

eit

WYDAWNICTWA NAUKOWO-TECHNICZNE
WARSZAWA

Zenon Fortuna
Bohdan Macukow
Janusz Wąsowski

Metody numeryczne

Wydanie szóste



Redaktor mgr LILIANA SZYMAŃSKA
Opracowanie graficzne TADEUSZ PIETRZYK
Redaktor techniczny IRENA MILEWSKA-BURCZYKOWA

518.12/681.3

W podręczniku podano wybrane metody stosowane przy obliczeniach komputerowych. Omówiono tu niektóre z ważniejszych metod interpolacyjnych i aproksymacyjnych, całkowanie numeryczne, metody rozwiązywania układów algebraicznych równań liniowych oraz równań różniczkowych. Zamieszczono wiele algorytmów obliczeniowych wraz z oceną ich efektywności. Materiał teoretyczny jest ilustrowany licznymi przykładami rachunkowymi, przykładami programów w językach Fortran i Pascal i ich schematów blokowych. Podręcznik jest przeznaczony dla studentów kierunków: elektronika, informatyka i telekomunikacja. Mogą z niego także korzystać studenci i inżynierowie innych specjalności.

1982 — wydanie pierwsze
1993 — wydanie drugie poprawione
1995 — wydanie trzecie
1998 — wydanie czwarte
2001 — wydanie piąte

© Copyright by Wydawnictwa Naukowo-Techniczne
Warszawa 1982, 1993

All Rights Reserved
Printed in Poland

Utwór w całości ani we fragmentach nie może być powielany ani rozpowszechniany za pomocą urządzeń elektronicznych, mechanicznych, kopiujących, nagrywających i innych, w tym również nie może być umieszczany ani rozpowszechniany w postaci cyfrowej zarówno w Internecie, jak i w sieciach lokalnych bez pisemnej zgody posiadacza praw autorskich.

Adres poczty elektronicznej: wnt@pol.pl
Strona WWW: www.wnt.com.pl

ISBN 83-204-2772-X

SPIS TREŚCI

	PRZEDMOWA DO WYDANIA DRUGIEGO	stronica	9
0	WSTĘPNE UWAGI O OBLICZENIACH NUMERYCZNYCH	stronica	11
0.1.	Własności zapisu zmiennopozycyjnego / 11		
0.2.	Błędy obliczeń / 13		
0.3.	Oszacowania błędów zaokrągleń / 16		
0.4.	Uwarunkowanie zadania i stabilność algorytmów / 20		
1	INTERPOLACJA	stronica	24
1.1.	Sformułowanie zagadnienia interpolacji / 24		
1.2.	Interpolacja za pomocą wielomianów / 25		
1.2.1.	Wzór interpolacyjny Lagrange'a / 26		
1.2.2.	Oszacowanie błędu wzoru interpolacyjnego / 34		
1.2.3.	Problem optymalnego doboru węzłów interpolacji / 36		
1.2.4.	Wzór interpolacyjny Newtona dla nierównych odstępów argumentu / 39		
1.2.5.	Różnice progresywne i różnice wsteczne / 45		
1.2.6.	Wzory interpolacyjne Newtona dla równoodległych wartości argumentu / 49		
1.2.7.	Zbieżność procesów interpolacyjnych / 58		
1.2.8.	Uwagi końcowe / 61		
1.3.	Interpolacja za pomocą funkcji sklejanych / 64		
1.3.1.	Określenie funkcji sklejanych / 64		
1.3.2.	Interpolacyjne funkcje sklelane stopnia trzeciego / 65		
2	APROKSYMACJA	stronica	73
2.1.	Wstęp / 73		
2.2.	Aproksymacja średniokwadratowa / 77		
2.2.1.	Aproksymacja wielomianowa / 79		
2.2.2.	Aproksymacja za pomocą wielomianów ortogonalnych / 84		

- 2.2.3. Aproksymacja trygonometryczna / 91
- 2.2.4. Szybka transformacja Fouriera / 97
- 2.2.5. Aproksymacja za pomocą funkcji sklepanych / 104
- 2.3. Aproksymacja jednostajna / 106
- 2.3.1. Metoda szeregów potęgowych / 107
- 2.3.2. Przybliżenia Padégo / 108
- 2.3.3. Szeregi Czebyszewa / 110
- 2.4. Uwagi końcowe / 113

3 PRZYBLIŻONE ROZWIĄZYWANIE RÓWNAŃ NIELINIOWYCH I ICH UKŁADÓW

stronica 115

- 3.1. Jedno równanie z jedną niewiadomą / 115
- 3.1.1. Metoda połowienia / 116
- 3.1.2. Reguła fałsi i metoda siecznych / 121
- 3.1.3. Metoda Newtona. Metody zmodyfikowane dla pierwiastków wielokrotnych / 126
- 3.2. Metody poszukiwania zer wielomianów / 133
- 3.2.1. Liczba pierwiastków rzeczywistych / 134
- 3.2.2. Lokalizacja zer rzeczywistych / 137
- 3.2.3. Metody przybliżonego obliczania zer rzeczywistych wielomianu / 139
- 3.2.4. Lokalizacja zer zespolonych / 140
- 3.2.5. Metody przybliżonego obliczania zer zespolonych wielomianu / 144
- 3.3. Uwagi o efektywności metod przybliżonego obliczania pierwiastków / 147
- 3.4. Układy równań nieliniowych / 150
- 3.4.1. Ogólne metody iteracyjne / 151
- 3.4.2. Metoda Newtona / 152
- 3.4.3. Metoda siecznych / 153
- 3.5. Poszukiwanie minimów funkcji jednej zmiennej / 155
- 3.5.1. Metody podziału / 156
- 3.5.2. Metoda optymalnych podziałów / 156
- 3.5.3. Metoda złotego podziału / 159

4 CAŁKOWANIE NUMERYCZNE

stronica 161

- 4.1. Wstęp / 161
- 4.1.1. Uwagi ogólne o całkowaniu numerycznym / 161
- 4.1.2. Ogólny wzór całkowania numerycznego / 162
- 4.2. Kwadratury z ustalonymi węzłami / 164
- 4.2.1. Kwadratury Newtona-Cotesa / 164
- 4.2.2. Kwadratury złożone Newtona-Cotesa / 168
- 4.2.3. Metoda Romberga / 171
- 4.3. Kwadratury Gaussa i kwadratury złożone Gaussa / 175
- 4.3.1. Kwadratury Gaussa / 175
- 4.3.2. Kwadratury złożone Gaussa / 185
- 4.4. Uwagi końcowe / 185

5 ROZWIĄZYWANIE UKŁADÓW ALGEBRAICZNYCH RÓWNAŃ LINIOWYCH

stronica 188

- 5.1. Wstęp / 188
- 5.2. Pojęcia podstawowe / 189

- 5.3. Metody dokładne / 192
- 5.3.1. Analiza błędów rozwiązywania / 194
- 5.3.2. Układy równań z macierzą trójkątną / 200
- 5.3.3. Metoda eliminacji Gaussa / 202
- 5.3.4. Metoda eliminacji Jordana / 216
- 5.3.5. Układy z macierzą symetryczną. Rozkłady LDL^T i LL^T / 217
- 5.3.6. Układy z macierzą trójdziagonalną / 220
- 5.3.7. Układy równań z macierzą zbliżoną do trójdziagonalnej / 223
- 5.3.8. Obliczanie wyznacznika i odwracanie macierzy / 225
- 5.3.9. Iteracyjne poprawianie rozwiązania / 230
- 5.4. Metody iteracyjne / 232
- 5.4.1. Metoda Jacobiego / 235
- 5.4.2. Metoda Gaussa-Seidla / 236
- 5.4.3. Metoda Czebyszewa / 240
- 5.4.4. Nakład obliczeń i testy stopu / 244
- 5.5. Układy równań z macierzami rzadkimi / 246
- 5.5.1. Organizacja pamięci / 247
- 5.5.2. Metody dokładne dla układów z macierzami rzadkimi / 248
- 5.5.3. Metody iteracyjne / 250
- 5.5.4. Metody blokowe / 250

6 OBLICZANIE WARTOŚCI WŁASNYCH I WEKTORÓW WŁASNYCH MACIERZY

stronica 253

- 6.1. Wstęp / 253
- 6.2. Pojęcia podstawowe / 254
- 6.3. Zaburzenia wartości i wektorów własnych / 257
- 6.4. Macierze o elementach rzeczywistych — metody ogólne / 265
- 6.4.1. Lokalizacja wartości własnych / 265
- 6.4.2. Znajdowanie wartości własnych przy użyciu wielomianu charakterystycznego / 267
- 6.4.3. Metoda potęgowa / 268
- 6.4.4. Algorytm QR dla macierzy Hessenberga / 270
- 6.4.5. Sprowadzanie macierzy do postaci Hessenberga / 273
- 6.4.6. Obliczanie wektorów własnych / 274
- 6.4.7. Algorytm rozkładu macierzy na iloczyn QR / 277
- 6.5. Macierze symetryczne / 283
- 6.5.1. Macierz trójdziagonalna symetryczna / 285
- 6.5.2. Sprowadzanie macierzy symetrycznej do postaci trójdziagonalnej / 291
- 6.6. Macierze wstęgowe / 293

7 METODY ROZWIĄZYWANIA ZAGADNIEŃ POCZĄTKOWYCH DLA RÓWNAŃ RÓŻNICZKOWYCH ZWYCZAJNYCH

stronica 295

- 7.1. Wstęp / 295
- 7.2. Metoda Eulera / 297
- 7.3. Metody różnicowe / 299
- 7.3.1. Ogólny wzór różnicowy / 299
- 7.3.2. Równanie dla błędu / 301
- 7.3.3. Stabilność i zbieżność / 306
- 7.3.4. Wyznaczanie praktyczne przydatnych wzorów różnicowych / 316
- 7.4. Metody typu Rungego-Kutty / 318

- 7.4.1. Wzór ogólny / 318
- 7.4.2. Stabilność metod Rungego-Kutty / 321
- 7.4.3. Metody rzędu czwartego / 323
- 7.4.4. Wybór kroku całkowania / 323
- 7.5. Metody ekstrapolacyjno-interpolacyjne / 325
- 7.5.1. Metody Hamminga / 327
- 7.5.2. Metoda Geara dla układów typu stiff / 331

8

METODY ROZWIĄZYWANIA ZAGADNIEŃ BRZEGOWYCH DLA RÓWNAŃ RÓŻNICZKOWYCH CZĄSTKOWYCH

stronica 336

- 8.1. Metoda różnicowa dla równania przewodnictwa cieplnego / 336
 - 8.1.1. Postawienie zagadnienia / 336
 - 8.1.2. Aproksymacja różnicowa / 336
 - 8.1.3. Stabilność i zbieżność / 340
 - 8.1.4. Przykład obliczeniowy / 351
 - 8.1.5. Schemat blokowy / 353
- 8.2. Metoda różnicowa dla równania drgań struny / 355
 - 8.2.1. Postawienie zagadnienia / 355
 - 8.2.2. Aproksymacja różnicowa / 355
- 8.3. Metoda różnicowa dla równania Poissona / 357
 - 8.3.1. Postawienie zagadnienia / 357
 - 8.3.2. Aproksymacje różnicowe równania Poissona / 357
 - 8.3.3. Aproksymacja różnicowa zagadnienia Dirichleta / 359
 - 8.3.4. Jednostajna zbieżność rozwiązania różnicowego / 364
- 8.4. Metoda prostych dla równania przewodnictwa cieplnego / 367
- 8.5. Metoda prostych dla równania drgań struny / 369
- Dodatek. Metoda elementu skończonego / 371
 - D.1. Pojęcia wstępne / 371
 - D.2. Metoda elementu skończonego / 373

LITERATURA / 378

SKOROWIDZ / 381

PRZEDMOWA DO WYDANIA DRUGIEGO

Celem autorów niniejszej książki jest wprowadzenie Czytelnika w dział matematyki stosowanej, zwanej metodami numerycznymi. Gwałtowny rozwój techniki obliczeniowej spowodował pojawienie się istnej lawiny metod, umożliwiających rozwiązywanie różnorodnych problemów z dziedziny techniki, medycyny, ekonomii itp. przy użyciu komputerów. Należy zdawać sobie sprawę z tego, że w zasadzie każde bardziej złożone zagadnienie wymaga opracowania indywidualnej metody jego rozwiązywania na komputerze. Niekiedy jednak różnica użytej metody w stosunku do metody standardowej polega na drobnych pozornie zmianach, takich jak nietypowe operowanie pamięcią operacyjną, zastosowanie przy pewnych pomocniczych obliczeniach podwójnej precyzji, zastosowanie innych testów zatrzymań. Zmiany takie mogą przesądzić na przykład o możliwości rozwiązania danego zagadnienia za pomocą aktualnie dostępnego komputera. Bywają jednak zagadnienia, do których rozwiązania należy opracować istotnie nową metodę. Powyższe względy powodują, iż stosowanie metod numerycznych jest po trosze sztuką, wymagającą dużej wprawy. Autorzy tej książki postawili sobie za cel przedstawienie wybranych działów metod numerycznych w ten sposób, by z jednej strony umożliwić Czytelnikowi rozwiązywanie typowych zagadnień — rozwiązywanie równań różniczkowych, całkowanie, rozwiązywanie układów równań algebraicznych — z drugiej zaś strony przez dokładniejsze omówienie działu algebry liniowej umożliwić samodzielne opracowywanie bardziej wyspecjalizowanych metod. Zakładamy, że Czytelnik jest obeznany z podstawowymi działami matematyki, jak algebra, teoria funkcji, rachunek różniczkowy i całkowy, w zakresie podanym w pracach [61], [72]—[74]. Ponadto do właściwego zrozumienia istoty większości algorytmów niezbędna jest umiejętność programowania w języku Pascal (lub Fortran). W obecnym, drugim wydaniu,

postaraliśmy się usunąć wszystkie dostrzeżone lub zasygnalizowane przez Czytelników błędy i niedokładności.

Ponieważ od pierwszego wydania minęło ponad 10 lat, a w okresie tym nastąpiła prawdziwa rewolucja w sprzęcie komputerowym oraz dostępnym oprogramowaniu, w obecnym wydaniu większość przykładowych programów została przerobiona na język Pascal.

W okresie tym ukazało się także na rynku wydawniczym kilka nowych, cennych pozycji mających charakter podręczników akademickich i książki te zamieściliśmy jako dodatkowe pozycje literatury.

Warszawa, w czerwcu 1992

*Zenon Fortuna
Bohdan Macukow
Janusz Wąsowski*

0

WSTĘPNE UWAGI O OBLICZENIACH NUMERYCZNYCH

0.1. WŁASNOŚCI ZAPISU ZMIENNOPOZYCYJNEGO

Aby posługiwać się maszyną cyfrową i rozwiązywać za jej pomocą nawet bardzo złożone problemy, niekoniecznie trzeba znać jej budowę i sposób reprezentowania liczb i znaków w tej maszynie. Wielu użytkownikom wystarczy znajomość jakiegoś języka programowania i sposób wprowadzania danych oraz wyprowadzania wyników. Pisząc jednak program, korzystając z gotowych algorytmów (procedur), zwłaszcza układając własne algorytmy obliczeniowe, należy znać pewne podstawowe dane techniczne używanej maszyny cyfrowej, mianowicie: wielkość pamięci operacyjnej maszyny, szybkość wykonywania operacji arytmetycznych i logicznych, zakres liczb dopuszczalnych podczas obliczeń oraz dokładność wykonywania podstawowych działań arytmetycznych na liczbach rzeczywistych.

Liczby rzeczywiste x są zapamiętywane w większości współczesnych maszyn cyfrowych w tzw. postaci zmiennopozycyjnej $x = M \times N^W$, gdzie M oznacza mantysę liczby x , a W — wykładnik części potęgowej. Podstawa N części potęgowej jest na ogół równa 2 lub 10, lecz np. w maszynie IBM 360 przyjęto $N = 16$. W zapisie zmiennopozycyjnym liczba rzeczywista jest więc przedstawiana za pomocą dwu grup bitów. Pierwsza grupa, tworząca mantysę M , jest interpretowana jako część ułamkowa; w większości maszyn cyfrowych jest spełniony warunek: $1/2 \leq M < 1$. Druga grupa, tworząca wykładnik W , jest interpretowana jako liczba całkowita, przy czym zakres liczb W decyduje o zakresie liczb rzeczywistych dopuszczalnych w danej maszynie.

Przykład. Jeżeli w zapisie dwójkowym liczbę M określa 5 bitów, a liczbę W trzy bity, przy czym pierwsze bity oznaczają znaki tych liczb (liczbę ujemną rozpoczyna bit o wartości 1), to zapis

$$x = \underset{M}{(1)1101} \underset{W}{(0)10}$$

oznacza w zapisie dwójkowym liczbę

$$-0,1101 \times 2^{+10} = -\left(\frac{1}{2} + \frac{1}{4} + \frac{0}{8} + \frac{1}{16}\right) \times 2^{+(1 \cdot 2 + 0 \cdot 1)}$$

czyli w zapisie dziesiętnym liczbę $-3,25$.
(10)

W zapisie zastosowanym w powyższym przykładzie można utworzyć tylko niektóre liczby dodatnie w zakresie od $0,0625$ do $7,5$, liczbę 0 i niektóre
(10) (10)

liczby ujemne w zakresie od $-0,0625$ do $-7,5$. Łatwo spostrzec, że są liczby
(10) (10)

rzeczywiste, których za pomocą tego zapisu nie można dokładnie przedstawić. Na przykład liczba $x = 0,2$ (w zapisie dziesiętnym) ma w zapisie dwójkowym
(10)

nieskończone rozwinięcie równe $0,0011(0011)\dots$. Najbliższa jej liczba, którą
(2)

można przedstawić stosując 5-bitową mantysę i 3-bitowy wykładnik, to liczba $x' = (0)1100 (1)10 \equiv 0,001100 = 0,1875$, a więc *błąd bezwzględny* $|x - x'|$
(2) (10)

wynosi $0,0125$. Podobnie liczbę $-6,25$ można najlepiej przybliżyć liczbą
(10) (10)

$(1)1101 (0)11 \equiv -110,1 = -6,5$, z błędem $0,25$. Widać więc, że błąd ten
(2) (10) (10)

silnie zależy od wielkości rozpatrywanej liczby. Dlatego wygodniej jest posługiwać się *błędem względnym*, określonym jako taka liczba ε' , że $x' = x(1 + \varepsilon')$. Dla liczby $0,2$ uzyskamy wówczas $\varepsilon' = -0,0625$, a dla liczby $-6,25$ uzyskamy $\varepsilon' = 0,04$.

Można uzasadnić, że $|\varepsilon'| \leq \varepsilon$ dla dowolnej liczby x mieszczącej się w zakresie liczb objętych zapisem, gdzie ε jest pewną ustaloną liczbą (niezależną od liczby x), tym mniejszą, im więcej bitów przeznaczono na reprezentowanie mantysy M . W rozpatrywanym poprzednio przykładzie można przyjąć $\varepsilon = 2^{-3} = 0,125$.

W dalszej części książki, zwłaszcza w rozdziałach 4 i 6, liczbę ε będziemy traktować jako parametr charakteryzujący dokładność obliczeniową maszyny (im mniejsza liczba ε , tym większa dokładność). We wszystkich prawie maszynach cyfrowych istnieje możliwość zwiększenia dokładności obliczeń. W tych miejscach książki, gdzie mówimy o zwiększonej dokładności odwołujemy się do sytuacji, gdy w przypadku programów napisanych w Turbo Pascalu konfiguracja sprzętu zawiera koprocessor arytmetyczny — czyli dostępne są rozszerzone typy rzeczywiste, a w przypadku programów napisanych w Fortranie — dostępny jest tryb obliczeń prowadzonych w tzw. *podwójnej precyzji* (DP — *double precision*). Dla mantysy i ewentualnie wykładnika przewidziana jest wówczas np. dwukrotnie większa liczba bitów. Powoduje to zmniejszenie parametru ε zgodnie z zależnością $\varepsilon_{DP} = \varepsilon^2$.

0.2. BŁĘDY OBLICZEŃ

Przy obliczeniach wykonywanych na maszynach cyfrowych występują trzy podstawowe rodzaje błędów: błędy wejściowe, błędy obcięcia i błędy zaokrągleń.

Błędy wejściowe (względnie *błędy danych wejściowych*) występują wówczas, gdy dane liczbowe wprowadzone do pamięci lub rejestrów maszyny cyfrowej odbiegają od dokładnych wartości tych danych. Błędy takie pojawiają się np. wtedy, gdy dane wejściowe są wynikiem pomiarów wielkości fizycznych, mierzonych z pewnymi błędami pomiaru. Bardziej jednak charakterystyczną przyczyną występowania błędów wejściowych jest skończona długość słów binarnych, reprezentujących liczby w maszynie, i nieuniknione w związku z tym wstępne zaokrąglanie. W poprzednim punkcie podano przykład zapisu zmiennopozycyjnego, w którym liczby 0,2 i $-6,25$ można było zapisać jedynie z nieuniknionym błędem względnym (nie przekraczającym liczby ε). Zauważmy na koniec, że wstępne zaokrąglanie występuje w przypadku wszystkich liczb niewymiernych, takich jak $\sqrt{2}$, π , e itp.

Błędy obcięcia powstają podczas obliczeń na skutek zmniejszania liczby działań. Postępujemy tak na ogół przy obliczaniu sum nieskończonych (szeregów). Chcąc np. obliczyć wartość wyrażenia e^x , równego szeregowi

$$1 + x + \frac{1}{2}x^2 + \dots + \frac{1}{N!}x^N + \dots$$

zastępujemy je sumą częściową $1 + x + \frac{1}{2}x^2 + \dots + \frac{1}{N!}x^N$ o odpowiednio dobranej wartości N . Zauważmy, że jeśli

N jest dostatecznie dużą liczbą, to uzyskana w ten sposób wartość liczby e^x może być równa wstępnie zaokrąglonej jej wartości, tzn. całkowity błąd popełniony przy obliczeniu e^x będzie równy błędowi wejściowemu. Na ogół jednak ze względu na długi czas obliczania tego rodzaju sum uwzględnia się niewielką liczbę składników szeregu, co powoduje pojawienie się dodatkowego błędu — błędu obcięcia. Błędy obcięcia powstają również często przy obliczaniu wielkości będących granicami. Jeżeli np. obliczamy wartość całki oznaczonej jako granicę sum przybliżających całkę o coraz to gęstszym podziale, to przyjęcie jednej z tych sum jako wartości bliskiej całce oznacza wprowadzenie błędu obcięcia. Podobnie zastąpienie wartości pochodnej funkcji jej ilorazem różnicowym powoduje powstanie błędu obcięcia.

W wielu przypadkach można uniknąć błędów wejściowych i błędów obcięcia, ograniczając np. dane do takich wartości, które możemy wprowadzać do rejestrów maszyny bez wstępnych zaokrągleń, lub np. zastępując obliczanie całki oznaczonej za pomocą sum obliczaniem jej wartości ze wzoru wykorzystującego całkę nieoznaczoną. W trakcie obliczeń pojawiają się jednak nowe błędy, zwane *błędami zaokrągleń*, których na ogół nie da się uniknąć, choć można je zmniejszyć ustalając umiejętnie sposób i kolejność wykonywania działań.

Przykład. Rozważmy taki sam jak w poprzednim przykładzie zapis zmiennopozycyjny z 5 bitami (wraz z bitem znaku) dla mantysy i 3 bitami dla wykładnika. Jeżeli mamy obliczyć sumę trzech liczb: 0,48, 0,24 i 0,12, to w rozważanym zapisie trzeba wstępnie zaokrąglić te liczby do wartości (0)1111 (1)01, (0)1111 (1)10, (0)1111 (1)11, czyli do liczb 0,46875, 0,234375, 0,1171875. Taki sposób zaokrąglania, polegający na odrzuceniu cyfr rozwinięcia binarnego wykraczających poza przyjętą liczbę bitów, nazywamy *urywaniem (chopping)*. Większość kompilatorów Fortranu właśnie w taki sposób dokonuje zaokrąglania liczb. Zamiast liczby x uzyskujemy wtedy liczbę $x' = x(1 + \varepsilon')$, przy czym błąd względny $\varepsilon' = \frac{x' - x}{x}$ jest nie większy co do wartości bezwzględnej od $\varepsilon = 2^{-t+1}$, gdzie t oznacza liczbę bitów przyjętych do reprezentowania w zapisie dwójkowym mantysy M . Istnieją jeszcze inne sposoby zaokrąglania, zbliżone np. do zaokrąglania powszechnie przyjętego w zapisie dziesiętnym ($0,249 \approx 0,2$, $0,250 \approx 0,3$, $0,251 \approx 0,3$ itp.). Przy takim sposobie zaokrąglania parametr ε jest dwukrotnie mniejszy (w zapisie dwójkowym), lecz wiąże się to z większą rozbudową rejestrów maszyny cyfrowej.

Obliczmy sumę liczb będących zaokrągleniem przez urywanie liczb 0,48, 0,24 i 0,12. Należy w tym celu zmienić wykładniki sumowanych składników w ten sposób, by były jednakowe. Sumując dwie pierwsze liczby mamy $(0)1111 (1)01 + (0)1111 (1)10 = (0)1111 (1)01 + (0)01111 (1)01$, przy czym liczbę $(0)01111 (1)01$ należy zaokrąglić do $(0)0111 (1)01$. Ostatnia suma nie może być zapisana w rejestrze, który przechowuje tylko 4 bity. Jednym z możliwych rozwiązań jest dodatkowa zmiana wykładników. Aby podkreślić, że działania są wykonywane w *technice zmiennopozycyjnej (floating point)* i wyniki na skutek zaokrągleń mogą różnić się od wyników dokładnych, obliczoną przy użyciu maszyny wartość a będziemy oznaczać symbolem $\text{fl}(a)$. Zatem

$$\text{fl}((0)1111 (1)01 + (0)0111 (1)01) = \text{fl}((0)0111 (1)00 + (0)0011 (1)00) = (0)1010 (1)00.$$

Podobnie

$$\text{fl}((0)1010 (1)00 + (0)1111 (1)11) = \text{fl}((0)1010 (1)00 + (0)0001 (1)00) = (0)1011 (1)00 = 0,1011 = 0,68125$$

Uzyskano wynik z błędem względnym około 20%.

Obliczmy powyższą sumę jeszcze raz, zmieniając kolejność składników:

$$\text{fl}((0)1111 (1)11 + (0)1111 (1)10) = \text{fl}((0)0111 (1)10 + (0)1111 (1)10) = \text{fl}((0)0011 (1)01 + (0)0111 (1)01) = (0)1010 (1)01$$

$$\text{fl}((0)1010 (1)01 + (0)1111 (1)01) = \text{fl}((0)0101 (1)00 + (0)0111 (1)00) = (0)1100 (1)00 = 0,1100 = 0,75$$

Wynik, który teraz uzyskaliśmy, jest dokładniejszy — błąd względny wynosi bowiem około 10%. W następnym punkcie podamy wyjaśnienie przyczyny zmniejszenia się powyższego błędu.

Na zakończenie ogólnych uwag o błędach zaokrągleń, podamy oszacowanie błędów względnych powstających podczas wykonywania pojedynczych działań arytmetycznych.

Lemat Wilkinsona. *Błędy zaokrągleń powstające podczas wykonywania działań zmiennopozycyjnych są równoważne zastępczemu zaburzeniu liczb, na których wykonujemy działania. W przypadku pojedynczych działań arytmetycznych otrzymujemy:*

$$\text{fl}(x_1 \pm x_2) = x_1(1 + \varepsilon_1) \pm x_2(1 + \varepsilon_2)$$

$$\text{fl}(x_1 x_2) = x_1(1 + \varepsilon_3) \cdot x_2 = x_1 \cdot x_2(1 + \varepsilon_3)$$

$$\text{fl}(x_1/x_2) = x_1(1 + \varepsilon_4)/x_2 = x_1/(x_2(1 + \varepsilon_5))$$

gdzie liczby ε_i , $i = 1, \dots, 5$, są niewiększe co do modułu od parametru ε , charakteryzującego dokładność maszyny.

U w a g a. Lemat powyższy stanowi podsumowanie szczegółowej analizy błędów zaokrągleń (patrz [68], s. 16—21).

Jako ilustrację tego lematu rozważmy ponownie obliczanie sumy liczb (0)1111 (1)01 i (0)1111 (1)10. Dokładny wynik wynosi 0,101101, natomiast $\text{fl}((0)1111 (1)01 + (0)1111 (1)10) = 0,1010$. Zamiast interpretować powstały błąd jako wynik zaokrąglenia (urwania), które wystąpiło podczas przekalowywania (zmiany wykładników) składników, można przyjąć w myśl lematu Wilkinsona, że samo działanie zostało wykonane dokładnie (bez błędu), ale na nieco zmienionych składnikach. Jeśli bowiem zamiast $x_1 = 0,46875$ i $x_2 = 0,234375$ przyjmiemy liczby $x'_1 = x_1(1 + \varepsilon_1)$, $x'_2 = x_2(1 + \varepsilon_2)$, gdzie $\varepsilon_1 = \varepsilon_2 = -\frac{1}{9}$, to $x'_1 + x'_2 = 0,1010$. Oczywiście $|\varepsilon_1| = |\varepsilon_2| < \varepsilon = 2^{-4+1} = \frac{1}{8}$.

Zauważmy, że wybór zaburzeń ε_1 i ε_2 nie jest na ogół jednoznaczny. W rozważanym przypadku sumowania można przyjąć np.

$$\varepsilon_1 = -\frac{1}{8}, \quad \varepsilon_2 = -\frac{1}{12} \quad \text{lub} \quad \varepsilon_1 = -\frac{5}{48} \cdots \left(|\varepsilon_1| < \frac{1}{8} \right), \quad \varepsilon_2 = -\frac{1}{8}.$$

Ta niejednoznaczność zastępczych zaburzeń jest często wygodna przy szacowaniu wielkości tych zaburzeń (skorzystamy z tego np. w rozdz. 5 dla oszacowania błędów powstających przy rozwiązywaniu układu równań liniowych).

Należy jeszcze zwrócić uwagę na zjawisko, zwane *redukcją cyfr znaczących*, występujące przy odejmowaniu liczb bliskich co do wartości bezwzględnej.

Przykład. Obliczmy $r = x_1 - x_2$, gdzie $x_1 = 0,50000$, $x_2 = 0,46875$. Wykonując obliczenia w opisany powyżej sposób, otrzymamy

$$\begin{aligned} \text{fl}(x_1 - x_2) &= \text{fl}((0)1000 (1)00 - (0)1111 (1)01) = \\ &= \text{fl}((0)1000 (1)00 - (0)0111 (1)00) = \\ &= (0)0001 (1)00 = (0)1000 (1)11 = 1/16 \end{aligned}$$

natomiast dokładna wartość różnicy r wynosi $1/32$. Błąd względny spowodowany skończoną liczbą bitów jest zatem równy 100%.

Uzyskany wynik pozornie nie jest zgodny z lematem Wilkinsona, z którego wynika, że błąd bezwzględny popełniony przy odejmowaniu jest mały (tu błąd bezwzględny wynosi $1/32$). Przy odejmowaniu bliskich sobie liczb mała jest jednak także dokładnie obliczona różnica, co powoduje możliwość wystąpienia dużego błędu względnego, o czym należy pamiętać obliczając np. ilorazy różnicowe przybliżające pochodne funkcji, wyznaczając

pierwiastki równania kwadratowego o dominującym współczynnikiem przy niewiadomej w pierwszej potęgze itp. Najprostszym środkiem zaradczym jest wówczas zwiększenie dokładności obliczeń (np. DP) — wtedy różnica obliczana jest bez błędu. Z drugiej strony niektóre maszyny cyfrowe są wyposażone w akumulator wykonujący standardowo wszystkie działania na słowach o podwójnej długości, co również zapobiega redukcji cyfr znaczących (patrz [68, s. 19]).

0.3. OSZACOWANIA BŁĘDÓW ZAOKRĄGLEŃ

Sumowanie liczb. Rozważmy sumowanie n liczb rzeczywistych: $x(1), x(2), \dots, x(n)$. Na maszynie sumę $s = x_1 + x_2 + \dots + x_n$ wyznaczamy stosując na ogół pętlę, na przykład:

```
.....
s := 0.0;
FOR i := 1 TO n DO
s := s + x(i);
.....
```

U w a g a: W wielu miejscach książki przy podawaniu prostych podprogramów, czy fragmentów programów napisanych w języku Turbo Pascal zakładamy, że w programie (lub podprogramie) zostały podane odpowiednie definicje typów i deklaracje zmiennych.

Oznaczmy przez s_k sumę k liczb $x(1), \dots, x(k)$. Wówczas $s_1 = x(1)$, $s_k = \text{fl}(s_{k-1} + x(k)) = s'_{k-1} + x'(k)$, gdzie liczby s'_{k-1} i $x'(k)$ różnią się od wartości s_{k-1} i $x(k)$ mnożnikami bliskimi 1, zgodnie z lematem Wilkinsona. Niech $s'_{k-1} = s_{k-1}(1 + \varepsilon_{k-1}^1)$, $x'(k) = x(k)(1 + \varepsilon_k^2)$, gdzie $|\varepsilon_{k-1}^1| \leq \varepsilon$ i $|\varepsilon_k^2| \leq \varepsilon$. Ponieważ wyznaczona ostatecznie wartość s_n jest poszukiwaną sumą s , więc

$$\begin{aligned} s = & x(1)(1 + \varepsilon_1^2)(1 + \varepsilon_2^1) \dots (1 + \varepsilon_{n-1}^1) + \\ & + x(2)(1 + \varepsilon_2^2)(1 + \varepsilon_3^1) \dots (1 + \varepsilon_{n-1}^1) + \\ & + x(3)(1 + \varepsilon_3^2)(1 + \varepsilon_4^1) \dots (1 + \varepsilon_{n-1}^1) + \dots + \\ & + x(n-1)(1 + \varepsilon_{n-1}^2)(1 + \varepsilon_n^1) + x(n)(1 + \varepsilon_n^2) \end{aligned}$$

Uzyskany wynik można interpretować w ten sposób, że wyznaczona suma s jest dokładną sumą zaburzonych składników, przy czym wielkość zaburzeń zależy od kolejności wykonywania sumowania. Nie można określić rzeczywistej wielkości mnożnika $(1 + \varepsilon_1^1)(1 + \varepsilon_2^1) \dots (1 + \varepsilon_{n-1}^1)$, którym należy zaburzyć wartość $x(1)$, bez podania pozostałych liczb $x(2), \dots, x(n)$, nie można też określić pozostałych mnożników bez uwzględniania wszystkich sumowanych liczb. Można jednak oszacować maksymalne możliwe zmiany składników, mianowicie:

$x(1)$ może zostać zmniejszony co najwyżej $(1 - \varepsilon)^n$ razy lub zwiększony co najwyżej $(1 + \varepsilon)^n$ razy;

takie samo oszacowanie uzyskujemy dla $x(2)$;

pozostałe liczby $x(i)$, $i = 3, \dots, n$, mogą być zmienione co najwyżej przez mnożniki $(1 - \varepsilon)^{n+1-i}$ lub $(1 + \varepsilon)^{n+1-i}$.

Najmniej zaburzony może być zatem ostatni składnik, bo tylko $1 - \varepsilon$ lub $1 + \varepsilon$ razy. Powyższe oszacowania stanowią podstawę reguły, że liczby należy sumować w kolejności wzrastających wartości bezwzględnych. Przykład podany w poprzednim punkcie potwierdza słuszność tej reguły, lecz należy zdawać sobie sprawę z tego, że w pewnych przypadkach sumowanie w innej kolejności może dać dokładniejszy wynik (gorsze jednak oszacowanie).

Obliczanie wartości wielomianu. Przy użyciu maszyn cyfrowych wartość wielomianu $w(x) = x^n + a_1 x^{n-1} + \dots + a_{n-1} x + a_n$ oblicza się stosując na ogół tzw. *schemat Hornera*:

$$\begin{aligned} w_1 &= x + a_1 \\ w_2 &= xw_1 + a_2 \\ &\dots\dots\dots \\ w_n &= xw_{n-1} + a_n \\ w(x) &= w_n \end{aligned}$$

odpowiadający obliczaniu wyrażenia

$$x \{ x [x \dots (x + a_1) + a_2] + \dots + a_{n-1} \} + a_n$$

Łatwo stwierdzić, że obliczając w ten sposób wartość wielomianu należy wykonać $M = n - 1$ mnożeń i $D = n$ dodawań. Obliczając natomiast tę wartość jako

$$\underbrace{xx \dots x}_n + a_1 \underbrace{x \dots x}_{n-1} + \dots + a_{n-1} x + a_n$$

należy wykonać łącznie $M = \frac{(n-1)(n+2)}{2}$ mnożeń i $D = n$ dodawań.

Schemat Hornera umożliwia znaczne zmniejszenie liczby działań arytmetycznych. Powstaje pytanie, która z powyższych metod daje lepsze oszacowanie błędów zaokrągleń. Otóż okazuje się, że jeśli wykonywać działania w kolejności przedstawionej powyżej, to oszacowanie wielkości błędów zaokrągleń jest identyczne dla obu metod.

Przykład. Dla wielomianu $x^3 + a_1 x^2 + a_2 x + a_3$ przy użyciu schematu Hornera uzyskujemy wartość

$$\begin{aligned} fl(w(x)) &= x(1 + \varepsilon_1) \{ x(1 + \varepsilon_2) [x(1 + \varepsilon_3) + a_1(1 + \varepsilon_4)] (1 + \varepsilon_5) + \\ &\quad + a_2(1 + \varepsilon_6) \} (1 + \varepsilon_7) + a_3(1 + \varepsilon_8) = \\ &= x^3(1 + \varepsilon_1)(1 + \varepsilon_2)(1 + \varepsilon_3)(1 + \varepsilon_5)(1 + \varepsilon_7) + \\ &\quad + a_1 x^2(1 + \varepsilon_1)(1 + \varepsilon_2)(1 + \varepsilon_4)(1 + \varepsilon_5)(1 + \varepsilon_7) + \\ &\quad + a_2 x(1 + \varepsilon_1)(1 + \varepsilon_6)(1 + \varepsilon_7) + a_3(1 + \varepsilon_8) \end{aligned}$$

Gdy zaś obliczymy wartość tego wielomianu jako $x \cdot x \cdot x + a_1 \cdot x \cdot x + a_2 \cdot x + a_3$, wówczas uzyskamy

$$\begin{aligned} \text{fl}(w(x)) = & x^3(1 + \varepsilon'_1)(1 + \varepsilon'_2)(1 + \varepsilon'_3)(1 + \varepsilon'_4)(1 + \varepsilon'_5) + \\ & + a_1 x^2(1 + \varepsilon'_6)(1 + \varepsilon'_7)(1 + \varepsilon'_8)(1 + \varepsilon'_4)(1 + \varepsilon'_5) + \\ & + a_2 x(1 + \varepsilon'_9)(1 + \varepsilon'_{10})(1 + \varepsilon'_5) + a_3(1 + \varepsilon'_{11}) \end{aligned}$$

Wynik obliczania $w(x)$ może być w obu przypadkach różny, lecz oszacowanie największej możliwej wartości błędu jest takie samo.

Błędy zaokrągleń w obliczeniach iteracyjnych. Wiele efektywnych algorytmów obliczeniowych polega na wyznaczeniu ciągu liczb zbieżnego do poszukiwanej wartości. Zagadnienie to rozpatrzmy na przykładzie wyznaczania wartości $y = \sqrt{x}$. Można to uczynić obliczając ciąg $y_0, y_1, y_2, \dots$, gdzie:

y_0 — dowolnie wybrana liczba dodatnia

$$y_i = \frac{1}{2} \left(y_{i-1} + \frac{x}{y_{i-1}} \right), \quad i = 1, 2, 3, \dots$$

Jeśli założyć, że działania są wykonywane dokładnie, to ciąg $y_0, y_1, y_2, \dots$

będzie zbieżny do liczby x . Jeśli bowiem $y_i = p_i \sqrt{x}$, to $y_{i+1} = \frac{1}{2} \left(p_i + \frac{1}{p_i} \right) \sqrt{x}$,

a ciąg $p_0 > 0, p_{i+1} = \frac{1}{2} \left(p_i + \frac{1}{p_i} \right), i = 0, 1, \dots$, jest zbieżny do 1.

Zespół działań potrzebnych do wyznaczania kolejnych wartości elementów ciągu $y_0, y_1, \dots$ nazywamy *iteracją*. O tym, czy w kolejnej iteracji nastąpiło przybliżenie do rozwiązania, decyduje wartość ułamka

$$q_i = \frac{y_{i+1} - \sqrt{x}}{y_i - \sqrt{x}}$$

W przypadku dokładnych działań, gdy $y_i = p_i \sqrt{x}$, mamy

$$q_i = \frac{p_i - 1}{2p_i}, \text{ więc } |q_i| < 1 \text{ dla } i = 0, 1, 2, \dots, \text{ jeśli tylko } p_i > \frac{1}{3}. \text{ Oznacza to}$$

malenie wartości bezwzględnej różnicy $y_i - \sqrt{x}$ począwszy od dostatecznie dużego numeru iteracji i . Gdy obliczenia wykonujemy na maszynie cyfrowej, wówczas

$$\text{fl}(y_{i+1}) = \frac{1}{2} \left[y_i(1 + \varepsilon_1^i) + \frac{x(1 + \varepsilon_2^i)(1 + \varepsilon_3^i)}{y_i} \right]$$

gdzie $|\varepsilon_k^i| \leq \varepsilon, k = 1, 2, 3$.

Zwróćmy uwagę, że mnożnik $1/2$ nie powoduje błędów zaokrągleń (zakładamy, że maszyna wykonuje działania w systemie dwójkowym). Dla liczby $y'_{i+1} = \text{fl}(y_{i+1})$ uzyskamy

$$q'_i = \frac{y'_{i+1} - \sqrt{x}}{y_i - \sqrt{x}} = \frac{p_i - 1}{2p_i} + \frac{p_i \varepsilon_1^i + \frac{\varepsilon_2^i + \varepsilon_3^i + \varepsilon_2^i \varepsilon_3^i}{p_i}}{2(p_i - 1)}$$

Pierwszy składnik tej sumy może mieć pomijalnie małą wartość, jeśli p_i jest dostatecznie bliskie 1, czyli jeśli y_i jest już dostatecznie dobrym przybliżeniem liczby $\sqrt{x}$. Drugi składnik jest spowodowany błędami zaokrągleń i może mieć tym większą wartość, im bliższe 1 jest p_i , a więc właśnie wtedy, gdy y_i dobrze przybliża $\sqrt{x}$. Jeśli na przykład $|p_i - 1| < \frac{3}{2}\varepsilon$, a liczby $\varepsilon_1^i, \varepsilon_2^i, \varepsilon_3^i$ są dodatnie i bliskie ε , to wartość q'_i przekroczy 1. Oznacza to, że na skutek błędów zaokrągleń nie uzyskamy lepszego przybliżenia rozwiązania w $(i+1)$ -ej iteracji niż w i -tej iteracji. W niektórych jednak przypadkach błędy zaokrągleń mogą pomóc w wyznaczeniu dokładnego rozwiązania. Jeśli bowiem wystąpi taki układ wartości liczb $p_i, \varepsilon_1^i, \varepsilon_2^i, \varepsilon_3^i$, że $q'_i = 0$, to oznacza to, że przypadkowo uzyskaliśmy $y_{i+1} = \sqrt{x}$, jakkolwiek teoretycznie żaden element ciągu $y_0, y_1, \dots$ nie może być równy $\sqrt{x}$ (jeśli $y_0 \neq \sqrt{x}$). Sytuacja taka zdarza się jednak w praktyce niezmiernie rzadko, gdyż niewielka zmiana początkowej wartości y_0 powoduje zmianę wartości y'_{i+1} do wartości różnej od $\sqrt{x}$ i zmiana ta jest na ogół duża. Na skutek występowania błędów zaokrągleń ciąg generowany metodą iteracyjną nie dąży do rozwiązania, lecz zaczyna oscylować wokół rozwiązania. Używa się w związku z tym pojęcia *maksymalnej granicznej dokładności*: jest to oszacowanie błędu rozwiązania, którego nie można zmniejszyć dla danej metody iteracyjnej.

W rozważanym przykładzie oszacowanie to uzyskamy obliczając mnożnik p_i , dla którego błąd w kolejnej iteracji nie zmniejszy się, czyli $|q_i| \geq 1$. Mamy

$$|q_i| = \left| \frac{p_i - 1}{2p_i} + \frac{p_i \varepsilon_1^i + \frac{\varepsilon_2^i + \varepsilon_3^i + \varepsilon_2^i \varepsilon_3^i}{p_i}}{2(p_i - 1)} \right| \approx \left| \frac{\varepsilon_1^i + \varepsilon_2^i + \varepsilon_3^i}{2(p_i - 1)} \right|$$

przy założeniu, że $|\varepsilon_k^i| \ll 1$ oraz $p_i \approx 1$. Najgorsze oszacowanie otrzymamy, gdy liczby $\varepsilon_k^i, k = 1, 2, 3$, są jednakowego znaku i równe co do wartości bezwzględnej liczbie ε . Zatem

$$\frac{3\varepsilon}{2|p_i - 1|} \geq 1$$

$$|p_i - 1| \leq \frac{3}{2}\varepsilon$$

$$1 - \frac{3}{2}\varepsilon \leq p_i \leq 1 + \frac{3}{2}\varepsilon$$

Oznacza to, że nie możemy zagwarantować lepszego przybliżenia $y_i = p_i \sqrt{x}$ liczby $\sqrt{x}$ niż z błędem $\frac{3}{2}\sqrt{x}\varepsilon$. Jest to właśnie maksymalna graniczna dokładność wyznaczenia liczby $\sqrt{x}$ rozważaną metodą iteracyjną.

0.4. UWARUNKOWANIE ZADANIA I STABILNOŚĆ ALGORYTMÓW

Zadanie numeryczne jest problemem polegającym na wyznaczeniu *wektora wyników* w na podstawie *wektora danych* a . Mówimy, że zadanie jest *dobrze postawione*, jeżeli wektor w jest jednoznacznie określony dla przyjętego wektora danych a . Niech D oznacza zbiór danych a , dla których zadanie jest dobrze postawione. Na zbiorze D istnieje zatem odwzorowanie W takie, że $w = W(a)$. Niech $\bar{w}$ oznacza obliczony numerycznie wektor wyników. W celu obliczenia wektora $\bar{w}$, należy sformułować *algorytm obliczeniowy* polegający na określeniu ciągu działań, które trzeba wykonać nad wektorem danych a i wynikami poprzednich działań, przy czym algorytm ten będzie *poprawnie sformułowany* wtedy, gdy liczba niezbędnych działań będzie skończona (choć może zależeć od wektora danych a).

Algorytm określa odwzorowanie WN takie, że $w = WN(a, \varepsilon)$. Parametr ε symbolizuje zależność wektora wyniku $\bar{w}$ od rodzaju arytmetyki maszyny cyfrowej (patrz p. 0.1 i 0.2). Należy podkreślić, że zbiór danych a , na którym odwzorowanie $WN(a, \varepsilon)$ jest określone, niekoniecznie jest równy zbiorowi D . Oznaczmy ten wzór przez DN i założmy, że $DN \cap D \neq \emptyset$. Warunek ten odpowiada wymaganiu, by istniały dane a , dla których istnieje rozwiązanie zadania ($a \in D$) i można wyznaczyć numeryczne rozwiązanie danym algorytmem ($a \in DN$). Rozpatrzmy to na przykładzie.

Dana jest liczba zespolona $a = x + iy$. Należy obliczyć $\frac{1}{a^2}$.

Algorytm I: — oblicz $t = y/x$ (tangens fazy liczby a)
 — oblicz $|a|^2 = x^2 + y^2$ (kwadrat modułu liczby a)
 — obliczyć $\operatorname{Re}\left(\frac{1}{a^2}\right) = \frac{1}{a^2} \frac{1-t^2}{1+t^2}$, $\operatorname{Im}\left(\frac{1}{a^2}\right) = \frac{1}{a^2} \frac{-2t}{1+t^2}$

Dyskusja:

Zadanie powyższe jest dobrze postawione, jeśli tylko $x^2 + y^2 \neq 0$. Zatem $D = \mathbb{R}^2 - \{(0, 0)\}$. Algorytm jest poprawnie sformułowany (11 niezbędnych działań), lecz nie dla każdej pary danych $(x, y) \neq 0$ można tym algorytmem znaleźć rozwiązanie. Podczas wykonywania działań mogą bowiem wystąpić zatrzymania pracy maszyny z powodu nadmiaru liczb zmiennopozycyjnych. Będzie tak oczywiście dla wartości $x = 0$, ale nie tylko, bo także dla wartości liczb x o tak małym module, że są zaokrąglane w maszynie do 0. Jeśli ponadto największa liczba zmiennopozycyjna w maszynie ma wartość rzędu 10^{80} , a przyjmimy $y = 10^{50}$ i $x = 10^{-50}$, to algorytm I zatrzyma się z powodu nadmiaru już przy pierwszym działaniu. Widzimy zatem, że $DN \subset D$, lecz $DN \neq D$.

Zwróćmy uwagę, że zbiór D nie zależy ani od dokładności działań wykonywanych przez maszynę cyfrową, ani od zakresu liczb, które można wprowadzać do rejestrów maszyny. Zbiór ten charakteryzuje bowiem jedynie

dane zadanie. Inaczej jest w przypadku zbioru $\mathbf{DN}$; jest on zależny od powyższych własności maszyny cyfrowej. Dla podkreślenia tego, będziemy oznaczać ten zbiór symbolem $\mathbf{DN}(\varepsilon)$. Przyjmiemy poza tym umowę, że jeśli zmieniamy własności arytmetyki maszyny tak, że ε maleje, to równocześnie wzrasta zakres liczb, które można wprowadzać do rejestrów maszyny bez sygnalizowania nadmiaru. Założenie to jest zgodne z najczęściej stosowaną metodą zmniejszania ε (zwiększania dokładności), polegającą na wydłużaniu słów maszynowych. Mówimy, że algorytm obliczeniowy jest *numerycznie stabilny*, jeżeli dla dowolnie wybranych danych $\mathbf{a}_0 \in \mathbf{D}$ istnieje taka dokładność obliczeń ε_0 , że dla $\varepsilon < \varepsilon_0$ mamy $\mathbf{a}_0 \in \mathbf{DN}(\varepsilon)$ oraz

$$\lim_{\varepsilon \rightarrow 0} WN(\mathbf{a}_0, \varepsilon) = W(\mathbf{a}_0)$$

Innymi słowy, algorytm jest numerycznie stabilny wtedy, gdy zwiększając dokładność obliczeń można wyznaczyć (i to z dowolną dokładnością) dowolne istniejące rozwiązanie zadania.

Z powodu błędów zaokrągleń przy zadanej dokładności maszyny $\varepsilon > 0$, jest na ogół $WN(\mathbf{a}, \varepsilon) \neq W(\mathbf{a})$. Interpretując odpowiednio błędy zaokrągleń, często jednak można dobrać takie *zaburzone* dane $\mathbf{a} + \delta\mathbf{a} \in \mathbf{D}$ (patrz: lemat Wilkinsona w p. 0.2), że $WN(\mathbf{a}, \varepsilon) = W(\mathbf{a} + \delta\mathbf{a})$. Wielkość zastępczego zaburzenia $\delta\mathbf{a}$ zależy od danych $\mathbf{a}$, od liczby wykonywanych działań (która ewentualnie zależy od $\mathbf{a}$) i od dokładności obliczeń. Na podstawie lematu Wilkinsona można przyjąć, że gdy $\varepsilon \rightarrow 0$, wówczas $\frac{\|\delta\mathbf{a}\|}{\|\mathbf{a}\|} \rightarrow 0$.

To, jak bardzo wynik $W(\mathbf{a} + \delta\mathbf{a})$ różni się od $W(\mathbf{a})$, zależy od rodzaju zadania i cechę tę nazywamy *uwarunkowaniem zadania*. Mówimy, że *zadanie jest stabilne*, jeżeli dla każdego $\mathbf{a} \in \mathbf{D}$

$$\lim_{\substack{\delta\mathbf{a} \rightarrow 0 \\ \mathbf{a} + \delta\mathbf{a} \in \mathbf{D}}} W(\mathbf{a} + \delta\mathbf{a}) = W(\mathbf{a})$$

czyli, jeżeli występuje ciągła zależność rozwiązania W od danych $\mathbf{a}$ w zbiorze danych $\mathbf{D}$, dla których zadanie jest dobrze postawione.

Niech δw oznacza błąd wyznaczonego rozwiązania, czyli $\delta w = WN(\mathbf{a}, \varepsilon) - W(\mathbf{a})$. Załóżmy, że można dobrać zastępcze zaburzenia $\delta\mathbf{a}$ takie, że $WN(\mathbf{a}, \varepsilon) = W(\mathbf{a} + \delta\mathbf{a})$ dla wszystkich $\mathbf{a} \in \mathbf{D}$. Zazwyczaj potrafimy oszacować wielkość $\frac{\|\delta\mathbf{a}\|}{\|\mathbf{a}\|}$ dla zastępczych zaburzeń danych $\mathbf{a}$ ze zbioru $\mathbf{D}$, tzn.

$$\frac{\|\delta\mathbf{a}\|}{\|\mathbf{a}\|} \leq C, \text{ gdzie } C \text{ zależy od } \mathbf{D} \text{ i } \varepsilon. \text{ Jeżeli można wskazać liczbę } B(\mathbf{a}) \text{ taką,}$$

$$\text{że } \frac{\|\delta w\|}{\|w\|} \leq B(\mathbf{a}) \frac{\|\delta\mathbf{a}\|}{\|\mathbf{a}\|}, \text{ to } B(\mathbf{a}) \text{ nazywamy wskaźnikiem uwarunkowania roz-}$$

wiązywanego zadania. Kres dolny liczb $B(\mathbf{a})$ nazywamy (nieuniknionym) *wzmocnieniem zaburzeń* danych. Wyznaczenie wzmocnienia zaburzeń jest na

ogół bardzo trudne, więc zadowalamy się informacją o uwarunkowaniu uzyskaną za pomocą obliczonej liczby $B(a)$. Jeśli wskaźnik $B(a)$ jest na tyle mały, że górne oszacowanie względnego błędu rozwiązania równe iloczynowi BC ma zadowalającą wartość, to mówimy, że zadanie *jest dobrze uwarunkowane*.

Zwróćmy uwagę, że ocena, czy zadanie jest dobrze, czy źle uwarunkowane, zależy nie tylko od danych a , lecz także od dokładności obliczeń. Gdy bowiem zwiększamy dokładność obliczeń (zmniejszamy ε), wówczas zmniejsza się liczba C , będąca górnym oszacowaniem $\frac{\|\delta a\|}{\|a\|}$. Ponieważ liczba $B(a)$

jest niezależna od ε (charakteryzuje ona jedynie zależność $W(a)$), więc BC maleje. Zadanie, które oceniliśmy jako źle uwarunkowane, może zatem okazać się zupełnie dobrze uwarunkowane po zwiększeniu dokładności obliczeń lub przy zastosowaniu innej (dokładniejszej) maszyny cyfrowej.

Wróćmy do rozwiązanego przykładu obliczania $1/a^2$ dla danej liczby zespolonej $a = x + iy$. Poprawnie sformułowany algorytm I nie jest numerycznie stabilny. Jeśli bowiem $x = 0$, to nie pomoże żaden wzrost dokładności obliczeń — istniejącego dla $y \neq 0$ rozwiązania nie można zatem wyznaczyć tym algorytmem. Rozpatrzmy wobec tego inny algorytm.

Algorytm II: — oblicz

$$r = \operatorname{Re}\left(\frac{1}{a^2}\right) = \frac{x^2 - y^2}{x^2 + y^2}$$

$$u = \operatorname{Im}\left(\frac{1}{a^2}\right) = \frac{-2xy}{x^2 + y^2}$$

Algorytm ten jest poprawnie sformułowany (9 działań) i jest numerycznie stabilny, co wynika z ciągłości odpowiednich wzorów przy założeniu $x^2 + y^2 \neq 0$. Wykonując obliczenia na maszynie cyfrowej uzyskamy zamiast wartości r i u takie liczby, jakie uzyskalibyśmy przy danych $x(1 + \varepsilon_1)$ i $y(1 + \varepsilon_2)$, przy czym jeśli $\varepsilon < 0,01$, to $\varepsilon_1 < 15\varepsilon$ i $\varepsilon_2 < 15\varepsilon$. Posłużenie się zatem algorytmem II daje błąd rozwiązania równoważny względnemu zaburzeniu danych mniejszemu niż 15ε , tzn. $\frac{\|\delta a\|}{\|a\|} < 15\varepsilon = C$. Oszacowanie to dotyczy wszystkich liczb $a \neq 0$. Ze wzoru Taylora mamy

$$\delta w = \frac{1}{(a + \delta a)^2} - \frac{1}{a^2} = \frac{-2\delta a}{(a + \kappa\delta a)^3}$$

dla pewnego $\kappa \in \langle 0; 1 \rangle$, więc

$$\frac{|\delta w|}{|w|} = \frac{2}{\left(1 - \kappa \frac{|\delta a|}{|a|}\right)^3} \frac{|\delta a|}{|a|}$$

Jako $B(a)$ można teraz przyjąć liczbę $4,0 > \frac{2}{(1 - 15\varepsilon)^3}$. Liczbę 4,0 można więc traktować jako wskaźnik uwarunkowania zadania obliczania $1/a^2$ dla zaburzeń $\frac{|\delta a|}{|a|} < 0,15$. Górne oszacowanie względnego błędu rozwiązania wynosi $BC = 60\varepsilon$. Jeżeli wymagamy dokładności obliczania $1/a^2$ co najmniej 1%, to przy posługiwaniu się maszyną, której $\varepsilon = 10^{-4}$, możemy powiedzieć, że zadanie to jest dobrze uwarunkowane. Jeśli jednak $\varepsilon = 10^{-3}$, to bez obliczenia wzmocnienia zaburzeń nie możemy ocenić, czy zadanie jest dobrze uwarunkowane.

INTERPOLACJA

1.1. SFORMUŁOWANIE ZAGADNIENIA INTERPOLACJI

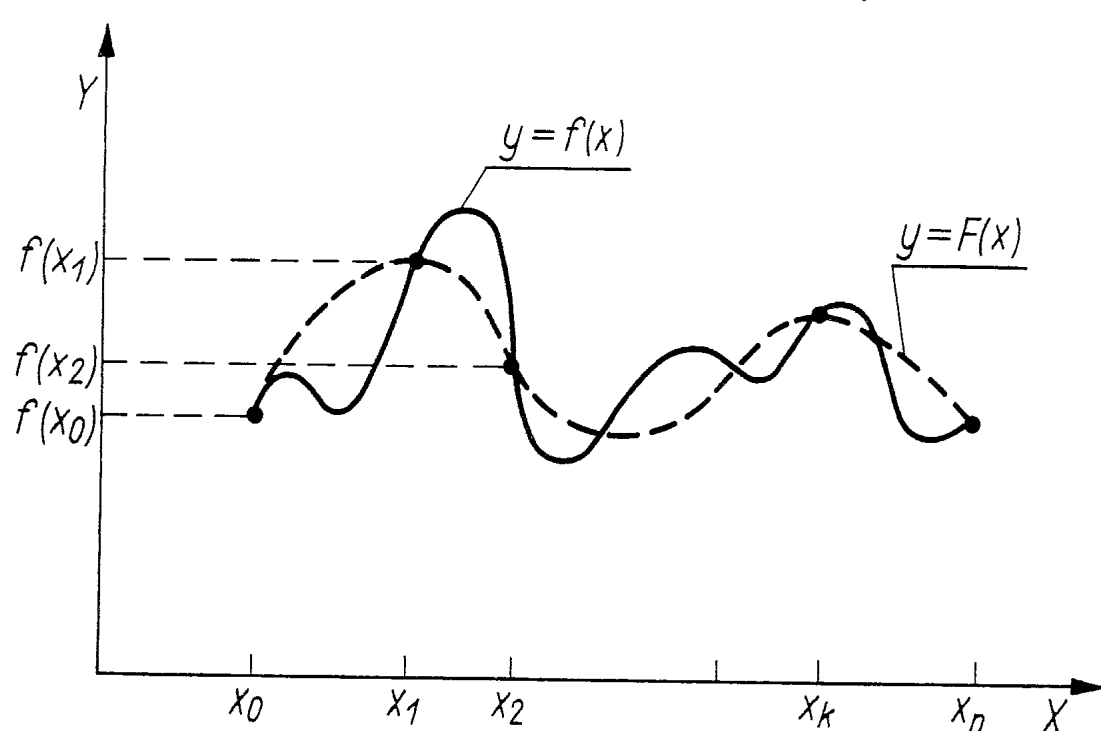
Zagadnienie interpolacji można sformułować następująco: na przedziale $\langle a; b \rangle$ danych jest $n + 1$ różnych punktów $x_0, x_1, \dots, x_n$ — które nazywamy *węzłami interpolacji*, oraz wartości pewnej funkcji $y = f(x)$ w tych punktach

$$f(x_0) = y_0, f(x_1) = y_1, \dots, f(x_n) = y_n \quad (1.1)$$

Zadaniem interpolacji jest wyznaczenie przybliżonych wartości funkcji w punktach nie będących węzłami oraz oszacowanie błędu tych przybliżonych wartości. W tym celu należy znaleźć funkcję $F(x)$, zwaną *funkcją interpolującą*, która w węzłach interpolacji przyjmuje takie same wartości co funkcja $y = f(x)$; patrz rys. 1.1.

Można powiedzieć, że interpolacja jest w pewnym sensie zadaniem odwrotnym do tablicowania funkcji. Przy tablicowaniu mając analityczną postać funkcji budujemy tablicę wartości, przy interpolacji natomiast na podstawie tablicy wartości funkcji określamy jej postać analityczną.

Funkcji interpolującej poszukuje się najczęściej jako funkcji pewnej określonej postaci. Mogą to być wielomiany algebraiczne, wielomiany



[illegible]

1.2. INTERPOLACJA ZA POMOCĄ WIELOMIANÓW

Tw. Istnieje dokładnie jeden wielomian interpolacyjny stopnia co najwyżej n ($n \geq 0$), który w punktach $x_0, x_1, \dots, x_n$ przyjmuje wartości $y_0, y_1, \dots, y_n$.

$$W_n(x) = a_0 + a_1 x + a_2 x^2 \dots + a_n x^n \quad (1.2)$$
$$\begin{array}{l} a_0 + a_1 x_0 + a_2 x_0^2 + \dots + a_n x_0^n = y_0 \\ a_0 + a_1 x_1 + a_2 x_1^2 + \dots + a_n x_1^n = y_1 \\ \\ a_0 + a_1 x_n + a_2 x_n^2 + \dots + a_n x_n^n = y_n \end{array} \quad (1.3)$$
$$A = \begin{bmatrix} 1 & x_0 & x_0^2 & x_0^3 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & x_1^3 & \dots & x_1^n \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_n & x_n^2 & x_n^3 & \dots & x_n^n \end{bmatrix} \quad (1.4)$$
$$D = \begin{vmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{vmatrix} \quad (1.5)$$

więc przy przyjętych założeniach, że $x_i \neq x_j$ dla $i \neq j$ zawsze

$$D = \prod_{0 \leq j < i \leq n} (x_i - x_j) \neq 0 \quad (1.6)$$

Zatem układ (1.3) ma dokładnie jedno rozwiązanie, a wartości a_i według twierdzenia Cramera są określone wzorem

$$a_i = \frac{1}{D} \sum_{j=0}^n y_j D_{ij} \quad (1.7)$$

gdzie D_{ij} ($j = 0, 1, \dots, n$) są kolejnymi dopełnieniami algebraicznymi elementów i -tej kolumny macierzy A .

Z twierdzenia Cramera wynika więc, że istnieje wielomian postaci (1.2) spełniający warunki (1.1) oraz że jest on wyznaczony jednoznacznie. Stopień wielomianu jest nie większy niż n (gdy $a_n = 0$ jest to oczywiście wielomian stopnia odpowiednio niższego).

1.2.1. WZÓR INTERPOLACYJNY LAGRANGE'A

Podstawiając (1.7) do (1.2) i grupując razem te składniki, w których występują jednakowe y_i , otrzymujemy

$$W_n(x) = y_0 \Phi_0(x) + y_1 \Phi_1(x) + \dots + y_n \Phi_n(x) \quad (1.8)$$

gdzie funkcje $\Phi_0(x), \Phi_1(x), \dots, \Phi_n(x)$ są wielomianami stopnia co najwyżej n .

Jak poprzednio, zakładamy, że węzły są rozmieszczone w dowolny sposób. Poszukujemy wielomianu $W_n(x)$ stopnia co najwyżej n określonego równaniem (1.8). Dla każdego $i = 0, 1, 2, \dots, n$ zachodzi zależność

$$W_n(x_i) = y_0 \Phi_0(x_i) + y_1 \Phi_1(x_i) + \dots + y_n \Phi_n(x_i) \quad (1.9)$$

skąd wynika, że

$$\Phi_j(x_i) = \begin{cases} 0, & \text{gdy } j \neq i \\ 1, & \text{gdy } j = i \end{cases} \quad (1.10)$$

Aby określić funkcje $\Phi_j(x)$ należy znaleźć wielomian stopnia n , który w punktach $x_0, x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n$ jest tożsamościowo równy zeru, a w punkcie x_j równa się jedności. Stąd

$$\Phi_j(x) = \lambda (x - x_0)(x - x_1) \dots (x - x_{j-1})(x - x_{j+1}) \dots (x - x_n) \quad (1.11)$$

a ponieważ $\Phi_j(x_j) = 1$, więc

$$1 = \lambda (x_j - x_0)(x_j - x_1) \dots (x_j - x_{j-1})(x_j - x_{j+1}) \dots (x_j - x_n) \quad (1.12)$$

Po wyrugowaniu stałej λ ze wzorów (1.11) i (1.12), otrzymujemy

$$\Phi_j(x) = \frac{(x - x_0)(x - x_1) \dots (x - x_{j-1})(x - x_{j+1}) \dots (x - x_n)}{(x_j - x_0)(x_j - x_1) \dots (x_j - x_{j-1})(x_j - x_{j+1}) \dots (x_j - x_n)} \quad (1.13)$$

zatem

$$\begin{aligned}
 W_n(x) &= y_0 \frac{(x-x_1)(x-x_2)\dots(x-x_n)}{(x_0-x_1)\dots(x_0-x_n)} + y_1 \frac{(x-x_0)(x-x_2)\dots(x-x_n)}{(x_1-x_0)(x_1-x_2)\dots(x_1-x_n)} + \\
 &+ \dots + y_n \frac{(x-x_0)(x-x_1)\dots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\dots(x_n-x_{n-1})} = \\
 &= \sum_{j=0}^n y_j \frac{(x-x_0)(x-x_1)\dots(x-x_{j-1})(x-x_{j+1})\dots(x-x_n)}{(x_j-x_0)(x_j-x_1)\dots(x_j-x_{j-1})(x_j-x_{j+1})\dots(x_j-x_n)} \quad (1.14)
 \end{aligned}$$

Przyjmując oznaczenie

$$\omega_n(x) = (x-x_0)(x-x_1)\dots(x-x_n) \quad (1.15)$$

możemy (1.14) zapisać w postaci

$$W_n(x) = \sum_{j=0}^n y_j \frac{\omega_n(x)}{(x-x_j) \left\{ \frac{\omega_n(x)}{x-x_j} \right\} \Big|_{x=x_j}} = \sum_{j=0}^n y_j \frac{\omega_n(x)}{(x-x_j) \omega'_n(x_j)} \quad (1.16)$$

gdzie $y_j = y(x_j)$, a $\omega'_n(x_j)$ jest wartością pochodnej wielomianu $\omega_n(x)$ w punkcie x_j (będącym zerem tego wielomianu), co można łatwo pokazać rozpatrując wzór Taylora w otoczeniu punktu x_j .

Otrzymany wielomian spełnia wszystkie warunki wymagane przy interpolacji. Wzór (1.14) nazywamy *wzorem interpolacyjnym Lagrange'a*. Na podstawie udowodnionego w p. 1.2 twierdzenia możemy stwierdzić, że wielomian ten jest jedynym wielomianem stopnia co najwyżej n .

Przykład 1. Znaleźć wielomian interpolacyjny, który w punktach $-2, 1, 2, 4$ przyjmuje wartości $3, 1, -3, 8$.

Stosujemy wzór interpolacyjny Lagrange'a (1.14) dla $n = 3$.

Mamy

$$\begin{aligned}
 W_3(x) &= 3 \cdot \frac{(x-1)(x-2)(x-4)}{(-2-1)(-2-2)(-2-4)} + 1 \cdot \frac{(x+2)(x-2)(x-4)}{(1+2)(1-2)(1-4)} - \\
 &- 3 \cdot \frac{(x+2)(x-1)(x-4)}{(2+2)(2-1)(2-4)} + 8 \cdot \frac{(x+2)(x-1)(x-2)}{(4+2)(4-1)(4-2)} = \\
 &= -\frac{1}{24}(x^3 - 7x^2 + 14x - 8) + \frac{1}{9}(x^3 - 4x^2 - 4x + 16) + \frac{3}{8}(x^3 - 3x^2 - 6x + 8) + \\
 &+ \frac{2}{9}(x^3 - x^2 - 4x + 4) = \frac{2}{3}x^3 - \frac{3}{2}x^2 - \frac{25}{6}x + 6
 \end{aligned}$$

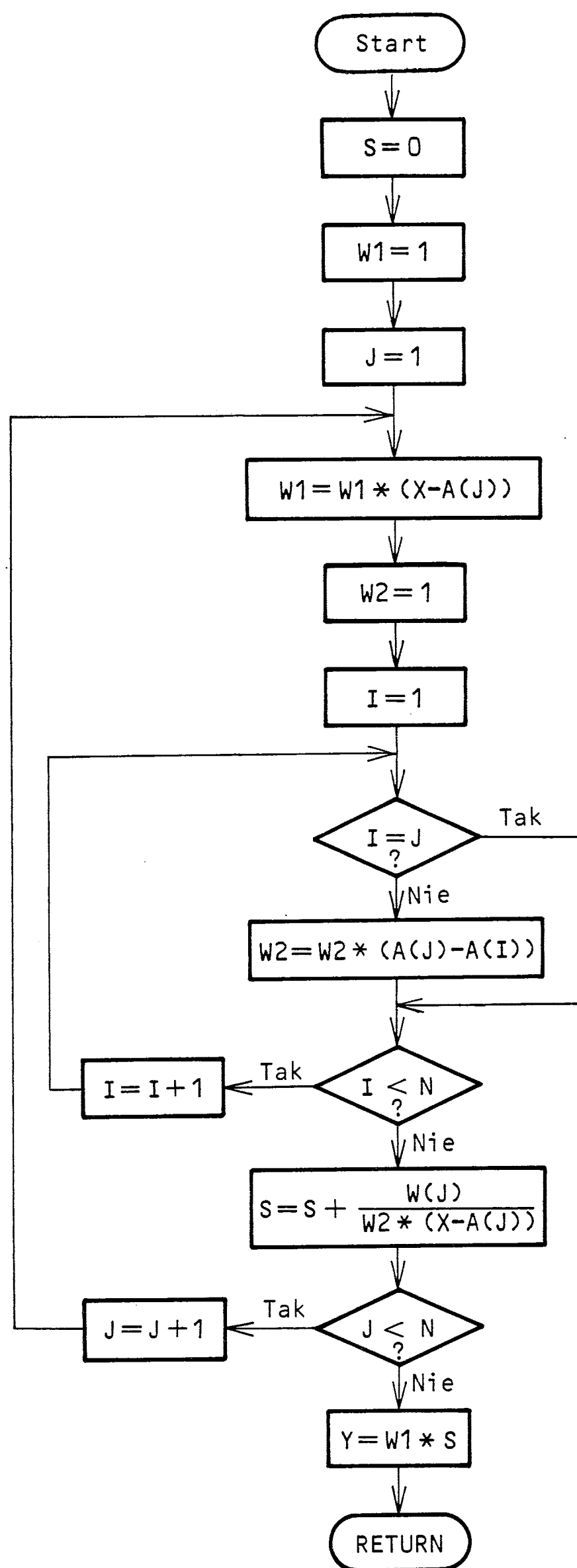
Przykład 2. Korzystając ze wzoru interpolacyjnego Lagrange'a (1.14) napisać program w Turbo Pascalu obliczania wartości wielomianu interpolacyjnego stopnia $n-1$ w dowolnym punkcie x leżącym wewnątrz przedziału $\langle x_1; x_n \rangle$ i różnym od węzła.

Na podstawie wzorów (1.14)–(1.16) mamy

$$W_n(x) = \omega_n(x) \sum_{j=1}^n y_j \frac{1}{\omega'_n(x_j)(x - x_j)}$$

Procedura LAGR wymaga umieszczenia wartości argumentów (węzłów) w tablicy A (o wymiarach $1 \times n$), a odpowiadających tym argumentom wartości funkcji w tablicy W (również o wymiarach $1 \times n$); $W1$ i $W2$ oznaczają odpowiednio wartości $\omega_n(x)$ i $\omega'_n(x)$ (wzory (1.15) i (1.16)) obliczone w punkcie x_j , X oznacza wartość argumentu wejściowego, Y — wartość wielomianu interpolacyjnego w punkcie X .

Schemat blokowy procedury LAGR jest przedstawiony na rys. 1.2.



Treść programu:

```

CONST max = ....;
TYPE zakres = 1..max;
    wektor = ARRAY[zakres] OF real;
.....
PROCEDURE LAGR(X:real; VAR A,W:wektor; VAR Y:real; N:zakres);
VAR S,W1,W2 : real;
    I,J : integer;
BEGIN
    S := 0.0;
    W1 := 1.0;
    FOR J := 1 TO N DO
        BEGIN
            W1 := W1 * (X - A[J]);
            W2 := 1;
            FOR I := 1 TO N DO
                IF NOT(I=J) THEN W2 := W2 * (A[J] - A[I]);
            S := S + W[J]/(W2 * (X - A[J]));
            END;
        Y := W1 * S;
    END;

```

U w a g a: Węzły zostały ponumerowane od 1 do N — zatem otrzymany wielomian będzie stopnia co najwyżej $n = N - 1$.

Wartość wielomianu interpolacyjnego Lagrange'a (stopnia n), opartego na $n + 1$ dowolnych (różnych) węzłach $x_0, x_1, \dots, x_n$ w danym punkcie $x \in \langle x_0, x_n \rangle$ można obliczyć stosując *iteracyjną metodę Aitkena*. Niech $W_{i,j}$ oznacza wielomian stopnia pierwszego, który w punktach x_i, x_j przyjmuje wartości y_i, y_j

$$W_{i,j}(x) = \frac{\begin{vmatrix} y_i & x_i - x \\ y_j & x_j - x \end{vmatrix}}{x_j - x_i}$$

Podobnie, $W_{i,j,k}(x)$ oznacza wielomian stopnia drugiego, który w trzech punktach x_i, x_j, x_k ($i \neq j \neq k$) przyjmuje wartości y_i, y_j, y_k

$$W_{i,j,k}(x) = \frac{\begin{vmatrix} W_{i,j}(x) & x_j - x \\ W_{i,k}(x) & x_k - x \end{vmatrix}}{x_k - x_j}$$

Ogólnie można wykazać, że

$$W_{1,2,\dots,k,m}(x) = \frac{\begin{vmatrix} W_{1,2,\dots,k-1,k}(x) & x_k - x \\ W_{1,2,\dots,k-1,m}(x) & x_m - x \end{vmatrix}}{x_m - x_k}$$

Przyjmijmy, że mamy dane dwie tablice: tablicę wartości argumentów x_i oraz tablicę wartości funkcji interpolowanej $y_i, i = 1, 2, \dots, n$. Aby znaleźć wartość Y funkcji interpolowanej w punkcie X różnym od podanych punktów x_i , obliczamy kolejno

$$W_{1,2} = \frac{1}{x_2 - x_1} \begin{vmatrix} y_1 & x_1 - X \\ y_2 & x_2 - X \end{vmatrix}$$

$$W_{1,2} = \frac{1}{x_3 - x_1} \begin{vmatrix} y_1 & x_1 - X \\ y_3 & x_3 - X \end{vmatrix}$$

$$W_{1,2,3} = \frac{1}{x_3 - x_2} \begin{vmatrix} W_{1,2} & x_2 - X \\ W_{1,3} & x_3 - X \end{vmatrix}$$

$$W_{1,2,4} = \frac{1}{x_4 - x_2} \begin{vmatrix} W_{1,2} & x_2 - X \\ W_{1,4} & x_4 - X \end{vmatrix}$$

$$W_{1,2,3,4} = \frac{1}{x_4 - x_3} \begin{vmatrix} W_{1,2,3} & x_3 - X \\ W_{1,2,4} & x_4 - X \end{vmatrix} \quad \text{itd.}$$

Kolejne wielomiany tworzą macierz trójkątną zgodnie z poniższym schematem, zwanym schematem *Aitkena*:

$$\begin{vmatrix} x_1 & y_1 & & & & \\ x_2 & y_2 & W_{1,2} & & & \\ x_3 & y_3 & W_{1,3} & W_{1,2,3} & & \\ x_4 & y_4 & W_{1,4} & W_{1,2,4} & W_{1,2,3,4} & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ x_n & y_n & W_{1,n} & W_{1,2,n} & W_{1,2,3,n} & \dots W_{1,2,3,\dots,n} \end{vmatrix}$$

Ostatni z wielomianów tego schematu jest szukanym wielomianem interpolacyjnym.

Przykład 3. Napisać program obliczania wartości wielomianu interpolacyjnego wykorzystując metodę interpolacji Aitkena-Lagrange'a.

Procedura obliczania wartości kolejnych wielomianów jest następująca. Oznaczmy wejściową tablicę argumentów przez A , a tablicę wartości funkcji przez W . Schemat generowania kolejnych elementów tablicy opisuje zależność

$$W(J) = \frac{W(I)(A(J) - X) - W(J)(A(I) - X)}{A(J) - A(I)}$$

$$I = 1, 2, \dots, J-1 \quad \text{dla} \quad J = 2, 3, \dots, n$$

Kolejno są obliczane wartości

$$W_{1,2}, W_{1,3}, W_{1,2,3}, W_{1,4}, W_{1,2,4}, W_{1,2,3,4} \text{ itd.}$$

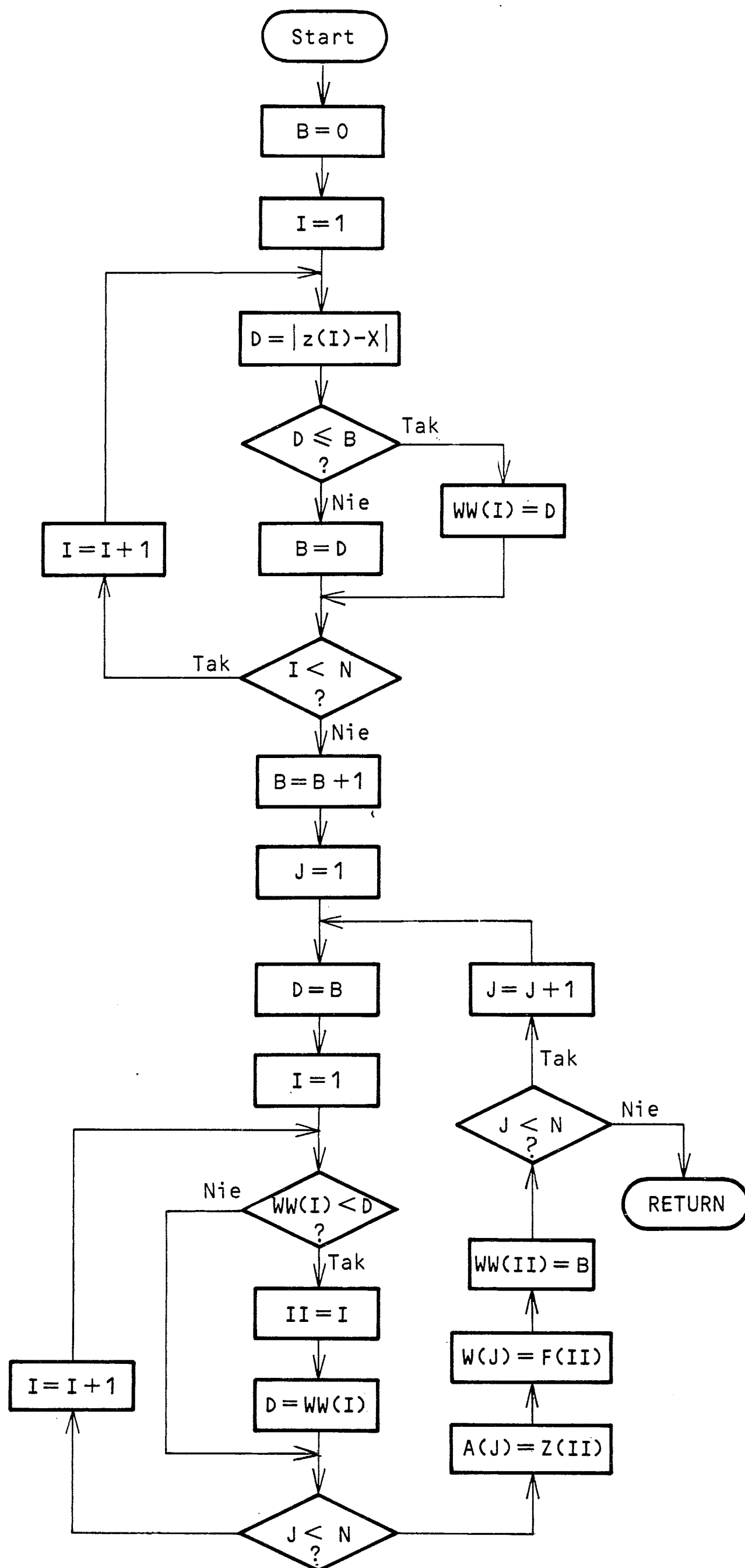
Ostatnia wartość J -tego wiersza jest zmagazynowana w $W(J)$. Wykonanie procedury zostaje przerwane, gdy:

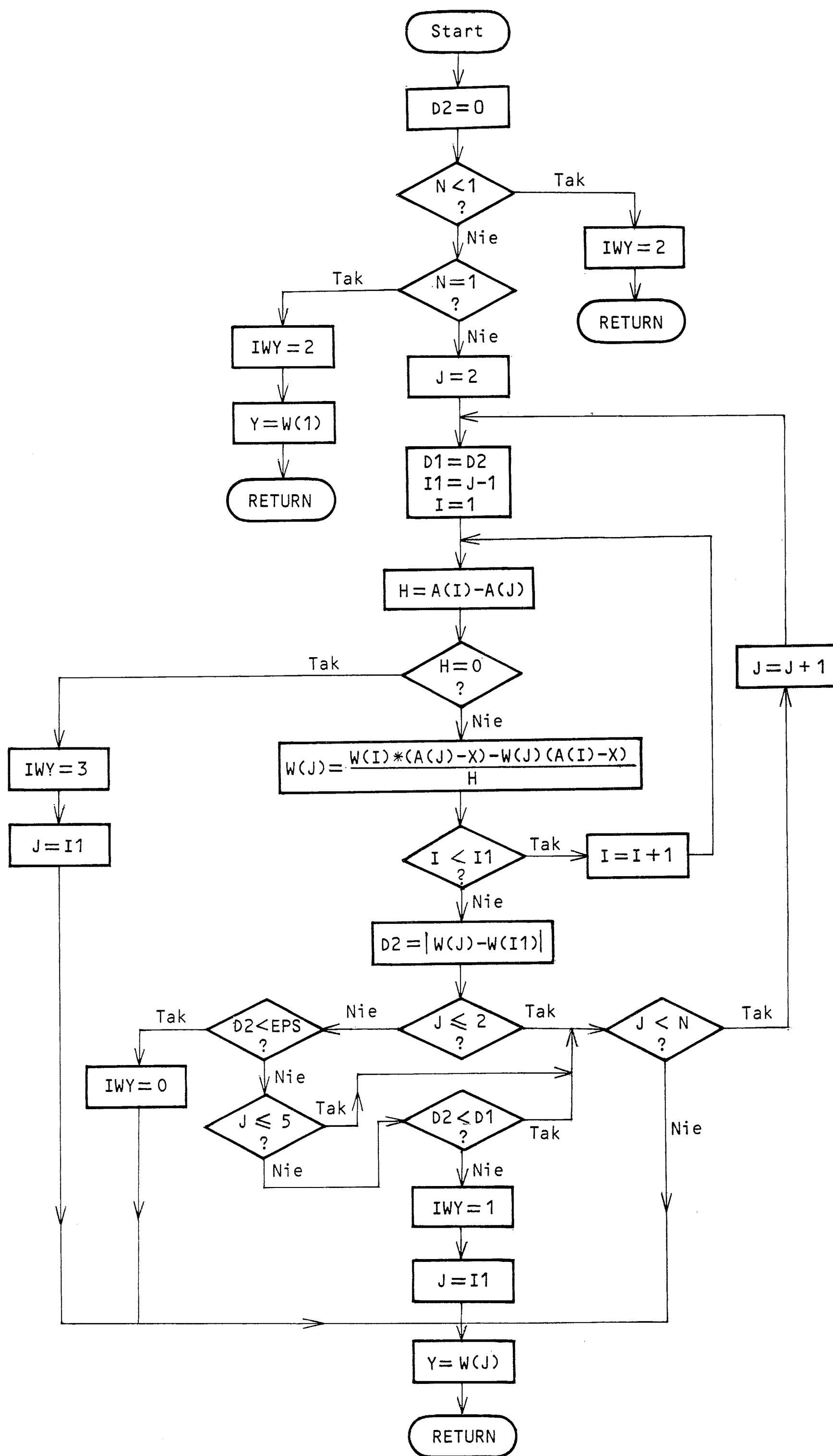
- 1) $n \leq 1$, wówczas zmienna $IWY = 2$;
- 2) różnica między dwiema kolejno obliczonymi wartościami na głównej przekątnej w tablicy schematu Aitkena jest mniejsza niż dana wartość EPS (badanie dla $J \geq 3$); wówczas zmienna $IWY = 0$;
- 3) różnica (wartość bezwzględna) zaczyna wzrastać, co wskazuje na pojawienie się błędu (badanie dla $J \geq 5$); zmienna $IWY = 1$.

Ostatecznie zmienna Y przyjmuje wartość $W(J)$, w przypadku gdy moduł różnicy

kolejnych elementów głównej przekątnej schematu Aitkena monotonicznie maleje, a wartość $W(J-1)$, gdy różnica ta zmienia się w sposób oscylacyjny.

U w a g a. Program wymaga, aby tablica $A(J)$ i odpowiadająca jej tablica $W(J)$ były uszeregowane tak, by $|A(I) - X| \geq |A(J) - X|$ dla $I > J$. Przy takim uporządkowaniu wartość wyrażenia $\omega_n(x)$ (wzór (1.15)), które występuje w reszcie, jest stale najmniejsza. Każde przybliżenie interpolacyjne ma wówczas tendencję do optymalności i zbieżność interpolacji jest





„z grubsza” najszybsza [50]. Uszeregowanie takie można wykonać na przykład za pomocą procedury SEL.

Procedura SEL ma na wejściu dwie tablice: tablicę argumentów Z i tablicę F , odpowiadających tym argumentom wartości funkcji, oraz wartość X . Na wyjściu w tablicach A i W są ustawione w wymaganym porządku argumenty i wartości funkcji. Tablica WW jest tablicą roboczą, a zmienna N określa wymiar tablic Z , F , A , W i WW . Najpierw oblicza się wektor WW , którego elementami są kolejne różnice $|Z(I) - X|$ ($I = 1, 2, \dots, N$), oraz wybiera maksymalny element B wektora WW . Następnie wybiera się najmniejszy element wektora WW i odpowiadającą mu zwartość $Z(II)$, umieszczając w tablicy A jako element $A(1)$, a to miejsce w tablicy WW zapelnia się wartością $B + 1$. Postępuje się tak cyklicznie aż do wyczerpania wszystkich elementów WW . Ustawia się równocześnie wektor W zgodnie z otrzymaną kolejnością. Schematy blokowe procedur SEL i LAGRAIT przedstawiono na rys. 1.3 i 1.4.

Dla obu procedur (SEL i LAGRAIT) zakładamy, że w bloku zewnętrznym znajdują się odpowiednie definicje typów i deklaracje zmiennych (patrz — procedura LAGR z przykładu 2).

Treść programu:

```

PROCEDURE SEL (x: REAL; VAR Z,F,WW,A,W: wektor; N: zakres);
VAR I,II,J: integer;
    B,D: real;
BEGIN
    B := 0.0;
    FOR I := 1 TO N DO
        BEGIN
            D := ABS(Z[I]) - X;
            IF D > B THEN B := D;
            WW[I] := D;
        END;
    B := B + 1.0;
    FOR J := 1 TO N DO
        BEGIN
            D := B;
            FOR I := 1 TO N DO
                IF WW[I] < D THEN
                    BEGIN
                        II := I;
                        D := WW[I];
                    END;
            A[J] := Z[II];
            W[J] := F[II];
            WW[II] := B;
        END;
    END;
PROCEDURE LAGRAIT(VAR A,W: wektor; VAR X,Y: real; N: zakres;
    EPS: real; VAR IWY: integer);
VAR D1,D2,H: real;
    I,J,I1: integer;
BEGIN
    IWY := 2;
    D2 := 0.0;
    IF N < 1 THEN EXIT;
    IF N = 1 THEN
        BEGIN
            Y := W[N];
            EXIT;
        END;

```

```

        END;
    FOR J := 2 TO N DO
        BEGIN
            D1 := D2;
            I1 := J-1;
            FOR I := 1 to I1 DO
                BEGIN
                    H := A[I] - A[J];
                    IF H = 0.0 THEN
                        BEGIN
                            IWY := 3;
                            Y := W[I1];
                            EXIT;
                        END;
                    W[J] := (W[I]*(X - A[J]) - W[J]*(X - A[I]))/H;
                END;
            D2 := ABS(W[J] - W[I1]);
            IF J > 2 THEN
                BEGIN
                    IF D2 <= EPS THEN
                        BEGIN
                            IWY := 0;
                            Y := W[J];
                            EXIT;
                        END;
                    IF J >= 5 THEN
                        IF D2 >= D1 THEN
                            BEGIN
                                IWY := 1;
                                Y := W[I1];
                                EXIT;
                            END;
                        END;
                    END;
                END;
            END;
        Y := W[J];
    END;

```

1.2.2. OSZACOWANIE BŁĘDU WZORU INTERPOLACYJNEGO

Wzór interpolacyjny (1.14) spełnia warunki (1.1). Powstaje pytanie, z jaką dokładnością wielomian ten przybliża funkcję $f(x)$ w pozostałych punktach leżących wewnątrz przedziału $\langle a; b \rangle$, czyli określenie wielkości błędu $\varepsilon(x)$ zdefiniowanego następująco:

$$\varepsilon(x) = f(x) - W_n(x) \quad (1.17)$$

Aby określić $\varepsilon(x)$ przyjmujemy, że funkcja $f(x)$ na rozpatrywanym przedziale $\langle a; b \rangle$ ma pochodne do rzędu $n+1$ włącznie. Wprowadzamy funkcję pomocniczą

$$\varphi(u) = f(u) - W_n(u) + K(u - x_0)(u - x_1) \dots (u - x_n) \quad (1.18)$$

w której K jest pewną stałą. Oczywiście $\varphi(x_0) = \varphi(x_1) = \dots = \varphi(x_n) = 0$. Współczynnik K dobieramy w taki sposób, aby pierwiastkiem funkcji $\varphi(u)$

był również punkt $\bar{x}$, różny od węzłów interpolacji. Stąd

$$K = \frac{f(\bar{x}) - W_n(\bar{x})}{(\bar{x} - x_0)(\bar{x} - x_1) \dots (\bar{x} - x_n)} = \frac{f(\bar{x}) - W_n(\bar{x})}{\omega_n(\bar{x})} \quad (1.19)$$

Mianownik wyrażenia (1.19) jest różny od zera dla każdego $\bar{x} \neq x_i$ ($x_0 < x_1 < \dots < x_n$). Funkcja $\varphi(u)$ ma więc $n + 2$ miejsc zerowych w przedziale $\langle a; b \rangle$, w punktach: $x_0, x_1, x_2, \dots, x_n, \bar{x}$. Na podstawie twierdzenia Rolle'a [72] możemy stwierdzić, że pochodna $\varphi'(u)$ funkcji $\varphi(u)$ ma co najmniej jedno miejsce zerowe w każdym z podprzedziałów, których końcami są miejsca zerowe funkcji $\varphi(u)$, a zatem wewnątrz przedziału o końcach $\min(x, x_0)$, $\max(x, x_n)$ ma co najmniej $n + 1$ miejsc zerowych: $\xi_0^{(1)}, \xi_1^{(1)}, \dots, \xi_n^{(1)}$. Stosując ponownie twierdzenie Rolle'a do pochodnej $\varphi'(u)$, stwierdzamy, że istnieje co najmniej n punktów $\xi_0^{(2)}, \xi_1^{(2)}, \dots, \xi_{n-1}^{(2)}$ takich, że

$$\varphi''(\xi_0^{(2)}) = \varphi''(\xi_1^{(2)}) = \dots = \varphi''(\xi_{n-1}^{(2)}) = 0 \quad (1.20)$$

Kontynuując to rozumowanie dochodzimy do wniosku, że istnieje co najmniej jeden taki punkt w przedziale $(\min(x, x_0); \max(x, x_n))$, że

$$\varphi^{(n+1)}(\xi) = 0 \quad (1.21)$$

Ponieważ

$$W_n^{(n+1)}(x) = 0 \quad (1.22)$$

$$\omega_n^{(n+1)}(x) = (n + 1)!$$

więc

$$\varphi^{(n+1)}(u) = f^{(n+1)}(u) - K(n + 1)! \quad (1.23)$$

Korzystając z (1.21), dla $u = \xi$ otrzymujemy

$$K = \frac{f^{(n+1)}(\xi)}{(n + 1)!}$$

stąd

$$f(x) - W_n(x) = \frac{f^{(n+1)}(\xi)}{(n + 1)!} \omega_n(x) \quad (1.24)$$

Jeżeli kres górny modułu $(n + 1)$ -szej pochodnej funkcji $f(x)$ na przedziale $\langle a; b \rangle$ oznaczymy przez

$$M_{n+1} = \sup_{x \in \langle a; b \rangle} |f^{(n+1)}(x)| \quad (1.25)$$

to

$$|f(x) - W_n(x)| \leq \frac{M_{n+1}}{(n + 1)!} |\omega_n(x)| \quad (1.26)$$

Wyrażenie (1.26) może służyć do oszacowania błędu bezwzględnego wzoru

interpolacyjnego, jeżeli tylko można oszacować wartość $f^{(n+1)}(\xi)$. Podamy teraz kilka przykładów takich oszacowań.

Przykład 1. Ocenic, z jaką dokładnością można obliczyć wartość $\ln 100,5$ przy użyciu wzoru interpolacyjnego Lagrange'a, jeżeli dane są wartości: $\ln 100$, $\ln 101$, $\ln 102$ i $\ln 103$.

Mamy

$$f(x) = \ln x, \quad n = 3, \quad a = 100, \quad b = 103, \quad f^{(4)}(x) = -\frac{6}{x^4}$$

$$M_4 = \sup_{x \in \langle 100; 103 \rangle} |f^{(4)}(x)| = 6/100^4$$

$$|\ln 100,5 - W(100,5)| \leq \frac{6}{100^4 \cdot 4!} \cdot 0,5 \cdot 0,5 \cdot 1,5 \cdot 2,5 \approx 2,344 \cdot 10^{-9}$$

Przykład 2. Z jaką dokładnością można obliczyć wartość $\sin \frac{\pi}{36}$ stosując wzór interpolacyjny Lagrange'a, jeżeli dane są wartości: $\sin 0$; $\sin \frac{\pi}{6}$, $\sin \frac{\pi}{4}$ i $\sin \frac{\pi}{3}$.

Mamy

$$f(x) = \sin x, \quad n = 3, \quad a = 0, \quad b = \frac{\pi}{3}, \quad f^{(4)}(x) = \sin x, \quad M_4 = \sqrt{3}/2$$

$$\left| \sin \frac{\pi}{36} - W_4\left(\frac{\pi}{36}\right) \right| \leq \frac{\sqrt{3}}{2} \cdot \frac{1}{4!} \left| \left(\frac{\pi}{36} - 0\right) \left(\frac{\pi}{36} - \frac{\pi}{6}\right) \left(\frac{\pi}{36} - \frac{\pi}{4}\right) \left(\frac{\pi}{36} - \frac{\pi}{3}\right) \right| \approx 9 \cdot 10^{-4}$$

1.2.3. Problem optymalnego doboru węzłów interpolacji

Jak widać ze wzoru (1.26), wielkość błędu zależy zarówno od wartości $f^{(n+1)}(\xi)$, jak i od wartości $\omega_n(x)$. Wartość czynnika $f^{(n+1)}(\xi)$ jest uzależniona tylko od własności funkcji $f(x)$ i nie mamy na nią wpływu, natomiast wartość $\omega_n(x)$ zależy od wyboru węzłów interpolacji. Zajmiemy się zatem zagadnieniem, w jaki sposób wybrać węzły interpolacji x_i , aby

$$\sup_{x \in \langle a; b \rangle} |\omega_n(x)| \tag{1.27}$$

miało jak najmniejszą wartość. Zagadnienie to zostało sformułowane przez rosyjskiego matematyka P.L. Czebyszewa jako zagadnienie znajdowania wielomianu algebraicznego najlepiej przybliżającego zero na zadanym przedziale. Aby można je było rozwiązać, skorzystajmy z wielomianów Czebyszewa

$$T_n(x) = \cos(n \arccos x) \tag{1.28}$$

Korzystając z prostych tożsamości trygonometrycznych, łatwo możemy wykazać, że wielomian $T_n(x)$ jest identyczny z pewnym wielomianem algebraicznym „zawężonym” do przedziału $\langle -1; 1 \rangle$. Można wykazać, że

$$T_0(x) = 1 \tag{1.29}$$

$$T_1(x) = \cos(\arccos x) = x$$

.....

oraz

$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x), \quad n = 2, 3, 4, \dots \quad (1.29a)$$

Każdy wielomian (1.28) stopnia n ma n różnych pierwiastków w punktach

$$x_m = \cos\left(\frac{2m+1}{2n}\pi\right), \quad m = 0, 1, 2, \dots, n-1 \quad (1.30)$$

zawartych między -1 i $+1$. Współczynnik przy najwyższej potędze w $T_n(x)$ jest równy 2^{n-1} . Ponieważ szukamy wielomianu, który przy najwyższej potędze będzie miał współczynnik równy jedności, więc

$$T_{n+1}^*(x) = \frac{1}{2^n} T_{n+1}(x) = (x - x_0)(x - x_1) \dots (x - x_n) \quad (1.31)$$

gdzie $x_m (m = 0, 1, 2, \dots, n)$ są pierwiastkami wielomianu $T_{n+1}(x)$. Gdy rozpatrywany przedział $\langle a; b \rangle$ jest przedziałem $\langle -1; 1 \rangle$, wówczas wielomianem algebraicznym stopnia $n+1$ takim, aby wyrażenie (1.27) miało najmniejszą wartość, jest wielomian

$$\omega_n(x) = \frac{1}{2^n} T_{n+1}(x)$$

i wówczas

$$\sup_{x \in \langle -1; 1 \rangle} |\omega_n(x)| = \frac{1}{2^n}$$

Jeżeli w przedziale $\langle -1; 1 \rangle$ za węzły interpolacji przyjmiemy zera wielomianu Czebyszewa, to można udowodnić, że

$$|f(x) - W_n(x)| \leq \frac{M_{n+1}}{2^n(n+1)!} \quad (1.31a)$$

Proste transformacje liniowe

$$x = \frac{1}{2}[(b-a)z + (b+a)]$$

oraz

$$z = \frac{1}{b-a}(2x - b - a)$$

dla $x \in \langle a; b \rangle$ i $z \in \langle -1; +1 \rangle$ pozwalają sprowadzić każdy przedział $\langle a; b \rangle$ do przedziału $\langle -1; 1 \rangle$ bez zmiany wartości błędu interpolacji (gdyż nie ulegają zmianie proporcje odpowiednich wartości funkcji), a także przenieść interpolację z przedziału $\langle -1; 1 \rangle$ do dowolnego przedziału $\langle a; b \rangle$, gdzie optymalnymi węzłami będą punkty

$$x_m = \frac{1}{2} \left[(b-a) \cos \frac{2m+1}{2n+2} \pi + (b+a) \right] \quad (1.32)$$

dla $m = 0, 1, 2, \dots, n$, przy czym nowe węzły x_m nie są rozmieszczone w równych odstępach, lecz są zagęszczone przy końcach przedziału.

Wielomian $T_{n+1}(x)$ jest wielomianem stopnia $n+1$, który na przedziale $\langle -1; 1 \rangle$ najlepiej przybliża zero i jest to jedyny taki wielomian stopnia $n+1$, którego współczynnik przy najwyższej potęgze jest równy jedności. Wybierając jako węzły liczby (1.32) otrzymamy oszacowanie błędu

$$|f(x) - W_n(x)| \leq \frac{M_{n+1}}{(n+1)!} \frac{(b-a)^{n+1}}{2^{n+1}}$$

Wielomian $W_n(x)$ wyznaczony przez te węzły nie daje (na ogół) najmniejszego błędu, lecz tylko najmniejszą wartość oszacowania (1.26). Węzły wielomianu o minimalnym błędzie (optymalnego wielomianu aproksymującego) są z reguły bardzo trudne do wyznaczenia.

Ćwiczenia

1. Dane są wartości funkcji $y = f(x)$ w trzech punktach: $y(0) = 1$, $y(1) = 1$ i $y(2) = 3$. Wyznaczyć wielomian stopnia drugiego, dla którego $W(0) = 1$, $W(1) = 1$ i $W(2) = 3$.

2. Znaleźć wielomian interpolacyjny Lagrange'a dla funkcji określonej tablicą wartości 1/1.

1/1

x	0	2	3	5	6
$f(x)$	1	3	2	5	6

3. Korzystając z wzoru interpolacyjnego Lagrange'a udowodnić, że

$$\sum_{m=0}^{2p+1} \frac{(-1)^m}{(2p+1-2m) \cdot m! (2p+1-2m)!} = (-1)^p \frac{2^{2p+1}}{[(2p+1)!]^2}$$

4. Obliczyć wartość $\sqrt{117}$ za pomocą wzoru interpolacyjnego Lagrange'a dla funkcji $y = \sqrt{x}$ i węzłów interpolacji: $x_0 = 100$, $x_1 = 121$ i $x_2 = 144$. Oszacować błąd.

5. a) Znaleźć wielomian interpolacyjny stopnia co najwyżej trzeciego z minimalnym oszacowaniem błędu dla funkcji $y = \sqrt{x}$ na przedziale $\langle 1; 9/4 \rangle$. W s k a z ó w k a. Jako węzły interpolacji należy przyjąć punkty: $1,25/16$, $16/9$ i $9/4$.

b) Oszacować resztę według wzoru (1.26).

c) Korzystając z otrzymanego wielomianu obliczyć $y(2)$, a następnie porównać otrzymany wynik z wartością $\sqrt{2} = 1,4142135$.

d) Korzystając z węzłów Czebyszewa znaleźć optymalny wielomian interpolacyjny dla tej funkcji — optymalny z punktu widzenia doboru węzłów. Oszacować błąd według wzoru (1.31a).

6. Niech $h = \max(x_{i+1} - x_i)$. Udowodnić, że

$$|f(x) - W_n(x)| \leq \frac{M_{n+1} h^{n+1}}{4(n+1)}$$

gdzie $M_{n+1} = \max_{x \in \langle x_0; x_n \rangle} |f^{(n+1)}(x)|$ dla $x \in \langle x_0; x_n \rangle$.

Odpowiedzi. 1. $W_2(x) = x^2 - x + 1$. 2. $W_4(x) = -\frac{11}{120}x^4 + \frac{73}{60}x^3 - \frac{601}{120}x^2 + \frac{413}{60}x + 1$. 3. Wskazówka. Rozpatrzeć funkcję $f(x) = 1$ dla $x \in \langle -1; +1 \rangle$ biorąc jako punkty $x_m^{(2p+1)} = -1 + 2m/(2p+1)$ dla $m = 0, 1, \dots, 2p+1$. 4. $10,815923, |R_2| \leq 1,1475 \cdot 10^{-3}$. 5. d) Optymalnymi węzłami są: $x_k = \frac{5}{8} \cos \frac{2k+1}{8} \pi + \frac{13}{8}$ dla $k = 0, 1, 2, 3$, czyli $x_0 = 2,20242$, $x_1 = 1,86418$, $x_2 = 1,38582$, $x_3 = 1,04758$.

W punkcie tym omówimy inny wzór służący do interpolacji za pomocą wielomianów algebraicznych. Do tego celu niezbędne jest wprowadzenie pojęcia ilorazów różnicowych.

$$\Delta x_i = x_{i+1} - x_i, \quad i = 0, 1, 2, \dots$$

nie są na ogół stałe. Wyrażenia

nazywamy *ilorazami różnicowymi pierwszego rzędu*.

Analogicznie definiujemy *ilorazy różnicowe rzędu drugiego*

$$\begin{aligned} f(x_0; x_1; x_2) &= \frac{f(x_1; x_2) - f(x_0; x_1)}{x_2 - x_0} \\ &\dots\dots\dots \\ f(x_{n-2}; x_{n-1}; x_n) &= \frac{f(x_{n-1}; x_n) - f(x_{n-2}; x_{n-1})}{x_n - x_{n-2}} \end{aligned} \tag{1.34}$$

Ogólnie iloraz różnicowy rzędu n tworzymy z ilorazu różnicowego rzędu $n - 1$ za pomocą wzoru rekurencyjnego

$$f(x_i; x_{i+1}; \dots; x_{i+n}) = \frac{f(x_{i+1}; x_{i+2}; \dots; x_{i+n}) - f(x_i; x_{i+1}; \dots; x_{i+n-1})}{x_{i+n} - x_i} \quad (1.35)$$

dla $n = 1, 2, \dots$ oraz $i = 0, 1, 2, \dots$

Z ilorazów różnicowych tworzy się zazwyczaj tablicę; zasadę jej tworzenia pokazuje tablica 1/2. Przykładowo, dla wielomianu $f(x) = x^3$ o węzłach interpolacji $x_0 = 0$, $x_1 = 2$, $x_2 = 3$, $x_3 = 5$ i $x_4 = 6$ tablica ilorazów różnicowych ma postać tablicy 1/3.

Założmy, że funkcja $f(x)$ jest określona tablicą wartości $(x_0, f(x_0))$, $((x_1, f(x_1)))$, ..., $((x_n, f(x_n)))$, gdzie $x_0, x_1, \dots, x_n$ są węzłami interpolacji. Niech $W_n(x)$ będzie wielomianem interpolacyjnym zapisanym wzorem (1.14),

1/2

x_i	$f(x_i)$	Ilorazy różnicowe				
		rzędu 1	rzędu 2	rzędu 3	rzędu 4	rzędu 5
x_0	$f(x_0)$					
x_1	$f(x_1)$	$f(x_0; x_1)$				
x_2	$f(x_2)$	$f(x_1; x_2)$	$f(x_0; x_1; x_2)$			
		$f(x_2; x_3)$	$f(x_1; x_2; x_3)$	$f(x_0; x_1; x_2; x_3)$		
x_3	$f(x_3)$		$f(x_2; x_3; x_4)$	$f(x_1; x_2; x_3; x_4)$	$f(x_0; x_1; x_2; x_3; x_4; x_5)$	
x_4	$f(x_4)$	$f(x_3; x_4)$		$f(x_2; x_3; x_4; x_5)$	$f(x_1; x_2; x_3; x_4; x_5)$	
x_5	$f(x_5)$	$f(x_4; x_5)$				

1/3

x_i	$f(x_i)$	$f(x_i; x_{i+1})$	$f(x_i; x_{i+1}; x_{i+2})$	$f(x_i; x_{i+1}; x_{i+2}; x_{i+3})$	$f(x_i; x_{i+1}; x_{i+2}; x_{i+3}; x_{i+4})$
0	0	4			
2	8	19	5	1	
3	27	49	10	1	0
5	125	91	14		
6	216				

spełniającym warunek

$$W_n(x_i) = f(x_i) \quad i = 0, 1, 2, \dots, n$$

Niech

$$W_n(x) = W_0(x) + [W_1(x) - W_0(x)] + [W_2(x) - W_1(x)] + \dots + [W_n(x) - W_{n-1}(x)] \quad (1.36)$$

Oczywiście różnica $W_k(x) - W_{k-1}(x)$ jest wielomianem, który można przedstawić w postaci

$$W_k(x) - W_{k-1}(x) = A(x - x_0)(x - x_1) \dots (x - x_{k-1}) \quad (1.37)$$

gdzie A jest pewną stałą. Aby ją obliczyć, podstawiamy $x = x_k$. Mamy wówczas

$$W_k(x_k) - W_{k-1}(x_k) = A(x_k - x_0)(x_k - x_1) \dots (x_k - x_{k-1})$$

a ponieważ $W_k(x_k) = f(x_k)$, więc

$$A = \frac{f(x_k)}{(x_k - x_0)(x_k - x_1) \dots (x_k - x_{k-1})} - \frac{\sum_{i=0}^{k-1} f(x_i) \frac{\omega_{k-1}(x_k)}{(x_k - x_i)\omega'_{k-1}(x_i)}}{(x_k - x_0)(x_k - x_1) \dots (x_k - x_{k-1})} \quad (1.38)$$

i korzystając ze wzoru (1.15), otrzymujemy

$$A = \frac{f(x_k)}{\omega_{k-1}(x_k)} + \sum_{i=0}^{k-1} \frac{f(x_i)}{(x_k - x_i)\omega'_{k-1}(x_i)} = \sum_{i=0}^k \frac{f(x_i)}{\omega_k(x_i)} \quad (1.39)$$

Z (1.37) i (1.39) mamy

$$W_k(x) - W_{k-1}(x) = \left(\sum_{i=0}^k \frac{f(x_i)}{\omega_k(x_i)} \right) \omega_{k-1}(x) \quad (1.40)$$

Dla zapisania wzoru (1.40) w dogodniejszej postaci za pomocą ilorazów różnicowych, udowodnimy następującą zależność dotyczącą ilorazu rzędu n :

$$\begin{aligned} f(x_i; x_{i+1}; \dots; x_{i+n}) &= \frac{f(x_i)}{(x_i - x_{i+1})(x_i - x_{i+2}) \dots (x_i - x_{i+n})} + \\ &+ \frac{f(x_{i+1})}{(x_{i+1} - x_i)(x_{i+1} - x_{i+2}) \dots (x_{i+1} - x_{i+n})} + \dots + \\ &+ \frac{f(x_{i+n})}{(x_{i+n} - x_i)(x_{i+n} - x_{i+1}) \dots (x_{i+n} - x_{i+n-1})} \end{aligned} \quad (1.41)$$

Wzór powyższy udowodnimy metodą indukcji.

DOWÓD. Dla $n = 1$ zależność jest spełniona, bowiem

$$f(x_i; x_{i+1}) = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = \frac{f(x_i)}{x_i - x_{i+1}} + \frac{f(x_{i+1})}{x_{i+1} - x_i}$$

Wykażemy teraz, że jeśli wzór (1.41) jest prawdziwy dla $n = k - 1$, to jest prawdziwy dla $n = k$. Mamy

$$\begin{aligned}
 f(x_i; x_{i+1}; x_{i+2}; \dots; x_{i+k}) &= \frac{f(x_{i+1}; x_{i+2}; \dots; x_{i+k}) - f(x_i; x_{i+1}; \dots; x_{i+k-1})}{x_{i+k} - x_i} = \\
 &= \frac{1}{x_{i+k} - x_i} \left\{ \frac{f(x_{i+1})}{(x_{i+1} - x_{i+2})(x_{i+1} - x_{i+3}) \dots (x_{i+1} - x_{i+k})} + \right. \\
 &\quad + \frac{f(x_{i+2})}{(x_{i+2} - x_{i+1})(x_{i+2} - x_{i+3}) \dots (x_{i+2} - x_{i+k})} + \\
 &\quad + \dots + \frac{f(x_{i+k})}{(x_{i+k} - x_{i+1})(x_{i+k} - x_{i+2}) \dots (x_{i+k} - x_{i+k-1})} - \\
 &\quad - \left[\frac{f(x_i)}{(x_i - x_{i+1})(x_i - x_{i+2}) \dots (x_i - x_{i+k-1})} + \right. \\
 &\quad + \frac{f(x_{i+1})}{(x_{i+1} - x_i)(x_{i+1} - x_{i+2}) \dots (x_{i+1} - x_{i+k-1})} + \\
 &\quad \left. \left. + \dots + \frac{f(x_{i+k-1})}{(x_{i+k-1} - x_i)(x_{i+k-1} - x_{i+1}) \dots (x_{i+k-1} - x_{i+k-2})} \right] \right\}
 \end{aligned}$$

W rozwinięciu tym składniki zawierające $f(x_i)$ oraz $f(x_{i+k})$ występują dokładnie jeden raz, natomiast składniki zawierające $f(x_j)$ (dla $j = i + 1, i + 2, \dots, i + k - 1$) występują po dwa razy. Grupując zatem te pary, otrzymujemy

$$\begin{aligned}
 &\frac{1}{x_{i+k} - x_i} \left[\frac{f(x_j)}{(x_j - x_{i+1})(x_j - x_{i+2}) \dots (x_j - x_{j-1})(x_j - x_{i+1}) \dots (x_j - x_{j+k})} - \right. \\
 &\quad \left. - \frac{f(x_j)}{(x_j - x_i)(x_j - x_{i+1}) \dots (x_j - x_{j-1})(x_j - x_{j+1}) \dots (x_j - x_{j+k-1})} \right] = \\
 &= \frac{f(x_j)}{(x_j - x_{i+1}) \dots (x_j - x_{j-1})(x_j - x_{j+1}) \dots (x_j - x_{j+k-1})} \frac{1}{x_{i+k} - x_i} \left[\frac{1}{x_j - x_{i+k}} - \frac{1}{x_j - x_i} \right] = \\
 &= \frac{f(x_j)}{(x_j - x_i)(x_j - x_{i+1}) \dots (x_j - x_{j-1})(x_j - x_{j+1}) \dots (x_j - x_{i+k})}
 \end{aligned}$$

dla $j = i + 1, i + 2, \dots, i + k - 1$. Ponieważ pozostały jeszcze dwa wyrazy

$$\frac{f(x_i)}{(x_i - x_{i+1})(x_i - x_{i+2}) \dots (x_i - x_{i+k})}$$

oraz

$$\frac{f(x_{i+k})}{(x_{i+k} - x_i)(x_{i+k} - x_{i+1}) \dots (x_{i+k} - x_{i+k-1})}$$

zatem otrzymujemy

$$f(x_i; x_{i+1}; \dots; x_{i+k}) = \sum_{j=i}^{j=i+k} \frac{f(x_j)}{(x_j - x_i)(x_j - x_{i+1}) \dots (x_j - x_{j-1})(x_j - x_{j+1}) \dots (x_j - x_{i+k})}$$

end.

Do dalszych przekształceń należy przyjąć $i = 0$, wówczas

$$f(x_0; x_1; \dots; x_n) = \sum_{j=0}^{j=n} \frac{f(x_j)}{(x_j - x_0)(x_j - x_1) \dots (x_j - x_{j-1})(x_j - x_{j+1}) \dots (x_j - x_n)}$$

dla $n = 1, 2, \dots$

i stąd

$$W_n(x) = f(x_0) + f(x_0; x_1) \omega_0(x) + f(x_0; x_1; x_2) \omega_1(x) + \dots + f(x_0; x_1; \dots; x_n) \omega_{n-1}(x) \quad (1.42)$$

Wzór interpolacyjny (1.42) nosi nazwę *wzoru interpolacyjnego Newtona dla nierównych odstępów argumentów* albo *wzoru interpolacyjnego Newtona z ilorazami różnicowymi*.

Przykład 1. Napisać wzór interpolacyjny dla funkcji $f(x)$ określonej tablicą wartości: $f(0) = 1, f(2) = 3, f(3) = 2, f(4) = 5, f(6) = 7$.

Sporządzamy tablicę ilorazów różnicowych dla tej funkcji (tabl. 1/4) i na podstawie (1.15), (1.42), otrzymujemy

$$W_4(x) = 1 + 1(x-0) - \frac{2}{3}(x-0)(x-2) + \frac{2}{3}(x-0)(x-2)(x-3) - \frac{2}{9}(x-0)(x-2)(x-3)(x-4) = -\frac{2}{9}x^4 + \frac{8}{3}x^3 - \frac{88}{9}x^2 + \frac{35}{3}x + 1$$

Przykład 2. Napisać w Turbo Pascalu procedurę obliczania ilorazów różnicowych funkcji $f(x)$, podanych w tablicy 1/2, potrzebnych do obliczeń przy posługiwaniu się wzorem interpolacyjnym Newtona z ilorazami różnicowymi.

1/4

x_i	$f(x_i)$	$f(x_i; x_{i+1})$	$f(x_i; x_{i+1}; x_{i+2})$	$f(x_i; x_{i+1}; x_{i+2}; x_{i+3})$	$f(x_i; x_{i+1}; x_{i+2}; x_{i+3}; x_{i+4})$
0	1	1			
2	3		$-\frac{2}{3}$		
		-1		$\frac{2}{3}$	
3	2		2		$-\frac{2}{9}$
		3		$-\frac{2}{3}$	
4	5		$-\frac{2}{3}$		
		1			
6	7				

Jak łatwo zauważyć, we wzorze (1.42) wykorzystuje się jedynie ilorazy różnicowe położone na przekątnej od $f(x_0)$ do $f(x_0; x_1; \dots; x_n)$. Procedura ILORAZY wymaga umieszczenia wartości argumentów (węzłów) w tablicy X o wymiarach $1 \times M$, gdzie $M = n + 1$, i odpowiadających im wartości funkcji w analogicznej tablicy W . Ponieważ przy korzystaniu ze wzoru (1.42) potrzeba nam tylko części ilorazów różnicowych, konstruowanie pełnej tablicy 1/2 byłoby nieekonomiczne, choćby ze względu na wykorzystanie pamięci maszyny; dla $n = 100$ wymagałoby to zbędnej rezerwacji pamięci około 10 tysięcy miejsc. Procedurę naszą napiszemy w taki sposób, aby po jej wykonaniu tablica W zawierała jedynie potrzebne ilorazy różnicowe w następującej kolejności

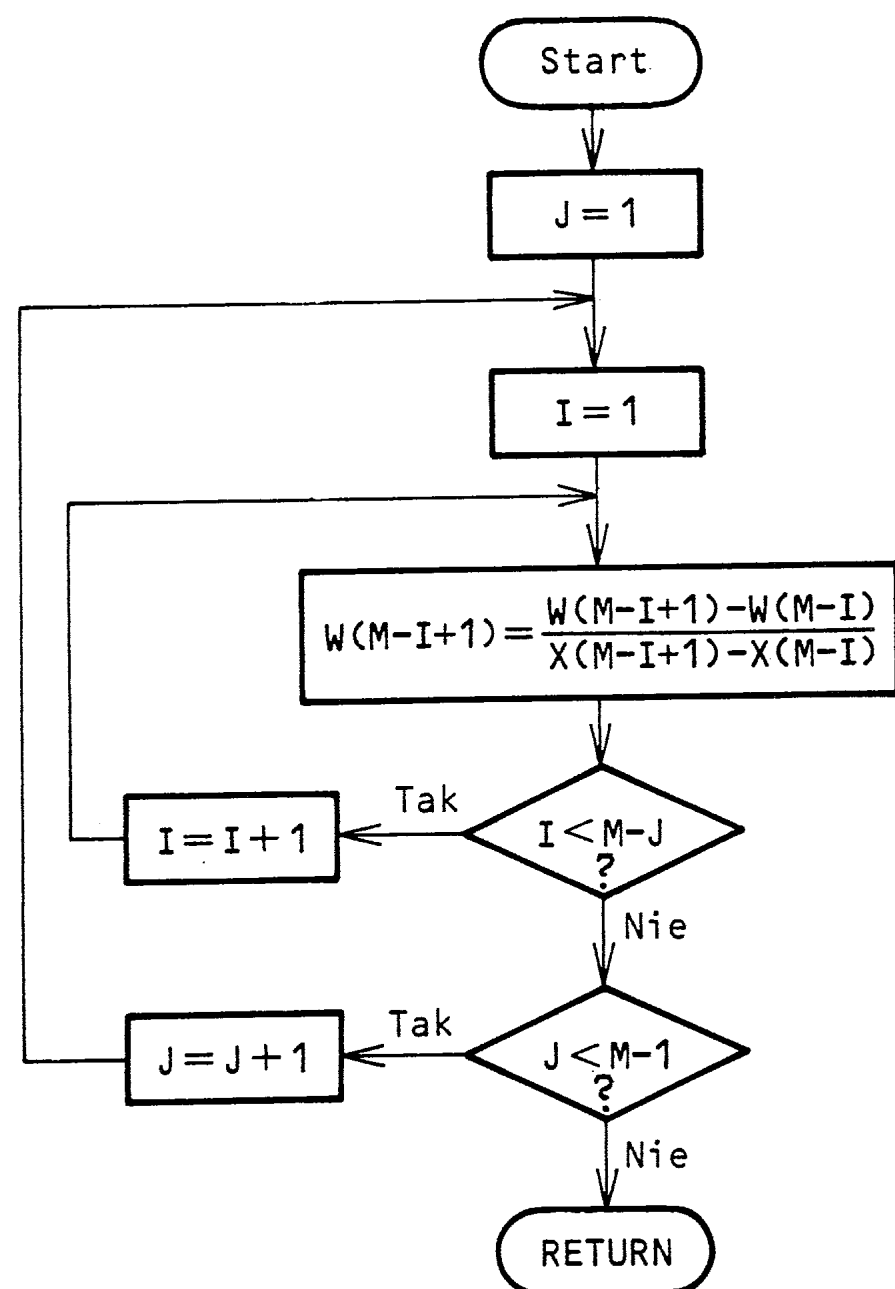
$$W(1) = f(x_0)$$

$$W(2) = f(x_0; x_1)$$

.....

$$W(M) = f(x_0; x_1; \dots; x_n)$$

Schemat blokowy procedury jest pokazany na rys. 1.5.



1.5

Przyjmując, że w programie są wprowadzone deklaracje typów i zmiennych identyczne jak w przykładach w p. 1.2.1, otrzymujemy następującą treść podprogramu:

```

PROCEDURE ILORAZY (VAR X,W: wektor; M: zakres);
VAR I,J : integer;
BEGIN
  FOR J := 1 TO M-1 DO
    FOR I := 1 TO M-J DO
      W[M-I+1] := (W[M-I+1] - W[M-I]) / (X[M-I+1] - X[M-I])
    END;
  END;

```

Ćwiczenia

1. Udowodnić, że jeżeli $f(x) = (x - x_0)(x - x_1) \dots (x - x_n)$, to $f(x_0; x_1; \dots; x_p) = 0$ dla $p < n$.
2. Za pomocą wzoru (1.42) znaleźć wielomian interpolacyjny dla funkcji podanej w ćwiczeniu 2 z p. 1.2.3.

3. Korzystając z podanej w przykładzie 2 procedury obliczania ilorazów różnicowych funkcji $f(x)$ napisać procedurę obliczania wartości wielomianu interpolacyjnego Newtona z ilorazami różnicowymi według wzoru (1.42) w dowolnym punkcie x (różnym od węzła).

Wskazówka. Szukany algorytm obliczeniowy jest bardzo podobny do algorytmu Hornera obliczania wartości wielomianu algebraicznego stopnia n uporządkowanego według potęg.

1.2.5. Różnice progresywne i różnice wsteczne

Na wartości zmiennej niezależnej nie nakładaliśmy dotychczas żadnych warunków. Rozpatrywane wartości argumentów $x_0, x_1, \dots, x_n$ (węzły interpolacji) mogły być wybierane zupełnie dowolnie. Obecnie przyjmujemy, że wartości te tworzą ciąg arytmetyczny i że przyrost argumentu jest wielkością stałą, tzn.

$$x_1 = x_0 + h, \quad x_2 = x_0 + 2h, \dots, x_n = x_0 + nh$$

Takie założenie bardzo upraszcza wzory interpolacyjne i ich wyprowadzenia.

Różnice progresywne. Niech będą dane wartości funkcji $f(x)$ w punktach $x_0, x_1, \dots, x_n$. Wyrażenie

$$\Delta f = f(x + h) - f(x) \quad (1.43)$$

nazywamy *różnicą progresywną rzędu pierwszego* funkcji f .

Podobnie definiujemy *różnice progresywne wyższych rzędów*

$$\Delta^n f = \Delta(\Delta^{n-1} f), \quad n = 2, 3, \dots \quad (1.44)$$

Na przykład

$$\begin{aligned} \Delta^2 f &= \Delta(\Delta f) = \Delta[f(x + h) - f(x)] = \\ &= [f(x + 2h) - f(x + h)] - [f(x + h) - f(x)] = \\ &= f(x + 2h) - 2f(x + h) + f(x) \end{aligned}$$

Wzór (1.44) można także zapisać w postaci

$$\begin{aligned} \Delta^0 f(x) &= f(x) \\ \Delta^n f(x) &= \Delta^{n-1} f(x + h) - \Delta^{n-1} f(x) \quad (1.44a) \\ n &= 1, 2, \dots \end{aligned}$$

Różnicę progresywną dowolnego rzędu funkcji f można więc przedstawić jako kombinację liniową wartości tej funkcji.

W przypadku wielomianów różnica rzędu pierwszego wielomianu rzędu n jest wielomianem stopnia $n - 1$.

Przykład 1. Obliczyć różnice progresywne wielomianu $W_4(x) = x^4 - x - 1$, przyjmując $h = 1$. Mamy

$$\Delta W = [(x + 1)^4 - (x + 1) - 1] - [x^4 - x - 1] = 4x^3 + 6x^2 + 4x$$

$$\Delta^2 W = [4(x+1)^3 + 6(x+1)^2 + 4(x+1)] - [4x^3 + 6x^2 + 4x] = 12x^2 + 24x + 14$$

$$\Delta^3 W = [12(x+1)^2 + 24(x+1) + 14] - [12x^2 + 24x + 14] = 24x + 36$$

$$\Delta^4 W = [24(x+1) + 36] - [24x + 36] = 24$$

$$\Delta^n W = 0 \quad \text{dla} \quad n \geq 5$$

Jak łatwo wykazać, w przypadku wielomianu n -tego stopnia n -ta różnica progresywna jest stała i ma postać

$$\Delta^n W(x) = n! a_0 h^n = \text{const}$$

Stosując interpolację funkcję $f(x)$ mamy określoną za pomocą tablicowanych wartości $y_i = f(x_i)$, a punkty $x_i (i = 0, 1, \dots, n)$ są rozmieszczone w równych odstępach. Skończone różnice progresywne ciągu y_i określa się następująco:

$$\Delta y_i = y_{i+1} - y_i$$

$$\Delta^2 y_i = \Delta(\Delta y_i) = \Delta y_{i+1} - \Delta y_i$$

$$\dots\dots\dots (1.45)$$

$$\Delta^n y_i = \Delta(\Delta^{n-1} y_i) = \Delta^{n-1} y_{i+1} - \Delta^{n-1} y_i$$

Można udowodnić (metodą indukcji), że

$$\begin{aligned} \Delta^n y_i &= \binom{n}{0} y_{i+n} - \binom{n}{1} y_{i+n-1} + \binom{n}{2} y_{i+n-2} - \dots - \\ &- (-1)^n \binom{n}{n-1} y_{i+1} + (-1)^n \binom{n}{n} y_i \end{aligned} \quad (1.46)$$

albo krócej

$$\Delta^n y_i = \sum_{j=0}^n (-1)^j \binom{n}{j} y_{i+n-j} \quad (1.46a)$$

Analogicznie dowodzi się, że każdą wartość $y_k (k = 1, 2, \dots, n)$ można obliczyć za pomocą różnic $\Delta y_0, \Delta^2 y_0, \dots, \Delta^k y_0$ oraz wartości y_0

$$y_k = y_0 + \binom{k}{1} \Delta y_0 + \binom{k}{2} \Delta^2 y_0 + \dots + \binom{k}{k} \Delta^k y_0 \quad (1.47)$$

$$k = 1, 2, \dots, n$$

Operator Δ ma wiele własności podobnych do własności operatora różniczkowego. Łatwo można sprawdzić za pomocą rachunku, że na przykład:

$$\Delta(g_k \pm f_k) = \Delta g_k \pm \Delta f_k$$

$$\Delta(a f_k) = a \Delta f_k$$

$$\Delta^m(\Delta^n f_k) = \Delta^{m+n} f_k$$

$$\Delta[f(x_k) \cdot g(x_k)] = f(x_{k+1}) \Delta g(x_k) + g(x_{k+1}) \Delta f(x_k)$$

$$\Delta \frac{f(x_k)}{g(x_k)} = \frac{\Delta f(x_k) g(x_k) - f(x_k) \Delta g(x_k)}{g(x_{k+1}) g(x_k)} \quad \text{dla} \quad g(x_{k+1}) g(x_k) \neq 0$$

$$\Delta c = 0 \quad (c = \text{const})$$

Obliczone różnice kolejnych rzędów najwygodniej jest układać w postaci trójkątnej tablicy różnic (tabl. 1/5).

Przykład 2. Sporządzić tablicę różnic progresywnych dla funkcji $y = x^3 - 1$ począwszy od wartości $x_0 = 0$, z krokiem 1, dla $n = 5$.

Korzystamy ze wzorów (1.45) i otrzymane wyniki zestawiamy w formie tablicy 1/6.

1/5

x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$	$\Delta^4 y_i$	$\Delta^5 y_i$
x_0	y_0	Δy_0				
x_1	y_1	Δy_1	$\Delta^2 y_0$	$\Delta^3 y_0$		
x_2	y_2	Δy_2	$\Delta^2 y_1$	$\Delta^3 y_1$	$\Delta^4 y_0$	
x_3	y_3	Δy_3	$\Delta^2 y_2$	$\Delta^3 y_2$	$\Delta^4 y_1$	$\Delta^5 y_0$
x_4	y_4	Δy_4	$\Delta^2 y_3$			
x_5	y_5					

1/6

x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$	$\Delta^4 y_i$	$\Delta^5 y_i$
0	-1	1				
1	0	7	6	6		
2	7	19	12	6	0	
3	26	37	18	6	0	0
4	63	61	24			
5	124					

Różnice wsteczne. Korzystając z przyjętych oznaczeń, *pierwszą różnicę wsteczną* funkcji $f(x)$ w punkcie x_i określamy jako pierwszą różnicę progresywną tej funkcji w punkcie x_{i-1} i oznaczamy symbolicznie: ∇y_i , $\nabla^1 y_i$, $\nabla f(x_i)$ albo $\nabla^1 f(x_i)$. Zatem

$$\nabla y_i = \Delta y_{i-1} = f(x_i) - f(x_{i-1}) \quad (1.48)$$

dla $i = 1, 2, \dots, n$

Ogólnie, *k-tą różnicę wsteczną* funkcji $f(x)$ w punkcie x_i określamy jako różnicę różnic $(k-1)$ -szych w punktach x_i oraz x_{i-1} , a więc

$$\nabla^k y_i = \nabla^{k-1} y_i - \nabla^{k-1} y_{i-1} \quad (1.49)$$

$k = 1, 2, \dots, n$

$i = k, k+1, \dots, n$

przy czym $\nabla^0 y_i = y_i$.

Można bez trudności wykazać, że

$$\nabla^k y_i = \Delta^k y_{i-k} \quad k = 0, 1, \dots, n \quad (1.50)$$

Tablica różnic wstecznych jest identyczna z tablicą różnic progresywnych — zmianie ulegają jedynie symbole różnic. Problem, czy korzystać z różnic progresywnych, czy też z różnic wstecznych jest uzależniony najczęściej od tego, w jakiej części przedziału $(x_0; x_n)$ chcemy badać funkcję $f(x)$. Jeżeli interesuje nas początek przedziału (otoczenie punktu x_0) korzystamy z różnic progresywnych i rozbudowujemy tablicę (analogicznie do tabl. 1/5) w kierunku rosnących wskaźników $x_i (x_0, x_1, \dots)$ tak daleko, jak to jest potrzebne. Jeśli natomiast interesuje nas zachowanie się funkcji $f(x)$ na otoczeniu końca przedziału, to korzystamy z różnic wstecznych i tablicę rozbudowujemy w kierunku malejących wskaźników $x_i (x_n, x_{n-1}, \dots)$. W tym drugim przypadku punkty, w których jest określona funkcja $f(x) (x_0, x_1, \dots, x_n)$ wygodnie jest oznaczać symbolami: $x_{-n}, x_{-n+1}, \dots, x_{-1}, x_0$. Sposób sporządzania tablicy różnic wstecznych pokazuje tablica 1/7.

1/7

x_i	y_i	∇y_i	$\nabla^2 y_i$	$\nabla^3 y_i$	$\nabla^4 y_i$	$\nabla^5 y_i$
x_{-6}	y_{-6}					
x_{-5}	y_{-5}	∇y_{-5}				
x_{-4}	y_{-4}	∇y_{-4}	$\nabla^2 y_{-4}$			
x_{-3}	y_{-3}	∇y_{-3}	$\nabla^2 y_{-3}$	$\nabla^3 y_{-3}$		
x_{-2}	y_{-2}	∇y_{-2}	$\nabla^2 y_{-2}$	$\nabla^3 y_{-2}$	$\nabla^4 y_{-2}$	
x_{-1}	y_{-1}	∇y_{-1}	$\nabla^2 y_{-1}$	$\nabla^3 y_{-1}$	$\nabla^4 y_{-1}$	$\nabla^5 y_{-1}$
x_0	y_0	∇y_0	$\nabla^2 y_0$	$\nabla^3 y_0$	$\nabla^4 y_0$	$\nabla^5 y_0$

Ćwiczenia

1. Wiadomo, że dla wielomianów (ogólniej, także dla funkcji klasy C^n) i dostatecznie małych przyrostów h różnice skończone danej funkcji, aż do rzędu n włącznie, zmieniają się równomiernie, a różnice rzędu n są ponadto stałe. Gdy w pewnym miejscu tablicy nastąpi zaburzenie regularności różnic, istnieje duże prawdopodobieństwo, że w tablicy powstała omyłka. Błąd ε wchodzi do kolejnych różnic $\Delta^k y$ (rzędu k) ze współczynnikami dwumianowymi o naprzemiennych znakach: $\binom{k}{0} \varepsilon, -\binom{k}{1} \varepsilon, \dots$. Wiedząc, że jedna z wartości funkcji w tablicy 1/8 jest zaburzona, znaleźć ten błąd i poprawić go.

W s k a z ó w k a. Należy skonstruować pomocniczą tablicę, w której wszystkie wartości funkcji $f(x)$ są zerami z wyjątkiem jednej, równej ε , a następnie należy zlokalizować nieregularność (np. zaburzenie w monotoniczności) w zachowaniu się różnic (w tym przypadku — różnic drugiego rzędu). Próba „porównania” danej tablicy różnic i tablicy pomocniczej ułatwia lokalizację omyłki.

2. Znaleźć omyłkę w tablicy 1/9 i poprawić odpowiednie różnice.

Odpowiedzi. 1. $y_5 = 51$. 2. Błędna jest wartość funkcji $f(1,4)$; powinno być $f(1,4) = 20,46$.

1/8

x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$	$\Delta^4 y_i$	$\Delta^5 y_i$	$\Delta^6 y_i$
1	3						
		6					
2	9		4				
		10		0			
3	19		4		-1		
		14		-1		5	
4	33		3		4		-15
		17		3		-10	
5	50		6		-6		20
		23		-3		10	
6	73		3		4		
		26		1			
7	99		4				
		30					
8	129						

1/9

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$	$\Delta^5 f$
1	17.500					
		-0.635				
1.1	18.135		0.07			
		-0.705		0		
1.2	18.840		0.07		-0.2	
		-0.775		0.2		-1.0
1.3	19.615		-0.13		0.8	
		-0.645		-0.6		2.0
1.4	20.260		0.47		-1.2	
		-1.115		0.6		-2.0
1.5	21.375		-0.13		0.8	
		-0.985		-0.2		1.0
1.6	22.360		0.07		-0.2	
		-1.055		0		
1.7	23.415		0.07			
		-1.125				
1.8	24.540					

1.2.6. Wzory interpolacyjne Newtona dla równoodległych wartości argumentu

Wzory interpolacyjne Newtona są stosowane do rozwiązywania tego samego zagadnienia interpolacyjnego co i wzór Lagrange’a. Niech będą dane wartości funkcji $f(x)$

$y_i = f(x_i), \quad i = 0, 1, \dots, n$

przy czym punkty x_i są rozmieszczone w jednakowych odstępach

$x_i = x_0 + ih, \quad i = 0, 1, 2, \dots, n; \quad h = \text{const}$

W punkcie 1.1.1 udowodniliśmy twierdzenie, że istnieje dokładnie jeden wielomian $W_n(x)$ stopnia nie większego niż n taki, że $W_n(x_i) = y_i$, $i = 0, 1, 2, \dots, n$. Obecnie zajmujemy się wyznaczeniem innej postaci wielomianu interpolacyjnego, który będzie oczywiście tożsamościowo równy wielomianowi otrzymanemu za pomocą wzoru interpolacyjnego Lagrange'a (1.14). W punkcie 1.2.4 podaliśmy wzór interpolacyjny Newtona przy założeniu dowolnego rozmieszczenia węzłów interpolacyjnych postaci

$$W_n(x) = f(x_0) + f(x_0; x_1) \omega_0(x) + f(x_0; x_1; x_2) \omega_1(x) + \dots + f(x_0; \dots; x_n) \omega_{n-1}(x) \quad (1.42)$$

gdzie $\omega_k(x) = (x - x_0)(x - x_1) \dots (x - x_k)$.

W przypadku węzłów równoległych ilorazy różnicowe kolejnych rzędów wynoszą odpowiednio

$$f(x_0) = a_0 = y_0$$

$$f(x_0; x_1) = a_1 = \frac{y_1 - y_0}{h}$$

$$f(x_0; x_1; x_2) = a_2 = \frac{\frac{y_2 - y_1}{h} - \frac{y_1 - y_0}{h}}{2h} = \frac{y_2 - 2y_1 + y_0}{2h^2} \quad (1.51)$$

.....

$$f(x_0; \dots; x_n) = \frac{y_n - ny_{n-1} + \binom{n}{2} y_{n-2} - \dots + (-1)^n y_0}{n! h^n}$$

Korzystając ze wzoru (1.45) na różnice progresywne rzędu n , otrzymujemy

$$a_0 = y_0, \quad a_1 = \Delta y_0 / h, \quad a_2 = \Delta^2 y_0 / 2h^2$$

ogólnie

$$a_k = \frac{\Delta^k y_0}{k! \cdot h^k}, \quad k = 0, 1, 2, \dots, n \quad (1.52)$$

przy czym $0! = 1$, a $\Delta^0 y_0 = y_0$. Podstawiając wyznaczone wartości współczynników do wzoru (1.42), otrzymujemy

$$W_n(x) = y_0 + \frac{\Delta y_0}{h}(x - x_0) + \frac{\Delta^2 y_0}{2! h^2}(x - x_0)(x - x_1) + \dots \\ \dots + \frac{\Delta^n y_0}{n! h^n}(x - x_0)(x - x_1) \dots (x - x_{n-1}) \quad (1.53)$$

Wzór (1.53) nosi nazwę *pierwszego wzoru interpolacyjnego Newtona*. Jest

to jak widać wielomian stopnia nie większego niż n , a ponadto

$$W_n(x_k) = y_k \quad \text{dla} \quad k = 0, 1, \dots, n$$

We wzorze (1.53) współczynniki a_i są kolejnymi różnicami progresywnymi podzielonymi przez odpowiednie silnie. Ponieważ różnice ze wzrostem ich rzędu zazwyczaj dość szybko maleją, więc bardzo często przy korzystaniu z tego wzoru nie uwzględnia się tych składników, których współczynniki są dostatecznie małe. Widać tutaj korzyści, jakie przynosi stosowanie wzoru (1.53) w porównaniu ze wzorem Lagrange'a (1.14), w którym każdy ze składników jest jednakowo istotny i nie można żadnego z nich zaniedbać.

W praktycznych zastosowaniach znacznie łatwiej jest stosować wielomian interpolacyjny Newtona (1.53), wprowadzając nową zmienną

$$q = \frac{x - x_0}{h}$$

Ponieważ

$$\frac{x - x_1}{h} = \frac{x - (x_0 + h)}{h} = q - 1$$

$$\frac{x - x_2}{h} = \frac{x - (x_0 + 2h)}{h} = q - 2$$

.....

$$\frac{x - x_{n-1}}{h} = \frac{x - [x_0 + (n-1)h]}{h} = q - (n-1)$$

więc wzór (1.53) można sprowadzić do postaci

$$\begin{aligned} W_n(x) = W_n(x_0 + qh) = & y_0 + \frac{q}{1!} \Delta y_0 + \frac{q(q-1)}{2!} \Delta^2 y_0 + \\ & + \dots + \frac{q(q-1) \dots (q-n+1)}{n!} \Delta^n y_0 \end{aligned} \quad (1.54)$$

Wielkość $q = (x - x_0)/h$ określa liczbę kroków potrzebnych do przejścia od punktu x_0 do punktu x . Wzór (1.54) wygodnie jest stosować do interpolacji funkcji $y = f(x)$ na otoczeniu wartości początkowej x_0 (dla $q < 1$). Przechodząc do przedziału $x_1 < x < x_2$, w którym q jest większe od jedności, wygodniej jest przyjąć za x_0 następny węzeł interpolacji, czyli x_1 .

Jeżeli posługujemy się trójkątną tablicą różnic, to do wzoru (1.54) wchodzi różnice „schodzące” w dół po odpowiedniej przekątnej. Dlatego też stosowanie tego wzoru wygodne jest dla górnej, początkowej części tablicy, gdzie mamy dostateczną liczbę różnic. Dla dolnej jej części wzór ten okazuje się mało przydatny lub wręcz nieprzydatny. Z tych względów wzór (1.54) nazywany jest także *wzorem interpolacyjnym Newtona na interpolację w przód*.

Przykład 1. Funkcja $f(x) = e^x$ dana jest na przedziale $\langle 3,5; 3,8 \rangle$ za pomocą tablicy wartości 1/10. Obliczyć: a) wartości funkcji w przedziale $\langle 3,50; 3,55 \rangle$ z krokiem $h = 0,01$, b) wartość funkcji w punkcie $x = 3,58$.

a) Sporządzamy tablicę różnic dla danych wartości funkcji (tabl. 1/11). Aby zbudować tablicę wartości w przedziale $\langle 3,50; 3,55 \rangle$ z krokiem $h = 0,01$, przyjmujemy $x_0 = 3,50$. Odpowiednie wartości $\Delta y_0, \Delta^1 y_0$ itd. leżą na najwyższej przekątnej tablicy 1/11. Ponieważ $qh = 0,05$, więc dla kolejnych wartości $x_i \in \langle 3,50; 3,55 \rangle$ odpowiednie wartości q_i wynoszą odpowiednio: dla $x_0 = 3,50, q_0 = 0$, dla $x_1 = 3,51, q_1 = 0,2$ itd. **U w a g a.** Aby wyznaczyć szukane wartości y dla wybranych q wygodnie jest zestawić pomocniczą tablicę 1/12.

b) W celu znalezienia wartości funkcji w punkcie $x = 3,58$ budujemy wielomian interpolacyjny dla tej funkcji. Ponieważ $x \in \langle 3,55; 3,60 \rangle$, jako x_0 należy przyjąć 3,55. Wtedy (patrz tabl. 1/11)

1/10

x	3.50	3.55	3.60	3.65	3.70	3.75	3.80
y	33.115	34.813	36.598	38.475	40.447	42.521	44.701

1/11

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	$\Delta^5 y$	$\Delta^6 y$
3.50	33.115	1.698					
3.55	34.813	1.785	0.087	0.005			
3.60	36.598	1.877	0.092	0.003	−0.002	0.006	
3.65	38.475	1.972	0.095	0.007	0.004	−0.007	−0.013
3.70	40.447	2.074	0.102	0.004	−0.003		
3.75	42.521	2.180	0.106				
3.80	44.701						

$y_0 = 34,813, \quad \Delta y_0 = \Delta y_1 = 1,785$ itd.

czyli

$$W_5(x) = 34,813 + q \cdot 1,785 + q(q - 1) \cdot 0,046 + q(q - 1)(q - 2) \cdot 0,0005 +$$

$$+ q(q - 1)(q - 2)(q - 3) \cdot 16,6 \cdot 10^{-5} + q(q - 1)(q - 2)(q - 3)(q - 4) \cdot 5,8 \cdot 10^{-5}$$

gdzie

$$q = \frac{x - 3,55}{0,05} = 20(x - 3,55)$$

Dla $x = 3,58, q = 20 \cdot 0,03 = 0,6$ i szukana wartość

$$f(3,58) = 35,865794$$

1/12

q	x	$q\Delta y_0$	$\frac{q(q-1)}{2}\Delta^2y_0$	$\frac{q(q-1)(q-2)}{3!}\Delta^3y_0$	$\frac{q(q-1)(q-2)(q-3)}{4!}\Delta^4y_0$	$\frac{q(q-1)(q-2)(q-3)(q-4)}{5!}\Delta^5y_0$	$\frac{q(q-1)(q-2)(q-3)(q-4)(q-5)}{6!}\Delta^6y_0$	$y = y_0 + (3) +$ $+ (4) + (5) +$ $+ (6) + (7) + (3)$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0	3.50	0	0	0	0	0	0	33.115
0.2	3.51	0.3396	-0.00696	0.00024	$0.672 \cdot 10^{-4}$	$0.15322 \cdot 10^{-3}$	0.26557	33.448364
0.4	3.52	0.6792	-0.01044	0.00032	$0.832 \cdot 10^{-4}$	$0.17971 \cdot 10^{-3}$	0.29852	33.784641
0.6	3.53	1.0188	-0.01044	0.00028	$0.672 \cdot 10^{-4}$	$0.13709 \cdot 10^{-3}$	0.21782	34.124062
0.8	3.54	1.3584	-0.00696	0.00016	$0.352 \cdot 10^{-4}$	$0.06758 \cdot 10^{-3}$	0.10250	34.466805
1.0	3.55	0	0	0	0	0	0	34.813

Jak już mówiliśmy, pierwszy wzór interpolacyjny Newtona jest nie-
dogodny do interpolacji w pobliżu końca tablicy. Stosuje się tu inny wzór,
który obecnie wyprowadzimy.

Przy niezmiennych założeniach szukamy wielomianu interpolacyjnego
zapisanego w postaci

$$W_n(x) = a_0 + a_1(x - x_n) + a_2(x - x_n)(x - x_{n-1}) + \dots + \\ + a_n(x - x_n)(x - x_{n-1}) \dots (x - x_1) \quad (1.55)$$

Współczynniki a_i , jak poprzednio, wyznaczamy ze wzorów na ilorazy
różnicowe

$$a_k = \frac{\Delta^k y_{n-k}}{k! h^k}, \quad k = 0, 1, 2, \dots, n \quad (1.56)$$

i korzystając z wprowadzonego pojęcia różnic wstecznych (zob. wzór (1.49)),
mamy

$$a_k = \frac{\nabla^k y_n}{k! h^k}, \quad k = 0, 1, 2, \dots, n \quad (1.57)$$

Podstawiając (1.56) do (1.55), otrzymujemy

$$W_n(x) = y_n + \frac{\Delta y_{n-1}}{1! h} (x - x_n) + \frac{\Delta^2 y_{n-2}}{2! h^2} (x - x_n)(x - x_{n-1}) + \dots + \\ + \frac{\Delta^n y_0}{n! h^n} (x - x_n)(x - x_{n-1}) \dots (x - x_1) \quad (1.58)$$

Wzór (1.58) nazywamy *drugim wzorem interpolacyjnym Newtona*.

Dla celów praktycznych wygodnie jest jak poprzednio wprowadzić
podstawienie

$$x = x_n + qh$$

i zapisać (1.58) w postaci

$$W_n(x) = y_n + q \Delta y_{n-1} + \frac{q(q+1)}{2!} \Delta^2 y_{n-2} + \dots + \\ + \frac{q(q+1)(q+2) \dots (q+n-1)}{n!} \Delta^n y_0 \quad (1.59)$$

Skorzystamy teraz ze wzorów (1.57) określających współczynniki a_k za
pomocą różnic wstecznych, a ponadto ponieważ z różnic wstecznych
korzystamy wtedy, gdy interesuje nas zachowanie się funkcji na końcowej
części przedziału, zmienimy numerację punktów x_i pisząc x_{i-n} zamiast
 x_i ($i = 0, 1, 2, \dots, n$). Wówczas

$$x_i = x_0 + ih, \quad i = 0, -1, -2, \dots, -n$$

i wzory (1.58), (1.59) przejdą odpowiednio we wzory

$$W_n(x) = y_0 + \frac{\nabla y_0}{h}(x - x_0) + \frac{\nabla^2 y_0}{2!h^2}(x - x_0)(x - x_{-1}) + \dots + \frac{\nabla^n y_0}{n!h^n}(x - x_0)(x - x_{-1}) \dots (x - x_{-n+1}) \quad (1.60)$$

w których punktom $x_{-n}, x_{-n+1}, \dots, x_{-1}, x_0$ odpowiadają punkty $x_0, x_1, \dots, x_{n-1}, x_n$. Po podstawieniu do (1.60)

$$q = \frac{x_0 - x}{h}$$

mamy

$$W_n(x) = y_0 - q\nabla y_0 + \frac{q(q-1)}{2}\nabla^2 y_0 - \dots + (-1)^n \frac{q(q-1)(q-2)\dots(q-n+1)}{n!}\nabla^n y_0 \quad (1.61)$$

Wzór (1.61) nazywa się często *wzorem interpolacyjnym Newtona na interpolację wstecz*.

Przykład 2. Korzystając z tablicy wartości funkcji $x = e^x$, podanej w poprzednim przykładzie, obliczyć wartość funkcji w punkcie $x = 3,76$.

Przyjmujemy $x_0 = 3,80$. Wówczas $y_0 = 44,701$ i odpowiednio $\nabla y_0 = 2,18$, $\nabla^2 y_0 = 0,106$, $\nabla^3 y_0 = 0,004$, $\nabla^4 y_0 = -0,003$, $\nabla^5 y_0 = -0,005$. Dla $x = 3,76$ mamy dodatnią wartość $q = 0,8$ i ze wzoru (1.61) otrzymujemy

$$\begin{aligned} W(3,76) &= 44,701 - q\nabla y_0 + \frac{q(q-1)}{2}\nabla^2 y_0 - \\ &\quad - \frac{q(q-1)(q-2)}{6}\nabla^3 y_0 + \frac{q(q-1)(q-2)(q-3)}{24}\nabla^4 y_0 - \\ &\quad - \frac{q(q-1)(q-2)(q-3)(q-4)}{120}\nabla^5 y_0 = 42,9485 \end{aligned}$$

Przykład 3. Niech $f(x) = \ln x$. Korzystając z tablicy wartości funkcji 1/13 obliczyć wartość przybliżoną $\ln 0,75$. Pokazać, jak zmienia się wynik przy zmianie liczby użytych różnic.

Budujemy tablicę różnic wstecznych (tabl. 1/14). Stąd mamy

$$\begin{aligned} W(x) &= y_0 - q\nabla y_0 + \frac{q(q-1)}{2}\nabla^2 y_0 - \frac{q(q-1)(q-2)}{6}\nabla^3 y_0 + \\ &\quad + \dots + \frac{q(q-1)(q-2)(q-3)(q-4)(q-5)}{620}\nabla^6 y_0 \end{aligned}$$

$$x_0 = 0,9, \quad q = \frac{x_0 - x}{h} = 1,5, \quad y_0 = 0,117783 \text{ itd.}$$

1/13

x	0.30	0.40	0.50	0.60	0.70	0.80	0.90
$\ln x$	−1.203973	−0.916291	−0.693147	−0.520826	−0.356675	−0.223144	−0.105361

1/14

x_i	y_i	∇y_i	$\nabla^2 y_i$	$\nabla^3 y_i$	$\nabla^4 y_i$	$\nabla^5 y_i$	$\nabla^6 y_i$
0.3	−1.203973	0.287682	−0.064538	0.023715	−0.011062	0.005959	−0.003534
0.4	−0.916291						
0.5	−0.693147	0.182321	−0.040823	0.012653	−0.005103	0.002425	
0.6	−0.510826	0.154151	−0.02817	0.00755			
0.7	−0.356675	0.133531	−0.02062	0.004872	−0.002678		
0.8	−0.223144	0.117783	−0.015748				
0.9	−0.105361						

1/15

Liczba użytych różnic	Wynik
0	−0.105361
1	−0.282035
2	−0.287940
3	−0.287635
4	−0.287697
5	−0.287668
6	−0.287696

W tablicy 1/15 są zestawione różne wyniki w zależności od liczby użytych różnic. Dokładna wartość $\ln 0,75 = -0,28768207$.

Ćwiczenia

1. Udowodnić wzór (1.50).
2. Wykazać, że pierwsza różnica (progresywna, wsteczna) wielomianu stopnia n jest wielomianem stopnia $(n - 1)$.
3. Udowodnić wzór (1.52).
4. Dana jest tablica wartości funkcji $y = \sin x$ dla $x \in \langle 15^\circ; 50^\circ \rangle$ z krokiem 5° (tabl. 1/16). Obliczyć wartości $\sin x$ dla kątów od 10° do 15° , co 1° .
5. Stosując wzór interpolacyjny Newtona znaleźć sumę kwadratów

$S_n = 1^2 + 2^2 + 3^2 + \dots + n^2$

liczb naturalnych od 1 do n .

1/16

x	15°	20°	25°	30°	35°	40°	45°	50°
$\sin x$	0.258819	0.342020	0.422618	0.5	0.573576	0.642788	0.707107	0.766044

1/17

x	$f(x)$
10.5	100
11.0	110
11.5	121
12.0	134
12.5	150

6. Dana jest tablica wartości funkcji $f(x)$ (tabl. 1/17). Obliczyć wartości $f(x)$ w przedziale $\langle 11,0; 11,5 \rangle$ z krokiem $h = 0,1$.

7. Wielomianem czynnikiem stopnia n nazywamy iloczyn n czynników, z których pierwszy równa się x , a każdy następny jest mniejszy od poprzedniego o stałą liczbę h

$$x_{df}^{[n]} = x(x-h)(x-2h)(x-3h)\dots[x-(n-1)h]$$

a) Udowodnić, że różnica skończona rzędu pierwszego wielomianu czynnika

$$\Delta x^{[n]} = nhx^{[n-1]}$$

gdzie $\Delta x = h$.

b) Udowodnić, że różnica skończona rzędu k

$$\Delta^k x^{[n]} = n(n-1)\dots[n-(k-1)]h^k x^{[n-k]}, \quad k = 1, 2, \dots, n$$

c) Zapisać pierwszy wzór interpolacyjny Newtona (1.54) za pomocą wielomianów czynnika.

d) Zapisać drugi wzór interpolacyjny Newtona (1.61) za pomocą wielomianów czynnika.

8. Wychodząc ze wzoru (1.51) wyprowadzić pierwszy wzór interpolacyjny Newtona bez korzystania z określenia ilorazów różnicowych.

9. a) Operator przesunięcia E definiujemy następująco: $EL(x) = L(x+h)$, $E^\alpha L(x) = L(x+\alpha h)$, $\alpha \in \mathbf{R}$. Udowodnić, że $E^2 L(x) = EEL(x)$, $E^p E^q L(x) = E^{p+q} L(x)$.

b) Ponieważ $EL(x_k) = Ey_k = y_{k+1} + \Delta y_k$ ($k = 0, 1, 2, \dots, n-1$), więc jeżeli $E = I + \Delta$ (czyli $Ey_k - (I + \Delta)y_k$), gdzie I jest operatorem identyczności ($Iy_k = y_k$), to $E^\alpha = (I + \Delta)^\alpha$. Udowodnić wzór

$$L(x) = (I + \Delta)^\alpha y_0, \quad \alpha = \frac{x - x_0}{h}$$

10. Wyznaczyć związek rekurencyjny między współczynnikami: a) wielomianu (1.54), b) wielomianu (1.61).

11. Napisać procedurę obliczania elementów tablicy różnic.

12. Napisać procedurę obliczania w dowolnym punkcie x wartości wielomianów (1.54) i (1.61).

Odpowiedzi. 4. Tablica 1/18. 5. $S_n = n(n+1)(2n+1)/6$. 6. Tablica 1/19.

1/18

x	10°	11°	12°	13°	14°
$\sin x$	0.173648	0.190809	0.207912	0.224951	0.241922

1/19

x	$f(x)$
11.0	110.000
11.1	112.088
11.2	114.224
11.3	116.416
11.4	118.672
11.5	121.000

$$7. c) W_n(x) = y_0 + \frac{\Delta y_0}{h}(x - x_0)^{[1]} + \frac{\Delta^2 y_0}{2 \cdot h^2}(x - x_0)^{[2]} + \dots + \frac{\Delta^n y_0}{n! h^n}(x - x_0)^{[n]}$$

$$d) W_n(x) = y_n + \frac{\Delta y_{n-1}}{h}(x - x_n)^{[1]} + \frac{\Delta^2 y_{n-2}}{2 \cdot h^2}(x - x_{n-1})^{[2]} + \dots + \frac{\Delta^n y_0}{n! h^n}(x - x_1)^{[n]}$$

1.2.7. Zbieżność procesów interpolacyjnych

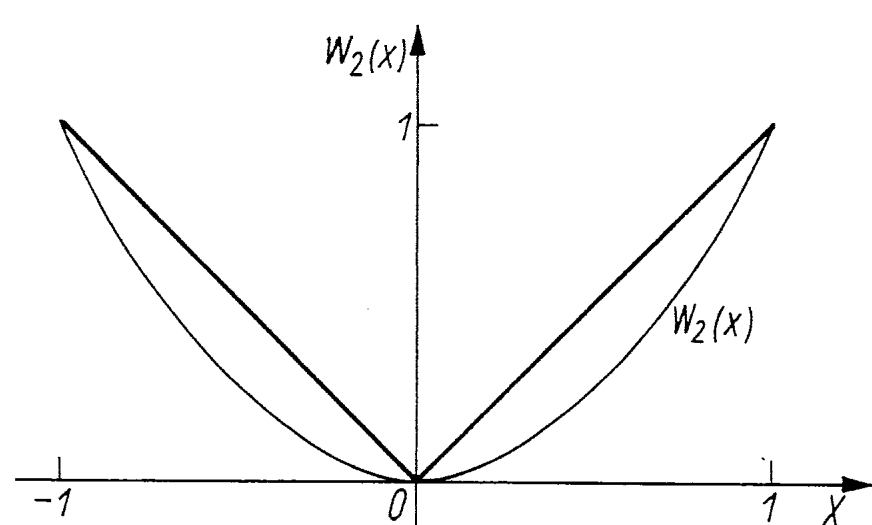
W omawianych zagadnieniach interpolacyjnych mieliśmy dane wartości funkcji $y = f(x)$ w $n + 1$ różnych punktach i poszukiwaliśmy wielomianu algebraicznego stopnia co najwyżej n takiego, że $W_n(x_i) = f(x_i)$, $i = 0, 1, 2, \dots, n$, oraz przybliżonych wartości funkcji $f(x)$ w punktach różnych od węzłów interpolacji. Powstaje pytanie, czy przybliżenie polepszy się, gdy zwiększymy liczbę węzłów? Intuicyjnie wydaje się oczywiste, że jeżeli będziemy mieli większą liczbę zmierzonych wartości funkcji, to można będzie dokładniej opisać matematycznie tę zależność funkcyjną. Nie zawsze jednak tak jest, zwłaszcza wtedy, gdy do opisu matematycznego stosujemy interpolację. Pogorszenie wyników interpolacji przy zwiększaniu liczby węzłów pokażemy na przykładzie funkcji $y = |x|$ na przedziale $\langle -1; +1 \rangle$. Znajdziemy wielomiany interpolacyjne i narysujemy ich wykresy dla $n = 2, 4, 10$ węzłów interpolacyjnych. Mamy

$$\text{dla } n = 2, x_0 = -1, h = 1, W_2(x) = x^2$$

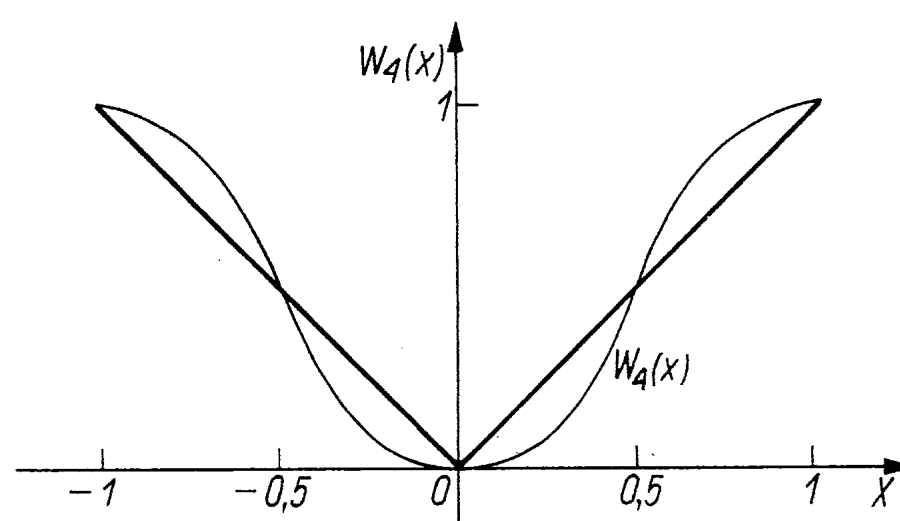
$$\text{dla } n = 4, x_0 = -1, h = 1/2, W_4(x) = -\frac{4}{3}x^4 + \frac{7}{3}x^2$$

$$\begin{aligned} \text{dla } n = 10, x_0 = -1, h = 1/5, W_{10}(x) = & \frac{390625}{5184}x^{10} - \\ & - \frac{1015625}{6048}x^8 + \frac{221875}{1728}x^6 + \frac{6835}{162}x^4 + \frac{11527}{1792}x^2 \end{aligned}$$

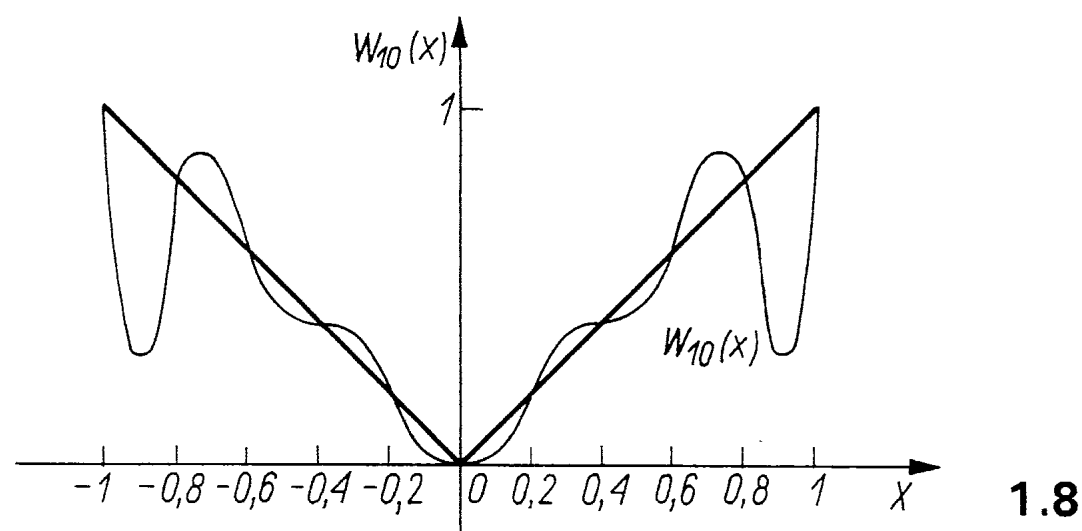
Na rysunkach 1.6, 1.7 i 1.8 są podane wykresy wielomianów interpolacyjnych. Widać wyraźnie, że początkowo ze wzrostem liczby węzłów n przybliżenie polepsza się, lecz przy dalszym wzroście n przybliżenie zaczyna się pogarszać, zwłaszcza przy końcach przedziału. Przybliżony wykres funkcji



1.6



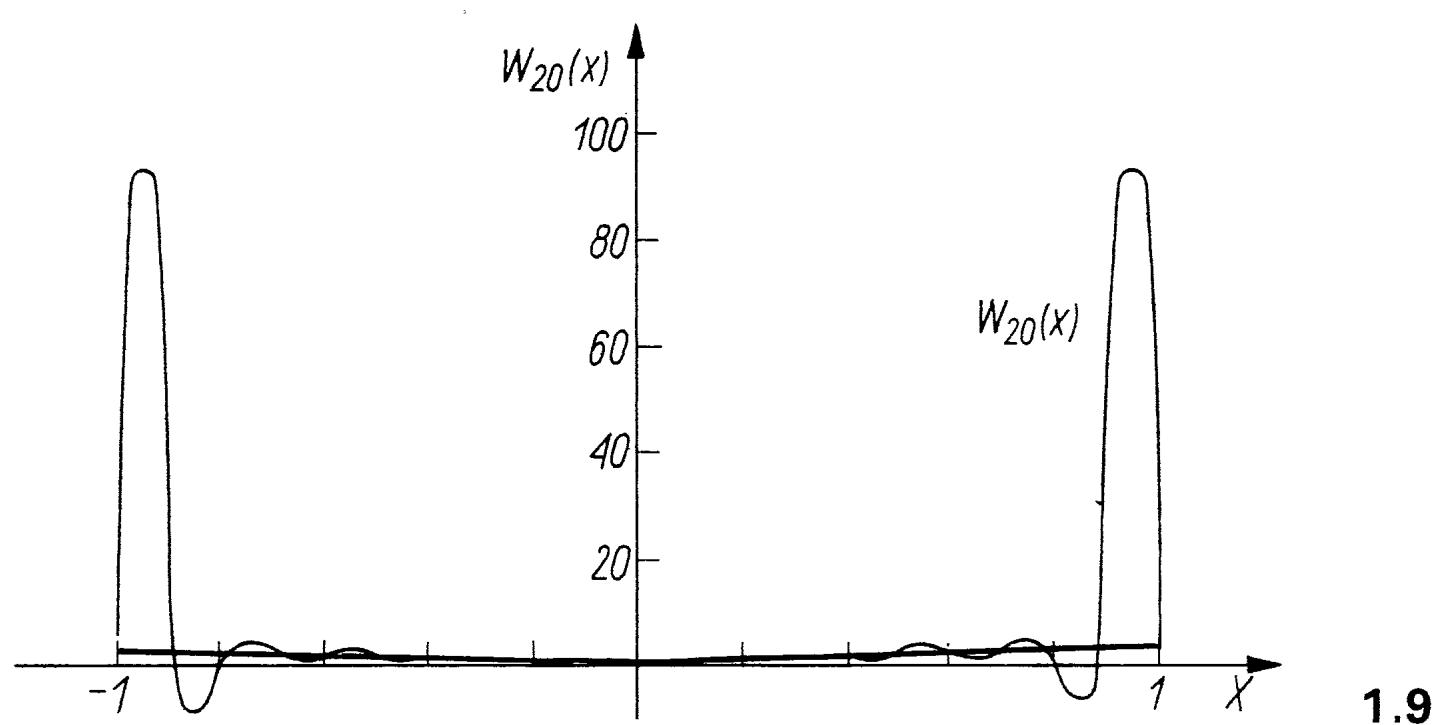
1.7



1.8

$y = |x|$ i jej wielomianu interpolacyjnego dwudziestego stopnia z węzłami -1 ; $-0,95$; $-0,90$; ...; $0,90$; $0,95$; 1 jest pokazany na rys. 1.9; skala na osi rzędnych jest tu sto razy mniejsza niż na osi odciętych. W tym przypadku odchylenia od interpolowanej funkcji są przy końcach przedziału przeszło 100 razy większe niż wartości funkcji.

Wadą funkcji $y = |x|$ użytej w powyższym przykładzie, a wybranej ze względu na jej prostotę, jest fakt, że w punkcie $x = 0$ nie ma ona pochodnej.



1.9

Można jednakże łatwo znaleźć funkcję nie mającą tej wady, np. $y = \frac{1}{1 + 25x^2}$, w przypadku której mamy do czynienia z tym samym zjawiskiem. Takie zachowanie się wielomianu interpolującego (zwłaszcza przy końcach przedziału interpolacji) jest zjawiskiem typowym dla interpolacji za pomocą wielomianów wysokich stopni przy stałych odległościach węzłów. Jest to tak zwane *zjawisko Rungego* — jest ono przykładem źle uwarunkowanego zadania. Interpolacja tego typu na środkowych częściach przedziału $\langle x_0; x_n \rangle$ jest natomiast bardzo dobra i bardzo użyteczna.

Również w przypadku interpolowania funkcji, której wykres znacznie różni się od wykresu wielomianu interpolującego (np. funkcji $y = \frac{1}{x}$), interpolacja przy dużej liczbie węzłów daje nie najlepsze wyniki (pojawiają się ekstrema w funkcji interpolującej). W takich przypadkach lepsze wyniki daje interpolacja przedziałowa (np. wielomianami stopnia drugiego).

Zajmijmy się teraz omówieniem zbieżności metody interpolacyjnej. Niech $f(x)$ oznacza funkcję interpolowaną, natomiast $f_n^*(x)$ funkcję interpolującą stopnia n (przechodzącą przez $n + 1$ równoodległych węzłów).

Def. Mówimy, że metoda interpolacyjna jest *zbieżna* do funkcji $f(x)$ w punkcie x_0 , gdy

$$\lim_{n \rightarrow \infty} f_n^*(x_0) = f(x_0) \quad (1.62)$$

Jeżeli ciąg $(f_n^*(x))$ jest jednostajnie zbieżny na przedziale $\langle a; b \rangle$ do funkcji $f(x)$, to mówimy, że metoda interpolacyjna jest *jednostajnie zbieżna* na przedziale $\langle a; b \rangle$.

Rozpatrzmy metodę interpolacyjną Lagrange'a. Mamy dane $n + 1$ punktów należących do przedziału $\langle a; b \rangle$ i wartości funkcji $f(x)$ w tych punktach. Zgodnie ze wzorem (1.16) wielomian interpolacyjny można zapisać jako

$$W_n(x) = \sum_{j=0}^n f(x_j) \frac{\omega_n(x)}{(x - x_j) \omega_n'(x_j)} \quad (1.16)$$

gdzie $\omega_n(x)$ jest określone zależnością (1.15). Faber i Bernstein [43] udowodnili, że dla każdego układu $n + 1$ węzłów $x_n \in \langle a; b \rangle$ istnieje funkcja ciągła na przedziale $\langle a; b \rangle$, dla której metoda interpolacyjna nie jest jednostajnie zbieżna na tym przedziale do tej funkcji. Nawet dla węzłów równoodległych i tak prostej funkcji, jak omawiana poprzednio funkcja $y = |x|$, Bernstein wykazał, że zastosowana metoda interpolacyjna jest rozbieżna w każdym punkcie przedziału $\langle -1; +1 \rangle$ z wyjątkiem punktów $-1, 0$ i $+1$.

Wykażemy teraz, że dla funkcji całkowitej $f(x)$ metoda interpolacyjna Lagrange'a jest przy dowolnym wyborze węzłów jednostajnie zbieżna na każdym skończonym przedziale. Funkcja $f(x)$ jest funkcją całkowitą, można ją więc przedstawić w postaci szeregu potęgowego

$$f(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k$$

zbieżnego dla każdego x . Reszta wzoru Lagrange'a jest określona zależnością (1.24) i można ją oszacować korzystając z (1.26). Ponieważ $f(x)$ jest funkcją całkowitą, więc

$$f^{(n+1)}(x) = \sum_{k=n+1}^{\infty} a_k k(k-1)(k-2) \dots (k-n) (x - x_0)^{k-(n+1)}$$

i przy założeniu, że $x_0 \in \langle a; b \rangle$ (jest to praktycznie bez znaczenia)

$$\begin{aligned} \sup_{x \in \langle a; b \rangle} |f^{(n+1)}(x)| &\leq \sum_{k=n+1}^{\infty} |a_k| k^{n+1} (b-a)^{k-(n+1)} = \\ &= \frac{(n+1)!}{(b-a)^{n+1}} \sum_{k=n+1}^{\infty} |a_k| \frac{k^{n+1}}{(n+1)!} (b-a)^k \end{aligned}$$

Z rozwinięcia e^k w szereg mamy

$$e^k > \frac{k^{n+1}}{(n+1)!}$$

zatem

$$\sup_{x \in (a; b)} |f^{n+1}(x)| < \frac{(n+1)!}{(b-a)^{n+1}} \sum_{k=n+1}^{\infty} |a_k| [e(b-a)]^k$$

Ponieważ

$$\lim_{n \rightarrow \infty} \sum_{k=n+1}^{\infty} |a_k| [e(b-a)]^k = 0$$

więc metoda interpolacyjna dla funkcji całkowitej $f(x)$ jest jednostajnie zbieżna. Ze sposobu dowodzenia wynika, że tezę twierdzenia można otrzymać również wtedy, gdy funkcja $f(x)$ jest rozwijalna w szereg Taylora w otoczeniu

punktu $x_0 = \frac{a+b}{2}$ o promieniu zbieżności $R > \frac{b-a}{2}e$.

Inne ciekawe wyniki dotyczące zbieżności metod interpolacyjnych można znaleźć w pracy B. SENDOWA [56].

1.2.8. Uwagi końcowe

Interpolację stosujemy w następujących przypadkach:

- przy obliczaniu wartości funkcji określonej za pomocą tablicy wartości w punkcie różnym od podanych w tej tablicy,
- przy zagęszczaniu tablic,
- przy obliczaniu poprawek, na przykład dla stablicowanych funkcji o rzadkim kroku,
- przy zastępowaniu funkcji zbyt skomplikowanych wielomianem odpowiedniego stopnia.

Dla typowych funkcji zagęszczanie tablic nie jest zbyt trudnym zadaniem, lecz dla funkcji rzadko używanych zagęszczanie jest jednym z zasadniczych problemów. Mając do dyspozycji dużą i szybką maszynę cyfrową niejednokrotnie ekonomiczniej jest prowadzić interpolację wielomianem stosunkowo wysokiego stopnia — aby zaoszczędzić miejsca w pamięci maszyny kosztem zwiększenia liczby niezbędnych operacji arytmetycznych.

Przy obliczaniu wartości wielomianu interpolacyjnego w jednym bądź kilku punktach problem wyboru postaci wzoru interpolacyjnego właściwie nie istnieje. Rodzaj wybranego wzoru, podobnie jak rozmieszczenie węzłów, ma wówczas wpływ jedynie na błąd obliczeń. Analizując wzór (1.26), określający maksymalny błąd (bezwzględny) popełniony przy interpolacji, widzimy, że jedynym czynnikiem, na który mamy wpływ, jest

$$\omega_n(x) = (x - x_0)(x - x_1)(x - x_2) \dots (x - x_n)$$

Gdy mamy możliwość wyboru rozmieszczenia węzłów i żądamy, aby błąd interpolacji był na całym przedziale jak najmniejszy, wówczas, jak wykazaliśmy w punkcie 1.2.3, jako węzły należy przyjąć punkty określone wzorami (1.32). Ogólnie należy starać się, aby węzły były rozłożone po obu stronach danego punktu x , i to w miarę symetrycznie.

Problem wyboru odpowiedniego wzoru interpolacyjnego ma szczególne znaczenie przy obliczeniach „ręcznych”, gdy nie możemy zbytnio zagęszczać węzłów. Stosuje się wówczas również inne (nie omawiane tutaj) wzory interpolacyjne, zawierające różnice położone nad i pod poziomym wierszem, w którym powinna znajdować się podana wartość zmiennej x . Są to tak zwane *wzory z różnicami centralnymi*; zaliczamy do nich wzory: Gaussa, Stirlinga i Bessela. Ze względu na ich stosunkowo małą przydatność do obliczeń prowadzonych na maszynach cyfrowych nie będziemy ich szerzej omawiać. Informacje dotyczące tych wzorów Czytelnik może znaleźć np. w [17].

Przy obliczeniach „ręcznych” w przypadku węzłów równoodległych stosuje się wzory zawierające różnice skończone, gdyż sporządzenie tablic takich różnic jest znacznie prostsze od obliczania wartości wielomianu zbudowanego na podstawie wzoru Lagrange’a (1.16).

Oszacowanie błędu ze wzoru (1.26) jest w praktyce bardzo uciążliwe ze względu na trudności oszacowania wartości $f^{(n+1)}(x)$. Przy stałych odległościach między węzłami, mamy

$$\frac{M_n |\omega_{n+1}(x)|}{(n+1)!} \leq M_{n+1} h^{n+1} \left| \frac{q(q-1)\dots(q-n)}{(n+1)!} \right|$$

a gdy pochodna mało się zmienia, wówczas

$$M_{n+1} h^{n+1} \approx |\Delta^{n+1} y_0|$$

i błąd interpolacji możemy oszacować wielkością wyrażenia

$$\left| \frac{q(q-1)(q-2)\dots(q-n)}{(n+1)!} \Delta^{n+1} y_0 \right|$$

czyli wielkością modułu dodatkowego wyrazu, o który $W_{n+1}(x)$ różni się od $W_n(x)$.

Przy obliczeniach na maszynach cyfrowych problem wyboru stopnia wielomianu interpolacyjnego rozwiązuje się najczęściej w następujący sposób. Gdy funkcja interpolowana dana jest za pomocą tablicy wartości albo oszacowanie błędu jest trudne, interpolację rozpoczynamy od $n = 1$ (interpolacja liniowa) i zwiększając stopniowo liczbę węzłów zwiększamy stopień wielomianu aż do ustabilizowania się istotnych dla nas miejsc dziesiętnych. Należy przy tym pamiętać o istnieniu maksymalnej granicznej dokładności [65], której na ogół nie można przekroczyć.

Wybrana postać wielomianu interpolacyjnego decyduje o algorytmie obliczania jego wartości, a więc i o błędzie otrzymanym w wyniku tych obliczeń. Można wykazać, że w przypadku stosowania wzoru (1.53) błąd ten jest najmniejszy, gdy

$$|x - x_0| \leq |x - x_1| \leq |x - x_2| \leq \dots \leq |x - x_n|$$

Stąd wniosek, że wzór (1.53) nadaje się najlepiej do interpolacji dla x znajdujących się na początku tablicy, czyli dla $x_0 < x < x_1$. Analogicznie wzór (1.58) najlepiej nadaje się do interpolacji dla x z końca tablicy, czyli dla $x_{n-1} < x < x_n$.

Podstawą do określenia liczby działań niezbędnych do wykonania obliczeń według danego wzoru (a zatem i czasochłonności tych obliczeń) jest zazwyczaj liczba mnożeń i dzielen. Czasy dodawania i odejmowania są bowiem zazwyczaj znacznie krótsze od czasów mnożeń i dzielen. We współczesnych, bardzo szybkich maszynach cyfrowych coraz częściej zdarza się, że czasy te są zbliżone do siebie.

Liczby działań arytmetycznych potrzebnych do obliczenia pojedynczej wartości wielomianu interpolacyjnego (stopnia n) są następujące:

- dla wielomianu postaci Lagrange'a $\frac{3}{2}n^2 + \frac{13}{2}n + 3$ działań, w tym $n^2 + 4n + 2$ mnożeń i dzielen,
- dla wielomianu postaci Newtona $\frac{3}{2}n^2 + \frac{9}{2}n$, w tym $\frac{1}{2}n^2 + \frac{3}{2}n$ mnożeń i dzielen,
- dla wzoru Aitkena $2n^2 + 3n + 1$ działań, w tym $\frac{1}{2}n^2 + \frac{1}{2}n$ mnożeń i dzielen.

Jak widać, iteracyjna metoda Aitkena jest nieco wolniejsza, lecz ze względu na jednolitość oraz łatwość zaprogramowania jest metodą często stosowaną. Ponadto, jeśli przy stosowaniu tej metody tak uporządkujemy (w ramach tablicy) dane węzły, aby wartości bezwzględne liczb $x - x_k$ wzrastały, to uzyskamy znaczne zwiększenie dokładności pierwszych przybliżeń.

Dotychczas zajmowaliśmy się wyznaczaniem wartości funkcji dla pewnych argumentów nie występujących w danej tablicy wartości. W praktyce czasami mamy do czynienia z zagadnieniem odwrotnym — wyznaczaniem wartości zmiennej niezależnej x , której odpowiada dana wartość funkcji nie występująca w tablicy wartości. Zagadnienie takie nazywamy *interpolacją odwrotną*. Jego rozwiązanie jest stosunkowo proste: korzystamy z któregoś ze znanych wzorów interpolacyjnych — oczywiście dostosowanego do zmienego kroku w tablicy — zamieniając miejscami zmienne x i y w tablicy i wzorze.

1.3. INTERPOLACJA ZA POMOCĄ FUNKCJI SKLEJANYCH

1.3.1. Określenie funkcji sklejanых

Niech w przedziale $\langle a; b \rangle$ danych będzie $(n + 1)$ punktów $x_0, x_1, \dots, x_n$, przy czym

$$a = x_0 < x_1 < \dots < x_{n-1} < x_n = b \quad (1.63)$$

Punkty $x_i, i = 0, 1, \dots, n$, określają pewien podział przedziału $\langle a; b \rangle$ na n podprzedziałów. Podział ten oznaczmy symbolem Δ_n .

Funkcję $s(x) = s(x, \Delta_n)$ określoną na przedziale $\langle a; b \rangle$ nazywamy *funkcją sklejaną* stopnia $m (m \geq 1)$; jeżeli:

1) $s(x)$ jest wielomianem stopnia co najwyżej m na każdym podprzedziale $(x_i; x_{i+1}), i = 0, 1, \dots, n - 1$

2) $s(x) \in C^{m-1}(\langle a; b \rangle)$

Punkty x_i nazywamy *węzłami funkcji sklejaney*.

Zbiór wszystkich funkcji sklejanых stopnia m o węzłach w punktach x_i oznaczmy $S_m(\Delta_n)$. Niech $s(x) \in S_m(\Delta_n)$. Na każdym przedziale $(x_i; x_{i+1}), i = 0, 1, \dots, n - 1$, funkcja $s(x)$ jest wielomianem stopnia co najwyżej m

$$s(x) = c_{im} x^m + c_{im-1} x^{m-1} + \dots + c_{i1} x + c_{i0} \quad (1.64)$$

$$x \in (x_i; x_{i+1})$$

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

W obliczeniach na maszynach cyfrowych funkcje bardzo często przybliża się funkcjami sklejanymi. Związane jest to z tym, że wartości tych funkcji są łatwe do wyznaczenia oraz że są one zbieżne dla licznych klas funkcji (jest to jedna z zalet funkcji sklejanых w porównaniu z wielomianami). W praktyce często stosuje się funkcje sklejaney stopnia trzeciego. Dla wielu zagadnień są one wystarczająco gładkie, szybkość ich zbieżności jest wystarczająca, a ponadto wyznacza się je stosunkowo łatwo. Postać interpolującej funkcji sklejaney w dużym stopniu zależy od zagadnienia. Często wygodnie jest przedstawić poszukiwaną funkcję $s(x) \in S_m(\Delta_n)$ w postaci kombinacji liniowej elementów bazy przestrzeni $S_m(\Delta_n)$.

Wyznamy teraz bazę przestrzeni funkcji $s(x)$ stopnia trzeciego z węzłami równoodległymi

$$x_1 = x_0 + ih, \quad h = \frac{b - a}{n}, \quad i = 0, 1, \dots, n$$

Oznaczmy dodatkowo przez $x_{-3}, x_{-2}, x_{-1}, x_{n+1}, x_{n+2}, x_{n+3}$ punkty $x_i = x_0 + ih$ dla $i = -3, -2, -1, n + 1, n + 2, n + 3$ i określmy funkcje

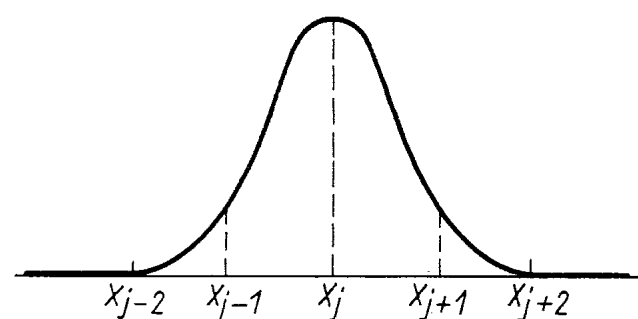
$\Phi_i^3(x)$, $i = -1, 0, 1, \dots, n, n+1$, w następujący sposób

$$\Phi_i^3(x) = \frac{1}{h^3} \begin{cases} (x - x_{i-2})^3 & \text{dla } x \in \langle x_{i-2}; x_{i-1} \rangle \\ h^3 + 3h^2(x - x_{i-1}) + 3h(x - x_{i-1})^2 - 3(x - x_{i-1})^3 & \text{dla } x \in \langle x_{i-1}; x_i \rangle \\ h^3 + 3h^2(x_{i+1} - x) + 3h(x_{i+1} - x)^2 - 3(x_{i+1} - x)^3 & \text{dla } x \in \langle x_i; x_{i+1} \rangle \\ (x_{i+2} - x)^3 & \text{dla } x \in \langle x_{i+1}; x_{i+2} \rangle \\ 0 & \text{dla pozostałych } x \in R \end{cases} \quad (1.65)$$

Funkcje (1.65) są klasy $C^2((-\infty; +\infty))$. W tablicy 1/20 podano wartości funkcji $\Phi_j^3(x)$ oraz jej pierwszej i drugiej pochodnej w punktach x_k dla $k = j-2, j-1, j, j+1, j+2$ (poza przedziałem $\langle x_{j-2}; x_{j+2} \rangle$ funkcja ta jest tożsamościowo równa zeru), a na rys. 1.10 pokazano jej wykres.

1/20

	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	$6/h^2$	$-12/h^2$	$6/h^2$	0



1.10

Tw. Funkcje $\bar{\Phi}_i(x)$, $i = -1, 0, 1, \dots, n+1$, określone na przedziale $\langle a; b \rangle$ w następujący sposób

$$\bar{\Phi}_i(x) = \Phi_i^3(x), \quad a \leq x \leq b \quad (1.66)$$

stanowią bazę przestrzeni funkcji sklejanych trzeciego stopnia $\mathcal{S}_3(\Delta_n)$.

Każdą funkcję $s(x) \in \mathcal{S}_3(\Delta_n)$ można więc przedstawić w postaci kombinacji liniowej

$$s(x) = \sum_{i=-1}^{n+1} c_i \Phi_i^3(x), \quad a \leq x \leq b \quad (1.67)$$

gdzie c_i są liczbami rzeczywistymi.

1.3.2. Interpolacyjne funkcje sklepane stopnia trzeciego

Omówimy teraz interpolację za pomocą funkcji sklejanych trzeciego stopnia. Niech $f(x) \in C(\langle a; b \rangle)$. Funkcję $s(x) \in \mathcal{S}_3(\Delta_n)$ nazywamy *interpolacyjną funkcją sklepaną stopnia trzeciego* dla funkcji $f(x)$, jeżeli

$$s(x_i) = f(x_i) = y_i, \quad i = 0, 1, \dots, n; \quad n \geq 2 \quad (1.68)$$

Jak wiemy, funkcja $s(x)$ stopnia trzeciego zależy od $(n + 3)$ parametrów. Tak więc interpolacyjne funkcje skleiane stopnia trzeciego mają dwa stopnie swobody, wobec czego nakładamy na nie dwa dodatkowe warunki. Wybór tych warunków zależy zarówno od własności funkcji $f(x)$, jak i od posiadanej informacji o tej funkcji.

Najczęściej rozpatrujemy następujące dodatkowe warunki

$$s'(a + 0) = \alpha_1, \quad s'(b - 0) = \beta_1 \quad (1.69)$$

lub

$$s''(a + 0) = \alpha_2, \quad s''(b - 0) = \beta_2 \quad (1.70)$$

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 możemy je przyjąć jako liczby występujące po prawych stronach warunków (1.69), (1.70). Na funkcję $s(x)$ możemy nakładać i inne warunki. Na przykład, jeżeli $f(x)$ jest funkcją okresową, o okresie $(b - a)$, to najczęściej żądamy, aby funkcję sklejaną $s(x)$ można było przedłużyć na przedział $(-\infty; +\infty)$ w taki sposób, że będzie ona funkcją okresową o okresie $(b - a)$, różniczkowalną w sposób ciągły. W takiej sytuacji na funkcję sklejaną nakładamy warunki

$$s^{(i)}(a + 0) = s^{(i)}(b - 0), \quad i = 1, 2 \quad (1.71)$$

Nasuują się następujące pytania: czy istnieje funkcja sklejana trzeciego stopnia spełniająca zależności (1.68) oraz jeden z warunków (1.69)—(1.71), i czy warunki te określają tę funkcję jednoznacznie? A jeżeli tak, to jak można wyznaczyć tę funkcję oraz jaki jest błąd interpolacji na przedziale $\langle a; b \rangle$?

Omówimy najpierw sposoby wyznaczania oraz problem istnienia i jednoznaczności interpolacyjnej funkcji sklejaney dla węzłów dowolnie rozłożonych, a pod koniec niniejszego punktu rozpatrzemy dokładniej przypadek węzłów równoległych. Oznaczmy

$$M_j = s''(x_j), \quad j = 0, 1, \dots, n \quad (1.72)$$

Zgodnie z określeniem funkcji sklejaney trzeciego stopnia, pochodna $s'(x)$ jest funkcją ciągłą na przedziale $\langle a; b \rangle$ i liniową na każdym podprzedziale $\langle x_{j-1}; x_j \rangle$, można więc przedstawić ją w postaci

$$s''(x) = M_{j-1} \frac{x_j - x}{h_j} + M_j \frac{x - x_{j-1}}{h_j}, \quad x \in \langle x_{j-1}; x_j \rangle \quad (1.73)$$

$$h_j = x_j - x_{j-1}$$

Całkując stronami równość (1.73), otrzymujemy (dla $x \in \langle x_{j-1}; x_j \rangle$)

$$s'(x) = -M_{j-1} \frac{(x_j - x)^2}{2h_j} + M_j \frac{(x - x_{j-1})^2}{2h_j} + A_j \quad (1.74)$$

$$s(x) = M_{j-1} \frac{(x_j - x)^3}{6h_j} + M_j \frac{(x - x_{j-1})^3}{6h_j} + A_j(x - x_{j-1}) + B_j \quad (1.75)$$

gdzie A_j, B_j są stałymi. Nakładając na (1.74) i (1.75) 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 . Mamy

$$B_j = y_{j-1} - M_{j-1} \frac{h_j^2}{6} \quad (1.76)$$

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

Nałożymy teraz warunek na $s(x)$, aby pochodna $s'(x)$ była funkcją ciągłą na $\langle a; b \rangle$. W tym celu obliczymy granice jednostronne

$$s'(x_j - 0) = \frac{h_j}{6} M_{j-1} + \frac{h_j}{3} M_j + \frac{y_j - y_{j-1}}{h_j} \quad (1.77)$$

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

i zażądamy, aby

$$s'(x_j - 0) = s'(x_j + 0), \quad j = 1, \dots, n-1 \quad (1.78)$$

Po podstawieniu (1.77) do (1.78) otrzymujemy układ $(n-1)$ równań

$$\frac{h_j}{6} M_{j-1} + \frac{h_j + h_{j+1}}{3} M_j + \frac{h_{j+1}}{6} M_{j+1} = \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j} \quad (1.79)$$

$$j = 1, 2, \dots, n-1$$

Równania te wygodnie jest zapisać w postaci

$$\mu_j M_{j-1} + 2M_j + \lambda_j M_{j+1} = d_j, \quad j = 1, \dots, n-1 \quad (1.80)$$

$$\text{gdzie: } \lambda_j = \frac{h_{j+1}}{h_j + h_{j+1}}, \quad \mu_j = 1 - \lambda_j,$$

$$d_j = \frac{6}{h_j + h_{j+1}} \left(\frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j} \right) = 6f(x_{j-1}; x_j; x_{j+1})$$

Jeżeli $M_j, j = 0, 1, \dots, n$, spełniają powyższy układ, to funkcja $s(x)$ określona wzorami (1.75), (1.76) na każdym podprzedziale $\langle x_{j-1}; x_j \rangle, j = 1, \dots, n$, jest interpolacyjną funkcją sklepaną stopnia trzeciego.

Do układu (1.80) należy dołączyć jeszcze dwa równania, wynikające ze spełnienia przez funkcję $s(x)$ jednego z dodatkowych warunków (1.69)—(1.71). Równania te mają postać:

dla warunków (1.69)

$$2M_0 + M_1 = d_0, \quad M_{n-1} + 2M_n = d_n \quad (1.81)$$

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

dla warunków (1.70)

$$M_0 = \alpha_2, \quad M_n = \beta_2 \quad (1.82)$$

dla warunków (1.71)

$$M_0 = M_n, \quad \lambda_n M_1 + \mu_n M_{n-1} + 2M_n = \tilde{d}_n \quad (1.83)$$

$$\text{gdzie } \lambda_n = \frac{h_1}{h_1 + h_n}, \quad \mu_n = 1 - \lambda_n, \quad \tilde{d}_n = \frac{6}{h_1 + h_n} \left(\frac{y_1 - y_0}{h_1} - \frac{y_n - y_{n-1}}{h_n} \right).$$

Zapiszemy teraz układ równań (1.80) i (1.81) w postaci macierzowej

$$\begin{bmatrix} 2 & 1 & 0 & \dots & 0 \\ \mu_1 & 2 & \lambda_1 & \dots & 0 \\ 0 & \mu_2 & 2 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & \dots & \dots & 2 & \lambda_{n-1} \\ 0 & \dots & \dots & 1 & 2 \end{bmatrix} \begin{bmatrix} M_0 \\ M_1 \\ \vdots \\ M_{n-1} \\ M_n \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_{n-1} \\ d_n \end{bmatrix} \quad (1.84)$$

W przypadku układu równań (1.80), (1.82) rugując dane wielkości $M_0 = \alpha_2$, $M_n = \beta_2$ z równań (1.80), otrzymamy także układ o macierzy trójdzielnej, natomiast w przypadku układu równań (1.80), (1.83) rugując M_0 ($M_0 = M_n$), otrzymamy

$$\begin{bmatrix} 2 & \lambda_1 & 0 & \dots & 0 & \mu_1 \\ \mu_2 & 2 & \lambda_2 & \dots & 0 & 0 \\ 0 & \mu_3 & 2 & \dots & 0 & 0 \\ \vdots & & & & \vdots & \\ 0 & 0 & 0 & & 2 & \lambda_{n-1} \\ \lambda_n & 0 & 0 & \dots & \mu_n & 2 \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_{n-1} \\ M_n \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-1} \\ \tilde{d}_n \end{bmatrix} \quad (1.85)$$

We wszystkich trzech przypadkach macierze współczynników układów, które należy rozwiązać, są macierzami silnie diagonalnie dominującymi (tzn. moduły elementów leżących na głównej przekątnej są większe od sumy modułów pozostałych elementów leżących w tym samym wierszu). Stąd wynika, że układy te mają jednoznaczne rozwiązanie (patrz wniosek z twierdz.

14 w p. 6.2), istnieje zatem dokładnie jedna interpolacyjna funkcja sklejana stopnia trzeciego, spełniająca jeden z dodatkowych warunków (1.69)—(1.71).

Dla układów z macierzami trójdziagonalnymi silnie diagonalnie dominującymi są opracowane niezawodne algorytmy ich rozwiązywania (patrz p. 5.3.6). Metody te można w sposób prosty uogólnić na przypadek układów postaci (1.85) (patrz p. 5.3.7), a więc wyznaczanie interpolacyjnych funkcji sklepanych w przedstawiony wyżej sposób nie nastęrcza numerycznych trudności.

Do obliczeń, w których korzystamy z interpolacyjnych funkcji sklepanych o przedstawionej wyżej postaci, są niezbędne następujące dane: węzły x_i , wartości drugiej pochodnej funkcji sklepanej $M_j = s''(x_j)$ oraz wartości funkcji $y_i = f(x_i)$. Przedstawiona wyżej postać funkcji $s(x)$ jest bardzo dogodna, zwłaszcza w takich obliczeniach, w których należy wielokrotnie liczyć nie tylko wartości tej funkcji, ale także jej drugiej pochodnej.

Zauważmy, że jeżeli węzły są równoodległe, to

$$\lambda_j = \mu_j = \frac{1}{2}$$

W ćwiczeniu 3 jest rozpatrzony inny sposób przedstawienia interpolacyjnej funkcji sklepanej. Jako niewiadome wielkości przyjmuje się tam $m_i = s'(x_i)$.

Omówimy teraz problem zbieżności rozpatrywanych w tym punkcie interpolacyjnych funkcji sklepanych. Określimy ciąg (Δ_k) podziałów przedziału $\langle a; b \rangle$

$$\Delta_k: a = x_0^{(k)} < x_1^{(k)} < \dots < x_{p_k}^{(k)} = b \quad (1.86)$$

w którym punkty $x_i^{(k)}$ są węzłami funkcji sklepanej dla podziału Δ_k . Przyjmiemy następujące oznaczenia: $\tilde{C}$ — zbiór funkcji okresowych o okresie $(b - a)$, ciągłych na przedziale $(-\infty; +\infty)$; $\tilde{C}^p$ — zbiór funkcji należących do $\tilde{C}$ i mających ciągłą p -tą pochodną; $s_{\Delta_k}(x)$ — funkcja sklejana stopnia trzeciego interpolująca funkcję $f(x)$ w węzłach $x_i^{(k)}$, $i = 0, 1, \dots, p_k$, spełniająca jeden z dodatkowych warunków (1.69)—(1.71) (dla podziału Δ_k prawe strony w (1.69), (1.70) oznaczmy odpowiednio $\alpha_1^k, \beta_1^k, \alpha_2^k, \beta_2^k$), oraz

$$\bar{h}_k = \max_{1 \leq i \leq p_k} (x_i^{(k)} - x_{i-1}^{(k)}), \quad \underline{h}_k = \min_{1 \leq i \leq p_k} (x_i^{(k)} - x_{i-1}^{(k)})$$

Założymy, że

$$\lim_{k \rightarrow \infty} \bar{h}_k = 0, \quad \bar{h}_k \underline{h}_k^{-1} \leq \beta \quad (1.87)$$

gdzie β nie zależy od k .

Szybkość zbieżności funkcji $s_{\Delta_k}(x)$ i jej pochodnych do funkcji interpolowanej $f(x)$ i jej pochodnych zależy głównie od następujących czynników: regularności funkcji $f(x)$ i dokładności, z jaką ta funkcja spełnia dodatkowe

warunki nałożone na funkcję $s_{\Delta_k}(x)$. Zwiększenie tej regularności i dokładności polepsza zbieżność.

Tw. Jeżeli

1) $f(x) \in \tilde{C}^q$, a $s_{\Delta_k}(x)$ jest funkcją sklejaną spełniającą dodatkowo warunek (1.71) lub

2) $f(x) \in C^q(\langle a; b \rangle)$, a $s_{\Delta_k}(x)$ jest funkcją sklejaną spełniającą dodatkowo jeden z warunków (1.69), (1.70)

to

$$\begin{aligned} f^{(p)}(x) - s_{\Delta_k}^{(p)}(x) &= O(\bar{h}_k^{q-p}) \\ \text{gdy } q &\leq 3, \quad 0 \leq p \leq q \\ f^{(p)}(x) - s_{\Delta_k}^{(p)}(x) &= O(\bar{h}_k^{4-p}) \\ \text{gdy } q &\geq 4, \quad 0 \leq p \leq 4 \end{aligned} \tag{1.88}$$

jednostajnie względem x na $\langle a; b \rangle^1$, przy dodatkowym założeniu, że

$$\alpha_1^k = f'(a), \quad \beta_1^k = f'(b) \quad \text{dla warunku (1.69)}$$

$$\alpha_2^k = f''(a), \quad \beta_2^k = f''(b) \quad \text{dla warunku (1.70)}$$

Dowód tego twierdzenia pomijamy; można go znaleźć np. w [2], [57].

Jeżeli dodatkowe warunki (1.69), (1.70) nie są spełnione dokładnie przez funkcję $f(x)$, to szybkość zbieżności może być bardzo mała lub może jej nie być wcale. Dokładny opis warunków, jakie powinny spełniać wielkości α_i^k , β_i^k , aby nie nastąpiło pogorszenie zbieżności, można znaleźć np. w [2], [57].

Rozpatrzmy teraz zagadnienie interpolacji, gdy w węzłach mamy dane tylko wartości funkcji $f(x)$. Niech: $\lambda_0(x)$, $\lambda_n(x)$ będą wielomianami interpolacyjnymi dla funkcji $f(x)$ z węzłami odpowiednio: x_0, x_1, x_2, x_3 oraz $x_{n-3}, x_{n-2}, x_{n-1}, x_n$ i niech $s_L(x) \in \mathcal{S}_3(\Delta_n)$ oznacza interpolacyjną funkcję sklejaną spełniającą dodatkowe warunki

$$\begin{aligned} s_L'(a+0) &= \alpha_1 = \lambda_0'(a) \\ s_L'(b-0) &= \beta_1 = \lambda_n'(b) \end{aligned} \tag{1.89}$$

Funkcję $s_L(x)$ nazywamy *interpolacyjną funkcją sklejaną Lagrange'a stopnia trzeciego*. Można udowodnić [59] że, jeżeli $f(x) \in C^4(\langle a; b \rangle)$, to

$$f(x) - s_L(x) = O(h_k^4), \quad a \leq x \leq b \tag{1.90}$$

Na zakończenie niniejszego punktu omówimy inny sposób wyznaczania interpolacyjnej funkcji sklejaney, w przypadku równoodległych węzłów

$$x_i = x_0 + ih$$

¹⁾ Uwaga: moduł lewej strony w (1.88) można oszacować przez niezależną od x prawą stronę, jako funkcję zmiennej $\bar{h}_k$.

$$h = \frac{b - a}{n}$$

$$i = 0, 1, \dots, n$$

Poszukiwać jej będziemy w postaci kombinacji liniowej (1.67)

$$s(x) = \sum_{i=-1}^{n+1} c_i \Phi_i^3(x) \quad (1.91)$$

$$a \leq x \leq b$$

gdzie $\Phi_i^3(x)$ są funkcjami określonymi wzorem (1.65), a c_i — liczbami rzeczywistymi.

Podstawmy (1.91) do (1.68). Na podstawie tablicy 1.20 możemy stwierdzić, że stałe c_i powinny spełniać następujący układ $(n + 1)$ równań

$$c_{i-1} + 4c_i + c_{i+1} = y_i \quad (1.92)$$

$$i = 0, 1, \dots, n$$

Omówimy dokładniej przypadek, gdy interpolacyjna funkcja sklejana spełnia dodatkowe warunki (1.69). Dwa równania, które należy dołączyć do układu (1.92), są wówczas następujące

$$\begin{aligned} -c_{-1} + c_1 &= \frac{h}{3}\alpha_1 \\ -c_{n-1} + c_{n+1} &= \frac{h}{3}\beta_1 \end{aligned} \quad (1.93)$$

Po wyeliminowaniu z układów (1.92), (1.93) współczynników c_{-1} , c_{n+1} , pozostałe współczynniki c_j , $j = 0, 1, \dots, n$ rozwinięcia (1.91) będą rozwiązaniem rozważanego układu. Macierz współczynników tego układu

$$\begin{bmatrix} 4 & 2 & & & & & \\ 1 & 4 & 1 & & & & 0 \\ & 1 & 4 & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & 0 & & 1 & 4 & 1 & \\ & & & & 2 & 4 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \\ c_n \end{bmatrix} = \begin{bmatrix} y_0 + \frac{h}{3}\alpha_1 \\ y_1 \\ \vdots \\ y_{n-1} \\ y_n - \frac{h}{3}\beta_1 \end{bmatrix} \quad (1.94)$$

jest macierzą trójdagonalną z elementami dominującymi na głównej przekątnej.

Wyznaczenie równań, które należy dołączyć do układu (1.92) w przypadku, gdy interpolacyjna funkcja sklejana spełnia dodatkowy warunek (1.70)

lub (1.71), a następnie przeprowadzenie dyskusji otrzymanego układu równań pozostawiamy Czytelnikowi (patrz ćwic. 4).

Jeżeli w dalszych obliczeniach, już po wyznaczeniu współczynników c_j , nie są potrzebne wartości funkcji y_i , to nie musimy ich pamiętać. Zauważmy ponadto, że obliczając wartość $s(x)$ w dowolnym punkcie $x \in \langle a; b \rangle$ wystarczy obliczyć wartości co najwyżej czterech funkcji $\Phi_j^3(x)$, ponieważ wartości pozostałych funkcji są równe zeru. Tak więc dla ustalonego x co najwyżej cztery składniki w sumie (1.91) są różne od zera.

Ćwiczenia

1. Udowodnić, że $S_m(\Delta_n)$ jest przestrzenią liniową.
2. Wykazać, że zbiór funkcji

$$\Phi^3 = \{\Phi_{-1}^3, \Phi_0^3, \dots, \Phi_n^3, \Phi_{n+1}^3\}$$

jest liniowo niezależny na $\langle a; b \rangle$. Czy jest on liniowo niezależny w $\mathbf{R}$?

3. Wyznaczyć interpolacyjną funkcję sklejaną $s(x) \in S_3(\Delta_n)$ spełniającą jeden z dodatkowych warunków (1.69)–(1.71), jeżeli jako niewiadome wielkości zostaną przyjęte $m_i = s'(x_i)$.

W s k a z ó w k a. Na przedziale $\langle x_{j-1}; x_j \rangle$ poszukiwana funkcja jest postaci

$$s(x) = m_{j-1} \frac{(x_j - x)^2(x - x_{j-1})}{h_j^2} - m_j \frac{(x - x_{j-1})^2(x_j - x)}{h_j^2} + \\ + y_{j-1} \frac{(x_j - x)^2[2(x - x_{j-1}) + h_j]}{h_j^3} + y_j \frac{(x - x_{j-1})[2(x_j - x) + h_j]}{h_j^3}$$

Należy nałożyć warunek ciągłości na $s''(x)$ w węzłach x_j ($j = 1, \dots, n-1$).

4. W punkcie 1.3.2 dla węzłów równoodległych poszukiwaliśmy interpolacyjnej funkcji sklejanego stopnia trzeciego w postaci kombinacji liniowej elementów bazy przestrzeni $S_3(\Delta_n)$. Funkcje te spełniały dodatkowy warunek (1.69). Rozpatrzyć przypadek, gdy są spełnione dodatkowe warunki (1.70) lub (1.71).

5. Wyznaczyć p występujące we wzorach

$$\alpha_1 = f'(a) + O(h^p), \quad \beta_1 = f'(b) + O(h^p)$$

gdzie α_1, β_1 są określone wzorem (1.89).

2

APROKSYMACJA

2.1. WSTĘP

W poprzednim rozdziale omówiliśmy zagadnienie interpolacji, czyli poszukiwanie pewnej funkcji interpolującej $F(x)$, która na danym, dyskretnym zbiorze argumentów pokrywa się z wartościami funkcji interpolowanej $f(x)$. Takie postępowanie jest często niedogodne. Przy interpolacji wielomianowej duża liczba węzłów interpolacji wymaga skonstruowania wielomianu interpolacyjnego wysokiego stopnia. Ponadto często mamy do czynienia z funkcją, której wartości na dyskretnym zbiorze argumentów są określone empirycznie (są np. wynikami pomiarów), a więc mogą być obarczone błędami, i wówczas żądanie dokładnego przyjmowania przez szukaną funkcję niedokładnych danych wartości nie ma sensu.

Zajmiemy się teraz zagadnieniem bardziej ogólnym, w którym warunek, aby funkcja dana i funkcja szukana przyjmowały dokładnie te same wartości na zbiorze danych z góry punktów węzłowych, nie musi być spełniony. 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 pojawienie się błędów, zwanych *błędami aproksymacji* (*przybliżenia*) i problem oszacowania tych błędów oraz ich wielkość mają istotny wpływ na wybór metody aproksymacyjnej.

Gdy zbiór, na którym mierzymy błąd aproksymacji, jest zbiorem dyskretnym, wówczas mówimy o aproksymacji *punktowej*, gdy zaś jest to przedział — aproksymację nazywamy *integralną*.

Niech $f(x)$ będzie funkcją, którą chcemy aproksymować, X — pewną przestrzenią liniową unormowaną (skończenie lub nieskończenie wymiarową; $f \in X$), a X_m — m -wymiarową podprzestrzenią liniową przestrzeni X .

Aproksymacja funkcji $f(x)$ polega na wyznaczeniu takich współczynników $a_0, a_1, \dots, a_m$ funkcji

$$F(x) = a_0 \varphi_0(x) + a_1 \varphi_1(x) + \dots + a_m \varphi_m(x) \quad (2.1)$$

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

Aproksymacja określona wzorem (2.1) nosi nazwę *aproksymacji liniowej* (wzór (2.1) nazywa się także *wielomianem uogólnionym*). W obliczeniach na maszynach cyfrowych duże znaczenie ma także *aproksymacja wymierna*, określona wzorem

$$F(x) = \frac{a_0 \varphi_0(x) + a_1 \varphi_1(x) + \dots + a_n \varphi_n(x)}{b_0 \psi_0(x) + b_1 \psi_1(x) + \dots + b_m \psi_m(x)} \quad (2.2)$$

gdzie $\varphi_i(x)$ i $\psi_j(x)$ ($i = 0, 1, \dots, n, j = 0, 1, \dots, m$) są elementami tej samej bazy k -wymiarowej podprzestrzeni liniowej ($k = \max(n, m)$), natomiast a_i, b_j są stałymi współczynnikami, które należy wyznaczyć.

Zajmiemy się teraz zagadnieniem wyboru odpowiedniej podprzestrzeni X_m i związanej z nią bazy (funkcji bazowych $\varphi_k(x)$). Jeżeli funkcja aproksymowana $f(x)$ jest funkcją ciągłą na przedziale $\langle a; b \rangle$ (jest elementem nieskończenie wymiarowej przestrzeni liniowej $C\langle a; b \rangle$), to funkcje $\varphi_k(x)$ będą elementami pewnej $m + 1$ wymiarowej podprzestrzeni $C\langle a; b \rangle$.

Jedną z często obieranych podprzestrzeni X_m jest podprzestrzeń funkcji trygonometrycznych z bazą

$$1, \sin x, \cos x, \sin 2x, \cos 2x, \dots, \sin kx, \cos kx \quad (2.3)$$

szczególnie przydatna, gdy aproksymowana funkcja $f(x)$ jest funkcją okresową. Inną, też często obieraną podprzestrzenią, prowadzącą do prostszej postaci funkcji aproksymującej $F(x)$, jest podprzestrzeń wielomianów stopnia co najwyżej m z bazą jednomianów

$$1, x, x^2, x^3, \dots, x^m \quad (2.4)$$

Interpolacja, omawiana w rozdziale 1, jest oczywiście także jedną z metod aproksymacji funkcji (nazywana jest ona często *aproksymacją interpolacyjną*). Rozpatrywaliśmy tam m.in. bazę jednomianów (2.4), która prowadziła do układu równań liniowych (1.4) z wyznacznikiem Vandermonde'a (1.5). Mimo prostoty działań na wielomianach baza ta ma jednak istotną wadę, jaką jest wrażliwość na błędy zaokrągleń; kumulujące się błędy w przypadku działań na małych oraz niewiele różniących się liczbach mogą całkowicie zniekształcić obliczenia. Przy interpolacji wzorami Newtona używaliśmy jako bazy wielomianów

$$\begin{aligned}
\varphi_0(x) &= 1 \\
\varphi_1(x) &= (x - x_0) \\
\varphi_2(x) &= (x - x_0)(x - x_1) \\
&\dots\dots\dots \\
\varphi_m(x) &= (x - x_0)(x - x_1)\dots(x - x_{m-1})
\end{aligned}$$

a przy interpolacji wzorem Lagrange'a wielomianów $\Phi_j(x)$ (zależności (1.11) i (1.13)).

Przekonaliśmy się w poprzednim rozdziale, że chociaż wszystkie te metody prowadzą w końcu do otrzymania tego samego wielomianu interpolacyjnego, wybór odpowiedniej bazy ma istotne znaczenie dla obliczeń wykonywanych na maszynie cyfrowej wpływając na ich koszt i dokładność.

Zajmiemy się teraz sposobem określania i doboru współczynników a_k we wzorze (2.1). Od kłopotów związanych z wrażliwością na błędy można częściowo się uwolnić wybierając funkcje bazowe mało na nie wrażliwe. W podprzestrzeni wielomianów stopnia co najwyżej m , określonych na przedziale $\langle -1; 1 \rangle$, bazą mogą być np. wielomiany Czebyszewa (patrz p. 1.2.3).

$$T_0, T_1(x), T_2(x), \dots, T_m(x) \quad (2.5)$$

czy też wielomiany Legendre'a [73]

$$L_0, L_1(x), L_2(x), \dots, L_m(x) \quad (2.6)$$

Przyjmijmy, że w przestrzeni X jest określona norma $\|\cdot\|$. Mówimy, że funkcja $F(x)$ dobrze przybliża funkcję $f(x)$, jeżeli norma

$$\|F(x) - f(x)\| \quad (2.7)$$

jest mała. Normą taką może być na przykład:

norma Czebyszewa

$$\|f\| = \sup_{\langle a; b \rangle} |f(x)|$$

norma L_2

$$\|f\|_2 = \left(\int_a^b |f(x)|^2 dx \right)^{\frac{1}{2}}$$

norma L_2 z wagą

$$\|f\|_{2,w} = \left(\int_a^b w(x) |f(x)|^2 dx \right)^{\frac{1}{2}}$$

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

Gdy funkcja $f(x)$ jest określona na dyskretnym zbiorze punktów, wówczas można dane wartości $f(x_i)$ przyjąć za współrzędne wektora f

i wówczas rozpatrujemy normę

$$\|f\| = \left(\sum_{i=0}^n [f(x_i)]^2 \right)^{\frac{1}{2}}$$

Zagadnienie najlepszej aproksymacji przy wybranych funkcjach bazowych $\varphi_k(x)$ sprowadza się do znalezienia wartości współczynników a_k takich, aby otrzymać minimum wyrażenia

$$\|f(x) - (a_0 \varphi_0(x) + \dots + a_m \varphi_m(x))\| \quad (2.8)$$

i aby istniało jedyne możliwe rozwiązanie tego zagadnienia ze względu na a_k .

Zajmiemy się w tym rozdziale dwoma rodzajami aproksymacji: aproksymacją średniokwadratową i aproksymacją jednostajną.

Aproksymacja średniokwadratowa. Dla funkcji $f(x)$ określonej na przedziale $\langle a; b \rangle$ poszukujemy minimum całki

$$\|F(x) - f(x)\| = \int_a^b w(x) [F(x) - f(x)]^2 dx \quad (2.9)$$

a dla funkcji $f(x)$ danej na dyskretnym zbiorze argumentów poszukujemy minimum sumy (*metoda najmniejszych kwadratów*)

$$\|F(x) - f(x)\| = \sum_{i=0}^n w(x_i) [F(x_i) - f(x_i)]^2 \quad (2.10)$$

$$w(x_i) \geq 0 \quad \text{dla} \quad i = 0, 1, \dots, n$$

Aproksymacja jednostajna. Dla funkcji $f(x)$ określonej na przedziale $\langle a; b \rangle$ poszukujemy funkcji $F(x)$ dającej najmniejsze maksimum różnicy między $F(x)$ a $f(x)$ na całym przedziale $\langle a; b \rangle$

$$\|F(x) - f(x)\| = \sup_{x \in \langle a; b \rangle} |F(x) - f(x)| \quad (2.11)$$

Zanim zajmiemy się bardziej szczegółowym omówieniem pewnych typowych rodzajów aproksymacji, podamy dwa sformułowane przez Weierstrassa twierdzenia (dowody tych twierdzeń Czytelnik może znaleźć np. u RALSTONA [50] czy ACHIEZERA [1]).

Tw. 1. *Jeżeli funkcja $f(x)$ jest ciągła na skończonym przedziale $\langle a; b \rangle$, to dla każdego ε dodatniego można dobrać takie n , że jest możliwe utworzenie wielomianu $P_n(x)$ stopnia n ($n = n(\varepsilon)$), który spełnia nierówność*

$$|f(x) - P_n(x)| < \varepsilon \quad (2.12)$$

na całym przedziale $\langle a; b \rangle$.

Tw. 2. *Jeżeli funkcja $f(x)$ jest funkcją ciągłą na $\mathbf{R}$ i okresową o okresie 2π , to dla każdego ε dodatniego istnieje wielomian trygonometryczny*

$$S_n(x) = a_0 + \sum_{k=1}^n (a_k \cos kx + b_k \sin kx), \quad n = n(\varepsilon) \quad (2.13)$$

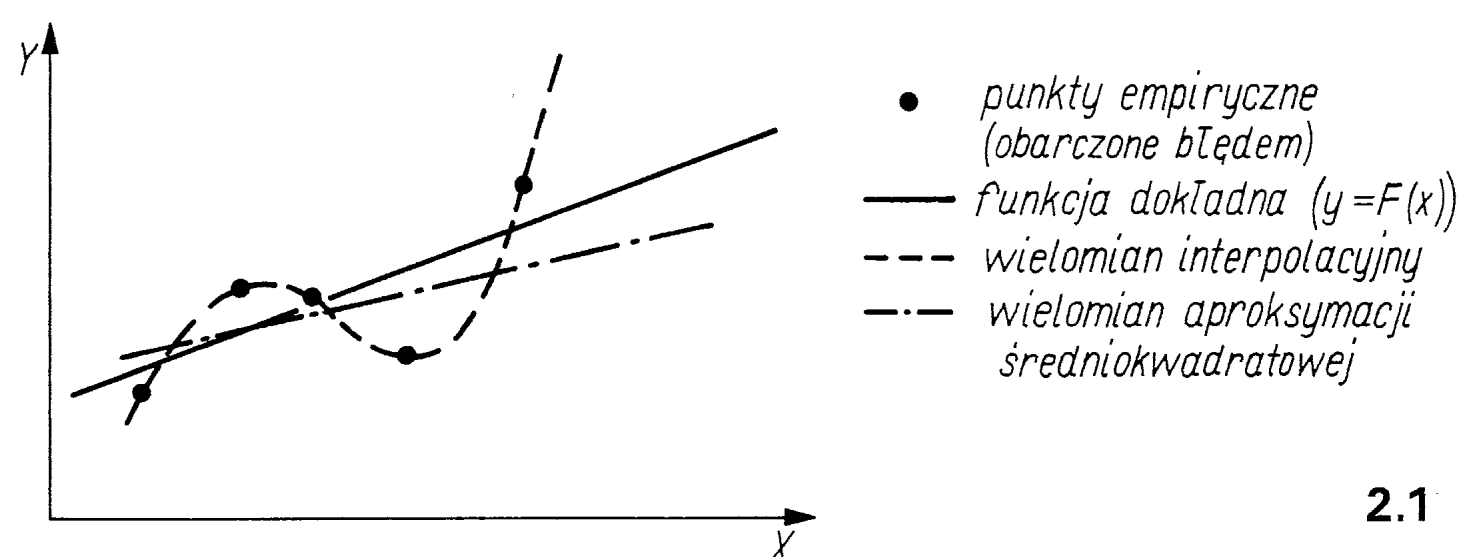
spełniający dla wszystkich x nierówność

$$|f(x) - S_n(x)| < \varepsilon$$

Twierdzenie 1 jest dodatkowym argumentem na poparcie aproksymacji wielomianowych, gdyż gwarantuje ono, że zawsze można znaleźć wielomian o dowolnie małym odchyleniu od funkcji $f(x)$ na przedziale $\langle a; b \rangle$.

2.2. APROKSYMACJA ŚREDNIOKWADRATOWA

W podrozdziale tym omówimy metody aproksymacji średniokwadratowej głównie w przypadku, gdy funkcja aproksymowana $f(x)$ jest znana jedynie na dyskretnym zbiorze argumentów X . Idea takiej aproksymacji jest zilustrowana na rys. 2.1.



Niech będzie dana funkcja $y = f(x)$, która na pewnym zbiorze X punktów $x_0, x_1, x_2, \dots, x_n$ przyjmuje wartości $y_0, y_1, y_2, \dots, y_n$. Wartości te możemy znać tylko w przybliżeniu, z pewnymi błędami (np. jako wyniki pomiarów obarczone błędami obserwacji), przy czym błędy te mogą być niejednokrotnie dosyć duże, co w istotny sposób wpływa na jakość aproksymacji. Będziemy poszukiwać takiej funkcji $F(x)$ przybliżającej daną funkcję $f(x)$, która umożliwi wygładzenie funkcji $f(x)$, tzn. pozwoli z zakłóconych błędami danych wartości funkcji przybliżanej otrzymać gładką funkcję przybliżającą, z dużym prawdopodobieństwem mało odchyłającą się od funkcji przybliżanej zarówno między węzłami, jak i w węzłach $x_0, x_1, \dots, x_n$, jeżeli tylko przyjmiemy, że funkcja przybliżana ma dość gładki przebieg.

Niech $\varphi_j(x), j = 0, 1, \dots, m$, będzie układem funkcji bazowych podprzestrzeni X_n . Poszukujemy wielomianu uogólnionego $F(x)$, będącego najlepszym przybliżeniem średniokwadratowym funkcji $f(x)$ na zbiorze $X = (x_j)$, tj. funkcji postaci

$$F(x) = \sum_{i=0}^m a_i \varphi_i(x) \quad (2.14)$$

przy czym współczynniki a_i są tak określone, aby wyrażenie (2.10) było

minimalne. Oznaczmy

$$H(a_0, a_1, \dots, a_m) = \sum_{j=0}^n w(x_j) \left[f(x_j) - \sum_{i=0}^m a_i \varphi_i(x_j) \right]^2 = \sum_{j=0}^n w(x_j) R_j^2 \quad (2.15)$$

gdzie funkcja wagowa $w(x)$ jest ustalona z góry i taka¹⁾, że $w(x_i) > 0$ dla $i = 0, 1, 2, \dots, n$, a R_j jest odchyleniem w punkcie x_j . Aby znaleźć współczynniki a_i obliczamy pochodne cząstkowe funkcji H względem zmiennych a_i . Z warunku (patrz metody poszukiwania ekstremów funkcji wielu zmiennych np. w [73, rozdz. I, § 8])

$$\frac{\partial H}{\partial a_k} = 0, \quad k = 0, 1, \dots, m \quad (2.16)$$

otrzymamy wówczas układ $m + 1$ równań liniowych z $m + 1$ niewiadomymi a_i

$$\frac{\partial H}{\partial a_k} = -2 \sum_{j=0}^n w(x_j) \left[f(x_j) - \sum_{i=0}^m a_i \varphi_i(x_j) \right] \varphi_k(x_j) = 0 \quad (2.17)$$

$$k = 0, 1, 2, \dots, m$$

zwany *układem normalnym*. Ponieważ funkcje $\varphi_j(x)$ tworzą bazę przestrzeni X_m , więc układ (2.17) ma wyznacznik różny od zera i rozwiązanie tego układu daje minimum funkcji H .

W zapisie macierzowym układ (2.17) przyjmuje postać

$$\mathbf{D}^T \mathbf{D} \mathbf{A} = \mathbf{D}^T \mathbf{f} \quad (2.18)$$

gdzie

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

Macierz współczynników układu (2.18) jest macierzą symetryczną i dodatnio określoną co zapewnia jednoznaczność rozwiązania.

Układ normalny (2.17) lub (2.18) można otrzymać z równania (2.14) po podstawieniu wartości punktów węzłowych x_i ($i = 0, 1, \dots, n$). Otrzymamy wówczas nadokreślony układ $n + 1$ równań z $m + 1$ niewiadomymi

$$\mathbf{D} \mathbf{A} = \mathbf{f} \quad (2.19)$$

z którego po pomnożeniu (lewostronnie) przez $\mathbf{D}^T$ otrzymujemy (2.18).

¹⁾ W następnych punktach będziemy przyjmować, że funkcja wagowa $w(x)$ ma stałą wartość, równą tożsamościowo jedności. W pewnych specjalnych przypadkach dobiera się inne funkcje wagowe. Na przykład, jeżeli chcemy, aby otrzymane przybliżenie było w pewnych punktach lepsze (bowiem wiemy, że w tych punktach wartości funkcji znamy z mniejszym błędem), to przyjmujemy w tych punktach większe wartości funkcji wagowej.

2.2.1. Aproksymacja wielomianowa

Jeżeli jako funkcje bazowe przyjmiemy ciąg jednomianów (x^i) , $i = 0, 1, \dots, m$, to wzór (2.17) przyjmie postać

$$\sum_{j=0}^n \left[f(x_j) - \sum_{i=0}^m a_i x_j^i \right] x_j^k = 0 \quad (2.20)$$

$$k = 0, 1, 2, \dots, m$$

Zmieniając kolejność sumowania, mamy

$$\sum_{j=0}^n f(x_j) x_j^k = \sum_{i=0}^m a_i \left(\sum_{j=0}^n x_j^{i+k} \right) \quad (2.21)$$

$$k = 0, 1, 2, \dots, m$$

i oznaczając

$$g_{ik} = \sum_{j=0}^n x_j^{i+k} \quad (2.22)$$

$$Q_k = \sum_{j=0}^n f(x_j) x_j^k$$

otrzymamy układ normalny (2.17) postaci

$$\sum_{i=0}^m a_i g_{ik} = Q_k, \quad k = 0, 1, 2, \dots, m \quad (2.23)$$

Można wykazać, że jeżeli punkty $x_0, x_1, \dots, x_n$ są różne i $m \leq n$, to wyznacznik układu (2.23) 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 $x_0, x_1, \dots, x_n$ i wówczas $H = 0$. W praktyce stopień wielomianu jest, i powinien być, znacznie mniejszy od liczby punktów x_k ($k = 0, 1, \dots, n$), wtedy bowiem korzystamy z dużej ilości informacji (np. wyników pomiarów) uzyskując równocześnie prostsze (niższego stopnia) funkcje aproksymujące.

Przykład 1. Niech dane będą wyniki pomiarów zestawione w tablicy 2/1. Poszukujemy zależności między x i y postaci

$$ax + by = 1 \quad (2.24)$$

przy czym zadanie polega na znalezieniu „najlepszych” współczynników a oraz b .

2/1

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

Przy tak sformułowanym zadaniu pojawia się problem: czy rozpatrywać odchylenia x , czy y ? Przedyskutujemy obydwa przypadki, pokazując jednocześnie, że prowadzą one do różnych wyników.

Gdy przyjmiemy, że wartości y są obarczone pewnymi błędami, wówczas minimalizujemy wyrażenie

$$\sum_{i=1}^n (y_i - \bar{y}_i)^2 \quad (2.25)$$

gdzie $\bar{y}_i$ są określone za pomocą wzoru (2.24) dla $i = 1, 2, \dots, n$. Jeżeli natomiast przyjmiemy, że wartości x są obarczone błędami, to powinniśmy minimalizować wyrażenie

$$\sum_{i=1}^n (x_i - \bar{x}_i)^2 \quad (2.26)$$

przy czym

$$a\bar{x}_i + by_i = 1, \quad i = 1, 2, \dots, n$$

Wartości x_i i y_i są każdorazowo brane z tablicy 2.1. Możliwe jest także, ale stosunkowo trudne, rozpatrywanie przypadku, gdy zarówno wartości x jak i y są obarczone błędami.

W przypadku minimalizacji wyrażenia (2.25) otrzymujemy układ normalny postaci

$$na_0 + \left(\sum_{i=1}^n x_i \right) a_1 = \sum_{i=1}^n y_i \quad (2.27)$$

$$\left(\sum_{i=1}^n x_i \right) a_0 + \left(\sum_{i=1}^n x_i^2 \right) a_1 = \sum_{i=1}^n x_i y_i$$

w którym $a_1 = -a/b$, $a_0 = 1/b$ i stąd

$$a_0 = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \quad (2.28)$$

$$a_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

Układ normalny (2.27) dla danych z tablicy 2/1 ma rozwiązanie

$$a_0 = 6/11 \quad \text{i} \quad a_1 = 7/11$$

i stąd równanie (2.24) ma postać

$$11y - 7x = 6 \quad (2.29)$$

Minimalizując w taki sam sposób wyrażenie (2.26), otrzymamy równanie

$$2x - 3y = -1 \quad (2.30)$$

Jak widać, równania (2.29) i (2.30) są różne, chociaż ich wykresy prawie się pokrywają.

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 wystar-

czająco 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 . Proces ten prowadzimy tak długo, jak długo ze wzrostem stopnia wielomianu funkcja H maleje w znaczny sposób. Niestety, ta ostatnia metoda ma istotną wadę. Wiadomo, że dla wartości $m \geq 6$ układ (2.23) jest układem źle uwarunkowanym (patrz p. 0.4), wskutek czego otrzymane wyniki obliczeń na maszynach cyfrowych mogą być tak bardzo zaburzone, iż nie nadają się do praktycznego wykorzystania przy aproksymacji.

Aby wyjaśnić problem złego uwarunkowania macierzy współczynników układu (2.23), rozpatrzmy następujący przykład. Niech wszystkie punkty x_i będą rozłożone w jednakowych odstępach w przedziale $\langle 0; 1 \rangle$. Liczby g_{ik} , określone wzorem (2.22), można dla dużych wartości m przybliżyć następująco

$$g_{ik} = \sum_{j=1}^m x_j^{i+k} \approx m \int_0^1 x^{i+k} dx = \frac{m}{i+k+1} \quad (2.31)$$

$$i, k = 0, 1, 2, \dots, m$$

Macierz współczynników układu (2.23), po podzieleniu przez m , możemy przybliżyć macierzą

$$A = \begin{bmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{m+1} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{m+2} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{m+1} & \frac{1}{m+2} & \cdots & \frac{1}{2m+1} \end{bmatrix} \quad (2.32)$$

Macierz odwrotna A^{-1} dla $m \geq 9$ ma elementy rzędu co najmniej $3 \cdot 10^{12}$, co powoduje, że przy rozwiązywaniu układu (2.32) błędy zaokrągleń są tak duże, iż wyniki praktycznie tracą sens.

Powyższy przykład przemawia przeciwko stosowaniu aproksymacji z funkcjami bazowymi typu jednomianów x^i z wyjątkiem przypadków, gdy m jest bardzo małe.

Przykład 2. Aby zilustrować kłopoty w przypadku aproksymacji wielomianami wyższych stopni rozpatrzmy następujący przykład. W tablicy 2.2 zestawiono wartości zmiennych x i y (zostały one wygenerowane jako tablica wartości wielomianu)

$$y = 40 + 10x + 5x^2 + 3x^3 + 2x^4 + x^5 + x^6$$

2/2

x	y	x	y
1	62	8	305080
2	232	9	606334
3	1330	10	1123640
4	5984	11	1966642
5	20590	12	3282352
6	57952	13	5262830
7	140642	14	8153584

Współczynniki układu normalnego utworzonego na podstawie tablicy 2/2 zmieniają się od 14 aż do 10^{14} . Wielomian aproksymacyjny stopnia szóstego, otrzymany po rozwiązaniu na maszynie cyfrowej, ma postać

$$\bar{y} = 7434,164 - 10135,4x + 4422,734x^2 - 856,697x^3 + 85,02283x^4 - 2,888804x^5 + 1,070349x^6$$

i jak widać bardzo się różni od wielomianu użytego do generowania danych. Różnice między wartościami y i $\bar{y}$ dla $x = 1, 2, \dots, 14$ wynoszą od 48 do 886.

Gdy obliczenia będziemy wykonywać ze zwiększoną dokładnością, wówczas

$$\bar{y} = 47,12713 + 8,56226x + 5,066897x^2 + 3,005381x^3 + 1,999317x^4 + 1,000023x^5 + 1,000000x^6$$

Ten wynik jest dużo lepszy, a różnice między y a $\bar{y}$ mają wartości od zera do 5,7.

Aby można było aproksymować wielomianami wyższych stopni, należy zastosować inną metodę rozwiązywania układu normalnego. W rozdziale 5 są podane pewne wskazówki dotyczące metod rozwiązywania układów równań, w których macierz współczynników ma wyznacznik bliski zeru. Można stosować przekształcenia Householdera albo inną z metod rozkładu macierzy na iloczyn **QR** (patrz p. 6.4.7), można także skorzystać z metody Goluba [30].

Innym sposobem umożliwiającym aproksymację wielomianami wyższych stopni jest zmiana bazy podprzestrzeni X_m przez zastąpienie bazy jednomianów (2.4) bazą złożoną z wielomianów ortogonalnych (np. (2.6)).

Przykład 3. Napisać program na aproksymację wielomianem stopnia trzeciego dla n węzłów funkcji.

Układ normalny (2.23) ma postać

$$\begin{aligned} na_0 + (\sum x_i) a_1 + (\sum x_i^2) a_2 + (\sum x_i^3) a_3 &= \sum y_i \\ (\sum x_i) a_0 + (\sum x_i^2) a_1 + (\sum x_i^3) a_2 + (\sum x_i^4) a_3 &= \sum x_i y_i \\ (\sum x_i^2) a_0 + (\sum x_i^3) a_1 + (\sum x_i^4) a_2 + (\sum x_i^5) a_3 &= \sum x_i^2 y_i \\ (\sum x_i^3) a_0 + (\sum x_i^4) a_1 + (\sum x_i^5) a_2 + (\sum x_i^6) a_3 &= \sum x_i^3 y_i \end{aligned}$$

gdzie wszystkie sumowania są od $i = 1$ do $i = n$. Wprowadźmy oznaczenia

$$S(1) = \sum x_i, \quad S(2) = \sum x_i^2, \quad \dots, \quad S(6) = \sum x_i^6$$

oraz

$$P(1) = \sum y_i, \quad P(2) = \sum x_i y_i, \quad P(3) = \sum x_i^2 y_i, \quad P(4) = \sum x_i^3 y_i$$

Program w pętli wczytuje kolejno punkt po punkcie (x i y) obliczając jednocześnie wartości wyrażeń $S(1)$ do $S(6)$ i $P(1)$ do $P(4)$. Po wczytaniu wszystkich n punktów zostaje zbudowana

dwuwymiarowa tablica, będąca macierzą współczynników układu normalnego (TAB). Następuje wywołanie procedury RUKL, która rozwiązuje układ umieszczając w tablicy A wartości pierwiastków będące szukanymi współczynnikami a_i ($i = 0, 1, 2, 3$). Procedura RUKL może być dowolną procedurą rozwiązywania układu n równań liniowych z n niewiadomymi. Ponieważ macierz układu normalnego jest symetryczna (i dodatnio określona), najbardziej efektywne byłyby procedury wykorzystujące metodę Cholesky'ego-Banachiewicza (rozkładu $A = LL^T$) albo metodę rozkładu $A = LDL^T$ (patrz p. 5.3.5), albo dowolna inna procedura wykorzystująca np. metodę eliminacji Gaussa (p. 5.3.5). Procedury takie znajdują się w standardowej bibliotece programów każdej maszyny cyfrowej.

```

PROGRAM APROKSYMACJA;
VAR i,j,n,k: integer;
    x,y: real;
    TAB: ARRAY [1..4,1..5] OF real;
    A,P: ARRAY [1..4] OF real;
    S,SUMA: ARRAY [1..6] OF real;
PROCEDURE RUKL (we,wy: pointer; wymx,wymy: integer);
.....
BEGIN {aproksymacja}
  write('podaj liczbe wezlow');
  readln(n:5);
  FOR j := 1 to 4 DO
    BEGIN
      A[j] := 0.0;
      P[j] := 0.0;
      FOR k := 1 TO 5 DO
        TAB[j,k] := 0.0
      END;
    END;
  FOR j := 1 TO 6 DO
    BEGIN
      S[j] := 0.0;
      SUMA[j] := 0.0
    END;
  FOR j := 1 TO n DO
    BEGIN
      write('podaj wartosci x i y dla kolejnego wezla');
      readln(x:10:4, y:10:4);
      S[1] := x;
      FOR i := 2 TO 6 DO S[i] := x * S[i-1];
      FOR i := 1 TO 6 DO SUMA[i] := SUMA[i] + S[i];
      P[1] := P[1] + y;
      FOR i := 2 TO 4 DO P[i] := P[i] + y * S[i-1]
      END;
      TAB[1,1] := n;
      FOR i := 1 TO 4 DO
        BEGIN
          TAB[i,5] := P[i];
          FOR j := 1 TO 4 DO
            BEGIN
              k := i + j;
              IF NOT (k=2) THEN TAB[i,j] := SUMA[k-2]
            END;
          END;
        END;
      RUKL ( TAB, A, 4, 5);
      writeln('poszukiwane wspolczynniki');
      FOR i := 1 TO 4 DO writeln (A[i]:12:6)
    END.

```

Ćwiczenia

1. Dla funkcji $y = f(x)$, określonej tablicą 2/3, znaleźć wielomian aproksymujący stopnia pierwszego.

2. Dla funkcji $y = f(x)$, określonej tablicą 2/4, znaleźć wielomian aproksymujący stopnia pierwszego.

3. Funkcja $y = f(x)$ jest określona tablicą 2/5. Znaleźć wielomian aproksymujący stopnia drugiego.

4. Funkcja $x = f(x)$ jest określona tablicą 2/6 (jak widać jest to tablica wartości funkcji $y = \sin x$).

a) Znaleźć wielomian aproksymujący stopnia drugiego i określić średni błąd aproksymacji.

b) Znaleźć wielomian aproksymujący stopnia czwartego.

c) Zwiększyć liczbę punktów do 17 (równoodległych): $x_k = k \frac{\pi}{32}$ ($k = 0, 1, \dots, 16$),

a następnie znaleźć dwa wielomiany aproksymujące stopnia drugiego odpowiednio na przedziałach $\left\langle 0; \frac{\pi}{4} \right\rangle$ i $\left\langle \frac{\pi}{4}; \frac{\pi}{2} \right\rangle$.

2/3

x	0	0.1	0.2	0.3	0.4	0.5	0.6
y	2.9	2.8	2.7	2.3	2.1	2.1	1.7

2/4

x	20.5	32.7	51.0	73.2	95.7
y	765	826	873	942	1032

2/5

x	0	0.2	0.4	0.6	0.8	1.0
y	1.026	0.768	0.648	0.401	0.272	0.193

2/6

x	0	$\pi/6$	$\pi/4$	$\pi/3$	$\pi/2$
y	0	1/2	$\sqrt{2}/2$	$\sqrt{3}/2$	1

Odpowiedzi. 1. $y = 2,97 - 2x$. 2. $y = 702,2 + 3,395x$. 3. $y = 0,999 - 1,006x + 0,21x^2$. 4. a) $y = -0,00495 + 1,168485x + 0,33462101x^2$, $\varepsilon_1 \approx 0,00982$; b) $y = 0,000122 + 0,996146x + 0,019752x^2 + 0,20319x^3 + 0,028549x^4$; $\varepsilon_2 \approx 0,00008$; c) $y = -0,002594 + 1,05517x - 0,188892x^2$ dla $x \in \left\langle 0; \frac{\pi}{4} \right\rangle$ lub $y = -0,166127 + 1,473219x - 0,465035x^2$ dla $x \in \left\langle \frac{\pi}{4}; \frac{\pi}{2} \right\rangle$; $\varepsilon_3 \approx 0,00133$.

2.2.2. Aproksymacja za pomocą wielomianów ortogonalnych

Jak wykazaliśmy w poprzednim punkcie, przy aproksymacji wielomianowej z funkcjami bazowymi (2.4) ze wzrostem stopnia wielomianu obliczenia stają się coraz bardziej pracochłonne, a ponadto ich wyniki są niepewne. Oczywiście, gdy do obliczeń wykorzystujemy maszynę cyfrową, dokładność wyników możemy w jakimś stopniu poprawić, na przykład przez zwiększenie liczby cyfr znaczących (zob. przykł. 2 w p. 2.2.1), lecz i ten sposób ma ograniczony zakres. Dodatkową trudnością jest fakt, że zmiana stopnia wielomianu przybliżającego wymaga ponownego rozwiązania układu normalnego, co też przemawia przeciwko stosowaniu ciągu funkcji $\varphi_i(x) = x^i$ do aproksymacji.

Obydwie te trudności można usunąć używając do aproksymacji wielomianów ortogonalnych¹⁾.

W punkcie tym omówimy problem aproksymacji za pomocą funkcji określonych na dyskretnym zbiorze punktów, a następnie aproksymację funkcji ciągłych.

Zdefiniujemy zatem ortogonalność funkcji na dyskretnym zbiorze punktów.

Def. Funkcje $f(x)$ i $g(x)$ nazywamy *ortogonalnymi na dyskretnym zbiorze punktów* $x_0, x_1, \dots, x_n$, jeśli

$$\sum_{i=0}^n f(x_i) g(x_i) = 0 \quad (2.33)$$

przy czym

$$\sum_{i=0}^n [f(x_i)]^2 > 0, \quad \sum_{i=0}^n [g(x_i)]^2 > 0$$

Analogicznie, ciąg funkcyjny

$$(\varphi_m(x)) \equiv \varphi_0(x), \varphi_1(x), \dots, \varphi_m(x), \dots \quad (2.34)$$

nazywamy *ortogonalnym na zbiorze punktów* $x_0, x_1, \dots, x_n$, jeśli

$$\sum_{i=0}^n \varphi_j(x_i) \varphi_k(x_i) = 0 \quad \text{dla } j \neq k \quad (2.35)$$

Zakładamy, że nie wszystkie punkty x_i są miejscami zerowymi rozpatrywanych funkcji $\varphi_j(x)$, tzn. $\sum_{i=0}^n \varphi_j^2(x_i) > 0$.

Zastosowanie do aproksymacji wielomianów ortogonalnych usuwa jedną z podstawowych trudności obliczeniowych pojawiających się przy aproksymacji wielomianami algebraicznymi — złe uwarunkowanie macierzy układu normalnego. Przy aproksymacji wielomianami ortogonalnymi macierz układu normalnego jest macierzą diagonalną, a jej elementy położone na głównej przekątnej wynoszą

$$a_{jj} = \sum_{i=0}^n \varphi_j^2(x_i)$$

Założmy, że mamy $n + 1$ równoodległych punktów x_i ($x_i = x_0 + ih$, $i = 0, 1, \dots, n$). Za pomocą przekształcenia liniowego

$$q = \frac{x - x_0}{h}$$

¹⁾ Pojęcie funkcji ortogonalnych oraz ortogonalnego ciągu funkcyjnego na przedziale $\langle a; b \rangle$ (dla przypadku funkcji całkowalnych w sensie Riemanna w tym przedziale) może Czytelnik znaleźć w [73, rozdz. III, § 11].

przeprowadzimy te punkty odpowiednio w kolejne liczby całkowite od 0 do n . Poszukujemy ciągu wielomianów

$$(F_i^{(n)}(q)) \equiv F_0^{(n)}(q), F_1^{(n)}(q), \dots, F_m^{(n)}(q) \quad (2.36)$$

(dolny indeks oznacza stopień, $m \leq n$) spełniających na zbiorze punktów $q = 0, 1, \dots, n$ warunek ortogonalności

$$\sum_{i=0}^n F_j^{(n)}(i) F_k^{(n)}(i) = 0 \quad \text{dla } j \neq k \quad (2.37)$$

przy czym

$$F_k^{(n)}(q) = a_0 + a_1 q + a_2 q(q-1) + a_3 q(q-1)(q-2) + \dots \\ \dots + a_k q(q-1) \dots (q-k+1) \quad (2.38)$$

W celu uproszczenia zapisu można użyć tutaj wprowadzonego w p. 1.2.6 pojęcia wielomianu czynnikowego

$$q^{[k]} = q(q-1) \dots (q-k+1)$$

Wówczas

$$F_k^{(n)}(q) = a_0 + a_1 q^{[1]} + a_2 q^{[2]} + \dots + a_k q^{[k]}$$

Często używa się *unormowanego* ciągu wielomianów (2.36), spełniających warunek

$$\hat{F}_k^{(n)}(0) = 1, \quad k = 0, 1, \dots, m \quad (2.39)$$

$$\hat{F}_k^{(n)}(q) = 1 + b_1 q^{[1]} + \dots + b_k q^{[k]}$$

Po stosunkowo prostych przekształceniach [18], [50], otrzymujemy związek

$$\hat{F}_k^{(n)}(q) = \sum_{s=0}^k (-1)^s \binom{k}{s} \binom{k+s}{s} \frac{q^{[s]}}{n^{[n]}} \quad (2.40)$$

$$k = 0, 1, 2, \dots, m$$

gdzie $\binom{k}{s}$ są symbolami Newtona. Wzór (2.40) określa *wielomiany Grama* stopni $k = 0, 1, 2, \dots, m$ dla $n+1$ równoodległych węzłów.

U w a g a. Wielomiany Grama zwane są również niekiedy wielomianami Czebyszewa, co prowadzi do nieporozumień.

Wzór aproksymacyjny oparty na wielomianach Grama, będący odpowiednikiem wzoru (2.14), ma postać

$$y_m(x) = \sum_{k=0}^m \frac{c_k}{s_k} \hat{F}_k^{(n)}(q) = \sum_{k=0}^m \frac{c_k}{s_k} \hat{F}_k^{(n)}\left(\frac{x-x_0}{h}\right) \quad (2.41)$$

dla $m \leq n$, gdzie

$$s_k = \sum_{q=0}^n [\hat{F}_k^{(n)}(q)]^2$$

$$c_k = \sum_{i=0}^n y_i \hat{F}_k^{(n)}(x_i)$$
(2.42)

Przykład 1. Dana jest tablica wartości funkcji $f(x)$ (tabl. 2/7). Znaleźć najlepszy, w sensie aproksymacji średniokwadratowej, wielomian przybliżający funkcję $f(x)$.

2/7

x	1	1.5	2	2.5	3
y	3	4.75	7	9.75	13

Ponieważ węzły są równoodległe, posłużymy się wielomianami Grama. Konstruujemy wielomian aproksymacyjny dla $n = 4$. Wartości $F_k(q)$, $q = 0, 1, 2, 3, 4$ obliczamy ze wzoru (2.40)

$$F_0(q) = 1, \quad F_1(q) = 1 - 2 \cdot \frac{q}{n}, \quad n \geq 1$$

$$F_2(q) = 1 - 6 \cdot \frac{q}{n} + 6 \cdot \frac{q(q-1)}{n(n-1)}, \quad n \geq 2$$

$$F_3(q) = 1 - 12 \cdot \frac{q}{n} + 30 \cdot \frac{q(q-1)}{n(n-1)} - 20 \cdot \frac{q(q-1)(q-2)}{n(n-1)(n-2)}, \quad n \geq 3$$

$$F_4(q) = 1 - 20 \cdot \frac{q}{n} + 90 \cdot \frac{q(q-1)}{n(n-1)} - 140 \cdot \frac{q(q-1)(q-2)}{n(n-1)(n-2)} +$$

$$+ 70 \cdot \frac{q(q-1)(q-2)(q-3)}{n(n-1)(n-2)(n-3)}, \quad n \geq 4$$

i zestawiamy w postaci tablicy (tabl. 2/8). Korzystając następnie ze wzorów (2.42), obliczamy wartości c_i , s_i oraz $b_i = c_i/s_i$ (tabl. 2/9) i ze wzoru (2.14) otrzymujemy

$$\left(\text{dla } q = \frac{x-1}{0,5} \right)$$

$$y(x) = \sum_{i=0}^4 b_i F_i(q) = x^2 + x + 1$$

2/8

q	x_i	y_i	$F_0(q)$	$F_1(q)$	$F_2(q)$	$F_3(q)$	$F_4(q)$
0	1	3	1	1	1	1	1
1	1.5	4.75	1	0.5	-0.5	-2	-4
2	2	7	1	0	-1	0	6
3	2.5	9.75	1	-0.5	-0.5	2	-4
4	3	13	1	-1	1	-1	1

2/9

C_i	37.5	-12.5	1.75	0	0
S_i	5	2.5	3.5	10	70
b_i	7.5	-5	0.5	0	0

Jeżeli punkty, w których są dane wartości funkcji nie są równoodległe, to korzystamy z pewnych związków rekurencyjnych [50, rozdz. VI, p. 4]. Niech mianowicie $(\varphi_j(x))$ będzie dowolnym ciągiem wielomianów ortogonalnych na zbiorze punktów x_i . Można udowodnić (np. indukcyjnie), że zachodzi związek

$$\varphi_{j+1}(x) = (x - \alpha_{j+1}) \varphi_j(x) - \beta_j \varphi_{j-1}(x) \quad \text{dla } j = 0, 1, \dots \quad (2.43)$$

$$\varphi_0(x) = 1, \quad \varphi_{-1}(x) = 0$$

gdzie stałe α_{j+1} i β_j są określone wzorami

$$\beta_j = \frac{\sum_{i=0}^n \varphi_j^2(x_i)}{\sum_{i=0}^n \varphi_{j-1}^2(x_i)} \quad (2.44)$$

$$\alpha_{j+1} = \frac{\sum_{i=0}^n x_i \varphi_j^2(x_i)}{\sum_{i=0}^n \varphi_j^2(x_i)}$$

Odpowiednikami wzorów (2.41) i (2.42) są teraz

$$y_m(x) = \sum_{k=0}^m b_k \varphi_k(x) \quad (2.45)$$

$$b_k = \frac{C_k}{S_k}$$

$$C_k = \sum_{i=1}^n y_i \varphi_k(x_i) \quad (2.46)$$

$$S_k = \sum_{i=0}^n \varphi_k^2(x_i)$$

Aproksymacja średniokwadratowa funkcji ciągłych. Przyjmijmy, że daną funkcję $f(x)$ ciągłą na przedziale $\langle a; b \rangle$ mamy aproksymować funkcję postaci (2.1)

$$P(x) = a_0 \varphi_0(x) + a_1 \varphi_1(x) + \dots + a_n \varphi_n(x)$$

gdzie $\varphi_0(x), \varphi_1(x), \dots, \varphi_n(x)$ są elementami bazy pewnej podprzestrzeni funkcji całkownych z kwadratem na przedziale $\langle a; b \rangle$. Podobnie jak w przypadku, gdy $f(x)$ była określona na dyskretnym zbiorze punktów, aproksymacja średniokwadratowa funkcji ciągłych polega na znalezieniu takiego ciągu współczynników $a_i (i = 0, 1, 2, \dots, n)$, aby otrzymać minimum normy (2.9). Przyjmijmy zatem, tak jak w przypadku aproksymacji dyskretnej, że funkcja wagowa jest stała i równa jedności

$$H_n = \int_a^b [P(x) - f(x)]^2 dx = \int_a^b \left[\sum_{i=0}^n a_i \varphi_i(x) - f(x) \right]^2 dx \quad (2.47)$$

Jak poprzednio, aby rozwiązać zadanie obliczamy pochodne cząstkowe $\frac{\partial H_n}{\partial a_i}$, $i = 0, 1, \dots, n$, przyrównujemy je do zera i z otrzymanego układu równań wyznaczamy współczynniki a_i .

Przykład 2. Znaleźć przybliżenie funkcji $f(x) = \sin x$ na przedziale $\langle 0; \pi/2 \rangle$ wielomianem stopnia co najwyżej drugiego $P(x) = a_0 + a_1x + a_2x^2$.

Układ równań ma postać

$$\begin{aligned} a_0 \int_0^{\pi/2} dx + a_1 \int_0^{\pi/2} x dx + a_2 \int_0^{\pi/2} x^2 dx &= \int_0^{\pi/2} \sin x dx \\ a_0 \int_0^{\pi/2} x dx + a_1 \int_0^{\pi/2} x^2 dx + a_2 \int_0^{\pi/2} x^3 dx &= \int_0^{\pi/2} x \sin x dx \\ a_0 \int_0^{\pi/2} x^2 dx + a_1 \int_0^{\pi/2} x^3 dx + a_2 \int_0^{\pi/2} x^4 dx &= \int_0^{\pi/2} x^2 \sin x dx \end{aligned}$$

czyli

$$\frac{\pi}{2} a_0 + \frac{\pi^2}{8} a_1 + \frac{\pi^3}{24} a_2 = 1$$

$$\frac{\pi^2}{8} a_0 + \frac{\pi^3}{24} a_1 + \frac{\pi^4}{64} a_2 = 1$$

$$\frac{\pi^3}{24} a_0 + \frac{\pi^4}{64} a_1 + \frac{\pi^5}{160} a_2 = -2$$

Rozwiązaniem powyższego układu są liczby

$$a_0 \approx 0,1338341, \quad a_1 \approx 0,5916223, \quad a_2 \approx 0,0463565$$

Zatem

$$P(x) \approx 0,134 + 0,59x + 0,05x^2$$

Obliczamy średni błąd aproksymacji

$$\begin{aligned} M^2 &= (b-a)^{-1} H_n(a_0, \dots, a_n) = \frac{2}{\pi} \int_0^{\pi/2} [\sin x - (a_0 + a_1x + a_2x^2)]^2 dx \approx \\ &\approx 0,00797, \text{ czyli } M \approx 0,089. \end{aligned}$$

Rachunki będą znacznie prostsze, jeżeli zamiast ciągu $\varphi_i(x) = x^i$ użyjemy ciągu funkcji ortogonalnych z wagą. Może to być na przykład omówiony w [73] ciąg wielomianów Legendre'a

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n, \quad n = 0, 1, 2, \dots \quad (2.48)$$

Przykład 3. Aproksymować funkcję z poprzedniego przykładu posługując się wielomianami Legendre'a.

Ponieważ na przedziale $\langle -1; 1 \rangle$ wielomiany Legendre'a tworzą ciąg ortogonalny z wagą 1, a funkcję $f(x) = \sin x$ mamy aproksymować na przedziale $\langle 0; \pi/2 \rangle$, wprowadzamy nową zmienną

$$t = \frac{4}{\pi}x - 1 \quad (2.49)$$

której wartości należą do przedziału $\langle -1; 1 \rangle$ dla $x \in \langle 0; \pi/2 \rangle$. Aproksymować będziemy zatem funkcję

$$f_1(t) = \sin \frac{\pi(t+1)}{4}$$

wielomianem

$$W(t) = a_0 P_0(t) + a_1 P_1(t) + a_2 P_2(t)$$

gdzie $P_0(t)$, $P_1(t)$ i $P_2(t)$ są wielomianami Legendre'a.

Odpowiednie współczynniki wynoszą

$$a_0 = \frac{1}{2} \int_{-1}^1 1 \cdot \sin \frac{\pi}{4}(t+1) dt = \frac{2}{\pi}$$

$$a_1 = \frac{3}{2} \int_{-1}^1 t \cdot \sin \frac{\pi}{4}(t+1) dt = \frac{24}{\pi^2} - \frac{6}{\pi}$$

$$a_2 = \frac{5}{2} \int_{-1}^1 \left(\frac{3}{2}t^2 - \frac{1}{2} \right) \sin \frac{\pi}{4}(t+1) dt = -\frac{480}{\pi^3} + \frac{120}{\pi^2} + \frac{10}{\pi}$$

Zatem

$$W(t) \approx 0,6366197 + 0,5218492x - 0,1390961x^2$$

Średni błąd aproksymacji

$$M^2 = \frac{1}{2} \left[\int_{-1}^1 \sin^2 \frac{\pi}{4}(t+1) dt - \|P_0\| a_0^2 - \|P_1\| a_1^2 - \|P_2\| a_2^2 \right] \approx$$

$$\approx \frac{1}{2} [1 - (0,8105692 + 0,181551 + 0,007739)] \approx 0,0000704$$

$$M \approx 0,00839$$

Jak widać, skutek użycia do aproksymacji wielomianów ortogonalnych uzyskaliśmy ponad dziesięciokrotne zmniejszenie średniego błędu aproksymacji.

Do aproksymacji używa się także układów funkcji ortogonalnych z wagą $w(x)$ będącą daną, dodatnią funkcją na przedziale $\langle a; b \rangle$. Warunek ortogonalności ma wówczas postać

$$\int_a^b w(x) \varphi_k(x) \varphi_n(x) dx = 0 \quad \text{dla} \quad k \neq n \quad (2.50)$$

a w wielomianie uogólnionym (2.1) dobieramy współczynniki a_i ($i = 0, 1, 2, \dots, n$) tak, aby norma (2.9) była najmniejsza dla danego ustalonego n .

Przykładem ortogonalnego układu funkcji z wagą na przedziale $\langle -1; 1 \rangle$ są wielomiany Czebyszewa (wzory (1.28), (1.29)) z funkcją wagową

$$w(x) = \frac{1}{\sqrt{1-x^2}} \quad (2.51)$$

Ćwiczenia

1. Funkcję $y = \sin x$ określoną na przedziale $\langle 0; \frac{\pi}{2} \rangle$ aproksymować wielomianem stopnia czwartego przy użyciu bazy wielomianów Legendre'a. Wskazówka. Wprowadzić nową zmienną, jak w przykładzie 3.

2. Funkcję $y = |x|$ aproksymować na przedziale $\langle -1; 1 \rangle$: a) wielomianem stopnia trzeciego, b) wielomianem stopnia piątego. Wskazówka. Szukamy $W_n(x) = \sum_{i=0}^n a_i P_i(x)$, gdzie $P_i(x)$ są odpowiednimi wielomianami Legendre'a.

Odpowiedzi. 1. $g(t) = \sum_{i=0}^4 a_i P_i(t)$, gdzie $P_i(t)$ — wielomiany Legendre'a; $a_0 \approx 0,6366$, $a_1 \approx 0,5218$, $a_2 \approx -0,1391$, $a_3 \approx -0,0221$, $a_4 \approx 0,0025$. 2. b) $W_5(x) = \frac{15}{128}(-7x^4 + 14x^2 + 1)$.

2.2.3. Aproksymacja trygonometryczna

W zagadnieniach aproksymacji często spotykamy się z przypadkiem, gdy funkcja $f(x)$ jest okresowa. Taką funkcję wygodniej i lepiej jest aproksymować nie wielomianami algebraicznymi, a wielomianami trygonometrycznymi (baza ((2.3))). Jeżeli $f(x)$ jest funkcją ciągłą o okresie 2π , wielomian trygonometryczny ma postać

$$Q_n(x) = \frac{a_0}{2} + \sum_{k=1}^n (a_k \cos kx + b_k \sin kx) \quad (2.52)$$

gdzie a_i, b_i są współczynnikami trygonometrycznymi Fouriera funkcji $f(x)$ względem ortogonalnego układu funkcji bazowych (2.3). Dokładne omówienie metody, wzory oraz różne szczególne przypadki Czytelnik może znaleźć w [73, rozdz. III, p. 11].

Bardziej interesujący jest przypadek, gdy funkcja $f(x)$ jest określona na dyskretnym zbiorze punktów. Przyjmijmy, że dane punkty są równoodległe, a ponadto, że ich liczba jest parzysta i wynosi $2L$. (W przypadku nieparzystej liczby punktów rozumowanie jest analogiczne). Niech

$$x_i = \frac{\pi i}{L} \quad \text{dla } i = 0, 1, \dots, 2L - 1 \quad (2.53)$$

Zbiór funkcji (2.3) jest ortogonalny nie tylko na przedziale $\langle 0; 2\pi \rangle$, ale i na

dyskretnym zbiorze (2.53), przy czym

$$\sum_{i=0}^{2L-1} \sin mx_i \cdot \sin kx_i = \begin{cases} 0 & \text{dla } m \neq k \\ L & \text{dla } m = k \neq 0 \\ 0 & \text{dla } m = k = 0 \end{cases}$$

$$\sum_{i=0}^{2L-1} \cos mx_i \cdot \cos kx_i = \begin{cases} 0 & \text{dla } m \neq k \\ L & \text{dla } m = k \neq 0 \\ 2L & \text{dla } m = k = 0 \end{cases} \quad (2.54)$$

$$\sum_{i=0}^{2L-1} \cos mx_i \cdot \sin kx_i = 0 \quad \text{dla } m, k \text{ — dowolnych}$$

m i k zmieniają się od 0 do L .

Przybliżeniem funkcji $f(x)$ na zbiorze (2.53) jest wielomian trygonometryczny

$$y_n(x) = \frac{1}{2}a_0 + \sum_{j=1}^n (a_j \cos jx + b_j \sin jx), \quad n < L \quad (2.55)$$

Współczynniki a_j i b_j wielomianu (2.55) wyznaczamy tak, aby suma kwadratów różnic

$$\sum_{i=0}^{2L-1} [f(x_i) - y_n(x_i)]^2 \quad (2.56)$$

była najmniejsza. Korzystając z warunków ortogonalności (2.54), otrzymujemy rozwiązanie układu normalnego (2.17) postaci

$$a_j = \frac{1}{L} \sum_{i=0}^{2L-1} f(x_i) \cos jx_i = \frac{1}{L} \sum_{i=0}^{2L-1} f(x_i) \cos \frac{\pi ij}{L}$$

$$b_j = \frac{1}{L} \sum_{i=0}^{2L-1} f(x_i) \sin jx_i = \frac{1}{L} \sum_{i=0}^{2L-1} f(x_i) \sin \frac{\pi ij}{L} \quad (2.57)$$

$$j = 1, 2, \dots, n$$

Przykład 1. Dla funkcji określonej tablicą 2/10 znaleźć trygonometryczne przybliżenie średniokwadratowe.

Aby ułatwić obliczenia, posłużymy się tablicą 2/11. Mamy 8 punktów, czyli $L = 4$, więc we wzorach (2.56) i (2.57) przyjmujemy $n = 3$. Zatem

$$y_3(x) = 2 - 0,9999 \sin x + 2 \cos 2x + 0,00001 \sin 3x$$

2/10

x_i	0	$\pi/4$	$\pi/2$	$3\pi/4$	π	$5\pi/4$	$3\pi/2$	$7\pi/4$
$f(x_i)$	4	1.2929	-1	1.2929	4	2.7071	1	2.7071

Błąd

$$\delta_3^2 = \sum_{i=0}^7 [f(x_i) - y_3(x_i)]^2 \approx 0,0000012$$

W rzeczywistości liczby zamieszczone w tablicy 2.10 są niewiele zaburzonymi wartościami funkcji $f^*(x) = 2 - \sin x + 2 \cos 2x$.

Przykład 2. Mamy daną tablicę wartości funkcji $y_i = y(x_i)$ w równoodległych punktach $x_i = \frac{2\pi i}{N}$ ($i = 1, 2, \dots, N$), przy czym dane te są wartościami pomiarowymi pewnego zjawiska, o którym wiadomo, że ma przebieg okresowy. Zgodnie z p. 2.2.3 w przypadku takim należy poszukiwać wielomianu trygonometrycznego stopnia K , aproksymującego dany układ wartości funkcji $y_i = y(x_i)$.

Funkcja aproksymująca ma postać

$$y(x) = a_0 + \sum_{i=1}^K \left(a_i \cos ix + b_i \sin ix \right) \quad (\text{P1})$$

Przyjmuje się, że stopień wielomianu K musi spełniać warunek

$$2K + 1 \leq N \quad (\text{P2})$$

Współczynniki Eulera-Fouriera w (P1) są określone wzorami

$$\begin{aligned} a_0 &= \frac{1}{N} \sum_{j=1}^N y(x_j) \\ a_i &= \frac{2}{N} \sum_{j=1}^N y(x_j) \cos \frac{2\pi ji}{N} \\ b_i &= \frac{2}{N} \sum_{j=1}^N y(x_j) \sin \frac{2\pi ji}{N} \quad i = 1, 2, \dots, K \end{aligned} \quad (\text{P3})$$

Algorytm programu

Stopień wielomianu określamy na podstawie zależności (P2) przy założeniu, że K jest liczbą naturalną

$$K \leq K_{\max} = \begin{cases} \frac{N-1}{2} & \text{dla } N \text{ — nieparzystego} \\ \frac{N-2}{2} & \text{dla } N \text{ — parzystego} \end{cases}$$

Obliczanie współczynników Eulera-Fouriera (P3) znacznie ułatwia spostrzeżenie, że są one średnimi arytmetycznymi pewnych N -wyrazowych ciągów. I tak odpowiednio:

a_0 jest średnią arytmetyczną ciągu $(y(x_i))$, $i = 1, 2, \dots, N$

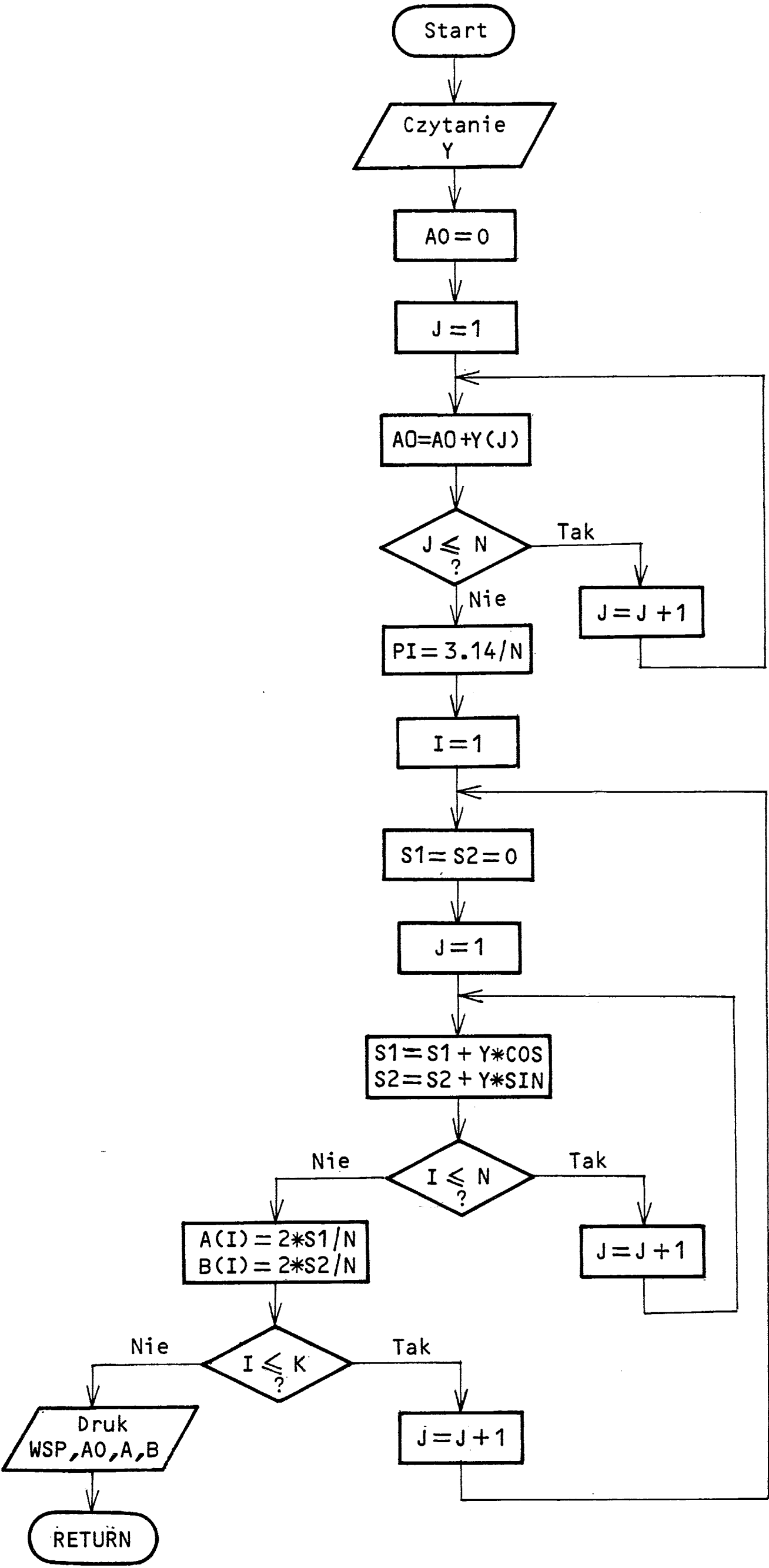
$a_j, j = 1, 2, \dots, K$, są podwojonymi średnimi arytmetycznymi ciągu

$$\left(y(x_i) \cos \frac{2\pi ij}{N} \right), \quad i = 1, 2, \dots, N$$

$b_j, j = 1, 2, \dots, K$ są podwojonymi średnimi arytmetycznymi ciągu

$$\left(y(x_i) \sin \frac{2\pi ij}{N} \right), \quad i = 1, 2, \dots, N$$

Na rysunku 2.2 jest przedstawiony schemat blokowy procedury.



Tekst procedury:

```

CONST max = ...;
TYPE zakres = 1..max;
   wektor= ARRAY [zakres] OF real;
.....
PROCEDURE APR(N,K: zakres; VAR Y,A,B: wektor);
VAR I,J: integer;
    PI,A0,S1,S2: real;
BEGIN
    write('podaj wartosci Y');
    FOR J := 1 TO N DO readln(Y[J]);
    A0 := 0.0;
    FOR J := 1 TO N DO A0 := A0 + Y[J];
    A0 := A0/N;
    PI := 4.0 * ArcTan(1.0)/N;
    FOR I := 1 TO K DO
        BEGIN
            S1 := 0.0;
            S2 := 0.0;
            FOR J := 1 TO N DO
                BEGIN
                    S1 := S1 + Y[J] * cos(PI * J * I);
                    S2 := S2 + Y[J] * sin(PI * J * I);
                END;
            A[I] := S1 * 2.0/N;
            B[I] := S2 * 2.0/N;
        END;
    writeln('wspolczynnik a0 = ',A0:10:4);
    writeln('      J      A[J]      B[J]');
    FOR J := 1 TO K DO
        writeln(j:3,'      ',A[J]:10:4,'      ',B[J]:10:4)
    END.

```

Ćwiczenia

1. Sformułować zagadnienie średniokwadratowej aproksymacji trygonometrycznej, w przypadku gdy liczba funkcji jest równa liczbie punktów pomiarowych (jest to interpolacja trygonometryczna).

2. Napisać wzory (2.54), (2.55) i (2.57) dla przypadku nieparzystej liczby punktów pomiarowych.

Odpowiedzi. 2. $x_i = 2\pi i/(2L + 1)$, $i = 0, 1, \dots, 2L$;

$$\sum_{i=0}^{2L} \sin jx_i \cdot \sin kx_i = \begin{cases} 0 & \text{dla } j \neq k \\ L + 1/2 & \text{dla } j = k \neq 0 \\ 0 & \text{dla } j = k = 0 \end{cases}$$

$$\sum_{i=0}^{2L} \cos jx_i \cdot \cos kx_i = \begin{cases} 0 & \text{dla } j \neq k \\ L + 1/2 & \text{dla } j = k \neq 0 \\ 2L + 1 & \text{dla } j = k = 0 \end{cases}$$

$$\sum_{i=0}^{2L} \cos jx_i \sin kx_i = 0, \quad j, k \text{ — dowolne}$$

$$a_j = \frac{2}{2L+1} \sum_{i=0}^{2L} f(x_i) \cos jx_i = \frac{2}{2L+1} \sum_{i=0}^{2L} f(x_i) \cos \frac{2\pi ij}{2L+1}, \quad j = 0, 1, \dots, n$$

$$b_j = \frac{2}{2L+1} \sum_{i=0}^{2L} f(x_i) \sin jx_i = \frac{2}{2L+1} \sum_{i=0}^{2L} f(x_i) \sin \frac{2\pi ij}{2L+1}, \quad j = 1, 2, \dots, n$$

2.2.4. SZYBKA TRANSFORMACJA FOURIERA

Transformacja (przekształcenie) Fouriera jest obecnie często stosowana w obliczeniach inżynierskich i dlatego rzeczą istotną jest określenie zakresu jej stosowania oraz złożoność obliczeniowa poszczególnych algorytmów. W wielu przypadkach ważne jest przekształcenie danego zadania w inne, mniej złożone lub prostsze obliczeniowo.

Każdą funkcję określoną tablicą wartości w n różnych punktach $x_\beta = \frac{2\pi\beta}{n}$, $\beta = 0, 1, \dots, n-1$, można przybliżać wielomianem trygonometrycznym postaci

$$F(x) = \sum_{k=0}^{n-1} c_k \exp(jkx), \quad j = \sqrt{+1} \quad (2.58)$$

lub

$$F(x) = \frac{a_0}{2} + \sum_{k=1}^m (a_k \cos kx + b_k \sin kx) + \frac{1}{2} \theta a_{m+1} \cos(m+1)x \quad (2.59)$$

gdzie

$$\theta = 1, \quad m = \frac{n-2}{2} \quad \text{dla } n \text{ — parzystych}$$

$$\theta = 0, \quad m = \frac{n-1}{2} \quad \text{dla } n \text{ — nieparzystych}$$

$$a_k = \frac{1}{2n} \sum_{\beta=0}^{n-1} f(x_\beta) \cos kx_\beta$$

$$b_k = \frac{1}{2n} \sum_{\beta=0}^{n-1} f(x_\beta) \sin kx_\beta$$

$$c_k = \frac{1}{n} \sum_{\beta=0}^{n-1} f(x_\beta) \exp(-jkx_\beta)$$

Oczywiście, gdy weźmiemy jedynie m_0 składników ($m_0 < n - 1$ dla funkcji (2.58) lub $m_0 < m$ dla funkcji (2.59)), otrzymamy wówczas wielomian trygonometryczny najlepszego przybliżenia średniokwadratowego spośród wszystkich wielomianów trygonometrycznych o tej samej liczbie składników.

Zadanie polega na znalezieniu albo wartości współczynników Fouriera

$$c_k = \frac{1}{n} \sum_{\beta=0}^{n-1} f(x_\beta) \exp(-jkx_\beta) \quad (2.60)$$

$$k = 0, 1, \dots, n-1$$

mając dane zespolone wartości funkcji $f(x)$ w punktach

$$x_\beta = \frac{2\pi\beta}{n}, \quad \beta = 0, 1, \dots, n-1 \quad (2.61)$$

albo dla znanych zespolonych wartości c_k na znalezieniu funkcji $F(x)$ określonej wzorem (2.58). Rozpatrzmy pierwszy przypadek.

Znajdowanie współczynników c_k . *Metody szybkiej transformacji Fouriera (FFT)* polegają na rozłożeniu liczby n na iloczyn czynników (różnych od jedności) i przeprowadzeniu transformacji Fouriera nad mniejszą liczbą wyrazów w każdym z czynników składowych.

Oznaczmy

$$w = \exp\left(-\frac{2\pi j}{n}\right), \quad a_\beta = \frac{1}{n} f\left(\frac{2\pi\beta}{n}\right) \quad (2.62)$$

skąd

$$c_k = \sum_{\beta=0}^{n-1} a_\beta w^{k\beta}, \quad k = 0, 1, \dots, n-1 \quad (2.63)$$

Przyjmijmy, że

$$n = r_1 r_2 \dots r_p \quad (2.64)$$

gdzie r_i są liczbami całkowitymi (różnymi od jedności), i oznaczmy

$$n_v = r_{v+1} r_{v+2} \dots r_p = \prod_{i=v+1}^p r_i \quad (2.65)$$

$$v = 0, 1, \dots, p-1$$

$$n_p = 1$$

Stąd

$$n = r_1 r_2 \dots r_v n_v$$

Korzystając z zapisu pozycyjnego można wykazać, że:

1. Każdą liczbę całkowitą $k \in \langle 0; n-1 \rangle$ można jednoznacznie zapisać jako

$$k = \alpha_1 n_1 + \alpha_2 n_2 + \dots + \alpha_{p-1} n_{p-1} + \alpha_p n_p$$

gdzie $0 \leq \alpha_i \leq r_i - 1$, czyli

$$k = \alpha_1(r_2 r_3 \dots r_p) + \alpha_2(r_3 \dots r_p) + \dots + \alpha_{p-1} r_p + \alpha_p \cdot 1 \quad (2.66)$$

przy czym współczynniki α_i są jednoznacznie określone i

$$\alpha_1 \in \{0, 1, 2, \dots, r_1 - 1\}$$

$$\alpha_2 \in \{0, 1, 2, \dots, r_2 - 1\}$$

$$\dots\dots\dots$$

$$\alpha_p \in \{0, 1, 2, \dots, r_p - 1\}$$

2. Dla każdej liczby całkowitej $\beta \in \langle 0; n - 1 \rangle$ ułamek $\frac{\beta}{n}$ ma jednoznaczny zapis

$$\frac{\beta}{n} = \frac{l_1}{n_0} + \frac{l_2}{n_1} + \dots + \frac{l_p}{n_{p-1}}$$

gdzie $0 \leq l_i \leq r_i - 1$, czyli

$$\frac{\beta}{n} = \frac{l_1}{r_1 \dots r_p} + \frac{l_2}{r_2 \dots r_p} + \dots + \frac{l_{p-1}}{r_{p-1} r_p} + \frac{l_p}{r_p} \quad (2.67)$$

przy czym współczynniki l_i są jednoznacznie określone i

$$l_1 \in \{0, 1, 2, \dots, r_1 - 1\}$$

$$l_2 \in \{0, 1, 2, \dots, r_2 - 1\}$$

$$\dots\dots\dots$$

$$l_p \in \{0, 1, 2, \dots, r_p - 1\}$$

Niech

$$k_v = \sum_{i=v+1}^p \alpha_i n_i \quad \text{oraz} \quad \frac{\beta_v}{n_v} = \sum_{i=v}^{p-1} \frac{l_{i+1}}{n_i} \quad (2.68)$$

$$k_v < n_v, \quad v = 0, 1, \dots, p$$

Stąd

$$\begin{aligned} \frac{k\beta}{n} &= \left(\sum_{i=1}^p \alpha_i n_i \right) \left(\sum_{v=0}^{p-1} \frac{l_{v+1}}{n_v} \right) = \sum_{v=0}^{p-1} \frac{l_{v+1}}{n_v} \left(\sum_{i=v+1}^p \alpha_i n_i \right) + M = \\ &= \sum_{v=0}^{p-1} \frac{l_{v+1} k_v}{n_v} + M \end{aligned} \quad (2.69)$$

gdzie M jest pewną stałą całkowitą. Podstawiając (2.69) do (2.62), otrzymujemy

$$\begin{aligned}
w^{k\beta} &= \exp \left[\left(-2\pi j \frac{k\beta}{n} \right) \right] = \exp \left[-2\pi j \left(\sum_{v=0}^{p-1} \frac{l_{v+1} k_v}{n_v} + M \right) \right] = \\
&= \prod_{v=0}^{p-1} \exp \left[-2\pi j \frac{l_{v+1} k_v}{n_v} - M 2\pi j \right] = \\
&= \prod_{v=0}^{p-1} \exp \left[-2\pi j \frac{l_{v+1} k_v}{n_v} \right] = \prod_{v=0}^{p-1} w_v^{\mu_v l_{v+1}}
\end{aligned} \tag{2.70}$$

przy czym

$$w_v = \exp \left[-\frac{2\pi j}{n_v} \right], \quad w_0 = w, \quad \exp[-M 2\pi j] = 1$$

Wzór (2.63) możemy teraz zapisać w postaci

$$\begin{aligned}
c_k &= c(l_1, l_2, \dots, l_p) = \\
&= \sum_{l_1=0}^{r_1-1} \sum_{l_2=0}^{r_2-1} \dots \sum_{l_p=0}^{r_p-1} c^{(0)}(l_1, \dots, l_p) w_0^{k_0 l_1} w_1^{k_1 l_2} \dots w_{p-1}^{k_{p-1} l_p}
\end{aligned} \tag{2.71}$$

Człony $c^{(k)}$, określone wzorem (2.71), nazywamy *współczynnikami manipulacyjnymi*.

Jak widać, szukaną transformatę Fouriera można obliczyć po wykonaniu p -kolejnych kroków iteracyjnych: Rozpoczynając od obliczenia

$$c^{(1)}(l_1, l_2, \dots, l_{p-1}, \alpha_p) = \sum_{l_p=0}^{r_p-1} c^{(0)}(l_1, l_2, \dots, l_p) w_{p-1}^{k_{p-1} l_p} \tag{2.72}$$

przy zachowaniu stałych $l_1, l_2, \dots, l_{p-1}$ (dla każdej z możliwych ich wartości), otrzymujemy n/r_p transformat Fouriera funkcji $f\left(\frac{2\pi\beta}{n}\right) = f(l_p)$, z których każda wymaga wykonania r_p^2 operacji. Podobnie obliczając

$$\begin{aligned}
c^{(2)}(l_1, l_2, \dots, l_{p-2}, \alpha_{p-1}, \alpha_p) &= \\
&= \sum_{l_{p-1}=0}^{r_{p-1}-1} c^{(1)}(l_1, l_2, \dots, l_{p-1}, \alpha_p) w_{p-2}^{k_{p-2} l_{p-1}}
\end{aligned} \tag{2.73}$$

(przy zachowaniu stałych $l_1, l_2, \dots, l_{p-2}$ dla każdej z ich możliwych wartości) otrzymujemy n/r_{p-1} transformat Fouriera funkcji $f(l_{p-1})$ wymagających r_{p-1}^2 operacji. Kontynuując to postępowanie w p -tym kroku, otrzymujemy

$$\begin{aligned}
c_k &= c^{(p)}(\alpha_1, \alpha_2, \dots, \alpha_p) = \\
&= \sum_{l_1=0}^{r_1-1} c^{(p-1)}(l_1, \alpha_2, \dots, \alpha_p) w_0^{k_0 l_1}
\end{aligned} \tag{2.74}$$

Ogólnie, cała procedura zamiast n^2 operacji, potrzebnych do wykonania obliczeń przy dyskretnej transformacji Fouriera wymaga jedynie wykonania $n(r_1 + r_2 + \dots + r_p)$ operacji.

Szczególnym przypadkiem omówionego algorytmu jest *metoda Cooleya-Tukeya* [13]. Stosuje się ją zwłaszcza wtedy, gdy n jest całkowitą potęgą dwójki, czyli $n = 2^p$. Oczywiście w razie potrzeby zawsze można do ciągu wartości $f(x_p)$ dodać odpowiednią liczbę zer, tak aby warunek ten był spełniony. Niech $\beta = 2\beta_1$, gdy β jest liczbą parzystą, albo $\beta = 2\beta_1 + 1$, gdy β jest liczbą nieparzystą $\left(0 \leq \beta_1 \leq \frac{n}{2} - 1\right)$. Wzór (2.63) zapisujemy w postaci

$$\begin{aligned} c_k &= \sum_{\beta_1=0}^{\frac{n}{2}-1} a_{2\beta_1} (w^2)^{k\beta_1} + \sum_{\beta_1=0}^{\frac{n}{2}-1} a_{2\beta_1+1} w^{k(2\beta_1+1)} = \\ &= \sum_{\beta_1=0}^{\frac{n}{2}-1} a_{2\beta_1} (w^2)^{k\beta_1} + \sum_{\beta_1=0}^{\frac{n}{2}-1} a_{2\beta_1+1} (w^2)^{k\beta_1} w^k \end{aligned} \quad (2.75)$$

Ponieważ $k = \frac{n}{2}\alpha + k_1$, $k_1 = 0, 1, \dots, \frac{n}{2} - 1$, $\alpha = 0, 1$, zatem

$$(w^2)^{k\beta_1} = (w^2)^{\left(\frac{n}{2}\alpha + k_1\right)\beta_1} = (w^2)^{\frac{\alpha n \beta_1}{2}} (w^2)^{k_1 \beta_1} = (w^n)^{\alpha \beta_1} (w^2)^{k_1 \beta_1} = (w^2)^{k_1 \beta_1}$$

bowiem $w^n = 1$. Oznaczając

$$\begin{aligned} \xi(k_1) &= \sum_{\beta_1=0}^{\frac{n}{2}-1} a_{2\beta_1} (w^2)^{k_1 \beta_1}, \quad k_1 = 0, 1, \dots, \frac{n}{2} - 1 \\ \eta(k_1) &= \sum_{\beta_1=0}^{\frac{n}{2}-1} a_{2\beta_1+1} (w^2)^{k_1 \beta_1} \end{aligned}$$

otrzymujemy

$$c_k = \xi(k_1) + w^k \eta(k_1), \quad k = 0, 1, \dots, n - 1 \quad (2.76)$$

przy czym

dla $k = 0, 1, \dots, \frac{n}{2} - 1$ mamy $k_1 = k$

dla $k = \frac{n}{2}, \dots, n - 1$ mamy $k_1 = k - \frac{n}{2}$

Dla n będącego całkowitą potęgą dwójki szybka transformacja Fouriera wymaga wykonania $2n \log_2 n$ działań zespolonych (zob. np. [3, rozdz. VII]).

Poniżej podajemy bardzo prosty program w Fortranie obliczania współczynników dyskretnej transformacji Fouriera na maszynach cyfrowych (według COOLEYA, LEWISA i WELCHA [12]). Pierwsza część programu (pętla DO 3) przeprowadza reorganizację danych, część druga (pętla DO 5) wykonuje obliczenia współczynników. Program ten wymaga wykonania $\log_2 n$ obliczeń wartości sinusa i cosinusa oraz n mnożeń przez w (razem $2n \log_2 n$). Przyjmijmy, że obliczenia prowadzimy na maszynie cyfrowej wykonującej około 1 miliona operacji na liczbach

zespolonych w ciągu jednej sekundy. W tablicy 2/12 podano przybliżone czasy rozwiązywania zadania przy użyciu tradycyjnego algorytmu obliczania dyskretnej transformacji Fouriera i FFT, opisaney w pracy M. JANKOWSKIEGO i H. WOŹNIAKOWSKIEGO [36], gdzie można znaleźć również wzory do obliczania FFT dla danych rzeczywistych. Dane programu: A — tablica zawierająca na wejściu dane wartości funkcji, a na wyjściu szukane współczynniki Fouriera, $M = \log_2 N$, gdzie N — liczba punktów.

2/12

M	Czas obliczeń	
	zwykły algor.	FFT
10	~ 1 s	$2 \cdot 10^{-2}$ s
15	~ 18 min	~ 1 s
20	~ 305 h	~ 42 s

Tekst programu

```

SUBROUTINE FFT(A,M)
COMPLEX A(1), U, T, W
N = 2**M
NW = N/2
NM = N - 1
J = 1
DO 3 I = 1, NM
IF (I .GE. J) GO TO 1
T = A(J)
A(J) = A(I)
A(I) = T
1 K = NW
2 IF (K .GE. J) GO TO 3
J = J - K
K = K/2
GO TO 2
3 J = J + K
DO 5 L = 1, M
LE = 2**L
LE1 = LE/2
U = (1., 0.)
ANG = 3.141593/LE1
W = CMPLX(COS(ANG), SIN(ANG))
DO 5 J = 1, LE1
DO 4 I = J, N, LE
IP = I + LE1
T = A(IP) * U
A(IP) = A(I) - T
4 A(I) = A(I) + T
5 U = U * W
RETURN
END

```

Stosowanie dyskretnej analizy Fouriera do znajdowania transmitancji układów, wyznaczania gęstości widmowej, gęstości widmowej mocy, funkcji autokorelacji, interkorelacji itp., czynią z niej niezwykle istotne narzędzie w obliczeniach inżynierskich, zwłaszcza dla elektroników i elektryków.

Szczegółowe omówienie tych zastosowań Czytelnik może znaleźć w książce J. BENDATA, A. PIERSOLA [9].

Omówimy jeszcze pokrótce algorytm szybkiego mnożenia wielomianów, pokazujący korzyści, jakie daje stosowanie szybkiej transformacji Fouriera.

Niech $\mathbf{a} = [a_0, a_1, \dots, a_{n-1}]^T$ będzie wektorem o elementach zespolonych, ciąg (w^k) , $k = 0, 1, \dots, n-1$, ciągiem pierwiastków stopnia n z jedności, A — macierzą kwadratową stopnia n taką, że $A[i+1, j+1] = w^{ij(\bmod n)}$, $i, j = 0, 1, \dots, n-1$. Wektor $\mathcal{F}(\mathbf{a}) = A \cdot \mathbf{a}$, którego i -ty element

$$f_i = \sum_{k=0}^{n-1} a_k w^{k \cdot i(\bmod n)}, \quad i = 0, 1, \dots, n-1 \quad (2.77)$$

nazywamy *dyskretną transformatą Fouriera wektora $\mathbf{a}$* .

Łatwo można wykazać, że macierz A jest nieosobliwa i element $A^{-1}[i, j] = \frac{1}{n} w^{-ij}$.

Wektor $\mathcal{F}^{-1}(\mathbf{a}) = A^{-1} \cdot \mathbf{a}$, którego i -ty element

$$f_i^{-1} = \frac{1}{n} \sum_{k=0}^{n-1} a_k w^{-(k \cdot i)(\bmod n)} \quad (2.78)$$

nazywamy *odwrotną dyskretną transformatą Fouriera wektora $\mathbf{a}$* . Oczywiście

$$\mathcal{F}^{-1} \mathcal{F}(\mathbf{a}) = \mathbf{a}.$$

Niech

$$P(x) = \sum_{i=0}^{n-1} a_i x^i$$

będzie wielomianem stopnia $n-1$. Wielomian ten można jednoznacznie opisać albo przez podanie wartości współczynników $a_0, a_1, \dots, a_{n-1}$, albo przez podanie wartości $P(x_k)$ w n różnych punktach x_k , $k = 0, 1, \dots, n-1$. Przyjmijmy, że ciąg punktów $(x_k) = (w^k)$. Obliczenie transformaty Fouriera wektora $\mathbf{a}$ jest równoważne przejściu z opisu wielomianu za pomocą współczynników a_k na opis za pomocą ciągu wartości $(\mathcal{F}(w^k))$. Transformacja odwrotna jest równoważna znajdowaniu współczynników przy znanych wartościach wielomianu w punktach — czyli interpolacji.

Zdefiniujemy teraz operację splotu dwóch wektorów. Niech

$$\mathbf{a} = [a_0, a_1, \dots, a_{n-1}]^T, \quad \mathbf{b} = [b_0, b_1, \dots, b_{n-1}]^T$$

*Splotem wektorów $\mathbf{a}$ i $\mathbf{b}$, oznaczanym symbolem $\mathbf{a} * \mathbf{b}$, nazywamy wektor $\mathbf{c} = [c_0, c_1, \dots, c_{2n-1}]^T$, przy czym*

$$c_i = \sum_{j=0}^{n-1} a_j b_{i-j} \quad (2.79)$$

$$a_k = b_k = 0 \quad \text{dla} \quad k < 0 \text{ lub } k \geq n, \quad c_{2n-1} = 0$$

Elementy wektora c są identyczne ze współczynnikami wielomianu, będącego iloczynem wielomianów

$$P(x) = \sum_{i=0}^{n-1} a_i x^i \quad \text{oraz} \quad Q(x) = \sum_{i=0}^{n-1} b_i x^i$$

zatem

$$P(x)Q(x) = \sum_{i=0}^{2n-2} a_j b_{i-j} x^i$$

Gdy wielomiany $P(x)$ i $Q(x)$ są opisane ciągami współczynników (a_k) i (b_k) , wówczas aby znaleźć wyrazy ciągu (c_k) możemy dokonać splotu wektorów a i b . Odwrotnie, jeśli $P(x)$ i $Q(x)$ są opisane przez ciągi ich wartości $(P(x_k))$ i $(Q(x_j))$, to aby znaleźć wartości opisujące ich iloczyn, wystarczy pomnożyć $P(x_k)Q(x_j)$, $k, j = 0, \dots, n-1$. Ponieważ iloczyn wielomianów stopnia $n-1$ jest wielomianem stopnia $2n-2$, więc do jego opisanie potrzeba znać wartości w $2n-2$ różnych punktach. Aby to uzyskać, wielomiany $P(x)$ i $Q(x)$ stopnia $n-1$ zastępujemy wielomianami $P^*(x)$ i $Q^*(x)$ stopnia $2n-1$, w których współczynniki a_k^* i b_k^* dla $k = 0, 1, \dots, n-1$ są identyczne ze współczynnikami a_k i b_k , natomiast współczynniki a_k^* i b_k^* dla $k = n, n+1, \dots, 2n-1$ są tożsamościowo równe zeru. Korzystając z wprowadzonych pojęć prostej i odwrotnej transformaty Fouriera, otrzymujemy

$$a^*b = \mathcal{F}^{-1}(\mathcal{F}(a) \cdot \mathcal{F}(b)) \quad (2.80)$$

W ten sposób zadanie wyznaczania współczynników iloczynu $P(x)Q(x)$ sprowadzamy do omówionego uprzednio zadania, które wymaga wykonania $n(r_1 + r_2 + \dots + r_p)$ działań, $n = r_1 r_2 r_3 \dots r_p$, podczas gdy „zwykły” algorytm mnożenia wielomianów wymaga n^2 działań.

Można także pokazać, że jeżeli zagadnienie znajdowania wielomianu interpolacyjnego stopnia n , opartego na $n+1$ węzłach, przy „klasycznych” algorytmach (patrz rozdz. 1) wymaga co najmniej n^2 działań, to użycie szybkiej transformacji Fouriera redukuje tę liczbę działań do $n \log_2^2 n$ (gdzie $n = 2^k$).

Szczegółowe omówienie (wraz z algorytmami obliczeniowymi) metod szybkiego mnożenia wielomianów, dzielenia wielomianów itp., można znaleźć w książce A. AHO, J. HOPCOFFA i J. ULLMANA [3].

2.2.5. Aproksymacja za pomocą funkcji sklejanych

W punkcie tym omówimy aproksymację średniokwadratową funkcji $f(x)$ za pomocą funkcji sklejanych stopnia trzeciego z równoodległymi węzłami $x_i = a + ih$, $h = \frac{b-a}{n}$. Przypominamy, że każdą funkcję $s(x) \in S_3(\Delta_n)$ można przedstawić w postaci kombinacji liniowej

$$s(x) = \sum_{i=-1}^{n+1} c_i \Phi_i^3(x), \quad a \leq x \leq b \quad (2.81)$$

gdzie $\Phi_i^3(x)$ są funkcjami określonymi wzorem (1.65).

Omówimy najpierw aproksymację średniokwadratową funkcji $f(x)$ na przedziale $\langle a; b \rangle$. Zgodnie z metodą opisaną w p. 2.2.2 współczynniki c_i we wzorze (2.81) dobierzemy tak, aby odchylenie kwadratowe

$$I = \int_a^b \left[f(x) - \sum_{i=-1}^{n+1} c_i \Phi_i^3(x) \right]^2 dx \quad (2.82)$$

było jak najmniejsze. W celu wyznaczenia minimum funkcji $I = I(c_{-1}, c_0, \dots, c_{n+1})$ obliczamy pochodne $\frac{\partial I}{\partial c_j}$, $j = -1, 0, \dots, n+1$, i porównujemy je do zera. Otrzymane równania tworzą układ $n+3$ równań z $n+3$ niewiadomymi c_i

$$\sum_{i=-1}^{n+1} c_i \int_a^b \Phi_i^3(x) \Phi_j^3(x) dx = \int_a^b f(x) \Phi_j^3(x) dx \quad (2.83)$$

$$j = -1, 0, 1, \dots, n+1$$

Funkcje $\Phi_i^3(x)$, $i = -1, 0, 1, \dots, n+1$, są liniowo niezależne na przedziale $\langle a; b \rangle$, więc układ (2.83) ma jednoznaczne rozwiązanie. Rozwiązanie to daje punkt, w którym funkcja I osiąga minimum.

Układ (2.83) możemy zapisać w postaci

$$\sum_{i=-1}^{n+1} a_{ij} c_i = \frac{1}{h} \int_a^b f(x) \Phi_j^3(x) dx \quad (2.84)$$

$$j = -1, 0, 1, \dots, n+1$$

$$\text{gdzie} \quad a_{ij} = \frac{1}{h} \int_a^b \Phi_i^3(x) \Phi_j^3(x) dx.$$

Macierz układu (2.84) jest macierzą symetryczną i wstęgową (tzn. pięciodiagonalną), a więc $a_{ij} = 0$ dla $|i - j| \geq 4$. Podamy tu wartości różnych od zera współczynników, zaczerpnięte z pracy [57]:

$$a_{-1,-1} = a_{n+1,n+1} = \frac{1}{7}, \quad a_{-1,0} = \frac{129}{140}, \quad a_{n+1,n} = \frac{6}{7}$$

$$a_{-1,1} = a_{n+1,n-1} = \frac{3}{7}, \quad a_{0,0} = a_{n,n} = \frac{302}{35}$$

$$a_{0,1} = a_{n,n-1} = \frac{531}{70}, \quad a_{1,1} = a_{n-1,n-1} = \frac{599}{35}$$

$$a_{k,k} = \frac{604}{35} \quad (2 \leq k \leq n-2), \quad a_{k,k+1} = \frac{1191}{140} \quad (1 \leq k \leq n-2)$$

$$a_{k,k+2} = \frac{6}{7} \quad (0 \leq k \leq n-2), \quad a_{k,k+3} = \frac{1}{140} \quad (-1 \leq k \leq n-2)$$

$$a_{i,j} = a_{j,i} \quad (i, j = -1, 0, 1, \dots, n+1)$$

Omówimy teraz aproksymację średniokwadratową funkcji $f(x)$ określonej na dyskretnym zbiorze punktów. Niech (x'_i) dla $i = 0, 1, \dots, n_1$; $n_1 > n+3$, będzie zbiorem punktów, w których dane są wartości funkcji $f(x)$. Minimalizujemy wyrażenie

$$J = \sum_{k=0}^{n_1} \left[f(x'_k) - \sum_{i=-1}^{n+1} c_i \Phi_i^3(x'_k) \right]^2 \quad (2.85)$$

W celu wyznaczenia współczynników c_i obliczamy pochodne cząstkowe funkcji $J = J(c_{-1}, c_0, \dots, c_{n+1})$ względem zmiennych c_j i przyrównujemy je do zera. Otrzymane równania tworzą układ

$$\sum_{i=-1}^{n+1} b_{ij} c_i = \sum_{k=0}^{n_1} f(x'_k) \Phi_j^3(x'_k), \quad j = -1, 0, 1, \dots, n+1 \quad (2.86)$$

$$\text{gdzie } b_{ij} = \sum_{k=0}^{n_1} \Phi_i^3(x'_k) \Phi_j^3(x'_k).$$

Jeżeli wyznacznik tego układu jest różny od zera, to rozwiązując go możemy wyznaczyć współczynniki c_j rozwinięcia (2.81). Można wykazać, że to rozwiązanie daje punkt, w którym funkcja J osiąga minimum. Ze wzoru określającego elementy b_{ij} widać, że macierz układu (2.86) jest macierzą symetryczną i wstęgową, mamy zatem $b_{ij} = 0$ dla $|i-j| \geq 4$.

Ćwiczenia

1. Wyjaśnić, dlaczego macierze układów (2.84) i (2.86) są macierzami wstęgowymi.

2.3. APROKSYMACJA JEDNOSTAJNA

Twierdzenie 1 z punktu 2.1 zapewnia, że dowolną funkcję $f(x)$ ciągłą na przedziale $\langle a; b \rangle$ możemy aproksymować jednostajnie za pomocą wielomianów z dowolnie dużą dokładnością. Jeżeli funkcję $f(x)$ można na przedziale $\langle a; b \rangle$ rozwinąć w szereg Taylora, to przybliżeniem może być odpowiednio obcięty szereg Taylora, w którym liczba wyrazów uzależniona jest od żądanej dokładności. Zagadnienie sprowadza się wówczas do znalezienia takich współczynników a_i wielomianu

$$W_n(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n \quad (2.87)$$

aby wyrażenie

$$E_n = \max_{x \in \langle a; b \rangle} |f(x) - W_n(x)| \quad (2.88)$$

było możliwe najmniejsze. Wielomian $W_n(x)$, dla którego E_n osiąga minimum nazywamy *n -tym wielomianem najlepszego przybliżenia jednostajnego funkcji $f(x)$* .

Przy poszukiwaniu wielomianu najlepszego przybliżenia jednostajnego korzysta się z *twierdzenia Borela*, który wykazał, że dla każdej funkcji $f(x)$ ciągłej na przedziale $\langle a; b \rangle$ i dowolnej liczby naturalnej n istnieje wielomian $W_n(x)$ będący jej najlepszym przybliżeniem jednostajnym, oraz z *twierdzenia Czebyszewa*, mówiącego, że wielomian taki jest tylko jeden [10, 18]. Ogólnej metody umożliwiającej znajdowanie wielomianu najlepszego przybliżenia jednostajnego stopnia n dla dowolnej funkcji $f(x)$ ciągłej na przedziale $\langle a; b \rangle$ nie ma. Metody takie podaje się dla pewnych szczególnych przypadków; są one dosyć pracochłonne i dlatego zazwyczaj rezygnujemy z poszukiwania wielomianu najlepszego przybliżenia aproksymując funkcję $f(x)$ wielomianami nie będącymi najlepszymi przybliżeniami jednostajnymi.

2.3.1. Metoda szeregów potęgowych

Jedną z metod aproksymacji jednostajnej jest metoda szeregów potęgowych. Niech dana funkcja $f(x)$, określona na przedziale $\langle a; b \rangle$, będzie rozwijalna na otoczeniu $Q(x_0; \delta)$ punktu x_0 o promieniu δ w szereg Taylora. Oczywiście

$$x_0 - \delta < a < b < x_0 + \delta$$

Rozpatrzmy wielomian stopnia n

$$\begin{aligned} W_n(x) = & f(x_0) + \frac{f'(x_0)}{1!} (x - x_0) + \frac{f''(x_0)}{2!} (x - x_0)^2 + \dots + \\ & + \frac{f^{(n)}(x_0)}{n!} (x - x_0)^n \end{aligned} \quad (2.89)$$

będący przybliżeniem funkcji $f(x)$ na przedziale $\langle a; b \rangle$. Korzystając ze wzoru Taylora z resztą Lagrange'a, mamy

$$f(x) - W_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!} (x - x_0)^{n+1}$$

gdzie c jest punktem leżącym między x i x_0 , a stąd otrzymujemy oszacowanie błędu aproksymacji

$$|f(x) - W_n(x)| \leq \frac{M_{n+1}}{(n+1)!} \delta^{n+1} \quad (2.90)$$

przy czym

$$|f^{(n+1)}(x)| \leq M_{n+1} \quad \text{dla} \quad x \in \langle a; b \rangle$$

Jeżeli funkcja $f(x)$ jest rozwijalna w szereg Taylora, to w przybliżeniu (2.89) można uzyskać dowolną dokładność dla dostatecznie dużego n .

Przykład. Korzystając ze wzoru (2.89), znaleźć przybliżenie funkcji $f(x) = \sin x$ na przedziale $\langle 0; \pi/2 \rangle$ wielomianem stopnia drugiego.

Przyjmijmy $x_0 = \pi/4$. Mamy

$$\begin{aligned} W_2(x) &= \sin \frac{\pi}{4} + \cos \frac{\pi}{4} \cdot \left(x - \frac{\pi}{4}\right) - \frac{1}{2} \sin \frac{\pi}{4} \cdot \left(x - \frac{\pi}{4}\right)^2 = \\ &= -0,35355x^2 + 1,262467x - 0,72061 \end{aligned}$$

Ponieważ $|f'''(x)| = |-\cos x| \leq 1$ dla $x \in \langle 0; \pi/2 \rangle$ i $\delta = \pi/4$, więc

$$|f(x) - W_2(x)| \leq \frac{1}{3!} \cdot \left(\frac{\pi}{4}\right)^3 < 0,081$$

Zatem maksymalny błąd aproksymacji jest nie większy niż 0,081.

Ponieważ oszacowanie błędu na podstawie nierówności (2.90) jest niezbyt dokładne często stosuje się inne metody pozwalające na dokładniejsze oszacowanie lub nawet na dokładne obliczanie odchylenia wielomianu $W_n(x)$ od funkcji $f(x)$.

2.3.2. Przybliżenia Padégo

Z punktu widzenia możliwości i wykorzystania maszyn cyfrowych do aproksymacji jednostajnej bardzo wielu funkcji, okazało się, że przybliżenia wymierne

$$R_{n,k}(x) = \frac{L_n(x)}{M_k(x)} \quad (2.91)$$

dają mniejsze błędy maksymalne niż przybliżenia wielomianami algebraicznymi stopnia N (gdzie $N = n + k$), zwłaszcza niż przybliżenia wzorami Taylora czy Maclaurina.

Zagadnienie sprowadza się do znalezienia $N + 1$ wartości współczynników wielomianów $L_n(x)$ i $M_k(x)$ takich, że

$$f(0) = \frac{L_n(0)}{M_k(0)} \quad \text{oraz że funkcje } f(x) \text{ i } \frac{L_n(x)}{M_k(x)}$$

mają w punkcie $x_0 = 0$ jak najwięcej równych pochodnych. Oczywiście dla $k = 0$ otrzymujemy przybliżenie funkcji $f(x)$ szeregiem potęgowym (Maclaurina).

Założmy, że funkcję $f(x)$ można na pewnym otoczeniu punktu $x_0 = 0$ rozwinąć w zbieżny szereg Maclaurina, a ponadto że funkcje $L_n(x)$ i $M_k(x)$ są względnie pierwsze. Niech

$$\begin{aligned} L_n(x) &= a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n \\ M_k(x) &= b_0 + b_1 x + b_2 x^2 + \dots + b_k x^k \end{aligned} \quad (2.92)$$

Ponieważ $b_0 \neq 0$ (iloraz $L_n(x)/M_k(x)$ musi być określony w zerze), więc przyjmijmy, że $b_0 = 1$. Rozwinięcie funkcji $f(x)$ w szereg Maclaurina ma postać

$$f(x) = \sum_{i=0}^{\infty} c_i x^i$$

Błąd aproksymacji wynosi

$$f(x) - \frac{L_n(x)}{M_k(x)} = \frac{\left(\sum_{i=0}^{\infty} c_i x^i\right) \left(\sum_{i=1}^k b_i x^i\right) - \sum_{i=0}^n a_i x^i}{\sum_{i=0}^k b_i x^i} \quad (2.93)$$

a ponieważ chcemy, aby zachodziło $f(0) = R_{n,k}(0) = 0$ oraz $f^{(m)}(0) - R_{n,k}^{(m)}(0) = 0$ dla $m = 1, 2, \dots, k+n$, trzeba by w liczniku wyrażenia (2.93) znikaly wszystkie wyrazy o potęgach x mniejszych od $n+k+1$. Mamy więc

$$\left(\sum_{i=0}^{\infty} c_i x^i\right) \left(\sum_{i=0}^k b_i x^i\right) - \sum_{i=0}^n a_i x^i = \sum_{i=n+k+1}^{\infty} d_i x^i \quad (2.94)$$

a stąd

$$\sum_{j=0}^k c_{n+k-s-j} \cdot b_j = 0 \quad \text{dla} \quad s = 0, 1, 2, \dots, k-1$$

$$b_0 = 1,$$

$$c_j = 0 \quad \text{dla} \quad j < 0$$

$$a_r = \sum_{j=0}^r c_{r-j} b_j \quad \text{dla} \quad r = 0, 1, 2, \dots, n \quad (2.95)$$

$$b_j = 0 \quad \text{dla} \quad j > k$$

Jeżeli układ (2.95) ma rozwiązanie, to otrzymamy współczynniki przybliżenia (2.91). Z analizy błędów wynika, że tak otrzymane przybliżenie wymierne ma stosunkowo małe błędy w pobliżu zera, które jednak rosną, gdy się od zera oddalamy. Zagadnieniem tym zajmiemy się bardziej szczegółowo w następnym punkcie.

Okazało się też, że dla większości aproksymowanych typowych funkcji w przypadku $n = k$ lub $n = k+1$, otrzymujemy najmniejszy błąd maksymalny (spośród wszystkich przybliżeń Padégo).

Przykład. Wyznamy przybliżenie funkcji $f(x) = e^x$ na przedziale $\langle -0,5; 0,5 \rangle$.

Niech $N = k + n = 3$. Z rozwinięciem e^x w szereg Maclaurina mamy: $c_0 = 1$, $c_1 = 1$, $c_2 = 1/2$ i $c_3 = 1/6$. Rozpatrzmy teraz kolejne przybliżenia Padégo:

a) $n = 3$, $k = 0$. W tym przypadku $R_{3,0}(x)$ jest sumą częściową szeregu Maclaurina

$$R_{3,0}(x) = \frac{1}{6}x^3 + \frac{1}{2}x^2 + x + 1$$

b) $n = 2, k = 1$. Układ (2.95) ma postać

$$\begin{aligned} \frac{1}{6} + b_1 &= 0 & \text{i stąd} & & b_0 &= 1 \\ a_0 &= 1 & & & b_1 &= -1/6 \\ a_1 &= 1 + b_1 & & & a_0 &= 1 \\ a_2 &= \frac{1}{2} + b_1 & & & a_1 &= 5/6 \\ b_0 &= 1 & & & a_2 &= 1/3 \end{aligned}$$

$$R_{2,1}(x) = \frac{\frac{1}{3}x^2 + \frac{5}{6}x + 1}{-\frac{1}{6}x + 1}$$

c) $n = 1, k = 2$. Układ (2.95) ma postać

$$\sum_{j=0}^2 c_{3-s-j} \cdot b_j = 0 \quad \text{dla} \quad s = 0, 1$$

$$a_r = \sum_{j=0}^r c_{r-j} \cdot b_j \quad \text{dla} \quad r = 0, 1$$

$$R_{1,2}(x) = \frac{\frac{2}{3}x + 1}{-\frac{1}{6}x^2 - \frac{1}{3}x + 1}$$

d) $n = 0, k = 3$

$$R_{0,3}(x) = \frac{1}{-\frac{1}{6}x^3 + \frac{1}{2}x^2 - x + 1}$$

W tablicy 2/13 są zestawione wartości funkcji $R_{n,k}(x)$ oraz różnic $e^x - R_{n,k}(x)$. Łatwo zauważyć, że błędy rosną w miarę oddalania się od punktu $x_0 = 0$.

2/13

x	$R_{3,0}$	$R_{2,1}$	$R_{1,2}$	$R_{0,3}$	$e^x - R_{3,0}$	$e^x - R_{2,1}$	$e^x - R_{1,2}$	$e^x - R_{0,3}$
-0.5	0.60416	0.6154	0.592586	0.607595	0.00237	-0.00887	0.013944	-0.001064
-0.3	0.7405	0.74285	0.737327	0.741015	0.00029	-0.00206	0.003463	-0.000225
-0.1	0.90483	0.90497	0.904687	0.904891	-0.00002	-0.00016	0.0001225	-0.000081
0.1	1.105166	1.10512	1.105347	1.105176	0.000034	0.00008	-0.000147	0.000024
0.3	1.3495	1.34734	1.35593	1.344688	0.0004	0.00254	-0.006032	0.005211
0.5	1.64583	1.636375	1.68422	1.655172	0.00287	0.12325	-0.03552	-0.006472

2.3.3. Szeregi Czebyszewa

Ze względu na przyjęte założenia o równości funkcji i ich pochodnych w zerze przybliżenia Padégo nie są najlepsze w sensie aproksymacji jednostajnej, bowiem na krańcach przedziału znacznie rośnie błąd. Podobnie zresztą przedstawia się sprawa przy przybliżeniach szeregami Maclaurina — błąd jest mały w pobliżu zera, lecz szybko rośnie w miarę oddalania się od niego.

Jeżeli zamiast potęg x jako funkcji przybliżającej użyjemy wielomianów, które w całym przedziale zachowują się stosunkowo równomiernie, możemy oczekiwać bardziej równomiernego rozkładu błędu. Do tego celu znakomicie nadaje się baza złożona z wielomianów Czebyszewa (2.5).

Jedną z metod jest przybliżenie funkcji $f(x)$ sumą częściową szeregu

$$f(x) \approx \frac{1}{2}c_0 + \sum_{j=1}^n c_j T_j(x) \quad (2.96)$$

gdzie

$$c_j = \frac{2}{\pi} \int_{-1}^1 \frac{f(x) T_j(x)}{\sqrt{1-x^2}} dx \quad (2.97)$$

przy czym $T_j(x)$ są określone wzorami (1.29).

Przykład. Znaleźć przybliżenie funkcji $f(x) = e^x$ szeregiem (2.96) dla $n = 3$. Mamy $f(x) \approx 0,177348x^3 + 0,54299x^2 + 0,997307x + 0,994571$

Błędy tego wielomianu w porównaniu z wielomianem Padégo dla kilku wybranych x są zestawione w tabeli 2/14.

2/14

x	$R_{3,0}$	T_3	$e^x - R_{3,0}$	$e^x - T_3$
0.5	1.64583	1.6511405	0.00287	0.0024
0.8	2.20533	2.2307329	0.02017	0.0052
1.0	2.66667	2.712216	0.0516	0.00606

Wielomianów Czebyszewa używa się także do budowy przybliżeń wymiernych

$$T_{n,k}(x) = \frac{\sum_{i=0}^n a_i T_i(x)}{\sum_{i=0}^k b_i T_i(x)} \quad (2.98)$$

Współczynniki a_i oraz b_i są dobierane w taki sposób, aby w liczniku wyrażenia

$$f(x) - T_{n,k}(x) = \frac{\left[\frac{1}{2}c_0 + \sum_{j=1}^{\infty} c_j T_j(x) \right] \sum_{i=0}^k b_i T_i(x) - \sum_{i=0}^n a_i T_i(x)}{\sum_{i=0}^k b_i T_i(x)} \quad (2.99)$$

znikały współczynniki przy $T_j(x)$ dla $j = 0, 1, 2, \dots, N$, zatem

$$\left[\frac{1}{2}c_0 + \sum_{j=1}^{\infty} c_j T_j(x) \right] \sum_{i=0}^k b_i T_j(x) - \sum_{i=0}^n a_i T_i(x) = \sum_{j=N+1}^{\infty} h_j T_j(x)$$

Po wymnożeniu i skorzystaniu z tożsamości

$$T_{i+j}(x) + T_{|i-j|}(x) = 2T_i(x) T_j(x) \quad (2.100)$$

otrzymujemy

$$a_0 = \frac{1}{2} \sum_{i=0}^k b_i c_i$$

$$a_r = \frac{1}{2} \sum_{i=0}^k b_i (c_{|i-r|} + c_{i+r}) \quad (2.101)$$

dla $r = 1, 2, \dots, k+n$

$a_r = 0$ dla $r > n$

(zazwyczaj przyjmuje się $b_0 = 1$).

Przykład 2. Znaleźć $T_{2,1}(x)$ i porównać z $R_{2,1}(x)$ dla funkcji $f(x) = e^x$. Mamy $n = 2, k = 1$ i $N = 3$. Układ równań (2.101) ma postać

$$a_0 = \frac{1}{2}(c_0 + b_1 c_1)$$

$$a_1 = c_1 + \frac{1}{2}b_1(c_0 + c_2)$$

$$a_2 = c_2 + \frac{1}{2}b_1(c_1 + c_3)$$

$$0 = c_3 + \frac{1}{2}b_1(c_2 + c_4)$$

$$b_0 = 1$$

i ma rozwiązanie: $b_0 = 1, b_1 = -0,3201585, a_0 = 1,0851256, a_1 = 0,6815317, a_2 = 0,0834572$. Stąd

$$T_{2,1}(x) = \frac{0,1669x^2 + 0,68153x + 1,00167}{-0,32016x + 1}$$

Porównanie $R_{2,1}(x)$ z $T_{2,1}(x)$ oraz powstające błędy są zestawione w tablicy 2/15.

2/15

x	$R_{2,1}(x)$	$T_{2,1}(x)$	$e^x - R_{2,1}$	$e^x - T_{2,1}$
0.1	1.105166	1.106936	0.000034	-0.001736
0.3	1.349500	1.350900	0.0004	-0.001
0.5	1.645830	1.647960	0.00287	0.00074
0.8	2.205330	2.223100	0.02	0.0024
1.0	2.66667	2.72137	0.0516	-0.003

Jak należało się spodziewać, przybliżenie za pomocą funkcji $T_{2,1}(x)$ w miarę oddalania się od punktu $x_0 = 0$ jest znacznie lepsze niż w przypadku $R_{2,1}(x)$.

Ćwiczenia

1. Porównać rozwinięcie funkcji $y = e^x$ w szereg Maclaurina i aproksymację szeregiem Czebyszewa $T_{3,0}(x)$ dla $x \in \langle -1; 1 \rangle$ (z krokiem 0,2).

2. Porównać rozwinięcie funkcji $y = \arctan x$ w szereg Maclaurina z przybliżeniem Padégo $R_{5,4}(x)$ dla $x \in \langle 0,2; 1,0 \rangle$ (z krokiem 0,2).
3. Znaleźć wielomiany $T_{1,2}(x)$ oraz $T_{0,3}(x)$ dla funkcji z przykładu 1. Znaleźć analogiczne wielomiany Padégo i porównać błędy.

Odpowiedzi. 1. Tablica 2/16, gdzie $W_3(x) = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3$; $T_{3,0}(x) = 0,9946 + 0,9971x + 0,543x^2 + 0,176x^3$. 2. Tablica 2/17, gdzie $W_9(x) = x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \frac{1}{7}x^7 + \frac{1}{9}x^9$;

$$R_{5,4}(x) = \frac{x + \frac{7}{9}x^3 + \frac{64}{945}x^5}{1 + \frac{10}{9}x^2 + \frac{5}{21}x^4}$$

2/16

x	e^x	$W_3(x)$	$\varepsilon \times 10^{-4}$	$T_{3,0}(x)$	$\varepsilon \times 10^{-4}$
-1.0	0.3679	0.3333	346	0.3629	50
-0.8	0.4493	0.4347	146	0.4535	42
-0.6	0.5488	0.5440	48	0.5535	47
-0.4	0.6703	0.6693	10	0.6713	10
-0.2	0.8187	0.8187	0	0.8155	32
0.0	1.0000	1.0000	0	0.9446	54
0.2	1.2214	1.2213	1	1.2172	42
0.4	1.4918	1.4907	11	1.4917	1
0.6	1.9221	1.8160	61	1.8267	46
0.8	2.2255	2.2053	202	2.2307	52
1.0	2.7183	2.6667	516	2.7123	60

2/17

x	$\arctan x$	$W_9(x)$	$\varepsilon \times 10^{-5}$	$R_{5,4}(x)$	$\varepsilon \times 10^{-5}$
0.2	0.19740	0.19740	0	0.19740	0
0.4	0.38051	0.38051	0	0.38051	0
0.6	0.54042	0.54067	-25	0.54042	0
0.8	0.67474	0.67982	-508	0.67477	-3
1.0	0.78540	0.83492	-4952	0.78558	-18

2.4. UWAGI KOŃCOWE

Często zdarza się, że dla funkcji określonej tablicą wartości albo wykresem (np. wyniki pomiarów) chcemy dobrać wyrażenie analityczne, które będzie w przybliżeniu opisywać daną funkcję. Proces dobierania takiego wzoru empirycznego składa się z dwóch części — wyboru postaci wzoru i określenia wartości liczbowych dla występujących w nim parametrów tak, aby przybliżenie to było jak najlepsze.

Czasami mamy pewne informacje dotyczące postaci poszukiwanego wzoru, ale najczęściej postać wzoru dobieramy „na oko” na zasadzie

podobieństwa wykresu analizowanej funkcji do znanych funkcji. Staramy się przy tym, aby funkcje te miały stosunkowo prostą postać. Takimi często używanymi wzorcami są między innymi:

$$y = ax^b + c$$

$$y = e^{ax^2 + bx + c}$$

$$y = ae^{bx} + c$$

$$y^2 = ax^2 + bx + c$$

$$y = ax^2 + bx + c$$

$$y = \frac{ax + b}{cx + d}$$

$$y = \frac{1}{ax^2 + bx + c}$$

$$y = \frac{x}{ax^2 + bx + c}$$

$$y = a + \frac{b}{x} + \frac{c}{x^2}$$

$$y = ax^b e^{cx}$$

$$y = ae^{bx} + ce^{dx} \quad \text{itd.}$$

Ponieważ przyjęta „metoda” określania podobieństwa wykresów jest bardzo niedokładna, po wybraniu postaci wzoru, ale przed doбором parametrów, sprawdza się jego przydatność za pomocą *metody wyrównywania*. Wielkości parametrów można określić albo w sposób dokładny omówioną uprzednio metodą najmniejszych kwadratów, albo w sposób przybliżony, *metodą przeciętnych*.

Dla pewnych prostych funkcji wskazówki dotyczące metody wyrównywania, omówienie metody przeciętnych oraz wykresy wybranych, prostych funkcji Czytelnik może znaleźć np. w [11]. Podobnie w pracy [42] są omówione pewne metody oraz podane wzory służące do obliczania wielomianów, ułamków prostych, funkcji wykładniczych, logarytmicznych, trygonometrycznych, cyklometrycznych, hiperbolicznych i funkcji do nich odwrotnych. Omówione są tam także różne sposoby przedstawiania funkcji elementarnych w postaci szeregów potęgowych, szeregów według wielomianów i wielomianów ortogonalnych, ułamków łańcuchowych, granic procesów iteracyjnych itp., a także wzory służące do przybliżonych obliczeń funkcji elementarnych z zadaną dokładnością.

3

PRZYBLIŻONE ROZWIĄZYWANIE RÓWNAŃ NIELINIOWYCH I ICH UKŁADÓW

3.1. JEDNO RÓWNANIE Z JEDNĄ NIEWIADOMĄ

W rozdziale tym zajmiemy się rozwiązywaniem równań nieliniowych oraz ich układów. Jak wiadomo, nie każde równanie można rozwiązać dokładnie. Przede wszystkim dotyczy to większości równań przestępnych. Wiadomo również, że nie można podać metody na rozwiązanie dowolnego równania algebraicznego stopnia wyższego niż czwarty. Dokładne rozwiązanie równania nie zawsze jest też konieczne, ponieważ w równaniach często mamy jedynie przybliżone wartości współczynników i zadanie dokładnego wyznaczenia wartości pierwiastków traci sens. Stąd wynika duże znaczenie metod przybliżonego rozwiązywania równań, w których zadanie wyznaczenia pierwiastków uważamy za praktycznie rozwiązane, jeśli potrafimy określić je z żadaną dokładnością i podać oszacowanie błędu. Rozpatrzmy zadanie przybliżonego obliczania pierwiastków rzeczywistych równania

$$f(x) = 0$$

gdy funkcja $f(x)$ jest ciągła na pewnym przedziale (skończonym lub nieskończonym). Większość metod przybliżonego rozwiązywania równań polega na poprawianiu kolejnych przybliżeń pierwiastków. Ponadto większość omówionych tu metod można stosować jedynie wtedy, gdy znamy przedział, w którym znajduje się pojedynczy pierwiastek rozwiązywanego równania (o takim przedziale mówi się, że jest *przedziałem izolacji* pierwiastka). Do wyznaczenia przedziałów izolacji można wykorzystać wykres funkcji $y = f(x)$ albo, jeżeli jest to trudne, równanie wyjściowe $f(x) = 0$ należy przekształcić do równoważnej mu postaci

$$\varphi(x) = \psi(x)$$

gdzie wykresy funkcji $y = \varphi(x)$ i $y = \psi(x)$ są prostsze do wykonania. Odcięte

punktów przecięcia tych wykresów będą pierwiastkami rozważanego równania.

Jeżeli funkcja $f(x)$ jest wielomianem algebraicznym, to istnieją metody analitycznego izolowania pierwiastków. W większości rozważanych przez nas przypadków poszukiwania przybliżonych rozwiązań, będziemy zakładali, że $f(x)$ jest klasy C^2 , a także będziemy przyjmować dodatkowe warunki na jej pierwszą i drugą pochodną.

Należy również pamiętać o oszacowaniu błędu uzyskiwanych przybliżeń. Mamy do czynienia z trzema źródłami tych błędów

- błąd przybliżenia wynikający ze stosowanej metody,
- wpływ błędów współczynników równania $f(x) = 0$ oraz
- dokładność obliczeń przybliżonych.

3.1.1. Metoda połowienia

Mamy dane równania

$$f(x) = 0 \tag{3.1}$$

przy czym o funkcji $f(x)$ zakładamy, że jest ciągła na przedziale domkniętym $\langle a; b \rangle$, wewnątrz którego znajduje się dokładnie jeden pierwiastek i na którego końcach wartości funkcji $f(x)$ mają przeciwne znaki ($f(a)f(b) < 0$). Zadanie polega na znalezieniu przybliżonej wartości pierwiastka równania (3.1). Aby go znaleźć, dzielimy przedział $\langle a; b \rangle$ na połowy punktem

$$x_1 = \frac{b + a}{2}$$

Jeżeli $f(x_1) = 0$, to x_1 jest szukany pierwiastkiem, jeśli zaś $f(x_1) \neq 0$, to z otrzymanych dwóch przedziałów $\langle a; x_1 \rangle$ i $\langle x_1; b \rangle$ wybieramy ten, na końcach którego funkcja $f(x)$ ma przeciwne znaki. Z kolei ten przedział dzielimy na dwie połowy, ponownie badamy wartość funkcji $f(x)$ w punkcie x_2 i znaki funkcji $f(x)$ na końcach otrzymanych przedziałów itd. W wyniku takiego postępowania po pewnej liczbie kroków albo otrzymamy pierwiastek dokładny $f(x_n) = 0$, albo ciąg przedziałów takich, że

$$f(x_i)f(x_{i+1}) < 0 \tag{3.2}$$

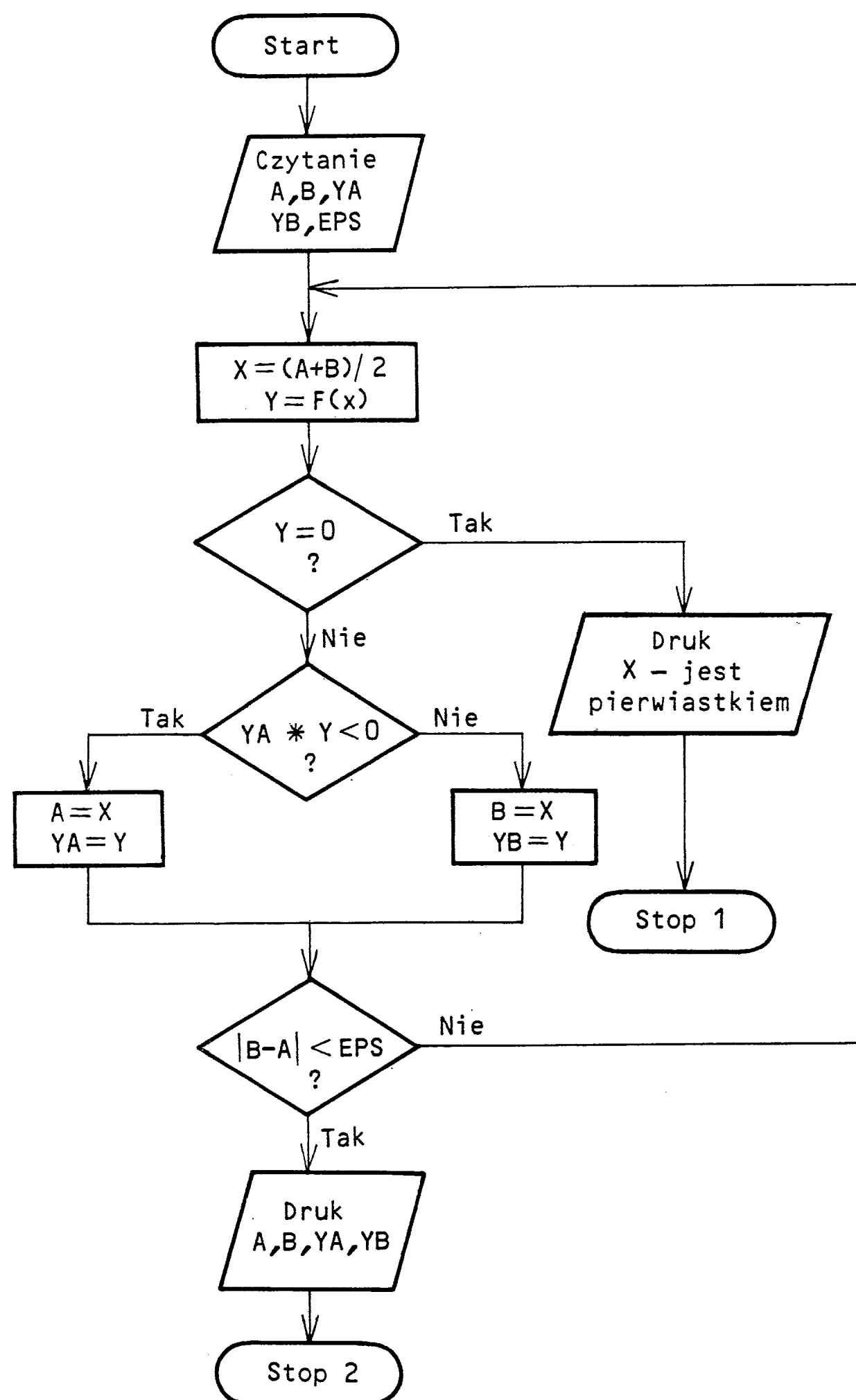
przy czym x_i oraz x_{i+1} są odpowiednio początkiem i końcem i -tego przedziału, a jego długość

$$|x_i - x_{i+1}| = \frac{1}{2^i} (b - a) \tag{3.3}$$

Ponieważ lewe końce ciągu przedziałów tworzą ciąg niemalejący i ograniczony z góry, a prawe końce ciąg nierosnący i ograniczony z dołu, więc z (3.3) wynika, że istnieje ich wspólna granica α . Ze względu na stosowaną metodę

obliczeniową ten sposób znajdowania pierwiastków nazywamy *metodą połowienia* (także *metodą równego podziału* albo *metodą bisekcji*).

Na rysunku 3.1 jest przedstawiony schemat blokowy postępowania według omówionej powyżej metody połowienia. Zmienne A i B określają końce przedziału w każdej z kolejnych iteracji; YA i YB są to wartości funkcji $y = f(x)$ odpowiednio w tych punktach, a zmienna EPS — to przyjęta dokładność, z jaką chcemy obliczyć pierwiastek. Gdy długość przedziału w danej iteracji (czyli wielkość $B - A$) jest mniejsza od wartości EPS , następuje przerwanie procedury obliczeniowej.



3.1

Przykład 1. Znaleźć metodą połowienia pierwiastek równania $x^3 + x^2 - 3x - 3 = 0$, położony w przedziale $\langle 1; 2 \rangle$.

Mamy $f(x) = x^3 + x^2 - 3x - 3$, zatem $f(1) = -4$ oraz $f(2) = 3$. Ponieważ pochodna $f'(x) = 3x^2 + 2x - 3$ jest na przedziale $\langle 1; 2 \rangle$ dodatnia, więc jest to przedział izolacji pierwiastka. Postępując zgodnie z omówioną metodą obliczamy kolejne przybliżenia szukanego pierwiastka. Wyniki obliczeń są zestawione w tablicy 3/1.

3/1

x	$f(x)$	x	$f(x)$
1.0	-4.0	1.6875	-0.40942
2.0	3.0	1.71875	-0.12487
1.5	-1.874	1.73437	0.02198
1.75	0.17187		
1.625	-0.94335	1.73205	-0.00000

Podstawową zaletą metody połowienia oprócz jej dużej prostoty jest pewność, że w każdej kolejnej iteracji szukany pierwiastek leży między dwiema wartościami zmiennej x , w których funkcja $f(x)$ zmienia znak. Wielokrotne powtarzanie obliczeń doprowadzić musi zatem do uzyskania wyniku z żadaną dokładnością (na rys. 3.1 dokładność obliczeń określa zmienna EPS). Oczywiście to założone EPS powinno być większe od maksymalnej granicznej dokładności obliczeń, w przeciwnym przypadku, gdy EPS będzie mniejsze, możemy otrzymać wynik jedynie z maksymalną graniczną dokładnością.

W pozostałych metodach przybliżonego rozwiązywania równań sprawa dokładności otrzymanych przybliżeń nie jest tak oczywista, a jej określenie jest znacznie trudniejsze.

Przykład 2. Napisać program na rozwiązywanie równania metodą połowienia.

Napişemy program w miarę uniwersalny, tak aby prosta wymiana podprogramu FUNCTION F pozwalała na rozwiązywanie dowolnego równania. Danymi wejściowymi do programu są x_1 i x_2 — końce pierwszego przedziału, przy czym na początku jest sprawdzany znak iloczynu $F(x_1) * F(x_2)$, który powinien być ujemny. Daną wejściową jest także EPS , określająca dokładność obliczeń.

Działanie programu jest przerywane gdy:

$$F(x_n) = 0, F(x_1) * F(x_2) > 0 \text{ lub gdy } |F(x_n)/F'(x_n)| < EPS,$$

przy czym wartość pochodnej $F'(x_n)$ zastępujemy ilorazem różnicowym.

Jako funkcję przykładową rozpatrzmy

$$F(x) = e^{-x} - \sin(\pi x/2)$$

```

PROGRAM bisekcja;
VAR x,x1,x2,fx,fx1,fx2,eps: real;
    j: integer;
FUNCTION F(x: real): real;
  CONST pi = 3.14159265;
  BEGIN
    F := exp(-x) - sin(pi*x/2);
  END;
BEGIN
  REPEAT
    readln(x1,x2,eps);
    fx1 := F(x1);
    fx2 := F(x2);
    IF fx1*fx2 > 0 THEN
      writeln('wartosci funkcji w koncach nie sa roznych znakow');
    
```



```

UNTIL fx1*fx2 <= 0;
IF fx1*fx2 = 0 THEN
  BEGIN
    IF fx1 = 0 THEN x := x1 ELSE x := x2;
    writeln('wielkosc wejscowa x = ',x:8:3,' jest pierwiastkiem');
    EXIT
  END;
{poczatek polowienia}
FOR j := 1 TO 50 DO
  BEGIN
    x := (x1+x2)/2;
    fx := F(x);
    fxx := (fx2-fx1)/(x2-x1);
    writeln(' w punkcie x=',x:8:3,' f(x) = ',fx:8:3);
    IF ABS(fx/fxx)- eps > 0 THEN
      BEGIN
        IF fx*fx1 < 0 THEN
          BEGIN
            x2 := x;
            fx2 := fx;
          END
        ELSE
          BEGIN
            IF fx*fx1 = 0 THEN
              BEGIN
                writeln('x = ',x:8:3,' jest pierwiastkiem');
                writeln('potrzeba bylo', j:4,' iteracji');
                EXIT
              END;
            x1 := x;
            fx1 := fx;
          END
        END
      ELSE
        BEGIN
          writeln('x = ',x:8:3,' jest pierwiastkiem');
          writeln('potrzeba bylo',j:4,' iteracji');
          EXIT
        END
      END;
    writeln('zbyt malo iteracji');
  END.

```

Na rysunku 3.2 przedstawiono schemat blokowy programu.

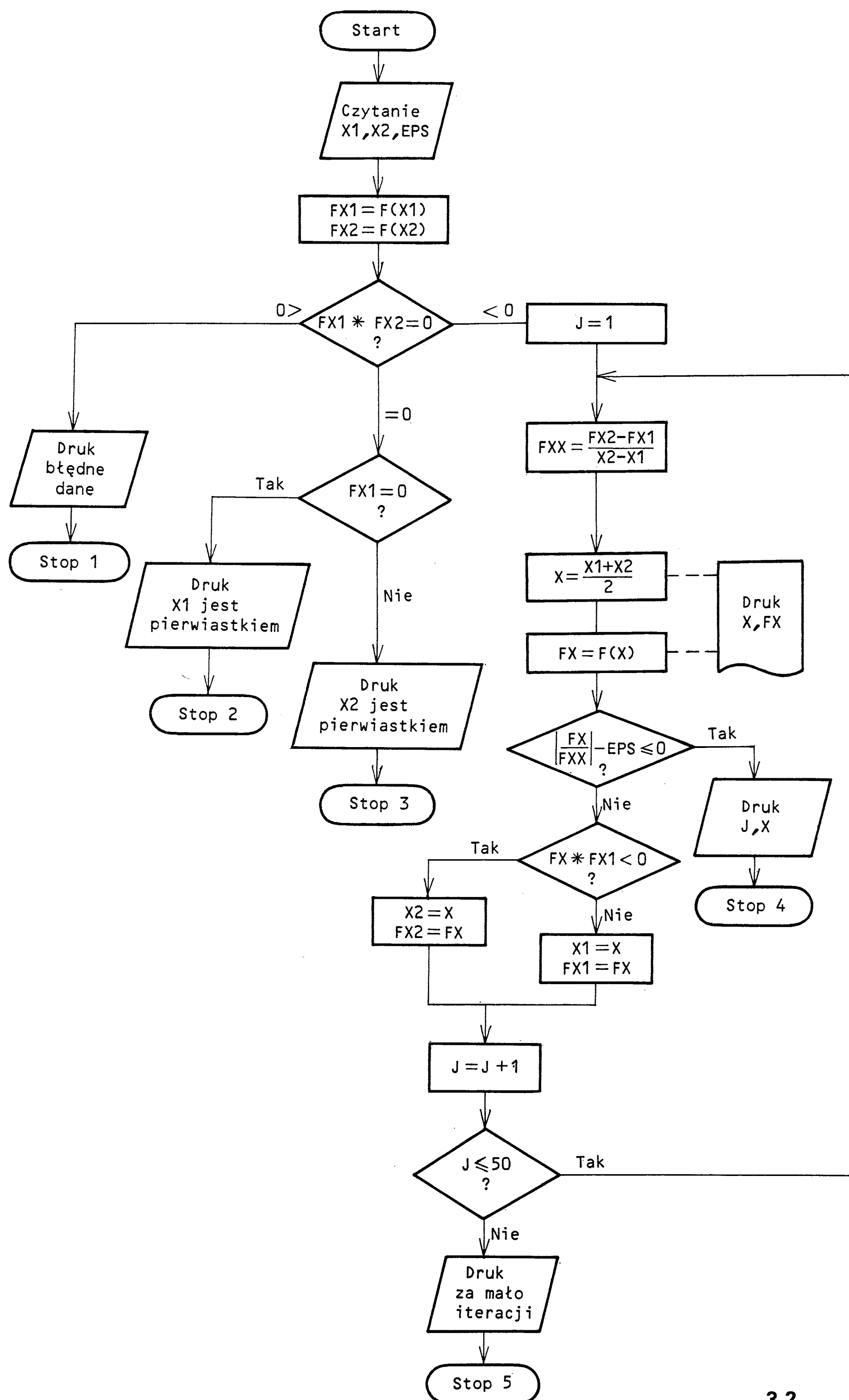
Ćwiczenia

1. Omówić metodę połowienia stosowaną do rozwiązywania równań nieliniowych z jedną niewiadomą.

2. Równanie $x^2 - 2 = 0$ ma pierwiastki $x_{1,2} = \pm \sqrt{2} \approx \pm 1,414214$. Stosując metodę połowienia, obliczyć dodatni pierwiastek tego równania zaczynając od przedziału $\langle 1; 2 \rangle$. Ile iteracji należy wykonać, aby obliczyć pierwiastek z dokładnością do czterech miejsc dziesiętnych? Jaki jest maksymalny błąd po tej liczbie iteracji?

3. Równanie $x^3 - x + 1 = 0$ ma tylko jeden pierwiastek rzeczywisty. (Dlaczego?). Znaleźć ten pierwiastek metodą połowienia.

4. Pokazać, że równanie $\sin x + x - 1 = 0$ ma pierwiastek w przedziale $\langle 0; 1 \rangle$. Ile trzeba



wykonać iteracji, aby metodą połowienia otrzymać przybliżoną wartość pierwiastka z błędem nie przekraczającym $0,5 \cdot 10^{-4}$?

5. Metodą połowienia znaleźć dodatni pierwiastek równania $x^2 - x - 1 = 0$ z dokładnością do 10^{-2} .

6. Metodą połowienia znaleźć wartość zmiennej x , dla której przecinają się wykresy funkcji $y = 3x$ i $y = e^x$. Wymagana dokładność obliczeń wynosi 10^{-4} .

7. Korzystając z metody połowienia znaleźć najmniejszy dodatni pierwiastek równania: a) $\lg x - x - 1 = 0$, b) $x^3 - x^2 - 2x + 1 = 0$, c) $2e^{-x} - \sin x = 0$.

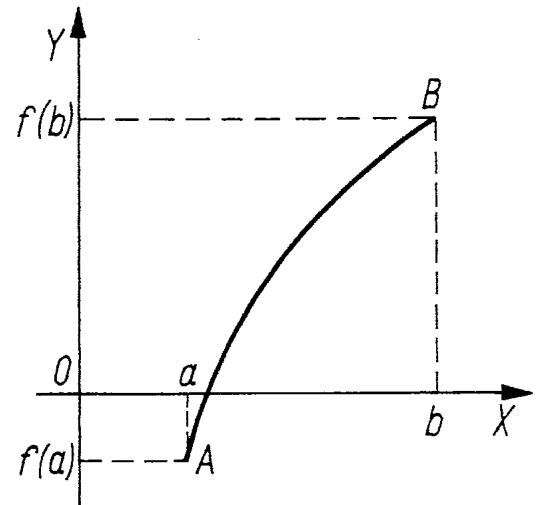
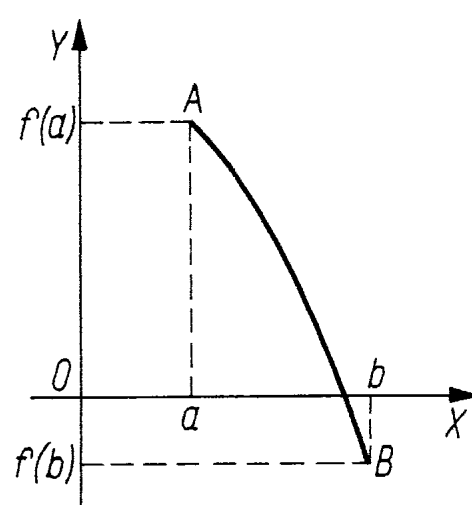
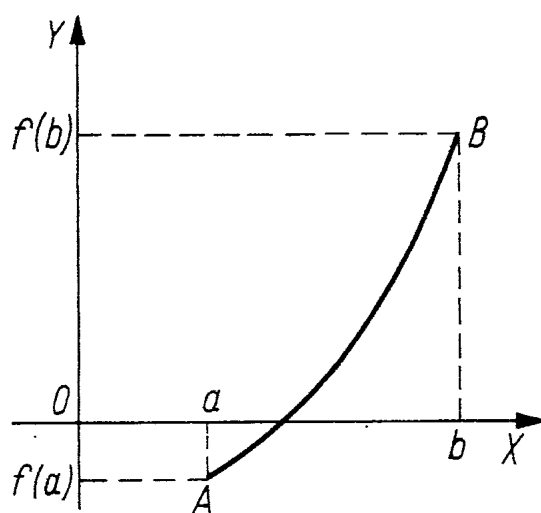
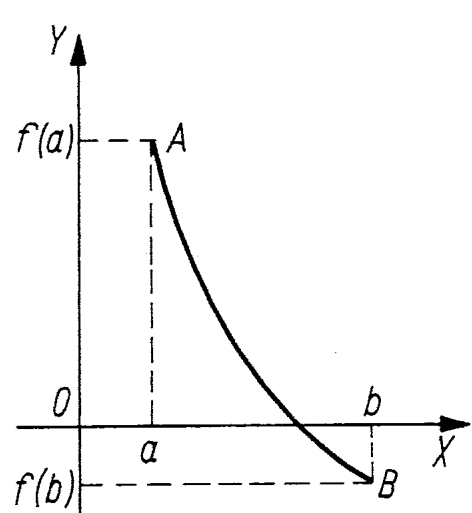
Odpowiedzi. 2. Trzynastą iteracją daje wynik $x = 1,4141844$. Maksymalny błąd w każdym kroku, określony jako różnica między dwiema ostatnimi iteracjami, jest równy $1/2^n$. Dla $n = 13$ mamy $(1/2)^{13} \approx 0,0001222$. 3. $x = 1,3247421$. 4. 14 iteracji. 5. $x \approx 1,6$. 6. $x = 0,6191$. 7. a) $x = 1,13226773$, b) $x = 0,4450$, c) $x = 0,92102480$.

3.1.2. Reguła fałsi i metoda siecznych

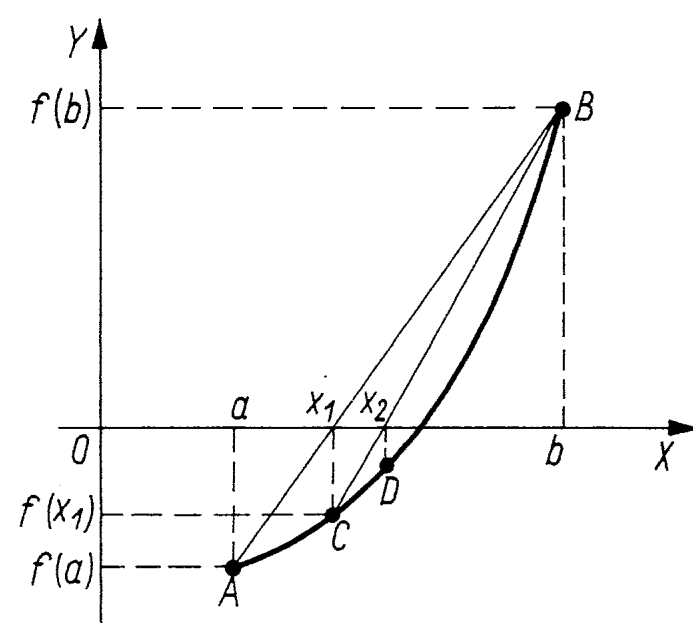
Nazwa metody pochodzi od łacińskich słów: *regula* — linia i *falsus* — fałszywy; jest to zatem metoda fałszywego założenia liniowości funkcji. Przyjmijmy, że w rozpatrywanym przedziale $\langle a; b \rangle$ równanie (3.1) ma dokładnie jeden pierwiastek, że jest to pierwiastek pojedynczy i że $f(a)f(b) < 0$. Ponadto niech $f(x)$ będzie na przedziale $\langle a; b \rangle$ funkcją klasy C^2 i niech jej pierwsza i druga pochodna mają stały znak na tym przedziale. Założenia te są potrzebne do oszacowania błędu i umożliwiają ustalenie stałego punktu iteracji. Wobec tych założeń wykres funkcji $y = f(x)$ może być tylko jednym z czterech rodzajów przedstawionych na rys. 3.3a—d.

Rozpatrzmy przypadek, gdy na przedziale $\langle a; b \rangle$ pochodne $f'(x)$ i $f''(x)$ są dodatnie (dla pozostałych przypadków rozumowanie jest analogiczne).

Przez punkty $A(a, f(a))$ i $B(b, f(b))$ prowadzimy cięciwę (rys. 3.4)



3.3



3.4

o równaniu

$$y - f(a) = \frac{f(b) - f(a)}{b - a} (x - a) \quad (3.4)$$

Odciętą x_1 punktu, w którym cięciwa AB przecina oś OX , przyjmujemy za pierwsze przybliżenie szukanego pierwiastka równania (3.1). Stąd

$$x_1 = a - \frac{f(a)}{f(b) - f(a)} (b - a) \quad (3.5)$$

Jeżeli $f(x_1) = 0$, to oczywiście x_1 jest szukanym pierwiastkiem α i zadanie zostało zakończone.

Założmy, że $f(x_1) \neq 0$. 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. 3.4). Odciętą x_2 punktu, w którym ta cięciwa przetnie oś OX da nam drugie przybliżenie pierwiastka α . Dla uproszczenia rozumowania przyjęliśmy, że na przedziale $\langle a; b \rangle$ pochodne $f'(x) > 0$ i $f''(x) > 0$. Przy przyjętych założeniach krzywa $y = f(x)$ jest skierowana wypukłością w dół, więc znajduje się poniżej cięciwy AB , a ponieważ $f(a) < 0$, zatem w kolejnych przybliżeniach punkt B pozostaje nieruchomy. (Gdyby np. $f'(x) > 0$ i $f''(x) < 0$ na przedziale $\langle a; b \rangle$, wówczas 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

$$x_1, x_2, x_3, \dots, x_n, \dots \quad (3.6)$$

określonego wzorem rekurencyjnym

$$x_0 = a, x_{k+1} = x_k - \frac{f(x_k)}{f(b) - f(x_k)} (b - x_k) \quad (3.7)$$

$$k = 1, 2, \dots, n$$

Można wykazać, że przy przyjętych założeniach ciąg (3.6) jest rosnący i ograniczony, a więc zbieżny, i że jego granicą jest szukany pierwiastek α . (Przy nieruchomym punkcie B , kolejne wyrazy ciągu (3.6) są mniejsze od α oraz $f(x_k) < 0$ dla każdego k). Przechodząc do granicy dla $n \rightarrow \infty$ z równości (3.7) otrzymujemy

$$g = g - \frac{f(g)}{f(b) - f(g)} (b - g)$$

przy czym $g = \lim_{n \rightarrow \infty} x_n$ ($a < g < b$). Stąd $f(g) = 0$, a ponieważ zakładaliśmy

istnienie tylko jednego pierwiastka w przedziale $\langle a; b \rangle$, więc $g \equiv \alpha$.

Błąd n -tego przybliżenia można ocenić korzystając z twierdzenia Lagran-

ge'a (o przyrostach)

$$f(x_n) - f(\alpha) = f'(c)(x_n - \alpha)$$

gdzie c jest zawarte w przedziale o końcach x_n i α . Ponieważ $f(\alpha) = 0$, więc

$$|x_n - \alpha| \leq \frac{|f(x_n)|}{m} \quad (3.8)$$

$$m = \inf_{x \in \langle a; b \rangle} |f'(x)| \quad (3.9)$$

Błąd bezwzględny przybliżenia x_n można oszacować znając wartości dwu kolejnych przybliżeń x_{n-1} i x_n [17]. Przyjmijmy, że kolejne przybliżenia funkcji przedstawionej na rys. 3.4 oblicza się według wzoru (3.7). Po przekształceniu mamy

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

a ponieważ $f(\alpha) = 0$, więc

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

i z twierdzenia Lagrange'a otrzymujemy

$$(\alpha - x_k) f'(\xi_k) = (x_{k+1} - x_k) f'(\bar{x}_k)$$

przy czym ξ_k leży w przedziale $(x_k; \alpha)$, a $\bar{x}_k \in (x_k; b)$. Po wykonaniu mnożenia i obustronnym dodaniu wyrażenia $-x_{k+1} f'(\xi_k)$, otrzymujemy

$$|\alpha - x_{k+1}| = \frac{|f'(\bar{x}_k) - f'(\xi_k)|}{|f'(\xi_k)|} |x_{k+1} - x_k| \quad (3.10)$$

Ponieważ $\xi_k \in \langle a; b \rangle$, $\bar{x}_k \in \langle a; b \rangle$ i $f'(x) > 0$, więc

$$|f'(\bar{x}_k) - f'(\xi_k)| \leq M - m$$

gdzie $m = \inf_{x \in \langle a; b \rangle} |f'(x)|$, $M = \sup_{x \in \langle a; b \rangle} |f'(x)|$, i ostatecznie

$$|\alpha - x_{k+1}| \leq \frac{M - m}{m} |x_{k+1} - x_k| \quad (3.11)$$

Jeżeli $M \leq 2m$, wówczas otrzymujemy wygodniejsze oszacowanie

$$|\alpha - x_{k+1}| \leq |x_{k+1} - x_k| \quad (3.11a)$$

Oszacowanie (3.11) jest zazwyczaj zbyt pesymistyczne, a ponadto wymaga znajomości m i M . Dla przybliżeń w niewielkim otoczeniu pierwiastka α możemy skorzystać z ilorazu wartości funkcji i pochodnej, który dla małych wartości $x_{k+1} - x_k$ zastępujemy ilorazem różnicowym

$$|\alpha - x_{k+1}| \sim \left| \frac{f(x_{k+1})}{f'(x_{k+1})} \right| \sim \left| \frac{x_{k+1} - x_k}{f(x_{k+1}) - f(x_k)} \right| \cdot |f(x_{k+1})| \quad (3.12)$$

Przykład 1. Znaleźć dodatni pierwiastek równania

$$x^3 + x^2 - 3x - 3 = 0$$

Sporządzamy szkic wykresu funkcji $f(x) = x^3 + x^2 - 3x - 3$, z którego wynika, że równanie to ma dokładnie jeden pierwiastek w przedziale $(1; 2)$. Funkcja ta spełnia żądane założenia, bowiem

$$f'(x) = 3x^2 + 2x - 3 \quad \text{oraz} \quad f''(x) = 6x + 2$$

są dodatnie dla $x > 1$. Mamy zatem do czynienia z przypadkiem pokazanym na rys. 3.4, przy czym $f(1) = -4$, a $f(2) = 3$.

Piszemy równanie cięciwy przechodzącej przez punkty $A(1, -1)$, $B(2, 3)$

$$y + 4 = \frac{3 + 4}{2 - 1}(x - 1)$$

i przyjmując $y = 0$, $x_1 = 1,57142$. Znajdujemy następnie $f(x_1) = -1,36449$. Ponieważ $f(x_1) < 0$, więc prowadzimy drugą cięciwę przez punkty $B(2, 3)$, $C(1,57142, -1,36449)$, podstawiamy $y = 0$ i otrzymujemy drugie przybliżenie

$$x_2 = 1,70540$$

Oceńmy błąd przybliżenia x_2 . Mamy

$$m = \inf_{x \in \langle 1; 2 \rangle} |3x^2 + 2x - 3| = 2$$

$$f(x_2) = -0,24784$$

i ze wzoru (3.8) znajdujemy

$$|x_2 - \alpha| < \frac{0,24784}{2} < 0,124$$

Ponieważ ciąg przybliżeń jest rosnący, więc

$$1,70540 < \alpha < 1,8294$$

Jeżeli takie przybliżenie jest wystarczające, to zadanie traktujemy jako rozwiązanie, jeżeli nie, to obliczamy następne przybliżenia według wskazanego wyżej schematu.

W tablicy 3/2 zestawiono kilka kolejnych przybliżeń rozwiązania tego równania. Dla przybliżenia x_5 błąd wynikający z oszacowania (3.11) nie przekracza 0,00297, a błąd wynikający z oszacowania (3.8) nie przekracza 0,00052.

3/2

x	$f(x)$
$a = 1$	-4
$b = 2$	3
$x_1 = 1,57142$	$-1,36449$
$x_2 = 1,70540$	$-0,24784$
$x_3 = 1,72788$	$-0,03936$
$x_4 = 1,73140$	$-0,00615$
$x_5 = 1,73194$	$-0,00105$

Podsumowując, metoda regula falsi jest zbieżna dla dowolnej funkcji ciągłej na przedziale $\langle a; b \rangle$ (oczywiście $f(a)f(b) < 0$), 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 $f''f > 0$, jest stałym punktem iteracji — wszystkie cięciwy przechodzą przez ten punkt. Proces iteracyjny kończymy, gdy dwa kolejne przybliżenia różnią się (co do modułu) o mniej niż zadane ε (wszystkie uwagi dotyczące dokładności są takie same jak w metodzie połowienia). Wadą omówionej metody jest jej stosunkowo powolna zbieżność. (Uwagi dotyczące szybkości zbieżności poszczególnych metod, ich rzędu itp. są omówione w p. 3.1.5 tego rozdziału).

Metodę regula falsi można znacznie ulepszyć, jeżeli zrezygnujemy z żądania, aby funkcja $f(x)$ miała w punktach wytyczających następną cięciwę różne znaki, natomiast do wyznaczenia $(n+1)$ -ego przybliżenia będziemy korzystali zawsze z punktów x_n i x_{n-1} . Mamy wówczas

$$x_{n+1} = x_n - \frac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})} \quad (3.13)$$

Ta metoda nosi nazwę *metody siecznych*. Jej zbieżność jest znacznie szybsza; niestety zdarzają się przypadki, gdy może ona 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). Program obliczeniowy powinien także zapewniać możliwość przerwania iteracji, jeżeli różnica między kolejnymi przybliżeniami zamiast maleć zaczyna szybko wzrastać. W takim przypadku należy przeprowadzić powtórную lokalizację pierwiastka — znacznie zawężając początkowy przedział izolacji.

Przykład 2. Stosując metodę siecznych znaleźć pierwiastek równania $x^3 + x^2 - 3x - 3 = 0$, położony w przedziale $(1; 2)$. Mamy (tabl. 3/3)

3/3

	x	$f(x)$
a	1	-4
b	2	-3
x_1	1.57142	-1.36449
x_2	1.70540	-0.24784
x_3	1.73513	0.02920
x_4	1.73199	0.000576
x_5	1.73193	

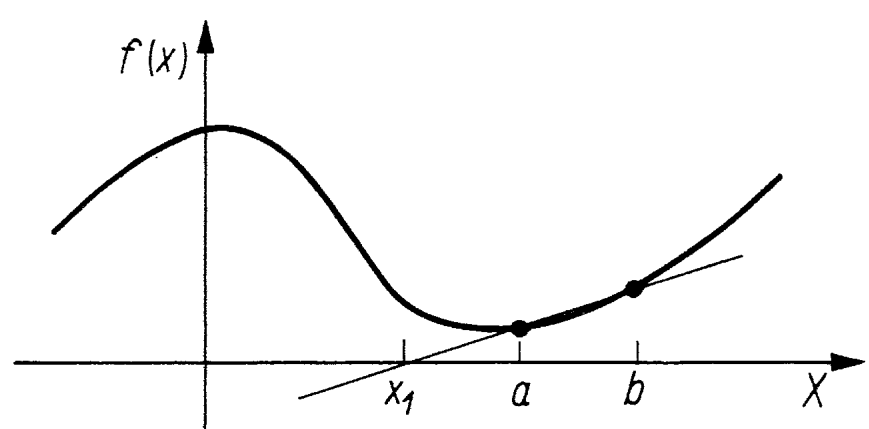
$$x_1 = 2 - \frac{3}{7}(2 - 1) = 1,57142$$

$$x_2 = 2 - \frac{3}{4,36449}(2 - 1,57142) = 1,70540$$

itd.

Jak widać, w metodzie siecznych znacznie szybciej niż w metodzie regula falsi otrzymujemy wynik przy tej samej zadanej dokładności. Zbieżność metody regula falsi w porównaniu z metodą siecznych jest znacznie wolniejsza w pobliżu pierwiastka.

Przy stosowaniu metody siecznych jest niezbędne, aby pierwsza iteracja zaczynała się z punktów, w których funkcja ma różne znaki, w przeciwnym razie możemy wykryć nieistniejący pierwiastek (rys. 3.5), co jest szczególnie niebezpieczne przy obliczeniach prowadzonych na maszynach cyfrowych.



3.5

Ćwiczenia

1. Omówić metodę regula falsi i ocenę błędu.
2. Jaka jest różnica pomiędzy metodą regula falsi a metodą siecznych?
3. Obliczyć metodą regula falsi przybliżony pierwiastek równania $\sin x + x - 1 = 0$ położony między $a = 0$ i $b = 1$. Obliczenia zakończyć, gdy $|\sin x_k + x_k - 1| < \frac{1}{2} \cdot 10^{-2}$.
4. Rozwiązać przykłady podane w ćwic. 7, p. 3.1.1 stosując metodę regula falsi.
5. Obliczyć metodą siecznych pierwiastek równania $e^x - 3x^2 = 0$ położony w pobliżu punktu $x_0 = -0,5$.
6. Znaleźć miejsce zerowe funkcji $y = 3x + \sin x - e^x$ w przedziale $\langle 0; 1 \rangle$ za pomocą metody siecznych.
7. Napisać schemat blokowy programu do obliczania przybliżonej wartości pierwiastka równania $2x^4 - 5x^2 + x + 1 = 0$ zadaną dokładnością ε .

Odpowiedzi. 3. 0,5110. 5. $x = -0,458962$. 6. $x = 0,3604$.

3.1.3. Metoda Newtona. Metody zmodyfikowane dla pierwiastków wielokrotnych

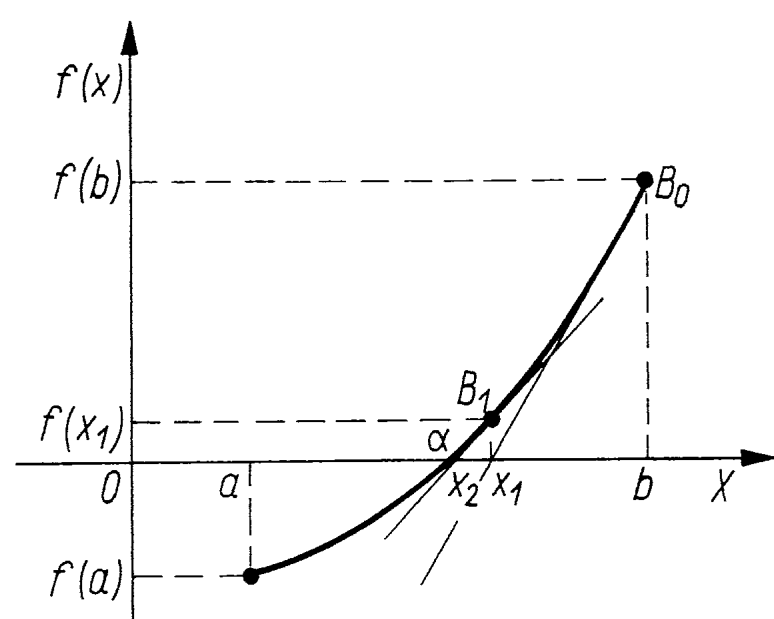
Przybliżoną wartość pierwiastka równania (3.1) przy założeniu, że na przedziale $\langle a; b \rangle$, w którym położony jest pierwiastek, funkcja $f(x)$ ma w końcach różne znaki oraz pierwsza i druga pochodna mają stały znak, można wyznaczyć też *metodą Newtona*, zwaną także *metodą stycznych*. W metodzie tej z tego końca przedziału $\langle a; b \rangle$, w którym funkcja $f(x)$ ma ten

sam znak co $f''(x)$ prowadzimy styczną do wykresu funkcji $y = f(x)$ i punkt x_1 , w którym ta styczna przecina oś OX przyjmujemy za pierwsze przybliżenie szukanego pierwiastka α . Oczywiście $f(x_1)$ ma ten sam znak co rzędna punktu, z którego wyprowadzona była styczna. Jeżeli otrzymane przybliżenie jest za mało dokładne, to z punktu o współrzędnych $(x_1, f(x_1))$ prowadzimy następną styczną. Punkt x_2 przecięcia tej stycznej z osią OX jest drugim przybliżeniem pierwiastka. W ten sposób otrzymujemy kolejne wyrazy ciągu przybliżeń

$x_1, x_2, \dots, x_n, \dots$

zbieżnego monotonicznie do α [17].

Na rysunku 3.6 pokazano geometryczną interpretację metody Newtona dla przypadku, gdy zarówno pierwsza, jak i druga pochodna funkcji $f(x)$ są



3.6

dodatnie na rozpatrywanym przedziale. Styczną prowadzimy z punktu B_0 (bowiem $f'(b) > 0$ i $f''(b-) > 0$). Jej równanie ma postać

$$y - f(b) = f'(b)(x - b)$$

Przyjmując $y = 0$, otrzymamy

$$x_1 = b - \frac{f(b)}{f'(b)} \quad (3.14)$$

Wykażemy, że x_1 leży bliżej α niż b . Korzystając ze wzoru Taylora mamy

$$f(\alpha) = f(b) + f'(b)(\alpha - b) + \frac{1}{2}f''(c)(\alpha - b)^2$$

gdzie $c \in (\alpha; b)$. Ponieważ $f(\alpha) = 0$, więc

$$\alpha = b - \frac{f(b)}{f'(b)} - \frac{1}{2} \frac{f''(c)}{f'(b)} (\alpha - b)^2 \quad (3.15)$$

i wobec (3.14)

$$\alpha - x_1 = -\frac{1}{2} \frac{f''(c)}{f'(b)} (\alpha - b)^2 < 0$$

gdyż przy przyjętych założeniach $f''(c) > 0$ i $f'(b) > 0$. Stąd $x_1 > \alpha$, a ponieważ z (3.14)

$$x_1 - b = -\frac{f(b)}{f'(b)} < 0$$

więc $x_1 < b$.

Z twierdzenia Lagrange'a mamy

$$f(x_1) - f(\alpha) = f'(c_1)(x_1 - \alpha)$$

przy czym $c_1 \in (\alpha, x_1)$. Ponieważ $f(\alpha) = 0$ i $f'(c_1) > 0$, więc $f(x_1) > 0$, a stąd wynika, że styczna poprowadzona w punkcie $B_1(x_1, f(x_1))$ będzie miała identyczne własności co styczna poprowadzona w punkcie B_0 . Z równania drugiej stycznej

$$y - f(x_1) = f'(x_1)(x - x_1)$$

otrzymamy

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} \quad (3.16)$$

Ogólnie, równanie stycznej w punkcie $B_n(x_n, f(x_n))$, $n = 0, 1, 2, \dots$, ma postać

$$y - f(x_n) = f'(x_n)(x - x_n) \quad (3.17)$$

co pozwala na podanie wzoru rekurencyjnego opisującego kolejne wyrazy ciągu przybliżeń

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad (3.18)$$

Tak jak w poprzednich metodach, ciąg przybliżeń jest ciągiem malejącym ($x_{n+1} < x_n$) ograniczonym z dołu ($\alpha < x_n$), a więc zbieżnym. Przechodząc do granicy w równości (3.18) dla $n \rightarrow \infty$ mamy

$$g = g - \frac{f(g)}{f'(g)}$$

czyli $f(g) = 0$, a więc $g = \alpha$.

Błąd n -tego przybliżenia można oszacować korzystając z nierówności (3.8)

$$|\alpha - x_n| \leq \frac{|f(x_n)|}{m}$$

Ponieważ zgodnie ze wzorem Taylora

$$\begin{aligned} f(x_n) &= f[x_{n-1} + (x_n - x_{n-1})] = f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) + \\ &\quad + \frac{1}{2} f''(\xi_{n-1})(x_n - x_{n-1})^2 \end{aligned}$$

przy czym ξ_{n-1} leży w przedziale o końcach x_n i x_{n-1} , więc ze wzoru (3.18) wynika, że

$$f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) = 0$$

czyli

$$|f(x_n)| \leq \frac{1}{2} M (x_n - x_{n-1})^2$$

gdzie $M = \sup_{x \in \langle a; b \rangle} |f''(x)|$. Stąd mamy

$$|\alpha - x_n| \leq \frac{1}{2} \frac{M}{m} (x_n - x_{n-1})^2 = \frac{M}{2m} \left(\frac{f(x_n)}{f'(x_n)} \right)^2 \quad (3.19)$$

Podobnie jak w przypadku metody regula falsi, dla x_{n+1} dostatecznie bliskich α mamy

$$|\alpha - x_{n+1}| \approx \left| \frac{f(x_n)}{f'(x_n)} \right|$$

co pozwala na przerywanie procedury iteracyjnej dla

$$|x_{n+1} - x_n| < \varepsilon$$

gdzie ε jest zadane z góry albo jest oszacowaniem błędu obarczającego $f(x_n)/f'(x_n)$.

Przykład 1. Stosując metodę Newtona obliczyć dodatni pierwiastek równania z przykładu 1 w p. 3.2.1.

Jako punkt początkowy wybieramy punkt $x_0 = 2$. Kolejnymi przybliżeniami są wówczas

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 2 - \frac{3}{13} \approx 1,76923$$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 1,76923 - \frac{0,36048}{9,92897} \approx 1,73292$$

$$x_3 \approx 1,73205$$

Gdybyśmy jako punkt początkowy wybrali $x_0 = 1$ proces iteracyjny też byłby zbieżny, ale znacznie wolniej (tabl. 3/4). Jak widać, dopiero od czwartego wyrazu ciąg przybliżeń zaczyna być szybko zbieżny.

3/4

i	x_i	$f(x_i)$	$f'(x_i)$
0	1.0	-4.0	2.0
1	3.0	24.0	30.0
2	2.2	5.888	15.92
3	1.83015	0.98899	10.7086
4	1.7578	0.05457	9.5354
5	1.73195		

Przy rozwiązywaniu równań metodą stycznych wygodnie jest skorzystać z następującego twierdzenia, które podamy tutaj bez dowodu.

Tw. Jeżeli mamy przedział $\langle a; b \rangle$ taki, że:

(i) $f(a)$ i $f(b)$ mają przeciwne znaki,

(ii) $f''(x)$ jest ciągła i nie zmienia znaku na $\langle a; b \rangle$,

(iii) styczne do krzywej $y = f(x)$ poprowadzone w punktach o odciętych a i b przecinają oś OX wewnątrz przedziału $\langle a; b \rangle$, wówczas równanie $f(x) = 0$ ma dokładnie jeden pierwiastek α w przedziale $\langle a; b \rangle$ i metoda Newtona jest zbieżna do α dla dowolnego punktu startowego $x_0 \in \langle a; b \rangle$.

Znanym przykładem zastosowania metody Newtona jest algorytm obliczania pierwiastka kwadratowego. Pierwiastek kwadratowy z liczby dodatniej c jest dodatnim pierwiastkiem równania

$$x^2 - c = 0$$

Mamy tutaj $f(x) = x^2 - c$ i $f'(x) = 2x$, zatem na podstawie wzoru (3.18)

$$x_{n+1} = x_n - \frac{x_n^2 - c}{2x_n} \quad (x_n \neq 0) \text{ i po uproszczeniu otrzymujemy}$$

$$x_{n+1} = \frac{1}{2} \left(x_n + \frac{c}{x_n} \right) \quad (3.20)$$

Wszystkie założenia powyższego twierdzenia są spełnione na przedziale $\langle a; b \rangle$ takim, że $0 < a < c^{1/2}$ i $b > \frac{1}{2}(a + c/a)$.

Dotychczasowe rozważania prowadziliśmy przy założeniu, że szukany pierwiastek α jest pojedynczy. Omówimy teraz pokrótce przydatność poszczególnych metod w przypadku, gdy α jest pierwiastkiem wielokrotnym. Przypomnimy najpierw definicję wielokrotności pierwiastka.

Def. Liczbę α nazywamy r -krotnym ($r \geq 2$) pierwiastkiem równania $f(x) = 0$ wtedy i tylko wtedy, gdy jest $(r - 1)$ -krotnym pierwiastkiem równania $f'(x) = 0$.

Jeżeli α jest pierwiastkiem o parzystej krotności, to metoda połowienia, omówiona w p. 3.1.1 traci sens, natomiast dla krotności nieparzystej ($r \geq 3$), gdy $f(a)f(b) < 0$ pozostaje słuszna. Podobnie przedstawia się sprawa dla metody regula falsi oraz metody siecznych, przy czym dla tej ostatniej obniża się rząd metody (uwagi dotyczące rzędu metody są opisane w p. 3.1.5). Metoda Newtona pozwala na znalezienie pierwiastka zarówno o nieparzystej, jak i o parzystej krotności, jeżeli tylko istnieje odpowiednie lewo- lub prawostronne sąsiedztwo szukanego pierwiastka o stałym znaku pierwszej i drugiej pochodnej, lecz tu także obniża się rząd metody.

Stosuje się również pewne modyfikacje omówionych metod pozwalające utrzymać ich rząd, a więc przyspieszające ich zbieżność. Jednym ze sposobów takiej modyfikacji jest przekształcanie zależności (3.18) do postaci

$$x_{n+1} = x_n - r \frac{f(x_n)}{f'(x_n)} \quad (3.21)$$

gdzie r jest krotnością pierwiastka (znaną a priori). Ponieważ na ogół krotność pierwiastka nie jest znana, często we wzorach (3.13) i (3.18) stosuje się podstawienie: zamiast funkcji $f(x)$ wprowadza się funkcję

$$u(x) = \frac{f(x)}{f'(x)} \quad (3.22)$$

Równanie $u(x) = 0$ ma identyczne pierwiastki jak równanie (3.1), ale wszystkie te pierwiastki są pojedyncze. Wzory (3.13) i (3.18) przyjmują wtedy postać

$$x_{n+1} = x_n - u(x_n) \frac{x_n - x_{n-1}}{u(x_n) - u(x_{n-1})} \quad (3.13a)$$

oraz

$$x_{n+1} = x_n - \frac{u(x_n)}{u'(x_n)} \quad (3.18a)$$

$$\text{gdzie } u'(x_n) = 1 - \frac{f''(x_n)}{f'(x_n)} u(x_n).$$

Przykład 2. Znaleźć pierwiastek równania

$$x^4 - 4x^3 - 5x^2 - 4x + 4 = 0$$

za pomocą: a) metody Newtona (wzór (3.18)), b) zmodyfikowanej metody Newtona dla $r = 2$ (wzór (3.21)) i c) zmodyfikowanej metody Newtona wykorzystującej podstawienie (3.22).

Mamy

$$f(x) = x^4 - 4x^3 + 5x^2 - 4x + 4 = (x - 2)^2(x^2 + 1)$$

$$f'(x) = 4x^3 - 12x^2 + 10x - 4$$

$$f''(x) = 12x^2 - 24x + 10$$

$$u(x) = \frac{f(x)}{f'(x)}$$

$$u'(x) = 1 - \frac{f(x)f''(x)}{[f'(x)]^2}$$

i jako x_0 przyjmujemy 3.

W tablicy 3/5 są zestawione wyniki otrzymane wszystkimi trzema metodami. Jak należało oczekiwać, zbieżność metod b) i c) jest znacznie szybsza niż metody a).

Przykład 3. Napisać program obliczania przybliżonej wartości pierwiastka równania $f(x) = 0$ z zastosowaniem metody Newtona (wzoru (3.18)).

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad \text{dla} \quad i = 0, 1, 2, \dots$$

3/5

x_i	Metoda a)	Metoda b)	Metoda c)
x_1	2.6156	2.2308	1.7963
x_2	2.3601	2.0183	1.9814
x_3	2.2007	2.00921	1.9997
x_4	2.1081	2.00462	
x_5	2.0562	2.00232	
x_6	2.0287		
x_7	2.0146		
x_8	2.00733		
x_9	2.00368		

Procedura iteracyjna jest przerywana, gdy po obliczeniu x_{i+1}

$$|\text{delta}| \leq EPS \quad \text{oraz} \quad |f(x_{i+1})| \leq 100 * EPS$$

gdzie EPS jest zadaną dokładnością a

$$\text{delta} = \begin{cases} \frac{x_{i+1} - x_i}{x_{i+1}} & \text{dla } |x_{i+1}| > 1 \\ x_{i+1} - x_i & \text{dla } |x_{i+1}| \leq 1 \end{cases}$$

Opis parametrów formalnych oraz zmiennych lokalnych podprogramu:

x — punkt startowy obliczania przybliżonej wartości pierwiastka, a po zakończeniu działania procedury — obliczona wartość pierwiastka,

EPS — dokładność obliczeń,

n — maksymalna liczba iteracji,

err — parametr wyjściowy, sygnalizacja błędów:

0 — brak błędów,

1 — nie osiągnięto wymaganej dokładności w n krokach iteracji,

2 — w trakcie obliczeń wartość pochodnej funkcji f była bliska lub równa zero,

$F(x, fx, fxx)$ — procedura zdefiniowana w bloku zewnętrznym programu obliczająca wartość funkcji fx i wartość pochodnej tej funkcji fxx w punkcie x .

Tekst programu:

```

PROCEDURE NEWTON(VAR x:real; eps:real; n:integer; VAR err:byte);
VAR t,fx,fxx,dx,delta: real;
    i: integer;
BEGIN
  err:=0;
  F(x,fx,fxx);
  { poczatek procedury iteracyjnej }
  FOR i := 1 TO n DO
    BEGIN
      IF fx = 0 THEN EXIT;
      IF (ABS(fxx) - 1.E-6) <= 0 THEN
        BEGIN
          err := 2;
          EXIT;
        END;
      dx := fx/fxx;
      t := x;
    
```

```

x := x - dx;
F(x, fx, fxx);
{ badanie dokladnosci }
delta := ABS(x - t);
IF ABS(x) > 1 THEN delta := delta/ABS(x);
IF ABS(delta) <= eps THEN
  IF ABS(fx) <= 100*eps THEN EXIT;
END;
err := 1;
END;

```

Ćwiczenia

1. Omówić metodę Newtona i podać błąd n -tego przybliżenia.
2. Udowodnić wzór (3.18).
3. Znaleźć pierwiastki równań z ćwiczenia 7 w p. 3.1.1 metodą Newtona.
4. Stosując metodę Newtona znaleźć pierwiastki równań: a) $x^3 - 2x + 1 = 0$, b) $x^4 + x^3 - 4x^2 - 3x + 3 = 0$,
5. Znaleźć dodatnie pierwiastki równania $e^x - 3x^2 = 0$ metodą Newtona.
6. Korzystając z metody Newtona znaleźć wzór rekurencyjny na obliczanie pierwiastka sześciennego z liczby dodatniej c . Wykonać kilka kroków obliczenia dla $c = 10$.
7. Stosując metodę Newtona do równania $x^n - c = 0$, wyprowadzić wzór na obliczanie pierwiastka n -tego stopnia z liczby dodatniej c .
8. Odwrotność liczby N można obliczyć za pomocą algorytmu iteracyjnego

$$x_{i+1} = x_i(2 - Nx_i)$$
a) Wyprowadzić ten algorytm stosując metodę Newtona do funkcji

$$f(x) = \frac{1}{x} - N$$

- b) Zaczynając z punktu $x_0 = 0,2$ obliczyć tym algorytmem odwrotność liczby $N = 4$.
9. Co się stanie, jeżeli we wzorze (3.20) wybierzemy $x_0 < 0$?
10. Równanie $x^2 - 4x + 4 = 0$ ma pierwiastek podwójny $x_0 = 2$. Zaczynając z punktu $x_1 = 1$ obliczyć przybliżoną wartość pierwiastka: a) metodą Newtona, b) zmodyfikowaną metodą Newtona według (3.21) i c) zmodyfikowaną metodą Newtona z podstawieniem (3.22).
11. Wyjaśnić, dlaczego metody regula falsi i metody siecznych nie można stosować do obliczania pierwiastków o parzystej krotności.
12. Narysować schemat blokowy programu obliczania przybliżonej wartości pierwiastka (pierwiastków) równania $f(x) = 0$ metodą Newtona.

Odpowiedzi. 4. a) $-1,618034$; b) $-1,732, -1,618, 0,618, +1,732$. 5. $0,91001; 3,73308$.

6. $x_{n+1} = \frac{1}{3}(2x_n + c/x_n^2)$. 7. $x_{r+1} = \frac{1}{n}[(n-1)x_r + c/x_r^{n-1}]$. 8. b) $x_0 = 0,2; x_1 = 0,24; x_2 = 0,2496; x_3 = 0,24999936, \dots$ 9. Otrzymamy procedurę zbieżną do $-\sqrt{c}$.

3.2. METODY POSZUKIWANIA ZER WIELOMIANÓW

Opisanymi w poprzednich punktach metodami można rozwiązywać dowolne równania zarówno algebraiczne, jak i przestępne, a po pewnych prostych modyfikacjach można również znajdować pierwiastki zespolone (z wyjątkiem metody regula falsi). W przypadku równań algebraicznych

$$f(z) = a_0 z^n + a_1 z^{n-1} + \dots + a_{n-1} z + a_n = 0 \quad (3.23)$$

jest wiele metod ułatwiających lokalizację jego pierwiastków oraz znajdowanie ich przybliżonych wartości. Metody te mają duże znaczenie praktyczne, istnieje bowiem bardzo wiele typowych problemów technicznych, w których jest niezbędne na przykład rozstrzygnięcie, czy wszystkie zera pewnego wielomianu położone są na płaszczyźnie Gaussa na lewo od osi urojonej, wewnątrz pewnego ograniczonego obszaru płaszczyzny Gaussa itp.

Zanim zajmiemy się omówieniem pewnych typowych metod znajdowania pierwiastków równania (3.23), podamy kilka twierdzeń oraz metod dotyczących liczby i rozmieszczenia zer wielomianu, i to zarówno rzeczywistych, jak i zespolonych. Szersze omówienie zasygnalizowanych tutaj zagadnień Czytelnik może znaleźć albo u RALSTONA [50], albo u TUROWICZA [62]. W większości rozważań będziemy przyjmować, że współczynniki a_i ($i = 0, 1, \dots, n$) wielomianu (3.23) są rzeczywiste.

3.2.1. Liczba pierwiastków rzeczywistych

Liczbę pierwiastków rzeczywistych w danym przedziale $(a; b)$ możemy określić orientacyjnie przez naszkicowanie wykresu funkcji $y = f(x)$. Pełną informację o liczbie pierwiastków rzeczywistych wielomianu (3.23) daje nam twierdzenie Sturma.

Dla danego wielomianu $f(x)$ budujemy tzw. *ciąg Sturma*

$$f_0(x), f_1(x), f_2(x), \dots, f_p(x) \quad (3.24)$$

w którym

$$f_0(x) \equiv f(x)$$

$$f_1(x) = f'(x)$$

$f_2(x)$ jest resztą z dzielenia $f_0(x)$ przez $f_1(x)$ wziętą ze znakiem przeciwnym

$f_3(x)$ jest resztą z dzielenia $f_1(x)$ przez $f_2(x)$ wziętą ze znakiem przeciwnym

.....

Zakładamy, że $f_{p+1}(x) \equiv 0$, a $f_p(x)$ jest ostatnią resztą różną od zera, czyli jest największym wspólnym dzielnikiem dla wielomianu $f_0(x)$ i jego pochodnej $f_1(x)$. Jeżeli ten największy wspólny dzielnik jest liczbą rzeczywistą różną od zera, oznacza to, że wielomian $f_0(x)$ nie ma zer wielokrotnych, gdy natomiast jest to wielomian stopnia k , oznacza to, że jego miejsce zerowe jest $(k + 1)$ -krotnym zerem wielomianu $f_0(x)$.

Oznaczamy przez $N(x_0)$ liczbę zmian znaku w ciągu (3.24), w punkcie $x = x_0$, w którym opuszczamy zera.

Tw. (Sturma). Jeżeli ciąg $(f_i(x))$, $i = 0, 1, \dots, p$, jest ciągiem Sturma na przedziale $(a; b)$ i $f_0(a)f_0(b) \neq 0$, to liczba różnych zer rzeczywistych wielomianu $f_0(x)$ leżących w tym przedziale jest równa $N(a) - N(b)$.

U w a g a. Jeżeli a jest skończone, to $N(a)$ definiujemy jako $N(a + \varepsilon)$, gdzie ε jest tak dobrane, aby żadna z funkcji ciągu $(f_i(x))$ nie zniknęła na przedziale $(a; a + \varepsilon)$; podobnie definiujemy $N(b)$, gdy b jest skończone. Jeżeli $a = -\infty$, to $N(a)$ definiujemy jako liczbę zmian znaku w ciągu $(\lim_{x \rightarrow -\infty} f_i(x))$; podobnie postępujemy, gdy $b = +\infty$.

Twierdzenie Sturma jest oczywiście prawdziwe dla przedziału $\langle a; b \rangle$, gdy $f_0(a)f_0(b) \neq 0$, a także gdy a lub b jest zerem pojedynczym wielomianu; należy tylko wtedy określić $N(x)$ jako liczbę zmian znaku w ciągu $(f_i(x))$, $i = 1, 2, \dots, p$.

Podamy tutaj jeszcze pewną praktyczną uwagę. Przy wyznaczaniu kolejnych wyrazów ciągu Sturma otrzymujemy często ułamkowe współczynniki. Ponieważ istotne dla nas są jedynie znaki tych wielomianów, a nie ich wartość, możemy więc każdy z nich pomnożyć przez odpowiednią dodatnią liczbę, tak aby otrzymać wielomian o współczynnikach całkowitych.

Przykład 1. Zastosować twierdzenie Sturma do określenia liczby pierwiastków rzeczywistych wielomianu $f(x) = x^3 + x^2 - x - 1$.

Obliczamy wyrazy ciągu Sturma

$$f_0(x) = x^3 + x^2 - x - 1$$

$$f_1(x) = 3x^2 + 2x - 1$$

$$f_2(x) = x + 1$$

Ponieważ $f_3(x) = 0$, więc równanie ma pierwiastek podwójny $x_0 = -1$. Dzieląc $f_0(x)$ przez $x + 1$ otrzymamy wielomian o pierwiastkach pojedynczych

$$\hat{f}_0(x) = x^2 - 1$$

$$\hat{f}_1(x) = 2x$$

$$\hat{f}_2(x) = 1$$

W tablicy 3/6 podano znaki $\hat{f}_i(x)$ dla kilku wybranych wartości x . Rozważany wielomian (po podzieleniu) ma 2 różne pierwiastki rzeczywiste dodatni i ujemny ($x_1 = -1$, $x_2 = 1$).

3/6

	$-\infty$	$+\infty$	0	+1	-1
$\hat{f}_0$	+	+	-	0	0
$\hat{f}_1$	-	+	0	+	-
$\hat{f}_2$	+	+	+	+	+
$N(x)$	2	0	1	0	1

Twierdzenie Sturma pozwala ściśle określić liczbę zer rzeczywistych w dowolnym zadanym przedziale, jednak wyznaczanie kolejnych wyrazów ciągu (3.24) wymaga niekiedy wykonania uciążliwych rachunków. Podamy teraz kilka metod znacznie prostszych w użyciu, lecz nie określających tak dokładnie liczby zer w danym przedziale.

Dla danego wielomianu $f(x)$ stopnia n tworzymy ciąg jego pochodnych

$$f(x), f'(x), f''(x), \dots, f^{(n)}(x) \quad (3.25)$$

Niech $M(x_0)$ oznacza liczbę zmian znaku w ciągu $(f^{(k)}(x))$, $k = 0, 1, \dots, n$, dla $x = x_0$.

Tw. (Fouriera). *Jeżeli $f(x)$ jest wielomianem stopnia n określonym w przedziale $(a; b)$ i $f(a)f(b) \neq 0$, to liczba zer wielomianu $f(x)$ w tym przedziale jest równa*

$$M(a) - M(b)$$

lub jest od tej liczby mniejsza o liczbę parzystą.

Przykład 2. Niech $f(x) = x^3 - 2x^2 - 5x + 5$. Tworzymy ciąg pochodnych

$$f(x) = x^3 - 2x^2 - 5x + 5$$

$$f'(x) = 3x^2 - 4x - 5$$

$$f''(x) = 6x - 4$$

$$f'''(x) = 6$$

Na podstawie tablicy 3/7 stwierdzamy, że wielomian ma jeden lub trzy pierwiastki ($M(-\infty) - M(+\infty) = 3$). Z warunku $M(-\infty) - M(0) = 1$ wynika, że istnieje pierwiastek ujemny, z warunku $M(0) - M(1) = 1$ wynika, że istnieje także pierwiastek w przedziale $(0; 1)$, a z $M(1) - M(3) = 1$, że trzeci pierwiastek należy do przedziału $(1; 3)$.

3/7

	$-\infty$	$+\infty$	0	1	3
f	—	+	+	—	+
f'	+	+	—	—	+
f''	—	+	—	+	+
f'''	+	+	+	+	+
$M(x)$	3	0	2	1	0

Znana jest także inna metoda określania liczby zer rzeczywistych wielomianu. Dla wielomianu

$$f(x) = a_0 x^n + a_1 x^{n-1} + \dots + a_{n-1} x + a_n, \quad a_0 \neq 0 \quad (3.26)$$

tworzymy ciąg wielomianów

$$f_0(x) = a_0$$

$$f_1(x) = a_0 x + a_1$$

$$f_2(x) = a_0 x^2 + a_1 x + a_2$$

.....

$$f_n(x) = f(x)$$

Niech $L(x_0)$ oznacza liczbę zmian znaku powyższego ciągu $(f_k(x))$, $k = 0, 1, \dots, n$, w punkcie $x = x_0$.

Tw. (Laguerre'a). *Jeżeli $f(x)$ jest wielomianem stopnia n określonym na przedziale $(a; b)$ i $f(a)f(b) \neq 0$, to liczba zer wielomianu $f(x)$ w tym przedziale jest równa*

$$L(a) - L(b)$$

lub jest od tej liczby mniejsza o liczbę parzystą.

Przypadkiem szczególnym tej metody jest tzw. *reguła Kartezjusza*, mówiąca, że *liczba dodatnich zer wielomianu (3.27) (z uwzględnieniem ich krotności) jest równa liczbie zmian znaków w ciągu współczynników $a_0, a_1, \dots, a_n$ lub jest od tej liczby mniejsza o liczbę parzystą.*

Chcąc znaleźć liczbę ujemnych zer wielomianu (3.26) konstruujemy wielomian $f(-x)$ i badamy liczbę jego dodatnich zer.

Przykład 3. Na podstawie reguły Kartezjusza dla wielomianu $f(x) = x^3 - 3x^2 - 5x + 5$ stwierdzamy, że jest jeden pierwiastek ujemny, natomiast liczba pierwiastków dodatnich, jest dwa albo zero.

Pełniejsze omówienie poruszanych tutaj metod wraz z dowodami twierdzeń można znaleźć w pracy [62].

Ćwiczenia

1. Podać twierdzenia określające liczbę pierwiastków rzeczywistych wielomianu.
2. Sformułować określenie k -krotnego pierwiastka funkcji $f(x)$.
3. Zbudować łańcuch Sturma dla wielomianu $f(x) = x^3 - 2x^2 - 5x + 5$.
4. Określić liczbę oraz lokalizację zer rzeczywistych wielomianu z ćwic. 3.
5. Zastosować twierdzenie Fouriera do wielomianu z ćwic. 3.
6. Korzystając ze wzorów (3.28) i (3.29) wyznaczyć przedział zawierający wszystkie dodatnie zera wielomianu $f(x) = x^7 + 5x^6 + 20x^5 - 12x^4 - 5x^3 - 2x^2 + 3x + 4$.

Odpowiedzi. 3. $f_0(x) = x^3 - 2x^2 - 5x + 5$, $f_1(x) = 3x^2 - 4x - 5$, $f_2(x) = 38x/9 - 35/9$, $f_3(x) = 8865/1444$. 4. Trzy pierwiastki położone odpowiednio w przedziałach: $(-2; -1)$, $(0; 1)$, $(3; 4)$. 6. $(1/2; 3)$.

3.2.2. Lokalizacja zer rzeczywistych

W wielu przypadkach bardzo istotna jest dla nas znajomość przedziału, w którym mieszczą się wszystkie rzeczywiste zera wielomianu (3.26). Problem można ograniczyć do wyznaczenia jedynie kresu górnego R dodatnich zer tego wielomianu, jeżeli bowiem wprowadzimy trzy pomocnicze równania

$$f_1(x) = x^n f\left(\frac{1}{x}\right) = 0$$

$$f_2(x) = f(-x) = 0 \tag{3.28}$$

$$f_3(x) = x^n f\left(\frac{-1}{x}\right) = 0$$

dla których kresy górne zer dodatnich są odpowiednio równe R_1, R_2, R_3 , to wszystkie dodatnie zera wielomianu (3.26) będą leżały w przedziale $(1/R_1; R)$, a ujemne w przedziale $(-R_2; -1/R_3)$.

Tw. (Lagrange'a). Niech $a_0 \neq 0$ i $a_k (k \geq 1)$ będzie pierwszym ujemnym współczynnikiem wielomianu (3.26). Wszystkie dodatnie zera tego wielomianu są mniejsze niż

$$R = 1 + \sqrt[k]{\frac{A}{|a_0|}} \quad (3.29)$$

gdzie A oznacza maksimum modułu ujemnych współczynników wielomianu.

Jeżeli wszystkie współczynniki wielomianu są nieujemne, to nie ma on zer dodatnich.

Należy pamiętać, że wyznaczenie przedziałów, w których mogą znajdować się zera rzeczywiste, nie oznacza bynajmniej, że wielomian ma takie zera. Ma je na pewno każdy wielomian nieparzystego stopnia.

Przykład 1. Wyznaczyć przedział zawierający wszystkie zera rzeczywiste wielomianu $f(x) = 2x^9 + x^7 - x^5 + 6x^3 - 12x^2 + 9$

Pierwszy ujemny współczynnik występuje przy x^5 , czyli jest to a_4 . Stąd mamy, że $k = 4$. Największy co do modułu ujemny współczynnik wynosi -12 , czyli $A = 12$. Mamy więc

$$R = 1 + \sqrt[4]{\frac{12}{2}} = 1 + \sqrt[4]{6} \approx 2,565$$

Zgodnie ze wzorami (3.28) obliczamy

$$\begin{aligned} f_1(x) &= x^9 f\left(\frac{1}{x}\right) = x^9 \left(\frac{2}{x^9} + \frac{1}{x^7} - \frac{1}{x^5} + \frac{6}{x^3} - \frac{12}{x^2} + 9 \right) = \\ &= 9x^9 - 12x^7 + 6x^6 - x^4 + x^2 + 2 \end{aligned}$$

$$f_2(x) = f(-x) = -2x^9 - x^7 + x^5 - 6x^3 - 12x^2 + 9$$

$$\begin{aligned} f_3(x) &= x^9 f\left(-\frac{1}{x}\right) = x^9 \left(\frac{2}{x^9} - \frac{1}{x^7} + \frac{1}{x^5} - \frac{6}{x^3} - \frac{12}{x^2} + 9 \right) = \\ &= 9x^9 - 12x^7 - 6x^6 + x^4 - x^2 + 2x^9 \end{aligned}$$

We wszystkich trzech przypadkach pierwszy ujemny współczynnik występuje przy x^7 , czyli jest to a_2 . Stąd mamy $k = 2$. Największy co do modułu ujemny współczynnik jest również we wszystkich trzech przypadkach równy -12 , czyli $A = 12$. Zatem

$$R_1 = 1 + \sqrt{\frac{12}{9}} \approx 2,155$$

$$R_2 = 1 + \sqrt{\frac{12}{2}} \approx 3,449$$

$$R_3 = 1 + \sqrt{\frac{12}{9}} \approx 2,155$$

Tak więc dodatnie zera należą do przedziału $(0,464; 2,565)$, a ujemne do przedziału $(-3,449; -0,464)$.

3.2.3. Metody przybliżonego obliczania zer rzeczywistych wielomianu

Istnieje wiele metod znajdowania pierwiastków rzeczywistych równania (3.23) (patrz np. [17, 50, 62]). Przytoczymy tutaj w skróconej formie *metodę iterowanego dzielenia*.

Dzieląc wielomian $f(x)$ przez czynnik liniowy $(x - x_j)$ otrzymamy

$$f(x) = (x - x_j)(b_0 x^{n-1} + b_1 x^{n-2} + \dots + b_{n-1}) + R_j \quad (3.30)$$

gdzie reszta R_j jest równa $f(x_j)$. Współczynniki b_k obliczamy ze wzoru rekurencyjnego

$$b_0 = a_0, \quad b_k = a_k + x_j b_{k-1} \quad \text{dla} \quad k = 1, 2, \dots, n \quad (3.31)$$

i stąd

$$R_j = b_n = a_n + x_j b_{n-1} \quad (3.32)$$

Powtórne podzielenie przez $(x - x_j)$ da nam

$$f(x) = (x - x_j)^2(c_0 x^{n-2} + c_1 x^{n-3} + \dots + c_{n-2}) + (x - x_j)R'_j + R_j \quad (3.33)$$

przy czym $R'_j = f'(x_j)$ oraz

$$c_0 = b_0 = a_0$$

$$c_k = b_k + x_j c_{k-1}, \quad k = 1, 2, \dots, n-1 \quad (3.34)$$

$$R'_j = c_{n-1} = b_{n-1} + x_j c_{n-2}$$

Do obliczania kolejnych wyrazów ciągu przybliżeń (x_k) , $k = 1, 2, \dots, j$, rzeczywistego pierwiastka równania (3.23) możemy zastosować:

metodę siecznych

$$x_{j+1} = x_j - \frac{R_j(x_j - x_{j+1})}{R_j - R_{j-1}} \quad (3.35)$$

(należy tutaj przyjąć początkowe przybliżenia x_1 i x_2)
lub metodę Newtona

$$x_{j+1} = x_j - \frac{R_j}{R'_j} \quad (3.36)$$

(należy tutaj przyjąć przybliżenie początkowe x_1). Oczywiście każdy ze wzorów (3.35) i (3.36) można stosować wielokrotnie, tworząc ciąg kolejnych przybliżeń dzielnika aż do uzyskania dwóch kolejnych wartości x_k i x_{k+1} możliwie bliskich.

Ćwiczenia

1. Korzystając ze wzorów (3.35) i (3.40) znaleźć rzeczywiste zero wielomianu $f(x) = x^3 - x - 1$, przyjmując jako przybliżenie początkowe odpowiednio: $x_1 = 1,2$ oraz $x_2 = 1,3$ w metodzie siecznych, a $x_1 = 1,3$ w metodzie Newtona.

3.2.4. Lokalizacja zer zespolonych

Podamy teraz kilka twierdzeń pozwalających oszacować moduł zespolonych zer wielomianu.

Tw. 1. (Cauchy'ego). *Niech*

$$f(z) = a_0 z^n + a_1 z^{n-1} + \dots + a_n, \quad a_0 \neq 0$$

oraz

$$F(z) = |a_0| z^n - \sum_{k=1}^n |a_k| z^{n-k} \quad (3.37)$$

Jeżeli α oznacza jedyne dodatnie zero rzeczywiste wielomianu $F(z)$, zaś $z_1, z_2, \dots, z_n$ są zerami wielomianu $f(z)$, to

$$|z_k| \leq \alpha \quad \text{dla} \quad k = 1, 2, \dots, n$$

U w a g a. Współczynniki a_k mogą być zespolone.

Tw. 2. *Niech $f(z) = \sum_{k=0}^n a_k z^{n-k}$, $a_0 = 1$, i niech β będzie dowolną liczbą*

dodatnią. Wszystkie zera wielomianu spełniają nierówność

$$|z_k| \leq \gamma \quad k = 1, 2, \dots, n \quad (3.38)$$

gdzie

$$\gamma = \max \left(1/\beta, \sum_{k=1}^n |a_k| \beta^{k-1} \right)$$

Przykład 1. Znaleźć oszacowanie modułów zer wielomianu

$$f(z) = z^4 + 3z^2 - 5z + 1$$

Szukane oszacowanie $\gamma = \max (1/\beta, 3\beta + 5\beta^2 + \beta^3)$. Trzeba tak dobrać β , aby γ było możliwie małe. Przyjmijmy $\beta = 1/2$. Mamy

$3 \cdot \frac{1}{2} + 5 \cdot \frac{1}{4} + \frac{1}{8} = 2,875$ oraz $1/\beta = 2$. Moduły zer badanego wielomianu $f(z)$ są mniejsze niż 2,9.

Tw. 3. *Niech $f(z) = \sum_{k=0}^n a_k z^{n-k}$, $a_0 \neq 0$. Wszystkie zera wielomianu $f(z)$*

spełniają nierówność

$$|z_k| \leq 1 + \max_{1 \leq k \leq n} \left| \frac{a_k}{a_0} \right|, \quad k = 1, 2, \dots, n \quad (3.39)$$

U w a g a. Współczynniki a_k mogą być zespolone.

Następne dwa twierdzenia dotyczą wielomianów o dodatnich rzeczywistych współczynnikach.

Tw. 4. *Moduły wszystkich zer wielomianu $f(z) = \sum_{k=0}^n a_k z^{n-k}$ spełniają nierówność*

$$\min\left(\frac{a_1}{a_0}, \frac{a_2}{a_1}, \dots, \frac{a_n}{a_{n-1}}\right) \leq |z_k| \leq \max\left(\frac{a_1}{a_0}, \frac{a_2}{a_1}, \dots, \frac{a_n}{a_{n-1}}\right) \quad (3.40)$$

Tw. 5. *Jeżeli współczynniki wielomianu $f(z) = \sum_{k=0}^n a_k z^{n-k}$ spełniają wa-*

runek $a_0 > a_1 > a_2 > \dots > a_n > 0$, to moduły wszystkich zer wielomianu $f(z)$ są mniejsze niż 1. Jeżeli $a_n > a_{n-1} > \dots > a_1 > a_0 > 0$, to moduły wszystkich zer wielomianu $f(z)$ są większe niż 1.

Często ważnym zagadnieniem jest dla nas lokalizacja zer wielomianu na płaszczyźnie zespolonej. Na przykład, badanie stabilności liniowych układów automatyki, równań różniczkowych zwyczajnych czy różniczkowo-całkowych wymaga rozstrzygnięcia, czy wszystkie zera danego wielomianu mają ujemną część rzeczywistą. Podamy teraz trzy kryteria: Michajłowa (graficzne), Routha i Hurwitza, określające warunki, przy których wielomian o współczynnikach rzeczywistych ma zera położone na lewo od osi urojonej.

Kryterium Michajłowa. *Warunkiem koniecznym i dostatecznym na to, aby zera wielomianu $f(z)$ miały ujemną część rzeczywistą, jest, by wektor o początku w punkcie $(0, 0)$ i końcu w punkcie $f(jt)$ zatoczył kąt $n\pi/2$, gdy t zmienia się od 0 do $+\infty$, przy czym krzywa zataczana przez koniec wektora nie może przechodzić przez początek układu.*

W celu sformułowania kryterium Routha rozważmy wielomian

$$f(z) = a_0 z^n + b_0 z^{n-1} + a_1 z^{n-2} + b_1 z^{n-3} + \dots \quad (3.41)$$

o współczynnikach rzeczywistych. Obliczamy wartości

$$c_k = a_{k+1} - \frac{a_0}{b_0} b_{k+1}, \quad k = 0, 1, \dots, r-1, \quad b_0 \neq 0$$

$$d_k = b_{k+1} - \frac{b_0}{c_0} c_{k+1}, \quad k = 0, 1, \dots, r-1, \quad c_0 \neq 0$$

$$e_k = c_{k+1} - \frac{c_0}{d_0} d_{k+1}, \quad k = 0, 1, \dots, r-2, \quad d_0 \neq 0$$

.....

i ustawiamy je w postaci tablicy

$$\begin{array}{l} a_0, a_1, a_2, \dots, a_r \\ b_0, b_1, b_2, \dots, b_r \\ c_0, c_1, c_2, \dots, c_{r-1} \\ d_0, d_1, d_2, \dots, d_{r-1} \\ e_0, e_1, e_2, \dots, e_{r-2} \\ \dots \end{array} \quad (3.42)$$

Kryterium Routha. Warunkiem koniecznym i dostatecznym na to, aby wielomian (3.41) miał zera o ujemnej części rzeczywistej, jest, by wszystkie elementy pierwszej kolumny tablicy (3.42) były różne od zera i miały ten sam znak.

Przykład 2. Zbadać, czy zera wielomianu $f(z) = z^4 + 5z^3 + 10z^2 + 10z + 4$ mają ujemne części rzeczywiste.

Tworzymy tablicę (3.42)

$$\begin{array}{r} 1 \quad 10 \quad 4 \\ 5 \quad 10 \\ 8 \quad 4 \\ 7,5 \\ 4 \end{array}$$

Ponieważ wszystkie elementy pierwszej kolumny są dodatnie, więc wszystkie zera wielomianu $f(z)$ mają ujemne części rzeczywiste.

Kryterium Hurwitza. Warunkiem koniecznym i dostatecznym na to, aby wielomian $f(z)$ stopnia n , $a_n > 0$, miał zera o ujemnych częściach rzeczywistych, jest, by wszystkie wyznaczniki:

$$\begin{aligned} D_1 &= a_{n-1}, \quad D_2 = \begin{vmatrix} a_{n-1} & a_n \\ a_{n-3} & a_{n-2} \end{vmatrix}, \\ D_3 &= \begin{vmatrix} a_{n-1} & a_n & 0 \\ a_{n-3} & a_{n-2} & a_{n-1} \\ a_{n-5} & a_{n-4} & a_{n-3} \end{vmatrix}, \dots \\ \dots, D_n &= \begin{vmatrix} a_{n-1} & a_n & 0 & \dots & 0 \\ a_{n-3} & a_{n-2} & a_{n-1} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ a_{1-n} & a_{2-n} & a_{3-n} & \dots & a_0 \end{vmatrix} \end{aligned} \quad (3.43)$$

gdzie $a_j = 0$ dla $j < 0$, były dodatnie.

Przykład 3. Stosując kryterium Hurwitza zbadać lokalizację zer wielomianu rozważanego w poprzednim przykładzie.

Mamy $n = 4$, $a_0 = 1$, $a_1 = 5$, $a_2 = 10$, $a_3 = 10$ i $a_4 = 4$. Zatem

$$D_1 = a_3 = 10$$

$$D_2 = \begin{vmatrix} a_5 & a_4 \\ a_1 & a_2 \end{vmatrix} = \begin{vmatrix} 10 & 4 \\ 5 & 10 \end{vmatrix} = 80$$

$$D_3 = \begin{vmatrix} a_3 & a_4 & 0 \\ a_1 & a_2 & a_3 \\ 0 & a_0 & a_1 \end{vmatrix} = \begin{vmatrix} 10 & 4 & 0 \\ 5 & 10 & 10 \\ 0 & 1 & 5 \end{vmatrix} = 300$$

$$D_4 = \begin{vmatrix} a_3 & a_4 & 0 & 0 \\ a_1 & a_2 & a_3 & a_4 \\ 0 & a_0 & a_1 & a_2 \\ 0 & 0 & 0 & a_0 \end{vmatrix} = \begin{vmatrix} 10 & 4 & 0 & 0 \\ 5 & 10 & 10 & 4 \\ 0 & 1 & 5 & 10 \\ 0 & 0 & 0 & 1 \end{vmatrix} = 300$$

Ponieważ wszystkie wyznaczniki są dodatnie, potwierdza się wynik osiągnięty w poprzednim przykładzie.

Jak wiadomo, badanie liczby zer wielomianu o ujemnej części rzeczywistej można sprowadzić do badania liczby jego zer w kole jednostkowym. Niech

$$f(z) = \sum_{k=0}^n a_k z^{n-k} = a_0 z^n + a_1 z^{n-1} + \dots + a_n \quad (3.44)$$

$$f^*(z) = z^n \bar{f}\left(\frac{1}{z}\right) = \sum_{k=0}^n \bar{a}_k z^k = a_0 + \bar{a}_1 z + \dots + \bar{a}_n z^n$$

kreska pozioma oznacza tu sprzężenie. Utwórzmy ciąg wielomianów

$$f_0(z), f_1(z), f_2(z), \dots, f_n(z)$$

przy czym

$$f_0(z) = f(z)$$

$$f_1(z) = \bar{a}_n f_0(z) - a_0 f_0^*(z)$$

$$f_2(z) = \bar{a}_{n-1}^{(1)} f_1(z) - a_0^{(1)} f_1^*(z)$$

i ogólnie

$$f_{j+1}(z) = \bar{a}_{n-1}^{(j)} f_j(z) - a_0^{(j)} f_j^*(z), \quad j = 0, 1, \dots, n-1$$

$$f_j(z) = \sum_{k=0}^{n-j} a_k^{(j)} z^{n-j-k}, \quad j = 0, 1, \dots, n \quad (3.45)$$

Wprowadźmy oznaczenia

$$\delta_{j+1} = |a_{n-j}^{(j)}|^2 - |a_0^{(j)}|^2$$

$$P_k = \delta_1 \delta_2 \dots \delta_k \quad (3.46)$$

Tw. 6. Jeśli dla wielomianu stopnia n wśród liczb $P_1, P_2, \dots, P_n$, określonych wzorami (3.46) jest s liczb ujemnych oraz $n - s$ liczb dodatnich, to liczba zer tego wielomianu leżących wewnątrz koła jednostkowego jest równa s .

U w a g a. Jeżeli wszystkie liczby $\delta_1, \delta_2, \dots, \delta_n$ są różne od zera, to wielomian $f(z)$ nie ma zer na okręgu $|z| = 1$. Wniosek odwrotny jest nieprawdziwy.

Przykład 15. Zbadać lokalizację zer wielomianu $f(z) = 4z^3 + z$. Mamy

$$f_0(z) = f(z) = 4z^3 + z$$

$$f_0^*(z) = 0 \cdot z^3 + 1 \cdot z^2 + 0 \cdot z + 4 = z^2 + 4$$

$$f_1(z) = 0 \cdot f_0(z) - 4f_0^*(z) = -4z^2 - 16$$

$$f_1^*(z) = -16z^2 + 0 \cdot z - 4 = -16z^2 - 4$$

$$f_2(z) = -16f_1(z) + 4f_1^*(z) = 64z^2 + 256 - 64z^2 - 16 = 240$$

$$\delta_1 = |a_3^{(0)}|^2 - |a_0^{(0)}|^2 = 0 - 16 = -16$$

$$\delta_2 = |a_2^{(1)}|^2 - |a_0^{(1)}|^2 = 256 - 16 = 240$$

$$\delta_3 = |a_1^{(2)}|^2 - |a_0^{(2)}|^2 = 240^2$$

Na podstawie drugiego ze wzorów (3.46) stwierdzamy, że $P_1 < 0$, $P_2 < 0$, $P_3 < 0$. Wszystkie P_k są ujemne, a więc wszystkie zera wielomianu leżą wewnątrz koła jednostkowego (jest tak istotnie, bowiem zerami są: $0, j/2, -j/2$).

Ćwiczenia

1. Oszacować moduł wszystkich zer wielomianu $f(z) = z^4 + z^3 + z - 2$.
2. Korzystając z kryterium Routha zbadać, czy wszystkie zera wielomianu $f(z) = z^4 + 7z^3 + 18z^2 + 20z + 8$ mają ujemną część rzeczywistą.
3. Zbadać istnienie zer zespolonych wielomianu $f(z) = 4z^4 - 2z^3 + 2z^2 - z$ wewnątrz koła jednostkowego $|z| = 1$.

3.2.5. Metody przybliżonego obliczania zer zespolonych wielomianu

Aby znaleźć zespolony pierwiastek równania (3.23), można oczywiście użyć wzorów (3.35) lub (3.36), z tym że działania należy wówczas wykonywać na liczbach zespolonych. Wiadomo [50], że zarówno metoda Newtona, jak i metoda siecznych są zbieżne dla pojedynczych pierwiastków zespolonych, jeżeli tylko początkowe przybliżenie jest wystarczająco dobre. Rozpatrzmy metodę Newtona dla z_k zespolonego. Mamy

$$z_i = x_i + jy_i$$

$$b_k = \gamma_k + j\delta_k \tag{3.47}$$

$$c_k = \varepsilon_k + j\eta_k$$

a więc wzory (3.32) i (3.34) przyjmują teraz postać

$$\gamma_0 = a_0, \quad \delta_0 = 0$$

$$\gamma_k = a_k + x_i \gamma_{k-1} - y_i \delta_{k-1}, \quad k = 1, 2, \dots, n$$

$$\delta_k = x_i \delta_{k-1} + y_i \gamma_{k-1}, \quad k = 1, 2, \dots, n$$

$$\varepsilon_0 = \gamma_0 = a_0, \quad \eta_0 = 0$$

$$\varepsilon_k = \gamma_k + x_i \varepsilon_{k-1} - y_i \eta_{k-1}, \quad k = 1, 2, \dots, n-1$$

$$\eta_{n-2} = 0$$

$$\eta_k = \delta_k + x_i \eta_{k-1} + y_i \varepsilon_{k-1}, \quad k = 1, 2, \dots, n-1$$

$$R_i = b_n = \gamma_n + j\delta_n$$

$$R'_i = c_{n-1} = \varepsilon_{n-1} + j\eta_{n-1} \tag{3.48}$$

a wzór (3.36) ma postać

$$x_{i+1} = x_i - \frac{\gamma_n \varepsilon_{n-1} + \delta_n \eta_{n-1}}{\varepsilon_{n-1}^2 + \eta_{n-1}^2} \quad (3.49)$$

$$y_{i+1} = y_i - \frac{\delta_n \varepsilon_{n-1} - \gamma_n \eta_{n-1}}{\varepsilon_{n-1}^2 - \eta_{n-1}^2}$$

Zamiast metody iteracyjnego dzielenia stosowanej w przypadku pierwiastków rzeczywistych, do znajdowania przybliżonych wartości pierwiastków zespolonych wielomianu można stosować *metodę obliczania kolejnych poprawek dla współczynników trójmianu kwadratowego — metodę Bairstowa* (wiąże się ona z zasadniczym twierdzeniem algebry o rozkładzie wielomianu rzeczywistego na czynniki rzeczywiste). Podamy najpierw wzory związane z dzieleniem wielomianu $f(z)$ przez czynnik kwadratowy $z^2 + p_i z + q_i$. Mamy

$$f(z) = (z^2 + p_i z + q_i)(b_0 z^{n-2} + b_1 z^{n-3} + \dots + b_{n-2}) + R_i z + S_i \quad (3.50)$$

gdzie

$$b_0 = a_0, \quad b_{-1} = 0$$

$$b_k = a_k - p_i b_{k-1} - q_i b_{k-2}, \quad k = 1, 2, \dots, n$$

$$R_i = a_{n-1} - p_i b_{n-2} - q_i b_{n-3} \equiv b_{n-1}$$

$$S_i = a_n - q_i b_{n-2} \equiv b_n + p_i b_{n-1}$$

Dzielać następnie (3.50) ponownie przez ten czynnik kwadratowy, otrzymujemy

$$f(z) = (z^2 + p_i z + q_i)^2 (c_0 z^{n-4} + c_1 z^{n-5} + \dots + c_{n-4}) + (R'_i z + S'_i)(z^2 + p_i z + q_i) + R_i z + S_i \quad (3.51)$$

gdzie

$$c_0 = b_0, \quad c_{-1} = 0$$

$$c_k = b_k - p_i c_{k-1} - q_i c_{k-2}, \quad k = 1, 2, \dots, n-2$$

$$R'_i = b_{n-3} - p_i c_{n-4} - q_i c_{n-5} \equiv c_{n-3}$$

$$S'_i = b_{n-2} - q_i c_{n-4} \equiv c_{n-2} + p_i c_{n-3}$$

Kolejne przybliżenia wyznacza się ze wzorów (patrz [44], [71])

$$p_{k+1} = p_k - \frac{b_{n-1} c_{n-2} - b_{n-3} c_{n-3}}{S} \quad (3.52)$$

$$q_{k+1} = q_i - \frac{b_{n-1} r - c_{n-2} b_n}{S}$$

gdzie

$$r = -p_k c_{n-2} + q_k c_{n-3}$$

$$S = c_{n-2}^2 - r c_{n-3}$$

Algorytm Q-D. Omówione powyżej metody wymagają określenia wartości początkowych dostatecznie bliskich wartości szukanego pierwiastka lub współczynników czynnika kwadratowego. Algorytm Q-D umożliwia znalezienie pierwiastków wielomianu bez potrzeby określania wartości początkowych.

Dla wielomianu (3.23) tworzymy tablicę czynników q i e , której pierwsze dwa wiersze (patrz tabl. 3/8) oblicza się ze wzorów

$$\begin{aligned} q^{(1)} &= -a_2/a_1, \quad \text{pozostałe } q^{(i)} = 0 \quad \text{dla } i = 2, \dots, n \\ e^{(i)} &= a_{i+2}/a_{i+1}, \quad i = 1, 2, \dots, n-1 \\ e^{(0)} &= e^{(n)} = 0 \end{aligned} \tag{3.53}$$

3/8

$e^{(0)}$	$q^{(1)}$	$e^{(1)}$	$q^{(2)}$	$e^{(2)}$	$q^{(3)}$	...	$e^{(n-1)}$	$q^{(n)}$	$e^{(n)}$
	$\frac{-a_1}{a_0}$		0	0		...		0	
0		$\frac{a_2}{a_1}$		$\frac{a_3}{a_2}$		...	$\frac{a_n}{a_{n-1}}$		0

W następnych wierszach tablicy wartości q oraz e oblicza się ze wzorów

$$\begin{aligned} q^{(i)} &= e^{(i)} - e^{(i-1)} + q^{(i)} \\ e^{(i)} &= (q^{(i+1)}/q^{(i)}) e^i \end{aligned} \tag{3.54}$$

w których wartości e i q po prawych stronach równości (3.54) są brane z poprzedzających wierszy tablicy. Obliczenia kontynuuje się tak długo, aż wartości e staną się bardzo bliskie zeru. Wówczas wartości kolejnych q^i są wartościami pierwiastków.

Jeżeli wielomian ma parę zespolonych, sprzężonych pierwiastków, to jedna z wartości e nie będzie dążyła do zera. W takim przypadku suma wartości q leżących po obu stronach tego e daje wartości $-r$, a iloczyn dwóch czynników q , lewego leżącego powyżej oraz prawego leżącego poniżej, daje wartości s czynnika kwadratowego $x^2 + rx + s$. W przypadku pierwiastka podwójnego postępuje się identycznie. Gdy chociaż jeden ze współczynników a_i jest równy zeru, wówczas nie można obliczyć q i e z pierwszych wierszy. W takim przypadku należy wprowadzić nową zmienną, np. $y = x - 1$ (dobór współczynnika jest dowolny).

Przykład. Stosując metodę Q-D znaleźć pierwiastki równania

$$x^4 - 6x^3 + 12x^2 - 19x + 12 = 0$$

W tablicy 3/9 zestawiono wyniki kolejnych obliczeń. Jak widać $q^{(1)}$ jest zbieżne do 4, $q^{(4)}$ do 1. Ponieważ $e^{(2)}$ nie jest zbieżne do zera, więc $q^{(2)}$ i $q^{(3)}$ reprezentują czynnik kwadratowy

$$r = -(q^{(2)} + q^{(3)}) = -(1,456 - 0,466) = -0,990$$

$$s = (-6,426)(-0,466) = 2,995$$

Czynnik kwadratowy ma postać

$$x^2 - 0,99x + 2,995$$

Zgodność, jak widać, jest zupełnie dobra, gdyż badany wielomian można rozłożyć na czynniki

$$x^4 - 6x^3 + 12x^2 - 19x + 12 = (x - 1)(x - 4)(x^2 - x + 3)$$

3/9

$e^{(0)}$	$q^{(1)}$	$e^{(1)}$	$q^{(2)}$	$e^{(2)}$	$q^{(3)}$	$e^{(3)}$	$q^{(4)}$	$e^{(4)}$
0	6	-2	0	-1.583	0	-0.632	0	0
0	4	-0.208	0.417	-3.610	0.951	-0.420	0.632	0
0	3.792	0.164	-2.985	5.008	4.141	-0.107	1.052	0
0	3.956	0.077	1.859	-2.624	-0.974	0.127	1.159	0
0	4.033	-0.016	-0.842	5.538	1.777	0.074	1.032	0
0	4.017	-0.019	4.712	-4.333	-3.687	-0.019	0.958	0
0	3.998	-0.002	0.398	-6.826	0.627	-0.030	0.977	0
0	4.000	0.003	-6.426	7.885	7.423	-0.004	1.007	0
0	4.003		1.456		-0.466		1.010	

3.3. UWAGI O EFEKTYWNOŚCI METOD PRZYBLIŻONEGO OBLICZANIA PIERWIASTKÓW

Dla użytkownika istotny jest problem jakości poszczególnych metod. Każdej metodzie można przyporządkować liczbę p (największą możliwą dla danej metody) — zwaną *wykładnikiem zbieżności* danej metody. Mówimy, że *metoda jest rzędu p* , jeżeli istnieje stała K taka, że dla dwu kolejnych przybliżeń x_k i x_{k+1} zachodzi nierówność

$$|x_{k+1} - x| \leq K|x_k - x|^p$$

Można porównywać rzędy metod określające szybkość ich zbieżności, można także porównywać ich *efektywność*, która określa zakres obliczeń koniecznych do otrzymania pierwiastka z żadaną dokładnością.

Miarą jakości metody w pobliżu dokładnego pierwiastka może być *wskaźnik efektywności*

$$E = p^{\frac{1}{k}}$$

gdzie p — rząd metody, a k — koszt jednej iteracji. Koszt jednej iteracji zależy głównie od tego, ile razy w każdym kroku trzeba obliczać wartość funkcji $f(x)$ lub jej pochodnych, natomiast mało zależy od operacji arytmetycznych wykonywanych na wartościach tych funkcji. Będziemy dalej przyjmować, że koszt obliczenia wartości funkcji $f(x)$ jest równy 1.

Metoda połowienia, podobnie jak metoda regula falsi są rzędu 1 i ich wskaźnik efektywności jest także równy 1. Metoda siecznych jest szybciej od nich zbieżna i jest rzędu $\frac{1}{2}(1 + \sqrt{5}) \approx 1,62$; tyle samo wynosi wskaźnik efektywności E dla tej metody. Dla metody Newtona, która jest rzędu 2, mamy $E = 2^{1/(1+k')}$, gdzie k' jest kosztem obliczania pochodnej $f'(x)$. Ralston [50] sugeruje, że gdy $k' < 0,44$, lepiej stosować metodę Newtona zamiast metody siecznych.

W przypadku pierwiastków wielokrotnych metoda Newtona jest rzędu 1, natomiast gdy znana jest krotność pierwiastka r i można korzystać ze wzoru (3.21) rząd metody wzrasta do 2. Podobnie podstawienie (3.22) poprawia rząd, lecz efektywność metody staje się nieco niższa.

W przypadku pierwiastków zespolonych metoda Newtona jest zbieżna, jeżeli tylko przybliżenie początkowe jest dostatecznie bliskie pierwiastka, a jej rząd zbieżności jest równy 2. Podobnie metoda siecznych jest dla dostatecznie bliskich przybliżeń początkowych zbieżna do pojedynczych pierwiastków zespolonych, a jej rząd jest równy $\frac{1}{2}(1 + \sqrt{5})$. Metoda Bairstowa jest metodą rzędu 2.

Ciekawe porównanie szybkości zbieżności dla metody siecznych i metody stycznych przedstawiono w tablicy 3/10. Są to wyniki rozwiązywania równania $x^3 + x + 10 = 0$ w przedziale $\langle -3; 1 \rangle$ z dokładnością $\varepsilon = 10^{-5}$. Wyraźnie można zaobserwować, jak w miarę zbliżania się do wartości pierwiastka zbieżność ciągu kolejnych przybliżeń w metodzie siecznych jest coraz powolniejsza (wynika to, między innymi, ze wzorów (3.11) i (3.19) na oszacowanie błędu w poszczególnych metodach).

3/10

Nr iteracji	Metoda siecznych	Metoda stycznych	Nr iteracji	Metoda siecznych	Metoda stycznych
1	−1.5714286	−2.285714286	8	−1.9996824	
2	−1.8361045	−2.032173457	9	−1.9998888	
3	−1.9406551	−2.000468841	10	−1.9999611	
4	−1.9789723	−2.000000101	11	−1.9999864	
5	−1.9926082		12	−1.9999952	
6	−1.9974089		13	−1.9999983	
7	−1.9990926				

Przy poszukiwaniu zer wielomianu po znalezieniu przybliżenia pierwiastka badanego wielomianu wygodnie jest podzielić ten wielomian przez czynnik liniowy zawierający ten pierwiastek i następne pierwiastki wyznaczać dla tak otrzymanego ilorazu. Zastąpienie wyjściowego wielomianu wielomianem niższego stopnia nazywa się *obniżeniem stopnia* lub *deflacją*. Korzyści ze stosowania deflacji polegają przede wszystkim na stopniowym zmniejszaniu pracochłonności przy wyznaczaniu następnych pierwiastków i związaną z tym oszczędnością pamięci maszyny, a także na uniknięciu powtórnego wyznaczenia tych samych pierwiastków. Często bowiem zdarza się, że nawet przy starcie z punktu położonego w pobliżu innego pierwiastka procedura iteracyjna doprowadza nas do uprzednio już znalezionego pierwiastka.

Wilkinson [68] wykazał, że błędy powstające przy deflacji są do pominięcia gdy:

- pierwiastki obliczamy kolejno ze wzrostem ich wartości bezwzględnych,
- każdy pierwiastek zostaje obliczony z maksymalną graniczną dokładnością.

W każdej z omawianych powyżej metod możemy osiągnąć maksymalną graniczną dokładność przybliżenia pierwiastka [38].

Dzieląc wielomian $f(z) = \sum_{i=0}^n a_i z^{n-i}$ przez czynnik liniowy $z - x$, mamy

$$\sum_{i=0}^n a_i z^{n-i} = \sum_{i=0}^{n-1} b_i z^{n-i-1} (z - x) = Q(z) (z - x)$$

Stąd otrzymujemy układ równań dla wyznaczania współczynników b_i wielomianu $Q(z)$

$$b_n = b_{-1} = 0$$

$$a_i = b_i - b_{i-1} x, \quad i = 0, 1, 2, \dots, n$$

Układ ten rozwiązujemy zazwyczaj albo zwykłym algorytmem Hornera (3.31), albo odwrotnym algorytmem Hornera (dla $x \neq 0$)

$$\begin{aligned} b_n &= 0 \\ b_{i-1} &= (b_i - a_i)/x, \quad i = n, n-1, \dots, 1 \end{aligned} \tag{3.55}$$

Można także wykorzystać oba te sposoby, wyznaczając dla dowolnego $k = 0, 1, 2, \dots, n$ współczynniki $b_0, b_1, \dots, b_k$ ze wzoru (3.31), a $b_{k+1}, b_{k+2}, \dots, b_n$ z (3.55) — jest to tzw. *sklejany algorytm Hornera*. W pracy [38] można znaleźć dane na temat „sklejania” algorytmu Hornera i oszacowanie zaburzeń zer wielomianu $Q(z)$ w wyniku deflacji, a także w wyniku p etapów deflacji. Wykazano tam, że zera wielomianu $R(z)$, otrzymanego po p etapach deflacji, są zerami zaburzonego wielomianu $f(z)$ z zaburzeniami rzędu $2^{-t}pn$, gdzie t — liczba cyfr mantysy binarnej, a n — stopień wielomianu (kolejność

wyznaczania przybliżeń pierwiastków nie ma specjalnego wpływu przy sklejaney deflacji).

W podobny sposób można znaleźć sklepany algorytm deflacyjny dla deflacji rzeczywistego wielomianu $f(z)$ z rzeczywistym czynnikiem kwadratowym (jest to ważne ze względu na stosowanie metody Bairstowa [47]). Problem deflacji — oczywisty w przypadku poszukiwania zer wielomianów — w przypadku równań przestępczych można sprowadzić np. dla metody Newtona do tzw. *pozornej deflacji*. Polega ona na formalnym podzieleniu lewej strony równania przez iloczyn czynników liniowych odpowiadających już znalezionym pierwiastkom i ponownego zastosowania metody Newtona do tak otrzymanej funkcji [68, s. 112]. W przypadku metody Newtona wzór na deflację pozorną ma postać

$$x_{i+1} = x_i - \left[f'(x_i)/f(x_i) - \sum_{n=1}^r \frac{1}{x - x_i} \right]^{-1}$$

r — liczba już znalezionych zer funkcji $f(x)$.

3.4. UKŁADY RÓWNAŃ NIELINIOWYCH

Istota rozwiązywania układu równań sprowadza się do znalezienia takiego rozwiązania $\alpha \in \mathbf{R}^n$, w którym odwzorowanie $F: \mathbf{R}^n \rightarrow \mathbf{R}^n$ przyjmuje wartość zero, tzn. $F(\alpha) = 0$ ¹⁾. W szczególnym przypadku gdy odwzorowanie F jest afiniczne, np. $F(x) = Ax - b$, uzyskujemy układ równań liniowych $Ax = b$, w którym A jest macierzą utworzoną ze współczynników układu, x — wektorem szukanych wartości, a b — wektorem wyrazów wolnych. Metody rozwiązywania układów równań liniowych są omówione w rozdziale 5: podano tam przegląd tzw. metod dokładnych i metod iteracyjnych poszukiwania rozwiązania α , przyjmując założenie o istnieniu i jednoznaczności rozwiązywania układu $Ax = b$, które sprowadza się do warunku $\det A \neq 0$. W przypadku gdy odwzorowanie F jest nieliniowe warunki istnienia rozwiązania układu równań $F(x) = 0$ są znacznie trudniejsze do sprawdzenia, a nawet nie można sformułować jednolitego kryterium istnienia rozwiązania bez założenia szczególnych własności odwzorowania F , takich jak różniczo-

¹⁾ Zwracamy uwagę Czytelnika, że metody poszukiwania rozwiązań układów równań są znacznie trudniejsze niż dla jednego równania. W celu krótkiego przedstawienia własności tych metod, wprowadza się zazwyczaj zapis wektorowy. Ponadto pojawia się tu potrzeba odróżnienia indeksów współrzędnych wektora od numerów nadawanych wektorom w kolejnych iteracjach. Przyjęto umowę, że numery iteracji są oznaczone indeksami u góry, a współrzędne wektora — indeksami u dołu. Zatem x_1 oznacza pierwszą współrzędną wektora x , a $x^{(1)}$ wektor obliczony w pierwszej iteracji.

walność, monotoniczność itp. W naszych rozważaniach będziemy zakładać a priori istnienie rozwiązania α układu równań $F(x) = 0$ i ograniczymy się jedynie do omówienia wybranych metod poszukiwania punktu α . Czytelnika zainteresowanego dalszymi szczegółami odsyłamy do prac [45], [69] i [23].

3.4.1. Ogólne metody iteracyjne

W obliczeniach numerycznych konstruujemy na ogół ciąg punktów $x^{(0)}, x^{(1)}, \dots$, zbieżny do rozwiązania α układu równań $F(x) = 0$. Jeżeli można wskazać takie odwzorowanie G , że $x^{(i)} = G(x^{(i-1)}, x^{(i-2)}, \dots, x^{(i-p)})$, to taką metodę iteracyjną nazywamy metodą *stacjonarną p-punktową*. W szczególnym przypadku, gdy $x^{(i)} = G(x^{(i-1)})$, metodę nazywamy *stacjonarną*, opuszczając dodatkowe określenie „1-punktową”. W pewnych przypadkach operator G może ulec modyfikacjom podczas iteracji — metodę nazywamy wówczas *niestacjonarną (p-punktową)* lub *metodą zmiennego operatora*.

Def. (Ostrowskiego). Niech $G: D \subset R^n \times R^n \times \dots \times R^n \rightarrow R^n$. Punkt α nazywamy *punktem przyciągania* metody iteracyjnej, jeżeli istnieje takie otoczenie U_α tego punktu, że wybierając punkty $x^{(-p+1)}, x^{(-p+2)}, \dots, x^{(0)}$ z tego otoczenia uzyskamy ciąg punktów $x^{(1)}, x^{(2)}, \dots \in D$, zbieżny do α .

Największe z tych otoczeń nazywamy *obszarem przyciągania* punktu α .

Lemat. Niech $G: D \subset R^n \rightarrow R^n$ i niech istnieją kula $K(\alpha, r) \subset D$ i stała $c < 1$ takie, że

$$\|Gx - \alpha\| < c \|x - \alpha\|, \quad \bigwedge x \in K$$

Wówczas dla dowolnego $x^{(0)} \in K$, ciąg $x^{(1)}, x^{(2)}, \dots$ generowany wzorem $x^{(i+1)} = G(x^{(i)})$ jest zawarty w K i jest zbieżny do α . Zbieżność ta jest co najmniej liniowa, tzn. $\|x^{(i+1)} - \alpha\| \leq c \|x^{(i)} - \alpha\|$, $i = 0, 1, \dots$.

Def. Odwzorowanie $G: R^n \rightarrow R^n$ nazywamy *różniczkowalnym w sensie Frécheta* w punkcie x , jeżeli istnieje taka macierz $A: R^n \rightarrow R^n$, że

$$\lim_{h \rightarrow 0} \frac{\|G(x+h) - G(x) - Ah\|}{\|h\|} = 0$$

przy dowolnym sposobie wyboru wektorów $h \rightarrow 0$. Macierz A nazywamy *pochodną Frécheta* odwzorowania G (lub *F-pochodną*) w punkcie x i oznaczamy ją $G'(x)$.

Tw. 1. (Ostrowskiego). Jeżeli *F-pochodna* odwzorowania $G: R^n \rightarrow R^n$ w punkcie α ma promień spektralny¹⁾ $\rho(G'(\alpha)) = \beta < 1$ oraz $G(\alpha) = \alpha$, to α jest *punktem przyciągania* metody iteracyjnej $x^{(i+1)} = G(x^{(i)})$.

DOWÓD. Można wykazać [45, s. 47], że dla dowolnego $\varepsilon > 0$ istnieje taka norma w R^n , że $\|G'(\alpha)\| \leq \beta + \varepsilon$. Z określenia *F-pochodnej* wynika, że istnieje taka kula $K(\alpha, r)$, że $\|G(x) - G(\alpha) - G'(\alpha)(x - \alpha)\| \leq \varepsilon \cdot \|x - \alpha\|$ dla $x \in K$. Zatem $\|G(x) - \alpha\| \leq \|G'(\alpha)(x - \alpha)\| +$

¹⁾ Promieniem spektralnym macierzy kwadratowej nazywamy największy spośród modułów wartości własnych tej macierzy.

$+\varepsilon\|\mathbf{x}-\boldsymbol{\alpha}\| \leq (\beta+2\varepsilon)\|\mathbf{x}-\boldsymbol{\alpha}\|$. Obierając dostatecznie małe ε , mamy $\beta+2\varepsilon < 1$ i z podanego powyżej lematu wynika, że przy dowolnym $\mathbf{x}^{(0)} \in K$ ciąg generowany wzorem $\mathbf{x}^{(i+1)} = G(\mathbf{x}^{(i)})$ jest zbieżny do $\boldsymbol{\alpha}$, cnd.

Twierdzenie 1 pozwala uzasadnić lokalną zbieżność wielu metod iteracyjnych stosowanych do wyznaczenia rozwiązania $\boldsymbol{\alpha}$ układu równań $F(\mathbf{x}) = 0$. Niech na przykład

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - A F(\mathbf{x}^{(i)}) \quad (3.56)$$

gdzie: $A: \mathbf{R}^n \rightarrow \mathbf{R}^n$ jest odpowiednio przyjętą macierzą¹⁾. Wzór (3.56) określa stacjonarną metodę iteracyjną z odwzorowaniem $G(\mathbf{x}) = \mathbf{x} - A F(\mathbf{x})$.

Założmy, że F jest funkcją różniczkowalną w sensie Frécheta w punkcie $\boldsymbol{\alpha}$. Wówczas $G'(\boldsymbol{\alpha}) = I - A F'(\boldsymbol{\alpha})$. Jeżeli $\varrho(I - A F'(\boldsymbol{\alpha})) < 1$, to istnieje takie otoczenie K punktu $\boldsymbol{\alpha}$, że jeśli $\mathbf{x}^{(0)} \in K$, to ciąg $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ generowany wzorem (3.56) jest zawarty w K oraz $\mathbf{x}^{(i)} \rightarrow \boldsymbol{\alpha}$. Warunek powyższy jest dostateczny (a także konieczny [45]) dla lokalnej zbieżności metody iteracyjnej (3.56) do punktu $\boldsymbol{\alpha}$. Istotna trudność polega jednak na tym, że na ogół nie znamy pochodnej $F'(\boldsymbol{\alpha})$, a więc nie mamy gwarancji, że warunek $\varrho(I - A F'(\boldsymbol{\alpha})) < 1$ jest spełniony. Trudność tę rozwiązuje omówiona poniżej metoda Newtona.

3.4.2. Metoda Newtona

Tw. 1. Niech funkcja $F(\mathbf{x})$ będzie różniczkowalna w sensie Frécheta w pewnym otoczeniu $K(\boldsymbol{\alpha}, r)$ punktu $\boldsymbol{\alpha}$, w którym $F(\boldsymbol{\alpha}) = 0$. Założmy, że pochodna $F'(\mathbf{x})$ jest ciągła w punkcie $\boldsymbol{\alpha}$, a pochodna $F'(\boldsymbol{\alpha})$ jest nieosobliwa. Wówczas punkt $\boldsymbol{\alpha}$ jest punktem przyciągania metody iteracyjnej

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - [F'(\mathbf{x}^{(i)})]^{-1} F(\mathbf{x}^{(i)}) \quad (3.57)$$

zwanej metodą Newtona.

SZKIC DOWODU. Założenie ciągłości pochodnej w punkcie $\boldsymbol{\alpha}$ i nieosobliwość pochodnej $F'(\boldsymbol{\alpha})$ gwarantują, że w pewnym otoczeniu tego punktu pochodna $F'(\boldsymbol{\alpha})$ jest nieosobliwa. Korzystając z definicji pochodnej Frécheta można uzasadnić, że pochodna odwzorowania $G(\mathbf{x}) = \mathbf{x} - [F'(\mathbf{x})]^{-1} F(\mathbf{x})$ w punkcie $\boldsymbol{\alpha}$ jest równa zero. Zatem z twierdzenia Ostrowskiego (patrz p. 3.4.1) wynika lokalna zbieżność metody Newtona.

Zauważmy, że metoda (3.57) jest prostym uogólnieniem na przypadek funkcji n zmiennych metody opisanej w p. 3.1.3. Jak poprzednio, przyjęto tu silne założenia dotyczące funkcji F (ciągła różniczkowalność, nieosobliwość pochodnej w punkcie $\boldsymbol{\alpha}$), a uzyskano jedynie gwarancję lokalnej zbieżności metody. Dodatkową niedogodnością jest konieczność rozwiązywania w każdej iteracji układu równań $F'(\mathbf{x}^{(i)}) d^{(i)} = F(\mathbf{x}^{(i)})$ w celu obliczenia $\mathbf{x}^{(i+1)} =$

¹⁾ W szczególnym przypadku, gdy $A = I$ (I — macierz jednostkowa), metodę nazywamy *metodą iteracji prostej*. Łatwo sprawdzić, że aby punkt $\boldsymbol{\alpha}$ był punktem przyciągania metody iteracji prostej, potrzeba i wystarcza, by w punkcie tym Jacobian $F'(\boldsymbol{\alpha})$ miał wartości własne o modułach w przedziale $(0; 2)$.

$= \mathbf{x}^{(i)} - d^{(i)}$, gdzie $d^{(i)} = [F'(\mathbf{x}^{(i)})]^{-1} F(\mathbf{x}^{(i)})$. Przy pewnych dodatkowych założeniach dotyczących funkcji F zbieżność metody Newtona jest jednak bardzo szybka.

Tw. 2. *Jeżeli odwzorowanie F spełnia warunki twierdzenia 1, a ponadto ciągłość pochodnej $F'(\mathbf{x})$ w punkcie α jest typu Höldera, tzn.*

$$\|F'(\mathbf{x}) - F'(\alpha)\| \leq H \|\mathbf{x} - \alpha\|^p \quad \text{dla} \quad \mathbf{x} \in K(\alpha, r), \quad p \in (0; 1]$$

to

$$\|\mathbf{x}^{(i+1)} - \alpha\| \leq C \|\mathbf{x}^{(i)} - \alpha\|^{1+p} \quad (3.58)$$

gdzie $C = 4H\|F'(\alpha)^{-1}\|$.

Uzasadnienie powyższego twierdzenia można znaleźć np. w [45].

Przy spełnieniu założeń tego twierdzenia metoda Newtona jest lokalnie zbieżna, tzn. ciąg $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$, uzyskany ze wzoru (3.57) jest zbieżny do α , jeśli tylko początkowy punkt $\mathbf{x}^{(0)}$ jest dostatecznie bliski punktowi α . Zatem z nierówności (3.58) wynika, że

$$\frac{\|\mathbf{x}^{(i+1)} - \alpha\|}{\|\mathbf{x}^{(i)} - \alpha\|} \leq C \|\mathbf{x}^{(i)} - \alpha\|^p \rightarrow 0$$

tzn. *zbieżność jest nadliniowa* (szybsza niż w przypadku dowolnego malejącego postępu geometrycznego).

W szczególnym przypadku, gdy $p = 1$, gdy pochodna $F'(\mathbf{x})$ jest ciągła w punkcie α w sensie Lipschitza, uzyskujemy *zbieżność kwadratową*, tzn.

$$\|\mathbf{x}^{(i+1)} - \alpha\| \leq C \|\mathbf{x}^{(i)} - \alpha\|^2$$

Wobec szybkiej lokalnej zbieżności i prostej formule (3.57) obliczania ciągu przybliżeń rozwiązania $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$, metoda Newtona znalazła szerokie zastosowanie w obliczeniach prowadzących na maszynach cyfrowych. Opracowano wiele modyfikacji tej metody, pozwalającej na osłabienie niektórych założeń gwarantujących zbieżność lokalną, a także opracowano sposoby zwiększania promienia zbieżności metody (czyli powiększania otoczenia punktu α , z którego można wybierać początkowe przybliżenia $\mathbf{x}^{(0)}$). W przypadku gdy liczba zmiennych n jest duża, istotną trudnością jest obliczanie pochodnych $F'(\mathbf{x})$ odwzorowania F . Wówczas stosuje się dyskretne warianty metody Newtona, w których zastępuje się pochodną $F'(\mathbf{x})$ jej przybliżeniem różnicowym [45]. Podobnie jak w przypadku funkcji jednej zmiennej, można jednak od razu zrezygnować z korzystania z pochodnych i zastosować metodę siecznych.

3.4.3. Metoda siecznych

Idea tej metody polega na przybliżeniu odwzorowania $F: \mathbf{R}^n \rightarrow \mathbf{R}^n$ odwzorowaniem afinicznym w ten sposób, by

$$F(\mathbf{y}^{(j)}) \approx A\mathbf{y}^{(j)} + \mathbf{b}, \quad j = 0, 1, \dots, n \quad (3.59)$$

a następnie na przyjęciu za przybliżenie rozwiązania α rozwiązania układu równań $Ax + b = 0$.

Niech ΔY oznacza macierz o kolumnach $y^{(1)} - y^{(0)}, y^{(2)} - y^{(1)}, \dots, y^{(n)} - y^{(n-1)}$, a ΔF — macierz o kolumnach $F(y^{(1)}) - F(y^{(0)}), F(y^{(2)}) - F(y^{(1)}), \dots, F(y^{(n)}) - F(y^{(n-1)})$. Z warunku (3.59) uzyskamy wówczas

$$\Delta F = A \cdot \Delta Y$$

Ponieważ rozwiązaniem układu równań $Ax + b = 0$ będzie punkt

$$x = -A^{-1}b = -A^{-1}(F(y) - Ay) = y - A^{-1}F(y)$$

więc jedna z możliwych wersji metody siecznych może przebiegać następująco:

0. Wybieramy punkty $x^{(-n)}, x^{(-n+1)}, \dots, x^{(0)}$ oraz przyjmujemy $i = 0$

1. Obliczamy macierze $\Delta F^{(i)}$ oraz $\Delta Y^{(i)}$, przyjmując

$$y^{(0)} = x^{(i-n)}, y^{(1)} = x^{(i-n+1)}, \dots, y^{(n)} = x^{(i)} \quad (3.60a)$$

2. Wyznaczamy punkty $x^{(i+1)}$ w sposób następujący

$$x^{(i+1)} = x^{(i)} - \Delta Y^{(i)}[\Delta F^{(i)}]^{-1} F(x^{(i)}) \quad (3.60b)$$

3. Przyjmujemy zamiast i numer $i + 1$ i wracamy do kroku 1

Wzory (3.60) określają tzw. $(n + 1)$ -punktową metodę siecznych, gdyż do wyznaczenia kolejnego punktu w ciągu $x^{(1)}, x^{(2)}, \dots$ jest potrzebne $(n + 1)$ poprzednich punktów i wartości odwzorowania F w tych punktach. Aby można było posłużyć się wzorem (3.60b), macierz $\Delta F^{(i)}$ musi być nieosobliwa. W przeciwnym przypadku należy w inny sposób wyznaczyć punkty $y^{(j)}, j = 0, 1, \dots, n$, rezygnując z części punktów $x^{(i-n)}, \dots, x^{(i)}$. W tych wybranych punktach y należy obliczyć wartość $F(y)$; nastąpi zatem wzrost nakładu obliczeń w danej iteracji. W skrajnym przypadku macierze ΔY i ΔF obliczamy bez korzystania z poprzednio wyznaczonych punktów $x^{(i-1)}, x^{(i-2)}, \dots, x^{(i-n)}$; przyjmujemy wówczas $y^{(n)} = x^{(i)}$, a brakujące punkty $y^{(n-1)}, \dots, y^{(0)}$ dobieramy np. zgodnie ze wzorem

$$y^{(n-j)} = x^{(i)} + he^{(j)}, \quad j = 1, 2, \dots, n \quad (3.61)$$

gdzie h jest pewną ustaloną liczbą, a wektory $e^{(j)}$ są wektorami przestrzeni R^n . Metoda siecznych sprowadza się wtedy do dyskretnej metody Newtona, w której pochodną $F'(x^{(i)})$ przybliżamy macierzą

$$\left[\frac{F(x^{(i)} + he^{(1)}) - F(x^{(i)})}{h}, \dots, \frac{F(x^{(i)} + he^{(n)}) - F(x^{(i)})}{h} \right]$$

Należy podkreślić, że w niektórych przypadkach nie można w żaden sposób wybrać punktów $y^{(0)}, y^{(1)}, \dots, y^{(n)}$ tak, by macierz ΔF była nieosobliwa. Jeśli np. $Fx = Bx + c$, gdzie B jest pewną macierzą osobliwą, a c pewnym wektorem, to $\Delta F = B \cdot \Delta Y$ jest macierzą osobliwą, a zatem przy

dowolnym układzie punktów y nie można obliczyć $x^{(i+1)}$ ze wzoru (3.60b). Z tego względu stosowane w praktyce algorytmy metody siecznych są uzupełnione dodatkowymi warunkami, gwarantującymi niezawodność metody (patrz np. [30]).

Ćwiczenia

1. Odwzorowanie $F: \mathbf{R}^n \rightarrow \mathbf{R}^n$ ma postać

$$F(x) = [f_1(x_1, x_2, \dots, x_n), f_2(x_1, x_2, \dots, x_n), \dots, f_n(x_1, x_2, \dots, x_n)]^T$$

Uzasadnić, że jeśli $f_n(x_1, x_2, \dots, x_n) = x_n$, to przy dowolnych punktach $y^{(n)}, y^{(n-1)}, \dots, y^{(0)}$ ciąg punktów $x^{(1)}, x^{(2)}, \dots$ generowany metodą siecznych ma tę własność, że n -te składowe punktów $x^{(1)}, x^{(2)}, \dots$ są równe zeru.

U w a g a. $(n+1)$ -punktową metodą siecznych, określoną wzorami (3.60), można w tym przypadku obliczyć co najwyżej n punktów, gdyż dla dalszych punktów macierz ΔF ma zerowy ostatni wiersz, więc jest osobliwa.

2. Niech macierze ΔY i ΔF mają jako kolumny wektory $y^{(1)} - y^{(0)}, y^{(2)} - y^{(1)}, \dots, y^{(n)} - y^{(n-1)}$ oraz $F(y^{(1)}) - F(y^{(0)}), F(y^{(2)}) - F(y^{(1)}), \dots, F(y^{(n)}) - F(y^{(n-1)})$, a macierze $\overline{\Delta Y}$ i $\overline{\Delta F}$ mają jako kolumny wektory $y^{(1)} - y^{(0)}, y^{(2)} - y^{(0)}, \dots, y^{(n)} - y^{(0)}$ oraz $F(y^{(1)}) - F(y^{(0)}), F(y^{(2)}) - F(y^{(0)}), \dots, F(y^{(n)}) - F(y^{(0)})$. Uzasadnić, że macierze ΔF i $\overline{\Delta F}$ są równocześnie osobliwe lub nieosobliwe oraz że w przypadku, gdy są nieosobliwe

$$\Delta Y (\Delta F)^{-1} = \overline{\Delta Y} (\overline{\Delta F})^{-1}$$

U w a g a. Powyższy wynik można uogólnić i wykazać, że porządek punktów $y^{(n)}, y^{(n-1)}, \dots, y^{(0)}$, za pomocą których obliczamy kolejny punkt $x^{(i+1)}$, nie jest istotny.

3.5. POSZUKIWANIE MINIMÓW FUNKCJI JEDNEJ ZMIENNEJ

W wielu przypadkach zadanie znalezienia minimum funkcji $f: \mathbf{R}^1 \rightarrow \mathbf{R}^1$ można sprowadzić do rozwiązania równania $f'(x) = 0$, gdzie $f': \mathbf{R}^1 \rightarrow \mathbf{R}^1$ oznacza pochodną funkcji f . Może się jednak zdarzyć, że wyznaczenie pochodnej jest zbyt trudne lub funkcja f nie jest różniczkowalna. Jeżeli funkcja f jest dostatecznie regularna (np. można ją lokalnie przybliżyć wielomianami niskiego rzędu), to można posłużyć się metodami aproksymacyjnymi. Jeśli jednak nie znamy własności funkcji, bezpieczniejsze są tzw. *metody podziału*, które omówimy poniżej.

Założmy, że na przedziale $\langle a; b \rangle$ funkcja f ma minimum w punkcie $\alpha \in \langle a; b \rangle$ i że na przedziale $\langle a; \alpha \rangle$ funkcja jest malejąca, a na przedziale $\langle \alpha; b \rangle$ funkcja jest rosnąca. Taką funkcję przyjęto nazywać *funkcją unimodalną*.

Lemat. Aby zlokalizować punkt α w przedziale $\langle a'; b' \rangle$ o mniejszej długości niż przedział $\langle a; b \rangle$, wystarczy obliczyć wartość funkcji w dwu punktach wewnątrz przedziału $\langle a; b \rangle$.

DOWÓD. Wybierzmy dwa dowolne punkty t_1 i t_2 takie, że $a < t_1 < t_2 < b$. Jeżeli $f(t_1) \leq f(t_2)$, to $\alpha \in \langle a; t_2 \rangle$. Jeżeli $f(t_1) > f(t_2)$, to $\alpha \in \langle t_1; b \rangle$.

3.5.1. Metody podziału

Podany powyżej lemat umożliwia wyznaczenie ciągu przedziałów $\langle a^{(i)}; b^{(i)} \rangle$, $i = 0, 1, \dots$, $a^{(0)} = a$, $b^{(0)} = b$ takich, że $\langle a^{(i+1)}; b^{(i+1)} \rangle \subset \langle a^{(i)}; b^{(i)} \rangle$. Jeśli tak wybieramy punkty, aby $b^{(i)} - a^{(i)} \rightarrow 0$, to na mocy twierdzenia Cantora ciągi $(a^{(i)})$ oraz $(b^{(i)})$ są zbieżne do jednego punktu, którym jest α .

Przedstawiamy kilka sposobów podziału przedziału $\langle a^{(i)}; b^{(i)} \rangle$ i porównamy je pod względem efektywności obliczeń. Ogólny schemat algorytmu jest przedstawiony na rys. 3.7. Liczba ϱ oznacza tu pożądaną dokładność wyznaczenia punktu α , tzn. chcemy uzyskać taki punkt t , by $\alpha \in \langle t - \varrho; t + \varrho \rangle$. Jako miarę efektywności metody, służącą do porównania algorytmów o różnych sposobach wyboru punktów podziału $t_1^{(i)}, \dots, t_k^{(i)}$, przyjmujemy liczbę obliczeń funkcji potrzebnych do wyznaczenia punktu t .

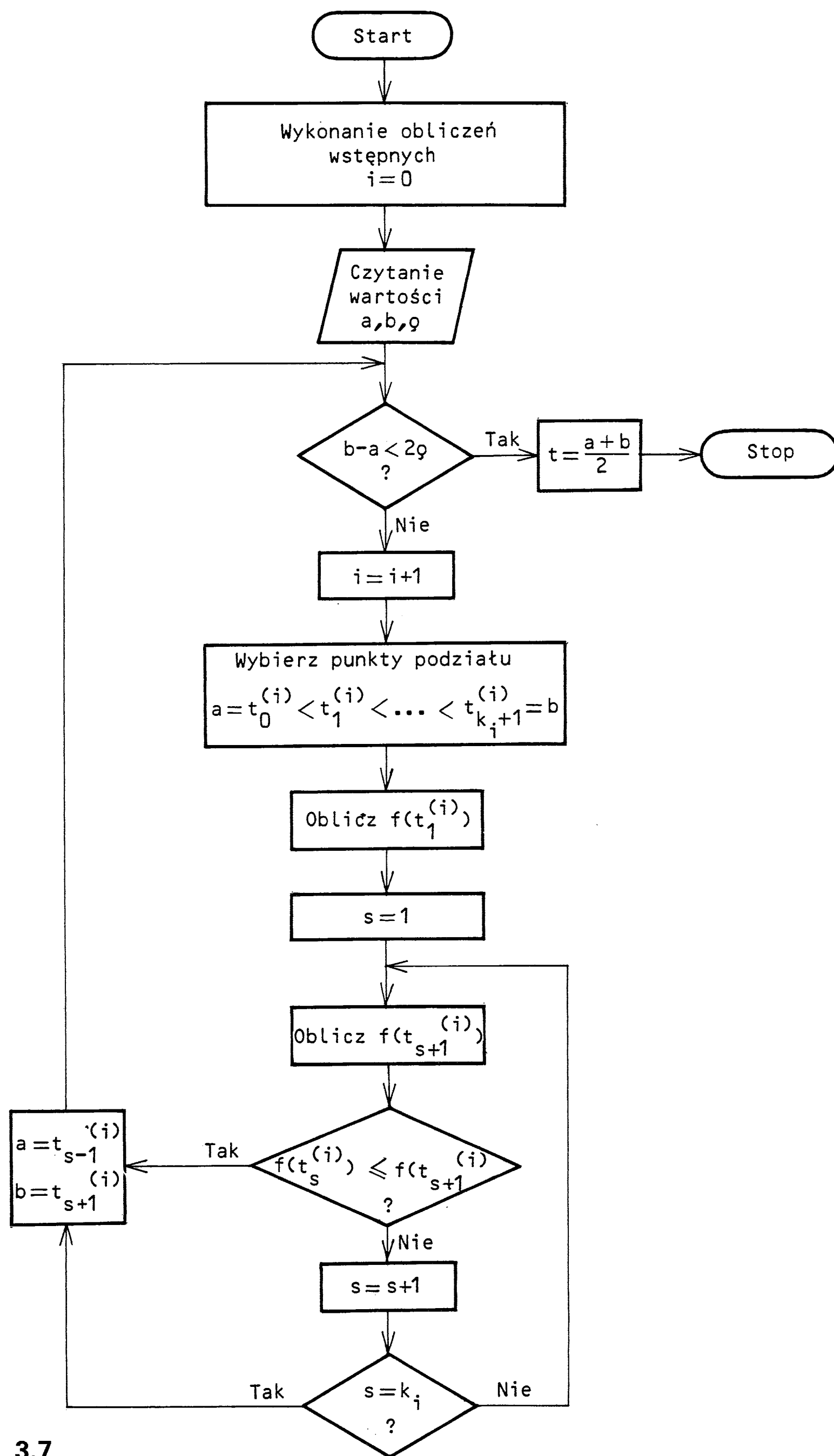
Metoda podziału na trzy części. Jako punkty podziału przyjmujemy $t_1^{(i)} = \frac{2}{3}a + \frac{1}{3}b$, $t_2^{(i)} = \frac{1}{3}a + \frac{2}{3}b$, które dzielą przedział $\langle a; b \rangle$ na trzy równe części. W każdej iteracji następuje zmniejszenie długości przedziału $3/2$ razy. Po I iteracjach uzyskamy zatem przedział o długości $b - a = \left(\frac{2}{3}\right)^I (b^{(0)} - a^{(0)})$ przy $2I$ obliczeniach funkcji. W metodzie tej każde obliczanie wartości funkcji zmniejsza przedział średnio $\sqrt{\frac{3}{2}}$ razy.

Metoda połowienia. Jako punkty podziału przyjmujemy $t_1^{(i)} = \frac{3}{4}a + \frac{1}{4}b$, $t_2^{(i)} = \frac{1}{2}a + \frac{1}{2}b$, $t_3^{(i)} = \frac{1}{4}a + \frac{3}{4}b$, które dzielą przedział $\langle a; b \rangle$ na cztery równe części. W każdej iteracji następuje zmniejszenie długości przedziału 2 razy. Po I iteracjach uzyskamy zatem przedział o długości $b - a = \left(\frac{1}{2}\right)^I (b^{(0)} - a^{(0)})$ przy $2I + 1$ obliczeniach funkcji w najgorszym przypadku (w pierwszej iteracji obliczamy 3 lub 2 wartości funkcji).

W metodzie połowienia każde obliczenie wartości funkcji zmniejsza przedział średnio 2 razy, choć w korzystnym przypadku jedno obliczenie wartości funkcji może zmniejszyć przedział 4 razy. Widzimy, że metoda połowienia jest bardziej ekonomiczna niż metoda podziału na trzy części. Nasuwa się naturalne pytanie, jak należy wybierać punkty podziału, aby osiągnąć t przy najmniejszej liczbie obliczeń wartości funkcji? Zajmiemy się zatem tym zagadnieniem.

3.5.2. Metoda optymalnych podziałów

Spośród metod podziału najmniejszej liczby obliczeń funkcji wymaga metoda korzystająca z liczb Fibonacciego (uzasadnił to Johnson, patrz np. [24]).



Przypominamy, że liczby Fibonacciego F_i są określone wzorem rekurencyjnym

$$F_0 = F_1 = 1, \quad F_i = F_{i-1} + F_{i-2}, \quad i = 2, 3, \dots$$

Na przykład: $F_7 = 21$, $F_{10} = 89$.

Algorytm optymalnego podziału jest następujący. Niech $c = \frac{b^{(0)} - a^{(0)}}{\varrho}$.

Znajdujemy takie N , aby $F_{N-1} < c \leq F_N$. Określamy

$$t_1^{(i)} = \frac{F_{N-i-1}}{F_{N-i+1}}(b - a) + a$$

$$t_2^{(i)} = \frac{F_{N-i}}{F_{N-i+1}}(b - a) + a, \quad i = 1, 2, \dots, N - 2$$

Zgodnie z rys. 3.7 w każdej iteracji obliczamy nowe punkty a i b w sposób następujący:

Jeżeli $f(t_1^{(i)}) \leq f(t_2^{(i)})$, to a pozostaje bez zmiany, $b = t_2^{(i)}$.

Jeżeli $f(t_1^{(i)}) > f(t_2^{(i)})$, to $a = t_1^{(i)}$, a b pozostaje bez zmiany.

Po i -tej iteracji długość przedziału $\langle a; b \rangle$ zostaje zmniejszona $\frac{F_{N-1}}{F_{N-i+1}}$ razy bez względu na to, która nierówność jest spełniona.

Po $(N - 2)$ iteracjach długość przedziału zostaje zmniejszona do wartości

$$(b - a) = \frac{F_{N-1}}{F_N} \cdot \frac{F_{N-2}}{F_{N-1}} \dots \frac{F_2}{F_3} (b^{(0)} - a^{(0)}) = 2 \cdot \frac{b^{(0)} - a^{(0)}}{F_N} \leq 2\varrho$$

Wykonano łącznie $N - 1$ obliczeń wartości funkcji.

Przykład. Znaleźć minimum funkcji $f(x) = |x|$ zlokalizowane w przedziale $\langle -4; 4 \rangle$. Pożądana dokładność $\varrho = 1$.

Przykład ten rozwiążemy dwiema metodami:

a) Metoda połowienia

$$t_1^{(1)} = -2, \quad t_2^{(1)} = 0, \quad t_3^{(1)} = 2$$

$$f(-2) < f(0), \quad f(2) > f(0) \Rightarrow \langle a; b \rangle = \langle -2; 2 \rangle$$

$$t_1^{(2)} = -1, \quad t_2^{(2)} = 0, \quad t_3^{(2)} = 1$$

$$f(-1) > f(0), \quad f(1) > f(0) \Rightarrow \langle -a; b \rangle = \langle 1; 1 \rangle \Rightarrow t = 0$$

$$f(t) = 0$$

Wykonano łącznie 5 obliczeń funkcji.

b) Metoda Johnsona

$$F_5 = 8 = \frac{b - a}{\varrho}, \quad \text{więc} \quad N = 5$$

$$t_1^{(1)} = \frac{3}{8} \cdot 8 + (-4) = -1, \quad t_2^{(1)} = \frac{5}{8} \cdot 8 + (-4) = 1$$

$$f(-1) = f(1) \Rightarrow \langle a; b \rangle = \langle -4; 1 \rangle$$

$$t_1^{(2)} = \frac{2}{5} \cdot 5 + (-4) = -2, \quad t_2^{(2)} = \frac{3}{5} \cdot 5 + (-4) = -1$$

$$f(-2) > f(-1) \Rightarrow \langle a; b \rangle = \langle -2; 1 \rangle$$

$$t_1^{(3)} = \frac{1}{3} \cdot 3 + (-2) = -1, \quad t_2^{(3)} = \frac{2}{3} \cdot 3 + (-2) = 0$$

$$f(-1) > f(0) \Rightarrow \langle a; b \rangle = \langle -1; 1 \rangle \Rightarrow t = 0$$

$$f(t) = 0$$

Wykonano łącznie 4 obliczenia wartości funkcji.

U w a g a. Ponieważ do zmniejszenia $\frac{F_N}{2}$ razy długości przedziału potrzeba $N - 1$ obliczeń wartości funkcji, więc jedno obliczenie wartości funkcji zmniejsza średnio przedział $\sqrt[N-1]{\frac{F_N}{2}}$ razy. Można wykazać, że $F_N = \left(\frac{1+\sqrt{5}}{2\sqrt{5}}\right)\left(\frac{1+\sqrt{5}}{2}\right)^N + \left(\frac{\sqrt{5}-1}{2\sqrt{5}}\right)\left(\frac{1-\sqrt{5}}{2}\right)^N$.
Zatem $\lim_{N \rightarrow \infty} \sqrt[N-1]{\frac{F_N}{2}} = \frac{1+\sqrt{5}}{2} \approx 1,61$.

Dla porównania wskaźnik ten dla metody podziału na trzy części wynosi $\sqrt[3]{\frac{3}{2}} \approx 1,22$, a dla metody połowienia $\sqrt{2} \approx 1,41$.

3.5.3. Metoda złotego podziału

Metoda ta polega na takim wyborze punktów podziału $t_1^{(i)}$ oraz $t_2^{(i)}$, aby przedział $\langle a; b \rangle$ zmniejszał swą długość po każdej iteracji tyle samo razy oraz by po wyznaczeniu punktów nowego podziału, tzn. $t_1^{(i+1)}$ i $t_2^{(i+1)}$, jeden z tych punktów pokrywał się z wyznaczonym punktem podziału w poprzedniej iteracji. Ma to na celu zmniejszenie obliczeń wartości funkcji, gdyż jedynie w pierwszej iteracji obliczamy dwie wartości funkcji, w następnych zaś już tylko jedną wartość funkcji. Zauważmy, że cechę tę miał omówiony poprzednio algorytm optymalny. Wymagania powyższe spełnia algorytm, w którym

$$t_2^{(i)} - a = b - t_1^{(i)} = \tau(b - a), \quad \tau \in (0; 1)$$

$$b - t_2^{(i)} = \tau(b - t_1^{(i)})$$

Stąd $\tau^2 + \tau - 1 = 0$, czyli $\tau = \frac{\sqrt{5}-1}{2} \approx 0,62$. Liczba τ jest stosunkiem

boków prostokąta, zwanego przez starożytnych Greków „złotym”, stąd nazwa metody.

W metodzie złotego podziału punkty podziału obliczamy ze wzorów

$$\begin{aligned} t_1^{(1)} &= a + (1 - \tau)(b - a) \\ t_2^{(1)} &= b - (1 - \tau)(b - a) \end{aligned} \quad (3.62a)$$

Jeżeli $f(t_1^{(i)}) \leq f(t_2^{(i)})$, to a pozostaje bez zmian, zaś

$$\begin{aligned} b &= t_2^{(i)} \\ t_2^{(i+1)} &= t_1^{(i)} \\ t_1^{(i+1)} &= a + (1 - \tau)(b - a) \end{aligned} \quad (3.62b)$$

Jeżeli $f(t_1^{(i)}) > f(t_2^{(i)})$, to b pozostaje bez zmian, zaś

$$\begin{aligned} a &= t_1^{(i)} \\ t_1^{(i+1)} &= t_2^{(i+1)} \\ t_2^{(i+1)} &= b - (1 - \tau)(b - a), \quad i = 1, 2, 3, \dots \end{aligned} \quad (3.62c)$$

We wzorach (3.62a) celowo zamiast mnożnika τ przyjęto mnożnik $1 - \tau$ (można bowiem obliczać $t_2^{(1)} = a + \tau(b - a)$), gdyż powoduje to mniejsze błędy zaokrągleń przy wyznaczaniu kolejnych punktów podziału.

Interesujące jest porównanie efektywności metody Johnsona i metody złotego podziału. Jeżeli zakładamy K obliczeń wartości funkcji (łącznie z wartością w uzyskanym ostatecznie punkcie t), to w metodzie Johnsona należy przyjąć $N = K + 1$ i wówczas $|t - \alpha| \leq \frac{b^{(0)} - a^{(0)}}{F_{K+1}}$. W metodzie złotego podziału uzyskamy zaś $|t - \alpha| \leq \frac{b^{(0)} - a^{(0)}}{2G_{K-1}}$ gdzie $G_i = \left(\frac{1}{\tau}\right)^{i-1}$. Korzystając z powyższych oznaczeń można uzasadnić, że

$$2G_{K-1} < F_{K+1} < 2G_K \quad (3.63)$$

Zatem, aby wyznaczyć t metodą złotego podziału z dokładnością nie gorszą niż metodą Johnsona, potrzeba co najwyżej jednego dodatkowego obliczenia wartości funkcji.

Ćwiczenia

1. Uzasadnić nierówność (3.63). Wskazać. Zastosować indukcję matematyczną.
2. Minimum funkcji $f(x) = |x|$ jest zlokalizowane w przedziale $\langle 0; 8 \rangle$. Wyznaczyć punkt minimum z dokładnością $\varrho = 1$ stosując metodę połowienia i metodę Johnsona. Dlaczego metoda połowienia okazała się w tym przypadku skuteczniejsza?

4

CAŁKOWANIE NUMERYCZNE

4.1. WSTĘP

4.1.1. Uwagi ogólne o całkowaniu numerycznym

W niniejszym rozdziale rozpatrzymy zagadnienie przybliżonego obliczania całki oznaczonej funkcji jednej zmiennej. Ponieważ wyznaczanie funkcji pierwotnej w wielu przypadkach jest bardzo trudne lub wręcz niemożliwe, stosowanie metod przybliżonych jest więc często konieczne. Ponadto, jeżeli funkcja podcałkowa jest określona za pomocą tablicy, to pojęcie funkcji pierwotnej traci sens i wówczas możemy obliczyć tylko przybliżoną wartość całki.

Gdy przedział całkowania jest skończony, wówczas wiele sposobów przybliżonego obliczania całek polega na tym, że funkcję podcałkową $F(x)$ zastępujemy interpolującą funkcją $\varphi(x)$, którą łatwo można całkować. Niech $\varphi(x)$ będzie wielomianem interpolacyjnym Lagrange'a dla funkcji $F(x)$ z węzłami interpolacji $x_0, x_1, \dots, x_N$ (patrz p. 1.2.1)

$$\varphi(x) = L_N(x) = \sum_{k=0}^N \Phi_k(x) F(x_k)$$

$$\Phi_k(x) = \frac{(x - x_0) \dots (x - x_{k-1})(x - x_{k+1}) \dots (x - x_N)}{(x_k - x_0) \dots (x_k - x_{k-1})(x_k - x_{k+1}) \dots (x_k - x_N)}$$

Podstawiając w miejsce funkcji podcałkowej $F(x)$ wielomian $\varphi(x)$, otrzymamy

$$\int_a^b F(x) dx \approx \int_a^b \varphi(x) dx = \sum_{k=0}^N A_k F(x_k), \quad A_k = \int_a^b \Phi_k(x) dx$$

Jeżeli $|F(x) - \varphi(x)| < \varepsilon$ i $x \in \langle a; b \rangle$, to

$$\left| \int_a^b F(x) dx - \sum_{k=0}^N A_k F(x_k) \right| = \left| \int_a^b (F(x) - \varphi(x)) dx \right| \leq \varepsilon (b - a)$$

Przedstawiony sposób pozwala zatem obliczyć całkę z dowolną dokładnością, jeśli tylko funkcję $F(x)$ można przybliżyć wielomianami z dowolną dokładnością; oprócz wielomianów można oczywiście stosować także inne funkcje interpolacyjne $\varphi(x)$, np. funkcje wymierne, trygonometryczne.

Podobny wzór przybliżonego całkowania otrzymamy, gdy przedział całkowania podzielimy na podprzedziały i w każdym podprzedziale zastosujemy powyższy sposób obliczania całki.

Jeżeli funkcja $F(x)$ ma osobliwości utrudniające dobre przybliżenie (np. gdy funkcja $F(x)$ jest nieograniczona lub jej pochodne niskiego rzędu nie istnieją w przedziale całkowania), to opisane powyżej postępowanie można zmodyfikować następująco. Funkcję podcałkową przedstawiamy w postaci iloczynu $F(x) = p(x)f(x)$, gdzie $f(x)$ jest już funkcją, którą można dobrze przybliżyć, a $p(x)$ ma wszelkie osobliwości funkcji podcałkowej utrudniające to przybliżenie. Niech teraz $\varphi(x)$ będzie wielomianem interpolacyjnym Lagrange'a dla funkcji $f(x)$ z węzłami interpolacji $x_0, x_1, \dots, x_N$. Podstawiając w miejsce $f(x)$ wielomian $\varphi(x)$, otrzymamy

$$\int_a^b F(x) dx = \int_a^b p(x)f(x) dx \approx \int_a^b p(x)\varphi(x) dx = \sum_{k=0}^N A'_k f(x_k)$$

$$A'_k = \int_a^b p(x)\Phi_k(x) dx$$

Ten zmodyfikowany sposób można zastosować i do obliczania całek na nieskończonym przedziale całkowania. Funkcja $p(x)$ określa wówczas szybkość malenia $F(x)$, gdy $x \rightarrow \pm \infty$.

Rozbicie funkcji podcałkowej $F(x)$ na iloczyn funkcji $f(x)$ i $p(x)$ niekoniecznie musi być związane z wydzieleniem osobliwości $F(x)$. W metodzie Gaussa (patrz p. 4.3.1) funkcję $p(x)$ dobiera się często ze względu na wybór odpowiedniego typu wielomianów ortogonalnych.

4.1.2. Ogólny wzór całkowania numerycznego

Do przybliżonego obliczania całek

$$I(f) = \int_a^b p(x)f(x) dx \tag{4.1}$$

będziemy stosować wzory postaci

$$S(f) = \sum_{k=0}^N A_k f(x_k), \quad x_k \in \langle a; b \rangle \tag{4.2}$$

przy czym współczynniki A_k nie zależą od funkcji $f(x)$. Wzór (4.2) nazywamy *kwadraturą*, a x_k — *węzłami kwadratury*.

Funkcję $p(x)$, występującą w (4.1) nazywamy *funkcją wagową* lub krótko *wagą*. Przy wyprowadzaniu wzorów postaci (4.2) wagę będziemy przyjmować za funkcję ustaloną. Z tego też powodu wzory te będziemy stosować do obliczania całek funkcji mających osobliwości tego samego typu, określonego wagą $p(x)$. Będziemy również zakładać, że waga jest nieujemna na przedziale całkowania i taka, że funkcja $p(x)W(x)$ jest całkowalna, gdy $W(x)$ jest wielomianem dowolnego stopnia.

Błąd przybliżenia całki (4.1) sumą (4.2), czyli *błąd metody*, oznaczmy jako

$$E(f) = I(f) - S(f) \quad (4.3)$$

Za *kryterium dokładności kwadratury* przyjmiemy zgodność $S(W)$ i $I(W)$, gdy W jest wielomianem.

Mówimy, że wzór (4.2) jest *rzędu r* , jeżeli:

1) $I(W) = S(W)$ dla wszystkich wielomianów $W(x)$ stopnia mniejszego niż r ,

2) istnieje wielomian $W(x)$ stopnia r taki, że $I(W) \neq S(W)$. Zakładać będziemy, że $r \geq 1$.

Współczynniki A_k rozpatrywanych kwadratur, poza omawianą w p. 4.2.3, obliczać będziemy według wzoru podanego w poprzednim punkcie. Będziemy przyjmować, że na całym przedziale całkowania lub na jego podprzedziałach funkcja $\varphi(x)$ jest wielomianem interpolacyjnym.

Podamy teraz bez dowodu *warunki zbieżności kwadratury* postaci (4.2) w klasie funkcji $C(\langle a; b \rangle)$, tzn. warunki, przy których $E(f) \rightarrow 0$, gdy $N \rightarrow \infty$. W ogólnym przypadku węzły i współczynniki w (4.2) zależą od N ; będziemy to zaznaczać pisząc x_k^N , A_k^N ($0 \leq k \leq N$).

Tw. *Kwadratura postaci (4.2) jest zbieżna dla każdej funkcji $f \in C(\langle a; b \rangle)$ wtedy i tylko wtedy, gdy:*

- 1) *jest ona zbieżna dla każdego wielomianu*
- 2) *istnieje liczba M niezależna od N taka, że*

$$B_N = \sum_{k=0}^N |A_k^N| \leq M \quad \text{dla} \quad N = 1, 2, \dots$$

U w a g a. Jeżeli A_k^N są nieujemne, to warunek 2) może być opuszczony. Istotnie, ponieważ wzór (4.2) jest dokładny dla funkcji $f(x) = 1$, więc

$$\sum_{k=0}^N A_k^N = \int_a^b p(x) dx \quad (4.4)$$

Dla nieujemnych współczynników A_k^N mamy $B_N = \int_a^b p(x) dx$.

4.2. KWADRATURY Z USTALONYMI WĘZŁAMI

Zakładamy, że węzły x_k ($k = 0, 1, \dots, N$) kwadratury (4.2) są już ustalone. W niniejszym punkcie zajmiemy się wyborem współczynników A_k , przy założeniu, że przedział całkowania jest skończony.

4.2.1. Kwadratury Newtona-Cotesa

Rozpatrzmy następujący sposób obliczania całki (4.1)

$$S(f) = I(L_N) = \sum_{k=0}^N A_k f(x_k), \quad A_k = \int_a^b p(x) \Phi_k(x) dx \quad (4.5)$$

gdzie L_N jest wielomianem interpolacyjnym Lagrange'a dla funkcji $f(x)$ z węzłami interpolacji $x_0, x_1, \dots, x_N$. Przypominamy, że

$$L_N(x) = \sum_{k=0}^N \Phi_k(x) f(x_k), \quad (4.6)$$

$$\Phi_k(x) = \frac{(x - x_0) \dots (x - x_{k-1})(x - x_{k+1}) \dots (x - x_N)}{(x_k - x_0) \dots (x_k - x_{k-1})(x_k - x_{k+1}) \dots (x_k - x_N)}$$

i że wzór na błąd przybliżenia $f(x)$ wielomianem $L_N(x)$ jest następujący

$$R_{N+1}(x) = f(x) - L_N(x) = \frac{1}{(N+1)!} \omega_{N+1}(x) f^{(N+1)}(\xi), \quad \xi \in (a; b) \quad (4.7)$$

przy czym $\omega_{N+1}(x) = (x - x_0) \dots (x - x_N)$.

Tw. *Rząd kwadratury (4.2) wynosi co najmniej $(N+1)$ wtedy i tylko wtedy gdy kwadratura ta jest postaci (4.5).*

DOWÓD. Załóżmy, że kwadratura (4.2) jest postaci (4.5). Niech $f(x)$ będzie dowolnym wielomianem stopnia co najwyżej N . Wielomian ten zapiszemy w postaci wielomianu interpolacyjnego Lagrange'a $L_N(x)$ z węzłami $x_0, \dots, x_N$. Mamy wówczas

$$f(x) \equiv L_N(x) \quad \text{oraz} \quad I(f) = I(L_N) = S(L_N) = S(f)$$

A zatem rząd kwadratury (4.5) wynosi co najmniej $N+1$.

Założmy teraz, że rząd kwadratury (4.2) wynosi co najmniej $N+1$. Jest więc ona na pewno dokładna dla wszystkich wielomianów stopnia mniejszego niż $(N+1)$, a zatem jest dokładna i dla wielomianów $f(x) = \Phi_i(x)$. Mamy

$$I(f) = \int_a^b p(x) \Phi_i(x) dx = \sum_{k=0}^N A_k \Phi_i(x_k) = A_i$$

czyli kwadratura (4.2) musi być postaci (4.5).

Kwadratury (4.5) z węzłami równoodległymi nazywamy *kwadraturami Newtona-Cotesa*. Wśród nich największe znaczenie praktyczne mają wzory, zwane *kwadraturami (wzorami) zamkniętymi*, w których końce przedziału są węzłami i waga $p(x) \equiv 1$. W dalszym ciągu rozważania nasze ograniczymy do takich wzorów. Mają one postać

$$S(f) = \sum_{k=0}^N A_k f_k \quad (4.8)$$

$$A_k = \int_a^b \Phi_k(x) dx = h \frac{(-1)^{N-k}}{k!(N-k)!} \int_0^N \frac{t(t-1)\dots(t-N)}{(t-k)} dt$$

$$\text{gdzie } h = \frac{b-a}{N}, \quad f_k = f(a + kh).$$

Przedstawimy teraz podstawowe własności kwadratur postaci (4.8).

1. Jeżeli N jest liczbą nieparzystą, to kwadratura (4.8) nie jest dokładna dla wszystkich wielomianów stopnia $(N+1)$ i $(N+1)$ jest jej rzędem dokładności. Gdy N jest liczbą parzystą, wówczas kwadratura (4.8) jest dokładna dla wielomianów stopnia $(N+1)$ (patrz ćwic. 1) i jej rząd wynosi $(N+2)$. Dla dowolnego N rząd kwadraturowy wynosi zatem

$$2[N/2] + 2 \quad (4.9)$$

gdzie $[N/2]$ oznacza całkowitą część liczby $N/2$.

2. Jeżeli $f \in C^r(\langle a; b \rangle)$, gdzie $r = 2[N/2] + 2$, to błąd kwadraturowy można przedstawić w postaci

$$E(f) = C_r f^{(r)}(\xi) \quad (4.10)$$

przy czym $\xi \in (a; b)$, a współczynnik C_r nie zależy od f .

3. Przy stosowaniu kwadratur (4.8) dla dużych N oszacowanie błędu nastręcza duże trudności. Z przedstawienia (4.10) wynika, że błąd kwadratury zależy od $f^{(r)}$, a jak wiadomo pochodne wysokiego rzędu trudno jest oszacować. Inną trudnością jest to, że pochodne te bardzo często rosną nieograniczenie; ten szybki wzrost może już wystąpić dla pochodnych niskiego rzędu. Ponadto dla dużych N kwadratura (4.8) jest mało odporna na niedokładności obliczeń na maszynach cyfrowych. Dla $N \geq 10$ wśród współczynników A_k zawsze występują współczynniki ujemne. Dla $N = 10, 15, 20$

suma $\frac{1}{b-a} \sum_{k=0}^N |A_k|$ wynosi w przybliżeniu odpowiednio: 3,1, 8,3, 560.

Wyjaśnimy pokrótce wpływ tego faktu na dokładność obliczania $S(f)$, przyjmując jako zasadę, że dysponujemy tylko przybliżonymi wartościami funkcji $f(x)$. Niech wartości $f(x_i)$ będą obliczone z błędem co do modułu nie przekraczającym ε . Wtedy moduł błędu sumy $S(f)$ szacuje się wielkością

$$\varepsilon \cdot \sum_{k=0}^N |A_k| \text{ i z (4.4) widać, że oszacowanie to jest najmniejsze, gdy współ-}$$

czynniki A_k są nieujemne. Tak więc strata dokładności przy obliczaniu $S(f)$ za pomocą wzoru (4.8) dla dużych N może być znaczna.

4. Zaznaczamy, że A_k zależą od N ($A_k = A_k^N$). Dla pewnych k mamy

$\lim_{N \leftarrow \infty} |A_k^N| = \infty$, a więc na podstawie twierdzenia z p. 4.1.1 metoda kwadratur Newtona-Cotesa nie jest zbieżna w klasie funkcji ciągłych. Można łatwo udowodnić (patrz ćwic. 2), że gdy ciąg wielomianów $L_N(x)$, określony wzorem (4.5), jest jednostajnie zbieżny do funkcji $f(x)$ na przedziale $\langle a; b \rangle$, wówczas metoda kwadratur jest zbieżna do funkcji $f(x)$ (powyższa własność jest prawdziwa dla dowolnych węzłów x_k). Z rozważań przeprowadzonych w p. 1.2.7 wynika, że jeżeli $f(x)$ jest rozwijalna w szereg Taylora o środku w punkcie $\frac{a+b}{2}$ i promieniu zbieżności większym niż $\frac{b-a}{2}e$, to ciąg $(L_N(x))$ jest jednostajnie zbieżny do $f(x)$ na przedziale $\langle a; b \rangle$. Dokładniejsze informacje o własnościach kwadratur postaci (4.8) może Czytelnik znaleźć np. w [40], [41] i [50].

Podamy teraz wzory (4.8) dla przybliżonego obliczania całek (4.1)–(4.3), gdy $N = 1$ i $N = 2$ oraz ich błędy.

Dla $N = 1$ mamy

$$A_0 = -h \int_0^1 (t-1) dt = \frac{1}{2}h$$

$$A_1 = h \int_0^1 t dt = \frac{1}{2}h$$

a więc

$$S(f) = \frac{1}{2}h(f_0 + f_1) \quad (4.11)$$

Jest to znany wzór *trapezów* (rys. 4.1).

Błąd tego wzoru na podstawie (4.7) wynosi

$$E(f) = \frac{1}{2!} \int_a^b (x-a)(x-b) f^{(2)}(\xi) dx \quad (4.12)$$

gdy $f \in C^2(\langle a; b \rangle)$. Funkcja $\omega_2(x) = (x-a)(x-b)$ nie zmienia znaku na przedziale $\langle a; b \rangle$, możemy zatem do (4.12) zastosować uogólnione twierdzenie o wartości średniej dla całek. Otrzymamy wówczas

$$E(f) = \frac{1}{2} f^{(2)}(\xi_1) \int_a^b (x-a)(x-b) dx = -\frac{1}{12} h^3 f^{(2)}(\xi_1) \quad (4.13)$$

przy czym $\xi_1 \in (a; b)$, $h = b - a$.

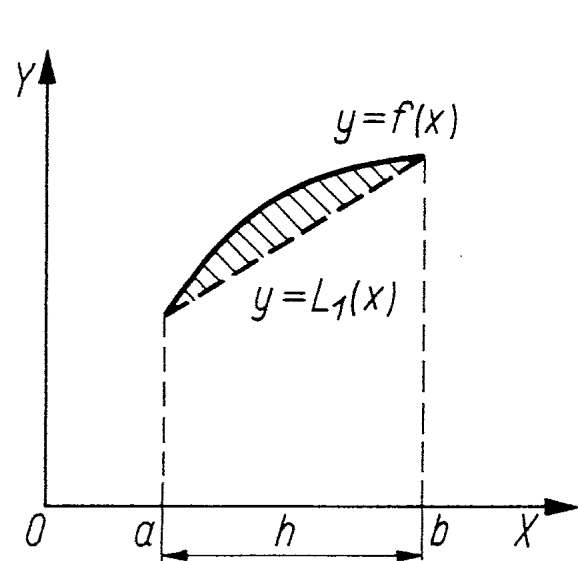
Dla $N = 2$ otrzymujemy

$$A_0 = \frac{1}{3}h, \quad A_1 = \frac{4}{3}h, \quad A_2 = \frac{1}{3}h$$

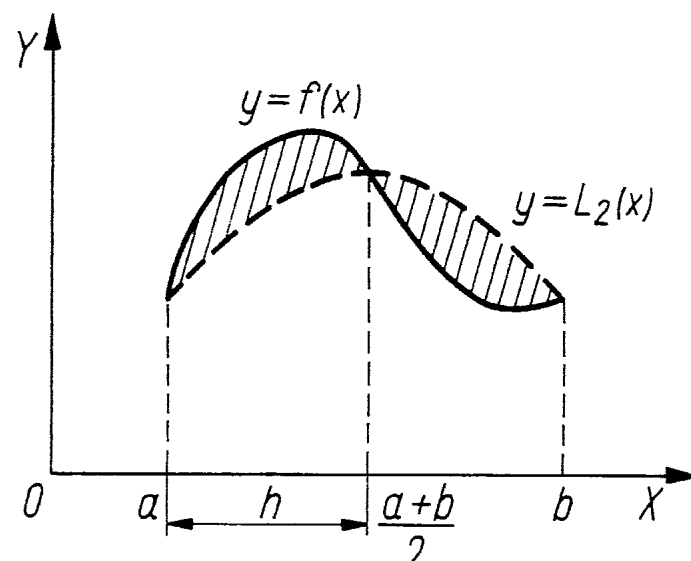
a więc

$$S(f) = \frac{1}{3}h(f_0 + 4f_1 + f_2) \quad (4.14)$$

Jest to znany wzór *parabol*, zwany też *wzorem Simpsona* (rys. 4.2).



4.1



4.2

Jak powiedzieliśmy powyżej, wzór (4.14) jest także dokładny dla wielomianów stopnia trzeciego. Niech $f \in C^4(\langle a; b \rangle)$ i niech $L_3(x)$ będzie wielomianem interpolacyjnym dla funkcji $f(x)$ z węzłami a , $\frac{a+b}{2}$, b i x_4 , gdzie x_4 jest dowolnym, ale różnym od pozostałych węzłów punktem z przedziału $\langle a; b \rangle$.

Mamy wówczas

$$f(x) - L_3(x) = \frac{1}{4!} (x-a) \left(x - \frac{a+b}{2} \right) (x-b) (x-x_4) f^{(4)}(\xi), \quad \xi \in (a; b) \quad (4.15)$$

Przechodząc w (4.15) do granicy z $x_4 \rightarrow \frac{a+b}{2}$, otrzymamy

$$f(x) - \tilde{L}_3(x) = \frac{1}{4!} (x-a) \left(x - \frac{a+b}{2} \right)^2 (x-b) f^{(4)}(\tilde{\xi}), \quad \tilde{\xi} \in \langle a; b \rangle \quad (4.16)$$

Błąd kwadratury (4.14) wynosi więc

$$E(f) = \frac{1}{4!} \int_a^b (x-a) \left(x - \frac{a+b}{2} \right)^2 (x-b) f^{(4)}(\tilde{\xi}) dx \quad (4.17)$$

Po zastosowaniu do (4.17) uogólnionego twierdzenia o wartości średniej dla całek, otrzymamy

$$E(f) = \frac{1}{4!} f^{(4)}(\xi_1) \int_a^b (x-a) \left(x - \frac{a+b}{2} \right)^2 (x-b) dx = -\frac{1}{90} h^5 f^{(4)}(\xi_1) \quad (4.18)$$

gdzie $\xi_1 \in \langle a; b \rangle$, $h = \frac{b-a}{2}$.

Kwadratury Newtona-Cotesa, w których końce przedziału nie są węzłami, mają mniejsze znaczenie praktyczne. Mają one zastosowanie przy rozwiązywaniu równań różniczkowych zwyczajnych.

Ćwiczenia

1. Udowodnić, że gdy N jest liczbą parzystą, to kwadratura Newtona-Cotesa (4.8) jest dokładna dla wszystkich wielomianów stopnia $(N + 1)$. W s k a z ó w k a. Przedstawić dowolny wielomian $W_{N+1}(x)$ stopnia $(N + 1)$ w postaci

$$W_{N+1}(x) = a_0 \left(x - \frac{a+b}{2} \right)^{N+1} + P_N(x), \text{ gdzie } P_N(x) \text{ jest wielomianem stopnia co najwyżej } N.$$

2. Udowodnić, że jeżeli ciąg wielomianów interpolacyjnych (4.6) jest jednostajnie zbieżny do $f(x)$ na przedziale $\langle a; b \rangle$, to metoda kwadratur (4.5) jest zbieżna dla funkcji $f(x)$.

3. Uzasadnić poprawność przejścia do granicy w (4.16), gdy $x_4 \rightarrow \frac{a+b}{2}$.

4.2.2. Kwadratury złożone Newtona-Cotesa

Z powodów, które przedstawiliśmy w poprzednim punkcie, kwadratur Newtona-Cotesa wysokich rzędów używa się rzadko. Dla N równych 1 i 2 współczynniki C_r we wzorze (4.10), określającym błąd kwadratury, wynoszą odpowiednio $\frac{1}{12} (b-a)^3$, $\frac{1}{2880} (b-a)^5$. Błąd tych kwadratur jest zatem proporcjonalny do pewnej potęgi długości przedziału całkowania i przy dużych przedziałach całkowania stosowanie kwadratur nawet niskiego rzędu może nie zapewnić żadnej dokładności. W praktyce, kwadratury te stosuje się najczęściej w następujący sposób:

1) przedział całkowania $\langle a; b \rangle$ dzielimy na pewną liczbę podprzedziałów,

2) w każdym podprzedziale stosujemy kwadraturę niskiego rzędu i wyniki sumujemy.

Kwadraturę, która jest sumą kwadratur na podprzedziałach, nazywamy *kwadraturą złożoną*.

Rozpatrzmy kwadraturę, dla której współczynnik C_r we wzorze (4.10) jest proporcjonalny do k -tej potęgi $(b-a)$. Po podzieleniu przedziału $\langle a; b \rangle$ na p części współczynnik ten zmniejszy się p^k razy. Błąd kwadratury złożonej będzie zatem w przybliżeniu p^{k-1} razy mniejszy. Tak więc dla $k > 1$ możemy przez zwiększanie liczby podprzedziałów dowolnie zmniejszać błąd.

Największe znaczenie praktyczne mają wzory złożone trapezów i parabol w przypadku równomiernego podziału przedziału całkowania.

Omówimy najpierw *wzór złożony trapezów*. Przedział całkowania $\langle a; b \rangle$ dzielimy na m części o długości $h = \frac{b-a}{m}$. W każdym przedziale stosujemy wzór trapezów (4.11) i wyniki sumujemy. Otrzymamy

$$S(f) = \sum_{k=0}^{m-1} \frac{1}{2} h (f_k + f_{k+1}) = h \left(\frac{1}{2} f_0 + f_1 + \dots + f_{m-1} + \frac{1}{2} f_m \right) \quad (4.19)$$

gdzie $f_i = f(a + ih)$.

Niech $f \in C^2(\langle a; b \rangle)$. Na podstawie (4.13) błąd złożonego wzoru trapezów wynosi

$$E(f) = -\frac{h^3}{12} \sum_{k=0}^{m-1} f^{(2)}(\xi_k) = -\frac{(b-a)^3}{12m^2} \cdot \frac{1}{m} \sum_{k=0}^{m-1} f^{(2)}(\xi_k) \quad (4.20)$$

przy czym $\xi_k \in (a + kh; a + (k+1)h)$.

Czynnik $\frac{1}{m} \sum_{k=0}^{m-1} f^{(2)}(\xi_k)$ we wzorze (4.20) jest średnią arytmetyczną wartości drugiej pochodnej w punktach ξ_k . Jest oczywiste, że wartość ta znajduje się między najmniejszą i największą wartością $f^{(2)}(x)$ na przedziale $\langle a; b \rangle$. Ponieważ pochodna $f^{(2)}(x)$ jest ciągła na $\langle a; b \rangle$, więc istnieje punkt $\xi \in \langle a; b \rangle$ taki, że

$$\frac{1}{m} \sum_{k=0}^{m-1} f^{(2)}(\xi_k) = f^{(2)}(\xi)$$

Zatem błąd złożonego wzoru trapezów można przedstawić w postaci

$$E(f) = -\frac{(b-a)^3}{12m^2} f^{(2)}(\xi) \quad (4.21)$$

Omówimy teraz wzór złożony parabol. I w tym przypadku dzielimy przedział całkowania $\langle a; b \rangle$ na m równych części o długości $h = \frac{b-a}{m}$, przy czym m jest parzyste. W przedziałach $\langle a; a+2h \rangle, \dots, \langle a+(m-2)h; b \rangle$ o długości $2h$ stosujemy wzór parabol, wyniki sumujemy i otrzymujemy

$$\begin{aligned} S(f) &= \frac{h}{3} \sum_{k=1}^{m/2} (f_{2k-2} + 4f_{2k-1} + f_{2k}) = \\ &= \frac{h}{3} [f_0 + f_m + 2(f_2 + f_4 + \dots + f_{m-2}) + 4(f_1 + f_3 + \dots + f_{m-1})] \end{aligned} \quad (4.22)$$

Niech $f \in C^4(\langle a; b \rangle)$. Błąd złożonego wzoru parabol wynosi

$$E(f) = -\frac{h^5}{90} \sum_{k=1}^{m/2} f^{(4)}(\xi_k)$$

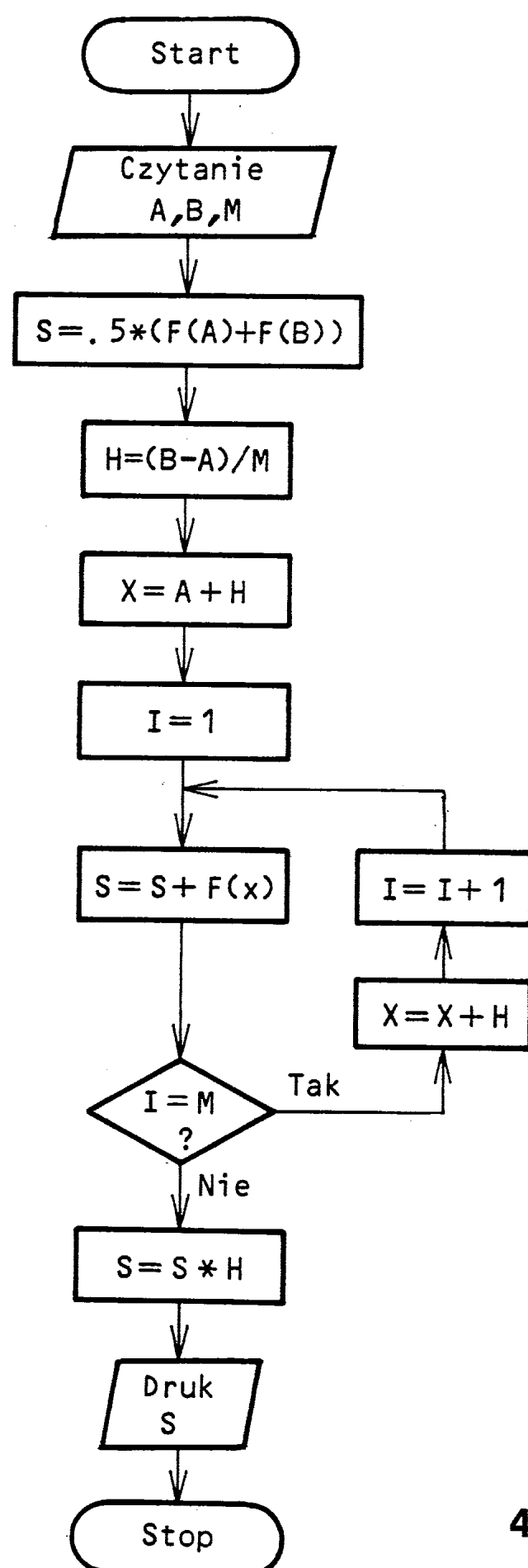
gdzie $\xi_k \in (a + 2(k-1)h; a + 2kh)$. Ponieważ pochodna $f^{(4)}(x)$ jest ciągła, więc istnieje punkt $\xi \in \langle a; b \rangle$ taki, że

$$\frac{2}{m} \sum_{k=1}^{m/2} f^{(4)}(\xi_k) = f^{(4)}(\xi)$$

Stąd mamy

$$E(f) = -\frac{(b-a)^5}{180m^4} f^{(4)}(\xi) \quad (4.23)$$

Na rysunku 4.3 jest przedstawiony schemat obliczania przybliżonej wartości całki na maszynie cyfrowej za pomocą wzoru trapezów dla ustalonego podziału (na M części) przedziału całkowania $\langle A; B \rangle$.



4.3

Ćwiczenia

1. Zapisać wzory złożone trapezów i parabol, gdy podział przedziału całkowania nie jest równomierny. Wyznaczyć oszacowanie tych wzorów.

2. Niech $S_i(f)$ oznacza wzór złożony parabol dla podziału przedziału całkowania na 2^i ($i = 0, 1, \dots$) równych części. Napisać program obliczania przybliżonej wartości całki w taki

sposób, aby $S_i(f)$ ($i = 0, 1, \dots$) były obliczane do momentu, gdy $|S_k(f) - S_{k-1}(f)| < \varepsilon$, gdzie ε jest z góryadaną wielkością. Obliczając $S_i(f)$ skorzystać z faktu, że węzły kwadratury $S_{i-1}(f)$ są węzłami $S_i(f)$.

3. Przybliżona wartość całki $\int_0^1 e^{-x^2} dx$ obliczona za pomocą złożonego wzoru parabol dla $h = 1/10$ wynosi 0,746825. Korzystając ze wzoru na błąd (4.23) wyznaczyć przedział, w którym znajduje się dokładna wartość całki.

Odpowiedzi. 3. (0,746818; 0,746830).

4.2.3. Metoda Romberga

Dzieląc przedział całkowania $\langle a; b \rangle$ na 2^i ($i = 0, 1, \dots$) równych części i wprowadzając oznaczenia

$$h_i = \frac{b-a}{2^i}, \quad x_{i,k} = a + kh_i, \quad f_{i,k} = f(x_{i,k})$$

możemy wzór złożony trapezów zapisać w postaci

$$T_{0,i} = h_i \left(\sum_{k=0}^{2^i} f_{i,k} - \frac{1}{2}(f(a) + f(b)) \right) \quad (4.24)$$

i wykazać, że błąd tego wzoru wynosi

$$I(f) - T_{0,i} = \sum_{j=1}^{\infty} C_{0,j} h_i^{2j} \quad (4.25)$$

przy czym współczynniki $C_{0,j}$ nie zależą od i (zależą one od a, b i f). Główny składnik tego błędu zawiera czynnik h_i^2 .

Określmy

$$T_{1,i} = T_{0,i+1} + \frac{T_{0,i+1} - T_{0,i}}{2^2 - 1} \quad (4.26)$$

Mamy wówczas

$$I(f) - T_{1,i} = \frac{1}{3} \sum_{j=1}^{\infty} C_{0,j} (2^{2j} h_{i+1}^{2j} - h_i^{2j})$$

Oznaczając $C_{1,j} = \frac{1}{3} \left(\frac{2^{2j}}{2^{2j}} - 1 \right) C_{0,j}$, mamy $C_{1,1} = 0$, a stąd

$$I(f) - T_{1,i} = \sum_{j=2}^{\infty} C_{1,j} h_i^{2j} \quad (4.27)$$

Jak widać, główny składnik tego wzoru zawiera czynnik h_i^4 . Można łatwo udowodnić, że $T_{1,i}$ jest wzorem złożonym parabol dla przedziału $\langle a; b \rangle$ podzielonego na 2^i części (patrz ćwic. 4).

Mając dwa przybliżenia całki uzyskane za pomocą złożonego wzoru trapezów ze wzoru (4.26) otrzymujemy trzecie przybliżenie, zapewne lepsze

— w (4.27) nie występuje h_i^2 , niemniej współczynniki rozwinięcia także uległy zmianie.

Opisany powyżej sposób poprawiania wyników jest szczególnym przypadkiem tzw. *ekstrapolacji Richardsona*, którą teraz krótko opiszemy. Rozpatrzmy metodę przybliżonego obliczania całki, w której obliczenia i błąd zależą od kroku h oraz niech $c_k h^k$ (c_k nie zależy od h) będzie głównym składnikiem błędu tej metody. Ekstrapolacja Richardsona polega na tym, że mając obliczone wartości s_1, s_2 dla dwóch różnych kroków h wyznaczamy takie trzecie przybliżenie s_3 , którego błąd nie zawiera już składnika z h^k . Nazwa ekstrapolacji pochodzi stąd, że jeśli $s_1 \neq s_2$, to s_3 nie należy do przedziału domkniętego o końcach s_1, s_2 (patrz ćwic. 2).

Rozpatrzmy teraz *metodę Romberga*, polegającą na tym, że obliczamy $T_{0,i}$ z krokiem h_i ($i = 0, 1, 2, \dots$) i eliminujemy kolejno pierwsze składniki w (4.25).

Określając ogólnie

$$T_{m,i} = T_{m-1,i+1} + \frac{T_{m-1,i+1} - T_{m-1,i}}{2^{2m} - 1} \quad (4.28)$$

łatwo można udowodnić, że

$$I(f) - T_{2,i} = \sum_{j=3}^{\infty} c_{2,j} h_i^{2j}$$

gdzie $c_{2j} = \frac{1}{15} c_{1,j} \left(\frac{16}{2^{2j}} - 1 \right)$. Można wykazać, że dla $m \geq 0$ główny składnik błędu $I(f) - T_{m,i}$ zawiera czynnik $h_i^{2(m+1)}$.

Z (4.28) wynika, że $T_{m,i}$ można przedstawić w postaci kombinacji liniowej $T_{0,i}, \dots, T_{0,m+i}$, a więc $T_{m,i}$ można zapisać w postaci

$$T_{m,i} = h \sum_{j=0}^{2^{m+i}} d_{m,j} f(a + jh) \quad (4.29)$$

gdzie $h = \frac{b-a}{2^{m+i}}$. Liczby $d_{m,j}$ są dodatnie dla wszystkich m, j . Kwadratury Romberga są zatem kwadraturami postaci (4.2) o dodatnich współczynnikach i węzłach równoodległych.

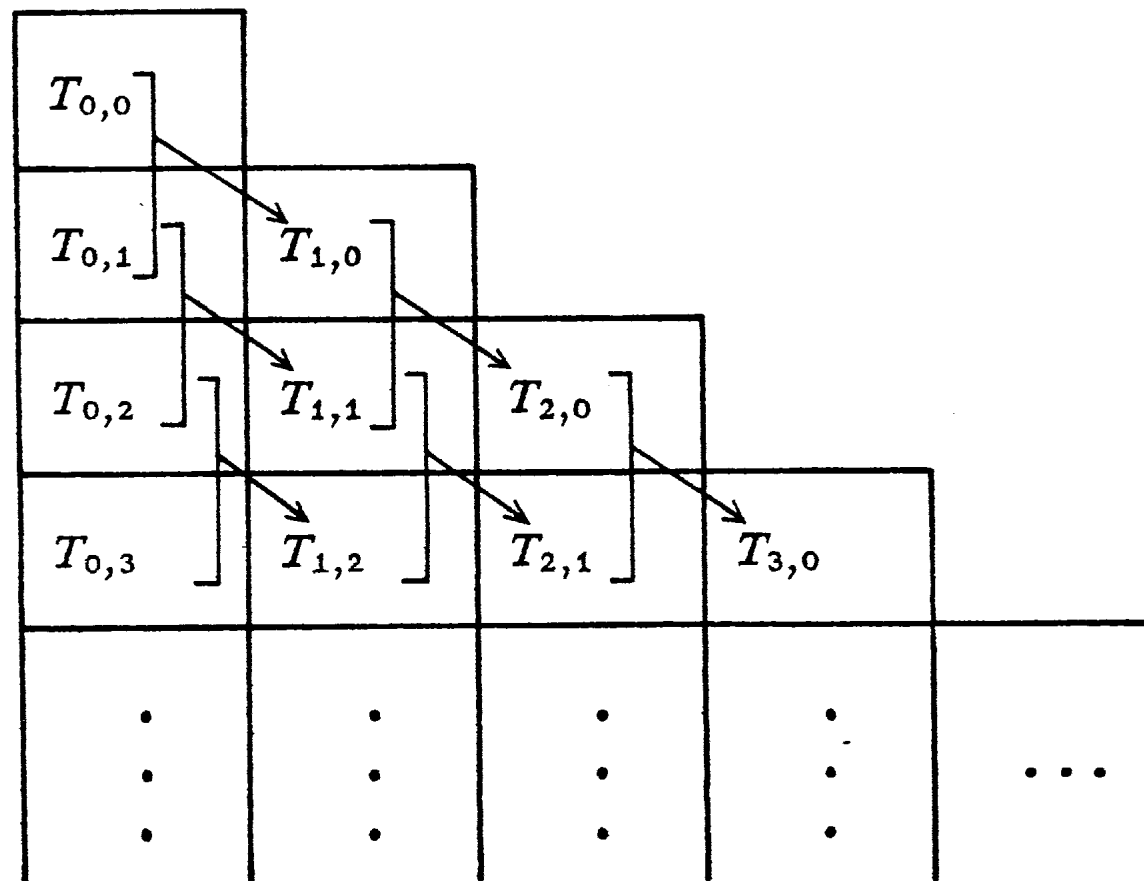
W przypadku gdy $f \in C^{(2m+2)}(\langle a; b \rangle)$, błąd kwadratury (4.29) można zapisać w postaci

$$I(f) - T_{m,i} = Ch^{2m+2} f^{(2m+2)}(\xi) \quad (4.30)$$

gdzie $\xi \in (a; b)$, C jest stałą, $h = \frac{b-a}{2^{m+i}}$.

Wielkości $T_{m,i}$ możemy zapisać w postaci nieskończonej tablicy (tabl. 4.1). Elementy pierwszej kolumny tej tablicy obliczamy ze złożonego wzoru trapezów (2.24). Po obliczeniu m pierwszych elementów z tej kolumny, ze

4/1



wzoru (4.28) możemy obliczyć m pierwszych wierszy tablicy. Można udowodnić, że ze zbieżności ciągu $(T_{0,m})$ do dokładnej wartości całki wynika zbieżność ciągu $(T_{m,0})$. W większości przypadków zbieżność tego ciągu jest szybsza niż zbieżność ciągu $(T_{0,m})$. Tak też będzie dla rozpatrywanej w p. 4.4 całki (4.61).

Dowód wzoru (4.30) jest podany w [7], dowody pozostałych zależności przedstawionych w tym punkcie można znaleźć w [50].

Opis standardowej procedury bibliotecznej. W standardowych procedurach bibliotecznych maksymalna liczba wierszy tablicy 4.1, które są wyznaczane, jest z góry ustalona. Niech R będzie tą liczbą. Jeżeli dwa kolejne elementy ciągu $(T_{m,0})$ spełniają nierówność

$$|T_{k,0} - T_{k-1,0}| \leq \varepsilon, \quad k < R \quad (4.31)$$

gdzie ε jest zadaną dokładnością, to obliczenia przerywa się wcześniej i wartość $T_{k,0}$ jest przyjmowana jako przybliżona wartość całki. Spełnienie warunku (4.31) nie gwarantuje, że moduł błędu przybliżenia $T_{k,0}$ nie przekroczy ε .

Do wykonania obliczeń wystarczy tablica $T(1 : R)$. Obliczenie wykonujemy wyznaczając kolejne wiersze tablicy 4.1. Wiersze te zapamiętujemy w tablicy T (patrz rys. 4.4). Mając wyznaczony i -ty wiersz, wiersz $i + 1$ -szy wyznaczamy ze wzoru

$$T(i+1) = T_{0,i}, \quad T(k) = T(k+1) + \frac{T(k+1) - T(k)}{4^{i-k+1} - 1}, \quad k = i, i-1, \dots, 1$$

Obliczając $T_{0,i}$ ($i = 1, 2, \dots$) korzystamy z faktu, że węzły kwadratury $T_{0,i-1}$ są węzłami kwadratury $T_{0,i}$

$$T_{0,i} = \frac{1}{2} T_{0,i-1} + h_i \sum_{k=1}^{2^i-1} f(a + h_i + (k-1)h_{i-1})$$

	T(1)	T(2)						T(i)	T(i+1)			T(R)
Wiersz 1-szy	$T_{0,0}$											
Wiersz 2-gi	$T_{1,0}$	$T_{0,1}$										
⋮												
Wiersz i-ty	$T_{i-1,0}$	$T_{i-2,1}$						$T_{0,i-1}$				
Wiersz (i+1)-szy	$T_{i,0}$	$T_{i-1,1}$						$T_{1,i-1}$	$T_{0,i}$			

4.4

Zakładamy, że w bloku głównym programu znajdują się definicje:

CONST max = ...;

TYPE wektor = ARRAY [1.. max] OF real;

oraz jest zdefiniowana funkcja FUN, która oblicza wartość funkcji podcałkowej w danym punkcie x

FUNCTION FUN (x: real): real;

Treść podprogramu:

```

PROCEDURE ROMBERG(a,b,eps:real; VAR r:integer; VAR y:real;
                  VAR T:wektor);
{ a,b - konce przedzialu calkowania;
  eps - zadana dokladnosc;
  r - wymiar pomocniczego wektora T, jest to maksymalna
    liczba wierszy tablicy 4.1,ktore zostana wyznaczone; po
    wykonaniu obliczen w r zapamietana jest liczba faktycznie
    wyznaczonych wierszy z tym, ze jezeli zadana dokladnosc
    nie zostala osiagnieta , to r =0;
  y - ostatnie obliczone T(k,o);
  T - pomocniczy wektor, po wykonaniu obliczen zapamietany
    jest ostatni obliczony wiersz tablicy 4.1 (patrz rys.4.4),
    z dokladnoscia do stalej multiplikatywnej (b-a);
  FUN - pomocnicz funkcja zdefiniowana w bloku zewnetrznym }

VAR i,j,i1,ii: integer;
    x,h,h1,s,q,e: real;
BEGIN
  T[1] := 0.5 * (FUN(a) + FUN(b));
  h := b - a;
  h1 := h;
  FOR i := 2 TO r DO
    BEGIN
      Y := T[1];
      s := 0.0;
      x := a + 0.5 * h1;
      FOR j := 1 TO jj DO
        BEGIN
          s := s + FUN(x);
          x := x + h1;
        
```



```

      END;
      h1 := 0.5 h1;
      T[i] := 0.5 * (T[i-1] + s/jj);
      q := 1.0;
      FOR j := 1 TO i-1 DO
        BEGIN
          q := 4 * q;
          i1 := i - j;
          T[i1] := T[i1+1] + (T[i1+1] - T[i1]) / (q-1.0);
          e := ABS(y - T[1]) * h;
          IF e <= eps THEN
            BEGIN
              r := i-1;
              y := h * T[1];
              EXIT;
            END;
          jj := jj + j;
        END;
      r := 0;
      y := h * T[1];
    END;
  END;

```

Ćwiczenia

1. Wykazać, że wzór (4.26) określający $T_{1,i}$ jest złożonym wzorem parabol dla podziału przedziału całkowania na 2^i części.

2. W ostatnim punkcie opisaliśmy krótko metodę ekstrapolacji Richardsona. Niech h_1, h_2 ($h_2 < h_1$) oznaczają krok całkowania, z jakim zostały obliczone wartości s_1, s_2 .

a) Wykazać, że

$$s_3 = s_2 - \frac{h_2^k}{h_1^k - h_2^k} (s_1 - s_2)$$

Udowodnić, że: 1) jeżeli $s_1 \neq s_2$, to s_3 nie należy do przedziału domkniętego o końcach s_1, s_2 ; 2) jeżeli $s_1 = s_2$, to $s_3 = s_1 = s_2$.

b) Niech s_1, s_2 oznaczają przybliżone wartości całki $\int_1^2 \frac{1}{x} dx$, otrzymane za pomocą wzoru parabol dla $h = 1/2, 1/4$. Obliczyć s_1, s_2 , a następnie s_3 . Czy s_3 lepiej przybliża dokładną wartość całki?

Odpowiedzi. 2b. $s_1 = 0,694444$; $s_2 = 0,693254$; $s_3 = 0,693214$; tak.

4.3. KWADRATURY GAUSSA I KWADRATURY ZŁOŻONE GAUSSA

4.3.1. Kwadratury Gaussa

W punkcie 4.2.1 rozpatrywaliśmy kwadratury postaci

$$S(f) = \sum_{k=0}^N A_k f(x_k) \quad (4.32)$$

z ustalonymi węzłami i pokazaliśmy, że jeżeli współczynniki A_k są określone wzorem

$$A_k = \int_a^b p(x) \Phi_k(x) dx \quad (4.33)$$

gdzie $p(x)$ jest wagą, a

$$\Phi_k(x) = \frac{(x - x_0) \dots (x - x_{k-1})(x - x_{k+1}) \dots (x - x_N)}{(x_k - x_0) \dots (x_k - x_{k-1})(x_k - x_{k+1}) \dots (x_k - x_N)}$$

to rząd kwadratury (4.32) wynosi co najmniej $(N + 1)$. W niniejszym punkcie przy ustalonej wadze $p(x)$ i liczbie N rozpatrzmy problem wyboru węzłów i współczynników tak, aby rząd kwadratury był jak najwyższy. Taką kwadraturę nazywamy *kwadraturą Gaussa*. Ponieważ mamy $2(N + 1)$ stałych do wyznaczenia, należy przypuszczać, że można je tak dobrać, aby kwadratura była dokładna dla wielomianów stopnia mniejszego niż $2(N + 1)$. Pokażemy, że taka kwadratura istnieje, jest jedna i jej rząd wynosi $2(N + 1)$. W punkcie 4.2.1 udowodniliśmy, że współczynniki każdej kwadratury rzędu co najmniej $(N + 1)$ są określone wzorami postaci (4.33). Tak więc problem sprowadza się do odpowiedniego wyboru węzłów x_k w kwadraturze (4.32).

Omówimy najpierw pewne własności wielomianów ortogonalnych, które są potrzebne przy wyznaczaniu kwadratur Gaussa. Niech ciąg wielomianów,

$$(P_n(x)) \equiv P_0(x), P_1(x), \dots, P_N(x), \dots \quad (4.34)$$

gdzie stopień $P_n(x)$ jest równy n , będzie ciągiem wielomianów ortogonalnych na przedziale $\langle a; b \rangle$ z wagą $p(x)$. Mamy więc

$$(P_r, P_s) = \int_a^b p(x) P_r(x) P_s(x) dx = 0 \quad \text{dla} \quad r \neq s$$

Tw. 1. *Wielomiany ortogonalne (4.34) mają tylko pierwiastki rzeczywiste jednokrotne, leżące w przedziale $(a; b)$.*

DOWÓD. Przypuśćmy, że pierwiastki wielomianu $P_k(x)$ nie spełniają powyższych warunków. Niech $\xi_1, \dots, \xi_m$ będą rzeczywistymi, różnymi, o nieparzystej krotności pierwiastkami wielomianu $P_k(x)$ z przedziału $(a; b)$. Mamy $m < k$. Oznaczmy $W(x) = (x - \xi_1) \dots (x - \xi_m)$ i przedstawmy $W(x)$ w postaci kombinacji liniowej wielomianów $P_0(x), \dots, P_m(x)$

$$W(x) = a_0 P_0(x) + \dots + a_m P_m(x)$$

Iloczyn skalarny wielomianów ortogonalnych

$$(W, P_k) = \sum_{i=0}^m a_i (P_i, P_k) = 0$$

Z drugiej strony $W(x) P_k(x) \geq 0$ na przedziale $\langle a; b \rangle$, a więc $(W, P_k) > 0$. Oznacza to, że nasze przypuszczenie okazało się fałszywe.

Przejdziemy teraz do wyznaczenia kwadratur Gaussa podając na wstępie kilka twierdzeń.

Tw. 2. 1) *Nie istnieje kwadratura postaci (4.32) rzędu wyższego niż $2(N + 1)$.*

2) Kwadratura (4.32), o współczynnikach A_k określonych wzorem (4.33), jest rzędu $2(N+1)$ wtedy i tylko wtedy, gdy x_k są pierwiastkami wielomianu $P_{N+1}(x)$ z ciągu (4.34).

DOWÓD. 1) Rozpatrzmy wielomian $W(x) = \omega_{N+1}^2(x)$ stopnia $2(N+1)$, gdzie $\omega_{N+1}(x) = (x - x_0) \dots (x - x_N)$. Mamy

$$I(W) = \int_a^b p(x) W(x) dx > 0, \quad S(W) = \sum_{k=0}^N A_k W(x_k) = 0$$

Istnieją zatem wielomiany stopnia $2(N+1)$, dla których kwadratura (4.32) nie jest dokładna, a to właśnie należało udowodnić.

2) Załóżmy, że kwadratura (4.32), (4.33) jest rzędu $2(N+1)$. Niech $W(x)$ będzie dowolnym wielomianem stopnia niższego niż $(N+1)$. Wielomian $R(x) = W(x) \omega_{N+1}(x)$ jest wielomianem stopnia mniejszego niż $2(N+1)$. Dla wielomianu $R(x)$ kwadratura (4.32) jest dokładna z założenia. Mamy zatem

$$\int_a^b p(x) W(x) \omega_{N+1}(x) dx = \sum_{k=0}^N A_k R(x_k) \quad (4.35)$$

Węzły x_k należy wobec tego tak wybrać, aby powyższy warunek był spełniony dla dowolnego wielomianu $W(x)$ stopnia niższego niż $(N+1)$. Przedstawmy wielomian $\omega_{N+1}(x)$ w postaci

$$\omega_{N+1}(x) = a_0 P_0(x) + \dots + a_{N+1} P_{N+1}(x), \quad a_{N+1} \neq 0 \quad (4.36)$$

i przyjmijmy

$$W(x) = a_0 P_0(x) + \dots + a_N P_N(x) \quad (4.37)$$

Po podstawieniu rozwinięć (4.36) i (4.37) do (4.35), otrzymamy

$$(W, \omega_{N+1}) = a_0 (P_0, P_0) + \dots + a_N (P_N, P_N) = 0$$

i stąd otrzymujemy $a_i = 0$, $i = 0, 1, \dots, N$. Na podstawie (4.36) mamy

$$\omega_{N+1}(x) = a_{N+1} P_{N+1}(x) \quad (4.38)$$

Udowodniliśmy zatem, że węzły x_k są pierwiastkami wielomianu $P_{N+1}(x)$.

Przeprowadzenie dowodu w drugą stronę pozostawiamy Czytelnikowi.

Tw. 3. Wszystkie współczynniki A_k w kwadraturach Gaussa są dodatnie.

DOWÓD. Rozpatrzmy wielomiany stopnia $2N$

$$R_i(x) = [(x - x_0) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_N)]^2, \quad i = 0, 1, \dots, N$$

Mamy

$$0 < \int_a^b p(x) R_i(x) dx = \sum_{k=0}^N A_k R_i(x_k) = A_i R_i(x_i)$$

stąd zaś wynika, że współczynniki A_k są dodatnie.

Zatem w odróżnieniu od kwadratur Newtona-Cotesa współczynniki kwadratur Gaussa są dodatnie dla dowolnego N i na podstawie twierdzenia i uwagi z p. 4.1.2 stwierdzamy, że metoda kwadratur Gaussa jest zbieżna dla każdej funkcji ciągłej na $\langle a; b \rangle$.

Kwadratury Gaussa są dokładne dla wielomianów stopnia $2N+1$. Niech $f \in C^{2N+2}(\langle a; b \rangle)$ i niech $L_{2N+1}(x)$ będzie wielomianem interpolacyj-

nym dla funkcji $f(x)$ z węzłami $x_0, x_1, \dots, x_N, \tilde{x}_0, \tilde{x}_1, \dots, \tilde{x}_N$, gdzie $x_0, x_1, \dots, x_N$ są pierwiastkami wielomianu $P_{N+1}(x)$; $\tilde{x}_i \neq x_k, \tilde{x}_i \neq \tilde{x}_j, \tilde{x}_i \in \langle a; b \rangle$, $i, k, j = 0, 1, \dots, N$. Mamy wówczas

$$f(x) - L_{2N+1}(x) = \frac{1}{(2N+2)!} (x - x_0) \dots (x - x_N) (x - \tilde{x}_0) \dots (x - \tilde{x}_N) f^{(2N+2)}(\xi) \quad (4.39)$$

gdzie $\xi \in (a; b)$. Przechodząc w (4.39) do granicy dla $\tilde{x}_i \rightarrow x_i (i = 0, 1, \dots, N)$, otrzymamy

$$f(x) - \tilde{L}_{2N+1}(x) = \frac{1}{(2N+2)!} \omega_{N+1}^2(x) f^{(2N+2)}(\bar{\xi}), \quad \bar{\xi} \in \langle a; b \rangle \quad (4.40)$$

Błąd kwadratury Gaussa wynosi

$$\begin{aligned} E(f) &= I(f) - S(f) = I(f) + S(\tilde{L}_{2N+1}) = \\ &= \frac{1}{(2N+2)!} \int_a^b p(x) \omega_{N+1}^2(x) f^{(2N+2)}(\xi) dx \end{aligned} \quad (4.41)$$

Po zastosowaniu uogólnionego twierdzenia o wartości średniej dla całek, otrzymamy

$$E(f) = \frac{1}{(2N+2)!} f^{(2N+2)}(\xi_1) \cdot \int_a^b p(x) \omega_{N+1}^2(x) dx, \quad \xi_1 \in (a; b) \quad (4.42)$$

Kwadratur Gaussa wysokiego rzędu, podobnie jak kwadratur Newtona-Cotesa, używa się rzadko i powody tego są takie same (trudności w szacowaniu pochodnych wysokiego rzędu oraz to, że pochodne te często rosną nieograniczenie; patrz p. 4.2.1). W praktyce najczęściej stosuje się kwadratury złożone Gaussa niskiego rzędu na każdym podprzedziale. Dokładniej omówimy je w p. 4.3.2.

Kwadratury Gaussa zależą od wagi $p(x)$. Wyprowadzimy obecnie wzory (oddzielnie dla przedziału skończonego i nieskończonego) z wagami, które mają duże znaczenie zarówno teoretyczne jak i praktyczne. Wzory na współczynniki i błędy podamy bez wyprowadzenia; wyprowadzenia te można znaleźć np. w [41], [58].

Kwadratury dla przedziału skończonego. Kwadraturę najwyższego rzędu z wagą $p(x) = 1$ nazywamy *kwadraturą Gaussa-Legendre'a*.

Niech $\langle a; b \rangle = \langle -1; 1 \rangle$. Dla wagi $p(x) = 1$ ciąg wielomianów ortogonalnych tworzą *wielomiany Legendre'a*

$$P_n(x) = \frac{1}{2^n \cdot n!} \frac{d^n}{dx^n} (x^2 - 1)^n \quad (4.43)$$

Współczynniki i błąd kwadratury Gaussa-Legendre'a wyrażają się w następujący sposób

$$A_k = -\frac{2}{(N+2)P_{N+2}(x_k)P'_{N+1}(x_k)} \quad (4.44)$$

$$E(f) = \frac{2^{2N+3}((N+1)!)^4}{(2N+3)((2N+2)!)^3} f^{(2N+3)}(\xi), \quad -1 < \xi < 1$$

gdzie x_k są zerami wielomianu $P_{N+1}(x)$.

W przypadku dowolnego przedziału całkowania $\langle a; b \rangle$ powyższe wzory stosujemy w następujący sposób. W obliczanej całce dokonujemy liniowej zamiany zmiennej całkowania

$$t = \frac{a+b}{2} + \frac{b-a}{2}x, \quad \int_a^b f(t) dt = \frac{b-a}{2} \int_{-1}^1 g(x) dx$$

gdzie $g(x) = f\left(\frac{a+b}{2} + \frac{b-a}{2}x\right)$, i do tak przekształconej całki stosujemy podane wyżej wzory. Ostatecznie otrzymujemy

$$\int_a^b f(t) dt \approx S(f) = \frac{b-a}{2} \sum_{k=0}^N A_k f(t_k) \quad (4.45)$$

gdzie $t_k = \frac{a+b}{2} + \frac{b-a}{2}x_k$, x_k — zera wielomianu $P_{N+1}(x)$, określonego wzorem (4.43), A_k — współczynniki określone wzorem (4.44).

Po dokonaniu zamiany zmiennej błąd kwadratury wyraża się wzorem

$$E(f) = \frac{(b-a)^{2N+3}((N+1)!)^4}{(2N+3)((2N+2)!)^3} f^{(2N+2)}(\xi), \quad a < \xi < b \quad (4.46)$$

Kwadratur wysokiego rzędu, o czym już mówiliśmy, używa się rzadko. W praktyce stosuje się kwadratury złożone ze wzorami niskiego rzędu na każdym podprzedziale. Tablica 4/2 podaje węzły i współczynniki kwadratur Gaussa-Legendre'a dla wartości $N = 1, 2, 3, 4$ ($\langle a; b \rangle = \langle -1; 1 \rangle$).

Wagę $p(x) = 1$ wybieramy wtedy, gdy funkcja podcałkowa jest funkcją dostatecznie gładką na przedziale domkniętym $\langle a; b \rangle$. Omówimy teraz zastosowanie kwadratur Gaussa do obliczania całek na przedziale skończonym, w przypadku gdy funkcja podcałkowa posiada osobliwości. Osobliwości te mogą polegać na tym, że funkcja podcałkowa jest nieograniczona na przedziale $\langle a; b \rangle$, jak i na tym, że dopiero jej pochodne są nieograniczone. Taką funkcję podcałkową przedstawiamy w postaci iloczynu $p(x)f(x)$, gdzie $p(x)$ zawiera wszelkie osobliwości funkcji podcałkowej, a $f(x)$ jest już funkcją dostatecznie gładką na przedziale $\langle a; b \rangle$. Przyjęcie funkcji $p(x)$ za funkcję

4/2

N	k	Węzły x_k	Współczynniki A_k
1	0; 1	$\mp 0.577350 = \mp 1/\sqrt{3}$	1
2	0; 2 1	$\mp 0.774597 = \mp \sqrt{3/5}$ 0	5/9 8/9
3	0; 3 1; 2	∓ 0.861136 ∓ 0.339981	0.347855 0.652145
4	0; 4 1; 3 2	∓ 0.906180 ∓ 0.538469 0	0.236927 0.478629 0.568889

wagową umożliwia usunięcie części osobliwej zarówno z sumy $S(f)$, jak i z błędu $E(f)$. Zakładamy, że $p(x)$ jest nieujemną, całkowalną funkcją na przedziale $\langle a; b \rangle$. Wadą takiego podejścia jest to, że dla danej wagi $p(x)$ należy zbudować wielomiany ortogonalne i wyznaczyć ich pierwiastki. Dla wielu wag najczęściej stosowanych w praktyce takie obliczenia zostały już wykonane. Nasze rozważania ograniczymy do przypadku, gdy osobliwości funkcji podcałkowej występują na końcach przedziału całkowania.

Niech $\langle a; b \rangle = \langle -1; 1 \rangle$. Rozpatrzmy najpierw kwadratury z wagą *Jacobiego* $p(x) = (1-x)^\alpha(1+x)^\beta$, gdzie $\alpha, \beta > -1$, zwane *kwadraturami Gaussa-Jacobiego*. Wielomianami ortogonalnymi z tą wagą są *wielomiany Jacobiego*

$$J_n(x; \alpha, \beta) = \frac{(-1)^n}{2^n \cdot n!} (1-x)^{-\alpha} (1+x)^{-\beta} \frac{d^n}{dx^n} [(1-x)^{\alpha+n} (1+x)^{\beta+n}] \quad (4.47)$$

Współczynniki i błąd kwadratury Gaussa-Jacobiego wyrażają się wzorami

$$A_k = - \frac{(2N + \alpha + \beta + 4) \cdot \Gamma(N + \alpha + 2) \cdot \Gamma(N + \beta + 2) \cdot 2^{\alpha + \beta}}{(N + \alpha + \beta + 2) \cdot \Gamma(N + \alpha + \beta + 2) \cdot J_{N+1}(x_k; \alpha, \beta) \cdot J_{N+2}(x_k; \alpha, \beta) (N + 2)!} \quad (4.48)$$

$$E(f) =$$

$$= \frac{\Gamma(N + \alpha + 2) \cdot \Gamma(N + \beta + 2) \cdot \Gamma(N + \alpha + \beta + 2) \cdot (N + 1)! \cdot 2^{2N + \alpha + \beta + 3}}{(2N + \alpha + \beta + 3) [\Gamma(2N + \alpha + \beta + 3)]^2 (2N + 2)!} f^{(2N+2)}(\xi)$$

gdzie $-1 \leq \xi \leq 1$, x_k są zerami wielomianu $J_{N+1}(x; \alpha, \beta)$, a Γ — oznacza funkcję gamma Eulera¹⁾.

Kwadratura Gaussa-Jacobiego, gdy $\langle a; b \rangle = \langle -1; 1 \rangle$, ma więc postać

$$\int_{-1}^1 (1-x)^\alpha (1+x)^\beta f(x) dx \approx S(f) = \sum_{k=0}^N A_k f(x_k) \quad (4.49)$$

¹⁾ $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt, \quad x > 0.$

gdzie x_k są zerami wielomianu $J_{N+1}(x; \alpha, \beta)$, a współczynniki A_k wyrażają się wzorem (4.48).

Podane wzory zależą od dwóch parametrów $\alpha, \beta > -1$. Dla $\alpha = \beta = 0$ podane wyżej wzory przechodzą we wzory Gaussa-Legendre'a, przyjmując zaś za α, β konkretne wartości otrzymamy wzory przybliżonego całkowania, uwzględniające ewentualne osobliwości postaci $(1-x)^\alpha(1+x)^\beta$ funkcji podcałkowej w punktach $x = \pm 1$.

Rozpatrzmy jeszcze szczególny przypadek, gdy $\alpha = \beta = -1/2$, czyli gdy $p(x) = (1-x^2)^{-1/2}$. Kwadratury z taką wagą nazywamy *kwadraturami Gaussa-Czebyszewa*. Ciąg wielomianów ortogonalnych tworzą w tym przypadku *wielomiany Czebyszewa pierwszego rodzaju* postaci

$$T_n(x) = \cos(n \arccos x)$$

a współczynniki i węzły tych kwadratur wynoszą odpowiednio

$$A_k = \frac{\pi}{N+1}, \quad x_k = \cos \frac{(2k+1)\pi}{2N+2} \quad (4.50)$$

Widzimy, że wszystkie współczynniki tej kwadratury są równe.

W przypadku dowolnego przedziału całkowania $\langle a; b \rangle$ przedział ten za pomocą liniowego przekształcenia sprowadzamy do przedziału $\langle -1; 1 \rangle$.

Przykład 1. Obliczyć całkę $\int_0^1 \frac{2(1+x)}{\sqrt{x(1-x)}} dx$ za pomocą kwadratury Gaussa-Czebyszewa

dla $N = 2$.

Wprowadzając nową zmienną $y = -1 + 2x$, otrzymujemy całkę

$$\int_{-1}^1 \frac{3+y}{\sqrt{1-y^2}} dy$$

Na podstawie (4.50) wyznaczamy węzły i współczynniki. Mamy

$$x_0 = \cos \frac{\pi}{6}, \quad x_1 = \cos \frac{3\pi}{6}, \quad x_2 = \cos \frac{5\pi}{6}, \quad A_0 = A_1 = A_2 = \frac{\pi}{3}$$

Zatem zgodnie ze wzorem (4.49)

$$\begin{aligned} \int_{-1}^1 \frac{2(1+x)}{\sqrt{x(1-x)}} dx &= \int_{-1}^1 \frac{3+y}{\sqrt{1-y^2}} dy \approx \frac{\pi}{3} \left[\left(3 + \cos \frac{\pi}{6} \right) + \left(3 + \cos \frac{3\pi}{6} \right) + \right. \\ &\quad \left. + \left(3 + \cos \frac{5\pi}{6} \right) \right] \approx 9,424778 \end{aligned}$$

a dokładna wartość tej całki wynosi 3π . W (4.48) występuje $f^{(6)}(\xi)$, a ponieważ $f(x) = 1+x$, a więc $E(f) = 0$. Pamiętając o błędach zaokrągleń stwierdzamy, że otrzymany wynik jest dokładnie taki, jaki powinien być.

Kwadratury Gaussa-Czebyszewa są szczególnym przypadkiem *kwadratur Czebyszewa* (patrz ćwicz. 2), tzn. kwadratur typu (4.32), w których

wszystkie współczynniki A_k są równe. Gdy np. wartości funkcji $f(x)$ są wynikami pomiarów i są one obarczone błędami, które są duże w porównaniu z błędem metody, wówczas może być wskazane zastosowanie kwadratury Czebyszewa. Można łatwo wykazać (patrz ćwic. 1), że jeżeli współczynniki A_k są równe, to oczekiwany błąd sumy $S(f)$ jest najmniejszy.

Kwadratury dla przedziału nieskończonego. Omówimy teraz zastosowanie kwadratur Gaussa do obliczania całek na przedziale nieskończonym, przy założeniu zbieżności tych całek. Rozpatrzmy przypadki, gdy przedział całkowania jest: 1) jednostronnie nieskończony, 2) obustronnie nieskończony.

1) Bez zmniejszenia ogólności rozważań możemy założyć, że tym przedziałem jest $\langle 0; \infty \rangle$. Rozpatrzmy wagę $p(x) = e^{-x}$, dla której ciągiem wielomianów ortogonalnych na przedziale $\langle 0; \infty \rangle$ są *wielomiany Laguerre'a*

$$L_n(x) = (-1)^n e^x \frac{d^n}{dx^n} (x^n e^{-x}) \quad (4.51)$$

Waga $p(x) = e^{-x}$ zapewnia zbieżność całki $\int_0^\infty p(x) W(x) dx$, gdy $W(x)$ jest wielomianem dowolnego stopnia, a więc całka (4.33) określająca współczynniki A_k jest zbieżna. Współczynniki A_k i błąd rozważanej kwadratury wyrażają się wobec tego wzorami

$$A_k = \frac{((N+1)!)^2}{L_{N+1}(x_k) L_{N+2}(x_k)}, \quad E(f) = \frac{((N+1)!)^2}{(2N+2)!} f^{(2N+2)}(\eta) \quad (4.52)$$

gdzie x_k są zerami wielomianu Laguerre'a $L_{N+1}(x)$, $\eta \in (0; \infty)$.

Wzór przybliżonego całkowania

$$\int_0^\infty e^{-x} f(x) dx \approx S(f) = \sum_{k=0}^N A_k f(x_k) \quad (4.53)$$

w którym węzły x_k są zerami wielomianu Laguerre'a $L_{N+1}(x)$, a współczynniki A_k wyrażają się wzorem (4.52), nazywamy *kwadraturą Gaussa-Laguerre'a*.

2) Na przedziale $(-\infty; +\infty)$ rozpatrzmy wagę $p(x) = e^{-x^2}$, dla której ciąg wielomianów ortogonalnych tworzą *wielomiany Hermite'a*

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2} \quad (4.54)$$

Podobnie jak poprzednio, waga $p(x) = e^{-x^2}$ zapewnia zbieżność całki (4.33) określającej współczynniki A_k . Mamy teraz

$$A_k = \frac{2^{N+2}(N+1)!}{H'_{N+1}(x_k) H_{N+2}(x_k)}, \quad E(f) = \frac{(N+1)! \sqrt{\pi}}{2^{N+1}(2N+2)!} f^{(2N+2)}(\eta) \quad (4.55)$$

gdzie x_k są zerami wielomianu Hermite'a $H_{N+1}(x)$, $\eta \in (-\infty; +\infty)$. Wzór przybliżonego całkowania

$$\int_{-\infty}^{+\infty} e^{-x^2} f(x) dx \approx S(f) = \sum_{k=0}^N A_k f(x_k)$$

(4.56)

w którym węzły x_k są zerami wielomianu $H_{N+1}(x)$, a współczynniki A_k są określone wzorem (4.55), nazywamy *kwadraturą Gaussa-Hermite’a*.
W tablicy 4/3 podano węzły i współczynniki dwóch ostatnio rozpatrzonych kwadratur, gdy $N = 1, 2, 3, 4$.

4/3

N	Kwadratury Gaussa-Laguerre’a			Kwadratury Gaussa-Hermite’a		
	k	Węzły x_k	Współczynniki A_k	k	Węzły x_k	Współczynniki A_k
1	0	0.585786	0.853553	0; 1	∓ 0.707107	0.886227
	1	3.414214	0.146447			
2	0	0.415775	0.711093	0; 2 1	∓ 1.224745	0.295409 1.181636
	1	2.294280	0.278518		0	
	2	6.289945	0.010389			
3	0	0.322548	0.603154	0; 3 1; 2	∓ 1.650680	0.081313 0.804914
	1	1.745761	0.357419		∓ 0.524648	
	2	4.536620	0.038888			
	3	9.395071	0.000539			
4	0	0.263560	0.521756	0; 4 1; 3 2	∓ 2.020183	0.019953 0.393619 0.945309
	1	1.413403	0.398667		∓ 0.958572	
	2	3.596426	0.075942		0	
	3	7.085810	0.003612			
	4	12.640801	0.000032			

Przykład 2. Obliczyć całkę $\int_0^{\infty} e^{-x} x^8 dx$ stosując kwadraturę Gaussa-Laguerre’a dla $N = 2$.
Posługując się tablicą 4.3, otrzymujemy
$$\int_0^{\infty} e^{-x} x^8 dx \approx 0,711093 \cdot (0,415775)^8 + 0,278518 \cdot (2,294280)^8 + 0,010389 \cdot (6,289945)^8 \approx 25667$$

Dokładna wartość całki wynosi $8! = 40320$.
Na podstawie wzoru (4.52) mamy

$$E(f) = 1008 \eta^2, \quad \eta \in (0; +\infty)$$

Tak znaczna różnica między przybliżoną a dokładną wartością całki jest spowodowana tym, że $f^{(6)}(x)$ nie jest ograniczona na przedziale całkowania $\langle 0; +\infty \rangle$.
Zauważmy, że w powyższym przykładzie dla $N = 4$ otrzymalibyśmy wynik dokładny, gdyż $f^{(10)}(x) \equiv 0$.

Z rozważonego przykładu wynika, że omówionych kwadratur dla nieskończonych przedziałów całkowania raczej nie należy stosować, jeśli pochodna we wzorze na błąd nie jest ograniczona na przedziale całkowania. W takich przypadkach zaleca się wykonanie obliczeń w następujący sposób:
— w przedziale całkowania wydziela się przedział skończony $\langle a; b \rangle$ taki,

aby wartość bezwzględna całki w pozostałej części przedziału była mniejsza od pewnej stałej dodatniej ε ,

— do obliczenia całki na przedziale $\langle a; b \rangle$ stosuje się wzór przeznaczony do obliczeń na przedziale skończonym.

Ćwiczenia

1. Niech będą dane wielkości $f_i (i = 1, \dots, n)$ obarczone błędem ε_i , przy czym $|\varepsilon_i| \leq \varepsilon$. Rozpatrzmy zachowanie się błędu sumy ważonej

$$S = \sum_{i=1}^n A_i f_i$$

przy warunku $\sum_{i=1}^n A_i = A > 0$, gdzie A jest ustaloną liczbą.

Zakładamy, że błędy ε_i są niezależnymi zmiennymi losowymi o tym samym rozkładzie równomiernym na przedziale $\langle -\varepsilon; \varepsilon \rangle$.

a) Pokazać, że rozkłady zmiennej losowej S , gdy $n \rightarrow \infty$, dążą do rozkładu normalnego

$$N(0, \sigma), \text{ gdzie } \sigma^2 = \frac{2}{3} \varepsilon^2 \sum_{i=1}^n A_i^2.$$

b) Udowodnić, że σ^2 osiąga minimum przy warunku $\sum_{i=1}^n A_i = A > 0$, jeżeli wszystkie współczynniki A_i są równe (obranie A_i w ten sposób, że σ^2 jest najmniejsze, prowadzi do statystycznie dobrego zachowania się błędu sumy S).

W s k a z ó w k a. Zastosować metodę mnożników Lagrange'a, tzn. zminimalizować

$$\frac{1}{2} \sum_{i=1}^n A_i^2 + \lambda \left(\sum_{i=1}^n A_i - A \right).$$

2. Napisać warunki, jakie powinny spełniać węzły x_i dla kwadratury Czebyszewa

$$\int_a^b p(x) f(x) dx \approx S(f) = A \sum_{k=0}^N f(x_k), \quad A = \frac{1}{N+1} \int_a^b p(x) dx$$

aby była ona dokładna dla wszystkich wielomianów stopnia nie wyższego niż $(N+1)$. Wyznaczyć te węzły, przyjmując:

- przedział całkowania $\langle -1; 1 \rangle$, wagę $p(x) \equiv 1$, $N = 1, 2$;
- przedział całkowania $(-\infty; +\infty)$, wagę $p(x) = e^{-x^2}$, $N = 1, 2$.

W s k a z ó w k a. Skorzystać z faktu, że wzór ma być dokładny dla wielomianów $t, t^2, \dots, t^{N+1}$.

3. Na podstawie tablicy 4.2 i wyników z poprzedniego ćwiczenia napisać wzory złożone: Gaussa-Legendre'a dla $N = 2$, Czebyszewa, dla $N = 1$ i 2 oraz wagi $p(x) \equiv 1$.

4. Obliczyć przybliżoną wartość całki $\int_{-\infty}^{+\infty} e^{-x^2} \cos x dx$ za pomocą wzoru Gaussa-Hermite'a (4.57) dla $N = 4$, a następnie oszacować błąd tego przybliżenia.

Odpowiedzi. 2. a) Dla $N = 1$ węzły $x_{1,2} = \mp \frac{1}{\sqrt{3}}$, dla $N = 2$ węzły $x_{1,3} = \mp \frac{1}{\sqrt{2}}$, $x_2 = 0$;

b) dla $N = 1$ węzły $x_{1,2} = \mp \frac{1}{\sqrt{2}}$, dla $N = 2$ węzły $x_{1,3} = \mp \frac{\sqrt{3}}{2}$, $x_2 = 0$.

4. $S(f) = 1,380390$; $|E(f)| \leq 2 \cdot 10^{-6}$.

4.3.2. Kwadratury złożone Gaussa

Niech przedział całkowania $\langle a; b \rangle$ będzie skończony. Mówiliśmy już w poprzednim punkcie, że kwadratur Gaussa wysokiego rzędu używa się rzadko, w praktyce zaś stosuje się kwadratury złożone ze wzorami niskiego rzędu w każdym podprzedziale.

Rozpatrzmy kwadratury Gaussa-Legendre'a. Podzielimy przedział całkowania $\langle a; b \rangle$ na n równych części. Jeżeli do obliczenia całki zastosujemy kwadraturę złożoną ze wzorami rzędu $(2N + 2)$ na każdym podprzedziale, to na podstawie (4.46) błąd otrzymanego wyniku, gdy $f \in C^{2N+2}(\langle a; b \rangle)$, będzie wynosił

$$E(f) = n \left(\frac{b-a}{n} \right)^{2N+3} \frac{((N+1)!)^4}{(2N+3)((2N+2)!)^3} f^{(2N+2)}(\xi_1) \quad (4.57)$$

$\xi_1 \in (a; b)$

Zwiększając liczbę podprzedziałów możemy zatem dowolnie zmniejszyć błąd.

Dla $N = 1$; $n = \frac{m}{2}$ (m — parzyste), na podstawie tablicy 4.2 otrzymujemy następujący wzór złożony

$$\int_a^b f(x) dx = \sum_{j=0}^{\frac{m}{2}-1} \int_{a+2hj}^{a+2h(j+1)} f(x) dx$$

$$S(f) = h \sum_{j=0}^{\frac{m}{2}-1} [f(a + h(1 - 1/\sqrt{3} + 2j)) + f(a + h(1 + 1/\sqrt{3} + 2j))] \quad (4.58)$$

przy czym $h = \frac{b-a}{m}$. Błąd tego wzoru wynosi

$$E(f) = \frac{(b-a)^5}{270m^4} f^{(4)}(\xi_1), \quad \xi_1 \in (a; b) \quad (4.59)$$

4.4. UWAGI KOŃCOWE

Podamy teraz pewne zalecenia co do wyboru omówionych w tym rozdziale kwadratur, gdy przedział całkowania jest skończony, a funkcja podcałkowa jest dana analitycznie i nie ma osobliwości.

Jak wiemy, błąd $E(f)$ kwadratury można przedstawić w postaci

$$E(f) = C_r f^{(r)}(\xi) \quad (4.60)$$

gdzie r jest rzędem kwadratury, C_r nie zależy od f , ξ należy do przedziału całkowania.

Rozpatrzmy dwie następujące sytuacje:

1. Umiemy obliczyć (oszacować) pochodne funkcji podcałkowej. Jeżeli umiemy obliczyć (oszacować) pochodne i nie rosną one zbyt szybko, to możemy zastosować kwadraturę (4.2) (może być ona wysokiego rzędu). Po wyborze typu kwadratury, rząd r wybieramy szacując (4.60). Możemy zastosować zarówno wzory Newtona-Cotesa, jak i Gaussa, przy czym wzory Gaussa są lepsze, ponieważ możemy osiągnąć potrzebną dokładność przy znacznie mniejszej liczbie działań. Istotnie, główną część obliczeń przy wyznaczaniu $S(f)$ stanowią obliczenia wartości funkcji podcałkowej, a wzory Gaussa dają większą dokładność przy mniejszej liczbie węzłów; ten sam rząd $r = 2N + 2$ mają kwadratury Gaussa dla $(N + 1)$ węzłów, a kwadratury Newtona-Cotesa dla $(2N + 1)$ węzłów. Ponadto współczynniki C_r w przypadku kwadratur Gaussa są co do modułu nieco mniejsze.

Rozpatrywana wyżej sytuacja w praktyce występuje rzadko. Jeśli nawet umiemy obliczyć (oszacować) pochodne, to bardzo często rosną one zbyt szybko. Jeśli tak jest (względnie umiemy szacować tylko pochodne niskiego rzędu), to możemy zastosować kwadraturę złożoną. Po wyborze typu kwadratury, szacując błąd wyznaczamy liczbę podprzedziałów, na jaką należy podzielić przedział całkowania, aby uzyskać potrzebną dokładność. Porównując wzory złożone Gaussa i Newtona-Cotesa niskiego rzędu widzimy, że pierwsze z nich nie są już wyraźnie lepsze. Istotnie, dla wzorów rzędu czwartego przy podziale przedziału całkowania na m równych części liczba węzłów w przypadku wzorów Gaussa i Newtona-Cotesa wynosi odpowiednio $2m$ i $m + 1$.

2. Nie umiemy oszacować pochodnych funkcji podcałkowej. Ustalając rząd wybranej kwadratury nie jesteśmy więc pewni, czy osiągniemy założoną dokładność. W takiej sytuacji najczęściej stosujemy kwadratury złożone niskiego rzędu w ten sposób, że zwiększamy liczbę podprzedziałów aż do osiągnięcia stabilizacji w kolejnych dwóch przybliżeniach.

Tablica 4/4 zawiera wyniki obliczeń całki.

4/4

m	Wzór złożony parabol	Wzór złożony Gaussa-Legendre'a (5.57)	Wzór złożony trapezów	Metoda Romberga
2	0.69444444	0.69230769	0.70833333	0.70833333
4	0.69325397	0.69307664	0.69702381	0.69444444
8	0.69315453	0.69314229	0.69412185	0.69317460
16	0.69314765	0.69314687	0.69339120	0.69314748
32	0.69314721	0.69314716	0.69320821	0.69314718
64	0.69314718	0.69314718	0.69316244	0.69314718
128	0.69314718	0.69314718	0.69315100	0.69314718

$$\int_1^2 \frac{1}{x} dx = \ln 2 \approx 0,69314718 \quad (4.61)$$

za pomocą wzorów złożonych: wzoru parabol (4.22), wzoru Gaussa-Legendre'a (4.58) rzędu 4, wzoru trapezów (4.19) oraz metodą Romberga dla podziału na $m = 2^i$ ($i = 1, \dots, 7$) równych części przedziału $\langle 1; 2 \rangle$.

Widzimy, że dwa pierwsze wzory złożone są szybciej zbieżne niż trzeci. Związane jest to z tym, że we wzorach (4.23), (4.59), określających błąd tych kwadratur, występuje czynnik $\frac{1}{m^4}$, natomiast w (4.21) czynnik ten wynosi $\frac{1}{m^2}$.

Metoda Romberga jest także szybciej zbieżna niż złożony wzór trapezów.

Jeżeli wykonujemy obliczenia kolejno dla podziału na $m = 2^1, 2^2, \dots$ równych części przedziału całkowania, to w przypadku wzorów (4.19) i (4.22) możemy wykorzystać fakt, że węzły z poprzedniego podziału są także węzłami w następnym podziale, natomiast w przypadku wzorów (4.58) tak nie jest. Z tego powodu wymienione dwie pierwsze kwadratury wymagają mniejszego nakładu obliczeń.

Jeżeli funkcja podcałkowa nie jest dana analitycznie, lecz jest określona za pomocą tablicy (zakładamy, że argumenty są równoodległe, ponieważ najczęściej w praktyce tak jest), to prawie na pewno stosujemy wzory Newtona-Cotesa. Zastosowanie wzorów Gaussa wymagałoby przeprowadzenia interpolacji. Przy wyborze wzoru całkowania kierujemy się takimi samymi przesłankami jak w przypadku funkcji danych analitycznie.

5

ROZWIĄZYWANIE UKŁADÓW ALGEBRAICZNYCH RÓWNAŃ LINIOWYCH

5.1. WSTĘP

Układ równań liniowych $Ax = b$, gdzie A jest daną macierzą o m wierszach i n kolumnach, b — wektorem m danych liczb, x — wektorem n niewiadomych, może mieć nieskończenie wiele rozwiązań, dokładnie jedno rozwiązanie lub nie mieć wcale rozwiązań (układ sprzeczny). Jakkolwiek warunki istnienia rozwiązania są teoretycznie proste i w wielu przypadkach istnieją gotowe wzory na wyznaczanie składowych wektora x , numeryczne wyznaczenie rozwiązania może być zadaniem bardzo trudnym. Istnieje wiele metod rozwiązywania układów równań i decyzja, którą z nich wybrać, powinna zależeć nie tylko od postaci macierzy A , ale także od specyfiki zagadnienia, które reprezentuje dany układ równań. Jeśli na przykład istnieje potrzeba wielokrotnego rozwiązywania układu równań przy różnych wartościach wektora b , to należy zastosować metodę, która w wyniku znalezienia pierwszego rozwiązania przekształci układ równań do prostszej postaci, ułatwiającej wyznaczanie dalszych rozwiązań (patrz metody z rozkładem LU , LL^T , LDL^T , QR). Metoda taka może jednak dawać rozwiązania o zbyt małej dokładności, co zmusi nas do zastosowania metody bardziej czasochłonnej, lecz dokładniejszej. Często szczególne własności macierzy A umożliwiają zastosowanie specjalnej metody, szybkiej i dokładnej (patrz p. 5.3.5 i 5.3.6). W przypadku układu wielu równań z wieloma niewiadomymi może okazać się niezbędne oszczędne operowanie pamięcią maszyny cyfrowej; korzystamy wówczas z metod typowych dla układów o macierzach rzadkich (patrz p. 5.5). Z powyższych względów przy szczegółowym omawianiu metod rozwiązywania układów równań będziemy się starali podać nakład obliczeń (liczbę mnożeń i dodawań), dokładność uzyskiwanego rozwiązania i liczbę zajętych komórek pamięci maszyny cyfrowej.

Należy pamiętać, że istnieją trzy przyczyny występowania błędów

w numerycznie wyznaczonym rozwiązaniu: błąd spowodowany niedokładnymi wartościami współczynników układu równań (błąd wejściowy), błąd powstający na skutek popełniania błędów podczas numerycznego wykonywania działań arytmetycznych (błąd zaokrągleń) oraz błąd wyznaczania rozwiązania spowodowany sposobem działania metody (błąd metody), gdy np. wyznaczamy ciąg wektorów $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ zbieżny do rozwiązania, lecz po skończonym czasie obliczeń przyjmujemy pewien wektor $\mathbf{x}^{(N)}$ jako przybliżone rozwiązanie. Przy szacowaniu błędu metody korzysta się z norm wektorów i macierzy. Poniżej zestawiono najważniejsze własności norm w przestrzeniach $\mathbf{R}^n$.

5.2. POJĘCIA PODSTAWOWE

W przestrzeni $\mathbf{R}^n$, której elementami są wektory $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$, można wprowadzić wiele norm wektorów, przy czym w obliczeniach numerycznych najczęściej są stosowane następujące normy:

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \dots + |x_n| \quad (5.1a)$$

$$\|\mathbf{x}\|_2 = (x_1^2 + x_2^2 + \dots + x_n^2)^{1/2} \quad (5.1b)$$

$$\|\mathbf{x}\|_\infty = \max \{|x_1|, |x_2|, \dots, |x_n|\} \quad (5.1c)$$

Dla dowolnego wektora $\mathbf{x} \in \mathbf{R}^n$ obowiązują nierówności

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n} \|\mathbf{x}\|_2 \leq n \|\mathbf{x}\|_\infty \quad (5.2)$$

można jednakże wskazać takie niezerowe wektory $\mathbf{x}$, dla których sąsiednie wyrażenia są równe, np. $\|\mathbf{x}\|_1 = \sqrt{n} \|\mathbf{x}\|_2$ dla $\mathbf{x} = [1, 1, \dots, 1]^T$.

U w a g a. W przestrzeniach $\mathbf{R}^n$ wszystkie normy wektorów są równoważne w tym sensie, że jeśli według pewnej normy ciąg wektorów $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ dąży do wektora zerowego, czyli ciąg liczb $\|\mathbf{x}^{(1)}\|, \|\mathbf{x}^{(2)}\|, \dots$ dąży do zera, to dla dowolnej innej normy również zachodzi zbieżność do zera.

Macierz A o m wierszach i n kolumnach może być traktowana jako operator liniowy przekształcający przestrzeń $\mathbf{R}^n$ w $\mathbf{R}^m$. Można zatem określić normę tego operatora korzystając z definicji

$$\|A\|_{pq} = \max_{\substack{\mathbf{x} \in \mathbf{R}^n \\ \mathbf{x} \neq 0}} \frac{\|A\mathbf{x}\|_q}{\|\mathbf{x}\|_p} \quad (5.3)$$

Indeksy p i q odpowiadają normom wektorów w przestrzeniach $\mathbf{R}^n$ i $\mathbf{R}^m$ odpowiednio, $\|A\|_{pq}$ jest normą macierzy indukowaną¹⁾ przez normy wektorów $\|\cdot\|_p$ i $\|\cdot\|_q$. Gdy $p = q$, przyjmujemy oznaczenie $\|A\|_p = \|A\|_{pq}$.

¹⁾ Norma macierzy indukowana przez normę wektorów bywa niekiedy oznaczana symbolem „lub” (np. $\text{lub}_2(A)$, $\text{lub}_1(A)$ itp.). Jest to skrót do określenia „lower upper bound”.

W szczególności $\|A\|_1$, $\|A\|_2$ i $\|A\|_\infty$ oznaczają normy macierzy A , gdy w obu przestrzeniach przyjęto odpowiednio normy wektorów $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_\infty$. Znane są zależności:

$$\|A\|_1 = \max_{j=1,2,\dots,n} \sum_{i=1}^m |a_{ij}| \quad (5.4a)$$

(maksymalna suma modułów w kolumnie)

$$\|A\|_2 = (\text{największa wartość własna macierzy } A^T A)^{\frac{1}{2}} \quad (5.4b)$$

$$\|A\|_\infty = \max_{i=1,2,\dots,m} \sum_{j=1}^n |a_{ij}| \quad (5.4c)$$

(maksymalna suma modułów w wierszu)

$$\|A\|_{1\infty} = \max_{i,j} |a_{ij}| \quad (5.4d)$$

(maksymalny moduł elementu)

Zwróćmy uwagę, że wartości $\|A\|_1$, $\|A\|_\infty$ i $\|A\|_{1\infty}$ można łatwo obliczyć numerycznie, wartość $\|A\|_2$ jest natomiast na ogół trudna do obliczenia (patrz rozdz. 6). Z drugiej strony norma $\|x\|_2$ ma naturalne zastosowanie w wielu zagadnieniach. Z tej przyczyny w przestrzeniach z normą wektorów typu $\|\cdot\|_2$ używa się często tzw. *euklidesowej normy macierzy* (*normy Schura*, *normy Frobeniusa*):

$$\|A\|_E = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2} \quad (5.4e)$$

która jakkolwiek nie jest indukowana przez żadną normę wektorów, spełnia z normą $\|\cdot\|_2$ *warunek zgodności*

$$\|Ax\|_2 \leq \|A\|_E \|x\|_2 \quad \text{dla każdego } x \in \mathbf{R}^n \quad (5.5)$$

Warunek zgodności spełnia każda indukowana norma macierzy z odpowiednią normą wektorów. Ponadto istnieją wówczas niezerowe wektory x takie, że $\|Ax\| = \|A\| \|x\|$. Na przykład dla $A = \begin{bmatrix} 3 & 2 \\ 2 & 0 \end{bmatrix}$ są to wektory $x^{(1)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $x^{(2)} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $x^\infty = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $x^{(1\infty)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ dla odpowiednich norm.

Tymczasem $\|A\|_E = \sqrt{17} > 4 = \|A\|_2$, więc dla każdego niezerowego $x \in \mathbf{R}^2$ norma $\|Ax\|_2 < \|A\|_E \|x\|_2$. Można zatem powiedzieć, że norma euklidesowa jest na ogół „zbyt duża”. Ogólnie, dla dowolnej macierzy A mamy

$$\|A\|_2 \leq \|A\|_E \leq \sqrt{n} \|A\|_2 \quad (5.6)$$

Dla dowolnych macierzy A i B , dla których jest określona macierz AB ,

wszystkie normy indukowane spełniają warunek

$$\|AB\|_{pq} \leq \|A\|_{rq} \|B\|_{pr} \quad (5.7)$$

dla dowolnych indeksów p, q, r .

Powyższa nierówność jest spełniona również dla normy euklidesowej. W szczególności

$$\|AB\|_p \leq \|A\|_p \|B\|_p$$

a więc

$$\|A^n\|_p \leq \|A\|_p^n \quad (5.8)$$

Dla dowolnej normy macierzy kwadratowej zgodnej z normą wektorów obowiązuje zależność

$$|\lambda_i| \leq \|A\| \quad (5.9)$$

gdzie λ_i jest dowolną wartością własną macierzy A (może to być liczba zespolona).

Wniosek

$$\|A\|_2^2 \leq \|A^T A\|_\infty \leq \|A^T\| \|A\|_\infty = \|A\|_1 \|A\|_\infty \quad (5.10)$$

Lemat 1. Jeżeli $\|M_{n \times n}\|_p < 1$, to macierz $I + M$ jest nieosobliwa i

$$\|(I + M)^{-1}\|_p \leq \frac{1}{1 - \|M\|_p} \quad \text{dla } p = 1, 2, \infty \quad (5.11)$$

(I oznacza macierz jednostkową).

DOWÓD. Gdyby macierz $I + M$ była macierzą osobliwą, to istniałby niezerowy wektor x taki, że $(I + M)x = 0$. Zatem $\|x\|_p = \|-Mx\|_p \leq \|M\|_p \|x\|_p$ — otrzymaliśmy sprzeczność.

Ponieważ $(I + M)^{-1}(I + M) = I$, czyli $(I + M)^{-1} + (I + M)^{-1}M = I$, więc

$$1 = \|I\|_p \geq \|(I + M)^{-1}\|_p - \|(I + M)^{-1}\|_p \|M\|_p \Rightarrow \|(I + M)^{-1}\|_p \leq \frac{1}{1 - \|M\|_p}$$

Liczbę $\max_{i=1,2,\dots,n} |\lambda_i|$, gdzie λ_i są wartościami własnymi macierzy A , nazywamy *promieniem spektralnym macierzy A* i oznaczamy $\varrho(A)$. Na podstawie (5.9) mamy zatem

$$\varrho(A) \leq \|A\|_p, \quad p = 1, 2, \infty, E \quad (5.12)$$

Lemat 2. Dla każdej liczby $\varepsilon > 0$ istnieje taka indukowana norma $\|\cdot\|_p$ macierzy A , że

$$\|A\|_p \leq \varrho(A) + \varepsilon$$

Dowód tego lematu można znaleźć np. w [45, s. 47].

Tw. Dla każdego rzeczywistego wektora x ciąg wektorów $Ax, A^2x, \dots, A^i x, \dots$, jest zbieżny do zera wtedy i tylko wtedy, gdy $\varrho(A) < 1$.

DOWÓD. Jeżeli $\varrho(A) < 1$, to przyjmując $\varepsilon = \frac{1 - \varrho(A)}{2}$ można wskazać taką normę $\|\cdot\|_p$ macierzy A , że $\|A\|_p \leq \frac{1 + \varrho(A)}{2} < 1$. Zatem $\|A^n x\|_p \leq \|A\|_p^n \cdot \|x\|_p \rightarrow 0$, czyli $A^n x \rightarrow 0$.

Przypuśćmy z kolei, że $\varrho(A) \geq 1$ i że dla wszystkich rzeczywistych wektorów x ciąg $Ax, A^2x, \dots$ jest zbieżny do zera. Niech λ będzie wartością własną macierzy A o największym module, tzn. $|\lambda| = \varrho(A)$, i niech $|\lambda| \geq 1$. Liczba λ może być zespolona, więc w ogólności odpowiada jej zespolony wektor własny v taki, że $Av = \lambda v$ ¹⁾.

Niech $v = a + ib$, $\lambda^n = \alpha_n + i\beta_n$. Ponieważ $|\lambda^n| = \sqrt{\alpha_n^2 + \beta_n^2}$, więc $\alpha_n^2 + \beta_n^2 \geq 1$. Mamy $A^n v = A^n a + iA^n b$ oraz $A^n v = \lambda^n v = (\alpha_n a - \beta_n b) + i(\beta_n a + \alpha_n b)$, stąd $A^n a = \alpha_n a - \beta_n b$, $A^n b = \beta_n a + \alpha_n b$. Zauważmy, że

$$\begin{aligned} \|A^n a\|_2^2 + \|A^n b\|_2^2 &= \langle \alpha_n a - \beta_n b, \alpha_n a - \beta_n b \rangle + \langle \beta_n a + \alpha_n b, \beta_n a + \alpha_n b \rangle = \\ &= (\alpha_n^2 + \beta_n^2)(\|a\|^2 + \|b\|^2) \geq (\|a\|^2 + \|b\|^2) > 0 \end{aligned}$$

a więc oba ciągi $(A^n a)$ i $(A^n b)$ nie mogą być równocześnie zbieżne do zera. Zatem jeden z wektorów a lub b jest przykładem rzeczywistego wektora x , dla którego ciąg $Ax, A^2x, \dots$ nie jest zbieżny do zera. Uzyskano sprzeczność z założeniem w dowodzie nie wprost, co uzasadnia prawdziwość twierdzenia.

U w a g a. Symbolem $\langle \cdot, \cdot \rangle$ oznaczono iloczyn skalarny wektorów, określony następująco:

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 + \dots + x_n y_n \quad (5.13)$$

Ćwiczenia

1. Wskazać niezerowe wektory x takie, by: $\|x\|_\infty = \|x\|_2$, $\|x\|_2 = \|x\|_1$, $\|x\|_1 = \sqrt{n} \|x\|_2$, $\|x\|_2 = \sqrt{n} \|x\|_\infty$.
2. Wykazać, że normy $\|\cdot\|_1$ i $\|\cdot\|_\infty$ są równoważne.
3. Wykazać, że $\|A\|_{1\infty} = \max_{i,j} |a_{ij}|$. W s k a z ó w k a. Jeśli dla pewnej liczby c mamy $\|Ax\| \leq c \|x\|$, to $\|A\| \leq c$. Aby liczba c była równa $\|A\|$, należy wskazać niezerowy wektor x taki, że $\|Ax\| = c \|x\|$.
4. Wskazać niezerowe wektory x takie, by: $\|Ax\|_1 = \|A\|_1 \|x\|_1$, $\|Ax\|_\infty = \|A\|_\infty \|x\|_\infty$, $\|Ax\|_\infty = \|A\|_{1\infty} \|x\|_1$.
5. Uzasadnić zależności (5.7) i (5.9).
6. Wykazać, że gdy macierz A jest symetryczna, wówczas $\|A\|_2 = \varrho(A)$.
7. Podać przykład macierzy A takiej, że $\|A\| > 1$, a $\varrho(A) < 1$.

5.3. METODY DOKŁADNE

Jeżeli rozwiązanie układu równań $Ax = b$ polega na takim przekształcaniu danych A i b , że przy założeniu dokładnie wykonywanych działań arytmetycznych po skończonej liczbie działań otrzymujemy rozwiązanie, to taką metodę rozwiązywania nazywamy *metodą dokładną*. Metody dokładne charakteryzu-

¹⁾ W wielu podręcznikach (patrz np. [70] lub [45]), dalsza część dowodu nie wprost polega na rozszerzeniu pojęcia normy wektorów na przypadek wektorów zespolonych, po czym korzysta się z warunku $\|A^n v\| = |\lambda|^n \cdot \|v\| \geq \|v\| > 0$. Ciąg $(A^n v)$ nie jest zatem zbieżny do zera. Nie stanowi to jednak sprzeczności z założeniem, że ciągi $(A^n x)$ są zbieżne do zera, bo wektory x są rzeczywiste, zaś wektor v w ogólności zespolony.

ją się małą liczbą obliczeń potrzebnych do wyznaczenia rozwiązania, jednak dla zadań źle uwarunkowanych numerycznie wyznaczone rozwiązanie może być obarczone dużym błędem. Można więc powiedzieć, że metody dokładne nie dają błędu metody, lecz mogą być *niestabilne* ze względu na błędy zaokrągleń (patrz p. 3.3). Ponadto przekształcanie macierzy A obciąża w dużym stopniu pamięć maszyny, zwłaszcza jeśli początkowe dane A i b należy przechowywać celem ostatecznego sprawdzenia.

Do metod dokładnych należy zaliczyć wyznaczanie rozwiązań na podstawie gotowych wzorów (wzorów Cramera); np. dla układu

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1 \\ a_{21}x_1 + a_{22}x_2 &= b_2 \end{aligned} \quad (5.14)$$

mamy

$$x_1 = \frac{a_{22}b_1 - a_{12}b_2}{a_{11}a_{22} - a_{21}a_{12}}, \quad x_2 = \frac{a_{11}b_2 - a_{21}b_1}{a_{11}a_{22} - a_{21}a_{12}}$$

Metoda ta oszczędza pamięć operacyjną maszyny, lecz dla większej liczby równań obliczenie wyznaczników wymaga dużej liczby mnożeń, a ponadto może dawać duży błąd w rozwiązaniu.

Dla przykładu rozwiążemy układ równań

$$\begin{aligned} 0,99x_1 + 0,70x_2 &= 0,54 \\ 0,70x_1 + 0,50x_2 &= 0,38 \end{aligned} \quad (5.15)$$

zakładając, że dysponujemy maszyną cyfrową z arytmetyką zmiennie-pozytywną (patrz p. 0.1), której słowo składa się z 2 cyfr dziesiętnych. Założymy ponadto, że mnożenie jest wykonywane w akumulatorze z podwójną precyzją, ale każdy wynik przed przesłaniem do pamięci jest zaokrąglany. Wynik a działania wykonanego na maszynie będziemy oznaczać $\text{fl}(a)$ (patrz p. 0.2).

Obliczamy wyznacznik główny $a_{11}a_{22} - a_{21}a_{12}$. Mamy

$$\begin{aligned} a_{11}a_{22} &= 0,99 \cdot 0,50 = 0,4950; & \text{fl}(a_{11}a_{22}) &= 0,50 \\ a_{21}a_{12} &= 0,70 \cdot 0,70 = 0,4900; & \text{fl}(a_{21}a_{12}) &= 0,49 \end{aligned}$$

$$\text{Zatem } \text{fl}(a_{11}a_{22} - a_{21}a_{12}) = 0,50 - 0,49 = 0,10 \cdot 10^{-1}.$$

Analogicznie $\text{fl}(a_{22}b_1 - a_{12}b_2) = \text{fl}(a_{22}b_1) - \text{fl}(a_{12}b_2) = \text{fl}(0,2700) - \text{fl}(0,2660) = 0,27 - 0,27 = 0,00$, a więc $\text{fl}(x_1) = \frac{0,00}{0,1 \cdot 10^{-1}} = 0,00$. Podobnie $\text{fl}(a_{11}b_2 - a_{21}b_1) = 0,00$ czyli $\text{fl}(x_2) = 0,00$, gdy tymczasem dokładne rozwiązanie wynosi: $x_1 = 0,80$, $x_2 = -0,36$.

Dla porównania wyznaczmy rozwiązanie inną metodą, eliminując niewiadomą x_1 z drugiego równania układu równań

$$\begin{aligned} 0,99x_1 + 0,70x_2 &= 0,54 \\ 0,70x_1 + 0,50x_2 &= 0,38 \end{aligned}$$

Obliczamy $\text{fl}\left(\frac{a_{21}}{a_{11}}\right) = \text{fl}(0,7070) = 0,71$. Układ przekształcamy odejmując od drugiego wiersza pierwszy wiersz pomnożony przez 0,71. Mamy wówczas

$$0,99x_1 + 0,70x_2 = 0,54$$

$$\text{fl}(0,50 - 0,70 \cdot 0,71)x_2 = \text{fl}(0,38 - 0,54 \cdot 0,71)$$

czyli

$$0,99x_1 + 0,70x_2 = 0,54$$

$$0,00x_2 = 0,00$$

Uzyskano układ nieoznaczony, posiadający nieskończenie wiele rozwiązań,

$$\text{np. } x_2 = 0,00, \text{ fl}(x_1) = \text{fl}\left(\frac{0,54}{0,99}\right) = 0,55.$$

Powyższy przykład dobitnie pokazuje, iż mimo stosowania metody dokładnej uzyskane rozwiązanie nie jest dokładne, a nawet nie stanowi zaokrąglenia rozwiązania dokładnego.

5.3.1. Analiza błędów rozwiązywania

Jeśli macierz A i wektor b zastąpimy danymi $A + \delta A$ i $b + \delta b$, to zamiast dokładnego rozwiązania x układu $Ax = b$, uzyskamy rozwiązanie dokładne $x + \delta x$. Wielkość δx błędu wejściowego zależy nie tylko od wartości zaburzeń δA i δb , ale także od tzw. uwarunkowania zadania (patrz p. 0.4). Rozpatrzmy oddzielnie trzy przypadki wpływu zaburzeń δb i δA dla kwadratowej i nieosobliwej macierzy A .

1. Niech $\delta b \neq 0$, a $\delta A = 0$, czyli

$$A(x + \delta x) = b + \delta b \Rightarrow \delta x = A^{-1}\delta b$$

Dla dowolnych norm wektorów δx i δb oraz indukowanej przez nie normy macierzy A^{-1} możemy napisać $\|\delta x\|_p \leq \|A^{-1}\|_{qp} \|\delta b\|_q$. Liczba $\|A^{-1}\|_{qp}$ stanowi miarę wpływu zaburzenia δb na zmianę rozwiązania δx . Interesuje nas jednak zazwyczaj wpływ względnego zaburzenia $\frac{\|\delta b\|_q}{\|b\|_q}$, dla $b \neq 0$. Mamy

$$\frac{\|\delta x\|_p}{\|x\|_p} \leq \frac{\|A^{-1}\|_{qp} \|b\|_q}{\|x\|_p} \frac{\|\delta b\|_q}{\|b\|_q} \quad (5.16)$$

Oznaczmy $k = \frac{\|A^{-1}\|_{qp} \|b\|_q}{\|x\|_p}$. Liczba k ma tę własność, że

$$\begin{aligned} \bigvee \delta \mathbf{b} \neq 0: \frac{\|\delta \mathbf{x}\|}{\|\mathbf{x}\|} &= k \frac{\|\delta \mathbf{b}\|}{\|\mathbf{b}\|} \\ \bigwedge \delta \mathbf{b} \in \mathbf{R}^n: \frac{\|\delta \mathbf{x}\|}{\|\mathbf{x}\|} &\leq k \frac{\|\delta \mathbf{b}\|}{\|\mathbf{b}\|} \end{aligned} \quad (5.17)$$

Liczbę k spełniającą warunki (5.17) nazywamy *wzmocnieniem zaburzenia*. Określona tak liczba k ma jednak tę wadę, że trudno ją obliczyć bez znajomości dokładnego rozwiązania $\mathbf{x}$ układu niezaburzonego $\mathbf{A}\mathbf{x} = \mathbf{b}$. Zauważmy jednak, że

$$k = \frac{\|\mathbf{A}^{-1}\|_{qp} \|\mathbf{A}\mathbf{x}\|_q}{\|\mathbf{x}\|_p} \leq \|\mathbf{A}^{-1}\|_{qp} \|\mathbf{A}\|_{pq} = K_{pq}$$

Liczbę K_{pq} nazywamy *wskaznikiem uwarunkowania* układu równań $\mathbf{A}\mathbf{x} = \mathbf{b}$ (w dalszej części książki będziemy opuszczać indeksy p i q wszędzie tam, gdzie nie spowoduje to niejednoznaczności). Jak widać, liczba ta jest niezależna od rozwiązania $\mathbf{x}$, a także od wektora $\mathbf{b}$. W konsekwencji względne zaburzenie $\frac{\|\delta \mathbf{x}\|}{\|\mathbf{x}\|}$ może być znacznie mniejsze od wartości $K \frac{\|\delta \mathbf{b}\|}{\|\mathbf{b}\|}$, stanowiącej jedynie zgrubne oszacowanie od góry.

Przykład 1. Układ równań

$$\begin{bmatrix} s & 0 \\ 0 & 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad s > 0 \quad (5.18)$$

ma rozwiązanie $\mathbf{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Przy zaburzeniu $\delta \mathbf{b} = \begin{bmatrix} \delta b_1 \\ \delta b_2 \end{bmatrix}$ powstaje zmiana rozwiązania

$$\delta \mathbf{x} = \begin{bmatrix} \frac{\delta b_1}{s} \\ \delta b_2 \end{bmatrix}. \text{ Niech } s \geq 1. \text{ Wówczas } \|\delta \mathbf{x}\|_\infty \leq \|\delta \mathbf{b}\|_\infty, \text{ a więc } \frac{\|\delta \mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \leq \frac{\|\delta \mathbf{b}\|_\infty}{\|\mathbf{b}\|_\infty}, \bigwedge \delta \mathbf{b} \in \mathbf{R}^n,$$

Stwierdzamy, że wzmocnienie zaburzenia $k = 1$, tzn. układ (5.18) jest mało wrażliwy na zaburzenia prawej strony układu równań, tymczasem $K = \|\mathbf{A}^{-1}\|_\infty \|\mathbf{A}\|_\infty = s$, a więc dla dużych s wskaźnik uwarunkowania sugeruje duże zmiany w rozwiązaniu.

2. Niech teraz $\delta \mathbf{b} = 0$, a $\delta \mathbf{A} \neq 0$. Wyznamy oszacowanie *różniczkowego wzmocnienia zaburzenia*, tzn. granicznej wartości k_0 współczynników k , spełniających zależności

$$\bigvee \delta \mathbf{A} \neq 0: \frac{\|\delta \mathbf{x}\|}{\|\mathbf{x}\|} = k \frac{\|\delta \mathbf{A}\|}{\|\mathbf{A}\|} \quad (5.19)$$

$$\frac{\|\delta \mathbf{x}\|}{\|\mathbf{x}\|} \leq k \frac{\|\delta \mathbf{A}\|}{\|\mathbf{A}\|}, \quad \bigwedge \delta \mathbf{A} \in \mathbf{R}^n \text{ takiego, że } \frac{\|\delta \mathbf{A}\|}{\|\mathbf{A}\|} \leq \delta$$

gdy $\delta \rightarrow 0$.

Wyznamy oszacowanie k_0 rozumując następująco. Równanie $\mathbf{A}\mathbf{x} = \mathbf{b}$ przy ustalonej wartości $\mathbf{b}$ można zapisać jako uwikłaną zależność $\mathbf{x}$ od

A w postaci równości $F(\mathbf{x}, A) = A\mathbf{x} - \mathbf{b} = 0$. Zapiszmy macierz A jako wektor $\mathbf{a} = [a_{11}, a_{21}, \dots, a_{n1}, a_{12}, a_{22}, \dots, a_{n2}, \dots, a_{nn}]^T$. Mamy zatem $F(\mathbf{x}, \mathbf{a}) = 0$ oraz dla każdego zaburzenia δA , dla którego macierz $A + \delta A$ jest nieosobliwa, $F(\mathbf{x} + \delta \mathbf{x}, \mathbf{a} + \delta \mathbf{a}) = 0$. Ponieważ $F_{\mathbf{x}}(\mathbf{x}, \mathbf{a}) = A$ jest macierzą nieosobliwą, więc na podstawie twierdzenia o pochodnej funkcji uwikłanej mamy $\frac{\partial \mathbf{x}}{\partial \mathbf{a}} = -[F_{\mathbf{x}}(\mathbf{x}, \mathbf{a})]^{-1} F_{\mathbf{a}}(\mathbf{x}, \mathbf{a})$. Z kolei, ponieważ $[F_{\mathbf{x}}(\mathbf{x}, \mathbf{a})]^{-1} = A^{-1}$,

a $F_{\mathbf{a}}(\mathbf{x}, \mathbf{a}) = [x_1 \mathbf{I}, x_2 \mathbf{I}, \dots, x_n \mathbf{I}]$, więc $\frac{\partial \mathbf{x}}{\partial \mathbf{a}} = -[x_1 A^{-1}, x_2 A^{-1}, \dots, x_n A^{-1}]$.

Można zatem wyznaczyć związek między skończonymi (lecz małymi) przyrostami $\delta \mathbf{x}$ i $\delta \mathbf{a}$:

$$\delta \mathbf{x} = -[x_1 A^{-1}, x_2 A^{-1}, \dots, x_n A^{-1}] \delta \mathbf{a} + o(\|\delta \mathbf{a}\|)$$

Zauważmy, że

$$\|[x_1 A^{-1}, x_2 A^{-1}, \dots, x_n A^{-1}]\|_{1\infty} = \max_{i,j,k} |x_k a_{ij}^{(-1)}| = \|\mathbf{x}\|_{\infty} \|A^{-1}\|_{1\infty}$$

Zatem różniczkowe wzmocnienie zanurzenia $\frac{\|\delta \mathbf{a}\|_1}{\|A\|_{1\infty}}$, dające względny błąd

$$\frac{\|\delta \mathbf{x}\|_{\infty}}{\|\mathbf{x}\|_{\infty}}, \text{ oszacowany jako}$$

$$\frac{\|\delta \mathbf{x}\|_{\infty}}{\|\mathbf{x}\|} \leq \|A^{-1}\|_{1\infty} \|A\|_{1\infty} \frac{\|\delta \mathbf{a}\|_1}{\|A\|_1} + o(\|\delta \mathbf{a}\|_1)$$

wynosi

$$k_0 = \|A^{-1}\|_{1\infty} \|A\|_{1\infty} \quad (5.20)$$

Reasumując, jeśli za miarę zaburzenia macierzy A przyjmiemy wielkość

$$\frac{\sum_{i=1}^n \sum_{j=1}^n |\delta a_{ij}|}{\max_{i,j} |a_{ij}|} = \frac{\|\delta \mathbf{a}\|_1}{\|A\|_1}, \text{ to wartość } k_0 \text{ oblicza się stosunkowo łatwo według}$$

wzoru (5.20), jest to bowiem iloczyn maksymalnych modułów elementów macierzy A^{-1} i A . Jeśli wówczas przyjąć zaburzenie δA takie, że

$$\frac{|\delta a_{ij}|}{\|A\|_{1\infty}} = \varrho \ll 1 \text{ dla elementu } a_{ij} \text{ znajdującego się w tym wierszu macierzy } A,$$

w którym dla macierzy A^{-1} znajduje się element o maksymalnym module, w kolumnie odpowiadającej maksymalnej co do modułu współrzędnej x_i rozwiązania $\mathbf{x}$, a pozostałe $\delta a_{ij} = 0$, to

$$\frac{\|\delta \mathbf{x}\|}{\|\mathbf{x}\|} = k_0 \frac{\|\delta \mathbf{a}\|_1}{\|A\|} + o(\varrho)$$

Przykład 2. Rozpatrując układ (5.18), otrzymujemy dla $s \leq 1$:

$$k_0 = \left\| \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} \right\|_{1\infty} \cdot \left\| \begin{bmatrix} s & 0 \\ 0 & 1 \end{bmatrix} \right\|_{1\infty} = \frac{1}{s}$$

W macierzy A^{-1} maksymalny element znajduje się w pierwszym wierszu, maksymalną współrzędną rozwiązania $\mathbf{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ jest x_2 , a więc rozpatrujemy zaburzenie elementu a_{12} , tzn.

$A = \begin{bmatrix} 0 & \delta \\ 0 & 0 \end{bmatrix}$. Mamy zatem

$$\begin{bmatrix} s & \delta \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 + \delta x_1 \\ 1 + \delta x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \Rightarrow \delta \mathbf{x} = \begin{bmatrix} -\frac{\delta}{s} \\ 0 \end{bmatrix} \Rightarrow \frac{\|\delta \mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = k_0 \frac{\|\delta \mathbf{a}\|_1}{\|A\|_{1\infty}}$$

3. Niech z kolei $\delta \mathbf{b} \neq 0$ i $\delta A \neq 0$, przy czym zakładamy, że zaburzenie δA jest na tyle małe, że macierz $A + \delta A$ jest nieosobliwa. Zatem

$$\begin{aligned} (A + \delta A)(\mathbf{x} + \delta \mathbf{x}) &= \mathbf{b} + \delta \mathbf{b} \Rightarrow \delta \mathbf{x} = -A^{-1}(\delta \mathbf{b} - \delta A \mathbf{x} - \delta A \delta \mathbf{x}) \Rightarrow \\ \Rightarrow \|\delta \mathbf{x}\|_p &\leq \|A^{-1}\|_{qp}(\|\delta \mathbf{b}\|_q + \|\delta A\|_{pq}\|\mathbf{x}\|_p + \\ &+ \|\delta A\|_{pq}\|\delta \mathbf{x}\|_p) \cdot (1 - \|A^{-1}\|_{qp}\|\delta A\|_{pq}) \|\delta \mathbf{x}\|_p \leq \|A^{-1}\|_{qp}(\|\delta \mathbf{b}\|_q + \|\delta A\|_{pq}\|\mathbf{x}\|_p) \end{aligned}$$

Przyjmijmy, iż zaburzenie δA jest na tyle małe, że $\|A^{-1}\|_{qp}\|A\|_{pq} = \varrho < 1$. Wtedy

$$\frac{\|\delta \mathbf{x}\|_p}{\|\mathbf{x}\|_p} \leq \frac{1}{1 - \varrho} \|A^{-1}\|_{qp} \|A\| \left(\frac{\|\delta \mathbf{b}\|_q}{\|\mathbf{b}\|_q} \frac{\|\mathbf{b}\|_q}{\|\mathbf{x}\|_p \|A\|} + \frac{\|\delta A\|_{pq}}{\|A\|} \right)$$

Jeżeli jako $\|A\|$ przyjmiemy $\|A\|_{pq}$, to z równości $A\mathbf{x} = \mathbf{b}$ otrzymamy

$$\frac{\|\mathbf{b}\|_q}{\|\mathbf{x}\|_p \|A\|_{pq}} \leq 1, \text{ skąd}$$

$$\frac{\|\delta \mathbf{x}\|_p}{\|\mathbf{x}\|_p} \leq \frac{1}{1 - \varrho} \|A^{-1}\|_{qp} \|A\|_{pq} \left(\frac{\|\delta \mathbf{b}\|_q}{\|\mathbf{b}\|_q} + \frac{\|\delta A\|_{pq}}{\|A\|_{pq}} \right) \quad (5.21)$$

Powyższe oszacowanie jest dosyć ogólne, obowiązuje bowiem dla dowolnego zaburzenia $\delta \mathbf{b}$ oraz dowolnego zaburzenia δA , jeśli tylko jest ono dostatecznie małe (tzn. $\|\delta A\|_{pq} \leq \frac{\varrho}{\|A^{-1}\|_{qp}}$). Wskaźnik uwarunkowania $K_{pq} = \|A^{-1}\|_{qp} \|A\|_{pq}$ stanowi dość dobre przybliżenie wartości wzmocnienia zaburzenia δA , lecz może wielokrotnie przekraczać wzmocnienie zaburzenia $\delta \mathbf{b}$. Przypuszcza się, że K jest równe różniczkowemu wzmocnieniu zaburzenia $\frac{\|\delta A\|_{pq}}{\|A\|_{pq}}$, lecz nie zostało to udowodnione. Nierówności (5.16) i (5.21) umożliwiają oszacowanie błędu wejściowego, jeśli nieznane niedokładności danych A i $\mathbf{b}$ potraktujemy jako zaburzenia $\delta \mathbf{b}$ i δA .

W podobny sposób można oszacować także wpływ błędów zaokrągleń.

Zgodnie z interpretacją Wilkinsona (patrz p. 0.1), błędy zaokrągleń możemy zastąpić pewnym wstępnym zaburzeniem danych. Innymi słowy, można przyjąć, że podczas rozwiązywania układu równań wszystkie działania są wykonywane dokładnie, lecz zamiast układu $Ax = b$ rozpatrujemy układ $(A + \delta A^{(z)})x = b + \delta b^{(z)}$. Okazuje się, że można przyjąć $\delta b^{(z)} = 0$ przy odpowiednim wyborze zaburzenia $\delta A^{(z)}$. Przy takim podejściu można oszacować względny błąd $\frac{\|\delta x\|}{\|x\|}$ spowodowany obliczeniami na maszynie cyfrowej, jeżeli wiemy, ile dla danej metody wynosi zastępcze zaburzenie $\frac{\|\delta A^{(z)}\|}{\|A\|}$.

Przykład 3. W przypadku rozwiązywania równań (5.15) na maszynie, której słowo składa się z 2 cyfr dziesiętnych, otrzymamy

$$\begin{aligned} \text{fl}(x_1) &= \text{fl}\left(\frac{a_{22}b_1 - a_{12}b_2}{a_{11}a_{22} - a_{21}a_{12}}\right) = \frac{\text{fl}(a_{22}b_1 - a_{12}b_2)(1 + \varepsilon_1)}{\text{fl}(a_{11}a_{22} - a_{21}a_{12})} = \\ &= \frac{[\text{fl}(a_{22}b_1)(1 + \varepsilon_2) - \text{fl}(a_{12}b_2)(1 + \varepsilon_3)](1 + \varepsilon_1)}{\text{fl}(a_{11}a_{22})(1 + \varepsilon_4) - \text{fl}(a_{21}a_{12})(1 + \varepsilon_5)} = \\ &= \frac{a_{22}b_1(1 + \varepsilon_6)(1 + \varepsilon_2)(1 + \varepsilon_1) - a_{12}b_2(1 + \varepsilon_7)(1 + \varepsilon_3)(1 + \varepsilon_1)}{a_{11}a_{22}(1 + \varepsilon_8)(1 + \varepsilon_4) - a_{21}a_{12}(1 + \varepsilon_9)(1 + \varepsilon_5)} \end{aligned}$$

Ponieważ liczby $\varepsilon_i (i = 1, 2, \dots, 9)$ są rzędu 10^{-2} (patrz p. 0.1), więc

$$\begin{aligned} \text{fl}(x_1) &= \frac{a_{22}(1 + E_1)b_1 - a_{12}(1 + E_2)b_2}{a_{11}a_{22}(1 + E_3) - a_{21}a_{12}(1 + E_4)} = \\ &= \frac{a_{22}(1 + E_1)b_1 - a_{12}(1 + E_2)b_2}{a_{11}\frac{1 + E_3}{1 + E_1} \cdot a_{22}(1 + E_1) - a_{21}\frac{1 + E_4}{1 + E_2} \cdot a_{12}(1 + E_2)} \end{aligned}$$

przy czym liczby E_1 i E_2 są rzędu $3 \cdot 10^{-2}$, a E_3 i E_4 — rzędu $2 \cdot 10^{-2}$. Wynik $\text{fl}(x_1)$ jest więc taki, jakbyśmy rozwiązywali dokładnie układ równań

$$\begin{bmatrix} a_{11}\frac{1 + E_3}{1 + E_1} & a_{12}(1 + E_2) \\ a_{21}\frac{1 + E_4}{1 + E_2} & a_{22}(1 + E_1) \end{bmatrix} x = b$$

czyli układ z zaburzoną macierzą A o macierz

$$\delta A^{(z)} = \begin{bmatrix} \delta a_{11}^{(z)} & \delta a_{12}^{(z)} \\ \delta a_{21}^{(z)} & \delta a_{22}^{(z)} \end{bmatrix}$$

w której liczby $\delta a_{11}^{(z)}$ i $\delta a_{21}^{(z)}$ są rzędu $8 \cdot 10^{-2}$, a liczby $\delta a_{12}^{(z)}$ i $\delta a_{22}^{(z)}$ — rzędu $3 \cdot 10^{-2}$. Przy obliczaniu oszacowań liczb $\delta a_{11}^{(z)}$ i $\delta a_{21}^{(z)}$ korzystaliśmy z następującego lematu.

Lemat. Jeżeli $|\varepsilon| < \frac{1}{2}$, to $\frac{1}{1 + \varepsilon} = 1 + \eta$, gdzie $|\eta| < 2|\varepsilon|$.

DOWÓD. $\eta = \frac{-\varepsilon}{1+\varepsilon}$, więc $|\eta| = \frac{1}{|1+\varepsilon|} |\varepsilon|$, a ponieważ $|1+\varepsilon| > 1 - |\varepsilon| > \frac{1}{2}$, więc $|\eta| < 2|\varepsilon|$.

Analogicznie jak $\text{fl}(x_1)$, znajdujemy

$$\text{fl}(x_2) = \frac{a_{11}(1+E_5)b_2 - a_{21}(1+E_6)b_1}{a_{11}(1+E_5) \cdot a_{22} \frac{1+E_7}{1+E_5} - a_{21}(1+E_6) \cdot a_{12} \frac{1+E_8}{1+E_6}}$$

a więc znowu x_2 jest dokładnym rozwiązaniem pewnego układu równań o macierzy A , zaburzonego o macierz $\delta A^{(z)}$, przy czym tym razem liczby $\delta a_{11}^{(z)}$ i $\delta a_{21}^{(z)}$ są rzędu $3 \cdot 10^{-2}$, a liczby $\delta a_{12}^{(z)}$ i $\delta a_{22}^{(z)}$ są rzędu $8 \cdot 10^{-2}$. W obu przypadkach oszacowanie zaburzenia $\frac{\|\delta A^{(z)}\|_\infty}{\|A\|_\infty}$ jest jednakowe i dla równań (5.15) wynosi $\frac{11 \cdot 10^{-2}}{1,69} < 7 \cdot 10^{-2}$.

Zwróćmy uwagę, że $\text{fl}(x_1)$ i $\text{fl}(x_2)$ potraktowaliśmy jako dokładne rozwiązania dwu różnych zaburzonych układów równań. Ponieważ jednak podaliśmy wspólne oszacowanie wielkości zaburzeń w obu przypadkach, wielkość zmiany rozwiązania można oszacować posługując się wskaźnikiem uwarunkowania: $K_\infty = \|A\|_\infty \|A^{-1}\|_\infty = 1,69 \cdot 338 = 571,22$, $\|x\|_\infty = 0,80$, zatem przy niekorzystnym zaburzeniu można spodziewać się błędu

$$\|\delta x\|_\infty \approx 571,22 \cdot 7 \cdot 10^{-2} \cdot 0,80 \approx 30$$

Tymczasem uzyskany błąd wyniósł „zaledwie” $\|\delta x\|_\infty = 0,80$.

K o m e n t a r z. Powyższy przykład ilustruje sygnalizowaną wcześniej własność wskaźnika uwarunkowania, iż prognoza błędów zaokrągleń jest przeprowadzona ze zbyt dużym nadmiarem. Przyczyny tego są dwie: po pierwsze oszacowanie (5.21) dotyczy najgorszego przypadku, tzn. zaburzenia $\delta A^{(z)}$ szczególnie wpływającego na zmianę rozwiązania. Z tego względu wskaźnik uwarunkowania jest zbyt duży dla większości zaburzeń $\delta A^{(z)}$. Po drugie zaś oszacowanie $\frac{\|\delta A^{(z)}\|_\infty}{\|A\|_\infty}$ również zostało przyjęte jako najgorsze z możliwych. Przyjmowano bowiem, że w każdym działaniu wystąpi zaokrąglenie (a przy odejmowaniu nie wystąpiło ono ani razu) oraz przyjęto, że wszystkie błędy ε_i , $i = 1, 2, \dots, 9$, są tego samego znaku i zastąpiono np. iloczyn $(1+\varepsilon_6)(1+\varepsilon_2)(1+\varepsilon_1)$ liczbą $1+E_1$, przyjmując $E_1 \approx \varepsilon_6 + \varepsilon_2 + \varepsilon_1$, a zatem jako liczbę rzędu $3 \cdot 10^{-2}$. Należy zatem pamiętać, że podawane w dalszej części rozdziału oszacowania błędów w poszczególnych metodach są na ogół zawyżone. Sens tego rodzaju oszacowań leży w tym, że potrafimy zagwarantować rozwiązanie danego układu równań zadaną dokładnością. Na przykład, jeśli w powyższym przykładzie przyjmiemy $t = 5$ miejsc dziesiętnych, to gwarantujemy, iż błąd $\frac{\|\delta x\|_\infty}{\|x\|_\infty}$ będzie mniejszy niż 4% (w rzeczywistości wtedy $\|\delta x\| = 0$).

Ćwiczenia

1. Oszacować błąd rozwiązania układu równań

$$\begin{aligned} \sqrt{2}x_1 + x_2 &= 0,50 \\ x_1 + \sqrt{2}x_2 &= 0,50 \end{aligned}$$

uzyskanego na maszynie cyfrowej o dwu miejscach dziesiętnych.

U w a g a. Wystąpią tu błędy zaokrągleń i błąd wejściowy, gdyż zamiast $\sqrt{2}$ możemy używać jedynie $0,14 \cdot 10^{+1}$.

Jeżeli macierz A układu n równań z n niewiadomymi $Ax = b$ jest macierzą trójkątną (górną lub dolną)¹⁾, to rozwiązanie x takiego układu można uzyskać wykonując małą liczbę działań arytmetycznych i przy małych błędach zaokrągleń. Ustalmy, że A jest macierzą trójkątną górną i aby istniało jednoznaczne rozwiązanie założmy, że wszystkie elementy na głównej diagonalu tej macierzy są różne od zera. Mamy zatem układ równań postaci

¹⁾ Macierz nazywamy trójkątną dolną (górną), jeżeli wszystkie elementy nad (pod) diagonalą są równe zeru.

(5.26)

$$\|\delta \mathbf{A}^{(z)}\|_1 \leq \varepsilon (2n + 1) g \quad (5.27)$$
$$\|\delta \mathbf{A}^{(z)}\|_{\infty} \leq \varepsilon (2n + 2) g \quad (5.28)$$

Ćwiczenia

1. Uzasadnić tezę, że nie opłaci się wstępnie dzielić równań przez elementy a_{ii} . W s k a - z ó w k a. Obliczyć liczbę mnożeń oraz dzielen i oszacować zastępcze zaburzenie $\delta \mathcal{A}^{(z)}$.
2. Pokazać, że przy wykonywaniu mnożeń i dzielen i ze zwiększoną dokładnością (np. w podwójnej precyzji) przy zachowaniu pojedynczej precyzji przy odejmowaniu — polepsza się jedynie nieznacznie oszacowanie (5.24).

W punkcie tym omówimy najczęściej stosowaną metodę do numerycznego rozwiązywania układów równań liniowych, metodę eliminacji Gaussa. Omówimy tu ponadto rozkład LU , metodę Doolittle'a, warianty stabilne metody eliminacji, błędy zaokrągleń, organizację programu (metoda Doolittle'a została włączona do tego punktu ze względu na jej bezpośredni związek z metodą eliminacji Gaussa).

Jeśli rozwiązujemy układ równań z pełną macierzą kwadratową A , to można rozwiązać go w dwóch etapach: sprowadzając układ do postaci trójkątnej (5.22), a następnie stosując wzory (5.23).

Etap pierwszy (zwany *etapem eliminacji zmiennych*). Dany jest układ równań $A^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$, postaci

(5.29)

Odejmując od i -tego wiersza tego układu, $i = 2, 3, \dots, n$, wiersz pierwszy pomnożony przez $a_{i1}^{(1)}/a_{11}^{(1)}$, otrzymujemy układ $A^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$ postaci

$$a_{11}^{(2)}x_1 + a_{12}^{(2)}x_2 + \dots + a_{1n}^{(2)}x_n = b_1^{(2)}$$

$$a_{22}^{(2)}x_2 + \dots + a_{2n}^{(2)}x_n = b_2^{(2)}$$

.....

$$a_{n2}^{(2)}x_2 + \dots + a_{nn}^{(2)}x_n = b_n^{(2)}$$

Wyeliminowaliśmy w ten sposób pierwszą niewiadomą x_1 z równań leżących w wierszach o numerach $i = 2, 3, \dots, n$. Podobnie eliminujemy x_2 z równań leżących w wierszach $3, 4, \dots, n$, odejmując od i -tego wiersza ($i = 3, 4, \dots, n$) wiersz drugi pomnożony przez $a_{i2}^{(2)}/a_{22}^{(2)}$. Postępując w ten sposób otrzymamy kolejne, przekształcone układy równań $A^{(3)}\mathbf{x} = \mathbf{b}^{(3)}$, $A^{(4)}\mathbf{x} = \mathbf{b}^{(4)}$, ..., i po $(n-1)$ eliminacjach uzyskamy trójkątny układ równań $A^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$ postaci

$$a_{11}^{(n)}x_1 + a_{12}^{(n)}x_2 + \dots + a_{1n}^{(n)}x_n = b_1^{(n)}$$

$$a_{22}^{(n)}x_2 + \dots + a_{2n}^{(n)}x_n = b_2^{(n)}$$

.....

$$a_{nn}^{(n)}x_n = b_n^{(n)}$$

Etap drugi (zwany *postępowaniem odwrotnym*). Otrzymany po przekształceniu układ rozwiązujemy zgodnie ze wzorami (5.23).

Opisany powyżej sposób rozwiązywania układu (5.29) nazywamy *metodą eliminacji Gaussa*.

Można obliczyć, że w celu wyznaczenia tą metodą rozwiązania $\mathbf{x}$, należy wykonać łącznie $M = \frac{1}{3}n^3 + n^2 - \frac{1}{3}n$ mnożeń i dzielen oraz $D = \frac{1}{3}n^3 + \frac{1}{2}n^2 - \frac{5}{6}n$ dodawań.

Jeśli nie musimy zachować danych $A^{(1)}$ i $\mathbf{b}^{(1)}$ do dalszych obliczeń lub sprawdzenia, to minimalna liczba zajętych komórek pamięci $P = n^2 + n$, w przeciwnym zaś przypadku $P' = 2P$. Należy podkreślić, że liczba wykonywanych działań w metodzie eliminacji Gaussa jest bez porównania mniejsza niż w metodzie wyznaczania rozwiązań za pomocą wzorów Cramera. Tam bowiem należało obliczyć $(n+1)$ wyznaczników, które w przypadku stosowania rozwijania wyznacznika według kolumny lub wiersza wymagają co najmniej $(n+1)!$ mnożeń. Zatem w przypadku 15 równań należałoby wykonać $4922789888000 \approx 5 \cdot 10^{12}$ mnożeń i maszyna cyfrowa wykonująca 10^6 mnożeń na sekundę musiałaby pracować nieprzerwanie ponad rok, natomiast w metodzie eliminacji Gaussa $M = 1345$ mnożeń, a więc maszyna wykona pracę w niecałe 0,01 sekundy.

Rozkład LU. W wielu zagadnieniach numerycznych dotyczących macierzy kwadratowych A , poszukujemy takich macierzy trójkątnych L i U , by $A = LU$. Metoda eliminacji Gaussa umożliwia wyznaczenie takiego rozkładu.

Zauważmy, że przekształcenie układu $A^{(1)}x = b^{(1)}$ do postaci $A^{(2)}x = b^{(2)}$ jest równoważne pomnożeniu obu stron układu (5.29) przez macierz

$$L^{(1)} = \begin{bmatrix} 1 & & & & \\ -l_{21} & 1 & & & \\ -l_{31} & 0 & 1 & & \\ \dots & \dots & \dots & \ddots & \\ -l_{n1} & 0 & 0 & \dots & 1 \end{bmatrix}, \quad l_{i1} = a_{i1}^{(1)}/a_{11}^{(1)}, \quad i = 2, 3, \dots, n$$

Mamy zatem dwie równości: $L^{(1)}A^{(1)} = A^{(2)}L^{(1)}b^{(1)} = b^{(2)}$.

Analogicznie otrzymujemy dwie następne równości: $L^{(2)}A^{(2)} = A^{(3)}$, $L^{(2)}b^{(2)} = b^{(3)}$, gdzie

$$L^{(2)} = \begin{bmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & -l_{32} & 1 & & \\ \dots & \dots & \dots & \ddots & \\ 0 & -l_{n2} & & \dots & 1 \end{bmatrix}, \quad l_{i2} = a_{i2}^{(2)}/a_{22}^{(2)}, \quad i = 3, 4, \dots, n$$

Postępując dalej w ten sposób otrzymujemy ostatecznie

$$L^{(n-1)}L^{(n-2)}\dots L^{(1)}A^{(1)} = A^{(n)}, \quad L^{(n-1)}L^{(n-2)}\dots L^{(1)}b^{(1)} = b^{(n)}$$

Macierze $L^{(i)}$, $i = 1, \dots, n-1$, są nieosobliwe, możemy zatem napisać

$$A^{(1)} = (L^{(1)})^{-1}(L^{(2)})^{-1}\dots(L^{(n-1)})^{-1}A^{(n)} \quad (5.30)$$

Zauważmy, że

$$(L^{(1)})^{-1} = \begin{bmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ l_{31} & 0 & 1 & & \\ \dots & \dots & \dots & \ddots & \\ l_{n1} & 0 & 0 & \dots & 1 \end{bmatrix}, \quad (L^{(2)})^{-1} = \begin{bmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & l_{32} & 1 & & \\ \dots & \dots & \dots & \ddots & \\ 0 & l_{n2} & 0 & \dots & 1 \end{bmatrix} \quad \text{itd.}$$

a ponadto

$$(\mathbf{L}^{(1)})^{-1}(\mathbf{L}^{(2)})^{-1} \dots (\mathbf{L}^{(n-1)})^{-1} = \begin{bmatrix} 1 & & & & 0 \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \dots & \dots & \dots & \dots & \cdot \\ l_{n1} & l_{n2} & l_{n3} & \dots & 1 \end{bmatrix}$$

Jeśli powyższą macierz oznaczmy symbolem $\mathbf{L}$, a macierz $\mathbf{A}^{(n)}$ symbolem $\mathbf{U}$, to rozkład (5.30) macierzy $\mathbf{A}^{(1)}$ na iloczyn możemy zapisać następująco

$$\mathbf{A}^{(1)} = \mathbf{L}\mathbf{U} \quad (5.30a)$$

gdzie $\mathbf{L}$ jest macierzą trójkątną dolną z jedynkami na diagonalu, a $\mathbf{U}$ — macierzą trójkątną górną.

U w a g a. Mogą istnieć nieosobliwe macierze $\mathbf{A}$, których nie da się rozłożyć na iloczyn $\mathbf{L}\mathbf{U}$. Na przykład, gdy $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$ to nie da się znaleźć takich macierzy

$$\mathbf{L} = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix} \quad \text{ i } \quad \mathbf{U} = \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}, \quad \text{ by } \quad \mathbf{A} = \mathbf{L}\mathbf{U}.$$

Z przeprowadzonych rozważań wynika, że jeżeli etap eliminacji zmiennych metodą Gaussa można wykonać do końca, to układ równań $\mathbf{A}\mathbf{x} = \mathbf{b}$ można zapisać jako $\mathbf{L}\mathbf{U}\mathbf{x} = \mathbf{b}$. Etap ten jest zatem równoważny przekształceniu wyjściowego układu do postaci $\mathbf{U}\mathbf{x} = \mathbf{L}^{-1}\mathbf{b} = \mathbf{b}^{(n)}$, a wyznaczenie rozwiązania $\mathbf{x}$ polega na rozwiązaniu dwu układów z macierzami trójkątnymi

$$\mathbf{L}\mathbf{y} = \mathbf{b}, \quad \mathbf{U}\mathbf{x} = \mathbf{y}$$

Jeśli znamy rozkład macierzy $\mathbf{A}$ na iloczyn $\mathbf{L}\mathbf{U}$, to dla wyznaczenia $\mathbf{x}$ wystarczy wykonać $M = \frac{1}{2}n(n-1) + \frac{1}{2}n(n+1) = n^2$ mnożeń i dzielen oraz $D = n^2 - n$ dodawań, a więc tyle samo co przy obliczaniu $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. Rozwiązując układ równań $\mathbf{A}\mathbf{x} = \mathbf{b}$ przy kilku różnych wektorach $\mathbf{b}$, możemy zatem zmniejszyć liczbę działań, zapamiętując zarówno macierz $\mathbf{U}$, jak i macierz $\mathbf{L}$ ułożoną z obliczanych ilorazów $a_{ik}^{(k)}/a_{kk}^{(k)}$, $i = k+1, k+2, \dots, n$. Warto zauważyć, że elementy l_{ij} , $j < i$, można wpisać w miejsce elementów a_{ij} bezpośrednio po obliczeniu. Oczywiście $l_{ii} = 1$, $i = 1, 2, \dots, n$, więc elementów na diagonalu nie musimy zapamiętywać. Po zakończeniu etapu eliminacji uzyskamy wówczas macierze $\mathbf{L}$ i $\mathbf{U}$ w tych komórkach pamięci maszyny, w których była uprzednio zapamiętana macierz $\mathbf{A}$, tzn.

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \Rightarrow \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ l_{21} & u_{22} & \dots & u_{2n} \\ \dots & \dots & \dots & \dots \\ l_{n1} & l_{n2} & \dots & u_{nn} \end{bmatrix}$$

Metoda Doolittle'a. Rozkład LU można uzyskać nie tylko metodą eliminacji Gaussa, ale także traktując równość $A = LU$ jako układ n^2 równań z n^2 niewiadomymi l_{ij} dla $i > j$ i niewiadomymi u_{ij} dla $i \leq j$.

Równania te wygodnie jest rozwiązywać na przemian wierszami i kolumnami. Pokażemy to na konkretnym przykładzie.

Przykład 1. Mamy

$$\begin{bmatrix} 3 & 3 & 0 \\ 0 & 2 & 2 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \cdot \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

Stąd

$$3 = u_{11}, \quad 3 = u_{12}, \quad 0 = u_{13}$$

$$0 = l_{21} \cdot 3 \quad 1 = l_{31} \cdot 3$$

$$2 = 0 \cdot 3 + u_{22}, \quad 2 = 0 \cdot 0 + u_{23}$$

$$0 = \frac{1}{3} \cdot 3 + l_{32} \cdot 2$$

$$1 = \frac{1}{3} \cdot 0 + \left(-\frac{1}{2}\right) \cdot 2 + u_{33}$$

i ostatecznie

$$\begin{bmatrix} 3 & 3 & 0 \\ 0 & 2 & 2 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{3} & -\frac{1}{2} & 1 \end{bmatrix} \cdot \begin{bmatrix} 3 & 3 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{bmatrix}$$

W przypadku ogólnym elementy l_{ij} oraz u_{ij} obliczamy następująco: dla $i = 1, 2, \dots, n$ obliczamy kolejno

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj}, \quad j = i, i+1, \dots, n \quad (5.31)$$

$$l_{ji} = \left(a_{ji} - \sum_{k=1}^{i-1} l_{jk} u_{ki} \right) / u_{ii}, \quad j = i+1, i+2, \dots, n$$

Przedstawiona metoda obliczania rozkładu LU nosi nazwę *metody Doolittle'a*. Liczba działań wynosi w niej $M = \frac{1}{3}n^3 - \frac{1}{3}n$, $D = \frac{1}{3}n^3 - \frac{1}{2}n^2 + \frac{1}{6}n$. Rozwiązując następnie układy równań $Ly = b$ i $Ux = y$, łącznie

otrzymujemy $M = \frac{1}{3}n^3 + n^2 - \frac{1}{3}n$, $D = \frac{1}{3}n^3 + \frac{1}{3}n^2 - \frac{5}{6}n$ działań, a więc tyle samo co w metodzie Gaussa.

Stosując wzory (5.31) można użyć podwójnej precyzji do obliczania u_{ij} oraz l_{ij} , a następnie liczby te zredukować ponownie do pojedynczej precyzji przed ostatecznym zapamiętaniem. Podwyższy to istotnie dokładność uzyskanego rozkładu LU , bez dodatkowego obciążenia pamięci maszyny. Jeśli natomiast działania będziemy wykonywać w takiej samej precyzji jak w metodzie Gaussa, to wszystkie błędy zaokrągleń będą identyczne w obu metodach.

Tak jak w metodzie Gaussa, elementy u_{ij} oraz l_{ij} można również zapamiętywać w komórkach pamięci poprzednio zajętych przez macierz A . Poniższy fragment programu realizuje tę wersję algorytmu (zakładamy, że w bloku programu zostały zadeklarowane zmienne $I, I1, J, K$ i N typu integer oraz zmienna A typu tablicowego)

```

.....
FOR J := 2 TO N DO
  A[J,1] := A[J,1]/A[1,1];
FOR I := 2 TO N DO
  BEGIN
    I1 := I - 1;
    FOR J := 1 TO N DO
      FOR K := 1 TO I1 DO
        A[I,J] := A[I,J] - A[I,K]*A[K,J];
      FOR J := I+1 TO N DO
        BEGIN
          FOR K := 1 TO I1 DO
            A[J,I] := A[J,I] - A[J,K]*A[K,I];
          A[J,I] := A[J,I]/A[I,I];
        END;
      END;
    END;
  END;

```

Warianty stabilne. Zarówno metoda Gaussa, jak i metoda Doolittle'a nie są niezawodne. Pokażemy to na przykładach.

Przykład 2. Rozważmy układ równań

$$\begin{bmatrix} 0 & 2 & 2 \\ 3 & 3 & 0 \\ 1 & 0 & 1 \end{bmatrix} x = \begin{bmatrix} b_2 \\ b_1 \\ b_3 \end{bmatrix}$$

uzyskany przez przestawienie dwu pierwszych wierszy układu równań z poprzedniego przykładu.

Macierz tego układu jest nieosobliwa, istnieje zatem dokładnie jedno rozwiązanie. Układu tego nie można jednak sprowadzić do postaci trójkątnej, ponieważ $a_{11} = u_{11} = 0$.

W podanym powyżej przypadku stosuje się pewną modyfikację metody eliminacji, zwaną *częściowym wyborem elementu podstawowego*. *Elementem podstawowym* nazywamy ten element macierzy A , za pomocą którego eliminujemy zmienną z dalszych równań. Dotychczas jako elementy pod-

stawowe w macierzach $A^{(k)}$, $k = 1, 2, \dots, n$, wybieraliśmy elementy leżące na diagonalu, $a_{kk}^{(k)}$. Stosując częściowy wybór elementu podstawowego wybieramy ten z elementów k -tej kolumny w k -tej macierzy, który ma największy moduł. Przez zmianę kolejności wierszy w macierzy można uzyskać element podstawowy leżący na diagonalu.

Przykład 3. Mamy

$$\begin{bmatrix} 0 & 2 & 2 \\ 1 & 0 & 1 \\ 3 & 3 & 0 \end{bmatrix} x = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \Rightarrow \begin{bmatrix} 3 & 3 & 0 \\ 1 & 0 & 1 \\ 0 & 2 & 2 \end{bmatrix} x = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 3 & 3 & 0 \\ 0 & -1 & 1 \\ 0 & 2 & 2 \end{bmatrix} x = \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix} \Rightarrow$$

$$\Rightarrow \begin{bmatrix} 3 & 3 & 0 \\ 0 & 2 & 2 \\ 0 & -1 & 1 \end{bmatrix} x = \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 3 & 3 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{bmatrix} x = \begin{bmatrix} 3 \\ 1 \\ 1,5 \end{bmatrix}$$

U w a g a. Przy obliczeniach na maszynie cyfrowej przestawianie wierszy nie jest konieczne (patrz uwagi na końcu tego podrozdziału).

Łatwo stwierdzić, że wersja algorytmu Gaussa z częściowym wyborem elementu podstawowego, zwana czasem *metodą Gaussa-Crouta*, którą w dalszej części będziemy oznaczać GCW, jest metodą niezawodną, tzn. przy założeniu dokładnych działań arytmetycznych nie nastąpi zatrzymanie procesu obliczeń z powodu dzielenia przez zero w przypadku, gdy istnieje jednoznaczne rozwiązanie układu równań. Gdyby bowiem jakiś wybrany element podstawowy był równy zero, oznaczałoby to liniową zależność wierszy, z których wybieraliśmy ten element, co przy założeniu nieosobliwości macierzy A jest niemożliwe. Częściowy wybór elementu podstawowego stosuje się również w metodzie Doolittle'a, w celu zapewnienia niezawodności tej metody.

Metodę GCW można stosować do równoczesnego rozwiązywania wielu układów równań różniących się jedynie wektorami wyrazów wolnych. Przekształcenia dokonywane na wierszach (mnożenie przez liczby i dodawanie) wykonuje się wtedy dla całej macierzy dołączonej, składającej się z kolumn macierzy A i kolumn wyrazów wolnych. W szczególności można w ten sposób wyznaczyć macierz odwrotną A^{-1} , gdy równocześnie rozwiązujemy n układów równań z wektorami wyrazów wolnych, równymi kolumnom macierzy jednostkowej. Macierz A^{-1} jest utworzoną wtedy przez kolejne wektory rozwiązań, stanowiące jej kolumny.

Należy zaznaczyć, że metoda GCW nie zawsze prowadzi do rozkładu $A = LU$.

Przykład 4. W przykładzie 3 metoda GCW określała macierze trójkątne

$$U = \begin{bmatrix} 3 & 3 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{bmatrix} \quad \text{oraz} \quad L = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{3} & -\frac{1}{2} & 1 \end{bmatrix}$$

natomiast rozważana macierz A nie jest równa iloczynowi LU , ponieważ

$$\begin{bmatrix} 0 & 2 & 2 \\ 1 & 0 & 1 \\ 3 & 3 & 0 \end{bmatrix} \neq \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{3} & -\frac{1}{2} & 1 \end{bmatrix} \cdot \begin{bmatrix} 3 & 3 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{bmatrix}$$

Otrzymane macierze L i U stanowią bowiem rozkład macierzy A , którą uzyskalibyśmy wykonując wstępnie wszystkie przestawienia wierszy, które wykonaliśmy w trakcie eliminacji zmiennych, czyli

$$\tilde{A} = \begin{bmatrix} 3 & 3 & 0 \\ 0 & 2 & 2 \\ 1 & 0 & 1 \end{bmatrix}$$

W celu dalszej analizy własności metody GCW zwróćmy uwagę na wielkość liczb obliczanych w trakcie przekształcania macierzy A . Metoda GCW gwarantuje, iż wszystkie elementy macierzy L nie przekraczają co do modułu liczby 1. Ponadto, ponieważ podczas kolejnych eliminacji zmiennych elementy macierzy przekształcamy zgodnie z zależnością

$$a_{ji}^{(k+1)} = a_{ji}^{(k)} - l_{jk} a_{kk}^{(k)} \quad \text{dla} \quad j > k, i > k$$

więc elementy kolejnych macierzy $A^{(1)}, A^{(2)}, \dots, A^{(n)}$ nie wzrastają szybciej niż 2 razy w każdej eliminacji, stąd

$$\|A^{(k+1)}\|_{1\infty} \leq 2 \cdot \|A^{(k)}\|_{1\infty}$$

Dokładne rozważania dotyczące szybkości wzrostu elementów dla metody GCW można znaleźć np. w [68] i [38].

Błędy zaokrągleń. Zgodnie ze spostrzeżeniem, iż metody Doolittle'a i Gaussa są równoważne numerycznie w tej samej precyzji, wyznaczenie rozwiązania x tymi metodami można rozłożyć na trzy etapy:

- (i) wyznaczenie rozkładu $\tilde{A} = LU$ metodą Gaussa lub Doolittle'a
- (ii) znalezienie rozwiązania układu $Ly = b$
- (iii) znalezienie rozwiązania układu $Ux = y$

W etapie (i) na skutek błędów zaokrągleń wyznaczone L i U spełniają równość $LU = \tilde{A} + E$, gdzie E jest błędem rozkładu. W etapach (ii) oraz (iii) znalezione rozwiązania y i x są dokładnymi rozwiązaniami układów $(L + \delta L)y = b$, $(U + \delta U)x = y$. Stąd $(L + \delta L)(U + \delta U)x = (L + \delta L)y = b$, czyli $(LU + L\delta U + \delta LU + \delta L\delta U)x = b$, a więc x jest rozwiązaniem dokładnym układu

$$(A + E + L\delta U + \delta LU + \delta L\delta U)x = b$$

Należy zatem wyznaczyć oszacowanie normy zaburzenia

$$\delta A^{(z)} = E + L\delta U + \delta LU + \delta L\delta U$$

Można wykazać [32], że elementy macierzy E spełniają warunki

$$|e_{ij}| \leq 3,01 \varepsilon n g_{ij} \quad (5.32a)$$

gdzie $g_{ij} = \max_{k=1,2,\dots,n} |a_{ij}^{(k)}|$. Oszacowanie to obowiązuje przy dowolnej strategii

wyboru elementów podstawowych. Dla metody GCW można ponadto wykazać (patrz np. [68, s. 153]), że

$$\begin{aligned} \|E\|_1 &\leq \varepsilon (n^2 - n) g \\ \|E\|_\infty &\leq \varepsilon (n^2 + n - 2) g \end{aligned} \quad (5.32b)$$

przy czym ε , jak zwykle, jest równe 2^{-t} lub 10^{-t} , a g oznacza największy moduł elementów macierzy $\tilde{A} = A^{(1)}, A^{(2)}, \dots, A^{(n)} = U$. W metodzie GCW mamy $g \leq 2^{n-1} \|A\|_{1\infty}$. Okazuje się jednak, że na ogół elementy macierzy A nie wzrastają tak szybko i można przyjąć w praktyce $g \leq 8 \|\tilde{A}\|_{1\infty}$. Ponieważ $\|\tilde{A}\|_1 \geq \|\tilde{A}\|_{1\infty}$ i $\|\tilde{A}\|_\infty \geq \|\tilde{A}\|_{1\infty}$, więc można przyjąć oszacowania

$$\frac{\|E\|_1}{\|A\|_1} \leq 8\varepsilon (n^2 - n) \quad (5.32c)$$

$$\frac{\|E\|_\infty}{\|A\|_\infty} \leq 8\varepsilon (n^2 + n - 2)$$

Wobec nierówności (5.26) po uwzględnieniu, że $|l_{ij}| \leq 1$, otrzymujemy

$$\|\delta L\|_1 \leq \varepsilon \left(\frac{1}{4}n^2 + \frac{5}{2}n - 1 \right)$$

$$\|\delta L\|_\infty \leq \varepsilon \left(\frac{1}{2}n^2 + \frac{5}{2}n - 3 \right)$$

$$\frac{\|\delta U\|_1}{\|A\|_1} \leq \varepsilon \left(\frac{1}{4}n^2 + \frac{5}{2}n \right)$$

$$\frac{\|\delta U\|_\infty}{\|A\|_\infty} \leq \varepsilon \left(\frac{1}{2}n^2 + \frac{5}{2}n - 2 \right)$$

Ponadto $\|U\|_1 \leq ng$ i $\|U\|_\infty \leq ng$, zatem np. dla normy $\|\cdot\|_\infty$ mamy

$$\frac{\|\delta A^{(z)}\|_\infty}{\|A\|_\infty} \leq \frac{\|E\|_\infty}{\|\tilde{A}\|_\infty} + \|L\|_\infty \frac{\|\delta U\|_\infty}{\|\tilde{A}\|_\infty} + \|\delta L\|_\infty \frac{\|U\|_\infty}{\|\tilde{A}\|_\infty} + o(\varepsilon)$$

skąd

$$\frac{\|\delta A^{(z)}\|_\infty}{\|A\|_\infty} \leq \varepsilon \left(\frac{9}{2}n^3 + \frac{61}{2}n^2 - 18n - 16 \right) + o(\varepsilon) \quad (5.33)$$

Ostatecznie, uwzględniając (5.21), otrzymujemy oszacowanie

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \leq \frac{\alpha}{1 - \alpha}, \quad \text{gdzie } \alpha = \varepsilon \cdot K_\infty \cdot O\left(\frac{9}{2}n^3\right) \quad (5.34)$$

Liczba K_∞ jest to wskaźnik uwarunkowania, równy $\|A\|_\infty \cdot \|A^{-1}\|_\infty$, a $O\left(\frac{9}{2}n^3\right)$ oznacza wielomian $\frac{9}{2}n^3 + \frac{61}{2}n^2 - 18n - 16$.

W różnych pracach można znaleźć oszacowania nieco inne niż oszacowanie (5.34), co wynika ze sposobu oszacowania zaburzeń, ale nie zmienia to istoty nierówności (5.34), która polega na tym, że błąd $\frac{\|\delta x\|_\infty}{\|x\|_\infty}$ można dowolnie zmniejszyć przez zmniejszenie ε (a więc przez powiększenie dokładności obliczeń). Algorytmy o tej własności będziemy nazywać *stabilnymi*.

Przykład 5. Ile równań można rozwiązać metodą GCW na maszynie cyfrowej pracującej w układzie dwójkowym z długością słowa $t = 33$, jeśli zakładamy, że $K_\infty \leq 10^6$? Dopuszczalny błąd w rozwiązaniu może wynosić najwyżej 10%.

Mamy

$$\alpha = \varepsilon K \cdot O\left(\frac{9}{2}n^3\right) \leq \frac{0,1}{1+0,1} \Leftarrow \frac{9}{2}n^3 + \frac{61}{2}n^2 - 18n - 16 \leq 10^{-6} \cdot 10^{10} \cdot 0,09 = 900 \Leftarrow n \leq 4$$

Można sprawdzić, że dla $n = 5$ uzyskamy $\frac{\alpha}{1-\alpha} > 0,1$, a wówczas oszacowanie (5.34) nie gwarantuje dobrego rozwiązania.

Tak silne ograniczenie liczby rozwiązywanych równań w powyższym przykładzie zostało spowodowane przede wszystkim przyjęciem dużej wartości wskaźnika uwarunkowania, $K \approx 10^6$. Że nie jest to przesadnie duża wartość dowodzi przykład tzw. *macierzy Hilberta* $A = \left[\frac{10}{i+j} \right]_{5 \times 5}$, której wskaźnik uwarunkowania $K_\infty > 1,5 \cdot 10^6$. Na szczęście, jak już wspominaliśmy poprzednio, oszacowanie (5.34) jest na ogół bardzo zawyżone. Jest to spowodowane własnościami wskaźnika uwarunkowania, a także zbyt zawyżonym oszacowaniem błędów zaokrągleń. Jeśli uwzględnić stochastyczny charakter błędów zaokrągleń, można przyjmować z dużym prawdopodobieństwem zamiast czynnika $O\left(\frac{9}{2}n^3\right)$ we wzorze (5.34) czynnik $\sqrt{O\left(\frac{9}{2}n^3\right)}$ (patrz [50 p. 1.4]); dla danych z przykładu można wówczas rozwiązać co najmniej 60 równań z 60 niewiadomymi. Oszacowanie (5.34) uzasadnia jednak tezę, że istnieje ograniczenie liczby równań możliwych do rozwiązania metodą GCW na konkretnej maszynie cyfrowej. Analogiczne oszacowania uzyskuje się także dla innych metod dokładnych.

Organizacja programu. Proces eliminacji zmiennych lub wyznaczanie rozkładu LU jest łatwo zaprogramować. Istotna trudność leży w umiejętnym wyborze elementów podstawowych. W szczególnie niekorzystnym przypadku przy zamianie kolejności wierszy w celu wyznaczenia elementów podstawowych, musimy wykonać $3(n-1)$ odczytów i zapisywać wierszy

o łącznej liczbie $\frac{3}{2}n^2 + \frac{9}{2}n - 6$ elementów. Istnieją co prawda wersje algorytmu GCW, w których unikamy zamiany wierszy, zapamiętując jedynie położenie elementów podstawowych (patrz np. [50 p. 9.3.3]), ale skrupulatne prześledzenie liczby odczytów i zapisywań wierszy dowodzi, że nie zmniejszamy w ten sposób liczby działań, a tylko komplikujemy w znacznym stopniu prosty algorytm. Dlatego w obecnie stosowanych algorytmach dokonuje się zamiany wierszy zarówno w metodzie GCW, jak i w metodzie Doolittle'a. Ponieważ w metodzie Doolittle'a na skutek nieumiejętnie prowadzonych obliczeń można niepotrzebnie zwiększyć liczbę wykonywanych działań, omówimy najważniejsze zagadnienia związane z wyborem elementu podstawowego.

Przykład 6. Obliczyć rozkład LU macierzy

$$A = \begin{bmatrix} 20 & 31 & 23 \\ 30 & 24 & 18 \\ 15 & 32 & 21 \end{bmatrix}$$

Aby skorzystać z metody Doolittle'a z częściowym wyborem elementu podstawowego wprowadzimy dodatkową kolumnę roboczą D i dodatkową kolumnę wskaźników przestawiania wierszy — na początku w kolumnie tej ustawiamy kolejne liczby naturalne 1, 2, ..., n (u nas 1, 2, 3). Mamy zatem

$$\begin{bmatrix} 20 & 31 & 23 \\ 30 & 24 & 18 \\ 15 & 32 & 21 \end{bmatrix} \quad \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

Wybór elementu podstawowego polega na takiej zmianie kolejności wierszy macierzy, by obliczony element u_{ii} , występujący w drugim ze wzorów (5.31), miał największą wartość. Oczywiście, ewentualne przestawienie dotyczy wierszy $i, i+1, \dots, n$ ($i = 1, 2, \dots, n-1$).

$i = 1$. Jeśli na pierwszej pozycji ustawimy wiersz pierwszy, drugi lub trzeci, to uzyskamy jako u_{11} odpowiednio liczby: $d_1 = 20, d_2 = 30, d_3 = 15$. Ponieważ największą z tych liczb jest 30, więc powinniśmy dokonać zamiany pierwszego i drugiego wiersza w macierzach A, D i w kolumnie wskaźników. Otrzymamy zatem macierze

$$\begin{bmatrix} 30 & 24 & 18 \\ 20 & 31 & 23 \\ 15 & 32 & 21 \end{bmatrix} \quad \begin{bmatrix} 30 \\ 20 \\ 15 \end{bmatrix} \quad \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$$

Korzystając z wzorów (5.31), znajdujemy

$$u_{11} = d_1 = 30, \quad u_{12} = a_{12} = 24, \quad u_{13} = a_{13} = 18$$

$$l_{21} = d_2/d_1 = \frac{2}{3}, \quad l_{31} = d_3/d_1 = \frac{1}{2}$$

Wartości te wpisujemy w odpowiednie miejsca macierzy A , uzyskując

$$\begin{bmatrix} 30 & 24 & 18 \\ 2 & & \\ 3 & 31 & 23 \\ 1 & & \\ 2 & 32 & 21 \end{bmatrix} \quad \begin{bmatrix} 30 \\ 20 \\ 15 \end{bmatrix} \quad \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$$

$i = 2$. Jeśli na drugiej pozycji ustawić wiersz drugi lub trzeci, to jako u_{22} uzyskamy odpowiednio liczby: $d_2 = a_{22} - a_{21}a_{12} = 31 - \frac{2}{3} \cdot 24 = 15$ lub $d_3 = a_{32} - a_{31}a_{12} = 32 - \frac{1}{2} \cdot 24 = 20$. Ponieważ większą z tych liczb jest 20, więc powinniśmy dokonać zamiany drugiego i trzeciego wiersza w macierzach A , D i w kolumnie wskaźników. Otrzymamy wówczas macierze

$$\begin{bmatrix} 30 & 24 & 18 \\ 1 & & \\ 2 & 32 & 21 \\ 2 & & \\ 3 & 31 & 23 \end{bmatrix} \quad \begin{bmatrix} 30 \\ 20 \\ 15 \end{bmatrix} \quad \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

i ze wzorów (5.31) znajdujemy

$$u_{22} = d_2 = 20, \quad u_{23} = a_{23} - a_{21}a_{13} = 21 - \frac{1}{2} \cdot 18 = 12$$

$$l_{32} = d_3/d_2 = \frac{3}{4}$$

Wartości te wpisujemy w odpowiednie miejsca macierzy A , uzyskując

$$\begin{bmatrix} 30 & 24 & 18 \\ 1 & & \\ 2 & 20 & 12 \\ 2 & 3 & \\ 3 & \frac{3}{4} & 23 \end{bmatrix} \quad \begin{bmatrix} 30 \\ 20 \\ 15 \end{bmatrix} \quad \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

$i = 3$. Ponieważ $i = n$, więc nie ma już przestawień. Obliczając

$$u_{33} = a_{33} - a_{31}a_{13} - a_{32}a_{23} = 23 - \frac{2}{3} \cdot 18 - \frac{3}{4} \cdot 12 = 2$$

otrzymujemy ostatecznie rozkład LU w postaci

$$\begin{bmatrix} 30 & 24 & 18 \\ 1 & & \\ 2 & 20 & 12 \\ 2 & 3 & \\ 3 & \frac{3}{4} & 2 \end{bmatrix} \quad \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

Kolumna wskaźników umożliwia odtworzenie przestawień wierszy wyjściowej macierzy A . Numery 2, 3, 1 oznaczają, że iloczyn LU jest macierzą, której kolejnymi wierszami są 2, 3 i 1 wiersz wyjściowej macierzy A .

Sprawdzamy:

$$LU = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{2}{3} & \frac{3}{4} & 1 \end{bmatrix} \cdot \begin{bmatrix} 30 & 24 & 18 \\ 0 & 20 & 12 \\ 0 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 30 & 24 & 18 \\ 15 & 32 & 21 \\ 20 & 31 & 23 \end{bmatrix}$$

U w a g a. Metoda Doolittle'a umożliwia zwiększenie dokładności obliczeń w porównaniu z metodą GCW. Można tu bowiem wykonywać wszystkie działania ze zwiększoną dokładnością. Dokładniej mówiąc liczby $d_1, d_2, \dots, d_n$ można obliczać w podwójnej precyzji, a zatem błędy pojedynczej precyzji wystąpią jedynie w chwili wpisywania obliczonych wartości do komórek, w których zapamiętane są elementy macierzy A . Zauważmy, że liczby l_{ji} są ilorazem odpowiednich wartości d_j i d_i , które obliczyliśmy uprzednio wyznaczając element podstawowy u_{ii} .

Podamy na zakończenie ważne przykłady macierzy, w stosunku do których nie trzeba stosować strategii częściowego wyboru elementów podstawowych.

Macierz kwadratową A nazywamy *diagonalnie dominującą*, jeżeli moduły elementów na diagonalu są niemniejsze od sumy modułów pozostałych elementów stojących w tym samym wierszu, tzn.

$$|a_{ii}| \geq \sum_{\substack{k=1 \\ k \neq i}}^n |a_{ik}|, \quad i = 1, 2, \dots, n$$

Na przykład

$$A = \begin{bmatrix} 3 & 1 & 0 \\ 0 & 2 & 2 \\ 4 & 1 & -5 \end{bmatrix}$$

jest diagonalnie dominująca.

Macierz A jest *silnie diagonalnie dominująca*, jeżeli powyższe nierówności są ostre.

Macierz A jest (*silnie*) *diagonalnie dominująca kolumnowo*, jeżeli A^T jest (*silnie*) diagonalnie dominująca, tzn.

$$|a_{ii}| \geq (>) \sum_{\substack{k=1 \\ k \neq i}}^n |a_{ki}|, \quad i = 1, 2, \dots, n$$

Tw. 2. Jeżeli macierz A jest nieosobliwa i diagonalnie dominująca kolumnowo, to przy eliminacji metodą GCW nie ma potrzeby przestawiania wierszy.

DOWÓD. Wystarczy pokazać, że po wyeliminowaniu N zmiennych, pozostałe wiersze i kolumny o numerach $N+1, N+2, \dots, n$ tworzą macierz diagonalnie dominującą kolumnowo. Dowód przeprowadzimy indukcyjnie. Macierz $A^{(1)}$ spełnia tę własność z założenia, zatem pierwszym elementem podstawowym jest $a_{11}^{(1)}$ i wszystkie mnożniki $l_{k1} = a_{k1}^{(1)}/a_{11}^{(1)}, k = 2, 3, \dots, n$, spełniają warunek $\sum_{k=2}^n |l_{k1}| \leq 1$. Po eliminacji pierwszej zmiennej uzyskamy z kolumn i wierszy

o numerach $2, 3, \dots, n$ macierz, której elementy są równe $a_{ki}^{(2)} = a_{ki}^{(1)} - l_{k1} a_{1i}^{(1)}$ dla $k = 2, 3, \dots, n$; $i = 2, 3, \dots, n$. Zatem na diagonalu tej macierzy elementy $|a_{ii}^{(2)}| \geq |a_{ii}^{(1)}| - |l_{i1}| \cdot |a_{1i}^{(1)}|$, a poza diagonalą elementy $|a_{ki}^{(2)}| \leq |a_{ki}^{(1)}| + |l_{k1}| \cdot |a_{1i}^{(1)}|$, skąd

$$|a_{ii}^{(2)}| - \sum_{\substack{k=2 \\ k \neq i}}^n |a_{ki}^{(2)}| \geq |a_{ii}^{(1)}| - |a_{1i}^{(1)}| \sum_{k=2}^n |l_{k1}| - \sum_{\substack{k=2 \\ k \neq i}}^n |a_{1i}^{(1)}| \geq 0$$

Macierz $A^{(2)}$ jest więc nadal diagonalnie dominująca kolumnowo.

Analogicznie uzasadniamy, że po eliminacji N zmiennych, uzyskamy macierz $A^{(N+1)}$ diagonalnie dominującą kolumnowo.

U w a g a. Własność $\sum_{k=i+1}^n |l_{ki}| \leq 1$, $i = 1, 2, \dots, (n-1)$, umożliwia istotne polepszenie oszacowania (5.26).

Nie potrzeba również dokonywać wyboru elementów podstawowych w przypadku, gdy macierz A jest *symetryczna i dodatnio określona*. Nie można wtedy co prawda uzyskać oszacowania $|l_{ij}| \leq 1$, ale za to dowodzi się (patrz [64]), że podczas eliminacji zmiennych nie następuje wzrost największych elementów macierzy $A^{(1)}, A^{(2)}, \dots, A^{(n)} = U$. Mamy wtedy $\|A^{(i)}\|_{1\infty} \leq \|A\|_{1\infty}$, $i = 1, 2, \dots, n$, a więc w szczególności $\|U\|_{1\infty} \leq \|A\|_{1\infty}$.

Często w praktyce spotykamy układy równań $Ax = b$ takie, że macierze A są symetryczne i silnie diagonalnie dominujące o dodatnich elementach na diagonalu (patrz p. 3). Wówczas z twierdzenia Gerschgorina (patrz p. 6.2) wynika, że są to macierze dodatnio określone, a zatem łączą w sobie pozytywne cechy obu powyższych przypadków. Oszacowania (5.33) i (5.34) można wtedy wzmocnić; dla normy $\|\cdot\|_{\infty}$ uzyskamy wówczas oszacowania

$$\|E\|_{\infty} \leq \varepsilon n A$$

$$\|\delta L\|_{\infty} \leq \varepsilon (2n + 2)$$

$$\|\delta U\|_{\infty} \leq \varepsilon \left(\frac{1}{2} n^2 + 5n - 4 \right) \|A\|_{\infty}$$

$$\|U\|_{\infty} \leq n \|A\|_{\infty}$$

skąd

$$\frac{\|\delta A^{(z)}\|_{\infty}}{\|A\|_{\infty}} \leq \varepsilon (3n^2 + 8n - 4) + o(\varepsilon) \quad (5.35)$$

a więc we wzorze (5.34) zamiast $O\left(\frac{9}{2}n^3\right)$ można przyjąć $O(3n^2)$.

Ćwiczenia

1. Znaleźć odpowiednik wzorów (5.31) wyznaczając rozkład $A = LU$, gdzie L jest macierzą trójkątną dolną, a U macierzą trójkątną górną z jedynkami na diagonalu. Uzasadnić, że taka metoda (zwana *metodą Crouta*) przy stosowaniu częściowego wyboru elementu podstawowego ma analogiczne do (5.34) oszacowanie błędów zaokrągleń. W s k a z ó w k a. Zastosować wzory Doolittle'a do macierzy A^T .

2. Znaleźć przykład macierzy $A_{n \times n}$ takiej, że w wyniku stosowania metody GCW największy element macierzy A wzrośnie 2^{n-1} razy. W s k a z ó w k a. Na przykład dla $n = 3$

$$A = \begin{bmatrix} a & 0 & a \\ -a & a & a \\ -a & -a & a \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} a & 0 & a \\ 0 & a & 2a \\ 0 & 0 & 4a \end{bmatrix}$$

Ogólnie ze wzoru (5.31) wynika, że musi być $l_{n1} = l_{n2} = \dots = l_{n,n-1} = -1$.

3. Wyprowadzić oszacowania błędów zaokrągleń dla metody Doolittle'a z częściowym wyborem elementu podstawowego i obliczaniu elementów d_i (patrz przykł. 6) ze zwiększoną dokładnością.

4. Uzasadnić, że gdy macierz A jest diagonalnie dominująca, wówczas nie musimy stosować częściowego wyboru elementu podstawowego w metodzie eliminacji Gaussa. W s k a - z ó w k a. Można skorzystać z rozkładu Crouta; patrz ćwic. 1.

5.3.4. Metoda eliminacji Jordana

Układ równań postaci $A^{(1)}x = b^{(1)}$, tzn.

$$a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \dots + a_{1n}^{(1)}x_n = b_1^{(1)}$$

$$a_{21}^{(1)}x_1 + a_{22}^{(1)}x_2 + \dots + a_{2n}^{(1)}x_n = b_2^{(1)}$$

.....

$$a_{n1}^{(1)}x_1 + a_{n2}^{(1)}x_2 + \dots + a_{nn}^{(1)}x_n = b_n^{(1)}$$

przekształcamy następująco. Pierwsze równanie dzielimy obustronnie przez $a_{11}^{(1)}$, a następnie od i -tego wiersza, $i = 2, 3, \dots, n$, odejmujemy wiersz pierwszy pomnożony przez $a_{i1}^{(1)}$, otrzymując układ $A^{(2)}x = b^{(2)}$, tzn.

$$x_1 + a_{12}^{(2)}x_2 + \dots + a_{1n}^{(2)}x_n = b_1^{(2)}$$

$$a_{22}^{(2)}x_2 + \dots + a_{2n}^{(2)}x_n = b_2^{(2)}$$

.....

$$a_{n2}^{(2)}x_2 + \dots + a_{nn}^{(2)}x_n = b_n^{(2)}$$

Następnie drugie równanie dzielimy obustronnie przez $a_{22}^{(2)}$ i od i -tego wiersza, $i = 1, 3, 4, \dots, n$, odejmujemy wiersz drugi pomnożony przez $a_{i2}^{(2)}$, otrzymując układ $A^{(3)}x = b^{(3)}$, tzn.

$$x_1 + a_{13}^{(3)}x_3 + \dots + a_{1n}^{(3)}x_n = b_1^{(3)}$$

$$x_2 + a_{23}^{(3)}x_3 + \dots + a_{2n}^{(3)}x_n = b_2^{(3)}$$

.....

$$a_{n3}^{(3)}x_3 + \dots + a_{nn}^{(3)}x_n = b_n^{(3)}$$

Po $(n - 1)$ eliminacjach otrzymujemy układ postaci

$$x_1 = b_1^{(n)}$$

$$x_2 = b_2^{(n)}$$

.....

$$x_n = b_n^{(n)}$$

a więc gotowe rozwiązanie.

Powyższa metoda, zwana metodą *eliminacji Jordana* lub *metodą eliminacji zupełnej* wymaga wykonania $M = \frac{1}{2}n^3 + \frac{1}{2}n^2$ i $D = \frac{1}{2}n^3 - \frac{1}{2}$ działań.

Liczba działań jest więc tu około półtora raza większa niż dla metody eliminacji Gaussa. Liczba zajętych komórek pamięci jest taka sama. Aby nie wystąpiło dzielenie przez zerowy element $a_{ii}^{(i)}$, należy zastosować odpowiedni wybór elementu podstawowego. Zaletą metody jest możliwość prostego określenia rozwiązania *obciętego układu równań*, otrzymanego na przykład z $k < n$ początkowych równań po przyjęciu zerowej wartości zmiennych $x_{k+1}, x_{k+2}, \dots, x_n$. Metoda ta umożliwia ponadto rozwiązywanie układu równań przy oszczędniejszym wykorzystaniu pamięci operacyjnej, co umożliwia rozwiązywanie układów o większej liczbie zmiennych (patrz np. [69, s. 424]). Istotną wadą jest tu jednak duży nakład obliczeń, który nie ulega zmniejszeniu przy ponownym rozwiązywaniu układu równań przy zmienionej prawej stronie układu (wektora $\mathbf{b}$). Nie można bowiem dla tej metody znaleźć odpowiednika rozkładu $A = LU$, jak dla metody Gaussa.

5.3.5. Układy z macierzą symetryczną. Rozkłady LDL^T i LL^T

Zauważmy, że dysponując rozkładem macierzy kwadratowej A na iloczyn $L\bar{U}$, gdzie L jest macierzą trójkątną dolną z jedynkami na diagonalu, a $\bar{U}$ — macierzą trójkątną górną z niezerowymi elementami na diagonalu, łatwo uzyskać rozkład macierzy $A = LDU$, gdzie L jest tą samą macierzą co poprzednio, D — macierzą diagonalną o elementach na diagonalu równych elementom na diagonalu macierzy $\bar{U}$, a U — macierzą trójkątną górną z jedynkami na diagonalu.

Rozkład $A = LDU$ można uzyskać metodą eliminacji Gaussa, bądź można opracować zwarte wzory analogiczne do (5.31). Rozkład taki jest jednoznaczny.

Jeśli macierz A jest symetryczna, to zarówno $LDU = A$, jak i $U^TDL^T = A^T = A$, a więc $U = L^T$. Dla macierzy symetrycznych istnieje zatem jednoznaczny rozkład

$$A = LDT^T$$

Wzory do wyznaczenia D i L są następujące (d_i oznaczają elementy d_{ii} macierzy diagonalnej D):

$$d_1 = a_{11}$$

Dla $i = 2, 3, \dots, n$ obliczamy na przemian

$$l_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} c_{ik} l_{jk} \right) / d_j$$

$$c_{ij} = d_j l_{ij}, \quad j = 1, 2, \dots, i-1 \quad (5.36)$$

$$d_i = a_{ii} - \sum_{k=1}^{i-1} c_{ik} l_{ik}$$

Przykład 1. Dla macierzy

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

mamy

$$\begin{aligned} i=1: & \quad d_1 = 1 \\ i=2: & \quad l_{21} = 1/1 = 1, \quad c_{21} = 1 \cdot 1 = 1, \quad d_2 = 0 - 1 \cdot 1 = -1 \\ i=3: & \quad l_{31} = 0/1 = 0, \quad c_{31} = 1 \cdot 0 = 0 \\ & \quad l_{32} = (1 - 0 \cdot 1)/(-1) = -1, \quad c_{32} = (-1) \cdot (-1) = 1, \\ & \quad d_3 = 1 - 0 \cdot 0 - 1(-1) = 2 \end{aligned}$$

Otrzymujemy zatem

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

W celu wyznaczenia rozkładu $A = LDL^T$ dla macierzy $A_{n \times n}$ trzeba wykonać $M = \frac{1}{6}n^3 + n^2 - \frac{7}{6}n$ i $D = \frac{1}{6}n^3 - \frac{1}{6}n$ działań. Układ równań $Ax = b$ możemy następnie zastąpić trzema układami równań:

$$L^T z = b, \quad Dy = z, \quad Lx = y$$

które rozwiązujemy po dodatkowych n^2 mnożeniach i $n^2 - n$ dodawaniach. Dla macierzy symetrycznej A można zatem znaleźć rozwiązanie x przy blisko dwa razy mniejszej liczbie działań niż przy stosowaniu metody Gaussa. Zauważmy także, że we wzorach (5.36) występują elementy a_{ij} , dla których $i \geq j$, a więc wystarczy zapamiętać macierz A w postaci dolnej połowy wraz z przekątną, czyli wystarczy zająć $\frac{1}{2}n^2 + \frac{1}{2}n$ komórek pamięci. Oczywiście dojdą wówczas dodatkowe operacje arytmetyczne spowodowane obliczaniem właściwego adresu elementów a_{ij} macierzy A , ale można sobie z tym poradzić analogicznie jak w przypadku wzorów (5.31). Reasumując, wskutek symetrii macierzy A oszczędzamy blisko dwukrotnie na czasie obliczeń i liczbie zajętych komórek pamięci.

Należy zdawać sobie sprawę z tego, że nawet dla nieosobliwej macierzy A we wzorach (5.36) może wystąpić $d_i = 0$, co zatrzyma obliczenia. Gdy jednak macierz A jest dodatnio określona, wszystkie liczby d_i będą dodatnie. Dla symetrycznych macierzy dodatnio określonych istnieje jeszcze jeden typ rozkładu na iloczyn macierzy trójkątnych, zwany *rozkładem Banachiewicza* (w wielu publikacjach anglosaskich rozkład ten nazywa się *rozkładem*

Cholesky'ego):

$$A = LL^T$$

Macierz L jest macierzą trójkątną dolną, przy czym na diagonalu mogą być elementy niekoniecznie równe 1. Ponieważ macierz $\bar{L} = -L$ również spełnia warunek $A = \bar{L}\bar{L}^T$ dla $A = LL^T$, więc rozkład taki nie jest jednoznaczny. Jeśli jednak założyć, że liczby na diagonalu macierzy L są dodatnie, to otrzymujemy jednoznaczne wzory obliczeniowe w sposób następujący. Obliczamy kolejno:

$$\begin{aligned} l_{ii} &= \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2}, \quad i = 1, 2, \dots, n \\ l_{ji} &= \left(a_{ji} - \sum_{k=1}^{i-1} l_{jk} l_{ik} \right) / l_{ii}, \quad j = i + 1, i + 2, \dots, n \end{aligned} \quad (5.37)$$

Przykład 2. Dla macierzy

$$A = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 3 & 6 \end{bmatrix}$$

znajdujemy:

$$i = 1: \quad l_{11} = \sqrt{4} = 2, \quad l_{21} = 2/2 = 1, \quad l_{31} = 2/2 = 1$$

$$i = 2: \quad l_{22} = \sqrt{5 - 1 \cdot 1} = 2, \quad l_{32} = (3 - 1 \cdot 1)/2 = 1$$

$$i = 3: \quad l_{33} = \sqrt{6 - 1 \cdot 1 - 1 \cdot 1} = 2$$

zatem

$$\begin{bmatrix} 4 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 3 & 6 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 2 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix}$$

W celu wyznaczenia rozkładu $A = LL^T$ dla macierzy $A_{n \times n}$ trzeba wykonać n pierwiastkowań, $M = \frac{1}{6}n^3 + \frac{1}{2}n^2 - \frac{2}{3}n$ mnożeń i $D = \frac{1}{6}n^3 - \frac{1}{2}n^2 + \frac{1}{3}n$ dodawań. Dla większości maszyn cyfrowych pierwiastkowanie jest tylko kilkakrotnie dłuższym działaniem niż mnożenie, a zatem znowu można przyjąć, że oszczędzamy blisko dwukrotnie na czasie obliczeń i liczbie zajętych komórek pamięci w porównaniu z metodą Gaussa (wystarczy zapamiętać dolną połowę macierzy A wraz z przekątną). Dodatnia określoność macierzy A gwarantuje, że elementy l_{ii} na diagonalu są niezerowe, a błąd wyznaczenia rozkładu $A = LL^T$ jest niewielki, zwłaszcza, jeśli działanie wykonujemy w zwiększonej precyzji (patrz np. [64]).

Błędy zaokrągleń w rozkładzie Banachiewicza. Ponieważ spełniony jest związek $\sum_{j=1}^i l_{ij}^2 = a_{ii}$, $i = 1, 2, \dots, n$, więc liczby l_{ij} mają ograniczoną wielkość,

Liczba działań wynosi: $M = 2n - 2$, $D = n - 1$, a liczba zajętych komórek pamięci $P = 3n - 2$.

U w a g a. Elementy l_i oraz u_i można zapamiętać w miejscach zajętych uprzednio przez elementy a_i oraz b_i macierzy T .

Aby znaleźć rozwiązanie x układu równań $Tx = d$, czyli $LUx = d$, należy rozwiązać kolejno dwa układy równań $Ly = d$ oraz $Ux = y$, według wzorów

$$\begin{aligned} y_1 &= d_1 \\ y_i &= d_i - l_i y_{i-1} \quad \text{dla } i = 2, 3, \dots, n \end{aligned} \quad (5.40a)$$

oraz

$$\begin{aligned} x_n &= y_n / u_n \\ x_i &= (y_i - c_i x_{i+1}) / u_i \quad \text{dla } i = n - 1, n - 2, \dots, 1 \end{aligned} \quad (5.40b)$$

Łącznie należy zatem dodatkowo wykonać $M = 3n - 2$, $D = 2n - 2$ działań.

U w a g a. Układ równań $Tx = d$ rozwiązujemy za pomocą wzorów (5.39) i (5.40) wówczas, gdy poszukujemy rozwiązań x dla co najmniej dwu różnych wektorów danych d . Jeśli jednak rozwiązujemy układ $Tx = d$ tylko raz, to nie musimy pamiętać elementów l_i , a więc otrzymamy wzory:

$$\begin{aligned} u_1 &= b_1, \quad y_1 = d_1 \\ l &= a_i / u_{i-1}, \quad u_i = b_i - l c_{i-1}, \quad y_i = d_i - l y_{i-1} \end{aligned} \quad (5.41a)$$

dla $i = 2, 3, \dots, n$, oraz

$$\begin{aligned} x_n &= y_n / u_n \\ x_i &= (y_i - c_i x_{i+1}) / u_i \quad \text{dla } i = n - 1, n - 2, \dots, 1 \end{aligned} \quad (5.41b)$$

Jak poprzednio, mamy tu $M = 5n - 4$, $D = 3n - 3$ działań. Bez dodatkowych założeń odnośnie macierzy T metoda może być niestabilna, w szczególności nawet dla nieosobliwej macierzy T może wystąpić $u_i = 0$, co zatrzyma obliczenia.

Przykład 1. Dla macierzy

$$T = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \quad \det T = -1$$

otrzymujemy $u_1 = 1$, $l_2 = 1$, $u_2 = 0$, a więc niemożliwe jest obliczenie $l_3 = 1/u_2$.

Jeżeli macierz T jest diagonalnie dominująca kolumnowo, tzn. $|b_1| \geq |a_2|$, $|b_i| \geq |c_{i-1}| + |a_{i+1}|$ dla $i = 2, 3, \dots, n - 1$, $|b_n| \geq |c_{n-1}|$, to zgodnie z twierdzeniem z p. 5.3.3 powyższy rozkład $T = LU$ jest równoważny wersji rozkładu z częściowym wyborem elementu podstawowego, która jest niezawodna. Ponadto elementy macierzy L spełniają warunek $|l_i| \leq 1$, $i = 2, 3, \dots, n$, co poprawia oszacowanie błędów zaokrągleń.

Jeśli macierz T jest diagonalnie dominująca, tzn. $|b_1| \geq |c_1|$, $|b_i| \geq |a_i| + |c_i|$, $i = 2, 3, \dots, n - 1$, $|b_n| \geq |a_n|$, to metoda eliminacji Gaussa również jest niezawodna (patrz ćwic. 4 z p. 5.3.3). Wówczas jednak często stosujemy inną wersję rozkładu $T = LU$, mianowicie metodę Crouta (ćwic. 1

z p. 5.3.3). W przypadku macierzy trójdzielnej T oznacza to obliczanie macierzy

$$L = \begin{bmatrix} l_1 & & & 0 \\ a_2 & & & \\ & \ddots & & \\ 0 & & a_n & l_n \end{bmatrix}, \quad U = \begin{bmatrix} 1 & u_1 & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & u_{n-1} \\ & & & & 1 \end{bmatrix}$$

Stąd otrzymujemy wzory analogiczne do wzorów (5.39)–(5.41), na przykład:

$$\begin{aligned} l &= b_1, \quad y_1 = d_1/l \\ u_i &= c_i/l, \quad l = b_{i+1} - u_i a_{i+1}, \quad y_{i+1} = (d_{i+1} - a_{i+1} y_i)/l \\ &\text{dla } i = 1, 2, \dots, n-1, \text{ oraz} \end{aligned}$$

$$\begin{aligned} x_n &= y_n \\ x_i &= y_i - u_i x_{i+1} \quad \text{dla } i = n-1, n-2, \dots, 1 \end{aligned}$$

Przy założeniu, że T jest macierzą diagonalnie dominującą powyższe wzory są niezawodne (są one równoważne wyborowi elementów podstawowych w kolejnych wierszach) i macierze L i U są diagonalnie dominujące.

Oszacowanie błędów zaokrągleń. W przypadku stosowania wzorów (5.41) dla macierzy diagonalnie dominującej kolumnowo przy analizie błędów zaokrągleń proces obliczeniowy dzielimy na dwa etapy: na oszacowanie błędu rozkładu $T = LU$, następnie zaś na oszacowanie błędu rozwiązania układów równań $Ly = d$, $Ux = y$. Ze wzorów (5.39) wyznaczamy L i U takie, że $LU = T + E$, przy czym dla norm $\|\cdot\|_1$ lub $\|\cdot\|_\infty$ zachodzą nierówności

$$\|E\|_p \leq 2\varepsilon \|T\|_p, \quad p = 1, \infty$$

Następnie ze wzorów (5.40) wyznaczamy rozwiązania zaburzonych układów $(L + \delta L)y = d$, $(U + \delta U)x = y$, przy czym

$$\|\delta L\|_p \leq 3\varepsilon, \quad \|L\|_p \leq 2\varepsilon$$

$$\|\delta U\|_p \leq 3\varepsilon \|T\|_p, \quad \|U\|_p \leq 2\varepsilon \|T\|_p$$

Otrzymane w ten sposób rozwiązanie x jest dokładnym rozwiązaniem układu równań $(T + \delta T)x = d$, przy czym $\|\delta T\|_p / \|T\|_p \leq 14\varepsilon$, skąd

$$\frac{\|\delta x\|_p}{\|x\|_p} \leq \frac{\alpha}{1 - \alpha}, \quad \alpha = 14\varepsilon K_p, \quad p = 1, \infty \quad (5.42)$$

Identyczne oszacowanie uzyskujemy dla metody Crouta.

Porównując oszacowania (5.42) z (5.34) dla metody GCW, spostrzegamy, że oszacowanie błędu wyznaczenia rozwiązania układu $Tx = d$ jest niezależne od wymiaru n . Aby więc zagwarantować rozwiązanie układu

równań z macierzą trójdagonalną T z dokładnością np. 10% wystarczy zażądać spełnienia warunku $K\varepsilon \leq 0,07$, czyli na maszynie z arytmetyką w systemie dwójkowym i długością słowa $t = 33$ bity, można rozwiązywać układy równań o wskaźniku uwarunkowania $K \leq 10^{10}$ bez względu na liczbę niewiadomych.

Dla trójdogonalnych macierzy T można również stosować rozkład $T = LDU$, a dla symetrycznych dodatnio określonych trójdogonalnych macierzy T rozkład $T = LDL^T$, ale nie uzyskamy istotnej poprawy dokładności, a liczba działań arytmetycznych nieco wzrasta.

Ćwiczenia

1. Wykazać, że jeśli macierz T jest diagonalnie dominująca, to wzory (5.39) są niezawodne bez stosowania wyboru elementu podstawowego.
2. Wyprowadzić wzory na obliczanie rozkładu $T = LL^T$.

3. Znaleźć rozkład LU macierzy $T = \begin{bmatrix} 2 & 2 & 0 \\ -2 & 0 & 2 \\ 0 & -2 & 0 \end{bmatrix}$.

5.3.7. Układy równań z macierzą zbliżoną do trójdogonalnej

Przy rozwiązywaniu równań różniczkowych cząstkowych spotykamy układy równań z macierzą postaci

$$\begin{bmatrix} b_1 & c_1 & . & . & . & . & . & . & . & p_1 \\ a_2 & b_2 & c_2 & & & & & & 0 & . \\ . & a_3 & b_3 & . & & & & & 0 & . \\ . & . & . & . & . & & & & . & . \\ . & . & 0 & . & . & . & . & . & . & c_{n-1} \\ q_1 & . & . & . & . & . & . & . & a_n & b_n \end{bmatrix}$$

W porównaniu z poprzednio rozpatrywanymi macierzami trójdogonalnymi występują tu dwa dodatkowe niezerowe elementy p_1 i q_1 . Okazuje się, że ta pozornie niewielka zmiana postaci macierzy powoduje ponad dwukrotny wzrost liczby działań arytmetycznych potrzebnych do rozwiązania układu równań, ponad 60% wzrost obciążenia pamięci oraz — co najważniejsze — istotne pogorszenie oszacowania błędów zaokrągleń powstających podczas rozwiązywania równań.

Dla uproszczenia przyjmiemy, że powyższą macierz można rozłożyć na iloczyn LU , stosując bądź metodę Gaussa, bądź Doolittle'a. W tym celu wystarczy założyć, np., że macierz A jest symetryczna i dodatnio określona lub nieosobliwa i diagonalnie dominująca kolumnowo (taka własność jest

spełniona w przypadku równań cząstkowych). Macierze L i U mają wówczas postać

$$L = \begin{bmatrix} 1 & & & & & \\ l_2 & \cdot & & & & 0 \\ \vdots & \cdot & \cdot & \cdot & & \\ 0 & & & l_{n-1} & \cdot & \\ & & & & \cdot & \\ q_2 & \cdot & \cdot & \cdot & q_{n-1} & q_n & 1 \end{bmatrix} \quad U = \begin{bmatrix} u_1 & c_1 & \cdot & \cdot & \cdot & p_1 \\ & \cdot & \cdot & & 0 & \vdots \\ & & \cdot & \cdot & \cdot & \vdots \\ & & & \cdot & c_{n-2} & p_{n-2} \\ & 0 & & \cdot & \cdot & p_{n-1} \\ & & & & \cdot & \\ & & & & & u_n \end{bmatrix}$$

Aby zapamiętać te macierze, należy zapamiętać łącznie $5n - 4$ elementów: $l_2, \dots, l_{n-1}, q_2, \dots, q_n, u_1, \dots, u_n, c_1, \dots, c_{n-2}, p_1, \dots, p_{n-1}$. Elementy te można wyznaczyć stosując metodę eliminacji Gaussa lub następujące wzory rekurencyjne:

$$\begin{aligned} u_1 &= b_1 & l_2 &= a_2/u_1, & q_2 &= -q_1/u_1 \\ u_2 &= b_2 - l_2 c_1, & l_3 &= a_3/u_2, & q_3 &= -q_2 c_1/u_2 \\ &\dots\dots\dots \end{aligned} \quad (5.43)$$

$$\begin{aligned} u_{n-2} &= b_{n-2} - l_{n-2} c_{n-3}, & l_{n-1} &= a_{n-1}/u_{n-2}, & q_{n-1} &= -q_{n-2} c_{n-3}/u_{n-2} \\ u_{n-1} &= b_{n-1} - l_{n-1} c_{n-2}, & q_n &= (a_n - q_{n-1} c_{n-2})/u_{n-1} \end{aligned}$$

$$\begin{aligned} p_2 &= -l_2 p_1, & p_3 &= -l_3 p_2, & \dots, & p_{n-2} &= -l_{n-2} p_{n-3}, & p_{n-1} &= c_{n-1} - l_{n-1} p_{n-2} \\ u_n &= b_n - p_1 q_2 - \dots - p_{n-1} q_n \end{aligned}$$

Łącznie należy zatem wykonać $M = 6n - 10$ i $D = 2n - 1$ działań arytmetycznych.

Jeśli chcemy teraz rozwiązać układ równań $Ax = d$, wystarczy rozwiązać kolejno dwa układy równań $Ly = d$ i $Ux = y$ według wzorów:

$$\begin{aligned} y_1 &= d_1 & x_n &= y_n/u_n \\ y_2 &= d_2 - l_2 y_1 & x_{n-1} &= (y_{n-1} - p_{n-1} x_n)/u_{n-1} \\ &\dots\dots\dots & x_{n-2} &= (y_{n-2} - c_{n-2} x_{n-1} - p_{n-2} x_n)/u_{n-2} \\ y_{n-1} &= d_{n-1} - l_{n-1} y_{n-2} & &\dots\dots\dots \\ y_n &= d_n - q_2 y_1 - \dots - q_n y_{n-1} & x_1 &= (y_1 - c_1 x_2 - p_1 x_n)/u_1 \end{aligned} \quad (5.44)$$

Dodatkowo należy zatem wykonać $M = 5n - 6$ i $D = 4n - 8$ działań arytmetycznych. Widzimy więc, że całkowity nakład obliczeń potrzebnych do rozwiązania układu $Ax = d$ wynosi $M = O(11n)$ i $D = O(6n)$, a więc przeszło dwukrotnie więcej niż dla macierzy trójdzielnej.

Oszacowanie błędów zaokrągleń. Przyjmując założenie, że macierz A jest symetryczna i diagonalnie dominująca, ze wzorów (5.43) wyznaczamy macierze L i U takie, że $LU = A + E$, przy czym $\|E\|_1 \leq 2n\|A\|_1 \varepsilon$. Z kolei ze

wzorów (5.44) wyznaczamy rozwiązanie układów $(L + \delta L)y = d$, $(U + \delta U)x = y$, przy czym

$$\begin{aligned}\|L\|_1 &\leq 2, & \|\delta L\|_1 &\leq n \|L\|_1 \varepsilon \\ \|U\|_1 &\leq 2 \|A\|_1, & \|\delta U\|_1 &\leq 5 \|U\|_1 \varepsilon\end{aligned}$$

Wyznaczone rozwiązanie x jest więc dokładnym rozwiązaniem zaburzonego układu równań: $(A + \delta A)x = d$, z oszacowaniem $\|\delta A\|_1 / \|A\|_1 \leq (6n + 20 + 20n\varepsilon)$, czyli w porównaniu z (5.42) nastąpiło istotne pogorszenie oszacowania, jakkolwiek nadal jest to wynik znacznie lepszy niż wynik uzyskany dla dowolnej macierzy (5.33).

5.3.8. Obliczanie wyznacznika i odwracanie macierzy

Przy rozwiązywaniu układów równań często korzystamy z wyznaczników macierzy, do obliczania których stosujemy reguły rozwijania względem kolumny lub wiersza lub też korzystamy z innych sposobów, np. z reguły Sarrusa. Tego rodzaju metody są praktycznie bezużyteczne, gdy korzystamy z maszyn cyfrowych, bowiem liczba mnożeń potrzebnych do obliczenia wyznacznika n -tego stopnia za pomocą rozwijania względem wiersza wynosi $n!$, więc w przypadku $n = 16$ byłoby to prawie $5 \cdot 10^{12}$ mnożeń; maszyna cyfrowa wykonująca 10^6 mnożeń na sekundę potrzebowałaby na to ponad rok nieprzerwanej pracy. Jeśli zachodzi konieczność obliczenia wyznacznika, najlepiej dokonać rozkładu macierzy $A = LU$ (stosując częściowy wybór elementu podstawowego), a wówczas $\det(A + E) = \det LU = \det L \det U = \det U$, ponieważ elementy na diagonalu macierzy L są równe 1 więc $\det L = 1$. Macierz E oznacza tu, jak poprzednio (patrz p. 5.3.3), macierz zaburzenia zastępującego błędy zaokrągleń powstające podczas obliczania elementów macierzy L i U .

Wyznacznik macierzy trójkątnej U jest równy iloczynowi elementów na diagonalu i można wyznaczyć go z dużą dokładnością wykonując $n - 1$ mnożeń. Jak jednak wiemy, obliczona wartość nie jest wyznacznikiem macierzy A , lecz macierzy zaburzonej $A + E$. Zgodnie z oszacowaniem (5.32c) norma zaburzenia E może być mała. Nie oznacza to jednak, że również błąd względny wyznaczenia wartości $\det A$ jest mały.

Przykład 1.

Niech $A = \begin{bmatrix} 2 & \varepsilon \\ 1 & \varepsilon \end{bmatrix}$, gdzie ε jest najmniejszą dodatnią liczbą, którą możemy zapisać w arytmetyce maszyny. Dokładna wartość wyznacznika macierzy A ma wartość ε . Jeśli jednak obliczyć macierz U z rozkładu $A = LU$, to uzyskamy

$$U = \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}$$

gdzie $u_{11} = 2$, $u_{12} = \varepsilon$, a $u_{22} = \varepsilon - \frac{\varepsilon}{2}$. Obliczając wartość u_{22} na maszynie cyfrowej uzyskamy

jednak $u_{22} = \varepsilon$, ponieważ $\frac{\varepsilon}{2}$ to zero w maszynie cyfrowej. Otrzymamy zatem $\det U = 2\varepsilon$, czyli obliczymy wartość $\det A$ ze 100% błędem. Tymczasem $\|E\|_\infty = \|LU - A\|_\infty = \frac{1}{2}\varepsilon$, więc błąd względny $\|E\|_\infty / \|A\|_\infty \approx \frac{1}{4}\varepsilon \ll 1$.

Powyższy przykład dowodzi znanej własności wyznaczników, iż małe zaburzenia elementów macierzy mogą spowodować duże zmiany wartości wyznacznika. Okazuje się, że zależą one od wielkości $K \frac{\|E\|}{\|A\|}$, gdzie K jest wskaźnikiem uwarunkowania macierzy A . Dokładniejsze oszacowania wielkości zmian wyznacznika podano niżej dla macierzy symetrycznej.

Omówimy obecnie krótko wyznaczanie macierzy odwrotnej A^{-1} metodą GCW. W tym celu należy wyznaczyć rozkład LU macierzy A (uwzględniając ewentualne przestawienia wierszy macierzy A), a następnie rozwiązać n razy układ równań $LUx^{(i)} = e^{(i)}$, $i = 1, 2, \dots, n$, gdzie $e^{(i)} = [0, \dots, 1, \dots, 0]^T$ jest i -tym wersorem w przestrzeni R^n . Rozwiązania $x^{(i)}$ są kolumnami macierzy A^{-1} , przy czym należy ustawić je w odpowiedniej kolejności, wynikającej z przestawień wierszy macierzy. Można też równocześnie rozwiązywać n układów równań z prawymi stronami równymi $e^{(i)}$, $i = 1, 2, \dots, n$.

Przykład 2. Obliczyć macierz

$$\begin{bmatrix} 1 & 0 & 1 \\ 3 & 3 & 0 \\ 0 & 2 & 2 \end{bmatrix}^{-1}$$

Etap 1. Wyznaczamy rozkład LU macierzy A . Należy w tym celu określić przestawienie wierszy, co uczynimy za pomocą dodatkowej kolumny numerów wierszy. Elementy macierzy L zapiszemy na miejscu odpowiednich elementów macierzy A . Mamy

$$\begin{bmatrix} 1 & 0 & 1 \\ 3 & 3 & 0 \\ 0 & 2 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \Rightarrow \text{po przestawieniu} \begin{bmatrix} 3 & 3 & 0 \\ 1 & 0 & 1 \\ 0 & 2 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \Rightarrow$$

po eliminacji pierwszej zmiennej z równań drugiego i trzeciego za pomocą mnożników $l_{21} = \frac{1}{3}$, $l_{31} = 0$

$$\begin{bmatrix} 3 & 3 & 0 \\ \frac{1}{3} & -1 & 1 \\ 0 & 2 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \Rightarrow \text{po przestawieniu} \begin{bmatrix} 3 & 3 & 0 \\ 0 & 2 & 2 \\ \frac{1}{3} & -1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} \Rightarrow$$

po eliminacji drugiej zmiennej z trzeciego równania za pomocą $l_{32} = -\frac{1}{2}$:

$$\begin{bmatrix} 3 & 3 & 0 \\ 0 & 2 & 2 \\ \frac{1}{3} & -\frac{1}{2} & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

Etap 2. Rozwiązujemy równanie $LU\mathbf{x}^{(i)} = \mathbf{e}^{(i)}$, $i = 1, 2, 3$, tzn.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{3} & -\frac{1}{2} & 1 \end{bmatrix} \cdot \begin{bmatrix} 3 & 3 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{bmatrix} \mathbf{x}^{(1)} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \text{ itd.}$$

Otrzymamy

$$\mathbf{x}^{(1)} = \begin{bmatrix} \frac{1}{6} \\ \frac{1}{6} \\ -\frac{1}{6} \end{bmatrix}, \quad \mathbf{x}^{(2)} = \begin{bmatrix} -\frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{bmatrix}, \quad \mathbf{x}^{(3)} = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$$

Numer 1 w wektorze przestawień $\begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$ zajmuje trzecią pozycję, więc jako kolumnę 1 macierzy A^{-1} należy przyjąć $\mathbf{x}^{(3)}$.

Numer 2 znajduje się na pierwszej pozycji, więc drugą kolumną A^{-1} jest $\mathbf{x}^{(1)}$. Numer 3 zajmuje drugą pozycję, zatem trzecią kolumną jest $\mathbf{x}^{(2)}$. Otrzymaliśmy

$$A^{-1} = \begin{bmatrix} \frac{1}{2} & \frac{1}{6} & -\frac{1}{4} \\ -\frac{1}{2} & \frac{1}{6} & \frac{1}{4} \\ \frac{1}{2} & -\frac{1}{6} & \frac{1}{4} \end{bmatrix}$$

Błąd wyznaczenia macierzy odwrotnej. Błąd wyznaczenia macierzy X odwrotnej do macierzy A można oszacować następująco (patrz np. [68, § 27 i 28]). Jeśli $E_1 = I - XA$ (błąd lewostronnej odwrotności), $E_2 = I - AX$ (błąd prawostronnej odwrotności), to dla $i = 1, 2$ mamy

$$\frac{\|E_i\|_\infty}{\|X\|} \leq \varepsilon g \left(2,005n^2 + n^3 + \frac{1}{4}n^4 \varepsilon \right)$$

gdzie g jest największym z elementów macierzy $A = A^{(1)}, A^{(2)}, \dots, A^{(n)} = U$ otrzymanych metodą eliminacji zmiennych. Na ogół $g \leq 8 \|A\|_{1\infty}$ (patrz p. 5.3.3).

Przy poszukiwaniu oszacowania błędu $X - A^{-1}$, wystarczy uwzględnić oszacowania błędów wektorów $\mathbf{x}^{(i)}$, znajdujących się jako rozwiązania układów $A\mathbf{x}^{(i)} = \mathbf{x}^{(i)}$. Ponieważ

$$\frac{\|\delta \mathbf{x}^{(i)}\|_\infty}{\|\mathbf{x}^{(i)}\|_\infty} \leq \frac{\alpha}{1 - \alpha}, \quad \alpha = \varepsilon K_\infty O\left(\frac{9}{2}n^3\right)$$

(patrz wzór (5.34)) zatem

$$\frac{\|X - A^{-1}\|_{1\infty}}{\|A^{-1}\|_{1\infty}} \leq \frac{\alpha}{1 - \alpha}, \quad \alpha = \varepsilon K_\infty \cdot O\left(\frac{9}{2}n^3\right)$$

Omówimy teraz sposób obliczania wyznacznika i macierzy odwrotnej przy założeniu, że macierz jest symetryczna i dodatnio określona. W tym przypadku zaleca się korzystać z rozkładu Banachiewicza $A = LL^T$, opisanego w p. 5.3.5. Mimo że obliczona macierz L spełnia ten rozkład z błędem E , tzn. $LL^T = A + E$, jednak oszacowanie $\|E\|_2 \leq \varepsilon n^{3/2} \|A\|_2$ (patrz (5.38)) powoduje, że wartości własne $\lambda_1, \dots, \lambda_n$ macierzy A oraz wartości własne $\gamma_1, \dots, \gamma_n$ macierzy LL^T (λ_1 i γ_1 są największymi wartościami własnymi, czyli $\|A\|_2 = \lambda_1$, $\|LL^T\|_2 = \gamma_1$) spełniają nierówności

$$-\varepsilon n^{3/2} \lambda_1 \leq \lambda_i - \gamma_i \leq \varepsilon n^{3/2} \lambda_1 \quad (5.45)$$

$$\text{skąd } \prod_{i=1}^n (\lambda_i - \varepsilon n^{3/2} \lambda_1) \leq \det(LL^T) \leq \prod_{i=1}^n (\lambda_i + \varepsilon n^{3/2} \lambda_1)$$

czyli

$$\prod_{i=1}^n \left(1 - \varepsilon n^{3/2} \frac{\lambda_1}{\lambda_i}\right) \leq \frac{\det(LL^T)}{\det A} \leq \prod_{i=1}^n \left(1 + \varepsilon n^{3/2} \frac{\lambda_1}{\lambda_i}\right) \quad (5.46)$$

Oznacza to, że błąd obliczenia wyznacznika zależy od wielkości liczby $\varepsilon n^{3/2} \lambda_{\max} / \lambda_{\min} = \varepsilon n^{3/2} K_2$.

Podobnie uzasadnia się (patrz np. [48, s. 36]), że przy wyznaczaniu macierzy odwrotnej $A^{-1} = (L^T)^{-1}(L^{-1})$ otrzymujemy

$$\frac{\|A^{-1} - (L^T)^{-1}(L^{-1})\|_2}{\|A^{-1}\|_2} \leq \varepsilon n^{3/2} K_2 \quad (5.47)$$

Należy podkreślić, że w przypadku symetrycznej i dodatnio określonej macierzy A nie powinno się stosować ogólnych metod (np. eliminacji Gaussa) zarówno do rozwiązywania równań $Ax = b$, jak i odwracania macierzy, lecz wskazane jest korzystanie z rozkładów $A = LL^T$ lub $A = LDL^T$. Wówczas bowiem nie tylko zmniejsza się liczba działań arytmetycznych, a więc mniej obciążamy pamięć maszyny, ale mamy lepsze oszacowania błędów zaokrągleń oraz numerycznie zapewnioną dodatnią określoność macierzy A^{-1} . Własność ta jest istotna w wielu zastosowaniach (np. przy wyznaczaniu kierunków poprawy w metodach optymalizacji), w których korzystamy z własności $\langle b, A^{-1}b \rangle > 0$, a która to własność może nie być zachowana przy innym sposobie obliczania. Korzystając z rozkładu $A = LL^T$ mamy $\langle b, A^{-1}b \rangle = \|L^{-1}b\|_2^2$, a zatem otrzymujemy liczbę dodatnią, jeśli tylko $L^{-1}b \neq 0$, i przy obliczaniu kwadratu normy możemy stosować zwiększoną dokładność.

równań $Ly = x$ wystarczy wykonać $n - 1$ mnożeń, jeśli skorzystamy z faktu, że jedynie $n - 1$ elementów pod diagonalą macierzy L_T jest niezerowych (patrz np. wzory (5.41)).

Ogólnie można pokazać, że jeśli nie obliczamy macierzy L^{-1} , lecz korzystamy z informacji, które elementy pod diagonalą są zerowe (traktujemy więc macierz L jako rzadką — patrz p. 5.5), to zmniejszamy liczbę działań tyle razy, ile razy więcej elementów ma pełna macierz trójkątna niż macierz L . Podobne wyniki otrzymujemy przy odwracaniu macierzy trójdzielnych i, ogólniej, macierzy wstęgowych.

Przykład 3. Dla macierzy trójdzielnej

$$T = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{bmatrix}$$

macierz odwrotna jest macierzą pełną

$$T^{-1} = \begin{bmatrix} 5 & -4 & 3 & -2 & 1 \\ -4 & 4 & -3 & 2 & -1 \\ 3 & -3 & 3 & -2 & 1 \\ -2 & 2 & -2 & 2 & -1 \\ 1 & -1 & 1 & -1 & 1 \end{bmatrix}$$

Aby obliczyć wektor $y = T^{-1}x$, należy zatem wykonać n^2 mnożeń, natomiast korzystając ze wzorów (5.41) przy rozwiązywaniu układu $Ty = x$ wystarczy wykonać $3n - 2$ mnożeń.

Ćwiczenia

1. Oszacować błąd obliczenia wyznacznika i macierzy odwrotnej dla macierzy $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ przeprowadzając obliczenia jak w przykładzie 2.

2. Z jaką dokładnością należy przeprowadzać obliczenia, aby korzystając z rozkładu $A = LL^T$ obliczyć wyznacznik macierzy Hilberta $A = \left[\frac{10}{i+j} \right]_{5 \times 5}$ z błędem mniejszym niż 1%.

Wskazówka. $K_2 > 1,5 \cdot 10^6$.

5.3.9. Iteracyjne poprawianie rozwiązania

Dotychczasowe rozważania uzasadniają stwierdzenie, że rozwiązanie układu równań $Ax = b$ uzyskane którąś z omówionych metod dokładnych jest na ogół obarczone pewnym błędem. Można przypuszczać, że błąd ów będziemy w stanie wykryć dokonując sprawdzenia, tzn. obliczając wektor reszt

$$r = b - Ax$$

i sprawdzając, czy jest on równy zeru. Że tak nie jest, przekonują nas dwa poniższe przykłady.

Przykład 1. Układ równań

$$\begin{aligned} 0,99x_1 + 0,70x_2 &= 0,54 \\ 0,70x_1 + 0,50x_2 &= 0,38 \end{aligned}$$

ma rozwiązanie dokładne $\hat{x} = \begin{bmatrix} 0,80 \\ -0,36 \end{bmatrix}$ (patrz p. 5.3). Jeśli obliczymy wektor reszt na maszynie z arytmetyką zmiennopozycyjną o dwu miejscach dziesiętnych w mantysie, dokonującą zaokrągleń, to nie tylko dla wektora $\hat{x}$, ale także dla innych 86 rozwiązań uzyskamy $r = b - Ax = 0$, np. dla $x^{(1)} = \begin{bmatrix} 0,10 \cdot 10^1 \\ -0,64 \end{bmatrix}$ lub $x^{(87)} = \begin{bmatrix} 0,02 \\ 0,74 \end{bmatrix}$, natomiast $\|x^{(1)} - \hat{x}\|_\infty = 0,28$, $\|x^{(87)} - \hat{x}\|_\infty = 1,1$.

Należy podkreślić, że obliczony numerycznie wyznacznik wg wzoru $\det A = \text{fl}[a_{22}a_{11} - a_{12}a_{21}] = 0,01$, a więc powinno istnieć tylko jedno rozwiązanie. Wskaźnik uwarunkowania $K_\infty = 571,22$.

Przykład 2. Układ równań

$$\begin{aligned} 0,99x_1 + 0,70x_2 &= 0,53 \\ 0,70x_1 + 0,00x_2 &= 0,35 \end{aligned}$$

ma rozwiązanie dokładne $\hat{x} = \begin{bmatrix} 0,50 \\ 0,05 \end{bmatrix}$, natomiast obliczony numerycznie wektor reszt $r = b - A\hat{x} = \begin{bmatrix} -0,01 \\ 0,00 \end{bmatrix}$, a więc jest różny od zera! Wskaźnik uwarunkowania $K_\infty \approx 5,82$.

Przytoczone przykłady sugerują, że sprawdzanie (obliczanie reszt) należy wykonywać ze zwiększoną dokładnością, gdyż w przeciwnym razie możemy błędnie wnioskować o tym, czy otrzymane rozwiązanie spełnia układ równań, czy nie. Załóżmy zatem, że dla wyznaczonego rozwiązania x potrafimy dokładnie obliczyć wektor reszt $r = b - Ax$. Wówczas przy dokładnych obliczeniach moglibyśmy wyznaczyć poprawkę δx taką, że $x + \delta x = \hat{x}$. Jak łatwo bowiem sprawdzić δx jest rozwiązaniem układu równań $A\delta x = r$. Jeśli stosowaliśmy metodę GCW i zapamiętaliśmy macierze L i U , to omawiany układ możemy szybko rozwiązać wykonując tylko n^2 mnożeń i $n^2 - n$ dodawań. Ale wówczas wyznaczymy jedynie przybliżoną poprawkę $\delta x + \delta(\delta x)$. Powstaje pytanie, jakie warunki powinny być spełnione, by dla poprawionego rozwiązania $\bar{x} = x + \delta x + \delta(\delta x)$ wektor reszt miał mniejszą normę?

Lemat. *Jeżeli wektor reszt $r = b - Ax$ jest obliczony dokładnie, poprawka δx została wyznaczona metodą GCW i jeśli*

$$\frac{1}{2} \|r\|_\infty \geq \varepsilon \|A\|_\infty \left(\frac{9}{2} n^3 + \frac{61}{2} n^2 - 18n - 16 \right) \|\delta x\|_\infty + \|Ax\|_\infty \varepsilon, \text{ to wówczas}$$

$$\|b - A\bar{x}\|_\infty \leq \frac{1}{2} \|b - Ax\|_\infty$$

U w a g a. Aby sprawdzić powyższy warunek trzeba obliczyć poprawkę δx . Jeśli chcemy tego uniknąć, możemy skorzystać z oszacowania

$\|\delta x\|_\infty = \|A^{-1}r\|_\infty \leq \|A^{-1}\|_\infty \|r\|_\infty$ i wtedy, jeśli $\|r\|_\infty \gg \|Ax\|_\infty \varepsilon$, uzyskamy zastępczy warunek

$$K_\infty \varepsilon O\left(\frac{9}{2}n^3\right) \leq \frac{1}{2} \quad (5.48)$$

tzn. warunek, który zgodnie z (5.34) gwarantuje, iż układ $Ax = b$ można rozwiązać z błędem

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \leq 1.$$

Powyższy lemat daje podstawę do następującej *metody iteracyjnego poprawienia rozwiązania*:

1. Układ $Ax = b$ rozwiązujemy metodą GCW i otrzymujemy rozwiązanie $x^{(1)}$ (przyjmujemy, iż warunek (5.48) jest spełniony).
2. Sprawdzamy rozwiązanie obliczając wektor reszt $r^{(1)}$.
3. Sprawdzamy, czy $\|r^{(1)}\|_\infty > \|Ax^{(1)}\|_\infty \varepsilon$; jeśli NIE, to stop, jeśli TAK, to wykonaj 4.
4. Obliczamy poprawkę $\delta x^{(1)}$ i wyznaczamy $x^{(2)} = x^{(1)} + \delta x^{(1)}$.
5. Sprawdzamy rozwiązanie obliczając wektor reszt $r^{(2)}$; dalej postępujemy jak poprzednio, wykonując kolejno czynności 3, 4 i 5.

U w a g a. Liczbę $\|Ax^{(i)}\|_\infty$ można zastąpić przez $\|b\|_\infty$, jeżeli wektor $x^{(i)}$ jest bliski wektorowi będącemu rozwiązaniem.

Powyższa metoda umożliwia wyznaczenie rozwiązania, dla którego wektor reszt ma współrzędne rzędu $\|b\|_\infty \varepsilon$. Metoda ta jest przykładem iteracyjnej metody rozwiązywania układu równań $Ax = b$, tzn. takiego sposobu konstruowania ciągu przybliżonych rozwiązań $x^{(1)}, x^{(2)}, \dots, x^{(i)}, \dots$ by malała miara błędu rozwiązania (tu $\|r^{(i)}\|_\infty = \|b - Ax^{(i)}\|_\infty$).

5.4. METODY ITERACYJNE

Rozważmy ciąg wektorów $x^{(0)}, x^{(1)}, \dots, x^{(i)}, \dots$ określony za pomocą wzoru $x^{(0)}$ — wektor dany

$$x^{(i+1)} = Mx^{(i)} + w, \quad i = 0, 1, \dots \quad (5.49)$$

gdzie M jest pewną macierzą kwadratową, w — wektorem.

Tw. 1. Ciąg określony wzorem (5.49) przy dowolnym wektorze $x^{(0)}$ jest zbieżny do jedyne go punktu granicznego wtedy i tylko wtedy, gdy $\rho(M) < 1$.

DOWÓD. Ponieważ $x^{(i+1)} = Mx^{(i)} + w = M(Mx^{(i-1)} + w) + w = \dots = M^{(i+1)}x^{(0)} + (M^i w + M^{(i-1)} w + \dots + w)$, więc gdy $\rho(M) < 1$, wówczas zgodnie z twierdzeniem z p. 5.2 mamy $M^{(i+1)}x^{(0)} \xrightarrow{i \rightarrow \infty} 0$. Szereg $w + Mw + \dots + M^i w + \dots$ jest wtedy zbieżny, gdyż spełnia

warunek Cauchy'ego zbieżności szeregu: $\lim_{m,k \rightarrow \infty} (M^k w + M^{(k+1)} w + \dots + M^m w) = 0$, można

bowiem, zgodnie z lematem 2 z p. 5.2, wybrać taką normę macierzy i wektorów, by $\|M\| < 1$, a wtedy szereg $\|w\| + \|M\| \|w\| + \dots + \|M\|^i \|w\| + \dots$ jest zbieżny i spełnia warunek Cauchy'ego.

Jeśli $\rho(M) \geq 1$, lecz przy dowolnym $x^{(0)}$ otrzymalibyśmy $x^{(i+1)} \rightarrow \hat{x}$, to punkt $\hat{x}$ spełniałby równanie $\hat{x} = M\hat{x} + w$, czyli wektory $e^{(i+1)} = x^{(i+1)} - \hat{x}$, określone rekurencyjnie wzorami $e^{(i+1)} = Me^{(i)}$, $e^{(0)} = x^{(0)} - \hat{x}$, $i = 0, 1, \dots$, powinny dążyć do zera. Tymczasem $e^{(i+1)} = M^{(i+1)}e^{(0)}$, zgodnie zaś z twierdzeniem z p. 5.2 można wskazać taki wektor $e^{(0)}$ (tzn. wybrać odpowiedni wektor $x^{(0)}$), że $e^{(i+1)}$ nie dąży do zera.

Z nierówności (5.12) wynika, że dla zbieżności ciągu określonego wzorem (5.49) wystarczy warunek $\|M\| < 1$ dla dowolnej normy macierzy spełniającej warunek zgodności (np. dla normy indukowanej). Nie jest to jednak warunek konieczny.

Przykład 1. Macierz $M = \begin{bmatrix} \frac{1}{2} & 8 \\ 9 & \frac{1}{2} \end{bmatrix}$ zapewnia zbieżność ciągu określonego wzorem (5.49),

gdyż $\rho(M) = \frac{1}{2}$, lecz $\|M\|_\infty = 8,5$, a $\|M\|_2 > 8$.

W przeprowadzonym powyżej dowodzie stwierdziliśmy, że jeśli ciąg $(x^{(i)})$ jest zbieżny do pewnego wektora $x^{(\infty)}$, to spełnia on równanie $x^{(\infty)} = Mx^{(\infty)} + w$. Jeśli ponadto założymy $\rho(M) < 1$, to równanie $x = Mx + w$ ma jednoznaczne rozwiązanie (patrz lematy podane w p. 5.2). Wynika stąd sposób konstruowania metod iteracyjnych stosowanych do rozwiązywania układu równań $Ax = b$. Wystarczy mianowicie tak dobrać macierz M , by był spełniony warunek zbieżności $\rho(M) < 1$ i warunek zgodności $\hat{x} = M\hat{x} + w$, gdzie $\hat{x}$ jest rozwiązaniem układu $Ax = b$. Teoretycznie wystarczy zatem wziąć dowolną macierz M taką, by $\rho(M) < 1$, a następnie obliczyć wektor $w = (I - M)A^{-1}b$ wynikający z warunku zgodności. Ponieważ wymagałoby to obliczania wyrażenia $A^{-1}b$, więc w praktyce postępujemy odwrotnie: przyjmujemy, że wektor w jest równy Nb , gdzie N jest pewną macierzą kwadratową. Z warunku zgodności mamy wówczas $M = I - NA$ i w ten sposób otrzymujemy *rodzinę metod iteracyjnych* postaci

$$x^{(i+1)} = (I - NA)x^{(i)} + Nb \quad (5.50)$$

umożliwiających wyznaczenie wektora $\hat{x}$, jeśli tylko $\rho(I - NA) < 1$.

Przy pewnych szczególnych własnościach macierzy A istnieją proste metody wyboru różnych macierzy N , gwarantujące spełnienie warunku $\rho(I - NA) < 1$ (patrz p. 5.4.1—5.4.3). Powstaje wtedy problem określenia *miary efektywności* różnych metod iteracyjnych. Oprócz liczby działań arytmetycznych niezbędnych do wyznaczenia kolejnych wektorów $x^{(i)}$, liczby wykorzystywanych komórek pamięci maszyny cyfrowej i wielkości błędów zaokrągleń, miarą efektywności różnych metod iteracyjnych jest szybkość malenia wektora błędu $e^{(i)} = x^{(i)} - x$. Dla metod postaci (5.49) wyprowadzono ścisłe związki między promieniem spektralnym macierzy M , a szybkością malenia wektorów $e^{(i)}$. Jest to złożone zagadnienie, ograniczymy się więc tu jedynie do przedstawienia zasadniczych wyników (obszerniej są one omówione w [70]).

Z równości $\mathbf{x}^{(i+1)} = \mathbf{M}\mathbf{x}^{(i)} + \mathbf{N}\mathbf{b}$ oraz $\hat{\mathbf{x}} = \mathbf{M}\hat{\mathbf{x}} + \mathbf{N}\mathbf{b}$ otrzymujemy $\mathbf{e}^{(i+1)} = \mathbf{M}\mathbf{e}^{(i)}$. Wynika stąd, że jeśli $\|\mathbf{M}\| > 1$, to norma błędu $\|\mathbf{e}^{(i)}\|$ nie musi maleć z iteracji na iterację mimo spełnienia warunku zbieżności.

Przykład 2. Niech

$$\mathbf{M} = \begin{bmatrix} \frac{1}{2} & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & 1 \\ 0 & & & & & \frac{1}{2} \end{bmatrix}_{n \times n}, \quad \mathbf{w} = \begin{bmatrix} -\frac{1}{2} \\ \vdots \\ -\frac{1}{2} \\ +\frac{1}{2} \end{bmatrix}$$

Mamy tu $\rho(\mathbf{M}) = \frac{1}{2}$, a więc przy dowolnym wektorze $\mathbf{x}^{(0)}$ ciąg generowany wzorem (5.49) dąży do $\hat{\mathbf{x}} = [1, 1, \dots, 1]^T$. Jeśli przyjmimy $\mathbf{x}^{(0)} = \mathbf{0}$, to $\|\mathbf{e}^{(0)}\|_\infty = 1$. Łatwo jednak sprawdzić, że $\|\mathbf{e}^{(1)}\|_\infty = \frac{3}{2}$, $\|\mathbf{e}^{(2)}\|_\infty = \left(\frac{3}{2}\right)^2, \dots, \|\mathbf{e}^{(n-2)}\|_\infty = \left(\frac{3}{2}\right)^{n-2}$. Także pierwsze współrzędne kolejnych wektorów $\mathbf{x}^{(i)}$ wykazują początkowo znaczne odchylenia od właściwych wartości $\hat{x}_1 = 1$, ponieważ $x_1^{(1)} = -\frac{1}{2}$, $x_1^{(2)} = -\frac{1}{2} - \frac{3}{2}$, ..., $x_1^{(n-2)} = -\frac{1}{2} - \left(\frac{3}{2}\right)^{n-3}$. Niemniej jednak dla dostatecznie dużych indeksów i błąd $\mathbf{e}^{(i)}$ będzie mały i będzie dążył do zera, choć niekoniecznie monotonicznie.

Zauważmy, że powyższe zjawisko może uniemożliwić numeryczne wyznaczenie rozwiązania. Jeśli np. w podanym przykładzie $n = 100$, to pierwsza współrzędna wektora $\mathbf{x}^{(98)}$ będzie co do modułu większa niż 10^{15} , co może spowodować nadmiar maszynowy. Jeszcze bardziej niebezpieczny może być w praktyce nadmierny wówczas wzrost błędów zaokrągleń. Jeśli bowiem w każdej iteracji zamiast wartości $\mathbf{M}\mathbf{x}^{(i)} + \mathbf{w}$ obliczamy $\mathbf{x}^{(i+1)} = \mathbf{M}\mathbf{x}^{(i)} + \mathbf{w} + \delta^{(i)}$, gdzie $\delta^{(i)}$ jest błędem zaokrągleń, to chociaż może to być bardzo małe zaburzenie wartości $\mathbf{x}^{(i+1)}$, to jednak po większej liczbie iteracji, otrzymamy

$$\mathbf{x}^{(i+1)} = \mathbf{M}^{(i+1)}\mathbf{x}^{(0)} + \mathbf{M}^{(i)}\mathbf{w} + \dots + \mathbf{w} + \delta^{(i)} + \mathbf{M}\delta^{(i-1)} + \dots + \mathbf{M}^i\delta^{(0)} \quad (5.51)$$

W skrajnym przypadku może się zdarzyć, że łączny błąd zaokrągleń $\delta^{(i)} + \mathbf{M}\delta^{(i-1)} + \dots + \mathbf{M}^i\delta^{(0)}$ spowoduje uzyskanie $\mathbf{x}^{(i+1)} = \mathbf{x}^{(0)}$. Wtedy powstanie cykl generujący ciąg wektorów, który nie jest zbieżny do rozwiązania. Przed sytuacją taką ustrzec się jest trudno.

W przykładzie 2 pokazaliśmy, że za miarę efektywności metod iteracyjnych nie można przyjąć stosunku $\|\mathbf{e}^{(i+1)}\|/\|\mathbf{e}^{(i)}\|$. Skoro jednak $\rho(\mathbf{M}) < 1$ gwarantuje zbieżność, musi zatem następować efektywne zmniejszenie błędu po pewnej liczbie iteracji. Można wykazać (patrz ćwic. 2), że dla każdej liczby $\kappa \in (0; 1)$ istnieje taka liczba naturalna $m(\kappa)$, że $\frac{\|\mathbf{e}^{(i+m(\kappa))}\|}{\|\mathbf{e}^{(i)}\|} \leq \kappa$ przy dowolnym wektorze $\mathbf{x}^{(0)}$. Można wówczas powiedzieć, że przez $m(\kappa)$ iteracji błędy malały

średnio $\frac{m(\kappa)}{\sqrt{\kappa}}$ razy w każdej iteracji. Jeżeli za $m(\kappa)$ przyjmujemy najmniejszą z liczb naturalnych, dla których $\frac{\|e^{(i+m(\kappa))}\|_2}{\|e^{(i)}\|_2} \leq \kappa$ przy dowolnym $x^{(0)}$, to można udowodnić, że istnieje granica $\lim_{\kappa \rightarrow 0+} \frac{m(\kappa)}{\sqrt{\kappa}} = \varrho(M)$ (patrz [70, s. 67]). Zatem

promień spektralny $\varrho(M)$ jest pewną *asymptotyczną miarą efektywności* metody iteracyjnej.

Ćwiczenia

1. Uzasadnić tezę, że jeśli $x^{(i+1)} = Mx^{(i)} + Nb$, $\varrho(M) < 1$, $\det M \neq 0$, $\|M\| > 1$ ($\|\cdot\|$ oznacza pewną normę indukowaną), to można wskazać wektor $x^{(0)}$ taki, że dla pewnego indeksu K , jest $\|e^{(K)}\| / \|e^{(K-1)}\| = \|M\| > 1$.

2. Uzasadnić tezę, że jeśli $\varrho(M) < 1$, to dla każdego $\kappa \in (0; 1)$ istnieje indeks $m(\kappa)$ taki, że $\|e^{m(\kappa)}\| \leq \kappa \|e^{(0)}\|$ przy dowolnym $x^{(0)}$. Wskazać a. $M^{(i)} \rightarrow 0$, więc $\|M^{(i)}\| \leq \kappa$ dla dostatecznie dużych i .

5.4.1. Metoda Jacobiego

Metoda Jacobiego jest szczególnym przypadkiem metod (5.50). Niech macierz A układu równań $Ax = b$ będzie taką macierzą, że można ją rozłożyć na składniki $L + D + U$, gdzie L jest macierzą poddiagonalną, D — macierzą diagonalną, a U — macierzą nadaddiagonalną; np.

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 4 & 0 & 0 \\ 7 & 8 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{bmatrix} + \begin{bmatrix} 0 & 2 & 3 \\ 0 & 0 & 6 \\ 0 & 0 & 0 \end{bmatrix}$$

Przyjmując w (5.50) $N = D^{-1}$, czyli $M = M_J = -D^{-1}(L + U)$, otrzymamy tzw. *wzór Jacobiego*

$$Dx^{(i+1)} = -(L + U)x^{(i)} + b, \quad i = 0, 1, 2, \dots \quad (5.52)$$

Aby móc zastosować wzór (5.52), należy uprzednio tak zmienić kolejność równań układu $Ax = b$, aby na diagonalu były elementy niezerowe. W przypadku gdy wymiar macierzy jest duży, wprowadzenie takiej zmiany kolejności równań jest trudne. Istnieją jednak metody automatycznego przestawiania wierszy macierzy A . Jedną z tych metod można krótko określić następująco:

- spośród kolumn, w których na diagonalu znajduje się element zerowy, wybieramy tę, w której jest największa liczba zer;
- w kolumnie tej wybieramy element o maksymalnym module i tak przestawiamy wiersze, by wybrany element znalazł się na diagonalu; wiersz ten ustalamy i pomijamy go w dalszym badaniu wartości elementów;
- spośród pozostałych kolumn, w których na diagonalu znajduje się element zerowy, wybieramy tę, w której jest największa liczba zer;
- dokonując kolejnych przestawień wierszy doprowadzamy ostatecznie wyjściową macierz do takiej postaci macierzy, w której na diagonalu są elementy niezerowe.

Przykład 1. Dana jest macierz

$$A = \begin{bmatrix} 0 & 0 & 1 & 2 \\ 2 & 1 & 0 & 2 \\ 7 & 3 & 0 & 1 \\ 0 & 5 & 0 & 0 \end{bmatrix}$$

Postępując zgodnie z powyższą instrukcją, uzyskamy:

$$\begin{bmatrix} 0 & 0 & 1 & 2 \\ 2 & 1 & 0 & 2 \\ 7 & 3 & 0 & 1 \\ 0 & 5 & 0 & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 7 & 3 & 0 & 1 \\ 2 & 1 & 0 & 2 \\ 0 & 0 & 1 & 2 \\ 0 & 5 & 0 & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 7 & 3 & 0 & 1 \\ 2 & 1 & 0 & 2 \\ 0 & 0 & 1 & 2 \\ 0 & 5 & 0 & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 7 & 3 & 0 & 1 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 2 & 1 & 0 & 2 \end{bmatrix}$$

Metoda Jacobiego, jakkolwiek niezawodna, bywa pracochłonna, jeśli chcemy ją zastosować w obliczeniach na maszynie cyfrowej, ponieważ do maszyny należy wprowadzić macierz A z niezerową już diagonalą.

Jeśli w każdej kolumnie macierzy A znajduje się taka sama liczba elementów zerowych (co jest typowe dla wielu rzadkich macierzy), to dla uzyskania macierzy z niezerową diagonalą, wystarczy postępować identycznie jak przy wyborze elementu podstawowego w metodzie GCW. Należy podkreślić, że zastosowanie procedury zamiany wierszy, nie gwarantuje zbieżności metody, tzn. warunku $\rho(D^{-1}(L + U)) < 1$. Jeśli jednak macierz A jest silnie diagonalnie dominująca lub silnie diagonalnie dominująca kolumnowo, to metoda Jacobiego jest zbieżna, natomiast dla dodatnio określonej macierzy symetrycznej A metoda Jacobiego niekoniecznie jest zbieżna (patrz przykł. 1 w następnym punkcie).

Ćwiczenia

1. Udowodnić, że jeśli A jest macierzą silnie diagonalnie dominującą, to $\|M_J\|_\infty < 1$.

$$2. \text{ Wykazać, że dla macierzy } A = \begin{bmatrix} 1 & -\frac{1}{2} & \frac{1}{2} \\ 1 & 1 & 1 \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{bmatrix} \text{ metoda Jacobiego nie gwarantuje}$$

zbieżności przy dowolnym $x^{(0)}$.

5.4.2. Metoda Gaussa-Seidla

Metoda Gaussa-Seidla jest także szczególnym przypadkiem metod (5.50). Jeśli jak poprzednio, $A = L + D + U$, to przyjmując $N = (D + L)^{-1}$, otrzymamy $M_{GS} = -(D + L)^{-1}U$.

Ciąg rozwiązań $x^{(1)}, x^{(2)}, \dots$ obliczamy w tym przypadku ze wzoru

$$Dx^{(i+1)} = -Lx^{(i+1)} - Ux^{(i)} + b, \quad i = 0, 1, 2, \dots \quad (5.53)$$

Na pierwszy rzut oka powstać może wątpliwość, w jaki sposób obliczyć prawą stronę równości (5.53) nie znając $x^{(i+1)}$. Łatwo jednak sprawdzić, że przy obliczaniu pierwszej współrzędnej wektora w równości tej nie wystąpi żadna

współrzędna wektora $\mathbf{x}^{(i+1)}$, zatem dzieląc tę równość przez $d_1 = a_{11}$ obliczymy $x_1^{(i+1)}$. Następnie przy obliczaniu $x_2^{(i+1)}$ potrafimy obliczyć prawą stronę równości, gdyż zależy ona jedynie od $x_1^{(i+1)}$ oraz $\mathbf{x}^{(i)}$ itd. W ten sposób przy obliczaniu kolejnej składowej wektora $\mathbf{x}^{(i+1)}$ korzystamy z obliczonych już poprzednio składowych, co w wielu przypadkach podnosi efektywność tego algorytmu w porównaniu z algorytmem Jacobiego.

Aby móc stosować metodę Gaussa-Seidla, należy przyjąć taką kolejność równań $A\mathbf{x} = \mathbf{b}$, by elementy na diagonalu macierzy A były niezerowe. Nie daje to jednak w ogólnym przypadku gwarancji zbieżności metody. Jeśli potrafimy uzasadnić, że metoda Jacobiego jest dla danej macierzy A zbieżna oraz macierz M_J ma nieujemne elementy, to także metoda Gaussa-Seidla jest zbieżna. Ponadto jeśli $0 < \rho(M_J) < 1$, to wówczas $\rho(M_{GS}) < \rho(M_J)$ a więc metoda ta jest zbieżna asymptotycznie szybciej (patrz np. [70, s. 70]). Jeśli macierz A jest silnie diagonalnie dominująca lub silnie diagonalnie dominująca kolumnowo, to metoda Gaussa-Seidla jest zbieżna, choć niekoniecznie szybciej niż metoda Jacobiego. Gdy macierz A jest symetryczna, wówczas metoda Gaussa-Seidla jest zbieżna wtedy i tylko wtedy, gdy A jest dodatnio określona. Podsumowując własności obu metod, metody Jacobiego i Gaussa-Seidla, należy zdawać sobie sprawę, że jakkolwiek w wielu przypadkach obie metody są zbieżne, przy czym metoda G-S jest szybciej zbieżna, to w ogólnym przypadku nie ma takiej zależności. Ilustracją tego są następujące dwa przykłady.

Przykład 1. Macierz

$$A = \begin{bmatrix} 1 & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & 1 & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & 1 \end{bmatrix}$$

jest dodatnio określona, więc $\rho(M_{GS}) < 1$. Tymczasem $\det(\lambda I - M_J) = \lambda^3 - \lambda + \frac{2}{\sqrt{3}}$ i wielomian charakterystyczny dla $\lambda = -1$ ma wartość dodatnią. Istnieje zatem wartość własna $\lambda_{\min} < -1$, czyli $\rho(M_J) > 1$.

Przykład 2. Macierz

$$A = \begin{bmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{bmatrix}$$

Zatem

$$M_J = \begin{bmatrix} 0 & -2 & 22 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{bmatrix}, \quad \rho(M_J) = 0$$

$$\mathbf{M}_{\text{GS}} = \begin{bmatrix} 0 & -2 & 2 \\ 0 & 2 & -3 \\ 0 & 4 & -2 \end{bmatrix}, \quad \varrho(\mathbf{M}_{\text{GS}}) = 2\sqrt{2} > 1$$

Analiza błędów zaokrągleń. Zgodnie ze wzorem (5.51) uzyskany numerycznie wektor $\tilde{\mathbf{x}}^{(i+1)}$ różni się od dokładnego wektora $\mathbf{x}^{(i+1)}$ o łączny błąd zaokrągleń

$$\tilde{\mathbf{x}}^{(i+1)} - \mathbf{x}^{(i+1)} = \delta^{(i)} + \mathbf{M}\delta^{(i-1)} + \dots + \mathbf{M}^{(i)}\delta^{(0)}$$

Błędy $\delta^{(j)}, j = 0, 1, 2, \dots, i$, są różne, lecz mają wspólne oszacowanie $\|\delta^{(j)}\| < \kappa$ dla dostatecznie dużych j , jeśli algorytm iteracyjny jest zbieżny (tzn. ciąg wektorów $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ jest zbieżny do wektora $\hat{\mathbf{x}}$, który jest dokładnym rozwiązaniem układu $\mathbf{A}\mathbf{x} = \mathbf{b}$). Można bowiem wtedy przyjąć

$$\frac{1}{2}\|\tilde{\mathbf{x}}\| < \|\mathbf{x}^{(j)}\| \leq \|2\tilde{\mathbf{x}}\|, \quad \text{gdy} \quad \hat{\mathbf{x}} \neq \mathbf{0}, \text{ skąd otrzymujemy}$$

oszacowanie

$$\frac{\|\tilde{\mathbf{x}}^{(i+1)} - \mathbf{x}^{(i+1)}\|}{\|\mathbf{x}^{(i+1)}\|} \leq (1 + \|\mathbf{M}\| + \|\mathbf{M}^{(2)}\| + \dots + \|\mathbf{M}^{(i)}\|) \frac{2\kappa}{\|\hat{\mathbf{x}}\|}.$$

Jeśli $\|\mathbf{M}\| < 1$, to

$$\frac{\|\tilde{\mathbf{x}}^{(i+1)} - \mathbf{x}^{(i+1)}\|}{\|\mathbf{x}^{(i+1)}\|} < \frac{1}{1 - \|\mathbf{M}\|} \frac{2\kappa}{\|\hat{\mathbf{x}}\|}$$

Przykład 3. Oszacujemy błąd, z jakim metodą Gaussa-Seidla znajdujemy rozwiązanie równania $\mathbf{A}\mathbf{x} = \mathbf{b}$, gdy $\mathbf{A}$ jest macierzą silnie diagonalnie dominującą.

Kolejne przybliżenia rozwiązania $\hat{\mathbf{x}}$ obliczamy ze wzoru (5.53). Mamy więc

$$x_1^{(i+1)} = \frac{1}{d_{11}}(-u_{12}x_2^{(i)} - u_{13}x_3^{(i)} - \dots - u_{1n}x_n^{(i)} + b_1)$$

$$x_2^{(i+1)} = \frac{1}{d_{22}}(-l_{21}x_1^{(i+1)} - u_{23}x_3^{(i)} - \dots - u_{2n}x_n^{(i)} + b_2)$$

.....

$$x_n^{(i+1)} = \frac{1}{d_{nn}}(-l_{n1}x_1^{(i+1)} - l_{n2}x_2^{(i+1)} - \dots - l_{n,n-1}x_{n-1}^{(i+1)} + b_n)$$

Zgodnie z rozważaniami przeprowadzonymi w p. 3.2, obliczone współrzędne $x_1^{(i+1)}, x_2^{(i+1)}, \dots, x_n^{(i+1)}$ są takie, jakie otrzymalibyśmy stosując dokładne działania w przypadku macierzy $\tilde{\mathbf{A}} = \mathbf{A} + \delta\mathbf{A}$, gdzie $\|\delta\mathbf{A}\|_{1\infty} \leq \varepsilon(n+2)\|\mathbf{A}\|_{1\infty}$. Ponieważ macierz $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$ jest diagonalnie dominująca, więc $\|\mathbf{A}\|_{1\infty} = \|\mathbf{D}\|_{1\infty} = \|\mathbf{D}\|_{\infty}$. Dla $\mathbf{M}_{\text{GS}} = -(\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}$, mamy więc

$$\begin{aligned} \|\tilde{\mathbf{M}}_{\text{GS}} - \mathbf{M}_{\text{GS}}\|_{\infty} &= \|(\mathbf{D} + \mathbf{L} + \mathbf{E}_1)^{-1}(\mathbf{U} + \mathbf{E}_2) - (\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}\|_{\infty} = \\ &= \|[(\mathbf{D} + \mathbf{L} + \mathbf{E}_1)^{-1} - (\mathbf{D} + \mathbf{L})^{-1}]\mathbf{U} + (\mathbf{D} + \mathbf{L} + \mathbf{E}_1)^{-1}\mathbf{E}_2\|_{\infty} \leq \\ &\leq \|\mathbf{I} + (\mathbf{D} + \mathbf{L})^{-1}\mathbf{E}_1\|_{\infty} \|\mathbf{M}_{\text{GS}}\|_{\infty} + \\ &+ \|(\mathbf{D} + \mathbf{L})^{-1}\|_{\infty} \|(\mathbf{I} + (\mathbf{D} + \mathbf{L})^{-1}\mathbf{E}_1)^{-1}\|_{\infty} \|\mathbf{E}_2\|_{\infty} \leq \\ &\leq \frac{\|(\mathbf{D} + \mathbf{L})^{-1}\|_{\infty} \|\mathbf{E}_1\|_{\infty}}{1 - \|(\mathbf{D} + \mathbf{L})^{-1}\|_{\infty} \|\mathbf{E}_1\|_{\infty}} \|\mathbf{M}_{\text{GS}}\|_{\infty} + \frac{\|(\mathbf{D} + \mathbf{L})^{-1}\|_{\infty} \|\mathbf{E}_2\|_{\infty}}{1 - \|(\mathbf{D} + \mathbf{L})^{-1}\|_{\infty} \|\mathbf{E}_1\|_{\infty}} \end{aligned}$$

Zauważmy, że

$$\|(\mathbf{D} + \mathbf{L})^{-1}\|_{\infty} = \|(\mathbf{I} + \mathbf{D}^{-1}\mathbf{L})^{-1}\mathbf{D}^{-1}\|_{\infty} \leq \|(\mathbf{I} + \mathbf{D}^{-1}\mathbf{L})^{-1}\|_{\infty} \|\mathbf{D}^{-1}\|_{\infty}$$

oraz $\|(\mathbf{I} + \mathbf{D}^{-1}\mathbf{L})^{-1}\|_{\infty} \leq n$.

Ostatnią nierówność można uzasadnić korzystając z faktu, że macierz $\mathbf{I} + \mathbf{D}^{-1}\mathbf{L} = \mathbf{C}$ jest trójkątna dolna z jedynkami na diagonalu, diagonalnie dominująca. Należy oszacować normę $\|\cdot\|_1$ każdego wiersza macierzy $\mathbf{C}^{-1}$, której ostatni wiersz jest rozwiązaniem układu równań $\mathbf{C}^T \mathbf{x} = \mathbf{e}_n$, skąd

$$\begin{aligned} x_n &= 1, \quad x_{n-1} = -c_{n,n-1}, \quad x_{n-2} = c_{n,n-2}c_{n,n-1} - \\ &- c_{n-1,n-2}, \dots, x_1 = -c_{n,1}(\dots) - c_{n-1,1}(\dots) - \dots - c_{2,1} \end{aligned}$$

Symbolami (...) oznaczono tu iloczyny odpowiednich elementów macierzy $\mathbf{C}$, które są mniejsze niż 1 co do wartości bezwzględnej. Stąd

$$\begin{aligned} |x_n| + |x_{n-1}| + \dots + |x_1| &\leq 1 + (|c_{n,n-2}| + \dots + \\ &+ |c_{n,1}|) + (|c_{n-1,n-2}| + \dots + |c_{n-1,1}|) + \dots + |c_{2,1}| < n \end{aligned}$$

Macierze $\mathbf{E}_1$ i $\mathbf{E}_2$ oznaczają odpowiednio dolną trójkątną i naddiagonalną część macierzy $\delta\mathbf{A}$. Oczywiście $\|\mathbf{E}_1\|_{\infty} \leq \|\delta\mathbf{A}\|_{\infty} \leq \varepsilon(n+2)\|\mathbf{A}\|_{\infty}$ oraz $\|\mathbf{E}_2\|_{\infty} \leq \varepsilon(n+2)\|\mathbf{A}\|_{\infty}$. Ponieważ mamy $\|\mathbf{M}_{\text{GS}}\|_{\infty} < 1$ (patrz ćwic. 1), więc z powyższych oszacowań otrzymamy

$$\|\tilde{\mathbf{M}}_{\text{GS}} - \mathbf{M}_{\text{GS}}\|_{\infty} \leq \frac{2\alpha}{1-\alpha}, \quad \alpha = \varepsilon \cdot O(2n^2) \|\mathbf{D}\|_{\infty} \|\mathbf{D}^{-1}\|_{\infty}$$

Aby teraz oszacować $\|\delta^{(i)}\|_{\infty} = \|\tilde{\mathbf{M}}_{\text{GS}}\mathbf{x}^{(i)} + \tilde{\mathbf{N}}\mathbf{b} - \mathbf{M}_{\text{GS}}\mathbf{x}^{(i)} - \mathbf{N}\mathbf{b}\|_{\infty}$, trzeba jeszcze ograniczyć $\|\tilde{\mathbf{N}}\mathbf{b} - \mathbf{N}\mathbf{b}\|$. Mamy

$$\begin{aligned} \|(\tilde{\mathbf{N}} - \mathbf{N})\mathbf{b}\|_{\infty} &= \|[(\mathbf{D} + \mathbf{L} + \mathbf{E}_1)^{-1} - (\mathbf{D} + \mathbf{L})^{-1}]\mathbf{A}\hat{\mathbf{x}}\|_{\infty} \leq \\ &\leq \|(\mathbf{I} + (\mathbf{D} + \mathbf{L})^{-1}\mathbf{E}_1)^{-1} - \mathbf{I}\|_{\infty} \|(\mathbf{D} + \mathbf{L})^{-1}(\mathbf{L} + \mathbf{D} + \mathbf{U})\|_{\infty} \|\hat{\mathbf{x}}\|_{\infty} \leq \\ &\leq \frac{\alpha}{1-\alpha} \|\mathbf{I} - \mathbf{M}_{\text{GS}}\|_{\infty} \|\hat{\mathbf{x}}\|_{\infty} \leq \frac{2\alpha}{1-\alpha} \|\hat{\mathbf{x}}\|_{\infty} \end{aligned}$$

Zatem

$$\|\delta^{(i)}\|_{\infty} \leq \|\tilde{\mathbf{M}}_{\text{GS}} - \mathbf{M}_{\text{GS}}\|_{\infty} \|\mathbf{x}^{(i)}\|_{\infty} + \|(\tilde{\mathbf{N}} - \mathbf{N})\mathbf{b}\|_{\infty} \leq \frac{6\alpha}{1-\alpha} \|\hat{\mathbf{x}}\|_{\infty}$$

dla tych i , dla których $\|\mathbf{x}^{(i)}\|_{\infty} \leq 2\|\hat{\mathbf{x}}\|_{\infty}$.

Ostatecznie, po uwzględnieniu (5.54), otrzymamy

$$\frac{\|\tilde{\mathbf{x}}^{(i+1)} - \mathbf{x}^{(i+1)}\|_{\infty}}{\|\mathbf{x}^{(i+1)}\|_{\infty}} \leq \frac{1}{1 - \|\mathbf{M}_{\text{GS}}\|_{\infty}} \frac{12\alpha}{1-\alpha} \quad (5.55)$$

gdzie $\alpha = \varepsilon \cdot O(2n^2) \|\mathbf{D}\|_{\infty} \|\mathbf{D}^{-1}\|_{\infty}$.

Porównując oszacowanie (5.55) z (5.34) dla metody Gaussa, widzimy, że stosując iteracyjną metodę Gaussa-Seidla możemy zyskać na dokładności rozwiązania, jeśli wskaźnik $\|\mathbf{D}\|_{\infty} \|\mathbf{D}^{-1}\|_{\infty}$ jest mały w porównaniu z K_{∞} .

Ćwiczenia

1. Udowodnić, że jeśli $\mathbf{A}$ jest macierzą silnie diagonalnie dominującą, to $\|\mathbf{M}_{\text{GS}}\|_{\infty} < 1$.
W s k a z ó w k a. Ponieważ $(\mathbf{L} + \mathbf{D})^{-1}\mathbf{U} = (\mathbf{I} + \mathbf{D}^{-1}\mathbf{L})^{-1}\mathbf{D}^{-1}\mathbf{U}$, więc można przyjąć do roz-

ważną macierz A taką, że $D = I$. Wyznaczając wiersze macierzy $(I + L)^{-1}U$ i grupując wyrazy można wykazać, że $\|\cdot\|_1$ każdego wiersza jest mniejsza niż 1.

2. Udowodnić, że gdy A jest macierzą silnie diagonalnie dominującą kolumnowo, to jakkolwiek wówczas nie musi być $\|M_{GS}\|_\infty < 1$, ani $\|M_{GS}\|_1 < 1$, to jednak $\rho(M_{GS}) < 1$.

Wskazówka. Niech $A = L + I + U$, a $\bar{A} = -|L| + I - |U|$, gdzie zapis $|L|$ i $|U|$ oznacza, że należy użyć modułów elementów tych macierzy. Metoda Jacobiego dla A jest zbieżna, bo $\|\bar{M}_J\|_1 < 1$. Macierz $\bar{M}_J = |L| + |U|$ ma nieujemne elementy, więc $\rho(\bar{M}_{GS}) < 1$. Macierz $\bar{M}_{GS} = (I - |L|)^{-1}|U|$, więc $\rho(M_{GS}) \leq \rho(\bar{M}_{GS})$ (patrz tw. 15 w p. 6.2).

3. Dla jakich wartości α układ równań

$$\begin{aligned} x + y + z &= 1 \\ \alpha z &= 0 \\ \alpha y + z &= 0 \end{aligned}$$

można rozwiązać metodą Gaussa-Seidla przyjmując dowolny punkt początkowy?

Odpowiedź. 3. $\alpha \neq 0$.

5.4.3. Metoda Czebyszewa

Zgodnie z (5.50) można metodą iteracyjną $x^{(i+1)} = (I - NA)x^{(i)} + Nb$ konstruować ciąg wektorów $x^{(1)}, x^{(2)}, \dots$ zbieżny do rozwiązania układu równań $Ax = b$ tak wybierając macierz $N = N_s$, by $\rho(I - N_s A) < 1$. Jeśli macierz A jest symetryczna i dodatnio określona, można zrobić to następująco.

Zakładamy, że N_s jest pewnym wielomianem macierzowym $w(A)$ stopnia $s - 1$ (tzn. macierzą uzyskaną za pomocą wielomianu $w(t)$ po podstawieniu zamiast argumentu t macierzy kwadratowej A). Wówczas M_s jest wielomianem stopnia s postaci $p_s(t) = 1 - w(t)t$ dla $t = A$. Wartości własne $m_1, m_2, \dots, m_n$ macierzy M_s jako wielomianu obliczanego dla macierzy symetrycznej A o rzeczywistych wartościach własnych $a_1, a_2, \dots, a_n$ spełniają warunek $m_i = p(a_i)$, $i = 1, 2, \dots, n$. Jeśli dla $0 < \alpha < \beta$ znamy przedział $\langle \alpha; \beta \rangle$, w którym leżą wszystkie liczby $a_1, a_2, \dots, a_n$, to aby uzyskać $\rho(M_s) < 1$, wystarczy zapewnić spełnienie warunku $|p(t)| < 1$ dla $t \in \langle \alpha; \beta \rangle$.

Formułujemy następujące zagadnienie: Znaleźć wielomian $w(t)$ stopnia $(s - 1)$ taki, by

$$\max_{t \in \langle \alpha; \beta \rangle} |1 - w(t)t| \tag{5.56}$$

było najmniejsze z możliwych. Jest to klasyczny problem rozważany w teorii aproksymacji, przy czym wielomiany $w(t)$ znajduje się za pomocą tzw. wielomianów Czebyszewa (patrz p. 1).

Wielomian będący rozwiązaniem zagadnienia (5.56) jest dany wzorem

$$p_s(t) = \frac{T_s\left(\frac{2t - (\beta + \alpha)}{\beta - \alpha}\right)}{T_s\left(-\frac{\beta + \alpha}{\beta - \alpha}\right)}$$

Zatem otrzymujemy

$$\text{dla } s = 1: \quad p_1(t) = \frac{2}{\beta + \alpha} t, \quad M_1 = I - \frac{2}{\beta + \alpha} A, \quad N = \frac{2}{\beta + \alpha} A,$$

$$\varrho(M_1) \leq \frac{\beta - \alpha}{\beta + \alpha}$$

$$\text{dla } s = 2: \quad p_2(t) = 1 - t \frac{8(\beta + \alpha)}{\beta^2 + 6\beta\alpha + \alpha^2} + t^2 \frac{8}{\beta^2 + 6\beta\alpha + \alpha^2},$$

$$\varrho(M_2) \leq \frac{(\beta - \alpha)^2}{(\beta + \alpha)^2 + 4\beta\alpha}$$

Po wyznaczeniu współczynników wielomianu $p_s(t)$, możemy następnie obliczyć macierze M_s i N_s oraz punkt $\mathbf{x}^{(i+1)}$ ze wzoru (5.49), ale postępowanie takie byłoby wielce kłopotliwe przy użyciu maszyn cyfrowych, gdyż należałoby obliczać kolejne potęgi tych macierzy i sumować je z odpowiednimi współczynnikami. Można jednak postawić sobie za cel wyznaczenie wektora $\mathbf{x}^{(i+1)}$ bez obliczania macierzy M_s i N_s . Jeśli przez $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(s)}$ oznaczmy wektory $\mathbf{x}^{(i+1)}$, które otrzymamy stosując wielomiany stopnia $s = 1, s = 2, \dots, s$, to wektory te możemy wyznaczyć ze związków rekurencyjnych postaci

$$\begin{aligned} \mathbf{y}^{(0)} &= \mathbf{x}^{(i)}, \quad \mathbf{y}^{(1)} = \mathbf{x}^{(i)} - \frac{2}{\beta + \alpha} (A\mathbf{x}^{(i)} - \mathbf{b}) \\ \mathbf{y}^{(k+1)} &= \mathbf{y}^{(k)} + \omega_k \omega_{k-1} (\mathbf{y}^{(k)} - \mathbf{y}^{(k-1)}) - \\ &\quad - \frac{2}{\beta + \alpha} (1 + \omega_k \omega_{k-1}) - (A\mathbf{y}^{(k)} - \mathbf{b}), \quad k = 1, 2, \dots, s-1 \end{aligned} \quad (5.57)$$

Liczby $\omega_0, \omega_1, \dots$ obliczamy ze wzorów

$$\begin{aligned} \omega_0 &= \frac{\beta - \alpha}{\beta + \alpha} \\ \omega_{k+1} &= \frac{1}{2 \frac{\beta + \alpha}{\beta - \alpha} - \omega_k} \end{aligned} \quad (5.58)$$

(patrz np. [26, s. 359]). Ponieważ $\mathbf{y}^{(s)}$ jest poszukiwanym wektorem $\mathbf{x}^{(i+1)}$, można połączyć podstawowe iteracje $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i)}, \mathbf{x}^{(i+1)}, \dots$ z dodatkowymi iteracjami uzyskując jednolite wzory za pomocą następującego algorytmu:

$$\begin{aligned} \text{Krok 0.} \quad i &= 0, \quad \omega_0 = \frac{\beta - \alpha}{\beta + \alpha}, \quad c = \frac{2}{\beta + \alpha}, \\ L &= 2 \frac{\beta + \alpha}{\beta - \alpha}, \quad \mathbf{x}^{(-1)} = \mathbf{0} \end{aligned}$$

Krok 1. $k = 0$, $\mathbf{x}^{(i)} = \mathbf{x}_0$ — dany wektor, $\omega_{i-1} = 0$, $\omega_i = \omega_0$

$$\begin{aligned} \text{Krok 2. } \mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} + \omega_i \omega_{i-1} (\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}) - \\ - c(1 + \omega_i \omega_{i-1}) - (A\mathbf{x}^{(i)} - \mathbf{b}), \quad \omega_{i+1} = \frac{1}{L - \omega_i} \end{aligned} \quad (5.59)$$

Krok 3. $k = k + 1$, $i = i + 1$. Jeśli $k < s$, to wracamy do kroku 2. Jeśli $k \geq s$, to przyjmujemy $\mathbf{x}_0 = \mathbf{x}^{(i)}$ i wracamy do kroku 1.

Tw. Jeżeli A jest macierzą symetryczną dodatnio określoną, której wartości własne leżą w przedziale $\langle \alpha; \beta \rangle$, $0 < \alpha < \beta$, to dla dowolnego $s \geq 1$ i dowolnego początkowego wektora $\mathbf{x}_0$, ciąg $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, ... generowany przez algorytm (5.59) przy braku błędów zaokrągleń dąży do rozwiązania równania $A\mathbf{x} = \mathbf{b}$.

DOWÓD. Weźmy najpierw pod uwagę wektory $\mathbf{x}^{(0)}$, $\mathbf{x}^{(s)}$, $\mathbf{x}^{(2s)}$, ... w których następuje powrót do kroku 1. Zgodnie z konstrukcją algorytmu (5.59) mamy $\mathbf{x}^{((k+1)s)} = \mathbf{M}_s \mathbf{x}^{(ks)} + \mathbf{N}_s \mathbf{b}$. Ponieważ macierze $\mathbf{M}_s$ i $\mathbf{N}_s$ spełniają warunek zgodności, więc $\hat{\mathbf{x}} = \mathbf{M}_s \hat{\mathbf{x}} + \mathbf{N}_s \mathbf{b}$. Odejmując ostatnie dwie równości uzyskamy $\mathbf{e}^{((k+1)s)} = \mathbf{M}_s \mathbf{e}^{(ks)}$, gdzie $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}$. Stąd indukcyjnie znajdujemy $\mathbf{e}^{(ks)} = (\mathbf{M}_s)^k \mathbf{e}^{(0)}$. Niech teraz indeks i znajduje się między liczbami ks oraz $(k+1)s$, tzn. $i = ks + p$, $1 \leq p \leq s$. Zatem $\mathbf{x}^{(i)} = \mathbf{M}_p \mathbf{x}^{(ks)} + \mathbf{N}_p \mathbf{b}$, skąd $\mathbf{e}^{(i)} = \mathbf{M}_p \mathbf{e}^{(ks)} = \mathbf{M}_p (\mathbf{M}_s)^k \mathbf{e}^{(0)}$, ponieważ $\|\mathbf{M}_p\|_2 \leq \|\mathbf{M}_1\|_2 \leq \frac{\beta - \alpha}{\beta + \alpha}$, a także $\|\mathbf{M}_s\|_2 \leq \|\mathbf{M}_1\|_2 \leq \frac{\beta - \alpha}{\beta + \alpha}$, więc $\|\mathbf{e}^{(i)}\|_2 \rightarrow 0$.

Z podanego powyżej twierdzenia można wyciągnąć następujące wnioski:

Gdy $s = 1$, wówczas $\mathbf{e}^{(i)} = (\mathbf{M}_1)^i \mathbf{e}^{(0)}$. Ciąg wektorów $\mathbf{e}^{(0)}$, $\mathbf{e}^{(1)}$, ... dąży zatem do zera z prędkością zależną od $\varrho(\mathbf{M}_1) \leq \frac{\beta - \alpha}{\beta + \alpha}$. Ponieważ $\mathbf{M}_1$ jest macierzą symetryczną, więc istnieje taki wektor własny $\mathbf{v}$ macierzy $\mathbf{M}_1$, że jeśli $\mathbf{e}^{(0)} = \mathbf{v}$, to $\mathbf{e}^{(i)} = (\varrho(\mathbf{M}_1))^i \mathbf{e}^{(0)}$. Oznacza to, iż nie możemy oczekiwać zbieżności lepszej niż ze współczynnikiem $\omega_0 = \frac{\beta - \alpha}{\beta + \alpha}$.

Gdy $s = \infty$, tzn. nie dokonujemy odnowy algorytmu przez powtarzanie kroku 1 co s iteracji, wówczas $\mathbf{e}^{(i)} = \mathbf{M}_i \mathbf{e}^{(0)}$. Dowodzi się, że wtedy $\varrho(\mathbf{M}_i) = \omega_0 \omega_1 \dots \omega_i$, a ponieważ ciąg wartości $\omega_0, \omega_1, \dots$ jest malejący i zbieżny do

liczby $\omega_\infty = \frac{\sqrt{\beta} - \sqrt{\alpha}}{\sqrt{\beta} + \sqrt{\alpha}}$ (patrz ćwic. 1.), więc nie możemy oczekiwać zbieżności

lepszej niż ze współczynnikiem ω_∞ . Gdy wskaźnik β/α jest duży (a jest on co najmniej równy wskaźnikowi $K_2 = \|A\|_2 \|A^{-1}\|_2$), wówczas oznacza to, że przy danej wartości błędu $\|\mathbf{e}\|_2$ zatrzymującej obliczenia zysk ze zmniejszenia

liczby iteracji wynosi około $\sqrt{\frac{\beta}{\alpha}}$ razy.

Z dyskusji tej wnosić by można, że należy stosować jedynie metodę bez odnowy ($s = \infty$). W praktyce trzeba jednak dokonywać odnowy algorytmu (5.59) ze względu na narastanie błędów zaokrągleń w rozwiązaniach $\mathbf{x}^{(i)}$. Błędy zaokrągleń w każdej iteracji nie są jednak na ogół duże, więc zaleca się stosować duże wartości s .

Uprozczone oszacowanie błędów zaokrągleń. Oszacowanie całkowitego błędu zaokrągleń, jak np. dla metody Gaussa-Seidla, jest dla metody Czebyszewa trudne, więc ograniczymy się do analizy błędów powstających w jednej iteracji. Pozwoli to nam uchwycić główne przyczyny powstawania błędów zaokrągleń w tej metodzie.

Założmy, że reszty $A\mathbf{x}^{(i)} - \mathbf{b}$ obliczamy z precyzją η (np. w podwójnej precyzji $\eta = \varepsilon^2$, $\varepsilon = 2^{-t}$ lub 10^{-t}). Błąd popełniany przy obliczaniu reszt jest wtedy rzędu $\eta(k\|A\|_2\|\mathbf{x}^{(i)}\|_2 + \|\mathbf{b}\|_2)$. Współczynnik k ma na celu uwzględnienie wpływu rodzaju macierzy A na wielkość błędów powstających przy wyznaczaniu iloczynów skalarnych podczas obliczania $A\mathbf{x}^{(i)}$ i jest on oszacowany od góry przez $(n+1)\sqrt{n}$, lecz gdy np. macierz A ma tylko po trzy niezerowe elementy w każdym wierszu (np. macierz trójdzielna) wówczas $k = 4\sqrt{n}$. Zatem zamiast $\mathbf{x}^{(i+1)}$ ze wzorów algorytmu (5.59) otrzymamy $\mathbf{fl}(\mathbf{x}^{(i+1)}) = \mathbf{x}^{(i+1)} + \boldsymbol{\delta}^{(i)}$, gdzie

$$\|\boldsymbol{\delta}^{(i)}\|_2 \leq \varepsilon \left(16 \frac{\|A\mathbf{x}^{(i)} - \mathbf{b}\|_2}{\beta + \alpha} + 7 \|\mathbf{x}^{(i)}\|_2 \right) + \eta(k+2) \frac{4\|A\|_2\|\mathbf{x}^{(i)}\|_2}{\beta + \alpha}$$

Aby otrzymać powyższe oszacowanie skorzystano z warunku $\omega_i \omega_{i-1} < 1$ oraz z warunku $\|\mathbf{b}\|_2 \leq 2\|A\|_2\|\mathbf{x}^{(i)}\|_2$ (jest on spełniony, jeśli $\mathbf{x}^{(i)}$ przybliża rozwiązanie $\hat{\mathbf{x}}$). Ponadto

$$\|A\mathbf{x}^{(i)} - \mathbf{b}\|_2 \leq \|A\mathbf{x}^{(i)}\|_2 + \|\mathbf{b}\|_2 \leq 3\|A\|_2\|\mathbf{x}^{(i)}\|_2$$

Przyjmujemy, że liczba α jako ograniczenie od dołu wartości własnych macierzy A nie różni się znacznie od $a_{\min} = 1/\|A^{-1}\|_2$; np. $\alpha > \frac{1}{2\|A^{-1}\|_2}$. Stąd

$$\frac{\|\boldsymbol{\delta}^{(i)}\|_2}{\|\mathbf{x}^{(i)}\|_2} \leq \varepsilon \frac{48K_2}{1 + \frac{\beta}{\alpha}} + 7\varepsilon + \eta K_2(8k+16) \frac{1}{1 + \frac{\beta}{\alpha}}$$

Zauważmy, że $\frac{\beta}{\alpha} > K_2$, czyli $K_2/(1 + \frac{\beta}{\alpha}) < 1$, więc

$$\frac{\|\boldsymbol{\delta}^{(i)}\|_2}{\|\mathbf{x}^{(i)}\|_2} \leq \varepsilon \left(55 + \frac{\eta}{\varepsilon}(8k+16) \right) \quad (5.60)$$

Widzimy zatem, że błędy zaokrągleń nie zależą w sposób istotny od wymiaru macierzy n . Zależność tę można dodatkowo zmniejszyć obliczając reszty $A\mathbf{x}^{(i)} - \mathbf{b}$ w podwójnej precyzji (wtedy $\frac{\eta}{\varepsilon}(8k+16) \ll 1$).

U w a g a. Ponieważ do działania algorytmu są niezbędne obliczenia zarówno wektora reszt $A\mathbf{x}^{(i)} - \mathbf{b}$, jak i wektora zmian rozwiązania $\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}$, więc wielkości te mogą służyć jako bieżące wskaźniki efektywności algorytmu i mogą decydować o momencie dokonania odnowy. Łatwo również użyć ich jako testów stopu $\|\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}\| < \varrho$ lub $\|A\mathbf{x}^{(i)} - \mathbf{b}\| < \varrho$ przy założonej z góry małej liczbie dodatniej ϱ .

Ćwiczenia

1. Wykazać, że ciąg liczb $\omega_0, \omega_1, \dots$ generowany wzorem rekurencyjnym

$$\omega_0 = \frac{\beta - \alpha}{\beta + \alpha}, \quad \omega_{i+1} = \frac{1}{2 \frac{\beta + \alpha}{\beta - \alpha} - \omega_i}$$

$i = 0, 1, 2, \dots$ jest malejący, gdy $0 < \alpha < \beta$. Wykazać ponadto, że $\frac{\sqrt{\beta} - \sqrt{\alpha}}{\sqrt{\beta} + \sqrt{\alpha}} < \omega_i \leq \frac{\beta - \alpha}{\beta + \alpha}$

oraz $\omega_i \xrightarrow{i \rightarrow \infty} \frac{\sqrt{\beta} - \sqrt{\alpha}}{\sqrt{\beta} + \sqrt{\alpha}}$. Wskazać o w k a. 1) $\frac{1}{L - a} < a \Leftrightarrow a \in (p_1; p_2)$, gdzie p_1 i p_2 są pierwiast-

kami równania $-a^2 + La - 1 = 0$. 2) Jeśli $p_1 < a < 1$, to $\frac{1}{L - a} > p_1$.

5.4.4. Nakład obliczeń i testy stopu

W typowej metodzie iteracyjnej w każdej iteracji wykonujemy około n^2 mnożeń (jeśli nie traktujemy macierzy A jako rzadkiej — patrz p. 5.5.3). Dotyczy to w równym stopniu metod typu (5.49), a więc np. metody Jacobiego lub Gaussa-Seidla, jak i metody Czebyszewa. Metody dokładne, takie jak eliminacja Gaussa, wymagają około $\frac{1}{3}n^3$ mnożeń do uzyskania rozwiązania. Wynika stąd, że aby metody iteracyjne były porównywalne z metodami dokładnymi pod względem nakładu obliczeń, możemy dopuścić około n iteracji. Tymczasem, jak pokazują proste przykłady, liczba iteracji potrzebnych do uzyskania zadowalającej dokładności rozwiązania powinna być bardzo duża.

Przykład. Układ równań

$$\begin{bmatrix} 1 & \frac{3}{4} \\ \frac{3}{4} & 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 448 \\ 448 \end{bmatrix} \text{ ma rozwiązanie } \mathbf{x} = \begin{bmatrix} 256 \\ 256 \end{bmatrix}. \text{ Jeśli jednak zastosować do jego rozwiąza-}$$

nia metodę Gaussa-Seidla, przyjmując $\mathbf{x}^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, to po ośmiu iteracjach błąd $\|\mathbf{x}^{(8)} - \hat{\mathbf{x}}_2\|$ będzie nadal większy niż 0,1. Wykonano więc $8 \cdot 4 = 32$ mnożenia, natomiast metoda eliminacji Gaussa wymaga łącznie 6 mnożeń.

Niezbędną liczbę iteracji I potrzebną do zmniejszenia początkowego błędu $\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\|_2$ do wartości ϱ , jest trudno przewidzieć z góry, niemniej jednak z oszacowań podanych w przykładzie 2 w p. 5.4 wynika, że jest to liczba rzędu

$$I_0 = \max \left\{ n \left[\left(\ln \frac{\varrho}{\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\|_2} / \ln \varrho(M) \right) \right] \right\} \quad (5.61)$$

gdzie symbol $[\cdot]$ oznacza funkcję „część całkowita z...”. W powyższym przykładzie $\varrho(M_{GS}) = 9/16$, $\|x^{(0)} - \hat{x}\|_2 \approx 362$, $\left\lceil \left(\frac{\ln 0,1}{362} / \ln 9/16 \right) \right\rceil = 8$, a więc wynik zgodny z rzeczywistością.

Liczba iteracji I_0 nie uwzględnia błędów zaokrągleń spowalniających zbieżność metod, zatem w praktyce po I_0 iteracjach błąd wyznaczonego rozwiązania może być jeszcze duży. Z tego względu korzystając z metod iteracyjnych nie ograniczamy z góry liczby iteracji (ewentualnie zakładając dużą wartość I w celu uniknięcia zbyt długich obliczeń lub zapętlenia), lecz stosujemy *testy stopu*, najczęściej typu

$$\|x^{(i+1)} - x^{(i)}\| < \varrho \quad (5.62a)$$

lub

$$\frac{1}{\|b\|} \|Ax^{(i+1)} - b\| < \varrho \quad (5.62b)$$

Jeśli norma macierzy A jest mała, to wartość reszty $\|Ax^{(i+1)} - b\| = \|A(x^{(i+1)} - \hat{x})\|$ może być mała, mimo dużego odchylenia rozwiązania $x^{(i+1)}$ od rozwiązania $\hat{x}$. Także mała wartość $\|x^{(i+1)} - x^{(i)}\|$ nie musi świadczyć o tym, że wektory $x^{(i+1)}$ i $x^{(i)}$ są bliskie $\hat{x}$. Dla metod postaci (5.49) mamy bowiem

$$\|x^{(i+1)} - x^{(i)}\| = \|e^{(i+1)} - e^{(i)}\| = \|Me^{(i)} - e^{(i)}\| = \|(I - M)e^{(i)}\|$$

a więc gdy macierz $I - M$ ma małą normę, to mimo dużego błędu $e^{(i)}$ wektory $x^{(i+1)}$ i $x^{(i)}$ mogą mało się różnić.

Zauważmy, że $\|I - M\| \geq \varrho(I - M) = 1 - \varrho(M)$, zatem test stopu (5.62a) bywa zawodny, gdy $\varrho(M)$ jest bliskie 1. Jeśli zatem obserwujemy powolną zbieżność algorytmu, to należy raczej stosować test (5.62b). Jeżeli potrafimy oszacować od góry $\|A^{-1}\|$, to zamiast testu (5.62b) można stosować warunek

$$\|A^{-1}\| \|Ax - b\| < \varrho \quad (5.62c)$$

Mamy bowiem $\|x - \hat{x}\| = \|A^{-1}(A(x - \hat{x}))\| \leq \|A^{-1}\| \|Ax - b\|$, zatem spełnienie warunku (5.62c) zapewnia, że $\|x - \hat{x}\| < \varrho$.

Testy (5.62b) i (5.62c) mogą być mało przydatne z powodu błędów zaokrągleń. Jeśli bowiem przyjąć, że w każdym wierszu a_i , $i = 1, 2, \dots, n$, macierzy A jest nie więcej niż k elementów ($k \leq n$), to każda współrzędna wektora $Ax - b$ jest obliczona z błędem oszacowanym od góry przez $\varepsilon[k(|a_i|, |x|) + |b_i|]$, $i = 1, 2, \dots, n$, gdzie $|a_i|$ oraz $|x|$ oznaczają wektory o składowych będących modułami składowych wektorów a_i oraz x . Stąd

$$\begin{aligned} \|\text{fl}(Ax - b) - (Ax - b)\|_2 &\leq \varepsilon(k \|A\|_2 \|x\|_2 + \|b\|_2) \leq \\ &\leq \varepsilon(k\sqrt{n} \|A\|_2 \|x\|_2 + \|b\|_2) \end{aligned}$$

Założmy, że $\frac{1}{2} \|\hat{\mathbf{x}}\|_2 \leq \|\mathbf{x}\|_2 \leq 2 \|\hat{\mathbf{x}}\|_2$, co przy warunku zbieżności metody iteracyjnej jest spełnione dla dostatecznie dużych numerów iteracji. Wówczas po uzyskaniu nierówności (5.62c) mamy gwarancję, że

$$\frac{\|\mathbf{x} - \hat{\mathbf{x}}\|_2}{\|\hat{\mathbf{x}}\|_2} \leq \frac{2\varrho}{\|\mathbf{x}\|_2} + \varepsilon K_2 (k\sqrt{n} + 2) \quad (5.63)$$

Nie wystarczy zatem przyjęcie małej liczby ϱ w tekście (5.62c), ale także wektor reszt należy obliczyć z dużą dokładnością (małe ε), co może być istotne zwłaszcza przy pełnych ($k \approx n$) i źle uwarunkowanych ($K_2 \gg 1$) macierzach.

Ćwiczenia

1. Układ równań $\begin{bmatrix} 1 & b \\ b & c \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ rozwiązujemy metodą Jacobiego. Jakie warunki muszą spełniać liczby b i c , by macierz tego układu była dodatnio określona oraz by zarówno testy stopu (5.62a), jak i (5.62b) mogły prowadzić do błędnych wniosków przy szacowaniu odchyłń rozwiązań $\mathbf{x}^{(0)}$, $\mathbf{x}^{(1)}$, ... od $\hat{\mathbf{x}}$? Wskazówka. Liczba $\varrho(\mathbf{M}_J)$ powinna być bliska 1, norma $\|\mathbf{M}^{-1}\|_2$ powinna być duża i np. dla pewnego wektora $\mathbf{e}$ powinno być $\mathbf{M}_J \mathbf{e} = \varrho(\mathbf{M}_J) \mathbf{e}$ oraz $\mathbf{A} \mathbf{e} = \frac{1}{\|\mathbf{A}^{-1}\|} \mathbf{e}$.

2. Uzasadnić nierówność $\|\mathbf{A}\|_2 \leq \sqrt{n} \|\mathbf{A}\|_1$ i wskazać przykład macierzy $\mathbf{A}$, dla której $\|\mathbf{A}\|_2 = \sqrt{n} \|\mathbf{A}\|_1$; macierz $|\mathbf{A}|$ oznacza macierz o elementach $|a_{ij}|$. Wskazówka. Posłużyć się nierównością (5.6).

5.5. UKŁADY RÓWNAŃ Z MACIERZAMI RZADKIMI

Jeżeli mamy układ równań $\mathbf{A}\mathbf{x} = \mathbf{b}$ o tej własności, że w każdym równaniu (lub przynajmniej w wielu równaniach) liczba niewiadomych jest znacznie mniejsza niż liczba wszystkich niewiadomych n , to w macierzy $\mathbf{A}$ występuje dużo zer. Taką macierz nazywamy *macierzą rzadką*; przykładami macierzy rzadkich są macierze diagonalne, trójkątne, trójdagonalne itp.

Do rozwiązywania układów równań z macierzą rzadką można stosować ogólne metody, takie jak metoda eliminacji Gaussa lub metoda Doolittle'a, jednakże powstały również wyspecjalizowane metody rozwiązywania takich układów równań, korzystające w istotny sposób z informacji o rzadkości macierzy. Stosując np. metodę opisaną w p. 5.3.2, służącą do rozwiązywania układów z macierzami trójkątnymi, lub metodę opisaną w p. 5.3.6, służącą do rozwiązywania układów z macierzami trójdagonalnymi, wykonujemy mniejszą liczbę działań arytmetycznych i zajmujemy mniej komórek pamięci maszyny cyfrowej niż w przypadku zastosowania metody eliminacji Gaussa. Gdy mówimy o rozwiązywaniu układów równań z macierzą rzadką, mamy na myśli metody wyspecjalizowane, które zostaną obecnie krótko omówione.

Macierze rzadkie umożliwiają rozwiązanie układu $\mathbf{A}\mathbf{x} = \mathbf{b}$ przy mniejszej liczbie działań arytmetycznych niż w przypadku układu o tej samej liczbie

równań z macierzą gęstą, umożliwiającą wyznaczenie tego rozwiązania z większą dokładnością oraz dają oszczędność na pamięci maszyny.

5.5.1. Organizacja pamięci

Często zdarza się, że macierz rzadka ma charakterystyczną strukturę, co ułatwia zapamiętanie jej niezerowych elementów. Jako przykład przypomnimy macierz trójdagonalną omawianą w p. 5.3.6, kiedy to zamiast n^2 komórek, które należy zapamiętać dla gęstej macierzy A , wystarczyło zapamiętać ciągi $a_2, a_3, \dots, a_n; b_1, b_2, \dots, b_n; c_1, c_2, \dots, c_{n-1}$ w $3n - 2$ komórkach. Również macierze trójkątne warto zapamiętać w postaci ciągu liczb, np. $l_{11}, l_{21}, \dots, l_{n1}, l_{22}, l_{32}, \dots, l_{n2}, \dots, l_{nn}$ zajmując $n(n + 1)/2$ komórek. Należy w tym celu zadeklarować odpowiednią macierz jednoindeksową, na przykład (w Fortranie)

```
.....
.....
NDIM = N * (N + 1) / 2
CALL METODA(NDIM, L, ..., ...)
.....
.....
SUBROUTINE METODA(NDIM, L, ..., ...)
DIMENSION L(NDIM)
.....
.....
```

Oczywiście, gdybyśmy zadeklarowali

```
.....
.....
DIMENSION L(N, N)
.....
.....
```

to jakkolwiek wyczytalibyśmy niezerowe wartości tylko do $n(n + 1)/2$ komórek, zarezerwowalibyśmy niepotrzebnie n^2 komórek. Możemy jednak wykorzystać górną część macierzy L do zapamiętania innej trójkątnej macierzy, a wówczas dwuindeksowa macierz L uprości (na ogół) program korzystający z jej elementów.

W ogólnym przypadku dla macierzy rzadkiej, w której nie potrafimy dopatrzeć się prostej prawidłowości w położeniu elementów niezerowych, stosuje się trzy typowe metody zapisu (patrz np. [53, s. 37]):

1) Zapamiętujemy kolumnami niezerowe elementy, np. $a_{11} a_{21} a_{12} a_{22} a_{32}$ $a_{23} a_{33} a_{43} a_{34} a_{44}$ oraz w oddzielnej macierzy całkowitoliczbowej numery wierszy, tu: 1, 2, 1, 2, 3, 2, 3, 4, 3, 4, a także znaczniki początku kolumn w podanych numerach wierszy (łącznie z nieistniejącą $(n + 1)$ -szą), tu: 1, 3, 6, 9, 11. Tu na przykład zapamiętano trójdagonalną macierz o wymiarze 4×4 .

2) Tak jak 1), lecz zamiast zapamiętywać numery wierszy, zapamiętujemy tylko pierwszy i ostatni w każdej kolumnie oraz dodatkowo słowami bitowymi utrwalamy położenie pośrednich niezerowych elementów. Na

przykład dla kolumny $a_{12} a_{32} a_{43} a_{72} a_{83}$ należy zapamiętać liczby 1 i 8 oraz kod bitowy 011001 wskazujący elementy $a_{32} a_{43} a_{72}$.

3) Elementy macierzy zapamiętujemy w dowolnej kolejności, dołączając do nich informacje o położeniu sąsiednich elementów w ich wierszach i kolumnach lub dodatkowe informacje, np. położenie ostatniego elementu w danej kolumnie.

O tym jaką metodę zapamiętywania niezerowych elementów przyjmiemy decyduje w gruncie rzeczy specyfika algorytmu, w którym będziemy używać rzadkiej macierzy. I tak na przykład pierwszy sposób kodowania jest bardzo prosty i oszczędny pod względem wykorzystania pamięci, ale jeśli dokonujemy eliminacji zmiennych za pomocą algorytmu Gaussa, spowoduje to pojawienie się nowych niezerowych i zerowych elementów, a zatem nastąpi wielokrotna reorganizacja zapamiętanych macierzy. Sposób trzeci jest wówczas lepszy, gdyż ewentualny nowy niezerowy element wystarczy po prostu dołączyć na końcu listy dotychczasowych elementów wraz z odpowiednimi informacjami dotyczącymi jego położenia w macierzy. Jeśli jednak przez cały czas działania algorytmu nie będziemy dokonywać zmian w macierzy A (tak jest np. w większości iteracyjnych metod rozwiązywania układu równań $Ax = b$), to sposób trzeci jest nieodpowiedni jako zbyt skomplikowany.

5.5.2. Metody dokładne dla układów z macierzami rzadkimi

Metoda eliminacji Gaussa, a także wywodzące się z niej metody typu LU , LDU , LDL^T , LL^T w ogólnym przypadku macierzy rzadkich nie znalazły jeszcze szerokiego zastosowania. Jest to spowodowane przede wszystkim brakiem skutecznych metod wyboru elementu podstawowego, które by z jednej strony zapewniały niezawodność i stabilność numeryczną (patrz p. 5.3.3), z drugiej zaś strony nie powodowały pojawienia się dużej liczby nowych niezerowych elementów. W niektórych jednak przypadkach warto skorzystać z metod stosowanych dla macierzy gęstych, np. gdy macierz A jest trójdzielna i diagonalnie dominująca kolumnowo. Wówczas podczas eliminacji zmiennych nie pojawiają się dodatkowe niezerowe elementy (patrz p. 5.3.6), lecz często powoduje to niekontrolowane zapełnianie macierzy $A^{(2)}$, $A^{(3)}$, ..., $A^{(n)} = U$, a także dużą liczbę niezerowych mnożników w macierzy L , co pokazuje poniższy przykład.

Przykład. Macierz A postaci

$$\begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 1 & \dots & 1 \end{bmatrix}$$

o nieparzystej liczbie n kolumn i wierszy, jest skonstruowana w ten sposób, że jej kolumny i wiersze składają się na przemian z samych jedynek lub zer, tzn.

$$\begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & & & & \\ \vdots & & & & \\ 1 & & & & \end{bmatrix} \quad \text{lub} \quad \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & & & & \\ \vdots & & & & \\ 0 & & & & \end{bmatrix}$$

Jest to macierz rzadka i ma tę własność, że łatwo tu uchwycić prawidłowość położenia niezerowych elementów i można je zapamiętać w $n(n+1)/2$ komórkach (zapełnienie około 50%). Jeśli jednak zastosujemy do rozwiązania układu równań $Ax = b$ metodę eliminacji Gaussa z częściowym wyborem elementu podstawowego, to uzyskamy rozkład $A = LU$, gdzie

$$L = \begin{bmatrix} 1 & & & & \\ 1 & 1 & & & 0 \\ 1 & 1 & 1 & & \\ \dots & & & & \\ 1 & 1 & 1 & \dots & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ & -1 & -1 & \dots & -1 \\ & & 1 & \dots & 1 \\ & & & \dots & \\ & 0 & & & 1 \end{bmatrix}$$

Jeśli zatem chcemy rozwiązywać równania $Ax = b$ przy różnych wektorach b korzystając z rozkładu LU , to nastąpi całkowite zapełnienie początkowo rzadkiej macierzy.

Pewnym uogólnieniem przypadku macierzy trójdzielnych są tzw. *macierze wstęgowe*. Są to macierze, których elementy niezerowe są skupione wokół diagonal. Zapisujemy to w postaci warunku

$$a_{ij} = 0 \quad \text{dla} \quad |i - j| > m$$

Zauważmy, że $m = 0$ odpowiada macierzy diagonalnej, $m = 1$ — macierzy trójdzielnej i odpowiednio $m = n - 1$ dopuszcza pełną macierz. Prawdziwe jest następujące twierdzenie.

Tw. *Jeśli macierz wstęgowa A o danej szerokości wstęgi $2m + 1$ jest diagonalnie dominująca kolumnowo lub jest dodatnio określoną macierzą symetryczną, to wybór elementów podstawowych kolejno pojawiających się na diagonal jest prawidłowy (tzn. są to liczby niezerowe), zagwarantowana jest stabilność numeryczna metody oraz uzyskane macierze L i U (a także wszystkie pośrednie $A^{(2)}, A^{(3)}, \dots, A^{(n)} = L$) są wstęgowe z tą samą liczbą m . Dla macierzy dodatnio określonej i symetrycznej można także zastosować metodę Choleskiego uzyskując rozkład $A = LL^T$, w którym dolna macierz trójkątna L jest wstęgowa z liczbą m .*

W przypadku macierzy wstęgowych możemy zatem korzystać z metod i wzorów podanych w p. 5.3.3 i 5.3.5, zapamiętując uzyskane rozkłady LU , LDL^T , LL^T w mniejszej liczbie komórek pamięci. Można także, jak w przypadku macierzy trójdzielnych wyprowadzić analogicznie uproszczone wzory rozwiązywania układów równań. Należy zdawać sobie także sprawę z faktu, że oszacowanie błędów zaokrągleń dla macierzy wstęgowych będzie zależeć nie od wymiaru n , ale od liczby m . Spotkaliśmy się z tym faktem już przy omawianiu macierzy trójdzielnych. Dowodzi się na przykład (patrz

[67, s. 60]), że rozkład Choleskiego umożliwia obliczenie takiej macierzy L , że $LL^T = A + E$ oraz

$$\|E\|_2 \leq \varepsilon (m + 1)^2 \|A\|_2 \quad (5.64)$$

Ćwiczenia

1. Jak można rozmieścić w pamięci macierz z podanego powyżej przykładu, by wykorzystać fakt, że jest ona rzadka?
2. Udowodnić, że jeśli $A_{3 \times 3}$ ma rozkład $A = LU$ taki, że wszystkie elementy $l_{ij}, i \geq j$, oraz $u_{ij}, i \leq j$, są niezerowe, to co najmniej 6 elementów macierzy A jest niezerowych. Podać przykład macierzy A o 6 niezerowych elementach, dla której L i U są pełnymi macierzami trójkątnymi.

5.5.3 Metody iteracyjne

W punkcie 5.4.5 omawialiśmy nakład obliczeń niezbędnych do uzyskania rozwiązania metodą iteracyjną i stwierdziliśmy, że dla macierzy pełnych jest on na ogół znacznie większy niż przy stosowaniu metod dokładnych. Jednak w przypadku macierzy rzadkich metody iteracyjne mogą być istotnie lepsze niż metody dokładne. Jeśli na przykład macierz M w metodzie iteracyjnej $\mathbf{x}^{(i+1)} = M\mathbf{x}^{(i)} + \mathbf{b}$ jest wstęgowa i $m_{ij} = 0$ dla $|i - j| > k$, to w każdej iteracji wykonujemy nie n^2 mnożeń, lecz nk mnożeń. Wówczas nakład obliczeń w porównaniu z metodą dokładną będzie zbliżony nie dla $I \approx n$ iteracji, ale dla $I \approx \frac{n^2}{k}$ iteracji, co w przypadku dużych wymiarów ($n \gg 1$) i wąskich macierzy ($k \ll n$) może wystarczać do znalezienia rozwiązania z wymaganą dokładnością. Nie nakład obliczeń jednak decyduje o użyteczności metod iteracyjnych dla macierzy rzadkich, lecz fakt, że podczas obliczeń nie musimy zmieniać położenia elementów macierzy A , a zatem w łatwy sposób zachowujemy jej rzadką strukturę. Możemy wówczas przyjąć jakąś prostą metodę zapamiętania macierzy A (patrz p. 5.5.1) oraz możemy korzystać z prostych algorytmów obliczeniowych. Nie można jednak kategorycznie stwierdzić, że w przypadku macierzy rzadkich metody iteracyjne są najbardziej wskazane, ponieważ w przypadkach szczególnych, gdy np. $\rho(M) \approx 1$, wymagana liczba iteracji może być bardzo duża (patrz wzór (5.61)), a także mogą nastąpić cykliczne zapętlenia lub nawet nadmiar maszynowy (patrz p. 5.4). Oczywiście środkiem zaradczym może się wtedy okazać wybór innej metody iteracyjnej. Przegląd stosowanych metod iteracyjnych można znaleźć w [22] lub [70].

5.5.4 Metody blokowe

Kompromis między prostotą metod iteracyjnych a efektywnością metod dokładnych stanowią tzw. *metody blokowe*. Przypominamy, że najprostsze metody iteracyjne, metoda Jacobiego i metoda Gaussa-Seidla, powstały w wyniku podziału macierzy A na sumę $A = L + D + U$; macierz D była diagonalna i wykorzystano to we wzorach obliczeniowych, w których

występuje odwracanie macierzy D . Można jednak za macierz D przyjąć inną, również łatwą do odwrócenia macierz, np. macierz trójdagonalną. Wówczas odpowiednikiem iteracyjnej metody Jacobiego byłaby metoda

$$Dx^{(i+1)} = -Lx^{(i)} - Ux^{(i)} + b \quad (5.65)$$

przy czym po obliczeniu prawej strony nastąpiłoby rozwiązanie układu równań $Dx^{(i+1)} = w$, zgodnie ze wzorami (5.39) i (5.40). Zauważmy, że wzory (5.39) wystarczy zastosować jedynie w pierwszej iteracji, a następnie korzystać tylko z wzorów (5.40) wykonując $3n - 2$ mnożeń.

Podobnie można zmodyfikować metodę Gaussa-Seidla. Są to jednak nadal metody iteracyjne i jakkolwiek przy poprawnym wyborze macierzy D można oczekiwać szybszej zbieżności (mniejszy promień spektralny odpowiedniej macierzy M), to w niekorzystnych przypadkach zbieżność będzie bardzo powolna, a nawet może być uniemożliwiona na skutek zwiększonych błędów zaokrągleń przy odwracaniu macierzy D .

W pewnych szczególnych przypadkach można dla macierzy rzadkich zaproponować metody dokładne, powstałe przez odpowiedni podział macierzy A na bloki. Załóżmy na przykład, że macierz A układu równań $Ax = k$ można podzielić następująco

$$\begin{bmatrix} B_1 & C_1 & & & 0 \\ A_2 & B_2 & C_2 & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot & C_{N-1} \\ 0 & & & \cdot & \cdot & \\ & & & A_N & B_N \end{bmatrix}$$

gdzie B_i , $i = 1, \dots, N$, są macierzami trójdagonalnymi, a $C_1, \dots, C_{N-1}$ oraz $A_2, \dots, A_N$ — macierzami diagonalnymi (taki przypadek występuje np. przy rozwiązywaniu równania Laplace'a lub Poissona z warunkami granicznymi Dirichleta metodami różnicowymi z równomiernym podziałem prostokątnym). Wówczas można zastosować metodę będącą prostym uogólnieniem metody (5.41). Mamy

$$U_1 = B_1, \quad Y_1 = K_1$$

Dla $i = 2, 3, \dots, N$ obliczamy macierze

$$L = A_i U_{i-1}^{-1}, \quad U_i = B_i - LC_{i-1}, \quad Y_i = K_i - LY_{i-1} \quad (5.66)$$

a następnie

$$X_N = U_N^{-1} Y_N$$

$$X_i = U_i^{-1}(Y_i - C_i Y_{i+1}), \quad i = N-1, \dots, 1$$

Przez K_i oznaczyliśmy tu podwektory k odpowiadające podziałowi na bloki

B_i , a przez Y_i podwektory pomocnicze o wymiarze równym wymiarowi K_i . Zwróćmy uwagę, że jakkolwiek macierze A_i oraz C_i są diagonalne (są to często macierze jednostkowe), to mimo że macierze B_i są trójdzielne, macierze L i U_i są na ogół gęste. Powyższą metodę można więc stosować ze względów praktycznych jedynie wtedy, gdy wymiary macierzy B_i nie są zbyt duże. Ponadto oszacowania błędów zaokrągleń będą silnie zależeć od wymiaru tych macierzy. Bliższe szczegóły dotyczące metod blokowych można znaleźć np. w [70].

Ćwiczenia

1. Macierz A podzielono na sumę $A = E + F + G$ w ten sposób, że odwrotność macierzy F łatwo jest obliczyć. Zaproponowano metodę iteracyjną analogiczną do metody Gaussa-Seidla, przyjmując $M = -(E + F)^{-1}G$. Jaki warunek powinna spełniać macierz E , by można było korzystać ze wzorów $Fx^{(i+1)} = -Ex^{(i+1)} - Gx^{(i)} + b$?

2. Obliczyć liczbę działań potrzebnych do znalezienia rozwiązania układu równań $Ax = k$ według wzorów (5.66).

6

OBLICZANIE WARTOŚCI WŁASNYCH I WEKTORÓW WŁASNYCH MACIERZY

6.1. WSTĘP

Macierz kwadratowa $A_{n \times n}$ może być interpretowana jako pewien sposób przekształcania n -wymiarowych wektorów, mianowicie każdemu wektorowi $x \in \mathbb{R}^n$ można przyporządkować wektor y równy Ax . Działanie operacji liniowej A na wektor x przejawia się na ogół w zmianie długości wektora x (norma $\|y\|_2$ nie jest na ogół równa normie $\|x\|_2$) oraz w zmianie kierunku wektora (wektor y nie jest zazwyczaj równoległy do wektora x); odróżnia to operację liniową A od mnożenia wektora x przez skalar. Jeśli dla pewnego niezerowego wektora v okaże się, że wektor Av jest równoległy do v , to mówimy, że wektor v jest *wektorem własnym* macierzy A . Istnieje zatem pewien skalar λ taki, że $Av = \lambda v$. Liczbę λ nazywamy *wartością własną* macierzy A . Widzimy więc, że wynik działania macierzy A na wektor własny sprowadza się do pomnożenia tego wektora przez liczbę. Własność ta powoduje, że w wielu zagadnieniach mechaniki, fizyki, chemii i matematyki pożądane jest wyznaczenie wektorów i wartości własnych, gdyż umożliwia to jakościowe uproszczenie rozwiązywania zagadnienia.

Sposób wyznaczenia numerycznego wektorów i wartości własnych macierzy A zależy nie tylko od jej cech szczególnych, ale także od tego, czy mamy wyznaczyć wszystkie wartości lub wektory własne, czy też tylko wartość własną o największym module lub kilka największych lub najmniejszych wartości własnych. Na przykład, stosując metodę Czebyszewa rozwiązywania układu równań $Ax = b$, trzeba na wstępie wskazać liczby α i β odpowiednio bliskie najmniejszej i największej wartości własnej macierzy A . W tym szczególnym przypadku nie musimy obliczać wartości własnych z dużą dokładnością, choć ma to pewien wpływ na szybkość zbieżności metody Czebyszewa.

Wartości własne macierzy A są pierwiastkami *wielomianu charakterys-*

tycznego $w(\lambda) \equiv \det(A - \lambda I)$ tej macierzy; I oznacza macierz jednostkową. Przy obliczeniach na maszynie cyfrowej rzadko jednak posługujemy się tym wielomianem, gdyż wyznaczenie jego współczynników z wymaganą dokładnością jest zwykle czasochłonne, a także rozwiązanie równania $w(\lambda) = 0$ jest numerycznie trudne (patrz p. 3.1.4). Dlatego na ogół macierz A przekształcamy do postaci dogodniejszej do wyznaczenia jej wartości własnych, a następnie korzystamy ze szczególnych cech takiej macierzy. Wymaga to dobrej znajomości wielu pojęć i twierdzeń z zakresu algebry liniowej. Najważniejsze z nich przytaczamy poniżej.

6.2. POJĘCIA PODSTAWOWE

Przyjmujemy, że elementy macierzy są liczbami rzeczywistymi, jeśli nie podkreślamy dodatkowo, że mogą to być liczby zespolone.

Def. 1. Niezerowy wektor x o składowych rzeczywistych lub zespolonych nazywamy *wektorem własnym* macierzy $A_{n \times n}$, jeżeli istnieje taka liczba λ (rzeczywista lub zespolona), że $Ax = \lambda x$. Liczbę λ nazywamy *wartością własną* macierzy A .

Tw. 1. Liczba λ jest wartością własną macierzy A wtedy i tylko wtedy, jeśli jest pierwiastkiem wielomianu charakterystycznego $\det(A - \lambda I)$ macierzy A .

Def. 2. Wektory własne i wartości własne macierzy transponowanej A^T nazywamy odpowiednio *lewostronnymi wektorami* i *lewostronnymi wartościami własnymi* macierzy A .

Def. 3. Pierwiastki $\lambda_1, \lambda_2, \dots, \lambda_n$ wielomianu charakterystycznego (z uwzględnieniem ewentualnych wielokrotności tych pierwiastków) nazywamy *widmem* macierzy $A_{n \times n}$. Zbiór liczb $\lambda_1, \lambda_2, \dots, \lambda_n$ oznaczamy $\text{Sp}(A)$.

Wniosek. Widmo macierzy A jest równe widmu lewostronnemu, czyli $\text{Sp}(A) = \text{Sp}(A^T)$.

Tw. 2. Niech λ_L i λ_P będą odpowiednio lewostronną wartością własną i wartością własną macierzy A , a x_L i x_P odpowiadającymi tym wartościom lewostronnym wektorem własnym i wektorem własnym. Jeżeli $\lambda_L \neq \lambda_P$, to wektory x_L i x_P są ortogonalne, czyli $\langle x_L, x_P \rangle = 0$.

DOWÓD. $\langle x_L, Ax_P \rangle = \lambda_P \langle x_L, x_P \rangle$ oraz $\langle x_L, Ax_P \rangle = \langle A^T x_L, x_P \rangle = \lambda_L \langle x_L, x_P \rangle$, więc $(\lambda_L - \lambda_P) \langle x_L, x_P \rangle = 0$.

Wniosek. Ponieważ dla macierzy symetrycznej $A = A^T$ wektory własne są zarazem lewostronnymi wektorami własnymi, więc wektory własne odpowiadające różnym wartościom własnym są ortogonalne.

Tw. 3. Jeżeli $p(t)$ jest wielomianem zmiennej t , liczby $\lambda_1, \lambda_2, \dots, \lambda_n$ tworzą widmo macierzy $A_{n \times n}$, to widmem macierzy $p(A)$ są liczby $p(\lambda_1), p(\lambda_2), \dots, p(\lambda_n)$. Wektory własne macierzy A są wektorami własnymi macierzy $p(A)$.

Tw. 4. (Cayleya-Hamiltona). *Jeżeli w (λ) jest wielomianem charakterystycznym macierzy A , to w (A) jest macierzą zerową.*

Def. 4. Mówimy, że macierze A i B są podobne, jeżeli istnieje taka nieosobliwa macierz P , zwana macierzą podobieństwa, że $P^{-1}AP = B$.

Tw. 5. *Jeżeli macierze A i B są podobne, to $\text{Sp}(A) = \text{Sp}(B)$.*

Def. 5. Macierz $Q_{m \times n}$, $m \geq n$, nazywamy ortogonalną, jeżeli $Q^T Q = I_{n \times n}$.

U w a g a. Macierz ortogonalna jest szczególnym przypadkiem macierzy unitarnej, tj. takiej macierzy S o elementach w ogólności zespolonych, że $S^* S = I$; symbol S^* oznacza tu macierz sprzężoną z S , czyli macierz transponowaną z zamianą elementów zespolonych na sprzężone.

Tw. 6. *Jeżeli macierz $Q_{n \times n}$ jest ortogonalna, to $Q Q^T = I$.*

Wniosek. Kolumny $k_1, k_2, \dots, k_n$ i wiersze $w_1, w_2, \dots, w_n$ macierzy ortogonalnej $Q_{n \times n}$ tworzą układy ortogonalne, tzn. $\langle k_i, k_j \rangle = \delta_{ij}$, $\langle w_i, w_j \rangle = \delta_{ij}$, gdzie δ_{ij} oznacza symbol Kroneckera.

Tw. 7. *Macierz symetryczna A jest ortogonalnie podobna do macierzy diagonalnej, tzn. istnieje macierz ortogonalna Q taka, że $Q^T A Q = D$, gdzie D jest macierzą diagonalną.*

Tw. 8. *Wartości własne macierzy symetrycznej A i jej wektory własne są rzeczywiste.*

Tw. 9. *Jeżeli macierz A ma różne wartości własne, to jest podobna do macierzy diagonalnej. Macierz podobieństwa P i macierz diagonalna $D = P^{-1}AP$ mogą być w ogólności zespolone.*

Tw. 10. *Każda macierz jest unitarnie podobna do macierzy trójkątnej (która może być macierzą zespoloną).*

Def. 6. Macierz postaci

$$J = \begin{bmatrix} \lambda & 1 & & 0 \\ & \cdot & \cdot & \\ & & \cdot & \cdot \\ & & & \cdot & 1 \\ 0 & & & & \lambda \end{bmatrix}_{k \times k}$$

nazywamy klatką Jordana rzędu k . Liczba λ może być zespolona.

Tw. 11. *Dla każdej macierzy A istnieją takie macierze P (być może zespolone), że*

$$P^{-1}AP = \begin{bmatrix} J_1 & & 0 \\ & J_2 & \\ & & \cdot & \\ 0 & & & \cdot & J_r \end{bmatrix} \quad (6.1)$$

gdzie macierze $J_1, J_2, \dots, J_r$ są klatkami Jordana.

Macierz (6.1) nazywamy *postacią kanoniczną Jordana* macierzy A .

Tw. 12 (Schura). *Suma kwadratów modułów wartości własnych jest ograniczona od góry przez kwadrat normy euklidesowej, tzn.*

$$\sum_{i=1}^n |\lambda_i|^2 \leq \|A\|_E^2$$

Tw. 13. *Dodanie do macierzy A macierzy jednostkowej pomnożonej przez skalar przesuwa widmo macierzy, gdyż dla dowolnej liczby c mamy*

$$\text{Sp}(A + cI) = \text{Sp}(A) + c$$

gdzie $\text{Sp}(A) + c$ oznacza zbiór liczb $\lambda_1 + c, \lambda_2 + c, \dots, \lambda_n + c$.

Tw. 14. (Gershgorina):

a) *Jeśli C_i oznaczają koła domknięte na płaszczyźnie zespolonej o środkach w punktach a_{ii} i promieniach równych sumie modułów elementów z danego wiersza spoza diagonal, tzn.*

$$C_i = \left\{ z : |z - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \right\}$$

to

$$\text{Sp}(A) \subset \bigcup_{i=1}^n C_i$$

b) *Jeżeli k kół C_i tworzy zbiór rozłączny z pozostałymi kołami, to w zbiorze tym leży dokładnie k wartości własnych macierzy A .*

Z twierdzenia 14a,b wynikają następujące wnioski:

Wniosek 1. *Jeżeli macierz A jest symetryczna i diagonalnie dominująca o nieujemnych elementach na diagonal, to jest nieujemnie określona. jeśli dodatkowo jest to macierz nieosobliwa, to jest ona dodatnio określona.*

Macierz symetryczna silnie diagonalnie dominująca jest dodatnio określona wtedy i tylko wtedy, gdy elementy na diagonal są dodatnie.

Wniosek 2. *W każdym kole Gershgorina rozłącznym z pozostałymi leży dokładnie jedna wartość własna.*

Tw. 15. *Jeżeli elementy macierzy A są niemniejsze niż moduły elementów macierzy B , to $\varrho(A) \geq \varrho(B)$.*

Dowód tego twierdzenia można znaleźć np. w [70, s. 47]. Symbolem $\mathbf{x}\mathbf{y}^T$ oznaczamy macierz powstałą przez pomnożenie wektora kolumnowego $\mathbf{x}$ przez wektor wierszowy $\mathbf{y}^T$ zgodnie z regułami mnożenia macierzy; jest to zatem macierz $\mathbf{x}\mathbf{y}^T$.

Tw. 16. *Niech wektor $\mathbf{x}$ o długości 1 (wg normy $\|\cdot\|_2$) będzie wektorem własnym odpowiadającym wartości własnej λ macierzy symetrycznej A . Jeżeli macierz A ma pozostałe wartości własne różne od λ , to macierz $B = A - c\mathbf{x}\mathbf{x}^T$ ma te same wektory i wartości własne co macierz A z wyjątkiem wartości własnej odpowiadającej wektorowi $\mathbf{x}$, równej $\lambda - c$.*

Ćwiczenia

1. Udowodnić twierdzenie 5. Jaki jest związek między wektorami własnymi macierzy A i B ?

2. Udowodnić twierdzenie 13.

3. Udowodnić, że jeśli macierz A jest silnie diagonalnie dominująca o dodatnich elementach na diagonalu i silnie diagonalnie dominująca kolumnowo, to jest dodatnio określona (tzn. $\forall x \neq 0, \langle x, Ax \rangle > 0$). Wskazówka. Zbadać własności macierzy $B = \frac{1}{2}(A + A^T)$.

4. Podać przykład macierzy symetrycznej diagonalnie dominującej i takiej, że przynajmniej w jednym wierszu element na diagonalu jest większy od sumy modułów pozostałych elementów w tym wierszu, a która nie jest dodatnio określona.

5. Czy macierz symetryczna dodatnio określona jest zawsze diagonalnie dominująca?

6. Udowodnić twierdzenie 16.

7. Udowodnić, że kwadratowa macierz ortogonalna Q ma normę spektralną równą 1, tzn.

$$\|Q\|_2 = 1. \text{ Wskazówka. } \|Q\|_2 = \max_{x \neq 0} \frac{\langle Qx, Qx \rangle}{\|x\|_2^2}.$$

8. Udowodnić, że gdy macierz Q jest ortogonalna i kwadratowa, wówczas $\|QA\|_2 = \|A\|_2, \|AQ\|_2 = \|A\|_2$. Wskazówka. $(QA)^T(QA) = A^T$. A więc $\|QA\|_2 = \|A\|_2, \|AQ\|_2 = \|A\|_2$.
 $= \max_{x \neq 0} \frac{\|AQx\|_2}{\|x\|_2}$, stąd $\|AQ\|_2 = \|A\|_2$

6.3. ZABURZENIA WARTOŚCI I WEKTORÓW WŁASNYCH

Wartości własne. Obliczona jakąś metodą wartość własna λ macierzy A jest na ogół obarczona pewnym błędem (wejściowym, zaokrągleń lub błędem metody, np. błędem obcięcia w metodach iteracyjnych). Błąd wejściowy i błąd zaokrągleń wygodnie jest interpretować jako wynik zastępczego zaburzenia macierzy A . Dowodzimy wówczas, że obliczona wartość własna $\bar{\lambda}$ jest dokładną wartością własną macierzy $\bar{A}$, różnej (na ogół) od A . Niech $\bar{\lambda} = \lambda + \delta\lambda$, $\bar{A} = A + \delta A$. Aby dowieść stabilności numerycznej wybranej metody znajdowania wartości własnych, wystarczy wykazać, że zastępcze zaburzenie δA dąży do zera przy zwiększeniu dokładności obliczeń (tzn. zmniejszeniu liczby ε ; patrz lemat w p. 5.3.1) oraz wykazać, że gdy δA dąży do zera, wówczas błąd $\delta\lambda$ również dąży do zera. W celu przedyskutowania wpływu zaburzenia δA macierzy A na położenie jej wartości własnych $\lambda \in \text{Sp}(A)$, podamy najpierw dwie definicje.

Def. 1. Odległością liczby z od zbioru $\text{Sp}(A) = \{\lambda_1, \lambda_2, \dots, \lambda_n\}$ nazywamy liczbę $\min_{i=1, \dots, n} |z - \lambda_i|$; będziemy ją oznaczać $\text{dist}(z, \text{Sp}(A))$.

Def. 2. Odległością zbiorów spektralnych $\text{Sp}(A)$ i $\text{Sp}(B)$ dwu macierzy $A_{n \times n}$ i $B_{n \times n}$ nazywamy liczbę $\min_p \max_{i=1, \dots, n} |\lambda_i - \gamma_i^{(p)}|$, gdzie λ_i są kolejnymi elementami uporządkowanego zbioru $\text{Sp}(A)$, a $\gamma_i^{(p)}$ — kolejnymi wartościami własnymi ze zbioru $\text{Sp}(B)$, które ustawiono w innym porządku, oznaczanym indeksem p ; odległość tę będziemy oznaczać $\|\text{Sp}(A) - \text{Sp}(B)\|_\infty$.

U w a g a. Powyższa definicja odległości zbiorów spektralnych może się wydawać niepotrzebnie skomplikowana na pierwszy rzut oka. Chodzi jednak o to, aby bliskość dwu zbiorów $\text{Sp}(A)$ i $\text{Sp}(B)$ oznaczała, że nie tylko np. w małym otoczeniu liczby λ_i istnieje jakiś element zbioru $\text{Sp}(B)$, lecz by wielokrotnej wartości własnej ze zbioru $\text{Sp}(A)$ odpowiadała odpowiednia ilość bliskich wartości własnych ze zbioru $\text{Sp}(B)$ i odwrotnie.

Przykład 1. Jeżeli $\text{Sp}(A) = \{1, 1, 2\}$, zaś $\text{Sp}(B) = \{2, 2, 1\}$, to $\|\text{Sp}(A) - \text{Sp}(B)\|_\infty = \max_{i=1,2,3} |\lambda_i - \gamma_i|$, $\lambda_i \in \{1, 1, 2\}$, $\gamma_i \in \{1, 2, 2\}$.

Zatem $\|\text{Sp}(A) - \text{Sp}(B)\|_\infty = 1$, mimo że $\text{dist}(1, \text{Sp}(B)) = 0$ i $\text{dist}(2, \text{Sp}(B)) = 0$.

Tw. 1 Jeżeli zaburzenie δA elementów macierzy A dąży do zera, to $\text{Sp}(A + \delta A)$ dąży do $\text{Sp}(A)$ w tym sensie, że $\|\text{Sp}(A + \delta A) - \text{Sp}(A)\|_\infty \rightarrow 0$.

Analogiczna teza jest treścią twierdzenia Ostrowskiego (patrz np. [65]). W dowodzie korzysta się tam z ciągłej zależności pierwiastków wielomianu $\lambda^n + a_{n-1}\lambda^{n-1} + \dots + a_0$ od wartości współczynników, która to własność jest dowodzona w ogólnej teorii funkcji analitycznych i jest wnioskiem z twierdzenia Rouchego (patrz np. [62]) oraz z faktu, że wielomiany charakterystyczne macierzy $A + \delta A$ i A mają dowolnie bliskie współczynniki, gdy $\delta A \rightarrow 0$.

Kolejne twierdzenie, które poniżej przytaczamy, precyzuje jak szybko wartości własne macierzy $A + \delta A$ dążą do wartości własnych macierzy A , jeśli jest ona podobna do macierzy diagonalnej.

Tw. 2. (Bauera-Fikego). Jeżeli macierz A jest podobna do macierzy diagonalnej D , tzn. istnieje nieosobliwa macierz podobieństwa P taka, że $P^{-1}AP = D$, to dla dowolnej wartości własnej γ macierzy $A + \delta A$ obowiązuje oszacowanie

$$\text{dist}(\gamma, \text{Sp}(A)) \leq \alpha \quad (6.2)$$

gdzie $\alpha = \|P^{-1}\|_2 \|P\|_2 \|\delta A\|_2$.

Wniosek 1. Liczbę $K_2(P) = \|P^{-1}\|_2 \|P\|_2$ można interpretować jako wskaźnik uwarunkowania zadania obliczenia wartości własnych macierzy A podobnej do diagonalnej.

Wniosek 2. Jeżeli macierz A jest symetryczna, to macierz P jest ortogonalna i liczba $K_2(P) = 1$. Zadanie wyznaczenia wartości własnych macierzy symetrycznej jest zatem zawsze (bez względu na układ wartości własnych) zadaniem dobrze uwarunkowanym.

Wniosek 3. Przekształcenie macierzy A do postaci $B = BAR^{-1}$ przy użyciu dowolnej nieosobliwej macierzy nie zmienia wprowadzie wartości własnych macierzy, ale może pogorszyć lub polepszyć uwarunkowanie wartości własnych. Dla macierzy A bowiem $K_2(P) = \|P^{-1}\|_2 \|P\|_2$, natomiast dla macierzy B wskaźnik $K_2 = \|(RP)^{-1}\|_2 \cdot \|RP\|_2$.

Wniosek 4. Jeżeli Q jest macierzą ortogonalną, to przekształcenie $B = QAQ^T$ nie zmienia uwarunkowania wartości własnych macierzy. Wtedy bowiem $\|QP\|_2 = \|P\|_2$ oraz $\|QP^{-1}\|_2 = \|P^{-1}Q^{-1}\|_2 = \|P^{-1}\|_2$ (patrz ćwic. 8).

Twierdzenie Bauera i Fikego uściślił Wilkinson (patrz [65]) badając

wrażliwość pojedynczych wartości własnych. Nierówność (6.2) charakteryzuje bowiem globalną drażliwość wartości własnych, a przecież niektóre z nich mogą być znacznie mniej drażliwe na zaburzenia elementów macierzy A .

Tw. 3 (Wilkinsona). *Niech macierz A będzie podobna do macierzy diagonalnej D z elementami na diagonalu $\lambda_1, \lambda_2, \dots, \lambda_n$. Niech λ_i będzie pojedynczą wartością własną, a $\mathbf{x}_j$ oraz $\mathbf{y}_j, j = 1, \dots, n$, niech będą odpowiednio wektorami własnymi i lewostronnymi wektorami własnymi. Wówczas wartość własna γ_i macierzy $A + \delta A$ dążąca do λ_i spełnia zależność*

$$\gamma_i = \lambda_i + \delta^1 \lambda_i + \delta^2 \lambda_i + o(\|\delta A\|^2) \quad (6.3)$$

gdzie

$$\delta^1 \lambda_i = \frac{\langle \mathbf{y}_i, \delta A \mathbf{x}_i \rangle}{\langle \mathbf{y}_i, \mathbf{x}_i \rangle}$$

$$\delta^2 \lambda_i = \frac{1}{\langle \mathbf{y}_i, \mathbf{x}_i \rangle} \sum_{\substack{j=1 \\ j \neq i}}^n \frac{\langle \mathbf{y}_j, \delta A \mathbf{x}_i \rangle \langle \mathbf{y}_i, \delta A \mathbf{x}_j \rangle}{\langle \mathbf{y}_j, \mathbf{x}_j \rangle (\lambda_i - \lambda_j)}$$

Wniosek. *Poprawka $\delta^1 \lambda_i$ jest członem rzędu pierwszego rozwinięcia wartości γ_i względem δA , a więc zamiast wskaźnika $K_2(\mathbf{P}) = \|\mathbf{P}^{-1}\|_2 \|\mathbf{P}\|_2$ można przyjąć dla wartości λ_i wskaźnik*

$$k_i = \frac{\|\mathbf{y}_i\|_2 \|\mathbf{x}_i\|_2}{\langle \mathbf{y}_i, \mathbf{x}_i \rangle} = \frac{1}{|\cos \theta_i|} \quad (6.4)$$

gdzie θ_i jest kątem między wektorami $\mathbf{x}_i, \mathbf{y}_i$.

Oczywiście, gdy A jest macierzą symetryczną, wówczas wektor $\mathbf{x}_i$ jest równoległy do wektora $\mathbf{y}_i$, więc $k_i = 1$. Można pokazać, że dla dowolnej macierzy podobnej do diagonalnej mamy $k_i \leq K_2$ (patrz ćwic. 9).

Oszacowanie (6.2) Bauera i Fikego można uogólnić na przypadek dowolnej macierzy A o postaci kanonicznej Jordana J . Niech k_i oznacza maksymalny rząd klatek Jordana, w których występuje wartość własna λ_i (np. gdy macierz A jest podobna do macierzy diagonalnej, to $k_i = 1, i = 1, 2, \dots, n$), a k niech oznacza maksymalny rząd klatek Jordana w postaci kanonicznej macierzy J . Oczywiście $k = \max \{k_1, k_2, \dots, k_n\}$. Przypomnijmy, że wartości własne $\gamma_1, \gamma_2, \dots, \gamma_n$ macierzy $A + \delta A$ zostały ponumerowane w ten sposób, by $\gamma_i \rightarrow \lambda_i$, gdzie λ_i jest i -tą wartością własną macierzy A .

Tw. 4. *Jeżeli $\mathbf{P}^{-1} A \mathbf{P} = J$, $\gamma \in \text{Sp}(A + \delta A)$, to dla dowolnego zaburzenia δA jest spełnione oszacowanie*

$$\text{dist}(\gamma, \text{Sp}(A)) \leq k \max(\alpha, \sqrt[k]{\alpha}) \quad (6.5a)$$

gdzie $\alpha = K_2(\mathbf{P}) \|\delta A\|_2$. Ponadto dla dostatecznie małych zaburzeń δA obowiązuje oszacowanie

$$\text{dist}(\gamma, \text{Sp}(A)) \leq (1 + (k - 1) \sqrt[k]{\alpha}) \sqrt[k]{\alpha} \quad (6.5b)$$

oraz

$$|\gamma_i - \lambda_i| \leq (1 + (k_i - 1) \sqrt[k_i]{\alpha}) \sqrt[k_i]{\alpha} \quad (6.5c)$$

U w a g a. Gdy $k = 1$, tzn. macierz A jest podobna do macierzy diagonalnej, wówczas uzyskujemy identyczne oszacowania jak (6.2). Ponadto z nierówności (6.5c) widać, że przy małych zaburzeniach δA oszacowanie wielkości zmian wartości własnej λ_i nie zależy od rzędu klatek Jordana pozostałych wartości własnych.

DOWÓD. Udowodnimy nierówność (6.5c). Mamy $P^{-1}AP = J$. Niech x będzie wektorem własnym macierzy $A + \delta A$ odpowiadającym wartości własnej γ i niech $y = P^{-1}x$. Zatem $(A + \delta A)x = \gamma x$, skąd $(J + P^{-1}(\delta A)P)y = \gamma y$. Oznaczmy $E = P^{-1}(\delta A)P$. Mamy więc $(J - \gamma I)y = -Ey$. Macierz

$$J - \gamma I = \begin{bmatrix} J_1 - \gamma I_1 & & & 0 \\ & \ddots & & \\ & & J_2 - \gamma I_2 & \\ 0 & & & \ddots \\ & & & & J_r - \gamma I_r \end{bmatrix}$$

gdzie macierze $J_m - \gamma I_m$, $m = 1, 2, \dots, r$, mają postać

$$\begin{bmatrix} \lambda_m - \gamma & 1 & & & 0 \\ & \lambda_m - \gamma & 1 & & \\ & & \ddots & \ddots & \\ 0 & & & \ddots & 1 \\ & & & & \lambda_m - \gamma \end{bmatrix}$$

Założmy na chwilę, że liczba γ jest różna od wszystkich wartości własnych $\lambda_1, \lambda_2, \dots, \lambda_n$. Wtedy macierze $J_m - \gamma I_m$ są nieosobliwe, a ich odwrotności są równe macierzom

$$(J_m - \gamma I_m)^{-1} = \begin{bmatrix} \frac{1}{(\lambda_m - \gamma)} & \frac{1}{(\lambda_m - \gamma)^2} & \cdots & \frac{(-1)^{k_m-1}}{(\lambda_m - \gamma)^{k_m}} \\ & \frac{1}{(\lambda_m - \gamma)} & \cdots & \frac{(-1)^{k_m-2}}{(\lambda_m - \gamma)^{k_m-1}} \\ & & \ddots & 1 \\ 0 & & & \frac{1}{(\lambda_m - \gamma)} \end{bmatrix}$$

Ponieważ

$$(J - \gamma I)^{-1} = \begin{bmatrix} (J_1 - \gamma I_1)^{-1} & & & 0 \\ & \ddots & & \\ & & (J_2 - \gamma I_2)^{-1} & \\ 0 & & & \ddots \\ & & & & (J_r - \gamma I_r)^{-1} \end{bmatrix}$$

oraz $y = -(J - \gamma I)^{-1} E y$, więc $1 \leq \|(J - \gamma I)^{-1}\|_2 \|E\|_2$. Dla dowolnej macierzy M mamy $\|M\|_2 \leq \sqrt{\|M\|_1 \|M\|_\infty}$, więc

$$\|(J - \gamma I)^{-1}\|_2 \leq \frac{1}{|\lambda - \gamma|} + \frac{1}{|\lambda - \gamma|^2} + \dots + \frac{1}{|\lambda - \gamma|^k}$$

dla każdego $\lambda \in \text{Sp}(A)$. Stąd, oznaczając $\varrho = \text{dist}(\gamma, \text{Sp}(A))$, otrzymamy

$$\varrho^k \leq (1 + \varrho + \varrho^2 + \dots + \varrho^{k-1}) \|E\|_2 \quad (6.6)$$

Zauważmy, że nierówność (6.6) jest spełniona także dla $\gamma \in \text{Sp}(A)$. Jeśli zatem $\varrho \leq 1$, to $\varrho^k \leq k \|E\|_2$, skąd $\varrho \leq \sqrt[k]{k} \cdot \sqrt[k]{\|E\|_2} \leq k \sqrt[k]{\alpha}$. Jeśli zaś $\varrho > 1$, to $1 + \varrho + \dots + \varrho^{k-1} \leq k \varrho^{k-1}$, skąd $\varrho < k \alpha$. Zatem

$$\varrho \leq k \max(\alpha, \sqrt[k]{\alpha}) \quad (6.7)$$

Ponadto dla dostatecznie małych zaburzeń mamy $\varrho \leq 1$, skąd

$$1 + \varrho + \varrho^2 + \dots + \varrho^{k-1} \leq 1 + (k-1)\varrho \leq \left(1 + \frac{(k-1)\varrho}{k}\right)^k.$$

Zatem z nierówności (6.6) i (6.7) otrzymujemy

$$\varrho \leq (1 + (k-1)\sqrt[k]{\alpha}) \sqrt[k]{\alpha}$$

Zauważmy ponadto, że gdy zaburzenie δA jest dostatecznie małe, wówczas normy $\|\cdot\|_1$ i $\|\cdot\|_\infty$ macierzy $(J - \gamma I)^{-1}$ są równe $\frac{1}{|\lambda_i - \gamma_j|} + \frac{1}{|\lambda_i - \gamma_j|^2} + \dots + \frac{1}{|\lambda_i - \gamma_j|^{k_i}}$, gdyż suma ta przekracza pozostałe sumy modułów elementów. Stąd uzyskujemy nierówność (6.5c) zamieniając k na k_i oraz ϱ na $|\lambda_i - \gamma_i|$, cnd.

U w a g a. Występowanie z prawej strony nierówności (6.5) pierwiastków $\sqrt[k]{\alpha}$ w znacznym stopniu pogarsza oszacowanie zaburzeń wartości własnych. Jeśli np. chcemy, aby wartości własne macierzy $A_{20 \times 20}$ nie zostały zaburzone więcej niż o 0,1, musimy przyjąć $k = 20$ i chcąc o tym wnioskować z nierówności (6.5b) musimy zapewnić spełnienie nierówności $\alpha < 10^{-20}$ (patrz ćwicz. 2). Jeśli jednak wiemy, że macierz A jest symetryczna, to można dopuścić zgodnie z (6.2), że $\|\delta A\|_2 = \alpha \leq 0,1$. Oszacowania (6.5) nie można jednak istotnie polepszyć bez założenia szczególnych własności macierzy A .

Przykład 2. Niech $A = \begin{bmatrix} \lambda & & & 0 \\ & \ddots & & \\ & & \ddots & \\ & & & 1 \\ & 0 & & \lambda \end{bmatrix}_{n \times n}$ oraz $\delta A = [\delta a_{ij}]$, gdzie $\delta a_{1n} = \alpha > 0$,

$\delta a_{ij} = 0$ dla pozostałych i, j . Oczywiście $\|\delta A\|_2 = \alpha$, zaś wartości własne γ macierzy $A + \delta A$ są pierwiastkami równania charakterystycznego $(\gamma - \lambda)^n - \alpha = 0$. Stąd $\gamma_i = \lambda + r_i \sqrt[n]{\alpha}$, $i = 1, 2, \dots, n$, gdzie $r_i = e^{j \frac{2\pi i}{n}}$ są zespolonymi pierwiastkami z jedności. Mamy więc $|\gamma_i - \lambda| = \sqrt[n]{\alpha}$.

Wektory własne. Niech macierz A ma wartości własne $\lambda_1, \lambda_2, \dots, \lambda_n$, a macierz $A + \delta A$ — wartości własne $\gamma_1, \gamma_2, \dots, \gamma_n$. Załóżmy, że wartość γ_i odpowiada wartości λ_i w tym sensie, że gdy $\delta A \rightarrow 0$, wówczas $\gamma_i \rightarrow \lambda_i$. Niech wektory x_i oraz y_i oznaczają wektory własne macierzy A i macierzy $A + \delta A$,

jednak $u_{22} = \varepsilon$, ponieważ $\frac{\varepsilon}{2}$ to zero w maszynie cyfrowej. Otrzymamy zatem $\det U = 2\varepsilon$, czyli obliczymy wartość $\det A$ ze 100% błędem. Tymczasem $\|E\|_\infty = \|LU - A\|_\infty = \frac{1}{2}\varepsilon$, więc błąd względny $\|E\|_\infty / \|A\|_\infty \approx \frac{1}{4}\varepsilon \ll 1$.

Powyższy przykład dowodzi znanej własności wyznaczników, iż małe zaburzenia elementów macierzy mogą spowodować duże zmiany wartości wyznacznika. Okazuje się, że zależą one od wielkości $K \frac{\|E\|}{\|A\|}$, gdzie K jest wskaźnikiem uwarunkowania macierzy A . Dokładniejsze oszacowania wielkości zmian wyznacznika podano niżej dla macierzy symetrycznej.

Omówimy obecnie krótko wyznaczanie macierzy odwrotnej A^{-1} metodą GCW. W tym celu należy wyznaczyć rozkład LU macierzy A (uwzględniając ewentualne przestawienia wierszy macierzy A), a następnie rozwiązać n razy układ równań $LUx^{(i)} = e^{(i)}$, $i = 1, 2, \dots, n$, gdzie $e^{(i)} = [0, \dots, 1, \dots, 0]^T$ jest i -tym wersorem w przestrzeni R^n . Rozwiązania $x^{(i)}$ są kolumnami macierzy A^{-1} , przy czym należy ustawić je w odpowiedniej kolejności, wynikającej z przestawień wierszy macierzy. Można też równocześnie rozwiązywać n układów równań z prawymi stronami równymi $e^{(i)}$, $i = 1, 2, \dots, n$.

Przykład 2. Obliczyć macierz

$$\begin{bmatrix} 1 & 0 & 1 \\ 3 & 3 & 0 \\ 0 & 2 & 2 \end{bmatrix}^{-1}$$

Etap 1. Wyznaczamy rozkład LU macierzy A . Należy w tym celu określić przestawienie wierszy, co uczynimy za pomocą dodatkowej kolumny numerów wierszy. Elementy macierzy L zapiszemy na miejscu odpowiednich elementów macierzy A . Mamy

$$\begin{bmatrix} 1 & 0 & 1 \\ 3 & 3 & 0 \\ 0 & 2 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \Rightarrow \text{po przestawieniu} \begin{bmatrix} 3 & 3 & 0 \\ 1 & 0 & 1 \\ 0 & 2 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \Rightarrow$$

po eliminacji pierwszej zmiennej z równań drugiego i trzeciego za pomocą mnożników $l_{21} = \frac{1}{3}$, $l_{31} = 0$

$$\begin{bmatrix} 3 & 3 & 0 \\ \frac{1}{3} & -1 & 1 \\ 0 & 2 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \Rightarrow \text{po przestawieniu} \begin{bmatrix} 3 & 3 & 0 \\ 0 & 2 & 2 \\ \frac{1}{3} & -1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} \Rightarrow$$

po eliminacji drugiej zmiennej z trzeciego równania za pomocą $l_{32} = -\frac{1}{2}$:

$$\begin{bmatrix} 3 & 3 & 0 \\ 0 & 2 & 2 \\ \frac{1}{3} & -\frac{1}{2} & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

Jeśli więc $\|\delta A\|_2 < d$, to

$$\varrho(y_i, x_i) = \|y_i^\perp\|_2 \leq \|\delta A\|_2 / (d - \|\delta A\|_2) \quad \text{cnd.}$$

Przykład 3. Dane są

$$A = \begin{bmatrix} 1+d & 0 \\ 0 & 1 \end{bmatrix}, \quad d > 0, \quad \delta A = \begin{bmatrix} 0 & \alpha \\ \alpha & 0 \end{bmatrix}$$

Mamy zatem $\text{Sp}(A) = \{1+d, 1\}$, przy czym dla wartości własnej $\lambda_2 = 1$ wektor własny $x_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Dla macierzy

$$A + \delta A = \begin{bmatrix} 1+d & \alpha \\ \alpha & 1 \end{bmatrix}$$

otrzymujemy wartości własne γ_1 i γ_2 , gdzie

$$\gamma_2 = 1 - \frac{2\alpha^2}{d + \sqrt{d^2 + 4\alpha^2}}$$

Wektor własny y_2 odpowiadający wartości własnej γ_2 spełnia równanie

$$\begin{bmatrix} 1+d-\gamma_2 & \alpha \\ \alpha & 1-\gamma_2 \end{bmatrix} y_2 = 0$$

a więc jeśli $y_2 = \begin{bmatrix} a \\ b \end{bmatrix}$, to $b = \frac{\alpha}{\gamma_2 - 1} a$. Stąd

$$\varrho(y_2, x_2) = \frac{|a|}{a^2 + b^2} = \frac{1}{\sqrt{1 + \left(\frac{\alpha}{\gamma_2 - 1}\right)^2}}$$

Ponieważ $\|\delta A\|_2 = |\alpha|$, więc

$$\varrho(y_2, x_2) = \frac{1}{d} k(\alpha) \|\delta A\|_2, \quad \text{gdzie} \quad k(\alpha) \xrightarrow{\alpha \rightarrow 0} 1$$

U w a g a. Powyższe twierdzenie i przykład dowodzą, że jako wskaźnik uwarunkowania zadania obliczania wektorów własnych macierzy symetrycznej należy przyjąć $K_w = \frac{1}{d}$.

Obecnie podamy oszacowania błędów wyznaczonych wartości własnych λ i wektorów własnych x , które można uzyskać przez sprawdzenie obliczając wektor reszt $r = Ax - \lambda x$. Niech $\bar{\lambda}$ oznacza obliczoną przez nas wartość własną macierzy A , będącą przybliżeniem poszukiwanej wartości własnej λ . Podobnie niech $\bar{x}$ oznacza obliczony wektor własny, a x niech będzie dokładnym wektorem własnym. Załóżmy, że wektor $\bar{x}$ ma długość jednostkową, czyli $\|\bar{x}\|_2 = 1$, oraz że obliczyliśmy dokładnie wektor $r = A\bar{x} - \bar{\lambda}\bar{x}$.

Tw. 6. (Wilkinsona). *Jeżeli macierz A jest symetryczna oraz $\|r\|_2 = \varrho$, to w przedziale $\langle \bar{\lambda} - \varrho; \bar{\lambda} + \varrho \rangle$ znajduje się co najmniej jedna wartość własna macierzy A . Jeżeli wartości własne macierzy A znajdują się poza tym*

przedziałem są odległe od liczby $\bar{\lambda}$ co najmniej o $d > \varrho$, to $\bar{\varrho}(\bar{x}, x) \leq \frac{\varrho}{d}$, gdzie $\bar{\varrho}(\bar{x}, x)$ jest błędem wektora własnego $\bar{x}$, obliczonym zgodnie z definicją, tzn. $\bar{\varrho}(\bar{x}, x) = \min_{x \in X} \|\bar{x} - x\|_2$; X jest podprzestrzenią rozpiętą na wektorach własnych macierzy A , odpowiadających wartościom własnym z przedziału $\langle \bar{\lambda} - \varrho; \bar{\lambda} + \varrho \rangle$.

U w a g a. Równość $r = A\bar{x} - \bar{\lambda}\bar{x}$ można zapisać także jako $(A - r)\langle \bar{x} \rangle \bar{x} = \bar{\lambda}\bar{x}$. Oznacza to, że liczba $\bar{\lambda}$ i wektor $\bar{x}$ są odpowiednio dokładną wartością własną i wektorem własnym macierzy $A + \delta A$, gdzie $\delta A = r \rangle \langle \bar{x}$. Ponieważ $\|r \rangle \langle \bar{x}\|_2 = \|r\|_2 \|\bar{x}\|_2 = \varrho$, więc z twierdzeń 2 i 5 uzyskamy oszacowania

$$\text{dist}(\bar{\lambda}, \text{Sp}(A)) \leq \varrho \quad \text{oraz} \quad \varrho(\bar{x}, x) \leq \frac{1}{d_1 - \varrho} \varrho$$

gdzie d_1 oznacza odległość wartości własnej λ macierzy A najbliższej liczbie $\bar{\lambda}$ od pozostałych liczb widma $\text{Sp}(A)$.

Na zakończenie tego punktu podamy oszacowanie wektora reszt $Ax - \lambda x$ obliczonego przy użyciu arytmetyki zmiennopozycyjnej dla dokładnych wartości λ i x macierzy A zaokrąglonych do wartości $\bar{\lambda}$ i $\bar{x}$. Przyjmujemy, że $\|\bar{x}\|_2 = 1$.

Mamy $Ax - \lambda x = 0$, a więc wektor reszt $A\bar{x} - \bar{\lambda}\bar{x}$ jest równy $\text{fl}(A\bar{x} - \bar{\lambda}\bar{x}) - (A\bar{x} - \bar{\lambda}\bar{x}) = r$. Składowa i -ta wektora r jest równa

$$\begin{aligned} \text{fl}(\langle a^{(i)}, \bar{x} \rangle - \bar{\lambda} \bar{x}_i) - (\langle a^{(i)}, x \rangle - \lambda x_i) = \\ = \{[a_1^{(i)} \bar{x}_1(1 + \varepsilon_1)(1 + \varepsilon_2) + a_2^{(i)} \bar{x}_2(1 + \varepsilon_3)(1 + \varepsilon_4)](1 + \varepsilon_5) + \\ + \dots + a_n^{(i)} \bar{x}_n(1 + \varepsilon_{3n-3})(1 + \varepsilon_{3n-2})\}(1 + \varepsilon_{3n-1}) + \lambda \bar{x}_1(1 + \varepsilon_{3n})(1 + \varepsilon_{3n+1}), \\ \|\varepsilon_j\| \leq 2^{-t}, \quad j = 1, 2, \dots, 3n + 1 \end{aligned}$$

Uwzględniając, że każda składowa $\bar{x}_j$ jako zaokrąglenie x_j jest równa $x_j(1 + \varepsilon'_j)$ oraz $\bar{\lambda} = \lambda(1 + \varepsilon'')$, a ponadto przyjmując, że długość słowa binarnego t jest na tyle duża, że $(n + 2) \cdot 2^{-t} < 0,1$, otrzymamy oszacowanie

$$|r_i| \leq \bar{\varepsilon}[(n + 2)\|A\|_\infty + 3|\lambda|] \leq \bar{\varepsilon}(n + 5)\|A\|_\infty$$

gdzie $\bar{\varepsilon} = 1,06\varepsilon$. Ponieważ $\|r\|_2 \leq \sqrt{n}\|r\|_\infty$, więc

$$\|r\|_2 \leq \bar{\varepsilon}(n^{\frac{3}{2}} + 5n^{\frac{1}{2}})\|A\|_\infty \quad (6.8)$$

Przypuśćmy jednak, że sprawdzanie (obliczanie r) wykonujemy w podwójnej precyzji, lecz dysponujemy wstępnie zaokrąglonymi wartościami $\bar{\lambda}$ i $\bar{x}$.

Wtedy

$$\|r\|_2 \leq \bar{\varepsilon}(n^{\frac{1}{2}}\|A\|_\infty + n^{\frac{1}{2}}2|\lambda|) \leq 3\bar{\varepsilon}n^{\frac{1}{2}}\|A\|_\infty \quad (6.9)$$

Oczywiście, nie można oczekiwać, by jakakolwiek metoda wyznaczania

wartości własnych λ i wektorów własnych x dawała lepsze oszacowania uzyskane w analogiczny sposób.

Ćwiczenia

1. Dowieść, że wskaźniki k_i (patrz wniosek po tw. 4) dla macierzy A podobnej do diagonalnej spełniają nierówność $|k_i| \leq \|P^{-1}\|_2 \|P\|_2$, gdzie P jest macierzą podobieństwa sprowadzającą macierz A do postaci diagonalnej. Wskazówka. Wektory x_i oraz y_i są równoległe do Pe_i oraz $P^{-T}e_i$, gdzie e_i są kolumnami macierzy jednostkowej.

2. Wykazać, że gdy prawe strony nierówności (6.5a) i (6.5b) są mniejsze niż $\varrho \in (0; 1)$ to $\alpha < \varrho^k$.

6.4. MACIERZE O ELEMENTACH RZECZYWISTYCH — METODY OGÓLNE

Przystępując do wyznaczania wartości własnych lub wektorów własnych macierzy A , należy starannie wybrać metodę obliczeniową. Nie opracowano jeszcze szybkich i dokładnych metod wyznaczania tych wielkości w przypadku ogólnym i aby dokonać trafnego wyboru metody, należy uwzględnić specyficzne cechy macierzy (symetrię, istnienie dominujących wartości własnych, istnienie pełnego układu wektorów własnych itp.), a także warunki danego zadania (czy należy wyznaczyć wszystkie wartości lub wektory własne, czy wystarczają przybliżone wielkości itp.). W szczególności w przypadku niektórych metod trzeba zdecydować na wstępie, czy możemy poszukiwać wartości własnych i wektorów własnych jako wielkości rzeczywistych, czy też należy stosować metody przystosowane do wyznaczania wielkości zespolonych. Rozważania nasze ograniczymy do macierzy o elementach rzeczywistych. Będziemy omawiać metody wyznaczania dowolnych wartości własnych, nawet jeśli są one zespolone (patrz p. 6.4.4), natomiast wektory własne będziemy wyznaczać jedynie przy pewnych dodatkowych założeniach.

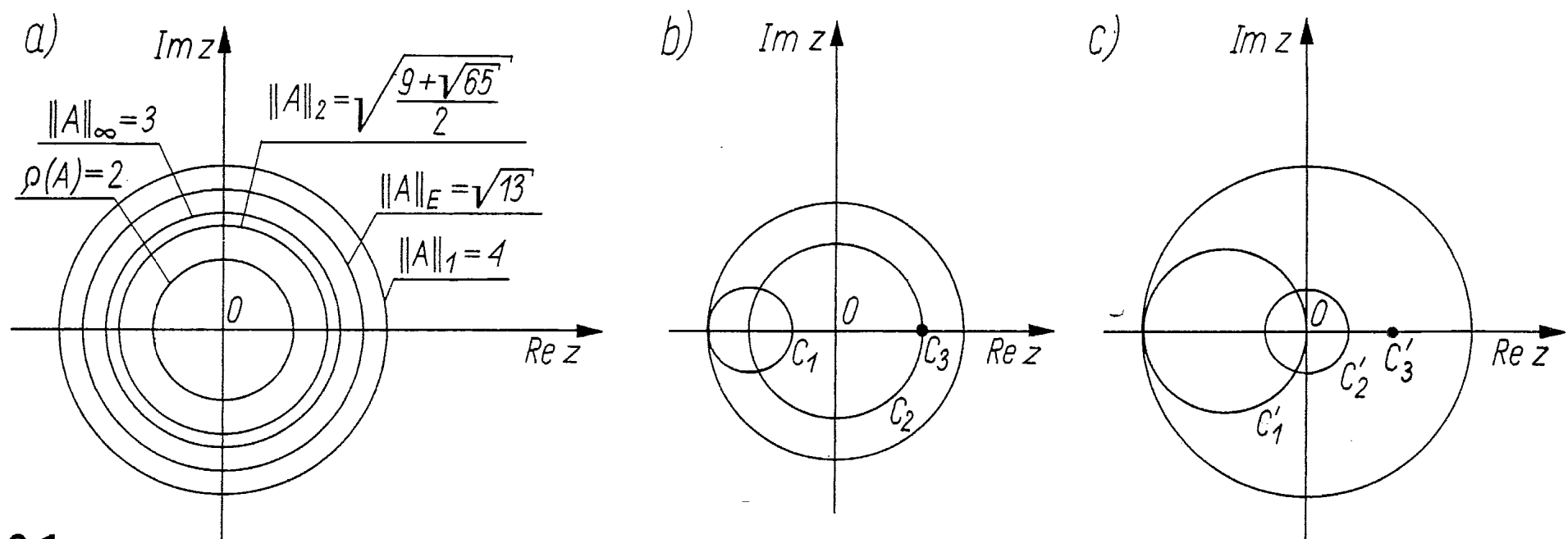
6.4.1. Lokalizacja wartości własnych

Wartości własne dowolnej macierzy A leżą na płaszczyźnie zespolonej w kole C_i o środku w 0 i promieniu równym promieniowi spektralnemu $\varrho(A)$. Ponieważ $\varrho(A) \leq \|A\|_p$, $p = 1, 2, \infty, E$, więc jeśli wystarczy nam bardziej zgrubne oszacowanie, to możemy przyjąć $\lambda_i \in \{z : |z| \leq \|A\|_p\}$. Dokładniejsze oszacowanie położenia wartości własnych niż $|\lambda_i| \leq \|A\|_\infty$ możemy otrzymać na podstawie twierdzenia 14a, b z p. 6.2, gdyż koła C_i są zawarte w kole $\{z : |z| \leq \|A\|_\infty\}$. Ponieważ $\text{Sp}(A) = \text{Sp}(A^T)$, więc zamiast oszacowania $|\lambda_i| \leq \|A\|_1$ można podać oszacowanie, w którym wykorzystuje się koła C'_i wyznaczone dla macierzy A^T . Są to koła o środkach w punktach a_{ii} i promieniach równych sumie modułów pozostałych elementów w i -tych kolumnach.

Przykład 1. Wskazać położenie wartości własnych macierzy

$$A = \begin{bmatrix} -2 & -1 & 0 \\ 2 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Na rysunku 6.1a przedstawiono koła o środkach w 0 i promieniach równych odpowiednio $\rho(A) < \|A\|_2 < \|A\|_\infty < \|A\|_E < \|A\|_1$, a na rys. 6.1b przedstawiono koła C_1, C_2, C_3 , o których mowa w twierdzeniu Gershgorina 14a. Koła te leżą wewnątrz koła o środku 0 i promieniu $\|A\|_\infty = 3$, więc nieco dokładniej lokalizują wartości własne. Widać na przykład, że żadna z tych wartości nie ma części urojonej większej co do modułu niż 2. Na rysunku 6.1c przedstawiono koła C'_1, C'_2, C'_3 odpowiadające macierzy A^T . Jedno z tych kół jest zdegenerowane (punkt $(2, 0)$) i jest odseparowane od dwu pozostałych kół. Z twierdzenia 14b wynika więc, że jedną z wartości własnych macierzy A jest $\lambda = 2$, dwie pozostałe leżą natomiast w obszarze $C'_1 \cup C'_2$.



6.1

Obliczane z wielomianu charakterystycznego wartości własne macierzy A są równe: $2, -1 - j, -1 + j$.

U w a g a. Dodatkową informację o położeniu wartości własnych można często uzyskać, jeśli korzystamy z twierdzeń Gershgorina w połączeniu z prostymi przekształceniami macierzy A przez podobieństwo $P^{-1}AP$. Dzieląc na przykład i -ty wiersz macierzy A przez liczbę $p_i \neq 0$, następnie zaś mnożąc i -tą kolumnę uzyskanej macierzy przez p_i , otrzymamy macierz A' podobną do A — o tych samych elementach na diagonalu co poprzednio, lecz o innych promieniach kół C_i .

Przykład 2. Macierz $A = \begin{bmatrix} -1 & 4 \\ 0,5 & 2 \end{bmatrix}$ ma koła C_1 i C_2 posiadające niepuste przecięcie.

Jeśli jednak podzielimy pierwszy wiersz przez $\frac{8}{3}$, a następnie pierwszą kolumnę pomnożymy przez

$\frac{8}{3}$, to otrzymamy macierz $A' = \begin{bmatrix} -1 & \frac{3}{2} \\ \frac{4}{3} & 2 \end{bmatrix}$, której koła C_1 i C_2 nie przecinają się. Wnioskujemy

stąd, że w każdym z nich znajduje się po jednej wartości własnej macierzy A , a więc są to liczby rzeczywiste.

Ćwiczenia

1. Uzasadnić, że jeśli macierz diagonalna $A_{5 \times 5}$ o wartościach własnych $-2, -1, 0, 1, 2$ jest zaburzona macierzą δA taką, że $\|\delta A\|_{1\infty} < 0,1$, to wartości własne macierzy $A + \delta A$ są nadal rzeczywiste.

2. Uzasadnić, że macierz silnie diagonalnie dominująca lub silnie diagonalnie dominująca kolumnowo nie może mieć wartości własnych o zerowych częściach rzeczywistych (w szczególności jest nieosobliwa).

6.4.2. Znajdowanie wartości własnych przy użyciu wielomianu charakterystycznego

Ponieważ wartości własne macierzy są pierwiastkami wielomianu charakterystycznego, pierwsze metody numeryczne służące do wyznaczania wartości własnych polegały w istocie na znajdowaniu zer tego wielomianu. Jedną z tego typu metod była *metoda Kryłowa* (patrz [41]), opracowana w 1931 r. Istotne znaczenie w tej metodzie miało obliczenie współczynników wielomianu charakterystycznego z zadaną dokładnością ϱ . Wydawało się bowiem, że po wyznaczeniu tych współczynników z małym błędem, także uzyskane następnie pierwiastki wielomianu charakterystycznego nie będą się wiele różnić od dokładnych wartości własnych. Okazało się jednak, że zagadnienie wyznaczenia pierwiastków wielomianu charakterystycznego jest na ogół znacznie gorzej uwarunkowane niż zagadnienie wyznaczenia wartości własnych macierzy (wartości własne mniej zależą od zmian elementów macierzy niż od zmian współczynników wielomianu charakterystycznego). Dlatego też nawet dla macierzy o dobrze uwarunkowanych wartościach własnych (tj. mało zmieniających się przy małych zaburzeniach elementów macierzy A) po wyznaczeniu zer wielomianu $\lambda^k + \alpha_{k-1}\lambda^{k-1} + \dots + \alpha_0$ nie możemy zakładać, iż przyjmując np. $\varrho < 10^{-6}$ obliczymy wartości własne macierzy A z błędem rzędu 10^{-6} , choć oczywiście przyjmując dostatecznie małe ϱ wyznaczymy je z dowolną dokładnością.

Podobnie stwierdzamy, że jeśli nawet zagadnienie wyznaczenia wartości własnych macierzy A jest dobrze uwarunkowane, to wyznaczenie zer równania $w(\lambda) = 0$ niekoniecznie jest dobrze uwarunkowanym zadaniem. W takich przypadkach metoda Kryłowa jest niekorzystna.

Przykład 1. Niech

$$A = \begin{bmatrix} 1,00 & 0,01 \\ 0,01 & 1,00 \end{bmatrix}$$

Jest to macierz symetryczna i zagadnienie wyznaczenia wartości własnych $\lambda_1 = 1,01$ i $\lambda_2 = 0,99$ jest zadaniem dobrze uwarunkowanym. Jeśli

$$\delta A = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 \\ \varepsilon_3 & \varepsilon_4 \end{bmatrix}$$

i $\lambda_1(\varepsilon)$ jest większą z wartości własnych macierzy $A + \delta A$, to

$$\lambda_1(\varepsilon) = 1,01 - \frac{1}{2} \cdot \varepsilon_1 + 1 \cdot \varepsilon_2 + 1 \cdot \varepsilon_3 - \frac{1}{2} \cdot \varepsilon_4 + O(\max\{|\varepsilon_1|, |\varepsilon_2|, |\varepsilon_3|, |\varepsilon_4|\})$$

Jeśli jednak wyznaczymy wielomian charakterystyczny $w(\lambda) = \lambda^2 - 2\lambda + 0,9999$ macierzy A , to większy z pierwiastków równania $\lambda^2 + (-2 + \delta_1) + (0,9999 + \delta_2) = 0$, spełnia warunek

$$\lambda_1(\delta) = 1,01 - 50,5\delta_1 - 50\delta_2 + O(\max\{|\delta_1|, |\delta_2|\})$$

Pierwiastek λ_1 jest więc 100 razy bardziej wrażliwy na zmiany współczynników wielomianu charakterystycznego niż na zmianę np. elementu a_{11} macierzy A .

6.4.3. Metoda potęgowa

Metoda potęgowa jest jedną z najprostszych metod wyznaczania pojedynczych wartości własnych i wektorów własnych macierzy A i często jest ona metodą bardzo skuteczną.

Algorytm tej metody jest następujący:

0. Dla $i = 0$, wybieramy dowolny wektor $\mathbf{x}_0$ taki, że $\|\mathbf{x}_0\|_\infty = 1$, oraz ustalamy ITM (maksymalną liczbę iteracji).

1. Obliczamy wektor $\mathbf{v}_{i+1} = A\mathbf{x}_i$ oraz $m_{i+1} = \|\mathbf{v}_{i+1}\|_\infty$.

Jeżeli $m_{i+1} = 0$, to Stop, w przeciwnym razie obliczamy

$$\mathbf{x}_{i+1} = \frac{\mathbf{v}_{i+1}}{m_{i+1}}$$

2. Jeśli $i + 1 \geq \text{ITM}$, to Stop, w przeciwnym razie $i = i + 1$, czyli powtarzamy etap 1.

Okazuje się, że gdy ciąg wektorów $\mathbf{x}_0, \mathbf{x}_2, \mathbf{x}_4, \dots, \mathbf{x}_{2j}, \dots$ jest zbieżny do wektora $\mathbf{x}$, wówczas także dla wartości $m_1, m_2, m_3, \dots$ jest zbieżny do pewnej liczby m . Ponadto, jeśli $\|\mathbf{x}_{j+1} - \mathbf{x}_j\| \rightarrow 0$, to $A\mathbf{x} = m\mathbf{x}$. W przeciwnym razie $\|\mathbf{x}_{j+1} - \mathbf{x}_j\|_\infty \rightarrow 2$ i $A\mathbf{x} = -m\mathbf{x}$.

W obu przypadkach znajdujemy wektor własny $\mathbf{x}$ taki, że $\|\mathbf{x}\|_\infty = 1$, oraz odpowiadającą mu wartość własną $\lambda = m$ lub $\lambda = -m$.

U w a g a. W literaturze spotyka się wiele wersji metody potęgowej: na przykład zamiast $m_{i+1} = \|\mathbf{v}_{i+1}\|_\infty$ bierzemy m_{i+1} równe współrzędnej wektora $\mathbf{v}_{i+1}$ o największej wartości bezwzględnej. Ponadto często bada się zbieżność ciągu $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots$, a nie podciągu o parzystych indeksach. Taka wersja może jednak nie być skuteczna nawet w przypadku bardzo prostej macierzy A .

Przykład 1. Mamy

$$A = \begin{bmatrix} -1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}, \quad \mathbf{x}_0 = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}. \quad \text{Ponieważ } A\mathbf{x}_0 = \begin{bmatrix} -2 \\ 2 \\ \frac{1}{2} \end{bmatrix}$$

więc gdybyśmy chcieli wybrać współrzędną o największym module, mielibyśmy kłopot, czy wybrać 2, czy -2 . Jeśli zaś wybierzemy dodatnie współrzędne, to otrzymamy następujący ciąg wektorów $\mathbf{x}_j, j = 0, 1, \dots$

$$\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} -1 \\ 1 \\ \frac{1}{4} \end{bmatrix}, \quad \begin{bmatrix} 1 \\ -1 \\ \frac{1}{16} \end{bmatrix}, \quad \begin{bmatrix} -1 \\ 1 \\ \frac{1}{64} \end{bmatrix} \text{ itd.}$$

Ciąg $x_0, x_1, x_2, \dots$ nie jest zbieżny, natomiast ciąg $x_0, x_2, x_4, \dots$ jest zbieżny do wektora

$$x = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \text{ który jest wektorem własnym i odpowiada mu wartość własna } \lambda = -2.$$

W ogólnym przypadku na pytanie, kiedy ciąg $x_0, x_2, x_4, \dots$ jest zbieżny, niełatwo jest odpowiedzieć. Gdy jednak macierz A jest podobna do macierzy diagonalnej, tzn. istnieje taka macierz P , że $P^{-1}AP = D$, przy czym na diagonalu macierzy D są liczby $\lambda_1, \lambda_2, \dots, \lambda_n$, wówczas prawdziwe jest następujące twierdzenie.

Tw. *Jeśli wśród wartości własnych $\lambda_1, \lambda_2, \dots, \lambda_n$ macierzy A nie ma różnych wartości własnych o równych modułach (tzn. nie można wskazać takich $\lambda_i \neq \lambda_j$, by $|\lambda_i| = |\lambda_j|$), to przy dowolnym wektorze startowym x_0 ciąg wektorów $x_0, x_2, x_4, \dots$ jest zbieżny (zakładamy przy tym, że działania są wykonywane dokładnie).*

DOWÓD. Wykażemy prawdziwość tego twierdzenia jedynie w przypadku symetrycznej macierzy A , gdyż wystarczy to do zrozumienia zasady działania metody. (Czytelnik z łatwością uogólni dowód bez zakładania symetrii macierzy A).

Jeżeli macierz $A_{n \times n}$ jest symetryczna, to istnieje zbiór wektorów własnych, tworzący bazę ortonormalną w R^n . Oznaczmy wektory odpowiadające wartościom własnym $\lambda_1, \lambda_2, \dots$ odpowiednio przez $e_1, e_2, \dots$. Załóżmy, że $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|$. Niech $x_0 = \alpha_1 e_1 + \alpha_2 e_2 + \dots + \alpha_n e_n$. Może się zdarzyć, że $\alpha_1 = \alpha_2 = \dots = \alpha_{i-1} = 0, \alpha_i \neq 0$. Może się też zdarzyć, że wartość własna λ_i jest wielokrotna i na przykład $\lambda_i = \lambda_{i+1} = \dots = \lambda_{i+k} = \lambda$, przy czym albo $i+k = n$, albo $\lambda_{i+k+1} \neq \lambda$. Ze względu na przyjęte założenia mamy $\frac{|\lambda_{i+k+j}|}{|\lambda|} < 1 \quad j = 1, 2, \dots, n-i-k$.

Jeśli $\lambda = 0$, to

$$x_0 = \alpha_i e_i + \dots + \alpha_n e_n$$

jest wektorem własnym dla $\lambda = 0$ i mamy $Ax_0 = 0, \|x_0\|_\infty = 1$. Jeśli $\lambda \neq 0$, to oznaczając $a = \alpha_i e_i + \dots + \alpha_{i+k} e_{i+k}, \delta_0 = \alpha_{i+k+1} e_{i+k+1} + \dots + \alpha_n e_n$ otrzymamy

$$A^L x_0 = \lambda^L \left(a + \left(\frac{i+k+1}{\lambda} \right)^L \alpha_{i+k+1} e_{i+k+1} + \dots + \left(\frac{\lambda_n}{\lambda} \right)^L \alpha_n e_n \right)$$

Wprowadźmy oznaczenie

$$\delta_L = \left(\frac{\lambda_{i+k+1}}{\lambda} \right)^L \alpha_{i+k+1} e_{i+k+1} + \dots + \left(\frac{\lambda_n}{\lambda} \right)^L \alpha_n e_n. \text{ Wówczas}$$

$$\|\delta_L\|_2 = \left[\left(\frac{\lambda_{i+k+1}}{\lambda} \right)^{2L} \alpha_{i+k+1}^2 + \dots + \left(\frac{\lambda_n}{\lambda} \right)^{2L} \alpha_n^2 \right]^{1/2} \leq \left| \frac{\lambda_{i+k+1}}{\lambda} \right|^L \|\delta_0\|_2 \rightarrow 0$$

Zatem $\frac{A^L x_0}{\lambda^L} \rightarrow a$ i $\frac{\|A^L x_0\|_\infty}{\lambda^L} \rightarrow \|a\|_\infty$. Ciąg wektorów $y_L = \frac{A^L x_0}{\|A^L x_0\|_\infty} (\operatorname{sgn} \lambda)^L$ jest więc zbieżny do

$$\frac{a}{\|a\|_\infty}.$$

Zauważmy, że $y_L = x_L (\operatorname{sgn} \lambda)^L$, $L = 1, 2, \dots$, zatem $x_2 = y_2$, $x_4 = y_4$, ..., czyli ciąg $x_2, x_4, \dots$, $x_{2j}, \dots$ jest zbieżny do $\frac{a}{\|a\|_\infty}$. Wektor $\frac{a}{\|a\|_\infty}$ jest wektorem własnym macierzy A odpowiadającym wartości własnej λ . Ciąg $m_1, m_2, m_3, \dots$ jest zbieżny do $|\lambda|$. Jeśli $\lambda > 0$, to $y_j = x_j$, $j = 1, 2, \dots$, więc $\|x_{j+1} - x_j\|_\infty \rightarrow 0$ oraz $m_i \rightarrow \lambda$. Jeśli $\lambda < 0$, to $y_{2j+1} = -x_{2j+1}$, więc $\|x_{j+1} - x_j\|_\infty \rightarrow 2$ oraz $m_i \rightarrow -\lambda$ cnd.

Analizując powyższy dowód łatwo spostrzec, że wektory δ_L , o które różnią się wektory x_L od wektora własnego macierzy A , maleją co najmniej $\frac{|\lambda_{i+k+1}|}{|\lambda|}$ razy w każdej iteracji. Najbardziej niekorzystny jest przypadek $\delta_0 = \alpha_{i+k+1} \cdot e_{i+k+1} \neq 0$, gdyż wtedy mamy $\frac{\|\delta_{L+1}\|}{\|\delta_L\|} = \frac{|\lambda_{i+k+1}|}{|\lambda|}$, $L = 0, 1, 2, \dots$, więc jeśli moduły $|\lambda_{i+k+1}|$ i $|\lambda|$ mało się różnią, czyli $\frac{|\lambda_{i+k+1}|}{|\lambda|} \approx 1$, to zbieżność wektorów $x_0, x_2, x_4, \dots$ oraz liczb $m_1, m_2, m_3, \dots$ może być bardzo powolna. W wielu przypadkach można znacznie polepszyć zbieżność przez przesunięcie wartości własnych, tzn. przez zastąpienie macierzy A macierzą $B = A - pI$ (patrz tw. 13 w p. 6.1). Może się też zdarzyć, że wartości własne macierzy A obliczamy inną metodą, a metodę potęgową wykorzystujemy jedynie do wyznaczenia wektorów własnych. Wtedy najlepszą zbieżność do wektora własnego odpowiadającego wartości własnej λ_1 uzyskamy, przyjmując $p = \frac{1}{2}(\lambda_i + \lambda_n)$, gdzie λ_i jest następną w kolejności po λ_1 wartością własną, różną od λ_1 . Najlepszą zaś zbieżność do wektora odpowiadającego wartości λ_n uzyskamy, przyjmując $p = \frac{1}{2}(\lambda_1 + \lambda_j)$, gdzie λ_j jest wartością własną poprzedzającą λ_n i różną od niej. Omówienie innych możliwości poprawienia efektywności zbieżności metody potęgowej można znaleźć np. w [65].

Ćwiczenia

1. W jaki sposób można obliczyć najmniejszą co do modułu wartość własną macierzy A metodą potęgową? Wskazówka. Najpierw obliczyć λ_1 , a następnie przyjąć $p = \lambda_1$.
2. Co uzyskamy przyjmując $p = \|A\|_q$, $q = 1, 2, \infty, E$?
3. Dlaczego w praktyce należy się spodziewać, że startując z dowolnego x_0 uzyskamy zbieżność do λ_1 ?

6.4.4. Algorytm QR dla macierzy Hessenberga

Macierzą *Hessenberga* (ściślej: macierzą Hessenberga górną) nazywamy macierz $H = T + U$, gdzie T jest macierzą trójdziagonalną, a U — macierzą trójkątną górną, innymi słowy $h_{ij} = 0$ dla $j < i - 1$. Każda macierz kwadratowa jest ortogonalnie podobna do macierzy Hessenberga (patrz p. 6.4.5), jeśli więc potrafimy wyznaczyć wartości własne macierzy H , to można wyznaczyć wartości własne dowolnych macierzy.

Podamy teraz najprostszą postać tzw. *algorytmu QR*. Każdą macierz kwadratową A można przedstawić w postaci iloczynu macierzy ortogonalnej

Q i macierzy trójkątnej górnej R , tzn. w postaci $A = QR$. W szczególnym przypadku, gdy macierz A jest nieosobliwa, kolumny macierzy Q można otrzymać w wyniku ortonormalizacji kolejnych kolumn macierzy A procedurą Grama-Schmidta, a kolumny macierzy R składają się wówczas ze współczynników rozwinięcia kolumn A w uzyskanej bazie ortonormalnej (algorytm otrzymywania rozkładu QR jest omówiony dokładnie w p. 6.4.7).

Algorytm QR wyznaczania wartości własnych polega na określeniu ciągu macierzy $H = H^{(1)}, H^{(2)}, \dots$ w sposób następujący:

$$\begin{aligned} H^{(i)} &= Q^{(i)} R^{(i)} \\ H^{(i+1)} &= R^{(i)} Q^{(i)}, \quad i = 1, 2, \dots \end{aligned} \quad (6.10)$$

Łatwo stwierdzić, że wszystkie macierze $Q^{(i)}$ oraz $H^{(i)}$, $i = 1, 2, \dots$, występujące we wzorach (6.10) są macierzami Hessenberga. Ponadto dowodzi się (patrz np. [46]), że gdy wartości własne macierzy H są rzeczywiste i różne (także co do modułu), wówczas elementy pod diagonalą macierzy $H^{(i)}$ zanikają i macierze te upodabniają się do macierzy trójkątnej. Nie wszystkie co prawda elementy macierzy ponad diagonalą dążą do ustalonych wartości, lecz co najważniejsze elementy na diagonalu są zbieżne. Ponieważ $H^{(i+1)} = Q^{(i)T} H^{(i)} Q^{(i)}$, więc wszystkie macierze $H^{(i)}$, $i = 1, 2, \dots$, są podobne. Zatem elementy na diagonalu dążą do wartości własnych macierzy H .

Jeśli macierz H ma pojedyncze wartości własne $\lambda_1, \lambda_2, \dots, \lambda_n$ takie, że są między nimi pary sprzężone, lecz są to jedyne pary wartości własnych o równych modułach, to nie wszystkie elementy pod diagonalą macierzy $H^{(i)}$ dążą do zera, lecz $h_{j,j-1}^{(i)} h_{j+1,j}^{(i)} \xrightarrow{i \rightarrow \infty} 0$. W granicy otrzymamy np. macierz

$$H^{(\infty)} = \begin{bmatrix} * & * & . & . & . \\ * & * & . & . & . \\ 0 & 0 & * & . & . \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix} \quad (6.11)$$

w której kropki oznaczają elementy niekoniecznie dążące do ustalonych wartości, natomiast pozostałe elementy oznaczone $*$ są zbieżne. Wartości własne macierzy H leżą wtedy bądź bezpośrednio na diagonalu, bądź są one wartościami własnymi dwuwymiarowych macierzy znajdujących się na diagonalu.

W ogólnym przypadku macierze $H^{(i)}$ nie muszą dążyć do postaci (6.11). Na przykład dla macierzy

$$H = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (6.12)$$

mamy $H^{(1)} = H^{(2)} = \dots = H^{(i)}$, gdyż macierz H jest ortogonalna, a w rozkładzie $H = QR$ macierz R jest jednostkowa. Zbieżność elementów na diagonalu bywa też często bardzo powolna; typowa sytuacja jest taka, że część elementów szybko dąży do swych wartości granicznych, część zaś mimo wielu iteracji wykazuje nadal duże zmiany wartości. Podobnie jak w metodzie potęgowej spowodowane to jest bliskim jedności stosunkiem modułów niektórych wartości własnych. Można temu często zaradzić stosując przesunięcia wartości własnych, tzn. obliczając ciąg macierzy $H^{(1)}, H^{(2)}, H^{(3)}, \dots$ według wzorów

$$\begin{aligned} H^{(i)} - k_i I &= Q^{(i)} R^{(i)} \\ H^{(i+1)} &= R^{(i)} Q^{(i)} + k_i I \end{aligned} \quad (6.13)$$

Wówczas nadal $H^{(i+1)} = Q^{(i)T} (H^{(i)} - k_i I) Q^{(i)} + k_i I = Q^{(i)T} H^{(i)} Q^{(i)}$, a więc macierze $H^{(i)}$, $i = 1, 2, \dots$, są podobne do macierzy H , natomiast przez odpowiednie dobranie przesunięcia k_i uzyskujemy szybszą zbieżność elementów macierzy $H^{(i)}$. Zazwyczaj wybiera się przesunięcia k_i, k_{i+1} równe otrzymanym już wartościom własnym, a więc $k_i = a_{nn}^{(i)}$, lub k_i oraz k_{i+1} równe wartościom własnym macierzy o wymiarze 2×2 , usytuowanej w prawym dolnym rogu macierzy $H^{(i)}$. Ponieważ mogą to być wtedy wartości zespolone, takie postępowanie prowadziło do rachunku na macierzach o zespolonych elementach.

Aby uniknąć liczb zespolonych można wykonywać w takim przypadku od razu dwie iteracje (patrz np. [46], [67] i [65]). Postępowanie to zawodzi jednak dla macierzy takich jak (6.12), lecz jeśli w jakiś sposób możemy stwierdzić, że mamy do czynienia z takim przypadkiem, to można uzyskać zbieżność wykonując wstępne przesunięcie, równe np. normie macierzy $H^{(i)}$.

Przykład. Jeżeli macierzy H jest dana wzorem (6.12), to $H^{(1)} = Q^{(1)} R^{(1)}$, gdzie $Q^{(1)} = H^{(1)}$, $R^{(1)} = I$, a więc orientujemy się, że ani według wzorów (6.10), ani (6.13) nie wyznaczymy wartości własnych macierzy H . Rozpatrzmy zatem macierz $\bar{H} = H - \|H\|I$, tzn. macierz postaci

$$\bar{H} = \begin{bmatrix} -1 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

Wówczas

$$\bar{H}^{(1)} = \bar{Q}^{(1)} \bar{R}^{(1)} = \begin{bmatrix} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & -\frac{1}{2\sqrt{3}} & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & -\frac{1}{2\sqrt{3}} & 0 \\ 0 & \frac{2}{\sqrt{6}} & -\frac{1}{2\sqrt{3}} & 0 \\ 0 & 0 & \frac{\sqrt{3}}{2} & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2} & -\frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & \frac{3}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \frac{-1}{\sqrt{6}} \\ 0 & 0 & \frac{2}{\sqrt{3}} & \frac{-2}{\sqrt{3}} \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

a więc

$$\bar{H}^{(2)} = \bar{R}^{(1)} Q^{(1)} = \begin{bmatrix} -2 & -\frac{1}{2\sqrt{3}} & -\frac{\sqrt{2}}{\sqrt{3}} & 0 \\ \frac{\sqrt{3}}{2} & -\frac{7}{\sqrt{6}} & -\frac{1}{2\sqrt{2}} & 0 \\ 0 & \frac{2\sqrt{2}}{3} & -\frac{2}{3} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Ponieważ $h_{4,3}^{(2)} = h_{4,4}^{(2)} = 0$, więc macierz $\bar{H}$ ma wartość własną $\lambda_4 = 0$. Opuszczamy czwarty wiersz i czwartą kolumnę macierzy $\bar{H}^{(2)}$, a pozostałe wartości własne macierzy $\bar{H}$ wyznaczamy posługując się macierzą

$$\bar{H} = \begin{bmatrix} -2 & -\frac{1}{2\sqrt{3}} & -\frac{\sqrt{2}}{\sqrt{3}} \\ \frac{\sqrt{3}}{2} & -\frac{7}{\sqrt{6}} & -\frac{1}{2\sqrt{2}} \\ 0 & \frac{2\sqrt{2}}{3} & -\frac{2}{3} \end{bmatrix}$$

której wartości własne są równe $\bar{\lambda}_3 = -1 - j$, $\bar{\lambda}_2 = -1 + j$, $\bar{\lambda}_1 = -2$. Przy takim rozkładzie wartości własnych ciąg macierzy $\bar{H}^{(i)}$ szybko dąży do macierzy o zerowym elemencie $\bar{h}_{3,2}$. Wtedy wyznaczymy wartość własną $\bar{\lambda}_1$ jako granicę elementów $\bar{h}_{3,3}^{(i)}$, natomiast $\bar{\lambda}_2$ i $\bar{\lambda}_3$ będą wartościami własnymi macierzy o wymiarze 2×2 , znajdującej się w lewym górnym rogu macierzy $\bar{H}^{(i)}$. Oczywiście wartości własne początkowej macierzy H obliczymy, dodając do liczb $\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3, \bar{\lambda}_4$ wstępne przesunięcie $k = \|H\| = 1$.

U w a g a. Metodę QR można stosować nie tylko dla macierzy Hessenberga, lecz dla dowolnych macierzy A . Wówczas jednak obliczenie rozkładu $A^{(i)} = Q^{(i)} R^{(i)}$ jest bardziej czasochłonne. Zaleca się zatem wstępnie przekształcić macierz A do postaci Hessenberga przez podobieństwo $P^{-1}AP = H$, a dopiero potem zastosować algorytm QR .

6.4.5. Sprowadzanie macierzy do postaci Hessenberga

Opracowano wiele sposobów przekształcania przez podobieństwo macierzy A do postaci Hessenberga. Ze względu na dokładność obliczeń najlepsze są trzy metody; są to *metoda Givensa* i *metoda Householdera* (analogicznie do metod omówionych w p. 6.5.2), w których stosuje się przekształcenia ortogonalne, oraz *metoda eliminacji Gaussa*.

Dwie pierwsze metody wymagają odpowiednio $\frac{10}{3}n^3 + O(n^2)$ oraz $\frac{5}{3}n^3 + O(n^2)$ mnożeń, przy czym w arytmetyce zmiennopozycyjnej otrzymujemy macierz H podobną do $A + E$, gdzie $\|E\|_E \leq k_1 \varepsilon n^{3/2} \|A\|_E$ dla metody Givensa oraz $\|E\|_E \leq k_2 \varepsilon n^2 \|A\|_E$ dla metody Householdera. Stałe k_1 i k_2 są to współczynniki rzędu 1, zależne od sposobu zaokrąglania na maszynie cyfrowej.

Podobnych oszacowań dokładności nie uzyskano dla metody eliminacji Gaussa. Duża użyteczność tej metody została jednak potwierdzona w praktyce i jest ona powszechnie stosowana ze względu na mały nakład obliczeń; liczba mnożeń w tej metodzie wynosi $\frac{2}{3}n^3 + O(n^2)$, a więc jest istotnie mniejsza niż dla poprzednich metod.

Opis metody eliminacji Gaussa. Dla danej macierzy $A_{n \times n}$ dokonujemy $(n-2)$ przekształceń uzyskując kolejno macierze $A = A^{(1)}, A^{(2)}, \dots, A^{(n-1)} = H$, gdzie H jest poszukiwaną macierzą Hessenberga podobną do A . Załóżmy, że mamy już wyznaczoną macierz $A^{(i)}$, której $(i-1)$ początkowych kolumn jest już równe $(i-1)$ kolumnom macierzy H . Przekształcenie prowadzące od macierzy $A^{(i)}$ do $A^{(i+1)}$ przebiega następująco:

1. Wybieramy element o największym module spośród elementów $a_{i+1,i}^{(i)}, a_{i+2,i}^{(i)}, \dots, a_{n,i}^{(i)}$. Jeśli jest kilka takich elementów, to przyjmujemy pierwszy z nich. Jeśli są to same zerowe elementy, to stwierdzamy, że $A^{(i+1)} = A^{(i)}$ i przystępujemy do kolejnego przekształcenia, tj. do wyznaczenia macierzy $A^{(i+2)}$. Gdy maksimum jest dodatnie i element o maksymalnym module znajduje się w k -tym wierszu ($k \geq i$), wówczas przestawiamy wiersze i kolumny o numerach k oraz $i+1$.

2. Obliczamy mnożniki $m_j^{(i)} = a_{j,i}^{(i)} / a_{i+1,i}^{(i)}, j = i+2, i+3, \dots, n$, przy czym zgodnie z wyborem elementu $a_{i+1,i}^{(i)}$, mamy $|m_j^{(i)}| \leq 1$, co zapewnia numeryczną stabilność przekształceń względem błędów zaokrągleń.

3. Od j -tego wiersza odejmujemy $(i+1)$ -szy wiersz pomnożony przez $m_j^{(i)}$ oraz do $(i+1)$ -szej kolumny dodajemy j -tą kolumnę pomnożoną przez $m_j^{(i)}$, $j = i+2, i+3, \dots, n$. Uzyskana macierz $A^{(i+1)}$ jest podobna do $A^{(i)}$, a elementy $a_{i+2,i}^{(i+1)}, a_{i+3,i}^{(i+1)}, \dots, a_{n,i}^{(i+1)}$ są równe zeru, co powoduje, że i -ta kolumna macierzy $A^{(i+1)}$ jest równa i -tej kolumnie macierzy H .

Ćwiczenia

1. Wykazać, że przekształcenie macierzy $A^{(i)}$ w macierz $A^{(i+1)}$ jest podobieństwem. Wskazać, że $A^{(i+1)} = (M^{(i)})^{-1} (P_{i+1,k})^{-1} A^{(i)} (P_{i+1,k}) (M^{(i)})$, gdzie $P_{i+1,k}$ jest macierzą przedstawień $(i+1)$ -szej kolumny i k -tej kolumny, zaś $m_{j,i+1}^{(i)} = m_j^{(i)}, j = i+2, i+3, \dots, n$; $m_{jl}^{(i)} = \delta_{ji}$ dla pozostałych elementów.

6.4.6. Obliczanie wektorów własnych

W poprzednich punktach omówiliśmy metody wyznaczania wartości własnych macierzy. Obecnie zajmiemy się wyznaczaniem wektorów własnych. Jeżeli znamy przekształcenie P takie, że $P^{-1}AP = B$, przy czym potrafimy obliczyć wektor własny macierzy B , na przykład $Bx = \lambda x$, to wektor $y = Bx$ będzie wektorem własnym macierzy A . Gdy macierz B jest macierzą trójkątną górną i ma różne wartości własne, wówczas wartości własnej $\lambda_i = b_{ii}$ odpowiada wektor $x^{(i)}$, który obliczamy ze wzorów:

$$\begin{aligned}
x_j^{(i)} &= 0, & j &= n, n-2, \dots, i+1 \\
x_i^{(i)} &= 1 \\
x_j^{(i)} &= -\left(\sum_{k=j+1}^i b_{jk} x_k^{(i)}\right)/(b_{jj} - b_{ii}), & j &= i-1, i-2, \dots, 1
\end{aligned} \tag{6.14}$$

Jeśli macierz B ma wielokrotną wartość własną λ , to wektor własny odpowiadający tej wartości można obliczyć jedynie dla najmniejszego indeksu i , dla którego $b_{ii} = \lambda$. Bez dodatkowych informacji o macierzy B (równoważnych stwierdzeniu, jakie klatki Jordana ma dana wartość własna λ) pozostałych wektorów własnych nie można obliczyć, zwłaszcza, że może się zdarzyć, iż wartości λ odpowiada tylko jeden wektor własny.

Metoda QR. Gdy obliczamy wartości własne macierzy A za pomocą rozkładu QR (patrz p. 5.4.4), wówczas wyznaczone macierze $Q^{(1)}, Q^{(2)}, \dots$ można wykorzystać do obliczenia wektorów własnych macierzy A . W tym celu wyznaczamy najpierw macierz P taką, że $P^{-1}AP = H$ (macierz Hessenberga), a następnie według wzorów (6.10) lub (6.13) obliczamy ciąg macierzy $H^{(i)}$. Aby obliczyć wektory własne, należy wyznaczyć równocześnie z macierzą $H^{(i+1)} = Q^{(i)T}H^{(i)}Q^{(i)}$ pomocniczą macierz $Z^{(i+1)} = Z^{(i)}Q^{(i)}$, $Z^{(1)} = P$. Wówczas, jeżeli stwierdzimy, że macierz $H^{(N)}$ jest dostatecznie bliska macierzy trójkątnej (tzn. elementy pod diagonalą są dostatecznie małymi liczbami), możemy obliczyć jej wektory własne $x^{(k)}$ według wzorów (6.14), a następnie obliczyć wektory własne $y^{(k)}$ macierzy A ze wzoru

$$y^{(k)} = Z^{(N)}x^{(k)} \tag{6.15}$$

Metoda LR. Jest to metoda bardzo podobna do metody QR . Poszukując wektorów własnych macierzy A obliczamy ciąg macierzy $A = A^{(1)}, A^{(2)}, \dots$ ze wzorów

$$\begin{aligned}
A^{(i)} &= L^{(i)}U^{(i)} \\
A^{(i+1)} &= U^{(i)}L^{(i)}, & i &= 1, 2, \dots
\end{aligned} \tag{6.16}$$

w których L i U są odpowiednio macierzą trójkątną dolną i górną. Macierze $A^{(i+1)} = (L^{(i)})^{-1}A^{(i)}L^{(i)}$ są podobne do A , jeśli więc ciąg $A^{(i)}$ dąży do macierzy trójkątnej, to możemy obliczyć wszystkie wartości własne macierzy A . Aby wyznaczyć wektory własne macierzy A należy obliczyć macierze $Z^{(i+1)} = Z^{(i)}L^{(i)}$, $Z^{(1)} = I$ i zastosować wzór (6.15).

Uzyskanie rozkładu $A^{(i)} = L^{(i)}U^{(i)}$ wymaga $\frac{1}{3}n^3 + O(n^2)$ mnożeń, a więc

3 razy mniejszej liczby działań niż w rozkładzie $A^{(i)} = Q^{(i)}R^{(i)}$. Sugerowałoby to przewagę metody LR nad QR , jednakże stosując przekształcenie QR popełniamy mniejsze błędy zaokrągleń. Gdy jednak macierz A jest symetryczna i dodatnio określona, wówczas przekształcenie LR jest stabilne numerycznie, zatem w porównaniu z metodą QR jest to metoda szybsza.

Podobnie jak w metodzie **QR** w celu przyspieszenia zbieżności ciągu macierzy stosujemy przesunięcie $A^{(i)} - k_i I$, a także sprowadzamy wstępnie macierz A do postaci Hessenberga, co zmniejsza liczbę obliczeń w dalszych iteracjach (każdy rozkład $H = LU$ wymaga $\frac{1}{2}n^2 + O(n)$ mnożeń).

Metoda iteracji odwrotnej. Istota działania metod **QR** i **LR** jest podobna do działania metody potęgowej (patrz p. 6.4.3), tzn. następuje zanikanie niektórych składowych wektorów (kolumn macierzy) na skutek różnych wartości własnych macierzy A . Zbieżność wektorów w metodzie potęgowej może być jednak powolna, np. ze współczynnikiem $\frac{|\lambda_{i+k+1}|}{|\lambda|} \approx 1$, natomiast zbieżność w metodach **QR** i **LR** przy właściwym wyborze przesunięć k_i (patrz np. [65]) bliskich wartości własnej λ_n (jej dobrym przybliżeniem jest a_{nn}^i) jest zależna od $\frac{|\lambda_n - k_i|}{|\lambda_{n-1} - k_i|}$, a więc następuje przyspieszenie zbieżności do zbieżności nadliniowej.

Odmianą metody potęgowej, umożliwiającą analogiczne przyspieszenie zbieżności, jest metoda iteracji odwrotnej o następującym algorytmie.

0. Dla $i = 0$ wybieramy dowolny wektor x_0 taki, że $\|x_0\|_\infty = 1$, i ustalamy ITM (maksymalną liczbę iteracji),

1. Obliczamy wektor $v_{i+1} = (A - k_i I)^{-1} x_i$, $m_{i+1} = \|v_{i+1}\|_\infty$ oraz $x_{i+1} = \frac{v_{i+1}}{m_{i+1}}$,

2. Jeżeli $i+1 \geq \text{ITM}$, to Stop. W przeciwnym razie $i = i+1$, więc powtarzamy etap 1.

Jeżeli wartościami własnymi macierzy A są liczby $\lambda_1, \lambda_2, \dots, \lambda_n$, przy czym $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|$, to wartościami własnymi macierzy $(A - k_i I)^{-1}$ są liczby $\frac{1}{\lambda_1 - k_i}, \frac{1}{\lambda_2 - k_i}, \dots, \frac{1}{\lambda_n - k_i}$. Dobierając odpowiednio przesunięcia k_i uzyskamy zbieżność wektorów $x^{(i)}$ do wektora własnego x ze współczynnikiem $\frac{|\lambda_n - k_i|}{|\lambda - k_i|}$, gdzie λ jest najbliższą liczbie λ_n wartością własną różną od niej. Ciąg liczb $m_1, m_2, \dots$ umożliwia ponadto wyznaczenie wartości własnej λ_n , analogicznie jak w metodzie potęgowej (patrz p. 6.4.3). Podobnie jak w przypadku przekształceń **QR** i **LR** przyjmując kolejno jako k_i wyznaczone przybliżenia wartości własnej λ_n , uzyskamy zbieżność nadliniową. Gdy jednak liczba k_i jest bliska wartości λ_n , wówczas macierz $(A - k_i I)$ jest macierzą źle uwarunkowaną, co jak wiemy prowadzi do dużych błędów przy wyznaczaniu wektora v_{i+1} . Można jednak uzasadnić, że jeśli tylko wektor własny x jest dobrze uwarunkowany, to błędy powstałe przy obliczaniu wektora v_{i+1} (tzn. przy rozwiązywaniu układu równań $(A - k_i I)v_{i+1} = x_i$) nie będą istotne, gdy zadbamy o takie wyznaczenie wektora v_{i+1} , by jego norma $\|v_{i+1}\|_\infty = m_{i+1}$ była duża.

Założmy na przykład, że $k_i = \lambda_n + e_n$ i że wyznaczony wektor $\mathbf{v}_{i+1}$ jest dokładnym rozwiązaniem układu $(\mathbf{A} - k_i \mathbf{I} + \mathbf{E})\mathbf{v}_{i+1} = \mathbf{x}_i$, gdzie $\|\mathbf{E}\|_\infty = e$. Wtedy

$$(\mathbf{A} - \lambda_n \mathbf{I})\mathbf{v}_{i+1} = (e_n \mathbf{I} - \mathbf{E})\mathbf{v}_{i+1} + \mathbf{x}_i$$

$$(\mathbf{A} - \lambda_n \mathbf{I}) \frac{\mathbf{v}_{i+1}}{\|\mathbf{v}_{i+1}\|_\infty} = (e_n \mathbf{I} - \mathbf{E}) \frac{\mathbf{v}_{i+1}}{\|\mathbf{v}_{i+1}\|_\infty} + \frac{\mathbf{x}_i}{\|\mathbf{v}_{i+1}\|_\infty}$$

skąd

$$\|(\mathbf{A} - \lambda_n \mathbf{I}) \mathbf{x}_{i+1}\|_\infty \leq (|e_n| + e) + \frac{1}{\|\mathbf{v}_{i+1}\|_\infty} \quad (6.17)$$

Z powyższego oszacowania wynika zatem, że należy co prawda stosować taką metodę rozwiązywania układu równań $(\mathbf{A} - k_i \mathbf{I})\mathbf{v}_{i+1} = \mathbf{x}_i$, by błąd $\|\mathbf{E}\|_\infty$ był mały, lecz fakt że współczynnik uwarunkowania macierzy $(\mathbf{A} - k_i \mathbf{I})$ jest bardzo duży nie jest zbyt groźny, jeżeli tylko $\|\mathbf{v}_{i+1}\|_\infty$ będzie duża. Uzyskamy wtedy mały wektor reszt $(\mathbf{A} - \lambda_n \mathbf{I})\mathbf{x}_{i+1}$, co przy dobrze uwarunkowanym wektorze $\mathbf{x}$ oznacza, że $\mathbf{x}_{i+1}$ jest dobrym jego przybliżeniem.

Ćwiczenia

1. Uzasadnić, że $\mathbf{y}^{(k)} = \mathbf{Z}^{(N)} \mathbf{x}^{(k)}$ jest wektorem własnym macierzy $\mathbf{A}$.
2. Jaką liczbę działań arytmetycznych należy wykonać w jednej iteracji w metodzie iteracji odwrotnej?

6.4.7. Algorytmy rozkładu macierzy na iloczyn $\mathbf{QR}$

U podstaw wielu metod korzystających z przekształcania macierzy za pomocą macierzy ortogonalnych (patrz def. 5 w p. 6.2) leży następujące twierdzenie.

Tw. 1. Każdą macierz $\mathbf{A}_{m \times n}$, $m \geq n$, posiadającą $k \leq n$ liniowo niezależnych kolumn można przedstawić w postaci iloczynu $\mathbf{A} = \mathbf{QR}$, gdzie $\mathbf{Q}_{m \times k}$ jest macierzą ortogonalną, a $\mathbf{R}_{k \times n}$ jest macierzą trapezoidalną górną.

Macierz $\mathbf{R}$ nazywamy trapezoidalną górną, jeżeli wszystkie elementy tej macierzy znajdujące się pod diagonalą są równe zeru, tzn. $r_{ij} = 0$ dla $i > j$.

Dowód twierdzenia 1 łatwo jest uzyskać stosując opisaną poniżej procedurę ortonormalizacji kolumn macierzy $\mathbf{A}$ metodą Grama-Schmidta. Wówczas kolumny macierzy $\mathbf{Q}$ są kolejno otrzymywanymi ortonormalnymi wektorami, a kolumny macierzy $\mathbf{R}$ — współczynnikami rozwinięcia kolumn macierzy $\mathbf{A}$ w bazie, jaką tworzą kolumny macierzy $\mathbf{Q}$.

Wniosek. Jeżeli macierz $\mathbf{A}_{m \times n}$, $m > n$, ma pełny rząd (tzn. $k = n$), to macierz $\mathbf{A}$ można przedstawić w postaci iloczynu $\mathbf{A} = \mathbf{QR}$, gdzie $\mathbf{Q}_{m \times n}$ jest macierzą ortogonalną, a $\mathbf{R}_{n \times n}$ jest macierzą trójkątną górną.

Tw. 2. Jeżeli $\mathbf{A}_{m \times n}$, $m \geq n$, ma pełny rząd, to istnieje jednoznaczny rozkład $\mathbf{A} = \mathbf{QR}$ taki, że elementy na diagonalu macierzy $\mathbf{R}$ są dodatnie.

Dowód patrz np. [65, rozdz. 4, s. 53].

Metoda Grama-Schmidta. Rozkład QR można wyznaczyć *metodą ortogonalizacji Grama-Schmidta* zastosowaną do kolumn, $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ macierzy A . Niech $A^{(1)}, A^{(2)}, \dots, A^{(n)}$ oznaczają macierze złożone z 1, 2, ..., n początkowych kolumn macierzy A . Załóżmy, że kolumna $\mathbf{a}_1$ jest niezerowa. Wtedy

$$A^{(1)} = Q^{(1)}R^{(1)}, \text{ gdzie}$$

$$Q^{(1)} = \begin{bmatrix} \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|_2} \end{bmatrix}, \quad R^{(1)} = [\|\mathbf{a}_1\|_2]_2$$

Jeżeli wyznaczyliśmy już rozkład $A^{(i)} = Q^{(i)}R^{(i)}$, to rozkład QR dla macierzy $A^{(i+1)}$ znajdziemy korzystając z następującego zapisu blokowego:

$$A^{(i+1)} = Q^{(i+1)}R^{(i+1)}$$

gdzie

$$A^{(i+1)} = [A^{(i)} \mathbf{a}_{i+1}], \quad Q^{(i+1)} = [Q^{(i)} \mathbf{q}_{i+1}],$$

$$R^{(i+1)} = \begin{bmatrix} R^{(i)} & \mathbf{r}_{i+1} \\ 0 & s_{i+1} \end{bmatrix}$$

Nieznane wektory $\mathbf{q}_{i+1}$ i $\mathbf{r}_{i+1}$ oraz liczbę s_{i+1} wyznaczamy ze wzorów:

$$\begin{aligned} \mathbf{r}_{i+1} &= (Q^{(i)})^T \mathbf{a}_{i+1} \\ \mathbf{p}_{i+1} &= \mathbf{a}_{i+1} - Q^{(i)} \mathbf{r}_{i+1} \\ s_{i+1} &= \|\mathbf{p}_{i+1}\|_2 \end{aligned} \tag{6.18}$$

Jeżeli $s_{i+1} = 0$, to $Q^{(i+1)} = Q^{(i)}$.

Jeżeli $s_{i+1} \neq 0$, to $\mathbf{q}_{i+1} = \frac{\mathbf{p}_{i+1}}{s_{i+1}}$.

Wzory (6.18) stosujemy dla $i = 1, 2, \dots, n - 1$.

Łatwo stwierdzić, że uzyskana tym sposobem macierz $Q^{(n)}$ jest macierzą ortogonalną, a $R^{(n)}$ — macierzą trapezoidalną górną oraz że $A = Q^{(n)}R^{(n)}$. Macierz $Q^{(n)}$ ma jednak tylko tyle kolumn, ile liniowo niezależnych kolumn miała macierz A . Przy stosowaniu algorytmu QR (patrz p. 6.4.4 i 6.4.6) należy zatem pamiętać, że obliczając macierz RQ za pomocą wzorów (6.10) lub (6.13) możemy otrzymać macierz o zredukowanym wymiarze. Oczywiście można sobie łatwo poradzić z tą pozorną „trudnością” na przykład uzupełniając zerami macierze Q i R do macierzy kwadratowych, ale powoduje to komplikację algorytmu.

Poważniejszym mankamentem metody Grama-Schmidta jest możliwość utraty własności ortogonalności w ciągu wyznaczanych macierzy $Q^{(1)}, Q^{(2)}, \dots, Q^{(n)} = Q$. Jest to spowodowane zbyt dużym zmniejszeniem dokładności obliczeń, w przypadku gdy kolumny macierzy A są prawie liniowo zależne. Można tego uniknąć stosując *reortogonalizację* wektora $\mathbf{p}_{i+1}$ względem kolumn macierzy $Q^{(i)}$ (patrz np. [16]), powoduje to jednak wzrost nakładu obliczeń.

Aby wyznaczyć rozkład $A = QR$ nieosobliwej macierzy kwadratowej A bez reortogonalizacji, należy wykonać $M = n^3 + n$ mnożeń, $D = n^3 - \frac{1}{2}n^2 - \frac{1}{2}n$ dodawań oraz n pierwiastkowań. Omówimy poniżej dwie z metod stosowanych do wyznaczania tego rozkładu.

Metoda Householdera. Na wstępie rozważmy następujące zagadnienie. Wyznaczyć macierz ortogonalną $H_{k \times k}$ taką, by dla danego wektora $z = [z_1, z_2, \dots, z_k]^T$ wynik działania H na z miał postać $\alpha \|z\|_2 \cdot [1, 0, \dots, 0]^T$, $\alpha = \pm 1$, tzn. $H z = \alpha \|z\|_2 e_1$.

Łatwo sprawdzić, że macierz

$$H = I - \frac{1}{\tau} u u^T \quad (6.19)$$

(definicję iloczynu zewnętrznego podano w p. 4.2), gdzie $u = z - \alpha \|z\|_2 e_1$, $\tau = \|z\|_2 (\|z\|_2 - \alpha e_1)$ dla $\|z\|_2 \neq 0$ oraz $\tau = 1$ dla $\|z\|_2 = 0$, spełnia ten warunek. Macierz postaci (6.19) zwana jest *macierzą Householdera*. Aby uniknąć nadmiernego wzrostu błędów zaokrągleń, spowodowanego dzieleniem przez liczbę τ , wybieramy na ogół $\alpha = -\text{sgn } z_1$, tzn. $\alpha = -1$ dla $z_1 \geq 0$ oraz $\alpha = +1$ dla $z_1 < 0$.

Określimy teraz ciąg macierzy $P^{(1)}, P^{(2)}, \dots, P^{(n-1)}$ takich, że

$$P^{(n-1)} P^{(n-2)} \dots P^{(2)} P^{(1)} A = R$$

$R_{n \times n}$ jest macierzą trójkątną górną.

Niech $P^{(1)} = H^{(1)}$, gdzie macierz $H^{(1)}$ sprowadza pierwszą kolumnę $a_1^{(1)}$ macierzy A do postaci $\alpha^{(1)} \|a_1^{(1)}\|_2 \cdot [1, 0, \dots, 0]^T_{1 \times n}$.

Niech $P^{(2)} = \begin{bmatrix} 1 & 0 & & \\ \vdots & & & \\ 0 & H^{(2)} & & \end{bmatrix}$, gdzie macierz $H^{(2)}$ sprowadza pierwszą kolumnę $a_1^{(2)}$ macierzy o wymiarze $(n-1) \times (n-1)$ (utworzonej przez kolumny i wiersze macierzy $P^{(1)}A$ o numerach 2, 3, ..., n) do postaci

$$\alpha^{(2)} \|a_1^{(2)}\|_2 \cdot [1, 0, \dots, 0]^T_{1 \times (n-1)} \quad \text{itd.}$$

$$\text{Ostatnia macierz } P^{(n-1)} = \begin{bmatrix} 1 & & & & \\ \cdot & & & & \\ & \cdot & & & \\ & & \cdot & & \\ & & & \cdot & \\ \text{---} & & & & 1 & \text{---} \\ & & & & & H^{(n-1)} \end{bmatrix}, \text{ gdzie } H^{(n-1)} \text{ sprowa-}$$

dza pierwszą kolumnę $a_1^{(n-1)}$ macierzy o wymiarze 2×2 (utworzonej przez kolumny i wiersze macierzy $P^{(n-2)} \dots P^{(2)} P^{(1)} A$ o numerach $(n-1)$ i n) do postaci

$$\alpha^{(n-1)} \|\mathbf{a}_1^{(n-1)}\|_2 \cdot [1, 0]^T$$

Ostatecznie $A = QR$, przy czym

$$Q = P^{(1)} P^{(2)} \dots P^{(n-1)}$$

$$R = Q^T A$$

Przykład 1. Wyznaczyć rozkład QR dla macierzy

$$A = \begin{bmatrix} 0 & -1 & -3 \\ 0 & 0 & -2 \\ 0 & -2 & -1 \end{bmatrix}$$

Mamy:

dla $i = 1$

$$\mathbf{a}_1^{(1)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow P^{(1)} = H^{(1)} = I_{3 \times 3}, \quad P^{(1)} A = \begin{bmatrix} 0 & -1 & -3 \\ 0 & 0 & -2 \\ 0 & -2 & -1 \end{bmatrix}$$

dla $i = 2$

$$\mathbf{a}_1^{(2)} = \begin{bmatrix} 0 \\ -2 \end{bmatrix} \Rightarrow H^{(2)} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad P^{(2)} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

$$P^{(2)} P^{(1)} A = \begin{bmatrix} 0 & -1 & -3 \\ 0 & -2 & -1 \\ 0 & 0 & -2 \end{bmatrix}$$

$$\text{Ostatecznie: } Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad R = \begin{bmatrix} 0 & -1 & -3 \\ 0 & -2 & -1 \\ 0 & 0 & -2 \end{bmatrix}$$

Liczba działań potrzebnych do wyznaczenia macierzy $P^{(1)}, P^{(2)}, \dots, P^{(n-1)}$ i R wynosi $M = \frac{2}{3}n^3 + 2n^2 - \frac{2}{3}n - 2$ mnożeń, $D = \frac{2}{3}n^3 + \frac{10}{3}n - 4$ dodawań oraz $n - 1$ pierwiastkowań.

Należy zwrócić uwagę, że aby uzyskać wynik działania macierzy H na dowolny wektor y , nie potrzeba obliczać wszystkich elementów macierzy H , lecz wystarczy skorzystać z równości $Hy = y - \frac{\langle u, y \rangle}{\tau} u$. Podobnie, jeśli nie musimy z jakichś względów obliczać macierzy Q za pomocą wyznaczania jej elementów, wystarczy zapamiętać wektory $u^{(i)}$ oraz liczby $\tau^{(i)}$ macierzy $P^{(i)}$, $i = n - 1, n - 2, \dots, 2, 1$, i dla dowolnego wektora y obliczyć $v = Qy$ według następującego algorytmu:

$$v_j = y_j, \quad j = 1, 2, \dots, n$$

Dla $i = n - 1, n - 2, \dots, 1$ obliczamy

$$\beta = 0$$

$$\begin{aligned} \beta &\Rightarrow \beta + u_j^{(i)} v_j, & j = n, n - 1, \dots, i \\ \beta &\Rightarrow \beta / \tau^{(i)} \end{aligned} \quad (6.20)$$

Dla $j = n, n - 1, \dots, i$ obliczamy

$$v_j \Rightarrow v_j - \beta u_j^{(i)}$$

Wymaga to $M = n^2 + 2n - 3$ i $D = n^2 - 1$ działań arytmetycznych. Jeśli jednak musimy obliczyć macierz $\mathbf{Q}$, to obliczamy kolumny $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n$ tej macierzy ze wzorów (6.20) przyjmując $\mathbf{y} = \mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$. Należy przy tym pamiętać, że $\mathbf{P}^{(1)} \mathbf{P}^{(2)} \dots \mathbf{P}^{(n-1)} \mathbf{e}_k = \mathbf{P}^{(1)} \mathbf{P}^{(2)} \dots \mathbf{P}^{(k)} \mathbf{e}_k$ dla $k \leq n - 1$. Wówczas liczba mnożeń i dodawań potrzebnych do obliczenia $\mathbf{Q}$ wynosi $\frac{2}{3}n^3 + O(n^2)$.

Metoda Givensa. Zasada działania tej metody jest analogiczna do zasady działania metody Householdera, tzn. za pomocą odpowiednio dobranych macierzy ortogonalnych sprowadzamy do zer elementy macierzy $\mathbf{A}$ znajdujące się pod diagonalą. Idea tej metody jest następująca. Wyznaczamy macierz ortogonalną $\mathbf{G}_{k \times k}$, zwaną *macierzą Givensa* taką, by dla danego wektora $\mathbf{z} = [z_1, z_2, \dots, z_k]^T$ wynik działania $\mathbf{G}$ na $\mathbf{z}$ był postaci $\alpha \|\mathbf{z}\|_2 \cdot [1, 0, \dots, 0]^T$, $\alpha = \pm 1$, tzn. $\mathbf{G}\mathbf{z} = \alpha \|\mathbf{z}\|_2 \mathbf{e}_1$.

Tym razem nie wyznaczamy macierzy $\mathbf{G}$ bezpośrednio, lecz przedstawiamy ją w postaci iloczynu elementarnych macierzy Givensa

$$\mathbf{G}^{(j)} = \begin{bmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & \ddots & & & & \\ & & & c^{(j)} & s^{(j)} & & \\ & & & s^{(j)} & -c^{(j)} & \dots & \\ & & 0 & & \ddots & \ddots & \\ & & & \ddots & \ddots & \ddots & 1 \end{bmatrix} \quad \begin{array}{l} \leftarrow j\text{-ty wiersz} \\ \uparrow \\ j\text{-ta kolumna} \end{array} \quad (6.21)$$

gdzie liczby $c^{(j)}$ i $s^{(j)}$ są tak dobrane, by po pomnożeniu macierzy $\mathbf{G}^{(j)}$ przez wektor $\mathbf{z}^{(j)} = [z_1^{(j)}, \dots, z_j^{(j)}, z_{j+1}^{(j)}, \dots, z_n^{(j)}]^T$ otrzymać wektor o zerowej $(j+1)$ -szej współrzędnej. Łatwo sprawdzić, że liczby $c^{(j)}$ i $s^{(j)}$ obliczone ze wzorów:

$$\begin{aligned} \varrho_1 &= \sqrt{(z_j^{(j)})^2 + (z_{j+1}^{(j)})^2} \\ \varrho_2 &= \alpha \varrho_1, \quad \alpha = \pm 1 \end{aligned} \quad (6.22)$$

$$\begin{cases} c^{(j)} = z_j^{(j)} / \varrho_2 \\ s^{(j)} = z_{j+1}^{(j)} / \varrho_2, \end{cases} \quad \text{jeżeli} \quad \varrho_1 \neq 0$$

$$\begin{cases} c^{(j)} = 1 \\ s^{(j)} = 0, \end{cases} \quad \text{jeżeli} \quad \varrho_1 = 0$$

spełniają ten warunek. Mamy bowiem

$$\begin{bmatrix} c^{(j)} & s^{(j)} \\ s^{(j)} & -c^{(j)} \end{bmatrix} \begin{bmatrix} z_j^{(j)} \\ z_{j+1}^{(j)} \end{bmatrix} = \varrho_2 \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Macierz $\mathbf{G}$ wyznaczamy jako iloczyn $\mathbf{G}^{(2)} \dots \mathbf{G}^{(n-1)} \mathbf{G}^{(n)}$, gdzie $\mathbf{G}^{(n)}$ sprowadza do zera element z_n wektora $\mathbf{z}$, macierz $\mathbf{G}^{(n-1)}$ sprowadza do zera element z_{n-1} wektora $\mathbf{G}^{(n)} \mathbf{z}$, ..., macierz $\mathbf{G}^{(2)}$ sprowadza do zera element z_2 wektora $\mathbf{G}^{(3)} \dots \mathbf{G}^{(n)} \mathbf{z}$.

Niech teraz macierz $\mathbf{G}_1 = \mathbf{G}_1^{(2)} \mathbf{G}_1^{(3)} \dots \mathbf{G}_1^{(n)}$ sprowadza pierwszą kolumnę $\mathbf{a}_1^{(1)}$ macierzy $\mathbf{A}$ do postaci

$$\alpha^{(1)} \|\mathbf{a}_1^{(1)}\|_2 \cdot [1, 0, \dots, 0]_{1 \times n}^T$$

a macierz $\mathbf{G}_2$ niech sprowadza pierwszą kolumnę $\mathbf{a}_1^{(2)}$ macierzy o wymiarze $(n-1) \times (n-1)$ (utworzonej przez kolumny i wiersze macierzy $\mathbf{G}_1 \mathbf{A}$ o numerach 2, 3, ..., n) do postaci

$$\alpha^{(2)} \|\mathbf{a}_1^{(2)}\| \cdot [1, 0, \dots, 0]_{1 \times (n-1)}^T \quad \text{itd.}$$

Ostatnia macierz $\mathbf{G}_{n-1}$ sprowadza pierwszą kolumnę $\mathbf{a}_1^{(n-1)}$ macierzy o wymiarze 2×2 (utworzonej przez kolumny i wiersze macierzy $\mathbf{G}_{n-2} \mathbf{G}_{n-3} \dots \mathbf{G}_1 \mathbf{A}$ o numerach $(n-1)$ i n) do postaci $\alpha^{(n-1)} \|\mathbf{a}_1^{(n-1)}\|_2 \cdot [1, 0]^T$. Ostatecznie $\mathbf{A} = \mathbf{QR}$, gdzie

$$\mathbf{Q} = \mathbf{G}_1^T \mathbf{G}_2^T \dots \mathbf{G}_{n-1}^T = (\mathbf{G}_1^{(n)} \dots \mathbf{G}_1^{(3)} \mathbf{G}_1^{(2)}) (\mathbf{G}_2^{(n-1)} \dots \mathbf{G}_2^{(3)} \mathbf{G}_2^{(2)}) \dots \mathbf{G}_{n-1}^{(2)}$$

$$\mathbf{R} = \mathbf{Q}^T \mathbf{A}$$

Przykład 2. Wyznaczyć rozkład $\mathbf{QR}$ macierzy

$$\mathbf{A} = \begin{bmatrix} 0 & -1 & -3 \\ 0 & 0 & -2 \\ 0 & -2 & -1 \end{bmatrix}$$

Mamy:

$$\text{dla } i = 1 \quad \mathbf{a}_1^{(1)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \mathbf{G}_1^{(3)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

$$\mathbf{G}_1^{(3)} \mathbf{a}_1^{(1)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \mathbf{G}_1^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$G_1 = G_1^{(2)} G_1^{(3)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \quad G_1 A = \begin{bmatrix} 0 & -1 & -3 \\ 0 & 0 & 2 \\ 0 & 2 & 1 \end{bmatrix}$$

$$\text{dla } i = 2 \quad a_1^{(2)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow G_2^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

$$G_2 = G_2^{(2)}, \quad G_2 G_1 A = \begin{bmatrix} 0 & -1 & -3 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix}$$

Ostatecznie

$$Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix}, \quad R = \begin{bmatrix} 0 & -1 & -3 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix}$$

Liczba działań potrzebnych do wyznaczenia ciągu $G_1^{(n)}, G_1^{(n-1)}, \dots, G_{n-1}^{(2)}$ i macierzy R wynosi $M = \frac{4}{3}n^3 - \frac{4}{3}n$ mnożeń, $D = \frac{2}{3}n^3 - \frac{1}{2}n^2 - \frac{1}{6}n$ dodawań oraz $\frac{1}{2}n^2 - \frac{1}{2}n$ pierwiastkowań. Można jednak tak zorganizować obliczenia (patrz [16], [29]), by nakład mnożeń i dodawań wynosił $n^3 + O(n^2)$. Oczywiście tego rodzaju zmiany warto wprowadzić jedynie wtedy, gdy mnożenie na używanej maszynie cyfrowej trwa istotnie dłużej niż dodawanie.

Ćwiczenia

1. Stosując wzory $y_1 = cz_1 + sz_2$, $y_2 = \mu(z_1 + y_1) - z_2$, gdzie $\mu = \frac{s}{c+1}$, uzasadnić, że zamiast 4 mnożeń i 2 dodawań przy obliczaniu iloczynu $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} c & s \\ s & -c \end{bmatrix} \cdot \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$ wystarczy wykonać 3 mnożenia i 3 dodawania.
2. Obliczyć liczbę działań potrzebnych do wyznaczenia macierzy Q metodą Givensa dla dowolnej macierzy $A_{n \times n}$.
3. Udowodnić, że macierz $G^{(n)} G^{(n-1)} \dots G^{(2)}$ jest transponowaną macierzą Hessenberga. Macierze $G^{(j)}$ są to macierze Givensa postaci (6.21).

6.5. MACIERZE SYMETRYCZNE

Obliczanie wartości własnych i wektorów własnych macierzy symetrycznych jest zagadnieniem znacznie prostszym niż w przypadku innych macierzy z kilku powodów:

1. Wskutek symetrii macierzy można dwukrotnie zmniejszyć liczbę komórek pamięci oraz nakład obliczeń przy wykonywaniu przekształceń macierzy.

2. Dodatkowe informacje o tym, że wartości własne i wektory własne są rzeczywiste pozwalają na uproszczenie stosowanych metod.

3. Znamy liczbę wektorów własnych i wiemy, że istnieje ich pełny wzajemnie ortogonalny układ.

4. Zagadnienie wyznaczania wartości własnych jest bardzo dobrze uwarunkowane (patrz p. 6.3).

5. Dla macierzy symetrycznych dodatnio określonych można stosować metodę Banachiewicza rozwiązywania układów równań i sprowadzania macierzy do iloczynu macierzy trójkątnych; metoda Banachiewicza jest bardzo dokładna i wymaga małej liczby działań.

Wobec powyższych własności macierzy symetrycznych wiele ogólnych metod poszukiwania wartości własnych i wektorów własnych ma zagwarantowaną skuteczność działania i na ogół wyznaczanie tych wielkości jest szybsze. Mówią o tym poniższe twierdzenia.

Tw. 1. *Jeżeli macierz A jest symetryczna i dodatnio określona, to przy dowolnym wektorze początkowym x_0 zbieżność metody potęgowej jest zagwarantowana.*

Tw. 2. *Jeżeli macierz A jest symetryczna i dodatnio określona, to algorytm QR generuje ciąg macierzy $A^{(1)}, A^{(2)}, \dots$, zbieżny do macierzy diagonalnej.*

W przypadku stosowania algorytmu QR macierze $A^{(i+1)} = R^{(i)} Q^{(i)} = Q^{(i)T} A^{(i)} Q^{(i)}$ zachowują swą symetrię, jeśli zaś stosować algorytm LR dla macierzy symetrycznych, to w ogólnym przypadku symetria nie jest zachowana. Dla macierzy dodatnio określonych można jednak zamiast dowolnego rozkładu $A^{(i)} = L^{(i)} U^{(i)}$ zastosować rozkład Banachiewicza $A^{(i)} = L^{(i)} L^{(i)T}$. Wtedy ciąg macierzy $A^{(1)}, A^{(2)}, \dots$ generowany algorytmem LR zachowuje symetrię, gdyż $A^{(i+1)} = L^{(i)T} L^{(i)}$. Ponadto prawdziwe jest następujące twierdzenie.

Tw. 3. *Jeżeli A jest macierzą symetryczną i dodatnio określoną, zaś $A = A^{(1)}, A^{(2)}, A^{(3)}, \dots, A^{(i)}$ oraz $\bar{A} = \bar{A}^{(1)}, \bar{A}^{(2)}, \bar{A}^{(3)}, \dots, \bar{A}^{(i)}$ są ciągami generowanymi odpowiednio przez algorytm QR oraz LR z rozkładem Banachiewicza, to $\bar{A}^{(2k+1)} = A^{(k+1)}$.*

U w a g a. Z twierdzenia 3 i 2 wynika zatem, że algorytm LR w wersji Banachiewicza jest zbieżny dla symetrycznej, dodatnio określonej macierzy (patrz np. [65]). Ponadto twierdzenie 3 uzasadnia opinię, że metoda QR jest szybciej zbieżna niż metoda LR , biorąc pod uwagę liczbę iteracji. Ponieważ jednak uzyskanie rozkładu LR wymaga blisko 3-krotnie mniejszej liczby działań niż uzyskanie rozkładu QR , obie metody są porównywalne pod względem nakładu obliczeń.

Obie metody należy stosować po uprzednim sprowadzeniu macierzy A do prostszej postaci. Ze względu na symetrię macierz A sprowadzamy zwykle do postaci trójdzielnej (patrz p. 6.5.2).

6.5.1. Macierz trójdagonalna symetryczna

Jeżeli

$$T = \begin{bmatrix} b_1 & a_2 & & & 0 \\ a_2 & b_2 & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot & a_n \\ 0 & & & \cdot & \cdot & a_n & b_n \end{bmatrix}$$

to pojedyncze wartości własne i wektory własne, jak i wszystkie wartości własne i wektory własne można obliczyć wykonując stosunkowo niewielką liczbę działań za pomocą różnych metod. Metody te są omówione w [6], [21], [54], [65], [66], [67], my zaś ograniczymy się do omówienia metody bisekcji i metody **QR**.

Metoda bisekcji. Jeżeli chcemy wyznaczyć pojedyncze wartości własne macierzy T , postępujemy następująco. Dla dowolnie wybranej liczby λ obliczamy wartość wielomianu charakterystycznego $w(\lambda) = \det(T - \lambda I)$. Ze względu na szczególną postać macierzy T można skorzystać ze wzoru rekurencyjnego:

$$\begin{aligned} \omega_0(\lambda) &= 1, & \omega_1(\lambda) &= b_1 - \lambda \\ \omega_i(\lambda) &= (b_i - \lambda)\omega_{i-1} - a_i^2\omega_{i-2}(\lambda) & i &= 2, 3, \dots, n \\ \omega(\lambda) &= \omega_n(\lambda) \end{aligned} \tag{6.23}$$

Dla obliczenia $\omega(\lambda)$ wykonujemy zatem $M = 3n - 3$ mnożeń i $D = 2n - 1$ dodawań. Korzystając z dodatkowej informacji, że zera wielomianu $\omega(\lambda)$ są rzeczywiste można by teraz skorzystać z jakiejś ogólnej metody poszukiwania zer $\omega(\lambda) = 0$ (patrz p. 3.2), lecz wygodniej jest skorzystać z następujących twierdzeń:

Tw. 1. *Jeżeli elementy $a_2, a_3, \dots, a_n$ są niezerowe, to wartości własne macierzy T są pojedyncze.*

Tw. 2. *Jeżeli elementy $a_2, a_3, \dots, a_n$ są niezerowe, to ciąg wartości $\omega_0(\lambda), \omega_1(\lambda), \dots, \omega_n(\lambda)$ spełnia warunki:*

- 1) *Jeżeli $\omega_i(\lambda) = 0$ dla pewnego $i < n$, to $\omega_{i-1}(\lambda)\omega_{i+1}(\lambda) < 0$.*
- 2) *Jeżeli $\omega_n(\lambda) \neq 0$, to liczba zmian znaków sąsiednich liczb $\omega_0(\lambda), \omega_1(\lambda), \dots, \omega_n(\lambda)$ jest równa liczbie wartości własnych macierzy T mniejszych od λ .*
- 3) *Jeżeli $\omega_n(\lambda) = 0$, to λ jest wartością własną macierzy T , a ponadto jest tyle wartości własnych mniejszych niż λ , ile nastąpiło zmian znaków w ciągu $\omega_0(\lambda), \omega_1(\lambda), \dots, \omega_{n-1}(\lambda)$.*

U w a g a. Aby uniknąć niejednoznaczności przy porównywaniu znaków liczb $\omega_i(\lambda)$ oraz $\omega_{i+1}(\lambda)$, w przypadku gdy któraś z nich jest równa zero, można przyjąć, że np. liczba zero ma znak dodatni.

W dowodach tych twierdzeń należy skorzystać z twierdzenia Sturma (patrz s. 134).

Przykład 1. Wyznaczyć wartości własne macierzy

$$T = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

Na podstawie (6.23) mamy:

Dla $\lambda = 0$

$$\omega_0(0) = 1, \quad \omega_1(0) = 1, \quad \omega_2(0) = -3, \quad \omega_3(0) = -4$$

Wystąpiła jedna zmiana znaku ($\omega_1(0)\omega_2(0) < 0$), więc macierz T ma jedną ujemną wartość własną.

Dla $\lambda = 1$

$$\omega_0(1) = 1, \quad \omega_1(1) = 0, \quad \omega_2(1) = -4, \quad \omega_3(1) = 0$$

Liczba $\omega_3(\lambda) = 0$, więc $\lambda = 1$ jest wartością własną macierzy T . Między liczbami $\omega_0(1)$, $\omega_1(1)$, $\omega_2(1)$ występuje jedna zmiana znaku, więc macierz T ma jedną wartość własną mniejszą niż 1.

Dla $\lambda = 3$

$$\omega_0(3) = 1, \quad \omega_1(3) = -2, \quad \omega_2(3) = 0, \quad \omega_3(3) = 2$$

Wystąpiły tu dwie zmiany znaku, więc macierz T ma dwie wartości własne mniejsze niż 3.

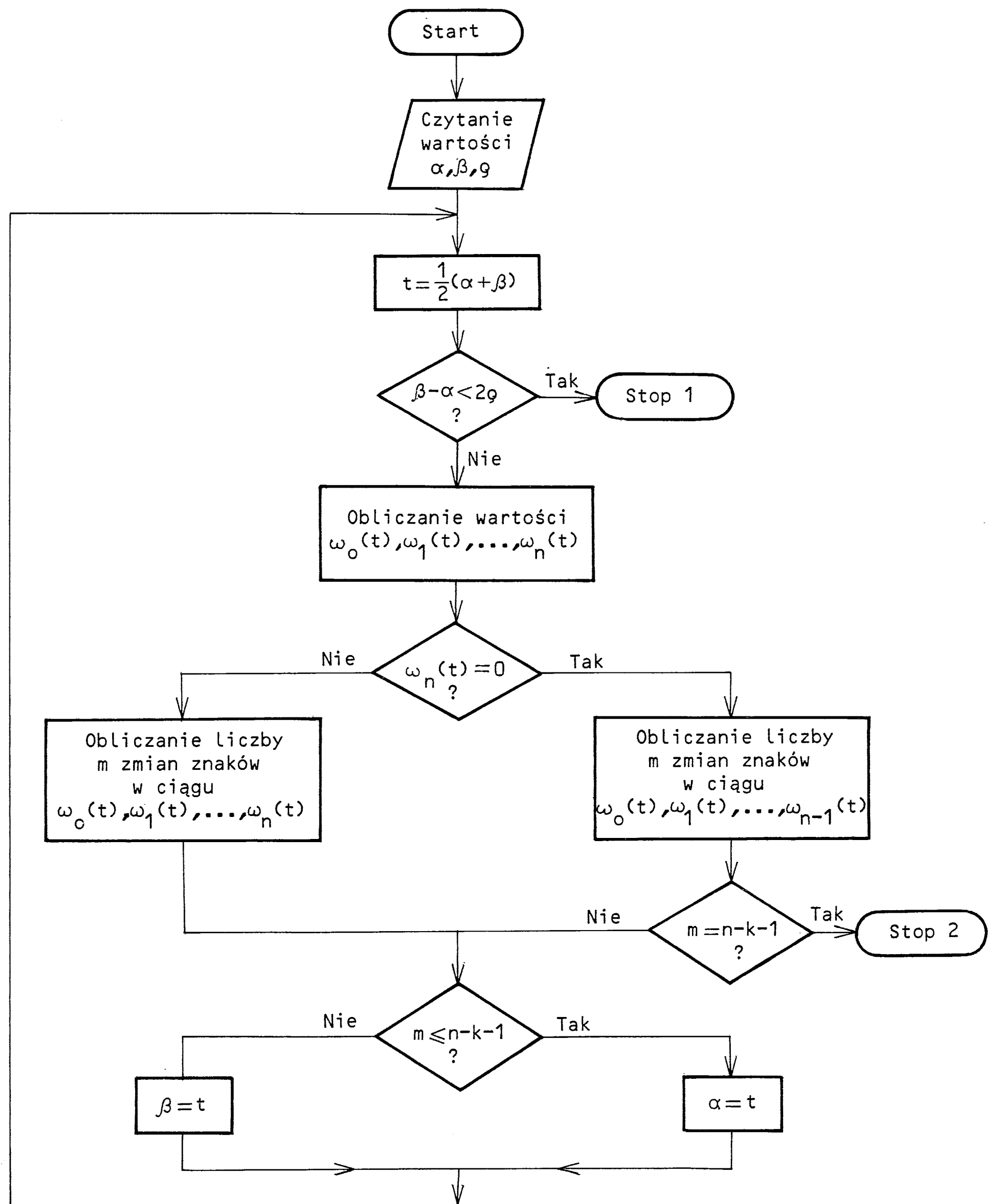
Macierz T ma wartości własne $\lambda_1 = 1 + \sqrt{5}$, $\lambda_2 = 1$, $\lambda_3 = 1 - \sqrt{5}$.

Posługując się twierdzeniem 2 można wyznaczyć spomiędzy wartości własnych $\lambda_1 > \lambda_2 > \dots > \lambda_n$ wybraną wartość λ_k stosując następujący algorytm (rys. 6.2).

Niech k oznacza numer wartości własnej, którą zamierzamy wyznaczyć, $\varrho < 0$ — dopuszczalny błąd obliczenia λ_k , a α i β ($\alpha < \beta$) — końce przedziału, w którym zlokalizowano wszystkie wartości własne macierzy T (np. $\alpha = -\|T\|$, $\beta = \|T\|$).

Jeżeli nastąpi Stop 1, to uzyskana wartość t jest równa liczbie λ_k z błędem nieprzekraczającym ϱ . Jeżeli nastąpi Stop 2, to uzyskana wartość t jest równa λ_k (z dokładnością do błędów zaokrągleń). Liczba iteracji IT potrzebna do wyznaczenia λ_k wynosi $IT = E\left(\log_2 \frac{\beta_0 - \alpha_0}{\varrho}\right)$, gdzie β_0 i α_0 — początkowe wartości liczb β i α . Jeśli wstępnie obliczymy liczbę 2ϱ oraz współczynniki a_2^2 , a_3^2 , ..., a_n^2 , potrzebne do wyznaczenia wartości $\omega_0(t)$, $\omega_1(t)$, ..., $\omega_n(t)$, to przy IT iteracjach trzeba wykonać $M = IT(2n - 1) + 1$ mnożeń.

Metoda powyższa jest bardzo dokładna. Jej jedynym mankamentem jest fakt osiągnięcia bardzo dużych (nadmiarowych) wartości elementów ciągu $\omega_0(\lambda)$, $\omega_1(\lambda)$, ..., $\omega_n(\lambda)$ nawet dla niezbyt dużych n , jeśli wartość λ odbiega znacznie od wartości własnych macierzy T . Z tego względu zamiast liczb



6.2

$\omega_0(\lambda), \dots, \omega_n(\lambda)$ lepiej jest obliczać liczby $p_1(\lambda) = \frac{\omega_1(\lambda)}{\omega_0(\lambda)}, \dots, p_n(\lambda) = \frac{\omega_n(\lambda)}{\omega_{n-1}(\lambda)}$, które można wyznaczyć ze wzoru rekurencyjnego

$$\begin{aligned} p_1(\lambda) &= b_1 - \lambda \\ p_i(\lambda) &= (b_i - \lambda) - a_i^2/p_{i-1}(\lambda), \quad i = 2, \dots, n \end{aligned} \quad (6.24)$$

Może się zdarzyć, że $p_{i-1}(\lambda) = 0$ i wówczas wzór (6.24) wymaga modyfikacji. Ciąg wartości $p_1(\lambda), p_2(\lambda), \dots, p_n(\lambda)$ można w takim przypadku obliczyć następująco:

0. $i = 1$
1. $p_i(\lambda) = b_i - \lambda$. Jeżeli $i = n$, to Stop.
2. Jeżeli $p_i(\lambda) \neq 0$, wykonujemy krok 3. W przeciwnym przypadku przyjmujemy $p_{i+1}(\lambda) = -1$ oraz $i \Rightarrow i + 2$. Jeżeli $i > n$, to Stop. W przeciwnym przypadku wracamy do kroku 1.
3. $i \Rightarrow i + 1$. Jeżeli $i > n$, to Stop. W przeciwnym przypadku $p_i(\lambda) = (b_i - \lambda) - a_i^2/p_{i-1}(\lambda)$; wracamy do kroku 2.

(6.25)

Poprzednio obliczana liczba zmian znaków w ciągu $\omega_0(\lambda), \omega_1(\lambda), \dots, \omega_n(\lambda)$ jest równa liczbie ujemnych wartości spośród $p_1(\lambda), p_2(\lambda), \dots, p_n(\lambda)$, zatem w tym przypadku przy IT iteracjach algorytm z rys. 6.2 wymaga co najwyżej $M = IT \times n + 1$ mnożeń i dzielení.

Dokładność wyznaczania wartości własnych metodą bisekcji. Cechą szczególną metody bisekcji jest to, że dokładność wyznaczonej wartości własnej nie zależy w sposób istotny od wartości liczb $\omega_0(\lambda), \omega_1(\lambda), \dots, \omega_n(\lambda)$ lub liczb $p_1(\lambda), \dots, p_n(\lambda)$, a jedynie od ich znaków. Uwzględnimy to analizując dokładność metody dla wzorów (6.25). W systemie zmiennopozycyjnym, scharakteryzowanym przez liczbę ε (zakładamy, że $\varepsilon < 0,01$), obliczona wartość p_i wynosi

$$\text{fl}(p_i(\lambda)) = b_i(1 + \varepsilon_1)(1 + \varepsilon_3) - \lambda(1 + \varepsilon_2)(1 + \varepsilon_3) - (a_i^2/p_{i-1}(\lambda))(1 + \varepsilon_4)(1 + \varepsilon_5)(1 + \varepsilon_6)$$

Bez zmiany znaku można tę liczbę podzielić przez $(1 + \varepsilon_2)(1 + \varepsilon_3)$, a zatem znaki liczb $\text{fl}(p_i(\lambda))$ są takie same jak znaki liczb

$$b_i \frac{(1 + \varepsilon_1)}{(1 + \varepsilon_2)} - \lambda - \frac{a_i^2}{p_{i-1}(\lambda)} \frac{(1 + \varepsilon_4)(1 + \varepsilon_5)(1 + \varepsilon_6)}{(1 + \varepsilon_2)(1 + \varepsilon_3)}$$

Decyzje, w jaki sposób dzielić kolejny przedział $\langle \alpha^{(i)}; \beta^{(i)} \rangle$ powinny być więc takie, jak w przypadku macierzy $T + \delta T$, przy czym elementy δT są ograniczone następująco:

$$\begin{aligned} |\delta b_i| &< 2,04\varepsilon |b_i| \\ |\delta a_i| &< 2,06\varepsilon |a_i| \end{aligned} \tag{6.26}$$

Zatem $\|\delta T\|_\infty < 2,06\varepsilon \|T\|_\infty$.

Można uzasadnić (śledząc wpływ ewentualnych błędnych decyzji spowodowanych innym położeniem wartości własnych macierzy T i $T + \delta T$), że jeśli wykonaliśmy m podziałów początkowego przedziału $\langle \alpha^{(0)}; \beta^{(0)} \rangle$, to odległość środka t ostatniego przedziału od λ_k można oszacować jako

$$|t - \lambda_k| < \|\delta T\|_\infty + (\beta^{(0)} - \alpha^{(0)}) \cdot 2^{-m-1}$$

Przyjmując np. $\beta^{(0)} = \|T\|_\infty$, $\alpha^{(0)} = -\|T\|_\infty$, otrzymamy

$$|t - \lambda_k| < (2,06\varepsilon + 2^{-m}) \|T\|_\infty \tag{6.27}$$

Z oszacowania (6.27) widać, że metoda bisekcji umożliwia wyznaczenie wartości własnych macierzy T z niezwykle dokładnością.

Wektory własne. Jeżeli obliczyliśmy wartość własną λ macierzy T , to wektor własny x odpowiadający tej wartości obliczamy ze wzorów

$$\begin{aligned} x_1 &= 1 \\ x_2 &= (\lambda - b_1)/a_2 \\ x_{i+1} &= [(\lambda - b_i)x_i - a_i x_{i-1}]/a_{i+1}, \quad i = 2, 3, \dots, n-1 \end{aligned} \quad (6.28)$$

U w a g a. Powyższe wzory dotyczą przypadku, gdy wszystkie liczby $a_2, a_3, \dots, a_n$ są niezerowe. Jeśli jednak któreś z nich są równe zeru, to dokonujemy podziału macierzy T na bloki, które są macierzami trójdiodagonalnymi. Wartości własne tych macierzy są wartościami własnymi macierzy T , a wektory własne tych macierzy są po uzupełnieniu zerami wektorami własnymi macierzy T . W ten sposób można wyznaczyć nawet wielokrotne wartości własne i odpowiadające im wektory własne.

Metoda QR. Jeżeli macierz T jest trójdiodagonalna, to aby uzyskać rozkład $T = QR$ konstruujemy ciąg macierzy ortogonalnych $P^{(1)}, P^{(2)}, \dots, P^{(n-1)}$ takich, że

$$P^{(n-1)}P^{(n-2)} \dots P^{(2)}P^{(1)}T = R$$

Macierz $P^{(1)}$ ma za zadanie sprowadzić do zera element t_{21} macierzy T , zatem może to być macierz Givensa $G^{(2)}$ lub macierz Householdera $H^{(1)} = I - \frac{u \rangle \langle u}{\tau}$ taka, że wektor u ma jedynie dwie początkowe składowe niezerowe. Zarówno macierz $G^{(2)}$, jak i macierz $H^{(1)}$ są postaci

$$\begin{bmatrix} a & b & & & 0 \\ b & -a & & & \\ & & 1 & & \\ & & & \ddots & \\ 0 & & & & \ddots & \\ & & & & & 1 \end{bmatrix}$$

i można wykazać, że różnią się ewentualnie jedynie znakiem elementów a i b (patrz ćwic. 1). Wobec tego metody Householdera i Givensa są w tym przypadku równoważne i mogą się różnić jedynie sposobem zapisywania elementów w pamięci maszyny cyfrowej. Wydaje się, że dogodniejsza jest do tego celu postać Givensa dla macierzy $P^{(1)}, P^{(2)}, \dots, P^{(n-1)}$. Wyznamy zatem macierze Givensa takie, że $G^{(n)}G^{(n-1)} \dots G^{(3)}G^{(2)}T = R$. Łatwo sprawdzić, że macierz $Q = G^{(2)} \dots G^{(n)}$ jest macierzą Hessenberga (górną) i na ogół wszystkie jej elementy są niezerowe (np. $q_{1n} = s^{(2)} \dots s^{(n)} \neq 0$, gdy elementy pod diagonalną macierzy T są niezerowe). Wobec tego zamiast obliczać elementy tej macierzy dogodniej jest zapamiętać ją w postaci $(n-1)$ par liczb $c^{(j)}, s^{(j)}$, $j = 2, 3, \dots, n$. Jeśli jednak zachodzi konieczność obliczenia macierzy Q , to ma ona postać

$$Q = \begin{bmatrix} p_1 \beta_1 & p_2 \beta_1 & \dots & p_n \beta_1 \\ s_2 & p_2 \beta_2 & \dots & p_n \beta_2 \\ & s_3 & \ddots & \\ 0 & & \ddots & s_n \\ & & & p_n \beta_n \end{bmatrix} \quad (6.29a)$$

przy czym liczby p_j i $\beta_j, j = 1, \dots, n$, można wyznaczyć (patrz [29]) następująco:

$$1. \quad p_1 = c_2, \quad \beta_1 = 1, \quad \alpha_1 = s_2, \quad \gamma_1 = s_2$$

2. Dla $j = 2, 3, \dots, n-1$ znajdujemy

$$p_j = c_{j+1} \alpha_{j-1}, \quad \gamma_j = s_{j+1}, \quad \beta_j = -c_j / \alpha_{j-1}, \quad (6.29b)$$

$$\alpha_j = s_{j+1} \alpha_{j-1}$$

$$3. \quad p_n = \alpha_{n-1}, \quad \beta_n = -c_n / p_n$$

Przykład 2. Pokażemy sposób wykorzystania wzorów (6.29) umożliwiających wyznaczenie elementów macierzy Q . Jeżeli

$$Q = \begin{bmatrix} c_2 & s_2 & 0 & 0 \\ s_2 & -c_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c_3 & s_3 & 0 \\ 0 & s_3 & -c_3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & c_4 & s_4 \\ 0 & 0 & s_4 & -c_4 \end{bmatrix}$$

to

$$p_1 = c_2, \quad \beta_1 = 1; \quad p_2 = c_3 s_2, \quad \beta_2 = -c_2 / s_2 \\ p_3 = c_4 s_2 s_3, \quad \beta_3 = -c_3 / (s_3 s_2); \quad p_4 = s_2 s_3 s_4, \quad \beta_4 = -c_4 / (s_2 s_3 s_4)$$

a zatem

$$Q = \begin{bmatrix} c_2 & c_3 s_2 & c_4 s_2 s_3 & s_2 s_3 s_4 \\ s_2 & -c_2 c_3 & -c_2 c_3 c_4 & -c_2 s_3 s_4 \\ 0 & s_3 & -c_3 c_4 & -c_3 s_4 \\ 0 & 0 & s_4 & -c_4 \end{bmatrix}$$

Przy posługiwaniu się wzorami (6.29) trzeba wykonać łącznie $M = \frac{1}{2}n^2 + \frac{5}{2}n - 5$ mnożeń.

Macierz R ma postać

$$R = \begin{bmatrix} r_1 & p_1 & q_1 & 0 \\ & \ddots & \ddots & \ddots \\ & & \ddots & q_{n-2} \\ 0 & & & p_{n-1} \\ & & & & r_n \end{bmatrix}$$

Jeśli więc stosujemy algorytm QR i obliczamy

$$\tilde{T} = RQ$$

to nie jest oczywiste, że uzyskana macierz jest nadal trójdzielna. Łatwo jednak sprawdzić, że iloczyn macierzy trójkątnej górnej i macierzy Hessenberga jest nadal macierzą Hessenberga. Ponadto $RQ = Q^T T Q$, więc rezultat mnożenia jest macierzą symetryczną, co jest możliwe tylko wtedy, gdy jest to macierz trójdzielna. Fakt, że macierz $\tilde{T}$ jest symetryczna można wykorzystać przy posługiwaniu się algorytmem QR , gdyż do wyznaczenia macierzy $\tilde{T}$ wystarczy obliczyć jej elementy na diagonalu i sąsiednie pod diagonalą. Jeśli obliczymy iloczyn $R G^{(2)} \dots G^{(n)}$, to otrzymamy macierz

$$\tilde{T} = \begin{bmatrix} r_1 c_2 + p_1 s_2 & * & & & & \\ r_2 s_2 & & \cdot & * & & 0 \\ & \cdot & & \cdot & * & \\ 0 & & \cdot & & \cdot & * \\ & & r_n s_n & -c_{n-1} r_{n-1} c_n + p_{n-1} s_n & & * \end{bmatrix} \quad (6.30)$$

Reasumując, aby wyznaczyć macierze $G^{(2)}, \dots, G^{(n)}$ i R trzeba wykonać $M = 9n - 9$ mnożeń, $D = 3n - 3$ dodawań i $(n - 1)$ pierwiastkowań. Nie uwzględniono przy tym obliczania elementów $q_1, \dots, q_{n-2}$ macierzy R , gdyż nie są one potrzebne do wyznaczenia macierzy $\tilde{T}$. Aby obliczyć macierz $\tilde{T} = R G^{(2)} \dots G^{(n)}$ według wzorów (6.30) trzeba wykonać $M = 4n - 1$ mnożeń i $D = n$ dodawań.

Obliczane elementy macierzy $R, G^{(2)}, \dots, G^{(n)}$ zajmują łącznie $4n - 3$ komórek pamięci.

Ćwiczenia

1. Wykazać, że jeśli macierz $Q_{2 \times 2}$ jest ortogonalna i symetryczna oraz dla danego wektora $z = [z_1, z_2]^T$ mamy $Qz = \|z\| \cdot [1, 0]^T$, to warunek $q_{11} \geq 0$ określa macierz Q jednoznacznie.

6.5.2. Sprowadzanie macierzy symetrycznej do postaci trójdzielnej

Jak już podkreśliliśmy poprzednio, przed przystąpieniem do wyznaczania wartości własnych lub wektorów własnych macierzy symetrycznej A , wskazane jest przekształcić tę macierz przez podobieństwo w symetryczną macierz trójdzielną. Okazuje się, że można znaleźć macierz ortogonalną P taką, że $PAP^T = T$. Macierz P wyznaczamy analogicznie jak w przypadku rozkładu QR , stosując metodę Householdera lub Givensa.

Metoda Householdera. Określamy ciąg macierzy $P^{(1)}, P^{(2)}, \dots, P^{(n-2)}$ takich, że:

$$P^{(1)} = \begin{bmatrix} 1 & 0 \\ 0 & H^{(1)} \end{bmatrix}$$

gdzie $H^{(1)}$ sprowadza wektor $\mathbf{z}^{(1)} = [a_{21}, a_{31}, \dots, a_{n1}]^T$ do postaci $\alpha^{(1)} \|\mathbf{z}^{(1)}\|_2 \times [1, 0, \dots, 0]_{1 \times n-1}^T$, i obliczamy macierz

$$M^{(2)} = P^{(1)} A P^{(1)} = \begin{bmatrix} b_1 & a_1 & 0 & \dots & 0 \\ a_1 & & & & \\ 0 & A_{(n-1) \times (n-1)}^{(2)} & & & \\ \vdots & & & & \\ 0 & & & & \end{bmatrix}$$

Następnie określamy macierz

$$P^{(2)} = \begin{bmatrix} 1 & 0 & \\ & 1 & 0 \\ & 0 & H^{(2)} \end{bmatrix}$$

gdzie $H^{(2)}$ sprowadza wektor $\mathbf{z}^{(2)} = [a_{21}^{(2)}, \dots, a_{(n-1)1}^{(2)}]^T$ do postaci $\alpha^{(2)} \|\mathbf{z}^{(2)}\|_2 \cdot [1, 0, \dots, 0]_{1 \times (n-2)}^T$ i obliczamy macierz

$$M^{(3)} = P^{(2)} M^{(2)} P^{(2)} = \begin{bmatrix} b_1 & & a_1 & | & 0 & \dots & 0 \\ a_1 & b_2 & & | & a_2 & 0 & \dots & 0 \\ 0 & a_2 & & | & & & & \\ \vdots & 0 & & | & A_{(n-2) \times (n-2)}^{(3)} & & & \\ \vdots & \vdots & & | & & & & \\ 0 & 0 & & | & & & & \end{bmatrix}$$

Po $(n-2)$ takich przekształceniach uzyskamy macierz $M^{(n-1)} = T$, gdzie T jest poszukiwaną macierzą trójdziagonalną.

Aby wykonać powyższe przekształcenia przy jak najmniejszej liczbie działań zauważmy, że za każdym razem przekształcamy jedynie macierze $A^{(i)}$, $i = 2, 3, \dots, n-1$. Wynik mnożenia $H^{(i)} A^{(i)} H^{(i)}$ jest macierzą symetryczną, więc wystarczy obliczyć jedynie elementy tej macierzy pod diagonalą i na diagonalu. Macierze $H^{(i)}$ są macierzami Householdera postaci $I - \frac{1}{\tau^{(i)}} \mathbf{u}^{(i)} \langle \mathbf{u}^{(i)}$,

więc aby obliczyć $H^{(i)} A^{(i)} H^{(i)}$ wystarczy wyznaczyć wektor $\mathbf{w}^{(i)} = \frac{1}{\tau^{(i)}} A^{(i)} \mathbf{u}^{(i)}$

oraz $\mathbf{v}^{(i)} = \mathbf{w}^{(i)} - \frac{1}{2\tau^{(i)}} \langle \mathbf{u}^{(i)}, \mathbf{w}^{(i)} \rangle \cdot \mathbf{u}^{(i)}$. Wówczas

$$H^{(i)} A^{(i)} H^{(i)} = A^{(i)} - \mathbf{u}^{(i)} \langle \mathbf{v}^{(i)} - \mathbf{v}^{(i)} \rangle \langle \mathbf{u}^{(i)} \quad (6.31)$$

Liczba działań potrzebnych do obliczenia macierzy T tą metodą wynosi $M =$

$= \frac{2}{3}n^3 + O_1(n^2)$ mnożeń, $D = \frac{2}{3}n^3 + O_2(n^2)$ dodawań i $(n-2)$ pierwiastkowań.

Analiza błędów zaokrągleń. Metoda Householdera umożliwia wyznaczenie macierzy T takiej, że

$$\|T - \hat{T}\|_E \leq 2k_1 n^2 \varepsilon (1 + k_1 \varepsilon n)^{2n} \|A\|_E \quad (6.32)$$

gdzie $\hat{T}$ oznacza dokładną postać trójdagonalną macierzy A , a stała k_1 zależy od sposobu zaokrąglania i dla typowej maszyny cyfrowej wynosi 20 (patrz [65], rozdz. 5).

Metoda Givensa. Postępujemy analogicznie jak w metodzie Householdera uzyskując ciąg macierzy $M^{(2)}, M^{(3)}, \dots, M^{(n-1)} = T$

Nakład obliczeń w tym przypadku wynosi: $M = \frac{4}{3}n^3 + O_1(n^2)$ mnożeń, $D = \frac{2}{3}n^3 + O_2(n^2)$ dodawań i $\frac{1}{2}n(n-1)$ pierwiastkowań.

Oszacowanie błędu wyznaczenia macierzy T :

$$\|T - \hat{T}\|_E \leq 12\varepsilon n^{3/2} (1 + 6\varepsilon)^{4n-7} \|A\|_E \quad (6.33)$$

U w a g a. Jakkolwiek ocena (6.33) jest lepsza niż (6.32), gdyż w (6.33) występuje mnożnik $n^{3/2}$, a w (6.32) mnożnik n^2 ; praktyczne obliczenia nie wskazują na to, by dokładność którejś z tych metod była lepsza. Co więcej, testowanie na dużej liczbie przykładów (patrz [65], rozdz. 5) wykazało, że oba oszacowania są znacznie zawyżone i w praktyce błąd jest rzędu $\varepsilon n^{1/2}$.

6.6. MACIERZE WSTĘGOWE

Macierze wstęgowe omawialiśmy w p. 5.5 jako przykład macierzy rzadkich. Z przytoczonego tam twierdzenia 5 wynika, że istnieje możliwość oszczędnego operowania pamięcią maszyny przy rozwiązywaniu układów równań z macierzą wstęgową, gdyż można wówczas uzyskać rozkłady LU lub LL^T takie, w których macierze trójkątne L i U są wstęgowe. Zatem taki algorytm LR dla macierzy wstęgowych zachowuje postać macierzy wstęgowej, gdyż łatwo sprawdzić, że iloczyn RL będzie nadal macierzą wstęgową. Bez stosowania przestawień wierszy algorytm LR bywa jednak niestabilny (patrz [65], rozdz. 8), lecz gdy macierz A jest symetryczna i dodatnio określona, wówczas stosując rozkład Banachiewicza uzyskamy zbieżny algorytm LR .

Z twierdzenia 3 z p. 6.5 wnioskujemy, że jeśli do macierzy wstęgowej zastosujemy algorytm QR , to uzyskiwane macierze będą również wstęgowe. Można także uzasadnić, że dla metody QR wystarczy założenie jedynie symetrii macierzy wstęgowej, aby zachować strukturę macierzy (patrz [65]). Zwracamy jednak uwagę na fakt, że macierz R uzyskiwana w rozkładzie QR macierzy wstęgowej A o liczbie m (tzn. o szerokości wstęgi $2m+1$) jest co

prawda macierzą wstęgową, ale ma $2m$ diagonalni ponad główną diagonalą. Na przykład w rozkładzie QR macierzy trójdagonalnej macierz R ma łącznie trzy diagonale (elementy ich oznaczyliśmy literami r, p, q). Do wyznaczenia iloczynu RQ wystarczyły jednak wartości elementów z dwu diagonalni. Podobnie jest w przypadku macierzy wstęgowej — do wyznaczenia iloczynu RQ wystarczą wartości elementów $m+1$ diagonalni, licząc od diagonalni głównej. Można zatem opracować wzory analogiczne do (6.30), dające oszczędność pamięci i zmniejszające liczbę działań arytmetycznych.

Jeżeli $m > 1$, to mniejszą liczbę działań arytmetycznych uzyskamy dla metody Householdera niż dla metody Givensa.

Zamiast macierzy Q należy przy tym zapamiętać odpowiednie macierze Householdera lub Givensa, których iloczyn jest równy Q . W przypadku macierzy Householdera $H^{(1)}, H^{(2)}, \dots, H^{(n-1)}$ do zapamiętania ich wystarczy $n(m+2) - \frac{m}{2}(m+1) - 2$ komórek. Macierz Q składa się na ogół z macierzy trójkątnej i z dodatkowych m diagonalni nad diagonalą główną, ma więc $\frac{1}{2}n^2 + \frac{1}{2}n(1+2m) - m^2 - m$ niezerowych elementów.

METODY ROZWIĄZYWANIA ZAGADNIEŃ POCZĄTKOWYCH DLA RÓWNAŃ RÓŻNICZKOWYCH ZWYCZAJNYCH

W rozdziale tym zajmiemy się rozwiązywaniem następującego zagadnienia początkowego. Znaleźć funkcje $y_i = y_i(x)$, $i = 1, \dots, m$, spełniające na przedziale $\langle x_0; b \rangle$ układ równań różniczkowych

i warunek początkowy

Zagadnienie powyższe możemy zapisać krótko w postaci wektorowej

gdzie

W przypadku $m = 1$ zagadnienie (7.3) przyjmuje postać jednego równania skalarnego

$$\frac{dy}{dx} = f(x, y), \quad y(x_0) = y_0 \quad (7.4)$$

Niech $\|\cdot\|$ oznacza normę w przestrzeni R^m , określoną wzorem (5.1a), tzn.

$$\|\mathbf{v}\| = |v_1| + \dots + |v_m|$$

gdzie $\mathbf{v} = [v_1, \dots, v_m]^T$.

Zakładamy, że

- 1) Funkcja $f(x, y)$ jest ciągła na zbiorze $D = \{(x, y) : x_0 \leq x \leq b, y \in R^m\}$
- 2) Funkcja $f(x, y)$ spełnia na zbiorze D warunek Lipschitza względem zmiennej y , tzn. istnieje stała L taka, że dla każdego $x \in \langle x_0; b \rangle$ i dowolnych $y, \tilde{y} \in R^m$ zachodzi nierówność

$$\|f(x, y) - f(x, \tilde{y})\| \leq L\|y - \tilde{y}\| \quad (7.5)$$

Przy powyższych założeniach można udowodnić, że na przedziale $\langle x_0; b \rangle$ istnieje dokładnie jedno rozwiązanie klasy C^1 zagadnienia (7.3).

Zauważmy, że nierówność (7.5) zachodzi, gdy funkcje $f_i(x, y_1, \dots, y_m)$ w każdym punkcie obszaru D mają pochodne cząstkowe $\frac{\partial f_i}{\partial y_j}$, $i, j = 1, \dots, m$, ograniczone na D .

W rozdziale tym omówimy tzw. *metody dyskretne* rozwiązywania zagadnienia (7.3), tzn. metody za pomocą których otrzymujemy przybliżone rozwiązania tylko dla dyskretnych wartości x_i ($i = 1, \dots, N$) zmiennej niezależnej x , przy czym ograniczymy się do omówienia najważniejszych z tych metod: *metod różnicowych* i *metod typu Rungego-Kutty*. Charakteryzują się one tym, że poszukiwane rozwiązanie y_{n+1} w punkcie x_{n+1} otrzymujemy za pomocą wcześniej obliczonych kilku wartości y_{n-j} , $j = 0, 1, \dots$, oraz wartości f w pewnych punktach.

Wprowadzimy teraz kilka oznaczeń, z których będziemy dalej korzystać. Niech $Y(x)$ oznacza dokładne rozwiązanie zagadnienia (7.3), a $y(x)$ rozwiązanie przybliżone. Ponadto wprowadzamy oznaczenia:

$$Y_i = Y(x_i), \quad Y'_i = \frac{dY(x_i)}{dx} = f(x_i, Y_i)$$

$$y_i = y(x_i), \quad y'_i = f(x_i, y_i) = f_i$$

gdzie $x_i \in \langle x_0, b \rangle$, $i = 1, 2, \dots, N$, są punktami, w których wyznaczamy przybliżone rozwiązania. Zauważmy, że wektor-funkcja y jest określona tylko w punktach x_i i oznaczenie $f(x_i, y_i)$ symbolem y'_i jest jedynie umowne.

W przypadku jednego równania różniczkowego ($m = 1$) będziemy używać tych samych oznaczeń tylko że nie wytłuszczonych.

Wyprowadzając wzory zarówno w metodach różnicowych, jak i w metodach typu Rungego-Kutty, będziemy zakładać, że punkty x_i są równoodległe:

$x_i = x_0 + ih$, gdzie h jest *krokiem całkowania*. W praktyce, podczas obliczeń można zmieniać krok całkowania, co omówimy dokładniej w p. 7.4.4 i 7.5.1.

Jak już wspomnieliśmy, rozwiązanie y_{n+1} wyznaczać będziemy za pomocą obliczonych wcześniej wartości $y_n, y_{n-1}, \dots, y_{n-k}$ ($k \geq 0$). Fakt ten zapiszemy w sposób następujący

$$A_n(h, y_{n-k}, \dots, y_n; y_{n+1}) = 0 \quad (7.6)$$

przy czym w omawianych metodach równanie (7.6) może być liniowe, jak też i nieliniowe względem niewiadomej y_{n+1} .

Po podstawieniu dokładnego rozwiązania Y_i do (7.6) otrzymamy równanie

$$A_n(h, Y_{n-k}, \dots, Y_n; Y_{n+1}) = T_n \quad (7.7)$$

Wielkość T_n nazywamy *błędem metody* powstałym przy przejściu od x_n do x_{n+1} .

Jeżeli błąd T_n jako funkcję zmiennej h można przedstawić w postaci

$$T_n = h^{p+1}\gamma_p + O(h^{p+2})^1 \quad (7.8)$$

gdzie stała $\gamma_p \neq 0$, to liczbę p będziemy nazywać *rzędem dokładności* lub, po prostu, *rzędem metody przybliżonej*.

W dalszych rozważaniach będziemy zakładać, nie zaznaczając tego za każdym razem, że dla rozpatrywanego zagadnienia początkowego (7.1), (7.2) są spełnione podane powyżej 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*²⁾, według której

$$y_{n+1} = y_n + hf(x_n, y_n), \quad n \geq 1 \quad (7.9)$$

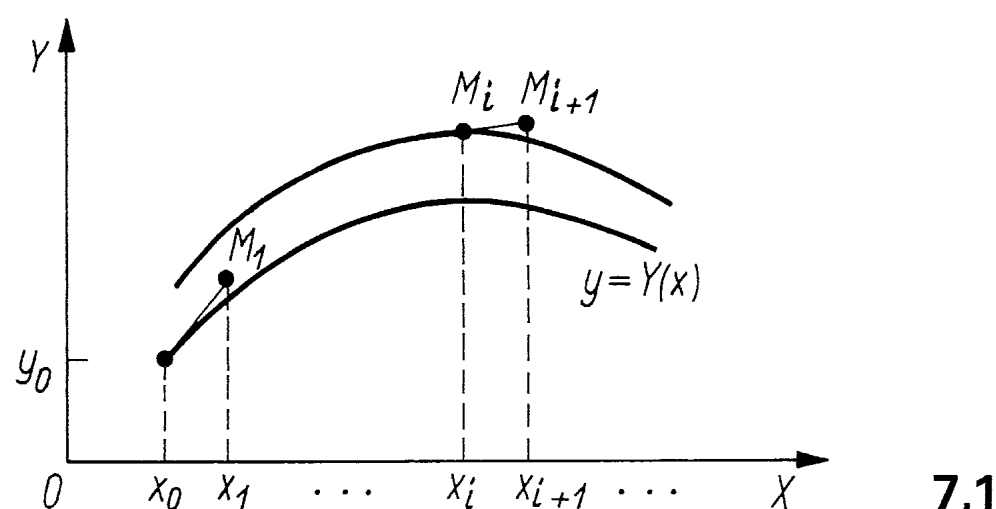
przy czym y_0 jest określone warunkiem początkowym podanym w (7.3). Ze względu na swoją prostotę metoda Eulera jest łatwa do analizy. Ograniczymy się tu do rozpatrzenia przypadku $m = 1$, czyli do zagadnienia (7.4).

Metoda Eulera ma prostą interpretację geometryczną, mianowicie odcinek $M_i M_{i+1}$ ma w punkcie $M_i = M_i(x_i, y_i)$ kierunek zgodny z kierunkiem

¹⁾ Symbol $O(h^{p+2})$ oznacza wektor m -wymiarowy, którego każda składowa jest postaci $O(h^{p+2})$.

²⁾ Metoda ta jest szczególnym przypadkiem zarówno metod różnicowych, jak i metod typu Rungego-Kutty.

stycznej do krzywej całkowej równania $y' = f(x, y)$, która przechodzi przez punkt M_i (rys. 7.1); z tego też powodu metoda Eulera jest nazywana często *metodą stycznych*.



Założmy, że $Y(x) \in C^2(\langle x_0; b \rangle)$. Rozwiązanie dokładne zagadnienia (7.4) spełnia równanie

$$Y_{n+1} = Y_n + hf(x_n, Y_n) + T_n \quad (7.10)$$

gdzie $T_n = \frac{1}{2}h^2 Y''(\xi_n)$, $\xi_n \in (x_n; x_{n+1})$.

Oszacowanie błędu $\varepsilon_n = Y_n - y_n$ metody Eulera.

Odejmując stronami (7.9) od (7.10), otrzymujemy

$$\varepsilon_{n+1} = \varepsilon_n + h[f(x_n, Y_n) - f(x_n, y_n)] + T_n \quad (7.11)$$

Stąd, korzystając z warunku Lipschitza (7.5), mamy

$$|\varepsilon_{n+1}| \leq (1 + hL)|\varepsilon_n| + |T_n|$$

Szacując następnie

$$|T_n| \leq \frac{1}{2}h^2 M_2 = T, \quad M_2 = \max_{\langle x_0; b \rangle} |Y''(x)|$$

znajdujemy

$$|\varepsilon_{n+1}| \leq (1 + hL)|\varepsilon_n| + T \quad (7.12)$$

Jest to rekurencyjne oszacowanie błędu $|\varepsilon_{n+1}|$ w zależności od $|\varepsilon_n|$. Powyższą nierówność zapiszemy w postaci

$$|\varepsilon_{n+1}| \leq a|\varepsilon_n| + b, \quad a > 1 \quad (7.13)$$

Mamy więc

$$|\varepsilon_1| \leq a|\varepsilon_0| + b, \quad |\varepsilon_2| \leq a^2|\varepsilon_0| + b(1 + a)$$

i ogólnie, co łatwo jest udowodnić za pomocą indukcji

$$|\varepsilon_n| \leq b \frac{a^n - 1}{a - 1} + a^n |\varepsilon_0|$$

Wracając do poprzednich oznaczeń, otrzymujemy

$$|\varepsilon_n| \leq T \frac{(1 + hL)^n - 1}{hL} + (1 + hL)^n |\varepsilon_0| \quad (7.14)$$

stąd, zaś na podstawie nierówności $1 + u < e^u$ dla $u > 0$, mamy

$$|\varepsilon_n| \leq \frac{T}{Lh} (e^{L(x_n - x_0)} - 1) + e^{L(x_n - x_0)} |\varepsilon_0|, \quad nh = x_n - x_0 \quad (7.15)$$

Podstawiając do (7.15) $\varepsilon_0 = 0$, $T = \frac{1}{2} h^2 M_2$, otrzymujemy dla metody Eulera następujące oszacowanie błędu

$$|\varepsilon_n| \leq \frac{1}{2} \frac{M_2}{L} h (e^{L(x_n - x_0)} - 1) \quad (7.16)$$

Z oszacowania (7.16) wynika, że dla $h \rightarrow 0$ w ustalonym punkcie x ($x - x_0 = nh$, $h \rightarrow 0$, $n \rightarrow \infty$) rozwiązanie y_n jest zbieżne do wartości dokładnej $Y(x)$ i szybkość tej zbieżności wynosi $O(h)$.

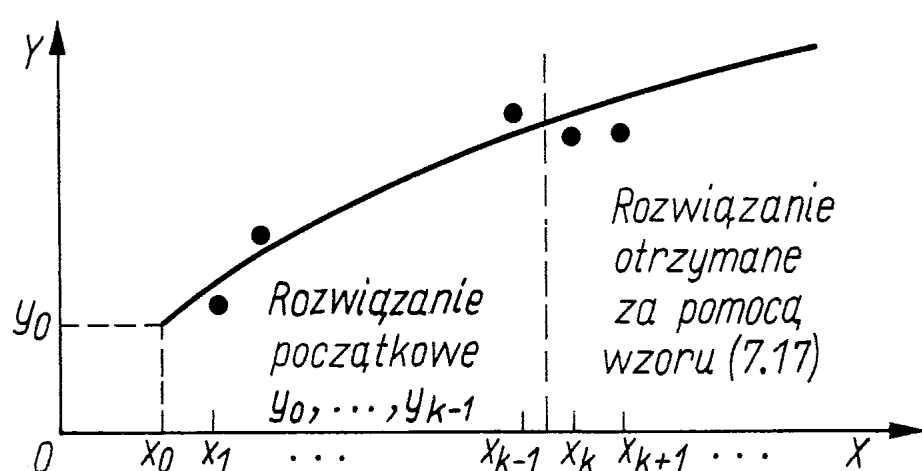
7.3. METODY RÓŻNICOWE

7.3.1. Ogólny wzór różnicowy

Metodę określoną wzorem

$$y_{n+1} = \sum_{i=1}^k a_i y_{n+1-i} + h \sum_{i=0}^k b_i f(x_{n+1-i}, y_{n+1-i}), \quad n \geq k-1 \quad (7.17)$$

gdzie współczynniki a_i , b_i są liczbami rzeczywistymi, a h jest krokiem całkowania nazywamy *metodą różnicową*. Zakładamy, że $|a_k| + |b_k| \neq 0$.



7.2

Jeżeli $k > 1$, to powyższą metodę możemy stosować do rozwiązywania zagadnienia (7.1), (7.2) tylko wtedy, gdy znamy wartości rozwiązania w punktach $x_1, \dots, x_{k-1}$ (rys. 7.2), bowiem warunek początkowy (7.2) daje nam rozwiązanie tylko w punkcie x_0 . Brakujące wartości rozwiązania wyznaczamy najczęściej za pomocą innych metod, które omówimy w p. 7.5.

Jeżeli $b_0 = 0$, to wzór (7.17) nazywamy *wzorem ekstrapolacyjnym (jawnym)*. Szukana wartość y_{n+1} , $n \geq k-1$, jest wówczas kombinacją liniową obliczonych wcześniej wartości $y_{n+1-k}, \dots, y_n$ oraz $f_{n+1-k}, \dots, f_n$, a więc wyznaczenie rozwiązania y_{n+1} jest łatwe.

Jeżeli $b_0 \neq 0$, to wzór (7.17) nazywamy *wzorem interpolacyjnym (uwikłanym)*. Szukana wartość y_{n+1} występuje po obu stronach (7.17). Gdy układ (7.1) jest nieliniowy (a tak jest najczęściej), to (7.17) jest nieliniowym układem ze względu na szukane y_{n+1} . Można udowodnić, że dla h spełniających nierówność $h|b_0|L < 1$ układ ten ma dokładnie jedno rozwiązanie y_{n+1} (dla dowolnych $y_n, y_{n-1}, \dots, y_{n+1-k}$). W dalszych rozważaniach będziemy zakładać, że krok całkowania h spełnia ten warunek.

Podamy teraz warunki na współczynniki a_i, b_i , przy których metoda różnicowa określona wzorem (7.17) jest rzędu p . Jeżeli w wyrażeniu

$$Y_{n+1} - \sum_{i=1}^k a_i Y_{n+1-i} - h \sum_{i=0}^k b_i Y'_{n+1-i} \quad (7.18)$$

rozwinimy Y_{n+1-i} według wzoru Taylora w otoczeniu punktu x_{n+1-k} , to wielkości A_s , określone w następujący sposób

$$A_0 = 1 - \sum_{i=1}^k a_i \quad (7.19)$$

$$A_s = \frac{k^s}{s!} - \sum_{i=1}^k \frac{(k-i)^s}{s!} a_i - \sum_{i=0}^k \frac{(k-i)^{s-1}}{(s-1)!} b_i, \quad s \geq 1, \quad 0^0 = 1$$

są współczynnikami przy potęgach h^s , po uporządkowaniu rozwinięcia względem potęg h . Jeżeli zatem

$$A_i = 0, \quad i = 0, 1, \dots, p, \quad A_{p+1} \neq 0 \quad (7.20)$$

to metoda różnicowa określona wzorem (7.17) jest rzędu p . Mówiąc, że metoda różnicowa jest rzędu p , będziemy rozumieli, że warunki te są spełnione.

Ponadto zauważmy, że jeżeli są spełnione warunki (7.20), to dla dostatecznie regularnego rozwiązania $Y(x)$ błąd (7.8) metody różnicowej można przedstawić w postaci

$$T_n = A_{p+1} h^{p+1} Y_{n+1-k}^{(p+1)} + O(h^{p+2}) \quad (7.21)$$

Przykład 1. Określić, dla jakich wartości a_1, b_0, b_1 metoda określona wzorem

$$y_{n+1} = a_1 y_n + h(b_0 y'_{n+1} + b_1 y'_n)$$

jest rzędu drugiego.

Warunki (7.20) mają postać

$$A_0 = 1 - a_1 = 0, \quad A_1 = 1 - b_0 - b_1 = 0, \quad A_2 = \frac{1}{2} - b_0 = 0$$

$$A_3 = \frac{1}{6} - \frac{1}{2} b_0 \neq 0$$

Zatem szukany wzór ma postać

$$y_{n+1} = y_n + \frac{1}{2}h(y'_{n+1} + y'_n)$$

Niech wektory $y_i = y_i(h)$, $i = 0, 1, \dots, k-1$, użyte do wyznaczenia rozwiązania y za pomocą wzoru (7.17) będą takie, że

$$\lim_{h \rightarrow 0} y_i(h) = y_0, \quad i = 0, 1, \dots, k-1 \quad (7.22)$$

gdzie y_0 jest wektorem początkowym (patrz wzór (7.3)).

Def. Metodę różnicową nazywamy *zbieżną*, jeżeli dla każdego zagadnienia początkowego (7.1), (7.2) rozwiązanie równania (7.17) jest takie, że

$$\lim_{\substack{h \rightarrow 0 \\ (n \rightarrow \infty)}} y_n = Y(x), \quad nh = x - x_0 \quad (7.23)$$

dla wszystkich $x \in \langle x_0; b \rangle$.

Warunek (7.22) jest oczywiście potrzebny, ponieważ wektory $y_1, \dots, y_{k-1}$ najczęściej wyznaczamy za pomocą innych metod i w związku z tym nie są one na ogół dokładne.

Tw. *Warunkiem koniecznym zbieżności metody różnicowej jest, aby*

$$\begin{aligned} A_0 &= 1 - \sum_{i=1}^k a_i = 0 \\ A_1 &= k - \sum_{i=1}^k (k-i) a_i - \sum_{i=0}^k b_i = 0 \end{aligned} \quad (7.24)$$

Metodę różnicową określoną wzorem (7.17), dla której warunki (7.24) są spełnione, nazywamy *zgodną* (z równaniem różniczkowym); wszystkie rozpatrywane w tym rozdziale metody są zgodne.

Warunek konieczny i wystarczający zbieżności metody różnicowej sformułujemy w p. 7.3.3.

7.3.2. Równanie dla błędów

W metodzie różnicowej na każdym pojedynczym kroku, tzn. przy przejściu od x_i do x_{i+1} , $i \geq k-1$, popełniamy dwa rodzaje błędów:

1) błąd metody powstały z zamiany równania różniczkowego równaniem różnicowym,

2) błędy powstałe przy numerycznym wyznaczaniu rozwiązania równania różnicowego (błędy zaokrągleń oraz błąd niedokładnego rozwiązania równania (7.17), w przypadku gdy jest to równanie nieliniowe).

Ponadto, aby rozpocząć obliczenia według wzoru (7.17) trzeba mieć wyznaczone przybliżone rozwiązanie w k początkowych punktach, podczas gdy w zagadnieniu (7.3) dane jest rozwiązanie tylko w punkcie x_0 . Naturalne

jest więc założenie, że wielkości $y_1, \dots, y_{k-1}$ nie są dokładne. Te wszystkie błędy mają oczywiście wpływ na to, jak obliczone rozwiązanie różnicowe $\tilde{y}_n$ przybliży rozwiązanie dokładne Y_n na całym przedziale. Przeprowadzimy zatem dyskusję błędu całkowitego, tzn. różnicy

$$\epsilon_n = Y_n - \tilde{y}_n \quad (7.25)$$

Rozwiązania $Y_n, \tilde{y}_n$ spełniają odpowiednio równania

$$Y_{n+1} = \sum_{i=1}^k a_i Y_{n+1-i} + h \sum_{i=0}^k b_i f(x_{n+1-i}, Y_{n+1-i}) + T_n \quad (7.26)$$

$$\tilde{y}_{n+1} = \sum_{i=1}^k a_i \tilde{y}_{n+1-i} + h \sum_{i=0}^k b_i f(x_{n+1-i}, \tilde{y}_{n+1-i}) - \delta_n \quad n \geq k-1 \quad (7.27)$$

gdzie δ_n jest błędem związanym z niedokładnym wyznaczaniem rozwiązania różnicowego na maszynie cyfrowej.

Odejmując stronami równanie (7.27) od (7.26) otrzymujemy

$$\epsilon_{n+1} = \sum_{i=1}^k a_i \epsilon_{n+1-i} + h \sum_{i=0}^k b_i \eta_{n+1-i} + B_n, \quad n \geq k-1 \quad (7.28)$$

przy czym $\eta_{n+1-i} = f(x_{n+1-i}, Y_{n+1-i}) - f(x_{n+1-i}, \tilde{y}_{n+1-i})$, a $B_n = T_n + \delta_n$.

Zakładając dodatkowo, że funkcje f_i , które są składowymi wektora f , posiadają ciągłe pochodne $\frac{\partial f_i}{\partial y_j}$ na obszarze D na podstawie twierdzenia o wartości średniej mamy

$$\eta_{n+1-i} = f_y(x_{n+1-i}; \tilde{y}_{n+1-i}; \epsilon_{n+1-i}; \theta_{n+1-i}) \epsilon_{n+1-i} \quad (7.29)$$

gdzie $f_y(x; \tilde{y}; \epsilon; \theta) = \left[\frac{\partial f_i}{\partial y_j}(x, \tilde{y} + \theta_i \epsilon) \right]_{i,j=1,\dots,m}$, $\theta = (\theta_1, \dots, \theta_m)$, $\theta_i \in (0; 1)$.

Po podstawieniu (7.29) do (7.28) ostatecznie otrzymujemy *równanie dla błędu całkowitego* ϵ_n , postaci

$$\begin{aligned} [I - hb_0 f_y(x_{n+1}; \tilde{y}_{n+1}; \epsilon_{n+1}; \theta_{n+1})] \epsilon_{n+1} = \\ = \sum_{i=1}^k [a_i I + hb_i f_y(x_{n+1-i}; \tilde{y}_{n+1-i}; \epsilon_{n+1-i}; \theta_{n+1-i})] \epsilon_{n+1-i} + B_n, \quad n \geq k-1 \end{aligned} \quad (7.30)$$

We wzorze tym $\epsilon_i = Y_i - y_i$, $i \leq k-1$, są błędami warunków początkowych, a I oznacza macierz jednostkową stopnia m .

W ogólnym przypadku wielkości $f_y(x_{n+1-i}; \tilde{y}_{n+1-i}; \epsilon_{n+1-i}; \theta_{n+1-i})$, T_n oraz δ_n zależą od n i wtedy równanie (7.30) nie jest łatwe do analizy. Dlatego też są konieczne pewne upraszczające założenia. Ponieważ w praktyce błąd B_n zmienia się nieznacznie przy zmianie n , więc dodatkowo założymy, że błąd ten nie zależy od n , i przyjmiemy, że $B_n = B$. Ponadto założymy, że

$$\Lambda = f_y = \text{const} \quad (7.31)$$

Warunek (7.31) stałości pochodnych $\frac{\partial f_i}{\partial y_i}$ jest równoważny temu, że prawą stronę układu równań różniczkowych (7.1) można przedstawić w postaci

$$f(x, y) = \Lambda y + g(x)$$

a ponieważ błąd całkowity ϵ_n określony równaniem (7.30) nie zależy od $g(x)$ więc warunek ten oznacza w istocie, że badamy błąd całkowity dla liniowego układu równań różniczkowych o stałych współczynnikach

$$y' = \Lambda y \quad (7.32)$$

przy założeniu, że błąd $B_n = B$ jest stały.

Przy powyższych założeniach równanie (7.30) możemy zapisać w postaci

$$[I - hb_0 \Lambda] \epsilon_{n+1} = \sum_{i=1}^k [a_i I + hb_i \Lambda] \epsilon_{n+1-i} + B \quad n \geq k-1 \quad (7.33)$$

Dalszą naszą dyskusję ograniczymy do omówienia rozwiązań równania postaci (7.33), które stanowi liniowy układ równań różnicowych o stałych współczynnikach. Odpowiadający temu układowi równań układ jednorodny ma postać

$$[I - hb_0 \Lambda] \epsilon_{n+1} = \sum_{i=1}^k [a_i I + hb_i \Lambda] \epsilon_{n+1-i} \quad (7.34)$$

Podamy teraz niektóre własności rozwiązań układów (7.33) i (7.34).

1. Rozwiązanie ogólne (tzn. każde rozwiązanie) ϵ_n układu równań (7.33) można przedstawić w postaci sumy

$$\epsilon_n = \epsilon_n^1 + \epsilon_n^2 \quad (7.35)$$

rozwiązania ogólnego ϵ_n^1 układu (7.34) i jakiegokolwiek rozwiązania układu (7.33).

2. Dowolna kombinacja liniowa rozwiązań układu (7.34) jest także rozwiązaniem układu (7.34).

3. Jeżeli $(\sum_{i=0}^k b_i) \det \Lambda \neq 0$ (zakładamy, że jest spełniony ten warunek),

to wśród rozwiązań układu (7.33) znajduje się rozwiązanie stałe

$$\epsilon = \epsilon_n^2 = \frac{-1}{h \sum_{i=0}^k b_i} \Lambda^{-1} B$$

Zajmiemy się z kolei wyznaczeniem rozwiązania ogólnego ϵ_n^1 układu (7.34) w przypadku $m = 1$ (jedno równanie). Przyjmując $\lambda = \Lambda$ otrzymujemy równanie

$$(1 - h\lambda b_0)\varepsilon_{n+1} = \sum_{i=1}^k (a_i + h\lambda b_i)\varepsilon_{n+1-i} \quad (7.36)$$

Rozwiązania równania (7.36) poszukujemy w postaci

$$\varepsilon_n = \mu^n \quad (7.37)$$

gdzie μ jest stałą liczbą. Po podstawieniu przewidywanego rozwiązania (7.37) do (7.36) stwierdzamy, że jest ono rozwiązaniem wtedy i tylko wtedy, gdy liczba μ jest pierwiastkiem równania algebraicznego stopnia k

$$(1 - h\lambda b_0)\mu^k = \sum_{i=1}^k (a_i + h\lambda b_i)\mu^{k-i} \quad (7.38)$$

Równanie (7.38) nazywamy *równaniem charakterystycznym dla równania* (7.36). Jego pierwiastkami mogą być zarówno liczby rzeczywiste, jak i zespolone, przy czym mogą być wśród nich pierwiastki wielokrotne. Niech $\mu_1, \dots, \mu_k$ oznaczają pierwiastki równania (7.38).

a) Pierwiastki $\mu_1, \dots, \mu_k$ są liczbami rzeczywistymi. Jeżeli wszystkie te pierwiastki są jednokrotne, to $\mu_1^n, \dots, \mu_k^n$ są rozwiązaniami równania (7.36) i rozwiązanie ogólne tego równania można przedstawić w postaci

$$\varepsilon_n^1 = c_1 \mu_1^n + \dots + c_k \mu_k^n \quad (7.39)$$

gdzie c_i są stałymi.

Rozpatrzmy teraz przypadek, gdy wśród pierwiastków $\mu_1, \dots, \mu_k$ znajdują się pierwiastki wielokrotne. Niech na przykład początkowe pierwiastki $\mu_1, \dots, \mu_s$ będą równe liczbie μ_0 , pozostałe zaś pierwiastki $\mu_{s+1}, \dots, \mu_k$, z których żaden nie równa się liczbie μ_0 , niech będą jednokrotne. Wówczas $\mu_0^n, n\mu_0^n, \dots, n^{s-1}\mu_0^n, \mu_{s+1}^n, \dots, \mu_k^n$ są rozwiązaniami równania (7.36) i jego rozwiązanie ogólne ma postać

$$\varepsilon_n^1 = (c_1 + c_2 n + \dots + c_s n^{s-1})\mu_0^n + c_{s+1}\mu_{s+1}^n + \dots + c_k \mu_k^n \quad (7.40)$$

Podobnie wyznacza się rozwiązanie ogólne równania (7.36) w przypadku, gdy równanie charakterystyczne (7.38) ma więcej pierwiastków wielokrotnych.

b) Wśród pierwiastków $\mu_1, \dots, \mu_k$ znajdują się pierwiastki zespolone. Niech na przykład liczba μ_1 będzie takim pierwiastkiem. Wówczas liczba np. μ_2 sprzężona z μ_1 też jest pierwiastkiem. Zapišemy te pierwiastki w postaci trygonometrycznej

$$\mu_{1,2} = |\mu_1|(\cos \theta \pm j \sin \theta)$$

Na podstawie wzoru Eulera mamy

$$\mu_{1,2}^n = |\mu_1|^n (\cos n\theta \pm j \sin n\theta)$$

a więc $|\mu_1|^n \cos n\theta, |\mu_1|^n \sin n\theta$ są rozwiązaniami równania (7.36). Jeżeli μ_1 jest pierwiastkiem jednokrotnym, a pozostałe pierwiastki są rzeczywiste i jednokrotne, to rozwiązanie ε_n^1 można przedstawić w postaci

$$\varepsilon_n^1 = (c_1 \cos n\theta + c_2 \sin n\theta) \cdot |\mu_1|^n + c_3 \mu_3^n + \dots + c_k \mu_k^n \quad (7.41)$$

Jeżeli wśród pierwiastków $\mu_1, \dots, \mu_k$ będzie więcej par pierwiastków zespolonych, to we wzorze (7.41) zwiększy się odpowiednio liczba składników, w których wystąpią funkcje trygonometryczne z tym, że składniki te będą takiej postaci jak w (7.40), tzn. pomnożone przez n^r jeżeli te pierwiastki będą wielokrotne.

Bez względu na to, jakie są pierwiastki μ_j równania charakterystycznego (7.38), rozwiązanie ε_n^1 , a tym samym też ε_n zależy od k stałych c_i . Stałe c_i są jednoznacznie określone przez błędy $\varepsilon_0, \varepsilon_1, \dots, \varepsilon_{k-1}$.

Rozpatrzone dla $m = 1$ rozwiązanie ogólne ε_n^1 ma bardzo podobną postać do rozwiązania ogólnego liniowego równania różniczkowego zwyczajnego jednorodnego rzędu k o stałych współczynnikach (patrz np. [74]). Pozostawiamy Czytelnikowi porównanie tych wzorów.

Zajmiemy się obecnie wyznaczeniem ε_n^1 w przypadku $m > 1$ (układ równań). Rozwiązania układu (7.34) poszukujemy w postaci

$$\varepsilon_n = \mu^n \mathbf{v} \quad (7.42)$$

gdzie μ jest liczbą stałą, a $\mathbf{v}$ stałym niezerowym wektorem. Po podstawieniu (7.42) do (7.34) stwierdzamy, że ε_n określone powyższym wzorem jest rozwiązaniem wtedy i tylko wtedy, gdy μ i $\mathbf{v}$ spełniają układ równań

$$\mathbf{M}(\mu) \mathbf{v} = \mathbf{0} \quad (7.43)$$

gdzie $\mathbf{M}(\mu)$ jest macierzą określoną w następujący sposób

$$\mathbf{M}(\mu) = [\mathbf{I} - hb_0 \mathbf{A}] \mu^k - \sum_{i=1}^k [a_i \mathbf{I} + hb_i \mathbf{A}] \mu^{k-i} \quad (7.44)$$

Zatem niezerowy wektor $\mathbf{v}$ istnieje wtedy i tylko wtedy, gdy μ jest pierwiastkiem równania

$$\det \mathbf{M}(\mu) = 0 \quad (7.45)$$

które jest *równaniem charakterystycznym dla układu* (7.34). Jest to równanie algebraiczne (względem μ) stopnia km .

Niech μ_j ($j = 1, 2, \dots, km$) oznaczają pierwiastki równania (7.45). Rozpatrzmy przypadek, gdy pierwiastki μ_j są rzeczywiste i jednokrotne. Przez $\mathbf{v}_j$ oznaczmy ustalony wektor spełniający równanie (7.43), gdy $\mu = \mu_j$ ($\mathbf{v}_j$ jest wektorem własnym macierzy $\mathbf{M}(\mu_j)$). W takiej sytuacji rozwiązanie ε_n^1 można przedstawić w postaci

$$\varepsilon_n^1 = \sum_{s=1}^{km} c_s \mu_s^n \mathbf{v}_s \quad (7.46)$$

gdzie c_s są stałymi.

Rozwiązania ε_n^1 gdy wśród pierwiastków μ_j znajdują się zespolone lub wielokrotne nie będziemy omawiać. Nadmieniamy tylko, że w ogólnym

przypadku rozwiązanie ε_n^1 może zawierać składniki podobne do tych, jakie występują w rozwiązaniu ogólnym dla $m = 1$ (patrz wzory (7.40), (7.41)). Ponadto ε_n^1 zawsze zależy od km stałych c_i i stałe te są jednoznacznie określone przez błędy $\varepsilon_0, \varepsilon_1, \dots, \varepsilon_{k-1}$.

W przypadku $m = 1$ pierwiastki μ_j zgodnie z (7.38) zależą od wielkości $h\lambda$, gdzie h — krok całkowania, $\lambda = \frac{\partial f}{\partial y}$. Gdy $m > 1$, wówczas μ_j zależą od macierzy $h\Lambda$. Wykażemy, że zależność ta wyraża się w ten sposób, że pierwiastki μ_j zależą w istocie od $h\lambda_j$, gdzie λ_j są wartościami własnymi macierzy Λ . Niech P oznacza macierz podobieństwa sprowadzającą macierz Λ do postaci kanonicznej Jordana (patrz tw. 11 w p. 6.2)

$$P^{-1}\Lambda P = \hat{\Lambda} \quad (7.47)$$

Mamy więc

$$\begin{aligned} \det M(\mu) &= \det (P^{-1}M(\mu)P) = \\ &= \det \left[(I - hb_0\hat{\Lambda})\mu^k - \sum_{i=1}^k (a_i I + hb_i\hat{\Lambda})\mu^{k-i} \right] \end{aligned} \quad (7.48)$$

Macierz $\hat{\Lambda}$ jest macierzą trójkątną i elementami na jej głównej przekątnej są wartości własne λ_j . Tak więc warunek $\det M(\mu) = 0$ jest równoważny temu, że μ jest pierwiastkiem co najmniej jednego z następujących równań algebraicznych

$$(1 - h\lambda_j b_0)\mu^k = \sum_{i=1}^k (a_i + h\lambda_j b_i)\mu^{k-i}, \quad j = 1, \dots, m \quad (7.49)$$

Zauważmy, że powyższe równania po podstawieniu $\lambda = \lambda_j$ są równaniami charakterystycznymi (7.38) dla $m = 1$, z tą jednak różnicą, że λ może być liczbą zespoloną ponieważ wartości własne λ_j mogą być zespolone.

Ćwiczenia

1. Dla liniowego równania różnicowego o stałych współrzędnych

$$y_{n+1} + ay_n + by_{n-1} = 0, \quad n \geq 1$$

udowodnić, że wzory (7.39)–(7.41) określają każde rozwiązanie tego równania.

7.3.3. Stabilność i zbieżność

Wzór (7.17) określający metodę różnicową zależy od stałych a_i, b_i , przy czym stałe te należy tak wybrać, aby rozwiązanie y_n miało pożądane własności. Z twierdzenia podanego w p. 7.3.1 wynika, że stałe a_i, b_i powinny spełniać warunki (7.24), tzn. metoda różnicowa powinna być zgodna. Zanim odpowiemy na pytanie, jakie jeszcze inne warunki powinny spełniać te stałe, rozpatrzmy następujący przykład.

Przykład. 1. Czy do rozwiązania zagadnienia

$$y' = \lambda y, \quad \lambda — \text{stała} \quad x \in \langle 0; 1 \rangle$$

$$y(0) = 1$$

można zastosować metodę różnicową określoną wzorem

$$y_{n+1} = -4y_n + 5y_{n-1} + h(4y'_n + 2y'_{n-1}) \quad (7.50)$$

Jak łatwo można sprawdzić, jest to metoda zgodna rzędu trzeciego.

Korzystając z danego równania różniczkowego, wzór (7.50) zapiszemy w postaci

$$y_{n+1} + 4(1 - h\lambda)y_n - (5 + 2h\lambda)y_{n-1} = 0 \quad (7.51)$$

Jest to liniowe równanie różnicowe o stałych współczynnikach, zatem rozwiązanie y_n spełnia równanie postaci (7.36). Na podstawie rozważań z poprzedniego punktu rozwiązaniem ogólnym równania (7.51) jest

$$y_n = c_1 \mu_1^n + c_2 \mu_2^n, \quad n \geq 0 \quad (7.52)$$

gdzie c_1, c_2 są stałymi, a μ_1, μ_2 — pierwiastkami równania

$$\mu^2 + 4(1 - h\lambda)\mu - (5 + 2h\lambda) = 0$$

Niech μ_2 oznacza większy co do modułu pierwiastek. Mamy

$$\lim_{h \rightarrow 0} \mu_2 = -5, \quad \lim_{h \rightarrow 0} \mu_1 = 1, \quad \lim_{\substack{h \rightarrow 0 \\ (h \rightarrow \infty)}} \mu_1^n = e^{\lambda x}, \quad nh = x \quad (7.53)$$

Ponieważ $y_0 = 1$, więc $c_1 = 1 - c_2$ i rozwiązanie y_n ma postać

$$y_n = (1 - c_2)\mu_1^n + c_2 \mu_2^n$$

Stała c_2 zależy od wyboru y_1 . Na przykład, dla $y_1 = e^{\lambda h}$ (wartość dokładnego rozwiązania) otrzymujemy $c_2 \neq 0$ oraz

$$\lim_{\substack{h \rightarrow 0 \\ (n \rightarrow \infty)}} |y_n| = \infty, \quad nh = x$$

dla każdego $x > 0$.

Zbieżność do rozwiązania dokładnego $Y(x) = e^{\lambda x}$ otrzymamy jedynie wtedy, gdy $c_2 = 0$. Oczywiście można tak dobrać y_1 ($y_1 = \mu_1$), że będzie spełniony ten warunek. Jednakże przy prowadzeniu obliczeń na maszynie cyfrowej z powodu błędów zaokrągleń człon $c_2 \mu_2^n$, zwany *rozwiązaniem pasożytniczym*, będzie występował w (7.52) i rozwiązanie to będzie rozbieżne.

7/1

n	Rozwiązanie dokładne $e^{-n/10}$	Rozwiązanie przybliżone $y_n = \lambda_1^n$	Rozwiązanie przybliżone y_n
0	1.	1.	1.
1	0.90483742	0.90483494	0.90483494
3	0.74081822	0.74081213	0.740811213
5	0.60653066	0.60652235	0.60652236
7	0.49658530	0.49657578	0.49657587
9	0.40656966	0.40655964	0.40656220
11	0.33287108	0.33286105	0.33293312
13	0.27253179	0.27252209	0.27455001
15	0.22313016	0.22312099	0.28018945

W tablicy 7/1 podano wyniki obliczeń wykonanych na maszynie Odra 1305 dla $\lambda = -1$, $h = \frac{1}{10}$. Ostatnia kolumna zawiera wartości y_n otrzymane za pomocą wzoru (7.50), gdy $y_0 = 1$, $y_1 = \mu_1$.

Z powyższego przykładu wynika, że na współczynniki a_i , b_i oprócz warunków postaci (7.24) należy nałożyć pewne dodatkowe warunki po to, aby wyeliminować niepożądane zjawisko, omówione wyżej.

Stabilność. Zbadamy teraz problem stabilności metod różnicowych, czyli zagadnienie, jak zachowuje się ε_n dla dużych n . W punkcie 7.3.1 zdefiniowaliśmy zbieżność w przypadku granicznym dla $h \rightarrow 0$. W konkretnych obliczeniach interesuje nas przede wszystkim to, jak zachowuje się ε_n dla ustalonych wartości h . Tym problemem zajmiemy się obecnie.

Jeżeli rozwiązanie dokładne $Y(x)$ maleje, to y_n będzie praktycznie przydatnym rozwiązaniem jedynie wtedy, gdy błąd ε_n nie będzie wzrastał, jeżeli natomiast rozwiązanie $Y(x)$ rośnie, to na ogół nie możemy żądać, aby błąd ε_n nie wzrastał. W tej sytuacji błąd nie powinien jednak wzrastać szybciej niż dokładne rozwiązanie. Ponieważ rozwiązanie y_n i błąd ε_n zależą od kroku całkowania h , istotnym zagadnieniem jest wyznaczenie przedziału zmienności h , przy którym y_n i ε_n spełniają powyższe ograniczenia. Przedział ten powinien być możliwie duży, można bowiem wtedy wykonywać obliczenia z możliwie dużym krokiem h (oczywiście w granicach utrzymania dokładności).

Równanie (7.30) określające ε_n , które wyprowadziliśmy w p. 7.3.2, w ogólnym przypadku, gdy funkcje f_i są dowolne oraz błąd B_n zależy od n , jest trudne do badania. Założymy zatem, jak poprzednio że błąd B_n jest stały i nie zależy od n , a macierz f_y zastąpimy stałą macierzą A . Wówczas równania określające ε_n są równaniami liniowymi o stałych współczynnikach. Rozpatrzmy oddzielnie przypadek $m = 1$ i $m > 1$.

a) $m = 1$. Zanim przejdziemy do badania ε_n , zbadamy rozwiązanie y_n otrzymane za pomocą wzoru (7.17), gdy równanie różniczkowe jest postaci

$$y' = \lambda y \quad (7.54)$$

Korzystając z tego równania wzór (7.17) możemy zapisać w następujący sposób

$$(1 - h\lambda b_0)y_{n+1} = \sum_{i=1}^k (a_i + h\lambda b_i)y_{n+1-i} \quad (7.55)$$

Równanie powyższe jest takiej samej postaci jak równanie (7.36), zatem rozwiązanie y_n będzie tej samej postaci co ε_n^1 , a więc będzie zależeć od pierwiastków równania charakterystycznego (7.38) (patrz wzory (7.39)–(7.41)).

W punkcie 7.3.1 zaznaczyliśmy, że wszystkie rozpatrywane w tym rozdziale metody różnicowe są zgodne (na podstawie podanego tam twierdzenia jest to warunek konieczny zbieżności). Zatem dla $h = 0$ równanie charakterystyczne (7.38) ma pierwiastek równy 1. Dla $h \neq 0$ niech $\mu_1 = \mu_1(h)$ oznacza ten pierwiastek równania (7.38), dla którego $\lim_{h \rightarrow 0} \mu_1(h) = 1$.

Założmy, że wszystkie pierwiastki μ_j równania (7.38) są jednokrotne i rzeczywiste. Zgodnie ze wzorem (7.39) rozwiązanie y_n ma wówczas postać

$$y_n = d_1 \mu_1^n + \dots + d_k \mu_k^n \quad (7.56)$$

gdzie d_i są stałymi. Jeżeli są spełnione warunki (7.20) (metoda różnicowa jest rzędu p), to pierwiastek $\mu_1 = \mu_1(h)$ będzie postaci (patrz ćwic. 1)

$$\mu_1 = e^{\lambda h} + O(h^{p+1}) \quad (7.57)$$

Mamy więc

$$\mu_1^n = (e^{\lambda h} + O(h^{p+1}))^n = e^{nh\lambda} + O(h^{p+1}) = e^{\lambda(x_n - x_0)} + O(h^{p+1}) \quad (7.58)$$

Współczynniki d_i we wzorze (7.56) są określone przez k warunków początkowych, potrzebnych do wyznaczenia rozwiązania równania (7.55). Zatem

$$d_1 \mu_1^i + \dots + d_k \mu_k^i = y_i, \quad i = 0, 1, \dots, k-1 \quad (7.59)$$

Z warunku początkowego $y(x_0) = y_0$ dla zagadnienia początkowego (7.4) dane mamy y_0 . Pozostałe wartości $y_1, \dots, y_{k-1}$ wyznaczamy za pomocą innych metod i w związku z tym są one obarczone błędami, ale będziemy zakładać, że jest spełniony warunek (7.22), tzn. $\lim_{h \rightarrow 0} y_i = y_0, i = 1, \dots, k-1$. Zatem

(7.59) jest układem k równań z niewiadomymi d_i . Na podstawie wzorów Cramera mamy

$$d_1 = d_1(h) = \frac{\begin{vmatrix} y_0 & 1 & \dots & 1 \\ y_1 & \mu_2 & \dots & \mu_k \\ \dots & \dots & \dots & \dots \\ y_{k-1} & \mu_2^{k-1} & \dots & \mu_k^{k-1} \end{vmatrix}}{\begin{vmatrix} 1 & 1 & \dots & 1 \\ \mu_1 & \mu_2 & \dots & \mu_k \\ \dots & \dots & \dots & \dots \\ \mu_1^{k-1} & \mu_2^{k-1} & \dots & \mu_k^{k-1} \end{vmatrix}} \quad (7.60)$$

Przechodząc do granicy, gdy $h \rightarrow 0$, pierwsze kolumny licznika i mianownika dążą do kolumn stałych, o elementach równych odpowiednio y_0 i 1. Stąd otrzymujemy

$$\lim_{h \rightarrow 0} d_1(h) = y_0 \quad (7.61)$$

Analogicznie dowodzimy, że

$$\lim_{h \rightarrow 0} d_i(h) = 0 \quad \text{dla} \quad j = 2, 3, \dots, k. \quad (7.62)$$

Zatem dla $h \rightarrow 0$ pierwszy składnik $d_1 \mu_1^n$ w (7.56) dąży do rozwiązania

dokładnego $Y(x) = y_0 e^{\lambda(x-x_0)}$ równania (7.54) z warunkiem początkowym $y(x_0) = y_0$. Pierwiastki $\mu_2, \dots, \mu_k$ nazywamy *pierwiastkami pasożytniczymi*, a sumę $\sum_{i=2}^k d_i \mu_i^n$ — *częścią pasożytniczą* rozwiązania.

Zauważmy, że dla $\lambda > 0$ i dostatecznie małych h mamy

$$\mu_1 = \mu_1(h) > 1 \quad (7.63)$$

Przypominamy, że błąd ε_n ma postać

$$\varepsilon_n = c_1 \mu_1^n + \dots + c_k \mu_k^n - \frac{B}{h\lambda \sum_{i=0}^k b_i} \quad (7.64)$$

gdzie c_i są stałymi.

Założmy, że h ma ustaloną wartość ($h > 0$). Jeżeli na pierwiastki μ_i nałożymy warunki $|\mu_i| \leq 1$, $i = 1, \dots, k$, to żaden ze składników $c_i \mu_i^n$ we wzorze (7.64) nie będzie wzrastał, gdy $n \rightarrow \infty$. Ale jeżeli rozwiązanie dokładne rośnie ($\lambda > 0$), to ze względu na (7.63) dla dostatecznie małych h nie możemy żądać, aby $|\mu_1| \leq 1$. Wtedy składnik $c_1 \mu_1^n$ w (7.64) rośnie nieograniczenie. Warunek $|\mu_i| \leq |\mu_1|$ zapewnia, że w (7.64) poszczególne składniki $c_i \mu_i^n$ dla $i \geq 2$, odpowiadające pierwiastkom pasożytniczym, nie będą wzrastać szybciej niż składnik $c_1 \mu_1^n$.

Metodę różnicową nazywamy *absolutnie stabilną*, gdy są spełnione warunki $|\mu_i| \leq 1$, a *względnie stabilną*, gdy $|\mu_i| \leq |\mu_1|$.

W przeprowadzonej powyżej dyskusji założyliśmy, że równanie charakterystyczne ma pierwiastki rzeczywiste i jednokrotne. W pozostałych sytuacjach założymy, że pierwiastki μ_i oprócz podanych wyżej ograniczeń spełniają dodatkowo następujący warunek: jeżeli $|\mu_j| = 1$ (absolutna stabilność), $|\mu_j| = |\mu_1|$ (względna stabilność) to μ_j jest pierwiastkiem jednokrotnym. Pozostawiamy Czytelnikowi sprawdzenie na podstawie podanych w p. 7.3.2 wzorów na ε_n , że powyższe warunki w ogólnym przypadku zapewniają odpowiednie własności błędu ε_n .

Ważnym praktycznym problemem jest wyznaczenie tych wartości kroku całkowania h , dla których metoda różnicowa jest absolutnie lub względnie stabilna. Pierwiastki μ_i są pierwiastkami równania charakterystycznego (7.38) i zależą od wielkości $h\lambda$. Wygodnie więc będzie posługiwać się definicją stabilności podaną w następujący sposób.

Def. 1. Metodę różnicową nazywamy *absolutnie stabilną na przedziale* $(\alpha; \beta)$, jeżeli dla wszystkich $h\lambda \in (\alpha; \beta)$ zachodzi nierówność

$$|\mu_j| \leq 1, \quad j = 1, \dots, k \quad (7.65)$$

oraz pierwiastki μ_j , dla których $|\mu_j| = 1$, są jednokrotne.

Def. 2. Metodę różnicową nazywamy *względnie stabilną na przedziale* $(\alpha; \beta)$, jeżeli dla wszystkich $h\lambda \in (\alpha; \beta)$ zachodzi nierówność

$$|\mu_j| \leq |\mu_1|, \quad j = 2, \dots, k \quad (7.66)$$

oraz pierwiastki μ_j , dla których $|\mu_j| = |\mu_1|$, są jednokrotne.

Uwaga 1. Dla $h = 0$ warunki absolutnej i względnej stabilności pokrywają się, ponieważ wtedy $\mu_1 = 1$. Warunki te dla $h = 0$ muszą zachodzić, jeżeli chcemy, aby metoda różnicowa była zbieżna, co wynika z podanego dalej twierdzenia.

Uwaga 2. Przedział absolutnej stabilności powinien być postaci $(\alpha; 0)$, $\alpha < 0$. Wówczas dla dowolnego $\lambda < 0$ można otrzymać metodę absolutnie stabilną dla dostatecznie małej wartości h .

Uwaga 3. Przedział względnej stabilności powinien być postaci $(\alpha; \beta)$, $\alpha < 0$, $\beta > 0$, lub $(0; \beta)$, $\beta > 0$. Wówczas dla dowolnego λ lub tylko dla $\lambda > 0$ można otrzymać metodę względnie stabilną dla dostatecznie małej wartości h .

Metodę absolutnie stabilną dla wszystkich $h\lambda < 0$ nazywamy *A-stabilną*.

Przykład 2. Wyznaczyć przedziały absolutnej i względnej stabilności metody rzędu drugiego określonej wzorem

$$y_{n+1} = y_{n-1} + 2hy'_n$$

Równanie charakterystyczne i jego pierwiastki są następujące

$$\mu^2 = 2h\lambda\mu + 1$$

$$\mu_1 = h\lambda + \sqrt{h^2\lambda^2 + 1}, \quad \mu_2 = -\frac{1}{\mu_1}.$$

Metoda jest względnie stabilna na przedziale $(0; +\infty)$; warunek $|\mu_i| \leq 1$, $i = 1, 2$, jest spełniony tylko dla $h = 0$.

b) $m > 1$. Zbadamy teraz rozwiązanie y_n otrzymane za pomocą metody różnicowej (7.17), gdy układ równań różniczkowych jest postaci (7.32) tzn. $y' = \Lambda y$.

Dla takiego układu równań różniczkowych wzór (7.17) można zapisać w następującej postaci

$$(I - hb_0\Lambda)y_{n+1} = \sum_{i=1}^k (a_i I + hb_i\Lambda)y_{n+1-i} \quad (7.67)$$

Rozwiązanie y_n jest określone zatem przez równanie tej samej postaci co błąd ϵ_n^1 (patrz wzór (7.34)). Postać ta zależy od pierwiastków μ_s równania charakterystycznego (7.45), tzn. równania $\det M(\mu) = 0$. Łącznie z krotnościami pierwiastków tych jest km . Na podstawie rozważań przeprowadzonych pod koniec p. 7.3.2 wiemy, że:

- zależność pierwiastków μ_s od macierzy Λ wyraża się w ten sposób, że pierwiastki te zależą w istocie od wartości własnych λ_j macierzy Λ ,
- μ_s są pierwiastkami równania charakterystycznego

$$(1 - h\lambda b_0)\mu^k = \sum_{i=1}^k (a_i + h\lambda b_i)\mu^{k-i} \quad (7.68)$$

gdy $\lambda = \lambda_j$, $j = 1, \dots, m$; dla dalszej dyskusji wygodnie będzie oznaczyć pierwiastki powyższego równania dla $\lambda = \lambda_j$ przez μ_{ij} , $i = 1, \dots, k$.

Podobnie jak dla $m = 1$, w rozwiązaniu y_n można wydzielić część, która

przy $h \rightarrow 0$ dąży do dokładnego rozwiązania, oraz tzw. część pasożytnicza. Rozważymy dokładniej przypadek, gdy wartości własne λ_j oraz dla $h \neq 0$ pierwiastki μ_{ij} są jednokrotne i rzeczywiste. Wówczas rozwiązanie dokładne

$$Y(x) = c_1 e^{\lambda_1(x-x_0)} v_1 + \dots + c_m e^{\lambda_m(x-x_0)} v_m \quad (7.69)$$

gdzie v_j jest wektorem własnym macierzy A , odpowiadającym wartości własnej λ_j , zaś c_i są stałymi zależnymi od warunku początkowego (7.2). Zgodnie ze wzorem (7.46) rozwiązanie y_n można przedstawić w postaci

$$y_n = \sum_{i=1}^k \sum_{j=1}^m d_{ij} \mu_{ij}^n v_{ij} \quad (7.70)$$

gdzie v_{ij} jest wektorem własnym macierzy $M(\mu_{ij})$, a d_{ij} są stałymi. Stałe te zależą od $y_0, \dots, y_{k-1}$. Dla $h = 0$ liczba $\mu = 1$ jest pierwiastkiem każdego z równań (7.68). Niech $\mu_{1,j} = \mu_{1,j}(h)$, $j = 1, 2, \dots, m$, dla $h \neq 0$ oznaczają te pierwiastki, dla których

$$\lim_{h \rightarrow 0} \mu_{1,j}(h) = 1$$

Jeżeli metoda różnicowa jest rzędu p to (patrz wzór (7.58))

$$\mu_{1,j}^n = e^{\lambda_j(x_n - x_0)} + O(h^{p+1})$$

Pomijając szczegółowe rozważania, które przeprowadziliśmy poprzednio, ograniczymy się do stwierdzenia, że częścią sumy (7.70), która dla $h \rightarrow 0$ dąży do rozwiązania dokładnego zagadnienia (7.1), (7.2) jest

$$\sum_{j=1}^m d_{1,j} \mu_{1,j}^n v_{1,j}$$

pozostała zaś część sumy (7.70) jest częścią pasożytniczą rozwiązania, a pierwiastki μ_{ij} ($i > 2$) są pierwiastkami pasożytniczymi.

Dla $m > 1$, zgodnie z podanymi na początku tego punktu własnościami rozwiązań układu (7.33), błąd

$$\varepsilon_n = \sum_{i=1}^k \sum_{j=1}^m c_{ij} \mu_{ij}^n v_{ij} - \frac{1}{h \sum_{i=0}^k b_i} A^{-1} B \quad (7.71)$$

Dla metody różnicowej warunek

$$|\mu_{ij}| \leq 1, \quad i = 1, \dots, k; \quad j = 1, \dots, m \quad (7.72)$$

zapewnia, że żaden ze składników w sumie (7.71) nie będzie wzrastał, gdy $n \rightarrow \infty$ — metodę taką nazywamy *absolutnie stabilną*. Natomiast warunek

$$|\mu_{ij}| \leq |\mu_{1j}|, \quad i = 1, \dots, k; \quad j = 1, \dots, m \quad (7.73)$$

zapewnia, że składniki $c_{ij}\mu_{ij}^n v_{ij}$ ($i > 2$) nie wzrastają szybciej niż składnik $c_{1j}\mu_{1j}^n v_{1j}$ — metodę taką nazywamy *względnie stabilną*.

Dotąd rozpatrywaliśmy przypadek, gdy μ_{ij} są jednokrotnymi i rzeczywistymi pierwiastkami równania (7.68). W pozostałych sytuacjach zakładamy, że μ_{ij} oprócz podanych wyżej ograniczeń spełniają dodatkowo następujący warunek: jeżeli $|\mu_{ij}| = 1$ (absolutna stabilność), $|\mu_{ij}| = |\mu_{1j}|$ (względna stabilność), to μ_{ij} jest pierwiastkiem jednokrotnym.

Zauważmy, że pierwiastki μ_{ij} zależą od wielkości $h\lambda$. W ogólnym przypadku wartości własne λ macierzy $\mathbf{A}$ mogą być liczbami zespolonymi. Dlatego w przypadku układu równań mówimy, że metoda różnicowa jest absolutnie lub względnie stabilna na obszarze Ω na płaszczyźnie zespolonej, jeżeli dla wszystkich $h\lambda \in \Omega$ są spełnione odpowiednio warunki (7.72), (7.73).

Metodę absolutnie stabilną dla wszystkich $h \operatorname{Re} \lambda < 0$ nazywamy *A-stabilną*.

Przykład 3. Wyznaczyć obszar absolutnej stabilności metody drugiego rzędu określonej wzorem

$$y_{n+1} = y_n + \frac{h}{2}(y'_{n+1} + y'_n)$$

Równanie charakterystyczne, którego pierwiastki należy wyznaczyć, ma postać

$$\left(1 - \frac{h}{2}\lambda\right)\mu = \left(1 + \frac{h}{2}\lambda\right)$$

Jest to równanie pierwszego stopnia. Dla każdego λ takiego, że $\operatorname{Re} \lambda < 0$ mamy $|\mu| < 1$. Rozważana metoda jest więc metodą A-stabilną.

Definicję absolutnej i względnej stabilności określiliśmy dla liniowych układów równań różniczkowych o stałych współczynnikach. W innych przypadkach dla każdego $n \geq k$ równanie różniczkowe $y' = f(x, y)$ zapiszemy w postaci

$$y' = f_n + f_y(x_n, y_n)(y - y_n) + O(h^2) \quad (7.74)$$

gdzie

$$f_n = f(x_n, y_n), \quad \text{zaś} \quad f_y(x_n, y_n) = \left[\frac{\partial f^i}{\partial y^j}(x_n, y_n) \right]_{i,j=1,\dots,m}$$

Prawa strona układu (7.74) jest więc rozwinięciem Taylora wokół punktu (x_n, y_n) . Jako macierz $\mathbf{A}$ będziemy przyjmować macierz $f_y(x_n, y_n)$, która zmienia się z kroku na krok. Wartości własne λ_j^n macierzy $\mathbf{A}$ określają h w ten sposób, że iloczyn $\lambda_j^n h$ powinien należeć do obszaru stabilności metody różnicowej. Na to, aby metoda była praktycznie przydatna (tzn. nie narzucała zbyt mocnych ograniczeń na h) dla możliwie szerokich klas równań różniczkowych, obszar stabilności powinien być możliwie duży.

W przypadku jednego równania Λ jest skalar i iloczyn $h\Lambda$ powinien należeć do przedziału stabilności.

Zależnościami między stabilnością układów $y' = f(x, y)$ i $y' = f_n + f_y(x_n, y_n)(y - y_n)$ zajmowano się w pracach [19], [35]; czytelnika interesującego się tymi problemami odsyłamy do tych prac.

Zbieżność. W punkcie 7.3.1 podaliśmy warunek konieczny zbieżności metody różnicowej (tym warunkiem jest, aby metoda była zgodna).

Tw. *Warunkiem koniecznym i wystarczającym zbieżności metody różnicowej zgodnej jest warunek, aby wszystkie pierwiastki μ_j wielomianu*

$$\varrho(\mu) = \mu^k - \sum_{i=1}^k a_i \mu^{k-i} \quad (7.75)$$

miały następujące własności;

- 1) $|\mu_j| \leq 1$,
- 2) jeżeli $|\mu_j| = 1$, to μ_j jest pierwiastkiem jednokrotnym.

Zauważmy, że dla $h = 0$:

— pierwiastki wielomianu charakterystycznego $\det M(\mu) = 0$ pokrywają się z pierwiastkami wielomianu $\varrho(\mu)$,

— warunek absolutnej i względnej stabilności pokrywa się z wyżej sformułowanym warunkiem koniecznym i wystarczającym zbieżności metody zgodnej.

Zatem domknięcie obszarów (przedziałów) absolutnej i względnej stabilności powinno zawierać zero.

Metodę różnicową, dla której są spełnione warunki (7.76) nazywamy *stabilną w sensie Dahlquista*.

Podamy teraz oszacowanie błędu całkowitego $\varepsilon_n = Y_n - \tilde{y}_n$ przy założeniu, że rozwiązanie zagadnienia (7.3) jest dostatecznie regularne.

Jeżeli rozwiązanie $Y(x)$ zagadnienia (7.3) na przedziale $\langle x_0; b \rangle$ ma ciągłe pochodne do rzędu $(p+1)$ włącznie, a metoda różnicowa określona wzorem (7.17), jest zbieżna i jest rzędu p , gdzie $p \geq 1$, to prawdziwe jest oszacowanie

$$\|\tilde{y}_n - Y_n\| \leq \left\{ \tilde{c}_1 \theta + \tilde{c}_2 (x_n - x_0) \left(\frac{\delta}{h} + \tilde{c}_3 \|Y^{(p+1)}\| h^p \right) \right\} e^{\tilde{c}_4 (x_n - x_0)}, \quad n \geq k \quad (7.77)$$

$$\text{gdzie } \theta = \max_{0 \leq i \leq k-1} \|y_i - Y_i\|, \quad \delta = \max_{n \geq k} \|\delta_n\|, \quad \|Y^{(p+1)}\| = \max_{x_0 \leq x \leq b} \|Y^{(p+1)}(x)\|,$$

a $\tilde{c}_i$ — stałe dodatnie (niezależne od h) dla dostatecznie małego h .

Dla metody Eulera ($k = 1$), gdy $\theta = \delta = 0$ i $m = 1$, oszacowanie takie wyprowadzono w p. 7.2.

Oszacowanie (7.77) błędu całkowitego składa się z trzech części związanych z trzema różnymi rodzajami błędów, z którymi mamy do czynienia przy stosowaniu metod różnicowych (7.17). Są to błędy: błędy rozwiązań początkowych $y_1, y_2, \dots, y_{k-1}$, błąd metody i błędy powstałe przy numerycznym wyznaczaniu rozwiązania równania różnicowego. Charakteryzują się one

wielkościami θ , h^p , δ/h , odpowiednio. Dla ustalonego h oszacowanie błędu rośnie wykładniczo wraz z długością przedziału, na którym szukamy rozwiązania (także w przypadku, gdy wszystkie działania są wykonywane dokładnie). Zauważmy, że ten szybki wzrost oszacowania jest niezależny od rzędu wielkości rozwiązania dokładnego $Y(x)$.

Okazuje się, że wśród metod zbieżnych są też takie, dla których błąd wzrasta nieograniczenie wtedy, gdy rozwiązanie maleje. Pokażemy to na następującym przykładzie.

Przykład 4. Wyznaczyć rozwiązanie zagadnienia $y' = -y$, $y(0) = 1$ metodą określoną wzorem (patrz przykł. 2)

$$y_{n+1} = y_{n-1} + 2hy'_n$$

z wartościami początkowymi y_0, y_1 spełniającymi warunki (7.22).

Zgodnie z ostatnim twierdzeniem metoda ta jest zbieżna. Nie jest ona jednak absolutnie stabilna dla żadnej wartości $h\lambda < 0$.

Korzystając z danego równania różniczkowego wzór określający stosowaną metodę zapiszemy w postaci

$$y_{n+1} + 2hy_n - y_{n-1} = 0$$

Rozwiązanie ogólne tego równania ma postać

$$y_n = c_1\mu_1^n + c_2\mu_2^n, \quad n > 0$$

gdzie c_1, c_2 są stałymi, $\mu_1 = -h + \sqrt{h^2 + 1}$, $\mu_2 = -h - \sqrt{h^2 + 1}$.

Mamy zatem

$$\lim_{\substack{h \rightarrow 0 \\ (n \rightarrow \infty)}} \mu_1^n = e^{-x}, \quad \lim_{\substack{h \rightarrow 0 \\ (n \rightarrow \infty)}} |\mu_2|^n = e^x, \quad nh = x$$

$$\begin{aligned} c_1 = c_1(h) \rightarrow 1, \quad c_2 = c_2(h) \rightarrow 0, \quad \text{gdy } h \rightarrow 0 \\ c_2(h) = 0 \quad \Leftrightarrow \quad y_1 = \mu_1, \quad \text{gdy } y_0 = 1 \end{aligned}$$

Uzyskane rozwiązanie jest więc zbieżne do dokładnego rozwiązania $Y(x) = e^{-x}$. Rozwiązanie to zawiera jednak człon $c_2\mu_2^n$ także wtedy, gdy $y_1 = e^{-h}$ (wartość dokładnego rozwiązania). Człon ten powoduje, że dla każdego ustalonego h ($h > 0$) błąd rośnie wykładniczo wraz z długością przedziału.

Rozpatrywana w powyższym przykładzie metoda należy do tych zbieżnych metod różnicowych, których wielomian charakterystyczny ma pierwiastek, np. $\mu_2 = \mu_2(h)$ ($\mu_2 \neq \mu_1$) o następujących własnościach: $\mu_2(0) \neq 1$, $|\mu_2(0)| = 1$, $|\mu_2(h)| > 1$ dla każdego $h\lambda < 0$. Obecność takiego pierwiastka może wywołać efekt wzrostu błędu zgodny z oszacowaniem (7.77).

Dowody twierdzenia o zbieżności oraz oszacowania (7.77), pełniejszą analizę zachowania się błędów można znaleźć w [34], [4].

Ćwiczenia

1. Zapisać pierwiastek $\mu_1 = \mu_1(h)$ w postaci $\mu_1(h) = 1 + \sum_{i=1}^p \beta_i h^i + O(h^{p+1})$, a następnie podstawiając tę postać do wzoru (7.38) i wykorzystując warunki (7.20) udowodnić (7.57).

7.3.4. Wyznaczanie praktyczne przydatnych wzorów różnicowych

W tablicy 7/2 zestawiono wzory różnicowe, które najczęściej są omawiane i stosowane w praktyce. Wzory 1—5 są wzorami jawnymi, a wzory 6—10 są 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 zbieżnych metod różnicowych zależność między k i rzędem metody p określa następujące twierdzenie.

Tw. Dla k parzystego maksymalny rząd metody zbieżnej wynosi $(k + 2)$, a dla k nieparzystego $(k + 1)$.

Dowód tego twierdzenia można znaleźć np. w [14].

Wzór różnicowy (7.17) zależy od $(2k + 1)$ współczynników a_i, b_i . Jeżeli współczynniki te określimy tak, aby:

- 1) rząd metody osiągnął zadaną wartość
- 2) spełnione były warunki (7.76) zapewniające zbieżność metody, to pozostanie pewna liczba wolnych parametrów, które można wykorzystać do osiągnięcia następujących celów:

7/2

	Wzór	p	A_{p+1}
1	$y_{n+1} = y_n + h y'_n$	1	$\frac{1}{2}$
2	$y_{n+1} = y_n + \frac{h}{2} (3 y'_n - y'_{n-1})$	2	$\frac{5}{12}$
3	$y_{n+1} = y_n + \frac{h}{12} (23 y'_n - 16 y'_{n-1} + 5 y'_{n-2})$	3	$\frac{3}{8}$
4	$y_{n+1} = y_n + \frac{h}{24} (55 y'_n - 59 y'_{n-1} + 37 y'_{n-2} - 9 y'_{n-3})$	4	$\frac{251}{720}$
5	$y_{n+1} = y_{n-3} + \frac{4}{3} h (2 y'_n - y'_{n-1} + 2 y'_{n-2})$	4	$\frac{14}{45}$
6	$y_{n+1} = y_n + h y'_{n+1}$	1	$-\frac{1}{2}$
7	$y_{n+1} = y_n + \frac{h}{2} (y'_{n+1} + y'_n)$	2	$-\frac{1}{12}$
8	$y_{n+1} = y_n + \frac{h}{12} (5 y'_{n+1} + 8 y'_n - y'_{n-1})$	3	$-\frac{1}{24}$
9	$y_{n+1} = y_n + \frac{h}{24} (9 y'_{n+1} + 19 y'_n - 5 y'_{n-1} + y'_{n-2})$	4	$-\frac{19}{720}$
10	$y_{n+1} = \frac{1}{8} (9 y_n - y_{n-2}) + \frac{3}{8} h (y'_{n+1} + 2 y'_n - y'_{n-1})$	4	$-\frac{1}{40}$

- a) zwiększenia obszarów (przedziałów), absolutnej i względnej stabilności,
- b) zmniejszenia współczynnika w T_n (patrz wzór (7.21)),
- c) łatwości wykonywania obliczeń (np. zerowe współczynniki).

Rozpatrzmy (za Hammingiem [33]) metodę różnicową określoną następującym wzorem

$$y_{n+1} = a_1 y_n + a_2 y_{n-1} + a_3 y_{n-2} + h(b_0 y'_{n+1} + b_1 y'_n + b_2 y'_{n-1}) \quad (7.78)$$

Mamy tu sześć wolnych parametrów. Żądając, aby metoda ta była czwartego rzędu (zgodnie z powyższym twierdzeniem jest to maksymalny rząd, jakiego możemy żądać), otrzymamy następujące równania (patrz ćwic. 1)

$$\begin{aligned} a_1 &= \frac{1}{8}(9 - 9a_2) & b_0 &= \frac{1}{24}(9 - a_2) \\ a_2 &= a_2, & b_1 &= \frac{1}{12}(9 + 7a_2) \\ a_3 &= -\frac{1}{8}(1 - a_2), & b_2 &= \frac{1}{24}(-9 + 17a_2) \end{aligned} \quad (7.79)$$

$$T_h = (-9 + 5a_2) h^5 y^{(5)}(\xi_n)/360$$

z parametrem a_2 . Współczynniki równania charakterystycznego

$$(1 - h\lambda b_0) \mu^3 = (a_1 + h\lambda b_1) \mu^2 + (a_2 + h\lambda b_2) \mu + a_3$$

zależą od parametru a_2 . W tej sytuacji wyznaczenie przedziału absolutnej i względnej stabilności jest trudne. Dla $h = 0$ warunki stabilności są spełnione dla $a_2 \in \langle -0,6; 1 \rangle$. Najwłaściwszym wydaje się wybór wartości a_2 leżącej w pobliżu środka tego przedziału, ponieważ wtedy wzór będzie stabilny w szerszym przedziale zmienności $h\lambda$ niż przy skrajnych wartościach a_2 . Hamming wybrał $a_2 = 0$. Otrzymaliśmy więc wzór 10 z tablicy 7.2. Równanie charakterystyczne dla tego wzoru ma postać

$$\left(\frac{3h\lambda}{8} - 1\right) \mu^3 = -\mu^2 \left(\frac{9}{8} + \frac{3h\lambda}{4}\right) + \mu \frac{3h\lambda}{8} + \frac{1}{8} \quad (7.80)$$

Przedziałem względnej stabilności dla $h\lambda$ jest $(-0,69; +\infty)$, a przedziałem absolutnej stabilności jest $(-2,6; 0)$.

Praktyczny sposób wykorzystania wzoru podanego przez Hamminga do wyznaczenia rozwiązania pojedynczych równań różniczkowych jak i układów będzie rozpatrzony w p. 7.5.1.

Na zakończenie przedstawimy krótko sposób otrzymywania wzorów różnicowych dla $m = 1$, polegający na wykorzystaniu wzorów Newtona-Cotesa (wzorów całkowania numerycznego). Całkując równanie $y' = f(x, y)$ na przedziale $\langle x_{n-j}; x_{n+1} \rangle$, otrzymujemy rozwiązanie dokładne

$$y(x_{n+1}) = y(x_{n-j}) + \int_{x_{n-j}}^{x_{n+1}} f(x, y(x)) dx \quad (7.81)$$

W miejsce funkcji podcałkowej (jest ona nieznana ponieważ nie znamy $y(x)$) podstawimy wielomian interpolacyjny $L(x)$, przechodzący przez punkty $(x_j, y_j'), j = n, n-1, \dots, n-l$. Wielomian ten zapiszemy w postaci Lagrange'a

$$L(x) = \sum_{i=0}^l f_{n-i} P_i(x)$$

przy czym wielomiany $P_i(x)$ nie zależą od f_j . Po podstawieniu $L(x)$ w miejsce $f(x, y(x))$ w (7.81) otrzymamy następujący wzór różnicowy

$$y_{n+1} = y_{n-j} + \sum_{i=0}^l f_{n-i} \int_{x_{n-j}}^{x_{n+1}} P_i(x) dx = y_{n-j} + h \sum_{i=0}^l b_i f_{n-i} \quad (7.82)$$

gdzie $b_i = \int_{-j}^1 P_i(x_n + th) dt, t = \frac{x - x_n}{h}$. Łatwo można sprawdzić, że współ-

czynniki b_i nie zależą od h (patrz ćwic. 2).

Wzór (7.82) jest wzorem ekstrapolacyjnym (jawnym). Przyjmując $j = 0$ otrzymujemy wzór Adamsa-Bashfortha. Wzór 5 z tablicy 7.2 można otrzymać przyjmując $j = 3, l = 2$.

Jeżeli do zbudowania wielomianu $L(x)$ oprócz $y'_n, \dots, y'_{n-l}$ weźmiemy (nieznana jeszcze) wartość y'_{n+1} , to otrzymamy wzór interpolacyjny (uwikłany). Przyjmując teraz $j = 0$ otrzymamy wzór Adamsa-Moultona.

Otrzymane w opisany powyżej sposób wzory są zbieżne. Niemniej jednak taki sposób wyprowadzenia wzorów różnicowych narzuca pewne ograniczenia na ich postać (wszystkie a_j , oprócz jednego, są zerami). Wychodząc natomiast z ogólnego wzoru różnicowego, podanego na początku tego punktu, możemy otrzymać wszystkie przydatne wzory różnicowe, a więc i takie, których nie można wyprowadzić z kwadratur Newtona-Cotesa.

Ćwiczenia

1. Udowodnić, że współczynniki a_i, b_i wzoru różnicowego (7.78) są określone wzorami (7.79), jeżeli zażądamy, aby wzór (7.78) określał metodę czwartego rzędu.
2. Udowodnić, że współczynniki b_i we wzorze (7.82) nie zależą od h .
3. Wyprowadzić wzory 2 i 7 z tablicy 7.2 korzystając z kwadratur Newtona-Cotesa.

7.4. METODY TYPU RUNGEGO-KUTTY

7.4.1. Wzór ogólny

Metodę określoną wzorem

$$y_{n+1} = y_n + \sum_{i=1}^s w_i K_i \quad (7.83)$$

gdzie:

$$K_1 = hf(x_n, y_n)$$

$$K_i = h f \left(x_n + a_i h, y_n + \sum_{j=1}^{i-1} b_{ij} K_j \right), \quad i > 1$$

w_i, a_i, b_{ij} — stałe

nazywamy *metodą Rungego-Kutty*.

Metodę tę możemy bezpośrednio stosować do rozwiązywania zagadnienia (7.1), (7.2), ponieważ w przeciwieństwie do metod różnicowych (7.17) ($k > 1$) do rozpoczęcia obliczeń wystarczy warunek początkowy (7.2).

Istotną częścią obliczeń przy wyznaczaniu rozwiązań y_{n+1} jest wyznaczenie funkcji $f(x, y)$, występującej z prawej strony układu równań różniczkowych. Stosując powyższe wzory musimy jednak w każdym kroku s razy obliczać $f(x, y)$. Udowodniono, że dla metod rzędu $p = 2, 3, 4$ najmniejsza liczba obliczeń $s = p$, a dla $p > 4$ najmniejsza liczba $s > p$. Tak więc dla metody rzędu p , należy tych obliczeń wykonać co najmniej p . W następnym punkcie omówimy metody ekstrapolacyjno-interpolacyjne oparte na omówionych w p. 7.3 wzorach różnicowych, które wymagają mniejszej liczby obliczeń $f(x, y)$, a więc są szybsze. Metody Rungego-Kutty mają duże znaczenie dla wyznaczania brakujących wartości wstępnych.

Rząd metody Rungego-Kutty wyznaczamy podobnie jak rząd metody różnicowej. Liczba p jest rzędem metody (7.83), jeżeli po podstawieniu rozwiązania dokładnego do (7.83) i rozwinięciu według wzoru Taylora wielkości: Y_{n+1} w otoczeniu punktu x_n oraz K_i (jako funkcji zmiennych x, Y) w otoczeniu punktu (x_n, Y_n) , współczynniki przy potęgach h^r będą jednakowe ($r = 1, \dots, p$), a przy h^{p+1} — różne.

Największe znaczenie praktyczne mają metody rzędu czwartego dla $s = 4$. Wyprowadzenie takich wzorów wymaga jednak żmudnych algebraicznych obliczeń, nie wystarczy bowiem porównać współczynników przy odpowiednich potęgach h , gdyż współczynniki te zależą od funkcji $f_i(x, y)$ i jej pochodnych cząstkowych, a w związku z tym należy dodatkowo porównać współczynniki przy odpowiednich wyrażeniach.

Omówimy najprostszy przypadek metod Rungego-Kutty, gdy $m = 1$ i $s = 2$ (zauważmy, że dla $s = 1$ i $w_1 = 1$ metoda określona wzorem (7.83) jest metodą Eulera, omówioną w p. 7.2). Mamy

$$Y_n + h Y_n' + \frac{h^2}{2} Y_n'' + O(h^3) = Y_n + h \left[w_1 f + w_2 \left(f + a_2 h \frac{\partial f}{\partial x} + \right. \right. \\ \left. \left. + b_{21} h f \frac{\partial f}{\partial y} \right) \right] \Big|_{(x_n, Y_n)} + O(h^3)$$

Podstawiając

$$Y_n' = f|_{(x_n, Y_n)}, \quad Y_n'' = \left(\frac{\partial f}{\partial x} + f \frac{\partial f}{\partial y} \right) \Big|_{(x_n, Y_n)}$$

i porównując współczynniki przy h^i , $i = 1, 2$, otrzymamy

$$w_1 + w_2 = 1$$

$$\frac{1}{2} \left(\frac{\partial f}{\partial x} + f \frac{\partial f}{\partial y} \right) = w_2 a_2 \frac{\partial f}{\partial x} + w_2 b_{21} f \frac{\partial f}{\partial y}$$

Ponieważ chcemy, aby wielkości w_i , a_i , b_{ij} były niezależne od $f(x, y)$, więc otrzymujemy trzy następujące równania

$$w_1 + w_2 = 1$$

$$a_2 w_2 = \frac{1}{2} \tag{7.84}$$

$$b_{21} = a_2$$

Ponadto, już po dość skomplikowanych obliczeniach możemy wyznaczyć γ_2 ze wzoru (7.8). Mamy

$$\gamma_2 = \left(\frac{1}{6} - \frac{1}{2} a_2^2 w_2 \right) D^2 f \Big|_{(x_n, Y_n)} + \frac{1}{6} f_y Df \Big|_{(x_n, Y_n)} \tag{7.85}$$

$$\text{gdzie } Df = \left(\frac{\partial}{\partial x} + f \frac{\partial}{\partial y} \right) f,$$

$$D^2 f = \left(\frac{\partial^2}{\partial x^2} + 2f \frac{\partial^2}{\partial x \partial y} + f^2 \frac{\partial^2}{\partial y^2} \right) f$$

Jeżeli dla funkcji $f(x, y)$ i jej pochodnych przyjmimy następujące oszacowanie

$$|f| \leq M, \quad \left| \frac{\partial^{i+j} f}{\partial x^i \partial y^j} \right| \leq \frac{N}{M^{j-1}}, \quad i+j \leq 2$$

to

$$|\gamma_2| \leq \left(4 \cdot \left| \frac{1}{6} - \frac{1}{2} a_2^2 w_2 \right| + \frac{1}{3} \right) MN \tag{7.86}$$

Układ (7.84) ma nieskończenie wiele rozwiązań zależnych od jednego parametru, na przykład od a_2 . Każde rozwiązanie tego układu daje nam metodę drugiego rzędu określoną wzorem

$$y_{n+1} = y_n + \left(1 - \frac{1}{2a_2} \right) K_1 + \frac{1}{2a_2} K_2, \quad a_2 \neq 0 \tag{7.87}$$

gdzie $K_1 = hf(x_n, y_n)$, $K_2 = hf(x_n + a_2 h, y_n + a_2 K_1)$. Parametr a_2 możemy tak wybrać, aby powyższy wzór miał dodatkowe własności, np. jak najmniejszą wartość oszacowania (7.86).

Pozostawiamy Czytelnikowi przeprowadzenie dyskusji dla przypadku układu ($m > 1$).

Widzimy, że błędu rozpatrywanej metody nie można przedstawić w tak prostej postaci jak dla metod różnicowych (patrz wzór (7.21)). Dla metod wyższego rzędu γ_p będzie jeszcze bardziej skomplikowane.

Dla dowolnego s mamy

$$Y_n + hY'_n + O(h^2) = Y_n + h(w_1 + \dots + w_s)f(x_n, Y_n) + O(h^2)$$

A zatem, jeżeli jest spełniony warunek

$$w_1 + \dots + w_s = 1 \quad (7.88)$$

to metoda (7.83) jest rzędu co najmniej pierwszego.

Ćwiczenia

1. Wyznaczyć wszystkie metody drugiego rzędu, gdy $s = 2$ i $m > 1$.

7.4.2. Stabilność metod Rungego-Kutty

Stabilność metod Rungego-Kutty, podobnie jak dla metod różnicowych, zbadamy dla układu liniowych równań różniczkowych, przy czym rozpatrzemy dokładniej przypadek $m = 1$. Jeżeli równanie różniczkowe jest postaci $y' = \lambda y$ (λ jest liczbą rzeczywistą) i jest spełniony warunek (7.88), to wzór (7.83) możemy zapisać w postaci

$$y_{n+1} = (1 + h\lambda + O(h^2))y_n \quad (7.89)$$

Równaniem charakterystycznym dla (7.89) jest wielomian pierwszego stopnia

$$\mu = 1 + h\lambda + O(h^2)$$

Rozwiązanie y_n ma zatem postać

$$y_n = (1 + h\lambda + O(h^2))^n y_0 = [e^{h\lambda} + O(h^2)]^n y_0 = [e^{\lambda(x_n - x_0)} + O(h^2)] y_0$$

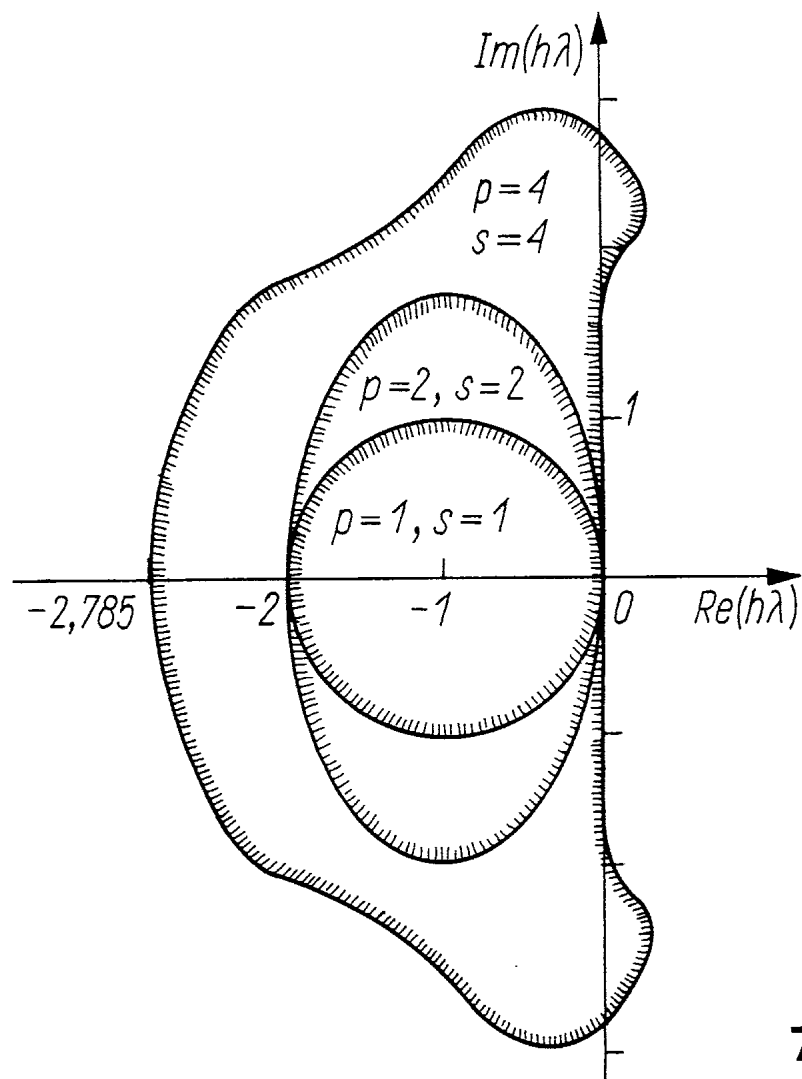
Tak więc y_n przy $h \rightarrow 0$ dąży do rozwiązania dokładnego $Y(x) = y_0 e^{\lambda(x - x_0)}$. W odróżnieniu od metod różnicowych (dla których $k > 1$), otrzymane rozwiązanie y_n nie posiada tzw. części pasożytniczej, dlatego dla metod Rungego-Kutty określamy tylko przedział absolutnej stabilności.

Dla metod drugiego rzędu ($p = 2$), gdy $s = 2$ równanie charakterystyczne ma postać

$$\mu = 1 + h\lambda + \frac{1}{2}h^2\lambda^2$$

Zauważmy, że nie zależy ono od parametru a_2 . Jak łatwo sprawdzić, przedział absolutnej stabilności jest następujący: $h\lambda \in (-2; 0)$.

W przypadku układu równań ($m > 1$), rozwiązanie y_n także nie posiada części pasożytniczej, a więc określamy tylko obszar absolutnej stabilności (przypomnimy, że w przypadku układu równań, λ jest liczbą zespoloną). Na



rysunku 7.3 podano obszar absolutnej stabilności dla metod Rungego-Kutty rzędu drugiego, gdy $s = 2$.

Na zakończenie rozważań o stabilności podamy, że wszystkie metody Rungego-Kutty rzędu co najmniej pierwszego są absolutnie stabilne dla dostatecznie małych h , ale żadna z metod nie jest metodą A -stabilną.

7/3

	Wzór klasyczny	Wzór „trzech ósmych”	Wzór Gilla	Wzór Ralstona
w_1	1/6	1/8	1/6	0.17476028
w_2	1/3	3/8	$(1 - 1/\sqrt{2})/3$	-0.55148053
w_3	1/3	3/8	$(1 + 1/\sqrt{2})/3$	1.20553547
w_4	1/6	1/8	1/6	0.17118478
a_2	1/2	1/3	1/2	0.4
a_3	1/2	2/3	1/2	0.45573726
a_4	1	1	1	1
b_{21}	1/2	1/3	1/2	0.4
b_{31}	0	-1/3	$(\sqrt{2} - 1)/2$	0.29697760
b_{32}	1/2	1	$1 - 1/\sqrt{2}$	0.15875966
b_{41}	0	1	0	0.21810038
b_{42}	0	-1	$-1/\sqrt{2}$	-3.05096470
b_{43}	1	1	$1 + 1/\sqrt{2}$	3.83286432

7.4.3. Metody rzędu czwartego

Omówimy krótko metody czwartego rzędu dla $s = 4$. W tym przypadku układ określający parametry w_i, a_i, b_{ij} zawiera 13 niewiadomych i 11 równań, zatem nie są one określone jednoznacznie. W tablicy 7/3 podano najczęściej stosowane w praktyce wzory. Ze wzorów tych jest wykorzystywany wzór klasyczny, ze względu na jego prostotę. Przy stosowaniu wzoru Gilla obliczenia można prowadzić w ten sposób, żeby wyrównywać gromadzące się błędy zaokrągleń ([51]) (otrzymujemy wówczas największą osiągalną dokładność obliczeń). Wzór Ralstona daje najlepsze oszacowanie błędu γ_4 ([50]) (takie oszacowanie dla γ_2 wyprowadziliśmy w poprzednim punkcie).

Równanie charakterystyczne dla tych metod ma postać

$$\mu = 1 + h\lambda + \frac{1}{2}h^2\lambda^2 + \frac{1}{6}h^3\lambda^3 + \frac{1}{24}h^4\lambda^4$$

Przedział absolutnej stabilności: $h\lambda \in (-2,785; 0)$; obszar absolutnej stabilności jest pokazany na rys. 7.3.

Dla błędu całkowitego prawdziwe jest takie samo oszacowanie, jak w przypadku metod różnicowych (patrz oszacowanie (7.77)) z tym, że należy przyjąć $k = 1$.

7.4.4. Wybór kroku całkowania

Przy wyborze kroku całkowania h należy brać pod uwagę stabilność i dokładność rozwiązania, przy czym:

a) Metody Rungego-Kutty nie są metodami A -stabilnymi, więc h powinno podlegać ograniczeniom, które zapewniają stabilność; zagadnienie to omówiliśmy w p. 7.3.3.

b) Dokładność otrzymanego wyniku zależy od błędu metody i błędów zaokrągleń. Ponieważ zaokrąglenia w porównaniu z błędem metody Rungego-Kutty są na ogół mało istotne, więc błąd metody uważać będziemy za główny składnik błędu całkowitego.

Wzór (7.83) jest napisany dla stałego h , natomiast w obliczeniach prowadzonych na maszynach cyfrowych stosuje się zmienny krok całkowania, postępując następująco. Początkową wartość h ustalamy tak, aby norma błędu metody była nie większa od ustalonej wielkości ε . Wielkość tę wybiera się najczęściej w ten sposób, że jeżeli szacowany błąd metody jest nie większy od ε , to rozwiązanie powinno mieć żadaną dokładność w przedziale, w którym je wyznaczamy. Jeżeli podczas obliczeń okaże się, że szacowany błąd metody jest większy od ε (rozwiązanie może nie osiągnąć żadnej dokładności), to zmniejszamy krok całkowania. Jeżeli szacowany błąd metody jest znacznie mniejszy od ε , to krok całkowania zwiększamy. Gdy wybrany w powyższy sposób krok h (wynika on z oszacowania błędu metody) nie zapewnia stabilności, wówczas należy go zmniejszyć tak, abyśmy mieli zapewnioną stabilność.

W metodach Rungego-Kutty do wyznaczania przybliżonej wartości błędu metody stosujemy tzw. *zasadę Rungego*. Przyjmiemy chwilowo, że $y_n = Y(x_n)$. Niech $y_{n+1}^{(1)}$ oznacza przybliżone rozwiązanie w punkcie $x_{n+1} = x_n + h$, gdy obliczenia zostały wykonane dwukrotnie z krokiem $h/2$. Zakładając, że błąd metody przy przejściu od x_n do $\bar{x}_n = x_n + \frac{h}{2}$ oraz od $\bar{x}_n$ do x_{n+1} jest prawie taki sam, otrzymamy na błąd rozwiązania $y_{n+1}^{(1)}$ następujące wyrażenie

$$Y_{n+1} - y_{n+1}^{(1)} \approx 2\gamma_p \left(\frac{h}{2}\right)^{p+1} \quad (7.90)$$

gdy p jest rzędem rozpatrywanej metody.

Jeżeli teraz przeprowadzimy powtórne obliczenia, ale od razu z krokiem równym h i przez $y_{n+1}^{(2)}$ oznaczmy otrzymane w ten sposób przybliżone rozwiązanie w punkcie x_{n+1} , to błąd rozwiązania $y_{n+1}^{(2)}$ będzie następujący

$$Y_{n+1} - y_{n+1}^{(2)} \approx \gamma_p h^{p+1} \quad (7.91)$$

Eliminując ze wzorów (7.90), (7.91) γ_p , otrzymamy

$$Y_{n+1} - y_{n+1}^{(1)} \approx \delta, \quad \delta = \frac{1}{2^p - 1} (y_{n+1}^{(1)} - y_{n+1}^{(2)}) \quad (7.92)$$

gdzie δ jest przybliżoną wielkością błędu metody (przy przejściu od x_n do x_{n+1}), gdy obliczenia wykonujemy dwukrotnie z krokiem $h/2$.

Podamy jeszcze sposób, w jaki jest wybierany krok całkowania w standardowych procedurach bibliotecznych opartych na metodach czwartego rzędu (weźmiemy pod uwagę tylko błąd metody). Krok początkowy h_0 i ε , tzn. górne oszacowanie błędu metody, są podawane przez użytkownika. Załóżmy, że mamy obliczone już rozwiązanie w punkcie x_n . Obliczenia w następnym etapie (tzn. 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 $h = h_0$). Ponieważ $Y(x_n)$ nie jest znane, δ wyznaczamy w wyżej opisany sposób z tym, że jako y_n przyjmujemy obliczone rozwiązanie w punkcie x_n . Za miarę błędu metody na tym etapie przyjmujemy liczbę

$$\delta = \|\delta\| = \frac{1}{15} \|y_{n+1}^{(1)} - y_{n+1}^{(2)}\| \quad (7.93)$$

Mogą wystąpić dwa przypadki:

a) $\delta \leq \varepsilon$. Obliczone rozwiązanie $y_{n+1}^{(1)}$ jest uznawane jako wystarczająco dokładne. Wówczas, jeśli ponadto $\delta < \varepsilon/50$, to krok h podwajamy, w przeciwnym zaś przypadku przechodzimy do następnego etapu.

b) $\delta > \varepsilon$. Obliczone rozwiązanie $y_{n+1}^{(1)}$ jest uznawane jako niewystarczająco dokładne. Krok całkowania jest połowiony i obliczenia są wykonywane jeszcze raz.

7.5. Metody ekstrapolacyjno-interpolacyjne

W podrozdziale 7.3 omówiliśmy zbieżność i stabilność metod różnicowych określonych ogólnym wzorem

$$y_{n+1} = \sum_{i=1}^k a_i y_{n+1-i} + h \sum_{i=0}^k b_i f(x_{n+1-i}, y_{n+1-i}) \quad (7.94)$$

Porównując współczynniki A_{p+1} (patrz ostatnia kolumna w tabl. 7/2), od których zależy błąd metody, widzimy, że współczynniki A_j dla wzorów interpolacyjnych ($b_0 \neq 0$) są znacznie mniejsze co do wartości bezwzględnej niż odpowiednie współczynniki dla wzorów ekstrapolacyjnych ($b_0 = 0$). I jest to ogólna zasada. Ponadto obszar stabilności, a zwłaszcza obszar absolutnej stabilności można wybrać większy dla wzorów interpolacyjnych. Ale przy stosowaniu wzorów interpolacyjnych musimy na każdym etapie całkowania rozwiązać równanie nieliniowe (jeśli układ (7.1) jest nieliniowy ze względu na y), podczas gdy posługiwanie się wzorami ekstrapolacyjnymi wymaga jedynie podstawienia do wzoru.

Przedstawimy teraz metodę opartą zarówno na wzorach interpolacyjnych, jak i ekstrapolacyjnych, w której wykorzystuje się zalety obydwu typów tych wzorów. Aby wyznaczyć y_{n+1} z (7.94) wtedy, gdy $b_0 \neq 0$, trzeba na ogół zastosować metodę iteracyjną. Najczęściej stosowana jest metoda iteracji prostej albo metoda Newtona (patrz p. 3.4.1, 3.4.2).

W naszym przypadku metoda iteracji prostej wyraża się wzorem

$$y_{n+1}^{(j+1)} = h b_0 f(x_{n+1}, y_{n+1}^{(j)}) + \omega_n \quad (7.95)$$

a metoda Newtona ma postać

$$y_{n+1}^{(j+1)} = y_n^{(j)} + \left[I - h b_0 \frac{\partial f}{\partial y}(x_n, y_n^{(j)}) \right]^{-1} (y_n^{(j)} - h b_0 f(x_n, y_n^{(j)}) - \omega_n) \quad (7.96)$$

gdzie

$$\omega_n = \sum_{i=1}^k (a_i y_{n+1-i} + h b_i f(x_{n+1-i}, y_{n+1-i})) \quad (7.97)$$

Stosując metodę iteracyjną powinniśmy być pewni, że ciąg $(y_{n+1}^{(j)})$ jest zbieżny do rozwiązania y_{n+1} . Zauważmy, że wprowadzone w p.7.3.1 ograniczenie $h|b_0|L < 1$ na krok całkowania h zapewnia zbieżność tego procesu w przypadku metody iteracji prostej.

Istotną częścią obliczeń w każdej iteracji jest obliczanie wartości $f(x, y)$, a w przypadku metody Newtona dodatkową trudnością jest konieczność obliczenia macierzy $\left[I - h b_0 \frac{\partial f}{\partial y}(x_n, y_n^{(j)}) \right]^{-1}$. Aby uniknąć konieczności wykonywania dużej liczby iteracji, należy wartość $y_{n+1}^{(0)}$ dobrać możliwie bliską wartości rozwiązania y_{n+1} . Okazało się, że $y_{n+1}^{(0)}$ najlepiej jest wyznaczać ze wzoru ekstrapolacyjnego, zwanego wtedy *wzorem wstępnym*. Wzór inter-

polacyjny, za pomocą którego poprawia się $y_{n+1}^{(0)}$, nazywamy wzorem korygującym.

Metodę polegającą na połączeniu wzorów ekstrapolacyjnych i interpolacyjnych w opisany powyżej sposób nazywamy *metodą ekstrapolacyjno-interpolacyjną* (w języku angielskim *predictor-corrector method*).

Ważnym zagadnieniem przy stosowaniu metod ekstrapolacyjno-interpolacyjnych jest wybór momentu przerwania iteracji. Zbyt wczesne przerwanie iterowania wzoru korygującego może spowodować, że otrzymamy wynik o dużo mniejszej dokładności niż spodziewaliśmy się. Jeżeli iteracje są kontynuowane zbyt długo, to wydłuża się niepotrzebnie czas rozwiązania. Podobnie jak w przypadku metod Rungego-Kutty dokładność wyznaczanego rozwiązania bada się szacując błąd metody, przy czym iteracje należy przerwać wtedy, gdy błędy spowodowane przez zakończenie iterowania wzoru korygującego są małe w porównaniu z błędem metody tego wzoru. Nie będziemy omawiać jak szacować błąd spowodowany zakończeniem iteracji, ponieważ w praktyce przy stosowaniu wzorów interpolacyjno-ekstrapolacyjnych wzór korygujący iterujemy pewną ustaloną liczbą razy — najczęściej 1 do 3 razy. Ale wtedy powstaje problem stabilności metody ponieważ zależy ona nie tylko od wzoru korygującego, ale także i od wzoru wstępnego. Aby wykazać, że tak jest, rozpatrzmy następującą metodę

$$y_{n+1} = y_n + hf(x_n, y_n) \text{ wzór wstępny} \quad (7.98)$$

$$y_{n+1} = y_n + \frac{h}{2}[f(x_{n+1}, y_{n+1}) + f(x_n, y_n)] \text{ wzór korygujący}$$

Jak łatwo sprawdzić, przedziały absolutnej stabilności tych wzorów wynoszą odpowiednio $(-2; 0)$, $(-\infty; 0)$. Jeżeli w tej metodzie będziemy wykonywać zawsze jedną iterację prostą (7.95) wzoru korygującego, to otrzymamy metodę

$$y_{n+1}^{(0)} = y_n + hf(x_n, y_n) \quad (7.99)$$

$$y_{n+1} = y_{n+1}^{(1)} = y_n + \frac{h}{2}[f(x_{n+1}, y_{n+1}^{(0)}) + f(x_n, y_n)]$$

czyli metodę Rungego-Kutty drugiego rzędu, której stabilność rozpatrywaliśmy w p. 7.4.2; przedział absolutnej stabilności tej metody wynosi $(-2; 0)$. Gdy zaś wykonywać będziemy zawsze dwie takie iteracje, to otrzymamy metodę

$$y_{n+1}^{(0)} = y_n + hf(x_n, y_n)$$

$$y_{n+1}^{(1)} = y_n + \frac{h}{2}[f(x_{n+1}, y_{n+1}^{(0)}) + f(x_n, y_n)] \quad (7.100)$$

$$y_{n+1} = y_{n+1}^{(2)} = y_n + \frac{h}{2}[f(x_{n+1}, y_{n+1}^{(1)}) + f(x_n, y_n)]$$

Równanie charakterystyczne w tym przypadku ma postać

$$\mu = 1 + h\lambda + \frac{1}{2}h^2\lambda^2 + \frac{1}{4}h^3\lambda^3$$

a przedział absolutnej stabilności w dalszym ciągu wynosi $(-2; 0)$.

7.5.1. Metoda Hamminga

Rozpatrzmy obecnie metodę ekstrapolacyjno-interpolacyjną opartą na wzorach Milne'a i Hamminga (wzory 5 i 10 w tablicy 7/2). Stosując metodę iteracji prostej do wzoru korygującego, otrzymujemy następujące wzory¹⁾

$$y_{n+1}^{(0)} = y_{n-3} + \frac{4}{3}h(2y'_n - y'_{n-1} + 2y'_{n-2}) \text{ — wzór wstępny} \quad (7.101)$$

$$y_{n+1}^{(j+1)} = \frac{1}{8}(9y_n - y_{n-2}) + \frac{3}{8}h((y_{n+1}^{(j)})' + 2y'_n - y'_{n-1}) \quad (7.102)$$

$(j = 0, 1, \dots)$ wzór korygujący

Rozwiązanie dokładne spełnia równania

$$Y_{n+1} = Y_{n-3} + \frac{4}{3}h(2Y'_n - Y'_{n-1} + 2Y'_{n-2}) + T_n^0 \quad (7.103)$$

$$Y_{n+1} = \frac{1}{8}(9Y_n - Y_{n-2}) + \frac{3}{8}h(Y'_{n+1} + 2Y'_n - Y'_{n-1}) + T_n \quad (7.104)$$

gdzie T_n^0 , T_n są błędami metody odpowiednio dla wzoru wstępnego i korygującego.

Nietrudno jest udowodnić, że błędy T_n^0 , T_n można przedstawić w postaci

$$T_n^0 = \frac{14}{45}h^5 Y^{(5)}(\xi_1), \quad T_n = -\frac{1}{40}h^5 Y^{(5)}(\xi_2) \quad (7.105)$$

gdzie

$$\xi_1 \in (x_{n-3}; x_{n+1}), \quad \xi_2 \in (x_{n-2}; x_{n+1})$$

Wprowadźmy oznaczenia: $\varepsilon_n = Y_n - y_n$, $\varepsilon'_n = Y'_n - y'_n$. Po odjęciu stronami wzoru (7.103) od (7.101) oraz wzoru (7.104) od wzoru (5) z tablicy 7/2, otrzymujemy

$$y_{n+1}^{(0)} = Y_{n+1} - \varepsilon_{n-3} - \frac{4}{3}h(2\varepsilon'_n - \varepsilon'_{n-1} + 2\varepsilon'_{n-2}) - T_n^0$$

$$y_{n+1} = Y_{n+1} - \frac{1}{8}(9\varepsilon_n - \varepsilon_{n-2}) - \frac{3}{8}h(\varepsilon'_{n+1} + 2\varepsilon'_n - \varepsilon'_{n-1}) - T_n$$

¹⁾ Ze względu na rozbudowaną postać wzorów omawianych w tym punkcie dla uproszczenia zapisu będziemy zamiast $f(x_i, y_i)$, $f(x_i, Y_i)$ pisać odpowiednio y'_i , Y'_i .

Odejmując stronami od siebie dwa ostatnie wzory, ostatecznie znajdujemy

$$y_{n+1} - y_{n+1}^{(0)} = -\frac{1}{8}(9\varepsilon_n - \varepsilon_{n-2}) + \varepsilon_{n-3} - h\left(\frac{3}{8}\varepsilon'_{n+1} + \frac{41}{12}\varepsilon'_n - \frac{41}{24}\varepsilon'_{n-1} + \frac{3}{8}\varepsilon'_{n-2}\right) + T_n^0 - T_n$$

Zakładając, że

- 1) ε_i zmieniają się wolno z kroku na krok
- 2) $h\varepsilon'_i$ są małe w porównaniu z błędem metody
- 3) $Y^{(5)}$ mało się zmienia między ξ_1 i ξ_2

otrzymamy

$$y_{n+1} - y_{n+1}^{(0)} \approx \frac{121}{360}h^5 Y^{(5)}(\xi) \quad (7.106)$$

Korzystając ze wzoru (7.105), możemy wyznaczyć błąd metody dla wzoru wstępnego i korygującego

$$T_n^0 \approx \frac{112}{121}(y_{n+1} - y_{n+1}^{(0)}), \quad T_n = -\frac{9}{121}(y_{n+1} - y_{n+1}^{(0)}) \quad (7.107)$$

Zrobimy teraz jeszcze jedno założenie. Założymy mianowicie, że różnica $(y_{n+1} - y_{n+1}^{(0)})$ mało się zmienia z kroku na krok. Mamy więc

$$T_n^0 \approx \frac{112}{121}(y_n - y_n^{(0)})$$

Zauważmy, że $y_{n+1}^{(0)}$ obliczyliśmy w ten sposób, że reszta T_n^0 została odrzucona we wzorze (7.103), zatem rozwiązanie

$$\bar{y}_{n+1}^{(0)} = y_{n+1}^{(0)} + \frac{112}{121}(y_n - y_n^{(0)}) \quad (7.108)$$

powinno lepiej przybliżać Y_{n+1} i tak na ogół jest.

Sposób wyznaczania rozwiązania taki, że:

$y_{n+1}^{(0)}$ obliczamy ze wzoru (7.101)

$\bar{y}_{n+1}^{(0)}$ obliczamy ze wzoru (7.108)

$\bar{y}_{n+1}^{(0)}$ poprawiamy iteracyjnie za pomocą wzoru korygującego (7.102), nazywamy *metodą Hamminga*.

Na ogół już po pierwszym zastosowaniu wzoru korygującego rozwiązanie $y_{n+1}^{(1)}$ wystarczająco dobrze przybliża rozwiązanie y_{n+1} . Ponadto rozwiązanie wzoru korygującego możemy poprawić, ponieważ umiemy obliczyć (oczywiście w przybliżeniu) błąd T_n . Metodę Hamminga można zatem stosować do obliczeń na maszynach cyfrowych w następujący sposób:

$$y_{n+1}^{(0)} = y_{n-3} + \frac{4}{3}h(2y'_n - y'_{n-1} + 2y'_{n-2}) \text{ — wzór wstępny}$$

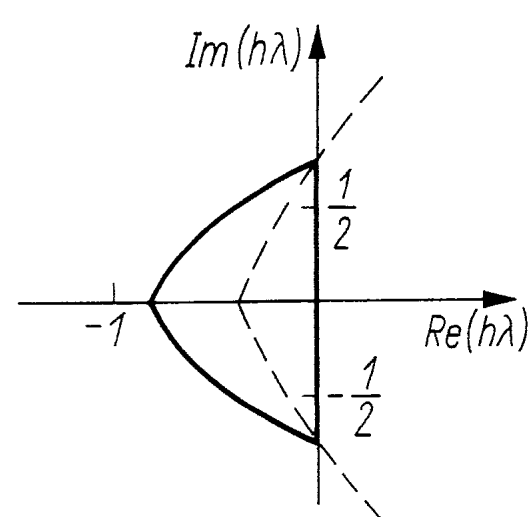
$$\bar{y}_{n+1}^{(0)} = y_{n+1}^{(0)} + \frac{112}{121}(y_n - y_n^{(0)}) \text{ — wzór modyfikujący}$$

$$(\bar{y}_{n+1}^{(0)})' = f(x_{n+1}, \bar{y}_{n+1}^{(0)}) \quad (7.109)$$

$$\bar{y}_{n+1} = \frac{1}{8}(9y_n - y_{n-2}) + \frac{3}{8}h((\bar{y}_{n+1}^{(0)})' + 2y_n' - y_{n-1}') \text{ — wzór korygujący}$$

$$y_{n+1} = \bar{y}_{n+1} - \frac{9}{121}(\bar{y}_{n+1} - y_{n+1}^{(0)}) \text{ — wzór końcowy}$$

Na rysunku 7.4 pokazano obszary absolutnej i względnej stabilności powyższej metody (linia ciągła oznacza brzeg obszaru absolutnej stabilności, przerywana — względnej stabilności).



7.4

Podamy teraz ograniczenia nałożone na krok całkowania h dla metody Hamminga, wynikające ze zbieżności metody iteracji prostej i stabilności. Warunek zbieżności metody iteracji prostej w zastosowaniu do wzoru Hamminga wynosi $|h\lambda_j| \leq \frac{8}{3}$, gdzie λ_j są wartościami własnymi macierzy Jacobiego $\left[\frac{\partial f(x_n, y_n)}{\partial y} \right]$, przy czym dla uzyskania natychmiastowej poprawy dokładności przybliżenia początkowego $y_n^{(0)}$ wartości $|h\lambda_j|$ powinny być znacznie mniejsze niż $\frac{8}{3}$, np. $\frac{1}{10} \cdot \frac{8}{3} \approx 0,26$. Ze względu na stabilność wartości $h\lambda_j$ powinny należeć do obszaru absolutnej stabilności, gdy $\text{Re } \lambda_j < 0$ oraz do obszaru względnej stabilności, gdy $\text{Re } \lambda_j > 0$, co jak widać z rys. 7.4 pokrywa się z poprzednim warunkiem.

Mając wzory wstępne i korygujące tego samego rzędu możemy zawsze wyznaczyć odpowiednie wzory modyfikujące i końcowe w opisany powyżej sposób.

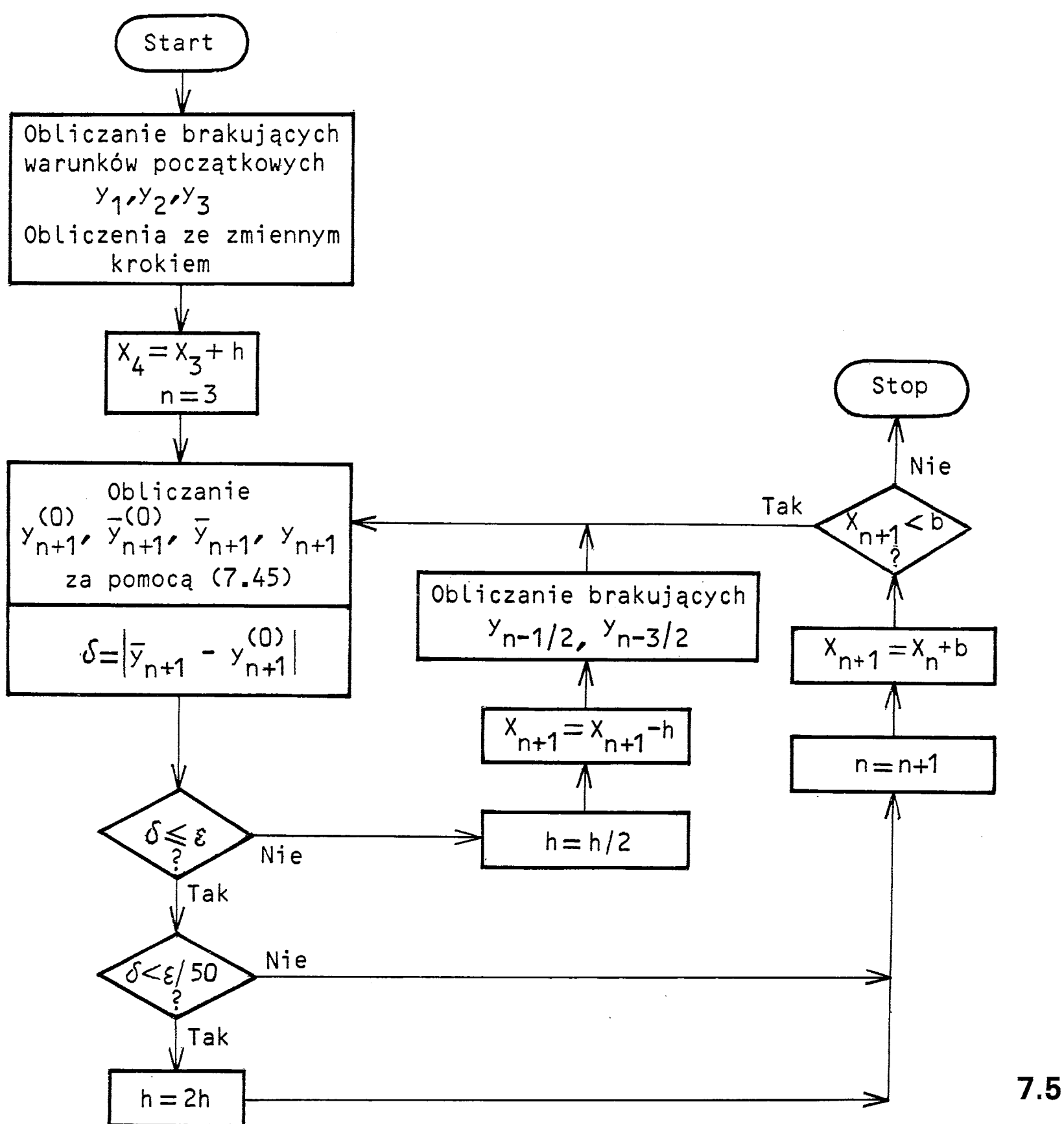
Jeśli wzory wstępne i korygujące są rzędu czwartego, to przy stosowaniu metod ekstrapolacyjno-interpolacyjnych wystarczy na ogół liczyć dwa razy $f(x, y)$ na każdym etapie: po modyfikacji i po wyznaczeniu rozwiązania. Rozpatrywane w p. 7.4.3 metody Rungego-Kutty czwartego rzędu, wymagały liczenia 4 razy wartości $f(x, y)$ na każdym etapie, a więc metody ekstrapolacyjno-interpolacyjne rzędu czwartego są od nich prawie dwukrotnie szybsze.

Wybór kroku całkowania. Idea doboru kroku całkowania w metodach ekstrapolacyjno-interpolacyjnych jest taka sama, jak dla metod Rungego-Kutty (patrz p. 7.4.4). Opiszemy dla przykładu działanie standardowej procedury bibliotecznej opartej na metodzie Hamminga (7.109).

Krok początkowy h i górne oszacowanie błędu metody ε są tu podawane przez użytkownika. Brakujące wielkości y_1, y_2, y_3 wyznaczamy za pomocą metody Rungego-Kutty rzędu czwartego. Przy ich wyznaczaniu stosowany jest zmienny krok całkowania. Pozostałe wartości $y_4, \dots$ są liczone metodą Hamminga. Krok, z jakim ostatecznie zostały wyznaczone warunki początkowe, jest przyjmowany jako krok początkowy h_0 w metodzie Hamminga.

Założmy, że mamy już obliczone rozwiązanie w punkcie x_n . Obliczenia na następnym etapie (tzn. przy przejściu od x_n do x_{n+1}) są następujące. Niech h oznacza krok całkowania z poprzedniego etapu obliczeń (gdy $n = 3$, to $h = h_0$). Za miarę błędu metody na tym etapie przyjmiemy liczbę

$$\delta = \|y_{n+1} - y_{n+1}^{(0)}\|$$



Wówczas, gdy:

a) $\delta \leq \varepsilon$, obliczone rozwiązanie y_{n+1} jest uznawane za wystarczająco dokładne, przy czym jeśli dodatkowo $\delta < \varepsilon/50$, to krok h podwajamy,

w przeciwnym przypadku krok nie ulega zmianie i procedura przechodzi do następnego etapu;

b) $\delta > \varepsilon$, obliczone rozwiązanie y_{n+1} jest uznawane za niewystarczająco dokładne, krok całkowania jest połówiony i obliczenia są wykonywane jeszcze raz z nowym krokiem.

Ponieważ metoda Hamminga wymaga stałego kroku, więc do wykonania powtórnych obliczeń może brakować pewnych wartości y . Brakujące wartości wyznaczamy najczęściej metodą Rungego-Kutty (w tym przypadku czwartego rzędu) albo za pomocą wzorów interpolacyjnych.

Jeżeli wybrany w powyższy sposób krok h (wynika on z szacowania błędu metody) nie zapewnia stabilności, to należy go zmniejszyć tak, aby zapewnić stabilność.

Na rysunku 7.5 podano schemat działania opisanej w tym punkcie procedury opartej na metodzie Hamminga (bez uwzględnienia ograniczeń wynikających ze stabilności).

Ćwiczenia

1. Wyprowadzić wzór modyfikujący dla wzorów 4 i 9, zamieszczonych w tablicy 7.3.

7.5.2. Metoda Geara dla układów typu stiff

Układy równań różniczkowych $y' = f(x, y)$ nazywamy *układami typu stiff* (*układami źle uwarunkowanymi*), gdy stosunek największej co do modułu wartości własnej macierzy Jacobiego $\left[\frac{\partial f_i}{\partial y_j} \right]$ do najmniejszej co do modułu wartości własnej jest znacznie większy od jedności. Z takimi układami spotykamy się przy opisie wielu zagadnień, np. przy omawianiu chemicznie reaktywnego gazu, chemicznych reakcji egzotermicznych, zachodzących w reaktorze rurowym, w dynamice procesów i sterowania, w teorii obwodów.

Trudności numeryczne jakie powstają przy wyznaczaniu rozwiązania układów typu stiff omówimy na przykładzie układów liniowych $y' = \Lambda y$. Rozpatrzmy następujące zagadnienie początkowe

$$y' = \Lambda y, \quad \Lambda = \begin{bmatrix} -667 & 333 \\ 666 & -334 \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad y(0) = \begin{bmatrix} 0 \\ 3 \end{bmatrix} \quad (7.110)$$

Mamy tu $\lambda_1 = -1$, $\lambda_2 = -1000$ oraz

$$\begin{aligned} y_1(x) &= e^{-x} - e^{-1000x} \\ y_2(x) &= 2e^{-x} + e^{-1000x} \end{aligned} \quad (7.111)$$

W przykładzie tym wartości własne λ_1, λ_2 są ujemne, więc rozwiązanie $y_1(x), y_2(x)$ maleje. Wyznaczając to rozwiązanie za pomocą metod różnicowych należy zatem wybrać h takie, aby $\lambda_i h$ należało do obszaru absolutnej stabilności. Na przykład dla metody Eulera mamy $|1 + \lambda_i h| \leq 1$ i stąd

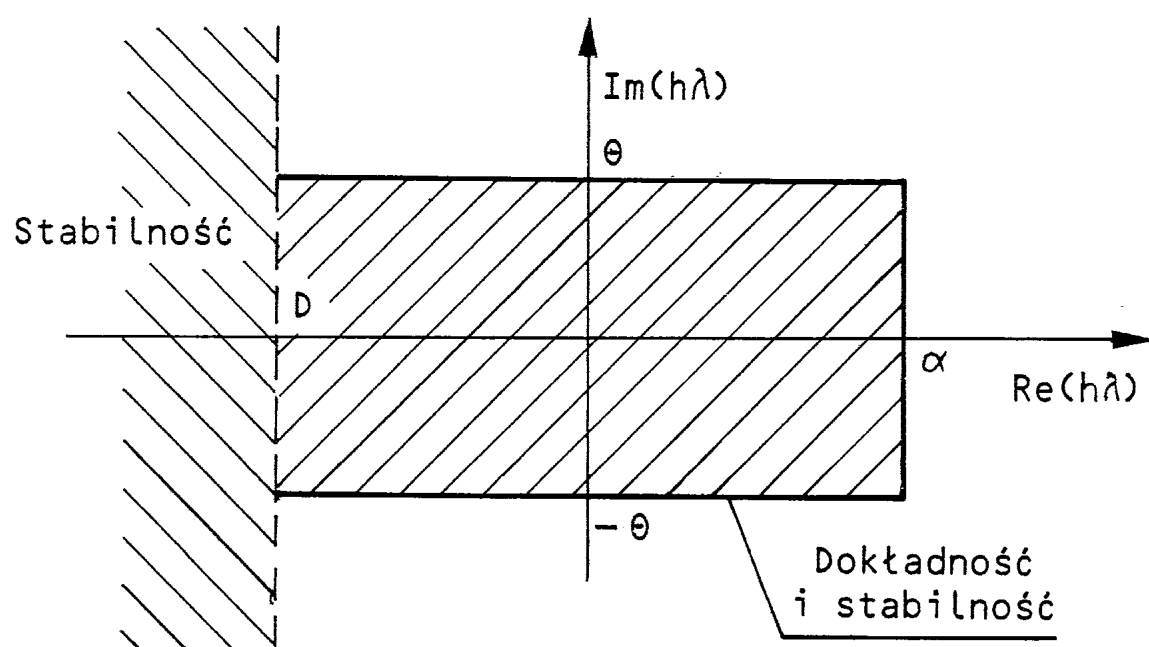
otrzymujemy $h \leq \frac{1}{500}$. Podobnie silne ograniczenie na h otrzymamy, gdy zastosujemy metodę Hamminga (patrz rys. 7.5). Zauważmy, że silne ograniczenie na h daje najmniejsza wartość własna $\lambda_2 = -1000$.

Błąd metody różnicowej, jak wiadomo, zależy od h i od pochodnych rozwiązania. Zauważmy, że części rozwiązania (7.111) odpowiadającej wartości własnej $\lambda_2 = -1000$ możemy nie rozpatrywać, gdy $\lambda_2 x \ll 0$, czyli praktycznie dla wszystkich dodatnich x . Część ta jest praktycznie równa zeru i nie ma wpływu na rozwiązanie. Taka sama sytuacja wystąpi w przypadku pochodnych rozwiązania (7.111). Stąd wyciągamy wniosek, że λ_2 nie ma praktycznego wpływu na błąd metody.

Z podobną sytuacją spotykamy się w przypadku dowolnego układu równań liniowych $y' = Ay$, gdy wszystkie wartości własne λ_i macierzy A mają ujemne części rzeczywiste. Wielkości $\lambda_i h$ powinny wówczas należeć do obszaru absolutnej stabilności, skąd otrzymujemy silne ograniczenia na h , jeżeli obszar absolutnej stabilności jest mały, jak np. dla metod Eulera lub Hamminga. O wielkości ograniczenia na h decydują wartości własne o najmniejszych częściach rzeczywistych, które nie mają wpływu na błąd metody.

W dalszym ciągu rozważymy przypadek, w którym wszystkie wartości własne λ_i mają ujemne części rzeczywiste, ponieważ wtedy problem wyboru h zapewniającego stabilność jest trudniejszy. Zatem metody Eulera i Hamminga nie nadają się do wyznaczenia rozwiązania liniowego układu typu stiff, bowiem silne ograniczenia na h wynikają wtedy jedynie z faktu, że obszar absolutnej stabilności tych metod jest mały. Do rozwiązywania takich zagadnień najlepiej byłoby zastosować metodę A -stabilną, tzn. metodę absolutnie stabilną dla każdego h , gdy $\operatorname{Re} \lambda < 0$. Wtedy nie otrzymalibyśmy żadnych ograniczeń na h wynikających ze stabilności. Ale udowodniono [15], że rząd A -stabilnej metody nie może przekraczać dwóch. Powstała zatem konieczność opracowania specjalnych metod dla układów typu stiff.

Jako alternatywę A -stabilności Gear (patrz [26]) wprowadził pojęcie stiff-stabilności. Mówimy, że metoda różnicowa (7.17) jest *stiff-stabilna* na płaszczyźnie zespolonej ($h\lambda$), jeżeli jest ona absolutnie stabilna w półpłaszczyźnie $\operatorname{Re}(h\lambda) < D$ oraz jest ona jednocześnie stabilna i dokładna w prostokącie $D < \operatorname{Re}(h\lambda) < \alpha$, $|\operatorname{Im}(h\lambda)| < \theta$ (rys. 7.6).



7.6

Równania (7.48) zapiszemy w postaci

$$\varrho(\mu) = h\lambda \cdot \sigma(\mu) \quad (7.112)$$

$$\text{gdzie } \varrho(\mu) = \mu^k - \sum_{i=1}^k a_i \mu^{k-i}, \sigma(\mu) = b_0 \mu^k + \sum_{i=1}^k b_i \mu^{k-i}$$

Dla stiff-stabilnej metody współczynnik b_0 jest różny od zera. W przeciwnym przypadku, wielomian $\sigma(\mu)$ byłby wielomianem niższego stopnia niż $\varrho(\mu)$ i wśród pierwiastków równania (7.112) znajdowałby się taki pierwiastek μ , że $|\mu| \rightarrow \infty$, gdy $|h\lambda| \rightarrow \infty$.

Gear poszukiwał stiff-stabilnej metody różnicowej rzędu k o postaci

$$y_{n+1} = \sum_{i=1}^k a_i y_{n+1-i} + hb_0 f(x_{n+1}, y_{n+1}) \quad (7.113)$$

Współczynniki a_i oraz b_0 wyznaczamy tak, aby metoda (7.113) była rzędu k . Należy więc rozwiązać układ równań

$$A_i = 0, \quad j = 0, 1, \dots, k \quad (7.114)$$

gdzie A_i są określone wzorami (7.19).

Dla przykładu rozpatrzmy przypadek $k = 1$. Układ równań (7.114) ma postać

$$\begin{aligned} A_0 &= 1 - a_1 = 0 \\ A_1 &= 1 - b_0 = 0 \end{aligned}$$

A zatem dla $k = 1$ otrzymujemy następujący wzór różnicowy

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}) \quad (7.115)$$

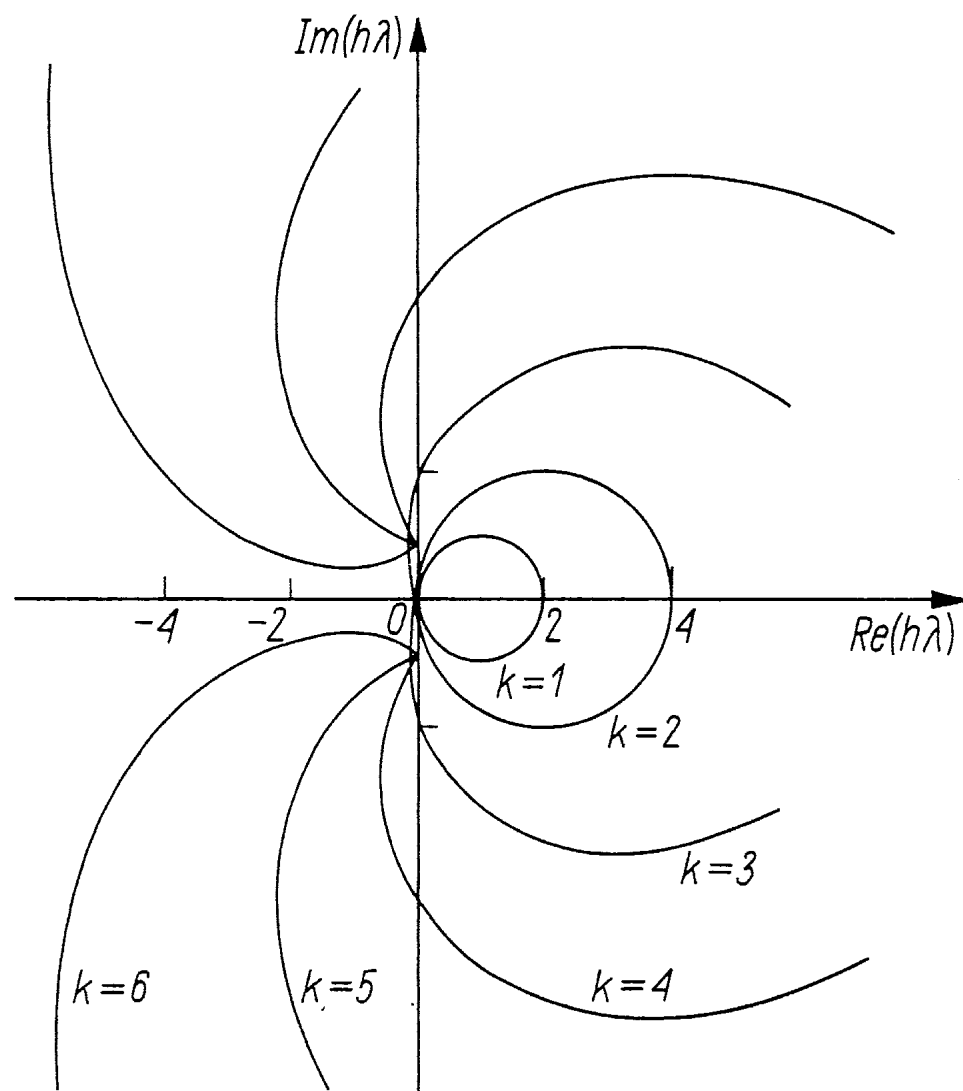
Gear wyznaczył współczynniki a_i , b_0 dla $k \leq 6$ i udowodnił, że są to stiff-stabilne metody. Obszar absolutnej stabilności wyznaczono w następujący sposób. Na płaszczyźnie $h\lambda$ zaznaczono zbiór punktów, w których pierwiastki μ_i równania (7.112) mają moduł równy jedności. Jest to krzywa o równaniu

$$h\lambda = \frac{\varrho(e^{i\theta})}{\sigma(e^{i\theta})}, \quad 0 \leq \theta \leq 2\pi \quad (7.116)$$

Na rysunku 7.7 podano wykresy takich krzywych dla $k \leq 6$.

Obszar znajdujący się na zewnątrz krzywej (7.116) jest obszarem absolutnej stabilności. Istotnie, dla żadnego z punktów $h\lambda$ tego obszaru pierwiastki μ_i nie mogą mieć modułu równego jedności. Ponadto, gdy $|h\lambda| \rightarrow \infty$, to wszystkie pierwiastki μ_i dążą do zera, ponieważ $\sigma(\mu) = \mu^k$. Pierwiastki μ_i są funkcjami ciągłymi ($h\lambda$), więc na zewnątrz krzywej (7.116) muszą być one co do modułu mniejsze od jedności.

Dla metody (7.115) równanie krzywej (7.116) ma postać $h\lambda = 1 + e^{-i\theta}$; jest to równanie okręgu o środku w punkcie 1 i promieniu 1.



7.7

W tablicy 7/4 podano współczynniki a_i, b_0 dla $k \leq 6$; Dla $k = 7, \dots, 12$ opisane metody nie są stiff-stabilne.

Jeśli układ równań różniczkowych jest typu stiff, to do wzorów interpolacyjnych nie stosujemy metody iteracji prostej, ponieważ ograniczenia na h , wynikające z konieczności zapewnienia szybkiej zbieżności, są wówczas bardzo silne

$$|hb_0 \lambda_i| \leq 1$$

7/4

k	1	2	3	4	5	6
b_0	1	$\frac{2}{3}$	$\frac{6}{11}$	$\frac{12}{25}$	$\frac{60}{137}$	$\frac{60}{147}$
a_1	1	$\frac{4}{3}$	$\frac{18}{11}$	$\frac{48}{25}$	$\frac{300}{137}$	$\frac{360}{147}$
a_2		$-\frac{1}{3}$	$-\frac{9}{11}$	$-\frac{36}{25}$	$-\frac{300}{137}$	$-\frac{450}{147}$
a_3			$\frac{2}{11}$	$\frac{16}{25}$	$\frac{200}{137}$	$\frac{400}{147}$
a_4				$-\frac{3}{25}$	$-\frac{75}{137}$	$-\frac{225}{147}$
a_5					$\frac{12}{137}$	$\frac{72}{147}$
a_6						$-\frac{10}{147}$

Należy wtedy stosować metodę Newtona, bowiem jak wykazała praktyka stosując tę metodę uzyskujemy szybką zbieżność przy słabszych ograniczeniach na h .

Na podstawie przedstawionych wyżej wzorów (dla $k \leq 6$) Gear opracował metodę ekstrapolacyjno-interpolacyjną automatycznego całkowania układów $y' = f(x, y)$. Obliczenia rozpoczynają się dla $k = 1$, przy czym w trakcie obliczeń może się zmieniać zarówno k jak i h . Celem tych zmian jest minimalizacja liczby działań przy zapewnieniu odpowiedniej dokładności obliczeń. Metody tej nie będziemy omawiać. Czytelnika interesującego się tą problematyką odsyłamy do prac [26, 27].

8

METODY ROZWIĄZYWANIA ZAGADNIEŃ BRZEGOWYCH DLA RÓWNAŃ RÓŻNICZKOWYCH CZĄSTKOWYCH

8.1. METODA RÓŻNICOWA DLA RÓWNANIA PRZEWODNICTWA CIEPLNEGO

8.1.1. Postawienie zagadnienia

Jako najprostszy przykład równania parabolicznego rozpatrzmy *równanie przewodnictwa cieplnego*

$$Lu = \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f(x, t) \quad (8.1)$$

Dla równania (8.1) typowe jest następujące zagadnienie. Znaleźć funkcję $u = u(x, t)$ spełniającą wewnątrz prostokąta $\bar{D} = \{(x, t) : 0 \leq x \leq l, 0 \leq t \leq T\}$ równanie (8.1), warunek początkowy

$$u(x, 0) = g(x), \quad 0 \leq x \leq l \quad (8.2)$$

i warunki brzegowe

$$u(0, t) = \varphi(t), \quad u(l, t) = \psi(t), \quad 0 \leq t \leq T \quad (8.3)$$

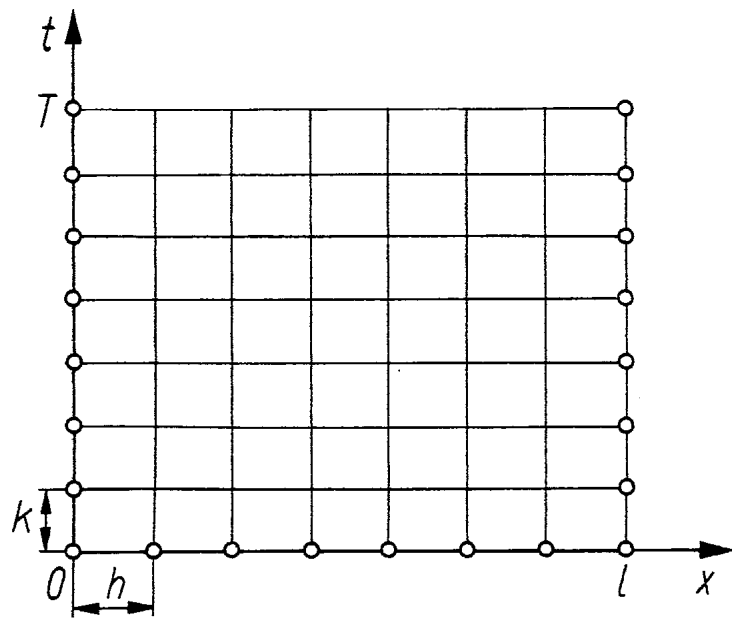
Zagadnienie powyższe nazywamy *zagadnieniem brzegowym pierwszego rodzaju* lub *zagadnieniem Dirichleta* dla równania (8.1). Zakładamy, że funkcja $f(x, t)$ jest ciągła na $\bar{D}$, funkcja $g(x)$ jest ciągła na przedziale $\langle 0; l \rangle$, a funkcje $\varphi(t)$, $\psi(t)$ są ciągłe na przedziale $\langle 0; T \rangle$.

8.1.2. Aproksymacja różnicowa

Przybliżonego rozwiązania zagadnienia (8.1)—(8.3) poszukujemy na zbiorze $\bar{D}$, który pokrywamy prostokątną siatką prostych

$$x = ih, \quad t = jk, \quad i, j \text{ — liczby całkowite} \quad (8.4)$$

gdzie $h = \frac{l}{N}$, $k = \frac{T}{M}$ (M, N — liczby całkowite) (rys. 8.1). Punkty przecięcia się prostych (8.4) nazywamy *węzłami siatki*, przy czym węzły leżące na prostych $x = 0$, $x = l$, $t = 0$ nazywamy *brzegowymi*, pozostałe zaś *wewnętrzne*. W dalszym ciągu rozpatrywać będziemy tylko te węzły, które należą do $\bar{D}$; zbiór tych węzłów oznaczmy przez $\bar{\omega}$.



8.1

Omówimy teraz przybliżenie pochodnych $\frac{\partial u}{\partial t}$, $\frac{\partial^2 u}{\partial x^2}$ wyrażeniami różnicowymi. Niech Λ oznacza *operator różnicowy* określony w następujący sposób

$$\Lambda u(x, t) = \frac{u(x + h, t) - 2u(x, t) + u(x - h, t)}{h^2} \quad (8.5)$$

Dla dostatecznie regularnej funkcji $u(x, t)$ mamy

$$\Lambda u(x, t) = \frac{\partial^2 u(x, t + k/2)}{\partial x^2} - \frac{k}{2} \frac{\partial^3 u(x, t + k/2)}{\partial x^2 \partial t} + O(h^2 + k^2) \quad (8.6)$$

$$\Lambda u(x, t + k) = \frac{\partial^2 u(x, t + k/2)}{\partial x^2} + \frac{k}{2} \frac{\partial^3 u(x, t + k/2)}{\partial x^2 \partial t} + O(h^2 + k^2)$$

$$\frac{u(x, t + k) - u(x, t)}{k} = \frac{\partial u(x, t + k/2)}{\partial t} + O(k^2) \quad (8.7)$$

Określimy z kolei *rodzinę operatorów różnicowych* $L_{h,k}^\sigma$, zależnych od parametru σ

$$\begin{aligned} L_{h,k}^\sigma u(x, t) &= \frac{u(x, t + k) - u(x, t)}{k} - (\sigma \Lambda u(x, t + k) + \\ &+ (1 - \sigma) \Lambda u(x, t)) \end{aligned} \quad (8.8)$$

Na podstawie wzorów (8.6) i (8.7) mamy

$$\begin{aligned} L_{h,k}^\sigma u(x, t) &= \frac{\partial u(x, t + k/2)}{\partial t} - \frac{\partial^2 u(x, t + k/2)}{\partial x^2} + \\ &+ \left(\frac{1}{2} - \sigma\right) k \frac{\partial^3 u(x, t + k/2)}{\partial x^2 \partial t} + O(h^2 + k^2) = \\ &= Lu(x, t + k/2) + \left(\frac{1}{2} - \sigma\right) k \frac{\partial^3 u(x, t + k/2)}{\partial x^2 \partial t} + O(h^2 + k^2) \end{aligned} \quad (8.9)$$

Przejdziemy teraz do opisu metody różnicowej dla zagadnienia (8.1)—(8.3), przy czym wprowadzimy następujące oznaczenia. Niech $u = u(x, t)$ oznacza rozwiązanie dokładne postawionego zagadnienia różniczkowego, a $v = v(x, t)$ rozwiązanie otrzymane metodą różnicową. Ponadto niech (i, j) oznacza węzeł siatki o współrzędnych (ih, jk) , a u_{ij} , v_{ij} , f_{ij} — wartości funkcji u , v , f w węźle (i, j) (z oznaczeń tych będziemy korzystać nie tylko przy omawianiu metody różnicowej dla równania przewodnictwa cieplnego, ale także przy opisie metody różnicowej dla równania struny i równania Poissona). Szukane rozwiązanie v wyznaczamy oczywiście w węzłach zbioru $\bar{\omega}$, przy czym w węzłach brzegowych będziemy korzystać z warunku początkowego (8.2) i warunków brzegowych (8.3); w węzłach $(i, 0)$, leżących na prostej $x = 0$, z warunku początkowego

$$v_{i,0} = g_i, \quad i = 1, \dots, N-1 \quad (8.10)$$

gdzie $g_i = g(ih)$, a w węzłach $(0, j)$, (N, j) — z warunków brzegowych

$$v_{0,j} = \varphi_j, \quad v_{N,j} = \psi_j, \quad j = 0, 1, \dots, M \quad (8.11)$$

gdzie $\varphi_j = \varphi(jk)$, $\psi_j = \psi(jk)$.

Dla każdego węzła wewnętrznego (i, j) rozpatrujemy równanie różnicowe zależne od parametru σ o postaci

$$L_{h,k}^\sigma v_{ij} = f_{i,j+1/2}, \quad i = 1, \dots, N-1; \quad j \geq 1 \quad (8.12)$$

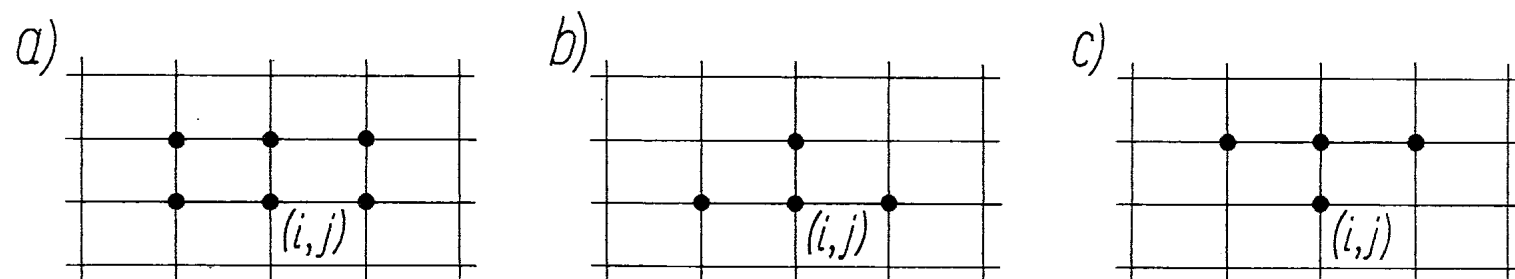
w którym $f_{i,j+1/2} = f(ih, jk + k/2)$. W ten sposób dla zagadnienia brzegowego (8.1)—(8.3) otrzymujemy układ równań różnicowych (8.10)—(8.12), zależnych od parametru σ . Układ ten jest układem algebraicznych równań liniowych, w których niewiadomymi są wartości funkcji v w węzłach zbioru $\bar{\omega}$, i w którym niewiadomych i równań jest tyle samo, co węzłów w zbiorze $\bar{\omega}$.

Zapiszemy teraz równania (8.12) w następującej postaci

$$\begin{aligned} -\alpha\sigma v_{i+1,j+1} + (1 + 2\alpha\sigma) v_{i,j+1} - \alpha\sigma v_{i-1,j+1} = \\ = \alpha(1 - \sigma) v_{i+1,j} + [1 - 2\alpha(1 - \sigma)] v_{ij} + \alpha(1 - \sigma) v_{i-1,j} + kf_{i,j+1/2} \end{aligned} \quad (8.13)$$

gdzie $\alpha = k/h^2$.

Dla $\sigma \neq 0$ i 1 powyższe równania wiążą ze sobą wartości v_{ij} w 6 węzłach, których wzajemne położenie jest takie, jak na rys. 8.2a. Dla $\sigma = 0$ lub $\sigma = 1$



8.2

równania te wiążą ze sobą wartości v_{ij} w 4 węzłach — ich wzajemne położenie jest pokazane na rys. 8.2b dla $\sigma = 0$, a na rys. 8.2c dla $\sigma = 1$. Jak widzimy, punkty te są węzłami z dwóch warstw: $t = jk$ i $t = (j + 1)k$ (przez warstwę $t = rk$ rozumiemy zbiór punktów postaci (ih, rk) , $i = 0, 1, \dots, N$).

Omówimy teraz problem istnienia i jednoznaczności rozwiązania v rozpatrywanego układu równań różnicowych oraz podamy sposób jego wyznaczania (jeśli istnieje i jest jednoznaczne). W węzłach brzegowych rozwiązanie v_{ij} wyznaczamy bezpośrednio ze wzorów (8.10), (8.11).

Dla $\sigma = 0$ na $(j + 1)$ -szej warstwie ($j \geq 0$) w węzłach wewnętrznych rozwiązanie otrzymamy bezpośrednio z równań (8.13)

$$v_{i,j+1} = \alpha v_{i+1,j} + (1 - 2\alpha) v_{ij} + \alpha v_{i-1,j} + f_{i,j+1/2}$$

Postępujemy tak kolejno warstwami dla $j = 0, 1, \dots, M - 1$. A zatem dla $\sigma = 0$ układ równań różnicowych ma jednoznaczne rozwiązanie i rozwiązanie to można wyznaczyć w bardzo prosty sposób.

W przypadku $\sigma \neq 0$, równania (8.13), $i = 1, 2, \dots, N - 1$, zapiszemy w postaci macierzowej

$$AV_{j+1} = F_{j+1} \quad (8.14)$$

gdzie

$$A = \begin{bmatrix} (1 + 2\alpha\sigma) & -\alpha\sigma & & & 0 \\ -\alpha\sigma & (1 + 2\alpha\sigma) & & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -\alpha\sigma \\ 0 & & & -\alpha\sigma & (1 + 2\alpha\sigma) \end{bmatrix}, \quad V_{j+1} = \begin{bmatrix} V_{1,j+1} \\ V_{2,j+1} \\ \\ V_{N-2,j+1} \\ V_{N-1,j+1} \end{bmatrix}$$

Dla $\sigma > 0$ macierz A jest macierzą silnie diagonalnie dominującą (tzn. moduły elementów na głównej przekątnej są większe od sumy modułów pozostałych elementów w tym samym wierszu). Na podstawie wniosku z twierdzenia Gershgorina (patrz p. 5.2) układy o takich macierzach mają jednoznaczne rozwiązanie. Zatem dla $\sigma > 0$ rozpatrywany układ równań różniczkowych (8.10)–(8.12) ma jednoznaczne rozwiązanie.

Dla $\sigma < 0$ macierz A nie jest macierzą silnie diagonalnie dominującą. Można udowodnić, że istnieją wartości $\sigma < 0$, dla których macierz A jest

macierzą osobliwą. Dla tych wartości σ układ równań różnicowych (8.10)–(8.12) nie ma jednoznacznego rozwiązania i takich wartości σ nie należy przyjmować. Do tego problemu powrócimy w następnym punkcie.

Dla $\sigma \neq 0$, jeżeli macierz A jest nieosobliwa rozwiązanie wyznaczamy kolejno warstwami z tym, że na każdej warstwie należy rozwiązać układ równań (8.14) o macierzy trójdzielnej. Układy te rozwiązujemy za pomocą metod omówionych w p. 5.3.6.

Niech $C_n^m(\bar{D})$ oznacza klasę funkcji, które na $\bar{D}$ mają m ciągłych pochodnych względem x i n ciągłych pochodnych względem t . Na podstawie wzoru (8.9) mamy wówczas

$$L_{h,k}^{1/2} u_{ij} = f_{i,j+1/2} + O(h^2 + k^2) \quad (8.15)$$

gdy $u \in C_3^4(\bar{D})$, oraz

$$L_{h,k}^\sigma u_{ij} = f_{i,j+1/2} + O(h^2 + k), \quad \sigma \neq 1/2 \quad (8.16)$$

gdy $u \in C_2^4(\bar{D})$. A zatem błąd zastąpienia równania różniczkowego (8.1) równaniami różnicowymi (8.13) dla $\sigma = 1/2$ wynosi $O(h^2 + k^2)$, gdy $u \in C_3^4(\bar{D})$, a dla $\sigma \neq 1/2$ błąd ten wynosi $O(h^2 + k)$, gdy $u \in C_2^4(\bar{D})$.

Najczęściej rozpatrywanymi w literaturze metodami różnicowymi dla zagadnienia (8.1)–(8.3), są metody dla $\sigma = 0, 1/2, 1$, przy czym gdy $\sigma = 0$, układ równań różnicowych jest najprostszy do rozwiązania, ale wtedy na siatkę należy nałożyć ograniczenie $k/h^2 \leq 1/2$ (patrz następny punkt), natomiast dla $\sigma = 1/2$ otrzymujemy najszybszą zbieżność rozwiązania różnicowego do rozwiązania dokładnego.

8.1.3. Stabilność i zbieżność

Stabilność. Zbadamy teraz problem stabilności względem prawej strony i warunku początkowego przedstawionej w poprzednim punkcie metody różnicowej, to znaczy zbadamy, jak zmiany funkcji $f(x, t)$ — występującej po prawej stronie równania różniczkowego (8.1) i funkcji $g(x)$ w warunku początkowym (8.2) — wpływają na rozwiązanie różnicowe v_{ij} .

Niech $v, \tilde{v}$ będą rozwiązaniami różnicowymi odpowiednio dla danych funkcji $f(x, t), g(x)$ i $\tilde{f}(x, t), \tilde{g}(x)$. Oznaczmy: $V = v - \tilde{v}$, $F = f - \tilde{f}$, $G = g - \tilde{g}$. Z liniowości równań różnicowych wynika, że wartości V_{ij} funkcji V spełniają równania

$$\begin{aligned} & -\alpha\sigma V_{i+1,j+1} + (1+2\alpha\sigma) V_{i,j+1} - \alpha\sigma V_{i-1,j+1} = \\ & = \alpha(1-\sigma) V_{i+1,j} + [1-2\alpha(1-\sigma)] V_{i,j} + \alpha(1-\sigma) V_{i-1,j} + kF_{i,j+1/2} \end{aligned}$$

$$\begin{aligned} i &= 1, 2, \dots, N-1; \quad j = 0, 1, \dots, M-1 \\ V_{i,0} &= G_i, \quad i = 1, 2, \dots, N-1 \\ V_{0,j} &= V_{N,j} = 0, \quad j = 0, 1, \dots, M \\ \text{gdzie } F_{i,j+1/2} &= F(ih, jk + k/2), \quad G_i = G(ih). \end{aligned} \quad (8.17)$$

Do prawej strony równania różniczkowego (8.1) oraz warunków początkowego i brzegowych, które będziemy określać wspólnym mianem jako dane wejściowe, wchodzi wielkości fizyczne, które na ogół są obarczone błędami. Błędy te mogą wynikać zarówno z pomiarów, jak i z uproszczeń przy układaniu modelu matematycznego badanego zjawiska. W następnym punkcie wykazemy, że dokładność rozwiązania różnicowego zależy od odstępów siatki h i k , i że dobre przybliżenie otrzymujemy dla dostatecznie gęstych siatek. Rozwiązanie różnicowe wyznacza się kolejno warstwami, więc do praktycznych obliczeń nie są przydatne takie metody, dla których zaburzenia funkcji f , g , φ , ψ , występujących w równaniach (8.1)–(8.3), powodują, że zmiany rozwiązania na kolejnych warstwach należących do zbioru $\bar{\omega}$ wzrastają nieograniczenie przy zagęszczaniu siatki w $\bar{D}$. W niniejszym rozdziale ograniczymy się do zbadania wpływu zmian funkcji f , g na rozwiązanie różnicowe, zakładając, że funkcje φ , ψ są dane dokładnie.

W celu określenia wielkości zmian rozwiązania na warstwie wprowadzamy normę $\|\cdot\|$ w klasie funkcji, których dziedziną są punkty $x_i = ih$, $i = 0, 1, \dots, N$.

Niech $Y = Y(x, t)$, a Y^r niech oznacza funkcję, której dziedziną są punkty x_i , przyjmującą wartości $Y(ih, rk)$ w punktach x_i , $i = 0, 1, \dots, N$. Zatem Y^r określa wielkość funkcji Y na warstwie $t = rk$.

Mówimy, że określona w poprzednim punkcie metoda różnicowa jest *stabilna względem prawej strony i warunku początkowego*, jeżeli istnieją stałe dodatnie A_1 , A_2 (niezależne od h , k) takie, że dla dowolnych funkcji G , F zachodzi

$$\|V^j\| \leq A_1 \|G\| + A_2 \cdot \max_{0 \leq s \leq j-1} \|F^{s+1/2}\|, \quad j = 1, \dots, M \quad (8.18)$$

gdzie V jest rozwiązaniem układu (8.17), a $\|\cdot\|$ jest normą (na warstwie) wprowadzoną wyżej.

Jeżeli powyższa własność zachodzi przy dodatkowym założeniu $F = f - \tilde{f} \equiv 0$ lub $G = g - \tilde{g} \equiv 0$, to taką metodę nazywamy odpowiednio metodą *stabilną względem warunku początkowego* lub metodą *stabilną względem prawej strony*.

Nierówności

$$\|V^j\| \leq A_1 \|G\| \quad (8.19)$$

$$\|V^j\| \leq A_2 \cdot \max_{0 \leq s \leq j-1} \|F^{s+1/2}\|, \quad j = 1, \dots, M \quad (8.20)$$

oznaczają więc odpowiednio stabilność względem warunku początkowego i stabilność względem prawej strony.

Zauważmy, że dla takich metod różnicowych stabilnych względem prawej strony i warunku początkowego małe zmiany funkcji g na warstwie $t = 0$ (mierzone wielkością normy na tej warstwie) oraz małe zmiany funkcji f na warstwach $t = (j+1/2)k$, $j = 0, 1, \dots, M-1$ (mierzone wielkością normy

na każdej warstwie) powodują małe zmiany rozwiązania różnicowego na wszystkich warstwach należących do zbioru $\bar{\omega}$ (mierzone wielkością normy na każdej warstwie) i zmiany te można oszacować niezależnie od wielkości h i k .

Zauważmy, że ze stabilności względem warunku początkowego lub względem prawej strony, gdy $\|\cdot\|$ jest dowolną normą (na warstwie), wynika jednoznaczność rozwiązania układu równań różnicowych.

Rozpatrzmy teraz normę określoną następująco

$$\|V\| = \|V\|_c = \max_{0 \leq i \leq N} |V(x_i)| \quad (8.21)$$

która jest różnicowym odpowiednikiem normy $\|V\| = \sup_{\langle 0; 1 \rangle} |V(x)|$. Korzystając z zasady maksimum udowodnimy nierówność

$$\|V^j\|_c \leq \|G\|_c + t_j \cdot \max_{0 \leq s \leq j-1} \|F^{s+1/2}\|, \quad j = 1, \dots, M \quad (8.22)$$

$$\text{dla } \sigma \geq 1 - \frac{h^2}{2k}, \quad 0 \leq \sigma \leq 1.$$

Rozpatrzmy w tym celu funkcje y, Y określone na zbiorze punktów x_i , $i = 0, 1, \dots, N$, przyjmując, że wartości tych funkcji w punkcie x_j wynoszą odpowiednio y_j, Y_j , $j = 0, 1, \dots, N$. Niech P będzie operatorem takim, że

$$Py_i = \alpha\sigma y_{i+1} - (1 + 2\alpha\sigma)y_i + \alpha\sigma y_{i-1} \quad i = 1, \dots, N-1, \quad \alpha\sigma > 0 \quad (8.23)$$

Tw. 1. (zasada maksimum). *Jeżeli $y \neq \text{const}$ oraz $Py_i \geq 0$ ($Py_i \leq 0$), to funkcja y nie może przyjmować największej dodatniej (najmniejszej ujemnej) wartości w punktach x_i , $i = 1, \dots, N-1$.*

DOWÓD. Przypuśćmy, że największą liczbą dodatnią jest jedna z wartości $y_1, \dots, y_{N-1}$. Z warunku $y \neq \text{const}$ wynika, że istnieje takie s , że $y_s = \max_{0 \leq i \leq N} y_i = M > 0$ oraz jedna z liczb y_{s-1}, y_{s+1} jest mniejsza od M . Mamy

$$\alpha\sigma y_{s+1} - (1 + 2\alpha\sigma)y_s + \alpha\sigma y_{s-1} < \alpha\sigma M - (1 + 2\alpha\sigma)M + \alpha\sigma M = -M < 0$$

co przeczy założeniu $Py_s \geq 0$. Dla drugiej części twierdzenia dowód jest analogiczny.

Tw. 2. *Jeżeli są spełnione nierówności*

$$\begin{aligned} |Py_i| &\leq -PY_i, \quad i = 1, \dots, N-1 \\ |y_0| &\leq Y_0, \quad |y_N| \leq Y_N \end{aligned}$$

to

$$|y_i| \leq Y_i, \quad i = 0, 1, \dots, N$$

DOWÓD. Z twierdzenia 1 wynika, że liczby $Y_i + y_i$ oraz $Y_i - y_i$ ($i = 1, \dots, N-1$) są nieujemne. A to już oznacza, że $|y_i| \leq Y_i$.

Tw. 3. *Jeżeli $PY_i = -H$ ($H \geq 0$), $i = 1, \dots, N-1$, $Y_0 = Y_N = 0$, to*

$$Y_i \leq H, \quad i = 0, 1, \dots, N \quad (8.24)$$

DOWÓD. Dla $i = 0, N$ nierówność (8.24) jest prawdziwa na podstawie założeń. Niech Y_s będzie największą wartością funkcji Y w punktach $x_i, i = 1, \dots, N-1$. Mamy

$$PY_s = \alpha\sigma(Y_{s+1} - Y_s) - \alpha\sigma(Y_s - Y_{s-1}) - Y_s = -H$$

a ponieważ $Y_{s+1} - Y_s \leq 0$ i $Y_s - Y_{s-1} \geq 0$, więc $Y_s \leq H$ co kończy dowód.

Obecnie przystąpimy do dowodu nierówności (8.22). Niech dla każdego $j = 0, 1, \dots, M-1$ wartości $\bar{v}_{ij}$ funkcji $\bar{v}$ spełniają równania

$$\alpha\sigma\bar{v}_{i+1,j+1} - (1 + 2\alpha\sigma)\bar{v}_{i,j+1} + \alpha\sigma\bar{v}_{i-1,j+1} = - \max_{0 \leq l \leq N} |V_{i,j}| - k \max_{0 \leq l \leq N} |F_{i,j+1/2}|$$

$$\bar{v}_{0,j} = \bar{v}_{N,j} = 0$$

Dla $\sigma \geq 1 - \frac{h^2}{2k}$, $0 \leq \sigma \leq 1$, na podstawie twierdzeń 2 i 3 mamy

$$|V_{i,j+1}| \leq \bar{v}_{i,j+1} \leq \max_{0 \leq l \leq N} |V_{i,j}| + k \max_{0 \leq l \leq N} |F_{i,j+1/2}| \quad (8.25)$$

Oszacowanie (8.25) zapiszemy w następującej postaci

$$\|V^{j+1}\|_c \leq \|V^j\|_c + k \|F^{j+1/2}\|_c$$

stąd otrzymujemy

$$\|V^{j+1}\|_c \leq \|V^0\|_c + k \sum_{s=0}^j \|F^{s+1/2}\|_c \leq \|G\|_c + t_{j+1} \max_{0 \leq s \leq j} \|F^{s+1/2}\|_c$$

co kończy dowód nierówności (8.22).

Z nierówności (8.22) wynika, że

$$\|V^j\|_c \leq \|G\|_c + T \max_{0 \leq s \leq j-1} \|F^{s+1/2}\|_c, \quad j = 1, \dots, M \quad (8.26)$$

Udowodniliśmy zatem stabilność względem prawej strony i warunku początkowego w normie $\|\cdot\|_c$ rozpatrywanej metody różnicowej dla przypadku, gdy parametr σ spełnia warunki $\sigma \geq 1 - \frac{h^2}{2k}$, $0 \leq \sigma \leq 1$. Gdy $\sigma = 0$,

wówczas $\alpha = \frac{k}{h^2} \leq 1/2$ — jest to ograniczenie nałożone na odstęp h, k siatki.

Podobnie dla $0 < \sigma < 1$ mamy $\alpha \leq \frac{1}{2(1-\sigma)}$, natomiast dla $\sigma = 1$ nie otrzymujemy żadnych ograniczeń na siatkę.

Korzystając z bardziej zaawansowanego aparatu matematycznego (teorii funkcji Greena, teorii potencjału), można udowodnić, że nierówność (8.22) zachodzi przy słabszych ograniczeniach nałożonych na σ, h i k (patrz np. [55]).

Dla metody różnicowej stabilnej względem prawej strony i warunku początkowego w normie $\|\cdot\|_c$ można na podstawie nierówności (8.18) wysnuć następujący wniosek.

Wniosek. *Jeżeli*

$$\max \left\{ \max_{\langle 0; l \rangle} |g(x) - \tilde{g}(x)|, \max_{\bar{D}} |f(x, y) - \tilde{f}(x, y)| \right\} < \delta \quad (8.27)$$

to istnieje stała A (niezależna od h, k) taka, że

$$\max_{(i,j) \in \bar{\omega}} |v_{ij} - \tilde{v}_{ij}| < A\delta \quad (8.28)$$

Istotnie, z (8.27) mamy

$$\|G\|_c < \delta, \quad \max_{0 \leq j \leq M-1} \|F^{j+1/2}\| < \delta$$

Stąd, na podstawie (8.18) otrzymujemy

$$\max_{0 \leq j \leq N} \|V^j\|_c < (A_1 + A_2) \delta$$

a więc możemy przyjąć, że $A = A_1 + A_2$.

Dla metody różnicowej stabilnej względem prawej strony i warunku początkowego w normie $\|\cdot\|_c$ małe co do wartości bezwzględnej zaburzenia prawej strony f równania (8.1) oraz funkcji g z warunku początkowego (8.2) powodują małe co do wartości bezwzględnej zmiany rozwiązania v w całym zbiorze $\bar{\omega}$.

Na podstawie nierówności (8.22) dla rozpatrzonego przez nas przypadku $\sigma \geq 1 - \frac{h^2}{2k}$, $0 \leq \sigma \leq 1$, mamy $A = 1 + T$.

Innym przykładem często rozpatrywanej normy jest norma $\|\cdot\|_{L_2}$, określona następująco

$$\|V\|_{L_2} = \sqrt{\langle V, V \rangle}, \quad \langle V, W \rangle = \sum_{i=1}^{N-1} h V(x_i) W(x_i) \quad (8.29)$$

Jest ona różnicowym odpowiednikiem normy $\|V\| = \sqrt{\int_0^1 V^2(x) dx}$. Zauważmy, że jeżeli $h \rightarrow 0$ oraz funkcja $V(x)$ jest ciągła na $\langle 0; l \rangle$, to $\|V\|_{L_2}$ dąży do $\sqrt{\int_0^1 V^2(x) dx}$.

Pokażemy, że ze stabilności metody różnicowej względem jej prawej strony i warunku początkowego w normie $\|\cdot\|_{L_2}$ nie wynika podany powyżej wniosek. Istotnie, z (8.27) i (8.18) mamy

$$\|G\|_{L_2} < \delta \sqrt{l}, \quad \max_{0 \leq j \leq M-1} \|F^{j+1/2}\|_{L_2} < \delta \sqrt{l}$$

$$\|V^j\|_{L_2} = \sqrt{\sum_{i=1}^{N-1} (v_{ij} - \tilde{v}_{ij})^2 h} < (A_1 + A_2) \delta \sqrt{l}$$

Z ostatniej nierówności nie wynika, że moduł różnicy $|V_{ij} - \tilde{V}_{ij}|$ musi być ograniczony na zbiorze $\bar{\omega}$ przez stałą niezależną od h dla dostatecznie małych h . Zatem dla metod stabilnych w normie $\|\cdot\|_{L_2}$ dla dostatecznie małych h małe

zmiany funkcji g i f mogą spowodować, że w pewnych węzłach rozwiązanie zmieni się znacznie, z tym, że suma $\sum_{i=1}^{N-1} (v_{ij} - \tilde{v}_{ij})^2 h$ będzie mała.

Przedstawimy teraz metodę rozdzielania zmiennych badania stabilności, prowadzącą do uzyskania oszacowania w normie $\|\cdot\|_{L_2}$. Zbadamy stabilność względem warunku początkowego, tzn. zbadamy zachowanie się rozwiązania V_{ij} równań (8.17) przy dodatkowym założeniu $F = 0$. Równania te przyjmą zatem postać

$$\begin{aligned} -\alpha\sigma V_{i+1,j+1} + (1+2\alpha\sigma) V_{i,j+1} - \alpha\sigma V_{i-1,j+1} = \\ = \alpha(1-\sigma) V_{i+1,j} + [1-2\alpha(1-\sigma)] V_{i,j} + \alpha(1-\sigma) V_{i-1,j} \end{aligned} \quad (8.30)$$

$$i = 1, \dots, N-1; \quad j = 0, 1, \dots, M-1$$

$$V_{i,0} = G_i, \quad i = 1, \dots, N-1 \quad (8.31)$$

$$V_{0,j} = V_{N,j} = 0, \quad j = 0, 1, \dots, M \quad (8.32)$$

W celu rozwiązania powyższego układu równań metodą rozdzielania zmiennych rozpatrzmy funkcje $V^{(s)}$

$$V_{ij}^{(s)} = \lambda_s^j \sin \frac{s\pi i}{N}, \quad s = 1, \dots, N-1 \quad (8.33)$$

gdzie λ_s jest stałą niezależną od i, j . Zauważmy, że każda z tych funkcji spełnia warunki brzegowe (8.32). Stałą λ_s dobieramy tak, aby wartości $V_{ij}^{(s)}$ spełniały równania (8.30). Po podstawieniu $V_{ij}^{(s)}$ do tych równań, mamy

$$\begin{aligned} -\alpha\sigma\lambda_s \left[\sin \frac{s\pi(i+1)}{N} - 2 \sin \frac{s\pi i}{N} + \sin \frac{s\pi(i-1)}{N} \right] + \lambda_s \sin \frac{s\pi i}{N} = \\ = \alpha(1-\sigma) \left[\sin \frac{s\pi(i+1)}{N} - 2 \sin \frac{s\pi i}{N} + \sin \frac{s\pi(i-1)}{N} \right] + \sin \frac{s\pi i}{N} \end{aligned}$$

i stąd wyznaczamy

$$\lambda_s = \frac{1 - \alpha(1-\sigma) 4 \sin^2 \frac{s\pi}{2N}}{1 + 4\alpha\sigma \sin^2 \frac{s\pi}{2N}} \quad (8.34)$$

A zatem funkcje $V_{ij}^{(s)}$ z λ_s określonym wzorem (8.34) spełniają równania (8.30), (8.31).

Rozwiązania V całego zagadnienia będziemy poszukiwać w postaci kombinacji liniowych funkcji $V^{(s)}$. Niech

$$V_{ij} = \sum_{s=1}^{N-1} a_s \lambda_s^j \sin \frac{s\pi i}{N} \quad (8.35)$$

Z liniowości równań (8.30) i (8.32) wynika, że V_{ij} spełniają te równania dla

dowolnych stałych a_s . Stałe te wybieramy tak, aby spełnione były równania (8.31). Mamy zatem

$$V_{i,0} = \sum_{s=1}^{N-1} a_s \sin \frac{s\pi i}{N} = G_i, \quad i = 1, \dots, N-1 \quad (8.36)$$

Mnożąc równanie $V_{i,0} = G_i$ stronami przez $\sin \frac{r\pi i}{N}$ oraz sumując wszystkie te równania dla $i = 1, \dots, N-1$, otrzymamy

$$a_r = \frac{2}{N} \sum_{i=1}^{N-1} G_i \sin \frac{r\pi i}{N}, \quad r = 1, \dots, N-1 \quad (8.37)$$

Zatem funkcja V określona wzorami (8.35), (8.34), (8.37) jest rozwiązaniem rozpatrywanego zagadnienia (8.30)–(8.32).

Ostatni wzór wynika natychmiast z następującej własności funkcji

$$X^s(x) = \sqrt{\frac{2}{l}} \sin s\pi x, \quad s = 1, \dots, N-1. \text{ Ciąg funkcji } X^1(x), X^2(x), \dots, X^{N-1}(x)$$

jest bowiem ortonormalny na zbiorze punktów $x_i = \frac{i}{N}$, $i = 1, \dots, N-1$, tzn.

$$\langle X^n, X^m \rangle = \sum_{i=1}^{N-1} h X^n(x_i) X^m(x_i) = \delta_{n,m}$$

Pozostawiamy Czytelnikowi porównanie otrzymanych wzorów (8.35), (8.37) dla funkcji V ze wzorem, jaki otrzymujemy dla rozwiązania zagadnienia brzegowego pierwszego rodzaju stosując metodę rozdzielania zmiennych Fouriera (patrz np. [74]).

Udowodnimy, że dla tych σ, h, k , dla których zachodzi $|\lambda_s| \leq 1$, $s = 1, \dots, N-1$, badana metoda różnicowa jest stabilna względem warunku początkowego w normie $\|\cdot\|_{L_2}$. Mamy

$$\|V^j\|_{L_2}^2 = \langle V^j, V^j \rangle = \frac{l}{2} \sum_{s,k=1}^{N-1} a_s a_k \lambda_s^j \lambda_k^j \langle X^s, X^k \rangle = \frac{l}{2} \sum_{s=1}^{N-1} a_s^2 \lambda_s^{2j} \quad (8.38)$$

Dla $j = 0$ otrzymujemy więc

$$\|G\|_{L_2}^2 = \frac{l}{2} \sum_{s=1}^{N-1} a_s^2$$

Jeżeli $|\lambda_s| \leq 1$, to

$$\|V^j\|_{L_2}^2 \leq \frac{l}{2} \sum_{s=1}^{N-1} a_s^2 = \|G\|_{L_2}^2 \quad (8.39)$$

Wyjaśnimy teraz, dla jakich σ, h, k zachodzi $|\lambda_s| \leq 1$, $s = 1, \dots, N-1$. Warunek $|\lambda_s| \leq 1$ zapiszemy w postaci

$$-1 \leq 1 - \frac{4\alpha \sin^2 \frac{2\pi}{2N}}{1 + 4\alpha\sigma \sin^2 \frac{s\pi}{2N}} \leq 1 \quad (8.40)$$

Po elementarnych przekształceniach stwierdzamy, że nierówności (8.40) zachodzą wtedy i tylko wtedy, gdy

$$\sigma \geq \frac{1}{2} - \frac{1}{4\alpha \sin^2 \frac{s\pi}{2N}} \quad (8.41)$$

Z kolei nierówność (8.41) zachodzi dla $s = 1, \dots, N-1$, gdy

$$\sigma \geq \frac{1}{2} - \frac{1}{4\alpha} = \frac{1}{2} - \frac{h^2}{4k} \quad (8.42)$$

Rozpatrzmy teraz przypadki:

a) $\sigma < 1/2$. Z (8.42) otrzymujemy następujące ograniczenie

$$\frac{k}{h^2} \leq \frac{1}{2(1-2\sigma)} \quad (8.43)$$

na odstępach h, k siatki. Zatem dla $\sigma < 1/2$ rozpatrywana metoda różnicowa jest stabilna względem warunku początkowego (w normie $\|\cdot\|_{L_2}$), jeśli h, k spełniają warunek (8.43). Dla $\sigma = 0$, dla którego wyznaczenie rozwiązania różnicowego jest najprostsze, otrzymujemy ograniczenie $\frac{k}{h^2} \leq 1/2$. Zauważ-

my, że dla $\sigma < 0$ ograniczenie (8.43) jest bardziej krępujące niż dla $\sigma = 0$.

b) $\sigma \geq 1/2$. Warunek (8.42) jest spełniony dla dowolnych h, k . Metoda różnicowa jest zatem stabilna względem warunku początkowego (w normie $\|\cdot\|_{L_2}$) dla dowolnych h, k .

Udowodnimy teraz, że dla $\sigma \leq \frac{1}{2} - \frac{1}{4\alpha} - \varepsilon$ (ε — dowolna stała dodatnia) rozpatrywana metoda różnicowa nie jest stabilna względem warunku początkowego w normie $\|\cdot\|_{L_2}$. Zauważmy, że dla $\sigma \leq \frac{1}{2} - \frac{1}{4\alpha} - \varepsilon$ i dostatecznie dużych N istnieje takie $s_0 < N$ (np. $s_0 = N-1$), że $|\lambda_{s_0}| > 1 + \varepsilon_1$, gdzie ε_1 nie zależy od h, k . Na podstawie wzoru (8.38) mamy

$$\|(V^{s_0})^j\|_{L_2}^2 = \lambda_{s_0}^{2j} \frac{l}{2} \geq \frac{l}{2} (1 + \varepsilon_1)^{2j} \rightarrow \infty \quad \text{dla} \quad k \rightarrow 0 \quad (8.44)$$

czyli badana metoda różnicowa nie jest stabilna względem warunku począt-

kowego (w normie $\|\cdot\|_{L_2}$). Zauważmy, że warunek (8.44) będzie także spełniony, jeżeli zamiast V^{s_0} weźmiemy $\frac{1}{R} V^{s_0}$, gdzie R jest dowolną liczbą rzeczywistą. Dobierając odpowiednio R można zaburzenie warunku początkowego $\frac{1}{R} \sin \frac{s_0 \pi i}{N}$ uczynić dowolnie małym. Nie należy więc przeprowadzać obliczeń, gdy $\sigma \leq \frac{1}{2} - \frac{1}{4\alpha} - \varepsilon$.

Zarówno dla normy $\|\cdot\|_c$ (patrz nierówność (8.22), gdy dodatkowo $F \equiv 0$), jak i dla normy $\|\cdot\|_{L_2}$ (patrz nierówność (8.39)), stabilność względem warunku początkowego rozpatrywanej metody różnicowej wynika z nierówności

$$\|V^j\| \leq \|G\| \quad (8.45)$$

Oszacowanie $\|V^j\|$ w zbiorze $\bar{\omega}$ nie zależy od T . Niech k będzie ustalone. Z (8.45) wynika, że zmiany rozwiązania różnicowego na wszystkich warstwach z $G = \{(x, t): 0 \leq x \leq l, t > 0\}$ mierzone normą na wszystkich warstwach można oszacować przez zaburzenia warunku początkowego na warstwie zerowej (mierzone normą na tej warstwie).

Rozpatrzmy przypadek $\sigma = 0$. Układ równań różnicowych można otrzymać rozbijając dyskretyzację na dwa etapy — najpierw względem zmiennej x w celu sprowadzenia równania przewodnictwa cieplnego do układu równań różniczkowych zwyczajnych, a następnie na zastosowaniu metody Eulera.

E t a p I. Oznaczmy $U_i(t) = u(x_i, t)$. Pochodną $\frac{\partial^2 u}{\partial x^2}$ na prostej $x = x_i$, $i = 1, \dots, N-1$, przybliżamy wyrażeniem

$$\frac{1}{h^2} [U_{i+1}(t) - 2U_i(t) + U_{i-1}(t)]$$

Dla zagadnienia brzegowego pierwszego rodzaju otrzymamy w ten sposób układ równań różniczkowych zwyczajnych

$$U_i'(t) = \frac{1}{h^2} [U_{i+1}(t) - 2U_i(t) + U_{i-1}(t)] + f(x_i, t) \quad i = 1, \dots, N-1$$

$$U_0(t) = \varphi(t), \quad U_N(t) = \psi(t), \quad 0 \leq t \leq T$$

z warunkami początkowymi

$$U_i(0) = g(x_i), \quad i = 1, \dots, N-1$$

Otrzymane zagadnienie początkowe zapiszemy w postaci macierzowej

$$U'(t) = \frac{1}{h^2} M U(t) + F(t) \quad (8.46)$$

$$U(0) = G$$

gdzie

$$M = \begin{bmatrix} -2 & 1 & & & 0 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ 0 & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}, \quad U(t) = \begin{bmatrix} U_1(t) \\ U_2(t) \\ \vdots \\ U_{N-1}(t) \end{bmatrix}$$

$$F(t) = \begin{bmatrix} f(x_1, t) - \frac{1}{h^2} \varphi(t) \\ f(x_2, t) \\ \vdots \\ f(x_{N-2}, t) \\ f(x_{N-1}, t) + \frac{1}{h^2} \psi(t) \end{bmatrix}, \quad G = \begin{bmatrix} g(x_1) \\ g(x_2) \\ \vdots \\ g(x_{N-1}) \end{bmatrix}$$

Etap II. Stosując opisaną w rozdziale 7 metodę Eulera do zagadnienia (8.46), otrzymujemy

$$U((j+1)k) = U(jk) + k \frac{1}{h^2} MU(jk) + kF(jk) \quad (8.47)$$

$$U(0) = G$$

czyli badaną metodę różnicową dla $\sigma = 0$. Składowe wektora $U(jk)$ są wartościami rozwiązania różnicowego na warstwie $t = jk$.

Na podstawie twierdzeń 8 i 14, podanych w p. 6.2, wartości własne λ_j macierzy M są rzeczywiste oraz $-4 \leq \lambda_j \leq 0$. Warunek absolutnej stabilności (patrz p. 7.3.3) dla metody Eulera (8.47) jest następujący

$$\left| 1 + k \frac{1}{h^2} \lambda_j \right| \leq 1$$

Stąd otrzymujemy ograniczenie $\frac{k}{h^2} \leq \frac{1}{2}$. Ograniczenie to przy ustalonym k jest zarówno warunkiem stabilności względem warunku początkowego rozpatrywanej metody różnicowej dla $\sigma = 0$, jak i absolutnej stabilności metody Eulera (8.47).

Przeprowadzenie analogicznej dyskusji dla $\sigma \neq 0$ pozostawimy Czytelnikowi (patrz ćwic. 2).

Zbieżność rozwiązania różnicowego. Z liniowości układu różnicowego (8.10)–(8.12) wynika, że błąd $V_{ij} = v_{ij} - u_{ij}$ jest rozwiązaniem układu (8.17), gdy $G_i = 0$, $F_{i,j+1/2} = L_{h,k}^\sigma(v_{ij} - u_{ij})$. Ze wzorów (8.6)–(8.9) oraz (8.15) i (8.16) wynika, że istnieją stałe c_i niezależne od h, k (zależą one od obszaru $\bar{D}$, rozwiązania u i jego pochodnych) takie, że

$$|F_{i,j+1/2}| \leq \begin{cases} c_1(h^2 + k^2), & \sigma = 1/2, u \in C_3^4(\bar{D}) \\ c_2(h^2 + k), & \sigma \neq 1/2, u \in C_2^4(\bar{D}) \end{cases}$$

Dla $\sigma \geq 1 - \frac{h^2}{2k}$, $0 \leq \sigma \leq 1$, na podstawie (8.26) otrzymujemy następujące oszacowanie błędu

$$\begin{aligned}
& \max_{(i,j) \in \bar{\omega}} |v_{ij} - u_{ij}| = \\
& = \max_j \|v^j - u^j\|_c \leq \begin{cases} Tc_1(h^2 + k^2), & \sigma = 1/2, u \in C_3^4(D) \\ Tc_2(h^2 + k), & \sigma \neq 1/2, u \in C_2^4(D) \end{cases} \quad (8.48)
\end{aligned}$$

stąd zaś wynika jednostajna zbieżność rozwiązania różnicowego v do rozwiązania u , gdy $h, k \rightarrow 0$.

Zauważmy, że dowodząc zbieżności korzystaliśmy tylko ze stabilności metody różnicowej względem prawej strony. Możemy zatem powiedzieć, że jeżeli metoda różnicowa jest stabilna względem prawej strony i norma (na warstwie) błędu $L_{h,k}(v_{ij} - u_{ij})$ (błędu zastąpienia równania różniczkowego równaniem różnicowym) dąży do zera, gdy $h, k \rightarrow 0$, to

$$\max_j \|v^j - u^j\| \rightarrow 0, \quad \text{gdy} \quad h, k \rightarrow 0 \quad (8.49)$$

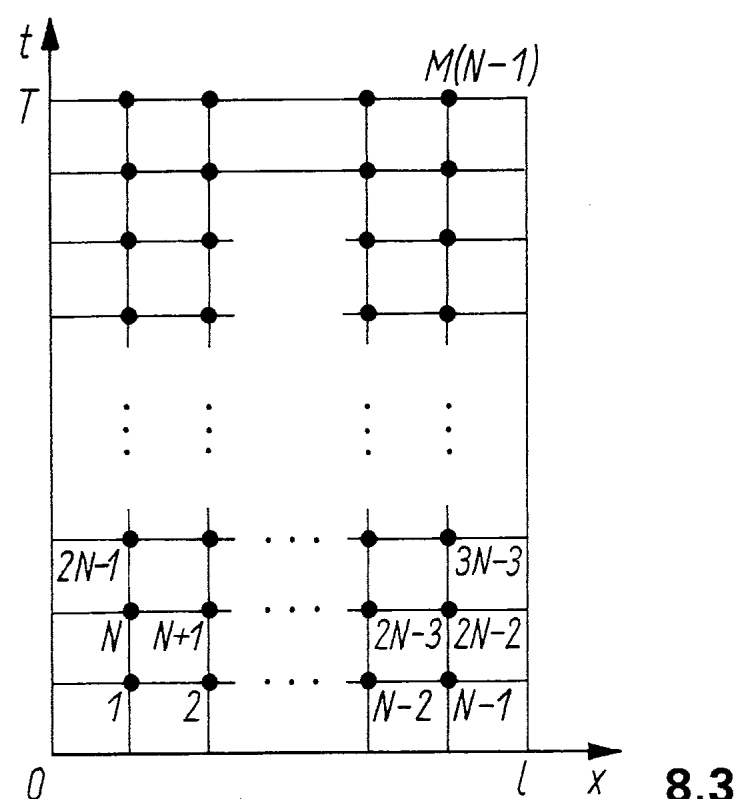
Jeżeli $\|\cdot\| = \|\cdot\|_c$, to warunek (8.49) zapewnia jednostajną zbieżność v do u .

W punkcie tym nie omawialiśmy problemu stabilności metody różnicowej względem prawej strony w normie $\|\cdot\|_{L_2}$, ale zauważmy, że wtedy warunek (8.49) nie zapewnia w tym przypadku jednostajnej zbieżności.

Kilka uwag końcowych. Rozwiązując zagadnienie brzegowe (8.1)–(8.3) za pomocą metod różnicowych, powinniśmy stosować tylko takie metody, które zapewniają zbieżność rozwiązania różnicowego v do rozwiązania dokładnego u zagadnienia brzegowego, gdy $h, k \rightarrow 0$, bowiem tylko w takim przypadku możemy otrzymać rozwiązanie różnicowe dostatecznie bliskie u . Ale powyższy warunek nie zapewnia tego, że przy praktycznym rozwiązywaniu zagadnień dla dostatecznie małych h i k zawsze otrzymamy dobre przybliżenie.

Dane wejściowe, czyli prawa strona równania różniczkowego oraz dane początkowe i brzegowe są na ogół obarczone błędami. Powinniśmy zatem stosować tylko takie metody różnicowe, które zapewniają, że zmiany rozwiązania różnicowego spowodowane zaburzeniami danych początkowych nie będą wzrastać nieograniczenie przy zagęszczaniu siatki. W punkcie 8.3.1 omówiliśmy problem stabilności względem prawej strony i warunku początkowego przedstawionej w p. 8.1.2 metody różnicowej. Rozpatrzyliśmy normy $\|\cdot\|_c$ i $\|\cdot\|_{L_2}$ (na warstwie). Dla $\sigma \geq 1 - \frac{h^2}{2k}$, $0 \leq \sigma \leq 1$, mamy zapewnioną stabilność w normie $\|\cdot\|_c$ i jednostajną zbieżność rozwiązania różnicowego v do rozwiązania dokładnego u .

Z powodu błędów zaokrągleń nie otrzymujemy dokładnego rozwiązania układu równań różnicowych. Błędy te mogą spowodować, że otrzymane rozwiązanie nie będzie praktycznie przydatne. Jeśli w zbiorze węzłów wewnętrznych ω węzły ponumerujemy tak jak na rys. 8.3, to układ równań różnicowych możemy zapisać w postaci macierzowej $AV = F$, gdzie i -tą



składową wektora niewiadomych V jest wartość szukanej funkcji w węźle o numerze i . Macierz A możemy wówczas przedstawić w postaci

$$A = \begin{bmatrix} A_1 & & & & & 0 \\ B_1 & A_2 & & & & \\ & B_2 & A_3 & & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \\ & 0 & & & A_{M-1} & \\ & & & & B_{M-1} & A_M \end{bmatrix} \quad (8.50)$$

w której macierze A_i oraz B_i są macierzami kwadratowymi stopnia $N-1$, a elementy na głównej przekątnej są dodatnie. Ze wzoru różnicowego (8.13) wynika wtedy bezpośrednio, że dla $\sigma \geq 1 - \frac{h^2}{2k}$, $0 \leq \sigma \leq 1$, elementy stojące na głównej przekątnej macierzy A są dominujące (tzn. suma modułów pozostałych elementów w wierszu nie przekracza modułu tego elementu), a ta własność macierzy A zapewnia, że dla omówionej w p. 8.1.2 metody rozwiązywania układu $AV = F$ błędy zaokrągleń mają niewielki wpływ na rozwiązanie różnicowe V .

8.1.4. Przykład obliczeniowy

Za pomocą przedstawionej w p. 8.1.2 metody różnicowej, gdy $\sigma = 0$, wyznaczmy rozwiązanie następującego zagadnienia

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 \leq x \leq 1, \quad t \geq 0$$

$$u(x, 0) = \sin \pi x, \quad u(0, t) = u(1, t) = 0$$

Na podstawie wzoru (8.35) rozwiązanie (dokładne) układu różnicowego jest następujące

$$v_{ij} = \left(1 - 4\alpha \sin^2 \frac{\pi}{2N}\right)^j \sin \frac{\pi i}{N} \quad (8.51)$$

Niech $\alpha = \frac{k}{h^2} = \text{const.}$ Ustalamy punkt (x, t) , gdzie $x = ih$, $t = jk$. Mamy

$$\lim_{\substack{h \rightarrow 0 \\ k \rightarrow 0}} v_{ij} = \lim_{\substack{h \rightarrow 0 \\ k \rightarrow 0}} \left(1 - 4\alpha \sin^2 \frac{\pi \sqrt{k}}{2\sqrt{\alpha}}\right)^{t/k} \sin \pi x = e^{-\pi^2 t} \sin \pi x$$

Widzimy więc, że przy zagęszczeniu siatki ($\alpha = \text{const}$) rozwiązanie (8.51) jest zbieżne do rozwiązania dokładnego $u(x, t) = e^{-\pi^2 t} \sin \pi x$.

W tablicy 8/1 zestawiono wyniki obliczeń wykonanych na maszynie Odra 1305 dla tego przykładu, dla wartości $\alpha = \frac{1}{2}$ i 2. Dla $\alpha = 2$ otrzymano rozwiązanie rozbieżne. Współczynniki a_k ($k \neq 1$) w sumie (8.35), które powinny być zerami wskutek błędów zaokrągleń mogą być różne od zera (choć bardzo małe). Dla $\alpha \geq \frac{1}{2} + \varepsilon$ ($\varepsilon > 0$) w sumie tej występują składniki

8/1

t	$(u-v)$ na prostej $x = \frac{1}{2}$	
	$h = 0.1, k = 0.005$ $\alpha = 0.5$	$h = 0.05, k = 0.005$ $\alpha = 2$
0.01	$0.15 \cdot 10^{-3}$	$0.21 \cdot 10^{-2}$
0.02	$0.27 \cdot 10^{-3}$	$0.38 \cdot 10^{-2}$
0.03	$0.37 \cdot 10^{-3}$	$0.51 \cdot 10^{-2}$
0.04	$0.45 \cdot 10^{-3}$	$0.62 \cdot 10^{-2}$
0.05	$0.51 \cdot 10^{-3}$	$0.88 \cdot 10^{-2}$
0.06	$0.55 \cdot 10^{-3}$	$0.90 \cdot 10^{-1}$
0.07	$0.58 \cdot 10^{-3}$	3.87
0.08	$0.60 \cdot 10^{-3}$	182.61
0.09	$0.61 \cdot 10^{-3}$	8654.44

rozbieżne (metoda nie jest stabilna), zatem otrzymano z maszyny rozwiązanie rozbieżne, natomiast dla $\alpha = 1/2$ takich składników nie ma, więc otrzymano rozwiązanie z oczekiwaną dokładnością.

8.1.5. Schemat blokowy

Podamy teraz schemat blokowy programu (rys. 8.4) wyznaczania rozwiązań v_{ij} , gdy $\sigma \neq 0$. Rozwiązania v_{ij} wyznaczamy kolejno warstwami (patrz p. 8.1.2) dla $t = jk$ ($j \geq 1$). Na każdej warstwie mamy do rozwiązania układu $(N-1)$ równań postaci (8.14). Niech j będzie ustaloną liczbą. Rozważany układ zapiszemy w następującej postaci

$$v_{i+1,j} - (2+s)v_{ij} + v_{i-1,j} = -r_{i,j} \quad s = \frac{h^2}{\sigma k}, \quad i = 1, \dots, N-1$$

$$v_{0,j} = \varphi_j, \quad v_{N,j} = \psi_j$$

gdzie

$$r_{i,j} = \frac{1-\sigma}{\sigma} v_{i+1,j-1} + \left[s - 2 \frac{1-\sigma}{\sigma} \right] v_{i,j-1} + \frac{1-\sigma}{\sigma} v_{i-1,j-1} - h^2 \sigma f_{i,j-1/2}$$

W rozdziale 5 omówiono metody rozwiązywania układów równań z macierzą trójdziagonalną. Zastosujemy teraz metodę określoną wzorami (5.40b) do naszego układu, tzn. rozwiązanie v_{ij} ($i = 1, \dots, N-1$) wyznaczymy ze wzoru rekurencyjnego

$$v_{N,j} = \psi_j$$

$$v_{i,j} = a_i v_{i+1,j} + b_{ij}, \quad i = N-1, \dots, 1$$

gdzie a_i , b_{ij} są określone w następujący rekurencyjny sposób

$$a_0 = 0, \quad a_i = \frac{1}{2+s-a_{i-1}}, \quad i = 1, \dots, N-1$$

$$b_{0,j} = \varphi_j, \quad b_{ij} = a_i(b_{i-1,j} + r_{i,j}), \quad i = 1, \dots, N-1$$

Ćwiczenia

1. Korzystając z zasady maksimum udowodnić, że układ (8.14) ma jednoznaczne rozwiązanie. Wskazać. Udowodnić, że rozwiązanie układu jednorodnego $AV_j = 0$ (patrz wzór (8.14)) nie może zawierać ani liczb dodatnich, ani liczb ujemnych.

2. Ustalić jakim metodom rozwiązywania zagadnienia (8.46) odpowiada określona w p. 8.1.2 metoda różnicowa, gdy $\sigma \neq 0$. Zbadać absolutną stabilność tych metod.

3. Rozpatrzyć zagadnienie brzegowe

$$y''(x) + r(x) y'(x) - p(x) y(x) = q(x), \quad a \leq x \leq b, \quad p(x) \geq 0$$

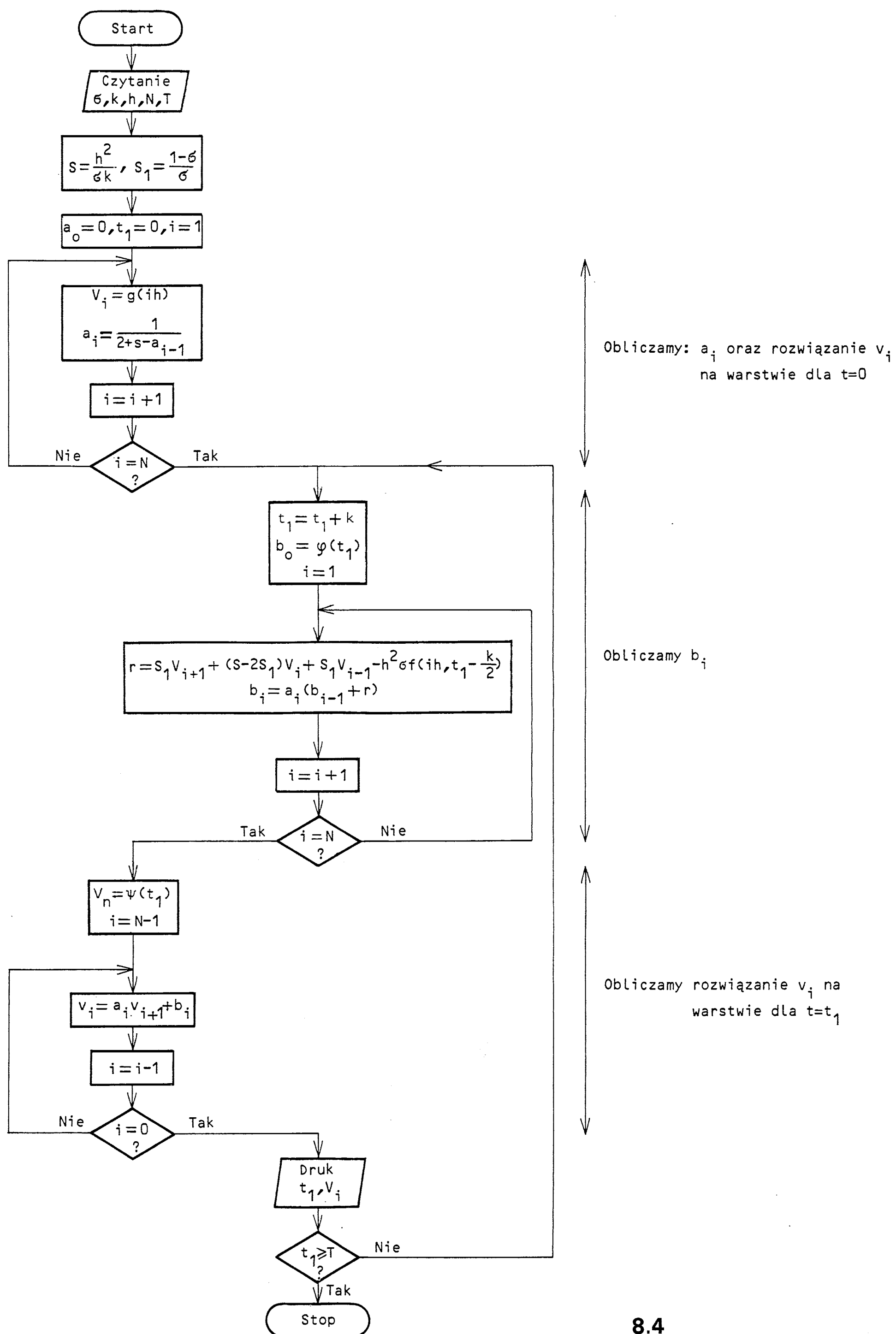
$$y(a) = \alpha, \quad y(b) = \beta$$

i metodę różnicową zastosowaną do rozwiązania tego zagadnienia

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + r_i \frac{y_{i+1} - y_{i-1}}{2h} - p_i y_i = q_i, \quad i = 1, \dots, N-1$$

$$y_0 = \alpha, \quad y_N = \beta$$

gdzie $x_i = ih$, $h = \frac{b-a}{N}$, N — całkowite. Wielkości y_i , r_i , p_i , q_i oznaczają odpowiednio $y(x_i)$, $r(x_i)$, $p(x_i)$, $q(x_i)$. Zbadać zbieżność tej metody. Wskazać. Zastosować zasadę maksimum.



8.2. METODA RÓŻNICOWA DLA RÓWNANIA DRGAŃ STRUNY

8.2.1. Postawienie zagadnienia

Jako najprostszy przykład równania hiperbolicznego rozpatrzmy *równanie drgań struny jednorodnej*

$$Lu = \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = f(x, t) \quad (8.52)$$

i następujące zagadnienie dla tego równania. Znaleźć funkcję $u = u(x, t)$ spełniającą wewnątrz prostokąta $\bar{D} = \{(x, t): 0 \leq x \leq l, 0 \leq t \leq T\}$ równanie (8.52) oraz warunki początkowe

$$u(x, 0) = g(x) \quad (8.53)$$

$$\frac{\partial u(x, 0)}{\partial t} = F(x), \quad 0 \leq x \leq l \quad (8.54)$$

i warunki brzegowe

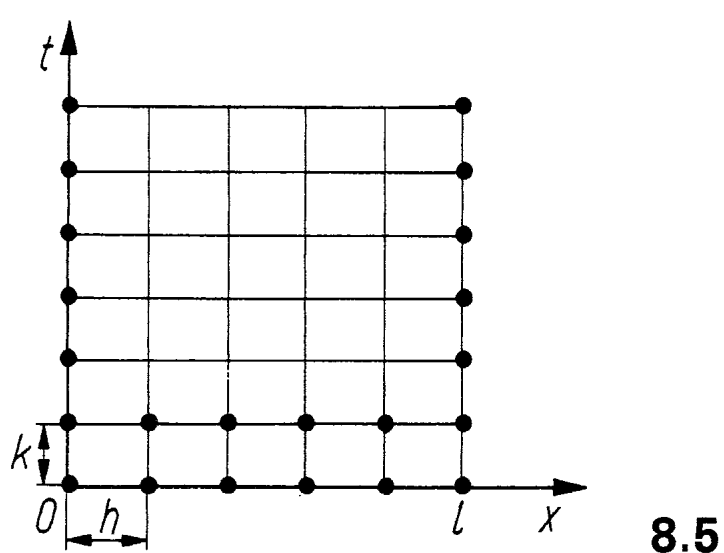
$$u(0, t) = \varphi(t), \quad u(l, t) = \psi(t), \quad 0 \leq t \leq T \quad (8.55)$$

Zagadnienie to nazywamy *zagadnieniem brzegowym pierwszego rodzaju* dla równania (8.52).

Zakładać będziemy, że funkcja $f(x, t)$ jest ciągła na $\bar{D}$, funkcje $g(x)$ i $F(x)$ są ciągłe na przedziale $\langle 0; l \rangle$, a funkcje $\varphi(t)$, $\psi(t)$ są ciągłe na $\langle 0; T \rangle$.

8.2.2. Aproksymacja różnicowa

W celu rozwiązania sformułowanego powyżej zagadnienia prostokąt $\bar{D}$ pokrywamy taką samą siatką, jak w przypadku opisanym w p. 8.1.2 i przyjmujemy taki sam zbiór węzłów $\bar{\omega} = \{(i, j): 0 \leq i \leq N, 0 \leq j \leq M\}$, natomiast zbiór ten podzielimy teraz inaczej. Węzły leżące na prostych $x = 0$, $x = l$,

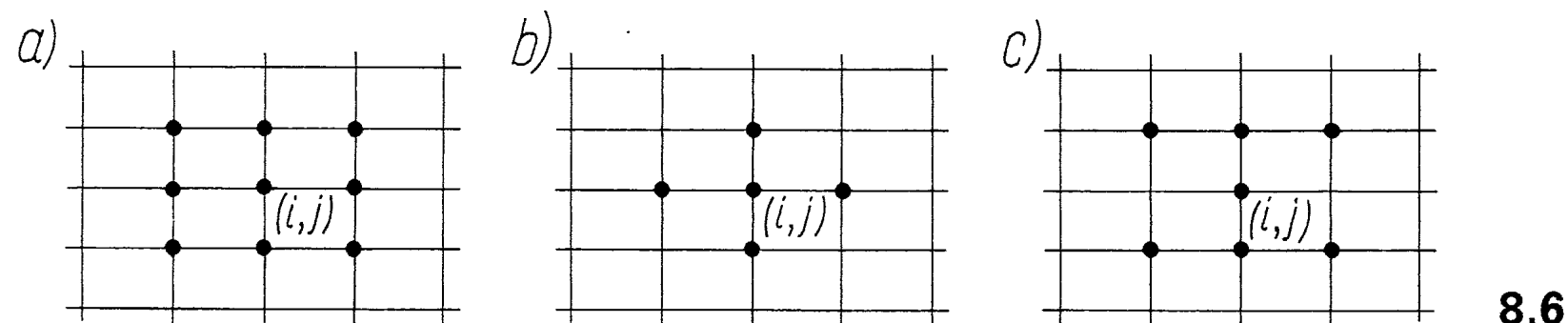


$t = 0$, $t = k$ zaliczymy do węzłów brzegowych, pozostałe zaś do wewnętrznych (rys. 8.5). Dla każdego węzła wewnętrznego (i, j) , $i = 1, \dots, N-1$; $j \geq 2$, rozpatrzmy następujące równanie różnicowe

$$L_{h,k}^{\sigma} v_{ij} = \frac{v_{i,j+1} - 2v_{i,j} + v_{i,j-1}}{k^2} - \Lambda(\sigma v_{i,j+1} + (1-2\sigma)v_{i,j} + \sigma v_{i,j-1}) = f_{ij} \quad (8.56)$$

gdzie Λ jest operatorem różnicowym określonym wzorem (8.5), a σ — parametrem.

Równanie powyższe wiąże ze sobą wartości v_{ij} w węzłach z trzech sąsiednich warstw. Wzajemne położenie tych węzłów dla $\sigma \neq 0$ i $\sigma \neq 1/2$ jest pokazane na rys. 8.6a, dla $\sigma = 0$ na rys. 8.6b, a dla $\sigma = 1/2$ na rys. 8.6c.



8.6

Gdy $u \in C_4^4(\bar{D})$, dla dowolnej stałej σ (σ nie zależy od h, k) mamy

$$L_{h,k}^{\sigma} u_{ij} = f_{ij} + O(h^2 + k^2) \quad (8.57)$$

Zatem dla $u \in C_4^4(\bar{D})$ błąd zastąpienia równania różniczkowego (8.52) równaniem różnicowym (8.57) wynosi $O(h^2 + k^2)$.

W węzłach brzegowych leżących na prostych $x = 0$, $x = l$, $t = 0$ rozwiązanie wyznaczamy z warunków brzegowych (8.55) i warunku początkowego (8.53). Mamy

$$v_{i,0} = g_i, \quad i = 1, \dots, N-1 \quad (8.58)$$

$$v_{0,j} = \varphi_j, \quad v_{N,j} = \psi_j, \quad j = 0, 1, \dots, M \quad (8.59)$$

gdzie $g_i = g(x_i)$, $\varphi_j = \varphi(t_j)$, $\psi_j = \psi(t_j)$.

Rozwiązanie w węzłach na prostej $t = k$ wyznaczamy wykorzystując drugi warunek początkowy (8.54). Mamy

$$\frac{u_{i,1} - u_{i,0}}{k} = \frac{\partial u_{i,0}}{\partial t} + \frac{k}{2} \frac{\partial^2 u_{i,0}}{\partial t^2} + O(k^2)$$

Zakładając, że równanie różniczkowe (8.52) jest spełnione na prostej $x = 0$ powyższy wzór zapiszemy w następującej postaci

$$\frac{\partial u_{i,0}}{\partial t} = \frac{u_{i,1} - u_{i,0}}{k} - \frac{k}{2} (g_i'' + f_{i,0}) + O(k^2) \quad (8.60)$$

Rozpatrzmy obecnie aproksymację warunku początkowego (8.54) równaniem

$$\frac{v_{i,1} - v_{i,0}}{k} = F_i + \frac{k}{2} (g_i'' + f_{i,0}), \quad i = 1, \dots, N-1 \quad (8.61)$$

Błąd zastąpienia równania (8.54) równaniem różnicowym (8.61) wynosi $O(k^2)$.

W węzłach (i, j) , $i = 1, \dots, N-1$; $j \geq 1$, rozwiązanie różnicowe wyznaczamy kolejno warstwami. Na pierwszej warstwie wartości $v_{i,1}$ obliczamy ze wzoru (8.61). Na następnych warstwach dla $\sigma = 0$ rozwiązanie $v_{i,j+1}$, $j \geq 1$, wyznaczamy bezpośrednio ze wzoru (8.56), a dla $\sigma \neq 0$ na każdej warstwie musimy rozwiązać układ równań

$$\left(\frac{1}{k^2} - \sigma\Lambda\right) v_{i,j+1} = r_{ij}$$

$$\text{gdzie } r_{ij} = \left(\frac{2}{k^2} + (1-2\sigma)\Lambda\right) v_{ij} + \left(\frac{1}{k^2} + \sigma\Lambda\right) v_{i,j-1} + f_{ij}.$$

Zagadnienia stabilności i zbieżności przedstawionej metody nie będziemy omawiać; Czytelnika interesującego się tymi problemami odsyłamy do książek [10], [55].

Ćwiczenia

1. Udowodnić, że dla $\sigma > 0$ układ (8.62) ma jednoznaczne rozwiązanie.
2. Zapisać rozpatrywane w tym punkcie równania różnicowe w postaci macierzowej w podobny sposób, jak równania różnicowe z p. 8.1.2 zostały zapisane w p. 8.1.3. Wyznaczyć σ , dla których macierz układu będzie macierzą diagonalnie dominującą.

8.3. METODA RÓŻNICOWA DLA RÓWNANIA POISSONA

8.3.1. Postawienie zagadnienia

Jako najprostszy przykład równania typu eliptycznego rozpatrzmy *równanie Poissona*

$$Lu = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y) \quad (8.63)$$

Zagadnieniem brzegowym pierwszego rodzaju (zagadnieniem *Dirichleta*) dla tego równania jest zagadnienie: znaleźć funkcję $u = u(x, y)$ spełniającą wewnątrz ograniczonego obszaru $\bar{D}$ o brzegu Γ równanie (8.63) oraz warunek brzegowy

$$u|_{\Gamma} = \varphi \quad (8.64)$$

gdzie φ — dana funkcja określona na Γ . Zakładamy, że funkcja $f(x, y)$ jest ciągła na $\bar{D}$, a funkcja φ jest ciągła na Γ .

8.3.2. Aproksymacje różnicowe równania Poissona

Rozpatrzmy różne aproksymacje różnicowe równania Poissona. Oznaczmy przez $0, 1, \dots$ węzły siatki, a przez u_j, v_j, f_j — wartości funkcji u, v, f w j -tym węźle.

Rozpatrzmy najpierw aproksymację dla siatki prostokątnej. W tym przypadku pochodne $\partial^2 v / \partial x^2$, $\partial^2 v / \partial y^2$ w węźle 0 (rys. 8.7) przybliżamy ilorazami

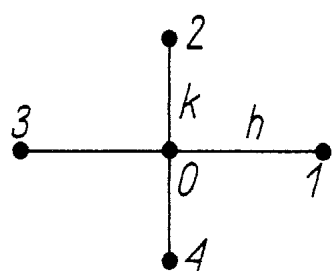
$$\frac{\partial^2 v_0}{\partial x^2} \sim \frac{v_1 - 2v_0 + v_3}{h^2}$$

$$\frac{\partial^2 v_0}{\partial y^2} \sim \frac{v_2 - 2v_0 + v_4}{k^2}$$

a całe równanie Poissona w węźle 0 przybliżamy równaniem różnicowym

$$L_{h,k} v_0 = \frac{v_1 - 2v_0 + v_3}{h^2} + \frac{v_2 - 2v_0 + v_4}{k^2} = f_0 \quad (8.65)$$

Jest to najczęściej stosowane równanie różnicowe aproksymujące równanie Poissona.



8.7

Gdy $u \in C^4(\bar{D})$ ($C^m(\bar{D})$ jest klasą funkcji mających na $\bar{D}$ ciągłe pochodne cząstkowe do rzędu m włącznie), wówczas

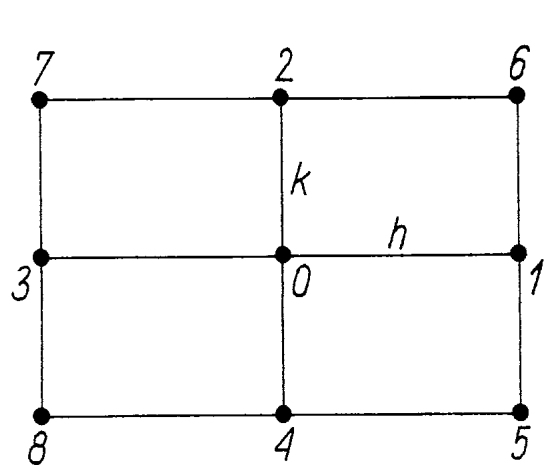
$$L_{h,k} u_0 = f_0 + O(h^2 + k^2) \quad (8.66)$$

Zatem błąd zastąpienia równania Poissona równaniem (8.65) wynosi $O(h^2 + k^2)$.

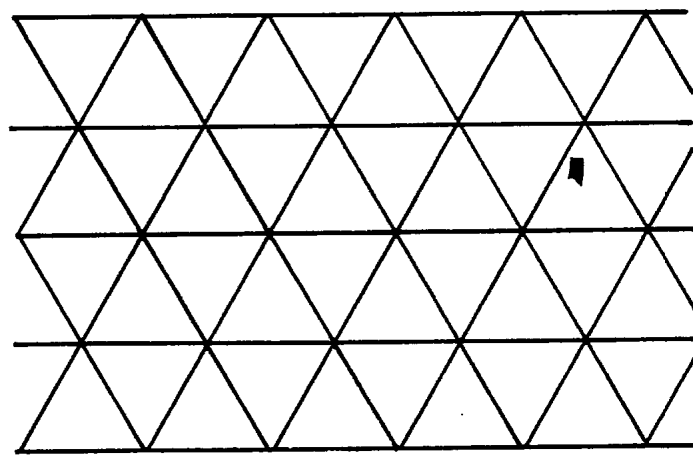
Stosowane są także równania aproksymujące dokładniej równanie Poissona. Takim równaniem jest np. równanie uwzględniające wartości funkcji w 9 punktach (rys. 8.8), które po odpowiednim pogrupowaniu wyrazów ma postać

$$\begin{aligned} \frac{5}{3} \left(\frac{1}{h^2} + \frac{1}{k^2} \right) v_0 &= \frac{1}{6} \left(\frac{5}{h^2} - \frac{1}{k^2} \right) (v_1 + v_3) + \frac{1}{6} \left(\frac{5}{k^2} + \frac{1}{h^2} \right) (v_2 + v_4) + \\ &+ \frac{1}{12} \left(\frac{1}{h^2} + \frac{1}{k^2} \right) (v_5 + v_6 + v_7 + v_8) + f_0 + \frac{h^2}{12} (f_1 - 2f_0 + f_3) + \\ &+ \frac{k^2}{12} (f_2 - 2f_0 + f_4) \end{aligned} \quad (8.67)$$

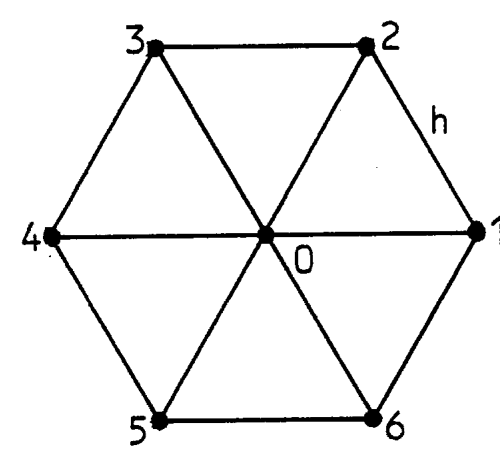
Po podstawieniu rozwiązania dokładnego do tego wzoru i rozwinięciu u_j, f_j według wzoru Taylora w otoczeniu węzła zerowego otrzymamy błąd zastąpienia równania Poissona równaniem (8.67), który wynosi $O(h^4 + k^4)$, gdy $u \in C^6(\bar{D})$.



8.8



8.9



8.10

Rozpatrzmy jeszcze regularną siatkę trójkątną (rys. 8.9), i równanie różnicowe

$$\frac{2}{3h^2}(v_1 + v_2 + v_3 + v_4 + v_5 + v_6 - 6v_0) = f_0 \quad (8.68)$$

przy czym węzły 0, 1, ..., 6 są rozłożone tak, jak na rys. 8.10.

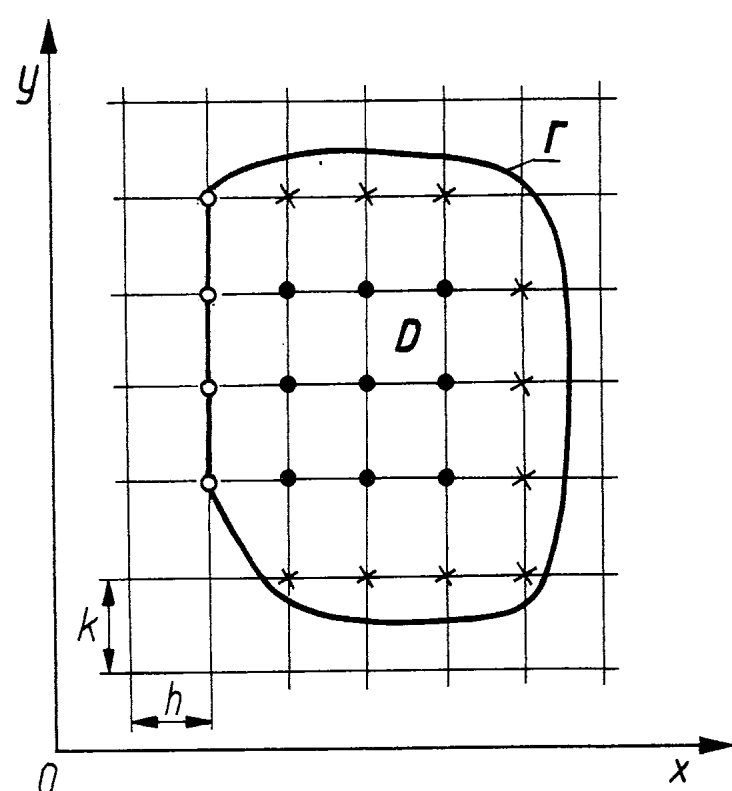
Podobnie jak poprzednio, stosując wzór Taylora znajdujemy błąd zastąpienia równania Poissona równaniem (8.68) — wynosi on $O(h^2)$. Dokładniejsze przybliżenie, równe $O(h^4)$ można uzyskać, jeżeli po prawej stronie wzoru (8.68) występuje nie tylko f_0 , ale i f_j w innych węzłach.

8.3.3. Aproksymacja różnicowa zagadnienia Dirichleta

Zbiór $D \cup \Gamma$, na którym poszukujemy rozwiązania zagadnienia Dirichleta (8.63), (8.64), pokrywamy prostokątną siatką prostych (rys. 8.11)

$$x = a + ih, \quad y = b + jk \quad (8.69)$$

Punkty przecięcia się prostych (8.69) nazywamy *węzłami siatki*. W dalszym ciągu rozpatrywać będziemy tylko te węzły, które należą do $D \cup \Gamma$.



8.11

Zbiór węzłów $\bar{\omega}$ (węzłów należących do $D \cup \Gamma$) dzielimy na dwa rozłączne podzbiory: zbiór węzłów wewnętrznych ω i zbiór węzłów brzegowych γ . Mówimy, że dwa węzły (i_1, j_1) oraz (i_2, j_2) są *sąsiednie*, jeżeli

$$\sum_{s=1}^2 |i_s - j_s| = 1 \quad (8.70)$$

Węzeł będziemy nazywać *wewnętrznym*, jeżeli każdy jego sąsiedni węzeł należy do $\bar{\omega}$; pozostałe węzły będziemy nazywać *brzegowymi*. Podzielimy z kolei zbiór γ na dwa podzbiory: $\gamma = \gamma_1 \cup \gamma_2$, gdzie γ_1 jest zbiorem węzłów brzegowych leżących na krzywej Γ , zaś γ_2 są pozostałymi węzłami brzegowymi. Na rysunku 8.11 znakami $\bullet$, $\circ$, $\times$ zostały oznaczone węzły odpowiednio ze zbiorów ω , γ_1 , γ_2 .

Założymy, że zbiór $\bar{\omega}$ jest *spójny*, tzn. że między każdymi dwoma węzłami ze zbioru ω można przejść po łańcuchu węzłów należących do ω , z których każde dwa kolejne są sąsiednie.

Jeżeli punkt $A \in \gamma_1$, to po prostu przyjmujemy

$$v(A) = \varphi(A) \quad (8.71)$$

i wartość $v(A)$ traktujemy jako już znaną.

Dla każdego węzła wewnętrznego (i, j) rozpatrujemy równanie różnicowe postaci (8.65). Mamy

$$L_{h,k} v_{ij} = \frac{v_{i+1,j} - 2v_{ij} + v_{i-1,j}}{h^2} + \frac{v_{i,j+1} - 2v_{ij} + v_{i,j-1}}{k^2} = f_{ij}, \quad (i, j) \in \omega \quad (8.72)$$

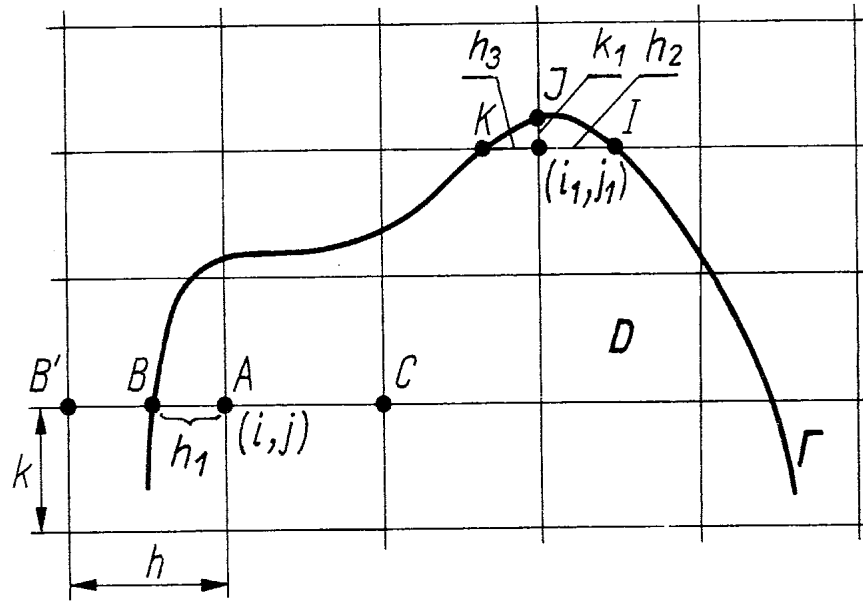
Dla każdego węzła brzegowego ze zbioru γ_2 piszemy równanie, w którym wykorzystujemy warunek brzegowy (8.64). Równania te można zapisywać w różny sposób. Najprostszy, ale i najmniej dokładny sposób polega na tym, że przyjmujemy

$$v_{ij} = \varphi(B_{ij}) \quad (i, j) \in \gamma_2 \quad (8.73)$$

gdzie $B_{i,j}$ jest najbliższym węzłowi (i, j) punktem zbioru Γ . Przenosząc więc wartość funkcji φ z punktu $B_{i,j}$ do węzła (i, j) popełniamy błąd $O(k+h)$. Sposób ten w praktyce nie jest często stosowany, ponieważ jest zbyt mało dokładny. Jego dokładność jest niewystarczająca dlatego, że popełniany błąd jest niższego rzędu niż błąd aproksymacji równania Poissona równaniem różnicowym (8.72). Później przekonamy się, że szybkość zbieżności do rozwiązania dokładnego zależy właśnie od tych dwóch błędów.

Wystarczająco dokładny sposób wyprowadzenia równań dla punktów ze zbioru γ_2 podał Collatz. Niech $A \in \gamma_2$. Istnieje wtedy węzeł B' sąsiedni z A i taki, że $B' \notin \bar{\omega}$. Niech B i C będą punktami leżącymi na linii siatki przechodzącej przez punkty A i B' takie, że $B \in \Gamma$ i B leży między punktami A i B' , a C jest sąsiedni z A . Załóżmy, że h, k są tak małe, iż punkt $C \in D$ (rys. 8.12). Wartość funkcji v w punkcie A wyznaczamy, stosując liniową interpolację między wartościami v w punktach B i C , czyli

$$v(A) = \frac{1}{h+h_1} (h_1 v(C) + h \varphi(B)) \quad (8.74)$$

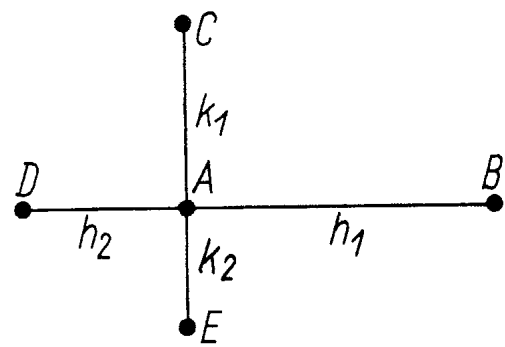


8.12

Otrzymany przy tym błąd wynosi $O(h^2 + k^2)$.

Dokładność tego sposobu jest wystarczająca, ale ma on następującą wadę. Otrzymany układ równań liniowych (tzn. równań (8.72)–(8.74)) nie jest układem symetrycznym, uważa się natomiast, że zagadnienie różnicowe powinno mieć podstawowe własności zagadnienia różniczkowego, a więc takie własności jak symetryczność, dodatnią określoność itp. Przedstawimy zatem inny sposób wyprowadzenia równań dla węzłów ze zbioru γ_2 , zapewniający symetryczność układu.

Niech A będzie dowolnym węzłem ze zbioru γ_2 . Rozpatrzmy punkty B, C, D, E (rys. 8.13). Zakładamy, że punkty te spełniają następujące warunki: $h_1, h_2 \leq h, k_1, k_2 \leq k$ i każdy z nich albo jest węzłem ze zbioru $\bar{\omega}$ sąsiadującym



8.13

z punktem A , albo jest punktem przecięcia się linii siatki z krzywą Γ . Dla każdego $A \in \gamma_2$ piszemy równanie różnicowe wiążące wartości rozwiązania v w pięciu punktach A, B, C, D, E :

$$l_{h,k} v(A) = \frac{1}{h} \left(\frac{v(B) - v(A)}{h_1} - \frac{v(A) - v(D)}{h_2} \right) + \frac{1}{k} \left(\frac{v(C) - v(A)}{k_1} - \frac{v(A) - v(E)}{k_2} \right) = f(A) \quad (8.75)$$

i podstawiamy w tym równaniu funkcję φ w miejsce v w punktach, które leżą na Γ . Wówczas dla węzłów $(i, j), (i_1, j_1)$ z rys. 8.12 otrzymamy następujące równania

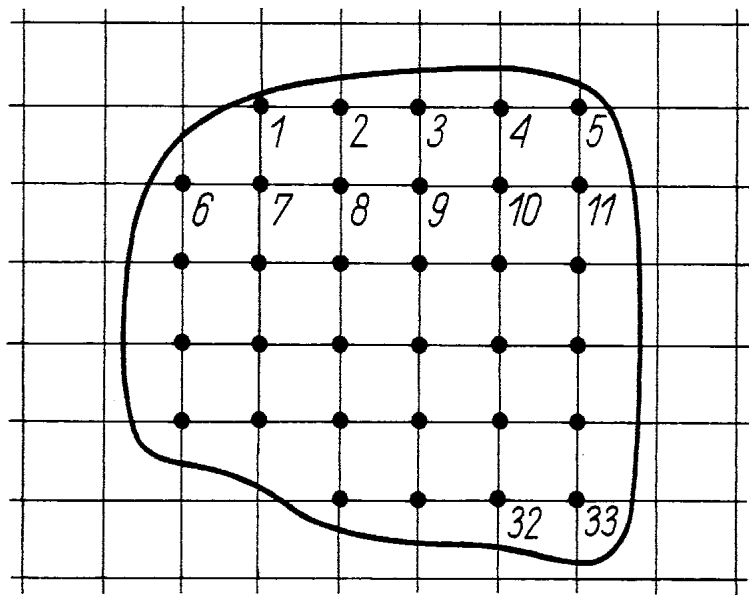
$$\frac{1}{k^2} (v_{i,j+1} - 2v_{i,j} + v_{i,j-1}) + \frac{1}{h} \left(\frac{v_{i+1,j} - v_{i,j}}{h} - \frac{v_{i,j} - \varphi(B)}{h_1} \right) = f_{ij}$$

$$\frac{1}{h} \left(\frac{\varphi(I) - v_{i_1, j_1}}{h_2} - \frac{v_{i_1, j_1} - \varphi(K)}{h_3} \right) + \frac{1}{k} \left(\frac{\varphi(J) - v_{i_1, j_1}}{k_1} - \frac{v_{i_1, j_1} - v_{i_1, j_1-1}}{k} \right) = f_{i_1, j_1}$$

Zauważmy, że jeżeli $h_1 = h_2 = h$, $k_1 = k_2 = k$, to równanie (8.75) staje się równaniem (8.72) aproksymującym równanie Poissona. Aproksymacja typu (8.75) jest często stosowana w praktyce.

W układach równań (8.72), (8.74) i (8.72), (8.75) niewiadomymi są wartości funkcji v w węzłach $\omega \cup \gamma_2$ i liczba niewiadomych jest równa liczbie równań (wartości v w węzłach leżących na Γ traktujemy jako wartości znane). Układy te zawsze posiadają jednoznaczne rozwiązania (patrz np. [4], [10], [55]).

Omówimy teraz własności układów równań (8.72), (8.74) i (8.72), (8.75). W zbiorze $\omega \cup \gamma_2$ wprowadzimy numerację przyporządkowując węzłom kolejne liczby naturalne — ponumerujemy je kolejno wierszami od lewej do prawej, tak jak na rys. 8.14.



8.14

Rozpatrzmy najpierw układ (8.72), (8.75). W równaniach tych w punktach leżących na Γ w miejsce v podstawiamy φ i zapisujemy je w postaci

$$\begin{aligned} -L_{h,k} v_{ij} &= F_{ij}, & (i, j) \in \omega \\ -l_{h,k} v_{ij} &= F_{ij}, & (i, j) \in \gamma_2 \end{aligned} \quad (8.76)$$

gdzie F zależy od f i φ . Równania (8.76) są równaniami postaci

$$\begin{aligned} v(A) \left[\frac{1}{h_1 h} + \frac{1}{h_2 h} + \frac{1}{k k_1} + \frac{1}{k k_2} \right] - \frac{1}{h_1 h} v(B) - \frac{1}{h h_2} v(D) - \\ - \frac{1}{k k_1} v(C) - \frac{1}{k k_2} v(E) = f(A), \quad h_1, h_2 \leq h; \quad k_1, k_2 \leq k \end{aligned} \quad (8.77)$$

przy czym wzajemne położenie punktów A, B, C, D, E jest takie, jak na rys. 8.13. Punkt A jest węzłem ze zbioru $\omega \cup \gamma_2$; jeżeli któryś z punktów C, D ,

E leży na Γ , to wtedy wartość funkcji v w tym punkcie traktujemy jako znaną i odpowiedni składnik należy przenieść na prawą stronę.

Układ (8.76) zapiszemy w postaci macierzowej $AV = F$, gdzie i -tą składową wektora V jest wartość szukanej funkcji v w węźle o numerze i . Jako i -te w kolejności równanie w układzie $AV = F$ napiszemy równanie dla węzła o numerze i . Macierz A ma postać

$$A = \begin{bmatrix} A_{11} & A_{12} & & & & 0 \\ A_{21} & A_{22} & A_{23} & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \cdot & A_{n-1,n} \\ 0 & & & & \cdot & & A_{n,n-1} & A_{n,n} \end{bmatrix} \quad (8.78)$$

w której A_{ij} są macierzami. Ze wzoru (8.77) bezpośrednio wynika, że w każdym wierszu macierzy A elementy stojące na głównej przekątnej są dodatnie i dominujące (tzn. suma modułów pozostałych elementów w wierszu nie przekracza wartości tego elementu). Ponadto z (8.77) wynika symetryczność macierzy A . Z twierdzenia 14 z p. 6.2 wynika, że macierz A jest nieujemnie określona. Jak wiemy, macierz A jest macierzą nieosobliwą. A zatem jest to macierz symetryczna i dodatnio określona.

Postępując w analogiczny sposób można również tak zapisać układ (8.72), (8.74), aby macierz A miała tę samą postać i była dodatnio określona.

Układy równań, które otrzymujemy dla zagadnienia Dirichleta są układami o dużej liczbie niewiadomych (kilkaset, a nawet kilka tysięcy), ale macierze tych układów są macierzami rzadkimi (tzn. duża liczba elementów tych macierzy jest zerami). Zauważmy, że w przypadku układów (8.72), (8.74) i (8.72), (8.75), a także (8.72), (8.73) w każdym wierszu macierzy A mamy co najwyżej 5 elementów różnych od zera. Metody rozwiązywania układów z macierzami rzadkimi są omówione w p. 5.5. Tutaj nadmieniamy tylko, że poważnym problemem przy rozwiązywaniu takich układów jest duże zapotrzebowanie na pamięć maszyny, często więc trzeba w takich przypadkach korzystać z pamięci zewnętrznej (dyski, bębny), co oczywiście komplikuje program i znacznie wydłuża czas liczenia.

Przeprowadzając analizę układu różnicowego należy rozpatrzyć następujące zagadnienia:

1) Zbieżność rozwiązania różnicowego do rozwiązania dokładnego, gdy $h, k \rightarrow 0$. Dla rozpatrywanych powyżej metod sprawy te omówimy w następnym punkcie.

2) Wpływ błędów zaokrągleń powstałych przy rozwiązywaniu układu na dokładność rozwiązania, przy czym gdy macierz A jest macierzą dodatnio określoną, wówczas dla omówionych w p. 5.5 metod rozwiązywania układu $AV = F$ błędy zaokrągleń mają niewielki wpływ na rozwiązanie różnicowe.

3) Wpływ zaburzeń funkcji f i φ , czyli prawej strony równania różniczkowego i warunku brzegowego na rozwiązanie różnicowe. W praktyce

funkcje f i φ są często obarczone błędami i dlatego powinno stosować się tylko takie metody, które zapewniają, że błędy te nie wypaczają rozwiązania różnicowego. Dla układu (8.72), (8.73) problemy te omówimy w następnym punkcie.

8.3.4. Jednostajna zbieżność rozwiązania różnicowego

Jeżeli $u \in C^4(\bar{D})$, to dla układu (8.72), (8.73) prawdziwe jest następujące oszacowanie błędu

$$\max_{(i,j) \in \bar{\omega}} |u_{ij} - v_{ij}| \leq N_1(h+k) \quad (8.79)$$

a dla układów (8.72), (8.74) i (8.72), (8.75)

$$\max_{(i,j) \in \bar{\omega}} |u_{ij} - v_{ij}| \leq N_2(h^2 + k^2) \quad (8.80)$$

gdzie N_1, N_2 są stałymi niezależnymi od h, k (patrz np. [4], [10], [55]). Rozwiązanie różnicowe v wymienionych układów jest jednostajnie zbieżne do rozwiązania dokładnego u zagadnienia Dirichleta (8.63), (8.64), gdy $h, k \rightarrow 0$.

Ze względu na małą dokładność rozwiązania układ (8.72), (8.73) jest rzadko stosowany do rozwiązywania zagadnienia Dirichleta, ale ze względu na jego prostą budowę jest on często analizowany. Korzystając z zasady maksimum w punkcie tym udowodnimy istnienie i jednoznaczność rozwiązania tego właśnie układu oraz oszacowanie (8.79). Oszacowanie (8.80) dla układów (8.72), (8.74) i (8.72), (8.75) dowodzimy podobnie.

Zapiszemy jeszcze raz wzory (8.72), (8.73)

$$\begin{aligned} L_{h,k} v_{i,j} &= f_{i,j}, & (i,j) \in \omega \\ v_{i,j} &= \varphi(B_{ij}) & (i,j) \in \gamma \end{aligned} \quad (8.81)$$

Punkt B_{ij} jest najbliższym węzłowi $(i,j) \in \gamma$ punktem zbioru Γ ; jeżeli węzeł $(i,j) \in \Gamma$, to wtedy jako B_{ij} przyjmujemy (i,j) .

Tw. 1. (zasada maksimum). *Jeżeli funkcja $v \neq \text{const}$ oraz $L_{h,k} v_{ij} \geq 0$ ($L_{h,k} v_{ij} \leq 0$) dla każdego węzła $(i,j) \in \omega$, to funkcja v nie może przyjmować największej dodatniej (najmniejszej ujemnej) wartości w węźle wewnętrznym ω .*

Tw. 2. *Jeżeli wartości dwóch funkcji v, Y spełniają nierówności*

$$\begin{aligned} |L_{h,k} v_{ij}| &\leq -L_{h,k} Y_{ij}, & (i,j) \in \omega \\ |v_{ij}| &\leq Y_{ij}, & (i,j) \in \gamma \end{aligned} \quad (8.82)$$

to

$$|v_{ij}| \leq Y_{ij}, \quad (i,j) \in \bar{\omega}$$

Dowody tych twierdzeń są analogiczne do dowodów twierdzeń 1 i 2, podanych w p. 8.1.3.

Wniosek. Układ równań (8.81) ma jednoznaczne rozwiązanie dla dowolnych funkcji f, φ .

DOWÓD. Aby tego dowieść, wystarczy pokazać, że odpowiadający układowi (8.81) układ jednorodny

$$\begin{aligned} L_{h,k} v_{ij} &= 0, & (i, j) \in \omega \\ v_{ij} &= 0, & (i, j) \in \gamma \end{aligned} \quad (8.83)$$

ma tylko rozwiązanie zerowe. Na podstawie twierdzenia 1 wśród rozwiązań v_{ij} układu (8.83) nie może być ani liczb dodatnich, ani liczb ujemnych. Zatem układ (8.83) ma tylko rozwiązanie zerowe, co kończy dowód.

Funkcję v przedstawimy teraz w postaci sumy funkcji

$$v = z + q$$

gdzie funkcje z, q są określone w następujący sposób

$$\begin{cases} L_{h,k} z_{i,j} = 0, & (i, j) \in \omega \\ z_{i,j} = \varphi(B_{i,j}), & (i, j) \in \gamma \end{cases} \quad \begin{cases} L_{h,k} q_{i,j} = f_{i,j}, & (i, j) \in \omega \\ q_{i,j} = 0, & (i, j) \in \gamma \end{cases}$$

Na podstawie twierdzenia 1 mamy

$$\max_{(i,j) \in \bar{\omega}} |z_{i,j}| \leq \max_{(i,j) \in \gamma} |z_{i,j}| \leq \max_{A \in \Gamma} |\varphi(A)| \quad (8.84)$$

Oszacowanie dla funkcji q wyprowadzimy na podstawie twierdzenia 2. Przyjmujemy

$$Y(x, y) = K(R^2 - x^2 - y^2) \quad (8.85)$$

gdzie R jest promieniem koła, zawierającego wewnątrz obszar D , a K jest stałą dodatnią. Stałą K dobierzemy tak, aby był spełniony warunek (8.82). Mamy $L_{h,k} Y_{i,j} = -4K$. Przyjmujemy zatem

$$K = \max_{(i,j) \in \omega} \frac{|f_{i,j}|}{4}$$

Funkcje q, Y spełniają założenia twierdzenia 2. Otrzymujemy więc następujące oszacowanie dla funkcji q

$$\max_{(i,j) \in \bar{\omega}} |q_{i,j}| \leq \max_{(i,j) \in \bar{\omega}} |Y_{i,j}| \leq \frac{R^2}{4} \max_{(i,j) \in \bar{\omega}} |f_{i,j}| \leq \frac{R^2}{4} \max_{A \in \bar{D}} |f(A)| \quad (8.86)$$

Na podstawie wzorów (8.84), (8.86) dla rozwiązania v układu (8.81) prawdziwa jest nierówność

$$\max_{(i,j) \in \bar{\omega}} |v_{i,j}| \leq \max_{A \in \Gamma} |\varphi(A)| + \frac{R^2}{4} \max_{A \in \bar{D}} |f(A)| \quad (8.87)$$

Zatem udowodniliśmy ciągłą zależność rozwiązania różnicowego v od funkcji f i φ .

Błąd $\varepsilon = u - v$ jest rozwiązaniem układu

$$\begin{aligned} L_{h,k} \varepsilon_{ij} &= F_{ij}, & (i, j) \in \omega \\ \varepsilon_{i,j} &= \psi_{ij}, & (i, j) \in \gamma \end{aligned} \quad (8.88)$$

gdzie F_{ij} , ψ_{ij} możemy oszacować w następujący sposób

$$|F_{ij}| \leq \frac{M_4}{12} (h^2 + k^2), \quad |\psi_{ij}| \leq M_1(k + h)$$

$$\text{przy czym } M_s = \max \left\{ \max_{\bar{D}} \left| \frac{\partial^s u}{\partial x^s} \right|, \max_{\bar{D}} \left| \frac{\partial^s u}{\partial y^s} \right| \right\}.$$

Na podstawie nierówności (8.87) oszacowanie błędu przyjmie postać

$$\max_{(i,j) \in \bar{\omega}} |u_{ij} - v_{ij}| \leq M_1(k + h) + \frac{R^2}{48} (h^2 + k^2) M_4 \quad (8.89)$$

Jeżeli D jest prostokątem, to siatkę możemy tak dobrać, aby cały brzeg Γ leżał na liniach tworzących siatkę ($\gamma = \gamma_1$). Wtedy $\psi_{ij} = 0$ i

$$\max_{(i,j) \in \bar{\omega}} |u_{ij} - v_{ij}| \leq \frac{R^2}{48} (h^2 + k^2) M_4 \quad (8.90)$$

Rozpatrzmy teraz wpływ zaburzeń funkcji f , φ na rozwiązanie różnicowego układu równań (8.72), (8.73). Niech v , $\tilde{v}$ będą rozwiązaniami tego układu odpowiednio dla funkcji f , φ i $\tilde{f}$, $\tilde{\varphi}$. Z liniowości równań różnicowych wynika, że wartości $(v_{ij} - \tilde{v}_{ij})$ są rozwiązaniem rozpatrywanego układu, gdy za f , φ przyjmiemy odpowiednio $f - \tilde{f}$, $\varphi - \tilde{\varphi}$. Z nierówności (8.87) otrzymujemy wówczas następujące oszacowanie zaburzenia rozwiązania różnicowego

$$\max_{(i,j) \in \bar{\omega}} |v_{ij} - \tilde{v}_{ij}| \leq \max_{A \in \Gamma} |\varphi(A) - \tilde{\varphi}(A)| + \frac{R^2}{4} \max_{A \in \bar{D}} |f(A) - \tilde{f}(A)| \quad (8.91)$$

Zatem małe co do wartości bezwzględnej zaburzenia funkcji f i φ powodują małe co do wartości bezwzględnej zmiany rozwiązania różnicowego w całym zbiorze $\bar{\omega}$ i zmiany te można oszacować niezależnie od h , k .

Tę samą własność mają także rozwiązania układów (8.72), (8.74) i (8.72), (8.75).

Ćwiczenia

1. Niech $D \cup \Gamma$ będzie prostokątem i niech siatka (8.69) będzie tak dobrana, że cały brzeg Γ leży na liniach tworzących tę siatkę. Jako węzły wewnętrzne przyjmujemy węzły leżące wewnątrz D . W każdym węźle wewnętrznym równanie Poissona zastępujemy równaniem różnicowym (8.67), a w węzłach brzegowych rozwiązanie wyznaczamy z warunku brzegowego (8.64):

$$\text{a) Udowodnić, że dla } \frac{1}{\sqrt{5}} \leq \frac{h}{k} \leq \sqrt{5} \text{ otrzymany układ równań różnicowych ma jedno-}$$

znaczne rozwiązanie.

W s k a z ó w k a. Udowodnić zasadę maksimum dla rozwiązania różnicowego.

b) Udowodnić, że macierz otrzymanego układu równań może być przedstawiona w postaci (8.78). Pokazać, że dla $\frac{1}{\sqrt{5}} \leq \frac{h}{k} \leq \sqrt{5}$ elementy na głównej przekątnej tej macierzy są dominujące.

8.4. METODA PROSTYCH DLA RÓWNANIA PRZEWODNICTWA CIEPLNEGO

Rozpatrzmy postawione w p. 8.1.1 zagadnienie brzegowe: znaleźć funkcję $u = u(x, t)$ spełniającą wewnątrz prostokąta $\bar{D} = \{(x, t): 0 \leq x \leq l, 0 \leq t \leq T\}$ równanie

$$Lu = \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f(x, t) \quad (8.92)$$

warunek początkowy

$$u(x, 0) = g_1(x), \quad 0 \leq x \leq l \quad (8.93)$$

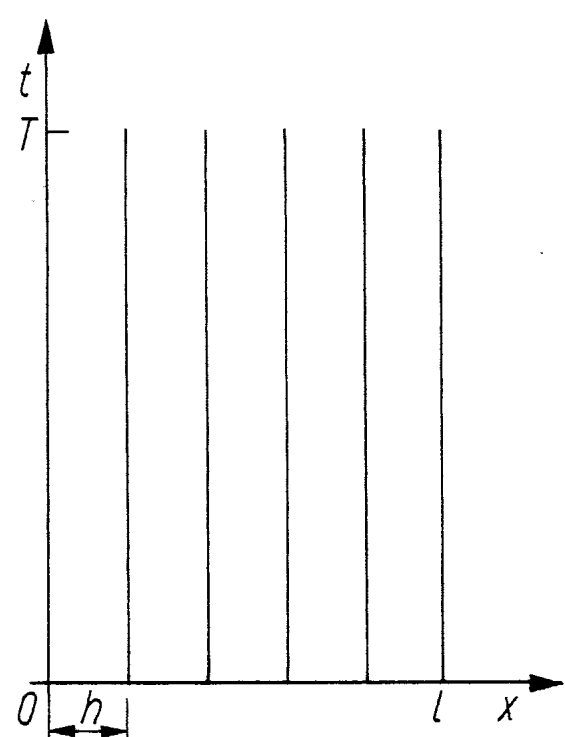
i warunki brzegowe

$$u(0, t) = \varphi(t), \quad u(l, t) = \psi(t), \quad 0 \leq t \leq T \quad (8.94)$$

przy czym przybliżonego rozwiązania powyższego zagadnienia poszukiwać będziemy za pomocą *metody prostych*. W tym celu zbiór $\bar{D}$ pokrywamy rodziną prostych

$$x = x_i = ih \quad (8.95)$$

gdzie $h = \frac{l}{N}$, N — liczba całkowita; $i = 0, 1, \dots, N$ (rys. 8.15).



8.15

Zastępując pochodną $\frac{\partial^2 u}{\partial x^2}$ na prostych $x = x_i$ ($i = 1, \dots, N-1$) wyrażeniem różnicowym

$$\frac{1}{h^2} [u(x_{i+1}, t) - 2u(x_i, t) + u(x_{i-1}, t)] \quad (8.96)$$

otrzymujemy następujący układ równań różniczkowych

$$L_h v_i(t) = v_i'(t) - \frac{1}{h^2} [v_{i+1}(t) - 2v_i(t) + v_{i-1}(t)] = f_i(t) \quad i = 1, \dots, N-1$$

gdzie $v_i(t)$ oznacza przybliżone rozwiązanie na prostej $x = x_i$, a $f_i(t) = f(x_i, t)$. Na podstawie warunków (8.93), (8.94) otrzymujemy

$$\begin{aligned} v_i(0) &= g_1(x_i), \quad i = 1, \dots, N-1 \\ v_0(t) &= \varphi(t), \quad v_N(t) = \psi(t), \quad 0 \leq t \leq T \end{aligned} \quad (8.97)$$

Równania (8.96), (8.97) zapiszemy w postaci macierzowej

$$V'(t) = \frac{1}{h^2} A V(t) + F(t), \quad V(0) = G_1 \quad (8.98)$$

gdzie

$$\begin{aligned} A &= \begin{bmatrix} -2 & 1 & & & 0 \\ 1 & -2 & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot \\ 0 & & & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}, \quad V(t) = \begin{bmatrix} v_1(t) \\ v_2(t) \\ \vdots \\ v_{N-2}(t) \\ v_{N-1}(t) \end{bmatrix}, \\ F(t) &= \begin{bmatrix} f_1(t) + \frac{1}{h^2} \varphi(t) \\ f_2(t) \\ \vdots \\ f_{N-2}(t) \\ f_{N-2}(t) + \frac{1}{h^2} \psi(t) \end{bmatrix}, \quad G_1 = \begin{bmatrix} g_1(x_1) \\ g_1(x_2) \\ \vdots \\ g_1(x_{N-2}) \\ g_1(x_{N-1}) \end{bmatrix} \end{aligned} \quad (8.99)$$

Przedstawiona w tym punkcie metoda prostych dla zagadnienia (8.92)–(8.94) polega na rozwiązaniu zagadnienia początkowego (8.98). Poszczególnymi składowymi wektora $V(t)$ są przybliżone wartości rozwiązania na prostych $x = x_i$.

Można udowodnić (patrz np. [10], [60]), że jeżeli $u \in C_2^4(\bar{D})$, to

$$|u(x_i, t) - v_i(t)| \leq Mh^2 \quad (8.100)$$

gdzie M jest stałą niezależną od h .

Rozwiązanie zagadnienia (8.98) jest więc zbieżne do rozwiązania zagadnienia (8.92)–(8.94) i szybkość tej zbieżności wynosi $O(h^2)$.

8.5. METODA PROSTYCH DLA RÓWNANIA DRGAŃ STRUNY

Rozpatrzmy obecnie metodę prostych dla przedstawionego w p. 8.2.1 zagadnienia brzegowego: znaleźć funkcję $u = u(x, t)$ spełniającą wewnątrz prostokąta $\bar{D} = \{(x, t): 0 \leq x \leq l, 0 \leq t \leq T\}$ równanie

$$Lu = \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = f(x, y) \quad (8.101)$$

warunki początkowe

$$u(x, 0) = g_1(x), \quad \frac{\partial u(x, 0)}{\partial t} = g_2(x), \quad 0 \leq x \leq l \quad (8.102)$$

i warunki brzegowe

$$u(0, t) = \varphi(t), \quad u(l, t) = \psi(t), \quad 0 \leq t \leq T \quad (8.103)$$

Tak samo jak w poprzednim punkcie, zbiór $\bar{D}$ pokrywamy rodziną prostych postaci (8.95) i pochodną $\frac{\partial^2 u}{\partial x^2}$ na prostych $x = ih$ zastępujemy wyrażeniem różnicowym (8.96).

Dla zagadnienia (8.101)–(8.103) otrzymamy układ równań postaci

$$L_h v_i(t) = v_i''(t) - \frac{1}{h^2} [v_{i+1}(t) - 2v_i(t) + v_{i-1}(t)] = f_i(t), \quad i = 1, \dots, N-1 \quad (8.104)$$

$$v_0(t) = \varphi(t), \quad v_N(t) = \psi(t), \quad 0 \leq t \leq T$$

z warunkami początkowymi

$$v_i(0) = g_1(x_i), \quad v_i'(0) = g_2(x_i), \quad i = 1, \dots, N-1 \quad (8.105)$$

gdzie $v_i(t)$ oznacza przybliżone rozwiązanie na prostej $x = x_i$, a $f_i(t) = f(x_i, t)$.

Zapiszemy zagadnienie (8.104), (8.105) w postaci macierzowej

$$\begin{aligned} V''(t) &= \frac{1}{h^2} AV(t) + F(t) \\ V(0) &= G_1, \quad V'(0) = G_2 \end{aligned} \quad (8.106)$$

gdzie A , V , F , G_1 są określone wzorami (8.99), a

$$G_2 = \begin{bmatrix} g_2(x_1) \\ \vdots \\ g_2(x_{N-1}) \end{bmatrix}$$

Można udowodnić (patrz np. [10]), że jeżeli $u \in C_4^4(\bar{D})$, to

$$|u(x_i, t) - v_i(t)| \leq M_1 h^2 \quad (8.107)$$

gdzie M_1 jest stałą niezależną od h .

Jak widać, rozwiązanie zagadnienia (8.106) jest zbieżne z rozwiązaniem zagadnienia (8.101)–(8.103) i szybkość tej zbieżności wynosi $O(h^2)$.

DODATEK

METODA ELEMENTU SKOŃCZONEGO

D.1. POJĘCIA WSTĘPNE

Rozpatrzmy przestrzeń liniową F , w której jest zdefiniowany iloczyn skalarny $\langle \cdot, \cdot \rangle$. Elementami tej przestrzeni są funkcje z określonymi w sposób naturalny działaniami dodawania funkcji i mnożenia funkcji przez liczbę. Niech $U \subset F$ będzie liniową podprzestrzenią F , w której jest określony ten sam iloczyn skalarny $\langle \cdot, \cdot \rangle$ co w F i niech A będzie liniowym operatorem różniczkowym określonym na przestrzeni U o wartościach należących do F . Załóżmy, że A jest operatorem symetrycznym i dodatnio określonym tzn.:

- 1) $\langle Au, v \rangle = \langle u, Av \rangle$ dla każdego $u, v \in U$
- 2) $\langle Au, u \rangle > 0$ dla wszystkich $u \in U$ z wyjątkiem $u = 0$.

Łatwo można wykazać, że wzór

$$\langle u, v \rangle_A = \langle Au, v \rangle \text{ dla każdego } u, v \in U \quad (\text{D.1})$$

określa nowy iloczyn skalarny w przestrzeni U .

Iloczyn skalarny $\langle u, v \rangle_A$ nazywać będziemy *iloczynem energetycznym względem operatora A* , a przestrzeń U z tym iloczynem skalarnym — *przestrzenią energetyczną operatora A* .

Normę $\|u\|_A = \sqrt{\langle u, u \rangle_A}$ nazywać będziemy *normą energetyczną*.

W celu wyjaśnienia wprowadzonych pojęć rozpatrzmy teraz następujące proste przykłady.

Przykład 1. Niech $U = C^1(\langle 0; 1 \rangle)$ oraz niech $u, v \in U$. Wzory

$$1) \langle u, v \rangle = \int_0^1 u(x) v(x) dx$$

$$2) \langle u, v \rangle = \int_0^1 [u'(x) v'(x) + u(x) v(x)] dx$$

$$3) \langle u, v \rangle = \int_0^1 \sigma(x) u(x) v(x) dx$$

gdzie $\sigma(x) > 0$ jest ustaloną funkcją ciągłą na przedziale $\langle 0; 1 \rangle$, określają iloczyny skalarne w przestrzeni U .

Przykład 2. Rozpatrzmy przestrzeń $F = C(\langle 0; 1 \rangle)$ z iloczynem skalarnym $\langle u, v \rangle = \int_0^1 u(x) v(x) dx$. Niech ponadto $U = \{u(x) \in C^2(\langle 0; 1 \rangle) : u(0) = u(1) = 0\}$ oraz $Au(x) = -u''(x) + \sigma(x) u(x)$, gdzie $\sigma(x)$ jest ustaloną funkcją ciągłą na przedziale $\langle 0; 1 \rangle$. Założymy, że $\sigma(x) \geq 0$ i $\sigma(x) \not\equiv 0$ na $\langle 0; 1 \rangle$. Jak łatwo można sprawdzić, U jest liniową podprzestrzenią F , zaś A jest liniowym operatorem różniczkowym określonym na przestrzeni U o wartościach należących do F .

Pokażemy, że operator A jest operatorem symetrycznym i dodatnio określonym. Dla każdego $u, v \in U$ mamy

$$\langle Au, v \rangle = \int_0^1 [-u''(x) + \sigma(x) u(x)] v(x) dx = - \int_0^1 u''(x) v(x) dx + \int_0^1 \sigma(x) u(x) v(x) dx \quad (D.2)$$

Całkując przez części, przekształcimy pierwszy składnik powyższej równości następująco

$$\begin{aligned} \int_0^1 u''(x) v(x) dx &= u'(x) v(x) \Big|_0^1 - \int_0^1 u'(x) v'(x) dx = \\ &= - \int_0^1 u'(x) v'(x) dx = -u(x) v'(x) \Big|_0^1 + \int_0^1 u(x) v''(x) dx = \int_0^1 u(x) v''(x) dx \end{aligned} \quad (D.3)$$

Stąd mamy

$$\langle Au, v \rangle = - \int_0^1 u(x) v''(x) dx + \int_0^1 \sigma(x) u(x) v(x) dx = \langle u, Av \rangle$$

Zatem operator A jest symetryczny. Z podanych powyżej przekształceń wynika ponadto, że

$$\langle Au, u \rangle = \int_0^1 (u'(x))^2 dx + \int_0^1 \sigma(x) (u(x))^2 dx \quad (D.4)$$

Stąd łatwo zauważyć, że $\langle Au, u \rangle > 0$ dla wszystkich funkcji $u(x) \in U$ z wyjątkiem $u(x) \equiv 0$. Udowodniliśmy zatem, że operator A jest dodatnio określony.

Przykład 3. Niech D będzie obszarem na płaszczyźnie oraz niech brzeg Γ tego obszaru będzie kawałkami gładkim konturem. Przyjmijmy, że $F = C(\bar{D})$ i $U = \{u \in C^2(\bar{D}) : u(P) = 0 \text{ dla } P \in \Gamma\}$. Niech $u(x, y), v(x, y) \in U$. Iloczyn skalarny określmy w następujący sposób: $\langle u, v \rangle = \int_D u(x, y) v(x, y) dx dy$. Ponadto, niech

$$Au = - \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right)$$

Widzimy zatem, że U jest liniową podprzestrzenią F . Pokażemy teraz, że operator A jest symetryczny i dodatnio określony. Mamy

$$\begin{aligned} \langle Au, v \rangle &= - \int_D \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) v dx dy = \\ &= \int_D \left[- \frac{\partial}{\partial x} \left(v \frac{\partial u}{\partial x} \right) - \frac{\partial}{\partial y} \left(v \frac{\partial u}{\partial y} \right) \right] dx dy + \int_D \left[\frac{\partial v}{\partial x} \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \frac{\partial u}{\partial y} \right] dx dy \end{aligned}$$

Na podstawie twierdzenia Greena pierwsza całka jest równa całce

$$\int_D \left(v \frac{\partial u}{\partial y} \right) dx + \left(-v \frac{\partial u}{\partial x} \right) dy$$

ta zaś jest równa zeru, ponieważ $v = 0$ na Γ . Otrzymaliśmy więc

$$\langle Au, v \rangle = \int_D \left[\frac{\partial u}{\partial x} \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial v}{\partial y} \right] dx dy \quad (D.5)$$

Funkcja u także znika na Γ . Powtarzając powyższe rozumowanie stwierdzamy, że iloczyn $\langle u, Av \rangle$ jest równy prawej stronie wzoru (D.5), a więc operator A jest symetryczny. Dodatnia określoność A wynika natychmiast z (D.5).

D.2. METODA ELEMENTU SKOŃCZONEGO

Niech przestrzenie U , F oraz operator różniczkowy A spełniają założenia z poprzedniego punktu i niech ponadto U_N oznacza N -wymiarową podprzestrzeń U .

Rozpatrzmy równanie różniczkowe

$$Au = f \quad (D.6)$$

gdzie f jest daną funkcją należącą do F , zakładając, że równanie to ma w przestrzeni U jednoznaczne rozwiązanie. Rozwiązanie równania (D.6) można przybliżyć funkcjami z przestrzeni U_N , przy czym jako przybliżone rozwiązanie równania (D.6) przyjmuje się taki element $u_N \in U_N$, że

$$\|u - u_N\|_A = \min_{\bar{u} \in U_N} \|u - \bar{u}\|_A \quad (D.7)$$

Można udowodnić (patrz np. [49]), że istnieje dokładnie jeden taki element u_N .

Powyższy sposób wyznaczania przybliżonego rozwiązania równania (D.6) nazywamy *metodą Rayleigha-Ritza*. Jeżeli elementami przestrzeni U_N są wielomianowe funkcje sklepane, to taką metodę nazywać będziemy *metodą elementu skończonego*, a funkcję u_N — *elementem skończonym* aproksymującym rozwiązanie u z przestrzeni U_N .

Podamy teraz algorytm wyznaczania elementu u_N . Rozpatrzmy ogólny przypadek, gdy elementami U_N są funkcje dowolnego typu, a niekoniecznie funkcje sklepane. Niech funkcje Φ_i ($i = 1, \dots, N$) stanowią bazę przestrzeni U_N . Każdy więc element $\bar{u} \in U_N$ można przedstawić w postaci

$$\bar{u} = c_1 \Phi_1 + \dots + c_N \Phi_N \quad (D.8)$$

gdzie c_i są liczbami rzeczywistymi. Elementu u_N poszukiwać będziemy w tej właśnie postaci. Mamy

$$\begin{aligned} g(c_1, \dots, c_N) &= \left\| u - \bar{u} \right\|_A^2 = \left\| u - \sum_{i=1}^N c_i \Phi_i \right\|_A^2 = \\ &= \left\langle u - \sum_{i=1}^N c_i \Phi_i, u - \sum_{i=1}^N c_i \Phi_i \right\rangle_A = \\ &= \langle u, u \rangle_A - 2 \sum_{j=1}^N c_j \langle u, \Phi_j \rangle_A + \sum_{i,j=1}^N c_i c_j \langle \Phi_i, \Phi_j \rangle_A \end{aligned}$$

Element z przestrzeni U_N , którego współczynnikami w rozwinięciu (D.8) są liczby $c_1^*, \dots, c_N^*$ takie że

$$g(c_1^*, \dots, c_N^*) = \min_{(c_1, \dots, c_N) \in R^N} g(c_1, \dots, c_N) \quad (\text{D.9})$$

jest poszukiwanym elementem u_N . W celu wyznaczenia minimum funkcji $g(c_1, \dots, c_N)$ obliczamy pochodne $\partial g / \partial c_k$, $k = 1, \dots, N$, i przyrównujemy je do zera

$$\frac{\partial g}{\partial c_k} = -2 \langle u, \Phi_k \rangle_A + 2 \sum_{j=1}^N c_j \langle \Phi_k, \Phi_j \rangle_A = 0, \quad k = 1, \dots, N \quad (\text{D.10})$$

Funkcja u jest rozwiązaniem równania $Au = f$, a więc

$$\langle u, \Phi_k \rangle_A = \langle Au, \Phi_k \rangle = \langle f, \Phi_k \rangle \quad (\text{D.11})$$

Po podstawieniu (D.11) do równań (D.10) otrzymamy

$$\sum_{j=1}^N \varphi_{k,j} c_j = \langle f, \Phi_k \rangle, \quad k = 1, \dots, N \quad (\text{D.12})$$

gdzie $\varphi_{k,j} = \langle \Phi_k, \Phi_j \rangle_A$. Równania (D.12) tworzą układ N równań z N niewiadomymi c_k , $k = 1, \dots, N$. Można udowodnić, że macierz tego układu jest macierzą nieosobliwą. Układ (D.12) posiada więc jednoznaczne rozwiązanie. Rozwiązanie to daje punkt, w którym funkcja g osiąga minimum.

Po ustaleniu więc bazy Φ_i , $i = 1, \dots, N$, w przestrzeni U_N (w przestrzeni, w której poszukujemy przybliżonego rozwiązania równania (D.6)), należy rozwiązać układ (D.12). Można udowodnić, że macierz układu (D.12) jest macierzą symetryczną i dodatnio określoną. Są to fakty ważne z numerycznego punktu widzenia. Dla takich układów jest opracowanych wiele efektywnych algorytmów ich rozwiązywania.

Metodę elementu skończonego można stosować do rozwiązywania zarówno równań różniczkowych zwyczajnych, jak i cząstkowych (częściej jest ona stosowana do rozwiązywania równań cząstkowych). W tej książce ograniczymy się do omówienia tej metody na przykładzie prostego równania różniczkowego zwyczajnego. Zastosowanie do równań cząstkowych wymaga omówienia wielomianowych funkcji sklejanых wielu zmiennych niezależnych, co wykracza poza zakres tej książki.

Rozpatrzmy następujące zagadnienie brzegowe: wyznaczyć funkcję $\bar{u}(x)$ spełniającą

$$\begin{aligned} A\bar{u}(x) &= -\bar{u}''(x) + \sigma(x)\bar{u}(x) = \bar{f}(x), & 0 \leq x \leq 1 \\ \bar{u}(0) &= \alpha, & \bar{u}(1) = \beta \end{aligned} \quad (\text{D.13})$$

gdzie $\sigma(x)$, $\bar{f}(x)$ — ustalone funkcje ciągłe na przedziale $\langle 0; 1 \rangle$, $\sigma(x) \geq 0$ i $\sigma(x) \not\equiv 0$ na $\langle 0; 1 \rangle$, a α, β — ustalone liczby rzeczywiste. Jak wiadomo, tak postawione zagadnienie ma jednoznaczne rozwiązanie $\bar{u}(x) \in C^2(\langle 0; 1 \rangle)$.

Zagadnienie (D.13) sprowadzimy do zagadnienia z jednorodnymi warunkami brzegowymi. Można to zrobić na przykład w następujący sposób:

niech $v(x) \in C^2(\langle 0; 1 \rangle)$ będzie ustaloną funkcją spełniającą warunki $v(0) = \alpha$, $v(1) = \beta$. Funkcja $u(x) = \bar{u}(x) - v(x)$ spełnia więc równania

$$Au(x) = -u''(x) + \sigma(x)u(x) = f(x), \quad 0 \leq x \leq 1 \quad (\text{D.14})$$

$$u(0) = 0, \quad u(1) = 0 \quad (\text{D.15})$$

gdzie $f(x) = \bar{f}(x) - Av(x)$.

Dla przykładu, przybliżonego rozwiązania zagadnienia (D.14), (D.15) poszukiwać będziemy w postaci funkcji sklejaney stopnia trzeciego z równoodległymi węzłami (patrz p. 1.3.2). Rozpatrzmy przestrzeń

$F = C(\langle 0; 1 \rangle)$ z iloczynem skalarnym

$$\langle u, v \rangle = \int_0^1 u(x) v(x) dx \quad i$$

$$U = \{u(x) \in C^2(\langle 0; 1 \rangle) : u(0) = u(1) = 0\}$$

Równanie (D.14) w przestrzeni U ma jednoznaczne rozwiązanie. Przestrzeń U jest liniową podprzestrzenią F , a operator A liniowym operatorem różniczkowym określonym na U o wartościach należących do F . Na podstawie przykładu 2 z poprzedniego punktu operator A jest operatorem symetrycznym i dodatnio określonym. Zatem wzór $\langle u, v \rangle_A = \langle Au, v \rangle$, $u, v \in U$, określa nowy iloczyn skalarny w przestrzeni U . Na podstawie (D.2), (D.3) iloczyn ten możemy zapisać w postaci

$$\langle Au, v \rangle = \int_0^1 [u'(x) v'(x) + \sigma(x) u(x) v(x)] dx \quad (\text{D.16})$$

Niech $\Delta_n: 0 = x_0 < x_1 < \dots < x_n = 1$; przypominamy, że w p. 1.3.1 symbolem $S_k(\Delta_n)$ oznaczyliśmy zbiór funkcji sklejanых k -tego stopnia z węzłami w punktach $x_i, i = 0, 1, \dots, n$. Załóżmy teraz dodatkowo, że węzły te są równoodległe, czyli $x_i = ih, h = \frac{1}{n}$, i przyjmijmy

$$X_n = \{s(x) \in S_3(\Delta_n) : s(0) = s(1) = 0\}$$

Przypominamy, że bazę przestrzeni $S_3(\Delta_n)$ stanowią funkcje $\bar{\Phi}_i(x)$, $i = -1, 0, 1, \dots, n+1$, określone na przedziale $\langle 0; 1 \rangle$ w następujący sposób

$$\bar{\Phi}_i(x) = \Phi_i^3(x), \quad 0 \leq x \leq 1$$

gdzie $\Phi_i^3(x)$ są funkcjami określonymi wzorem (1.65). Wykres funkcji $\Phi_i^3(x)$ jest podany na rys. 1.10.

Widzimy, że $\Phi_i^3(0) = \Phi_j^3(1) = 0$ dla $j = 2, 3, \dots, n-2$. Zatem $\bar{\Phi}_j(x) \in U_N, j = 2, \dots, n-2$, i funkcje te weźmiemy jako część bazy U_N . Pozostałe funkcje $\bar{\Phi}_i(x), i = -1, 0, 1, n-1, n, n+1$, nie należą do U_N , ponieważ nie znikają w punktach $x = 0, 1$. Można udowodnić (patrz [49]), że funkcje $\bar{\Phi}_i(x), i = 0, 1, \dots, n$, określone następująco

$$\begin{aligned}
\tilde{\Phi}_0(x) &= \bar{\Phi}_0(x) - 4\bar{\Phi}_{-1}(x), & \tilde{\Phi}_1(x) &= \bar{\Phi}_0(x) - 4\bar{\Phi}_1(x) \\
\tilde{\Phi}_i(x) &= \bar{\Phi}_i(x), & i &= 2, \dots, n-2 \\
\tilde{\Phi}_{n-1}(x) &= \bar{\Phi}_n(x) - 4\bar{\Phi}_{n-1}(x), & \tilde{\Phi}_n(x) &= \bar{\Phi}_n(x) - 4\bar{\Phi}_{n+1}(x)
\end{aligned} \tag{D.17}$$

stanowią bazę przestrzeni U_N . Wymiar tej przestrzeni wynosi więc $n+1$.

Funkcje $\tilde{\Phi}_i(x)$, $i = 0, 1, \dots, n$, wybierzemy jako bazę, w której przedstawimy poszukiwany element skończony u_N . Mamy więc

$$u_N(x) = c_0 \tilde{\Phi}_0(x) + \dots + c_n \tilde{\Phi}_n(x) \tag{D.18}$$

Współczynniki rozwinięcia c_i , $i = 0, 1, \dots, n$, są rozwiązaniem układu równań

$$\sum_{j=1}^n \varphi_{ij} c_j = \langle f, \tilde{\Phi}_i \rangle, \quad i = 0, 1, \dots, n \tag{D.19}$$

w którym

$$\begin{aligned}
\varphi_{ij} &= \int_0^1 [-\tilde{\Phi}_i''(x) + \sigma(x) \tilde{\Phi}_i(x)] \tilde{\Phi}_j(x) dx = \\
&= \int_0^1 [\tilde{\Phi}_i'(x) \tilde{\Phi}_j'(x) + \sigma(x) \tilde{\Phi}_i(x) \tilde{\Phi}_j(x)] dx
\end{aligned}$$

Macierz tego układu jest macierzą symetryczną dodatnio określoną i wstęgową

$$\varphi_{ij} = 0, \quad \text{gdy} \quad |i-j| \geq 4 \tag{D.20}$$

Mamy więc do obliczenia około $8n$ całek. Gdy funkcje σ , f są na tyle skomplikowane, że całek tych nie można dokładnie obliczyć, do obliczania ich możemy zastosować wzory przybliżone. Wystarczy zastosować kwadratury złożone tego samego rzędu dokładności co błąd $\|u - u_N\|$. Można udowodnić (patrz [49]), że jeżeli $f \in C^2(\langle 0; 1 \rangle)$, to istnieje stała K niezależna od h taka, że

$$\|u - u_N\| \leq Kh^3 \tag{D.21}$$

gdzie $\|\cdot\| = \max_{x \in \langle 0; 1 \rangle} |\cdot|$.

Ćwiczenia

1. Udowodnić, że jeżeli U jest przestrzenią liniową oraz $\langle u, v \rangle_i$ ($u, v \in U$, $i = 1, \dots, n$) oznaczają iloczyny skalarne w U , to wzór

$$\langle u, v \rangle = \sum_{i=1}^n \langle u, v \rangle_i$$

określa nowy iloczyn skalarny w U .

2. Weźmy przestrzeń $F = C(\langle 0; 1 \rangle)$ i $U = \{u(x) \in C^2(\langle 0; 1 \rangle) : u(0) = u(1) = 0\}$ z iloczynem skalarnym $\langle u, v \rangle = \int_0^1 u(x) v(x) dx$, określmy operator różniczkowy $A : U \rightarrow F$

w następujący sposób

$$Au(x) = -(\sigma_1(x) u'(x))' + \sigma_2(x) u(x)$$

i załóżmy, że $\sigma_1 \in C^1(\langle 0; 1 \rangle)$, $\sigma_2 \in C(\langle 0; 1 \rangle)$ oraz że $\sigma_1(x) > 0$, $\sigma_2(x) \geq 0$, $\sigma_2(x) \not\equiv 0$ na przedziale $\langle 0; 1 \rangle$.

Czy operator A jest operatorem symetrycznym i dodatnio określonym?

3. Udowodnić, że funkcje $\tilde{\Phi}_i(x)$, $i = 0, 1, \dots, n$, określone wzorem (D.17) są liniowo niezależne na przedziale $\langle 0; 1 \rangle$.

LITERATURA

1. ACHIEZER N.I.: *Teoria aproksymacji*. Warszawa, PWN 1957.
2. AHLBERG J.H., NILSON E.N., WALSH J.L.: *Theory of splines and their applications*. London Academic Press 1967 (przekład rosyjski: *Теория сплайнов и ее приложения*. Москва, Мир 1972).
3. АНО А.В., ХОПСРОФТ J.E., УЛЛМАН J.D.: *Projektowanie i analiza algorytmów komputerowych*. Warszawa, PWN 1983.
4. БАБУШКА И., ВИЧАСЕК Э., ПРАГЕР М.: *Численные процессы решения дифференциальных уравнений*. Москва, Мир 1969.
5. БАХВАЛОВ Н.С.: *Численные методы*. Москва, Наука 1973.
6. BARTH W., MARTIN R.S., WILKINSON J.H.: Calculation of the eigenvectors of a symmetric tridiagonal matrix by the method of bisection. *Num. Math.* 1967, **9**, s. 386–393.
7. BAUER F.L., RUTISHAUSER H., STIEFEL E.: New aspects in numerical quadrature. *Proc. of Symp. in Applied Math.* Vol. 15, Providence, 1963.
8. BAUER F.L., FIKE C.T.: Norms and exclusion theorem. *Num. Math.* 1960, **2**, s. 137–141.
9. BENDAT S., PIERSOL A.G.: *Metody analizy pomiaru sygnałów losowych*. Warszawa, PWN 1977.
10. БЕРЕЗИН И.С., ЖИДКОВ Н.П.: *Методы вычислений*. Т. 1, ч. 2, Москва, Физматгиз 1962.
11. BRONSZTEJN J.N., SIEMIENDIAJEW K.A.: *Matematyka. Poradnik encyklopedyczny*. Wyd. 6, Warszawa, PWN 1976.
12. COOLEY J.W., LEWIS P.A.W., WELCH P.D.: The fast Fourier transform and its applications. *IEE Trans. on Educ.* March 1969, vol. E-12, s. 27–34.
13. COOLEY J.W., TUKEY J.W.: An algorithm for the machine calculation of complex Fourier series. *Math. of Comput.* April 1965, **19**, 90, s. 297–302.
14. DAHLQUIST G.: Stability equations for some numerical methods for ordinary differential equations. *Proc. Symposium for Applied Math.* 1963, vol. 15, s. 147–158.
15. DAHLQUIST G.: A special stability problem for linear multistep methods. *BIT*. 1963, **3**, s. 27–43.
16. DANIEL J.W., GRAGG W.B., KAUFMAN L., STEWART G.W.: Reorthogonalization and stable algorithms for updating the Gram-Schmidt *QR* factorization. *Math. of Comp.* 1976, **30**, 136, s. 772–795.
17. DEMIDOWICZ B.L., MARON I.A.: *Metody numeryczne*. Т. 1, Warszawa PWN 1965.
18. DEMIDOWICZ B.L., MARON I.A., SZUWAŁOWA E.J.: *Metody numeryczne*. Т. 2, Warszawa, PWN 1965.
19. DISTEFANO G.P.: Stability of numerical integration techniques. *AICH J.* 1968, **14**, s. 946.

20. DORN W.S., MC CRACKEN D.D.: *Numerical methods with Fortran IV case studies*. New York, John Wiley 1972.
21. DUBRULLE A., MARTIN R.S., WILKINSON J.H.: The implicit QD -algorithm. *Num. Math.* 1968, **12**, s. 377–383.
22. EVANS D.J.: Iterative sparse matrix algorithms. In.: *Software for numerical mathematics*. London, Academic Press 1974, s. 29–47.
23. FINDEISEN W., SZYMANOWSKI J., WIERZBICKI A.: *Metody obliczeniowe optymalizacji*. Warszawa, PWN 1977.
24. ГАБАСОВ Р., КИРИСЛОВА Р.М.: *Методы оптимизации*. Минск, БГУ 1975.
25. GANTMACHER F.R.: *The theory of matrices*. New York, Chelsea 1960.
26. GEAR C.W.: *The automatic integration of stiff ordinary differential equations*, *Information processing*, 68, ed. A.I.H. Morrel, Amsterdam, North Holland, s. 193–197, 1969.
27. GEAR C.W.: *Numerical initial value problem in ordinary differential equations*. Englewood Cliffs, N.J., Prentice-Hall 1971.
28. GERALD C.F.: *Applied numerical analysis*. New York, Addison-Wesley 1973.
29. GILL P.E., GOLUB G.H., MURRAY W., SAUNDERS M.A.: Methods for modifying matrix factorization. 1974, *Math. of Comp.* vol. 28, 126, s. 505–535.
30. GOLUB G.: Numerical methods for solving linear least squares problems. *Num. Math.* 1965, **7**, 3, s. 206–216.
31. GOULT, HOSKINS, MILNER, PRATT: *Computational methods in linear algebra*. Stanley Thornes-Publ. 1974.
32. GRAGG W.B., STEWART G.W.: A stable variant of the secant method for solving nonlinear equations. *SIAM J. Num. Anal.*, vol. 13, 1976.
33. HAMMING R.W.: Stable predictor-corrector methods for ordinary differential equations. *Journ. Assoc. Comput. Math.* 1965, **6**, s. 37–47.
34. HENRICI P.: *Error propagation for difference methods*. New York, John Wiley 1963.
35. HILDEBRAND F.B.: *Introduction to numerical analysis*. New York, McGraw-Hill 1956.
36. JANKOWSKI M., WOŹNIAKOWSKI H.: O złożoności obliczeniowej w analizie numerycznej. *Mat. Stos.* 1975, **5**, s. 5–27.
37. KATO T.: *Perturbation theory for linear operators*. Berlin, Springer 1966.
38. KIEŁBASIŃSKI A.: Zagadnienie deflacji wielomianowej. *Mat. Stos.* 1974, **3**, s. 93–106.
39. KOWALCZYK B.: *Macierze i ich zastosowania*. Warszawa, WNT 1976.
40. КРЫЛОВ В.И., БОБКОВ В.В., МОНАСТЫРСКИЙ П.И.: *Вычисленные методы*. Т. 1, Москва, Наука 1976.
41. КРЫЛОВ В.И., ШУЛЬГИНА Л.Т.: *Справочная книга по численному интегрированию*. Москва, Наука 1966.
42. LUSTERNIK L.A., CZERWONIENKIS O.A., JANPOLSKI A.R.: *Obliczanie funkcji elementarnych*. Warszawa, PWN 1966.
43. НАТАНСОН И.Р.: *Конструктивная теория функций*. Москва, 1949.
44. *Nowoczesne metody numeryczne*. National Physical Lab. Tedd. Middlesex. Warszawa, PWN 1965.
45. ORTEGA J.M., RHEINBOLDT W.C.: *Iterative solution of nonlinear equations in several variables*. New York, Academic Press 1970.
46. PARLETT B.N.: Global convergence of the basis QR -algorithm on Hessenber matrices. *Math. of Comp.* 1968, **22**, s. 803–817.
47. PETERS G., WILKINSON J.H.: Practical problems arising in the solution of polynomial equations. *J. Inst. Math. Appl.* 1971, **8**.
48. PHILLIPS G.M.: *Theory and applications of numerical analysis*. New York, Academic Press 1973.
49. PRENTER P.M.: *Splines and variational methods*. New York, John Wiley 1975.
50. RALSTON A.: *Wstęp do analizy numerycznej*. Warszawa, PWN 1975.
51. RALSTON A., WILF H.S.: *Mathematical methods for digital computers*. New York, John Wiley 1960.

52. REID J.K.: A note on the stability of Gaussian elimination. *J. Inst. Math. Appl.* 1971, **8**, s. 374.
53. REID J.K. In: *Software for Numerical Mathematics*. Ed. Evans, London, Academic Press s. 29–47.
54. REINSCH C., BAUER F.L.: Rational *QR*-transformation with Newton shift for symmetric tridiagonal matrices. *Num. Math.* 1968, **2**, 246–267.
55. САМАРСКИЙ А.А.: *Введение в теорию разностных схем*. Москва, Наука 1971.
56. SENDOW B.: *Stare i nowe w metodach numerycznych*. Warszawa, PWN 1976.
57. СТЕЧКИН С.Б., СУББОТИН Ю.Н.: *Сплэйны в вычислительной математике*. Москва, Наука 1976.
58. STROUD A.H., SEKREST D.: *Gaussian quadrature formulas*. Englewood Cliffs, N. J., Prentice-Hall 1966, Series in automatic computation.
59. SCHWARTZ B.K., VARGA R.S.: Error bounds for spline and *L*-spline interpolation. *J. Approx. Theory*. 1972, **6**, s. 6–49.
60. TICHONOW A.N., SAMARSKI A.A.: *Równania fizyki matematycznej*. Warszawa, PWN 1963.
61. TRAJDOS T.: *Matematyka*. Cz. 3. Wyd. 5, Warszawa, WNT 1992.
62. TUROWICZ A.: *Geometria zer wielomianów*. Warszawa, PWN 1967.
63. WILKINSON J.H.: Error analysis of direct methods of matrix inversion. *J. Assoc. Comp. Mach.* 1961, **8**, s. 281–330.
64. WILKINSON J.H.: Note on matrices with a very *IU*-Conditioned eigenproblem. *Num. Math.* 1972, **19**, s.176–178.
65. WILKINSON J.H.: *The algebraic eigenvalue problem*. Oxford, Clarendon Press 1965.
66. WILKINSON J.H.: Calculation of the eigenvectors of a symmetric tridiagonal matrix inverse iteration. *Num. Math.* 1962, **4**, s. 368–376.
67. WILKINSON J.H., REINSCH C.: *Linear Algebra-Handbook for Automatic Computation*. Berlin, Springer 1971.
68. WILKINSON J.H.: *Błędy zaokrągleń w procesach algebraicznych*. Warszawa, PWN 1967.
69. WOŹNIAKOWSKI H.: Numerical stability for solving nonlinear equations. *Num. Math.* 1977, **27**, s. 373–390.
70. VARGA R.S.: *Matrix Iterative Analysis*. Englewood Cliffs, N.Y., Prentice-Hall 1965.
71. ZUBER R.: *Metody numeryczne i programowanie*. Warszawa, WSiP 1975.
72. ŻAKOWSKI W., DECEWICZ G.: *Matematyka*. Cz. 1. Wyd. 11, Warszawa, WNT 1991.
73. ŻAKOWSKI W., KOŁODZIEJ W.: *Matematyka*. Cz. 2. Wyd. 9, Warszawa, WNT 1992.
74. ŻAKOWSKI W., LEKSIŃSKI W.: *Matematyka*. Cz. 4. Wyd. 7, Warszawa, WNT 1984.

Literatura pomocnicza w języku polskim

1. DAHLQUIST G., BJORCK A.: *Metody numeryczne*. Warszawa, PWN 1983.
2. DRYJA M., JANKOWSCY J. i M.: *Przegląd metod i algorytmów numerycznych*. Cz. II. Wyd. 2, Warszawa, WNT 1988.
3. JANKOWSCY J. i M.: *Przegląd metod i algorytmów numerycznych*. Cz. I, Wyd. 2, Warszawa, WNT 1988.
4. STOER J.: *Wstęp do metod numerycznych*. T. 1, Warszawa, PWN 1990.
5. STOER J., BULIRSCH R.: *Wstęp do metod numerycznych*. T. 2, Warszawa, PWN 1990.

SKOROWIDZ

- Adamsa-Bashfortha wzór różnicowy 318
- Adamsa-Moultona wzór różnicowy 318
- Aitkena schemat 30
- algorytm Q-D 146
 - szybkiego mnożenia wielomianów 103
- aproksymacja 73
 - średniokwadratowa 76
 - – wielomianowa 79
 - – trygonometryczna 91
 - – funkcjami sklejanymi 104
 - jednostajna 76, 106
- Bairstowa metoda 145
- Banachiewicza rozkład 218
- Bauera-Fikego twierdzenie 258
- bisekcji metoda 285
- błędy obliczeń 13
 - – oszacowanie 16
- Cauchy’ego twierdzenie 140
- ciąg Sturma 134
- Czebyszewa kwadratura 181
 - metoda 240
 - wielomian (pierwszego rodzaju) 36, 181
- Deflacja 149
- Doolitte’a metoda 206
- dyskretna transformata Fouriera 103
- Ekstrapolacja Richardsona 172
- ekstrapolacyjno-interpolacyjna metoda 325
- Eulera metoda 297
- Fibonacciego liczby 156
- Fouriera dyskretna transformata 103
 - szybka transformata 97
 - twierdzenie 136
- Frécheta pochodna 151
- funkcja sklejana 64
- Gaussa-Czebyszewa kwadratura 181
- Gaussa-Hermite’a kwadratura 183
- Gaussa-Jacobiego kwadratura 180
- Gaussa kwadratura 175
 - – złożona 185
- Gaussa-Laguerre’a kwadratura 182
- Gaussa-Legendre’a kwadratura 178
- Gaussa metoda eliminacji 202
- Gaussa-Seidla metoda 236
- Geara metoda 331
- Gershgorina twierdzenie 256
- Givensa macierz 281
 - metoda 281, 293
- Gram-Schmidta ortonormalizacja 278
- Gram wielomian 86
- Hamminga metoda 328
 - wzór różnicowy 317, 327
- Hermite’a wielomian 182
- Hessenberga macierz 270
- Hilberta macierz 211
- Householdera macierz 279
 - metoda 279, 291
- Hurwitza kryterium 142

- Iloczyn energetyczny 371
- iloraz różnicowy 39
- interpolacja 24
 - odwrotna 63
- interpolacyjna funkcja sklejana stopnia trzeciego 65
 - – – Lagrange’a stopnia trzeciego 70
- Jacobiego metoda 235
 - wielomian 180
- Jordana klatka 255
 - metoda eliminacji 216
 - postać kanoniczna 256
- Kartezjusza reguła 137
- klatka Jordana 255
- Kryłowa metoda 267
- kryterium Hurwitza 142
 - Michajłowa 141
 - Routha 142
- kwadratura 162
 - Czebyszewa 181
 - Gaussa 175
 - – złożona 185
 - Gaussa-Czebyszewa 181
 - Gaussa-Hermite’a 183
 - Gaussa-Jacobiego 180
 - Gaussa-Laguerre’a 182
 - Gaussa-Legendre’a 178
 - Newtona-Cotesa 164
 - – złożona 168
- Lagrange’a twierdzenie 138
 - wzór interpolacyjny 26
 - – – oszacowanie błędu 34
- Laguerre’a twierdzenie 136
 - wielomian 182
- Legendre’a wielomian 89, 178
- lemat Wilkinsona 14
- liczby Fibonacciego 156
- Macierz diagonalnie dominująca 214
 - – – kolumnowo 214
 - Givensa 281
 - Hessenberga 270
 - Hilberta 211
 - Householdera 279
 - ortogonalna 255
 - wstęgowa 249
- metoda Bairstowa 145
 - bisekcji 285
 - Czebyszewa 240
 - Doolittle’a 206
 - ekstrapolacyjno-interpolacyjna 325
 - elementu skończonego 373
 - eliminacji Gaussa 202
 - – – z częściowym wyborem elementu podstawowego 208
 - – Jordana (eliminacji zupełnej) 216
 - Gaussa-Seidla 236
 - Geara 331
 - Givensa 281, 293
 - Hamminga 328
 - Householdera 279, 291
 - iteracji odwrotnej 276
 - iteracyjna 151
 - iterowanego dzielenia 139
 - Jacobiego 235
 - Kryłowa 267
 - LR 275
 - ortonormalizacji Grama-Schmidta 278
 - podziału 156
 - – optymalnego 156
 - – złotego 159
 - połowienia 116
 - potęgowa 268
 - prostych 367, 369
 - QR 275, 287
 - reguła falsi 121
 - Romberga 171
 - Rungego-Kutty 318
 - różnicowa (dla równań różniczkowych zwyczajnych) 299
 - – absolutnie stabilna 310, 312
 - – A-stabilna 311, 313
 - – stabilna w sensie Dahlquista 314
 - – stiff-stabilna 332
 - – względnie stabilna 310, 313
 - – (dla równania przewodnictwa cieplnego) 336
 - – stabilna względem prawej strony 341
 - – – warunku początkowego 341
 - – dla równania drgań struny 355
 - – dla równania Poissona 357
 - siecznych 125, 153
 - stycznych (Newtona) 126, 152
 - – zmodyfikowana 131
 - szeregów potęgowych 107
 - zgodna 301
- metody blokowe 250
- Michajłowa kryterium 141
- Milne’a wzór różnicowy 316

- Newtona-Cotesa kwadratura 164
– – – złożona 168
Newtona wzór interpolacyjny 43, 50, 54
norma energetyczna 371
– macierzy 189
– wektora 189
- Odległość zbiorów spektralnych 257
Ostrowskiego twierdzenie 151
- Pade’go przybliżenia 108
pochodna Frécheta 151
potęgowa metoda 268
promień spektralny macierzy 191
przedział izolacji pierwiastka 115
przestrzeń energetyczna operatora 371
- Rayleigha-Ritza metoda 373
reguła Kartezjusza 137
Richardsona ekstrapolacja 172
Romberga metoda 171
Routha kryterium 142
rozkład LU 203
– Banachiewicza 218
Rungego-Kutty metoda 318
Rungego zjawisko 59
różnice progresywne 45
– wsteczne 47
Schemat Aitkena 30
– Hornera 17
Schura twierdzenie 256
splot wektorów 103
stabilność algorytmu 20
Sturma ciąg 134
– twierdzenie 134
szybka transformata Fouriera 97
- Układ normalny 78
układ typu stiff (źle uwarunkowany) 331
uwarunkowanie zadania 21
- Waga (funkcja wagowa) 163
wartość własna macierzy 253
wektor własny macierzy 253
węzły interpolacji 24
– funkcji sklepanej 64
widmo macierzy 254
wielomian Czebyszewa (pierwszego rodzaju)
36, 181
– Grama 86
– Hermite’a 182
– Jacobiego 180
– Laguerre’a 182
– Legendre’a 89, 178
Wilkinsona lemat 14
– twierdzenie 259, 263
wskaźnik uwarunkowania 21
wzór interpolacyjny Lagrange’a 26
– – Newtona dla nierównych odstępów argumentów 43
– – – pierwszy 50
– – – drugi 54
– – oszacowanie błędu 34
– parabol 167
– – złożony 169
– trapezów 166
– – złożony 169
– różnicowy Adamsa-Bashfortha 318
– – Adamsa-Maultona 318
– – Hamminga 317, 327
– – Milne’a 316
- Zasada maksimum 342, 364
zjawisko Rungego 59

WYDAWNICTWA NAUKOWO-TECHNICZNE

ul. Mazowiecka 2/4, 00-048 Warszawa

tel. 826-72-71 do 79

Dział Handlowy

tel. 827-56-87, fax 826-82-93

e-mail: marketing@wnt.com.pl

WNT. Warszawa 2002

Wydanie VI

Ark. wyd. 26,3. Ark. druk. 24,0

Symbol MF/83759/WNT

Cieszyńska Drukarnia Wydawnicza