Liczby zmiennoprzecinkowe i błędy

Michał Goliński

Elementy metod numerycznych

Kontakt

- Michał Goliński
- pokój B3-10
- tel.: 829 53 62
- `http://golinski.faculty.wmi.amu.edu.pl/`
- golinski@amu.edu.pl

Plan wykładu

1 Liczby zmiennoprzecinkowe

Plan wykładu

1 Liczby zmiennoprzecinkowe

2 Błędy

Plan wykładu

1 Liczby zmiennoprzecinkowe

2 Błędy

3 Współczynnik uwarunkowania zadania

Liczby zmiennoprzecinkowe

Każda liczbę niezerową można przedstawić jednoznacznie w postaci

$$\pm m \cdot 2^c,$$

gdzie $m \in [1, 2)$ (mantysa) i $c \in \mathbb{Z}$ (cecha). Liczby takie nazywamy zmiennoprzecinkowymi i współczesne komputery takie liczby przechowują i wykonują na nich obliczenia. Pamięć nie jest nieograniczona – w praktyce przechowujemy skończoną ilość cyfr mantysy oraz wartość cechy też jest jakoś ograniczona. Dla przykładu:

$$\begin{aligned}
2 &\longrightarrow 0\ 10000000000\ 00000000000000000000000000000000000000000000000000 \\
1/3 &\longrightarrow 0\ 01111111101\ 01010101010101010101010101010101010101010101010101 \\
-1/5 &\longrightarrow 1\ 01111111100\ 10011001100110011001100110011001100110011001100110
\end{aligned}$$

Kolejne grupy cyfr to: znak, zakodowana cecha i ułamkowe cyfry mantysy.

IEEE 754

Standardem liczb zmiennoprzecinkowych implementowanym w komputerach jest IEEE 754. To w tym standardzie były zapisane liczby na poprzednim slajdzie. Poza zaprezentowanymi liczbami tzw. znormalizowanymi, standard ten definiuje zapis dla:

- ± 0
- $\pm \infty$
- bardzo bliskich zeru liczb subnormalnych (zwiększamy zakres kosztem precyzji)
- NaN – Not a number – sygnalizuje programiście wykonanie nieprawidłowej operacji

Standard definiuje liczby pojedynczej precyzji (`float` w C/C++) i podwójnej precyzji (`double`). Oprócz tego nakłada na implementację pewne obowiązki dotyczące dokładności prowadzonych działań na tych liczbach. Niestety, standard ten nie jest dostępny za darmo.

Zaokrąglenia

Ze względu na używanie liczb binarnych, większość skończonych ułamków dziesiętnych nie ma skończonych rozwinięć zmiennoprzecinkowych. Ponieważ przechowujemy tylko skończenie wiele cyfr, prowadzi to do błędów zaokrągleń, a w następstwie do powstawania błędów. Dla przykładu, jeżeli na papierze obliczenia dają 0, to może się okazać, że w komputerze dostajemy w wyniku małą ale niezerową liczbę. Podobnie wynik w arytmetyce zmiennoprzecinkowej może zależeć od kolejności wykonywanych działań.

Błędy względne i bezwzględne

Mając wartośc dokładną x i wartość przybliżoną $\overline{x}$ mówimy, że popełniamy **błąd bezwzględny**

$$|x - \overline{x}|$$

oraz **błąd względny**

$$\left| \frac{x - \overline{x}}{x} \right|.$$

Błędy względne i bezwzględne

Mając wartośc dokładną x i wartość przybliżoną $\overline{x}$ mówimy, że popełniamy **błąd bezwzględny**

$$|x - \overline{x}|$$

oraz **błąd względny**

$$\left| \frac{x - \overline{x}}{x} \right|.$$

W praktyce znacznie bardziej istotne są błędy względne.

Źródła błędów

Nigdy nie możemy zakładać, że nasz program działa na danych dokładnych odpowiadających rzeczywistości. W szczególności błędy to np.:

- błędy zaokrągleń,
- błędy pomiarów,
- błędy wykonania.

Programista może kontrolować (w ograniczonym zakresie) błędy zaokrągleń. Pozostałe rodzaje błędów są poza naszą kontrolą. Należy tak konstruować programy, by pojawiające się błędy nie wypaczyły za bardzo wyniku.

Błędy zaokrągleń dla liczb zmiennoprzecinkowych

Błędy zaokrągleń dla stosowanych w komputerach liczb zmiennoprzecinkowych są ograniczone przez tzw. dokładnośc maszynową, oznaczaną z reguły przez ε. Dokładność maszynowa to najmniejsza liczba taka, że (w obliczeniach na danej maszynie):

$$1 + \varepsilon \neq 1.$$

Liczby rzeczywiste są przybliżane przez liczby zmiennoprzecinkowe z dokładnością względną $\varepsilon/2$, to znaczy, że

$$fl(x) = x(1 + \delta), \text{gdzie } |\delta| < \varepsilon/2.$$

Dla liczb pojedynczej precyzji standardu IEEE 754 mamy $\varepsilon = 2^{-23}$, dla liczb podwójnej precyzji: $\varepsilon = 2^{-52}$.

Utrata precyzji

Z utratą precyzji mamy do czynienia, gdy odejmujemy dwie bliskie siebie liczby.

Utrata precyzji

Z utratą precyzji mamy do czynienia, gdy odejmujemy dwie bliskie siebie liczby.

Theorem

Jeżeli $x > y$ są znormalizowanymi liczbami zmiennoprzecinkowymi, oraz $2^{-q} \leq 1 - y/x \leq 2^{-p}$, to przy obliczaniu $x - y$ tracimy co najmniej p i co najwyżej q bitów precyzji (to znaczy: co najmniej ostatnich p bitów mantysy będzie zerami).

Utrata precyzji – przykład

Obliczymy na komputerze $(1 + 1/999999) - (1 + 1/1000000)$.
Wynik ma reprezentację:

0 01111010111 00011001100000000000000000000000000000000000000000000000

Wynik dla $(1/999999) - (1/1000000)$ to:

0 01111010111 00011001011110011010101010000101000000000000000000000000

Wynik dla $1/(999999 \cdot 1000000)$ to:

0 01111010111 00011001011110011010101010000101001111011010100000110

Utrate precyzji – przykład cd.

W pierwszych dwóch obliczeniach mamy odpowiednio:

$$2^{-40} \leq 1 - \frac{y_1}{x_1} \cong 10^{-12} \leq 2^{-39}$$
$$2^{-20} \leq 1 - \frac{y_2}{x_2} \cong 10^{-6} \leq 2^{-19}$$

W pierwszym przykładzie tracimy więc ok. 40 bitów precyzji, w drugim 20 bitów precyzji.

Współczynnik uwarunkowania zadania mierzy jak bardzo błąd danych zadania obliczeniowego przenosi się na błąd względny wyniku. Definiujemy go jako stosunek tych dwóch wartości.

Współczynnik uwarunkowania zadania mierzy jak bardzo błąd danych zadania obliczeniowego przenosi się na błąd względny wyniku. Definiujemy go jako stosunek tych dwóch wartości. Zadanie, dla którego ten współczynnik jest mały nazywamy dobrze uwarunkowanym, zadanie dla którego jest duży nazywamy źle uwarunkowanym. Zadania dla którego współczynnik ten jest wręcz nieskończony – źle postawionym. Duży współczynnik oznacza, że stworzenie programu, który w zadowalający sposób obliczy co trzeba może okazać się niemożliwe.

Współczynnik uwarunkowania zadania mierzy jak bardzo błąd danych zadania obliczeniowego przenosi się na błąd względny wyniku. Definiujemy go jako stosunek tych dwóch wartości.
Zadanie, dla którego ten współczynnik jest mały nazywamy dobrze uwarunkowanym, zadanie dla którego jest duży nazywamy źle uwarunkowanym. Zadania dla którego współczynnik ten jest wręcz nieskończony – źle postawionym. Duży współczynnik oznacza, że stworzenie programu, który w zadowalający sposób obliczy co trzeba może okazać się niemożliwe.
Dla przykładu rozważymy współczynnik uwarunkowania zadania obliczenia wartości funkcji w zadanym punkcie i współczynnik uwarunkowania zadania znalezienia rozwiązania układu równań liniowych.

Obliczenie wartości funkcji

Załóżmy, że chcemy obliczyć wartośc funkcji $f(x)$, jednak dane mają błąd względny δ Zanczy to, że w istocie obliczymy wartość $f(x(1+\delta))$.

Obliczenie wartości funkcji

Załóżmy, że chcemy obliczyć wartośc funkcji $f(x)$, jednak dane mają błąd względny δ Zanczy to, że w istocie obliczymy wartość $f(x(1+\delta))$. Stąd błąd względny wynosi:

$$\left| \frac{f(x) - f(x(1+\delta))}{f(x)} \right|$$

Obliczenie wartości funkcji

Załóżmy, że chcemy obliczyć wartośc funkcji $f(x)$, jednak dane mają błąd względny δ Zanczy to, że w istocie obliczymy wartość $f(x(1+\delta))$. Stąd błąd względny wynosi:

$$\left|\frac{f(x) - f(x(1+\delta))}{f(x)}\right| = \left|\frac{f'(\xi)\delta x}{f(x)}\right|$$

Obliczenie wartości funkcji

Załóżmy, że chcemy obliczyć wartośc funkcji $f(x)$, jednak dane mają błąd względny δ Zanczy to, że w istocie obliczymy wartość $f(x(1+\delta))$. Stąd błąd względny wynosi:

$$\left|\frac{f(x)-f(x(1+\delta))}{f(x)}\right| = \left|\frac{f'(\xi)\delta x}{f(x)}\right| \cong \left|\frac{f'(x)\delta x}{f(x)}\right|$$

Obliczenie wartości funkcji

Załóżmy, że chcemy obliczyć wartośc funkcji $f(x)$, jednak dane mają błąd względny δ Zanczy to, że w istocie obliczymy wartość $f(x(1+\delta))$. Stąd błąd względny wynosi:

$$\left|\frac{f(x)-f(x(1+\delta))}{f(x)}\right| = \left|\frac{f'(\xi)\delta x}{f(x)}\right| \cong \left|\frac{f'(x)\delta x}{f(x)}\right| = \left|\frac{f'(x)x}{f(x)}\right| |\delta|.$$

Dla funkcji wielu zmiennych $F(x_1,\ldots,x_n)$ współczynnik ten wynosi:

$$\sum_{k=1}^{n}\left|\frac{\frac{\partial F}{\partial x_k}(x_1,\ldots,x_n)x_k}{F(x_1,\ldots,x_n)}\right|.$$

Przykład

Znajdź współczynnik uwarunkowania zadania obliczenia pola powierzchni pokoju o wymiarach $a \times b$.

Przykład

Znajdź współczynnik uwarunkowania zadania obliczenia pola powierzchni pokoju o wymiarach $a \times b$.
Pole jest funkcją dwóch zmiennych: $P(x, y) = xy$. Korzystamy ze wzoru:

$$\left| \frac{ba}{ab} \right|$$

Przykład

Znajdź współczynnik uwarunkowania zadania obliczenia pola powierzchni pokoju o wymiarach $a \times b$.
Pole jest funkcją dwóch zmiennych: $P(x, y) = xy$. Korzystamy ze wzoru:

$$\left|\frac{ba}{ab}\right| + \left|\frac{ab}{ab}\right|$$

Przykład

Znajdź współczynnik uwarunkowania zadania obliczenia pola powierzchni pokoju o wymiarach $a \times b$.
Pole jest funkcją dwóch zmiennych: $P(x, y) = xy$. Korzystamy ze wzoru:

$$\left|\frac{ba}{ab}\right| + \left|\frac{ab}{ab}\right| = 2.$$

Przykład

Znajdź współczynnik uwarunkowania zadania obliczenia pola powierzchni pokoju o wymiarach $a \times b$.
Pole jest funkcją dwóch zmiennych: $P(x, y) = xy$. Korzystamy ze wzoru:

$$\left|\frac{ba}{ab}\right| + \left|\frac{ab}{ab}\right| = 2.$$

Przy wymiarach $1 \times 1m$ i tolerancji 1% dostajemy najmniejsze pole $0{,}9801m^2$ i największe pole $1{,}0201m^2$, więc widać, że tolerancja wynosi ok. 2%.

Rozwiązywanie układów równań

Załóżmy, że interesuje nas rozwiązanie układu $Ax = b$. Przypuśćmy, że macierz A jest dokładna, jedyne błędy dotyczą wyrazów wolnych b. Rozwiązaniem układu jest $x = A^{-1}b$. Błąd względny popełniany przy przybliżonych danych $\overline{b}$ jest równy (użyjemy bliżej niesprecyzowanej normy $\|\cdot\|$ dla wektorów):

$$\frac{\left\|A^{-1}b - A^{-1}\overline{b}\right\|}{\|A^{-1}b\|} = \frac{\left\|A^{-1}(b-\overline{b})\right\|}{\|A^{-1}b\|} = \frac{\|b\|}{\|A^{-1}b\|}\frac{\left\|A^{-1}(b-\overline{b})\right\|}{\|b\|}$$
$$\leq \|A\|\left\|A^{-1}\right\|\frac{\left\|b-\overline{b}\right\|}{\|b\|}.$$

Wyrażenie $\|A\|\left\|A^{-1}\right\|$ nazywamy współczynnikiem uwarunkowania macierzy. Macierze dla których ten wskaźnik jest duży sprawiają problemy w obliczeniach numerycznych.