

Obliczenia naukowe

Wykład nr 10

Paweł Zieliński

Katedra Podstaw Informatyki,
Wydział Informatyki i Telekomunikacji,
Politechnika Wrocławska



Literatura

Literatura podstawowa

- [1] D. Kincaid, W. Cheney, *Analiza numeryczna*, WNT, 2005.
- [2] M. T. Heath, *Scientific Computing, An Introductory Survey*, The McGraw-Hill Companies, 1997.
- [3] A. Bjorck, G. Dahlquist, *Metody numeryczne*, PWN, 1987.
- [4] M. Dryja, J. i M. Jankowscy, *Przegląd metod i algorytmów numerycznych*, cz. 2, WNT, 1988.
- [5] Z. Fortuna, B. Macukow, J. Wąsowski, *Metody numeryczne*, WNT 1998.
- [6] J. i M. Jankowscy, *Przegląd metod i algorytmów numerycznych*, cz. 1, WNT, 1988.
- [7] J. Stoer, R. Bulirsch, *Wstęp do analizy numerycznej*, PWN, 1987.

Literatura uzupełniająca

- [8] Kiełbasiński, H. Schwetlick, *Numeryczna algebra liniowa*, WNT, 1993.

Znaczna część wykładu została przygotowana na podstawie książki [6].



Algorytmy numerycznie poprawne

Pytanie: Jak dokładny może być dla danego zadania wynik obliczony w arytmetyce *fl*?

Rozważmy zadania polegające na obliczaniu dla danych $\mathbf{d} = (d_1, d_2, \dots, d_n)$ wyniku $\mathbf{w} = (w_1, w_2, \dots, w_m)$

$$\varphi(\mathbf{d}) = \mathbf{w},$$

gdzie $\varphi : \mathcal{D}_0 \rightarrow \mathbb{R}^m$, $\mathcal{D}_0 \subset \mathbb{R}^n$, φ jest ciągłym odwzorowaniem.



Algorytmy numerycznie poprawne

Definicja

Mówimy o algorytmie $\mathcal{A}$ w klasie zadań $\{\varphi : \mathcal{D}\}$, gdzie $\mathcal{D} \subset \mathcal{D}_0$, jeśli $\mathcal{A}$ oznacza algorytm obliczania (dokładnego) wyniku $\varphi(\mathbf{d}) = \mathbf{w}$ dla dowolnych danych $\mathbf{d} \in \mathcal{D}$. Jest to obliczanie w arytmetyce klasycznej.

Przez $fl(\mathcal{A}(\mathbf{d}))$ rozumiemy realizację algorytmu $\mathcal{A}$ w arytmetyce fl polegającą na:

- zastąpieniu danych liczbowych ich reprezentacjami zmiennopozycyjnymi,
- wykonaniu operacji arytmetycznych w arytmetyce fl .

Zakłada się, że dane $\mathbf{d} \in \mathcal{D}$ i wyniki $\mathbf{w} = \varphi(\mathbf{d})$ reprezentowane są z błędami nie większymi od ρ_d i ρ_w

$$\|\mathbf{d} - rd(\mathbf{d})\| \leq \rho_d \|\mathbf{d}\|, \quad \|\mathbf{w} - rd(\mathbf{w})\| \leq \rho_w \|\mathbf{w}\|,$$

gdzie $|\rho_d| \leq C2^{-t}$, $|\rho_w| \leq D2^{-t}$, $C = \mathcal{O}(1)$, $D = \mathcal{O}(1)$.



Algorytmy numerycznie poprawne

Definicja

Algorytm $\mathcal{A}$ nazywamy numerycznie poprawnym w klasie zadań $\{\varphi : \mathcal{D}\}$ jeśli istnieją stałe K_d i K_w (wskaźniki kumulacji) takie, że dla każdego $\mathbf{d} \in \mathcal{D}$ istnieje dla dostatecznie silnej arytmetyki (tj. dostatecznie dużego t) takie $\tilde{\mathbf{d}} \in \mathcal{D}_0$, że

$$\begin{aligned} \|\mathbf{d} - \tilde{\mathbf{d}}\| &\leq K_d \rho_d \|\mathbf{d}\| \\ \|fl(\mathcal{A}(\mathbf{d})) - \varphi(\tilde{\mathbf{d}})\| &\leq K_w \rho_w \|\varphi(\tilde{\mathbf{d}})\|. \end{aligned}$$

Interpretacja. Dla dowolnego zadania rozważanej klasy, tj. zadania wyznaczenia elementu $\varphi(\mathbf{d})$, rozwiązanie $fl(\mathcal{A}(\mathbf{d}))$ obliczone algorytmem numerycznie poprawnym różni się od rozwiązania dokładnego $\varphi(\tilde{\mathbf{d}})$ co najwyżej K_w razy więcej niż poziom błędu jego reprezentacji. Jednocześnie $\varphi(\tilde{\mathbf{d}})$ jest dokładnym rozwiązaniem zadania o danych $\tilde{\mathbf{d}}$ różniących się od danych początkowych $\mathbf{d}$ co najwyżej K_d razy więcej niż poziom błędu ich reprezentacji.



Algorytmy numerycznie poprawne

Przykład

Zbadajmy numeryczną poprawność algorytmu obliczania iloczynu skalarnego

$$S(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^n a_i b_i, \quad \mathbf{a} = (a_1, a_2, \dots, a_n), \quad \mathbf{b} = (b_1, b_2, \dots, b_n).$$

Rozpatrzmy algorytm $\mathcal{A}(\mathbf{a}, \mathbf{b})$

```
S  $\leftarrow$  0;  
for k  $\leftarrow$  1 to n do  
    S  $\leftarrow$  S + ai * bi;  
end for  
return S;
```



Algorytmy numerycznie poprawne

Realizujemy algorytm $\mathcal{A}(\mathbf{a}, \mathbf{b})$ w arytmetyce fl , czyli zastępujemy dane ich reprezentacjami zmiennopozycyjnymi

$$\hat{a}_i = rd(a_i) = a_i(1 + \alpha_i)$$

$$\hat{b}_i = rd(b_i) = b_i(1 + \beta_i), \text{ gdzie } |\alpha_i|, |\beta_i| \leq 2^{-t}, \quad i = 1, \dots, n$$

i wykonujemy obliczenia w arytmetyce fl

$$\begin{aligned} fl(\mathcal{A}(\mathbf{a}, \mathbf{b})) &= (\dots (\hat{a}_1 \hat{b}_1 (1 + \epsilon_1) + \hat{a}_2 \hat{b}_2 (1 + \epsilon_2)) (1 + \delta_2) + \\ &+ \dots + \hat{a}_n \hat{b}_n (1 + \epsilon_n)) (1 + \delta_n) \\ &= \sum_{k=1}^n a_k (1 + \alpha_k) b_k (1 + \beta_k) (1 + \epsilon_k) \prod_{j=k}^n (1 + \delta_j) \end{aligned}$$

gdzie $\delta_1 = 0$, $1 + E_k = (1 + \beta_k)(1 + \epsilon_k) \prod_{j=k}^n (1 + \delta_j)$.



Algorytmy numerycznie poprawne

Dokładny wynik zadania dla zaburzonych danych jest postaci

$$S(\tilde{\mathbf{a}}, \tilde{\mathbf{b}}) = \sum_{k=1}^n \tilde{a}_k \tilde{b}_k$$

$$\tilde{a}_i = a_i(1 + \alpha_i), \quad \tilde{b}_i = b_i(1 + E_i), \quad i = 1, \dots, n$$

$$\|\mathbf{a} - \tilde{\mathbf{a}}\|_2 = \|(\mathbf{a}_1\alpha_1, \dots, \mathbf{a}_n\alpha_n)\|_2 \leq \max_{1 \leq k \leq n} |\alpha_k| \|\mathbf{a}\|_2 = 2^{-t} \|\mathbf{a}\|_2$$

$$(K_{d_1} = 1, \rho_{d_1} = 2^{-t})$$

$$\|\mathbf{b} - \tilde{\mathbf{b}}\|_2 = \|(\mathbf{b}_1 E_1, \dots, \mathbf{b}_n E_n)\|_2 \leq \max_{1 \leq k \leq n} |E_k| \|\mathbf{b}\|_2 =$$

$$(n+1)2^{-t} \|\mathbf{b}\|_2$$

$$(K_{d_2} = n+1, \rho_{d_2} = 2^{-t})$$

$$|S(\tilde{\mathbf{a}}, \tilde{\mathbf{b}}) - fl(\mathcal{A}(\mathbf{a}, \mathbf{b}))| = \left| \sum_{k=1}^n \tilde{a}_k \tilde{b}_k - \sum_{k=1}^n \tilde{a}_k \tilde{b}_k \right| = 0$$

Wskaźnik kumulacji ze względu na wynik $K_w = 0$.



Algorytmy numerycznie poprawne

Dąży się, aby wskaźniki kumulacji były jak najmniejsze. W praktyce za algorytmy numerycznie poprawne uznaje się algorytmy mające wskaźniki kumulacji rzędu liczby działań arytmetycznych potrzebnych do rozwiązania zadania. Stąd algorytm obliczania iloczynu skalarnego jest numerycznie poprawny.

Warto zauważyć, że pominięcie błędów reprezentacji danych spowoduje zmniejszenie wskaźników kumulacji K_{d_1} i K_{d_2} o jeden. Zatem w analizie numerycznej poprawności algorytmu błędy reprezentacji danych pomija się.



Algorytmy numerycznie poprawne

Przykład

Zbadajmy numeryczną poprawność algorytmu Hornera obliczania wartości wielomianu $w(x)$

$$\begin{aligned}w(x) &= a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0 \\ &= (\dots ((a_n x + a_{n-1})x + a_{n-2})x + \cdots + a_1)x + a_0.\end{aligned}$$

Algorytm Hornera $\mathcal{A}(\mathbf{a}, x)$

```
 $w_n \leftarrow a_n;$   
for  $k \leftarrow n - 1$  downto 0 do  
     $w_i \leftarrow w_{i+1} * x + a_i;$   
end for  
return  $w_0;$ 
```



Algorytmy numerycznie poprawne

Zakładamy, że dane zadania reprezentowane są dokładnie, czyli $x = rd(x)$, $a_i = rd(a_i)$, $i = 0, \dots, n$.

Realizacja algorytmu Hornera $\mathcal{A}(\mathbf{a}, x)$ w arytmetyce fl jest następująca

$$\tilde{w}_n = a_n$$

$$\tilde{w}_i = fl(\tilde{w}_{i+1} * x + a_i) = (\tilde{w}_{i+1}x(1 + \epsilon_i) + a_i)(1 + \delta_i)$$

$$i = n-1, n-2, \dots, 0, \quad \text{gdzie } |\epsilon_i|, |\delta_i| \leq 2^{-t}.$$

Z powyższego dostajemy

$$\begin{aligned} w_0 &= ((\dots((a_n x(1 + \epsilon_{n-1}) + a_{n-1})(1 + \delta_{n-1})x(1 + \epsilon_{n-2}) + \\ &+ a_{n-2})(1 + \delta_{n-2})x(1 + \epsilon_{n-3}) + \dots + a_1)(1 + \delta_1)x(1 + \epsilon_0) + \\ &+ a_0)(1 + \delta_0) = \end{aligned}$$



Algorytmy numerycznie poprawne

$$\begin{aligned} &= a_n(1 + \epsilon_{n-1})(1 + \delta_{n-1}) \cdots (1 + \epsilon_0)(1 + \delta_0)x^n + \\ &+ a_{n-1}(1 + \epsilon_{n-2})(1 + \delta_{n-1}) \cdots (1 + \epsilon_0)(1 + \delta_0)x^{n-1} + \cdots \\ &+ a_1(1 + \epsilon_0)(1 + \delta_1)(1 + \delta_0) + a_0(1 + \delta_0) = \sum_{k=0}^n a_k(1 + E_k)x^k \\ &= \sum_{k=0}^n \tilde{a}_k x^k, \end{aligned}$$

gdzie $1 + E_k = (1 + \delta_k) \prod_{i=0}^{k-1} (1 + \epsilon_i)(1 + \delta_i)$ przyjmując $\delta_n = 0$.

$$\|\mathbf{a} - \tilde{\mathbf{a}}\|_2 = \|(a_0 E_0, \dots, a_n E_n)\|_2 \leq \max_{0 \leq k \leq n} |E_k| \|\mathbf{a}\|_2 = (2n+1)2^{-t} \|\mathbf{a}\|_2$$

$$|w(\tilde{\mathbf{a}}, x) - fl(\mathcal{A}(\mathbf{a}, x))| = \left| \sum_{k=0}^n \tilde{a}_k x^k - \sum_{k=0}^n \tilde{a}_k x^k \right| = 0$$

$K_d = 2n + 1$, $\rho_d = 2^{-t}$, $K_w = 0$. Stąd $\mathcal{A}(\mathbf{a}, x)$ jest numerycznie poprawny

