Ćwiczenia z ANALIZY NUMERYCZNEJ (M)

Lista M 1

8 października 2015 r.

- **M1.1.** I punkt Niech B będzie liczbą naturalną większą od 1. Wykazać, że każda niezerowa liczba rzeczywista x ma jednoznaczne przedstawienie w postaci znormalizowanej $x = smB^c$, gdzie s jest znakiem liczby x, c liczbą całkowitą (cechq), a m liczbą z przedziału [1, B), zwaną mantysq.
- **M1.2.** I punkt Dla danych: naturalnej liczby t oraz niezerowej liczby rzeczywistej $x=s\,m\,2^c$, gdzie s jest znakiem liczby x, c liczbą całkowitą, a m liczbą z przedziału [1,2), o rozwinięciu dwójkowym $m=1+\sum_{k=1}^{\infty}e_{-k}2^{-k}$, w którym $e_{-k}\in\{0,1\}$ dla $k\geqslant 1$, definiujemy zaokrąglenie liczby x do t+1 cyfr za pomocą wzoru

$$\operatorname{rd}(x) := s \, \bar{m} \, 2^c$$
,

gdzie $\bar{m} = 1 + \sum_{k=1}^{t} e_{-k} 2^{-k} + e_{-t-1} 2^{-t}$.

Wykazać, że

$$|\operatorname{rd}(x) - x| \leq 2^{c} \mathsf{u},$$

gdzie $u := 2^{-t-1}$ jest precyzją arytmetyki.

- **M1.3.** 1,5 punktu Niech x będzie dowolną niezerową liczbą rzeczywistą. Wykazać, że błąd względny zaokrąglenia liczby x nie przekracza u/(1 + u).
- M1.4. 1 punkt le jest liczb zmiennopozycyjnych w arytmetyce *single*, a ile w arytmetyce *double* (wg standardu IEEE 754)?
- **M1.5.** 1 punkt Załóżmy, że $|\alpha_j| \le u$ i $\rho_j \in \{-1, +1\}$ dla $j = 1, 2, \dots, n$ oraz że n u < 1, gdzie $u := 2^{-t-1}$. Wykazać, że zachodzi równość

$$\prod_{j=1}^{n} (1 + \alpha_j)^{\rho_j} = 1 + \theta_n,$$

gdzie θ_n jest wielkością spełniającą nierówność $|\theta_n| \leqslant \gamma_n$, gdzie z kolei

$$\gamma_n := \frac{n\mathsf{u}}{1 - n\mathsf{u}}.$$

M1.6. 1,5 punktu Załóżmy, że $|\alpha_j| \le u$ dla $j=1,2,\ldots,n$ oraz że nu < 0.01. Wykazać, że zachodzi równość

$$\prod_{j=1}^{n} (1 + \alpha_j) = 1 + \eta_n,$$

gdzie

$$|\eta_n| \leqslant 1.01nu$$
.

M1.7. 1 punkt Załóżmy, że x,y są liczbami maszynowymi, tzn. rd(x)=x, rd(y)=y, takimi, że 0< y< x. Wykazać, że jeśli

$$2^{-q} \leqslant 1 - \frac{y}{r} \leqslant 2^{-p}$$

(p i q sa całkowite), to

 $p \leq \text{liczba bitów straconych przy odejmowaniu } x - y \leq q.$