

Zwięzły kurs analizy numerycznej

Jacek Kabziński
Jarosław Kacerka
Maciej Krawiecki
Krzysztof Marzjan
Przemysław Mosiołek
Rafał Zawiślak

Politechnika Łódzka
Łódź 2018

Zwięzły kurs analizy numerycznej

Jacek Kabziński
Jarosław Kacerka
Maciej Krawiecki
Krzysztof Marzjan
Przemysław Mosiołek
Rafał Zawiślak

Politechnika Łódzka
Łódź 2018

Recenzent: **dr Bożenna Szkopińska**

Redaktor Naukowy Wydziału Elektrotechniki, Elektroniki,
Informatyki i Automatyki
prof. dr hab. inż. Piotr Ostalczyk

Autorzy rozdziałów

- 1 i 2 – Jacek Kabziński i Rafał Zawiślak
- 3 i 9 – Jacek Kabziński i Maciej Krawiecki
- 4 i 5 – Jacek Kabziński i Krzysztof Marzjan
- 6 i 7 – Jacek Kabziński i Przemysław Mosiołek
- 8 – Jacek Kabziński i Jarosław Kacerka

Okładka powstała w wyniku numerycznej analizy nierówności
$$(1 + 2x + 3y) \cos^2 \left(\frac{\pi}{2} (2x - 3(y + 0,55)) \right) < 3.$$

© Copyright by Politechnika Łódzka 2018

WYDAWNICTWO POLITECHNIKI ŁÓDZKIEJ
90-924 Łódź, ul. Wólczańska 223
tel. 42-631-20-87, 42-631-29-52
fax 42-631-25-38
e-mail: zamowienia@info.p.lodz.pl
www.wydawnictwa.p.lodz.pl

ISBN 978-83-7283-877-3

Nakład 300 egz. Ark druk. 19,5. Papier offset. 80 g 70 x 100
Druk ukończono w marcu 2018 r.
Wykonano w drukarni „Quick-Druk” s.c. 90-562 Łódź, ul. Łąkowa 11
Nr 2234

Spis treści

Przedmowa	7
1. Cyfry, liczby i błędy – podstawy analizy numerycznej	11
1.1. Systemy liczbowe.....	11
1.2. Binarna reprezentacja zmiennoprzecinkowa	16
1.3. Arytmetyka zmiennopozycyjna.....	18
1.4. Błędy w obliczeniach numerycznych.....	21
1.5. Błędy skrótów i zaokrągleń	24
1.6. Cyfry poprawne i znaczące	27
1.7. Przenoszenie się błędów w obliczeniach numerycznych	28
1.8. Uwarunkowanie zadania numerycznego	33
1.9. Stabilność numeryczna algorytmu	37
1.10. Złożoność obliczeniowa algorytmu.....	40
2. Rozwiązywanie układów równań liniowych i rozkład trójkątny macierzy kwadratowej.....	45
2.1. Układy równań liniowych	45
2.2. Eliminacja Gaussa	47
2.3. Kontrola poprawności obliczeń w eliminacji Gaussa.....	57
2.4. Złożoność obliczeniowa eliminacji Gaussa.....	62
2.5. Zastosowania rozkładu trójkątnego	66
2.6. Błędy rozwiązania układu równań liniowych metodą eliminacji Gaussa	70
2.7. Inne metody rozwiązywania układów równań liniowych	75
3. Aproksymacja i interpolacja	77
3.1. Modelowanie na podstawie danych cyfrowych.....	77
3.2. Liniowe zadanie aproksymacji średniokwadratowej.....	79
3.3. Wielomiany Czebyszewa	84
3.4. Aproksymacja jednostajna	86
3.5. Interpolacja wielomianowa	90
3.6. Ocena jakości interpolacji – reszta wzoru interpolacyjnego i zjawisko Rungego	98
3.7. Odcinkowa interpolacja wielomianowa	105
3.8. Interpolacja funkcji wielu zmiennych	108
3.9. Obliczanie wartości wielomianu	112
4. Różniczkowanie numeryczne i ekstrapolacja Richardsona	119
4.1. Podstawowe wzory różniczkowania numerycznego	119
4.2. Numeryczne przybliżenie drugiej pochodnej	121
4.3. Dokładniejsze wzory przybliżające pochodną	122
4.4. Różniczkowanie funkcji wielu zmiennych.....	123
4.5. Błędy zaokrągleń w różniczkowaniu numerycznym.....	124
4.6. Iterowana ekstrapolacja Richardsona	126

5. Całkowanie numeryczne.....	133
5.1. Kwadratury proste i złożone	133
5.2. Kwadratury Newtona-Cotesa	134
5.3. Kwadratury Gaussa	136
5.4. Kwadratury złożone	138
5.5. Kwadratury adaptacyjne.....	141
6. Iteracyjne metody rozwiązywania równań nieliniowych.....	143
6.1. Właściwości metod iteracyjnych	143
6.2. Metoda bisekcji.....	147
6.3. Metoda iteracji prostej	148
6.4. Metoda Newtona-Raphsona.....	149
6.5. Metoda siecznych	162
6.6. Metoda <i>regula falsi</i>	164
6.7. Odwrotna interpolacja kwadratowa (Inverse Quadratic Interpolation – IQI)	165
6.8. Złożone metody rozwiązywania równań nieliniowych	167
6.9. Uwarunkowanie pierwiastków równań nieliniowych.....	167
6.10. Układy równań nieliniowych	174
6.11. Metoda Newtona-Raphsona dla układów równań.....	175
6.12. Metoda Broydena	177
6.13. Rozwiązywanie układów równań nieliniowych drogą minimalizacji	179
6.14. Iteracyjne metody rozwiązywania układów równań liniowych	180
7. Pierwiastki wielomianów.....	183
7.1. Operacje na wielomianach	183
7.2. Deflacja	185
7.3. Metoda Newtona-Raphsona i jej warianty	185
7.4. Inne podejścia do wyznaczania pierwiastków wielomianów	191
7.5. Kombinowane algorytmy wyznaczania pierwiastków wielomianu	195
7.6. Uwarunkowanie pierwiastków wielomianów	195
8. Wartości i wektory własne.....	197
8.1. Definicje.....	197
8.2. Uwarunkowanie wartości własnych.....	200
8.3. Wyznaczanie wartości własnych z wielomianu charakterystycznego.....	202
8.4. Metoda potęgowa	204
8.5. Metoda QR wyznaczania wartości własnych	207
8.6. Wartości szczególne macierzy	211
8.7. Zastosowania wartości własnych i szczególnych.....	214
9. Równania różniczkowe zwyczajne	225
9.1. Zagadnienie początkowe	225
9.2. Numeryczne rozwiązanie zagadnienia początkowego.....	229
9.3. Liniowe równania różniczkowe.....	234
9.4. Schematy różnicowe jednokrokowe niskiego rzędu i ich najważniejsze cechy	236
9.5. Metody Rungego-Kutty.....	247

9.6. Sterowanie długością kroku w metodach jednokrokowych	250
9.7. Metody wielokrokowe	253
9.8. Metody Adamsa.....	259
9.9. Metody wstecznego różniczkowania	264
9.10. Jak dopasować metodę numerycznego rozwiązania zagadnienia początkowego do specyfiki zadania?	268
Dodatki	
D1. Liczby i wektory	275
D2. Podstawy rachunku macierzowego	285
D3. Elementy analizy matematycznej.....	291
Bibliografia	295
Indeks terminów	297
Indeks nazwisk.....	303

Przedmowa

W polskiej terminologii używa się nazw **analiza numeryczna** i **metody numeryczne** wymiennie. Przedmiot analizy numerycznej można zdefiniować dwojako:

- po pierwsze, to nauka zajmująca się rozwiązywaniem problemów matematycznych metodami arytmetycznymi,
- po drugie, to sztuka doboru spośród wielu możliwych procedur takiej, która jest „najlepiej” dostosowana do rozwiązania konkretnego zadania.

Wiele problemów matematycznych i inżynierskich jest dokładnie rozwiązywanym metodami analizy matematycznej, z użyciem abstrakcyjnych pojęć i narzędzi. Metoda numeryczna prowadzi do uzyskania przybliżonego rozwiązania takich zadań sposobami, które są dostępne maszynie cyfrowej, czyli przede wszystkim przez wykonywanie operacji arytmetycznych. Nie wystarczy procedurę numeryczną zaprojektować, trzeba jeszcze zbadać jej właściwości. W takim rozumieniu (projektowanie metod numerycznych i badanie ich właściwości) analizę numeryczną można uważać za część matematyki, a do jej wyników dochodzi się, stosując aparat pojęciowy i narzędzia matematyki.

Z reguły, dla jednego problemu dysponujemy kilkoma metodami numerycznymi. Trzeba wybrać jedną z nich i zaimplementować. To znaczy, biorąc pod uwagę zasoby sprzętowe i oprogramowanie, które są do dyspozycji, utworzyć algorytm, który zrealizuje obliczenia i wygeneruje wynik, który świadomie zaakceptujemy. Przy wyborze i implementacji metody numerycznej bierze się pod uwagę liczne czynniki i przesłanki, wśród których jest specyfika urządzenia cyfrowego realizującego obliczenia, sposób reprezentacji danych wejściowych i wyników, czas obliczeń, zasoby pamięci i wiele innych, wraz z najważniejszą: świadomością celu wykonywanych obliczeń. Wielu uważa, że ta sztuka stosowania metod numerycznych, łącząca znajomość ich matematycznie udowodnionych właściwości, zrozumienie rozwiązywanego problemu, wiedzę o sprzęcie, algorytmikę i programowanie, to nowa dziedzina określana terminem *scientific computing* albo *computational science*.

Analiza numeryczna jest niezbędna we wszystkich naukach, w których wykorzystujemy cyfrowe urządzenia liczące. W każdej sytuacji, w której stosujemy maszynę cyfrową do wykonania obliczeń, ciąży na nas obowiązek analizy dokładności i przydatności otrzymanego wyniku.

Materiał przedstawiony w tej książce to wybrane przez autorów metody numeryczne prezentowane studentom różnych kierunków studiów inżynierskich w trakcie wykładów, ćwiczeń i laboratoriów. Mamy nadzieję, że przygotowane w zwartej postaci podstawowe informacje z analizy numerycznej ułatwią studentom usystematyzowanie i przyswojenie wiedzy. Uczestnictwo w wykładzie,

ćwiczeniach i laboratoriach umiejscowi tę wiedzę w kontekście zastosowań oraz pozwoli na rozwinięcie praktycznych umiejętności.

Chcielibyśmy podkreślić, że opanowanie i stosowanie analizy numerycznej obejmuje nie tylko poznanie i zrozumienie arsenału metod i sposobów na uzyskanie przybliżonych rozwiązań problemów matematycznych i inżynierskich. Chcielibyśmy zwrócić uwagę studentów i czytelników tej książki na kilka podstawowych zasad i kluczowych pojęć, które stanowią o istocie analizy numerycznej:

1. Wszelkie metody numeryczne obarczone są błędami różnego rodzaju, pochodzenia i znaczenia. Bez analizy tych błędów, zrozumienia ich źródeł i wagi w otrzymanym wyniku metoda numeryczna jest bezużyteczna.
2. Zbieżność wyników metody numerycznej do dokładnego rezultatu jest jej kluczową cechą. Badanie tej zbieżności jako funkcji parametrów metody pozwala na świadome stosowanie metody numerycznej w sposób dopasowany do naszych celów.
3. Złożoność numeryczna algorytmów realizujących metodę numeryczną jest miarą zasobów obliczeniowych koniecznych do ich wykonania i ma kluczowe znaczenie dla oceny możliwości zastosowania tej metody w konkretnej sytuacji.
4. Uwarunkowanie problemów numerycznych, czyli miara wrażliwości wyniku na zmiany lub błędy danych wejściowych, decyduje o tym, które zadania można rozwiązać, a których należy unikać.
5. Metody numeryczne powinny być stosowane tak, by uzyskać rozwiązanie spełniające narzucone wymagania przy jak najmniejszej liczbie wykonanych operacji arytmetycznych.
6. Obliczeń numerycznych nie wolno wykonywać bezmyślnie, a ich wyników akceptować bezrefleksyjnie.

Będziemy się odnosić do tych zasad w kolejnych rozdziałach.

Poza przydatnością poszczególnych metod numerycznych do rozwiązania powszechnie spotykanych problemów inżynierskich, nauka analizy numerycznej jest doskonałą szkołą myślenia algorytmicznego. Zapraszamy do lektury i samodzielnych, aktywnych studiów.

Książka składa się z dziewięciu rozdziałów i trzech dodatków. Dodatki zawierają podstawy algebry i analizy matematycznej, które mieszczą się w materiale wykładanym zwykle na pierwszych semestrach studiów inżynierskich. Zostały zebrane w jednym miejscu wiadomości, do których bezpośrednio odwołano się w tekście.

Każdy z dziewięciu rozdziałów stanowi odrębną całość i może być czytany oddzielnie. Twierdzenia, definicje i przykłady zaznaczono w tekście. Definicje i twierdzenia porządkują podstawowe fakty, które są wykorzystane w konstrukcji metod numerycznych. Te twierdzenia, które umieszczono bez dowodu zaopa-

trzone w odsyłacze do podającej je literatury. Wyprowadzenia metod numerycznych, analiza ich właściwości i dyskusja przydatności stanowią główny nurt książki i nie zamykano ich w formie twierdzeń. Przykłady ilustrujące podstawowe pojęcia i właściwości są proste, często tak by umożliwić ich natychmiastowe przeliczenie przez czytelnika. Podano też przykłady ilustrujące wyjątki i sytuacje awaryjne metod numerycznych. Algorytmy dodatkowe i pomocnicze umieszczono w ramkach wydzielających je z tekstu.

Zamieszczona bibliografia zawiera przede wszystkim tytuły w języku polskim, które mogą być wykorzystane do pogłębienia wiadomości. Podano też te pozycje anglojęzyczne, z których zaczerpnięto twierdzenia niedostępne w polskiej literaturze. Nie było celem autorów kompletowanie pełnej bibliografii przedmiotu, bo ta byłaby bardzo obszerna.

Autorzy

1. Cyfry, liczby i błędy – podstawy analizy numerycznej

1.1. Systemy liczbowe

Arytmetyka (łac. arithmetica, gr. ἀριθμητική od ἀριθμός – liczba) jest jednym z najstarszych działów matematyki zajmującym się liczbami, sposobem ich zapisu i operacjami na liczbach.

Działania arytmetyczne to dodawanie, odejmowanie, mnożenie i dzielenie. Jednostka arytmetyczno-logiczna komputera wykonuje właśnie działania arytmetyczne (na liczbach, które są reprezentowane w maszynie cyfrowej, a ponadto realizuje operacje logiczne, przesunięcia bitowe, czasem dzielenie modulo).

System liczbowy to zbiór reguł zapisu i nazewnictwa liczb. Współcześnie używa się **systemów pozycyjnych**, które posługują się skończonym zbiorem znaków, zwanych **cyframi**. W zależności od swojej pozycji w ciągu reprezentującym liczbę, cyfra oznacza wielokrotność odpowiedniej potęgi liczby p nazywanej **podstawą systemu**. Potęgi podstawy o wykładniku $0, 1, 2, \dots$ odpowiadają kolejnym pozycjom na lewo, a potęgi o wykładniku $-1, -2, -3, \dots$ – pozycjom na prawo od znaku rozdzielającego (przecinka lub kropki) część całkowitą od części ułamkowej. Oddzielnie trzeba zapisać znak liczby. Cyfry w systemie pozycyjnym o podstawie p to kolejno $0, 1, 2, \dots, p - 1$.

Przykład 1.1

Liczbę 103,45 zapisaną w dziesiętnym systemie liczbowym (systemie którego podstawą p jest liczba 10) odczytujemy jako:

$$103,45 = 1 \cdot 10^2 + 0 \cdot 10^1 + 3 \cdot 10^0 + 4 \cdot 10^{-1} + 5 \cdot 10^{-2}$$

czyli: sto trzy i czterdzieści pięć setnych.

Liczbę L w systemie pozycyjnym o podstawie p można przedstawić jako:

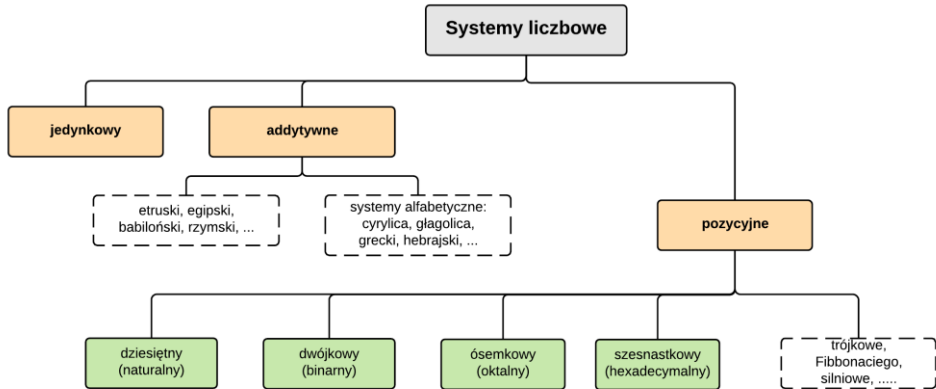
$$L = \sigma \sum_{i \in \{c_1, c_1 - 1, \dots, 1, 0, -1, \dots, -c_0\}} d_i p^i \quad (1.1)$$

gdzie: $\sigma \in \{-1, +1\}$ – jest **znakiem liczby** L , p jest liczbą naturalną oznaczającą **podstawę systemu pozycyjnego** ($p > 1$), d_i jest liczbą całkowitą nieujemną, mniejszą od p , czyli **cyfrą** liczby L , $c_1 + 1$ jest liczbą cyfr na lewo, a c_0 liczbą cyfr na prawo od znaku rozdzielającego **cyfry całkowite** od **cyfr ułamkowych**.

Najbardziej popularne systemy pozycyjne zestawiono w tabeli 1.1, a miejsce systemów pozycyjnych w strukturze wszystkich systemów zapisu liczb znanych w historii matematyki pokazano na rys. 1.1.

Tabela 1.1. Podstawowe systemy liczbowe

System	Podstawa	Cyfry
dziesiętny	$p = 10$	$D = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$
binarny	$p = 2$	$D = \{0, 1\}$
ósemkowy	$p = 8$	$D = \{0, 1, 2, 3, 4, 5, 6, 7\}$
hexadecymalny	$p = 16$	$D = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F\}$



Rys. 1.1. Systemy liczbowe

Z racji reprezentacji liczb w pamięci komputerów za pomocą bitów, najbardziej naturalnym systemem w informatyce jest dwójkowy system liczbowy (system, którego podstawą jest $p = 2$). W pionierskich czasach komputeryzacji ważną rolę odgrywał również system ósemkowy. Ze względu na specyfikę architektury komputerów, gdzie często najszybszy jest dostęp do adresów parzystych, albo podzielnych przez 4, 8 czy 16, często używany jest również szesnastkowy system liczbowy. Sprawdza się on szczególnie przy zapisie dużych liczb, takich jak adresy pamięci. System szesnastkowy przyjął się również w językach programowania stron WWW (HTML, CSS), gdzie służy np. do określania kolorów poszczególnych elementów, kodowania znaków specjalnych. Natomiast naturalny dla ludzi system dziesiętny został wprowadzony do informatyki dopiero wraz z powstaniem języków programowania wyższego poziomu, których celem było jak największe ułatwienie tworzenia algorytmów. Liczba przedstawiona w systemie dziesiętnym jest jednak za każdym razem przeliczana do jej binarnej reprezentacji w pamięci maszyny.

Zapis liczby w systemie pozycyjnym wymaga zawsze podania podstawy systemu. Jeżeli podstawa nie jest podana, to przyjmujemy, że liczba jest zapisana w systemie dziesiętnym. Tak więc zapisujemy

$$12,5 = 1100,1_2 = 14,4_8 = C,8_{16} .$$

Przedstawienie dowolnej liczby rzeczywistej w postaci (1.1) wymagałoby możliwości użycia nieskończenie wielu cyfr. Jeśli w przedstawieniu (1.1) zostaną ustalone wartości c_0 i c_1 (liczba cyfr przed i po przecinku), to mówimy o **reprezentacji stałoprzecinkowej (stałopozycyjnej)**.

Jeśli natomiast w przedstawieniu liczby L ustalimy łączną liczbę użytych cyfr t , to stosując postać

$$L = \sigma p^c \sum_{i=1}^t d_i p^{-i}. \quad (1.2)$$

uzyskamy tak zwaną **reprezentację zmiennoprzecinkową (zmiennopozycyjną)**.

Przedstawienie to jest niejednoznaczne dopóki nie założymy, że $d_1 > 0$. Postać liczby (1.2), w której pierwsza cyfra ułamkowa $d_1 > 0$ nazywa się **znormalizowaną**.

Sumę $m = \sum_{i=1}^t d_i p^{-i}$ nazywamy **mantysą** liczby L , liczbę c jej **cechą**, zaś $\sigma \in \{-1, +1\}$ – znakiem. Oczywiście w praktycznych realizacjach zarówno zakres cechy jak i mantysy liczby musi być ograniczony.

Definicja 1.1 (zbioru liczb zmiennoprzecinkowych)

Dla danej

- podstawy $p \geq 2$ i cyfr $D \in \{1, 2, \dots, p - 1\}$,
- długości mantysy $t \in \mathbb{N}$,
- ograniczeń cechy $c_{min} < 0 < c_{max}$,

definiujemy zbiór liczb

$$\begin{aligned} \Phi &= \Phi(p, t, c_{min}, c_{max}) = \\ &= \left\{ \begin{array}{l} \sigma p^c \sum_{i=1}^t d_i p^{-i} \\ d_i \in D, d_1 > 0, c_{min} \leq c \leq c_{max}, c \in \mathbb{Z}, \sigma \in \{-1, +1\} \end{array} \right\} \quad (1.3) \\ &\cup \{0\} \end{aligned}$$

gdzie \mathbb{Z} jest zbiorem liczb całkowitych, nazywany **znormalizowanym zbiorem liczb zmiennoprzecinkowych (lub zmiennopozycyjnych)**. Zbiór $\hat{\Phi} = \hat{\Phi}(p, t, c_{min}, c_{max})$ opisany jak wyżej, ale z dopuszczeniem $d_1 = 0$, gdy $c = c_{min}$, nazywany jest **nieznormalizowanym zbiorem liczb zmiennoprzecinkowych**.

Jako że stosowanie konkretnego zbioru liczb zmiennoprzecinkowych wiąże się zwykle z typem maszyny cyfrowej realizującej obliczenia, liczby te są nazywane **maszynowymi** lub **mającymi reprezentację maszynową**.

Normalizacja zbioru liczb zmiennoprzecinkowych gwarantuje, że reprezentacja liczby L w postaci (1.2) jest jednoznaczna.

Analiza postaci liczby w znormalizowanym zbiorze liczb zmiennoprzecinkowych, pozwala opisać ich podstawowe właściwości:

1. Dla dowolnej mantysy m zachodzi $m = \sum_{i=1}^t d_i p^{-i} \geq p^{-1}$ (bo $d_1 > 0$), więc dla dowolnej dodatniej liczby zmiennoprzecinkowej $p^c \sum_{i=1}^t d_i p^{-i} \geq p^{c \min^{-1}}$. Dla dowolnej cyfry mamy $d_i \leq p - 1$, więc dla dowolnej dodatniej liczby zmiennopozycyjnej $p^c \sum_{i=1}^t d_i p^{-i} \leq p^{c \max} \sum_{i=1}^t (p - 1) p^{-i} = p^{c \max} (1 - p^{-t})$. Tak więc najmniejszym i największym dodatnim elementem w zbiorze Φ dodatnich, znormalizowanych liczb zmiennopozycyjnych są liczby

$$x_{\min} = p^{c \min^{-1}} \text{ i } x_{\max} = p^{c \max} (1 - p^{-t}). \quad (1.4)$$

Zbiór liczb zmiennoprzecinkowych możliwych do przedstawienia w maszynie cyfrowej o określonej **precyzji** (długości mantysy) i długości cechy jest jedynie podzbiorem liczb rzeczywistych:

$$\Phi \subset [-x_{\max}, -x_{\min}] \cup \{0\} \cup [x_{\min}, x_{\max}]. \quad (1.5)$$

2. Rozważmy wszystkie dodatnie, znormalizowane liczby zmiennopozycyjne, które mają tę samą cechę c : $L = p^c \sum_{i=1}^t d_i p^{-i}$. Najmniejszą z nich jest liczba $p^c p^{-1} = p^{c-1}$ (bo $d_1 > 0$), kolejną $p^{c-1} + p^c p^{-t} = p^{c-1} + p^{c-t}$, następną $p^{c-1} + p^c 2p^{-t} = p^{c-1} + 2p^{c-t}$ itd. Największą z tych liczb będzie taka, której wszystkie cyfry mają najwyższe wartości, czyli $p^c \sum_{i=1}^t (p - 1) p^{-i} = p^c (1 - p^{-t}) = p^{c-1} + (p^t - p^{t-1} - 1) p^{c-t}$.

Oznacza to, że w każdym przedziale postaci $[p^{c-1}, p^c]$ mamy jednakową ilość $M = p^t - p^{t-1}$ równoodległe rozmieszczonych liczb zmiennoprzecinkowych, odległych od siebie o stałą

$$\Delta_c = p^{c-t}, \quad (1.6)$$

a zatem:

$$\Phi \cap [p^{c-1}, p^c] = \{(p^{-1} + j p^{-t}) p^c, j = 0, 1, \dots, M - 1\}. \quad (1.7)$$

Wniosek: Liczby zmiennoprzecinkowe nie są równoodległe – im większa cecha, tym większa odległość między sąsiednimi liczbami.

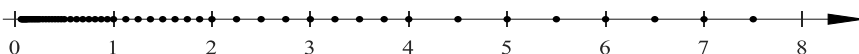
Przykład 1.2

Rozważmy zbiór liczb binarnych, o 4-cyfrowej mantysie i cesze z przedziału $[-3,3]$: $\Phi(2, 4, -3,3)$. W tym przypadku $M = 2^4 - 2^3 = 8$. Przeliczając 56 dodatnich liczb binarnych reprezentowanych w tym zbiorze na ich odpowiedniki dziesiętne uzyskamy liczby podane w tabeli 1.2.

Tabela 1.2. Reprezentacja dziesiętna zbioru $\Phi(2, 4, -3,3)$

Mantysa	Cecha						
	$c = -3$	$c = -2$	$c = -1$	$c = 0$	$c = 1$	$c = 2$	$c = 3$
$0,1000_2$	0,0625	0,125	0,25	0,5	1	2	4
$0,1001_2$	0,0703125	0,140625	0,28125	0,5625	1,125	2,25	4,5
$0,1010_2$	0,078125	0,15625	0,3125	0,625	1,25	2,5	5
$0,1011_2$	0,0859375	0,171875	0,34375	0,6875	1,375	2,75	5,5
$0,1100_2$	0,09375	0,1875	0,375	0,75	1,5	3	6
$0,1101_2$	0,1015625	0,203125	0,40625	0,8125	1,625	3,25	6,5
$0,1110_2$	0,109375	0,21875	0,4375	0,875	1,75	3,5	7
$0,1111_2$	0,1171875	0,234375	0,46875	0,9375	1,875	3,75	7,5
	$\Delta_{-3} = 2^{-7}$	$\Delta_{-2} = 2^{-6}$	$\Delta_{-1} = 2^{-5}$	$\Delta_0 = 2^{-4}$	$\Delta_1 = 2^{-3}$	$\Delta_2 = 2^{-2}$	$\Delta_3 = 2^{-1}$

Na rys. 1.2 pokazano ten zbiór na osi liczbowej.



Rys. 1.2. Dodatnie liczby ze zbioru $\Phi(2, 4, -3,3)$

Jak widać liczby zmiennoprzecinkowe nie są równomiernie rozmieszczone na osi liczbowej. Najmniejszą dodatnią liczbą zmiennoprzecinkową jest $x_{min} = 0,0625$. Możemy stwierdzić, że względnie duży jest przedział $(-x_{min}, x_{min})$, w którym jedyną liczbą zmiennoprzecinkową jest 0.

Zastosowanie nieznormalizowanego zbioru liczb $\hat{\Phi}(2, 4, -3,3)$ pozwala zwiększyć ilość dostępnych wartości o kolejne 7 pokazanych w tabeli 1.3 i zmniejszyć długość przedziału $(-x_{min}, x_{min})$.

Tabela 1.3. Reprezentacja dziesiętna liczb ze zbioru $\widehat{\Phi}(2, 4, -3, 3)$, które nie występowały w zbiorze $\Phi(2, 4, -3, 3)$

Mantysa	Cecha $c = -3$
$0,0001_2$	0,0078125
$0,0010_2$	0,015625
$0,0011_2$	0,0234375
$0,0100_2$	0,03125
$0,0101_2$	0,0390625
$0,0110_2$	0,046875
$0,0111_2$	0,0546875
	$\Delta_{-3} = 2^{-4} \cdot 2^{-3}$

Parametrem charakteryzującym system liczb zmiennoprzecinkowych jest tak zwany **epsilon maszynowy** – *eps*. Jest to odległość między liczbą 1 a następną liczbą zmiennopozycyjną. Liczba 1 jest zapisywana z cechą równą 1, więc zgodnie z (1.6)

$$1 + eps = p^1(p^{-1}) + p^1p^{-t}, \quad (1.8)$$

czyli

$$eps = p^{1-t}. \quad (1.9)$$

Zgodnie z (1.6), odległość dowolnej liczby rzeczywistej $x \in [p^{c-1}, p^c)$ od najbliższej liczby zmiennoprzecinkowej nie przekracza $\frac{1}{2}p^{c-t}$, odległość względna nie przekracza więc

$$\frac{\frac{1}{2}p^{c-t}}{|x|} \leq \frac{\frac{1}{2}p^{c-t}}{p^{c-1}} = \frac{1}{2}p^{1-t} = \frac{1}{2}eps. \quad (1.10)$$

Tak więc, dokładność względna reprezentacji zmiennoprzecinkowej zależy tylko od długości mantysy. Błąd zaokrąglenia liczby x do najbliższej liczby zmiennoprzecinkowej nie przekracza

$$|\Delta_x| \leq \frac{eps}{2} |x|. \quad (1.11)$$

1.2. Binarna reprezentacja zmiennoprzecinkowa

W celu minimalizacji problemów związanych z kodowaniem liczb oraz by ujednotoczyć sposób wykonywania obliczeń na maszynach cyfrowych, został wprowadzony

dzony standard IEEE754¹ dotyczący arytmetyki binarnej oraz standard IEEE854 dla maszyn cyfrowych o podstawie 10.

Tabela 1.4 ilustruje zestawienie najczęściej używanych systemów liczb zmiennoprzecinkowych w formacie binarnym zgodnym z normą IEEE754 stosowanych w różnych rozwiązaniach technicznych.

Tabela 1.4. Popularne systemy liczb zmiennoprzecinkowych w formacie binarnym

Nazwa	Długość	Bity mantysy	Bity cechy	c_{min}	c_{max}
single (pojedynczej precyzji)	32	24	8	-126	127
double (podwójnej precyzji)	64	53	11	-1022	1023
quadruple (poczwórnej precyzji)	128	113	15	-16382	16383

Po przeliczeniu poszczególnych wartości granicznych dla powyższych systemów do systemu dziesiętnego i zapisaniu ich w postaci $x = 10^c \cdot (0, d_1 d_2 \dots d_r)$, gdzie d_i oznacza i -tą cyfrę, uzyskamy parametry liczb podane w tabeli 1.5.

Tabela 1.5. Zakres i precyzja liczb zmiennoprzecinkowych

Nazwa	Ilość cyfr rozwinięcia dziesiętnego r	Najmniejsza dodatnia liczba x_{min}	Największa dodatnia liczba x_{max}	Epsilon maszynowy eps
single (pojedynczej precyzji)	7	$1,18 \cdot 10^{-38}$	$3,4 \cdot 10^{38}$	$2^{-23} = 1,19 \cdot 10^{-7}$
double (podwójnej precyzji)	16	$2,2 \cdot 10^{-308}$	$1,8 \cdot 10^{308}$	$2^{-52} = 2,22 \cdot 10^{-16}$
quadruple (poczwórnej precyzji)	34	$1,32 \cdot 10^{-4932}$	$1,92 \cdot 10^{4932}$	$2^{-112} = 1,93 \cdot 10^{-34}$

Przyjrzyjmy się dokładniej najczęściej stosowanemu formatowi, czyli formatowi podwójnej precyzji, reprezentującemu zbiór $\Phi(2, 53, -1022, 1023)$ – liczb binarnych o 53-bitowej mantysie i cesze z przedziału $[-1022, 1023]$. Maszynową reprezentację takiej liczby pokazano na rys. 1.3.

¹ IEEE = The Institute of Electrical and Electronics Engineers, organizacja ta opracowała szereg norm-i zaleceń dotyczących m.in. techniki komputerowej.

rt|gl"y {f gplg"t|glut»y "y"ct{wo qo gw|g"q"5"dk"y "rqt»y pcpkw| "f ai q ek " o cpv u{ "rt| { "rt| gejqy {y cpkw|cp{ej "y "r co k ek0D€f "qdrle|cpk|lpp{ej "hwpncl" grgo gpwtp{ej "hwpncl"r qv i qy c."y {mef ple|c."m|ct{wo {"hwpnclg"vt {i qpgo g/ vt {el pg-|plg|lgvltgi wqy cp{ 'pqto . "crg"fr"rqto cw»y "šukpi ngö'kšf qwdngö'plg'r|t|g/ mtce|c"qp|c|y {el cl'lgf pgl'lgf pqumk'qucvplg|'r qf cpgl'e {ht {"*r q'cpi kgrumw'3"wr "ó" wpk|lp"rurvr meg+0Rqpcf vq|'pqto c"KGG976"y {o ci c."cd {"t gcrk| qy cpq" d { € "v|y 0' uqr plqy g'plgf qr g'plgplg" *cpi 0i tef wcn|wpf gtl qy +0Q| pce|c"vq." g'y {pkn|qr gtccl' |d {v'o c€."d {"rt|gf ucy k "i q"y "r qucek"rle|d {"|pqto crk|qy cpqgl."plg"o q g"d {" vcmqy cp {" lcnq| |gtq." fqr »nk' r q|quclg" y " tglgutcej " ct {wo qo gw w0' P qto c" KGG976"fgłpkw|g'e|vgt {"t {d {"|cqnt i rnpk."crg"fqo { np{o "lgu'vcln| y cpq" |cq/ mt i rnpk|dcpn|g|tun|g" *cpi 0'dcpn|gtu"tqwpf lpi ."tqwpf lpi "vq"pgctguv'gxgp+0P qto c" KGG976"plg|lgv'f qun|pc€."y "u|e|gi »rpq ek'plg"i y ctpwlg'pcf cn" g"o cu| {p {" |i qf pg"|"pk "d f "fr" v'ej "uco {ej "fcp{ej "qdrle|c "vg"uco g"y {pknk "lgf pcn| el ekqy q"plgf qi qf pq "c"y {pkn"|"d|cnw'f qucvgel pgi q"y ur ctekc"ct {wo gv nk| |o kppqr t| gelpny gl| g'utqp {"1 | |m»y "rtqi tco qy cpk0'

Rtqy cf| e"qdrle| gpk"y "ct{wo gv eg" |o kppqr t| gelpny gl|"pcrg {"|fey c "uqdlg" ur tcy ." g'r qo ko q"dtf|q"ukp{ej "qdqut|g "pcmef cp{ej "rt|gl "pqto {"KGG976" e| {"KGG: 76."plg"u "ur g'plqpg'r qf ucy qy g'cmulqo cv {"ct {wo gv nk|rle|d't|ge| {y k' u'v'ej ."p'e'r t| {mef "f qf cy cpk|plg|lgv'f| k€plgo "€e|p{o ."c'y 'r gy p{ej "untclp{ej " y {r cf ncej "o q g'd| "pctw| qpc'pcy gvtq|f |kgrpq "o pq gpk'y| i n f go "f qf cy c/ plc" *pkm»tg" uctu|g" v|r {"o cu| {p" e {htqy {ej ."rt|gf" y rtqy cf| gplgo "pqto {" KGG976"pctw| c€"pcy gv'rtcy c'r t| go kppq ek'o pq gpk"e| {"f qf cy cpk+0'

Y {pkn|qdrle|g " *y tc| " | |cqnt i rnpkgo + "y "ct {wo gv eg" |o kppqr t| gelpny gl" q| pce|c"uk "u{o dqrgo "fl(*)0'

Vc"uco c"rle|dc"y "u{uogo kg" f| kguk v{p{o "k'dkpctp{o "o c't» p "rle|d "e {ht" w€o nq/ y {ej 0'Y u| {un|g" w€o nk'f| kguk vpg."m»t {ej "plg" f c'uk " |crkuc "y "r qucek" $\frac{x}{2k}$ " d f " o k€"r q'r t| grle| gpkw'pc"u{uogo "dkpctp{ "plgunq| el qpg'tq| y lpk ekg" w€o nqy g0'

Przykład 1.3

W€o $gn^{\frac{1}{10}}$ | cr kucp {"y "u{uogo kg" f| kguk v{p{o "o c'unq| el qp "tgrt|g| gpvccl "r qucek" 2.30F qnqpw| e'lg|q'r t| grle| gpk'pc"u{uogo "dkpctp{ "f qucvplgo {"w€o gn|qnt guqy {" 0,1 = 0,0(0011)₂."eq"r qy qf wlg."y " |crg pq ek'qf "rt| |1 vgi q"y "fcp{o "u{uogo kg" ur quqdwh|qf qy cpk."nup|ge| pq "lgi q'unt»egpk" *q| cqnt i rnpk+fq' rle|d {"e {ht"y {/ pkncl egl"|"w {y cpq|f " ai q ek'o cpv u{0'Y {nupw| e"qdrle| gpk"|" |cuquqy cpkgo " rle|d| o kppqr t| gelpny {ej "r qlgf |pel gl'r tge{| |lk'r q'r qy »tp{o "r t| grle| gpkw'pc" u{uogo "f| kguk v{p {"qnc| wlg'uk ." g'y "r co k ek'o co {"y ctvq "2.322222236; 2338i " |co kcu'v| ctvq ek'2.3" v'v' nq'9'r qr tcy p{ej "e {ht" f| kguk v{p {ej "ó'r cwt| "vcd030' +0"

Przykład 1.4

Korzystając z tabeli 1.2, obliczymy wartość wyrażenia $\left(\frac{1}{10} + \frac{1}{5}\right) + \frac{1}{6} = \frac{7}{15}$ stosując maszynę z arytmetyką binarnych liczb zmiennoprzecinkowych o 4-cyfrowej mantysie, dokonującą zaokrąglenia z pięciobitowej reprezentacji w arytmometrze. Poszukując najlepszej reprezentacji binarnej, otrzymujemy $\frac{1}{10} \approx 0,1101_2 \cdot 2^{-3}$, $\frac{1}{5} \approx 0,1101_2 \cdot 2^{-2}$.

Po dodaniu w arytmometrze, otrzymujemy

$0,1101_2 \cdot 2^{-3} + 0,1101_2 \cdot 2^{-2} = 0,01101_2 \cdot 2^{-2} + 0,1101_2 \cdot 2^{-2} = 1,00111_2 \cdot 2^{-2}$, co ze względu na 4-cyfrową mantysę zostanie zaokrąglone do wartości $0,1010_2 \cdot 2^{-1} \approx \frac{3}{10}$.

Po dodaniu liczby $\frac{1}{6} \approx 0,1011_2 \cdot 2^{-2}$, otrzymujemy

$0,1010_2 \cdot 2^{-1} + 0,1011_2 \cdot 2^{-2} = 0,1010_2 \cdot 2^{-1} + 0,01011_2 \cdot 2^{-1} = 0,11111_2 \cdot 2^{-1}$, co zostanie zaokrąglone do $0,11111_2 \cdot 2^{-1} \approx 0,1000_2 \cdot 2^0$.

Błąd jaki popełniliśmy stosując do obliczeń maszynę binarną o 4 cyfrowej mantysie jest więc równy

$$\frac{7}{15} - 0,1000_2 = 0,4(6) - 0,5 = -0,0(3).$$

Stanowi to 7,14% liczby $\frac{7}{15}$.

Przykład 1.5 (naruszenie łączności dodawania)

Rozważmy maszynę cyfrową dziesiętną, w której długość mantysy wynosi 5, a wyniki są poprawnie zaokrąglone.

- Rozważmy sumę $4,6743 + 0,000032167 + 0,000049132$. Łatwo można sprawdzić, że:

$$\begin{aligned} &(4,6743 + 0,000032167) + 0,000049132 = \\ &fl(fl(4,6743 + 0,000032167) + 0,000049132) = \\ &= fl(4,6743 + 0,000049132) = 4,6743 \end{aligned}$$

zaś:

$$\begin{aligned} &4,6743 + (0,000032167 + 0,000049132) = \\ &fl(4,6743 + fl(0,000032167 + 0,000049132)) = \\ &= fl(4,6743 + 0,000081299) = 4,6744. \end{aligned}$$

- Suma algebraiczna $3,2417 + 0,0004567 + 0,00004876 - 3,2417$ obliczana w naturalnym porządku składników będzie równa

$$\begin{aligned}
 & fl(fl(fl(3,2417 + 0,0004567) + 0,00004876) - 3,2417) = \\
 & = fl(fl(fl(3,24174567) + 0,00004876) - 3,2417) = \\
 & = fl(fl(3,2417 + 0,00004876) - 3,2417) = \\
 & = fl(fl(3,24174876) - 3,2417) = \\
 & = fl(3,2417 - 3,2417) = 0,
 \end{aligned}$$

pomimo, że na pierwszy rzut oka widać, że dokładna wartość jest dodatnia.

Powyższe przykłady mogą wydawać się nieco sztuczne, jednak podobne sytuacje mogą występować w praktyce, na przykład: przy obliczaniu iloczynów skalarnych wektorów o znacznej liczbie składników błąd względny wyniku może okazać się znaczny.

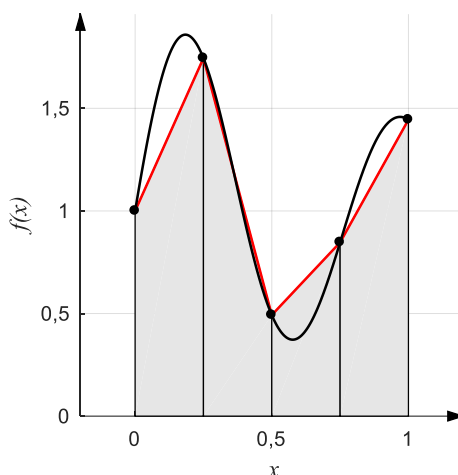
1.4. Błędy w obliczeniach numerycznych

Analiza numeryczna proponuje metody rozwiązania problemu matematycznego (lub zapisanego językiem matematyki problemu inżynierskiego, ekonomicznego, medycznego itp.) na drodze obliczeń arytmetycznych. Przykładem takiego problemu może być obliczenie całki oznaczonej $I = \int_0^1 f(x)dx$.

Uzyskanie rozwiązania oryginalnego problemu wymaga zastosowania metod korzystających z pojęć abstrakcyjnych i umiejętności analitycznych. Rozwiązanie można utożsamiać z wykonaniem odwzorowania Z „danych wejściowych D ” na „dane wyjściowe W ” $W = Z(D)$. Zarówno dane wejściowe i wyniki mogą mieć charakter abstrakcyjny (nie liczbowy), na przykład rozwiązaniem zadania wyznaczenia pochodnej funkcji $\ln(x)$ jest funkcja $\frac{1}{x}$.

Metoda numeryczna to innymi słowy sposób postępowania prowadzący do wyniku, który jest przybliżeniem dokładnego rozwiązania postawionego zadania. Na przykład, jedną z możliwych metod obliczenia całki oznaczonej jest podział przedziału całkowania na podprzedziały o długości h i przybliżenia całki na każdym z podprzedziałów polem odpowiedniego trapezu, tak jak to pokazano na rys. 1.4.

Metoda numeryczna korzysta z danych wejściowych D_N , które muszą mieć postać skończonego zbioru liczb rzeczywistych i daje wyniki W_N o takim samym charakterze. Oryginalne zadanie Z musi być więc przeformułowane, by móc uzyskać i odpowiednio zinterpretować jego numeryczne rozwiązanie W_N . Przekształcenie danych wejściowych D_N w wyniki W_N realizuje algorytm Z_N przez odwzorowanie $W_N = Z_N(D_N)$, który daje precyzyjny, jednoznaczny przepis wykonania wszystkich operacji. Metoda numeryczna jest często wyposażona w parametry (jak parametr h w przykładzie dotyczącym całkowania), od których zależą jej właściwości.



Rys. 1.4. Przykład numerycznej metody obliczenia całki oznaczonej

Algorytm Z_N jest realizowany w konkretnej arytmetyce zmiennopozycyjnej. **Numeryczna realizacja algorytmu Z_{Nfl}** obejmuje zamianę danych wejściowych D_N na ich reprezentację zmiennoprzecinkową D_{Nfl} i wykonanie obliczeń zgodnie z regułami stosowanej arytmetyki zmiennopozycyjnej. Realizacja numeryczna algorytmu doprowadza więc do wyników W_{Nfl} , które są liczbami maszynowymi: $W_{Nfl} = Z_{Nfl}(D_{Nfl})$.

Wynik obliczeń wykonanych przez maszynę cyfrową różni się od wyniku dokładnego. Mówimy, że jest obciążony błędem. Przyczyny błędów są wielorakie. Poniżej sklasyfikujemy ich podstawowe źródła:

- **Błąd danych wejściowych** – Jeżeli dane, które przyjmujemy do obliczeń pochodzą z pomiarów, to są obciążone błędem wynikającym z dokładności zastosowanych urządzeń pomiarowych. Jeżeli dane wejściowe są wynikiem wcześniejszych obliczeń numerycznych, są obciążone błędem tych obliczeń.
- Podstawowe znaczenie ma powszechny i dotyczący praktycznie wszystkich obliczeń składnik błędu danych wejściowych zwany **błędem reprezentacji**. Jest to błąd wynikający z faktu, że nie każdą liczbę rzeczywistą można przedstawić w postaci maszynowej. Zamiana reprezentacji liczb z systemu dziesiętnego na system dwójkowy, używany w przeważającej większości maszyn cyfrowych, powoduje z reguły, że skończone rozwinięcie staje się nieskończonym, więc niereprezentowalnym maszynowo. **Błąd reprezentacji jest związany z numeryczną realizacją algorytmu.**
- **Błędy skrócenia (ucięcia/zaokrąglenia) wyników operacji arytmetycznych** – U podłoża tych błędów tkwi także skończona reprezentacja liczb rzeczywistych.

stych w maszynach cyfrowych. Występują one, gdy wynik operacji arytmetycznej czy obliczenia funkcji na argumentach reprezentowalnych maszynowo nie daje się dokładnie przedstawić w postaci maszynowej. Niekiedy, w przypadku gdy maszyna cyfrowa nie zaokrągliła, ale ucina nie mieszczące się w rejestrach rozwinięcie wyniku, mówi się o błędzie ucięcia (nie mylić z błędem obcięcia), jest to jednak coraz rzadsze, gdyż większość maszyn cyfrowych realizuje obliczenia zmiennoprzecinkowe zgodnie z normą IEEE754, a więc dokonując zaokrągleń. **Błąd skrócenia jest związany z numeryczną realizacją algorytmu.**

- **Błędy metody lub inaczej obcięcia** – Metoda numeryczna jest zwykle (choć nie zawsze) obciążona błędem wynikającym z jej założeń i uproszczeń, zależnym od parametrów metody. Błąd ten pojawia się nawet wtedy, gdy wszystkie obliczenia są wykonane w dokładny sposób, bez zaokrągleń. Suma pól trapezów z rysunku 1.4 różni się od dokładnej wartości całki. Często wynik metody numerycznej można przedstawić w postaci $W_N = W + S(h)$, gdzie W jest wartością dokładną, a $S(h)$ jest sumą zbieżnego szeregu (potęgowego) pewnego parametru metody h i $\lim_{h \rightarrow 0} S(h) = 0$. Błąd metody wynika więc z odrzucenia (obcięcia) $S(h)$ i przyjęcia $W_N \approx W$ i dlatego jest nazywany błędem obcięcia. **Błąd metody jest związany z algorytmem Z_N , a nie z jego numeryczną realizacją.**

Oprócz tych błędów, które są obiektem badań analizy numerycznej występują i inne, których aparatem metod numerycznych nie zbadamy. Na przykład:

- **Błędy wnoszone przez uproszczenie modelu matematycznego.**
- **Błędy człowieka.**

Niezależnie od źródła błędów zawsze można przyjąć, że w obliczeniach numerycznych mamy do czynienia z **wartością dokładną a i przybliżoną** (obliczoną, zaokrągloną itp.) \tilde{a} .

Definicja 1.2 (błędu bezwzględnego i względnego)

Błędem bezwzględnym Δ_a nazywamy różnicę wartości przybliżonej i wartości dokładnej:

$$\Delta_a \stackrel{\text{def}}{=} \tilde{a} - a. \quad (1.13)$$

Błędem względnym ε_a nazywamy stosunek błędu bezwzględnego do wartości dokładnej

$$\varepsilon_a \stackrel{\text{def}}{=} \frac{\Delta_a}{a} = \frac{\tilde{a} - a}{a} = \frac{\tilde{a}}{a} - 1, \quad a \neq 0. \quad (1.14)$$

Definicje te zawierają wartość dokładną, więc nie mogą być bezpośrednio stosowane do wyznaczania błędu. **Z reguły nie znamy nawet znaku błędu i możemy jedynie szacować jego moduł.**

Przykład 1.6

Z faktu, że wartość przybliżona $\tilde{a} = 1,22$ powstała z poprawnego zaokrąglenia wartości dokładnej a do dwóch cyfr po przecinku wynika jedynie, że $|\Delta_a| \leq 0,005$. Nie wiadomo, czy wartość dokładna jest większa, czy mniejsza od wartości przybliżonej.

Błąd bezwzględny przybliżenia $\tilde{\pi} = 3,14$ liczby π można podać z dowolną dokładnością i wiadomo, że jest on ujemny.

Bezpośrednio z definicji można wyprowadzić przydatne nierówności, które pozwalają szacować błędy na podstawie wartości przybliżonej.

Z definicji (1.13) wartość dokładna znajduje się w przedziale $\tilde{a} - |\Delta_a| \leq a \leq \tilde{a} + |\Delta_a|$. Jeżeli $\tilde{a} - |\Delta_a| > 0$, to $|\tilde{a} - |\Delta_a|| \leq |a|$, a jeśli $\tilde{a} + |\Delta_a| < 0$, to $|\tilde{a} + |\Delta_a|| \leq |a|$, więc

$$|\varepsilon_a| = \left| \frac{\Delta_a}{a} \right| \leq \begin{cases} \frac{|\Delta_a|}{\tilde{a} - |\Delta_a|} & \text{jeśli } \tilde{a} - |\Delta_a| > 0 \\ \frac{|\Delta_a|}{-\tilde{a} - |\Delta_a|} & \text{jeśli } \tilde{a} + |\Delta_a| < 0 \end{cases}. \quad (1.15)$$

Jeżeli $0 < |\varepsilon_a| < 1$, to z (1.13) i (1.14) wynika

$$|\Delta_a| \leq \frac{|\varepsilon_a|}{1 - |\varepsilon_a|} |\tilde{a}|. \quad (1.16)$$

1.5. Błędy skrótów i zaokrągleń

Skrócenie liczby oznacza zastąpienie jej inną wartością, która jest jej w przybliżeniu równa, ale ma krótszą, prostszą lub bardziej pożądaną reprezentację.

Przykład 1.7

Skróceniem będzie:

- zastąpienie 125,4478 zł kwotą 125,45 zł,
- zastąpienie ułamka $\frac{312}{937}$ ułamkiem $\frac{1}{3}$,
- zastąpienie wyrażenia $\sqrt{2}$ wartością 1,414,
- generalizacja wyniku – zamiast powiedzieć 127481 wyborców powiemy, że było około 130 tys. wyborców.

Często wartość przybliżona \tilde{a} powstaje przez skrócenie wartości dokładnej a do t cyfr mantysy. Różne sposoby skrócenia liczby przedstawiono na rys. 1.5.



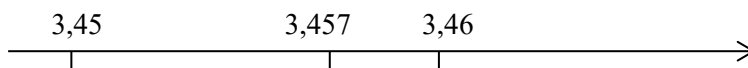
Rys. 1.5. Różne sposoby skrócenia liczby do t cyfr mantysy

Najprostszą i zarazem najgorszą z możliwych metod jest metoda polegająca na odrzuceniu wszystkich cyfr, które nie mieszczą się w pożądanej reprezentacji liczby (skrócenie przez ucięcie). W takim przypadku zawsze zachodzi $|\tilde{a}| < |a|$, co w przypadku wielokrotnych operacji arytmetycznych prowadzi do kumulacji błędów. Zdecydowanie lepszą metodą jest skrócenie przez zaokrąglenie.

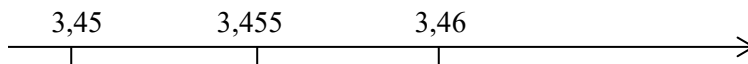
Zaokrąglenie polega na zastąpieniu zaokrąglanej liczby jej najbliższym „sąsiadem” mającym t cyfr mantysy, czyli sprowadza się do wyboru tej jednej z dwóch liczb mających t cyfr mantysy, dla której moduł błędów bezwzględnych zaokrąglenia będzie mniejszy.

Przykład 1.8

Załóżmy, że chcemy zaokrąglić liczbę 3,457 do dwóch cyfr po przecinku.



Oczywiste jest, że bliższym rozwiązaniem (i jedynym słusznym w tym przypadku) jest „prawe sąsiedztwo”, tak więc $3,457 \approx 3,46$. Co by jednak było, gdybyśmy chcieli jak poprzednio zaokrąglić do dwóch miejsc po przecinku liczbę 3,455.



W tym przypadku „obaj sąsiedzi” są równo odlegli, musimy w inny sposób określić, który z nich będzie lepszym przybliżeniem, czyli zdecydować czy $3,455 \approx 3,45$, czy $3,455 \approx 3,46$.

W przypadku gdy zaokrąglana wartość znajduje się dokładnie w połowie przedziału (moduł błędów bezwzględnych zaokrąglenia jest taki sam w obu przypadkach) stosuje się jedną z poniższych metod.

- Zaokrąglenie w górę (zaokrąglenie w kierunku do plus nieskończoności) oznacza, że wartość zaokrąglana jest zastąpiona wartością większą („sąsiad z prawej strony”). *Taki sposób zaokrąglania jest niesymetryczny to znaczy, że po zaokrągleniu liczb x i $-x$ otrzymamy liczby o różnych modułach.*
- Zaokrąglenie w dół (zaokrąglenie w kierunku do minus nieskończoności) oznacza, że wartość zaokrąglana jest zastąpiona wartością mniejszą („sąsiad z lewej strony”).
- Zaokrąglenie w kierunku od zera (zaokrąglenie w kierunku do nieskończoności): wartość środkowa jest zaokrąglana do jej sąsiada o większym module. *Metoda ta traktuje dodatnie i ujemne wartości symetrycznie.*
- Zaokrąglenie w kierunku do zera: wartość środkowa jest zaokrąglana do jej sąsiada o mniejszym module.
- Zaokrąglenie do cyfry parzystej (zaokrąglenie: bankierskie, bezstronne, statystyczne, holenderskie, nieparzysto-parzyste): ostatnia cyfra po zaokrągleniu musi być parzysta. *Wartości ujemne i dodatnie są traktowane symetrycznie.*

Zaokrąglenie można zdefiniować, wykorzystując definicję liczb zmiennoprzecinkowych.

Definicja 1.3 (zaokrąglenia do liczby maszynowej)

Dla zbioru liczb zmiennoprzecinkowych $\Phi(p, t, c_{\min}, c_{\max})$ z parzystą podstawą p , zaokrągleniem do t cyfr mantysy nazywamy wynik działania funkcji

$rd: \{x: x_{\min} \leq |x| \leq x_{\max}\} \rightarrow \Phi \subset R$ określonej przez

$$rd(x) = \begin{cases} \sigma\left(\sum_{k=1}^t d_k p^{-k}\right) p^c & \text{dla } d_{t+1} \leq \frac{p}{2} - 1 \\ \sigma\left(\sum_{k=1}^t d_k p^{-k} + p^{-t}\right) p^c & \text{dla } d_{t+1} \geq \frac{p}{2} \end{cases} \quad (1.17)$$

$rd(x)$ oznacza wartość $x = \sigma(\sum_{k=1}^{\infty} d_k p^{-k}) p^c$ zaokrągloną do t cyfr mantysy.

Dla danego zbioru liczb zmiennoprzecinkowych $\Phi(p, t, c_{\min}, c_{\max})$, dla każdego rzeczywistego $x: x_{\min} \leq |x| \leq x_{\max}$ zachodzi

$$rd(x) \in \Phi(p, t, c_{\min}, c_{\max}),$$

$$\frac{|rd(x) - x|}{|x|} \leq \frac{1}{2} p^{1-t} = \frac{1}{2} eps, \quad (1.18)$$

przy czym epsilon maszynowy eps został zdefiniowany w (1.9).

Definicja 1.4 (ucięcia do liczby maszynowej)

Dla zbioru liczb zmiennoprzecinkowych $\Phi(p, t, c_{min}, c_{max})$, **ucięciem** do t cyfr matysy nazywamy wynik działania funkcji $tc: \{x: x_{min} \leq |x| \leq x_{max}\} \rightarrow \Phi \subset R$ określonej przez

$$tc(x) = \sigma \left(\sum_{k=1}^t d_k p^{-k} \right) p^c, \quad (1.19)$$

$tc(x)$ oznacza wartość $x = \sigma(\sum_{k=1}^{\infty} d_k p^{-k}) p^c$ uciętą do t cyfr.

Dla danego zbioru liczb zmiennoprzecinkowych $\Phi(p, t, c_{min}, c_{max})$, dla każdego $x: x_{min} \leq |x| \leq x_{max}$ zachodzi

$$tc(x) \in \Phi(p, t, c_{min}, c_{max}), \frac{|tc(x) - x|}{|x|} \leq eps = p^{1-t}. \quad (1.20)$$

1.6. Cyfry poprawne i znaczące

Z uwagi na zapis liczby w zmiennoprzecinkowych systemach liczbowych można przyjąć, że informacja o liczbie jest zawarta w jej cyfrach ułamkowych.

Dla uproszczenia rozważmy system dziesiętny (dla pozostałych systemów obliczeniowych pojęcia te są definiowane analogicznie) i liczby postaci $(-1)^s \cdot (0, f) \cdot 10^c$, gdzie $s \in \{0,1\}$ definiuje znak, f oznacza cyfry mantysy (rozwinienia dziesiętnego), a c jest cechą.

Cyfry rozwinięcia dziesiętnego, poczynając od stojącej na pozycji 10^{-1} , są nazywane **cyframi ułamkowymi**.

Porównajmy dwie liczby: $x = 0,1214$ i $y = 0,0053$. Obie mają tyle samo cyfr ułamkowych, ale do przedstawienia liczby y (gdybyśmy ją zapisali w postaci wykładniczej $y = 0,53 \cdot 10^{-2}$) wystarczyłyby tylko 2. Możemy zatem zdefiniować pojęcie **cyfr istotnych** – czyli wszystkich cyfr w reprezentacji liczby z pominięciem poprzedzających je cyfr 0.

Jeśli liczba dziesiętna \tilde{a} przybliża wartość dokładną a , to mówimy, że \tilde{a} ma t **poprawnych** cyfr ułamkowych, jeśli zachodzi

$$|\tilde{a} - a| \leq \frac{1}{2} 10^{-t}. \quad (1.21)$$

W systemie o podstawie p odpowiednikiem (1.21) jest nierówność

$$|\tilde{a} - a| \leq \frac{1}{2} p^{-t}. \quad (1.22)$$

Z pojęciem cyfr poprawnych wiąże się pojęcie **poprawnego zaokrąglenia**. Jeśli liczba a została poprawnie zaokrągloną do liczby \tilde{a} mającej t cyfr ułamkowych, to znaczy, że wszystkie te cyfry są poprawne, więc błąd zaokrąglenia spełnia nierówność (1.21).

Wszystkie cyfry uławkowe począwszy od pierwszej różnej od zera, aż do stojącej na pozycji 10^{-t} nazywamy **cyframi znaczącymi**, inaczej **cyfry znaczące to wszystkie poprawne cyfry istotne**.

Przykład 1.9

Rozważmy następujące liczby:

$$x = 0,532146793, y = 0,5320211522, x - y = 0,000125641.$$

Jeśli powtórzmy te obliczenia w systemie dziesiętnym z 5-cyfrową mantysą (zaokrąglając wartości do 5 cyfr mantysy), otrzymamy:

$$rd(x) = 0,53215, rd(y) = 0,53202, rd(x) - rd(y) = 0,00013.$$

Składniki mają po 5 cyfr znaczących, a wynik tylko 2. Tylko 2 cyfry mantysy zależą od cyfr w x i y . Trzy zera tylko uzupełniają cyfry mantysy do 5.

Błąd względny wyniku wynosi:

$$\left| \frac{x - y - (rd(x) - rd(y))}{x - y} \right| = \left| \frac{0,000125641 - 0,00013}{0,000125641} \right| \approx 0,035$$

podczas gdy błędy względne składników nie przekraczały $5 \cdot 10^{-5}$.

Zjawisko pokazane w przykładzie 1.9, polegające na tym, że liczba cyfr znaczących w wyniku jest znacznie mniejsza niż w argumentach, jest nazywane **utrata cyfr znaczących**. Należy tak prowadzić obliczenia numeryczne, by unikać utraty cyfr znaczących. *Wykorzystanie do dalszych obliczeń wyniku, w którym wystąpiła utrata cyfr znaczących, powoduje duże błędy bezwzględne w kolejnych operacjach.*

1.7. Przenoszenie się błędów w obliczeniach numerycznych

Większość obliczeń naukowo-technicznych jest złożonym, wieloetapowym procesem, a na ogół każdy etap obliczeń wnosi dodatkowe błędy wynikające z zaokrągleń lub innych źródeł. Zazwyczaj dane wejściowe, często pochodzące z pomiarów, także obarczone są błędem. Ich wprowadzanie do komputera wnosi błędy reprezentacji wynikające z faktu, że przywykliśmy posługiwać się układem dziesiętnym, natomiast większość maszyn cyfrowych działa w systemie dwójkowym. W związku z powyższym, dla wiarygodności uzyskanych wyników kapitalne znaczenie ma rzetelna analiza błędów, czyli określenie maksymalnego błędów wyniku na podstawie oszacowań błędów danych wejściowych, użytych metod obliczeniowych i arytmetyki zastosowanej maszyny cyfrowej.

Najprostszą, niestety również najbardziej żmudną metodą analizy propagacji błędów, jest metoda „krok po kroku” zwana inaczej analizą przedziałową. Algorytm dzieli się na kolejne etapy (działania elementarne, obliczenia wartości funkcji itp.).

Na podstawie oszacowania danych wejściowych i przybliżonej wartości uzyskanej na rozważanym etapie obliczeń szuka się przedziału, który zawiera dokładną wartość. Poprzez analizę kolejnych etapów dochodzi się do przedziału, w którym znajduje się dokładna wartość obliczanego rozwiązania. Metodę analizy przedziałowej można w „prosty” sposób zastosować jedynie, jeśli analizowane etapy dotyczą operacji monotonicznych względem wszystkich argumentów. Metoda daje wtedy nierówności dokładne.

Następna metoda wykorzystuje szybkie przybliżone szacowanie błędu względnie prostych operacji algorytmu (działań elementarnych, obliczeń wartości funkcji itp.) w oparciu o wzory przybliżone. Oszacowania te (wzory uproszczone) można wyprowadzić bezpośrednio z definicji błędów (1.13) i (1.14). Wyprowadzenia oszacowań modułu błędu względnego kilku podstawowych operacji zestawiono w tabeli 1.6. W celu wyprowadzenia nierówności dane wartości przybliżone i odpowiadające im błędy względne oznaczono: \tilde{x}_1 , ε_1 , \tilde{x}_2 , ε_2 , \tilde{x} , ε . Otrzymane nierówności są przybliżone – prawdziwe z dokładnością do błędu malejącego wraz z malejącymi błędami względnymi.

Tabela 1.6. Oszacowania błędu względnego

Operacja	Wyprowadzenie	Użyteczna nierówność
Iloczyn $y = x_1 x_2$	$\varepsilon_y = \frac{\tilde{x}_1 \tilde{x}_2}{x_1 x_2} - 1 = \frac{x_1(1 + \varepsilon_1)x_2(1 + \varepsilon_2)}{x_1 x_2} - 1 =$ $= (1 + \varepsilon_1)(1 + \varepsilon_2) - 1 =$ $= 1 + \varepsilon_1 + \varepsilon_2 + \varepsilon_1 \varepsilon_2 - 1 \approx \varepsilon_1 + \varepsilon_2$	$ \varepsilon_y \lesssim \varepsilon_1 + \varepsilon_2 $
Iloraz $y = \frac{x_1}{x_2}$	$\varepsilon_y = \frac{\tilde{x}_1 x_2}{x_1 \tilde{x}_2} - 1 = \frac{x_1(1 + \varepsilon_1)x_2}{x_1 x_2(1 + \varepsilon_2)} - 1 =$ $= \frac{(1 + \varepsilon_1)}{(1 + \varepsilon_2)} - 1 = \frac{(\varepsilon_1 - \varepsilon_2)}{(1 + \varepsilon_2)} \approx \varepsilon_1 - \varepsilon_2$	$ \varepsilon_y \lesssim \varepsilon_1 + \varepsilon_2 $
Pierwiastek $y = \sqrt{x}$	$\varepsilon_y = \frac{\sqrt{\tilde{x}}}{\sqrt{x}} - 1 = \frac{\sqrt{x(1 + \varepsilon)}}{\sqrt{x}} - 1 = \sqrt{1 + \varepsilon} - 1 =$ $= 1 + \frac{1}{2}\varepsilon - \frac{1}{8}\varepsilon^2 + \dots - 1 \approx \frac{1}{2}\varepsilon$	$ \varepsilon_y \lesssim \frac{1}{2} \varepsilon $

Tabela 1.6 (cd.)

Suma, różnica $y = x_1 \pm x_2$	$\varepsilon_y = \frac{\tilde{x}_1 \pm \tilde{x}_2}{x_1 \pm x_2} - 1 = \frac{x_1(1 + \varepsilon_1) \pm x_2(1 + \varepsilon_2)}{x_1 \pm x_2} - 1 =$ $= \frac{x_1 \varepsilon_1}{x_1 \pm x_2} \pm \frac{x_2 \varepsilon_2}{x_1 \pm x_2}$	$ \varepsilon_y \lesssim \left \frac{x_1}{x_1 \pm x_2} \right \varepsilon_1 + \left \frac{x_2}{x_1 \pm x_2} \right \varepsilon_2 $
<p>Wykorzystano definicje (1.13), (1.14) i wynikające z nich tożsamości</p> $\tilde{a} = a + \Delta_a = a + \varepsilon_a a = (1 + \varepsilon_a)a, \varepsilon_a = \frac{\Delta_a}{a} = \frac{\tilde{a} - a}{a} = \frac{\tilde{a}}{a} - 1, \quad a \neq 0.$ <p>Symbol \lesssim oznacza, że nierówność jest spełniona, jeśli liczby $\varepsilon_1 , \varepsilon_2 , \varepsilon$ są dostatecznie małe.</p>		

Otrzymane nierówności pozwalają na szybkie szacowanie błędu względnego dużej liczby szeregowo wykonywanych operacji. Następnie można wykorzystać nierówność (1.16) do oszacowania błędu bezwzględnego wyniku.

Przykład 1.10

Nierówność $|\varepsilon_y| \lesssim \left| \frac{x_1}{x_1 \pm x_2} \right| |\varepsilon_1| + \left| \frac{x_2}{x_1 \pm x_2} \right| |\varepsilon_2|$ wyprowadzona w tabeli 1.6 dla błędu względnego sumy (lub różnicy) jest inną ilustracją zjawiska utraty cyfr znaczących. Jeśli współczynnik $\left| \frac{x_i}{x_1 \pm x_2} \right|$ będzie duży (gdy wynik jest znacznie mniejszy od składników), błąd względny wyniku może być duży mimo małego błędu względnego składników.

Kolejną przybliżoną metodą szacowania błędu obliczeń numerycznych, które można przedstawić jako obliczanie wartości $y = f(x_1, x_2, \dots, x_n)$ funkcji wielu zmiennych jest wykorzystanie tzw. formuły różniczki zupełnej.

Jeśli $\Delta_f = f(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n) - f(x_1, x_2, \dots, x_n)$, $\Delta_{x_i} = \tilde{x}_i - x_i$, $\varepsilon_i = \frac{\Delta_{x_i}}{x_i}$ to

$$\Delta_f \approx \sum_{i=1}^n \left[\frac{\partial}{\partial x_i} f(x_1, x_2, \dots, x_n) \right] \Bigg|_{\substack{x_1 = \tilde{x}_1 \\ \vdots \\ x_n = \tilde{x}_n}} \Delta_{x_i}, \quad (1.23)$$

skąd wynika przybliżone oszacowanie

$$|\Delta_f| \lesssim \sum_{i=1}^n \left| \left[\frac{\partial}{\partial x_i} f(x_1, x_2, \dots, x_n) \right] \Bigg|_{\substack{x_1 = \tilde{x}_1 \\ \vdots \\ x_n = \tilde{x}_n}} \right| |\Delta_{x_i}|. \quad (1.24)$$

Błąd względny otrzymujemy z wyrażenia

$$\begin{aligned} \varepsilon_f &= \frac{\Delta_f}{f(x_1, x_2, \dots, x_n)} \\ &\approx \sum_{i=1}^n \frac{x_i}{f(x_1, x_2, \dots, x_n)} \left[\frac{\partial}{\partial x_i} f(x_1, x_2, \dots, x_n) \right] \Bigg|_{\substack{x_1=\tilde{x}_1 \\ \vdots \\ x_n=\tilde{x}_n}} \frac{\Delta_{x_i}}{x_i}, \end{aligned} \quad (1.25)$$

gdzie $x_i \neq 0, f(x_1, x_2, \dots, x_n) \neq 0,$

co prowadzi do nierówności

$$|\varepsilon_f| \lesssim \sum_{i=1}^n \left| \frac{x_i}{f(x_1, x_2, \dots, x_n)} \right| \left| \left[\frac{\partial}{\partial x_i} f(x_1, x_2, \dots, x_n) \right] \Bigg|_{\substack{x_1=\tilde{x}_1 \\ \vdots \\ x_n=\tilde{x}_n}} \right| \left| \frac{\Delta_{x_i}}{x_i} \right|. \quad (1.26)$$

Czynnik $\left| \frac{x_i}{f(x_1, x_2, \dots, x_n)} \right|$, w którym występują wartości dokładne należy zastąpić wyrażeniem zawierającym wartości przybliżone, tak by nie zmienić zwrotu nierówności (1.26). Na przykład dla $x_i - |\Delta_{x_i}| > 0, f(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n) - |\Delta_f| > 0,$ gdzie $|\Delta_f|$ jest oszacowaniem obliczonym w (1.24) mamy $\left| \frac{x_i}{f(x_1, x_2, \dots, x_n)} \right| \leq \frac{x_i + |\Delta_{x_i}|}{f(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n) - |\Delta_f|}$.

Metoda różniczki zupełnej jest metodą przybliżoną. Jej stosowanie wymaga obliczenia pochodnych cząstkowych, jest więc dość ograniczone. Wynika z niej ważny, choć dość oczywisty wniosek: **obliczanie wartości funkcji, których pochodne są duże jest bardziej wrażliwe na błędy danych wejściowych.**

Od końca lat sześćdziesiątych rozwijane są co raz bardziej zaawansowane metody **arytmetyki przedziałowej**, które pozwalają na uzyskanie oszacowań błędów wyników skomplikowanych obliczeń. W roku 2015 sformułowano normę IEEE 1788-2015 regulującą modele i zasady obliczeń arytmetyki przedziałowej w zgodzie ze standardami definiującymi arytmetykę zmiennopozycyjną.

Do analizy przenoszenia się błędów obliczeń numerycznych stosuje się też metody **geometrii obliczeniowej** – działu algorytmiki rozwijającego metody pozwalające wykonywać działania na obiektach geometrycznych, takich jak zbiory punktów, odcinków, wielokątów czy też okręgów.

Przykład 1.11

Obliczmy wartość wyrażenia $z = \frac{\sqrt{a}}{a+b}$ dla wartości przybliżonych $\tilde{a} = 4,0$ i $\tilde{b} = 3,5,$ które powstały poprzez poprawne zaokrąglenie wartości dokładnych.

Przybliżoną wartością wyniku jest $\tilde{z} = \frac{\sqrt{4,0}}{4,0+3,5} = 0,26(6).$

- Korzystając z analizy przedziałowej „krok po kroku”, otrzymujemy: ponieważ \tilde{a} jest poprawnie zaokrąglone $a \in (3,95, 4,05)$, a zatem $\sqrt{a} \in (1,9875, 2,0125)$, analogicznie $b \in (3,45, 3,55)$, zatem $a + b \in (7,40, 7,60)$, wykonując dzielenie (szacując zakładamy najmniejszą wartość licznika i największą mianownika i analogicznie największą wartość licznika i najmniejszą mianownika) otrzymujemy $z = \frac{\sqrt{a}}{a+b} \in (0,2615, 0,2720)$.

Powyższe kroki są równoważne zapisowi

$$\frac{\sqrt{3,95}}{4,05 + 3,55} \leq z \leq \frac{\sqrt{4,05}}{3,95 + 3,45}$$

$$0,2615 \leq z \leq 0,2720$$

który daje prawdziwe, choć konserwatywne nierówności (przecież a nie może jednocześnie być równe 3.95 (tak jak przyjęto w prawej części powyższej nierówności) i 4.05 (jak przyjęto w lewej części nierówności)).

Błąd bezwzględny spełnia nierówność

$$|\Delta_z| \leq \max\{0,2720 - 0,26(6), 0,26(6) - 0,2615\} =$$

$$= \max\{0,0053, 0,0052\} = 0,0053$$

$$0,0053 > 0,5 \cdot 10^{-2},$$

więc szacując błąd w ten sposób, musimy przyjąć, że w otrzymanym wyniku \tilde{z} jest tylko jedna cyfra poprawna.

- Korzystając ze wzorów uproszczonych, otrzymujemy następujące wartości:

$$\tilde{a} = 4,0, |\Delta_a| \leq 0,05 \text{ i } |\varepsilon_a| = \frac{|\Delta_a|}{|a|} \leq \frac{0,05}{3,95} \approx 0,0162 \approx 1,62\%,$$

$$\text{zatem } |\varepsilon_{\sqrt{a}}| \approx \frac{1}{2} |\varepsilon_a| \leq 0,0081 = 0,81\%, \quad \tilde{b} = 3,5, |\Delta_b| = 0,05,$$

$$|\Delta_{a+b}| \leq |\Delta_a| + |\Delta_b| = 0,1 \text{ oraz } |\varepsilon_{a+b}| \leq \frac{0,1}{7,4} = 0,0135 \approx 1,35\%.$$

Jeśli błędy licznika i mianownika potraktujemy niezależnie (co nie jest prawdą), to

$$|\varepsilon_z| \lesssim |\varepsilon_{\sqrt{a}}| + |\varepsilon_{a+b}| = 0,0162 + 0,0135 = 0,0297 = 2,97\%.$$

Wykorzystanie nierówności (1.16) daje

$$|\Delta_z| \leq \frac{|\varepsilon_z|}{1-|\varepsilon_z|} |\tilde{z}| = \frac{0,0297}{1-0,0297} 0,2(6) = 0,0306 \cdot 0,2(6) = 0,0082,$$

a zatem oszacowanie tą metodą jest znacznie bardziej pesymistyczne.

- Korzystając z metody różniczki zupełnej, otrzymujemy

$$|\Delta_z| \lesssim \left| \frac{\partial z}{\partial a} \right|_{\substack{a=4,0 \\ b=3,5}} |\Delta_a| + \left| \frac{\partial z}{\partial b} \right|_{\substack{a=4,0 \\ b=3,5}} |\Delta_b| = \left| \frac{\frac{a+b}{2\sqrt{a}} - \sqrt{a}}{(a+b)^2} \right|_{\substack{a=4,0 \\ b=3,5}} |\Delta_a| + \left| -\frac{\sqrt{a}}{(a+b)^2} \right|_{\substack{a=4,0 \\ b=3,5}} |\Delta_b| = (0,002(2) + 0,035(5)) \cdot 0,05 \approx 0,0019,$$

wynik mocno odbiegający od poprzednich metod.

- Dokładna analiza funkcji $z = \frac{\sqrt{a}}{a+b}$ pozwala stwierdzić, że na zbiorze $a \in (3,95, 4,05)$, $b \in (3,45, 3,55)$ przyjmuje ona wartość $z_{\min} = 0,2648$ i $z_{\max} = 0,2686$.

Daje to błąd bezwzględny

$$|\Delta_z| \leq \max\{0,2686 - 0,26(6), 0,26(6) - 0,2648\} \\ = \max\{0,0019, 0,0019\} = 0,0019$$

i względny

$$|\varepsilon_z| \leq \frac{0,0019}{0,2648} = 0,0072 = 0,72\%.$$

W rozpatrywanym przykładzie metoda różniczki zupełnej dała najbardziej poprawne oszacowanie błędu. Metoda analizy przedziałowej i metoda szacowania błędu względnego dały bardzo zachowawcze wyniki. Spowodowane to jest niemonotonicznością funkcji z i występowaniem tego samego argumentu w liczniku i mianowniku – co zostało zignorowane przy stosowaniu tych metod.

1.8. Uwarunkowanie zadania numerycznego

Uwarunkowanie zadania obliczeniowego to właściwość, która charakteryzuje wpływ zaburzeń w danych wejściowych na wynik, niezależnie od wybranego algorytmu obliczeń.

Przypuśćmy, że zadanie numeryczne jest opisane odwzorowaniem $f: R^n \rightarrow R^m$.

Dokładne dane wejściowe x zostaną zaokrąglone do liczb \tilde{x} (np. do najbliższych liczb zmiennoprzecinkowych), na których zostaną przeprowadzone obliczenia w maszynie cyfrowej. Można powiedzieć, że liczby maszynowe \tilde{x} reprezentują zbiór DWE (Danych Wejściowych) wszystkich liczb, które mogą być zaokrąglone do \tilde{x} . Zadanie $f: R^n \rightarrow R^m$ odwzorowuje ten zbiór w zbiór wyników DWY (Danych Wyjściowych). Wpływ zaburzeń danych wejściowych na wynik może być oceniany przez stosunek odpowiednich „miar” zbiorów DWY do DWE.

Jeżeli niewielkie zaburzenia danych wejściowych powodują duże zmiany wyniku, to zadanie numeryczne będziemy nazywać **źle uwarunkowanym**.

Liczbową miarą uwarunkowania zadania jest wskaźnik (współczynnik) uwarunkowania. Wskaźnik uwarunkowania może dotyczyć błędu bezwzględnego albo względnego i może być definiowany w oparciu o różne normy w przestrzeni danych wejściowych i wyników.

Definicja (wskaźników uwarunkowania zadania numerycznego)

Bezwzględnym wskaźnikiem uwarunkowania zadania $f: R^n \rightarrow R^m$ w sensie normy $\|\cdot\|$ nazywamy najmniejszą liczbę $\kappa_{\|abs\|}$ taką, że

$$\|f(\tilde{x}) - f(x)\| \leq \kappa_{\|abs\|} \|\tilde{x} - x\| \quad \text{dla } \tilde{x} \rightarrow x. \quad (1.27)$$

Względny wskaźnikiem uwarunkowania zadania $f: R^n \rightarrow R^m$ w sensie normy $\|\cdot\|$ nazywamy najmniejszą liczbę $\kappa_{\|rel\|}$ taką, że

$$\frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} \leq \kappa_{\|rel\|} \frac{\|\tilde{x} - x\|}{\|x\|} \quad \text{dla } \tilde{x} \rightarrow x, \quad \|f(x)\|, \|x\| \neq 0. \quad (1.28)$$

Oba zdefiniowane wskaźniki uwarunkowania mają charakter lokalny, to znaczy ich wartość zależy od argumentu x . Uwarunkowanie względne ma większe znaczenie dla oceny właściwości zadania obliczeniowego i terminy „uwarunkowanie” oraz „wskaźnik uwarunkowania” będą odnosić się do definicji (1.28), jeśli nie powiedziano inaczej.

Jeżeli odwzorowanie $f: R^n \rightarrow R^m$ jest różniczkowalne, to po wprowadzeniu oznaczeń $\tilde{x} = x + \delta x$, $f(\tilde{x}) = f(x) + \delta f$ można przekształcić definicję (1.28) do postaci

$$\kappa_{\|rel\|} = \lim_{\delta x \rightarrow 0} \sup_{\delta x} \frac{\|\delta f\|}{\|\delta x\|} \bigg/ \frac{\|f(x)\|}{\|x\|} = \frac{\left\| \frac{\partial f}{\partial x} \right\|}{\|f(x)\|} = \frac{\left\| \frac{\partial f}{\partial x} \right\| \|x\|}{\|f(x)\|}, \quad (1.29)$$

gdzie $\frac{\partial f}{\partial x}$ jest Jacobianem (macierzą pochodnych cząstkowych odwzorowania f , a norma Jacobianu jest zgodna z zastosowaną normą wektorową (patrz definicja 2.1), czyli spełnia nierówność

$$\left\| \frac{\partial f}{\partial x} \delta x \right\| \leq \left\| \frac{\partial f}{\partial x} \right\| \|\delta x\|. \quad (1.30)$$

Przykład 1.12

Zadanie polega na sumowaniu trzech liczb zaokrąglonych do drugiego miejsca po przecinku, czyli $f(x) = \sum_{i=1}^3 x_i$. Obliczymy wskaźnik uwarunkowania tego zadania numerycznego (posługując się normą $\|\cdot\|_\infty$) dla danych

$$\tilde{x}_a = [1,00 \ 2,00 \ 3,00]^T \text{ oraz } \tilde{x}_b = [-1,00 \ 0,03 \ 1,00]^T.$$

W obu przypadkach, ze względu na zastosowane zaokrąglenie, możemy przyjąć oszacowanie $(|\Delta_{x_i}| \leq 0,005$. Zbiór DWE jest więc sześcianem o środku w punkcie \tilde{x} i długości krawędzi $2 \cdot 0,005$, a zbiór DWY przedziałem o środku $\tilde{y} = \sum_{i=1}^3 \tilde{x}_i$ i długości $3 \cdot 2 \cdot 0,005$. A zatem, przyjmując na potrzeby obliczeń:

$$\|x\| := \|x\|_\infty \quad (1.31)$$

otrzymujemy:

$$|f(\tilde{x}) - f(x)| \leq 3\|\tilde{x} - x\|_\infty, \quad (1.32)$$

więc $\kappa_{\|abs\|_\infty} = 3$.

Wektor pochodnych cząstkowych odwzorowania $f(x)$ jest równy $\frac{\partial f}{\partial x} = [1 \ 1 \ 1]$.

Jego norma: $\left\| \frac{\partial f}{\partial x} \right\|_\infty = 3$, (patrz wzór (2.89)) więc $\kappa_{\|rel\|_\infty} = \frac{3\|x\|_\infty}{|f(x)|}$, czyli dla \tilde{x}_a można przybliżyć $\kappa_{\|rel\|_\infty} \approx \frac{3 \cdot 3}{6} = 1,5$ albo oszacować $\kappa_{\|rel\|_\infty} \leq \frac{3 \cdot 3,005}{5,985} = 1,5$, zaś dla \tilde{x}_b można przybliżyć $\kappa_{\|rel\|_\infty} \approx \frac{3 \cdot 1}{0,03} = 100$, albo oszacować $\kappa_{\|rel\|_\infty} \leq \frac{3 \cdot 1,005}{0,015} = 201$.

W praktyce najczęściej posługujemy się euklidesową normą wektorową

$$\|x\| = \|x\|_2 = \left(\sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}}. \quad (1.33)$$

Ze wzoru (1.29) można łatwo obliczyć uwarunkowanie prostych operacji matematycznych, na przykład:

- dla funkcji: $f(x) = \sqrt{x} \Rightarrow \kappa_{\|rel\|_2} = \frac{\frac{1}{2\sqrt{x}}|x|}{\frac{1}{\sqrt{x}}} = \frac{1}{2}$, z czego wynika, że uwarunkowanie pierwiastkowania jest dobre i nie zależy od wartości x ,

- dla funkcji:

$$f(x_1, x_2) = x_1 - x_2 \Rightarrow \kappa_{\|rel\|_2} = \frac{\|[1, -1]\|_2 \sqrt{x_1^2 + x_2^2}}{|x_1 - x_2|} = \sqrt{2} \frac{\sqrt{x_1^2 + x_2^2}}{|x_1 - x_2|},$$

czyli odejmowanie staje się źle uwarunkowane, jeśli $x_1 \rightarrow x_2$!

- dla funkcji: $f(a, b) = \langle a, b \rangle = a^T b$ gdzie: $a = [a_1 \ a_2 \ \dots \ a_n]^T$,
 $b = [b_1 \ b_2 \ \dots \ b_n]^T$

$$\begin{aligned} f(a, b) &= \sum_{i=1}^n a_i b_i \Rightarrow \kappa_{\|rel\|_2} \\ &= \frac{\|[b_1 \ \dots \ b_n \ a_1 \ \dots \ a_n]\| \cdot \|[a_1 \ \dots \ a_n \ b_1 \ \dots \ b_n]^T\|}{|\sum_{i=1}^n a_i b_i|} \\ &= \frac{\sum_{i=1}^n (a_i^2 + b_i^2)}{|\sum_{i=1}^n a_i b_i|}, \end{aligned}$$

czyli obliczanie iloczynu skalarnego staje się źle uwarunkowane, jeśli wektory są prawie ortogonalne!

Należy pamiętać, że wskaźnik uwarunkowania pomnożony przez błąd danych wejściowych daje jedynie przybliżenie błędu wyniku, tym dokładniejsze im mniejszy jest błąd argumentu (danych wejściowych).

Jeżeli wyznaczenie pochodnych funkcji $f(x)$ opisującej zadanie obliczeniowe nie jest możliwe, to można szacować wskaźnik uwarunkowania, posługując się definicją (1.28).

Jeśli rozważane błędy danych wejściowych są jedynie błędami reprezentacji zmiennoprzecinkowej liczb, to dla normy euklidesowej i dla normy $\|x\| := \|x\|_\infty$ zachodzi

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq eps, \tag{1.34}$$

(eps – epsilon maszynowy stosowanego systemu liczb zmiennopozycyjnych).
 Wtedy

$$\frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} \leq \kappa_{\|rel\|} \cdot eps \quad \text{dla } \tilde{x} \rightarrow x, \quad \|f(x)\|, \|x\| \neq 0. \tag{1.35}$$

Błąd bezwzględny spowodowany zmiennoprzecinkową reprezentacją danych wejściowych to

$$\|f(\tilde{x}) - f(x)\| \leq \|f(x)\| \cdot \kappa_{\|rel\|} \cdot eps, \tag{1.36}$$

a jeśli dodamy do niego składnik odpowiadający błędowi zmiennoprzecinkowej reprezentacji wyniku $\|f(x)\| \cdot eps$, to otrzymamy wielkość nazywaną **błędem nieuniknionym**

$$E_{nu} = \|f(x)\| \cdot (\kappa_{\|rel\|} + 1) \cdot eps. \quad (1.37)$$

Błąd nieunikniony zależy zarówno od danych wejściowych, od uwarunkowania zadania jak i od stosowanej arytmetyki zmiennopozycyjnej.

1.9. Stabilność numeryczna algorytmu

Dobre lub złe uwarunkowanie zadania obliczeniowego jest jego właściwością niezależną od metody numerycznej, czy algorytmu, który będzie to zadanie realizował. Wskaźnik uwarunkowania opisuje wrażliwość wyniku na zaburzenie danych wejściowych, a nie na błędy zaokrągleń w operacjach zmiennoprzecinkowych wykonywanych w trakcie realizacji algorytmu i ich propagację w kolejnych etapach algorytmu.

Wiadomo, że to samo zadanie numeryczne można zrealizować stosując różne algorytmy obliczeń. Na przykład zadanie $f(a, b) = a^2 - b^2$ można zrealizować według algorytmu

$$A_1[f(a, b)] = a \cdot a - b \cdot b \quad (1.38)$$

albo

$$A_2[f(a, b)] = (a + b) \cdot (a - b). \quad (1.39)$$

Obydwa algorytmy są **równoważne matematycznie**, nie musi to jednak oznaczać ich **równoważności numerycznej**.

Oznaczmy przez $fl\{A[f(x)]\}$ ostateczny (czyli policzony w sposób przybliżony) wynik otrzymamy po numerycznej realizacji algorytmu. To znaczy realizacji, w której wszystkie liczby (dane wejściowe i współczynniki algorytmu) zostały zastąpione ich reprezentacjami zmiennopozycyjnymi, a wszystkie operacje zostały wykonane zgodnie z arytmetyką zmiennopozycyjną danej maszyny.

Miarą jakości algorytmu rozwiązywania określonego zadania numerycznego jest między innymi to, jak duży jest błąd generowany przez algorytm (po jego numerycznej realizacji) w porównaniu z błędami wynikającymi z niedokładnej reprezentacji danych wejściowych.

Jeśli małe błędy danych wejściowych kumulują się i rosną w czasie obliczeń, powodując, że wynik jest poważnie zaburzony (ewidentnie nieprawdziwy lub nawet nieskończony), mówimy o **numerycznej niestabilności algorytmu**.

Wykorzystując nierówność (1.36), możemy powiedzieć, że algorytm będziemy nazywać **stabilnym**, jeśli istnieje taka stała K , nazywana **współczynnikiem stabilności**, dla której (w pewnym zakresie danych wejściowych) spełniona jest nierówność

$$\frac{\|fl\{A[f(x)]\} - f(x)\|}{\|f(x)\|} \leq K \cdot \kappa_{rel} \cdot eps. \quad (1.40)$$

Innymi słowy: błąd względny po numerycznej realizacji stabilnego algorytmu jest K razy większy od błędu względnego powodowanego tylko zaburzeniem danych wejściowych.

W przypadku stabilnego algorytmu możemy dowolnie poprawiać dokładność wyniku, zmniejszając błąd reprezentacji zmiennopozycyjnej, bowiem z (1.40) wynika

$$\lim_{eps \rightarrow 0} \frac{\|fl\{A[f(x)]\} - f(x)\|}{\|f(x)\|} = 0. \quad (1.41)$$

Zestawienie nierówności (1.36) i (1.40) pozwala na podanie równoważnego warunku stabilności algorytmu:

$$\|fl\{A[f(x)]\} - f(x)\| \leq K \cdot E_{nu}, \quad (1.42)$$

a więc: **algorytm stabilny jest obciążony błędem co najwyżej K razy większym od błędu nieuniknionego.**

By zilustrować powyższe zagadnienia, rozważmy przykład porównujący działanie dwóch algorytmów obliczeniowych.

Przykład 1.13

Niech zadanie numeryczne polega na obliczeniu $f(x) = x^n$ dla $x = \frac{1}{3}$. Wskaźnik uwarunkowania jest równy $\kappa_{rel} = \frac{|nx^{n-1}| \cdot |x|}{|x^n|} = n$. Zadanie to można zrealizować, posługując się dwoma algorytmami rekurencyjnymi:

$$A_1 \left\{ f_0 = 1, f_{m+1} = \frac{1}{3} f_m, m = 0, 1, \dots, n-1 \right\} \quad (1.43)$$

oraz

$$A_2 \left\{ f_0 = 1, f_1 = \frac{1}{3}, f_{m+1} = \frac{13}{3} f_m - \frac{4}{3} f_{m-1}, m = 1, 2, \dots, n-1 \right\}. \quad (1.44)$$

Oba algorytmy są równoważne matematycznie, gdyż jak łatwo zauważyć, jeśli $f_m = \frac{1}{3^m}$, $f_{m-1} = \frac{1}{3^{m-1}}$, to $f_{m+1} = \frac{13}{3} \frac{1}{3^m} - \frac{4}{3} \frac{1}{3^{m-1}} = \frac{1}{3^{m-1}} \left(\frac{13}{3} \cdot \frac{1}{3} - \frac{4}{3} \right) = \frac{1}{3^{m+1}}$.

Jeżeli dane wejściowe zostały zaburzone przez zaokrąglenie $\frac{1}{3} \approx 0,3333333333333333$, to z samego zaburzenia danych wejściowych można spodziewać się błędu względnego około

$$n \frac{3 \cdot 10^{-16}}{1/3} = 9n \cdot 10^{-16}.$$

W tabeli 1.7 podano wartości obliczone w Excelu za pomocą obydwóch algorytmów.

Tabela 1.7 Wartości $\left(\frac{1}{3}\right)^n$ obliczone za pomocą algorytmów A_1 i A_2

n	algorytm A_1	algorytm A_2	n	algorytm A_1	algorytm A_2
0	1,0000000000000000	1,0000000000000000	13	0,000000627225474	0,000000623130666
1	0,3333333333333333	0,3333333333333333	14	0,000000209075158	0,000000192695925
2	0,1111111111111111	0,1111111111111110	15	0,000000069691719	0,000000004174787
3	0,037037037037037	0,037037037037033	16	0,000000023230573	-0,000000238837156
4	0,012345679012346	0,012345679012330	17	0,000000007743524	-0,000001040527392
5	0,004115226337449	0,004115226337386	18	0,000000002581175	-0,000004190502491
6	0,001371742112483	0,001371742112233	19	0,000000000860392	-0,000016771474272
7	0,000457247370828	0,000457247369828	20	0,000000000286797	-0,000067089051857
8	0,000152415790276	0,000152415786277	21	0,000000000095599	-0,000268357259018
9	0,000050805263425	0,000050805247430	22	0,000000000031866	-0,001073429386602
10	0,000016935087808	0,000016935023827	23	0,00000000010622	-0,004293717663251
11	0,000005645029269	0,000005644773344	24	0,00000000003541	-0,017174870691952
12	0,000001881676423	0,000001880652721	25	0,00000000001180	-0,068699482780791

Jak widać, wyniki uzyskane za pomocą algorytmu A_2 są drastycznie niedokładne. Nie jest to spowodowane uwarunkowaniem zadania, lecz niestabilnością algorytmu.

Błąd obciążający f_m jest mnożony przez $\frac{13}{3}$, a więc błąd obciążający f_1 równy $3 \cdot 10^{-16}$ przenosi się do f_m pomnożony przez $\left(\frac{13}{3}\right)^{m-1}$. Do tego dochodzą błędy kolejnych elementów ciągu rekurencyjnego obciążające f_m z odpowiednimi mnożnikami. Już w f_2 widać utratę jednej cyfry poprawnej, w f_{13} są tylko dwie cyfry poprawne $\left(\left(\frac{13}{3}\right)^{12} \approx 43839457\right)$, a dalsze wyniki są już zupełnie niepoprawne.

Pojęcie stabilności algorytmu przedstawione wyżej można uznać za wynik tak zwanej **analizy progresywnej**. W analizie progresywnej badamy zbiór danych wyjściowych po uwzględnieniu wszystkich zaburzeń danych wejściowych oraz błędów algorytmu (oznaczymy go \overline{DWY}). Porównanie zbioru DWY zdefiniowanego wyżej ze zbiorem \overline{DWY} określa stabilność algorytmu w sensie analizy progresywnej. Wskaźnik stabilności jest definiowany jako czynnik, który obrazuje

powiększenie przez działanie algorytmu błędu nieuniknionego (związanego z uwarunkowaniem zadania i dokładnością maszynową).

Pojęcie stabilności algorytmu można też wprowadzić, używając tak zwanej **analizy wstecznej**. Koncepcja analizy wstecznej wywodzi się od Wilkinsona. Polega na przekształceniu błędów algorytmu wstecz, aż do wejścia i interpretacji ich jako błędów danych wejściowych. W przeciwieństwie do analizy progresywnej analiza wsteczna nie wymaga wcześniejszego badania uwarunkowania zadania. ***Stabilność w sensie analizy wstecznej jest zwykle silniejsza od stabilności w sensie analizy progresywnej.***

1.10. Złożoność obliczeniowa algorytmu

Często stawiamy sobie pytanie, który algorytm jest lepszy. Jednym z kryteriów oceny może być analiza propagacji błędu, innym pomiar zasobów potrzebnych do realizacji algorytmu. **Złożoność obliczeniowa algorytmu** rozumiana jest powszechnie jako miara wzrostu niezbędnego nakładu obliczeń przy wzroście wymiaru zadania (wzroście liczby danych wejściowych), albo przy zaostrzeniu wymagań dotyczących dokładności wyniku. System komputerowy rozwiązujący określony problem posiada do swej dyspozycji dwa podstawowe zasoby:

- pamięć (ang. memory, space),
- czas (ang. time).

Konsekwentnie definiuje się więc dwie niezależne miary:

- złożoność pamięciową (ang. space computational complexity/space complexity),
- złożoność czasową (ang. time computational complexity/time complexity).

Złożoność pamięciowa określa liczbę komórek pamięci, która będzie zajęta przez dane oraz wyniki pośrednie tworzone w trakcie pracy algorytmu. Wyrażana jest w liczbie bajtów lub w liczbie zmiennych (typów elementarnych) jako funkcji rozmiaru (liczby) danych wejściowych.

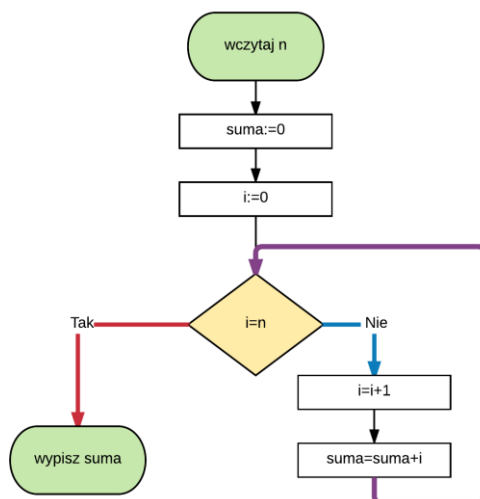
Z kolei **złożoność czasową** rozumiemy jako ilość czasu niezbędnego do rozwiązania danego problemu w zależności od rozmiaru (liczby) danych wejściowych. Jako miarę złożoności czasowej przyjmuje się zwykle liczbę operacji dominujących, które należy wykonać dla n danych wejściowych, aby otrzymać rozwiązanie problemu. Przez **operację dominującą** rozumiemy operację, której realizacja w decydujący sposób wpływa na czas wykonania całego algorytmu – np. jeśli wykonanie mnożenia jest wielokrotnie dłuższe niż innych operacji (jak było to w starszych maszynach cyfrowych), to liczba koniecznych mnożeń jest dobrą miarą złożoności obliczeniowej.

Współczesne komputery potrzebują mniej więcej tyle samo czasu na wykonanie: dodawania, odejmowania, mnożenia czy też dzielenia liczb zmiennoprzecinkowych. Dlatego wszystkie te operacje określane są wspólną nazwą **operacji zmiennoprzecinkowych**, a ich łączna liczba, której jednostką jest flop (ang. floating point operation), decyduje o czasie wykonania obliczeń.

W przypadku niektórych maszyn cyfrowych czas wykonania mnożenia jest znacząco różny od czasu wykonania dodawania. Dla takich realizacji, określając złożoność czasową, podaje się osobno liczbę mnożeń i dodawań.

Przykład 1.14

Obliczmy złożoność czasową algorytmu obliczającego sumę kolejnych liczb naturalnych od 1 do n w pętli przedstawionej na rysunku poniżej



Rys. 1.6. Algorytm sumowania n liczb w pętli

Sumując czasy wykonania poszczególnych operacji, otrzymamy

$$T(n) = t_1 + 2t_2 + (n + 1)t_3 + t_4 + n(t_5 + t_6 + t_7 + t_8) + t_9,$$

gdzie t_1 – czas wczytania liczby n , t_2 – czas inicjalizacji zmiennych, t_3 – czas operacji porównania, t_4 – czas operacji „idź do” (w przypadku Tak), t_5 – czas operacji „idź do” (w przypadku Nie), t_6, t_7 – czasy operacji dodawania, t_8 – czas operacji „idź do”, t_9 – czas potrzebny na wypisanie sumy. Grupując, otrzymujemy

$$T(n) = n(t_3 + t_5 + t_6 + t_7 + t_8) + t_1 + 2t_2 + t_3 + t_4 + t_9,$$

co po podstawieniu $t_a = t_3 + t_5 + t_6 + t_7 + t_8$ i $t_b = t_1 + 2t_2 + t_3 + t_4 + t_9$ daje nam liniową zależność $T(n) = t_a n + t_b$. Dla dużych n wartość t_b można

pominąć i przyjąć $T(n) \cong t_a n$. Taka złożoność obliczeniowa jest wyrażona w jednostkach czasu i zależy od czasu wykonania poszczególnych operacji algorytmu na konkretnym komputerze.

Drugim sposobem określenia złożoności czasowej jest wyznaczenie w algorytmie operacji dominującej i zliczenie liczby jej wykonań. W powyższym algorytmie za operację dominującą możemy przyjąć jeden obieg pętli sumującej liczby. Pozostałe operacje traktujemy jako nieistotne (tzn. przyjmujemy, że dla dużych n ich czas wykonania jest pomijalnie mały w porównaniu z czasem wykonania wszystkich operacji dominujących). Otrzymujemy wtedy $T_{flop}(n) \cong an$, gdzie a oznacza ilość operacji zmiennoprzecinkowych przypadających na realizację jednego obiegu pętli. Jak widać otrzymany wynik jest analogiczny do poprzedniego – złożoność obliczeniowa jest liniową funkcją n . Wyrażając złożoność obliczeniową przez liczbę operacji uniezależniliśmy wynik od konkretnej maszyny na której realizujemy algorytm.

Przykład 1.15

Policzmy ilość działań elementarnych (mnożeń i dodawań) niezbędną do obliczenia wartości wielomianu stopnia n w punkcie x :

- wykonując obliczenia w tradycyjny sposób, to znaczy korzystając z postaci potęgowej wielomianu, otrzymujemy dla przykładowego wielomianu 4 stopnia 10 mnożeń i 4 dodawania,

$$-2x^4 + 3x^3 + 5x^2 + 2x + 7 = -2 \cdot x \cdot x \cdot x \cdot x + 3 \cdot x \cdot x \cdot x + 5 \cdot x \cdot x + 2 \cdot x + 7,$$

- wykorzystując wyniki pośrednie $x^2 = x \cdot x$ i $x^3 = x^2 \cdot x$, możemy ograniczyć ilość działań elementarnych do 7 mnożeń i 4 dodawania,

$$-2x^4 + 3x^3 + 5x^2 + 2x + 7 = -2 \cdot x^3 \cdot x + 3 \cdot x^3 + 5 \cdot x^2 + 2 \cdot x + 7,$$

- przedstawiając wielomian z wykorzystaniem tzw. **schematu Hornera**, obliczymy wartość wielomianu, wykonując jedynie 4 mnożenia i 4 dodawania:

$$-2x^4 + 3x^3 + 5x^2 + 2x + 7 = (((-2 \cdot x + 3) \cdot x + 5) \cdot x + 2) \cdot x + 7.$$

Uogólniając, dochodzimy do wniosku, że obliczenie wartości wielomianu stopnia n z postaci potęgowej wymaga $T_{flop-p} = \frac{(n+3)n}{2}$ operacji elementarnych, zaś zastosowanie schematu Hornera jedynie $T_{flop-h} = 2n$ operacji elementarnych (połowa przypada na mnożenie, połowa na dodawanie).

W literaturze można spotkać pojęcie **klasy złożoności obliczeniowej** (ang. computational complexity class), posługując się symbolem Landau'a (patrz dodatek D3). Klasa złożoności obliczeniowej definiuje zbiór zadań obliczeniowych, do rozwiązania których potrzebna jest podobna ilość zasobów. Na przykład:

- $O(1)$ – stała klasa czasowej złożoności obliczeniowej – występuje, gdy algorytm wykonuje stałą liczbę operacji bez względu na rozmiar danych n ,
- $O(n)$ – liniowa klasa czasowej złożoności obliczeniowej – występuje, gdy algorytm wykonuje stałą liczbę operacji dla każdej danej; liczba operacji rośnie liniowo z liczbą danych,
- $O(n^2)$ – kwadratowa klasa czasowej złożoności obliczeniowej – występuje, gdy dla n -tej danej, algorytm wykonuje proporcjonalną do n liczbę operacji; liczba operacji rośnie z kwadratem liczby danych.

Istnieją też algorytmy mieszczące się w innych klasach złożoności obliczeniowej:

- $O(\log(n))$ – logarytmiczna,
- $O(n \log(n))$ – liniowo-logarytmiczna,
- $O(2^n)$ – wykładnicza złożoność obliczeniowa.

Znajomość klasy złożoności obliczeniowej czasowej i pamięciowej algorytmu pozwala informatykowi przewidywać jego zachowanie się dla różnych zestawów danych oraz dobrać algorytmy dla określonych sytuacji.

Wracając do przykładu 1.15 możemy powiedzieć, że algorytm 1 – obliczenie wartości wielomianu stopnia n z postaci potęgowej, jest algorytmem o złożoności $O(n^2)$, a algorytm 3 wykorzystujący schemat Hornera jedynie $O(n)$.

Ponieważ często zużycie zasobów w algorytmie jest uzależnione od postaci przetwarzanych danych, definiuje się również pojęcia **złożoności pesymistycznej** $T_W(n)$, **złożoności optymistycznej** $T_O(n)$ i **złożoności oczekiwanej** (średniej) $T_A(n)$.

Przykład 1.16

Przeszukując losowy ciąg n liczb całkowitych w celu znalezienia pierwszej liczby ujemnej, możemy wykonać algorytm o złożoności:

- $T_W(n) = n$ – jeśli w ciągu nie ma liczb ujemnych lub będzie ona ostatnią liczbą ciągu,
- $T_O(n) = 1$ – jeśli liczba ujemna będzie pierwszą liczbą ciągu,
- $T_A(n) = \frac{1}{2}n$ – jeśli oceniamy złożoność na podstawie średniej z wielu wykonań algorytmu i położenie liczby ujemnej jest losowe z rozkładem równomiernym.

Można zatem mówić o złożoności pesymistycznej, optymistycznej i średniej. Złożoność pesymistyczna określa zużycie zasobów dla najbardziej niekorzystnego zestawu danych, złożoność optymistyczna określa zużycie zasobów dla najkorzystniejszego zestawu danych, zaś złożoność średnia określa zużycie zasobów dla typowego (uśrednionego pod względem odpowiednich cech) zestawu danych.

2. Rozwiązywanie układów równań liniowych i rozkład trójkątny macierzy kwadratowej

Ten rozdział jest poświęcony podstawowym metodom rozwiązywania układu n równań liniowych z n niewiadomymi. Układy równań liniowych pojawiają się w rozlicznych problemach inżynierskich, logistycznych czy ekonomicznych. Liczba niewiadomych może być przy tym bardzo duża, co oznacza, że stosowanie metod znanych z podstawowego kursu algebry, jak na przykład wzory Cramera nie jest uzasadnione. Potrzebne są metody znajdujące rozwiązania obarczone niewielkim błędem, o umiarkowanej złożoności obliczeniowej, które mogą być stosowane dla n rzędu setek tysięcy. Istotna jest zarówno minimalizacja liczby wykonywanych operacji elementarnych – czyli czasu obliczeń jak i błędu, którym będzie obarczone rozwiązanie. Konieczność rozwiązania układu równań liniowych pojawia się też w innych problemach numerycznych, które będą omawiane w kolejnych rozdziałach. Zostaną tam wykorzystane wnioski i metody wyprowadzone w tym rozdziale.

2.1. Układy równań liniowych

Układ n równań liniowych z n niewiadomymi x_i , $i = 1, 2, \dots, n$ można zapisać w postaci skalarnej

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2 \\ \dots \quad \quad \quad \dots \quad \quad \dots \quad \quad \dots \quad \quad \dots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n = b_n \end{cases}, \quad (2.1)$$

lub odpowiadającej jej postaci macierzowej

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}. \quad (2.2)$$

Jeżeli macierze współczynników i niewiadomych zostaną oznaczone przez:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}, \quad (2.3)$$

to zapis układu równań upraszcza się do postaci

$$Ax = b. \quad (2.4)$$

Układ ten będzie miał dokładnie jedno rozwiązanie wtedy i tylko wtedy, gdy macierz współczynników A jest nieosobliwa, czyli

$$\det A \neq 0. \quad (2.5)$$

Korzystając z właściwości macierzy odwrotnej

$$A^{-1}A = AA^{-1} = I, \quad (2.6)$$

i mnożąc lewostronnie obie strony równania (2.4) przez A^{-1} , można przedstawić jego jedyne rozwiązanie jako

$$x = A^{-1}b. \quad (2.7)$$

Wzór (2.7) i wykorzystanie wzoru

$$A^{-1} = \frac{1}{\det A} \operatorname{adj}A, \quad (2.8)$$

gdzie macierz dołączona $\operatorname{adj}A$ (transponowana macierz dopełnień algebraicznych) została zdefiniowana w dodatku D2, do obliczenia macierzy odwrotnej nie jest efektywną numerycznie metodą rozwiązania równania (2.4). Odwrotnie, to metody rozwiązywania układu równań liniowych dostarczają wydajnych sposobów obliczania wyznacznika i odwracania macierzy.

Skutecznym sposobem rozwiązywania układu równań (2.4) jest przekształcenie go do równoważnej mu (czyli mającej to samo rozwiązanie) postaci

$$Tx = \hat{b}, \quad (2.9)$$

w której macierz współczynników T jest macierzą trójkątną (górną lub dolną). Taki układ równań jest nazywany **układem trójkątnym**. Rozwiązanie trójkątnego układu równań sprowadza się do rozwiązania (przez podstawienie) n równań liniowych z jedną niewiadomą. Rozwiązanie trójkątnego układu równań jest metodą bezpośrednią, wyznaczającą rozwiązanie w skończonej liczbie operacji. **Otrzymane w ten sposób rozwiązanie będzie obarczone jedynie błędem zaokrąglenia**. Oczywiście należy pamiętać, że będzie on się kumulował w miarę prowadzonych obliczeń, czyli w najmniejszym stopniu będzie dotyczył niewiadomych wyznaczonych na początku, w największym – na końcu.

Jeżeli macierz współczynników układu równań jest macierzą trójkątną dolną

$$\begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{bmatrix}, \quad (2.10)$$

to pierwszą niewiadomą wyznaczmy z pierwszego równania. Następnie podstawimy ją do drugiego i wyznaczmy z niego drugą niewiadomą. Postępując tak dalej, aż do ostatniej niewiadomej, znajdziemy rozwiązanie całego układu równań. Poszukiwane rozwiązanie można zapisać w postaci wzoru

$$x_i = \frac{1}{l_{ii}} [b_i - \sum_{k=1}^{i-1} l_{ik} x_k], \quad i = 1, 2, \dots, n. \quad (2.11)$$

Jeżeli macierz współczynników układu równań jest macierzą trójkątną górną

$$\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{bmatrix}, \quad (2.12)$$

to proces obliczania i podstawiania należy rozpocząć od ostatniego równania i zmiennej x_n , a następnie kontynuować dla równań $n-1, n-2, \dots, 1$. Algorytm ten pozwala wyznaczyć rozwiązanie w oparciu o wzór

$$x_i = \frac{1}{u_{ii}} \left[b_i - \sum_{k=i+1}^n u_{ik} x_k \right], \quad i = n, n-1, \dots, 1. \quad (2.13)$$

Metoda podstawienia będzie także bardzo skuteczna w przypadku dużych układów równań, o ile potrafimy je sprowadzić do postaci trójkątnej.

2.2. Eliminacja Gaussa

Jeżeli można stosunkowo łatwo rozwiązać układ równań liniowych, gdy macierz współczynników jest macierzą trójkątną, to należałoby dowolny układ równań sprowadzić do równoważnego (czyli mającego to samo rozwiązanie), w takiej właśnie postaci. Algorytm służący do przekształcenia układu równań liniowych do równoważnej postaci z trójkątną górną macierzą współczynników U nosi nazwę **eliminacji Gaussa** (nazwa pochodzi od nazwiska niemieckiego matematyka Karola Fryderyka Gaussa). Operacje, które zostaną wykorzystane do przekształcania układu równań liniowych to:

- mnożenie równania przez liczbę różną od zera,
- dodanie do jednego równania innego pochodzącego z tego samego układu równań,

- zamiana kolejności równań.

Żadna z tych operacji nie zmienia rozwiązania układu równań.

Zilustrujemy działanie algorytmu eliminacji Gaussa na przykładzie układu trzech równań z trzema niewiadomymi:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3 \end{cases} \quad (2.14)$$

a w postaci macierzowej

$$Ax = b$$

$$A := \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad b := \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}. \quad (2.15)$$

Załóżmy, że $a_{11} \neq 0$.

Żeby wyzerować współczynnik a_{21} , stojący przy zmiennej x_1 w drugim równaniu, odejmiemy od tego równania równanie pierwsze pomnożone przez liczbę $m_{21} = \frac{a_{21}}{a_{11}}$. Analogicznie, żeby wyzerować współczynnik a_{31} przy zmiennej x_1 w trzecim równaniu, odejmiemy od tego równania równanie pierwsze pomnożone przez liczbę $m_{31} = \frac{a_{31}}{a_{11}}$.

W wyniku tych operacji, które składają się na pierwszy etap eliminacji Gaussa, współczynniki przy zmiennych x_2 , x_3 i wyrazy wolne w drugim i trzecim równaniu zmieniają się (co zaznaczono górnymi indeksami), a układ równań przybierze postać

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 = b_2^{(2)}, \\ a_{32}^{(2)}x_2 + a_{33}^{(2)}x_3 = b_3^{(2)} \end{cases} \quad (2.16)$$

bądź równoważną mu postać macierzową

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2^{(2)} \\ b_3^{(2)} \end{bmatrix}. \quad (2.17)$$

Jeżeli $a_{22}^{(2)} \neq 0$, współczynnik $a_{32}^{(2)}$ stojący przy x_2 w trzecim równaniu można wyzerować, odejmując od trzeciego równania równanie drugie pomnożone przez liczbę $m_{32} = \frac{a_{32}^{(2)}}{a_{22}^{(2)}}$. Jest to drugi etap eliminacji Gaussa.

W efekcie otrzymuje się równoważny układowi (2.14) i (2.16) układ równań

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 = b_2^{(2)}, \\ a_{33}^{(3)}x_3 = b_3^{(3)} \end{cases}, \quad (2.18)$$

w którym, w każdym równaniu, patrząc od dołu do góry, mamy jedynie o jedną niewiadomą więcej. W reprezentacji macierzowej oznacza to, że macierz współczynników jest macierzą trójkątną górną.

$$Ux = b^{(3)}, \quad U = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} \\ 0 & 0 & a_{33}^{(3)} \end{bmatrix}, \quad b^{(3)} = \begin{bmatrix} b_1 \\ b_2^{(2)} \\ b_3^{(3)} \end{bmatrix}. \quad (2.19)$$

Kończy to eliminację Gaussa i pozwala przystąpić do rozwiązywania układu trójkątnego omówioną powyżej metodą podstawienia.

Z przedstawionego przykładu wynikają następujące spostrzeżenia, które posłużą do uogólnienia algorytmu.

Spostrzeżenie 2.1

Do przeprowadzenia eliminacji Gaussa używamy tylko współczynników układu równań, a kolejne etapy eliminacji Gaussa można interpretować jako ciąg odpowiednich przekształceń macierzy, który rozpocznie się od macierzy zawierającej wszystkie 3×4 współczynniki układu równań:

$$\begin{aligned} A^{(1)} := [A \ b] &= \begin{bmatrix} a_{11} & a_{12} & a_{13} & b_1 \\ a_{21} & a_{22} & a_{23} & b_2 \\ a_{31} & a_{32} & a_{33} & b_3 \end{bmatrix} \\ \rightarrow A^{(2)} &:= \begin{bmatrix} a_{11} & a_{12} & a_{13} & b_1 \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & b_2^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & b_3^{(2)} \end{bmatrix} \\ \rightarrow A^{(3)} &:= \begin{bmatrix} a_{11} & a_{12} & a_{13} & b_1 \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & b_2^{(2)} \\ 0 & 0 & a_{33}^{(3)} & b_3^{(3)} \end{bmatrix} \\ &= [U \ b^{(3)}]. \end{aligned} \quad (2.20)$$

Spostrzeżenie 2.2

Wykonanie przedstawionego algorytmu eliminacji Gaussa jest możliwe tylko przy założeniu, że w każdym etapie współczynnik $a_{ii}^{(i)}$, przez który dzielimy, obliczając mnożniki m_{ij} jest niezerowy. Współczynnik ten jest nazywany **elementem głównym**, a przedstawiona procedura – algorytmem **eliminacji Gaussa bez wyboru elementu głównego**. Jeżeli którykolwiek z elementów głównych byłby równy zeru, to algorytm eliminacji Gaussa bez wyboru elementu głównego kończy się niepowodzeniem.

Spostrzeżenie 2.3

Jeżeli element główny jest niezerowy, ale ma bardzo mały moduł, to sytuacja też nie jest korzystna. Dzielenie przez liczbę bliską zeru spowoduje bowiem znaczny wzrost błędu bezwzględnego wynikającego z błędów zaokrągleń dzielnika i pogorszy dokładność rozwiązania. Dodatkowo mały element główny jest zwykle obciążony dużym błędem względnym, co powoduje wzrost błędu w kolejnych etapach. Warto zatem doprowadzić do sytuacji, w której moduł elementu głównego byłby w miarę duży.

Spostrzeżenie 2.4

Można łatwo zmodyfikować algorytm eliminacji Gaussa, żeby zabezpieczyć się przed trudnościami zauważonymi wyżej. Jako że macierz współczynników A jest nieosobliwa, wśród elementów jej pierwszej kolumny musi być co najmniej jeden

niezerowy. Można z pierwszej kolumny $\begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix}$ wybrać element o największym mo-

dule i zmienić kolejność równań (czyli zamienić miejscami dwa odpowiednie wiersze w macierzy $A^{(1)}$), tak by był on elementem głównym. Z kolei, w drugim

etapie należy wybrać element o największym module spośród $\begin{bmatrix} a_{22}^{(2)} \\ a_{32}^{(2)} \end{bmatrix}$ i odpowied-

nio zmienić kolejność równań (czyli zamienić miejscami dwa odpowiednie wiersze w macierzy $A^{(2)}$).

Tak zmodyfikowany algorytm nazywa się **eliminacją Gaussa z częściowym (kolumnowym) wyborem elementu głównego**. W efekcie moduły wszystkich mnożników m_{ij} są nie większe od jedynki. Oczywiście, można by poszukiwać elementu o największym module w całej macierzy A w pierwszym etapie i w podmacierzy

$\begin{bmatrix} a_{22}^{(2)} & a_{23}^{(2)} \\ a_{32}^{(2)} & a_{33}^{(2)} \end{bmatrix}$ w etapie drugim. Taki algorytm jest nazywany **eliminacją Gaussa**

z pełnym wyborem elementu głównego. Jednak takie postępowanie wymagałoby nie tylko zamiany kolejności równań, ale także zmiany numeracji niewiadomych.

Jest to bardzo uciążliwe, a z reguły częściowy wybór elementu głównego zapewnia dostateczną dokładność rozwiązania.

Spostrzeżenie 2.5

W przypadku eliminacji Gaussa bez wyboru elementu głównego kolejne etapy algorytmu można przedstawić jako lewostronne mnożenie macierzy $A^{(i)}$ przez odpowiednią macierz współczynników:

$$A^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ -m_{21} & 1 & 0 \\ -m_{31} & 0 & 1 \end{bmatrix} A^{(1)}, \quad (2.21)$$

$$A^{(3)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -m_{32} & 1 \end{bmatrix} A^{(2)}. \quad (2.22)$$

Przy oznaczeniach „macierzy mnożników”:

$$L_1 := \begin{bmatrix} 1 & 0 & 0 \\ -m_{21} & 1 & 0 \\ -m_{31} & 0 & 1 \end{bmatrix} \text{ i } L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -m_{32} & 1 \end{bmatrix} \quad (2.23)$$

łatwo obliczyć, że:

$$L_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & 0 & 1 \end{bmatrix}, \quad L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & m_{32} & 1 \end{bmatrix}, \quad (2.24)$$

zaś iloczyn ich odwrotności jest równy

$$L = L_1^{-1} L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & m_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{bmatrix}. \quad (2.25)$$

W takim razie

$$A^{(3)} = L_2 L_1 A^{(1)} \Rightarrow A^{(1)} = [A \quad b] = \underline{L}^{-1} \underline{L}^{-1} A^{(3)} = L A^{(3)}, \quad (2.26)$$

$$[A \ b] = L[U \ b^{(3)}], \quad (2.27)$$

czyli

$$A = LU. \quad (2.28)$$

Przeprowadzone obliczenia doprowadziły do wyznaczenia macierzy trójkątnej dolnej L z jedynkami na głównej przekątnej i macierzy trójkątnej górnej U , takich, że ich iloczyn jest równy macierzy A . Takie przedstawienie macierzy kwadratowej A będziemy nazywać jej **rozkładem trójkątnym**.

Oczywiście do wyznaczenia rozkładu trójkątnego macierzy kwadratowej A nie ma potrzeby poddawać eliminacji Gaussa kolumny współczynników prawych stron b . Algorytm eliminacji Gaussa służy więc nie tylko do rozwiązywania układów równań liniowych, ale także do wyznaczania rozkładu trójkątnego macierzy kwadratowej. Zastosowania rozkładu trójkątnego zostaną przedstawione w podrozdziale 2.5.

Spostrzeżenie 2.6

Jeżeli w trakcie eliminacji Gaussa wykonamy przestawienie wierszy, to także otrzymamy macierz trójkątną dolną L z jedynkami na przekątnej i macierz trójkątną górną U . W tym przypadku jednak $LU = \tilde{A}$ gdzie macierz \tilde{A} powstaje z macierzy A przez takie same przestawienia wierszy, jak te wykonane w trakcie eliminacji Gaussa.

Operację zamiany kolejności wierszy można przedstawić w postaci lewostronnego mnożenia przez tak zwaną **macierz permutacji** P , która jest macierzą powstałą z macierzy jednostkowej przez odpowiednie przestawienie wierszy. Na przykład

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{21} & a_{22} & a_{23} \\ a_{11} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad (2.29)$$

czyli lewostronne mnożenie przez macierz $P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (powstałą z macierzy

jednostkowej przez zamianę pierwszego wiersza z drugim) realizuje zamianę pierwszego i drugiego wiersza. Macierz P z równości (2.29) można interpretować także jako efekt zamiany pierwszej i drugiej kolumny w macierzy jednostkowej. Przy takiej interpretacji prawostronne mnożenie

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{12} & a_{11} & a_{13} \\ a_{22} & a_{21} & a_{23} \\ a_{32} & a_{31} & a_{33} \end{bmatrix}, \quad (2.30)$$

realizuje zamianę pierwszej i drugiej kolumny macierzy A .

Każda macierz permutacji jest macierzą ortogonalną (patrz dodatek D2) zachodzi więc zależność

$$P^T P = I. \quad (2.31)$$

Jeżeli macierze P_1, P_2 reprezentują przestawienia wierszy w kolejnych etapach eliminacji Gaussa, to

$$U = L_2 P_2 L_1 P_1 A. \quad (2.32)$$

Z uwagi na (2.31) można zapisać

$$P_2 L_1 P_1 = P_2 L_1 (P_2^T P_2) P_1. \quad (2.33)$$

Oznaczmy przez

$$\hat{L}_1 := P_2 L_1 P_2^T. \quad (2.34)$$

Założmy, że w drugim etapie eliminacji nastąpiła zamiana drugiego wiersza z trzecim. Po podstawieniu macierzy L_1 z równania (2.23) otrzymujemy

$$\hat{L}_1 := P_2 L_1 P_2^T = \begin{bmatrix} 1 & 0 & 0 \\ -m_{31} & 1 & 0 \\ -m_{21} & 0 & 1 \end{bmatrix}. \quad (2.35)$$

Tak więc, zastosowanie zamiany wierszy w drugim etapie eliminacji Gaussa powoduje tylko adekwatną zamianę mnożników z etapu pierwszego, natomiast struktura macierzy \hat{L}_1 jest taka sama jak macierzy L_1 .

Z (2.35), (2.33) i (2.32) otrzymujemy

$$U = L_2 \hat{L}_1 P_2 P_1 A. \quad (2.36)$$

Ponadto, tak jak w (2.23), (2.24):

$$\begin{aligned} L &= (L_2 \hat{L}_1)^{-1} = \hat{L}_1^{-1} L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ m_{31} & 1 & 0 \\ m_{21} & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & m_{32} & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ m_{31} & 1 & 0 \\ m_{21} & m_{32} & 1 \end{bmatrix}. \end{aligned} \quad (2.37)$$

Jak widać, jeżeli była stosowana zamiana wierszy oraz w macierzy L uwzględniono odpowiednią zamianę mnożników, to

$$LU = P_2 P_1 A \quad (2.38)$$

i macierz $P = P_2 P_1$ reprezentuje wszystkie zamiany wierszy dokonane w trakcie eliminacji Gaussa. Poczynione spostrzeżenia z analizy przypadku trzech równań można uogólnić i sformułować algorytm eliminacji Gaussa z częściowym wyborem elementu głównego dla przypadku n równań z n niewiadomymi.

Algorytm eliminacji Gaussa z częściowym wyborem elementu głównego rozwiązujący układ n równań z n niewiadomymi

Rozwiązujemy układ równań $Ax = b$. Punktem wyjściowym algorytmu jest macierz $A^{(1)} := [A \quad b]$ o n wierszach i $n + 1$ kolumnach. W kroku k dysponujemy macierzą $A^{(k)}$, której elementy oznaczmy jednolicie $a_{ij}^{(k)}$, $i = 1, \dots, n$, $j = 1, \dots, n + 1$, i w której elementy pod przekątną, w pierwszych $k - 1$ kolumnach są zerowe:

$$a_{ij}^{(k)} = 0, \quad i > j, \quad j = 1, \dots, k - 1. \quad (2.39)$$

Postępowanie w k -tym kroku przekształcające $A^{(k)} \rightarrow A^{(k+1)}$:

- wybrać wiersz $r \geq k$, tak że

$$|a_{rk}^{(k)}| = \max_{k \leq i \leq n} |a_{ik}^{(k)}|, \quad (2.40)$$

- przestawić (zamienić miejscami) wiersze k i r , przestawienie zapamiętać,
- obliczyć mnożniki

$$m_{ik} = \frac{a_{ik}^{(k)}}{a_{rk}^{(k)}}, \quad i = k + 1, k + 2, \dots, n, \quad (2.41)$$

- obliczyć nowe elementy macierzy $A^{(k+1)}$

$$\begin{aligned} a_{ij}^{(k+1)} &= a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)}, \\ i &= k + 1, k + 2, \dots, n, \\ j &= k + 1, k + 2, \dots, n + 1, \end{aligned} \quad (2.42)$$

- wyzerować odpowiednie współczynniki w kolumnie k

$$a_{ik}^{(k+1)} = 0, \quad i = k + 1, k + 2, \dots, n. \quad (2.43)$$

Po wykonaniu $n - 1$ etapów algorytmu otrzymujemy macierz $A^{(n)} = [U \quad b^{(n)}]$:

Rozkład trójkątny macierzy w przypadku zamiany wierszy w trakcie eliminacji Gaussa

Jeżeli w trakcie eliminacji Gaussa wykonano przestawienia wierszy, to $LU = \tilde{A}$, gdzie macierz \tilde{A} powstała z macierzy A przez przestawienia tych samych wierszy.

Operację zamiany kolejności wierszy można przedstawić w postaci lewostronnego mnożenia przez macierz permutacji P , która powstała z macierzy jednostkowej przez odpowiednie przestawienie wierszy. Podobnie jak dla $n = 3$, każda macierz permutacji jest macierzą ortogonalną: $P^T P = I$, a prawostronne mnożenie przez macierz permutacji powoduje odpowiednią zamianę kolumn.

Jeżeli macierze P_1, \dots, P_{n-1} reprezentują przestawienia wierszy w kolejnych etapach eliminacji Gaussa, to można zapisać

$$U = L_{n-1}P_{n-1} \cdots L_2P_2L_1P_1A. \quad (2.49)$$

Postępowanie analogiczne do przypadku $n = 3$, prowadzi do równości

$$LU = P_{n-1} \cdots P_1A = PA, \quad (2.50)$$

gdzie w macierzy L uwzględniono odpowiednią zamianę mnożników zaś macierz $P = P_{n-1} \cdots P_1$ reprezentuje wszystkie zamiany wierszy dokonane w trakcie eliminacji Gaussa.

Częściowy wybór elementu głównego w eliminacji Gaussa nie wymaga operacji arytmetycznych, a jedynie porównywania liczb, które można zaimplementować bez przesyłania do pamięci. **Jest on stosowany rutynowo by zmniejszyć wpływ błędów zaokrągleń, choć jest konieczny jedynie w przypadku zerowego elementu głównego.** Odpowiedź na pytanie, kiedy można przeprowadzić eliminację Gaussa bez konieczności przestawiania wierszy daje twierdzenie 2.1.

Twierdzenie 2.1 (o rozkładzie trójkątnym macierzy – *Dahlquist, Björck, 1983*):

$$\text{Jeśli: } A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, A_k = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix}, \det A_k \neq 0,$$

$k = 1, 2, \dots, n - 1$, to istnieje dokładnie jeden rozkład $A = LU$, taki że L jest macierzą trójkątną dolną z jedynekami na przekątnej, a U jest macierzą trójkątną górną.

Rozkład taki jak w twierdzeniu 2.1 można uzyskać, przeprowadzając eliminację Gaussa bez przestawiania wierszy (bez wyboru elementu głównego).

2.3. Kontrola poprawności obliczeń w eliminacji Gaussa

W czasach kiedy duże układy równań liniowych rozwiązywano ręcznie, istotne było wyeliminowanie pomyłek i błędów człowieka. Do kontroli poprawności obliczeń stosowano chętnie następującą właściwość eliminacji Gaussa, opisaną w twierdzeniu 2.2.

Twierdzenie 2.2 (o sumach kontrolnych):

Jeżeli do macierzy $A^{(1)}$ dołączymy dodatkową kolumnę utworzoną przez zsumowanie wszystkich kolumn macierzy $A^{(1)}$ i otrzymaną w ten sposób macierz $A^{(1s)}$ poddamy eliminacji Gaussa, to właściwość polegająca na tym, że ostatnia kolumna jest sumą poprzednich jest zachowana na każdym etapie eliminacji Gaussa.

Dowód: Niech $s_i^{(1)} := \sum_{j=1}^{n+1} a_{ij}^{(1)}$, $i = 1, 2, 3, \dots, n$, $A^{(1s)} = [A^{(1)} \quad s]$.

W pierwszym etapie eliminacji Gaussa obliczamy:

$$a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i1} a_{1j}^{(1)}, \quad i = 2, 3, \dots, n, \quad j = 2, 3, \dots, n+2 \text{ oraz } s_i^{(2)} = s_i^{(1)} - m_{i1} s_1^{(1)}, \quad i = 2, 3, \dots, n. \text{ Wtedy } s_i^{(2)} = \sum_{j=1}^{n+1} a_{ij}^{(1)} - m_{i1} \sum_{j=1}^{n+1} a_{1j}^{(1)} = \sum_{j=1}^{n+1} [a_{ij}^{(1)} - m_{i1} a_{1j}^{(1)}] = \sum_{j=1}^{n+1} a_{ij}^{(2)}, \quad i = 2, 3, \dots, n.$$

W kolejnych etapach rozumowanie jest analogiczne.

Można więc kontrolować prawidłowość obliczeń, sprawdzając na każdym etapie, czy odpowiednie elementy kolumny sum kontrolnych są rzeczywiście równe sumie poprzedzających je elementów macierzy $A^{(i)}$. Zgodność nie gwarantuje poprawności obliczeń – można wszak pomylić się dwa razy tak, by efekty pomyłek się znosiły, ale niezgodność jest pewnym dowodem błędu.

Obecnie metoda ta (nazywana czasem metodą sum kontrolnych) może sygnalizować etap, na którym następuje utrata dokładności obliczeń na skutek dużego błędu zaokrążeń (może też być stosowana przez ćwiczących swe umiejętności studentów).

Przykład 2.1 (skuteczność wyboru elementu głównego)

Rozwiążmy układ trzech równań z trzema niewiadomymi

$$\begin{cases} 2,4759x_1 + 1,6235x_2 + 4,6231x_3 = 0,0647 \\ 1,4725x_1 + 0,9589x_2 - 1,3253x_3 = 1,0475 \\ 2,6951x_1 + 2,8965x_2 - 1,4794x_3 = -0,6789. \end{cases}$$

Obliczenia wykonujemy metodą eliminacji Gaussa bez wyboru elementu głównego. Obliczenia prowadzone są zgodnie z algorytmem (2.39-2.43). Wynik każdej z operacji algorytmu (obliczenie mnożnika, obliczenie współczynnika równania) jest zaokrąglany do czterech miejsc po przecinku. Oznacza to, że także w wynikach zaakceptujemy rozbieżności między wartościami obliczonymi a dokładnymi na czwartym miejscu po przecinku. W celu sprawdzenia poprawności obliczeń dodajemy kolumnę sum kontrolnych.

$$A^{(1s)} = [A \ b \ s] = \begin{bmatrix} 2,4759 & 1,6235 & 4,6231 & 0,0647 & 8,7872 \\ 1,4725 & 0,9589 & -1,3253 & 1,0475 & 2,1536 \\ 2,6951 & 2,8965 & -1,4794 & -0,6789 & 3,4333 \end{bmatrix}.$$

W pierwszym etapie eliminacji obliczamy mnożniki

$$m_{21} = \frac{1,4725}{2,4759} = 0,5947, \quad m_{31} = \frac{2,6951}{2,4759} = 1,0885$$

i otrzymujemy:

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -0,5947 & 1 & 0 \\ -1,0885 & 0 & 1 \end{bmatrix},$$

$$A^{(2s)} = \begin{bmatrix} 2,4759 & 1,6235 & 4,6231 & 0,0647 & 8,7872 \\ 0 & -0,0066 & -4,0747 & 1,0090 & -3,0721 \\ 0 & 1,1293 & -6,5116 & -0,7493 & -6,1316 \end{bmatrix}.$$

Nie ma potrzeby obliczać wartości elementów $a_{21}^{(2)}, a_{31}^{(2)}$, które powinny być zerami. Sumując pierwsze cztery kolumny macierzy $A^{(2s)}$ otrzymujemy

$$SUM_1 = \begin{bmatrix} * \\ -3,0723 \\ -6,1316 \end{bmatrix},$$

co daje akceptowalną niezgodność z kolumną sumy kontrolnej na czwartej pozycji po przecinku.

W drugim etapie eliminacji Gaussa elementem głównym będzie $-0,0066$. Jest to liczba o module niewiele większym od stosowanej dokładności obliczeń. Wyznaczamy mnożnik $m_{32} = \frac{1,1293}{-0,0066} = -171,1061$ i otrzymujemy:

$$L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 171,1061 & 1 \end{bmatrix} \Rightarrow L = \begin{bmatrix} 1 & 0 & 0 \\ 0,5947 & 1 & 0 \\ 1,0885 & -171,1061 & 1 \end{bmatrix},$$

$$A^{(3s)} = \begin{bmatrix} 2,4759 & 1,6235 & 4,6231 & 0,0647 & 8,7872 \\ 0 & -0,0066 & -4,0747 & 1,0090 & -3,0721 \\ 0 & 0 & -703,7176 & 171,8968 & -531,7866 \end{bmatrix},$$

a sumując cztery pierwsze elementy ostatniego wiersza w $A^{(3s)}$, obliczamy:

$$SUM_2 = \begin{bmatrix} * \\ * \\ -531, \mathbf{8208} \end{bmatrix}.$$

Liczby $-531, \mathbf{7866}$ i $-531, \mathbf{8208}$ są zgodne dopiero po zaokrągleniu do pierwszego miejsca po przecinku, co wskazuje na pogorszenie dokładności na tym etapie obliczeń.

Rozwiązując metodą podstawienia trójkątny układ równań

$$Ux = b^{(3)},$$

gdzie:

$$U = \begin{bmatrix} 2,4759 & 1,6235 & 4,6231 \\ 0 & -0,0066 & -4,0748 \\ 0 & 0 & -703,7176 \end{bmatrix}, b^{(3)} = \begin{bmatrix} 0,0647 \\ 1,0090 \\ 171,8968 \end{bmatrix},$$

otrzymujemy

$$\begin{cases} x_1 = +1,8286 \\ x_2 = -2,0531. \\ x_3 = -0,2443 \end{cases}$$

Dokonując sprawdzenia dokładności otrzymanego rozwiązania przez obliczenie tak zwanej **reszty** $r = Ax - b$ dla otrzymanych x_1, x_2, x_3 , otrzymujemy

$$r = \begin{bmatrix} 2,4759 & 1,6235 & 4,6231 \\ 1,4725 & 0,9589 & -1,3253 \\ 2,6951 & 2,8965 & -1,4794 \end{bmatrix} \cdot \begin{bmatrix} 1,8286 \\ -2,0531 \\ -0,2443 \end{bmatrix} - \begin{bmatrix} 0,0647 \\ 1,0475 \\ -0,6789 \end{bmatrix} = \begin{bmatrix} 0,0001 \\ 0,0002 \\ -0,0218 \end{bmatrix}.$$

Niezerowe cyfry na drugim miejscu po przecinku świadczą o utracie dokładności obliczeń. Wykonano niewiele operacji arytmetycznych, przyczyną błędu nie jest więc kumulacja małych zaokrągleń w długim łańcuchu operacji, a „wzmocnienie” błędu zaokrągleń przez wymnożenie przez stosunkowo duży mnożnik $m_{32} = -171,1061$, który był efektem małego elementu głównego w drugim etapie eliminacji. Efekt małego elementu głównego byłby jeszcze wyraźniej widoczny, gdyby była stosowana arytmetyka zmiennoprzecinkowa z pięcioma cyframi mantysy, zamiast reprezentacji stałopozycyjnej z czterema cyframi po przecinku. Wtedy, np., mnożnik m_{32} byłby dodatkowo zaokrąglany: $m_{32} = -171,1061 \approx -171,11$.

Mały element główny pojawia się w ostatnim etapie eliminacji Gaussa, więc w mniejszym stopniu dotyczy rozkładu trójkątnego macierzy A . Gdybyśmy przeliczyli różnicę między macierzą współczynników A i iloczynem macierzy LU , to otrzymujemy

$$A - LU = \begin{bmatrix} 0,0000 & 0,0000 & 0,0000 \\ 0,0001 & 0,0000 & 0,0000 \\ 0,0001 & 0,0000 & -0,0001 \end{bmatrix}.$$

Obliczmy jeszcze raz rozwiązanie układu równań, stosując te same zasady zaokrągleń, tym jednak razem zgodnie z algorytmem eliminacji Gaussa z częściowym wyborem elementu głównego.

Analizując wartości bezwzględne elementów z pierwszej kolumny macierzy współczynników, dochodzimy do wniosku, że przed pierwszym krokiem eliminacji musimy zamienić miejscami trzeci z pierwszym wierszem. Macierz permutacji przyjmuje postać

$$P_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$

otrzymujemy zatem

$$P_1 A^{(1s)} = P_1 [A \ b \ s] = \begin{bmatrix} 2,6951 & 2,8965 & -1,4794 & -0,6789 & 3,4333 \\ 1,4725 & 0,9589 & -1,3253 & 1,0475 & 2,1536 \\ 2,4759 & 1,6235 & 4,6231 & 0,0647 & 8,7872 \end{bmatrix}.$$

W pierwszym kroku eliminacji, otrzymujemy mnożniki

$$m_{21} = \frac{1,4725}{2,6951} = 0,5464, \quad m_{31} = \frac{2,4759}{2,6951} = 0,9187$$

i stąd

$$\hat{L}_1 = \begin{bmatrix} 1 & 0 & 0 \\ -0,9187 & 1 & 0 \\ -0,5464 & 0 & 1 \end{bmatrix}$$

$$i A^{(2s)} = \begin{bmatrix} 2,6951 & 2,8965 & -1,4794 & -0,6789 & 3,4333 \\ 0 & -0,6237 & -0,5170 & 1,4185 & 0,2776 \\ 0 & -1,0375 & 5,9822 & 0,6884 & 5,6330 \end{bmatrix}.$$

Sumując pierwsze cztery kolumny macierzy $A^{(2s)}$, otrzymujemy

$$SUM_2 = \begin{bmatrix} * \\ 0,2778 \\ 5,6331 \end{bmatrix},$$

co daje dopuszczalną niezgodność z kolumną sumy kontrolnej na ostatniej pozycji po przecinku. Przed drugim etapem eliminacji Gaussa konieczna jest kolejna zamiana wierszy. Zamieniamy wiersze drugi i trzeci. Macierz permutacji przyjmuje więc postać

$$P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Wykonując odpowiednie przestawienie, otrzymujemy

$$P_2 A^{(2s)} = \begin{bmatrix} 2,6951 & 2,8965 & -1,4794 & -0,6789 & 3,4333 \\ 0 & -1,0375 & 5,9822 & 0,6884 & 5,6330 \\ 0 & -0,6237 & -0,5170 & 1,4185 & 0,2776 \end{bmatrix}.$$

Pamiętamy też o zamianie mnożników związanych z przestawionymi wierszami:

$$m_{21} = 0,9187, \quad m_{31} = 0,5464.$$

W drugim kroku eliminacji Gaussa, otrzymujemy:

$$L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -0,6012 & 1 \end{bmatrix} \Rightarrow L = \begin{bmatrix} 1 & 0 & 0 \\ 0,9187 & 1 & 0 \\ 0,5464 & 0,6012 & 1 \end{bmatrix},$$

$$A^{(3s)} = \begin{bmatrix} 2,6951 & 2,8965 & -1,4794 & -0,6789 & 3,4333 \\ 0 & -1,0375 & 5,9822 & 0,6884 & 5,6330 \\ 0 & 0 & -4,1135 & 1,0046 & -3,1090 \end{bmatrix}.$$

Sumując po raz kolejny pierwsze cztery elementy ostatniego wiersza macierzy $A^{(3s)}$, mamy

$$SUM = \begin{bmatrix} * \\ * \\ -3,1089 \end{bmatrix},$$

co (w przeciwieństwie do wersji metody bez wyboru elementu głównego) nie wskazuje na utratę dokładności. Rozwiązując metodą podstawienia trójkątny układ równań

$$Ux = b^{(3)},$$

gdzie:

$$U = \begin{bmatrix} 2,6951 & 2,8965 & -1,4794 \\ 0 & -1,0375 & 5,9822 \\ 0 & 0 & -4,1135 \end{bmatrix}, \quad b^{(3)} = \begin{bmatrix} -0,6789 \\ 0,6884 \\ 1,0046 \end{bmatrix},$$

otrzymujemy

$$\begin{cases} x_1 = +1,8405 \\ x_2 = -2,0716 \\ x_3 = -0,2442 \end{cases}$$

Po obliczeniu reszty, otrzymujemy

$$r = \begin{bmatrix} 2,4759 & 1,6235 & 4,6231 \\ 1,4725 & 0,9589 & -1,3253 \\ 2,6951 & 2,8965 & -1,4794 \end{bmatrix} \cdot \begin{bmatrix} 1,8405 \\ -2,0716 \\ -0,2442 \end{bmatrix} - \begin{bmatrix} 0,0647 \\ 1,0472 \\ -0,6788 \end{bmatrix} = \begin{bmatrix} 0,0001 \\ -0,0002 \\ 0,0000 \end{bmatrix}.$$

Otrzymana różnica ma niezerowe cyfry jedynie na ostatnim miejscu po przecinku potwierdza korzyści płynące z wyboru elementu głównego.

2.4. Złożoność obliczeniowa eliminacji Gaussa

Wybierając algorytm, który ma realizować określone zadanie, kierujemy się zwykle analizą jego wydajności (kosztu obliczeniowego), czy też błędów które może generować. Spróbujmy zatem porównać czasową złożoność obliczeniową omawianych algorytmów. Do jej oszacowania dla algorytmów związanych z rozwiązywaniem układów równań liniowych będą przydatne dwa wzory

$$\sum_{k=1}^{n-1} k = \frac{n(n-1)}{2}, \quad \sum_{k=1}^{n-1} k^2 = \frac{n(n-1)(2n-1)}{6}. \quad (2.51)$$

Rozwiązanie trójkątnego układu równań:

Dla wyznaczenia k -tej zmiennej (w kolejności obliczania, a nie zgodnie z numeracją w równaniu) $k = 1, \dots, n$ wykonujemy:

- podstawienie $k - 1$ zmiennych obliczonych poprzednio ($k - 1$ mnożeń),
- $k - 1$ dodawań (składniki z podstawionymi zmiennymi i współczynniki po prawej stronie),
- i jedno dzielenie.

Oczywiście mogą wystąpić sytuacje szczególne, np. kiedy współczynniki są zerowe lub równe jeden. Dlatego też łączną liczbę mnożeń i dodawań można oszacować jako

$$N_{T*} \approx \sum_{k=1}^n k = \frac{n(n+1)}{2} = \frac{1}{2}n^2 + \frac{1}{2}n, \quad (2.52)$$

$$N_{T+} \approx \sum_{k=1}^n (k-1) = \frac{n(n+1)}{2} - n = \frac{1}{2}n^2 - \frac{1}{2}n.$$

Tak więc rozwiązanie trójkątnego układu równań należy do klasy zadań rzędu $O(n^2)$.

Wybór elementu głównego:

Częściowy wybór elementu głównego można potraktować jako n -krotne przeszukiwanie zbioru co najwyżej n liczb. Złożoność obliczeniowa takiego zadania jest rzędu $O(n^2)$. Należy jednak pamiętać, że operacje wykonywane w procesie porównywania liczb są znacznie szybsze od operacji arytmetycznych. Pełny wybór elementu głównego oznaczałby n -krotne przeszukiwanie zbioru co najwyżej n^2 liczb, czyli zadanie o złożoności $O(n^3)$. Poniesienie tego kosztu nie jest uzasadnione i częściowy wybór elementu głównego jest powszechnie akceptowanym standardem.

Eliminacja Gaussa:

W przedstawionych algorytmach eliminację Gaussa prowadzi się na macierzy o liczbie kolumn:

- $p = n$ (w celu otrzymania rozkładu trójkątnego macierzy),
- $p = n + 1$ (rozwiązując układ równań przez przekształcanie współczynników lewej i prawej strony $[A \ b]$),
- $p = n + 2$ (jeżeli dodatkowo dodamy kolumnę sum kontrolnych $[A \ b \ s]$).

W k -tym etapie ($k = 1, \dots, n - 1$):

- obliczamy $n - k$ mnożników ($n - k$ dzielen),
- mnożymy $p - k$ elementów k -tego wiersza przez $n - k$ mnożników ($(p - k)(n - k)$ mnożeń),
- wykonujemy $p - k$ odejmowań w każdym z $n - k$ równań ($(p - k)(n - k)$ odejmowań),

razem $(p - k + 1)(n - k)$ mnożeń i dzielen oraz $(p - k)(n - k)$ odejmowań.

Łączną liczbę mnożeń i dzielen dla $p = n$ można obliczyć jako

$$\begin{aligned}
 N_{GE(p=n)*} &= \sum_{k=1}^{n-1} (n-k+1)(n-k) \\
 &= \sum_{k=1}^{n-1} (n^2 + n - (2n+1)k + k^2) \\
 &= (n^2 + n)(n-1) - (2n+1) \frac{n(n-1)}{2} \\
 &\quad + \frac{n(n-1)(2n-1)}{6} \\
 &= n^3 - n - \frac{1}{2}(2n^3 - n^2 - n) \\
 &\quad + \frac{1}{6}(2n^3 - 3n^2 + n) = \frac{1}{3}n^3 - \frac{1}{3}n.
 \end{aligned} \tag{2.53}$$

Dodawanie jest o

$$\sum_{k=1}^{n-1} (n-k) = n(n-1) - \frac{n(n-1)}{2} = \frac{1}{2}n^2 - \frac{1}{2}n \tag{2.54}$$

mniej niż mnożeń, czyli jest ich

$$N_{GE(p=n)+} = \frac{1}{3}n^3 - \frac{1}{2}n^2 + \frac{1}{6}n. \tag{2.55}$$

Zwiększenie p o jeden spowoduje wzrost liczby i mnożeń i dodawań o

$$\Delta N = \sum_{k=1}^{n-1} (n-k) = \frac{1}{2}n^2 - \frac{1}{2}n. \tag{2.56}$$

Tak więc, wybierając standardową metodę rozwiązania układu równań liniowych bez sum kontrolnych ($p = n + 1$) wykonamy łącznie przy eliminacji Gaussa i rozwiązaniu trójkątnego układu równań

$$\begin{aligned}
 N_{GE(p=n+1)T*} &= \left(\left(\frac{1}{3}n^3 - \frac{1}{3}n \right) + \left(\frac{1}{2}n^2 - \frac{1}{2}n \right) \right) \\
 &\quad + \left(\frac{1}{2}n^2 + \frac{1}{2}n \right) = \frac{1}{3}n^3 + n^2 - \frac{1}{3}n
 \end{aligned} \tag{2.57}$$

mnożeń oraz

$$\begin{aligned}
 N_{GE(p=n+1)T+} &= \left(\left(\frac{1}{3}n^3 - \frac{1}{2}n^2 + \frac{1}{6}n \right) + \left(\frac{1}{2}n^2 - \frac{1}{2}n \right) \right) \\
 &\quad + \left(\frac{1}{2}n^2 - \frac{1}{2}n \right) = \frac{1}{3}n^3 + \frac{1}{2}n^2 - \frac{5}{6}n
 \end{aligned} \tag{2.58}$$

dodawań.

Oczywiście w przypadku dużych n decydujące są składniki z najwyższymi potęgami.

Alternatywną metodą rozwiązywania układu równań liniowych jest wykorzystanie przeprowadzanego najpierw rozkładu trójkątnego macierzy współczynników A . Otrzymujemy wtedy macierze P, L, U , takie że

$$PA = LU, \quad (2.59)$$

a następnie zamiast równania $Ax = b$ rozwiązujemy równanie

$$LUx = \tilde{b}, \quad (2.60)$$

gdzie kolumna $\tilde{b} = Pb$ powstaje z kolumny b po takich samych zamianach wierszy jak przeprowadzane w trakcie eliminacji Gaussa wykonane dla otrzymania rozkładu trójkątnego. Technicznie rozwiązanie tego równania jest realizowane przez rozwiązanie dwóch trójkątnych układów równań:

$$Ly = \tilde{b}, \quad (2.61)$$

który rozwiązuje się przez podstawianie w przód (od pierwszego równania) oraz

$$Ux = y, \quad (2.62)$$

który rozwiązuje się przez podstawianie wstecz (od ostatniego równania). Macierz L ma jedynki na głównej przekątnej, więc przy rozwiązaniu układu (2.61) nie wykonamy dzieleni. Łącznie, przy takim postępowaniu wykonamy

$$\begin{aligned} N_{GE(p=n)2T*} &= \left(\frac{1}{3}n^3 - \frac{1}{3}n\right) + \left(\frac{1}{2}n^2 - \frac{1}{2}n\right) + \left(\frac{1}{2}n^2 + \frac{1}{2}n\right) \\ &= \frac{1}{3}n^3 + n^2 - \frac{1}{3}n \end{aligned} \quad (2.63)$$

mnożeń oraz

$$\begin{aligned} N_{GE(p=n)2T+} &= \left(\frac{1}{3}n^3 - \frac{1}{2}n^2 + \frac{1}{6}n\right) + 2\left(\frac{1}{2}n^2 - \frac{1}{2}n\right) \\ &= \frac{1}{3}n^3 + \frac{1}{2}n^2 - \frac{5}{6}n \end{aligned} \quad (2.64)$$

dodawań.

Przedstawione wyliczenia można podsumować w następujący sposób:

Złożoność obliczeniowa eliminacji Gaussa to $O(n^3)$ operacji zmiennopozycyjnych. Dodatkowa kolumna poddawana eliminacji Gaussa to dodatkowe $O(n^2)$ operacji.

Efektywna implementacja eliminacji Gaussa nie jest prostym problemem. Istotna jest nie tylko liczba operacji arytmetycznych, ale także koszt pobrania i zapisu danych w pamięci. Na szczęście dostępne pakiety obliczeniowe oferują dobre rozwiązania i wystarczy w świadomy sposób z nich korzystać. Zwykle sprawdza się najpierw strukturę macierzy współczynników i stosuje się algorytmy dostosowane do szczególnych przypadków. Na przykład instrukcja $x = A \setminus b$ w Matlabie oznacza, że zostanie kolejno sprawdzone, czy macierz A jest prostokątna, trójkątna, trójkątna z przestawionymi wierszami, symetryczna, dodatnio określona, w postaci Hessenberga, a dopiero gdy żaden tych szczególnych przypadków nie występuje zostanie wyznaczony rozkład trójkątny metodą eliminacji Gaussa i rozwiązane dwa trójkątne układy równań. Podobnie działa funkcja $\text{linsolve}(A, b)$.

Przykład 2.2

W tabeli 2.1 podano czasy wykonania instrukcji $x = A \setminus b$ oraz $x = \text{linsolve}(A, b)$, które w Matlabie odpowiadają metodzie eliminacji Gaussa i wykorzystaniu rozkładu trójkątnego. Macierz współczynników A była losową, a b kolumną jedynek:

$$A = \text{rand}(n, n); \quad b = \text{ones}(n, 1);$$

Czasy podane w tabeli 2.1 są faktycznie proporcjonalne do n^3 . W obu metodach stosowany jest ten sam algorytm podstawowy, więc różnice między pierwszym a drugim wierszem wynikają z innych aspektów niż złożoność obliczeniowa algorytmu.

Tabela 2.1. Czas (sekundy) rozwiązania układu równań liniowych w przykładowym uruchomieniu programu Matlab

n	100	500	1000	5000	10000
$x = A \setminus b$	0,0009	0,0089	0,0352	1,7591	12,1740
$x = \text{linsolve}(A, b)$	0,0003	0,0089	0,0378	1,7994	13,1682

2.5. Zastosowania rozkładu trójkątnego

Rozkład trójkątny macierzy ma szereg zastosowań w różnych problemach numerycznych. W tym podrozdziale podano najważniejsze z nich.

Obliczanie wyznacznika

Wykorzystanie rozkładu trójkątnego jest podstawową metodą obliczania wyznacznika macierzy kwadratowej. Dla dowolnej macierzy trójkątnej (z definicji wyznacznika – patrz dodatek D2) wyznacznik jest równy iloczynowi elementów na jej głównej przekątnej. Jak wiadomo zamiana wierszy w macierzy powoduje jedynie zmianę znaku wyznacznika. Stąd jeśli podczas rozkładu LU nastąpiła zamiana wierszy możemy zapisać

$$\det(A) = (-1)^s \det(PA), \quad (2.65)$$

gdzie s jest liczbą wykonanych zamian wierszy. Uwzględniając równanie (2.50) i korzystając z twierdzenia Cauchy'ego (patrz dodatek D2.2), otrzymujemy

$$\det(A) = (-1)^s \det(LU) = (-1)^s \det(L)\det(U), \quad (2.66)$$

Macierze L i U są trójkątne, dodatkowo w przypadku macierzy L wszystkie elementy na diagonalu są równe jedności, więc ostatecznie wystarczy wymnożyć elementy na głównej przekątnej macierzy U :

$$\det(A) = (-1)^s u_{11} u_{22} \cdots u_{nn}. \quad (2.67)$$

Rozwiązanie wielu układów równań o tej samej macierzy współczynników

Czasami zachodzi potrzeba rozwiązania wielu układów równań różniących się tylko prawą stroną

$$Ax_1 = b_1, Ax_2 = b_2, \dots, Ax_k = b_k, \quad (2.68)$$

co można zapisać łącznie

$$A[x_1 \ x_2 \ \dots \ x_k] = [b_1 \ b_2 \ \dots \ b_k]. \quad (2.69)$$

Równanie (2.68) można rozwiązać, wykonując eliminację Gaussa na macierzy $[A \ b_1 \ b_2 \ \dots \ b_k]$ i wyznaczając rozwiązania k układów trójkątnych. Można też wyznaczyć rozkład trójkątny macierzy A , to jest macierze P, L, U ($PA = LU$), a następnie rozwiązać trójkątne układy równań

$$Ly_1 = Pb_1, Ux_1 = y_1, Ly_2 = Pb_2, Ux_2 = y_2, \dots, Ly_k = Pb_k, Ux_k = y_k, \quad (2.70)$$

czyli równania:

$$\begin{aligned} L[y_1 \ y_2 \ \dots \ y_k] &= P[b_1 \ b_2 \ \dots \ b_k], \\ U[x_1 \ x_2 \ \dots \ x_k] &= [y_1 \ y_2 \ \dots \ y_k]. \end{aligned} \quad (2.71)$$

Odwracanie macierzy

Sposób postępowania, przedstawiony wyżej, może być przydatny przy obliczaniu macierzy odwrotnej. Macierz odwrotna do macierzy A , to taka macierz X , że

$$AX = I. \quad (2.72)$$

Jeżeli przez x_1, x_2, \dots, x_n oznaczymy kolumny macierzy odwrotnej X , a przez e_1, e_2, \dots, e_n kolumny macierzy jednostkowej, to macierzowe równanie (2.71) można zapisać (pisząc równości dla kolejnych kolumn prawej i lewej strony) w postaci n układów równań liniowych

$$Ax_1 = e_1, Ax_2 = e_2, \dots, Ax_n = e_n \quad (2.73)$$

i obliczyć poszczególne kolumny macierzy odwrotnej, rozwiązując trójkątne układy równań:

$$Ly_1 = Pe_1, Ux_1 = y_1, Ly_2 = Pe_2, Ux_2 = y_2, \dots, Ly_n = Pb_n, Ux_n = y_n. \quad (2.74)$$

Najbardziej efektywnym sposobem obliczania macierzy odwrotnej będzie bezpośrednie wykorzystanie równości uzyskanej z rozkładu trójkątnego. Jeśli $PA = LU$, to

$$(PA)^{-1} = (LU)^{-1} \Rightarrow A^{-1}P^{-1} = U^{-1}L^{-1} \Rightarrow A^{-1} = U^{-1}L^{-1}P. \quad (2.75)$$

Prawostronne mnożenie przez macierz P realizuje zamiany kolumn analogiczne do zamian wierszy wykonanych przy wyznaczaniu rozkładu trójkątnego, a macierze U i L łatwo odwrócić, posługując się otwartymi wzorami:

$$z_{ij} = (L^{-1})_{ij} \Rightarrow z_{ij} = \frac{1}{l_{ii}} \left[\delta_{ij} - \sum_{k=j}^{i-1} l_{ik}z_{kj} \right] \quad i = j, j+1, \dots, n, \quad (2.76)$$

$$q_{ij} = (U^{-1})_{ij} \Rightarrow q_{ij} = \frac{1}{u_{ii}} \left[\delta_{ij} - \sum_{k=i+1}^j u_{ik}q_{kj} \right] \quad i = j, j-1, \dots, 1, \quad (2.77)$$

w których $\delta_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$.

Z powyższych wzorów wynika, że odwrotność macierzy trójkątnej górnej jest macierzą trójkątną górną i analogicznie odwrotność macierzy trójkątnej dolnej jest macierzą trójkątną dolną. Ponadto dla elementów na diagonalu zachodzi:

$$(L^{-1})_{ii} = \frac{1}{l_{ii}} \quad i = 1, 2, \dots, n, \quad (2.78)$$

$$(U^{-1})_{ij} = \frac{1}{u_{ij}} \quad i = 1, 2, \dots, n. \quad (2.79)$$

Przykład 2.3

Skoro to algorytm eliminacji Gaussa służy do wyznaczenia macierzy odwrotnej nie należy rozwiązywać układu równań liniowych $Ax = b$ obliczając $x = A^{-1}b$. W tabeli 2.2 zestawiono czasy wykonania instrukcji $x = A \setminus b$ oraz $x = \text{inv}(A) * b$ w Matlabie. Tym razem jednak macierz współczynników A była losową macierzą ortogonalną (co zabezpieczało przed przypadkowym wyborem macierzy źle uwarunkowanej), zaś b kolumną jedynek:

```
A = orth(rand(n,n)); b = ones(n,1); .
```

Czas potrzebny na zastosowanie wzoru $x = A^{-1}b$ jest kilkukrotnie dłuższy od czasu koniecznego do przeprowadzenia eliminacji Gaussa.

Tabela 2.2. Czas (sekundy) rozwiązania układu równań liniowych w przykładowym uruchomieniu programu Matlab

n	100	500	1000	5000
$x = A \setminus b$	0,0004	0,0054	0,0260	1,7743
$x = \text{inv}(A) * b$	0,0007	0,0118	0,0632	5,1373

Przykład 2.4

W tabeli 2.3 podano czasy wykonania instrukcji $X = \text{inv}(A)$ oraz $X = A \setminus \text{eye}(\text{size}(A));$ i $X = \text{linsolve}(A, \text{eye}(\text{size}(A)));$ w Matlabie. Macierz współczynników A była losową macierzą ortogonalną: $A = \text{orth}(\text{rand}(n,n)); .$

Czasy wykonania wszystkich trzech algorytmów są podobne i kilkukrotnie dłuższe od czasu potrzebnego na wykonanie pojedynczej eliminacji Gaussa (tabela 2.2).

Tabela 2.3. Czas (sekundy) procedur obliczających odwrotność macierzy w przykładowym uruchomieniu programu Matlab

n	100	500	1000	5000
$X = \text{inv}(A)$	0,0007	0,0074	0,0521	6,6376
$X = A \setminus \text{eye}(\text{size}(A));$	0,0008	0,0124	0,0715	9,2357
$X = \text{linsolve}(A, \text{eye}(\text{size}(A)));$	0,0013	0,0111	0,0720	8,1114

2.6. Błędy rozwiązania układu równań liniowych metodą eliminacji Gaussa

Rozwiązywanie układu równań liniowych drogą eliminacji Gaussa nie jest obarczone błędem metody. Jedyne błędy zaokrągleń mogą spowodować, że otrzymane rozwiązanie będzie odbiegać od dokładnego. Niestety, czasem wpływ błędów zaokrągleń jest znaczny i trudno go uniknąć.

Żeby oszacować wpływ zmiany (błędu) współczynników na otrzymane rozwiązanie, trzeba użyć narzędzi pozwalających na ocenę, czy wektor lub macierz są „duże” lub „małe”. W tym celu zostaną zdefiniowane odpowiednie normy macierzy.

Normy wektorów i macierzy

Normą wektora $x \in R^n$ jest nazywane dowolne odwzorowanie $x \rightarrow \|x\| \in R_+$ spełniające warunki:

$$\begin{aligned} \|x\| &\geq 0 \text{ i } (\|x\| = 0 \Leftrightarrow x = 0), \\ \forall c \in R \quad \|cx\| &= |c| \cdot \|x\|, \\ \|x_1 + x_2\| &\leq \|x_1\| + \|x_2\|. \end{aligned} \tag{2.80}$$

Powszechnie stosuje się normę euklidesową (inaczej normę L_2)

$$\|x\|_2 = \left(\sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}}, \tag{2.81}$$

ale można posługiwać się też normą L_p dla $p \geq 1$:

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}, \tag{2.82}$$

normą L_1 (normą Manhattan):

$$\|x\|_1 = \sum_{i=1}^n |x_i|, \tag{2.83}$$

normą L_∞ :

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|. \tag{2.84}$$

Jeżeli zostanie już ustalona norma wektorowa $\|x\|$, to norma macierzowa, która będzie stosowana w tym rozdziale, zostanie zdefiniowana w specjalny sposób.

Definicja 2.1 (zgodnej normy macierzy)

Normą macierzową indukowaną przez normę wektorową (zgodną z normą wektorową) będzie nazywana norma zdefiniowana jako:

$$\|A\| := \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}. \quad (2.85)$$

Po prawej stronie we wzorze definicyjnym znajdują się normy dwóch wektorów: x i Ax , więc norma macierzy A jest poprawnie określona.

Bezpośrednio z definicji wynika, że gdy $\|A\| < \infty$, to

$$\|Ax\| \leq \|A\| \cdot \|x\| \quad (2.86)$$

dla dowolnego wektora x . Także dla dwóch kwadratowych macierzy A, B zachodzi

$$\|AB\| \leq \|A\| \cdot \|B\|. \quad (2.87)$$

Z normą wektorową L_1 jest zgodna norma macierzowa

$$\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|, \quad (2.88)$$

z normą L_∞ norma

$$\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|, \quad (2.89)$$

a z normą euklidesową norma obliczana jako największa wartość osobliwa² macierzy

$$\|A\|_2 = \max_i \sigma_{max}. \quad (2.90)$$

Norma $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$ (tak zwana norma Frobeniusa) nie jest zgodna (w sensie (2.85)) z euklidesową (ani żadną inną) normą wektorową!

² Nieosobliwą macierz kwadratową można przedstawić w postaci $A = UDV^T$, gdzie ortogonalna macierz U jest zbudowana z wektorów własnych macierzy $A^T A$, ortogonalna macierz V – z wektorów własnych macierzy AA^T , a diagonalna macierz D zawiera na przekątnej dodatnie liczby nazywane wartościami szczególnymi (osobliwymi) macierzy A . Jeżeli jakakolwiek wartość szczególna (osobliwa) jest równa zero, to A jest osobliwa. Wartości własne i szczególne zostały dokładniej opisane w rozdziale 8.

Współczynnik uwarunkowania macierzy

Rozwiązujemy układ równań $Ax = b$. Jeżeli zamiast wektora współczynników b zastosujemy zaburzony (np. na skutek błędów zaokrągleń) wektor $b + \delta b$, to rozwiązanie zmieni się na $x + \delta x$:

$$A(x + \delta x) = b + \delta b. \tag{2.91}$$

Wtedy

$$A\delta x = \delta b \implies \delta x = A^{-1}\delta b. \tag{2.92}$$

Z nierówności

$$\|\delta x\| \leq \|A^{-1}\| \cdot \|\delta b\| \text{ i } \|b\| \leq \|A\| \cdot \|x\|, \tag{2.93}$$

które wynikają bezpośrednio z definicji zgodnej normy macierzowej, mamy

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\| \cdot \|\delta b\|}{\left(\frac{\|b\|}{\|A\|}\right)} = \|A\| \|A^{-1}\| \frac{\|\delta b\|}{\|b\|}. \tag{2.94}$$

Po lewej stronie tej nierówności mamy wyrażenie $\frac{\|\delta x\|}{\|x\|}$, które jest odpowiednikiem względnego błędu rozwiązania, po prawej $\frac{\|\delta b\|}{\|b\|}$ jest odpowiednikiem względnego błędu danych wejściowych.

Definicja 2.2 (wskaźnika uwarunkowania macierzy)

Współczynnik

$$\text{cond}(A) := \|A\| \|A^{-1}\| \tag{2.95}$$

jest nazywany **współczynnikiem (wskaźnikiem) uwarunkowania macierzy A** .

Wskaźnik uwarunkowania decyduje o tym, jak bardzo błąd danych wejściowych jest „wzmocniany”, przenosząc się na błąd wyniku.

Każda indukowana norma macierzy jednostkowej jest równa 1, więc z (2.87) wynika, że $\text{cond}(A) \geq 1$.

Jeżeli wskaźnik uwarunkowania będzie duży, to błąd rozwiązania będzie duży, nawet jeśli błąd danych wejściowych jest mały!

Jeżeli zmiana rozwiązania na $x + \delta x$ jest wynikiem zmiany współczynników lewej strony na $A + \delta A$, to można wyprowadzić

$$\begin{aligned} \delta x = -A^{-1}\delta A(x + \delta x) &\Rightarrow \|\delta x\| \leq \|A^{-1}\|\|\delta A\|\|x + \delta x\| \\ &\Rightarrow \frac{\|\delta x\|}{\|x + \delta x\|} \leq \|A\|\|A^{-1}\| \frac{\|\delta A\|}{\|A\|}. \end{aligned} \quad (2.96)$$

Jeżeli zostaną uwzględnione zaburzenia (błędy) we współczynnikach obu stron układu równań, to dla dostatecznie małych $\|\delta A\|$ otrzymuje się

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A\|\|A^{-1}\|}{1 - \|A^{-1}\|\|\delta A\|} \left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right). \quad (2.97)$$

W każdym wypadku to współczynnik (wskaźnik) uwarunkowania macierzy A decyduje o tym jak bardzo błąd danych wejściowych jest „wzmacniany”, przenosząc się na błąd wyniku.

Jeżeli $\varepsilon = eps/2$ oznacza względny błąd arytmetyki zmiennopozycyjnej ($eps = 2,22 \cdot 10^{-16}$ dla liczb zmiennopozycyjnych podwójnej precyzji), to błąd rozwiązania można oszacować jako

$$\frac{\|\delta x\|}{\|x\|} \lesssim 2\varepsilon \cdot \text{cond}(A). \quad (2.98)$$

Wartość wskaźnika uwarunkowania zależy od zastosowanej normy macierzy, np. dla normy L_2 będzie to iloraz największej do najmniejszej wartości szczególnej macierzy. Jeżeli iloraz ten jest duży, to wskaźnik uwarunkowania pozostanie duży niezależnie od wybranej normy.

Wskaźnik uwarunkowania macierzy diagonalnej jest równy ilorazowi największego do najmniejszego modułu jej elementów. Jeśli wszystkie będą równe, to wskaźnik uwarunkowania będzie równy 1, nawet jeśli elementy te będą bliskie zeru. Jeśli wystąpią duże dysproporcje między liczbami na diagonalu, to wskaźnik uwarunkowania będzie wysoki.

Macierz o wysokim wskaźniku uwarunkowania jest w pewnym sensie bliska macierzy osobliwej. Wśród wartości własnych macierzy o wysokim wskaźniku uwarunkowania znajdują się liczby o bardzo małym module. Dokładne wyznaczenie wskaźnika uwarunkowania wymagałoby dokładnego wyznaczenia A^{-1} , a to przy wysokim wskaźniku uwarunkowania może być kłopotliwe. Dlatego też zwykle szacuje się tylko wartość wskaźnika uwarunkowania macierzy.

Oczywiste jest, że nie ma sensu rozwiązywanie układu równań liniowych, jeśli współczynnik uwarunkowania macierzy A jest wysoki. Problemem nie jest wtedy metoda numeryczna, którą wybrano, ale właściwości samego układu równań. Jeżeli natrafiamy na taki problem, to najlepszą strategią jest modyfikacja zadania, która pozwoli na zmniejszenie wskaźnika uwarunkowania. Nie zawsze można zaproponować sposób takiej modyfikacji, ale dobrą strategią może być zrównoważenie współczynników układu – doprowadzenie do tego, by wszystkie były tego samego rzędu. Można to uzyskać przez pomnożenie poszczególnych równań przez niezerowe współczynniki β_i oraz przeskalowanie niewiadomych: $x_j = \alpha_j x'_j$. W efekcie otrzymamy nowy układ równań:

$$A'x' = b', \quad A' = D_\beta A D_\alpha, \quad b' = D_\beta b, \quad x = D_\alpha x', \quad (2.99)$$

gdzie diagonalne macierze $D_\alpha = \text{diag}\{\alpha_i\}$, $D_\beta = \text{diag}\{\beta_i\}$ zawierają odpowiednie współczynniki na głównej przekątnej. Warto, by współczynniki te były potęgami podstawy stosowanego systemu pozycyjnego (czyli najczęściej potęgami 2), żeby uniknąć zaokrąglania.

Przykład 2.5 (Dahlquist, Björk, 1983).

Obliczenia wykonano w arytmetyce podwójnej precyzji. Niech macierzą współczynników będzie $A = \begin{bmatrix} 1,2969 & 0,8648 \\ 0,2161 & 0,1441 \end{bmatrix}$. Macierz b dobrano tak, by $b = \begin{bmatrix} 1,2969 & 0,8648 \\ 0,2161 & 0,1441 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 4,3234 \\ 0,7204 \end{bmatrix}$, czyli $x = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ jest dokładnym rozwiązaniem równania $Ax = b$. Rozwiązaniem równania $Ax = b - 10^{-8} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ jest $\hat{x} = \begin{bmatrix} 2,7207 \\ 0,9192 \end{bmatrix}$, a więc bardzo odległe od rozwiązania równania niezaburzonego, mimo że reszta obliczona dla tego rozwiązania $r = A\hat{x} - b = 10^{-8} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ jest tego rzędu, co zaburzenie w macierzy b . W tabeli 2.4 podano rozwiązania uzyskane po zaburzeniu pojedynczych elementów macierzy A i normy odpowiadającej im reszty.

Tabela 2.4. Rozwiązania uzyskane po zaburzeniu pojedynczych elementów macierzy A przez dodanie 10^{-8}

	a_{11}	a_{12}	a_{21}	a_{22}
$\hat{x} =$	$\begin{bmatrix} 1,74817 \\ 2,3778 \end{bmatrix}$	$\begin{bmatrix} 1,6324 \\ 2,5513 \end{bmatrix}$	$\begin{bmatrix} 14,7929 \\ -17,1849 \end{bmatrix}$	$\begin{bmatrix} 2,7530 \\ 0,8707 \end{bmatrix}$
$\ r\ $	$1,7 \cdot 10^{-8}$	$2,6 \cdot 10^{-8}$	$1,5 \cdot 10^{-7}$	$0,9 \cdot 10^{-8}$

Jak widać, zaburzenie dowolnego współczynnika powoduje zmianę rozwiązania około 10^8 razy większą od zaburzenia. Przyczyną jest wskaźnik uwarunkowania macierzy, który w normie euklidesowej wynosi $\text{cond}(A) = 2,4973 \cdot 10^8$.

2.7. Inne metody rozwiązywania układów równań liniowych

Metoda eliminacji Gaussa i wykorzystanie rozkładu trójkątnego są standardowymi metodami rozwiązywania dobrze uwarunkowanych układów równań liniowych. Przeglądając literaturę opisującą algorytmy numeryczne można znaleźć wiele konkurencyjnych metod eliminacji. Przykładowo eliminacja Gaussa-Jordana, która polega na równoważnym przekształceniu układu o współczynnikach $A^{(1)} := [A \ b]$ do układu o współczynnikach $A^{(n)} := [I \ b^{(n)}]$. Wymaga ona jednak większych nakładów obliczeniowych o ok. 50% operacji elementarnych nie oferując jednocześnie zmniejszenia wpływu błędów zaokrągleń. Jeżeli poczyni się dodatkowe założenia o macierzy współczynników, to można zaproponować metody bardziej oszczędne od eliminacji Gaussa. Przykładem może być dla macierzy dodatnio określonych³, metoda Choleskiego prowadząca do rozkładu $A = LL^T$, gdzie L jest macierzą trójkątną dolną. W przypadku szczególnych postaci macierzy współczynników, zwłaszcza w przypadku **macierzy rzadkich** (to jest dużych macierzy, które zawierają znaczną liczbę np. 90% współczynników zerowych), bardziej skuteczne od eliminacji Gaussa mogą okazać się również metody iteracyjne.

³ Macierz symetryczna A jest dodatnio określona, jeśli $\forall_{x \neq 0} x^T A x > 0$.

3. Aproxymacja i interpolacja

3.1. Modelowanie na podstawie danych cyfrowych

Niech funkcja $f(x)$ opisuje zależność, o której zebraliśmy informacje w wybranych $m + 1$ punktach x_i , $i = 0, 1, \dots, m$, zwanych **węzłami (punktami węzłowymi)**, czyli znane są wartości

$$f_i = f(x_i), \quad i = 0, \dots, m. \quad (3.1)$$

Niekiedy zbiór wszystkich punktów węzłowych $x_S = \{x_i, i = 0, \dots, m\}$ nazywa się **siatką**. Na podstawie tych informacji chcemy zbudować model analityczny odtwarzający zależność $f(x)$.

Z takim problemem spotykamy się, gdy postać analityczna $f(x)$ nie jest znana, a dane pochodzą np. z przeprowadzonych pomiarów. W analizie numerycznej częstszym i bardziej interesującym przypadkiem jest sytuacja, w której postać $f(x)$ jest znana, jednak sposób obliczania wartości $f(x)$ dla konkretnych argumentów jest uciążliwy, złożony obliczeniowo (czyli wymaga dużej liczby operacji elementarnych). Próbujemy wtedy zaproponować bardziej „przyjazny” model zależności $f(x)$, godząc się z tym, że będzie on obarczony pewnym błędem przybliżenia i że do jego wyznaczenia konieczne będzie obliczenie wartości $f(x)$ w punktach węzłowych. W zamian za to chcielibyśmy otrzymać funkcję $f^*(x)$, która będzie „dobrze” przybliżała $f(x)$ (przy tym pojęcie „dobrego” przybliżenia powinno być precyzyjnie zdefiniowane), a ponadto:

- będzie łatwo ją opisać skończonym zbiorem rzeczywistych parametrów,
- będzie łatwo obliczać jej wartości w dowolnym punkcie,
- będzie łatwo ją różniczkować i całkować,
- będzie łatwo wykonywać operacje arytmetyczne na funkcjach przybliżających – obliczać ich sumę lub iloczyn.

Od razu nasuwa się klasa funkcji, które spełniają te wymagania – to wielomiany. Bardzo często to właśnie w klasie wielomianów określonego stopnia poszukiwana jest funkcja przybliżająca. Czasem jednak zastosowanie wielomianów nie jest uzasadnione. Jeżeli wiemy, że przybliżana funkcja jest okresowa, to będziemy poszukiwać także okresowej funkcji przybliżającej – wtedy sięgniemy np. do tzw. wielomianów trygonometrycznych. Jeżeli funkcja $f(x)$ posiada osobliwości (nieciągłości), które chcemy odtworzyć w modelu, to funkcja przybliżająca może być np. funkcją wymierną.

Każdy problem przybliżenia danych numerycznych (3.1) wymaga określenia klasy (zbioru) funkcji, z której wybierzemy funkcję przybliżającą $f^*(x)$. Co więcej klasa ta powinna być sparametryzowana przez skończony wektor parametrów $c \in R^p$ (można więc zapisać dowolną funkcję z tej klasy jako $f^*(x, c)$, podkreślając jej zależność od parametrów). Ostateczny wybór funkcji przybliżającej polega na wyborze „najlepszych” parametrów c_0 , czyli wybraną funkcję przybliżającą można zapisać jako $f^*(x) = f^*(x, c_0)$. Nie jest więc konstruktywne np. poszukiwanie funkcji przybliżającej w klasie wszystkich funkcji ciągłych, a jest np. w klasie wielomianów stopnia nie wyższego niż 5.

Do konstrukcji funkcji przybliżającej $f^*(x)$ można podejść na dwa sposoby. Jeżeli uważamy, że we wszystkich punktach siatki funkcja przybliżająca powinna być równa przybliżanej, to podstawą jej konstrukcji są równania

$$f^*(x_i) = f_i = f(x_i), \quad i = 0, \dots, m. \quad (3.2)$$

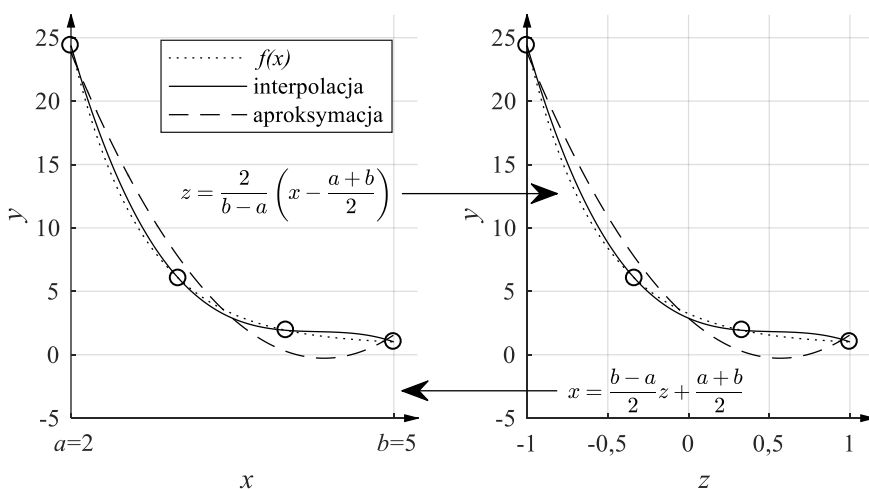
Taki sposób konstrukcji funkcji przybliżającej nazywany jest **interpolacją**, a funkcję przybliżającą **interpolantem** lub **funkcją interpolującą**. Jeżeli warunków narzuconych w węzłach jest dokładnie tyle, co parametrów w opisie klasy rozważanych funkcji interpolujących, to parametry wybranej funkcji będą z reguły określone w sposób jednoznaczny. Czasem do warunków (3.2) dodaje się dodatkowe wymagania np. dotyczące pochodnej funkcji interpolującej. Oczywiście mamy nadzieję, że zaproponowana funkcja interpolująca przybliży funkcję $f(x)$ także między węzłami i badaniu jakości tego przybliżenia poświęcimy dużo uwagi, ale wymagania dotyczące zachowania funkcji interpolującej między węzłami nie wchodzi do standardowo sformułowanego zadania interpolacji. Rozwiązanie zadania interpolacji polega na wyznaczeniu takich parametrów c_0 , że funkcja $f^*(x, c_0)$ spełnia zadane warunki algebraiczne (3.2).

Drugi sposób konstrukcji funkcji przybliżającej $f^*(x)$ – **aproksymacja** – zakłada, że powinna to być funkcja z określonej klasy (zbioru funkcji), która będzie minimalizowała błąd przybliżenia informacji zawartej w zależnościach (3.1). Każde zadanie aproksymacji wymaga więc:

- 1 – określenia klasy funkcji $f^*(x, c)$, w której poszukujemy funkcji aproksymującej,
- 2 – określenia wyrażenia $J(x_S, c)$, którego argumentami są węzły siatki x_S i parametry funkcji aproksymującej c , za pomocą którego będzie mierzony błąd aproksymacji.

Rozwiązanie zadania aproksymacji polega na rozwiązaniu problemu optymalizacji: znalezieniu takich parametrów c^* , że $\forall_c J(x_S, c^*) \leq J(x_S, c)$. Jeżeli zmieni się istotnie którykolwiek z elementów 1, 2 w zadaniu aproksymacji, to zmieni się też, i to diametralnie, sposób rozwiązania tego zadania.

Zadanie przybliżania danych numerycznych można, a często jest to korzystne, poprzedzić etapem normalizacji tych danych. Np. w przypadku modelowania funkcji jednej zmiennej normalizuje się siatkę, tak by wszystkie węzły znalazły się w przedziale $[-1, 1]$. Oczywiście po rozwiązaniu zadania należy powrócić do oryginalnej skali zmiennej niezależnej. Wszystkie te operacje opierają się na liniowych przekształceniach zmiennej niezależnej, nie są więc uciążliwe.



Rys. 3.1. Interpolacja i aproksymacja: linia kropkowana – funkcja przybliżana, linia kreskowa – aproksymacja wielomianem stopnia 2 metodą najmniejszych kwadratów, linia ciągła – interpolacja wielomianem stopnia 3. Z lewej zadanie rozwiązane na oryginalnym zestawie danych, z prawej po normalizacji zmiennej niezależnej do przedziału $[-1, 1]$

3.2. Liniowe zadanie aproksymacji średniokwadratowej

Niech funkcje $\varphi_i(x)$, $i = 0, \dots, n$ (nazywane **funkcjami bazowymi**) będą znane, wybrane przez projektanta. Funkcja aproksymująca jest wybierana spośród wszystkich kombinacji liniowych funkcji bazowych:

$$f^*(x) = \sum_{i=0}^n c_i \varphi_i(x). \quad (3.3)$$

Błąd aproksymacji jest mierzony wartością wyrażenia

$$J(x_S, c) = \sum_{i=0}^m (f^*(x_i) - f_i)^2 w_i, \quad (3.4)$$

w którym $w_i > 0$ są współczynnikami wagowymi doboranymi przez projektanta dla zróżnicowania „ważności” poszczególnych węzłów x_i . Sumowanie we wzorze (3.4) rozciąga się na wszystkie węzły. Tak postawione zadanie aproksymacji jest

nazywane **liniowym zadaniem aproksymacji średniokwadratowej**. Liniowe – bo liniowa jest zależność funkcji aproksymującej od szukanych współczynników c_i , a średniokwadratowy jest błąd aproksymacji (3.4).

Po podstawieniu funkcji aproksymującej (3.3) do wyrażenia (3.4) otrzymamy kwadratową funkcję nieznaną współczynników c_i . Rozwiązanie liniowego zadania aproksymacji średniokwadratowej polega więc na znalezieniu minimum kwadratowej funkcji wielu zmiennych c_i . Funkcja ta jest nieujemna i wypukła, więc minimum takie istnieje, jest jedyne i znajduje się w punkcie, w którym zerują się pochodne cząstkowe funkcji $J(x_S, c)$ względem współczynników c_i .

Rozważmy zadanie, w którym występują tylko dwie funkcje bazowe φ_0, φ_1 i w związku z tym dwa współczynniki c_0, c_1 . Mamy wtedy

$$J(x_S, c) = \sum_{i=0}^m (c_0 \varphi_0(x_i) + c_1 \varphi_1(x_i) - f_i)^2 w_i, \quad (3.5)$$

$$\begin{aligned} \frac{\partial J(x_S, c)}{\partial c_0} &= \sum_{i=0}^m 2\varphi_0(x_i)(c_0 \varphi_0(x_i) + c_1 \varphi_1(x_i) - f_i)w_i \\ &= 2c_0 \sum_{i=0}^m \varphi_0(x_i)^2 w_i \\ &\quad + 2c_1 \sum_{i=0}^m \varphi_0(x_i)\varphi_1(x_i)w_i - 2 \sum_{i=0}^m \varphi_0(x_i)f_i w_i \end{aligned} \quad (3.6)$$

i podobnie

$$\begin{aligned} \frac{\partial J(x_S, c)}{\partial c_1} &= 2c_1 \sum_{i=0}^m \varphi_0(x_i)\varphi_1(x_i)w_i + 2c_1 \sum_{i=0}^m \varphi_1(x_i)^2 w_i \\ &\quad - 2 \sum_{i=0}^m \varphi_1(x_i)f_i w_i. \end{aligned} \quad (3.7)$$

Z przyrównania pochodnych (3.6) i (3.7) do zera wynikają równania, które można zapisać łącznie:

$$\begin{aligned} & \begin{bmatrix} \sum_{i=0}^m \varphi_0(x_i)^2 w_i & \sum_{i=0}^m \varphi_0(x_i) \varphi_1(x_i) w_i \\ \sum_{i=0}^m \varphi_0(x_i) \varphi_1(x_i) w_i & \sum_{i=0}^m \varphi_1(x_i)^2 w_i \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} \\ & = \begin{bmatrix} \sum_{i=0}^m \varphi_0(x_i) f_i w_i \\ \sum_{i=0}^m \varphi_1(x_i) f_i w_i \end{bmatrix}. \end{aligned} \quad (3.8)$$

Jest to układ równań liniowych, który trzeba rozwiązać, by wyznaczyć współczynniki funkcji aproksymującej, rozwiązującego liniowe zadanie aproksymacji średniokwadratowej.

W przypadku $n + 1$ funkcji bazowych (i $n + 1$ nieznanymi współczynnikami) postępowanie jest analogiczne. Aby skrócić i uprościć zapis wprowadza się następujące oznaczenia.

Definicja 3.1 (iloczynu skalarnego w liniowym zadaniu aproksymacji średniokwadratowej)

Dla dowolnych funkcji f, g przy danej siatce węzłów i współczynnikach wagowych **iloczynem skalarnym** funkcji f, g nazywa się liczbę

$$\langle f, g \rangle := \sum_{i=0}^m f(x_i) g(x_i) w_i. \quad (3.9)$$

Jeżeli $\langle f, g \rangle = 0$ to funkcje $f(\cdot)$, $g(\cdot)$, nazywamy **ortogonalnymi**.

Jeżeli $\langle f_i, f_j \rangle = 0$ dla $i \neq j$ i $\langle f_i, f_i \rangle \neq 0$ to funkcje $f_i(\cdot)$, $i = 1, 2, \dots$ nazywamy **układem (rodziną) funkcji ortogonalnych**.

Nazwa iloczyn skalarny dla operatora $\langle f, g \rangle$ nie jest przypadkowa – ma on wszystkie (poza dodatnią określonością) właściwości iloczynu skalarnego w przestrzeni euklidesowej. Jak wynika z definicji (3.9), wartość iloczynu skalarnego zależy od wyboru węzłów i współczynników wagowych. Tak więc, dwie funkcje mogą być ortogonalne na jednym układzie węzłów, a na innym nie.

Korzystając z wprowadzonej notacji i rozumując analogicznie jak w przypadku dwóch funkcji bazowych, można wyprowadzić rozwiązanie liniowego zadania aproksymacji średniokwadratowej.

Twierdzenie 3.1 (o rozwiązaniu liniowego zadania aproksymacji średniokwadratowej – *Dahlquist, Björck, 1983*):

Jeżeli funkcje bazowe są tak wybrane, że wektory $\varphi_i(x_S) = [\varphi_i(x_0), \dots, \varphi_i(x_m)]$, $i = 0, 1, \dots, n$ są liniowo niezależne (dodatek D1), to liniowe zadanie aproksymacji średniokwadratowej ma jedyne rozwiązanie. Rozwiązanie to spełnia układ równań:

$$\begin{bmatrix} \langle \varphi_0, \varphi_0 \rangle & \langle \varphi_1, \varphi_0 \rangle & \cdots & \langle \varphi_n, \varphi_0 \rangle \\ \langle \varphi_0, \varphi_1 \rangle & \langle \varphi_1, \varphi_1 \rangle & \cdots & \langle \varphi_n, \varphi_1 \rangle \\ \cdots & \cdots & \cdots & \cdots \\ \langle \varphi_0, \varphi_n \rangle & \langle \varphi_1, \varphi_n \rangle & \cdots & \langle \varphi_n, \varphi_n \rangle \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} \langle f, \varphi_0 \rangle \\ \langle f, \varphi_1 \rangle \\ \cdots \\ \langle f, \varphi_n \rangle \end{bmatrix}. \quad (3.10)$$

Jeżeli funkcje bazowe są rodziną funkcji ortogonalnych, to rozwiązanie upraszcza się do:

$$c_i = \frac{\langle f, \varphi_i \rangle}{\langle \varphi_i, \varphi_i \rangle}, \quad i = 0, \dots, n. \quad (3.11)$$

Układ równań (3.10) nazywany jest **układem równań normalnych**, a macierz jego współczynników **gramianem**.

Wybór funkcji bazowych i węzłów jest najważniejszym etapem poszukiwania funkcji aproksymującej. Wybór ten wcale nie musi być jednoznaczny. Jeśli planujemy aproksymację wielomianem stopnia nie większego niż n , to możemy wybrać funkcje bazowe

$$\varphi_0(x) = 1, \quad \varphi_1(x) = x, \dots, \quad \varphi_n(x) = x^n, \quad (3.12)$$

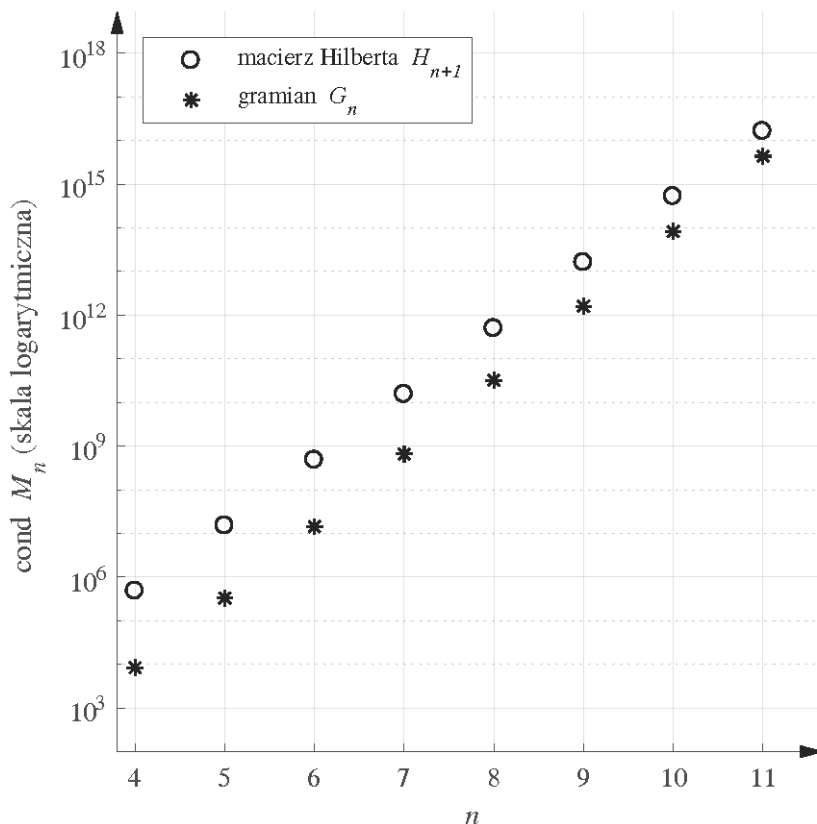
ale także i -tą funkcją bazową może być dowolny wielomian stopnia i . Wybór funkcji bazowych ma bezpośredni wpływ na wskaźnik uwarunkowania gramianu, a więc rzutuje na błąd rozwiązania układu równań normalnych.

Przykład 3.1

Obliczymy gramian funkcji bazowych (3.12) na przedziale $[0, 1]$ dla $n + 1$ węzłów równoodległych. Wprost z definicji 3.1 wynika, że

$$\begin{aligned} G_n &= [g_{ij}]_{i,j=0}^n, \quad g_{ij} = \langle \varphi_i, \varphi_j \rangle = \sum_{k=0}^n \left(\frac{k}{n}\right)^i \left(\frac{k}{n}\right)^j = \sum_{k=0}^n \left(\frac{k}{n}\right)^{i+j} \\ &= \frac{1}{n^{i+j}} \sum_{k=0}^n k^{i+j}. \end{aligned}$$

Jako że $\frac{m}{n^{m+1}} \sum_{k=0}^n k^m \xrightarrow[n \rightarrow \infty]{} 1$, to otrzymana macierz G_n jest zbliżona do macierzy Hilberta (czyli macierzy H_{n+1} o elementach $h_{i,j} = \frac{1}{i+j-1}, i, j = 1, \dots, n+1$) o wymiarach $(n+1) \times (n+1)$ pomnożonej przez $n+1$, która jest podręcznikowym przykładem macierzy źle uwarunkowanej. Wykres na rysunku 3.2 pokazuje współczynniki uwarunkowania takich macierzy dla n od 4 do 11.



Rys. 3.2. Wskaźnik uwarunkowania gramianu G_n i macierzy Hilberta H_{n+1}

Rozważany gramian jest nieznacznie lepiej uwarunkowany od macierzy Hilberta, ale osiąga współczynnik uwarunkowania 10^{15} już przy $n = 11$, co praktycznie uniemożliwia numeryczne rozwiązanie układu równań normalnych przy obliczeniach w podwójnej precyzji (format double IEEE754/854). Uwarunkowanie gramianu rośnie przy tym szybciej i zbliża się do uwarunkowania macierzy Hilberta.

Jak wynika z zależności (3.11) doskonały byłby wybór ortogonalnej rodziny funkcji bazowych. Jeżeli celem jest aproksymacja wielomianem, to istnieje kilka rodzin wielomianów ortogonalnych, z których można skorzystać. Jedną z nich są wielomiany Czebyszewa. Znajdują one wiele zastosowań i wielokrotnie będziemy się odwoływać do ich właściwości.

3.3. Wielomiany Czebyszewa

Definicja i podstawowe własności.

Rodzina wielomianów Czebyszewa jest określona na przedziale $[-1, 1]$ wzorem otwartym

$$T_n(x) = \cos(n \arccos x) \quad -1 \leq x \leq 1, \quad n = 0, 1, \dots, \quad (3.13)$$

albo za pomocą wzoru rekurencyjnego

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \quad (3.14)$$

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

Podstawowe właściwości wielomianów Czebyszewa, to:

1. Współczynnik przy najwyższej potędze w wielomianie $T_n(x)$ jest równy 2^{n-1} dla $n = 1, 2, \dots$
2. $T_n(-x) = (-1)^n T_n(x)$.
3. Wielomian $T_{n+1}(x)$ ma $n + 1$ pierwiastków w punktach

$$x_k = \cos \frac{(2k+1)\pi}{2(n+1)}, \quad k = 0, 1, \dots, n, \quad n = 0, 1, \dots. \quad (3.15)$$

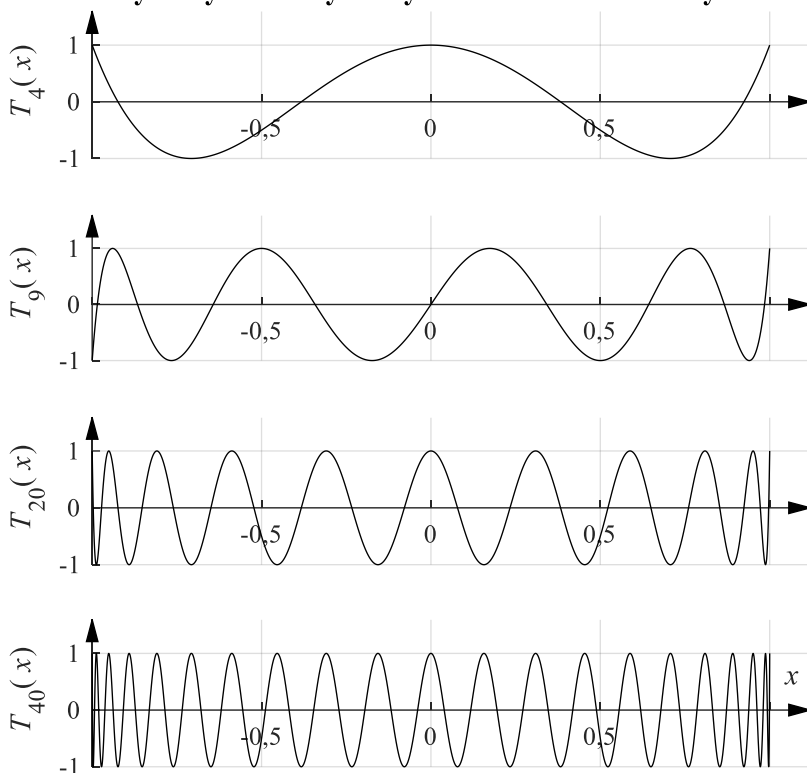
Punkty te są nazywane węzłami Czebyszewa I rodzaju.

4. Wielomian $T_n(x)$ ma $n + 1$ ekstremów lokalnych w przedziale $[-1, 1]$ (wliczając końce przedziału, które są ekstremami warunkowymi – nie są ekstremami $T_n(x)$) rozpatrywanego jako funkcja $T_n(x)$ w punktach

$$x_k = \cos \frac{k\pi}{n}, \quad k = 0, 1, \dots, n, \quad n = 1, 2, \dots, \quad (3.16)$$

zwanych węzłami Czebyszewa II rodzaju.

Wykresy kilku wybranych wielomianów Czebyszewa



Rys. 3.3. Wykres kilku wielomianów Czebyszewa

Ortogonalność wielomianów Czebyszewa

Twierdzenie 3.2 (o ortogonalności wielomianów Czebyszewa – Dahlquist, Björck, 1983)

Układ wielomianów $T_0(x), T_1(x), \dots, T_n(x)$ jest ortogonalny względem wag $w_i = 1$ i węzłów x_i , które są zerami wielomianu $T_{n+1}(x)$:

$$\langle T_i, T_j \rangle = \begin{cases} 0 & \text{dla } i \neq j \\ \frac{n+1}{2} & \text{dla } i = j \neq 0 \\ n+1 & \text{dla } i = j = 0 \end{cases} \quad (3.17)$$

Własność min-max wielomianów Czebyszewa

Definicja 3.2

Wielomian będziemy nazywać **monicznym** gdy jego współczynnik przy najwyższej potędze jest równy 1.

Definicja 3.3

Dla funkcji f ciągłej w przedziale $[a, b]$ definiujemy normę $\|f\|_{[a,b]} = \max_{x \in [a,b]} |f(x)|$.

Twierdzenie 3.3 (własność min-max wielomianów Czebyszewa)

Jeśli P jest wielomianem monicznym stopnia $n > 0$, to

$$\|P\|_{[-1,1]} \geq \|2^{1-n}T_n\|_{[-1,1]} = 2^{1-n}.$$

Dowód: (przez doprowadzenie do sprzeczności)

Przypuśćmy, że $\|P\|_{[-1,1]} < 2^{1-n} \Leftrightarrow \forall x \in [-1,1] |P(x)| < 2^{1-n}$. Rozważmy wielomian moniczny $2^{1-n}T_n$. Dla $n+1$ wartości $x_k = \cos \frac{k\pi}{n}$, $k = 0, 1, \dots, n$ w przedziale $[-1, 1]$ mamy $2^{1-n}T_n(x_k) = 2^{1-n} \cos\left(n \arccos\left(\cos \frac{k\pi}{n}\right)\right) = 2^{1-n} \cos(k\pi) = \begin{cases} 2^{1-n} & \text{dla parzystych } k \\ -2^{1-n} & \text{dla nieparzystych } k \end{cases}$

Wynika stąd, że różnica $2^{1-n}T_n - P$ ma w każdym z punktów x_k , $k = 0, 1, \dots, n$ taki sam znak jak $2^{1-n}T_n$, a więc zmienia znak tak, że ma co najmniej n pierwiastków w przedziale $[-1, 1]$, co jest sprzeczne z tym, że stopień wielomianu $2^{1-n}T_n - P$ jest mniejszy od n (bo współczynniki przy najwyższej, n -tej potędze w obu wielomianach są równe 1).

3.4. Aproksymacja jednostajna

Słynnym twierdzeniem w teorii funkcji rzeczywistych jest twierdzenie Weierstrassa:

Twierdzenie 3.4 (Weierstrassa – Kincaid, Cheney, 2006)

Jeżeli funkcja $f(x)$ jest ciągła w skończonym przedziale $[a, b]$, to dla każdego $\varepsilon > 0$ istnieje wielomian $P_n(x)$ stopnia n , taki że dla każdego $x \in [a, b]$, $|f(x) - P_n(x)| < \varepsilon$.

Twierdzenie to wprost prowokuje do postawienia zadania aproksymacji wielomianowej w następujący sposób:

Niech funkcja aproksymująca będzie wielomianem stopnia nie większego niż n :

$$f^*(x) = \sum_{i=0}^n c_i x^i, \quad (3.18)$$

a błąd aproksymacji będzie mierzony wartością

$$J(c) = \max_{x \in [a,b]} |f^*(x) - f(x)|. \quad (3.19)$$

Twierdzenie Weierstrassa gwarantuje, że tak zdefiniowany błąd aproksymacji można dowolnie zmniejszyć – podnosząc stopień wielomianu aproksymacyjnego. W praktycznych obliczeniach zadanie upraszcza się do minimalizacji wskaźnika jakości przybliżenia

$$J(x_S, c) = \max_i |f^*(x_i) - f_i|, \quad (3.19a)$$

a stopień n wielomianu f^* jest mniejszy od liczby węzłów.

Ta zmiana (w stosunku do liniowego zadania aproksymacji średniokwadratowej) wyrażenia służącego do oceny błędu ma daleko idące konsekwencje dla sposobu rozwiązania zadania. W wyrażeniu (3.18) jak i (3.19a) znajdują się dwie funkcje nieróżniczkowalne – moduł i maksimum, nie można więc obliczyć pochodnych cząstkowych i przyrównać do zera, tak jak było to w liniowym zadaniu aproksymacji średniokwadratowej. Dowody twierdzenia Weierstrassa nie dają użytecznego sposobu konstrukcji wielomianu aproksymacyjnego dla zadanego błędu ε . W niektórych wersjach wykorzystuje się przybliżenia oparte na wielomianach Bernsteina, które bardzo wolno, ale jednostajnie zbiegają do przybliżanej funkcji, a wiele dowodów spotykanych w literaturze jest zupełnie niekonstruktywnych.

Podstawą do praktycznego wyznaczania przybliżeń optymalnych, czyli wielomianów zadanego stopnia minimalizujących (3.19), jest poniższe twierdzenie.

Twierdzenie 3.5 (o alternansie – *Paszkowski, 1975*)

Jeżeli domknięty i ograniczony zbiór F zawiera co najmniej $n + 2$ punkty, to wielomian W (co najwyżej n -tego stopnia) jest wtedy i tylko wtedy n -tym wielomianem optymalnym dla funkcji f , ciągłej na przedziale $[-1, 1]$, gdy istnieje podzbiór $\{x_0, x_1, \dots, x_{n+1}\} \subset F$ (zwany alternansem) taki, że $x_0 < x_1 < \dots < x_{n+1}$ i że różnice $f(x_k) - W(x_k)$, ($k = 0, 1, \dots, n + 1$) są równe na przemian $\|f - W\|_\infty^F$ i $-\|f - W\|_\infty^F$, gdzie $\|\cdot\|_\infty^F$ oznacza normę supremum funkcji na zbiorze F , tzn. $\|g\|_\infty^F = \sup_{x \in F} |g(x)|$.

Taka forma twierdzenia daje podstawę do poszukiwania zarówno przybliżeń optymalnych w sensie wskaźnika jakości (3.19), jak i (3.19a). W obydwu przypadkach stosuje się numeryczne metody iteracyjne, poszukując przybliżenia, o właściwościach opisanych w twierdzeniu o alternansie. W praktyce stosuje się niekiedy **algorytm Remeza**, służący do obliczania optymalnych przybliżeń wielomianowych i przydatny do syntezy filtrów cyfrowych, ale w wielu sytuacjach wygodniejsze może być skorzystanie z poniższego twierdzenia 3.6.

Twierdzenie 3.6 (*Trefethen, 2013*)

Niech funkcja f będzie ciągła na przedziale $[-1, 1]$, p_n oznacza jej wielomian interpolacyjny stopnia n zbudowany na węzłach Czebyszewa II rodzaju, a p_n^* optymalne w sensie normy supremum przybliżenie wielomianem stopnia $n \geq 1$. Wtedy zachodzi nierówność:

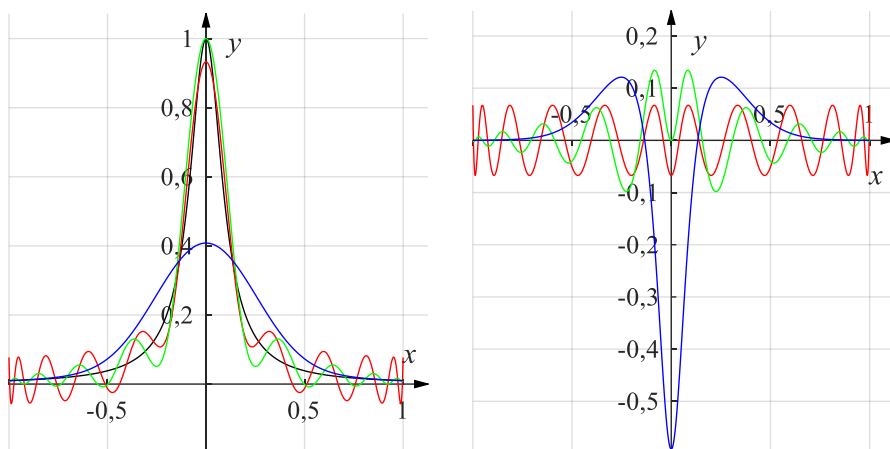
$$\|f - p_n\|_{[-1,1]} \leq \left(2 + \frac{2}{\pi} \log(n+1)\right) \|f - p_n^*\|_{[-1,1]}. \quad (3.20)$$

Wyrażenie w nawiasie w (3.20) rośnie bardzo wolno ze wzrostem n , na przykład dla $n = 10000$ wynosi około 7,863548449829773, a wartość 10 przekracza dopiero dla $n = 286751$. Zatem – ponieważ mało realne jest posługiwanie się wielomianami stopnia większego niż 200000, poza szczególnymi przypadkami, najłatwiej posłużyć się wielomianem interpolacyjnym z węzłami Czebyszewa (II rodzaju), zamiast poszukiwać przybliżeń optymalnych. Tak więc przybliżeniem rozwiązania zadania aproksymacji jednostajnej jest rozwiązanie zadania interpolacji (omówione dokładnie w podrozdziale 3.5) przy szczególnym wyborze węzłów.

Przykład 3.2

Dla funkcji $f(x) = \frac{1}{1+100x^2}$ wyznaczono (iteracyjnie, algorytmem Remeza) przybliżenie optymalne wielomianem stopnia 20 oznaczony $p_{20}^*(x)$, interpolant stopnia 20 z węzłami Czebyszewa II rodzaju - $p_{20}(x)$ i przybliżenie wielomianami Bernsteina stopnia 20 zgodnie ze wzorem

$$b_{20}(x) = \sum_{k=0}^{20} f\left(\frac{2k-20}{20}\right) \binom{20}{k} \left(\frac{x+1}{2}\right)^k \left(\frac{1-x}{2}\right)^{20-k}.$$

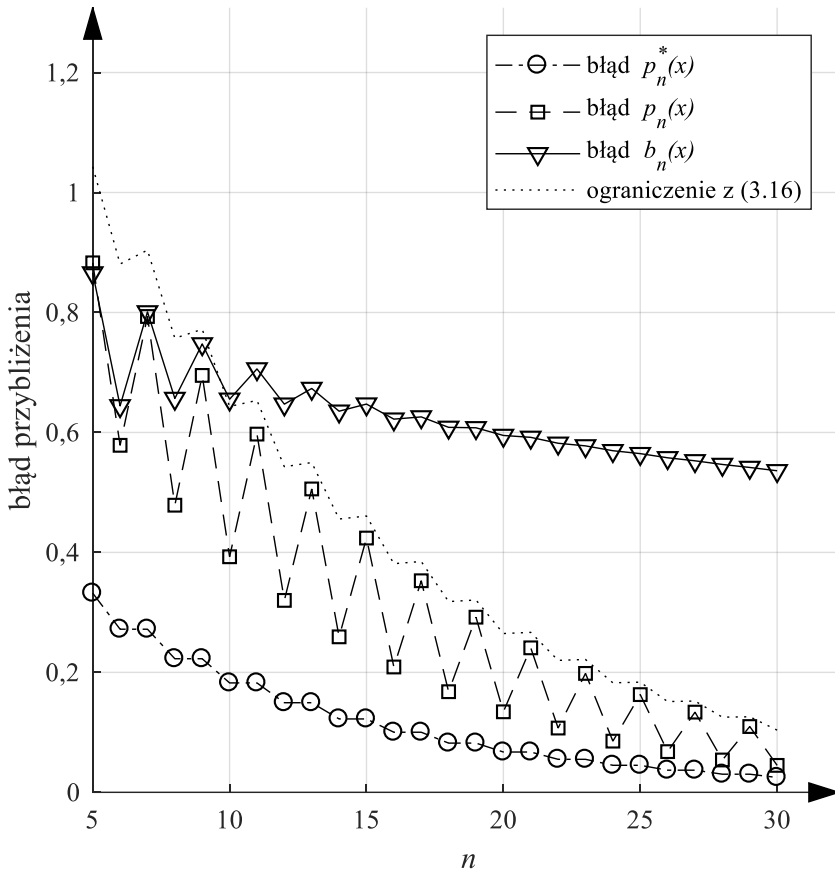


Rys. 3.4. Po lewej porównanie przybliżenia optymalnego stopnia 20 (linia czerwona), interpolantu Czebyszewa stopnia 20 (linia zielona) i przybliżenia wielomianami Bernsteina (linia niebieska) na tle funkcji przybliżanej (linia czarna). Po prawej błędy tych przybliżeń (ten sam dobór kolorów linii)

Obliczenia powtórzono dla wielomianów stopnia od 5 do 30, uzyskując wartości normy supremum błędu, przedstawione na rys. 3.5.

Na rysunku 3.5 widać wyraźnie bardzo wolną zbieżność przybliżeń wielomianami Bernsteina. Rzuci się w oczy oscylacyjny przebieg błędu przybliżenia interpolacyjnego, co wynika z faktu, że funkcja interpolowana jest funkcją parzystą z wyraźnym „pikiem” w zerze, co skutkuje większym błędem dla przybliżeń interpolacyjnych, dla których zero nie jest węzłem interpolacji, a więc nieparzystego stopnia⁴. Bliższe przeanalizowanie wykresu pozwala stwierdzić, że także przybliżenia optymalne stopnia nieparzystego nie są lepsze od przybliżeń stopnia parzystego, mniejszego o 1. Jest to specyficzna cecha przybliżanej funkcji.

⁴ Wielomian interpolacyjny stopnia n ma $n+1$ węzłów, więc stopnia nieparzystego ma parzystą liczbę węzłów. Dlatego n we wzorze (3.16) jest nieparzyste, a więc argument funkcji cosinus nie może być równy $\pi/2$.



Rys. 3.5. Porównanie błędów (w sensie wskaźnika (3.19)) dla przybliżenia optymalnego, interpolantu Czebyszewa i przybliżenia wielomianami Bernsteina w zależności od stopnia wielomianu

3.5. Interpolacja wielomianowa

Istnienie i jednoznaczność wielomianu interpolacyjnego

Poszukujemy wielomianu interpolacyjnego dla funkcji $f(x)$, na siatce węzłów

$$x_i, f_i = f(x_i), \quad i = 0, \dots, n. \quad (3.21)$$

Kwestię istnienia i jednoznaczności takiego wielomianu rozstrzyga twierdzenie:

Twierdzenie 3.7 (istnienie i jednoznaczność wielomianu interpolacyjnego):

Dla każdych, różnych $n + 1$ węzłów istnieje dokładnie jeden wielomian interpolacyjny (czyli spełniający warunki $P(x_i) = f_i$, $i = 0, \dots, n$) stopnia nie większego niż n .

Dowód:

Istnienie: wynika bezpośrednio z dowolnego sposobu konstrukcji wielomianu interpolacyjnego pokazanego dalej.

Jednoznaczność:

Założmy istnienie dwóch wielomianów interpolacyjnych P oraz Q , każdy stopnia nie wyższego niż n , czyli $P(x_i) = f_i$, $Q(x_i) = f_i$ dla $i = 0, 1, \dots, n$. Wtedy $P - Q$ jest też wielomianem stopnia nie wyższego niż n . Niezerowy wielomian stopnia nie wyższego niż n ma co najwyżej n pierwiastków. Wielomian $P - Q$ zeruje się w $n + 1$ punktach x_i , $i = 0, 1, \dots, n$, musi być więc wielomianem zerowym.

Jednoznaczność wielomianu interpolacyjnego nie zmienia faktu, że może on być wyznaczany i zapisywany na różne sposoby.

Wzór interpolacyjny Vandermonde'a

Podstawową postacią wielomianu jest postać potęgowa:

$$P(x) = c_n x^n + c_{n-1} x^{n-1} + \dots + c_1 x + c_0. \quad (3.22)$$

Można powiedzieć, że wybrano $n + 1$ liniowo niezależnych wielomianów

$$B = \{\varphi_0(x) = 1, \quad \varphi_1(x) = x, \quad \varphi_2(x) = x^2, \dots, \varphi_n(x) = x^n\} \quad (3.23)$$

– bazę w przestrzeni wielomianów stopnia nie wyższego niż n , i zapisano wielomian $P(x)$ jako liniową kombinację wielomianów z tej bazy. Jeżeli zbierzemy razem równości $P(x_i) = f_i$, $i = 0, \dots, n$, to otrzymamy układ równań liniowych względem nieznanych współczynników c_0, \dots, c_n :

$$\begin{bmatrix} x_0^n & x_0^{n-1} & \cdots & x_0 & 1 \\ x_1^n & x_1^{n-1} & \cdots & x_1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n-1}^n & x_{n-1}^{n-1} & \cdots & x_{n-1} & 1 \\ x_n^n & x_n^{n-1} & \cdots & x_n & 1 \end{bmatrix} \begin{bmatrix} c_n \\ c_{n-1} \\ \vdots \\ c_1 \\ c_0 \end{bmatrix} = \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \\ f_n \end{bmatrix}. \quad (3.24)$$

Macierz V współczynników w tym równaniu nosi nazwę macierzy Vandermonde'a. Metodą indukcji można pokazać, że wyznacznik macierzy Vandermonde'a jest równy

$$\begin{aligned} \det(V) &= \prod_{k < j} (x_k - x_j) \\ &= (x_0 - x_1)(x_0 - x_2)(x_1 - x_2) \dots (x_0 - x_n) \dots (x_{n-1} - x_n), \end{aligned} \quad (3.25)$$

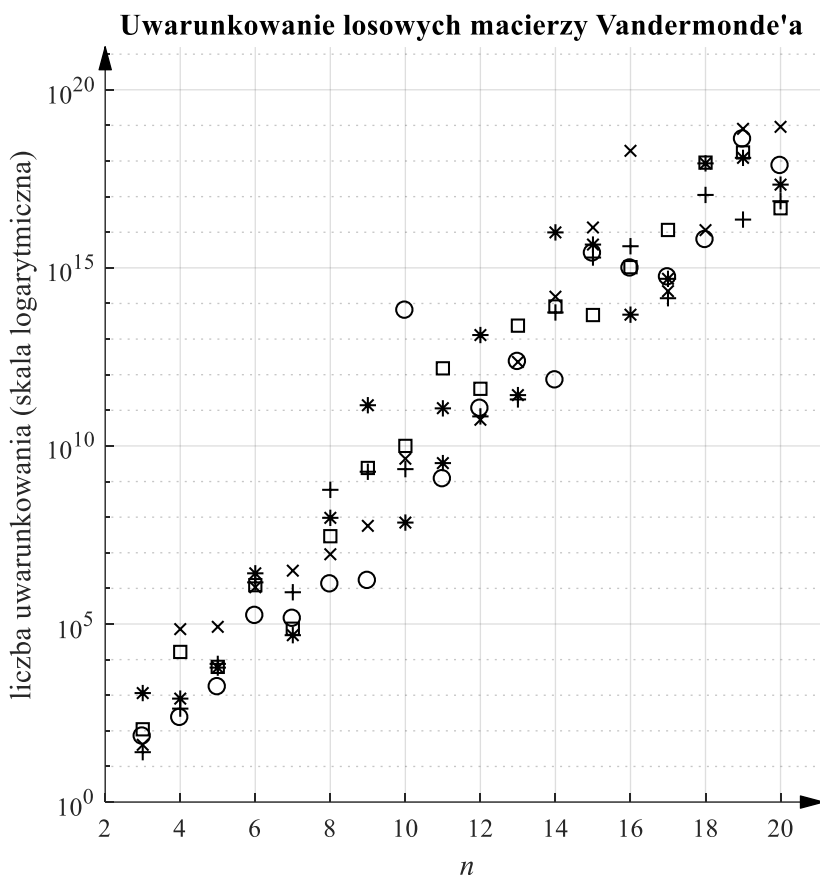
czyli macierz V jest nieosobliwa, jeśli tylko węzły interpolacji są różne. Układ równań (3.24) ma więc jednoznaczne rozwiązanie (wykazaliśmy istnienie i (ponownie) jednoznaczność wielomianu interpolacyjnego).

Niestety, wskaźnik uwarunkowania macierzy V rośnie bardzo szybko z jej wymiarem (a więc ze stopniem wielomianu interpolacyjnego) i dokładne rozwiązanie układu (3.24), przestaje być możliwe (rys. 3.6).

Równanie (3.24) można zapisać jako

$$\begin{bmatrix} \varphi_n(x_0) & \cdots & \varphi_0(x_0) \\ \vdots & \ddots & \vdots \\ \varphi_n(x_n) & \cdots & \varphi_0(x_n) \end{bmatrix} \begin{bmatrix} c_n \\ \vdots \\ c_0 \end{bmatrix} = \begin{bmatrix} f_0 \\ \vdots \\ f_n \end{bmatrix}. \quad (3.26)$$

Można próbować zaradzić niekorzystnemu zjawisku wysokiego wskaźnika uwarunkowania w równaniu (3.24) zmieniając bazę B (3.21), która posłużyła do przedstawienia wielomianu interpolacyjnego.



Rys. 3.6. Wskaźnik uwarunkowania macierzy Vandermonde'a. Dla każdego wymiaru n obliczono wskaźnik macierzy V dla pięciu losowo wygenerowanych rozkładów węzłów w przedziale $[0,1]$

Wzór interpolacyjny Lagrange'a

Weźmy wielomian zapisany wzorem

$$P(x) = \sum_{i=0}^n f_i \prod_{\substack{k=0 \\ k \neq i}}^n \frac{x - x_k}{x_i - x_k}. \quad (3.27)$$

Każdy z wielomianów

$$\varphi_i(x) = \prod_{\substack{k=0 \\ k \neq i}}^n \frac{x - x_k}{x_i - x_k} \quad (3.28)$$

jest wielomianem stopnia n jako iloczyn n wielomianów stopnia pierwszego, przyjmuje wartość 1 w węzle $x_i : \varphi_i(x_i) = 1$, i wartość zero w każdym innym węzle: $k \neq i \Rightarrow \varphi_i(x_k) = 0$. Wielomian $P(x)$ jest więc wielomianem stopnia nie wyższego niż n , spełniającym warunki $P(x_i) = f_i, i = 0, \dots, n$, czyli wielomianem interpolacyjnym.

Jak widzimy, przy bazie (3.28) macierz $\begin{bmatrix} \varphi_0(x_0) & \cdots & \varphi_n(x_0) \\ \vdots & \ddots & \vdots \\ \varphi_0(x_n) & \cdots & \varphi_n(x_n) \end{bmatrix}$ jest macierzą

jednostkową i współczynniki reprezentacji wielomianu interpolacyjnego w tej bazie są wartościami w węzłach. Z uwagi na dość dużą liczbę mnożeń wzór interpolacyjny Lagrange'a nie jest najdogodniejszy.

Interpolacja przez rodzinę trójkątną wielomianów

Newton zaproponował przedstawienie wielomianu interpolacyjnego przy użyciu bazy, która nazywa się **rodziną trójkątną wielomianów**:

$$\begin{aligned} \varphi_0(x) &= 1 \\ \varphi_1(x) &= (x - x_0) \\ \varphi_2(x) &= (x - x_0)(x - x_1) \\ &\dots \end{aligned} \tag{3.29}$$

$$\begin{aligned} \varphi_n(x) &= (x - x_0)(x - x_1) \cdots (x - x_{n-1}) \\ P(x) &= c_n \varphi_n(x) + c_{n-1} \varphi_{n-1}(x) + \cdots + c_1 \varphi_1(x) + c_0. \end{aligned} \tag{3.30}$$

Pozwala to na kolejne wyznaczanie współczynników:

$$\begin{aligned} f_0 &= P(x_0) = c_0 \Rightarrow c_0 = f_0, \\ f_1 &= P(x_1) = c_1(x_1 - x_0) + c_0 \Rightarrow c_1 = \frac{f_1 - c_0}{x_1 - x_0}, \\ f_2 &= P(x_2) = c_2(x_2 - x_0)(x_2 - x_1) + c_1(x_2 - x_0) + c_0 \Rightarrow c_2 = \dots \end{aligned} \tag{3.31}$$

lub na zapisanie układu równań

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & x_1 - x_0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_n - x_0 & \cdots & (x_n - x_0)(x_n - x_1) \cdots (x_n - x_{n-1}) \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \cdots \\ c_n \end{bmatrix} = \begin{bmatrix} f_0 \\ f_1 \\ \cdots \\ f_n \end{bmatrix}. \tag{3.32}$$

Jak widać, przy takim wyborze bazy do reprezentacji wielomianu interpolacyjnego trzeba rozwiązać trójkątny układ równań, co bardzo ułatwia znalezienie rozwiązania.

Czasem używa się do wyznaczania wielomianu interpolacyjnego metodą Newtona następującego schematu rekurencyjnego:

Jeżeli zdefiniujemy:

$$\begin{aligned} \Delta_1 &= \frac{f_1 - f_0}{x_1 - x_0}, \quad \Delta_2 = \frac{f_2 - f_1}{x_2 - x_1}, \quad \dots, \quad \Delta_n = \frac{f_n - f_{n-1}}{x_n - x_{n-1}}, \\ \Delta_2^{(2)} &= \frac{\Delta_2 - \Delta_1}{x_2 - x_0} = \frac{\left(\frac{f_2 - f_1}{x_2 - x_1} - \frac{f_1 - f_0}{x_1 - x_0}\right) \cdot 1}{(f_2 - f_1)(x_1 - x_0) - (f_1 - f_0)(x_2 - x_1)}, \\ &= \frac{(x_2 - x_1)(x_1 - x_0)(x_2 - x_0)}{(f_2 - f_1)(x_1 - x_0) - (f_1 - f_0)(x_2 - x_1)}, \\ \Delta_3^{(2)} &= \frac{\Delta_3 - \Delta_2}{x_3 - x_1}, \dots, \Delta_i^{(2)} = \frac{\Delta_i - \Delta_{i-1}}{x_i - x_{i-2}}, \quad i = 2, 3, \dots, n-2, \\ \Delta_3^{(3)} &= \frac{\Delta_3^{(2)} - \Delta_2^{(2)}}{x_3 - x_0}, \dots \end{aligned} \quad (3.33)$$

to

$$P(x) = \Delta_n^{(n)} \varphi_n(x) + \Delta_{n-1}^{(n-1)} \varphi_{n-1}(x) + \dots + \Delta_2^{(2)} \varphi_2(x) + \Delta_1 \varphi_1(x) + f_0 \quad (3.34)$$

i współczynniki można obliczyć według schematu przedstawionego w tabeli 3.1, w którym zielona strzałka oznacza pierwszy, a czerwona drugi argument odejmowania, współczynniki potrzebne do konstrukcji wielomianu (3.34) znajdują się na przekątnej tabeli 3.1.

Tabela 3.1. Schemat metody Newtona tworzenia wielomianu interpolacyjnego na przykładzie wielomianu stopnia 5

	$\begin{matrix} - & \times \\ + & \rightarrow \end{matrix}$	$\frac{1}{x_i - x_{i-1}}$	$\frac{1}{x_i - x_{i-2}}$	$\frac{1}{x_i - x_{i-3}}$	$\frac{1}{x_i - x_{i-4}}$	$\frac{1}{x_i - x_{i-5}}$
x_0	$f(x_0)$					
x_1	$f(x_1)$	Δ_1				
x_2	$f(x_2)$	Δ_2	$\Delta_2^{(2)}$			
x_3	$f(x_3)$	Δ_3	$\Delta_3^{(2)}$	$\Delta_3^{(3)}$		
x_4	$f(x_4)$	Δ_4	$\Delta_4^{(2)}$	$\Delta_4^{(3)}$	$\Delta_4^{(4)}$	
x_5	$f(x_5)$	Δ_5	$\Delta_5^{(2)}$	$\Delta_5^{(3)}$	$\Delta_5^{(4)}$	$\Delta_5^{(5)}$

Rekurencyjne metody tworzenia wielomianów interpolacyjnych

Rekurencyjne metody budowania wielomianu interpolacyjnego opierają się na twierdzeniu 3.8.

Twierdzenie 3.8 (rekurencyjne tworzenie wielomianów interpolacyjnych)

Niech $P_{i_0, i_1, \dots, i_k}(x)$ oznacza wielomian stopnia nie wyższego niż k , spełniający warunki interpolacji w węzłach o numerach i_0, i_1, \dots, i_k : $P_{i_0, i_1, \dots, i_k}(x_{i_j}) = f_{i_j}$ $j = 0, \dots, k$.

Obowiązuje następująca zależność rekurencyjna

$$P_i(x) = f_i \quad i = 0, \dots, n, \quad (3.35)$$

$$P_{i_0, i_1, \dots, i_k}(x) = \frac{(x - x_{i_0})P_{i_1, i_2, \dots, i_k}(x) - (x - x_{i_k})P_{i_0, i_1, \dots, i_{k-1}}(x)}{x_{i_k} - x_{i_0}} = \quad (3.36)$$

$$\frac{\begin{vmatrix} x-x_{i_0} & P_{i_0, i_1, \dots, i_{k-1}}(x) \\ x-x_{i_k} & P_{i_1, i_2, \dots, i_k}(x) \end{vmatrix}}{x_{i_k} - x_{i_0}},$$

która rozpoczyna się wielomianami (3.35) (stopnia 0) i pozwala na wyznaczenie wielomianu stopnia k z dwóch wielomianów stopnia $k - 1$, zgodnie z (3.36).

Dowód:

$$\begin{aligned} P_{i_0, i_1, \dots, i_k}(x_{i_0}) &= \frac{(x_{i_0} - x_{i_0})P_{i_1, i_2, \dots, i_k}(x_{i_0}) - (x_{i_0} - x_{i_k})P_{i_0, i_1, \dots, i_{k-1}}(x_{i_0})}{x_{i_k} - x_{i_0}} = \\ &= P_{i_0, \dots, i_{k-1}}(x_{i_0}) = f_{i_0} \\ P_{i_0, i_1, \dots, i_k}(x_{i_k}) &= \frac{(x_{i_k} - x_{i_0})P_{i_1, i_2, \dots, i_k}(x_{i_k}) - (x_{i_k} - x_{i_k})P_{i_0, i_1, \dots, i_{k-1}}(x_{i_k})}{x_{i_k} - x_{i_0}} = \\ &= P_{i_1, \dots, i_k}(x_{i_k}) = f_{i_k} \end{aligned}$$

dla $0 < j < k$:

$$\frac{P_{i_0, i_1, \dots, i_k}(x_{i_j}) = \frac{(x_{i_j} - x_{i_0})P_{i_1, i_2, \dots, i_k}(x_{i_j}) - (x_{i_j} - x_{i_k})P_{i_0, i_1, \dots, i_{k-1}}(x_{i_j})}{x_{i_k} - x_{i_0}} = \frac{(x_{i_j} - x_{i_0})f_{i_j} - (x_{i_j} - x_{i_k})f_{i_j}}{x_{i_k} - x_{i_0}} = f_{i_j}.$$

Zgodnie z tym twierdzeniem, rozpoczynając od wielomianów stopnia zerowego (stałych), spełniających warunki interpolacji w jednym węźle, można zbudować wielomiany stopnia pierwszego, których wykresy przechodzą przez dwa węzły, następnie stopnia drugiego – spełniające równania trzech węzłów, i tak dalej aż do wielomianu interpolacyjnego stopnia n . Poszczególne metody różnią się strategią kolejnego wyboru węzłów. Np. **metoda Aitkena** wykorzystuje schemat podany w tabeli 3.2.

i	x_i	$x_i - x$	$y_i = P_i(x)$	$P_{0,i}(x)$	$P_{0,i,1}(x)$	\dots	$P_{0,i,2,i}(x)$	\dots	$P_{0,i,1,\dots,m}(x)$
0	x_0	$x_0 - x$	$y_0 = P_0(x)$	$\frac{x_0 - x}{x_1 - x} \frac{P_0(x)}{P_1(x)}$					
1	x_1	$x_1 - x$	$y_1 = P_1(x)$	$\frac{x_0 - x}{x_2 - x} \frac{P_0(x)}{P_2(x)}$	$\frac{x_1 - x}{x_2 - x} \frac{P_{0,1}(x)}{P_{0,2}(x)}$				
2	x_2	$x_2 - x$	$y_2 = P_2(x)$	$\frac{x_0 - x}{x_3 - x} \frac{P_0(x)}{P_3(x)}$	$\frac{x_1 - x}{x_3 - x} \frac{P_{0,1}(x)}{P_{0,3}(x)}$		$\frac{x_2 - x}{x_3 - x} \frac{P_{0,1,2}(x)}{P_{0,1,3}(x)}$		
3	x_3	$x_3 - x$	$y_3 = P_3(x)$	\vdots	\vdots		\vdots		
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots		\vdots		
m	x_m	$x_m - x$	$y_m = P_m(x)$	$\frac{x_0 - x}{x_m - x} \frac{P_0(x)}{P_m(x)}$	$\frac{x_1 - x}{x_m - x} \frac{P_{0,1}(x)}{P_{0,m}(x)}$		$\frac{x_2 - x}{x_m - x} \frac{P_{0,1,2}(x)}{P_{0,1,m}(x)}$		$\frac{x_{m-1} - x}{x_m - x} \frac{P_{0,1,\dots,m-1}(x)}{P_{0,1,\dots,m-2,m}(x)}$

Rys. 3.7. Schemat metody Aitkena

Tabela 3.2. Schemat rekurencyjnej metody tworzenia wielomianu interpolującego na przykładzie wielomianu stopnia 4

		Wielomiany utworzone z $P_0(x)$	Wielomiany utworzone z $P_{0,1}(x)$	Wielomiany utworzone z $P_{0,1,2}(x)$	Wielomiany utworzone z $P_{0,1,2,3}(x)$
x_0	$P_0(x) = f_0$				
x_1	$P_1(x) = f_1$	$P_{0,1}(x)$			
x_2	$P_2(x) = f_2$	$P_{0,2}(x)$	$P_{0,1,2}(x)$		
x_3	$P_3(x) = f_3$	$P_{0,3}(x)$	$P_{0,1,3}(x)$	$P_{0,1,2,3}(x)$	
x_4	$P_4(x) = f_4$	$P_{0,4}(x)$	$P_{0,1,4}(x)$	$P_{0,1,2,4}(x)$	$P_{0,1,2,3,4}(x)$

Rekurencyjne metody tworzenia wielomianów interpolacyjnych są najczęściej używane nie do wyznaczania analitycznej postaci wielomianu interpolacyjnego, ale do obliczania jego wartości w wybranym punkcie x . Schemat metody Aitkena można wtedy przedstawić jak na rys. 3.7, wykorzystując wyznacznik w liczniku wzoru (3.36).

3.6. Ocena jakości interpolacji – reszta wzoru interpolacyjnego i zjawisko Rungego

Wielomian interpolacyjny jest konstruowany po to, by zastępował interpolowaną funkcję w pewnym określonym przedziale, nie tylko w węzłach, ale przede wszystkim między węzłami. Zastosowanie twierdzenia o wartości średniej pozwala na wyprowadzenie następującego wzoru na różnicę między interpolowaną funkcją a wielomianem interpolacyjnym.

Twierdzenie 3.9 (o reszcie wzoru interpolacyjnego)

Jeżeli funkcja $f(x)$ ma ciągłe pochodne do rzędu $n + 1$, a $P(x)$ jest jej wielomianem interpolacyjnym (stopnia n), to w dowolnym punkcie x :

$$R(x) = f(x) - P(x) = \frac{1}{(n + 1)!} f^{(n+1)}(\xi) \prod_{i=0}^n (x - x_i), \quad (3.37)$$

gdzie ξ jest pewnym punktem z najmniejszego przedziału domkniętego $[a, b]$ zawierającego x, x_0, \dots, x_n .

Dowód:

Wystarczy przeprowadzić dowód w przypadku, gdy x nie jest węzłem interpolacji, gdyż w przeciwnym razie obie strony równości (3.37) są zerami. Oznaczmy $w(t) = \prod_{i=0}^n (t - x_i)$, $\psi(t) = f(t) - P(t) - \lambda w(t)$, gdzie λ jest liczbą dobraną tak, że $\psi(x) = 0$, a więc $\lambda = \frac{f(x) - P(x)}{w(x)}$.

Przy takim wyborze λ funkcja $\psi(t)$ jest funkcją klasy $C^{n+1}[a, b]$ przyjmującą wartość zero w $n + 2$ punktach x, x_0, x_1, \dots, x_n . Na mocy twierdzenia o wartości średniej (Rolle'a, dodatek D3) zastosowanego na przedziałach wyznaczonych przez kolejne zera funkcji $\psi(t)$, jej pochodna $\psi'(t)$ ma $n + 1$ różnych zer w (a, b) , $\psi''(t)$ ma n różnych zer w (a, b) , i tak dalej, aż ostatecznie $\psi^{(n+1)}(t)$ ma tam co najmniej jedno zero, które oznaczamy ξ . Ponieważ zachodzi $\psi^{(n+1)}(t) = f^{(n+1)}(t) - P^{(n+1)}(t) - \lambda w^{(n+1)}(t)$, więc uwzględniając, że $P(t)$ jest wielomianem stopnia n , zatem jego $n + 1$ -sza pochodna jest tożsamościowo równa zeru, a $w(t)$ jest wielomianem monicznym stopnia $n + 1$, więc jego $n + 1$ -sza pochodna jest stałą $(n + 1)!$. Możemy napisać $\psi^{(n+1)}(t) = f^{(n+1)}(t) - \lambda(n + 1)!$. Ze sposobu w jaki na wstępie wybraliśmy λ wynika $0 = \psi^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \lambda(n + 1)! = f^{(n+1)}(\xi) - \frac{f(x) - P(x)}{w(x)}(n + 1)!$ co jest równoważne (3.37).

Wyrażenie $R(x)$ jest często nazywane **resztą wzoru interpolacyjnego**. Postać wzoru (3.37) pozwala na analizę wpływu poszczególnych parametrów na jakość interpolacji.

Podstawowym problemem jest odpowiedź na pytanie, czy zwiększenie liczby wykorzystanych węzłów, a tym samym podniesienie stopnia n wielomianu interpolacyjnego poprawia jakość interpolacji, czyli zmniejsza resztę wzoru interpolacyjnego? Mogłaby to sugerować obecność wyrażenia $(n + 1)!$ w mianowniku wzoru (3.37), jednak od liczby węzłów zależą też pozostałe czynniki wzoru (3.37): $f^{(n+1)}(\xi)$ oraz $N(x) = \prod_{i=0}^n (x - x_i)$. Nawet w przypadku funkcji o „ładnym” przebiegu wysokie pochodne mogą osiągać duże wartości, a nawet być nieograniczone. Z kolei czynnik $N(x)$ zależy nie tylko od stopnia wielomianu interpolacyjnego, ale i od sposobu rozmieszczenia węzłów.

Załóżmy, że siatka węzłów została znormalizowana do przedziału $[-1, 1]$ i przypomnijmy wprowadzoną już normę

$$\|f\|_{[-1,1]} = \max_{x \in [-1,1]} |f(x)|. \quad (3.38)$$

Resztę wzoru interpolacyjnego na przedziale $[-1, 1]$ można oszacować w następujący sposób:

$$\|f(x) - P(x)\|_{[-1,1]} \leq \frac{1}{(n+1)!} \|f^{(n+1)}(x)\|_{[-1,1]} \left\| \prod_{i=0}^n (x - x_i) \right\|_{[-1,1]}. \quad (3.39)$$

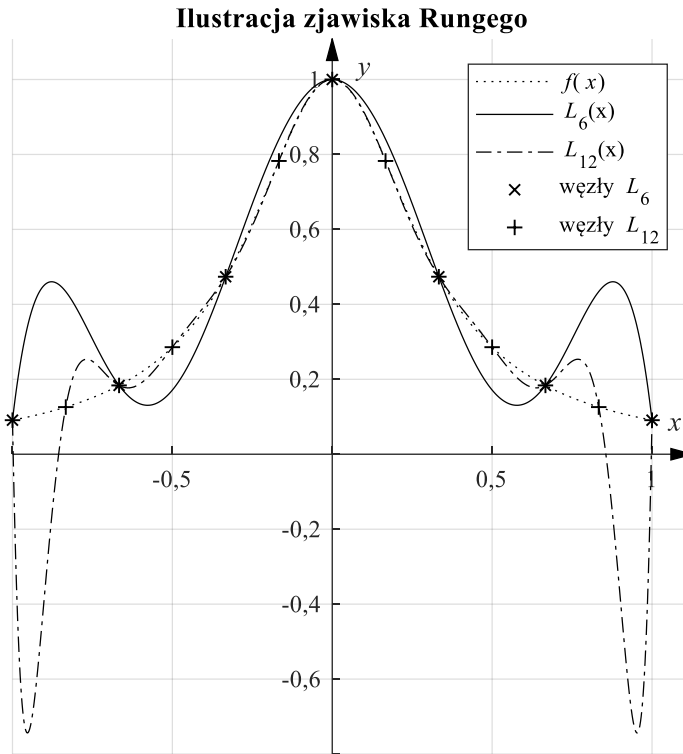
$N(x) = \prod_{i=0}^n (x - x_i)$ jest monicznym wielomianem stopnia $n + 1$.

Zgodnie z twierdzeniem 3.3 o własności min-max wielomianów Czebyszewa, dla dowolnych x_i zachodzi nierówność $\|N(x)\|_{[-1,1]} \geq 2^{-n}$ i norma $\|N(x)\|_{[-1,1]}$ osiąga wartość najmniejszą gdy $\prod_{i=0}^n (x - x_i) = 2^{-n} T_{n+1}(x)$, to jest gdy x_i są pierwiastkami wielomianu $2^{-n} T_{n+1}(x)$, czyli $x_i = \cos \frac{(2i+1)\pi}{2(n+1)}$, $i = 0, 1, \dots, n$.

Wynika stąd, że wybór węzłów interpolacji, które są pierwiastkami wielomianu Czebyszewa stopnia $n + 1$, minimalizuje $\|N(x)\|_{[-1,1]}$, a tym samym niekorzystny wpływ czynnika $N(x)$ na jakość interpolacji. Wybór węzłów położonych równoodlegle jest zdecydowanie mniej korzystny, co pokazują poniższe przykłady.

Przykład 3.3 (Rungego)

Poddano interpolacji funkcję $f(x) = \frac{1}{1+10x^2}$ na przedziale $[-1, 1]$ wielomianami stopnia 6 i 12 z węzłami równoodległymi. Wyniki interpolacji zostały pokazane na rysunku 3.8.



Rys. 3.8. Wielomiany interpolacyjne stopnia 6 i 12 funkcji $f(x) = \frac{1}{1+10x^2}$

Widać, że o ile w środku przedziału wielomiany interpolacyjne dobrze odwierciedlają przebieg funkcji interpolowanej, o tyle blisko krańców przedziału przyjmują dość odległe wartości, i zjawisko to nasila się ze wzrostem stopnia wielomianu.

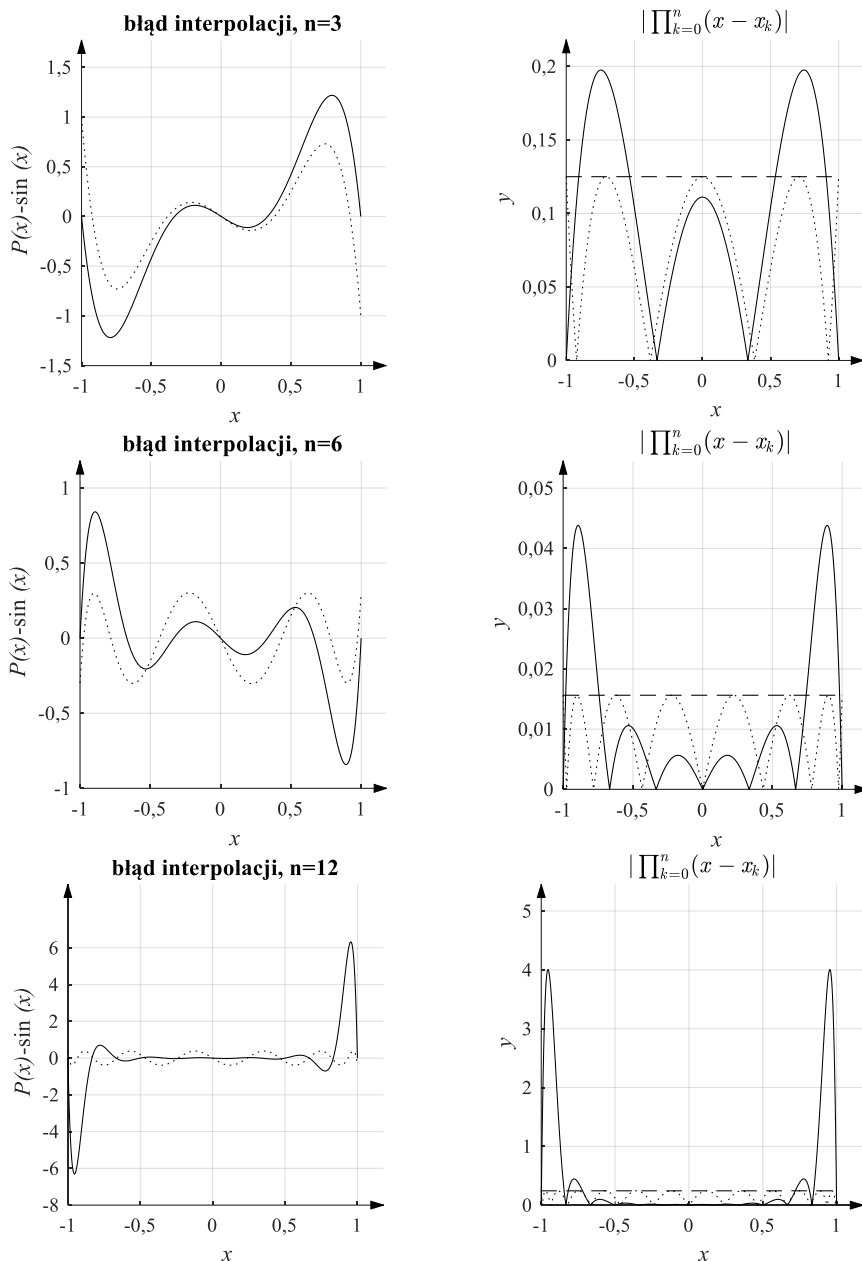
Przykład 3.4

Znajdziemy wielomian interpolacyjny funkcji $\sin(x)$ stosując węzły $-1, -1/3, 1/3, 1$ i oszacujemy błąd interpolacji.

x	-1	$-1/3$	$1/3$	1
$y = \sin x$	$-0,8414709848079$	$-0,3271946967962$	$0,3271946967962$	$0,8414709848079$

Dowolna metoda interpolacji wyznacza wielomian

$$P(x) = -0,1576272437781 x^3 + 0,9990982285860 x.$$



Rys. 3.9. Porównanie interpolacji funkcji sinus wielanami różnych stopni, z węzłami równoodległymi (linia ciągła) i węzłami Czebyszewa I rodzaju (linia kropkowa). Z lewej błąd całkowity, z prawej wartość $|\prod_{i=0}^n (x - x_i)|$, a poziome linie kreskowe odpowiadają wartości 2^{-n}

Zgodnie ze wzorem (3.37) zachodzi: $\sin(x) - P(x) = \frac{1}{4!} \sin^{(4)}(\xi) \prod_{i=0}^3 (x - x_i)$.
 Obliczymy kolejne pochodne: $\sin^{(4)}(\xi) = \cos^{(3)}(\xi) = (-\sin(\xi))'' = (-\cos(\xi))' = \sin(\xi)$, a więc:

$$|\sin(x) - P(x)| \leq \frac{1}{4!} \max_{-1 \leq \xi \leq 1} |\sin(\xi)| \max_{-1 \leq x \leq 1} \left| \prod_{i=0}^3 (x - x_i) \right|,$$

$$|\sin(x) - P(x)| \leq \frac{1}{24} 0,8414709848079 \cdot 0,1975308641975$$

$$\leq 0,006925687117760.$$

Oszacowanie to jest konserwatywne, bo w rzeczywistości

$$\max_{-1 \leq x \leq 1} |\sin(x) - P(x)| \approx 0,001218168482132.$$

Stosując węzły Czebyszewa, dostajemy:

$$P(x) = -0,1585048936070 x^3 + 0,9989828358483 x.$$

Analogiczne szacowanie daje nierówność

$$|\sin(x) - P(x)| \leq \frac{1}{24} 0,8414709848079 \cdot 0,125 \leq 0,004382661379208.$$

Ponownie jest to konserwatywne oszacowanie, bo w rzeczywistości

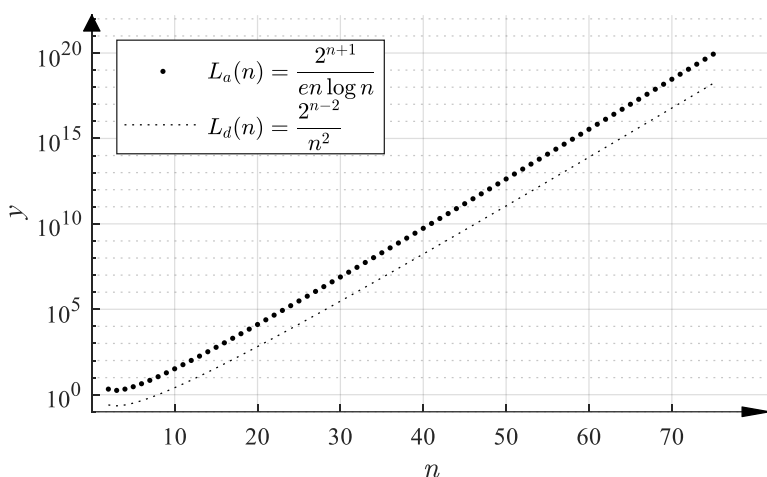
$$\max_{-1 \leq x \leq 1} |\sin(x) - P(x)| \approx 9,930425666558739 \cdot 10^{-4}.$$

Jak widać, użycie węzłów Czebyszewa poprawia oszacowanie błędu interpolacji, ale w obu przypadkach jest ono konserwatywne.

Zaobserwowany w powyższych przykładach gwałtowny wzrost reszty wzoru interpolacyjnego (czyli pogorszenie jakości interpolacji) na krańcach przedziału zawierającego równoodległe rozmieszczone węzły jest nazywany **zjawiskiem Rungego** (zilustrowanym na rysunku 3.8). Można mu przeciwdziałać, zagęszczając węzły na krańcach przedziału.

Interpolacja – przynajmniej z węzłami równoodległymi – za pomocą wielomianu wysokiego stopnia ($n > 5$) ma istotne wady. Należy do nich także złe uwarunkowanie zależności między wartościami w węzłach a współczynnikami otrzymanego wielomianu.

Maksymalna zmiana wartości wielomianu interpolacyjnego (między węzłami), spowodowana zmianą wartości interpolowanej funkcji w węzle może być co najmniej $L_a(n) = \frac{2^{n-2}}{n^2}$, a asymptotycznie (kiedy n rośnie do ∞) nawet $L_a(n) = \frac{2^{n+1}}{ne \log n}$ razy większa.



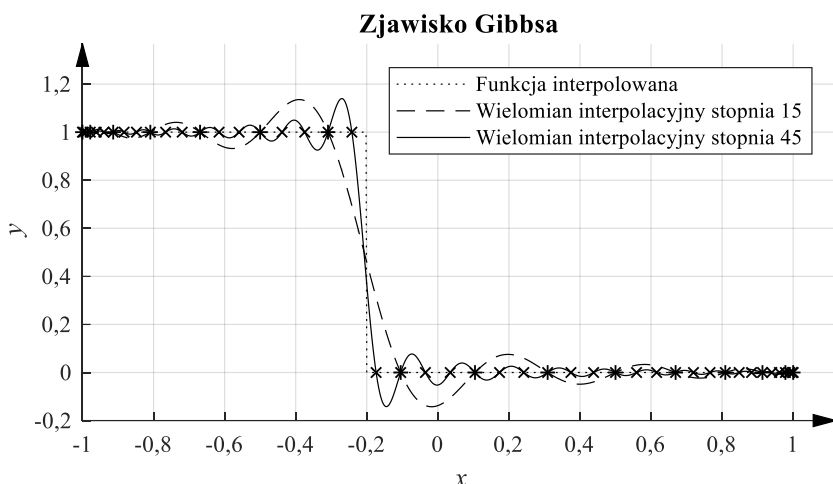
Rys. 3.10. Wartości y wyrażenia $L_d(n)$ oraz $L_a(n)$ będących oszacowaniami krotności stosunku maksymalnej zmiany wartości wielomianu interpolacyjnego z węzłami równoodległymi do maksymalnej zmiany wartości funkcji w węzłach interpolacji

Przykład 3.5

Innym fenomenem związanym z interpolacją jest tzw. zjawisko Gibbsa występujące przy interpolacji funkcji nieciągłej. Ponieważ wielomiany są funkcjami gładkimi nietrudno przewidzieć, że próba interpolacji funkcji nieciągłej może prowadzić do niepowodzenia. Na rysunku 3.11 przedstawiono interpolację funkcji

$$f(x) = \begin{cases} 1 & \text{dla } x < -\frac{1}{5} \\ 0 & \text{dla } x \geq -\frac{1}{5} \end{cases} \text{ wielomianami stopnia 15 i 45 z węzłami Czebyszewa}$$

II rodzaju (ekstremum wielomianu Czebyszewa). Widać wyraźnie, że błąd w pobliżu nieciągłości nie maleje ze wzrostem stopnia wielomianu interpolacyjnego, a jedynie rejon niezgodności przybliża się do punktu nieciągłości. **Zjawisko Gibbsa** dotyczy nie tylko interpolacji wielomianowej, ale także przybliżeń wielomianami trygonometrycznymi – rozwinięcia w szereg Fouriera.



Rys. 3.11. Ilustracja efektu Gibbsa w interpolacji wielomianowej

3.7. Odcinkowa interpolacja wielomianowa

Dobór jednego wielomianu na przedziale zawierającym wszystkie węzły stwarza trudności w przypadku liczby węzłów większej niż kilka. Alternatywnym rozwiązaniem jest podział przedziału interpolacji na podprzedziały zawierające niewielką liczbę węzłów i interpolacja na takim podprzedziale wielomianem niskiego stopnia.

Interpolacja wykorzystująca dwa sąsiednie węzły

Interpolacja wykorzystująca dwa węzły na każdym podprzedziale i liniowe wielomiany interpolacyjne pozwala na skonstruowanie ciągłej funkcji interpolującej, zdefiniowanej między parą sąsiednich węzłów wzorem:

$$x \in [x_i, x_{i+1}] \Rightarrow L(x) = f_i + (x - x_i) \frac{f_{i+1} - f_i}{x_{i+1} - x_i}. \quad (3.40)$$

Funkcja $L(x)$ jest ciągła, ale jej pochodna nie (jest stała między parą sąsiednich węzłów i zmienia się skokowo w węźle). Zgodnie ze wzorem (3.37), na każdym z podprzedziałów o długości h

$$\begin{aligned} \|f(x) - L(x)\|_{[x_i, x_{i+1}]} &\leq \frac{1}{2} \|f'''(x)\|_{[x_i, x_{i+1}]} \|(x - x_i)(x - x_{i+1})\|_{[x_i, x_{i+1}]} \\ &\leq \frac{h^2}{2} \|f'''(x)\|_{[x_i, x_{i+1}]} \cdot \end{aligned} \quad (3.41)$$

Wartość $|f(x) - L(x)|$ może zostać dowolnie zmniejszona przez zagęszczenie węzłów, ale interpolacja odcinkowo liniowa nie zachowa gładkiego kształtu funkcji $f(x)$.

Jeżeli w węzłach są znane nie tylko wartości interpolowanej funkcji $f_i = f(x_i)$, ale i jej pochodne $f'_i = f'(x_i)$, to można poszukać wielomianu $\varphi_i(x)$, który dla węzłów x_i, x_{i+1} spełni warunki

$$\varphi_i(x_i) = f_i, \varphi_i(x_{i+1}) = f_{i+1}, \varphi'_i(x_i) = f'_i, \varphi'_i(x_{i+1}) = f'_{i+1}. \quad (3.42)$$

To cztery równania, czyli wielomian $\varphi_i(x)$ musi mieć co najmniej 4 współczynniki, więc musi być wielomianem sześciennym. Taki sposób interpolacji to odcinkowa interpolacja **sześciennymi wielomianami Hermite'a**. Otrzymana funkcja interpolująca

$$x \in [x_i, x_{i+1}] \Rightarrow L(x) = \varphi_i(x) \quad (3.43)$$

ma ciągłą pochodną w całym przedziale interpolacji (x_0, x_n) .

Interpolacja za pomocą sześciennych funkcji sklejaných

Przy wykorzystaniu odcinkowej interpolacji wielomianami sześciennymi można uzyskać funkcję interpolacyjną, której nawet druga pochodna będzie ciągła.

Rozpatrzmy $n + 1$ węzłów $x_i, i = 0, \dots, n$ dzielących przedział $[x_0, x_n]$ na n podprzedziałów $[x_i, x_{i+1}]$. Skonstruujemy rodzinę n wielomianów sześciennych $\varphi_i(x), x \in [x_i, x_{i+1}], i = 0, \dots, n - 1$. Musimy więc wyznaczyć $4n$ współczynników wielomianów $\varphi_i(x)$, a w tym celu potrzebujemy $4n$ równań:

- warunki interpolacji $\varphi_i(x_i) = f_i, i = 0, \dots, n$ dają $n + 1$ równań,
- warunki równości wielomianów w węzłach wewnętrznych $\varphi_i(x_{i+1}) = \varphi_{i+1}(x_{i+1}), i = 0, \dots, n - 2$ dają $n - 1$ równań,
- warunki zgodności pochodnych w węzłach wewnętrznych $\varphi'_i(x_{i+1}) = \varphi'_{i+1}(x_{i+1}), \varphi''_i(x_{i+1}) = \varphi''_{i+1}(x_{i+1}), i = 0, \dots, n - 2$ dają $2(n - 1)$ równań,

mamy więc $4n - 2$ równań. Brakujące 2 równania trzeba narzucić, na przykład zakładając, że $\varphi''_0(x_0) = \varphi''_{n-1}(x_n) = 0$ albo w przypadku gdy $f_0 = f_n$ potraktować skrajne węzły x_0, x_n jak węzeł wewnętrzny, lub narzucić warunki ciągłości trzeciej pochodnej w wybranych węzłach.

Wszystkie warunki zebrane razem prowadzą do układu $4n$ równań liniowych.

W zastosowaniach, w których jest bardzo wiele węzłów może być wygodne posłużyć się innym przedstawieniem funkcji sześciennych niż za pomocą czterech współczynników. Można na przykład zastosować zależność

$$s(x_{j-1} + h_j t) = (1-t) \left(s_{j-1} + \frac{1}{6} h_j^2 t ((1-t)^2 - 1) s_{j-1}'' \right) + t \left(s_j + \frac{1}{6} h_j^2 (t^2 - 1) s_j'' \right) \quad (3.44)$$

z warunkami ciągłości

$$h_j s_{j-1}'' + 2(h_j + h_{j+1}) s_j'' + h_{j+1} s_{j+1}'' = 6 \left(\frac{1}{h_{j+1}} (s_{j+1} - s_j) - \frac{1}{h_j} (s_j - s_{j-1}) \right), \quad (3.45)$$

gdzie $s_j, j = 0, 1, \dots, n$ są (znanymi) wartościami funkcji sklejaney (a także interpolowanej) w węzłach, $s_j'', j = 0, 1, \dots, n$ (nieznanymi) wartościami drugiej pochodnej funkcji sklejaney w węzłach, $h_j = x_j - x_{j-1}, j = 1, \dots, n$, a $t \in [0, 1]$, $t = \frac{x - x_{j-1}}{h_j}$. Takie podejście redukuje liczbę nieznanymi parametrów i równań do rozwiązania (do równań ciągłości), a złożoność obliczeniowa tego zadania – pod warunkiem uwzględnienia specjalnej postaci równań (3.45) – jest proporcjonalna do n .

Interpolacja przy pomocy **funkcji sklejaneych** może być uogólniona na wielomiany wyższego stopnia spełniające warunki ciągłości pochodnych wyższego rzędu.

Interpolacja funkcjami sklejanymi oferuje zachowanie kształtu, wymaga umiarkowanych nakładów obliczeniowych, daje funkcje interpolujące, które zbiegają do funkcji interpolowanej $f(x)$ przy zagęszczaniu węzłów, a także jej pierwsza i druga pochodna zbiega do pierwszej i drugiej pochodnej funkcji $f(x)$. Te wszystkie zalety powodują, że interpolacja funkcjami sklejanymi znajduje bardzo liczne zastosowania.

Błąd interpolacji sześciennymi funkcjami sklejanymi opisuje twierdzenie 3.10.

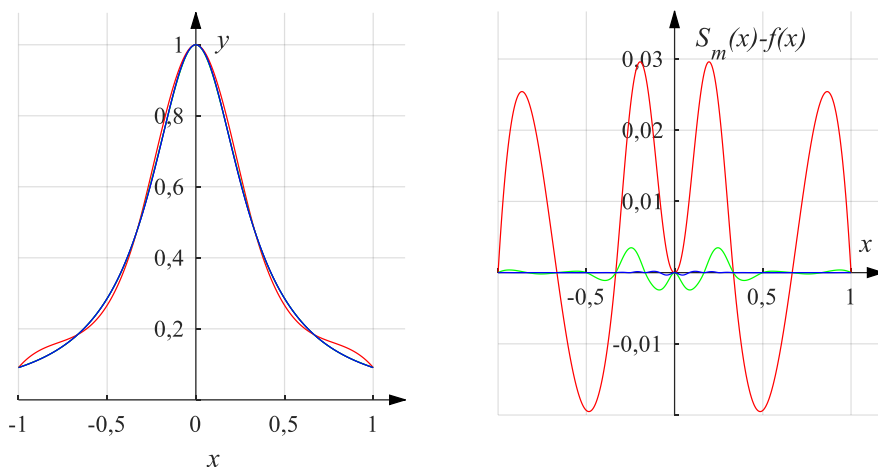
Twierdzenie 3.10 (Jankowska, Jankowski, 1981)

Jeżeli funkcja $f : [a, b] \rightarrow R$ ma drugą pochodną ciągłą i ograniczoną na przedziale $[a, b]$ oraz $M = \max_{x \in [a, b]} f''(x)$, $a = x_0 < x_1 < x_2 < \dots < x_{N-1} < x_N = b$ jest podziałem przedziału $[a, b]$, $s(\cdot)$ funkcją sklejaną sześcienną interpolującą $f(\cdot)$ w węzłach x_0, x_1, \dots, x_N , to zachodzi

$$\max_{x \in [a, b]} |s(x) - f(x)| \leq 5M \max_{i \in \{1, 2, \dots, N\}} (x_i - x_{i-1})^2. \quad (3.46)$$

Przykład 3.6

Dla tej samej funkcji, na której zademonstrowane zostało zjawisko Rungego obliczono przybliżenia interpolacyjne funkcjami sklejanymi sześciennymi $S_m(x)$ z $m = 7, 13$ i 25 równoodległymi węzłami, przyjmując zerową wartość drugiej pochodnej na krańcach przedziału. Na rys. 3.11 przedstawiono wykresy tych przybliżeń oraz błędów interpolacji.



Rys. 3.12. Przybliżenia funkcjami sklejanymi sześciennymi (po lewej) i ich błędy (po prawej). Linia czerwona dla $m = 7$ węzłów, zielona dla $m = 13$ węzłów, niebieska dla $m = 25$ węzłów równoodległych

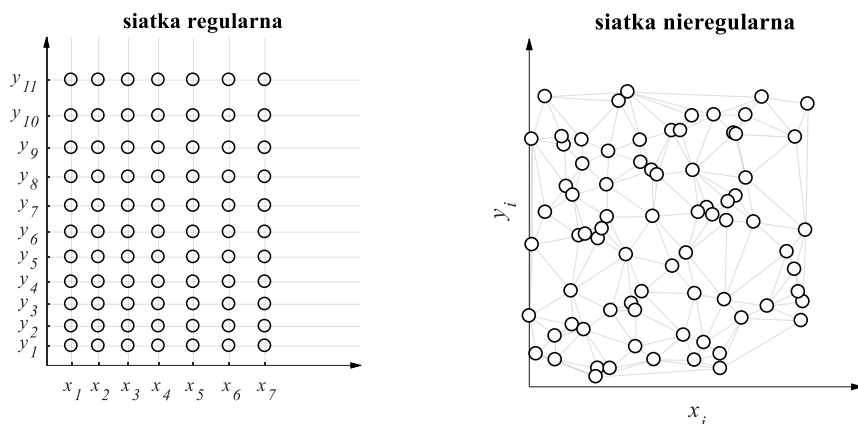
Jak widać przybliżenie funkcją sklejaną nie wykazuje efektów analogicznych do zjawiska Rungego: już dla 7 węzłów przybliżenie jest dokładniejsze niż wielomianem 6 stopnia, a dla 13 i 25 węzłów wykres funkcji sklejaney pokrywa się z funkcją interpolowaną. Wykresy błędów interpolacji pokazują bardzo szybką zbieżność przybliżenia.

3.8. Interpolacja funkcji wielu zmiennych

W pewnych zastosowaniach konieczne jest przybliżenie funkcji o wielu argumentach. Dokładniej przedstawimy koncepcję interpolacji wielomianowej funkcji dwu zmiennych – w przypadku większej liczby wymiarów wykorzystuje się analogiczne metody.

Definicja 3.4

Jednomianem dwóch zmiennych x, y jest nazywane wyrażenie postaci $x^i y^j$, gdzie i, j są dowolnymi, nieujemnymi liczbami całkowitymi. Stopniem jednomianu jest nazywana liczba $d = i + j$. Dowolna kombinacja liniowa jednomianów jest nazywana wielomianem zmiennych x, y , a stopniem wielomianu jest nazywany największy stopień jednomianu wchodzącego w jego skład.



Rys. 3.13. Ilustracja regularnej i nieregularnej dwuwymiarowej siatki węzłów. Węzły oznaczono kółeczkami. W przypadku siatki nieregularnej zaznaczono triangulację obszaru

Jeżeli dwuwymiarowa siatka węzłów jest regularna (rys 3.13), to jest:

- węzłami są punkty zlokalizowane w prostokącie $[x_0, x_n] \times [y_0, y_m]$ o współrzędnych (x_i, y_j) , $i = 0, \dots, n$, $j = 0, \dots, m$,
- wartościami interpolowanej funkcji w węzłach są $f_{i,j} := f(x_i, y_j)$,

to metody interpolacji jednowymiarowej można dość łatwo uogólnić. Metody interpolacji wykorzystywały bazę wielomianów $\varphi_i(x)$. Dla metody Vandermonde'a były to funkcje (3.23), Lagrange'a – (3.28), Newtona – (3.29). Baza $(m + 1)(m + 1)$ wielomianów dwu zmiennych postaci

$$\omega_{i,j}(x, y) := \varphi_i(x)\varphi_j(y) \quad (3.47)$$

pozwala na uogólnienie metod interpolacji jednowymiarowej. Na przykład dwuwymiarowy wzór Lagrange'a ma postać:

$$P(x, y) = \sum_{i,j} f_{i,j} \prod_{\substack{k=0 \\ k \neq i}}^n \frac{x - x_k}{x_i - x_k} \prod_{\substack{k=0 \\ k \neq j}}^m \frac{y - y_k}{y_j - y_k}. \quad (3.48)$$

Wielomian

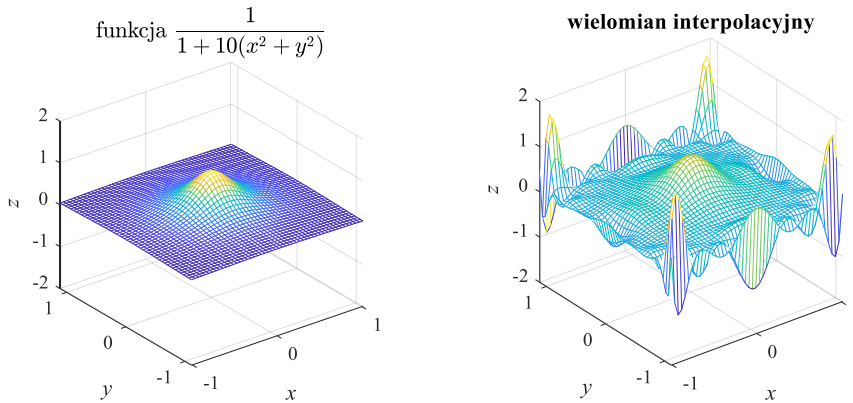
$$\omega_{i,j}(x,y) := \prod_{\substack{k=0 \\ k \neq i}}^n \frac{x - x_k}{x_i - x_k} \prod_{\substack{k=0 \\ k \neq j}}^m \frac{y - y_k}{y_j - y_k} \quad (3.49)$$

jest wielomianem stopnia co najwyżej $n \cdot m$, przyjmuje wartość 1 w węźle o współrzędnych (x_i, y_j) , i wartość 0 we wszystkich pozostałych węzłach.

Oczywiście wielomianowa interpolacja wielowymiarowa nie jest wolna od zjawiska Rungego.

Przykład 3.7

Dokonano interpolacji funkcji $f(x,y) = \frac{1}{1+10(x^2+y^2)}$ na regularnej siatce równoodległych węzłów. Na obszarze prostokątnym $[-1; 1] \times [-1,1; 1,1]$ użyto 9 węzłów w osi x i 13 węzłów w osi y . Wielomian interpolacyjny zgodny ze wzorem (3.48) obliczono następnie na siatce 50×50 równoodległych węzłów na prostokącie $[-1,01; 1,01] \times [-1,11; 1,11]$. Wynik pokazano na rysunku 3.13. Widoczny jest wyraźnie efekt analogiczny jak w przypadku jednowymiarowym.



Rys. 3.14. Zjawisko Rungego przy interpolacji funkcji dwóch zmiennych

Na liczbę wymiarów większą niż jeden można też przenieść metody interpolacji odcinkowej. Podobnie jak w przypadku jednowymiarowym funkcja interpolująca powstaje wtedy przez połączenie wielu wielomianów, każdy z nich „odpowiada” za jeden obszar elementarny. W przypadku jednowymiarowym takim obszarem elementarnym był przedział.

W przypadku danych dwuwymiarowych odpowiednikiem przedziału będzie prostokąt $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$. Nie istnieje płaszczyzna przechodząca przez 4 dowolne punkty przestrzeni, więc wielomianem spełniającym warunki interpolacji

będzie iloczyn dwóch wielomianów liniowych (funkcja biliniowa), czyli wielomian postaci:

$$p(x, y) = ax + by + cxy + d. \quad (3.50)$$

Cztery współczynniki takiego wielomianu trzeba wyznaczyć z czterech warunków

$$\begin{aligned} p(x_i, y_j) &= f_{i,j}, \\ p(x_{i+1}, y_j) &= f_{i+1,j}, \\ p(x_i, y_{j+1}) &= f_{i,j+1}, \\ p(x_{i+1}, y_{j+1}) &= f_{i+1,j+1}. \end{aligned} \quad (3.51)$$

W przypadku dwuwymiarowej interpolacji Hermite'a (wielomianami bi-sześciennymi) dysponujemy informacją o wartościach funkcji i jej pochodnych cząstkowych $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial^2 f}{\partial x \partial y}$. Konstruujemy wielomian

$$p(x, y) = \sum_{i=0}^3 \sum_{j=0}^3 a_{i,j} x^i y^j, \quad (3.52)$$

który ma 16 współczynników $a_{i,j}$. Dysponujemy:

- czterema równaniami (3.51) dla wartości wielomianu w wierzchołkach,
- czterema równaniami dla wartości każdej z trzech pochodnych $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial^2 f}{\partial x \partial y}$.

Łącznie mamy więc 16 równań i 16 niewiadomych, co pozwala na wyznaczenie równania powierzchni bi-sześciennej. Otrzymana funkcja interpolująca będzie miała ciągłe pochodne $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial^2 f}{\partial x \partial y}$ we wszystkich wewnętrznych węzłach regularnej siatki.

Jeżeli siatka węzłów nie jest regularna (rys. 3.13), to nie można rozdzielić zmiennych tak jak w bi-wielomianach. Obszarem elementarnym nie może być prostokąt. Można wtedy wykorzystać podział obszaru interpolacji na trójkąty (**triangulacja**), z których każdy wyznaczony jest przez trzy węzły tworzące jego wierzchołki. Każdy wielokąt można przedstawić w postaci sumy (mnogościowej) takich trójkątów. Trzy punkty w przestrzeni wyznaczają jednoznacznie płaszczyznę, można więc dla każdego z elementarnych trójkątów zbudować wielomian liniowy

$$p(x, y) = ax + by + c, \quad (3.53)$$

o trzech współczynnikach spełniających trzy warunki interpolacji w trzech wierzchołkach A, B, C o współrzędnych $(x_A, y_A), (x_B, y_B), (x_C, y_C)$:

$$p(x_A, y_A) = f(x_A, y_A), p(x_B, y_B) = f(x_B, y_B), p(x_C, y_C) = f(x_C, y_C). \quad (3.54)$$

W przypadku wielomianów stopnia $n > 1$ trzeba wyznaczyć większą liczbę współczynników (dla $n = 2$ jest to 6 współczynników, dla $n = 3 - 10$) i dodatkową swobodę można wykorzystać do zapewnienia ciągłości pochodnych funkcji interpolującej.

3.9. Obliczanie wartości wielomianu

Wyznaczenie wartości wielomianu wysokiego stopnia we wskazanym punkcie x może być zadaniem stwarzającym trudności numeryczne.

Zastosowanie **postaci potęgowej** $P(x) = c_n x^n + c_{n-1} x^{n-1} + \dots + c_1 x + c_0$ wymaga nie tylko wykonania n mnożeń przez współczynniki, ale także obliczania (wysokich) potęg x . Daje to $1 + 2 + \dots + n = \frac{n(n+1)}{2}$ mnożeń oraz n dodawań.

Metoda (schemat) Hornera bazuje na przedstawieniu wielomianu w innej formie. Wielomian $P(x)$ stopnia 3 można zapisać w postaci:

$$P(x) = c_3 x^3 + c_2 x^2 + c_1 x + c_0 = (c_3 x + c_2)x^2 + c_1 x + c_0 = ((c_3 x + c_2)x + c_1)x + c_0. \quad (3.55)$$

Prowadzi to do rekurencji:

$$b_3 := c_3, b_2 := b_3 x + c_2, b_1 := b_2 x + c_1, b_0 := b_1 x + c_0 = P(x). \quad (3.56)$$

Dla wielomianu stopnia n można zapisać:

$$b_n := c_n, \dots, b_{i-1} := b_i x + c_{i-1}, \dots, b_0 := b_1 x + c_0 = P(x). \quad (3.57)$$

Ten algorytm wymaga tylko n mnożeń oraz n dodawań.

Jeżeli znane są pierwiastki (miejsca zerowe) wielomianu $x_i, i = 1, \dots, n$, to najbardziej dogodną formą do wyznaczenia wartości wielomianu jest postać iloczynowa

$$P(x) = c_n (x - x_1) \dots (x - x_n). \quad (3.58)$$

Na ogół jednak pierwiastki wielomianu nie są znane i dlatego najdogodniejszą metodą obliczania wartości wielomianu interpolacyjnego jest tzw. **formuła barycentryczna**, wywodząca się wprost ze wzoru interpolacyjnego Lagrange'a. Jeśli we wzorze (3.27) wyciągniemy przed znak sumy iloczyn $\ell(x) = \prod_{i=0}^n (x - x_i)$, to otrzymamy

$$P(x) = \ell(x) \sum_{k=0}^n f_k \frac{\omega_k}{x - x_k}, \quad (3.59)$$

gdzie

$$\omega_k = \frac{1}{\prod_{i=0, i \neq k}^n (x_i - x_k)}, \quad (3.60)$$

czyli **pierwszą formę formuły barycentrycznej**. Drugą postać (**drugą formułą barycentryczną**) uzyskujemy, zauważając że

$$\ell(x) \sum_{k=0}^n \frac{\omega_k}{x - x_k} = 1, \quad (3.61)$$

więc

$$P(x) = \frac{\sum_{k=0}^n f_k \frac{\omega_k}{x - x_k}}{\sum_{k=0}^n \frac{\omega_k}{x - x_k}}. \quad (3.62)$$

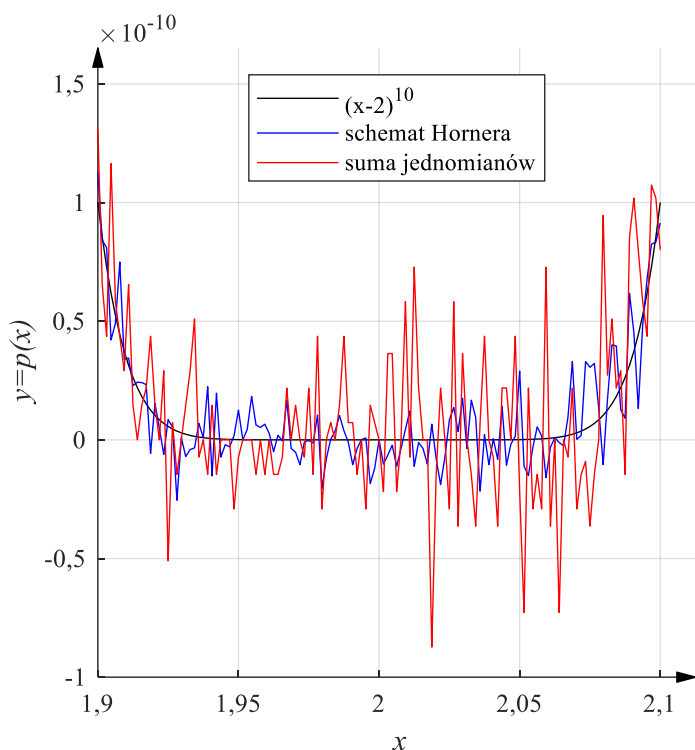
Na pierwszy rzut oka wzór (3.62) wygląda na przedstawiający funkcję wymierną, a co gorsza traci sens, gdy argument jest jednym z węzłów. We współczesnych maszynach cyfrowych implementujących standard IEEE754/854 można łatwo zidentyfikować taki przypadek, testując, czy wyrażenie $\frac{\omega_k}{x-x_k}$ stało się nieskończonością. Jeśli tak się stało dla pewnego k , to w miejsce wyniku formuły barycentrycznej należy przyjąć wartość funkcji interpolowanej w odpowiednim węźle f_k . Jak pokażemy w przykładzie formuła barycentryczna jest bezkonkurencyjnym sposobem obliczania wartości wielomianu interpolacyjnego.

Przykład 3.8

Obliczono wielomian $p(x) = (x - 2)^{10} = x^{10} - 20x^9 + 180x^8 - 960x^7 + 3360x^6 - 8064x^5 + 13440x^4 - 15360x^3 - 5120x^2 + 1024 =$

$$= \left(\left(\left(\left(\left(\left(\left(\left(\left(\left((x - 20)x + 180 \right) x - 960 \right) x + 3360 \right) x - 8064 \right) x + 13440 \right) x - 15360 \right) x - 5120 \right) x + 1024 \right) \right) \right) \right) \right) \right) \right) \right) x$$

dla 129 wartości argumentu w otoczeniu liczby 2, wykorzystując wszystkie trzy równoważne matematycznie postacie wielomianu. Wynik przedstawia rys. 3.15. Z punktu widzenia obliczeń numerycznych zastosowane postaci wielomianu nie są równoważne: wariant pierwszy (postać iloczynowa) jest bezkonkurencyjny, ale na ogół niepraktyczny w przypadku wielomianów będących wynikiem aproksymacji czy interpolacji, gdzie pierwiastki nie są znane. Z kolei sumowanie jedno- mianów okazuje się zupełnie niepraktyczne: przy „naiwnej” implementacji wymaga największej ilości operacji, a przy tym jest źródłem największych błędów zaokrągleń. Schemat Hornera okazuje się użytecznym kompromisem.



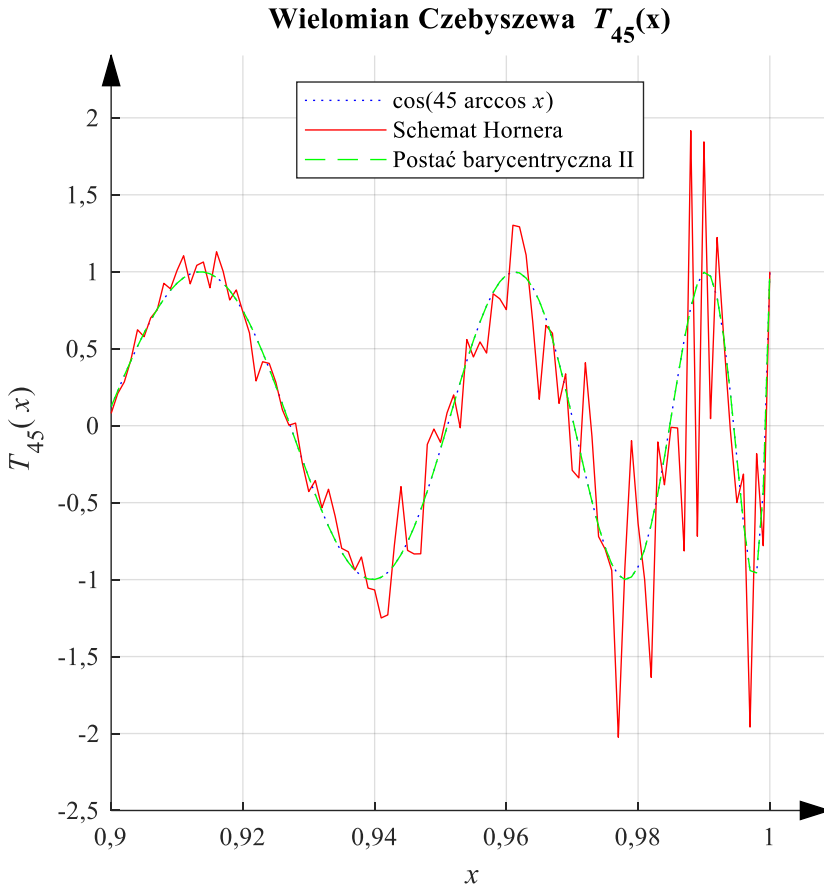
Rys. 3.15. Wartości wielomianu $p(x)$ obliczone z postaci iloczynowej, metodą Hornera i z postaci potęgowej

Przykład 3.9

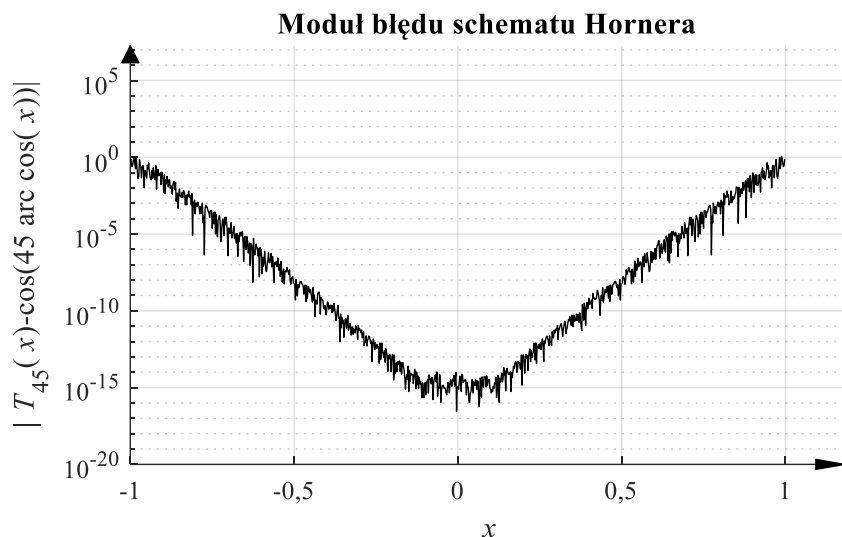
Obliczono wielomian Czebyszewa $T_{45}(x)$ stosując:

- schemat Hornera,
- drugą formę formuły barycentrycznej zastosowanej do węzłów Czebyszewa II rodzaju.

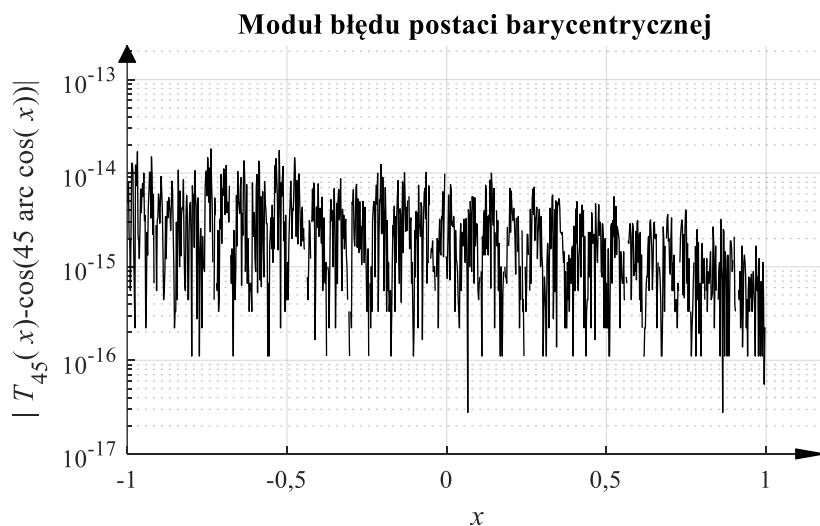
Wszystkie rezultaty porównano z wynikiem otrzymanym ze wzoru (3.13). Efekty eksperymentu przedstawia rys. 3.16.



Rys. 3.16. Wartości wielomianu Czebyszewa $y = T_{45}(x)$ obliczone ze wzoru $y = \cos(45 \arccos x)$, metodą Hornera i z postaci barycentrycznej. Jest to powiększony fragment wykresu dotyczący przedziału $[0,9; 1]$. Wykres w przedziale $[-1; -0,9]$ jest symetryczny. W przedziale $(-0,9; 0,9)$ błąd jest praktycznie niewidoczny na wykresie w skali liniowej



Rys. 3.17. Moduł błędu obliczenia wielomianu Czebyszewa $T_{45}(x)$ schematem Hornera wyrażony w skali logarytmicznej



Rys. 3.18. Moduł błędu obliczenia wielomianu Czebyszewa $T_{45}(x)$ formułą barycentryczną, wyrażony w skali logarytmicznej

Jak widać schemat Hornera blisko krańców przedziału wykazuje błąd przekraczający 1, czyli większy od normy supremum obliczanego wielomianu! Powodem są

błędy zaokrąglenia oraz duże wartości współczynników wielomianu (współczynnik wiodący jest równy 2^{44}). Formuła barycentryczna daje błąd na poziomie kilku epsilonów maszynowych. Formuła barycentryczna jest wzorem numerycznie poprawnym we wnętrzu przedziału interpolacji pod warunkiem, że jest stosowana na specjalnym układzie węzłów⁵, na przykład na węzłach Czebyszewa, albo Legendre'a.

Istnieją biblioteki pozwalające na obliczenia z użyciem wielomianów interpolacyjnych Czebyszewa do różnych języków do obliczeń naukowych (ChebFun do Matlab-a, pychebfun do Pythona, czy ApproxFun.jl do języka Julia) operujące wielomianami interpolacyjnymi stopnia dochodzącego do tysięcy, przy obliczeniach w zwykłej arytmetyce IEEE754/854 z użyciem typu double.

Wniosek wypływający stąd jest taki, że w przypadku, gdy konieczne jest stosowanie węzłów równoodległych lub prawie równoodległych odpowiednim wyborem są funkcje sklejące, ale jeśli można wybrać węzły interpolacji, to wybór węzłów Czebyszewa w połączeniu z formułą barycentryczną umożliwia wygodne i bezpieczne stosowanie wielomianów interpolacyjnych bardzo wysokiego stopnia.

⁵ Chodzi o węzły dające niewielką tzw. stałą Lebesgue'a. Przez stałą Lebesgue'a rozumiemy normę operatora, który funkcji ciągłej przyporządkowuje jej wielomian interpolacyjny, indukowaną przez normę supremum. Inaczej mówiąc stała Lebesgue'a obrazuje relację między maksymalną wartością funkcji a maksymalną wartością jej wielomianu interpolującego. W przypadku węzłów równoodległych rośnie ona bardzo szybko (wykładniczo) z liczbą węzłów.

4. Różniczkowanie numeryczne i ekstrapolacja Richardsona

4.1. Podstawowe wzory różniczkowania numerycznego

Jak wiadomo pochodna funkcji $f(x)$ w punkcie x_0 jest zdefiniowana jako

$$f'(x_0) = \lim_{h \rightarrow 0} \frac{f(x_0+h) - f(x_0)}{h}. \quad (4.1)$$

Podstawowe wzory służące do numerycznego przybliżania pochodnej na podstawie dyskretnych danych o wartościach funkcji $f(x)$ można (dla funkcji $f(x)$ mającej skończone pochodne dowolnego stopnia) wyprowadzić i przeanalizować na podstawie wzoru Taylora:

$$f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2}{2!}f''(x_0) + \frac{h^3}{3!}f^{(3)}(x_0) + \dots \quad (4.2)$$

Jeżeli w definicji pochodnej pominiemy operację obliczania granicy, otrzymamy przybliżenie pochodnej w punkcie x_0 nazywane **różnicą progresywną** (bo obliczamy różnicę wartości funkcji, wykonując krok „w przód” o długości h):

$$D_P(h) = \frac{f(x_0 + h) - f(x_0)}{h}. \quad (4.3)$$

Parametr $h > 0$ tego przybliżenia będzie decydował o wielkości błędu metody. Z przekształcenia wzoru Taylora otrzymujemy

$$D_P(h) = \frac{f(x_0+h) - f(x_0)}{h} = f'(x_0) + \frac{h}{2!}f''(x_0) + \frac{h^2}{3!}f^{(3)}(x_0) + \dots, \quad (4.4)$$

czyli różnica progresywna $D_P(h)$ składa się z dokładnej wartości pochodnej i błędu metody

$$ED_P(h) = D_P(h) - f'(x_0) = \frac{h}{2!}f''(x_0) + \frac{h^2}{3!}f^{(3)}(x_0) + \dots \quad (4.5)$$

Wartość przybliżona jest obliczana dla małych wartości h , znacznie mniejszych od 1, więc największym składnikiem błędu $ED_P(h)$ (tzw. częścią główną błędu) będzie składnik, w którym występuje h w najniższej potęgce, czyli $\frac{h}{2}f''(x_0)$. Błąd metody dąży do zera, gdy h dąży do zera i zbieżność ta jest (w przybliżeniu) liniowa. Mówiąc jeszcze inaczej, w przypadku różnicy progresywnej błąd metody jest w przybliżeniu proporcjonalny do h .

Jeżeli wykonamy krok wstecz o długości $h > 0$ przy obliczaniu różnicy wartości funkcji, to otrzymamy przybliżenie pochodnej w punkcie x_0 nazywane **różnicą wsteczną**:

$$D_B(h) = \frac{f(x_0) - f(x_0 - h)}{h}. \quad (4.6)$$

Jeżeli we wzorze Taylora zastąpimy h przez $-h$, to otrzymamy

$$f(x_0 - h) = f(x_0) - hf'(x_0) + \frac{h^2}{2!}f''(x_0) - \frac{h^3}{3!}f^{(3)}(x_0) + \dots, \quad (4.7)$$

a przekształcając ten wzór, można obliczyć błąd metody dla różnicy wstecznej

$$ED_B(h) = D_B(h) - f'(x_0) = -\frac{h}{2!}f''(x_0) + \frac{h^2}{3!}f^{(3)}(x_0) - \dots. \quad (4.8)$$

Częścią główną błędu jest $-\frac{h}{2}f''(x_0)$, jest to więc błąd przeciwnego znaku niż w różnicy progresywnej, ale także liniowy względem długości kroku h .

Jeżeli błędy metody dla różnicy progresywnej i wstecznej są przeciwnych znaków, to może, biorąc średnią z tych przybliżeń, dostaniemy dokładniejszy wynik? Przybliżenie pochodnej

$$D_C(h) = \frac{1}{2}(D_P(h) + D_B(h)) = \frac{f(x_0 + h) - f(x_0 - h)}{2h} \quad (4.9)$$

nazywane jest **różnicą centralną**. Błąd metody dla różnicy centralnej można wyznaczyć, korzystając z obu postaci szeregu Taylora (4.2) i (4.7):

$$\begin{aligned} D_C(h) &= \frac{f(x_0 + h) - f(x_0 - h)}{2h} = \\ &= \frac{1}{2h} \left\{ \left(f(x_0) + hf'(x_0) + \frac{h^2}{2!}f''(x_0) + \frac{h^3}{3!}f^{(3)}(x_0) + \dots \right) - \right. \\ &\quad \left. - \left(f(x_0) - hf'(x_0) + \frac{h^2}{2!}f''(x_0) - \frac{h^3}{3!}f^{(3)}(x_0) + \dots \right) \right\} = \\ &= f'(x_0) + \frac{h^2}{3!}f^{(3)}(x_0) + \frac{h^4}{5!}f^{(5)}(x_0) + \dots, \end{aligned} \quad (4.10)$$

więc

$$ED_C(h) = D_C(h) - f'(x_0) = \frac{h^2}{3!}f^{(3)}(x_0) + \frac{h^4}{5!}f^{(5)}(x_0) + \dots. \quad (4.11)$$

Błąd metody w przypadku różnicy centralnej jest w przybliżeniu proporcjonalny do h^2 , a więc dla tych samych h znacznie mniejszy niż dla różnicy progresywnej lub wstecznej, przy takim samym nakładzie obliczeń.

4.2. Numeryczne przybliżenie drugiej pochodnej

Wzór Taylora (4.2) dostarcza także informacji o wyższych pochodnych funkcji $f(x)$ w punkcie x_0 . Dla obliczenia drugiej pochodnej należy we wzorze (4.2) „pozbyć się” składnika zawierającego pierwszą pochodną. Na przykład tak: zastępując h przez $2h$, dostajemy

$$f(x_0 + 2h) = f(x_0) + 2hf'(x_0) + \frac{4h^2}{2!}f''(x_0) + \frac{8h^3}{3!}f^{(3)}(x_0) + \dots \quad (4.12)$$

Wzór (4.2) pomnożony przez 2

$$2f(x_0 + h) = 2f(x_0) + 2hf'(x_0) + 2\frac{h^2}{2!}f''(x_0) + 2\frac{h^3}{3!}f^{(3)}(x_0) + \dots \quad (4.13)$$

można odjąć od (4.12):

$$f(x_0 + 2h) - 2f(x_0 + h) = -f(x_0) + h^2f''(x_0) + 6\frac{h^3}{3!}f^{(3)}(x_0) + \dots, \quad (4.14)$$

czyli:

$$D_{2P}(h) := \frac{f(x_0+2h) - 2f(x_0+h) + f(x_0)}{h^2} = f''(x_0) + 6\frac{h}{3!}f^{(3)}(x_0) + \dots \quad (4.15)$$

Wyrażenie (4.15), zwane różnicą progresywną drugiego rzędu, jest przybliżeniem drugiej pochodnej funkcji $f(x)$ w punkcie x_0 , obciążonym błędem metody proporcjonalnym do h .

Podobne przekształcenia można wykonać, by otrzymać różnicę wsteczną i centralną drugiego rzędu:

$$D_{2B}(h) := \frac{f(x_0) - 2f(x_0-h) + f(x_0-2h)}{h^2}, \quad (4.16)$$

$$D_{2C}(h) := \frac{f(x_0+h) - 2f(x_0) + f(x_0-h)}{h^2}. \quad (4.17)$$

Podobnie jak w przypadku pierwszej pochodnej błąd metody dla $D_{2B}(h)$ jest w przybliżeniu proporcjonalny do h , a błąd metody dla $D_{2C}(h)$ do h^2 .

Przybliżenia drugiej pochodnej wymagają wykorzystania wartości funkcji w trzech kolejnych punktach, nakład obliczeń jest więc większy niż w przypadku wzorów przybliżających pierwsze pochodne.

4.3. Dokładniejsze wzory przybliżające pochodną

Powróćmy do wzoru (4.4) przybliżającego pochodną w punkcie x_0 różnicą progresywną. Wynika z niego:

$$f'(x_0) = \frac{f(x_0+h)-f(x_0)}{h} - \frac{h}{2!} f''(x_0) - \frac{h^2}{3!} f^{(3)}(x_0) - \dots \quad (4.18)$$

Z kolei z (4.15)

$$\frac{h}{2} f''(x_0) = \frac{f(x_0+2h)-2f(x_0+h)+f(x_0)}{2h} - 3 \frac{h^2}{3!} f^{(3)}(x_0) - \dots, \quad (4.19)$$

co podstawione do (4.18) daje:

$$f'(x_0) = \frac{f(x_0+h)-f(x_0)}{h} - \frac{f(x_0+2h)-2f(x_0+h)+f(x_0)}{2h} - 4 \frac{h^2}{3!} f^{(3)}(x_0) - \dots \quad (4.20)$$

Tak więc, przybliżenie pierwszej pochodnej $f'(x_0)$ przez

$$D_{P+}(h) = \frac{-f(x_0 + 2h) + 4f(x_0 + h) - 3f(x_0)}{2h} \quad (4.21)$$

jest obarczone błędem metody proporcjonalnym do h^2 . Wykorzystanie drugiej pochodnej we wzorze Taylora, a więc i większej liczby wartości funkcji, pozwoliło na zmniejszenie błędu metody z proporcjonalnego do h do proporcjonalnego do h^2 .

Podobny sposób postępowania można zastosować do wyprowadzenia dokładniejszych wersji wzoru z różnicą wsteczną i centralną. Ten ostatni ma postać

$$D_{C+}(h) = \frac{-f(x_0 + 2h) + 8f(x_0 + h) - 8f(x_0 - h) + f(x_0 - 2h)}{12h} \quad (4.22)$$

i jest obarczony błędem metody proporcjonalnym do h^4 .

Wyprowadzone wzory wymagają wykorzystania wartości funkcji w trzech lub czterech punktach, nakład obliczeń jest więc większy niż w przypadku podstawowych wzorów przybliżających pierwsze pochodne.

4.4. Różniczkowanie funkcji wielu zmiennych

W przypadku funkcji dwu zmiennych x, y przybliżenia pierwszych pochodnych cząstkowych w punkcie (x_0, y_0) dokonuje się tak samo jak dla funkcji jednej zmiennej i można wyprowadzić analogiczny zestaw wzorów. Na przykład wzory z zastosowaniem różnicy centralnej do przybliżenia pochodnych $\frac{\partial f(x,y)}{\partial x}$ i $\frac{\partial f(x,y)}{\partial y}$ w punkcie (x_0, y_0) to:

$$\begin{aligned} D_{Cx}(h) &= \frac{f(x_0+h, y_0) - f(x_0-h, y_0)}{2h}, \\ D_{Cy}(h) &= \frac{f(x_0, y_0+h) - f(x_0, y_0-h)}{2h}. \end{aligned} \quad (4.23)$$

Podobnie będzie z wzorami przybliżającymi drugie pochodne. Wariant wzorów z różnicą centralną przybliżających pochodne $\frac{\partial^2 f(x,y)}{\partial x^2}$ i $\frac{\partial^2 f(x,y)}{\partial y^2}$ w punkcie (x_0, y_0) to:

$$\begin{aligned} D_{2Cx}(h) &= \frac{f(x_0+h, y_0) - 2f(x_0, y_0) + f(x_0-h, y_0)}{h^2}, \\ D_{2Cy}(h) &= \frac{f(x_0, y_0+h) - 2f(x_0, y_0) + f(x_0, y_0-h)}{h^2}. \end{aligned} \quad (4.24)$$

Pozostaje jeszcze wyprowadzenie wzoru przybliżającego pochodną mieszaną $\frac{\partial^2 f(x,y)}{\partial x \partial y}$ w punkcie (x_0, y_0) . Niech h_x, h_y będą krokami w kierunku x i y odpowiednio. Wtedy

$$\begin{aligned} D_{2Cxy}(h_x, h_y) &= \frac{f(x_0+h_x, y_0+h_y) - f(x_0+h_x, y_0-h_y) - f(x_0-h_x, y_0+h_y) + f(x_0-h_x, y_0-h_y)}{4h_x h_y}, \end{aligned} \quad (4.25)$$

a po uproszczeniach:

$$D_{2Cxy}(h_x, h_y) = \frac{f(x_0+h_x, y_0+h_y) - f(x_0+h_x, y_0-h_y) - f(x_0-h_x, y_0+h_y) + f(x_0-h_x, y_0-h_y)}{4h_x h_y}. \quad (4.26)$$

W analogiczny sposób można postępować w przypadku funkcji większej liczby zmiennych.

4.5. Błędy zaokrągleń w różniczkowaniu numerycznym

Dotychczas wyprowadzono wzory pokazujące zachowanie błędu metody dla różnych przybliżeń pochodnych w funkcji parametru h . Błędy te maleją proporcjonalnie do h (dla różnicy progresywnej i wstecznej), proporcjonalnie do h^2 (dla różnicy centralnej), czy nawet do h^4 (wzór 4.22). Jednak w realnych obliczeniach obok błędu metody wystąpi błąd zaokrągleń. Na przykład we wzorze na różnicę centralną

$$D_C(h) = \frac{f(x_0 + h) - f(x_0 - h)}{2h} \quad (4.27)$$

wartości funkcji $f(x_0 + h)$ i $f(x_0 - h)$ będą obciążone błędem wynikającym z zaokrągleń wykonanych przy zmiennoprzecinkowych operacjach koniecznych od obliczenia wartości funkcji f . Jeśli przyjmiemy że błąd bezwzględny (wynikający z zaokrągleń) obliczenia każdej wartości funkcji f ma moduł nie większy od Δ_f , to nawet jeśli operacje odejmowania i dzielenia we wzorze (4.27) uznamy za dokładne, wartość $D_C(h)$ będzie obciążona błędem zaokrągleń $RD_C(h)$, którego moduł spełnia nierówność

$$|RD_C(h)| \leq \frac{2\Delta_f}{2h} = \frac{\Delta_f}{h}. \quad (4.28)$$

Tak więc, oszacowanie błędu wynikającego z zaokrągleń jest odwrotnie proporcjonalne do h i rośnie do nieskończoności dla $h \rightarrow 0!$ Jeśli h jest dostatecznie małe to liczby $f(x_0 + h)$ i $f(x_0 - h)$ mogą być zaokrąglone do tej samej liczby zmiennoprzecinkowej i stosowanie takich wartości h nie ma sensu.

Jak obliczono we wzorze (4.11) błąd metody $ED_C(h)$ spełnia

$$|ED_C(h)| \approx \frac{h^2}{6} |f^{(3)}(x_0)|. \quad (4.29)$$

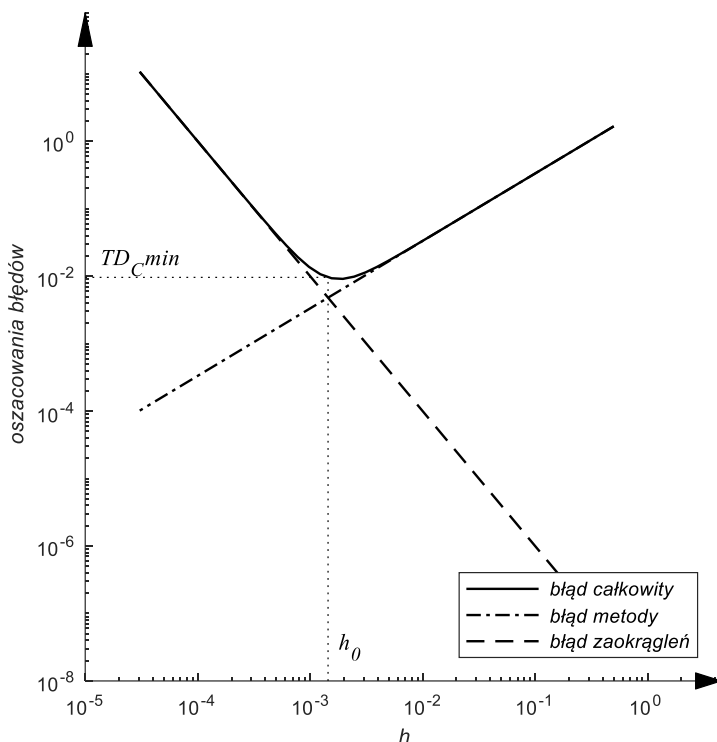
Jeżeli trzecia pochodna funkcji f jest ograniczona: $|f^{(3)}(x)| \leq M$, to można spodziewać się błędu całkowitego $TD_C(h)$, którego moduł spełnia

$$|TD_C(h)| \lesssim \frac{\Delta_f}{h} + \frac{h^2}{6} M. \quad (4.30)$$

Można wyznaczyć taką wartość parametru $h = h_0$, dla której oszacowanie błędu całkowitego będzie minimalne. Po przyrównaniu pochodnej prawej strony w (4.30) do zera, dostajemy

$$-\frac{\Delta_f}{h^2} + 2\frac{h}{6}M = 0 \implies h_0 = \sqrt[3]{\frac{3\Delta_f}{M}}, \quad (4.31)$$

co daje wartość oszacowania $TD_{Cmin} = 0,5 \sqrt[3]{9M\Delta_f^2}$. Przebieg oszacowań błędu zaokrągleń, błędu metody i błędu całkowitego dla różnicy centralnej dla $\Delta_f = 10^{-8}$, $M = 10$, $h = 2^{-i}$, $i = 1, 2, \dots, 15$ pokazano na rys. 4.1.

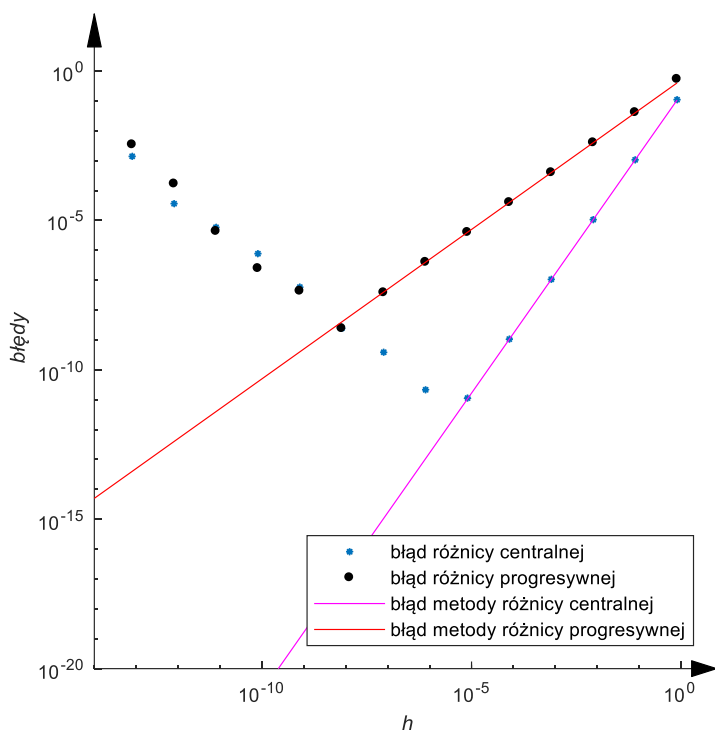


Rys. 4.1. Oszacowania błędów w różnicy centralnej

Oczywiście, wybór parametru h poniżej wartości h_0 nie ma sensu – mimo zmniejszenia błędu metody – błąd całkowity będzie większy niż dla $h = h_0$.

Wszystkie przybliżenia pochodnej wyprowadzone w tym rozdziale mają podobną strukturę: w liczniku występują obliczone wartości różniczkowanej funkcji, a mianownik maleje do zera dla $h \rightarrow 0$. Także w każdym przypadku błąd metody maleje do zera dla $h \rightarrow 0$. Przebieg błędu spowodowanego zaokrągleniami, błędu metody i błędu całkowitego będzie miał podobny charakter do przebiegu wyznaczonego dla różnicy centralnej. W każdym przypadku wybór parametru h poniżej pewnej wartości h_0 spowoduje wzrost błędu całkowitego.

Na rys. 4.2 przedstawiono błędy w przybliżonym obliczaniu pochodnej funkcji $f(x) = e^x$ w punkcie $x_0 = 1$, za pomocą różnicy progresywnej i centralnej.



Rys. 4.2. Błędy w numerycznym obliczaniu pochodnej funkcji $f(x) = e^x$ w punkcie $x_0 = 1$

Takie zachowanie błędów zaokrągleń, błędów metody i błędów całkowitych jest typowe dla wielu problemów numerycznych. Sposób przedstawiony w następnym rozdziale pozwala na poprawę dokładności wyniku bez znacznego zwiększania nakładu obliczeń.

4.6. Iterowana ekstrapolacja Richardsona

Zadanie numerycznego różniczkowania opisane w tym rozdziale odpowiada pewnemu generalnemu schematowi, który można streścić w następujący sposób:

Do obliczenia pewnej wielkości stosuje się metodę numeryczną z parametrem h . Wynikiem jej działania (np. zastosowania wzoru przybliżającego – jak w przypadku różniczkowania numerycznego lub wykonania algorytmu realizującego metodę numeryczną) jest wyznaczona wartość $F(h)$. Błąd metody maleje do zera dla $h \rightarrow 0$, więc wartością dokładną jest $F(0)$. Trudności obliczeniowe rosną, gdy h maleje i uniemożliwiają zastosowanie tak małej wartości h , by błąd wyniku był dostatecznie mały.

W przypadku różniczkowania numerycznego te „trudności” to rosnący błąd spowodowany zaokrągleniami. W przypadku innych problemów może to być rosnący czas obliczeń lub zapotrzebowanie na pamięć.

Założymy, że znamy wykładniki $p_1 < p_2 < p_3 \dots$ parametru h , które występują w rozwinięciu $F(h)$ w szereg potęgowy

$$F(h) = a_0 + a_1 h^{p_1} + a_2 h^{p_2} + a_3 h^{p_3} \dots \quad (4.32)$$

W przypadku metod różniczkowania numerycznego takiej wiedzy dostarczyła analiza metody z zastosowaniem rozwinięcia w szereg Taylora. Na przykład dla różnicy progresywnej, ze wzoru (4.4) wynika, że w rozwinięciu występują wszystkie potęgi począwszy od pierwszej, czyli $p_1 = 1, p_2 = 2, p_3 = 3 \dots$, a dla różnicy centralnej, ze wzoru (4.10) dowiadujemy się, że w rozwinięciu (4.32) wystąpią tylko potęgi parzyste: $p_1 = 2, p_2 = 4, p_3 = 6 \dots$.

Wartość $a_0 = F(0)$ we wzorze (4.32) jest poszukiwaną wartością dokładną, a $EF(h) = a_1 h^{p_1} + a_2 h^{p_2} + a_3 h^{p_3} + \dots$ błędem metody obarczającym przybliżenie $F(h)$. Największym składnikiem tego błędu (częścią główną błędu) jest ten, w którym występuje h w najniższej potędze (bo $h \ll 1$), czyli $a_1 h^{p_1}$, przy tym to wykładnik p_1 decyduje o wielkości błędu.

Idea ekstrapolacji Richardsona opiera się na następującym rozumowaniu: trzeba obliczyć przybliżenia $F(h)$ dla kilku wartości h mieszczących się w zakresie, w którym trudności obliczeniowe nie są jeszcze dotkliwe i na podstawie uzyskanej w ten sposób informacji o przebiegu $F(h)$ „odgadnąć” czyli ekstrapolować wartość $a_0 = F(0)$. W metodzie tej kolejne wartości parametru h tworzą ciąg geometryczny o ilorazie $q > 1$:

$$\left\{ h_0 = h, h_1 = \frac{h_0}{q}, h_2 = \frac{h_1}{q} = \frac{h_0}{q^2}, h_3 = \frac{h_2}{q} = \frac{h_0}{q^3}, \dots \right\} \quad (4.33)$$

Twierdzenie 4.1

Jeśli $F_1(h) := F(h) = a_0 + a_1 h^{p_1} + a_2 h^{p_2} + \dots$ i zastosujemy wzór rekurencyjny:

$$F_{m+1}(h) = F_m(q^{-1}h) + \frac{F_m(q^{-1}h) - F_m(h)}{q^{p_m - 1}}, \quad m = 1, 2, 3, \dots, \quad (4.34)$$

gdzie $q > 1$, to

$$F_{m+1}(h) = a_0 + a_{m+1}^{(m+1)} h^{p_{m+1}} + a_{m+2}^{(m+1)} h^{p_{m+2}} + \dots \quad (4.35)$$

Dowód: (z zastosowaniem indukcji zupełnej)

Dla $m = 0$ teza jest spełniona na mocy założenia.

Przypuśćmy, że $F_m(h) = a_0 + a_m^{(m)} h^{p_{m+1}} + a_{m+1}^{(m)} h^{p_{m+1}} + \dots$ (założenie indukcyjne). Chcemy udowodnić, że $F_{m+1}(h) = a_0 + a_{m+1}^{(m+1)} h^{p_{m+1}} + a_{m+2}^{(m+1)} h^{p_{m+2}} + \dots$. Istotnie:

$$\begin{aligned}
 F_{m+1}(h) &= F_m(q^{-1}h) + \frac{F_m(q^{-1}h) - F_m(h)}{q^{p_m} - 1} = \\
 &= a_0 + a_m^{(m)} q^{-p_m} h^{p_m} + a_{m+1}^{(m)} q^{-p_{m+1}} h^{p_{m+1}} + \dots \\
 &+ \frac{[a_0 + a_m^{(m)} q^{-p_m} h^{p_m} + a_{m+1}^{(m)} q^{-p_{m+1}} h^{p_{m+1}} + \dots] - [a_0 + a_m^{(m)} h^{p_m} + a_{m+1}^{(m)} h^{p_{m+1}} + \dots]}{q^{p_m} - 1} \\
 &= a_0 + a_m^{(m)} \underbrace{\left[q^{-p_m} + \frac{q^{-p_m} - 1}{q^{p_m} - 1} \right]}_{=0} h^{p_m} + a_{m+1}^{(m)} \underbrace{\left[q^{-p_{m+1}} + \frac{q^{-p_{m+1}} - 1}{q^{p_m} - 1} \right]}_{=a_{m+1}^{(m+1)}} h^{p_{m+1}} + \dots = \\
 &= a_0 + a_{m+1}^{(m+1)} h^{p_{m+1}} + a_{m+2}^{(m+1)} h^{p_{m+2}} + \dots
 \end{aligned}$$

Na mocy tego twierdzenia, po m iteracjach otrzymamy (zgodnie z wzorem (4.34)) przybliżenie tej samej wartości poprawnej $a_0 = F_{m+1}(0)$, ale błąd metody

$$EF_{m+1}(h) = a_{m+1}^{(m+1)} h^{p_{m+1}} + a_{m+2}^{(m+1)} h^{p_{m+2}} + \dots \quad (4.36)$$

będzie teraz mniej więcej proporcjonalny do $h^{p_{m+1}}$, a nie do h^{p_1} , jak to było przed wykonaniem iteracji (4.34). To, że mogą zmienić się współczynniki w rozwinięciu błędu metody w szereg potęgowy, nie ma tak istotnego znaczenia jak to, że $h^{p_{m+1}} \ll h^{p_1}$.

Można powiedzieć, że jednokrotne wykonanie iteracji (4.34) eliminuje jeden składnik w rozwinięciu błędu (4.32). Wynika stąd, że szczególnie efektywnie będą współpracowały z ekstrapolacją Richardsona te metody (wzory) $F(h)$, które w rozwinięciu (4.32) nie mają wszystkich, kolejnych potęg parametru h . Tak jest, na przykład, w przypadku wzoru na różnicę centralną przybliżającego pochodną – w rozwinięciu błędu wystąpią tylko potęgi parzyste: $p_1 = 2$, $p_2 = 4$, $p_3 = 6 \dots$. Tak więc, różnica centralna ma nie tylko mniejszy błąd metody od różnicy progresywnej lub wstecznej, ale i bardziej efektywnie współpracuje z ekstrapolacją Richardsona.

Przeanalizujmy dokładniej pierwsze iteracje ekstrapolacji Richardsona:
 pierwsza iteracja:

$$F_1(h) := F(h), \quad F_2(h) = F_1(q^{-1}h) + \underbrace{\frac{F_1(q^{-1}h) - F_1(h)}{q^{p_1} - 1}}_{\text{poprawka}}, \quad (4.37)$$

druga iteracja:

$$F_2(q^{-1}h) = F_1(q^{-2}h) + \underbrace{\frac{F_1(q^{-2}h) - F_1(q^{-1}h)}{q^{p_1} - 1}}_{\text{poprawka}},$$

$$F_3(h) = F_2(q^{-1}h) + \underbrace{\frac{F_2(q^{-1}h) - F_2(h)}{q^{p_2} - 1}}_{\text{poprawka}}. \quad (4.38)$$

Jeżeli obliczenia są wykonywane ręcznie, to wygodnie zorganizować je w tabeli, jak poniżej:

k	0		1		2		3
m		$+\frac{\Delta}{q^{p_1}-1} =$		$+\frac{\Delta}{q^{p_2}-1} =$		$+\frac{\Delta}{q^{p_3}-1} =$	
0	$F_1(h_0)$						
1	$F_1(q^{-1}h_0)$	$+\frac{F_1(q^{-1}h_0) - F_1(h_0)}{q^{p_1} - 1}$	$F_2(h_0)$				
2	$F_1(q^{-2}h_0)$	$+\frac{F_1(q^{-2}h_0) - F_1(q^{-1}h_0)}{q^{p_1} - 1}$	$F_2(q^{-1}h_0)$	$+\frac{F_2(q^{-1}h_0) - F_2(h_0)}{q^{p_2} - 1}$	$F_3(h_0)$		
3	$F_1(q^{-3}h_0)$	$+\frac{F_1(q^{-3}h_0) - F_1(q^{-2}h_0)}{q^{p_1} - 1}$	$F_2(q^{-2}h_0)$	$+\frac{F_2(q^{-2}h_0) - F_2(q^{-1}h_0)}{q^{p_2} - 1}$	$F_3(q^{-1}h_0)$	$+\frac{F_3(q^{-1}h_0) - F_3(h_0)}{q^{p_3} - 1}$	$F_4(h_0)$

Jeżeli poprawka dodawana w kolejnej iteracji jest mniejsza od maksymalnego błędu, który chcemy uzyskać (lub od błędu, którym są obarczone dane wejściowe), to wynik iteracji należy odrzucić i jako najbardziej dokładny rezultat zaakceptować wynik poprzedniej iteracji.

Początkowa wartość parametru $h_0 = h$ musi być wybrana tak, by prawdziwe było rozwinięcie w szereg potęgowy (4.32). Na przykład dla różniczkowania funkcji f w punkcie x_0 metodą różnicy centralnej oznacza to, że funkcja f musi być gładka w przedziale $[x_0 - h, x_0 + h]$.

Przykład 4.1

Oblicz pochodną funkcji $f(x) = e^x$ w punkcie $x_0 = 1$. Przyjmij krok początkowy $h = 0,8$.

Obliczamy różnicę centralną: $D_C(h) = \frac{f(x_0+h)-f(x_0-h)}{2h}$ (obliczenia w całym przykładzie były wykonywane w podwójnej precyzji (typ double IEEE754), a wyniki zapisano z zachowaniem 5 cyfr po przecinku):

$$D_C(h)|_{h=0,8} = \frac{e^{1,8}-e^{0,2}}{1,6} = 3,01765, \quad D_C(h)|_{h=0,4} = \frac{e^{1,4}-e^{0,6}}{0,8} = 2,79135,$$

$$D_C(h)|_{h=0,2} = \frac{e^{1,2}-e^{0,8}}{0,4} = 2,73644, \quad D_C(h)|_{h=0,1} = \frac{e^{1,1}-e^{0,9}}{0,2} = 2,72281.$$

Przybliżone wartości pochodnej obliczone różnicą centralną zostały wpisane do pierwszej kolumny w tabeli ekstrapolacji.

k	0		1		2		3
m		$\frac{\Delta}{2^2-1} = \frac{\Delta}{3}$		$\frac{\Delta}{2^4-1} = \frac{\Delta}{15}$		$\frac{\Delta}{2^6-1} = \frac{\Delta}{63}$	
0	3,01765						
1	2,79135	$+(-0,07543) =$	2,71592				
2	2,73644	$+(-0,01830) =$	2,71814	$+0,00015 =$	2,71828		
3	2,72281	$+(-0,00454) =$	2,71827	$+0,00001 =$	2,71828	$+0,00000 =$	2,71828

W kolejnej tabeli zostały zapisane błędy ekstrapolacji:

$k \backslash m$	0	1	2	3
0	0,2993711			
1	0,0730696	-0,0023642		
2	0,0181582	-0,0001457	0,0000022	
3	0,0045327	-0,0000091	$0,3 \cdot 10^{-7}$	$-0,3 \cdot 10^{-9}$

Obliczyliśmy $\frac{d}{dt} e^x \Big|_{x=1} = 2,71828$.

Należy zaznaczyć, że aby ekstrapolacja była skuteczna, tzn. zmniejszała błąd metody, krok początkowy nie może być zbyt mały. W ekstrapolacji nie można zmniejszyć błędu zaokrągleń jeśli ten będzie dominował.

5. Całkowanie numeryczne

5.1. Kwadratury proste i złożone

Problem obliczenia całki oznaczonej

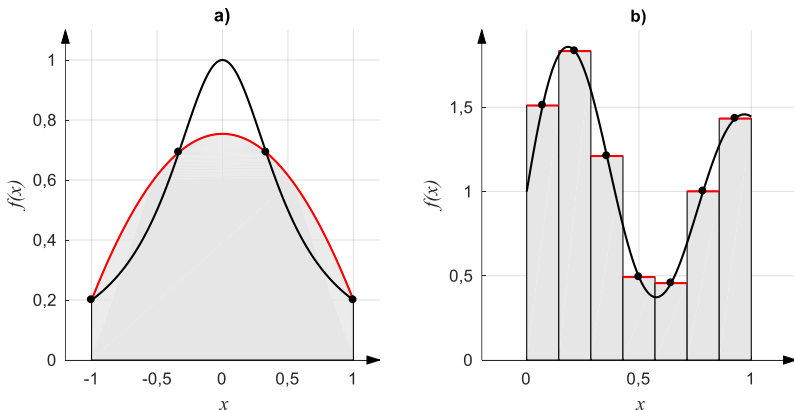
$$I = \int_a^b f(x)dx \quad (5.1)$$

z ciągłej funkcji $f(x)$ na ograniczonym przedziale $[a, b]$ pojawia się w licznych zastosowaniach inżynierskich. Koncepcje numerycznego rozwiązania tego problemu bazują na dość oczywistym pomysśle zastąpienia funkcji $f(x)$ przez taką funkcję przybliżającą, którą można bez trudu scałkować. Najczęściej tą funkcją jest wielomian, a sposobem jego konstrukcji interpolacja. Powstają w ten sposób wzory nazywane **kwadraturami**:

$$\int_a^b f(x)dx \approx \sum_{k=0}^n w_k f(x_k), \quad (5.2)$$

w których liczby w_k nazywane są **współczynnikami kwadratury**, a punkty $x_k \in [a, b]$ **węzłami kwadratury**. Jeżeli kwadratura zbudowana jest przez całkowanie jednego wielomianu interpolacyjnego na całym przedziale $[a, b]$, to nazywamy ją **kwadraturą prostą**.

Z **kwadraturą złożoną** mamy do czynienia jeżeli zastosowano koncepcję interpolacji odcinkowej, to jest podzielono przedział całkowania $[a, b]$ na podprzedziały i na każdym z tych podprzedziałów zastosowano wielomian interpolacyjny, scałkowano go i zsumowano otrzymane przybliżenia całek na podprzedziałach. Ideę kwadratur prostych i złożonych pokazano na rysunkach 5.1a i b.



Rys. 5.1. Przykład kwadratury prostej (a) i złożonej (b)

5.2. Kwadratury Newtona-Cotesa

Jeżeli w przedziale całkowania $[a, b]$ wybrano $n + 1$ równoodległych węzłów $x_k \in [a, b]$ i zbudowano na nich wielomian interpolacyjny $P_n(x)$ stopnia n , stosując wzór interpolacyjny Lagrange'a, to otrzymana kwadratura nosi nazwę kwadratury Newton-Cotesa. Współczynniki kwadratur Newtona-Cotesa nie zależą od całkowanej funkcji i mogą być wyznaczone z góry. We wzorze

$$\int_a^b f(x)dx \approx \int_a^b P_n(x)dx = \frac{b-a}{n_s} \sum_{i=0}^n \sigma_i f_i, \quad f_i = f(x_i) = P_n(x_i) \quad (5.3)$$

są one zapisane tak, by liczby σ_i były całkowite. Dla kolejnych stopni wielomianu interpolacyjnego zestawiono współczynniki kwadratur Newtona-Cotesa w tabeli 5.1.

Tabela 5.1. Współczynniki kwadratur Newtona-Cotesa

n	σ_i	n_s	błąd	nazwa
1	1 1	2	$h^3 \frac{1}{12} f^{(2)}(\xi)$	wzór trapezów
2	1 4 1	6	$h^5 \frac{1}{90} f^{(4)}(\xi)$	wzór Simpsona
3	1 3 3 1	8	$h^5 \frac{3}{80} f^{(4)}(\xi)$	wzór "trzech ósmych"
4	7 32 12 32 7	90	$h^7 \frac{8}{945} f^{(6)}(\xi)$	wzór Milne'a
5	19 75 50 50 75 19	288	$h^7 \frac{275}{12096} f^{(6)}(\xi)$	–
6	41 216 27 272 27 216 41	840	$h^9 \frac{9}{1400} f^{(8)}(\xi)$	wzór Weddle'a
$h = b - a$ – długość przedziału,		ξ – punkt pośredni		

Oszacowania błędów podanych w tabeli można wyprowadzić, na przykład, korzystając ze wzoru na resztę wielomianu interpolacyjnego (3.35). Dla przykładu, dla prostej kwadratury trapezów mamy:

$$\int_a^b f(x)dx - \int_a^b P_1(x)dx = \int_a^b \frac{1}{2} f^{(2)}(\xi)(x-a)(x-b)dx, \quad (5.4)$$

a po zastosowaniu podstawienia $z = \frac{1}{h}(x-a)$, $dx = h dz$:

$$\begin{aligned} \int_a^b f(x)dx - \int_a^b P_1(x)dx &= \frac{1}{2}f^{(2)}(\xi)h \int_0^1 z(z-1)h^2 dz \\ &= \frac{h^3}{2}f^{(2)}(\xi) \left[\frac{z^3}{3} - \frac{z^2}{2} \right]_0^1 = \frac{h^3}{12}f^{(2)}(\xi). \end{aligned} \quad (5.5)$$

Dla wielomianu stopnia 8,10,11 i wyższych współczynniki kwadratur Newtona-Cotesa nie tylko rosną, ale występują wśród nich ujemne i dodatnie liczby. Dodawanie i odejmowanie coraz większych liczb, by otrzymać tę samą wartość całki jest prostym sposobem na numeryczną niestabilność algorytmu.

Za pomocą kwadratury Newtona-Cotesa o $n + 1$ węzłach można scałkować bez błędu metody dowolny wielomian stopnia n (jeżeli $f(x)$ jest wielomianem stopnia n , a $P_n(x)$ wielomianem interpolacyjnym zbudowanym na $n + 1$ węzłach, to $P_n(x) \equiv f(x)$).

Ta własność sugeruje inny (niż całkowanie wzoru Lagrange'a) sposób wyznaczenia współczynników kwadratur Newtona-Cotesa. Jeżeli kwadratura o współczynnikach $w_i, i = 0, \dots, n$ ma obliczyć bez błędu metody każdą z całek $\int_{-1}^1 x^i dx, i = 0, \dots, n$, to musi być prawdziwy układ równań

$$\begin{aligned} w_0x_0^0 + w_1x_1^0 + \dots + w_nx_n^0 &= \int_{-1}^1 x^0 dx = \left[\frac{x^1}{1} \right]_{-1}^1 = \frac{1 - (-1)}{1}, \\ w_0x_0^1 + w_1x_1^1 + \dots + w_nx_n^1 &= \int_{-1}^1 x^1 dx = \left[\frac{x^2}{2} \right]_{-1}^1 = \frac{1^2 - (-1)^2}{2}, \\ &\dots\dots\dots \\ w_0x_0^n + w_1x_1^n + \dots + w_nx_n^n &= \int_{-1}^1 x^n dx = \left[\frac{x^{n+1}}{n+1} \right]_{-1}^1 = \frac{1 - (-1)^{n+1}}{n+1}. \end{aligned} \quad (5.6)$$

W postaci macierzowej może być zapisany jako:

$$\begin{bmatrix} x_0^0 & x_1^0 & \dots & x_n^0 \\ x_0^1 & x_1^1 & \dots & x_n^1 \\ \vdots & \vdots & \vdots & \vdots \\ x_0^n & x_1^n & \dots & x_n^n \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} \frac{1 - (-1)}{1} \\ \frac{1 - (-1)^2}{2} \\ \vdots \\ \frac{1 - (-1)^{n+1}}{n+1} \end{bmatrix}. \quad (5.7)$$

Macierz współczynników tego układu jest transponowaną macierzą Vandermonde'a i jak wiadomo (rozdział 3) jej wskaźnik uwarunkowania rośnie bardzo szybko z wymiarem, czyli liczbą węzłów kwadratury. Potwierdza to trudności na

jakie napotykamy, próbując stosować kwadratury Newtona-Cotesa z dużą liczbą węzłów.

5.3. Kwadratury Gaussa

Założona równoodległość węzłów w kwadraturach Newtona-Cotesa była bez wątpienia ograniczeniem. Odejście od tego założenia pozwala na skonstruowanie kwadratur, które przy $n + 1$ węzłach potrafią bez błędu scałkować wielomian stopnia większego niż n . Zilustrujmy to na przykładzie kwadratury wykorzystującej dwa węzły, czyli prostej kwadratury trapezów:

$$T = w_0 f(x_0) + w_1 f(x_1). \quad (5.8)$$

Przyjmijmy, że przedziałem całkowania jest $[a, b] = [-1, 1]$. Nie jest to istotnym ograniczeniem, bo wystarczy liniowo przeskalować zmienną niezależną. Jeśli założymy, że $x_0 = -1, x_1 = 1$, to równania (5.6, 5.7) mają dwie niewiadome:

$$\begin{bmatrix} x_0^0 & x_1^0 \\ x_0^1 & x_1^1 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \quad (5.9)$$

i rozwiązaniem

$$\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad (5.10)$$

czyli dostaliśmy prostą kwadraturę trapezów Newtona-Cotesa.

Jeżeli potraktujemy węzły x_0, x_1 jak niewiadome, to równanie (5.9) ma 4 niewiadome. Można do niego dołączyć kolejne dwa równania z układu (5.6, 5.7):

$$\begin{bmatrix} x_0^0 & x_1^0 \\ x_0^1 & x_1^1 \\ x_0^2 & x_1^2 \\ x_0^3 & x_1^3 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 2/3 \\ 0 \end{bmatrix} \quad (5.11)$$

i rozwiązując taki układ nieliniowych równań z czterema niewiadomymi wyznaczyć:

$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad x_0 = -\sqrt{\frac{1}{3}}, x_1 = \sqrt{\frac{1}{3}}. \quad (5.12)$$

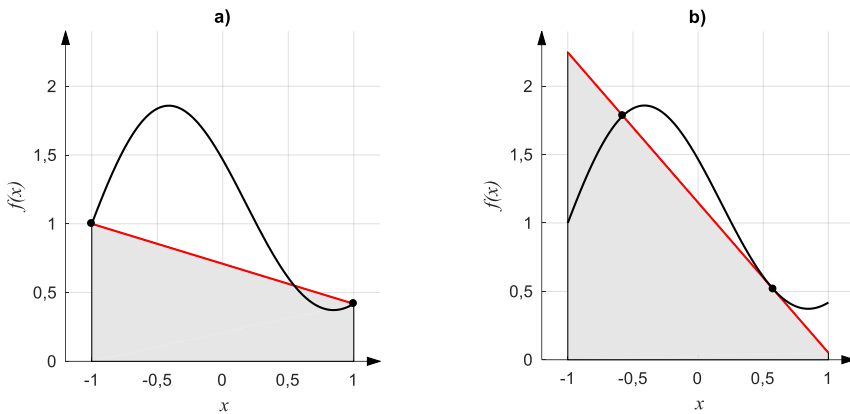
Tak więc, kwadratura

$$T = f\left(-\sqrt{\frac{1}{3}}\right) + f\left(\sqrt{\frac{1}{3}}\right) \quad (5.13)$$

pozwala bez błędu metody scałkować na przedziale $[-1, 1]$ każdy wielomian f , stopnia nie wyższego niż 3.

Na rysunku (5.2) porównano kwadratury o dwóch węzłach dla całki

$$\int_0^1 (1 + e^{-0,5x} \sin 5x) dx = 0,5 \int_{-1}^1 \left(1 + e^{-0,5\frac{z+1}{2}} \sin\left(5\frac{z+1}{2}\right)\right) dz.$$



Rys. 5.2. Porównanie kwadratury Newtona-Cotesa (a) i kwadratury Gaussa, (b) o dwóch węzłach

Przykład 5.1

Oblicz wartość całki $\int_{-0,25}^{0,25} e^x dx$, stosując kwadraturę Newtona-Cotesa oraz Gaussa o dwóch węzłach.

Dla pierwszej kwadratury mamy:

$$\int_{-0,25}^{0,25} e^x dx \approx 0,5 \left(\frac{e^{-0,25}}{2} + \frac{e^{0,25}}{2} \right) = 0,5157066,$$

dla drugiej kwadratury

$$\int_{-0,25}^{0,25} e^x dx = 0,25 \int_{-1}^1 e^{0,25z} dz \approx 0,25 \left(e^{-\frac{0,25}{\sqrt{3}}} + e^{\frac{0,25}{\sqrt{3}}} \right) = 0,5052174.$$

Wartość dokładna całki $\int_{-0,25}^{0,25} e^x dx = 2 \sinh 0,25 = 0,5052246 \dots$

Przedstawiony sposób rozumowania można uogólnić na większą niż 2 liczbę węzłów i wyprowadzić kwadratury Gaussa, które przy n węzłach pozwalają całkować bez błędu metody każdy wielomian stopnia nie wyższego niż $2n - 1$. Ważne jest przy tym, że współczynniki kwadratury Gaussa są zawsze dodatnie. Współczynniki i węzły kwadratur Gaussa są stabilizowane i nie ma potrzeby rozwiązywania nieliniowego układu równań przy każdym zastosowaniu kwadratury. Błąd metody dla kwadratur Gaussa na przedziale $[-1,1]$ można oszacować wyrażeniem $\frac{2^{2n+3}[(n+1)!]^4}{(2n+3)[(2n+2)!]^3} f^{(2n+2)}(\xi)$ (Ralston, 1983).

Wszystkie te uwagi nie zmieniają faktu, że stosowanie kwadratur prostych z wysoką liczbą węzłów stwarza istotne problemy numeryczne związane z błędami zaokrągleń.

5.4. Kwadratury złożone

Kwadratury złożone wykorzystują podział przedziału całkowania na n podprzedziałów (najczęściej równej długości) i interpolację funkcji podcałkowej na każdym z podprzedziałów wielomianem niskiego stopnia. Złożone kwadratury prostokątów, trapezów i Simpsona są zilustrowane na rysunku 5.3.

Złożona kwadratura trapezów jest opisana wzorem

$$\int_a^b f(x) dx \approx \frac{h}{2} \sum_{i=0}^{n-1} [f(x_i) + f(x_{i+1})] = T(h). \quad (5.14)$$

Jeżeli obliczenia są wykonywane ręcznie, to wygodniejszy jest wzór uwzględniający występowanie $f(x_i)$ dla $i = 1, \dots, n - 1$ w dwóch sąsiednich trapezach:

$$T(h) = h \left[\frac{f(a)}{2} + f(a+h) + \dots + f(b-h) + \frac{f(b)}{2} \right]. \quad (5.15)$$

Błąd metody dla kwadratury trapezów na pojedynczym podprzedziale jest, zgodnie z (5.5), równy

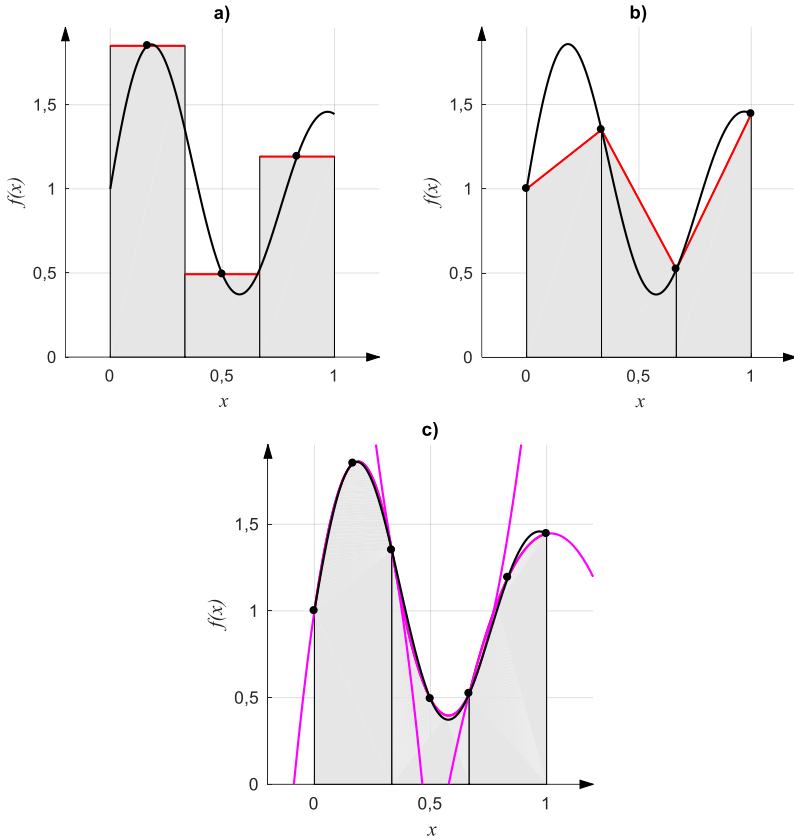
$$E_{Ti} = \frac{h^3}{12} f^{(2)}(\xi_i), \quad (5.16)$$

gdzie ξ_i jest pewnym punktem w podprzedziale $[x_i, x_{i+1}]$. Zsumowanie tych błędów po wszystkich podprzedziałach daje błąd złożonej kwadratury trapezów:

$$E_T = \frac{h^3}{12} \sum_{i=0}^{n-1} f^{(2)}(\xi_i) = \frac{h^2 nh \sum_{i=0}^{n-1} f^{(2)}(\xi_i)}{12 n} \quad (5.17)$$

$$= \frac{(b-a)h^2}{12} f^{(2)}(\xi),$$

przy czym ostatnia równość wynika z tego, że ciągła funkcja $f^{(2)}(x)$ osiąga średnią z wartości $f^{(2)}(\xi_i)$ w pewnym punkcie pośrednim ξ .



Rys. 5.3. Idea konstrukcji złożonych kwadratur Newtona-Cotesa: a) prostokątów (wariant punktu środkowego) – na każdym podprzedziale funkcja podcałkowa przybliżona stałą, b) trapezów – wielomianem liniowym, c) Simpsona – wielomianem kwadratowym (parabole narysowano także poza przedziałem w którym są całkowane)

W przypadku złożonej kwadratury Simpsona błąd metody (wyprowadzony w analogiczny sposób) wynosi:

$$E_S = \frac{(b-a)h^4}{180} f^{(4)}(\xi). \quad (5.18)$$

Z przedstawionych oszacowań błędów wynika, że złożona kwadratura trapezów ma większy błąd metody od złożonej kwadratury prostokątów i od złożonej kwadratury Simpsona. Podstawową zaletą złożonej kwadratury trapezów jest to, że rozwinięcie jej błędu metody w szereg potęgowy względem długości podprzeździału h zawiera tylko parzyste potęgi h :

$$T(h) = \int_a^b f(x)dx + a_1h^2 + a_2h^4 + a_3h^6 + \dots \quad (5.19)$$

Pozwala to efektywnie zastosować złożoną kwadraturę trapezów z ekstrapolacją Richardsona. Metoda całkowania utworzona w ten sposób nosi nazwę metody Romberga i przy co najmniej dwukrotnej ekstrapolacji jest dokładniejsza i od złożonej kwadratury prostokątów i od złożonej kwadratury Simpsona.

Przykład 5.2

Oblicz metodą Romberga całkę $\int_1^3 \frac{dx}{x}$.

Stosujemy wzór trapezów: $T(h) = h \left[\frac{f(a)}{2} + f(a+h) + \dots + f(b-h) + \frac{f(b)}{2} \right]$.

Wybieramy krok początkowy $h_0 = 2$, a kolejne wartości wyznaczmy stosując połowienie: $h_1 = 1$, $h_2 = 0,5$ oraz $h_3 = 0,25$. Obliczone wartości przybliżeń całki po zaokrągleniu do piątej cyfry po przecinku to:

$$T(h_0) = h_0 \left[\frac{f(a)}{2} + \frac{f(b)}{2} \right] = 2 \cdot \left[\frac{f(1)}{2} + \frac{f(3)}{2} \right] = 2 \cdot \left(\frac{1}{2} + \frac{1}{6} \right) = 1,33333,$$

$$\begin{aligned} T(h_1) &= h_1 \left[\frac{f(a)}{2} + f(a+h_1) + \frac{f(b)}{2} \right] = 1 \cdot \left[\frac{f(1)}{2} + f(2) + \frac{f(3)}{2} \right] \\ &= 1 \cdot \left(\frac{1}{2} + \frac{1}{2} + \frac{1}{6} \right) = 1,16667, \end{aligned}$$

$$\begin{aligned} T(h_2) &= h_2 \left[\frac{f(a)}{2} + f(a+h_2) + f(a+2h_2) + f(b-h_2) + \frac{f(b)}{2} \right] \\ &= 0,5 \cdot \left[\frac{f(1)}{2} + f(1,5) + f(2) + f(2,5) + \frac{f(3)}{2} \right] \\ &= 0,5 \cdot \left(\frac{1}{2} + \frac{2}{3} + \frac{1}{2} + \frac{2}{5} + \frac{1}{6} \right) = 1,11667, \end{aligned}$$

$$\begin{aligned}
T(h_3) &= h_3 \left[\frac{f(a)}{2} + \sum_{i=1}^6 f(a + ih_3) + f(b - h_3) + \frac{f(b)}{2} \right] = \\
&= 0,25 \cdot \left[\frac{f(1)}{2} + f(1,25) + f(1,5) + f(1,75) + f(2) + f(2,25) + f(2,5) \right] \\
&\quad + f(2,75) + \frac{f(3)}{2} \\
&= 0,25 \cdot \left(\frac{1}{2} + \frac{4}{5} + \frac{2}{3} + \frac{4}{7} + \frac{1}{2} + \frac{4}{9} + \frac{2}{5} + \frac{4}{11} + \frac{1}{6} \right) = 1,10321.
\end{aligned}$$

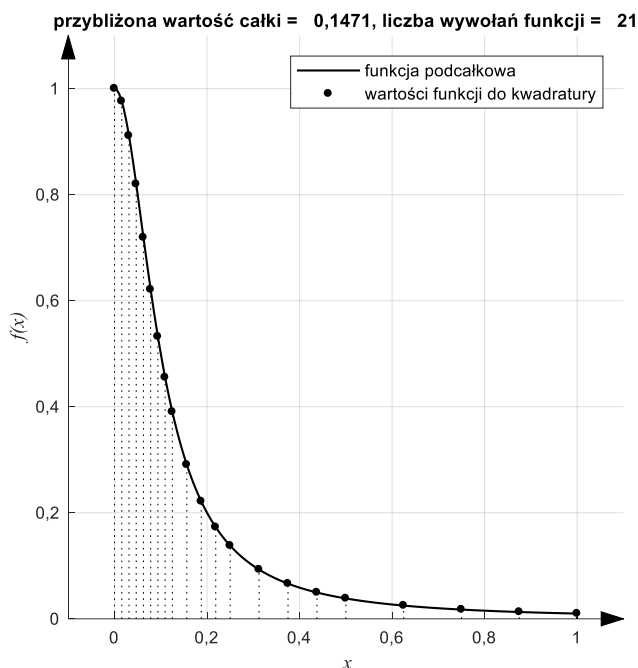
Obliczone wartości wpisujemy do pierwszej kolumny tabeli, w której będą zapisane kolejne iteracje ekstrapolacji.

k	0		1		2		3
m		$\frac{\Delta}{2^2 - 1} = \frac{\Delta}{3}$		$\frac{\Delta}{2^4 - 1} = \frac{\Delta}{15}$		$\frac{\Delta}{2^6 - 1} = \frac{\Delta}{63}$	
0	1,33333						
1	1,16667	$+(-0,05555) =$	1,11111				
2	1,11667	$+(-0,01667) =$	1,10000	$+(-0,00074) =$	1,0992		
3	1,10321	$+(-0,00449) =$	1,09873	$+(-0,00008) =$	1,0986	$+(-0,00001) =$	1,09863

Ostatecznie uzyskaliśmy $\int_1^3 \frac{dx}{x} \approx 1,09863$, podczas gdy wartość dokładna to $\int_1^3 \frac{dx}{x} = \ln 3 = 1,0986122886681$. Błąd całkowity przybliżenia nie przekracza więc $1,8 \cdot 10^{-5}$.

5.5. Kwadratury adaptacyjne

Podprzedziały, na które dzieli się przedział całkowania w kwadraturach złożonych nie muszą mieć stałej długości. Jeżeli funkcja podcałkowa jest wolnozmienna (prawie stała), można ją dokładnie scałkować przybliżając wielomianem niskiego stopnia na dłuższym podprzedziale, jeśli zmienia się gwałtownie, podprzedziały powinny być krótkie. Dobór długości podprzedziału można zautomatyzować – powierzyć procedurze całkowania.



Rys. 5.4. Ilustracja kwadratury adaptacyjnej – wykorzystano kwadratury Simpsona na kolejnych podprzedziałach

Kwadratura adaptacyjna estymuje błąd całkowania na bieżącym podprzedziale i decyduje o skróceniu lub wydłużeniu następnego podprzedziału. Estymacja błędów wykorzystuje różne sposoby, na przykład porównanie wyników obliczenia całki dla kroku o długości h i $h/2$. Jeśli oszacowany błąd jest akceptowalny, to wynik otrzymany dla tego podprzedziału jest aprobowany. Jeśli błąd jest zbyt duży następuje kolejne połowienie długości podprzedziału, aż do uzyskania wyniku spełniającego narzucone wymagania co do dokładności. W końcu podejmowana jest decyzja co do długości następnego podprzedziału. Metody szacowania błędów i doboru długości podprzedziału zostaną omówione dokładniej przy rozwiązywaniu równań różniczkowych zwyczajnych.

Oczywiście szacowanie błędów na każdym z podprzedziałów wymaga dodatkowych obliczeń, ale możliwość wydłużenia długości podprzedziałów w obszarach spokojnej zmienności funkcji podcałkowej sprawia, że kwadratury adaptacyjne są bardzo efektywne.

6. Iteracyjne metody rozwiązywania równań nieliniowych

6.1. Właściwości metod iteracyjnych

Nazwa tych metod pochodzi od słowa *iteratio*, czyli powtórzenie. W metodach iteracyjnych powtarza się proces numeryczny w celu ulepszenia wcześniejszych wyników. Każdy etap takiej metody, czyli iteracja wyznacza kolejne, w założeniu bardziej dokładne przybliżenie szukanego rozwiązania, korzystając z wyniku poprzedniego etapu. Ciąg przybliżeń otrzymanych z kolejnych iteracji powinien dążyć do dokładnego rozwiązania, ale może być ono osiągnięte w granicy, dla liczby iteracji dążącej do nieskończoności.

W każdej metodzie iteracyjnej muszą być określone:

- 1 – warunki początkowe, pozwalające wykonać pierwszą iterację,
- 2 – równanie lub algorytm opisujące jakie obliczenia należy wykonać w każdej iteracji,
- 3 – kryterium zatrzymania, które pozwoli zdecydować, że otrzymane przybliżenie jest wystarczająco dokładne i można przerwać wykonywanie iteracji.

Nie każdy proces iteracyjny musi być zbieżny. Ta sama metoda może w przypadku jednego problemu generować zbieżny ciąg iteracji, a w przypadku innego – rozbieżny. Pokazano to na rysunku 6.2. Zbieżność ciągu iteracji może też zależeć od wyboru punktu startowego.

Definicja 6.1

Metoda iteracyjna jest **zbieżna lokalnie** do rozwiązania dokładnego a , jeżeli istnieje takie otoczenie a , że dla każdego warunku początkowego x_0 z tego otoczenia ciąg przybliżeń generowanych w kolejnych iteracjach x_i zbiega do a dla liczby iteracji $i \rightarrow \infty$. Jeżeli zbieżność zachodzi dla dowolnych warunków początkowych to mówimy, że metoda jest **zbieżna globalnie**.

Z reguły istnieje wiele metod iteracyjnych pozwalających znaleźć rozwiązanie i przydatne jest narzędzie, które pozwala porównywać szybkość zbieżności tych metod. Taką miarą szybkości zbieżności metody iteracyjnej może być rząd metody.

Definicja 6.2

Niech x_i będzie ciągiem kolejnych przybliżeń zbieżnej metody iteracyjnej ($\lim_{i \rightarrow \infty} x_i = a$). Jeżeli istnieje liczba $p \geq 1$ i stała $C > 0$, takie, że

$$\lim_{i \rightarrow \infty} \frac{|x_{i+1} - a|}{|x_i - a|^p} = C \neq 0, \quad C < 1 \quad \text{gdy } p = 1 \quad (6.1)$$

to mówimy, że **metoda jest rzędu p w punkcie a** . Liczba C jest nazywana **stałą asymptotyczną błędu**.

Z równości (6.1) wynika, że dla dużej liczby wykonanych iteracji i :

$$|x_{i+1} - a| \approx C|x_i - a|^p. \quad (6.2)$$

Po lewej stronie (6.2) mamy błąd po wykonaniu i -tej iteracji $|x_{i+1} - a|$, a po prawej błąd przed wykonaniem i -tej iteracji $|x_i - a|$. Proces jest zbieżny, a liczba wykonanych iteracji duża, można więc przyjąć, że błędy te są mniejsze od 1. W takiej sytuacji to wykładnik p ma decydujące znaczenie w relacji między błędem po i przed iteracją, czyli dla szybkości zbieżności. Jeżeli $p = 1$, to mówimy o **zbieżności liniowej**. Wtedy stała asymptotyczna błędu decyduje o szybkości zbieżności – błąd po każdej iteracji maleje liniowo ze współczynnikiem C . Jeżeli $p = 2$, to zbieżność nazywamy **kwadratową**. Błąd po wykonaniu iteracji jest wtedy proporcjonalny do kwadratu błędu przed iteracją. Zbieżność kwadratowa jest znacznie szybsza od liniowej, o czym łatwo się przekonać, obliczając wartości błędów po kolejnych iteracjach dla przypadków $C = 0,5$; $p = 1$ i $C = 1$, $p = 2$ dla tego samego błędu początkowego, np. 0,1. Oczywiście rząd metody p może przyjmować także inne wartości np. ułamkowe.

Rząd metody ma charakter lokalny i jest związany z rozwiązaniem a . Np. ta sama metoda może zbiegać kwadratowo do jednego, a liniowo do innego pierwiastka tego samego równania.

Rząd metody „mierzy” szybkość zbieżności liczbą iteracji: dwie metody o tym samym rzędzie zbieżności, o podobnych stałych asymptotycznych błędu, startujące z tego samego przybliżenia początkowego, będą potrzebowały podobnej liczby iteracji dla osiągnięcia tej samej dokładności. Jeżeli potrafimy określić „koszt” K wykonania jednej iteracji, mierzony np. czasem obliczeń, liczbą operacji, czy jakkolwiek inaczej, to można zdefiniować **wskaźnik efektywności metody**:

$$E = p^{\frac{1}{K}}. \quad (6.3)$$

Dwie metody o tym samym wskaźniku efektywności, startujące z tego samego przybliżenia początkowego, będą wymagały podobnego kosztu całkowitego dla osiągnięcia tej samej dokładności, choć liczba iteracji może być różna.

Rząd zbieżności metody można określić, badając właściwości równania opisującego każdą iterację:

Twierdzenie 6.1

Jeżeli równaniem iteracji jest $x_{i+1} = \Phi(x_i)$ i kolejne pochodne spełniają warunek $\Phi^{(k)}(a) = 0$, $k = 1, \dots, p-1$ oraz pochodna $\Phi^{(p)}(a)$ jest ograniczona, to metoda jest rzędu p .

Dowód:

Po rozwinięciu funkcji $\Phi(x)$ w szereg Taylora w otoczeniu punktu a , otrzymujemy:

$$x_{i+1} = \Phi(x_i) = \Phi(a) + (x_i - a)\Phi'(a) + \frac{(x_i - a)^2\Phi''(a)}{2!} + \dots \\ + \frac{(x_i - a)^p\Phi^{(p)}(a)}{p!} + O(|x_i - a|^{p+1}),$$

czyli

$$\lim_{i \rightarrow \infty} \frac{|x_{i+1} - a|}{|x_i - a|^p} = \frac{\Phi^{(p)}(a)}{p!}. \quad (6.4)$$

Kryteria zatrzymania pracy metody iteracyjnej mogą być określone bardzo prosto – np. przez podanie maksymalnej liczby iteracji, mogą być też uzależniane w automatyczny sposób od oczekiwanej dokładności wyniku, na przykład jak w metodzie przedstawionej w ramce 6.1.

Ramka 6.1 Przykład kryterium zatrzymania

Niech równaniem iteracji, które opisuje zbieżny do a proces iteracyjny będzie $x_{i+1} = \Phi(x_i)$. Oczywiście zachodzi $a = \Phi(a)$. Na skutek błędów zaokrągleń w i -tej iteracji obliczamy $x_{i+1} = \Phi(x_i) + \delta_i$, gdzie δ_i oznacza błąd z jakim obliczana jest wartość $\Phi(x_i)$. Mamy $x_{i+1} - a = \Phi(x_i) - \Phi(a) + \delta_i$, a z twierdzenia o wartości średniej (dodatek D3) istnieje taki punkt pośredni z_i , że $x_{i+1} - a = \Phi'(z_i)(x_i - a) + \delta_i$. Jeżeli po obu stronach odejmiemy $\Phi'(z_i)(x_{i+1} - a)$, to otrzymamy $(1 - \Phi'(z_i))(x_{i+1} - a) = \Phi'(z_i)(x_i - x_{i+1}) + \delta_i$, co prowadzi do nierówności

$$|x_{i+1} - a| \leq \left| \frac{\Phi'(z_i)}{1 - \Phi'(z_i)} \right| |x_i - x_{i+1}| + \frac{|\delta_i|}{|1 - \Phi'(z_i)|}. \quad (\text{R6.1.1})$$

Jeżeli w punkcie z_i zachodzi $|\Phi'(z_i)| \leq m < 1$, to z (R6.1.1) wynika

$$|x_{i+1} - a| \leq \frac{m}{1 - m} |x_i - x_{i+1}| + \frac{|\delta_i|}{1 - m}. \quad (\text{R6.1.2})$$

Drugi składnik po prawej stronie nierówności (R6.1.2) to błąd zależny jedynie od stosowanej arytmetyki (błąd $|\delta_i|$) i od metody iteracyjnej (funkcja $\Phi(x)$).

Pierwszy składnik można zmniejszać wykonując kolejne iteracje. Oszacowanie pochodnej m można przybliżyć wyrażeniem

$$m \approx \hat{m} = \frac{|\Phi(x_i) - \Phi(x_{i-1})|}{|x_i - x_{i-1}|} = \frac{|x_{i+1} - x_i|}{|x_i - x_{i-1}|}. \quad (\text{R6.1.3})$$

Obliczenia powinny być więc zatrzymane, gdy $\hat{m} < 1$ oraz $\frac{\hat{m}}{1 - \hat{m}} |x_i - x_{i+1}| < \varepsilon$, gdzie ε jest parametrem określającym wymaganą dokładność wyniku. Warunki te można zapisać łącznie w postaci

$$(x_{i+1} - x_i)^2 < \varepsilon (|x_i - x_{i-1}| - |x_{i+1} - x_i|) \quad (\text{R6.1.4})$$

i ten warunek może być kryterium zatrzymania iteracji.

Jeśli iteracje zostały zakończone, to z (R6.1.2) wynika, że $|x_{i+1} - a| \leq \varepsilon + \frac{|\delta_i|}{1 - m}$, czyli na błąd otrzymanego wyniku ma wpływ błąd obliczeń $|\delta_i|$ z ostatniej iteracji. Wcześniejsze błędy $|\delta_{i-1}|, |\delta_{i-2}|, \dots$ nie mają znaczenia, jeśli tylko nie były na tyle duże, żeby doprowadzić do utraty zbieżności metody.

W tym rozdziale metody iteracyjne będą stosowane do rozwiązywania nieliniowych równań algebraicznych, przy czym przez rozwiązanie równania $f(x) = 0$ będzie rozumiane znalezienie dowolnego, rzeczywistego pierwiastka równania, leżącego w przedziale $[a, b]$.

6.2. Metoda bisekcji

Szukamy rzeczywistego pierwiastka równania $f(x) = 0$. Weźmy przedział $[a, b]$, na końcach którego $f(x)$ jest różnego znaku: $f(a)f(b) < 0$. Jeśli $f(x)$ jest ciągła, to na mocy własności Darboux (dodatek D3), osiąga wartość zero wewnątrz $[a, b]$. Połowiąc przedział $[a, b]$ i badając znak funkcji na końcach przedziałów, a następnie wybierając do kolejnej iteracji ten przedział, w którym funkcja zmienia znak, zawężamy przedział zawierający pierwiastek równania $f(x) = 0$. Po przeprowadzeniu każdej iteracji otrzymujemy przedział dwa razy krótszy, a w przypadku przzerwania obliczeń wynikiem jest środek ostatniego przedziału. Ponieważ prowadzimy obliczenia w arytmetyce zmiennopozycyjnej nie znajdziemy pewnie punktu, w którym dokładnie $f(x) = 0$. Naszym celem będzie znalezienie przedziału o długości nie przekraczającej zadanej dokładności obliczeń (mogą to być dwie sąsiednie liczby zmiennoprzecinkowe), w którym $f(x)$ zmienia znak.

Przykład 6.1

Należy znaleźć miejsce zerowe funkcji

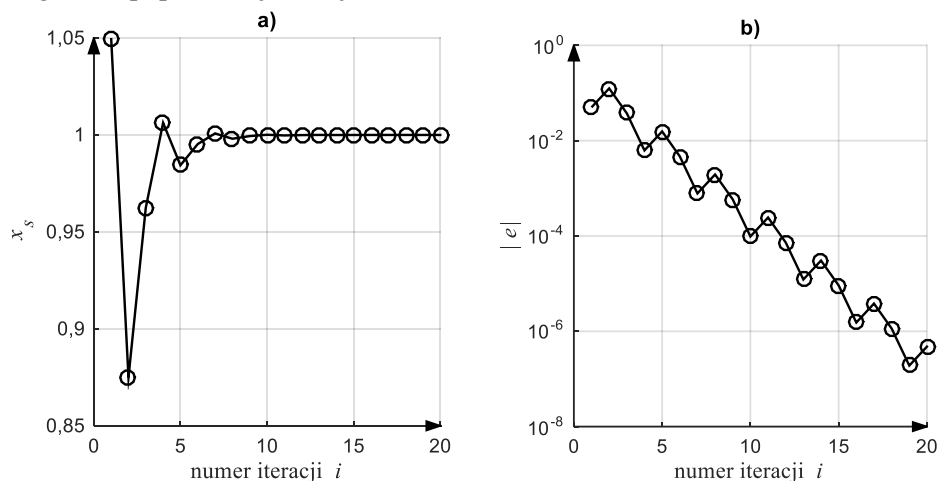
$$f(x) = 2e^{x-1} - x - 1. \quad (6.5)$$

Jak łatwo sprawdzić, miejscem zerowym podanej funkcji jest $x = 1$ (drugim pierwiastkiem jest $x_0 \approx -0,5936$). Zastosujemy metodę bisekcji z przedziałem startowym $[x_a, x_b] = [0,7; 1,4]$. Wyniki zostały przedstawione w tabeli 6.1 (błąd $e = 1 - x_s$ jest błędem całkowitym, na który składa się błąd metody i błąd zaokrąglenia) oraz na rysunku 6.1.

Tabela 6.1. Zestawienie wyników metody bisekcji dla pierwszych siedmiu iteracji

n	x_a	x_b	$x_s = \frac{1}{2}(x_a + x_b)$	$f(x_a)$	$f(x_b)$	$f(x_s)$	$e = 1 - x_s$
1	0,7	1,4	1,05	-	+	+	-0,05
2	0,7	1,05	0,875	-	+	-	0,125
3	0,875	1,05	0,9625	-	+	-	0,0375
4	0,9625	1,05	1,00625	-	+	+	-0,00625
5	0,9625	1,00625	0,984375	-	+	-	0,015625
6	0,984375	1,00625	0,9953125	-	+	-	0,0046875
7	0,9953125	1,00625	1,00078125	-	+	+	-0,00078125

W tabeli pogrubioną czcionką zaznaczono kraniec przedziału, który uległ zmianie względem poprzedniej iteracji.



Rys. 6.1. Kolejne przybliżenia pierwiastka x_s (a) oraz błąd $|e| = |1 - x_s|$ (b) w zależności od numeru iteracji i

Metoda bisekcji jest zbieżna liniowo. W każdej iteracji błąd rozwiązania maleje w przybliżeniu dwukrotnie. Funkcja $f(x)$ jest jednokrotnie wywoływana w każdej iteracji, ale istotny jest tylko znak, a nie dokładna wartość.

6.3. Metoda iteracji prostej

Metoda jest przeznaczona do rozwiązywania równań postaci $x = g(x)$. W każdej iteracji wykonuje się operację

$$x_{i+1} = g(x_i). \quad (6.6)$$

Jeżeli dokładnym rozwiązaniem jest $a = g(a)$, to wykorzystując twierdzenie o wartości średniej dla funkcji $g(x)$ można napisać

$$a - x_{i+1} = g(a) - g(x_i) = g'(\xi)(a - x_i), \quad (6.7)$$

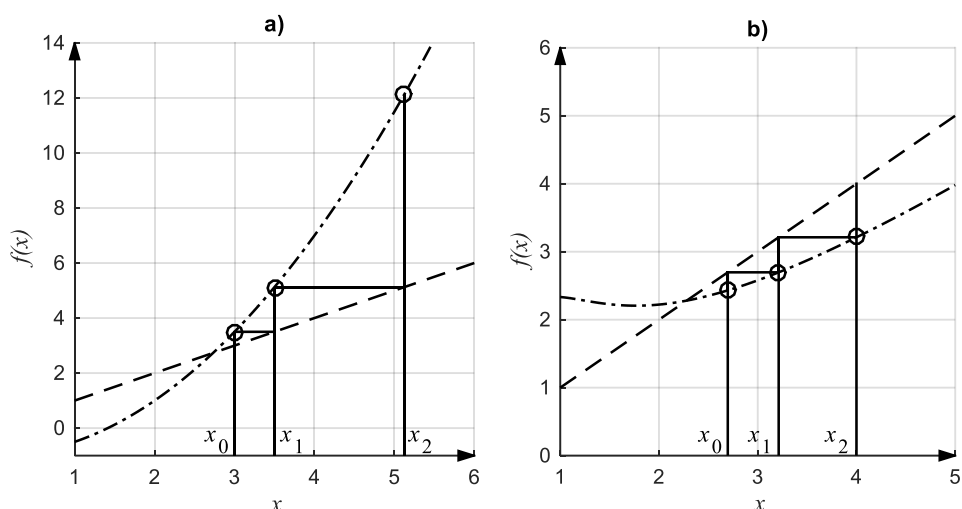
gdzie ξ jest pewnym punktem leżącym pomiędzy a i x_i . Zależność (6.7) podaje związek między błędem po wykonaniu i -tej iteracji $a - x_{i+1}$, a błędem przed tą iteracją $a - x_i$. Z (6.7) oraz z twierdzenia 6.2 wynika, że jeżeli dla każdego punktu ξ w otoczeniu a (zawierającym x_0):

- $|g'(\xi)| < 1$, to proces iteracyjny jest zbieżny, a jeśli ponadto $g'(\xi) > 0$, to błąd nie zmienia znaku,
- $|g'(\xi)| > 1$, to proces iteracyjny jest rozbieżny.

Metoda iteracji prostej jest więc zbieżna liniowo, ze stałą asymptotyczną błędów

$$C \leq \max_{\xi \in I} |g'(\xi)|, \quad x_0, a \in I. \quad (6.8)$$

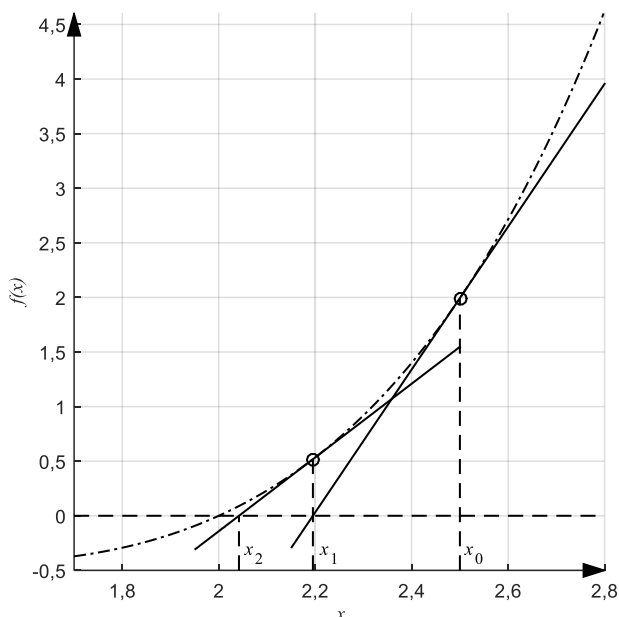
Na rysunku 6.2 pokazano przykład rozbieżnego i zbieżnego procesu iteracji prostej.



Rys. 6.2. Przykład rozbieżnego (a) i zbieżnego (b) procesu iteracji – linia kreskowa prosta $y = x$, linia kropka-kreska funkcja $g(x)$

6.4. Metoda Newtona-Raphsona

Metoda Newtona-Raphsona jest chyba najbardziej popularną techniką rozwiązywania równań nieliniowych postaci $f(x) = 0$. W każdej iteracji funkcja $f(x)$ jest zastępowana przybliżeniem liniowym i rozwiązywane jest odpowiednie równanie liniowe. Geometrycznie odpowiada to wystawieniu stycznej w punkcie x_i i znalezieniu punktu przecięcia tej stycznej z osią x , tak jak to pokazano na rysunku 6.3.



Rys. 6.3. Idea działania metody Newtona: linia ciągła – styczne, linia kreska-kropka – funkcja $f(x)$

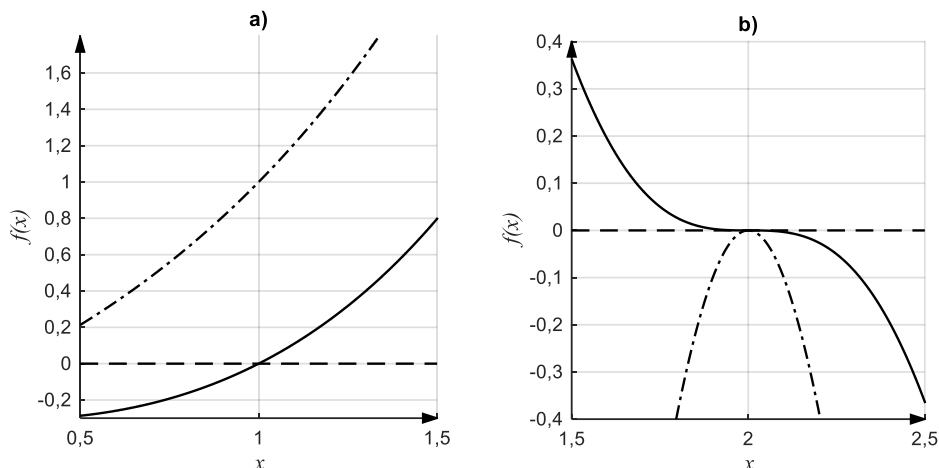
Jeżeli skorzystamy z rozwinięcia funkcji $f(x)$ w szereg Taylora w otoczeniu aktualnego przybliżenia x_i , w którym znajdzie się rozwiązanie równania a , to można zapisać:

$$f(a) = 0 = f(x_i) + (a - x_i)f'(x_i) + \frac{(a - x_i)^2}{2!}f''(x_i) + \frac{(a - x_i)^3}{3!}f^{(3)}(x_i) + \dots \quad (6.9)$$

Pominięcie składników nieliniowych (na prawo od drugiego znaku +) prowadzi do równania liniowego, którego rozwiązaniem nie będzie wprowadzany poszukiwany pierwiastek równania nieliniowego a , ale jego kolejne przybliżenie x_{i+1} . W każdej iteracji rozwiązywane jest więc równanie

$$0 = f(x_i) + (x_{i+1} - x_i)f'(x_i) \Rightarrow x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}. \quad (6.10)$$

Rząd zbieżności metody Newtona-Raphsona można zbadać, wykorzystując twierdzenie 6.1 o rzędzie zbieżności. Trzeba rozróżnić między przypadkiem pojedynczego (kiedy $f'(a) \neq 0$) a wielokrotnego (kiedy $f'(a) = 0$) pierwiastka równania.



Rys. 6.4. Wykres funkcji z pojedynczym (a) i wielokrotnym pierwiastkiem (b) (o nieparzystej krotności): linia ciągła – funkcja $f(x)$, linia kropka-kreska – pochodna $f'(x)$

W przypadku pojedynczego pierwiastka, po zróżniczkowaniu funkcji $\Phi(x) = x - \frac{f(x)}{f'(x)}$ opisującej każdą iterację otrzymuje się:

$$\Phi'(a) = \left[1 - \left(\frac{f'(x)}{f'(x)} \right)^2 + \frac{f(x)f''(x)}{[f'(x)]^2} \right] \Bigg|_{x=a} = 0, \quad (6.11)$$

czyli, zgodnie z twierdzeniem 6.1, rząd zbieżności wynosi 2.

Kwadratowa zbieżność metody Newtona-Raphsona w przypadku pojedynczych pierwiastków jest jej podstawową zaletą. Istnieją metody o szybszej zbieżności (wyższym rzędzie i wskaźniku efektywności), ale kwadratowa zbieżność jest naprawdę wystarczająca. Liczba cyfr poprawnych rośnie dwukrotnie z każdą iteracją, więc po kilku iteracjach natrafiamy na barierę dokładności stosowanej arytmetyki zmiennopozycyjnej.

W przypadku wielokrotnego pierwiastka pochodną funkcji $\Phi(x)$ trzeba wyznaczyć w inny sposób. Najpierw wyodrębnia się czynnik związany z m -krotnym pierwiastkiem:

$$f(x) = (x - a)^m g(x), \quad g(a) \neq 0, \quad (6.12)$$

co daje

$$f'(x) = m(x - a)^{m-1} g(x) + (x - a)^m g'(x), \quad (6.13)$$

a następnie różniczkuje funkcję:

$$\Phi(x) = x - \frac{(x-a)^m g(x)}{m(x-a)^{m-1}g(x) + (x-a)^m g'(x)}, \quad (6.14)$$

co daje

$$\Phi'(a) = 1 - \frac{1}{m}. \quad (6.15)$$

W przypadku wielokrotnych pierwiastków zbieżność metody Newtona-Raphsona jest liniowa, stała asymptotyczna błędu $C = 1 - \frac{1}{m}$ zależy od krotności pierwiastka. Zawsze będzie korzystne takie sformułowanie problemu, które zapewni jednokrotność poszukiwanego pierwiastka!

Przykład 6.2

Rozważmy funkcje

$$f_1(x) = 2e^{x-1} - x - 1, \quad (6.16)$$

$$f_2(x) = \arctg(x-2) - (x-2). \quad (6.17)$$

Funkcja f_1 ma pierwiastek jednokrotny $x_I = 1$, funkcja f_2 ma pierwiastek wielokrotny w punkcie $x_{II} = 2$. Wykresy funkcji zostały przedstawione na rysunku 6.4.

Zastosowanie metod Newtona-Raphsona wymaga znajomości pochodnych funkcji f_1 i f_2

$$f_1'(x) = 2e^{x-1} - 1, \quad (6.18)$$

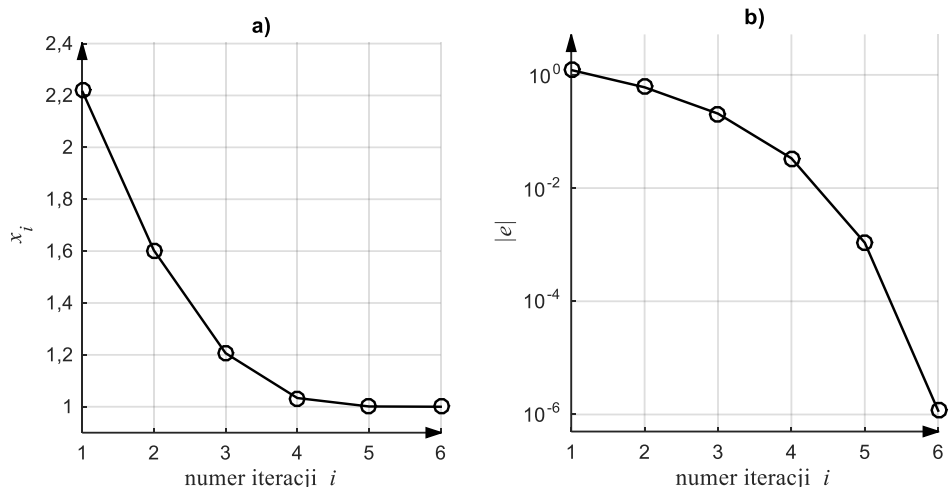
$$f_2'(x) = -\frac{(x-2)^2}{(x-2)^2 + 1}. \quad (6.19)$$

Z zależności (6.19) widać, że w punkcie $x_{II} = 2$ będącym pierwiastkiem $f_2(x)$ pochodna się zeruje, czyli $f_2'(x_{II}) = 0$. Także druga pochodna zeruje się w tym punkcie, a dopiero trzecia jest różna od zera. Pierwiastek jest więc trzykrotny.

W tabeli 6.2 oraz na rysunku 6.5 przedstawiono wyniki kolejnych iteracji dla funkcji f_1 . Błąd e jest błędem całkowitym, na który składa się błąd metody i błąd zaokrążeń.

Tabela 6.2. Wyniki rozwiązania równania $f_1(x) = 0$ dla sześciu pierwszych iteracji

i	x_i	x_{i-1}	$f_1(x_{i-1})$	$f_1'(x_{i-1})$	$e = x_i - x_{i-1}$
1	2,2177	3,0000	3,5413	5,7591	-1,2177
2	1,6028	2,2177	1,0517	2,6545	$-6,0282 \cdot 10^{-1}$
3	1,2066	1,6028	$2,5242 \cdot 10^{-1}$	1,4590	$-2,0663 \cdot 10^{-1}$
4	1,0336	1,2066	$3,4767 \cdot 10^{-2}$	1,0684	$-3,3623 \cdot 10^{-2}$
5	1,0011	1,0336	$1,0834 \cdot 10^{-3}$	1,0022	$-1,0822 \cdot 10^{-3}$
6	1,0000	1,0011	$1,1694 \cdot 10^{-6}$	1,0000	$-1,1694 \cdot 10^{-6}$

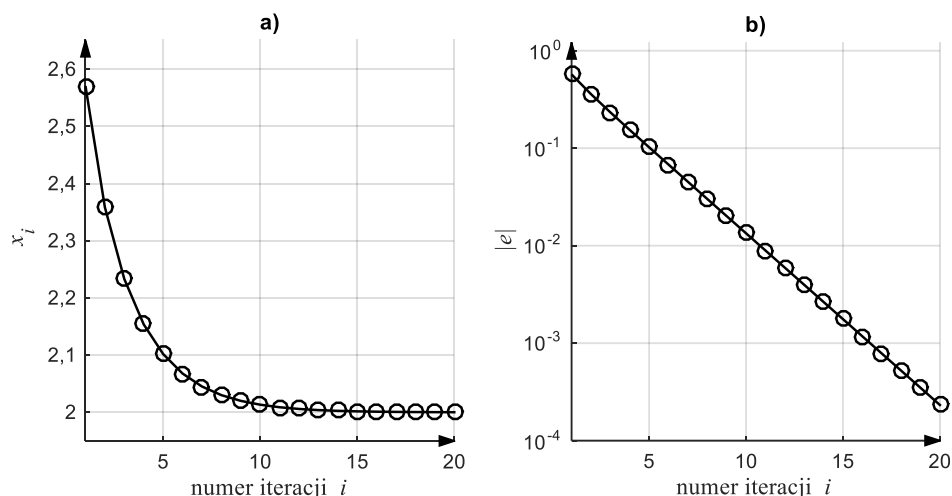


Rys. 6.5. Przybliżona wartość obliczanego pierwiastka x_i (a) oraz moduł błędu bezwzględnego $|e| = |x_i - x_{i-1}|$ (b) w kolejnych iteracjach. Pojedynczy pierwiastek funkcji $f_1(x)$

W tabeli 6.3 oraz na rysunku 6.6 przedstawiono wyniki dla funkcji f_2 . Tu także błąd e jest błędem całkowitym, na który składa się błąd metody i błąd zaokrążeń.

Tabela 6.3. Wyniki rozwiązania równania $f_2(x) = 0$

i	x_i	x_{i-1}	$f_2(x_{i-1})$	$f_2'(x_{i-1})$	$e = x_{i1} - x_{i-1}$
1	2,570796	3,000000	-0,052127	-0,245743	-0,570796
2	2,358677	2,570796	-0,014293	-0,113985	-0,358677
3	2,233282	2,358677	-0,004099	-0,051612	-0,233282
4	2,153867	2,233282	-0,001197	-0,023128	-0,153867
5	2,102097	2,153867	-0,000353	-0,010316	-0,102097
6	2,067924	2,102097	-0,000104	-0,004592	-0,067924



Rys. 6.6. Przybliżona wartość obliczanego pierwiastka x_i (a) oraz moduł błędu bezwzględnego $|e| = |x_i - x_{i-1}|$ (b) w kolejnych iteracjach. Wielokrotny pierwiastek funkcji $f_2(x)$

Jak widać na rysunkach 6.6 oraz w tabeli 6.3 zbieżność metody jest wolna dla pierwiastka wielokrotnego.

W celu przyspieszenia zbieżności można wykorzystać informację o krotności pierwiastka: jak wynika z (6.12, 6.13) a , n -krotny pierwiastek funkcji $f(x)$ będzie $(n - 1)$ -krotnym pierwiastkiem jej pochodnej $f'(x)$, a więc pojedynczym pierwiastkiem funkcji

$$h(x) = \frac{f(x)}{f'(x)}, \quad (6.20)$$

co wynika z tożsamości $f(x) = (x - a)^n g(x)$, $f'(x) = (x - a)^{n-1} [ng(x) + (x - a)g'(x)] = (x - a)^{n-1} \hat{g}(x)$, $\frac{f(x)}{f'(x)} = \frac{(x-a)^n g(x)}{(x-a)^{n-1} \hat{g}(x)} = (x - a) \frac{g(x)}{\hat{g}(x)}$. Zastosowanie metody Newtona-Raphsona do równania $h(x) = 0$ pozwala określić równanie iteracji jako

$$x_{i+1} = x_i - \frac{h(x_i)}{h'(x_i)} = x_i - \frac{f(x_i)f'(x_i)}{[f'(x_i)]^2 - f(x_i)f''(x_i)}. \quad (6.21)$$

Przykład 6.3

Rozważmy ponownie funkcję

$$f(x) = f_2(x) = \arctg(x - 2) - (x - 2). \quad (6.22)$$

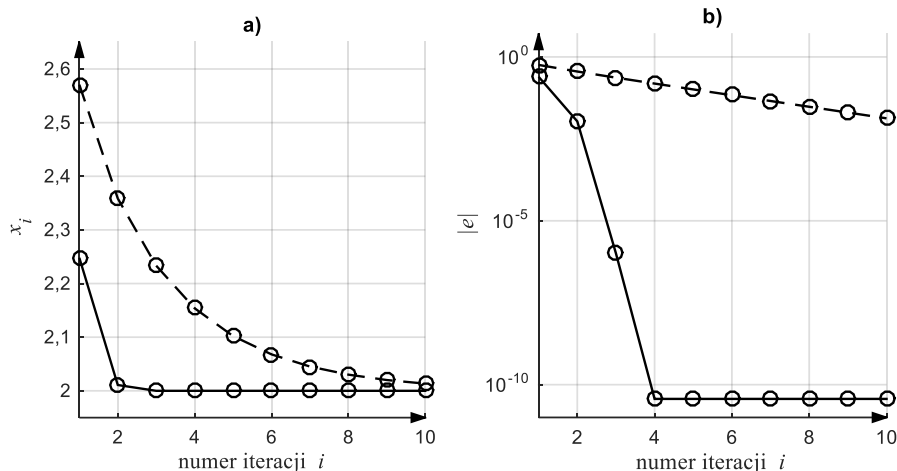
Definiujemy funkcję $h(x)$ jako

$$h(x) = -\frac{[\operatorname{arctg}(x-2) - (x-2)][(x-2)^2 + 1]}{(x-2)^2}. \quad (6.23)$$

Mimo składnika $(x-2)^2$ w mianowniku (6.23), po dokonaniu skróceń okazuje się, że 2 należy do dziedziny $h(x)$ i jest jej miejscem zerowym. W tabeli 6.4 oraz na rysunku 6.7 przedstawiono wyniki dla funkcji h .

Tabela 6.4. Wyniki rozwiązania równania $h(x) = 0$

i	x_i	x_{i-1}	$h(x_{i-1})$	$h'(x_{i-1})$	$e = x_{i1} - x_{i-1}$
1	2,248062	3,000000	0,084671	0,356915	-0,248062
2	2,010832	2,248062	0,003611	0,333380	-0,010832
3	2,000001	2,010832	0,000001	1,000001	-0,000001
4	2,000000	2,000000	0,000000	1,000000	-0,000000



Rys. 6.7. Porównanie szybkości zbieżności w przypadku pierwiastka dwukrotnego: rozwiązanie równania $f_2(x) = 0$ – linia kreskowa oraz równania $h(x) = 0$ – linia ciągła

Przekształcenie równania $f_2(x) = 0$ do $h(x) = 0$ spowodowało przyspieszenie zbieżności. Widoczna na rysunku 6.7b stagnacja błędów po czwartej iteracji (linia ciągła) jest wynikiem osiągnięcia dokładności wynikającej z zastosowanej reprezentacji zmiennoprzecinkowej (błędów zaokrągleń).

Metoda Newtona-Raphsona została wyprowadzona przy założeniu, że w otoczeniu poszukiwanego pierwiastka istnieją i są ciągłe kolejne pochodne funkcji $f(x)$. Dla wielu funkcji okazuje się, że zbieżność metody Newtona-Raphsona ma charakter lokalny, czyli zachodzi jedynie wtedy, gdy przybliżenie początkowe zostanie

wybrane dostatecznie blisko wyznaczanego pierwiastka. Kilka przykładów sytuacji, w których metoda Newtona-Raphsona nie jest zbieżna, lub zbiega bardzo wolno przedstawiono w przykładzie 6.4.

Przykład 6.4

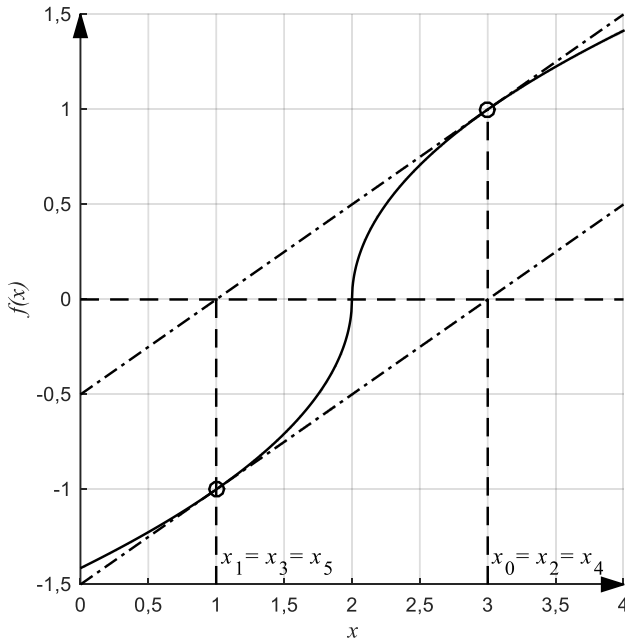
1. Rozważmy funkcję

$$f(x) = \text{sign}(x - 2)\sqrt{|x - 2|}. \tag{6.24}$$

Cechą charakterystyczną funkcji (6.24) jest nieskończona wartość pierwszej pochodnej w wyznaczanym pierwiastku $x = 2$. Na rysunku 6.7 i w tabeli 6.5 zostały przedstawione wyniki kolejnych iteracji.

Tabela 6.5. Wyniki iteracji dla różnych warunków początkowych

i	x_i dla $x_0 = 1,95$	x_i dla $x_0 = 1,5$	x_i dla $x_0 = 3$
1	2,05	2,5	1
2	1,95	1,5	3
3	2,05	2,5	1
4	1,95	1,5	3



Rys. 6.8. Graficzne przedstawienie kolejnych iteracji: linia ciągła – funkcja $f(x)$, linia kreska-kropka – styczne

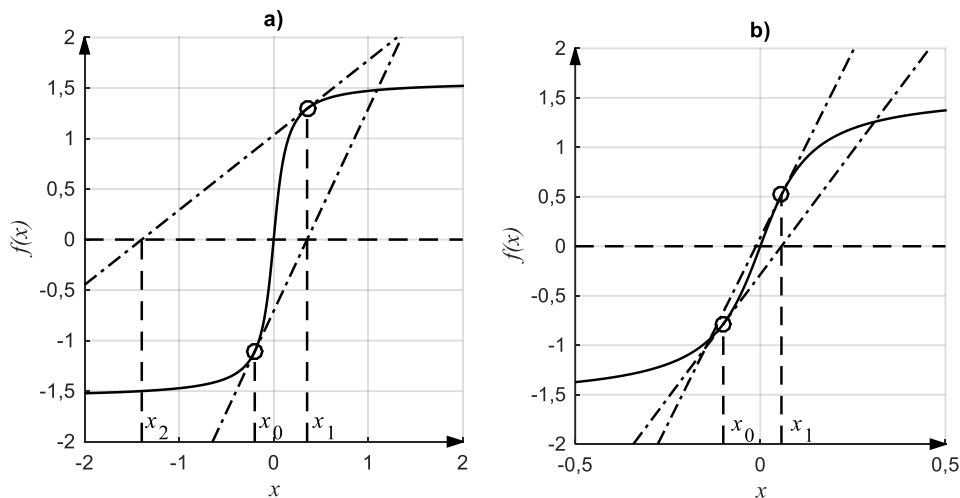
2. Rozważamy funkcję

$$f(x) = \operatorname{arctg}(40x). \quad (6.25)$$

Pochodna funkcji (6.25) dąży do zera przy oddalaniu się od pierwiastka i to jest przyczyną rozbieżności procesu iteracyjnego, jeśli punkt startowy zostanie wybrany daleko od wyznaczanego pierwiastka. W tabeli 6.6 i na rysunku 6.8 przedstawiono wyniki iteracji dla dwóch warunków początkowych.

Tabela 6.6. Wyniki iteracji dla różnych warunków początkowych

i	x_i dla $x_0 = -0,2$	x_i dla $x_0 = -0,1$
1	0,35357435890	0,057079632679490
2	1,39509590869	-0,011685990399891
3	27,93440665336	0,000106102211704
4	-1220,169989179	-0,000000000079631



Rys. 6.9. Graficzne przedstawienie kolejnych iteracji dla procesu rozbieżnego (a) i zbieżnego (b): linia ciągła – funkcja $f(x)$, linia kropka-kreska – styczne

3. Rozważmy funkcję

$$f(x) = x^{10} - 1. \quad (6.26)$$

Funkcję (6.26) cechują dwie właściwości: bardzo płaskie minimum dla $x = 0$ (dziewięć pierwszych pochodnych tej funkcji zeruje się w tym punkcie) oraz szybki wzrost wartości funkcji dla $x > 1$. W tabeli 6.7 przedstawiono wyniki iteracji dla punktu startowego $x_0 = 0,5$.

Tabela 6.7. Wyniki kolejnych iteracji dla funkcji (6.25)

i	x_i	$f(x_{i-1})$	$f'(x_{i-1})$	$e = 1 - x_i$
1	51,6500	0,9990	0,0195	-50,6500
2	46,4850	$1,3511 \cdot 10^{17}$	$0,2616 \cdot 10^{17}$	-45,4850
3	41,8365	$4,7112 \cdot 10^{16}$	$1,0135 \cdot 10^{16}$	-40,8365
4	37,6529	$1,6427 \cdot 10^{16}$	$3,9264 \cdot 10^{16}$	-36,6529
20	6,9771	$7,8407 \cdot 10^8$	$1,0114 \cdot 10^9$	5,9771
30	2,4328	$2,0826 \cdot 10^4$	$7,7047 \cdot 10^4$	1,4328
33	1,7738	$8,8243 \cdot 10^2$	$4,4829 \cdot 10^3$	0,7738
34	1,5970	$3,0742 \cdot 10^2$	$1,7387 \cdot 10^3$	0,5970
35	1,4388	$1,0693 \cdot 10^2$	$6,7580 \cdot 10^2$	0,4388
36	1,2987	$3,7021 \cdot 10^1$	$2,6426 \cdot 10^2$	0,2987
37	1,1784	$1,2650 \cdot 10^1$	$1,0510 \cdot 10^2$	0,1784
38	1,0835	4,1613	$4,3801 \cdot 10^1$	0,0835
39	1,0237	1,2268	$2,0555 \cdot 10^1$	0,0237
40	1,0023	0,2635	$1,0210 \cdot 10^1$	0,0023
41	1,0000	0,0234	$1,0002 \cdot 10^1$	0,0000

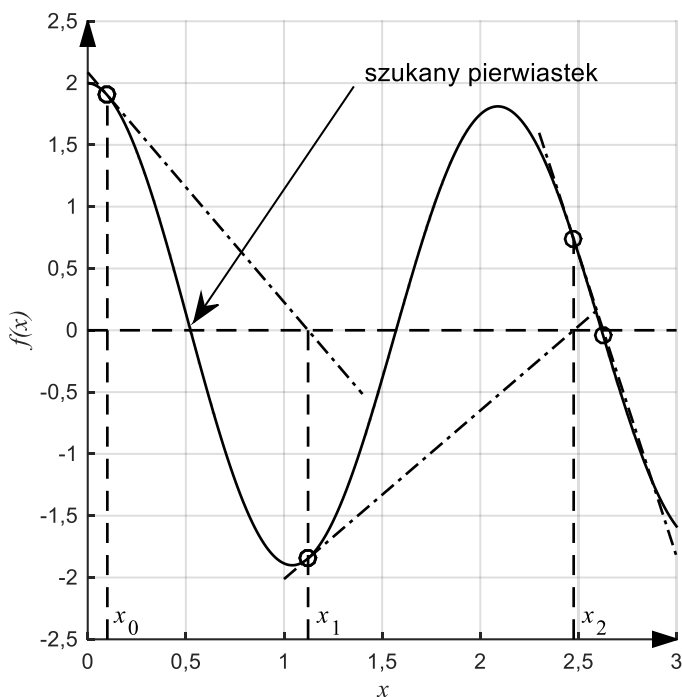
Wyniki zamieszczone w tabeli 6.7 pokazują bardzo wolną początkową zbieżność. Spowodowana jest ona przez dwie właściwości podane wcześniej. Mała wartość pochodnej w punkcie startowym powoduje przeskok do obszaru, gdzie pochodna funkcji jest bardzo duża. Skutkuje to wolną zbieżnością procesu iteracyjnego.

4. Rozważmy funkcję

$$f(x) = (e^{-0,1x} + 1) \cos(3x). \quad (6.27)$$

Celem jest znalezienie pierwiastka równania $f(x) = 0$, którym jest $x \approx 0,5236$.

Na rysunku 6.9 przedstawiono wyniki iteracji dla punktu startowego $x_0 = 0,1$.



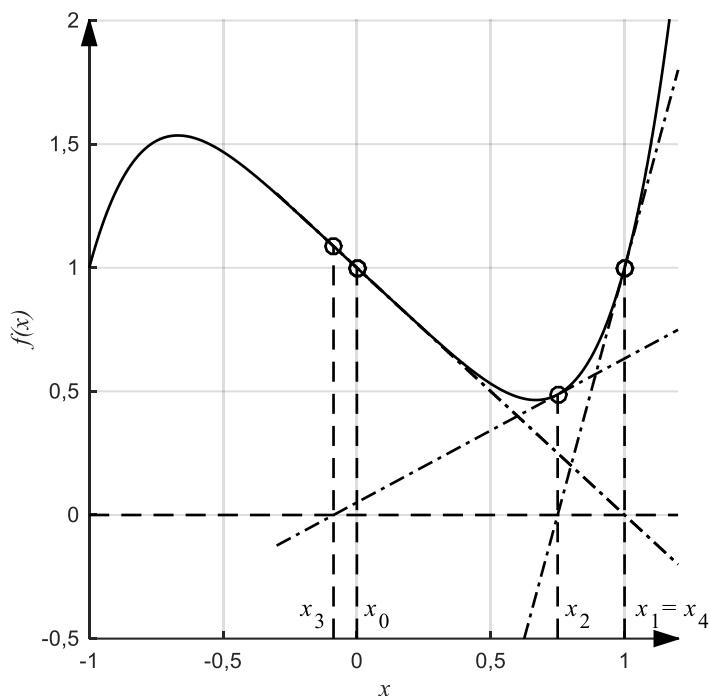
Rys. 6.10. Wyniki procesu iteracji dla funkcji (6.27): linia ciągła – funkcja $f(x)$, linia kropka-kreska – styczne

Jak widać na wykresie, iteracje rozpoczynające się w pobliżu pierwiastka, który chcemy znaleźć doprowadzają do innego pierwiastka.

5. Rozważmy funkcję

$$f(x) = x^5 - x + 1. \quad (6.28)$$

Miejszem zerowym funkcji jest $x \approx -1,1673$. Punktem startowym będzie $x_0 = 0$. Wyniki zostały przedstawione na rysunku 6.10.



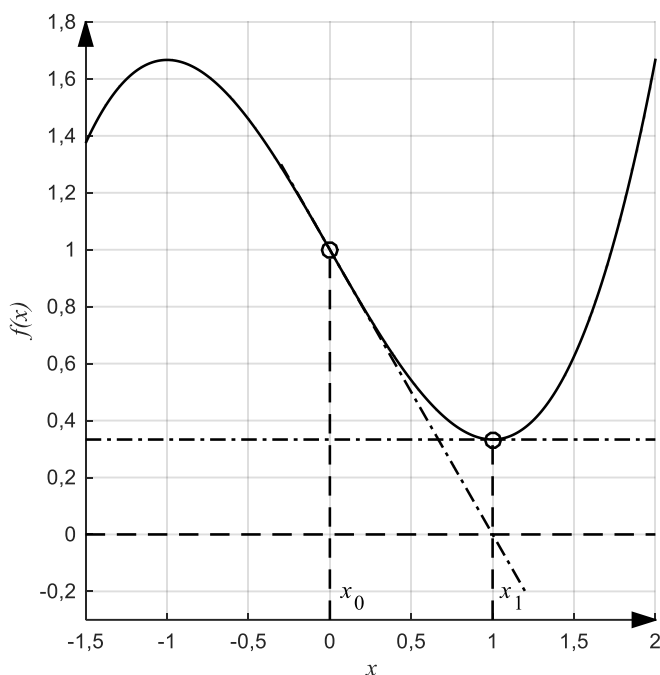
Rys. 6.11. Wyniki procesu iteracji dla funkcji (6.28): linia ciągła – funkcja $f(x)$, linia kropka-kreska – styczne

Otrzymany wykres pokazuje, że proces iteracji utyka w sąsiedztwie lokalnego minimum funkcji.

6. Jako ostatnią rozważmy funkcję

$$f(x) = \frac{x^3}{3} - x + 1. \quad (6.29)$$

Punktem startowym będzie $x_0 = 0$. Wyniki zostały przedstawione na rysunku 6.11.



Rys. 6.12. Wyniki procesu iteracji dla funkcji (6.29): linia ciągła – funkcja $f(x)$, linia kropka-kreska – styczne

Wyniki przedstawione na rysunku 6.12 obrazują przypadek kiedy po kolejnej iteracji trafiamy w punkt, w którym pochodna się zeruje i proces iteracji musi zostać przerwany.

Przypadki 1-6 obrazują najczęstsze problemy jakie można napotkać w trakcie zastosowań metod Newtona-Raphsona.

Lokalny charakter zbieżności metody Newtona-Raphsona jest jej największą wadą. Przy pewnych założeniach metoda jest zbieżna dla dowolnego punktu startowego, tak jak w twierdzeniu 6.2.

Twierdzenie 6.2 (o zbieżności metody Newtona-Raphsona – Kincaid, Cheney, 2006)

Jeżeli funkcja $f(x)$ jest dwukrotnie różniczkowalna i jej pochodne są ciągłe, jest rosnąca, wypukła i ma pierwiastek, to ten pierwiastek jest jedyny i metoda Newtona-Raphsona generuje ciąg zbieżny do tego pierwiastka dla dowolnego punktu początkowego.

Istnieją modyfikacje metody Newtona-Raphsona dające zbieżność globalną. Dwie z nich to:

- „tłumiona” (ang. damped) metoda Newtona,
- metoda Levenberga-Marquardta.

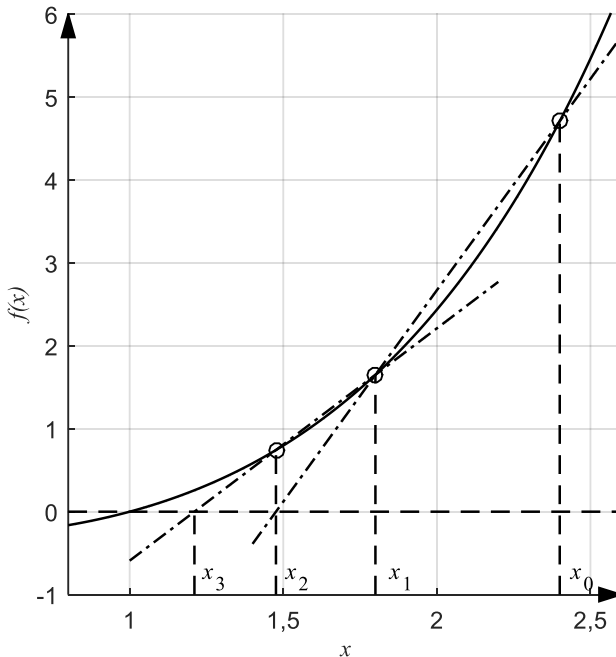
Idea obu tych metod bazuje na zmniejszeniu kroku, jaki jest wykonywany w każdej iteracji. Dla „tłumionej” metody Newtona wzór iteracyjny ma postać:

$$x_{i+1} = x_i - \alpha_k \frac{f(x_i)}{f'(x_i)}, \quad (6.30)$$

gdzie parametr α_k jest odpowiedzialny za redukcję kroku w każdej iteracji, w sposób zapewniający spełnienie nierówności $|f(x_{i+1})| \leq |f(x_i)|$.

6.5. Metoda siecznych

Czasem wyznaczenie pochodnej $f'(x)$, która jest konieczna w metodzie Newtona-Raphsona stwarza trudności. Niekiedy też obliczenie wartości pochodnej zajmuje znacznie więcej czasu od obliczenia wartości funkcji. W metodzie siecznych zamiast stycznymi posługujemy się siecznymi, co wymaga dwóch punktów startowych, ale eliminuje konieczności obliczania pochodnej.

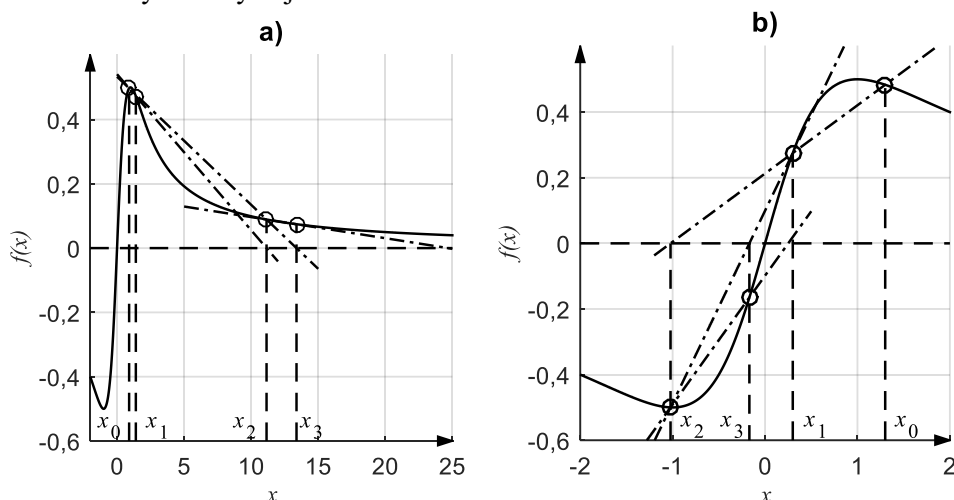


Rys. 6.13. Idea działania metody siecznych: linia ciągła – funkcja $f(x)$, linia kropka-kreska – sieczne

Jeżeli we wzorze opisującym iterację metody Newtona zastąpimy pochodną ilorazem różnicowym, otrzymamy opis iteracji w metodzie siecznych:

$$\begin{aligned} x_{i+1} &= x_i - \frac{f(x_i)}{f'(x_i)} \approx x_i - \frac{f(x_i)(x_i - x_{i-1})}{f(x_i) - f(x_{i-1})} \\ &= \frac{f(x_i)x_{i-1} - f(x_{i-1})x_i}{f(x_i) - f(x_{i-1})}. \end{aligned} \quad (6.31)$$

Metoda siecznych dla funkcji $f(x)$ z ciągłą drugą pochodną w pojedynczym pierwiastku ma rząd zbieżności $p = \frac{1+\sqrt{5}}{2} = 1,618 \dots$, jest więc wolniejsza od metody Newtona-Raphsona. Jeżeli jednak w czasie koniecznym na wykonanie jednej iteracji metody Newtona-Raphsona można wykonać dwie iteracje metody siecznych, bo czas obliczenia wartości pochodnej jest dłuższy od czasu wyznaczenia wartości funkcji, to proces używający metody siecznych będzie szybszy. Zbieżność metody siecznych jest także lokalna.



Rys. 6.14. Rozbieżna (a) i zbieżna (b) metoda siecznych: linia ciągła – funkcja $f(x)$, linia kropka-kreska – sieczne

Wzór (6.31) może być niestabilny numerycznie (generować bardzo duże wartości na skutek błędów zaokrągleń), jeśli jego mianownik dąży do zera szybciej niż licznik. Warto więc zapewnić, by mianownik był możliwie duży. Tak będzie jeśli punkty do poprowadzenia siecznej będą wybierane zawsze po różnych stronach pierwiastka.

6.6. Metoda *regula falsi*

Metoda *regula falsi* (czyli „falszywej prostej” a w terminologii angielskiej także „false position”) konstruuje sieczne między punktami, w których funkcja $f(x)$ jest różnego znaku.

Punktami początkowymi dla i -tej iteracji są:

$$x_i, \quad a_i, \quad f(x_i)f(a_i) < 0. \quad (6.32)$$

Następnie oblicza się punkt przecięcia siecznej z osią:

$$\mu_i = \frac{f(x_i)a_i - f(a_i)x_i}{f(x_i) - f(a_i)} \quad (6.33)$$

i wybiera punkty do następnej iteracji:

$$\left. \begin{matrix} x_{i+1} = \mu_i \\ a_{i+1} = a_i \end{matrix} \right\} \Leftarrow f(x_i)f(\mu_i) > 0, \quad \left. \begin{matrix} x_{i+1} = \mu_i \\ a_{i+1} = x_i \end{matrix} \right\} \Leftarrow f(x_i)f(\mu_i) < 0. \quad (6.34)$$

Algorytm w tej postaci wykazuje dużo podobieństw do metody bisekcji – jego zbieżność jest też liniowa. Jeżeli $f(x)$ jest ściśle rosnąca, lub ściśle malejąca, w rozpatrywanym przedziale, to jeden z punktów, przez które prowadzimy sieczną pozostaje zawsze ten sam, a to spowalnia zbieżność (problem retencji). Istnieje wiele modyfikacji metody *regula falsi*, które mają zapobiegać retencji. Polegają na zastąpieniu wartości $f(a_i)$ we wzorze (6.33) przez $mf(a_i)$, jeżeli miałyby być $a_i = a_{i+1}$. Współczynnik redukcyjny m można wybierać w różny sposób np.:

- $m = 1/2$ (algorytm Illinois),
- $m = 1 - \frac{f(\mu_i)}{f(x_i)}$ jeśli ta liczba jest dodatnia albo $m = 1/2$ (algorytm Andersona-Björka).

Te i inne modyfikacje metody *regula falsi* pozwalają uzyskać super-liniową zbieżność (rzęd metody $1 < p < 2$) i sprawiają, że metody z tej rodziny skutecznie konkurują z metodą Newtona-Raphsona.

Przykład 6.5

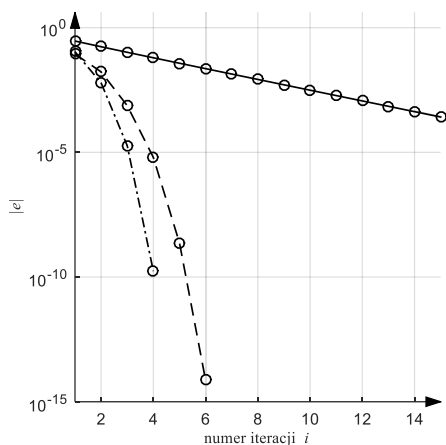
Rozważmy równanie $f(x) = 0$ z funkcją

$$f(x) = \ln(x). \quad (6.35)$$

Do rozwiązania podanego równania zostaną zastosowane metody: *regula falsi* z punktami startowymi $x_0 = 1,5$; $a_0 = 0,1$; siecznych z punktami startowymi $x_0 = 1,5$; $x_1 = 1,4$ oraz Newtona Raphsona z punktem startowym $x_0 = 0,5$.

Tabela 6.8. Porównanie wyników procesu iteracji trzema różnymi metodami

i	x_i m. regula falsi	x_i m. siecznych	x_i m. Newtona Raphsona
1	1,2903838152186443	0,9123086931019401	0,8918023378377534
2	1,1717237344854457	1,0168244522562710	0,9939233060488715
3	1,1027118628128103	1,0007469101108153	0,9999814993830553
4	1,0618689390277005	0,9999937350769671	0,999999998288626
5	1,0374292973063631	1,0000000023393793	1,0000000000000000
10	1,0031082427215972	1,0000000000000074	1,0000000000000000
15	1,0002604497871654	1,0000000000000000	1,0000000000000000
20	1,0000218404498959	1,0000000000000000	1,0000000000000000

Rys. 6.15. Przebieg błędów bezwzględnych dla trzech metod iteracyjnych: linia ciągła – metoda *regula falsi*, kreskowa – siecznych, kreska-kropka – Newtona-Raphsona

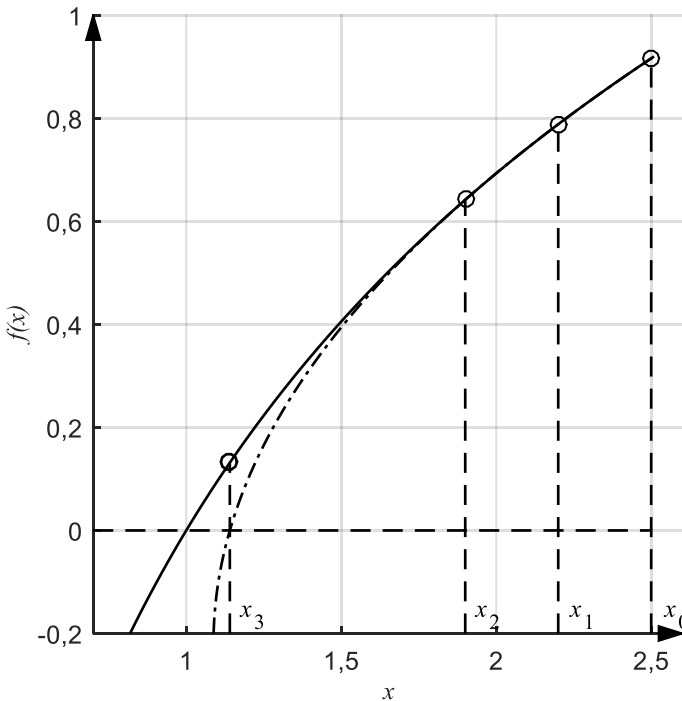
6.7. Odwrotna interpolacja kwadratowa (Inverse Quadratic Interpolation – IQI)

Dotychczas przedstawione metody rozwiązywania równań algebraicznych posługują się liniowym przybliżeniem nieliniowego równania. Równania kwadratowe można rozwiązywać równie skutecznie jak równania liniowe, dlategoż by nie skorzystać z lokalnego przybliżenia funkcji $f(x)$ parabolą?

Przypuśćmy, że mamy 3 wartości argumentu x : a , b i c , i odpowiadające im wartości funkcji y : $f(a)$, $f(b)$ i $f(c)$. Możemy interpolować te wartości wielomianem stopnia 2 i przyjmując za kolejne przybliżenie punkt, w którym parabola przecina oś x . Ale może zdarzyć się, że parabola nie przecina osi x – trójmian kwadratowy nie ma pierwiastków rzeczywistych. Zamiast budować wielomian interpolacyjny

stopnia 2 względem x możemy zbudować taki wielomian względem y (oznacmy go $P(y)$) – jego wykresem będzie „odwrócona” parabola. Taka parabola zawsze przetnie oś x i punkt przecięcia ($x = P(0), y = 0$) będzie następnym przybliżeniem w metodzie iteracyjnej (rys. 6.16). Kolejne przybliżenie pierwiastka określa wzór iteracyjny

$$\begin{aligned}
 x_{i+1} = & \frac{f_{i-1}f_i}{(f_{i-2} - f_{i-1})(f_{i-2} - f_i)} x_{i-2} \\
 & + \frac{f_{i-2}f_i}{(f_{i-1} - f_{i-2})(f_{i-1} - f_i)} x_{i-1} \\
 & + \frac{f_{i-1}f_{i-2}}{(f_i - f_{i-2})(f_i - f_{i-1})} x_i.
 \end{aligned}
 \tag{6.36}$$



Rys. 6.16. Zasada działania metody IQI: linia ciągła – funkcja $f(x)$, linia kropka-kreska – odwrotna parabola

Rząd zbieżności tej metody jest bliski 2 ($p = 1,8$), nie wymaga ona obliczania wartości pochodnej, ale jest także lokalnie zbieżna.

6.8. Złożone metody rozwiązywania równań nieliniowych

Przedstawione metody rozwiązywania równań nieliniowych są albo niezawodne i wolne (jak metoda bisekcji), albo lokalnie zbieżne i szybkie – jak metoda Newtona-Raphsona czy metoda odwrotnej interpolacji kwadratowej. Rozsądnym posunięciem jest połączenie różnych metod w jeden algorytm, tak by można było korzystać z zalety każdej z nich. Metoda o wysokim rzędzie zbieżności powinna być uruchamiana w pobliżu dokładnego rozwiązania, a od metody o dużym obszarze zbieżności należy rozpocząć obliczenia.

Przykładem takiego postępowania jest **algorytm Brendta**:

1 Startujemy od a i b takich, że $f(a)$ i $f(b)$ są różnych znaków.

2 Budujemy sieczną, której przecięcie z osią x daje punkt c między a i b .

3 Powtarzamy następujące kroki, dopóki $|b - a| < \text{eps} \cdot b$ lub $f(c) = 0$:

A Porządkujemy a, b i c , tak by:

- $f(a)$ i $f(b)$ były różnych znaków,
- $|f(b)| \leq |f(a)|$,
- c było poprzednią wartością b .

B Jeśli $c \neq a$, wykonujemy krok IQI.

C Jeśli $c = a$, wykonujemy krok metody siecznych.

D Jeśli wynik kroku IQI lub kroku metody siecznych jest wewnątrz $[a, b]$, akceptujemy go.

E Jeśli wynik kroku IQI lub kroku metody siecznych jest poza $[a, b]$ stosujemy bisekcję.

6.9. Uwarunkowanie pierwiastków równań nieliniowych

Błąd metody każdego ze sposobów obliczania pierwiastka równania nieliniowego zależy od liczby wykonanych iteracji. Jeżeli potrafimy oszacować błąd początkowy to można obliczyć liczbę iteracji konieczną do osiągnięcia zadanej dokładności.

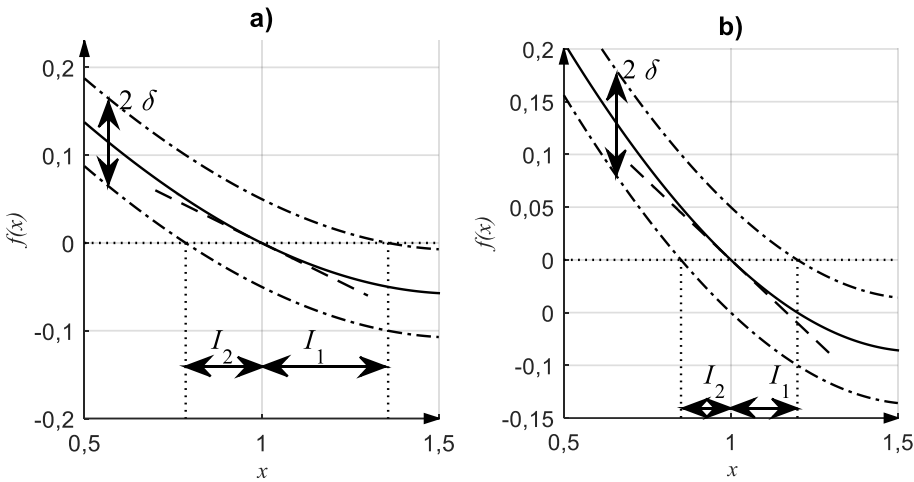
Każda z metod rozwiązania równania $f(x) = 0$ wymaga obliczenia wartości funkcji $f(x)$ w kolejnych iteracjach. Jest to oczywiście związane z błędem zaokrągleń. Jeżeli błąd obliczenia wartości funkcji nie przekracza δ , to każdą liczbę z przedziałów I_1, I_2 zaznaczonych na rysunku 6.17 można uznać za rozwiązanie. Wartość błędu (bezwzględnego) Δ zależy od δ , ale także od stromości funkcji $f(x)$ w pobliżu rozwiązania (rys. 16.7).

Jeżeli x_n jest wynikiem ostatniej iteracji a a dokładnym pierwiastkiem (niech $x_n < a$), funkcja f jest ciągła na $[x_n, a]$ i różniczkowalna w (x_n, a) , to na mocy

twierdzenia Lagrange'a o wartości średniej istnieje punkt $c \in [x_n, a]$, taki że $x_n - a = \frac{f(x_n)}{f'(c)}$. Można więc napisać

$$|x_n - a| \leq \frac{|f(x_n)|}{\min_{x_n, a} |f'(c)|} \approx \frac{\delta}{|f'(a)|} = \Delta. \quad (6.37)$$

Błąd wyznaczonego pierwiastka jest wprost proporcjonalny do błędowi obliczenia wartości funkcji, a odwrotnie proporcjonalny do modułu pochodnej funkcji $f(x)$ w pobliżu wyznaczonego pierwiastka.



Rys. 6.17. Wpływ wartości pochodnej funkcji $f'(a)$ na rozmiar przedziału Δ : moduł pochodnej funkcji na rys. a) jest mniejszy niż na rys. b), co pokazuje nachylenie stycznej (linia kreskowa)

Δ w wyrażeniu (6.37) jest oszacowaniem pierwszego rzędu błędzi $|x_n - a|$ i spełnia nierówność

$$\Delta \leq \max(I_1, I_2), \quad (6.38)$$

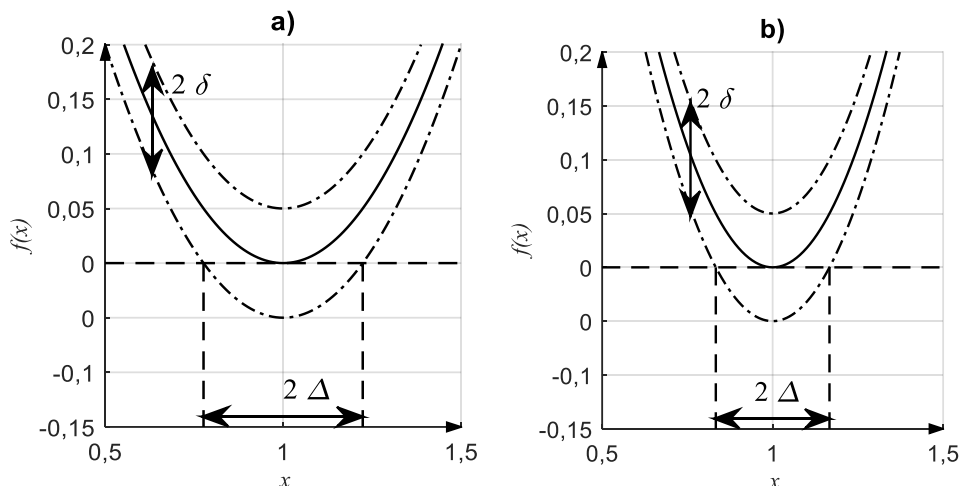
gdzie I_1, I_2 spełniają równania

$$f(a + I_1) = -\delta, \quad f(a - I_2) = \delta.$$

Przedstawiona analiza i rysunki 6.17 dotyczą przypadku jednokrotnych pierwiastków. Jak widać na rysunkach 6.18, inaczej będzie w przypadku pierwiastków wielokrotnych. Dla pierwiastka o krotności m zamiast wzoru (6.37) obowiązuje

$$\Delta = \left(\frac{\delta m!}{|f^{(m)}(a)|} \right)^{\frac{1}{m}}, \quad (6.39)$$

co oznacza większą wrażliwość pierwiastków wielokrotnych na błędy w obliczeniu $f(x)$.



Rys. 6.18. Wpływ wartości drugiej pochodnej funkcji $f''(a)$ na rozmiar przedziału Δ dla pierwiastka dwukrotnego: moduł drugiej pochodnej na rys. a) jest mniejszy niż na rys. b)

Zwykle błąd w obliczeniu funkcji $f(x)$ wynika z błędów (zaokrążeń) współczynników występujących w wyrażeniu określającym rozważaną funkcję, np. dla wielomianu są to współczynniki stojące przy kolejnych potęgach. Może być wtedy przedstawiony w postaci

$$\delta = \sigma |g(x)|, \quad (6.40)$$

gdzie σ reprezentuje błąd współczynnika a $g(x)$ składnik, który występuje z tym współczynnikiem. Zamiast wzoru (6.37) mamy wtedy

$$\Delta = |x_n - a| \approx \frac{\sigma |g(a)|}{|f'(a)|}. \quad (6.41)$$

Można też dla oceny uwarunkowania znalezione pierwiastka posłużyć się współczynnikiem wzmocnienia błędu

$$\frac{\Delta/|a|}{|\sigma g(a)|/|g(a)|} = \frac{|g(a)|}{|af'(a)|} \quad (6.42)$$

Inny sposób oceny wrażliwości pierwiastka daje następujące rozumowanie:

Rozważmy, zamiast równania $f(x) = 0$ równanie zaburzone

$$f(x) + \sigma g(x) = 0$$

i niech $\alpha(\sigma)$ oznacza pierwiastek tego równania w funkcji parametru zaburzenia σ , więc $\alpha_0 = \alpha(0)$ jest pierwiastkiem równania $f(x) = 0$. Istnienie $\alpha(\sigma)$ wynika z twierdzenia o funkcji uwikłanej. Zmiana pierwiastka będzie w przybliżeniu proporcjonalna do pochodnej $K = \left. \frac{d\alpha(\sigma)}{d\sigma} \right|_{\sigma=0}$. Różniczkowanie zaburzonego równania daje

$$\left. \frac{d}{d\sigma} [f(\alpha(\sigma)) + \sigma g(\alpha(\sigma))] \right|_{\sigma=0} = 0 \Rightarrow K = -\frac{g(\alpha(0))}{f'(\alpha(0))}.$$

Gdyby pierwiastek α_0 zbliżał się do innego, czyli stawałby się pierwiastkiem wielokrotnym, to $f'(\alpha(0))$ dążyłoby do zera, czyli współczynnik K dążyłoby do nieskończoności. Kolejny raz demonstrujemy tu kłopoty z obliczeniem wielokrotnych pierwiastków.

Przykład 6.6

Rozważmy wielomian

$$\begin{aligned} P(x) &= (x - 1)(x - 2) \dots (x - 14)(x - 15) = \\ &= a_{15}x^{15} + a_{14}x^{14} + \dots + a_1x + a_0, \end{aligned} \tag{6.43}$$

gdzie współczynniki a_k zebrano w tabeli 6.9.

Tabela 6.9. Współczynniki a_k wielomianu (6.43)

k	a_k	k	a_k
15	1	7	54631129553
14	-120	6	-272803210680
13	6580	5	1009672107080
12	-218400	4	-2706813345600
11	4899622	3	5056995703824
10	-78558480	2	-6165817614720
9	928095740	1	4339163001600
8	-8207628000	0	-1307674368000

W całym przykładzie pierwiastki wielomianu obliczano metodą opisaną w rozdziale 8.3. W przypadku wielomianu (6.43), gdy danymi wejściowymi były współczynniki z tabeli 6.9 całkowite błędy obliczonych pierwiastków nie przekraczały $3 \cdot 10^{-5}$ i zostały zapisane w tabeli 6.10.

Tabela 6.10. Błędy całkowite pierwiastków x_i wielomianu (6.43)

x_i	błąd całkowity	x_i	błąd całkowity
15	$0,002689686127155 \cdot 10^{-4}$	7	$0,025010476534248 \cdot 10^{-4}$
14	$-0,023103696964455 \cdot 10^{-4}$	6	$-0,005585864979452 \cdot 10^{-4}$
13	$0,086474202074527 \cdot 10^{-4}$	5	$0,000826836208390 \cdot 10^{-4}$
12	$-0,187230083739109 \cdot 10^{-4}$	4	$-0,000076775816815 \cdot 10^{-4}$
11	$0,261611922809379 \cdot 10^{-4}$	3	$0,000004117759467 \cdot 10^{-4}$
10	$-0,248652738559230 \cdot 10^{-4}$	2	$-0,000000113142828 \cdot 10^{-4}$
9	$0,164835190989976 \cdot 10^{-4}$	1	$0,000000001105782 \cdot 10^{-4}$
8	$-0,076803157806893 \cdot 10^{-4}$		

Wpływ zaburzenia współczynników wielomianu na jego pierwiastki przeanalizowano w trzech przypadkach.

1. Rozważmy zaburzenie współczynnika $a_{14} = -120$. Zaburzenie jest reprezentowane przez składnik $\sigma g(x) = \sigma x^{14}$.

Współczynnik K obliczony dla pierwiastka $x_i = i, i = 1, 2, \dots, 15$ będzie równy

$$K_i = -\frac{g(x_i)}{f'(x_i)} = -\frac{x_i^{14}}{\prod_{j=1, j \neq i}^{15} (x_i - x_j)}.$$

Dla pierwiastka $x_1 = 1$ wynosi on $K_1 = -\frac{1}{(1-2)(1-3)\dots(1-15)} = 1,1471 \cdot 10^{-11}$,

a dla pierwiastka $x_{15} = 15$ wynosi on $K_{15} = -\frac{15^{14}}{(15-1)(15-2)\dots(15-14)} = -3,3486 \cdot$

10^5 więc pierwiastek x_{15} jest dużo bardziej wrażliwy (źle uwarunkowany) na zmianę współczynnika a_{14} niż pierwiastek x_1 (pierwiastek dobrze uwarunkowany). Widać to także w tabelach 6.11 i 6.12, w których pokazano obliczone pierwiastki wielomianu z zaburzonym współczynnikiem a_{14} . W obu tabelach widać, że zmiana zaburzenia nawet o 5 rzędów wielkości (z $\sigma_1 = 10^{-10}$ do $\sigma_1 = 10^{-5}$) powoduje nieznaczną zmianę pierwiastka x_1 (błąd względny zmienił się z wartości $1,3011 \cdot 10^{-13}$ na $3,8192 \cdot 10^{-14}$ i jest właściwie taki sam jak w przypadku wyznaczania pierwiastka wielomianu niezaburzonego – tabela 6.10), natomiast dla pierwiastka x_{15} błąd względny zmienia się z $2,2449 \cdot 10^{-6}$ na $5,1615 \cdot 10^{-2}$, podczas gdy przy obliczeniach z wielomianu niezaburzonego wynosił $1,79 \cdot 10^{-8}$.

2. Jeśli zaburzenie dotyczy współczynnika przy x , to jest opisane przez $\sigma g(x) = \sigma x$ i odpowiednie współczynniki K wynoszą

$$K_i = -\frac{g(x_i)}{f'(x_i)} = -\frac{x_i}{\prod_{j=1, j \neq i}^{15} (x_i - x_j)},$$

$$K_1 = -\frac{1}{(1-2)(1-3)\dots(1-15)} = 1,1471 \cdot 10^{-11},$$

$$K_{15} = -\frac{15}{(15-1)(15-2)\dots(15-14)} = -1,7206 \cdot 10^{-10}.$$

Wartości współczynników K_1 i K_{15} wskazują, że oba pierwiastki x_1 i x_{15} są mało wrażliwe na zmiany współczynnika a_1 (są dobrze uwarunkowane).

3. Ponownie wprowadźmy do współczynnika a_{14} zaburzenie $\sigma_1 = 10^{-10}$. Rozważmy dwa skrajne pierwiastki $x_1 = 1$ i $x_{15} = 15$. Wyznaczmy dla nich oszacowanie błędu bezwzględnego wywołanego zaburzeniem z zależności (6.41):

$$\Delta_1 \approx \frac{10^{-10} \cdot 1^{14}}{|P'(1)|} = 1,1470 \cdot 10^{-21}, \quad \Delta_{15} \approx \frac{10^{-10} \cdot 15^{14}}{|P'(15)|} = 3,3486 \cdot 10^{-5}. \quad (6.44)$$

Z tabeli 6.11 odczytujemy całkowite błędy bezwzględne $1,3 \cdot 10^{-13}$ i $3 \cdot 10^{-5}$ odpowiednio, podczas gdy przy obliczeniach z niezaburzonych współczynników wynosiły one $1,3 \cdot 10^{-13}$ i $3 \cdot 10^{-7}$. Przy zaburzeniu $\sigma_1 = 10^{-10}$ zmiany pierwiastków są więc niewielkie.

Tabela 6.11. Pierwiastki wielomianu z zaburzonym współczynnikiem a_{14} o wartość $\sigma_1 = 10^{-10}$

1,000000000000130	5,999999610252106	11,00044704333125
1,999999999991920	7,000003275663333	11,99945540553272
3,000000000178619	7,999979797751623	13,00041530798098
3,99999997033527	9,000085097570430	13,99982015336151
5,000000037768315	9,999760599918055	15,00003367376527

Jak widać w powyższej tabeli pierwiastki uległy zmianie w niewielkim stopniu. Można także zaobserwować, że pierwiastki o większym module są bardziej wrażliwe na zmiany współczynników wielomianu.

Teraz wprowadźmy do współczynnika a_{14} zaburzenie $\sigma_1 = 10^{-5}$. Rozważmy dwa skrajne pierwiastki $x_1 = 1$ i $x_{15} = 15$. Wyznaczmy dla nich oszacowanie błędu z zależności (6.41):

$$\Delta_1 \approx \frac{10^{-5} \cdot 1^{14}}{|P'(1)|} = 1,1470 \cdot 10^{-21}, \quad \Delta_{15} \approx \frac{10^{-5} \cdot 15^{14}}{|P'(15)|} = 3,3486. \quad (6.45)$$

Wartości te zgadzają się z danymi podanymi w tabeli 6.12.

Tabela 6.12. Pierwiastki wielomianu z zaburzonym współczynnikiem a_{14} o wartość $\sigma_1 = 10^{-5}$

0,9999999999999618	5,982864222364555	11,2814126882532 - 2,127109368267351j
1,999999999978124	7,248220442633894 - 0,3005981537291766j	11,2814126882532 + 2,127109368267351j
3,000000049695138	7,248220442633894 + 0,3005981537291766j	14,1501903395076- 1,835664246386915j
3,999988797733152	8,94129207708574 - 1,345216085825933j	14,1501903395076+ 1,835664246386915j
5,000702553458658	8,94129207708574 + 1,345216085825933j	15,77422328180954

Zaburzenie miejsca zerowego jest pięć rzędów wielkości większe od zaburzenia pojedynczego współczynnika. W rzeczywistości miejsca zerowe tak zaburzonego wielomianu stają się nawet zespolone (tabela 6.12).

Sytuacja komplikuje się dodatkowo, jeśli na skutek zaburzenia pierwiastki zbliżają się do siebie. O dalszych zmianach będą wtedy decydowały wyższe pochodne $\left. \frac{d^k \alpha(\sigma)}{d\sigma^k} \right|_{\sigma=0}$. Dotyczy to przede wszystkim tych pierwiastków, dla których odległość od „sąsiadów” jest mniejsza od ich modułów.

Przykład 6.7

Rozważmy wielomian

$$\begin{aligned}
 P(x) &= 135660x^4 - 255877x^3 - 2390x^2 + 232088x \\
 &\quad - 109440 = \\
 &= 17 \cdot 19 \cdot 20 \cdot 21 \left(x + \frac{20}{21}\right) \left(x - \frac{16}{17}\right) \left(x - \frac{18}{19}\right) \left(x - \frac{19}{20}\right).
 \end{aligned} \tag{6.46}$$

Zdefiniujmy dwa „słabo” zaburzone wielomiany

$$\begin{aligned}
 P_1(x) &= P(x) - 0,1 = 135660x^4 - 255877x^3 - 2390x^2 + \\
 &\quad + 232088x - 109440,1,
 \end{aligned} \tag{6.47}$$

$$\begin{aligned}
 P_2(x) &= P(x) + 0,1 = 135660x^4 - 255877x^3 - 2390x^2 + \\
 &\quad + 232088x - 109439,9.
 \end{aligned} \tag{6.48}$$

Wielomiany $P_1(x)$ i $P_2(x)$ mogą być określone jako „słabo” zaburzone ponieważ względna zmiana współczynnika a_0 jest mniejsza niż 10^{-6} . W tabeli 6.13 zamieszczono wyznaczone pierwiastki wielomianów $P_1(x)$ i $P_2(x)$.

Tabela 6.13. Pierwiastki wielomianów (6.46) – (6.48)

$P(x)$	$P_1(x)$	$P_2(x)$
-0,9523809523809523	-0,9523808446660023	-0,9523810600958650
0,9411764705882353	0,950361390079782 + 0,0056291690812024j	0,942135368071371 + 0,0053667563451038j
0,9473684210526315	0,950361390079782 - 0,0056291690812024j	0,942135368071371 - 0,0053667563451038j
0,9500000000000000	0,9378220037663492	0,9542742632130336

Zamieszczony przykład potwierdza, że najbardziej wrażliwe na zmiany współczynników wielomianu są pierwiastki, które leżą blisko siebie (odległość między nimi jest mniejsza niż ich moduły). W zaburzonych wielomianach $P_1(x)$ i $P_2(x)$ dwa z bliskich sobie pierwiastków rozszczepiły się na pierwiastki zespolone (wiersz trzeci i czwarty w tabeli 6.13). Pierwiastek o wartości ujemnej, którego odległość od pozostałych pierwiastków jest „duża” (większa niż jego moduł) zmienia się nieznacznie (błąd względny mniejszy niż 0,00002%).

6.10. Układy równań nieliniowych

Układ n równań nieliniowych z n niewiadomymi można zapisać w postaci skalarnej:

$$f_i(x_1, x_2, \dots, x_n) = 0, \quad i = 1, \dots, n, \quad (6.49)$$

lub macierzowej:

$$F(X) = 0, \quad X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad F(\cdot) = \begin{bmatrix} f_1(\cdot) \\ f_2(\cdot) \\ \vdots \\ f_n(\cdot) \end{bmatrix}. \quad (6.50)$$

Podobnie jak w przypadku pojedynczego równania rozwiązaniem będzie dowolny, rzeczywisty wektor X^* spełniający równanie.

Jeżeli równanie (6.50) można zapisać w równoważnej formie:

$$G(X) = X, \quad X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad G(\cdot) = \begin{bmatrix} g_1(\cdot) \\ g_2(\cdot) \\ \vdots \\ g_n(\cdot) \end{bmatrix}, \quad (6.51)$$

to można zastosować do niego metodę iteracji prostej. Metoda opiera się na twierdzeniu o punkcie stałym, które przy powyższych oznaczeniach można sformułować:

Twierdzenie 6.3 (o punkcie stałym, *Kincaid, Cheney, 2006*)

Jeżeli funkcje g_i są ciągłe i istnieje liczba $0 \leq c < 1$, taka że dla każdego X, Y zachodzi

$$\|G(X) - G(Y)\| \leq c\|X - Y\|, \quad (6.52)$$

to istnieje dokładnie jedno rozwiązanie X^* równania $G(X) = X$. Iteracje $X_{i+1} = G(X_i)$ zbiegają do X^* dla dowolnych warunków początkowych X_0 , a błąd po i iteracjach spełnia nierówność:

$$\|X^* - X_i\| \leq \frac{c^i}{1-c} \|X_i - X_0\|. \quad (6.53)$$

Jest to znane **twierdzenie Banacha** o punkcie stałym odwzorowania zwężającego (które obowiązuje w przestrzeniach metrycznych zupełnych, a tu zostało podane w wersji ograniczonej do przestrzeni R^n , która jest przestrzenią metryczną zupełną) i jego dowód można znaleźć w wielu podręcznikach⁶.

Zbieżność metody iteracji prostej jest liniowa, ze stałą asymptotyczną błędu c . Jeżeli wyjściowym równaniem jest równanie (6.50) to istnieje wiele sposobów zbudowania równoważnego równania w postaci (6.51). W szczególności może to być każde równanie

$$X = G(X) = X - AF(x), \quad (6.54)$$

w którym A jest nieosobliwą macierzą kwadratową. Przez dobór macierzy A można wpływać na zbieżność iteracji.

6.11. Metoda Newtona-Raphsona dla układów równań

Rozwinięcie w szereg Taylora i liniowe przybliżenie nieliniowego odwzorowania istnieją nie tylko w przypadku jednowymiarowym, kiedy funkcja $f(x)$ odwzorowuje przestrzeń R w R , ale i w przypadku wielowymiarowym, kiedy $F(X)$ odwzorowuje przestrzeń R^n w R^n . Odpowiednikiem wzoru (6.9) w przypadku wielowymiarowym jest

$$0 = F(X^*) = F(X_i) + F'(X_i)(X^* - X_i) + \frac{1}{2!} \begin{bmatrix} (X^* - X_i)^T H_1(X_i)(X^* - X_i) \\ \vdots \\ (X^* - X_i)^T H_n(X_i)(X^* - X_i) \end{bmatrix} + \dots, \quad (6.55)$$

⁶ Więcej informacji np. w książce Andrzej Granas, James Dugundji, Fixed Point Theory (2003) Springer-Verlag.

gdzie $F'(X)$ oznacza macierz pierwszych pochodnych (macierz Jacobiego)

$$F'(X) = \begin{bmatrix} \frac{\partial f_1(x_1, \dots, x_n)}{\partial x_1} & \frac{\partial f_1(x_1, \dots, x_n)}{\partial x_2} & \dots & \frac{\partial f_1(x_1, \dots, x_n)}{\partial x_n} \\ \frac{\partial f_2(x_1, \dots, x_n)}{\partial x_1} & \frac{\partial f_2(x_1, \dots, x_n)}{\partial x_2} & \dots & \frac{\partial f_2(x_1, \dots, x_n)}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n(x_1, \dots, x_n)}{\partial x_1} & \frac{\partial f_n(x_1, \dots, x_n)}{\partial x_2} & \dots & \frac{\partial f_n(x_1, \dots, x_n)}{\partial x_n} \end{bmatrix}, \quad (6.56)$$

a $H_k(X)$ macierz drugich pochodnych funkcji $f_k(x_1, x_2, \dots, x_n)$ (macierz Hessego):

$$H_k(X) = \begin{bmatrix} \frac{\partial^2 f_k(x_1, \dots, x_n)}{\partial x_1^2} & \frac{\partial^2 f_k(x_1, \dots, x_n)}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 f_k(x_1, \dots, x_n)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f_k(x_1, \dots, x_n)}{\partial x_2 \partial x_1} & \frac{\partial^2 f_k(x_1, \dots, x_n)}{\partial x_2^2} & \dots & \frac{\partial^2 f_k(x_1, \dots, x_n)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f_k(x_1, \dots, x_n)}{\partial x_n \partial x_1} & \frac{\partial^2 f_k(x_1, \dots, x_n)}{\partial x_n \partial x_2} & \dots & \frac{\partial^2 f_k(x_1, \dots, x_n)}{\partial x_n^2} \end{bmatrix}. \quad (6.57)$$

Jeżeli zignorujemy składniki zawierające pochodne rzędu wyższego niż 1, to otrzymamy liniowe przybliżenie równania (6.55), którego rozwiązaniem będzie kolejne przybliżenie rozwiązania dokładnego:

$$0 = F(X_i) + F'(X_i)(X_{i+1} - X_i). \quad (6.58)$$

Rozwiązanie tego równania można zapisać w postaci odpowiadającej skalarnemu wzorowi opisującemu iteracje metody Newtona-Raphsona:

$$X_{i+1} = X_i - [F'(X_i)]^{-1}F(X_i), \quad (6.59)$$

ale wykonując kolejne iteracje nie odwracamy macierzy pochodnych $F'(X_i)$, ale rozwiązujemy, najczęściej metodą eliminacji Gaussa, względem $D_{i+1} := X_{i+1} - X_i$ równanie:

$$F'(X_i)D_{i+1} = -F(X_i) \quad (6.60)$$

i obliczamy

$$X_{i+1} = X_i + D_{i+1}. \quad (6.61)$$

Metoda Newtona-Raphsona w przypadku układów równań jest więc dość kosztowna obliczeniowo – wymaga rozwiązania układu n równań liniowych w każdej iteracji oraz wcześniejszego wyznaczenia współczynników tych równań, do czego jest potrzebne obliczenie wartości pochodnych cząstkowych. Istnieją warianty metody Newtona-Raphsona, w których macierz $F'(X_i)$ jest modyfikowana co kilka iteracji.

Metoda Newtona-Raphsona w przypadku układów równań jest także zbieżna lokalnie i kwadratowo w tym sensie, że $\|X^* - X_{i+1}\|$ jest (z grubsza) proporcjonalne do $\|X^* - X_i\|^2$.

6.12. Metoda Broydena

W pierwszej metodzie Broydena obliczamy X_{i+1} analogicznie jak w metodzie Newtona-Raphsona:

$$J_i D_{i+1} = -F(X_i), \quad X_{i+1} = X_i + D_{i+1}, \quad (6.62)$$

ale zamiast Jakobianu $F'(X_i)$ używane jest jego iteracyjnie poprawiane przybliżenie

$$J_{i+1} = J_i + \frac{1}{D_{i+1}^T D_{i+1}} [F(X_{i+1}) - F(X_i) - J_i D_{i+1}] D_{i+1}^T. \quad (6.63)$$

Metoda wymaga początkowego przybliżenia Jakobianu J_0 .

W drugiej metodzie Broydena, żeby uniknąć rozwiązywania układu równań liniowych w każdej iteracji, używa się przybliżenia odwrotności Jakobianu:

$$\begin{aligned} X_{i+1} &= X_i - B_i F(X_i), \\ D_{i+1} &= X_{i+1} - X_i, \end{aligned} \quad (6.64)$$

$$B_{i+1} = B_i + \frac{1}{D_{i+1}^T B_i [F(X_{i+1}) - F(X_i)]} [D_{i+1} - B_i (F(X_{i+1}) - F(X_i))] D_{i+1}^T B_i$$

Metoda wymaga początkowego przybliżenia odwrotności Jakobianu B_0 .

Obie metody Broydena są zbieżne lokalnie, super-liniowo ($1 < p < 2$). W obu też znajomość dobrych przybliżeń początkowych Jakobianu lub jego odwrotności ma decydujące znaczenie.

Przykład 6.8

Rozważmy układ równań

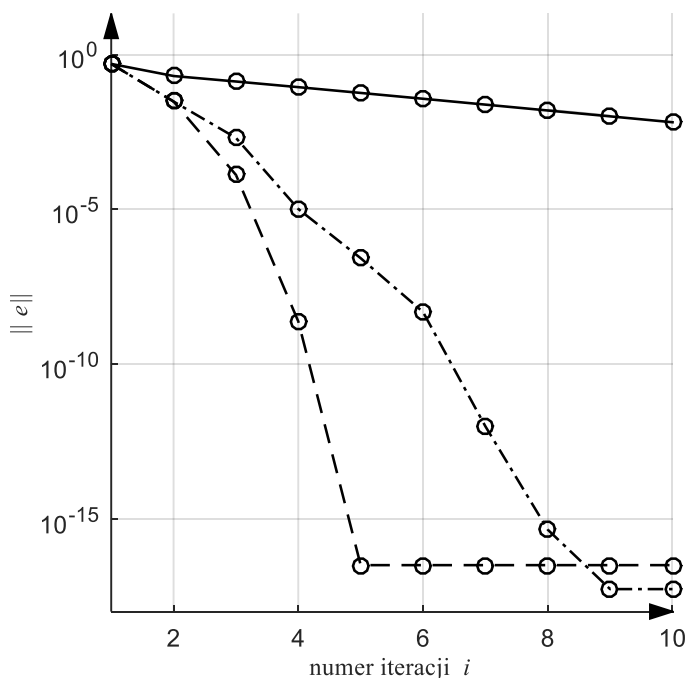
$$\begin{aligned} f_1(x, y) &= xy - x + y + 3 = 0 \\ f_2(x, y) &= x^2 + 2x + y^2 - 2y - 15 = 0. \end{aligned} \quad (6.65)$$

Dokładnym rozwiązaniem układu równań (6.65) jest para liczby $x = 0, y = -3$. Do rozwiązania układu równań (6.65) zostały zastosowane trzy metody: iteracji prostej, Newtona-Raphsona oraz druga metoda Broydena.

Wyniki kolejnych iteracji zostały przedstawione w tabeli 6.15 oraz na rysunku 6.19.

Tabela 6.15. Wartości normy błędu $\|e_i\| = \sqrt{(x_i - x)^2 + (y_i - y)^2}$ w kolejnych iteracjach

i	metoda iteracji prostej	metoda Newtona Raphsona	metoda Broydena
0	0,5000000000000000	0,5000000000000000	0,5000000000000000
1	0,203049255108213	0,030417628009613	0,030417628009613
2	0,134918896620933	0,000129524401242	0,002036710419999
3	0,088628047418504	0,000000002372461	0,000010239495889
4	0,057661675202227	0,000000000000000	0,000000263433444
5	0,037365661237009	0,000000000000000	0,000000004803276



Rys. 6.19. Wykres normy błędu $\|e_i\| = \sqrt{(x_i - x)^2 + (y_i - y)^2}$ dla trzech metod: linia ciągła – metoda iteracji prostej, kreskowa – metoda Newtona Raphsona, kropka-kreska – metoda Broydena

6.13. Rozwiązywanie układów równań nieliniowych drogą minimalizacji

Punkt X^* , który jest rozwiązaniem równania $F(X) = 0$ jest jednocześnie punktem minimum skalarnej funkcji

$$g(X) = F^T(X)F(X) = \sum_{i=1}^n f_i^2(x_1, \dots, x_n). \quad (6.66)$$

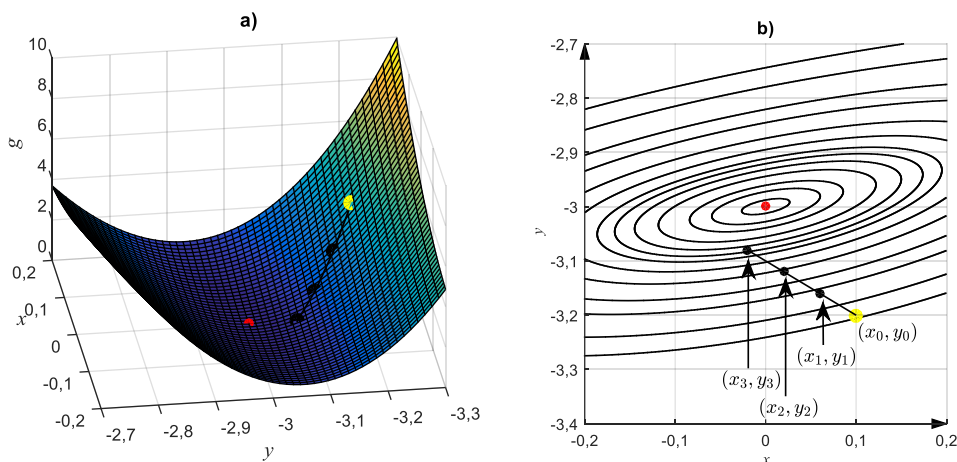
Istnieje wiele iteracyjnych metod poszukiwania minimum funkcji wielu zmiennych. Jedne z nich wykorzystują pochodne funkcji $g(X)$ (metody: gradientu prostego, najszybszego spadku, gradientu sprzężonego, Newtona-Raphsona, Marquardta), innym wystarczy obliczanie wartości $g(X)$ (metody: Rosenbrocka, Hooke'a-Jeeves'a, simpleks, Neldera i Meada, złotego podziału, interpolacyjne). Odrębną grupę stanowią algorytmy ewolucyjne, które są odmianą poszukiwania stochastycznego. Poszukiwanie minimum jest szczególnie skuteczne, jeśli minimum to jest jedyne i funkcja $g(X)$, przynajmniej w okolicy minimum, jest wypukła, czego wystarczającym warunkiem jest by jej hesjan był dodatnio określony.

Przykład 6.9

Rozważmy ponownie układ równań (6.65) z przykładu 6.8. Zdefiniujmy funkcję $g(X) = g(x, y)$ zgodnie ze wzorem (6.66):

$$g(x, y) = (xy - x + y + 3)^2 + (x^2 + 2x + y^2 - 2y - 15)^2, \quad (6.67)$$

którą przedstawiono na rysunku 6.20. Do poszukiwania minimum funkcji (6.67) użyto metody Hooke'a-Jeeves'a.



Rys. 6.20. Wykres funkcji $g(X)$ (a) oraz jej poziomic (b), żółta kropka – punkt startowy, czerwona – rozwiązanie dokładne, czarne – wynik kolejnych iteracji

6.14. Iteracyjne metody rozwiązywania układów równań liniowych

Metody iteracyjne mogą być także stosowane w przypadku układu równań liniowych rozważanych w rozdziale drugim, czyli

$$Ax = b. \quad (6.68)$$

Metody opierają się na twierdzeniu, że ciąg określony równaniem iteracyjnym

$$x_{i+1} = Mx_i + w \quad (6.69)$$

przy dowolnym wektorze x_0 jest zbieżny do jedyne punktu granicznego x_* spełniającego równanie

$$x_* = Mx_* + w \quad (6.70)$$

wtedy i tylko wtedy, gdy macierz M spełnia warunek

$$\lambda_{max}(M) < 1, \quad (6.71)$$

gdzie M jest pewną macierzą kwadratową, a w – wektorem.

Dla równania (6.68) równanie iteracyjne (6.69) jest określone wzorem

$$x_{i+1} = (I - NA)x_i + Nb, \quad (6.72)$$

gdzie macierz pomocnicza N jest wybierana w sposób zapewniający spełnienie warunku (6.71), czyli

$$\lambda_{max}(I - NA) < 1. \quad (6.73)$$

Wzór iteracyjny (6.72) określa całą rodzinę metod iteracyjnych do której należą na przykład:

- **Metoda Jacobiego**, w której równanie iteracji ma postać:

$$x_{i+1} = -D^{-1}(L + U)x_i + D^{-1}b, \quad (6.74)$$

- **Metoda Gaussa-Seidela**, w której równanie iteracji ma postać:

$$x_{i+1} = -(D + L)^{-1}Ux_i + (D + L)^{-1}b, \quad (6.75)$$

- **Metoda nadrelaksacji**, która jest zmodyfikowaną metodą Gaussa-Seidla o równaniu

$$x_{i+1} = (1 - \omega)x_i - \omega(D + L)^{-1}Ux_i + \omega(D + L)^{-1}b, \quad 1 < \omega < 2, \quad (6.76)$$

gdzie macierze: D – diagonalna, L – macierz trójkątna dolna z zerami na głównej przekątnej, U – macierz trójkątna górna z zerami na głównej przekątnej spełniają równanie

$$A = L + D + U. \quad (6.77)$$

Powyższe metody wymagają wstępnej zmiany kolejności równań w taki sposób, aby na diagonalu macierzy A znajdowały się tylko elementy niezerowe.

Odrębną klasę stanowią metody:

- Metoda najszybszego spadku określona wzorem

$$x_{i+1} = x_k + \frac{(b - Ax_k)^T (b - Ax_k)}{(b - Ax_k)^T A (b - Ax_k)} (b - Ax_k), \quad (6.78)$$

- Metoda najmniejszego residuum

$$x_{i+1} = x_k + \frac{(b - Ax_k)^T A^T (b - Ax_k)}{(b - Ax_k)^T A^T A (b - Ax_k)} (b - Ax_k). \quad (6.79)$$

Dla dużych układów równań nakład obliczeń metod iteracyjnych jest na ogół znacznie większy niż dla metod przedstawionych w rozdziale drugim. Jednak w przypadku układów równań, w których macierz A jest tak zwaną **macierzą rzadką** (duża liczba – np. 95% zerowych elementów macierzy) metody iteracyjne mogą stanowić alternatywę dla metod klasycznych z rozdziału drugiego. Duże układy równań z macierzami rzadkimi występują w analizie wielu problemów technicznych, takich jak: badanie sieci elektrycznych, modele systemów ekonomicznych, zjawisko dyfuzji, promieniowania, elastyczności itp. oraz w numerycznych metodach rozwiązywania równań różniczkowych cząstkowych.

7. Pierwiastki wielomianów

7.1. Operacje na wielomianach

Z zasadniczego twierdzenia algebry (dodatek D1) wynika, że wielomian $P(x)$ stopnia $n > 0$, o rzeczywistych współczynnikach ma n pierwiastków, czyli liczb spełniających równanie $P(x) = 0$. Pierwiastki mogą być pojedyncze lub wielokrotne. Suma krotności wszystkich pierwiastków jest równa stopniowi wielomianu. Pierwiastki wielomianu o rzeczywistych współczynnikach mogą być rzeczywiste lub zespolone parami sprzężone.

Problem numerycznego wyznaczania pierwiastków wielomianu polega na obliczeniu wszystkich pierwiastków wielomianu $P(x)$. Jest to więc nieco inne zagadnienie niż rozwiązywanie równania nieliniowego, gdzie był poszukiwany jeden, rzeczywisty pierwiastek, ale także rozwiązuje się je metodami iteracyjnymi.

Obok obliczania wartości wielomianu, co było omówione w rozdziale 3, do wyznaczania pierwiastków wielomianu będą używane operacje wyznaczania pochodnej wielomianu i dzielenia wielomianu przez wielomian. Schemat Hornera omówiony w rozdziale 3 jest przydatny także w tych zadaniach. Metoda opiera się na przedstawieniu wielomianu w postaci

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_2 x^2 + a_1 x + a_0 = \left(\left(\dots \left((a_n x + a_{n-1}) x + a_{n-2} \right) x + \dots + a_2 \right) x + a_1 \right) x + a_0, \quad (7.1)$$

czyli wartość $P(x)$ można obliczyć za pomocą wzoru rekurencyjnego

$$\begin{aligned} b_n &:= a_n, \\ b_{n-1} &:= a_{n-1} + b_n x, \\ &\dots \\ b_i &:= a_i + b_{i+1} x, \\ &\dots \\ b_0 &:= a_0 + b_1 x = P(x). \end{aligned} \quad (7.2)$$

Schemat obliczania wartości pochodnej można prześledzić na przykładzie wielomianu sześciennego. Pochodna wielomianu może być przedstawiona, po wykorzystaniu praw różniczkowania iloczynu, w następujący sposób:

$$\begin{aligned}
 P'(x) &= \frac{d}{dx}(a_3x^3 + a_2x^2 + a_1x + a_0) = \\
 &= \frac{d}{dx}[(a_3x + a_2)x + a_1]x + a_0 = \\
 &= \left[\frac{d}{dx}((a_3x + a_2)x + a_1) \right]x + ((a_3x + a_2)x + a_1) = \\
 &= \left[\left[\frac{d}{dx}(a_3x + a_2) \right]x + (a_3x + a_2) \right]x + ((a_3x + a_2)x + a_1) = \\
 &= [[a_3]x + (a_3x + a_2)]x + ((a_3x + a_2)x + a_1).
 \end{aligned} \tag{7.3}$$

Zgodnie z (7.2), wartość pochodnej $P'(x)$ wielomianu w punkcie x można obliczyć wyznaczając kolejno wartości: $b_2 = a_3x + a_2$, $b_1 = b_2x + a_1 = (a_3x + a_2)x + a_1$, a w końcu (dla wielomianu sześciennego):

$$P'(x) = [[b_3]x + b_2]x + b_1. \tag{7.4}$$

Wartość wielomianu $P(x)$ i pochodnej $P'(x)$ w tym samym punkcie x mogą być obliczane we wspólnym schemacie rekurencyjnym:

$$\begin{aligned}
 d_n &:= b_n := a_n, \\
 &\dots \\
 b_i &:= a_i + b_{i+1}x, \quad d_i := b_i + d_{i+1}x, \\
 &\dots \\
 b_1 &:= a_1 + b_2x, \quad d_1 := b_1 + d_2x = P'(x), \\
 b_0 &:= a_0 + b_1x = P(x).
 \end{aligned} \tag{7.5}$$

Dzielenie wielomianu przez czynnik liniowy można opisać wzorami:

$$\begin{aligned}
 P(x) &= a_nx^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0 = \\
 &= (x - z_0)(b_{n-1}x^{n-1} + b_{n-2}x^{n-2} + \dots + b_1x + b_0) + R(z_0), \\
 b_n &= 0, \quad b_k = a_{k+1} + z_0b_{k+1}, \quad k = n - 1, n - 2, \dots, 0, \\
 R(z_0) &= a_0 + z_0b_0.
 \end{aligned} \tag{7.6}$$

Podobne wzory można wyprowadzić dla dzielenia przez wielomian wyższego stopnia. Przy dzieleniu przez trójmian kwadratowy mamy:

$$\begin{aligned}
 P(x) &= a_nx^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0 = \\
 &= (x^2 + rx + q)(b_{n-2}x^{n-2} + b_{n-3}x^{n-3} + \dots + b_1x + b_0) + \\
 &\quad + A(r, q)x + B(r, q), \\
 b_n &= b_{n-1} = 0, \quad b_k = a_{k+2} - rb_{k+1} - qb_{k+2}, \\
 &\quad k = n - 2, n - 3, \dots, 0, \\
 A(r, q) &= a_1 - rb_0 - qb_1, \quad B(r, q) = a_0 - qb_0.
 \end{aligned} \tag{7.7}$$

7.2. Deflacja

Jeżeli został znaleziony rzeczywisty pierwiastek z_0 wielomianu $P(x)$, to zgodnie z twierdzeniem Bézouta (dodatek D1) wielomian dzieli się bez reszty przez czynnik liniowy $x - z_0$. Jeżeli znaleziono pierwiastek zespolony $z_1 = p + jq$, to drugim pierwiastkiem wielomianu jest $z_2 = p - jq$ i wielomian dzieli się bez reszty przez trójmian kwadratowy $(x - z_1)(x - z_2) = (x - p - jq)(x - p + jq) = x^2 + (-2p)x + (p^2 + q^2)$. Takie dzielenie, czyli obliczenie współczynników wielomianu $P_1(x) = \frac{P(x)}{x - z_0}$ lub $P_1(x) = \frac{P(x)}{x^2 + (-2p)x + (p^2 + q^2)}$ nazywa się **deflacją**. Wykonać ją można, korzystając ze wzorów (7.6) lub (7.7). Kolejnych pierwiastków wielomianu $P(x)$ poszukuje się, wyznaczając pierwiastki wielomianu $P_1(x)$.

Korzyści z przeprowadzenia deflacji są oczywiste. Znaleziony pierwiastek (pierwiastki) usunięto z wielomianu $P_1(x)$, nie ma więc niebezpieczeństwa ponownego, przypadkowego trafienia do niego. Wielomian $P_1(x)$ jest stopnia o 1 lub 2 mniejszego od wielomianu $P(x)$, więc poszukiwania kolejnych pierwiastków będą prostsze.

Niestety, deflacja jest także źródłem dodatkowych błędów. Jeżeli pierwiastek wielomianu $P(x)$ został wyznaczony metodą iteracyjną, do przeprowadzenia deflacji nie dysponujemy dokładną wartością pierwiastka, a jedynie jej przybliżeniem. Współczynniki wielomianu $P_1(x)$ otrzymanego po deflacji będą więc obciążone błędem zaokrągleń. Jak pokazano w rozdziale 6, uwarunkowanie pierwiastków wielomianu przy zmianach jego współczynników może być bardzo złe. Każde wykonanie deflacji wprowadza dodatkowy błąd zaokrągleń, który będzie kumulował się w błędach kolejno wyznaczanych pierwiastków.

Istnieją różne sposoby zmniejszania niekorzystnego wpływu deflacji na błąd wyznaczonych pierwiastków. Pierwiastki, które powinny być wyznaczone najdokładniej, powinny być obliczane jako pierwsze. Można na przykład starać się wyznaczać pierwiastki w kolejności rosnących modułów. Można też pierwiastki obliczone z użyciem deflacji traktować jak przybliżenia początkowe do kolejnego procesu iteracyjnego, przeprowadzonego już na całym wielomianie wyjściowym $P(x)$. Proces bywa nazywany „**wygładzaniem pierwiastków**” (ang. roots polishing). Istnieją też specjalne metody pozwalające uniknąć deflacji, a zabezpieczyć się przed ponownym znajdowaniem tego samego pierwiastka.

7.3. Metoda Newtona-Raphsona i jej warianty

Zgodnie z metodą **Newtona-Raphsona** poszukujemy pierwiastka wielomianu $P(x)$, iterując zgodnie z równaniem

$$x_{i+1} = x_i - \frac{P(x_i)}{P'(x_i)}. \quad (7.8)$$

Do obliczenia wartości wielomianu i pochodnej można użyć schematu (7.5). Jeżeli przybliżenie początkowe x_0 jest liczbą rzeczywistą, to iteracje (7.8), jeśli będą zbieżne, to do rzeczywistego pierwiastka. Trzeba następnie przeprowadzić deflację według wzoru (7.6) i kontynuować obliczenia dla wielomianu stopnia o 1 mniejszego.

Jeżeli chcemy wyznaczyć zespolony pierwiastek, to przybliżenie początkowe x_0 musi być liczbą zespoloną. Jeżeli iteracje (7.8) będą zbieżne do zespolonego pierwiastka, to trzeba przeprowadzić deflację zgodnie ze wzorem (7.7) i kontynuować obliczenia dla wielomianu stopnia o 2 mniejszego.

Dystans do poszukiwanego pierwiastka na płaszczyźnie zespolonej pozwala ocenić twierdzenie 7.1.

Twierdzenie 7.1 (o szacowaniu pierwiastków wielomianu – *Kincaid, Cheney, 2006*):

Jeżeli x_i i x_{i+1} są kolejnymi przybliżeniami pierwiastka wielomianu $P(x)$ stopnia n wyznaczonymi metodą Newtona-Raphsona, to istnieje pierwiastek tego wielomianu leżący na płaszczyźnie zespolonej w odległości od x_i nie przekraczającej $n|x_i - x_{i+1}|$.

Metoda Newtona-Raphsona jest zbieżna kwadratowo (dla pojedynczych pierwiastków) ale lokalnie. Przy poszukiwaniu zespolonych zer wielomianów ta lokalna zbieżność jest całkiem skomplikowanym zjawiskiem, jak pokazano w przykładzie 7.1.

Przykład 7.1

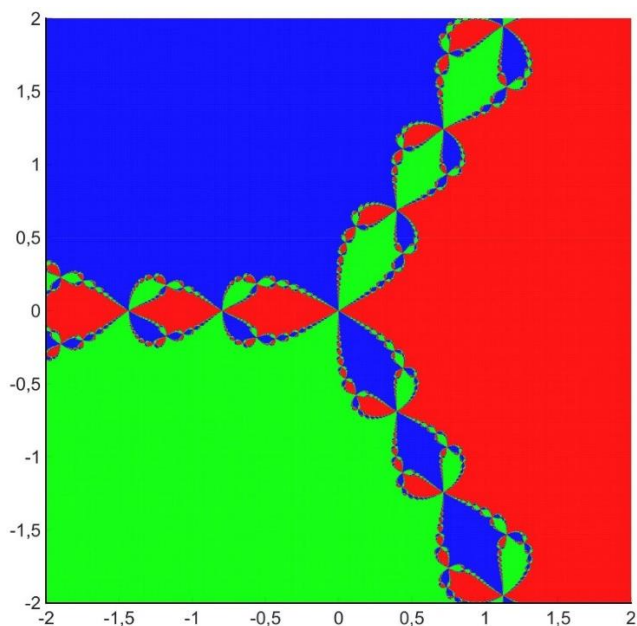
Rozważmy dwa równania wielomianowe

$$x^3 - 1 = 0, \quad (7.9)$$

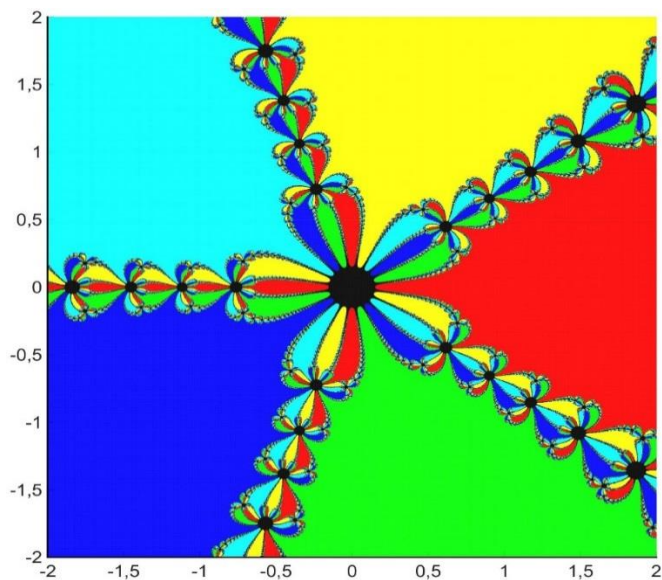
$$x^5 - 1 = 0. \quad (7.10)$$

Miejscami zerowymi równania (7.9) są $z_1 = 1$, $z_{2,3} = -0,5 \pm \frac{\sqrt{3}}{2}j$, a równania (7.10) $z_1 = 1$, $z_{2,3} \approx 0,3090 \pm 0,9511j$, $z_{4,5} \approx -0,8090 \pm 0,5878j$.

Zbiorem przyciągania danego pierwiastka z_* będziemy nazywali zbiór takich wartości początkowych x_0 , dla których metoda Newtona, określona zależnością (7.8), jest zbieżna do pierwiastka z_* . Na rysunku 7.1 i 7.2 pokazano na płaszczyźnie zespolonej zbiory przyciągania dla wszystkich pierwiastków równań (7.9) i (7.10).



Rys. 7.1. Zbiory przyciągania pierwiastków równania (7.9): czerwony – pierwiastek z_1 , niebieski – pierwiastek z_2 , zielony – pierwiastek z_3



Rys. 7.2. Zbiory przyciągania pierwiastków równania (7.10): czerwony – pierwiastek z_1 , żółty – pierwiastek z_2 , zielony – pierwiastek z_3 , jasno niebieski – pierwiastek z_4 , niebieski – pierwiastek z_5

Na rysunku 7.2 czarny obszar oznacza zbiór warunków początkowych, dla których metoda Newtona nie osiągnęła żadnego pierwiastka po założonej maksymalnej liczbie iteracji przy założonej dokładności. Zbiór punktów początkowych, dla których metoda Newtona-Raphsona nie jest zbieżna jest nazywany zbiorem Julii wielomianu $P(x)$ od nazwiska matematyka, który opisał go na początku XX wieku. Jeżeli wszystkie pierwiastki są pojedyncze, to ich zbiory przyciągania są otwarte, a zbiór Julii składa się z brzegów zbiorów przyciągania. Dla obu równań zbiory przyciągania tworzą złożoną strukturę nazywaną fraktalem (obiekt samopodobny). Na dowolnie powiększonym fragmencie fraktala zawierającym fragmenty co najmniej dwóch zbiorów przyciągania widzimy fraktal o tej samej strukturze. O skomplikowanej strukturze sąsiedztwa zbiorów przyciągania świadczy fakt, że każdy punkt brzegowy zbioru przyciągania z rysunku 7.1 jest jednocześnie punktem brzegowym trzech innych zbiorów przyciągania.

Metoda Maehly’ego jest modyfikacją metody Newtona, pozwalającą na uniknięcie niekorzystnych skutków deflacji.

Jeżeli w pierwszym etapie iteracje (7.8) doprowadziły do wyznaczenia pierwiastka z_1 , zamiast obliczyć poprzez deflację współczynniki wielomianu $P_1(x) = \frac{P(x)}{x-z_1}$ i poszukiwać kolejnego pierwiastka, iterując zgodnie z równaniem $x_{i+1} = x_i - \frac{P_1(x_i)}{P_1'(x_i)}$, należy skorzystać ze wzoru

$$P_1'(x) = \frac{P'(x)}{x - z_1} - \frac{P(x)}{(x - z_1)^2} \quad (7.11)$$

i kolejnego pierwiastka poszukiwać poprzez iteracje:

$$x_{i+1} = x_i - \frac{P_1(x_i)}{P_1'(x_i)} = x_i - \frac{P(x_i)}{P'(x_i) - \frac{P(x_i)}{x_i - z_1}}. \quad (7.12)$$

Wzór (7.12) jest bardziej złożony od wzoru (7.8), ale występują w nim tylko współczynniki wielomianu wyjściowego $P(x)$.

Jeżeli zostały znalezione pierwiastki z_1, \dots, z_k , to z uwagi na to, że

$$\begin{aligned} & \frac{d}{dx} \frac{P(x)}{(x - z_1) \dots (x - z_k)} \\ &= \frac{P'(x)}{(x - z_1) \dots (x - z_k)} - \frac{P(x)}{(x - z_1) \dots (x - z_k)} \sum_{j=1}^k \frac{1}{x - z_j} \end{aligned} \quad (7.13)$$

kolejny pierwiastek będzie wyznaczany w iteracjach:

$$\begin{aligned}
 x_{i+1} &= x_i - \frac{\frac{P(x_i)}{(x_i - z_1) \dots (x_i - z_k)}}{\frac{P'(x_i)}{(x_i - z_1) \dots (x_i - z_k)} - \frac{P(x_i)}{(x_i - z_1) \dots (x_i - z_k)} \sum_{j=1}^k \frac{1}{x_i - z_j}} \\
 &= x_i - \frac{P(x_i)}{P'(x_i) - P(x_i) \sum_{j=1}^k \frac{1}{x_i - z_j}}.
 \end{aligned} \tag{7.14}$$

Za podobną do metody Maehly'ego można uznać metodę **Abertha-Ehrlicha**, w której rozwiązuje się iteracyjnie układ równań

$$F_k(x) = \frac{P(x)}{(x-z_1)\dots(x-z_{k-1})(x-z_{k+1})\dots(x-z_n)} = 0, \tag{7.15}$$

$k = 1, \dots, n,$

przy czym rozwiązaniem k -tego równania jest pierwiastek z_k wielomianu $P(x)$. Zastosowanie iteracji newtonowskich do każdego z tych równań daje proces iteracyjny

$$x_k^{(i+1)} = x_k^{(i)} - \frac{F_k(x_k^{(i)})}{F'_k(x_k^{(i)})} = x_k^{(i)} - \frac{P(x_k^{(i)})}{P'(x_k^{(i)}) - P(x_k^{(i)}) \sum_{\substack{j=1 \\ j \neq k}}^n \frac{1}{x_k^{(i)} - z_j}}, \tag{7.16}$$

gdzie (i) oznacza numer iteracji, a ciąg $x_k^{(i)}$ zbiega do pierwiastka z_k . Jeżeli pierwiastek zostanie zastąpiony w (7.16) aktualnym przybliżeniem otrzymamy schemat iteracyjny, który można realizować dla wyznaczenia wszystkich jednocześnie:

$$x_k^{(i+1)} = x_k^{(i)} - \frac{F_k(x_k^{(i)})}{F'_k(x_k^{(i)})} = x_k^{(i)} - \frac{P(x_k^{(i)})}{P'(x_k^{(i)}) - P(x_k^{(i)}) \sum_{\substack{j=1 \\ j \neq k}}^n \frac{1}{x_k^{(i)} - x_j^{(i)}}}, \tag{7.17}$$

gdzie $k = 1, \dots, n$ oznacza numer pierwiastka, a $(i) = 0, 1, 2, 3, \dots$ numer iteracji.

Metoda Bairstowa poszukuje współczynników trójmianu kwadratowego $x^2 + rx + q$, który dzieli wielomian $P(x)$ bez reszty. Pierwiastki takiego trójmianu będą wtedy jednocześnie pierwiastkami wielomianu $P(x)$. Zgodnie ze wzorami (7.7) musimy rozwiązać układ dwóch równań z dwiema niewiarymymi r, q :

$$\begin{aligned}
 A(r, q) &= a_1 - rb_0 - qb_1 = 0, \\
 B(r, q) &= a_0 - qb_0 = 0.
 \end{aligned} \tag{7.18}$$

Ten układ równań można rozwiązać metodą Newtona-Raphsona. Zgodnie z wzorami (6.34), (6.35) należy przeprowadzić następujące iteracje:

$$\begin{aligned} \left[\begin{array}{cc} \frac{\partial A(r,q)}{\partial r} & \frac{\partial A(r,q)}{\partial q} \\ \frac{\partial B(r,q)}{\partial r} & \frac{\partial B(r,q)}{\partial q} \end{array} \right]_{\substack{r=r_i \\ q=q_i}} \begin{bmatrix} D_{i+1,1} \\ D_{i+1,2} \end{bmatrix} &= \begin{bmatrix} A(r_i, q_i) \\ B(r_i, q_i) \end{bmatrix}, \\ \begin{bmatrix} r_{i+1} \\ q_{i+1} \end{bmatrix} &= \begin{bmatrix} r_i \\ q_i \end{bmatrix} + \begin{bmatrix} D_{i+1,1} \\ D_{i+1,2} \end{bmatrix}. \end{aligned} \quad (7.19)$$

Wzory (7.7) można wykorzystać do obliczenia pochodnych cząstkowych i rozwiązać równania (7.19), odwracając macierz Jacobiego. W efekcie otrzymuje się kompletny schemat iteracji:

$$\begin{aligned} b_n &= b_{n-1} = 0, \quad b_k = a_{k+2} - r_i b_{k+1} - q_i b_{k+2}, \\ &\quad k = n-2, n-3, \dots, 0, \\ A(r_i, q_i) &= a_1 - r_i b_0 - q_i b_1, \quad B(r_i, q_i) = a_0 - q_i b_0, \\ d_{n-1} &= d_{n-2} = 0, \quad d_k = b_{k+1} - r_i d_{k+1} - q_i d_{k+2}, \\ &\quad k = n-3, n-4, \dots, 0, -1, \\ \begin{bmatrix} r_{i+1} \\ q_{i+1} \end{bmatrix} &= \begin{bmatrix} r_i \\ q_i \end{bmatrix} - \begin{bmatrix} d_{-1} & d_0 \\ -q_i d_0 & d_{-1} + r_i d_0 \end{bmatrix}^{-1} \begin{bmatrix} A(r_i, q_i) \\ B(r_i, q_i) \end{bmatrix}. \end{aligned} \quad (7.20)$$

Wszystkie metody wykorzystujące iteracje Newtona-Raphsona charakteryzują się kwadratową (dla pojedynczych pierwiastków), ale lokalną zbieżnością. Ta lokalna zbieżność może mieć bardzo złożony charakter, jak pokazano w przykładach 7.1 i 7.2, i mogą występować różne osobliwości poszczególnych metod. Na przykład metoda Bairstowa może być niestabilna dla wielomianów nieparzystego stopnia mających tylko jeden pierwiastek rzeczywisty.

Metoda Laguerre'a wykorzystuje schemat iteracyjny:

$$x_{i+1} = x_i - \frac{n}{A \pm \sqrt{(n-1)(nB - A^2)}}, \quad (7.21)$$

gdzie

$$A = \frac{P'(x_i)}{P(x_i)}, \quad B = \left(\frac{P'(x_i)}{P(x_i)} \right)^2 - \frac{P''(x_i)}{P(x_i)}.$$

Znak w mianowniku wyrażenia (7.21) wybieramy w taki sposób, żeby wartość bezwzględna mianownika była jak największa. Konieczna jest znajomość drugiej pochodnej, którą trzeba obliczać w każdej iteracji.

7.4. Inne podejścia do wyznaczania pierwiastków wielomianów

Metody iteracyjne przedstawione w poprzednim rozdziale wymagały w każdym kroku wyznaczenia pochodnej wielomianu. Istnieją metody iteracyjne wymagające informacji wyłącznie o wartości wielomianu. Niektóre z tych metod to:

- **Metoda Duranda-Knera** (pierwszy raz podana przez Weierstrassa i ponownie odkryta w XX wieku niezależnie przez Duranda i Knera) wykorzystuje przedstawienie wielomianu w postaci iloczynowej

$$P(x) = (x - z_1) \dots (x - z_n). \quad (7.22)$$

Pierwiastek z_k wielomianu $P(x)$ można obliczyć jako

$$z_k = x - \frac{P(x)}{(x-z_1)\dots(x-z_{k-1})(x-z_{k+1})\dots(x-z_n)}, \quad k = 1, \dots, n. \quad (7.23)$$

Układ równań (7.23) można rozwiązywać metodą iteracji prostej, według równań

$$z_k^{(i+1)} = z_k^{(i)} - W_k^{(i)}, \quad (7.24)$$

gdzie

$$W_k^{(i)} = \frac{P(z_k^{(i)})}{\prod_{\substack{j=1 \\ j \neq k}}^n (z_k^{(i)} - z_j^{(i)})} \quad (7.25)$$

i $k = 1, \dots, n$ oznacza numer pierwiastka, a $(i) = 0, 1, 2, 3, \dots$ numer iteracji.

Metoda Duranda-Knera wyznacza wszystkie pierwiastki jednocześnie, więc i jej rząd zbieżności trzeba rozumieć nieco inaczej niż w przypadku iteracji dążących do jednego pierwiastka. Metoda startująca z dostatecznie dokładnym, n -elementowym wektorem przybliżeń początkowych $z^{(0)}$ o różnych składowych generuje w i -tej iteracji n -elementowy wektor przybliżeń $z^{(i)}$. Kwadratowa zbieżność w sensie średniej (ang. quadratic-like mean convergence – QLMC), którą charakteryzuje się metoda, oznacza, że dla każdego z pierwiastków z_k wielomianu $P(x)$ (niekoniecznie pojedynczego) istnieje zbiór indeksów I , taki że średnia $\bar{z}_k^{(i+1)}$ z liczb $\{z_j^{(i+1)}, j \in I\}$ spełnia $\left| z_k - \bar{z}_k^{(i+1)} \right| = O\left(\max_{1 \leq k \leq n} |z_k - z_k^{(i)}|^2 \right)$.

- **Metoda Böerscha-Supana** wykorzystuje schemat iteracyjny:

$$z_k^{(i+1)} = z_k^{(i)} - \frac{W_k^{(i)}}{1 + \sum_{\substack{j=1 \\ j \neq k}}^n \frac{W_j^{(i)}}{(z_k^{(i)} - z_j^{(i)})}}, \quad (7.26)$$

gdzie $W_k^{(i)}$ jest określone zależnością (7.25).

Metoda ta cechuje się lokalną, sześcienną zbieżnością w podobnym sensie jak metoda Duranda-Kernera kwadratową.

- **Metoda Noureina** wykorzystuje schemat iteracyjny:

$$z_k^{(i+1)} = z_k^{(i)} - \frac{W_k^{(i)}}{1 + \sum_{\substack{j=1 \\ j \neq k}}^n \frac{W_j^{(i)}}{(z_k^{(i)} - W_k^{(i)} - z_j^{(i)})}}, \quad (7.27)$$

gdzie $W_k^{(i)}$ jest określone zależnością (7.25).

- **Metoda Müllera** jest uogólnieniem metody siecznych wykorzystującym interpolację kwadratową. Metoda potrzebuje trzech punktów startowych x_i, x_{i-1}, x_{i-2} . Schemat iteracyjny jest określony równaniem

$$x_{i+1} = x_i - (x_i - x_{i-1}) \frac{2C}{B \pm \sqrt{B^2 - 4AC}}, \quad (7.28)$$

gdzie

$$q = \frac{x_i - x_{i-1}}{x_{i-1} - x_{i-2}}, \quad A = qP(x_i) - q(1 + q)P(x_{i-1}) + q^2P(x_{i-2}),$$

$$B = (2q + 1)P(x_i) - (q + 1)^2P(x_{i-1}) + q^2P(x_{i-2}), \quad C = (1 + q)P(x_i).$$

Znak w mianowniku wyrażenia (7.28) wybieramy w taki sposób, żeby wartość bezwzględna mianownika była jak największa.

- **Metoda Lehmera-Schura** wykorzystuje **test Schura-Cohna**, który pozwala stwierdzić, czy w jednostkowym okręgu na płaszczyźnie zespolonej znajduje się pierwiastek wielomianu $P(x)$, czy nie. Test daje tylko odpowiedź „tak” lub „nie” (poza osobliwymi przypadkami, kiedy jest nierozstrzygnięty) i nie daje żadnej informacji o liczbie ani położeniu pierwiastków w okręgu. Test Schura-Cohna został przedstawiony w ramce R7.1.

Ramka R7.1 Test Schura-Cohna

Rozważmy wielomian w dziedzinie zespolonej

$$P^*(z) = a_n z^n + a_{n-1} z^{n-1} + \dots + a_2 z^2 + a_1 z + a_0. \quad (\text{R7.1.1})$$

Współczynniki $a_i, i = 1, \dots, n$ mogą być rzeczywiste lub zespolone.

Przez $P^*(z)$ oznaczymy wielomian

$$P^*(z) = \bar{a}_0 z^n + \bar{a}_1 z^{n-1} + \dots + \bar{a}_{n-2} z^2 + \bar{a}_{n-1} z + \bar{a}_n \quad (\text{R7.1.2})$$

gdzie $\bar{a}_i, i = 1, \dots, n$ oznacza liczbę sprzężoną do a_i .

Definiujemy nowy wielomian $P_1(z)$ jako

$$P_1(z) = T[P(z)] = \bar{a}_0 P(z) - a_n P^*(z) \quad (\text{R7.1.3})$$

i dalej rekurencyjnie

$$P_j(z) = T^j[P(z)] = T^j \left[T^{j-1}[P(z)] \right]. \quad (\text{R7.1.4})$$

Stopień wielomianu $P_1(z)$ jest mniejszy od stopnia wielomianu $P(z)$. Test prowadzimy według algorytmu:

A. Jeżeli $f(0) = 0$ to $z = 0$ jest pierwiastkiem. W przeciwnym wypadku przechodzimy do B.

B. Jeżeli $T[P(0)] = |a_0|^2 - |a_n|^2 < 0$ to pierwiastek znajduje się w kole jednostkowym. W przeciwnym wypadku przechodzimy do C.

C. Obliczamy $T^j[P(z)], j = 2, \dots, k$ aż do uzyskania $T^k[P(0)] < 0$, wtedy pierwiastek znajduje się w kole jednostkowym lub $T^k[P(0)] = 0$, wtedy żaden pierwiastek nie leży w kole jednostkowym, jeżeli $T^{k-1}[P(z)]$ jest stałą.

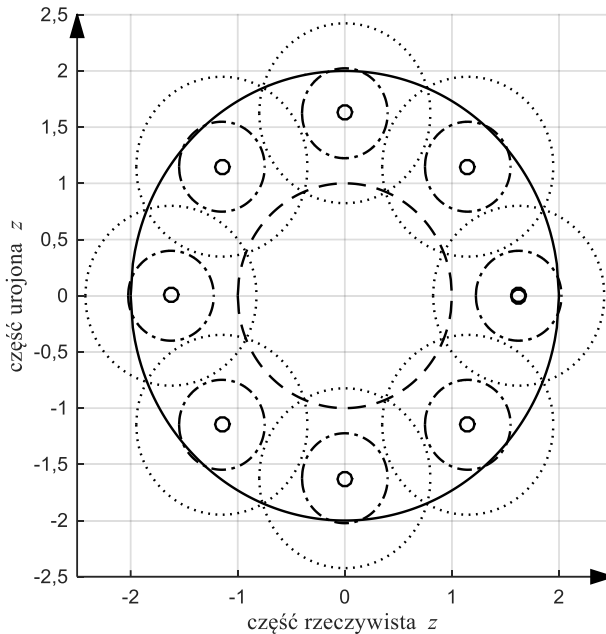
Test Schura-Cohna można zastosować do sprawdzenia, czy w dowolnym okręgu na płaszczyźnie zespolonej znajduje się pierwiastek wielomianu $P(z)$. Wystarczy skorzystać z faktu, że jeżeli wielomian $P(z)$ ma pierwiastek wewnątrz koła $|z - c| < r$ (c – środek koła, r – promień koła) to wielomian $Q(z) = P(rz + c)$ ma pierwiastek w kole jednostkowym.

Metoda Lehmera-Schura polega na pokrywaniu badanego obszaru okręgami i stosowaniu testu Schura-Cohna do kolejnych okręgów. Należy tak zaplanować konstruowanie kolejnych okręgów, żeby zminimalizować ich liczbę i nie pozostawić niezbadanych obszarów. Schemat konstrukcji okręgów pokazano na rysunku 7.3. Niech $R = 1$. Zaczynamy od sprawdzenia czy istnieje pierwiastek wewnątrz koła o promieniu $2R$ (linia ciągła, rys. 7.3). Jeśli stwierdzimy brak pierwiastka należy ponownie podwoić promień poszukiwań. Jeżeli stwierdzamy istnienie pierwiastka

w kole $2R$ to następnie sprawdzamy okrąg o promieniu R (linia kreskowa, rys. 7.3). Jeśli w kole R znajduje się pierwiastek, połowimy promień poszukiwań do momentu znalezienia okręgu, w którym nie ma pierwiastka. Jeżeli stwierdziliśmy brak pierwiastka w kole R oznacza to, że pierwiastek znajduje się pierścieniu ograniczonym okręgami o promieniach R i $2R$. Pierścień ten należy pokryć kołami. W algorytmie Lehmera-Shura tworzy się 8 takich kół o środkach

$$z_i = \frac{3R}{2 \cos\left(\frac{\pi}{8}\right)} e^{\frac{jk\pi}{4}}, \quad i = 0, 1, \dots, 7. \quad (7.29)$$

i o promieniu $\frac{4}{5}R$ (linia kreska-kropka, rys. 7.3). Stosujemy test do pierwszego z tych kół i sprawdzamy czy znajduje się w nim pierwiastek. Jeżeli NIE to przechodzimy do następnego koła pokrywającego pierścień. Jeśli TAK to połowimy promień i sprawdzamy koło o promieniu $\frac{2}{5}R$ (linia kropkowa, rys. 7.3). Od tego momentu postępujemy tak, jak w przypadku okręgów R i $2R$. Poszukiwania prowadzimy tak długo, aż promień okręgu poszukiwań będzie mniejszy niż założona dokładność. Przybliżeniem pierwiastka jest środek tego okręgu.



Rys. 7.3. Przykład pokrycia poszukiwanego obszaru okręgami w metodzie Lehmera-Schura: linia ciągła – okrąg o promieniu $2R$, linia kreskowa – okrąg o promieniu R , linia kreska-kropka – okrąg o promieniu $\frac{2}{5}R$, linia kropkowa – okrąg o promieniu $\frac{4}{5}R$, 'o' – środki okręgów według zależności (7.29)

Metoda jest wolno zbieżna, ale niezawodna. Istnieją też narzędzia (twierdzenie Rouché'a i twierdzenie Gerszgorina) pozwalające oszacować obszar, w którym leżą wszystkie pierwiastki wielomianu i który powinien być pokryty okręgami. Łatwy sposób oszacowania modułu największego pierwiastka wielomianu daje twierdzenie 7.2.

Twierdzenie 7.2 (o szacowaniu pierwiastków wielomianu – Kincaid, Cheney, 2006):

Wszystkie pierwiastki wielomianu $P(z) = a_n z^n + a_{n-1} z^{n-1} + \dots + a_2 z^2 + a_1 z + a_0$ leżą na płaszczyźnie zespolonej w kole o środku w 0, i promieniu

$$R = 1 + \frac{1}{|a_n|} \max_{1 \leq k < n} |a_k|.$$

7.5. Kombinowane algorytmy wyznaczania pierwiastków wielomianu

Wśród opisanych metod są wolne i zawsze zbieżne (np. metoda Lehmera-Schura) i szybsze, ale zbieżne lokalnie. Podobnie jak w przypadku rozwiązywania równań nieliniowych można łączyć metody o różnych właściwościach, wykorzystując zalety każdej z nich. Metoda zawsze zbieżna powinna w pierwszych iteracjach przygotować przybliżenie początkowe dla metody szybko zbieżnej. Jeżeli ta okaże się zbieżna, dokończymy obliczenia, jeśli nie powrócimy do metody zawsze zbieżnej, żeby poprawić przybliżenie początkowe dla metody szybko zbieżnej.

7.6. Uwarunkowanie pierwiastków wielomianów

Jak to już podkreślono w rozdziale 6 pierwiastki wielomianu mogą być bardzo wrażliwe na zmiany jego współczynników.

Znaczenie dla uwarunkowania pierwiastków ma też ich krotność, a właściwie nie krotność w ścisłym sensie matematycznym, ale dobre rozdzielenie w sensie numerycznym, to jest dostateczna odległość między sąsiednimi pierwiastkami w stosunku do ich modułów – przykład 6.7 z rozdziału 6.

Zawsze należy pamiętać o deflacji jako o kolejnym źródle błędów kumulujących się dla kolejno wyznaczanych pierwiastków.

8. Wartości i wektory własne

8.1. Definicje

Rzeczywista macierz kwadratowa $A \in R^{n \times n}$ reprezentuje przekształcenie przestrzeni liniowej C^n w C^n – każdemu wektorowi $x \in C^n$ przyporządkowuje wektor $y = A \cdot x \in C^n$. W przestrzeni argumentów tego przekształcenia istnieją takie szczególne wektory, które w jego wyniku nie zmieniają kierunku, a jedynie zostają pomnożone przez pewną liczbę – czyli zmieniają swoją normę lub inaczej: zostają przeskalowane.

Każdy, niezerowy wektor $x \in C^n$ spełniający wraz z liczbą $s \in C$ zależność:

$$A \cdot x = s \cdot x \quad (8.1)$$

nazywamy (prawym) **wektorem własnym** macierzy A , a „współczynnik skali” s – **wartością własną** związaną z wektorem własnym x .

Jeżeli wektor własny jest rzeczywisty, to i wartość własna musi być rzeczywista, jeśli wartość własna jest zespolona, to i wektor własny jest zespolony. Jeżeli x jest zespolonym wektorem własnym, to

$$A \cdot \bar{x} = \overline{s \cdot x} = \bar{s} \cdot \bar{x}, \quad (8.2)$$

więc także wektor sprzężony \bar{x} jest wektorem własnym macierzy A , a \bar{s} jest związaną z nim wartością własną. Jeżeli x jest wektorem własnym macierzy A , to dla dowolnej niezerowej liczby c

$$A \cdot (cx) = s \cdot (cx), \quad (8.3)$$

więc i wektor cx jest wektorem własnym macierzy A odpowiadającym tej samej wartości własnej.

Przykład 8.1

Jeśli przez macierz $A = \begin{bmatrix} 1 & 3 \\ 0 & 4 \end{bmatrix}$ przemnożymy wektory: $x_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $x_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $x_3 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ uzyskamy: $A \cdot x_1 = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$, $A \cdot x_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $A \cdot x_3 = \begin{bmatrix} 5 \\ 4 \end{bmatrix}$. Wektory x_1 oraz x_2 nie zmieniły kierunku – w ich przypadku mnożenie przez macierz A sprowadziło się do mnożenia przez skalar (równy odpowiednio 4 i 1). Są to więc wektory własne macierzy A . W przypadku x_3 zmianie uległa zarówno długość wektora jak i jego kierunek – nie jest wektorem własnym macierzy A .

Równanie (8.1) można zapisać w równoważnej postaci:

$$(sI - A)x = 0. \quad (8.4)$$

Ma ono niezerowe rozwiązanie wtedy i tylko wtedy, gdy:

$$\det(sI - A) = 0. \quad (8.5)$$

Wyznacznik $\det(sI - A)$ jest wielomianem zmiennej s o rzeczywistych współczynnikach. Stopień tego wielomianu jest równy wymiarowi macierzy A . Wielomian ten nazywamy **wielomianem charakterystycznym**, a równanie (8.5) – **równaniem charakterystycznym**. Wartości własne macierzy są zatem pierwiastkami równania charakterystycznego. Wynika stąd, że uwzględniając ich krotności jako pierwiastków, jest ich n , oraz że mogą być rzeczywiste lub zespolone parami sprzężone.

Zbiór wszystkich wartości własnych macierzy nazywamy **widmem macierzy**.

Przykład 8.2

Dla macierzy z przykładu 1:

$$\det\left(sI - \begin{bmatrix} 1 & 3 \\ 0 & 4 \end{bmatrix}\right) = \det\left(s \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 3 \\ 0 & 4 \end{bmatrix}\right) = \det\left(\begin{bmatrix} s-1 & -3 \\ 0 & s-4 \end{bmatrix}\right) = (s-1) \cdot (s-4) = 0.$$

Pierwiastkami równania są: 4 i 1.

Dwie macierze kwadratowe A i B nazywamy **podobnymi**, jeśli istnieje taka nieosobliwa ($\det(P) \neq 0$) macierz P , że:

$$B = P^{-1}AP. \quad (8.6)$$

Przekształcenie (8.6) macierzy A w B nazywa się **przekształceniem przez podobieństwo**.

Po pomnożeniu obu stron równania (8.1) przez P^{-1} otrzymujemy:

$$P^{-1}Ax = sP^{-1}x, \quad (8.7)$$

czyli:

$$BP^{-1}x = sP^{-1}x, \quad (8.8)$$

zatem s jest także wartością własną macierzy B , natomiast odpowiadającym jej wektorem własnym jest $P^{-1}x$. Przekształcenie przez podobieństwo nie zmienia widma macierzy.

Zbiór n równań (8.1) dla kolejnych wartości i i wektorów własnych można łącznie zapisać w postaci:

$$AX = X\Lambda, \quad (8.9)$$

gdzie:

- $X = [x_1, x_2, \dots, x_n]$ jest macierzą, której kolumnami są wektory własne,

$$\bullet \quad A = \begin{bmatrix} s_1 & 0 & 0 & \dots & 0 \\ 0 & s_2 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & s_n \end{bmatrix} = \text{diag}(s_1, s_2, \dots, s_n)$$

jest macierzą diagonalną, o kolejnych wartościach własnych na przekątnej.

Jeśli istnieje n liniowo niezależnych wektorów własnych x_1, \dots, x_n (warunkiem dostatecznym jest, by wszystkie wartości własne były różne), to macierz X jest nieosobliwa i można zapisać:

$$X^{-1}AX = A. \quad (8.10)$$

Stosując przekształcenie przez podobieństwo za pomocą X , uzyskujemy macierz diagonalną. O macierzy A mówimy wówczas, że jest **diagonalizowalna**.

Jeżeli macierz X spełniająca równanie (8.9) jest osobliwa (nie można wskazać n liniowo niezależnych wektorów własnych), to macierz A nie jest diagonalizowalna. Może tak być, jeśli wartości własne są wielokrotne.

Z przedstawionych zależności wynika jasno, że problemy wyznaczenia wartości i wektorów własnych są ze sobą ściśle związane.

Jeśli znana jest wartość własna s , to do wyznaczenia wektora własnego można skorzystać z równania (8.1) lub (8.4). W przypadku jednokrotnej wartości własnej, jest to układ, w którym można wyodrębnić $n - 1$ równań liniowo niezależnych, określających zależności między współrzędnymi wektora własnego. Układ ten ma nieskończenie wiele rozwiązań, co wynika z (8.3) i pozwala na wyznaczenie kierunku wektora własnego. Można dodać do niego równanie normalizujące długość wektora własnego albo ustalające jedną składową i w ten sposób jednoznacznie wyznaczyć wektor własny.

Przykład 8.3

Wiadomo, że dla macierzy A z przykładu 8.1 wartością własną jest 4. Można zatem napisać:

$\begin{bmatrix} 1 & 3 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} = 4 \begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix}$, gdzie x_{11} i x_{21} to pierwsza i druga współrzędna wektora własnego związanego z wartością własną $s_1 = 4$. Z powyższego równania macierzowego uzyskuje się układ dwóch równań:

$$x_{11} + 3x_{21} = 4x_{21},$$

$$4x_{21} = 4x_{21},$$

czyli ostatecznie zależność: $x_{11} = x_{21}$, którą musi spełniać wektor własny. Swobodę w wyborze składowych wektora własnego można wykorzystać do jego normalizacji (np. przez narzucenie długości = 1) albo do dowolnego wyboru jednej ze składowych. Najprostszym rozwiązaniem będzie $x_{11} = 1$ oraz $x_{21} = 1$.

Równanie (8.1) pozwala również na wyznaczenie wartości własnej na podstawie znanego wektora własnego. Po pomnożeniu obu stron równania przez transponowany i sprzężony wektor własny:

$$\bar{x}^T \cdot A \cdot x = \bar{x}^T \cdot s \cdot x = s \cdot \bar{x}^T \cdot x \quad (8.11)$$

uzyskujemy:

$$s = \frac{\bar{x}^T \cdot A \cdot x}{\bar{x}^T \cdot x} = \frac{\bar{x}^T \cdot A \cdot x}{\|x\|^2}. \quad (8.12)$$

Zamiast wektora \bar{x}^T można użyć w równaniu (8.11) dowolnego wektora y^T , jeśli tylko $y^T x \neq 0$. Wtedy

$$s = \frac{y^T \cdot A \cdot x}{y^T \cdot x}. \quad (8.13)$$

Wektor y można dobrać tak, by iloczyn $y^T \cdot x$ nie był mały, a jednocześnie by ograniczyć liczbę wykonywanych operacji arytmetycznych. Często wybiera się wektor y , w którym wszystkie elementy są zerami, z wyjątkiem jedynki znajdującej się w tym wierszu, w którym występuje element wektora x o największym module.

Jeżeli znane są wszystkie wektory własne, to można wyznaczyć wszystkie wartości własne z równania (8.10).

Istnieją macierze, których wartości własne są widoczne na pierwszy rzut oka: w macierzy trójkątnej, a w szczególności w macierzy diagonalnej są to liczby na głównej przekątnej. Niestety, jak wynika ze wzoru (8.10), przekształcenie przez podobieństwo do postaci diagonalnej wymaga znajomości wektorów własnych, a ich wyznaczenie z równań (8.1) wymaga znajomości wartości własnych. Nie istnieje też algorytm, który przekształcałby macierz przez podobieństwo do postaci trójkątnej w skończonej liczbie operacji. Analityczne wyznaczenie wartości własnych z równania charakterystycznego (8.5) jest możliwe tylko dla niskich wymiarów macierzy A (zgodnie z twierdzeniem Abela (dodatek D1) jest to możliwe dla $n \leq 4$). Tak więc, metody numerycznego wyznaczania wartości własnych muszą być metodami iteracyjnymi.

8.2. Uwarunkowanie wartości własnych

Rozważmy uwarunkowanie wektora wartości własnych, czyli jego wrażliwość na zmianę (błąd) w danych wejściowych, czyli w elementach macierzy A . Podobnie jak w rozdziale 2 będzie tu używana norma wektorowa i zgodna z nią (indukowana) norma macierzowa. Niech ΔA będzie taką zmianą diagonalizowalnej macierzy A . Użycie macierzy $A + \Delta A$ zamiast A w równaniu (8.10) spowoduje,

że zamiast macierzy A z wartościami własnymi na przekątnej otrzymamy macierz $A + \Delta A$:

$$X^{-1}(A + \Delta A)X = A + \Delta A. \quad (8.14)$$

Macierz ΔA nie musi być diagonalna, ale jej norma jest dobrą miarą błędu, jakim będą obciążone wartości własne na skutek zaburzenia ΔA . Z wagi na to, że

$$X^{-1}AX + X^{-1}\Delta AX = A + \Delta A \Rightarrow \Delta A = X^{-1}\Delta AX, \quad (8.15)$$

dostajemy

$$\|\Delta A\| \leq \|X^{-1}\| \cdot \|\Delta A\| \cdot \|X\| = \text{cond}(X)\|\Delta A\|, \quad (8.16)$$

(wskaźnik uwarunkowania macierzy $\text{cond}(X)$ został omówiony w rozdziale 2). Uwarunkowanie wartości własnych zależy więc od wskaźnika uwarunkowania macierzy zbudowanej z wektorów własnych.

Jeżeli macierz wektorów własnych jest ortogonalna, czyli $X^{-1} = X^T$, to $\text{cond}(X) = 1$ i zadanie wyznaczenia wartości własnych jest dobrze uwarunkowane. Tak jest w przypadku, gdy A jest macierzą symetryczną. Jeżeli natomiast wektory własne będą zbliżać się do wspólnego kierunku, to wskaźnik uwarunkowania macierzy X rośnie. Tak jest w przypadku wartości własnych bliskich sobie, słabo rozdzielonych. Wyznaczenie wartości własnych macierzy niediagonalizowalnej jest bardzo źle uwarunkowane.

Można także wyprowadzić zależności opisujące uwarunkowanie pojedynczej wartości własnej. Jeżeli w zależności (8.13) zastosujemy $A + \Delta A$ zamiast A , to dostaniemy

$$s + \Delta s = \frac{y^T \cdot (A + \Delta A) \cdot x}{y^T \cdot x} = \frac{y^T \cdot A \cdot x}{y^T \cdot x} + \frac{y^T \cdot \Delta A \cdot x}{y^T \cdot x}, \quad (8.17)$$

a stąd i z nierówności Cauchy'ego-Schwarza, w przypadku rzeczywistych wektorów y, x , dostaniemy

$$|\Delta s| = \left| \frac{y^T \cdot \Delta A \cdot x}{y^T \cdot x} \right| \leq \frac{1}{\cos \angle(y, x)} \|\Delta A\|. \quad (8.18)$$

Dla wektora $y = x$ mamy $\frac{1}{\cos \angle(y, x)} = 1$ i dobre uwarunkowanie wartości własnej, wybór wektora y bliskiego prostopadłemu do x prowadzi do złego uwarunkowania.

Problem uwarunkowania wektorów własnych jest bardziej złożony. Z równania (8.4) wyznacza się tylko kierunek wektora własnego i to uwarunkowanie tego kierunku jest istotne. Uwarunkowanie problemu wyznaczania wektora własnego będzie złe, jeżeli dwie wartości własne będą źle rozdzielone (położone blisko siebie) lub jeśli kąt tego wektora z innym wektorem własnym będzie mały.

8.3. Wyznaczanie wartości własnych z wielomianu charakterystycznego

Ponieważ wartości własne to pierwiastki wielomianu charakterystycznego, to można najpierw wyznaczyć współczynniki tego wielomianu, a potem znaleźć jego pierwiastki jedną z metod opisanych w rozdziale 7.

Niech wielomianem charakterystycznym macierzy A będzie

$$\det(sI - A) = s^n + b_{n-1}s^{n-1} + \dots + b_1s + b_0. \quad (8.19)$$

Na mocy twierdzenia Cayley'a-Hamiltona (dodatek D2) zachodzi

$$A^n + b_{n-1}A^{n-1} + \dots + b_1A + b_0I = 0. \quad (8.20)$$

Po prawostronnym pomnożeniu obu stron równania (8.20) przez wektor y dostajemy

$$[A^{n-1}y : A^{n-2}y : \dots : Ay : y] \begin{bmatrix} b_{n-1} \\ \vdots \\ b_1 \\ b_0 \end{bmatrix} = -A^n y. \quad (8.21)$$

Wyznaczenie współczynników wielomianu charakterystycznego tą metodą, która nosi nazwę **metody Kryłowa**, polega więc na rozwiązaniu układu równań liniowych (8.21). Wektor y musi być tak dobrany, by wskaźnik uwarunkowania macierzy $[A^{n-1}y : A^{n-2}y : \dots : Ay : y]$ był niewielki.

W dalszej kolejności należy wyznaczyć pierwiastki wielomianu charakterystycznego, korzystając z metod opisanych w rozdziale 7. Zadanie to jest zwykle znacznie gorzej uwarunkowane niż zadanie wyznaczenia wartości własnych macierzy. Błędy wyznaczania wartości własnych z wielomianu nie wynikają z działania metod wyszukujących pierwiastki, lecz z niedokładności wyznaczenia współczynników wielomianu. Dlatego też, to nie zadanie wyznaczania wartości własnych sprowadza się do zadania obliczania pierwiastków wielomianu, a odwrotnie – pierwiastki wielomianu można wyznaczać, obliczając wartości własne odpowiedniej macierzy.

Dla danego wielomianu:

$$p(x) = x^n + a_{n-1}x^{n-1} + \dots + a_2x^2 + a_1x + a_0 \quad (8.22)$$

można wyznaczyć tzw. **macierz stowarzyszoną**:

$$C = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & -a_0 \\ 1 & 0 & 0 & \dots & 0 & -a_1 \\ 0 & 1 & 0 & \dots & 0 & -a_2 \\ 0 & 0 & 1 & \dots & 0 & -a_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -a_{n-1} \end{bmatrix}, \quad (8.23)$$

$$C^T = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -a_0 & -a_1 & -a_2 & \dots & -a_{n-1} \end{bmatrix},$$

dla której (8.22) jest wielomianem charakterystycznym. Można zatem znaleźć pierwiastki wielomianu, wyznaczając wartości własne macierzy C .

Taka metoda szukania pierwiastków wielomianu zastosowana jest m.in. w programie Matlab, gdzie polecenie `roots` tworzy macierz stowarzyszoną i dla niej wywołuje polecenie `eig` szukające wartości własnych.

Przykład 8.4

Wyznaczyć wartości własne macierzy: $A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 2 \\ 2 & 0 & 3 \end{bmatrix}$, korzystając z metody Kryłowa.

Przyjmujemy dowolny wektor y , np. $y = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ i obliczamy:

$b_2A^2y + b_1Ay + b_0y = -A^3y$, uzyskując:

$$\begin{bmatrix} 4 \\ 30 \\ 37 \end{bmatrix} b_2 + \begin{bmatrix} 2 \\ 8 \\ 11 \end{bmatrix} b_1 + \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} b_0 = - \begin{bmatrix} 8 \\ 104 \\ 119 \end{bmatrix} \text{ lub } \begin{bmatrix} 4 & 2 & 1 \\ 30 & 8 & 2 \\ 37 & 11 & 3 \end{bmatrix} \begin{bmatrix} b_2 \\ b_1 \\ b_0 \end{bmatrix} = - \begin{bmatrix} 8 \\ 104 \\ 119 \end{bmatrix}.$$

Stosując np. eliminację Gaussa, otrzymujemy wartości: $b_2 = -6$, $b_1 = 11$, $b_0 = -6$, zatem wielomian charakterystyczny to: $w(x) = x^3 - 6x^2 + 11x - 6$.

Dalej należy wyznaczyć pierwiastki wielomianu np. korzystając ze wzorów na rozwiązanie równania 3 rzędu czy korzystając z metod opisanych w rozdziale 7.

Przykład 8.5

Nawet w przypadku niewielkich macierzy posłużenie się wielomianem charakterystycznym daje większy błąd wyznaczania wartości własnych. Weźmy macierz $A = \begin{bmatrix} 1 + 10^{-7} & 0 \\ 0 & 1 \end{bmatrix}$. Jej wartościami własnymi są 1 i $1 + 10^{-7}$ i każda metoda bezpośrednio wyznaczająca wartości własne wyznaczy je z błędem nieprzekraczającym błędu zaokrąglenia do najbliższej liczby zmiennoprzecinkowej – eps_M (w Matlabie liczby 1 i $1 + 10^{-7}$ oraz wynik polecenia `eig(A)` są zaokrąglane do tych samych liczb zmiennopozycyjnych). Tymczasem obliczenie wielomianu charakterystycznego (w Matlabie polecenie `charpoly(A)`) daje współczynniki $[1; -2 - 10^{-7}; 1 + 10^{-7}]$, co przez polecenie `roots` prowadzi do wyniku obciążonego błędem $0,3 \cdot 10^{-8}$, a więc rzędu \sqrt{eps} .

Zaburzenie macierzy A przez dodanie $\begin{bmatrix} 0 & 10^{-8} \\ 10^{-8} & 0 \end{bmatrix}$ zmienia wynik procedury `eig(A)` o składnik rzędu 10^{-8} , podczas gdy zmiana wyrazu wolnego wielomianu charakterystycznego o 10^{-8} powoduje, że obliczone pierwiastki stają się zespolone z częścią urojoną około $10^{-4} = \sqrt{10^{-8}}$.

8.4. Metoda potęgowa**Przykład 8.6**

Macierz $A = \begin{bmatrix} 1 & 4 \\ 0 & 4 \end{bmatrix}$ ma dwie wartości własne $s_1 = 4$, $s_2 = 1$. Z wartością własną 4 jest związany wektor własny $x_1 = [1; 0,75]^T$. Wybierzmy (przypadkowy) wektor $v_0 = [1; 1]^T$ i wielokrotnie pomnożmy go przez macierz A , jednocześnie normalizując wektor przez podzielenie przez największy element. Otrzymujemy ciąg:

$$v_0 = [1; 1]^T; \frac{Av_0}{\|Av_0\|_\infty} = \frac{[5; 4]^T}{5} = [1; 0,8]^T,$$

$$\frac{A^2v_0}{\|A^2v_0\|_\infty} = \frac{[21; 16]^T}{21} = [1 \ 0,7619]^T, \dots$$

$$\frac{A^5v_0}{\|A^5v_0\|_\infty} = \frac{[1365; 1024]^T}{1365} = [1 \ 0,7502]^T, \dots$$

Uzyskiwany ciąg wektorów coraz bardziej zbliża się do kierunku wektora $x_1 = [1 \ 0,75]^T$ – czyli wektora własnego związanego z wartością własną o największym module.

Z właściwości zilustrowanej przez przykład 8.6 korzysta iteracyjna metoda wyznaczania dominującej wartości własnej nazywana metodą potęgową.

Założmy, że macierz A ma dominującą (to znaczy o największym module) wartość własną s_1 i jest diagonalizowalna, to znaczy, że ma n niezależnych liniowo wektorów własnych $x_i, i = 1, \dots, n$. Jeżeli A jest macierzą rzeczywistą, to i dominująca wartość własna musi być pojedyncza i rzeczywista (bo dwie liczby zespolone sprzężone mają takie same moduły).

Weźmy dowolny wektor $v_0 \in R^n$, który nie jest ortogonalny do wektora własnego x_1 związanego z dominującą wartością własną s_1 (wybierając wektor v_0 losowo z prawdopodobieństwem jeden spełnimy ten warunek). Wektor v_0 można przedstawić w postaci liniowej kombinacji wektorów własnych $x_i, i = 1, \dots, n$:

$$v_0 = \sum_{i=1}^n a_i x_i. \quad (8.24)$$

Z równości $A \cdot x_i = s_i \cdot x_i$ wynika, że wektor $v_1 = Av_0$ będzie równy:

$$v_1 = Av_0 = \sum_{i=1}^n a_i A x_i = \sum_{i=1}^n a_i s_i x_i, \quad (8.25)$$

wektor $v_2 = Av_1 = A^2 v_0$:

$$v_2 = Av_1 = \sum_{i=1}^n a_i s_i A x_i = \sum_{i=1}^n a_i s_i^2 x_i, \quad (8.26)$$

a wektor $v_m = Av_{m-1} = A^m v_0$:

$$v_m = Av_{m-1} = \sum_{i=1}^n a_i s_i^m x_i. \quad (8.27)$$

Jeżeli wyciągniemy przed nawias s_1^m , to otrzymamy

$$v_m = s_1^m \left(a_1 x_1 + \sum_{i=2}^n a_i \frac{s_i^m}{s_1^m} x_i \right). \quad (8.28)$$

Każda z liczb $\frac{s_i}{s_1}, i = 2, \dots, n$ (być może zespolonych) ma moduł mniejszy od 1 (bo s_1 jest dominującą wartością własną), czyli każda z liczb $\left(\frac{s_i}{s_1}\right)^m, i = 2, \dots, n$ dąży do zera dla rosnącego m . Tak więc, składnik $\sum_{i=2}^n a_i \frac{s_i^m}{s_1^m} x_i$ w (8.28) zanika,

a wektor v_m odtwarza coraz dokładniej kierunek wektora własnego x_1 . Jeżeli w formule (8.13) zastąpimy wektor własny przez przybliżający go wektor v_m , to otrzymamy

$$s_1 = \lim_{m \rightarrow \infty} \frac{y^T A A^m v_0}{y^T A^m v_0} = \lim_{m \rightarrow \infty} \frac{y^T v_{m+1}}{y^T v_m} \approx \frac{y^T v_{m+1}}{y^T v_m}. \quad (8.29)$$

Wektor y^T może składać się z zer poza jedynką w miejscu odpowiadającym elementowi v_m o największym module.

Szybkość zbieżności metody potęgowej zależy od tego, jak szybko zbiegają do zera liczby $\left(\frac{s_i}{s_1}\right)^m$. Jeżeli moduł wartości własnej s_1 jest wielokrotnie większy od modułów pozostałych wartości własnych, zbieżność jest szybka, jeśli niewiele większy – wolniejsza.

By ustrzec się operowania coraz większymi (lub mniejszymi) liczbami, warto normalizować wektory v_m , stosując w każdym kroku dodatkowe podstawienie

$$v_m := \frac{v_m}{\|v_m\|_\infty}, \quad (8.30)$$

tak jak to było w przykładzie 8.6. Nie zmienia to niczego w odtwarzaniu wektora własnego, bo przecież liczy się tylko jego kierunek.

Jeśli już zostanie wyznaczona dominująca wartość własna s_1 i odpowiadający jej wektor własny x_1 , i planujemy wyznaczenie kolejnej wartości własnej, to konieczne jest wyeliminowanie znalezionej wartości własnej i zastąpienie jej taką, która na pewno nie stanie się dominująca. Jednym ze sposobów takiej „redukcji” jest wprowadzenie macierzy

$$A_1 = A - s_1 x_1 z^T, \quad z^T x_1 = 1, \quad (8.31)$$

która ma wartości własne macierzy A , poza wartością s_1 , w miejscu której jest zero. Jeżeli macierz A_1 spełnia założenia metody (ma dominującą, pojedynczą wartość własną), to można rozpocząć poszukiwanie kolejnej wartości własnej.

Zastosowanie metody potęgowej jest dość skutecznie ograniczone założeniem o pojedynczej, dominującej wartości własnej.

Metody potęgowej można używać do poprawiania dokładności przybliżenia wartości własnej otrzymanego inną metodą. Ten algorytm nosi też nazwę **odwrotnej metody potęgowej**. Przypuśćmy, że dysponujemy „dobrym” przybliżeniem σ wartości własnej s_i , to jest takim że:

$$\forall j \neq i \quad |\sigma - s_i| < |\sigma - s_j|. \quad (8.32)$$

Wtedy $(\sigma - s_i)^{-1}$ jest dominującą wartością własną macierzy $(\sigma I - A)^{-1}$, stosujemy więc metodę potęgową do niej:

$$v_m = (\sigma I - A)^{-1} v_{m-1} \Rightarrow (\sigma I - A) v_m = v_{m-1}. \quad (8.33)$$

W każdej iteracji musimy rozwiązać układ równań liniowych (8.33), ale macierz współczynników jest zawsze ta sama i równa $\sigma I - A$, wystarczy więc tylko raz przeprowadzić jej rozkład trójkątny.

8.5. Metoda QR wyznaczania wartości własnych

Istnieje cała grupa metod iteracyjnych wykorzystujących przekształcenie przez podobieństwo do wyznaczenia numerycznych przybliżeń wartości własnych. Zgodnie z (8.6) iteracje w takiej metodzie przebiegają według schematu

$$A_0 = A, \quad A_{i+1} = P_i^{-1} A_i P_i, \quad \det P_i \neq 0. \quad (8.34)$$

Ich celem jest sprowadzenie macierzy A_i do postaci, w której wartości własne są widoczne na głównej przekątnej, na przykład do postaci trójkątnej. Jeżeli macierz przekształcenia przez podobieństwo jest ortogonalna, czyli $P_i = Q_i$, $Q_i^{-1} = Q_i^T$, to nie tylko unikamy kłopotliwego wyznaczania macierzy odwrotnej, ale zachowujemy w kolejnych iteracjach uwarunkowanie wartości własnych macierzy. Wiele z metod wykorzystujących iteracyjne przekształcenie przez podobieństwo jest dedykowanych macierzom symetrycznym, dla których problem wyznaczania wartości własnych jest łatwiejszy i zawsze dobrze uwarunkowany. Uniwersalnym sposobem jest tak zwana **metoda QR**.

Dla macierzy rzeczywistej A istnieje ortogonalna macierz Q i trójkątna górna R , takie że $A = QR$.

Algorytm wyznaczania tych macierzy, lub inaczej rozkładu QR albo faktoryzacji QR, wymaga skończonej liczby operacji, podobnie jak eliminacja Gaussa. Algorytm wyznaczania rozkładu QR pokazano w ramce R8.1.

Ramka R8.1 Algorytm wyznaczania rozkładu QR:

Dla danej macierzy A ($n \times n$) uzyskuje się rozkład na dwie macierze: ortogonalną macierz Q oraz trójkątną górną R :

$$A = QR.$$

Kolejne kolumny macierzy A są oznaczone przez A_j i prowadzi się obliczenia rekurencyjne wyznaczając elementy r_{jj} na diagonalu macierzy R i kolumny q_j macierzy Q :

$$y_1 = A_1,$$

$$r_{1,1} = \|y_1\|_2,$$

$$q_1 = \frac{y_1}{r_{11}},$$

$$y_2 = A_2 - q_1 q_1^T A_2,$$

$$r_{2,2} = \|y_2\|_2,$$

$$q_2 = \frac{y_2}{r_{22}}$$

i dalej dla $j = 3, \dots, n$:

$$y_j = A_j - q_1 q_1^T A_j - q_2 q_2^T A_j - \dots - q_{j-1} q_{j-1}^T A_j,$$

$$r_{j,j} = \|y_j\|_2,$$

$$q_j = \frac{y_j}{r_{j,j}}.$$

następnie oblicza się pozostałe elementy macierzy R , dla $i = 3, \dots, n$, $j = i, i + 1, \dots, n$

$$r_{ij} = q_i^T A_j.$$

Iteracje przy wyznaczaniu wartości własnych metodą QR przebiegają w następujący sposób:

Krok 1: $A_0 = A$.

Krok 2: wyznaczamy rozkład QR macierzy A_i :

$$A_i = Q_i R_i. \quad (8.35)$$

Krok 3: wyznaczamy macierz A_{i+1} dla kolejnej iteracji:

$$A_{i+1} = R_i Q_i. \quad (8.36)$$

Jeśli tak postąpimy, to $A_{i+1} = R_i Q_i = Q_i^T Q_i R_i Q_i = Q_i^T A_i Q_i$, a więc w każdej iteracji dokonujemy przekształcenia przez podobieństwo z ortogonalną macierzą przekształcenia Q_i .

Z jednej strony mamy

$$A_{i+1} = Q_i^T \dots Q_2^T Q_1^T A Q_1 Q_2 \dots Q_i, \quad (8.37)$$

z drugiej

$$R_{i+1} = Q_{i+1}^T A_{i+1} = Q_{i+1}^T Q_i^T \dots Q_2^T Q_1^T A Q_1 Q_2 \dots Q_i, \quad (8.38)$$

jest macierzą trójkątną górną, której wartości własne są ulokowane na przekątnej. Z porównania wzorów (8.37) i (8.38) widać, że jeżeli iteracje metody QR są zbieżne, to musi być

$$\lim_{i \rightarrow \infty} A_i = \lim_{i \rightarrow \infty} R_i, \quad (8.39)$$

więc macierz A_i dąży do macierzy trójkątnej górnej z wartościami własnymi macierzy A na przekątnej. Pełne wyjaśnienie zbieżności metody QR i jej właściwości jest dość skomplikowane i na potrzeby tego skryptu trzeba zadowolić się tym, dość heurystycznym, argumentem.

W implementacjach metody QR stosuje się dodatkowe elementy, które zmniejszają nakład obliczeń lub przyspieszają zbieżność.

Macierz A można wstępnie przekształcić do postaci prawie trójkątnej górnej, w której elementy niezerowe występują na głównej przekątnej, nad nią i bezpośrednio pod nią. Postać taka jest nazywana też postacią Hessenberga, a algorytm, który do niej prowadzi ma skończoną liczbę operacji i oczywiście zachowuje wartości własne. Pokazano go w ramce R8.2.

Ramka R8.2 Algorytm sprowadzania do postaci Hessenberga

Dla rzeczywistej macierzy kwadratowej A ($n \times n$) przyjmujemy, że: $A_0 = A$.
Następnie, rekurencyjnie dla $i = 1, \dots, n - 2$:

- wyznaczamy wektory:
 x_i – utworzony przez $n - i$ dolnych elementów z i -tej kolumny macierzy

$$w_i = [\|x\|_2, 0, \dots, 0]^T - \text{wektor } (n - i) \times 1,$$

$$v_i = w_i - x_i,$$

- wyznaczamy tzw. odbicie Householdera U_i przekształcający x_i w w_i (czyli $U_i x = w$):

$$U_i = I - \frac{2v_i v_i^T}{v_i^T v_i},$$

- konstruujemy macierz H_i :

$$H_i = \begin{bmatrix} I_{i \times i} & 0 \\ 0 & U_i \end{bmatrix},$$

- obliczamy:

$$A_i = H_i A_{i-1} H_i^T.$$

Macierz A przekształconą do postaci Hessenberga uzyskujemy, stosując przekształcenie przez podobieństwo:

$$H = P A P^T,$$

gdzie: $P = H_{n-2} \cdot \dots \cdot H_1$ oraz $P^T = H_1^T \cdot \dots \cdot H_{n-2}^T$.

Metoda QR zachowuje postać Hessenberga w każdej iteracji, nie ma więc potrzeby obliczania elementów, które będą zerami.

Przyspieszenie zbieżności można uzyskać, modyfikując iteracje do postaci:

Krok 1: $A_0 = A$.

Krok 2: wyznaczamy rozkład QR macierzy $A_i - \mu_i I$:

$$A_i - \mu_i I = Q_i R_i. \quad (8.40)$$

Krok 3: $A_{i+1} = R_i Q_i + \mu_i I$.

Istnieje kilka sposobów doboru przesunięcia μ_i , według ogólnej zasady, że μ_i powinno być jak najbliższej wyznaczonej najmniejszej wartości własnej. Na przykład μ_i jest elementem z prawego, dolnego narożnika macierzy A_i (przesunięcie Rayleigha).

Metoda QR, choć powstała w latach 60. XX wieku, jest nadal najpowszechniej i najskuteczniej stosowaną metodą wyznaczania wartości własnych macierzy niesymetrycznych. Jest zaimplementowana w procedurze `eig` Matlaba i w wielu innych pakietach obliczeniowych.

Przykład 8.7

Obliczamy wartości własne macierzy A , której wielomianem charakterystycznym jest wielomian $P(x) = (x - 1)(x - 2) \dots (x - 14)(x - 15)$ rozważany już w przykładzie 6.6, gdzie stwierdziliśmy złe uwarunkowanie jego pierwiastków. Macierz taką utworzono, obliczając $A = P^T S P$, gdzie P jest macierzą symetryczną i ortogonalną, której elementami są $P(i, j) = \frac{2}{\sqrt{2n+1}} \sin \frac{2ij\pi}{2n+1}$, $n = 15$ (polecenie `gallery('orthog', n, 2)` Matlaba), a S macierzą diagonalną z liczbami $1, 2, \dots, 15$ na przekątnej. Polecenie `eig(A)`, czyli metoda QR oblicza wszystkie wartości własne z błędem nie przekraczającym 10^{-14} , wynikającym przede wszystkim z błędów zaokrągleń przy mnożeniu $A = P^T S P$. Przejście przez wielomian charakterystyczny, czyli polecenie `roots(charpoly(A))` powoduje wzrost błędów obliczonych wartości własnych do około 10^{-5} . Zaburzenie elementu $A(15,1)$ składnikiem 10^{-8} powoduje zmianę wyników `eig(A)` rzędu 10^{-9} , podczas gdy zmiana wyrazu przy 14 potędze w obliczonym wielomianie charakterystycznym o 10^{-8} daje zmianę wyniku procedury `roots`, (obliczanie pierwiastków wielomianu przez wyznaczenie wartości własnych macierzy stowarzyszonej) rzędu 10^{-2} .

8.6. Wartości szczególne macierzy

Dla każdej macierzy prostokątnej A o m wierszach i n kolumnach, rzędu $r \leq \min(m, n)$, istnieją macierze ortogonalne⁷ $U \in R^{m \times m}$ i $V \in R^{n \times n}$ oraz (pseudo)diagonalna macierz $\Sigma \in R^{m \times n}$, takie że

$$A = U \Sigma V^T, \quad \Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}, \quad D = \text{diag}\{\sigma_1, \sigma_2, \dots, \sigma_r\}, \quad (8.41)$$

przy czym liczby $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ są określone jednoznacznie i nazywane są wartościami szczególnymi macierzy A . Przedstawienie macierzy w postaci (8.41) jest nazywane **rozkładem (faktoryzacją) szczególnym** (według wartości szczególnych, rozkładem SVD) macierzy A .

Kolumny v_i , $i = 1, \dots, n$ macierzy V są nazywane **prawymi wektorami szczególnymi**, a kolumny u_i , $i = 1, \dots, m$ macierzy U – **lewymi wektorami szczególnymi**.

⁷ Unitarne jeśli A jest macierzą zespoloną.

Wymiary zerowych bloków w macierzy Σ zależą od relacji między liczbami r, m, n . Jeśli A jest kwadratową macierzą pełnego rzędu ($r = n = m$), to wszystkie te bloki znikają.

Rozkład szczególny można też zapisać w postaci:

$$A = U\Sigma V^T = \sum_{i=1}^r \sigma_i u_i v_i^T. \quad (8.42)$$

Wzór (8.42) przedstawia macierz A w postaci sumy r macierzy pierwszego rzędu. Każdą z nich można zapisać w postaci dwóch wektorów i jednej liczby. Jeżeli niektóre wartości szczególne są znacznie mniejsze od pozostałych, to można pominąć odpowiednie składniki sumy (8.42) i zredukować informację konieczną do reprezentacji danych zawartych w A , godząc się na ich przybliżoną reprezentację.

Składniki sumy (8.42) noszą nazwę **składowych głównych**. Jeżeli wybierzemy q z nich – odpowiadających największym wartościom szczególnym, to macierz

$$A_q = \sum_{i=1}^q \sigma_i u_i v_i^T \quad (8.43)$$

jest najlepszą aproksymacją rzędu q macierzy A według normy euklidesowej, a norma błędu tej aproksymacji jest równa kolejnej wartości szczególnej: $\|A - A_q\| = \sigma_{q+1}$. Te fakty są podstawą metody analizy danych nazywanej **analizą składowych głównych** (ang. principal component analysis – PCA).

Z uwagi na to, że

$$A^T A = V\Sigma U^T U\Sigma V^T = V\Sigma^2 V^T \Rightarrow V^T (A^T A) V = \Sigma^2, \quad (8.44)$$

to kwadraty wartości szczególnych są wartościami własnymi macierzy $A^T A$, a prawe wektory szczególne są wektorami własnymi macierzy $A^T A$.

Jeżeli macierz kwadratowa A jest nieosobliwa ($r = n$), to wszystkie jej wartości szczególne A są dodatnie. Dla wyznacznika nieosobliwej macierzy kwadratowej zachodzi:

$$|\det(A)| = \sigma_1 \sigma_2 \dots \sigma_n. \quad (8.45)$$

Rozkład szczególny pozwala na łatwe odwrócenie nieosobliwej macierzy kwadratowej:

$$A = U\Sigma V^T \Rightarrow A^{-1} = V\Sigma^{-1} U^T. \quad (8.46)$$

Rozkład szczególny jest także formą diagonalizacji macierzy. Podobnie jak dla wartości własnych macierzy kwadratowej obowiązywało równanie (8.9) $AX = X\Lambda$, to dla wartości szczególnych mamy

$$AV = U\Sigma, \quad A^T U = V\Sigma. \quad (8.47)$$

Oba te równania „diagonalizują” macierz A , ale

- rozkład szczególny istnieje dla każdej macierzy, nie tylko kwadratowej i diagonalizowalnej,
- rozkład szczególny stosuje inne wektory przekształcenia z prawej i lewej strony macierzy A (inne wektory bazy dla dziedziny i obrazu przekształcenia liniowego reprezentowanego przez macierz A) i obie te bazy są zawsze ortogonalne (nie tylko w przypadku macierzy symetrycznych).

Ponieważ przekształcenia do rozkładu szczególnego wykorzystują macierze ortogonalne, uwarunkowanie zależności między elementami macierzy A a wartościami szczególnymi będzie zawsze poprawne. Rozkład szczególny istnieje zawsze i jest dobrze uwarunkowany, ale jego wyznaczenie jest bardziej kosztowne obliczeniowo niż wyznaczenie wartości własnych.

Wykorzystanie zależności (8.44) do wyznaczenia wartości szczególnych nie jest korzystne z uwagi na złe uwarunkowanie małych wartości własnych macierzy $A^T A$. Wskaźnik uwarunkowania macierzy $A^T A$ może być nawet większy od kwadratu wskaźnika uwarunkowania macierzy A . Zamiast tego wyznacza się wartości własne macierzy blokowej

$$H = \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix} \in R^{(n+m) \times (n+m)}. \quad (8.48)$$

Jest to symetryczna macierz, która ma tyle dodatnich wartości własnych ile wynosi rząd macierzy A . Pozostałymi wartościami własnymi macierzy H są liczby przeciwne i zera. Jeżeli $x = \begin{bmatrix} w \\ v \end{bmatrix} \in R^{n+m}$ jest wektorem własnym macierzy H odpowiadającym jej wartości własnej s , to

$$\begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = s \begin{bmatrix} v \\ w \end{bmatrix} \implies A^T w = sv, \quad Av = sw \quad (8.49)$$

oraz

$$A^T Av = sA^T w = s^2 v, \quad (8.50)$$

czyli s^2 jest wartością własną macierzy $A^T A$, a v odpowiadającym jej wektorem własnym. Należy wyznaczyć dodatnie wartości własne macierzy H , a z pierwszych n składowych odpowiadających im wektorów własnych utworzyć prawe wektory szczególne. Tak więc wyznaczenie wartości szczególnych macierzy A odbywa się

poprzez wyznaczenie wartości własnych macierzy H . Inne zastosowania wartości własnych i szczególnych zostaną przedstawione w podrozdziale 8.7.

8.7. Zastosowania wartości własnych i szczególnych

Wyznaczenie wartości własnych pomaga rozwiązać liczne problemy matematyczne i inżynierskie. W rozdziale 9 pokazano, że wyznaczenie wartości i wektorów własnych pozwala na pełne określenie rozwiązania liniowego układu równań różniczkowych zwyczajnych pierwszego stopnia. Jest to podstawowy model zachowania w funkcji czasu liniowych układów dynamicznych, bardzo szeroko wykorzystywany w automatyce, elektrotechnice, ekonomii, farmakologii itd.

Poniżej możemy zasygnalizować jedynie niektóre zastosowania wartości własnych w problemach związanych z przetwarzaniem informacji.

Wyznaczanie wskaźnika Google PageRank⁸ dla rankingu istotności stron www

Dla każdego zapytania do wyszukiwarki zwracana jest lista stron, które:

- pasują do zapytania (dopasowanie dotyczy nie tylko słów kluczowych użytych na stronie, ale również np. preferencji i profilu użytkownika),
- posortowane są według malejącej wartości wskaźnika PageRank, który ma być wskaźnikiem istotności stron i jest obliczany co kilkanaście tygodni dla wszystkich indeksowanych stron internetowych.

Wartość PageRank jest uzależniona od struktury połączeń między wszystkimi zaindeksowanymi stronami (ilość stron to $3 \cdot 10^{13}$ w 2014 roku⁹, prawdopodobnie ok. $5 \cdot 10^{13}$ w 2017 roku). Połączenia między stronami można zapisać w postaci rzadkiej macierzy G , w której $g_{ij} = 1$ oznacza, że do strony i -tej jest link ze strony j -tej.

Suma $r_i = \sum_j g_{ij}$, to ilość linków przychodzących dla i -tej strony, a $c_j = \sum_i g_{ij}$ – wychodzących z j -tej strony. Do każdej strony można przypisać pewną liczbę P_i oznaczającą istotność strony, równą ważonej sumie istotności stron P_j prowadzących do strony i -tej, przy czym waga jest odwrotnością liczby linków wychodzących z j -tej strony

$$P_i = \sum_{j:g_{ij} \neq 0} \frac{P_j}{c_j}. \quad (8.51)$$

⁸ <http://www.sirgroane.net/google-page-rank/>,
<https://www.mathworks.com/content/dam/mathworks/mathworks-dot-com/moler/exm/chapters/pagerank.pdf>

⁹ <http://www.statisticbrain.com/total-number-of-pages-indexed-by-google/>

Równania (8.51) można zapisać łącznie jako:

$$P = HP, \quad (8.52)$$

gdzie:

$$h_{ij} = \begin{cases} \frac{g_{ij}}{c_j}, & c_j \neq 0 \\ 0, & c_j = 0 \end{cases}, \quad P = \begin{bmatrix} \vdots \\ P_i \\ \vdots \end{bmatrix}. \quad (8.53)$$

Wektor ważności stron P jest zatem wektorem własnym macierzy H związanym z wartością własną równą 1.

Kilka dodatkowych założeń przyjętych w algorytmie PageRank prowadzi do modyfikacji równania (8.52).

Po pierwsze: przyjęto, że jeśli dana strona nie ma linków wychodzących, to jej istotność zostanie rozdzielona równo, między wszystkie strony, stąd: $h_{ij} = \frac{1}{n}$ jeżeli $c_j = 0$:

$$h_{ij} = \begin{cases} \frac{g_{ij}}{c_j}, & c_j \neq 0 \\ \frac{1}{n}, & c_j = 0 \end{cases}. \quad (8.54)$$

Po drugie: dodatkowo uwzględniono model zachowania użytkownika, który może nie skorzystać z linku, a wejść na stronę bezpośrednio (np. wpisując adres). W tym celu tworzona jest macierz A

$$A = pH + (1 - p) \frac{1}{n} \mathbf{1}_{n \times n}. \quad (8.55)$$

Jeżeli użytkownik otwiera stronę internetową, to może zrobić to przez użycie linku z innej strony lub bezpośrednio (wpisując adres). Przyjmuje się, że $p = 0,85$ to prawdopodobieństwo kliknięcia linku, a $(1 - p)$ to prawdopodobieństwo bezpośredniego otworzenia strony. Tak więc, w macierzy A występują dwa składniki: pierwszy – według którego użytkownik z prawdopodobieństwem p porusza się zgodnie z macierzą H i drugi – według którego użytkownik przechodzi do losowej strony spośród n dostępnych.

Po uwzględnieniu tych wszystkich modyfikacji modelu do wyznaczenia wektora istotności stron trzeba rozwiązać równanie

$$P = AP, \quad (8.56)$$

zamiast równania (8.52).

Macierz A ma elementy z zakresu $[0,1)$, a sumy elementów w kolumnach równe są 1. Z teorii macierzy o dodatnich elementach¹⁰ wynika istnienie niezerowego rozwiązania x równania $x = Ax$, którego składowe są dodatnie. W PageRank przyjęto, że wektor P wyznaczony z równania (8.56) jest dodatkowo normalizowany, tak by: $\sum_i P_i = 1$. Elementy tak uzyskanego wektora są oszacowaniem istotności poszczególnych stron. Rozwiązanie równania (8.56), czyli obliczenie wektora własnego można uzyskać na kilka sposobów. Na przykład:

Metoda 1. Wektor P jest rozwiązaniem jednorodnego równania: $(I - A)P = 0$, czyli wektorem własnym macierzy A , związanym z wartością własną równą 1.

Dla niewielkiej ilości stron wektor P można obliczyć za pomocą metody potęgowej, zaczynając np. od przybliżonego rozwiązania: $P_i = 1/n$.

W praktyce takie rozwiązanie nie mogłoby być stosowane z uwagi na wielkość macierzy A .

Metoda 2. Ponieważ macierz G jest macierzą rzadką, więc korzystne jest takie rozwiązanie, w którym ta właściwość zostanie wykorzystana. Macierz A może zostać zapisana jako:

$$A = pGD + ez^T, \quad (8.57)$$

gdzie:

- D – macierz diagonalna o elementach: $d_{ii} = \begin{cases} \frac{1}{c_j}, & c_j \neq 0 \\ 0, & c_j = 0 \end{cases}$,
- e – wektor o n elementach równych 1,
- z – wektor o elementach: $z_j = \begin{cases} \frac{1-p}{n}, & c_j \neq 0 \\ \frac{1}{n}, & c_j = 0 \end{cases}$.

Część ez^T odpowiada składowej związanej z przejściem na stronę bez użycia linku.

Równanie $P = AP$ można zapisać jako:

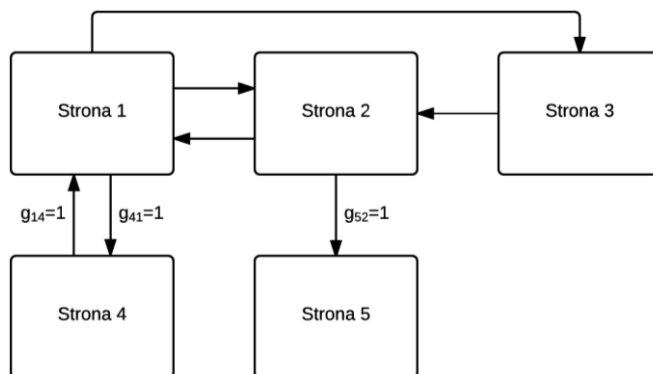
$$(I - pGD)P = ez^T P, \quad (8.58)$$

gdzie $z^T P$ jest nieznaną liczbą. Dla $p \in (0,1)$ macierz $I - pGD$ będzie nieosobliwa i można znaleźć rozwiązanie względem P , np. przyjmując, że $z^T P = 1$. Znaleziony wektor P należy znormalizować.

¹⁰ Henryk Minc, *Nonnegative matrices*, John Wiley&Sons, New York, 1988, ISBN 0-471-83966-3

Przykład 8.8

Schemat połączeń między stronami www pokazano na rysunku 8.1. Strona 1 zawiera link do stron 2, 3, 4, a strona 5 nie ma żadnych linków wychodzących.



Rys. 8.1. Schemat połączeń między stronami www

Temu schematowi odpowiada macierz połączeń:

$$G = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

i macierz A :

$$A = \begin{bmatrix} 0,03 & 0,4550 & 0,03 & 0,88 & 0,20 \\ 0,3133 & 0,03 & 0,88 & 0,03 & 0,20 \\ 0,3133 & 0,03 & 0,03 & 0,03 & 0,20 \\ 0,3133 & 0,03 & 0,03 & 0,03 & 0,20 \\ 0,03 & 0,4550 & 0,03 & 0,03 & 0,20 \end{bmatrix},$$

której elementy obliczane są zgodnie z równaniem (8.55), np.

$$g_{12} = 1, c_2 = 2, h_{12} = \frac{1}{2}, a_{12} = 0,85 * h_{12} + 0,15 * \frac{1}{5} = 0,455$$

$$g_{25} = 0, c_5 = 0, h_{25} = \frac{1}{5}, a_{25} = 0,85 * h_{25} + 0,15 * \frac{1}{5} = 0,2$$

W metodzie 2 po przekształceniach uzyskuje się układ równań odpowiadający (8.58):

$$\begin{bmatrix} 1 & -0,425 & 0 & -0,85 & 0 \\ -0,2833 & 1 & -0,85 & 0 & 0 \\ -0,2833 & 0 & 1 & 0 & 0 \\ -0,2833 & 0 & 0 & 1 & 0 \\ 0 & -0,425 & 0 & 0 & 1 \end{bmatrix} \cdot P = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},$$

skąd po rozwiązaniu i znormalizowaniu:

$$P = \begin{bmatrix} 0,2890 \\ 0,2603 \\ 0,1407 \\ 0,1407 \\ 0,1694 \end{bmatrix}.$$

Strony będą prezentowane wg malejącej wartości z wektora P , zatem kolejno strona 1, 2, 5, 3, 4. Wyniki wyszukiwania dla konkretnej frazy będą zawierały pasujące strony w tej kolejności.

Ukryte indeksowanie semantyczne (Latent Semantic Indexing – LSI)

Ukryte indeksowanie semantyczne to technika poszukiwania informacji „słów” w „dokumentach”, w której buduje się tzw. macierz wektorów dokumentów (macierz łączącą słowa z dokumentami, w których występują). Macierz taka jest macierzą rzadką i o bardzo dużych wymiarach. Korzystając z rozkładu szczególnego, obniża się wymiarowość macierzy – uwzględniane są składowe główne odpowiadające największym wartościom szczególnym, co prowadzi do odkrycia zależności między słowami o podobnym znaczeniu i utworzenie nowych „pseudo słów” będących ich kombinacją. Nowe słowa lepiej oddają semantyczną zawartość dokumentów. Nowy opis pozwala m.in. na wyszukiwanie dokumentów dla danego słowa – znajdowane będą również dokumenty, które tego słowa nie zawierają, ale ich tematyka jest zbliżona, a jednocześnie ilość przeszukiwanych danych zostaje zmniejszona.

Na przykład weźmy poniższe zdania – będzie to zbiór dokumentów:

- 1) całkowanie numeryczne korzysta z kwadratur,
- 2) skrypt dotyczący metod numerycznych,
- 3) skrypt shell może ułatwiać pracę programisty,
- 4) numeryczne rozwiązywanie równań różniczkowych,
- 5) metody numeryczne w pracy inżyniera,
- 6) równania różniczkowe w automatyce.

Wyrazy występujące w dokumencie zostały zmienione tak, by w macierzy wektorów nie występowały wielokrotnie w różnych odmianach, np. „równania” zamiast „równań”.

- 1) całkowanie numeryczne korzystać z kwadratura,
- 2) skrypt dotyczący metody numeryczne,
- 3) skrypt shell może ułatwiać praca programista,
- 4) numeryczne rozwiązywanie równanie różniczkowe,
- 5) metody numeryczne w praca inżynier,
- 6) równania różniczkowe w automatyka.

Wiersze macierzy wektorów dokumentów A odpowiadają słowom, kolumny dokumentom, a jej elementy (w najprostszym przypadku) będą jedynkami lub zerami, w zależności od tego czy dane słowo występuje lub nie w danym dokumencie. Definiujemy wektor q , będący wektorem o długości równej ilości wszystkich słów i mający jedynki w miejscach odpowiadających wyszukiwanym słowom. Przykład struktury A i q odpowiadający poszukiwaniu dokumentów związanych z terminem „całkowanie” ($q = [0,1,0,0,\dots,0]^T$) to:

$$\begin{array}{l}
 \text{słowa} = \left[\begin{array}{l}
 \text{automatyka} \\
 \text{całkowanie} \\
 \text{dotyczący} \\
 \text{inżynier} \\
 \text{korzystać} \\
 \text{kwadratura} \\
 \text{metoda} \\
 \text{może} \\
 \text{numeryczne} \\
 \text{praca} \\
 \text{programista} \\
 \text{równania} \\
 \text{różniczkowe} \\
 \text{rozwiązywanie} \\
 \text{shell} \\
 \text{skrypt} \\
 \text{ułatwić} \\
 \text{w} \\
 \text{z}
 \end{array} \right], \quad A = \underbrace{\begin{bmatrix}
 0 & 0 & 0 & 0 & 0 & 1 \\
 1 & 0 & 0 & 0 & 0 & 0 \\
 0 & 1 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 1 & 0 \\
 1 & 0 & 0 & 0 & 0 & 0 \\
 1 & 0 & 0 & 0 & 0 & 0 \\
 0 & 1 & 0 & 0 & 1 & 0 \\
 0 & 0 & 1 & 0 & 0 & 0 \\
 1 & 1 & 0 & 1 & 1 & 0 \\
 0 & 0 & 1 & 0 & 1 & 0 \\
 0 & 0 & 1 & 0 & 0 & 0 \\
 0 & 0 & 0 & 1 & 0 & 1 \\
 0 & 0 & 0 & 1 & 0 & 1 \\
 0 & 0 & 0 & 1 & 0 & 0 \\
 0 & 0 & 1 & 0 & 0 & 0 \\
 0 & 1 & 1 & 0 & 0 & 0 \\
 0 & 0 & 1 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 1 & 1 \\
 1 & 0 & 0 & 0 & 0 & 0
 \end{bmatrix}}_{\text{dokumenty}}, \quad q = \begin{bmatrix}
 0 \\
 1 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0 \\
 0
 \end{bmatrix}
 \end{array}$$

Mamy więc $n = 6$ dokumentów i $m = 19$ słów, wymiar macierzy A jest $m \times n$. Kolumna $[a_{ij}]_{i=1,\dots,m}$ macierzy A jest nazywana wektorem j -tego dokumentu – zawiera informacje o tym, które słowa są zawarte w dokumencie. Jeżeli wynik operacji $q^T [a_{ij}]_{i=1,\dots,m}$ jest niezerowy, czyli wektory zapytań i dokumentu nie są prostopadłe, to w j -tym dokumencie znajdują się słowa z zapytania. Im ten wynik jest większy, czyli im bardziej wektor zapytań zbliża się do wektora dokumentu, tym więcej słów z zapytania jest zawartych w dokumencie. Wynik działania $q^T A$

ujawnia wszystkie dokumenty, w których występują słowa z zapytania – w tym przykładzie *całkowanie*. Dla pytającego istotne jest zredukowanie liczby dokumentów i wybranie najbardziej istotnych dla zapytania. Z drugiej strony słowo *całkowanie* wystąpiło razem ze słowami *kwadratura* i *numeryczne*, należałoby odkryć ten związek i dołączyć odpowiednie dokumenty. Temu wszystkiemu służy rozkład szczególny i analiza składowych głównych.

Po zbudowaniu macierzy A wyznaczany jest jej rozkład szczególny $U\Sigma V^T$. Wybierane jest $k < n$ największych wartości szczególnych i – przy założeniu, że w obliczonym rozkładzie wartości szczególne są ułożone od największej do najmniejszej – wyznaczane macierze $U' \in R^{m \times k}$ (z k pierwszych kolumn macierzy U) oraz odpowiednio $\Sigma' \in R^{k \times k}$ (podmacierz $k \times k$ z lewego górnego rogu macierzy Σ) i $V'^T \in R^{k \times n}$ (k pierwszych wierszy macierzy V^T). Kolumny macierzy V' są (podrozdział 8.6) prawymi wektorami szczególnymi macierzy A związanymi z k największymi wartościami szczególnymi A . k -elementowe kolumny V'^T (albo wiersze macierzy V') składają się ze współrzędnych wektorów dokumentów w zredukowanej, k -wymiarowej przestrzeni. Z kolei zredukowany, k -elementowy wektor zapytania obliczany jest jako $q'^T = q^T U' (\Sigma')^{-1}$. Analizę kończy sprawdzenie „bliskości” wektora q' i kolejnych wierszy V'_i , $i = 1, \dots, n$ macierzy V' poprzez obliczenie (dla każdego dokumentu) tzw. współczynnika podobieństwa (*cosine similarity*), czyli:

$$\cos \angle(q', V'_i) = \frac{V'_i q'}{\|q'\|_2 \|V'_i\|_2}.$$

Dokumenty, dla których wartość podobieństwa jest największa, zawierają informacje dotyczące poszukiwanego hasła.

W naszym przykładzie (dla $k = 3$) wynik dla słowa „całkowanie” to:

$\cos \angle(q', V'_i)$	zdanie=dokument
0,98	całkowanie numeryczne korzysta z kwadratur
0,59	skrypt dotyczący metody numeryczne
0,3	metody numeryczne w praca inżynier
-0,1	numeryczne rozwiązywanie równania różniczkowe
-0,37	skrypt shell może ułatwiać praca programista
-0,52	równania różniczkowe w automatyka

Jak łatwo zauważyć, na szczycie listy wyników jest jedyne zdanie, w którym pojawiło się słowo „całkowanie”, dalej wyniki związane z metodami numerycznymi

i na końcu zdania mające (w świetle dostarczonych informacji) mniejszy związek z poszukiwanym hasłem.

W rzeczywistych zastosowaniach do elementów macierzy A stosowane są wagi uwzględniające m.in. ilość wszystkich wystąpień danego słowa we wszystkich dokumentach i w danym dokumencie.

Kompresja obrazu

Rozkład szczególny można wykorzystać do **kompresji obrazu**. Dla uproszczenia przyjmiemy obraz w odcieniach szarości (dla obrazu kolorowego należałoby powtórzyć operacje dla każdego z trzech kolorów) zapisany w formie macierzy M o wymiarze $(m \times n)$, gdzie poszczególne wartości odpowiadają znormalizowanym jasnościom pikseli.

W najprostszym przypadku algorytm będzie następujący:

- 1) dla macierzy M należy obliczyć rozkład szczególny: $M = U\Sigma V^T$, gdzie wymiary macierzy to $U - m \times m$, $\Sigma - m \times n$, $V - n \times n$; w macierzy Σ , podobnie jak w przypadku LSI, wartości szczególne powinny być ułożone od największej do najmniejszej,
- 2) wybrać wartość $r \leq \min\{m, n\}$ – ilość uwzględnianych składowych głównych, od której będzie zależeć jakość wynikowego obrazu oraz stopień kompresji,
- 3) wyznaczyć macierze $U' - m \times r$, $\Sigma' - r \times r$, $V' - n \times r$, biorąc pierwsze r kolumn (i pierwsze r wierszy w przypadku Σ') z odpowiednich macierzy,
- 4) obliczyć $M' = U'\Sigma'V'^T$.

Do obliczenia macierzy M' , która tak jak M ma n wierszy i m kolumn potrzebne jest $r(m + n + 1)$ liczb (elementów macierzy U', Σ', V'), w porównaniu do $m \cdot n$ dla oryginalnej macierzy M .

Skompresowany obraz (macierz M') uwzględnia r pierwszych składników sumy ze wzoru (8.42) – jest najlepszym przybliżeniem rzędu r macierzy M według normy euklidesowej, a norma błędu tej aproksymacji jest równa kolejnej wartości szczególnej: $\|M - M'_q\| = \sigma_{r+1}$. Pomijane są te składniki sumy, których wpływ na elementy macierzy jest najmniejszy.

Przykład 8.9

Zdjęcie pokazane na rysunku 8.2a o wymiarach 1085x1200 skompresowano, przyjmując $r = 100$ (rys. 8.2b), $r = 25$ (rys. 8.2c), $r = 10$ (rys. 8.2d), zmniejszając ilość liczb opisujących zdjęcie o odpowiednio: 82%, 96% i 98%. (np. w pierwszym przypadku zamiast 1302000 liczb, zapamiętywane jest: $100(1085+1200+1) = 228600$).

a)



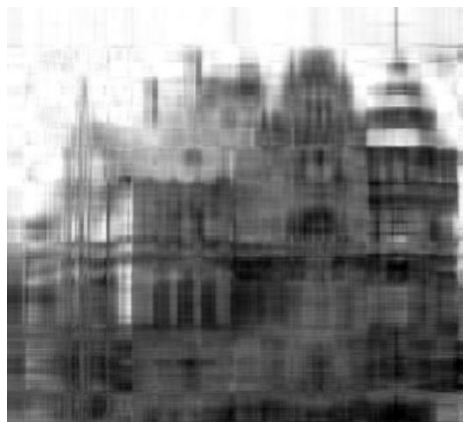
b)



c)

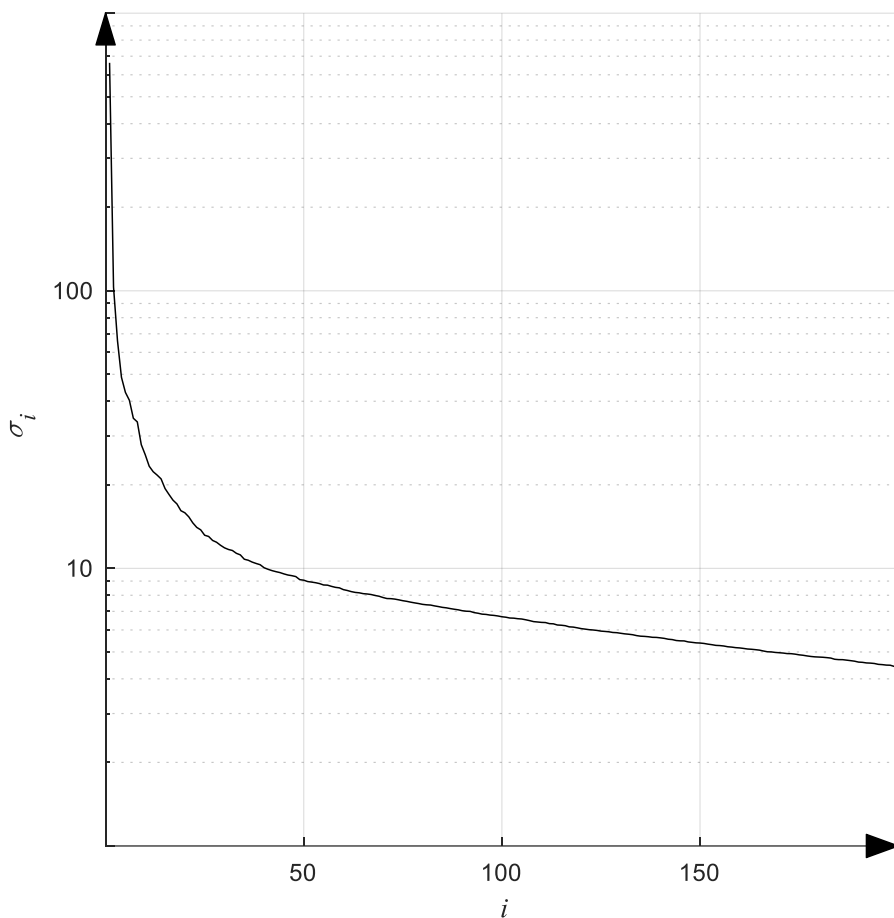


d)



Rys. 8.2. Kompresja obrazu z wykorzystaniem rozkładu SVD

Z drugiej strony, błąd kolejnych aproksymacji rośnie i wynosi $\sigma_{101} = 6,67$, $\sigma_{26} = 13,03$, $\sigma_{11} = 23,34$, co jest widoczne na rysunkach 8.2b-d. Największe wartości szczególne macierzy M pokazano na rysunku 8.3.



Rys. 8.3. 200 największych wartości szczególnych macierzy M

Z rysunku 8.3 wynika, że dobrym kompromisem między jakością obrazu a kompresją danych byłaby aproksymacja z zachowaniem $40 < r < 80$ wartości szczególnych.

9. Równania różniczkowe zwyczajne

9.1. Zagadnienie początkowe

Najbardziej ogólną postacią **równania różniczkowego zwyczajnego** rzędu n jest wyrażenie:

$$F\left(\frac{d^n y}{dx^n}, \dots, \frac{dy}{dx}, y, x\right) = 0, \quad (9.1)$$

w którym występuje nieznaną funkcją $y(x)$ zmiennej niezależnej x i jej pochodne. W wielu zastosowaniach zmienną niezależną jest czas ($x = t$). W takim przypadku pochodne przyjęto oznaczać kropkami $\dot{y} = \frac{dy}{dt}$, $\ddot{y} = \frac{d^2 y}{dt^2}$ itd.

Rozwiązaniem ogólnym równania różniczkowego zwyczajnego rzędu n nazywamy rodzinę wszystkich funkcji $y(x)$, klasy co najmniej C^n , spełniających równanie (9.1). Rozwiązanie ogólne można sparametryzować, stosując n stałych parametrów.

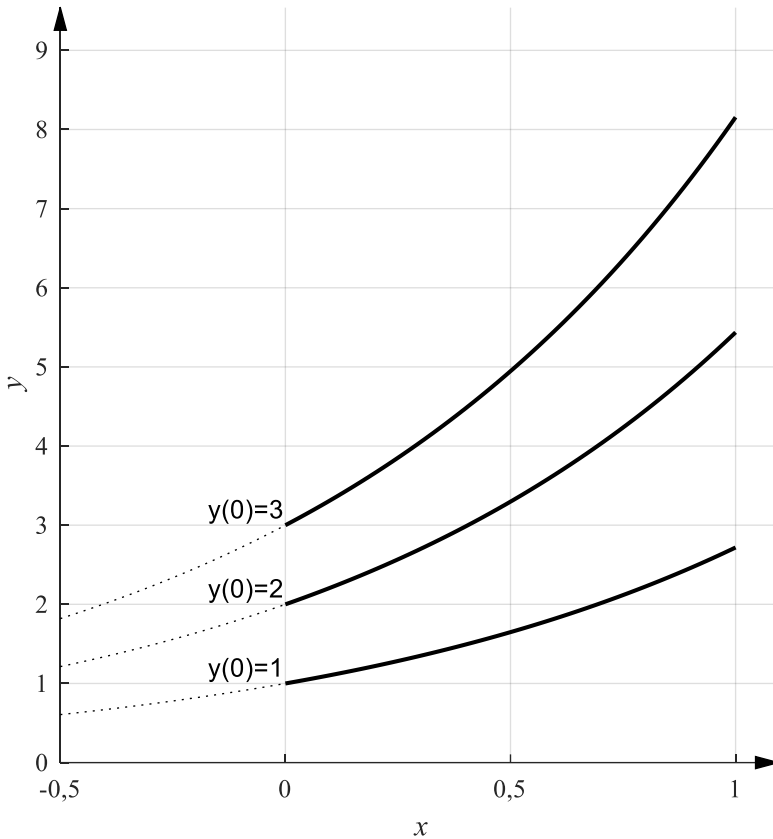
Jeśli w rozwiązaniu ogólnym, wszystkie te parametry zostaną ustalone otrzymamy **rozwiązanie szczególne**. Najczęściej odbywa się to przez podanie warunków na wartości funkcji i pochodnych dla wybranej wartości zmiennej niezależnej x (w wybranym punkcie).

Równanie rzędu n można sprowadzić (pod pewnymi warunkami) do układu n równań stopnia pierwszego, wtedy rozwiązaniem jest funkcja $y: R \rightarrow R^n$. Numerycznie można poszukiwać tylko szczególnych (nie ogólnych) rozwiązań równań różniczkowych. Rozwiązanie równania rzędu n lub też, co równoważne, układu n równań rzędu pierwszego, zależy od n stałych wyznaczanych z warunków, jakie spełniać ma poszukiwana funkcja. Jeśli wszystkie te warunki zadane są w jednym punkcie, czyli dla jednej wartości zmiennej niezależnej, mówimy, że dany jest problem początkowy, zwany inaczej **zagadnieniem początkowym** albo **problemem Cauchy'ego**:

Szukamy różniczkowalnej funkcji $y(x)$ spełniającej warunki:

$$\frac{dy}{dx} = f(y(x), x), \quad y(x_0) = y_0 \quad (9.2)$$

dla zadanych x_0 i y_0 .



Rys. 9.1. Rozwiązania szczególne równania $\frac{dy}{dx} = y$ dla różnych warunków początkowych

Nie w każdym przypadku zagadnienie początkowe ma rozwiązanie. Żeby zapewnić istnienie rozwiązania trzeba dokonać pewnych założeń dotyczących funkcji f występującej w równaniu różniczkowym.

Definicja 9.1

Funkcja $f: R^n \rightarrow R^n$ spełnia w zbiorze $S \subset R^n$ lokalnie **warunek Lipschitza** (jest lokalnie lipschitzowska), jeśli dla każdego punktu $y \in S$ istnieje otoczenie $S_y \subset S$, w którym

$$\forall_{z \in S_y} \|f(y) - f(z)\| \leq K \|y - z\| \quad (9.3)$$

dla pewnej stałej K nazywanej stałą Lipschitza. Funkcja $f: R^n \rightarrow R^n$ spełnia na zbiorze $S \subset R^n$ warunek Lipschitza jeśli warunek (9.3) jest spełniony dla każdego $y, z \in S$ z tą samą stałą K . Jeżeli $S = R^n$, to funkcja f spełnia warunek Lipschitza globalnie.

Funkcja $f: R \rightarrow R$ spełniająca warunek Lipschitza na przedziale I jest ciągła i różniczkowalna prawie wszędzie na I , a moduł jej pochodnej jest ograniczony przez stałą Lipschitza, jest więc także jednostajnie ciągła.

Jeżeli $f: R^n \rightarrow R^n$ jest w wypukłym zbiorze $S \subset R^n$ różniczkowalna i istnieje stała K , taka że $\left\| \frac{\partial f}{\partial y} \right\| \leq K$, to funkcja f spełnia na zbiorze S warunek Lipschitza ze stałą K .

O istnieniu rozwiązania zagadnienia początkowego rozstrzygają następujące twierdzenia. Zwykle są one formułowane przy założeniu ciągłości względem x , można je jednak uogólnić na przypadek odcinkowej ciągłości, tak jak to sformułowano w twierdzeniach 9.1, 9.2. Takie sformułowanie jest przydatne np. w analizie układów sterowania.

Twierdzenie 9.1 (o lokalnym istnieniu i jednoznaczności rozwiązania równania różniczkowego – Khalil, 2002)

Jeżeli funkcja $f(y, x)$ jest odcinkowo ciągła względem x i spełnia warunek Lipschitza lokalnie w punkcie y_0 , jednostajnie względem $x \in [x_0, x_k)$, to znaczy:

$$\forall_{z, y \in S_{y_0}} \forall_{x \in [x_0, x_k)} \|f(y, x) - f(z, x)\| \leq K \|y - z\|, \quad (9.4)$$

to istnieje $\varepsilon > 0$, takie że równanie (9.2) ma jedyne rozwiązanie na przedziale $[x_0, x_0 + \varepsilon]$.

Do istnienia przynajmniej jednego rozwiązania zagadnienia początkowego (9.2) wystarcza ciągłość funkcji $f(y, x)$, natomiast przykładem na to, że ciągłość nie jest wystarczająca dla jednoznaczności rozwiązania jest równanie $\frac{dy}{dx} = \sqrt[3]{y}$ z warunkiem $y(0) = 0$, którego rozwiązaniami są $y(x) = 0$ i $y(x) = \left(\frac{2}{3}x\right)^{\frac{3}{2}}$.

Twierdzenie 9.1 gwarantuje istnienie rozwiązania na „małym” przedziale zmiennej niezależnej – przedział $[x_0, x_0 + \varepsilon]$ może być znacznie mniejszy od przedziału $[x_0, x_k)$. Do uzyskania „przedłużalności” rozwiązania – to jest zapewnienia jego istnienia na zadanym, (niekoniecznie małym) przedziale zmiennej niezależnej, trzeba wzmocnić warunek Lipschitza, tak był spełniony globalnie. Opisuje to kolejne twierdzenie.

Twierdzenie 9.2 (o przedłużalności rozwiązania – *Khalil, 2002*)

Jeżeli funkcja $f(y, x)$ jest odcinkowo ciągła względem x , i spełnia warunek Lipschitza globalnie, jednostajnie względem $x \in [x_0, x_k]$, to znaczy:

$$\forall y, z \in \mathbb{R}^n \quad \forall x \in [x_0, x_k] \quad \|f(y, x) - f(z, x)\| \leq K\|y - z\|, \quad (9.5)$$

oraz $\|f(y_0, x)\| \leq k < \infty$, to zagadnienie początkowe (9.2) ma jedyne rozwiązanie na przedziale $[x_0, x_k]$.

Przykładem równania, którego rozwiązanie nie jest dowolnie przedłużalne jest $\frac{dy}{dx} = -y^2$ z warunkiem $y(0) = -1$, którego rozwiązaniem jest funkcja $y(x) = \frac{1}{x-1}$ dążąca do nieskończoności gdy x dąży do 1 z lewej strony.

Rozwiązania numeryczne są konstruowane na ograniczonym przedziale $[x_0, x_k]$ dla takich równań, dla których istnieje rozwiązanie na przedziale $[x_0, x_k]$, zgodnie z powyższymi twierdzeniami.

Przykład 9.1 (Sprowadzenie równania wyższego rzędu do układu równań rzędu pierwszego).

Należy rozwiązać zagadnienie początkowe

$$\ddot{y} - \dot{y}x = y^2\dot{y} + y, \quad y(0) = y_0, \quad \dot{y}(0) = y_1, \quad \ddot{y}(0) = y_2.$$

Ponieważ zdecydowana większość standardowych procedur rozwiązywania numerycznego równań różniczkowych dotyczy wektorowych równań rzędu pierwszego (poza tym istnieje pewna ilość procedur dla równań wektorowych drugiego rzędu, które mają zastosowanie w mechanice i astronomii, a procedury rozwiązujące bezpośrednio równanie n -tego rzędu są rzadkością), więc równanie to trzeba przekształcić do standardowej postaci.

Typowym podejściem (aczkolwiek nie jedynym możliwym) jest posłużenie się tzw. współrzędnymi fazowymi. Oznaczmy $z_1 = y, z_2 = \dot{y}, z_3 = \ddot{y}$. Zachodzą więc równości

$$\dot{z}_1 = \dot{y} = z_2,$$

$$\dot{z}_2 = \ddot{y} = z_3,$$

$$\dot{z}_3 = \ddot{y} = z_1^2 z_2 + z_1 + x z_3$$

oraz $z_1(0) = y_0, z_2(0) = y_1, z_3(0) = y_2$. Otrzymaliśmy więc układ równań, który można zapisać w postaci wektorowej

$$\dot{z} = f(x, z), \quad f(x, z) = \begin{bmatrix} z_2 \\ z_3 \\ z_1^2 z_2 + z_1 + x z_3 \end{bmatrix}, \quad z(0) = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix}.$$

Wyprowadzony układ równań różniczkowych pierwszego rzędu, ma tę właściwość, że pierwsza współrzędna jego rozwiązania jest rozwiązaniem wyjściowego równania trzeciego rzędu i odwrotnie: rozwiązanie wyjściowego równania po uwzględnieniu definicji z_1 i z_2 jest rozwiązaniem otrzymanego układu równań.

Taki układ równań może być bezpośrednio rozwiązywany przez standardowe procedury numeryczne.

9.2. Numeryczne rozwiązanie zagadnienia początkowego

Numeryczne rozwiązanie zagadnienia początkowego polega na wyznaczeniu kolejnych wartości y_i , $i = 0, 1, 2, \dots$, które w wybranych punktach zmiennej niezależnej x_i , $i = 0, 1, 2, \dots$ mają przybliżać wartości dokładnego rozwiązania $y(x_i)$, $i = 0, 1, 2, \dots$. Ta część algorytmu rozwiązującego zagadnienie początkowe, która odpowiada za generowanie kolejnych wartości y_i , $i = 0, 1, 2, \dots$ jest nazywana **schematem różnicowym**. Przejście do kolejnej wartości x_i , y_i to **krok schematu różnicowego**, a odległość między kolejnymi dwiema wartościami x_i to **długość kroku schematu różnicowego**, najczęściej oznaczana przez $h > 0$.

W praktyce, rozwiązania numeryczne poszukujemy zawsze na ograniczonym przedziale $[x_0, x_k]$ i trzeba je uzyskać po skończonej liczbie operacji arytmetycznych, więc liczba kroków i otrzymanych wartości przybliżonych y_i musi być skończona.

Schematy różnicowe dzieli się na grupy posługując się różnymi kryteriami:

Ze względu na przyrost zmiennej niezależnej stosowany na poszczególnych etapach obliczeń:

- schemat różnicowy **stałokrokowy**, w którym przyrost zmiennej niezależnej jest taki sam w każdym kroku,
- schemat różnicowy **zmiennokrokowy**, w którym stosujemy sterowanie długością kroku (przyrostem zmiennej niezależnej), dostosowując ją w każdym kroku do wymagań dotyczących dokładności otrzymanego rozwiązania.

Ze względu na ilość informacji wykorzystywanej do wykonania jednego kroku

- schemat różnicowy **jednokrokowy** – do obliczenia y_{i+1} wykorzystuje informację o y_i ,
- schemat różnicowy **k -krokowy (wielokrokowy)** – do obliczenia y_{i+1} wykorzystuje informację o $y_i, y_{i-1}, \dots, y_{i-k+1}$.

Ze względu na postać równań wykorzystywanych na pojedynczym etapie obliczeń:

- schemat różnicowy **jawny (otwarty)** podaje jawną zależność, z której można obliczyć y_{i+1} ,
- schemat różnicowy **niejawny (zamknięty)** podaje równanie (algebraiczne, nieliniowe), które trzeba rozwiązać, żeby obliczyć y_{i+1} .

Po zastosowaniu schematu różnicowego otrzymujemy dyskretny ciąg wartości, które przybliżają rozwiązanie dokładne w wybranych punktach. Wartość przybliżoną między punktami x_i obliczamy najczęściej korzystając z interpolacji odcinkowo-liniowej, niekiedy wielomianowej (np. metody BDF opisane w podrozdziale 9.8).

Oprócz zastosowania schematu różnicowego, to jest obliczenia przybliżonych wartości rozwiązania, konieczna jest analiza błędu otrzymywanego rozwiązania. W metodach zmiennokrokowych taka analiza jest dokonywana automatycznie, w każdym kroku i jej wnioski służą do akceptacji lub odrzucenia otrzymanego rozwiązania i do pojęcia decyzji o długości następnego kroku.

Każde równanie różniczkowe (9.2) można zapisać w równoważnej postaci równania całkowego, a schematy różnicowe mają wiele wspólnego z metodami całkowania numerycznego. Z tego powodu metody numeryczne rozwiązywania równań różniczkowych nazywa się także metodami całkowania równań różniczkowych.

Rozważmy schemat jednokrokowy, z krokiem h , na przedziale $[x_0, x_0 + Nh]$. Niech $x_n = x_0 + nh$, a y_n niech będzie rozwiązaniem przybliżonym odpowiadającym dokładnemu rozwiązaniu $y(x_n)$. Na całkowity błąd, jakim jest obarczone rozwiązanie przybliżone składają się, jak w każdej metodzie numerycznej, błąd zaokrągleń i błąd metody. Błąd zaokrągleń zależy przede wszystkim od arytmetyki zmiennopozycyjnej, w której prowadzone są obliczenia, a błąd metody scharakteryzujemy w tym rozdziale. W przypadku numerycznego rozwiązywania równania różniczkowego można mówić o kilku rodzajach błędów metody. Precyzują to poniższe definicje, odnosząc się także do zależności błędów schematu różnicowego od długości kroku.

Definicja 9.2 (błędów metody rozwiązywania zagadnienia początkowego):

Błędem schematu różnicowego nazwiemy ciąg

$$\varepsilon_h(x_n) = y(x_n) - y_n, \quad n = 1, 2, \dots, N. \quad (9.6)$$

Błędem globalnym liczbę:

$$\varepsilon_h = \max_n \|y(x_n) - y_n\|. \quad (9.7)$$

Błąd globalny jest **zbieżny** jeśli $\lim_{h \rightarrow 0} \varepsilon_h = 0$, a **zbieżny z rzędem p** jeśli dodatkowo

$$\varepsilon_h < Ch^p \quad (9.8)$$

dla pewnej stałej C .

Jeżeli oznaczymy \bar{y}_{n+1} rozwiązanie otrzymane po wykonaniu kroku o długości h z punktu początkowego $(y(x_n), x_n)$, to **błędem lokalnym** nazwiemy

$$r_{n+1}(h) = y(x_{n+1}) - \bar{y}_{n+1}. \quad (9.9)$$

Jeżeli błąd lokalny metody można przedstawić w postaci rozwinięcia w szereg

$$r_{n+1}(h) = \varphi(y(x_n), x_n)h^{p+1} + O(h^{p+2}), \quad (9.10)$$

to mówimy, że **metoda jest rzędu p** .

Jeżeli błąd lokalny można przedstawić w postaci

$$r_{n+1}(h) = h\tau_{n+1}(h),$$

to $\tau_{n+1}(h)$ jest nazywany **lokalnym błędem odcięcia**.

Jeżeli na początku kroku znaleźliśmy się w punkcie \bar{y}_{n+1} , różnym od dokładnej wartości rozwiązania $y(x_{n+1})$, to w następnym kroku, jeśli nie jest obciążony żadnymi błędami, będziemy poruszać się wzdłuż rozwiązania lokalnego wybranego przez warunek początkowy $y(x_{n+1}) = \bar{y}_{n+1}$, a nie wzdłuż rozwiązania wyróżnionego przez warunek $y(x_0) = y_0$. Błąd lokalny obrazuje zachowanie schematu różnicowego w tym jednym kroku. Im wyższy rząd metody tym metoda jest dokładniejsza, a błąd popełniany w jednym kroku mniejszy (długość kroku h jest znacznie mniejsza od 1). Rząd metody można definiować także przez lokalny błąd odcięcia. Dla metody rzędu p $\tau_{n+1}(h) = O(h^p)$ dla $h \rightarrow 0$.

Jeżeli $\tau_{n+1}(h) \rightarrow 0$ dla $h \rightarrow 0$, to mówimy, że schemat różnicowy jest **zgodny** z zadaniem początkowym, które rozwiązuje. Jawny jednokrokowy schemat różnicowy można zapisać w zwartej formie $y_{n+1} = y_n + \Phi_f(y_n, x_n, h)h$ (Φ_f jest nazywana **funkcją przyrostową lub inkrementalną**). Zgodnie z (9.9) $y(x_{n+1}) = y(x_n) + \Phi_f(y(x_n), x_n, h)h + h\tau_{n+1}(h)$, więc $\frac{y(x_{n+1}) - y(x_n)}{h} = \Phi_f(y(x_n), x_n, h) + \tau_{n+1}(h)$. Jeżeli schemat różnicowy jest zgodny, to

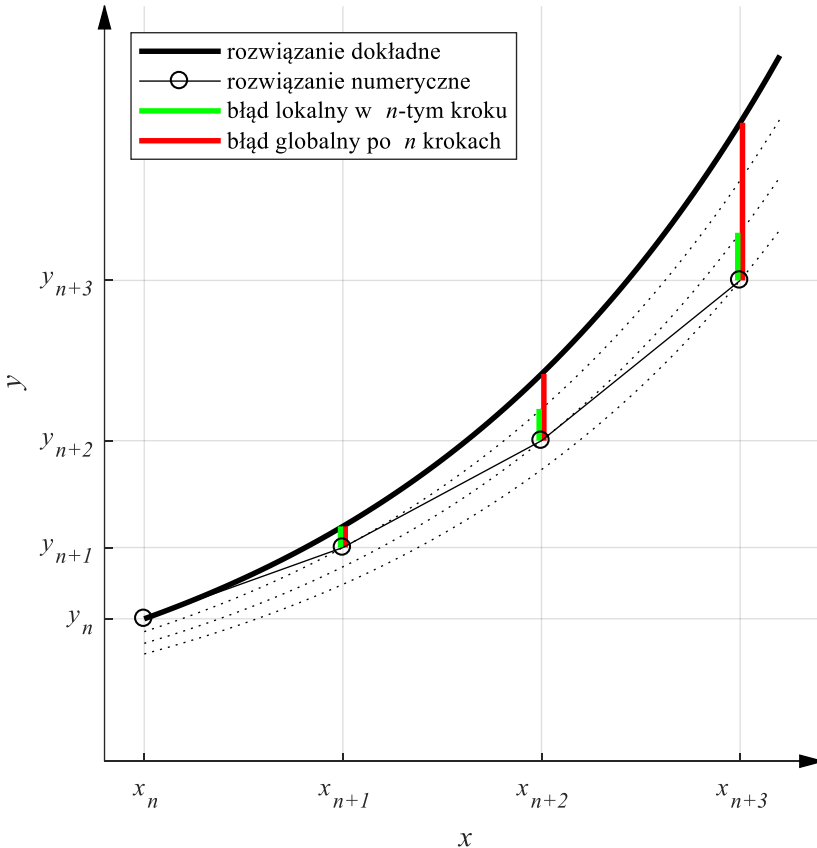
$$f(y(x_n), x_n) = \left. \frac{dy}{dx} \right|_{x=x_n} = \lim_{h \rightarrow 0} \frac{y(x_{n+1}) - y(x_n)}{h} = \lim_{h \rightarrow 0} \Phi_f(y(x_n), x_n, h).$$

Można więc powiedzieć, że zgodny schemat różnicowy odtwarza dla $h \rightarrow 0$ równanie różniczkowe, które rozwiązuje numerycznie. Rozwiązanie numeryczne odbywa się ze skończoną, niezerową długością kroku, więc choć uzyskane metodą zgodną, będzie różniło się od rozwiązania dokładnego. Schemat różnicowy pierwszego rzędu (według definicji 9.2) jest zgodny.

Błąd schematu różnicowego i maksimum jego normy, czyli błąd globalny odnoszą się do rozbieżności między rozwiązaniem numerycznym, a rozwiązaniem dokładnym z warunkiem początkowym $y(x_0) = y_0$, którego szukamy. Wielkość błędu

globalnego jest wynikiem kumulowania się błędów lokalnych, ale wpływ na błąd globalny ma także charakter zmienności rodziny rozwiązań równania różniczkowego dla różnych warunków początkowych, to znaczy to czy rozwiązania te zbliżają się czy oddalają od siebie ze wzrostem zmiennej niezależnej.

Na rysunku 9.2 pokazano błąd schematu różnicowego, błędy lokalne i błąd globalny.



Rys. 9.2. Błąd lokalny i globalny schematu różnicowego na przykładzie jawnej metody Eulera (rozwiązanie równania $\frac{dy}{dx} = y, y(0) = 1$ z krokiem $h = 1$)

Przy rozwiązywaniu równania różniczkowego szacowany jest błąd lokalny. Związek między błędem lokalnym a globalnym i uzasadnienie tego, że ograniczając błąd lokalny panujemy jednocześnie nad błędem globalnym daje poniższe twierdzenie.

Twierdzenie 9.3 (o zbieżności rozwiązania numerycznego – *Krupowicz, 1986*)

Jeżeli dokładne rozwiązanie zagadnienia początkowego jest gładkie, schemat różnicowy jest zgodny i jest rzędu $p \geq 1$ (to znaczy błąd lokalny spełnia warunek (9.10), to błąd globalny jest zbieżny z rzędem p (to znaczy błąd globalny spełnia warunek (9.8)).

Definicja 9.3

Schemat różnicowy z krokiem h będzie nazywany **stabilnym** jeżeli z ograniczoneści rozwiązania dokładnego wynika ograniczoność rozwiązania przybliżonego, przy liczbie kroków $n \rightarrow \infty$.

Aby schemat różnicowy był stabilny zgodnie z definicją 9.3, to błąd metody nie może narastać w kolejnych krokach. Załóżmy, że w kolejnych krokach otrzymujemy $y(x_0) = y_0$, $y(x_1) = y_1 + \varepsilon_1$, $y(x_2) = y_2 + \varepsilon_2, \dots$, gdzie $y(x_n)$ jest rozwiązaniem dokładnym, y_n – przybliżeniem, a ε_n błędem po n -tym kroku. Jeżeli potrafimy przedstawić błąd w postaci

$$\varepsilon_{n+1} \approx G\varepsilon_n, \quad n = 1, 2, \dots, \quad (9.11)$$

to macierz G (nazywana **macierzą wzmocnienia błędu**), która jest zależna od długości kroku h , decyduje o stabilności metody. Żeby błąd nie narastał musi ona mieć wszystkie wartości własne o module mniejszym od 1. Jest to znany warunek stabilności liniowego równania różnicowego (9.11).

Stabilność schematu różnicowego jest wymaganiem elementarnym. Schemat niestabilny będzie po prostu generował co raz większe wartości z rosnącą liczbą wykonanych kroków.

W teorii metod numerycznych sformułowano wiele definicji stabilności schematów różnicowych. Różnią się one kontekstami i szczegółami definicji, mogą być inne dla schematów jednokrokowych i wielokrokowych. W przypadku schematów jednokrokowych szczególnie popularne jest pojęcie stabilności absolutnej (lub bezwzględnej), odwołujące się do równania liniowego pierwszego rzędu, jako do problemu testowego.

Definicja 9.4

Jeżeli $y(x)$ jest rozwiązaniem dokładnym równania $\frac{d}{dx}y = \lambda y$, $y(0) = 1$, gdzie λ może być liczbą zespoloną, to schemat różnicowy z krokiem h generujący rozwiązanie przybliżone y_n jest **absolutnie stabilny**, jeżeli wartości przybliżone y_n pozostają ograniczone dla $n \rightarrow \infty$.

Rozwiązanie przybliżone y_n zależy od parametru równania λ i długości kroku h . **Obszarem stabilności absolutnej** schematu różnicowego nazywamy podzbiór płaszczyzny zespolonej $z = \lambda h$, w którym jest spełniony warunek z definicji 9.4. Jeżeli obszar ten zawiera lewą półpłaszczyznę zmiennej zespolonej, to mówimy, że schemat różnicowy jest **A-stabilny**, jeżeli sektor kątowy lewej półpłaszczyzny to **$A(\alpha)$ -stabilny**. Jeżeli schemat jest A-stabilny, to ograniczone rozwiązanie numeryczne uzyskamy przy dowolnie długim kroku. Jeśli obszar stabilności jest ograniczony, to zakres długości kroku generujących ograniczone rozwiązania numeryczne jest też ograniczony. Dokładniejszy opis konsekwencji ograniczonego obszaru stabilności absolutnej przedstawimy w podrozdziale 9.4.

9.3. Liniowe równania różniczkowe

Równania różniczkowe postaci

$$\frac{d}{dx}y = Ay, \quad y(0) = y_0, \quad (9.12)$$

gdzie A jest stałą macierzą współczynników o wymiarze $n \times n$, nazywamy **liniowym, stacjonarnym równaniem różniczkowym** (układem liniowych równań różniczkowych). To, że współczynniki równania nie zależą od zmiennej x , pozwala bez utraty ogólności rozumowania rozwiązywać równanie z warunkiem początkowym określonym dla $x = 0$.

Równań liniowych nie trzeba rozwiązywać numerycznie – można otrzymać analityczną postać rozwiązania. Wyprowadzimy rozwiązanie równania (9.12) przy założeniu, że macierz A ma różne wartości własne $\lambda_i, i = 1, \dots, n$. Z rozdziału 8

wiadomo, że wtedy $V^{-1}AV = \Lambda$, gdzie $\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \lambda_n \end{bmatrix}$, V jest macierzą,

której kolumnami są (liniowo niezależne) wektory własne odpowiadające wartościom własnym λ_i . Jeżeli wprowadzimy w równaniu (9.12) zamiannę zmiennych

$$Vz(x) = y(x), \quad (9.13)$$

to otrzymamy jego nową postać:

$$\frac{d}{dx}z(x) = V^{-1}AVz(x), \quad z(0) = V^{-1}y(0) = V^{-1}y_0 = z_0. \quad (9.14)$$

Macierz $V^{-1}AV = \Lambda$ jest diagonalna, więc układ równań (9.14) można zapisać w postaci równań skalarnych

$$\frac{d}{dx}z_i(x) = \lambda_i z_i(x), \quad z_i(0) = z_{i_0}, \quad i = 1, 2, \dots, n. \quad (9.15)$$

Dla takich równań znamy rozwiązanie:

$$z_i(x) = e^{\lambda_i x} z_{i0}, \quad i = 1, 2, \dots, n, \quad (9.16)$$

albo w postaci macierzowej

$$z(x) = \begin{bmatrix} e^{\lambda_1 x} & 0 & \dots & 0 \\ 0 & e^{\lambda_2 x} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & e^{\lambda_n x} \end{bmatrix} z(0). \quad (9.17)$$

Wracając do oryginalnego równania dostajemy:

$$y(x) = Vz(x) = V \begin{bmatrix} e^{\lambda_1 x} & 0 & \dots & 0 \\ 0 & e^{\lambda_2 x} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & e^{\lambda_n x} \end{bmatrix} V^{-1}y(0). \quad (9.18)$$

Jeżeli wiersze macierzy V^{-1} oznaczymy przez $V^{-1} = \begin{bmatrix} w_1^T \\ w_2^T \\ \vdots \\ w_n^T \end{bmatrix}$, to (9.18) można

zapisać jako

$$y(x) = \sum_{i=1}^n e^{\lambda_i x} v_i w_i^T y(0). \quad (9.19)$$

Rozwiązanie stacjonarnego równania liniowego zostało przedstawione w postaci sumy n składników. Każdy z nich jest nazywany **modem**. Każdy mod jest związany z jedną wartością własną: jest wektorem o kierunku wektora własnego v_i (bo $w_i^T y(0)$ jest skalarzem), zmieniającym się zgodnie z funkcją $e^{\lambda_i x}$. Jeżeli λ_i jest liczbą rzeczywistą przebieg $e^{\lambda_i x}$ jest aperiodyczny (nie występują w nim oscylacje), rosnący do nieskończoności jeśli $\lambda_i > 0$, dążący do zera jeśli $\lambda_i < 0$ i stały dla $\lambda_i = 0$. Jeśli λ_i jest liczbą zespoloną, to istnieje sprzężona do niej wartość własna, powiedzmy λ_{i+1} . Suma dwóch modów związanych z tymi wartościami własnymi jest rzeczywista (co można wyprowadzić ze wzoru Eulera), ale zawiera oscylacje o pulsacji związanej z częścią urojoną $\text{Im}\{\lambda_i\}$. Amplituda tych oscylacji rośnie do nieskończoności gdy $\text{Re}\{\lambda_i\} > 0$, maleje do zera gdy $\text{Re}\{\lambda_i\} < 0$, a jest stała gdy $\text{Re}\{\lambda_i\} = 0$. Im dalej na lewo od osi urojonych na płaszczyźnie Gaussa leży wartość własna λ_i , tym szybciej zanika mod związany z tą wartością własną.

Liniowe równania różniczkowe są używane jako test dla metod numerycznych, i w takim kontekście będziemy się do nich odwoływać. Dokładniejszą analizę rozwiązania liniowego równania różniczkowego, także w przypadku wielokrotnych wartości własnych macierzy współczynników, można znaleźć w wielu podręcznikach.

9.4. Schematy różnicowe jednokrokowe niskiego rzędu i ich najważniejsze cechy

Jawna metoda Eulera – przykład metody o ograniczonym obszarze stabilności

Jeżeli w równaniu różniczkowym (9.2) zastąpimy pochodną ilorazem różnicowym w następujący sposób:

$$\frac{dy}{dx} = f(y(x), x) \Rightarrow \frac{y_{n+1} - y_n}{h} \approx f(y_n, x_n) \quad (9.20)$$

to otrzymamy schemat różnicowy

$$y_{n+1} = y_n + f(y_n, x_n)h \quad (9.21)$$

nazywany jawną **metodą Eulera**. Na przykład, w przypadku układu liniowych równań różniczkowych jawna metoda Eulera jest opisana równaniem $y_{n+1} = y_n + hAy_n = (I + hA)y_n$.

Zgodnie z definicją, jeżeli błąd lokalny metody można przedstawić w postaci rozwinięcia w szereg $r_{n+1}(h) = \varphi(y(x_n), x_n)h^{p+1} + O(h^{p+2})$, to metoda ma rząd zbieżności p . W jawnej metodzie Eulera błąd lokalny jest równy:

$$\begin{aligned} r_{n+1}(h) &= y(x_{n+1}) - y_{n+1} = y(x_{n+1}) - y(x_n) - hf(y(x_n), x_n) = \\ &= h \left[\frac{y(x_{n+1}) - y(x_n)}{h} - \frac{dy(x)}{dx} \Big|_{x=x_n} \right] = h \left(\frac{h}{2!} y''(x_n) + \dots \right) = O(h^2), \end{aligned} \quad (9.22)$$

czyli rząd zbieżności metody Eulera jest równy 1.

Jawna metoda Eulera jest zgodna, dla $h \rightarrow 0$ $\frac{y_{n+1} - y_n}{h} \rightarrow \frac{dy}{dx} \Big|_{x=x_n} = f(y(x_n), x_n)$.

Macierz wzmocnienia błędu można wyprowadzić w metodzie Eulera w następującym ciągu przekształceń wynikającym z rozwinięcia różnicy progresywnej w szereg Taylora:

$$\begin{aligned} &\frac{y(x_{n+1}) - y(x_n)}{h} \\ &= \frac{dy(x)}{dx} \Big|_{x=x_n} + \frac{h}{2!} y''(x_n) + \dots = f(y(x_n), x_n) + \frac{h}{2!} y''(x_n) + \dots, \end{aligned} \quad (9.23)$$

czyli po uwzględnieniu, że $y_{n+1} + \varepsilon_{n+1} = y(x_{n+1})$, $y_n + \varepsilon_n = y(x_n)$, mamy

$$y_{n+1} + \varepsilon_{n+1} - y_n - \varepsilon_n = hf(y_n + \varepsilon_n, x_n) + \frac{h^2}{2!} y''(x_n) + \dots \quad (9.24)$$

Po podstawieniu $y_{n+1} = y_n + f(y_n, x_n)h$ dostajemy

$$y_n + hf(y_n, x_n) + \varepsilon_{n+1} - y_n - \varepsilon_n = hf(y_n + \varepsilon_n, x_n) + \frac{h^2}{2!}y''(x_n) + \dots, \quad (9.25)$$

skąd można wyliczyć

$$\varepsilon_{n+1} = \varepsilon_n + h[f(y_n + \varepsilon_n, x_n) - f(y_n, x_n)] + \frac{h^2}{2!}y''(x_n) + \dots \quad (9.26)$$

Ponowne wykorzystanie szeregu Taylora daje

$$\varepsilon_{n+1} = \left(I + h \frac{\partial f}{\partial y} \Big|_{x=x_n, y=y_n} \right) \varepsilon_n + O(\varepsilon_n^2) + O(h^2), \quad (9.27)$$

czyli macierz wzmocnienia błędu w jawnej metodzie Eulera jest równa

$$G = I + h \frac{\partial f}{\partial y} \Big|_{\substack{x=x_n \\ y=y_n}}.$$

„Testowym” równaniem dla którego określamy stabilność metody jest układ liniowych równań różniczkowych stopnia pierwszego:

$$\frac{d}{dx}y = Ay \quad (9.28)$$

i wtedy $G = I + h \frac{\partial f}{\partial y} \Big|_{x=x_n, y=y_n} = I + hA$. Jeśli liczby (między którymi mogą być zespolone) λ_i są wartościami własnymi macierzy A , to wartościami własnymi macierzy G są liczby $\gamma_i = 1 + h\lambda_i$. Moduł tych wartości własnych jest mniejszy od 1 ($|\gamma_i| < 1$) wtedy i tylko wtedy, gdy liczby zespolone $h\lambda_i$ znajdują się wewnątrz okręgu o promieniu 1 i środku $-1 + j0$ na płaszczyźnie zespolonej. Jest to jednocześnie **obszar stabilności absolutnej** tej metody (zdefiniowany w definicji 9.4). Faktycznie, rozwiązaniem numerycznym równania z definicji 9.4 jest ciąg $y_{n+1} = y_n + h\lambda y_n = (1 + h\lambda)y_n$, który będzie ograniczony, jeśli $|1 + h\lambda| \leq 1$.

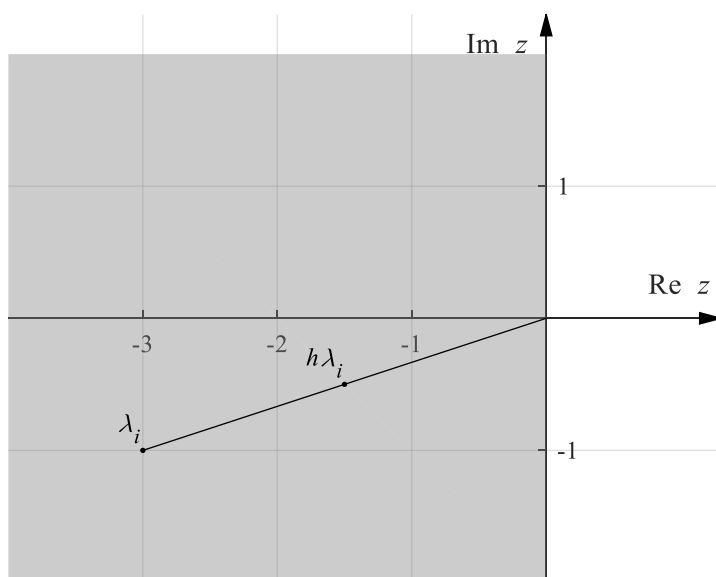
Obszar stabilności określa z jaką maksymalną długością kroku można prowadzić obliczenia przy danym rozkładzie wartości własnych macierzy A . Im dalej na lewo od osi urojonych leży wartość własna, to jest im szybciej zanika składowa rozwiązania równania (9.28), tym mniejsza długość kroku jest konieczna dla zapewnienia stabilności metody. Ta mała długość kroku musi być utrzymana przez cały przedział, na którym rozwiązujemy równanie, nawet wtedy gdy „szybki” składnik przestał już być praktycznie widoczny w rozwiązaniu.

Równania, w rozwiązaniu których występuje szybko zmieniający się składnik, obok składników zmieniających się znacznie wolniej, nazywamy **sztywnymi**. Pomimo, że pierwszy raz pojęcie to pojawiło się w literaturze na początku lat pięćdziesiątych XX wieku, przez ponad 60 lat nie udało się sformułować ścisłej i powszechnie akceptowanej definicji sztywności zagadnienia początkowego. Wciąż przytacza się – jako „najbardziej pragmatyczną »definicję«” sformułowanie Curtissa i Hirschfeldera z 1952 roku: „sztywne równania (różniczkowe) to takie dla których pewne metody niejawne, w szczególności BDF (opisane w podrozdziale 9.8) działają lepiej, zwykle nieporównanie lepiej niż jakiegokolwiek metody jawne”. Aby przewidzieć i wybrać odpowiednią metodę, potrzebna jest jednak „miara sztywności”, którą można wyznaczyć na podstawie równania różniczkowego bez eksperymentowania z wieloma metodami. Tego rodzaju miarą „sztywności” problemu (aczkolwiek niedoskonałą¹¹) w przypadku równania liniowego (9.28) może być $S = \frac{\operatorname{Re}\{\lambda_M\}}{\operatorname{Re}\{\lambda_m\}}$, gdzie λ_M oznacza wartość własną macierzy A leżącą najdalej od osi urojonych w lewej półpłaszczyźnie, a λ_m – najbliższej. W przypadku równania nieliniowego odnosi się to do wartości własnych macierzy Jacobiego.

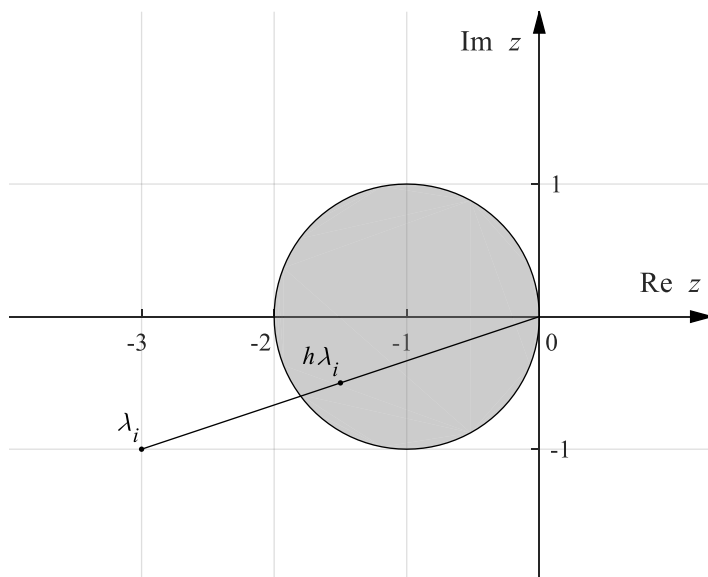
W każdej metodzie o ograniczonym obszarze stabilności długość kroku musi być dobrana do najszybciej zmieniającego się składnika rozwiązania i krok nie może być wydłużony w całym przedziale, w którym wyznaczamy rozwiązanie przybliżone. W przypadku równań sztywnych będzie to szczególnie uciążliwe, bo długość kroku stosowna dla najszybszego składnika może być wielokrotnie mniejsza od długości kroku akceptowanej przy wolniejszych składnikach.

Dużą zaletą schematu różnicowego byłby otwarty na lewą półpłaszczyznę, nieograniczony obszar stabilności absolutnej. Pozwoliłoby to na dobór dowolnej długości kroku bez utraty stabilności – nie byłoby zagrożenia przerwania obliczeń z powodu osiągnięcia ograniczeń stosowanej arytmetyki zmiennopozycyjnej.

¹¹ Odwoływanie się tylko do części rzeczywistej wartości własnych jakobianu jest dyskusyjne, bo nie wyklucza bardzo dużych części urojonych przy małej ujemnej części rzeczywistej, które wiążą się z szybko oscylującą, ale wolno zanikającą składową rozwiązania, która będzie wymagała małego kroku od wszystkich metod.



Rys. 9.3. Lewa półpłaszczyzna płaszczyzny Gaussa, to pożądany obszar stabilności metody (A-stabilność). Jak widać w takim przypadku kwestia stabilności rozwiązania nie nakłada ograniczeń na długość kroku



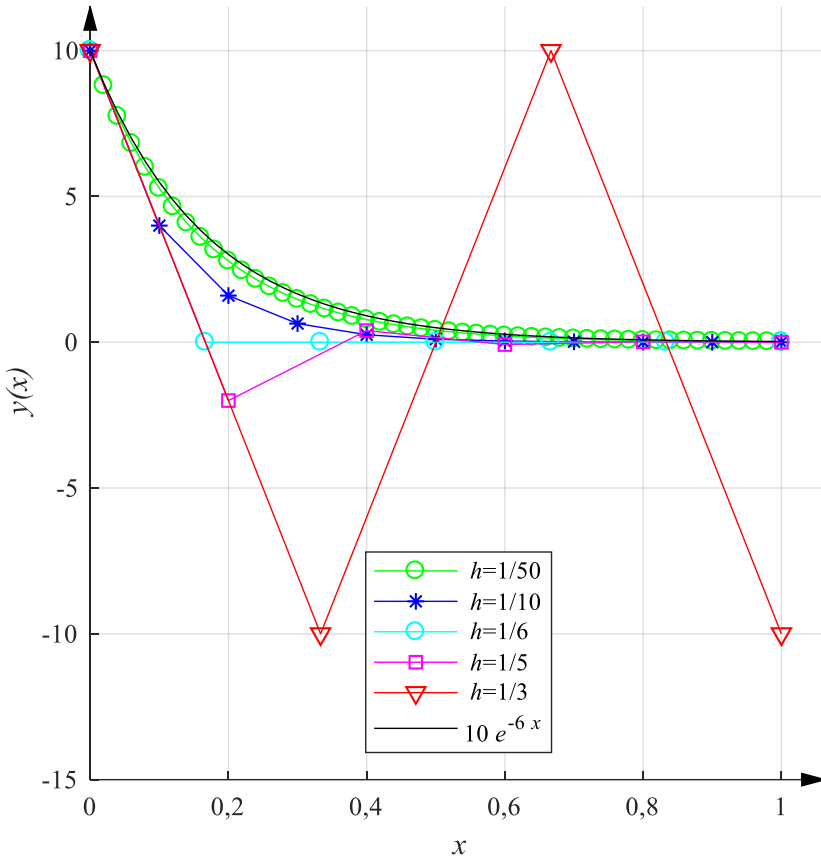
Rys. 9.4. Obszar stabilności absolutnej jawnej metody Eulera. Obszar jest ograniczony, co zmusza do ograniczania długości kroku

Przykład 9.2

Zastosowanie jawnej metody Eulera do równania

$$\dot{y} = -6y, \quad y(0) = 10.$$

Przy długości kroku $h = 1/50, h = 1/10, h = 1/6, h = 1/5, h = 1/3$, na przedziale $[0, 1]$ uzyskano wyniki przedstawione na rysunku 9.5.



Rys. 9.5. Rozwiązania równania $\frac{dy}{dx} = -6y$, $y(0) = 10$ jawną metodą Eulera z różnymi długościami kroku na tle rozwiązania dokładnego $y = 10e^{-6x}$

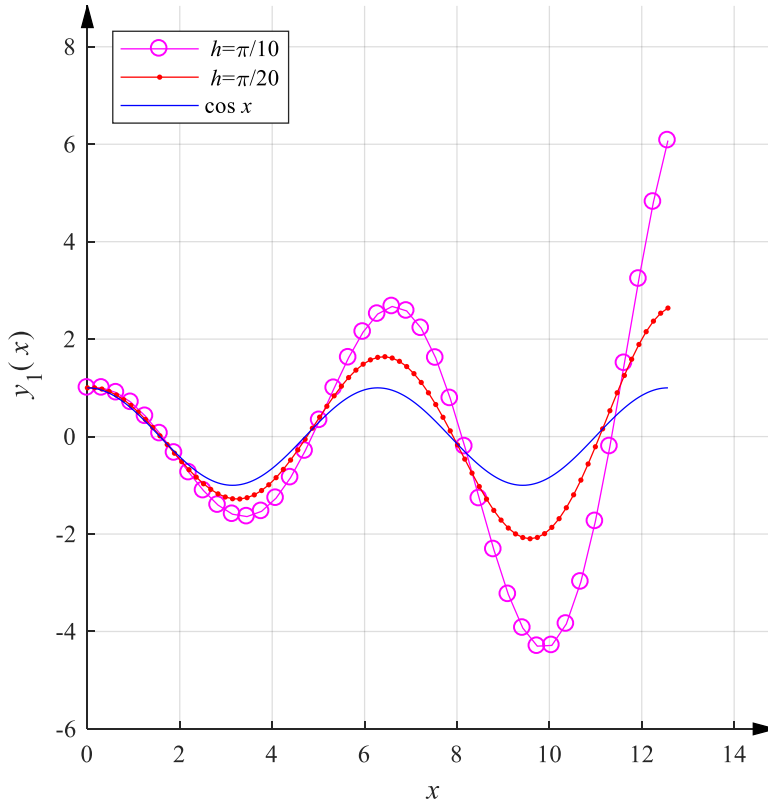
Jak widać pozostawanie iloczynu $-6h$ w obszarze stabilności gwarantuje ograniczoną rozbieżność rozwiązania (dla $h = 1/3$ oscyluje ono między 10 i -10), ale już dla $h = 1/6$ rozwiązanie nie tylko jest niedokładne, ale przestaje oddawać charakter rozwiązania dokładnego.

Przykład 9.3

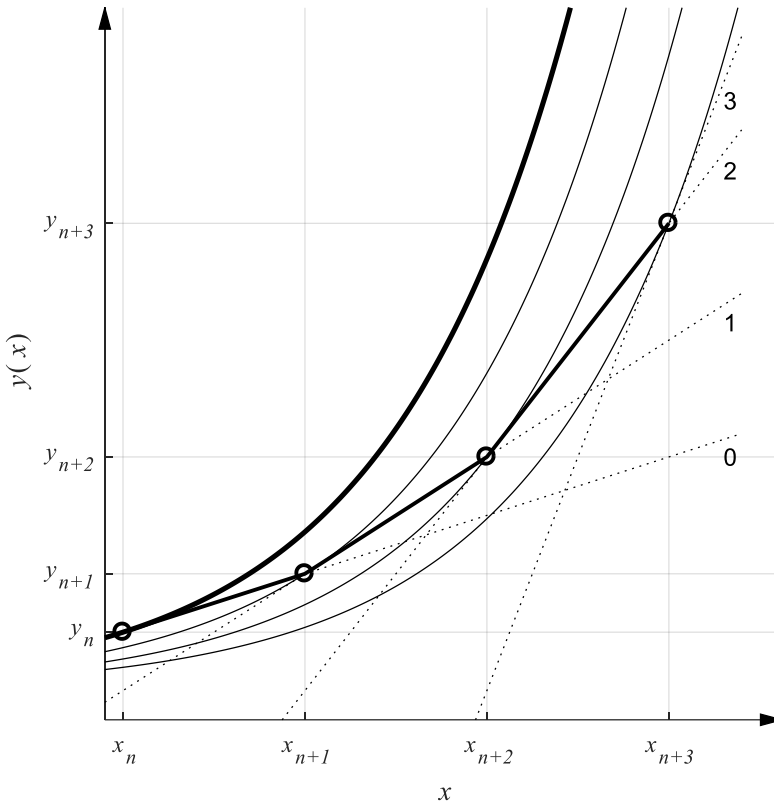
Obszar stabilności absolutnej jawnej metody Eulera ma jedyny punkt wspólny z osią liczb urojonych w zerze, co oznacza, że metodą tą nie da się uzyskać ograniczonych rozwiązań równania oscylatora nietłumionego, czyli układu opisanego równaniami:

$$\begin{cases} \dot{y}_1 = y_2 \\ \dot{y}_2 = -y_1 \end{cases}, \quad y_1(0) = 1, \quad y_2(0) = 0.$$

Wartościami własnymi macierzy współczynników są liczby urojone $\pm j$ i dla żadnej wartości $h \pm jh$ nie znajdzie się w obszarze stabilności. Na rysunku 9.6 zaprezentowano pierwszą współzrzedną rozwiązania numerycznego uzyskanego jawną metodą Eulera w przedziale $[0, 4\pi]$ z krokiem $h = \pi/10$ i $h = \pi/20$ na tle rozwiązania dokładnego $y_1(x) = \cos x$. Jak widać amplituda oscylacji narasta w odróżnieniu od rozwiązania dokładnego.



Rys. 9.6. Rozwiązania równań oscylatora nietłumionego uzyskane jawną metodą Eulera (składowa $y_1(x)$)



Rys. 9.7. Interpretacja geometryczna jawnej metody Eulera. Rozwiązanie dokładne – linia pogrubiona, rozwiązania przybliżone (kółka) uzyskane w kolejnych krokach w kierunku stycznej (linie kropkowane) do rozwiązania lokalnego

Niejawna metoda Eulera

Jeżeli w równaniu różniczkowym (9.2) zastąpimy pochodną ilorazem różnicowym, stosując różnicę wsteczną z punktu (x_{n+1}, y_{n+1}) :

$$\frac{dy}{dx} = f(y(x), x) \Rightarrow \frac{y_n - y_{n+1}}{-h} \approx f(y_{n+1}, x_{n+1}) \quad (9.29)$$

to otrzymamy schemat różnicowy

$$y_{n+1} = y_n + hf(y_{n+1}, x_{n+1}) \quad (9.30)$$

nazywany **niejawną metodą Eulera**.

Metoda niejawna wymaga (w ogólnym przypadku) wyznaczenia y_{n+1} przez rozwiązanie nieliniowego równania algebraicznego (9.30), w każdym kroku schematu różnicowego. W przypadku liniowego układu równań różniczkowych jest to

równanie $y_{n+1} = y_n + hAy_{n+1}$, czyli $(I - hA)y_{n+1} = y_n$, czyli układ liniowych równań algebraicznych.

Podobnie jak jawna metoda Eulera, metoda niejawna jest zgodna i jest metodą pierwszego rzędu. Stosując metodę niejawną, nie poprawiamy więc dokładności metody, ale uzyskujemy znaczne korzyści w zakresie stabilności schematu różnicowego i możliwości doboru długości kroku. Macierz wzmocnienia błędu można wyznaczyć w następującym ciągu przekształceń, wykorzystując, jak poprzednio, rozwinięcie w szereg Taylora. Z zależności

$$\begin{aligned} \frac{-y(x_{n+1}) + y(x_n)}{-h} &= \left. \frac{dy(x)}{dx} \right|_{x=x_{n+1}} + \frac{h}{2!} y''(x_{n+1}) + \dots \\ &= f(y(x_{n+1}), x_{n+1}) + \frac{h}{2!} y''(x_{n+1}) + \dots \end{aligned} \quad (9.31)$$

można wyznaczyć

$$-y_{n+1} - \varepsilon_{n+1} + y_n + \varepsilon_n = -hf(y_{n+1} + \varepsilon_{n+1}, x_{n+1}) - \frac{h^2}{2!} y''(x_{n+1}) - \dots, \quad (9.32)$$

a po podstawieniu $y_{n+1} = y_n + f(y_{n+1}, x_{n+1})h$

$$\begin{aligned} -y_n - hf(y_{n+1}, x_{n+1}) - \varepsilon_{n+1} + y_n + \varepsilon_n \\ = -hf(y_{n+1} + \varepsilon_{n+1}, x_{n+1}) - \frac{h^2}{2!} y''(x_{n+1}) - \dots \end{aligned} \quad (9.33)$$

Stąd można obliczyć

$$\varepsilon_{n+1} = \varepsilon_n + h[f(y_{n+1} + \varepsilon_{n+1}, x_{n+1}) - f(y_{n+1}, x_{n+1})] + O(h^2), \quad (9.34)$$

a po ponownym skorzystaniu z szeregu Taylora

$$\begin{aligned} \left(I - h \left. \frac{\partial f}{\partial y} \right|_{x=x_{n+1}, y=y_{n+1}} \right) \varepsilon_{n+1} &= \varepsilon_n + O(\varepsilon_{n+1}^2) + O(h^2), \\ \varepsilon_{n+1} &= \left(I - h \left. \frac{\partial f}{\partial y} \right|_{x=x_{n+1}, y=y_{n+1}} \right)^{-1} \varepsilon_n + O(\varepsilon_{n+1}^2) + O(h^2). \end{aligned} \quad (9.35)$$

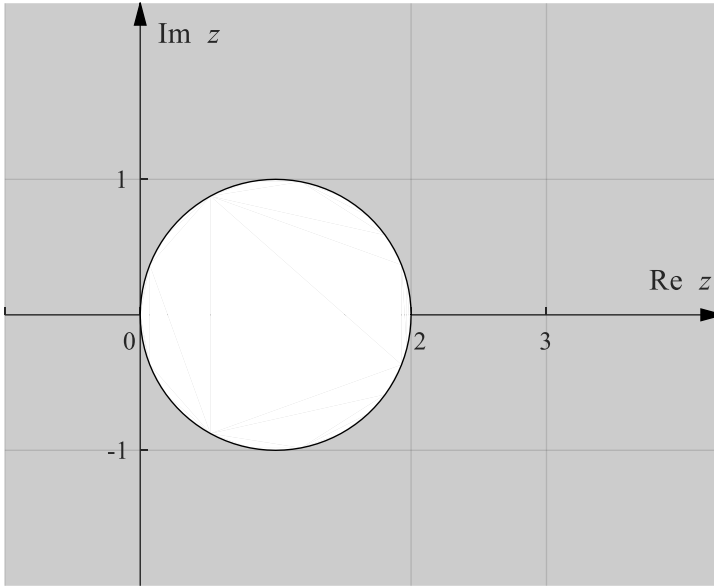
Ostatecznie macierz wzmocnienia błędu to

$$G = \left(I - h \left. \frac{\partial f}{\partial y} \right|_{x=x_{n+1}, y=y_{n+1}} \right)^{-1}. \quad (9.36)$$

Jeżeli rozważymy, jak poprzednio „testowy” układ liniowych równań różniczkowych $\frac{d}{dx}y = Ay$, to

$$G = \left(I - h \left. \frac{\partial f}{\partial y} \right|_{x=x_{n+1}, y=y_{n+1}} \right)^{-1} = [I - hA]^{-1}. \quad (9.37)$$

Jeśli λ_i są wartościami własnymi macierzy A , to wartościami własnymi macierzy G są liczby $\gamma_i = \frac{1}{1-h\lambda_i}$, których moduł będzie mniejszy od 1 ($|\gamma_i| < 1$) wtedy i tylko wtedy, gdy $h\lambda_i$ znajdzie się na zewnątrz okręgu o promieniu 1 i środku $1+j0$ na płaszczyźnie zespolonej. Obszar na zewnątrz okręgu pokazanego na rysunku 9.8 jest jednocześnie obszarem stabilności absolutnej niejawniej metody Eulera.

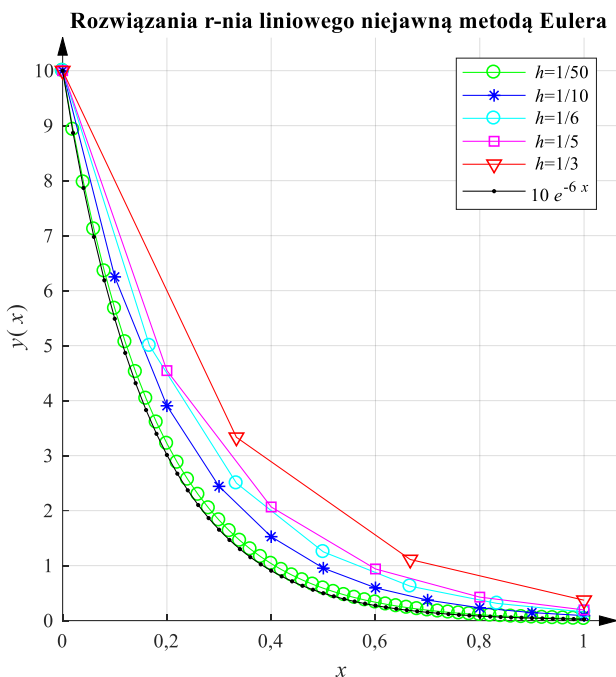


Rys. 9.8. Obszar stabilności absolutnej niejawniej metody Eulera

Jeżeli wszystkie wartości własne λ_i mają ujemne części rzeczywiste, to niejawną metodą Eulera jest stabilna dla dowolnego h . Nie oznacza to oczywiście, że wynik będzie dokładny dla dowolnej długości kroku.

Przykład 9.4

Dla niejawniej metody Eulera powtórzono obliczenia wykonane poprzednio jawną metodą Eulera dla równania $\dot{y} = -6y$ z warunkiem początkowym $y(0) = 10$. Uzyskano wyniki przedstawione na rysunku 9.9.



Rys. 9.9. Rozwiązania równania $\dot{y} = -6y$ niejawną metodą Eulera z różnymi długościami kroku na tle rozwiązania dokładnego $y(x) = 10e^{-6x}$

Jak widać, nawet dla dużych długości kroku charakter rozwiązania jest odtwarzany prawidłowo, chociaż dokładność zdecydowanie się pogarsza. Nie należy oczekiwać, że podobna sytuacja ma miejsce dla dowolnego równania. Podobnie, jak w przypadku metody jawnej rozwiązania numeryczne równania liniowego oscylatora nietłumionego nie odtwarzają charakteru przebiegów dokładnych, ale w tym przypadku uzyskujemy zawsze rozwiązanie zanikające do zera. Nawet w przypadku niektórych układów niestabilnych, przy wybranych długościach kroku uzyskuje się rozwiązania zanikające do zera.

Modyfikacje metody Eulera – metoda punktu środkowego i metoda Heuna

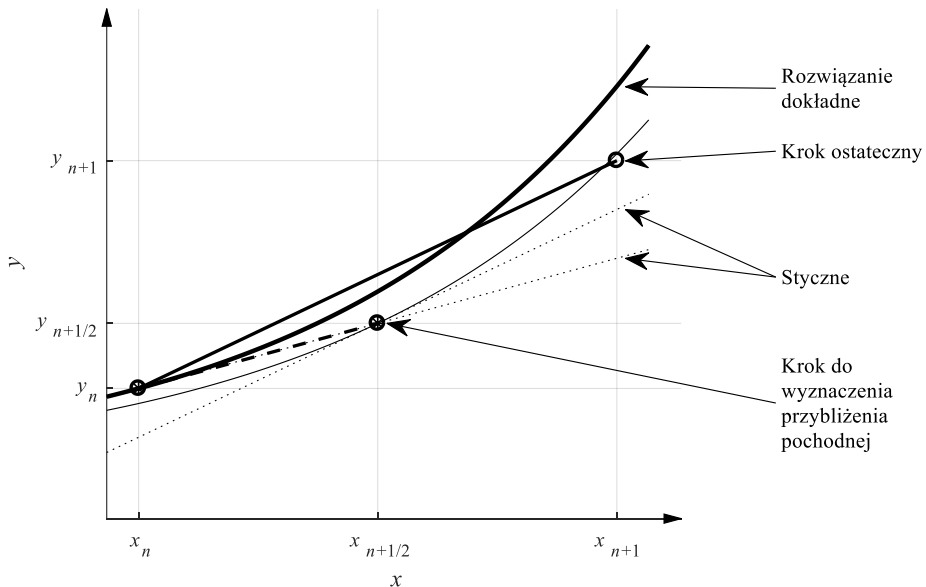
Źródłem błędu metody Eulera jest to, że przyjmuje ona stałą wartość pochodnej rozwiązania (w metodzie jawnej w punkcie początkowym kroku) i wykonuje krok w kierunku wyznaczonym przez tę pochodną (rys. 9.7). W rzeczywistości pochodna $\frac{d}{dx}y$ zmienia się tak, jak zmienia się funkcja $f(y(x), x)$ w trakcie kroku. Aktualizacja wartości pochodnej w innych punktach przedziału $[x_n, x_{n+1}]$ i przyjęcie uśrednionej wartości do ostatecznego kroku dałaby zapewne dokładniejszą metodę. Pomysł ten jest zrealizowany w wielu różnych metodach jednokrokowych. Jedną z nich jest zmodyfikowana jawna metoda Eulera (metoda punktu środkowego). W metodzie tej:

- 1) wyznaczamy pochodną rozwiązania w początkowym punkcie przedziału,
- 2) wykonujemy pół kroku w kierunku określonym przez tę pochodną,
- 3) określamy wartość pochodnej w środku przedziału,
- 4) wykonujemy pełny krok o długości h z punktu x_n, y_n w kierunku wyznaczonym przez zmodyfikowaną, poprawioną pochodną.

Matematycznie można opisać te kroki wzorami:

$$\begin{aligned}
 1) \quad & f_n = f(y_n, x_n) \\
 2) \quad & y_{n+\frac{1}{2}} = y_n + f_n \frac{h}{2} \\
 3) \quad & f_{n+\frac{1}{2}} = f(y_{n+\frac{1}{2}}, x_n + \frac{h}{2}) \\
 4) \quad & y_{n+1} = y_n + f_{n+\frac{1}{2}} h,
 \end{aligned}
 \tag{9.38}$$

a interpretację graficzną podano na rysunku 9.10.



Rys. 9.10. Schemat działania jawnej metody punktu środkowego. Linia pogrubiona – krok w kierunku stycznej do rozwiązania lokalnego w punkcie $(y_{n+\frac{1}{2}}, x_{n+\frac{1}{2}})$

Zmodyfikowana metoda Eulera jest metodą rzędu drugiego, jest więc dokładniejsza od metody Eulera. Jej obszar stabilności absolutnej pokazano na rys. 9.12 dla $r = 2$. Metoda ma wersję niejawną.

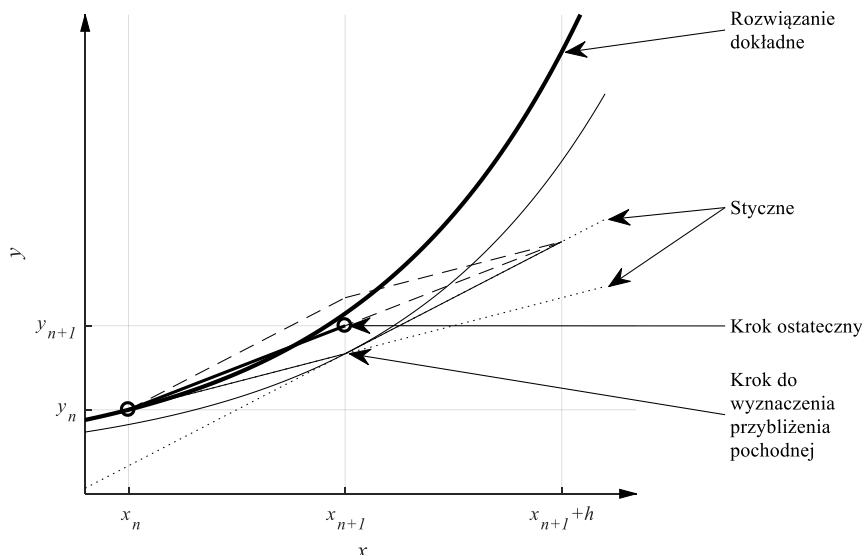
Podobną modyfikacją metody Eulera jest **metoda Heuna** (jawna metoda trapezów). Postępowanie w metodzie Heuna jest następujące:

- 1) obliczamy pochodną rozwiązania w lewym krańcu przedziału,
- 2) wykonujemy krok metody Eulera do prawego krańca przedziału,
- 3) obliczamy pochodną w osiągniętym punkcie na prawym krańcu przedziału,
- 4) wykonujemy jeszcze raz krok metody Eulera do prawego krańca przedziału w kierunku wyznaczonym przez średnią z obu obliczonych pochodnych.

Pokazano to na rysunku 9.11, a matematyczny opis metody Heuna tworzą wzory:

$$\begin{aligned}
 1) \quad & k_1 = f(y_n, x_n), \\
 2) \quad & k_2 = f(y_n + hk_1, x_{n+1}), \\
 3) \quad & y_{n+1} = y_n + h \frac{k_1 + k_2}{2}.
 \end{aligned}
 \tag{9.39}$$

Podobnie jak metoda punktu środkowego metoda Heuna jest drugiego rzędu i można zastosować ją w wersji niejawnej.



Rys. 9.11. Schemat działania metody Heuna. Linia pogrubiona – krok w kierunku „uśrednionym” ze stycznych do rozwiązań lokalnych w punktach (y_n, x_n) i $(y_n + hk_1, x_{n+1})$

9.5. Metody Rungego-Kutty

Jeżeli uśrednienie obliczeń pochodnej z dwóch różnych punktów przynosi poprawę dokładności metody, to czemu by nie uśredniać większej liczby obliczonych

pochodnych rozwiązania? Na tym pomysle opierają się najbardziej popularne metody jednokrokowe rozwiązywania równań różniczkowych – metody Rungego-Kutty.

Ogólną formę jawnego schematu jednokrokowego można zapisać równaniem

$$y_{n+1} = y_n + \Phi_f(y_n, x_n, h)h, \quad (9.40)$$

a w postaci niejawnego równaniem

$$y_{n+1} = y_n + \Phi_f(y_{n+1}, y_n, x_n, h)h, \quad (9.41)$$

przy czym Φ_f (funkcja **przyrostowa lub inkrementalna**) reprezentuje właśnie tę uśrednioną pochodną. Jeżeli uśrednienie jest wykonywane z r wartości, to mówimy o schemacie **r -etapowym**.

W przypadku **r -etapowej metody Rungego-Kutty** ogólne wzory przyjmują postać:

$$\Phi_f(y_n, x_n, h) = \sum_{i=1}^r c_i K_i(y_n, x_n, h), \quad (9.42)$$

$$K_i(y_n, x_n, h) = f\left(y_n + h \sum_{j=1}^r b_{ij} K_j, x_n + \alpha_i h\right), \quad i = 1, 2, \dots, r,$$

gdzie c_i, b_{ij}, α_i są współczynnikami schematu, przy czym $\sum_{i=1}^r c_i = 1$ jest warunkiem zgodności metody. Jeśli schemat jest jawny (otwarty), to $b_{ij} = 0$ dla $j \geq i$:

$$K_1 = f(y, x), \quad K_i(y, x, h) = f\left(y + h \sum_{j=1}^{i-1} b_{ij} K_j, x + h\alpha_i\right), \quad (9.43)$$

$$i = 2, \dots, r.$$

Metoda Eulera może być uważana za jednoetapową metodę Rungego-Kutty, metoda punktu środkowego i metoda Heuna za metodę dwuetapową.

W przypadku czteroetapowego, jawnego schematu Rungego-Kutty wzory opisujące tok obliczeń wyglądają tak:

$$\begin{aligned} K_1 &= f(y_n, x_n), \\ K_2 &= f\left(y_n + \frac{1}{2}K_1, x_n + \frac{1}{2}h\right), \\ K_3 &= f\left(y_n + \frac{1}{2}K_2, x_n + \frac{1}{2}h\right), \\ K_4 &= f(y_n + K_3, x_n + h), \\ y_{n+1} &= y_n + \frac{1}{6}h(K_1 + 2K_2 + 2K_3 + K_4). \end{aligned} \quad (9.44)$$

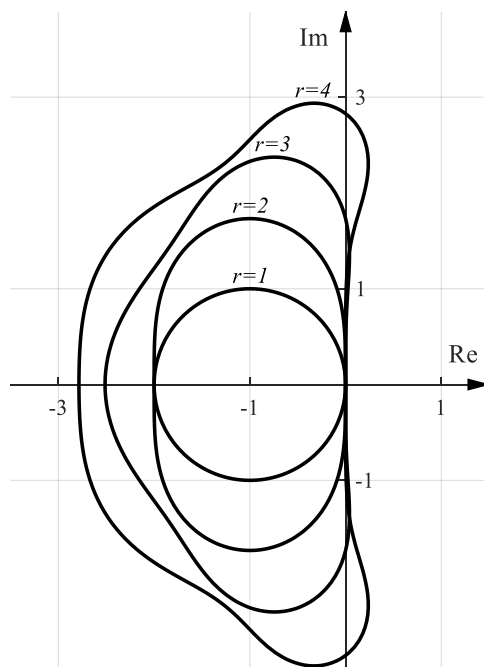
Współczynniki w metodach jednokrokowych nie są przypadkowe, ale dobiera się je tak by zapewnić odpowiedni rząd metody. Rząd jawnej metody Rungego-Kutty nigdy nie jest większy od liczby etapów: $p \leq r$ (tabela 9.1).

Tabela 9.1. Rząd metody Rungo-Kutty w funkcji liczby etapów

Rząd metody RK	1	2	3	4	5	6	7	8
Minimalna liczba etapów	1	2	3	4	6	7	9	11

Początkowo, zwiększenie liczby etapów o 1 podnosi o 1 rząd metody, żeby jednak uzyskać metodę piątego rzędu potrzeba przynajmniej 6 etapów. Z tą kwestią wiążą się twierdzenia, zwane barierami Butchera, stwierdzające, że dla $p \geq 7$ nie istnieją metody rzędu p o $p + 1$ etapach, a dla $p \geq 8$ nie istnieją metody rzędu p o $p + 2$ etapach. Uzasadnia to szczególną popularność jawnych metod czteroetapowych – aby uzyskać metodę rzędu wyższego niż czwarty, trzeba wyliczyć więcej punktów pośrednich niż wynosi rząd metody. Z drugiej strony do bardzo specyficznych zastosowań opracowano 17-etapową metodę rzędu 10.

Obszary stabilności absolutnej jawnych metod Rungego-Kutty pokazano na rysunku 9.12.



Rys. 9.12. Obszary stabilności absolutnej jawnych metod Rungego-Kutty rzędu od 1 do 4

Jawne metody RK mają ograniczony obszar stabilności absolutnej – tylko metody niejawne mogą mieć nieograniczony obszar stabilności.

Można udowodnić, że wszystkie jawne r -etapowe metody rzędu p wymagają takiej samej liczby obliczeń prawej strony równania (wartości funkcji $f(y, x)$), mają taki sam rząd i dla tych samych r i p takie same obszary stabilności. Jednak uzyskiwane wyniki numeryczne są różne dla różnych metod. W ocenie użytkownika jedna z metod może być wygodniejsza w zastosowaniach od innych.

Niejawne metody jednokrokowe są bardziej kosztowne obliczeniowo. Wymagają rozwiązywania nieliniowego układu równań algebraicznych (9.41) w każdym kroku schematu różnicowego. Rozwiązanie dokładne tego równania algebraicznego nie jest zwykle dostępne. Można zastąpić je, tak zwanym **rozwiązaniem samouzgodnionym** (self-consistent), w następujący sposób:

Jeżeli mamy metodę jawną rzędu p opisaną wzorem $y_{n+1} = y_n + \Phi_f(y_n, x_n, h)h$ i analogiczną metodę niejawną: $y_{n+1} = y_n + \Psi_f(y_{n+1}, y_n, x_n, h)h$, to obliczamy:

$$\begin{aligned} y_{n+1}^{pred} &= y_n + \Phi_f(y_n, x_n, h)h, \\ y_{n+1}^{kor} &= y_n + \Psi_f(y_{n+1}^{pred}, y_n, x_n, h)h. \end{aligned} \quad (9.45)$$

Pierwszy ze wzorów (9.45) to tzw. **predyktor**, drugi – **korektor**.

Numeryczne rozwiązanie (9.45) nie jest tym samym, co dokładne rozwiązanie równania (9.41), można je jednak poprawiać metodą iteracyjną – z uwagi na strukturę tego równania może to być metoda iteracji prostej:

$$y_{n+1}^{kor(i+1)} = y_n + \Psi_f(y_{n+1}^{kor(i)}, y_n, x_n, h)h. \quad (9.46)$$

Jeżeli punkt startowy do tych iteracji był dokładny, to znaczy predyktor był metodą tego samego rzędu co korektor, to wystarcza kilka iteracji korektora, żeby uzyskać dobre przybliżenie rozwiązania równania (9.41). Można w ten sposób konstruować grupę metod, znanych jako **metody predyktor-korektor**. Dla skuteczności metody iteracji prostej do rozwiązania równania konieczne jest jednak utrzymywanie małego kroku obliczeń. Jeśli ograniczamy liczbę iteracji korektora do 2-3, zamiast iteracji do zbieżności, to nie są to metody niejawne i nie mają z reguły tak dużych obszarów stabilności jak metody bazujące na dokładnym rozwiązaniu równania (9.41) metodą Newtona czy uproszczoną metodą Newtona.

9.6. Sterowanie długością kroku w metodach jednokrokowych

Każda metoda rozwiązywania równania różniczkowego powinna wykonać jak najmniej obliczeń przy spełnieniu nałożonych wymagań co do dokładności rozwiązania. Zwykle użytkownik chciałby ograniczyć do zadanej wielkości błąd globalny, rozwiązania numerycznego. Twierdzenie 9.4 uzasadnia przekonanie, że można wpływać na błąd globalny kontrolując błąd lokalny. Nie jest to zresztą jedyne

twierdzenie dotyczące tej kwestii. Do schematów jednokrokowych odnosi się jego wersja:

Twierdzenie 9.4 (o zbieżności schematu jednokrokowego – Plato, 2003):

Jeżeli metoda jednokrokowa jest rzędu $p \geq 1$ i funkcja przyrostowa spełnia warunek Lipschitza $|\Phi_f(y, x, h) - \Phi_f(\tilde{y}, x, h)| \leq L_\Phi |y - \tilde{y}|$ dla każdego x w zwartym przedziale, w którym wyznaczane jest rozwiązanie numeryczne, to błąd globalny można oszacować przez $\max_{l=0,1,\dots,n} |y_l - y(x_l)| \leq Ch_{\max}^p$, $h_{\max} = \max_{l=0,1,\dots,n-1} (x_{l+1} - x_l)$, czyli zgodnie z (9.8) błąd globalny jest zbieżny z rzędem p .

W każdym kroku schematu różnicowego można oszacować błąd lokalny i porównać go z narzuconymi wymaganiami. Jeżeli oszacowany błąd jest mały, to można podjąć decyzję o wydłużeniu kolejnego kroku, jeżeli jest duży, to trzeba odrzucić otrzymane przybliżenie i skrócić krok.

Metoda połowienia kroku

Prostą metodą oszacowania błędu jest tak zwana **metoda połowienia kroku**. Załóżmy, że mamy do czynienia ze schematem różnicowym rzędu p . Jeśli wykonamy krok o długości h z punktu $(x_n, y(x_n))$ i otrzymamy punkt $(x_n + h, y_{n+1})$, to błąd lokalny można zapisać w postaci:

$$y(x_n + h) - y_{n+1} = \varphi h^{p+1} + \dots \{\text{wyraży wyższego rzędu}\}, \quad (9.47)$$

gdzie φ jest współczynnikiem rozwinięcia zależnym od punktu $x_n, y(x_n)$. Jeżeli z tego samego punktu wykonamy dwa kroki o długości $h/2$ każdy: $x_n, y(x_n) \xrightarrow{\frac{h}{2}} \xrightarrow{\frac{h}{2}} y_{n+\frac{1}{2}+\frac{1}{2}}$, to otrzymamy inną wartość przybliżoną, oznaczoną przez $y_{n+\frac{1}{2}+\frac{1}{2}}$, a błąd tego przybliżenia można zapisać jako:

$$y(x_n + h) - y_{n+\frac{1}{2}+\frac{1}{2}} = 2\varphi \left(\frac{h}{2}\right)^{p+1} + \dots \{\text{wyraży wyższego rzędu}\}. \quad (9.48)$$

Równania (9.47) i (9.48) pozwalają wyrugować nieznaną wartość dokładną i oszacować błąd lokalny:

$$E_{\text{oszacowany}} = \varphi h^{p+1} \approx \frac{2^p}{2^{p-1}} (y_{n+\frac{1}{2}+\frac{1}{2}} - y_{n+1}). \quad (9.49)$$

Ta metoda szacowania błędu jest dość kosztowna – wykonanie dwu kroków „testujących” wymaga dodatkowych obliczeń wartości funkcji $f(y(x), x)$. Na przykład dla czteroetapowej metody Rungego-Kutty byłoby to 7 dodatkowych

obliczeń prawej strony równania (4 na jeden krok, w tym jedno wspólne dla obu kroków).

Zagnieżdżone metody Rungego-Kutty

Weźmy dwa schematy różnicowe: pierwszy, po wykonaniu kroku, oblicza przybliżenie z_{n+1} i jest rzędu p , drugi – przybliżenie w_{n+1} i jest rzędu $q > p$. Błędy lokalne obu schematów to:

$$\begin{aligned} y(x_n + h) - z_{n+1} &= \varphi h^{p+1} + \dots, \\ y(x_n + h) - w_{n+1} &= \gamma h^{q+1} \dots \end{aligned} \quad (9.50)$$

Po odjęciu od siebie tych równości i przyjęciu $\varphi \approx \gamma$ można wyliczyć oszacowany błąd lokalny

$$E_{oszacowany} = \varphi h^{p+1} \approx \frac{w_{n+1} - z_{n+1}}{(h^{p+1} - h^{q+1})h^{-p-1}} = \frac{w_{n+1} - z_{n+1}}{1 - h^{q-p}}. \quad (9.51)$$

Można wybrać dwa schematy Rungego-Kutty, tak skonstruowane, że różnią się jedynie wagami (współczynniki b_{ij} we wzorze (9.43)), mają takie same punkty pośrednie (współczynniki K_i we wzorze (9.43)) a ich rzędy różnią się o jeden $q = p + 1$.

Wtedy

$$E_{oszacowany} = \varphi h^{p+1} \approx \frac{w_{n+1} - z_{n+1}}{1-h} \approx w_{n+1} - z_{n+1}. \quad (9.52)$$

Na przykład **metoda Rungego-Kutty-Fehlberga** stosuje dwa schematy Rungego-Kutty $m + 1$ i m -etapowy z odpowiednio dobranymi współczynnikami. Schemat m -etapowy jest rzędu $p = m$, a schemat $m + 1$ etapowy jest rzędu $p + 1$. Na przykład algorytm RKF4/5 jest opisany wzorami:

$$\begin{aligned} K_1 &= f(y_n, x_n), \\ K_2 &= f(y_n + \frac{1}{4}K_1, x_n + \frac{1}{4}h), \\ K_3 &= f(y_n + \frac{3}{32}K_2 + \frac{9}{32}K_2, x_n + \frac{3}{8}h), \\ K_4 &= f(y_n + \frac{1932}{2197}K_1 - \frac{7200}{2197}K_2 + \frac{7296}{2197}K_3, x_n + \frac{12}{13}h), \\ K_5 &= f(y_n + \frac{439}{216}K_1 - 8K_2 + \frac{3680}{513}K_3 - \frac{845}{4104}K_4, x_n + h), \\ K_6 &= f(y_n - \frac{8}{27}K_1 + 2K_2 + \frac{3544}{2565}K_3 - \frac{1859}{4104}K_4, x_n - \frac{11}{40}K_5 + \frac{h}{2}), \\ z_{n+1} &= y_n + h \left(\frac{25}{216}K_1 + \frac{1408}{2565}K_3 + \frac{2197}{4104}K_4 - \frac{1}{5}K_5 \right), \\ w_{n+1} &= y_n + h \left(\frac{16}{135}K_1 + \frac{6656}{12825}K_3 + \frac{28561}{56430}K_4 - \frac{9}{50}K_5 + \frac{2}{55}K_6 \right). \end{aligned} \quad (9.53)$$

Sześć obliczeń wartości prawej strony równania wystarcza do oszacowania błędu.

Podobnie działają **metody Shampine’a-Bogackiego** – schemat 3- i 2-etapowy 3(2), **Casha-Karpa** – 5(4), **Dormanda-Prince’a** – 5(4).

Po oszacowaniu błędu należy podjąć decyzję co do długości kroku. Oszacowano błąd dla kroku o długości h i wiadomo, że błąd jest proporcjonalny do h^{p+1} (9.52). Jeżeli zostanie narzucona wartość maksymalnego błędu E_{max} , to chcielibyśmy, żeby w następnym kroku o długości h_{nowy} taki właśnie błąd uzyskać. Z tej proporcjonalności wynika

$$\left(\frac{h}{h_{nowy}}\right)^{p+1} = \frac{|E_{oszacowany}|}{E_{max}} \Rightarrow h_{nowy} = \sqrt[p+1]{\frac{E_{max}}{|E_{oszacowany}|}} h. \quad (9.54)$$

Na przykład dla algorytmu RKF4/5 we wzorze (9.54) będziemy mieli pierwiastek piątego stopnia. Jeżeli błąd oszacowany jest większy od E_{max} , to z (9.54) wynika, że krok trzeba skrócić i powtórzyć obliczenia. Jeśli błąd oszacowany jest mniejszy od E_{max} , to z (9.54) wynika, że następny krok będzie dłuższy. Czasem wprowadza się pewne „współczynniki bezpieczeństwa”, na przykład przyjmuje się, że współczynnik zmiany długości kroku będzie równy:

$$\alpha = \min\left(0,9 \sqrt[p+1]{\frac{E_{max}}{|E_{oszacowany}|}}, 3\right), \quad (9.55)$$

czyli skrócimy krok 10% bardziej niż to wynika ze wzoru (9.54), a wydłużymy co najwyżej trzykrotnie.

Wartość E_{max} występującą w tych relacjach w większości algorytmów ustala użytkownik przez podanie dwóch parametrów ε_{max} i Δ_{max} :

$$E_{max} = y_{odn} \varepsilon_{max} + \Delta_{max}, \quad (9.56)$$

gdzie y_{odn} może być największym modułem dotychczas obliczonego rozwiązania numerycznego, lub np. średnią z modułów dwóch ostatnich wartości.

9.7. Metody wielokrokowe

Jawna metoda Eulera jest metodą jednokrokową, wymagającą tylko jednego obliczenia prawej strony równania w jednym kroku. Wzór opisujący jawną metodę Eulera (9.21) jest liniowy względem argumentów y_n i $f_n = f(y_n, x_n)$. Ale metoda Eulera jest metodą pierwszego rzędu, więc niedokładną. Drogą do uzyskania metody o większej dokładności jest wykorzystanie informacji pochodzącej nie z jednego (jak w metodzie Eulera) a z kilku punktów rozwiązania.

Metody Rungego-Kutty zachowują prostą strukturę metody jednokrokowej i wykorzystują informację z kilku punktów do lepszego przybliżenia pochodnej

rozwiązania przez funkcję przyrostową (inkrementalną) Φ_f . Tracą przy tym liniowość metody Eulera i wymagają większej liczby obliczeń wartości funkcji f w jednym kroku.

Metody wielokrokowe, podobnie jak metoda Eulera, wymagają jednego obliczenia wartości funkcji f w jednym kroku, zachowują liniowość względem f_n , y_n , a poprawę dokładności uzyskują przez to, że do obliczenia przybliżenia y_{n+1} wykorzystują wartości z kilku poprzednich kroków.

Przykład 9.5a

Przykładem jawnej metody wielokrokowej jest schemat różnicowy uzyskany przez zastąpienie pochodnej w równaniu (9.2) przez różnicę centralną (patrz rozdział 4, wzór (4.9)). Otrzymujemy wtedy

$$\begin{aligned} \frac{y(x_{n+1}) - y(x_{n-1}))}{2h} &= f(y(x_n), x_n) \Rightarrow y_{n+1} \\ &= y_{n-1} + 2hf(y_n, x_n). \end{aligned} \quad (9.57)$$

Schemat ten jest zdefiniowany dla $n \geq 1$ i wymaga dwóch wartości y_0 i y_1 do rozpoczęcia działania. Warunek początkowy w (9.2) daje tylko y_0 , więc działanie metody wielokrokowej (dwukrokowej w tym przypadku) trzeba poprzedzić krokiem metody jednokrokowej.

W ogólnej postaci metodę $(k + 1)$ -**krokową** można zapisać przy pomocy równania

$$\begin{aligned} y_{n+1} &= \sum_{j=0}^k a_j y_{n-j} + \\ &h \sum_{j=0}^k b_j f(y_{n-j}, x_{n-j}) + hb_{-1} f(y_{n+1}, x_{n+1}), \end{aligned} \quad (9.58)$$

które można stosować dla $n = k, k + 1, \dots$. W przypadku metody jawnej $b_{-1} = 0$, a zawsze $a_k \neq 0$ i $b_k \neq 0$.

Błąd lokalny metody jest zdefiniowany jako

$$\begin{aligned} r_{n+1}(h) &= y(x_{n+1}) - \sum_{j=0}^k a_j y(x_{n-j}) \\ &+ h \sum_{j=-1}^k b_j f(y_{n-j}, x_{n-j}), \end{aligned} \quad (9.59)$$

i tak jak w definicji 9.2, jeżeli błąd lokalny metody można przedstawić w postaci rozwinięcia w szereg $r_{n+1}(h) = \varphi(y(x_n), x_n)h^{p+1} + O(h^{p+2})$, to mówimy,

że metoda jest rzędu p . Jeżeli $\tau_{n+1}(h) = \frac{r_{n+1}(h)}{h} \rightarrow 0$ dla $h \rightarrow 0$, to mówimy, że metoda jest **zgodna** z zadaniem początkowym, które rozwiązuje.

Zbieżność i zgodność metod wielokrokowych można badać korzystając z rozwinięcia w szereg Taylora rozwiązania y i funkcji f , co daje:

$$\begin{aligned} y(x_{n-j}) &= y(x_n) - jhf(y(x_n), x_n) + O(h^2), \\ f(y(x_{n-j}), x_{n-j}) &= f(y(x_n), x_n) + O(h). \end{aligned} \tag{9.60}$$

Podstawienie tych wyrażeń do (9.59) pozwala udowodnić twierdzenie 9.5.

Twierdzenie 9.5 (o zgodności schematu wielokrokowego – *Quarteroni, 2000*)

Metoda wielokrokowa (9.58) jest zgodna wtedy i tylko wtedy, gdy

$$\sum_{j=0}^k a_j = 1 \text{ i } \sum_{j=-1}^k b_j - \sum_{j=0}^k ja_j = 1. \tag{9.61}$$

Wykorzystanie wyrazów wyższego rzędu w rozwinięciach w szereg Taylora (9.60) pozwala przeprowadzić dowód twierdzenia o zbieżności metody wielokrokowej.

Twierdzenie 9.6 (o zbieżności schematu wielokrokowego – *Quarteroni, 2000*)

Jeżeli rozwiązanie zagadnienia początkowego $y(x)$ ma ciągle pochodne do rzędu $p + 1$ włącznie, metoda wielokrokowa spełnia warunek (9.61), to jest rzędu p wtedy i tylko wtedy, gdy

$$\sum_{j=0}^k (-j)^i a_j + i \sum_{j=-1}^k (-j)^{i-1} b_j = 1 \text{ dla } i = 2, \dots, p. \tag{9.62}$$

Przykład 9.5b

Dla metody (9.57) mamy $k = 1$, $a_0 = 0$, $a_1 = 1$, $b_{-1} = 0$, $b_0 = 2$, $b_1 = 0$. Warunek (9.61) ($0 + 1 = 1$ i $0 + 2 + 0 - (0 + 1) = 1$) jest spełniony, tak jak i warunek (9.62) dla $i = 2$ ($(0 + (-1)^2 \cdot 1) + 2(0 + 0^1 \cdot 2 + 0) = 1$). Metoda (9.57) jest więc metodą drugiego rzędu.

Równanie (9.58), którego rozwiązaniem jest ciąg y_n , $n = k, k + 1, \dots$, można traktować jak równanie różnicowe rzędu $k + 1$. Rozwiązanie takiego równania można wyznaczać krok po kroku lub szukać postaci ogólnej rozwiązania. Służą do tego różne sposoby, na przykład transformata Z, która tak jak transformata Laplace'a zamienia równanie różniczkowe w równanie algebraiczne, sprowadza równanie różnicowe do równania algebraicznego. Do zbadania zachowania roz-

wiązania równania (9.58) można zastosować teorię stabilności równań różnicowych. Wiadomo, że o właściwościach rozwiązania równania (9.58) decyduje w znacznym stopniu tak zwana składowa swobodna, która jest rozwiązaniem równania różnicowego

$$y_{n+1} - \sum_{j=0}^k a_j y_{n-j} = 0. \quad (9.63)$$

Równanie (9.63) można interpretować jako opis działania schematu różnicowego (9.58) rozwiązującego równanie różniczkowe

$$\frac{dy}{dx} = 0, \quad y(0) = 1. \quad (9.64)$$

Równanie (9.63) jest liniowym równaniem różnicowym (z tego powodu mówimy o **liniowych metodach wielokrokowych**). Z teorii liniowych równań różnicowych (np. z zastosowania transformaty Z do rozwiązania równania (9.63)) wiadomo, że rozwiązanie równania (9.63) zależy od pierwiastków jego wielomianu charakterystycznego

$$P_0(z) = z^{k+1} - a_0 z^k - \dots - a_{k-1} z - a_k. \quad (9.65)$$

Wiemy, że pierwiastków tych jest (z uwzględnieniem ich krotności) $k + 1$, że mogą być rzeczywiste lub zespolone parami sprzężone. Warunkiem na ograniczoną rozbieżność rozwiązania równania (9.63), więc i na ograniczoną numeryczną rozbieżność rozwiązania równania (9.64) metodą (9.58), jest, by wszystkie pierwiastki wielomianu $P_0(z)$ leżały w kole jednostkowym na płaszczyźnie zespolonej, a pierwiastki leżące na okręgu były pojedyncze. Warunek ten jest nazywany **warunkiem położenia pierwiastków**, a wielokrokowy schemat różnicowy który go spełnia **zero-stabilnym** (albo D-stabilnym). Warunek położenia pierwiastków nie zależy od długości kroku, a zero-stabilność opisuje poprawne (stabilne) zachowanie schematu różnicowego przy zaburzeniach wartości początkowych i prawej strony równania (9.58).

Przykład 9.5c

Metoda (9.57), w której $k = 1$, $a_0 = 0$, $a_1 = 1$, $b_{-1} = 0$, $b_0 = 2$, $b_1 = 0$, ma wielomian charakterystyczny $P_0(z) = z^2 - a_0 z - a_1 = z^2 - 1$. Jego pierwiastkami są liczby ± 1 , metoda jest więc zero-stabilna. Jeżeli stosujemy ją do rozwiązania równania (9.64) z warunkiem $y_0 = y_1 = 1$, to otrzymamy (zgodnie z algorytmem $y_{n+1} = y_{n-1}$):

$n = 1$	$n = 2$	$n = 3$	$n = 4$...
$y_2 = y_0 = 1$	$y_3 = y_1 = 1$	$y_4 = y_2 = 1$	$y_5 = y_3 = 1$...

Jeśli jednak zaburzymy warunek początkowy, na przykład $y_0 = 1$, $y_1 = 1 + \varepsilon$, to dostajemy:

$n = 1$	$n = 2$	$n = 3$	$n = 4$...
$y_2 = y_0 = 1$	$y_3 = y_1 = 1 + \varepsilon$	$y_4 = y_2 = 1$	$y_5 = y_3 = 1 + \varepsilon$...

Błąd nie jest wzmacniany i rozwiązanie pozostaje ograniczone, ale także błąd nie jest korygowany.

Jeżeli zastosujemy metodę (9.58) do rozwiązania równania testowego

$$\frac{d}{dx}y = \lambda y, \quad y(0) = 1, \quad (9.66)$$

gdzie λ może być liczbą zespoloną, to otrzymujemy równanie

$$y_{n+1} = \sum_{j=0}^k a_j y_{n-j} + h\lambda \sum_{j=-1}^k b_j y_{n-j}, \quad (9.67)$$

a więc także liniowe, jednorodne równanie różnicowe

$$(1 - h\lambda b_{-1})y_{n+1} - \sum_{j=0}^k (a_j + h\lambda b_j)y_{n-j} = 0. \quad (9.68)$$

Jego wielomianem charakterystycznym jest

$$P_1(z) = (1 - h\lambda b_{-1})z^{k+1} - (a_0 + h\lambda b_0)z^k - \dots - (a_{k-1} + h\lambda b_{k-1})z - (a_k + h\lambda b_k). \quad (9.69)$$

Współczynniki wielomianu $P_1(z)$ zależą od $\mu := h\lambda$, więc i jego pierwiastki zależą od μ . Podobnie jak poprzednio, rozwiązanie równania (9.67) jest ograniczone, więc i ograniczone jest numeryczne rozwiązanie równania (9.66) metodą (9.58), jeśli wszystkie pierwiastki wielomianu $P_1(z)$ leżały w kole jednostkowym na płaszczyźnie zespolonej, a pierwiastki leżące na okręgu były pojedyncze. Zbiór S tych $\mu = h\lambda$ na płaszczyźnie zespolonej, dla których warunek położenia pierwiastków wielomianu $P_1(z)$ jest spełniony nazywamy **obszarem stabilności absolutnej** wielokrokowej metody (9.58). Jeżeli lewa półpłaszczyzna zmiennej zespolonej jest zawarta w zbiorze S , to metodę nazywamy **A-stabilną**, a jeśli sektor kątowy lewej półpłaszczyzny, to **A(α)-stabilną**.

Jeżeli $\lambda = 0$, to równanie (9.67) przechodzi w (9.63), a wielomian $P_1(z)$ w $P_0(z)$, więc zero-stabilność jest równoważna temu, że $0 \in S$.

Przykład 9.5d

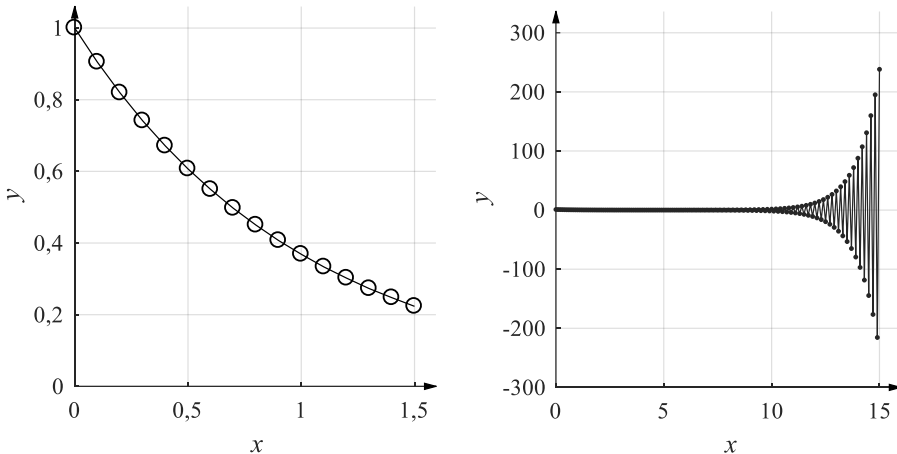
W przypadku metody (9.57), w której $k = 1$, $a_0 = 0, a_1 = 1, b_{-1} = 0, b_0 = 2, b_1 = 0$, wielomian $P_1(z) = (1 - h\lambda b_{-1})z^2 - (a_0 + h\lambda b_0)z - (a_1 + h\lambda b_1) = z^2 - 2h\lambda z - 1 = z^2 - 2\mu z - 1$. Ze wzorów Vieté'a wynika, że iloczyn jego pierwiastków wynosi -1 , więc albo jeden z nich ma moduł większy od 1, albo obydwa są liczbami zespolonymi o module 1, których iloczyn równy jest -1 a suma 2μ , czyli dla pewnych rzeczywistych a i b , takich że $a^2 + b^2 = 1$ zachodzi $z_1 = a + jb$, $z_2 = \frac{-1}{a+jb} = -a + jb$, $2\mu = z_1 + z_2 = 2jb \Rightarrow \mu = h\lambda = jb, -1 \leq b \leq 1$.

Obszar stabilności jest więc odcinkiem na osi urojonych. Obszar stabilności o pustym wnętrzu oznacza, że dla dowolnie małego $h > 0$, przy dowolnym $\lambda \neq 0$ rozwiązanie numeryczne będzie niestabilne. Schemat różnicowy (9.57) rozwiązujący równanie (9.66) ma postać

$$(1 - h\lambda b_{-1})y_{n+1} - \sum_{j=0}^k (a_j + h\lambda b_j)y_{n-j} = 0$$

$$y_{n+1} = (a_0 + h\lambda b_0)y_n + (a_1 + h\lambda b_1)y_{n-1}, \quad y_{n+1} = 2h\lambda y_n + y_{n-1}.$$

Wykonane według niego obliczenia dla $h = 0,1$, $\lambda = -1$ (pierwiastki $P_1(z)$: $-1,1050$ i $0,9050$), przyjmując $y_0 = y(0) = 1$ i $y_1 = y(0,1) = e^{-0,1}$ (tzn. wartość dokładnego rozwiązania po jednym kroku), dają rozwiązanie pokazane na rysunku 9.13.



Rys. 9.13. Pierwszych 15 (po lewej) i 150 kroków (po prawej) schematu (9.57) z krokiem $h = 0,1$ dla równania $\frac{dy}{dx} = -y$, $y(0) = 1$. Jak widać, początkowe kroki odtwarzają dobrze rozwiązanie dokładne, ale w długim przedziale pojawia się oscylujące, nieograniczone rozwiązanie pasożytnicze

Analiza stabilności metod wielokrokowych doprowadziła do udowodnienia twierdzeń określających ich możliwości zwanych „barierami Dahlquista”.

Twierdzenie 9.7 (pierwsza i druga bariera Dahlquista – *Quarteroni, 2000*)

1. Nie istnieje zero-stabilna liniowa metoda k -krokowa rzędu wyższego niż $k + 1$, jeśli k jest nieparzyste, a $k + 2$ jeśli k jest parzyste.
2. Liniowa, jawna metoda wielokrokowa nie może być ani A -stabilna, ani $A(\alpha)$ -stabilna.
3. Nie istnieje A -stabilna liniowa metoda wielokrokowa rzędu wyższego niż 2.

9.8. Metody Adamsa

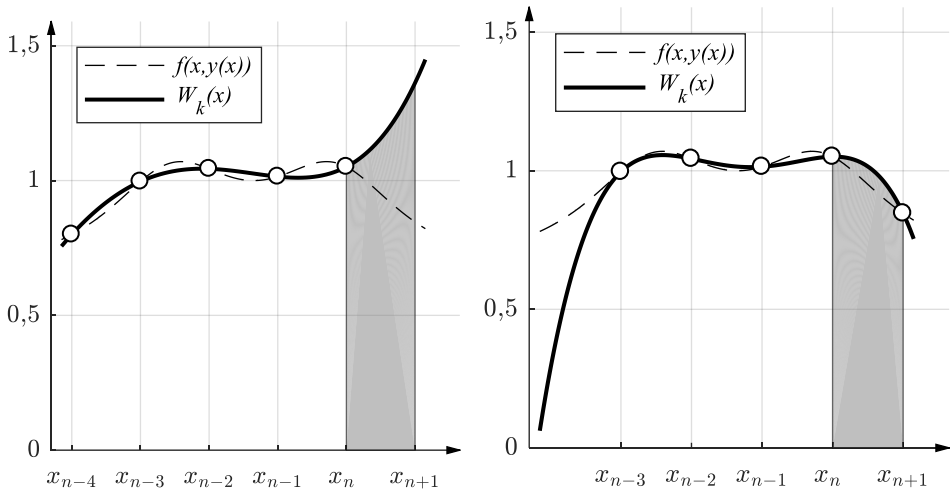
Każde zagadnienie początkowe można przedstawić w postaci równoważnego równania całkowego. Jeżeli punktem początkowym jest $x_n, y(x_n)$, to całkując równanie $\frac{dy}{dx} = f(y(x), x)$ na przedziale $[x_n, x_{n+1}]$, otrzymuje się

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(y(x), x) dx, \quad (9.70)$$

co dałoby schemat różnicowy

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} f(y(x), x) dx. \quad (9.71)$$

Funkcję $f(y(x), x)$ można przybliżyć wielomianem interpolacyjnym (który potrafimy scałkować), umieszczając węzły w punktach rozwiązania numerycznego równania różniczkowego. Tak skonstruowane schematy różnicowe noszą nazwę **metod Adamsa**. Jeżeli wykorzystamy węzeł x_{n+1}, y_{n+1} otrzymamy metodę niejawną (**Adamsa-Moultona**), jeśli tylko węzły poprzednie – metodę jawną (**Adamsa-Bashfortha**), ale ponieważ wykorzystuje się wtedy wielomian interpolacyjny do ekstrapolacji, więc nie należy oczekiwać nadzwyczajnych właściwości metod jawnych. Konstrukcję metod Adamsa zobrazowano na rysunku 9.14.



Rys. 9.14. Konstrukcja wielomianu w metodach Adamsa-Bashforta (po lewej) i Adamsa-Moultona (po prawej)

Całka z wielomianu interpolacyjnego zależy tylko od wartości funkcji $f(y, x)$ w węzłach, czyli punktach rozwiązania numerycznego. Wzory opisujące metody Adamsa wykorzystujące wielomian interpolacyjny stopnia $k - 1$, można zapisać w postaci:

$$y_{n+1} = y_n + h \sum_{j=1}^k b_j f(y_{n-j+1}, x_{n-j+1}), \quad (9.72)$$

dla metody jawnej oraz

$$y_{n+1} = y_n + h \sum_{j=0}^{k-1} \beta_j f(y_{n-j+1}, x_{n-j+1}) \quad (9.73)$$

dla metody niejawnej. Współczynniki w tych wzorach zależą tylko od stopnia wielomianu, można je obliczyć i stabilizować. W przypadku metod jawnych wykorzystujących k węzłów (a więc wielomian interpolacyjny stopnia $k - 1$) uzyskuje się metodę k -krokową. Odmienna sytuacja występuje dla metod niejawnych, gdzie dla $k = 1$ i $k = 2$ uzyskuje się odpowiednio niejawny wzór Eulera i niejawny wzór trapezów, a więc metody jednokrokowe. Dla pozostałych k , metoda Adamsa-Moultona wykorzystująca k węzłów (a więc wielomian interpolacyjny stopnia $k - 1$) jest $(k - 1)$ -krokowa.

Na przykład dla metody jawnej dla $k = 3$ węzłami są punkty x_n, x_{n-1}, x_{n-2} , każdy odległy o h od sąsiedniego. Wielomianem interpolacyjnym jest (przy oznaczeniu $f_n = f(y_n, x_n)$):

$$\begin{aligned}
 P(x) &= f_n \frac{(x - x_{n-1})(x - x_{n-2})}{(x_n - x_{n-1})(x_n - x_{n-2})} \\
 &\quad + f_{n-1} \frac{(x - x_n)(x - x_{n-2})}{(x_{n-1} - x_n)(x_{n-1} - x_{n-2})} \\
 &\quad + f_{n-2} \frac{(x - x_n)(x - x_{n-1})}{(x_{n-2} - x_n)(x_{n-2} - x_{n-1})} = \quad (9.74) \\
 &= f_n \frac{1}{2h^2} (x - x_n + h)(x - x_n + 2h) - f_{n-1} \frac{1}{h^2} (x - x_n)(x - \\
 &\quad x_n + 2h) + f_{n-2} \frac{1}{2h^2} (x - x_n)(x - x_n + h).
 \end{aligned}$$

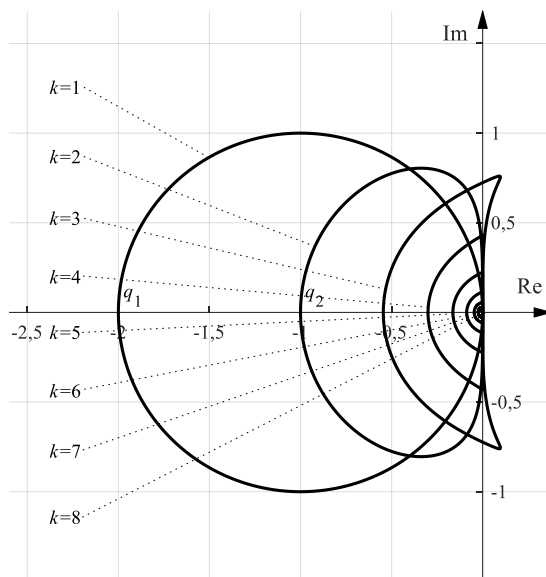
Po scałkowaniu tego wielomianu i porównaniu z (9.72), otrzymujemy

$$\begin{aligned}
 hb_1 &= \frac{1}{2h^2} \int_{x_n}^{x_n+h} (x - x_n + h)(x - x_n + 2h) dx = \frac{1}{2h^2} \frac{23}{6} h^3 = \frac{23}{12} h, \\
 hb_2 &= -\frac{1}{h^2} \int_{x_n}^{x_n+h} (x - x_n)(x - x_n + 2h) dx = -\frac{1}{h^2} \frac{4}{3} h^3 = -\frac{4}{3} h, \quad (9.75) \\
 hb_3 &= \frac{1}{2h^2} \int_{x_n}^{x_n+h} (x - x_n)(x - x_n + h) dx = \frac{1}{2h^2} \frac{5}{6} h^3 = \frac{5}{12} h,
 \end{aligned}$$

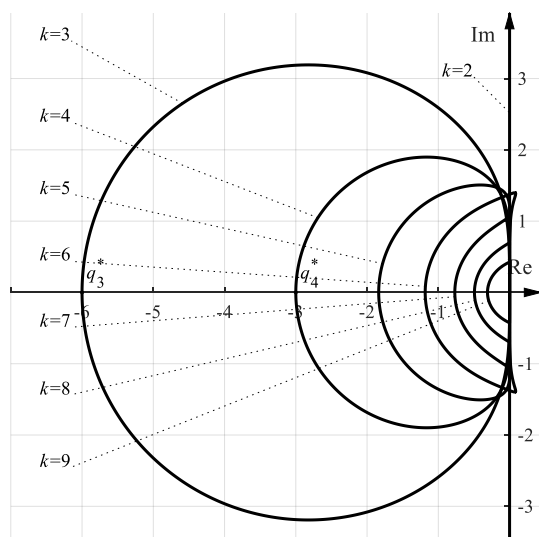
czyli $y_{n+1} = y_n + h \left[\frac{23}{12} f(y_n, x_n) - \frac{4}{3} f(y_{n-1}, x_{n-1}) + \frac{5}{12} f(y_{n-2}, x_{n-2}) \right]$.

Metody Adamsa są zgodne. Rząd k -krokowej metody Adamsa (jawnej lub niejawnej) jest równy k .

Obszary stabilności absolutnej metod Adamsa pokazano na rysunkach 9.15 i 9.16.



Rys. 9.15. Obszary stabilności absolutnej metod Adamsa-Bashfорта (jawnych). Obszarem stabilności jest wnętrze konturu



Rys. 9.16. Obszary stabilności absolutnej metod Adamsa-Moultona (niejawnych). Obszarem stabilności jest wnętrze konturu, poza przypadkiem metody Adamsa-Moultona dla $k = 2$, która jest identyczna z niejawnym wzorem trapezów, więc jest A -stabilna (jej obszar stabilności jest całą lewą półpłaszczyzną). Rysunek nie uwzględnia metody Adamsa-Moultona dla $k = 1$, która jest tożsama z niejawnym wzorem Eulera, którego obszar stabilności absolutnej jest przedstawiony na rysunku 9.8

Tabela 9.2. Lewy skrajny punkt obszarów stabilności metod Adamsa

Liczba węzłów k	1	2	3	4	5	6	7	8
q_k Adamsa-Bashfortha	-2	-1	$-\frac{6}{11}$	$-\frac{3}{10}$	$-\frac{90}{551}$	$-\frac{5}{57}$	$-\frac{1890}{40633}$	$-\frac{945}{38716}$
q_k^* Adamsa-Moultona	$-\infty$	$-\infty$	-6	-3	$-\frac{90}{49}$	$-\frac{45}{38}$	$-\frac{1890}{2459}$	$-\frac{35}{71}$

Obszar stabilności absolutnej metod jawnych jest wyraźnie mniejszy niż niejawnych, a ze wzrostem rzędu metody dość szybko maleje. W konsekwencji metody Adamsa-Bashfortha (tzn. jawne) wykorzystywane są głównie pomocniczo do obliczania punktu startowego dla wariantu niejawnego (Adamsa-Moultona). Przedstawione obszary stabilności metod niejawnych są obliczone przy założeniu dokładnego rozwiązania równania nieliniowego w każdym kroku. W praktyce, podobnie jak to pokazano przy metodach jednokrokowych, metody Adamsa mogą być stosowane w schemacie predyktor-korektor. Predyktorem jest metoda jawna, a korektorem metoda niejawna o tym samym rzędzie. Wykonuje się kilka (co najwyżej 3-4) iteracji poprawiających przybliżenie otrzymane z metody jawnej. Metody takie nie są metodami niejawnymi – nie wykorzystują dokładnego rozwiązania nieliniowego równania algebraicznego, które pojawia się w metodzie niejawnej, a zadowolają się rozwiązaniem przybliżonym. Obszar stabilności absolutnej metody predyktor-korektor jest pośredniego rozmiaru między obszarami składających się na taką metodę metody jawnej i niejawnej, przy czym szczegóły zależą od liczby kroków korekcji. Pomimo, że dwie metody Adamsa-Moultona są A-stabilne realizacja w formie metod predyktor-korektor z rozwiązywaniem równania nieliniowego metodą iteracji prostej nie prowadzi do metod o nieograniczonym obszarze stabilności absolutnej.

Współczynniki metod Adamsa są, jak w (9.75), wyprowadzone przy założeniu, że długości kroków są jednakowe. Sterowanie długością kroku w metodach Adamsa wymaga wyprowadzenia oddzielnych wzorów, uwzględniających zależność współczynników od długości wszystkich wykorzystanych kroków i oszacowania błędu w ostatnim kroku. Jest to możliwe, ale znacznie bardziej skomplikowane niż w przypadku metod jednokrokowych. Ponadto dla stabilności potrzeba, aby długość kroku nie zmieniała się zbyt często.

9.9. Metody wstecznego różniczkowania

W metodach **wstecznego różniczkowania (BDF – backward differentiation formula)** przybliżamy rozwiązanie $y(x)$ równania różniczkowego wielomianem interpolacyjnym (stopnia k) $W(x)$ zbudowanym na węzłach x_{n-i}, y_{n-i} , gdzie $i = -1, 0, 1, \dots, k-1$. Mamy tu dla $i = -1$ węzeł (x_{n+1}, y_{n+1}) z poszukiwaną wartością y_{n+1} oraz k węzłów, w których wartości $y_n, y_{n-1}, \dots, y_{n-k+1}$ są znane. Następnie obliczana jest pochodna tego wielomianu w węźle x_n dla metod jawnych, a w węźle x_{n+1} dla niejawnych.

W przypadku metody jawnej wykorzystanie równości

$$y'(x_n) = f(y_n, x_n) = W'(x_n) \quad (9.76)$$

pozwala na jawne wyliczenie y_{n+1} i prowadzi do wzorów postaci

$$y_{n+1} = \sum_{j=0}^{k-1} \tilde{\alpha}_j y_{n-j} + h\tilde{\beta}_0 f(y_n, x_n). \quad (9.77)$$

Dla $k = 1$ uzyskuje się jawny wzór Eulera. Dla $k = 2$ wielomianem interpolacyjnym zbudowanym na węzłach $(x_{n+1}, y_{n+1}), (x_n, y_n), (x_{n-1}, y_{n-1})$ jest

$$\begin{aligned} W(x) = & y_n \frac{x - x_{n+1}}{x_n - x_{n+1}} \frac{x - x_{n-1}}{x_n - x_{n-1}} \\ & + y_{n+1} \frac{x - x_{n-1}}{x_{n+1} - x_n} \frac{x - x_{n-1}}{x_{n+1} - x_{n-1}} \\ & + y_{n-1} \frac{x - x_{n+1}}{x_{n-1} - x_n} \frac{x - x_{n+1}}{x_{n-1} - x_{n+1}}, \end{aligned} \quad (9.78)$$

co po uwzględnieniu, że $x_{n+1} - x_n = h = x_n - x_{n-1}, x_{n+1} - x_{n-1} = 2h$, daje

$$\begin{aligned} W(x) = & -\frac{y_n}{h^2} (x - x_{n+1})(x - x_{n-1}) \\ & + \frac{y_{n+1}}{2h^2} (x - x_n)(x - x_{n-1}) \\ & + \frac{y_{n-1}}{2h^2} (x - x_n)(x - x_{n+1}). \end{aligned} \quad (9.79)$$

Pochodną tego wielomianu jest

$$\begin{aligned} W'(x) = & -\frac{y_n}{h^2} (2x - x_{n+1} - x_{n-1}) + \frac{y_{n+1}}{2h^2} (2x - x_n - x_{n-1}) + \\ & \frac{y_{n-1}}{2h^2} (2x - x_n - x_{n+1}), \end{aligned} \quad (9.80)$$

skąd, po wykorzystaniu warunku (9.76) i uwzględnieniu, że sąsiednie węzły są odległe o h od siebie, otrzymujemy

$$W'(x_n) = \frac{y_{n+1}}{2h} - \frac{y_{n-1}}{2h} = f(y_n, x_n). \quad (9.81)$$

Jest to równanie kroku od x_{n-1} do x_{n+1} jawnej metody punktu środkowego, przy czym punktem środkowym jest x_n (porównaj z (9.38), (9.57)):

$$y_{n+1} = y_{n-1} + 2hf(y_n, x_n). \quad (9.82)$$

Dla $k = 3$ po analogicznym wyprowadzeniu otrzymuje się metodę opisaną równaniem

$$\frac{1}{3}y_{n+1} + \frac{1}{2}y_n - y_{n-1} + \frac{1}{6}y_{n-2} = hf(y_n, x_n),$$

która jest niestabilna. Także pozostałe ($k > 3$) metody jawne są niestabilne, więc bezużyteczne. Wartościowymi metodami wstecznego różniczkowania są metody niejawne. W przypadku metody niejawnej zastosowanie warunku

$$y'(x_{n+1}) = f(y_{n+1}, x_{n+1}) = W'(x_{n+1}) \quad (9.83)$$

daje wzór postaci

$$y_{n+1} = \sum_{j=0}^{k-1} \alpha_j y_{n-j} + h\beta_{-1}f(y_{n+1}, x_{n+1}) \quad (9.84)$$

czyli nieliniowe równanie algebraiczne względem y_{n+1} .

Na przykład, jeżeli wybierzemy do interpolacji dwa punkty odległe o h ($k = 1$), to wielomianem interpolacyjnym jest

$$W(x) = y_n \frac{x - x_{n+1}}{x_n - x_{n+1}} + y_{n+1} \frac{x - x_n}{x_{n+1} - x_n}. \quad (9.85)$$

Po obliczeniu jego pochodnej i wykorzystaniu warunku (9.70), dostajemy

$$W'(x) = \frac{y_n}{x_n - x_{n+1}} + \frac{y_{n+1}}{x_{n+1} - x_n} = \frac{1}{h}(y_{n+1} - y_n), \quad (9.86)$$

czyli

$$\frac{1}{h}(y_{n+1} - y_n) = f(y_{n+1}, x_{n+1}) \Rightarrow y_{n+1} = y_n + hf(y_{n+1}, x_{n+1}). \quad (9.87)$$

Dwupunktowa metoda wstecznego różniczkowania jest więc tożsama z niejawną metodą Eulera. W podobny sposób można wyprowadzić kolejne wzory. Dla metody 3-punktowej ($k = 2$) otrzymujemy to samo, co w przypadku jawnym, wyrażenie na wielomian $W(x)$ (9.79) i jego pochodną $W'(x)$ (9.80), ale odmienny warunek $W'(x_{n+1}) = f(y_{n+1}, x_{n+1})$, prowadzi do

$$\begin{aligned}
 W'(x)|_{x=x_{n+1}} &= \left(-\frac{y_n}{h^2}(2x - x_{n+1} - x_{n-1}) + \frac{y_{n+1}}{2h^2}(2x - x_n - x_{n-1}) \right. \\
 &\quad \left. + \frac{y_{n-1}}{2h^2}(2x - x_n - x_{n+1}) \right) \Big|_{x=x_{n+1}} \\
 &= -2\frac{y_n}{h} + \frac{3}{2}\frac{y_{n+1}}{h} + \frac{1}{2}\frac{y_{n-1}}{h} = f(y_{n+1}, x_{n+1}),
 \end{aligned} \tag{9.88}$$

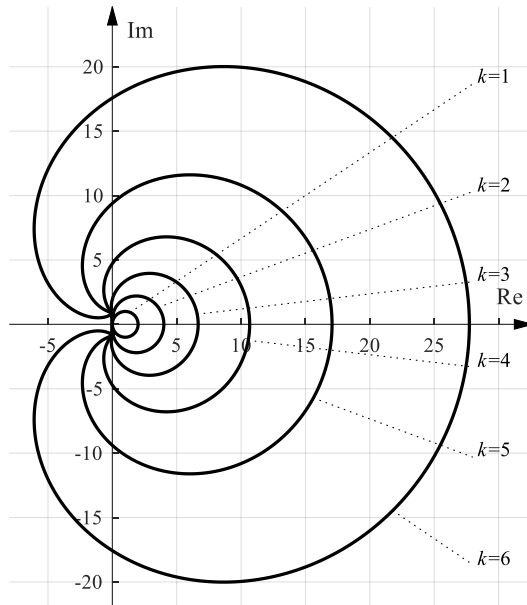
czyli

$$\frac{3}{2}y_{n+1} - 2y_n + \frac{1}{2}y_{n-1} = hf(y_{n+1}, x_{n+1}). \tag{9.89}$$

Ostatecznie dostaje się:

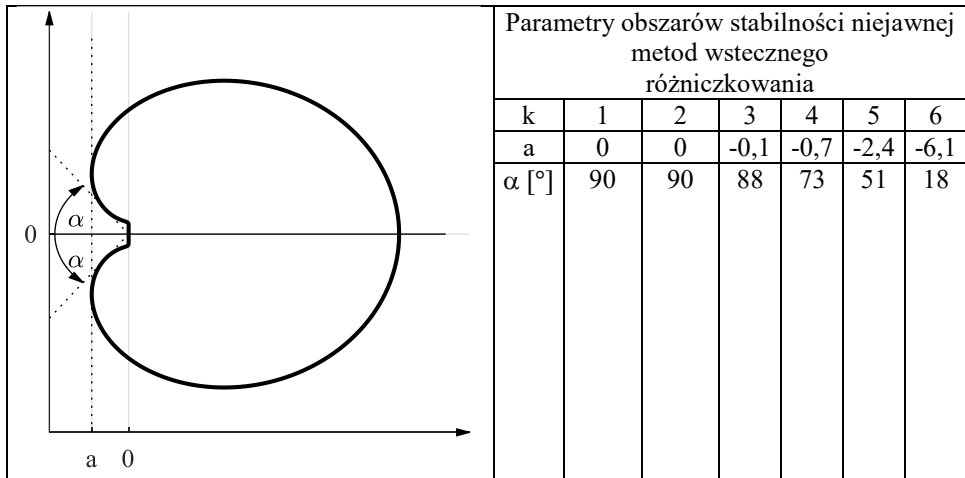
$$y_{n+1} = \frac{4}{3}y_n - \frac{1}{3}y_{n-1} + \frac{2}{3}hf(y_{n+1}, x_{n+1}). \tag{9.90}$$

Niejawna metoda wstecznego różniczkowania wykorzystująca $k + 1$ węzłów jest metodą rzędu k . Obszary stabilności absolutnej pokazano na rysunku 9.17. Dla $k > 6$ niejawne metody wstecznego różniczkowania są niestabilne.



Rys. 9.17. Obszary stabilności absolutnej niejawnych metod wstecznego różniczkowania. Obszarem stabilności jest **zewnątrz** wykreślonego dla danego k konturu

Tabela 9.3. Parametry obszarów stabilności niejawniej metod wstecznego różniczkowania



Jak widać z tabeli 9.3 A-stabilne są tylko metody niskiego rzędu, a kąt α metody rzędu 6 jest bardzo mały.

Stosowanie metod wstecznego różniczkowania ze zmienną długością kroku jest kłopotliwe – wymaga ponownego wyprowadzenia wzorów uwzględniających różne długości poszczególnych kroków.

Koncepcję z metod wstecznego różniczkowania wykorzystują metody **numerycznego różniczkowania** (NDF – **numerical differentiation formula**). Jeśli oznaczymy różnicę wsteczną

$$\nabla y_n = y_n - y_{n-1}, \quad \nabla^2 y_n = \nabla y_n - \nabla y_{n-1}, \dots \quad (9.91)$$

to zamiast równania (9.83) można zapisać:

$$\sum_{m=1}^k \frac{1}{m} \nabla^m y_{n+1} = hf(y_{n+1}, x_{n+1}). \quad (9.92)$$

Składnik $\frac{1}{h} \sum_{m=1}^k \frac{1}{m} \nabla^m y_{n+1}$ odpowiada pochodnej $W'(x_{n+1})$ wielomianu interpolacyjnego. Faktycznie, np. dla $k = 2$ otrzymujemy $\sum_{m=1}^2 \frac{1}{m} \nabla^m y_{n+1} = \frac{1}{2} \nabla^2 y_{n+1} + \nabla y_{n+1} = \frac{1}{2} (\nabla y_{n+1} - \nabla y_n) + \nabla y_{n+1} = \frac{3}{2} (y_{n+1} - y_n) - \frac{1}{2} (y_n - y_{n-1}) = \frac{3}{2} y_{n+1} - 2y_n + \frac{1}{2} y_{n-1}$, dokładnie jak w równaniu (9.89).

Równanie (9.92) rozwiązuje się iteracyjnie uproszczoną metodą Newtona (pochodna jest obliczana tylko w pierwszej iteracji), wychodząc od warunku początkowego

$$y_{n+1}^{(0)} = \sum_{m=1}^k \nabla^m y_n. \quad (9.93)$$

Ponieważ we wzorze (9.93) wykorzystuje się i tak o 1 więcej wartości rozwiązania numerycznego w poprzedzających krokach niż we wzorze (9.92), to w latach siedemdziesiątych ubiegłego wieku zaproponowano modyfikację wzoru (9.92) tak, aby zmniejszyć stałą błędów tych metod BDF, które są A-stabilne, albo powiększyć obszar stabilności pozostałych. Modyfikacja równania (9.92) do postaci

$$\sum_{m=1}^k \frac{1}{m} \nabla^m y_{n+1} = hf(y_{n+1}, x_{n+1}) + \kappa \gamma_k (y_{n+1} - y_{n+1}^{(0)}), \quad (9.94)$$

gdzie $\gamma_k = \sum_{j=1}^k \frac{1}{j}$ pozwala zwiększyć efektywność o około jedną czwartą. Parametr κ bywał dobierany eksperymentalnie. W programie Matlab metody numerycznego różniczkowania są realizowane przez procedurę `ode15s`. Zastosowano tam dobrane w eksperymentach numerycznych wartości stałej κ podane w tabeli 9.4.

Tabela 9.4. Porównanie metod wstecznego różniczkowania i numerycznego różniczkowania

Rząd k	Wsp. κ	Przyrost efektywności	BDF – kąt α	NDF – kąt α	Zmiana
1	-0,1850	26%	90°	90°	0%
2	-1/9	26%	90°	90°	0%
3	-0,0823	26%	86°	80°	-7%
4	-0,0415	12%	73°	66°	-10%
5	0	0%	51°	51°	0%

9.10. Jak dopasować metodę numerycznego rozwiązania zagadnienia początkowego do specyfiki zadania?

Przedstawiliśmy reprezentatywny wybór stosowanych współcześnie metod numerycznego rozwiązywania równań różniczkowych zwyczajnych. Metody te są cały czas rozwijane i wciąż pojawiają się nowe lub udoskonalone warianty. Nie wspomnieliśmy o metodach wykorzystujących ekstrapolację Richardsona jak w metodzie całkowania Romberga (algorytm Bulirsha-Stoera), ani o specjalnych metodach dedykowanych równaniom drugiego rzędu. Nie omawialiśmy silnie rozwijających się metod rozwiązywania równań, w których warunki brzegowe

podane są w kilku różnych punktach, a nie tylko na początku przedziału ani o układach równań różniczkowo-algebraicznych. I tak pojawiło się tu dość dużo metod, by zadać pytanie: którą z nich wybrać do rozwiązania konkretnego problemu? Oczywiście, gdyby istniała metoda najlepsza w każdej sytuacji dla każdego równania, to tylko ona byłaby używana. Tak nie jest – każda z przedstawionych metod może być doskonała dla jednych równań i przeciętna lub niedopuszczalna dla innych. Wybierając metodę, musimy uwzględnić szereg okoliczności związanych z posiadanym sprzętem, czasem przeznaczonym na obliczenia, dokładnością danych wejściowych, możliwościami oprogramowania i stosowanej arytmetyki, ale przede wszystkim musimy wiedzieć, do czego będą użyte otrzymane rozwiązania numeryczne i jakie, w związku z tym, powinny spełniać wymagania.

Najczęściej stosujemy pakiety oprogramowania zawierające gotowe algorytmy rozwiązywania równań różniczkowych zwyczajnych. Wtedy zadaniem użytkownika jest wybór metody (schematu różnicowego i algorytmu sterowania długością kroku) i wprowadzenie parametrów decydujących o dokładności rozwiązania. Poniższe zasady i przykłady mają ułatwić ten wybór, sprawić, że będzie dokonywany świadomie.

1. Metody stałokrokowe są używane wyłącznie jeśli wymaga tego specyfika stosowanej maszyny cyfrowej. Na przykład, równania różniczkowe będące częścią algorytmu sterowania implementowanego w sterowniku ze stałym czasem próbkowania (w tym przypadku zmienną niezależną jest czas i długość kroku jest równa okresowi próbkowania) będą rozwiązywane metodą stałokrokową. Czas taktowania procesora i złożoność algorytmu mogą też wymuszać zastosowanie prostych schematów różnicowych. W czasie testów symulacyjnych takiego układu trzeba upewnić się, że rozwiązanie otrzymywane metodą stałokrokową jest wystarczająco dokładne.
2. Jeżeli zadanie nie jest sztywne, można stosować jawne metody jednokrokowe z odpowiednim algorytmem sterowania długością kroku. Ważne jest odpowiednie dobranie rzędu metody do oczekiwanej dokładności wyniku. Jeżeli oczekujemy rozwiązania numerycznego z małym błędem, to wybór metody niskiego rzędu będzie niecelowy, bo wiele wartości przybliżonych będzie odrzucane, zostanie wymuszona mała długość kroku i łączny nakład obliczeń będzie większy niż w przypadku metody wyższego rzędu, mimo że liczba operacji wykonywanych w pojedynczym kroku będzie mniejsza.
3. Jeżeli rozwiązujemy równanie sztywne należy stosować metody niejawne, o nieograniczonym obszarze stabilności absolutnej, najlepiej A-stabilne, pozwalające na znaczne wydłużanie kroku po zaniknięciu szybkiej składowej rozwiązania.

Przykład 9.6

Obliczono, metodami zmiennokrokowymi dostępnymi w pakiecie Matlab, rozwiązanie zagadnienia początkowego:

$$\frac{dy}{dx} = A(-y + \cos x), \quad y(0) = 0.$$

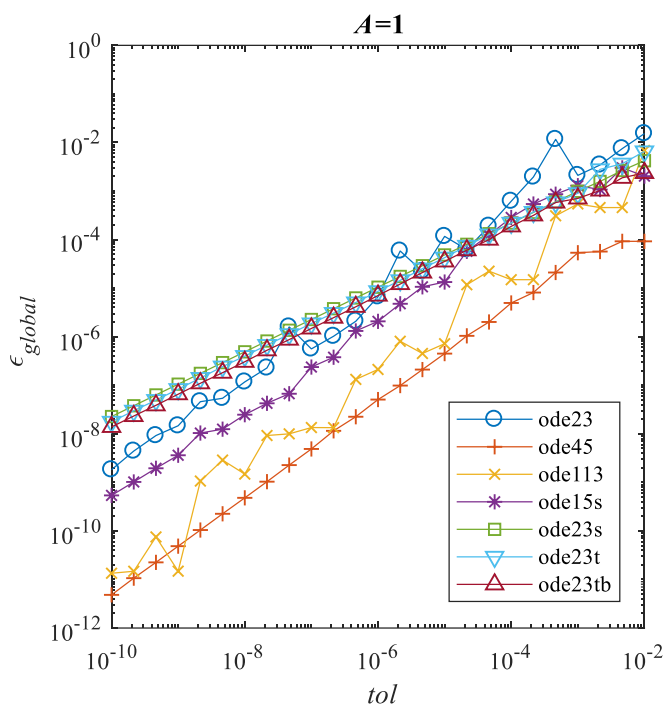
Można sprawdzić bezpośrednim rachunkiem, że $y(x) = \frac{A \sin x + A^2 \cos x - A^2 e^{-Ax}}{A^2 + 1}$ jest rozwiązaniem tego zagadnienia początkowego dla dowolnej wartości parametru A .

Znajomość rozwiązania dokładnego umożliwia porównanie błędów globalnych obliczonych rozwiązań przybliżonych. Eksperymenty numeryczne przeprowadzono dla dwóch wartości parametru A . Dla każdej z nich powtórzono wielokrotnie obliczenia dla przedziału zmiennej niezależnej $x \in \left[0, \frac{\pi}{2}\right]$ z różnymi wartościami parametru tol , służącego do wyboru dokładności (maksymalnej wartości błędu względnego w mechanizmie sterowania długością kroku), wybieranego w zakresie od 10^{-10} do 10^{-2} . Wykreślono $\max_n \{|y_n - y(x_n)|\}$ (co odpowiada błędowi globalnemu rozwiązania) w funkcji założonej dokładności tol , oraz nakład obliczeń mierzony liczbą wywołań funkcji obliczającej prawą stronę równania różniczkowego w funkcji uzyskanego błędu globalnego.

W teście użyto metod:

- ode23 – jawna, zagnieżdżona metoda typu Rungego-Kutty opracowana przez Shampine’a-Bogackiego,
- ode23t – niejawna metoda trapezów,
- ode23tb – niejawna, zagnieżdżona metoda Rungego-Kutty,
- ode23s – metoda Rungego-Kutty-Rosenbrock’a,
- ode45 – opracowana przez Dormand’a-Prince’a jawna, zagnieżdżona metoda Rungego-Kutty,
- ode113 – oparta o metody Adamsa-Bashforta i Adamsa-Moultona metoda predyktor-korektor,
- ode15s – oparta o metody numerycznego różniczkowania (z możliwością zmiany na metodę wstecznego różniczkowania).

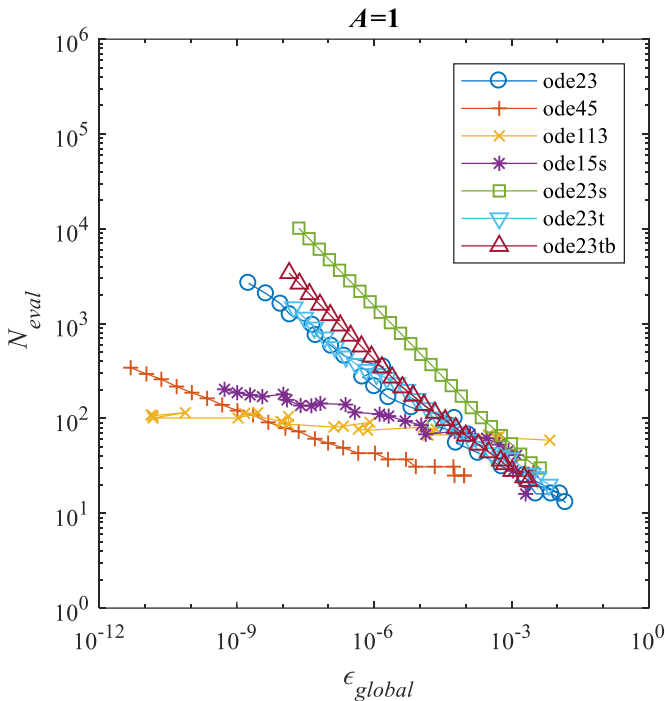
Liczby w nazwach funkcji informują o rzędzie metody: np. ode23 to para metod zagnieżdżonych drugiego i trzeciego rzędu, a ode113 może automatycznie zmieniać rząd od 1 do 13; podobnie ode15s może zmieniać rząd od 1 do 5 (można także ograniczyć zakres zmian, wykluczając metody wyższych rzędów).



Rys. 9.18. Wyniki eksperymentu numerycznego dla układu z $A = 1$. Pokazano rzeczywiście uzyskany błąd globalny ϵ_{global} w funkcji parametru tol

Jak widać na rysunku 9.18, w przybliżeniu zachowana jest proporcjonalność błędu do parametru tol dla metod wyższych rzędów (choć z różnymi współczynnikami proporcjonalności), ale metody niskiego rzędu osiągają tylko nieco lepszą dokładność niż $tol^{\frac{2}{3}}$. Te wyniki są reprezentatywne dla rozważanego zagadnienia początkowego przy różnych wartościach parametru A .

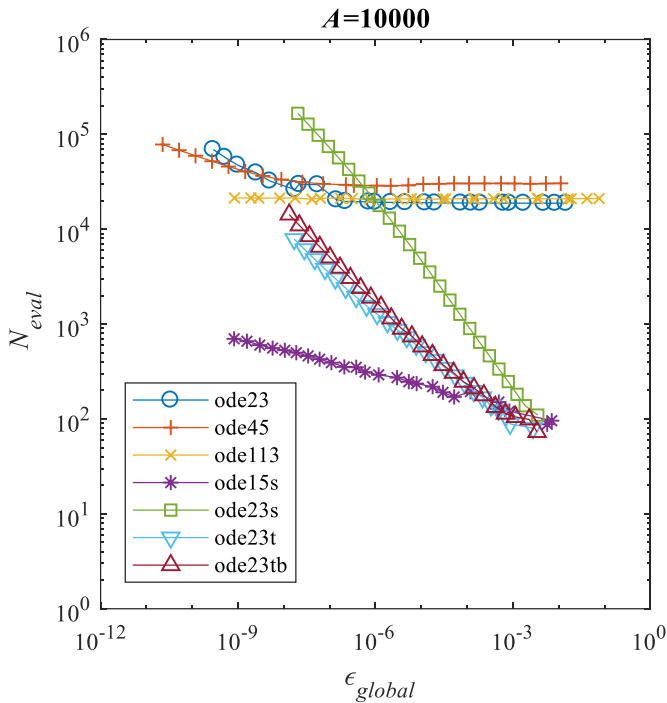
Na rysunku 9.19 przedstawiono wykresy nakładu obliczeń w funkcji osiągniętej dokładności. Nakład obliczeń mierzony był liczbą wywołań procedury obliczającej prawą stronę równania różniczkowego. Wykresy te w istocie są parametryczne – parametrem była wielkość tol .



Rys. 9.19. Nakład obliczeń N_{eval} (mierzony ilością obliczeń prawej strony równania różniczkowego) w funkcji błędu obliczeń dla $A = 1$

Jeśli żądamy dużej dokładności, to nakład obliczeń jest zdecydowanie najmniejszy przy metodach, które mają (ode45), względnie mogą osiągać¹² (ode113, ode15s) wysoki rząd. Tylko przy $tol > 10^{-4}$ pozostałe metody zbliżają się do efektywności wymienionych wyżej, a w wyjątkowych przypadkach ją przewyższają. Na wykresie widać wyraźnie różne nachylenia linii odpowiadających metodom różnych rzędów: w przybliżeniu nakład obliczeń jest odwrotnie proporcjonalny do błędów w potęgze $1/p$, gdzie p oznacza rząd metody (maksymalny jeśli zmiana rzędu odbywa się automatycznie jak w metodach ode113 i ode15s). Jakkolwiek można odnieść wrażenie, że dla małej dokładności metoda ode15s jest efektywniejsza od ode113, różnica jest niewielka i w praktyce rzeczywista efektywność może zależeć od nieuwzględnionych w tym zestawieniu narzutów obliczeniowych związanych z realizacją metody, a nie obliczaniem prawej strony równania różniczkowego.

¹² Niektóre metody – w pakiecie Matlab są to ode113 i ode15s – mogą zmieniać automatycznie nie tylko długość kroku, ale także rząd.



Rys. 9.20. Wyniki eksperymentu numerycznego dla układu o znacznej sztywności ($A=10000$)

Zwiększenie parametru (do $A = 10000$, wyniki zaprezentowano na rysunku 9.20) powoduje, że nakład obliczeń procedury ode113 opartej o metody Adamsa, podobnie jak ode23, ode45 realizującej jawne metody Rungego-Kutty praktycznie przestaje zależeć od tolerancji i ustala się na poziomie wyraźnie wyższym niż dla jakiegokolwiek metody przy $A = 1$. Bardzo nieznacznie wzrasta nakład obliczeń wymagany przez metody numerycznego różniczkowania (ode15s) i A-stabilną niejawną metodę trapezów (ode23t). Także charakter zależności nakładu obliczeń od osiągniętego błędu dla pozostałych metod niejawnych nie ulega zmianie, chociaż sam nakład rośnie prawie o rząd wielkości. Metody jawne: ode23, ode45 oraz posiadająca ograniczony obszar stabilności metoda ode113 wymagają praktycznie takiego samego nakładu obliczeń niezależnie od założonej tolerancji i rzeczywiście osiągniętego błędu: wskazuje to na ograniczenie długości kroku przez wymóg stabilności, a nie odtworzenia wolnozmiennego rozwiązania, a także zapewne znaczny odsetek kroków odrzuconych przez mechanizm automatycznego doboru.

Należy podkreślić, że we wszystkich przypadkach, niezależnie od wartości parametru A było obliczane to samo wolnozmiennne rozwiązanie, a zmiana parametru modyfikowała rodzinę rozwiązań w pobliżu obliczanej trajektorii, ale powodowało to wyraźną zmianę w funkcjonowaniu poszczególnych metod.

Z zaprezentowanych wyników eksperymentów wynikają następujące wnioski:

- parametr tolerancji służący do sterowania pracą mechanizmu automatycznej zmiany długości kroku tylko w bardzo grubym przybliżeniu odpowiada wartości rzeczywiście osiągniętego błędu,
- w przypadku zagadnień początkowych które nie są sztywne, ma sens użycie metod wysokiego rzędu, zwłaszcza gdy wymagamy dużej dokładności obliczonego rozwiązania,
- w przypadku zagadnień sztywnych należy sięgać po metody o nieograniczonym obszarze stabilności absolutnej, co eliminuje wszystkie metody jawne i metody Adamsa.

Należy pamiętać, że cała teoria leżąca u podstawy oszacowań wielkości błędu i opartej na niej automatycznego doboru długości kroku (i rzędu w niektórych przypadkach), zakłada istnienie ciągłych pochodnych cząstkowych prawej strony równania różniczkowego przynajmniej do rzędu odpowiadającego rzędowi metody. Jeśli chcemy stosować metody zmiennokrokowe do równań, które nie spełniają tego warunku, musimy odrzucić metody wielokrokowe, które pracują wtedy nieefektywnie, niekiedy z minimalnym dostępnym rzędem. W opisanych okolicznościach może być konieczne zastosowanie specjalnych środków jak wykrywanie sytuacji specjalnych (*ang.* „*event handling*”) do „przekraczania” powierzchni nieciągłości. Podobnie, metody wielokrokowe nie są godne polecenia przy rozwiązywaniu równań o prawych stronach nieciągłych względem zmiennej niezależnej, z czym mamy na przykład do czynienia przy symulacji układów sterowania z regulatorami cyfrowymi.

D1. Liczby i wektory

Ciało

Ciałem nazywamy strukturę $(K, +, \cdot, 1, 0)$, w której zbiór K zawiera co najmniej dwa elementy oznaczone symbolami $1, 0$, (nazywane elementem jednostkowym i zerowym), a działania $(+), (\cdot)$ spełniają warunki:

$$\begin{aligned} \forall_{a,b \in K} \quad a + b &= b + a, \\ \forall_{a,b,c \in K} \quad a + (b + c) &= (a + b) + c, \\ \forall_{a \in K} \quad a + 0 &= a, \\ \forall_{a \in K} \quad \exists_{b \in K} \quad a + b &= 0, \end{aligned} \tag{D1.1}$$

$$\begin{aligned} \forall_{a,b \in K} \quad a \cdot b &= b \cdot a, \\ \forall_{a,b,c \in K} \quad a \cdot (b \cdot c) &= (a \cdot b) \cdot c, \\ \forall_{a \in K} \quad a \cdot 1 &= a, \\ \forall_{a \in K} \quad a \neq 0 &\Rightarrow \exists_{b \in K} \quad a \cdot b = 1, \\ \forall_{a,b,c \in K} \quad a \cdot (b + c) &= (a \cdot b) + (a \cdot c). \end{aligned} \tag{D1.2}$$

Ciałem jest na przykład zbiór liczb rzeczywistych z działaniami dodawania i mnożenia.

Ciało liczb zespolonych

Oznaczmy przez $C = R \times R$ iloczyn kartezyjański zbioru liczb rzeczywistych. Elementami zbioru C są uporządkowane pary liczb rzeczywistych.

Liczba zespolona z to para liczb rzeczywistych $z = (a_1, a_2)$. Pierwszy element pary nazywamy **częścią rzeczywistą** liczby z i oznaczmy $a_1 = \operatorname{Re} z$, natomiast drugi element pary nazywamy **częścią urojoną** liczby z i oznaczmy $a_2 = \operatorname{Im} z$.

Równość par rozumiemy w naturalny sposób

$$(a_1, a_2) = (b_1, b_2) \Leftrightarrow a_1 = b_1 \wedge a_2 = b_2. \tag{D1.3}$$

Zdefiniujemy dodawanie i mnożenie par następująco:

$$(a_1, a_2) + (b_1, b_2) = (a_1 + b_1, a_2 + b_2), \tag{D1.4}$$

$$(a_1, a_2) \cdot (b_1, b_2) = (a_1 b_1 - a_2 b_2, a_1 b_2 + a_2 b_1). \tag{D1.5}$$

Można sprawdzić, że $(C, +, \cdot)$ jest ciałem. Elementem neutralnym dodawania (czyli zerem) jest element (para) $(0, 0)$. Elementem przeciwnym do (a_1, a_2) jest element $(-a_1, -a_2)$.

Elementem neutralnym mnożenia (jedyнкą) jest para $(1,0)$. Natomiast elementem odwrotnym do elementu $(a_1, a_2) \neq (0,0)$ jest

$$(a_1, a_2)^{-1} = \left(\frac{a_1}{a_1^2 + a_2^2}, \frac{-a_2}{a_1^2 + a_2^2} \right) \quad (\text{D1.6})$$

co łatwo sprawdzić z (D1.5).

Widać, że

$$(a, 0) + (b, 0) = (a + b, 0), \quad (a, 0) \cdot (b, 0) = (ab, 0), \quad (\text{D1.7})$$

czyli liczby zespolone postaci $(a, 0)$ można utożsamiać z liczbami rzeczywistymi. Przyjmując zasadę utożsamiania struktur izomorficznych, możemy powiedzieć, że **ciało liczb rzeczywistych $(R, +, \cdot)$ jest podciałem ciała liczb zespolonych $(C, +, \cdot)$.**

Liczbę zespoloną $(0,1)$ nazywamy jednostką urojoną i oznaczamy ją przez $j = (0,1)$ lub $i = (0,1)$. Oznaczenie j jest używane przez elektrotechników, żeby nie myliło się z popularnym oznaczeniem wartości chwilowej prądu.

Zauważmy, że $j^2 = (0,1) \cdot (0,1) = (-1,0)$, czyli $j = (0,1)$ jest pierwiastkiem kwadratowym z liczby -1 (spełnia równanie $z^2 + 1 = 0$).

Podobnie liczba $-j = (0, -1)$ jest pierwiastkiem algebraicznym z liczby -1 .

Postać algebraiczna liczby zespolonej. Każdą liczbę zespoloną z można jednoznacznie przedstawić w postaci

$$z = a + j b; \quad a, b \in R. \quad (\text{D1.8})$$

Taką postać nazywamy **postacią dwumienną**. Działania na liczbach zespolonych zapisanych w postaci dwumiennej wykonujemy tak jak na dwumianach, pamiętając tylko, że $j^2 = -1$ i nie musimy pamiętać wzoru definiującego mnożenie oraz dzielenie (mnożenie przez element odwrotny), na przykład: $(1 + 2j)(3 - j) = 3 - j + 6j - 2j^2 = 5 + 5j$.

Płaszczyzna liczb zespolonych

Liczbę zespoloną $z = (a, b) = a + j b$ interpretujemy jako wektor na płaszczyźnie. Wektory te potrafimy dodawać i odejmować tak jak wektory na płaszczyźnie euklidesowej.

Modulem liczby zespolonej $z = a + j b$ nazywamy liczbę rzeczywistą $|z| = \sqrt{a^2 + b^2}$. Jedyнкą liczbą zespoloną o module równym 0 jest liczba $(0,0)$.

Z wektorowej interpretacji liczb zespolonych z_1, z_2 i z nierówności trójkąta wynika, że

$$|z_1 + z_2| \leq |z_1| + |z_2|.$$

Liczbę zespoloną $\bar{z} = (a, -b) = a - jb$ nazywamy **liczbą sprzężoną** do liczby $z = (a, b) = a + jb$.

Prawdziwe są następujące związki

$$\begin{aligned} \bar{\bar{z}} &= z, \\ \overline{z_1 + z_2} &= \bar{z}_1 + \bar{z}_2, \\ \overline{z_1 \cdot z_2} &= \bar{z}_1 \cdot \bar{z}_2, \\ z \cdot \bar{z} &= |z|^2. \end{aligned} \tag{D1.9}$$

Postać trygonometryczna liczby zespolonej

Każdą liczbę zespoloną różną od (0,0) można zapisać w postaci

$$z = a + jb = |z| \left(\frac{a}{|z|} + j \frac{b}{|z|} \right) = |z| \left(\frac{a}{\sqrt{a^2+b^2}} + j \frac{b}{\sqrt{a^2+b^2}} \right), \tag{D1.10}$$

a ponieważ $\left(\frac{a}{\sqrt{a^2+b^2}} \right)^2 + \left(\frac{b}{\sqrt{a^2+b^2}} \right)^2 = 1$, to można oznaczyć $\frac{a}{\sqrt{a^2+b^2}} = \cos\varphi$, $\frac{b}{\sqrt{a^2+b^2}} = \sin\varphi$ i wtedy

$$z = a + jb = |z| \left(\frac{a}{|z|} + j \frac{b}{|z|} \right) = |z|(\cos\varphi + j\sin\varphi). \tag{D1.11}$$

Postać (D1.11) nazywamy **postacią trygonometryczną** liczby zespolonej $z = (a, b) = a + jb$. Kąt φ nazywamy **argumentem** liczby zespolonej i oznaczamy jako $\varphi = \text{Arg}\{z\}$. Argument jest wyznaczony z dokładnością do wielokrotności kąta 2π .

Argument $\varphi \in (-\pi, \pi]$ nazywamy **argumentem głównym** i oznaczamy jako $\varphi = \text{arg}\{z\}$. Argument główny jest wyznaczony jednoznacznie. Liczbie (0,0) nie przypisujemy argumentu. Jest ona jednoznacznie wyznaczona przez swój moduł.

Postać wykładnicza liczby zespolonej

Postać wykładniczą liczby zespolonej otrzymuje się po wykorzystaniu, tzw. wzoru Eulera

$$e^{j\varphi} = \cos\varphi + j\sin\varphi, \tag{D1.12}$$

który podaje związek między funkcją wykładniczą a funkcjami trygonometrycznymi. Każdą z tych funkcji można zdefiniować w postaci sumy szeregu i wyprowadzić wzór (D1.11), sumując szeregi. Można też uzasadnić go w następujący sposób:

Niech $F(\varphi) = e^{-j\varphi}(\cos \varphi + j \sin \varphi)$.

Wtedy $F(0) = e^{-j0}(\cos 0 + j \sin 0) = 1(1 + j0) = 1$,

$$\frac{d}{d\varphi}F(\varphi) = e^{-j\varphi}(-\sin \varphi + j \cos \varphi) - j(\cos \varphi + j \sin \varphi) = 0,$$

czyli $F(\varphi) = e^{-j\varphi}(\cos \varphi + j \sin \varphi) \equiv 1$. To znaczy, że $e^{j\varphi} = \cos \varphi + j \sin \varphi$.

Wzór Eulera i (D1.11) pozwalają napisać

$$z = a + jb = |z| \left(\frac{a}{|z|} + j \frac{b}{|z|} \right) = |z|(\cos \varphi + j \sin \varphi) = |z|e^{j\varphi}. \quad (D1.13)$$

Oczywiście jest

$$\arg\{|z|e^{j\varphi}\} = \varphi, \quad \left| |z|e^{j\varphi} \right| = |z|, \quad |e^{j\varphi}| = 1. \quad (D1.14)$$

Dla liczb zespolonych zapisanych w postaci wykładniczej łatwo można podać moduł i argument. Postać ta w bardzo dobry sposób obrazuje mnożenie, dzielenie liczb zespolonych. Od razu widać, że w **wyniku mnożenia otrzymamy liczbę, której moduł będzie równy iloczynowi modułów czynników, a argument równy sumie argumentów czynników.**

Ze wzoru Eulera wynika tożsamość:

$$e^{j\pi} + 1 = 0. \quad (D1.15)$$

Tożsamość Eulera jest często nazywana najpiękniejszym wzorem matematycznym. Wykorzystane są w niej trzy działania arytmetyczne: dodawanie, mnożenie i potęgowanie. Tożsamość łączy pięć fundamentalnych stałych matematycznych: liczbę 0, liczbę 1, liczbę π , liczbę e , liczbę j – jednostkę urojoną liczb zespolonych. Każde z działań oraz każda ze stałych użyte są dokładnie raz.

Pierwiastki wielomianu jednej zmiennej

Wielomianem stopnia n zmiennej x nazywane jest wyrażenie algebraiczne $P(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_2 x^2 + a_1 x + a_0$, $a_n \neq 0$. Jeżeli współczynniki wielomianu $a_n, a_{n-1}, \dots, a_2, a_1, a_0$ są liczbami rzeczywistymi, wielomian nazywamy rzeczywistym, jeżeli zespolonymi – zespolonym. Wielomian moniczny, to wielomian o współczynniku przy najwyższej potędze równym 1.

Pierwiastek wielomianu $P(x)$ to taka liczba p (zespolona lub rzeczywista), dla której dwumian $x - p$ dzieli bez reszty wielomian $P(x)$. **Miejszem zerowym** funkcji wielomianowej $y = P(x)$ nazywa się taką wartość zmiennej, dla której wartość funkcji wielomianowej wynosi 0, innymi słowy jest to rozwiązanie **równania algebraicznego** $P(x) = 0$.

Twierdzenie Bézouta: Liczba a jest pierwiastkiem wielomianu $P(x)$ wtedy i tylko wtedy gdy jest miejscem zerowym funkcji $P(x)$.

Zbiór miejsc zerowych funkcji wielomianowej pokrywa się więc ze zbiorem pierwiastków odpowiadającego jej wielomianu.

Krotnością pierwiastka wielomianu $P(x)$ nazywa się największą liczbę naturalną taką, że wielomian $P(x)$ dzieli się bez reszty przez wielomian $(x - p)^k$. Jeżeli pierwiastek ma krotność równą co najmniej 2, to nazywa się go **pierwiastkiem wielokrotnym** (dwu-, trzy-, cztero-, pięciokrotnym itd.), jeżeli wynosi ona 1, nazywa się go **jednokrotnym** lub **pojedynczym**.

Zasadnicze twierdzenie algebry: Każdy wielomian zespolony stopnia dodatniego ma co najmniej jeden pierwiastek w ciele liczb zespolonych.

Z zasadniczego twierdzenia algebry i twierdzenia Bézouta wynika, że każdy wielomian moniczny o rzeczywistych lub zespolonych współczynnikach, może być przedstawiony w postaci iloczynu zespolonych wielomianów liniowych:

$$P(x) = \underbrace{(x - p_1) \dots (x - p_1)}_{k_1 \text{ razy}} \dots \underbrace{(x - p_l) \dots (x - p_l)}_{k_l \text{ razy}}, \quad \sum_{i=1}^l k_i = n, \quad (\text{D1.16})$$

gdzie l jest liczbą różnych pierwiastków wielomianu, k_i – krotnością i -tego pierwiastka, n – stopniem wielomianu.

Wielomian drugiego stopnia $P(x) = x^2 + bx + c$, o rzeczywistych współczynnikach, może mieć

- dwa pierwiastki pojedyncze, rzeczywiste: $x_1 = \frac{-b - \sqrt{b^2 - 4c}}{2}$, $x_2 = \frac{-b + \sqrt{b^2 - 4c}}{2}$, jeśli $b^2 > 4c$,
- pierwiastek rzeczywisty, podwójny: $x_1 = \frac{-b}{2}$ jeśli $b^2 = 4c$,
- parę pierwiastków zespolonych sprzężonych: $x_1 = \frac{-b - j\sqrt{-b^2 + 4c}}{2}$,
 $x_2 = \frac{-b + j\sqrt{-b^2 + 4c}}{2}$, jeśli $b^2 < 4c$.

Wielomian rzeczywisty można rozłożyć na iloczyn wielomianów rzeczywistych co najwyżej drugiego stopnia. Czynniki nieliniowe mają wtedy postać $x^2 + bx + c$, przy czym $b^2 < 4c$.

Twierdzenie Abela-Ruffiniego: – pierwiastki równania algebraicznego $P(x) = 0$ stopnia wyższego niż 4 nie dają się wyrazić w postaci zależności od współczynników równania za pomocą skończonej liczby operacji polegających na wykonaniu czterech działań algebraicznych i pierwiastkowania.

Wektory i przestrzenie liniowe

Pod pojęciem wektora $x \in R^n$ jest tutaj rozumiana uporządkowana kolumna liczb rzeczywistych $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$. Dla zbioru wszystkich wektorów n -wymiarowych określone są działania dodawania:

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \Rightarrow x + y = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix} \quad (\text{D1.17})$$

i mnożenia przez liczbę rzeczywistą a :

$$ax = \begin{bmatrix} ax_1 \\ \vdots \\ ax_n \end{bmatrix}. \quad (\text{D1.18})$$

Definicja przestrzeni liniowej (wektorowej) nad ciałem K

Niech $(K, +, \dots, 1, 0)$ będzie ciałem (np. ciałem liczb rzeczywistych lub zespolonych), którego elementy nazywane są skalarami, a ono samo – ciałem skalarów. Przestrzenią liniową bądź **wektorową nad ciałem K** nazywa się zbiór V , którego elementy nazywane są wektorami z dwoma działaniami dwuargumentowymi: *dodawaniem wektorów*: $V \times V \rightarrow V$ oznaczanym $(v, w) \rightarrow v + w$ i *mnożeniem przez skalar* $K \times V \rightarrow V$ oznaczanym $(a, w) \rightarrow aw$, które spełniają poniższe aksjomaty.

- 1) Dodawanie wektorów jest łączne: dla dowolnych $u, v, w \in V$ zachodzi $u + (v + w) = (u + v) + w$.
- 2) Dodawanie wektorów jest przemienne: dla dowolnych $v, w \in V$ jest $v + w = w + v$.
- 3) Dodawanie wektorów ma element neutralny (zerowy): istnieje taki element nazywany wektorem zerowym, że $v + 0 = v$ dla dowolnego $v \in V$.
- 4) Dodawanie wektorów ma elementy przeciwne: dla każdego $v \in V$ istnieje element $w \in V$ nazywany wektorem przeciwnym do v , taki że $v + w = 0$

- 5) Mnożenie przez skalar jest rozdzielne względem dodawania wektorów: dla każdego $a \in K$ oraz $v, w \in V$ zachodzi $a(v + w) = av + aw$.
- 6) Mnożenie przez wektor jest rozdzielne względem dodawania skalarów: dla każdego $a, b \in K$ oraz $v \in V$ zachodzi $(a + b)v = av + bv$
- 7) Mnożenie przez skalar jest zgodne z mnożeniem skalarów: dla dowolnych $a, b \in K$ oraz $v \in V$ jest $a(bv) = (a \cdot b)v$.
- 8) Mnożenie przez skalar ma element neutralny: dla dowolnego $v \in V$ jest $1v = v$, gdzie oznacza 1 element neutralny mnożenia w ciele K

Jeżeli zbiór V wektorów zostanie zdefiniowany jako podzbiór w R^n , a działania dodawania i mnożenia przez skalar jak w (D1.17,18), jeżeli V jest domknięty ze względu na operacje dodawania wektorów i mnożenia przez skalar (to znaczy, że wynik tych działań należy do V) to jest przestrzenią liniową (rzeczywistą lub nad ciałem liczb rzeczywistych). Przestrzenią liniową jest więc zbiór wszystkich wektorów $x \in R^n$, a także na przykład zbiór wszystkich wektorów postaci $y = ax$, $a \in R$.

Wektor postaci:

$$y = \sum_{i=1}^k a_i x_i, \quad a_i \in R, \quad x_i \in R^n \quad (\text{D1.19})$$

nazywany jest **kombinacją liniową** wektorów x_i .

Jeżeli prawdziwa jest implikacja

$$0 = \sum_{i=1}^k a_i x_i \implies a_i = 0, \quad i = 1, \dots, k, \quad (\text{D1.20})$$

to mówimy, że wektory x_i , $i = 1, \dots, k$ są **liniowo niezależne** (żadnego wektora ze zbioru wektorów liniowo niezależnych nie można przedstawić w postaci liniowej kombinacji pozostałych).

Zbiór S wszystkich kombinacji liniowych k wektorów $x_i \in R^n$, $i = 1, \dots, k$ jest przestrzenią liniową. Nazywamy ją **przestrzenią rozpiętą na wektorach** $x_i \in R^n$, $i = 1, \dots, k$, co zapisujemy

$$S = \text{span}\{x_1, \dots, x_k\}. \quad (\text{D1.21})$$

Ze zbioru $\{x_1, \dots, x_k\}$ rozpinającego przestrzeń liniową S można wyjąć podzbiór $B \subset \{x_1, \dots, x_k\}$, taki że: zbiór B jest zbiorem wektorów liniowo niezależnych i rozpinają przestrzeń S . Zbiór ten nazywamy **bazą przestrzeni** S . Każda przestrzeń liniowa ma bazę, wszystkie bazy tej samej przestrzeni liniowej są równoliczne. Jeśli baza składa się z $n < \infty$ elementów, to n nazywamy **wymiarem przestrzeni**

S i mówimy, że przestrzeń jest **n -wymiarowa**, jeśli zbiór elementów bazy jest nieskończony to mówimy o przestrzeni nieskończenie wymiarowej. Wymiar przestrzeni wektorowej zależy od ciała, nad którym przestrzeń ta jest rozważana.

Przestrzeń wszystkich wektorów $x \in R^n$ jest przestrzenią n -wymiarową, a jej bazą jest zbiór wektorów

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \dots, \quad e_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}. \quad (D1.22)$$

Oczywiście nie jest to jedyna baza przestrzeni R^n , ale każda z baz składa się z n niezależnych liniowo wektorów.

Przestrzeń wszystkich wektorów postaci $y = ax$, $a \in R$, gdzie x jest wektorem niezerowym, jest jednowymiarowa, a jej bazą jest wektor x . Nie jest to jedyna baza tej przestrzeni, ale każda z baz składa się z jednego, niezerowego wektora.

Symbol $\|x\|$ oznacza euklidesową normę wektora x : $\|x\| = \sqrt{x^T x} = \sqrt{\sum_{i=1}^n x_i^2}$.

Iloczynem skalarnym wektorów $x, y \in R^n$ nazywamy liczbę

$$\langle x, y \rangle = x^T y, \quad (D1.23)$$

czyli

$$\|x\| = \sqrt{x^T x} = \sqrt{\langle x, x \rangle}. \quad (D1.24)$$

Przestrzeń liniową R^n z iloczynem skalarnym (D1.23) nazywamy **przestrzenią euklidesową**, a z normą (D1.24) unormowaną przestrzenią euklidesową.

Wszystkie powyższe definicje, konstrukcje i stwierdzenia można uogólnić na przypadek wektorów zespolonych $x \in C^n$, wprowadzając operacje ich dodawania i mnożenia przez liczby zespolone. Wszystko, co powiedziano wyżej o przestrzeni liniowej rzeczywistej (nad ciałem liczb rzeczywistych) pozostaje prawdziwe dla przestrzeni liniowej zespolonej (nad ciałem liczb zespolonych). Bazą przestrzeni wszystkich wektorów $x \in C^n$ nad ciałem liczb zespolonych jest zbiór wektorów (D1.22). Zauważmy, że bazą przestrzeni C^1 nad ciałem liczb zespolonych jest np. wektor $1 = 1 + j0$, czyli przestrzeń ta jest jednowymiarowa, ale bazą przestrzeni liczb zespolonych nad ciałem liczb rzeczywistych są wektory 1 i j , czyli ta przestrzeń jest dwuwymiarowa.

Iloczyn skalarny w przestrzeni C^n jest zdefiniowany jako

$$\langle x, y \rangle = \bar{x}^T y, \quad (D1.25)$$

czyli

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\bar{x}^T x} = \sqrt{\sum_{i=1}^n |x_i|^2}. \quad (\text{D1.26})$$

Przestrzeń unitarna to zespolona przestrzeń liniowa, w której został określony iloczyn skalarny.

Nierówność Cauchy'ego-Schwarza

Dla wektorów x, y w przestrzeni unitarnej zachodzi

$$\langle x, y \rangle^2 \leq \langle x, x \rangle \langle y, y \rangle \quad (\text{D1.27})$$

przy czym równość zachodzi wtedy i tylko wtedy, gdy wektory x, y są liniowo zależne.

Równoważną postacią nierówności (D1.27) jest

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\|. \quad (\text{D1.28})$$

D2. Podstawy rachunku macierzowego

Terminologia

Macierz jest prostokątną tablicą liczb nazywanych elementami macierzy. Położenie każdego z elementów w wierszu i kolumnie macierzy jest określone parą wskaźników (indeksów). Rozważamy tu *macierze prostokątne* o rzeczywistych *elementach*, choć większość definicji i twierdzeń można uogólnić na przypadek macierzy zespolonych.

Element w *i*-tym wierszu i *j*-tej kolumnie macierzy A będzie oznaczany $a_{i,j}$ lub $(A)_{i,j}$.

Liczba wierszy i kolumn określa *wymiar* macierzy.

Macierz jest nazywana *kwadratową*, jeśli liczba wierszy jest taka sama jak kolumn.

Macierzą transponowaną do macierzy A nazywamy macierz A^T taką, że $(A^T)_{i,j} = (A)_{j,i}$.

Macierz A jest nazywana *symetryczną*, jeśli $A^T = A$ (oczywiście macierz symetryczna musi być kwadratowa).

Macierz A jest nazywana *diagonalną*, jeśli $(A)_{i,j} = 0$ dla $i \neq j$. Diagonalna macierz kwadratowa o elementach $a_{i,i}$ na głównej przekątnej może być oznaczana $\text{diag } a_{i,i}$.

Macierz kwadratowa A jest nazywana *trójprzekątniową*, jeśli $(A)_{i,j} = 0$ dla $|i - j| > 1$ (elementy niezerowe mogą znajdować się na głównej przekątnej, bezpośrednio pod nią i bezpośrednio nad nią).

Macierz kwadratowa A jest nazywana *trójkątną górną*, jeśli $(A)_{i,j} = 0$ dla $i > j$ (elementy niezerowe mogą znajdować się na głównej przekątnej i nad nią).

Macierz kwadratowa A jest nazywana *prawie trójkątną górną* albo *macierzą Hessenberga*, jeśli $(A)_{i,j} = 0$ dla $i > j + 1$ (elementy niezerowe mogą znajdować się na głównej przekątnej, bezpośrednio pod nią i powyżej głównej przekątnej).

Macierz kwadratowa A jest nazywana *trójkątną dolną*, jeśli $(A)_{i,j} = 0$ dla $i < j$ (elementy niezerowe mogą znajdować się na głównej przekątnej i pod nią).

Macierz kwadratowa A jest nazywana *prawie trójkątną dolną*, jeśli $(A)_{i,j} = 0$ dla $i < j + 1$ (elementy niezerowe mogą znajdować się na głównej przekątnej, bezpośrednio nad nią i poniżej głównej przekątnej).

Macierz kwadratowa diagonalna, mająca wszystkie elementy na przekątnej równe 1 jest nazywana **macierzą jednostkową** i oznaczana przez I_n , jeśli jej wymiar jest równy n .

Jeśli A i B są takimi macierzami (niekoniecznie kwadratowymi), że zachodzi $AB = I$, to B jest nazywana **prawą odwrotnością** A , a A jest nazywana **lewą odwrotnością** B .

Jeśli macierz kwadratowa A ma prawą odwrotność B , to ta odwrotność jest jedyną i $BA = AB = I$. W tym przypadku B jest nazywana **odwrotnością** A i jest oznaczana A^{-1} . Macierz kwadratowa A , która ma odwrotność jest nazywana **nieosobliwą**.

Macierz kwadratowa A jest nazywana **normalną**, jeśli $A^T A = AA^T$.

Normalna macierz A jest nazywana **ortogonalną**, jeśli $A^T A = AA^T = I$ lub $A^T = A^{-1}$. Oczywiście macierz ortogonalna jest nieosobliwa.

Wyznaczniki

Niech A będzie macierzą kwadratową o wymiarze $n > 1$. Oznaczmy przez $A_{i,j}$ macierz otrzymaną z A przez usunięcie i -tego wiersza i j -tej kolumny.

Wyznacznikiem macierzy A jest liczba $\det A$ zdefiniowana przez poniższą zależność rekurencyjną:

- Jeśli $A = [a]$ (A jest skalar (liczbą)) to $\det A = a$.
- Jeśli $A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$, $n > 1$, to $\det A = \sum_{i=1}^n (-1)^{i+n} a_{in} \det A_{i,n}$.

Twierdzenie (rozwiniecie Laplace'a)

Niech $A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$ będzie macierzą o wymiarze $n \times n$, ($n \geq 2$). Dla dowolnego $1 \leq j \leq n$ zachodzi

$$\det A = \sum_{i=1}^n (-1)^{j+i} a_{ji} \det A_{j,i}$$

$$\det A = \sum_{i=1}^n (-1)^{i+j} a_{ij} \det A_{i,j}$$
(D2.1)

Wzory te są nazywane rozwinięciem Laplace'a względem j -ego wiersza lub j -ej kolumny, odpowiednio.

Wyznacznik $\det A_{ij}$ jest nazywany *minorem* A , liczba $(-1)^{i+j} \det A_{ij}$ jest nazywana *dopełnieniem algebraicznym elementu* a_{ij} , a transponowana macierz dopełnień algebraicznych jest nazywana macierzą dołączoną i oznaczana $\text{adj } A$: $[\text{adj } A]_{i,j} = (-1)^{i+j} \det A_{j,i}$.

Wyznacznik macierzy A o wymiarze $n \times n$ ma następujące właściwości:

- $\det A^T = \det A$,
- $\det(cA) = c^n \det A$,
- przestawienie dwóch wierszy (lub dwóch kolumn) macierzy powoduje zmianę znaku wyznacznika,
- jeśli jeden z wierszy (lub jedna z kolumn) zawiera tylko elementy zerowe, to wyznacznik jest równy 0,
- jeśli wiersze (kolumny) są liniowo zależne, to wyznacznik jest równy 0,
- dodanie do wiersza (kolumny) liniowej kombinacji pozostałych wierszy (kolumn) nie zmienia wartości wyznacznika.

Twierdzenie Cauchy'ego

Dla dowolnych kwadratowych macierzy A i B :

$$\det(AB) = \det A \cdot \det B. \quad (\text{D2.2})$$

Macierz odwrotna a wyznacznik

Odwrotność macierzy można obliczyć z zależności:

$$A^{-1} = \frac{1}{\det A} \text{adj } A, \quad (\text{D2.3})$$

gdzie $\text{adj } A$ oznacza transponowaną macierz dopełnień algebraicznych elementów macierzy A . Stąd stwierdzenia "macierz A jest nieosobliwa" i " $\det A \neq 0$ " są równoważne.

Odwrotność macierzy A ma następujące właściwości:

$$(A^{-1})^{-1} = A, \quad (AB)^{-1} = B^{-1}A^{-1}, \quad \det(A^{-1}) = \frac{1}{\det A}. \quad (\text{D2.4})$$

Rząd

Rzędem macierzy A jest nazywana liczba liniowo niezależnych wierszy A . Jest ona równa liczbie liniowo niezależnych kolumn. Może być też określona jako

wymiar *podprzestrzeni liniowej rozpiętej na wierszach (kolumnach) macierzy*, czyli podprzestrzeni liniowej złożonej ze wszystkich wektorów, które są liniowymi kombinacjami wierszy (kolumn) macierzy A . Jest oznaczana przez $\text{rank}(A)$.

Wymiar największego niezerowego minora (podwyznacznika) macierzy jest równy jej rzędowi. Dla macierzy kwadratowej A o wymiarze n stwierdzenia “ A jest nieosobliwa”, “ $\det A \neq 0$ ” oraz “ $\text{rank}(A) = n$.” są równoważne.

Wartości i wektory własne macierzy

Niezerowy wektor x (rzeczywisty lub zespolony) jest nazywany **wektorem własnym** macierzy kwadratowej A odpowiadającym jej **wartości własnej** s (która może być liczbą rzeczywistą lub zespoloną), jeśli

$$Ax = sx, \quad (\text{D2.5})$$

lub równoważnie

$$(sI - A)x = 0. \quad (\text{D2.6})$$

Niezerowe rozwiązanie tego równania jest możliwe wtedy i tylko wtedy, gdy macierz $sI_n - A$ jest osobliwa (to jest $\det(sI_n - A) = 0$). Z rozwinięcia Laplace’a i indukcji zupełnej wynika, że wyznacznik $\det(sI_n - A)$ jest wielomianem stopnia n względem s . Wielomian ten (oznaczymy go $p(s) = s^n + b_{n-1}s^{n-1} + \dots + b_2s^2 + b_1s + b_0$) jest nazywany **wielomianem charakterystycznym** macierzy A . Wartości własne można równoważnie zdefiniować jako pierwiastki wielomianu $p(s)$ wyznaczane z równania:

$$s^n + b_{n-1}s^{n-1} + \dots + b_2s^2 + b_1s + b_0 = 0, \quad (\text{D2.7})$$

nazywanego równaniem charakterystycznym macierzy A . Wiadomo, że wielomian stopnia n ma n (rzeczywistych lub zespolonych, parami sprzężonych) pierwiastków s_1, s_2, \dots, s_n , (gdzie pierwiastki wielokrotne wypisano wielokrotnie, zgodnie z ich krotnością). Tak więc macierz A ma n (różnych lub nie) wartości własnych.

Jeśli macierz jest trójkątna (górną lub dolną), to wartości własne są elementami leżącymi na głównej przekątnej.

Każdy wektor własny x_i odpowiadający wartości własnej s_i (o krotności μ_i) spełnia równanie $Ax_i = s_i x_i$, jest więc określony z dokładnością do czynnika skalującego, to znaczy, jeśli x_i jest wektorem własnym, to także ax_i jest wektorem własnym dla dowolnego $a \neq 0$.

Twierdzenie o liniowej niezależności wektorów własnych

Wektory własne x_i ($i = 1, 2, \dots, n$) odpowiadające różnym wartościom własnym s_i ($i = 1, 2, \dots, n$) są liniowo niezależne.

Wartości własne macierzy symetrycznych ortogonalnych

Wszystkie wartości własne macierzy symetrycznej są rzeczywiste.

Wszystkie wartości własne macierzy ortogonalnej mają moduł równy 1.

Przekształcenie przez podobieństwo

Dwie macierze A i B nazywamy *podobnymi*, jeśli istnieje nieosobliwa macierz T taka, że $B = T^{-1}AT$

Macierze podobne mają ten sam wielomian charakterystyczny, czyli te same wartości własne (o takich samych krotnościach). Jeśli x_i jest wektorem własnym związanym z wartością własną s_i macierzy A , to Tx_i jest wektorem własnym związanym z tą samą wartością własną s_i macierzy B . Istotnie:

z twierdzenia Cauchy'ego i z tożsamości $\det(T^{-1}) = \frac{1}{\det T}$ wynika

$$\det(sI - B) = \det(sT^{-1}T - T^{-1}AT) = \det[T^{-1}(sI - A)T] = \det(T^{-1}) \det(sI - A) \det T = \det(sI - A). \quad (\text{D2.8})$$

Twierdzenie Cayley'a-Hamiltona

Każda macierz spełnia swoje równie charakterystyczne, to znaczy jeśli $s^n + b_{n-1}s^{n-1} + \dots + b_2s^2 + b_1s + b_0 = 0$ jest równaniem charakterystycznym macierzy A , to

$$A^n + b_{n-1}A^{n-1} + \dots + b_2A^2 + b_1A + b_0I = 0. \quad (\text{D2.9})$$

Diagonalizacja

Rozważmy macierz $\Lambda = \begin{bmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & s_n \end{bmatrix}$ w której elementy niezerowe są rzeczywiste lub zespolone parami sprzężone i macierz A podobną do Λ . Istnieje więc nieosobliwa macierz T taka, że $\Lambda = T^{-1}AT$. Wartości własne macierzy podobnych są jednakowe, to znaczy że wartościami własnymi macierzy A są liczby s_1, \dots, s_n . Po przepisaniu zależności $AT = T\Lambda$ kolumna po kolumnie otrzymuje się $At_i = s_it_i$ gdzie $t_i, i = 1, \dots, n$ są kolumnami macierzy T . Tak więc kolumny T muszą być wektorami własnymi macierzy A . Macierz T jest nieosobliwa, więc wektory własne macierzy A muszą być liniowo niezależne, czyli tworzyć bazę przestrzeni C^n nad ciałem liczb zespolonych. Rozumowanie powyższe można odwrócić i pokazać, że: każda macierz kwadratowa A o wymiarze $n \times n$

mająca n niezależnych liniowo wektorów własnych jest podobna do macierzy diagonalnej.

Macierz ta jest nazywana **kanoniczną postacią diagonalną** macierzy A .

Każda macierz kwadratowa A o wymiarze $n \times n$ mająca n pojedynczych wartości własnych ma oczywiście n niezależnych liniowo wektorów własnych, jest więc podobna do macierzy diagonalnej. Macierz podobną do macierzy diagonalnej nazywamy **diagonalizowalną**.

Nie każda macierz kwadratowa jest podobna do macierzy diagonalnej.

Klatką Jordana związaną z liczbą s_i nazwiemy macierz postaci

$$J_{ij} = \begin{bmatrix} s_i & 1 & 0 & \cdots & 0 \\ 0 & s_i & 1 & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & & \ddots & 1 \\ 0 & 0 & 0 & \cdots & s_i \end{bmatrix}, \quad (\text{D2.10})$$

a **blokiem Jordana** macierz blokową postaci

$$J_i = \begin{bmatrix} J_{i1} & 0 & \cdots & 0 \\ 0 & J_{i2} & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & J_{id_i} \end{bmatrix}, \quad (\text{D2.11})$$

gdzie J_{ij} $j = 1, \dots, d_i$ są klatkami Jordana związanymi z tą samą liczbą s_i , a pozostałe elementy są zerami. Jak widać, każdy blok Jordana ma tylko jedną wartość własną s_i .

Twierdzenie o postaci kanonicznej Jordana: Każda macierz kwadratowa jest podobna do macierzy blokowej, która ma na głównej przekątnej bloki Jordana, a poza nią bloki zerowe. Macierz tę

$$J = \begin{bmatrix} J_1 & 0 & \cdots & 0 \\ 0 & J_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & J_k \end{bmatrix}, \quad (\text{D2.12})$$

nazywamy **postacią kanoniczną Jordana** macierzy A .

D3. Elementy analizy matematycznej

Poniżej zebrano podstawowe definicje i twierdzenia analizy matematycznej, do których odwoływano się przy badaniu właściwości metod numerycznych.

Ciągłość

Funkcja $f: R^n \rightarrow R^m$ jest **ciągła** w punkcie x , jeżeli dla każdego ciągu x_i zbieżnego do x ciąg $f(x_i)$ zbiega do $f(x)$ lub równoważnie, jeśli

$$\forall \varepsilon > 0 \exists \delta > 0 \forall z \in R^n: \|x - z\| < \delta \Rightarrow \|f(x) - f(z)\| < \varepsilon. \quad (D3.1)$$

Funkcja jest ciągła w zbiorze $S \subset R^n$ jeśli jest ciągła w każdym punkcie $x \in S$. Funkcja $f: R^n \rightarrow R^m$ jest **jednostajnie ciągła** na zbiorze $S \subset R^n$ jeżeli

$$\forall \varepsilon > 0 \exists \delta > 0 \forall x, z \in S: \|x - z\| < \delta \Rightarrow \|f(x) - f(z)\| < \varepsilon. \quad (D3.2)$$

Ciągłość jednostajna jest właściwością definiowaną na zbiorze S , a nie w każdym punkcie zbioru S , jak ciągłość w sensie definicji (D3.1). Ciągłość jednostajna jest mocniejszym warunkiem niż ciągłość w każdym punkcie zbioru S , ale jeśli zbiór S jest domknięty i ograniczony, to obie te właściwości są równoważne.

Funkcja $f: R \rightarrow R^m$ jest **odcinkowo ciągła** w przedziale $I \subset R$, jeśli jest ciągła w każdym punkcie dowolnego ograniczonego podprzedziału $I_0 \subset I$ z wyjątkiem skończonego zbioru punktów nieciągłości x_i , w każdym z punktów nieciągłości obie granice jednostronne funkcji f istnieją i są ograniczone.

Własność Darboux: Jeżeli $f: [a, b] \rightarrow R$ jest funkcją ciągłą oraz $f(a) < f(b)$, to dla każdego $p \in (f(a), f(b))$ istnieje taki punkt $c \in [a, b]$, że $f(c) = p$. W szczególności: jeżeli $f: [a, b] \rightarrow R$ jest funkcją ciągłą oraz $f(a)f(b) < 0$ to istnieje taki punkt $c \in [a, b]$, że $f(c) = 0$.

Różniczkowalność

Funkcja $f: R \rightarrow R^m$ jest różniczkowalna w punkcie x , jeśli granica

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \quad (D3.3)$$

istnieje i jest skończona. Granica ta jest nazywana **pochoďną** funkcji f w punkcie x .

Funkcja $f: R^n \rightarrow R^m$ $f(x) = \begin{bmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_m(x_1, \dots, x_n) \end{bmatrix}$ jest różniczkowalna (w sposób ciągły) w punkcie x , jeśli wszystkie pochodne $\frac{\partial f_i}{\partial x_j}(x)$ istnieją (są ciągłe). Macierz

$J_f = \frac{\partial f}{\partial x} := \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \dots & \frac{\partial f_1}{\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \dots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix}$ będzie nazywana **macierzą Jacobiego**. Re-

prezentuje ona odwzorowanie liniowe z przestrzeni R^n do R^m . Wyjątkowo, dla funkcji $f: R^n \rightarrow R$ przyjęto oznaczenia $\nabla f = \frac{\partial f}{\partial x^T} = \left[\frac{\partial f}{\partial x_1} \quad \dots \quad \frac{\partial f}{\partial x_n} \right]$, $\frac{\partial f}{\partial x} = \nabla f^T$.

Jeżeli funkcja $f: R \rightarrow R$ ma ograniczoną pochodną w przedziale $I \subset R$, to jest jednostajnie ciągła na I .

Warunek Lipschitza

Funkcja $f: R^n \rightarrow R^n$ spełnia w zbiorze $S \subset R^n$ lokalnie **warunek Lipschitza** (jest lokalnie lipschitz'owska), jeśli dla każdego punktu $x \in S$ istnieje otoczenie $S_x \subset S$, w którym

$$\forall_{z \in S_x} \|f(x) - f(z)\| \leq K \|x - z\| \quad (\text{D3.4})$$

dla pewnej stałej K nazywanej stałą Lipschitza. Funkcja $f: R^n \rightarrow R^n$ spełnia na zbiorze $S \subset R^n$ warunek Lipschitza, jeśli warunek (D3.4) jest spełniony dla każdych $x, z \in S$ z tą samą stałą K . Jeżeli $S = R^n$, to funkcja f spełnia warunek Lipschitza globalnie.

Funkcja $f: R \rightarrow R$ spełniająca warunek Lipschitza na przedziale I jest ciągła i różniczkowalna prawie wszędzie na I , a moduł jej pochodnej jest ograniczony przez stałą Lipschitza, jest więc także jednostajnie ciągła.

Jeżeli $f: R^n \rightarrow R^n$ jest w wypukłym zbiorze $S \subset R^n$ różniczkowalna i istnieje stała K , taka że $\left\| \frac{\partial f}{\partial x} \right\| \leq K$, to funkcja f spełnia na zbiorze S warunek Lipschitza ze stałą K .

Funkcja $f: R^n \times [a, b] \rightarrow R^n$ spełnia warunek Lipschitza lokalnie w punkcie x_0 , **jednostajnie** względem $t \in [a, b]$, jeśli istnieje otoczenie S_{x_0} i stała K takie, że:

$$\forall_{z, y \in S_{x_0}} \forall_{t \in [a, b]} \|f(y, t) - f(z, t)\| \leq K \|y - z\|. \quad (\text{D3.5})$$

Oznaczenia przestrzeni funkcyjnych

Mówimy, że funkcja f zmiennej rzeczywistej należy do przestrzeni funkcji n -krotnie różniczkowalnych w przedziale $[a, b]$ i oznaczamy $f \in C^n[a, b]$, jeśli jej n -ta pochodna istnieje i jest ciągła w przedziale $[a, b]$.

Mówimy, że funkcja f , której dziedziną jest cały zbiór liczb rzeczywistych R należy do przestrzeni funkcji n -krotnie różniczkowalnych i oznaczamy $f \in C^n(R)$, jeśli jej n -ta pochodna istnieje i jest ciągła wszędzie w R .

$C^0(R) := C(R)$ oznacza przestrzeń funkcji ciągłych. Dla kolejnych n zachodzi:

$$C^\infty(R) \subset \dots \subset C^{n+1}(R) \subset C^n(R) \subset \dots \subset C(R). \quad (D3.6)$$

Twierdzenie Lagrange'a o wartości średniej

Jeżeli funkcja $f: [a, b] \rightarrow R$ jest ciągła w przedziale $[a, b]$ i różniczkowalna w przedziale (a, b) , to istnieje punkt $c \in (a, b)$, taki że

$$f'(c) = \frac{f(b) - f(a)}{b - a}. \quad (D3.7)$$

Szczególny przypadek twierdzenia Lagrange'a jest znany pod nazwą twierdzenie Rolle'a.

Twierdzenie Rolle'a

Jeżeli funkcja $f: [a, b] \rightarrow R$ jest ciągła w przedziale $[a, b]$ i różniczkowalna w przedziale (a, b) i jeżeli $f(a) = f(b)$, to istnieje punkt $c \in (a, b)$, taki że $f'(c) = 0$.

Wzór Taylora

Wzór Taylora pozwala przedstawić zachowanie nieliniowej, gładkiej funkcji $f(x)$ w otoczeniu wybranego argumentu c .

Jeśli $f \in C^n[a, b]$ i jeśli $\frac{d^{n+1}}{dx^{n+1}} f(x)$ istnieje w (a, b) , to dla dowolnych $x, c \in [a, b]$

$$f(x) = \sum_{k=0}^n \frac{1}{k!} f^{(k)}(c)(x - c)^k + E_n(x), \quad (D3.8)$$

gdzie dla pewnego ξ (zależnego od x) leżącego między c a x

$$E_n(x) = \frac{1}{(n + 1)!} f^{(n+1)}(\xi)(x - c)^{n+1} \quad (D3.9)$$

($E_n(x)$ to **reszta wzoru Taylora w postaci Lagrange'a**). Jeśli $c = 0$ – to wzór Taylora jest nazywany wzorem **Maclaurina**.

Jeśli $f \in C^{n+1}[a, b]$, to dla dowolnych $x, x + h \in [a, b]$

$$f(x + h) = \sum_{k=0}^n \frac{h^k}{k!} f^{(k)}(x) + E_n(h), \quad (\text{D3.10})$$

gdzie, dla pewnego ξ leżącego między x a $x + h$

$$E_n(h) = \frac{h^{n+1}}{(n+1)!} f^{(n+1)}(\xi). \quad (\text{D3.11})$$

Biorąc $n \rightarrow \infty$ we wzorze Taylora (jeśli granica istnieje), otrzymujemy **rozwinięcie funkcji $f(x)$ w szereg Taylora**:

$$f(x) = \sum_{k=0}^{\infty} \frac{1}{k!} f^{(k)}(c)(x - c)^k \quad (\text{D3.12})$$

lub

$$f(x + h) = \sum_{k=0}^{\infty} \frac{h^k}{k!} f^{(k)}(x). \quad (\text{D3.13})$$

Symbole Landau'a

Symbole Landau'a pozwalają krótko opisać szybkość zbieżności funkcji.

Jeśli $\lim_{h \rightarrow 0} f(h) = 0$ i $\lim_{h \rightarrow 0} g(h) = 0$, $g(h) \neq 0$, to oznaczamy:

- $f(h) = O(g(h))$ jeśli istnieje stała $C \neq 0$ taka, że $\lim_{h \rightarrow 0} \frac{f(h)}{g(h)} = C$ (f dąży do 0 tak szybko jak g),
- $f(h) = o(g(h))$ jeśli $\lim_{h \rightarrow 0} \frac{f(h)}{g(h)} = 0$ (f dąży do 0 szybciej niż g).

Podobnie można zdefiniować symbol Landau'a dla wartości argumentu rosnących do nieskończoności. Jeśli istnieje takie $n_0 \in \mathbb{N}$ oraz takie $c \in \mathbb{R}$, że dla każdego $n \geq n_0$ prawdziwa jest nierówność $f(n) \leq cg(n)$, to mówimy, że funkcja f jest rzędu funkcji g i oznaczamy $f(n) = O(g(n))$. Funkcja $f(n)$ może dążyć do nieskończoności, ale „nie szybciej” niż $g(n)$.

Bibliografia

- Dahlquist, Germund; Björck, Åke (1983), *Metody numeryczne*. PWN Warszawa, 1983. ISBN 83-01-04276-1.
- Dryja, Maksymilian; Jankowska, Janina; Jankowski, Michał (1982) „Przegląd metod i algorytmów numerycznych”, Część 2. WNT Warszawa, 1982. ISBN 83-204-0352-9.
- Fortuna, Zenon; Macukow, Bohdan; Wąsowski, Janusz (1998), *Metody numeryczne*. WNT Warszawa 1982, 1993, 1995, 1998, ISBN 83-204-1875-5.
- Fortuna, Zenon; Macukow, Bohdan; Wąsowski, Janusz (2017), *Metody numeryczne*, Wydanie VII, Wydawnictwo Naukowe PWN, Warszawa 2017, ISBN 9788301193126.
- Jankowska, Janina; Jankowski, Michał (1981), *Przegląd metod i algorytmów numerycznych*. Część 1, WNT Warszawa, 1981. ISBN 83-204-0226-3.
- Kącki, Edward; Małolepszy, Andrzej; Romanowicz, Alicja (1997), *Metody numeryczne dla inżynierów*, Wydawnictwo Politechniki Łódzkiej, 1997, ISBN 83-87198-32-3.
- Khalil, Hassan. K. (2002), *Nonlinear Systems*, Third Edition, Prentice Hall, ISBN-10: 0130673897.
- Kiełbasiński, Andrzej; Schwetlick, Hubert (1992), *Numeryczna algebra liniowa. Wprowadzenie do obliczeń zautomatyzowanych*, WNT Warszawa, 1992, ISBN 83-204-1260-9.
- Kincaid, David; Cheney, Ward (2006), *Analiza numeryczna*, WNT Warszawa, 2006, ISBN 83-204-3078-X.
- Krupowicz, Andrzej (1986), *Metody numeryczne zagadnień początkowych równań różniczkowych zwyczajnych*, PWN Warszawa, 1986, ISBN 83-01-06638-5.
- Krupka, Jerzy; Morawski, Roman Z.; Opalski, Leszek J. (1997), *Metody numeryczne dla studentów elektroniki i technik informacyjnych*, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa, 1997.
- Krupka, Jerzy; Morawski, Roman Z.; Opalski, Leszek J. (1999), *Wstęp do metod numerycznych dla studentów elektroniki i technik informacyjnych*, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa, 1999. ISBN 83-7207-150-0.
- Paszkowski, Stefan (1975), *Zastosowania numeryczne wielomianów i szeregów Czebyszewa*, PWN Warszawa, 1975.
- Plato, Robert, *Concise Numerical Mathematics*, American Mathematical Society, 2003, ISBN 0-8218-3414-2.
- Ralston, Anthony (1983), *Wstęp do analizy numerycznej*, PWN Warszawa 1983, ISBN 83-01-01626-4.
- Quarteroni, Alfio Riccardo Sacco, Fausto Saleri, *Numerical Mathematics*, Springer-Verlag New York, Inc., 2000, ISBN 0-387-98959-5.
- Stoer, Josef (1979), *Wstęp do metod numerycznych*, tom 1, PWN Warszawa, 1979, ISBN 83-01-00077-5.

Stoer, Josef; Bulirsch, Roland (1980), *Wstęp do metod numerycznych*, tom 2, PWN Warszawa, 1980, ISBN 83-01-00078-3.

Stoer, Josef; Bulirsch, Roland (1987), *Wstęp do analizy numerycznej*, PWN Warszawa, 1987, ISBN 83-01-06505-2.

Szmurło, Robert; Markiewicz, Tomasz; Wincenciak Stanisław (2015), *Metody numeryczne. Wykłady na Wydziale Elektrycznym Politechniki Warszawskiej (eBook)*, Wydanie 1, 2015, ISBN: 978-83-7814-414-4, ISBN wersji drukowanej 978-83-7814-220-1.

Tatjewski, Piotr (2013), *Metody numeryczne*, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 2013, ISBN 978-83-7814-160-0.

Trefethen Lloyd N., *Approximation Theory and Approximation Practice*, SIAM 2013, ISBN 978-1-611972-39-9.

Indeks terminów

algorytm Brendta 167
algorytm Remeza 88
analiza numeryczna 7
analiza progresywna 39
analiza składowych głównych 220
analiza wsteczna 40
aproksymacja 78
aproksymacja jednostajna 86
aproksymacja średniokwadratowa 80, 82, 87
arytmetyka 11
arytmetyka przedziałowa 31
arytmetyka zmiennoprzecinkowa 59
błąd bezwzględny 23
błąd danych wejściowych 22
błąd globalny schematu różnicowego 230
błąd lokalny schematu różnicowego 230, 231
błąd metody (obcięcia) 23
błąd nieunikniony 36
błąd reprezentacji 22
błąd schematu różnicowego 230
błąd względny 23
błędy skrócenia (ucięcia/zaokrąglenia) 22
cecha, cecha liczby zmiennoprzecinkowej 12–14, 17, 18, 27
cyfra 11
cyfry całkowite 11
cyfry istotne 27
cyfry poprawne 27
cyfry ułamkowe 11, 27
cyfry znaczące 28
deflacja 184, 185
działania arytmetyczne 11
ekstrapolacja Richardsona 119
element główny 50
eliminacja Gaussa 47
eliminacja Gaussa bez wyboru elementu głównego 50
eliminacja Gaussa z częściowym (kolumnowym) wyborem elementu głównego 50
eliminacja Gaussa z pełnym wyborem elementu głównego 50
epsilon maszynowy 16
formuła barycentryczna 112
fraktal 188
funkcja interpolująca 78

funkcja przyrostowa (inkrementalna) schematu różnicowego 231, 248
funkcja sklejana 107
funkcje bazowe 79
Google PageRank 214
gramian 82
iloczyn skalarny funkcji 81
interpolacja odcinkowa 105-7
interpolacja wielomianami sześciennymi Hermite'a 106
interpolacja wielomianowa 90
interpolacja 78
interpolant 78
klasa złożoności obliczeniowej 42
kompresja obrazu 221
korektor 250
krok schematu różnicowego 229
kryterium zatrzymania metody iteracyjnej 143, 146
kwadratura adaptacyjna 141
kwadratura Gaussa 136
kwadratura Newtona–Cotesa 134
kwadratura prosta 133
kwadratura prostokątów 138-140
kwadratura trapezów 138-140
kwadratura złożona 133, 138
liczba stałoprzecinkowa 13
liczba zmiennoprzecinkowa 14
liczba maszynowa (mająca reprezentację maszynową) 14
liczba zmiennoprzecinkowa znormalizowana 13
liczba binarna 12
lokalny błąd odcięcia schematu różnicowego 231
macierz diagonalizowalna 199
macierz Hessenberga 210, 285
macierz permutacji 52
macierz rzadka 75, 181
macierz stowarzyszona 203
macierz Vandermonde'a 92-4, 109
macierz wzmocnienia błędu 233
macierze podobne 198
mantysa 13
mantysa liczby zmiennoprzecinkowej 12-14
metoda Dormanda–Prince'a 253
metoda (schemat) Hornera 112
metoda Abertha–Ehrlicha 189
metoda Adamsa 259
metoda Adamsa–Moultona 259

- metoda Adamsa–Bashfortha 259
- metoda Aitkena 96
- metoda Bairstowa 189
- metoda bisekcji 147
- metoda Boerscha-Supana 191
- metoda Broydena 177
- metoda Casha–Karpa 253
- metoda Duranda–Kenera 191
- metoda Eulera jawna 236
- metoda Eulera niejawna 242
- metoda Gaussa–Seidela 180
- metoda Heuna 247
- metoda iteracji prostej 148
- metoda iteracyjna 143
- metoda iteracyjna zbieżna globalnie 143
- metoda iteracyjna zbieżna lokalnie 143
- metoda Jacobiego 180
- metoda Kryłowa 202
- metoda Laguerre’a 190
- metoda Lehmera-Schura 192
- metoda Maehly’ego 188
- metoda Müllera 192
- metoda nadrelaksacji 180
- metoda Newtona tworzenia wielomianu interpolacyjnego 94
- metoda Newtona-Raphsona 149, 175, 185
- metoda Noureina 192
- metoda numeryczna 7, 21
- metoda numerycznego różniczkowania 267
- metoda połowienia kroku 251
- metoda potęgowa 204
- metoda predyktor–korektor 250
- metoda punktu środkowego 245, 247, 248
- metoda QR 207
- metoda *regula falsi* 164
- metoda rekurencyjna budowania wielomianu interpolacyjnego 95
- metoda Romberga 140
- metoda Rungego–Kutty 247, 248
- metoda Rungego–Kutty m-etapowa 252
- metoda Rungego–Kutty–Fehlberga 252
- metoda Shampine’a–Bogackiego 253
- metoda wielokrokowa liniowa 256
- metoda wstecznego różniczkowania 264
- metody minimalizacji 179
- mod 235

nierówność Cauchy’ego–Schwarza 201, 283
norma macierzowa indukowana przez normę wektorową 71
numeryczna realizacja algorytmu 22
obszarem stabilności absolutnej 234, 237, 257
odwrotna interpolacja kwadratowa 165
odwrotna metoda potęgowa 206
operacja dominująca 40
operacja zmiennoprzecinkowa 40
podstawa systemu 11
podstawa sytemu liczbowego 11
podstawa systemu pozycyjnego 11
poprawne zaokrąglenia 27
postać potęgowa wielomianu 112
precyzja 14
predyktor 250, 263
problemem Cauchy’ego 225
przekształcenie przez podobieństwo 198
przenoszenie (propagacja) błędu 28
reprezentacja stałoprzecinkowa 13
reprezentacja zmiennoprzecinkowa 14
reszta układu równań 59
reszta wzoru interpolacyjnego 99
rozkład QR 208, 210
rozkład szczególny 211
rozkład trójkątny 52
rozwiązanie ogólne 225
rozwiązanie samouzgodnione 250
rozwiązanie szczególne 225
równanie charakterystyczne 198
równanie różniczkowe liniowe, stacjonarne 234
równanie różniczkowe zwyczajne 225
równanie sztywne 238
różnica centralna 120
różnica progresywna 119
różnica wsteczna 119
różniczkowanie numeryczne 119
rząd metody iteracyjnej 144
rząd schematu różnicowego 231
schemat Hornera 42, 112
schemat różnicowy 229
schemat różnicowy absolutnie stabilny 233
schemat różnicowy $A(\alpha)$ -stabilny 257
schemat różnicowy A -stabilny 257
schemat różnicowy jawny(otwarty) 230

schemat różnicowy jednokrokowy/wielokrokowy 229, 236
schemat różnicowy niejawny (zamknięty) 230
schemat różnicowy stabilny 233
schemat różnicowy stałokrokowy/zmiennokrokowy 229
schemat różnicowy zerostabilny 256
schemat różnicowy zgodny 231
siatka 77
składowe główne 212
skrócenie liczby 24
stabilność numeryczna algorytmu 37
stała asymptotyczna błędu metody iteracyjnej 144
symbole Landau'a 294
system liczbowy 11
system liczbowy pozycyjny 11
test Schura–Cohna 192, 193
triangulacja 111
trójkątna rodzina wielomianów 94
twierdzenia o istnieniu, jednoznaczności i przedłużalności rozwiązania równania różniczkowego 227–8
twierdzenie Abela–Ruffiniego 280
twierdzenie Banacha 175
twierdzenie Bézouta 185, 279
twierdzenie Cauchy'ego 287
twierdzenie Cayley'a–Hamiltona 289
twierdzenie Lagrange'a o wartości średniej 168, 293
twierdzenie o alternansie 87
twierdzenie o postaci kanonicznej Jordana 290
twierdzenie Rolle'a 293
twierdzenie Weierstrassa 87
ucięcie liczby 27
układ (rodzina) funkcji ortogonalnych 81
układ równań liniowych 45
układ równań normalnych 82
układ trójkątny 46
ukryte indeksowanie semantyczne 218
uwarunkowanie zadania numerycznego 33
wartość własna macierzy 197
warunek Lipschitza 226
warunek położenia pierwiastków 256
wektor szczególny (prawy/lewy) macierzy 211
wektor własny macierzy 197
węzeł (punkt węzłowy) 77
węzły Czebyszewa I i II rodzaju 84, 88, 103
węzły kwadratury 133

widmo macierzy 198
wielomian charakterystyczny 198, 202
wielomian moniczny 86
wielomiany Czebyszewa 84
własność Darboux 291
wskaźnik efektywności metody iteracyjnej 144
wskaźnik uwarunkowania 34, 72–73, 75, 82–83, 92–93, 135, 201–202, 213
współczynnik (wskaźnik) uwarunkowania macierzy 72
współczynnik stabilności 37
współczynniki kwadratury 133
wygładzanie pierwiastków 185
wzór interpolacyjny Lagrange’a 109
wzór interpolacyjny Vandermonde’a 91
wzór Taylora 293–294
zagadnienie początkowe 225
zagnieżdżone metody Rungego-Kutty 252
zaokrąglenie liczby 25
zasadnicze twierdzenie algebry 279
zbieżność kwadratowa metody iteracyjnej 144
zbieżność liniowa metody iteracyjnej 144
zbiór liczb zmiennoprzecinkowych nieznormalizowany 13
zbiór liczb zmiennoprzecinkowych znormalizowany 13
zjawisko Gibbsa 104
zjawisko Rungego 103
złożoność czasowa 40
złożoność obliczeniowa algorytmu 40
złożoność oczekiwana 43
złożoność optymistyczna 43
złożoność pamięciowa 40
złożoność pesymistyczna 43
znak liczby 11

Indeks nazwisk

Nazwisko	Imię	Żyjący	Strona
Abel	Niels Henrik	1802-1829	200
Adams	John Couch	1819-1892	259, 260, 261, 263, 273, 274
Aitken	Alexander Craig "Alec"	1895-1967	96, 97, 98
Bashforth	Francis	1819-1912	259, 263
Bernstein	Siergiej Natanowicz	1880-1968	87, 88, 89, 90
Bézout	Étienne	1730-1783	185, 279
Björk	A° e	współcześnie	164
Bogacki	Przemysław	współcześnie	253, 270
Börsch-Supan	Axel	1954-	192
Brendt	Richard Peirce	1946-	167
Broyden	Charles George	1933-2011	177, 178
Butcher	John Charles	1933-	249
Cash	Jeff R.	współcześnie	253
Cauchy	Augustin Louis	1789-1857	67, 201, 225, 283, 287, 289
Cayley	Artur	1821-1895	202, 289
Choleski	André-Louis	1875-1918	75
Cohn	Arthur	1894–1940	192, 193
Cotes	Roger	1682-1716	134, 135, 136, 137, 139
Curtiss	Charles F.	współcześnie	238
Czebyszew	Pafnucy Lwowicz	1821-1894	104, 114, 115, 116, 117, 235
Darboux	Jean Gaston	1842-1917	147, 291
Dormand	John R.	współcześnie	253, 270
Durand	Émile	1911-1999	191, 192
Euklides	z Aleksandrii	365 p.n.e. - 300 p.n.e.	212, 221, 282
Euler	Leonhard	1701-1783	232, 235, 236, 237, 239-248, 253, 254, 260, 262, 264, 265, 277, 278
Fehlberg	Erwin	1911-1990	252

Gauss	Carl Friedrich	1777-1855	47-70, 75, 136-138, 176, 180, 203, 207, 235, 239
Gerszgorin	Siemion Aranowicz	1901-933	194
Gibbs	Josiah Willard	1839-1903	104, 105
Gram	Jørgen Pedersen	1850-1916	82
Hamilton	William Rowan	1805-1865	202, 289
Hermite	Charles	1822-1901	106, 111
Hessenberg	Karl	1904-1959	66, 209, 210, 285
Heun	Karl	1859-1929	238
Hilbert	David	1861-1943	83
Hirschfelder	Joseph Oakland	1911-1990	238
Hooke	Robert	współcześnie	179
Horner	William George	1786-1837	42, 43, 112, 114, 115, 116, 183
Householder	Alston Scott	1904-1993	210
Jacobi	Carl Gustav Jacob	1804-1851	35, 175, 180, 190, 238, 292
Jeeves	Terry Allen	współcześnie	179
Jordan	Marie Ennemond Camille	1838-1922	75, 290
Karp	Allan H.	współcześnie	253
Kerner	Immo O.	współcześnie	191, 192
Kutta	Martin Wilhelm	1867-1944	247, 248, 249, 252, 254, 270
Lagrange	Joseph Louis	1736-1813	93-94, 109, 112, 134-135, 168, 293-294
Landau	Edmund Georg Hermann (Yehezkel)	1877-1938	42, 294
Lebesgue	Henri	1875-1941	117
Lehmer	Derrick Henry	1905-1991	192-195
Lipschitz	Rudolf Otto Sigismund	1832-1903	226, 227, 228, 251, 292
Maclaurin	Colin	1698-1746	294
Maehly	Hans J.	współcześnie	188, 189
Mead	Roger	współcześnie	179
Moulton	Forest Ray	1872-1952	259, 260, 262, 263, 270
Müller	David Eugene	1924-2008	192

Indeks nazwisk

Nelder	John Ashworth	1924-2010	179
Newton	Isaac	1642-1726?/27?	94-95, 109, 134-136, 149-152, 154-156, 161-165, 167, 175-179, 185-186, 188-190, 250, 268
Nourein	Amal	współcześnie	192
Prince	Pete J.	współcześnie	253, 270
Raphson	Joseph	1648-1715	149-150, 154, 156, 161-165, 167, 175-179, 185-186, 188-190
Rayleigh	John William Strutt	1842-1919	210
Remez	Ewgenij Jakowlewicz	1895-1975	88
Richardson	Lewis Fry	1881-1953	119, 126-129, 140, 268
Rolle	Michel	1652-1719	99, 293
Romberg	Werner	1909-2003	140, 268
Rosenbrock	Howard Harry	1920-2010	179
Rouché	Eugène	1832-1910	194
Ruffini	Paolo	1765-1822	280
Runge	Carl David Tolmé	1856-1927	98, 100, 103, 108, 110, 247-149, 252, 254, 270, 273
Schur	Issai	1875-1941	192-195
Seidel	Philipp Ludwig	1821-1896	180
Shampine	Lawrence F.	współcześnie	253, 270
Taylor	Brook	1685-1731	119-122, 127, 145, 150, 175, 236-237, 243, 255, 293-294
Vandermonde	Alexandre-Théophile	1735-1796	91-93, 109, 136
Weierstrass	Karl	1815-1897	86-87, 191
Wilkinson	James Hardy	1919-1986	40



ISBN 978-83-7283-877-3