmy wartość całki przez $I(f)$.

$$T(1, 1) = I(f) + K_1 h^2 + O(h^4)$$

$$T(2, 1) = I(f) + K_1(\frac{h}{2})^2 + O(h^4).$$

Pomijając wyrażenia $O(h^4)$ otrzymamy układ równań, który możemy rozwiązać dla K_1 oraz $I(f)$. Wartość, którą oznaczamy przez $T(2, 2)$ będzie lepszym przybliżeniem:

$$T(2, 2) = T(2, 1) + \frac{T(2, 1) - T(1, 1)}{3}.$$

Stąd wynika, że $I(f) = T(2, 2) + O(h^4)$. Załóżmy, że obliczamy kolejne przybliżenia $T(3, 1)$ przy użyciu reguły trapezów z 4 przedziałami. Tak jak poprzednio, możemy użyć ekstrapolacji Richardsona z $T(2, 1)$ i $T(3, 1)$, aby uzyskać nowe przybliżenie $T(3, 2)$, które jest dokładnie rzędu $O(h^4)$. Teraz mamy dwa przybliżenia $T(2, 2)$ oraz $T(3, 2)$, które spełniają:

$$T(2, 2) = I(f) + L_2 h^4 + O(h^6)$$

$$T(3, 2) = I(f) + L_2(\frac{h}{2})^4 + O(h^6)$$

dla pewnej stałej L_2. Wynika stąd, że możemy stosować ponownie ekstrapolacje Richardsona do tych przybliżeń, aby otrzymać nowe przybliżenie z dokładnością do $O(h^6)$. Kontynuując ten proces, możemy otrzymać tak wysoki rząd dokładności jaki chcemy. Uogólniając:

$$T(n, m) = T(n, m-1) + \frac{T(n, m-1) + T(n-1, m-1)}{4^{n-1} - 1}$$

Weźmy wzór trapezów i podstawmy $n := 2^n - 1$: i podzielmy przedział $[a, b]$ na równe podprzedziały, otrzymamy:

$$T(n, 1) = \frac{h_n}{2}[f(a) + 2\sum_{i=1}^{2^{n-1}-1} f(a + ih_n) + f(b)], \quad h_n = \frac{b-a}{2^n}.$$

Wyprowadźmy teraz rekurencje na $T(n, 1)$ rozbijając sumowanie na dwie sumy zawierające odpowiednio wyrazy nieparzyste i parzyste:

$$T(j, 1) = \frac{h_n}{2}[f(a) + \sum_{i=1}^{2^{n-2}} f(a + (2i-1)h_n) + 2\sum_{i=1}^{2^{n-2}-1} f(a + 2ih_n) + f(b)]$$

$$= \frac{h_n}{2}[f(a) + \sum_{i=1}^{2^{n-2}-1} f(a + 2ih_n) + f(b)] + \frac{h_n}{2}[2\sum_{i=1}^{2^{n-2}} f(a + (2i-1)h_n)]$$

$$= \frac{1}{2}T(n-1, 1) + h_n \sum_{i=1}^{2^{n-2}} f(a + (2i-1)h_n.$$

Sumy $T(j, 1)$ obliczamy rekurencyjnie tak, aby uniknąć wielokrotnego obliczania wartości funkcji w tych samych punktach.

Z powyższych obserwacji możemy sformułować algorytm Romberga w pseudokodzie:

Metoda Romberga

$h = b - a$

for $n = 1, 2, ..., N$ **do**

$T(n,1) = \frac{1}{2}T(n-1,1) + h\sum_{i=1}^{2^{n-2}} f(a + (2i-1)h$

for $m = 2, 3, ..., n$ **do**

$T(n,m) = T(n,m-1) + \frac{T(n,m-1)+T(n-1,m-1)}{4^{n-1}-1}$

output $n, m, T(n,m)$

end

$h = \frac{h}{2}$

end

2 Analiza błędu

Rozwińmy ze wzoru Taylora funkcję f dla $x = a$:

$$\int_a^b f(x)dx = \int_a^b \Big[f(a)+(x-a)f'(a)+\frac{(x-a)^2}{2!}f''(a)+\frac{(x-a)^3}{3!}f'''(a)+...\Big]dx = hf(a)+\frac{h^2}{2!}f'(a)+\frac{h^3}{3!}f'''(a)+...$$

Analogicznie dla $x = b$:

$$\int_a^b f(x)dx = \int_a^b \Big[f(b)+(x-b)f'(a)+\frac{(x-b)^2}{2!}f''(b)+\frac{(x-b)^3}{3!}f'''(b)+...\Big]dx = hf(b)+\frac{h^2}{2!}f'(b)+\frac{h^3}{3!}f'''(b)+...$$

Dodając obydwie równości stronami i dzieląc przez 2 otrzymujemy:

$$\int_a^b f(x)dx = \frac{h}{2}\Big[f(a)+f(b)\Big] + \frac{h^2}{4}\Big[f'(a)-f'(b)\Big] + \frac{h^3}{12}\Big[f''(a)+f''(b)\Big] + ...$$

Powtórzmy całą procedurę z taką różnicą, że będziemy rozwijać funkcje f'.
Dla $x = a$:

$$f'(x) = f'(a) + (x-a)f''(a) + \frac{(x-a)^2}{2!}f'''(a) + \frac{(x-a)^3}{3!}f^{(4)}(a) + ... \qquad (i)$$

oraz $x = b$:

$$f'(x) = f'(b) + (x-b)f''(a) + \frac{(x-b)^2}{2!}f'''(b) + \frac{(x-b)^3}{3!}f^{(4)}(b) + ... \qquad (ii)$$

Podstawiająca do (i) $x = b$ mamy:

$$f'(b) = f'(a) + hf''(a) + \frac{h^2}{2!}f'''(a) + \frac{h^3}{3!}f^{(4)}(a) + ...$$

Podstawiająca do (ii) $x = a$ mamy:

$$f'(a) = f'(b) - hf''(b) + \frac{h^2}{2!}f'''(b) - \frac{h^3}{3!}f^{(4)}(b) + ...$$

Wyznaczając z dwóch poprzednich równości $f''(a)+f''(b)$ otrzymujemy:

$$f''(a)+f''(b)=\frac{2}{h}\Big[f'(b)-f'(a)\Big]-\frac{h}{2}\Big[f'''(a)-f'''(b)\Big]-\frac{h^2}{6}\Big[f^{(4)}(a)+f^{(4)}(b)\Big]+...$$

Postepując podobnie można wyznaczyć:

$$f^{(4)}(a)+f^{(4)}(b)=\frac{2}{h}\Big[f'''(b)-f'''(a)\Big]+...$$

Zauważmy, że teraz nasza całka przybiera postać:

$$\int_a^b f(x)dx=\frac{h}{2}\Big[f(a)+f(b)\Big]+\frac{h^2}{12}\Big[f'(a)-f'(b)\Big]-\frac{h^4}{720}\Big[f'''(a)-f'''(b)\Big]+...$$

Podzielmy przedział $[a,b]$ na równe podprzedziały:

$$\int_a^b f(x)dx=\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(x)dx=h\Big[\frac{f(x_0)}{2}+f(x_1)+...+f(x_{n-1})+\frac{f(x_n)}{2}\Big]+\frac{h^2}{12}\Big[f'(x_0)-f'(x_n)\Big]-\frac{h^4}{720}\Big[f'''(x_0)-f'''(x_n)\Big]+...$$

Otrzymaliśmy złożony wzór trapezów wraz z błędem, który generuje ta metoda.

3 Opis eksperymentu oraz analiza wyników

Obliczenia zostały wykonane w języku Julia (wersja 1.5.2.) przy użyciu 64-bitowej arytmetyki (64 bity przeznaczone na reprezentację mantysy).

Przyjrzyjmy się funkcji $cosx$. Będziemy przybliżać wartość całki:

$$\int_0^{\frac{\pi}{2}} cosx\ dx$$

Wiadomo z elementarnych przekształceń, że:

$$\int_0^{\frac{\pi}{2}} cosx\ dx=sinx|_0^{\frac{\pi}{2}}=sin\frac{\pi}{2}-sin0=1$$

Spróbujmy obliczyć tą całkę opisanymi wcześniej metodami.

3.1 Porównanie wyników

Obliczmy wartość funkcji $compositeTrapezoid(n)$ dla $n=2,4,...,64$. Wyniki znajdują się w poniższej tabeli:

n	$compositeTrapezoid(n)$	$error$
2^1	$7.8539816339744828\cdot 10^{-1}$	$2.15\cdot 10^{-1}$
2^2	$9.4805944896851990\cdot 10^{-1}$	$5.19\cdot 10^{-2}$
2^3	$9.8711580097277540\cdot 10^{-1}$	$1.29\cdot 10^{-2}$
2^4	$9.9678517188616966\cdot 10^{-1}$	$3.21\cdot 10^{-3}$
2^5	$9.9919668048507226\cdot 10^{-1}$	$8.03\cdot 10^{-4}$
2^6	$9.9979919432001874\cdot 10^{-1}$	$2.01\cdot 10^{-4}$

Tablica 1: Wyniki funkcji $compositeTrapezoid(n)$ dla wybranych wartości n.

Zauważmy, że błąd funkcji bardzo wolno zbiega do zera.

Zastosujmy więc metodę Romberga. Obliczamy wartości funkcji $R(n, m)$ dla $n, m = 1, 2, ..., 6$.
Wyniki wartości funkcji znajdujących się na przekątnej tablicy Romberga:

n	$R(n, n)$
1	$7.8539816339744828 \cdot 10^{-1}$
2	1.0022798774922104
3	$9.9999156547299273 \cdot 10^{-1}$
4	1.0000000081440208
5	$9.9999999999801692 \cdot 10^{-1}$
6	1.0000000000000002

Tablica 2: Wyniki funkcji $R(n, m)$ dla wybranych wartości n.

Tablica przekątniowa błędów metody Romberga:

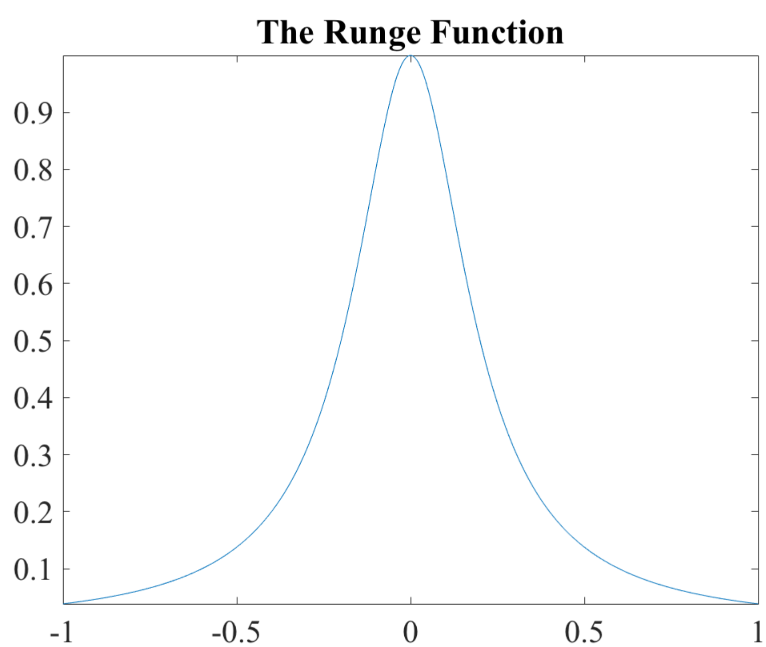
$2.15 \cdot 10^{-1}$					
$5.19 \cdot 10^{-2}$	$2.28 \cdot 10^{-3}$				
$1.29 \cdot 10^{-2}$	$1.35 \cdot 10^{-4}$	$8.43 \cdot 10^{-6}$			
$3.21 \cdot 10^{-3}$	$8.30 \cdot 10^{-6}$	$1.24 \cdot 10^{-7}$	$8.14 \cdot 10^{-9}$		
$8.03 \cdot 10^{-4}$	$5.17 \cdot 10^{-7}$	$1.90 \cdot 10^{-9}$	$2.98 \cdot 10^{-11}$	$1.98 \cdot 10^{-12}$	
$2.01 \cdot 10^{-4}$	$3.23 \cdot 10^{-8}$	$2.96 \cdot 10^{-11}$	$1.15 \cdot 10^{-13}$	$1.78 \cdot 10^{-15}$	$2.22 \cdot 10^{-16}$

Zauważmy, że najlepsze przybliżenie posiada błąd równy $\mathbf{2.22 \cdot 10^{-16}}$. Dla metody trapezów wyniósł on $\mathbf{2.01 \cdot 10^{-4}}$. Zatem ich iloraz to aż ponad $\mathbf{10^6}$.

3.2 Funkcja Rungego

Spójrzmy na kolejny przykład. Będzie to fukcja Rungego:

$$\int_{-1}^{1} \frac{1}{25x^2 + 1}$$



Obliczmy wartość funkcji $compositeTrapezoid(n)$ dla $n = 2, 4, ..., 1024$. Wyniki znajdują się w poniższej tabeli:

n	$compositeTrapezoid(n)$	$error$
2^1	$7.6923076923076927 \cdot 10^{-2}$	$4.72 \cdot 10^{-1}$
2^2	$1.0384615384615385 \cdot 10^{0}$	$4.89 \cdot 10^{-1}$
2^3	$6.5716180371352784 \cdot 10^{-1}$	$1.08 \cdot 10^{-1}$
2^4	$5.5689787382316402 \cdot 10^{-1}$	$7.54 \cdot 10^{-3}$
2^5	$5.4922232360879353 \cdot 10^{-1}$	$1.38 \cdot 10^{-4}$
2^6	$5.4931218845096019 \cdot 10^{-1}$	$4.81 \cdot 10^{-5}$
...	...	...
2^{10}	$5.4936011867707291 \cdot 10^{-1}$	$1.88 \cdot 10^{-7}$

Tablica 3: Wyniki funkcji $compositeTrapezoid(n)$ dla wybranych wartości n.

Błąd funkcji również bardzo wolno zbiega do 0.

Zastosujmy ponownie metodę Romberga. Obliczamy wartości funkcji $R(n, m)$ dla $n, m = 1, 2, ..., 10$. Wyniki wartości funkcji znajdujących się na przekątnej tablicy Romberga:

n	$R(n,n)$
1	0.0769231
2	1.35897
3	0.474801
4	0.523803
5	0.548706
6	0.549546
...	...
10	0.549360

Tablica 4: Wyniki funkcji $R(n, m)$ dla wybranych wartości n.

Tablica przekątniowa błędów metody Romberga:

$4.7 \cdot 10^{-1}$								
$4.89 \cdot 10^{-1}$	$8.09 \cdot 10^{-1}$							
$1.07 \cdot 10^{-1}$	$1.92 \cdot 10^{-2}$	$7.45 \cdot 10^{-2}$						
$7.53 \cdot 10^{-3}$	$2.58 \cdot 10^{-2}$	$2.63 \cdot 10^{-2}$	$2.55 \cdot 10^{-2}$					
$1.37 \cdot 10^{-4}$	$2.69 \cdot 10^{-3}$	$1.15 \cdot 10^{-3}$	$7.51 \cdot 10^{-4}$	$6.53 \cdot 10^{-4}$				
$4.81 \cdot 10^{-5}$	$1.81 \cdot 10^{-5}$	$1.60 \cdot 10^{-4}$	$1.81 \cdot 10^{-4}$	$1.84 \cdot 10^{-4}$	$1.85 \cdot 10^{-4}$			
$1.20 \cdot 10^{-5}$	$9.09 \cdot 10^{-9}$	$1.20 \cdot 10^{-6}$	$1.32 \cdot 10^{-6}$	$2.04 \cdot 10^{-6}$	$2.22 \cdot 10^{-6}$	$2.27 \cdot 10^{-6}$		
$3.00 \cdot 10^{-6}$	$5.21 \cdot 10^{-10}$	$5.05 \cdot 10^{-11}$	$1.90 \cdot 10^{-8}$	$1.38 \cdot 10^{-8}$	$1.19 \cdot 10^{-8}$	$1.13 \cdot 10^{-8}$	$1.12 \cdot 10^{-8}$	
$7.52 \cdot 10^{-7}$	$3.26 \cdot 10^{-11}$	$1.90 \cdot 10^{-14}$	$8.21 \cdot 10^{-13}$	$7.37 \cdot 10^{-11}$	$8.73 \cdot 10^{-11}$	$9.03 \cdot 10^{-11}$	$9.10 \cdot 10^{-11}$	$9.11 \cdot 10^{-11}$
$1.88 \cdot 10^{-7}$	$2.03 \cdot 10^{-12}$	$2.22 \cdot 10^{-16}$	$1.11 \cdot 10^{-16}$	$3.33 \cdot 10^{-15}$	$6.88 \cdot 10^{-14}$	$9.01 \cdot 10^{-14}$	$9.57 \cdot 10^{-14}$	$9.70 \cdot 10^{-14}$

Na podstawie powyższych wyników łatwo zauważyć, że metoda Romberga jest w obu przypadkach zdecydowanie szybciej zbieżna do dokładnego wyniku niż złożona metoda trapezów.

$a_1 = a$, $a_k = 1$, $a_{i+1} = floor(a_i/2)$. Niech $a = \sum_{i=1}^{k} 2^{i-1}\overline{a}_i$, $\overline{a}_k \in \{0, 1\}$, czyli zapis a w postaci binarnej. Wtedy $a_n = \sum_{i=1}^{n} 2^{i-1}\overline{a}_{i+(k-n)}$, czyli $a_n = \overline{a}_k\overline{a}_{k-1} \ \ldots \ \overline{a}_n$.

$b_1 = b$, $b_{i+1} = 2b_i \Rightarrow b_i = 2^i b$

Dowod: $\sum_{i=1,nieparzyste(a_i)}^{k} b_i = \sum_{i=1}^{k} \overline{a}_i 2^i b = b \sum_{i=1}^{k} 2^i \overline{a}_i = ab.$