Arytmetyka fl

(Dwa terminy ćwiczeń)

Zadanie 1 Rozpatrzmy arytmetykę fl opartą na 3 bitach na cechę i 3 bitach na mantysę. Dobierz b tak aby zakres liczb był możliwie zrównoważony tzn $b = 3$. Znajdź najmniejszą i największą liczbę dodatnią wśród takich liczb w arytmetyce fl. Określ największą i najmniejszą różnicę między dwoma sąsiadującymi liczbami tak określonej arytmetyki fl.

Zadanie 2 Powtórz poprzednie zadanie ale dla arytmetyki pojedynczej (23 bitów mantysy, 8 bitów cechu i 1 bit na znak, bias 127) i podwójnej precyzji (52 bitów mantysy, 11 bitów cechu i 1 bit na znak, bias 1023). Znajdź taką liczbę w arytmetyce double, że po dodaniu do jedynki i zaokrągleniu daje jedynkę.

Zadanie 3 Przybliżeniem współczynnika uwarunkowania względnego obliczania wartości funkcji $f : [a, b] \to R$ w punkcie x jest $|f'(x)||x|/|f(x)|$. Policz to uwarunkowanie dla

- $\sin(x)$,
- $\cos(x)$,
- $\exp(x)$,
- $1 + x^2$,
- $x + \sqrt{1 + x^2}$.
- $\sqrt{x + 1} - \sqrt{x}$
- $1 - \cos(x)$

Określ kiedy w arytmetyce pojedynczej precyzji obliczanie funkcji nie ma sensu, tzn. dla zaburzonych danych możemy oczekiwać że błąd względny może być większy niż jeden.

Rozwiązanie dla $f(x) = \exp(x)$. Wtedy przybliżony wsp. uwarunkowania wynosi $|x|$ czyli błąd względny $|f(x + \delta x) - f(x)|/|f(x)| \leq |x|\nu + O(\nu^2)$ pomijając człony rzędu $O(\nu^2)$ zatem dla pojedynczej precyzji jeśli $|x|$ jest rzędu $\frac{1}{\nu} = 2^{23}$ błąd względny może być większy od jeden. Można sprawdzić czy takie liczby mieszczą się w zakresie liczb pojedynczej precyzji.

Zadanie 4 Wyraź wynik w fl algorytmu sumowania liczb n liczb a_i (sumujemy po kolei) i pokaż że jest on numerycznie poprawny.

Rozwiązanie chcemy obliczyć $F(\vec{a}) = \sum_i a_i$ algorytm $s = a_1$;

$for\ k = 2 : n, s = s + a_i; endfor$

W 1 kroku $fl(s) = (a_1 + a_2)(1 + \epsilon_2)$ z $|\epsilon_2| \leq \nu$, w drugim: $fl(s) = ((a_1 + a_2)(1 + \epsilon_2) + a_3)(1 + \epsilon_3)$ z $|\epsilon_3| \leq \nu$, i tak dalej w końcu $fl(s) = (\ldots(a_1 + a_2)(1 + \epsilon_2) + \ldots a_{n-1})(1 + \epsilon_{n-1}) + a_n))(1 + \epsilon_n)$ z $|\epsilon_k| \leq \nu$ $k = 2, \ldots, n$. Zauważmy że $\pi_{j=k}^{n}(1 + \epsilon_i) = 1 + \mu_k$ z $|\mu_k| \leq (n-1)\nu + O(\nu^2)$ zatem wynik algorytmu to $F(\tilde{a})$ dla $\tilde{a} = (\tilde{a}_1, \ldots, \tilde{a}_n)^T$ z $a_k = a(1 + \mu_k)$. Widzimy, że $\|\vec{a} - \tilde{a}\|_\infty = \max_j |a_j - \tilde{a}_j| \leq \max_j |\mu_j||a_j| \leq (n-1)\nu\|\vec{a}\|_\infty + O(\nu^2)$.

Zadanie 5 Mnożenie wektora przez macierz.Czy klasyczny algorytm tzn liczymy $y_i = w_i^T * x$ dla w_i. i-tego wiersza macierzy A uzyskując $y = Ax$ klasycznym algorytmem dla iloczynu skalarnego wykonany w arytmetyce fl podwójnej precyzji zwróci $\hat{y} = (A + E) * x$ dla $E = (e_{ij})$ i $|e_{ij}| \leq n10^{-12}|a_{ij}| + O(\nu^2)$ z $A = (a_{ij})$.

Zadanie 6 Pokaż że jeśli $\hat{x}$ to zaburzony po współrzędnych wektor x tzn. $\frac{|x_i - \hat{x}_i|}{|x_i|} \leq \nu$ to $\frac{\|x_i - \hat{x}_i\|_p}{\|x_i\|_p}\nu$ dla $p = \infty$. Podaj kontrprzykłady dla $p = \infty$ że odwrotna zależność nie jest prawdziwa.

Rozwiązanie Wprost ze wzoru na norme. Kontrprzykład $x = (1, 10^{-k})^T$ i $\hat{x} = (1, 0)^T$. Norma max x równa jeden, a $\frac{\|x - \hat{x}\|_\infty}{\|x\|_\infty} = 10^{-k}$ a $\frac{|x_2 - \hat{x}_2|}{|x_2|} = 1$.

Zadanie 7 Policz z definicji uwarunkowanie względne obliczania $f(x) = 1 - x^2$ pomijając człony rzędu ν^2. Sprawdź który z algorytmów obliczania $f(x)$ wprost czy ze wzoru ($f(x) = (1-x)(1+x)$ daje mniejszy błąd względny dla $x = rd(x)$.

Zadanie 8 Ropatrzmy funkcję $F(a, b) = a^2 - b^2$.

- Policz z definicji uwarunkowanie względne obliczania $a^2 - b^2$ względem a.
- Policz wzór na błąd bezwzględny wyniku algorytmu obliczania $a^2 - b^2$: $x_1 = a * a; x_2 = b * b; w = x_1 - x_2$ wykonanego w arytmetyce zmiennopozycyjnej.
- Policz wzór na błąd bezwzględny wyniku algorytmu obliczania $a^2 - b^2$: $x_1 = (a + b); x_2 = (a - b); w = x_1 * x_2$ wykonanego w arytmetyce zmiennopozycyjnej.
- Czy oba powyższe algorytmy są numerycznie poprawne?

Zadanie 9 Policz uwarunkowanie względne obliczania wartości funkcji $\sum_k x_k$ w normie $\|\vec{y}\|_\infty = \max_k |y_k|$.

Zadanie 10 Chcemy policzyć pierwiastek kwadratowy dla liczby w fl tzn $x = 2^{c-b}(1+m)$ dla $m \in [0, 1)$. Widzimy, że realnie wystarczy policzyć pierwiastek z $y = 1 + m$ (dlaczego?).

Ile iteracji metody Herona dla $y \in [1, 2)$ z $x_0 = 1.2$ potrzeba aby mieć dokładność względną na poziomie arytmetyki podwójnej precyzji tzn. 2^{-52}.

(przypominam że pokazaliśmy że błąd metody Herona (czyli metody Newtona zastosowanej do równania $x^2 - y = 0$) spełnia: $x_{n+1} - \sqrt{y} = \frac{1}{2x_n}(x_n - \sqrt{y}^2)$.

Zadanie 11 Chcemy policzyć $1/x$ dla liczby w fl tzn $x = 2^{c-b}(1 + m)$ dla $m \in [0, 1)$. Widzimy, że realnie wystarczy policzyć odwrotność z $y = 1+m$ (dlaczego?).

Ile iteracji metody Newtona zastosowanej do funkcji $1/x - y$ z $y \in [1, 2)$ i $x_0 = 0.75$ (można ją zaimplementować bez dzielenia) potrzeba aby mieć dokładność względną na poziomie arytmetyki podwójnej precyzji tzn. 2^{-52}. Przypominam że pokazaliśmy że błąd metody Newtona zastosowanej do równania $1/x - y = 0$ wynosi $|x_{n+1} - 1/y| = y * (x_n - 1/y)^2$.

Zadanie 12 Jak policzyć z rozwinięcia w szereg $\exp(x)$ dla $x = 2^c m$ czyli liczby w fl. ($m = k * log2 + r$ z $k \in \mathbb{N}, r \in [0, log2)$, $\exp(m) = 2^k \exp(r)$).