Sprawozdanie – zadanie numeryczne nr. 1
Such Jarosław

Dołączony program należy uruchomić poleceniem „make run”

Treść zadania:

Napisz program wyliczający przybliżenie pochodnej ze wzorów

a) $D_h \equiv \frac{f(x+h)-f(x)}{h}$
b) $D_h \equiv \frac{f(x+h)-f(x-h)}{2h}$

Przeanalizuj jak zachowuje się błąd $|Dhf(x) - f'(x)|$ dla funkcji $f(x) = cos(x)$ oraz punktu x = 0.3 przy zmianie parametru h dla różnych typów zmiennoprzecinkowych (float, double). Wykreśl $|Dhf(x) - f'(x)|$ w funkcji h w skali logarytmicznej. Poeksperymentuj również używając innych funkcji.

Wprowadzenie:

Jako pierwsze zadanie należało napisać program, który korzystając z podanych wzorów obliczy przybliżenie pochodnej dla funkcji cos(x) w punkcie 0.3. Następnie trzeba przeanalizować błędy powstałe przy obliczeniach, dla dwóch typów zmiennoprzecinkowych: float oraz double.

Przy obliczaniu przybliżenia pochodnej za pomocą programu mamy do czynienia z dwoma rodzajami błędów. Pierwszy z nich dotyczy błędu dla dużego h (błąd systematyczny), drugi natomiast to błąd dla małego h (błąd w zaokrągleniu).

Zastosujmy teraz rozwinięcie w szereg Taylora dla funkcji f(x+h):

$$f(x+h) = f(x) + f'(x) * h + \frac{1}{2} f''(\epsilon) * h^2 \ldots$$

$$f(x+h) - f(x) - f'(x) * h = \tfrac{1}{2} f''(\epsilon) * h^2 \ldots$$

$$\frac{f(x+h) - f(x)}{h} - f'(x) \approx \frac{1}{2} f''(\epsilon) * h$$

Po obliczeniach otrzymujemy wzór na pochodną następującej postaci z błędem, którym jest wyrażenie z minusem.

$$f'(x) = \frac{f(x+h) - f(x)}{h} - \frac{1}{2} f''(\epsilon) * h$$

Teraz jeżeli uwzględnimy błąd przybliżeń otrzymujemy następujące wzory:

$$\bar{f}(x+h) = f(x+h) * (1 + \epsilon_1)$$

$$\bar{f}(x) = f(x) * (1 + \epsilon_2)$$

Jeśli posłużymy się powyższymi wzorami możemy obliczyć interesujący nas błąd:

$$\frac{\bar{f}(x+h)-\bar{f}(x)}{h}=\frac{f(x+h)*(1+\epsilon_1)-f(x)*(1-\epsilon_2)}{h}$$

$$=\frac{f(x+h)-f(x)}{h}+\frac{f(x+h)*\epsilon_1}{h}-\frac{f(x)*\epsilon_2}{h}$$

Dokonujemy niezbędnych oszacowań:

$$\approx f'(x)+\frac{1}{2}f''(x)*h+\frac{f(x+h)}{h}*\epsilon_1-\frac{f(x)}{h}*\epsilon_2$$

Dokonując dalszych obliczeń, przenosimy $f'(x)$ na przeciwną stronę i korzystamy z własności bezwzględnej oraz jej własności:

$$\left|\frac{\bar{f}(x+h)-\bar{f}(x)}{h}-f'(x)\right|\approx\left|\frac{1}{2}f''(x)*h+\frac{f(x+h)}{h}*\epsilon_1-\frac{f(x)}{h}*\epsilon_2\right|$$

Gdzie wyrażenie $\left|\frac{\bar{f}(x+h)-\bar{f}(x)}{h}-f'(x)\right|$ jest naszym szukanym błędem obliczeniowym E(h).

Jeśli teraz podstawimy wyliczone wyżej wyrażenia otrzymamy:

$$E(h)=\left|\frac{1}{2}f''(x)*h+\frac{f(x+h)}{h}*\epsilon_1-\frac{f(x)}{h}*\epsilon_2\right|\leq\frac{1}{2}|f''(x)|*h+\frac{|f(x+h)|}{h}*|\epsilon_1|+\frac{|f(x)|}{h}*|\epsilon_2|\approx\frac{1}{2}|f''(x)|+\frac{\epsilon}{h}(|f(x)|+|f(x)|)$$

Czyli po uproszczeniu:

$$\left|\frac{\bar{f}(x+h)-\bar{f}(x)}{h}-f'(x)\right|\leq\frac{1}{2}|f''(x)|*h+\frac{2\epsilon}{h}*|f(x)|$$

Czyli:

$$E(h)\leq\frac{1}{2}|f''(x)|*h+\frac{2\epsilon}{h}*|f(x)|$$

Możemy jeszcze wyliczyć $\hat{h}$, dla którego błąd obliczeniowy będzie najmniejszy. W tym celu obliczymy miejscem zerowej pochodnej $E'(x)$.

$$E'(x)=\frac{1}{2}|f''(x)|-\frac{2\epsilon*|f(x)|}{h^2}=0$$

Po prostych obliczeniach otrzymujemy wzór na h, dla którego błąd obliczeniowy jest najmniejszy:

$$\hat{h}=\sqrt{\frac{4\epsilon*|f(x)|}{|f''(x)|}}$$

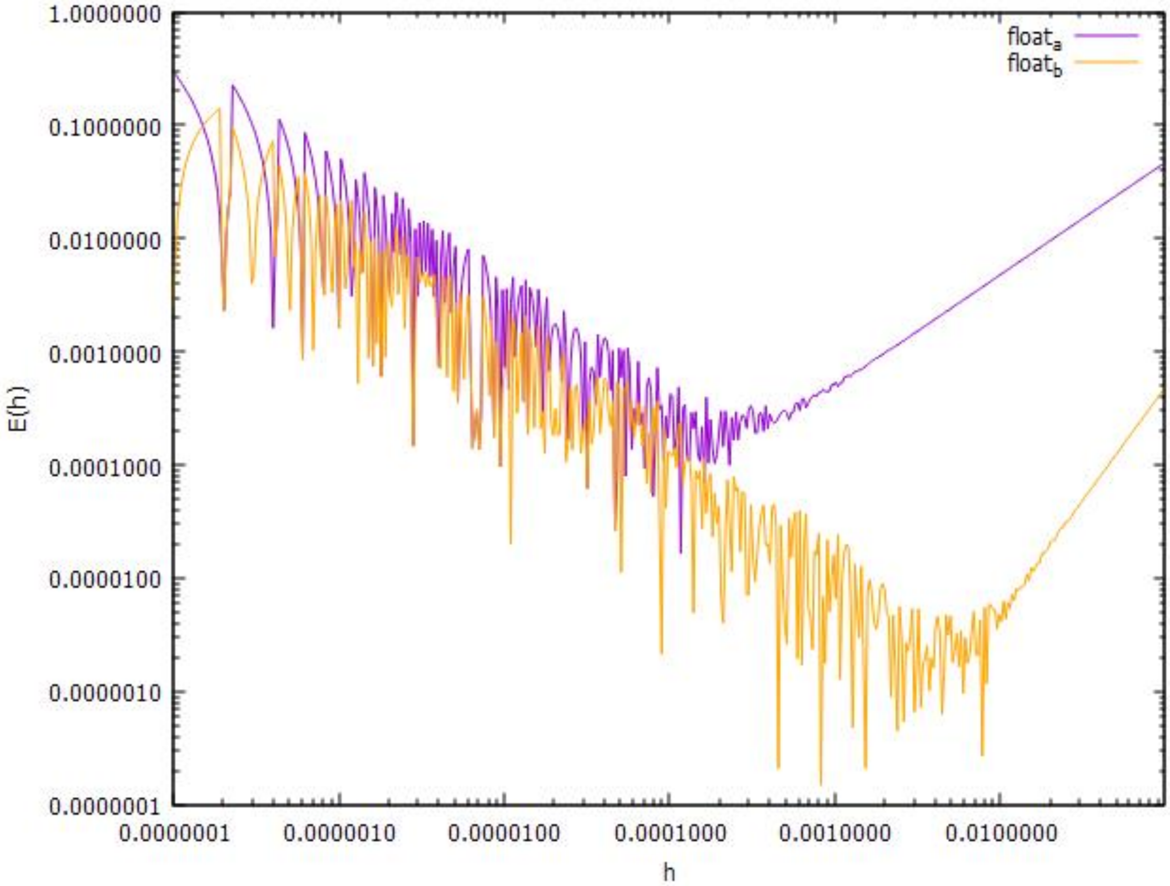
W zależności od wielkości h otrzymujemy zależności:

Dla h, takiego że: $h\gg\hat{h}\approx\frac{1}{2}|f''(x)|*h$

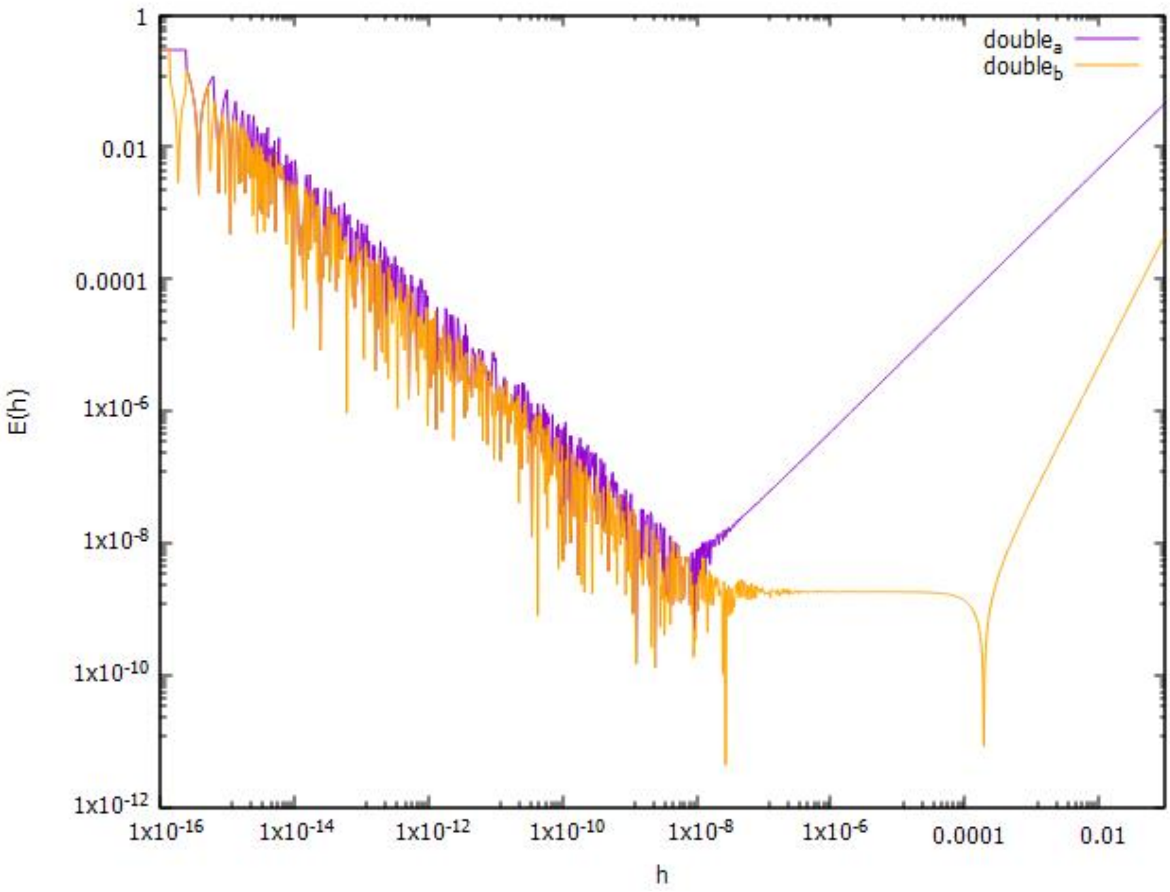
Dla h, takiego że: $h\leq\hat{h}\approx\frac{2\epsilon}{h}*|f(x)|$

Wykresy zależności błędu obliczeniowego E(h) od h prezentują się następująco:

Dla typu float



Dla typu double



Omówienie wyników:

Analizując otrzymane w wyniku działania programu wykresy możemy zauważyć iż faktycznie dzięki wzorowi, który wyliczyliśmy dla $\hat{h}$ widzimy że w jego okolicy błąd obliczeniowy będzie najmniejszy. Dodatkowo porównując obliczenia pomiędzy podpunktami a) oraz b) zauważamy że ten drugi sposób liczy przybliżenie pochodnej z mniejszym błędem. Wykresy dla typów double natomiast są podobne do tych dla typów float, ale cechują się większą precyzją spowodowaną tym że typ double przechowuje dane na większej ilości bitów niż typ float. Typ float używa do tego 32 bitów natomiast typ double 64 bitów.