Metody numeryczne

Wojciech Kubiak

19 czerwca 2019

1 Arytmetyka zmiennopozycyjna, standard IEEE

- **Liczby maszynowe:** liczby rzeczywiste które można reprezentować w komputerze.
- **Reprezentacja liczby zmiennoprzecinkowej:** w arytmetyce podwójnej precyzji jest to ciąg 64 bitów, w pojedyńczej jest to ciąg 32 bitów.
- **Epsilon:** jest to najmniejsza dodatnia liczba spełniająca równanie $1 + \mathcal{E} \neq 1$
- **Wykładnik po przesunięciu "shift":** Jest to liczba zapisana na 8 bitach może być z zakresu -126 do 127. Używa się go do kodowania z nadmiarem.

2 Nadmiar i niedomiar

- **Nadmiar:** jeżeli w czasie obliczeń liczba x (np. wynik operacji arytmetycznych) znajdzie się poza dopuszczalnym zakresem liczb. Kończy działanie programu.
- **Niedomiar:** automatycznie zapamiętywana jest liczba x jako zero bez przerwania działania programu.

3 Błąd względny, bezwzględny, zaokrąglenie do najbliższej liczby parzystej

- **Błąd względny:** $|\frac{a-\tilde{a}}{a}|$
- **Błąd bezwzględny:** $|a - \tilde{a}|$
- **Metoda zaokrąglenia do najbliższej pażystej:** redukuje błąd całkowity obliczeń z uwagi na statycznie równą liczbę zaokrągleń w góre i w dół.

4 Źródła błędów

- **zaokrąglenia** (związane z pracą w arytmetyce o skończonej pozycji)
- **obcięcia** (obliczeń do skończonej liczby kroków)
- **niepewność danych** (pojawiająca się w przypadku pracy na danych związanych z problemiami praktycznymi np. pomiarami fizycznymi)

5 Uwarunkowanie zadania

- **Definicja:** Jeżeli niewielka zmiana danych wejściowych powoduje duże błędy w wyniku mówimy że nasze zadanie jest źle uwarunkowane. Wielkość charakteryzująca wpływ tych zaburzeń nazywamy wskaźnikiem uwarunkowania zadania (ang. condition numbers).
- **Dla funkcji wielu zmiennych** $cond(f(x, y, \dots)) = \frac{\partial f}{\partial x} \cdot \frac{x}{f(x,y)} + \frac{\partial f}{\partial y} \cdot \frac{y}{f(x,y)} \dots$
- **Dla iloczynu skalarnego wektorów** $cond(x, y) = \frac{\langle |x|,|y| \rangle}{|\langle x,y \rangle|}$
- **Dla macierzy** $cond(A) = \|A\| \cdot \|A^{-1}\|$

6 Stabilność numeryczna algorytmu

- **Definicja:** algorytm jest niestabilny numerycznie jeżeli wprowadza duże błędy w dobrze uwarunkowanym zadaniu.
- **Jak stworzyć algorytm stabilny numerycznie ?**
  1. Unikaj odejmowania wielkości obarczonych błędem (o ile to możliwe).
  2. Minimalizuj rozmiar wyników pośrednich względem wielkości rozwiązania.
  3. Upewnij się, że metody obliczeń są równoważne numerycznie (nie tylko matematycznie).
  4. Przedstawiaj aktualne wyrażenie jako **nowa wartość** = **poprzednia wartość** + **mała korekta** jeżeli mała korekta może być obliczona z dużą liczbą cyfr znaczących.

7 Ilorazy różnicowe

- **Iloraz różnicowy rzędu k:** $f[x_0, x_1, \dots, x_k] = \sum_{k=0}^{i} \left[\frac{f(x_k)}{\prod_{j=0, j \neg k}^{i} (x_k - x_j)} \right]$
- **Twierdzenie:** Wartość ilorazu różnicowego $f[x_0, x_1, \dots, x_k]$ nie zmienia się bezwzględu na permutacje argumentów $x_0, x_1, \dots, x_k$.

8 Zalety postaci Lagrange'a i Newtona wielomianu interpolacyjnego

- **Zalety Newtona:** jest możliwość dodania nowego punktu bez zmiany wcześniej obliczonych wartości

9 Błąd interpolacji

- **Twierdzenie:** Jeżeli p jest wielomianem stopnia co najwyżej n, interpolującym funkcję f w $n+1$ parami różnych węzłach $x_0, x_1, \dots, x_k$ należących do przedziału [a, b] i jeżeli f^{n+1} jest ciągła to dla każdego x z [a,b] istnieje ξ z (a,b) taki, że

  $$p(x) - f(x) = \omega_{n+1}(x) \frac{f^{n+1}(\xi)}{n+1}!$$

  gdzie $\omega_{n+1}(x) = (x - x_0)(x - x_1) \dots (x - x_n)$ inczej:

  $$|f(x) - L(x)| = \frac{M_{n+1}}{(n+1)!} \cdot (x - x_0)(x - x_1) \dots (x - x_n)$$

  gdzie $M_{n+1} = max_{x \in [a,b]} |f^{n+1}(x)|$

10 Optymalne węzły interpolacji

- **Twierdzenie:** Z twierzenia o błędzie interpolacyjnym wynika, że wielkość błędu zależy od $f(x)$ i od $\omega_{n+1}(x)$, który to wielomian zależy od doboru węzłów interpolacji. Zatem można wybrać węzły $x_0, x_1, \ldots, x_k$ minimalizujące

$$\omega_{max} = max_{x \in [a,b]} |(x - x_0)(x - x_1) \ldots (x - x_n)|$$

11 Funkcje sklejane

- **Funkcja sklejana stopnia k:** Funkcję S nazywamy funkcją sklejaną stopnia k jeżeli:
  1. $[a, b]$ jest dziedziną funckji S
  2. $S, S', S'', \ldots, S^{(k-1)}$ są funkcjami ciągłymi na $[a, b]$
  3. Istnieje taki podział przedziału $a = t_0 < t_1 < \cdots < t_n = b$ dla którego S jest wielomianem stopnia co najwyżej k na każdym popprzedziale $[t_i, t_{i+1}]$

12 Kwadratury

- **Prosty wzór trapezów:**

$$S(f) = \frac{b-a}{2}(f(a) + f(b))$$

- **Prosty wzór Simpsona - parabol:**

$$S(f) = \frac{b-a}{6}(f(a) + 4f(\frac{a+b}{2}) + f(b))$$

13 Metody iteracyjne

- **Metoda bisekcji (poławiania przedziału):**

$$x_k = \frac{a_k + b_k}{2}$$

- **Metoda Newtona (Newtona-Raphsona, stycznych):**

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

- **Metoda siecznych:**

$$x_{k+1} = x_k - (\frac{x_{k-1-x_k}}{f(x_{k-1}) - f(x_k)})f(x_k)$$

- **Rząd metody:** załóżmy, że ciąg przybliżeń $x_{n+1} = F(x_n)$ jest zbieżny do α, tzn.

$$lim_{i \to \infty} x_i = \alpha$$

Mówimy, że w punkcie α metoda iteracyjna ma wykładnik zbieżności p, jeżeli istnieje taka rzeczywista liczba $p \geqslant 1$, że

$$\lim_{i \to \infty} \frac{|\mathcal{E}_{i+1}|}{|\mathcal{E}|^p} = \lim_{i \to \infty} \frac{|x_{i+1} - \alpha|}{|x_i - \alpha|^p} = c \neq 0$$

14 Rozkład LU macierzy

- **Obliczanie wyznacznika:**

$$det(A) = det(L) \cdot det(U)$$

- **Obliczanie odwrotności macierzy:**

$$A^-1 = (LU)^{-1} = L^{-1} \cdot U^{-1}$$

15 Metody iteracyjne rozwiązywania układów algebraicznych

- **Uwaga** aby zapewnić istnienie rozwiązania układu $Ax = b$ dla dowolnego b zakładamy, że macierz A jest nieosobliwa
- Chcemy rozwiązać $Ax = b$. Niech $A = L + D + U$ gdzie:
  - A macierz przedstawiająca nasz układ(wiersze składają się ze współczynników naszego układu)
  - b macierz przedstawiająca nasze wartości(kolumna składająca się z wartości)
  - L jest **macierzą dolno trójkątną** z zerami na głównej przekątnej
  - D jest **macierzą diagonalną**
  - U jest **macierzą górno trójkątną** z zerami na głównej przekątnej
  - **macierz nieosobliwa** jest to taka której wyróżnik jest różny od zera $det(A)\neg 0$
- **Metoda Jacobiego:** zapisujemy nasz układ równań $Ax = b$ w następujący sposób

  $$(L + D + U)x = b$$

  co jest równoważne układowi

  $$Dx = -(L + U) + b$$

  . Jeżeli D jest macierzą nieosobliwą możemy otrzymać metodę

  $$x^{(k)} = -D^{-1}(L + U)x^{k-1} + D^{-1}b$$

  Twierdzenie: Jeżeli A jest silnie diagonalnie dominująca, to dla każdego wektora początkowego $x^{(0)}$ metoda Jacobiego tworzy ciąg zbieżny do rozwiązania układu $Ax = b$
- **Metoda Gaussa-Seidela:** zapisujemy nasz układ równań $Ax = b$ w następujący sposób

  $$(L + D)x = -Ux + b$$

  Jeżeli $L + D$ jest macierzą nieosobliwą, to otrzymujemy

  $$x^{(k)} = -(L + D)^{-1}Ux^{(k-1)} + (L + D)^{-1}b$$

  Twierdzenie: Jeżeli A jest silnie diagonalnie dominująca, to dla każdego wektora początkowego $x^{(0)}$ metoda Gaussa-Seidela tworzy ciąg zbieżny do rozwiązania układu $Ax = b$
- **Metoda SOR (nadrelaksacji):** Niech $A = L + D + U$ i $\omega \in \mathbb{R}$ jest postaci

  $$x^{(k)} = G_\omega x^{(k-1)} + w_\omega$$

  gdzie

  $$G_\omega = (D + \omega L)^{-1}((1 - \omega)D - \omega U)$$

  i

  $$w_\omega = \omega(D + \omega L)^{-1}b$$

  . **Twierdzenie:** Niech A będzie macierzą o dodatnich elementach diagolnych oraz $0 < \omega < 2$. Metoda nadrelaksacji jest zbieżna dla dowolnego wektora początkowego $x^{(0)}$ wtedy i tylko wtedy, gdy A jest symetryczna i dodatnio określona.
- **Twierdzenie:** Jeżeli $||I - C^{-1}|| < 1$ dla pewnej normy indukowanej macierzy, to ciąg określony równaniem $Cx^{(k)} = (C - A)x^{(k-1)} + b$ jest zbieżny do rozwiązania układu $Ax = b$ dla dowolnego wektora początkowego $x^{(0)}$.