

Sprawozdanie z obliczeń naukowych i metod numerycznych

lista 2

Bartosz Chodorowski <166142@student.pwr.wroc.pl>

Wrocław, 23 listopada 2009

1 Zadanie 1

1.1 Treść zadania

Wyznacz najmniejszą liczbę dodatnią w komputerze ϵ (epsilon maszynowy) postaci $\epsilon = 2^{-q}$, dla której $1 + \epsilon \neq 1$. Wyznacz precyzję arytmetyki 2^{-t} (*unit roundoff*), gdzie t jest liczbą cyfr mantysy, oraz najmniejszą liczbę dodatnią η reprezentowaną w komputerze. Jaki jest związek między q i t ? Czemu w komputerze równa się $1 + \eta$? Obliczenia wykonaj dla różnych typów arytmetyk (`Real`, `Double`, `Extended`).

1.2 Wstęp teoretyczny

Precyzją arytmetyki nazywamy najmniejszy możliwy błąd względny jaki komputer może popełnić podczas jej konwersji do liczby maszynowej.

Liczby zmiennopozycyjne (zmiennoprzecinkowe) reprezentowane są za pomocą trzech wartości: znaku S , wykładnika E i mantysy M :

$$x = S \cdot M \cdot 2^E$$

Znak S zapisywany jest na jednym bicie. Ilość bitów przeznaczonych do zapisu wykładnika i mantysy zależy od użytej arytmetyki. Wykładnik E jest liczbą całkowitą, mantysa M jest liczbą wymierną z przedziału $[1, 2)$ – zapisywane są jedynie bity znajdujące się po przecinku, gdyż nie ma sensu pamiętać jedynki reprezentującej jedność, gdyż jest ona zawsze obecna w owej reprezentacji. Jeśli jednak bity służące do reprezentacji wykładnika są wyzerowane, wówczas mantysa reprezentuje liczby z przedziału $[0, 1)$ (wówczas przed przecinkiem nie ma jedynki). Można w ten sposób zapisać bardzo małe liczby.

1.3 Wyniki

W ramach rozwiązania zadania wykonano program w języku C++, który dał następujące wyniki:

```

1 float arithmetic:
2 Epsilon: 1.19209e-07 = 2^(-23)
3 Unit roundoff: 5.96046e-08 = 2^(-24)
4 Etha: 1.4013e-45 = 2^(-149)
5
6 double arithmetic:
7 Epsilon: 2.22045e-16 = 2^(-52)
8 Unit roundoff: 1.11022e-16 = 2^(-53)
9 Etha: 4.94066e-324 = 2^(-1074)
10
11 long double arithmetic:
12 Epsilon: 1.0842e-19 = 2^(-63)
13 Unit roundoff: 5.42101e-20 = 2^(-64)
14 Etha: 3.6452e-4951 = 2^(-16445)

```

1.4 Wnioski

Otrzymane wyniki wyjaśnia sposób, w jaki komputer reprezentuje liczby zmiennopozycyjne.

Najmniejsza liczba większa od jedynki, którą można reprezentować, zawiera wykładnik o wartości 0 oraz mantysę zawierającą same zera i jedynkę na ostatnim bicie. Liczba ta wynosi więc $1 + 2^{-t}$ gdzie t jest liczbą bitów użytych do zapisu mantysy. Na tej podstawie $\epsilon = 2^{-t}$.

Precyzja arytmetyki wynosi $\frac{\epsilon}{2}$. Jest to oczywiste, gdyż iloraz dwóch kolejnych liczb maszynowych wynosi $1 + \epsilon$, natomiast liczba z przedziału pomiędzy nimi zostanie zawsze zaokrąglona do tej bliższej, błąd względny nie może być wówczas większy niż połowa epsilon maszynowego.

Najmniejsza możliwa liczba reprezentowana w arytmetyce `float` wynika ze sposobu zapisu wykładnika. Najmniejszym możliwym wykładnikiem jest -126. Mantysa pamiętana jest wówczas bez jedynki przed przecinkiem i reprezentuje liczbę $2^{-t} = 2^{-23}$. Stąd $\eta_{float} = 2^{-126} \cdot 2^{-23} = 2^{-149}$. Dla arytmetyki `double` i `long double` η wyznacza się analogicznie.

2 Zadanie 11

2.1 Treść zadania

Przybliżoną wartość pochodnej funkcji $f(x)$ w punkcie x można obliczyć z wzoru:

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

Skorzystaj z tego wzoru do obliczenia przybliżonej wartości pochodnej funkcji $f(x) = \sin x + \cos 3x$ w punkcie $x = 1$ dla $h = 2^{-n}$ ($n = 0, 1, 2, \dots, 40$). Jak wytłumaczyć, że od pewnego momentu zmniejszenie wartości h nie poprawia przybliżenia wartości pochodnej? Jak zachowują się wartości $1 + h$? Obliczone przybliżenia pochodnej porównaj z dokładną wartością pochodnej.

2.2 Wstęp teoretyczny

Podany wzór wymusza odejmowanie bliskich liczb (to znaczy liczb, których różnica jest mała w stosunku do nich samych). Powoduje to utratę cyfr znaczących – jest to konsekwencją zapisu liczb zmiennoprzecinkowych.

2.3 Wyniki

W ramach rozwiązania zadania wykonano skrypt w programie octave, który dał następujące wyniki:

```
1 ans =
2
3 Columns 1 through 8:
4    2.01799    1.87044    1.10779    0.62324    0.37040    0.24344    0.18010    0.14849
5
6 Columns 9 through 16:
7    0.13271    0.12482    0.12088    0.11891    0.11793    0.11743    0.11719    0.11707
8
9 Columns 17 through 24:
10   0.11700    0.11697    0.11696    0.11695    0.11695    0.11694    0.11694    0.11694
11
12 Columns 25 through 32:
13   0.11694    0.11694    0.11694    0.11694    0.11694    0.11694    0.11694    0.11694
14
15 Columns 33 through 40:
16   0.11694    0.11694    0.11694    0.11694    0.11694    0.11693    0.11694    0.11688
17
18 Columns 41 through 48:
19   0.11682    0.11694    0.11719    0.11621    0.11719    0.11328    0.10938    0.12500
20
21 Columns 49 through 56:
22   0.09375    0.12500    0.00000    0.00000   -0.50000    0.00000    0.00000    0.00000
23
24 Columns 57 through 61:
25   0.00000    0.00000    0.00000    0.00000    0.00000
```

2.4 Wnioski

Prawdziwa wartość pochodnej danej funkcji w punkcie 1 wynosi w przybliżeniu 0,11694. Wyniki wskazują na to, iż wyrazy 22 - 38 najdokładniej przybliżają daną wartość. Kolejne wyrazy ciągu są coraz mniej dokładne ze względu na utratę cyfr znaczących wspomnianą we wstępie teoretycznym. Od wyrazu 54 (czyli dla $h \leq 2^{-53}$, gdyż n -ty wyraz ciągu liczony jest dla $h = 2^{n-1}$) w ciągu pojawiają się same zera, gdyż h jest mniejsze od epsilon maszynowego – w efekcie liczniku pojawia się zero.