Metody numeryczne w przykładach

Podręczniki – Politechnika Lubelska

Politechnika Lubelska
Wydział Elektrotechniki i Informatyki
ul. Nadbystrzycka 38A
20-618 Lublin

Beata Pańczyk
Edyta Łukasik
Jan Sikora
Teresa Guziak

Metody numeryczne w przykładach

Politechnika Lubelska
Lublin 2012

Recenzent:
dr hab. Stanisław Grzegórski, prof. Politechniki Lubelskiej

Redakcja i skład:
Beata Pańczyk
Edyta Łukasik

Publikacja wydana za zgodą Rektora Politechniki Lubelskiej

ISBN: 978-83-63569-14-3

Wydawca: Politechnika Lubelska
ul. Nadbystrzycka 38D, 20-618 Lublin
Realizacja: Biblioteka Politechniki Lubelskiej
Ośrodek ds. Wydawnictw i Biblioteki Cyfrowej
ul. Nadbystrzycka 36A, 20-618 Lublin
tel. (81) 538-46-59, email: wydawca@pollub.pl
www.biblioteka.pollub.pl
Druk: TOP Agencja Reklamowa Agnieszka Łuczak
www.agencjatop.pl

Elektroniczna wersja książki dostępna w Bibliotece Cyfrowej PL www.bc.pollub.pl
Nakład: 100 egz.

Spis treści

Wstęp

Metody numeryczne są obecnie przedmiotem ujętym w standardach kształcenia studentów uczelni technicznych i przede wszystkim do nich jest adresowany ten podręcznik. Znajomość metod numerycznych umożliwia właściwe wykorzystanie gotowych pakietów obliczeniowych (Matlab, Maple, Mathematica itp.), jak również daje niezbędne podstawy do samodzielnego rozwiązywania specyficznych, coraz bardziej złożonych problemów inżynierskich. Wymaga to z jednej strony świadomości istoty rozwiązywanych zagadnień, z drugiej zaś znajomości metod służących do ich rozwiązywania.

Treść niniejszego podręcznika stanowią wybrane zagadnienia z teorii i praktyki metod numerycznych. Teoretyczne podstawy bazują na pozycjach klasycznych [1, 4, 5, 8, 9, 10], które obejmują znacznie więcej materiału, niż można przedstawić w trakcie trzydziestogodzinnego wykładu. Niniejszy podręcznik zawiera tylko wyselekcjonowane informacje, które są omawiane na wykładach. Autorzy ograniczyli się do niezbędnych elementów teorii, bardziej koncentrując się na przykładach, dobranych w taki sposób, aby jak najprościej zobrazować działanie omawianej metody numerycznej. Niektóre przykłady obliczeń zostały dodatkowo przedstawione za pomocą tabel i rysunków. Rozdziały 3-7 zawierają odpowiednio opracowane zbiory zadań. Samodzielne rozwiązanie tych zadań pomoże Czytelnikowi w utrwaleniu prezentowanego materiału a dzięki zamieszczonym odpowiedziom umożliwi ich weryfikację.

Większość rozdziałów opracowano wykorzystując materiał z podręcznika [2], z którego usunięto rozdziały ukierunkowane na zastosowanie technik numerycznych w elektrotechnice a pozostałe uzupełniono zestawem bardziej uniwersalnych przykładów, bazując na wykładach prowadzonych w ostatnim dziesięcioleciu ze studentami kierunku Informatyka na Wydziale Elektrotechniki i Informatyki Politechniki Lubelskiej.

Cenne wskazówki i wnikliwe uwagi recenzenta dr hab. Stanisława Grzegórskiego, prof. Politechniki Lubelskiej, przyczyniły się do podniesienia merytorycznej jakości podręcznika.

Wszelkie uwagi i propozycje prosimy kierować do autorów na adres *b.panczyk@pollub.pl* lub *e.lukasik@pollub.pl.*

Autorzy

1. Błędy obliczeń numerycznych

1.1. Wstęp

W teorii metod numerycznych zasadniczą rolę odgrywa zrozumienie ograniczeń danej metody, co jest z kolei ściśle związane z określeniem błędu obliczeniowego. W niniejszym rozdziale przedstawimy podstawowe definicje i problemy dotyczące obliczeń numerycznych.

Przez ***zadanie numeryczne*** rozumiemy jasny i jednoznaczny opis powiązania funkcjonalnego między danymi wejściowymi (zmienne niezależne) i danymi wyjściowymi (szukanymi wynikami). ***Algorytm*** dla danego zadania numerycznego jest z definicji pełnym opisem poprawnie określonych operacji przekształcających dopuszczalne dane wejściowe na dane wyjściowe. „Operacje" oznaczają tu działania arytmetyczne i logiczne. Dla danego zadania numerycznego można rozważać wiele różnych algorytmów. Algorytmy te mogą dawać wyniki o bardzo różnej dokładności [1].

1.2. Podstawowe pojęcia szacowania błędów

Mówiąc o błędach numerycznych należy poznać podstawowe pojęcia z nimi związane, które krótko omówiono w kolejnych podrozdziałach [1].

1.2.1. Źródła błędów

Do źródeł błędów można zaliczyć:

a) błędy danych wejściowych (gdy wykorzystujemy dane zaokrąglone, pochodzące np. z wcześniejszych obliczeń lub gdy dane wejściowe są wynikiem pomiarów wielkości fizycznych obarczonych błędami pomiarowymi),
b) błędy zaokrągleń w czasie obliczeń (związane z odpowiednią reprezentacją liczby - patrz rozdział 1.3),
c) błędy obcięcia (gdy proces obliczania granicy jest przerywany przed osiągnięciem wartości granicznej - np. ograniczenie szeregu nieskończonego do skończonej liczby składników, aproksymacja pochodnej za pomocą ilorazu różnicowego, obliczanie wartości

całki oznaczonej jako granicy sum przybliżających ją podziałów itp.),

d) uproszczenie modelu matematycznego (przyjęcie założeń upraszczających),
e) błędy programisty.

1.2.2. Błędy względne i bezwzględne

Załóżmy, że wartość x jest reprezentowana jako $\tilde{x}$. Wówczas:

- ***błąd bezwzględny*** reprezentacji jest równy $\tilde{x} - x$.
- ***błąd względny*** [%] reprezentacji jest równy $\frac{\tilde{x} - x}{x} \cdot 100\%$, $x \neq 0$.

Przyjmijmy, że zapis $x = \tilde{x} \pm \varepsilon$ oznacza, że $|\tilde{x} - x| \leq \varepsilon$. Wartość $\varepsilon = \max|x - \tilde{x}|$ nazywamy ***maksymalnym błędem bezwzględnym*** lub ***błędem granicznym***.

Mówiąc o liczbie ***cyfr istotnych*** w ułamku dziesiętnym nie uwzględnia się zer na początku tego ułamka, gdyż określają one tylko pozycję kropki dziesiętnej. Natomiast ***cyfry ułamkowe*** są to wszystkie cyfry po kropce dziesiętnej, także ewentualne zera.

Jeśli $|\tilde{x} - x| \leq \frac{1}{2} \cdot 10^{-t}$, to mówimy, że $\tilde{x}$ ma t ***poprawnych cyfr ułamkowych***. Cyfry istotne występujące aż do pozycji t-tej po kropce nazywamy ***cyframi znaczącymi***.

Przykład 1.1.

W kolejnych przykładach podano liczby odpowiednich cyfr:
0.00147 – 5 cyfr ułamkowych, 3 cyfry istotne,
12.34 – 2 cyfry ułamkowe, 4 cyfry istotne,
0.001234 ±0.000004 – 5 cyfr poprawnych, 3 cyfry znaczące,
0.001234 ± 0.000006 – 4 cyfry poprawne, 2 cyfry znaczące.

1.2.3. Przenoszenie się błędów

Przenoszenie się błędów numerycznych najlepiej zobrazuje przykład 1.2.

Przykład 1.2.

Niech $x_1 = 2.31 \pm 0.02, x_2 = 1.42 \pm 0.03$.

Obliczmy różnicę:

$x_1 - x_2$: 2.33 -1.39 = 0.94, 2.29 -1.45 = 0.84,

czyli:

$x_1 - x_2$ = 0.89 ± 0.05.

Ogólnie mamy:

$$(\tilde{x}_1 - \varepsilon_1) - (\tilde{x}_2 + \varepsilon_2) \le x_1 - x_2 \le \tilde{x}_1 + \varepsilon_1 - (\tilde{x}_2 - \varepsilon_2).$$

Zatem:

$$x_1 + x_2 = \tilde{x}_1 + \tilde{x}_2 \pm (\varepsilon_1 + \varepsilon_2),$$

gdzie:

$\varepsilon_1, \varepsilon_2$ są maksymalnymi błędami bezwzględnymi.

Błąd bezwzględny sumy i różnicy jest więc równy: $\varepsilon_1 + \varepsilon_2$.

W przypadku obliczania iloczynu lub ilorazu przenoszenie się błędów przedstawimy za pomocą błędów względnych.

Niech *r* będzie rzeczywistym błędem względnym, tzn.

$$\tilde{x} = x + rx = x(1 + r).$$

Weźmy:

$$\tilde{x}_1 = x_1(1 + r_1), \tilde{x}_2 = x_2(1 + r_2).$$

Wówczas:

$$\tilde{x}_1\tilde{x}_2 = x_1x_2(1 + r_1)(1 + r_2), \frac{\tilde{x}_1}{\tilde{x}_2} = \frac{x_1(1 + r_1)}{x_2(1 + r_2)}.$$

Błąd względny iloczynu jest zatem równy:

$$(1 + r_1)(1 + r_2) - 1 = 1 + r_1 + r_2 + r_1r_2 - 1 \approx r_1 + r_2,$$

jeśli tylko $|r_1| << 1, |r_2| << 1$.

Błąd względny ilorazu jest równy:

$$\frac{(1+r_1)}{(1+r_2)}-1=\frac{1+r_1-1-r_2}{1+r_2}=\frac{r_1-r_2}{1+r_2}\approx r_1-r_2\text{, jeśli } |r_2|<<1.$$

1.3. Reprezentacja stałopozycyjna i zmiennopozycyjna

Reprezentacja stałopozycyjna operuje na ustalonej liczbie cyfr ułamkowych – wszystkie liczby rzeczywiste skraca się do *t* cyfr ułamkowych. Długość słowa maszynowego jest zwykle stała (np. *s* cyfr), więc dopuszcza się tylko liczby z przedziału: $\left(-10^{s-t}, 10^{s-t}\right)$.

Reprezentacja zmiennopozycyjna operuje natomiast na ustalonej liczbie cyfr istotnych.

W przypadku reprezentacji stałopozycyjnej liczbę całkowitą *l* przedstawiamy za pomocą ***rozwinięcia dwójkowego*** w postaci:

$$l=s\sum_{i=0}^{n}b_i 2^i\text{, gdzie } b_n\neq 0 \text{ dla } l\neq 0, s=\pm 1,\ b_i=0 \text{ lub } 1.$$

Jeśli $n < t$, to liczba *l* jest reprezentowana w rozpatrywanej arytmetyce „znak-moduł” *t*-cyfrowej. Liczby dokładnie reprezentowane w tej arytmetyce należą do przedziału: $\left\langle -2^t+1, 2^t-1\right\rangle$.

Liczbę rzeczywistą $x\neq 0$ można przedstawić także w postaci zmiennopozycyjnej:

$$x=s2^c m,$$

gdzie:

$s=\pm 1$ (znak liczby),

c – ***cecha*** liczby (liczba całkowita),

m – ***mantysa*** liczby, $m\in<0.5,1)$ tzn. rozwinięcie dwójkowe tej liczby jest takie, że pierwsza cyfra po przecinku jest różna od zera.

Ponadto:

d – bitów przeznacza się na reprezentację mantysy,

t-d-1 – bitów przeznacza się na reprezentację cechy.

Rozwinięcie dwójkowe mantysy jest na ogół nieskończone:

$$m = \sum_{i=1}^{\infty} b_{-i} 2^{-i} , \tag{1.1}$$

dlatego zapamiętuje się tylko d-początkowych cyfr dwójkowych, stosując zokrąglenie (d+1)-go bitu.

Jeśli:

$$m_d = \sum_{i=1}^{d} b_{-i} 2^{-i} + b_{(-d-1)} 2^{-d} , \tag{1.2}$$

to zakłada się, że mantysa m została prawidłowo zaokrąglona do d cyfr dwójkowych.

Przykład 1.3.

Niech $d = 5$, $x = 0.66$.

Wyznaczając reprezentację dwójkową liczby x, po pięciu krokach otrzymujemy:

x=0.	66
1	32
0	64
1	28
0	56
1	12

Wynik rozwinięcia dwójkowego, czyli: 0.010101, zwiększamy o 0.00001 (ponieważ $d = 5$), w celu otrzymania prawidłowo zaokrąglonej mantysy:

0.010101
+0.00001
0.01011 (prawidłowo zaokrąglona mantysa).

Reprezentację zmiennopozycyjną liczby x oznaczamy rd(x), tzn:

$$\mathrm{rd}(x) = s2^{c} m_d . \tag{1.3}$$

Z (1.1) i (1.2) otrzymujemy:

$$\left| m - m_d \right| \le \frac{1}{2} 2^{-d} \quad (\textbf{\textit{błąd bezwzględny reprezentacji danych}}). \tag{1.4}$$

Z (1.3) dla $x \neq 0$ mamy:

$$\left|\frac{\mathrm{rd}(x)-x}{x}\right| \leq 2^{-d} \quad (\textbf{\textit{błąd względny reprezentacji danych}}) \tag{1.5}$$

lub inaczej rd(x) = $x(1+\varepsilon)$, gdzie $|\varepsilon| \leq 2^{-d}$.

Liczba cyfr cechy decyduje o zakresie liczb zmiennopozycyjnych. Liczba cyfr mantysy decyduje o dokładności liczb zmiennopozycyjnych. Cecha:

$c \in \langle c_{\min}, c_{\max} \rangle$, gdzie $c_{\min} = -c_{\max}$, $c_{\max} = 2^{t-d-1}$.

Jeśli cecha danej liczby $c < c_{\min}$, występuje ***niedomiar pozycyjny*** (liczba jest reprezentowana za pomocą samych zer, co powoduje dużą niedokładność i na ogół przerwanie obliczeń lub przyjmuje się, że jest to liczba równa zero). Jeśli $c > c_{\max}$ - występuje ***nadmiar pozycyjny*** i też na ogół przerwanie systemowe programu.

1.4. Błędy zaokrągleń obliczeń zmiennopozycyjnych

Ze względu na ograniczoną długość słów binarnych konieczne jest zaokrąglanie obliczanych wartości, co powoduje pojawienie się ***błędów zaokrągleń***. Zaokrąglanie występuje w przypadku reprezentacji wszystkich liczb niewymiernych (tj. o nieskończonym rozwinięciu dwójkowym), takich jak np. π czy liczba Eulera.

Przykład 1.4.

Dzielenie dwóch liczb wymiernych prowadzi często do konieczności zaokrąglenia powstałej w wyniku dzielenia liczby niewymiernej,np.:

$\frac{1}{3} = 0.333333... \approx 0.333333$,

$\frac{1}{6} = 0.666666... \approx 0.166667$.

Jeśli x i y są dwiema liczbami zmiennopozycyjnymi to wyniki dodawania, odejmowania, mnożenia i dzielenia są zapamiętywane przez maszynę po zaokrągleniu lub ucięciu.

Wyniki tych działań oznacza się jako:
fl(*a*+*b*),
fl(*a*-*b*),
fl(*a***b*),
fl(*a*/*b*).

Wyniki te są odpowiednio równe zaokrąglonej lub uciętej wartości dokładnego wyniku działania oraz:

$$\boldsymbol{fl}(\boldsymbol{a} \bullet \boldsymbol{b}) = \mathrm{rd}(\boldsymbol{a} \bullet \boldsymbol{b})(1+\boldsymbol{\varepsilon}), \text{ gdzie } |\varepsilon| \leq 2^{-d}, \tag{1.6}$$

a znak • oznacza jeden z symboli działania: +, -, *, /.

Działania $fl(a \bullet b)$ mają, do pewnego stopnia inne własności niż dokładne działania arytmetyczne. Dla dodawania zmiennopozycyjnego nie zachodzi na ogół łączność, co pokazuje przykład 1.5.

Przykład 1.5.

Rozważmy dodawanie $a+b+c$ z siedmiocyfrowymi mantysami dla:

$a = 0.1234567 \cdot 10^0, b = 0.4711325 \cdot 10^4, c = -b.$

Obliczenia wykonamy zmieniając kolejność obliczeń zmiennopozycyjnych:

a) $fl(a + fl(b+c))$,

b) $fl(fl(a+b)+c)$.

Ad a)

$fl(b+c) = 0,$

$fl(a + fl(b+c)) = a = 0.1234567 \cdot 10^0.$

Ad b)

a	0.0000123	$4567 \cdot 10^4$
$+\ b$	0.4711325	$\cdot 10^4$
$fl(a+b)$	0.4711448	$\cdot 10^4$
$+c$	-0.4711325	$\cdot 10^4$
$fl(fl(a+b)+c)$	0.0000123	$\cdot 10^4$

Otrzymujemy:

$fl(a + fl(b+c)) = 0.1234567 \cdot 10^0.$

$fl(fl(a+b)+c) = 0.0000123 \cdot 10^4 = 0.1230000 \cdot 10^0.$

Jak widać nie jest spełnione prawo łączności dodawania w obliczeniach zmiennopozycyjnych bowiem:

$$fl(a+fl(b+c)) \neq fl(fl(a+b)+c).$$

Podczas wszelkich operacji arytmetycznych jesteśmy narażeni na kumulowanie się błędów a nawet na ich powstawanie wskutek ograniczonej dokładności reprezentacji wyniku działań w pamięci komputera. Przykładowo dodając do siebie bardzo dużą i bardzo małą liczbę w wyniku możemy otrzymać wartość większej zamiast sumy tych liczb, co pokazuje przykład 1.6.

Przykład 1.6.

Błąd wynikający z niewystarczająco dużej mantysy można pokazać na przykładzie obliczania sumy dwóch liczb:

a = 231 000 000.0

b = 0.000 000 384

$a+b$ = 231 000 000.000 000 384

Jeśli długość mantysy wynosi np. 10 cyfr znaczących, to dla dodawania zmiennopozycyjnego otrzymamy:

$fl(a+b)$ = 231 000 000.0

czyli:

$fl(a+b) = a$ (!!!).

Wskutek powyższego może pojawić się problem np. pętli nieskończonej, jeśli w kolejnych iteracjach warunek zakończenia bazuje na dodawaniu bardzo małej do bardzo dużej wartości.

Błędy wynikające z reprezentacji liczb można zmniejszyć ustalając umiejętnie ***sposób i kolejność wykonywanych działań***. Wpływ arytmetyki zmiennopozycjnej na wynik obliczeń w zależności od zastosowanego algorytmu pokazuje przykład 1.7.

Przykład 1.7.

Dla danych a, b obliczyć wartość wyrażenia $w=a^2-b^2$. Zakładamy, że mantysa jest reprezentowana na d bitach oraz błędy reprezentacji wynoszą $\varepsilon_i \leq 2^{-d}$.

Obliczenia wykonamy dwoma algorytmami.

Algorytm 1

Obliczamy kolejno: *x=a*a, y=b*b, w=x-y*.

Z wzoru (1.6) otrzymujemy:

$$fl(x) = (a*a)(1+\varepsilon_1),$$
$$fl(y) = (b*b)(1+\varepsilon_2),$$
$$fl(w) = [(a*a)(1+\varepsilon_1) - (b*b)(1+\varepsilon_2)](1+\varepsilon_3) =$$
$$= (a^2 - b^2)\left[\frac{a^2\varepsilon_1 - b^2\varepsilon_2}{a^2 - b^2} + 1\right](1+\varepsilon_3) = (a^2 - b^2)(1+\eta_1),$$

gdzie:

$$(1+\eta_1) = \left[1 + \frac{a^2\varepsilon_1 - b^2\varepsilon_2}{a^2 - b^2}\right](1+\varepsilon_3) =$$
$$= \left[1 + \frac{a^2\varepsilon_1 - b^2\varepsilon_2}{a^2 - b^2}\right] + \varepsilon_3 + O(\varepsilon_1, \varepsilon_1, \varepsilon_1),$$

a $O(\varepsilon_1, \varepsilon_1, \varepsilon_1)$ jest pomijalnie małe.

Jeśli $a^2 \approx b^2$ (mianownik bliski 0) oraz ε_1 i ε_2 mają przeciwne znaki to błąd η_1 może być bardzo duży.

Algorytm 2

Tym razem obliczamy kolejno: *x=a+b, y=a-b, w=x*y*.

Z wzoru (1.6) otrzymujemy:

$$fl(x) = (a+b)(1+\varepsilon_1),$$
$$fl(y) = (a-b)(1+\varepsilon_2),$$
$$fl(w) = [(a+b)(1+\varepsilon_1)(a-b)(1+\varepsilon_2)](1+\varepsilon_3) = (a^2 - b^2)(1+\eta_2),$$

gdzie:

$$(1+\eta_2) = (1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3) =$$
$$= 1 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_1\varepsilon_2 + \varepsilon_1\varepsilon_3 + \varepsilon_3\varepsilon_2 + \varepsilon_1\varepsilon_2\varepsilon_3 \approx 1 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3.$$

W tym przypadku, niezależnie od wartości *a* i *b*, mamy zawsze:

$\eta_2 \le 3 \cdot 2^{-d}$.

Błąd dla drugiego algorytmu jest mniejszy i nie zależy od wartości *a* i *b*.

1.5. Algorytm numerycznie stabilny i poprawny

Wiemy już, że jeśli nawet argumenty działania matematycznego wyrażają się w komputerze dokładnie to nie jest pewne, że wynik tego działania również będzie dokładny. W rozdziale 1.4 pokazalismy, że wynik każdej operacji arytmetycznej jest obarczony błędem reprezentacji liczby zmiennopozycyjnej. W przypadku algorytmów niestabilnych prowadzi to często do wyników niezgodnych z oczekiwaniami nawet co do znaku.

Niestabilność numeryczna powstaje wówczas, kiedy mały błąd numeryczny w trakcie dalszych obliczeń powiększa się (np. przemnaża się) i powoduje duży błąd wyniku. Niestabilność numeryczną możemy łatwo zaobserwować obliczając ciąg całek pokazany w przykładzie 1.8.

Przykład 1.8.

Obliczyć dla $n = 0,1,..,15$ całki: $y_n = \int_0^1 \frac{x^n}{x+5}\,\mathrm{d}x$.

Zauważmy, że:

$$y_n + 5y_{n-1} = \int_0^1 \frac{x^n + 5x^{n-1}}{x+5}\,\mathrm{d}x = \int_0^1 \frac{x^{n-1}(x+5)}{x+5}\,\mathrm{d}x = \frac{x^n}{n}\Big|_0^1 = \frac{1}{n}.$$

Otrzymujemy wobec tego wzór rekurencyjny:

$$y_n + 5y_{n-1} = \frac{1}{n},$$

na podstawie którego zbudujemy dwa algorytmy. Oba algorytmy zaimplementowano w języku C a przykładowe obliczenia realizowano z zastosowaniem reperezentacji zmiennopozycyjnej pojedynczej precyzji (wartości rzeczywiste reprezentowane na 4 bajtach).

Algorytm 1

Korzystając z wzoru:

$$y_n = \frac{1}{n} - 5y_{n-1},$$

mamy:

$$y_0 = \int_0^1 \frac{\mathrm{d}x}{x+5} = \ln(x+5)\,\Big|_0^1 = \ln 6 - \ln 5 \approx 0.182\,32156.$$

Przyjmując $y_0 = 0.182321556$, obliczamy kolejno:

$y_1 = 1 - 5y_0 \approx 0.0884$,

$y_2 = 1/2 - 5y_1 \approx 0.0580$,

$y_3 = 1/3 - 5y_2 \approx 0.0431$,

.....

$y_{10} = 1/9 - 5y_9 \approx 0.0040$,

$y_{11} \approx 1/10 - 5y_{10} \approx -0.3071$ (wynik błędny – wartość całki oznaczonej ujemna?).

Powodem otrzymania złego wyniku jest to, że błąd zaokrąglenia ε wartości y_0 jest mnożony przez -5 przy obliczaniu y_1. Tak więc wartość y_1 jest obarczona błędem -5ε. Ten błąd tworzy błąd 25ε w y_2, -125ε w y_3 itd. Nakładają się na to błędy zaokrąglenia popełniane w kolejnych krokach obliczeń, mające jednak stosunkowo małe znaczenie. Podstawiając $y_0 = \ln 6 - \ln 5$ popełniamy mniejszy błąd zaokrąglenia, który także powoduje duże zniekształcenie wyniku obliczeń y_i, dla $i > 16$. Oczywiście otrzymywane wyniki zależą także od precyzji, z jaką przeprowadzano obliczenia. Dla podwójnej precyzji obliczeń (reprezentacja liczby na 8 bajtach), wyraźnie błędne wyniki występują dla $i > 20$.

Algorytm 2

Ten sam ciąg całek możemy wyznaczyć inaczej. Jeśli przekształcimy wzór na zależność rekurencyjną tak, żeby obliczać w kolejnych iteracjach elementy ciągu w drugą stronę, mamy:

$$y_{n-1} = \frac{1}{5n} - \frac{1}{5} y_n .$$

Dzięki temu w każdym kroku błąd będzie dzielony przez -5. Ponieważ y_n maleje gdy n rośnie, możemy przypuszczać, że dla dużych n, y_n maleje wolno. Wobec tego przyjmując $y_{16} \approx y_{17}$ i korzystając z wzoru:

$$y_{16} = \frac{1}{5 \cdot 17} - \frac{1}{5} y_{17},$$

otrzymujemy:

$$y_{16} \approx \frac{1}{5 \cdot 17} - \frac{1}{5} y_{16} \Rightarrow y_{16} \approx \frac{1}{5 \cdot 17} \cdot \frac{5}{6} \approx 0.009803921 .$$

Następnie obliczamy:

$y_{15} = \frac{1}{5 \cdot 16} - \frac{1}{5} y_{16} \approx 0.0113,$

$y_{14} \approx 0.0120,$

$y_{13} \approx 0.0129,$

.....

$y_0 \approx 0.18232156$ (wynik poprawny).

W przypadku algorytmu 1 mamy do czynienia z niestabilnością numeryczną, bowiem małe błędy popełniane na pewnym etapie obliczeń rosną w następnych etapach i istotnie zniekształcają ostateczne wyniki (nawet co do znaku!).

Maksymalny przewidywalny błąd wynikły wyłącznie z przeniesienia błędu reprezentacji danych na wynik obliczeń nazywamy ***optymalnym poziomem błędu*** danego zadania w arytmetyce *t*-cyfrowej.

Algorytm ***stabilny*** gwarantuje otrzymanie wyniku akceptowanego z poziomem błędu tego samego rzędu, co optymalny poziom błędu.

Rozwiązanie obliczone algorytmem numerycznie ***poprawnym*** jest nieco zaburzonym rozwiązaniem zadania o nieco zaburzonych danych, tzn. jeśli dane są obarczone błędem, to i wynik jest obarczony porównywalnym błędem.

Stabilność jest minimalną własnością, jakiej wymagamy od algorytmu, ***poprawność*** maksymalną własnością jakiej możemy oczekiwać od zastosowanego algorytmu [1, 4, 5].

1.6. Uwarunkowanie zadania obliczeniowego

Okazuje się, że powszechna intuicja, że małe zaburzenia danych powinny dawać małe zaburzenia wyniku, nie znajduje potwierdzenia nawet w prostych przypadkach. Umiejętność oceny jakościowego ***wpływu zaburzenia danych na wynik*** jest podstawą w obliczeniach numerycznych.

Wrażliwość rozwiązania na dane początkowe określa tzw. ***uwarunkowanie zadania numerycznego***.

Zadanie jest źle uwarunkowane, jeśli małe (względne) zmiany w danych początkowych wywołują duże (względne) zmiany wyników. Zadanie źle

uwarunkowane obarczone jest dużymi błędami wyników niezależnie od obranej metody rozwiązywania.

Przypuśćmy, że zadanie obliczeniowe polega na wyznaczeniu *f(x)* dla danego *x*. Jak bardzo będzie odległe $f(\tilde{x})$, gdy $x \approx \tilde{x}$?

Rozważa się dwa przypadki:

- ***uwarunkowanie względne:*** jak względne zaburzenie danych wpływa na błąd względny wyniku:
  $$\frac{\|f(x)-f(\tilde{x})\|}{\|f(x)\|} \le cond_{rel}(f,x)\cdot\frac{\|x-\tilde{x}\|}{\|x\|}.$$
  Najmniejszy mnożnik $cond_{rel}(f,x)$ spełniający powyższą nierówność nazywamy ***współczynnikiem uwarunkowania (względnego)*** zadania obliczenia *f(x)* dla danego *x*.
- ***uwarunkowanie bezwzględne***: jak bezwzględne zaburzenie danych wpływa na błąd bezwzględny wyniku:
  $$\|f(x)-f(\tilde{x})\| \le cond_{abs}(f,x)\cdot\|x-\tilde{x}\|.$$
  Najmniejszy mnożnik $cond_{abs}(f,x)$ spełniający powyższą nierówność nazywamy ***współczynnikiem uwarunkowania (bezwzględnego)*** zadania obliczenia *f(x)* dla danego *x*.

Symbol $\|\cdot\|$ w powyższych nierównościach w ogólnym przypadku oznacza ***normę,*** czyli miarę pewnej odległości. Dla liczb rzeczywistych normą może być wartość bezwzględna. Przykłady norm dla wektorów i macierzy podano w rozdziale 2.2.4, natomiast normy definiowane jako miary odległości pomiędzy funkcjami rozważane są w rozdziale 3.5.1.

Mówimy, że zadanie *f(x)* jest

- ***dobrze uwarunkowane*** w punkcie *x*, gdy $cond(f,x) \approx 1$,
- ***źle uwarunkowane*** w punkcie *x*, gdy $cond(f,x) >> 1$,
- ***źle postawione*** w punkcie *x*, gdy $cond(f,x) = +\infty$.

2. Podstawy rachunku macierzowego

2.1.Wstęp

W metodach numerycznych wiodącą rolę odgrywają operacje macierzowe. Celowym jest wobec tego przypomnienie niezbędnych pojęć i definicji związanych z algebrą liniową [1].

W zastosowaniach matematyki do rozmaitych zagadnień naukowych bardzo często wykorzystuje się najprostszy typ operatorów - ***operatory liniowe***.

Oznaczmy pewną daną macierz kwadratową jako **A**. Podstawowymi problemami algebry liniowej będą:

- rozwiązanie układu równań **Ax** = **b**,
- rozwiązanie zadania własnego, czyli określenie wartości własnych λ_k i wektorów własnych $\mathbf{x}_k$ takich, że $\mathbf{A}\mathbf{x}_k = \lambda_k \mathbf{x}_k$ dla $k = 1,2,...,n$.

Rozwiązywanie układów równań liniowych jest zadaniem występującym często w różnych problemach inżynierskich. Nawet układy równań nieliniowych rozwiązuje się często przybliżając je ciągami układów liniowych (np. w metodzie Newtona – rozdział 5.5).

2.2. Podstawowe pojęcia algebry liniowej

Macierzą **A** nazywamy układ *m* x *n* liczb rzeczywistych lub zespolonych.

Liczby te są zgrupowane w tablicę:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}, \qquad (2.1)$$

gdzie *m* x *n* oznacza, że macierz ma *m* wierszy i *n* kolumn.

Definicję macierzy **A** można też zapisać krócej:

$$\mathbf{A} = (a_{ij}), \quad i = 1,2,...,m,\ j = 1,2,...,n.. \tag{2.2}$$

Jeśli $n = 1$, to macierz składa się tylko z jednej kolumny i nazywa się ***wektorem kolumnowym***, który oznaczamy jako:

$$\mathbf{x} = (x_i), \quad i = 1,2,...,m. \tag{2.3}$$

Dla $m = n$ macierz **A** nazywamy ***macierzą kwadratową***, natomiast n - ***stopniem macierzy kwadratowej***.

Jednym z rodzajów macierzy kwadratowych są ***macierze przekątniowe* (*diagonalne*)**, które mają wartości różne od zera tylko na głównej przekątnej (***diagonali***):

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & ... & 0 \\ 0 & d_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & d_n \end{bmatrix} = \mathrm{diag}(d_1, d_2, ..., d_n). \tag{2.4}$$

Natomiast szczególnym przypadkiem macierzy **D** jest ***macierz jednostkowa*** $\mathbf{I}_n$ stopnia n, określona wzorem:

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & 0 & 1 \end{bmatrix} = \mathrm{diag}(1,1,...,1) = (\delta_{ij}), \tag{2.5}$$

gdzie (δ_{ij}) jest ***symbolem*** (***deltą***) ***Kroneckera***:

$$\delta_{ij} = \begin{cases} 0 \ \text{dla} \ i \neq j, \\ 1 \ \text{dla} \ i = j. \end{cases} \tag{2.6}$$

Macierz jednostkową często określa się samym symbolem **I.**

Macierze **A** i **B** są sobie ***równe*** (**A** = **B**), jeśli mają takie same wymiary i ich wszystkie wyrazy są równe:

$a_{ij} = b_{ij}$ dla $i = 1,2,...,m;\ j = 1,2,...,n.$

Iloczyn macierzy **A** i liczby α jest macierzą $\alpha\ \mathbf{A} = (\alpha\ a_{ij})$.

Suma dwóch macierzy o takich samych wymiarach (**C** = **A** + **B**) jest macierzą o elementach $c_{ij} = a_{ij} + b_{ij}$.

Iloczyn dwóch macierzy: **A** (*m* x *p*) i **B** (*p* x *n*) jest macierzą **C** (*m* x *n*) o elementach obliczanych ze wzoru:

$$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj} \,. \tag{2.7}$$

Mnożenie macierzy na ogół nie jest przemienne, czyli:

$$\mathbf{AB} \neq \mathbf{BA}.$$

Inne własności mnożenia to:

$$\begin{aligned} &\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}, \\ &\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}. \end{aligned} \tag{2.8}$$

Transpozycją $\mathbf{A}^{\mathrm{T}}$ macierzy **A** nazywamy macierz, której wiersze są kolumnami macierzy **A**.

Jeśli $\mathbf{B} = \mathbf{A}^{\mathrm{T}}$, to $b_{ij} = a_{ji}$.

Wektor kolumnowy $\mathbf{x} = (x_1, x_2, ..., x_n)^{\mathrm{T}}$ jest transpozycją pewnego wektora wierszowego.

W przypadku transpozycji iloczynu macierzy występuje tożsamość:

$$(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}. \tag{2.9}$$

Macierzą sprzężoną $\overline{\mathbf{A}}$ nazywamy macierz zespoloną, której każdy element został zastąpiony liczbą z nim sprzężoną. Macierz $\overline{\mathbf{A}}^{\mathrm{T}}$ oznacza się symbolem $\mathbf{A}^{\mathrm{H}}$.

Macierz trójkątna ma postać:

$$\mathbf{L} = \begin{bmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{bmatrix} \text{ lub } \mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & r_{22} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & r_{nn} \end{bmatrix}, \tag{2.10}$$

przy czym **L** jest ***macierzą trójkątną lewą*** (***dolną***), a R - ***prawą*** (***górną***).

Sumy i iloczyny macierzy trójkątnych dolnych są także macierzami trójkątnymi dolnymi, a sumy i iloczyny macierzy trójkątnych górnych są macierzami trójkątnymi górnymi.

Wyznacznik macierzy kwadratowej **A** stopnia *n* ma symbol det(**A**):

$$\det(\mathbf{A}) = \det\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}. \tag{2.11}$$

Przy obliczaniu wyznacznika obowiązują wzory:

$$\text{dla } n = 1 \ \det(\mathbf{A}) = a_{11}, \tag{2.12}$$

$$\text{dla } n > 1 \ \det(\mathbf{A}) = a_{11}A_{11} - a_{12}A_{12} + \dots + (-1)^{n+1} a_{1n}A_{1n}, \tag{2.13}$$

gdzie A_{1k} $(k = 1,2,\dots,n)$ oznacza wyznacznik stopnia ***n***-1, który powstaje przez skreślenie pierwszego wiersza i ***k***-tej kolumny z macierzy **A**.

Dla dowolnej macierzy kwadratowej **A** obowiązują następujące reguły:

- wartość wyznacznika nie zmieni się, jeśli do wiersza (kolumny) doda się iloczyn innego wiersza (lub innej kolumny) przez liczbę,
- wyznacznik macierzy trójkątnej jest równy iloczynowi elementów głównej przekątnej: $\det(\mathbf{L}) = l_{11}l_{22}\dots l_{nn}$, $\det(\mathbf{R}) = r_{11}r_{22}\dots r_{nn}$,
- przestawienie dwóch wierszy lub dwóch kolumn zmienia jedynie znak wyznacznika,
- $\det(\mathbf{A}) = \det(\mathbf{A}^{\mathrm{T}})$,
- $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$.

Podwyznacznikiem A_{ik}, czyli ***minorem*** macierzy odpowiadającym elementowi a_{ik}, nazywamy wyznacznik podmacierzy stopnia *n*-1, która powstaje z danej macierzy przez opuszczenie *i*-tego wiersza i *k*-tej kolumny.

Macierz **A** jest nazywana ***nieosobliwą***, jeśli $\det(\mathbf{A}) \neq 0$, w przeciwnym wypadku określa się ją jako ***osobliwą***. Do każdej macierzy nieosobliwej istnieje macierz ***odwrotna*** $\mathbf{A}^{-1}$ spełniająca zależność $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. W przypadku odwrotności iloczynu spełniona jest tożsamość: $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

Macierz symetryczna jest równa swojej transpozycji ($\mathbf{A} = \mathbf{A}^{\mathrm{T}}$). Iloczyn dwóch macierzy symetrycznych **A** i **B** jest symetryczny tylko pod warunkiem, że **AB** = **BA**.

Macierz symetryczną nazywa się ***dodatnio określoną***, jeśli związana z nią forma kwadratowa $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ spełnia warunek $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0$ dla każdego rzeczywistego $\mathbf{x} \neq \mathbf{0}$.

Macierzą ortogonalną **Q** nazywamy macierz m x n spełniającą zależność $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$. W przypadku, gdy $m = n$, mamy $\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}^{-1}$ oraz $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}}$.

Rzeczywistym macierzom symetrycznym i ortogonalnym odpowiadają zespolone ***macierze hermitowskie*** o zależności $\mathbf{A}^{\mathrm{H}} = \mathbf{A}$ oraz ***macierze unitarne***, dla których $\mathbf{U}^{\mathrm{H}}\mathbf{U} = \mathbf{I}$.

2.2.1. Macierze blokowe

Dowolną macierz **A** można przedstawić jako ***macierz blokową***, zbudowaną z pewnej liczby macierzy o mniejszych wymiarach:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \ldots & \mathbf{A}_{1n} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \ldots & \mathbf{A}_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{A}_{m1} & \mathbf{A}_{m2} & \ldots & \mathbf{A}_{mn} \end{bmatrix}, \tag{2.14}$$

gdzie $\mathbf{A}_{ij}$ jest macierzą o wymiarach $p_i \mathrm{x} q_j$.

Najbardziej interesujący jest przypadek, kiedy macierze na przekątnej $\mathbf{A}_{ii}$ są kwadratowe. W takim przypadku macierz **A** również musi być kwadratowa, a $p_i = q_i$ $(i = 1,2,\ldots,n)$. Dodawanie i mnożenie takich macierzy blokowych wykonuje się tak samo, jak gdyby bloki były liczbami.

Na przykład, dla **C** = **AB** istnieje zależność:

$$\mathbf{C}_{ij} = \sum_{k=1}^{n} \mathbf{A}_{ik}\mathbf{B}_{kj} \,. \tag{2.15}$$

Macierzą blokowo-przekątniową nazywa się macierz, którą można zapisać w postaci blokowej jako:

$$\mathbf{A} = \mathrm{diag}(\mathbf{A}_{11}, \mathbf{A}_{22}, \ldots, \mathbf{A}_{nn}),$$

przy czym macierze $\mathbf{A}_{ii}$ muszą być kwadratowe.

Analogicznie definiuje się ***macierz blokowo-trójkątną***. Dla macierzy blokowo-trójkątnej lewej wyznacznik oblicza się ze wzoru:

$$\det(\mathbf{L}) = \det(\mathbf{L}_{11})\det(\mathbf{L}_{22})...\det(\mathbf{L}_{nn})$$

i podobnie dla macierzy blokowo-trójkątnej prawej.

Macierze blokowe mogą być także wykorzystywane do upraszczania obliczeń na liczbach zespolonych. Dowolną macierz zespoloną można zastąpić macierzą rzeczywistą, dwukrotnie większą. Ta sama reguła odnosi się także do wektorów.

Rozpatrzmy macierz zespoloną **A** i wektor zespolony **x**: $\mathbf{A} = \mathbf{B}+j\mathbf{C}$, $\mathbf{x} = \mathbf{y}+j\mathbf{z}$, gdzie $j = \sqrt{-1}$, **B** i **C** - macierze rzeczywiste, **y** i **z** - wektory rzeczywiste. Można wtedy zapisać, że:

$$\tilde{\mathbf{A}} = \begin{bmatrix} \mathbf{B} & -\mathbf{C} \\ \mathbf{C} & \mathbf{B} \end{bmatrix}, \tilde{\mathbf{x}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix}, \tag{2.16}$$

gdzie $\tilde{\mathbf{A}}$, $\tilde{\mathbf{x}}$ oznaczają rzeczywiste odpowiedniki **A** i **x**.

2.2.2. Przestrzenie liniowe wektorowe

Wektor możemy zdefiniować jako uporządkowany zbiór *n* liczb (rzeczywistych lub zespolonych) o postaci:

$(x_1, x_2, ..., x_n)$.

Zbiór wszystkich takich wektorów tworzy ***przestrzeń wektorową*** R_n (dla liczb rzeczywistych) lub C_n (dla liczb zespolonych) o wymiarze *n*. Jeśli założymy, że liczby x_i są współrzędnymi w układzie prostokątnym, to ***iloczyn skalarny*** wektorów **x** i **y** w przypadku zespolonym określa się wzorem:

$$(\mathbf{x}, \mathbf{y}) = (\bar{x}_1 y_1 + \bar{x}_2 y_2 + ... + \bar{x}_n y_n). \tag{2.17}$$

Jeśli **x** jest wektorem kolumnowym, to iloczyn skalarny (**x**,**y**) można uznać za przypadek szczególny mnożenia macierzy:

$$(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\mathrm{H}}\mathbf{y}.$$

Dla wektorów rzeczywistych ten sam wzór przyjmie postać:

$$(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\mathrm{T}}\mathbf{y}.$$

Wektor:

$$\mathbf{y} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + ... + c_k\mathbf{x}_k ,$$

nazywany jest ***kombinacją liniową*** wektorów $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k$.

Wektory te są nazywane ***liniowo niezależnymi***, jeśli równość:

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + ... + c_k\mathbf{x}_k = \mathbf{0}$$

zachodzi tylko pod warunkiem, że $c_1 = c_2 = ... = c_k = 0$.

W przeciwnym wypadku wektory te określa się jako ***liniowo zależne***.

Zbiór wszystkich możliwych kombinacji liniowych wektorów $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k$ nazywa się ***podprzestrzenią liniową*** R_k w R_n. Mówi się, że podprzestrzeń ta jest ***rozpięta*** na wektorach $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k$.

W przestrzeni R_n istnieje co najwyżej *n* wektorów liniowo niezależnych. Dowolny zbiór *n* takich wektorów: $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$ tworzy ***bazę*** przestrzeni R_n.

Każdy wektor **x** z R_n można więc wyrazić jako:

$$\mathbf{x} = \alpha_1\mathbf{y}_1 + \alpha_2\mathbf{y}_2 + ... + \alpha_n\mathbf{y}_n , \tag{2.18}$$

gdzie $\alpha_1, \alpha_2, ..., \alpha_n$ nazywa się ***współrzędnymi*** tego wektora względem bazy $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$.

Jako najprostszy przykład bazy w R_n można rozpatrzyć *n* kolumn macierzy jednostkowej $\mathbf{I} = (\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n)$.

Dowolną macierz **A** można uważać za zbudowaną z wektorów kolumnowych lub wierszowych:

$$\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_n), \mathbf{A} = (\mathbf{a}_1^{\mathrm{T}}, \mathbf{a}_2^{\mathrm{T}}, ..., \mathbf{a}_m^{\mathrm{T}})^{\mathrm{T}} , \tag{2.19}$$

gdzie $\mathbf{a}_i = (a_{1i}, a_{2i}, ..., a_{mi})$.

Największa liczba niezależnych liniowo wektorów kolumnowych w macierzy **A** jest równa największej liczbie niezależnych liniowo wektorów wierszowych w **A**. Jeśli liczba ta wynosi *r*, to oznacza ona ***rząd macierzy*** **A**, określany jako:

rank(**A**) = r,

przy czym $r \le \min(m,n)$.

W szczególnym przypadku, jeśli $r = m = n$, macierz **A** jest nieosobliwa.

Układ *m* równań z *n* niewiadomymi postaci:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = b_2 \\ ..... \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n = b_m \end{cases} \tag{2.20}$$

w symbolice macierzowej może zostać zapisany jako:

$\mathbf{Ax} = \mathbf{b}$.

Jeśli $\mathbf{b} = \mathbf{0}$, układ taki nazywa się ***jednorodnym***. Układ jednorodny ma zawsze trywialne rozwiązanie $\mathbf{x} = \mathbf{0}$.
Jeśli $\text{rank}(\mathbf{A}) = r < n$, to jednorodny układ $\mathbf{Ax} = \mathbf{0}$ ma *n*-*r* rozwiązań niezależnych liniowo.

Ten sam układ równań można także zapisać w postaci:

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + ... + x_n\mathbf{a}_n = \mathbf{b}. \tag{2.21}$$

Wektor **b** można zatem określić jako kombinację liniową wektorów kolumnowych macierzy **A**. Wynika z tego warunek rozwiązalności układu równań:

$\mathbf{Ax} = \mathbf{b},\ \mathbf{b} \in R(\mathbf{A})$,

gdzie $R(\mathbf{A})$ oznacza przestrzeń rozpiętą na kolumnach macierzy **A**.

Ten sam warunek można zapisać jako:

$\text{rank}(\mathbf{A},\mathbf{b}) = \text{rank}(\mathbf{A})$.

Jeśli $r = m = n$, to $R(\mathbf{A}) = R_n$ i warunek rozwiązalności jest spełniony przez każdy wektor **b**. W takim przypadku rozwiązanie jest jednoznacznie określone wzorem:

$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

Uwagi te są także prawdziwe w przypadku zespolonym.

2.2.3. Wartości własne

Jeśli dla danej liczby λ i wektora $\mathbf{x} \neq \mathbf{0}$ spełniona jest równość:

$\mathbf{Ax} = \lambda\mathbf{x}$,

to λ nazywa się ***wartością własną macierzy*** **A**, a **x** - ***wektorem własnym*** odpowiadającym wartości λ.

Równanie $\mathbf{Ax}=\lambda\mathbf{x}$ można także zapisać w postaci:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$$

i jest to układ jednorodny względem **x**.

Układ ten ma rozwiązanie $\mathbf{x} \neq \mathbf{0}$ tylko pod warunkiem, że:

$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$.

Równanie:

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

nazywane jest ***równaniem charakterystycznym*** macierzy **A**. Ponieważ jest ono równaniem *n*-tego stopnia względem λ, więc ma dokładnie *n* pierwiastków rzeczywistych lub zespolonych: $\lambda_1,\lambda_2,...,\lambda_n$ licząc krotności. Zbiór wartości własnych $(\lambda_1,\lambda_2,...,\lambda_n)$ macierzy kwadratowej **A** nazywamy ***widmem*** tej macierzy.

Jeśli **C** jest macierzą nieosobliwą to macierz:

$$\mathbf{B} = \mathbf{C}^{-1}\mathbf{AC},$$

nazywa się ***podobną*** do **A**, a przekształcenie **A** w **B** - ***przekształceniem macierzy A przez podobieństwo***. Jeśli λ jest wartością własną, a **x** -\odpowiednim wektorem własnym macierzy **A**, to zachodzą następujące zależności:

$$\mathbf{Ax} = \lambda\mathbf{x} \Rightarrow (\mathbf{C}^{-1}\mathbf{AC})\mathbf{C}^{-1}\mathbf{x} = \lambda\mathbf{C}^{-1}\mathbf{x}. \qquad (2.22)$$

Wynika z tego, że λ jest także wartością własną macierzy **B**, a wektorem własnym tej macierzy jest $\mathbf{y} = \mathbf{C}^{-1}\mathbf{x}$.

Każda macierz kwadratowa ma *n* wartości własnych i wektorów własnych: $\mathbf{Ax}_i = \lambda_i\mathbf{x}_i \quad i = 1,2,...,n$. Tę samą zależność można także zapisać krócej:

$$\mathbf{AX} = \mathbf{\Lambda X}, \text{ dla } \mathbf{X} = (\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n), \quad \Lambda = \mathrm{diag}(\lambda_1,\lambda_2,...,\lambda_n). \qquad (2.23)$$

Jeśli wektory własne $\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n$ są niezależne liniowo, to macierz **X** jest nieosobliwa, z czego wynika równość $\mathbf{\Lambda} = \mathbf{X}^{-1}\mathbf{AX}$. Poprzez przekształcenie macierzy **A** przez podobieństwo za pomocą **X** uzyskuje się macierz przekątniową. Taką macierz **A** nazywa się wtedy ***diagonalizowalną***.

Jeśli wszystkie wartości własne są różne ($\lambda_i \neq \lambda_j$ dla $i \neq j$), to wektory własne są zawsze liniowo niezależne. Istnieją jednak macierze z wielokrotnymi wartościami własnymi, których nie można diagonalizować. Najprostszym przykładem takiej macierzy jest:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}. \tag{2.24}$$

Jeżeli elementy macierzy **A** są liczbami rzeczywistymi, to jej wartości własne są liczbami rzeczywistymi lub zespolonymi parami sprzężonymi. Jeśli natomiast macierz **A** jest rzeczywista i symetryczna, to jej wartości własne są liczbami rzeczywistymi. Wektory własne $\mathbf{x}_i$ i $\mathbf{x}_j$ odpowiadające dwóm różnym wartościom własnym λ_i i λ_j są ortogonalne, czyli $\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j = 0$. Zawsze można tak dobrać wektory własne, odpowiadające wielokrotnej wartości własnej, aby były ortogonalne. Na przykład każdy wektor jest wektorem własnym macierzy jednostkowej **I**, odpowiadającym wartości własnej 1.

Z zależności $\mathbf{Ax} = \lambda\mathbf{x}$ wynika, że:

$$(\mathbf{A} + c\mathbf{I})\mathbf{x} = (\lambda + c)\mathbf{x},$$

$$\mathbf{A}^2\mathbf{x} = \lambda^2\mathbf{x}, \tag{2.25}$$

$$\mathbf{A}^{-1}\mathbf{x} = \frac{1}{\lambda}\mathbf{x} \text{ dla } \lambda \neq 0.$$

Oznacza to, że macierze o postaci $\mathbf{A}^{\pm n}$ mają wartości własne $\lambda^{\pm n}$, a macierz $\mathbf{A} + c\mathbf{I}$ - wartości własne $\lambda + c$.

Ogólnie, jeśli $P(z) = a_0 z^n + a_1 z^{n-1} + \ldots + a_n$, to macierz $\mathbf{P}(\mathbf{A})$ ma wartości własne $P(\lambda)$.

2.2.4. Normy wektorów i macierzy

Normą macierzy lub wektora nazywamy liczbę nieujemną, będącą w pewnym sensie miarą wielkości tej macierzy lub wektora.

Normę wektora definiujemy jako:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}. \tag{2.26}$$

Często używa się dwóch szczególnych przypadków:

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} \text{ dla } p = 2 \text{ (}\textbf{\textit{norma euklidesowa)}} \tag{2.27}$$

$$\|\mathbf{x}\|_\infty = \max_i |x_i| \text{ dla } p \to \infty \text{ (}\textbf{\textit{norma maksymalna}}\text{)} \tag{2.28}$$

Normy wektorów muszą mieć następujące własności:

$$\|\mathbf{x}\| > 0 \text{ dla } \mathbf{x} \neq \mathbf{0},\ \|\mathbf{x}\| = 0 \text{ dla } \mathbf{x} = \mathbf{0},$$

$$\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|, \text{ gdzie } \alpha \text{ - dowolna liczba,} \tag{2.29}$$

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|.$$

Natomiast norma macierzy musi spełniać następujące warunki:

$$\|\mathbf{A}\| > 0 \text{ dla } \mathbf{A} \neq \mathbf{0},\ \|\mathbf{A}\| = 0 \text{ dla } \mathbf{A} = \mathbf{0},$$

$$\|\alpha\mathbf{A}\| = |\alpha| \cdot \|\mathbf{A}\|, \tag{2.30}$$

$$\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|,$$

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|.$$

Trzecia nierówność we wzorach (2.29) i (2.30) nazywana jest ***nierównością trójkąta***, natomiast ostatnia nierówność we wzorze (2.30) - ***nierównością Schwarza***. Z tego ostatniego warunku wynika w szczególności, że:

$$\|\mathbf{A}^m\| \leq \|\mathbf{A}\|^m. \tag{2.31}$$

Normę euklidesową macierzy określa się wzorem:

$$\|\mathbf{A}\|_2 = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}. \tag{2.32}$$

Jeśli norma macierzy i norma wektora są tak ze sobą związane, że spełniona jest nierówność:

$$\|\mathbf{Ax}\| \le \|\mathbf{A}\| \cdot \|\mathbf{x}\| . \tag{2.33}$$

to te dwie normy nazywa się ***zgodnymi***.

Dla każdej normy wektora istnieje zgodna z nią norma macierzy. Jest to tzw. norma macierzy ***indukowana przez normę wektora***:

$$\|\mathbf{A}\| = \max \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|}, \text{ dla } \mathbf{x} \neq \mathbf{0} . \tag{2.34}$$

Normę macierzy indukowaną przez normę maksymalną wektora (2.28) oblicza się ze wzoru:

$$\|\mathbf{A}\|_{\infty} = \max_i \sum_{j=1}^{n} |a_{ij}| . \tag{2.35}$$

3. Interpolacja i aproksymacja

3.1. Wstęp

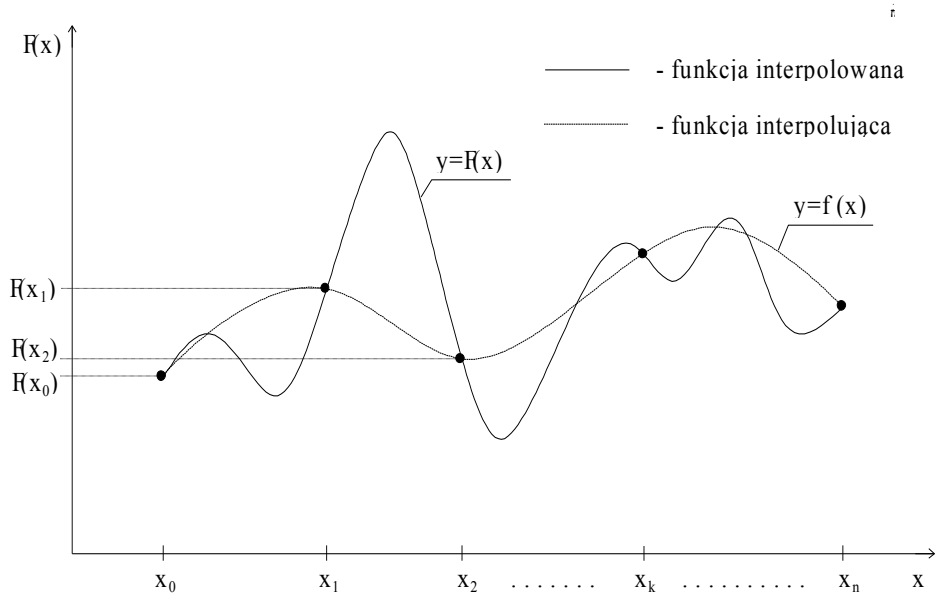
Wiele zjawisk fizycznych jest opisywanych przez funkcje, których postaci nie znamy, ale potrafimy obliczyć lub zmierzyć wartości tych funkcji oraz ich pochodnych dla określonych wartości argumentu. Na przykład, możemy dysponować zbiorem pomiarów wartości pewnych parametrów uzyskanych w określonych chwilach czasowych.

Interpolacją nazywamy postępowanie prowadzące do znalezienia wartości pewnej funkcji $f(x)$ w dowolnym punkcie przedziału (x_0, x_n) na podstawie znanych wartości tej funkcji w punktach $x_0, x_1, ..., x_n$, zwanych ***węzłami interpolacji*** ($x_0 < x_1 < ... < x_n$) [1, 4, 8, 9, 10].

Postępowanie prowadzące do znalezienia wartości funkcji $f(x)$ w punkcie leżącym poza przedziałem (x_0, x_n) nazywamy ***ekstrapolacją***. Interpolację lub ekstrapolację stosujemy najczęściej w następujących przypadkach:

- gdy nie znamy samej funkcji $f(x)$, a tylko jej wartości w pewnych punktach (tak przeważnie bywa w naukach doświadczalnych);
- gdy obliczanie wartości pewnej funkcji $F(x)$ bezpośrednio z określającego ją wzoru nastręcza zbyt duże trudności rachunkowe; wtedy zastępujemy ją prostszą funkcją $f(x)$, o której zakładamy, że w punktach $x_0, x_1, ..., x_n$ ma te same wartości co funkcja $F(x)$; w tym przypadku $F(x)$ nazywamy ***funkcją interpolowaną***, zaś $f(x)$ ***funkcją interpolującą*** (rys. 3.1).

Funkcji interpolującej poszukuje się zwykle w pewnej określonej postaci. Najczęściej zakłada się, że jest to wielomian lub funkcja wymierna. Przedmiotem naszych rozważań będzie interpolacja wielomianami algebraicznymi, wielomianami trygonometrycznymi oraz funkcjami sklejanymi. Obecnie stosuje się albo proste metody, jak interpolacja liniowa czy kwadratowa, albo też bardziej złożone, wymagające użycia komputera, jak np. interpolacja za pomocą funkcji sklejanych.



Rys. 3.1. Interpretacja geometryczna zagadnienia interpolacji

Wzory interpolacyjne są punktem wyjścia do wyprowadzenia wielu metod stosowanych w różnych działach metod numerycznych (rozwiązywanie równań, różniczkowanie i całkowanie numeryczne, numeryczne rozwiązywanie równań różniczkowych zwyczajnych).

3.2. Interpolacja wielomianowa

3.2.1. Jednoznaczność rozwiązania zagadnienia interpolacyjnego

Wielomianem interpolacyjnym $W_n(x)$ nazywamy wielomian stopnia co najwyżej n, który w punktach $x_0, x_1, ..., x_n$ spełnia warunki interpolacji:

$$W_n(x_i)=y_i \quad \text{dla} \quad i=0,1,2,\ldots,n. \tag{3.1}$$

Twierdzenie 3.1.

Istnieje dokładnie jeden wielomian interpolacyjny, który w punktach $x_0, x_1, ..., x_n$ przy założeniu, że $x_i \neq x_j \; dla \; i \neq j$ spełnia warunki interpolacji (3.1).

Węzły interpolacji $x_0, x_1, ..., x_n$ mogą być rozmieszczone w zupełnie dowolny sposób a szukany wielomian można zapisać w postaci:

$$\boldsymbol{W_n(x)=a_0+a_1x+a_2x^2+\ldots+a_nx^n}. \tag{3.2}$$

Korzystając z definicji wielomianu interpolacyjnego otrzymujemy układ n+1 równań z n+1 niewiadomymi współczynnikami $a_0, a_1, ..., a_n$:

$$\begin{cases} a_0 + a_1x_0 + a_2x_0^2 + \cdots + a_nx_0^n = y_0 \\ a_0 + a_1x_1 + a_2x_1^2 + \cdots + a_nx_1^n = y_1 \\ \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_0 + a_1x_n + a_2x_n^2 + \cdots + a_nx_n^n = y_n \end{cases} \tag{3.3}$$

Macierz współczynników tego układu jest ***macierzą Vandermonde'a*** postaci:

$$\mathbf{A} = \begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{bmatrix}, \tag{3.4}$$

zaś wyznacznik D macierzy **A** ma postać:

$$\mathrm{D} = \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{vmatrix}. \tag{3.5}$$

Przy założeniach, że $x_i \neq x_j$ dla $i \neq j$, mamy zawsze $\mathrm{D} = \prod_{1 \le j < i \le n} (x_i - x_j) \neq 0$.

Zatem układ (3.3) ma dokładnie jedno rozwiązanie, a wartości a_i według twierdzenia Cramera są określone wzorem:

$$a_i = \frac{1}{D} \sum_{j=0}^{n} y_j D_{ij}\,, \tag{3.6}$$

gdzie D_{ij} (j = 0, 1, 2, ..., n) są kolejnymi dopełnieniami algebraicznymi elementów i-tej kolumny wyznacznika D.

3.2.2. Wielomian interpolacyjny Lagrange'a

Twierdzenie 3.2.

Wielomian $W_n(x)$ postaci (3.7) jest wielomianem interpolacyjnym dla dowolnego wyboru n+1 węzłów interpolacji x_0, x_1, ..., x_n takich, że $x_i \neq x_j$ dla $i \neq j$.

$$\begin{aligned} W_n(x) = y_0 \frac{(x-x_1)(x-x_2)...(x-x_n)}{(x_0-x_1)(x_0-x_2)...(x_0-x_n)} + \\ + y_1 \frac{(x-x_0)(x-x_2)...(x-x_n)}{(x_1-x_0)(x_1-x_2)\cdot....\cdot(x_1-x_n)} + ... + \\ + y_n \frac{(x-x_0)(x-x_1)...(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)...(x_n-x_{n-1})} = \sum_{j=0}^{n} y_j \prod_{\substack{k=0 \\ k\neq j}}^{n} \frac{(x-x_k)}{(x_j-x_k)}. \end{aligned} \tag{3.7}$$

Przyjmując oznaczenia:

$\omega_k(x) = (x - x_0)(x - x_1)...(x - x_n)$,

$\omega_n'(x_j)$ - wartość pochodnej wielomianu $\omega_n(x)$ w punkcie x_j będącym zerem tego wielomianu,

wielomian interpolacyjny $W_n(x)$ można zapisać w postaci:

$$W_n(x) = \sum_{j=0}^{n} y_j \frac{\omega_n(x)}{(x-x_j)\omega_n'(x_j)}. \tag{3.8}$$

Wzór (3.7) nazywamy ***wzorem interpolacyjnym Lagrange'a*** opartym na węzłach x_0, x_1, ..., x_n. Wielomian ten jest wielomianem stopnia co najwyżej n i jest jednoznacznym rozwiązaniem zadania interpolacyjnego dla dowolnego wyboru n+1 różnych węzłów interpolacji.

Przykład 3.1.

Znaleźć wielomian interpolacyjny Lagrange'a, który w punktach:
-2, -1, 0, 2 przyjmuje odpowiednio wartości 0, 1, 1, 2.
Obliczyć także przybliżoną wartość funkcji danej powyższymi wartościami w punkcie $x=1$.

Stosując wzór Lagrange'a (3.7) dla $n = 3$ otrzymujemy:

$$W_3(x) = 0\frac{(x+1)(x-0)(x-2)}{(-2+1)(-2-0)(-2-2)} + 1\frac{(x+2)(x-0)(x-2)}{(-1+2)(-1-0)(-1-2)} +$$
$$+1\frac{(x+2)(x+1)(x-2)}{(0+2)(0+1)(0-2)} + 2\frac{(x+2)(x+1)(x-0)}{(2+2)(2+1)(2-0)} =$$
$$= \frac{1}{3}(x^3 - 4x) - \frac{1}{4}(x^3 + x^2 - 4x - 4) + \frac{1}{12}(x^3 + 3x^2 + 2x) =$$
$$= \frac{1}{6}x^3 - \frac{1}{6}x + 1.$$

Używając tego wielomianu można teraz interpolować wartości funkcji *f(x)* w punktach przedziału [-2,2]. Przybliżona wartość funkcji *f* w punkcie 1, to wartość wielomianu $W_3(1)$:

$$f(1) \cong W_3(1) = \frac{1}{6}\cdot 1^3 - \frac{1}{6}\cdot 1 + 1 = 1.$$

Wielomian interpolacyjny stosuje się również do obliczania jego wartości w punktach nie należących do przedziału obejmującego punkty $x_0, x_1, ..., x_n$ i wtedy mamy do czynienia z ekstrapolacją. Niezależnie od tego czy punkt należy do tego przedziału czy znajduje się poza nim należy, do oceny błędu, posłużyć się tzw. twierdzeniem o reszcie aproksymacji wielomianem interpolacyjnym.

Niech *W(x)* będzie wielomianem interpolacyjny Lagrange'a oraz $\mathrm{I} = \mathrm{I}[x_0, x_1, ..., x_n]$ oznacza przedział zawierający węzły interpolacji $x_0, x_1, ..., x_n$.

Twierdzenie o reszcie interpolacyjnej

Dla każdej funkcji $f \in C^{n+1}(\mathrm{I})$ i każdego $x \in \mathrm{I}$, istnieje taki punkt $\xi \in \mathrm{I}$, że:

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

gdzie :

$$p(x) = \prod_{k=0}^{n}(x - x_k).$$

a funkcja ***R(x)=f(x)-W(x)*** jest tzw. ***resztą interpolacji***.

Dla dowolnego ***M***>0 rozważmy teraz klasę funkcji:

$$C_M^{n+1}[a,b]=\left\{ f \in C^{n+1}[a,b]: \bigwedge_{x\in[a,b]} \left| f^{(n+1)}(x)\right| \le M \right\}.$$

Jeśli węzły interpolacji ***x0***, ***x1***, ..., ***xn*** należą do przedziału [***a,b***] i $f \in C^{n+1}[a,b]$ wtedy:

$$\left|R(x)\right| \le \frac{M}{(n+1)!}\left\|p(x)\right\|,$$

gdzie $\left\|p(x)\right\| = \max_{a\le x\le b}\left|p(x)\right|$ jest normą jednostajną wielomianu ***p(x).***

3.2.3. Wzór interpolacyjny Newtona

Wzór interpolacyjny Lagrange'a jest niewygodny do stosowania w przypadku, gdy zmienia się liczba węzłów. Wtedy do wyznaczenia wielomianu określonego stopnia trzeba powtarzać obliczenia od początku. Zatem poprzez dodanie nowych węzłów interpolacyjnych nie można modyfikować wcześniej wyznaczonego wielomianu Lagrange'a. Wzór interpolacyjny Newtona równoważny wielomianowi Lagrange'a usuwa tę niedogodność.

Niech $x_0, x_1, ..., x_n$, będą węzłami interpolacji, w których wielomian interpolacyjny przyjmuje odpowiednio wartości $y_0, y_1, ..., y_n$. Można wówczas zdefiniować wyrażenia zwane ***ilorazami różnicowymi***:

- pierwszego rzędu:

$$[x_0, x_1] = \frac{y_1 - y_0}{x_1 - x_0},$$

$$[x_1, x_2] = \frac{y_2 - y_1}{x_2 - x_1},$$

……

$$[x_{n-1}, x_n] = \frac{y_n - y_{n-1}}{x_n - x_{n-1}},$$

- drugiego rzędu (analogicznie):

$$[x_0, x_1, x_2] = \frac{[x_1, x_2] - [x_0, x_1]}{x_2 - x_0},$$

$$[x_1, x_2, x_3] = \frac{[x_2, x_3] - [x_1, x_2]}{x_3 - x_1},$$

.....

$$[x_{n-2}, x_{n-1}, x_n] = \frac{[x_{n-1}, x_n] - [x_{n-2}, x_{n-1}]}{x_n - x_{n-2}},$$

- ***k***-tego rzędu:

$$[x_i, x_{i+1}, ..., x_{i+k}] = \frac{[x_{i+1}, x_{i+2}, ..., x_{i+k}] - [x_i, x_{i+1}, ..., x_{i+k-1}]}{x_{i+k} - x_i} \quad \text{dla } \boldsymbol{k} = 1, 2, ...$$

oraz $\boldsymbol{i} = 0, 1, 2, ...$.

Z ilorazów różnicowych tworzy się zazwyczaj tablicę (tabela 3.1).

Tabela 3.1. Tablica ilorazów różnicowych

x_i	y_i	rzędu 1	rzędu 2	rzędu 3	rzędu 4	rzędu 5
x_0	y_0					
		$[x_0,x_1]$				
x_1	y_1		$[x_0,x_1,x_2]$			
		$[x_1,x_2]$		$[x_0,x_1,x_2,x_3]$		
x_2	y_2		$[x_1,x_2,x_3]$		$[x_0,x_1,x_2,x_3,x_4]$	
		$[x_2,x_3]$		$[x_1,x_2,x_3,x_4]$		. . .
x_3	y_3		$[x_2,x_3,x_4]$		$[x_1,x_2,x_3,x_4,x_5]$	
		$[x_3,x_4]$		$[x_2,x_3,x_4,x_5]$		
x_4	y_4		$[x_3,x_4,x_5]$			
		$[x_4,x_5]$				
x_5	y_5					

Twierdzenie 3.3.

Jeśli $y_0, y_1, ..., y_n$ są wartościami wielomianu stopnia n, to iloraz różnicowy rzędu n+1 jest tożsamościowo równy zeru tzn. :

$$[x, x_0, x_1, ..., x_n] \equiv 0. \tag{3.9}$$

Korzystając z twierdzenia 3.3 oraz z definicji ilorazu różnicowego, można napisać:

$$[x, x_0, x_1, ..., x_n] = \frac{[x_o, x_1, ..., x_n] - [x, x_0, ..., x_{n-1}]}{x_n - x} \equiv 0 .$$

Wynika stąd, że:

$$[x, x_0, x_1, ..., x_{n-1}] = [x_0, x_1, ..., x_n]$$

a więc:

$$\frac{[x_0, x_1, ..., x_{n-1}] - [x, x_0, ..., x_{n-2}]}{x_{n-1} - x} = [x_0, x_1, ..., x_n],$$

czyli:

$$[x, x_0, x_1, ..., x_{n-2}] = [x_0, x_1, ..., x_n] \cdot (x - x_{n-1}) + [x_0, x_1, ..., x_{n-1}].$$

Korzystając dalej analogicznie z definicji ilorazów różnicowych, otrzymujemy ***wzór interpolacyjny Newtona*** z ilorazami różnicowymi:

$$\begin{aligned} y = W_n(x) = y_0 &+ [x_0, x_1](x - x_0) \\ &+ [x_0, x_1, x_2](x - x_0)(x - x_1) + ... \\ &+ [x_0, x_1, ..., x_n](x - x_0)(x - x_1) \cdot ... \cdot (x - x_{n-1}). \end{aligned} \tag{3.10}$$

Należy zwrócić uwagę, iż do wyznaczenia wzoru na wielomian interpolacyjny Newtona wykorzystuje się tylko ilorazy różnicowe zaczynające się od węzła x_0, czyli te, które w tabeli 3.1 znajdują się jako pierwsze w każdej kolumnie. Jednak obliczenie całej tabeli ilorazów jest konieczne i nieuniknione.

Wzór interpolacyjny Newtona ma tę własność, że rozszerzenie zadania interpolacji o dodatkowy węzeł sprowadza się do obliczenia i dołączenia do poprzednio wyznaczonego wielomianu dodatkowego składnika.

Przykład 3.2.

Znaleźć wielomian interpolacyjny Newtona, który w punktach: -2, -1, 0, 2, 4 przyjmuje odpowiednio wartości -1, 0, 5, 99, -55.

Najpierw dla $n = 4$ musimy obliczyć tablicę ilorazów różnicowych taką, jak pokazana została w tabeli 3.2.

Tabela 3.2. Przykładowa tablica ilorazów różnicowych dla przykładu 3.2

x_i	y_i	rzędu 1	rzędu 2	rzędu 3	rzędu 4
-2	-1				
		1			
-1	0		2		
		5		3	
0	5		14		-2
		47		-9	
2	99		-31		
		-77			
4	-55				

Obliczenie ilorazów różnicowych rzędu 1-go ma postać:

$$[x_0, x_1] = \frac{y_1 - y_0}{x_1 - x_0} = \frac{0-(-1)}{-1-(-2)} = 1,$$
$$[x_1, x_2] = \frac{y_2 - y_1}{x_2 - x_1} = \frac{5-0}{0-(-1)} = 5,$$
$$[x_2, x_3] = \frac{y_3 - y_2}{x_3 - x_2} = \frac{99-(5)}{2-0} = 47,$$
$$[x_3, x_4] = \frac{y_4 - y_3}{x_4 - x_3} = \frac{-55-99}{4-2} = -77.$$

Na podstawie ilorazów rzędu 1-go i wartości węzłów możemy teraz obliczyć ilorazy różnicowe rzędu 2-go:

$$[x_0, x_1, x_2] = \frac{[x_1,x_2]-[x_0,x_1]}{x_2 - x_0} = \frac{5-1}{0-(-2)} = 2,$$
$$[x_1, x_2, x_3] = \frac{[x_2,x_3]-[x_1,x_2]}{x_3 - x_1} = \frac{47-5}{2-(-1)} = 14,$$
$$[x_2, x_3, x_4] = \frac{[x_3,x_4]-[x_2,x_3]}{x_4 - x_2} = \frac{-77-47}{4-0} = -31.$$

W dalszej kolejności musimy obliczyć ilorazy różnicowe rzędu 3-go:

$$[x_0, x_1, x_2, x_3] = \frac{[x_1,x_2,x_3]-[x_0,x_1,x_2]}{x_3 - x_0} = \frac{14-2}{2-(-2)} = 3,$$
$$[x_1, x_2, x_3, x_4] = \frac{[x_2,x_3,x_4]-[x_1,x_2,x_3]}{x_4 - x_1} = \frac{-31-14}{4-(-1)} = -9.$$

a następnie rzędu 4-go:

$$[x_0, x_1, x_2, x_3, x_4] = \frac{[x_1,x_2,x_3,x_4]-[x_0,x_1,x_2,x_3]}{x_4-x_0} = \frac{-9-3}{4-(-2)} = -2.$$

Dla ***n=4*** wielomian interpolacyjny Newtona ma postać ze wzoru (3.10):

$$W_4(x) = y_0 + [x_0, x_1] \cdot (x - x_0) + [x_0, x_1, x_2] \cdot (x - x_0)(x - x_1) +$$
$$+[x_0, x_1, x_2, x_3] \cdot (x - x_0)(x - x_1)(x - x_2) +$$
$$+[x_0, x_1, x_2, x_3, x_4](x - x_0)(x - x_1)(x - x_2)(x - x_3)$$

A zatem otrzymujemy wielomian o współczynnikach:

$$W_4(x) = -1 + 1(x + 2) + 2(x + 2)(x + 1) + 3(x + 2)(x + 1) +$$
$$-2(x + 2)(x + 1)x(x - 2) = -1 + x + 2 + 2(x^2 + 3x + 2) +$$
$$+3(x^3 + 3x^2 + 2x) - 2(x^4 + x^3 - 4x^2 - 4x) = -2x^4 + x^3 +$$
$$+19x^2 + 21x + 5$$

Przykład 3.3.

Wyznacz przybliżoną wartość funkcji *f(x)* stosując interpolację Newtona w punkcie *x=3*. Węzły interpolacji to: -4, -3, -1, 0, 2, 4 a wartości funkcji w tych punktach wynoszą odpowiednio: 2, -1, -37, -58, 464, -3382.

Ilorazy różnicowe mają wartości jak pokazano w tabeli 3.3.

Wielomian interpolacyjny Newtona ma postać:

$$\begin{aligned} W_5(x) = 2 &- 3(x + 4) - 5(x + 4)(x + 3) \\ &+ 1(x + 4)(x + 3)(x + 1) \\ &+ 3(x + 4)(x + 3)(x + 1)x \\ &- 3(x + 4)(x + 3)(x + 1)x(x - 2) = \\ &= -3x^5 - 15x^4 + 16x^3 + 138x^2 + 89x - 58 \end{aligned}$$

Przybliżoną wartość funkcji w punkcie $x=3$ możemy obliczyć stosując otrzymany wielomian, gdyż $3 \in [-4,4]$ i wynosi ona:

$$f(3) \cong W_5(3) = -3 \cdot 3^5 - 15 \cdot 3^4 + 16 \cdot 3^3 + 138 \cdot 3^2 + 89 \cdot 3 - 58 = \quad = -61 .$$

Tabela 3.3. Tablica ilorazów różnicowych dla przykładu 3.3

x_i	y_i	rzędu 1	rzędu 2	rzędu 3	rzędu 4	rzędu 5
-4	2					
		-3				
-3	-1		-5			
		-18		1		
-1	-37		-1		3	
		-21		19		-3
0	-58		94		-21	
		261		-128		
2	464		-546			
		-1923				
4	-3382					

3.3. Interpolacja trygonometryczna

Interpolację trygonometryczną stosuje się do wyznaczania funkcji okresowych, często sinusoidalnych, np. funkcji opisujących sygnały elektryczne lub drgania w mechanice.

Zakładamy, że dana jest funkcja f zmiennej rzeczywistej o wartościach zespolonych, okresowa o okresie 2π, czyli dla każdego x:

$$f(x+2\pi) = f(x) .$$

Jeśli funkcja g byłaby funkcją okresową o okresie t, tzn. dla każdego y:

$$g(y+t) = g(y) ,$$

to dokonując podstawienia $x = \frac{2\pi}{t} y$ otrzymamy funkcję okresową:

$$f(x) = g\left(\frac{xt}{2\pi} \right)$$ o okresie 2π.

Można więc, bez zmniejszenia ogólności, rozpatrywać tylko funkcje o okresie 2π.

Zadanie interpolacji trygonometrycznej polega na znalezieniu, dla danej funkcji f, wielomianu trygonometrycznego postaci:

$$F_n(x) = \sum_{m=0}^{n} c_m e^{jmx} = \sum_{m=0}^{n} c_m (\cos(mx) + j\sin(mx)), \tag{3.11}$$

gdzie $j = \sqrt{-1}$. Wielomian ten w n+1 różnych punktach x_k, k=0, 1, ..., n, $x_k \neq x_l$ dla $k \neq l$, z przedziału $(0,2\pi)$ przyjmuje te same wartości, co funkcja f. Tzn. dla k=0,1,...,n:

$$F_n(x_k) = f(x_k). \tag{3.12}$$

Twierdzenie 3.4.
Zadanie interpolacji trygonometrycznej ma dokładnie jedno rozwiązanie.

Warunki interpolacji można zapisać w postaci układu n+1 równań liniowych z n +1 niewiadomymi c_0, c_1, ..., c_n :

$$\sum_{m=0}^{n} c_m z_k^m = f(x_k), \text{ dla } k = 0, 1, ..., n,$$

gdzie $z_k = e^{jx_k}$.

Macierz tego układu jest macierzą Vandermonde'a i jest nieosobliwa, gdyż jej wyznacznik (3.4) nie zeruje się na mocy założenia, że węzły x_k są różne. Zatem zadanie interpolacyjne ma jednoznaczne rozwiązanie.

Potrzeba wyznaczania współczynników wielomianu interpolacyjnego funkcji f opartego na dowolnych węzłach pojawia się w praktyce bardzo rzadko. Z tego powodu ograniczymy się tylko do przypadku węzłów równoodległych:

$$x_k = \frac{2k\pi}{n+1}, \; k = 0, 1, ..., n. \tag{3.13}$$

Przy tym założeniu rozwiązanie zadania interpolacyjnego upraszcza się w istotny sposób.

Funkcje e^{jmx} (m=0, 1, ..., n) tworzą układ ortogonalny w sensie iloczynu skalarnego:

$$(f,g)=\sum_{k=0}^{n} f(x_k)g(x_k), \tag{3.14}$$

ponieważ:

$$(e^{jlx},e^{jmx})=\sum_{k=0}^{n} e^{jlx_k}e^{-jmx_k}=\sum_{k=0}^{n} e^{j(l-m)\frac{2k\pi}{n+1}}=$$
$$=\begin{cases}\dfrac{e^{j(l-m)2\pi}-1}{e^{j(l-m)\frac{2\pi}{n+1}}-1}=0 & \text{dla} \quad l\neq m,\\ n+1 & \text{dla} \quad l=m.\end{cases} \tag{3.15}$$

Z tej własności wynika kolejne twierdzenie.

Twierdzenie 3.5.

Współczynniki wielomianu trygonometrycznego (3.11) interpolującego funkcję f, opartego na węzłach (3.13) są równe:

$$c_m=\frac{(f,e^{jmx})}{n+1}=\frac{1}{n+1}\sum_{k=0}^{n} f(x_k)e^{-jmx_k}. \tag{3.16}$$

Z założenia (3.12) wynikają równości:

$$(F_n,e^{jmx})=\sum_{k=0}^{n} F_n(x_k)e^{-jmx_k}=\sum_{k=0}^{n} f(x_k)e^{-jmx_k}=(f,e^{jmx}),$$

a z własności iloczynu skalarnego oraz ze wzoru (3.15) otrzymujemy:

$$(F_n,e^{jmx})=(\sum_{l=0}^{n} c_l e^{jlx},e^{jmx})=\sum_{l=0}^{n} c_l(e^{jlx},e^{jmx})=(n+1)c_m.$$

Współczynniki c_m określone wzorem (3.16) są równe współczynnikom rozwinięcia Fouriera funkcji f względem iloczynu skalarnego (3.14). Stąd zadanie wyznaczania współczynników wielomianu trygonometrycznego interpolującego funkcję f nazywane jest ***dyskretną analizą Fouriera*** [1,2].

Wielomian trygonometryczny stosuje się często w postaci (3.17), szczególnie przydatnej w przypadku interpolacji funkcji o wartościach rzeczywistych.

Twierdzenie 3.6.

Trygonometryczny wielomian interpolacyjny funkcji *f*, oparty na węzłach (3.13) może być przedstawiony w następującej postaci:

$$F_n(x) = \frac{1}{2}a_0 + \sum_{m=1}^{p}(a_m \cos mx + b_m \sin mx) + \frac{1}{2}\delta a_{p+1}\cos(p+1)x \qquad (3.17)$$

przy czym:

dla *n* parzystego δ=0, p=0.5n;

dla *n* nieparzystego δ=1, p=0.5(n-1),

współczynniki a_m oraz b_m mają postać:

$$a_m = \frac{2}{n+1}\sum_{k=0}^{n} f(x_k)\cos mx_k, \quad m = 0,1,2,...,p,$$
$$b_m = \frac{2}{n+1}\sum_{k=0}^{n} f(x_k)\sin mx_k, \quad m = 1,2,...,p. \qquad (3.18)$$

Wielomian (3.17) jest trygonometrycznym odpowiednikiem wzoru Lagrange'a.

Można sprawdzić, że jeśli $f(x)$ jest ***funkcją parzystą***, tzn. $f(x)=f(-x)$, to b_m=0 dla każdego m.

Jeśli natomiast $f(x)$ jest ***funkcją nieparzystą*** tzn. $f(x) = -f(-x)$, to $a_m = 0$ dla każdego m.

Istnieją również bardziej efektywne algorytmy obliczania współczynników wielomianu $F_n(x)$, np. algorytm Goertzela czy algorytm Reinscha [5].

Przykład 3.4.

Znaleźć trygonometryczny wielomian interpolacyjny $F_4(x)$ przechodzący przez węzły interpolacji (3.13) i dla $n = 4$ przybliżający funkcję daną w postaci dyskretnych wartości $f_0 = f_1 = f_2 = 0$, $f_3 = f_4 = 1$.

Zgodnie ze wzorem (3.17) dla p=0.5, $n = 2$ mamy:

$$F_4(x) = \frac{1}{2}a_0 + \sum_{m=1}^{2}(a_m \cos mx + b_m \sin mx) =$$
$$= 0.5a_0 + a_1\cos x + b_1 \sin x + a_2 \cos 2x + b_2 \sin 2x.$$

Wyznaczamy węzły interpolacji:

$x_0 = 0,\ x_1 = 2/5\pi,\ \ x_2 = 4/5\pi,\ \ \ x_3 = 6/5\pi,\ \ x_4 = 8/5\pi,$

zaś ze wzorów (3.18) otrzymujemy:

$$a_m = \frac{2}{5}\sum_{k=0}^{4} f(x_k)\cos mx_k, m = 0,1,2,$$

$$b_m = \frac{2}{5}\sum_{k=0}^{4} f(x_k)\sin mx_k, m = 1,2.$$

Zatem:

$$a_0 = \frac{2}{5}\sum_{k=0}^{4} f(x_k) = \frac{2}{5}(1+1) = \frac{4}{5} = 0.8,$$

$$a_1 = \frac{2}{5}\sum_{k=0}^{4} f(x_k)\cos x_k = \frac{2}{5}(-0.809 + 0.309) = -0.200,$$

$$a_2 = \frac{2}{5}\sum_{k=0}^{4} f(x_k)\cos 2x_k = \frac{2}{5}(0.309 - 0.809) = -0.200,$$

$$b_1 = \frac{2}{5}\sum_{k=0}^{4} f(x_k)\sin x_k = \frac{2}{5}(-0.558 - 0.951) = -0.616,$$

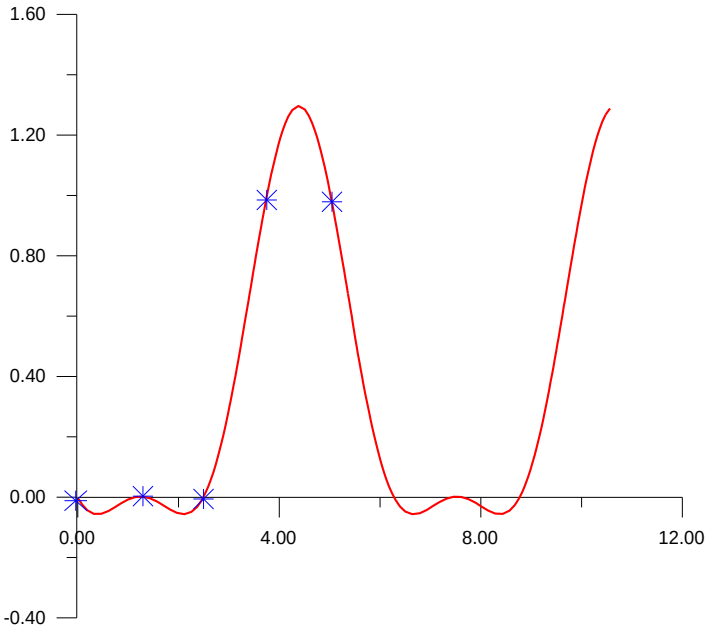
$$b_2 = \frac{2}{5}\sum_{k=0}^{4} f(x_k)\sin 2x_k = \frac{2}{5}(0.951 - 0.588) = 0.145.$$

Ostatecznie:

$F_4(x) = 0.8 - 0.2\cos x - 0.616\sin x - 0.2\cos 2x + 0.145\sin 2x.$

Można sprawdzić, że:
$F_4(x_0)=0,\ F_4(x_1)=1.00062,\ F_4(x_2)=0.99997,\ F_4(x_3)=0.00002,$
$F_4(x_4)=0.00621.$

Ilustracja graficzna przykładu 3.4 (rys. 3.2) pokazuje, że funkcja interpolacyjna jest funkcją okresową o okresie 2π. Wartości funkcji $F_4(x)$ pokrywają się w węzłach interpolacji z wartościami danymi.



Rys. 3.2. Wykres funkcji interpolacyjnej z przykładu 3.4

3.4. Funkcje sklejane

3.4.1. Określenie funkcji sklejanych

Niech w przedziale $<a, b>$ danych będzie $(n+1)$ punktów $x_0, x_1, \ldots, x_n$, przy czym $a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$. Punkty x_i, $i = 1, ..., n$ określają pewien podział przedziału $<a, b>$ na n podprzedziałów. Podział ten oznaczymy symbolem Δ_n.

Funkcję $S(x)=S(x, \Delta_n)$ określoną na przedziale $<a, b>$ nazywamy ***funkcją sklejaną*** stopnia m ($m \geq 1$), jeżeli:

- $S(x)$ jest wielomianem stopnia co najwyżej m na każdym podprzedziale (x_i, x_{i+1}), $i = 0, 1, ..., n-1$,
- $S(x)$ i jej pochodne stopnia $1, 2, ..., m-1$ są ciągłe w rozpatrywanych przedziałach.

Zbiór wszystkich funkcji sklejanych stopnia m o węzłach w punktach x_i oznaczymy $S_m(\Delta_n)$. Jeśli $S(x) \in S_m(\Delta_n)$, to na każdym przedziale

(x_i, x_{i+1}), $i = 0, 1, ..., n$ funkcja $S(x)$ jest wielomianem stopnia co najwyżej m:

$$S(x) = c_{i,m}x^m + c_{im-1}x^{m-1} + ... + c_{i,1}x + c_{i0} \text{ dla } x \in (x_i, x_{i+1}). \quad (3.19)$$

Mamy więc $n(m+1)$ dowolnych stałych c_{ij}. Żądanie ciągłości pochodnych rzędu 0, 1, ..., m-1 w każdym węźle wewnętrznym x_i daje $m(n-1)$ warunków. Tak więc funkcja $S(x)$ zależy od:

$n(m+1) - m(n-1) = n+m$ parametrów.

Dowolne funkcje bardzo często przybliża się funkcjami sklejanymi. Wiąże się to z łatwością wyznaczania ich wartości oraz ze zbieżnością dla licznych klas funkcji. W praktyce często stosuje się funkcje sklejane stopnia trzeciego, które dla wielu zagadnień są wystarczająco gładkie, a szybkość ich zbieżności do funkcji aproksymowanej jest zadowalająca.

3.4.2. Interpolacyjne funkcje sklejane stopnia trzeciego

Funkcję $S(x) \in S_3(\Delta_n)$ nazywamy interpolacyjną funkcją sklejaną stopnia trzeciego dla funkcji $f(x)$, jeżeli $S(x_i) = f(x_i) = y_i$, $i = 0, 1,..., n$, $n>1$.
Funkcja $S(x)$ stopnia trzeciego zależy od $n+3$ parametrów. Ponieważ dane są wartości funkcji $f(x_i)=y_i$ w $n+1$ punktach, to na interpolacyjne funkcje sklejane stopnia trzeciego należy nałożyć dwa dodatkowe warunki:

$$S'(a+0) = \alpha_1, \quad S'(b-0) = \beta_1 \quad (3.20)$$

lub:

$$S''(a+0) = \alpha_2, \quad S''(b-0) = \beta_2, \quad (3.21)$$

gdzie $\alpha_1, \beta_1, \alpha_2, \beta_2$ są ustalonymi liczbami rzeczywistymi.

Jeżeli funkcja $f(x)$ ma pochodne w punktach a, b i są one znane, to można je przyjąć jako liczby występujące po prawych stronach warunków (3.20) i (3.21).

Wyznaczymy interpolacyjną funkcję sklejaną dla węzłów dowolnie rozłożonych.

Oznaczmy:

$M_j = S''(x_j)$, $j = 0, 1, ..., n$.

Zgodnie z określeniem funkcji sklejanej trzeciego stopnia, pochodna $S''(x)$ jest funkcją ciągłą na przedziale $\langle a, b \rangle$ i liniową na podprzedziale (x_{j-1}, x_j).

Można więc przedstawić ją w postaci:

$$S''(x) = M_{j-1}\frac{x_j - x}{h_j} - M_j\frac{x_{j-1} - x}{h_j}, \tag{3.22}$$

przy czym $x \in < x_{j-1}, x_j >$, $h_j = x_j - x_{j-1}$.

Całkując stronami (3.22) otrzymujemy:

$$S'(x) = -M_{j-1}\frac{(x_j - x)^2}{2h_j} + M_j\frac{(x - x_{j-1})^2}{2h_j} + A_j,$$

$$S(x) = M_{j-1}\frac{(x_j - x)^3}{6h_j} + M_j\frac{(x - x_{j-1})^3}{6h_j} + A_j(x - x_{j-1}) + B_j, \tag{3.23}$$

gdzie A_j, B_j są stałymi. Nakładając na $S(x)$ warunki interpolacji:

$$S(x_{j-1}) = M_{j-1}\frac{h_j^2}{6} + B_j = y_{j-1},$$

$$S(x_j) = M_j\frac{h_j^2}{6} + A_j h_j + B_j = y_j,$$

wyznaczamy stałe A_j i B_j:

$$B_j = y_{j-1} - M_{j-1}\frac{h_j^2}{6},$$

$$A_j = \frac{y_j - y_{j-1}}{h_j} - \frac{h_j}{6}(M_j - M_{j-1}). \tag{3.24}$$

Żądamy aby pochodna $S'(x)$ była funkcją ciągłą na $<a, b>$. W tym celu obliczamy granice jednostronne:

$$S'(x_j - 0) = \frac{h_j}{6}M_{j-1} + \frac{h_j}{3}M_j + \frac{y_j - y_{j-1}}{h_j},$$

$$S'(x_j + 0) = -\frac{h_{j+1}}{3}M_j - \frac{h_{j+1}}{6}M_{j+1} + \frac{y_{j+1} - y_j}{h_{j+1}} \tag{3.25}$$

i nakładamy warunek:

$$S'(x_j - 0) = S'(x_j + 0), \quad j = 1, 2, ..., n\text{-}1. \tag{3.26}$$

Po podstawieniu (3.25) do (3.26) otrzymujemy układ (n-1) równań:

$$\frac{h_j}{6} M_{j-1} + \frac{h_j + h_{j+Á1}}{3} M_j + \frac{h_{j+1}}{6} M_{j+1} = \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j}, \qquad (3.27)$$

przy czym j = 1, 2, ..., n-1.

Równania (3.27) można zapisać w postaci:

$$\mu_j M_{j-1} + 2M_j + \lambda_j M_{j+1} = d_j, \quad j = 1, 2, ..., n\text{-}1, \qquad (3.28)$$

gdzie:

$$\lambda_j = \frac{h_{j+1}}{h_j + h_{j+1}},$$

$$\mu_j = 1 - \lambda_j,$$

$$d_j = \frac{6}{h_j + h_{j+1}} \left(\frac{y_{j+1} - y_j}{h_j} - \frac{y_j - y_{j-1}}{h_{j-1}} \right) = 6[x_{j-1}, x_j, x_{j+1}].$$

Jeżeli M_j, j = 0, 1, ..., n spełniają powyższy układ, to funkcja $S(x)$ określona wzorami (3.23) i (3.24) jest na każdym podprzedziale $x \in < x_{j-1}, x_j >$, j = 1, 2, ..., n interpolacyjną funkcją sklejaną stopnia trzeciego. Do układu (3.28) należy dołączyć jeszcze dwa równania, wynikające ze spełnienia przez funkcję $S(x)$ jednego z dodatkowych warunków (3.20) lub (3.21). Równania te dla warunków (3.20) mają postać:

$$2M_0 + M_1 = d_0, \quad M_{n-1} + 2M_n = d_n,$$

gdzie:

$$d_0 = \frac{6}{h_1} \left(\frac{y_1 - y_0}{h_1} - \alpha_1 \right), \quad d_n = \frac{6}{h_n} (\beta_1 - \frac{y_n - y_{n-1}}{h_n}).$$

zaś dla warunków (3.21) są postaci:

$$M_0 = \alpha_{2,}, M_n = \beta_2.$$

Układ równań zapisany w postaci macierzowej ma postać:

$$\begin{bmatrix} 2 & 1 & 0 & ... & 0 \\ \mu_1 & 2 & \lambda_1 & ... & 0 \\ 0 & \mu_2 & 2 & ... & 0 \\ ... & & & & ... \\ 0 & ... & ... & 2 & \lambda_{n-1} \\ 0 & ... & ... & 1 & 2 \end{bmatrix} \begin{bmatrix} M_0 \\ M_1 \\ \\ \\ M_{n-1} \\ M_n \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ \\ \\ d_{n-1} \\ d_n \end{bmatrix}. \tag{3.29}$$

Macierz współczynników układu jest macierzą diagonalną o elementach na głównej przekątnej znacznie przewyższających co do modułu sumę modułów pozostałych elementów wiersza. Stąd wynika, że układ (3.29) ma jednoznaczne rozwiązanie [4]. Istnieje zatem jedna funkcja interpolująca stopnia trzeciego, spełniająca jeden z warunków (3.20) lub (3.21).

Do wyznaczania interpolacyjnych funkcji sklejanych o przedstawionej wyżej postaci, są niezbędne następujące dane: węzły x_i, wartości drugiej pochodnej funkcji sklejanej $M_j = S''(x)$ oraz wartości funkcji $y_i = f(x_i)$.

Często wygodnie jest przedstawić poszukiwaną funkcję $S(x)$ w postaci kombinacji liniowej elementów bazy przestrzeni $S_m(\Delta_m)$.

Wyznaczymy bazę przestrzeni funkcji $S(x)$ stopnia trzeciego z węzłami równoodległymi $x_i = x_0 + ih$, $h = \dfrac{b-a}{n}$, $i = 0,1,..,n$. Dodatkowo przez $x_{-3}, x_{-2}, x_{-1}, x_{n+1}, x_{n+2}, x_{n+3}$ oznaczmy punkty $x_i = x_0 + ih$ dla i = -3, -2, -1, n+1, n+2, n+3 i określmy funkcje $\Phi_i^3(x)$, i = -1, 0, 1, ..., n, n+1 za pomocą wzoru:

$$\Phi_j^3(x) =$$

$$= \frac{1}{h^3} \begin{cases} (x - x_{j-2})^3, & x \in < x_{j-2;} x_{j-1} >, \\ h^3 + 3h^2(x - x_{j-1}) + 3h(x - x_{j-1})^2 - 3(x - x_{j-1})^3, & x \in < x_{j-1;}; x_j >, \\ h^3 + 3h^2(x_{j+1} - x) + 3h(x_{j+1} - x)^2 - 3(x_{j+1} - x)^3, & x \in < x_j; x_{j+1} >, \\ (x_{j+2} - x)^3, & x \in < x_{j+1} : x_{j+2} >, \\ 0 \qquad \text{dla pozostałych} & x \in R. \end{cases} \tag{3.30}$$

Funkcje te są klasy C^2. W tablicy 3.4 podane są wartości funkcji $\Phi_j^3(x)$ oraz jej pierwszej i drugiej pochodnej w punktach x_k dla $k = j\text{-}2, j\text{-}1, j, j+1, j+2$. Poza przedziałem $<x_{j-2}, x_{j+2}>$ funkcja ta jest tożsamościowo równa zeru.

Tabela 3.4. Wartości funkcji $\Phi_j^3(x)$ i jej pochodnych

	x_{j-2}	x_{j-1}	x_j	x_{j+1}	x_{j+2}
$\Phi_j^3(x)$	0	1	4	1	0
$(\Phi_j^3(x))'$	0	$3/h$	0	$-3/h$	0
$(\Phi_j^3(x))''$	0	$3/h^2$	$-12/h^2$	$6/h^2$	0

Twierdzenie 3.7.

Funkcje $\Phi_i^3(x)$, $i = -1, 0, 1, ..., n+1$, określone na przedziale $<a, b>$ stanowią bazę przestrzeni funkcji sklejanych trzeciego stopnia. Każdą funkcję $S(x)$ można przedstawić w postaci kombinacji liniowej:

$$S(x) = \sum_{i=-1}^{n+1} c_i \Phi_i^3(x), a \le x \le b, \tag{3.31}$$

gdzie $\Phi_i^3(x)$ są określone wzorem (3.30), c_i są liczbami rzeczywistymi.

W przypadku węzłów równoodległych szukamy interpolacyjnej funkcji sklejanej w postaci kombinacji liniowej (3.31). Na podstawie tablicy 3.2 można stwierdzić, że stałe c_i muszą spełniać układ $(n+1)$ równań:

$$c_{i-1} + 4c_i + c_{i+1} = y_i, \quad i = 0, 1, ..., n. \tag{3.32}$$

Jeśli funkcja spełnia dodatkowe warunki (3.20), to dodatkowe dwa równania będą następujące:

$$-c_{-1} + c_1 = \frac{h}{3}\alpha_1 - c_{n-1} + c_{n+1} = \frac{h}{3}\beta_1.$$

Po wyeliminowaniu z układu współczynników c_{-1} oraz c_{n+1}, pozostałe współczynniki c_j , j = 0, 1, ..., n rozwinięcia będą rozwiązaniem układu:

$$\begin{bmatrix} 4 & 2 & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 & 0 \\ 0 & 1 & 4 & ... & 0 & 0 \\ 0 & 0 & ... & ... & ... & 0 \\ 0 & 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 2 & 4 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ ... \\ ... \\ c_{n-1} \\ c_n \end{bmatrix} = \begin{bmatrix} y_0 + h\alpha_1 / 3 \\ y_1 \\ ... \\ ... \\ y_{n-1} \\ y_n + h\beta_1 / 3 \end{bmatrix}, \tag{3.33}$$

którego macierz współczynników jest macierzą trójdiagonalną o dominujących elementach na głównej przekątnej. Układ ma więc jednoznaczne rozwiązanie.

Przykład 3.5.

Mając dane węzły interpolacyjne jak w tabeli 3.5, znaleźć sześcienną funkcję sklejaną.

Tabela 3.5. Dane do przykładu 3.5

i	0	1	2	3
x_i	1	3	5	8
y_i	2	4	7	9

Korzystając z programu opracowanego na podstawie rozważań z bieżącego rozdziału, otrzymuje się następującą funkcję sklejaną:

a) w przedziale $< x_0 , x_1 > = < 1, 3 >$:

$S(x) = 2.00+0.84(x\text{-}1.00)+0.00(x\text{-}1.00)^2+0.04(x\text{-}1.00)^3$

b) w przedziale $< x_1 , x_2 > = < 3, 5 >$:

$S(x) = 4.00+1.33(x\text{-}3.00)+0.25(x\text{-}3.00)^2-0.08(x\text{-}3.00)^3$

c) w przedziale $< x_2 , x_3 > = < 5, 8 >$:

$S(x) = 7.00+1.35(x\text{-}5.00)-0.24(x\text{-}5.00)^2+0.00(x\text{-}5.00)^3$

Przykład 3.6.

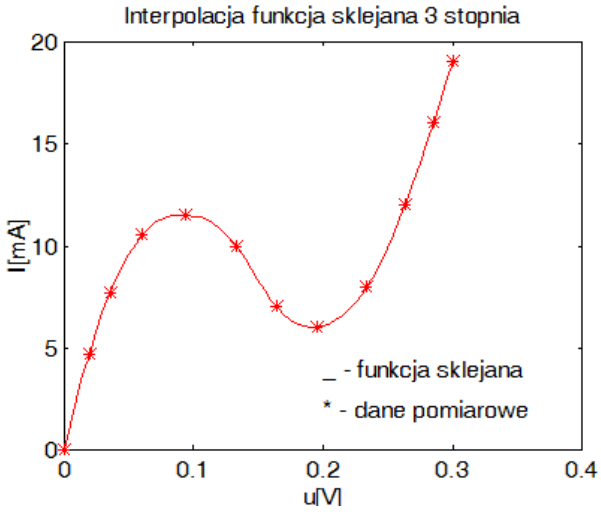
Zastosować interpolację funkcją sklejaną trzeciego stopnia do aproksymacji charakterystyki prądowo-napięciowej diody tunelowej,

której charakterystyka przedstawiona jest w tabeli 3.6. Rolę zmiennej x pełni wartość napięcia U a rolę y – wartość natężenia prądu I. Graficzne wyniki interpolacji wielomianową funkcją sklejaną prezentuje rys. 3.3.

Tabela 3.4. Charakterystyka diody tunelowej U=f(I)

Lp.	***U*[V]**	***I*[mA]**
1	0.000	0.0
2	0.020	4.7
3	0.036	7.7
4	0.060	10.5
5	0.094	11.5
6	0.133	10.0

Lp.	***U*[V]**	***I*[mA]**
7	0.164	7.0
8	0.196	6.0
9	0.234	8.0
10	0.264	12.0
11	0.285	16.0
12	0.300	19.0



Rys. 3.3. Wyniki interpolacji funkcją sklejaną dla danych z przykładu 3.6.

3.5. Aproksymacja

3.5.1. Sformułowanie zagadnienia aproksymacji

Zadanie aproksymacyjne może być sformułowane bardzo różnie. W klasycznym przypadku dla danej funkcji f spośród funkcji ustalonej klasy poszukuje się funkcji F (też ustalonej klasy), która w określonym sensie najlepiej przybliża f.

Innym zadaniem jest wyznaczenie, możliwie niskim kosztem, przybliżenia F funkcji f z zadaną dokładnością. Można również stawiać problem aproksymacji nie jednej, ale całej klasy funkcji, funkcjami innej klasy. Rozwiązania tak różnie postawionych zadań są oczywiście różne, nie istnieje więc jedna „optymalna" aproksymacja [1, 4, 5, 8, 9, 10].

Funkcję $f(x)$, znaną lub określoną tablicą wartości, będziemy aproksymować (zastępować) inną funkcją $F(x)$, zwaną ***funkcją aproksymującą*** lub ***przybliżeniem*** funkcji $f(x)$. Oczywiście przybliżenie takie powoduje powstanie błędów aproksymacji.

Niech $f(x)$ będzie funkcją, którą chcemy aproksymować, X - pewną przestrzenią liniową unormowaną (tzn. określona jest w niej funkcja nazywana normą) zaś X_{m+1} – $(m+1)$-wymiarową podprzestrzenią liniową przestrzeni X.

Aproksymacja funkcji $f(x)$ polega na wyznaczeniu takich współczynników $a_0, a_1, a_2, ..., a_m$ funkcji:

$$F(x) = a_0\varphi_0(x) + a_1\varphi_1(x) + ... + a_m\varphi_m(x),$$

aby spełniała ona pewne warunki (np. minimalizowała normę różnicy $||f(x) - F(x)||$), przy czym $\varphi_0, \varphi_1, ..., \varphi_m$ są funkcjami bazowymi $m+1$ wymiarowej podprzestrzeni liniowej X_{m+1}.

Wybór odpowiedniej podprzestrzeni X_m i związanej z nią bazy (funkcji bazowych $\varphi_k(x)$) jest zagadnieniem istotnym ze względu na numeryczny koszt rozwiązania i błędy zaokrągleń.

Często obieraną podprzestrzenią X_{m+1} jest:

- podprzestrzeń funkcji trygonometrycznych z bazą:
  $1, \sin x, \cos x, \sin 2x, \cos 2x, ..., \sin kx, \cos kx,$
  szczególnie przydatna, gdy aproksymowana funkcja $f(x)$ jest funkcją okresową;
- podprzestrzeń wielomianów stopnia co najwyżej m z bazą jednomianów:
  $1, x, x^2, x^3, ..., x^m$.

Mimo prostoty działań na wielomianach, baza ta ma istotną wadę - wrażliwość na błędy zaokrągleń; kumulujące się błędy w przypadku działań na małych oraz na niewiele różniących się liczbach mogą całkowicie zniekształcić obliczenia.

- podprzestrzeń wielomianów stopnia co najwyżej *m*, określonych na przedziale <-1, 1> z bazą wielomianów Czebyszewa opisanych dalej wzorem (3.65):
  $T_0(x)$, $T_1(x)$, $T_2(x)$, ..., $T_m(x)$,
- czy też z bazą wielomianów Legendre'a wzór (3.56):
  $L_0(x)$, $L_1(x)$, $L_2(x)$, ..., $L_m(x)$.

Zagadnienie aproksymacji przy wybranych funkcjach bazowych $\varphi_k(x)$ sprowadza się do jednoznacznego wyznaczenia wartości współczynników a_k, zapewniających minimum normy $\|f(x) - F(x)\|$, czyli:

$$\|\boldsymbol{f}(\boldsymbol{x}) - \left(a_0\varphi_0(x) + a_1\varphi_1(x) + ... + a_m\varphi_m(x)\right)\| .$$

Norma jest tu rozumiana w sensie miary odległości między dwoma funkcjami. Najczęściej stosowane normy w aproksymacji to:

- norma jednostajna (Czebyszewa) (wzór 3.34),
- norma L_2 (wzór 3.35),
- norma średniokwadratowa (wzór 3.36).

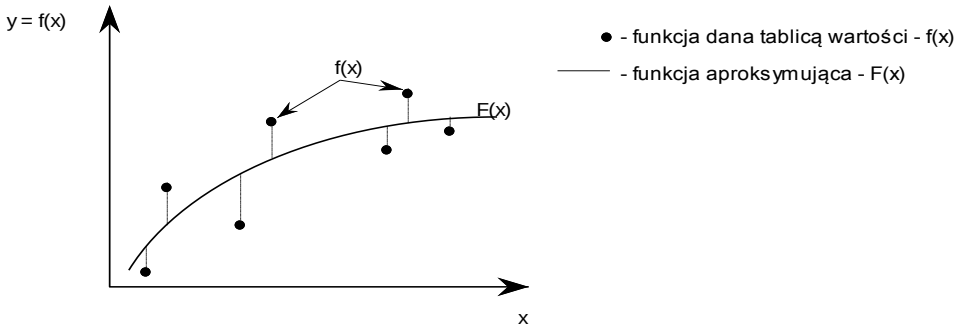
W zależności od stosowanej normy mówimy odpowiednio o aproksymacji jednostajnej (Czebyszewa), aproksymacji z normą L_2, aproksymacji średniokwadratowej.

Aproksymacja w przypadku normy Czebyszewa

Dla funkcji $f(x)$ określonej na przedziale $<a, b>$ poszukujemy funkcji $F(x)$ zapewniającej najmniejsze maksimum różnicy między $F(x)$ a $f(x)$ na całym przedziale $< a, b >$:

$$\|F(x) - f(x)\| = \sup_{x \in <a,b>} |F(x) - f(x)| . \qquad (3.34)$$

Aproksymacja taka nazywa się ***aproksymacją jednostajną***. Polega ona na takim wyznaczeniu funkcji $F(x)$, aby największa odległość jej wartości od wartości funkcji danej $f(x)$ była jak najmniejsza (rys. 3.4). Odległość ta określa jednocześnie maksymalny błąd bezwzględny z jakim funkcja $F(x)$ przybliża daną funkcję $f(x)$.



Rys. 3.4. Interpretacja graficzna aproksymacji jednostajnej

Aproksymacja w przypadku normy L_2 z wagą

Dla funkcji $f(x)$ określonej i ciągłej na przedziale $<a, b>$ poszukujemy minimum całki:

$$\|F(x) - f(x)\| = \int_a^b w(x)\left[F(x) - f(x)\right]^2 dx, \tag{3.35}$$

gdzie $w(x)$ jest ciągłą nieujemną funkcją wagową, dodatnią poza zbiorem miary zero.

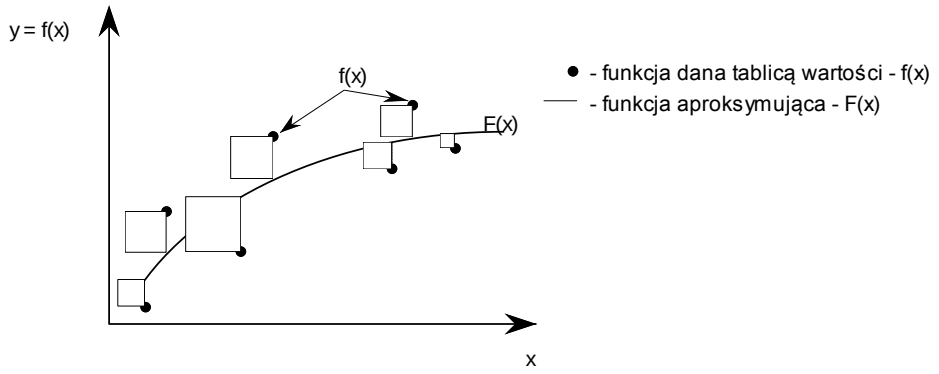
Natomiast dla funkcji $f(x_i)$, danej na dyskretnym zbiorze argumentów, poszukujemy minimum sumy (***metoda najmniejszych kwadratów***):

$$\|F(x) - f(x)\|^2 = \sum_{i=0}^{n} w(x_i)\left[F(x_i) - f(x_i)\right]^2, \tag{3.36}$$

przy czym $w(x_i)$ jest ***funkcją wagową*** taką, że $w(x_i) \geq 0$ dla $i = 0, 1, ..., n$.

Aproksymacja taka nazywa się ***aproksymacją średniokwadratową***. Polega ona na takim wyznaczeniu funkcji $F(x)$, aby suma kwadratów odległości jej wartości od wartości danej funkcji $f(x)$ była jak najmniejsza (rys. 3.5).

Aproksymacja średniokwadratowa znacznie lepiej od aproksymacji jednostajnej „eliminuje" duże błędy przypadkowe (np. wynikające z pomyłek przy pomiarach).



Rys. 3.5. Interpretacja graficzna aproksymacji średniokwadratowej

3.5.2. Aproksymacja średniokwadratowa

Niech będzie dana funkcja $y = f(x)$, która na pewnym zbiorze **X** punktów: $x_0, x_1, x_2, ..., x_n$ przyjmuje wartości $y_0, y_1, y_2, ..., y_n$. Wartości te mogą być przybliżone, obarczone pewnymi błędami (np. błędami obserwacji pomiarowych). Należy znaleźć funkcję $F(x)$ mało odchylającą się od danej funkcji $f(x)$ zarówno między węzłami, jak i w węzłach $x_0, x_1, x_2, ..., x_n$, która przybliżałaby daną funkcję tak, aby ją wygładzić.

Niech $\varphi_j(x)$, $j = 0, 1, 2, ..., m$, będzie układem funkcji bazowych podprzestrzeni X_m. Poszukujemy wielomianu uogólnionego postaci:

$$F(x) = a_0\varphi_0(x) + a_1\varphi_1(x) + ... + a_m\varphi_m(x) \tag{3.37}$$

lub:

$$F(x) = \sum_{i=0}^{m} a_i\varphi_i(x), \tag{3.38}$$

będącego najlepszym przybliżeniem średniokwadratowym funkcji $f(x)$, przy czym współczynniki a_i są tak określone, aby wyrażenie (3.36) było minimalne.

Oznaczmy:

$$H(a_0, a_1, a_2, ..., a_m) = \\ = \sum_{j=0}^{n} w(x_j)\left[f(x_j) - \sum_{i=0}^{m} a_i\varphi_i(x_j) \right]^2 = \sum_{j=0}^{n} w(x_j)R_j^2, \tag{3.39}$$

gdzie $w(x)$ jest ustaloną z góry funkcją wagową taką, że $w(x_i) \geq 0$ dla $i = 0, 1, 2, ..., n$, zaś R_i jest odchyleniem w punkcie x_i. Najczęściej przyjmuje się, że funkcja wagowa $w(x)$ ma stałą wartość, równą tożsamościowo jedności, można jednak dobrać inną funkcję wagową (np. jeżeli wartości funkcji $f(x)$ w pewnych punktach znane są z mniejszym błędem, to w celu otrzymania lepszego przybliżenia przyjmuje się w tych punktach większe wartości funkcji wagowej).

W celu znalezienia takich współczynników a_k, dla których funkcja H osiąga minimum, obliczamy pochodne cząstkowe względem zmiennych a_k i przyrównujemy je do zera:

$$\frac{\partial H}{\partial a_k} = 0 \qquad k = 0,1,2,...,m\,.$$

Otrzymujemy układ m+1 równań z m+1 niewiadomymi a_k, $k = 0, 1, 2, ..., m$:

$$\frac{\partial H}{\partial a_k} = -2\sum_{j=0}^{n} w(x_j)\left[f(x_j) - \sum_{i=0}^{m} a_i \varphi_i(x_j) \right] \varphi_k(x_j) = 0\,, \tag{3.40}$$

zwany ***układem normalnym***. Ponieważ funkcje $\varphi_j(x)$ tworzą bazę przestrzeni X_m, zatem wyznacznik układu (3.40) jest różny od zera i jednoznaczne rozwiązanie tego układu zapewnia minimum funkcji H.

W zapisie macierzowym układ (3.40) przyjmuje postać:

$$\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{A} = \mathbf{D}^{\mathrm{T}}\mathbf{f}, \tag{3.41}$$

gdzie:

$$\mathbf{D} = \begin{bmatrix} \varphi_0(x_0) & ..... & ..... & \varphi_m(x_0) \\ \varphi_0(x_1) & ..... & ..... & \varphi_m(x_1) \\ ..... & ..... & ..... & ..... \\ \varphi_0(x_n) & ..... & ..... & \varphi_m(x_n) \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}.$$

Macierz współczynników układu jest macierzą symetryczną i dodatnio określoną, co zapewnia jednoznaczność rozwiązania.

Układ (3.40) lub (3.41) powstaje z równania (3.37) po podstawieniu wartości punktów węzłowych x_i, $i = 0, 1, 2, ..., n$. Otrzymujemy wówczas nadokreślony układ n+1 równań z m+1 niewiadomymi $\mathbf{DA} = \mathbf{f}$, z którego po pomnożeniu (lewostronnie) przez $\mathbf{D}^{\mathrm{T}}$ dochodzi się do (3.41).

Jeżeli za funkcje bazowe $\varphi_j(x)$ przyjmuje się ciąg jednomianów $1, x, x^2, x^3, ..., x^m$, to wzór (3.40) można zapisać w postaci:

$$\sum_{j=0}^{n}\left[f\left(x_j\right)-\sum_{i=0}^{m} a_i x_j^i \right] x_j^k = 0 \; k = 0,1,2,...,m,$$

lub po przekształceniu:

$$\sum_{j=0}^{n} f(x_j) x_j^k = \sum_{i=0}^{m} a_i \left(\sum_{j=0}^{n} x_j^{i+k} \right) \quad k = 0, 1, 2, ..., m. \tag{3.42}$$

Oznaczając:

$$g_{ik} = \sum_{j=0}^{n} x_j^{i+k}, \; \varsigma_k = \sum_{j=0}^{n} f\left(x_j\right) x_j^k,$$

otrzymujemy układ normalny (3.40) postaci:

$$\sum_{i=0}^{m} a_i g_{ik} = \varsigma_k \;\; k = 0,1,2,...,m \tag{3.43}$$

lub:

$$\begin{array}{lllll}
a_0(n+1)+ & a_1\sum x_j + & \ldots & + a_m \sum x_j^m & = \sum f\left(x_j\right) \\
a_0\sum x_j + & a_1\sum x_j^2 + & \ldots & + a_m \sum x_j^{m+1} & = \sum f\left(x_j\right) x_j \\
a_0\sum x_j^2 + & a_1\sum x_j^3 + & \ldots & + a_m \sum x_j^{m+2} & = \sum f\left(x_j\right) x_j^2 \\
\ldots\ldots\ldots & \ldots\ldots\ldots\ldots & \ldots & \ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots \\
a_0\sum x_j^m + & a_1\sum x_j^{m+1} + & \ldots & + a_m \sum x_j^{2m} & = \sum f\left(x_j\right) x_j^m
\end{array}$$

gdzie wszystkie sumowania wykonywane są od j=0 do j=n.

Można wykazać, że jeżeli punkty $x_0, x_1, x_2, ..., x_n$ są różne i $m \le n$, to wyznacznik układu (3.43) jest różny od zera, a więc układ ten ma jednoznaczne rozwiązanie. Jeżeli $m = n$, to wielomian aproksymacyjny $F(x)$ pokrywa się z wielomianem interpolacyjnym dla punktów $x_0, x_1, x_2, ..., x_n$ i wówczas H=0. W praktyce stopień wielomianu m jest i powinien być znacznie niższy od liczby punktów n, wtedy bowiem korzystamy z dużej ilości informacji (np. wyników pomiarów) uzyskując równocześnie prostsze (niskiego stopnia) funkcje aproksymujące.

Wielomian aproksymujący daną funkcję $f(x)$ w sensie najmniejszych kwadratów powinien mieć stopień na tyle wysoki, aby dostatecznie przybliżać aproksymowaną funkcję, a jednocześnie stopień ten powinien być wystarczająco niski, aby wielomian ten wygładzał losowe błędy wynikające np. z pomiarów. W praktyce stopień wielomianu określamy a priori na podstawie analizy modelu fizycznego badanego zjawiska bądź też przeprowadzamy aproksymację kolejno wielomianami coraz to wyższych stopni i obliczamy odchylenia funkcji H.

Dla $m \geq 6$ układ (3.43) jest źle uwarunkowany, wskutek czego otrzymane wyniki obliczeń mogą być tak bardzo zaburzone, iż nie nadają się do praktycznego wykorzystania.

Niech x_i będą rozłożone w jednakowych odstępach w przedziale $< 0, 1 >$. Liczby g_{ik} występujące we wzorze (3.43) można dla dużych m przybliżyć następująco:

$$g_{ik} = \sum_{j=1}^{m} x_j^{i+k} \approx m \int_0^1 x^{i+k} dx = \frac{m}{i+k+1}, \; i, k = 0, 1, 2, ..., m.$$

Macierz współczynników układu (3.43) ma postać:

$$\mathbf{A} = \begin{bmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{m+1} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{m+2} \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ \frac{1}{m+1} & \frac{1}{m+2} & \cdots & \frac{1}{2m+1} \end{bmatrix}.$$

Elementy macierzy odwrotnej $\mathbf{G}^{-1}$ są rzędu $3 \cdot 10^{12}$, co powoduje błędy zaokrągleń tak duże, że wyniki praktycznie tracą sens. Zatem stosowanie aproksymacji z funkcjami bazowymi typu jednomianów x^i ma sens jedynie dla małych m ($m < 6$).

W celu aproksymacji danej funkcji wielomianami wyższych stopni należy:

- zastosować specjalną metodę rozwiązywania układów równań, których macierz współczynników ma wyznacznik bliski zeru;
- zwiększyć precyzję (dokładność) wykonywania obliczeń;
- zastąpić bazę jednomianów x^i bazą złożoną z wielomianów ortogonalnych.

Przykład 3.7.

W tabeli 3.5 dane są wyniki pewnych pomiarów. Metodą najmniejszych kwadratów znaleźć funkcję liniową, która najlepiej aproksymuje podane dane.

Tabela 3.5. Wyniki pomiarów do przykładu 3.7

j	0	1	2	3	4	5	6	7
x_j	1	3	4	6	8	9	11	14
$f(x_j)$	1	2	4	4	5	7	8	9

W celu znalezienia funkcji liniowej, aproksymującej dane z tabeli, należy wyznaczyć funkcję postaci (3.37):

$$F(x) = y(x) = a_0\varphi_0(x) + a_1\varphi_1(x) = a_0 + a_1 x = \sum_{i=0}^{1} a_i\varphi_i(x_j)$$

dla j=0,1,...,7 oraz $\varphi_0(x) = x^0$, $\varphi_1(x) = x^1$.

Określając funkcję H, zgodnie z (3.39), otrzymujemy:

$$H(a_0, a_1) = \sum_{j=0}^{7}\left[f(x_j) - F(x_j)\right]^2 = \sum_{j=0}^{7}\left[f(x_j) - a_0\,\varphi_0(x_j) - a_1\varphi_1(x_j)\right]^2 =$$

$$= \sum_{j=0}^{7}\left[f(x_j) - a_0 - a_1 x_j\right]^2$$

W celu wyznaczenia szukanych współczynników a_0, a_1 obliczamy pochodne cząstkowe funkcji H względem zmiennych a_k oraz przyrównujemy je (patrz wzór 3.40) do zera:

$$\frac{\partial H}{\partial a_k} = -2\sum_{j=0}^{7}\left[f(x_j) - \sum_{i=0}^{m=1} a_i\varphi_i(x_j)\right]\varphi_k(x_j) = 0$$

W ten sposób otrzymujemy układ dwóch równań liniowych z dwiema niewiadomymi:

$$\begin{cases} \dfrac{\partial H}{\partial a_0} = -2\displaystyle\sum_{j=0}^{7}\left[f(x_j) - \sum_{i=0}^{m=1} a_i\,\varphi_i(x_j)\right]\varphi_0(x_j) = 0 \\ \dfrac{\partial H}{\partial a_1} = -2\displaystyle\sum_{j=0}^{7}\left[f(x_j) - \sum_{i=0}^{m=1} a_i\,\varphi_i(x_j)\right]\varphi_1(x_j) = 0 \end{cases}.$$

Podstawiając do powyższego układu $\varphi_0(x) = x^0 = 1, \varphi_1(x) = x$ oraz dzieląc obustronnie oba równania przez (-2) mamy:

$$\begin{cases} \sum_{j=0}^{7} \left[f(x_j) - a_0 - a_1 x_j \right] \cdot 1 = 0 \\ \sum_{j=0}^{7} \left[f(x_j) - a_0 - a_1 x_j \right] \cdot x_j = 0 \end{cases}.$$

Podstawiając następnie za x_j i $f(x_j)$, $j=0,...,7$ wartości z tabeli 3.5 pierwsze równanie powyższego układu przyjmie postać:

$$(1 - a_0 - 1 \cdot a_1) + (2 - a_0 - 3 \cdot a_1) + (4 - a_0 - 4 \cdot a_1) + (4 - a_0 - 6 \cdot a_1) +$$
$$+ (5 - a_0 - 8 \cdot a_1) + (7 - a_0 - 9 \cdot a_1) + (8 - a_0 - 11 \cdot a_1) + (9 - a_0 - 14 \cdot a_1) = 0$$

a drugie:

$$(1 - a_0 - 1 \cdot a_1) \cdot 1 + (2 - a_0 - 3 \cdot a_1) \cdot 3 + (4 - a_0 - 4 \cdot a_1) \cdot 4 +$$
$$+ (4 - a_0 - 6 \cdot a_1) \cdot 6 + (5 - a_0 - 8 \cdot a_1) \cdot 8 + (7 - a_0 - 9 \cdot a_1) \cdot 9 +$$
$$+ (8 - a_0 - 11 \cdot a_1) \cdot 11 + (9 - a_0 - 14 \cdot a_1) \cdot 14 = 0$$

Po dalszych uproszczeniach otrzymujemy:

$$\begin{cases} 40 - 8\ a_0 - 56 \cdot a_1 = 0 \\ 364 - 56 a_0 - 524 \cdot a_1 = 0 \end{cases}$$

$$\begin{cases} 8\ a_0 + 56 \cdot a_1 = 40 \\ 56 a_0 + 524 \cdot a_1 = 364 \end{cases}$$

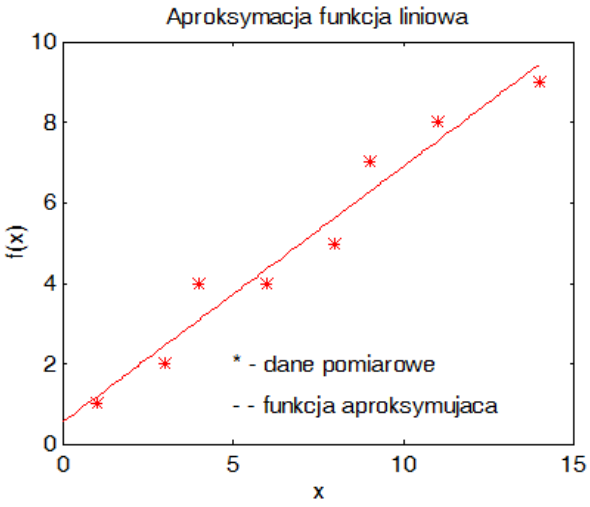
$$\begin{cases} a_0 + 7 \cdot a_1 = 5 \\ 14 a_0 + 131 \cdot a_1 = 91 \end{cases}.$$

Rozwiązaniem układu jest:

$$\begin{cases} a_0 = \frac{6}{11} \\ a_1 = \frac{7}{11} \end{cases}.$$

Poszukiwana funkcja $F(x)$ ma wobec tego postać: $F(x) = \frac{6}{11} + \frac{7}{11} x$.

Graficzną reprezentację przykładu demonstruje rys. 3.6.



Rys. 3.6. Ilustracja graficzna przykładu 3.7

Przykład 3.8.

Dla danych z tabeli 3.6 znaleźć metodą najmniejszych kwadratów funkcję postaci:

$y = a_0\sqrt{x} + a_1.$

Tabela 3.6. Dane do przykładu 3.8

x_i	1	4	9	25	36
y_i	-6	-9	-12	-18	-21

Po sprowadzeniu problemu do postaci liniowej:

$y = a_0\bar{x} + a_1$, gdzie $\bar{x}_i = \sqrt{x_i}$ dla i=0,1,…4,

można zastosować metodę dla funkcji liniowej czyli można od początku wyprowadzać funkcję *H*, liczyć jej pochodne i przyrównywać je do zera albo po prostu wykorzystać wzór (3.43) dla *m=1* (liczba naszych funkcji bazowych) oraz *n=4* (liczba węzłów aproksymacji):

$$\begin{cases} a_0(4+1) + a_1\sum_{j=0}^{4}\bar{x}_j = \sum_{j=0}^{4} f(x_j) \\ a_0\sum_{j=0}^{4}\bar{x}_j + a_1\sum_{j=0}^{4}(\bar{x}_j)^2 = \sum_{j=0}^{4} f(x_j)\bar{x}_j \end{cases}.$$

Nastepnie należy policzyć odpowiednie sumy występujące w tym układzie równań:

$$\sum_{j=0}^{4} \bar{x}_j = \sum_{j=0}^{4} \sqrt{x_j} = 1 + 2 + 3 + 5 + 6 = 17,$$

$$\sum_{j=0}^{4} (\bar{x}_j)^2 = \sum_{j=0}^{4} (\sqrt{x_j})^2 = \sum_{j=0}^{4} x_j = 1 + 4 + 9 + 25 + 36 = 75,$$

$$\sum_{j=0}^{4} f(x_j) = -6 - 9 - 12 - 18 - 21 = -66,$$

$$\sum_{j=0}^{4} f(x_j)\bar{x}_j = \sum_{j=0}^{4} f(x_j)\sqrt{x_j} == -6 \cdot 1 - 9 \cdot 2 - 12 \cdot 3 - 18 \cdot 5 - 21 \cdot 6 = -276.$$

Ostatecznie otrzymujemy układ równań postaci:

$$\begin{cases} 5a_0 + 17a_1 = -66 \\ 17a_0 + 75 \cdot a_1 = -276 \end{cases}.$$

Rozwiązaniem układu są liczby $a_0 = -3$ i $a_1 = -3$ czyli szukana funkcja ma postać:

$$y = -3\sqrt{x} - 3.$$

Przykład 3.9.

Dla danych doświadczalnych z tabeli 3.7 znaleźć metodą najmniejszych kwadratów krzywą typu hiperboli.

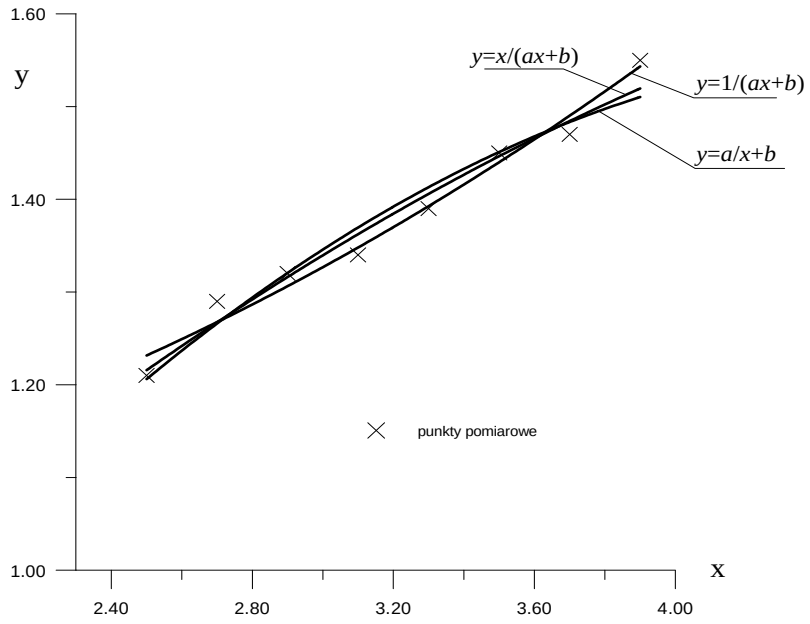
Tabela 3.7. Dane do przykładu 3.9

x_i	2,5	2,7	2,9	3,1	3,3	3,5	3,7	3,9
y_i	1,21	1,29	1,32	1,34	1,39	1,45	1,47	1,55

Poszukujemy funkcji aproksymujących typu:

a) $y = a/x + b$,
b) $y = 1/(ax+b)$,
c) $y = x/(ax+b)$.

W każdym przypadku zadanie sprowadza się do problemu liniowego. Wyniki obliczeń przedstawione są na rys. 3.7.



Rys. 3.7. Ilustracja graficzna do przykładu 3.9.

W wielu zagadnieniach technicznych często stosowaną funkcją przybliżającą jest sinus hiperboliczny lub cosinus hiperboliczny, które definiujemy następująco:

$$sh(x) = \frac{e^x - e^{-x}}{2},\ ch(x) = \frac{e^x + e^{-x}}{2}.$$

Przykładowo n par punktów pomiarowych indukcji magnetycznej i natężenia pola elektromagnetycznego (B_1,H_1), (B_2,H_2), ... (B_n,H_n) tworzy doświadczalną krzywą magnesowania. Stosując metodę najmniejszych kwadratów można znaleźć funkcję:

$H = a_0\ \mathrm{sh}(a_1 B)$,

tak, aby zminimalizować wyrażenie:

$$S(a_0, a_1) = \sum_{i=1}^{n} w_i \left(H_i - a_0 \mathrm{sh}(a_1 B_i)\right)^2 , \tag{3.44}$$

gdzie w_i jest wagą statystyczną i - tego punktu, $i = 1,2,...n$.

W tym celu należy wyznaczyć pochodne cząstkowe funkcji S względem a_0 i a_1 oraz przyrównać je do zera. Otrzymamy układ równań:

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = \sum_{i=1}^{n} w_i a_0 \mathrm{sh}^2(a_1 B_i) - \sum_{i=1}^{n} w_i H_i \mathrm{sh}(a_1 B_i) = 0 \\ \dfrac{\partial S}{\partial a_1} = \sum_{i=1}^{n} w_i H_i B_i \mathrm{ch}(a_1 B_i) - a_0 \sum_{i=1}^{n} w_i B_i \mathrm{sh}(a_1 B_i)\mathrm{ch}(a_1 B_i) = 0 \end{cases} . \tag{3.45}$$

Z pierwszego z równań (3.45) wyznaczamy współczynnik a_0:

$$a_0 = \frac{\sum_{i=1}^{n} w_i H_i \mathrm{sh}(a_1 B_i)}{\sum_{i=1}^{n} w_i \mathrm{sh}^2(a_1 B_i)}, \tag{3.46}$$

i po wstawieniu do drugiego równania otrzymujemy równanie z niewiadomą a_1:

$$\begin{aligned} &\left(\sum_{i=1}^{n} w_i H_i B_i \mathrm{ch}(a_1 B_i)\right)\left(\sum_{i=1}^{n} w_i \mathrm{sh}^2(a_1 B_i)\right) - \\ &-\left(\sum_{i=1}^{n} w_i H_i \mathrm{sh}(a_1 B_i)\right)\left(\sum_{i=1}^{n} w_i B_i \mathrm{sh}(a_1 B_i)\mathrm{ch}(a_1 B_i)\right) = 0. \end{aligned} \tag{3.47}$$

Równanie (3.47) rozwiązuje się stosując jedną z metod omówionych w rozdziale 4.

Zamiast minimalizowania błędu bezwzględnego (3.44) można minimalizować błąd względny:

$$S_{\mathrm{wzgl}}(a_0, a_1) = \sum_{i=1}^{n} \left[\frac{w_i \left(H_i - a_0 \mathrm{sh}(a_1 B_i)\right)}{H_i} \right]^2 . \tag{3.48}$$

Podobnie jak w przypadku błędu bezwzględnego, uzyskuje się równanie z niewiadomą a_1.

W przypadku aproksymacji średniokwadratowej funkcji $f(x)$ ciągłej na przedziale $\langle a, b\rangle$ poszukuje się funkcji:

$$F(x) = a_0\varphi_0(x) + a_1\varphi_1(x) + \ldots + a_m\varphi_m(x),$$

gdzie φ_0, φ_1, ..., φ_m są elementami bazy pewnej podprzestrzeni funkcji całkowalnych z kwadratem na przedziale $\langle a, b\rangle$.

Aproksymacja średniokwadratowa funkcji ciągłych polega na znalezieniu takiego ciągu współczynników a_i (i = 0, 1, 2, ..., m), aby otrzymać minimum normy (3.35). W celu rozwiązania zadania należy utworzyć układ m+1 równań z m+1 niewiadomymi a_i :

$$\frac{\partial H}{\partial a_k} = 0, \quad k = 0,1,2,...,m \text{, gdzie:}$$

$$H(a_0, a_1, ..., a_m) = \int_a^b w(x)\left[F(x) - f(x)\right]^2 dx =$$

$$= \int_a^b w(x)\left[\sum_{i=0}^{m} a_i \varphi_i(x) - f(x)\right]^2 dx.$$

Rozwiązanie tego układu wyznaczy poszukiwaną funkcję aproksymującą.

Przykład 3.10.

Znaleźć aproksymację doświadczalnej krzywej magnesowania obwodu magnetycznego dla danych zestawionych w tabeli 3.8, za pomocą funkcji sinus hiperboliczny.

Tabela 3.8. Dane wejściowe do aproksymacji krzywej magnesowania

Lp.	*B*[T]	*H*[A/m]	Waga
1	0.00	0	1
2	0.80	40	1
3	1.00	135	1
4	1.20	340	1
5	1.40	720	1
6	1.50	1000	1
7	1.60	1320	1
8	1.70	1650	1
9	1.80	2500	1

Lp.	*B*[T]	*H*[A/m]	Waga
10	1.90	4000	1
11	2.00	7500	1
12	2.10	15200	1
13	2.15	20000	1
14	2.20	26200	1
15	2.25	33000	1
16	2.30	41500	1
17	2.35	50000	1
18	2.40	61000	1
19	2.45	74500	1

Stosując metodę najmniejszych kwadratów do funkcji *sinh* o współczynnikach a_0, a_1 , tzn.:

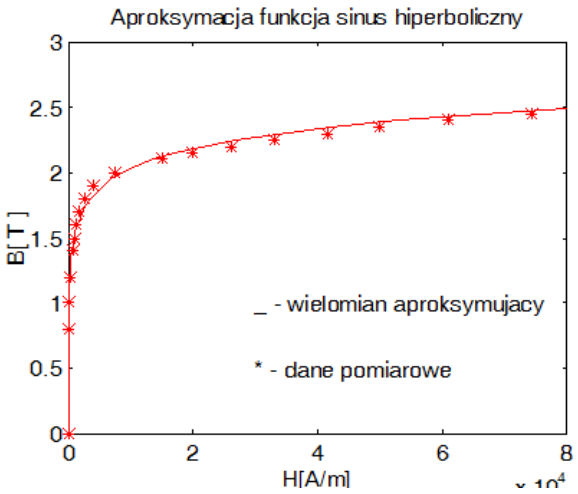
H=a_0 sinh(a_1B),

otrzymano następujące wartości współczynników:

a_0=1.9966 A/m,

a_1=4.5328 1/T.

Wyniki aproksymacji przedstawiono graficznie na rys. 3.8.



Rys. 3.8. Wykres wyników aproksymacji krzywej magnesowania

3.6. Wielomiany ortogonalne

Aproksymacja z funkcjami bazowymi typu x^i powoduje, że wraz ze wzrostem stopnia wielomianu obliczenia stają się coraz bardziej pracochłonne, a ponadto ich wyniki są niepewne. Ponadto zmiana stopnia wielomianu przybliżającego wymaga ponownego rozwiązywania układu normalnego (3.43), co też przemawia przeciwko stosowaniu bazy funkcji $\varphi_i(x) = x^i$ do aproksymacji. Obie te trudności można usunąć używając do aproksymacji ***wielomianów ortogonalnych***.

Ciąg $P_0, P_1, ..., P_n$, gdzie P_k ($k = 0, 1, ..., n$) jest wielomianem stopnia dokładnie k, nazywamy:

a) ciągiem wielomianów ortogonalnych na przedziale [a, b] z wagą p, jeśli tworzą one układ ortogonalny w przestrzeni $\boldsymbol{L^2p}$ [$\boldsymbol{a}$, $\boldsymbol{b}$], tzn.

$$\left(P_k, P_l\right) = \int_a^b P_k(x) P_l(x)\, p(x)\mathrm{d}x = 0 \text{ dla } k \neq l, k, l = 0,1,...,n \qquad (3.49)$$

b) ciągiem wielomianów ortogonalnych na zbiorze dyskretnym $\{x_1, x_2, ..., x_N\}$ z wagą p, jeśli tworzą one układ ortogonalny w przestrzeni $L^2_{p,N}$, tzn.:

$$\left(P_k, P_l\right) = \sum_{i=1}^{N} P_k(x_i) P_l(x_i)\, p(x_i) = 0 \text{ dla } k \neq l, k, l = 0,1,...,n, \qquad (3.50)$$

gdzie $n \leq N$-1, zaś $p(x_i)$, $i = 1, 2, ..., N$ są danymi liczbami dodatnimi.

W przypadku a) ciąg $P_0, P_1, ..., P_n$ może być nieskończony, zaś w przypadku b) jest skończony. Tam, gdzie nie będzie to istotne, nie będziemy rozróżniać wielomianów ortogonalnych w sensie a) i b), mówiąc po prostu o wielomianach ortogonalnych.

Wielomiany ortogonalne $P_0, P_1, ..., P_n$ tworzą bazę przestrzeni liniowej W_n wielomianów stopnia nie wyższego niż n.

Twierdzenie 3.8.

W przestrzeni L^2_p [a, b] lub w przestrzeni $l^2_{p,N}$, ciąg wielomianów ortogonalnych jest wyznaczony jednoznacznie z dokładnością do mnożników liczbowych.

Wielomiany ortogonalne spełniają także zależność rekurencyjną, tzw. regułę trójczłonową.

Twierdzenie 3.9.

Wielomiany ortogonalne P_k (k = 0, 1, ..., n) spełniają zależność:

$$\begin{aligned} &P_{-1}(x) = 0, \\ &P_0(x) = a_0, \\ &P_k(x) = (\alpha_k x + \beta_k)P_{k-1}(x) + \gamma_k P_{k-2}(x), \quad k = 1,2,...,n. \end{aligned} \tag{3.51}$$

gdzie:

$$\begin{aligned} &\alpha_k = \frac{a_k}{a_{k-1}} \neq 0, \ \beta_k = -\frac{\alpha_k (xP_{k-1}, P_{k-1})}{(P_{k-1}, P_{k-1})}, \\ &\gamma_k = -\frac{a_k}{a_{k-1}} \frac{(P_{k-1}, P_{k-1})}{(P_{k-2}, P_{k-2})}, \ \ \gamma_k \neq 0, \end{aligned} \tag{3.52}$$

a_k oznaczają współczynniki wielomianu $P_k(x) = a_k x^k + \ldots$.

Z definicji k - ty wielomian ortogonalny P_k jest dokładnie stopnia k, zatem jego współczynnik $a_k \neq 0$. Jeśli $\alpha_k = a_k / a_{k-1}$, to $P_k - \alpha_k P_{k-1}(x)$ jest wielomianem stopnia k-1 i można go przedstawić w postaci:

$$P_k - \alpha_k x P_{k-1} = \sum_{i=0}^{k-1} b_i P_i \, . \tag{3.53}$$

W przestrzeniach $L^2_p[a, b]$ i $l^2_{p,N}$ zachodzi równość:

$$(xP_{k-1}, P_j) = (P_{k-1}, xP_j) \, .$$

Zatem, na podstawie twierdzenia 3.8, dla j<k-2 mamy:

$$(P_k - \alpha_k x P_{k-1}, P_j) = (P_k, P_j) - \alpha_k (P_{k-1}, xP_j) \, .$$

Z zależności (3.52) otrzymuje się:

$$(P_k - \alpha_k x P_{k-1}, P_j) = \sum_{i=0}^{k-1} b_i \left(P_i, P_j\right) = b_j \left(P_j, P_j\right),$$

czyli $b_j = 0$ dla $j < k - 2$. Wzór (3.53) można zatem zapisać w postaci:

$$P_k = \alpha_k \, x \, P_{k-1} + \beta_k \, P_{k-1} + \gamma_k \, P_{k-2} \, , \tag{3.54}$$

gdzie $\beta_k = b_{k-1}$, $\gamma_k = b_{k-2}$. Stałe te można wyznaczyć mnożąc skalarnie obie strony równości (3.54) przez P_{k-1}:

$$0 = (P_k \, , P_{k-1} \,) = \alpha_k \, (xP_{k-1}, P_{k-1}) + \beta_k \, (P_{k-1}, P_{k-1}),$$

skąd otrzymuje się wartość β_k jak we wzorach (3.52).

Analogicznie wyznacza się wartość γ_k dla $k = 2, 3, ..., n$.

W przypadku wielomianów ortogonalnych na zbiorze dyskretnym $\{x_1, x_2, ..., x_N\}$ zależność rekurencyjna (3.51) jest spełniona tylko dla $k \le N$-1. Zauważmy, że wielomian stopnia N:

$$a_N(x - x_1)(x - x_2) ... (x - x_N), \tag{3.55}$$

zeruje się w każdym z punktów x_k, skąd wynika, że jest prostopadły do wszystkich wielomianów ortogonalnych P_j niższego stopnia. Tym samym N - ty wielomian ortogonalny P_N musiałby być postaci (3.55), ponieważ z twierdzenia 3.8 wiadomo, że jest on wyznaczany jednoznacznie z dokładnością do stałego czynnika. Dla wielomianu:

$$P_N = a_N(x - x_1)(x - x_2) ... (x - x_N),$$

zachodzi równość $(P_N, P_N) = 0$, co dowodzi, że jest on zerowym elementem przestrzeni zerowej $l^2_{p,N}$, nie może więc być elementem układu ortogonalnego. Jest to oczywiste, bo przestrzeń $l^2_{p,N}$ ma wymiar N, nie może więc zawierać układu liniowo niezależnego o więcej niż N elementach.

Reguła trójczłonowa (3.51) umożliwia konstrukcję wielomianów ortogonalnych. Są one wyznaczone jednoznacznie z dokładnością do mnożników liczbowych. Możemy więc dowolnie ustalić wartości współczynników a_k. Biorąc $a_k = 1$ ($k = 0, 1, ..., n$) otrzymujemy $\alpha_k = 1$, a tym samym prostszą postać wzorów (3.52).

Znając już wielomiany $P_0 = 1$, $P_1, P_2, ..., P_{k-1}$, wyznaczamy P_k z zależności:

$$P_k(x) = (x+\beta_k)P_{k-1}(x) + \gamma_k P_{k-2}(x).$$

Koszt otrzymania kolejnego wielomianu ortogonalnego jest równy kosztowi obliczenia dwóch iloczynów skalarnych (xP_{k-1}, P_{k-1}) oraz (P_{k-1}, P_{k-1}) - iloczyn (P_{k-2}, P_{k-2}) musiał być obliczony wcześniej przy wyznaczaniu wielomianu P_{k-1}. Znalezienie n wielomianów ortogonalnych z reguły trójczłonowej wymaga zatem obliczenia $2n - 2$ iloczynów skalarnych.

Twierdzenie 3.10.

Niech $\{P_n\}_{n=0}^{\infty}$ będzie ciagiem wielomianów ortogonalnych w przedziale (a,b) z wagą p. Wówczas wielomian P_n, n=1,2, ... ma n zer rzeczywistych, pojedynczych, leżących w przedziale (a,b).

Twierdzenie 3.11.

W przypadku wielomianów ortogonalnych układ równań w aproksymacji średniokwadratowej sprowadza się do macierzy diagonalnej.

W obliczeniach numerycznych nie powinno się wielomianów ortogonalnych i ich kombinacji liniowych, reprezentować względem bazy $1, x, x^2..., x^n$. Cała informacja o wielomianach ortogonalnych powinna wyrażać się współczynnikami α_k, β_k, γ_k formuły trójczłonowej i ewentualnie normami $\| P_i \|^2 = (P_i, P_i)$. Z formuły trójczłonowej należy korzystać także przy obliczaniu wartości wielomianu ortogonalnego w danym punkcie i przy wyznaczaniu wartości ich kombinacji liniowej.

Przykładami ciągów wielomianów ortogonalnych są wielomiany ***Legendre'a***, ***Hermite'a***, ***Grama*** i ***Czebyszewa***.

Wielomiany Legendre'a

Wielomiany Legendre'a są zdefiniowane wzorami:

$$P_0(x)=1,\ P_k(x)=\frac{1}{2^k k!}\frac{d^k}{dx^k}(x^2-1)\quad \text{dla } k=1,2,\ldots\ , \tag{3.56}$$

i tworzą one ciąg wielomianów ortogonalnych na przedziale [-1, 1] z funkcją wagową $p(x)$=1 oraz spełniają zależność rekurencyjną:

$$P_0(x)=1,\ P_k(x)=\frac{2k-1}{k}xP_{k-1}(x)-\frac{k-1}{k}P_{k-2}(x)\quad \text{dla } k=1,2,\ldots \tag{3.57}$$

Korzystając z (3.57) można pokazać, że:

$$P_k(x)=\frac{1}{2k}\sum_{i=0}^{[k/2]}(-1)^i\binom{k}{i}\binom{2k-2i}{k}x^{k-2i}. \tag{3.58}$$

Ponadto:

$$\|P_k\|^2=\int_{-1}^{1}(P_k(x))^2\,\mathrm{d}x=\frac{2}{2k-1}. \tag{3.59}$$

Wielomiany Hermite'a

Wielomiany Hermite'a są określone wzorem:

$$H_0(x)=1,\ H_k(x)=(-1)^k e^{x^2}\frac{d^k}{dx^k}e^{-x^2}\ \text{dla}\ k=1,2,... \tag{3.60}$$

lub w postaci jawnej:

$$H_k(x)=k!\sum_{i=0}^{[k/2]}(-1)^i\frac{(2x)^{k-2i}}{i!(k-2i)!}. \tag{3.61}$$

Tworzą one ciąg wielomianów ortogonalnych w przestrzeni $L^2{}_p(-\infty,+\infty)$, z funkcją wagową $p(x)=e^{-x^2}$.

Reguła trójczłonowa w tym przypadku ma postać:

$$H_0(x)=1,\ \ H_k(x)=2xH_{k-1}(x)-(2k-2)H_{k-2}(x),\ k=1,2,.... \tag{3.62}$$

Norma $||H_k||$ jest równa:

$$\|H_k\|^2=\int_{-\infty}^{+\infty}e^{-x^2}\left(P_k(x)\right)^2\mathrm{d}x=\sqrt{\pi}\,2^k\,k! \tag{3.63}$$

Wielomiany Grama

Jeśli w przestrzeni $l^2_{p,N}$, punkty $x_1, x_2, ..., x_N$, położone są w równej odległości (bez zmniejszenia ogólności można założyć, że leżą one w przedziale [-1, 1], czyli:

$$x_i=\frac{2(i-1)}{N-1}-1,$$

zaś wagi $p(x_i)$ są równe jedności, to ciągiem wielomianów ortonormalnych na zbiorze dyskretnym $\{x_1, x_2, ..., x_N\}$, a więc ortonormalnych w $l^2_{p,N}$, są wielomiany Grama $G_0, G_1, ..., G_{N-1}$:

$$(G_k,G_l)=\sum_{i=1}^{N}G_k(x_i),\ G_l(x_i)=\begin{cases}0 & \text{dla}\ k\neq l\\ 1 & \text{dla}\ k=l\end{cases}\quad k,l=0,1,...,n.$$

Spełniają one zależność rekurencyjną:

$$G_0(x)=\frac{1}{\sqrt{N}},$$

$$G_k(x) = \alpha_k x G_{k-1}(x) - \gamma_k G_{k-2}(x), k = 1,2,..., N-2, \tag{3.64}$$

przy czym:

$$\alpha_k = \frac{N-1}{k}\sqrt{\frac{4k^2-1}{N^2-k^2}},\ \gamma_k = \frac{\alpha_k}{\alpha_{k-1}}.$$

Dla k znacznie mniejszych od $\sqrt{N}$ wielomiany Grama G_k są bliskie wielomianów Legendre'a P_k. Natomiast dla k istotnie większych od $\sqrt{N}$ wielomiany G_k mają dużą normę jednostajną w przedziale [-1, 1], a ich wartości silnie oscylują między punktami x_i.

Wielomiany Czebyszewa

Wielomiany Czebyszewa pierwszego rodzaju są zdefiniowane wzorem:

$$T_k(x) = \frac{\left(x+\sqrt{x^2-1}\right)^k + \left(x-\sqrt{x^2-1}\right)^k}{2} \quad \text{dla } k = 0,1,2,... \ . \tag{3.65}$$

Dla $|x| \le 1$, podstawiając $x = \cos t$ (tzn. $t = \arccos x$) dostajemy:

$$T_k(x) = \frac{(\cos t + j\sin t)^k + (\cos t - j\sin t)^k}{2} = \cos kt,\ j = \sqrt{-1}, \tag{3.66}$$

a zatem:

$$T_k(x) = \cos(k \arccos x). \tag{3.67}$$

Z (3.67) oraz z tożsamości trygonometrycznej:

$\cos kt + \cos(k\text{-}2)t = 2\cos t \cos(k\text{-}1)t$,

wynika zależność rekurencyjna dla wielomianów Czebyszewa:

$$T_0(x) = 1,\ T_1(x) = x,\ T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x),\ k = 2,3,.... \tag{3.68}$$

Łatwo sprawdzić, że dla $|x| > 1$ wielomiany Czebyszewa spełniają także równość (3.68).

Mamy więc:

$T_0(x) = 1$,

$T_1(x) = x$,

$T_2(x) = 2x^2 - 1$,

$T_3(x) = 4x^3 - 3x$,

$T_4(x) = 8x^4 - 8x^2 + 1$,

... itd.

Z definicji (3.65) wynika, że wielomiany Czebyszewa stopnia parzystego są funkcjami parzystymi, a stopnia nieparzystego - funkcjami nieparzystymi tzn. $T_k(-x) = (-1)^k T_k(x)$.

Twierdzenie 3.12.

Wielomiany Czebyszewa T_k, $k = 0, 1, ...$ tworzą układ ortogonalny w przestrzeni $L^2_p[-1, 1]$, z funkcją wagową:

$$p(x) = \frac{1}{\sqrt{1-x^2}}.$$

Ponadto wielomian T_k, $k = 1, 2, ...$ w przedziale $[-1, 1]$ ma $k+1$ punktów ekstremalnych y_m:

$$y_m = \cos\frac{m\pi}{k} \quad m = 0,1,...,k,$$ w których $T_k(y_m) = (-1)^m$.

Współczynnik przy najwyższej potędze k-tego wielomianu Czebyszewa jest równy 2^{k-1}.

Przykład 3.11.

Zanalizować wartości wielomianów stopnia 10 przybliżających funkcję $f(x) = \frac{1}{25x^2 + 1}$ na siatce:

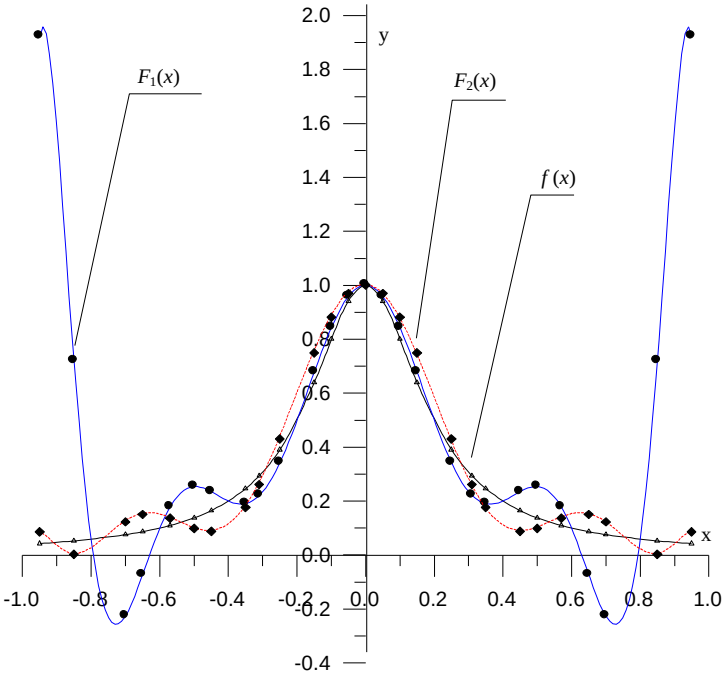
a) węzłów równoodległych $x_k = -1 + \frac{2k}{10}$, $k = 0, 1, ..., 10$,

b) węzłów Czebyszewa $x_k = \cos\frac{2k+1}{11}\frac{\pi}{2}$, $k = 0, 1, ..., 10$.

Wyniki obliczeń są zestawione w tabeli 3.9 i przedstawione na rys. 3.9.

Tabela 3.9. Wartości wielomianów aproksymujących z przykładu 3.11

Węzły interpolacji x_i	**Wartości funkcji $f(x_i)$**	**Wartości wielomianu aproksymującego na siatce węzłów:**	
		równoodległych $F_1(x_i)$	**Czebyszewa $F_2(x_i)$**
-0.95	0.042440	1.923631	0.085535
-0.85	0.052459	0.719459	0.002516
-0.7	0.075472	-0.226196	0.122414
-0.65	0.086486	-0.072604	0.150329
-0.57	0.109619	0.178626	0.136558
-0.5	0.137931	0.253755	0.098672
-0.45	0.164948	0.234969	0.087856
-0.35	0.246154	0.190580	0.176362
-0.31	0.293902	0.221343	0.260728
-0.25	0.390244	0.342641	0.429591
-0.15	0.64	0.67899	0.748897
-0.1	0.8	0.843407	0.881247
-0.05	0.941176	0.958627	0.969191
0.0	1.0	1.0	1.0
0.05	0.941176	0.958627	0.969191
0.1	0.8	0.843407	0.881247
0.15	0.64	0.67899	0.748897
0.25	0.390244	0.342641	0.429591
0.31	0.293902	0.221343	0.260728
0.35	0.246154	0.190580	0.176362
0.57	0.109619	0.178626	0.136558
0.65	0.086486	-0.072604	0.150329
0.7	0.075472	-0.226196	0.122414
0.85	0.052459	0.719459	0.002516
0.95	0.042440	1.923631	0.085535



Rys. 3.9. Ilustracja graficzna przykładu 3.11

3.7. Zadania do samodzielnego rozwiązania

Zadanie 3.1.

Dla danych z tabeli 3.10 wyznaczyć wielomian interpolacyjny Lagrange'a i obliczyć przybliżoną wartość funkcji w punktach 0 oraz 7.

Tabela 3.10. Dane do zadania 3.1

x_i	-2	-1	1	3	4
f_i	-90	-80	72	80	-180

Odp.
$W_4(x) = -x^4 - 7x^3 + 13x^2 + 83x - 16$,
f(0) ≈ *–16*,
f(7) – nie można interpolować w punkcie *x*=7, gdyż leży on poza zakresem węzłów.

Zadanie 3.2.

Dla danych z tabeli 3.11 obliczyć, wykorzystując wielomian interpolacyjny Lagrange'a, przybliżoną wartość funkcji w punktach 2, -1.

Tabela 3.11. Dane do zadania 3.2

x_i	-3	0	1	4	5
f_i	1344	120	96	-252	-320

Odp.
$W_4(x) = 3x^4 - 23x^3 + 29x^2 - 33x + 120,$
$f(2) \approx 34,\ f(-1) \approx 208.$

Zadanie 3.3.

Dla danych z tabeli 3.12 wyznaczyć tablicę ilorazów różnicowych wykorzystywanych w interpolacji Newtona.

Tabela 3.12. Dane do zadania 3.3

x_i	-3	-1	1	2	3
f_i	1	-5	5	-14	-161

Odp.
$[x_0,x_1] = -3,\ [x_1,x_2] = 5,\ [x_2,x_3] = -19,\ [x_3,x_4] = -147,$
$[x_0,x_1,x_2] = 2,\ [x_1,x_2,x_3] = -8,\ [x_2,x_3,x_4] = -64,$
$[x_0,x_1,x_2,x_3] = -2,\ [x_1,x_2,x_3\ x_4] = -2,$
$[x_0,x_1,x_2,x_3,x_4] = -2.$

Zadanie 3.4.

Dla danych z tabeli 3.13 wyznaczyć wielomian interpolacyjny Newtona i obliczyć przybliżoną wartość funkcji w punktach 1 oraz -1.

Tabela 3.13. Dane do zadania 3.4

x_i	-4	-2	0	2	5
f_i	-2	-4	-22	-152	1123

Odp.
$W_4(x) = 2x^4 + 6x^3 - 22x^2 - 61x - 22,$
$f(1) \approx -97,$
$f(-1) \approx 13.$

Zadanie 3.5.

Wyznaczyć funkcję liniową, która w sensie metody najmniejszych kwadratów aproksymuje dane z tabeli 3.14.

Tabela 3.14. Dane do zadania 3.5

x_i	1	2	3	6	9
f_i	-4	-6	-8	-14	-20

Odp. $y(x) = -2x - 2.$

Zadanie 3.6.

Wyznaczyć funkcję postaci $y=a/x+b$, która w sensie metody najmniejszych kwadratów aproksymuje dane z tabeli 3.15.

Tabela 3.15. Dane do zadania 3.6

x_i	1	1/5	1/9	1/13	1/15
f_i	2	-6	-14	-22	-26

Odp. $y(x) = \frac{-2}{x} + 4.$

Zadanie 3.7.

Sprawdzić, czy ciąg wielomianów: $P_0 = 1, P_1 = x - 1, P_2 = 3x^2 + 4$ jest ciągiem wielomianów ortogonalnych na przedziale [-1,1] z funkcją wagową $p(x) = x + 1.$

Odp. Wielomiany te nie stanowią ciągu ortogonalnego, ponieważ $(P_0, P_1) \neq 0.$

4. Metody rozwiązywania układów równań liniowych

4.1. Wstęp

W rozdziale tym przedstawimy metody skończone oraz iteracyjne [1, 4, 5, 8, 9, 10] rozwiązywania układów równań liniowych postaci:

$$\mathbf{Ax} = \mathbf{b}, \tag{4.1}$$

gdzie:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$ jest macierzą n x n,

$$\mathbf{b} = \begin{bmatrix} b_1 \\ \dots \\ b_2 \\ b_n \end{bmatrix}$$ danym wektorem,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \dots \\ x_2 \\ x_n \end{bmatrix}$$ szukanym rozwiązaniem układu równań liniowych (4.1).

4.2. Metody skończone

4.2.1. Eliminacja Gaussa

Metoda eliminacji Gaussa jest najczęściej stosowaną skończoną metodą numerycznego rozwiązywania nieosobliwych układów równań liniowych.

Układ n równań liniowych (4.1) można zapisać w postaci macierzowej:

$$\mathbf{A}^{(1)}\mathbf{x} = \mathbf{b}^{(1)} \tag{4.2}$$

lub bezpośrednio:

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + ... + a_{1n}^{(1)}x_n = b_1^{(1)} \\ a_{21}^{(1)}x_1 + a_{22}^{(1)}x_2 + ... + a_{2n}^{(1)}x_n = b_2^{(1)} \\ ..... \\ a_{n1}^{(1)}x_1 + a_{n2}^{(1)}x_2 + ... + a_{nn}^{(1)}x_n = b_n^{(1)} \end{cases}. \tag{4.3}$$

Indeks w nawiasie oznacza tu kolejny krok realizowamy w wyniku działania metody eliminacji Gaussa.

W metodzie wyróżnia się dwa etapy:

- postępowanie proste,
- postępowanie odwrotne.

Pierwszy etap eliminacji (eliminacja w przód, czyli tzw. ***postępowanie proste***) sprowadza układ do postaci górnie trójkątnej.

Pierwszy krok metody polega na odjęciu od i-tego wiersza układu (4.3) (i=2,3,...,n), wiersza pierwszego pomnożonego odpowiednio przez:

$$m_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}.$$

Po wykonaniu pierwszego kroku eliminacji otrzymujemy układ:

$$\mathbf{A}^{(2)}\mathbf{x} = \mathbf{b}^{(2)} \tag{4.4}$$

lub:

$$\begin{cases} a_{11}^{(2)}x_1 + a_{12}^{(2)}x_2 + ... + a_{1n}^{(2)}x_n = b_1^{(2)} \\ \qquad a_{22}^{(2)}x_2 + ... + a_{2n}^{(2)}x_n = b_2^{(2)} \\ \qquad ..... \\ \qquad a_{n2}^{(2)}x_2 + ... + a_{nn}^{(2)}x_n = b_n^{(2)} \end{cases}. \tag{4.5}$$

Wyeliminowaliśmy w ten sposób niewiadomą x_1 z równań leżących w wierszach o numerach i=2, 3,..., n.

Podobnie, w kroku drugim, eliminujemy zmienną x_2 z równań leżących w wierszach 3, 4,..., n, odejmując od i-tego wiersza (i=3, 4,..., n), wiersz drugi pomnożony przez m_{i2}, gdzie kolejne mnożniki wyznaczane są ze wzoru:

$$m_{ij} = a_{ij}^{(k)} / a_{jj}^{(k)} \qquad k = 1,...,n,\ i=2,...,n,\ j = 1,...,n-1. \tag{4.6}$$

Postępując w ten sposób otrzymamy przekształcony układ równań:

$$\mathbf{A}^{(3)}\mathbf{x} = \mathbf{b}^{(3)}.$$

Po wykonaniu n-1 eliminacji uzyskamy trójkątny układ równań:

$$\mathbf{A}^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$$

czyli:

$$\begin{cases} a_{11}^{(n)}x_1 + a_{12}^{(n)}x_2 + ... + a_{1n}^{(n)}x_n = b_1^{(n)} \\ \qquad\quad a_{22}^{(n)}x_2 + ... + a_{2n}^{(n)}x_n = b_2^{(n)} \\ \qquad\quad . \qquad\quad ..... \\ \qquad\qquad\qquad\quad a_{nn}^{(n)}x_n = b_n^{(n)} \end{cases}. \tag{4.7}$$

W drugim etapie rozwiązania (***postępowanie odwrotne*** lub ***postępowanie wsteczne***) w celu znalezienia rozwiązania układu równań, korzysta się z uzyskanej (w wyniku postępowania prostego) macierzy trójkątnej górnej (4.6) i wzorów rekurencyjnych :

$$x_n = \frac{b_n^{(n)}}{a_{nn}^{(n)}},\ x_i = \frac{b_i^{(n)} - \sum_{k=i+1}^{n} a_{i,k}^{(n)} x_k}{a_{ii}^{(n)}},\ i = n-1,...,1. \tag{4.8}$$

Metoda Gaussa zapewnia uzyskanie wyników z niewielkim błędem, jeżeli tylko wartości współczynników ostatecznie zredukowanego układu równań leżących na głównej przekątnej nie są bliskie zeru. Gdyby moduł któregoś z dzielników był mały w porównaniu z innymi współczynnikami, to mógłby powstać znaczny błąd numeryczny. Istnienie zera na przekątnej wykluczałoby rozwiązanie układu (patrz przykład 4.3). Aby tego uniknąć, stosuje się jedną z metod ***wyboru elementu głównego*** [1]:

- wybór częściowy elementu głównego,
- wybór pełny elementu głównego.

Wybór częściowy elementu głównego polega na tym, że w *k*-tym kroku eliminacji wybiera się ten element *k*-tej kolumny albo *k*-tego wiersza macierzy, który ma największy moduł tzn.: $a_{rk}^{(k)} = \max_{k \le i \le n} \left| a_{ik}^{(k)} \right|$, a następnie dokonuje się przestawienia wierszy o numerze *k* z wierszem o numerze *r* lub $a_{kr}^{(k)} = \max_{k \le i \le n} \left| a_{ki}^{(k)} \right|$ oraz przestawienia kolumny numer *k* z kolumną numer *r*. Należy przy tym pamiętać o jednoczesnej zmianie kolejności zmiennych w wektorze wynikowym.

Wybór pełny elementu głównego polega na tym, że w *k*-tym kroku eliminacji wybiera się największy co do modułu element $a_{rs}^{(k)}, k \le r \le n,\ k \le s \le n$ tzn.: $a_{rs}^{(k)} = \max_{k \le i,j \le n} \left| a_{ij}^{(k)} \right|$ a następnie dokonuje się przestawienia wierszy *k* i *r* oraz kolumny *k* i *s*.

W praktyce wybór częściowy elementu głównego zwykle wystarcza i ze względu na znacznie większy koszt poszukiwania, rzadko stosuje się wybór pełny. Z drugiej strony jednak ***metoda Gaussa jest numerycznie poprawna tylko w przypadku pełnego wyboru elementu głównego***.

Łączna liczba operacji w metodzie eliminacji Gaussa wynosi około $(1/3)n^3 + (1/2)n^2$. Ponieważ samo rozwiązanie końcowego układu trójkątnego wymaga $(1/2)n^2$ operacji, toteż dla dużego *n* najkosztowniejszą częścią obliczeń jest redukcja do postaci trójkątnej.

Przykłady 4.1-4.3 przedstawiają różne układy demonstrujące przypadek, kiedy częściowy wybór elementu głównego nie jest wystarczający.

Przykład 4.1.

Rozwiązać metodą eliminacji Gaussa z wyborem elementu podstawowego w kolumnie układ równań:

$$\begin{cases} -x_1 - 5x_2 + 0.5x_3 + 5.5x_4 = 9.5 \\ -2x_1 \qquad\quad - x_3 + 3x_4 = -3 \\ -1.5x_1 - 1.25x_2 + 0.5x_3 - 0.75x_4 = -1.5 \\ \qquad -1.25x_2 + 0.5x_3 + 5.5x_4 = 9.5 \end{cases}$$

Postępowanie proste

Na początek tworzymy macierz współczynników 4*x*4 powiększoną o wektor wyrazów wolnych tego układu:

$$A^{(1)} = \begin{bmatrix} -1 & -5 & 0.5 & 5.5 & 9.5 \\ -2 & 0 & -1 & 3 & -3 \\ -1.5 & -1.25 & 0.5 & -0.75 & -1.5 \\ 0 & -1.25 & 0.5 & 5.5 & 9.5 \end{bmatrix}.$$

Krok pierwszy eliminacji Gaussa

Chcemy zastosować wybór elementu podstawowego w kolumnie więc w pierwszym kroku należy wybrać element największy co do modułu z kolumny pierwszej – u nas jest to element a_{21}=*-2*, który stoi w wierszu drugim, dlatego zamieniamy wiersz 1 z wierszem 2. Macierz ma postać:

$$A^{(1)} = \begin{bmatrix} -2 & 0 & -1 & 3 & -3 \\ -1 & -5 & 0.5 & 5.5 & 9.5 \\ -1.5 & -1.25 & 0.5 & -0.75 & -1.5 \\ 0 & -1.25 & 0.5 & 5.5 & 9.5 \end{bmatrix}.$$

Następnie wyznaczamy mnożniki m_{ij} ze wzoru (4.7a) dla *j=1, i=2,3,4* oraz podajemy od razu wzory na obliczenie nowych wierszy macierzy współczynników z wykorzystaniem policzonych mnożników:

$$m_{21} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{-1}{-2} = \frac{1}{2} \rightarrow w_2^{(2)} = w_2^{(1)} - m_{21} w_1^{(1)} = w_2^{(1)} - \frac{1}{2} w_1^{(1)},$$

$$m_{31} = \frac{a_{31}^{(1)}}{a_{11}^{(1)}} = \frac{-1.5}{-2} = \frac{3}{4} \rightarrow w_3^{(2)} = w_3^{(1)} - m_{31} w_1^{(1)} = w_3^{(1)} - \frac{3}{4} w_1^{(1)},$$

$$m_{41} = \frac{a_{41}^{(1)}}{a_{11}^{(1)}} = \frac{0}{-2} = 0 \rightarrow w_4^{(2)} = w_4^{(1)} - m_{41} w_1^{(1)} = w_4^{(1)} - 0 w_1^{(1)} = w_4^{(1)}.$$

Wykonujemy pierwszy krok według powyższych wzorów (wiersz pierwszy oczywiście nie ulega zmianie):

$$A^{(2)} = \begin{bmatrix} -2 & 0 & -1 & 3 & -3 \\ 0 & -5 & 1 & 4 & 11 \\ 0 & -1.25 & 1.25 & -3 & 0.75 \\ 0 & -1.25 & -0.25 & 2 & 2.75 \end{bmatrix}.$$

Krok drugi eliminacji Gaussa

Stosujemy wybór elementu podstawowego w kolumnie więc w drugim kroku należy wybrać element największy co do modułu z kolumny drugiej i wierszy od drugiego do czwartego (bez elementu tej kolumny stojącego w wierszu pierwszym) – u nas jest to element a_{22}=*-5* stojący w wierszu drugim – wobec czego nie dokonujemy żadnej zamiany.

Analogiczne obliczenia, jak w kroku pierwszym, wykonujemy w celu otrzymania mnożników dla *j=2* oraz *i=3,4* i przekształcenia wierszy:

$$m_{32} = \frac{a_{32}^{(2)}}{a_{22}^{(2)}} = \frac{-1.25}{-5} = \frac{1}{4} \;\rightarrow\; w_3^{(3)} = w_3^{(2)} - m_{32} w_2^{(2)} = w_3^{(2)} - \frac{1}{4} w_2^{(2)},$$

$$m_{42} = \frac{a_{42}^{(2)}}{a_{22}^{(2)}} = \frac{-1.25}{-5} = \frac{1}{4} \;\rightarrow\; w_4^{(3)} = w_4^{(2)} - m_{42} w_2^{(2)} = w_4^{(2)} - \frac{1}{4} w_2^{(2)}.$$

W wyniku przekształceń opisanych powyższymi wzorami otrzymujemy macierz:

$$A^{(3)} = \begin{bmatrix} -2 & 0 & -1 & 3 & -3 \\ 0 & -5 & 1 & 4 & 11 \\ 0 & 0 & 1 & -4 & -2 \\ 0 & 0 & -0.5 & 1 & 0 \end{bmatrix}.$$

Krok trzeci eliminacji Gaussa

Stosując wybór elementu podstawowego w kolumnie, w trzecim kroku należy wybrać element największy co do modułu z kolumny trzeciej i wierszy od trzeciego do czwartego (bez elementów tej kolumny stojących w wierszu pierwszym i drugim) – w naszym przykładzie jest to element $a_{33}=1$ stojący w wierszu trzecim, wobec czego ponownie nie trzeba dokonywać żadnej zamiany a mnożnik wyznaczamy ze wzoru:

$$m_{43} = \frac{a_{43}^{(3)}}{a_{33}^{(3)}} = \frac{-0.5}{1} = -\frac{1}{2} \;\rightarrow\; w_4^{(4)} = w_4^{(3)} - m_{43} w_3^{(3)} = w_4^{(3)} + \frac{1}{2} w_3^{(3)}$$

W wyniku przekształcenia otrzymujemy macierz:

$$A^{(4)} = \begin{bmatrix} -2 & 0 & -1 & 3 & -3 \\ 0 & -5 & 1 & 4 & 11 \\ 0 & 0 & 1 & -4 & -2 \\ 0 & 0 & 0 & -1 & -1 \end{bmatrix},$$

która jest efektem końcowym eliminacji prostej.

Postępowanie odwrotne

Po wykonaniu trzech kroków eliminacji Gaussa otrzymaliśmy macierz górną trójkątną, dla której należy przeprowadzić postępowanie odwrotne zmierzające do rozwiązania tak otrzymanego układu równań:

$$\begin{cases} -2x_1 + 0x_2 - x_3 + 3x_4 = -3 \\ -5x_2 + x_3 + 4x_4 = 11 \\ x_3 - 4x_4 = -2 \\ -x_4 = -1 \end{cases}.$$

Ten etap jest już bardzo prosty: z równania ostatniego wyliczamy zmienną x_4, z trzeciego – x_3 itd., aż dojdziemy do równania pierwszego.

W ten sposób uzyskujemy rozwiązanie:

- z równania czwartego $x_4 = 1$;
- z równania trzeciego $x_3 = -2 + 4x_4 = 2$;
- z równania drugiego $x_2 = \frac{11-4x_4-x_3}{-5} = \frac{11-4-2}{-5} = -1$;
- z równania pierwszego $x_1 = \frac{-3-3x_4+x_3}{-2} = \frac{-3-3+2}{-2} = 2$.

Zamiana wierszy nie pociąga za sobą zamiany zmiennych w wektorze rozwiązań. Zatem rozwiązaniem podanego układu równań jest wektor:
$x = [2 \quad -1 \quad 2 \quad 1]^T$.

Przykład 4.2.

Metodą eliminacji Gaussa z wyborem elementu podstawowego w wierszu rozwiązać układ równań:

$$\begin{cases} -2x_1 - 4x_2 + 5x_3 - 2x_4 = -5 \\ 6.5x_1 + 2x_2 + 1.25x_3 + 3.5x_4 = 1.75 \\ 1.75x_1 + 7.25x_2 - 8.75x_3 + 5.5x_4 = 5 \\ 3.25x_1 - 2.75x_2 - 3.75x_3 + x_4 = 6 \end{cases}.$$

Początkowo macierz układu jest postaci:

$$A^{(1)} = \begin{bmatrix} -2 & -4 & 5 & -2 & -5 \\ 6.5 & 2 & 1.25 & 3.5 & 1.75 \\ 1.75 & 7.25 & -8.75 & 5.5 & 5 \\ 3.25 & -2.75 & -3.75 & 1 & 6 \end{bmatrix}.$$

Postępowanie proste

Krok pierwszy

Chcemy zastosować wybór elementu podstawowego w wierszu - w pierwszym kroku należy wybrać element największy co do modułu z wiersza pierwszego. W naszym przykładzie jest to element $a_{13}=5$, który stoi w kolumnie trzeciej, dlatego zamieniamy kolumnę 1 z kolumną 3.

Macierz po zmianach ma postać:

$$A^{(1)} = \begin{bmatrix} 5 & -4 & -2 & -2 & -5 \\ 1.25 & 2 & 6.5 & 3.5 & 1.75 \\ -8.75 & 7.25 & 1.75 & 5.5 & 5 \\ -3.75 & -2.75 & 3.25 & 1 & 6 \end{bmatrix}.$$

Następnie wyznaczamy mnożniki:

$$m_{21} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{1.25}{5} = \frac{1}{4} \rightarrow w_2^{(2)} = w_2^{(1)} - m_{21}w_1^{(1)} = w_2^{(1)} - \frac{1}{4}w_1^{(1)},$$

$$m_{31} = \frac{a_{31}^{(1)}}{a_{11}^{(1)}} = \frac{-8.75}{5} = \frac{-7}{4} \rightarrow w_3^{(2)} = w_3^{(1)} - m_{31}w_1^{(1)} = w_3^{(1)} + \frac{7}{4}w_1^{(1)},$$

$$m_{41} = \frac{a_{41}^{(1)}}{a_{11}^{(1)}} = \frac{-3.75}{5} = \frac{-3}{4} \rightarrow w_4^{(2)} = w_4^{(1)} - m_{41}w_1^{(1)} = w_4^{(1)} + \frac{3}{4}w_1^{(1)}$$

i po obliczeniach uzyskujemy macierz:

$$A^{(2)} = \begin{bmatrix} 5 & -4 & -2 & -2 & -5 \\ 0 & 3 & 7 & 4 & 3 \\ 0 & 0.25 & -1.75 & 2 & -3.75 \\ 0 & -5.75 & 1.75 & -0.5 & 2.25 \end{bmatrix}.$$

Krok drugi

Element największy co do modułu z wiersza drugiego jest to element $a_{23}=7$, który stoi w kolumnie trzeciej, dlatego zamieniamy kolumnę 2 z kolumną 3. Macierzpo zamianie ma postać:

$$A^{(2)} = \begin{bmatrix} 5 & -2 & -4 & -2 & -5 \\ 0 & 7 & 3 & 4 & 3 \\ 0 & -1.75 & 0.25 & 2 & -3.75 \\ 0 & 1.75 & -5.75 & -0.5 & 2.25 \end{bmatrix}.$$

Wyznaczamy kolejne mnożniki:

$$m_{32} = \frac{a_{32}^{(2)}}{a_{22}^{(2)}} = \frac{-1.75}{7} = -\frac{1}{4} \rightarrow w_3^{(3)} = w_3^{(2)} - m_{32}w_2^{(2)} = w_3^{(2)} + \frac{1}{4}w_2^{(2)},$$

$$m_{42} = \frac{a_{42}^{(2)}}{a_{22}^{(2)}} = \frac{1.75}{7} = \frac{1}{4} \rightarrow w_4^{(3)} = w_4^{(2)} - m_{42}w_2^{(2)} = w_4^{(2)} - \frac{1}{4}w_2^{(2)}$$

i uzyskujemy macierz:

$$A^{(3)} = \begin{bmatrix} 5 & -2 & -4 & -2 & -5 \\ 0 & 7 & 3 & 4 & 3 \\ 0 & 0 & 1 & 3 & -3 \\ 0 & 0 & -6.5 & -1.5 & 1.5 \end{bmatrix}.$$

Krok trzeci

Element największy co do modułu z wiersza trzeciego jest to element $a_{34}=3$, który stoi w kolumnie czwartej, dlatego zamieniamy kolumnę 3 z kolumną 4.

Macierz po zamianie ma postać:

$$A^{(3)} = \begin{bmatrix} 5 & -2 & -2 & -4 & -5 \\ 0 & 7 & 4 & 3 & 3 \\ 0 & 0 & 3 & 1 & -3 \\ 0 & 0 & -1.5 & -6.5 & 1.5 \end{bmatrix}.$$

Wyznaczamy mnożnik:

$$m_{43} = \frac{a_{43}^{(3)}}{a_{33}^{(3)}} = \frac{-1.5}{3} = -\frac{1}{2} \;\rightarrow\; w_4^{(4)} = w_4^{(3)} - m_{43} w_3^{(3)} = w_4^{(3)} + \frac{1}{2} w_3^{(3)}$$

i ostatecznie otrzymujemy macierz:

$$A^{(4)} = \begin{bmatrix} -5 & -2 & -2 & -4 & -5 \\ 0 & 7 & 4 & 3 & 3 \\ 0 & 0 & 3 & 1 & -3 \\ 0 & 0 & 0 & -6 & 0 \end{bmatrix}.$$

Postępowanie odwrotne

Zamiana kolumn pociąga za sobą zamiany w wektorze rozwiązań:

- wyjściowa sytuacja dla numerów zmiennych to: [1, 2, 3, 4];
- pierwsza zamiana dotyczyła kolumn 1 i 3: [3, 2, 1, 4];
- druga zamiana dotyczyła kolumn 2 i 3: [3, 1, 2, 4];
- trzecia zamiana dotyczyła kolumn 3 i 4: [3, 1, 4, 2].

Dla takiej kolejności w wektorze zmiennych wykonujemy postępowanie odwrotne:

$$\begin{cases} -5x_3 - 2x_1 - 2x_4 - 4x_2 = -5 \\ 7x_1 + 4x_4 + 3x_2 = 3 \\ 3x_4 + x_2 = -3 \\ -6x_2 = 0 \end{cases}.$$

Szukanym rozwiązaniem jest:

$x_2 = 0, x_4 = -1, x_1 = 1, x_3 = -1.$

Zatem rozwiązaniem podanego układu równań jest wektor:

$x = [1 \quad 0 \quad -1 \quad -1]^T.$

Przykład 4.3.

Metodą eliminacji Gaussa bez wyboru elementu podstawowego rozwiązać podany układ równań:

$$\begin{cases} -x_1 - 2x_2 + 4x_3 + 2x_4 = -4 \\ -4x_3 + 2x_4 = 6 \\ -0.5x_1 - x_2 + 2x_3 + 5x_4 = 2 \\ -0.5x_1 - x_2 - 2x_3 + 8x_4 = 9 \end{cases}.$$

Macierz wyjściowa ma postać:

$$A^{(1)} = \begin{bmatrix} -1 & -2 & 4 & 2 & -4 \\ 0 & 0 & -4 & 2 & 6 \\ -0.5 & -1 & 2 & 5 & 2 \\ -0.5 & -1 & -2 & 8 & 9 \end{bmatrix}.$$

Krok pierwszy

Dokonujemy obliczeń:

$$m_{21} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{0}{-1} = 0 \;\rightarrow\; w_2^{(2)} = w_2^{(1)} - m_{21} w_1^{(1)} = w_2^{(1)},$$

$$m_{31} = \frac{a_{31}^{(1)}}{a_{11}^{(1)}} = \frac{-0.5}{-1} = \frac{1}{2} \;\rightarrow\; w_3^{(2)} = w_3^{(1)} - m_{31} w_1^{(1)} = w_3^{(1)} - \frac{1}{2} w_1^{(1)},$$

$$m_{41} = \frac{a_{41}^{(1)}}{a_{11}^{(1)}} = \frac{-0.5}{-1} = \frac{1}{2} \rightarrow w_4^{(2)} = w_4^{(1)} - m_{41} w_1^{(1)} = w_4^{(1)} - \frac{1}{2} w_1^{(1)}$$

i otrzymujemy macierz:

$$A^{(1)} = \begin{bmatrix} -1 & -2 & 4 & 2 & -4 \\ 0 & 0 & -4 & 2 & 6 \\ 0 & 0 & 2 & 5 & 2 \\ 0 & 0 & -4 & 7 & 11 \end{bmatrix}.$$

Po pierwszym kroku metody eliminacji Gaussa element a_{22} jest równy zero, co powoduje niemożliwość prowadzenia dalszych obliczeń. Metodę eliminacji Gaussa bez wyboru elementu podstawowego należy w tym miejscu zakończyć. Nie uzyskaliśmy rozwiązania układu. Nie oznacza to jednak, że macierz jest osobliwa.

Ten sam układ można spróbować rozwiązać metodą eliminacji Gaussa z wyborem elementu podstawowego. W tym przypadku również ta metoda nie da rozwiązania, co oznacza, że macierz jest osobliwa i układ jest sprzeczny.

4.2.2. Eliminacja Gaussa-Jordana

Metoda eliminacji Gaussa-Jordana jest modyfikacją metody eliminacji Gaussa. Układ *n* równań liniowych (4.2), tzn.

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + ... + a_{1n}^{(1)}x_n = b_1^{(1)} \\ a_{21}^{(1)}x_1 + a_{22}^{(1)}x_2 + ... + a_{2n}^{(1)}x_n = b_2^{(1)} \\ ..... \\ a_{n1}^{(1)}x_1 + a_{n2}^{(1)}x_2 + ... + a_{nn}^{(1)}x_n = b_n^{(1)} \end{cases}, \tag{4.9}$$

drogą elementarnych transformacji przekształcany jest tak, że zarówno powyżej, jak i poniżej głównej przekątnej występują współczynniki zerowe.

Układ równań (4.9) przekształcamy zgodnie z poniższym algorytmem. Pierwsze równanie dzielimy obustronnie przez $a_{11}^{(1)}$ a następnie od *i*-tego wiersza (*i*=2,3...,*n*) odejmujemy wiersz pierwszy pomnożony przez $a_{i1}^{(1)}$, otrzymując układ równań:

$$\mathbf{A}^{(2)}\mathbf{x} = \mathbf{b}^{(2)}, \tag{4.10}$$

postaci:

$$\begin{cases} x_1 + a_{12}^{(2)}x_2 + ... + a_{1n}^{(2)}x_n = b_1^{(2)} \\ \quad a_{22}^{(2)}x_2 + ... + a_{2n}^{(2)}x_n = b_2^{(2)} \\ \quad ..... \\ \quad a_{n2}^{(2)}x_2 + ... + a_{nn}^{(2)}x_n = b_n^{(2)} \end{cases}. \tag{4.11}$$

Następnie drugie równanie dzielimy obustronnie przez $a_{22}^{(2)}$ i od *i*-tego wiersza (*i*=1,3,4,...,*n*), odejmujemy wiersz drugi pomnożony przez $a_{i2}^{(2)}$, otrzymując układ:

$$\mathbf{A}^{(3)}\mathbf{x} = \mathbf{b}^{(3)}, \tag{4.12}$$

postaci:

$$\begin{cases} x_1 + a_{13}^{(3)} x_3 + \cdots + a_{1n}^{(3)} x_n = b_1^{(3)} \\ x_2 + a_{23}^{(3)} x_3 + \cdots + a_{2n}^{(3)} x_n = b_2^{(3)} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n3}^{(3)} x_3 + \cdots + a_{nn}^{(3)} x_n = b_n^{(3)} \end{cases} \tag{4.13}$$

Po n-1 eliminacjach otrzymuje się układ o macierzy diagonalnej:

$$\begin{cases} x_1 = b_1^{(n)} \\ x_2 = b_2^{(n)} \\ \cdots\cdots\cdots\cdots \\ x_n = b_n^{(n)} \end{cases} \tag{4.14}$$

a zatem już gotowe rozwiązanie.

Metoda ta zwana metodą ***eliminacji Jordana*** lub ***metodą eliminacji zupełnej*** wymaga wykonania około półtora raza więcej działań, niż w przypadku metody eliminacji Gaussa. Liczba zajętych komórek pamięci jest taka sama. Aby nie wystąpiło dzielenie przez zerowy element $a_{ii}^{(i-1)}$, należy zastosować odpowiedni wybór elementu podstawowego. Zaletą redukcji Gaussa-Jordana jest możliwość rozwiązywania układu równań „obciętego" do k początkowych równań i k niewiadomych (w przypadku, gdy dokonujemy redukcji bez wyboru elementów podstawowych). Metoda ta umożliwia ponadto rozwiązywanie układu równań przy oszczędniejszym wykorzystaniu pamięci operacyjnej, co pozwala na rozwiązywanie układów o większej liczbie zmiennych. Istotną wadą jest tu jednak duży nakład obliczeń, który nie ulega zmniejszeniu podczas ponownego rozwiązywania układu równań przy zmienionej prawej stronie układu wektora **b** (4.9).

4.2.3. Rozkład LU

W wielu zagadnieniach numerycznych dotyczących macierzy kwadratowych **A**, poszukujemy takich macierzy trójkątnych **L** i **U** (zobacz wzór 2.10), by: **A**=**LU** (jest to tzw. ***rozkład trójkątny***, albo ***rozkład LU***). Metoda eliminacji Gaussa umożliwia wyznaczenie takiego rozkładu.

Układ:

Ax=b

jest równoważny układowi:

LUx=b,

który rozpada się na dwa układy trójkątne:

Ly=b

i

Ux=y.

Znając czynniki **L** i **U** można rozwiązać układ **Ax=b** kosztem $2n^2/2 = n^2$ operacji (eliminacja Gaussa wymaga natomiast $n^3/3$ operacji).

Niech **A** będzie macierzą wymiaru *n* x *n* i niech $\mathbf{A}_k$ oznacza macierz *k* x *k*, utworzoną z elementów początkowych *k* wierszy i kolumn z **A**. Jeśli $\det(\mathbf{A}_k) \neq 0$ (*k*=1,2,...,*n*-1), to istnieje jedyny rozkład **A=LU** na czynniki takie, że:

- macierz $\mathbf{L} = \lfloor l_{ij} \rfloor$ jest macierzą dolnie trójkątną i ma elementy $l_{ii} = 1$ (*i*=1,...,*n*)
- macierz $\mathbf{U} = \lceil u_{ij} \rceil$ jest macierzą górnie trójkątną.

Zauważmy ponadto, że macierz **U** jest końcową macierzą trójkątną otrzymaną za pomocą eliminacji Gaussa. Aby otrzymać macierz **L**, należy zachować mnożniki $m_{ik} = a_{ik}^{(k)} / a_{kk}^{(k)}$, które określa się w eliminacji Gaussa (4.7a) tak, że $a_{ik}^{(k+1)}$ staje się zerem (można więc na miejscu $a_{ik}^{(k)}$ wpisać m_{ik} tak jak to pokazano na schemacie 4.15). Wtedy $l_{ik} = m_{ik}$. Nie trzeba też pamiętać jedynek z głównej przekątnej macierzy **L**, dlatego nie jest potrzebna dodatkowa pamięć.

Efekt rozkładu **LU** macierzy **A** można zobrazować schematem:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1,n-1} & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2,n-1} & a_{2n} \\ \dots & & \dots & & \dots \\ \dots & & & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{n,n-1} & a_{nn} \end{bmatrix} \Rightarrow \begin{bmatrix} \underline{u_{11}} & u_{12} & \dots & u_{1,n-1} & u_{1n} \\ m_{21} & \underline{u_{22}} & & u_{2,n-1} & u_{2n} \\ \dots & & \dots & & \dots \\ \dots & & & \dots & \dots \\ m_{n1} & m_{n2} & \dots & m_{n,n-1} & \underline{u_{nn}} \end{bmatrix}. \quad (4.15)$$

Bezpośrednie wzory na elementy macierzy **U** oraz **L** przedstawiają się następująco dla *i=1,2,..,n* oraz *j=1,2,..,n*:

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj} \qquad dla \quad i \le j, \tag{4.16}$$

$$u_{ij} = 0 \qquad dla \quad i > j,$$

$$l_{ji} = \frac{\left(a_{ji} - \sum_{k=1}^{i-1} l_{jk}u_{ki}\right)}{u_{ii}} \qquad dla \quad i < j,$$

$$l_{ij} = 1 \qquad dla \quad i = j, \tag{4.17}$$

$$l_{ij} = 0 \qquad dla \quad i > j.$$

Wzory te należy stosować naprzemiennie dla obu macierzy, tzn. na początku obliczamy pierwszy wiersz macierzy **U**: u_{11}, u_{12}, ..., u_{1n} i kolejno pierwszą kolumnę macierzy L: l_{21}, l_{31}, ..., l_{n1}, a następnie drugi wiersz macierzy **U** i drugą kolumnę macierzy **L** itd.
Znajomość rozkładu **LU** macierzy **A** jest niezbędna do obliczenia wyznacznika macierzy i znalezienia macierzy odwrotnej $\mathbf{A}^{-1}$.

Przykład 4.4.

Metodą rozkładu **LU** rozwiązać układ równań:

$$\begin{cases} 3x_1 - 4x_2 + 4x_3 - 4x_4 = -9 \\ 1.5x_1 - x_2 + 2x_3 - 2x_4 = -3.5 \\ 1.5x_1 - 0.5x_2 - 3x_4 = -2 \\ 4.5x_1 - 5.5\,x_2 + 4x_3 - 9x_4 = -14 \end{cases}.$$

Macierz wyjściowa ma postać:

$$A = \begin{bmatrix} 3 & -4 & 4 & -4 \\ 1.5 & -1 & 2 & -2 \\ 1.5 & -0.5 & 0 & -3 \\ 4.5 & -5.5 & 4 & -9 \end{bmatrix}.$$

Dla $i = 1$ obliczamy pierwszy wiersz macierzy **U**:

$u_{11} = a_{11} = 3,$
$u_{12} = a_{12} = -4,$
$u_{13} = a_{13} = 4,$
$u_{14} = a_{14} = -4.$

oraz pierwszą kolumnę macierzy **L**:

$$m_{21} = \frac{a_{21}}{u_{11}} = \frac{1.5}{3} = 0.5,$$
$$m_{31} = \frac{a_{31}}{u_{11}} = \frac{1.5}{3} = 0.5,$$
$$m_{41} = \frac{a_{41}}{u_{11}} = \frac{4.5}{3} = 1.5.$$

Dla $i = 2$ obliczamy drugi wiersz macierzy **U** (począwszy od głownej przekątnej) oraz drugą kolumnę macierzy **L** (poniżej głównej przekątnej):

$$u_{22} = a_{22} - m_{21} \cdot u_{12} = -1 - 0.5 \cdot (-4) = 1,$$
$$u_{23} = a_{23} - m_{21} \cdot u_{13} = 2 - 0.5 \cdot 4 = 0,$$
$$u_{24} = a_{24} - m_{21} \cdot u_{14} = -2 - 0.5 \cdot (-4) = 0,$$
$$m_{32} = \frac{a_{32} - m_{31} \cdot u_{12}}{u_{22}} = \frac{-0.5 - 0.5 \cdot (-4)}{1} = 1.5,$$
$$m_{42} = \frac{a_{42} - m_{41} \cdot u_{12}}{u_{22}} = \frac{-5.5 - 1.5 \cdot (-4)}{1} = 0.5.$$

Kolejne elementy macierzy **U** oraz **L** obliczamy ze wzorów:

$$u_{33} = a_{33} - m_{31} \cdot u_{13} - m_{32} \cdot u_{23} = 0 - 0.5 \cdot 4 - 1.5 \cdot 0 = -2,$$
$$u_{34} = a_{34} - m_{31} \cdot u_{14} - m_{32} \cdot u_{24} = -3 - 0.5 \cdot (-4) - 1.5 \cdot 0 = -1$$
$$m_{43} = \frac{a_{43} - m_{41} \cdot u_{13} - m_{42} \cdot u_{23}}{u_{33}} = \frac{4 - 1.5 \cdot 4 - 0.5 \cdot 0}{-2} = 1.$$

Ostatecznie macierze **L** i **U** są postaci:

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 0 \\ 0.5 & 1.5 & 1 & 0 \\ 1.5 & 0.5 & 1 & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 3 & -4 & 4 & -4 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -2 & -1 \\ 0 & 0 & 0 & -2 \end{bmatrix}.$$

Kolejny etap obliczeń to rozwiązanie dwóch układów równań:

Ly = **b** i **Ux** = **y**.

Rozwiązując pierwszy układ równań ***Ly* = *b*** otrzymujemy:

$$\begin{cases} y_1 = -9 \\ 0.5y_1 + y_2 = -3.5 \\ 0.5y_1 + 1.5y_2 + y_3 = -2 \\ 1.5y_1 + 0.5\,y_2 + y_3 + y_4 = -14 \end{cases} \Rightarrow \quad y = \begin{bmatrix} -9 \\ 1 \\ 1 \\ -2 \end{bmatrix}.$$

Rozwiązując drugi układ $Ux = y$ mamy:

$$\begin{cases} 3x_1 - 4x_2 + 4x_3 - 4x_4 = -9 \\ x_2 = -9 \\ -2x_3 - x_4 = 1 \\ x_4 = -2 \end{cases} \Rightarrow x = \begin{bmatrix} 1 \\ 1 \\ -1 \\ 1 \end{bmatrix}.$$

Ostatecznie rozwiązaniem podanego układu równań jest wektor:

$x = [1 \quad 1 \quad -1 \quad 1]^T$.

4.2.4. Rozkład Choleskiego

Przy założeniu, że macierz **A** o wymiarze *n* x *n* jest macierzą symetryczną dodatnio określoną, dekompozycja **LU** tej macierzy ma dużo prostszą postać i nazywa się ją ***rozkładem Choleskiego***.

Dla takiej macierzy wszystkie minory główne są dodatnie i rozkłada się ona jednoznacznie na czynniki trójkątne:

$$\boldsymbol{A = L \cdot L^T} \tag{4.18}$$

gdzie macierz **L** ma postać:

$$L = \begin{bmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{bmatrix}.$$

Elementy macierzy **L** obliczamy według wzorów:

$$l_{ss} = \sqrt{a_{ss} - \sum_{k=1}^{s-1} (l_{sk})^2} \qquad dla \quad s = 1,2,\dots,n \tag{4.19a}$$

$$l_{is} = \frac{(a_{is} - \sum_{k=1}^{s-1} l_{ik} l_{sk})}{u_{ii}} \qquad dla \quad i = s+1,\dots,n \tag{4.19b}$$

Macierz **L** obliczamy kolumnami. W rozkładzie **LU** należało policzyć n^2 elementów, natomiast w rozkładzie Choleskiego wystarczy policzyć $n(n+1)/2$. Elementy macierzy $\mathbf{L}^{\mathrm{T}}$ mogą być przechowywane w macierzy **L**, więc niepotrzebna jest dodatkowa pamięć.

Rozwiązanie układu równań **Ax**=**b** jest równoważne układowi $\mathbf{LL^Tx}$=**b**, który rozpada się na dwa układy trójkątne: **Ly**=**b** i $\mathbf{L^Tx}$=**y**. Koszt rozwiązania (liczba mnożeń) wynosi w tym przypadku 1/6 n^3.

Przykład 4.5.

Wykorzystując rozkład Choleskiego rozwiązać układ równań postaci:

$$\begin{cases} x_1 + 2x_2 + 3x_3 = 1 \\ 2x_1 + 8x_2 + 10x_3 = 3 \\ 3x_1 + 10x_2 + 22x_3 = 7 \end{cases}.$$

Na początku sprawdzamy, czy macierz układu

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 8 & 10 \\ 3 & 10 & 22 \end{bmatrix}$$

jest dodatnio określona. W tym celu należy obliczyć minory główne:

$$A_1 = det[1] = 1, \quad A_2 = det\begin{bmatrix} 1 & 2 \\ 2 & 8 \end{bmatrix} = 4, \quad A_3 = \det(A) = 36.$$

Wszystkie minory główne są dodatnie co gwarantuje, że macierz **A** jest dodatnio określona i można zastosować rozkład Choleskiego. Obliczamy elementy macierzy **L**:

$$l_{11} = \sqrt{a_{11}} = 1,$$

$$l_{21} = \frac{a_{21}}{l_{11}} = \frac{2}{1} = 2,$$

$$l_{31} = \frac{a_{31}}{l_{11}} = \frac{3}{1} = 3,$$

$$l_{22} = \sqrt{a_{22} - \sum_{k=1}^{1}(l_{2k})^2} = \sqrt{a_{22} - (l_{21})^2} = \sqrt{8 - 2^2} = 2,$$

$$l_{32} = \frac{a_{32} - \sum_{k=1}^{1} l_{3k} \cdot l_{2k}}{l_{22}} = \frac{a_{32} - l_{31} \cdot l_{21}}{l_{22}} = \frac{10 - 3 \cdot 2}{2} = 2,$$

$$l_{33} = \sqrt{a_{33} - \sum_{k=1}^{2}(l_{3k})^2} = \sqrt{a_{33} - (l_{31})^2 - (l_{32})^2} = \sqrt{22 - 3^2 - 2^2} = 3$$

Macierz **L** ma zatem postać:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 3 & 2 & 3 \end{bmatrix}.$$

Po wyznaczeniu macierzy L należy rozwiązać dwa układy równań: **Ly**=**b** i $\mathbf{L}^T\mathbf{x}$=**y**.

Rozwiązujemy pierwszy układ **Ly** = **b**:

$$\begin{cases} y_1 = 1 \\ 2y_1 + 2y_2 = 3 \\ 3y_1 + 2y_2 + 3y_3 = 7 \end{cases} \Rightarrow y = \begin{bmatrix} 1 \\ 0.5 \\ 1 \end{bmatrix}.$$

Rozwiązujemy drugi układ $\mathbf{L}^{\mathrm{T}}\mathbf{x}=\mathbf{y}$:

$$\begin{cases} x_1 + 2x_2 + 3x_3 = 1 \\ 2x_2 + 2x_3 = 0.5 \\ 3x_3 = 1 \end{cases} \Rightarrow x = \begin{bmatrix} 1/6 \\ -1/12 \\ 1/3 \end{bmatrix}.$$

Rozwiązaniem podanego układu równań jest zatem wektor:

$$x = \left[\frac{1}{6}, \quad -\frac{1}{12}, \quad \frac{1}{3}\right]^T.$$

4.2.5. Rozkład QR metodą Householdera

Niech dane będą :

$A \in R^{m \times n}, \quad rank(A) = n \leq m \; i \; b \in R^m$.

Jeśli $n<m$ mówimy, że układ równań **Ax**=**b** jest nadokreślony, tzn. mamy więcej równań (m) niż niewiadomych (n). Taki układ nie zawsze posiada rozwiązanie, ale zawsze można znaleźć przybliżone rozwiązanie zgodne z pewnymi założeniami.

Zadanie wygładzania liniowego polega na znalezieniu wektora **x***, który minimalizuje wektor residualny (wektor reszty) $\mathbf{r} = \mathbf{b} - \mathbf{Ax}$ tzn.:

$$\|\mathrm{b} - \mathrm{Ax}^*\|_2 = \min_{\mathrm{x}} \|\mathrm{b} - \mathrm{Ax}\|_2. \tag{4.20}$$

Zadanie wygładzania liniowego jest więc uogólnieniem rozwiązywania kwadratowych układów równań liniowych. Metoda Householdera jest jedną z metod znajdowania rozwiązania dla układów równań prostokątnych.

Odbicia Householdera

Dla danego wektora $\mathrm{w} \in \mathrm{R}^{\mathrm{m}}$ o normie $\|\mathrm{w}\|_2 = \sqrt{\mathrm{w}^{\mathrm{T}}\mathrm{w}} = 1$, odbicie (macierz) Householdera definiujemy jako:

$$\mathrm{H} = \mathrm{I} - 2\mathrm{w}\mathrm{w}^{\mathrm{T}}. \tag{4.21}$$

Zauważmy, iż:

$$\mathrm{Hx} = \mathrm{x} - 2(\mathrm{w}^{\mathrm{T}}\mathrm{x})\mathrm{w} = \mathrm{x} - 2\mathrm{r}, \tag{4.22}$$

gdzie ***r*** jest rzutem prostopadłym ***x*** na kierunek wektora ***w***. Ze wzoru (4.22) wynika twierdzenie 4.1.

Twierdzenie 4.1

Przekształcenie Householdera przyporządkowuje wektorowi **x** jego odbicie lustrzane względem hiperpłaszczyzny prostopadłej do wektora ***w***.

Twierdzenie 4.2

Odbicia Householdera są przekształceniami symetrycznymi i ortogonalnymi (tj. niezmieniającymi długości wektora), tzn.

$$\mathrm{H}^{\mathrm{T}} = \mathrm{H} = \mathrm{H}^{-1}. \tag{4.23}$$

Przekształecenie Householdera stosuje się do przeprowadzenia wektora $x \neq 0$ na kierunek innego wektora niezerowego **e**, czyli:

$$\mathrm{Hx} = \alpha \mathrm{e}. \tag{4.24}$$

Załóżmy, że $\|e\|_2 = 1$. W szczególności dla ***e=e1*** otrzymujemy wzory:

$$\mathrm{H} = \mathrm{I} - \frac{1}{\beta}\mathrm{uu}^{\mathrm{T}}, \tag{4.25}$$

gdzie

$$\mathrm{u_i} = \begin{cases} \mathrm{x_1} + \mathrm{sign}(\mathrm{x_1})\|\mathrm{x}\|_2, & \mathrm{i} = 1 \\ \mathrm{x_i}, & 2 \leq \mathrm{i} \leq \mathrm{m} \end{cases} \tag{4.26}$$

$$\beta = \frac{1}{2}\|\mathrm{u}\|_2^2 = \|\mathrm{x}\|_2^2 + \mathrm{sign}(\mathrm{x_1})\mathrm{x_1}\|\mathrm{x}\|_2^2 \tag{4.27}$$

oraz sign(*t*) oznacza znak liczby *t*.

Współczynnik α ze wzoru (4.24) ma wartość:

$$\alpha = -sign(x_1)\|x\|_2. \tag{4.27a}$$

Rozkład QR

Odbić Householdera można użyć do uzyskania rozkładu macierzy $A \in R^{m \times n}$ na iloczyn ortogonalno-trójkątny.

Twierdzenie 4.3

Każdą macierz $\mathrm{A} \in \mathrm{R}^{\mathrm{m} \times \mathrm{n}}$ dla m≥n, której $\mathrm{rank(A)} = \mathrm{n} \leq \mathrm{m}$ czyli taką, której kolumny są liniowo niezależne, można przedstawić w postaci:

$$\mathrm{A} = \mathrm{Q} \cdot \mathrm{R}, \tag{4.28}$$

gdzie $Q \in R^{m\times m}$ ma kolumny ortogonalne, a $R \in R^{m\times n}$ jest macierzą uogólnioną trójkątną górną.

Niech:

$$\mathrm{A}^{(0)} = \left[\mathrm{a}_1^{(0)}, \mathrm{a}_2^{(0)}, \ldots, \mathrm{a}_\mathrm{n}^{(0)}\right] \tag{4.29}$$

gdzie a_j oznacza j-tą kolumnę macierzy A.

Niech $\mathbf{H_1}$ przekształca $a_1^{(0)}$ na wersor $e_1 \in R^m$.

Wtedy

$$\mathrm{A}^{(1)} = \mathrm{H}_1\mathrm{A} = \left[\mathrm{H}_1\mathrm{a}_1^{(1)}, \mathrm{H}_1\mathrm{a}_2^{(1)}, \ldots, \mathrm{H}_1\mathrm{a}_\mathrm{n}^{(1)}\right] \tag{4.30}$$

$$\mathrm{A}^{(1)} = \begin{bmatrix} \mathrm{r}_{11} & \mathrm{r}_{12} & \ldots & \mathrm{r}_{1\mathrm{n}} \\ 0 & \mathrm{a}_2^{(1)} & \ldots & \mathrm{a}_\mathrm{n}^{(1)} \end{bmatrix} \tag{4.31}$$

W kolejnym kroku wybieramy przekształcenie Householdera $\overline{H}_2$ tak, aby przekształcało ono wektor $a_2^{(1)}$ na kierunek wersora $e_1 \in R^{m-1}$. Wtedy przyjmujemy:

$$\mathrm{H}_2 = \begin{bmatrix} 1 & 0 \\ 0 & \overline{\mathrm{H}}_2 \end{bmatrix}. \tag{4.32}$$

Pomnożenie macierzy $\mathrm{A}^{(1)}$ z lewej strony przez H_2 spowoduje wyzerowanie drugiej kolumny macierzy poniżej elementu $a_{2,2}^{(1)}$, przy czym pierwszy wiersz i pierwsza kolumna macierzy pozostaną niezmienione. Zatem:

$$\mathrm{A}^{(2)} = \mathrm{H}_2\mathrm{H}_1\mathrm{A} = \begin{bmatrix} \mathrm{r}_{11} & \mathrm{r}_{12} & \mathrm{r}_{13} & \ldots & \mathrm{r}_{1\mathrm{n}} \\ 0 & \mathrm{H}^{(2)}\mathrm{a}_2^{(1)} & \mathrm{H}^{(2)}\mathrm{a}_3^{(1)} & \ldots & \mathrm{H}^{(2)}\mathrm{a}_\mathrm{n}^{(1)} \end{bmatrix} \tag{4.33}$$

i postać ostateczna:

$$\mathrm{A}^{(2)} = \begin{bmatrix} \mathrm{r}_{11} & \mathrm{r}_{12} & \mathrm{r}_{13} & \ldots & \mathrm{r}_{1\mathrm{n}} \\ 0 & \mathrm{r}_{22} & \mathrm{r}_{23} & \ldots & \mathrm{r}_{2\mathrm{n}} \\ 0 & 0 & \mathrm{a}_3^{(2)} & \ldots & \mathrm{a}_\mathrm{n}^{(2)} \end{bmatrix}. \tag{4.34}$$

Po wykonaniu *n* kroków otrzymujemy macierz:

$$\mathrm{A}^{(\mathrm{n})} = \mathrm{H}_\mathrm{n} \ldots \mathrm{H}_2\mathrm{H}_1\mathrm{A} \tag{4.35}$$

postaci:

$$A^{(n)} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & 0 & r_{22} & \cdots & r_{2n} \\ & \vdots & \vdots & \ddots & \vdots \\ & 0 & 0 & \cdots & H^{(n)}a_n^{(n-1)} \end{bmatrix}. \quad (4.36)$$

gdzie **H(*n*)** przekształca $a_n^{(n-1)}$ na wersor $e_1 \in R^1$.

Oznaczmy

$$R_1 = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{bmatrix}. \quad (4.37)$$

Wtedy:

$$R = \begin{bmatrix} R_1 \\ 0 \end{bmatrix}. \quad (4.38)$$

jest uogólnioną macierzą trójkątną górną wymiaru *m*×*n* taką, że:

$r_{ij} = 0\ dla\ i > j.$

Skoro:

$$H_n \dots H_2 H_1 A = R, \quad (4.39)$$

to podstawiając:

$$Q = H_1 H_2 \dots H_n \quad (4.40)$$

dostajemy rozkład macierzy **A** na iloczyn macierzy ortogonalnej **Q** i macierzy górnie trójkątnej **R** (4.28).

Koszt metody Householdera to $2n^2(m - \frac{n}{3})$. Dla m=n daje to $\frac{4}{3}n^3$ czyli dwa razy więcej niż metoda eliminacji Gaussa. Metoda Householdera jest numerycznie poprawna, idealnie nadaje się do obliczeń równoległych. Niewątpliwą jej zaletą jest fakt, iż możemy ją stosować w przypadku układów prostokątnych. Dla układów nadokreślonych otrzymujemy rozwiązanie minimalizujące sumę kwadratów, a w przypadku układów niedookreślonych rozwiązanie o minimalnej normie. Metoda ta posiada także więszą dokładnośc obliczeń, w porównaniu z metodą Gaussa.

Przykład 4.5a.

Wykorzystując odbicie Householdera przeprowadzić wektor $\vec{a} = [-4,0,3]^T$ na wektor $\vec{b} = [*,0,0]^T$.

Najpierw policzymy długość wektora $\vec{a}$:

$\|\bar{a}\|_2 = \sqrt{(-4)^2 + 0 + 3^2} = 5$

i

$\text{sign}(a_1) = \text{sign}(-4) = -1$

Następnie należy obliczyć współczynnik α ze wzoru (4.27a):

$\alpha = -(-1) \cdot 5 = 5.$

W tym przypadku, nie jest konieczne obliczanie całego **H**. Otrzymujemy wektor $\vec{b} = [5,0,0]^T$.

4.2.6. Wyznaczanie macierzy odwrotnej

Przy założeniu, że macierz **A** o wymiarze *n* x *n* jest macierzą nieosobliwą, możemy wyznaczyć rozkład trójkątny macierzy **A=LU.**
Tzn.:

$$\mathbf{A} = \begin{bmatrix} \underline{1} & 0 & \dots & 0 & 0 \\ l_{21} & \underline{1} & & 0 & 0 \\ \dots & & \dots & & \dots \\ \dots & & & \dots & \dots \\ l_{n1} & l_{n2} & \dots & l_{n,n-1} & \underline{1} \end{bmatrix} \begin{bmatrix} \underline{u_{11}} & u_{12} & \dots & u_{1,n-1} & u_{1n} \\ 0 & \underline{u_{22}} & & u_{2,n-1} & u_{2n} \\ \dots & & \dots & & \dots \\ \dots & & & \dots & \dots \\ 0 & 0 & \dots & 0 & \underline{u_{nn}} \end{bmatrix}$$

i obliczyć macierz odwrotną do **A** jako iloczyn macierzy odwrotnych do **U** i do **L**, czyli:

$\mathbf{A}^{-1} = \mathbf{U}^{-1} \cdot \mathbf{L}^{-1}$.

Macierz odwrotna do **L** jest także macierzą dolnotrójkątną:

$$\mathbf{L}^{-1} = \begin{bmatrix} \underline{1} & 0 & \dots & 0 & 0 \\ l'_{21} & \underline{1} & & 0 & 0 \\ \dots & & \dots & & \dots \\ \dots & & & \dots & \dots \\ l'_{n1} & l'_{n2} & \dots & l'_{n,n-1} & \underline{1} \end{bmatrix}$$ i $\mathbf{L}\,\mathbf{L}^{-1}=\mathbf{I}$.

Elementy l'_{ij} macierzy odwrotnej $\mathbf{L}^{-1}$ wyznaczamy ze wzorów:

$$l'_{ij} = -l_{ij} - \sum_{k=j+1}^{i-1} l_{ik} \cdot l'_{kj} \qquad dla \quad i > j$$
$$l'_{ij} = 1 \qquad dla \quad i = j \tag{4.41}$$
$$l'_{ij} = 0 \qquad dla \quad i < j$$

gdzie $i = 1, ..., n,\ j = 1, ..., n$.

Macierz odwrotna do **U** jest macierzą górnotrójkątną:

$$\mathbf{U}^{-1} = \begin{bmatrix} \underline{u'_{11}} & u'_{12} & \dots & u'_{1,n-1} & u'_{1n} \\ 0 & \underline{u'_{22}} & & u'_{2,n-1} & u'_{2n} \\ \dots & & \dots & & \dots \\ \dots & & & \dots & \dots \\ 0 & 0 & \dots & 0 & \underline{u'_{nn}} \end{bmatrix}$$ i $\mathbf{U}\,\mathbf{U}^{-1}=\mathbf{I}$.

Elementy u'_{ij} macierzy odwrotnej $\mathbf{U}^{-1}$ wyznaczamy ze wzorów:

$$u'_{ij} = \frac{\left(-\sum_{k=i+1}^{n} u_{ik} \cdot u'_{kj}\right)}{u_{ii}} \qquad dla \quad i < j,$$
$$u'_{ij} = 1/u_{ij} \qquad dla \quad i = j, \tag{4.42}$$
$$u'_{ij} = 0 \qquad dla \quad i > j,$$

gdzie $i = 1, ..., n,\ j = 1, ..., n$.

Gdy **L** i **U** są znane, metoda wymaga $\frac{1}{6}n^3 + \frac{1}{6}n^3 + \frac{1}{3}n^3 = \frac{2}{3}n^3$ operacji. Ponieważ rozkład trójkątny macierzy **A** wymaga $n^3/3$ operacji, to ogólny koszt odwracania macierzy wynosi n^3 operacji.

4.2.7. Obliczanie wyznacznika macierzy

W celu obliczenia wyznacznika macierzy **A** dokonujemy rozkładu macierzy **A** metodą **LU**. Jest to pomocne dlatego, że wyznacznik macierzy trójkątnej jest równy iloczynowi elementów tej macierzy stojących na głównej przekątnej.

Jeżeli więc ***A=LU***, wówczas:

$$\det \mathbf{A} = \det \mathbf{L} \cdot \det \mathbf{U} = 1 \cdot \det \mathbf{U} = u_{11}u_{22} \cdot \dots \cdot u_{nn}.$$

W przypadku przestawiania wierszy macierzy **A**, wyznacznik ten należy pomnożyć przez $(-1)^s$, gdzie s jest łączną liczbą przestawień wierszy.

4.3. Metody iteracyjne

W poprzednich podrozdziałach przedstawiono najważniejsze metody skończone rozwiązywania układów równań liniowych. Realizacja tych metod dla układu n x n wymaga n^2 komórek w pamięci operacyjnej oraz wykonania rzędu n^3 działań arytmetycznych. Jeśli więc tylko n nie jest zbyt duże i układ jest dostatecznie dobrze uwarunkowany, to rozwiązanie zadania nie nastręcza żadnych trudności.

W praktyce obliczeniowej pojawiają się jednak dosyć często układy liniowe, których wymiar n jest rzędu 10^3 lub nawet większy. Podstawowym źródłem takich zadań są metody przybliżonego rozwiązywania równań różniczkowych cząstkowych. Prawie zawsze macierze wielkich układów liniowych nie są macierzami pełnymi, ale rzadkimi, tzn. mają niewiele elementów niezerowych. Jednym ze sposobów rozwiązywania wielkich układów równań jest stosowanie metod iteracyjnych, które zakładają jedynie możliwość mnożenia dowolnego wektora przez macierz układu (lub przez macierz od niej pochodzącą). Jeśli macierz jest rozrzedzona, to mnożenie takie wymaga wykonania około n działań arytmetycznych a nie n^2, jak w przypadku ogólnym.

Zaletą metod iteracyjnych jest również możliwość wyznaczenia przybliżenia rozwiązania z zadaną dokładnością, niekiedy kosztem istotnie mniejszym od kosztu metod skończonych. Dla niektórych zadań metody iteracyjne są więc efektywniejsze.

Jedną z najprostszych metod iteracyjnych jest ***metoda iteracji prostej***. Polega ona na przejściu od danego układu równań liniowych (4.1) do równoważnego (tzn. mającego te same rozwiązania) układu:

$$\mathbf{x} = \mathbf{B}\mathbf{x}+\mathbf{c}. \tag{4.43}$$

Sposób wyznaczenia macierzy **B** i wektora **c** zależy od rodzaju stosowanej metody iteracyjnej (patrz rozdział 4.3.1, 4.3.2, 4.3.3).

Znając (4.43) wyznaczamy ciąg $\left\{\mathbf{x}^{(i)}\right\}, i = 1,2,...$ kolejnych przybliżeń rozwiązania $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ ze wzoru:

$$\mathbf{x}^{(i+1)} = \mathbf{B}\mathbf{x}^{(i)} + \mathbf{c}, \ i = 0,1,... \tag{4.44}$$

Odejmując stronami równanie (4.43) od równania (4.44) otrzymujemy:

$$\mathbf{x}^{(i+1)} - \mathbf{x} = \mathbf{B}(\mathbf{x}^{(i)} - \mathbf{x}), \ \ i = 0,1,... \tag{4.45}$$

a stąd:

$$\mathbf{x}^{(i+1)} - \mathbf{x} = \mathbf{B}(\mathbf{x}^{(i)} - \mathbf{x}) = \mathbf{B}\mathbf{B}(\mathbf{x}^{(i-1)} - \mathbf{x})... = \mathbf{B}^{i+1}(\mathbf{x}^{(0)} - \mathbf{x}), \tag{4.46}$$

gdzie:

$$\mathbf{B}^{i+1} = \mathbf{B} \cdot \mathbf{B} \cdot ... \cdot \mathbf{B} \ (i+1 \text{ razy}), \ \boldsymbol{i}=0,1,... \ . \tag{4.47}$$

Przechodząc do norm otrzymujemy oszacowania:

$$\| \ \mathbf{x}^{(i+1)} - \mathbf{x} \ \| \le \| \ \mathbf{B}^{i+1} \ \| \ \| \ (\mathbf{x}^{(0)} - \mathbf{x}) \ \| \le \| \ \mathbf{B} \ \|^{i+1} \ \| \ (\mathbf{x}^{(0)} - \mathbf{x}) \ \|. \tag{4.48}$$

Istotną rzeczą jest znajomość warunku wystarczającego zbieżności metody iteracyjnej. Wystarczającym warunkiem na to, aby ciąg $\left\{\mathbf{x}^{(i)}\right\}, i = 1,2,...$ zdefiniowany wzorem (4.44) był zbieżny do rozwiązania układu (4.1) jest, aby dowolna norma macierzy **B** była mniejsza od jedności. Dla wielu metod sprawdzenie nierówności $\|\mathbf{B}\| \le 1$ jest możliwe.

Metody Jacobiego, Gaussa-Seidela, nadrelaksacji (SOR) są wariantami metody iteracji prostej.

4.3.1. Metoda Jacobiego

Przedstawmy macierz wyjściowego układu w postaci:

$$\mathbf{A}=\mathbf{L}+\mathbf{D}+\mathbf{U}, \tag{4.49}$$

gdzie **D** jest macierzą diagonalną, **L** - macierzą dolną trójkątną, a **U** - macierzą górną trójkątną o zerowych elementach diagonalnych.

Przykład 4.6.

Rozkład przykładowej macierzy według wzoru (4.49) jest postaci:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 4 & 0 & 0 \\ 7 & 8 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{bmatrix} + \begin{bmatrix} 0 & 2 & 3 \\ 0 & 0 & 6 \\ 0 & 0 & 0 \end{bmatrix}.$$

Rozwiązywany układ **Ax**=**b** można zapisać jako:

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

a stąd:

$$\mathbf{Dx}=-(\mathbf{L}+\mathbf{U})\mathbf{x}+\mathbf{b}. \tag{4.50}$$

Jeśli macierz **D** jest nieosobliwa (jest tak np. w przypadku macierzy symetrycznej dodatnio określonej) to możemy przejść do układu równoważnego:

$$\mathbf{x} = \mathbf{B}_{\mathrm{J}}\,\mathbf{x} + \mathbf{c}, \tag{4.51}$$

gdzie:

$$\begin{aligned} \mathbf{B}_{\mathrm{J}} &= -\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U}), \\ \mathbf{c} &= \mathbf{D}^{-1}\mathbf{b}. \end{aligned} \tag{4.52}$$

Przekształcając równanie (4.50) zapisane w postaci:

$\mathbf{Dx}^{(i+1)}=-(\mathbf{L}+\mathbf{U})\mathbf{x}^{(i)}+\mathbf{b}$,

otrzymujemy:

$$a_{kk}x_k^{(i+1)} = -\left(\sum_{j=1}^{k-1} a_{kj}x_j^{(i)} + \sum_{j=k+1}^{n} a_{kj}x_j^{(i)}\right) + b_k, k = 1,2,...,n.$$

Po dalszych przekształceniach dochodzimy do zależności:

$$x_k^{(i+1)} = \frac{-\sum_{j=1, j\neq k}^{n} a_{kj}x_j^{(i)} + b_k}{a_{kk}}, \text{ dla } a_{kk} \neq 0, k = 1,2,...,n. \tag{4.53}$$

Jako początkowe przybliżenie wybiera się często wektor $\mathbf{x}^{(0)}=\mathbf{0}$.

Warunek konieczny i dostateczny zbieżności jest spełniony m.in. gdy **A** jest nieredukowalna i diagonalnie dominująca.

Macierz A jest ***nieredukowalna***, jeżeli poprzez przestawienie wierszy i kolumn nie można jej sprowadzić do postaci blokowej górnej trójkątnej.

Macierz **A** o wymiarze *n* x *n* nazywamy ***diagonalnie dominującą***, jeśli dla *i*=1,2,...,*n* zachodzi nierówność $|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{ij}|$.

Przykład 4.7.

Zgodnie z definicją macierz:

$$\mathbf{A} = \begin{bmatrix} 3 & 1 & 0 \\ 0 & 2 & 2 \\ 4 & 1 & -5 \end{bmatrix}$$

jest macierzą diagonalnie dominującą, ponieważ dla każdego *i*=1,2,3 zachodzi $|a_{ii}| \geq \sum_{j=1, j\neq i}^{3} |a_{ij}|$.

Jeśli **A** jest macierzą nieredukowalną i diagonalnie dominującą, to **A** jest nieosobliwa i wszystkie elementy diagonalne macierzy **A** są różne od zera.

Również macierze symetryczne dodatnio określone są nieosobliwe, a wszystkie ich elementy diagonalne są dodatnie.

4.3.2. Metoda Gaussa-Seidela

Załóżmy, że znamy już przybliżenie $\mathbf{x}^{(i)} = \left[x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}\right]^{\mathrm{T}}$. W metodzie Gaussa-Seidla następne przybliżenie $\mathbf{x}^{(i+1)}$ wyznacza się tak, aby jego kolejne współrzędne $x_k^{(i+1)}$, (*k*=1,2,...*n*) spełniały równania:

$$a_{k1}x_1^{(i+1)} + \ldots + a_{kk}x_k^{(i+1)} + a_{k,k+1}x_{k+1}^{(i)} + \ldots + a_{kn}x_n^{(i)} = b_k \qquad (4.54)$$

Korzystając z przedstawienia (4.49) macierzy **A**, możemy te zależności zapisać w następujący sposób:

$$(\mathbf{L} + \mathbf{D})\mathbf{x}^{(i+1)} + \mathbf{U}\mathbf{x}^{(i)} = \mathbf{b},$$

a stąd:

$$\mathbf{x}^{(i+1)} = \mathbf{B}_{\mathrm{GS}}\mathbf{x}^{(i)} + \mathbf{c}, \qquad (4.55)$$

gdzie:

$$\begin{aligned}\mathbf{B}_{\mathrm{GS}} &= -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{U},\\ \mathbf{c} &= (\mathbf{L}+\mathbf{D})^{-1}\mathbf{b}.\end{aligned} \tag{4.56}$$

Warunkiem na to, żeby macierz $\mathbf{B}_{\mathrm{GS}}$ była dobrze określona, jest niezerowość wszystkich elementów diagonalnych macierzy **A**.

Przekształcając równanie (4.54), otrzymujemy następującą zależność pomiędzy współrzędnymi kolejnych przybliżeń:

$$x_k^{(i+1)} = \frac{-\sum_{j=1}^{k-1} a_{kj} x_j^{(i+1)} - \sum_{j=k+1}^{n} a_{kj} x_j^{(i)} + b_k}{a_{kk}}, k = 1,2,...,n. \tag{4.57}$$

Metoda Gaussa-Seidela, jako ulepszenie metody Jacobiego, zachowuje te same warunki zbieżności. Jeżeli macierz **A** jest dodatnio określona to metoda Gaussa-Seidela jest zbieżna dla dowolnego wektora początkowego [8, 10].

Metodę Gaussa-Seidela stosuje się niemal wyłącznie do układów z macierzą diagonalnie dominującą gdyż w wielu praktycznych zastosowaniach jest to łatwy do spełnienia warunek gwarantujący zbieżność metody.

4.3.3. Metoda SOR (nadrelaksacji)

Modyfikacja metody Gaussa-Seidela, przyspieszająca zbieżność konstruowanego ciągu (4.57), polega na przemnożeniu poprawki obliczanej w (4.57) przez odpowiednio dobraną liczbę ω.

Ponownie korzystając z wzoru (4.49) mamy:

$$\begin{aligned}&\omega(\mathbf{L}+\mathbf{D}+\mathbf{U})\mathbf{x} = \omega\mathbf{b} \Rightarrow\\ &\mathbf{D}\mathbf{x} = \mathbf{D}\mathbf{x} - \omega\left[(\mathbf{L}+\mathbf{D}+\mathbf{U})\mathbf{x} - \mathbf{b}\right] \Rightarrow\\ &(\mathbf{D}+\omega\mathbf{L})\mathbf{x}^{(i+1)} = (1-\omega)\mathbf{D}\mathbf{x}^{(i)} - \omega\mathbf{U}\mathbf{x}^{(i)} + \omega\mathbf{b}\end{aligned} \tag{4.58}$$

lub w równoważnej postaci:

$$\mathbf{D}\mathbf{x}^{(i+1)} = (1-\omega)\mathbf{D}\mathbf{x}^{(i)} - \omega(\mathbf{L}\mathbf{x}^{(i+1)} + \mathbf{U}\mathbf{x}^{(i)} - \mathbf{b}). \tag{4.59}$$

Przekształcając dalej otrzymujemy zależność:

$$\mathbf{x}^{(i+1)} = \mathbf{B}_{\omega}\mathbf{x}^{(i)} + \mathbf{c}, \tag{4.60}$$

gdzie:

$$\mathbf{B}_{\omega} = (\mathbf{D} + \omega\mathbf{L})^{-1}((1-\omega)\mathbf{D} - \omega\mathbf{U}),$$

$$\mathbf{c} = \omega(\mathbf{D} + \omega\mathbf{L})^{-1}\mathbf{b}.$$

Wzór (4.60) można zapisać w postaci:

$$x_k^{(i+1)} = x_k^{(i)} + \omega \frac{-\sum_{j=1}^{k-1} a_{kj} x_j^{(i+1)} - \sum_{j=k}^{n} a_{kj} x_j^{(i)} + b_k}{a_{kk}}, k = 1,2,...,n, \qquad (4.61)$$

i dalej:

$$a_{kk} x_k^{(i+1)} = (1-\omega) a_{kk} x_k^{(i)} - \omega \left(\sum_{j=1}^{k-1} a_{kj} x_j^{(i+1)} + \sum_{j=k}^{n} a_{kj} x_j^{(i)} - b_k \right). \qquad (4.62)$$

Dla ω=1 jest to metoda SOR (ang. *successive over relaxation*). Zwiększając współczynnik ϖ, można próbować przyspieszać jej zbieżność. Parametr ω może przyjmować wartości co najwyżej z przedziału (0, 2), gdyż dla pozostałych wartości metoda może nie być zbieżna dla pewnych przybliżeń początkowych.

Metody Jacobiego i Gaussa-Seidela można ewentualnie stosować do układów bardzo dobrze uwarunkowanych. Znacznie efektywniejsze, szczególnie dla zadań o dużym wskaźniku uwarunkowania, jest użycie metody SOR lub metody Czebyszewa.

Przykład 4.8.

Dla układu:

$$\begin{cases} 4x_1 - x_2 = 2 \\ -x_1 + 4x_2 - x_3 = 6 \\ -x_2 + 4x_3 = 2 \end{cases}$$

obliczono kilka przybliżeń metodami Jacobiego, Gaussa-Seidela i SOR. We wszystkich metodach przyjęto $\mathbf{x}^{(1)} = \mathbf{0}$.

Metoda Jacobiego

Dla metody Jacobiego, korzystając ze wzoru (4.50) :

$\mathbf{D}\mathbf{x}^{(i+1)} = -(\mathbf{L}+\mathbf{U})\mathbf{x}^{(i)} + \mathbf{b}$,

a stąd:

$\mathbf{x}^{(i+1)} = -\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\mathbf{x}^{(i)}+\mathbf{D}^{-1}\mathbf{b}$.

Obliczamy kolejno:

$$\mathbf{x}^{(2)} = -\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\mathbf{x}^{(1)}+\mathbf{D}^{-1}\mathbf{b} =$$

$$= -\begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix}\begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix}\begin{bmatrix} 2 \\ 6 \\ 2 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.5 \\ 0.5 \end{bmatrix},$$

$$\mathbf{x}^{(3)} = -\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\mathbf{x}^{(2)}+\mathbf{D}^{-1}\mathbf{b} =$$

$$= -\begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix}\begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix}\begin{bmatrix} \frac{1}{2} \\ \frac{3}{2} \\ \frac{1}{2} \end{bmatrix} + \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix}\begin{bmatrix} 2 \\ 6 \\ 2 \end{bmatrix} = \begin{bmatrix} 0.875 \\ 1.75 \\ 0.875 \end{bmatrix},$$

$\mathbf{x}^{(4)} = -\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\mathbf{x}^{(3)}+\mathbf{D}^{-1}\mathbf{b}$=[0.9375, 1.9375, 0.9375],

$\mathbf{x}^{(5)} = -\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\mathbf{x}^{(4)}+\mathbf{D}^{-1}\mathbf{b}$=[0.9844,1.9688,0.9844].

Metoda Gaussa-Seidela

Korzystając z metody Gaussa-Seidela określonej wzorami (4.55), (4.56):

$\mathbf{x}^{(i+1)} = -(\mathbf{D}+\mathbf{L})^{-1}\mathbf{U}\mathbf{x}^{(i)}+(\mathbf{D}+\mathbf{L})^{-1}\mathbf{b}$,

wyznaczamy kolejne przybliżenia:

$\mathbf{x}^{(2)} = -(\mathbf{D}+\mathbf{L})^{-1}\mathbf{U}\mathbf{x}^{(1)}+(\mathbf{D}+\mathbf{L})^{-1}\mathbf{b}$=[0.5000, 1.6250, 0.9062],

$\mathbf{x}^{(3)} = -(\mathbf{D}+\mathbf{L})^{-1}\mathbf{U}\mathbf{x}^{(2)}+(\mathbf{D}+\mathbf{L})^{-1}\mathbf{b}$=[0.9062, 1.9531, 0.9883],

$\mathbf{x}^{(4)} = -(\mathbf{D}+\mathbf{L})^{-1}\mathbf{U}\mathbf{x}^{(3)}+(\mathbf{D}+\mathbf{L})^{-1}\mathbf{b}$=[0.9883, 1.9941, 0.9985],

$\mathbf{x}^{(5)} = -(\mathbf{D}+\mathbf{L})^{-1}\mathbf{U}\mathbf{x}^{(4)}+(\mathbf{D}+\mathbf{L})^{-1}\mathbf{b}$=[0.9885, 1.9993, 0.9998].

Metoda SOR

W przypadku wykorzystania metody SOR zdefiniowanej wzorem (4.61):

$$\mathbf{x}^{(i+1)} == (\mathbf{D}+\omega\mathbf{L})^{-1}((1-\omega)\mathbf{D}-\omega\mathbf{U})\mathbf{x}^{(i)} + \omega(\mathbf{D}+\omega\mathbf{L})^{-1}\mathbf{b},$$

dla ω**=1.2** otrzymujemy kolejno:

$\mathbf{x}^{(2)}$= [0.6000, 1.9800, 1.1940],

$\mathbf{x}^{(3)}$= [1.0740, 2.0844, 0.9865],

$\mathbf{x}^{(4)}$= [1.0105, 1.9822, 0.9974],

$\mathbf{x}^{(5)}$= [0.9926, 2.0005, 1.0007],

natomiast dla ω**=1.1** mamy:

$\mathbf{x}^{(2)}$= [0.5500, 1.8013, 1.0453],

$\mathbf{x}^{(3)}$= [0.9903, 2.0297, 1.0036],

$\mathbf{x}^{(4)}$= [1.0091, 2.0005, 0.9998],

$\mathbf{x}^{(5)}$= [0.9998, 1.9997, 0.9999].

Dokładne rozwiązanie układu jest równe:
x=[1, 2, 1].

Z powyższych zestawień wynika więc, że metoda Jacobiego jest zbieżna najwolniej, natomiast najszybciej zbieżna jest metoda SOR w przypadku ω=1.1.

4.3.4. Metoda Czebyszewa

Zajmiemy się teraz metodą rozwiązywania wielkich układów równań liniowych **Ax**=**b** o symetrycznej dodatnio określonej macierzy **A** wymiaru *n* x *n*.

Chcemy skonstruować ciąg $\{x^i\}$ przybliżeń rozwiązania $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, spełniający równości :

$$\mathbf{x}^{(i)} - \mathbf{x} = W_i(\mathbf{A})(\mathbf{x}^{(0)} - \mathbf{x}), \tag{4.63}$$

gdzie W_i jest wielomianem stopnia nie większego niż *i*, a $\mathbf{x}^{(0)}$ jest danym przybliżeniem początkowym **x**.

Przekształcając (4.63) otrzymujemy zależność:

$$\mathbf{x}^{(i)} = W_i(\mathbf{A})\mathbf{x}^{(0)} - (W_i(\mathbf{A}) - \mathbf{I})\mathbf{x}.$$

W celu wyznaczenia wektora $\mathbf{x}^{(i)}$ należy obliczyć $(W_i(\mathbf{A}) + \mathbf{I})\mathbf{x}$.
Wektor ten jest równy:

$$(W_i(\mathbf{A}) + \mathbf{I})\mathbf{x} = (a_i^{(i)}\mathbf{A}^i + a_{i-1}^{(i)}\mathbf{A}^{i-1} + ... + a_1^{(i)}\mathbf{A})\mathbf{x} + (a_0^{(i)} - 1)\mathbf{x} =$$
$$= (a_i^{(i)}\mathbf{A}^{i-1} + a_{i-1}^{(i)}\mathbf{A}^{i-2} + ... + a_1^{(i)}\mathbf{I})\mathrm{b} + (a_0^{(i)} - 1)\mathbf{x}.$$

Musimy zatem założyć, że $a_0^{(i)} = 1$, co odpowiada warunkowi $W_i(0) = 1$. Z zależności (4.63) wynika oszacowanie:

$$\left\|\mathbf{x}^{(i)} - \mathbf{x}\right\|_2 \le \left\|W_i(\mathbf{A})\right\|_2 \left\|\mathbf{x}^{(0)} - \mathbf{x}\right\|_2.$$

Aby zapewnić jak najlepszą zbieżność ciągu $\{\mathbf{x}_i\}$, musimy wybrać wielomian W_i o możliwie małej normie $\left\|W_i(\mathbf{A})\right\|_2$. Ponieważ **A** jest z założenia macierzą symetryczną, to $\left\|W_i(\mathbf{A})\right\|_2 = \max\left|\mathbf{W}_i(\lambda)\right|$, gdzie λ należy do zbioru wartości własnych macierzy **A**. Na ogół nie znamy wartości własnych macierzy **A**, a jedynie pewien przedział <*a*,*b*> zawierający te wartości (0<*a*<*b*). Wówczas: $\left\|W_i(\mathbf{A})\right\|_2 \le \left\|W_i\right\| = \max\left|W_i(\lambda)\right|$.

Nie znając wartości własnych macierzy **A** potrzebnych do minimalizacji $\|W_i(\mathbf{A})\|_2$, będziemy minimalizowali $\|W_i\|$, do czego wystarcza nam znajomość przedziału <*a*, *b*>.

Twierdzenie 4.4. (o wielomianach Czebyszewa)

Niech *a*, $|a| > 1$ i $b \neq 0$ będą zadanymi liczbami. Spośród wszystkich wielomianów w_k stopnia *k* spełniających równość $w_k(a) = b$ najmniejszą normę ma wielomian :

$$w_k = \frac{b}{T_k(a)} T_k ,$$

gdzie T_k jest *k*-tym wielomianem Czebyszewa określonym wzorem:

$$T_k(x) = \tfrac{1}{2}[(x + \sqrt{x^2 - 1})^k + (x - \sqrt{x^2 - 1})^k . \tag{4.64}$$

Z powyższego twierdzenia wynika, że w klasie wielomianów stopnia co najwyżej *i*, spełniających warunek $W_i(0) = 1$, najmniejszą normę ma wielomian:

$$W_i(\lambda) = T_i(f(\lambda))/T_i(f(0)), \tag{4.65}$$

gdzie:

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

Metoda (4.63), w której wielomiany określone są przez (4.65), nosi nazwę ***metody Czebyszewa***.

Zbieżność ciągu $\{x^{(i)}\}$ konstruowanego w metodzie Czebyszewa jest określona nierównością:

$$\left\|\mathbf{x}^{(i)} - \mathbf{x}\right\|_2 \leq 2\left(\frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}}\right)\left\|\mathbf{x}^{(0)} - \mathbf{x}\right\|_2 . \tag{4.66}$$

W celu skonstruowania ciągu $\left\{\mathbf{x}^{(i)}\right\}$ w metodzie Czebyszewa zdefiniujmy:

$$t_i = T_i(f(0)) = T_i((b+a)/(b-a)) .$$

Z zależności wielomianów Czebyszewa wiadomo, że spełniają one następującą zależność rekurencyjną (wzór 3.68):

$T_0(\lambda) = 1,$

$T_1(\lambda) = \lambda,$

$T_{i+1}(\lambda) = 2\lambda\, T_i(\lambda) - T_{i-1}(\lambda),\ i = 1,2,\ldots.$

Dla i=1 mamy zatem:

$$\mathbf{x}^{(1)} - \mathbf{x} = W_1(\mathbf{A})(\mathbf{x}^{(0)} - \mathbf{x}) = \frac{T_1(f(\mathbf{A}))}{t_1}(\mathbf{x}^{(0)} - \mathbf{x}) =$$

$$= \frac{T_1\left(\frac{b+a}{b-a} - 2\frac{\mathbf{A}}{b-a}\right)}{T_1\left(\frac{b+a}{b-a}\right)}(\mathbf{x}^{(0)} - \mathbf{x}) = \frac{\left(\frac{b+a}{b-a} - 2\frac{\mathbf{A}}{b-a}\right)}{\left(\frac{b+a}{b-a}\right)}(\mathbf{x}^{(0)} - \mathbf{x}) =$$

$$= \left(\mathbf{I} - \frac{2}{b+a}\mathbf{A}\right)(\mathbf{x}^{(0)} - \mathbf{x}),$$

a stąd:

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - 2\mathbf{r}^{(0)}/(b+a), \tag{4.67}$$

gdzie $\mathbf{r}^{(0)} = \mathbf{A}\mathbf{x}^{(0)} - \mathbf{b}$.

Dla $i \geq 1$ możemy zapisać następujący związek:

$$\frac{T_{i+1}(f(\mathbf{A}))}{t_{i+1}} = \frac{2t_i}{t_{i+1}} f(\mathbf{A}) \frac{T_i(f(\mathbf{A}))}{t_i} - \frac{t_{i-1}}{t_{i+1}} \frac{T_{i-1}(f(\mathbf{A}))}{t_{i-1}},$$

czyli:

$$W_{i+1}(\mathbf{A}) = \frac{2t_i}{t_{i+1}} f(\mathbf{A}) W_i(\mathbf{A}) - \frac{t_{i-1}}{t_{i+1}} W_{i-1}(\mathbf{A}).$$

i dalej [5]:

$$\mathbf{x}^{(i+1)} - \mathbf{x} = \frac{2t_i}{t_{i+1}}\left(\frac{b+a}{b-a}\mathbf{I} - \frac{2}{b-a}\mathbf{A}\right)(\mathbf{x}^{(i)} - \mathbf{x}) - \frac{t_{i-1}}{t_{i+1}}(\mathbf{x}^{(i-1)} - \mathbf{x}).$$

Uwzględniając zależność $t_{i+1} = 2t_i(b+a)/(b-a) - t_{i-1}$ oraz wzór (4.67), otrzymujemy zależność rekurencyjną:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} + (p_{i-1}(\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}) - \mathbf{r}^{(i)})/q_i,\ i = 0,1,\ldots \tag{4.68}$$

gdzie:

$\mathbf{r}^{(i)} = \mathbf{A}\mathbf{x}^{(i)} - \mathbf{b}$,

$$p_{-1} = 0, \quad p_{i\text{-}1} = \frac{(b-a)}{4}\frac{t_{i\text{-}1}}{t_i},$$

$$q_0 = (b+a)/2, \quad q_i = \frac{(b-a)}{4}\frac{t_{i\text{-}1}}{t_i}\ .$$

Główną częścią kosztu wyznaczenia kolejnego przybliżenia $\mathbf{x}^{(i+1)}$ jest koszt mnożenia wektora $\mathbf{x}^{(i)}$ przez macierz **A**, wykonywanego przy obliczaniu wektora $\mathbf{r}^{(i)}$. Realizacja metody Czebyszewa wymaga pamiętania dwóch poprzednich przybliżeń $\mathbf{x}^{(i)}$ i $\mathbf{x}^{(i+1)}$, a więc co najmniej 3*n* miejsc w pamięci.

W praktyce do rozwiązywania wielkich układów liniowych często stosuje się połączenie metody Czebyszewa z jedną z metod gradientowych, omawianych w następnym rozdziale.

4.3.5. Metody gradientowe

Podobnie jak poprzednio zakładamy, że macierz **A** rozwiązywanego układu jest symetryczna i dodatnio określona.

Dla określenia rozważanej klasy metod rozwiązywania takich układów potrzebna jest nam norma $\|\mathbf{x}\|_{\mathrm{B}}$ zdefiniowana równością:

$$\|\mathbf{x}\|_{\mathrm{B}} = \sqrt{(\mathbf{Bx}, \mathbf{x})}, \text{gdzie } (\mathbf{Bx}, \mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{Bx}\ ,$$

natomiast **B** jest dowolną macierzą symetryczną, dodatnio określoną i taką, że **AB**=**BA** (jeśli **AB**=**BA** to macierz B jest ***macierzą komutującą*** z macierzą **A).**

W ***metodach gradientowych*** konstruuje się ciąg przybliżeń $\{\mathbf{x}^{(i)}\}$ rozwiązania $\mathbf{x} = \mathbf{A}\mathbf{x}^{(i)} - \mathbf{b}$ ze wzorów:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - c_i\mathbf{r}^{(i)}, \quad \mathbf{r}^{(i)} = \mathbf{A}\mathbf{x}^{(i)} - \mathbf{b}\,. \tag{4.69}$$

Współczynniki c_i dobiera się tak, aby zminimalizować błąd $\|\mathbf{x}^{(i+1)} - \mathbf{x}\|_{\mathrm{B}}$, tzn. $\|\mathbf{x}^{(i+1)} - \mathbf{x}\| = \inf_{\mathrm{c}} \|\mathbf{x}^{(i)} - \mathbf{x} - c\mathbf{r}^{(i+1)}\|_{\mathrm{B}}$. Minimalizacja ta ma charakter lokalny. W danym kroku dla przybliżenia $\mathbf{x}_{\mathrm{i}}$ szukamy jego najlepszej poprawki w kierunku wektora $\mathbf{r}^{(i)}$. Można sprawdzić, że:

$$c_i = \frac{(\mathbf{r}^{(i)}, \mathbf{B}(\mathbf{x}^{(i)} - \mathbf{x}))}{(\mathbf{r}^{(i)}, \mathbf{Br}^{(i)})}. \tag{4.70}$$

Tak określony współczynnik c_i potrafimy obliczyć jedynie dla pewnych macierzy **B**. Jest tak np. dla $\mathbf{B} = \mathbf{A}^p$, gdzie p jest liczbą naturalną. Wtedy bowiem zachodzi równość:

$$(\mathbf{r}^{(i)}, \mathbf{B}(\mathbf{x}^{(i)} - \mathbf{x})) = (\mathbf{r}^{(i)}, \mathbf{A}^{\text{p-1}}\mathbf{r}^{(i)}).$$

Dla **B**=**A** metoda (4.48)-(4.49) nosi nazwę ***metody najszybszego spadku***. W każdym jej kroku minimalizowana jest wielkość:

$$\left\|\mathbf{x}_{i+1} - \mathbf{x}\right\|_A = \sqrt{(\mathbf{A}(\mathbf{x}_{i+1} - \mathbf{x}), (\mathbf{x}_{i+1} - \mathbf{x}))} = \sqrt{(\mathbf{r}_{i+1}, \mathbf{x}_{i+1} - \mathbf{x})}\,.$$

W ogólnym przypadku dowodzi się, że:

$$\left\|\mathbf{x}^{(i+1)} - \mathbf{x}\right\|_B \le \left(\frac{\text{cond}(\mathbf{A}) - 1}{\text{cond}(\mathbf{A}) + 1}\right)\left\|\mathbf{x}^{i} - \mathbf{x}\right\|_B,$$

gdzie cond(**A**) oznacza ***liczbę uwarunkowania macierzy*** **A** definiowaną jako:

$$\text{cond}(\mathbf{A}) = \left\|\mathbf{A}\right\|\left\|\mathbf{A}^{-1}\right\|.$$

Metoda gradientowa może więc definiować ciąg przybliżeń $\{\mathbf{x}^{(i)}\}$ bardzo wolno zbieżny do rozwiązania **x**, dla zadań źle uwarunkowanych (duża liczba uwarunkowania). Wydaje się więc rzeczą naturalną, że do konstrukcji metod szybciej zbieżnych można dojść w podobny sposób, jak w przypadku metody Czebyszewa.

Rozważmy ponownie metody iteracyjne $\mathbf{x}^{(i)} - \mathbf{x} = W_i(\mathbf{A})(\mathbf{x}^{(0)} - \mathbf{x})$, gdzie W_i jest wielomianem stopnia co najwyżej *i,* spełniającym warunek $W_i(0) = 1$. Wybierając odpowiednie wielomiany W_i minimalizujące błąd $\left\|\mathbf{x}^{(i)} - \mathbf{x}\right\|_{\text{B}}$, otrzymujemy rekurencyjną definicję tworzonego ciągu $\{\mathbf{x}^{(i)}\}$:

$$\begin{aligned}
&\mathbf{z}^{(i)} = \mathbf{x}^{(i)} - c_i\mathbf{r}^{(i)},\\
&\mathbf{r}^{(i)} = \mathbf{A}\mathbf{x}^{(i)} - \mathbf{b},\\
&\mathbf{x}^{(i+1)} = \mathbf{z}^{(i)} - u_i\mathbf{y}^{(i)},\\
&\mathbf{y}^{(i)} = \mathbf{x}^{(i-1)} - \mathbf{z}^{(i)},
\end{aligned} \qquad (4.71)$$

gdzie:

$$c_i = \frac{\left(\mathbf{r}^{(i)}, \mathbf{B}(\mathbf{x}^{(i)} - \mathbf{x})\right)}{(\mathbf{r}^{(i)}, \mathbf{B}\mathbf{r}^{(i)})},$$

$$u_0 = 0,\ u_i = \frac{(\mathbf{y}^{(i)}, \mathbf{B}(\mathbf{z}^{(i)} - \mathbf{x}))}{(\mathbf{y}^{(i)}, \mathbf{B}\mathbf{y}^{(i)})} \text{ dla } i = 1,2,... \tag{4.72}$$

Każdy krok metody (4.71) składa się z dwóch etapów. Najpierw jednym krokiem metody gradientowej (4.69) wyznaczamy $\mathbf{z}^{(i)}$, czyli możliwie najlepiej poprawiamy przybliżenie $\mathbf{x}^{(i)}$ w kierunku wyznaczonym przez wektor $\mathbf{r}^{(i)}$. W drugim etapie postępujemy tak samo z $\mathbf{z}^{(i)}$, zmieniając kierunek poprawki na $\mathbf{y}^{(i)}$.

Zależnie od wyboru macierzy $\mathbf{B}$ definiuje się różne warianty omawianej metody. Przyjęcie $\mathbf{B} = \mathbf{A}^p$, gdzie p=0,1,2 wydaje się wyczerpywać wszystkie przypadki o praktycznym znaczeniu.

Dla $\mathbf{B} = \mathbf{A}^0 = \mathbf{I}$ metoda nosi nazwę ***metody minimalnych błędów***, gdyż wielkością minimalizowaną jest $\left\|\mathbf{x}^{(i)} - \mathbf{x}\right\|_2$. Odnosi się to tylko do przypadku układu **Ax**=**b**, gdzie $\mathbf{A} = \mathbf{M}^{\mathrm{T}}\mathbf{M}$, a $\mathbf{M}$ jest znaną macierzą nieosobliwą. Założenie takie jest potrzebne, aby móc wyznaczyć współczynniki c_i i u_i określone wzorami (4.72).

Dla $\mathbf{B} = \mathbf{A}^1$ otrzymujemy ***metodę sprzężonych gradientów*** (ang. *conjugate gradient*, C-G), minimalizującą $\left\|\mathbf{A}^{1/2}(\mathbf{x}^{(i)} - \mathbf{x})\right\|_2$. Metoda ta daje dokładne rozwiązanie po n iteracjach, ale jest numerycznie nie poprawna, a to powoduje, że bardzo szybko następuje liniowa zależność gradientów. Jednym ze sposobów radzenia sobie z tą dolegliwością jest stosowanie tzw. RESTART-u co kilka lub kilkanaście iteracji.

Wariant metody dla $\mathbf{B} = \mathbf{A}^2$, w którym minimalizowane są wektory residualne $\mathbf{r}^{(i)}$ (gdyż $\left\|\mathbf{x}^{(i)} - \mathbf{x}\right\|_{\mathrm{A}^2} = \left\|\mathbf{A}(\mathbf{x}^{(i)} - \mathbf{x})\right\|_2$), nazywany jest ***metodą minimalnych residuów***.

Metody gradientowe nie są polecane ze względu na numeryczną niestabilność. Dobre wyniki daje natomiast połączenie metod gradientowych z numerycznie stabilną metodą Czebyszewa.

4.4. Macierze specjalne

Wiele problemów inżynierskich prowadzi do rozwiązywania układów równań liniowych z tzw. macierzą rzadką.

Macierzą rzadką nazywamy macierz zawierającą dużo zer. Miarą rzadkości macierzy jest stosunek liczby jej elementów niezerowych do

ogólnej liczby elementów: *s=liczba_zer/liczba_elementów* Przykładami macierzy rzadkich są macierze wstęgowe, diagonalne, trójdiagonalne, trójkątne.

Wiele zagadnień (np. metody numeryczne służące do rozwiązywania równań różniczkowych cząstkowych) prowadzi do układów liniowych rzadkich, w których elementy niezerowe są rozmieszczone wzdłuż głównej przekątnej.

Ogólnie macierz **A** taką, że $a_{ij}=0$ jeżeli $j > i+p$ lub $i > j+q$ nazywa się ***macierzą wstęgową*** o szerokości wstęgi $w = p+q$. Liczba niezerowych elementów w dowolnym wierszu lub kolumnie macierzy **A** nie przewyższa w, a ogólna liczba niezerowych elementów jest mniejsza od $w \times n$, gdzie n oznacza stopień macierzy **A**.

Dla macierzy symetrycznej istnieje jeszcze inna definicja: szerokość wstęgi wynosi m wtedy i tylko wtedy, gdy $a_{ij}\neq 0$ jeżeli $|i-j|\leq m$. Jeżeli $w=0$, otrzymujemy macierz diagonalną, jeżeli $p=q=1$, otrzymujemy macierz trójdiagonalną.

Do rozwiązywania układów z macierzami rzadkimi można stosować zarówno metody omówione do tej pory (np. eliminację Gaussa), jak też metody korzystające w istotny sposób z rzadkości macierzy. Metody takie pozwalają rozwiązać układ równań z macierzą rzadką wykonując znacznie mniej działań arytmetycznych, niż w przypadku układu o tej samej liczbie równań z macierzą gęstą. Umożliwiają także wyznaczenie rozwiązania z większą dokładnością i przy oszczędniejszym wykorzystaniu pamięci komputera.

Technika macierzy rzadkich pozwala na:

- oszczędne gospodarowanie pamięcią przez zapamiętanie tylko niezerowych współczynników układu, ich pozycji w macierzy oraz minimum danych umożliwiających efektywne docieranie do tych elementów. Wówczas zajętość pamięci jest proporcjonalna tylko do liczby elementów niezerowych;
- wykonywanie operacji tylko na elementach niezerowych macierzy i wektora prawych stron, prowadzące do zmniejszenia liczby operacji w stosunku równym mierze rzadkości;
- skuteczny i szybki wybór elementów podstawowych przy zachowaniu rzadkości macierzy.

4.4.1. Reprezentacja macierzy w strukturach danych

Ortogonalne listy powiązane

Najważniejszym systemem reprezentacji macierzy w strukturach danych jest struktura ortogonalnych list powiązanych. System list powiązanych stanowi dwukierunkowo związany wskaźnikami system struktur. Każda struktura reprezentuje jeden element niezerowy. Zawiera ona następujące dane:

- *value* - wartość elementu niezerowego,
- *row* - numer wiersza elementu,
- *col* - numer kolumny elementu,
- *next_in_row* - wskaźnik do takiej samej struktury reprezentującej następny element niezerowy w wierszu,
- *next_in_col* - wskaźnik do takiej samej struktury reprezentującej następny element niezerowy w kolumnie.

Ortogonalny dwukierunkowy charakter list polega na tym, że od każdego elementu macierzy poprzez wskaźnik można dostać się do następnego elementu w wierszu oraz następnego elementu w kolumnie.

Aby znaleźć elementy pierwsze w wierszach lub kolumnach potrzebne są dodatkowo dwie tablice:

- *first_in_row*[*n*] - zawierająca wskaźniki na struktury w listach reprezentujące pierwsze elementy w każdym wierszu,
- *first_in_col*[*n*] - zawierająca wskaźniki na struktury w listach reprezentujące pierwsze elementy w każdej kolumnie.

Dodatkowo tworzy się tablicę *diag*[*n*] zawierającą wskaźniki na struktury reprezentujące elementy z głównej przekątnej macierzy. Wskaźnik zerowy oznacza brak elementu następnego w liście.

Wektory s,u (dla macierzy symetrycznych)

Dolny trójkąt macierzy symetrycznej **A** stopnia *n* zapamiętuje się wiersz po wierszu. W każdym wierszu pamięta się tylko elementy począwszy od pierwszego elementu niezerowego aż do głównej przekątnej (wraz z zerami jeżeli są w tej części wiersza). Zapamiętane elementy macierzy tworzą wektor $\mathbf{s}=(s_1,s_2,...,s_{Un})$. Wraz z nimi zapamiętuje się wektor wskaźników: $\mathbf{u}=(u_1,u_2,...,u_n)$, gdzie wartość u_i wskazuje na pozycję *i*-tego elementu w wektorze **s.**

Przykład 4.9.

Macierz symetryczną:

$$\mathbf{A} = \begin{bmatrix} \mathbf{[25]} & 3 & 0 & 0 & 0 \\ \mathbf{[\ 3} & \mathbf{21]} & 2 & 4 & 0 \\ 0 & \mathbf{[2} & \mathbf{23]} & 0 & 0 \\ 0 & \mathbf{[4} & \mathbf{0} & \mathbf{22]} & 1 \\ 0 & 0 & 0 & \mathbf{[1} & \mathbf{20]} \end{bmatrix},$$

zapamiętujemy w postaci wektorów **s** i **u**:

s = ([25], [3, 21], [2, 23], [4, 0, 22], [1, 20]),

u = (1, 3, 5, 8, 10).

$i = 1$ $i = 2$ $i = 3$ $i = 4$ $i = 5$

W tym przykładzie przyjęty sposób pamiętania macierzy **A** nie przyniósł żadnej oszczędności, ale inaczej byłoby np. dla macierzy 100 x 100 o 400 niezerowych elementach.

Graf

Strukturę macierzy symetrycznej rzadkiej można wyrazić za pomocą ***grafu***, w którym wierzchołki *i,j* są połączone łukiem wtedy i tylko wtedy, gdy $a_{ij} \neq 0$. Mówimy wtedy, że wierzchołki *i,j* są sąsiednie. Liczba łuków wychodzących z danego wierzchołka nazywa się stopniem. Permutacja oznacza zmianę symboli wierzchołków. Dla macierzy niesymetrycznej elementy a_{ij} i a_{ji} nie muszą być jednocześnie zerami. W takim przypadku trzeba posługiwać się grafem zorientowanym, w którym każdy łuk ma wyróżniony kierunek.

Wektory AN, JA, IA.

W pewnych zastosowaniach występują bardzo duże macierze niesymetryczne rzadkie o wysoce nieregularnym rozmieszczeniu niezerowych elementów. Można wtedy opisać strukturę danych macierzy rzadkiej **A** za pomocą trzech wektorów **AN, JA, IA.** Wektor **AN** zawiera niezerowe elementy z kolejnych wierszy (zer występujących między nimi nie trzeba pamiętać). Dla elementu **AN**(*k*) w **JA**(*k*) podaje się numer kolumny, w której ten element znajduje się w macierzy **A**. Natomiast **IA**(*i*) zawiera pozycję pierwszego elementu *i*-tego wiersza z macierzy **A** w tablicach **JA** i **AN**.

Przykład 4.10.

Macierz rzadką 6 *x* 6 przedstawimy za pomocą odpowiednich wektorów **AN**, **JA** i **IA**.

$$\mathbf{A}=\begin{bmatrix} 7 & 0 & -3 & 0 & -1 & 0 \\ 2 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ -3 & 0 & 0 & 5 & 0 & 0 \\ 0 & -1 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & -2 & 0 & 6 \end{bmatrix},$$

Odpowiednie wektory mają postać:
AN=(7, -3, -1, 2, 8, 1, -3, 5, -1, 4, -2, 6),
JA =(1, 3, 5, 1, 2, 3, 1, 4, 2, 5, 4, 6),
IA =(1, 4, 6, 7, 9, 11, 13).

4.4.2. Metody dokładne dla układów z macierzami rzadkimi

Metoda eliminacji Gaussa oraz metoda LU nie znalazły szerokiego zastosowania do rozwiązywania układów z macierzami rzadkimi. Spowodowane jest to przede wszystkim brakiem skutecznych metod wyboru elementu podstawowego, które z jednej strony zapewniałyby niezawodność i stabilność numeryczną, z drugiej zaś strony nie powodowałyby pojawienia się dużej liczby nowych elementów niezerowych. W niektórych przypadkach warto jednak skorzystać z metod stosowanych dla macierzy gęstych (np. dla układów z macierzami trójdiagonalnymi).

Układy o macierzach wstęgowych

Wiele zagadnień prowadzi do układów liniowych rzadkich, w których elementy niezerowe są rozmieszczone wzdłuż głównej przekątnej. Proces eliminacji Gaussa nie zmienia struktury macierzy wstegowej. Jeśli nie przestawia się wierszy ani kolumn to czynniki trójkątne $\mathbf{L}=(l_{ij})$ i $\mathbf{U}=(u_{ij})$ są macierzami wstęgowymi takimi, że:

$l_{ij}=0$, jeśli $j>i$ lub $i>j+q$,
$u_{ij}=0$, jeśli $j>i+p$ lub $i>j$.

Jeżeli dokonujemy częściowego wyboru elementu głównego, to w macierzy **L** szerokość wstęgi nie zmienia się, natomiast szerokość wstęgi w **U** będzie taka jak w **A**, czyli:

$u_{ij}=0$, jeśli $j>i+p+q$ lub $i>j$.

Rozwiązywanie układów o macierzach trójdiagonalnych

Częstym przypadkiem macierzy wstęgowej jest macierz, w której p=q=1, tzn. trójprzekątniowa (trójdiagonalna). Układ równań o takiej macierzy rozwiązuje się znacznie szybciej i prościej. Jeśli istnieje jej rozkład trójkątny, to można go przedstawić w postaci:

$$\begin{bmatrix} a_1 & c_1 & & & 0 \\ b_2 & a_2 & c_2 & & \\ & \cdots & \cdots & \cdots & \\ & & b_{n-1} & a_{n-1} & c_{n-1} \\ 0 & & & b_n & a_n \end{bmatrix} = \begin{bmatrix} 1 & & & & 0 \\ z_2 & 1 & & & \\ & & \cdots & & \\ & & z_{n-1} & 1 & \\ 0 & & & z_n & 1 \end{bmatrix} \begin{bmatrix} y_1 & c_1 & & & 0 \\ & y_2 & c_2 & & \\ & & \cdots & \cdots & \\ & & & y_{n-1} & c_{n-1} \\ 0 & & & & y_n \end{bmatrix}$$

gdzie:

$y_1 = a_1,$

$z_k = b_k / y_{k-1}$ (jesli tylko $y_{k-1} \neq 0$),

$y_k = a_k - z_k c_{k-1}, k = 2,3,...,n.$

Jeśli $y_{k-1} = 0$ dla pewnego k to musimy skorzystać z omówionej już metody Gaussa-Jordana lub LU.

Następnie rozwiązujemy układ **Ax**=**f** (gdzie **f** jest znanym wektorem prawej strony) stosując podstawienie wprzód i wstecz :

$g_1 = f_1, g_i = f_i - z_i g_{i-1}, i = 2,3,...,n,$

$x_n = g_n / y_n, x_i = (g_i - c_i x_{i+1}) / y_i, i = n-1,...,1.$

Łączna liczba działań arytmetycznych wynosi tu *tylko* 3(n-1) dodawań i mnożeń, i 2n-2 dzieleń.

4.4.3. Rozwiązywanie układów równań liniowych - wnioski

W przypadku macierzy pełnych, liczba obliczeń potrzebna do uzyskania rozwiązania metodą iteracyjną jest zwykle znacznie większa niż przy stosowaniu metod dokładnych. Jednak w przypadku macierzy rzadkich metody iteracyjne mogą być lepsze niż metody dokładne. Nie nakład obliczeń jednak decyduje o tym (chociaż zwykle jest mniejszy), lecz fakt, że podczas obliczeń nie zmieniamy położenia elementów macierzy **A** układu równań **Ax=b**, a zatem zachowujemy jej rzadką strukturę. Możemy wówczas przyjąć jedną z podanych metod zapamiętywania macierzy **A** oraz korzystać z prostych algorytmów obliczeniowych. Nie możemy jednak kategorycznie stwierdzić, że w przypadku macierzy

rzadkich metody iteracyjne są bardziej wskazane, ponieważ w szczególnych przypadkach liczba iteracji może być bardzo duża.

4.5. Przykłady obliczeniowe

Przykład 4.11.

Znaleźć macierz odwrotną $\mathbf{A}^{-1}$ macierzy:

$$A = \begin{bmatrix} 2 & 1 & 2 \\ 1 & 2 & 3 \\ 4 & 1 & 2 \end{bmatrix}$$

a) rozwiązując układ **AX=I** z częściowym wyborem elementów głównych,
b) znajdując rozkład trójkątny i korzystając z wzoru $\mathbf{A}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1}$.

Po obliczeniu macierzy odwrotnej metodą **AX=I** szukana macierz odwrotna ma postać:

$$\begin{bmatrix} -0.5 & -4{,}547473508910^{-13} & 0.5 \\ -5 & 2 & 2 \\ 3.5 & -1 & -1.5 \end{bmatrix}.$$

Po skorzystaniu z metody $\mathbf{A^{-1} = U^{-1} L^{-1}}$ szukana macierz odwrotna ma postać:

$$\begin{bmatrix} -0.5 & 0 & 0.5 \\ -5 & 2 & 2 \\ 3.5 & -1 & -1.5 \end{bmatrix}.$$

W obydwu metodach wywoływana jest ta sama procedura eliminacji Gaussa z częściowym wyborem elementów głównych, ale jak widać dokładność metody pierwszej (**A X = I**) jest nieco mniejsza.

Przykład 4.12.

a) Pokazać, że macierz symetryczna:

$$\mathbf{A}=\begin{bmatrix}10 & 7 & 8 & 7\\ 7 & 5 & 6 & 5\\ 8 & 6 & 10 & 9\\ 7 & 5 & 9 & 10\end{bmatrix}$$ jest dodatnio określona,

b) Wyznaczyć macierz trójkątną **R** taką, że $\mathbf{A}=\mathbf{R}^{\mathrm{T}}\mathbf{R}$.

Ad a)

Z symetrycznej wersji eliminacji Gaussa:

$$m_{ik}=\frac{a_{ki}^{(k)}}{a_{kk}^{(k)}},\quad a_{ij}^{(k+1)}=a_{ij}^{(k)}-m_{ik}a_{kj}^{(k)}\ \ (i=k+1,\ k+2,\ \ldots,n\ ,\ j=i,\ i+1,\ \ldots,\ n),$$

otrzymujemy kolejno następujące zredukowane macierze (wystarcza tylko przekształcać ich części górne trójkątne):

$$\begin{bmatrix}10 & 7 & 8 & 7\\ & 5 & 6 & 5\\ & & 10 & 9\\ & & & 10\end{bmatrix},\begin{bmatrix}0.1 & 0.4 & 0.1\\ & 3.6 & 3.4\\ & & 5.1\end{bmatrix},\begin{bmatrix}2 & 3\\ & 5\end{bmatrix},\begin{bmatrix}0.5\end{bmatrix}.$$

Ponieważ elementy główne: 10; 0.1; 2; 0.5 są dodatnie, więc macierz **A** jest dodatnio określona.

Ad b)

Z twierdzenia o rozkładzie trójkątnym dla macierzy **A** symetrycznej dodatnio określonej istnieje jedyna macierz trójkątna górna **R** o dodatnich elementach na głównej przekątnej i taka, że $\mathbf{A}=\mathbf{R}^{\mathrm{T}}\mathbf{R}$. Z twierdzenia o rozkładzie LU wynika, że:

A=**LU**,

gdzie $u_{11}=a_{11}>0$ i $u_{kk}=\dfrac{\det(A_k)}{\det(A_{k-1})}>0$, ($k$=2,3, ... ,$n$).

Wprowadzając macierz przekątniową $\mathbf{D}=\operatorname{diag}(u_{11},u_{22},\ldots,u_{nn})$, możemy napisać rozkład:

$\mathbf{A}=\mathbf{LDD}^{-1}\mathbf{U}=\mathbf{LDU}'$,

gdzie $\mathbf{U}'=\mathbf{D}^{-1}\mathbf{U}$, a macierze **L** i **U** są trójkątne, mają jedynki na głównej przekątnej i są jednoznacznie określone.

Z symetrii **A** wynika, że:

$$\mathbf{A} = \mathbf{A}^{\mathrm{T}} = (\mathbf{U}')^{\mathrm{T}}\mathbf{D}\mathbf{L}^{\mathrm{T}},$$

czyli:

$$\mathbf{L}^{\mathrm{T}} = \mathbf{U}' = \mathbf{D}^{-1}\mathbf{U}.$$

Przyjmując $\mathbf{R} = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}$, gdzie macierz przekątniowa $\mathbf{D}^{-\frac{1}{2}}$ ma dodatnie elementy $u_{kk}^{-\frac{1}{2}}$, otrzymujemy:

$$\mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{U}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{U} = \mathbf{L}\mathbf{U} = \mathbf{A}.$$

W symetrycznej eliminacji Gaussa nie trzeba pamiętać mnożników. Układ $\mathbf{A}\mathbf{x} = \mathbf{b}$ w tym przypadku rozkłada się na dwa układy trójkątne: $\mathbf{U}^{\mathrm{T}}\mathbf{y} = \mathbf{b}$ i $\mathbf{U}\mathbf{x} = \mathbf{D}\mathbf{y}$ (unikamy w ten sposób pierwiastkowania). Wobec tego $\mathbf{R} = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}$, jest macierzą trójkątną górną o dodatnich elementach na głównej przekątnej taką, że $\mathbf{A}=\mathbf{R}^{\mathrm{T}}\mathbf{R}$.

$$\mathbf{R} = \begin{bmatrix} 10 & 7 & 8 & 7 \\ 0 & 0.1 & 0.4 & 0.1 \\ 0 & 0 & 2 & 3 \\ 0 & 0 & 0 & 0.5 \end{bmatrix};$$

$$\mathbf{D}^{-\frac{1}{2}} = \mathrm{diag}(10^{-\frac{1}{2}}, 10^{\frac{1}{2}}, 2^{-\frac{1}{2}}, 2^{\frac{1}{2}}).$$

Przykład 4.13.

W układzie równań $\mathbf{A}\mathbf{x}=\mathbf{b}$ dane są:

$$\mathbf{A} = \begin{bmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0.217 \\ 0.254 \end{bmatrix}.$$

Dokładnym rozwiązaniem jest wektor:

$$\mathbf{x}^{\mathrm{T}} = [1, -1].$$

Dodatkowo dane są dwa rozwiązania przybliżone:

$$\mathbf{x}_1^{\mathrm{T}} = [0.999, -1.001], \quad \mathbf{x}_2^{\mathrm{T}} = [0.341, -0.087].$$

Obliczyć wartości residuów $\mathbf{r}(\mathbf{x}_1)$, $\mathbf{r}(\mathbf{x}_2)$ oraz wyznaczyć wskaźnik uwarunkowania cond(**A**), korzystając z normy maksimum oraz z macierzy:

$$\mathbf{A}^{-1} = \begin{bmatrix} 659000 & -563000 \\ -913000 & 780000 \end{bmatrix}.$$

Jeżeli **x'** jest obliczonym rozwiązaniem układu **Ax = b**, to wektor residuum ma postać:

r = b – Ax'.

Residuum $\mathbf{r_1} = \mathbf{b} - \mathbf{Ax_1}'$ dla rozwiązania przybliżonego $\mathbf{x}_1$:

$$\mathbf{r}_1 = \begin{bmatrix} 0.217 \\ 0.254 \end{bmatrix} - \begin{bmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{bmatrix} \begin{bmatrix} 0.999 \\ -1.001 \end{bmatrix} =$$

$$= \begin{bmatrix} 0.217 \\ 0.254 \end{bmatrix} - \begin{bmatrix} 0.780 \cdot 0.999 + 0.563 \cdot (-1.001) \\ 0.913 \cdot 0.999 + 0.659 \cdot (-1.001) \end{bmatrix} = \begin{bmatrix} 1.24 \cdot 10^{-3} \\ 1.57 \cdot 10^{-3} \end{bmatrix}.$$

Residuum $\mathbf{r_2} = \mathbf{b} - \mathbf{Ax_2}'$ dla rozwiązania przybliżonego $\mathbf{x}_2$:

$$\mathbf{r}_2 = \begin{bmatrix} 0.217 \\ 0.254 \end{bmatrix} - \begin{bmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{bmatrix} \begin{bmatrix} 0.341 \\ -0.087 \end{bmatrix} =$$

$$= \begin{bmatrix} 0.217 \\ 0.254 \end{bmatrix} - \begin{bmatrix} 0.780 \cdot 0.341 + 0.563 \cdot (-0.087) \\ 0.913 \cdot 0.341 + 0.659 \cdot (-0.087) \end{bmatrix} = \begin{bmatrix} 1 \cdot 10^{-6} \\ 0 \end{bmatrix}.$$

Z powyższego wynika, że nie zawsze mniejsza wartość residuum odpowiada lepszemu rozwiązaniu. Może się tak zdarzyć w przypadku macierzy o dużym wskaźniku uwarunkowania (liczba warunkowa), tak jak ma to miejsce w analizowanym przypadku.

Wskaźnik uwarunkowania cond(**A**) dla macierzy określa się wzorem (patrz rozdział 4.2.5) $\mathrm{cond}(\mathbf{A}) = \|\mathbf{A}\| \, \|\mathbf{A}^{-1}\|$. W przykładzie rozważamy normę maksimum, definiowaną jako:

$$\| \mathbf{A} \|_{\infty} = \max_{1 \le i \le n} \sum_{j=1}^{n} | a_{ij} | .$$

Normy macierzy **A** oraz macierzy $\mathbf{A}^{-1}$ są równe:

$$\| \mathbf{A} \| = \max_{1 \le i \le n} \sum_{j=1}^{n} | a_{ij} | = (| 0.780 | + | 563000 |; | 91300 | + | 0.659 |) =$$

$$= \max(1.343; 1.572) = 1.572$$

$$\| \mathbf{A}^{-1} \| = \max_{1 \le i \le n} \sum_{j=1}^{n} | a_{ij} | = (| 659000 | + | 563000 |; | 913000 | + | 780000 |) =$$

$$= \max(1.343; 1693000) = 1693000$$

a wskaźnik uwarunkowania cond(**A**):

$$\mathrm{cond}(\mathbf{A}) = 1.572 \cdot 1693000 = 2661396 .$$

Przykład 4.14.

Obliczyć przybliżone rozwiązanie układu **Ax**=**b**, gdzie:

$$\mathbf{A}=\begin{bmatrix} 0.95 & 0.07 & 0 & 0 & 0.05 & 0.01 \\ 0.07 & 0.95 & 0.07 & 0 & 0 & 0.04 \\ 0 & 0.007 & 0.95 & 0.06 & 0 & 0 \\ 0 & 0 & 0.06 & 0.95 & 0.06 & 0 \\ 0.05 & 0 & 0 & 0.06 & 0.95 & 0.06 \\ 0.01 & 0.04 & 0 & 0 & 0.06 & 0.95 \end{bmatrix},\ \mathbf{b}=\begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}.$$

Obliczenia wykonać dla kilku wybranych metod iteracyjnych. Jako kryterium końca obliczeń przyjąć: $\|r_k\| \le \varepsilon$.

Obliczenia przeprowadzono dla następujących dokładności ε: 0.001, 0.0001, 0.00001, 0.000001, przyjmując początkowe przybliżenia rozwiązania $\mathbf{x}^{(0)} = [x_1, x_2, \ldots, x_6]$ równe kolejno:

[0,0,0,0,0,0], [10,10,10,10,10,10], [100,...,100], [10000,...,10000].

Tabela 4.1. przedstawia porównanie wybranych metod pod względem liczby koniecznych iteracji przy danej dokładności bezwzględnej i danych początkowych przybliżeniach rozwiązania.

Na podstawie tabeli 4.1, dla rozpatrywanej macierzy **A** najszybciej zbieżna jest metoda Czebyszewa. Metody Jacobiego i Gaussa-Seidela są porównywalne, przy czym niewielką przewagę ma metoda Gaussa-Seidela.

Tabela 4.1. Wyniki otrzymane w przykładzie 4.14

Początkowe przybliżenie $\mathbf{x}^{(0)}$	**Liczba iteracji w metodzie**		
	Jacobiego	Gaussa-Seidela	Czebyszewa
bezwzględna dokładność ε= 0,001			
[0,0,0,0,0,0]	5	5	2
[10,10,...,10]	7	6	2
[100,100,...,100]	8	7	2
[1000,...,1000]	9	9	3
bezwzględna dokładność ε= 0,0001			
[0,0,0,0,0,0]	6	6	2
[10,10,...,10]	8	7	2
[100,100,...,100]	9	9	3
[1000,...,1000]	10	10	3
bezwzględna dokładność ε= 0,00001			
[0,0,0,0,0,0]	7	7	2
[10,10,...,10]	9	9	3
[100,100,...,100]	10	10	3
[1000,...,1000]	11	11	3
bezwzględna dokładność ε= 0,000001			
[0,0,0,0,0,0]	9	9	3
[10,10,...,10]	10	10	3
[100,100,...,100]	11	11	3
[1000,...,1000]	13	12	3

Przykład 4.15.

Dana jest macierz **A** i wektor **b** postaci jak poniżej**.**:

$$\mathbf{A}=\begin{bmatrix} 3 & -1 & 0 & 0 & 0 & -1 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & 0 \\ 0 & -1 & 0 & -1 & 3 & -1 \\ -1 & 0 & 0 & 0 & -1 & 3 \end{bmatrix}, \mathbf{b}=\begin{bmatrix} 1 \\ 2 \\ 3 \\ 0 \\ 2 \\ 3 \end{bmatrix}.$$

Porównać zbieżność metod Jacobiego i Gaussa-Seidela w zależności od dokładności obliczeń. Jako kryterium końca obliczeń przyjąć: $\|\mathbf{r}_k\| \le 1$.

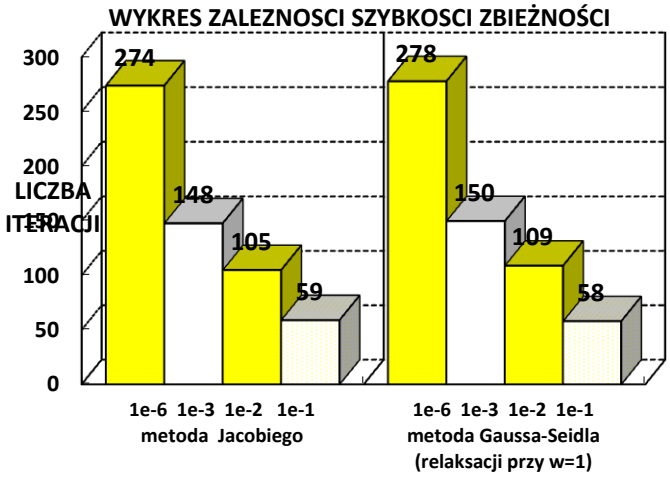
Rozwiązanie równania **Ax = b** jest następujące:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 2.6339285714 \\ 3.7857142857 \\ 3.0089285714 \\ 2.24107142286 \\ 3.7142857143 \\ 3.116071286 \end{bmatrix}$$

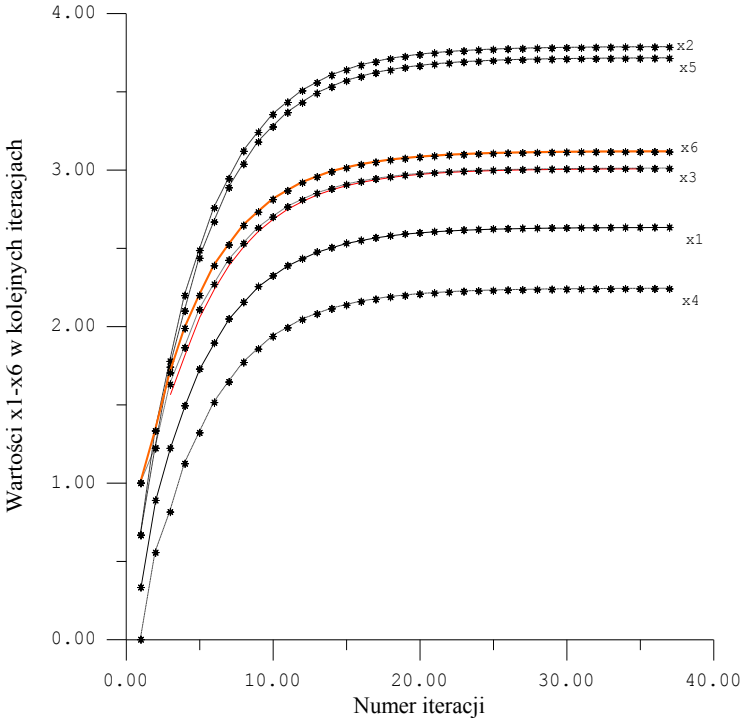
W celu lepszego zobrazowania różnic szybkości zbieżności metod iteracyjnych otrzymane wyniki zestawiono w tabeli 4.2 i na wykresach (rysunki 4.1 - 4.2).

Na podstawie przeprowadzonych obliczeń stwierdzono, że:

- największa co do wartości bezwzględnej wartość własna analizowanej macierzy wynosi 0.85292 i jest mniejsza od jedności a tym samym spełnia warunek konieczny i dostateczny zbieżności metod iteracyjnych;
- największa co do wartości bezwzględnej norma macierzy wynosi 0.9265 i ponieważ jej moduł jest mniejszy od jedności więc warunek wystarczający zbieżności metod iteracyjnych jest również spełniony;
- szybsza zbieżność metody Gaussa-Seidela jest prawdą w przypadku, gdy punkt startowy $\mathbf{x}^{(0)}$ jest w pobliżu rozwiązania, w przeciwnym razie szybciej zbieżna jest metoda Jacobiego;
- dla osiągnięcia dziesięciokrotnie większej dokładności potrzeba około dwa razy większej liczby iteracji;
- różnice w szybkości zbieżności poszczególnych metod są minimalne szczególnie przy obliczeniach z dużą dokładnością ε.



Rys. 4.1. Zależność liczby iteracji od dokładności obliczeń dla metod Jacobiego i Gaussa-Seidela



Rys. 4.2. Proces zbieżności dla metod iteracyjnych

Tabela 4.2. Wyniki otrzymane w przykładzie 4.15

Dokładność *eps*	Wektor startowy $x^{(0)}$	Liczba iteracji potrzebna do osiągnięcia dokładności ε metodą	
		Jacobiego	Gaussa-Seidela
0.1	[0,0,0,0,0,0]	6	5
	[10,10,...,10]	6	6
	[100,100,...,100]	18	18
	[1000,...,1000]	29	29
	suma iter.	59	58
0.01	[0,0,0,0,0,0]	16	15
	[10,10,...,10]	17	20
	[100,100,...,100]	31	32
	[1000,...,1000]	41	42
	suma iter.	105	109
0.001	[0,0,0,0,0,0]	26	25
	[10,10,...,10]	29	30
	[100,100,...,100]	41	42
	[1000,...,1000]	52	53
	suma iter.	148	150
10^{-6}	[0,0,0,0,0,0]	58	57
	[10,10,...,10]	60	62
	[100,100,...,100]	73	74
	[1000,...,1000]	83	85
	suma iter.	274	278

Przykład 4.16.

Przedmiotem tego zadania jest wyznaczenie macierzy **B** z równania macierzowego **AB** = **C,** gdzie macierz **A** jest macierzą kwadratową *n* x *n,* macierze **B** i **C** mogą być macierzami prostokątnymi *n* x *m*. Należy znaleźć algorytm wyznaczenia macierzy **B** przy wykorzystaniu eliminacji Gaussa i przeanalizować jego własności.

Klasyczny algorytm eliminacji Gaussa służy do znajdowania rozwiązania układu równań liniowych **Ax=b,** czyli wyznacza wektor rozwiązań **x** przy danej macierzy kwadratowej **A** i wektorze wyrazów wolnych **b**.

W przypadku równania **AB** = **C**, aby wyznaczyć macierz **B** stosując klasyczne zasady rachunku macierzowego, należałoby rozwiązać następujące równanie: $\mathbf{B} = \mathbf{A}^{-1}\mathbf{C}$. Na ogół jest to kłopotliwe i czasochłonne

ze względu na konieczność dokonania odwrócenia macierzy **A**. Niezależnie od zastosowanego algorytmu odwracania macierzy, należy dodatkowo dokonać mnożenia macierzy odwróconej oraz macierzy **C**. Liczba wykonanych podstawowych operacji matematycznych jest więc dosyć duża. Analizując sposób mnożenia dwóch macierzy możemy zauważyć interesującą nas zależność: tj. każda pojedyncza kolumna macierzy **C** jest wynikiem mnożenia macierzy **A** przez odpowiadającą kolumnę macierzy **B**, a pozostałe kolumny nie wpływają na jej postać. Praktycznie, mnożenia dwóch macierzy można dokonać w *m* niezależnych procesach (gdzie *m* jest liczbą kolumn macierzy **B** i **C**). Odwracając zagadnienie: znalezienie dowolnej kolumny macierzy **B** wymaga rozwiązania układu równań liniowych z prawą stroną równą odpowiadającej kolumnie w macierzy **C**. Przypadek ten doskonale można wykorzystać do eliminacji Gaussa: rozwiązując *m* układów równań liniowych otrzymamy *m* wektorów odpowiedzi, które będą szukanymi kolumnami macierzy **B**.

Najprostszym rozwiązaniem zagadnienia byłoby *m*-krotne wywołanie eliminacji Gaussa ze zmieniającym się wektorem wyrazów wolnych **b**. Takie postawienie problemu zaprzeczałoby wcześniejszym rozważaniom na temat ilości dokonywanych operacji matematycznych. Podstawą eliminacji Gaussa jest przekształcenie macierzy **A** do równoważnej macierzy górnej i przekształcenie to jest niezależne od prawej strony równania. Natomiast *m*-krotne wywołanie eliminacji powodowałoby *m*-krotne wykonywanie tych samych obliczeń przy przekształcaniu macierzy **A**. Dlatego w algorytmie podczas eliminacji Gaussa należy modyfikować całą macierz **C**, dzięki czemu liczba działań jest minimalna.

W przypadku wystąpienia elementu zerowego na przekątnej macierzy **A** podczas eliminacji, należy dokonać przestawienia wiersza z pierwszym napotkanym w kolumnie elementem niezerowym. Nie wpływa to na rozwiązanie (tak jak w zwykłym układzie równań liniowych przestawianie wierszy nie ma wpływu na wynik).

Analiza liczby działań elementarnych algorytmu

Pod terminem „działania elementarne" należy rozumieć dodawanie, odejmowanie, mnożenie i dzielenie tj. działania elementarne z punktu widzenia matematyki, a nie procesora. Przedstawiona analiza jest więc analizą ogólną, ale daje pewien obraz skomplikowania algorytmu.

W przykładzie tym przedstawiona zostanie analiza algorytmu w porównaniu do podstawowego algorytmu mnożenia dwóch macierzy.

Wybór ten wydaje się być odpowiedni, gdyż:

- mnożenie macierzy jest łatwe do wykonania i łatwo będzie porównać wyniki ilościowe działań;
- możliwe będzie wyciągnięcie wniosków co do skuteczności algorytmu w porównaniu ze wspomnianym klasycznym sposobem wyznaczania macierzy **B** tj. według wzoru $\mathbf{B} = \mathbf{A}^{-1}\,\mathbf{C}$.

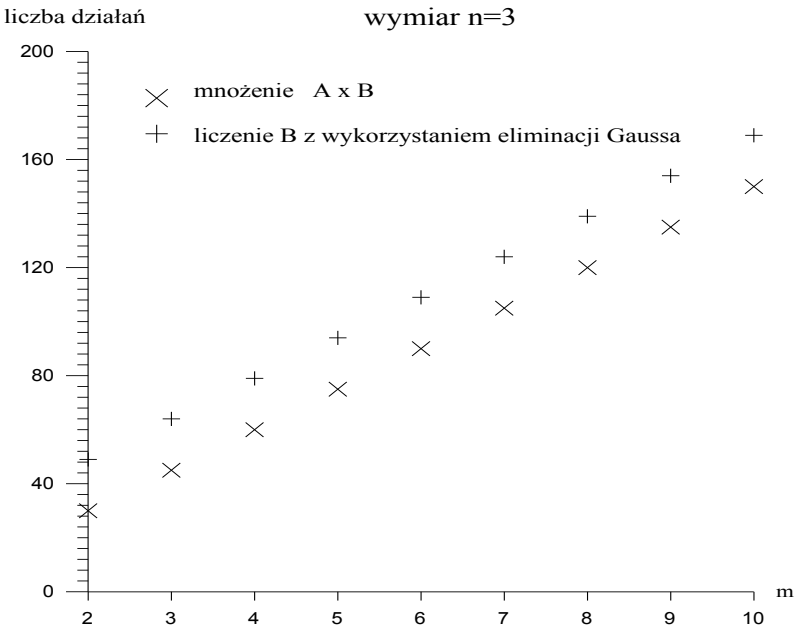
Po dokładnej analizie wykonywanych działań przeprowadzonej za pomocą odpowiedniego programu, okazuje się, że dla macierzy $\mathbf{A}_{nxn}$, $\mathbf{B}_{nxm}$, $\mathbf{C}_{nxm}$:

- mnożenie **A B** wymaga wykonania: n^2m elementarnych mnożeń oraz $(n^2-n)m$ elementarnych dodawań;
- wyznaczenie **B** przy wykorzystaniu eliminacji Gaussa wymaga wykonania $mn+(n^2-n)/2$ elementarnych dzieleń, $(n^2-n)(3m+n+1)/3$ elementarnych mnożeń oraz $(n^2-n)(3m+n+1)/3$ elementarnych odejmowań.

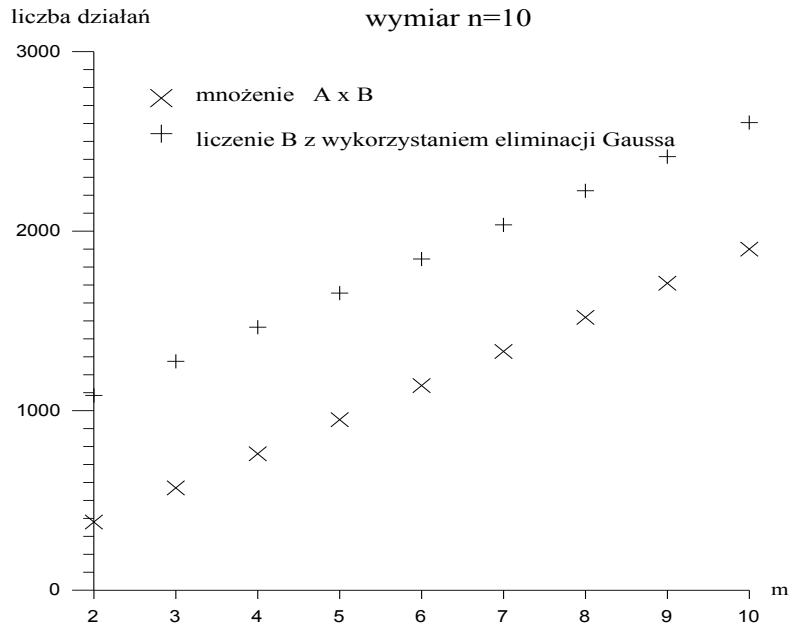
W celu ułatwienia porównania, na rysunkach 4.3 i 4.4 przeprowadzono graficzną analizę liczby działań (suma działań bez rozróżniania mnożeń, dodawań i odejmowań) w zależności od m przy stałym n, lub w zależności od n przy stałym m.

Wykresy na rysunku 4.3a i 4.3b (przy stałym n) wykazują, że dla danego n różnica działań jest w przybliżeniu stała, stąd dla dużych wartości m różnica ta procentowo nie jest duża (różnica ta rośnie wraz z wartością n).

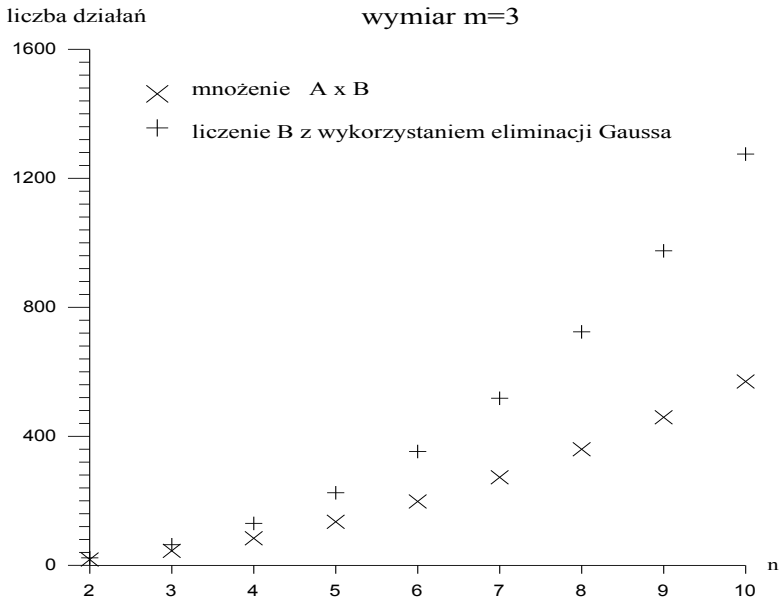
Wykresy na rysunku 4.4a i 4.4b (przy stałym m) wykazują, że dla małego m (np. 3), przy większych wartościach n różnica staje się duża, bardziej opłacalny jest więc przypadek z większą wartością m (np. 10).



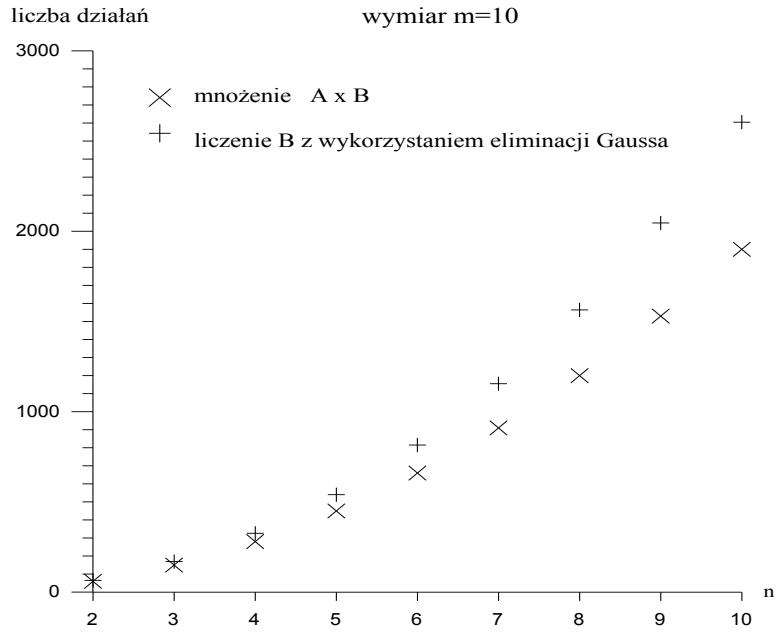
Rys. 4.3a. Porównanie efektywności metod rozwiązywania równania AB=C dla n=3



Rys. 4.3b. Porównanie efektywności metod rozwiązywania równania **AB**=**C** dla n =10



Rys. 4.4a. Porównanie efektywności metod rozwiązywania równania **AB**=**C** dla *m*=3



Rys. 4.4b. Porównanie efektywności metod rozwiązywania równania **AB**=**C** dla *m*=10

Reasumując, zaprezentowane wykresy dowodzą przydatności algorytmu pod względem liczby działań dla większych wartości *m*, aczkolwiek dla innych wartości liczba działań także nie jest zbyt duża.

Każde działanie arytmetyczne obarczone jest błędem zaokrąglenia wyniku, ale błąd względny dla wszystkich działań jest taki sam. Niebezpieczeństwo tkwi w odejmowaniu liczb zbliżonych do siebie – następuje redukcja cyfr najbardziej znaczących. Dodatkowo problem polega na tym, że elementy przez które dzielimy są obarczone dużym błędem wytworzonym we wcześniejszych iteracjach (zobacz przykład 1.9). Pewną poprawę dokładności można uzyskać stosując znany w eliminacji Gaussa wybór elementu podstawowego.

Przykład 4.16

Wyprowadzić wzory eliminacji Gaussa dla macierzy pięciodiagonalnej. Dla macierzy pięciodiagonalnej **A**:

$$\mathbf{A}=\begin{bmatrix} c_1 & d_1 & e_1 & & & & \\ b_2 & c_2 & d_2 & e_2 & & & \\ a_3 & b_3 & c_3 & d_3 & e_3 & & \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ & & a_{n-2} & b_{n-2} & c_{n-2} & d_{n-2} & e_{n-2} \\ & & & a_{n-1} & b_{n-1} & c_{n-1} & d_{n-1} \\ & & & & a_n & b_n & c_n \end{bmatrix}$$

rozkład **LU** jest postaci:

$$\mathbf{L}=\begin{bmatrix} 1 & & & & & & \\ l_2 & 1 & & & & & \\ m_3 & l_3 & 1 & & & & \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ & & m_{n-2} & l_{n-2} & 1 & & \\ & & & m_{n-1} & l_{n-1} & 1 & \\ & & & & m_n & l_n & 1 \end{bmatrix},\qquad \mathbf{U}=\begin{bmatrix} u_1 & w_1 & e_1 & & & & \\ & u_2 & w_2 & e_2 & & & \\ & & u_3 & w_3 & e_3 & & \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ & & & & u_{n-2} & w_{n-2} & e_{n-2} \\ & & & & & u_{n-1} & w_{n-1} \\ & & & & & & u_n \end{bmatrix}$$

gdzie:

$$u_1 = c_1,\ w_1 = d_1,$$

$$l_2 = \frac{b_2}{u_1},$$

$$u_2 = c_2 - l_2 \cdot w_1,$$
$$w_2 = d_2 - l_2 \cdot e_1,$$

$$m_i = \frac{a_i}{u_{i-2}},$$
$$l_i = \frac{b_i - m_i \cdot w_{i-2}}{u_{i-1}},$$
$$u_i = c_i - m_i \cdot e_{i-2} - l_i \cdot w_{i-1},$$
$$w_i = d_i - l_i \cdot e_{i-1} \quad \text{dla } i = 3,4,...,n.$$

Z rozwiązania dwu układów równań **Ly = f** i **Ux = y (Ax=f -> LUx=f)** otrzymujemy:

$$y_1 = f_1,$$
$$y_2 = f_2 - l_2 \cdot y_1,$$
$$y_i = f_i - l_i \cdot y_{i-1} - m_i \cdot y_{i-2} \quad \text{dla } i = 3,4,...,n$$

oraz:

$$x_n = \frac{y_n}{u_n},$$
$$x_{n-1} = \frac{(y_{n-1} - w_{n-1} \cdot x_n)}{u_{n-1}},$$
$$x_i = \frac{(y_i - w_i \cdot x_{i+1} - e_i \cdot x_{i+2})}{u_i} \qquad \text{dla} \quad i = n - 2, n - 3,...,1.$$

4.6. Metoda SVD rozwiązywania układów równań nadokreślonych

W obliczeniach praktycznych często pojawia się problem rozwiązania nadokreślonego układu równań liniowych:

$$\mathbf{A}\,\mathbf{x} = \mathbf{b}, \tag{4.73}$$

gdzie:

A - macierz *m* x *n*,

$\mathbf{b} = \left[b_1, b_2, ..., b_m\right]^{\mathrm{T}}$ - wektor prawej strony równania,

$\mathbf{x} = \left[x_1, x_2, ..., x_n\right]^{\mathrm{T}}$ - szukane rozwiązanie,

który w wyniku błędów pomiarowych może być również układem sprzecznym. Jeden ze sposobów rozwiązywania takiego układu bazuje na omówionej wcześniej metodzie Householdera (rozdział 4.2.5).

Sposobem rozwiązania takiego problemu jest znalezienie wektora $\mathbf{x}^*$, o możliwie małej normie euklidesowej, który dla zadanej macierzy **A** i wektora **b** minimalizuje normę euklidesową wektora residualnego:

r=**b**-**Ax**,

tzn.

$$\left\|\mathbf{b}-\mathbf{A}\mathbf{x}\right\|_2 = \min_{\mathbf{y}\in R^n}\left\|\mathbf{b}-\mathbf{A}\mathbf{y}\right\| \ , \quad \left\|\mathbf{x}^*\right\|_2 = \min\left\|\mathbf{x}\right\|_2, \tag{4.74}$$

gdzie ostatnie minimum brane jest po wszystkich wektorach **x** spełniających poprzednią równość. Jest to tzw. liniowe zadanie najmniejszych kwadratów (LZNK). Do LZNK prowadzi wiele różnych problemów, przede wszystkim aproksymacyjnych.

Przy wyznaczaniu postaci analitycznej rozwiązania liniowego zadania najmniejszych kwadratów i badaniu jego własności, korzystamy z twierdzenia o rozkładzie dowolnej macierzy prostokątnej na iloczyn macierzy ortogonalnej, diagonalnej i ortogonalnej. Mówi ono, że dla dowolnej macierzy $\mathbf{A} \in \mathrm{R}^{\mathrm{m\times n}}(\mathrm{m} \geq \mathrm{n})$ istnieją macierze ortogonalne $\mathbf{U} \in \mathrm{R}^{\mathrm{m\times m}}$ i $\mathbf{V} \in \mathrm{R}^{\mathrm{n\times n}}$ takie, że:

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}, \tag{4.75}$$

gdzie:

$$\mathbf{S} = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ . & . & . & . \\ 0 & 0 & 0 & \sigma_n \\ . & . & . & . \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \in R_{mxn},$$

$\sigma_1 \geq \sigma_2 ... \geq \sigma_k > \sigma_{k+1} = \sigma_{k+2} = ... = \sigma_n = 0$,

k jest rzędem macierzy **A**.

Wielkości σ_i nazywamy ***wartościami szczególnymi* (*osobliwymi*)** macierzy **A** , a rozkład (4.75) ***rozkładem według wartości szczególnych*** (ang. *singular value decomposition*, SVD). Pierwiastki σ są pierwiastkami wartości własnych macierzy $\mathbf{A}^T\mathbf{A}$, a kolumny macierzy **V** odpowiadającymi im ortonormalnymi wektorami własnymi tej macierzy. Z kolei kolumny **U** są wektorami własnymi $\mathbf{A}\mathbf{A}^T$. Widzimy stąd, że wartości szczególne są określone jednoznacznie, natomiast macierze **U** i **V** nie.

Korzystając z wartości szczególnych macierzy **A**, jej liczbę warunkową (patrz rozdział 4.2.5) można obliczyć ze wzoru:

$$cond(\mathbf{A}) = \frac{\sigma_1}{\sigma_k}.$$

Znając rozkład (4.75) można łatwo wyznaczyć rozwiązanie LZNK:

$$\mathbf{x}^* = \mathbf{A}^+\mathbf{b}, \tag{4.76}$$

gdzie:

$\mathbf{A}^+ = \mathbf{V}\mathbf{S}^+\mathbf{U}^T$ nazywana jest ***macierzą pseudoodwrotną*** do **A** (lub czasami macierzą odwrotną w sensie Moore'a -Penrose'a),

$$S^+ = \text{diag}\left(\frac{1}{\sigma_1},...,\frac{1}{\sigma_k},0,...,0\right) \in R_{nxm}. \tag{4.77}$$

Dla nieosobliwej macierzy kwadratowej zachodzi równość:

$$\mathbf{A}^+ = \mathbf{A}^{-1}.$$

Ze względu na fakt, że algorytm SVD jest dość skomplikowany, a jest dostępny jako gotowa metoda w pakietach do obliczeń numerycznych (Mathematica, Matlab), nie omawiamy tego algorytmu, a zainteresowanych odsyłamy do literatury [6].

4.7. Zadania do samodzielnego rozwiązania

Zadanie 4.1.

Metodą eliminacji Gaussa bez wyboru elementu podstawowego rozwiązać układy równań:

a) $\begin{cases} 2x_1 + x_3 = 5 \\ -3x_2 - x_3 - x_4 = 1 \\ 2x_1 + 5x_3 + 3x_4 = 12 \\ 3x_1 - 1.5x_2 + 5x_3 + 4.5x_4 = 17 \end{cases}$

b) $\begin{cases} -2x_1 - 2x_2 - 3x_3 = 5 \\ -2x_1 + x_2 - 3x_3 - 3x_4 = 2 \\ -2x_1 - 2x_2 - 4x_3 - 2x_4 = 6 \\ -x_1 + 2x_2 - 3x_3 - 2x_4 = 1 \end{cases}$

c) $\begin{cases} -3x_1 + 3x_2 - 3x_3 + x_4 = 10 \\ -3x_1 + 6x_2 - x_3 - 2x_4 = 16 \\ -1.5x_1 + 4.5x_2 - 2.5x_3 - 0.5x_4 = 13 \\ 1.5x_2 - 0.5x_3 - 3.5x_4 = 1 \end{cases}$.

Odp.

Rozwiązaniem układu a) jest wektor $x = [2, -1, 1, 1]^T$.
Rozwiązaniem układu b) jest wektor $x = [0, -1, -1, 0]^T$.
Rozwiązaniem układu c) jest wektor $x = [0, 3, 0, 1]^T$.

Zadanie 4.2.

Metodą eliminacji Gaussa z wyborem elementu podstawowego w kolumnie rozwiązać układy równań:

a) $\begin{cases} -1.5x_1 + 0.5x_2 + x_4 = -0.5 \\ 3x_1 - x_2 + 5x_3 - 5x_4 = -6 \\ 1.5x_1 - 4.5x_2 + 4.5x_3 - 1.5x_4 = 0 \\ 1.5x_1 + 0.5x_2 + 7x_3 - 1.75x_4 = -14.75 \end{cases}$

b) $\begin{cases} 0.25x_1 + 2.75x_2 + 7.5x_3 + 1.25x_4 = 10.25 \\ -0.5x_1 - 3.5x_2 - 3.5x_3 - 4x_4 = -7 \\ x_1 - x_2 + x_3 - 2x_4 = 0 \\ 2.5x_3 - 4x_4 = 2.5 \end{cases}$

c) $\begin{cases} x_1 + 0.25x_2 - 4.25x_3 + 6.25x_4 = 3.5 \\ 0.25x_2 - 0.75x_3 + 4.75x_4 = -3 \\ 4x_1 + 4x_2 + 4x_3 + 2x_4 = -2 \\ -3x_1 - 2x_2 - 6.5x_4 = 1.5 \end{cases}$.

Odp.

Rozwiązaniem układu a) jest wektor $x = [-1 \quad -2 \quad -2 \quad -1]^T$.
Rozwiązaniem układu b) jest wektor $x = [0 \quad 1 \quad 1 \quad 0]^T$.
Rozwiązaniem układu c) jest wektor $x = [1 \quad 1 \quad -2 \quad -1]^T$.

<u>Zadanie 4.3.</u>

Metodą eliminacji Gaussa z wyborem elementu podstawowego w wierszu rozwiązać układy równań:

a) $\begin{cases} -3x_1 + x_2 - 2x_3 - 5x_4 = -10 \\ 1.5x_1 + 5.5x_2 + 6x_3 - 2.5x_4 = 5 \\ -11.5x_1 + 9.5x_2 - 0.5x_3 - 10x_4 = -22 \\ -8.5x_1 + 17.5x_2 + 10x_3 - 1.25x_4 = 0.25 \end{cases}$

b) $\begin{cases} 5x_1 + 3x_2 + 4x_3 + 7x_4 = 30 \\ 5x_2 + x_3 + 7x_4 = 21 \\ 13.75x_1 - 2.5x_2 + 8.25x_3 + 7x_4 = 41.75 \\ 12.5x_1 - 2x_2 + 13.5x_3 + 7x_4 = 51.5 \end{cases}$

c) $\begin{cases} -2x_1 + 2x_2 - 3x_3 - 6x_4 = 4 \\ 2.5x_1 + 4.5x_2 + 3.25x_3 - 7.5x_4 = 9 \\ -5x_1 + 2x_2 - 1.5x_3 - 3x_4 = 4 \\ 3.25x_1 + 4.25x_2 + 3.25x_3 + 3x_4 = 8.5 \end{cases}$.

Odp.

Rozwiązaniem układu a) jest wektor $x = [1 \quad 0 \quad 1 \quad 1]^T$.
Rozwiązaniem układu b) jest wektor $x = [1 \quad 1 \quad 2 \quad 2]^T$.
Rozwiązaniem układu c) jest wektor $x = [0 \quad 2 \quad 0 \quad 0]^T$.

Zadanie 4.4.

Metodą Gaussa-Jordana rozwiązać układy równań:

a) $\begin{cases} 4x_1 - x_3 = -4 \\ -3x_2 + 2x_3 + 2x_4 = -2 \\ 4x_1 - 1.5x_2 + 3x_3 + 3x_4 = -1 \\ 2x_1 + x_3 + 3x_4 = 4 \end{cases}$

b) $\begin{cases} -3x_1 + 3x_2 - 3x_4 = 0 \\ -3x_1 + 6x_2 + 4x_3 - 3x_4 = 5 \\ -4.5x_1 + 7.5x_2 + 7x_3 - 7.5x_4 = 14 \\ -4.5x_1 + 7.5x_2 + 4x_3 - 2.5x_4 = 3 \end{cases}$

Odp.

Rozwiązaniem układu a) jest wektor $x = [-1 \quad 2 \quad 0 \quad 2]^T$.
Rozwiązaniem układu b) jest wektor $x = [0 \quad -1 \quad 2 \quad -1]^T$.

Zadanie 4.5.

Metodą rozkładu LU rozwiązać układy równań:

a) $\begin{cases} 2x_1 - 2x_2 + 2x_3 + 4x_4 = 0 \\ 2x_1 + x_2 + 3x_3 + 2x_4 = 0 \\ 2x_1 + x_2 + 7x_3 = -6 \\ 1.5x_2 + 6.5x_3 - 5x_4 = -10 \end{cases}$

b) $\begin{cases} 3x_1 - 2x_2 + 4x_3 - 3x_4 = 7 \\ 3x_1 - 4x_2 + 5x_3 - 4x_4 = 7 \\ -2x_2 - 2x_3 = -2 \\ -2x_2 - 3.5x_3 - 1.5x_4 = 1 \end{cases}$

Odp.

Rozwiązaniem układu a) jest wektor $x = [0 \quad 1 \quad -1 \quad 1]^T$.
Rozwiązaniem układu b) jest wektor $x = [1 \quad 1 \quad 0 \quad -2]^T$.

Zadanie 4.6.

Metodą Choleskiego rozwiązać układy równań:

a) $\begin{cases} 9x_1 - 6x_2 + 6x_3 - 6x_4 = 15 \\ -6x_1 + 13x_2 - 7x_3 + 10x_4 = -22 \\ 6x_1 - 7x_2 + 14x_3 - 12x_4 = 11 \\ -6x_1 + 10x_2 - 12x_3 + 13x_4 = -17 \end{cases}$

b) $\begin{cases} 4x_1 + 8x_2 + 4x_3 + 8x_4 = 0 \\ 8x_1 + 20x_2 - 2x_3 + 12x_4 = 34 \\ 4x_1 - 2x_2 + 54x_3 + 8x_4 = -80 \\ 8x_1 + 12x_2 + 8x_3 + 40x_4 = -84 \end{cases}$

c) $\begin{cases} 25x_1 + 15x_2 + 20x_3 - 15x_4 = -75 \\ 15x_1 + 13x_2 + 20x_3 - 15x_4 = -65 \\ 20x_1 + 20x_2 + 48x_3 - 4x_4 = -76 \\ -15x_1 - 15x_2 - 4x_3 + 59x_4 = 137 \end{cases}$.

Odp.

Rozwiązaniem układu a) jest wektor $x = [1 \quad -1 \quad -1 \quad -1]^T$.
Rozwiązaniem układu b) jest wektor $x = [1 \quad 3 \quad -1 \quad -3]^T$.
Rozwiązaniem układu c) jest wektor $x = [-1 \quad 0 \quad -1 \quad 2]^T$.

Zadanie 4.7.

Dokonać rozkładu QR dla macierzy:

a) $A = \begin{bmatrix} 1 & 3 & 1 & 2 \\ 1 & 5 & 0 & 4 \\ 2 & 7 & 1 & 1 \\ 3 & 2 & 3 & 2 \end{bmatrix}$

b) $A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 3 \end{bmatrix}$.

Odp.

a) $Q = \begin{bmatrix} -0.2582 & -0.1923 & 0.8702 & 0.3730 \\ -0.2582 & -0.5317 & -0.4739 & 0.6528 \\ -0.5164 & -0.5543 & 0.0041 & -0.6528 \\ -0.7746 & 0.6108 & -0.1348 & 0.0933 \end{bmatrix}$

$$R = \begin{bmatrix} -3.8730 & -7.2296 & -3.0984 & -3.6148 \\ 0 & -5.8936 & 1.0859 & -1.8438 \\ 0 & 0 & 0.4698 & -0.4208 \\ 0 & 0 & 0 & 2.8908 \end{bmatrix}$$

b) $Q = \begin{bmatrix} -0.5774 & 0 & -0.8165 \\ -0.5774 & -0.7071 & 0.4082 \\ -0.5774 & -0.7071 & 0.4082 \end{bmatrix}$

$$R = \begin{bmatrix} -1.7321 & -1.7321 \\ 0 & 2.8284 \\ 0 & 0 \end{bmatrix}.$$

Zadanie 4.8.

Rozwiązać układ równań z zadania 4.7 metodą Householdera dla prawych stron równych odpowiednio:

a) $b = [7{,}5 \quad 11 \quad 16{,}5 \quad 8{,}5]^T$;
b) $b = [3 \quad 1 \quad -1]^T$.

Odp.

Rozwiązaniem układu a) jest wektor $x = [1 \quad 2 \quad 0.5 \quad 0]^T$.
Rozwiązaniem układu b) jest wektor $x = [1.5 \quad -0.5]^T$.

5. Rozwiązywanie równań i układów równań nieliniowych

5.1. Wstęp

Na ogół pierwiastki równania nieliniowego:

$$\boldsymbol{f}(\boldsymbol{x}) = 0 \tag{5.1}$$

nie dają wyrazić się za pomocą wzoru analitycznego. Dlatego duże znaczenie mają metody przybliżonego rozwiązywania równań. Są to ***metody kolejnych przybliżeń*** pierwiastka czyli ***metody iteracyjne***. Polegają one na tym, że startując od jednej początkowej wartości pierwiastka czyli punktu startowego x_0 konstruuje się ciąg punktów x_1, x_2, x_3, ... zbieżny do tego pierwiastka. W niektórych metodach potrzebne są dwa pierwsze przybliżenia pierwiasta. W metodach tych zadanie znalezienia pierwiastków uważamy za wykonane, jeśli potrafimy określić je z żądaną dokładnością i podać oszacowanie błędu. Trzeba jednak pamiętać, że większość metod przybliżonego rozwiązywania równań można stosować jedynie wtedy, gdy znany jest przedział, w którym znajduje się pojedynczy pierwiastek, czyli tzw. ***przedział izolacji***.

Do najbardziej popularnych metod znajdowania pierwiastków równań nieliniowych zaliczamy metodę bisekcji, metodę regula-falsi, metodę siecznych, metodę Newtona i jej modyfikacje [1, 4, 5, 7, 8, 9, 10].

5.2. Metoda bisekcji

O funkcji $f(x)$ z równania (5.1) zakładamy, że:

- jest ciągła na przedziale domkniętym $<a,b>$;
- w punktach a i b wartości funkcji $f(x)$ mają przeciwne znaki, tzn. $f(a)f(b)<0$;

W przypadku metody bisekcji (inaczej zwanej też metodą połowienia) nie musimy zakładać monotoniczności funkcji na przedziale domkniętym $<a,b>$. Metoda bisekcji znajduje jeden pierwiastek, nawet jeśli w przedziale $<a,b>$ jest tych pierwiastków wiele. Metoda nie korzysta

z własności funkcji i jej przebiegu wewnątrz badanego przedziału - wystarcza jej informacja o znaku funkcji na jego krańcach. Stosując metodę bisekcji pierwiastek możemy wyznaczyć z dowolną zadaną dokładnością ε.

Pierwszy krok metody bisekcji polega na podziale przedziału $< a, b>$ na połowę punktem:

$$x_1 = (a + b) / 2.$$

Jeżeli wartość funkcji w tym punkcie jest bardzo bliska zeru tj. $|f(x_1)| < eps$, to x_1 jest szukanym pierwiastkiem. W przeciwnym wypadku z otrzymanych dwóch przedziałów $< a, x_1>$ i $< x_1, b>$ wybieramy ten, na końcach którego funkcja $f(x)$ ma przeciwne znaki tj.:

- jeżeli $f(a)*f(x_1)<0$, wtedy a się nie zmienia ale $b=x_1$;
- w przeciwnym razie $a=x_1$, podczas gdy b nie ulega zmianie.

W drugim kroku metody, wybrany do dalszych obliczeń przedział, ponownie dzielimy na połowę i dostajemy punkt:

$$x_2 = (a + b) / 2.$$

Po raz kolejny badamy wartość funkcji $f(x)$, tym razem w punkcie x_2 i znaki funkcji $f(x)$ na końcach przedziałów oraz wybieramy jeden z nich do dalszych obliczeń tj:

- jeżeli $f(a)*f(x_2)<0$, wtedy a się nie zmienia, $b=x_2$;
- w przeciwnym razie $a=x_2$, b się nie zmienia.

W wyniku takiego postępowania po pewnej liczbie kroków otrzymamy ciąg przedziałów takich, że $f(a_i) \cdot f(b_i) < 0$, przy czym a_i oraz b_i są odpowiednio początkiem i końcem i-tego przedziału, a jego długość wynosi:

$$|b_i - a_i| = \frac{1}{2^i}(b - a). \tag{5.2}$$

Lewe końce ciągu przedziałów tworzą ciąg niemalejący i ograniczony z góry a prawe końce ciąg nierosnący i ograniczony z dołu, więc z (5.2) wynika, że istnieje ich wspólna granica α. Ze względu na stosowaną metodę obliczeniową, ten sposób znajdowania pierwiastków nazywamy metodą bisekcji (także ***metodą połowienia*** lub ***metodą równego podziału***).

Obliczenia kończymy gdy przedział $<a_i,b_i>$ jest już odpowiednio mały czyli: $|b_i-a_i| < 2\varepsilon$, ponieważ przybliżenie pierwiastka $\bar{x} = 0.5(a+b)$ spełnia najczęściej stosowane kryterium końca obliczeń:

$$\left|\bar{x} - x^*\right| \le \varepsilon ,$$

gdzie x^* jest rozwiązaniem równania.

Innym, ale stosowanym w ostateczności kryterium końca obliczeń jest sprawdzanie, czy wartość funkcji w punkcie x_i jest dostatecznie małą, czyli $|f(x_i)| < eps$.

Podstawową zaletą metody bisekcji oprócz jej dużej prostoty i łatwej implementacji jest pewność, że w każdej kolejnej iteracji szukany pierwiastek leży między dwiema wartościami, dla których funkcja $f(x)$ zmienia znak. Zawsze więc dojdziemy do szukanego rozwiązania. Jest to metoda na ogół wolniej zbieżna niż metoda Newtona, ale nie oznacza to, że jest wolno zbieżna. Przy rozwiązywaniu wielkich układów równań liniowych potrzeba niekiedy wykonać kilkaset tysięcy iteracji, a w metodzie bisekcji aby zyskać 6 cyfr znaczących wystarcza 20 iteracji.

Przykład 5.1.

Metodą bisekcji znaleźć rzeczywisty pierwiastek równania:

$x^3 + x - 1 = 0$

z dokładnością do 0.01.

Łatwo sprawdzić, że pierwiastek ten znajduje się w przedziale $<0,1>$. Mamy bowiem:

$f(0) = -1$ oraz $f(1) = 1$,

czyli:

$f(0)*f(1)<0$.

Pochodna: $f'(x) = 3x^2 + 1$ jest w przedziale $<0,1>$ dodatnia oraz $f(x)$ jako wielomian jest funkcją ciągłą, zatem $<0,1>$ jest przedziałem izolacji pierwiastka. Zadana dokładność obliczeń $eps = 0.01$.

Pierwszy krok obliczeń:

$a=0, b=1$,

$x_1 = (a+b)/2 = (0+1)/2 = 0.5$.

Wartość funkcji w znalezionym punkcie x_1:

$f(x_1) = f(0.5) = -0.375$.

Mamy $f(x_1)*f(b)<0$ więc przyjmujemy:
$a = x_1 = 0.5$,
$b = 1$,
a ponieważ $|b-a|>2eps$ to należy kontynuować obliczenia.

Drugi krok obliczeń:
$a=0.5, b=1$,
$x_2 = (a+b)/2 = (0.5+1)/2 = 0.75$
oraz:
$f(x_2) = f(0.75) = 0.172$.
Ponieważ $f(a)*f(x_2)<0$ więc:
$a = 0.5$,
$b = x_2 = 0.75$ i ponownie $|b-a|>2eps$, czyli kontynuujemy obliczenia.

Trzeci krok obliczeń:
$a=0.5, b=0.75$,
$x_3 = (a+b)/2 = (0.5+0.75)/2 = 0.625$
$f(x_3) = f(0.625) = -0.131$.
Skoro $f(x_3)*f(b)<0$ to:
$a = x_3 = 0.625$,
$b = 0.75$ i nadal $|b-a|>2eps$.

W analogiczny sposób wykonujemy kolejne kroki. Obliczenia kończymy po wykonaniu kroku 7 a znaleziony pierwiastek jest równy x_7=0.6796875. Dokładne wyniki obliczeń kolejnych przybliżeń szukanego pierwiastka dla pierwszych 7 kroków umieszczone są w tabeli 5.1.

Tabela 5.1. Wyniki obliczeń do przykładu 5.1

i	x_i	$f(x_i)$	i	x_i	$f(x_i)$
1	0.5	-0.375	5	0.65625	-0.0611268
2	0.75	0.171875	6	0.671875	-0.0248299
3	0.625	-0.1308593	7	0.6796875	-0.0063138
4	0.6875	0.0124511			

5.3. Metoda regula falsi

Nazwa metody pochodzi od łacińskich słów: *regula* - linia i *falsus* - fałszywy. Jest to zatem metoda fałszywego założenia liniowości funkcji.

Zakładamy, że w rozpatrywanym przedziale <*a*,*b*> funkcja *f(x)* spełnia założenia:

- jest funkcją klasy C^2 na przedziale domkniętym <*a*,*b*>;
- w punktach *a* i *b* wartości funkcji *f*(*x*) mają przeciwne znaki, tzn. *f*(*a*)*f*(*b*)<0;
- pierwsza pochodna funkcji *f(x)* ma na przedziale <*a*,*b*> stały znak, różny od zera;
- druga pochodna funkcji *f(x)* ma na przedziale <*a*,*b*> stały znak, różny od zera.

Spełnienie wymienionych warunków gwarantuje zbieżność metody oraz, że wewnątrz badanego przedziału znajduje się dokładnie jeden pierwiastek.

Z założeń tych wynika, że wykres funkcji $y = f(x)$ może mieć jedną z czterech postaci przedstawionych na rysunkach 5.1a-5.1d.

Rozpatrzmy przypadek, gdy w przedziale <*a*,*b*> pochodne $f'(x)$ i $f''(x)$ są dodatnie (rys. 5.2, dla pozostałych przypadków rozumowanie jest analogiczne).

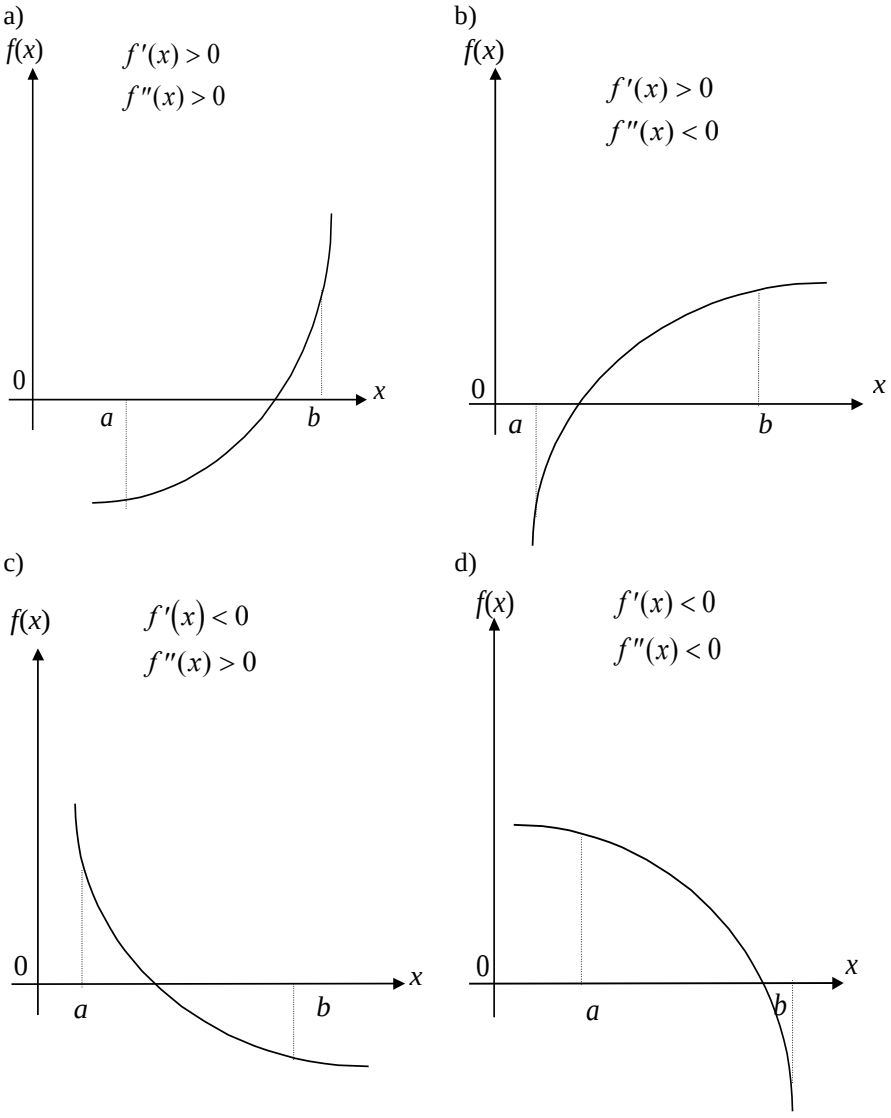
Przez punkty A(*a*, *f*(*a*)) i B(*b*, *f*(*b*)) prowadzimy cięciwę o równaniu:

$$y - f(a) = \frac{f(b) - f(a)}{b - a}(x - a).$$

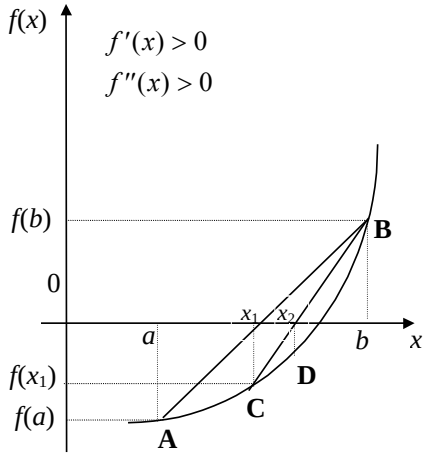
Odciętą x_1 punktu, w którym cięciwa AB przecina oś 0*x*, przyjmuje się jako pierwsze przybliżenie szukanego pierwiastka równania (5.1). Stąd:

$$x_1 = a - \frac{f(a)}{f(b) - f(a)}(b - a).$$

Jeżeli $|f(x_1)|<eps$ (zadana dokładność), to oczywiście x_1 jest szukanym przybliżeniem pierwiastka α i zadanie jest zakończone.



Rys. 5.1. Wykres funkcji $f(x)$ w przedziale $< a, b >$ w zależności od znaku pierwszej i drugiej pochodnej funkcji



Rys. 5.2. Interpretacja geometryczna metody regula falsi, jeśli $f'(x) > 0$ i $f''(x) > 0$

Załóżmy, że $|f(x_1)| > eps$. Jeżeli przybliżenie x_1 nie jest wystarczająco dokładne, to przez punkt $C(x_1, f(x_1))$ oraz przez ten z punktów A, B, którego rzędna ma przeciwny znak niż $f(x_1)$, prowadzimy następną cięciwę (rys.5.2). Odcięta x_2 punktu, w którym ta cięciwa przetnie oś $0x$, da nam drugie przybliżenie pierwiastka α. Dla uproszczenia rozumowania przyjęliśmy, że $f'(x) > 0$ oraz $f''(x) > 0$ w przedziale $< a, b >$, co oznacza, że funkcja $y = f(x)$ jest wypukła i w kolejnych przybliżeniach punkt B pozostaje nieruchomy. Dla funkcji $f(x)$ takiej, że $f'(x) > 0$ i $f''(x) < 0$ w przedziale $< a, b >$, nieruchomy byłby punkt A. Jeżeli przybliżenie x_2 jest nadal niewystarczające, to przez punkty B i $D(x_2, f(x_2))$ prowadzimy trzecią cięciwę, co daje nam trzecie przybliżenie x_3 itd.

W ten sposób otrzymujemy kolejne wyrazy ciągu przybliżeń pierwiastka $x_1, x_2, x_3, ..., x_n$ określonego wzorem rekurencyjnym:

$$x_0 = a,\ x_{k+1} = x_k - \frac{f(x_k)}{f(b) - f(x_k)}(b - x_k), \quad k = 1, 2, ..., \boldsymbol{n}. \tag{5.3}$$

Można wykazać, że przy przyjętych założeniach ciąg ten jest rosnący i ograniczony, a więc zbieżny oraz, że jego granicą jest szukany pierwiastek α. Jeśli nieruchomy jest punkt B, to kolejne wyrazy ciągu przybliżeń są mniejsze od α oraz $f(x_k) < 0$ dla każdego k).

Przechodząc do granicy dla $n \to \infty$ z równości (5.3) otrzymujemy:

$$g = g - \frac{f(g)}{f(b) - f(g)}(b - g),$$

gdzie $g = \lim_{n\to\infty} x_n$, $a < g < b$.

Stąd $f(g) = 0$, a przy założeniu istnienia tylko jednego pierwiastka w przedziale $< a, b >$ mamy $g \equiv \alpha$.

Błąd bezwzględny przybliżenia x_n można oszacować znając dwa kolejne przybliżenia x_{n-1} i x_n oraz korzystając z twierdzenia Lagrange'a o przyrostach [4]:

$$|\alpha - x_{k+1}| \le \frac{M - m}{m}|x_{k+1} - x_k|, \tag{5.4}$$

przy czym:

$$m = \inf_{x\in(a,b)} |f'(x)|, \quad M = \sup_{x\in(a,b)} |f'(x)|.$$

Jeśli $M \le 2m$, to:

$$|\alpha - x_{k+1}| \le |x_{k+1} - x_k|. \tag{5.5}$$

W niewielkim otoczeniu pierwiastka α można przyjąć, że:

$$|\alpha - x_{k+1}| \cong \left| \frac{x_{k+1} - x_k}{f(x_{k+1}) - f(x_k)} f(x_{k+1}) \right|. \tag{5.6}$$

Metoda regula falsi jest zbieżna dla dowolnej funkcji ciągłej w przedziale $< a, b >$ ($f(a)f(b) < 0$), jeżeli tylko pierwsza pochodna tej funkcji jest ograniczona i różna od zera w otoczeniu pierwiastka. Jeżeli druga pochodna nie zmienia znaku w rozpatrywanym przedziale, to ten koniec przedziału, w którym $f''(x)f(x) > 0$, jest stałym punktem iteracji - wszystkie cięciwy przechodzą przez ten punkt. Wadą metody jest jej stosunkowo powolna zbieżność.

Przykład 5.2.

Metodą regula falsi znaleźć rzeczywisty pierwiastek równania:

3x-cos x-1=0.

Badając funkcję występującą w równaniu możemy stwierdzić, że w przedziale $< 0.25, 0.75 >$ ma ona dokładnie jedno miejsce zerowe

(w badanym przedziale $f'(x) > 0$). Kolejne przybliżenia obliczone według wzoru (5.3) znajdują się w tabeli 5.2.

5.4. Metoda siecznych

Metodę regula falsi można znacznie ulepszyć, jeżeli zrezygnuje się z żądania, aby w punktach wytyczających kolejną cięciwę funkcja *f(x)* miała różne znaki, natomiast do wyznaczenia (n+1)-szego przybliżenia wykorzysta się punkty x_n oraz x_{n-1}. Wzór (5.3) przyjmie wówczas postać:

$$x_{n+1} = x_n - \frac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})},\ n = 1, 2, \ldots . \tag{5.7}$$

Metoda (5.7) nosi nazwę metody siecznych. Jej zbieżność jest znacznie szybsza, niż metody regula falsi. Niestety, zdarzają się przypadki, gdy może nie być zbieżna, np. gdy początkowe przybliżenia nie leżą dostatecznie blisko pierwiastka. W metodzie tej istotne znaczenie ma maksymalna graniczna dokładność wynikająca z przyjętej arytmetyki. Gdy bowiem różnica $x_{n+1} - x_n$ jest tego samego rzędu, co oszacowanie błędu, jakim jest obarczona, następne przybliżenie może już być całkowicie błędne. Dlatego też za dodatkowe kryterium przerwania iteracji należy przyjmować wartości $|f(x_n)|$ tak, aby tworzyły one ciąg malejący (w końcowej fazie obliczeń). Iteracja powinna być przerwana, jeżeli różnica między kolejnymi przybliżeniami zamiast maleć zaczyna szybko wzrastać. W takim przypadku należy przeprowadzić powtórną lokalizację pierwiastka znacznie zawężając początkowy przedział jego izolacji.

Tabela 5.2. Wyniki obliczeń do przykładu 5.2

x_i	$f(x_i)$
a = 0.25	-1.218912
b = 0.75	0.518311
x_1 = 0.600819	-0.022416
x_2 = 0.607003	-0.000352
x_3 = 0.607100	-0.000006
x_4 = 0.607101	-0.000002
x_5 = 0.607101	

Przykład 5.3.

Stosując metodę siecznych znaleźć pierwiastek równania:
$x^3 + x^2 - 3x - 3 = 0$ w przedziale < 1, 2 >.

Funkcja występująca w równaniu ma w przedziale < 1, 2 > dokładnie jeden pierwiastek. Kolejne przybliżenia tego pierwiastka obliczamy zgodnie ze wzorem (5.7). Wyniki obliczeń zapisane są w tabeli 5.3.

Tabela 5.3. Wyniki obliczeń do przykładu 5.3

i	x_i	$f(x_i)$
0	1	-4
1	2	3
2	1.57142	-1.36449
3	1.70540	-0.24784
4	1.73513	0.02920
5	1.73199	0.000576
6	1.73193	

5.5. Metoda Newtona-Raphsona

Metoda Newtona-Raphsona, zwana także ***metodą Newtona*** lub ***metodą stycznych,*** należy do metod iteracyjnych. Dla zadania jednowymiarowego, tzn. jednego równania, w metodzie tej dla znalezienia następnego punktu iteracji korzysta się tylko z jednego punktu startowego x_0. Jeśli wartość funkcji dla $x = x_0$ jest różna od zera, to w punkcie o współrzędnych (x_0, $f(x_0)$) prowadzi się styczną do wykresu funkcji (rys. 5.3).

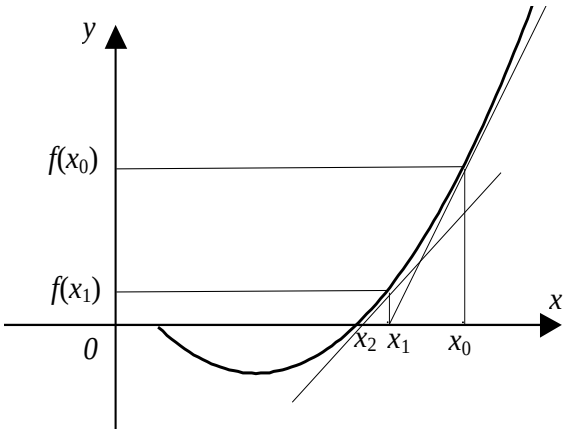
Punkt przecięcia tej stycznej z osią $0x$ stanowi pierwsze przybliżenie x_1 szukanego pierwiastka. Następnie w punkcie o współrzędnych (x_1, $f(x_1)$) prowadzi się kolejną styczną. Punkt przecięcia tej stycznej z osią $0x$ jest drugim przybliżeniem pierwiastka x_2. W ten sposób otrzymuje się kolejne wyrazy ciągu przybliżeń x_1, x_2, x_3, ...

Wzór rekurencyjny opisujący obliczanie tych wyrazów ma postać:

$$x_{k+1} = x_k + h_k, \; h_k = -\frac{f(x_k)}{f'(x_k)}, \qquad k = 0,1,.... \tag{5.8}$$

lub pisząc krócej:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0,1,.... \tag{5.9}$$



Rys. 5.3. Interpretacja geometryczna metody Newtona-Raphsona

Obliczenia kończy się, gdy:

$$\left|h^k\right| = \left|x^{k+1} - x^k\right| < eps\,. \tag{5.10}$$

przy czym *eps* oznacza zadaną z góry dokładność i jest oszacowaniem błędu wartości $f(x_n)/f'(x_n)$.

Warto zauważyć, że jeżeli we wzorze (5.9) za $f'(x_k)$ wstawimy iloraz różnicowy $\frac{f(x_k)-f(x_{k-1})}{x_k-x_{k-1}}$ to otrzymamy metodę siecznych.

Wybór punktu startowego x_0 jest bardzo istotny i może decydować o zbieżności ciągu kolejnych przybliżeń. Jeżeli wystartujemy odpowiednio blisko od rozwiązania, wtedy metoda Newtona jest lokalnie kwadratowo zbieżna czyli jej ***wykładnik zbieżności*** wynosi p=2, jeśli $f'(\alpha) \neq 0$.

Wykładnik zbieżności metody iteracyjnej z definicji jest taką liczbą p>1, że:

$$\left\|x_{k+1} - \alpha\right\| \leq c\left\|x_k - \alpha\right\|^p, \; 0 < c < \infty\,.$$

Gdy $c \approx 1$, to w metodzie Newtona, w każdym kroku (z wyjątkiem początkowych) podwaja się liczba cyfr dokładnych w przybliżeniu pierwiastka.

W metodzie siecznych wykładnik zbieżności wynosi:

$$p = \frac{\sqrt{5}+1}{2},$$

skąd wynika szybsza zbieżność metody Newtona.

Przykład 5.4.

Zbadać z dokładnością do 1*e*-5 rozwiązanie równania $y = e^x + 2x + 1$ metodą Newtona wybierając jako punkt startowy:

a) x_0=0,
b) x_0=5.

Dla badanej funkcji mamy:

$f(x) = e^x + 2x + 1$

$f'(x) = e^x + 2$

Ad a)

Wykonamy obliczenia dla punktu startowego x_0=0. Obliczamy przybliżenia rozwiązania korzystając z wzoru 5.8 i wartości funkcji w obliczonych punktach:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 0 - \frac{f(0)}{f'(0)} = 0 - \frac{2}{3} = -0.666667.$$

Wartość funkcji w otrzymanym punkcie x_1 wynosi:

$$f(x_1) = f(-0.666667) = 0.18008379.$$

Ponieważ $|f(x_1)| > 1e-5$ należy policzyć kolejny krok metody:

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = -0.666667 - \frac{f(-0.666667)}{f'(-0.666667)} = -0.738316$$

Kontynuując obliczenia otrzymujemy:

$$f(x_2) = f(-0.738316) = 0.00128692,$$

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)} = -0.738835,$$
$$f(x_3) = f(-0.738835) = 0.00000100.$$

Po wykonaniu trzeciego kroku następuje koniec obliczeń (osiągnięto zadaną dokładność 0.00001) a znaleziony pierwiastek wynosi:
x = -0.738835.

Ad b)

Podobne obliczenia wykonamy dla punktu startowego x_0=5.
Otrzymujemy kolejno:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 5 - \frac{f(5)}{f'(5)} = 5 - \frac{159.413159}{150.413159} = 3.940165,$$
$$f(x_1) = f(3.940165) = 60.3074059,$$
$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 2.811385,$$
$$f(x_2) = f(2.811385) = 23.255709,$$
$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)} = 1.563288,$$
$$f(x_3) = f(1.563288) = 8.901072,$$
$$x_4 = x_3 - \frac{f(x_3)}{f'(x_3)} = 0.249379,$$
$$f(x_4) = f(0.249379) = 2.781987,$$
$$x_5 = x_4 - \frac{f(x_4)}{f'(x_4)} = -0.597954,$$
$$f(x_5) = f(-0.597954) = 0.354403,$$
$$x_6 = x_5 - \frac{f(x_5)}{f'(x_5)} = -0.736792,$$
$$f(x_6) = f(-0.736792) = 0.005063,$$
$$x_7 = x_6 - \frac{f(x_6)}{f'(x_6)} = -0.738835,$$
$$f(x_7) = f(-0.738835) = 0.0000010.$$

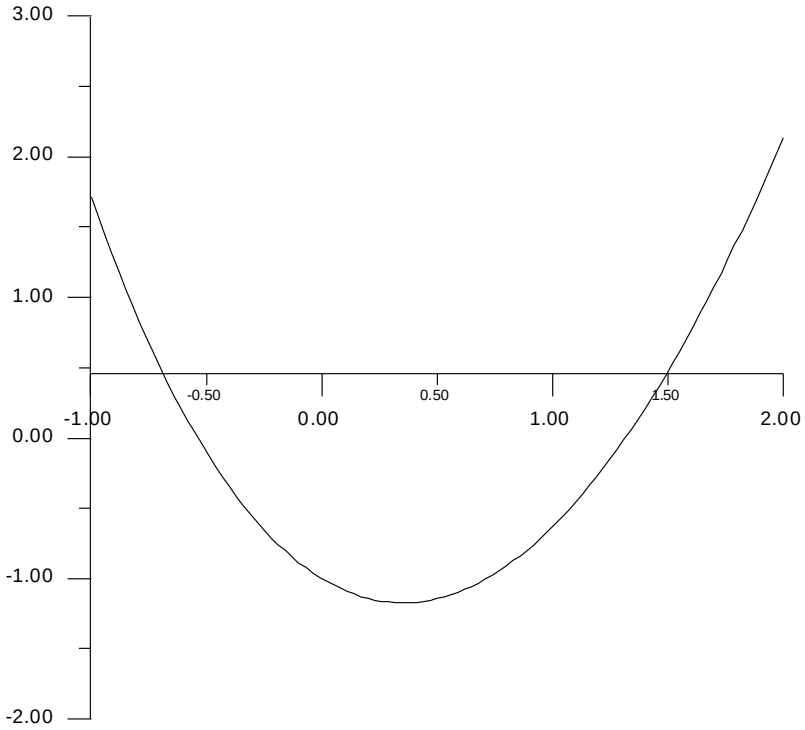
Po wykonaniu siódmego kroku następuje koniec obliczeń. Pierwiastek wynosi: *x = -0.738835*.

Przykład 5.5.

Zbadać zależność rozwiązania równania od wyboru punktu startowego w metodzie Newtona dla równania:

$x_0\ y = e^{-x} + x^2 - 2$.

Wykres funkcji $y(x)$ przedstawia rys. 5.4.



Rys. 5.4. Wykres funkcji z przykładu 5.4.

Badana funkcja ma dwa rzeczywiste pierwiastki $x_1 > 0$ oraz $x_2 < 0$. Wyniki obliczeń tych pierwiastków z różną dokładnością i dla różnych punktów startowych (niekiedy bardzo odległych od szukanych pierwiastków) zebrane są w tabeli 5.4.

Tabela 5.4. Wyniki obliczeń do przykładu 5.5

x_0	-1.0		10.0		100.0	
dokładność	0.001	0.000001	0.001	0.000001	0.001	0.000001
liczba iteracji	4	5	6	7	10	11
	-0.635	-0.635	5.1	5.1	50.1	50.1
	-0.534	-0.543	2.74	2.74	25.02	25.02
	-0.537	-0.537	1.71	1.71	12.5	12.5
	-0.537	-0.5373	1.36	1.369	6,35	6,35
		-0.5372	1.31	1.317	3.33	3.33
			1.3159	1.3169	1.95	1.95
				1.3159	1.43	1.43
					1.32	1.32
					1.315	1.315
					1.3159	1.3159
						1.315973

Z tabeli 5.4 wynika, że osiągnięcie dokładności tysiąc razy większej wymaga wykonania tylko jednego kroku więcej. Warto zwrócić uwagę na to, że wartość kolejnego przybliżenia najbardziej zmienia się w pierwszych iteracjach. Wybór punktu startowego ma także znaczenie - np. dla $x_0 = -1$ potrzeba 4 iteracji, zaś 11 iteracji dla $x_0 = 100$.

W wyniku przeprowadzonych obliczeń stwierdzono, że wartość pochodnej $f'(x)$ począwszy od drugiej iteracji zmienia się nieznacznie. Z wyrażenia na h_k wynika, że pochodną $f'(x)$ można obliczać z taką dokładnością względną, z jaką oblicza się $f(x)$. Można więc nie wyznaczać $f'(x)$ w każdej iteracji. Znacznie przyspiesza to proces iteracyjny (szczególnie dla układów równań nieliniowych), nie wpływając znacząco na jego zbieżność.

Uproszczone wzory (5.8) przyjmują postać:

$$x_{k+1} = x_k + h_k, \; h_k = -\frac{f(x_k)}{f'(x_0)}, \; k = 0,1,.... \tag{5.11}$$

Jeśli poszukujemy pierwiastka równania (5.1) metodą Newtona w przedziale jego izolacji $< a, b >$, to warunkiem koniecznym zbieżności jest, aby punkt startowy x_0 znajdował się w tym przedziale. Metoda Newtona zapewnia zbieżność procesu iteracyjnego wtedy, gdy w przedziale izolacji pierwiastka pierwsza i druga pochodna funkcji $f(x)$ nie zmieniają znaku (rys. 5.1a-5.1d).

Punkt startowy x_0 należy wybierać następująco:

$x_0 = a$, jeśli $f(x)f''(x) < 0$ - przypadek b) i c),

$x_0 = b$, jeśli $f(x)f''(x) > 0$ - przypadek a) i d).

Zwiększenie czasu obliczeń lub niemożność znalezienia pierwiastka ma miejsce wówczas, gdy wybierzemy punkt początkowy:

$x_0 = b$, jeśli $f(x)f''(x) < 0$ - przypadek b) i c),

$x_0 = a$, jeśli $f(x)f''(x) > 0$ - przypadek a) i d).

Rys. 5.5 ilustruje przypadki niewłaściwego doboru punktu startowego x_0 w metodzie Newtona, co może doprowadzić do znalezienia innego pierwiastka (rys. 5.5a) lub przerwać obliczenia na skutek znalezienia punktu x_1 poza przedziałem określoności funkcji (rys. 5.5b) lub w nieskończoności (rys. 5.5c).

Przykład 5.6.

Stosując różne wzory iteracyjne rozwiązać równanie $x + \ln x = 0$. Pokazać, że szybkość zbieżności zależy od zastosowanego wzoru.

Metoda bisekcji

Przyjmujemy:

- początek przedziału: 0.34,
- koniec przedziału: 0.98,
- dokładność: 0.0001.

Znaleziony pierwiastek: $x = 0.567109$, liczba iteracji: 13.

Metoda regula falsi

Przyjmujemy:

- początek przedziału: 0.34,
- koniec przedziału: 0.98,
- dokładność: 0.0001.

Znaleziony pierwiastek: $x = 0.567307$, liczba iteracji: 13.

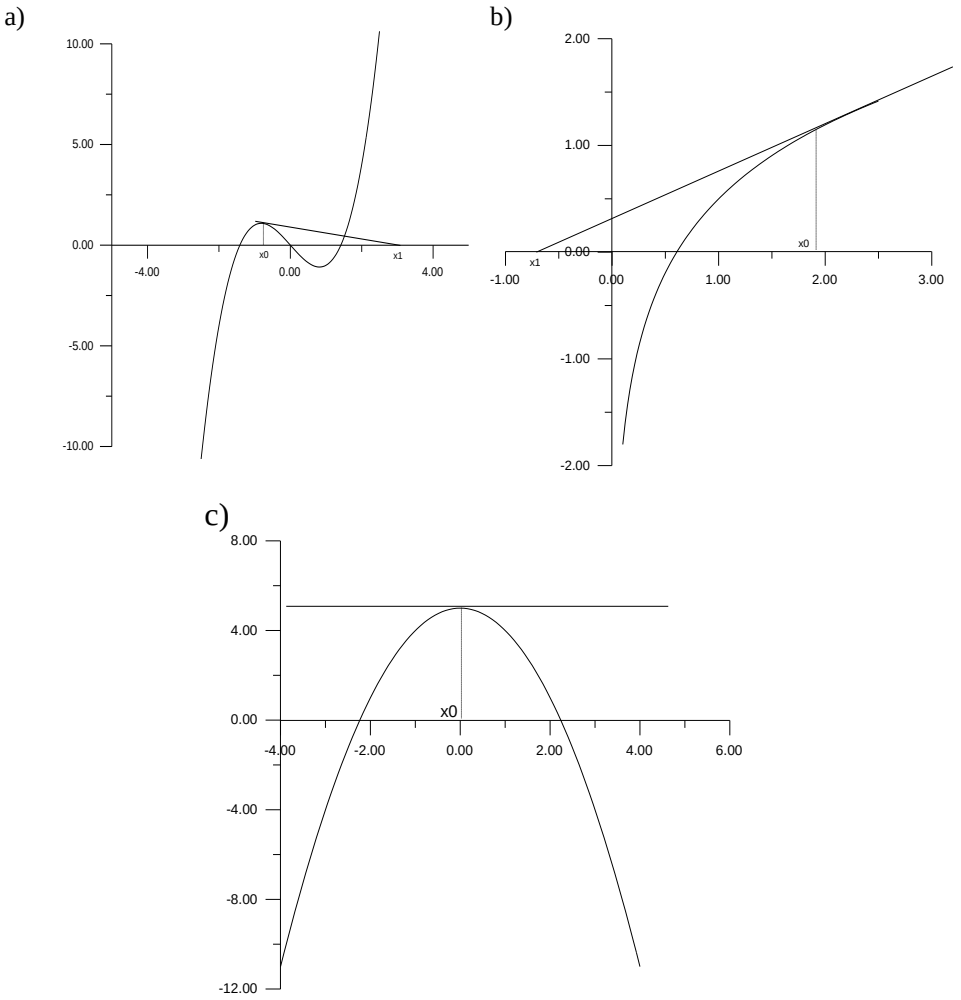
Metoda Newtona

Przyjmujemy:

- punkt startowy: 1,
- dokładność: 0.0001.

Znaleziony pierwiastek: $x = 0.567143$, liczba iteracji: 4.

Z przeprowadzonych obliczeń wynika, że najszybszą zbieżnością charakteryzuje się metoda Newtona, co jest zgodne z teorią.



Rys. 5.5. Przykłady niewłaściwego doboru punktu startowego w metodzie Newtona

5.6. Układy równań nieliniowych

W celu rozwiązania układu *n* równań nieliniowych:

$$\begin{cases} f_1(x_1, x_2 ... x_n) = 0 \\ f_2(x_1, x_2 ... x_n) = 0 \\ ..... \\ f_n(x_1, x_2 ... x_n) = 0 \end{cases} \tag{5.12}$$

metodę Newtona-Raphsona uogólnia się na *n*-wymiarów.

Zbiór argumentów $x_1, x_2, x_3, ...$ rozpatrujemy jako *n* - wymiarowy wektor postaci: $\mathbf{X} = [x_1, x_2 ... x_n]^{\mathrm{T}}$. Analogicznie zbiór funkcji $f_1, f_2, f_3, ...$ traktujemy jako *n*-wymiarowy wektor $\mathbf{F} = [f_1, f_2 ... f_n]^{\mathrm{T}}$. Wtedy układ równań (5.12) można zapisać w postaci:

$$\mathbf{F(X)} = 0, \tag{5.13}$$

zaś iteracyjny wzór Newtona-Raphsona przyjmie postać:

$$\begin{cases} \mathbf{J}(\mathbf{X}_{i+1})\mathbf{h} = -\mathbf{F}(\mathbf{X}_i) \\ \mathbf{X}_{i+1} = \mathbf{X}_i + \mathbf{h} \end{cases}, \tag{5.14}$$

gdzie **h** jest wektorem poprawki a **J**(**X**) jest obliczoną dla **X** ***macierzą Jacobiego***:

$$\mathbf{J(X)} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ .... & ... & ... & ... \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}.$$

W przypadku układu równań problem zbieżności jest trudniejszy, niż w przypadku pojedynczego równania. Dlatego bardzo ważne jest dobre przybliżenie startowe.

Wzór (5.14) można też zapisać w postaci:

$$\mathbf{J}(\mathbf{X}_i)(\mathbf{X}_{i+1} - \mathbf{X}_i) = -\mathbf{F}(\mathbf{X}_i). \tag{5.15}$$

Oznaczając $\mathbf{Z}_{i+1} = (\mathbf{X}_{i+1} - \mathbf{X}_i)$ otrzymujemy:

$$\mathbf{J}(\mathbf{X}_i)\mathbf{Z}_{i+1} = -\mathbf{F}(\mathbf{X}_i). \tag{5.16}$$

Algorytm rozwiązywania układu jest następujący:

1. Dla przybliżenia zerowego $\mathbf{X}_0$ obliczamy $\mathbf{F}(\mathbf{X}_0)$ oraz $\mathbf{J}(\mathbf{X}_0)$.
2. Stosując dowolną metodę rozwiązywania równań liniowych (np. eliminacji Gaussa), rozwiązujemy układ (5.16), otrzymując poprawkę $\mathbf{Z}_1$.
3. Sprawdzamy, czy $\mathbf{Z}_1$ spełnia narzucony warunek dokładności rozwiązania ($\|\mathbf{Z}_i\| \le \varepsilon$). Jeżeli tak, to przybliżenie zerowe jest rozwiązaniem. W przeciwnym przypadku przechodzimy do kroku czwartego.
4. Dodajemy poprawkę $\mathbf{Z}_1$ do $\mathbf{X}_0$ otrzymując $\mathbf{X}_1 = \mathbf{X}_0 + \mathbf{Z}_1$.
5. Z nową wartością **X** wracamy do kroku 1, obliczając kolejną poprawkę i nowe przybliżenie **X,** tak długo aż zostanie spełniony warunek dokładności.

W przypadku braku zbieżności przerywamy wykonywanie algorytmu. Wskaźnikiem rozbieżności jest wzrost lub oscylacja **Z** w kolejnych iteracjach. Zwykle wystarcza wykonanie niewielkiej liczby iteracji w celu zorientowania się, czy kontynuacja obliczeń jest celowa. Jeśli nie, to należy zadać nowe warunki początkowe lub zastosować inną metodę.

Przykład 5.7.

Znaleźć dwa pierwsze przybliżenia rozwiązania układu równań postaci:

$$f_1(x_1, x_2) = x_1^{\,2} - 2x_2^{\,2} = 0,$$
$$f_2(x_1, x_2) = 2x_1x_2 - 3 = 0.$$

Obliczenia wykonać z dokładnością *eps*=0.001. Jako przybliżenie początkowe przyjąć wektor:

$$\mathbf{X}_0 = \begin{bmatrix} 1{,}3 \\ 1{,}1 \end{bmatrix}.$$

Obliczamy $\mathbf{F}(\mathbf{X}_0)=\begin{bmatrix}-0{,}73\\-0{,}14\end{bmatrix}$, a następnie tworzymy macierz Jacobiego:

$$\mathbf{J}(\mathbf{X})=\begin{bmatrix}2x_1 & -4x_2\\ 2x_2 & 2x_1\end{bmatrix}.$$

Zatem:

$$\mathbf{J}(\mathbf{X}_0)=\begin{bmatrix}2{,}6 & -4{,}4\\ 2{,}2 & 2{,}6\end{bmatrix},$$

Rozwiązujemy układ równań (np. metodą eliminacji Gaussa):

$$\mathbf{J}(\mathbf{X}_0)\mathbf{Z}_1=-\mathbf{F}(\mathbf{X}_0),\ \text{czyli:}\begin{bmatrix}2{,}6 & 4{,}4\\ 2{,}2 & 2{,}6\end{bmatrix}\begin{bmatrix}h_0\\h_1\end{bmatrix}=-\begin{bmatrix}-0{,}73\\-0{,}14\end{bmatrix}$$

i otrzymujemy poprawkę $\mathbf{Z}_1=[h_0, h_1]^T=[0{,}1529\ \ -0{,}0755]$.

Obliczamy:

$$\mathbf{X}_1=\mathbf{X_0}+\mathbf{Z_1}=\begin{bmatrix}1{,}3\\1{,}1\end{bmatrix}+\begin{bmatrix}0{,}1529\\-0{,}0755\end{bmatrix}=\begin{bmatrix}1{,}4529\\1{,}0245\end{bmatrix}.$$

i sprawdzamy warunek zakończenia obliczeń:

$$\|\mathbf{Z}_1\|=0.1706>0.001.$$

Kryterium końca nie jest jeszcze spełnione, wobec czego wykonujemy drugą iterację:

$$\mathbf{F}(\mathbf{X}_1)=\begin{bmatrix}0{,}0120\\-0{,}0231\end{bmatrix},$$

$$\mathbf{J}(\mathbf{X}_1)=\begin{bmatrix}2{,}9058 & -4{,}0978\\ 2{,}0489 & 2{,}9058\end{bmatrix}.$$

Rozwiązujemy układ równań:

$$\mathbf{J}(\mathbf{X}_1)\mathbf{Z}_2=-\mathbf{F}(\mathbf{X}_1),\ \text{czyli:}\begin{bmatrix}2{,}9058 & -4{,}0978\\ 2{,}0489 & 2{,}9058\end{bmatrix}\begin{bmatrix}h_0\\h_1\end{bmatrix}=-\begin{bmatrix}0{,}0120\\-0{,}0231\end{bmatrix}$$

i otrzymujemy poprawkę $\mathbf{Z}_2=[h_0, h_1]^T=[0{,}0036\ \ 0{,}0054]$.

Obliczamy:

$$\mathbf{X}_2 = \mathbf{X}_1 + \mathbf{Z}_2 = \begin{bmatrix} 1{,}4565 \\ 1{,}0299 \end{bmatrix}$$

i ponownie sprawdzamy warunek zakończenia obliczeń:

$\|\mathbf{Z}_2\| = 0.0065 > 0.001.$

Obliczenia kontynuujemy do momentu kiedy $\|\mathbf{Z}_i\| < 0.001.$

Dokładnym rozwiązaniem układu jest:

$$x_1 = \sqrt{\frac{3}{\sqrt{2}}} \approx 1{,}45648, \quad x_2 = \sqrt{\frac{3}{2\sqrt{2}}} \approx 1{,}02988.$$

5.7. Zadania do samodzielnego rozwiązania

Zadanie 5.1.

Metodą Newtona wyznaczyć wzór na przybliżoną wartość:

a) $\sqrt{a} \quad dla\ a > 1,$
b) $\sqrt[s]{a} \quad dla\ a > 1,\ s \in N,\ s > 2.$

Odp.

a) $x_{k+1} = \frac{1}{2}x_k + \frac{a}{2x_k};$
b) $x_{k+1} = \frac{s-1}{s}x_k + \frac{a}{s(x_k)^{s-1}}.$

Zadanie 5.2.

Wyznaczyć metodą Newtona przybliżoną wartość podanego pierwiastka z zadaną dokładnością oraz podać konieczną do wykonania liczbę kroków. Obliczenia przeprowadzić dla punktu startowego 1.0 oraz 2.0.

a) $\sqrt{7}, \ \varepsilon = 0.01;$
b) $\sqrt{20}, \ \varepsilon = 0.001;$
c) $\sqrt[3]{11}, \ \varepsilon = 0.01;$
d) $\sqrt[3]{11}, \ \varepsilon = 0.0001;$
e) $\sqrt[4]{25}, \ \varepsilon = 0.01;$

f) $\sqrt[4]{25},\ \varepsilon = 0.00001.$

Odp.

a) $x_0 = 1.0$ to $x_4 = 2.64577$; $x_0 = 2.0$ to $x_3 = 2.64575$.
b) $x_0 = 1.0$ to $x_5 = 4.47214$; $x_0 = 2.0$ to $x_4 = 4.47214$.
c) $x_0 = 1.0$ to $x_5 = 2.22414$; $x_0 = 2.0$ to $x_3 = 2.22428$.
d) $x_0 = 1.0$ to $x_6 = 2.22398$; $x_0 = 2.0$ to $x_3 = 2.22398$.
e) $x_0 = 1.0$ to $x_8 = 2.23607$; $x_0 = 2.0$ to $x_3 = 2.23607$.
f) $x_0 = 1.0$ to $x_9 = 2.23607$; $x_0 = 2.0$ to $x_4 = 2.23607$.

Zadanie 5.3.

Metodą regula falsi oraz metodą siecznych wyznaczyć dodatni pierwiastek równania:

$$\sin x - \frac{1}{2}x = 0$$

z dokładnością do:

a) 0.01;
b) 0.0001.

Dla każdej metody podać konieczną liczbę kroków.

Odp.

a). Metoda regula falsi: x = 1.89320, liczba kroków = 5.
Metoda siecznych: x = 1.89242, liczba kroków = 3.
b). Metoda regula falsi: x = 1.89546, liczba kroków = 9.
Metoda siecznych: x = 1.89543, liczba kroków = 4.

6. Całkowanie numeryczne

6.1. Wstęp

W niniejszym rozdziale podane zostaną metody numerycznego całkowania funkcji (tzw. kwadratury) [1, 4, 7, 8, 9, 10]. Będą to metody, które jako wynik dadzą przybliżoną wartość liczbową całki postaci:

$$\int_a^b f(x)dx. \tag{6.1}$$

Zakładamy, że jest to całka właściwa tzn. ma skończone granice całkowania oraz funkcja podcałkowa *f(x)* jest ciągła w przedziale całkowania. Obliczenie całki (6.1) nastąpi przy użyciu kwadratur postaci:

$$K_n(f) = \sum_{i=0}^{n} A_i f(x_i). \tag{6.2}$$

Współczynniki kwadratury A_i oraz węzły x_i są niezależne od funkcji podcałkowej *f(x)*. Obliczenie wartości całki (6.1) będzie polegało na aproksymacji funkcji podcałkowej za pomocą łatwo całkowalnej funkcji. W rozdziale podane zostaną dwa rodzaje kwadratur:

- kwadratury Newtona-Cotesa, które oparte są na przybliżeniu funkcji podcałkowej wielomianami stopnia pierwszego (metoda trapezów) oraz stopnia drugiego (metoda Simpsona);
- kwadratury Gaussa, które polegają na wykorzystaniu wielomianów ortogonalnych i takim wyborze punktów x_i oraz współczynników A_i, aby kwadratura była dokładna dla wszystkich wielomianów możliwie najwyższego stopnia.

6.2. Kwadratury Newtona-Cotesa

Niech $l_i(x)$ oznaczają wielomiany fundamentalne Lagrange'a zdefiniowane wzorem:

$$l_i(x) = \prod_{\substack{j=0, \\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j}, \quad i = 0, 1, 2, \dots, n. \tag{6.3}$$

Kwadraturę (6.2) ze współczynnikami A_i zdefiniowanymi wzorem:

$$A_i = \int_a^b l_i(x)dx, \tag{6.4}$$

nazywamy kwadraturą interpolacyjną. Można nią przybliżać wartość całki (6.1). Pozostaje tylko kwestia, z jakim błędem będzie podany wynik. Kwadraturę interpolacyjną $K_n(f)$ z węzłami równoodległymi nazywamy kwadraturą Newtona-Cotesa.

6.2.1. Wzór trapezów

Całkę (6.1) możemy przybliżać jako pole trapezu wyznaczonego przez granice całkowania. Dla przypadku przedstawionego na rys.6.1 kwadratura wyraża się wzorem:

$$K_1(f) = \frac{b-a}{2}(f(a) + f(b)) \tag{6.5}$$

Dla dwóch węzłów $x_0=a$ i $x_1=b$ błąd takiej kwadratury jest dość duży. Dlatego dokonuje się podziału przedziału całkowania na n równej długości podprzedziałów $[x_i, x_{i+1}]$. Krok między węzłami wynosi wtedy:

$$h = \frac{b-a}{n}, \tag{6.6}$$

natomiast węzły spełniają zależność:

$$x_i = a + ih = x_{i-1} + h \quad dla \quad i = 0,1,\dots,n. \tag{6.7}$$

W takim przypadku przybliżenie całki (6.1) jest sumą pól trapezów wyznaczonych przez kolejne podprzedziały:

$$T_n = \sum_{i=0}^{n-1} \frac{h}{2}\left(f(x_i) + f(x_{i+1})\right) = h\left(\frac{1}{2}f(x_0) + \sum_{i=1}^{n-1} f(x_i) + \right.$$
$$\left. + \frac{1}{2}f(x_n)\right)$$

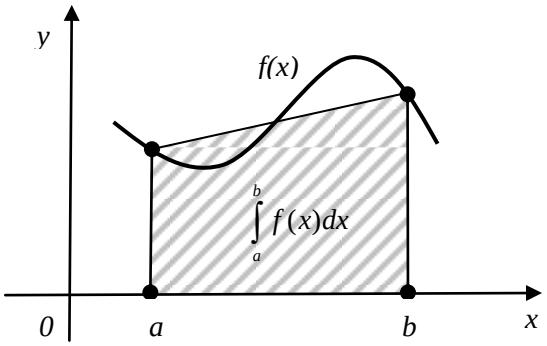
Ostatecznie **wzór złożony trapezów** można zapisać w postaci:

$$T_n = h\left(\frac{1}{2}f(x_0) + f(x_1) + f(x_2) + \cdots + f(x_{n-1}) + \frac{1}{2}f(x_n)\right). \tag{6.8}$$

Błąd tej kwadratury jest rzędu $O(h^2)$ i wynosi dokładnie:

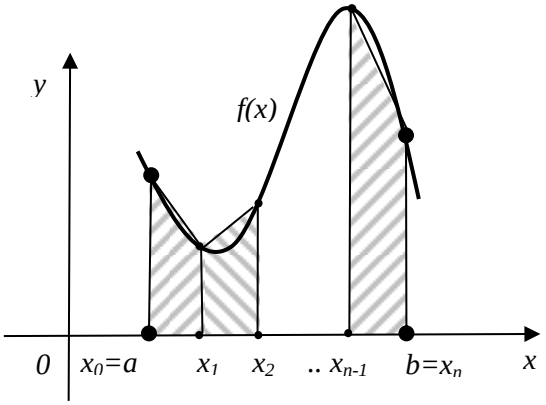
$$-\frac{(b-a)h^2}{12}f''(\xi), \quad gdzie \ \xi \in (a,b). \tag{6.9}$$

Wynika on z twierdzenia o reszcie w interpolacji wielomianowej.



Rys. 6.1. Wzór trapezów dla n=1

Interpretację geometryczną złożonego wzoru trapezów przedstawia rys. 6.2.



Rys. 6.2. Wzór złożony trapezów

Przykład 6.1.

Obliczyć wartość całki $\int_0^1 \sqrt{1+x}dx$ stosując wzór złożony trapezów z krokiem h=1/3.

Wynik obliczony w sposób analityczny to I=1.21895. Liczba podprzedziałow wynosi:

$$n = \frac{b-a}{h} = \frac{1-0}{\frac{1}{3}} = 3,$$

a liczba węzłów jest równa (n+1) czyli w naszym przykładzie 4.

Węzły mają wartości:
$x_0 = 0$ – lewy przedział całkowania,
$x_1 = x_0 + h = 1/3,$
$x_2 = x_1 + h = 2/3,$
$x_3 = x_2 + h = 1$ – prawy przedział całkowania.

Wzór na kwadraturę to:

$$T_3 = h\sum_{i=0}^{3} f(x_i) = h\sum_{i=0}^{3}\sqrt{1+x_i}.$$

Wykonując obliczenia dla naszego przypadku mamy:

$$T_3 = h\left(\tfrac{1}{2}f(x_0) + f(x_1) + f(x_2) + \tfrac{1}{2}f(x_3)\right) =$$

$$= \frac{1}{3}\left(\tfrac{1}{2}\sqrt{1+0} + \sqrt{1+\tfrac{1}{3}} + \sqrt{1+\tfrac{2}{3}} + \tfrac{1}{2}\sqrt{1+1}\right) = 1.21760$$

Wartość błędu wynosi dla n=3 wynosi:

$$\varepsilon = \left|\frac{T_3 - I}{I}\right| \cdot 100\% = \left|\frac{1.21760 - 1.21895}{1.21895}\right| \cdot 100\% = 11.08\%$$

Oczywiście dla mniejszego kroku, czyli większej liczby węzłów, wartość błędu też się zmniejszy i wynik będzie dokładniejszy.

6.2.2. Wzór Simpsona

Całkę (6.1) możemy przybliżać również jako pole pod parabolą przechodzącą przez punkty $x_0=a$, $x_1=(a+b)/2$, $x_2=b$, czyli:

$$K_2(f) = \frac{b-a}{6}\left(f(a) + 4f(\frac{a+b}{2}) + f(b)\right). \qquad (6.10)$$

Przy podziale przedziału całkowania na $n/2$ równej długości podprzedziały $[x_i, x_{i+1}]$, przy założeniu parzystości liczby n, otrzymujemy wzór złożony.

Wzór ten ma postać:

$$S = \sum_{i=0}^{n-1} \frac{2h}{6}\left(f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2i+2})\right)$$

$$= \frac{h}{3}\{[f(x_0) + + 4f(x_1) + f(x_2)] + [f(x_2) + 4f(x_3) + f(x_4)] + [f(x_4) + + 4f(x_5) + + f(x_6)] + \cdots + [f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)]\}$$

Ostatecznie wzór złożony Simsona to:

$$S = \frac{h}{3}\left[f(x_0) + 4\left(f(x_1) + f(x_3) + \cdots + f\left(x_{n-1}\right)\right) + 2\left(f(x_2) + + f(x_4) + \cdots + f\left(x_{n-2}\right)\right) + f(x_n)\right] \quad (6.11)$$

Błąd tej kwadratury jest rzędu $O(h^4)$ i wynosi dokładnie:

$$-\frac{(b-a)h^4}{180} f^{(4)}(\xi), \quad gdzie\ \xi \in (a,b). \quad (6.12)$$

Przykład 6.2.

Obliczyć wartość całki $\int_0^1 \sqrt{1+x}dx$ z przykładu 6.1 za pomocą wzoru złożonego Simpsona.

Wartość kroku w tej metodzie musi ulec zmianie – nie może wynosić h=1/3, bo wówczas n nie jest liczba parzystą. Weźmy h=1/4 i wtedy n=4. Węzły wynoszą odpowiednio: 0, ¼, ½, ¾, 1.

Wzór złożony Simpsona dla tego zadania ma postać:

$$S_4 = \frac{h}{3}\left[f(x_0) + 4\left(f(x_1) + f(x_3)\right) + 2\left(f(x_2)\right) + f(x_4)\right] =$$

$$= \frac{1}{12}\left[f(0) + 4\left(f\left(\tfrac{1}{4}\right) + f\left(\tfrac{3}{4}\right)\right) + 2f\left(\tfrac{1}{2}\right) + f(1)\right] =$$

$$= \frac{1}{12}\left[\sqrt{1} + 4\left(\sqrt{\tfrac{5}{4}} + \sqrt{\tfrac{7}{4}}\right) + 2\sqrt{\tfrac{3}{2}} + \sqrt{2}\right] = 1.21895$$

Przy obliczeniach analitycznych, już dla 5 węzłów osiągamy dokładność wyniku 10^{-5}.

Przykład 6.3.

Stosując wzór złożony Simpsona obliczyć przybliżoną wartość ln(7).

W tym celu należy skorzystać z zależności:

$$\ln(a) = \int_1^a \frac{1}{x} dx \quad dla \quad a > 1.$$

Przyjmijmy krok obliczeń h=0.2. Wtedy:

$$n = \frac{7-1}{0.2} = 30.$$

Węzły mają wartości:

$$x_0 = 1, \;\; x_i = x_{i-1} + 0.2 \;\; dla \;\; i = 1,2,\ldots,30.$$

Wzór złożony Simpsona przyjmuje w tym przykładzie postać:

$$S_{30} = \frac{h}{3}[f(x_0) + 4\sum_{i=0}^{14} f(x_{2i+1}) + 2\sum_{i=1}^{14} f(x_{2i}) + f(x_{30})].$$

Funkcja podcałkowa ma postać:

$f(x_i) = \frac{1}{x_i}$.

Węzły wyrażają się zależnościami:

$$x_{2i} = 1 + h \cdot 2i,$$

$$x_{2i+1} = 1 + h \cdot (2i + 1).$$

Ostatecznie

$$S_{30} = \frac{0.2}{3}\left[1 + 4\sum_{i=0}^{14} \frac{1}{1.2+0.4i} + 2\sum_{i=1}^{14} \frac{1}{1+0.4i} + \frac{1}{7}\right] = 1.94596.$$

Wynik dokładny to 1.94591.

Ostateczna postać wzoru złożonego Simpsona przybliżającego ln(7) dla dowolnego parzystego n to:

$$S_n = \frac{h}{3}\left[1 + 4\sum_{i=0}^{n/2} \frac{1}{1+h\cdot(2\cdot i+1)} + 2\sum_{i=1}^{n/2} \frac{1}{1+h\cdot 2\cdot i} + \frac{1}{7}\right]$$

Zagęszczanie węzłów daje następujące wyniki:

- dla n=40 wynik S_{40}=1.94593;
- dla n=50 wynik S_{50}=1.94592;

- dla n=60 wynik S_{60}=1.94591 (wartość jak dla wyniku dokładnego).

Należy nadmienić, iż kwadratury Newtona-Cotesa nie są numerycznie poprawne tzn.mogą zdarzyć się przypadki, gdy błąd wytworzony w trakcie obliczeń może znacznie przekroczyć wartość obliczonej całki.

6.3. Kwadratury Gaussa

Kwadratury Gaussa oparta na *n+1* punktach ma być dokładna dla wielomianów: $1, x, x^2, \ldots, x^{2n+1}$. Powstaje układ *2(n+1)* równań nieliniowych z *2(n+1)* niewiadomymi. Okazuje się, że pierwiastki tego układu są miejscami zerowymi odpowiednich wielomianów ortogonalnych. Dzięki temu układ, który należy rozwiązać, redukuje się do *n+1* równań liniowych, z którego wyznacza się współczynniki kwadratury.

Definicja i twierdzenia dotyczące wielomianów ortogonalnych znajdują się w rozdziale 3.6.

Rozważmy teraz całkę I_p określoną wzorem:

$$I_p = \int_a^b \omega(x) f(x) dx, \tag{6.13}$$

gdzie funkcja wagowa ω jest dodatnia na [a,b].

Twierdzenie 6.1

Kwadraturą Gaussa postaci (6.2) opartą na *n+1* węzłach o maksymalnym rzędzie, równym *2n+1*, jest kwadratura interpolacyjna, której węzłami są pierwiastki *(n+1)*-go wielomianu ortogonalnego na [*a,b*] z wagą *ω*.

Dla kwadratur Gaussa węzły kwadratury są zatem pierwiastkami odpowiedniego wielomianu ortogonalnego.

Twierdzenie 6.2

Współczynniki kwadratury Gaussa wyrażają się wzorami:

$$A_i = \frac{a_{n+1}}{a_n} \frac{\|p_n\|^2}{p'_{n+1}(x_i) p_n(x_i)}, \quad i = 0,1,\ldots,n, \tag{6.14}$$

gdzie a_k jest współczynnikiem przy najwyższej potędze *k*-tego wielomianu ortogonalnego p_k, a $x_0, x_1, \ldots, x_n$ są pierwiastkami $(n+1)$-ego wielomianu ortogonalnego.

Twierdzenie 6.3

Jeśli $f(x) \in C^{2n+2}[a,b]$, to reszta kwadratury Gaussa wyraża się wzorem:

$$R_n(f) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!a_{n+1}^2}\int_a^b \omega(x)P_{n+1}^2(x)dx \tag{6.15}$$

W szczególnym przypadku, dla funkcji wagowej $\omega(x) \equiv 1$, współczynniki kwadratury A_i i węzły x_i dobiera się tak, aby kwadratura była dokładna dla wszystkich wielomianów $w(x)$ stopnia nie większego niż $2n+2$. Wtedy kwadratura (6.2) jest kwadraturą Gaussa przybliżającą całkę:

$$\int_{-1}^{1} f(x)dx. \tag{6.16}$$

Jest to osiągnięte, gdy spełniony jest układ:

$$\sum_{i=0}^{n} A_i(x_i)^k = \int_{-1}^{1} x^k dx \quad dla \quad k = 0,1,\ldots,2n+1, \tag{6.17}$$

w którym za funkcję podcałkową przyjmuje się wielomiany bazowe.

Po wyliczeniu prawych stron układ (6.17) można zapisac krócej:

$$\sum_{i=0}^{n} A_i(x_i)^k = \frac{1}{k+1}[1-(-1)^{k+1}] dla \quad k = 0,1,\ldots,2n+1. \tag{6.18}$$

Do obliczenia całki w dowolnym przedziale $[a, b]$ dokonujemy liniowego przekształcenia przedziału $[a, b]$ na przedział [-1, 1]. Wtedy:

$$x = \frac{b-a}{2}t + \frac{b+a}{2}. \tag{6.19}$$

a całka po przekształceniu ma postać:

$$\int_a^b f(x)dx = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b-a}{2}t + \frac{b+2}{2}\right)dt. \tag{6.20}$$

Zalety kwadratur Gaussa:

1. Możemy przy ich użyciu obliczać przybliżone wartości całek osobliwych bardzo często występujące np. w fizyce.
2. Mają one dużo wyższą dokładność.

3. Dla *(n+1)* punktów kwadratura Gaussa jest dokładna dla wielomianów stopnia *2n+1*, gdy kwadratury Newtona-Cotesa są dokładne tylko dla wielomianów *n*-tego stopnia.

Przykład 6.4.

Obliczyć współczynniki i węzły kwadratury Gaussa dla przedziału [-1,1] z funkcją wagową $\omega(x) \equiv 1$ opartej na dwóch węzłach.

Przyjmując *n=1*, mamy kwadraturę rzędu 4. Należy rozwiązać układ równań dla *k* od *0* do *2n+1=3* zgodnie ze wzorem (6.17):

k = 0: $A_0 \cdot (x_0)^0 + A_1 \cdot (x_1)^0 = 2,$

k = 1: $A_0 \cdot (x_0)^1 + A_1 \cdot (x_1)^1 = 0,$

k = 2: $A_0 \cdot (x_0)^2 + A_1 \cdot (x_1)^2 = \frac{2}{3},$

k = 3: $A_0 \cdot (x_0)^3 + A_1 \cdot (x_1)^3 = 0.$

Rozwiązaniem są liczby:

$A_0 = A_1 = 1$

oraz

$x_0 = -\frac{1}{\sqrt{3}}\ i\ x_1 = \frac{1}{\sqrt{3}}.$

Kwadratura dwupunktowa Gaussa dla przedziału [-1,1] i wagi $\omega(x) \equiv 1$ ma postać:

$$K_1(f) = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

i przybliża ona całkę:

$$\int_{-1}^{1} f(x)dx.$$

Przykład 6.5.

Obliczyć współlczynniki i węzły kwadratury Gaussa dla przedziału [-1,1] z funkcją wagową $\omega(x) \equiv 1$ opartej na trzech węzłach. Dla *n=2* będzie to kwadratura rzędu 5.

Należy rozwiązać układ równań dla *k* od *0* do *5* zgodnie ze wzorem (6.16). Rozwiązaniem są liczby:

$A_0 = A_2 = 5/9,\quad A_1 = 8/9$.

oraz

$$x_0 = -\sqrt{\tfrac{3}{5}},\ \ x_1 = 0\ \ i\ \ x_2 = \sqrt{\tfrac{3}{5}}.$$

Kwadratura trzypunktowa Gaussa dla przedziału [-1,1] i wagi: $\omega(x) \equiv 1$ ma postać:

$$K_2(f) = \tfrac{5}{9}\cdot f\left(-\sqrt{\tfrac{3}{5}}\right) + \tfrac{8}{9}\cdot f(0) + \tfrac{5}{9}\cdot f\left(\sqrt{\tfrac{3}{5}}\right).$$

i przybliża ona całkę:

$$\int_{-1}^{1} f(x)dx.$$

Przykład 6.6.

Sprawdzić, czy kwadratura postaci:

$$K = \tfrac{b-a}{3}\left[f(a) + f\left(\tfrac{2a+b}{3}\right) + f\left(\tfrac{a+2b}{3}\right)\right]$$

jest kwadraturą Gaussa przybliżającą wartość całki $\int_a^b f(x)dx$.

K jest kwadraturą o trzech węzłach czyli *n*=2. Trzeba sprawdzić czy zachodzi wzór (6.15) dla:

$$A_0 = A_1 = A_2 = \tfrac{b-a}{3}$$

i węzłów:

$$x_0 = a,\ \ x_1 = \tfrac{2a+b}{3},\ \ \ x_2 = \tfrac{a+2b}{3}.$$

Sprawdzamy, dla *k* = 0, czy zachodzi równość:

$$\sum_{i=0}^{2} A_i (x_i)^0 = \int_a^b x^0 dx\ \ .$$

Licząc stronami otrzymujemy:

$$L = \sum_{i=0}^{2} A_i (x_i)^0 = \sum_{i=0}^{2} A_i \cdot 1 = A_0 + A_1 + A_2 = 3\cdot\tfrac{b-a}{3} = b - a.$$

$$P = \int_a^b x^0 dx = b - a.$$

Dla *k* = 0 zachodzi równość (*L*=*P*).

Podobne sprawdzenie należy zatem przeprowadzić dla *k*=1,2,…,5.

Dla k = 1 otrzymujemy:

$\sum_{i=0}^{2} A_i (x_i)^1 = \int_a^b x^1 dx$.

Licząc stronami mamy:

$$L = \sum_{i=0}^{2} A_i (x_i)^1 = A_0 \cdot x_0 + A_1 \cdot x_1 + A_2 \cdot x_2 = \frac{b-a}{3} \cdot \left(a + \frac{2a+b}{3} + \right.$$
$$\left. + \frac{a+2b}{3} \right) = \frac{b-a}{3} \cdot \frac{6a+3b}{3} = \frac{b-a}{3} \cdot (2a + b)$$

$$P = \int_a^b x^1 dx = \frac{1}{2}(b^2 - a^2).$$

Wobec powyższego L≠P dla k=1 - zatem kwadratura ta nie jest kwadraturą przybliżającą podaną całkę, gdyż nie jest dokładna. Dla kolejnych k nie ma już potrzeby sprawdzania.

Z przykładu 6.6. wynika, iż węzły i wagi nie mogą być dowolnie wybrane w danym przedziale całkowania. Stąd cały proces wyboru tych danych w kwadraturach Gaussa ma istotne uzasadnienie.

W praktyce nie znajduje się jednak punktów i wag Gaussa z układu (6.15), tylko wykorzystuje się wielomiany ortogonalne. Wartości współczynników A_i oraz węzłów x_i dla poszczególnych wartości n oraz odpowiednich wielomianów ortogonalnych są stablicowane, ogólnie dostępne i nie oblicza się ich każdorazowo od nowa, gdyż generowałoby to bardzo duży koszt.

W kolejnych rozdziałach przedstawimy cztery najczęściej używane rodzaje wielomianów ortogonalnych i odpowiadające im kwadratury Gaussa.

6.3.1. Kwadratury Gaussa-Hermite'a

Wielomiany ortogonalne Hermite'a mają postać:

$$H_0(x) = 1, \quad H_1(x) = 2x,$$
$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x). \tag{6.19}$$

Funkcja wagowa dla tych wielomianów to:

$$\omega(x) = e^{-x^2},$$

a przedział całkowania (-∞,∞).

Zatem całkę o wskazanej postaci możemy przybliżać kwadraturami Gaussa-Hermite'a:

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)dx \approx \sum_{i=0}^{n} A_i f(x_i), \tag{6.20}$$

w której węzły x_i są pierwiastkami wielomianu ortogonalnego Hearmite'a stopnia (n+1) a wartości A_i - odpowiadającymi im współczynnikami. W tabeli 6.1 przedstawiono wartości tych współczynników dla wielomianów stopnia od 1 do 4.

Tabela 6.1. Wartości węzłów i wag kwadratur Gaussa-Hermite'a

n	i	x_i	A_i
1	0	-0.707107	0.886227
	1	0.707107	0.886227
2	0	-1.224745	0.295409
	1	0	1.181636
	2	1.224745	0.295409
3	0	-1.650680	0.081313
	1	-0.534648	0.804914
	2	0.534648	0.804914
	3	1.650680	0.081313
4	0	-2.020183	0.019953
	1	-0.958572	0.393619
	2	0	0.945309
	3	0.958572	0.393619
	4	2.020183	0.019953

6.3.2. Kwadratury Gaussa-Laguerre'a

Wielomiany ortogonalne Laguerre'a mają postać:

$$L_n(x) = e^x \frac{d}{dx^n}(x^n e^{-x}). \tag{6.21}$$

Funkcja wagowa dla tych wielomianów to:

$$\omega(x) = e^{-x},$$

a przedział całkowania $[0,\infty)$.

Zatem całkę o wskazanej postaci możemy przybliżać kwadraturami Gaussa-Laguerre'a:

$$\int_0^\infty e^{-x} f(x)dx \approx \sum_{i=0}^{n} A_i f(x_i), \tag{6.22}$$

w której węzły x_i są pierwiastkami wielomianu ortogonalnego Laguerre'a stopnia (n+1) a wartości A_i odpowiadającymi im współczynnikami. W tabeli 6.2 przedstawiono wartości tych współczynników dla wielomianów stopnia od 1 do 5.

Tabela 6.2. Wartości węzłów i wag kwadratur Gaussa- Laguerre'a

n	i	x_i	A_i
1	0	0.585789	0.853553
	1	3.414214	0.146447
2	0	0.415775	0.711093
	1	2.294280	0.278518
	2	6.289945	0.010389
3	0	0.322548	0.603154
	1	1.745761	0.357419
	2	4.536620	0.038888
	3	2.395071	0.000539
4	0	0.263560	0.521756
	1	1.413403	0.398667
	2	3.596426	0.075942
	3	7.085810	0.003612
	4	12.640801	0.000032

6.3.3. Kwadratury Gaussa-Czebyszewa

Wielomiany ortogonalne Czebyszewa są postaci:

$$T_0(x) = 1,\ \ T_1(x) = x,$$
$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x). \tag{6.23}$$

Funkcja wagowa dla tych wielomianów to:

$$\omega(x) = 1/\sqrt{1-x^2},$$

a przedział całkowania [-1,1].

Zatem całkę o wskazanej postaci możemy przybliżać kwadraturami Gaussa-Czebyszewa:

$$\int_{-1}^{1} f(x)\frac{1}{\sqrt{1-x^2}}dx \approx \sum_{i=0}^{n} A_i f(x_i), \tag{6.24}$$

w której węzły x_i są pierwiastkami wielomianu ortogonalnego Czebyszewa stopnia (n+1) a wartości A_i są odpowiadającymi im współczynnikami.

Wartości węzłów i współczynników dla kwadratur Gaussa-Czebyszewa wyrażają się wzorami:

$$x_i = cos\left(\frac{2i-1}{2n}\pi\right),\ A_i = \frac{\pi}{n}. \tag{6.25}$$

6.3.4. Kwadratury Gaussa-Legendre'a

Wielomiany ortogonalne Legendre'a:

$$P_0(x) = 1,\ \ P_1(x) = 1,$$

$$P_{n+1}(x) = \frac{2n+1}{n+1}xP_n(x) - \frac{n}{n+1}P_{n-1}(x). \tag{6.26}$$

Funkcja wagowa dla tych wielomianów to:

$$\omega(x) = 1,$$

przedział całkowania [-1,1]. Zatem całkę (6.27) możemy przybliżać kwadraturami Gaussa- Legendre'a

$$\int_{-1}^{1} f(x)dx \approx \sum_{i=0}^{n} A_i f(x_i), \tag{6.27}$$

w której węzły x_i są pierwiastkami wielomianu ortogonalnego Legendre'a stopnia (n+1) a wartości A_i odpowiadającymi im współczynnikami. W tabeli 6.3 przedstawiono wartości tych współczynników dla wielomianów stopnia od 1 do 5.

Obliczenie wartości kwadratury przybliżającej odpowiednią całkę sprowadza się więc tylko do obliczenia sumy $\sum_{i=0}^{n} A_i f(x_i)$, w której współczynniki A_i oraz węzły x_i należy pobrać z odpowiedniej tabeli. Pozostaje tylko policzenie w tych węzłach wartości funkcji podcałkowej.

Tabela 6.3. Wartości węzłów i wag kwadratur Gaussa-Legendre'a

n	i	x_i	A_i
1	0	-0.577350	1
	1	0.577350	1
2	0	-0.774597	5/9
	1	0	8/9
	2	0.774597	5/9
3	0	-0.861136	0.347855
	1	-0.339981	0.652145
	2	0.339981	0.652145
	3	0.861136	0.347855
4	0	-0.906180	0.236927
	1	-0.538469	0.478629
	2	0	0.568889
	3	0.538469	0.478629
	4	0.906180	0.236927

6.4. Zadania do samodzielnego rozwiązania

Zadanie 6.1.

Obliczyć przybliżoną wartość całki

$$\int_0^3 -x^4 + 2.01x^2 + 1dx,$$

stosując wzór złożony trapezów oraz wzór złożony Simpsona dla liczby węzłów równej:

a) 7;
b) 10;
c) 13;
d) 25;
e) 31.

Odp.

Uwaga: *n = liczba węzłów – 1*.

a) T_6 = -35.5025 S_6 = -33.535;
b) T_9 = -34.3971 S_9 – zbyt duży błąd, gdyż n nie jest parzyste;
c) T_{12} = -34.0093 S_{12} = -33.5116;

d) T_{24} = -33.6349 S_{24} = -33.5101;
e) T_{30} = -33.5899 S_{30} = -33.51.

Zadanie 6 2.

Obliczyć przybliżoną wartość całki:

$\int_0^1 e^{-x} + xsin(x)dx$

stosując wzór złożony trapezów oraz wzór złożony Simpsona dla liczby węzłów równej:

a) 7;
b) 21.

Odp.

a) T_6 = 0.937954 S_6 = 0.933279;
b) T_6 = 0.933709 S_6 = 0.933289.

Zadanie 6.3.

Obliczyć wzorem złożonym Simpsona z dokładnością do 10^{-3} pole pod jednym łukiem:

a) sinusoidy;
b) cosinusoidy.

Odp.

a) S_6 = 2.00086;
b) S_6 = 2.00086.

Zadanie 6.4.

Obliczyć dwupunktową kwadraturą Gaussa przybliżoną wartość całki:

$\int_{-\sqrt{3}+1}^{\sqrt{3}+1} e^{\frac{x}{2}}(\sin(x) - 1)dx.$

Odp. $K_1 = -2.1591.$

Zadanie 6.5.

Obliczyć trzypunktową kwadraturą Gaussa przybliżoną wartość całki:

$\int_{-\sqrt{5+\sqrt{3}}}^{\sqrt{5+\sqrt{3}}} 4x^4 - 2x^2 dx.$

Odp. $K_2 = 745.356.$

Zadanie 6.6.

Obliczyć dwupunktową oraz trzypunktową kwadraturą Gaussa przybliżone wartości całek:

a) $\int_1^3 \frac{1}{x} dx$;
b) $\int_0^1 x^2 e^x dx$;
c) $\int_0^{2\pi} x sin(x) dx$;
d) $\int_0^1 \sin(\pi x)\, dx$.

Odp.

a) $K_1 = 1.09091$, $K_2 = 1.09804$
b) $K_1 = 0.711942$, $K_2 = 0.718252$
c) $K_1 = 1.97996$, $K_2 = 1.89115$
d) $K_1 = 0.616191$, $K_2 = 0.637062$.

Zadanie 6.7.

Wykorzystując odpowiednie wielomiany ortogonalne (a dokładnie ich pierwiastki i odpowiadające im wagi) obliczyć przybliżone wartości całek:

a) $\int_{-\infty}^{\infty} e^{-x^2} x^2 dx$, kwadraturą Gaussa-Hermite'a dla *n*=4;
b) $\int_{-\infty}^{\infty} e^{-x^2} \cos(x) dx$, kwadraturą Gaussa-Hermite'a dla *n*=4;
c) $\int_0^{\infty} e^{-x} x^5 dx$, kwadraturą Gaussa-Laguerre'a dla *n*=4;
d) $\int_0^{\infty} \frac{sinx}{x} e^{-2x} dx$, kwadraturą Gaussa-Laguerre'a dla *n*=3;
e) $\int_{-1}^{1} \frac{x^2}{\sqrt{1-x^2}} dx$, kwadraturą Gaussa-Czebyszewa dla *n*=60;
f) $\int_{-1}^{1} (1-x^2)^{3/2} cos(x) dx$, kwadraturą Gaussa-Czebyszewa dla *n*=100;
g) $\int_{-1}^{1} \frac{x^3}{sin(x)cos(x)} dx$, kwadraturą Gaussa-Legendre'a dla *n*=3;
h) $\int_{-1}^{1} \frac{1}{1+x^2} dx$, kwadraturą Gaussa-Legendre'a dla *n*=4.

Odp.

a) K = 0.886224;
b) K = 1.38039;
c) K = 122.789;
d) K = 0.464436;

e) K = 1.62312;
f) K = 1.08294;
g) K = 1.06186;
h) K = 1.57117.

7. Rozwiązywanie równań i układów równań różniczkowych zwyczajnych

7.1. Wstęp

W rozdziale tym zajmiemy się rozwiązaniem ***zagadnienia początkowego***, polegającego na znalezieniu funkcji $y_i = y_i(x)$, $i = 1,2,\ldots,m$, spełniających w przedziale $<x_0,b>$ układ równań różniczkowych:

$$\begin{cases} \dfrac{dy_1}{dx} = f_1(x, y_1,\ldots y_m) \\ \dfrac{dy_2}{dx} = f_2(x, y_1,\ldots y_m) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \dfrac{dy_m}{dx} = f_m(x, y_1,\ldots y_m) \end{cases} \tag{7.1}$$

z warunkiem początkowym:

$$y_i(x_0) = y_{i,0},\ i = 1,2,\ldots,m. \tag{7.2}$$

Zagadnienie powyższe możemy zapisać krótko w postaci wektorowej:

$$\frac{d\mathbf{y}}{dx} = \mathbf{f}(x,\mathbf{y}), \quad \mathbf{y}(x_0) = \mathbf{y}_0, \tag{7.3}$$

gdzie:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \ldots \\ y_m \end{bmatrix},\ \mathbf{f} = \begin{bmatrix} f_1 \\ \ldots \\ f_m \end{bmatrix},\ \mathbf{y_0} = \begin{bmatrix} y_{1,0} \\ \ldots \\ y_{m,0} \end{bmatrix}.$$

Równanie (7.3) nazywamy równaniem różniczkowym zwyczajnym rzędu pierwszego. Wykorzystując tę właśnie postać można łatwo opisać metody numeryczne rozwiązywania układów - tak samo, jak pojedynczych równań [1, 4, 5, 8, 9, 10].

W przypadku m=1 zagadnienie (7.3) sprowadza się do jednego równania skalarnego:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f(x, y), \quad y(x_0) = y_0 . \tag{7.4}$$

W dalszych rozważaniach będziemy zakładać, że:

1. Funkcje $f_i(x, y_1, \ldots, y_m),\ i = 1,2,\ldots,m$, jako funkcje ($m$+1) zmiennych są ciągłe w zbiorze:
   $\mathbf{D} = \{(x, y_1, y_2, \ldots, y_m): x_0 \le x \le b, -\infty < y_i < +\infty, i = 1,2,\ldots,m\}$.
2. Funkcje $f_i(x, y_1, ..., y_m),\ i = 1,2,...,m$ spełniają w zbiorze **D** warunek ***Lipschitza*** względem zmiennych $y_j,\ j = 1,2,\ldots,m$, tzn. istnieje skończona liczba L (***stała Lipschitza***) taka, że dla każdego $x \in < x_0, b >$ i dowolnych $y_j, \tilde{y}_j$ zachodzą nierówności (7.5).

$$\left| f_i(x, y_1, ..., y_m) - f_i(x, \tilde{y}_1, ..., \tilde{y}_m) \right| \le L \sum_{j=1}^{m} \left| y_j - \tilde{y}_j \right| . \tag{7.5}$$

Nierówność (7.5) jest spełniona, gdy w każdym punkcie obszaru **D** funkcje $f_i(x, y_1, ..., y_m),\ i = 1,2,...,m$ mają pochodne cząstkowe:

$$\frac{\partial f_i}{\partial y_j},\quad i, j = 1,2,...,m, \text{ ograniczone w } \mathbf{D}.$$

Przy powyższych założeniach można udowodnić, że w przedziale $< x_0, b >$ istnieje dokładnie jedno rozwiązanie klasy C^1 zagadnienia (7.3).
Przy rozwiązywaniu równań różniczkowych metodami numerycznymi ważne jest, że każde równanie rzędu wyższego niż 1 można wyrazić jako układ równań różniczkowych rzędu pierwszego postaci (7.1) lub (7.3).

Przykład 7.1.

Rozpatrzmy równanie różniczkowe rzędu drugiego:

$$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + \frac{\mathrm{d}y}{\mathrm{d}x} - (y - y^3) = \sin(x),$$

które można przedstawić w postaci:

$$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} = f\left(x, y, \frac{\mathrm{d}y}{\mathrm{d}x}\right),$$

gdzie:

$$f(x, y, \frac{\mathrm{d}y}{\mathrm{d}x}) = \sin(x) + (y - y^3) - \frac{\mathrm{d}y}{\mathrm{d}x}.$$

Wprowadzając następujące zmienne:

$y_1 = y,$

$$y_2 = \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y_1}{\mathrm{d}x},$$

równanie powyższe można zapisać w postaci układu równań różniczkowych zwyczajnych rzędu pierwszego:

$$\begin{cases} \dfrac{\mathrm{d}y_1}{\mathrm{d}x} = y_2 \\ \dfrac{\mathrm{d}y_2}{\mathrm{d}x} = \sin(x) + (y_1 - y_1^3) - y_2 \end{cases}.$$

Przykład 7.2.

Rozważmy równanie różniczkowe rzędu trzeciego postaci:

$$\frac{\mathrm{d}^3 y}{\mathrm{d}x^3} = f(x, y, \frac{\mathrm{d}y}{\mathrm{d}x}, \frac{\mathrm{d}^2 y}{\mathrm{d}x^2}),$$

z warunkami początkowymi:

$$y(0) = \eta_1, \quad y'(0) = \frac{\mathrm{d}y}{\mathrm{d}x}(0) = \eta_2, \quad y''(0) = \frac{\mathrm{d}^2 y}{\mathrm{d}x^2}(0) = \eta_3.$$

Po podstawieniu:

$$y_1 = y,\ y_2 = \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y_1}{\mathrm{d}x},\ y_3 = \frac{\mathrm{d}^2 y}{\mathrm{d}x^2} = \frac{\mathrm{d}y_2}{\mathrm{d}x},$$

równanie powyższe przekształca się na układ:

$$\begin{cases} \dfrac{\mathrm{d}y_1}{\mathrm{d}x} = y_2 \\ \dfrac{\mathrm{d}y_2}{\mathrm{d}x} = y_3 \\ \dfrac{\mathrm{d}y_3}{\mathrm{d}x} = f(x, y_1, y_2, y_3) \end{cases}$$

z warunkami początkowymi:

$y_1(0) = \eta_1,\ y_2(0) = \eta_2,\ y_3(0) = \eta_3.$

Otrzymano więc układ równań różniczkowych rzędu pierwszego.

Przykład 7.3.

Jeśli w równaniu różniczkowym (7.3) zamiast zmiennej x wstawimy t (czas), to równanie to będzie opisywać wektor prędkości cząsteczki jako funkcję wektora jej położenia y. Tak więc równanie różniczkowe określa w tym przypadku ***pole wektorowe***. Rozwiązanie równania opisuje ruch cząsteczki w takim polu.

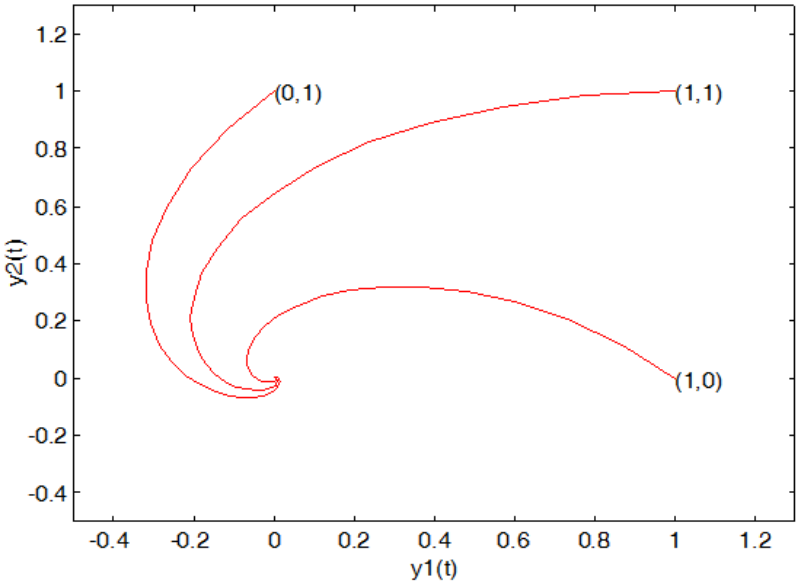
Rozważmy układ:

$$\begin{cases} \dfrac{dy_1}{dt} = -(y_1 + y_2) \\ \dfrac{dy_2}{dt} = y_1 - y_2 \end{cases}$$

z warunkami początkowymi:

$y_1(0) = \eta_1,\ y_2(0) = \eta_2$.

Dla różnych warunków początkowych otrzymujemy tu całą ***rodzinę rozwiązań***. Ruch cząsteczki w danym polu prędkości jest jednoznacznie określony przez jej położenie początkowe. Kilka takich krzywych pokazano na rys. 7.1.



Rys. 7.1. Rozwiązanie zagadnienia z przykładu 7.3 dla różnych warunków początkowych przy $t \in < 0{,}100 >$

Rysunek 7.1. obrazuje trzy przypadki:

a) $y_1(0)=1,\ y_2(0)=0$,

b) $y_1(0)=1,\ y_2(0)=1$,

c) $y_1(0)=0,\ y_2(0)=1$

i sugeruje, że jeśli wybiera się dostatecznie krótkie przedziały czasu, to zasada „*przesunięcie=przyrost czasu x średnia prędkość*" pozwala budować krok po kroku rozwiązanie przybliżone. Jest to podstawowa zasada większości metod całkowania numerycznego równań różniczkowych.

Mniej lub bardziej wyszukane konstrukcje „*średniej prędkości*" w rozpatrywanym przedziale czasu dają różne metody rozwiązania, określane niekiedy ***symulacją dynamiczną*** lub ***ciągłą układu***.

Podstawowe definicje i oznaczenia

W rozdziale tym omówimy tzw. ***metody dyskretne*** rozwiązywania zagadnienia (7.3). Są to metody, za pomocą których otrzymujemy przybliżone rozwiązania tylko dla dyskretnych wartości $x_i, i=1,2,...,N$ zmiennej niezależnej *x*. Ograniczymy się tu do omówienia najważniejszych z tych metod: ***metod różnicowych*** i ***metod typu Rungego-Kutty***. Polegają one na tym, że poszukiwane rozwiązanie $\mathbf{y}_{n+1}$, będące przybliżoną wartością funkcji *y(x)* w punkcie x_{n+1}, obliczane jest metodą iteracyjną w *n*+1 krokach z wykorzystaniem wartości $\mathbf{y}_{n-j}, j=0,1,...$ oraz wartości $\mathbf{f}(x_i,\mathbf{y}_i)$ w obliczonych wcześniej punktach dla *n=0,1,...,N-1*.

Niech $\mathbf{Y}(x)$ oznacza dokładne rozwiązanie zagadnienia (7.3), a $\mathbf{y}(x)$ rozwiązanie przybliżone. Do dalszych rozważań wprowadzimy następujące oznaczenia:

$$\mathbf{Y}_i=\mathbf{Y}_i(x),$$

$$\mathbf{Y}_i'=\frac{\mathrm{d}\mathbf{Y}(x_i)}{\mathrm{d}x}=\mathbf{f}(x_i,\mathbf{Y}_i),$$

$$\mathbf{y}_i=\mathbf{y}_i(x),$$

$$\mathbf{y}_i'=\frac{\mathrm{d}\mathbf{y}(x_i)}{\mathrm{d}x}=\mathbf{f}(x_i,\mathbf{y}_i),$$

gdzie $x_i \in <x_0,b>,\ i=1,2,\ldots,N$, są punktami, w których wyznaczamy przybliżone rozwiązania. Zauważmy, że wektor (funkcja) **y** jest określony

tylko w punktach x_i i oznaczenie $\mathbf{f}(x_i, \mathbf{y}_i)$ symbolem $\mathbf{y}'_i$ jest jedynie umowne.

W przypadku jednego równania różniczkowego (m=1) stosujemy te same oznaczenia, wstawiając zamiast wektorów wartości skalarne. Wprowadzając wzory zarówno w metodach różnicowych, jak i w metodach typu Rungego-Kutty zakładamy, że punkty x_i są równoodległe:

$$x_i = x_0 + ih,$$

gdzie h jest ***krokiem całkowania***.

Rozwiązanie $\mathbf{y}_{n+1}$ wyznacza się za pomocą obliczonych wcześniej wartości $\mathbf{y}_n, \mathbf{y}_{n-1}, \ldots, \mathbf{y}_{n-k}$ $(k \geq 0)$, co można zapisać w postaci:

$$\mathbf{A}_n(h, \mathbf{y}_{n-k}, \ldots, \mathbf{y}_n; \mathbf{y}_{n+1}) = 0, \tag{7.6}$$

przy czym równanie (7.6) może być liniowe lub nieliniowe względem niewiadomej $\mathbf{y}_{n+1}$. Po podstawieniu dokładnego rozwiązania $\mathbf{Y}_i$ do (7.6), otrzymujemy równanie:

$$\mathbf{A}_n(h, \mathbf{Y}_{n-k}, \ldots, \mathbf{Y}_n; \mathbf{Y}_{n+1}) = \mathbf{T}_n. \tag{7.7}$$

Wielkość $\mathbf{T}_n$ nazywamy ***błędem metody*** powstałym przy przejściu od x_n do x_{n+1}. Jeżeli błąd $\mathbf{T}_n$ jako funkcję zmiennej h można przedstawić w postaci:

$$\mathbf{T}_n = h^{p+1}\gamma_p + \mathbf{O}(h^{p+2}), \tag{7.8}$$

gdzie $\gamma_p \neq 0$, to liczbę p nazywa się ***rzędem dokładności*** lub po prostu ***rzędem metody przybliżonej.***

Metody numeryczne pozwalają zwykle określić rozwiązanie:

$$\mathbf{y}(x_k) = \mathbf{y}_k$$

w sposób przybliżony w stosunku do wartości właściwej $\mathbf{Y}_k$. Wielkość:

$$\varepsilon = \|\mathbf{y}_k - \mathbf{Y}_k\|$$

przedstawia ***błąd całkowity*** dla $x = x_k$, na który składa się ***błąd obcięcia***, czyli błąd algorytmiczny zależny od charakteru algorytmu użytego do obliczenia $\mathbf{y}(x)$, a także ***błąd zaokrąglenia*** (błąd maszynowy spowodowany skończoną długością słowa maszynowego). Oba rodzaje błędów kumulują się w kolejnych krokach i stąd, w celu porównania

algorytmów, zamiast błędu całkowitego wygodniej jest używać błędu lokalnego.

Błąd lokalny przy $x=x_1$ definiuje się podobnie:

$$\varepsilon_1 = \|\mathbf{y}_1 - \mathbf{Y}_1\|,$$

przy założeniu, że wartość **y** dla poprzedniego kroku była dokładną wartością funkcji $y(x)$ w punkcie x_1.

Algorytm nazywamy ***numerycznie stabilnym***, gdy lokalny błąd zaokrąglenia maleje ze wzrostem liczby kroków. W przeciwnym wypadku algorytm jest niestabilny i nie przedstawia sobą żadnej wartości praktycznej.

W dalszych rozważaniach zakładamy, że dla rozpatrywanego zagadnienia początkowego (7.1), (7.2), są spełnione podane wcześniej warunki istnienia i jednoznaczności rozwiązania. Założenia te są istotne nie tylko dla istnienia i jednoznaczności rozwiązania, ale także dla ***zbieżności*** rozpatrywanych metod.

7.2. Metoda Eulera

Jedną z najprostszych metod rozwiązywania zagadnienia (7.1), (7.2) jest ***metoda Eulera***, będąca szczególnym przypadkiem zarówno metod różnicowych jak i metod typu Rungego-Kutty.

Rozważmy najpierw przypadek tylko jednego równania różniczkowego (7.4):

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f(x, y), \quad y(x_0) = y_0 .$$

Równanie to dla każdego punktu (x, y) określa nachylenie stycznej do rozwiązania przechodzącej przez dany punkt. Kierunek stycznej zmienia się w sposób ciągły od punktu do punktu, ale najprostsza aproksymacja polega na tym, że rozwiązanie bada się tylko dla pewnych wartości $x=x_0$, x_0+h, x_0+2h, x_0+3h, ... oraz przyjmuje się, że wartość $\mathrm{d}y/\mathrm{d}x$ jest stała między sąsiednimi punktami. Wobec tego równanie aproksymuje się łamaną (rys. 7.2) o wierzchołkach $(0,y_0)$, (h,y_1), $(2h,y_2)$,... , gdzie:

$$\frac{y_{n+1} - y_n}{h} = f(nh, y_n) .$$

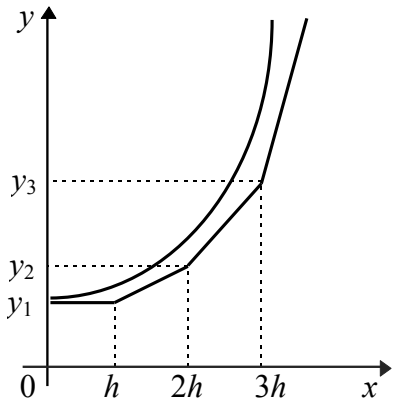
Otrzymujemy stąd prosty wzór rekurencyjny określający metodę Eulera:

$$y_{n+1} = y_n + hf(x_0 + nh, y_n), n = 0,1,... \tag{7.9}$$

Metoda Eulera ma prostą interpretację geometryczną (rys. 7.3). Odcinek $M_i M_{i+1}$ ma w punkcie $M_i = M_i(x_i,y_i)$ kierunek zgodny z kierunkiem stycznej do krzywej całkowej równania $y' = f(x, y)$, która przechodzi przez punkt M_i. Z tego też powodu metoda Eulera jest nazywana często metodą stycznych.

W przypadku układu równań różniczkowych postaci (7.3) metoda Eulera przyjmuje postać ogólną:

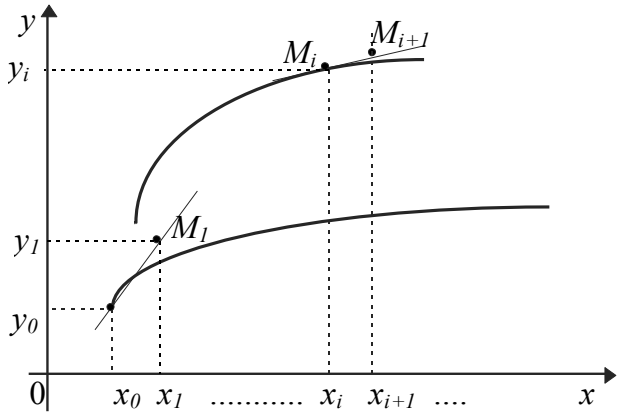
$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(x_n, \mathbf{y}_n), n = 0,1,... \ . \tag{7.10}$$



Rys. 7.2. Aproksymacja rozwiązania łamaną

W metodzie Eulera dla $h \to 0$ w ustalonym punkcie x (x-x_0=nh, $h \to 0, n \to \infty$) rozwiązanie $\mathbf{y}_n$ jest zbieżne do wartości dokładnej $\mathbf{Y}(x)$ i szybkość tej zbieżności wynosi O(h) [4].

Podstawową wadą metody jest fakt, że aby uzyskać dużą dokładność obliczeń, długość kroku musi być bardzo mała, co z kolei prowadzi do wydłużenia czasu obliczeń.



Rys. 7.3. Interpretacja geometryczna metody Eulera

7.3. Metody typu Rungego-Kutty

Metodę określoną wzorem:

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \sum_{i=1}^{s} w_i \mathbf{k}_i \,, \tag{7.11}$$

gdzie:

$$\mathbf{k}_1 = h\mathbf{f}(x_n, \mathbf{y}_n),$$

$$\mathbf{k}_i = h\mathbf{f}(x_n + a_i h, \mathbf{y}_n + \sum_{j=1}^{i-1} b_{ij}\mathbf{k}_j), i > 1,$$

w_i, a_i, b_{ij} - stałe,

nazywamy ***metodą Rungego-Kutty***.

Metodę tę możemy bezpośrednio stosować do rozwiązania zagadnienia (7.1), (7.2), ponieważ w przeciwieństwie do metod różnicowych (patrz rozdział 7.4), do rozpoczęcia obliczeń wystarczy warunek początkowy (7.2).

Istotną częścią obliczeń przy wyznaczaniu rozwiązań $\mathbf{y}_{n+1}$ jest wyznaczenie funkcji **f**(*x*,**y**), występującej po prawej stronie układu równań różniczkowych. Stosując powyższe wzory należy jednak w każdym kroku *s* razy obliczać **f**(*x*,**y**). Dla metody rzędu *p*, należy tych obliczeń wykonać co najmniej *p*.

Największe znaczenie praktyczne mają metody rzędu czwartego dla s=4. W tabeli 7.1 zestawiono wartości współczynników dla najbardziej znanych metod typu (7.11).

Zaletą podwyższenia rzędu algorytmu jest możliwość znacznego wydłużenia kroku h przy zachowaniu tej samej dokładności. Wadą jest znaczne zwiększenie liczby punktów pośrednich obliczeń, które w dodatku nie są wykorzystywane w następnych krokach. Ponadto znaczną trudność stanowi w metodzie precyzyjne oszacowanie długości kroku h dla uzyskania założonej wartości lokalnego błędu obcięcia.

Podamy jeszcze sposób wyboru kroku całkowania w standardowych procedurach bibliotecznych, opartych na metodach czwartego rzędu. Krok początkowy h_0 i dokładność ε są podawane przez użytkownika. Załóżmy, że mamy już rozwiązanie w punkcie x_n. Obliczenia w następnym etapie (tj. przy przejściu od x_n do x_{n+1}) są następujące. Niech h oznacza krok całkowania z poprzedniego etapu obliczeń (jeżeli n=1, to przyjmujemy h=h_0). Za miarę błędu metody na tym etapie przyjmujemy liczbę:

$$\delta = \frac{1}{15}\left\|\mathbf{y}_{n+1}^{(1)} - \mathbf{y}_{n+1}^{(2)}\right\|, \tag{7.12}$$

gdzie:

$\mathbf{y}_{n+1}^{(1)}$ - oznacza przybliżone rozwiązanie w punkcie x_{n+1}=x_n+h, gdy obliczenia wykonane zostały dwukrotnie z krokiem h/2;

$\mathbf{y}_{n+1}^{(2)}$ - przybliżone rozwiązanie w punkcie x_{n+1} liczone od razu z krokiem h.

Mogą wystąpić dwa przypadki:

a) $\delta \le \varepsilon$ - obliczone rozwiązanie $\mathbf{y}_{n+1}^{(1)}$ jest uznawane jako wystarczająco dokładne. Wówczas, jeśli ponadto $\delta \le \varepsilon / 50$, to krok h podwajamy, w przeciwnym razie przechodzimy do następnego etapu.

b) $\delta > \varepsilon$ - obliczone rozwiązanie jest uznawane jako niewystarczająco dokładne. Krok całkowania jest połowiony i obliczenia są wykonywane jeszcze raz.

Tabela 7.1. Wartości współczynników w metodach typu Rungego-Kutty

Rząd metody	Stałe w_i	Wartości współczynników k_i	Metoda
1	$w_1=1$	$\mathbf{k}_1 = h\mathbf{f}(x_n, \mathbf{y}_n)$	***Eulera***
2	$w_1=w_2=1/2$	$\mathbf{k}_1 = h\mathbf{f}(x_n, \mathbf{y}_n)$ $\mathbf{k}_2 = h\mathbf{f}(x_n + h, \mathbf{y}_n + \mathbf{k}_1)$	***Heuna*** (ulepszona *Eulera*)
3	$w_1=w_3=1/6$ $w_2=2/3$	$\mathbf{k}_1 = h\mathbf{f}(x_n, \mathbf{y}_n)$ $\mathbf{k}_2 = h\mathbf{f}(x_n + 0.5h, \mathbf{y}_n + 0.5\mathbf{k}_1)$ $\mathbf{k}_3 = h\mathbf{f}(x_n + h, \mathbf{y}_i - \mathbf{k}_1 + 2\mathbf{k}_2)$	pokrewna metodzie ***Simpsona***
4	$w_1=w_4=1/6$ $w_2=w_3=1/3$	$\mathbf{k}_1 = h\mathbf{f}(x_n, \mathbf{y}_n)$ $\mathbf{k}_2 = h\mathbf{f}(x_n + 0.5h, \mathbf{y}_n + 0.5\mathbf{k}_1)$ $\mathbf{k}_3 = h\mathbf{f}(x_n + 0.5h, \mathbf{y}_n + 0.5\mathbf{k}_2)$ $\mathbf{k}_3 = h\mathbf{f}(x_n + 0.5h, \mathbf{y}_n + 0.5\mathbf{k}_2)$ $\mathbf{k}_4 = h\mathbf{f}(x_n + h, \mathbf{y}_n + \mathbf{k}_3)$	klasyczna ***Rungego-Kutty***

Przykład 7.4.

Znaleźć rozwiązanie równania $y'(x) = y$ z warunkiem początkowym $y(0)=1$ metodą Rungego-Kutty pierwszego rzędu (czyli metodą Eulera) dla $x \in \langle 0, 0.4\rangle$ z krokiem $h = 0.1$.

Metoda Eulera ma postać:

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(x_n, \mathbf{y}_n), n = 0,1,..., N.$$

W zadaniu dane są:
$f(x,y) = y$,
$x_0=0$, $b=0.4$, $y_0=1$, $h=0.1$ liczba kroków $N = \frac{b-x_0}{h} = \frac{0.4-0}{0.1} = 4$.
Obliczamy kolejno:

$$y_1 = y_0 + hf(x_0, y_0) = 1 + 0.1f(0,1) = 1 + 0.1 \cdot 1 = 1.1$$
$$x_1 = x_0 + h = 0 + 0.1 = 0.1$$
$$y_2 = y_1 + hf(x_1, y_1) = 1.1 + 0.1f(0.1,1.1) = 1.1 + 0.1 \cdot 1.1 = 1.21$$
$$x_2 = x_1 + h = 0.1 + 0.1 = 0.2$$

$$y_3 = y_2 + hf(x_2, y_2) = 1.21 + 0.1f(0.2,1.21) = 1.21 + 0.1 \cdot 1.21 = 1.331$$

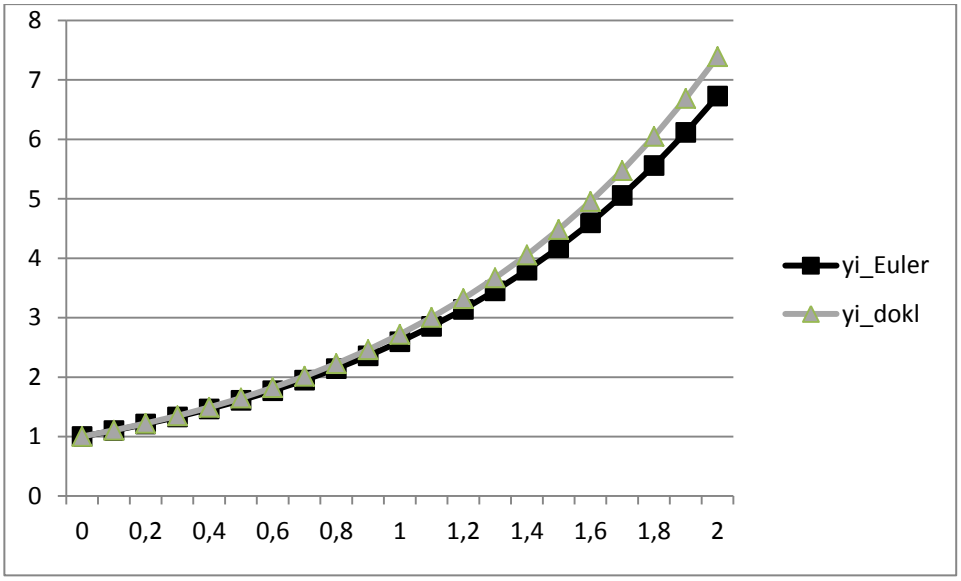
$$x_3 = x_2 + h = 0.2 + 0.1 = 0.3$$

$$y_4 = y_3 + hf(x_3, y_3) = 1.331 + 0.1f(0.3,1.331) = = 1.331 + 0.1 \cdot 1.331 = 1.4641$$

Rozwiązaniem analitycznym przedstawionego równania różniczkowego jest funkcja:

$y(x) = e^x$.

Na rysunku 7.4 zaprezentowane zostały wyniki dla rozwiązania dokładnego i wyniki otrzymane w metodzie Eulera na przedziale <0,2>.



Rys. 7.4. Porównanie wyników z metody Eulera z rozw. dokładnym

Przykład 7.5.

Znaleźć rozwiązanie równania $y'(x) = \cos x - \sin x - y$ z warunkiem początkowym $y(0)=2$ metodą Rungego-Kutty drugiego rzędu (czyli metodą ulepszoną Eulera) dla $x \in <0, 0.3>$ z krokiem $h = 0.1$.

Metoda ulepszona Eulera ma postać:

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{2}\left[\mathbf{f}(x_n, \mathbf{y}_n) + \mathbf{f}(x_n + h, \mathbf{y}_n + h\mathbf{f}(x_n, \mathbf{y}_n))\right], n = 0,1,..., N.$$

W zadaniu dane są:

$$f(x, y) = \cos x - \sin x - y$$

x_0=0, b=0.3, y_0=2, h=0.1, liczba korków $N = \frac{b-x_0}{h} = \frac{0.3-0}{0.1} = 3$.

Obliczamy kolejno:

$$f(x_0, y_0) = f(0,2) = \cos(0) - \sin(0) - 2 = -1,$$

$$\begin{aligned} y_1 = y_0 + \frac{h}{2}[f(x_0, y_0) + f(x_0 + h, y_0 + hf(x_0, y_0))] \\ = 2 + 0.05[f(0,2) + f(0 + 0.1,2 + 0.1f(0,2))] \\ = 2 + 0.05 \cdot [3 + f(0.1,2 + 0.1 \cdot 3] \\ = 2 + 0.05 \cdot [3 + f(0.1,2.3)] \\ = 2 + 0.05 \cdot [3 + \cos(0.1) - \sin(0.1) - 2.3] \\ = 1.88935 \end{aligned}$$

$$x_1 = x_0 + h = 0 + 0.1 = 0.1$$

$$y_2 = y_1 + \frac{h}{2}[f(x_1, y_1) + f(x_1 + h, y_1 + hf(x_1, y_1))] = 1.77802$$

$$x_2 = x_1 + h = 0.1 + 0.1 = 0.2$$

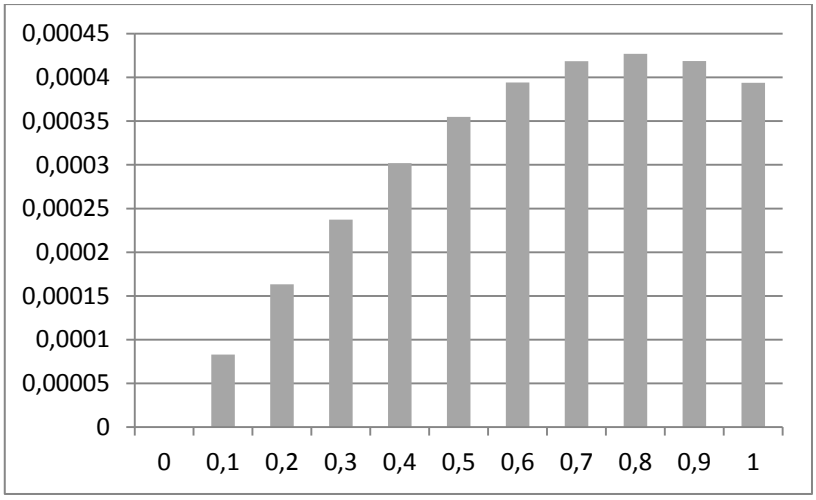
$$y_3 = y_2 + \frac{h}{2}[f(x_2, y_2) + f(x_2 + h, y_2 + hf(x_2, y_2))] = 1.66538$$

W efekcie otrzymaliśmy:

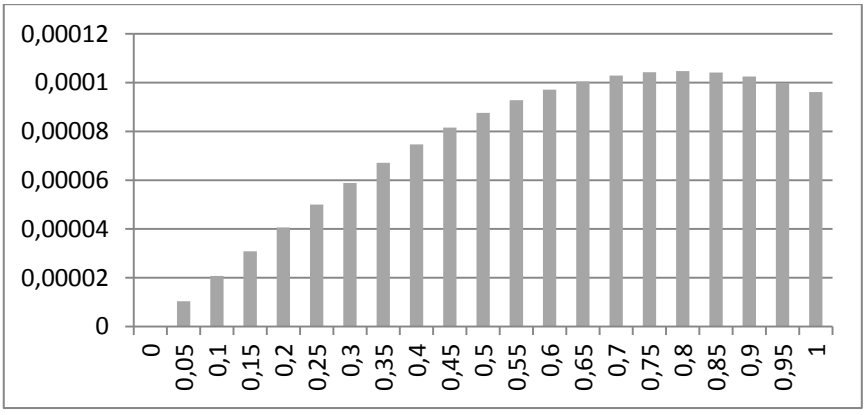
$$y_1 \approx y(x_1) = y(0.1)$$
$$y_2 \approx y(x_2) = y(0.2)$$
$$y_3 \approx y(x_3) = y(0.3).$$

Rozwiązaniem analitycznym przedstawionego równania różniczkowego jest funkcja $y(x) = \cos(x) - \sin(x) - y$.

Na rysunku 7.5 i 7.6 zaprezentowane zostały wartości równe różnicy rozwiązania dokładnego i wyników otrzymanych odpowiednio w metodzie ulepszonej Eulera dla kroku h=0.1 oraz h=0.05.



Rys. 7.5. Błąd w metodzie ulepszonej Eulera dla h=0.1



Rys. 7.6. Błąd w metodzie ulepszonej Eulera dla h=0.05

Przykład 7.6.

Znaleźć rozwiązanie równania $y'(x) = x + y$ z warunkiem początkowym $y(0)=1$ metodą Rungego-Kutty czwartego rzędu dla $x \in <0, 0.2>$ z krokiem $h = 0.1$.

Wzory Rungego-Kutty czwartego rzędu mają postać (tabela 7.1):

$$y_{i+1} = y_i + (k_1 + 2k_2 + 2k_3 + k_4)/6 \qquad i = 0, 1, 2, ...$$

przy czym:

$$k_1 = hf(x_i, y_i),$$

$$k_2 = hf(x_i + 0.5h, y_i + 0.5k_1),$$

$$k_3 = hf(x_i + 0.5h, y_i + 0.5\, k_2),$$

$$k_4 = hf(x_i + h, y_i + k_3),$$

We wzorach tych:

$$y_{i+1} = y(x_{i+1}) = y(x_0 + (i + 1)h).$$

Dla $i = 0$ obliczamy:

$$y1 = y(x1) = y(x0 + h) = y(0.1)$$

$$k1 = h(x0 + y0) = 0.1(0+1) = 0.1$$

$$k2 = 0.1(0+0.1/2+1+0.1/2) = 0.11$$

$$k3 = 0.1(0+0.1/2+1+0.11/2) = 0.1105$$

$$k4 = 0.1(0+0.1+1+0.1105) = 0.12105$$

$$y1 = y0 + 1/6(k1 + 2k2 + 2k3 + k4) =$$

$$= 1+(0.1+2\cdot0.11+2\cdot0.1105+0.12105)/6 = 1.110341.$$

Analogiczne obliczenia przeprowadzamy dla $i = 1$ - dla $x_2 = x_0 + 2h$ obliczamy $y_2 = y(x_2)$. Wartości rozwiązania zestawiono w tabeli 7.2.

Tabela 7.2. Wyniki obliczeń dla przykładu 7.4

i	x	y	$k=0.1(x+y)$
0	$x_0=0$ $x_0+0.5h=0.05$ $x_0+0.5h=0.05$ $x_0+h=0.1$	$y_0=1$ $y_0+0.5k_1=1.05$ $y_0+0.5k_2=1.055$ $y_0+k_3=1.1105$	$k_1=0.1$ $k_2=0.11$ $k_3=0.1105$ $k_4=0.12105$
1	$x_1=0.1$ $x_1+0.5h=0.15$ $x_1+0.5h=0.15$ $x_1+h=0.2$	$y_1=1.11034$ $y_1+0.5k_1=1.170857$ $y_1+0.5k_2=1.17638285$ $y_1+k_3=1.242978285$	$k_1=0.121034$ $k_2=0.1320857$ $k_3=0.132638285$ $k_4=0.1442978285$
2	$x_2=0.2$	$y_2=1.242803$	

7.4. Metody różnicowe (wielokrokowe)

W ***metodzie jednokrokowej*** (np. w metodzie Eulera, metodach typu Rungego-Kutty) do wyznaczenia kolejnego przybliżenia $\mathbf{y}_{n+1}$ wystarcza znajomość tylko poprzedniego $\mathbf{y}_n$. W ***metodzie wielokrokowej*** korzystamy z k obliczonych wcześniej wartości $\mathbf{y}_{n+1-k},\ldots,\mathbf{y}_n$ ($k>1$).

Metodę określoną wzorem:

$$\mathbf{y}_{n+1}=\sum_{i=1}^{k}a_i\mathbf{y}_{n+1-i}+h\sum_{i=0}^{k}b_i\mathbf{f}_i(x_{n+1-i};\mathbf{y}_{n+1-i}),\ n\geq k-1, \tag{7.13}$$

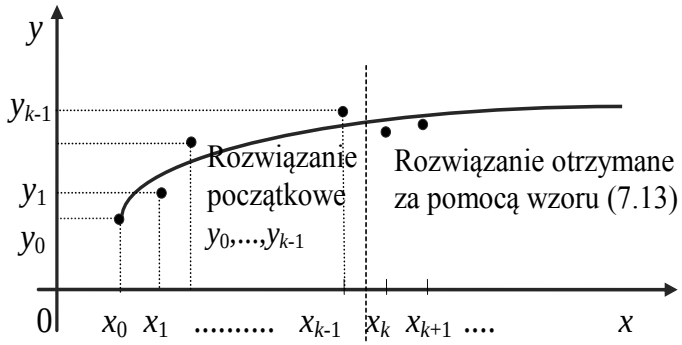
gdzie współczynniki a_i, b_i są liczbami rzeczywistymi, a h jest krokiem całkowania, nazywamy ***metodą różnicową*** (***wielokrokową***).

Jeżeli $k>1$, to powyższą metodę możemy stosować do rozwiązywania zagadnienia (7.1), (7.2) tylko wtedy, gdy znamy wartości rozwiązania w punktach $x_1, x_2, \ldots, x_{k-1}$ (rys. 7.7), bowiem warunek początkowy (7.2) daje nam rozwiązanie tylko w punkcie x_0.

Do wyznaczenia punktów startowych wykorzystuje się zwykle algorytmy jednokrokowe, np. algorytmy Rungego-Kutty, które powtórzone k-krotnie przy ujemnej wartości h pozwalają wyznaczyć wymagane k punktów startowych.

Jeśli $b_0=0$, to wzór (7.13) nazywamy ***wzorem ekstrapolacyjnym*** (***jawnym***). Szukana wartość $\mathbf{y}_{n+1}$, $n\geq k-1$, jest wówczas kombinacją liniową obliczonych wcześniej wartości $\mathbf{y}_{n+1-k},\ldots,\mathbf{y}_n$ oraz $\mathbf{f}_{n+1-k},\ldots,\mathbf{f}_n$, a więc wyznaczenie rozwiązania $\mathbf{y}_{n+1}$ jest stosunkowo proste.

Jeżeli natomiast $b_0 \neq 0$, to wzór (7.13) nazywamy ***wzorem interpolacyjnym*** (***uwikłanym***). Ponieważ szukana wartość $\mathbf{y}_{n+1}$ występuje po obu stronach równania (7.13), więc na ogół wyznaczamy ją stosując metody iteracyjne. Jeżeli jednak układ (7.1) jest liniowy, to $\mathbf{y}_{n+1}$ można wyznaczyć bezpośrednio.



Rys. 7.7. Interpretacja geometryczna wzoru (7.13)

W tabeli 7.3 zestawiono wzory różnicowe spełniające warunki zbieżności (twierdzenia dotyczące zbieżności metod różnicowych można znaleźć w pracy [4]), które najczęściej są omawiane i stosowane w praktyce. Wzory 1-5 są wzorami ekstrapolacyjnymi (jawnymi), a wzory 6-10 są interpolacyjne (uwikłane). Wzory 1-4 są typu ***Adamsa-Bashfortha***, wzór 5 jest typu ***Milne'a***. Wzory 6-9 są typu ***Adamsa-Moultona***, a wzór 10 podał ***Hamming***.

Dla uzyskania maksymalnej efektywności algorytmu całkowania numerycznego pożądany jest dobór optymalnego rzędu algorytmu oraz kroku całkowania. Przy ich doborze należy mieć na względzie zarówno stabilność algorytmu, jak i dopuszczalny błąd całkowania [6].

Tabela 7.3. Zestawienie najczęściej stosowanych wzorów różnicowych

Wzór	Rząd metody *p*
1. $\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{y}'_n$ (ekstrapolacyjny Eulera)	1
2. $\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{2}(3\mathbf{y}'_n - \mathbf{y}'_{n-1})$	2
3. $\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{12}(23\mathbf{y}'_n - 16\mathbf{y}'_{n-1} + 5\mathbf{y}'_{n-2})$	3
4. $\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{24}(55\mathbf{y}'_n - 59\mathbf{y}'_{n-1} + 37\mathbf{y}'_{n-2} - 9\mathbf{y}'_{n-3})$	4
5. $\mathbf{y}_{n+1} = \mathbf{y}_{n-3} + \frac{4}{3}h(2\mathbf{y}'_n - \mathbf{y}'_{n-1} + 2\mathbf{y}'_{n-2})$	4
6. $\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{y}'_{n+1}$ (interpolacyjny Eulera)	1
7. $\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{2}(\mathbf{y}'_{n+1} + \mathbf{y}'_n)$	1
8. $\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{12}(5\mathbf{y}'_{n+1} + 8\mathbf{y}'_n - \mathbf{y}'_{n-1})$	3
9. $\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{24}(9\mathbf{y}'_{n+1} + 19\mathbf{y}'_n - 5\mathbf{y}'_{n-1} + \mathbf{y}'_{n-2})$	4
10. $\mathbf{y}_{n+1} = \frac{1}{8}(9\mathbf{y}_n - \mathbf{y}_{n-2}) + \frac{3}{8}h(\mathbf{y}'_{n+1} + 2\mathbf{y}'_n - \mathbf{y}'_{n-1})$	4

7.5. Metoda Geara dla układów sztywnych

Układy równań różniczkowych (7.3) nazywamy ***układami sztywnymi*** (***źle uwarunkowanymi***), gdy stosunek największej co do modułu wartości własnej macierzy Jacobiego (patrz rozdział 5.6) do najmniejszej co do modułu wartości własnej jest znacznie większy od jedności.

Z takimi układami spotykamy się przy opisie wielu zagadnień inżynierskich np. w dynamice procesów i sterowania.

Macierz Jacobiego ma postać:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial f_1}{\partial y_1} & \frac{\partial f_1}{\partial y_2} & \cdots & \frac{\partial f_1}{\partial y_m} \\ \frac{\partial f_2}{\partial y_1} & \frac{\partial f_2}{\partial y_2} & \cdots & \frac{\partial f_2}{\partial y_m} \\ \cdots & \cdots & & \cdots \\ \frac{\partial f_m}{\partial y_1} & \frac{\partial f_m}{\partial y_2} & \cdots & \frac{\partial f_m}{\partial y_m} \end{bmatrix} \tag{7.14}$$

Aby rozwiązać efektywnie sztywny układ równań należy zastosować algorytm wielokrokowy o zmiennym rzędzie i zmiennym kroku całkowania, przy czym zmiana *h* odbywać się będzie w bardzo szerokim zakresie. Powstają przy tym trudne problemy stabilności algorytmu, zwłaszcza przy dużej wartości kroku *h*.

Do rozwiązania sztywnych układów równań zwykle stosowane są ***algorytmy Geara***.

Algorytm Geara rzędu *k* jest algorytmem wielokrokowym uwikłanym, określanym wzorem (7.13) o $p=k$-1 oraz współczynnikach $b_0 = b_1 = ... = b_{k-1} = 0$. Wartość $\mathbf{y}_{n+1}$ jest określona wówczas w sposób następujący:

$$\mathbf{y}_{n+1} = \sum_{i=0}^{k-1} a_i \mathbf{y}_{n-i} + h b_{-1} f(x_{n+1}, \mathbf{y}_{n+1}) . \tag{7.15}$$

Przykładowe wzory algorytmów Geara rzędu od jednego do czterech zestawiono w tabeli 7.4 [6]. Algorytm Geara rzędu pierwszego jest poznanym wcześniej algorytmem interpolacyjnym Eulera. Podobnie jak algorytm Adamsa-Bashfortha, algorytm Geara rzędu *k*-tego wymaga *k* wartości startowych, a więc jest algorytmem *k*-krokowym.

Badania stabilności bezwzględnej algorytmów Geara wykazują, że obszar stabilności bezwzględnej jest znacznie szerszy niż najlepszego dotąd algorytmu Adamsa-Moultona oraz spełnia wszystkie wymagane warunki potrzebne do efektywnego i dokładnego rozwiązania sztywnych równań różniczkowych.

Tabela 7.4. Algorytmy Geara rzędu 1-4

Wzór	Rząd metody
1. $\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{y}'_{n+1}$ (interpolacyjny Eulera)	1
2. $\mathbf{y}_{n+1} = \frac{4}{3}\mathbf{y}_n - \frac{1}{3}\mathbf{y}_{n-1} + \frac{2}{3}h\mathbf{y}'_{n+1}$	2
3. $\mathbf{y}_{n+1} = \frac{18}{11}\mathbf{y}_n - \frac{9}{11}\mathbf{y}_{n-1} + \frac{2}{11}\mathbf{y}_{n-2} + \frac{6}{11}h\mathbf{y}'_{n+1}$	3
4. $\mathbf{y}_{n+1} = \frac{48}{25}\mathbf{y}_n - \frac{36}{25}\mathbf{y}_{n-1} + \frac{16}{25}\mathbf{y}_{n-2} - \frac{3}{25}\mathbf{y}_{n-3} + \frac{12}{25}h\mathbf{y}'_{n+1}$	4

7.6. Zadania do samodzielnego rozwiązania

Zadanie 7.1.

Metodą Eulera znaleźć rozwiązanie równania podanych równań różniczkowych z warunkiem początkowym dla x z podanego przedziału oraz z podanym krokiem h:

a) $y' = 2y,\ \ y(0) = 1,\ \ x \in \langle 0,2\rangle,\ \ h = 0.25;$
b) $y' = \frac{x}{y^2},\ \ y(1) = 0,\ \ x \in \langle 1,2\rangle,\ \ h = 0.2;$
c) $y' = 2xe^x,\ \ y(0) = 0,\ \ x \in \langle 0,2\rangle,\ \ h = 0.5;$
d) $y' = 2xe^x,\ \ y(0) = 0,\ \ x \in \langle 0,2\rangle,\ \ h = 0.1;$
e) $y' = e^{-x} - 2x,\ \ y(0) = -1,\ \ x \in \langle 0,1\rangle,\ \ h = 0.2;$
f) $y' = (2x + ysin(x))/\cos(x),\ \ y(0) = 0,\ \ x \in \langle 0,2\rangle,\ \ h = 0.2.$

Odp.

Tablice z wartościami funkcji dla kolejnych wartości x:

a) F=[1, 1.25, 2.25, 3.375, 5.063, 7.594, 11.391, 17.086, 25.629];
b) F=[1, 1.24, 1.422, 1.58, 1.724, 1.859];
c) F=[0, 0.824, 3.543, 10.265, 25.043];
d) F=[0, 0.022, 0.071, 0.152, 0.271, 0.436, 0.655, 0.937, 1.293, 1.736, 2.279, 2.94, 3.737, 4.691, 5.827, 7.171, 8.756, 10.617, 12.795, 15.336, 18.291];
e) F=[-1, -0.916, -0.942, -1.072, -1.303, -1.629];
f) F=[0, 0.082, 0.262, 0.589, 1.169, 2.274, 4.769, 13.593, -101.389, -17.642, -11.855].

Zadanie 7.2.

Ulepszoną metodą Eulera znaleźć rozwiązanie równania podanych równań różniczkowych z warunkiem początkowym dla *x* z podanego przedziału oraz z podanym krokiem *h*:

a) $y' = \frac{y}{x} + x,\ \ y(1) = 2,\ \ x \in \langle 1{,}2\rangle,\ \ h = 0.1;$
b) $y' = x + xy + y + 1,\ \ y(-1) = 1,\ \ x \in \langle -1{,}1\rangle,\ \ h = 0.25;$
c) $y' = \frac{x^3 cosx + 2y}{x},\ \ y(\pi) = \pi^2,\ \ x \in \langle \pi, 2\pi\rangle,\ \ h = \frac{\pi}{10};$
d) $y' = y - 2x^2 e^{-x},\ \ y(0) = 0.5,\ \ x \in \langle 0{,}2\rangle,\ \ h = 0.2;$
e) $y' = \frac{y}{x}(lny - lnx),\ \ y(1) = e^2,\ \ x \in \langle 1{,}3\rangle,\ \ h = 0.1.$

Odp.

Tablice z wartościami funkcji dla kolejnych wartości *x*:

a) F=[2, 2.301, 2.623, 2.964, 3.326, 3.707, 4.108, 4.53, 4.971, 5.433, 5.914];
b) F=[1, 1.195, 1.564, 2.185, 3.205, 4.897, 7.777, 12.851, 22.157];
c) F=[9.87, 7.851, 5.587, 3.777, 3.271, 4.947, 9.588, 17.737, 29.574, 44.821, 62.704];
d) F=[0.5, 0.581, 0.643, 0.68, 0.687, 0.663, 0.608, 0.522, 0.406, 0.258, 0.078];
e) F=[7.389, 8.743, 10.273, 12.001, 13.947, 16.137, 18.598, 21.36, 24.457, 27.926, 31.807, 36.146, 40.991, 46.398, 52.428, 59.146, 66.626, 74.949, 84.205, 94.491, 105.916].

Zadanie 7.3.

Wyznaczyć przybliżone wartości rozwiązań równań różniczkowych z warunkiem początkowym metodą ulepszoną Eulera w podanym punkcie z zadanym krokiem.

a) $y' = 2xy^2 - 2xy,\ \ y(0) = 0.5,\ \ h = 0.1,\ \ x = 2;$
b) $y' = (x - y)^2 + 1,\ \ y(1) = -1,\ \ h = 0.1,\ \ x = 2;$
c) $y' = -sinx\sqrt{y},\ \ y(\pi) = 0,\ \ h = \frac{\pi}{5},\ \ x = \frac{3}{2}\pi.$

Odp.

a) $y(2) \approx 1.1;$
b) $y(2) \approx 1.1;$
c) $y(\frac{3\pi}{2}) \approx 1.1.$

Bibliografia

1. Dahlquist G., Björck A.: „Metody numeryczne”. PWN, Warszawa, 1983.
2. Guziak T., Kamińska A., Pańczyk B., Sikora J., „Metody numeryczne w elektrotechnice”, Wydawnictwo Politechniki Lubelskiej, 2002.
3. Engeln-Mullges G., Uhlig F.: „Numerical Recipes in C”. Springer, 1996.
4. Fortuna Z., Macukow B., Wąsowski J.: „Metody numeryczne”. Warszawa, WNT, 1993.
5. Jankowscy J. i M.: „Przegląd metod i algorytmów numerycznych”. WNT, Warszawa, 1988.
6. Osowski S. “Komputerowe metody analizy i optymalizacji obwodów elektrycznych”, Wydawnictwo Politechniki Warszawskiej, Warszawa 1993.
7. Press W.H., Flannery B.P., Teukolsky S.A., Vetterling W.T.: „Numerical Recipes”. Cambridge University Press, Cambridge, 1985.
8. Ralston A.: „Wstęp do analizy numerycznej”. PWN, Warszawa, 1975.
9. Stoer J.: „Wstęp do metod numerycznych”. PWN, Warszawa, 1979.
10. Stoer J., Bulirsch R.: „Wstęp do analizy numerycznej”. PWN, Warszawa, 1987.
11. Wit R.: „Metody programowania nieliniowego”. WNT, Warszawa, 1986.

Indeks

L

M

N

O

P

R

S

T

U

W

Z