DAVID KINCAID
WARD CHENEY

analiza numeryczna

W przekładzie
i pod redakcją
STEFANA PASZKOWSKIEGO

Wydawnictwa Naukowo-Techniczne
Warszawa

Dane o oryginale

Numerical Analysis. Mathematics of Scientific Computing

Third Edition

David Kincaid, Ward Cheney

The University of Texas at Austin

Recenzent wydania polskiego *prof. dr hab. Krystyna Ziętak*

Okładkę i strony tytułowe projektował *Wojciech J. Steifer*
Redaktor techniczny *Grażyna Miazek*
Korekta *Joanna Ożóg, Gabriela Szpunar*
Skład i łamanie preTEXt

Wydawnictwa Naukowo-Techniczne
00-048 Warszawa, ul. Mazowiecka 2/4
tel. (0-22) 826 72 71, e-mail: wnt@wnt.pl
www.wnt.pl

ISBN 83-204-3078-X

Książkę poświęcamy naszym Rodzicom,
Sarah i Robertowi B. Kincaidom
oraz (in memoriam) Carleton i Elliottowi Cheneyom

Spis treści

Od tłumacza i redaktora przekładu

Dostępne w Polsce podręczniki analizy numerycznej, jak (włączone do bibliografii) książki Ralstona, Dahlquista i Björcka, Stoera i Bulirscha oraz Dryi i Jankowskich, ukazały się ponad dwadzieścia lat temu. Nowsza jest książka Kiełbasińskiego i Schwetlicka, ale ta dotyczy tylko metod numerycznych algebry liniowej. Potem ukazywały się co najwyżej skrypty wydawane przez niektóre uczelnie i ukierunkowane głównie na dydaktykę. Tymczasem analiza numeryczna jest dziedziną matematyki nie tylko stale stosowaną, ale i rozwijającą się nieustannie, o czym świadczą chociażby coraz to nowe czasopisma naukowe jej poświęcone. Dlatego dobrze się stało, że Wydawnictwa Naukowo-Techniczne postanowiły wydać niniejszy podręcznik.

Jest to przekład trzeciego wydania amerykańskiego, w którym można znaleźć – w porównaniu z podręcznikami wymienionymi wyżej – sporo nowości. Wiele tematów ujęto inaczej, na co innego położono nacisk. Jak zwykle, dobór materiału i sposób jego prezentacji wynika z indywidualnych poglądów i zainteresowań autorów. Nawet tak obszerna książka nie może choćby w skrócie objąć wszystkich najważniejszych działów i metod analizy numerycznej. Żałuję na przykład, że teoria i praktyka przyspieszania zbieżności – tematu ważnego w całej analizie numerycznej – zasłużyły tylko na skąpe wzmianki. Mam nadzieję, że w przyszłości uda się zapełnić tę i inne luki. Na razie zaś, jako tłumacz i redaktor książki, spodziewam się, że jej lektura będzie pożyteczna i dla studentów różnych dyscyplin, i dla osób, które w swej pracy zawodowej posługują się metodami numerycznymi.

Tłumacząc książkę, usunąłem bardzo dużo drobnych usterek różnego typu. Jako redaktor przekładu wprowadziłem – za zgodą autorów – sporo zmian nienaruszających oczywiście autorskiej koncepcji książki, ale mających na celu ułatwienie jej lektury i uwzględniających wiedzę polskich czytelników. Skróciłem więc lub pominąłem szczególnie elementarne rozważania, zbędne dla czytelników ze standardowym poziomem wiedzy matematycznej. Dowody kilku twierdzeń zastąpiłem prostszymi lub bardziej

naturalnymi. W kilku miejscach przestawiłem fragmenty tekstu, np. poprzedziłem twierdzenie używanymi w nim lematami.

Przejrzałem krytycznie listy zadań (a jest ich w książce wyjątkowo dużo!) do poszczególnych podrozdziałów, usunąłem zadania powtarzające się lub wyraźnie banalne, dostosowałem – gdzie było to wskazane – porządek zadań do kolejności wprowadzanych pojęć, metod i twierdzeń. Aby ułatwić czytelnikom lekturę, wprowadziłem w obrębie każdego podrozdziału wspólną numerację twierdzeń, wniosków, przykładów itd. Oryginalny podrozdział 6.11, dający tylko bardzo ogólnikowe wiadomości o ułamkach łańcuchowych, zastąpiłem za zgodą autorów nieco szerszym podrozdziałem o aproksymacji wymiernej, który zawiera wstępne wiadomości nie tylko o takich ułamkach, ale również o interpolacji wymiernej i aproksymacji Padégo.

Bibliografia w oryginale zawiera ok. 360 pozycji niecytowanych w tekście. Usunąłem je, bo w wielu przypadkach na podstawie tytułów nie można nawet ustalić, z jakimi dziedzinami te prace się wiążą. Przed przedmową autorów umieściłem krótką listę symboli używanych dalej, a wyjaśnianych w różnych miejscach książki.

W przekładzie pominąłem dodatek zawierający komentowaną listę adresów kilkudziesięciu stron z informacjami o towarzystwach naukowych, czasopismach, programach, wykładach itd. z dziedziny analizy numerycznej; informacje na ten temat zawiera przypis do przedmowy autorów. Z podobnych powodów usunąłem też z podrozdz. 2.1 fragment o hipotetycznym komputerze Marc-32, a z dalszych partii książki wszelkie wzmianki o nim (zresztą w końcowych rozdziałach były one coraz rzadsze). Tam, gdzie jest to potrzebne, czytelnicy są informowani o precyzji stosowanej arytmetyki.

Jedną z najważniejszych metod numerycznych algebry liniowej jest metoda Cholesky'ego rozkładu macierzy na czynniki. Czytelnicy znający tę metodę z innych podręczników zauważyli zapewne, że nie cytuje się w nich oryginalnej pracy Cholesky'ego. Profesor Claude Brezinski uprzejmie przekazał mi wyniki swych poszukiwań w bibliotekach i archiwach. Teraz już wiadomo, kim był André Louis Cholesky oraz gdzie i kiedy opublikował swą metodę; notabene jest bardzo prawdopodobne, że pochodził on z rodziny polskich emigrantów Cholewskich, osiadłych we Francji w XIX w. Metoda ta w wielu polskich publikacjach nosi imię Tadeusza Banachiewicza, który wynalazł ją niezależnie, ale znacznie później.

Profesor Krystynie Ziętak zawdzięczam wiele istotnych uwag dotyczących zarówno oryginału, jak i przekładu, głównie rozdziałów 2–5. Żałuję, że z równie istotnych powodów nie wszystkie takie uwagi mogłem uwzględnić. Kilka osób pomogło mi rozstrzygnąć wątpliwości terminologiczne i inne; byli to: prof. Edward Neuman, dr Andrzej Wakulicz i mgr inż. Tymon

Godzwon. Dzięki uprzejmości personelu Biblioteki Instytutu Matematycznego PAN w Warszawie udało mi się sprawdzić kilka istotnych informacji. Należne podziękowania składam również mgr Liliannie Szymańskiej kierującej Redakcją Matematyki i Fizyki WNT, mgr Małgorzacie Jachymek z tejże Redakcji oraz p. Grażynie Miazek z Redakcji Technicznej WNT za współpracę przy nadaniu przekładowi jego ostatecznej, oby optymalnej postaci.

Wrocław, kwiecień 2005 *Stefan Paszkowski*

Oznaczenia i konwencje

$\mathbb{N}$	zbiór liczb naturalnych $1, 2, 3, \ldots$
$\mathbb{Z}$	zbiór liczb całkowitych $\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$
$\mathbb{R}$	zbiór liczb rzeczywistych
$\mathbb{C}$	zbiór liczb zespolonych
$\Re z$, $\Im z$	część rzeczywista i urojona liczby zespolonej z
$\mathbb{R}^n$	przestrzeń n-wymiarowa punktów $(x_1, x_2, \ldots, x_n)$ o współrzędnych rzeczywistych (wektorów o takich składowych)
$\mathbb{C}^n$	przestrzeń n-wymiarowa punktów $(x_1, x_2, \ldots, x_n)$ o współrzędnych zespolonych (wektorów o takich składowych)
$\times$	symbol iloczynu kartezjańskiego
$\mathcal{O}, o$	duże i małe o; symbole używane w porównywaniu elementów ciągów lub wartości funkcji
$:=$	równe z definicji
$\mapsto$	symbol używany w definicji funkcji (odwzorowania); np. $x \mapsto x^2 - 1$
exp	symbol funkcji wykładniczej: $\exp x := e^x$
C, C^m, C^∞	przestrzeń funkcji zmiennej rzeczywistej (odpowiednio: ciągłych, mających m-tą pochodną ciągłą, mających wszystkie pochodne ciągłe)
A^T	macierz transponowana względem macierzy A
$\langle u, v \rangle$	iloczyn skalarny wektorów u i v
δ_{jk}	*delta Kroneckera*, równa 1 dla $j = k$ i 0 dla $j \neq k$
sgn	*signum* (znak): $\operatorname{sgn} x := -1, 0, 1$ odpowiednio dla $x < 0$, $x = 0$ i $x > 0$
$(x)_n$	*symbol Pochhammera*: $(x)_n := x(x+1)\ldots(x+n-1)$ dla $n \in \mathbb{N}$, $(x)_0 := 1$

$m_{10}c$ symbol liczby $m \times 10^c$ (m – ułamek dziesiętny skończony, c – liczba całkowita)

$\sum_{k=m}^{n} s_k := 0$ dla $m = n - 1$

$\prod_{k=m}^{n} s_k := 1$ dla $m = n - 1$

Przedmowa

Książka powstawała przez wiele lat z notatek do wykładów z matematyki i informatyki na ostatnim roku studiów podstawowych lub na studiach magisterskich na University of Texas w Austin. Te wykłady wprowadzały studentów w dziedzinę algorytmów i metod najczęściej potrzebnych w obliczeniach naukowych. Zwracaliśmy uwagę zarówno na matematyczne podstawy tych metod, jak i na ich algorytmiczne aspekty. Słuchaczami byli studenci studiów podstawowych z matematyki, nauk technicznych lub ścisłych i informatyki oraz studenci różnych typów studiów magisterskich. Fragmenty książki stanowiły też podstawę wielu wykładów poświęconych poszczególnym działom analizy numerycznej, takim jak rozwiązywanie numeryczne równań różniczkowych, metody numeryczne algebry liniowej i teoria aproksymacji. Nasze podejście polegało zawsze na traktowaniu przedmiotu z matematycznego punktu widzenia; pokazywaliśmy szeroką paletę twierdzeń, dowodów i ciekawych pomysłów. Wynika stąd wiele procedur numerycznych i intrygujących problemów analizy numerycznej. Oczywiście, motywował nas obszar praktycznych zastosowań tej dziedziny, narzucający wybór tematów i sposób ich ujęcia. Tak na przykład, w pewnych działach bardziej pouczające jest rozważanie podstaw teoretycznych niż próba szczegółowego analizowania algorytmów. W innych przypadkach jest na odwrót i studenci wiele wynoszą, programując samodzielnie proste algorytmy i testując je; popieramy jednak bez zastrzeżeń korzystanie ze starannie sprawdzonego oprogramowania, na przykład z bibliotek programów, dotyczącego problemów wynikających z zastosowań.

Treść tej książki i treść naszego bardziej elementarnego podręcznika *Numerical Mathematics and Computing* (wyd. 4, Brooks/Cole) częściowo się pokrywają. Jednak wymieniony podręcznik jest adresowany do studentów mających skromniejsze przygotowanie matematyczne (a często także mniej zapału do studiowania teoretycznych aspektów przedmiotu). Zestaw tematów jest tam inny, a żadnego z nich nie analizuje się zbyt głęboko. Niniejsza książka zaś jest przeznaczona do wykładu dającego bardziej akademickie ujęcie przedmiotu; pewne tematy są opisane szczegółowo. Tu i ówdzie

poruszamy zagadnienia, które nie znalazły wcześniej swego miejsca w standardowych podręcznikach na tym poziomie. Do tej kategorii należą metody wielosiatkowe, procedury interpolacji funkcji wielu zmiennych, metody homotopii (lub kontynuacji), równania różniczkowe z opóźnionym argumentem i optymalizacja.

Algorytmy w niniejszej książce wyrażamy w symbolice zawierającej nie tylko wzory, ale i dodatkowe elementy. Czytelnik może łatwo przetłumaczyć taki algorytm na dowolny typowy język programowania. Sądzimy, że studenci najlepiej nauczą się metod numerycznych i zrozumieją je, widząc, jak algorytmy wynikają z teoretycznych rozważań, oraz pisząc i testując programy komputerowe. Zapewne nie będą one zawierały wszystkich skomplikowanych procedur i wyszukanych sposobów kontroli, jakimi charakteryzują się programy biblioteczne. Przykłady bibliotek oprogramowania można znaleźć w dodatku A[1]. W wielu zastosowaniach odwołanie się do tych bibliotek jest znacznie bardziej sensowne niż pisanie własnych programów.

Ważną częścią składową książki (i niezbędną dla celów dydaktycznych) jest obfitość zadań do rozwiązywania przez studentów; znajdą tu oni dwa rodzaje zadań: analityczne i komputerowe. Te ostatnie z kolei dzielą się na takie, gdzie student ma napisać własny program i takie, gdzie należy zastosować istniejące oprogramowanie. Uważamy, że oba rodzaje praktyki komputerowej są konieczne. Z jednej strony, użycie cudzego programu nie zawsze jest banalnym ćwiczeniem, nawet wtedy, gdy jest on dobrze udokumentowany, jak to bywa w dużych bibliotekach czy pakietach. Z drugiej strony, studenci zdobywają głębszą wiedzę o algorytmie, jeśli go sami programują i testują, niż wtedy, gdy tylko korzystają z programu bibliotecznego. W większości przypadków zadania komputerowe wymagają stosowania arytmetyki zmiennopozycyjnej z liczbami co najmniej 32-bitowymi.

Oprogramowanie opisane w książce, erratę do niej i pewne pomoce dydaktyczne można znaleźć w Internecie. Wydawca[2] udostępnia też zbiór rozwiązań zadań wykładowcom, którzy zechcą oprzeć się na tej książce.

Trzecie wydanie zawiera nowe zadania, inny jest też ich układ. Usunęliśmy wszystkie błędy zauważone w poprzednich wydaniach. Bibliografia została zaktualizowana[3]. Stronę graficzną zaprojektowano na nowo. Wpro-

[1] Nie ma go w polskim wydaniu, autorzy uznali bowiem, że obecnie stał się on już zbędny wobec łatwości wyszukiwania w Internecie potrzebnych informacji. Tłumacz pominął więc kilka następnych wzmianek o tym dodatku (*przyp. tłum.*).

[2] Chodzi o wydawcę oryginału, tzn. wydawnictwo BROOKS/COLE (*przyp. red. WNT*).

[3] W polskim wydaniu rozszerzono ją o pozycje polskich autorów, przekłady na język polski kilku ważnych prac pominiętych przez autorów i pewne książki wydrukowane po ukazaniu się trzeciego wydania oryginału. W bibliografii i odwołaniach do niej te dodatkowe pozycje są oznaczone gwiazdką $*$, poprzedzającą rok wydania (*przyp. tłum.*).

wadziliśmy wiele innych ulepszeń. W szczególności w tym nowym wydaniu dodaliśmy rozdział o optymalizacji z takimi tematami, jak metody spadku, algorytmy przybliżania kwadratowego, algorytm Neldera-Meada, symulowane wyżarzanie, algorytmy genetyczne, optymalizacja Pareto i programowanie wypukłe.

Standardowy wykład jednosemestralny można oprzeć na wybranych fragmentach rozdziałów 1–4 i 6–8. Wykład dwusemestralny może obejmować fragmenty rozdziałów 1–9 i dodatkowe interesujące tematy. Rozdziały 4 i 5 można uważać za niezależny od poprzednich krótki podręcznik numerycznej algebry liniowej. Ze względu na obszerny zakres tematów lektura pewnych podrozdziałów stawia czytelnikom większe wymagania. Na ogół podrozdziały takie zamieściliśmy na końcu rozdziałów, aby na początku lektury nie zniechęcać czytelnika, który według własnego uznania może trudniejsze partie pominąć.

Podziękowania

Z radością wyrażamy naszą wdzięczność wielu osobom, które towarzyszyły nam przy pisaniu tej książki.

Pierwsze wydanie

Pomoc administracyjną zapewnili nam: Sheri Brice, Margaret Combs, Jan Duffy, Katherine Mueller i Jenny Tsao z University of Texas w Austin. Przede wszystkim chcemy podkreślić rolę Margaret Combs z Wydziału Matematyki, która przepisywała w TEX-u niezliczone wersje każdego podrozdziału i cierpliwie przygotowywała nowe, gdy tylko były potrzebne jako skrypty w kolejnych latach. Żaden problem techniczny nie był dla niej zbyt trudny, tak świetnie opanowała umiejętność dopasowywania makr TEX-a do nietypowych celów. Jest tu właściwe miejsce na nasze publiczne podziękowania dla Donalda Knutha za to, co zrobił dla społeczności uczonych, tworząc system składu drukarskiego TEX.

Thomas A. Atchison, Frederick J. Carter, Philip Crooke, Jim D'Archangelo, R.S. Falk, J.R. Hubbard, Patrick Lang, Giles Wilson Maloof, A.K. Rigler, F. Schumann, A.J. Worsey i Charles Votaw byli dociekliwymi krytykami wstępnych wersji rękopisu; doceniamy zgłaszane przez nich sugestie. Jesteśmy też winni podziękowania kilku osobom za pomoc techniczną i krytyczne przejrzenie rękopisu; byli to: Victoria Hunter, Carole Kincaid, Tad Liszka, Rio Hirowati Shariffuddin i Laurette Tuckerman. David Young służył zawsze uwagą i radą. Bardzo nam też pomogli studenci starszych lat, którzy prowadzili ćwiczenia do naszych wykładów. Oto oni: David Bruce, Nai-ching Chen, Ashok Hattangady, Ru Huang, Wayne Joubert, Irina Mukherjee, Bill Nanry, Tom Oppe, Marcos Raydan, Malathi Ramdas, John Respess, Phien Tran, Linda Tweedy i Baba Vemuri. Ściśle z nami współpracowali i służyli pomocą wydawcy i redaktorzy techniczni z Brooks/Cole Publishing Company. W szczególności z przyjemnością dziękujemy za wsparcie Jeremy'emu Hayhurstowi i Marlene Thom. Stacey Sawyer z firmy Sawyer

and Williams odpowiadał za staranne przygotowanie maszynopisu, a Ralph Youngen z Amerykańskiego Towarzystwa Matematycznego zapewnił pomoc techniczną i nadzór nad przekształceniem plików TEX-owych w końcową drukowaną postać.

Drugie wydanie

Recenzentami, którym jesteśmy wdzięczni za cenne uwagi, byli: Dan Boley z University of Minnesota, Min Chen z Pennsylvania State University, John Harper z University of Rochester, Ramon Moore z Ohio State University, Yves Nievergelt z Eastern Washington University i Elinor Velasquez z University of California-Berkeley. Szczególnie dziękujemy Ronovi Boisvertowi za wyjaśnienie nam różnych kategorii oprogramowania matematycznego i przykłady podane w dodatku. Chcemy też podziękować wszystkim, którzy zadali sobie trud i przekazali nam uwagi lub poprawki do pierwszego wydania. Wśród tych osób są: Victor M. Afram, Roger Alexander, A. Awwal, Carl de Boor, T. P. Brown, James Caveny, George J. Davis, Hakan Ekblom, Mariano Gasca, Bill Gearhart, Patrick Goetz, Gary L. Gray, Bob Gregorac, Katherine Hua Guo, Cecilia Jea, Liz Jessup, Grant Keady, Baker Kearfott, Junjiang Lei, Teck C. Lim, Julio Lopez, C. Lu, Taketomo Mitsui, Irina Mukherjee, Teresa Perez, Robert Piche, Sherman Riemenschneider, Maria Teresa Rodriquez, Ulf Roennow, Larry Schumaker, Wei-Chang Shann, Christopher J. van Wyk, Kang Zhao i Mark Zhou.

Trzecie wydanie

Chcemy podziękować wszystkim, którzy przekazali nam propozycje i poprawki do drugiego wydania. Byli to: Eyal Arian, Carl de Boor, Yung-Ming Chang, Antonella Cupillari, Paul Eenigenburg, Leopoldo P. Franca, Henry Greenside, R. J. Gregorac, Scott A. King, Robert Piche, N. Shamsundar, Topi Urponen i Yuan Xu. Jesteśmy wdzięczni Patrickowi Goetzowi, Shashankowi Khandelwalowi i przede wszystkim Durene Ngo, którzy pomagali przygotować nowe wydanie.

Nasze wyrazy podziękowania należą się też następującym instytucjom za pomoc techniczną i stworzenie doskonałych warunków pracy na komputerach: Center for Numerical Analysis, Texas Institute of Computational and Applied Mathematics oraz dwóm wydziałom University of Texas w Austin: Computer Sciences Department i Mathematics Department.

Wydawca Brooks/Cole-Thomson Learning i jego redaktorzy techniczni, a szczególnie Bob Pirtle, Janet Hill i Molly Nance, służyli nam wszelką pomocą, gdy przygotowywaliśmy to poprawione wydanie. Don DeLand, Leslie Galen, Joe Albrecht z firmy Integre Technical Publishing Company wykonali świetną robotę, za którą im dziękujemy.

Będziemy wdzięczni czytelnikom, którzy zechcą się z nami skontaktować, za wszystkie komentarze, sugestie, pytania, uwagi krytyczne lub poprawki.

David Kincaid
Ward Cheney

Czym jest analiza numeryczna?

Analiza numeryczna obejmuje *tworzenie*, *badanie* i *analizę* algorytmów, których celem jest otrzymywanie rozwiązań numerycznych różnorodnych zadań matematycznych. Często analizę numeryczną nazywa się *matematyką obliczeń naukowych*[1].

Badane przez nas algorytmy są nieuchronnie przeznaczone do stosowania na szybkich komputerach i dlatego pewien decydujący etap musi poprzedzać rozwiązanie zadania: trzeba napisać *program*, aby przekazać w tej postaci algorytm komputerowi. To jest oczywiście nietrywialny problem, ale mamy teraz do wyboru tyle komputerów i języków programowania, że programowanie pozostaje poza obrębem analizy numerycznej w ścisłym sensie tego terminu.

Poza rozwiązywaniem numerycznym zadań matematycznych komputery mają oczywiście wiele innych zastosowań, jak komunikacja, tworzenie wielkich baz danych, gry, „surfowanie" w sieci, pisanie powieści, rachunkowość itd. Rozwiązywanie zadań *matematycznych* numerycznie na komputerze – to *obliczenia naukowe*. Tworzenie odpowiednich algorytmów (procedur) i badanie ich własności – to matematyka obliczeń naukowych.

Konstrukcja algorytmu jest nierzadko stymulowana przez konstruktywny dowód w matematyce. W klasycznej analizie są często stosowane metody niekonstruktywne, ale te na ogół nie prowadzą do algorytmów. Na przykład, twierdzenia dotyczące istnienia i jednoznaczności dowodzi się przypuszczając, że są one fałszywe i dochodząc poprzez logiczne rozumowanie do sprzeczności. Nie każdy jednak konstruktywny dowód prowadzi do efektywnego algorytmu. Trudność, która wtedy się pojawia, bierze się stąd, że rozwiązanie *analityczne* danego zadania może być zupełnie odmienne od rozwiązania *numerycznego*. To pierwsze może być bezużyteczne, gdy prowadzi do wolnej zbieżności albo zmusza do długotrwałych obliczeń.

---

[1] W polskiej literaturze matematycznej ta nazwa nie jest używana, natomiast dawne podręczniki analizy numerycznej (np. wymienione w dodatkowej bibliografii od tłumacza) miały w tytule *metody numeryczne* (*przyp. tłum.*).

Jako przykład luki między twierdzeniem o istnieniu i rozwiązaniem numerycznym zadania rozważmy wszechobecne równanie macierzowe $Ax = b$. Wiemy, że ma ono jedyne rozwiązanie, jeśli tylko macierz A jest nieosobliwa. To jednak jest niewielką pociechą, gdy stajemy przed wielkim układem liniowym zawierającym dane empiryczne i chcemy znaleźć numerycznie jego *przybliżone* rozwiązanie.

W tej książce będziemy z reguły zaczynać każdy temat od tych podstawowych zadań matematycznych, które pojawiają się często w zastosowaniach praktycznych. Aby dojść do algorytmu rozwiązania takiego zadania, będziemy musieli przejść przez pewne rozważania analityczne. Algorytmy są zwykle dane w postaci programu napisanego w pewnym fikcyjnym języku programowania[2)].

Na końcu może być podana dodatkowa analiza algorytmu, która ma ułatwić zrozumienie jego własności, takich jak zbieżność lub odporność na błędy zaokrągleń. Ta analiza błędów może przybierać formę analizy *bezpośredniej* lub analizy *pozornych równoważnych zaburzeń*[3)].

Za każdym ważnym zadaniem matematycznym stoją zawsze zastosowania fizyczne. Przykładem może być zadanie przewodnictwa cieplnego. Temperaturą w metalowym pręcie, dla różnych warunków brzegowych, rządzą równania matematyczne, które muszą być spełnione w każdym punkcie i w każdej chwili. Zasadniczym równaniem może być *równanie przewodnictwa cieplnego*

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t}.$$

Jest to równanie różniczkowe cząstkowe typu parabolicznego, liniowe, drugiego rzędu. Opisuje ono rozchodzenie się ciepła w pręcie przy pewnych założeniach o rzeczywistym zadaniu fizycznym. W równaniu x jest zmienną przestrzenną, t oznacza czas, a $u = u(x,t)$ jest temperaturą. Rozwiązując zadanie modelowe na komputerze, *dyskretyzujemy* obszar czasoprzestrzeni wprowadzając siatkę punktów i szukamy rozwiązania w każdym z nich. Pochodne cząstkowe w równaniu można aproksymować za pomocą różnic skończonych, np. tak:

$$\frac{\partial v(x,t)}{\partial t} \approx \frac{1}{k}[v(x,t+k) - v(x,t)],$$
$$\frac{\partial^2 v(x,t)}{\partial x^2} \approx \frac{1}{k^2}[v(x+h,t) - 2v(x,t) + v(x-h,t)].$$

2) W oryginale tak wyrażony program jest nazywany *pseudocode*, ale w polskiej terminologii taki termin wyszedł już z użycia (*przyp. tłum.*).

3) W oryginale jest to odpowiednio *forward* i *backward error analysis*. Oba te rodzaje analizy błędów definiuje i stosuje Wilkinson [1963]; zob. też Higham [2002] (*przyp. tłum.*).

Wielkości h i k są tu odległościami sąsiednich punktów siatki, odpowiednio w kierunku t i w kierunku x. Zmiana symbolu funkcji na v przypomina, że zamiast pierwotnego modelowego zadania rozwiązujemy jego przybliżenie. Po zastąpieniu pochodnych cząstkowych podanymi wyrażeniami przybliżonymi i uproszczeniach dochodzimy do równania liniowego w każdym punkcie (x_i, t_j) siatki. Używając prostszego oznaczenia v_{ij} zamiast $v(x_i, t_j)$, wyrażamy to równanie w postaci

$$v_{i,j+1} = sv_{i-1,j} + (1 - 2s)v_{ij} + sv_{i+1,j},$$

gdzie $s = k/h^2$. Dzięki niemu możemy tworzyć rozwiązanie numeryczne, posuwając się krok po kroku w kierunku t. Taką procedurę nazywamy *jawną*, gdyż nowe wartości $v_{i,j+1}$ wyrażają się jawnym wzorem przez poprzednie wartości $v_{i-1,j}$, v_{ij} i $v_{i+1,j}$. Metoda jest bardzo elegancka i trudno przewidywać jakieś kłopoty w jej zastosowaniu. A jednak zarówno analiza jak i doświadczenie numeryczne dowodzą, że metoda ma fatalną wadę! Dlatego przechodzimy do metody *niejawnej*. Polega ona na wyznaczeniu wszystkich wartości jednocześnie ze szczególnego typu układu liniowego

$$V_{j+1} = AV_j.$$

A jest tu pewną macierzą trójprzekątniową, a $V_j = (v_{j1}, v_{j2}, \ldots, v_{jn})$. Dla każdej takiej metody trzeba zbadać jej stabilność, dzięki czemu określamy dopuszczalne wielkości dla h i k i ustalamy rodzaj zbieżności. Tu metoda jawna sprawuje się kiepsko. Szczegóły można znaleźć w rozdz. 9.

ROZDZIAŁ 1

Narzędzia matematyczne

1.0. Wstęp

Rozdział zaczyna się od przeglądu kilku ważnych tematów analizy matematycznej, które będą potrzebne w następnych rozdziałach. Zachęcamy czytelników, aby bez wahania pominęli te tematy, które są już im znane. Zapewne niektórzy z nich przejdą od razu do rozdz. 2.

1.1. Podstawowe pojęcia i wzór Taylora

Zaczynamy od przeglądu pewnych ważnych pojęć analizy. Ktoś może zapytać: Dlaczego trzeba rozważać takie kwestie, skoro interesują nas przede wszystkim obliczenia naukowe i algorytmy numeryczne? Dobra znajomość podstawowych pojęć matematycznych jest potrzebna dla zrozumienia większości algorytmów numerycznych. Wzór Taylora w różnych wariantach odgrywa fundamentalną rolę w wielu procedurach numerycznych i stanowi doskonały punkt wyjścia do studiowania obliczeń naukowych, gdyż nie odwołuje się do zaawansowanych pojęć matematycznych.

Granica, ciągłość i pochodna

Jeśli f jest funkcją zmiennej rzeczywistej i ma wartości rzeczywiste, to jej *granica* w punkcie c jest określona (jeśli istnieje) w następujący sposób:

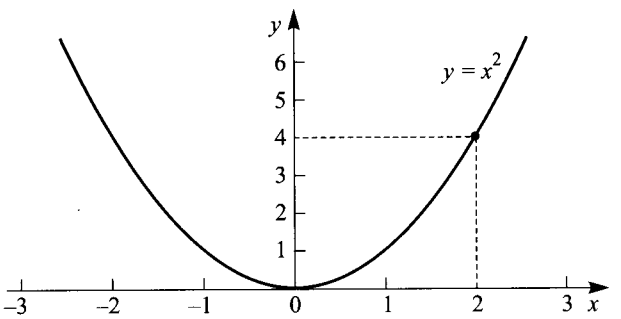
równość

$$\lim_{x \to c} f(x) = L$$

oznacza, że dla każdego ε dodatniego istnieje takie δ dodatnie, iż odległość między $f(x)$ i L jest mniejsza od ε, jeśli odległość między x i c jest dodatnia i mniejsza od δ, czyli

$$|f(x) - L| < \varepsilon, \quad \text{jeśli} \quad 0 < |x - c| < \delta.$$

Jeśli żadna liczba L nie ma takiej własności, to granica funkcji f w c nie istnieje.



RYS. 1.1. $y = f(x) := x^2$

Korzystając z definicji granicy, można sprawdzić, że np.

$$\lim_{x \to 2} x^2 = 4$$

(rys. 1.1). Natomiast funkcja

$$g(x) := \frac{|x|}{x} = \begin{cases} 1, & \text{jeśli} \quad x > 0 \\ -1, & \text{jeśli} \quad x < 0 \end{cases}$$

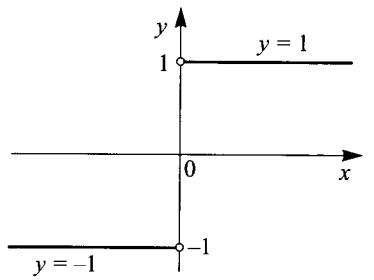
(rys. 1.2) nie ma granicy w punkcie 0 [1)].

Jeśli f jest określona tylko w pewnym podzbiorze X osi rzeczywistej, to definicję granicy modyfikujemy tak, że $|f(x) - L| < \varepsilon$, jeśli tylko $x \in X$ i $0 < |x - c| < \delta$.

Mówimy, że funkcja f jest *ciągła* w c, jeśli

$$\lim_{x \to c} f(x) = f(c).$$

[1)] Ma ona jednak w punkcie 0 granicę *lewostronną*, równą -1, i *prawostronną*, równą 1 (*przyp. tłum.*).



RYS. 1.2. $y = g(x) := |x|/x$

Tak więc funkcja $f(x) = x^2$ jest ciągła w punkcie 2, natomiast funkcja $|x|/x$ nie jest ciągła w 0 i to niezależnie od tego, jak byśmy ją określili w tym punkcie. Wynika to z poprzednich uwag.

Poniższe twierdzenie, intuicyjnie oczywiste, wyraża tzw. *własność Darboux funkcji ciągłych*:

TWIERDZENIE 1.1.1. *Funkcja ciągła f w przedziale $[a, b]$ przyjmuje w nim wszystkie wartości zawarte między $f(a)$ i $f(b)$.*

Pochodną funkcji f w c (jeśli istnieje) określamy wzorem

$$f'(c) := \lim_{x \to c} \frac{f(x) - f(c)}{x - c}.$$

Ponieważ ta granica istnieje nie dla każdej funkcji i nie w każdym punkcie, więc i pochodna nie zawsze istnieje. Jeśli dla f istnieje $f'(c)$, to mówimy, że f jest *różniczkowalna* w c. W takim przypadku f jest na pewno ciągła w c. Twierdzenie przeciwne nie jest jednak prawdziwe. Jeśli na przykład

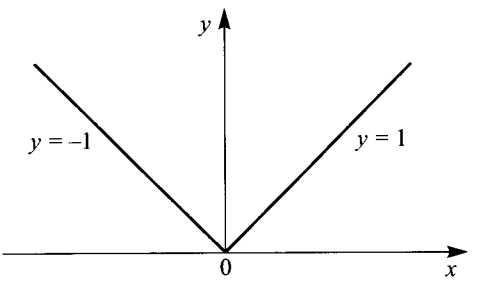
$$f(x) = |x|,$$

to $f'(0)$ nie istnieje; zob. wykres tej funkcji na rys. 1.3.

Zbiór wszystkich funkcji ciągłych na całej prostej rzeczywistej $\mathbb{R}$ oznaczamy $C(\mathbb{R})$. Zbiór funkcji, dla których pochodna f' jest wszędzie ciągła, oznaczamy $C^1(\mathbb{R})$. Tę przestrzeń tworzą więc wszystkie funkcje różniczkowalne na całej prostej rzeczywistej. Ponieważ z różniczkowalności funkcji w punkcie wynika jej ciągłość w tymże punkcie, więc $C^1(\mathbb{R}) \subset C(\mathbb{R})$. Zbiór $C^1(\mathbb{R})$ jest podzbiorem *właściwym* zbioru $C(\mathbb{R})$, gdyż liczne funkcje ciągłe nie są różniczkowalne; przykładem jest funkcja $f(x) = |x|$.

Symbolem $C^2(\mathbb{R})$ oznaczamy zbiór wszystkich funkcji, których druga pochodna jest wszędzie ciągła. Rozumując jak poprzednio, wnioskujemy, że

$$C^2(\mathbb{R}) \subset C^1(\mathbb{R}) \subset C(\mathbb{R}).$$



RYS. 1.3. $y = f(x) := |x|$

Określony właśnie zbiór jest podzbiorem właściwym zbioru $C^1(\mathbb{R})$, bo istnieją funkcje różniczkowalne tylko raz; tak jest dla $f(x) = x^2 \sin(1/x)$.

Podobnie definiujemy, dla dowolnej liczby naturalnej n, $C^n(\mathbb{R})$ jako zbiór wszystkich funkcji mających n-tą pochodną ciągłą. Na koniec, $C^\infty(\mathbb{R})$ jest zbiorem funkcji mających wszystkie pochodne ciągłe. Jest

$$C^\infty(\mathbb{R}) \subset \ldots \subset C^2(\mathbb{R}) \subset C^1(\mathbb{R}) \subset C(\mathbb{R}).$$

Znaną funkcją z $C^\infty(\mathbb{R})$ jest $f(x) = e^x$.

W ten sam sposób określamy $C^n[a,b]$ jako zbiór funkcji f, dla których $f^{(n)}$ istnieje i jest ciągła w przedziale domkniętym $[a,b]$.

Wzór Taylora

Jest to ważny wzór dotyczący funkcji z $C^n[a,b]$, którym posługujemy się bardzo często w rozważaniach z analizy numerycznej i badaniu algorytmów numerycznych.

TWIERDZENIE 1.1.2. *Jeśli $f \in C^n[a,b]$ i jeśli $f^{(n+1)}$ istnieje w przedziale otwartym (a,b), to dla dowolnych punktów c i x z przedziału domkniętego $[a,b]$*

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(c)(x-c)^k + E_n(x), \tag{1.1.1}$$

gdzie dla pewnego punktu ξ leżącego między c i x

$$E_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi)(x-c)^{n+1}.$$

Wyrażenie $E_n(x)$ nazywamy *resztą Lagrange'a* wzoru Taylora. Słowa „leży między" użyte w twierdzeniu należy rozumieć tak, że albo $c < \xi < x$, albo $x < \xi < c$, zależnie od wartości c i x (przypadek $x = c$ można pominąć).

W ważnym szczególnym przypadku, gdy $c = 0$, wzór (1.1.1) nazywamy *wzorem Maclaurina* dla $f(x)$:

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(0)x^k + E_n(x), \text{ gdzie } E_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi)x^{n+1}.$$

Uwzględniając we wzorze Taylora lub Maclaurina nieskończenie wiele składników, otrzymujemy (jeśli to przejście graniczne jest dopuszczalne) *szereg Taylora*. Można go otrzymać dla wielu ważnych funkcji takich jak

$$\sin x = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^{2k+1} \qquad (-\infty < x < \infty),$$

$$\cos x = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k} \qquad (-\infty < x < \infty),$$

$$\log(1+x) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} x^k \qquad (-1 < |x| < 1),$$

$$\frac{1}{1+x} = \sum_{k=0}^{\infty} (-1)^k x^k \qquad (-1 < x < 1).$$

Są to przykłady *szeregów potęgowych* (zob. podrozdz. 6.7).

Przykład 1.1.3. Korzystając z tw. 1.1.2, podać wzór Taylora dla

$$f(x) := \log x$$

przyjmując, że $a = 1$, $b = 2$ i $c = 1$.

Rozwiązanie. Trzeba obliczyć pochodne funkcji $f(x) = \log x$. Mamy $f'(x) = x^{-1}$, $f''(x) = -x^{-2}$, $f'''(x) = 2x^{-3}$ itd., a ogólnie

$$f^{(k)}(x) = (-1)^{k-1}(k-1)!x^{-k} \quad (k \geqslant 1).$$

Oczywiście dla $x = 1$ mamy równości

$$f^{(k)}(1) = (-1)^{k-1}(k-1)! \quad (k \geqslant 1),$$

jest też $f^{(0)}(1) = f(1) = \log 1 = 0$. Uwzględniając to wszystko we wzorze Taylora (1.1.1), otrzymujemy wyrażenie

$$\log x = \sum_{k=1}^{n} \frac{(-1)^{k-1}}{k} (x-1)^k + E_n(x) \quad (1 \leqslant x \leqslant 2),$$

gdzie

$$E_n(x) = \frac{(-1)^n}{n+1} \xi^{-(n+1)} (x-1)^{n+1} \quad (1 < \xi < x). \qquad \blacksquare$$

W wyrażeniu otrzymanym dla $\log x$ suma $\sum_{k=1}^{n}$ jest wielomianem. Można go interpretować jako proste przybliżenie bardziej skomplikowanej funkcji $\log x$. Ostatni składnik w tymże wyrażeniu, czyli $E_n(x)$, jest błędem tego przybliżenia, informującym nas, jak bardzo różni się ono od $\log x$. Zauważmy, że ten składnik nie jest wielomianem, gdyż ξ zależy od x w sposób niewielomianowy.

Wzór Taylora pozwala obliczać przybliżone wartości funkcji w konkretnych punktach. We wzorze dla logarytmu, tj.

$$\log x = (x-1) - \frac{1}{2}(x-1)^2 + \frac{1}{3}(x-1)^3 - \ldots + \frac{(-1)^{n-1}}{n}(x-1)^n + E_n(x),$$

mamy

$$|E_n(x)| = \frac{1}{n+1}\xi^{-(n+1)}(x-1)^{n+1} < \frac{1}{n+1}(x-1)^{n+1},$$

gdyż $1 < \xi$ i $\xi^{-(n+1)} < 1$.

Przykład 1.1.4. Ile składników wielomianu we wzorze Taylora z przykł. 1.1.3 jest potrzebnych, aby obliczyć $\log 2$ z dokładnością do 10^{-8}?

Rozwiązanie. Dla $x = 2$ mamy

$$\log 2 = \sum_{k=1}^{n} \frac{(-1)^{k-1}}{k} + E_n(2),$$

gdzie $|E_n(2)| < 1/(n+1)$. Założoną dokładność otrzymamy wybierając takie n, żeby wartość bezwzględna błędu $E_n(2)$ nie przewyższała 10^{-8}. Ma zatem być $1/(n+1) \leqslant 10^{-8}$, czyli $n+1 \geqslant 10^8$. Musimy więc wziąć co najmniej sto milionów składników (!), gdy chcemy znaleźć $\log 2$ z żądaną dokładnością. Wnioskujemy stąd, że wzór Taylora nie daje dobrego sposobu obliczania $\log 2$ i potrzebna jest jakaś inna metoda. Dodajmy jednak, że jeśli stosujemy ten sam wzór dla $\log 1.5$, to powyższą dokładność zapewniają już 22 składniki (zob. zad. 1). ■

W rozumowaniach matematycznych stosujemy często szczególny przypadek wzoru Taylora dla $n = 0$. Odpowiada mu tzw. *twierdzenie o wartości średniej*:

Twierdzenie 1.1.5. *Jeśli $f \in C[a,b]$ i jeśli f' istnieje w przedziale otwartym (a,b), to dla x i c z przedziału domkniętego $[a,b]$ jest*

$$f(x) = f(c) + f'(\xi)(x-c),$$

gdzie ξ leży między c i x.

Dla $x = b$ i $c = a$ wynika stąd ważna równość

$$f(b) - f(a) = f'(\xi)(b - a), \quad \text{gdzie} \quad a < \xi < b,$$

a to daje wyrażenie przybliżone dla $f'(x)$:

$$f'(x) \approx \frac{f(x + h) - f(x - h)}{2h};$$

rozważamy je w podrozdz. 7.1.

Szczególnym przypadkiem twierdzenia o wartości średniej jest *twierdzenie Rolle'a*:

Twierdzenie 1.1.6. *Jeśli f jest ciągła w $[a, b]$, jeśli f' istnieje w (a, b) i jeśli $f(a) = f(b)$, to dla pewnego ξ z przedziału otwartego (a, b) jest $f'(\xi) = 0$.*

W obu ostatnich twierdzeniach może istnieć więcej niż jeden punkt ξ spełniający odpowiednie związki.

W podrozdziale 7.6 będzie nam potrzebny inny wariant wzoru Taylora, różniący się od poprzedniego wyrażeniem reszty w postaci całkowej:

Twierdzenie 1.1.7. *Jeśli $f \in C^{n+1}[a, b]$, to dla dowolnych punktów x i c przedziału domkniętego $[a, b]$ jest*

$$f(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(c)(x - c)^k + R_n(x),$$

gdzie

$$R_n(x) = \frac{1}{n!} \int_c^x f^{(n+1)}(t)(x - t)^n \, dt.$$

Inne warianty wzoru Taylora

Jeszcze inny wariant obu części składowych wzoru Taylora można otrzymać, zmieniając w (1.1.1) x na $x + h$ oraz c na x:

Twierdzenie 1.1.8. *Dla funkcji f z $C^{n+1}[a, b]$ oraz dowolnych punktów x i $x + h$ przedziału domkniętego $[a, b]$ jest*

$$f(x + h) = \sum_{k=0}^{n} \frac{h^k}{k!} f^{(k)}(x) + E_n(h), \tag{1.1.2}$$

gdzie

$$E_n(h) = \frac{h^{n+1}}{(n+1)!} f^{(n+1)}(\xi);$$

punkt ξ *leży między* x *i* $x+h$.

Przykład 1.1.9. Za pomocą wzoru Taylora wyrazić A^{x+h} i znaleźć wartość przybliżoną dla $10^{1.0001}$.

Rozwiązanie. Dla $f(x) = A^x$ jest $f^{(n)}(x) = A^x(\log A)^n$. Z (1.1.2) wynika, że

$$A^{x+h} = A^x\Big(1 + \sum_{k=1}^{n} \frac{h^k}{k!}(\log A)^k\Big) + E_n(h).$$

Dla $A = 10$, $x = 1$ i $h = 10^{-4}$ jest zatem

$$\begin{aligned} 10^{1.001} &= 10(1 + 10^{-4}(\log 10) + \tfrac{1}{2} \cdot 10^{-8}(\log 10)^2 + \ldots) \approx \\ &\approx 10(1 + 2.30258\,5093 \times 10^{-4} + 2.650949 \times 10^{-8}) \approx \\ &\approx 10.00230\,285. \end{aligned}$$

■

Wzór Taylora odnosi się także do funkcji, których argumenty i wartości są wektorami. Ściślej, jeśli f jest odwzorowaniem z $\mathbb{R}^n$ w $\mathbb{R}^m$, to są znane wyrażenia dla $f(x+h)$ przez $f(x)$, $f'(x)$, $f''(x)$ itd. Oczywiście główna trudność polega tu na właściwym określeniu pochodnych. Ten temat jest omawiany w wielu podręcznikach; zob. na przykład Bartle [1976], Smith [1971], Dieudonné [1960], Rudin [*1998]. Niżej ograniczono się do szczególnego przypadku użytecznego dalej.

Dla funkcji $f\colon \mathbb{R}^2 \to \mathbb{R}$ wzór Taylora upraszcza się, jeśli stosujemy w nim szczególną symbolikę:

Twierdzenie 1.1.10. *Jeśli funkcja* f *należy do* $C^{n+1}([a,b] \times [c,d])$ *i jeśli punkty* (x,y) *i* $(x+h, y+k)$ *leżą w prostokącie* $[a,b] \times [c,d] \subseteq \mathbb{R}^2$, *to*

$$f(x+h, y+k) = \sum_{i=0}^{n} \frac{1}{i!} \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^i f(x,y) + E_n(h,k), \qquad (1.1.3)$$

gdzie

$$E_n(h,k) = \frac{1}{(n+1)!} \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^{n+1} f(x+\theta h, y+\theta k) \quad (0 < \theta < 1).$$

Symbolikę zastosowaną w twierdzeniu wyjaśniają przykłady:

$$\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^0 f(x,y)=f(x,y),$$
$$\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^1 f(x,y)=\left(h\frac{\partial f}{\partial x}+k\frac{\partial f}{\partial y}\right)(x,y),$$
$$\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^2 f(x,y)=\left(h^2\frac{\partial^2 f}{\partial x^2}+2hk\frac{\partial^2 f}{\partial x\partial y}+k^2\frac{\partial^2 f}{\partial y^2}\right)(x,y),\ldots$$

Oznaczenia: $f_x=\partial f/\partial x$, $f_y=\partial f/\partial y$, $f_{xx}=\partial^2 f/\partial x^2$, $f_{xy}=\partial^2 f/(\partial x\partial y)$, $f_{yy}=\partial^2 f/\partial y^2$ upraszczają postać początkowych składników wzoru (1.1.3):

$$f(x+h,y+k)=f+(hf_x+kf_y)+\tfrac{1}{2}(h^2f_{xx}+2hkf_{xy}+k^2f_{yy})+\ldots$$

Po prawej stronie funkcja f i jej pochodne cząstkowe są obliczane w punkcie (x,y).

Przykład 1.1.11. Jakie są początkowe składniki wzoru Taylora dla funkcji $f(x,y)=\cos(xy)$?

Rozwiązanie. Dla tej funkcji

$$\frac{\partial f}{\partial x}=-y\sin(xy),\quad \frac{\partial f}{\partial y}=-x\sin(xy),$$
$$\frac{\partial^2 f}{\partial x^2}=-y^2\cos(xy),\quad \frac{\partial^2 f}{\partial x\partial y}=-xy\cos(xy)-\sin(xy),$$
$$\frac{\partial^2 f}{\partial y^2}=-x^2\cos(xy).$$

Dlatego ze wzoru (1.1.3) dla $n=1$ wynika, że

$$\cos[(x+h)(y+k)]=\cos(xy)-hy\sin(xy)-kx\sin(xy)+E_1(h,k),$$

gdzie reszta E_1 jest sumą trzech składników:

$$-\frac{1}{2}h^2(y+\theta k)^2\cos[(x+\theta h)(y+\theta k)]-$$
$$-hk\{(x+\theta h)(y+\theta k)\cos[(x+\theta h)(y+\theta k)]+\sin[(x+\theta h)(y+\theta k)]\}-$$
$$-\frac{1}{2}k^2(x+\theta h)^2\cos[(x+\theta h)(y+\theta k)].$$ ■

ZADANIA 1.1

1. (a) Znaleźć szereg Taylora w 0 dla funkcji $f(x) = \log(x+1)$. Podać dwa wyrażenia reszty wzoru Taylora.

(b) Wyznaczyć najmniejszą liczbę składników szeregu niezbędną do obliczenia $\log 1.5$ z błędem mniejszym od 10^{-8}.

(c) Wyznaczyć liczbę składników potrzebną do obliczenia $\log 1.6$ z błędem równym co najwyżej 10^{-10}.

2. Udowodnić, że jeśli f jest różniczkowalna w x, to

$$\lim_{h\to 0} \frac{f(x+h) - f(x-h)}{2h} = f'(x).$$

Pokazać, że dla pewnych funkcji, które nie są różniczkowalne w x, powyższa granica jednak istnieje. (Zob. Eggermont [1988] albo następne zadanie).

3. Sprawdzić, czy następujące zdanie jest prawdziwe, czy fałszywe: Jeśli f jest różniczkowalna w x, to dla $\alpha \neq 1$

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x+\alpha h)}{h - \alpha h}.$$

4. Znaleźć szereg Taylora dla $f(x) = \cosh x$ w otoczeniu punktu $c = 0$.

5. Jeśli szeregu dla $\log x$ obciętego po składniku z $(x-1)^{1000}$ użyto do obliczenia $\log 2$, to jak można oszacować błąd?

6. Niech k będzie liczbą naturalną i niech $0 < \alpha < 1$. Do jakich klas $C^n(\mathbb{R})$ należy funkcja $x^{k+\alpha}$?

7. Udowodnić, że jeśli $f \in C^n(\mathbb{R})$, to $f' \in C^{n-1}(\mathbb{R})$, a $\int_a^x f(t)\,dt$ należy do $C^{n+1}(\mathbb{R})$.

8. Wykazać, że jeśli $f \in C^n(\mathbb{R})$ i $f(x_0) = f(x_1) = \ldots = f(x_n) = 0$, gdzie $x_0 < x_1 < \ldots < x_n$, to $f^{(n)}(\xi) = 0$ dla pewnego $\xi \in (x_0, x_n)$. Wskazówka: Zastosować n razy twierdzenie Rolle'a.

9. Dla małych x używa się często przybliżenia $\sin x \approx x$. Korzystając ze wzoru Taylora, oszacować błąd tego przybliżenia. Dla jakich x przybliżenie jest dokładne aż do sześciu cyfr dziesiętnych?

10. Jak dobre jest przybliżenie $\cos x \approx 1 - \frac{1}{2}x^2$ dla małych x? Dla jakich x to przybliżenie daje poprawne wyniki zaokrąglone do trzech cyfr dziesiętnych?

11. Korzystając ze wzoru Taylora dla $n = 2$, wykazać, że nierówność $1 + x < e^x$ jest prawdziwa dla wszystkich x rzeczywistych różnych od 0.

12. Ile składników szeregu dla e^x trzeba uwzględnić, aby obliczyć e^2 z dokładnością do czterech cyfr dziesiętnych (po zaokrągleniu)?

13. Znaleźć wzór Taylora z resztą dla $f(x) = \log x$ w otoczeniu punktu e. Założyć, że $|x - e| < 1$ i że trzeba uzyskać dokładność $0.5_{10}{-}1$. Ile składników wzoru trzeba uwzględnić?

14. Znaleźć wzór Taylora dla funkcji $f(x) = e^{2x}\sin x$ i $n = 2$ w otoczeniu punktu $\pi/2$.

15. Znaleźć wzór Taylora dla $f(x) = \exp(\cos x)$ i $n = 2$ w otoczeniu punktu π.

16. Zakładając, że $|x| < \frac{1}{2}$ i stosując wzór Taylora, znaleźć najlepsze oszacowanie z góry dla: (**a**) $|\cos x - (1 - x^2/2)|$, (**b**) $|\sin x - x(1 - x^2/6)|$.

17. Ile składników trzeba uwzględnić w szeregu

$$e = \sum_{k=0}^{\infty} \frac{1}{k!},$$

aby otrzymać e z błędem nie większym od $0.6_{10}{-}20$?

18. Znaleźć dwa początkowe składniki wzoru Taylora dla $x^{1/5}$ w otoczeniu punktu 32. Jak dokładne będzie wynikające stąd przybliżenie dla pierwiastka piątego stopnia z 31.999999?

19. Wyznaczyć resztę Lagrange'a wzoru Taylora dla funkcji $f(x) = \cos x$, gdy $n = 2$ i $c = \pi/2$. Jak małe musi być $|x - \pi/2|$, żeby wartość bezwzględna tej reszty nie przewyższała $0.5_{10}{-}4$?

20. W przykładzie 1.1.3 ilustrującym wzór Taylora otrzymano resztę Lagrange'a. Porównać z nią odpowiednią resztę całkową.

1.2. Rząd zbieżności i inne podstawowe pojęcia

W obliczeniach numerycznych, szczególnie na superszybkich komputerach, często się zdarza, że zamiast ostatecznej odpowiedzi otrzymujemy ciąg przybliżonych wyników, zwykle coraz dokładniejszych. Zbieżność ciągów jest ważnym problemem i powrócimy do niego w dalszym ciągu. Tutaj zaś prezentujemy tylko kilka pojęć wprowadzających w temat.

Ciągi zbieżne

Rozważmy wyidealizowaną sytuację, w której wynikiem zadania jest tylko jedna liczba rzeczywista. Może to być pierwiastek skomplikowanego równania albo wartość całki oznaczonej, której nie potrafimy obliczyć analitycznie. W takich przypadkach program komputerowy może generować ciąg liczb rzeczywistych $x_1, x_2, \ldots$, które są tylko *przybliżeniami* dokładnego wyniku.

Piszemy

$$\lim_{n\to\infty} x_n = L,$$

jeśli każdemu dodatniemu ε odpowiada takie rzeczywiste r, że $|x_n - L| < \varepsilon$ dla każdego całkowitego $n > r$.

Jest, na przykład,

$$\lim_{n\to\infty} \frac{n+1}{n} = 1.$$

Mniej banalny przykład wiąże się z definicją

$$e := \lim_{n\to\infty} \left(1+\frac{1}{n}\right)^n$$

ważnej liczby niewymiernej e. Oto niektóre elementy ciągu $x_n = (1+1/n)^n$:

$$x_1 = 2.000000,\ x_{10} = 2.593742,\ x_{30} = 2.674319,$$
$$x_{50} = 2.691588,\ x_{1000} = 2.716924.$$

Ten ciąg jest zbieżny bardzo wolno, skoro ma granicę $e = 2.71828\,18\ldots$, a jego tysięczny element przybliża ją z błędem $0.00135\,8$. Wykonując obliczenia w podwójnej precyzji, można postawić hipotezę, że

$$\frac{|x_{n+1}-e|}{|x_n-e|} \to 1.$$

Taka zbieżność (zwana *logarytmiczną*) jest znacznie wolniejsza od określonej dalej zbieżności liniowej.

Rozważmy teraz ciąg

$$x_{n+1} = \frac{x_{n-1}x_n+1}{x_{n-1}+x_n}.$$

Dla wybranych dwóch wartości początkowych mamy

$$x_0 = 0.0,\ x_1 = 2.00,\ x_2 = 0.5,\ x_3 = 0.8,\ x_4 = 1.076923077,$$
$$x_5 = 0.991803278,\ x_6 = 0.999695214,\ x_7 = 1.000001254.$$

Można więc przypuszczać, że ten ciąg jest szybko zbieżny (do punktu 1). Jest tak rzeczywiście, gdyż

$$\frac{x_{n+1}-1}{x_n-1} = \frac{x_{n-1}-1}{x_{n-1}+x_n} \to 0$$

Taką zbieżność nazywamy *nadliniową*.

Jeszcze szybciej jest zbieżny ciąg określony wzorami:

$$x_1 = 2, \qquad x_{n+1} = \frac{1}{2}x_n + \frac{1}{x_n} \quad (n \geqslant 1).$$

Jego początkowe elementy są następujące:

$$x_1 = 2.000000,\ x_2 = 1.500000,\ x_3 = 1.416667,\ x_4 = 1.414216.$$

Granicą ciągu jest $\sqrt{2} = 1.41421\,3562\ldots$, a jego szybka zbieżność wynika z równości

$$\frac{x_{n+1}-\sqrt{2}}{(x_n-\sqrt{2})^2} = \frac{1}{2x_n} \to \frac{1}{2\sqrt{2}} \approx 0.354.$$

Mamy tu do czynienia ze zbieżnością *kwadratową*.

Rząd zbieżności

Opisując prędkość, z jaką ciąg dąży do granicy, używamy specjalnej terminologii. Niech $\{x_n\}$ będzie ciągiem liczb rzeczywistych lub zespolonych, zbieżnym do granicy x^*. Mówimy, że zbieżność jest co najmniej *liniowa*, jeśli istnieją stała $c < 1$ i liczba całkowita N takie, że

$$|x_{n+1} - x^*| \leqslant c|x_n - x^*| \quad (n \geqslant N).$$

Zbieżność jest co najmniej *nadliniowa*, jeśli istnieją ciąg $\{\varepsilon_n\}$ zbieżny do 0 i liczba całkowita N takie, że

$$|x_{n+1} - x^*| \leqslant \varepsilon_n|x_n - x^*| \quad (n \geqslant N).$$

Zbieżność jest co najmniej *kwadratowa*, jeśli istnieją stała dodatnia C (niekoniecznie mniejsza od 1) i liczba całkowita N takie, że

$$|x_{n+1} - x^*| \leqslant C|x_n - x^*|^2 \quad (n \geqslant N).$$

Ogólniej, jeśli istnieją stała dodatnia C, stała $\alpha > 1$ i liczba całkowita N takie, że

$$|x_{n+1} - x^*| \leqslant C|x_n - x^*|^\alpha \quad (n \geqslant N),$$

to mamy zbieżność co najmniej *rzędu* α. Do wprowadzonych tu pojęć wrócimy w rozdz. 3 [2)].

Symbole $\mathcal{O}$ i o

Poznamy teraz pewne typowe sposoby porównywania dwóch ciągów lub dwóch funkcji. Zacznijmy od ciągów.

Niech $\{x_n\}$ i $\{\alpha_n\}$ będą dwoma różnymi ciągami. Piszemy

$$x_n = \mathcal{O}(\alpha_n),$$

jeśli istnieją stałe C i n_0 takie, że $|x_n| \leqslant C|\alpha_n|$ dla każdego $n \geqslant n_0$. Mówimy, że x_n jest równe „$\mathcal{O}$ dużemu" od α_n.

[2)] Z numerycznego punktu widzenia ciąg zbieżny co najmniej nadliniowo pozwala na ogół dostatecznie łatwo obliczyć jego granicę. Natomiast ciągi zbieżne liniowo, dla których stała c jest bliska 1, a tym bardziej ciągi zbieżne logarytmicznie, sprawiają wiele kłopotów. W takich przypadkach obliczenie granicy w rozsądnym czasie i z żądaną dokładnością może wymagać zastosowania jednej z metod *przyspieszania zbieżności*. Należy do nich metoda Δ^2 Aitkena opisana w podrozdz. 5.1 (*przyp. tłum.*).

Relacja

$$x_n = o(\alpha_n)$$

oznacza, że $\lim_{n\to\infty}(x_n/\alpha_n) = 0$. Mówimy w takim przypadku, że x_n jest równe „o małemu" od α_n. Aby jednak uniknąć możliwego dzielenia przez zero, określamy ściśle symbol o: dla pewnego ciągu $\{\varepsilon_n\}$ liczb nieujemnych zbieżnego do 0 jest $|x_n| \leqslant \varepsilon_n|\alpha_n|$.

Te pojęcia pozwalają porównywać z grubsza dwa ciągi. Jest to często potrzebne dla ciągów zbieżnych do zera. Jeśli $x_n \to 0$, $\alpha_n \to 0$ i $x_n = \mathcal{O}(\alpha_n)$, to ciąg $\{x_n\}$ dąży do 0 co najmniej tak szybko jak $\{\alpha_n\}$. Jeśli $x_n = o(\alpha_n)$, to ciąg $\{x_n\}$ jest zbieżny do 0 szybciej niż $\{\alpha_n\}$.

Oto kilka przykładów:

$$\frac{n+1}{n^2} = \mathcal{O}\left(\frac{1}{n}\right), \quad \frac{5}{n} + e^{-n} = \mathcal{O}\left(\frac{1}{n}\right), \tag{1.2.1}$$

$$\frac{1}{n\log n} = o\left(\frac{1}{n}\right), \quad \frac{1}{n} = o\left(\frac{1}{\log n}\right), \quad e^{-n} = o\left(\frac{1}{n^2}\right). \tag{1.2.2}$$

Przykład 1.1.4 prowadzi do wniosku, że

$$\log 2 - \sum_{k=1}^{n-1} \frac{(-1)^{k-1}}{k} = \mathcal{O}\left(\frac{1}{n}\right).$$

Dla $n \to \infty$ podane wyżej sumy są zbieżne bardzo wolno do $\log 2$. Natomiast równość

$$e^x - \sum_{k=0}^{n-1} \frac{1}{k!}x^k = \mathcal{O}\left(\frac{1}{n!}\right) \quad (|x| \leqslant 1)$$

ilustruje bardzo szybką zbieżność pewnych sum do e^x.

Wprowadzone oznaczenia są używane nie tylko dla ciągów. Możemy na przykład napisać, że

$$\sin x = x - \frac{x^3}{6} + \mathcal{O}(x^5) \quad (x \to 0).$$

Rozumiemy to tak, że istnieją otoczenie punktu 0 i stała C takie, że w tym otoczeniu

$$\left|\sin x - x + \frac{x^3}{6}\right| \leqslant C|x|^5.$$

Można to sprawdzić, korzystając ze wzoru Taylora dla $n = 4$ i $f(x) = \sin x$.

Równość

$$f(x) = \mathcal{O}(g(x)) \quad (x \to \infty)$$

oznacza, że istnieją takie stałe r i C, że $|f(x)| \leqslant C|g(x)|$ dla każdego $x \geqslant r$. Jest na przykład

$$\sqrt{x^2+1} = \mathcal{O}(x) \quad (x \to \infty),$$

gdyż $\sqrt{x^2+1} \leqslant 2x$ dla $x \geqslant 1$. Używając oznaczeń $f(x) = \mathcal{O}(g(x))$ lub $f(x) = o(g(x))$, trzeba koniecznie podać punkt, do którego dąży x. Jest na przykład $x^{-2} = o(x^{-1})$ dla $x \to \infty$, ale $x^{-1} = o(x^{-2})$ dla $x \to 0$.

Ogólniej, piszemy

$$f(x) = \mathcal{O}(g(x)) \quad (x \to x^*),$$

jeśli istnieją stała C i otoczenie punktu x^* takie, że w tym otoczeniu jest $|f(x)| \leqslant C|g(x)|$. Podobnie, równość

$$f(x) = o(g(x)) \quad (x \to x^*)$$

jest równoważna temu, że $\lim_{x \to x^*}[f(x)/g(x)] = 0$.

Twierdzenie o wartości średniej dla całek

W analizie numerycznej jest często potrzebne następujące twierdzenie o wartości średniej:

Twierdzenie 1.2.1. *Jeśli funkcje rzeczywiste u i v są ciągłe w $[a,b]$ i jeśli $v \geqslant 0$, to istnieje punkt $\xi \in [a,b]$ taki, że*

$$\int_a^b u(x)v(x)\,dx = u(\xi)\int_a^b v(x)dx.$$

Kresy dolny i górny

Przypomnimy teraz inne ważne pojęcia, które często występują w analizie numerycznej. Są to *kres górny* (*supremum*) i *kres dolny* (*infimum*). Niech S będzie zbiorem niepustym i ograniczonym z dołu liczb rzeczywistych, czyli takim, że dla pewnej liczby rzeczywistej a jest

$$a \leqslant x \quad \text{dla wszystkich} \quad x \in S.$$

Wśród liczb a o tej własności istnieje największa; nazywamy ją *kresem dolnym* zbioru S i oznaczamy symbolem $\inf S$.

Podobnie, jeśli S jest zbiorem niepustym i ograniczonym z góry liczb rzeczywistych, czyli takim, że dla pewnej liczby rzeczywistej b jest

$$x \leqslant b \quad \text{dla wszystkich} \quad x \in S,$$

to wśród liczb b o tej własności istnieje najmniejsza; nazywamy ją *kresem górnym* zbioru S i oznaczamy symbolem $\sup S$.

Istnienie kresu dolnego i kresu górnego zbioru ograniczonego liczb rzeczywistych jest jedną z jego głębszych charakterystyk; na przykład zbiór liczb wymiernych nie ma takiej własności.

Jeśli f jest funkcją, to symbol $\sup_{x\in A} f(x)$ oznacza $\sup\{f(x)\colon x \in A\}$. Możemy na przykład wykazać, że

$$\sup_{0<x<\pi/6} \sin x = \frac{1}{2}.$$

Funkcje jawne i uwikłane

Funkcje są zwykle definiowane za pomocą jawnego wyrażenia, które pozwala obliczać wartość funkcji dla dowolnego argumentu. Przykładem jest definicja

$$f(x) := \sqrt{7x^3 - 2x}.$$

Istnieje jednak wiele innych metod definiowania funkcji, np. za pomocą równania różniczkowego, całki lub szeregu nieskończonego. Można na przykład określić poprawnie funkcję $y = f(x)$ za pomocą równania różniczkowego z warunkiem początkowym:

$$y' = 1 + \sin y, \quad y(0) = 0.$$

Inny przykład, to tzw. *funkcja błędu*, oznaczana $\operatorname{erf} x$ i określona przez całkę:

$$\operatorname{erf} x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

W tym podrozdziale rozważamy funkcje *uwikłane*, tj. zdefiniowane w sposób niejawny. Rozumiemy to tak, że dana jest funkcja G dwóch zmiennych i że z równania $G(x, y) = 0$ chcemy odtworzyć y jako funkcję x. W pewnych przypadkach możemy rozwikłać to równanie i otrzymać $y = f(x)$. Na przykład, z równania

$$y^2 + 3xy - 7 = 0$$

wynika, że

$$y = \frac{1}{2}\left(-3x \pm \sqrt{9x^2 + 28}\right),$$

co daje dwie jawne funkcje. Podobnie, z równania

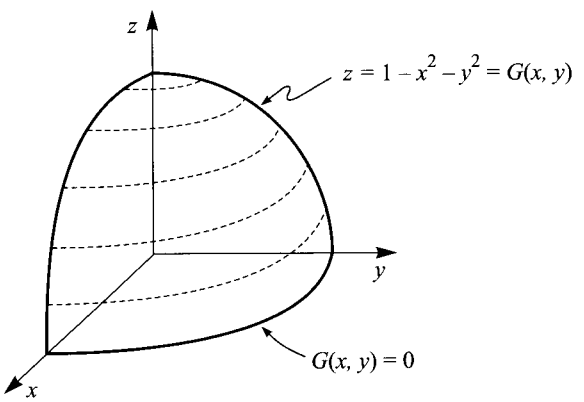
$$\sin(y+7) = x^3 - 2$$

wynika funkcja określona jawnym wzorem:

$$y = \arcsin\left(x^3 - 2\right) - 7;$$

w istocie jest tu wiele funkcji, gdyż można wybrać różne *gałęzie* funkcji odwrotnej do sinusa.

Ogólnie, jeśli G jest daną funkcją, a (x_0, y_0) jest punktem takim, że $G(x_0, y_0) = 0$, to spodziewamy się, że w jego pobliżu są inne punkty spełniające równanie $G(x, y) = 0$. Dlatego y jest zapewne funkcją zmiennej x w pewnym otoczeniu punktu x_0. Typową sytuację pokazuje rys. 1.4.



RYS. 1.4. Powierzchnia G przecina płaszczyznę xy wzdłuż krzywej $G(x, y) = 0$

Poniższe ważne twierdzenie opisuje taką sytuację.

TWIERDZENIE 1.2.2. *Jeśli G jest funkcją dwóch zmiennych rzeczywistych, określoną i różniczkowalną w sposób ciągły w pewnym otoczeniu takiego punktu (x_0, y_0), w którym $G = 0$ i $\partial G/\partial y \neq 0$, to istnieje liczba dodatnia δ i funkcja f mająca pochodną ciągłą dla $|x - x_0| < \delta$ i taka, że $f(x_0) = y_0$ i $G(x, f(x)) = 0$ dla $|x - x_0| < \delta$.*

PRZYKŁAD 1.2.3. Czy równanie

$$x^7 + 2y^8 - y^3 = 0$$

określa y jako funkcję zmiennej x o pochodnej ciągłej w pewnym otoczeniu punktu $x = -1$?

Rozwiązanie. Żeby odpowiedzieć na to pytanie, przyjmijmy, że

$$G(x,y) = x^7 + 2y^8 - y^3$$

i $(x_0, y_0) = (-1, 1)$. Jest zatem $G(x_0, y_0) = 0$ i

$$\frac{\partial G}{\partial y} = 16y^7 - 3y^2,$$

skąd $\partial G/\partial y = 13$ w punkcie (x_0, y_0). Na mocy tw. 1.2.2 y jest funkcją zmiennej x, mającą pochodną ciągłą w pobliżu punktu $x_0 = -1$. ■

Jeśli f jest funkcją uwikłaną określoną za pomocą równania $G(x,y) = 0$, to dla x z pewnego przedziału jest $G(x, f(x)) = 0$. Pochodną $f'(x)$ możemy obliczyć w sposób znany z analizy matematycznej. Różniczkujemy mianowicie względem x obie strony równania $G(x,y) = 0$, pamiętając o tym, że y jest funkcją zmiennej x. Dlatego otrzymujemy równanie

$$\frac{\partial G}{\partial x} + \frac{\partial G}{\partial y}\frac{dy}{dx} = 0,$$

skąd

$$\frac{dy}{dx} = -\frac{\partial G}{\partial x} \Big/ \frac{\partial G}{\partial y},$$

czyli

$$f'(x) = -\frac{\partial G}{\partial x} \Big/ \frac{\partial G}{\partial y}.$$

Pochodna wyraża się przez x i y [3)].

Przykład 1.2.4. Jaka jest wartość dy/dx w punkcie $(2,1)$, jeśli $y(x)$ jest funkcją uwikłaną określoną za pomocą równania $x^3 - y^7 + 4x^2 + y^4 - 24 = 0$?

Rozwiązanie. Różniczkowanie opisane wyżej daje równanie

$$3x^2 - 7y^6\frac{dy}{dx} + 8x + 4y^3\frac{dy}{dx} = 0.$$

Dla $x = 2$ i $y = 1$ (dane równanie jest wtedy spełnione) wynika stąd, że

$$12 - 7\frac{dy}{dx} + 16 + 4\frac{dy}{dx} = 0$$

i że $dy/dx = 28/3$. ■

W podrozdziale 3.2 omawia się pewne zadania numeryczne związane z funkcjami uwikłanymi.

[3)] W podobny sposób można znaleźć także pochodne wyższych rzędów funkcji uwikłanej i dzięki temu obliczyć dobre przybliżenia jej wartości (*przyp. tłum.*).

ZADANIA 1.2

1. Obliczanie elementów $x_n = (1 + 1/n)^n$ ciągu pozwala wnioskować, że ten ciąg jest ściśle rosnący. Udowodnić, że tak jest. Wskazówki: Po pierwsze, jeśli $\log f(x)$ jest funkcją rosnącą, to $f(x)$ ma tę samą własność. Po drugie, jeśli $f'(x) > 0$, to f jest funkcją rosnącą. Na koniec, $\log t$ można określić jako $\int_1^x t^{-1}\,dt$.
2. (cd.). Wykazać, że elementy ciągu z poprzedniego zadania są mniejsze od 3.
3. Udowodnić, że jeśli $0 < \theta < 1$, to $(1 + a\theta^n)/(1 + a\theta^{n-1})$ dąży do 1 liniowo, gdy $n \to \infty$.
4. Wyznaczyć najlepszą wartość całkowitą k w równaniu

   $$\operatorname{arctg} x = x + \mathcal{O}(x^k), \quad \text{gdy} \quad x \to 0.$$

5. Niech ciąg $\{x_n\}$ będzie określony rekurencyjnie wzorem $x_{n+1} = F(x_n)$, gdzie funkcja F ma ciągłą pochodną. Zakładając, że $\{x_n\} \to x$ dla $n \to \infty$ i $F'(x) = 0$, wykazać, że

   $$x_{n+2} - x_{n+1} = o(x_{n+1} - x_n).$$

   Wskazówka: Zastosować twierdzenie o wartości średniej.
6. Udowodnić, że każdą dostatecznie regularną funkcję można przybliżyć w przedziale długości h za pomocą wielomianu stopnia n z błędem równym $\mathcal{O}(h^{n+1})$, gdy $h \to 0$.
7. Rozważyć szereg

   $$e^{\operatorname{tg} x} = 1 + x + \frac{x^2}{2!} + \frac{3x^3}{3!} + \frac{9x^4}{4!} + \ldots \qquad (|x| < \pi/2).$$

   Zachowując trzy składniki szeregu, oszacować jego resztę używając symbolu $o(\,)$ z najlepszym całkowitym wykładnikiem dla $x \to 0$.
8. Powtórzyć poprzednie zadanie, ale dla szeregu

   $$\log \frac{\operatorname{tg} x}{x} = \frac{x^2}{3} + \frac{7x^4}{90} + \frac{62x^6}{2835} + \ldots \qquad (0 < |x| < \pi/2)$$

   i używając symbolu $\mathcal{O}(\,)$.
9. Zbadać, dla jakich γ i δ całkowitych wykropkowane składniki w szeregu

   $$\log(1 + x) = \sum_{k=1}^{n-1} (-1)^{k-1} \frac{x^k}{k} + \ldots$$

   można dla $x \to 0$ zmienić na $\mathcal{O}(x^\gamma)$ albo $o(x^\delta)$.
10. Czy dla niżej podanych par (x_n, α_n) jest prawdą, że $x_n = \mathcal{O}(\alpha_n)$ dla $n \to \infty$?
    (a) $x_n = 5n^2 + 9n^3 + 1$, $\alpha_n = n^2$
    (b) $x_n = 5n^2 + 9n^3 + 1$, $\alpha_n = 1$
    (c) $x_n = \sqrt{n+3}$, $\alpha_n = 1$
    (d) $x_n = 5n^2 + 9n^3 + 1$, $\alpha_n = n^3$
    (e) $x_n = \sqrt{n+3}$, $\alpha_n = 1/n$

11. Sprawdzić, które z następujących zdań są prawdziwe (w każdym $n \to \infty$):

(**a**) $(n+1)/n^2 = o(1/n)$ (**d**) $1/(n \log n) = o(1/n)$

(**b**) $(n+1)/\sqrt{n} = o(1)$ (**e**) $e^n/n^5 = \mathcal{O}(1/n)$

(**c**) $1/(\log n) = \mathcal{O}(1/n)$

12. Wyrażenia e^h, $(1-h^4)^{-1}$, $\cos h$ i $1 + \sin h^3$ mają tę samą granicę dla $h \to 0$. Przedstawić każde z nich w postaci

$$f(h) = c + \mathcal{O}(h^\alpha) = c + o(h^\beta),$$

wybierając najlepsze α i β całkowite.

13. (cd.). Jaka jest granica i jaka jest szybkość zbieżności dla $h \to 0$ wyrażenia

$$\frac{1}{h^2}[(1+h) - e^h]?$$

Przedstawić wyrażenie w postaci podanej w poprzednim zadaniu.

14. Wykazać, że następujące zdania są fałszywe:

(**a**) $e^x - 1 = \mathcal{O}(x^2)$ dla $x \to 0$.

(**b**) $x^{-2} = \mathcal{O}(\operatorname{ctg} x)$ dla $x \to 0$.

(**c**) $\operatorname{ctg} x = o(x^{-1})$ dla $x \to 0$.

15. Niech będzie $\{a_n\} \to 0$ i $\lambda > 1$. Wykazać, że $\sum_{k=0}^{n} a_k \lambda^k = o(\lambda^n)$ dla $n \to \infty$.

16. Czy poniższe dwa zdania są równoważne?

(**a**) $|f(x)| = \mathcal{O}(|x|^{-n-\varepsilon})$ dla pewnego $\varepsilon > 0$, gdy $|x| \to \infty$.

(**b**) $|f(x)| = o(|x|^{-n})$.

17. Udowodnić, że $x_n = x + o(1)$ wtedy i tylko wtedy, gdy $\lim_{n\to\infty} x_n = x$.

18. Wykazać, że wszystkie równości (1.2.1) i (1.2.2) są poprawne.

19. Wykazać, że dla ustalonego n

$$\sum_{k=0}^{n} x^k = 1/(1-x) + o(x^n) \quad (x \to 0).$$

20. Wyznaczyć najlepsze β całkowite takie, że dla ustalonego n

$$\frac{1}{1-x} = 1 + x + x^2 + \ldots + x^n + \mathcal{O}(x^\beta) \quad (0 < x < 1),$$

gdy $x \to 0$. Zrobić to samo dla $o(x^\beta)$. Czy w tym przypadku istnieje najlepsze β całkowite?

21. Wykazać, że jeśli $x_n = \mathcal{O}(\alpha_n)$, to $cx_n = \mathcal{O}(\alpha_n)$.

22. Wykazać, że jeśli $x_n = \mathcal{O}(\alpha_n)$, to $x_n/(\log n) = o(\alpha_n)$.

23. Znaleźć najlepsze k całkowite takie, że $\cos x - 1 + x^2/2 = \mathcal{O}(x^k)$ dla $x \to 0$.

24. Udowodnić, że jeśli $x_n = o(\alpha_n)$, to $x_n = \mathcal{O}(\alpha_n)$. Wykazać, że twierdzenie przeciwne nie jest prawdziwe.

25. Udowodnić, że jeśli $x_n = \mathcal{O}(\alpha_n)$ i $y_n = \mathcal{O}(\alpha_n)$, to $x_n + y_n = \mathcal{O}(\alpha_n)$.

26. Udowodnić, że jeśli $x_n = o(\alpha_n)$ i $y_n = o(\alpha_n)$, to $x_n + y_n = o(\alpha_n)$.

27. Wykazać, że $x^r = \mathcal{O}(e^x)$ $(x \to \infty)$ dla dowolnego $r > 0$.

28. Wykazać, że $\log x = \mathcal{O}(x^r)$ $(x \to \infty)$ dla dowolnego $r > 0$.

29. Udowodnić, że jeśli $\alpha_n \to 0$, $x_n = \mathcal{O}(\alpha_n)$ i $y_n = \mathcal{O}(\alpha_n)$, to $x_n y_n = o(\alpha_n)$.

30. Udowodnić, że jeśli $x_n = \mathcal{O}(\alpha_n)$, to $\alpha_n^{-1} = \mathcal{O}(x_n^{-1})$. Wykazać, że ta własność zachowuje się po zmianie relacji $\mathcal{O}$ na o.

31. Stosując twierdzenie o wartości średniej dla całek, wykazać, że dla pewnego $y \in (0, \pi/2)$ jest

$$\int_0^{\pi/2} e^x \cos x \, dx = e^y.$$

32. Pokazać na przykładzie, że w tw. 1.2.1 nie można usunąć założenia o ciągłości funkcji u.

33. Obliczyć wartości wyrażeń: (**a**) $\sup_{x \in \mathbb{R}} \operatorname{arctg} x$, (**b**) $\sup_{x \geqslant 0} e^{-x}$, (**c**) $\inf_{x \in \mathbb{R}} e^{-x}$, (**d**) $\sup_{x \in \mathbb{R}} (x^2 + 1)^{-1}$.

34. Wyrazić w jawnej postaci dwie funkcje określone w sposób uwikłany za pomocą równania

$$(x^3 - 1)y + e^x y^2 + \cos x - 1 = 0.$$

35. Rozwiązania równań różniczkowych otrzymuje się często w postaci uwikłanej. Pokazać, że równanie

$$2x^3 y^2 + x^2 y + e^x = c$$

opisuje rozwiązanie równania różniczkowego

$$\frac{dy}{dx} = -(6x^2 y^2 + 2xy + e^x)/(4x^3 y + x^2).$$

36. W astronomii jest używane równanie Keplera: $x - y + \varepsilon \sin y = 0$, gdzie parametr ε należy do przedziału $[0, 1]$. Wykazać, że dla każdego x rzeczywistego istnieje y rzeczywiste spełniające to równanie. Wykazać, że dla $0 \leqslant \varepsilon < 1$ pochodna dy/dx jest wszędzie ciągła. Wskazówka: Napisać równanie w postaci $x = y - \varepsilon \sin y$ i zbadać własności prawej strony dla $y \to +\infty$ i $y \to -\infty$. W drugiej części zadania zastosować twierdzenie o funkcji uwikłanej.

37. Znaleźć takie x, dla których równanie

$$y - \log(x + y) = 0$$

określa y jako funkcję uwikłaną zmiennej x. Obliczyć dy/dx.

ZADANIA KOMPUTEROWE 1.2

K1. Rozważmy związek rekurencyjny

$$x_0 = 1, \qquad x_1 = c, \qquad x_{n+1} = x_n + x_{n-1} \quad (n \geqslant 1).$$

Można wykazać, że jeśli $c = (1 + \sqrt{5})/2$, to

$$x_n = \left(\frac{1+\sqrt{5}}{2}\right)^n.$$

Podobnie, jeśli $c = (1 - \sqrt{5})/2$, to

$$x_n = \left(\frac{1-\sqrt{5}}{2}\right)^n,$$

a jeśli $c = 1$, to

$$x_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^{n+1}.$$

Dla wszystkich $n = 1, 2, \ldots, 30$ obliczyć x_n zarówno ze wzoru rekurencyjnego, jak i z jawnego wyrażenia. Wytłumaczyć wyniki. Ten wzór określa (dla $c = 1$) sławny *ciąg Fibonacciego*.

1.3. Równania różnicowe

Algorytmy numeryczne służą często do obliczania ciągów liczb. Dlatego jest celowe naszkicować tu teorię przestrzeni liniowych ciągów. Będzie ona potrzebna w rozdz. 8, w którym analizuje się metody liniowe wielokrokowe rozwiązywania równań różniczkowych. Zajmujemy się tą teorią już tu, gdyż jej zrozumienie nie wymaga dużej wiedzy matematycznej.

Podstawowe pojęcia

W dalszym ciągu V będzie oznaczać zbiór wszystkich ciągów nieskończonych, takich jak

$$x = \{x_1, x_2, x_3, \ldots\}, \quad y = \{y_1, y_2, y_3, \ldots\}.$$

Formalnie rzecz biorąc, ciąg jest funkcją o wartościach zespolonych, określoną na zbiorze $\mathbb{N} = (1, 2, \ldots)$ liczb naturalnych. Tylko dla wygody wartość takiej funkcji oznaczamy symbolem x_n, a nie $x(n)$.

W zbiorze V określamy dwa działania:

$$x + y := \{x_1 + y_1, x_2 + y_2, x_3 + y_3, \ldots\},$$
$$\lambda x := \{\lambda x_1, \lambda x_2, \lambda x_3, \ldots\}.$$

Krócej wyrażamy to tak:

$$(x + y)_n := x_n + y_n,$$
$$(\lambda x)_n := \lambda x_n.$$

Zbiór V zawiera element zerowy, mianowicie $0 = \{0, 0, \ldots\}$. Dzięki przyjętym definicjom V staje się przestrzenią wektorową. Jest ona nieskończeniewymiarowa, bo następujący układ wektorów z V jest liniowo niezależny:

$$v^{(1)} = \{1, 0, 0, 0, \ldots\},$$
$$v^{(2)} = \{0, 1, 0, 0, \ldots\},$$
$$v^{(3)} = \{0, 0, 1, 0, \ldots\}, \ \ldots$$

Interesują nas operatory liniowe $L\colon V \to V$. Jednym z najważniejszych jest operator *przesunięcia* o symbolu E, określony wzorem

$$Ex := \{x_2, x_3, x_4, \ldots\}, \quad \text{gdy} \quad x = \{x_1, x_2, x_3, \ldots\}.$$

Tak więc

$$(Ex)_n = x_{n+1}.$$

Operator E można stosować wielokrotnie, co daje ogólnie równość

$$(E^k x)_n = x_{n+k} \quad (k \geqslant 0).$$

Oczywiście E^0 jest operatorem identyczności (o symbolu I), czyli

$$(E^0 x)_n = (Ix)_n = x_n.$$

W pozostałej części tego podrozdziału ograniczamy się do operatorów liniowych, które można wyrazić jako kombinacje liniowe potęg operatora E [4]. Nazywamy je operatorami *różnicowymi* (o stałych współczynnikach i skończonego rzędu). Ich ogólna postać jest zatem taka:

$$L = \sum_{i=0}^{m} c_i E^i. \tag{1.3.1}$$

[4] Jednym z nich jest bardzo często stosowany operator Δ; zob. zad. 14 (*przyp. tłum.*).

Stąd wynika, że operatory różnicowe tworzą podprzestrzeń liniową zbioru wszystkich operatorów liniowych z V do V. Potęgi operatora E są bazą tej podprzestrzeni.

Zauważmy, że L w (1.3.1) jest wielomianem względem E. Możemy więc napisać, że

$$L = p(E),$$

gdzie p jest *wielomianem charakterystycznym* operatora L, określonym wzorem

$$p(\lambda) = \sum_{i=0}^{m} c_i \lambda^i.$$

Zadaniem, które tu rozważymy, jest wyznaczenie wszystkich rozwiązań równania *różnicowego liniowego jednorodnego* $Lx = 0$, gdzie L jest operatorem postaci (1.3.1). Jest on liniowy, więc zbiór $\{x\colon Lx = 0\}$, zwany *jądrem* operatora L, jest podprzestrzenią liniową przestrzeni V. Równanie $Lx = 0$ możemy uważać za rozwiązane, jeśli znajdziemy bazę jądra.

Aby zobaczyć, czego można się spodziewać w ogólnym przypadku, rozważmy konkretny przykład operatora L taki, że $m = 2$, $c_0 = 2$, $c_1 = -3$ i $c_2 = 1$. Wtedy równanie $Lx = 0$ ma postać

$$x_{n+2} - 3x_{n+1} + 2x_n = 0 \quad (n \geqslant 1). \tag{1.3.2}$$

Konstrukcja ciągów spełniających to równanie jest bardzo prosta. Istotnie, możemy wybrać dowolnie x_1 i x_2, a następnie obliczać $x_3, x_4, \ldots$ z (1.3.2). Daje to na przykład ciągi

$$\{1, 0, -2, -6, -14, -30, \ldots\},$$
$$\{1, 1, 1, 1, \ldots\},$$
$$\{2, 4, 8, 16, \ldots\}.$$

Trudno zgadnąć, jak wyraża się każdy element pierwszego ciągu. Natomiast dla drugiego i trzeciego jest oczywiście $x_n = \lambda^n$, gdzie odpowiednio $\lambda = 1$ i $\lambda = 2$. Jest naturalnym pytanie, czy istnieją inne podobne rozwiązania. Podstawiając $x_n = \lambda^n$ do (1.3.2), otrzymujemy

$$\lambda^{n+2} - 3\lambda^{n+1} + 2\lambda^n = 0,$$
$$\lambda^n(\lambda^2 - 3\lambda + 2) = 0,$$
$$\lambda^n(\lambda - 1)(\lambda - 2) = 0.$$

Te elementarne rozważania dowodzą, że istnieje jeszcze dokładnie jedno rozwiązanie szukanego typu, mianowicie $\{0, 0, \ldots\}$. Nazywamy je rozwiązaniem *trywialnym*. Wydaje się teraz, że rozwiązania u i v określone odpowiednio wzorami $u_n = 1$ i $v_n = 2^n$ tworzą bazę przestrzeni rozwiązań równania (1.3.2). Aby to sprawdzić, rozważmy dowolne rozwiązanie x tego równania. Szukamy stałych α i β takich, że $x = \alpha u + \beta v$, tzn. $x_n = \alpha u_n + \beta v_n$ dla wszystkich n. W szczególności dla $n = 1$ i $n = 2$ ma być

$$x_1 = \alpha + 2\beta, \quad x_2 = \alpha + 4\beta.$$

Ten układ określa jednoznacznie α i β, gdyż wyznacznik jego macierzy

$$\begin{bmatrix} 1 & 2 \\ 1 & 4 \end{bmatrix}$$

jest różny od 0 (jest to przykład wyznacznika Vandermonde'a określonego w podrozdz. 6.1). Można udowodnić przez indukcję, że $x_n = \alpha u_n + \beta v_n$ dla wszystkich n. Istotnie, jeśli taka równość zachodzi dla wskaźników mniejszych od n, to

$$\begin{aligned} x_n &= 3x_{n-1} - 2x_{n-2} = 3(\alpha u_{n-1} + \beta v_{n-1}) - 2(\alpha u_{n-2} + \beta v_{n-2}) = \\ &= \alpha(3u_{n-1} - 2u_{n-2}) + \beta(3v_{n-1} - 2v_{n-2}) = \alpha u_n + \beta v_n. \end{aligned}$$

Ten przykład ilustruje przypadek wielomianu charakterystycznego o pierwiastkach pojedynczych [5].

Pierwiastki pojedyncze

Twierdzenie 1.3.1. *Jeśli λ jest pierwiastkiem wielomianu p, to ciąg $\{\lambda, \lambda^2, \lambda^3, \ldots\}$ spełnia równanie różnicowe $p(E)x = 0$. Jeśli wszystkie pierwiastki wielomianu p są pojedyncze i różne od 0, to każde rozwiązanie tego równania jest kombinacją liniową tych szczególnych rozwiązań.*

Dowód. Jeśli λ jest dowolną liczbą zespoloną i $u = \{\lambda, \lambda^2, \lambda^3, \ldots\}$, to $Eu = \lambda u$, gdyż

$$(Eu)_n = u_{n+1} = \lambda^{n+1} = \lambda u_n.$$

[5] Będziemy używać terminu *pierwiastek*, gdy rzecz dotyczy wielomianu i terminu *zero* dla innych (bardziej ogólnych) funkcji. W matematyce wyższej w obu przypadkach mówi się o zerach; pierwszy termin był w użyciu dawniej.

Stosując wielokrotnie operator E, otrzymujemy równość $E^i u = \lambda^i u$. Jest tak również dla $i = 0$. Jeśli zatem $p(\lambda) = \sum_{i=0}^{m} c_i \lambda^i$, to

$$p(E)u = \Big(\sum_{i=0}^{m} c_i E^i\Big)u = \sum_{i=0}^{m} c_i (E^i u) = \sum_{i=0}^{m} c_i \lambda^i u = p(\lambda)u.$$

Jeśli $p(\lambda) = 0$, to $p(E)u = 0$, co należało udowodnić.

Niech teraz wszystkie pierwiastki $\lambda_1, \lambda_2, \ldots, \lambda_m$ wielomianu p będą pojedyncze i różne od 0. Każdemu z λ_k odpowiada rozwiązanie

$$u^{(k)} = \{\lambda_k, \lambda_k^2, \lambda_k^3, \ldots\}$$

równania różnicowego $p(E)x = 0$. Niech x oznacza jego dowolne rozwiązanie. Chcemy je wyrazić w postaci $x = \sum_{k=1}^{m} a_k u^{(k)}$. Dla m początkowych elementów ciągów ma zatem być

$$x_i = \sum_{k=1}^{m} a_k \lambda_k^i \quad (1 \leqslant i \leqslant m). \tag{1.3.3}$$

Macierz kwadratowa stopnia m, o elementach λ_k^i jest nieosobliwa, gdyż w przeciwnym razie zachodziłyby nietrywialne równości

$$\sum_{i=1}^{m} b_i \lambda_k^i = 0, \quad \text{czyli} \quad \sum_{i=1}^{m} b_i \lambda_k^{i-1} = 0 \qquad (1 \leqslant k \leqslant m),$$

z których by wynikało, że wielomian stopnia $m-1$ ma m pierwiastków. Tak więc równania (1.3.3) określają jednoznacznie współczynniki $a_1, a_2, \ldots, a_m$. Pozostaje wykazać, że (1.3.3) są spełnione dla wszystkich i. Niech będzie $z = x - \sum_{k=1}^{m} a_k u^{(k)}$. Stąd $p(E)z = 0$, czyli równoważnie $\sum_{i=0}^{m} c_i z_{n+i} = 0$ dla wszystkich n. Inaczej mówiąc,

$$z_{n+m} = -c_m^{-1}(c_0 z_n + c_1 z_{n+1} + \ldots + c_{m-1} z_{n+m-1}) \quad (n \geqslant 1). \tag{1.3.4}$$

Zauważmy, że $c_m \neq 0$, gdyż wielomian p ma m różnych pierwiastków, a więc jest stopnia m. Ponieważ $z_i = 0$ dla $i = 1, 2, \ldots, m$, więc stosując wielokrotnie równość (1.3.4), wnioskujemy, że $z_{m+1} = z_{m+2} = \ldots = 0$. ■

Pierwiastki wielokrotne

Pozostaje rozwiązać równanie różnicowe $p(E)x = 0$ wtedy, gdy p ma pierwiastki wielokrotne. Niech będzie $x(\lambda) = \{\lambda, \lambda^2, \lambda^3, \ldots\}$. W dowodzie tw. 1.3.1 wykazaliśmy, że dla dowolnego wielomianu p jest

$$p(E)x(\lambda) = p(\lambda)x(\lambda).$$

Różniczkowanie stronami względem λ daje równość

$$p(E)x'(\lambda) = p'(\lambda)x(\lambda) + p(\lambda)x'(\lambda).$$

Niech λ będzie pierwiastkiem co najmniej podwójnym wielomianu p. Wtedy $p(\lambda) = p'(\lambda) = 0$ i z dwóch ostatnich równości wynika, że nie tylko $x(\lambda)$, ale i $x'(\lambda) = \{1, 2\lambda, 3\lambda^2, \ldots\}$ jest rozwiązaniem równania różnicowego. Jeśli $\lambda \neq 0$, to te dwa rozwiązania są liniowo niezależne, gdyż

$$\begin{vmatrix} \lambda & \lambda^2 \\ 1 & 2\lambda \end{vmatrix} \neq 0,$$

czyli po obcięciu ciągów do dwóch początkowych elementów wynikająca stąd para wektorów w $\mathbb{R}^2$ jest liniowo niezależna.

Kontynuując to rozumowanie, możemy udowodnić, że jeśli λ jest k-krotnym pierwiastkiem wielomianu p, to następujące ciągi spełniają równanie różnicowe $p(E)x = 0$:

$$\begin{aligned} x(\lambda) &:= \{\lambda, \lambda^2, \lambda^3, \ldots\}, \\ x'(\lambda) &:= \{1, 2\lambda, 3\lambda^2, \ldots\}, \\ x''(\lambda) &:= \{0, 2, 6\lambda, \ldots\}, \ldots, \\ x^{(k-1)}(\lambda) &:= \frac{d^{k-1}}{d\lambda^{k-1}}\{\lambda, \lambda^2, \lambda^3, \ldots\}. \end{aligned}$$

Twierdzenie 1.3.2. *Niech p będzie wielomianem takim, że $p(0) \neq 0$. Wtedy dla każdego k-krotnego pierwiastka λ tego wielomianu do bazy jądra operatora $p(E)$ należą określone wyżej ciągi $x(\lambda), x'(\lambda), \ldots, x^{(k-1)}(\lambda)$.*

Przykład 1.3.3. Znaleźć ogólne rozwiązanie równania różnicowego

$$4x_{n+3} + 7x_{n+2} + 2x_{n+1} - x_n = 0.$$

Rozwiązanie. Jest to równanie $p(E)x = 0$, gdzie

$$p(\lambda) := 4\lambda^3 + 7\lambda^2 + 2\lambda - 1 = (\lambda + 1)^2(4\lambda - 1).$$

Ten wielomian ma pierwiastek podwójny -1 i pierwiastek pojedynczy $\frac{1}{4}$. Bazowymi rozwiązaniami są ciągi

$$\begin{aligned} x(-1) &:= \{-1, 1, -1, 1, \ldots\}, \\ x'(-1) &:= \{1, -2, 3, -4, \ldots\}, \\ x(\tfrac{1}{4}) &:= \{\tfrac{1}{4}, \tfrac{1}{16}, \tfrac{1}{64}, \ldots\}, \end{aligned}$$

a ogólne rozwiązanie wyraża się wzorem

$$x = \alpha x(-1) + \beta x'(-1) + \gamma x(\tfrac{1}{4}).$$

Inaczej mówiąc,

$$x_n = \alpha(-1)^n + \beta(-1)^{n-1}n + \gamma(\tfrac{1}{4})^n. \qquad \blacksquare$$

Równania różnicowe stabilne

Ciąg $x = \{x_1, x_2, \ldots\}$ należący do przestrzeni V nazywamy *ograniczonym*, jeśli istnieje taka stała c, że $|x_n| \leqslant c$ dla wszystkich n. Równanie różnicowe $p(E)x = 0$ nazywamy *stabilnym*, jeśli jego wszystkie rozwiązania są ograniczone. Równanie (1.3.2) nie jest stabilne, gdyż dla jednego z jego rozwiązań $x_n = 2^n$. (Uwarunkowanie i stabilność rozważane w innych miejscach książki nie mają związku z pojęciem stabilności równania różnicowego).

Prosta metoda ustalenia, czy równanie różnicowe jest stabilne, wynika z poniższego twierdzenia.

Twierdzenie 1.3.4. *Jeśli wielomian p jest taki, że $p(0) \neq 0$, to równanie różnicowe $p(E)x = 0$ jest stabilne wtedy i tylko wtedy, gdy wszystkie pierwiastki tego wielomianu leżą w kole $|z| \leqslant 1$, a pierwiastki wielokrotne w kole $|z| < 1$.*

Dowód. Załóżmy, że wielomian p ma własności opisane w twierdzeniu. Niech λ będzie jego pierwiastkiem. Wtedy jednym z rozwiązań równania różnicowego jest $x(\lambda) = \{\lambda, \lambda^2, \lambda^3, \ldots\}$. Ponieważ $|\lambda| \leqslant 1$, więc ten ciąg jest ograniczony. Jeśli λ jest pierwiastkiem wielokrotnym, to co najmniej pierwszy z ciągów $x'(\lambda), x''(\lambda), \ldots$ też spełnia równanie różnicowe. W tym przypadku zakładamy, że $|\lambda| < 1$. Znana reguła de l'Hospitala pozwala wykazać, że

$$\lim_{n\to\infty} n^k\lambda^n = 0 \quad (k \geqslant 0),$$

czyli każdy z ciągów $x'(\lambda), x''(\lambda), \ldots$ jest ograniczony (zob. zad. 20).

Załóżmy teraz, że wielomian p nie ma własności podanych w twierdzeniu. Jeśli pierwiastek λ wielomianu p jest taki, że $|\lambda| > 1$, to ciąg $x(\lambda)$ nie jest ograniczony. Jeśli p ma pierwiastek wielokrotny λ taki, że $|\lambda| \geqslant 1$, to ciąg $x'(\lambda)$ nie jest ograniczony, gdyż jego elementy spełniają nierówność

$$|x_n| = |n\lambda^{n-1}| = n|\lambda|^{n-1} \geqslant n. \qquad \blacksquare$$

Przykład 1.3.5. Sprawdzić, czy równanie różnicowe z przykł. 1.3.3 jest stabilne.

Rozwiązanie. Wiemy, że dla tego równania wielomian p ma pierwiastek podwójny -1 i pojedynczy $\frac{1}{4}$. Równanie jest zatem *niestabilne*. ■

Przykład równania różnicowego o zmiennych współczynnikach można znaleźć w teorii funkcji Bessela. *Funkcja Bessela* J_n jest określona wzorem

$$J_n(x) := \frac{1}{\pi}\int_0^\pi \cos(x\sin\theta - n\theta)\,d\theta.$$

Z tej definicji wynika od razu, że $|J_n(x)| \leqslant 1$. Nie jest tak oczywisty (ale jest prawdziwy) wzór rekurencyjny

$$J_n(x) = 2(n-1)x^{-1}J_{n-1}(x) - J_{n-2}(x).$$

Jeśli dla pewnego x znamy wartości $J_0(x)$ i $J_1(x)$, to stosując ten wzór możemy obliczyć $J_2(x), J_3(x), \ldots, J_n(x)$. To postępowanie staje się jednak niestabilne i nieskuteczne dla $2n > |x|$, gdyż nieuniknione błędy zaokrągleń są mnożone przez czynnik $2nx^{-1}$, a ten wtedy może być duży (zob. zad. K2).

Dalsze informacje o obliczaniu funkcji za pomocą związków rekurencyjnych można znaleźć w następujących publikacjach: Abramowitz i Stegun [1964, s. xiii], Cash [1979], Gautschi [1961, 1967, 1975], Wimp [1984].

Zadania 1.3

1. Wyrazić pierwszy z trzech ciągów podanych po (1.3.2) przez dwa pozostałe.
2. Niech p będzie wielomianem stopnia m. Czy przestrzeń rozwiązań równania $p(E)x = 0$ ma zawsze wymiar m?
3. Niech p będzie wielomianem stopnia m i niech będzie $p(0) \neq 0$. Udowodnić, że jeśli ciąg x zawiera m kolejnych zerowych elementów i $p(E)x = 0$, to $x = 0$.
4. Czy operator E jest iniektywny? Czy ma on prawą lub lewą odwrotność? Czy jest suriektywny? Odpowiedzieć na te same pytania dla operatora F określonego wzorami $(Fx)_n = x_{n-1}$ i $F(x)_1 = 0$. Wyjaśnić związek między E i F. Przypuśćmy, że V jest określone na nowo jako przestrzeń wszystkich funkcji na zbiorze liczb całkowitych i zdefiniujmy F wzorem $(Fx)_n = x_{n-1}$. Czy to zmienia odpowiedzi na postawione pytania?
5. Jakie są wartości własne i wektory własne operatora E?
6. Rozważmy szereg nieskończony $\sum_{n=1}^{\infty} x_n v^{(n)}$. Co można powiedzieć o jego zbieżności? Wykazać, że $x = \sum_{n=1}^{\infty} x_n v^{(n)}$ w sensie zbieżności punktowej.
7. Niech układ $(v^{(1)}, v^{(2)}, \ldots)$ ciągów określonych na początku podrozdz. 1.3 będzie bazą zbioru V. Wykazać, że sumę $\sum_{i=0}^{m} c_i E^i$ można wyrazić za pomocą macierzy nieskończonej.

8. (cd.). Wykazać, że dowolne dwa operatory postaci opisanej w poprzednim zadaniu komutują.

9. Wykazać, że jeśli L_1 i L_2 są kombinacjami liniowymi potęg operatora E oraz jeśli $L_1 x = 0$, to $L_1 L_2 x = 0$.

10. Opracować pełną teorię równania różnicowego $E^r x = 0$.

11. Podać bazy złożone z ciągów o elementach rzeczywistych dla przestrzeni rozwiązań każdego z poniższych równań różnicowych.

(**a**) $(4E^0 - 3E^2 + E^3)x = 0$

(**b**) $(3E^0 - 2E + E^2)x = 0$

(**c**) $(2E^6 - 9E^5 + 12E^4 - 4E^3)x = 0$

12. Wykazać, że jeśli wielomian p ma współczynniki rzeczywiste i jeśli z jest rozwiązaniem (zespolonym) równania $p(E)z = 0$, to ciąg sprzężony względem z oraz części rzeczywista i urojona ciągu z też spełniają to równanie.

13. Rozwiązać równania: (**a**) $x_{n+1} - nx_n = 0$, (**b**) $x_{n+1} - x_n = n$, (**c**) $x_{n+1} - x_n = 2$.

14. Określmy operator Δ wzorem

$$\Delta x = \{x_2 - x_1, x_3 - x_2, x_4 - x_3, \ldots\}.$$

Wykazać, że $E = I + \Delta$. Wykazać, że jeśli p jest wielomianem, to

$$p(E) = p(I) + p'(I)\Delta + \frac{1}{2}p''(I)\Delta^2 + \frac{1}{3!}p'''(I)\Delta^3 + \ldots + \frac{1}{m!}p^{(m)}(I)\Delta^m.$$

15. (cd.). Udowodnić, że jeśli $x = \{\lambda, \lambda^2, \lambda^3, \ldots\}$, a p jest wielomianem, to $p(\Delta)x = p(\lambda - 1)x$. Opisać sposób rozwiązania równania wyrażonego w postaci $p(\Delta)x = 0$.

16. (cd.). Pokazać, że

$$\Delta^n = (-1)^n[E^0 - nE + \tfrac{1}{2}n(n-1)E^2 - \tfrac{1}{3!}n(n-1)(n-2)E^3 + \ldots + (-1)^n E^n].$$

17. Podać kompletny dowód tw. 1.3.2.

18. Podać jądro operatora $p(E)$, gdy p jest wielomianem i $p(0) = 0$.

19. Niech dla $\lambda \in \mathbb{C}$ będzie $x(\lambda) := \{\lambda, \lambda^2, \ldots\}$. Wykazać, że jeśli liczby zespolone $\lambda_1, \lambda_2, \ldots, \lambda_m$ są niezerowe i parami różne, to ciągi $x(\lambda_1), x(\lambda_2), \ldots, x(\lambda_m)$ są niezależne liniowo w V.

20. Udowodnić, że jeśli $\mu \in (0, \infty)$ i $|\lambda| < 1$, to $\lim_{n\to\infty} n^\mu \lambda^n = 0$.

21. Udowodnić tw. 1.3.4, odrzucając założenie, że $p(0) \neq 0$.

22. Sprawdzić, czy równanie różnicowe $x_n = x_{n-1} + x_{n-2}$ jest stabilne.

23. Udowodnić, że jeśli x spełnia równanie różnicowe $p(E)x = 0$, to Ex ma tę samą własność.

24. Wykazać, że ogólne rozwiązanie równania $x_n = 2(x_{n-1} + x_{n-2})$ ma postać $z_n = \alpha(1 + \sqrt{3})^n + \beta(1 - \sqrt{3})^n$. Wykazać, że rozwiązanie z początkowymi wartościami $x_1 = 1$ i $x_2 = 1 - \sqrt{3}$ otrzymujemy dla $\alpha = 0$ i $\beta = (1 - \sqrt{3})^{-1}$.

ZADANIA KOMPUTEROWE 1.3

K1. Jednym z rozwiązań równania różnicowego $x_{n+2} - 2x_{n+1} - 2x_n = 0$ jest ciąg o elementach $x_n = (1-\sqrt{3})^{n-1}$, które są na przemian dodatnie i ujemne i które dążą do 0. Obliczyć i wydrukować 100 początkowych liczb x_n, korzystając ze wzoru $x_{n+2} = 2(x_{n+1} + x_n)$ dla $x_1 = 1$ i $x_2 = 1 - \sqrt{3}$. Wyjaśnić, skąd bierze się osobliwe zachowanie wyników.

K2. Przyjmując, że $J_0(1) = 0.76519\,76866$ i $J_1(1) = 0.44005\,05857$ są wartościami funkcji Bessela, obliczyć wartości $J_2(1), J_3(1), \ldots, J_{20}(1)$ za pomocą wzoru rekurencyjnego podanego na końcu podrozdz. 1.3. Dlaczego jest oczywiste, że wyniki są obarczone dużymi błędami?

K3. Sprawdzić, czy równanie $4x_{n+2} - 8x_{n+1} + 3x_n = 0$ jest stabilne. Znaleźć jego ogólne rozwiązanie. Dla $x_0 = 0$ i $x_1 = -2$ obliczyć x_{100} w najbardziej ekonomiczny sposób.

K4. Obliczyć elementy do setnego włącznie rozwiązania szczególnego z zad. 24 trzema metodami:

(**a**) x_n obliczane wprost ze związku rekurencyjnego.

(**b**) $y_n = \beta(1 - \sqrt{3})^n$.

(**c**) $z_n = \alpha(1 + \sqrt{3})^n + \beta(1 - \sqrt{3})^n$, gdzie α jest równe precyzji arytmetyki w użytym komputerze (zob. podrozdz. 2.1).

Porównać wyniki.

K5. Rozwiązać numerycznie równanie $x_{n+2} - (\pi + \pi^{-1})x_{n+1} + x_n = 0$ dla $x_0 = 1$ i $x_1 = \pi$. Znaleźć błąd względny obliczonego x_{50}. Powtórzyć obliczenia dla $x_1 = \pi^{-1}$ i wyjaśnić, dlaczego błędy względne w tych dwóch przypadkach się różnią.

ROZDZIAŁ 2

Arytmetyka komputerowa

2.0. Wstęp

W tym rozdziale wyjaśniamy, na czym polega arytmetyka zmiennopozycyjna i opisujemy podstawowe fakty dotyczące błędów zaokrągleń, które mogą zakłócać wyniki obliczeń. Omawiamy także przyczyny redukcji liczby cyfr znaczących (np. odejmowanie dwóch prawie identycznych liczb) i sposoby zapobiegania temu niebezpiecznemu zjawisku. Na koniec, dajemy przegląd pewnych algorytmów stabilnych lub niestabilnych i zadań źle lub dobrze uwarunkowanych.

2.1. Arytmetyka zmiennopozycyjna

Większość komputerów pracuje w układzie dwójkowym, a nie dziesiętnym, którego używamy na co dzień. Liczba 2 jest podstawą układu dwójkowego w takim samym sensie, jak liczba 10 jest podstawą układu dziesiętnego. Aby to zrozumieć, przypomnijmy sobie najpierw znany nam sposób wyrażania liczb. Symbol liczby rzeczywistej, na przykład 427.325, w *układzie dziesiętnym* znaczy tyle, że

$$427.325 = 4 \times 10^2 + 2 \times 10^1 + 7 \times 10^0 + 3 \times 10^{-1} + 2 \times 10^{-2} + 5 \times 10^{-3}.$$

Suma po prawej stronie zawiera potęgi podstawy 10 i cyfry, którymi mogą być 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Jeśli dopuścimy, że na prawo od kropki dziesiętnej może występować nieskończenie wiele cyfr, to każda liczba rzeczywista

da się wyrazić tak jak wyżej, wraz z odpowiednim znakiem (+ lub −). Na przykład, liczbę $-\pi$ wyrażamy tak:

$$-\pi = -3.14159\,26535\,89793\,23846\,26433\,8\ldots$$

Ostatnia podana tu cyfra 8 oznacza 8×10^{-26}.

W układzie dwójkowym stosuje się tylko dwie cyfry: 0 i 1; nazywamy je *bitami*. Typową liczbę wyrażoną w *układzie dwójkowym* (czyli *binarnym*) interpretujemy podobnie jak wyżej liczbę w układzie dziesiętnym. Jest na przykład

$$(1001.11101)_2 = 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 + \\ + 1 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3} + 0 \times 2^{-4} + 1 \times 2^{-5}.$$

Ta sama liczba w układzie dziesiętnym wyraża się jako 9.90625.

Ogólniej, każda liczba naturalna $\beta > 1$ może być *podstawą* (*bazą*) *układu pozycyjnego*. Liczby w tym układzie reprezentujemy używając β cyfr $0, 1, \ldots, \beta - 1$. Jeśli z kontekstu nie wynika, w jakim układzie wyrażamy liczbę N, to możemy to uściślić za pomocą symbolu $(N)_\beta$ zastosowanego już wyżej dla $\beta = 2$. Warto dodać, że inne często stosowane podstawy to $\beta = 8$ (daje układ *ósemkowy*, w którym używa się cyfr $0, 1, \ldots 7$) i $\beta = 16$ (układ *szesnastkowy* z cyframi oznaczanymi tradycyjnie $0, 1, \ldots, 9$, A, B,...,F). Te układy są związane w oczywisty sposób z układem dwójkowym: trójka (czwórka) bitów daje jedną cyfrę ósemkową (szesnastkową).

Ponieważ typowy komputer pracuje wewnętrznie w układzie dwójkowym, a komunikuje się z ludźmi w przyjętym przez nich układzie dziesiętnym, więc musi on stosować procedury *konwersji*, czyli przejścia od jednego układu do drugiego. Odbywa się to na wejściu i wyjściu. Zazwyczaj użytkownik nie ma do czynienia z tymi konwersjami, powinien jednak wiedzieć, że każda z nich może powodować pewne błędy.

Komputery nie potrafią operować na liczbach rzeczywistych mających dowolną liczbę cyfr. Dokładność, z jaką można te liczby przedstawiać, zależy od długości słów w komputerze. Jednak nawet tak prosta liczba jak 1/10 nie może być zapamiętana dokładnie w jakimkolwiek komputerze dwójkowym, gdyż jej rozwinięcie dwójkowe jest nieskończone:

$$\tfrac{1}{10} = (0.0\,0011\,0011\,0011\,0011\ldots)_2.$$

Dlatego, jeśli liczba 0.1 zostanie zapamiętana (w przybliżeniu) w postaci dwójkowej, a następnie wydrukowana z 40 cyframi dziesiętnymi, to wynik może być następujący:

0.10000 00014 90116 11938 47656 25000 00000 00000.

Zwykle nie zauważamy tego błędu konwersji, gdyż standardowy format drukowania, uwzględniający specyfikę komputera, daje np. 0.10000000.

Zaokrąglanie

Zaokrąglanie będzie omawiane szczegółowo dalej. Tu zajmujemy się nim tylko o tyle, o ile jest to istotne w obliczeniach ręcznych lub na kalkulatorze. Wyniki pośrednie obliczeń mają na ogół coraz więcej cyfr, a liczba cyfr znaczących pozostaje stała lub maleje. Na przykład, iloczyn liczb mających po osiem cyfr po kropce ma tam cyfr szesnaście.

Zaokrąglanie jest ważnym pojęciem w obliczeniach naukowych. Rozważmy dodatnią liczbę dziesiętną x postaci $\square.\square\square\square\ldots\square\square\square$ z m cyframi po kropce. Sposób *zaokrąglenia* liczby x do n cyfr dziesiętnych ($n < m$) zależy od wartości $(n+1)$-szej cyfry. Jeśli jest nią 0, 1, 2, 3 lub 4, to n-tej cyfry nie zmieniamy, a pozostałe odrzucamy. Jeśli natomiast $(n+1)$-szą cyfrą jest 5, 6, 7, 8 lub 9, to po odrzuceniu cyfr jak wyżej dodajemy do liczby 10^{-n}. (Jeśli $(n+1)$-szą cyfrą jest 5, to przyjmuje się niekiedy inny sposób zaokrąglenia, mianowicie zaokrąglenie w górę tylko wtedy, gdy daje ono parzystą n-tą cyfrę – to zdarza się mniej więcej w połowie przypadków. Dla uproszczenia przyjmujemy powyższą jednolitą regułę.)

Oto kilka przykładów poprawnego zaokrąglenia liczb siedmiocyfrowych do czterech cyfr po kropce:

$$\begin{aligned} 0.1735499 &\to 0.1735 \\ 0.9999500 &\to 1.0000 \\ 0.4321409 &\to 0.4321 \end{aligned}$$

Jeśli liczba dodatnia x jest zaokrąglona do przybliżenia $\tilde{x}$ mającego n cyfr po kropce, to

$$|x - \tilde{x}| \leqslant \tfrac{1}{2} \times 10^{-n}. \tag{2.1.1}$$

Istotnie, jeśli $(n+1)$-szą cyfrą liczby x jest 0, 1, 2, 3 lub 4, to $x = \tilde{x} + \varepsilon$, gdzie $0 \leqslant \varepsilon < \frac{1}{2} \times 10^{-n}$ i (2.1.1) zachodzi; jeśli zaś tą cyfrą jest 5, 6, 7, 8 lub 9, to $\tilde{x} = \hat{x} + 10^{-n}$, gdzie $\hat{x}$ powstaje z x przez odrzucenie wszystkich cyfr począwszy od $(n+1)$-szej. Wtedy $x = \hat{x} + \delta \times 10^{-n}$, gdzie $\frac{1}{2} \leqslant \delta < 1$, czyli $\tilde{x} - x = (1 - \delta) \times 10^{-n}$ i (2.1.1) zachodzi.

Dla liczby dziesiętnej $x > 0$ jej *obciętym* n-cyfrowym przybliżeniem jest liczba $\hat{x}$ określona wyżej. To przybliżenie jest takie, że

$$|x - \hat{x}| < 10^{-n}.$$

Istotnie, $x = \hat{x} + \delta \times 10^{-n}$, gdzie $0 \leqslant \delta < 1$, więc $|x - \hat{x}| < 10^{-n}$. Jak widać, obcięcie może spowodować dwa razy większy błąd niż zaokrąglenie.

Reprezentacja zmiennopozycyjna liczb

W układzie dziesiętnym każdą liczbę rzeczywistą $x \neq 0$ można wyrazić w postaci *zmiennopozycyjnej*:

$$x = \pm r \times 10^n,$$

gdzie $r \in [1, 10)$ (w innym wariancie $r \in [0.1, 1)$), a n jest liczbą całkowitą (dodatnią, ujemną lub zerem)[1]. Jeśli $x = 0$, to $r = 0$ (a n może być dowolne); w przeciwnym razie n dobieramy tak, aby r leżało w wybranym przedziale. Liczbę r nazywamy *mantysą*, a n – *cechą* liczby x.

W układzie dwójkowym postać zmiennopozycyjna liczby x różnej od 0 jest opisana podobnym wzorem:

$$x = \pm q \times 2^m, \tag{2.1.2}$$

gdzie $q \in [1, 2)$ (w innym wariancie $q \in [0.5, 1)$), a m jest liczbą całkowitą. Przyjmuje się, że mantysa q i cecha m są wyrażone w układzie dwójkowym.

W typowym komputerze liczby są reprezentowane tak jak to opisano wyżej, jednak z koniecznymi ograniczeniami na q i m, które wynikają z ustalonej długości słowa. Opiszemy teraz taki sposób wyrażania liczb, który jest wzorowany na standardzie IEEE (Nr 754 z 1985 r.) arytmetyki zmiennopozycyjnej[2]. Wtedy pierwszy bit może sygnalizować znak (+ lub −) liczby, 8 następnych bitów – cechę m zwiększoną o 127, a pozostałe 23 bity są przeznaczone na część ułamkową mantysy q. Ta informacja wymaga pewnych uściśleń.

Po pierwsze, przyjmujemy, że cecha jest liczbą całkowitą z przedziału $[-126, 127]$. Wtedy $m + 127 \in [1, 254]$. Na ośmiu bitach można zapamiętać liczby całkowite od 0 czyli $(00\,000\,000)_2$ do 255 czyli $(11\,111\,111)_2$. Wyżej tych dwóch skrajnych wartości nie wykorzystano, bo rezerwuje się je dla specjalnych celów; wyjaśni się to nieco dalej.

Po drugie, mantysa liczby różnej od zera ma z założenia część całkowitą równą 1, więc nie warto na nią tracić miejsca. Dlatego właśnie zapamiętuje się tylko bity części ułamkowej mantysy. Dla najmniejszej mantysy, równej 1, wszystkie 23 bity tej części słowa są zerami, największą jest $2 - 2^{-23}$ i wtedy te bity są jedynkami.

Wiadomo już, że wszystkie liczby, które można zapamiętać w komputerze i na których można wykonywać działania arytmetyczne, mieszczą się

[1] W dalszym ciągu liczbę $r \times 10^n$ będziemy na ogół pisać w postaci $r_{10}n$ (*przyp. tłum.*).

[2] IEEE – Institute of Electrical and Electronic Engineers (*przyp. tłum.*).

w przedziale od $2^{-126} \approx 1.2_{10}{-}38$ do $(2-2^{-23}) \times 2^{127} \approx 3.4_{10}38$ lub w analogicznym przedziale dla liczb ujemnych; do tego dochodzi wyróżniona liczba 0. Prócz tego, część ułamkowa mantysy ma co najwyżej 23 bity. Każdą liczbę spełniającą te warunki nazywamy *liczbą maszynową* (nieco archaiczna, ale wygodna nazwa). Jest oczywiste, że nie każda liczba rzeczywista jest liczbą maszynową – nawet „prosta" liczba 1/100 nie da się dokładnie wyrazić w opisany wyżej sposób. Jeśli taka liczba ma być wprowadzona do komputera lub jeśli jest wynikiem działania arytmetycznego, to zastąpienie jej przez możliwie bliską liczbę maszynową wywołuje nieuchronny błąd. Szczegóły będą wyjaśnione nieco dalej.

Powróćmy jeszcze do standardu IEEE. Przyjęto tam, że zero w pojedynczej precyzji ma dwie postaci, $+0$ i -0, reprezentowane w komputerze jako słowa $(00000000)_{16}$ i $(80000000)_{16}$. Zero w jednej z tych wersji jest w szczególności skutkiem działania, którego dokładny wynik jest wprawdzie różny od 0, ale ma zbyt małą cechę: $m < -126$. Mówimy, że wtedy powstał *niedomiar*.

Nieskończoność ma również dwie odmiany, $+\infty$ i $-\infty$, wyrażone odpowiednio słowami $(7F800000)_{16}$ i $(FF800000)_{16}$. Nieskończoność może być skutkiem działania, którego dokładny wynik ma zbyt dużą cechę: $m > 127$. Mówimy, że wtedy powstał *nadmiar*. Nieskończoność jest traktowana jako bardzo duża liczba. Przypuśćmy np., że x jest liczbą maszynową z przedziału $(0, \infty)$. Wtedy wynikiem każdego z działań $x + \infty$, $x \times \infty$ i ∞/x jest $+\infty$, natomiast x/∞ daje $+0$ (tu ∞ rozumiemy jako $+\infty$). Podobne informacje dotyczą wielkości $-\infty$. Jeśli jednak nieskończoność miałaby być argumentem działania, które nie ma sensu, to w wielu komputerach wykonanie programu jest automatycznie przerywane. Natomiast niedomiar jest na ogół co najwyżej sygnalizowany użytkownikowi.

Dodatkowo uznano za potrzebne kodowanie symbolu `NaN` (od słów angielskich *Not a Number*); sygnalizuje on, że pewne działanie, np. $0/0$, $\infty - \infty$ lub $x +$ `NaN`, jest niewykonalne. `NaN` jest reprezentowane przez słowo komputerowe, w którym na 8 bitach przeznaczonych dla cechy występują same jedynki i gdzie co najmniej jeden z 23 bitów odpowiadających mantysie jest jedynką.

Powyższe informacje, choć częściowo tylko przykładowe, ułatwiają zrozumienie tego, z jakimi liczbami mamy do czynienia, stosując w obliczeniach konkretny typ *arytmetyki zmiennopozycyjnej*. Może on zależeć od stosowanego komputera (ten pracuje z reguły w układzie dwójkowym) i używanego języka programowania. Języki takie jak Pascal, Fortran i C++ pozwalają na stosowanie liczb zmiennopozycyjnych kilku typów (zresztą liczb całkowitych też). Wybierając jeden z nich, uwzględniamy przede wszystkim następujące informacje:

(a) Jakie są dopuszczalne wartości cechy m? To określa z grubsza zakres liczb maszynowych, tj. ich najmniejszą (jeśli pominąć zero) i największą wartość bezwzględną.

(b) Jaka jest długość części słowa przeznaczonego na mantysę q (bez jej znaku)? Ta informacja określa z grubsza dokładność obliczeń. Jeśli ta część składa się z t bitów (przy takiej konwencji jak w przykładzie, gdzie było $t = 23$), to liczbę 2^{-t-1} nazywamy *precyzją arytmetyki* dla danego typu komputera i stosowanej przez nas arytmetyki. W dalszym ciągu ta wielkość będzie oznaczana symbolem ε.

Oczywiście warto też wiedzieć, czy działania arytmetyczne na liczbach danego typu są wykonywane sprzętowo, czy programowo, gdyż w tym drugim przypadku czas działania może być znacznie dłuższy. W obliczeniach nie są natomiast istotne szczegóły reprezentacji liczb zmiennopozycyjnych w pamięci komputera – to np., czy cecha zajmuje początkowe, czy ostatnie bity w słowie (układzie bajtów) przeznaczonym na liczbę, albo jak jest pamiętany znak liczby.

Nie wdając się w szczegóły, warto tu podkreślić, że w obliczeniach naukowych taka długość mantysy, jaką podano wyżej w przykładzie, może nas nie zadowalać. Wtedy pewne obliczenia powinny być wykonywane co najmniej w *podwójnej precyzji*. Jeśli np. liczba zmiennopozycyjna w pojedynczej precyzji jest pamiętana jako słowo 32-bitowe, to analogiczna liczba w podwójnej precyzji zajmuje dwa takie słowa, dzięki czemu mantysa jest ponad dwukrotnie dłuższa. Także zakres możliwych cech może być znacznie większy. Działania wykonywane w podwójnej precyzji są też co najmniej dwa razy wolniejsze, gdyż na ogół są programowane, a nie wykonywane sprzętowo.

Rozkład liczb zmiennopozycyjnych w komputerze jest nierównomierny. Między kolejnymi potęgami dwójki znajduje się tyle samo liczb maszynowych. Dlatego znaczna ich część skupia się w pobliżu zera, ale pewne otoczenie zera stanowi lukę – liczb maszynowych (tzw. znormalizowanych) tam nie ma.

Na zakończenie tego fragmentu podajemy listę książek i artykułów poświęconych standardom IEEE i pojęciom z nimi związanym: ANSI/IEEE [1985, 1987], Cody [1988], Coonen [1981], Fosdick [1993], Hough [1981], Overton [2001], Raimi [1969] i Scott [1985]; zob. też Biernat [*2001].

Liczby rzeczywiste i liczby maszynowe

Jak już podkreślono, nie każda liczba rzeczywista x jest liczbą maszynową. Trzeba więc na ogół zastąpić x jakąś bliską liczbą maszynową. Tę ostatnią można wybierać na kilka sposobów.

Niech będzie $x > 0$ i $x = q \times 2^m$, gdzie $1 \leqslant q < 2$. Stąd

$$x = (1.a_1a_2\ldots)_2 \times 2^m,$$

gdzie wszystkie a_i są równe 0 lub 1. Jeśli mantysy liczb maszynowych mają t bitów po kropce, to bliska względem x taka liczba powstaje przez odrzucenie zbędnych bitów $a_{t+1}, a_{t+2}, \ldots$ Taką czynność nazywamy *obcięciem*. Daje ono liczbę

$$x_- = (1.a_1a_2\ldots a_t)_2 \times 2^m.$$

Zauważmy, że $x_- \leqslant x$. Inna bliska liczba maszynowa, leżąca na prawo od x, powstaje przez *zaokrąglenie w górę*; odrzucamy więc jak przedtem zbędne bity, ale do q dodajemy jedynkę na ostatniej zachowanej pozycji. Daje to liczbę

$$x_+ = [(1.a_1a_2\ldots a_t)_2 + 2^{-t}] \times 2^m.$$

Jest oczywiście $x_+ - x_- = 2^{m-t}$.

Najbliższą względem x liczbą maszynową, oznaczaną symbolem $\mathrm{fl}(x)$, jest bliższa jej z liczb x_- i x_+. Jeśli bliższa jest pierwsza z nich, to $\mathrm{fl}(x) = x_-$ i

$$|x - \mathrm{fl}(x)| \leqslant \frac{1}{2}|x_+ - x_-| = 2^{m-t-1}.$$

W przeciwnym razie $\mathrm{fl}(x) = x_+$ i

$$|x - \mathrm{fl}(x)| \leqslant \frac{1}{2}|x_+ - x_-| = 2^{m-t-1}.$$

W obu przypadkach *błąd względny* reprezentacji maszynowej liczby x szacujemy tak:

$$\left|\frac{x - \mathrm{fl}(x)}{x}\right| \leqslant \frac{2^{m-t-1}}{2^m q} = \frac{1}{q}2^{-t-1} \leqslant 2^{-t-1}.$$

Wprowadziwszy wielkość $\delta = (x^* - x)/x$ wnioskujemy stąd, że

$$\mathrm{fl}(x) = x(1+\delta), \quad \text{gdzie} \quad |\delta| \leqslant \varepsilon. \tag{2.1.3}$$

Użyto tu wprowadzonego nieco wcześniej symbolu ε precyzji arytmetyki.

Przykład 2.1.1. Jaka jest postać dwójkowa liczby $x = 2/3$ w przykładowej arytmetyce, w której $t = 23$? Jakie są dla niej dwie bliskie liczby maszynowe x_- i x_+? Którą z tych liczb jest $\mathrm{fl}(x)$? Jaki jest błąd zaokrąglenia, bezwzględny i względny, wywołany zamianą x na $\mathrm{fl}(x)$?

Rozwiązanie. Aby odpowiedzieć na pierwsze pytanie, napiszmy równość

$$\tfrac{2}{3} = (0.a_1a_2a_3\ldots)_2.$$

Mnożenie przez 2 daje

$$\tfrac{4}{3} = (a_1.a_2a_3\ldots)_2.$$

Część całkowita obu stron jest równa $1 = a_1$. Po jej odjęciu mamy

$$\tfrac{1}{3} = (0.a_2a_3a_4\ldots)_2.$$

Powtarzając te czynności, wnioskujemy ostatecznie, że

$$x = \tfrac{2}{3} = (0.101010\ldots)_2 = (1.010101\ldots)_2 \times 2^{-1}.$$

Dwie bliskie liczby maszynowe są równe

$$x_- = (1.0101\ldots010)_2 \times 2^{-1},$$
$$x_+ = (1.0101\ldots011)_2 \times 2^{-1},$$

gdzie po kropce mamy 23 bity. Pierwszą liczbę daje obcięcie, drugą – zaokrąglenie w górę.

Aby stwierdzić, którą z liczb x_- i x_+ należy wybrać jako $\mathrm{fl}(x)$, obliczamy różnice

$$x - x_- = (0.1010\ldots)_2 \times 2^{-24} = \tfrac{2}{3} \times 2^{-24},$$
$$x_+ - x = (x_+ - x_-) - (x - x_-) = 2^{-24} - \tfrac{2}{3} \times 2^{-24} = \tfrac{1}{3} \times 2^{-24}.$$

Dlatego $\mathrm{fl}(x) = x_+$ i błąd bezwzględny zaokrąglenia jest równy

$$|\mathrm{fl}(x) - x| = \tfrac{1}{3} \times 2^{-24},$$

a błąd względny zaokrąglenia jest równy

$$\frac{|\mathrm{fl}(x) - x|}{|x|} = \frac{\frac{1}{3} \times 2^{-24}}{\frac{2}{3}} = 2^{-25}.$$ ■

Równość (2.1.3) dotyczy – ogólniej – komputerów, w których stosuje się układ pozycyjny o podstawie β, a liczby zmiennopozycyjne mają n-cyfrowe mantysy. Wtedy $\varepsilon = \frac{1}{2}\beta^{1-n}$ w przypadku zaokrąglenia i $\varepsilon = \beta^{1-n}$ w przypadku obcięcia. W nowoczesnych komputerach wartość ε zmienia się w szerokich granicach wobec różnych długości słowa, różnych układów pozycyjnych, rodzajów zaokrąglenia itd. Długość słowa waha się od 64 lub 60 bitów dla dużych komputerów, stosowanych w obliczeniach naukowych, przez

32 lub 36 bitów dla komputerów średniej wielkości, do 16 bitów w niektórych komputerach osobistych. Kalkulatory osobiste mogą mieć jeszcze mniejszą dokładność. Wiele komputerów stosuje arytmetykę dwójkową, ale szesnastkowa i ósemkowa też są używane. Stosuje się różne rodzaje zaokrąglenia; pewne kompilatory pozwalają nawet użytkownikowi wybrać jeden z tych rodzajów.

Działania arytmetyczne na liczbach zmiennopozycyjnych

Rozważymy teraz skutki wykonywania działań arytmetycznych na liczbach zmiennopozycyjnych. Dokładny wynik takiego działania na ogół nie jest taką liczbą. Niech symbol $\odot$ oznacza jedno z czterech działań arytmetycznych. Niech x i y będą liczbami maszynowymi. Mamy obliczyć $x \odot y$. Przyjmujemy, że komputer działa tak, że po wykonaniu tego działania mantysa wyniku jest normalizowana (tzn. sprowadzana do właściwego przedziału, np. $[1, 2)$) i zaokrąglana, a cecha odpowiednio korygowana. Tak więc w istocie wynikiem działania jest $\mathrm{fl}(x \odot y)$.

Przykład 2.1.2. Dla ilustracji tych czynności rozważmy komputer działający na liczbach dziesiętnych zmiennopozycyjnych z mantysą pięciocyfrową, należącą do przedziału $[0.1, 1)$ (nie jest istotne, czy wybieramy ten przedział, czy $[1, 10)$). Należy znaleźć wyniki dodawania $(+)$, odejmowania $(-)$, mnożenia $(\times)$ i dzielenia $(/)$ liczb maszynowych

$$x = 0.31426_{10}3, \quad y = 0.92577_{10}5$$

oraz ich błędy względne.

Rozwiązanie. Przyjmujemy, że „surowe" wyniki są umieszczane w akumulatorze podwójnej długości:

$$\begin{aligned} x + y &= 0.92891\,26000_{10}5, \\ x - y &= -0.92262\,74000_{10}5, \\ x \times y &= 0.29093\,24802_{10}8, \\ x/y &\approx 0.33945\,79647_{10}-2. \end{aligned}$$

Po zaokrągleniu mantys do pięciu cyfr komputer zapamiętuje te wyniki jako

$$\begin{aligned} \mathrm{fl}(x + y) &= 0.92891_{10}5, \\ \mathrm{fl}(x - y) &= -0.92263_{10}5, \\ \mathrm{fl}(x \times y) &= 0.29093_{10}8, \\ \mathrm{fl}(x/y) &= 0.33946_{10}-2. \end{aligned}$$

Błędy względne tych liczb są odpowiednio równe: $2.8_{10}-6$, $2.8_{10}-6$, $8.5_{10}-6$ i $6.0_{10}-6$; wszystkie one są mniejsze od 10^{-5}. ■

W dobrze zaprojektowanym komputerze wyniki czterech działań arytmetycznych spełniają związki

$$\mathrm{fl}(x \odot y) = (x \odot y)(1+\delta), \quad \text{przy czym} \quad |\delta| \leqslant \varepsilon,$$

gdzie ε jest precyzją arytmetyki (ewentualne drobne odchylenia od tej własności są tak nieznaczne, że można je pominąć). Tak więc precyzja arytmetyki szacuje z góry *błąd względny* wyniku dowolnego działania arytmetycznego. Jeśli x i y mogą nie być liczbami maszynowymi, to powyższa równość komplikuje się:

$$\mathrm{fl}(\mathrm{fl}(x) \odot \mathrm{fl}(y)) = (x(1+\delta_1) \odot y(1+\delta_2))(1+\delta_3), \quad |\delta_i| \leqslant \varepsilon.$$

Współczesne komputery wykonują działania w specjalnych rejestrach, które są dłuższe od słów przeznaczonych na liczby maszynowe. Wyniki działań mają więc przejściowo dużą dokładność, którą zapewniają dodatkowe *bity chroniące*. Oczywiście te wyniki są następnie zaokrąglane, co daje zwykłe liczby maszynowe. Liczba bitów chroniących i inne szczegóły mogą być dla każdego typu komputera inne i często trudno uzyskać informacje na ten temat. Można dowiedzieć się więcej o tym z książki Sternbenza [1974] i licznych artykułów w czasopismach komputerowych; zob. także Feldstein i Turner [1986], Gregory [1980], Rall [1965], Overton [2001], Scott [1985] oraz Waser i Flynn [1982].

Błąd względny wyrażeń arytmetycznych

Korzystając z już otrzymanych wyników, możemy badać wpływ błędów zaokrągleń na wartości bardziej złożonych wyrażeń arytmetycznych. Niech najpierw x, y i z będą liczbami maszynowymi. Chcemy obliczyć $x(y+z)$.

Mamy ciąg równości

$$\begin{aligned}
\mathrm{fl}[x(y+z)] &= [x\,\mathrm{fl}(y+z)](1+\delta_1) = && |\delta_1| \leqslant \varepsilon \\
&= [x(y+z)(1+\delta_2)](1+\delta_1) = && |\delta_2| \leqslant \varepsilon \\
&= x(y+z)(1+\delta_1+\delta_2+\delta_1\delta_2) \approx \\
&\approx x(y+z)(1+\delta_1+\delta_2) = \\
&= x(y+z)(1+\delta_3) && |\delta_3| \leqslant 2\varepsilon
\end{aligned}$$

Iloczyn $\delta_1\delta_2$ usunięto jako pomijalnie mały w porównaniu z ε.

Przekonamy się teraz na prostym przykładzie, jak można badać błąd względny wyniku długich obliczeń. Z grubsza mówiąc, poniższe twierdzenie orzeka, że jeśli sumujemy $n+1$ liczb dodatnich (czyli wykonujemy n dodawań), to błąd względny nie przewyższa $n\varepsilon$.

Twierdzenie 2.1.3. *Jeśli $x_0, x_1, \ldots, x_n$ są liczbami maszynowymi dodatnimi, to błąd względny sumy*

$$\sum_{i=0}^{n} x_i$$

obliczanej w zwykły sposób jest równy co najwyżej $(1+\varepsilon)^n - 1 \approx n\varepsilon$.

Dowód. Niech $S_k = x_0 + x_1 + \ldots + x_k$ i niech S_k^* będzie wartością tej sumy obliczoną przez komputer. Wzory rekurencyjne dla tych wielkości są następujące:

$$S_0 = x_0, \quad S_{k+1} = S_k + x_{k+1},$$
$$S_0^* = x_0, \quad S_{k+1}^* = \mathrm{fl}(S_k^* + x_{k+1}).$$

Wprowadzamy jeszcze oznaczenia

$$\rho_k = \frac{S_k^* - S_k}{S_k}, \quad \delta_k = \frac{S_{k+1}^* - (S_k^* + x_{k+1})}{S_k^* + x_{k+1}}.$$

Tak więc $|\rho_k|$ jest błędem względnym przybliżenia S_k^* dokładnej k-tej sumy S_k, a $|\delta_k|$ – takim błędem dla przybliżenia sumy $S_k^* + x_{k+1}$ za pomocą wielkości $\mathrm{fl}(S_k^* + x_{k+1})$. Z tych definicji wynika, że

$$\begin{aligned}\rho_{k+1} &= (S_{k+1}^* - S_{k+1})/S_{k+1} = \\ &= [(S_k^* + x_{k+1})(1+\delta_k) - (S_k + x_{k+1})]/S_{k+1} = \\ &= \{[S_k(1+\rho_k) + x_{k+1}](1+\delta_k) - (S_k + x_{k+1})\}/S_{k+1} = \\ &= \delta_k + \rho_k(S_k/S_{k+1})(1+\delta_k).\end{aligned}$$

Ponieważ $S_k < S_{k+1}$ i $|\delta_k| \leqslant \varepsilon$, więc

$$|\rho_{k+1}| \leqslant \varepsilon + |\rho_k|(1+\varepsilon) = \varepsilon + \theta|\rho_k|,$$

gdzie $\theta = 1 + \varepsilon$. Mamy zatem ciąg nierówności

$$|\rho_0| = 0, \quad |\rho_1| \leqslant \varepsilon, \quad |\rho_2| \leqslant \varepsilon + \theta\varepsilon,$$
$$|\rho_3| \leqslant \varepsilon + \theta(\varepsilon + \theta\varepsilon) = \varepsilon(1 + \theta + \theta^2), \ldots,$$

a ogólnie

$$|\rho_n| = \varepsilon(1 + \theta + \ldots + \theta^{n-1}) = \varepsilon(\theta^n - 1)/(\theta - 1) = (1+\varepsilon)^n - 1.$$

Z wzoru dwumiennego Newtona wynika, że

$$(1+\varepsilon)^n - 1 = 1 + \binom{n}{1}\varepsilon + \binom{n}{2}\varepsilon^2 + \ldots + \binom{n}{n}\varepsilon^n - 1 \approx n\varepsilon. \qquad \blacksquare$$

ZADANIA 2.1

Zadania 1–9 dotyczą przykładowej reprezentacji zmiennopozycyjnej, w której mantysy mają 23 bity po kropce, czyli jest $\varepsilon = 2^{-24}$. W pozostałych zadaniach ε jest dowolne albo podane explicite.

1. Czy poniższe liczby są liczbami maszynowymi?
(**a**) 10^{40}, (**b**) $2^{-1} + 2^{-26}$, (**c**) $\frac{1}{5}$, (**d**) $\frac{1}{3}$, (**e**) $\frac{1}{256}$.

2. Jeśli liczba $\frac{1}{10}$ jest poprawnie zaokrąglona do znormalizowanej liczby dwójkowej $(1.a_1a_2\ldots a_{23})_2 \times 2^m$, to jaki jest błąd zaokrąglenia – bezwzględny i względny?

3. Znaleźć najbliższą liczbę maszynową dla: (**a**) $\frac{4}{5}$, (**b**) $\frac{2}{7}$, i błąd względny przybliżenia za pomocą takiej liczby.

4. Niech będzie $x = (1.11\ldots111000\ldots)_2 \times 2^{16}$, gdzie część ułamkowa zaczyna się od 26 jedynek, po których następują zera. Wyznaczyć: x_-, x_+, $\mathrm{fl}(x)$, $x - x_-$, $x_+ - x$, $x_+ - x_-$ i $|x - \mathrm{fl}(x)|/\varepsilon$.

5. Znaleźć $\mathrm{fl}(x)$ i $|x - \mathrm{fl}(x)|$ dla $x = 2^{16} + 2^{-8} + 2^{-9} + 2^{-10}$.

6. Jaka jest dokładna wartość różnicy $\mathrm{fl}(x) - x$, gdy $x = \sum_{n=1}^{26} 2^{-n}$?

7. Dla $x = 2^{12} + 2^{-12}$ znaleźć liczby maszynowe x_- i x_+.

8. Dla $x = 2^3 + 2^{-19} + 2^{-22}$ znaleźć liczby maszynowe x_- i x_+. Obliczyć $\mathrm{fl}(x)$ oraz błędy bezwzględny $|x - \mathrm{fl}(x)|$ i względny $|x - \mathrm{fl}(x)|/x$. Sprawdzić, że ten ostatni nie przewyższa ε.

9. Ile liczb maszynowych znajduje się między kolejnymi potęgami liczby 2? Ile jest wszystkich liczb maszynowych?

10. Znaleźć liczbę maszynową leżącą najbliżej liczby 1/9 po jej prawej stronie w arytmetyce zmiennopozycyjnej z mantysami mającymi 43 bity po kropce.

11. Jaka jest precyzja arytmetyki fl w komputerze dziesiętnym (pamiętającym liczby w postaci $x = \pm r \times 10^n$, gdzie $\frac{1}{10} \leqslant r < 1$), gdy mantysy mają 12 cyfr?

12. Znaleźć oszacowanie z góry błędu względnego ilorazu $(a \times b)/(c \times d)$, gdzie a, b, c, d – liczby maszynowe.

13. Jakim błędem względnym zaokrąglenia może być obarczony iloczyn n liczb maszynowych? Jak zmieni się odpowiedź na to pytanie, jeśli liczby nie muszą być maszynowe (ale mają dopuszczalną wielkość)?

14. Komputer oblicza $z_1, z_2, \ldots$ dla danych $x, a_1, a_2, \ldots$ za pomocą wzorów

$$z_1 = a_1, \qquad z_n = xz_{n-1} + a_n \quad (n \geqslant 2)$$

(jest to tzw. *schemat Hornera*). Wykazać, że jeśli dane są liczbami maszynowymi, to liczby z_n można interpretować jako dokładne wyniki działań na zaburzonych danych. Wyrazić oszacowania tych zaburzeń przez precyzję arytmetyki fl.

15. Jeśli komputer nie zaokrągla poprawnie liczb, ale tylko odrzuca zbędne bity, to jaka jest precyzja arytmetyki fl?

16. Niech $x_1, x_2, \ldots, x_n$ będą dodatnimi liczbami maszynowymi, S_n ich dokładną sumą, a S_n^* sumą obliczoną przez komputer. Udowodnić, że jeśli $x_{i+1} \geqslant \varepsilon S_i$ dla każdego i, to

$$|S_n^* - S_n|/S_n \leqslant (n-1)\varepsilon.$$

17. Niech $S_n = x_1 + x_2 + \ldots + x_n$ będzie sumą liczb maszynowych i niech S_n^* będzie wartością tej sumy obliczoną przez komputer. Wtedy $S_n^* = \mathrm{fl}(S_{n-1}^* + x_n)$. Udowodnić, że

$$S_n^* \approx S_n + S_2\delta_2 + \ldots + S_n\delta_n, \quad \text{gdzie} \quad |\delta_k| \leqslant \varepsilon.$$

18. W twierdzeniu 2.1.3 występuje różnica $(1+\varepsilon)^n - 1$. Wykazać, że jeśli $n\varepsilon < 0.01$, to jest ona mniejsza od 0.01006.

19. Co można powiedzieć o błędzie względnym zaokrąglenia sumy n liczb maszynowych? (Odrzucić założenie, że te liczby są dodatnie; taki przypadek był zbadany w tw. 2.1.3).

20. Udowodnić, że równość (2.1.3) można nieznacznie poprawić:

$$\left|\frac{x - \mathrm{fl}(x)}{x}\right| \leqslant \frac{\varepsilon}{1+\varepsilon}.$$

21. Wykazać, że $\mathrm{fl}(x) = x/(1+\delta)$, gdzie $|\delta| \leqslant \varepsilon$.

22. x, y i z są liczbami maszynowymi. Sprawdzić, które z poniższych związków są prawdziwe.

(a) $\mathrm{fl}(xy) = xy(1+\delta)$
(b) $\mathrm{fl}(x+y) = (x+y)(1+\delta)$
(c) $\mathrm{fl}(xy) = xy/(1+\delta)$
(d) $|\mathrm{fl}(xy) - xy| \leqslant |xy| \times \varepsilon$
(e) $\mathrm{fl}(x+y+z) = (x+y+z)(1+\delta)$

23. Znaleźć liczby rzeczywiste x i y, dla których $\mathrm{fl}(x \odot y) \neq \mathrm{fl}(\mathrm{fl}(x) \odot \mathrm{fl}(y))$. Zrobić to dla każdego z czterech działań arytmetycznych, przyjmując, że w komputerze dziesiętnym mantysy są pięciocyfrowe.

24. Znaleźć takie liczby maszynowe x, y, z, że $\mathrm{fl}[\mathrm{fl}(xy)z] \neq \mathrm{fl}[x\,\mathrm{fl}(yz)]$ (stąd wynika, że mnożenie maszynowe nie jest łączne).

25. Wykazać, że jeśli x i y są liczbami maszynowymi takimi, że $|y| \leqslant \frac{1}{2}\varepsilon|x|$, to $\mathrm{fl}(x+y) = x$.

26. Znaleźć liczbę rzeczywistą x taką, że $\mathrm{fl}(x) = x(1+\delta)$ dla możliwie największego $|\delta|$. Czy ta wielkość może być równa ε?

27. Wykazać, że jeśli x jest liczbą maszynową, to dla dowolnego naturalnego k jest $\mathrm{fl}(x^k) = x^k(1+\delta)^{k-1}$.

28. Jakie liczby mają reprezentację skończoną w układzie dwójkowym, ale nie mają jej w układzie dziesiętnym?

29. Niech $x \in (-\infty, 0)$ będzie liczbą maszynową. Jakie wartości w standardowej arytmetyce IEEE daje komputer dla $(-\infty) + x$, $(-\infty)x$, $x/(-\infty)$ i $(-\infty)/x$?

30. Niech liczba $\varepsilon = \frac{1}{2}\beta^{1-n}$ odnosi się do komputera, który działa w układzie pozycyjnym z podstawą β, w którym liczby zmiennopozycyjne mają n-cyfrowe mantysy i w którym stosuje się poprawne zaokrąglanie. Dla jakich β to ε jest liczbą maszynową? Jeśli nią jest, to czy jest to najmniejsza liczba maszynowa spełniająca nierówność $\mathrm{fl}(1+\varepsilon) > 1$?

Zadania komputerowe 2.1

K1. Napisać program, który oblicza wartość ε precyzji arytmetyki zmiennopozycyjnej dla liczb określonego typu. Czy będzie to wartość dokładna czy przybliżona? Wskazówka: Wyznaczyć najmniejszą dodatnią liczbę maszynową ε postaci 2^{-k}, dla której $1.0 + \varepsilon \neq 1.0$.

K2. Napisać program znajdujący największą i najmniejszą liczbę maszynową.

K3. (a) Znaleźć taką liczbę maszynową x, że $1 < x < 2$ i $x(1/x) \neq 1$, czyli $\mathrm{fl}(x\,\mathrm{fl}(1/x)) \neq 1$.

(b) Za pomocą prymitywnej metody przeszukiwania znaleźć najmniejszą taką liczbę. (Edelman [1994] pokazał, jak można to przeszukiwanie znacznie skrócić stosując metody analityczne).

2.2. Błędy bezwzględne i względne. Utrata cyfr znaczących

Gdy liczbę rzeczywistą x przybliżamy inną liczbą x^*, to *błąd bezwzględny* tego przybliżenia jest z definicji równy

$$|x - x^*|,$$

a *błąd względny* jest (dla $x \neq 0$) równy

$$\left|\frac{x - x^*}{x}\right|.$$

W pomiarach niemal zawsze istotny jest ten drugi błąd. Informacja o błędzie bezwzględnym jest rzadko użyteczna, gdy nic nie wiemy o rzędzie wielkości liczby. (Błąd bezwzględny równy 1 m w określeniu odległości Jowisza od

Ziemi byłby zadziwiająco mały, ale któż by chciał, aby chirurg popełnił taki błąd, wybierając miejsce cięcia!).

Błąd względny stosowaliśmy już, badając błędy zaokrąglenia. Nierówność

$$\left|\frac{x - \mathrm{fl}(x)}{x}\right| \leqslant \varepsilon$$

dotyczy właśnie takiego błędu spowodowanego przybliżeniem liczby rzeczywistej x bliską maszynową liczbą zmiennopozycyjną.

Utrata cyfr znaczących

Wprawdzie błędy zaokrąglenia są nieuniknione i trudno nimi sterować, ale inne typy błędów są pod naszą kontrolą. Analiza numeryczna wymaga rozumienia błędów różnych rodzajów i panowania nad nimi. Tu zajmiemy się pewnym typem błędów, wynikającym często z niestarannego programowania.

Jako przykład sytuacji, w której mogą pojawiać się duże błędy względne, rozważmy odejmowanie dwóch bliskich liczb. Niech będzie na przykład

$$x = 0.37214\,78693, \quad y = 0.37202\,30572, \quad x - y = 0.00012\,48121.$$

Jeśli te obliczenia byłyby wykonane na komputerze dziesiętnym z pięciocyfrowymi mantysami, to mielibyśmy takie równości:

$$\mathrm{fl}(x) = 0.37215, \quad \mathrm{fl}(y) = 0.37202, \quad \mathrm{fl}(x) - \mathrm{fl}(y) = 0.00013.$$

Ta różnica ma więc mniej, w porównaniu z odjemną i odjemnikiem, *cyfr znaczących*[3]. Inaczej mówiąc, różnica została obliczona z bardzo dużym błędem względnym:

$$\left|\frac{x - y - [\mathrm{fl}(x) - \mathrm{fl}(y)]}{x - y}\right| = \left|\frac{0.00012\,48121 - 0.00013}{0.00012\,48121}\right| \approx 0.04.$$

Kiedy tylko komputer musi przesuwać w lewo mantysę aby otrzymać liczbę zmiennopozycyjną, na końcu mantysy pojawiają się zera. Te zera są obce, tj. nie wynikają z różnicy dokładnych wartości x i y. Wprawdzie $\mathrm{fl}(x) - \mathrm{fl}(y)$ w stosowanym tu komputerze jest reprezentowane przez 0.13000×10^{-3}, ale zera na końcu mantysy tylko ją uzupełniają do pięciu cyfr.

[3] Ten termin, użyty już w tytule podrozdz. 2.2, nie jest określony. Warto więc przypomnieć jego sens. Niech $\tilde{a}$ będzie wartością przybliżoną pewnej wielkości a; przyjmujemy, że $a \neq 0$ i $\tilde{a} \neq 0$. Jeśli te liczby są wyrażone w układzie pozycyjnym przy podstawie β (najczęściej $\beta = 2$ albo $\beta = 10$) i jeśli t jest największą liczbą całkowitą taką, że $|a - \tilde{a}| \leqslant \frac{1}{2}\beta^{-t}$, to wszystkie cyfry przybliżenia $\tilde{a}$, począwszy od pierwszej różnej od 0 aż do cyfry mnożonej w rozwinięciu tej liczby przez β^{-t}, nazywamy *cyframi znaczącymi* tego przybliżenia. Ich liczba jest związana z błędem względnym (*przyp. tłum.*).

Odejmowanie bliskich wielkości

Gdy tylko to jest możliwe, powinniśmy unikać ryzykownego odejmowania bliskich sobie liczb. Staranny programista jest na to uczulony. Aby zilustrować możliwe postępowanie w takich przypadkach, rozważmy pewien przykład.

Przykład 2.2.1. Instrukcja

$$y \leftarrow \sqrt{x^2+1}-1$$

powoduje dla małych x odejmowanie bliskich liczb i zmniejszenie liczby cyfr znaczących. Jak można temu zaradzić?

Rozwiązanie. Przekształćmy wyrażenie y:

$$y = \left(\sqrt{x^2+1}-1\right)\frac{\sqrt{x^2+1}+1}{\sqrt{x^2+1}+1} = \frac{x^2}{\sqrt{x^2+1}+1}.$$

Tu już nie ma niebezpiecznego odejmowania. Dlatego zamiast pierwotnej instrukcji należy zastosować następującą:

$$y \leftarrow x^2/(\sqrt{x^2+1}+1).$$ ■

Ciekawe jest następujące pytanie: ile dokładnie bitów znaczących tracimy, obliczając różnicę $x-y$ bliskich liczb? Ścisła odpowiedź zależy od konkretnych wartości x i y. Można jednak otrzymać ogólne obustronne oszacowanie, wyrażone przez wielkość $|1-y/x|$, która jest wygodną miarą bliskości liczb x i y. Twierdzenie dotyczy każdego komputera dwójkowego.

Twierdzenie 2.2.2. *Jeśli liczby maszynowe x i y są takie, że $x > y > 0$ i*

$$2^{-q} \leqslant 1 - \frac{y}{x} \leqslant 2^{-p}$$

(p i q są całkowite), to liczba bitów znaczących straconych przy odejmowaniu $x-y$ jest równa co najmniej p i co najwyżej q.

Dowód. Udowodnimy tu lewą część powyższej nierówności; prawą część zostawiamy jako ćwiczenie. Zgodnie z założeniem przyjmujemy, że

$$x = r \times 2^n \quad (\tfrac{1}{2} \leqslant r < 1), \qquad y = s \times 2^m \quad (\tfrac{1}{2} \leqslant s < 1)$$

(w podrozdz. 2.1 częściej przyjmowano, że mantysa liczby dodatniej należy do przedziału $[1,2)$, ale zmiana warunku nie wpływa na postać twierdzenia).

Ponieważ $x > y$, więc przed odejmowaniem komputer w razie potrzeby przesuwa mantysę liczby y tak, aby zrównać cechy tych liczb; inaczej mówiąc, wyraża y w postaci

$$y = (2^{m-n}s)2^n.$$

Mamy więc

$$x - y = (r - 2^{m-n}s)2^n.$$

Tymczasowa mantysa tej liczby jest równa

$$r - 2^{m-n}s = r\left(1 - \frac{2^m s}{2^n r}\right) = r\left(1 - \frac{y}{x}\right) < 2^{-p}.$$

Poprawna mantysa wynika z tej liczby przez jej przesunięcie co najmniej o p bitów w lewo. Na końcu mantysy pojawi się więc co najmniej p obcych zer, co oznacza utratę co najmniej p bitów znaczących. ■

Przykład 2.2.3. Rozważmy podstawienie

$$y \leftarrow x - \sin x.$$

Ponieważ dla małych x jest $\sin x \approx x$, więc wtedy obliczenie y powoduje utratę cyfr znaczących. Jak można tego uniknąć?

Rozwiązanie. Alternatywna postać różnicy $x - \sin x$ wynika z szeregu potęgowego dla $\sin x$:

$$y = x - \sin x = x - \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots\right) =$$
$$= \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} - \ldots$$

Dla x bliskich 0 można użyć sumy częściowej tego szeregu, skąd wynika na przykład takie podstawienie:

$$y \leftarrow (x^3/6)\{1 - (x^2/20)[1 - (x^2/42)(1 - x^2/72)]\}.$$

Jeśli zmienna x przebiega szerszy przedział, to zależnie od jej wartości wybieramy tę wersję instrukcji podstawienia, która jest lepsza w danym przypadku. ■

W powyższym przykładzie trzeba by jeszcze zbadać dokładniej, dla jakich x mamy wykonywać każdą ze wspomnianych instrukcji podstawienia. Na mocy tw. 2.2.2 utrata bitów znaczących, gdy używamy pierwszej instrukcji, ogranicza się do co najwyżej jednego bitu, gdy

$$1 - \frac{\sin x}{x} \geqslant \frac{1}{2}.$$

Używając kalkulatora, sprawdzamy, że tak jest dla $|x| \geqslant 1.9$. Dlatego wtedy obliczamy różnicę $x - \sin x$, a dla $|x| < 1.9$ stosujemy sumę częściową szeregu. Możemy sprawdzić, że w najgorszym razie (dla $x = 1.9$) siedem początkowych składników tego szeregu daje y z błędem mniejszym od 10^{-9} (zob. zad. 3).

W pewnych obliczeniach skutki utraty cyfr znaczących można zmniejszyć, stosując podwójną precyzję. Można to ograniczyć do szczególnie ważnych fragmentów obliczeń i dzięki temu skrócić ich czas. Trzeba bowiem wiedzieć, że działania w podwójnej precyzji są wykonywane od 2 do 4 razy wolniej, gdyż na ogół są programowane, a nie realizowane sprzętowo.

Drastyczne zmniejszenie liczby cyfr znaczących występuje też wtedy, gdy obliczamy wartości pewnych funkcji dla bardzo dużych argumentów. Przykładem jest cosinus – funkcja okresowa taka, że

$$\cos(x + 2n\pi) = \cos x \quad (n \text{ całkowite}).$$

Dzięki temu obliczanie $\cos x$ dla dowolnego x można poprzedzić zredukowaniem argumentu do przedziału $[0, 2\pi]$. Procedury dostępne w komputerach wykorzystują tę możliwość. Dodatkowo można tu zastosować inne własności cosinusa:

$$\cos(-x) = \cos x = -\cos(\pi - x).$$

Na przykład, obliczając $\cos x$ dla $x = 33278.21$ redukujemy najpierw argument według wzoru

$$y = 33278.21 - 5296 \times 2\pi = 2.46.$$

W y zachowano tylko dwie cyfry po kropce, przyjmując, że taka była dokładność dla x. Tak więc zredukowany argument ma trzy cyfry znaczące, chociaż pierwotnie było ich siedem. Dlatego także wartość cosinusa ma co najwyżej trzy cyfry znaczące. Niech nas nie zwiedzie to, że iloczyn $5296 \times 2\pi$ można obliczyć z dowolną dokładnością. Nie powinniśmy też wierzyć pozornie dużej dokładności drukowanej wartości $\cos x$. Podprogram traktuje argument jako mający pełną dokładność maszynową, ale tu tak nie jest. Jeśli więc

komputer daje wynik

$$\cos 2.46 = -0.77657\,02835,$$

to i tak pewne w nim są tylko trzy początkowe cyfry.

Arytmetyka przedziałowa

Metodę kontrolowania wielkości błędów zaokrąglenia daje *arytmetyka przedziałowa*. Zamiast na liczbach operuje się w niej na przedziałach, które zawierają dokładne wartości. W idealnym przypadku te przedziały, także dla końcowych wyników, są bardzo małe. Koszt działania na przedziałach (a nie na zwykłych liczbach maszynowych) jest jednak znaczny i dlatego arytmetykę przedziałową stosujemy tylko wtedy, gdy dokładne informacje o wynikach mają szczególne znaczenie. Może być też trudno zahamować rozszerzanie się kolejno tworzonych przedziałów, które coraz bardziej odbiegają od sensownych oszacowań wyników. W ostatnich latach opracowano pakiety programów stymulujące użycie arytmetyki przedziałowej w obliczeniach. Arytmetyka ta doczekała się wielu publikacji, a nawet własnego czasopisma. Oto książki jej poświęcone: Alefeld i Grigorieff [1980], Alefeld i Herzberger [1983], Kulisch i Miranker [1981] oraz Moore [1966, 1979]. Aktualne wyniki badań można znaleźć w Internecie.

Zadania 2.2

1. Jeśli chcemy, aby obliczenie różnicy $y = \sqrt{x^2+1} - 1$ powodowało zmniejszenie dokładności najwyżej o 2 bity, to jak trzeba ograniczyć wartości x?
2. O ile bitów zmniejsza się dokładność różnicy $x - \sin x$ dla $x = \frac{1}{2}$?
3. Wykorzystując postać reszty we wzorze Taylora, wykazać, że w przykł. 2.2.3 uzyskanie błędu poniżej 10^{-9} wymaga użycia co najmniej siedmiu składników szeregu.
4. O ile bitów zmniejsza się dokładność różnicy $1 - \cos x$ dla $x = \frac{1}{4}$?
5. (cd.). Dla funkcji z poprzedniego zadania znaleźć szereg potęgowy zapewniający pełną dokładność wyników.
6. Znaleźć tożsamość trygonometryczną, która pozwala obliczać $1 - \cos x$ dla małych x z pełną dokładnością przy użyciu procedur systemowych dla $\sin x$ lub $\cos x$. (Możliwe są dwa dobre rozwiązania).
7. Znaleźć sposób obliczania $\sqrt{x^4+4} - 2$ bez niepotrzebnej straty dokładności.
8. Opierając się na definicji $\sinh x = \frac{1}{2}(e^x - e^{-x})$, zaproponować rozsądny sposób obliczania wartości tej funkcji.

9. Równanie kwadratowe $ax^2 + bx + c = 0$ rozwiązujemy, stosując wzór

$$x = (-b \pm \sqrt{b^2 - 4ac})/(2a).$$

Jeśli $4ac$ jest małe w porównaniu z b^2, to ten wzór dla jednego z pierwiastków powoduje utratę dokładności, gdyż wtedy

$$\sqrt{b^2 - 4ac} \approx |b|.$$

Zaproponować metodę usunięcia tej trudności.

10. Zaproponować takie metody obliczeń poniższych wyrażeń, które nie powodują niepotrzebnej utraty dokładności.
(**a**) $\sqrt{x^2+1} - x$, (**b**) $\log x - \log y$, (**c**) $x^{-3}(\sin x - x)$, (**d**) $\sqrt{x+2} - \sqrt{x}$,
(**e**) $e^x - e$, (**f**) $\log x - 1$, (**g**) $(\cos x - e^{-x})/\sin x$, (**h**) $\sin x - \operatorname{tg} x$,
(**i**) $\sinh x - \operatorname{tgh} x$, (**j**) $\log(x + \sqrt{x^2+1})$ (funkcja odwrotna względem $\sinh x$).

11. Dla dowolnego $x_0 > -1$ ciąg określony wzorem rekurencyjnym

$$x_{n+1} = 2^{n+1}\left(\sqrt{1 + 2^{-n}x_n} - 1\right)$$

jest zbieżny do $\log(x_0 + 1)$ (zob. Henrici [1962, s. 243]). Przekształcić ten wzór tak, aby uniknąć straty dokładności.

12. Które z poniższych wyrażeń warto wykorzystać, obliczając $\operatorname{tg} x - \sin x$ dla x bliskich 0?
(**a**) $(\sin x)(1/\cos x - 1)$, (**b**) $\frac{1}{2}x^3$, (**c**) $(\sin x)/\cos x - \sin x$,
(**d**) $(x^2/2)(1 - x^2/12)\operatorname{tg} x$, (**e**) $\frac{1}{2}x^2 \operatorname{tg} x$, (**f**) $\operatorname{tg} x \sin^2 x/(\cos x + 1)$.

13. Znaleźć sposoby obliczania wartości poniższych funkcji bez istotnej utraty dokładności.
(**a**) $(1-x)/(1+x) - 1/(1+3x)$, (**b**) $\sqrt{1+1/x} - \sqrt{1-1/x}$, (**c**) $e^x - \cos x - \sin x$.

14. Rozważyć obliczanie e^{-x} dla $x > 0$ za pomocą szeregu

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \dots$$

Zasugerować lepszy sposób obliczeń, przyjmując, że nie dysponujemy funkcją standardową dla e^x.

15. Dla jakich wartości x obliczanie $f(x) = 1 + \cos x$ powoduje redukcję dokładności? Jak można tego uniknąć?

16. Rozważmy funkcję $f(x) = x^{-1}(1 - \cos x)$.
(**a**) Jaka definicja wartości $f(0)$ zapewnia ciągłość tej funkcji?
(**b**) W pobliżu jakich punktów użycie powyższej definicji powoduje utratę dokładności?
(**c**) Jak można usunąć wadę z części (b) zadania? Nie odwoływać się do wzoru Taylora.
(**d**) Nowe wyrażenie otrzymane w części (c) zadania powoduje być może złe w skutkach odejmowanie w jakimś innym punkcie. Usunąć i tę wadę.

17. Dla jakich θ przybliżenie $\sin\theta \approx \theta$ ma co najmniej trzy cyfry dziesiętne (zaokrąglone) dokładne?

18. Jak dobre jest przybliżenie $\cos x \approx 1$ dla małych x? Jak małe musi być x, żeby błąd tego przybliżenia nie przewyższał $0.5_{10}{-}8$?

19. Zakładając, że x jest małe, znaleźć sposób możliwie dokładnego obliczania wartości funkcji $f(x) = x + e^x - e^{2x}$.

20. Znaleźć sposób obliczania, dla małych x, dokładnych wartości funkcji

$$f(x) = (e^{\operatorname{tg} x} - e^x)/x^3.$$

21. Wyjaśnić, dlaczego w przypadku obliczania prawej strony równości przybliżonej

$$x - \sin x \approx (x^3/6)[1 - (x^2/20)(1 - x^2/42)]$$

odejmowania nie grożą utratą dokładności.

22. Załóżmy, że sumę szeregu nieskończonego $\sum_{n=1}^{\infty} x_n$ mamy znaleźć z błędem bezwzględnym mniejszym od ε. Czy mamy prawo zakończyć sumowanie składników, gdy stają się one mniejsze od ε? Odpowiedź zilustrować przykładem szeregu $\sum_{n=1}^{\infty} 0.99^n$.

23. (cd.). Rozważyć pytanie z poprzedniego zadania, zakładając, że składniki x_n są na przemian dodatnie i ujemne i że ciąg $\{|x_n|\}$ maleje do 0. (Powołać się na odpowiednie twierdzenie z analizy matematycznej).

24. Wykazać, że jeśli $x > 2^{25}\pi$ jest liczbą maszynową w komputerze z precyzją arytmetyki $\varepsilon = 2^{-24}$, to nie można obliczyć $\cos x$ nawet z jedną cyfrą dokładną.

Zadania komputerowe 2.2

K1. Napisać i wykonać program obliczający

$$f(x) = \sqrt{x^2+1} - 1, \quad g(x) = x^2/(\sqrt{x^2+1} + 1)$$

dla $x = 8^{-1}, 8^{-2}, 8^{-3}, \ldots$ Chociaż $f = g$, to komputer daje różne wyniki. Które z nich są wiarygodne, a które nie?

K2. Napisać i sprawdzić procedurę, która dla liczby maszynowej x oblicza wartość $y = x - \sin x$ z dokładnością bliską maksymalnej dla wybranej precyzji.

K3. Obliczyć i wydrukować wartości funkcji:

$$f(x) = x^8 - 8x^7 + 28x^6 - 56x^5 + 70x^4 - 56x^3 + 28x^2 - 8x + 1,$$
$$g(x) = (((((((x-8)x + 28)x - 56)x + 70)x - 56)x + 28)x - 8)x + 1,$$
$$h(x) = (x-1)^8$$

w 101 równoodległych punktach, od 0.99 do 1.01. Każdą z tych funkcji należy obliczać korzystając z jej podanej wyżej postaci, bez żadnych przekształceń. Funkcje są identyczne. Wyjaśnić, dlaczego niektóre drukowane wartości są ujemne, choć nie powinno tak być. Jeśli to możliwe, wykreślić te trzy funkcje w pobliżu punktu 1 w powiększonej skali na osi wartości (zob. Rice [1992, s. 43]).

K4. Napisać i sprawdzić procedurę obliczającą możliwie dokładne wartości różnicy $1 - \cos x$ dla $-\pi \leqslant x \leqslant \pi$. Zastosować wzór Taylora w pobliżu zera i podprogram dla cosinusa w przeciwnym razie. Przedziały, w których należy stosować jedną z metod, wybrać tak, aby tracić najwyżej jeden bit.

K5. Napisać i sprawdzić procedurę obliczającą $f(x) = x^{-2}(1 - \cos x)$. Uniknąć redukcji cyfr znaczących przy odejmowaniu i, oczywiście, zadbać o wyjątkowy punkt $x = 0$.

K6. Interesującym doświadczeniem numerycznym jest obliczenie iloczynu skalarnego wektorów x i y, równych odpowiednio:

(2.71828 1828, −3.14159 2654, 1.41421 3562, 0.57721 56649, 0.30102 99957),
(1486.2497, 8 78366.9879, −22.37492, 47 73714.647, 0.00018 5049).

Iloczyn skalarny $\sum_{i=1}^{5} x_i y_i$ obliczyć czterema różnymi sposobami:

(a) sumując składniki w naturalnym porządku;

(b) sumując składniki od ostatniego do pierwszego;

(c) sumując oddzielnie dodatnie iloczyny od największego do najmniejszego, a ujemne iloczyny od najmniejszego do największego i dodając te dwie sumy częściowe;

(d) postępując podobnie jak wyżej, ale ze zmianą porządku sumowania na przeciwny.

W każdym z przypadków (a)–(d) wykonać obliczenia w pojedynczej i podwójnej precyzji. Porównać osiem otrzymanych liczb z wartością $1.006571_{10}-9$ iloczynu skalarnego, poprawną do siedmiu cyfr znaczących. Wyjaśnić te wyniki.

K7. (cd.). Powtórzyć zad. K6, skreślając końcową cyfrę 9 w x_4 i końcową cyfrę 7 w x_5. Jak ta mała zmiana danych wpływa na wyniki?

K8. Obliczyć $f(40545, 70226)$, gdzie $f(x, y) = 9x^4 - y^4 + 2y^2$. Obliczenia wykonać na dwa sposoby:

(a) Wykonać działania na liczbach całkowitych i liczbach zmiennopozycyjnych w pojedynczej i podwójnej precyzji.

(b) Po sprawdzeniu, że $f(x, y) = (3x^2 - y^2 + 1)(3x^2 + y^2 - 1) + 1$, postąpić jak w części (a), ale używając tego wyrażenia.

2.3. Algorytmy stabilne i niestabilne. Uwarunkowanie

Zajmiemy się teraz innym tematem przewijającym się przez analizę numeryczną: podziałem algorytmów na *stabilne* i takie, które tej własności nie mają. Innym ważnym pojęciem, omówionym nieco dalej, jest *uwarunkowanie zadań numerycznych.*

Niestabilność numeryczna

Mówiąc niezbyt ściśle, algorytm numeryczny określamy jako *niestabilny*, jeśli małe błędy popełnione w jakimś etapie obliczeń rosną w następnych etapach i poważnie zniekształcają ostateczne wyniki.

Aby wyjaśnić to pojęcie, posłużymy się przykładem. Rozważmy ciąg liczb rzeczywistych określony wzorem rekurencyjnym

$$x_0 = 1, \quad x_1 = \tfrac{1}{3}, \qquad x_{n+1} = \tfrac{13}{3}x_n - \tfrac{4}{3}x_{n-1} \quad (n \geqslant 1). \tag{2.3.1}$$

Łatwo sprawdzić, że te związki generują ciąg o elementach

$$x_n = \left(\tfrac{1}{3}\right)^n. \tag{2.3.2}$$

Istotnie, to wyrażenie jest zgodne z (2.3.1) dla $n = 0$ i $n = 1$, a jeśli jest tak dla $n \leqslant m$, to i dla $n = m + 1$, gdyż

$$x_{m+1} = \tfrac{13}{3}x_m - \tfrac{4}{3}x_{m-1} = \tfrac{13}{3}\left(\tfrac{1}{3}\right)^m - \tfrac{4}{3}\left(\tfrac{1}{3}\right)^{m-1} = \left(\tfrac{1}{3}\right)^{m-1}\left(\tfrac{13}{9} - \tfrac{4}{3}\right) = \\ = \left(\tfrac{1}{3}\right)^{m+1}.$$

Jeśli ciąg $\{x_n\}$ obliczamy korzystając z definicji (2.3.1), to pewne wyniki są horrendalnie niedokładne. Oto kilkanaście z nich, obliczonych w arytmetyce zmiennopozycyjnej z 24-cyfrowymi mantysami:

$x_0 = 1.00000\,00,$
$x_1 = 0.33333\,33$ (7 cyfr znaczących),
$x_2 = 0.11111\,12$ (6 cyfr znaczących),
$x_3 = 0.03703\,73$ (5 cyfr znaczących),
$x_4 = 0.01234\,66$ (4 cyfry znaczące),
$x_5 = 0.00411\,87$ (3 cyfry znaczące),
$x_6 = 0.00138\,57$ (2 cyfry znaczące),
$x_7 = 0.00051\,31$ (1 cyfra znacząca),
$x_8 = 0.00037\,57$ (brak cyfr znaczących),
$x_9 = 0.00094\,37,$
$x_{10} = 0.00358\,87,$
$x_{11} = 0.01429\,27,$
$x_{12} = 0.05715\,02,$
$x_{13} = 0.22859\,39,$

$x_{14} = 0.9143735,$

$x_{15} = 3.6574953$ (wartość fałszywa, błąd względny 10^8).

Algorytm jest więc oczywiście niestabilny. Wynika to stąd, że gdy obliczamy x_{n+1}, to błąd elementu x_n jest mnożony przez 13/3, czyli np. błąd wartości x_1 może przenieść się na x_{15} z mnożnikiem $(13/3)^{14}$. Ponieważ błąd bezwzględny wartości x_1 jest rzędu 10^{-8}, a $(13/3)^{14} \approx 10^9$, więc tylko ta część błędu elementu x_{15}, za którą jest odpowiedzialne x_1, może wynosić około 10. Do tego dochodzą błędy popełniane przy obliczaniu $x_2, x_3, \ldots$, które też przenoszą się na x_{15} z mnożnikiem $(13/3)^k$ dla odpowiednich k.

Inny sposób wyjaśnienia, dlaczego wyniki są tak złe, wynika z faktu, że w (2.3.1) mamy równanie różnicowe o ogólnym rozwiązaniu

$$x_n = A\left(\tfrac{1}{3}\right)^n + B \cdot 4^n,$$

gdzie stałe A i B są określone przez wartości początkowe x_0 i x_1 (teorię równań różnicowych liniowych naszkicowano w podrozdz. 1.3). Wprawdzie chcemy znaleźć rozwiązanie szczególne (2.3.2), dla którego $A = 1$ i $B = 0$, ale nie możemy uniknąć „zarażenia" go przez niechciany składnik 4^n i to on może zdominować całe rozwiązanie.

To, czy algorytm jest czy nie jest numerycznie stabilny, ustalamy operując błędami względnymi. Tak więc duże bezwzględnie błędy wyników można zaakceptować, jeśli i wyniki są duże. Niech w ostatnim przykładzie wartościami początkowymi będą $x_0 = 1$ i $x_1 = 4$. Chociaż błędy przenoszą się i rosną jak przedtem, to szukanym rozwiązaniem jest teraz $x_n = 4^n$ i wyniki obliczeń mają siedem cyfr znaczących. Oto trzy z nich:

$$x_1 = 4.000006, \quad x_{10} = 1.048576_{10}6, \quad x_{20} = 1.099512_{10}12.$$

W tym przypadku poprawne wartości są na tyle duże, że „zagłuszają" błędy. Błędy bezwzględne są bez wątpienia duże (jak i przedtem), ale są pomijalne względnie.

Inny przykład niestabilności numerycznej wiąże się z obliczaniem całek

$$y_n = \int_0^1 x^n e^x \, dx \quad (n \geqslant 0). \tag{2.3.3}$$

Całkowanie przez części zastosowane do y_{n+1} daje związek rekurencyjny

$$y_{n+1} = e - (n+1)y_n. \tag{2.3.4}$$

Stąd i z oczywistej równości $y_0 = e-1$ wynika, że $y_1 = e-y_0 = e-(e-1) = 1$.

Startując od wartości $y_1 = 1$ i stosując wzór (2.3.4), obliczamy $y_2, y_3, \ldots, y_{15}$ podobnie jak w przypadku (2.3.1). Daje to m.in. wyniki

$$y_2 = 0.7182817, \quad y_{11} = 1.4224553, \quad y_{15} = 39711.43.$$

Dwa ostatnie na pewno nie są poprawne. Rzeczywiście, wobec (2.3.3) jest $y_1 > y_2 > \ldots > 0$ i $\lim_{n\to\infty} y_n = 0$ (dla $0 < x < 1$ wyrażenie x^n maleje do 0). Dlatego z (2.3.4) wynika, że $\lim_{n\to\infty}(n+1)y_n = e$.

W tym przykładzie błąd wielkości y_2 wynoszący δ pewnych jednostek jest mnożony przez 3, gdy obliczamy y_3. Ten błąd 3δ jest mnożony przez 4, gdy obliczamy y_4. Wynikający stąd błąd 12δ jest mnożony przez 5, gdy obliczamy y_5 itd. Dlatego przy obliczaniu y_{10} błąd może być rzędu $\frac{1}{2}\,10!\delta \approx 2 \cdot 10^6\delta$. Dla y_{20} analogiczna wielkość wynosi już $10^{18}\delta$. W stosowanej arytmetyce jest $\delta \approx 2^{-23}$, więc $10^{18}\delta \approx 10^{10}$. Jak widać, błędy kompletnie zniekształcają poprawne wartości y_n zbieżne do 0 tak jak liczby $1/(n+1)$.

Uwarunkowanie

Pojęcie *uwarunkowania* wiąże się, z grubsza mówiąc, z wrażliwością rozwiązania zadania na małe zmiany danych początkowych. Zadanie jest *źle uwarunkowane*, jeśli małe zmiany tych danych wywołują duże zmiany wyników. Dla pewnych typów zadań można zdefiniować wskaźnik uwarunkowania. Jeśli jest on duży, to wiemy, że zadanie jest źle uwarunkowane. Przykłady można znaleźć dalej, tu ograniczymy się do elementarnych uwag. Ilustrują one bardzo ważny fakt: wskaźnik uwarunkowania determinuje w znacznej mierze przebieg rozwiązywania danego zadania niezależnie od użytej w tym celu metody. Krótko mówiąc, jeśli jest on duży, to spodziewajmy się kłopotów!

Niech naszym zadaniem będzie po prostu obliczenie wartości funkcji f w punkcie x. Stawiamy następujące pytanie: jak małe zaburzenie wartości x wpływa na $f(x)$? Aby na to odpowiedzieć, korzystamy z twierdzenia o wartości średniej, z którego wynika, że

$$f(x+h) - f(x) = f'(\xi)h \approx hf'(x).$$

Jeśli zatem $|f'(x)|$ nie jest zbyt duże, to wpływ na $f(x)$ zaburzenia argumentu x jest bezwzględnie mały. Zwykle jednak w takim zadaniu istotny jest błąd względny. Dodając do x wielkość h, mamy zaburzenie względne argumentu równe h/x. To zaburzenie zmienia $f(x)$ na $f(x+h)$ i wielkość względna zaburzenia wartości funkcji jest równa

$$\frac{f(x+h)-f(x)}{f(x)} \approx \frac{hf'(x)}{f(x)} = \left[\frac{xf'(x)}{f(x)}\right]\left(\frac{h}{x}\right).$$

Dlatego *wskaźnikiem uwarunkowania* w dyskutowanym zadaniu jest wartość bezwzględna ilorazu $xf'(x)/f(x)$.

Przykład 2.3.1. Jaki jest wskaźnik uwarunkowania dla funkcji arc sin?

Rozwiązanie. Jeśli $f(x) = \arcsin x$, to

$$\frac{xf'(x)}{f(x)} = \frac{x}{\sqrt{1-x^2}\arcsin x}.$$

Dla x bliskich 1 jest $\arcsin x \approx \pi/2$ i wskaźnik uwarunkowania równy tam w przybliżeniu $2x/(\pi\sqrt{1-x^2})$ jest bardzo duży. Wobec tego dla $x \approx 1$ małe błędy względne wartości x prowadzą do dużych błędów względnych dla $\arcsin x$. ■

Rozważmy teraz zadanie wyznaczenia zera (pierwiastka) funkcji f; metody służące do tego są opisane w rozdz. 3. Niech f i g będą dwiema funkcjami klasy C^2 w otoczeniu pierwiastka r funkcji f.

Zakładamy, że r jest pierwiastkiem pojedynczym, czyli że $f'(r) \neq 0$. Stawiamy pytanie: jeśli zaburzamy funkcję f, zmieniając ją na $F = f + \varepsilon g$, to jaki jest pierwiastek funkcji F? Niech będzie on równy $r + h$; znajdziemy przybliżone wyrażenie dla h. Zaburzenie h jest takie, że $F(r+h) = 0$, czyli

$$f(r+h) + \varepsilon g(r+h) = 0.$$

Ponieważ f, $g \in C^2$, więc możemy użyć wzoru Taylora dla F:

$$f(r) + hf'(r) + \tfrac{1}{2}h^2 f''(\xi) + \varepsilon[g(r) + hg'(r) + \tfrac{1}{2}h^2 g''(\eta)] = 0.$$

Skreślając składniki z h^2 i uwzględniając, że $f(r) = 0$, otrzymujemy

$$h \approx -\varepsilon \frac{g(r)}{f'(r) + \varepsilon g'(r)} \approx -\varepsilon \frac{g(r)}{f'(r)}.$$

Przykład 2.3.2. Aby zilustrować powyższy wniosek, rozważmy klasyczny przykład podany przez Wilkinsona. Niech będzie

$$f(x) = \prod_{k=1}^{20}(x-k) = (x-1)(x-2)\ldots(x-20), \quad g(x) = x^{20}.$$

Pierwiastkami wielomianu f są oczywiście liczby naturalne $1, 2, \ldots, 20$. Jak wpływa na pierwiastek $r = 20$ zaburzenie funkcji f składnikiem εg (czyli zmiana współczynnika przy x^{20} z 1 na $1 + \varepsilon$)?

Rozwiązanie. Oto odpowiedź:

$$h \approx -\frac{g(20)}{f'(20)}\varepsilon = -\frac{20^{20}}{19!}\varepsilon \approx -10^9\varepsilon.$$

Tak więc zaburzenie $f(x)$ o εx^{20} może zaburzyć pierwiastek 20 nawet o $10^9\varepsilon$. Stąd wynika, że pierwiastki wielomianu f są skrajnie czułe na zakłócenia współczynników. (Zob. zad. K5). ■

Jeszcze innym typem zadania, w którym występuje wskaźnik uwarunkowania, jest rozwiązywanie układów równań liniowych $Ax = b$, szczegółowo opisane w podrozdz. 4.4. Ściślej, w tym zadaniu jest istotny *wskaźnik uwarunkowania* macierzy A, oznaczany symbolem $\varkappa(A)$ i określony jako iloczyn norm macierzy A i jej odwrotności:

$$\varkappa(A) = \|A\| \cdot \|A^{-1}\|.$$

Jeśli rozwiązanie równania $Ax = b$ jest niezbyt czułe na małe zmiany macierzy A i prawej strony b, to mówimy, że macierz A jest *dobrze uwarunkowana*. Taką sytuację mamy, gdy wskaźnik $\varkappa(A)$ jest niezbyt duży. W przeciwnym razie A jest *źle uwarunkowana* i obliczone rozwiązanie x równania $Ax = b$ trzeba traktować z dużą dozą sceptycyzmu.

Macierz Hilberta n-tego stopnia, H_n, o elementach $h_{ij} = 1/(i+j-1)$ dla $1 \leqslant i \leqslant n$ i $1 \leqslant j \leqslant n$, była przez lata przedmiotem intensywnych badań. Ich historię opisują Hestenes i Todd [1991, rozdz. IX]. Ta macierz jest często używana w testach jako przykład macierzy źle uwarunkowanej. Istotnie, $\varkappa(H_n) \approx ce^{3.5n}$ (gdy wskaźnik odpowiada normie $\|A\| = \max_{i,j}|a_{ij}|$), a więc ten wskaźnik bardzo szybko rośnie wraz z n. Ciekawym testem jest rozwiązywanie na komputerze układu $H_n x = b$ dla rosnących wartości n i dla $b_i = \sum_{j=1}^{n} h_{ij}$. Rozwiązaniem powinien być wektor $x = (1, 1, \ldots, 1)$. Dla małych n taki właśnie wektor otrzymujemy, ale w miarę wzrostu n dokładność drastycznie spada. Sprawdzono, że jeśli n jest w przybliżeniu równe liczbie cyfr dziesiętnych w mantysach, to składowe obliczonego rozwiązania nie mają zapewne ani jednej cyfry dokładnej.

Macierz Hilberta pojawia się w aproksymacji średniokwadratowej, gdy dla danej funkcji f szukamy minimum całki

$$\int_0^1 \left[\sum_{j=0}^{n} a_j x^j - f(x)\right]^2 dx.$$

Przyrównanie do 0 pochodnych cząstkowych tego wyrażenia względem a_i daje układ *równań normalnych*

$$\sum_{j=0}^{n} a_j \int_0^1 x^{i+j} dx = \int_0^1 x^i f(x)\, dx \qquad (0 \leqslant i \leqslant n).$$

Ponieważ całka po lewej stronie jest równa

$$\left.\frac{x^{i+j+1}}{i+j+1}\right|_0^1 = \frac{1}{i+j+1},$$

więc macierzą tego układu jest H_{n+1}. Funkcje x^i tworzą bardzo źle uwarunkowaną bazę przestrzeni wielomianów stopnia niewiększego od n. Dobrą bazą jest natomiast układ wielomianów ortogonalnych; zob. podrozdz. 6.8.

ZADANIA 2.3

1. Funkcja

$$f(x) = \alpha x^{12} + \beta x^{13}$$

jest taka, że $f(0.1) = 6.06_{10}{-}13$ i $f(0.9) = 0.03577$. Wyznaczyć α i β oraz zbadać czułość tych parametrów na małe zmiany wartości funkcji f w podanych punktach.

2. Rozwiązać analitycznie poniższe równanie różnicowe z danymi warunkami początkowymi:

$$x_0 = 1, \quad x_1 = 0.9, \quad x_{n+1} = -0.2x_n + 0.99x_{n-1}.$$

Nie obliczając rozwiązania $\{x_n\}$ ze wzoru rekurencyjnego, przewidzieć, czy te obliczenia będą stabilne.

3. *Całki wykładnicze* E_n określamy wzorem

$$E_n(x) = \int_1^\infty (e^{xt}t^n)^{-1}dt \qquad (n \geqslant 0,\ x > 0).$$

Spełniają one równanie

$$nE_{n+1}(x) = e^{-x} - xE_n(x).$$

Czy znając funkcję $E_1(x)$, można zastosować to równanie do obliczania $E_2(x)$, $E_3(x), \ldots$ z dobrą dokładnością? Wskazówka: Sprawdzić, czy $E_n(x)$ jest funkcją rosnącą czy malejącą wskaźnika n.

4. Wskaźnik uwarunkowania dla funkcji $f(x) = x^\alpha$ nie zależy od x. Jaki on jest?

5. Jakie są wskaźniki uwarunkowania dla poniższych funkcji? Gdzie są one duże? (**a**) $\log x$, (**b**) $\sin x$, (**c**) e^x, (**d**) $x^{-1}e^x$, (**e**) $\arccos x$.

6. Rozważyć przykład podany w tekście, gdzie $y_{n+1} = e - (n+1)y_n$. Ile cyfr dziesiętnych należy zachowywać obliczając $y_1, y_2, \ldots, y_{20}$, żeby ostatnia z tych wielkości miała pięć cyfr znaczących?

7. Związek rekurencyjny $x_n = 2x_{n-1} + x_{n-2}$ ma ogólne rozwiązanie postaci

$$x_n = A\lambda^n + B\mu^n.$$

Obliczywszy λ i μ sprawdzić, czy ten związek daje dobrą metodę obliczania x_n dla dowolnych wartości początkowych x_0 i x_1.

8. W zadaniu 1.2.K1 zdefiniowano ciąg Fibonacciego $\{x_n\}$ i podano, jak się wyrażają jego elementy dla $c = (1 - \sqrt{5})/2$. Czy w tym przypadku związek rekurencyjny określający ciąg daje stabilną metodę obliczania wielkości x_n?

Zadania komputerowe 2.3

K1. *Funkcje Bessela* Y_n spełniają taki sam związek rekurencyjny jak funkcje J_n określone na końcu podrozdz. 1.3. Wartości początkowe są tu jednak inne. Dla $x = 1$ jest

$$Y_0(1) = 0.08825\,69642,$$
$$Y_1(1) = -0.78121\,28213.$$

Stosując związek rekurencyjny, obliczyć $Y_2(1), Y_3(1), \ldots, Y_{20}(1)$. Spróbować ocenić wiarygodność wyników.

Wskazówka: Liczby $|Y_n(1)|$ powinny szybko rosnąć. Może uda się udowodnić nierówność typu $|Y_n(1)/Y_{n-1}(1)| > n$?

K2. Niech będzie

$$x_n = \int_0^1 t^n (t+5)^{-1} dt.$$

Wykazać, że $x_0 = \log 1.2$ i $x_n = n^{-1} - 5x_{n-1}$ dla $n \geqslant 1$. Stosując ten wzór rekurencyjny, obliczyć wielkości $x_0, x_1, \ldots, x_{10}$ i oszacować dokładność ostatniej z nich.

K3. (cd.). Znaleźć sposób dokładniejszego obliczenia x_{20}, stosując na przykład do funkcji podcałkowej wzór Taylora. Po wyznaczeniu x_{20} z pełną dokładnością maszynową zastosować rekurencję wstecz, tj. obliczyć kolejno $x_{19}, x_{18}, \ldots, x_0$. Czy otrzymane x_0 jest poprawne? A inne x_n? Czy związek rekurencyjny stosowany wstecz ma inne własności, a jeśli tak, to dlaczego?

K4. Funkcje Bessela $J_0(x), J_1(x), \ldots$ można obliczać, stosując wzór rekurencyjny z podrozdz. 1.3 dla malejących n. Ściślej, dla pewnego N przyjmujemy próbne wartości $J_{N+1}(x) = 0$ i $J_N(x) = 1$, a następnie stosujemy wzór

$$J_{n-1}(x) = \frac{2n}{x} J_n(x) - J_{n+1}(x) \qquad (n = N, N-1, \ldots, 1).$$

Obliczone wartości skalujemy, czyli zmieniamy J_n na λJ_n z takim λ, aby była spełniona tożsamość

$$J_0^2(x) + 2\sum_{n=1}^{\infty} J_n^2(x) = 1.$$

Oczywiście N musi być na tyle duże, żeby założenie $J_{N+1}(x) = 0$ było usprawiedliwione. Ten sposób obliczeń zaproponował J. C. P. Miller. Zastosować go dla $N = 51$ i $x = 1$.

K5. („Perfidny" wielomian, Wilkinson [1984])

(a) Korzystając z dostępnego programu obliczania zer (także zespolonych) funkcji, znaleźć wszystkie pierwiastki wielomianu

$$\begin{aligned} P(x) = {} & x^{20} - 210x^{19} + 20615x^{18} - 12\,56850x^{17} + 533\,27946x^{16} - \\ & -16722\,80820x^{15} + 4\,01717\,71630x^{14} - 75\,61111\,84500x^{13} + \\ & +1131\,02769\,95381x^{12} - 13558\,51828\,99530x^{11} + \\ & +1\,30753\,50105\,40395x^{10} - 10\,14229\,98655\,11450x^{9} + \\ & +63\,03081\,20992\,94896x^{8} - 311\,33364\,31613\,90640x^{7} + \\ & +1206\,64780\,37803\,73360x^{6} - 3599\,97951\,79476\,07200x^{5} + \\ & +8037\,81182\,26450\,51776x^{4} - 12870\,93124\,51509\,88800x^{3} + \\ & +13803\,75975\,36407\,04000x^{2} - 8752\,94803\,67616\,00000x + \\ & +2432\,90200\,81766\,40000. \end{aligned}$$

Powyższy wielomian jest identyczny z wielomianem Wilkinsona

$$p(x) = \prod_{k=1}^{20} (x - k)$$

z przykł. 2.2.3. Sprawdzić otrzymane zera z_k dla $1 \leqslant k \leqslant 20$, obliczając $|P(z_k)|$, $|p(z_k)|$ i $|z_k - k|$. Skomentować wyniki.

(b) Powtórzyć osiem razy część (a), zmieniając współczynnik przy x^{20} w P na $1 + \varepsilon$, gdzie $\varepsilon = 10^{-2k}$ i $k = 8, 7, \ldots, 1$. Czy otrzymane wyniki są zgodne z informacjami z przykł. 2.2.3?

(c) Wilkinson wykazał, że zmiana współczynnika -210 przy x^{19} w wielomianie P na $-210 - 2^{-23}$ powoduje zmianę zer 16 i 17 na parę liczb zespolonych $16.73\ldots \pm 2.812\ldots i$. Potwierdzić tę informację.

Nota historyczna. We wczesnych latach czterdziestych XX w. Jim Wilkinson jako młody matematyk pracował z Alanem Turingiem i innymi nad nowym komputerem o nazwie *Pilot ACE* w National Physical Laboratory w Wielkiej Brytanii. Testując ten komputer, napisał program obliczający pierwiastki tego wielomianu metodą Newtona. Nie przewidując trudności, zaczął obliczenia dla $x_0 = 21$ i spodziewał się natychmiastowej zbieżności do największego pierwiastka 20. Gdy te oczekiwania nie spełniły się, zbadał zadanie głębiej. Ten test numeryczny wraz z wieloma innymi doprowadził go do stworzenia zupełnie nowej dziedziny analizy numerycznej – *analizy pozornych zaburzeń*. Dodatkowe szczegóły o dziele Wilkinsona można znaleźć w pracy Foxa [1987].

ROZDZIAŁ 3

Rozwiązywanie równań nieliniowych

3.0. Wstęp

Rozdział 3 jest poświęcony lokalizacji pierwiastków równań (czyli zer funkcji). Jest to zadanie często potrzebne. Tu zajmujemy się rozwiązywaniem równań nieliniowych lub układów takich równań, tj. obliczaniem takiego rzeczywistego x, że $f(x) = 0$ albo takiego $X = (x_1, x_2, \ldots, x_n)$, że $F(X) = 0$. W tych równaniach jedna lub więcej zmiennych występuje nieliniowo. Odmienne zadanie, a mianowicie rozwiązywanie układów liniowych postaci $Ax = b$, rozważamy w rozdz. 4.

Ogólne zadanie, sformułowane w najprostszym przypadku funkcji rzeczywistej jednej zmiennej, jest następujące: dla danej funkcji $f\colon \mathbb{R} \to \mathbb{R}$ znaleźć wartości x, dla których $f(x) = 0$. Rozważymy kilka standardowych procedur służących do rozwiązania tego zadania.

Przykłady równań nieliniowych można znaleźć w wielu zastosowaniach. W teorii dyfrakcji światła są potrzebne pierwiastki równania

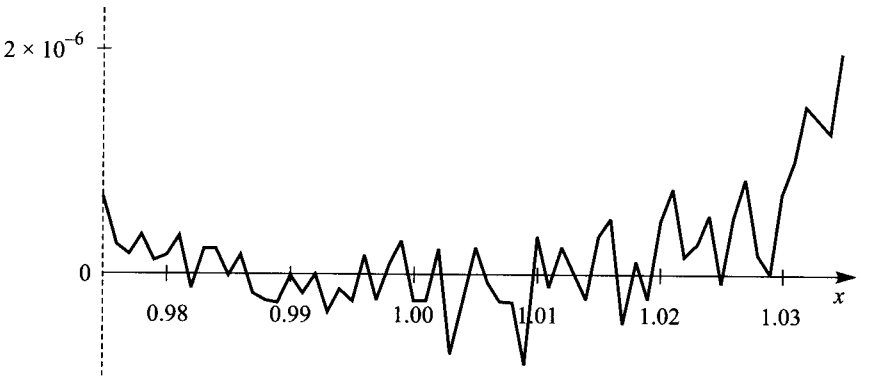
$$x - \operatorname{tg} x = 0.$$

Obliczanie orbit planet wymaga rozwiązywania równania Keplera

$$x - a \sin x = b$$

dla wielu wartości a i b.

Wyznaczaniem zer funkcji zajmowano się przez setki lat; w tym czasie znaleziono wiele metod związanych z tym zadaniem. Zaczynamy ten rozdział od trzech prostych ale użytecznych metod: metody bisekcji, metody Newtona i metody siecznych. Następnie zajmiemy się ogólną teorią metod punktu stałego i metod kontynuacji. Rozważymy też specjalne metody służące do obliczania pierwiastków wielomianów.



RYS. 3.1. Wykres wielomianu $p_4(x)$

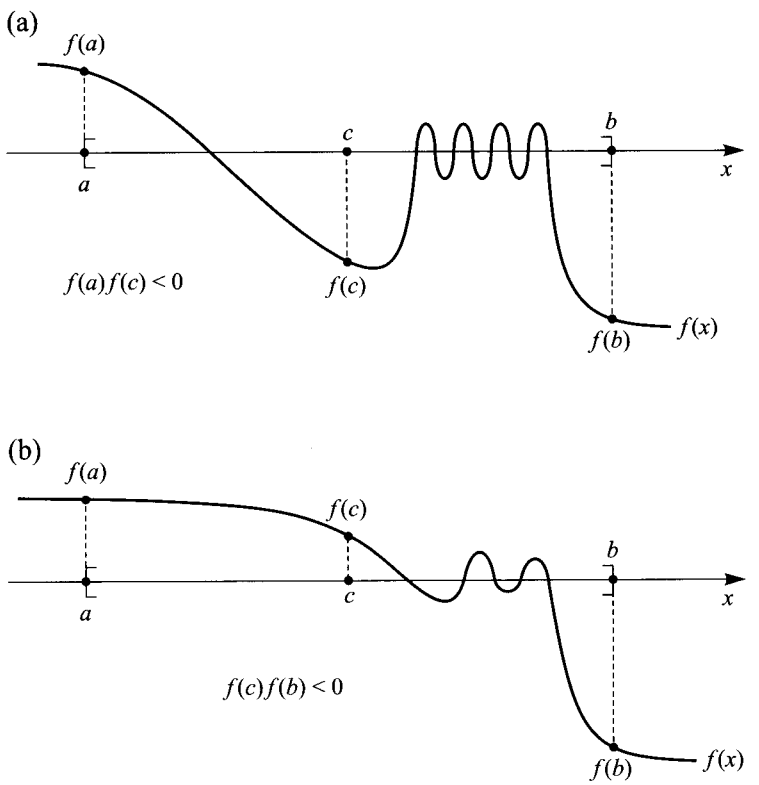
Używając komputera do znalezienia przybliżonego zera funkcji, możemy otrzymać wiele takich wartości, jeśli nawet dokładne zero jest jedyne. Doskonale to ilustruje przykład wielomianu

$$p_4(x) := x^4 - 4x^3 + 6x^2 - 4x + 1.$$

Jeśli przypomnimy sobie, że ten wielomian rozkłada się na czynniki tak, że $p_4(x) = (x-1)^4$, to stanie się oczywiste, że jedynym (poczwórnym) zerem jest liczba 1. Przypuśćmy, że znamy tylko pierwotną postać tego wielomianu i że tego zera nie zauważyliśmy. Użyjmy komputera, w którym precyzja ε arytmetyki wynosi 2^{-24} (zob. podrozdz. 2.1) i obliczmy wartości wielomianu od 0.975 co 0.001 do 1.035. Te wartości zmieniają wiele razy znak, co zdaje się wskazywać na istnienie wielu zer. Rysunek 3.1 pokazuje wykres łamanej łączącej otrzymane punkty. Zamiast gładkiej krzywej otrzymano zygzakowatą linię. Każdą wartość z przedziału $[0.981, 1.026]$ można uznać za przybliżenie prawdziwego rozwiązania. Powody tkwią w przyjętym sposobie obliczania wartości wielomianu, ograniczonej dokładności użytej arytmetyki i wynikających stąd błędach zaokrągleń. Powyższy przykład, podobny do tego, jaki podali Conte i de Boor [1980, s. 73], ilustruje jedno z niebezpieczeństw, jakie grożą nam, gdy obliczamy pierwiastki równań.

3.1. Metoda bisekcji (połowienia przedziału)

Jeśli f jest funkcją ciągłą w przedziale $[a,b]$ i jeśli $f(a)f(b) < 0$, a więc f zmienia znak w $[a,b]$, to ta funkcja musi mieć zero w (a,b). Jest to konsekwencja własności Darboux funkcji ciągłych (tw. 1.1.1).



RYS. 3.2. (a) Metoda bisekcji wybiera lewy podprzedział. (b) Metoda bisekcji wybiera prawy podprzedział

Oto jak *metoda bisekcji* (znana też jako *metoda połowienia przedziału*) korzysta z tej własności. Jeśli $f(a)f(b) < 0$, to obliczamy $c = \frac{1}{2}(a+b)$ i sprawdzamy, czy $f(a)f(c) < 0$. Jeśli tak, to f ma zero w $[a,c]$; wtedy pod b podstawiamy c. W przeciwnym razie jest $f(c)f(b) < 0$; wtedy pod a podstawiamy c. W obu przypadkach nowy przedział $[a,b]$, dwa razy krótszy od poprzedniego, zawiera zero funkcji f, więc postępowanie można powtórzyć. Rysunki 3.2(a) i 3.2(b) pokazują oba przypadki przy założeniu, że $f(a) > 0 > f(b)$. Dzięki tym rysunkom rozumiemy, dlaczego metoda bisekcji daje jedno zero funkcji, a nie wszystkie zawarte w $[a,b]$. Oczywiście, jeśli $f(a)f(c) = 0$, to $f(c) = 0$ i zero zostało znalezione. Błędy za-

okrągleń powodują jednak, że otrzymanie zerowej wartości $f(c)$ jest mało prawdopodobne. Dlatego równość $f(c) = 0$ nie powinna stanowić kryterium zakończenia obliczeń. Trzeba dopuścić pewną rozsądną tolerancję, godząc się np. na to, żeby było $|f(c)| < 10^{-5}$, gdy precyzja arytmetyki wynosi $\varepsilon = 2^{-24}$.

Przykład 3.1.1. Za pomocą metody bisekcji znaleźć pierwiastek równania $e^x = \sin x$ najbliższy 0.

Rozwiązanie. Wykreśliwszy z grubsza funkcje e^x i $\sin x$, zauważamy łatwo, że to równanie nie ma pierwiastków dodatnich i że pierwiastek ujemny najbliższy 0 znajduje się w przedziale $[-4, -3]$. Stosując metodę bisekcji na komputerze z $\varepsilon = 2^{-24}$ i zaczynając od tego przedziału, otrzymujemy wyniki podane niżej:

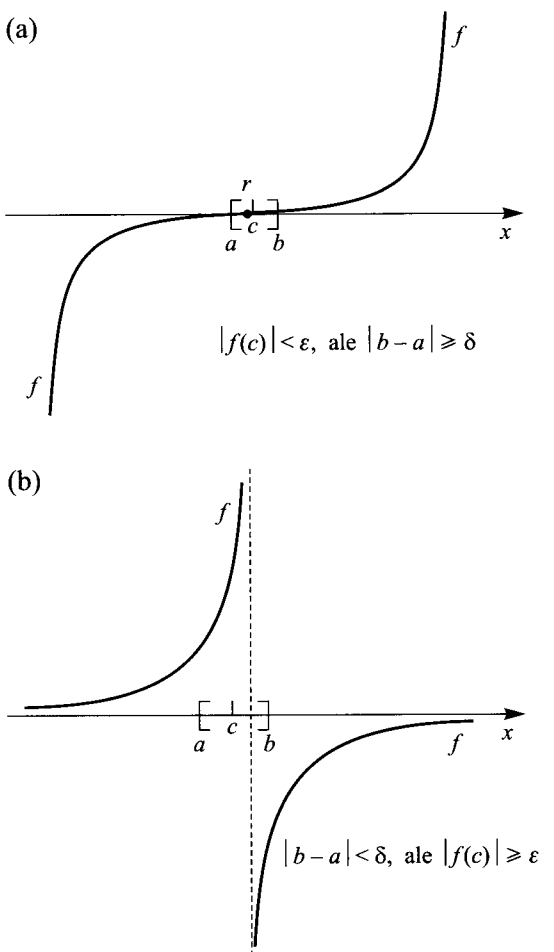
k	c	$f(c)$
1	−3.5000	-0.321
2	−3.2500	$-0.694_{10}-1$
3	−3.1250	$0.605_{10}-1$
4	−3.1875	$0.625_{10}-1$
.....		
13	−3.1829	$0.122_{10}-3$
14	−3.1830	$0.193_{10}-4$
15	−3.1831	$-0.124_{10}-4$
16	−3.1831	$0.345_{10}-5$

■

Algorytm bisekcji

Pewne fragmenty programu bisekcji podanego niżej wymagają dodatkowych wyjaśnień. Po pierwsze, punkt środkowy c obliczamy za pomocą instrukcji $c \leftarrow a + (b - a)/2$, a nie instrukcji $c \leftarrow (a + b)/2$, gdyż w obliczeniach numerycznych lepiej jest obliczać nową wielkość, dodając do poprzedniej małą poprawkę. Forsythe, Malcolm i Moler [1977, s. 162] podają przykład, w którym punkt środkowy obliczany jako $(a + b)/2$ na komputerze ze skończoną precyzją wychodzi poza przedział $[a, b]$! Po drugie, zmianę znaku wartości funkcji jest lepiej badać za pomocą nierówności $\operatorname{sgn}(w) \neq \operatorname{sgn}(u)$ zamiast $wu < 0$, gdyż w tym drugim przypadku wykonujemy zbędne mnożenie i możemy spowodować niedomiar lub nadmiar. Po trzecie, e jest oszacowaniem błędu zgodnym z następnym twierdzeniem. Zauważmy wreszcie, że program uwzględnia trzy kryteria zakończenia obliczeń. Po pierwsze, M jest maksymalną liczbą kroków dopuszczoną przez użytkownika. Takie zabezpieczenie trzeba zawsze uwzględnić, aby

usunąć ryzyko niekończących się obliczeń. Prócz tego obliczenia są przerywane, gdy błąd jest dostatecznie mały lub gdy $f(c)$ jest dostatecznie bliskie 0. Służą do tego parametry δ i ε. Można łatwo podać przykłady, w których jedno z dwóch ostatnich kryteriów jest spełnione, a drugie nie. Istotnie, rozważmy rysunki 3.3(a) i 3.3(b). Na pierwszym z nich wykres funkcji jest płaski w pobliżu jej zera, co świadczy o tym, że to zero jest wielokrotne. Metoda bisekcji z trudem wyznacza takie zero z dużą precyzją. Funkcja z rys. 3.3(b) jest oczywiście nieciągła, ale sprawdzenie, że tak jest, może być w praktyce trudne. Są to oczywiście przypadki patologiczne, ale uwzględnienie trzech kryteriów daje algorytm pewny w działaniu.



RYS. 3.3. (a) Kryterium $|b-a| < \delta$ zawodzi. (b) Kryterium $|f(c)| < \varepsilon$ zawodzi

Program bisekcji można napisać tak:

input $a, b, M, \delta, \varepsilon$
$u \leftarrow f(a)$
$v \leftarrow f(b)$
$e \leftarrow b - a$
output a, b, u, v
if $\operatorname{sgn}(u) = \operatorname{sgn}(v)$ **then stop**
for $k = 1$ **to** M **do**
    $e \leftarrow e/2$
    $c \leftarrow a + e$
    $w \leftarrow f(c)$
    output k, c, w, e
    if $|e| < \delta$ **or** $|w| < \varepsilon$ **then stop**
    if $\operatorname{sgn}(w) \neq \operatorname{sgn}(u)$ **then**
        $b \leftarrow c$
        $v \leftarrow w$
    else
        $a \leftarrow c$
        $u \leftarrow w$
    end if
end do

Analiza błędu

Badając metodę bisekcji, oznaczmy kolejne otrzymywane przedziały symbolami $[a_0, b_0]$, $[a_1, b_1]$ itd. Oczywiście

$$a_0 \leqslant a_1 \leqslant a_2 \leqslant \ldots \leqslant b_0,$$
$$b_0 \geqslant b_1 \geqslant b_2 \geqslant \ldots \geqslant a_0,$$
$$b_{n+1} - a_{n+1} = \tfrac{1}{2}(b_n - a_n) \quad (n \geqslant 0).$$

Ciąg $\{a_n\}$ jako niemalejący i ograniczony z góry jest zbieżny. Podobnie, jest zbieżny ciąg $\{b_n\}$. Z ostatniej równości wynika, że

$$b_n - a_n = 2^{-n}(b_0 - a_0).$$

Stąd

$$\lim_{n\to\infty} b_n - \lim_{n\to\infty} a_n = \lim_{n\to\infty} 2^{-n}(b_0 - a_0) = 0.$$

Niech będzie

$$r := \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n.$$

Przejście do granicy w nierówności $0 \geqslant f(a_n)f(b_n)$ daje $0 \geqslant [f(r)]^2$, skąd $f(r) = 0$.

Jeśli obliczenia przerwano po znalezieniu przedziału $[a_n, b_n]$, to pierwiastek równania na pewno w nim się znajduje. Najlepszym przybliżeniem tego pierwiastka nie jest ani a_n, ani b_n, ale środek przedziału:

$$c_n := (a_n + b_n)/2,$$

a błąd tego przybliżenia szacujemy tak:

$$|r - c_n| \leqslant \frac{1}{2}(b_n - a_n) = 2^{-(n+1)}(b_0 - a_0).$$

Następujące twierdzenie podsumowuje tę dyskusję:

Twierdzenie 3.1.2. *Jeśli przedziały* $[a_0, b_0]$, $[a_1, b_1]$,... *są tworzone metodą bisekcji, to granice* $\lim_{n\to\infty} a_n$ *i* $\lim_{n\to\infty} b_n$ *istnieją, są identyczne i równe zeru funkcji* f. *Jeśli* $r := \lim_{n\to\infty} c_n$, *gdzie* $c_n := \frac{1}{2}(a_n + b_n)$, *to*

$$|r - c_n| \leqslant 2^{-(n+1)}(b_0 - a_0).$$

Przykład 3.1.3. Niech metoda bisekcji startuje od przedziału $[50, 63]$. Ile najwyżej kroków trzeba wykonać, aby otrzymać pierwiastek z błędem względnym 10^{-12}?

Rozwiązanie. Żądanie dotyczące błędu względnego oznacza, że ma być

$$|r - c_n|/|r| \leqslant 10^{-12}.$$

Wiemy, że $r \geqslant 50$, wystarczy więc zapewnić spełnienie warunku

$$|r - c_n|/50 \leqslant 10^{-12}.$$

Na mocy tw. 3.1.2 wystarczy, żeby było

$$2^{-(n+1)} \times (13/50) \leqslant 10^{-12}.$$

Jest tak dla $n \geqslant 37$. ■

Zadania 3.1

1. Znaleźć pierwiastek dodatni równania

$$x^2 - 4x \sin x + (2 \sin x)^2 = 0$$

z dokładnością do dwóch cyfr znaczących. Użyć kalkulatora.

2. Rozważyć metodę bisekcji startującą od przedziału $[1.5, 3.5]$.

(**a**) Jaka jest długość przedziału w n-tym kroku metody?

(**b**) Jaka może być maksymalna odległość między pierwiastkiem r i środkiem tego przedziału?

3. Czy możemy obliczyć pierwiastek z błędem bezwzględnym mniejszym od 10^{-6}, używając metody bisekcji (z precyzją arytmetyki 2^{-24}), jeśli startujemy od przedziału $[128, 129]$?

4. Udowodnić nierówność

$$n \geqslant \frac{\log(b_0 - a_0) - \log \varepsilon}{\log 2} - 1$$

dla takiej liczby n kroków metody bisekcji, że $|r - c_n| \leqslant \varepsilon$.

5. Udowodnić nierówność

$$n \geqslant \frac{\log(b_0 - a_0) - \log \varepsilon - \log a_0}{\log 2} - 1$$

dla takiej liczby n kroków metody bisekcji, że pierwiastek będzie wyznaczony z błędem względnym $\leqslant \varepsilon$. Założyć, że $a_0 > 0$.

6. (cd.). Co zmienia się w poprzednim zadaniu, jeśli $a_0 < 0 < b_0$?

7. Ile kroków w metodzie bisekcji startującej od przedziału $[2, 3]$ trzeba wykonać, aby obliczyć pierwiastek z błędem bezwzględnym $< 10^{-6}$? Odpowiedzieć na to samo pytanie dla błędu względnego. Co będzie w obu przypadkach, jeśli precyzja arytmetyki wynosi 2^{-24}?

8. Niech będzie $c_n := \frac{1}{2}(a_n + b_n)$, $r := \lim_{n\to\infty} c_n$ i $e_n := r - c_n$, gdzie przedziały $[a_n, b_n]$ $(n \geqslant 0)$ są przedziałami tworzonymi metodą bisekcji stosowaną do funkcji ciągłej f.

(**a**) Wykazać, że $e_n \leqslant 2^{-n}(b_1 - a_1)$.

(**b**) Czy jest prawdą, że $|e_0| \geqslant |e_1| \geqslant \ldots$?

(**c**) Wykazać, że $|c_n - c_{n+1}| = 2^{-n-2}(b_0 - a_0)$.

(**d**) Wykazać, że $a_m \leqslant b_n$ dla dowolnych m i n.

(**e**) Wykazać, że r jest jedynym elementem zbioru $\bigcap_{n=0}^{\infty}[a_n, b_n]$.

(**f**) Wykazać, że dla każdego n jest $[a_n, b_n] \supset [a_{n+1}, b_{n+1}]$.

9. Sprawdzić, które z poniższych zdań mogą być fałszywe, gdy metodę bisekcji stosujemy do funkcji ciągłej:

(**a**) $|r - 2^{-1}(a_n + b_n)| \leqslant 2^{-n}(b_0 - a_0)$ $(n \geqslant 0)$

(**b**) $|r - 2^{-1}(a_{n+1} + b_{n+1})| \leqslant |r - 2^{-1}(a_n + b_n)|$ $(n \geqslant 0)$

(**c**) $|r - c_n| < |r - c_{n-1}|$ $(n \geqslant 1)$

10. W metodzie bisekcji przedział $[a_{n-1}, b_{n-1}]$ jest dzielony na połowy i jedna z nich staje się następnym przedziałem $[a_n, b_n]$. Przyjąć, że $d_n = 0$, jeśli jest to lewa połowa i $d_n = 1$ w przeciwnym razie. Wyrazić pierwiastek obliczony metodą bisekcji przez liczby $d_1, d_2, \ldots$ Wskazówka: Rozważyć najpierw przypadek, gdy $[a_0, b_0] = [0, 1]$ i pomyśleć o reprezentacji binarnej pierwiastka.

11. Używając oznaczeń z poprzednich zadań, znaleźć wzór wiążący a_n, b_n i c_n z d_n.

12. Podać przykład, w którym $a_0 < a_1 < a_2 < \ldots$ lub wykazać, że to jest niemożliwe.

13. Podać przykład, w którym $a_0 = a_1 < a_2 = a_3 < a_4 = a_5 < a_6 = \ldots$

14. Czy w metodzie bisekcji istnieje $\lim_{n\to\infty} |r - c_{n+1}|/|r - c_n|$?

15. Wykazać, że c obliczone metodą bisekcji jest tym punktem, w którym odcinek łączący punkty $(a, \operatorname{sgn} f(a))$ i $(b, \operatorname{sgn} f(b))$ przecina oś x.

ZADANIA KOMPUTEROWE 3.1

K1. Napisać i sprawdzić podprogram lub procedurę implementującą algorytm bisekcji. Do sprawdzenia użyć następujących funkcji i przedziałów:

(a) $x^{-1} - \operatorname{tg} x$ w $[0, \pi/2]$ (c) $2^{-x} + e^x + 2\cos x - 6$ w $[1, 3]$

(b) $x^{-1} - 2^x$ w $[0, 1]$ (d) $x - \operatorname{tg} x$ w $[1, 2]$

K2. Znaleźć pierwiastek równania

$$x^8 - 36x^7 + 546x^6 - 4536x^5 + 22449x^4 - 67284x^3 + \\ +118124x^2 - 109584x + 40320 = 0$$

w przedziale $[5.5, 6.5]$. Powtórzyć obliczenia po zmianie współczynnika 36 na 36.001.

K3. Znaleźć liczby a i b $(a < b)$ takie, że teoretycznie równoważne czynności $c \leftarrow (a + b)/2$ i $c \leftarrow a + 0.5(b - a)$ dają na użytym komputerze różne wyniki. Wybrać przykład, w którym nie wystąpi ani niedomiar, ani nadmiar.

K4. Napisać i sprawdzić rekursywną postać algorytmu bisekcji.

3.2. Metoda Newtona

Metoda Newtona jest ogólną procedurą, którą można zastosować w wielu różnych sytuacjach. Jej szczególny wariant odnoszący się do lokalizacji zer funkcji rzeczywistych jest też nazywany *metodą Newtona-Raphsona*. Na ogół metoda Newtona jest szybsza od metod bisekcji i siecznych, gdyż jej zbieżność jest kwadratowa, a nie liniowa bądź nadliniowa. Gdy tylko przybliżenia tworzone metodą Newtona są dostatecznie bliskie pierwiastka, staje się ona tak szybko zbieżna, że zaledwie kilka dodatkowych przybliżeń daje już maksymalną dokładność. Niestety, metoda nie zawsze jest zbieżna. Dlatego często używa się jej w kombinacji – już numerycznie globalnie zbieżnej – z jakąś wolniejszą metodą.

Jak w podrozdz. 3.1 mamy funkcję f, której zera należy wyznaczyć numerycznie. Niech r będzie takim zerem, a x jego przybliżeniem. Jeśli f'' istnieje, to na mocy twierdzenia Taylora

$$0 = f(r) = f(x + h) = f(x) + hf'(x) + \mathcal{O}(h^2),$$

gdzie $h = r - x$. Jeśli h jest małe (czyli jeśli x jest bliskie r), to jest rozsądne pominięcie składnika $\mathcal{O}(h^2)$ i rozwiązanie otrzymanego równania względem h. Daje to $h = -f(x)/f'(x)$. Jeśli x jest przybliżeniem r, to $x - f(x)/f'(x)$ powinno być lepszym przybliżeniem tego zera. Dlatego z definicji *metoda Newtona* zaczyna od przybliżenia x_0 zera r i polega na rekurencyjnym stosowaniu wzoru

$$x_{n+1} := x_n - \frac{f(x_n)}{f'(x_n)} \quad (n \geqslant 0). \tag{3.2.1}$$

Algorytm Newtona

Oto prosty algorytm stosujący M kroków metody Newtona dla danej wartości początkowej x:

input x, M
$y \leftarrow f(x)$
output $0, x, y$
for $k = 1$ **to** M **do**
    $x \leftarrow x - y/f'(x)$
    $y \leftarrow f(x)$
    output k, x, y
end do

Bardziej szczegółowy algorytm uwzględnia kryteria zakończenia obliczeń:

input $x_0, M, \delta, \varepsilon$
$v \leftarrow f(x_0)$
output $0, x_0, v$
if $|v| < \varepsilon$ **then stop**
for $k = 1$ **to** M **do**
    $x_1 \leftarrow x_0 - v/f'(x_0)$
    $v \leftarrow f(x_1)$
    output k, x_1, v
    if $|x_1 - x_0| < \delta$ **or** $|v| < \varepsilon$ **then stop**
    $x_0 \leftarrow x_1$
end do

Program komputerowy oparty na którymś z tych algorytmów musi zawierać procedury służące do obliczania $f(x)$ i $f'(x)$.

Przykład 3.2.1. Metodą Newtona, używając arytmetyki podwójnej precyzji, znaleźć zero ujemne funkcji $f(x) = e^x - 1.5 - \operatorname{arctg} x$.

Rozwiązanie. Podany wyżej algorytm zastosowano na komputerze, w którym liczby w pojedynczej precyzji mają mantysy 48-bitowe; podwójna precyzja to mantysy 96-bitowe, co odpowiada ok. 28 cyfrom dziesiętnym.

Zaprogramowano obliczanie funkcji f i $f'(x) = e^x - (1+x^2)^{-1}$. Jako punkt początkowy wybrano $x_0 = -7$. Wyniki programu są następujące:

k	x	$f(x)$
0	−7.00000 00000 00000 00000 00000 0	$-0.702_{10}-1$
1	−10.67709 61766 40013 99296 98438 6	$-0.226_{10}-1$
2	−13.27916 73756 32712 90859 78631 9	$-0.437_{10}-2$
3	−14.05365 58542 69238 73474 83175 3	$-0.239_{10}-3$
4	−14.10110 99568 66413 47616 31270 6	$-0.800_{10}-6$
5	−14.10126 97709 39415 94621 57950 6	$-0.901_{10}-11$
6	−14.10126 97727 39968 42508 30031 4	$-0.114_{10}-20$
7	−14.10126 97727 39968 42531 15512 2	0.000
8	−14.10126 97727 39968 42531 15512 2	0.000

Te wyniki pokazują szybką zbieżność ciągu przybliżeń: w każdym kroku liczba poprawnych cyfr wyniku z grubsza się podwaja. Dalej wyjaśni się, dlaczego tak jest. ■

Interpretacja graficzna

Zanim zbadamy teoretyczne podstawy metody Newtona, zajmijmy się jej graficzną interpretacją. Wiemy już, że ta metoda opiera się na *linearyzacji funkcji*, tj. zastąpieniu f funkcją liniową. Jest nią suma dwóch początkowych składników we wzorze Taylora dla f. Jeśli zatem

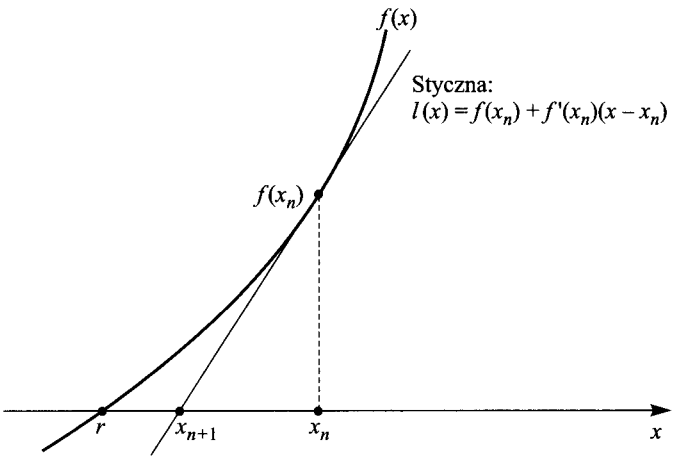
$$f(x) = f(c) + f'(c)(x-c) + \frac{1}{2!}f''(c)(x-c)^2 + \ldots,$$

to linearyzacja (w punkcie c) daje funkcję liniową

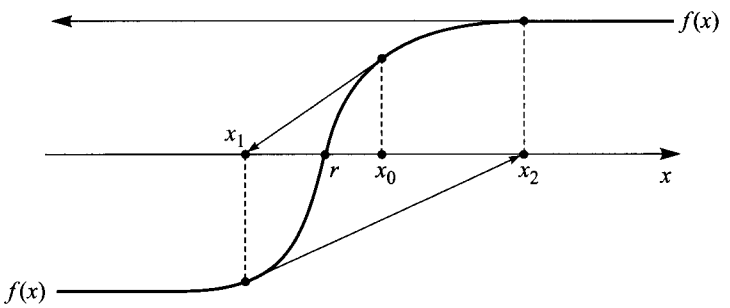
$$l(x) = f(c) + f'(c)(x-c).$$

Przybliża ona dobrze f w pobliżu c. Istotnie, $l(c) = f(c)$ i $l'(c) = f'(c)$, czyli w c funkcja liniowa ma tę samą wartość i to samo nachylenie jak f. Tak więc w metodzie Newtona konstruujemy styczną do wykresu funkcji f w punkcie bliskim r i znajdujemy punkt, w którym ta styczna przecina oś x; zob. rys. 3.4.

Pamiętając o tej graficznej interpretacji, możemy łatwo wyobrazić sobie taką funkcję i taki punkt początkowy, dla których metoda Newtona zawodzi. Taki przypadek pokazano na rys. 3.5. Kształt wykresu funkcji jest tu taki, że dla pewnych punktów początkowych ciąg $\{x_n\}$ jest rozbieżny. Wobec tego każde twierdzenie o zbieżności metody Newtona musi zakładać, że x_0 jest dostatecznie bliskie zera funkcji albo że jej wykres ma pewien określony kształt.



RYS. 3.4. Interpretacja geometryczna metody Newtona



RYS. 3.5. Przykład rozbieżności metody Newtona

Analiza błędu

Zbadajmy teraz błędy metody Newtona. Przez *błąd* rozumiemy tu wielkość

$$e_n = x_n - r.$$

Załóżmy, że funkcja f'' jest ciągła i że r jest zerem *pojedynczym* funkcji f, tj. $f(r) = 0 \neq f'(r)$. Z definicji iteracji w metodzie Newtona wynika, że

$$\begin{aligned} e_{n+1} &= x_{n+1} - r = x_n - \frac{f(x_n)}{f'(x_n)} - r = \\ &= e_n - \frac{f(x_n)}{f'(x_n)} = \frac{e_n f'(x_n) - f(x_n)}{f'(x_n)}. \end{aligned} \tag{3.2.2}$$

Na mocy wzoru Taylora

$$0 = f(r) = f(x_n - e_n) = f(x_n) - e_n f'(x_n) + \frac{1}{2} e_n^2 f''(\xi_n),$$

gdzie ξ_n jest liczbą zawartą między x_n i r. Stąd

$$e_n f'(x_n) - f(x_n) = \frac{1}{2} f''(\xi_n) e_n^2.$$

Podstawiając to do (3.2.2), otrzymujemy

$$e_{n+1} = \frac{1}{2} \frac{f''(\xi_n)}{f'(x_n)} e_n^2 \approx \frac{1}{2} \frac{f''(r)}{f'(r)} e_n^2 = C e_n^2. \tag{3.2.3}$$

Przypuśćmy, że określone tu C jest równe w przybliżeniu 1 i że $e_n \approx 10^{-4}$. Wtedy z (3.2.3) wynika, że $e_{n+1} \approx 10^{-8}$ i $e_{n+2} \approx 10^{-16}$. Jest zaskakujące, że zaledwie parę dodatkowych iteracji wystarczy, aby otrzymać więcej niż dokładność maszynową!

Równość (3.2.3) mówi nam, że e_{n+1} jest z grubsza równe pewnej stałej pomnożonej przez e_n^2. Taka korzystna sytuacja jest nazywana *zbieżnością kwadratową*. Dzięki niej właśnie w wielu zastosowaniach każda iteracja metody Newtona podwaja liczbę cyfr dokładnych przybliżenia.

Możemy już zbadać zbieżność metody. Wobec (3.2.3) pomysł dowodu jest prosty: jeśli e_n jest małe, a czynnik $\frac{1}{2} f''(\xi_n)/f'(x_n)$ nie jest zbyt duży, to e_{n+1} jest mniejsze co do modułu od e_n. Określmy wielkość $c(\delta)$ zależną od δ wzorem

$$c(\delta) = \frac{\max\limits_{|x-r| \leqslant \delta} |f''(x)|}{2 \min\limits_{|x-r| \leqslant \delta} |f'(x)|} \qquad (\delta > 0).$$

Wybieramy δ na tyle małe, aby mianownik po prawej stronie był dodatni, a dodatkowo, jeśli to konieczne, zmniejszamy δ tak, żeby było $\delta c(\delta) < 1$. To jest możliwe, gdyż jeśli δ dąży do 0, to $c(\delta)$ dąży do $\frac{1}{2}|f''(r)/f'(r)|$, czyli $\delta c(\delta)$ dąży do 0. Ustaliwszy δ, przyjmujemy $\rho := \delta c(\delta)$. Przypuśćmy, że zaczynamy iteracje Newtona od punktu x_0 takiego, że $|x_0 - r| \leqslant \delta$. Stąd $|e_0| \leqslant \delta$ i $|\xi_0 - r| \leqslant \delta$. Dlatego, wobec definicji wielkości $c(\delta)$, mamy

$$\tfrac{1}{2}|f''(\xi_0)/f'(x_0)| \leqslant c(\delta),$$

a równość (3.2.3) daje

$$|x_1 - r| = |e_1| \leqslant e_0^2 c(\delta) = |e_0||e_0|c(\delta) \leqslant |e_0|\delta c(\delta) = |e_0|\rho < |e_0| \leqslant \delta.$$

To pokazuje, że następny punkt, czyli x_1, także leży nie dalej od r niż o δ.

Korzystając wielokrotnie z udowodnionej już nierówności, wnioskujemy, że

$$|e_1| \leqslant \rho|e_0|,$$
$$|e_2| \leqslant \rho|e_1| \leqslant \rho^2|e_0|,$$
$$|e_3| \leqslant \rho|e_2| \leqslant \rho^3|e_0|, \ldots$$

Ogólniej,

$$|e_n| \leqslant \rho^n|e_0|.$$

Ponieważ $0 \leqslant \rho < 1$, więc $\lim_{n\to\infty} \rho^n = 0$, czyli $\lim_{n\to\infty} e_n = 0$. Ostatecznie daje to następujące twierdzenie:

Twierdzenie 3.2.2. *Niech r będzie zerem pojedynczym funkcji f i niech jej druga pochodna f'' będzie ciągła. Wtedy istnieje takie otoczenie punktu r i taka stała C, że jeśli metoda Newtona startuje z tego otoczenia, to kolejne punkty są coraz bliższe r i takie, że*

$$|x_{n+1} - r| \leqslant |C(x_n - r)^2| \qquad (n \geqslant 0).$$

W pewnych przypadkach metoda Newtona jest zbieżna dla dowolnego punktu początkowego:

Twierdzenie 3.2.3. *Jeśli f należy do $C^2(\mathbb{R})$, jest rosnąca, wypukła i ma zero, to jest ono jedyne, a metoda Newtona daje ciąg do niego zbieżny dla dowolnego punktu początkowego.*

Dowód. Przypomnijmy, że funkcja f jest *wypukła*, jeśli $f''(x) > 0$ dla każdego x. Ponieważ f jest rosnąca, więc $f' > 0$ w $\mathbb{R}$. Wobec (3.2.3) jest $e_{n+1} > 0$, czyli $x_n > r$ dla $n \geqslant 1$. Z tej samej własności funkcji f wynika, że $f(x_n) > f(r) = 0$. Dlatego, na mocy (3.2.2), $e_{n+1} < e_n$. Tak więc ciągi $\{e_n\}$ i $\{x_n\}$ są malejące i ograniczone z dołu (odpowiednio przez 0 i r). Dzięki temu granice $e^* = \lim_{n\to\infty} e_n$ i $x^* = \lim_{n\to\infty} x_n$ istnieją. Z (3.2.2) wynika, że $e^* = e^* - f(x^*)/f'(x^*)$, czyli $f(x^*) = 0$ i $x^* = r$. ■

Przykład 3.2.4. Znaleźć efektywny sposób obliczania pierwiastków kwadratowych.

Rozwiązanie. Niech będzie $R > 0$ i $x = \sqrt{R}$. Wobec tego x jest pierwiastkiem równania $x^2 - R = 0$. Gdy stosujemy metodę Newtona (3.2.1) do funkcji $f(x) = x^2 - R$, wzór iteracyjny można wyrazić w postaci

$$x_{n+1} := \frac{1}{2}\left(x_n + \frac{R}{x_n}\right).$$

Ten wzór, poprzedzony redukcją zakresu liczby R, jest stosowany w podprogramach pierwiastkowania. (Wzór jest bardzo stary; przypisuje się go Heronowi, greckiemu inżynierowi i architektowi, który żył między 100. rokiem p.n.e. i 100. rokiem n.e.). Jeśli np. chcemy obliczyć $\sqrt{17}$ i zaczynamy od punktu $x_0 = 4$, to kolejne przybliżenia są następujące:

$$x_1 = 4.12,$$
$$x_2 = 4.123106,$$
$$x_3 = 4.12310\,56256\,177,$$
$$x_4 = 4.12310\,56256\,17660\,54982\,14098\,56$$

(podano tu tylko poprawne początkowe cyfry). Wartość x_4 ma 28 cyfr dokładnych. Jak widać, liczba cyfr znaczących podwaja się w każdej iteracji. ■

Funkcje uwikłane

Ciekawym zastosowaniem metody Newtona jest obliczanie wartości funkcji uwikłanych, czyli określonych w sposób niejawny. Z twierdzenia 1.2.2 dotyczącego takich funkcji wiadomo, że przy dość słabych założeniach równanie $G(x, y) = 0$ określa y jako funkcję zmiennej x. Dla ustalonego x można je rozwiązać względem y, stosując metodę Newtona. Dla sensownego punktu początkowego y_0 określamy $y_1, y_2, \ldots$ wzorem

$$y_{k+1} := y_k - G(x, y_k) \Big/ \frac{\partial G}{\partial y}(x, y_k) .$$

Tej metody można użyć do zbudowania tablicy funkcji $y(x)$. Jeśli tablica zawiera pozycję (x_n, y_n) i chcemy znaleźć bliską pozycję (x_{n+1}, y_{n+1}), to w metodzie Newtona zaczynamy od punktu (x_{n+1}, y_n). Ponieważ $G(x_n, y_n) = 0$, a x_{n+1} jest bliskie x_n, więc możemy się spodziewać, że $G(x_{n+1}, y_n)$ jest małe i że parę kroków metody Newtona wystarczy, aby poprawić y_n i osiągnąć dokładną równość $G(x_{n+1}, y_{n+1}) = 0$.

Przykład 3.2.5. Zbudować tablicę wartości y, które dla danych x spełniają równanie $G(x, y) = 0$, gdzie $G(x, y) = 3x^7 + 2y^5 - x^3 + y^3 - 3$. Począwszy od $x = 0$, zwiększać x o 0.1 aż do $x = 10$.

Rozwiązanie. Niech będzie $x = 0$ i $y = 1$ (wtedy $G(x, y) = 0$). Przypuszczamy, że cztery kroki metody Newtona wystarczą, aby uzyskać pełną dokładność komputerową. W algorytmie M jest liczbą kroków dla zmiennej x, a N jest liczbą iteracji w metodzie Newtona. Program oparty na algo-

rytmie będzie wymagał dwóch podprogramów lub procedur – do obliczania $G(x,y)$ i $\partial G/\partial y$ (w przykładzie ta pochodna nie zależy od x). Tu mamy

$$G(x,y) = 3x^7 + 2y^5 - x^3 + y^3 - 3,$$
$$\frac{\partial G}{\partial y}(x,y) = 10y^4 + 3y^2.$$

Algorytm jest następujący:

$x \leftarrow 0$; $y \leftarrow 1$; $h \leftarrow 0.1$; $M \leftarrow 100$; $N \leftarrow 4$
output $0, x, y, G(x,y)$
for $i = 1$ **to** M **do**
  $x \leftarrow x + h$
  for $j = 1$ **to** N **do**
    $y \leftarrow y - G(x,y) \Big/ \frac{\partial G}{\partial y}(x,y)$
  end do
  output $i, x, y, G(x,y)$
end do

Niżej podano niektóre wartości funkcji uwikłanej obliczone za pomocą tego algorytmu:

i	x	y	$G(x,y)$
0	0.0	1.000000	0.00
1	0.1	1.000077	0.00
2	0.2	1.000612	$0.89_{10}-15$
.....			
20	2.0	−2.810639	$-0.82_{10}-10$
.....			
80	8.0	−19.92365	$0.56_{10}-9$
.....			
99	9.9	−26.85618	$0.12_{10}-7$
100	10.0	−27.23685	$-0.15_{10}-8$

■

Układy równań nieliniowych

Metoda Newtona dla układów równań nieliniowych opiera się na pomyśle zastosowanym już do pojedynczych równań. Tak więc równania linearyzujemy, a następnie rozwiązujemy; w razie potrzeby robimy to wielokrotnie. Dla ilustracji rozważmy dwa równania z dwiema niewiadomymi:

$$f_1(x_1, x_2) = 0,$$
$$f_2(x_1, x_2) = 0.$$

Zakładając, że (x_1, x_2) jest przybliżonym rozwiązaniem tego układu, obliczmy poprawki h_1 i h_2 takie, żeby $(x_1 + h_1, x_2 + h_2)$ było lepszym przybliżeniem. Używając tylko liniowych członów rozwinięcia Taylora funkcji dwóch zmiennych (tw. 1.1.10), otrzymujemy równości

$$\begin{aligned} 0 = f_1(x_1 + h_1, x_2 + h_2) &\approx f_1(x_1, x_2) + h_1 \frac{\partial f_1}{\partial x_1} + h_2 \frac{\partial f_1}{\partial x_2}, \\ 0 = f_2(x_1 + h_1, x_2 + h_2) &\approx f_2(x_1, x_2) + h_1 \frac{\partial f_2}{\partial x_1} + h_2 \frac{\partial f_2}{\partial x_2}. \end{aligned} \tag{3.2.4}$$

Wszystkie pochodne cząstkowe tu występujące są obliczane w (x_1, x_2). Układ (3.2.4) składa się z dwóch równań liniowych względem h_1 i h_2. Jego macierzą jest *jakobian*

$$J = \begin{bmatrix} \partial f_1/\partial x_1 & \partial f_1/\partial x_2 \\ \partial f_2/\partial x_1 & \partial f_2/\partial x_2 \end{bmatrix}.$$

Rozwiązanie układu (3.2.4) istnieje, jeśli macierz J jest nieosobliwa. Wtedy jest ono równe

$$\begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = -J^{-1} \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{bmatrix}.$$

Tak więc metoda Newtona dla układu dwóch równań nieliniowych wyraża się wzorem

$$\begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{bmatrix} := \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \end{bmatrix} + \begin{bmatrix} h_1^{(k)} \\ h_2^{(k)} \end{bmatrix},$$

gdzie układ liniowy

$$J \begin{bmatrix} h_1^{(k)} \\ h_2^{(k)} \end{bmatrix} = - \begin{bmatrix} f_1\left(x_1^{(k)}, x_2^{(k)}\right) \\ f_2\left(x_1^{(k)}, x_2^{(k)}\right) \end{bmatrix},$$

którego macierzą jest jakobian, rozwiązujemy metodą eliminacji Gaussa opisaną w rozdz. 4. Może to być kłopotliwe, gdy macierz J jest prawie osobliwa.

Większe układy równań rozwiązuje się w taki sam sposób. Układ równań

$$f_i(x_1, x_2, \ldots, x_n) = 0 \qquad (1 \leqslant i \leqslant n)$$

wyrażamy prościej jako

$$F(X) = 0,$$

gdzie $X = (x_1, x_2, \ldots, x_n)$ i $F = (f_1, f_2, \ldots, f_n)$. Odpowiednikiem równania (3.2.4) jest

$$0 = F(X + H) \approx F(X) + F'(X)H, \tag{3.2.5}$$

gdzie $H = (h_1, h_2, \ldots, h_n)$, a $F'(X)$ jest jakobianem $J(X)$, tj. następującą macierzą kwadratową stopnia n:

$$F'(X) := \begin{bmatrix} \partial f_1/\partial x_1 & \partial f_1/\partial x_2 & \ldots & \partial f_1/\partial x_n \\ \partial f_2/\partial x_1 & \partial f_2/\partial x_2 & \ldots & \partial f_2/\partial x_n \\ \ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots \\ \partial f_n/\partial x_1 & \partial f_n/\partial x_2 & \ldots & \partial f_n/\partial x_n \end{bmatrix}.$$

Wektor poprawek H wyraża się wzorem wynikającym z (3.2.5):

$$H = -F'(X)^{-1}F(X).$$

W praktyce jednak wyznaczamy H, stosując do (3.2.5) metodę eliminacji Gaussa, bo obliczanie odwrotności macierzy jest bardziej kosztowne. Ostatecznie, metodę Newtona dla układu n równań nieliniowych z n niewiadomymi opisuje wzór

$$X^{(k+1)} := X^{(k)} + H^{(k)}, \tag{3.2.6}$$

gdzie $H^{(k)}$ jest rozwiązaniem układu

$$F'(X^{(k)})H^{(k)} = -F(X^{(k)}). \tag{3.2.7}$$

Przykład 3.2.6. Zaczynając od punktu $(1, 1, 1)$, wykonać sześć iteracji metodą Newtona, aby znaleźć pierwiastek układu nieliniowego

$$\begin{aligned} xy &= z^2 + 1, \\ xyz + y^2 &= x^2 + 2, \\ e^x + z &= e^y + 3. \end{aligned}$$

Rozwiązanie. Niech będzie

$$F(X) = \begin{bmatrix} f_1(x_1, x_2, x_3) \\ f_2(x_1, x_2, x_3) \\ f_3(x_1, x_2, x_3) \end{bmatrix} = \begin{bmatrix} x_1x_2 - x_3^2 - 1 \\ x_1x_2x_3 - x_1^2 + x_2^2 - 2 \\ e^{x_1} - e^{x_2} + x_3 - 3 \end{bmatrix}.$$

Obliczenie pochodnych cząstkowych daje jakobian

$$F'(X) = \begin{bmatrix} x_2 & x_1 & -2x_3 \\ x_2x_3 - 2x_1 & x_1x_3 + 2x_2 & x_1x_2 \\ e^{x_1} & -e^{x_2} & 1 \end{bmatrix}.$$

Dla punktu początkowego $X^{(0)} = (1, 1, 1)$ stosujemy metodę Newtona opisaną wzorami (3.2.6) i (3.2.7). Daje to następujące wyniki:

n	x_1	x_2	x_3
0	1.0000000	1.0000000	1.0000000
1	2.1893260	1.5984751	1.3939006
2	1.8505896	1.4442514	1.2782240
3	1.7801611	1.4244359	1.2392924
4	1.7776747	1.4239609	1.2374738
5	1.7776719	1.4239605	1.2374711
6	1.7776719	1.4239605	1.2374711

■

Rozwiązywanie układów równań nieliniowych jest często zadaniem niebanalnym. Standardowym podręcznikiem tej dziedziny jest książka Ortegi i Rheinboldta [1970]; zob. też Rheinboldt [1974], Ostrowski [1966], Byrne i Hall [1973], Schnabel i Frank [1984], Eaves, Gould, Peitgen i Todd [1983] oraz Allgower, Glasshoff i Peitgen [1981]. Zbieżność metody Newtona dla układów równań badają Goldstein [1967] oraz Ortega i Rheinboldt [1970].

ZADANIA 3.2

1. Używając kalkulatora, wykonać cztery iteracje metody Newtona dla wielomianu $p(x) := 4x^3 - 2x^2 + 3$ i punktu początkowego $x_0 = -1$.
2. Metodę Newtona stosujemy do funkcji $f(x) = x^2 - q$, gdzie $q > 0$. Wykazać, że jeśli przybliżenie x_n ma k cyfr dokładnych po kropce dziesiętnej, to x_{n+1} ma co najmniej $2k - 1$ takich cyfr, jeśli tylko $q > 0.006$ i $k \geqslant 1$.
3. Udowodnić, że jeśli metodę Newtona stosujemy do funkcji f o drugiej pochodnej ciągłej i jeśli $f(r) = 0 \neq f'(r)$, to granica $\lim_{n\to\infty} e_{n+1}e_n^{-2}$ istnieje i jest równa $f''(r)/[2f'(r)]$. Jak można z tego skorzystać w programie do sprawdzania, czy zbieżność jest kwadratowa?
4. Zinterpretować wzór iteracyjny $x_{n+1} := 2x_n - x_n^2 y$ jako opis metody Newtona dla pewnej funkcji.
5. Aby obliczać odwrotności liczb bez dzielenia, możemy rozwiązać równanie $x = 1/R$ znajdując zero funkcji $f(x) = x^{-1} - R$. Napisać krótki algorytm obliczania $1/R$ metodą Newtona zastosowaną do f. Nie używać w nim dzielenia ani potęgowania. Jakie są odpowiednie punkty początkowe, gdy R jest dodatnie?
6. Zaprojektować wzór iteracyjny Newtona służący do obliczania $\sqrt[3]{R}$ dla $R > 0$. Zbadać graficznie wybraną funkcję f, aby sprawdzić, dla jakich punktów początkowych metoda będzie zbieżna.
7. Zaprojektować algorytm Newtona obliczania pierwiastka piątego stopnia z dowolnej liczby dodatniej.

8. Jeśli metodę Newtona stosujemy do $f(x) = x^2 - 1$ i $x_0 = 10^{10}$, to ile kroków trzeba wykonać, aby otrzymać pierwiastek z dokładnością do 10^{-8}? (Zadanie rozwiązać analitycznie, a nie numerycznie).

9. Niech r będzie podwójnym zerem funkcji f, czyli niech $f(r) = f'(r) = 0 \neq f''(r)$. Wykazać, że jeśli funkcja f'' jest ciągła, to w metodzie Newtona mamy $e_{n+1} \approx \frac{1}{2} e_n$ (zbieżność jest liniowa).

10. Znaleźć najmniejszą liczbę dodatnią, która użyta jako punkt początkowy w metodzie Newtona dla funkcji $f(x) := \operatorname{arctg} x$ daje ciąg rozbieżny przybliżeń.

11. Udowodnić, że dla każdego (rzeczywistego) punktu początkowego metoda Newtona jest rozbieżna, jeśli:
(**a**) $f(x) := x^2 + 1$ (**b**) $f(x) := 7x^4 + 3x^2 + \pi$

12. Funkcja $f(x) = x^2 + 1$ ma zera na płaszczyźnie zespolonej w punktach $x = \pm i$. Czy istnieje punkt początkowy *rzeczywisty*, dla którego metoda Newtona byłaby zbieżna do któregokolwiek z tych zer? Jakie punkty początkowe zespolone byłyby dobre?

13. Znaleźć warunki dotyczące α, które gwarantują, że wzór iteracyjny

$$x_{n+1} = x_n - \alpha f(x_n)$$

daje ciąg zbieżny liniowo do zera funkcji f, jeśli punkt początkowy leży blisko tego zera.

14. Wykazać, że jeśli r jest k-krotnym zerem funkcji f, to zbieżność kwadratową metody Newtona można utrzymać, wprowadzając następującą modyfikację:

$$x_{n+1} = x_n - kf(x_n)/f'(x_n).$$

15. (cd.). Jak, stosując metodę Newtona, można wykryć zero wielokrotne dzięki badaniu punktów $(x_n, f(x_n))$?

16. Rozważyć wariant metody Newtona, wymagający obliczenia wartości pochodnej tylko w jednym punkcie:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}.$$

Znaleźć C i s takie, że $e_{n+1} = Ce_n^s$.

17. (*Metoda Steffensena*). Rozważmy wzór iteracyjny

$$x_{n+1} = x_n - f(x_n)/g(x_n),$$

gdzie

$$g(x) := \left[f\big(x + f(x)\big) - f(x)\right]/f(x).$$

Wykazać, że wynikająca stąd metoda jest przy odpowiednich założeniach zbieżna kwadratowo.

18. *Metoda Halleya* rozwiązywania równania $f(x) = 0$ korzysta ze wzoru iteracyjnego

$$x_{n+1} = x_n - \frac{f_n f'_n}{(f'_n)^2 - (f_n f''_n)/2},$$

gdzie $f_n = f(x_n)$ itd. Wykazać, że jest to metoda równoważna metodzie Newtona zastosowanej do funkcji $f/\sqrt{f'}$.

19. Zastosować metodę Newtona w następujących przypadkach:

(**a**) Układ

$$xy - z^2 = 1, \qquad xyz - x^2 + y^2 = 2, \qquad e^x - e^y + z = 3,$$

punkt początkowy $(0, 0, 1)$, jedna iteracja.

(**b**) Układ

$$4x_1^2 - x_2^2 = 0, \qquad 4x_1x_2^2 - x_1 = 1,$$

punkt początkowy $(0, 1)$, dwie iteracje.

(**c**) Układ

$$xy^2 + x^2y + x^4 = 3, \qquad x^3y^5 - 2x^5y - x^2 = -2,$$

punkt początkowy $(1, 1)$, dwie iteracje.

ZADANIA KOMPUTEROWE 3.2

K1. Napisać krótki program rozwiązywania równania $x^3 + 3x = 5x^2 + 7$ metodą Newtona. Wykonać dziesięć kroków, zaczynając od $x_0 = 5$.

K2. Napisać program rozwiązujący równanie $x = \operatorname{tg} x$ metodą Newtona. Znaleźć pierwiastki bliskie punktów 4.5 i 7.7.

K3. (cd.). Napisać i sprawdzić program obliczania dziesięciu początkowych pierwiastków równania $x = \operatorname{tg} x$. (To zadanie jest znacznie trudniejsze od poprzedniego). Nota historyczna: Jeśli $\lambda_1, \lambda_2, \ldots$ są wszystkimi dodatnimi pierwiastkami tego równania, to $\sum_{i=1}^{\infty} \lambda_i^{-2} = 1/10$ (*Amer. Math. Monthly*, październik 1986, s. 660).

K4. Niech będzie $f(x) := x^{-2} \operatorname{tg} x$. Obliczając zera funkcji f' metodą Newtona, znaleźć najmniejszy dodatni punkt, w którym f osiąga minimum.

K5. Równanie $2x^4 + 24x^3 + 61x^2 - 16x + 1 = 0$ ma dwa pierwiastki bliskie 0.1. Wyznaczyć je metodą Newtona.

K6. W przykładzie 3.2.1 zbadać czułość pierwiastka na zakłócenie stałego składnika.

K7. Napisać program dla metody Newtona w arytmetyce zespolonej. Sprawdzić go dla podanych niżej funkcji i punktów początkowych.

(a) $f(z) := z^2 + 1, \quad z = 3 + i$

(b) $f(z) := z + \sin z - 3, \quad z = 2 - i$

(c) $f(z) := z^4 + z^2 + 2 + 3i, \quad z = 1$

K8. Wielomian $p(x) := x^3 + 94x^2 - 389x + 294$ ma zera: 1, 3 i -98. Punkt $x_0 = 2$ powinien więc być dobrym punktem początkowym do wyznaczenia metodą Newtona któregoś z mniejszych zer. Wykonać obliczenia i wyjaśnić, co one dają.

K9. Zastosować metodę Newtona w dziedzinie zespolonej do funkcji $f(z) := z^4 - 1$ i do każdego punktu początkowego $(0.1j, 0.1ki)$ z koła $|z| < 2$ na płaszczyźnie zespolonej. Oznaczyć wspólnym kolorem te punkty siatki, które generują ciągi zbieżne do tego samego pierwiastka. Przedstawić wyniki w kolorze na ekranie monitora lub na wydruku.

K10. Wykonać pięć iteracji metodą Newtona w dziedzinie zespolonej dla funkcji $f(z) := 1 + z^2 + e^z$ i punktu początkowego $z_0 = -1 + 4i$.

K11. Znaleźć cztery zera funkcji z poprzedniego zadania o coraz większych modułach. Jak można się dowiedzieć, że są to zera o najmniejszych modułach i że nie pominięto żadnego?

K12. Napisać program dla metody Steffensena (zob. zad. 17) i sprawdzić go dla równania z zad. K2.

K13. Używając metody Newtona, znaleźć rozwiązania poniższych układów nieliniowych.

(a) $4y^2 + 4y + 52x = 19, \quad 169x^2 + 3y^2 + 111x - 10y = 10$

(b) $x + e^{-x} + y^3 = 0, \quad x^2 + 2xy - y^2 + \operatorname{tg} x = 0$

(c) $1 + x^2 - y^2 + e^x \cos y = 0, \quad 2xy + e^x \sin y = 0$; punkt początkowy: $x_0 = -1$, $y_0 = 4$. Czy to zadanie wiąże się z zad. K10 i czy obliczenia w obu przypadkach przebiegają podobnie?

3.3. Metoda siecznych

Przypomnijmy, że metodę iteracyjną Newtona określa wzór

$$x_{n+1} := x_n - \frac{f(x_n)}{f'(x_n)}. \tag{3.3.1}$$

Jedną z jej wad jest to, że wymaga ona obliczania wartości pochodnej funkcji f, której zera szukamy. Aby usunąć tę wadę, proponowano wiele metod. Jedną z nich jest metoda Steffensena (zad. 3.2.17), opisana wzorem

$$x_{n+1} := x_n - \frac{[f(x_n)]^2}{f(x_n + f(x_n)) - f(x_n)}.$$

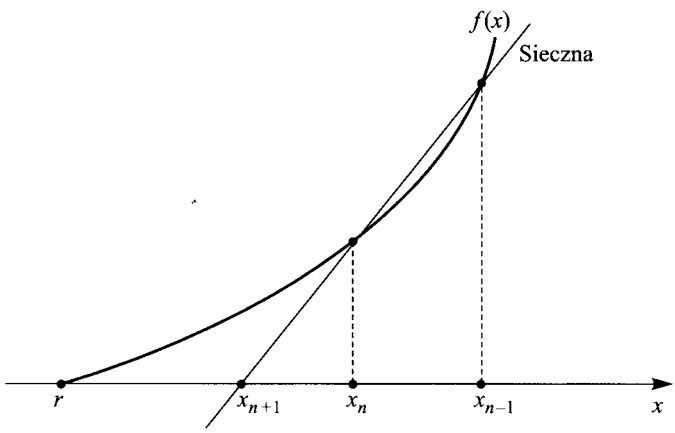
Inny pomysł polega na zastąpieniu w (3.3.1) pochodnej $f'(x)$ *ilorazem różnicowym*, na przykład takim:

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}.$$

Ta równość przybliżona, która wynika wprost z definicji pochodnej, daje *metodę siecznych* opisaną wzorem

$$x_{n+1} := x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \qquad (n \geqslant 1). \tag{3.3.2}$$

Ponieważ x_{n+1} wyraża się przez x_n i x_{n-1}, więc potrzebne są dwa punkty początkowe. Każde nowe x_{n+1} wymaga jednak obliczenia tylko jednej nowej wartości funkcji f (w metodzie Steffensena trzeba znaleźć dwie).



RYS. 3.6. Interpretacja geometryczna metody siecznych

Interpretacja graficzna metody siecznych różni się od tejże interpretacji metody Newtona tylko tym, że styczną do krzywej zastępujemy teraz sieczną; zob. rys. 3.6.

Przykład 3.3.1. Zastosować metodę siecznych do funkcji

$$f(x) = x^3 - \sinh x + 4x^2 + 6x + 9.$$

Rozwiązanie. Szkic wykresu funkcji sugeruje istnienie zera między 7 i 8, więc te liczby uznajemy za przybliżenia początkowe x_0 i x_1. Obliczenia wykonane na komputerze z precyzją arytmetyki równą 2^{-24} dają następujące wyniki:

n	x_n	$f(x_n)$
0	8.00000	$-0.665_{10}3$
1	7.00000	$0.417_{10}2$
2	7.05895	$0.208_{10}2$
3	7.11764	$-0.183_{10}1$
4	7.11289	$0.710_{10}-1$
5	7.11306	$0.244_{10}-3$
6	7.11306	$0.191_{10}-4$

■

Algorytm siecznych

W poniższym algorytmie zmodyfikowano nieco metodę siecznych tak, aby kolejne wartości funkcji miały moduły nierosnące.

input $a, b, M, \delta, \varepsilon$
$fa \leftarrow f(a)$; $fb \leftarrow f(b)$
output $0, a, fa$
output $1, b, fb$
for $k = 2$ **to** M **do**
  if $|fa| > |fb|$ **then**
    $a \leftrightarrow b$; $fa \leftrightarrow fb$
  end if
  $s \leftarrow (b - a)/(fb - fa)$
  $b \leftarrow a$
  $fb \leftarrow fa$
  $a \leftarrow a - fa * s$
  $fa \leftarrow f(a)$
  output k, a, fa
  if $|fa| < \varepsilon$ **or** $|b - a| < \delta$ **then stop**
end do

Zauważmy, że ten program przestawia końce a i b przedziału (co sygnalizuje symbol $\leftrightarrow$), gdy wymaga tego utrzymanie nierówności $|f(a)| \leqslant |f(b)|$. Dzięki temu, począwszy od drugiego kroku, moduły wartości funkcji w punktach x_n nie rosną.

Analiza błędu

Zbadamy teraz błędy metody siecznych. Program uwzględnia przestawianie przybliżeń. Niżej jednak dla uproszczenia tę czynność pomijamy.

Z definicji (3.3.2) metody siecznych wynika dla $e_n = x_n - r$, że

$$e_{n+1} = x_{n+1} - r = [f(x_n)x_{n-1} - f(x_{n-1})x_n]/[f(x_n) - f(x_{n-1})] - r =$$
$$= [f(x_n)e_{n-1} - f(x_{n-1})e_n]/[f(x_n) - f(x_{n-1})].$$

Wyłączamy tu czynnik $e_n e_{n-1}$ i wprowadzamy czynnik $x_n - x_{n-1}$ w liczniku i mianowniku:

$$e_{n+1} = \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \frac{f(x_n)/e_n - f(x_{n-1})/e_{n-1}}{x_n - x_{n-1}} e_n e_{n-1}. \tag{3.3.3}$$

Na mocy twierdzenia Taylora

$$f(x_n) = f(r + e_n) = f(r) + e_n f'(r) + \frac{1}{2} e_n^2 f''(r) + \mathcal{O}(e_n^3).$$

Ponieważ $f(r) = 0$, więc

$$f(x_n)/e_n = f'(r) + \frac{1}{2} e_n f''(r) + \mathcal{O}(e_n^2).$$

Po zmianie wskaźnika n na $n-1$ daje to równość

$$f(x_{n-1})/e_{n-1} = f'(r) + \frac{1}{2} e_{n-1} f''(r) + \mathcal{O}(e_{n-1}^2).$$

Odejmujemy stronami dwie ostatnie równości:

$$f(x_n)/e_n - f(x_{n-1})/e_{n-1} = \frac{1}{2}(e_n - e_{n-1}) f''(r) + \mathcal{O}(e_{n-1}^2).$$

Ponieważ $x_n - x_{n-1} = e_n - e_{n-1}$, więc

$$\frac{f(x_n)/e_n - f(x_{n-1})/e_{n-1}}{x_n - x_{n-1}} \approx \frac{1}{2} f''(r).$$

Natomiast pierwszy iloraz w (3.3.3) jest w przybliżeniu równy $1/f'(r)$. Wykazaliśmy więc, że

$$e_{n+1} \approx \frac{f''(r)}{2f'(r)} e_n e_{n-1} = C e_n e_{n-1}. \tag{3.3.4}$$

Jest to równanie podobne do równania (3.2.3) odpowiadającego metodzie Newtona. Aby ustalić charakter zbieżności metody siecznych, przyjmijmy, że zachodzi następująca równość asymptotyczna:

$$|e_{n+1}| \sim A|e_n|^\alpha, \tag{3.3.5}$$

gdzie A jest stałą dodatnią. Ta równość, definiująca *zbieżność rzędu* α, oznacza, że iloraz $|e_{n+1}|/(A|e_n|^\alpha)$ dąży do 1, gdy $n \to \infty$. Mamy więc

$$|e_n| \sim A|e_{n-1}|^\alpha, \quad \text{skąd} \quad |e_{n-1}| \sim (A^{-1}|e_n|)^{1/\alpha}. \tag{3.3.6}$$

W (3.3.4) wyrażamy wartości asymptotyczne wielkości $|e_{n+1}|$ i $|e_{n-1}|$ zgodnie z (3.3.5) i (3.3.6):

$$A|e_n|^\alpha \sim |C||e_n|A^{-1/\alpha}|e_n|^{1/\alpha}.$$

Po przekształceniu daje to równość

$$A^{1+1/\alpha}|C|^{-1} \sim |e_n|^{1-\alpha+1/\alpha}. \tag{3.3.7}$$

Lewa strona tego związku jest niezerową stałą, a $e_n \to 0$, więc musi być $1 - \alpha + 1/\alpha = 0$, czyli $\alpha = (1 + \sqrt{5})/2 \approx 1.62$ (wybieramy pierwiastek dodatni). Tak więc zbieżność metody siecznych jest *nadliniowa*, tj. lepsza od liniowej. Możemy teraz wyznaczyć A, gdyż prawa strona (3.3.7) jest równa 1. Ponieważ $1 + 1/\alpha = \alpha$, więc

$$A = |C|^{1/(1+1/\alpha)} = |C|^{1/\alpha} = |C|^{\alpha-1} \approx |C|^{0.62} = \left|\frac{f''(r)}{2f'(r)}\right|^{0.62}.$$

Dla tego A mamy ostateczną równość dotyczącą metody siecznych:

$$|e_{n+1}| \approx A|e_n|^{(1+\sqrt{5})/2}.$$

Ponieważ $(1 + \sqrt{5})/2 \approx 1.62 < 2$, więc metoda siecznych jest zbieżna wolniej od metody Newtona, ale szybciej od metody bisekcji. Zauważmy jednak, że każdy krok metody siecznych wymaga obliczenia tylko jednej wartości funkcji, a w metodzie Newtona trzeba obliczyć dwie takie wartości, mianowicie $f(x)$ i $f'(x)$. W obu metodach najbardziej kosztowne jest właśnie obliczanie wartości funkcji i w tym sensie para kroków metody siecznych jest porównywalna z jednym krokiem metody Newtona. Dla dwóch kroków pierwszej metody jest zaś

$$|e_{n+2}| \sim A|e_{n+1}|^\alpha \sim A^{1+\alpha}|e_n|^{\alpha^2} = A^{1+\alpha}|e_n|^{(3+\sqrt{5})/2}.$$

To jest znacznie lepszy wynik niż dla zbieżnej kwadratowo metody Newtona, gdyż $(3+\sqrt{5})/2 \approx 2.62$. Oczywiście, dwa kroki metody siecznych wymagają więcej pracy na iterację.

Trzy rozważone już metody (bisekcji, Newtona i siecznych) ilustrują ogólne zjawisko w analizie numerycznej: konflikt między szybkością i wiarygodnością. Szybkość jest bezpośrednio związana z kosztem obliczeń. Dla pewnych zadań, wymagających bardzo długich obliczeń (na przykład rozwiązywania numerycznego równań różniczkowych cząstkowych), o jakości programu decyduje szybkość jego działania. W oprogramowaniu, które ma być stosowane przez wielu różnych użytkowników, istotne są *wiarygodność*

i *odporność*. Algorytm albo podprogram jest *odporny*, jeśli radzi sobie z bardzo rozmaitymi sytuacjami numerycznymi bez interwencji użytkownika. W ciągu wielu lat czyniono herkulesowe wysiłki, aby stworzyć biblioteki programów ogólnego użytku, mających obie te cechy. Dwa dobre przykłady takich bibliotek to zbiory IMSL [1995] i NAG [1995].

W obliczeniach naukowych zamiast własnych programów lepiej stosować znane pakiety programów; wyjątkiem są szczególne zastosowania. Najlepsze programy obliczające pierwiastki (zera) są raczej złożone, gdyż muszą zapewniać zbieżność zarówno globalną, jak i lokalnie szybką. Jak już wspomniano, te dwa cele niekiedy wzajemnie się kłócą.

Istniejących metod i ich odmian jest tyle, że niestety nie można uwzględnić ich wszystkich w tej książce. Pominięto tu na przykład interesujące algorytmy, które opracowali Brent [1973], Dekker [1969] i Le [1985]. Łączą one zalety metod bisekcji i siecznych i wymagają tylko założenia $f(a)f(b) \leqslant 0$ o funkcji f, której zero mamy znaleźć. Brent łącząc wspomniane metody, stosuje odwrotną interpolację kwadratową, aby otrzymać procedurę bardziej odporną. Algorytm Le jest kombinacją metody bisekcji z metodami rzędu drugiego lub trzeciego, w których pochodne wyraża się w przybliżeniu przez wartości samej funkcji. Czytelnik zainteresowany tymi metodami może znaleźć ich pełny opis w pracach cytowanych wyżej. Ponieważ odpowiednie programy są długie i skomplikowane, warto ściągnąć je przez Internet. Nasza ogólna rada – aby używać, jeśli tylko to możliwe, znanych i sprawdzonych programów – pozostaje tu w mocy. Zwykle takie programy są dostępne bezpłatnie i były starannie zaprojektowane i sprawdzone. Możemy np. połączyć się ze stroną o adresie[1] `http://gams.nist.gov` („Guide to Available Mathematical Software" czyli „Przewodnik po oprogramowaniu matematycznym"). Programy są uporządkowane tematycznie; przechodząc w dół od F. *Rozwiązywanie równań nieliniowych* przez F1. *Pojedyncze równania* do F1b. *Równania niewielomianowe*, znajdujemy wersję algorytmu Brenta w języku C, służącą do szukania minimum lub zera funkcji jednej zmiennej w danym przedziale.

ZADANIA 3.3

1. Wykazać, że wzór określający metodę siecznych można napisać w postaci

$$x_{n+1} = \frac{f(x_n)x_{n-1} - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}.$$

Wyjaśnić, dlaczego w praktyce ten wzór jest gorszy od (3.3.2).

[1] Podany adres internetowy dotyczy oryginału książki, tzn. wydania w jęz. ang. (*przyp. red. WNT*).

2. Wykazać dla metody siecznych, że jeśli $x_n \to q$, gdy $n \to \infty$, i jeśli $f'(q) \neq 0$, to q jest zerem funkcji f.

3. Jak zamiana przybliżeń w programie metody siecznych wpływa na analizę błędu? Przedstawić szczegóły zmodyfikowanej analizy.

4. Korzystając ze wzoru Taylora dla $f(x+h)$ i $f(x+k)$, otrzymać następujące przybliżenie dla $f'(x)$:

$$f'(x) \approx \frac{k^2 f(x+h) - h^2 f(x+k) + (h^2 - k^2) f(x)}{(k-h)kh}.$$

5. Relacja $x_n \sim y_n$ asymptotycznej równości elementów dwóch ciągów oznacza, że $\lim_{n\to\infty} x_n/y_n = 1$. Udowodnić, że jeśli $x_n \sim y_n$, $u_n \sim v_n$ i $c \neq 0$, to:
(**a**) $cx_n \sim cy_n$, (**b**) $x_n^c \sim y_n^c$, (**c**) $x_n u_n \sim y_n v_n$, (**d**) jeśli dodatkowo $y_n \sim u_n$, to $x_n \sim v_n$, (**e**) $y_n \sim x_n$.

6. (Wielomian bardzo wysokiego stopnia). *Renta* jest funduszem powstającym z wpłat (niekoniecznie identycznych) wnoszonych w regularnych okresach. W pewnej formie renty fundusz jest oprocentowany ze stałą stopą procentową r, liczoną na okres. Odsetki są doliczane na końcu każdego okresu. Niech wpłaty wynoszą $a_1, a_2, \ldots$ Niech V_i oznacza łączną wysokość renty obliczoną tuż po dokonaniu wpłaty a_i. Stąd $V_1 = a_1$ i

$$V_i = V_{i-1}(1+r) + a_i \quad (i = 2, 3, \ldots).$$

Czynnik $1+r$ odpowiada doliczeniu odsetek do V_{r-1}. Wykazać, że zachodzi równość $V_n = \sum_{i=1}^{n} a_i x^{n-i}$, gdzie $x = 1+r$; zob. też zad. K6.

7. Dwóch ludzi stosuje różne strategie oszczędzania przez 44 lata. Pierwszy oszczędza 1000 $ rocznie przez sześć lat, a przez pozostałych 38 lat do zgromadzonej sumy są tylko doliczane odsetki. Drugi nie wpłaca nic w ciągu początkowych sześciu lat, a potem oszczędza 1000 $ rocznie. Po 44 latach oszczędności są takie same. Założyć, że stopa procentowa w obu przypadkach jest taka sama; odsetki są doliczane co roku. Jakie jest oprocentowanie i jaka suma jest zgromadzona na każdym rachunku?

Zadania komputerowe 3.3

K1. Napisać podprogram realizujący metodę siecznych dla funkcji f i danych dwóch punktów początkowych. Sprawdzić go dla poniższych funkcji.
(**a**) $\sin(x/2) - 1$, (**b**) $e^x - \operatorname{tg} x$, (**c**) $x^3 - 12x^2 + 3x + 1$.

K2. Zaprogramować i sprawdzić ulepszoną metodę siecznych wynikającą z użycia w metodzie Newtona przybliżonej wartości $f'(x)$ podanej w zad. 4 zamiast wartości dokładnej. Są teraz potrzebne trzy punkty początkowe. Dwa z nich mogą być dowolne, a trzeci można wyznaczyć metodą siecznych.

K3. Napisać procedury dla metod bisekcji, Newtona i siecznych, działających dla dowolnej funkcji F. W każdym przypadku wywołując procedurę należy podać maksymalną liczbę kroków M akceptowaną przez użytkownika i wymaganą dokładność (ε i δ, jak w programie z tego podrozdziału). Należy użyć pojedynczej precyzji.

(**a**) Sprawdzić te procedury dla funkcji $f(x) := \operatorname{arctg} x - 2x/(1+x^2)$. Zadbać o obliczenie zera dodatniego z pełną dokładnością maszynową.

(**b**) Połączyć dwie z napisanych procedur, tak aby uzyskać mieszaną metodę o dobrych cechach globalnej i lokalnej.

K4. Wyszukać opracowany już program, rozwiązujący równanie $f(x) = 0$ bez użycia pochodnych, i sprawdzić jego działanie dla poniższych funkcji i przedziałów:
(**a**) $x^{20} - 1$ w $[0, 10]$, (**b**) $\operatorname{tg} x - 30x$ w $[1, 1.57]$, (**c**) $x^2 - (1-x)^{10}$ w $[0, 1]$, (**d**) $x^3 + 10^{-4}$ w $[-0.75, 0.5]$, (**e**) $x^{19} + 10^{-4}$ w $[-0.75, 0.5]$, (**f**) x^5 w $[-1, 10]$, (**g**) x^9 w $[-1, 10]$, (**h**) xe^{-x^2} w $[-1, 4]$ (zob. Nerinckx i Haegemans [1976]).

K5. Zaprogramować i sprawdzić metodę siecznych na przykładzie 3.3.1. Powtórzyć obliczenia, wybierając 3 i 10 jako punkty początkowe. Wyjaśnić, co się wtedy dzieje.

K6. (zob. zad. 6). Wpłacono 60 miesięcznych rat. W latach od pierwszego do piątego te raty wyniosły odpowiednio: 200 \$, 275 \$, 312 \$, 380 \$ i 400 \$. Tuż po ostatniej wpłacie na koncie znalazło się 24738 \$. Jakie było miesięczne oprocentowanie wpłat? Użyć metody siecznych do znalezienia zera wielomianu, który wynika z tych założeń.

3.4. Punkty stałe i metody iteracyjne

Metody Newtona i Steffensena są przykładami procedur, w których dla danego równania $f(x) = 0$ ciąg punktów jest obliczany według wzoru

$$x_{n+1} := F(x_n) \qquad (n \geqslant 0), \tag{3.4.1}$$

gdzie F wyraża się przez f. Algorytm określony takim wzorem nazywamy *metodą iteracyjną*[2]. W metodzie Newtona

$$F(x) := x - \frac{f(x)}{f'(x)},$$

a w metodzie Steffensena

$$F(x) := x - \frac{[f(x)]^2}{f(x + f(x)) - f(x)}.$$

Można by zacytować wiele innych przykładów metod iteracyjnych. Niżej naszkicowano ich ogólną teorię.

[2] Ściślej, metodą iteracyjną *stacjonarną* (tj. taką, że F nie zależy od n), *jednopunktową* (nowy punkt x_{n+1} zależy bezpośrednio tylko od jednego poprzedniego punktu; nie jest tak w przypadku metody siecznych). Autorzy używają terminu *iteracja funkcyjna* (*przyp. tłum.*).

Wzór (3.4.1) może generować ciągi rozbieżne. Nas interesują głównie przypadki, w których granica $\lim_{n\to\infty} x_n$ istnieje i jest skończona. Załóżmy, że tak jest:

$$\lim_{n\to\infty} x_n = s.$$

Jak s wiąże się z F? Jeśli F jest ciągła, to

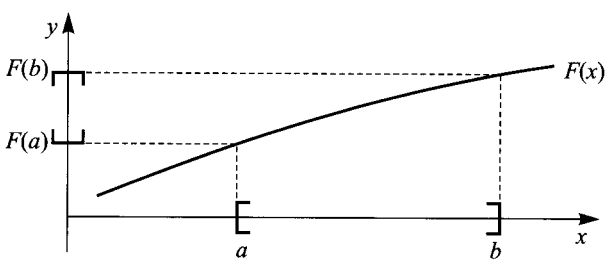
$$F(s) = F\left(\lim_{n\to\infty} x_n\right) = \lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} x_{n+1} = s.$$

Dlatego $F(s) = s$ i s nazywamy *punktem stałym* funkcji F.

Wiele problemów matematycznych można sprowadzić do poszukiwania punktu stałego pewnej funkcji. Bardzo ciekawych zastosowań dostarczają równania różniczkowe, teoria optymalizacji i inne dziedziny. Zwykle funkcja F, której punktów stałych szukamy, odwzorowuje jedną przestrzeń wektorową w inną. Zbadamy najprostszy przypadek, w którym F odwzorowuje pewien zbiór domknięty $C \subset \mathbb{R}$ w siebie. Twierdzenie, które udowodnimy, dotyczy *odwzorowań zwężających* czyli *kontrakcji*. Tak nazywamy odwzorowanie (albo funkcję) F, jeśli istnieje liczba $\lambda \in [0, 1)$ taka, że

$$|F(x) - F(y)| \leqslant \lambda |x - y| \tag{3.4.2}$$

dla dowolnych punktów x i y z dziedziny F. Funkcja zwężająca F przekształca więc odległość między x i y na mniejszą odległość między $F(x)$ i $F(y)$; zob. rys. 3.7.



RYS. 3.7. Przykład odwzorowania zwężającego

TWIERDZENIE 3.4.1. *Niech C będzie podzbiorem domkniętym osi rzeczywistej. Jeśli F jest odwzorowaniem zwężającym zbioru C w siebie, to F ma jedyny punkt stały. Ponadto ten punkt stały jest granicą każdego ciągu otrzymanego za pomocą wzoru* (3.4.1) *z punktu początkowego $x_0 \in C$.*

Dowód. Własność (3.4.2) wraz ze wzorem (3.4.1) daje związek

$$|x_n - x_{n-1}| = |F(x_{n-1}) - F(x_{n-2})| \leqslant \lambda |x_{n-1} - x_{n-2}|.$$

Korzystając z niego wielokrotnie, dostajemy

$$|x_n - x_{n-1}| \leqslant \lambda |x_{n-1} - x_{n-2}| \leqslant \ldots \leqslant \lambda^{n-1} |x_1 - x_0|. \tag{3.4.3}$$

Ponieważ

$$x_n = x_0 + (x_1 - x_0) + (x_2 - x_1) + \ldots + (x_n - x_{n-1}),$$

więc ciąg $\{x_n\}$ jest zbieżny wtedy i tylko wtedy, gdy jest zbieżny szereg

$$\sum_{n=1}^{\infty} (x_n - x_{n-1}).$$

Aby udowodnić jego zbieżność, wystarczy zrobić to samo dla szeregu

$$\sum_{n=1}^{\infty} |x_n - x_{n-1}|.$$

To zaś jest łatwe, gdyż możemy użyć kryterium porównawczego i nierówności otrzymanej wcześniej:

$$\sum_{n=1}^{\infty} |x_n - x_{n-1}| \leqslant \sum_{n=1}^{\infty} \lambda^{n-1} |x_1 - x_0| = \frac{1}{1-\lambda} |x_1 - x_0|.$$

Ciąg $\{x_n\}$ jest zatem zbieżny; niech s będzie jego granicą. Jak już zauważono, $F(s) = s$. (Zauważmy, że funkcja zwężająca F jest ciągła). Jeśli x i y są punktami stałymi, to

$$|x - y| = |F(x) - F(y)| \leqslant \lambda |x - y|,$$

a ponieważ $\lambda < 1$, więc $|x - y| = 0$, czyli istnieje tylko jeden punkt stały. Zauważmy na koniec, że otrzymane s należy do C, gdyż s jest granicą ciągu punktów z C. ■

Twierdzenie 3.4.1 jest prawdziwe dla odwzorowania dowolnej przestrzeni metrycznej zupełnej w siebie[3)].

[3)] Udowodnił to Stefan Banach (*przyp. tłum.*).

Przykład 3.4.2. Wykazać, że ciąg $\{x_n\}$ z tw. 3.4.1 spełnia warunek Cauchy'ego zbieżności (przypomnijmy, że jest on następujący: dla dowolnego $\varepsilon > 0$ istnieje N całkowite takie, że $|x_n - x_m| < \varepsilon$, jeśli tylko $n, m \geqslant N$).

Rozwiązanie. Jeśli $n \geqslant m \geqslant N$, to z nierówności trójkąta i z (3.4.3) wynika, że

$$\begin{aligned}|x_n - x_m| &\leqslant |x_n - x_{n-1}| + |x_{n-1} - x_{n-2}| + \ldots + |x_{m+1} - x_m| \leqslant \\ &\leqslant \lambda^{n-1}|x_1 - x_0| + \lambda^{n-2}|x_1 - x_0| + \ldots + \lambda^m|x_1 - x_0| = \\ &= \lambda^m|x_1 - x_0|(1 + \lambda + \lambda^2 + \ldots + \lambda^{n-1-m}) \leqslant \\ &\leqslant \lambda^N|x_1 - x_0|(1 + \lambda + \lambda^2 + \ldots) = \lambda^N|x_1 - x_0|(1 - \lambda)^{-1}.\end{aligned}$$

Dla każdego $\varepsilon > 0$ istnieje N takie, że $|x_n - x_m| < \varepsilon$, gdy $n, m \geqslant N$. Możemy bowiem wybrać N tak duże, żeby było $\lambda^N|x_1 - x_0|(1 - \lambda)^{-1} < \varepsilon$. ■

Przykład 3.4.3. Udowodnić, że ciąg $\{x_n\}$ określony rekurencyjnie wzorami

$$x_0 := -15, \qquad x_{n+1} := 3 - \frac{1}{2}|x_n| \qquad (n \geqslant 0)$$

jest zbieżny.

Rozwiązanie. Funkcja $F(x) = 3 - \frac{1}{2}|x|$ jest zwężająca, gdyż wobec nierówności trójkąta

$$|F(x) - F(y)| = |(3 - \tfrac{1}{2}|x|) - (3 - \tfrac{1}{2}|y|)| = \tfrac{1}{2}||y| - |x|| \leqslant \tfrac{1}{2}|y - x|.$$

Na mocy tw. 3.4.1 określony wyżej ciąg jest zbieżny do jedynego punktu stałego funkcji F, którym oczywiście jest liczba 2. W twierdzeniu zbiorem C może być dowolnie szeroki przedział. ■

Przykład 3.4.4. Stosując tw. 3.4.1, udowodnić, że funkcja

$$F(x) = 4 + \tfrac{1}{3}\sin 2x$$

ma punkt stały i obliczyć go.

Rozwiązanie. Z twierdzenia o wartości średniej wynika, że

$$|F(x) - F(y)| = \frac{1}{3}|\sin 2x - \sin 2y| = \frac{2}{3}|\cos 2\zeta||x - y| \leqslant \frac{2}{3}|x - y|$$

dla pewnego ζ między x i y. To pokazuje, że F jest zwężająca, przy czym $\lambda = \frac{2}{3}$. Na mocy tw. 3.4.1 F ma punkt stały. Program obliczenia tego punktu stałego może opierać się na następującym algorytmie, który poleca wykonać 20 iteracji dla punktu początkowego 4:

$x \leftarrow 4$; $M \leftarrow 20$
for $k = 1$ **to** M **do**
    $x \leftarrow 4 + \frac{1}{3}\sin 2x$
    output k, x
end do

Program produkuje 20 wierszy wyników; niektóre pokazano niżej:

k	x
1	4.3297861
2	4.2308951
3	4.2736338
.....	
14	4.2614830
15	4.2614840
16	4.2614836
.....	
20	4.2614837

Punkt stały z dokładnymi siedmioma cyframi po kropce znajduje się w ostatnim wierszu. ■

Analiza błędu

Zbadajmy teraz błędy metody iteracyjnej. Zakładamy, że F ma punkt stały s i że ciąg $\{x_n\}$ jest określony wzorem $x_{n+1} := F(x_n)$. Niech będzie

$$e_n := x_n - s.$$

Jeśli pochodna F' istnieje i jest ciągła, to na mocy twierdzenia o wartości średniej

$$x_{n+1} - s = F(x_n) - F(s) = F'(\zeta_n)(x_n - s)$$

czyli

$$e_{n+1} = F'(\zeta_n)e_n,$$

gdzie ζ_n leży między x_n i s. Warunek $|F'(x)| < 1$ spełniony dla każdego x zapewnia, że moduły błędów maleją. Jeśli e_n jest małe, to ζ_n leży blisko s i $F'(\zeta_n) \approx F'(s)$. Można oczekiwać szybkiej zbieżności, gdy $F'(s)$ jest bliskie 0. Idealną sytuację mamy dla $F'(s) = 0$. Wtedy trzeba uwzględnić dodatkowe składniki szeregu Taylora. Aby rozważyć jednocześnie wszystkie możliwe przypadki, załóżmy, że dla pewnej liczby naturalnej q jest

$$F^{(k)}(s) = 0 \quad (1 \leqslant k < q), \qquad F^{(q)}(s) \neq 0.$$

Stąd i ze wzoru Taylora w punkcie s wynika, że

$$e_{n+1} = x_{n+1} - s = F(x_n) - F(s) = F(s + e_n) - F(s) = \frac{1}{q!} e_n^q F^{(q)}(\zeta_n).$$

Wobec tego z równości $\lim_{n\to\infty} x_n = s$ wynika, że

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^q} = \frac{1}{q!} F^{(q)}(s).$$

Zgodnie z definicją podaną w podrozdz. 1.2 liczba q jest rzędem zbieżności (do s) ciągu $\{x_n\}$.

Jeśli na przykład $F'(s) = 0$ i $F''(s) \neq 0$, to $q = 2$ i mamy

$$e_{n+1} = \frac{1}{2} e_n^2 F''(\zeta_n).$$

To się kojarzy z metodą Newtona (por. (3.2.3)). Istotnie, jest w niej

$$F(x) = x - \frac{f(x)}{f'(x)},$$

a wtedy

$$F'(x) = 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{[f'(x)]^2} = \frac{f(x)f''(x)}{[f'(x)]^2}.$$

Ponieważ punkt stały funkcji F jest zerem funkcji f, więc $F'(s) = 0$. Ponadto

$$F''(x) = \frac{[f'(x)]^2[f(x)f'''(x) + f''(x)f'(x)] - 2f(x)f''(x)f'(x)f''(x)}{[f'(x)]^4},$$

więc $F''(s) = f''(s)/f'(s)$ i na ogół ta wielkość jest różna od 0.

ZADANIA 3.4

1. Wykazać, że poniższe funkcje są zwężające w podanych obok przedziałach. Wyznaczyć najlepsze wartości λ z nierówności (3.4.2).
(**a**) $(1+x^2)^{-1}$, dowolny przedział; (**b**) $\frac{1}{2}x$, $[1,5]$; (**c**) $\operatorname{arctg} x$, dowolny przedział domknięty niezawierający zera; (**d**) $|x|^{3/2}$, $|x| \leqslant \frac{1}{3}$.

2. Jeśli funkcja F jest zwężająca z $[a,b]$ w $[a,b]$ i $x_{n+1} = F(x_n)$ dla $x_0 \in [a,b]$, to $|x_n - s| \leqslant C\lambda^n$ dla pewnego C (s jest punktem stałym dla F). Udowodnić to i podać oszacowanie z góry dla C.

3. Wykazać, że jeśli $F\colon [a,b] \to \mathbb{R}$, F' jest ciągła i $|F'(x)| < 1$ w $[a,b]$, to funkcja F jest zwężająca. Czy taka F musi mieć punkt stały? Czy własność opisana w pierwszym zdaniu zachowuje się po zmianie przedziału na otwarty?

4. Udowodnić, że jeśli F jest odwzorowaniem ciągłym przedziału $[a, b]$ w siebie, to F musi mieć punkt stały. Zbadać, czy ta własność się utrzymuje po zmianie przedziału na $\mathbb{R}$.

5. Niech F będzie odwzorowaniem zwężającym przedziału $[a, b]$ w siebie, a s jego punktem stałym. Czy stąd, że $a \leqslant x \leqslant b$ i $|F(x) - x| < \varepsilon$, wynika nierówność $|x - s| < \varepsilon$? Wykazać, że $|x - s| < \varepsilon(1 - \lambda)^{-1}$, gdzie λ jest stałą z (3.4.2).

6. Udowodnić, że jeśli f jest ciągła w $[a, b]$ i taka, że $a \leqslant f(a)$, $f(b) \leqslant b$, to ma punkt stały w $[a, b]$. Zauważmy, że nie zakłada się, iż $a \leqslant f(x) \leqslant b$ dla wszystkich $x \in [a, b]$.

7. Jaki jest najsłabszy warunek dotyczący przedziału $[c, d]$, zapewniający, że każde ciągłe odwzorowanie $[a, b]$ w $[c, d]$ ma punkt stały?

8. Udowodnić albo obalić następujące twierdzenie: Jeśli $F\colon \mathbb{R} \to [a, b]$ i jeśli F jest zwężająca w $[a, b]$, to F ma jedyny punkt stały, który można otrzymać metodą iteracyjną dla dowolnego rzeczywistego punktu początkowego.

9. Podać przykłady funkcji, które nie mają punktu stałego, choć mają następujące własności:

(**a**) $f\colon [0,1] \to [0,1]$, (**b**) $f\colon (0,1) \to (0,1)$ i f jest ciągła, (**c**) $f\colon A \to A$, gdzie $A = [0,1] \cup [2,3]$ i f jest ciągła, (**d**) $f\colon \mathbb{R} \to \mathbb{R}$ i f jest ciągła.

10. Niech F ma ciągłą pochodną w przedziale otwartym i punkt stały s w tym przedziale. Wykazać, że jeśli $|F'(x)| < 1$, to ciąg określony za pomocą metody iteracyjnej jest zbieżny do s dla punktów początkowych dostatecznie bliskich s. Wskazówka: Wybrać λ takie, że $|F'(s)| < \lambda < 1$ i rozważyć taki przedział o środku w s, w którym $|F'(x)| < \lambda$.

11. Wykazać, że dla $x \in [0, \pi]$ i $0 < \varepsilon < 1$ równanie Keplera $x = y - \varepsilon \sin y$ (znane z astronomii) ma rozwiązanie y. Zinterpretować to zadanie jako dotyczące punktu stałego.

12. Wprowadzamy do kalkulatora pewną liczbę i naciskamy wielokrotnie klawisz funkcji cos. Jakich wyników można się spodziewać?

13. Rozważyć metodę iteracyjną określoną przez funkcję $F(x) = x + f(x)g(x)$, gdzie $f(r) = 0$ i $f'(r) \neq 0$. Znaleźć ścisłe warunki, jakie powinna spełniać funkcja g, aby ta metoda była zbieżna sześciennie do r dla dostatecznie bliskich punktów początkowych. Co otrzymamy? Obmyślić dowód.

14. Wykazać, że ciąg tworzony według wzoru $x_{n+1} = F(x_n)$ jest zbieżny, jeśli $|F'(x)| \leqslant \lambda < 1$ w przedziale $[x_0 - \rho, x_0 + \rho]$, gdzie $\rho = |F(x_0) - x_0|/(1 - \lambda)$.

15. Jakie szczególne własności musi mieć funkcja f, aby zastosowana do niej metoda Newtona była zbieżna sześciennie do zera tej funkcji?

16. Jaki wzór iteracyjny wynika z próby znalezienia punktu stałego funkcji F metodą Newtona, zastosowaną do równania $F(x) - x = 0$?

17. Jeśli f' jest ciągła i dodatnia w $[a, b]$ oraz $f(a)f(b) < 0$, to f ma dokładnie jedno zero w (a, b). Udowodnić to i wykazać, że dla pewnego λ to zero można otrzymać, stosując dla $F(x) = x + \lambda f(x)$ metodę iteracyjną.

18. Niech p będzie liczbą dodatnią. Znaleźć wartość wyrażenia

$$x = \sqrt{p + \sqrt{p + \sqrt{p + \dots}}}$$

rozumianą jako granica $\lim_{n\to\infty} x_n$, gdzie $x_1 := \sqrt{p}$, $x_{n+1} := \sqrt{p + x_n}$.

19. Niech będzie $p > 1$. Znaleźć wartość ułamka łańcuchowego (podrozdz. 6.11)

$$x = \cfrac{1}{p + \cfrac{1}{p + \cfrac{1}{p + \dots}}}$$

rozumianą podobnie jak w poprzednim zadaniu.

20. Za pomocą odpowiednio dobranej metody iteracyjnej obliczać pierwiastki równania kwadratowego $x^2 + px + q = 0$.

21. Udowodnić, że funkcja F określona wzorem $F(x) = 4x(1 - x)$ odwzorowuje przedział $[0, 1]$ na siebie, ale nie jest zwężająca. Wykazać, że ma ona punkt stały. Dlaczego nie przeczy to twierdzeniu o odwzorowaniu zwężającym?

22. Jaki rząd zbieżności ma metoda iteracyjna zastosowana do funkcji $F(x) = 2 + (x - 2)^4$ dla punktu początkowego $x = 2.5$? Znaleźć zakres wartości początkowych, dla których metoda jest zbieżna. Zauważyć, że 2 jest punktem stałym.

23. Niech będzie $F(x) = 2x - qx^2$, gdzie $\frac{1}{2} \leqslant q \leqslant 1$. Dla jakiego przedziału można być pewnym, że metoda iteracyjna dla F daje ciąg zbieżny do punktu stałego? (Zadanie wiąże się z zad. 3.2.4).

24. Metodę iteracyjną stosujemy do $F(x) = x^2 + x - 2$. Jeśli daje ona ciąg zbieżny liczb dodatnich, to jaka jest jego granica i jaki był punkt początkowy?

25. Wykazać, że poniższe funkcje są zwężające w podanych zbiorach, ale nie mają tam punktów stałych. Dlaczego to nie przeczy twierdzeniu o odwzorowaniu zwężającym?

(**a**) $F(x) = 3 - x^2$ w $\left[-\frac{1}{4}, \frac{1}{4}\right]$, (**b**) $F(x) = -x/2$ w $[-2, -1] \cup [1, 2]$.

26. Zero funkcji f jest równe punktowi stałemu funkcji $F(x) = x - f(x)/f'(x)$. Ten punkt możemy znaleźć, rozwiązując równanie $F(x) - x = 0$ metodą Newtona. Jaki jest wtedy wzór generujący ciąg $\{x_n\}$?

27. Czy zastosowanie metody iteracyjnej do funkcji $f(x) = \frac{1}{2}(1 + x^2)^{-1}$ dla $x_0 = 7$ daje ciąg zbieżny? Jeśli tak, to jaka jest jego granica?

28. Wykazać, że dla funkcji $f(x) = 2 + x - \operatorname{arctg} x$ jest $|f'(x)| < 1$ i że f nie ma punktu stałego. Wyjaśnić, dlaczego to nie jest sprzeczne z twierdzeniem o odwzorowaniu zwężającym.

29. Wykazać, że funkcja F taka, że $F(x) = 10 - 2x$, ma punkt stały. Dla dowolnego x_0 przyjmujemy, że $x_{n+1} = F(x_n)$ $(n \geqslant 0)$. Znaleźć jawne wyrażenie dla x_n. Wykazać, że ciąg $\{x_n\}$ jest zbieżny tylko dla szczególnej wartości x_0. Jaka ona jest? Dlaczego nie ma tu sprzeczności z twierdzeniem o odwzorowaniu zwężającym?

30. W którym z przedziałów: $[\frac{1}{2},\infty)$, $[\frac{1}{8},1]$, $[\frac{1}{4},2]$, $[0,1]$, $[\frac{1}{5},\frac{3}{2}]$ funkcja $f(x) := \sqrt{x}$ jest zwężająca?

31. Udowodnić, że metoda obliczania $\sqrt{R}$ za pomocą wzoru

$$x_{n+1} := \frac{x_n(x_n^2 + 3R)}{3x_n^2 + R}$$

ma rząd równy 3.

32. Funkcja F jest nazywana *kontrakcją iterowaną*, jeśli

$$|F(F(x)) - F(x)| \leqslant \lambda|F(x) - x| \qquad (\lambda < 1).$$

Udowodnić, że każda kontrakcja jest kontrakcją iterowaną, ale kontrakcja iterowana nie musi być ani kontrakcją, ani funkcją ciągłą.

33. Procesy iteracyjne na ogół nie wyrażają się prostym wzorem $x_{n+1} := F(x_n)$ dla $F\colon \mathbb{R} \to \mathbb{R}$. Może być np. $F\colon \mathbb{R}^2 \to \mathbb{R}^2$. Wykazać, że tak właśnie jest dla metody bisekcji i metody siecznych. W obu przypadkach podać jawne wyrażenie dla F.

3.5. Obliczanie pierwiastków wielomianów

Metody opisane w poprzednich podrozdziałach, a w szczególności metodę Newtona, można oczywiście stosować do wielomianów. Szukając pierwiastków (zer) wielomianów powinniśmy jednak, jeśli to możliwe, uwzględniać specjalną strukturę tych funkcji. Dodatkową komplikacją w przypadku wielomianów jest to, że często chcemy obliczać także pierwiastki zespolone (nawet jeśli wielomian ma współczynniki rzeczywiste) lub wszystkie pierwiastki danego wielomianu. Dlatego poszukiwanie pierwiastków wielomianów budzi szczególne zainteresowanie już niemal od 400 lat.

Zaczniemy od pewnych ważnych wyników teoretycznych, zresztą w większości zapewne znanych czytelnikom. Wielomian piszemy w postaci

$$p(z) = a_n z^n + a_{n-1} z^{n-1} + \ldots + a_1 z + a_0, \tag{3.5.1}$$

gdzie współczynniki a_k i zmienna z mogą być zespolone. Jeśli $a_n \neq 0$, to p ma *stopień* n. Szukamy pierwiastków wielomianu p. Czy jednak one istnieją? Na to pytanie odpowiada poniższe twierdzenie, zwane *zasadniczym twierdzeniem algebry*, a udowodnione po raz pierwszy przez Gaussa w 1799 r. Natomiast podany tu dowód pochodzi z r. 1938 od Hardy'ego; zob. Hardy [1960, dodatek I] i Fefferman [1967].

TWIERDZENIE 3.5.1. *Każdy wielomian różny od stałej ma co najmniej jeden pierwiastek w ciele $\mathbb{C}$ liczb zespolonych.*

Dowód. Niech p będzie wielomianem różnym od stałej. Chcemy wykazać, że $p(z_0) = 0$ dla pewnego $z_0 \in \mathbb{C}$. Ponieważ p nie jest stałą, więc $|p(z)| \to \infty$, gdy $|z| \to \infty$. Niech D będzie takim kołem o środku w 0, że poza nim $|p(z)| \geqslant |p(0)|$. Niech z_0 będzie punktem, w którym jest osiągnięte $\inf_{z\in D}|p(z)|$. Ponieważ $0 \in D$, więc $|p(z_0)| \leqslant |p(0)|$. Wobec tego $|p(z_0)| \leqslant |p(z)|$ dla wszystkich $z \in \mathbb{C}$. Niech będzie $q(z) := p(z + z_0)$. Chcemy wykazać, że $q(0) = 0$, czyli $p(z_0) = 0$. Wyrażamy q w postaci $q(z) = c_0 + c_j z^j + \ldots + c_n z^n = c_0 + c_j z^j + z^{j+1} r(z)$, gdzie $c_j \neq 0$, a r jest wielomianem (być może równym 0). Chcemy teraz udowodnić, że $c_0 = 0$. Przypuśćmy, że $c_0 \neq 0$. Niech w będzie dowolną liczbą zespoloną, dla której $c_j w^j = -c_0$. Definiujemy $N := \sup_{0<\varepsilon<1}|r(\varepsilon w)|$. Wybieramy $\varepsilon \in (0,1)$ tak małe, żeby było $\varepsilon|w|^{j+1}N < |c_0|$. Wtedy poniższy ciąg nierówności prowadzi do sprzeczności:

$$\begin{aligned}|q(\varepsilon w)| &\leqslant |c_0 + c_j\varepsilon^j w^j| + \varepsilon^{j+1}|w^{j+1}||r(\varepsilon w)| = \\ &= |c_0 - c_0\varepsilon^j| + \varepsilon^j\varepsilon|w|^{j+1}N < |c_0|(1-\varepsilon^j) + \varepsilon^j|c_0| = \\ &= |c_0| = |q(0)| = |p(z_0)| \leqslant |p(z_0 + \varepsilon w)| = |q(\varepsilon w)|.\end{aligned}$$

■

Historię tego twierdzenia i prób jego udowodnienia opisuje Kline [1972, s. 597–606]. Zwykły nowoczesny dowód jest wnioskiem z twierdzenia Liouville'a; zob. na przykład Bak i Newman [1982], Henrici [1974], Ahlfors [1966] i inne podręczniki analizy zespolonej.

Zasadnicze twierdzenie algebry nie zapewnia istnienia pierwiastków rzeczywistych. Przykład tak prostego wielomianu jak z^2+1 pokazuje, że nawet wielomian o wszystkich współczynnikach rzeczywistych może nie mieć pierwiastków rzeczywistych.

Jeśli wielomian p stopnia $n \geqslant 1$ podzielimy przez czynnik liniowy $z-c$, to otrzymamy iloraz q i resztę r. Iloraz jest wielomianem stopnia $n-1$, a reszta – liczbą zespoloną. Wykonaną czynność opisuje równanie

$$p(z) = (z-c)q(z) + r.$$

Stąd wynika (po podstawieniu $z = c$), że $p(c) = r$. Ten fakt jest znany jako *twierdzenie o reszcie*. Jeśli c jest pierwiastkiem wielomianu p, to $r = 0$ i mamy równość

$$p(z) = (z-c)q(z).$$

Tak więc $z-c$ jest czynnikiem wielomianu $p(z)$. Ten wniosek jest nazywany *twierdzeniem o czynniku*.

Wiemy, że $p(z) = (z-r_1)q_1(z)$, gdzie r_1 jest dowolnym pierwiastkiem p. Na mocy zasadniczego twierdzenia algebry q_1 ma, jeśli jego stopień jest

dodatni, pierwiastek, np. r_2. Tak więc $q_1(z)$ ma czynnik $z - r_2$ i możemy napisać, że $p(z) = (z - r_1)(z - r_2)q_2(z)$. Kontynuacja takiego rozumowania musi się zakończyć, gdyż stopnie kolejnych q_k zmniejszają się o 1 w każdym kroku. Tak więc q_n jest stałą i dochodzimy ostatecznie do równości

$$p(z) = (z - r_1)(z - r_2)\dots(z - r_n)q_n.$$

Dowodzi to, że wielomian p stopnia n rozkłada się na iloczyn n czynników liniowych, z których każdy odpowiada pewnemu pierwiastkowi tego wielomianu. Oczywiście p nie może mieć innych pierwiastków, gdyż jeśli z jest dowolną liczbą zespoloną różną od $r_1, r_2, \dots, r_n$, to iloczyn $\prod_{k=1}^{n}(z - r_k)$ jest różny od 0. Ponieważ pewne pierwiastki r_k mogą być sobie równe, więc wielomian stopnia n ma co najwyżej n różnych pierwiastków. *Krotność* pierwiastka r jest liczbą czynników $z-r$ występujących w rozkładzie pierwiastka. Łącząc te uwagi z zasadniczym twierdzeniem algebry, otrzymujemy następujący wniosek:

TWIERDZENIE 3.5.2. *Wielomian stopnia n ma dokładnie n pierwiastków na płaszczyźnie zespolonej, gdy każdy z nich jest liczony tyle razy, ile wynosi jego krotność.*

Często chcielibyśmy wiedzieć, jak z grubsza pierwiastki wielomianu są rozmieszczone na płaszczyźnie zespolonej. Mamy tu zadanie *lokalizacji pierwiastków*. Odnosi się do niego wiele wyników znanych z publikacji, w tym reguła Descartesa znaków; zob. na przykład książki Younga i Gregory'ego [1972, § 5.5] oraz Stoera i Bulirscha [1980, § 5.5], monografię Mardena [1966], tom I trzytomowej monografii Henriciego [1974] i książkę Turowicza [*1967].

Następujące twierdzenie daje łatwe do obliczenia oszacowanie z góry modułów pierwiastków:

TWIERDZENIE 3.5.3. *Wszystkie pierwiastki wielomianu* (3.5.1) *leżą w kole otwartym o środku w punkcie* 0 *płaszczyzny zespolonej i promieniu*

$$\rho := 1 + |a_n|^{-1} \max_{0 \leqslant k < n} |a_k|.$$

Dowód. Niech będzie $c = \max_{0 \leqslant k < n} |a_k|$, skąd $c|a_n|^{-1} = \rho - 1$. Jeśli $c = 0$, to twierdzenie jest oczywiście prawdziwe. W przeciwnym razie jest $\rho > 1$ i dla $|z| \leqslant \rho$ mamy nierówność

$$|p(z)| \geqslant |a_n z^n| - |a_{n-1}z^{n-1} + \ldots + a_0| \geqslant |a_n z^n| - c\sum_{k=0}^{n-1}|z|^k >$$
$$> |a_n z^n| - c|z|^n(|z|-1)^{-1} = |a_n z^n|[1 - c|a_n|^{-1}(|z|-1)^{-1}] \geqslant$$
$$\geqslant |a_n z^n|[1 - c|a_n|^{-1}(\rho - 1)^{-1}] = 0. \quad \blacksquare$$

Przykład 3.5.4. Znaleźć koło o środku w punkcie 0, zawierające wszystkie pierwiastki wielomianu $p(z) = z^4 - 4z^3 + 7z^2 - 5z - 2$.

Rozwiązanie. Na mocy tw. 3.5.3 to koło ma promień

$$\rho = 1 + |a_4|^{-1} \max_{0 \leqslant k < 4} |a_k| = 8. \quad \blacksquare$$

Inne wiadomości o lokalizacji pierwiastków wielomianu (3.5.1) daje funkcja

$$s(z) := z^n p(1/z) = z^n[a_n(1/z)^n + a_{n-1}(1/z)^{n-1} + \ldots + a_0] =$$
$$= a_n + a_{n-1}z + a_{n-2}z^2 + \ldots + a_0 z^n,$$

czyli wielomian stopnia niewiększego od n, o współczynnikach takich jak w p, ale uporządkowanych odwrotnie. Oczywiście dla różnej od zera liczby zespolonej z_0 warunki $p(z_0) = 0$ i $s(1/z_0) = 0$ są równoważne. Wynika stąd poniższy wniosek.

Twierdzenie 3.5.5. *Jeśli wszystkie pierwiastki wielomianu s leżą w kole $|z| \leqslant \rho$, to wszystkie niezerowe pierwiastki wielomianu p leżą poza kołem $|z| < \rho^{-1}$.*

Przykład 3.5.6. Znaleźć koło o środku w punkcie 0, w którym wielomian p z przykł. 3.5.4 nie ma żadnego pierwiastka.

Rozwiązanie. Wielomian s z tw. 3.5.5 jest następujący:

$$s(z) = -2z^4 - 5z^3 + 7z^2 - 4z + 1.$$

Na mocy tw. 3.5.3 wszystkie pierwiastki tego wielomianu leżą w kole $|z| < \rho$, gdzie $\rho = 1 + |a_4|^{-1} \max_{0 \leqslant k < 4} |a_k| = \frac{9}{2}$; z tw. 3.5.5 wynika zatem, że pierwiastki wielomianu p leżą poza kołem o promieniu $\frac{2}{9}$. Ostatecznie wnioskujemy, że wszystkie te pierwiastki leżą w pierścieniu $\frac{2}{9} < |z| < 8$ na płaszczyźnie zespolonej. $\blacksquare$

Schemat Hornera

Rozważmy proste zadanie: obliczanie wartości $v = p(z_0)$ wielomianu (3.5.1) stopnia n o danych współczynnikach a_k w danym punkcie z_0. Oczywisty algorytm, czyli obliczanie potęg wielkości z_0, ich mnożenie przez współczynni-

ki wielomianu i sumowanie tych iloczynów, jest zbyt kosztowny, bo wymaga $2n-1$ mnożeń i n dodawań. Natomiast wyrażając wielomian w postaci

$$a_0 + z(a_1 + z(a_2 + \ldots + z(a_{n-1} + a_n z)\ldots)),$$

otrzymujemy od razu taki algorytm obliczania v, w którym wystarczy wykonać n mnożeń i tyleż dodawań:

input $n, (a_i\colon\ 0 \leqslant i \leqslant n), z_0$
$v \leftarrow a_n$
for $k = n-1$ **to** 0 **step** -1 **do**
    $v \leftarrow a_k + z_0 v$
end do
output v

Ten algorytm jest nazywany tradycyjnie *schematem Hornera*. Jego warianty są użyteczne także w innych zadaniach. Jednym z nich jest dzielenie wielomianu $p(z)$ przez dwumian $z - z_0$, a ściślej – wyznaczenie *ilorazu* $q(z)$ i *reszty* z tego dzielenia. Jest ona równa $p(z_0)$, co wynika z tożsamości

$$p(z) = (z - z_0)q(z) + p(z_0). \tag{3.5.2}$$

Przedstawmy wielomian q (jego stopień jest oczywiście o 1 mniejszy od stopnia wielomianu p) w postaci

$$q(z) = b_0 + b_1 z + \ldots + b_{n-1} z^{n-1}.$$

Współczynniki obu stron równości (3.5.2) przy tych samych potęgach zmiennej z muszą być identyczne. Wynikają stąd równości

$$b_{n-1} = a_n, \quad b_{n-2} = a_{n-1} + z_0 b_{n-1}, \ \ldots,$$
$$b_0 = a_1 + z_0 b_1, \quad p(z_0) = a_0 + z_0 b_0.$$

Jak widać, współczynniki wielomianu q i – na końcu – wartość $p(z_0)$ obliczamy, wykonując dokładnie te same działania, które występują w podanym wyżej algorytmie. Różnica polega tylko na tym, że teraz zapamiętujemy także te wyniki pośrednie, które dają współczynniki b_k:

input n, $(a_i\colon\ 0 \leqslant i \leqslant n)$, z_0
$b_{n-1} \leftarrow a_n$
for $k = n-1$ **to** 0 **step** -1 **do**
    $b_{k-1} \leftarrow a_k + z_0 b_k$
end do
output $(b_i\colon\ -1 \leqslant i \leqslant n-1)$

W tej wersji wartością $p(z_0)$ jest b_{-1}. Stosując powyższy algorytm w obliczeniach ręcznych, rozmieszczamy często dane i wyniki w następujący sposób:

	a_n	a_{n-1}	$\dots$	a_1	a_0
z_0		$z_0 b_{n-1}$	$\dots$	$z_0 b_1$	$z_0 b_0$
	b_{n-1}	b_{n-2}	$\dots$	b_0	$\boxed{b_{-1}}$

Liczba w ramce jest równa $p(z_0)$.

Przykład 3.5.7. Za pomocą schematu Hornera obliczyć $p(3)$, gdzie wielomian p jest określony w przykł. 3.5.4.

Rozwiązanie. Wielkości występujące w obliczeniach ustawiamy zgodnie z powyższą sugestią:

	1	−4	7	−5	−2
3		3	−3	12	21
	1	−1	4	7	$\boxed{19}$

Jest zatem $p(3) = 19$ i możemy napisać, że

$$p(z) = (z-3)(z^3 - z^2 + 4z + 7) + 19. \qquad \blacksquare$$

Schemat Hornera stosujemy również w *deflacji*. Jest to proces usuwania z wielomianu jego czynnika liniowego[4]. Liczba z_0 jest pierwiastkiem wielomianu p wtedy i tylko wtedy, gdy $z - z_0$ jest czynnikiem tego wielomianu. Pozostałe pierwiastki wielomianu p są $n-1$ pierwiastkami ilorazu $p(z)/(z - z_0)$.

Przykład 3.5.8. Dla wielomianu p z przykładu 3.5.4 wykonać deflację, usuwając jego pierwiastek 2.

Rozwiązanie. Tworzymy tablicę liczb jak w poprzednim przykładzie.

	1	−4	7	−5	−2
2		2	−4	6	2
	1	−2	3	1	$\boxed{0}$

Wobec tego

$$z^4 - 4z^3 + 7z^2 - 5z - 2 = (z-2)(z^3 - 2z^2 + 3z + 1).$$

Ostatnia kolumna tablicy jest w istocie zbędna i może co najwyżej służyć do kontroli poprawności obliczeń, które powinny się zakończyć znalezieniem liczby 0. $\blacksquare$

[4] Niech p ma tylko pierwiastki rzeczywiste. Stosując wielokrotnie deflację, obliczamy po jednym pierwiastku wielomianów coraz niższych stopni, co zmniejsza koszt obliczeń. Jest jednak ważne, aby te pierwiastki wyznaczyć w kolejności rosnących wartości bezwzględnych; zob. Dryja, Jankowscy [*1982, s. 111] i prace tam cytowane (*przyp. tłum.*).

Wyznaczywszy wielomian q i resztę $p(z_0)$ (zob. (3.5.2)) za pomocą schematu Hornera, można w ten sam sposób znaleźć iloraz r i resztę $q(z_0)$ z dzielenia q przez $z - z_0$. Iterując te czynności dla wielomianu $p(z)$ postaci (3.5.1), otrzymujemy współczynniki c_k takie, że

$$\begin{aligned} p(z) &= a_n z^n + a_{n-1} z^{n-1} + \ldots + a_0 = \\ &= c_n (z - z_0)^n + c_{n-1} (z - z_0)^{n-1} + \ldots + c_0, \end{aligned}$$

czyli określające rozwinięcie Taylora wielomianu w otoczeniu punktu z_0.

Przykład 3.5.9. Znaleźć rozwinięcie Taylora wielomianu z przykładu 3.5.4 w punkcie $z_0 = 3$.

Rozwiązanie. Obliczenia wyrażamy za pomocą następującej tablicy:

	1	−4	7	−5	−2
3		3	−3	12	21
	1	−1	4	7	19
3		3	6	30	
	1	2	10	37	
3		3	15		
	1	5	25		
3		3			
	1	8			

Liczby w ramkach dają nowe wyrażenie wielomianu p:

$$p(z) = (z-3)^4 + 8(z-3)^3 + 25(z-3)^2 + 37(z-3) + 19. \qquad \blacksquare$$

Algorytm opisany wyżej nazywamy *kompletnym schematem Hornera*. Odpowiedni program działa tak, że współczynniki c_k znajdują się po jego wykonaniu na miejscu danych a_k.

input n, $(a_i\colon\ 0 \leqslant i \leqslant n)$, z_0
for $j = 0$ **to** $n - 1$ **do**
  for $k = n - 1$ **to** j **step** -1 **do**
    $a_k \leftarrow a_k + z_0 a_{k+1}$
  end do
end do
output $(a_i\colon\ 0 \leqslant i \leqslant n)$

Jak wiemy, metodę iteracyjną Newtona dla równania $f(x) = 0$ opisuje wzór

$$z_{k+1} := z_k - \frac{f(z_k)}{f'(z_k)}.$$

Jeśli funkcja f jest wielomianem p, to powyższy program okrojony do wartości $j = 0, 1$ daje $a_0 = p(z_0)$ i $a_1 = p'(z_0)$, czyli wielkości potrzebne w metodzie Newtona. W tym jednak przypadku warto przerobić ten fragment programu, łącząc etapy dla $j = 0$ i $j = 1$. Prócz tego rezygnujemy z zapisywania wyników na miejscu danych, gdyż współczynniki wielomianu będą potrzebne także później. Program daje $\alpha = p(z_0)$ i $\beta = p'(z_0)$ dla wielomianu p określonego jak w (3.5.1) i dla danego z_0:

input n, $(a_i\colon\ 0 \leqslant i \leqslant n)$, z_0
$\alpha \leftarrow a_n$
$\beta \leftarrow 0$
for $k = n - 1$ **to** 0 **step** -1 **do**
    $\beta \leftarrow \alpha + z_0\beta$
    $\alpha \leftarrow a_k + z_0\alpha$
end do
output α, β

Jeśli ten program (bez pierwszego wiersza) wywołujemy instrukcją

$$\texttt{horner}\big(n, (a_i\colon\ 0 \leqslant i \leqslant n), z_0, \alpha, \beta\big),$$

to program wykonujący M kroków metody Newtona dla danego wielomianu i punktu początkowego z_0 może wyglądać tak:

input n, $(a_i\colon\ 0 \leqslant i \leqslant n)$, z_0, M, ε
for $j = 1$ **to** M **do**
    call horner$\big(n, (a_i\colon\ 0 \leqslant i \leqslant n), z_0, \alpha, \beta\big)$
    $z_1 \leftarrow z_0 - \alpha/\beta$
    output α, β, z_1
    if $|z_1 - z_0| < \varepsilon$ **stop**
    $z_0 \leftarrow z_1$
end do

Przykład 3.5.10. Zastosować metodę Newtona do wielomianu z przykł. 3.5.4 dla punktu początkowego $z_0 = 0$.

Rozwiązanie. Dla $z_0 = 0$ podany już algorytm daje wartości $p(0) = -2$ i $p'(0) = -5$. Nową wartością z jest

$$z_1 = z_0 - \frac{p(z_0)}{p'(z_0)} = 0 - \frac{-2}{-5} = -0.4.$$

Wykonanie algorytmu na komputerze z precyzją arytmetyki równą 2^{-24} daje następujące wyniki:

k	$p(z_k)$	$p'(z_k)$	z_k
1	−2.00000	−5.00000	−0.40000
2	1.40160	−12.77600	−0.29029
3	1.46322	−10.17322	−0.27591
4	0.00226	−9.86030	−0.27568
5	0.00000	−9.85537	−0.27568

Ciąg $\{z_k\}$ jest więc szybko zbieżny do pierwiastka −0.27568. ■

Twierdzenie 3.5.11. *Niech x_k i x_{k+1} będą kolejnymi przybliżeniami pierwiastka otrzymanymi metodą Newtona dla wielomianu p stopnia n. Wtedy ten wielomian ma na płaszczyźnie zespolonej pierwiastek odległy od x_k co najwyżej o $n|x_k - x_{k+1}|$.*

Dowód. Niech $r_1, r_2, \ldots, r_n$ będą pierwiastkami wielomianu p. Wtedy $p(z) = c\prod_{j=1}^{n}(z - r_j)$. Poprawka w metodzie Newtona wynosi $-p(z)/p'(z)$. Pochodną wielomianu p wyrażamy w postaci

$$p'(z) = c\sum_{k=1}^{n}\prod_{i=1, i\neq k}^{n}(z - r_i) = \sum_{k=1}^{n} p(z)/(z - r_k) = p(z)\sum_{k=1}^{n}(z - r_k)^{-1}.$$

Chcemy wykazać, że dla dowolnego z (a więc i dla x_k) istnieje wskaźnik j, dla którego $|z - r_j| \leqslant n|p(z)/p'(z)|$. W przeciwnym razie dla każdego j mielibyśmy nierówność $|z - r_j| > n|p(z)/p'(z)|$, a stąd wynikałoby, że

$$|z - r_j|^{-1} < \frac{1}{n}|p'(z)/p(z)| = \frac{1}{n}\left|\sum_{k=1}^{n}(z - r_k)^{-1}\right| \leqslant \frac{1}{n}\sum_{k=1}^{n}|z - r_k|^{-1}.$$

To jest jednak niemożliwe, gdyż średnia n liczb nie może być większa od każdej z nich. ■

Powyższe twierdzenie podał Bodewig [1946].

Metoda Bairstowa

Jeśli wielomian ma tylko współczynniki rzeczywiste, to i tak jego pierwiastki mogą być zespolone. Możemy wtedy obliczać je parami, używając tylko arytmetyki rzeczywistej. Służy do tego *metoda Bairstowa*, która będzie podana nieco dalej.

Dla dowolnej liczby zespolonej $z = x + iy$ *liczbą sprzężoną* jest $\bar{z} = x - iy$. W dalszym ciągu wykorzystamy następujący podstawowy fakt:

Twierdzenie 3.5.12. *Jeśli wielomian p ma współczynniki rzeczywiste, a w jest jego pierwiastkiem nierzeczywistym, to również $\bar{w}$ jest pierwiastkiem wielomianu p, a iloczyn $(z-w)(z-\bar{w})$ jest jego czynnikiem kwadratowym o współczynnikach rzeczywistych.*

Dowód. Niech będzie $p(z) = a_n z^n + a_{n-1}x^{n-1} + \ldots + a_1 z + a_0$, gdzie wszystkie a_k są rzeczywiste. Z założenia wynika, że

$$0 = a_n w^n + a_{n-1} w^{n-1} + \ldots + a_1 w + a_0.$$

Obliczamy sprzężenie obu stron, używając na przemian reguł $\overline{z_1 + z_2} = \bar{z}_1 + \bar{z}_2$ i $\overline{z_1 z_2} = \bar{z}_1 \bar{z}_2$. Ponieważ wszystkie a_k są rzeczywiste, więc wynik ma postać

$$0 = a_n \bar{w}^n + a_{n-1} \bar{w}^{n-1} + \ldots + a_1 \bar{w} + a_0.$$

Dlatego $\bar{w}$ jest pierwiastkiem wielomianu p. Ponieważ w nie jest rzeczywiste, więc w i $\bar{w}$ są różnymi pierwiastkami i p ma czynnik kwadratowy

$$(z-w)(z-\bar{w}) = z^2 - (w+\bar{w})z + w\bar{w},$$

którego współczynniki $-(w+\bar{w})$ i $w\bar{w}$ są rzeczywiste. ■

Jak już wiemy, pierwiastki nierzeczywiste wielomianu rzeczywistego tworzą pary sprzężone, którym odpowiadają jego czynniki kwadratowe rzeczywiste. Warto zatem szukać takich czynników. Można to robić za pomocą metody Newtona.

Twierdzenie 3.5.13. *Dzielenie wielomianu*

$$p(z) := a_n z^n + a_{n-1} z^{n-1} + \ldots + a_0$$

przez wielomian kwadratowy $z^2 - uz - v$ daje iloraz i resztę, równe odpowiednio

$$q(z) := b_n z^{n-2} + b_{n-1} z^{n-3} + \ldots + b_3 z + b_2,$$
$$r(z) := b_1(z-u) + b_0,$$

których współczynniki można obliczać rekurencyjnie według wzorów

$$b_{n+1} = b_{n+2} = 0, \qquad b_k = a_k + u b_{k+1} + v b_{k+2} \qquad (n \geqslant k \geqslant 0).$$

Dowód. Wielomiany p, q i r wiąże relacja

$$p(z) = q(z)(z^2 - uz - v) + r(z),$$

a bardziej konkretnie – związek

$$\sum_{k=0}^{n} a_k z^k = \Big(\sum_{k=2}^{n} b_k z^{k-2}\Big)(z^2 - uz - v) + b_1(z-u) + b_0.$$

Przyrównujemy współczynniki przy z^k obu stron tej równości:

$$\begin{aligned} a_k &= b_k - ub_{k+1} - vb_{k+2} \quad (0 \leqslant k \leqslant n-2),\\ a_{n-1} &= b_{n-1} - ub_n,\\ a_n &= b_n. \end{aligned}$$

Dla b_{n+1} i b_{n+2} równych z definicji 0 pierwsze równanie dla $k = n-1, n$ usuwa potrzebę wyróżniania drugiego i trzeciego. ■

Ograniczmy teraz rozważania do przypadku, w którym wszystkie współczynniki a_k są rzeczywiste. Szukamy czynnika kwadratowego rzeczywistego, czyli w ostatnim twierdzeniu u i v mają być rzeczywiste. Współczynniki b_0 i b_1 obliczane w procesie dzielenia zależą od u i v, piszemy zatem $b_0 = b_0(u,v)$ i $b_1 = b_1(u,v)$. Wielomian q ma być czynnikiem wielomianu p, więc reszta r ma znikać:

$$b_0(u,v) = 0, \qquad b_1(u,v) = 0.$$

W metodzie Bairstowa ten układ dwóch równań rozwiązujemy metodą Newtona. Potrzebne pochodne cząstkowe

$$c_k := \frac{\partial b_k}{\partial u}, \quad d_k := \frac{\partial b_{k-1}}{\partial v} \qquad (0 \leqslant k \leqslant n)$$

otrzymujemy, różniczkując związek rekurencyjny z tw. 3.5.13. Daje to dodatkowe związki:

$$\begin{aligned} c_k &= b_{k+1} + uc_{k+1} + vc_{k+2} \quad (c_{n+1} = c_n = 0),\\ d_k &= b_{k+1} + ud_{k+1} + vd_{k+2} \quad (d_{n+1} = d_n = 0). \end{aligned}$$

Ponieważ są one identyczne, więc wystarczy rozważać pierwszy z nich. Zarys postępowania jest następujący: Przypisujemy wartości początkowe parametrom u i v. Szukamy takich poprawek, oznaczanych δu i δv, które by spełniały równania

$$b_0(u+\delta u, v+\delta v) = b_1(u+\delta u, v+\delta v) = 0.$$

Jak w podrozdz. 3.2, linearyzujemy te równania, co daje

$$b_0(u,v) + \frac{\partial b_0}{\partial u}\delta u + \frac{\partial b_0}{\partial v}\delta v = 0,$$

$$b_1(u,v) + \frac{\partial b_1}{\partial u}\delta u + \frac{\partial b_1}{\partial v}\delta v = 0.$$

Uwzględniając poprzednie uwagi, piszemy ten układ w postaci

$$\begin{bmatrix} c_0 & c_1 \\ c_1 & c_2 \end{bmatrix} \begin{bmatrix} \delta u \\ \delta v \end{bmatrix} = -\begin{bmatrix} b_0(u,v) \\ b_1(u,v) \end{bmatrix}.$$

Rozwiązanie tego układu wyraża się wzorami:

$$\delta u = (c_1 b_1 - c_2 b_0)/J, \quad \delta v = (c_1 b_0 - c_0 b_1)/J, \quad J = c_0 c_2 - c_1^2.$$

J jest tu jakobianem pary funkcji nieliniowych $b_0(u,v)$ i $b_1(u,v)$.

Oto program, który startując z danego punktu (u,v), wykonuje M kroków metody Bairstowa. Jest on zgodny z powyższym opisem.

input $n, (a_i\colon\ 0 \leqslant i \leqslant n), u, v, M$
$b_n \leftarrow a_n$
$c_n \leftarrow 0$
$c_{n-1} \leftarrow a_n$
for $j = 1$ **to** M **do**
  $b_{n-1} \leftarrow a_{n-1} + ub_n$
  for $k = n-2$ **to** 0 **step** -1 **do**
    $b_k \leftarrow a_k + ub_{k+1} + vb_{k+2}$
    $c_k \leftarrow b_{k+1} + uc_{k+1} + vc_{k+2}$
  end do
  $J \leftarrow c_0 c_2 - c_1^2$
  $u \leftarrow u + (c_1 b_1 - c_2 b_0)/J$
  $v \leftarrow v + (c_1 b_0 - c_0 b_1)/J$
  output j, u, v, b_0, b_1
end do

Przykład 3.5.14. Wyznaczyć czynnik kwadratowy rzeczywisty wielomianu z poprzednich przykładów, stosując metodę Bairstowa dla punktu początkowego $(u,v) = (3,-4)$.

Rozwiązanie. W programie komputerowym wzorowanym na algorytmie podanym wyżej użyto arytmetyki podwójnej precyzji. Siódma iteracja dała następujące wyniki:

$$u = 2.27568\,22036\,510, \quad v = -3.62736\,50847\,118,$$
$$b_0 = -0.2_{10}{-14}, \quad b_1 = 0.0.$$

Ponieważ b_0 i b_1 są praktycznie zerami, więc możemy zaakceptować u i v jako przybliżone wartości współczynników czynnika kwadratowego $z^2 - uz - v$.

Łącząc wyniki tego i poprzednich przykładów, otrzymujemy rozkład na czynniki danego wielomianu:

$$z^4 - 4z^3 + 7z^2 - 5z - 2 = (z-2)(z+0.276)(z^2 - 2.28z + 3.63).$$

Współczynniki zaokrąglono tu do trzech cyfr. Używając ich dokładniejszych wartości, można obliczyć dwa pierwiastki zespolone wielomianu p. Są one równe

$$1.13784\,11018\,255 \pm 1.52731\,22508\,866\,i. \qquad \blacksquare$$

Aby uzupełnić analizę metody Bairstowa, musimy ustalić, przy jakich rozsądnych założeniach jakobian J nie znika w szukanym rozwiązaniu.

Twierdzenie 3.5.15. *Jeśli pierwiastki czynnika $z^2 - u_0 z - v_0$ wielomianu p są zarazem pierwiastkami pojedynczymi tego ostatniego, to w punkcie (u_0, v_0) jakobian dla metody Bairstowa jest różny od 0.*

Dowód. W każdym kroku procesu jest

$$p(z) = (z^2 - uz - v)q(z) + b_1(z-u) + b_0.$$

Obliczenie pochodnych cząstkowych względem u i v daje równania

$$0 = -zq(z) + (z^2 - uz - v)\frac{\partial q}{\partial u} - b_1 + \frac{\partial b_1}{\partial u}(z-u) + \frac{\partial b_0}{\partial u},$$

$$0 = -q(z) + (z^2 - uz - v)\frac{\partial q}{\partial v} + \frac{\partial b_1}{\partial v}(z-u) + \frac{\partial b_0}{\partial v}.$$

Niech wielomian $z^2 - u_0 z - v_0$ ma pierwiastki z_1 i z_2. Stąd $p(z_1) = 0$, $p(z_2) = 0$, $b_1 = 0$ i $b_0 = 0$. Przyjmijmy, że w poprzednich równaniach $u = u_0$, $v = v_0$ i $z = z_1$ albo $z = z_2$. Wynikają stąd cztery równania:

$$0 = -z_j q(z_j) + c_1(z_j - u_0) + c_0 \qquad (j = 1, 2),$$

$$0 = -q(z_j) + c_2(z_j - u_0) + c_1 \qquad (j = 1, 2).$$

Skorzystano tu z oznaczeń i rozważań podanych po dowodzie tw. 3.5.13. Wyrażamy te równania w postaci macierzowej:

$$\begin{bmatrix} c_0 & c_1 \\ c_1 & c_2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ z_1 - u_0 & z_2 - u_0 \end{bmatrix} = \begin{bmatrix} z_1 q(z_1) & z_2 q(z_2) \\ q(z_1) & q(z_2) \end{bmatrix}.$$

Aby wykazać, że macierz Jacobiego jest nieosobliwa, wystarczy udowodnić, że taka jest macierz po prawej stronie tego równania. Jej wyznacznik jest równy $(z_1 - z_2)q(z_1)q(z_2)$. Ponieważ z_1 i z_2 są pierwiastkami pojedynczymi wielomianu p, więc nie mogą one być pierwiastkami wielomianu q. Jest też $z_1 \neq z_2$, czyli wspomniany wyznacznik nie znika. $\blacksquare$

Opis metody Bairstowa zaczerpnięto z książki Henriciego [1964].

Metoda Laguerre'a

Przejdźmy teraz do *metody Laguerre'a* obliczania pierwiastków wielomianu p stopnia n. Używa się jej w wielu nowych pakietach oprogramowania, gdyż jest raczej odporna, a jej zbieżność jest sześcienna w otoczeniu każdego pierwiastka pojedynczego. Jest to metoda iteracyjna, a przejście od danego przybliżenia z pierwiastka do nowego przybliżenia następuje według wzorów

$$A := -p'(z)/p(z), \quad B := A^2 - p''(z)/p(z),$$
$$C := n^{-1}[A \pm \sqrt{(n-1)(nB - A^2)}], \quad z_{\text{nowe}} := z + 1/C.$$

Znak w definicji wielkości C wybieramy tak, żeby $|C|$ było jak największe.

Poniższe *twierdzenie Kahana* [1967] jest dla metody Laguerre'a tym, czym dla metody Newtona było tw. 3.5.11.

Twierdzenie 3.5.16. *Jeśli p jest wielomianem stopnia n, z jest dowolną liczbą zespoloną, a C jest określone jak wyżej, to p ma pierwiastek na płaszczyźnie zespolonej odległy od z co najwyżej o $\sqrt{n}/|C|$.*

Dowód. Niech $r_1, r_2, \ldots, r_n$ będą, jak w tw. 3.5.11, pierwiastkami wielomianu p. Możemy założyć, że $p(x) = \prod_{j=1}^{n}(x - r_j)$, tj. że współczynnik wiodący wielomianu jest równy 1, bo wyrażenia dla A, B, C nie zależą od wartości tego współczynnika. Niech będzie $u_j = (r_j - z)^{-1}$. Różniczkowanie p daje znany już wzór

$$p'(z) = \sum_{j=1}^{n} \prod_{k=1,\, k\neq j}^{n} (z - r_k) = \sum_{j=1}^{n} p(z)/(z - r_j) = -p(z) \sum_{j=1}^{n} u_j.$$

Stąd

$$A = -\frac{p'(z)}{p(z)} = \sum_{j=1}^{n} u_j.$$

Różniczkując to wyrażenie i korzystając z definicji B, wyrażamy tę wielkość w postaci

$$B = \frac{-p(z)p''(z) + [p'(z)]^2}{[p(z)]^2} = \sum_{j=1}^{n} (r_j - z)^{-2} = \sum_{j=1}^{n} u_j^2.$$

Z definicji C wynika, że

$$(nC - A)^2 = (n-1)(nB - A^2). \tag{3.5.3}$$

Niech jeszcze będzie $D = (A - C)/(n-1)$, czyli

$$A = (n-1)D + C. \tag{3.5.4}$$

Z (3.5.3) wynika, że

$$n^2C^2 - 2nCA + A^2 = (n-1)nB - nA^2 + A^2,$$
$$(n-1)B = nC^2 - 2CA + A^2. \tag{3.5.5}$$

W ostatniej równości zastępujemy A przez $(n-1)D + C$:

$$\begin{aligned}(n-1)B &= nC^2 + [(n-1)D + C][(n-1)D - C] = \\ &= nC^2 + (n-1)^2D - C^2 = (n-1)C^2 + (n-1)^2D^2,\end{aligned}$$
$$B = C^2 + (n-1)D^2. \tag{3.5.6}$$

Z (3.5.5), dzięki (3.5.4) i (3.5.6), wynika, że

$$\begin{aligned}nB - A^2 &= nC^2 - 2CA + B = \\ &= nC^2 - 2C[(n-1)D + C] + C^2 + (n-1)D^2 = \\ &= (n-1)C^2 - 2(n-1)CD + (n-1)D^2 = (n-1)(C-D)^2.\end{aligned}$$

To równanie wraz z wyrażeniami dla A i B daje relacje

$$\begin{aligned}(n-1)|C-D|^2 &= |nB - A^2| = n^{-1}|n^2B - 2nA^2 + nA^2| = \\ &= n^{-1}\Big|\sum_{j=1}^{n}(n^2u_j^2 - 2nAu_j + A^2)\Big| = \\ &= n^{-1}\Big|\sum_{j=1}^{n}(nu_j - A)^2\Big| \leqslant n^{-1}\sum_{j=1}^{n}|nu_j - A|^2 = \\ &= n^{-1}\sum_{j=1}^{n}[n^2|u_j|^2 - 2n\Re(\bar{A}u_j) + |A|^2] = \\ &= n\sum_{j=1}^{n}|u_j|^2 - 2\bar{A}A + |A|^2 \leqslant n^2\max_j|u_j|^2 - |A|^2 = \\ &= n^2\max_j|u_j|^2 - |C - D + nD|^2 = \\ &= n^2\max_j|u_j|^2 - |C-D|^2 - 2n\Re[\bar{D}(C-D)] - n^2|D|^2.\end{aligned}$$

Dlatego

$$n|C-D|^2 \leqslant n^2\max_j|u_j|^2 - n^2|D|^2 - 2n\Re(\bar{D}C) + 2n|D|^2,$$
$$|C-D|^2 \leqslant n\max_j|u_j|^2 - n|D|^2 - 2\Re(\bar{D}C) + 2|D|^2,$$

$$|C|^2 - 2\Re(C\bar{D}) + |D|^2 \leqslant n \max_j |u_j|^2 - n|D|^2 - 2\Re(\bar{D}C) + 2|D|^2,$$
$$|C|^2 + (n-1)|D|^2 \leqslant n \max |u_j|^2 = n / \min_j |z - r_j|^2,$$

a stąd

$$\min_j |z - r_j|^2 \leqslant n/[|C|^2 + (n-1)|D|^2]$$

i ostatecznie otrzymujemy oszacowanie

$$\min_j |z - r_j| \leqslant \frac{\sqrt{n}}{\sqrt{|C|^2 + (n-1)|D|^2}},$$

lepsze od tego, które podano w twierdzeniu. ■

Powyższy dowód podał Kahan [1967].

Oto prosty program realizujący metodę Laguerre'a dla danego wielomianu i wybranego punktu początkowego z_0:

input $n, (a_i\colon\ 0 \leqslant i \leqslant n), z_0, M, \varepsilon$
for $k = 1$ **to** M **do**
  $\alpha \leftarrow a_n$
  $\beta \leftarrow 0$
  $\gamma \leftarrow 0$
  for $j = n-1$ **to** 0 **step** -1 **do**
    $\gamma \leftarrow z_0\gamma + \beta$
    $\beta \leftarrow z_0\beta + \alpha$
    $\alpha \leftarrow z_0\alpha + a_j$
  end do
  $A \leftarrow -\beta/\alpha$
  $B \leftarrow A^2 - 2\gamma/\alpha$
  $C \leftarrow \left[A \pm \sqrt{(n-1)(nB - A^2)}\right]/n$
  $z_1 \leftarrow z_0 + 1/C$
  output k, z
  if $|z_1 - z_0| < \varepsilon$ **then stop**
  $z_0 \leftarrow z_1$
end do

Wielkościom $p(z_0)$, $p'(z_0)$ i $\frac{1}{2}p''(z_0)$ odpowiadają wyżej α, β i γ.

Lemat 3.5.17. *Niech $v_1, v_2, \ldots, v_n$ będą dowolnymi liczbami rzeczywistymi i niech będzie $\alpha := \sum_{i=1}^n v_i$, $\beta := \sum_{i=1}^n v_i^2$. Wtedy liczby v_j leżą w przedziale domkniętym o końcach*

$$n^{-1}\left[\alpha \pm \sqrt{(n-1)(n\beta - \alpha^2)}\right]. \tag{3.5.7}$$

Dowód. Wystarczy udowodnić, że w określonym wyżej przedziale leży v_1. Posłużymy się przy tym nierównością Cauchy'ego-Schwarza:

$$\Big(\sum_{i=1}^{m} x_i y_i\Big)^2 \leqslant \Big(\sum_{i=1}^{m} x_i^2\Big)\Big(\sum_{j=1}^{m} y_j^2\Big).$$

Wynika z niej, że

$$\begin{aligned}\alpha^2 - 2\alpha v_1 + v_1^2 = (\alpha - v_1)^2 &= (v_2 + v_3 + \ldots + v_n)^2 \leqslant \\ &\leqslant (1^2 + 1^2 + \ldots + 1^2)(v_2^2 + v_3^2 + \ldots + v_n^2) = \\ &= (n-1)(v_2^2 + v_3^2 + \ldots + v_n^2) = (n-1)(\beta - v_1^2),\end{aligned}$$

czyli po oczywistych przekształceniach

$$nv_1^2 - 2\alpha v_1 + \alpha^2 - (n-1)\beta \leqslant 0.$$

To pokazuje, że funkcja kwadratowa $q(x) := nx^2 - 2\alpha x + \alpha^2 - (n-1)\beta$ jest taka, iż $q(v_1) \leqslant 0$. Dla dużych $|x|$ jest oczywiście $q(x) > 0$. Dlatego v_1 leży między dwoma pierwiastkami wielomianu q, a są nimi punkty (3.5.7). ■

Lemat 3.5.18. *Niech p będzie wielomianem rzeczywistym stopnia n, mającym pierwiastki rzeczywiste $r_1, r_2, \ldots, r_n$. Dla dowolnego x rzeczywistego i różnego od wszystkich r_j liczby $(x - r_j)^{-1}$ leżą w przedziale o końcach*

$$[np(x)]^{-1}\Big\{p'(x) \pm \sqrt{[(n-1)p'(x)]^2 - n(n-1)p(x)p''(x)}\Big\}. \tag{3.5.8}$$

Dowód. Jest $p(x) = c\prod_{j=1}^{n}(x - r_j)$. Wyrażenie (3.5.8) nie zależy od c; niech będzie $c = 1$. Jak w dowodzie tw. 3.5.16 mamy, dla $v_j = (x - r_j)^{-1}$, $\alpha = \sum_{j=1}^{n} v_j$ i $\beta = \sum_{j=1}^{n} v_j^2$, relacje

$$p'(x)/p(x) = \sum_{j=1}^{n} v_j = \alpha,$$

$$\{[p'(x)]^2 - p(x)p''(x)\}/[p(x)]^2 = \sum_{j=1}^{n} v_j^2 = \beta.$$

Na mocy lem. 3.5.17 wszystkie liczby v_j leżą w przedziale o końcach (3.5.7). Podstawienie tam wyrażeń otrzymanych dla α i β daje (3.5.8). ■

Twierdzenie 3.5.19. *Jeśli wielomian rzeczywisty p ma wszystkie pierwiastki rzeczywiste, to ciąg generowany za pomocą algorytmu Laguerre'a dla dowolnego punktu początkowego jest zbieżny monotonicznie do jednego z nich.*

Dowód. Niech pierwiastki r_j wielomianu p będą uporządkowane tak, że $r_1 \leqslant r_2 \leqslant \ldots \leqslant r_n$. Niech x nie będzie pierwiastkiem. Na mocy lematu 3.5.18 wszystkie liczby $(x - r_i)^{-1}$ leżą w przedziale o końcach

$$u(x) = [p'(x) + w(x)]/[np(x)], \qquad v(x) = [p'(x) - w(x)]/[np(x)],$$

gdzie

$$w(x) = \sqrt{[(n-1)p'(x)]^2 - n(n-1)p(x)p''(x)}.$$

Rozważmy najpierw przypadek, w którym $r_j < x < r_{j+1}$ dla pewnego j. Przypuśćmy, że $p(y) > 0$ w (r_j, r_{j+1}); dla $p(y) < 0$ rozumowania są podobne. Mamy wtedy

$$v(x) \leqslant (x - r_{j+1})^{-1} < 0 < (x - r_j)^{-1} \leqslant u(x).$$

Stąd wynika, że

$$r_j \leqslant x - \frac{1}{u(x)} < x < x - \frac{1}{v(x)} \leqslant r_{j+1}.$$

Jeśli więc zaczynamy od punktu $x \in (r_j, r_{j+1})$ i, jak w metodzie Laguerre'a, obliczamy $x - 1/u(x)$ i $x - 1/v(x)$, to te dwa nowe punkty leżą w $[r_j, r_{j+1}]$. Oczywiście, jeśli którykolwiek z nich jest końcem tego przedziału, to pierwiastek wielomianu p został już znaleziony. W przeciwnym razie dwa nowe punkty leżą w (r_j, r_{j+1}) i obliczanie pierwiastka można kontynuować. Formalnie rzecz biorąc, określamy dwa ciągi $\{y_k\}$ i $\{z_k\}$ wzorami

$$\begin{aligned} y_0 &= x, \quad y_{k+1} = y_k - 1/u(y_k) \qquad (k \geqslant 1), \\ z_0 &= x, \quad z_{k+1} = z_k - 1/v(z_k) \qquad (k \geqslant 1). \end{aligned}$$

Powyższe rozważania pokazują, że

$$r_j \leqslant y_{k+1} < y_k < \ldots < y_1 < x < z_1 < \ldots < z_k < z_{k+1} \leqslant r_{j+1}.$$

Ponieważ ciąg $\{y_k\}$ maleje i jest ograniczony z dołu przez r_j, więc jest zbieżny: $y_k \searrow y$. Ze wzoru iteracyjnego wynika, że

$$1/u(y_k) = y_k - y_{k+1} \to 0,$$

więc $u(y_k) \to \infty$. Wzór dla u pokazuje, że $p(y_k) \to 0$, czyli $p(y) = 0$ i $y = r_j$. W podobny sposób stwierdzamy, że $z_k \nearrow r_{j+1}$.

Pozostaje rozważyć przypadek, gdy punkt początkowy nie leży między dwoma pierwiastkami; jest więc na przykład $x > r_n$. Przypuśćmy też, że $p(y) > 0$ w (r_n, ∞). Rozumując jak poprzednio, otrzymujemy nierówność

$$0 < (x - r_n)^{-1} \leqslant u(x),$$

stąd zaś wynika, że $r_n \leqslant x - 1/u(x) < x$. Ciąg $\{y_n\}$ określony jak wyżej maleje i jest zbieżny do r_n. ■

Rozważania dotyczące metody Laguerre'a można znaleźć w następujących artykułach i książkach: Bodewig [1946], van der Corput [1946], Durand [1960], Foster [1981], Galeone [1977], Householder [1970], Kahan [1967], Ostrowski [1966], Parlett [1964], Ralston [*1971], Redish [1974], Wilkinson [1965].

Metoda Newtona w dziedzinie zespolonej

Dla wielomianów o współczynnikach zespolonych metodę Newtona trzeba programować w arytmetyce zespolonej. Znalazłszy jeden pierwiastek, należy wykonać deflację, zaprogramowaną w tejże arytmetyce. Metodę Newtona można wtedy zastosować do zredukowanego wielomianu. Te czynności powtarzamy aż do znalezienia wszystkich pierwiastków. Głębsza analiza i doświadczenie dowodzą, że na ogół taka procedura daje zadowalające wyniki, jeśli tylko uwzględnimy dwa zalecenia:

1. Pierwiastki należy obliczać w kolejności rosnących modułów.
2. Każdy pierwiastek zredukowanego wielomianu obliczony metodą Newtona trzeba od razu poprawić stosując tę samą metodę do pierwotnego wielomianu; startuje się przy tym z uzyskanego przybliżenia. Dopiero wtedy wykonujemy następną deflację.

Zainteresowani czytelnicy znajdą w pracach Wilkinsona [1984] oraz Petersa i Wilkinsona [1971] dalsze porady dotyczące tej ogólnej strategii.

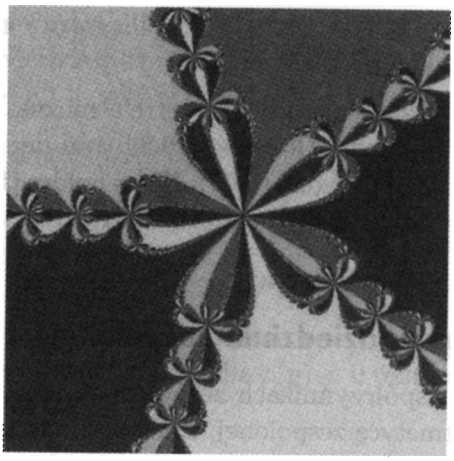
Można by opisać wiele innych metod obliczania pierwiastków wielomianów. Metoda Laguerre'a jest jednak szczególnie atrakcyjna, gdyż jej globalna zbieżność ma korzystne własności. Odporną metodę nadającą się do szerokiego zastosowania opracowali Jenkins i Traub [1970a]. Inne niewymienione jeszcze prace to: Allgower, Glasshoff i Peitgen [1981], Gautschi [1979, 1984], Henrici [1974], Householder [1970], Ostrowski [1966], Jenkins i Traub [1970b], Marden [1966], Smale [1981], Stoer i Bulirsch [1980], Ralston i Rabinowitz [1978] oraz Traub [1964].

Z metodą Newtona w dziedzinie zespolonej wiąże się rys. 3.8. Dodajmy do niego krótki komentarz.

Niech p będzie wielomianem stopnia co najmniej drugiego i niech ξ będzie jednym z jego pierwiastków. Metoda Newtona startująca od punktu z płaszczyzny zespolonej tworzy ciąg określony wzorami

$$z_0 := z, \quad z_{n+1} := z_n - p(z_n)/p'(z_n) \qquad (n \geqslant 0).$$

Jeśli $\lim_{n\to\infty} z_n = \xi$, to mówimy, że punkt początkowy z jest *przyciąganym* przez ξ. Zbiór wszystkich punktów z przyciąganych przez ξ nazywamy



RYS. 3.8. Zbiory przyciągania dla $p(z) = z^5 + 1$

zbiorem przyciągania dla ξ. Każdy pierwiastek wielomianu p ma swój zbiór przyciągania. Te zbiory są parami rozłączne, gdyż ciąg zbieżny do jednego pierwiastka nie może być zbieżny do innego. Pewne liczby zespolone nie należą do żadnego zbioru przyciągania; są to punkty początkowe, dla których metoda Newtona nie jest zbieżna. Te wyjątkowe punkty tworzą *zbiór Julii* wielomianu p, nazwany tak na cześć francuskiego matematyka (Gaston Julia), który w 1918 r. opublikował ważną pracę na ten temat. Jeśli wszystkie pierwiastki wielomianu p są pojedyncze, to zbiory przyciągania są otwarte, a zbiór Julii składa się z brzegów tych zbiorów.

Na rysunku 3.8 pokazano zbiory przyciągania dla wielomianu $p(z) = z^5 + 1$, który ma pierwiastki

$$\omega_k = \cos(\tfrac{2}{5}\pi k) + i\sin(\tfrac{2}{5}\pi k) \qquad (k = 0, 1, 2, 3, 4).$$

Aby otrzymać ten rysunek, wygenerowaliśmy dużą liczbę punktów siatki pokrywającej pewien obszar kwadratowy na płaszczyźnie zespolonej. Dla każdego z nich wykonaliśmy zgrubny test, aby określić, do którego zbioru przyciągania ten punkt należy. Test polegał na obliczaniu do dwudziestu początkowych przybliżeń w metodzie Newtona i sprawdzeniu, czy któreś z nich leży w odległości nie większej od 0.25 od jednego z pierwiastków. Jeśli tak jest, to następne przybliżenia są już zbieżne kwadratowo do tego pierwiastka. Wynika to z ogólnej teorii zbieżności metody Newtona w dziedzinie zespolonej. Powstała w ten sposób lista punktów siatki należących do poszczególnych zbiorów przyciągania. Każdemu z pięciu zbiorów przypisano inny kolor. Te zbiory (ściślej, punkty do nich należące) wyświetlono na kolorowym ekranie i wydrukowano również w kolorach. Te kolorowe zbiory

mają szczególną strukturę *fraktalną*. Polega ona na tym, że na powiększonym fragmencie płaszczyzny zawierającym dwa zbiory widzimy powtórzone ogólne wzorce. Utrzymuje się to przy ponownym powiększaniu. Co więcej, każdy punkt brzegowy dowolnego z pięciu zbiorów jest zarazem punktem brzegowym trzech innych zbiorów!

W ostatnich latach opublikowano wiele artykułów i książek o fraktalach i chaosie, m.in.: Barnsley [1988], Curry, Garnett i Sullivan [1983], Dewdney [1988], Glieck [1987], Kudrewicz [*1993], Mandelbrot [1982], Peitgen, Jürgens i Saupe [*1995-6], Peitgen i Richter [1986], Peitgen, Saupe i Haeseler [1984], Pickover [1988], Sander [1987]. Można tam znaleźć dodatkowe informacje.

ZADANIA 3.5

1. Przyjmując definicję krotności pierwiastka podaną w tekście, udowodnić, że jeśli z jest m-krotnym pierwiastkiem wielomianu p, to

$$p(z) = p'(z) = \ldots = p^{(m-1)}(z) = 0, \qquad p^{(m)}(z) \neq 0.$$

2. (cd.). Udowodnić twierdzenie przeciwne do tego z poprzedniego zadania.

3. Dla wielomianu $p(z) := 3z^5 - 7z^4 - 5z^3 + z^2 - 8z + 2$ znaleźć koła o środku w punkcie 0:

(**a**) zawierające wszystkie pierwiastki,

(**b**) nie zawierające żadnego pierwiastka.

4. Czy tw. 3.5.3 może dać promień najmniejszego koła o środku w punkcie 0, zawierającego wszystkie pierwiastki danego wielomianu?

5. Za pomocą algorytmu Hornera znaleźć $p(4)$ dla wielomianu z zad. 3.

6. Dla wielomianu z zad. 3 znaleźć rozwinięcie Taylora w punkcie $z_0 = 4$.

7. Napisać taki algorytm deflacji wielomianu $p(z)$ gdy jest dany jego pierwiastek z_0, który oblicza współczynniki zredukowanego wielomianu według rosnących potęg zmiennej, tj. począwszy od składnika stałego.

8. Dla wielomianu z zad. 3 i $z_0 = 4$ obliczyć metodą Newtona z_1.

9. Dla wielomianu $p(z) = 9z^4 - 7z^3 + z^2 - 2z + 5$ znaleźć $p(6)$, $p'(6)$ i następne przybliżenie obliczone metodą Newtona startującą z $z = 6$.

10. Udowodnić, że każdy wielomian o współczynnikach rzeczywistych można rozłożyć na iloczyn czynników liniowych lub kwadratowych, mających współczynniki rzeczywiste.

11. Do wielomianu z zad. 3 zastosować metodę Bairstowa, startując z punktu $(u, v) = (3, 1)$. Obliczyć poprawki δu i δv.

12. Sprawdzić związki rekurencyjne i wartości początkowe dla c_k i d_k podane w opisie metody Bairstowa.

13. Czy metoda Bairstowa daje ciąg (u_k, v_k) zbieżny kwadratowo?

14. W opisie metody Laguerre'a wielkości A i B są funkcjami zmiennej z i zależą od danego wielomianu p. Niech r będzie jego pierwiastkiem. Wykazać, że funkcje A i B dla $p(z)/(z-r)$ są równe odpowiednio

$$A+(z-r)^{-1}, \qquad B-(z-r)^2.$$

15. Wykazać, że jeśli wielomian p ma współczynniki i pierwiastki rzeczywiste, to

$$(n-1)[p'(x)]^2 \geqslant np(x)p''(x).$$

Zadania komputerowe 3.5

K1. Napisać program, który dla danych współczynników wielomianu p i danego punktu z_0 oblicza wartości $p(z_0)$, $p'(z_0)$ i $p''(z_0)$. Program ma zawierać tylko jedną pętlę. Sprawdzić go dla wielomianu z zad. 3 i $z_0 = 4$.

K2. Napisać program realizujący metodę Newtona dla wielomianu o współczynnikach zespolonych, danego punktu początkowego z płaszczyzny zespolonej i danej liczby iteracji. Sprawdzić ten program dla wielomianu z zad. 3 i $z_0 = 3-2i$.

K3. Napisać program, który stosując metodę Newtona i deflację, znajduje wszystkie pierwiastki wielomianu. Sprawdzić ten program dla „perfidnego" wielomianu Wilkinsona z zad. 2.3.K5.

K4. Napisać i sprawdzić program obliczający metodą Laguerre'a wszystkie pierwiastki wielomianu zespolonego, w kolejności rosnących modułów. Stosować deflację i tą samą metodą dla pierwotnego wielomianu poprawiać każdy pierwiastek otrzymany z wielomianu zredukowanego. W celu sprawdzenia programu obliczyć wszystkie pierwiastki wielomianu z przykł. 3.5.4 oraz wielomianu

$$x^8 - 36x^7 + 546x^6 - 4536x^5 + 22449x^4 - 67284x^3 + \\ +118124x^2 - 109584x + 40320.$$

Te ostatnie są równe dokładnie $1, 2, \ldots, 8$. Powtórzyć obliczenia po zmianie współczynnika przy x^7 na -37. Zauważyć, że małe zakłócenia współczynników mogą radykalnie zmienić pierwiastki, tj. że pierwiastki są niestabilnymi funkcjami współczynników.

K5. Zmodyfikować metodę Laguerre'a tak, żeby współczynniki ilorazu wielomianów, obliczane za pomocą schematu Hornera, były zapamiętywane. Wtedy nie trzeba oddzielnie obliczać współczynników zredukowanego wielomianu po znalezieniu pierwiastka. Sprawdzić program dla wielomianu

$$z^4 - (10+26i)z^3 - (216-190i)z^2 + (1140+636i)z - (72-68i).$$

K6. Stosując metodę Newtona do wielomianu $p(z) = z^3 - 1$, znaleźć trzy bliskie punkty początkowe (rozmieszczone co 0.01) takie, żeby otrzymane ciągi przybliżeń były zbieżne do różnych pierwiastków. Łącząc odcinkami kolejne przybliżenia, pokazać, na przykład na ekranie monitora, jaką drogę przebywają te przybliżenia.

K7. Napisać program, który metodą Newtona w dziedzinie zespolonej oblicza zbiory przyciągania dla pierwiastków wielomianu $f(z) = z^8+1$ tak, jak to opisano w tym podrozdziale. Wyświetlać wyniki na ekranie monitora.

3.6. Metody homotopii i kontynuacji

W tym podrozdziale rozważamy problem szukania pierwiastków równania

$$f(x) = 0, \tag{3.6.1}$$

gdzie f jest odwzorowaniem pewnej przestrzeni liniowej na inną taką przestrzeń: $f\colon X \to Y$. Problem jest na tyle ogólny, że obejmuje układy równań algebraicznych, równania różniczkowe i całkowe itd. Metody tu rozważane opierają się na innej strategii niż te z poprzednich podrozdziałów; to nowe podejście wymaga rozwiązywania numerycznego układów równań różniczkowych zwyczajnych, co w tej książce jest tematem podrozdz. 8.6. Jedno z zastosowań metody odnosi się do programowania liniowego; ono też jest omawiane dalej, w podrozdz. 10.3.

Podstawowe pojęcia

Głównym pomysłem *metody kontynuacji* jest potraktowanie danego zadania jako szczególnego przypadku jednoparametrowej rodziny zadań, z parametrem t przebiegającym przedział $[0,1]$. Zadanie dane ma odpowiadać wartości $t = 1$, a inne zadanie, dla $t = 0$, powinno mieć znane rozwiązanie. Jeśli na przykład pewne równanie $g(x) = 0$ ma znane rozwiązanie, to możemy przyjąć, że

$$h(t,x) := tf(x) + (1-t)g(x). \tag{3.6.2}$$

Następnym etapem jest wybór punktów $t_0, t_1, \ldots, t_m$ takich, że

$$0 = t_0 < t_1 < t_2 < \ldots < t_m = 1.$$

Próbujemy rozwiązać każde z równań $h(t_i, x) = 0$ dla $0 < i \leqslant m$. Jeśli stosujemy przy tym jakąś metodę iteracyjną (np. metodę Newtona), to jest rozsądne traktować rozwiązanie dla i-tego kroku jako przybliżenie początkowe w obliczaniu rozwiązania w $(i+1)$-szym kroku. Takie postępowanie można uważać za remedium na kłopot, z którym mamy do czynienia stosując metodę Newtona, a mianowicie na konieczność posiadania dobrego punktu początkowego.

Równość (3.6.2), dzięki której pierwotne zadanie (3.6.1) należy do szerszej rodziny zadań, jest przykładem *homotopii* wiążącej dwie funkcje f i g. Ogólniej, homotopią może być dowolny związek ciągły f z g. Formalnie rzecz biorąc, homotopią wiążącą dwie funkcje $f, g\colon X \to Y$, jest odwzorowanie ciągłe

$$h\colon [0,1] \times X \to Y$$

takie, że $h(0,x) = g(x)$ i $h(1,x) = f(x)$. Jeśli takie odwzorowanie istnieje, to mówimy, że funkcje f i g są *homotopijne*. Jest to relacja równoważności dla odwzorowań ciągłych z X do Y, gdzie X i Y mogą być dowolnymi przestrzeniami topologicznymi.

Prosta, często używana w metodzie kontynuacji homotopia jest opisana wzorem

$$h(t,x) := tf(x) + (1-t)[f(x) - f(x_0)] = f(x) + (t-1)f(x_0). \quad (3.6.3)$$

x_0 jest tu dowolnym punktem przestrzeni X; jest oczywiste, że x_0 jest rozwiązaniem zadania dla $t = 0$.

Jeśli równanie $h(t,x) = 0$ ma jedyny pierwiastek dla każdego $t \in [0,1]$, to jest on funkcją parametru t i możemy określić $x(t)$ jako jedyny element przestrzeni X, dla którego $h(t,x(t)) = 0$. Zbiór

$$\{x(t)\colon\ 0 \leqslant t \leqslant 1\} \quad (3.6.4)$$

można uważać za łuk krzywej w X, sparametryzowany za pomocą t. Ten łuk prowadzi od znanego punktu $x(0)$ do rozwiązania $x(1)$ naszego zadania. Metodą kontynuacji próbujemy wyznaczyć tę krzywą, obliczając punkty $x(t_0), x(t_1), \ldots, x(t_m)$ leżące na niej.

Jeśli funkcja $t \mapsto x(t)$ jest różniczkowalna i jeśli h jest różniczkowalna, to twierdzenie o funkcji uwikłanej pozwala nam obliczyć $x'(t)$. Idąc tym tropem, możemy opisać krzywą (3.6.4) za pomocą równania różniczkowego. Dla dowolnej homotopii mamy $0 = h(t,x(t))$. Różniczkując obie strony względem t, otrzymujemy

$$0 = h_t(t,x(t)) + h_x(t,x(t))x'(t),$$

gdzie wskaźniki oznaczają pochodne cząstkowe. Stąd

$$x'(t) = -[h_x(t,x(t))]^{-1} h_t(t,x(t)). \quad (3.6.5)$$

Jest to równanie różniczkowe dla x. Wartość początkowa jest z założenia znana. Całkując to równanie (zwykle numerycznie), dochodzimy do szukanego rozwiązania $x(1)$.

Przykład 3.6.1. Niech będzie $X = Y = \mathbb{R}^2$. Przyjmujemy, że

$$f(x) = \begin{bmatrix} \xi_1^2 - 3\xi_2^2 + 3 \\ \xi_1\xi_2 + 6 \end{bmatrix}, \qquad x = (\xi_1, \xi_2) \in \mathbb{R}^2.$$

Rozwiązanie. Właściwa tu homotopia jest dana wzorem (3.6.3), gdzie przyjmujemy, że $x_0 = (1,1)$. Pochodne występujące po prawej stronie równania (3.6.5), są następujące:

$$h_x = f'(x) = \begin{bmatrix} \partial f_1/\partial \xi_1 & \partial f_1/\partial \xi_2 \\ \partial f_2/\partial \xi_1 & \partial f_2/\partial \xi_2 \end{bmatrix} = \begin{bmatrix} 2\xi_1 & -6\xi_2 \\ \xi_2 & \xi_1 \end{bmatrix},$$

$$h_t = f(x_0) = \begin{bmatrix} f_1(x_0) \\ f_2(x_0) \end{bmatrix} = \begin{bmatrix} 1 \\ 7 \end{bmatrix}.$$

Odwrotnością $f'(x)$ jest

$$h_x^{-1} = [f'(x)]^{-1} = \frac{1}{\Delta} \begin{bmatrix} \xi_1 & 6\xi_2 \\ -\xi_2 & 2\xi_1 \end{bmatrix}, \quad \text{gdzie} \quad \Delta = 2\xi_1^2 + 6\xi_2^2.$$

Ścieżka prowadząca z punktu x_0, jest określona równaniem różniczkowym (3.6.5). Ściślej, jest to para równań różniczkowych zwyczajnych:

$$\begin{bmatrix} \xi_1' \\ \xi_2' \end{bmatrix} = -\frac{1}{\Delta} \begin{bmatrix} \xi_1 & 6\xi_2 \\ -\xi_2 & 2\xi_1 \end{bmatrix} \begin{bmatrix} 1 \\ 7 \end{bmatrix} = -\frac{1}{\Delta} \begin{bmatrix} \xi_1 + 42\xi_2 \\ 14\xi_1 - \xi_2 \end{bmatrix}.$$

Ten układ można rozwiązać numerycznie (stosując jedną z metod opisanych w rozdz. 8) w przedziale $0 \leqslant t \leqslant 1$; dla $t = 1$ daje to punkt $(-2.961, 1.978)$. Zauważmy, że f ma pierwiastek $(-3, 2)$.

Aby zakończyć obliczenia, możemy teraz użyć metody Newtona startującej z punktu otrzymanego metodą homotopii. Metoda Newtona zmienia przybliżenie x pierwiastka na $x - \delta$, gdzie poprawka δ wyraża się wzorem

$$\delta = [f'(x)]^{-1} f(x).$$

W tym przykładzie wektor δ jest taki, że

$$\begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} \xi_1 & 6\xi_2 \\ -\xi_2 & 2\xi_1 \end{bmatrix} \begin{bmatrix} \xi_1^2 - 3\xi_2^2 + 3 \\ \xi_1\xi_2 + 6 \end{bmatrix}.$$

Trzy kroki metody Newtona dają następujące wyniki:

k	ξ_1	ξ_2
0	−2.96100 00000 00	1.97800 00000 00
1	−3.00025 32813 14	2.00032 02744 78
2	−3.00000 00057 80	2.00000 00378 24
3	−3.00000 00000 00	2.00000 00000 00

■

Poniższe twierdzenie Ortegi i Rheinboldta [1970] podaje warunki zapewniające skuteczność metody homotopii.

TWIERDZENIE 3.6.2. *Jeśli funkcja* $f\colon \mathbb{R}^n \to \mathbb{R}^n$ *jest różniczkowalna w sposób ciągły i jeśli* $\|[f'(x)]^{-1}\| \leqslant M$ *na* $\mathbb{R}^n$, *to dla każdego* $x_0 \in \mathbb{R}^n$ *istnieje jedyna krzywa* $\{x(t)\colon 0 \leqslant t \leqslant 1\}$ *w* $\mathbb{R}^n$ *taka, że* $f(x(t)) + (t-1)f(x_0) = 0$ *dla* $0 \leqslant t \leqslant 1$. *Funkcja* $t \mapsto x(t)$ *jest rozwiązaniem różniczkowalnym w sposób ciągły zagadnienia początkowego* $x' = -[f'(x)]^{-1}f(x)$, $x(0) = x_0$.

Wykreślanie ścieżki

Inny sposób wykreślania ścieżki $x(t)$ opisali Garcia i Zangwill [1981]. Zaczynamy od równania $h(t,x) = 0$, zakładając, że $x \in \mathbb{R}^n$ i $t \in [0,1]$. Wektor $y \in \mathbb{R}^{n+1}$ jest określony jako

$$y = (t, \xi_1, \xi_2, \ldots, \xi_n),$$

gdzie $\xi_1, \xi_2, \ldots, \xi_n$ są składowymi wektora x. Dlatego nasze równanie przybiera prostszą postać $h(y) = 0$. Każda składowa wektora y, w tym t, może być teraz funkcją zmiennej niezależnej s, piszemy więc $h(y(s)) = 0$. Różniczkując to równanie względem s, otrzymujemy podstawowe równanie różniczkowe

$$h'(y(s))y'(s) = 0. \tag{3.6.6}$$

s zmienia się od 0, jak t. Wartością początkową dla x jest $x(0) = x_0$. Tak więc dla równania różniczkowego (3.6.6) są dostępne odpowiednie wartości początkowe.

Ponieważ f i g są odwzorowaniami $\mathbb{R}^n$ w $\mathbb{R}^n$, więc h jest odwzorowaniem $\mathbb{R}^{n+1}$ w $\mathbb{R}^n$. Dlatego pochodna $h'(y)$ jest reprezentowana przez macierz A o rozmiarach $n \times (n+1)$. Wektor $y(s)$ ma $n+1$ składowych, które oznaczamy teraz symbolami $\eta_1, \eta_2, \ldots, \eta_{n+1}$. Powołując się na lem. 3.6.4, możemy znaleźć inną postać równania (3.6.6):

$$\eta_j' = (-1)^{j+1} \det A_j \qquad (1 \leqslant j \leqslant n+1), \tag{3.6.7}$$

gdzie A_j jest macierzą kwadratową stopnia n, wynikającą z A przez wykreślenie j-tej kolumny; symbol det oznacza wyznacznik macierzy. Zilustrujmy ten formalizm za pomocą zadania z przykł. 3.6.1.

PRZYKŁAD 3.6.3. Dla f i x_0 jak w poprzednim przykładzie mamy

$$h(t,x) = \begin{bmatrix} \xi_1^2 - 3\xi_2^2 + 2 + t \\ \xi_1\xi_2 - 1 + 7t \end{bmatrix}.$$

Rozwiązanie. Równanie różniczkowe (3.6.6) ma teraz postać

$$\begin{bmatrix} \partial h_1/\partial t & \partial h_1/\partial \xi_1 & \partial h_1/\partial \xi_2 \\ \partial h_2/\partial t & \partial h_2/\partial \xi_1 & \partial h_2/\partial \xi_2 \end{bmatrix} \begin{bmatrix} t' \\ \xi_1' \\ \xi_2' \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

czyli

$$\begin{bmatrix} 1 & 2\xi_1 & -6\xi_2 \\ 7 & \xi_2 & \xi_1 \end{bmatrix} \begin{bmatrix} t' \\ \xi_1' \\ \xi_2' \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

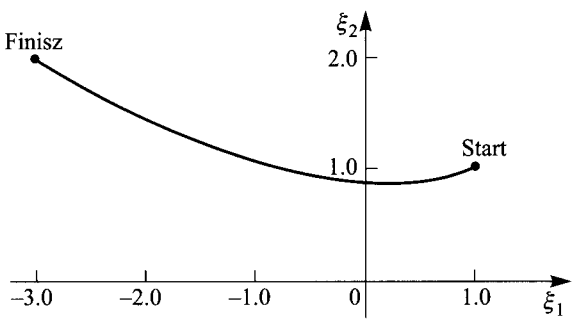
Jest jednak wygodniej zastosować związek (3.6.7) i wyrazić równania różniczkowe w postaci

$$\begin{aligned} t' &= 2\xi_1^2 + 6\xi_2^2, & t(0) &= 0, \\ \xi_1' &= -\xi_1 - 42\xi_2, & \xi_1(0) &= 1, \\ \xi_2' &= -14\xi_1 + \xi_2, & \xi_2(0) &= 1. \end{aligned}$$

Pochodne w tym układzie oblicza się względem s. Całkowanie numeryczne daje następujące dwa punkty:

$$\begin{aligned} s &= 0.087, & t &= 0.969, & \xi_1 &= -2.94, & \xi_2 &= 1.97, \\ s &= 0.088, & t &= 1.010, & \xi_1 &= -3.02, & \xi_2 &= 2.01. \end{aligned}$$

Każdego z nich można użyć jako punktu początkowego w metodzie Newtona, jak to zrobiono w poprzednim przykładzie. Ścieżkę generowaną przez tę homotopię pokazuje rys. 3.9. ■



RYS. 3.9. Krzywa generowana w przykł. 3.6.3

Wadą metody zastosowanej w powyższym przykładzie jest to, że a priori nie znamy wartości s odpowiadającej wartości $t = 1$. W praktyce może to wymagać wielu prób obliczeń.

LEMAT 3.6.4. *Niech A będzie macierzą $n \times (n+1)$. Rozwiązanie równania jednorodnego $Ax = 0$ jest dane wzorem $x_j = (-1)^j \det A_j$, gdzie A_j jest macierzą A bez j-tej kolumny.*

Dowód. Wybierzmy dowolny, na przykład i-ty, wiersz macierzy A i dołączmy jego kopię nad A. Daje to macierz kwadratową B stopnia $n+1$, oczywiście osobliwą, skoro jej dwa wiersze są identyczne. Rozwijając wyznacznik macierzy B względem górnego wiersza, otrzymujemy

$$0 = \det B = \sum_{j=1}^{n+1} (-1)^{j+1} a_{ij} \det A_j = -\sum_{j=1}^{n+1} a_{ij} x_j.$$

Ponieważ jest tak dla $i = 1, 2, \ldots, n$, więc $Ax = 0$. ■

Związki z metodą Newtona

Związki między metodami homotopii i Newtona są ściślejsze niż może się wydawać na pierwszy rzut oka. Zacznijmy od homotopii

$$h(t, x) = f(x) - e^{-t} f(x_0).$$

Parametr t zmienia się tu od 0 do ∞. Szukamy krzywej (ścieżki) $x = x(t)$, na której

$$0 = h(t, x(t)) = f(x(t)) - e^{-t} f(x_0).$$

Jak zwykle, różniczkowanie względem t prowadzi do równania różniczkowego opisującego ścieżkę:

$$\begin{aligned} 0 &= f'(x(t))x'(t) + e^{-t} f(x_0) = f'(x(t))x'(t) + f(x(t)), \\ x'(t) &= -[f'(x(t))]^{-1} f(x(t)). \end{aligned} \tag{3.6.8}$$

Rozwiązując to równanie różniczkowe metodą Eulera (opisaną w podrozdz. 8.2) z krokiem 1, otrzymujemy wzór

$$x_{n+1} = x_n - [f'(x_n)]^{-1} f(x_n),$$

który oczywiście określa metodę Newtona. Rzecz jasna, można oczekiwać, że rozwiązując równanie (3.6.7) bardziej wyrafinowaną metodą numeryczną i ze zmiennym krokiem, otrzymamy lepsze wyniki (te tematy są omawiane w rozdz. 8).

Programowanie liniowe

Metodę homotopii można zastosować także w zadaniach programowania liniowego (rozważanych w podrozdz. 10.3). Takie podejście prowadzi w naturalny sposób do algorytmu Karmarkara [1984]. Komentując w tym kontekście metodę homotopii, wzorujemy się ściśle na opisie, który podali Brophy i Smith [1988]. Czytelnik, który zechce zbadać te pomysły dokładniej, może z pożytkiem wykorzystać inne publikacje.

Rozważmy standardowe zadanie programowania liniowego:
znaleźć maksimum $c^{\mathsf{T}}x$
przy ograniczeniach $Ax = b$ i $x \geqslant 0$.

Wyżej jest $c \in \mathbb{R}^n$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, a A jest macierzą $m \times n$. Zaczynamy od jakiegokolwiek *punktu dopuszczalnego* $x^{(0)}$, spełniającego z definicji podane wyżej ograniczenia. *Zbiór dopuszczalny* jest określony wzorem

$$\mathcal{F} = \{x \in \mathbb{R}^n\colon\ Ax = b \quad \text{i} \quad x \geqslant 0\}.$$

Zamierzamy przesuwać się od $x^{(0)}$ kolejno do innych punktów, pozostając zawsze w $\mathcal{F}$ i zwiększając *funkcję celu* $c^{\mathsf{T}}x$. Jest oczywiste, że jeśli przesuwamy się z $x^{(0)}$ do $x^{(1)}$, to różnica $x^{(1)} - x^{(0)}$ musi leżeć w jądrze macierzy A (zob. przypis na s. 141), tj. ma być $A(x^{(1)} - x^{(0)}) = 0$. Powinniśmy szukać krzywej $t \mapsto x(t)$ w zbiorze dopuszczalnym, startując z $x^{(0)}$ i zmierzając do rozwiązania zadania na ekstremum. Ma przy tym być

1. $x(t) \geqslant 0$ dla $t \geqslant 0$.
2. $Ax(t) = b$ dla $t \geqslant 0$.
3. $c^{\mathsf{T}}x(t)$ rośnie dla $t \geqslant 0$.

Krzywą określimy za pomocą zagadnienia początkowego

$$x' = f(x), \qquad x(0) = x^{(0)}. \tag{3.6.9}$$

Zadaniem, jakie przed nami stoi, jest wyznaczenie właściwego f. Aby spełnić warunek **1**, postępujemy tak, żeby wówczas gdy jakaś składowa x_i zbliża się do 0, prędkość $x_i'(t)$ jej zmiany także dążyła do 0. Można to osiągnąć przyjmując, że dla macierzy przekątniowej

$$D(x) = \operatorname{diag}(x_1, x_2, \ldots, x_n)$$

i dla pewnej funkcji ograniczonej G jest

$$f(x) = D(x)G(x). \tag{3.6.10}$$

W takim przypadku z (3.6.9) i (3.6.10) wynika, że

$$x_i' = x_i G_i(x)$$

i oczywiście $x_i' \to 0$, jeśli $x_i \to 0$.

Aby spełnić warunek **2**, wystarczy zażądać, żeby było $Ax' = 0$. Ponieważ $x' = f = DG$, więc wymagamy, żeby było $ADG = 0$. Najwygodniej jest to osiągnąć, przyjmując, że $G = PH$, gdzie H jest dowolną funkcją, a P jest rzutem ortogonalnym na jądro macierzy AD.

Wreszcie, aby spełnić warunek **3**, powinniśmy wybrać H tak, aby $c^{\mathsf{T}}x(t)$ rosło. Chcemy zatem, żeby było

$$0 < \frac{d}{dt}\left(c^{\mathsf{T}}x(t)\right) = c^{\mathsf{T}}x' = c^{\mathsf{T}}f(x) = c^{\mathsf{T}}DG = c^{\mathsf{T}}DPH.$$

Właściwą postacią H jest Dc, gdyż wtedy, dla $v = Dc$, jest

$$c^{\mathsf{T}}DPH = c^{\mathsf{T}}DPDc = v^{\mathsf{T}}Pv = \langle v, Pv\rangle = \\ = \langle v - Pv + Pv, Pv\rangle = \langle Pv, Pv\rangle \geqslant 0.$$

Ostateczna wersja naszego zagadnienia początkowego jest następująca:

$$x' = D(x)P(x)D(x)c, \qquad x(0) = x^{(0)}. \tag{3.6.11}$$

Teoretycznym wyrażeniem dla P jest

$$P = I - (AD)^{\mathsf{T}}[(AD)(AD)^{\mathsf{T}}]^{-1}AD. \tag{3.6.12}$$

Jest to sensowne, jeśli macierz $B = AD$ ma pełny rząd, tak że BB^{T} jest nieosobliwa. Stąd znów wynika żądanie $x_i > 0$ dla każdej składowej. Tak więc punkty $x(t)$ powinny pozostawać we wnętrzu zbioru $\{x\colon x \geqslant 0\}$. W szczególności tak trzeba wybrać $x^{(0)}$. W praktyce Pv obliczamy nie za pomocą (3.6.12), ale rozwiązując równanie $BB^{\mathsf{T}}z = Bv$ i zauważając, że

$$Pv = v - B^{\mathsf{T}}z.$$

Zagadnienie początkowe (3.6.11) nie musi być rozwiązywane bardzo dokładnie. Można użyć w tym celu wariantu metody Eulera. Przypomnijmy, że dla zadania (3.6.9) metoda Eulera przesuwa rozwiązanie zgodnie ze wzorem

$$x(t + \delta) = x(t) + \delta x'(t) = x(t) + \delta f(x).$$

Korzystając ze wzoru tego typu, generujemy ciąg wektorów $x^{(0)}, x^{(1)}, \ldots$ za pomocą wzoru

$$x^{(k+1)} = x^{(k)} + \delta_k f(x^{(k)}).$$

Chciałoby się wybrać możliwie duże δ_k, z zastrzeżeniem, że $x^{(k+1)} \in \mathcal{F}$, ale dałoby to punkt $x^{(k+1)}$ mający co najmniej jedną składową zerową. Jak wspomniano wcześniej, powodowałoby to inne trudności. Wydaje się, że w praktyce jest sensowne wybierać δ_k jako 9/10 maksymalnego możliwego kroku. Ten ostatni można łatwo obliczyć; jest to maksymalne δ, dla którego $x^{(k+1)} \geqslant 0$. (Ograniczenie $Ax = b$ jest automatycznie spełnione).

ZADANIA 3.6

1. Metodą homotopii użytą w przykł. 3.6.3 rozwiązać układ równań

$$x - 2y + y^2 + y^3 - 4 = -x - y + 2y^2 - 1 = 0,$$

startując z punktu $(0, 0)$. (Wszystkie obliczenia można wykonać nie odwołując się do metod numerycznych).

2. Rozważyć homotopię $h(t, x) = tf(x) + (1 - t)g(x)$, gdzie

$$f(x) = x^2 - 5x + 6, \qquad g(x) = x^2 - 1.$$

Wykazać, że wtedy nie ma ścieżki prowadzącej od zera funkcji g do zera funkcji f.

3. Niech $y = y(s)$ będzie funkcją różniczkowalną z $\mathbb{R}$ do $\mathbb{R}^n$, spełniającą równanie różniczkowe (3.6.6). Zakładając, że $h(y(0)) = 0$, wykazać równość $h(y(s)) = 0$.

4. Jaki układ równań różniczkowych określa ścieżkę, gdy metodę homotopii z przykł. 3.6.3 stosujemy do układu

$$\sin x + \cos y + e^{xy} = \operatorname{arctg}(x + y) - xy = 0$$

i punktu początkowego $(0, 0)$? Program komputerowy wyznaczania rozwiązania może być pouczający.

5. Wykazać, że homotopia jest relacją równoważności dla odwzorowań ciągłych jednej przestrzeni topologicznej na inną.

6. Czy funkcje $f(x) = \sin x$ i $g(x) = \cos x$ są homotopijne?

7. Czy odwzorowania $f(x) = 0$, $g(x) = 2$ przedziału $[0, 1]$ w zbiór $[0, 1] \cup [2, 3]$ są homotopijne?

ROZDZIAŁ 4

Rozwiązywanie układów równań liniowych

4.0. Wstęp

Tematem tego rozdziału są algorytmy ogólnego użytku (m.in. iteracyjne) służące do rozwiązywania układów równań liniowych. Zbadamy też błędy wyników obliczeń komputerowych oraz sposoby kontrolowania i redukcji takich błędów.

Głównym celem rozdziału jest rozważanie aspektów numerycznych rozwiązywania układów równań liniowych postaci

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1, \\ &a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n. \end{aligned}$$

Jest to układ n równań z n niewiadomymi $x_1, x_2, \ldots, x_n$. Wielkości a_{ij} i b_i są danymi liczbami rzeczywistymi; rozwiązanie $x_1, x_2, \ldots, x_n$ jest też rzeczywiste.

Wygodnym narzędziem upraszczającym opis układów równań są macierze. Powyższy układ można wyrazić w postaci

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{bmatrix},$$

a po oznaczeniu odpowiednich macierzy symbolami A, x i b – w postaci

$$Ax = b.$$

4.1. Algebra macierzy

Zapoznamy się teraz z podstawowymi pojęciami teorii macierzy. Uzupełniający materiał znajdzie się w następnych podrozdziałach, tam gdzie to będzie potrzebne[1]. Większości czytelników wystarczy zapewne pobieżne przejrzenie tego podrozdziału.

Podstawowe pojęcia

Macierz jest prostokątną tablicą liczb, np. jedną z następujących:

$$\begin{bmatrix} 3.0 & 1.1 & -0.12 \\ 6.2 & 0.0 & 0.15 \\ 0.6 & -4.0 & 1.3 \\ 9.3 & 2.1 & 8.2 \end{bmatrix}, \quad \begin{bmatrix} 3 & 6 & \frac{11}{7} & -17 \end{bmatrix}, \quad \begin{bmatrix} 3.2 \\ -4.7 \\ 0.11 \end{bmatrix}.$$

Są to odpowiednio macierze: 4×3, 1×4 i 3×1. Określamy w ten sposób *rozmiar* macierzy: na pierwszym miejscu liczbę jej *wierszy* (linii poziomych), a na drugim – liczbę *kolumn* (linii pionowych). Macierz $1 \times n$ nazywamy także *wektorem wierszowym*, a macierz $m \times 1$ – *wektorem kolumnowym* albo po prostu *wektorem*[2]. Macierz $n \times n$ nazywamy macierzą *kwadratową*, a liczbę n – jej *stopniem* (jest on zatem określony tylko dla takich właśnie macierzy).

[1] W tym rozdziale, poza fragmentem podrozdz. 4.6, autorzy ograniczają się do dziedziny rzeczywistej, jednak większość definicji pozostaje poprawna w dziedzinie zespolonej, której dotyczy rozdz. 5 (*przyp. tłum.*).

[2] Wektor kolumnowy o składowych $x_1, x_2, \dots, x_n$, czyli $[x_1\ x_2\ \dots\ x_n]^T$, będzie – tradycyjnie – oznaczany prościej: $(x_1, x_2, \dots, x_n)$. Te właśnie wektory są elementami przestrzeni $\mathbb{R}^n$ lub, w przypadku zespolonym, $\mathbb{C}^n$ (*przyp. tłum.*).

Jeśli A jest macierzą, to jej element znajdujący się na przecięciu i-tego wiersza i j-tej kolumny oznaczamy symbolami a_{ij} lub $(A)_{ij}$. Jeśli więc A jest pierwszą z trzech powyższych macierzy, to $(A)_{32} = -4.0$. Macierz o elementach a_{ij} jest często oznaczana symbolem (a_{ij}).

Dla danej macierzy A macierz *transponowana*, oznaczana symbolem A^{T}, jest z definicji taka, że $(A^{\mathsf{T}})_{ij} = a_{ji}$. Tak więc w tym samym przykładzie jest

$$A^{\mathsf{T}} = \begin{bmatrix} 3.0 & 6.2 & 0.6 & 9.3 \\ 1.1 & 0.0 & -4.0 & 2.1 \\ -0.12 & 0.15 & 1.3 & 8.2 \end{bmatrix}.$$

Macierz A taką, że $A = A^{\mathsf{T}}$, nazywamy macierzą *symetryczną*; oczywiście jest ona kwadratowa.

Macierz stopnia n

$$I := \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

nazywamy macierzą *jednostkową*. Ma ona własność $IA = A = AI$ dla dowolnej macierzy A stopnia n[3)].

Jeśli A jest macierzą, a λ skalarem (czyli tu: liczbą rzeczywistą), to macierz λA jest zdefiniowana wzorem $(\lambda A)_{ij} := \lambda a_{ij}$. Oczywiście, $-A$ znaczy to samo, co $(-1)A$. Jeśli $A = (a_{ij})$ i $B = (b_{ij})$ są macierzami $m \times n$, to ich *suma* $A + B$ jest określona wzorem $(A + B)_{ij} := a_{ij} + b_{ij}$. Jeśli A jest macierzą $m \times p$, a B jest macierzą $p \times n$, to ich *iloczyn* AB jest macierzą $m \times n$ o elementach

$$(AB)_{ij} := \sum_{k=1}^{p} a_{ik} b_{kj} \qquad (1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant n)$$

(trzeba pamiętać, że mnożenie macierzy – w przeciwieństwie do mnożenia liczb – nie jest przemienne, tj. zwykle $AB \neq BA$, a nawet istnienie jednego z tych iloczynów nie implikuje istnienia drugiego). Oto kilka przykładów

[3)] Ogólniej, macierz $A = (a_{ij})$ (kwadratowa lub nie) jest *przekątniowa*, jeśli $a_{ij} = 0$ dla $i \neq j$. Macierz przekątniową kwadratową o elementach a_{ii} na jej *głównej przekątnej* oznacza się symbolem $\mathrm{diag}(a_{ii})$. Macierz *trójprzekątniowa* jest taką macierzą kwadratową, że $a_{ij} = 0$, jeśli tylko $|i - j| > 1$. Macierz *trójkątna górna* (*dolna*) jest taka, że $a_{ij} = 0$ dla $i > j$ (odpowiednio, dla $i < j$) (*przyp. tłum.*).

tych działań na macierzach:

$$3\begin{bmatrix}1 & 3\\ 2 & -1\\ 4 & -4\end{bmatrix} = \begin{bmatrix}3 & 9\\ 6 & -3\\ 12 & -12\end{bmatrix},$$

$$\begin{bmatrix}1 & 3\\ 2 & -1\\ 4 & -4\end{bmatrix} + \begin{bmatrix}6 & 0\\ 3 & -7\\ 8 & 2\end{bmatrix} = \begin{bmatrix}7 & 3\\ 5 & -8\\ 12 & -2\end{bmatrix},$$

$$\begin{bmatrix}2 & 1 & 3\\ 1 & 5 & -6\\ 2 & 1 & 5\\ 0 & 1 & -2\end{bmatrix}\begin{bmatrix}1 & 0\\ -5 & 4\\ 0 & 3\end{bmatrix} = \begin{bmatrix}-3 & 13\\ -24 & 2\\ -3 & 19\\ -5 & -2\end{bmatrix}.$$

Dla układów równań liniowych ważnym pojęciem jest ich *równoważność*. Niech będą dane dwa układy n równań z n niewiadomymi:

$$Ax = b, \qquad Bx = d.$$

Takie układy są *równoważne*, jeśli mają identyczne rozwiązania. Tak więc, zamiast pewnego układu równań możemy rozwiązywać dowolny, równoważny mu układ; nie stracimy przez to żadnego rozwiązania, żadne nowe też się nie pojawi. Ten prosty pomysł jest podstawą stosowanych procedur numerycznych. Chcąc mianowicie rozwiązać dany układ równań, przekształcamy go za pomocą pewnych elementarnych operacji na prostszy układ równoważny i dopiero ten rozwiązujemy.

Operacje elementarne zapowiedziane w poprzednim akapicie dzielą się na trzy rodzaje (w ich opisie $\mathcal{E}_i$ oznacza i-te równanie układu):

1. Przestawienie dwóch równań w układzie: $\mathcal{E}_i \leftrightarrow \mathcal{E}_j$.
2. Pomnożenie obu stron równania przez liczbę różną od 0: $\lambda\mathcal{E}_i \to \mathcal{E}_i$.
3. Dodanie stronami do równania wielokrotności innego równania, czyli $\mathcal{E}_i + \lambda\mathcal{E}_j \to \mathcal{E}_i$, gdzie $i \neq j$.

TWIERDZENIE 4.1.1. *Jeśli układ równań wynika z innego układu przez ciąg skończony operacji elementarnych, to te dwa układy są równoważne.*

Dowód. Wystarczy rozważyć skutek zastosowania pojedynczej operacji. Niech przekształca ona układ $Ax = b$ w układ $Bx = d$. W przypadku operacji typu **1** te dwa układy składają się z tych samych równań, chociaż inaczej uporządkowanych. Dlatego rozwiązanie x pierwszego układu spełnia drugi układ i na odwrót. W przypadku operacji typu **2** i-temu równaniu

$$a_{i1}x_1 + \ldots + a_{in}x_n = b_i \tag{4.1.1}$$

pierwszego układu odpowiada i-te równanie

$$\lambda a_{i1}x_1 + \ldots + \lambda a_{in}x_n = \lambda b_i$$

drugiego układu. Te równania mogą być spełnione tylko jednocześnie, gdyż z założenia $\lambda \neq 0$. Na koniec rozważmy operację typu **3**, wykonaną na równaniu (4.1.1) i równaniu

$$a_{j1}x_1 + \ldots + a_{jn}x_n = b_j. \tag{4.1.2}$$

Daje ona następujące i-te równanie drugiego układu:

$$(a_{i1} + \lambda a_{j1})x_1 + \ldots + (a_{in} + \lambda a_{jn})x_n = b_i + \lambda b_j; \tag{4.1.3}$$

j-te równanie oczywiście nie zmienia się. Jeśli x spełnia pierwszy układ, a więc m.in. jego i-te i j-te równanie, to spełnia też równanie (4.1.3). Z drugiej strony, jeśli x jest rozwiązaniem układu $Bx = d$, to zachodzi (4.1.2) i (4.1.3). Odejmując stronami od (4.1.3) równanie (4.1.2) pomnożone przez λ, dostajemy i-te równanie (4.1.1) pierwszego układu. ■

Własności macierzy

Jeśli A i B są takimi macierzami, że $AB = I$, to mówimy, że B jest *prawą odwrotnością* macierzy A, natomiast A jest *lewą odwrotnością* macierzy B. Mamy np.

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

Z tego przykładu wynika, że jeśli macierz ma prawą odwrotność, to nie musi być ona określona jednoznacznie. Prostsza sytuacja, jak zobaczymy niżej, ma miejsce dla macierzy kwadratowych.

Twierdzenie 4.1.2. *Macierz kwadratowa ma co najwyżej jedną prawą odwrotność.*

Dowód. Niech będzie $AB = I$, gdzie A, B, I są macierzami stopnia n. Niech $A^{(j)}$ oznacza j-tą kolumnę macierzy A, a $I^{(k)}$ – k-tą kolumnę macierzy I. Z równości $AB = I$ wynika, że

$$\sum_{j=1}^{n} b_{jk}A^{(j)} = I^{(k)} \qquad (1 \leqslant k \leqslant n).$$

Każda kolumna macierzy I jest więc kombinacją liniową kolumn macierzy A. Ponieważ kolumny macierzy jednostkowej są bazą przestrzeni $\mathbb{R}^n$, więc to samo jest prawdą dla kolumn macierzy A. To zaś oznacza, że współczynniki b_{jk} w powyższych równościach są określone jednoznacznie. ■

Twierdzenie 4.1.3. *Jeśli A i B są macierzami kwadratowymi takimi, że $AB = I$, to $BA = I$.*

Dowód. Niech będzie $C = BA - I + B$. Wtedy

$$AC = ABA - AI + AB = A - A + I = I.$$

Dlatego C (jak i B) jest prawą odwrotnością macierzy A. Na mocy tw. 4.1.2 jest $B = C$, czyli $BA = I$. ■

Z dwóch ostatnich twierdzeń wynika, że jeśli macierz kwadratowa A ma prawą odwrotność B, to jest ona jedyna i $BA = AB = I$. Macierz B nazywamy wtedy krócej *odwrotnością* macierzy A albo macierzą *odwrotną* względem A, o tej ostatniej zaś mówimy, że jest *nieosobliwa*. Oczywiście wtedy B też jest nieosobliwa, a jej odwrotnością jest A. Wyrażamy to, pisząc $B = A^{-1}$ i $A = B^{-1}$. Oto przykład takiej zależności:

$$\begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Jeśli macierz A jest nieosobliwa, to układ równań $Ax = b$ ma rozwiązanie $x = A^{-1}b$. Gdy znamy już odwrotność A^{-1}, to ostatnia równość pozwala znaleźć wektor x. Nie należy jednak obliczać tej odwrotności tylko w tym celu. Dalej opisano inne metody rozwiązywania układu $Ax = b$, które są bardziej efektywne i dają dokładniejsze wyniki.

Operacje elementarne, o których już była mowa, można interpretować – jak zaraz zobaczymy – jako mnożenie macierzy. *Macierzą elementarną* jest taka macierz stopnia n, która powstaje z macierzy jednostkowej stopnia n przez wykonanie którejś z operacji elementarnych. Te operacje, wyrażone za pomocą wierszy A_s macierzy A, są następujące:

1. Zamiana dwóch wierszy: $A_s \leftrightarrow A_t$.
2. Mnożenie wiersza przez niezerową stałą: $\lambda A_s \to A_s$.
3. Dodanie do wiersza wielokrotności innego wiersza: $A_s + \lambda A_t \to A_s$.

Każdą z tych operacji można wykonać, mnożąc macierz A z lewej strony przez pewną macierz elementarną. Oto przykłady ilustrujące trzy typy operacji:

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

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

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ \lambda a_{21} + a_{31} & \lambda a_{22} + a_{32} & \lambda a_{23} + a_{33} \end{bmatrix}.$$

Aby wykonać na A ciąg operacji elementarnych, wprowadzamy odpowiednie macierze elementarne $E_1, E_2, \ldots, E_m$ i tworzymy macierz

$$E_m E_{m-1} \ldots E_2 E_1 A.$$

Jeśli macierz A jest nieosobliwa, to stosując do niej ciąg operacji elementarnych można ją zredukować do macierzy jednostkowej:

$$E_m E_{m-1} \ldots E_2 E_1 A = I.$$

Stąd wynika, że $A^{-1} = E_m E_{m-1} \ldots E_2 E_1$. Inaczej mówiąc, odwrotność A^{-1} można otrzymać stosując do macierzy I ten sam ciąg operacji. Oto przykład pokazujący obliczanie macierzy odwrotnej: korzystając z macierzy elementarnych

$$E_1 := \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad E_2 := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix},$$

$$E_3 := \begin{bmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad E_4 := \begin{bmatrix} 1 & 0 & -3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

obliczamy

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 3 & 3 \\ 2 & 4 & 7 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I,$$

$$E_1 A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 0 \\ 2 & 4 & 7 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = E_1 I,$$

$$E_2 E_1 A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} = E_2 E_1 I,$$

$$E_3E_2E_1A = \begin{bmatrix} 1 & 0 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 3 & -2 & 0 \\ -1 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} = E_3E_2E_1I,$$

$$E_4E_3E_2E_1A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 9 & -2 & -3 \\ -1 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} = E_4E_3E_2E_1I = A^{-1}.$$

TWIERDZENIE 4.1.4. *Dla macierzy kwadratowej A stopnia n następujące własności są równoważne:*

1. *Istnieje odwrotność macierzy A, czyli ta macierz jest nieosobliwa.*
2. *Wyznacznik macierzy A jest różny od 0*[4].
3. *Wiersze macierzy A tworzą bazę przestrzeni $\mathbb{R}^n$.*
4. *Kolumny macierzy A tworzą bazę przestrzeni $\mathbb{R}^n$.*
5. *Odwzorowanie $\mathbb{R}^n$ na $\mathbb{R}^n$ określone przez macierz A jest iniekcją (jest wzajemnie jednoznaczne).*
6. *Odwzorowanie z* **5** *jest suriekcją (odwzorowanie „na").*
7. *Z równości $Ax = 0$ wynika, że $x = 0$.*
8. *Dla każdego $b \in \mathbb{R}^n$ istnieje dokładnie jedno $x \in \mathbb{R}^n$ takie, że $Ax = b$.*
9. *Macierz A jest iloczynem macierzy elementarnych.*
10. *Liczba 0 nie jest wartością własną macierzy A*[5].

Ważnym rodzajem macierzy są macierze dodatnio określone. Macierz A jest z definicji *dodatnio określona*, jeśli $x^TAx > 0$ dla każdego niezerowego wektora $x \in \mathbb{R}^n$. Taką własność ma na przykład macierz

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix},$$

gdyż

$$x^TAx = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = (x_1 + x_2)^2 + x_1^2 + x_2^2 > 0,$$

jeśli tylko $x_1 \neq 0$ lub $x_2 \neq 0$. Występujący wyżej iloczyn x^TAx nazywamy *formą kwadratową*. Z zadań 15–17 wynika, że badając macierze dodatnio określone możemy ograniczyć się do macierzy symetrycznych. W takim

[4] Autorzy przyjmują, że pojęcie wyznacznika jest znane. Można zapoznać się z nim w wielu podręcznikach algebry (*przyp. tłum.*).

[5] Odwzorowanie wymienione w **5** polega na przekształceniu każdego wektora x na wektor Ax. Wartości własne macierzy są zdefiniowane w podrozdz. 4.6 (*przyp. tłum.*).

przypadku wszystkie wartości własne macierzy (określone dalej) są rzeczywiste i dodatnie. Dodajmy, że na ogół nie jest łatwo ustalić na podstawie definicji, czy macierz jest dodatnio określona, gdyż wymaga to sprawdzenia pewnej własności dla każdego $x \neq 0$.

Nieco szerszą klasę tworzą macierze *dodatnio półokreślone*, tj. takie, że $x^{\mathsf{T}}Ax \geqslant 0$ dla każdego $x \in \mathbb{R}^n$.

Macierze blokowe

Często jest wygodnie podzielić macierze na podmacierze czyli *bloki* i mnożyć macierze tak, jakby te bloki były liczbami: [6)]

$$\left[\begin{array}{cc|ccc|c} 1 & 2 & 1 & -1 & 0 & 1 \\ \hline -1 & 1 & 1 & 0 & -1 & 1 \\ 0 & 1 & -1 & 1 & 0 & 1 \\ 1 & -1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 2 & 1 & 0 \end{array}\right] \left[\begin{array}{ccc|cc} 1 & 0 & 1 & 2 & 1 \\ -1 & 1 & 2 & 0 & 1 \\ \hline 1 & 0 & 1 & 1 & 2 \\ -1 & 1 & 0 & 0 & 1 \\ 2 & 1 & 0 & -2 & 1 \\ 0 & 1 & 1 & -1 & 1 \end{array}\right] =$$

$$= \left[\begin{array}{ccc|cc} 1 & 2 & 7 & 2 & 5 \\ \hline -3 & 1 & 3 & 0 & 2 \\ -3 & 3 & 2 & -2 & 1 \\ 4 & 0 & -1 & 0 & 1 \\ 2 & 3 & 2 & 1 & 6 \end{array}\right].$$

Oznaczając powyższe bloki prostszymi symbolami, można tę równość napisać tak:

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}.$$

Można sprawdzić, że $C_{ij} = \sum_{s=1}^{2} A_{is}B_{sj}$. W szczególności

$$C_{11} = A_{11}B_{11} + A_{12}B_{21} =$$

$$= \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 2 \end{bmatrix} + \begin{bmatrix} 1 & -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \\ 2 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 7 \end{bmatrix}.$$

[6)] *Macierz blokowa* powstaje przez podział na części liniami poziomymi lub pionowymi przecinającymi całą macierz. Dlatego bloki leżące obok siebie mają tę samą liczbę wierszy, a bloki leżące jeden nad drugim – tę samą liczbę kolumn (*przyp. tłum.*).

Aby sformułować ogólny wynik dotyczący mnożenia macierzy blokowych, przyjmijmy, że macierze A, B i C są w następujący sposób podzielone na bloki:

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \dots & \dots & \dots & \dots \\ A_{m1} & A_{m2} & \dots & A_{mn} \end{bmatrix}, \qquad B = \begin{bmatrix} B_{11} & B_{12} & \dots & B_{1k} \\ B_{21} & B_{22} & \dots & B_{2k} \\ \dots & \dots & \dots & \dots \\ B_{n1} & B_{n2} & \dots & B_{nk} \end{bmatrix},$$

$$C = \begin{bmatrix} C_{11} & C_{12} & \dots & C_{1k} \\ C_{21} & C_{22} & \dots & C_{2k} \\ \dots & \dots & \dots & \dots \\ C_{m1} & C_{m2} & \dots & C_{mk} \end{bmatrix}.$$

TWIERDZENIE 4.1.5. *Jeśli każdy z iloczynów $A_{is}B_{sj}$ jest określony i jeśli $C_{ij} = \sum_{s=1}^{n} A_{is}B_{sj}$, to $C = AB$.*

Dowód. Niech A_{ij} będzie macierzą rozmiaru $m_i \times n_j$, a B_{ij} – macierzą $\hat{m}_i \times \hat{n}_j$. Ponieważ iloczyn $A_{is}B_{sj}$ istnieje, więc $n_s = \hat{m}_s$ dla każdego s. Stąd wynika, że C_{ij} jest macierzą $m_i \times \hat{n}_j$. Rozważmy teraz dowolny element c_{ij} macierzy C. Przypuśćmy, że c_{ij} leży w bloku C_{rs}, a dokładniej w jego p-tym wierszu i q-tej kolumnie. Stąd

$$i = m_1 + m_2 + \ldots + m_{r-1} + p, \tag{4.1.4}$$

$$j = \hat{n}_1 + \hat{n}_2 + \ldots + \hat{n}_{s-1} + q. \tag{4.1.5}$$

Dzięki temu

$$c_{ij} = (C_{rs})_{pq} = \Big(\sum_{t=1}^{n} A_{rt}B_{ts}\Big)_{pq} = \sum_{t=1}^{n}(A_{rt}B_{ts})_{pq} =$$

$$= \sum_{t=1}^{n}\sum_{\alpha=1}^{n_t}(A_{rt})_{p\alpha}(B_{ts})_{\alpha q}.$$

Wobec (4.1.4) elementy $(A_{rt})_{p\alpha}$ leżą w i-tym wierszu macierzy A i zapełniają go w całości, gdyż $1 \leqslant t \leqslant n$ i $1 \leqslant \alpha \leqslant n_t$. Podobnie rozumując, stwierdzamy na mocy (4.1.5), że elementy $(B_{ts})_{\alpha q}$ zapełniają, w naturalnym porządku, j-tą kolumnę macierzy B. Dlatego

$$c_{ij} = \sum_{\beta=1}^{\nu}(A)_{i\beta}(B)_{\beta j} = (AB)_{ij},$$

gdzie ν jest liczbą kolumn macierzy A i zarazem liczbą wierszy macierzy B (przed ich podziałem na bloki). ■

ZADANIA 4.1 [7]

1. Czy lewa odwrotność macierzy może być różna od jej prawej odwrotności?
2. Wykazać, że operacja elementarna pierwszego typu jest równoważna złożeniu czterech operacji dwóch pozostałych typów, w których wystarczy dopuścić wartości $\lambda = \pm 1$.
3. Czy tw. 4.1.1 jest prawdziwe dla układów, w których liczba równań różni się od liczby niewiadomych?
4. Udowodnić, że skutki każdej operacji elementarnej można usunąć za pomocą operacji tego samego typu.
5. Udowodnić, że dodawanie i mnożenie macierzy trójkątnych górnych (dolnych) oraz mnożenie takich macierzy przez skalar dają macierze tego samego typu.
6. Wykazać, że jeśli macierz trójkątna górna (dolna) jest nieosobliwa, to jej odwrotność jest trójkątna górna (odpowiednio, dolna). Wykazać, że jeśli dana macierz powyższego typu ma jedynki na głównej przekątnej, to tę samą własność ma jej odwrotność.
7. Rozważmy układ równań liniowych $Ax = b$, gdzie A jest macierzą $m \times n$, x – wektorem $n \times 1$, a b – wektorem $m \times 1$. Niech $A_1, A_2, \ldots, A_n$ będą kolumnami macierzy A. Udowodnić, że układ ma rozwiązanie wtedy i tylko wtedy, gdy b należy do przestrzeni rozpiętej na tych kolumnach. Udowodnić, że jeśli są one liniowo niezależne, to układ ma co najwyżej jedno rozwiązanie.
8. Niech $E(p, q, \lambda)$, gdzie $p \neq q$, będzie macierzą, która powstaje z macierzy I przez dodanie q-tego wiersza pomnożonego przez λ do p-tego wiersza. Wykazać, że:
   (a) dla dowolnej macierzy A iloczyn $E(p, q, \lambda)A$ wynika z A przez taką samą jej zmianę,
   (b) dla dowolnej macierzy A iloczyn $AE(p, q, \lambda)$ wynika z A przez dodanie p-tej kolumny pomnożonej przez λ do q-tej kolumny,
   (c) $E(p, q, \lambda)^{-1} = E(p, q, -\lambda)$.

   (W (a) i (b) rozmiary macierzy A i E muszą pozwalać na ich mnożenie).
9. Udowodnić, że macierz, w której każdy wiersz i każda kolumna zawierają dokładnie jeden element niezerowy, jest nieosobliwa.

---

[7] Podajemy definicje dodatkowych pojęć z teorii macierzy, potrzebne w tych zadaniach i dalej. Do *przestrzeni wartości* macierzy A rozmiaru $m \times n$ należy każdy wektor $y \in \mathbb{R}^m$ taki, że dla pewnego $x \in \mathbb{R}^n$ jest $y = Ax$. Jest to podprzestrzeń liniowa przestrzeni $\mathbb{R}^m$. Jej wymiar (równy liczbie niezależnych liniowo kolumn [lub wierszy] macierzy A) nazywamy *rzędem* macierzy A i oznaczamy rank A. Dla macierzy stopnia n rząd jest równy n wtedy i tylko wtedy, gdy jest ona nieosobliwa. *Jądrem* macierzy A rozmiaru $m \times n$ nazywamy zbiór wszystkich wektorów $x \in \mathbb{R}^n$ takich, że $Ax = 0$. Jądro macierzy kwadratowej nieosobliwej zawiera tylko wektor zerowy. Ogólnie, dla macierzy rozmiaru $m \times n$, jest to podprzestrzeń liniowa przestrzeni $\mathbb{R}^n$. Jej wymiar, zwiększony o rank A, jest równy n (*przyp. tłum.*).

10. Czy poniższe macierze są dodatnio określone?

$$\text{(a)}\ \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \qquad \text{(b)}\ \begin{bmatrix} 4 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 4 \end{bmatrix}.$$

11. Dla jakich a macierz

$$A = \begin{bmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{bmatrix}$$

jest dodatnio określona?

12. Czy z dodatniej określoności macierzy A wynika ta sama własność dla A^{-1}?

13. Znaleźć warunki na a, b i c gwarantujące, że macierz $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$ jest dodatnio półokreślona.

14. Udowodnić, że $(AB)^{\mathsf{T}} = B^{\mathsf{T}}A^{\mathsf{T}}$, jeśli tylko iloczyn AB istnieje.

15. Macierz kwadratowa A jest *skośnosymetryczna*, jeśli $A^{\mathsf{T}} = -A$. Udowodnić, że jeśli A jest taką macierzą, to $x^{\mathsf{T}}Ax = 0$ dla każdego x.

16. (cd.). Udowodnić, że elementy przekątniowe macierzy skośnosymetrycznej są równe 0. Ponadto udowodnić, że jeśli stopień tej macierzy jest nieparzysty, to jej wyznacznik znika.

17. (cd.). Dla macierzy kwadratowej A definiujemy $A_0 := \frac{1}{2}(A+A^{\mathsf{T}})$ i $A_1 := \frac{1}{2}(A-A^{\mathsf{T}})$. Udowodnić, że macierz A_0 jest symetryczna, A_1 jest skośnosymetryczna, $A = A_0+A_1$ i $x^{\mathsf{T}}Ax = x^{\mathsf{T}}A_0x$ dla każdego x. To tłumaczy, dlaczego rozważając formy kwadratowe, możemy ograniczyć się do macierzy symetrycznych.

18. Podać przykład macierzy symetrycznej A, której wszystkie elementy są dodatnie i dla której pewne wartości $x^{\mathsf{T}}Ax$ są ujemne.

19. Niech

$$A := \begin{bmatrix} B & C \\ 0 & I \end{bmatrix}$$

będzie macierzą blokową, w której wszystkie bloki są kwadratowe stopnia n. Udowodnić, że jeśli macierz $B - I$ jest nieosobliwa, to dla $k \geqslant 1$ jest

$$A^k = \begin{bmatrix} B^k & (B^k - I)(B-I)^{-1}C \\ 0 & I \end{bmatrix}.$$

20. (cd.). Odwołując się do poprzedniego zadania, znaleźć strukturę blokową macierzy A^k, gdy

$$A = \begin{bmatrix} B & 0 \\ C & I \end{bmatrix}.$$

21. Niech A będzie macierzą nieosobliwą stopnia n i niech u i v będą wektorami z $\mathbb{R}^n$. Znaleźć warunki konieczne i dostateczne na to, żeby macierz

$$\begin{bmatrix} A & u \\ v^T & 0 \end{bmatrix}$$

była nieosobliwa i znaleźć wyrażenie dla jej odwrotności, jeśli istnieje.

22. Niech D będzie następującą macierzą blokową:

$$\begin{bmatrix} A & B \\ C & I \end{bmatrix}.$$

Udowodnić, że jeśli różnica $A - BC$ jest nieosobliwa, to taka jest również macierz D.

23. (cd.). Udowodnić mocniejszy wynik: że wymiar jądra macierzy D jest nie większy od wymiaru jądra macierzy $A - BC$.

ZADANIA KOMPUTEROWE 4.1

Czytelnikom, których interesują eksperymenty obliczeniowe, sugerujemy projekt dzielący się na kilka etapów i polegający na napisaniu procedur wykonujących podstawowe zadania algebry liniowej. Zestaw tych procedur pozwoliłby rozwiązywać układy równań liniowych, rozkładać macierze na czynniki różnych typów oraz obliczać wartości i wektory własne. Kolejnymi etapami projektu są następujące zadania komputerowe: K1, 4.2.K2, 4.2.K4, wszystkie z podrozdz. 4.3 i zad. 4.4.K1.

K1. Napisać i sprawdzić następujące procedury:

(a) `Store`(n, x, y), która zastępuje wektor y o n składowych takimże wektorem x: $y \leftarrow x$.

(b) `Prod`(m, n, A, x, y), która mnoży macierz A rozmiaru $m \times n$ z prawej strony przez wektor x o n składowych i zapamiętuje wynik jako wektor y o m składowych: $y \leftarrow Ax$.

(c) `Mult`(k, m, n, A, B, C), która oblicza $C \leftarrow AB$ dla macierzy A rozmiaru $k \times n$, B rozmiaru $m \times n$; C jest więc macierzą $k \times n$.

(d) `Dot`(n, x, y, a), która w arytmetyce podwójnej precyzji oblicza iloczyn skalarny wektorów x i y o n składowych w pojedynczej precyzji i podstawia wynik, skrócony do pojedynczej precyzji, pod a: $a \leftarrow \sum_{i=1}^{n} x_i y_i$.

4.2. Rozkłady LU

Rozważmy układ n równań liniowych z n niewiadomymi, który – jak już wiemy – można napisać w postaci

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

Oznaczając występujące tu macierze symbolami A, x i b, wyrażamy ten układ po prostu tak:

$$Ax = b.$$

Układy łatwe do rozwiązania

Przyjrzymy się najpierw szczególnym układom, które można łatwo rozwiązać. Załóżmy na przykład, że macierz kwadratowa A stopnia n jest przekątniowa:

$$\begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{bmatrix}.$$

Taki układ rozkłada się na n oddzielnych równań i ma rozwiązanie

$$x = \begin{bmatrix} b_1/a_{11} \\ b_2/a_{22} \\ \dots \\ b_n/a_{nn} \end{bmatrix}.$$

Jeśli dla pewnego i jest $a_{ii} = 0$ i $b_i = 0$, to x_i może być dowolną liczbą. Jeśli natomiast $a_{ii} = 0$ i $b_i \neq 0$, to układ nie ma rozwiązania.

Załóżmy teraz, że macierz A jest trójkątna dolna:

$$\begin{bmatrix} a_{11} & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{bmatrix}.$$

Aby go rozwiązać, załóżmy, że $a_{ii} \neq 0$ dla każdego i. Wtedy z pierwszego równania wyznaczamy x_1. Znając już tę niewiadomą i podstawiając ją do drugiego równania, możemy wyznaczyć x_2. Postępując tak dalej, wyznaczamy kolejno $x_1, x_2, \dots, x_n$ – właśnie w tym porządku. Stosowany przy tym algorytm nazywamy *podstawianiem w przód*:

input $n, (a_{ij}), (b_i)$
for $i = 1$ **to** n **do**
    $x_i \leftarrow \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j\right)/a_{ii}$
end do
output (x_i)

Ten sam pomysł można zastosować, gdy macierz układu jest *trójkątna górna*, tj. gdy jej elementy poniżej głównej przekątnej znikają:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{bmatrix}.$$

Znów załóżmy, że $a_{ii} \neq 0$ dla $1 \leqslant i \leqslant n$. Algorytm rozwiązania takiego układu nazywamy *podstawianiem wstecz*:

input $n, (a_{ij}), (b_i)$
for $i = n$ **to** 1 **step** -1 **do**
    $x_i \leftarrow \left(b_i - \sum_{j=i+1}^{n} a_{ij} x_j\right) / a_{ii}$
end do
output (x_i)

W podobny sposób można łatwo rozwiązywać także inne układy równań, takie mianowicie, których macierz powstaje z macierzy trójkątnej przez przestawienie wierszy. Do tej klasy należy m.in. układ

$$\begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & a_{23} \\ a_{31} & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}. \tag{4.2.1}$$

Odpowiednie przestawienie tych równań dałoby układ z macierzą trójkątną dolną:

$$\begin{bmatrix} a_{31} & 0 & 0 \\ a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_3 \\ b_1 \\ b_2 \end{bmatrix},$$

który można rozwiązać znaną już metodą. Możemy też jednak rozwiązywać pierwotny układ (4.2.1), chociaż nie w naturalnym porządku równań: 1, 2, 3, ale w kolejności 3, 1, 2.

Spróbujmy opisać dokładniej własność macierzy pokazaną wyżej na przykładzie. Niech pewien wiersz macierzy A, np. p_1-szy, ma zera na pozycjach $2, 3, \dots, n$. Niech inny, p_2-gi, wiersz ma zera na pozycjach $3, 4, \dots, n$ itd. Jeśli tak jest, to wierszy o numerach $p_1, p_2, \dots$ używamy odpowiednio do obliczenia $x_1, x_2, \dots$ Gdybyśmy ustawili wiersze w kolejności $p_1, p_2, \dots, p_n$, to powstałaby macierz trójkątna dolna.

Jak można rozwiązać układ $Ax = b$, jeśli A jest spermutowaną macierzą trójkątną dolną lub górną? Załóżmy, że *permutacja* $(p_1, p_2, \dots, p_n)$ układu $(1, 2, \dots, n)$ jest już w jakiś sposób wyznaczona. Modyfikując znany już algorytm, otrzymujemy algorytm podstawiania w przód dla *spermutowanego układu o macierzy trójkątnej dolnej*:

input $n, (a_{ij}), (b_i), (p_i)$
for $i = 1$ **to** n **do**
    $x_i \leftarrow \left(b_{p_i} - \sum_{j=1}^{i-1} a_{p_i j} x_j\right) / a_{p_i i}$
end do
output (x_i)

Tu macierz A jest z założenia taka, że $a_{p_i j} = 0$ dla $j > i$ oraz $a_{p_i i} \neq 0$ dla każdego i. Natomiast algorytm podstawiania wstecz dla *spermutowanego układu o macierzy trójkątnej górnej* jest następujący:

input $n, (a_{ij}), (b_i), (p_i)$
for $i = n$ **to** 1 **step** -1 **do**
    $x_i \leftarrow \left(b_{p_i} - \sum_{j=i+1}^{n} a_{p_i j} x_j\right) / a_{p_i i}$
end do
output (x_i)

Tu trzeba założyć, że $a_{p_i j} = 0$ dla $j < i$ oraz $a_{p_i i} \neq 0$ dla każdego i.

Cztery podane już algorytmy są zbyt ogólnikowe, żeby można je było bezpośrednio przetłumaczyć na typowy język programowania. Niżej podano dla ostatniego algorytmu bardziej szczegółowy ciąg instrukcji:

input $n, (a_{ij}), (b_i), (p_i)$
for $i = n$ **to** 1 **step** -1 **do**
    $s \leftarrow b_{p_i}$
    for $j = i + 1$ **to** n **do**
        $s \leftarrow s - a_{p_i j} x_j$
    end do
    $x_i \leftarrow s / a_{p_i i}$
end do
output (x_i)

Rozkłady LU

Przypuśćmy, że macierz A można wyrazić jako iloczyn macierzy trójkątnej dolnej L i trójkątnej górnej U: $A = LU$. Wtedy rozwiązywanie układu równań $Ax = b$ dzieli się na dwa etapy:
rozwiązywanie $Lz = b$ względem z,
rozwiązywanie $Ux = z$ względem x.
Wiemy już, że jest to łatwe, gdyż każdy z tych układów ma macierz trójkątną.

Pokażemy teraz, jak można znaleźć rozkład $A = LU$, jeśli tylko w którymś momencie obliczeń nie wystąpi dzielenie przez 0. Nie każda macierz ma taki rozkład i kłopot z tym związany przedyskutujemy nieco później.

Dla danej macierzy kwadratowej A stopnia n szukamy więc macierzy

$$L := \begin{bmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{bmatrix}, \qquad U := \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ 0 & u_{22} & \dots & u_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & u_{nn} \end{bmatrix}$$

takich, że

$$A = LU.$$

Jeśli one istnieją, to mówimy, że A ma *rozkład LU*. Okazuje się, że równość $A = LU$ nie określa czynników L i U jednoznacznie. Istotnie, dla każdego i można wybrać dowolną wartość różną od 0 dla jednej z liczb l_{ii}, u_{ii} (ale nie dla obu). Byłoby np. naturalne przyjąć, że $l_{ii} = 1$ dla $i = 1, 2, \dots, n$. Wtedy L jest *jedynkowa trójkątna dolna*. Inny oczywisty wybór jest taki, że $u_{ii} = 1$ dla każdego i. Wtedy U jest *jedynkowa trójkątna górna*. Te dwa przypadki są szczególnie ważne.

Aby opracować algorytm dla rozkładu LU, zaczynamy od wzoru na mnożenie macierzy:

$$a_{ij} = \sum_{s=1}^{n} l_{is} u_{sj} = \sum_{s=1}^{\min(i,j)} l_{is} u_{sj}. \tag{4.2.2}$$

Wykorzystano tu równości $l_{is} = 0$ dla $s > i$ oraz $u_{sj} = 0$ dla $s > j$.

Każdy etap obliczeń wyznacza jeden nowy wiersz macierzy U i jedną nową kolumnę macierzy L. W k-tym kroku możemy więc założyć, że znamy już początkowych $k-1$ wierszy macierzy U i tyleż początkowych kolumn macierzy L (dla $k = 1$ założenie jest oczywiście spełnione). Przyjmując w (4.2.2), że $i = j = k$, mamy

$$a_{kk} = \sum_{s=1}^{k-1} l_{ks} u_{sk} + l_{kk} u_{kk}.$$

Stąd, ustaliwszy wartość jednego z elementów u_{kk}, l_{kk}, wyznaczamy drugi z nich. Znając już oba, stosujemy równość (4.2.2) do obliczenia k-tego wiersza macierzy U (wtedy $i = k$) i k-tej kolumny macierzy L (wtedy $j = k$):

$$a_{kj} = \sum_{s=1}^{k-1} l_{ks} u_{sj} + l_{kk} u_{kj} \qquad (k+1 \leqslant j \leqslant n), \tag{4.2.3}$$

$$a_{ik} = \sum_{s=1}^{k-1} l_{is} u_{sk} + l_{ik} u_{kk} \qquad (k+1 \leqslant i \leqslant n). \tag{4.2.4}$$

Jeśli $l_{kk} \neq 0$, to na podstawie (4.2.3) obliczamy elementy u_{kj}. Podobnie, jeśli $u_{kk} \neq 0$, to (4.2.4) pozwala obliczyć elementy l_{ik}. Warto zauważyć, że te dwie czynności można wykonywać równolegle (tzn. w tym samym czasie). Pewne komputery to umożliwiają, co daje istotną oszczędność czasu; Kincaid i Oppe [1988] podają szczegóły takiego postępowania. Obliczenie k-tego wiersza macierzy U i k-tej kolumny macierzy L w opisany wyżej sposób zamyka k-ty krok obliczeń. W obliczeniach występuje dzielenie przez l_{kk} i u_{kk}; jeśli któraś z tych liczb znika, to na ogół, choć nie zawsze, nie można zakończyć obliczeń; zob. zad. 38.

Algorytm opisany wyżej jest znany jako *rozkład Doolittle'a*, jeśli macierz L ma być taka, że $l_{ii} = 1$ dla $1 \leqslant i \leqslant n$ i jako *rozkład Crouta*, gdy macierz U ma być taka, że $u_{ii} = 1$ dla $1 \leqslant i \leqslant n$. Gdy $U = L^{\mathsf{T}}$, więc $l_{ii} = u_{ii}$ dla $1 \leqslant i \leqslant n$, mówimy o *rozkładzie Cholesky'ego*[8]. Metodę Cholesky'ego opiszemy szczegółowo później, gdyż wymaga specjalnych założeń o macierzy A: powinna ona być rzeczywista, symetryczna i dodatnio określona.

Który z wymienionych rozkładów jest lepszy? Powiemy tylko, że każdy wiąże się z innym wariantem eliminacji Gaussa (podrozdz. 4.3) i że trzeba je wszystkie poznać, aby w pełni zrozumieć rozważane tu zagadnienie.

Algorytm dowolnego rozkładu LU jest następujący:

input $n, (a_{ij})$
for $k = 1$ **to** n **do**
  Wybrać wartość niezerową dla l_{kk} albo u_{kk}
    i wyznaczyć drugą z tych liczb z równości
    $l_{kk}u_{kk} = a_{kk} - \sum_{s=1}^{k-1} l_{ks}u_{sk}$
  for $j = k+1$ **to** n **do**
    $u_{kj} \leftarrow \left(a_{kj} - \sum_{s=1}^{k-1} l_{ks}u_{sj}\right)/l_{kk}$
    $l_{jk} \leftarrow \left(a_{jk} - \sum_{s=1}^{k-1} l_{js}u_{sk}\right)/u_{kk}$
  end do
end do
output $(l_{ij}), (u_{ij})$

Przypomnijmy, że można równocześnie obliczać k-ty wiersz macierzy U i k-tą kolumnę macierzy L; zob. zad. 35.

Szczegóły algorytmów rozkładu macierzy na czynniki trójkątne dobiera się tak, aby jak najlepiej wykorzystać właściwości nowoczesnych komputerów dużej mocy. Aby opracować algorytmy skalowalne, wykorzystujące

[8] W dalszym ciągu używa się też określenia *metoda* (lub *algorytm*) *Doolittle'a*, *Crouta*, *Cholesky'ego* rozkładu na czynniki. Ta ostatnia metoda jest też nazywana metodą *pierwiastków kwadratowych*. Jej opis znajduje się w rękopisie Cholesky'ego z 1910 r., odnalezionym niedawno w archiwach paryskiej Ecole Polytechnique przez C. Brezinskiego. On też ustalił, że metodę, z powołaniem się na Cholesky'ego, opublikował Benoit w *Bull. Géodésique* **2**(1924), s. 67–77. Metodę odkrył na nowo Tadeusz Banachiewicz, polski astronom, w latach 1938–39 (*przyp. tłum.*).

szczególne własności komputerów o architekturze typu MIMD (*multiple-instruction multiple-data*), stosuje się różne techniki. I tak np., algorytmy operujące na macierzach podzielonych na bloki pozwalają zmniejszyć częstość przesyłania danych między różnymi poziomami pamięci. Specjalne wersje rozproszone podstawowych podprogramów algebry liniowej (BLAS) służą jako bloki obliczeniowe bądź komunikacyjne. Te możliwości wykorzystano w takich bibliotekach programów jak LAPACK i ScaLAPACK. Odpowiednie informacje podają Anderson i in. [1995] oraz Dongarra i Walker [1995].

Dla rozkładu $A = LU$ z (4.2.2) wynika n^2 równań z $n^2 + n$ niewiadomymi elementami tych czynników; n elementów trzeba ustalić w znany już sposób. Inny ich wybór może dać równania nieliniowe (zob. zad. 32).

Algorytm rozkładu Doolittle'a jest następujący:

input $n, (a_{ij})$
for $k = 1$ **to** n **do**
  $l_{kk} \leftarrow 1$
  for $j = k$ **to** n **do**
    $u_{kj} \leftarrow a_{kj} - \sum_{s=1}^{k-1} l_{ks} u_{sj}$
  end do
  for $i = k + 1$ **to** n **do**
    $l_{ik} \leftarrow (a_{ik} - \sum_{s=1}^{k-1} l_{is} u_{sk}) / u_{kk}$
  end do
end do
output $(l_{ij}), (u_{ij})$

Przykład 4.2.1. Znaleźć rozkłady Doolittle'a, Crouta i Cholesky'ego macierzy

$$A := \begin{bmatrix} 60 & 30 & 20 \\ 30 & 20 & 15 \\ 20 & 15 & 12 \end{bmatrix}.$$

Rozwiązanie. Posługując się podanym wyżej algorytmem, otrzymujemy rozkład Doolittle'a:

$$A = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{3} & 1 & 1 \end{bmatrix} \begin{bmatrix} 60 & 30 & 20 \\ 0 & 5 & 5 \\ 0 & 0 & \frac{1}{3} \end{bmatrix} \equiv LU.$$

Zamiast obliczać bezpośrednio dwa pozostałe rozkłady, można je otrzymać z poprzedniego. Umieszczając elementy przekątniowe macierzy U w macierzy przekątniowej D, możemy napisać, że

$$A = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{3} & 1 & 1 \end{bmatrix} \begin{bmatrix} 60 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \equiv LD\hat{U}.$$

Przyjmując $\hat{L} = LD$, otrzymujemy rozkład Crouta:

$$A = \begin{bmatrix} 60 & 0 & 0 \\ 30 & 5 & 0 \\ 20 & 5 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \equiv \hat{L}\hat{U}.$$

Aby otrzymać rozkład Cholesky'ego z iloczynu $LD\hat{U}$, wyrażamy macierz D w postaci $D^{1/2}D^{1/2}$ i mnożymy pierwszy czynnik przez L, a drugi przez $\hat{U}$:

$$A = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{3} & 1 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{60} & 0 & 0 \\ 0 & \sqrt{5} & 0 \\ 0 & 0 & \frac{1}{3}\sqrt{3} \end{bmatrix} \begin{bmatrix} \sqrt{60} & 0 & 0 \\ 0 & \sqrt{5} & 0 \\ 0 & 0 & \frac{1}{3}\sqrt{3} \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} =$$

$$= \begin{bmatrix} \sqrt{60} & 0 & 0 \\ \frac{1}{2}\sqrt{60} & \sqrt{5} & 0 \\ \frac{1}{3}\sqrt{60} & \sqrt{5} & \frac{1}{3}\sqrt{3} \end{bmatrix} \begin{bmatrix} \sqrt{60} & \frac{1}{2}\sqrt{60} & \frac{1}{3}\sqrt{60} \\ 0 & \sqrt{5} & \sqrt{5} \\ 0 & 0 & \frac{1}{3}\sqrt{5} \end{bmatrix} \equiv \tilde{L}\tilde{L}^{\mathsf{T}}. \qquad \blacksquare$$

W dalszym ciągu pisząc o rozkładzie LU, będziemy mieli na uwadze jego najważniejszy typ, a mianowicie rozkład Doolittle'a $A = LU$, w którym L jest macierzą jedynkową trójkątną dolną, a U – macierzą trójkątną górną.

Oto warunek dostateczny istnienia rozkładu LU macierzy kwadratowej:

Twierdzenie 4.2.2. *Jeśli wszystkie minory główne macierzy kwadratowej A są nieosobliwe, to ma ona rozkład LU.*

Dowód. Zacznijmy od przypomnienia, że k-tym *minorem głównym* macierzy A jest macierz

$$A_k = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \dots & \dots & \dots & \dots \\ a_{k1} & a_{k2} & \dots & a_{kk} \end{bmatrix}.$$

W algorytmie Doolittle'a można zmienić porządek obliczeń, a mianowicie kolejno dla $k = 1, 2, \dots, n$ obliczać elementy $l_{k1}, l_{k2}, \dots, l_{kk}$ macierzy L i elementy $u_{1k}, u_{2k}, \dots, u_{kk}$ macierzy U. Jest to wykonalne dla $k = 1$ i daje $l_{11} = 1$ i $u_{11} = a_{11} = \det A_1 \neq 0$. Załóżmy, że dla pewnego k naturalnego udało się obliczyć k początkowych wierszy macierzy L i k początkowych kolumn macierzy U. Daje to rozkład minora A_k na czynniki L_k (jedynkowy trójkątny dolny) i U_k (trójkątny górny). Ponieważ

$$0 \neq \det A_k = (\det L_k)(\det U_k) = u_{11}u_{22}\dots u_{kk},$$

więc $u_{11}, u_{22}, \ldots, u_{kk} \neq 0$ i za pomocą algorytmu Doolittle'a można wykonać następny etap obliczeń, w którym występują dzielenia przez te wielkości, czyli znaleźć L_{k+1} i U_{k+1}. Prowadzi to ostatecznie do wniosku, że rozkład LU macierzy A istnieje. ■

Rozkład Cholesky'ego

Jak już wspomniano, André Louis Cholesky zaproponował jedną z wersji rozkładu macierzy na czynniki, typu LL^{T}, użyteczną w pewnych sytuacjach. Udowodnił on następujący wynik:

Twierdzenie 4.2.3. *Jeśli macierz A jest rzeczywista, symetryczna i dodatnio określona, to ma ona jedyny rozkład na czynniki $A = LL^{\mathsf{T}}$, gdzie L jest macierzą trójkątną dolną o elementach dodatnich na głównej przekątnej.*

Dowód. Przypomnijmy, że macierz A jest symetryczna i dodatnio określona, jeśli $A = A^{\mathsf{T}}$ i $x^{\mathsf{T}}Ax > 0$ dla każdego wektora $x \neq 0$. Stąd od razu wynika, że ta macierz jest nieosobliwa, gdyż nie może przekształcić żadnego wektora $x \neq 0$ na 0 (tw. 4.1.4, własność **7**). Co więcej, rozważając szczególne wektory postaci $x = (x_1, x_2, \ldots, x_k, 0, \ldots, 0)$, wnioskujemy, że również minory wiodące główne macierzy A są dodatnio określone. Na mocy tw. 4.2.2 ma ona rozkład LU. Z symetrii macierzy wynika, że

$$LU = A = A^{\mathsf{T}} = U^{\mathsf{T}}L^{\mathsf{T}},$$

wobec czego

$$U(L^{\mathsf{T}})^{-1} = L^{-1}U^{\mathsf{T}}.$$

Lewa strona tej równości jest macierzą trójkątną górną, a prawa – trójkątną dolną (zob. zad. 4.1.6). Dlatego obie strony są równe pewnej macierzy przekątniowej D. Ponieważ $U(L^{\mathsf{T}})^{-1} = D$, więc $U = DL^{\mathsf{T}}$ i $A = LDL^{\mathsf{T}}$. Zgodnie z zad. 19 macierz D jest dodatnio określona, czyli jej elementy d_{ii} są dodatnie. Niech będzie $D^{1/2} := \operatorname{diag}(\sqrt{d_{ii}})$. Wtedy $A = \tilde{L}\tilde{L}^{\mathsf{T}}$, gdzie $\tilde{L} = LD^{1/2}$. Daje to rozkład Cholesky'ego. Dowód jego jednoznaczności pozostawiamy jako zadanie. ■

W przypadku rozkładu LU Cholesky'ego jest $U = L^{\mathsf{T}}$. Dlatego element przekątniowy czynnika L obliczamy tu według wzoru

$$l_{kk} = \Big(a_{kk} - \sum_{s=1}^{k-1} l_{ks}^2\Big)^{1/2}, \tag{4.2.5}$$

a cały algorytm jest następujący:

input $n, (a_{ij})$
for $k = 1$ **to** n **do**
$l_{kk} \leftarrow \left(a_{kk} - \sum_{s=1}^{k-1} l_{ks}^2\right)^{1/2}$
for $i = k+1$ **to** n **do**
$l_{ik} \leftarrow \left(a_{ik} - \sum_{s=1}^{k-1} l_{is} l_{ks}\right)/l_{kk}$
end do
end do
output (l_{ij})

Twierdzenie 4.2.3 zapewnia, że $l_{kk} > 0$. Zauważmy, że (4.2.5) daje oszacowanie

$$a_{kk} = \sum_{s=1}^{k} l_{ks}^2 \geqslant l_{kj}^2 \qquad (j \leqslant k),$$

z którego wynika, że

$$|l_{kj}| \leqslant \sqrt{a_{kk}} \qquad (1 \leqslant j \leqslant k).$$

Tak więc każdy element macierzy L jest ograniczony z góry przez pierwiastek kwadratowy z odpowiedniego elementu przekątniowego macierzy A. Dzięki temu elementy obliczanej macierzy nie są duże w stosunku do elementów danej macierzy, i to nawet bez wyboru elementów głównych (to pojęcie jest opisane w następnym podrozdziale).

Dodajmy, że w algorytmach Cholesky'ego i Doolittle'a iloczyny skalarne wektorów powinny być obliczane z podwójną precyzją; unikamy wtedy nadmiernych błędów zaokrągleń; zob. zad. komputerowe 2.2.K6 [9].

ZADANIA 4.2 [10]

1. Udowodnić, że algorytmy podstawiania wstecz i w przód w wersji z permutacjami pozwalają rozwiązać każdy układ $Ax = b$ z macierzą nieosobliwą A.
2. (cd.). Znaleźć liczbę działań arytmetycznych wykonywanych w tych algorytmach.
3. Korzystając z równości $UU^{-1} = I$, opracować algorytm obliczania odwrotności macierzy trójkątnej górnej. Założyć, że U^{-1} istnieje, tj. że elementy przekątniowe macierzy U są niezerowe.

[9] Jeśli arytmetyka podwójnej precyzji nie jest dostępna, to można zalecić sumowanie składników w pojedynczej precyzji, ale za pomocą algorytmu Gilla-Möllera. Wtedy oszacowanie błędu sumy jest niezależne od liczby jej składników; zob. Jankowscy [1981*, s. 30] (*przyp. tłum.*).

[10] We wszystkich zadaniach tego podrozdziału symbole L, U i D oznaczają odpowiednio macierz trójkątną dolną, trójkątną górną i przekątniową. Dlatego niżej podaje się explicite tylko dodatkowe założenia o tych macierzach (*przyp. tłum.*).

4. Znaleźć liczbę mnożeń i/lub dzieleń potrzebnych do odwrócenia macierzy jedynkowej trójkątnej dolnej.
5. Niech A będzie macierzą stopnia n, a $(p_1, p_2, \ldots, p_n)$ – taką permutacją zbioru $(1, 2, \ldots, n)$, że i-ty wiersz macierzy zawiera niezerowe elementy tylko w kolumnach $p_1, p_2, \ldots, p_i$ $(i = 1, 2, \ldots, n)$. Napisać algorytm rozwiązania układu $Ax = b$.
6. Niech układ $Ax = b$ ma następującą własność: istnieją takie permutacje $(p_1, p_2, \ldots, p_n)$ i $(q_1, q_2, \ldots, q_n)$ zbioru $(1, 2, \ldots, n)$, że dla każdego i równanie o numerze p_i zawiera tylko zmienne $x_{q_1}, x_{q_2}, \ldots, x_{q_i}$. Napisać efektywny algorytm rozwiązywania takiego układu.
7. Sprawdzić poprawność poniższego algorytmu rozwiązania układu $Ux = b$.

for $j = n$ **to** 1 **step** -1 **do**
  $x_j \leftarrow b_j / u_{jj}$
  for $i = 1$ **to** $j - 1$ **do**
    $b_i \leftarrow b_i - u_{ij} x_j$
  end do
end do

8. Opracować algorytmy rozwiązywania układu liniowego $Ax = b$ w następujących przypadkach:
   (a) $a_{ij} = 0$ dla $i + j \leqslant n$,
   (b) $a_{ij} = 0$ dla $i + j > n + 1$.
9. Udowodnić, że jeśli macierz nieosobliwa A ma rozkład LU, to jest on jedyny pod warunkiem, że czynnik L jest macierzą jedynkową.
10. Udowodnić, że dla macierzy $A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$ nie istnieje rozkład LU (to nie jest prosta konsekwencja tw. 4.2.2).
11. Pokazać, że każda macierz postaci $A = \begin{bmatrix} 0 & a \\ 0 & b \end{bmatrix}$ ma rozkład LU. Pokazać, że jeśli nawet L jest macierzą *jedynkową*, to rozkład nie jest jedyny. (To zadanie wraz z następnym ilustruje *zasadę Olgi Taussky*: Jeśli jakieś przypuszczenie dotyczące macierzy jest fałszywe, to na ogół można to wykazać już na przykładzie macierzy 2×2).
12. Pokazać, że każda z niżej podanych macierzy ma rozkład LU. Czy tak jest przy dodatkowym założeniu, że L jest macierzą jedynkową?
   (a) $A = \begin{bmatrix} 0 & 0 \\ a & b \end{bmatrix}$, (b) $A = \begin{bmatrix} a & 0 \\ b & 0 \end{bmatrix}$, (c) $A = \begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix}$.
13. Udowodnić, że jeśli macierz A jest nieosobliwa i ma rozkład LU, to jej wszystkie minory główne są też nieosobliwe.
14. Sprawdzić, czy jeśli A ma rozkład LU, gdzie L jest macierzą jedynkową, to ma też taki rozkład z macierzą U jedynkową.
15. Podać algorytm odwracania macierzy A, której rozkład LU jest znany. Odwołać się do zad. 3 i zad. K2.

16. Udowodnić, że jeśli wszystkie minory wiodące główne macierzy A są nieosobliwe, to $A = LDU$, gdzie macierze L i U są jedynkowe.

17. (cd.). Przy dodatkowym założeniu, że A jest symetryczna, wykazać istnienie jej rozkładu LDL^T, gdzie L jest jedynkowa.

18. (cd.). Dla macierzy z poprzedniego zadania napisać algorytm konstrukcji podanego tam rozkładu; algorytm powinien być z grubsza dwukrotnie szybszy od standardowego algorytmu rozkładu na czynniki. **Uwaga**: Algorytm może zawieść, jeśli pewne minory główne macierzy A są osobliwe. (Ta modyfikacja algorytmu Cholesky'ego nie wymaga pierwiastkowania).

19. Udowodnić, że A jest dodatnio określona, a B nieosobliwa wtedy i tylko wtedy, gdy iloczyn BAB^T jest dodatnio określony.

20. Dla macierzy

$$A = \begin{bmatrix} 2 & 6 & -4 \\ 6 & 17 & -17 \\ -4 & -17 & -20 \end{bmatrix}$$

znaleźć bezpośrednio (tj. nie korzystając z podanych dalej algorytmów eliminacji) rozkład $A = LDL^T$ z macierzą L jedynkową.

21. Opracować algorytm bezpośredniego obliczania rozkładu $A = UL$, gdzie macierz L jest jedynkowa. Podać algorytm rozwiązywania układu $ULx = b$.

22. Znaleźć rozkład LU (U – macierz jedynkowa) dla

$$A = \begin{bmatrix} 3 & 0 & 1 \\ 0 & -1 & 3 \\ 1 & 3 & 0 \end{bmatrix}.$$

23. Znaleźć rozkład LL^T dla macierzy $A = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix}$.

24. Znaleźć bezpośrednio rozkład LL^T (L – macierz z dodatnimi elementami na głównej przekątnej) dla

$$A = \begin{bmatrix} 4 & \frac{1}{2} & 1 \\ \frac{1}{2} & \frac{17}{16} & \frac{1}{4} \\ 1 & \frac{1}{4} & \frac{33}{64} \end{bmatrix}.$$

25. Znaleźć rozkład LU macierzy $A = \begin{bmatrix} 1 & 5 \\ 3 & 16 \end{bmatrix}$, w którym obie macierze L i U są jedynkowe. Powtórzyć zadanie, zmieniając 16 na 15.

26. Dla macierzy

$$A = \begin{bmatrix} 6 & 10 & 0 \\ 12 & 26 & 4 \\ 0 & 9 & 12 \end{bmatrix}$$

znaleźć rozkład LU z macierzą L o elementach przekątniowych równych 2.

27. Dla macierzy A symetrycznej i dodatnio określonej udowodnić jednoznaczność rozkładu $A = LL^T$ takiego, że elementy przekątniowe czynnika L są dodatnie.

28. Udowodnić, że jeśli macierz A jest symetryczna, to w jej rozkładzie LU kolumny czynnika L są wielokrotnościami wierszy czynnika U.

29. Dla $A = \begin{bmatrix} 1 & 5 \\ 3 & 17 \end{bmatrix}$ znaleźć wszystkie rozkłady LU i UL z macierzą L jedynkową.

30. Z definicji *P-macierz* jest taka, że $a_{ij} = 0$ dla $i + j \leqslant n$, a *Q-macierz* jest P-macierzą, w której dodatkowo $a_{i,n-i+1} = 1$ dla $i = 1, 2, \ldots, n$. Znaleźć *rozkład PQ* macierzy $A = \begin{bmatrix} 3 & 15 \\ -1 & -1 \end{bmatrix}$.

31. (cd.). Opracować algorytm rozkładu PQ danej macierzy oraz algorytm rozwiązywania układu $PQx = b$.

32. Wykazać, że jeśli w rozkładzie LU macierzy 2×2 ustalamy elementy l_{22} i u_{22}, to równania służące do wyznaczenia pozostałych elementów czynników L i U są nieliniowe.

33. Udowodnić, że jeśli macierz A jest symetryczna i dodatnio półokreślona, to ma rozkład LL^T.

34. Znaleźć warunki konieczne i dostateczne na to, żeby macierz $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$ miała rozkład LL^T.

35. Niech $X_{i:j,k}$ oznacza część k-tej kolumny, od elementu i-tego do j-tego, macierzy X. Niech $X_{k,i:j}$ będzie analogiczną częścią k-tego wiersza tej macierzy.

(**a**) Pokazać, że równości (4.2.3) można wyrazić w postaci

$$U_{k,k+1:n} = (A_{k,k+1:n} - L_{k,1:k-1}M)/l_{kk},$$

gdzie M jest macierzą o wierszach $U_{i,k+1:n}$ dla $1 \leqslant i \leqslant k-1$.

(**b**) W podobny sposób przekształcić równości (4.2.4).

Obliczenia rozważane w tym zadaniu mają postać $y \leftarrow y - Mx$ i mogą być bardzo efektywnie wykonane na superkomputerze wektorowym. (Szczegóły – zob. Kincaid i Oppe [1988] oraz Oppe i Kincaid [1988]).

36. (**a**) Napisać *wersję wierszową* algorytmu Doolittle'a, obliczającą w k-tym kroku k-ty wiersz macierzy L i U (ten krok ma zatem dawać kolejno elementy $l_{k1}, l_{k2}, \ldots, l_{k,k-1}, u_{kk}, \ldots, u_{kn}$).

(**b**) Napisać *wersję kolumnową* tegoż algorytmu, która w k-tym kroku oblicza k-tą kolumnę macierzy U i L (ten krok ma zatem dawać kolejno elementy $u_{1k}, u_{2k}, \ldots, u_{kk}, l_{k+1,k}, \ldots, l_{nk}$).

37. Sprawdzić, czy jeśli macierz osobliwa ma rozkład Doolittle'a, to jest on jedyny.

38. Znaleźć wszystkie rozkłady Doolittle'a macierzy

$$A = \begin{bmatrix} 2 & 1 & -2 \\ 4 & 2 & -1 \\ 6 & 3 & 11 \end{bmatrix}.$$

W tym przykładzie algorytm działa, chociaż $u_{22} = 0$.

39. Jak wyraża się wyznacznik macierzy A, jeśli jest znany jej rozkład: (**a**) Doolittle'a, (**b**) Cholesky'ego?

40. Niech będzie

$$A = \begin{bmatrix} 25 & 0 & 0 & 0 & 1 \\ 0 & 27 & 4 & 3 & 2 \\ 0 & 54 & 58 & 0 & 0 \\ 0 & 108 & 116 & 0 & 0 \\ 100 & 0 & 0 & 0 & 24 \end{bmatrix}.$$

Znaleźć najogólniejszy rozkład LU tej macierzy z czynnikiem L jedynkowym. Wykazać, że metoda Doolittle'a daje jeden z tych rozkładów.

41. Korzystając z tw. 4.2.3, udowodnić równoważność: macierz symetryczna A jest dodatnio określona wtedy i tylko wtedy, gdy istnieje układ wektorów liniowo niezależnych $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ w $\mathbb{R}^n$ taki, że $a_{ij} = (x^{(i)})^{\mathsf{T}} x^{(j)}$.

42. Macierz A symetryczna i dodatnio określona ma pierwiastek kwadratowy X ($X^2 = A$) o tejże własności. Znaleźć X dla $A = \begin{bmatrix} 13 & 10 \\ 10 & 17 \end{bmatrix}$.

43. Macierz A o elementach a_{ij} równych zeru dla $j < i-1$ i dla $j > i$ jest nazywana *macierzą Stieltjesa*. Zaprojektować efektywny algorytm jej odwracania.

44. Opracować algorytm odwracania macierzy o elementach takich, że $a_{ij} = 0$ dla $i + j \leqslant n$.

45. Niech A będzie macierzą symetryczną o nieujemnych minorach głównych. Czy macierz $A + \varepsilon I$ dla $\varepsilon > 0$ ma tę samą własność?

46. Sprawdzić, czy macierz symetryczna jest dodatnio półokreślona wtedy i tylko wtedy, gdy wyznaczniki jej wszystkich minorów głównych są nieujemne.

Zadania komputerowe 4.2

K1. Rozważyć macierz symetryczną trójprzekątniową dodatnio określoną

$$A = \begin{bmatrix} 136.01 & 90.86 & 0.0 & 0.0 \\ 90.86 & 98.81 & -67.59 & 0.0 \\ 0.0 & -67.59 & 132.01 & 46.26 \\ 0.0 & 0.0 & 46.26 & 177.17 \end{bmatrix}.$$

Znaleźć jej następujące rozkłady: (**a**) LU, gdzie L jest jedynkowa; (**b**) LDU, gdzie L i U są jedynkowe; (**c**) LU, gdzie U jest jedynkowa; (**d**) LL^{T}.

K2. Opracować efektywny algorytm odwracania macierzy A trójkątnej dolnej stopnia n. Sugestia: Wykorzystać to, że ta odwrotność też jest trójkątna dolna. Zaprogramować algorytm i sprawdzić go dla macierzy o elementach $a_{ij} = (i+j)^2$ $(i \geqslant j)$ i $n = 10$. Dla kontroli odwrotności obliczyć AA^{-1}.

K3. Stosując rozkład Cholesky'ego, rozwiązać układ

$$0.05x_1 + 0.07x_2 + 0.06x_3 + 0.05x_4 = 0.23$$
$$0.07x_1 + 0.10x_2 + 0.08x_3 + 0.07x_4 = 0.32$$

$$0.06x_1 + 0.08x_2 + 0.10x_3 + 0.09x_4 = 0.33$$
$$0.05x_1 + 0.07x_2 + 0.09x_3 + 0.10x_4 = 0.31.$$

K4. Zaprogramować algorytm ogólnego rozkładu LU, zakładając, że tablica D zawiera dane elementy przekątniowe, a związana z nią tablica boolowska wskazuje, w którym z czynników L, U te liczby mają się znaleźć. Sprawdzić program dla kilku macierzy Hilberta o elementach $a_{ij} = (i+j-1)^{-1}$. W każdym przypadku wyznaczyć rozkłady Doolittle'a, Crouta i Cholesky'ego oraz jeden lub więcej rozkładów określonego wyżej typu.

4.3. Eliminacja Gaussa z wyborem elementów głównych

Opisany w podrozdz. 4.2 rozkład LU macierzy można traktować jako pewną wersję eliminacji Gaussa. Jej tradycyjna wersja będzie opisana niżej i powiązana z tamtą. Następnie zajmiemy się modyfikacjami koniecznymi do uzyskania zadowalających programów. Słowa „równanie" i „wiersz macierzy" układu będą używane zamiennie.

Dlaczego opisano rozkłady Doolittle'a, Crouta i Cholesky'ego, skoro eliminacja Gaussa dobrze działa? W epoce przedkomputerowej każda z tych procedur miała jakieś zalety w porównaniu z innymi. W miarę doskonalenia komputerów i ich oprogramowania te drobne różnice zanikały. Dlatego opis rozkładów Doolittle'a i Crouta ma przede wszystkim uzasadnienie historyczne. Natomiast algorytm Cholesky'ego działa szczególnie dobrze w odniesieniu do macierzy symetrycznych dodatnio określonych.

Podstawowa eliminacja Gaussa

Eliminację Gaussa poznamy na przykładzie prostego układu czterech równań z tyluż niewiadomymi:

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 34 \\ 27 \\ -38 \end{bmatrix}.$$

W pierwszym kroku postępowania odejmujemy stronami pierwsze równanie pomnożone przez 2 od drugiego, pomnożone przez $\frac{1}{2}$ odejmujemy od trzeciego i pomnożone przez -1 od czwartego. Liczby 2, $\frac{1}{2}$ i -1 nazywamy *mnożnikami* dla pierwszego kroku eliminacji, a liczbę 6 używaną jako dzielnik przy ich obliczaniu – *elementem głównym*. Po wykonaniu pierwszego

kroku układ przybiera postać

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & -12 & 8 & 1 \\ 0 & 2 & 3 & -14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ 21 \\ -26 \end{bmatrix}.$$

Pierwszy wiersz, chociaż użyty w obliczeniach, nie zmienił się. Jest to *wiersz główny* dla pierwszego kroku. W drugim kroku wierszem głównym jest drugi wiersz, a elementem głównym liczba -4. Odejmujemy ten wiersz pomnożony przez 3 od trzeciego i pomnożony przez $-\frac{1}{2}$ od czwartego. Liczby -4 i $-\frac{1}{2}$ są zatem mnożnikami w drugim kroku. Wynik jest następujący:

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 4 & -13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ -21 \end{bmatrix}.$$

W ostatnim kroku odejmujemy trzecie równanie pomnożone przez 2 od czwartego; element główny i mnożnik są równe 2:

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ -3 \end{bmatrix}. \tag{4.3.1}$$

Otrzymaliśmy układ o macierzy trójkątnej górnej, równoważny pierwotnemu w tym sensie, że oba układy mają to samo rozwiązanie. Nowy układ można łatwo rozwiązać, wyznaczając niewiadome od ostatniej do pierwszej. Daje to rozwiązanie

$$x = \begin{bmatrix} 1 \\ -3 \\ -2 \\ 1 \end{bmatrix}.$$

Mnożniki, których użyliśmy przekształcając układ, są elementami macierzy $L = (l_{ij})$ jedynkowej trójkątnej dolnej:

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ \frac{1}{2} & 3 & 1 & 0 \\ -1 & -\frac{1}{2} & 2 & 1 \end{bmatrix}.$$

Zauważmy, że każdy mnożnik występuje tu na pozycji tego zera w macierzy układu, do którego powstania posłużył. Końcowy układ (4.3.1) ma macierz trójkątną górną $U = (u_{ij})$:

$$U = \begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{bmatrix}.$$

Te dwie macierze dają rozkład LU macierzy A złożonej ze współczynników pierwotnego układu:

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ \frac{1}{2} & 3 & 1 & 0 \\ -1 & -\frac{1}{2} & 2 & 1 \end{bmatrix} \begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{bmatrix}.$$

Nietrudno wykazać, że tak musi być. Wiemy jak U powstaje z A. Odwróciwszy to postępowanie, możemy otrzymać A z U. Oznaczmy wiersze tych macierzy odpowiednio symbolami A_j i U_j. Eliminacja daje nam np. $U_2 = A_2 - 2A_1$. Stąd $A_2 = 2A_1 + U_2 = 2U_1 + U_2$. Współczynniki 2 i 1 znajdują się w drugim wierszu macierzy L. Podobnie, działania dające trzeci wiersz wyrażają się wzorem $U_3 = (A_3 - \frac{1}{2}A_1) - 3U_2$, czyli mamy $A_3 = \frac{1}{2}A_1 + 3U_2 + U_3 = \frac{1}{2}U_1 + 3U_2 + U_3$. Współczynniki $\frac{1}{2}$, 3 i 1 znajdują się więc w L w trzecim wierszu itd.

Aby ściśle opisać algorytm Gaussa, interpretujemy go jako sekwencję $n-1$ kroków dających ciąg macierzy:

$$A^{(1)} := A \to A^{(2)} \to \ldots \to A^{(n)}.$$

Macierz $A^{(k)}$ powstaje w $(k-1)$-szym kroku. Pokazano ją niżej. Linie obramowują k-ty wiersz i poprzedzają k-tą kolumnę, wyróżnione w tym kroku:

$$\left[\begin{array}{ccc|ccc|c} a_{11}^{(k)} & \ldots & a_{1,k-1}^{(k)} & a_{1k}^{(k)} & \ldots & a_{1j}^{(k)} & \ldots & a_{1n}^{(k)} \\ \vdots & \ddots & \vdots & \vdots & & \vdots & & \vdots \\ 0 & \ldots & a_{k-1,k-1}^{(k)} & a_{k-1,k}^{(k)} & \ldots & a_{k-1,j}^{(k)} & \ldots & a_{k-1,n}^{(k)} \\ \hline 0 & \ldots & 0 & a_{kk}^{(k)} & \ldots & a_{kj}^{(k)} & \ldots & a_{kn}^{(k)} \\ \hline 0 & \ldots & 0 & a_{k+1,k}^{(k)} & \ldots & a_{k+1,j}^{(k)} & \ldots & a_{k+1,n}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ 0 & \ldots & 0 & a_{ik}^{(k)} & \ldots & a_{ij}^{(k)} & \ldots & a_{in}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ 0 & \ldots & 0 & a_{nk}^{(k)} & \ldots & a_{nj}^{(k)} & \ldots & a_{nn}^{(k)} \end{array}\right].$$

Naszym celem jest wyjaśnić jak $A^{(k+1)}$ wynika z $A^{(k)}$. Aby otrzymać zera w k-tej kolumnie pod elementem głównym $a_{kk}^{(k)}$, odejmujemy odpowiednie wielokrotności k-tego wiersza od wierszy leżących niżej (wiersze $1, 2, \ldots, k$ nie zmieniają się):

$$a_{ij}^{(k+1)} := \begin{cases} a_{ij}^{(k)} & (i \leqslant k) \\ a_{ij}^{(k)} - (a_{ik}^{(k)}/a_{kk}^{(k)})a_{kj}^{(k)} & (i \geqslant k+1,\ j \geqslant k+1) \\ 0 & (i \geqslant k+1,\ j \leqslant k). \end{cases} \tag{4.3.2}$$

Dlatego przyjmujemy, że $U = A^{(n)}$ i określamy L wzorem

$$l_{ik} = \begin{cases} a_{ik}^{(k)}/a_{kk}^{(k)} & (i \geqslant k+1) \\ 1 & (i = k) \\ 0 & (i \leqslant k-1). \end{cases} \tag{4.3.3}$$

Równość $A = LU$ z macierzą jedynkową trójkątną dolną L i trójkątną górną U opisuje standardowy rozkład Gaussa macierzy A. Zarówno z (4.3.2) i (4.3.3), jak i z wcześniejszego przykładu numerycznego, wynika jasno, że proces eliminacji załamuje się, jeśli tylko któryś z elementów głównych znika. Możemy teraz udowodnić następujące twierdzenie:

Twierdzenie 4.3.1. *Jeśli wszystkie elementy główne $a_{kk}^{(k)}$ obliczane w opisany wyżej sposób są różne od 0, to $A = LU$.*

Dowód. Zauważmy, że $a_{ij}^{(k+1)} = a_{ij}^{(k)}$, jeśli $i \leqslant k$ lub $j \leqslant k-1$. Zauważmy też, że $u_{kj} = a_{kj}^{(n)} = a_{kj}^{(k)}$ i, na koniec, że $l_{ik} = 0$ dla $k > i$ i $u_{kj} = 0$ dla $k > j$. Niech teraz będzie $i \leqslant j$. Wtedy

$$\begin{aligned} (LU)_{ij} &= \sum_{k=1}^{n} l_{ik}u_{kj} = \sum_{k=1}^{i} l_{ik}u_{kj}^{(k)} = \sum_{k=1}^{i} l_{ik}a_{kj}^{(k)} = \\ &= \sum_{k=1}^{i-1} l_{ik}a_{kj}^{(k)} + l_{ii}a_{ij}^{(i)} = \sum_{k=1}^{i-1}(a_{ik}^{(k)}/a_{kk}^{(k)})a_{kj}^{(k)} + a_{ij}^{(i)} = \\ &= \sum_{k=1}^{i-1}(a_{ij}^{(k)} - a_{ij}^{(k+1)}) + a_{ij}^{(i)} = a_{ij}^{(1)} = a_{ij}. \end{aligned}$$

Podobnie, jeśli $i > j$, to

$$(LU)_{ij} = \sum_{k=1}^{j} l_{ik}a_{kj}^{(k)} = \sum_{k=1}^{j}(a_{ij}^{(k)} - a_{ij}^{(k+1)}) = a_{ij}^{(1)} - a_{ij}^{(j+1)} = a_{ij}^{(1)} = a_{ij},$$

gdyż $a_{ij}^{(k)} = 0$ dla $i \geqslant j+1$ i $k \geqslant j+1$. ∎

Algorytm opisanej już podstawowej eliminacji Gaussa jest – przy założeniu jak w tw. 4.3.1 – następujący:

input $n, (a_{ij})$
for $k = 1$ **to** $n - 1$ **do**
  for $i = k + 1$ **to** n **do**
    $z \leftarrow a_{ik}/a_{kk}$
    $a_{ik} \leftarrow 0$
    for $j = k + 1$ **to** n **do**
      $a_{ij} \leftarrow a_{ij} - z a_{kj}$
    end do
  end do
end do
output (a_{ij})

Mnożniki dobrano tak, żeby wyzerować wszystkie elementy leżące pod główną przekątną macierzy. Obliczać ich nie trzeba, wystarczy instrukcja $a_{ik} \leftarrow 0$.

Znaczenie elementów głównych

Algorytm Gaussa w najprostszej, już opisanej wersji nie jest zadowalający, gdyż zawodzi dla układów, które w istocie można łatwo rozwiązać. Aby to uzasadnić, rozważmy trzy proste przykłady. Oto pierwszy:

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Wspomniana wersja zawodzi, gdyż już pierwszy element główny jest równy 0 (zob. też zad. 4.2.10).

Rozważmy teraz układ

$$\begin{bmatrix} \varepsilon & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \tag{4.3.4}$$

gdzie ε jest małą liczbą różną od 0. Układ na pierwszy rzut oka nie powinien sprawiać takich kłopotów jak poprzedni. Stosując do (4.3.4) algorytm Gaussa, otrzymujemy następujący układ o macierzy trójkątnej górnej:

$$\begin{bmatrix} \varepsilon & 1 \\ 0 & 1 - \varepsilon^{-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 - \varepsilon^{-1} \end{bmatrix}.$$

Ma on rozwiązanie dokładne

$$x_2 = (2 - \varepsilon^{-1})/(1 - \varepsilon^{-1}), \qquad x_1 = (1 - x_2)\varepsilon^{-1}.$$

W obliczeniach komputerowych, gdy ε jest dostatecznie małe, wartością obu różnic $2-\varepsilon^{-1}$ i $1-\varepsilon^{-1}$ jest $-\varepsilon^{-1}$, czyli otrzymujemy $x_2=1$, a wobec tego $x_1=0$. Jednak rozwiązanie dokładne jest inne:

$$x_1=\frac{1}{1-\varepsilon}\approx 1, \qquad x_2=\frac{1-2\varepsilon}{1-\varepsilon}\approx 1.$$

Można zatem zaakceptować obliczone x_2, ale otrzymane x_1 jest zupełnie fałszywe!

Przypomnijmy, że przed odejmowaniem liczb zmiennopozycyjnych 2 i ε^{-1} trzeba wyrównać ich cechy, a to wymaga przesunięcia mantysy pierwszej liczby. Jeśli np. używany komputer pracuje w układzie dziesiętnym, a mantysy są siedmiocyfrowe i jeśli $\varepsilon=10^{-8}$, to

$$\varepsilon^{-1}=0.1000000_{10}9, \qquad 2=0.2000000_{10}1=0.000000002_{10}9,$$

$2-\varepsilon^{-1}=-0.099999998_{10}9$ i ostatecznym wynikiem odejmowania jest w komputerze $-0.1000000_{10}9=-\varepsilon^{-1}$.

Ostatni przykład pokazuje, iż kłopoty sprawia w istocie nie to, że współczynnik a_{11} jest mały, ale że jest on taki w porównaniu z innymi elementami pierwszego wiersza. Rozważmy układ równoważny poprzedniemu:

$$\begin{bmatrix}1 & \varepsilon^{-1}\\ 1 & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}\varepsilon^{-1}\\ 2\end{bmatrix}.$$

Zastosowanie zwykłej eliminacji Gaussa daje układ

$$\begin{bmatrix}1 & \varepsilon^{-1}\\ 0 & 1-\varepsilon^{-1}\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}\varepsilon^{-1}\\ 2-\varepsilon^{-1}\end{bmatrix}.$$

Rozwiązując go na komputerze, otrzymujemy znów $x_2=1$ i $x_1=0$; wartość x_1 i tym razem jest fałszywa.

Trudności, na jakie napotkaliśmy w tych przykładach, znikają po przestawieniu równań. W szczególności dla (4.3.4) daje to układ

$$\begin{bmatrix}1 & 1\\ \varepsilon & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}2\\ 1\end{bmatrix}.$$

Zastosujmy tu eliminację Gaussa:

$$\begin{bmatrix}1 & 1\\ 0 & 1-\varepsilon\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}2\\ 1-2\varepsilon\end{bmatrix}.$$

Obliczone x_2 jest znów równe 1, ale x_1 obliczamy ze wzoru $x_1 = 2 - x_2$, więc $x_1 = 1$.

Powyższe proste przykłady prowadzą do wniosku, że dobry algorytm musi uwzględniać przestawianie równań układu, gdy wymagają tego okoliczności. W istocie nie przenosimy wierszy macierzy w pamięci komputera, bo to przedłużałoby obliczenia. Zamiast tego wybieramy inaczej wiersze główne – nie w naturalnym porządku $1, 2, \ldots, n-1$, ale kolejno wiersze o wskaźnikach $p_1, p_2, \ldots, p_{n-1}$, gdzie $(p_1, p_2, \ldots, p_n)$ jest pewną permutacją zbioru $(1, 2, \ldots, n)$; jej wybór uzasadnimy nieco później. Tak więc w pierwszym kroku odejmujemy wielokrotności wiersza p_1 od wszystkich innych, o wskaźnikach $p_2, p_3, \ldots, p_n$. W drugim kroku odejmujemy wielokrotności wiersza p_2 od tych o wskaźnikach $p_3, \ldots, p_n$ itd.

Oto algorytm, który dla danej tablicy permutacji p wykonuje eliminację:

input $n, (a_{ij}), (p_i)$
for $k = 1$ **to** $n - 1$ **do**
  for $i = k + 1$ **to** n **do**
    $z \leftarrow a_{p_i k} / a_{p_k k}$
    $a_{p_i k} \leftarrow 0$
    for $j = k + 1$ **to** n **do**
      $a_{p_i j} \leftarrow a_{p_i j} - z a_{p_k j}$
    end do
  end do
end do
output (a_{ij})

Algorytm ten różni się od poprzedniego tylko pod jednym względem: pierwszy wskaźnik każdego elementu tablicy a zależy od wyboru permutacji. Oczywiście, dla permutacji identycznej ($p_i = i$) nowy algorytm redukuje się do poprzedniego.

Skalowany wybór wierszy głównych

Opiszemy teraz algorytm rozwiązywania układu $Ax = b$ z macierzą kwadratową A stopnia n, nazywany eliminacją Gaussa *ze skalowanym wyborem wierszy głównych*. Algorytm dzieli się na dwa etapy.

Pierwszym z nich jest rozkład stosowany tylko do macierzy A. Ściślej, wynikiem tego etapu jest znalezienie (w sposób opisany niżej) pewnej permutacji $(p_1, p_2, \ldots, p_n)$ zbioru $(1, 2, \ldots, n)$ i rozkładu LU iloczynu PA, gdzie macierz kwadratowa P jest taka, że $(P)_{ij} = \delta_{p_i j}$ (inaczej mówiąc, P powstaje z I przez przestawienie wierszy zgodnie z p). Rozkład $PA = LU$ otrzymujemy za pomocą algorytmu zmodyfikowanej eliminacji Gaussa, opi-

sanego w poprzednim fragmencie tego podrozdziału (z tą jednak różnicą, że permutacja p nie jest z góry dana).

W drugim etapie obliczeń rozwiązujemy kolejno dwa układy: $Lz = Pb$ i $Ux = z$. Pierwszy układ ma macierz trójkątną dolną, a jego prawa strona różni się od b porządkiem składowych. Z tym zastrzeżeniem mamy tu do czynienia z podstawianiem w przód opisanym już wcześniej. Rozwiązanie tego układu wpisujemy na miejscu wektora b. Natomiast podstawianie wstecz zastosowane do układu $Ux = b$ daje kolejno $x_n, x_{n-1}, \ldots, x_1$.

Trzeba teraz określić dokładnie sposób tworzenia permutacji p w pierwszym etapie obliczeń. Zaczynamy od wyznaczenia *skali* każdego wiersza, tj. wielkości

$$s_i = \max_{1 \leqslant j \leqslant n} |a_{ij}| \qquad (1 \leqslant i \leqslant n).$$

Zapamiętujemy je w tablicy s i korzystamy z nich w dalszym ciągu. Tworzymy też wstępny wariant $(1, 2, \ldots, n)$ permutacji $(p_1, p_2, \ldots, p_n)$.

W pierwszym kroku rozkładu wyznaczamy pierwszy wiersz główny – ten mianowicie, dla którego iloraz $|a_{i1}|/s_i$ jest największy. Niech jego wskaźnikiem będzie j. W tablicy p przestawiamy p_1 z p_j, czyli teraz p_1 będzie wskaźnikiem wybranego wiersza głównego. Stąd $|a_{p_1 1}|/s_{p_1} \geqslant |a_{i1}|/s_i$ dla $1 \leqslant i \leqslant n$. Następnie odejmujemy odpowiednie wielokrotności wiersza głównego od pozostałych wierszy macierzy A, aby wyzerować w nich elementy jej pierwszej kolumny. W dalszych obliczeniach wiersz p_1 nie zmienia się.

Ogólniej, przypuśćmy, że po $k-1$ przekształceniach macierzy A jesteśmy gotowi do zerowania elementów jej k-tej kolumny. Porównując liczby $|a_{p_i k}|/s_{p_i}$ dla $k \leqslant i \leqslant n$ wybieramy największą z nich. Niech j będzie wskaźnikiem takiej liczby. Przestawiamy p_k z p_j w tablicy p i odejmujemy wiersz p_k pomnożony przez $a_{p_i k}/a_{p_k k}$ od wiersza p_i $(k+1 \leqslant i \leqslant n)$.

Zobaczymy teraz, jak pierwszy etap obliczeń przebiega dla macierzy

$$A := \begin{bmatrix} 2 & 3 & -6 \\ 1 & -6 & 8 \\ 3 & -2 & 1 \end{bmatrix}.$$

Na początku jest $p = (1, 2, 3)$ i $s = (6, 8, 3)$. Aby wybrać pierwszy wiersz główny, porównujemy ilorazy $2/6$, $1/8$ i $3/3$. Największy jest trzeci z nich $(j = 3)$ i ten wiersz będzie pierwszym wierszem głównym. Dlatego przestawiamy p_1 z p_3, co daje $p = (3, 2, 1)$. Od wierszy pierwszego i drugiego odejmujemy takie wielokrotności trzeciego wiersza, aby w tamtych wyzerować elementy pierwszej kolumny. Przekształcona macierz A jest następująca:

$$\begin{bmatrix} \boxed{\frac{2}{3}} & \frac{13}{3} & -\frac{20}{3} \\ \boxed{\frac{1}{3}} & -\frac{16}{3} & \frac{23}{3} \\ 3 & -2 & 1 \end{bmatrix}.$$

Liczby w ramkach na pozycjach a_{11} i a_{21} są mnożnikami.

W następnym kroku wybór wiersza głównego zależy od liczb

$$|a_{p_2 2}|/s_{p_2} = \tfrac{16}{3}/8 = \tfrac{2}{3},$$
$$|a_{p_3 2}|/s_{p_3} = \tfrac{13}{3}/6 = \tfrac{13}{18}.$$

Większa jest ta ostatnia. Dlatego $j = 3$ i przestawiamy p_2 z p_3, co daje $p = (3, 1, 2)$. Nowa macierz jest następująca:

$$\begin{bmatrix} \boxed{\frac{2}{3}} & \frac{13}{3} & -\frac{20}{3} \\ \boxed{\frac{1}{3}} & \boxed{-\frac{16}{13}} & -\frac{7}{13} \\ 3 & -2 & 1 \end{bmatrix}.$$

Na miejscu a_{22} występuje w niej mnożnik użyty w przekształceniu.

Powyższe obliczenia dały macierz permutacji

$$P := \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

i rozkład LU macierzy PA:

$$PA = \begin{bmatrix} 3 & -2 & 1 \\ 2 & 3 & -6 \\ 1 & -6 & 8 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{2}{3} & 1 & 0 \\ \frac{1}{3} & -\frac{16}{13} & 1 \end{bmatrix} \begin{bmatrix} 3 & -2 & 1 \\ 0 & \frac{13}{3} & -\frac{20}{3} \\ 0 & 0 & -\frac{7}{13} \end{bmatrix}.$$

Algorytm rozkładu dla eliminacji Gaussa ze skalowanym wyborem wierszy głównych jest następujący:

input $n, (a_{ij})$
for $i = 1$ **to** n **do**
    $p_i \leftarrow i$
    $s_i \leftarrow \max_{1 \leqslant j \leqslant n} |a_{ij}|$
end do
for $k = 1$ **to** $n - 1$ **do**
    wybór takiego $j \geqslant k$, że
        $|a_{p_j k}|/s_{p_j} \geqslant |a_{p_i k}|/s_{p_i}$ dla $i = k, k+1, \ldots, n$
    $p_k \leftrightarrow p_j$

```
        for i = k + 1 to n do
            z ← a_{p_i k}/a_{p_k k}; a_{p_i k} ← z
            for j = k + 1 to n do
                a_{p_i j} ← a_{p_i j} − z a_{p_k j}
            end do
        end do
    end do
    output (a_{ij}), (p_i)
```

Zauważmy, że algorytm zapamiętuje mnożniki na miejscu zerowanych elementów macierzy A. Dzięki temu w A znajdują się wszystkie wielkości potrzebne do odtworzenia rozkładu LU. Stosując algorytm, musimy więc pamiętać, że niszczy on pierwotną macierz. Jeśli byłaby ona potrzebna później, to trzeba by zapamiętać ją w dodatkowej tablicy.

Aby rozwiązać układ $Ax = b$ po znalezieniu rozkładu $PA = LU$, stosujemy następujący algorytm:

```
input n, (a_{ij}), (p_i), (b_i)
for k = 1 to n − 1 do
    for i = k + 1 to n do
        b_{p_i} ← b_{p_i} − a_{p_i k} b_{p_k}
    end do
end do
for i = n to 1 step −1 do
    x_i ← (b_{p_i} − Σ_{j=i+1}^{n} a_{p_i j} x_j)/a_{p_i i}
end do
output (x_i)
```

Jego pierwsza część rozwiązuje układ $Lz = Pb$ z macierzą trójkątną dolną. Druga część oblicza x z układu $Ux = z$ o macierzy trójkątnej górnej. W obu przypadkach trzeba oczywiście używać tablicy permutacji p.

Pełny wybór elementów głównych

Częściowy wybór elementów głównych polega na tym, że k-ty z nich jest tym spośród $n - k + 1$ elementów dolnej części k-tej kolumny macierzy $A^{(k)}$, który ma największą wartość bezwzględną. To określa wiersz główny. W wyborze *skalowanym*, opisanym wcześniej, zamiast elementów porównuje się ich ilorazy przez czynniki skalujące s_j. Inną strategią jest *pełny wybór elementów głównych*. W tym przypadku badamy $(n - k + 1)^2$ elementów w dolnej prawej części macierzy $A^{(k)}$, ewentualnie z uwzględnieniem skal. Na ogół uważa się, że pełny wybór, oczywiście bardziej kosztowny, nie ma istotnych zalet w porównaniu z wyborem częściowym.

Rozkłady $PA = LU$

Jak już wiemy, (skalowany) wybór wierszy głównych włączony do algorytmu Gaussa daje rozkład LU macierzy PA, gdzie P jest pewną macierzą permutacji. Poznamy teraz formalny dowód tego faktu, wzorowany na dowodzie tw. 4.3.1. Dowód nie zależy od strategii użytej do wyboru elementów głównych.

Niech $p_1, p_2, \ldots, p_n$ będą wskaźnikami kolejnych wierszy głównych. Dla $A^{(1)} := A$ określamy rekurencyjnie macierze $A^{(2)}, A^{(3)}, \ldots, A^{(n)}$ wzorem

$$a_{p_i j}^{(k+1)} = \begin{cases} a_{p_i j}^{(k)} & (i \leqslant k \text{ lub } i > k > j) \\ a_{p_i j}^{(k)} - \left(a_{p_i k}^{(k)} \,/\, a_{p_k k}^{(k)}\right) a_{p_k j}^{(k)} & (i > k,\ j > k) \\ a_{p_i k}^{(k)} \,/\, a_{p_k k}^{(k)} & (i > k,\ j = k). \end{cases}$$

Twierdzenie 4.3.2. *Niech macierze: permutacji P, trójkątna górna U i jedynkowa trójkątna dolna L będą określone odpowiednio wzorami $(P)_{ij} := \delta_{p_i j}$, $u_{ij} := a_{p_i j}^{(n)}$ dla $j \geqslant i$ oraz $l_{ij} := a_{p_i j}^{(n)}$ dla $j < i$. Wtedy $PA = LU$.*

Dowód. Z definicji macierzy $A^{(k)}$ wynika, że

$$u_{kj} = a_{p_k j}^{(n)} = a_{p_k j}^{(k)} \qquad (j \geqslant k),$$
$$l_{ik} = a_{p_i k}^{(n)} = a_{p_i k}^{(k+1)} = a_{p_i k}^{(k)} / a_{p_k k}^{(k)} \qquad (i \geqslant k).$$

Istotnie, wiersz p_k macierzy ustala się w k-tym kroku, a k-tą kolumnę – w $(k+1)$-szym kroku. Powyższy wzór dla l_{ik} jest poprawny dla $i = k$, bo daje wtedy wartość 1. Załóżmy teraz, że $i \leqslant j$. Wtedy

$$(LU)_{ij} = \sum_{k=1}^{i} l_{ik} u_{kj} = \sum_{k=1}^{i-1} \left(a_{p_i k}^{(k)} / a_{p_k k}^{(k)}\right) a_{p_k j}^{(k)} + l_{ii} a_{p_i j}^{(i)} =$$
$$= \sum_{k=1}^{i-1} \left(a_{p_i j}^{(k)} - a_{p_i j}^{(k+1)}\right) + a_{p_i j}^{(i)} = a_{p_i j}^{(1)} = a_{p_i j}.$$

Jeśli natomiast $i > j$, to

$$(LU)_{ij} = \sum_{k=1}^{j} l_{ik} u_{kj} = \sum_{k=1}^{j-1} \left(a_{p_i k}^{(k)} / a_{p_k k}^{(k)}\right) a_{p_k j}^{(k)} + \left(a_{p_i j}^{(j)} / a_{p_j j}^{(j)}\right) a_{p_j j}^{(j)} =$$
$$= \sum_{k=1}^{j-1} \left(a_{p_i j}^{(k)} - a_{p_i j}^{(k+1)}\right) + a_{p_i j}^{(j)} = a_{p_i j}^{(1)} = a_{p_i j}.$$

Z drugiej strony

$$(PA)_{ij} = \sum_{k=1}^{n} (P)_{ik} a_{kj} = \sum_{k=1}^{n} \delta_{p_i k} a_{kj} = a_{p_i j}.$$

W ten sposób wykazaliśmy, że dla wszystkich par (i, j) jest $(PA)_{ij} = (LU)_{ij}$. ■

Koszt obliczeń

Aby oszacować koszt rozwiązywania układu równań liniowych, obliczymy, ilu działań arytmetycznych wymagają etapy rozkładu i rozwiązywania. Ponieważ czas wykonania mnożenia i dzielenia jest zwykle podobny, ale wyraźnie dłuższy niż dla dodawania i odejmowania, przyjęło się mierzyć koszt obliczeń liczbą *długich działań* (w skrócie op) czyli par mnożenie–dodawanie itp. Rozważmy pierwszy krok ($k = 1$) w procesie rozkładu; dotyczy on macierzy stopnia n. Określenie wskaźnika p_1 wiersza głównego wymaga n dzieleń (n op). Dla każdego z $n - 1$ wierszy o wskaźnikach $p_2, \ldots, p_n$ obliczamy mnożnik (1 op) i odejmujemy wielokrotność wiersza p_1 od wiersza p_i ($2 \leqslant i \leqslant n$). Zera w pierwszej kolumnie nie obliczamy. Tak więc utworzenie jednego nowego wiersza kosztuje n op, a $n - 1$ wierszy kosztuje $n(n-1)$ op. Łącznie z obliczeniem p_1 daje to n^2 op.

Dalszy ciąg rozkładu można uważać za powtórzenie pierwszego kroku, ale dla macierzy coraz niższego stopnia. Dlatego cały rozkład wymaga

$$n^2 + (n-1)^2 + \ldots + 3^2 + 2^2 = \tfrac{1}{3}n^3 + \tfrac{1}{2}n^2 + \tfrac{1}{6}n - 1 \approx \tfrac{1}{3}n^3 + \tfrac{1}{2}n^2$$

długich działań.

Analiza drugiego etapu algorytmu pokazuje, że rozwiązywanie układu $Lz = Pb$ rozpada się na $n - 1$ kroków. W pierwszym z nich wykonujemy $n - 1$ długich działań, w drugim $n - 2$ itd., czyli łącznie

$$(n-1) + (n-2) + \ldots + 1 = \tfrac{1}{2}n^2 - \tfrac{1}{2}n.$$

Rozwiązywanie układu $Ux = z$ wymaga w kolejnych krokach jednego, dwóch, ..., n długich działań, a więc łącznie jest ich

$$1 + 2 + 3 + \ldots + n = \tfrac{1}{2}n^2 + \tfrac{1}{2}n,$$

a w całym drugim etapie n^2 op. Uwzględniając to wszystko, formułujemy następujący nieco ogólniejszy wynik:

Twierdzenie 4.3.3. *Jeśli eliminację Gaussa wykonano ze skalowanym wyborem elementów głównych, to rozwiązanie m układów $Ax = b$ o wspólnej macierzy A i m różnych wektorach b wymaga wykonania około*

$$\tfrac{1}{3}n^3 + (m + \tfrac{1}{2})n^2$$

długich działań.

Aby uzasadnić tezę tego twierdzenia, zauważmy, że jeśli mamy rozwiązać m układów $Ax^{(i)} = b^{(i)}$ dla $i = 1, 2, \ldots, m$, to rozkład $PA = LU$ wykonujemy tylko raz. Układy z macierzami trójkątnymi trzeba już rozwiązywać oddzielnie i to kosztuje mn^2 op. Stąd wynika liczba podana w twierdzeniu. Zauważmy, że odwrotność A^{-1} można obliczyć rozwiązując n układów $Ax^{(i)} = e_i$ (daje to jej poszczególne kolumny). Na mocy tw. 4.3.3 dla $m = n$ koszt tych obliczeń wynosi $\mathcal{O}(\frac{4}{3}n^3)$ op. Oczywiście, aby rozwiązać układ $Ax = b$, nie należy obliczać odwrotności A^{-1} – byłoby to zbyt kosztowne!

Macierze dominujące przekątniowo

Dla pewnych macierzy eliminacja Gaussa może być bezpiecznie wykonywana bez wyboru elementów głównych. Jest tak w szczególności dla *macierzy dominujących przekątniowo*, tzn. takich, że

$$|a_{ii}| > \sum_{j=1,\, j\neq i}^{n} |a_{ij}| \qquad (1 \leqslant i \leqslant n). \tag{4.3.5}$$

Oto przykład macierzy o tej własności:

$$\begin{bmatrix} 4 & -1 & 0 & -1 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix}.$$

Takie macierze wynikają w naturalny sposób z dyskretyzacji równań różniczkowych cząstkowych, wiążą się też z funkcjami sklejanymi itd.

Jeśli macierz układu równań ma określoną wyżej własność, to w pierwszym kroku eliminacji Gaussa wierszem głównym może być pierwszy wiersz, gdyż wtedy na mocy (4.3.5) element główny a_{11} jest różny od 0. Chcemy oczywiście wiedzieć, czy w drugim kroku taką samą rolę może odgrywać drugi wiersz. Poniższe twierdzenie wyjaśnia tę wątpliwość.

Twierdzenie 4.3.4. *Eliminacja Gaussa bez wyboru elementów głównych zachowuje dominację przekątniową macierzy.*

Dowód. Wystarczy rozważyć skutki wykonania pierwszego kroku eliminacji Gaussa, gdyż następne kroki są podobne, choć dotyczą macierzy niższych stopni. Zakładamy więc, że A jest macierzą stopnia n, dominującą przekątniowo. Ponieważ pierwszy krok tworzy zera w pierwszej kolumnie i nie zmienia tegoż wiersza, więc trzeba wykazać, że dla $i = 2, 3, \ldots, n$ jest

$$|a_{ii}^{(2)}| > \sum_{j=2,\, j\neq i}^{n} |a_{ij}^{(2)}|.$$

Wyrażamy to przez elementy macierzy A:

$$|a_{ii} - (a_{i1}/a_{11})a_{1i}| > \sum_{j=2,\, j\neq i}^{n} |a_{ij} - (a_{i1}/a_{11})a_{1j}|.$$

Zamiast tej nierówności udowodnimy mocniejszą:

$$|a_{ii}| - |(a_{i1}/a_{11})a_{1i}| > \sum_{j=2,\, j\neq i}^{n} [|a_{ij}| + |(a_{i1}/a_{11})a_{1j}|].$$

Jest ona równoważna nierówności

$$|a_{ii}| - \sum_{j=2,\, j\neq i}^{n} |a_{ij}| > \sum_{j=2}^{n} |(a_{i1}/a_{11})a_{1j}|.$$

Z dominacji przekątniowej w i-tym wierszu wynika, że

$$|a_{ii}| - \sum_{i=2,\, j\neq i}^{n} |a_{ij}| > |a_{i1}|,$$

więc wystarczy udowodnić, że

$$|a_{i1}| \geqslant \sum_{j=2}^{n} |(a_{i1}/a_{11})a_{1j}|.$$

Tak istotnie jest dzięki dominacji przekątniowej w pierwszym wierszu:

$$|a_{11}| > \sum_{j=2}^{n} |a_{1j}| \implies 1 > \sum_{j=2}^{n} |a_{1j}/a_{11}|. \qquad \blacksquare$$

Wniosek 4.3.5. *Każda macierz dominująca przekątniowo jest nieosobliwa i ma rozkład LU.*

Dowód. Na mocy tw. 4.3.4 i 4.3.1 macierz A dominująca przekątniowo ma rozkład LU, gdzie L jest jedynkowa trójkątna dolna. Zgodnie z ostatnim twierdzeniem macierz U jest dominująca przekątniowo, więc jej elementy przekątniowe są różne od zera. Dlatego L i U są nieosobliwe. ■

Wniosek 4.3.6. *Niech macierz będzie dominująca przekątniowo. Jeśli w eliminacji Gaussa ze skalowanym wyborem wierszy głównych tablicę skal oblicza się na nowo po każdym kroku, to wiersze główne mają naturalny porządek:* $1, 2, \ldots, n$, *czyli można pominąć czynności określające wybór tych wierszy.*

Dowód. Wobec tw. 4.3.4 wystarczy wykazać, że pierwszy wiersz główny ma wskaźnik 1, tj. że

$$|a_{11}|/s_1 > |a_{i1}|/s_i \qquad (2 \leqslant i \leqslant n).$$

Dzięki dominacji przekątniowej wiemy, że $|a_{ii}| = \max_j |a_{ij}| = s_i$ dla każdego i. Stąd $|a_{11}|/s_1 = 1$. Dla $i \geqslant 2$

$$|a_{i1}| \leqslant \sum_{j=1,\, j\neq i}^{n} |a_{ij}| < |a_{ii}| = s_i,$$

czyli $|a_{i1}|/s_i < 1$. ■

W twierdzeniu 4.2.3 wykazano, że macierz A symetryczna i dodatnio określona ma jedyny rozkład Cholesky'ego: $A = LL^{\mathsf{T}}$ (jest to szczególny przypadek rozkładu LU). W tym przypadku wybór wierszy głównych jest zbędny, gdyż dzięki nierówności $|l_{ij}| \leqslant \sqrt{a_{ii}}$ udowodnionej w podrozdz. 4.2 elementy macierzy L są stosunkowo małe w porównaniu z elementami A.

Układy trójprzekątniowe

Macierze układów występujących w zastosowaniach mają często szczególną strukturę. Rozwiązując takie układy, warto zazwyczaj stosować algorytmy, które biorą ją pod uwagę. Rozważymy jako przykład układ o macierzy A trójprzekątniowej, tj. takiej, że $a_{ij} = 0$ dla $|i - j| > 1$. Wtedy w i-tym wierszu dla $1 < i < n$ tylko trzy elementy ($a_{i,i-1}$, a_{ii} i $a_{i,i+1}$) są różne od 0, a w pierwszym i ostatnim wierszu tylko dwa. Do zapamiętania tych niezerowych elementów wystarczy użyć trzech wektorów i układ o takiej macierzy wyrażamy w następujący sposób:

$$\begin{bmatrix} d_1 & c_1 & & & & \\ a_1 & d_2 & c_2 & & & \\ & a_2 & d_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & a_{n-2} & d_{n-1} & c_{n-1} \\ & & & & a_{n-1} & d_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix}. \tag{4.3.6}$$

Elementy macierzy nie pokazane wyżej są zerami.

Załóżmy, że ta macierz nie wymaga stosowania wyboru elementów głównych, bo np. jest symetryczna dodatnio określona. Używamy wtedy zwykłej eliminacji Gaussa z dodatkiem jednoczesnego przetwarzania prawej strony (wektora b). W pierwszym kroku od wiersza 2 odejmujemy taką wielokrotność wiersza 1, żeby wyzerować element na pozycji zajmowanej dotąd przez a_1. Zauważmy, że d_2 i b_2 zmieniają się, ale nie c_2. Mnożnikiem jest a_1/d_1. Tak więc w tym kroku wykonujemy podstawienia

$$d_2 \leftarrow d_2 - (a_1/d_1)c_1, \qquad b_2 \leftarrow b_2 - (a_1/d_1)b_1.$$

Wszystkie następne kroki eliminacji w przód są dokładnie takie same. Podstawianie wstecz zaczyna się od czynności

$$x_n \leftarrow b_n/d_n,$$

drugim krokiem jest

$$x_{n-1} \leftarrow (b_{n-1} - c_{n-1}x_n)/d_{n-1},$$

pozostałe podstawienia są podobne. Kompletny algorytm (o nazwie `tri`) jest następujący:

input $n, (a_i), (b_i), (c_i), (d_i)$
for $i = 2$ **to** n **do**
    $d_i \leftarrow d_i - (a_{i-1}/d_{i-1})c_{i-1}$
    $b_i \leftarrow b_i - (a_{i-1}/d_{i-1})b_{i-1}$
end do
$x_n \leftarrow b_n/d_n$
for $i = n - 1$ **to** 1 **step** -1 **do**
    $x_i \leftarrow (b_i - c_i x_{i+1})/d_i$
end do
output (x_i)

ZADANIA 4.3 [11)]

1. Napisać algorytm rozwiązywania takiego układu $Ax = b$, że dla danej permutacji p zbioru $(1, 2, \ldots, n)$ i dla każdego i:
   (a) równanie p_i-te zawiera tylko niewiadomą x_i,
   (b) równanie i-te zawiera tylko niewiadomą x_{p_i},
   (c) równanie p_i-te zawiera tylko zmienną x_{p_i}.

2. Napisać algorytm rozwiązywania układu $Ax = b$ poprawny przy założeniu, że dla danych permutacji p i q zbioru $(1, 2, \ldots, n)$ i dla każdego i:
   (a) równanie p_i-te zawiera tylko niewiadomą x_{q_i},
   (b) zmienne $x_{q_1}, x_{q_2}, \ldots, x_{q_{i-1}}$ nie występują w p_i-tym równaniu,
   (c) zmienna x_{q_i} występuje tylko w równaniach o wskaźnikach $p_1, p_2, \ldots, p_i$.

3. Wykazać, że wzór (4.3.2) definiujący eliminację Gaussa można również wyrazić w postaci

$$a_{ij}^{(k+1)} := \begin{cases} a_{ij}^{(k)} & (i \leqslant k \text{ lub } j < k) \\ a_{ij}^{(k)} - (a_{ik}^{(k)}/a_{kk}^{(k)})a_{kj}^{(k)} & (i > k,\ j \geqslant k). \end{cases}$$

4. Układ $Ax = b$ warto niekiedy zmodyfikować, wprowadzając nowe zmienne $y_i = d_i x_i$, gdzie d_i są dodatnie. Jeśli x_i są wielkościami fizycznymi, to oznacza to zmianę jednostek, w których wyrażają się x_i. Jeśli np. x_1 ma być mierzone nie w centymetrach, ale w metrach, to $y_1 = 10^{-2}x_1$. W symbolice macierzowej mamy tu wzór $y = Dx$, gdzie $D = \operatorname{diag}(d_i)$. Nowym układem równań jest $AD^{-1}y = b$. Jeśli $d_j = \max_{1 \leqslant i \leqslant n} |a_{ij}|$, to przekształcenie nazywamy *wyważeniem kolumn*. Włączyć je do obu etapów algorytmu rozwiązywania układu $Ax = b$.

5. Jeśli znamy czynnik U rozkładu LU macierzy A, to jaki algorytm daje L?

6. Rozwiązać dwukrotnie każdy z podanych układów, stosując najpierw zwykłą eliminację Gaussa i uzyskując rozkład $A = LU$, a następnie stosując ten algorytm, ale ze skalowanym wyborem wierszy głównych, co ma dać rozkład $PA = LU$.

(a) $$\begin{bmatrix} -1 & 1 & -4 \\ 2 & 2 & 0 \\ 3 & 3 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ \frac{1}{2} \end{bmatrix}$$

(b) $$\begin{bmatrix} 1 & 6 & 0 \\ 2 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix}$$

[11)] We wszystkich zadaniach tego podrozdziału $p = (p_1, p_2, \ldots, p_n)$ jest pewną permutacją zbioru $(1, 2, \ldots, n)$, a P odpowiadającą jej macierzą permutacji, o elementach $(P)_{ij} = \delta_{p_i j}$ (*przyp. tłum.*).

(c) $\begin{bmatrix} -1 & 1 & 0 & -3 \\ 1 & 0 & 3 & 1 \\ 0 & 1 & -1 & -1 \\ 3 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \\ 3 \\ 1 \end{bmatrix}$

(d) $\begin{bmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 4 & 10 \\ 3 & -13 & 3 & 3 \\ -6 & 4 & 2 & -18 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ -10 \\ -39 \\ -16 \end{bmatrix}$

(e) $\begin{bmatrix} 1 & 0 & 2 & 1 \\ 4 & -9 & 2 & 1 \\ 8 & 16 & 6 & 5 \\ 2 & 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 2 \\ 14 \\ -3 \\ 0 \end{bmatrix}$

7. Niech A będzie dowolną macierzą stopnia n. Jak wyrażają się macierze: PA, AP, P^{-1} i PAP^{-1}?

8. Niech dla macierzy A stopnia n czynniki skalujące $s_i := \max_{1 \leqslant j \leqslant n} |a_{ij}|$ będą dodatnie. Niech B będzie macierzą o elementach a_{ij}/s_j. Udowodnić, że eliminacja w przód zastosowana do A i B daje tę samą macierz L. Znaleźć wzory wiążące końcowe A i B (po eliminacji).

9. Wykazać, że w eliminacji Gaussa z pełnym wyborem elementów głównych mnożniki należą do przedziału $[-1, 1]$; zob. zad. K1.

10. Niech eliminacja w przód zastosowana do macierzy A stopnia n daje macierz B i permutację p. Udowodnić, że rozkład LU iloczynu PA można otrzymać tak: dla $C = PB$ przyjmujemy $(L)_{ij} = (C)_{ij}$ dla $j < i$ i $(U)_{ij} = (C)_{ij}$ dla $i \leqslant j$ (prócz tego $(L)_{ij} = 0$ dla $j > i$, $(L)_{ii} = 1$, $(U)_{ij} = 0$ dla $i > j$.)

11. Wykonać eliminację w przód ze skalowanym wyborem wierszy głównych dla każdej z macierzy

(a) $\begin{bmatrix} 2 & -2 & -4 \\ 1 & 1 & -1 \\ 3 & 7 & 5 \end{bmatrix}$, (b) $\begin{bmatrix} 3 & 7 & 3 \\ 1 & \frac{7}{3} & 4 \\ 4 & \frac{4}{3} & 0 \end{bmatrix}$, (c) $\begin{bmatrix} -9 & 1 & 17 \\ 3 & 2 & -1 \\ 6 & 8 & 1 \end{bmatrix}$, (d) $\begin{bmatrix} 1 & -2 & 3 \\ 2 & -4 & 2 \\ 3 & -5 & -1 \end{bmatrix}$.

Podać skale s_i, końcową permutację p i końcową tablicę A, z mnożnikami na właściwych miejscach.

12. Zaprojektować algorytm rozwiązywania układu $Ax = b$ za pomocą eliminacji Gaussa, bez wyboru elementów głównych, dostosowany do przypadku, gdy $a_{ij} = 0$ dla $i > j + 1$. Znaleźć liczbę wykonywanych operacji.

13. Znaleźć liczbę operacji dla algorytmu w tekście, rozwiązującego układ o macierzy trójprzekątniowej.

14. Zmodyfikować wspomniany wyżej algorytm, odwracając kolejność przekształcania równań i obliczania niewiadomych.

15. Rozwiązać układ $Ax = b$ dla $b = (100, 1)^\top$ i dla każdej z macierzy

$$A_1 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 & 1 \\ 1 & 0.01 \end{bmatrix}$$

(Stoer i Bulirsch [1980, s. 185]). To zadanie pokazuje, że rozwiązanie układu równań może być *niestabilne* względem zaburzeń danych.

16. Jakie jest rozwiązanie układu

$$\varepsilon x_1 + 2x_2 = 4, \quad x_1 - x_2 = -1$$

otrzymanego numerycznie za pomocą algorytmu Gaussa bez wyboru elementów głównych, jeśli ε jest dostatecznie małe w porównaniu z precyzją arytmetyki?

17. Rozwiązać układ

$$\begin{bmatrix} -9 & 1 & 17 \\ 3 & 2 & -1 \\ 6 & 8 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 9 \\ -3 \end{bmatrix},$$

stosując eliminację Gaussa z pełnym wyborem elementów głównych (por. zad. K1).

18. Rozwiązać układ

$$\begin{aligned} 0.2641x_1 + 0.1735x_2 + 0.8642x_3 &= -0.7521 \\ 0.9411x_1 + 0.0175x_2 + 0.1463x_3 &= \ \ 0.6310 \\ -0.8641x_1 - 0.4243x_2 + 0.0711x_3 &= \ \ 0.2501 \end{aligned}$$

za pomocą eliminacji Gaussa: (**a**) bez wyboru elementów głównych, (**b**) ze skalowanym wyborem wierszy głównych.

19. (**a**) Wykazać, że eliminacja Gaussa bez wyboru elementów głównych, zastosowana do macierzy symetrycznej A, daje $l_{i1} = a_{1i}/a_{11}$.

(**b**) Stąd wywnioskować, że po usunięciu pierwszego wiersza i tejże kolumny z $A^{(2)}$ otrzymujemy macierz symetryczną stopnia $n-1$, a więc nie trzeba obliczać jej elementów pod główną przekątną. Wykazać przez indukcję, że to uproszczenie jest możliwe w każdym z następnych kroków obliczania rozkładu.

(**c**) Wykazać, że koszt opisanych wyżej obliczeń jest prawie dwukrotnie mniejszy niż dla macierzy niesymetrycznych.

(**d**) Wykorzystać to uproszczenie, rozwiązując układ

$$\begin{aligned} 0.6428x_1 + 0.3475x_2 - 0.8468x_3 &= \ \ 0.4127 \\ 0.3475x_1 + 1.8423x_2 + 0.4759x_3 &= \ \ 1.7321 \\ -0.8468x_1 + 0.4759x_2 + 1.2147x_3 &= -0.8621. \end{aligned}$$

20. Dla macierzy

$$\begin{bmatrix} 0 & 4 & 25 & 79 \\ 9 & 7 & 39 & 89 \\ 0 & 16 & 2 & 99 \\ 0 & 6 & 6 & 49 \end{bmatrix}$$

wskazać element, który będzie użyty jako pierwszy element główny w eliminacji Gaussa ze skalowanym wyborem wierszy głównych, gdy tablicą skal jest $s = (80, 89, 160, 30)$.

21. Podać macierz, która powstaje po zastosowaniu eliminacji w przód ze skalowanym wyborem wierszy głównych do macierzy

$$\begin{bmatrix} 2 & -2 & -4 \\ 1 & 1 & -1 \\ 3 & 7 & 5 \end{bmatrix}.$$

Wynik ma zawierać mnożniki na odpowiednich pozycjach.

22. Znaleźć wyznacznik

$$\begin{vmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{vmatrix},$$

nie korzystając z jego rozwinięcia Laplace'a (czyli wyrażenia przez wyznaczniki niższego stopnia).

23. Stosując eliminację Gaussa ze skalowanym wyborem elementów głównych do macierzy:

$$A = \begin{bmatrix} 3 & 2 & -1 \\ 6 & 6 & 2 \\ -1 & 1 & 3 \end{bmatrix},$$

otrzymać rozkład $PA = LDU$, gdzie L, U i D są odpowiednio macierzami: jedynkową trójkątną dolną, jedynkową trójkątną górną i przekątniową.

24. Znaleźć algorytm rozwiązywania układu $Ax = b$ przy założeniu, że tylko te elementy a_{ij} są różne od 0, dla których $|i - j| \leqslant 1$ lub $(i, j) = (1, n)$, lub $(i, j) = (n, 1)$. Zastosować eliminację Gaussa bez wyboru elementów głównych.

25. Niech A będzie macierzą stopnia n, dominującą przekątniowo w kolumnach, tj. taką, że

$$\sum_{i=1,\, i\neq j}^{n} |a_{ij}| < |a_{jj}| \qquad (1 \leqslant j \leqslant n).$$

Sprawdzić, czy eliminacja Gaussa bez wyboru elementów głównych zachowuje tę własność.

26. Dla macierzy A dominującej przekątniowo *nadwyżka w i-tym wierszu* jest określona wzorem

$$e_i = |a_{ii}| - \sum_{j=1,\, j\neq i}^{n} |a_{ij}|.$$

Wykazać, że w dowodzie tw. 4.3.4 jest

$$|a_{ii} - a_{i1}a_{1i}/a_{11}| \geqslant \sum_{j=2,\, j\neq i}^{n} |a_{ij} - a_{i1}a_{1j}/a_{11}| + e_1,$$

czyli nie zmniejsza się ona w eliminacji Gaussa.

27. Wykazać, że $P^{-1} = P^{\mathsf{T}}$.

28. Macierze A i B mają odpowiednio wymiary $n \times n$ i $n \times m$. Ilu mnożeń i dzieleń wymaga rozwiązanie układu $AX = B$ za pomocą eliminacji Gaussa ze skalowanym wyborem wierszy głównych? To samo pytanie dla $B = I$.

29. Załóżmy, że skale są obliczane na nowo przed każdym krokiem eliminacji Gaussa ze skalowanym wyborem wierszy głównych. Udowodnić, że dla macierzy symetrycznej i dominującej przekątniowo te same wyniki daje eliminacja bez wyboru elementów głównych.

30. Niech skale będą określone wzorem $s_i := \sum_{j=1}^{n} |a_{ij}|$. Udowodnić, że wtedy eliminacja Gaussa ze skalowanym wyborem elementów głównych, zastosowana do macierzy dominującej przekątniowo, daje takie same wyniki jak bez tego wyboru.

31. Sprawdzić, czy eliminacja Gaussa bez wyboru elementów głównych zachowuje własność macierzy opisaną nierównościami

$$0 \neq |a_{ii}| \geqslant \sum_{j=1,\, j\neq i}^{n} |a_{ij}| \qquad (1 \leqslant i \leqslant n).$$

32. (**a**) Udowodnić, że obliczenie wyznacznika za pomocą jego rozwinięcia Laplace'a wymaga wykonania $(n-1)n!$ op.

(**b**) Udowodnić, że rozwiązując układ n równań liniowych za pomocą wzorów Cramera, trzeba wykonać $(n^2-1)n!$ op.

(**c**) *Metoda Gaussa-Jordana* (w najprostszej wersji, tj. bez wyboru elementów głównych) polega na tym, że w k-tym kroku wielokrotności k-tego wiersza są odejmowane od wszystkich innych wierszy; mnożniki wybieramy tak, aby wyzerować tam składniki z x_k. Dlatego metoda daje układ o macierzy przekątniowej, a nie trójkątnej górnej, jak w eliminacji Gaussa. Udowodnić, że ta metoda wymaga wykonania $\frac{1}{2}n(n+1)^2 \approx \frac{1}{2}n^3$ op, czyli jest o ok. 50% bardziej kosztowna od eliminacji Gaussa.

33. Dla macierzy trójprzekątniowej A, jak w (4.3.6), przyjmujemy dodatkowo, że $c_0 = 0$ i $a_n = 0$. Wykazać, że jeśli A jest *dominująca przekątniowo w kolumnach*, tj. jeśli

$$|d_i| > |a_i| + |c_{i-1}| \qquad (1 \leqslant i \leqslant n),$$

to zwykła eliminacja Gaussa jest – w teorii – skuteczna, gdyż żaden element główny nie znika.

34. Sprawdzić, czy jeśli macierz A jest trójprzekątniowa, to PAP^{-1} ma tę samą własność.

Zadania komputerowe 4.3

K1. Napisać program eliminacji Gaussa z pełnym wyborem elementów głównych; będą w nim potrzebne dwa wektory permutacji.

K2. W związku z zad. 19 napisać program obliczający rozkład macierzy symetrycznej bez wyboru elementów głównych.

K3. Napisać program eliminacji Gaussa ze skalowanym wyborem wierszy głównych i sprawdzić go dla macierzy z zadań 15, 17, 18 i 19d.

K4. Napisać i sprawdzić programy uwzględniające w eliminacji Gaussa wyważanie kolumn (zob. zad. 4).

K5. Napisać i sprawdzić programy rozwiązujące układy $Ax = b$ i $y^T A = c^T$. Należy ograniczyć się do jednego rozkładu (macierzy A, ze skalowanym wyborem elementów głównych). Dwa oddzielne podprogramy mają obliczać x i y.

K6. Napisać i sprawdzić program rozwiązujący układ $Ax = b$. Zastosować wyważanie kolumn (zob. zad. 4), analogicznie określone wyważanie wierszy i pełny wybór elementów głównych.

K7. Napisać procedurę `GaussJ`(n, A, b, x, p, s, d), która rozwiązuje układ $Ax = b$ z macierzą stopnia n metodą Gaussa-Jordana (zad. 32c), jednak uzupełnioną o wyważenie kolumn na początku (zad. 4) i skalowany wybór wierszy głównych (p_k-ty wiersz ma służyć do eliminacji niewiadomej x_k z pozostałych równań). Mnożniki niezbędne w tym postępowaniu należy zapamiętać w tablicy d, gdyż będą potrzebne na końcu do obliczenia x.

K8. Napisać i sprawdzić wersję rekurencyjną skalowanej eliminacji Gaussa, z obliczaniem skal przed każdym krokiem.

4.4. Normy i analiza błędów

Badając błędy w zadaniach numerycznych dotyczących wektorów, używamy ich norm. Wektory, którymi będziemy się zajmować, należą do $\mathbb{R}^n$, ale normę można określić w dowolnej przestrzeni wektorowej.

Normy wektorów

W przestrzeni wektorowej V *norma* jest funkcją $\|\cdot\|$ określoną na V, o wartościach rzeczywistych nieujemnych, która ma trzy własności:

$$\|x\| > 0 \quad \text{dla } x \neq 0,\ x \in V, \tag{4.4.1}$$

$$\|\lambda x\| = |\lambda|\|x\| \quad \text{dla } \lambda \in \mathbb{R},\ x \in V, \tag{4.4.2}$$

$$\|x + y\| \leqslant \|x\| + \|y\| \quad \text{dla } x, y \in V \quad \text{(nierówność trójkąta)}. \tag{4.4.3}$$

Możemy uważać $\|x\|$ za *długość* albo *wielkość* wektora x. Norma wektora uogólnia pojęcie wartości bezwzględnej (modułu) $|r|$ liczby rzeczywistej lub zespolonej r. Najbardziej znaną normą w $\mathbb{R}^n$ jest *norma euklidesowa* (norma l_2) określona wzorem

$$\|x\|_2 := \Big(\sum_{i=1}^{n} x_i^2\Big)^{1/2}, \qquad \text{gdzie } x = (x_1, x_2, \ldots, x_n)^T.$$

Odpowiada ona naszemu intuicyjnemu pojęciu długości. Wskaźnik 2 odróżnia tylko tę normę od innych, używanych w analizie numerycznej. Najprostszą z nich jest norma l_∞ określona wzorem

$$\|x\|_\infty := \max_{1\leqslant i\leqslant n} |x_i|. \tag{4.4.4}$$

Trzecią ważną normą w $\mathbb{R}^n$ jest norma l_1 określona wzorem

$$\|x\|_1 := \sum_{i=1}^{n} |x_i|.$$

Przykład 4.4.1. Obliczyć trzy określone wyżej normy dla wektorów

$$x := (4,4,-4,4), \quad v := (0,5,5,5), \quad w := (6,0,0,0).$$

Rozwiązanie. Wyniki podano niżej:

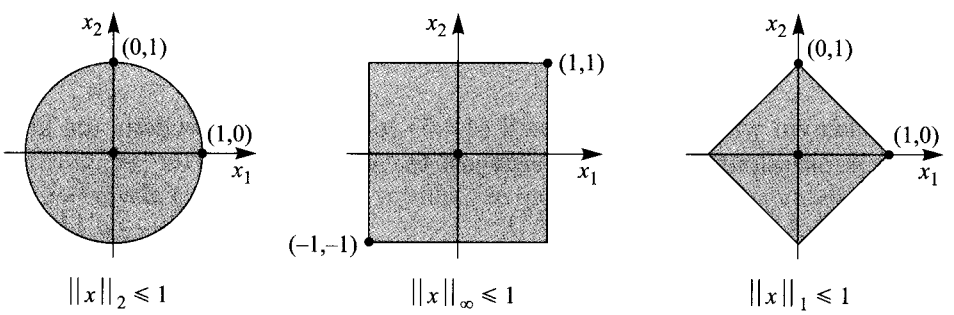
	$\|\cdot\|_1$	$\|\cdot\|_2$	$\|\cdot\|_\infty$
x	16	8	4
v	15	8.66	5
w	6	6	6

■

Aby lepiej zrozumieć sens wprowadzonych norm, rozważmy przestrzeń $\mathbb{R}^2$. Rysunek 4.1 pokazuje dla każdej z nich zbiór

$$\{x\colon\ x \in \mathbb{R}^2,\ \|x\| \leqslant 1\}.$$

Jest to *kostka* (*kula*) *jednostkowa* w tej dwuwymiarowej przestrzeni.



RYS. 4.1. Kostki jednostkowe w $\mathbb{R}^2$ dla trzech norm

Normy macierzy

Przejdźmy teraz do określenia normy macierzy. Ogólnie rzecz biorąc, wystarczy, żeby taka norma spełniała warunki (4.4.1)–(4.4.3). Właściwsze jednak są definicje powiązane z normami wektorów. Dla ustalonej normy $\|\cdot\|$ wektora *indukowana* przez nią *norma macierzy* kwadratowej A stopnia n jest określona wzorem

$$\|A\| := \sup_{\|u\|=1} \{\|Au\| \colon u \in \mathbb{R}^n\} \tag{4.4.5}$$

(ogólniej, tak określa się normę macierzy $m \times n$).

Twierdzenie 4.4.2. *Dla dowolnej normy $\|\cdot\|$ w $\mathbb{R}^n$ wzór (4.4.5) określa normę w przestrzeni liniowej macierzy stopnia n.*

Dowód. Mamy sprawdzić, że norma (4.4.5) spełnia aksjomaty (4.4.1)–(4.4.3). Po pierwsze, jeśli $A \neq 0$, to A ma co najmniej jedną kolumnę, np. $A^{(j)}$, niezerową. Niech x będzie j-tym wektorem jednostkowym. Oczywiście $x \neq 0$, a wektor $v := x/\|x\|$ ma normę 1. Dlatego z definicji (4.4.5) wynika, że

$$\|A\| \geqslant \|Av\| = \frac{\|Ax\|}{\|x\|} = \frac{\|A^{(j)}\|}{\|x\|} > 0.$$

Następnie, z własności (4.4.2) normy wektora wynika, że

$$\|\lambda A\| = \sup_{\|u\|=1} \|\lambda Au\| = |\lambda| \sup_{\|u\|=1} \|Au\| = |\lambda|\,\|A\|.$$

Korzystając z nierówności trójkąta dla normy wektorów i z zad. 51, wnioskujemy, że

$$\begin{aligned} \|A+B\| &= \sup_{\|u\|=1} \|(A+B)u\| \leqslant \sup_{\|u\|=1} (\|Au\| + \|Bu\|) \leqslant \\ &\leqslant \sup_{\|u\|=1} \|Au\| + \sup_{\|u\|=1} \|Bu\| = \|A\| + \|B\|. \end{aligned}$$

■

Ważnym wnioskiem z definicji (4.4.5) (a właściwie powodem, dla którego tak określono normę macierzy) jest to, że

$$\|Ax\| \leqslant \|A\|\,\|x\| \qquad (x \in \mathbb{R}^n). \tag{4.4.6}$$

Istotnie, jest tak dla $x = 0$. Jeśli zaś $x \neq 0$, to wektor $v := x/\|x\|$ ma normę 1 i na mocy (4.4.5) jest

$$\|A\| \geqslant \|Av\| = \frac{\|Ax\|}{\|x\|}.$$

Niech np. normą wektorową będzie $\|\cdot\|_\infty$ z (4.4.4). Jaką normę macierzy ona indukuje? Wynika to z następującego rozumowania:

$$\|A\|_\infty = \sup_{\|u\|_\infty=1} \|Au\|_\infty = \sup_{\|u\|_\infty=1} \max_{1\leqslant i\leqslant n} |(Au)_i| = \max_{1\leqslant i\leqslant n} \sup_{\|u\|_\infty=1} |(Au)_i| =$$

$$= \max_{1\leqslant i\leqslant n} \sup_{\|u\|_\infty=1} \Big|\sum_{j=1}^{n} a_{ij}u_j\Big| = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|. \tag{4.4.7}$$

Korzystamy tu z faktu, że w złożeniu dwóch maksimów można je przestawiać (zob. zad. 52). Korzystamy też z tego, że kres górny wyrażenia $|\sum_{j=1}^{n} a_{ij}u_j|$ dla ustalonego i oraz $\|u\|_\infty = 1$ jest osiągnięty dla $u_j = 1$, gdy $a_{ij} \geqslant 0$ i $u_j = -1$ w przeciwnym razie.

W ten sposób udowodniliśmy następujące twierdzenie:

Twierdzenie 4.4.3. *Norma macierzowa indukowana przez normę wektorową* (4.4.4) *wyraża się wzorem*

$$\|A\|_\infty = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|.$$

Norma macierzowa indukowana przez dowolną normę wektorową ma – oprócz (4.4.1)–(4.4.3) – jeszcze inne własności, m.in. następujące:

$$\|I\| = 1, \qquad \|AB\| \leqslant \|A\|\,\|B\|. \tag{4.4.8}$$

Pierwsza z nich wynika wprost z definicji (4.4.5), a druga z (4.4.5) i (4.4.6).

Inną ważną normą macierzową jest norma *spektralna*, indukowana przez normę euklidesową wektorów, czyli określona wzorem

$$\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2.$$

W twierdzeniu 5.4.10 udowodnimy, że

$$\|A\|_2 = \max_{1\leqslant i\leqslant n} |\sigma_n|,$$

gdzie σ_i są wartościami szczególnymi macierzy A określonymi po tw. 5.4.1. Z podanych tam informacji wynika, że normę spektralną można określić równoważnym wzorem

$$\|A\|_2 = \sqrt{\rho(A^\mathsf{T}A)},$$

gdzie *promień spektralny* $\rho(A^\mathsf{T}A)$ macierzy $A^\mathsf{T}A$ jest z definicji jej największą wartością własną.

Wskaźnik uwarunkowania

Zobaczymy teraz, do czego się przydają określone już normy. Rozważmy układ $Ax = b$ o macierzy kwadratowej nieosobliwej.

Przykład 4.4.4. Jeśli macierz A^{-1} jest zaburzona, co zmienia ją na macierz B, to zaburzenie przenosi się na rozwiązanie $x = A^{-1}b$, zamiast którego otrzymujemy wektor $\tilde{x} = Bb$. Jak duże jest zaburzenie rozwiązania, mierzone bezwzględnie i względnie?

Rozwiązanie. Dla dowolnej normy wektorowej i indukowanej przez nią normy macierzowej wielkość zaburzenia bezwzględnego wynika z nierówności

$$\|x - \tilde{x}\| = \|x - Bb\| = \|x - BAx\| = \|(I - BA)x\| \leqslant \|I - BA\| \, \|x\|.$$

Stąd zaś wynika oszacowanie dla *zaburzenia względnego*:

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leqslant \|I - BA\|. \qquad \blacksquare$$

Przykład 4.4.5. Załóżmy, że zamiast b mamy wektor zaburzony $\tilde{b}$. Niech x i $\tilde{x}$ spełniają odpowiednio równania $Ax = b$ i $A\tilde{x} = \tilde{b}$. Jak różnią się x i $\tilde{x}$, bezwzględnie i względnie?

Rozwiązanie. Jeśli macierz A jest nieosobliwa, to

$$\|x - \tilde{x}\| = \|A^{-1}b - A^{-1}\tilde{b}\| = \|A^{-1}(b - \tilde{b})\| \leqslant \|A^{-1}\| \, \|b - \tilde{b}\|.$$

Jest to oszacowanie zaburzenia bezwzględnego. Zaburzenie względne wynika z dalszych rozumowań (poprawnych, gdy $b \neq 0$):

$$\|x - \tilde{x}\| \leqslant \|A^{-1}\| \, \|b - \tilde{b}\| = \|A^{-1}\| \, \|Ax\| \frac{\|b - \tilde{b}\|}{\|b\|} \leqslant$$

$$\leqslant \|A^{-1}\| \, \|A\| \, \|x\| \frac{\|b - \tilde{b}\|}{\|b\|}.$$

Stąd

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leqslant \varkappa(A) \frac{\|b - \tilde{b}\|}{\|b\|}, \tag{4.4.9}$$

gdzie

$$\varkappa(A) := \|A\| \cdot \|A^{-1}\|. \qquad \blacksquare$$

Wielkość $\varkappa(A)$ nazywamy *wskaźnikiem uwarunkowania* macierzy A. Zależy on od wybranej normy macierzy, a więc pośrednio od wybranej normy wektorowej; jeśli trzeba ten wybór wyraźnie zaznaczyć, to używamy symboli $\varkappa_\infty(A)$, $\varkappa_2(A)$ itp. W każdym przypadku jest $\varkappa(A) \geqslant 1$ (zob. zad. 38). Zawsze też, na mocy (4.4.9), błąd względny obliczonego rozwiązania $\tilde{x}$ nie przewyższa iloczynu wskaźnika uwarunkowania $\varkappa(A)$ przez błąd względny prawej strony. Z (4.4.9) wynika, że jeśli ten wskaźnik nie jest zbyt duży, to małe zaburzenia prawej strony b nieznacznie zaburzają rozwiązanie x.

Sens wskaźnika zbadamy na przykładzie macierzy

$$A = \begin{bmatrix} 1 & 1+\varepsilon \\ 1-\varepsilon & 1 \end{bmatrix}, \qquad A^{-1} = \varepsilon^{-2} \begin{bmatrix} 1 & -1-\varepsilon \\ -1+\varepsilon & 1 \end{bmatrix},$$

gdzie $\varepsilon > 0$. Z (4.4.7) wynika, że $\|A\|_\infty = 2+\varepsilon$ i $\|A^{-1}\|_\infty = \varepsilon^{-2}(2+\varepsilon)$, skąd $\varkappa_\infty(A) = [(2+\varepsilon)/\varepsilon]^2 > 4/\varepsilon^2$. Jeśli $\varepsilon \leqslant 0.01$, to $\varkappa_\infty(A) > 40\,000$. W tym przypadku małe zaburzenie względne wektora b może spowodować 40 000 razy większe zaburzenie względne rozwiązania układu $Ax = b$.

Rozwiązując numerycznie układ równań $Ax = b$, otrzymujemy zamiast rozwiązania dokładnego x jego przybliżenie $\tilde{x}$. Aby sprawdzić jego dokładność, porównujemy $A\tilde{x}$ z b, a ściślej – obliczamy *wektor residualny*

$$r := b - A\tilde{x}.$$

Różnicę

$$e := x - \tilde{x}$$

między dokładnym rozwiązaniem x i jego przybliżeniem $\tilde{x}$ nazywamy *wektorem błędu*. Oba te wektory są powiązane ważną zależnością:

$$Ae = r.$$

Zauważmy, że $\tilde{x}$ jest dokładnym rozwiązaniem układu $A\tilde{x} = \tilde{b}$ z zaburzoną prawą stroną $\tilde{b} := b - r$. Znajdziemy teraz związki między błędami względnymi wektorów $\tilde{x}$ i $\tilde{b}$, tj. między wielkościami $\|x - \tilde{x}\|/\|x\|$ i $\|b - \tilde{b}\|/\|b\| = \|r\|/\|b\|$. Poniższe twierdzenie pokazuje, że ważną rolę odgrywa tu wskaźnik uwarunkowania $\varkappa(A)$.

Twierdzenie 4.4.6. *Wektory residualny i błędu oraz wskaźnik uwarunkowania spełniają nierówność*

$$\frac{1}{\varkappa(A)} \frac{\|r\|}{\|b\|} \leqslant \frac{\|e\|}{\|x\|} \leqslant \varkappa(A) \frac{\|r\|}{\|b\|}.$$

Dowód. Prawa część tej nierówności jest równoważna temu, że

$$\|e\|\,\|b\| \leqslant \|A\|\,\|A^{-1}\|\,\|r\|\,\|x\|,$$

a tak jest, gdyż

$$\|e\|\,\|b\| = \|A^{-1}r\|\,\|Ax\| \leqslant \|A^{-1}\|\,\|r\|\,\|A\|\,\|x\|$$

(w istocie mamy tu nierówność (4.4.9)). Lewą część nierówności z twierdzenia można napisać w postaci

$$\|r\|\,\|x\| \leqslant \|A\|\,\|A^{-1}\|\,\|b\|\,\|e\|,$$

a to wynika natychmiast stąd, że

$$\|r\|\,\|x\| = \|Ae\|\,\|A^{-1}b\| \leqslant \|A\|\,\|e\|\,\|A^{-1}\|\,\|b\|. \qquad \blacksquare$$

Macierz A o dużym wskaźniku $\varkappa(A)$ nazywamy *źle uwarunkowaną*. Dla takich macierzy rozwiązanie układu $Ax = b$ może być bardzo czułe na małe zmiany wektora b. Inaczej mówiąc, aby wyznaczyć x z pewną dokładnością, musimy znać znacznie dokładniejsze b. Jeśli wskaźnik uwarunkowania jest niezbyt duży, to macierz jest *dobrze uwarunkowana*[12].

Zadania 4.4

1. Udowodnić, że dowolna norma wektorowa ma następujące własności:
   (a) $\|0\| = 0$, (b) $\|x+y\| \geqslant |\|x\| - \|y\||$,
   (c) $\|\sum_{i=1}^m x^{(i)}\| \leqslant \sum_{i=1}^m \|x^{(i)}\|$ dla dowolnych wektorów $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$.
2. Podać przykład normy w $\mathbb{R}^2$, której wartość dla $(1,0)$ byłaby równa 2, a dla $(1,1)$ równa 1.
3. Czy istnieje taka norma w $\mathbb{R}^2$, że $\|(1,0)\| = \|(0,1)\| = \|(\frac{1}{3}, \frac{1}{3})\|$?
4. Niech $\|\cdot\|$ będzie normą w $\mathbb{R}^n$. Udowodnić, że wzór

   $$\|x\|' := \sup_{u \in \mathbb{R}^n,\, \|u\|=1} u^{\mathsf{T}} x$$

   określa pewną normę. Udowodnić, że powtórzenie przejścia od $\|\cdot\|$ do $\|\cdot\|'$ powoduje powrót do pierwotnej normy, tj. że $(\|\cdot\|')' = \|\cdot\|$. Udowodnić, że dla dowolnych $x, y \in \mathbb{R}^n$ jest $|x^{\mathsf{T}} y| \leqslant \|x\|\,\|y\|'$.
5. Niech $\|\cdot\|$ będzie normą wektorową w $\mathbb{R}^n$, a A macierzą stopnia n. Podać ścisłe warunki, zapewniające, że także $\|Ax\|$ jest normą w tejże przestrzeni.

[12] Autorzy rozważają skutki zaburzenia macierzy A^{-1} (przykład 4.4.4) lub prawej strony układu $Ax = b$. Ten układ rozwiązujemy jednak na ogół nie korzystając z obliczonej macierzy odwrotnej. Dlatego istotne jest również to, jak zaburzenia samej macierzy A wpływają na rozwiązanie układu; zob. Dryja, Jankowscy [*1982, s. 23] (*przyp. tłum.*).

6. Wykazać, że zbiór $H := \{x \in \mathbb{R}^n\colon \|x-a\|_2 = \|x-b\|_2\}$ jest hiperpłaszczyzną, tj. przekształceniem przez przesunięcie przestrzeni liniowej wymiaru $n-1$, ale że na ogół tak nie jest dla innych norm. Zilustrować to dla $n=2$.

7. Wykazać, że każda z norm: $\|\cdot\|_\infty$, $\|\cdot\|_2$, $\|\cdot\|_1$ ma własności (4.4.1)–(4.4.3).

8. Wykazać, że $\|x\|_\infty \leqslant \|x\|_2 \leqslant \|x\|_1$ dla każdego $x \in \mathbb{R}^n$ i że równość jest tu możliwa nawet dla pewnych wektorów niezerowych.

9. Znaleźć wszystkie wektory x takie, że odpowiednio:
(**a**) $\|x\|_\infty = \|x\|_1$, (**b**) $\|x\|_\infty = \|x\|_2$, (**c**) $\|x\|_1 = \|x\|_2$.

10. Wykazać, że $\|x\|_1 \leqslant n\|x\|_\infty$ i $\|x\|_2 \leqslant \sqrt{n}\,\|x\|_\infty$ dla każdego $x \in \mathbb{R}^n$.

11. Wykazać, że norma wektora musi zależeć od jego wszystkich składowych.

12. Sprawdzić, czy poniższe wyrażenia definiują normę w $\mathbb{R}^n$.
(**a**) $\sum_{i=1}^n |x_i|^3$
(**b**) $\left(\sum_{i=1}^n |x_i|^{1/2}\right)^2$
(**c**) $\max\{|x_1-x_2|, |x_1+x_2|, |x_3|, |x_4|, \ldots, |x_n|\}$
(**d**) $\sum_{i=1}^n 2^{-i}|x_i|$

13. Niech będzie

$$A := \begin{bmatrix} 4 & -3 & 2 \\ -1 & 0 & 5 \\ 2 & 6 & -2 \end{bmatrix}.$$

Wśród wektorów x takich, że $\|x\|_\infty \leqslant 1$, znaleźć ten, dla którego $\|Ax\|_\infty$ jest maksymalne. Obliczyć wartość $\|A\|_\infty$.

14. *Norma ważona* l_∞ wektora $x \in \mathbb{R}^n$ jest określona wzorem

$$\|x\| := \max_{1 \leqslant i \leqslant n} w_i|x_i|,$$

gdzie *wagi* $w_1, w_2, \ldots, w_n$ są ustalonymi liczbami dodatnimi. Udowodnić, że spełnia ona warunki (4.4.1)–(4.4.3) i znaleźć normę indukowaną macierzy.

15. Dla każdego $p \geqslant 1$ wzór $\|x\|_p := \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ określa pewną normę (dowód podaje np. Bartle [1976, s. 61]). Udowodnić, że dla każdego $x \in \mathbb{R}^n$

$$\lim_{p\to\infty} \|x\|_p = \|x\|_\infty,$$

co uzasadnia stosowanie symbolu $\|\cdot\|_\infty$.

16. Niech $\|\cdot\|$ będzie normą w przestrzeni wektorowej V. Dla $x, y \in V$ przyjmujemy, że $d(x,y) := \|x-y\|$. Sprawdzić, że d ma następujące własności:
(**a**) $d(x,x) = 0$, (**b**) $d(x,y) = d(y,x)$, (**c**) $d(x,y) > 0$ dla $x \neq y$,
(**d**) $d(x,y) \leqslant d(x,z) + d(z,y)$. (Taką funkcję d nazywamy *metryką*).

17. Dla wektora $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ określamy jego wartość bezwzględną jako wektor $(|x_1|, |x_2|, \ldots, |x_n|)$. Prócz tego dla wektorów x i y definiujemy relację $\leqslant$; jest $x \leqslant y$, jeśli $x_i \leqslant y_i$ dla $i = 1, 2, \ldots, n$. Udowodnić, że każda z norm $\|\cdot\|_1$, $\|\cdot\|_2$ i $\|\cdot\|_\infty$ jest taka, że jeśli $|x| \leqslant |y|$, to $\|x\| \leqslant \|y\|$.

18. Udowodnić, że każdej normie wektorowej i indukowanej przez nią normie macierzowej dowolnej macierzy A stopnia n odpowiada wektor $x \neq 0$ taki, że $\|Ax\| = \|A\| \, \|x\|$.

19. Czy normy indukowane macierzy spełniają równość $\|AB\| = \|BA\|$?

20. Udowodnić, że jeśli macierz A ma nietrywialny punkt stały (tzn. jeśli $Ax = x$ dla pewnego $x \neq 0$), to $\|A\| \geqslant 1$ dla dowolnej normy indukowanej.

21. Wykazać, że norma indukowana $\|A\|$ jest najmniejszą liczbą M taką, że $\|Ax\| \leqslant M\|x\|$ dla każdego x.

22. *Norma Frobeniusa* macierzy A stopnia n jest określona wzorem

$$\|A\|_F = \Big(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2\Big)^{1/2}.$$

Udowodnić, że spełnia ona warunki (4.4.1)–(4.4.3) i sprawdzić, czy jest indukowana przez jakąś normę wektorową. Zrobić to samo, przyjmując, że $\|A\| = \max_{1\leqslant i,\, j\leqslant n}|a_{ij}|$.

23. Czy wartość każdej normy indukowanej jest równa 1 dla macierzy permutacji?

24. Udowodnić, dla dowolnej normy wektorowej i indukowanej przez nią normy macierzowej, następujące twierdzenie: jeśli dla pewnego $\theta > 0$ i dla każdego wektora x macierz kwadratowa A spełnia nierówność $\|Ax\| \geqslant \theta\|x\|$, to jest ona nieosobliwa i taka, że $\|A^{-1}\| \leqslant \theta^{-1}$.

25. (cd.). Udowodnić, że macierz A dominująca przekątniowo ma własność podaną w poprzednim zadaniu. Jakie θ odpowiada normie $\|\cdot\|_\infty$?

26. Udowodnić, że dla dowolnej macierzy nieosobliwej A istnieje takie $\delta > 0$, że $A + E$ jest nieosobliwa, jeśli tylko $\|E\| < \delta$, gdzie $\|\cdot\|$ jest dowolną normą macierzy.

27. (cd.). Wykazać, że dla dowolnej normy wektorowej i indukowanej przez nią normy macierzy w poprzednim zadaniu można przyjąć

$$\delta := \inf_{\|x\|=1} \|Ax\|.$$

28. Dla normy macierzowej z tw. 4.4.3 sprawdzić, czy $\|AB\|_\infty = \|A\|_\infty \|B\|_\infty$. Jak jest w szczególnym przypadku $A = B$?

29. Udowodnić, że

$$\|A\|_2 = \max_{\|x\|_2=1,\, \|y\|_2=1} |y^{\mathrm{T}} Ax|.$$

30. Wykazać, że norma wektorowa $\|\cdot\|_1$ indukuje normę macierzową

$$\|A\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|.$$

31. Korzystając z wyników zadań 8 i 10, udowodnić, że

$$n^{-1}\|A\|_2 \leqslant n^{-1/2}\|A\|_\infty \leqslant \|A\|_2 \leqslant n^{1/2}\|A\|_1 \leqslant n\|A\|_2.$$

32. Niech A będzie macierzą $m \times n$. Jeśli $x \in \mathbb{R}^n$, to $Ax \in \mathbb{R}^m$. Można przyjąć, że A określa odwzorowanie liniowe przestrzeni $\mathbb{R}^n$ z normą $\|\cdot\|_1$ na przestrzeń $\mathbb{R}^m$ z normą $\|\cdot\|_\infty$. Jak należy wtedy rozumieć $\|A\|$?

33. (cd.). Powtórzyć poprzednie rozumowania, zamieniając miejscami normy $\|\cdot\|_1$ i $\|\cdot\|_\infty$.

34. Udowodnić, że wzór

$$\|A\| := \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|$$

określa normę macierzy stopnia n. Wykazać, że nie jest ona indukowana przez jakąkolwiek normę wektorową. Czy warunki (4.4.8) są tu spełnione?

35. Udowodnić, że $\varkappa(\lambda A) = \varkappa(A)$ $(\lambda \neq 0)$.

36. Udowodnić, że $\varkappa(A) = \sup_{\|x\|=\|y\|} \|Ax\|/\|Ay\|$.

37. Stosując normę $\|\cdot\|_1$ z zad. 30, obliczyć wskaźnik uwarunkowania macierzy $\begin{bmatrix} 1 & 1+\varepsilon \\ 1-\varepsilon & 1 \end{bmatrix}$.

38. Udowodnić, że wskaźnik uwarunkowania macierzy nieosobliwej jest nie mniejszy od 1.

39. Dla jakich macierzy wskaźnik uwarunkowania jest równy 1?

40. Obliczyć wskaźnik uwarunkowania $\varkappa_\infty$ macierzy $\begin{bmatrix} 7 & 8 \\ 9 & 10 \end{bmatrix}$.

41. Znaleźć wskaźnik uwarunkowania $\varkappa_\infty$ macierzy stopnia n trójkątnej dolnej, mającej na głównej przekątnej elementy 1, a pod nią -1.

42. Udowodnić, że wskaźnik uwarunkowania spełnia nierówność

$$\varkappa(AB) \leqslant \varkappa(A)\varkappa(B).$$

43. Obliczyć wskaźniki uwarunkowania $\varkappa_1$, $\varkappa_2$ i $\varkappa_\infty$ dla następujących macierzy:

(a) $\begin{bmatrix} a+1 & a \\ a & a-1 \end{bmatrix}$, (b) $\begin{bmatrix} 0 & 1 \\ -2 & 0 \end{bmatrix}$, (c) $\begin{bmatrix} \alpha & 1 \\ 1 & 1 \end{bmatrix}$.

44. Podać przykład macierzy dobrze uwarunkowanej, mającej bardzo mały wyznacznik.

45. Niech $C = (c_{ij})$ będzie odwrotnością macierzy A. Wykazać, że dla układu $Ax = b$ zaburzenie δ wielkości b_j powoduje zaburzenie $c_{ij}\delta$ niewiadomej x_i.

46. (cd.). Wykazać, że zaburzenie δ elementu a_{jk} zaburza x_i z grubsza o $-c_{ij}x_k\delta$.

47. Jako wskaźnik uwarunkowania macierzy A stopnia n stosuje się czasem wielkość

$$M(A) := n \max_{1 \leqslant i,j \leqslant n} |a_{ij}| \max_{1 \leqslant i,j \leqslant n} |c_{ij}|,$$

gdzie $C = A^{-1}$. Udowodnić, że jeśli $Ax = b$ i jeśli tylko jedna składowa wektora b jest zaburzona, np. o ε, to zaburzone rozwiązanie $\tilde{x}$ jest takie, że

$$\frac{\|x - \tilde{x}\|_\infty}{\|x\|_\infty} \leqslant M(A) \frac{|\varepsilon|}{\|b\|_\infty}.$$

48. Udowodnić, że dla każdej macierzy nieosobliwej A nierówność (4.4.9) staje się równością dla pewnych wektorów b i $\tilde{b}$ (oczywiście chcemy, żeby było $b \neq 0$ i $b \neq \tilde{b}$). Wskazówka: Prześledzić dowód tej nierówności i sprawdzić, kiedy w przekształceniach może wystąpić równość.

49. Rozwiązując układ $Ax = b$ z macierzą $A = \begin{bmatrix} 1 & 2 \\ 1 & 2.01 \end{bmatrix}$, przewidzieć wpływ małych zaburzeń wektora b na rozwiązanie x. Sprawdzić te przewidywania na konkretnym przykładzie: $b = (4, 4)$ i $\tilde{b} = (3, 5)$.

50. Udowodnić, że jeśli macierz A jest nieosobliwa, to istnieje macierz osobliwa B taka, że $\|B - A\|_2 = \|A^{-1}\|_2$.

51. Udowodnić, że dla dowolnych funkcji f i g o wartościach rzeczywistych

$$\sup[f(x) + g(x)] \leqslant \sup f(x) + \sup g(x).$$

52. (**a**) Udowodnić, że dla dowolnych zbiorów A i B i funkcji rzeczywistej ograniczonej, określonej na $A \times B$ jest

$$\sup_{a \in A} \sup_{b \in B} f(a, b) = \sup_{b \in B} \sup_{a \in A} f(a, b).$$

(**b**) Pokazać na przykładzie, że na ogół nie można przestawiać kresu górnego i kresu dolnego.

(**c**) Wykazać, że

$$\sup_{a \in A} \inf_{b \in B} f(a, b) \leqslant \inf_{b \in B} \sup_{a \in A} f(a, b).$$

ZADANIA KOMPUTEROWE 4.4

K1. Napisać procedury obliczające normę $\|x\|_\infty$ wektora x i indukowaną przez nią normę macierzy kwadratowej.

4.5. Szeregi Neumanna i poprawianie iteracyjne

Ważnym zastosowaniem norm jest ścisłe określenie zbieżności w przestrzeni wektorowej V. Jeśli wprowadzono w niej normę $\|\cdot\|$, to para $(V, \|\cdot\|)$ jest *przestrzenią liniową unormowaną*. Ciąg wektorów $v^{(1)}, v^{(2)}, \ldots$ z tej przestrzeni jest *zbieżny* do v, jeśli

$$\lim_{k \to \infty} \|v^{(k)} - v\| = 0.$$

Jest to zgodne z naszym intuicyjnym przekonaniem, że odległości między wektorami $v^{(k)}$ i granicą v powinny dążyć do 0, gdy $k \to \infty$.

Rozważmy przykład w $\mathbb{R}^4$. Jeśli

$$v^{(k)} := (3 - k^{-1}, -2 + k^{-1/2}, (k+1)k^{-1}, e^{-k}), \qquad v := (3, -2, 1, 0),$$

to

$$v^{(k)} - v = (-k^{-1}, k^{-1/2}, k^{-1}, e^{-k}).$$

Dla normy $\|\cdot\|_\infty$ określonej w podrozdz. 4.4 jest oczywiście $\|v^{(k)} - v\|_\infty \to 0$, gdy $k \to \infty$. Wobec tego v jest granicą ciągu $\{v^{(k)}\}$ w przestrzeni liniowej unormowanej $(\mathbb{R}^4, \|\cdot\|_\infty)$.

W tym miejscu warto przypomnieć (bez dowodu) ważny wynik dotyczący takich przestrzeni: dowolne dwie normy w skończeniewymiarowej przestrzeni wektorowej określają zbieżność w ten sam sposób. Sprawdziwszy więc, że $\|v^{(k)} - v\|_\infty \to 0$, wiemy bez dodatkowych obliczeń, że dla dowolnej normy w $\mathbb{R}^4$ jest $\|v^{(k)} - v\| \to 0$. To twierdzenie nie stosuje się jednak do nieskończeniewymiarowych przestrzeni liniowych unormowanych (zob. zad. 22).

A oto inny ważny fakt dotyczący skończeniewymiarowych przestrzeni liniowych unormowanych: każdy ciąg spełniający *warunek Cauchy'ego* jest zbieżny. Inaczej mówiąc, jeśli ciąg $\{v^{(k)}\}$ jest taki, że

$$\lim_{k\to\infty} \sup_{i,j\geqslant k} \|v^{(i)} - v^{(j)}\| = 0,$$

to ma on granicę.

Zastosujemy te wiadomości do wektorów z $\mathbb{R}^n$ i macierzy stopnia n. W poniższych twierdzeniach $\|\cdot\|$ jest dowolną normą w $\mathbb{R}^n$ i ten sam symbol oznacza normę indukowaną macierzy, określoną w podrozdz. 4.4.

Twierdzenie 4.5.1. *Jeśli A jest macierzą stopnia n taką, że $\|A\| < 1$, to macierz $I - A$ jest nieosobliwa i*

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k. \tag{4.5.1}$$

Szereg po prawej stronie (4.5.1) nazywamy *szeregiem Neumanna.*

Dowód. Przypuśćmy, że macierz $I - A$ jest osobliwa. Wtedy istnieje wektor x taki, że $\|x\| = 1$ i $(I - A)x = 0$. Stąd wynika, że

$$1 = \|x\| = \|Ax\| \leqslant \|A\|\,\|x\| = \|A\|,$$

co przeczy założeniu. Udowodnimy teraz, że ciąg sum częściowych szeregu Neumanna jest zbieżny do $(I - A)^{-1}$:

$$\sum_{k=0}^{m} A^k \to (I - A)^{-1} \qquad \text{dla } m \to \infty.$$

Wystarczy sprawdzić, że

$$(I-A)\sum_{k=0}^{m} A^k \to 0 \qquad \text{dla } m \to \infty. \tag{4.5.2}$$

Lewa strona jest równa

$$(I-A)\sum_{k=0}^{m} A^k = \sum_{k=0}^{m}(A^k - A^{k+1}) = A^0 - A^{m+1} = I - A^{m+1}.$$

Ponieważ $\|A^{m+1}\| \leqslant \|A\|^{m+1} \to 0$ dla $m \to \infty$, więc zachodzi (4.5.2). ■

Twierdzenie 4.5.1 praktycznie bez zmian obowiązuje w teorii operatorów liniowych ciągłych w dowolnej przestrzeni Banacha. Ma ono ważne skutki, zarówno praktyczne, jak i teoretyczne. Zauważmy, że z (4.5.1) wynika oszacowanie

$$\|(I-A)^{-1}\| \leqslant \sum_{k=0}^{\infty}\|A^k\| \leqslant \sum_{k=0}^{\infty}\|A\|^k = \frac{1}{1-\|A\|}.$$

Przykład 4.5.2. Stosując szereg Neumanna, obliczyć odwrotność macierzy

$$B := \begin{bmatrix} 0.9 & -0.2 & -0.3 \\ 0.1 & 1.0 & -0.1 \\ 0.3 & 0.2 & 1.1 \end{bmatrix}.$$

Rozwiązanie. Niech będzie

$$A := I - B = \begin{bmatrix} 0.1 & 0.2 & 0.3 \\ -0.1 & 0.0 & 0.1 \\ -0.3 & -0.2 & -0.1 \end{bmatrix}.$$

Ponieważ $\|A\|_\infty = 0.6$, więc szereg Neumanna $\sum_{k=0}^{\infty} A^k$ jest zbieżny do B^{-1}. Stosując algorytm z zad. 19, obliczamy początkowe sumy częściowe:

$$\sum_{k=0}^{0} A^k = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \sum_{k=0}^{1} A^k = \begin{bmatrix} 1.1 & 0.2 & 0.3 \\ -0.1 & 1.0 & 0.1 \\ -0.3 & -0.2 & 0.9 \end{bmatrix},$$

$$\sum_{k=0}^{2} A^k = \begin{bmatrix} 1.00 & 0.16 & 0.32 \\ -0.14 & 0.96 & 0.06 \\ -0.28 & -0.24 & 0.80 \end{bmatrix}, \ldots,$$

$$\sum_{k=0}^{19} A^k = \begin{bmatrix} 1.00000000 & 0.14285714 & 0.28571429 \\ -0.12500000 & 0.96428571 & 0.05357143 \\ -0.25000000 & -0.21428571 & 0.82142857 \end{bmatrix}.$$

Ostatnia suma daje elementy macierzy B^{-1} z ośmioma dokładnymi cyframi po kropce. ■

Niżej podano inny wariant tw. 4.5.1:

TWIERDZENIE 4.5.3. *Macierze kwadratowe A i B takie, że $\|I - AB\| < 1$, są nieosobliwe i*

$$A^{-1} = B\sum_{k=0}^{\infty}(I - AB)^k, \qquad B^{-1} = \sum_{k=0}^{\infty}(I - AB)^k A.$$

Dowód. Na mocy tw. 4.5.1 macierz AB jest nieosobliwa i

$$(AB)^{-1} = \sum_{k=0}^{\infty}(I - AB)^k.$$

Stąd

$$A^{-1} = BB^{-1}A^{-1} = B(AB)^{-1} = B\sum_{k=0}^{\infty}(I - AB)^k,$$

$$B^{-1} = B^{-1}A^{-1}A = (AB)^{-1}A = \sum_{k=0}^{\infty}(I - AB)^k A.$$ ■

Poprawianie iteracyjne

Jeśli $x^{(0)}$ jest przybliżonym rozwiązaniem równania $Ax = b$, to dokładne rozwiązanie x wyraża się wzorem

$$x = x^{(0)} + A^{-1}(b - Ax^{(0)}) = x^{(0)} + e^{(0)},$$

gdzie $e^{(0)} := A^{-1}(b - Ax^{(0)})$ jest wektorem błędu. Natomiast wektor residualny odpowiadający przybliżeniu $x^{(0)}$ określamy wzorem $r^{(0)} := b - Ax^{(0)}$. Ten wektor możemy obliczyć. Aby znaleźć $e^{(0)}$, nie musimy odwracać macierzy A, gdyż wystarczy rozwiązać równanie

$$Ae^{(0)} = r^{(0)}.$$

Te uwagi prowadzą do procedury numerycznej, zwanej *poprawianiem iteracyjnym* rozwiązania.

Przypuśćmy, że równanie $Ax = b$ rozwiązano numerycznie, stosując np. eliminację Gaussa (podrozdz. 4.3). Ponieważ błędy zaokrągleń powodują, że wynik $x^{(0)}$ jest tylko przybliżony, więc obliczamy następnie $r^{(0)}$, $e^{(0)}$ i $x^{(1)}$, stosując relacje

$$r^{(0)} = b - Ax^{(0)}, \qquad Ae^{(0)} = r^{(0)}, \qquad x^{(1)} = x^{(0)} + e^{(0)}.$$

Iterowanie tych czynności daje lepsze przybliżenia $x^{(2)}, x^{(3)}, \ldots$ Ściślej, aby ta procedura była skuteczna, musimy obliczać składowe $b_i - \sum_{j=1}^{n} a_{ij}x_j^{(k)}$ wektora residualnego $r^{(k)}$ w podwójnej precyzji, bo wtedy unikamy utraty cyfr znaczących powodowanej odejmowaniem. (Przypomnijmy, że w idealnej sytuacji $r^{(k)} = 0$, więc wyżej musi występować różnica prawie identycznych liczb).

Przykład 4.5.4. Poprawiać iteracyjnie rozwiązanie układu

$$\begin{bmatrix} 420 & 210 & 140 & 105 \\ 210 & 140 & 105 & 84 \\ 140 & 105 & 84 & 70 \\ 105 & 84 & 70 & 60 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 875 \\ 539 \\ 399 \\ 319 \end{bmatrix}.$$

Rozwiązanie. Stosując eliminację Gaussa ze skalowanym wyborem wierszy głównych, znajdujemy rozkład macierzy układu, a następnie za pomocą podstawiania wstecz znajdujemy przybliżone rozwiązanie

$$x^{(0)} = (0.99998\,0,\ 1.00013\,7,\ 0.99967\,0,\ 1.00021\,5).$$

Kilka kroków poprawiania iteracyjnego daje następujące wyniki:

$$x^{(1)} = (0.99999\,4,\ 1.00006\,9,\ 0.99983\,1,\ 1.00011\,0),$$
$$x^{(2)} = (0.99999\,6,\ 1.00004\,6,\ 0.99989\,1,\ 1.00007\,0),$$
$$x^{(3)} = (0.99999\,3,\ 1.00008\,0,\ 0.99981\,2,\ 1.00012\,1),$$
$$x^{(4)} = (1.00000\,0,\ 1.00000\,6,\ 0.99998\,4,\ 1.00001\,1).$$

Dokładnym rozwiązaniem jest $x = (1, 1, 1, 1)$. Ponieważ operowano na liczbach z siedmiocyfrowymi mantysami, więc końcowy wynik możemy uznać za zupełnie dobry. ■

Aby zbadać teoretycznie ten algorytm, przyjmijmy, że wektor $x^{(0)}$ znaleziono ze wzoru

$$x^{(0)} := Bb,$$

gdzie B jest przybliżoną odwrotnością macierzy A. Proces iteracyjny można więc opisać wzorem

$$x^{(k+1)} := x^{(k)} + B(b - Ax^{(k)}) \quad (k \geqslant 0). \tag{4.5.3}$$

Pokażemy teraz, że te wektory są sumami częściowymi szeregu Neumanna, a to pozwoli udowodnić, że ciąg $\{x^{(k)}\}$ jest zbieżny do rozwiązania układu $Ax = b$ (jeśli działania w (4.5.3) są wykonywane dokładnie).

Jeśli B jest na tyle dobrym przybliżeniem macierzy A^{-1}, że $\|I - AB\| < 1$, to na mocy tw. 4.5.3

$$A^{-1} = B\sum_{k=0}^{\infty}(I - AB)^k$$

i dokładne rozwiązanie układu $Ax = b$ wyraża się wzorem

$$x = B\sum_{k=0}^{\infty}(I - AB)^k b. \tag{4.5.4}$$

Twierdzenie 4.5.5. *Jeśli* $\|I - AB\| < 1$, *to metoda poprawiania iteracyjnego według wzoru* (4.5.3) *daje ciąg wektorów*

$$x^{(m)} := B\sum_{k=0}^{m}(I - AB)^k b \qquad (m \geqslant 0),$$

czyli ciąg sum częściowych szeregu (4.5.4) *jest zbieżny do* x.

Dowód. Ponieważ $x^{(0)} = Bb$, więc twierdzenie jest prawdziwe dla $m = 0$. Przyjmując, że jest tak dla ustalonego $m \geqslant 0$, obliczamy

$$\begin{aligned} x^{(m+1)} &= x^{(m)} + B(b - Ax^{(m)}) = \\ &= B\sum_{k=0}^{m}(I - AB)^k b + Bb - BAB\sum_{k=0}^{m}(I - AB)^k b = \\ &= B\left[b + (I - AB)\sum_{k=0}^{m}(I - AB)^k b\right] = B\sum_{k=0}^{m+1}(I - AB)^k b, \end{aligned}$$

czyli twierdzenie pozostaje prawdziwe po zmianie m na $m + 1$. ■

Zbieżność ciągu $\{x^{(m)}\}$ do x można też wykazać bezpośrednio, ale przy innym założeniu. Z (4.5.3) wynika, że

$$x^{(m+1)} - x = x^{(m)} - x + B(Ax - Ax^{(m)}) = (I - BA)(x^{(m)} - x).$$

Stąd

$$\begin{aligned}\|x^{(m+1)} - x\| &\leqslant \|I - BA\| \, \|x^{(m)} - x\| \leqslant \\ &\leqslant \|I - BA\|^2 \, \|x^{(m-1)} - x\| \leqslant \ldots \\ \ldots &\leqslant \|I - BA\|^m \, \|x^{(0)} - x\|\end{aligned}$$

i błędy $\|x^{(m)} - x\|$ dążą do 0 dla $m \to \infty$, jeśli tylko $\|I - BA\| < 1$. Rozwiązując zad. 8, można się dowiedzieć, czy nierówności $\|I - AB\| < 1$ i $\|I - BA\| < 1$ są równoważne.

Wyważanie

Dla układów równań liniowych o szczególnie kłopotliwych macierzach procedury rozkładu i rozwiązywania znane z podrozdz. 4.3 można wzbogacić o dodatkowe czynności. Oto pięć takich technik:

1. Wstępne wyważanie wierszy[13].
2. Wstępne wyważanie kolumn.
3. Pełny wybór elementów głównych.
4. Wyważanie lub skalowanie w każdym kroku eliminacji.
5. Poprawianie iteracyjne końcowego rozwiązania.

Wyważanie wierszy polega na dzieleniu wszystkich elementów każdego wiersza macierzy współczynników przez ten z nich, który ma największą wartość bezwzględną; inaczej mówiąc, elementy i-tego wiersza mnożymy przez $r_i := 1/\max_{1 \leqslant j \leqslant n} |a_{ij}|$ $(1 \leqslant i \leqslant n)$. Nowe elementy $\tilde{a}_{ij}$ spełniają warunek $\max_{1 \leqslant j \leqslant n} |\tilde{a}_{ij}| = 1$. W praktyce, gdy komputer pracuje w układzie dwójkowym, bardziej celowe jest przyjąć, że r_i jest równe tej z liczb 2^m (m całkowite), która jest najbliższa wartości $1/\max_{1 \leqslant j \leqslant n} |a_{ij}|$. Wtedy unikamy dodatkowych błędów zaokrąglenia przy przejściu od a_{ij} do $\tilde{a}_{ij}$. Ponieważ i-tym równaniem układu jest

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad \text{czyli} \quad \sum_{j=1}^{n} (r_i a_{ij}) x_j = r_i b_i,$$

[13] W oryginale punkty **1**, **2** i **4** zaczynają się słowem *preconditioning*, które nigdzie nie jest zdefiniowane. Tylko z kontekstu wynika, że autorzy mają na myśli wstępne przekształcenie macierzy A lub wektora b przed procesem eliminacji lub analogiczne przekształcenia macierzy tworzonych w poszczególnych krokach eliminacji, wykonywane przed każdym jej krokiem. Te przekształcenia mają poprawiać własności układu (zmniejszać wskaźnik uwarunkowania macierzy) i tym samym zwiększać dokładność wyników (*przyp. tłum.*).

więc także b_i trzeba pomnożyć przez r_i. Dlatego liczby r_i należy przechować na czas rozkładu, aby później mogły być użyte w fazie rozwiązywania układu. W symbolice macierzowej wyważanie wierszy opisujemy równaniem

$$(RA)x = Rb, \qquad \text{gdzie } R := \operatorname{diag}(r_i).$$

Wyważanie kolumn określamy podobnie: j-tą kolumnę mnożymy przez $c_j := 1/\max_{1\leqslant i\leqslant n}|a_{ij}|$ $(1 \leqslant j \leqslant n)$, a raczej – co jest bardziej wskazane – przez bliską tej wartości liczbę 2^m. Pierwotne równania zmieniają się na następujące:

$$\sum_{j=1}^{n}(c_j a_{ij})\left(\frac{x_j}{c_j}\right) = b_i \qquad (1 \leqslant i \leqslant n).$$

Faza rozwiązywania daje wielkości x_j/c_j, które trzeba jeszcze pomnożyć przez c_j; te ostatnie liczby należy więc zapamiętać do końca obliczeń. W symbolice macierzowej wyważanie kolumn jest opisane równaniem

$$(AC)(C^{-1}x) = b, \qquad \text{gdzie } C := \operatorname{diag}(c_j).$$

Pełny wybór elementu głównego na początku obliczeń polega na znalezieniu największego co do modułu elementu macierzy. Określa on zarówno pierwszy wiersz główny, jak i pierwszą taką kolumnę, której elementy będą zerowane w czasie eliminacji. Te kolumny będą zatem wybierane nie w naturalnym porządku $1, 2, \ldots, n$, ale zgodnie z dokładniejszą strategią. Wymaga to użycia dwóch tablic permutacji; jedna zawiera wskaźniki wierszy, a druga – wskaźniki kolumn zawierających kolejne elementy główne (zob. zad. 4.3.4 i 4.3.K1).

Czwartą z wymienionych technik jest wyważanie lub skalowanie w każdym kroku obliczania rozkładu macierzy; daje to bardziej logiczną strukturę programu.

Piątą technikę (poprawianie iteracyjne rozwiązania) omówiono już wcześniej.

Zalety wstępnego wyważania wierszy i kolumn są czasem wątpliwe. Możemy się przekonać, że wyważenie wierszy i następująca po nim eliminacja Gaussa z nieskalowanym ich wyborem jest w istocie tym samym co taka eliminacja, ale z wyborem skalowanym. Dlatego rozsądną strategią jest wyważenie kolumn i następująca po nim eliminacja tego ostatniego typu.

Poniższy przykład pokazuje różnicę między wyważaniem najpierw wierszy, a potem kolumn:

$$\begin{bmatrix} 1 & 10^8 \\ 2 & 0 \end{bmatrix} \longrightarrow \begin{bmatrix} 10^{-8} & 1 \\ 1 & 0 \end{bmatrix} \longrightarrow \begin{bmatrix} 10^{-8} & 1 \\ 1 & 0 \end{bmatrix},$$

a wyważaniem w odwrotnej kolejności:

$$\begin{bmatrix} 1 & 10^8 \\ 2 & 0 \end{bmatrix} \longrightarrow \begin{bmatrix} \frac{1}{2} & 1 \\ 1 & 0 \end{bmatrix} \longrightarrow \begin{bmatrix} \frac{1}{2} & 1 \\ 1 & 0 \end{bmatrix}.$$

Pierwszy wariant polega na wstępnym i końcowym skalowaniu układu za pomocą macierzy przekątniowych, np. R i B:

$$((RA)B)(B^{-1}x) = Rb,$$

a drugi wyraża się wzorem

$$(S(AC))(C^{-1}x) = Sb,$$

gdzie macierze C i S są przekątniowe.

ZADANIA 4.5 [14)]

1. Udowodnić, że zbiór macierzy nieosobliwych stopnia n jest: (**a**) otwarty, (**b**) gęsty w przestrzeni wszystkich macierzy tego stopnia. Tak więc, (**a**) jeśli A jest nieosobliwa, to istnieje takie ε dodatnie, że każda macierz B spełniająca warunek $\|A - B\| < \varepsilon$ jest także nieosobliwa; (**b**) dla dowolnej macierzy A i dowolnego $\varepsilon > 0$ istnieje macierz nieosobliwa B taka, że $\|A - B\| < \varepsilon$.

2. Udowodnić, że jeśli macierz A jest nieosobliwa i $\|A - B\| < \|A^{-1}\|^{-1}$, to również B jest nieosobliwa.

3. Udowodnić, że jeśli $\|A\| < 1$, to

$$\frac{1}{1 + \|A\|} \leqslant \|(I + A)^{-1}\| \leqslant \frac{1}{1 - \|A\|}.$$

4. Wykazać, że jeśli $\|I - AB\| = \varepsilon < 1$, to

$$\|A^{-1} - B\| \leqslant \frac{\varepsilon}{1 - \varepsilon} \|B\|.$$

5. Udowodnić, że jeśli macierz A jest nieosobliwa, to dla dowolnego B

$$\|B - A^{-1}\| \geqslant \frac{\|I - AB\|}{\|A\|}.$$

6. Udowodnić, że jeśli macierz E ma dostatecznie małą normę (jak małą?), to

$$\|(I - E)^{-1} - (I + E)\| \leqslant 3\|E\|^2.$$

[14)] We wszystkich zadaniach macierze A, B, ... są kwadratowe, a ich norma jest indukowana przez pewną normę wektorową, co pozwala korzystać np. z własności (4.4.8) (*przyp. tłum.*).

7. Sprawdzić, czy z nierówności $1 = \|A\| > \|B\|$ wynika, że macierz $A - B$ jest nieosobliwa.

8. Sprawdzić, czy nierówności $\|I - AB\| < 1$ i $\|I - BA\| < 1$ są równoważne.

9. Wykazać, że jeśli macierz A jest nieosobliwa i $\|A - B\| < \|A^{-1}\|^{-1}$, to

$$B^{-1} = A^{-1} \sum_{k=0}^{\infty} (I - BA^{-1})^k.$$

10. Rozwinąć A^{-1} w szereg, zakładając, że $\|I - \alpha A\| < 1$ dla pewnej liczby α.

11. Udowodnić, że jeśli $\inf_{\lambda \in \mathbb{R}} \|I - \lambda A\| < 1$, to macierz A jest nieosobliwa.

12. Udowodnić, że jeśli $\|I - AB\| < 1$, to macierz BA jest nieosobliwa. Czy można to uogólnić na przypadek macierzy niekwadratowych?

13. Udowodnić, że jeśli $\|I - cA^n\| < 1$ dla pewnego c i pewnej liczby naturalnej n, to A jest nieosobliwa.

14. Udowodnić, że jeśli dla pewnego wielomianu p takiego, że $p(0) = 0$, jest $\|I - p(A)\| < 1$, to macierz A jest nieosobliwa. Uogólnić to twierdzenie na wielomian o współczynnikach macierzowych.

15. Wykazać, że jeśli dla pewnego wielomianu p jest $|p(0)| + \|I - p(A)\| < 1$, to macierz A jest nieosobliwa.

16. Wykazać, że dla ustalonej macierzy A operacja $x \mapsto Ax$ jest ciągła, tj. że jeśli ciąg $\{x^{(k)}\}$ jest zbieżny do x, to $\{Ax^{(k)}\} \to Ax$.

17. Udowodnić, że dla macierzy nieosobliwej A jest $\|Ax\| \geqslant \|x\| \, \|A^{-1}\|^{-1}$.

18. Udowodnić, że jeśli $\|I - AB\| < 1$, to $2B - BAB$ jest lepszym od B przybliżeniem macierzy A^{-1} w tym sensie, że iloczyn $A(2B - BAB)$ jest bliższy I.

19. Niech będzie $B_k = \sum_{j=0}^{k} A^j$. Wykazać, że te macierze można obliczać rekurencyjnie według wzorów $B_0 = I$, $B_{k+1} = I + AB_k$.

20. Udowodnić, że jeśli ciąg punktów przestrzeni liniowej unormowanej jest zbieżny, to spełnia warunek Cauchy'ego.

21. Rozważyć układ $\begin{bmatrix} 1 & 2 \\ 1+\delta & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3+\delta \end{bmatrix}$ dla małych $\delta > 0$.

(**a**) Niech $\tilde{x} = (3, 0)$ będzie jego przybliżonym rozwiązaniem. Porównać normę $\|\cdot\|_\infty$ wektora residualnego z takąż normą wektora błędu. Co stąd wynika?

(**b**) Wyznaczyć wskaźnik uwarunkowania $\varkappa_\infty(A)$. Co się dzieje, gdy $\delta \to 0$?

(**c**) Wykonać jeden krok poprawiania iteracyjnego wektora $\tilde{x}$.

22. Niech V będzie przestrzenią wszystkich funkcji ciągłych na $[0, 1]$. Dwie ważne normy w V są następujące:

$$\|x\|_\infty = \max_{0 \leqslant t \leqslant 1} |x(t)|, \qquad \|x\|_1 = \int_0^1 |x(t)| \, dt.$$

Wykazać, że dla funkcji $x_n(t) = t^n$ jest $\|x_n\|_\infty = 1$ i $\|x_n\|_1 \to 0$ dla $n \to \infty$. Stąd wynika, że te normy generują różne pojęcia zbieżności.

ZADANIA KOMPUTEROWE 4.5

K1. Napisać i sprawdzić następujące procedury ulepszające algorytmy rozwiązywania układu $Ax = b$ podane w tekście:

(**a**) Wyważanie kolumn.

(**b**) Wyważanie wierszy.

(**c**) Pełny wybór elementów głównych.

(**d**) Dwa kroki poprawiania iteracyjnego na końcu obliczeń. Alternatywnie, napisać oddzielną procedurę, która dla danych A, x i b poprawia rozwiązanie dwukrotnie (albo m-krotnie). Residua $b_i - \sum_{j=1}^{n} a_{ij}x_j$ powinny być obliczane w podwójnej precyzji i zaokrąglane do pojedynczej. Pamiętać, że pierwotna postać macierzy A i wektora b jest potrzebna do obliczania tych residuów.

Do testów użyć trzech układów. W pierwszym $n = 10$,

$$a_{ij} := (i/11)^j \qquad (1 \leqslant i, j \leqslant n),$$
$$b_i := i[1 - (i/11)^{10}]/(33 - 3i) \qquad (1 \leqslant i \leqslant n).$$

W drugim $n = 4$, $a_{ij} := 1/(2n - i - j + 1)$ i $b_i := \sum_{j=1}^{n} a_{ij}$. Rozwiązaniem jest $x := (1, 1, 1, 1)$. Trzeci test różni się od drugiego tylko tym, że $n = 10$. Przewidzieć drukowanie wyników pośrednich pokazujących szczegóły obliczeń, w tym rozwiązania pierwotne i poprawione.

K2. Używając macierzy testowej A stopnia trzeciego (takiej, że pewna norma indukowana $\|A\|$ jest mniejsza od 1) i metod z poprzedniego zadania, obliczyć $B = \sum_{j=0}^{20} A^j$ i sprawdzić, czy $(I - A)B \approx I$.

K3. Rozwiązać poniższy układ i do wyniku zastosować trzy kroki poprawiania iteracyjnego. Wydrukować r, e i x po każdej iteracji.

$$\begin{bmatrix} 60 & 30 & 20 \\ 30 & 20 & 15 \\ 20 & 15 & 12 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 110 \\ 65 \\ 47 \end{bmatrix}.$$

4.6. Rozwiązywanie układów metodami iteracyjnymi

Algorytm Gaussa i jego warianty są określane jako metody *bezpośrednie* rozwiązywania zadania macierzowego $Ax = b$. Każdy z nich po skończonej liczbie kroków daje rozwiązanie x, które byłoby dokładne, gdyby nie błędy zaokrągleń.

Metoda *iteracyjna* działa inaczej: tworzy ciąg wektorów zbieżny do rozwiązania. Obliczenia przerywamy, gdy rozwiązanie przybliżone osiągnęło wymaganą dokładność lub po ustalonej liczbie iteracji.

Dla wielkich układów, złożonych z tysięcy równań, metody iteracyjne górują często nad metodami bezpośrednimi szybkością działania i wymaganiami co do pamięci. Jeśli żądana dokładność nie jest zbyt duża, to niekiedy można ją uzyskać kosztem stosunkowo niewielu iteracji. Metody iteracyjne są też często efektywne dla *układów rzadkich*, tj. takich, w których elementy macierzy są na ogół zerami. Dzięki temu można ją pamiętać w szczególnym, oszczędnym formacie. W pewnych przypadkach – np. wtedy, gdy rozwiązujemy numerycznie równania różniczkowe cząstkowe – tej macierzy wcale nie trzeba pamiętać. Każdy jej wiersz tworzy się tylko na czas, gdy jest to potrzebne. Inną zaletą metod iteracyjnych jest to, że są zazwyczaj stabilne; błędy zaokrągleń są wygaszane w dalszych obliczeniach.

Aby dać ogólne wyobrażenie o metodach iteracyjnych, opiszemy teraz dwie z nich, najpierw na najprostszym przykładzie.

Przykład 4.6.1. Jak można rozwiązać iteracyjnie układ

$$\begin{bmatrix} 7 & -6 \\ -8 & 9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ -4 \end{bmatrix}?$$

Rozwiązanie. Najprostsza procedura wynika z wyrażenia i-tej niewiadomej z i-tego równania:

$$x_1^{(k)} = \tfrac{6}{7} x_2^{(k-1)} + \tfrac{3}{7}, \qquad x_2^{(k)} = \tfrac{8}{9} x_1^{(k-1)} - \tfrac{4}{9}.$$

Te wzory opisują *metodę* lub *iterację Jacobiego*[15]. Jej początkiem jest wybór przybliżeń początkowych $x_1^{(0)}$ i $x_2^{(0)}$; w braku lepszych mogą to być zera. Powyższe równania generują dokładniejsze – przynajmniej tego byśmy chcieli – przybliżenia $x_1^{(1)}$ i $x_2^{(1)}$. Procedurę powtarzamy albo ustaloną liczbę razy, albo do osiągnięcia odpowiedniej dokładności wektora $(x_1^{(k)}, x_2^{(k)})$. Niżej podano wartości wybranych przybliżeń, otrzymanych metodą Jacobiego:

k	$x_1^{(k)}$	$x_2^{(k)}$
0	0.00000	0.00000
10	0.14865	−0.19820
20	0.18682	−0.24909
30	0.19662	−0.26215
40	0.19913	−0.26551
50	0.19978	−0.26637

Ten proces iteracyjny można by oczywiście zmodyfikować tak, żeby ostatnio obliczone $x_1^{(k)}$ było od razu użyte do wyznaczenia $x_2^{(k)}$:

[15] Jest też używana nazwa *iteracja prosta* (*przyp. tłum.*).

$$x_1^{(k)} = \tfrac{6}{7}x_2^{(k-1)} + \tfrac{3}{7}, \qquad x_2^{(k)} = \tfrac{8}{9}x_1^{(k)} - \tfrac{4}{9}.$$

Jest to *metoda* lub *iteracja Gaussa-Seidela*. W przykładzie daje ona następujące wyniki:

k	$x_1^{(k)}$	$x_2^{(k)}$
0		0.00000
10	0.21978	−0.24909
20	0.20130	−0.26531
30	0.20009	−0.26659
40	0.20001	−0.26666
50	0.20000	−0.26667

(wartość $x_1^{(0)}$ wyżej pominięto, gdyż procedura z niej nie korzysta). Można przypuszczać, że obie metody iteracyjne dają ciągi przybliżeń zbieżne do dokładnego rozwiązania $(\frac{1}{5}, -\frac{4}{15})$ i że druga z nich jest zbieżna szybciej. Zauważmy też, że w przeciwieństwie do metod bezpośrednich dokładność wyników zależy tu od momentu przerwania procesu iteracyjnego. ■

Naszkicowane wyżej dwie metody są szczególnymi przypadkami ogólnej metody iteracyjnej i będą nieco dalej ściśle określone i zbadane.

Ogólna metoda iteracyjna

Poznamy teraz ogólną definicję metody iteracyjnej, służącej do rozwiązywania układu $Ax = b$. Dla ustalonej macierzy Q wyrażamy ten układ w równoważnej postaci

$$Qx = (Q - A)x + b. \tag{4.6.1}$$

Sugeruje to, aby proces iteracyjny opisać równaniem

$$Qx^{(k)} = (Q - A)x^{(k-1)} + b \qquad (k \geqslant 1). \tag{4.6.2}$$

Wektor początkowy $x^{(0)}$ może być dowolny; oczywiście warto wykorzystać jakąś przybliżoną informację o dokładnym rozwiązaniu x. Powiemy, że metoda iteracyjna oparta na (4.6.2) jest zbieżna, jeśli ciąg $\{x^{(k)}\}$ jest zbieżny do x dla dowolnego wektora początkowego $x^{(0)}$.

Macierz Q powinna spełniać dwa warunki:

1. Obliczanie przybliżeń $x^{(k)}$ jest łatwe.
2. Ciąg $\{x^{(k)}\}$ jest szybko zbieżny do rozwiązania.

Tak więc układy równań liniowych z macierzą Q powinny być łatwo rozwiązywalne. Zobaczymy, że drugi warunek jest spełniony, jeśli Q^{-1} jest dobrym przybliżeniem macierzy A^{-1}.

Zauważmy najpierw, że jeśli ciąg $\{x^{(k)}\}$ jest zbieżny, to musi dążyć do x. Istotnie, przechodząc w (4.6.2) do granicy i uwzględniając ciągłość operacji algebraicznych, otrzymujemy układ (4.6.1), czyli $Ax = b$.

Aby zapewnić istnienie rozwiązania układu $Ax = b$ dla dowolnego b, zakładamy, że macierz A jest nieosobliwa. To samo zakładamy o Q, dzięki czemu z (4.6.2) można obliczyć wektor $x^{(k)}$. Mamy więc prawo zastosować w rozważaniach teoretycznych wzór

$$x^{(k)} := (I - Q^{-1}A)x^{(k-1)} + Q^{-1}b. \tag{4.6.3}$$

Natomiast w praktyce numerycznej wektor $x^{(k)}$ oblicza się niemal zawsze z (4.6.2) bez użycia odwrotności Q^{-1}.

Ponieważ dokładne rozwiązanie spełnia podobne do (4.6.3) równanie

$$x = (I - Q^{-1}A)x + Q^{-1}b, \tag{4.6.4}$$

więc x jest punktem stałym odwzorowania

$$x \mapsto (I - Q^{-1}A)x + Q^{-1}b.$$

Odejmując stronami (4.6.4) od (4.6.3), otrzymujemy

$$x^{(k)} - x = (I - Q^{-1}A)(x^{(k-1)} - x).$$

Dla dowolnej normy wektorowej i indukowanej przez nią normy macierzowej daje to nierówność

$$\|x^{(k)} - x\| \leqslant \|I - Q^{-1}A\| \, \|x^{(k-1)} - x\|,$$

a z niej wynika, że

$$\|x^{(k)} - x\| \leqslant \|I - Q^{-1}A\|^k \, \|x^{(0)} - x\|.$$

Jeśli zatem $\|I - Q^{-1}A\| < 1$, to

$$\lim_{k\to\infty} \|x^{(k)} - x\| = 0$$

dla dowolnego $x^{(0)}$. Zauważmy, że to założenie implikuje nieosobliwość macierzy $Q^{-1}A$ i A. W ten sposób udowodniono

TWIERDZENIE 4.6.2. *Jeśli $\|I - Q^{-1}A\| < 1$ dla pewnej normy indukowanej macierzy, to ciąg określony równaniem (4.6.2) jest zbieżny do rozwiązania układu $Ax = b$ dla dowolnego wektora początkowego $x^{(0)}$.*

Zakładając, że norma $\delta := \|I - Q^{-1}A\|$ jest mniejsza od 1, możemy bezpiecznie zakończyć proces iteracyjny, gdy tylko wielkość $\|x^{(k)} - x^{(k-1)}\|$ jest dostatecznie mała. Istotnie, można udowodnić (zob. zad. 11), że

$$\|x^{(k)} - x\| \leqslant \frac{\delta}{1-\delta}\|x^{(k)} - x^{(k-1)}\|.$$

Metoda Richardsona

Jedną z metod iteracyjnych jest *metoda Richardsona*, w której Q jest macierzą jednostkową:

$$x^{(k)} := (I - A)x^{(k-1)} + b = x^{(k-1)} + r^{(k-1)},$$

gdzie $r^{(k-1)} := b - Ax^{(k-1)}$ jest określonym już w podrozdz. 4.4 wektorem residualnym. Zgodnie z tw. 4.6.2 ta metoda daje w granicy rozwiązanie, jeśli dla pewnej normy indukowanej jest $\|I - A\| < 1$. W zadaniach 15 i 16 podano dwie klasy macierzy A o takiej własności.

Algorytm, który wykonuje M kroków metodą Richardsona, jest następujący:

input $n, (a_{ij}), (b_i), (x_i), M$
for $k = 1$ **to** M **do**
  for $i = 1$ **to** n **do**
    $r_i \leftarrow b_i - \sum_{j=1}^{n} a_{ij}x_j$
  end do
  for $i = 1$ **to** n **do**
    $x_i \leftarrow x_i + r_i$
  end do
  output $k, (x_i), (r_i)$
end do

Przykład 4.6.3. Wykonać 100 kroków metodą Richardsona, zaczynając od $x = (0, 0, 0)$, dla układu

$$\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{3} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \frac{11}{18} \\ \frac{11}{18} \\ \frac{11}{18} \end{bmatrix}.$$

Rozwiązanie. Oto wybrane wyniki otrzymane tą metodą:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.00000	0.00000	0.00000
1	0.61111	0.61111	0.61111
10	0.27950	0.27950	0.27950
40	0.33311	0.33311	0.33311
80	0.33333	0.33333	0.33333

■

Metoda Jacobiego

Innym przykładem ilustrującym ogólną teorię jest *metoda Jacobiego*, w której Q jest macierzą przekątniową o elementach a_{ii} takich, jak w A. Wtedy

$(Q^{-1}A)_{ij} = a_{ij}/a_{ii}$ i ta macierz ma jedynki na głównej przekątnej. Dlatego

$$\|I - Q^{-1}A\|_\infty = \max_{1\leqslant i\leqslant n} \sum_{j=1,\, j\neq i}^{n} |a_{ij}/a_{ii}|. \tag{4.6.5}$$

Twierdzenie 4.6.4. *Jeśli macierz A jest dominująca przekątniowo, to dla dowolnego wektora początkowego metoda Jacobiego tworzy ciąg zbieżny do rozwiązania układu $Ax = b$.*

Dowód. Wynikająca z założenia nierówność

$$|a_{ii}| > \sum_{j=1,\, j\neq i}^{n} |a_{ij}|$$

wraz z (4.6.5) prowadzi do wniosku, że $\|I - Q^{-1}A\|_\infty < 1$ i metoda Jacobiego jest zbieżna na mocy tw. 4.6.2. ■

Algorytm oparty na metodzie Jacobiego jest następujący (M jest liczbą kroków, które należy wykonać):

input $n, (a_{ij}), (b_i), (x_i), M$
for $k = 1$ **to** M **do**
  for $i = 1$ **to** n **do**
    $u_i \leftarrow \left(b_i - \sum_{j=1,\, j\neq i}^{n} a_{ij}x_j\right)/a_{ii}$
  end do
  for $i = 1$ **to** n **do**
    $x_i \leftarrow u_i$
  end do
  output $k, (x_i)$
end do

Ten algorytm i inne podobne można ulepszyć, wykonując wszystkie dzielenia przed rozpoczęciem iteracji. Służą do tego instrukcje wykonywane na początku programu:

for $i = 1$ **to** n **do**
  $d \leftarrow 1/a_{ii}$
  $b_i \leftarrow db_i$
  for $j = 1$ **to** n **do**
    $a_{ij} \leftarrow da_{ij}$
  end do
end do

Dzięki temu główna instrukcja podstawienia upraszcza się do postaci

$$u_i \leftarrow b_i - \sum_{j=1,\, j\neq i}^{n} a_{ij}x_j.$$

Te zmiany algorytmu można interpretować inaczej: pierwotny układ $Ax = b$ zmieniamy na

$$D^{-1}Ax = D^{-1}b,$$

gdzie $D = \operatorname{diag}(a_{ii})$. Można też uniknąć dzieleń, przygotowując układ w inny sposób, np. przeskalowując go dwustronnie, czyli nadając mu postać

$$(D^{-1/2}AD^{-1/2})(D^{1/2}x) = D^{-1/2}b,$$

gdzie $D^{\pm 1/2} := \operatorname{diag}(a_{ii}^{\pm 1/2})$ (zakładamy, że elementy a_{ii} są dodatnie). Zauważmy, że jeśli macierz A jest symetryczna, to takie przeskalowanie nie narusza tej własności. W wielu metodach iteracyjnych jakieś proste przygotowanie układu do stosowania metody iteracyjnej może znacznie polepszyć jej efektywność.

Ogólna metoda iteracyjna

Następnym tematem jest teoria dowolnej metody iteracyjnej, opisanej wzorem

$$x^{(k)} := Gx^{(k-1)} + c, \tag{4.6.6}$$

gdzie G jest daną macierzą stopnia n, a c wektorem z $\mathbb{R}^n$. Oczywiście, równanie (4.6.2) prowadzi do takiego wzoru, mianowicie dla

$$G := I - Q^{-1}A, \qquad c := Q^{-1}b.$$

Chcemy sprawdzić, jakie warunki powinna spełniać macierz G, żeby metoda opisana wzorem (4.6.6) była zbieżna dla dowolnego wektora początkowego. Najpierw jednak trzeba poznać pomocnicze pojęcia i ich własności (omówione zresztą bardziej szczegółowo w podrozdz. 5.0).

Wartość własna macierzy A jest z definicji taką liczbą zespoloną λ, że macierz $A - \lambda I$ jest osobliwa, tj. że istnieje *wektor własny* $x \neq 0$, spełniający równanie $Ax = \lambda x$. Wartości własne są więc pierwiastkami równania

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

Promień spektralny macierzy A jest określony wzorem

$$\rho(A) := \max\{|\lambda| \colon \det(A - \lambda I) = 0\},$$

czyli jest to promień najmniejszego koła o środku w punkcie 0 na płaszczyźnie zespolonej, zawierającego wszystkie wartości własne macierzy A.

Macierz A jest *podobna* do macierzy B, jeśli istnieje taka macierz nieosobliwa S, że $S^{-1}AS = B$; oczywiście B jest podobna do A, gdyż $SBS^{-1} = A$. Stąd wynika, że A i B mają te same wartości własne (tw. 5.0.2). Łatwo też zauważyć, że wartościami własnymi macierzy trójkątnej są jej elementy przekątniowe.

Twierdzenie 4.6.5. *Każda macierz kwadratowa jest podobna do pewnej macierzy trójkątnej górnej (być może zespolonej) o elementach pozaprzekątniowych dowolnie małych.*

Dowód. Na mocy tw. Schura 5.2.3 dowolna macierz kwadratowa A jest podobna do pewnej macierzy trójkątnej górnej $T = (t_{ij})$, której elementy mogą być zespolone. Niech będzie $0 < \varepsilon < 1$ i $D = \operatorname{diag}(\varepsilon, \varepsilon^2, \ldots, \varepsilon^n)$. Łatwo sprawdzić, że $(D^{-1}TD)_{ij} = t_{ij}\varepsilon^{j-i}$. Elementy tej macierzy pod główną przekątną znikają, a nad nią, czyli dla $j > i$, są takie, że

$$|t_{ij}\varepsilon^{j-i}| \leqslant \varepsilon |t_{ij}|.$$

To oszacowanie z góry może być dowolnie małe. ■

Twierdzenie 4.6.6. *Zachodzi równość*

$$\rho(A) = \inf_{\|\cdot\|} \|A\|,$$

gdzie kres dolny bierze się po wszystkich normach indukowanych macierzy.

Dowód. Aby wykazać, że $\rho(A) \leqslant \inf_{\|\cdot\|} \|A\|$, rozważmy dowolną wartość własną macierzy A. Niech odpowiada jej wektor własny x. Dla dowolnej normy wektorowej i indukowanej przez nią normy macierzowej jest

$$|\lambda| \|x\| = \|\lambda x\| = \|Ax\| \leqslant \|A\| \, \|x\|,$$

czyli $|\lambda| \leqslant \|A\|$. Dlatego $\rho(A) \leqslant \|A\|$ i wzięcie kresu dolnego daje zapowiedzianą nierówność.

Aby udowodnić, że $\rho(A) \geqslant \inf_{\|\cdot\|} \|A\|$, zauważmy, że na mocy tw. 4.6.5 dla każdego $\varepsilon > 0$ istnieje macierz nieosobliwa S taka, iż $S^{-1}AS = D + T$, gdzie D jest przekątniowa, a T ma niezerowe elementy tylko nad główną przekątną i spełnia warunek $\|T\|_\infty \leqslant \varepsilon$. Stąd

$$\|S^{-1}AS\|_\infty = \|D + T\|_\infty \leqslant \|D\|_\infty + \|T\|_\infty.$$

Ponieważ na głównej przekątnej macierzy D występują wartości własne λ_i macierzy A, więc

$$\|D\|_\infty = \max_{1\leqslant i\leqslant n} |\lambda_i| = \rho(A), \qquad \|S^{-1}AS\|_\infty \leqslant \rho(A) + \varepsilon.$$

Zgodnie z zad. 1 funkcja $\|\cdot\|'_\infty$ określona wzorem

$$\|A\|'_\infty := \|S^{-1}AS\|_\infty$$

jest pewną normą indukowaną macierzy A. Wiemy, że $\|A\|'_\infty \leqslant \rho(A) + \varepsilon$ i tym bardziej $\inf_{\|\cdot\|}\|A\| \leqslant \rho(A)+\varepsilon$. Ponieważ ε jest dowolną liczbą dodatnią, więc daje to żądaną nierówność. ■

Dowiemy się teraz, dla jakich macierzy G metoda iteracyjna (4.6.6) jest zbieżna.

Twierdzenie 4.6.7. *Warunkiem koniecznym i dostatecznym na to, żeby dla każdego wektora początkowego $x^{(0)}$ wzór*

$$x^{(k)} := Gx^{(k-1)} + c \qquad (k \geqslant 1)$$

generował ciąg zbieżny do $(I - G)^{-1}c$, jest nierówność $\rho(G) < 1$.

Dowód. Niech będzie $\rho(G) < 1$. Na mocy tw. 4.6.6 istnieje taka norma indukowana macierzy, że $\|G\| < 1$. Obliczamy

$$x^{(1)} = Gx^{(0)} + c, \qquad x^{(2)} = G^2x^{(0)} + Gc + c, \qquad \ldots,$$

a ogólnie

$$x^{(k)} = G^kx^{(0)} + \sum_{j=0}^{k-1} G^jc. \tag{4.6.7}$$

Dla tej normy wektorowej, która indukuje użytą wyżej normę macierzy, jest

$$\|G^kx^{(0)}\| \leqslant \|G^k\|\,\|x^{(0)}\| \leqslant \|G\|^k\|x^{(0)}\| \to 0 \qquad \text{dla } k \to \infty.$$

Z twierdzenia 4.5.1 wynika, że

$$\sum_{j=0}^{\infty} G^jc = (I - G)^{-1}c,$$

więc przejście do granicy $k \to \infty$ w (4.6.7) daje równość

$$\lim_{k\to\infty} x^{(k)} = (I - G)^{-1}c.$$

Przypuśćmy teraz, że jest przeciwnie: $\rho(G) \geqslant 1$. Wybieramy u i λ tak, że $Gu = \lambda u$, $|\lambda| \geqslant 1$, $u \neq 0$. Dla $c := u$ i $x^{(0)} := 0$ z (4.6.7) wynika, że

$$x^{(k)} = \sum_{j=0}^{k-1} G^j u = \sum_{j=0}^{k-1} \lambda^j u.$$

Jeśli $\lambda = 1$, to $x^{(k)} = ku$ i ten ciąg jest rozbieżny dla $k \to \infty$. Tak samo jest dla $\lambda \neq 1$, bo wtedy $x^{(k)} = (\lambda^k - 1)(\lambda - 1)^{-1}u$. ■

Wniosek 4.6.8. *Jeśli $\rho(I - Q^{-1}A) < 1$, to dla dowolnego $x^{(0)}$ wzór iteracyjny* (4.6.2) *daje ciąg zbieżny do rozwiązania układu $Ax = b$.*

Metoda Gaussa-Seidela

Zbadamy teraz dokładniej *metodę Gaussa-Seidela*, użytą na początku tego podrozdziału w bardzo prostym przykładzie. Jest to metoda (4.6.2) w przypadku, gdy Q jest częścią trójkątną dolną (wraz z główną przekątną) macierzy A.

Twierdzenie 4.6.9. *Jeśli macierz A jest dominująca przekątniowo, to metoda Gaussa-Seidela jest zbieżna dla dowolnego wektora początkowego.*

Dowód. Wobec wniosku 4.6.8 wystarczy wykazać, że

$$\rho(I - Q^{-1}A) < 1.$$

Niech λ będzie dowolną wartością własną macierzy $I - Q^{-1}A$, a x odpowiadającym jej wektorem własnym. Nie ograniczając ogólności, możemy założyć, że $\|x\|_\infty = 1$. Jest

$$(I - Q^{-1}A)x = \lambda x, \quad \text{czyli} \quad Qx - Ax = \lambda Qx.$$

Stąd i z określenia macierzy Q wynika, że

$$-\sum_{j=i+1}^{n} a_{ij}x_j = \lambda \sum_{j=1}^{i} a_{ij}x_j \qquad (1 \leqslant i \leqslant n).$$

To zaś po przegrupowaniu składników daje równości

$$\lambda a_{ii}x_i = -\lambda \sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}x_j \qquad (1 \leqslant i \leqslant n).$$

Niech wskaźnik i będzie taki, że $|x_i| = 1 \geqslant |x_j|$ dla wszystkich j. Stąd

$$|\lambda|\,|a_{ii}| \leqslant |\lambda| \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|.$$

Ponieważ A jest macierzą dominującą przekątniowo, więc

$$|\lambda| \leqslant \Big(\sum_{j=i+1}^{n} |a_{ij}| \Big) \Big(|a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}| \Big)^{-1} < 1. \qquad \blacksquare$$

Algorytm metody Gaussa-Seidela jest następujący:

input $n, (a_{ij}), (b_i), (x_i), M$
for $k = 1$ **to** M **do**
  for $i = 1$ **to** n **do**
    $x_i \leftarrow \big(b_i - \sum_{j=1,\, j\neq i}^{n} a_{ij} x_j\big)/a_{ii}$
  end do
output $k, (x_i)$
end do

Zauważmy, że w tej metodzie nowe wartości x_i natychmiast po ich znalezieniu zastępują stare i są używane przy obliczaniu x_{i+1} itd. Inaczej działa metoda Jacobiego: tam nowe przybliżenia składowych rozwiązania (oznaczane w algorytmie symbolem u_i) są wykorzystywane dopiero w następnej iteracji i można je obliczać jednocześnie. Dzięki temu metoda Jacobiego lepiej nadaje się do obliczeń w komputerach pracujących równolegle lub wektorowo. Zauważmy też, że efektywność metody Gaussa-Seidela można zwiększyć, odpowiednio przekształcając układ przed jej zastosowaniem.

Przykład 4.6.10. Zastosować metodę Gaussa-Seidela dla $x^{(0)} = (0, 0, 0)$ do układu

$$\begin{bmatrix} 2 & -1 & 0 \\ 1 & 6 & -2 \\ 4 & -3 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ -4 \\ 5 \end{bmatrix}.$$

Rozwiązanie. Układ skalujemy, czyli wyrażamy w postaci $D^{-1}Ax = D^{-1}b$, gdzie $D := \operatorname{diag}(a_{ii})$:

$$\begin{bmatrix} 1 & -\frac{1}{2} & 0 \\ \frac{1}{6} & 1 & -\frac{1}{3} \\ \frac{1}{2} & -\frac{3}{8} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -\frac{2}{3} \\ \frac{5}{8} \end{bmatrix}.$$

Stąd

$$x_1^{(k)} = \tfrac{1}{2}x_2^{(k-1)} + 1,$$
$$x_2^{(k)} = -\tfrac{1}{6}x_1^{(k)} + \tfrac{1}{3}x_3^{(k-1)} - \tfrac{2}{3},$$
$$x_3^{(k)} = -\tfrac{1}{2}x_1^{(k)} + \tfrac{3}{8}x_2^{(k)} + \tfrac{5}{8}.$$

Wybrane przybliżenia otrzymane z tych równań podano niżej; ostatnie jest już dokładne:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
1	1.000000	−0.833333	−0.187500
5	0.622836	−0.760042	0.028566
10	0.620001	−0.760003	0.029998
13	0.620000	−0.760000	0.030000

■

Metoda nadrelaksacji (SOR)

Następny ważny przykład metody iteracyjnej jest znany jako *nadrelaksacja* i oznaczany skrótem SOR (od nazwy angielskiej *successive overrelaxation*). Ogólna teoria tej metody odnosi się do dziedziny zespolonej, więc najpierw przypomnimy kilka potrzebnych pojęć.

Liczbę zespoloną γ można wyrazić w postaci $\gamma = \alpha + \beta i$, gdzie α i β są rzeczywiste, $i^2 = -1$. *Liczbą sprzężoną* względem γ jest $\bar{\gamma} := \alpha - \beta i$. *Modułem* liczby γ jest $|\gamma| := \sqrt{\alpha^2 + \beta^2} = \sqrt{\gamma\bar{\gamma}}$.

Dla macierzy $A := (a_{ij})$ o dowolnej liczbie wierszy i kolumn oraz o elementach zespolonych definiujemy *macierz sprzężoną* $A^{\mathsf{H}} := (\bar{a}_{ji})$[16]. Przestrzeń wektorów o n składowych zespolonych oznaczamy symbolem $\mathbb{C}^n$. Te wektory identyfikujemy z macierzami jednokolumnowymi. W $\mathbb{C}^n$ *iloczyn skalarny* wektorów x i y jest określony wzorem

$$\langle x, y\rangle := y^{\mathsf{H}}x = \sum_{i=1}^{n} x_i\bar{y}_i.$$

Łatwo spostrzec, że

$$\langle x, x\rangle > 0 \quad \text{dla} \quad x \neq 0,$$
$$\langle x, \lambda y\rangle = \bar{\lambda}\langle x, y\rangle,$$

[16] Symbol A^{H} stosują np. Golub i van Loan [1989] oraz Kiełbasiński i Schwetlick [*1992], ale niektórzy autorzy używają symbolu A^*; tak też jest w oryginale tej książki (*przyp. tłum.*).

$\langle x, y\rangle = \overline{\langle y, x\rangle}$,
$\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle \qquad (\alpha,\ \beta$ – liczby),
$\langle x, Ay\rangle = \langle A^{\mathsf{H}}x, y\rangle$.

Normą euklidesową wektora $x \in \mathbb{C}^n$ jest liczba

$$\|x\|_2 := \sqrt{\langle x, x\rangle} = \sqrt{x^{\mathsf{H}}x} = \Big(\sum_{i=1}^{n} |x_i|^2\Big)^{1/2}.$$

Macierz A jest *hermitowska*, jeśli $A = A^{\mathsf{H}}$ i *dodatnio określona* (*półokreślona*), jeśli $\langle Ax, x\rangle > 0$ (odpowiednio, $\langle Ax, x\rangle \geqslant 0$) dla każdego $x \neq 0$. Zauważmy, że dla macierzy hermitowskiej jest $\langle Ax, y\rangle = \langle x, Ay\rangle$.

Metoda nadrelaksacji jest kolejnym przypadkiem szczególnym metody iteracyjnej służącej do rozwiązania układu $Ax = b$ i opisanej równością (4.6.2). Zakładamy teraz, że występująca tam macierz Q jest równa $\alpha D - C$, gdzie α jest parametrem rzeczywistym, D – macierzą hermitowską dodatnio określoną, a macierz C jest taka, że $C + C^{\mathsf{H}} = D - A$. Warunki zbieżności tej metody podano niżej.

Twierdzenie 4.6.11. *Jeśli macierz A jest hermitowska dodatnio określona, a Q jest nieosobliwa i jeśli $\alpha > \frac{1}{2}$, to dla dowolnego wektora początkowego metoda nadrelaksacji jest zbieżna.*

Dowód. Jak w poprzednim dowodzie, chcemy wykazać, że promień spektralny macierzy $G := I - Q^{-1}A$ jest mniejszy od 1. Niech λ będzie jej wartością własną, a x odpowiednim wektorem własnym. Przyjmujemy, że $y := (I - G)x$. Łatwo sprawdzić, że

$$y = x - Gx = x - \lambda x = Q^{-1}Ax, \tag{4.6.8}$$
$$Q - A = (\alpha D - C) - (D - C - C^{\mathsf{H}}) = \alpha D - D + C^{\mathsf{H}}. \tag{4.6.9}$$

Korzystając z (4.6.8), stwierdzamy, że

$$(\alpha D - C)y = Qy = Ax, \tag{4.6.10}$$

a to wraz z (4.6.9) i (4.6.8) daje ciąg równości

$$(\alpha D - D + C^{\mathsf{H}})y = (Q - A)y = Ax - Ay = A(x - y) = \\ = A(x - Q^{-1}Ax) = AGx.$$

Stąd i z (4.6.10) wynika, że

$$\alpha\langle Dy, y\rangle - \langle Cy, y\rangle = \langle Ax, y\rangle,$$

$$\alpha\langle y, Dy\rangle - \langle y, Dy\rangle + \langle y, C^{\mathsf{H}}y\rangle = \langle y, AGx\rangle.$$

Sumujemy stronami te równości uwzględniając, że $\langle Cy, y\rangle = \langle y, C^{\mathsf{H}}y\rangle$:

$$\begin{aligned} 2\alpha\langle Dy, y\rangle - \langle y, Dy\rangle &= \langle Ax, y\rangle + \langle y, AGx\rangle, \\ (2\alpha - 1)\langle Dy, y\rangle &= \langle Ax, y\rangle + \langle y, AGx\rangle. \end{aligned} \tag{4.6.11}$$

Skorzystano tu z tego, że macierz D jest hermitowska. Ponieważ $y = (1-\lambda)x$ i $Gx = \lambda x$, więc z (4.6.11) wynika, że

$$\begin{aligned} (2\alpha - 1)|1 - \lambda|^2\langle Dx, x\rangle &= (1 - \bar{\lambda})\langle Ax, x\rangle + \bar{\lambda}(1 - \lambda)\langle x, Ax\rangle = \\ &= (1 - |\lambda|^2)\langle Ax, x\rangle \end{aligned} \tag{4.6.12}$$

(macierz A jest hermitowska). Jeśli $\lambda \neq 1$, to lewa strona (4.6.12) jest dodatnia. Prawa strona też musi być dodatnia, więc $|\lambda| < 1$. Jeśli natomiast $\lambda = 1$, to $y = (1 - \lambda)x = 0$ i wobec (4.6.10) jest $Ax = 0$. To przeczyłoby nierówności $\langle Ax, x\rangle > 0$ spełnionej dla każdego $x \neq 0$. Tak więc $\rho(G) < 1$ i metoda nadrelaksacji jest zbieżna. ■

W metodzie nadrelaksacji przyjmuje się zwykle, że $D = \operatorname{diag}(a_{ii})$, a $-C$ jest częścią trójkątną dolną, bez głównej przekątnej, macierzy A. W twierdzeniu 4.6.11 nie ma jednak takiego założenia. Warto też ostrzec czytelników, że w publikacjach zamiast α występuje zwykle iloraz $1/\omega$. Warunek podany w twierdzeniu znaczy więc, że $0 < \omega < 2$. Problem wyboru parametru ω zapewniającego najszybszą zbieżność metody nadrelaksacji rozważają Young [1971], Varga [1962], Hageman i Young [1981], Wachspress [1966], Isaacson i Keller [1966] oraz wielu innych autorów.

Macierze określające metody iteracyjne

Przypuśćmy, że $A := (a_{ij})$ rozłożono na składniki według wzoru

$$A = D - C_L - C_U,$$

gdzie $D := \operatorname{diag}(a_{ii})$, a $-C_L$ i $-C_U$ są odpowiednio częścią trójkątną dolną i górną, obie bez głównej przekątnej, macierzy A. W innym wariancie występują macierze blokowo przekątniowe i blokowo trójkątne. W dyskretyzacji równań różniczkowych cząstkowych pierwszy wariant rozkładu odpowiada pojedynczym punktom siatki, a drugi – ich układom.

Niżej streszczono informacje o konkretnych metodach iteracyjnych opisanych w tym podrozdziale.

Metoda Richardsona:

$$Q := I, \quad G := I - A,$$
$$x^{(k)} := (I - A)x^{(k-1)} + b.$$

Metoda Jacobiego:

$$Q := D, \quad G := D^{-1}(C_L + C_U),$$
$$Dx^{(k)} := (C_L + C_U)x^{(k-1)} + b.$$

Metoda Gaussa-Seidela:

$$Q := D - C_L, \quad G := (D - C_L)^{-1}C_U,$$
$$(D - C_L)x^{(k)} := C_U x^{(k-1)} + b.$$

Nadrelaksacja (SOR):

$$Q := \omega^{-1}(D - \omega C_L), \quad G := (D - \omega C_L)^{-1}[\omega C_U + (1 - \omega)D],$$
$$(D - \omega C_L)x^{(k)} := \omega(C_U x^{(k-1)} + b) + (1 - \omega)Dx^{(k-1)}.$$

Nadrelaksacja symetryczna (SSOR):

$$Q := [\omega(2 - \omega)]^{-1}(D - \omega C_L)D^{-1}(D - \omega C_U),$$
$$G := (D - \omega C_U)^{-1}[\omega C_L + (1 - \omega)D](D - \omega C_L)^{-1}[\omega C_U + (1 - \omega)D],$$
$$(D - \omega C_L)x^{(k-1/2)} := \omega(C_U x^{(k-1)} + b) + (1 - \omega)Dx^{(k-1)},$$
$$(D - \omega C_U)x^{(k)} := \omega(C_L x^{(k-1/2)} + b) + (1 - \omega)Dx^{(k-1/2)}.$$

W ostatniej metodzie, niedyskutowanej wcześniej, każda iteracja zawiera *nadrelaksację w przód*, w której niewiadome wyznacza się w pewnym porządku i *nadrelaksację wstecz*, w której ten porządek jest odwrotny. Odpowiedź na pytanie, jak wybrać optymalne parametry metod SOR i SSOR, jest raczej skomplikowana i dyskutować jej nie będziemy. Metodę SOR rozumie się wyżej tak, jak w komentarzu po tw. 4.6.11, podobnie SSOR. Warto też zauważyć, że metoda Gaussa-Seidela jest szczególnym przypadkiem metody nadrelaksacji dla $\omega = 1$.

Ekstrapolacja

Zbieżność liniowych metod iteracyjnych można poprawić, stosując pewną ogólną technikę zwaną *ekstrapolacją*. Rozważmy znów wzór iteracyjny

$$x^{(k)} := Gx^{(k-1)} + c. \tag{4.6.13}$$

Określamy jednoparametrową rodzinę metod iteracyjnych, zależną od $\gamma \neq 0$, wzorem

$$x^{(k)} := \gamma(Gx^{(k-1)} + c) + (1 - \gamma)x^{(k-1)} = G_\gamma x^{(k-1)} + \gamma c, \tag{4.6.14}$$

gdzie

$$G_\gamma := \gamma G + (1-\gamma)I.$$

Metoda (4.6.13) wynika stąd dla $\gamma = 1$.

Jeśli metoda określona wzorem (4.6.14) jest zbieżna do pewnego x, to przejście do granicy daje równość

$$x = \gamma(Gx + c) + (1-\gamma)x,$$

czyli $x = Gx + c$, gdyż $\gamma \neq 0$. Tak więc ta ogólna metoda daje rozwiązanie tego samego równania dla każdego dopuszczalnego γ. Jeśli $G := I - Q^{-1}A$ i $c := Q^{-1}b$, to równanie $x = Gx + c$ jest identyczne z $Ax = b$.

Chcąc wyznaczyć optymalne γ, posługujemy się pomocniczym twierdzeniem o wartościach własnych.

Twierdzenie 4.6.12. *Jeśli λ jest wartością własną macierzy A, a p jest wielomianem, to $p(\lambda)$ jest wartością własną macierzy $p(A)$.*

Dowód. Niech będzie $Ax = \lambda x$ i $x \neq 0$. Wtedy $A^2x = \lambda Ax = \lambda^2 x$. Przez indukcję dowodzi się, że $A^k x = \lambda x$ $(k \geqslant 0)$, czyli λ^k jest wartością własną macierzy A^k. Jeśli $p(z) = \sum_{k=0}^{m} c_k z^k$, to

$$p(A)x = \sum_{k=0}^{m} c_k A^k x = \sum_{k=0}^{m} c_k \lambda^k x = p(\lambda)x. \qquad \blacksquare$$

Na mocy tw. 4.6.7 warunkiem koniecznym i dostatecznym zbieżności metody ekstrapolacyjnej (4.6.14) jest nierówność $\rho(G_\gamma) < 1$. Przypuśćmy, że nie znamy dokładnie wartości własnych macierzy G, wiemy jednak, że leżą one w przedziale $[a, b]$ na osi rzeczywistej. Zgodnie z tw. 4.6.12 wartości własne macierzy $G_\gamma := \gamma G + (1-\gamma)I$ leżą w przedziale o końcach $\gamma a + 1 - \gamma$ i $\gamma b + 1 - \gamma$. Poniższe twierdzenie wyjaśnia, kiedy można wybrać γ tak, żeby było $\rho(G_\gamma) < 1$.

Twierdzenie 4.6.13. *Jeśli o wartościach własnych macierzy G wiadomo tylko, że leżą w przedziale rzeczywistym $[a, b]$ i jeśli $1 \notin [a, b]$, to najlepszą wartością parametru γ jest $2/(2 - a - b)$. Dla niej jest $\rho(G_\gamma) \leqslant 1 - |\gamma| d$, gdzie $d := \min\{|a-1|, |b-1|\}$.*

Dowód. Niech $\Lambda(A)$ będzie zbiorem wartości własnych macierzy A. Wiadomo, że

$$\rho(G_\gamma) = \max_{\lambda \in \Lambda(G_\gamma)} |\lambda| = \max_{\lambda \in \Lambda(G)} |\gamma\lambda + 1 - \gamma| \leqslant \max_{a \leqslant \lambda \leqslant b} |\gamma\lambda + 1 - \gamma|.$$

Ponieważ $1 \notin [a, b]$, więc albo $a > 1$, albo $b < 1$. Podamy dowód w drugim przypadku; pierwszy zostawiamy jako ćwiczenie. Mamy więc $a \leqslant b < 1$ i $d = 1 - b$. Niech będzie $\gamma := 2/(2 - a - b)$. Wtedy $\gamma > 0$ i każda wartość własna λ macierzy G_γ spełnia nierówność

$$\gamma a + 1 - \gamma \leqslant \lambda \leqslant \gamma b + 1 - \gamma. \tag{4.6.15}$$

Prawa część tego oszacowania daje nierówność $\lambda \leqslant 1 + \gamma(b - 1) = 1 - \gamma d$, a z lewej wynika, że

$$\lambda \geqslant \gamma(a + b - 2) + 1 + \gamma(1 - b) = -1 + \gamma d.$$

W ten sposób wykazaliśmy, że $|\lambda| \leqslant 1 - \gamma d$ i $\rho(G_\gamma) \leqslant 1 - \gamma d$.

Wybrane γ jest w tym sensie optymalne, że jego przesunięcie w prawo zmniejsza liczbę szacującą w (4.6.15) z dołu wartość λ, a przesunięcie w lewo zwiększa analogiczne oszacowanie z góry. W obu przypadkach otrzymane oszacowanie dla $\rho(G_\gamma)$ pogorszyłoby się. ■

Warto zauważyć, że opisaną już procedurę ekstarpolacji można zastosować nawet do metod rozbieżnych. Istotne jest tylko to, żeby wartości własne macierzy G były rzeczywiste i leżały po jednej stronie punktu 1.

Niech wartości własne $\lambda_1, \lambda_2, \ldots, \lambda_n$ macierzy A będą rzeczywiste. Definiujemy wielkości

$$m(A) := \min_i \lambda_i, \quad M(A) := \max_i \lambda_i.$$

W twierdzeniu 4.6.13 możemy więc przyjąć, że $a = m(G)$ i $b = M(G)$.

Przykład 4.6.14. Wyznaczyć promień spektralny dla optymalnej ekstrapolowanej metody Richardsona.

Rozwiązanie. W metodzie Richardsona $Q = I$ i $G = I - A$. Jeśli A ma tylko wartości własne rzeczywiste, to jest tak i dla G. Z twierdzenia 4.6.12 wynika, że

$$M(G) = 1 - m(A), \qquad m(G) = 1 - M(A).$$

Jeśli $m(A) > 0$ albo $M(A) < 0$, to można poprawić zbieżność metody Richardsona. Optymalne γ, obliczone według tw. 4.6.13, jest równe

$$\gamma = \frac{2}{m(A) + M(A)},$$

a ponieważ $d = m(A)$, więc promień spektralny jest równy

$$\rho(G_\gamma) = \frac{M(A) - m(A)}{M(A) + m(A)}.$$

■

Przykład 4.6.15. Wyznaczyć promień spektralny dla optymalnej ekstrapolowanej metody Jacobiego i układu przeskalowanego.

Rozwiązanie. Mamy tu metodę Richardsona dla układu $D^{-1}Ax = D^{-1}b$, gdzie $D := \operatorname{diag}(a_{ii})$ (zob. zad. 12). Z przykładu 4.6.14 wynika, że jeśli $m(D^{-1}A) > 0$ albo $M(D^{-1}A) < 0$, to można przyspieszyć zbieżność tej metody. Ponadto dla optymalnego $\gamma := 2/[m(D^{-1}A) + M(D^{-1}A)]$ promień spektralny macierzy opisującej iterację jest równy

$$\rho(G_\gamma) = \frac{M(D^{-1}A) - m(D^{-1}A)}{M(D^{-1}A) + m(D^{-1}A)}.$$ ■

Metoda Czebyszewa

Znacznie ogólniejszy sposób przyspieszania zbieżności liniowych metod iteracyjnych jest nazywany *metodą Czebyszewa*. Jak przedtem, rozważamy metodę typu

$$x^{(k)} := Gx^{(k-1)} + c,$$

która ma dać rozwiązanie x równania $x = Gx + c$. Po jej k-tym kroku znamy wektory $x^{(0)}, x^{(1)}, \ldots, x^{(k)}$. Chcielibyśmy znaleźć taką ich kombinację liniową, która byłaby lepszym od $x^{(k)}$ przybliżeniem dla x. Szukamy jej w postaci

$$u^{(k)} := \sum_{i=0}^{k} a_i^{(k)} x^{(i)}, \quad \text{gdzie} \quad \sum_{i=0}^{k} a_i^{(k)} = 1.$$

W znany już sposób obliczamy

$$u^{(k)} - x = \sum_{i=0}^{k} a_i^{(k)}(x^{(i)} - x) = \sum_{i=0}^{k} a_i^{(k)} G^i(x^{(0)} - x) = P(G)(x^{(0)} - x), \tag{4.6.16}$$

gdzie $P(z) := \sum_{i=0}^{k} a_i^{(k)} z^i$. Wynika stąd oszacowanie dla normy:

$$\|u^{(k)} - x\| \leqslant \|P(G)\| \, \|x^{(0)} - x\|.$$

Norma macierzowa jest tu indukowana przez wektorową, a ta jest dowolna. Na mocy tw. 4.6.6 kres dolny normy $\|P(G)\|$ wzięty po wszystkich normach indukowanych jest równy $\rho(P(G))$; chcemy, żeby ta wielkość była jak najmniejsza. Jeśli wartości własne μ_i macierzy G leżą w pewnym zbiorze ograniczonym S na płaszczyźnie zespolonej, to zgodnie z tw. 4.6.12

$$\rho(P(G)) = \max_{1 \leqslant i \leqslant n} |P(\mu_i)| \leqslant \|P\|_S,$$

gdzie

$$\|P\|_S := \max_{z\in S}|P(z)|. \tag{4.6.17}$$

Należy wybrać wielomian P, dla którego ta norma jest minimalna przy założeniu, że $\sum_{i=0}^{k} a_i = 1$, czyli $P(1) = 1$. Jest to typowy problem teorii aproksymacji. Jego rozwiązanie jest znane tylko dla pewnych S.

Jeśli S jest przedziałem $[a, b]$ na osi rzeczywistej, nie zawierającym punktu 1, to rozwiązaniem jest wielomian Czebyszewa I rodzaju z odpowiednio przesuniętą i przeskalowaną zmienną. Klasyczny wielomian Czebyszewa T_k jest jedynym wielomianem p stopnia k minimalizującym normę $\|p\|_{[-1,1]}$ przy założeniu, że jego współczynnik wiodący (przy z^k) jest równy 2^{k-1}. Własności tych wielomianów są tematem podrozdz. 6.1. Tu są istotne dwie z nich: wielomiany Czebyszewa wyrażają się wzorem

$$T_k(z) = \cos(n \arccos z) \tag{4.6.18}$$

i spełniają związek rekurencyjny

$$T_0(z) = 1, \quad T_1(z) = z, \quad T_k(z) = 2zT_{k-1}(z) - T_{k-2}(z) \qquad (k \geqslant 2). \tag{4.6.19}$$

Załóżmy teraz, że wartości własne macierzy G są zawarte w takim przedziale $[a, b]$, że $b < 1$. Chcemy znaleźć

$$\min_{P_k(1)=1} \|P_k\|_{[a,b]},$$

gdzie P_k jest wielomianem ustalonego stopnia k. Łatwo przewidzieć, że wynik wyraża się przez wielomiany Czebyszewa. Uzasadniają to dwa poniższe lematy, w których Π_k oznacza zbiór wielomianów stopnia k.

Lemat 4.6.16. *Jeśli $p \in \Pi_k$ i $p(\beta) = 1$ dla pewnego $\beta \in \mathbb{R} \setminus (-1, 1)$, to $\|p\|_{[-1,1]} \geqslant |\alpha|$, gdzie $\alpha := 1/T_k(\beta)$.*

Dowód. Przypuśćmy, że p spełnia założenia, ale nie tezę. Z (4.6.18) wynika, że $\|T_k\|_{[-1,1]} = 1$ i że w punktach $t_i := \cos(i\pi/k)$ $(0 \leqslant i \leqslant k)$ wielomian T_k osiąga wartości ekstremalne w przedziale $[-1, 1]$:

$$T_k(t_i) = \cos(k \arccos t_i) = \cos i\pi = (-1)^i.$$

Stąd

$$(-1)^i(\operatorname{sgn}\alpha)[\alpha T_k(t_i) - p(t_i)] \geqslant |\alpha| - \|p\|_{[-1,1]} > 0.$$

To pokazuje, że wielomian $\alpha T_k - p$ w $k+1$ punktach $t_0, t_1, \ldots, t_k$ ma wartości na przemian dodatnie i ujemne, a więc ma co najmniej k zer w przedziale $(-1, 1)$. Ten wielomian znika też w punkcie β. Jest to jednak niemożliwe, gdyż jego stopień nie przewyższa k. ■

Lemat 4.6.17. *Jeśli* $a < b < 1$, $p \in \Pi_k$ *i* $p(1) = 1$, *to* $\|p\|_{[a,b]} \geqslant 1/T_k(w(1))$, *gdzie*

$$w(t) := (2t - b - a)/(b - a).$$

Nierówność dla normy staje się równością dla $p(t) := T_k(w(t))/T_k(w(1))$.

Dowód. Niech będzie $\beta := w(1)$ i $p(t) := q(w(t))$, gdzie $q \in \Pi_k$. Zauważmy, że $1 = p(1) = q(w(1)) = q(\beta)$ i $\beta > 1$. Z lematu 4.6.16 wynika, że $\|q\|_{[-1,1]} \geqslant 1/T_k(\beta)$, czyli $\|p\|_{[a,b]} \geqslant 1/T_k(\beta)$. Jeśli $p(t) = T_k(w(t))/T_k(w(1))$, to oczywiście $p(1) = 1$ i $\|p\|_{[a,b]} = \|T_k\|_{[-1,1]}/T_k(w(1)) = 1/T_k(w(1))$. ■

Lemat 4.6.18. *Wielomiany*

$$P_k(t) := T_k(w(t))/T_k(w(1))$$

spełniają związek rekurencyjny

$$P_0(t) = 1, \quad P_1(t) = (2t - b - a)/(2 - b - a),$$
$$P_k(t) = \rho_k P_1(t) P_{k-1}(t) + (1 - \rho_k) P_{k-2}(t) \qquad (k \geqslant 2),$$

gdzie dla $\alpha := [2w(1)]^{-2}$ *jest*

$$\rho_1 := 2, \quad \rho_k := (1 - \alpha\rho_{k-1})^{-1} \qquad (k \geqslant 2).$$

Dowód. Niech będzie $\beta_k = T_k(w(1))$. Ze wzoru rekurencyjnego (4.6.19) i równości $\beta_k P_k(t) = T_k(w(t))$ wynika, że

$$\begin{aligned} P_k(t) &= \beta_k^{-1} T_k(w(t)) = \beta_k^{-1}[2w(t)T_{k-1}(w(t)) - T_{k-2}(w(t))] = \\ &= 2\beta_k^{-1}\beta_{k-1} w(t) P_{k-1}(t) - \beta_k^{-1}\beta_{k-2} P_{k-2}(t) = \\ &= 2\beta_k^{-1}\beta_{k-1} w(1) P_1(t) P_{k-1}(t) - \beta_k^{-1}\beta_{k-2} P_{k-2}(t). \end{aligned}$$

Użyta tu równość $w(t) = w(1)P_1(t)$ wynika z postaci wielomianu T_1. Jest wygodnie wprowadzić oznaczenie $\rho_k := 2\beta_k^{-1}\beta_{k-1} w(1) = \alpha^{-1/2}\beta_k^{-1}\beta_{k-1}$. Stosując ponownie wzór (4.6.19), obliczamy

$$\begin{aligned} \beta_k &= T_k(w(1)) = 2w(1)T_{k-1}(w(1)) - T_{k-2}(w(1)) = \alpha^{-1/2}\beta_{k-1} - \beta_{k-2}, \\ 1 &= 2w(1)\beta_k^{-1}\beta_{k-1} - \beta_k^{-1}\beta_{k-2} = \rho_k - \beta_k^{-1}\beta_{k-2} \end{aligned}$$

i otrzymany wzór rekurencyjny dla P_k można uprościć do postaci

$$P_k = \rho_k P_1 P_{k-1} + (1-\rho_k)P_{k-2}.$$

Pozostaje sprawdzić, że współczynniki ρ_k też można obliczać rekurencyjnie:

$$\begin{aligned}\rho_k &= \alpha^{-1/2}\beta_k^{-1}\beta_{k-1} = \alpha^{-1/2}\beta_{k-1}(\alpha^{-1/2}\beta_{k-1} - \beta_{k-2})^{-1} = \\ &= \alpha^{-1/2}(\alpha^{-1/2} - \beta_{k-1}^{-1}\beta_{k-2})^{-1} = \alpha^{-1}(\alpha^{-1} - \alpha^{-1/2}\beta_{k-1}^{-1}\beta_{k-2})^{-1} = \\ &= \alpha^{-1}(\alpha^{-1} - \rho_{k-1})^{-1} = (1-\alpha\rho_{k-1})^{-1}.\end{aligned}$$ ■

LEMAT 4.6.19. *Wektory $u^{(k)}$ zdefiniowane w metodzie Czebyszewa można obliczać, dla dowolnego wektora początkowego $u^{(0)}$, dla $\gamma := 2/(2-b-a)$ i wielkości ρ_k określonych w lem. 4.6.18, stosując wzory*

$$u^{(1)} = \gamma(Gu^{(0)} + c) + (1-\gamma)u^{(0)},$$
$$u^{(k)} = \rho_k[\gamma(Gu^{(k-1)} + c) + (1-\gamma)u^{(k-1)}] + (1-\rho_k)u^{(k-2)} \qquad (k \geqslant 2).$$

Dowód. Jeśli

$$u^{(k)} = \sum_{i=0}^{k} a_i^{(k)} x^{(i)}, \qquad P_k(t) = \sum_{i=0}^{k} a_i^{(k)} t^i$$

(druga równość wynika z definicji wielomianu P_k), to $u^{(0)} = x^{(0)}$, a $u^{(1)}$ wyraża się jak w dowodzonym lemacie. Z (4.6.16) i lem. 4.6.18 wynika, że dla rozwiązania x równania $x = Gx + c$ jest

$$\begin{aligned}u^{(k)} - x &= P_k(G)(u^{(0)} - x) = \\ &= [\rho_k P_1(G)P_{k-1}(G) + (1-\rho_k)P_{k-2}(G)](u^{(0)} - x) = \\ &= \rho_k P_1(G)(u^{(k-1)} - x) + (1-\rho_k)(u^{(k-2)} - x), \\ u^{(k)} &= \rho_k P_1(G)u^{(k-1)} + (1-\rho_k)u^{(k-2)} + \rho_k[I - P_1(G)]x.\end{aligned}$$

Ponieważ $P_1(G) = \gamma G + (1-\gamma)I$, więc ostatni składnik jest równy $\rho_k\gamma c$, a całe wyrażenie dla $u^{(k)}$ przybiera szukaną postać. ■

Jak w analizie metody ekstrapolacji, możemy oszacować z góry promień spektralny macierzy $P_k(G)$:

$$\rho(P_k(G)) = \max_{\lambda\in\Lambda(P_k(G))}|\lambda| = \max_{\lambda\in\Lambda(G)}|P_k(\lambda)| \leqslant \|P_k\|_{[a,b]} = 1/T_k(w(1)).$$

Skorzystaliśmy tu z lem. 4.6.17. Obliczając wartość tego oszacowania, można skorzystać z wyniku zad. 33, co daje wzory

$$t := w(1), \qquad b := t + \sqrt{t^2-1}, \qquad \frac{1}{T_k(w(1))} = \frac{2}{b^n + b^{-n}}.$$

Można wykazać, że metoda Czebyszewa jest szybciej zbieżna od metody ekstrapolacji. Dodatkowe szczegóły podają Hageman i Young [1981] oraz Kincaid i Young [1979].

Przykład 4.6.20. Rozwiązać układ

$$\begin{bmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -4 \\ 0 \\ 4 \\ -4 \end{bmatrix},$$

wyrażając go jak w metodzie Jacobiego i stosując metodę Czebyszewa.

Rozwiązanie. Zaczynamy od wyrażenia układu w postaci $D^{-1}Ax = D^{-1}b$, jak w metodzie Jacobiego. Daje to macierz o wartościach własnych w przedziale $[-\frac{1}{2}, \frac{1}{2}]$. Dla wektora początkowego $u^{(0)} := (0, 0, 0, 0)$ po 10 iteracjach otrzymujemy przybliżone rozwiązanie

$$u := (-0.999996, -0.500002, 0.500002, -0.999996). \quad \blacksquare$$

Algorytm metody Czebyszewa można napisać tak [17]:

input u, a, b, M, δ
$\gamma \leftarrow 2/(2-b-a)$
$\alpha \leftarrow [\frac{1}{2}(b-a)/(2-b-a)]^2$
output $0, u$
call Extrap(γ, n, G, c, u, v)
output $1, v$
$\rho \leftarrow 1/(1-2\alpha)$
call Cheb$(\rho, \gamma, n, G, c, u, v)$
output $2, u$
for $k = 3$ **to** M **step** 2 **do**
    $\rho \leftarrow 1/(1-\rho\alpha)$
    call Cheb$(\rho, \gamma, n, G, c, v, u)$
    output k, v
    $\rho \leftarrow 1/(1-\rho\alpha)$
    call Cheb$(\rho, \gamma, n, G, c, u, v)$
    output $k+1, u$
    if $\|u - v\|_\infty < \delta$ **then stop**
end do

[17] Autorzy odstępują tu od konwencji stosowanej wcześniej w programach i w instrukcjach wejścia i wyjścia używają symboli u, v tablic (wektorów) zamiast (u_i), (v_i). Także w instrukcjach podstawienia występują wektory i macierze (np. G), a nie ich składowe lub elementy (*przyp. tłum.*).

Obliczenia zaczynają się od wywołania poniższej procedury `Extrap`, która oblicza $u^{(1)}$ według wzoru z lem. 4.6.19:

procedure `Extrap`(γ, n, G, c, u, v)
$v \leftarrow \gamma c + (1-\gamma)u + \gamma G u$
return

Drugą użytą wyżej procedurą jest `Cheb`, która oblicza $u^{(k)}$ dla $k \geqslant 2$ za pomocą $u^{(k-1)}$ i $u^{(k-2)}$ według wzoru z tegoż lematu; jej dwa różne wywołania zapewniają użycie właściwych danych bez zamiany wektorów u i v.

procedure `Cheb`$(\rho, \gamma, n, G, c, u, v)$
$u \leftarrow \rho[\gamma c + (1-\gamma)v] + (1-\rho)u + \rho\gamma G v$
return

ZADANIA 4.6

1. Wykazać, że jeśli $\|\cdot\|$ jest pewną indukowaną normą macierzy, to jest nią także $\|A\|' := \|SAS^{-1}\|$, gdzie S jest macierzą nieosobliwą.
2. Wykazać, że jeśli $f(z) := \sum_{j=-m}^{m} c_j z^j$ i jeśli λ jest wartością własną macierzy A, to $f(\lambda)$ jest także wartością dla $f(A)$.
3. Udowodnić, że jeśli macierz A jest nieosobliwa i $|\lambda| < \|A^{-1}\|^{-1}$, to λ nie jest jej wartością własną. W nierówności występuje dowolna norma indukowana macierzy.
4. Określając Q jak w metodzie Gaussa-Seidela, udowodnić, że jeśli macierz A jest dominująca przekątniowo, to $\|I - Q^{-1}A\|_\infty < 1$.
5. Znaleźć jawną postać macierzy $G = I - Q^{-1}A$ dla metody Gaussa-Seidela, gdy

$$A := \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}.$$

6. Scharakteryzować rodzinę macierzy nieosobliwych A stopnia n, dla których jeden krok metody Gaussa-Seidela dla zerowego wektora początkowego daje rozwiązanie układu $Ax = b$.
7. Podać przykład macierzy A, dla której metoda Gaussa-Seidela zastosowana do układu $Ax = b$ jest zbieżna, chociaż A nie jest dominująca przekątniowo.
8. Niech macierz A będzie dominująca przekątniowo i niech Q będzie jej częścią trójkątną dolną, jak w metodzie Gaussa-Seidela. Udowodnić, że $\rho(I - Q^{-1}A)$ nie przewyższa największego z ilorazów

$$\Big(\sum_{j=i+1}^{n} |a_{ij}|\Big) \Big/ \Big(|a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}|\Big).$$

9. Wykazać, że podstawowa procedura iteracyjna określona wzorem (4.6.2) jest równoważna następującej: dla danego $x^{(k)}$ obliczamy $r^{(k)} := b - Ax^{(k)}$, wyznaczamy $z^{(k)}$ z równania $Qz^{(k)} = r^{(k)}$ i definiujemy $x^{(k+1)} := x^{(k)} + z^{(k)}$.

10. (cd.). Wykazać, że

$$r^{(k+1)} = (I - AQ^{-1})r^{(k)}, \qquad z^{(k+1)} = (I - Q^{-1}A)z^{(k)}.$$

11. Udowodnić, że jeśli $\delta := \|I - Q^{-1}A\| < 1$, to

$$\|x^{(k)} - x\| \leqslant \frac{\delta}{1-\delta}\|x^{(k)} - x^{(k-1)}\|.$$

12. Udowodnić, że metoda Richardsona zastosowana do układu $Ax = b$ po podzieleniu w nim i-tego równania przez a_{ii} daje to samo co metoda Jacobiego bez takiej zmiany.

13. Wyjaśnić, dlaczego w dowodzie tw. 4.6.5 nie możemy przejść do granicy $\varepsilon \to 0$ i uznać, że A jest podobna do macierzy przekątniowej.

14. Udowodnić, że jeśli macierz A jest dominująca przekątniowo, a Q jest określona jak w metodzie Jacobiego, to $\rho(I - Q^{-1}A) < 1$.

15. Udowodnić, że jeśli

$$a_{ii} = 1 > \sum_{j=1,\, j\neq i}^{n} |a_{ij}| \qquad (1 \leqslant i \leqslant n),$$

to metoda Richardsona jest zbieżna.

16. (cd.). Rozważyć poprzednie zadanie, zmieniając założenie na następujące:

$$a_{jj} = 1 > \sum_{i=1,\, i\neq j}^{n} |a_{ij}| \qquad (1 \leqslant j \leqslant n).$$

17. (cd.). Wykazać, że dla macierzy A z zad. 15 następująca metoda jest zbieżna:

for $k = 1$ **to** ...
  for $i = 1$ **to** n **do**
    $x_i \leftarrow x_i + b_i - \sum_{j=1}^{n} a_{ij}x_j$
  end do
end do

18. Którą z własności normy ma promień spektralny ρ, a której nie ma? Podać dowody i przykłady.

19. Wykazać, że zbiór macierzy trójkątnych górnych, ustalonego stopnia n, jest przestrzenią wektorową. Udowodnić, że promień spektralny ρ jest w tej przestrzeni *pseudonormą*, tj. ma wszystkie własności normy z tym wyjątkiem, że dla $A \neq 0$ może być $\rho(A) = 0$.

20. Udowodnić, że $\rho(A) < 1$ wtedy i tylko wtedy, gdy $\lim_{k\to\infty} A^k x = 0$ dla każdego x.

21. Czy istnieje taka macierz A, że $\rho(A) < \|A\|$ dla każdej normy indukowanej?

22. Czy nierówność $\rho(AB) \leqslant \rho(A)\rho(B)$ zachodzi dla dowolnych macierzy stopnia n? Czy odpowiedź będzie taka sama, jeśli ograniczymy się do macierzy trójkątnych górnych?

23. Wykazać, że jeśli macierze A i B są nieosobliwe, to $\rho(AB) = \rho(BA)$.

24. Udowodnić, że wszystkie wartości własne macierzy hermitowskiej są rzeczywiste. Wskazówka: Rozważyć $\langle x, Ax\rangle$ i $\langle Ax, x\rangle$.

25. Udowodnić, że jeśli macierz A jest nieosobliwa, to AA^{H} jest dodatnio określona.

26. Udowodnić, że jeśli macierz jest dodatnio określona, to jej wartości własne są dodatnie.

27. Wykazać, że twierdzenia z poprzedniego zadania nie można odwrócić.

28. Udowodnić, że jeśli A jest dodatnio określona, to takie są również macierze $A^2, A^3, \ldots$ i $A^{-1}, A^{-2}, \ldots$

29. Czy zbiór macierzy hermitowskich stopnia n tworzy przestrzeń wektorową nad $\mathbb{C}$?

30. Jakie warunki ma spełniać macierz przekątniowa D, żeby dla dowolnej macierzy dodatnio określonej A tę samą własność miał iloczyn DA?

31. Udowodnić, że macierz dodatnio określona A jest hermitowska, używając definicji tej pierwszej, czyli spełnienia nierówności $x^{\mathsf{H}}Ax > 0$ dla każdego $x \neq 0 \in \mathbb{C}^n$. W szczególności przyjęcie takiej definicji (zamiast $x^{\mathsf{T}}Ax > 0$ dla każdego $x \neq 0 \in \mathbb{R}^n$) oznaczałoby, że macierz rzeczywista dodatnio określona musi być symetryczna.

32. Udowodnić tw. 4.6.13 dla $a > 1$.

33. Udowodnić, że $T_n(t) = \frac{1}{2}(b^n + b^{-n})$, gdzie $b = t + \sqrt{t^2 - 1}$.

34. Jak upraszcza się metoda Czebyszewa, jeśli stosuje się ją do metody Jacobiego?

Zadania komputerowe 4.6

K1. Zaprogramować metodę Gaussa-Seidela i sprawdzić ją dla układów

(a)
$$\begin{aligned} 3x + y + z &= 5 \\ x + 3y - z &= 3 \\ 3x + y - 5z &= -1 \end{aligned}$$

(b)
$$\begin{aligned} 3x + y + z &= 5 \\ 3x + y - 5z &= -1 \\ x + 3y - z &= 3 \end{aligned}$$

Zbadać, co tu daje zastosowanie zwykłej eliminacji Gaussa bez wyboru elementów głównych.

K2. Zastosować metodę Gaussa-Seidela do układu

$$\begin{aligned} 0.96326x + 0.81321y &= 0.88824 \\ 0.81321x + 0.68654y &= 0.74988 \end{aligned}$$

dla wektora początkowego (0.33116, 0.70000). Wyjaśnić, co to dało.

4.7. Metody najszybszego spadku i sprzężonych gradientów

Metody opisane w tym podrozdziale odnoszą się do układów $Ax = b$ z macierzą A rzeczywistą i dodatnio określoną, stopnia n. Zakładamy więc, że

$$A^{\mathsf{T}} = A, \quad x^{\mathsf{T}}Ax > 0 \qquad \text{dla} \quad x \neq 0.$$

Używany dalej iloczyn skalarny wektorów x i y z przestrzeni $\mathbb{R}^n$ jest określony wzorem

$$\langle x, y\rangle := x^{\mathsf{T}}y = \sum_{i=1}^{n} x_i y_i.$$

Ma on następujące oczywiste własności:

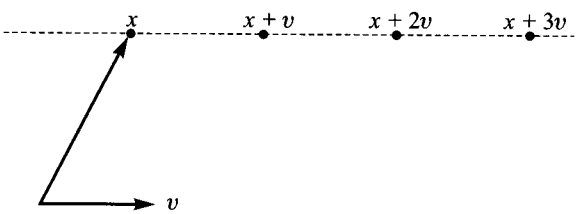
1. $\langle x, y\rangle = \langle y, x\rangle$.
2. $\langle x, \alpha y + \beta z\rangle = \alpha\langle x, y\rangle + \beta\langle x, z\rangle$ dla dowolnych rzeczywistych α i β.
3. $\langle x, x\rangle \geqslant 0$ i $\langle x, x\rangle = 0$ wtedy i tylko wtedy, gdy $x = 0$.
4. $\langle x, Ay\rangle = \langle A^{\mathsf{T}}x, y\rangle$.

Wobec własności **1** można w **2** i **3** przestawić argumenty iloczynu skalarnego[18].

Sprawdzimy najpierw, że pewne zadania numeryczne związane z A i b są równoważne.

Lemat 4.7.1. *Jeśli macierz A jest symetryczna i dodatnio określona, to rozwiązywanie układu $Ax = b$ jest równoważne obliczaniu minimum formy kwadratowej*

$$q(x) := \langle x, Ax\rangle - 2\langle x, b\rangle.$$



RYS. 4.2. Przykład promienia

[18] Warto zwrócić uwagę na różnice – i podobieństwa – definicji iloczynu skalarnego dla wektorów $\mathbb{R}^n$ (zob. wyżej) i $\mathbb{C}^n$ (podrozdz. 4.6 – fragment o metodzie nadrelaksacji – i podrozdz. 5.0). W szczególności tu zamiast operacji A^{H} występuje jej wariant rzeczywisty A^{T} (*przyp. tłum.*).

Dowód. Sprawdźmy najpierw, jak funkcja q zmienia się wzdłuż ustalonego promienia, tj. dla argumentu postaci $x + tv$, gdzie x i v są ustalonymi wektorami, a t przebiega wartości rzeczywiste. Rysunek 4.2 pokazuje (linia przerywana) przykład takiego promienia. Uwzględniając równość $A^{\mathsf{T}} = A$, łatwo obliczyć

$$\begin{aligned} q(x+tv) &= \langle x+tv, A(x+tv)\rangle - 2\langle x+tv, b\rangle = \\ &= \langle x, Ax\rangle + t\langle x, Av\rangle + t\langle v, Ax\rangle + t^2\langle v, Av\rangle - 2\langle x, b\rangle - 2t\langle v, b\rangle = \\ &= q(x) + 2t\langle v, Ax\rangle - 2t\langle v, b\rangle + t^2\langle v, Av\rangle = \\ &= q(x) + 2t\langle v, Ax - b\rangle + t^2\langle v, Av\rangle. \end{aligned}$$

Zauważmy, że współczynnik przy t^2 jest dodatni. Dlatego forma q na promieniu ma minimum, a nie maksimum. Obliczamy pochodną względem t otrzymanego wyrażenia:

$$\frac{d}{dt}\, q(x+tv) = 2\langle v, Ax - b\rangle + 2t\langle v, Av\rangle.$$

Znika ona w punkcie

$$\hat{t} := \frac{\langle v, b - Ax\rangle}{\langle v, Av\rangle}, \tag{4.7.1}$$

a wartość minimalna tam osiągnięta jest równa

$$\begin{aligned} q(x+\hat{t}v) &= q(x) + \hat{t}[2\langle v, Ax - b\rangle + \hat{t}\langle v, Av\rangle] = \\ &= q(x) + \hat{t}[2\langle v, Ax - b\rangle + \langle v, b - Ax\rangle] = \\ &= q(x) - \hat{t}\langle v, b - Ax\rangle = q(x) - \frac{\langle v, b - Ax\rangle^2}{\langle v, Av\rangle}. \end{aligned}$$

Tak więc przejście od x do $x + \hat{t}v$ nie zmniejsza wartości formy q tylko wtedy, gdy wektor v jest ortogonalny względem residuum $b - Ax$, tzn. gdy $\langle v, b - Ax\rangle = 0$. Jeśli x nie jest rozwiązaniem układu $Ax = b$, to istnieje wiele wektorów v takich, że $(v, b - Ax) \neq 0$. Zatem takie x nie minimalizuje formy q. Jeśli natomiast $Ax = b$, to nie ma promienia wychodzącego z x, na którym q miałoby wartość mniejszą od $q(x)$, czyli q osiąga minimum w x. ■

Przeprowadzony dowód sugeruje pewną metodę iteracyjną rozwiązywania układu $Ax = b$. Startuje ona z pewnego wektora początkowego $x^{(0)}$. W k-tym kroku metody ($k = 0, 1, \ldots$) zgodnie z pewną regułą wybieramy kierunek $v^{(k)}$ promienia i wektor

$$x^{(k+1)} := x^{(k)} + t_k v^{(k)}, \tag{4.7.2}$$

gdzie

$$t_k := \frac{\langle v^{(k)}, b - Ax^{(k)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle}$$

dający minimum formy q na tym promieniu.

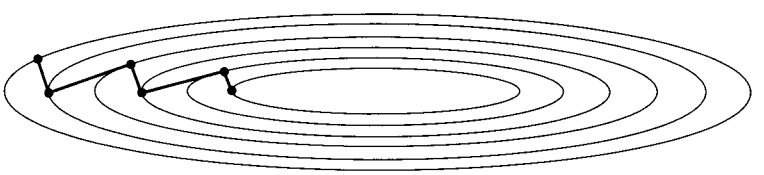
Dodajmy, że wzór (4.7.2) opisuje wiele metod iteracyjnych; różnią się one sposobem wyznaczania liczb t_k i wektorów $v^{(k)}$. Jeśli $\|v^{(k)}\| = 1$, to t_k jest miarą odległości $x^{(k+1)}$ od $x^{(k)}$.

Metoda najszybszego spadku

Do metod określonych wyżej należy *metoda najszybszego spadku*, w której wektorem $v^{(k)}$ jest gradient, ze zmienionym znakiem, formy q w $x^{(k)}$. Trzeba tu przypomnieć, że *gradient* funkcji $g\colon \mathbb{R}^n \to \mathbb{R}$ jest z definicji wektorem o składowych $\partial g/\partial x_i$ $(i = 1, 2, \ldots, n)$. W takim przypadku gradient jest proporcjonalny do wektora residualnego $r^{(k)} := b - Ax^{(0)}$ (zob. zad. 3). Algorytm najszybszego spadku jest więc następujący (M oznacza liczbę wykonywanych kroków):

input x, A, b, M
output $0, x$
for $k = 1$ **to** M **do**
    $v \leftarrow b - Ax$
    $t \leftarrow \langle v, v \rangle / \langle v, Av \rangle$
    $x \leftarrow x + tv$
    output k, x
end do

Uwzględniono tu oczywisty fakt, że kolejny wektor $x^{(k+1)}$ można zapamiętać na miejscu poprzedniego; to samo dotyczy wektorów kierunkowych $v^{(k)}$.



RYS. 4.3. Interpretacja geometryczna metody najszybszego spadku

Wadą metody najszybszego spadku jest jej wolna zbieżność. Można to odczytać z rys. 4.3. Pokazuje on dla zadania dwuwymiarowego krzywe, na których forma kwadratowa q ma tę samą wartość.

Metody sprzężonych kierunków

Rodzina *metod sprzężonych kierunków* realizuje opisaną już strategię minimalizacji funkcji kwadratowej wzdłuż kolejnych promieni. Zwykle ich kierunki są wyznaczane w trakcie obliczeń, ale najpierw rozważymy przypadek, gdy wybiera się je z góry.

Niech A będzie macierzą symetryczną dodatnio określoną stopnia n. Załóżmy, że znamy wektory $u^{(1)}, u^{(2)}, \ldots, u^{(n)}$, które są *$A$-ortonormalne*, tj.

$$\langle u^{(i)}, Au^{(j)} \rangle = \delta_{ij} \qquad (1 \leqslant i, j \leqslant n).$$

Ta własność jest oczywistym uogólnieniem zwykłej ortonormalności. Zobaczymy, że jeśli takie właśnie wektory wyznaczają kierunki promieni używanych w stopniowej minimalizacji formy q, to n-ty krok daje ostateczne rozwiązanie. Przed ścisłym sformułowaniem tego faktu warto zauważyć, że A-ortonormalność można wyrazić jednym równaniem

$$U^{\mathsf{T}} AU = I,$$

gdzie kolumnami macierzy U stopnia n są wektory $u^{(1)}, u^{(2)}, \ldots, u^{(n)}$. Stąd od razu wynika, że macierze A i U są nieosobliwe i że te wektory tworzą bazę przestrzeni $\mathbb{R}^n$.

Twierdzenie 4.7.2. *Jeśli układ $(u^{(1)}, u^{(2)}, \ldots, u^{(n)})$ jest A-ortonormalny i jeśli dla dowolnego $x^{(0)} \in \mathbb{R}^n$ wektory $x^{(i)}$ są określone wzorem*

$$x^{(i)} := x^{(i-1)} + \langle b - Ax^{(i-1)}, u^{(i)} \rangle u^{(i)} \qquad (1 \leqslant i \leqslant n),$$

to $Ax^{(n)} = b$.

Dowód. Dla $t_i := \langle b - Ax^{(i-1)}, u^{(i)} \rangle$ powyższy wzór rekurencyjny ma prostszą postać

$$x^{(i)} := x^{(i-1)} + t_i u^{(i)}.$$

Stąd i z A-ortonormalności wnioskujemy, że

$$Ax^{(i)} = Ax^{(i-1)} + t_i Au^{(i)},$$
$$Ax^{(n)} = Ax^{(0)} + t_1 Au^{(1)} + \ldots + t_n Au^{(n)},$$
$$\langle Ax^{(n)} - b, u^{(i)} \rangle = \langle Ax^{(0)} - b, u^{(i)} \rangle + t_i.$$

Przekształcamy jeszcze t_i:

$$t_i = \langle b - Ax^{(i-1)}, u^{(i)} \rangle = \langle b - Ax^{(0)}, u^{(i)} \rangle + \sum_{j=1}^{i-1} \langle Ax^{(j-1)} - Ax^{(j)}, u^{(i)} \rangle =$$
$$= \langle b - Ax^{(0)}, u^{(i)} \rangle - \sum_{j=1}^{i-1} \langle t_j Au^{(j)}, u^{(i)} \rangle = \langle b - Ax^{(0)}, u^{(i)} \rangle.$$

Tak więc różnica $Ax^{(n)} - b$ jest ortogonalna (w zwykłym sensie) względem $u^{(1)}, u^{(2)}, \ldots, u^{(n)}$, a zatem musi być zerem. ■

Opiszemy teraz konkretną realizację tej metody. Wzór

$$\langle x, y\rangle_A := \langle x, Ay\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i y_j$$

określa – co łatwo sprawdzić – iloczyn skalarny, tj. funkcję zmiennych x, y, mającą własności podane na początku tego podrozdziału. Temu iloczynowi odpowiada norma określona wzorem

$$\|x\|_A^2 := \langle x, x\rangle_A.$$

W podrozdziale 5.3 zdefiniowano algorytm Grama-Schmidta, który przekształca dany układ wektorów liniowo niezależnych na układ ortonormalny $\{u^{(1)}, u^{(2)}, \ldots, u^{(n)}\}$. Przyjmijmy, że używamy w tym algorytmie określonego wyżej iloczynu skalarnego. Jeśli wyjściowy układ tworzą wektory jednostkowe $e^{(1)}, e^{(2)}, \ldots, e^{(n)}$, to wzór (5.3.2) się upraszcza:

$$u^{(i)} := \|v^{(i)}\|_A^{-1} v^{(i)}, \quad \text{gdzie} \quad v^{(i)} := e^{(i)} - \sum_{j<i} (Au^{(j)})_i u^{(j)}.$$

Oczywiście każde $u^{(i)}$ jest kombinacją liniową wektorów $e^{(1)}, e^{(2)}, \ldots, e^{(i)}$, wobec czego macierz U, której kolumnami są wektory $u^{(1)}, u^{(2)}, \ldots, u^{(n)}$, jest trójkątna górna. Wspomniana wcześniej równość $U^{\mathsf{T}}AU = I$ daje wzór $A = (U^{\mathsf{T}})^{-1}U^{-1}$, czyli rozkład typu LU. Dzięki temu układ $Ax = b$ można rozwiązać, stosując wariant eliminacji Gaussa bez wyboru elementów głównych.

W praktyce numerycznej są wygodniejsze układy *A-ortogonalne*. Nazywamy tak układ wektorów niezależnych liniowo $v^{(1)}, v^{(2)}, \ldots$, dla którego $\langle v^{(i)}, Av^{(j)}\rangle = 0$ $(i \neq j)$. Daje on układ A-ortonormalny $u^{(i)} := \|v^{(i)}\|_A^{-1} v^{(i)}$.

Twierdzenie 4.7.3. *Jeśli macierz A stopnia n jest symetryczna i dodatnio określona, a układ wektorów $\{v^{(1)}, v^{(2)}, \ldots, v^{(n)}\}$ jest A-ortogonalny, to dla dowolnego $x^{(0)}$ wzór*

$$x^{(i)} := x^{(i-1)} + \frac{\langle b - Ax^{(i-1)}, v^{(i)}\rangle}{\langle v^{(i)}, Av^{(i)}\rangle} v^{(i)} \qquad (1 \leqslant i \leqslant n)$$

określa takie wektory, że $Ax^{(n)} = b$.

Metoda sprzężonych gradientów

Metoda sprzężonych gradientów (Hestenes i Stiefel [1952]) należy do rodziny metod sprzężonych kierunków. Kierunki $v^{(i)}$ z tw. 4.7.3 wybiera się tu w trakcie obliczeń w taki szczególny sposób, że reszty $r^{(i)} := b - Ax^{(i)}$ tworzą układ ortogonalny w zwykłym sensie, tzn. jest $\langle r^{(i)}, r^{(j)} \rangle = 0$ dla $i \neq j$.

Metodę sprzężonych gradientów można przedkładać nad zwykłą eliminację Gaussa, jeśli macierz A jest duża i rzadka. W teorii metoda daje rozwiązanie układu $Ax = b$ złożonego z n równań po n krokach. W praktyce jednak, dla układów źle uwarunkowanych, błędy zaokrągleń uniemożliwiają znalezienie w tylu krokach dostatecznie dokładnego rozwiązania. To sprawiło, że entuzjazm wywołany opublikowaniem metody później znacznie osłabł. Jednak około 20 lat po publikacji, gdy metodę sprzężonych gradientów zaczęto traktować jako metodę iteracyjną, generującą ciąg wektorów zbieżny do rozwiązania, doceniono jej wartość. Okazało się nawet, że dla macierzy dobrze uwarunkowanych zadowalające rozwiązanie otrzymujemy po mniejszej od n liczbie kroków. Historię metody opisują Golub i O'Leary [1989].

Pierwsza, „surowa", wersja programu dla metody sprzężonych gradientów jest następująca:

input $A, b, x^{(0)}, M, \varepsilon$
$r^{(0)} \leftarrow b - Ax^{(0)}$
$v^{(0)} \leftarrow r^{(0)}$
output $0, x^{(0)}, r^{(0)}$
for $k = 0$ **to** $M - 1$ **do**
  if $v^{(k)} = 0$ **then stop**
  $t_k \leftarrow \|r^{(k)}\|_2^2 / \langle v^{(k)}, Av^{(k)} \rangle$
  $x^{(k+1)} \leftarrow x^{(k)} + t_k v^{(k)}$
  $r^{(k+1)} \leftarrow r^{(k)} - t_k Av^{(k)}$
  if $\|r^{(k+1)}\|_2^2 < \varepsilon$ **then stop**
  $s_k \leftarrow \|r^{(k+1)}\|_2^2 / \|r^{(k)}\|_2^2$
  $v^{(k+1)} \leftarrow r^{(k+1)} + s_k v^{(k)}$
  output $k + 1, x^{(k+1)}, r^{(k+1)}$
end do

Jeśli $r^{(k)} = 0$, to – teoretycznie – $x^{(k)}$ jest rozwiązaniem układu $Ax = b$. Algorytm wymaga pamiętania czterech wektorów: $x^{(k)}$, $r^{(k)}$, $v^{(k)}$ i $Av^{(k)}$. Pamiętanie pełnej macierzy A nie jest konieczne. Koszt jednej iteracji jest niewielki; trzeba obliczyć iloczyn macierzy przez wektor oraz dwa iloczyny skalarne: $\langle v^{(k)}, Av^{(k)} \rangle$ i $\|r^{(k+1)}\|_2^2$. Jest to dobrze widoczne w drugiej wersji programu:

input $A, b, x, M, \varepsilon, \delta$
$r \leftarrow b - Ax$

$v \leftarrow r$
$c \leftarrow \langle r, r\rangle$
for $k = 1$ **to** M **do**
  if $\langle v, v\rangle < \delta$ **then stop**
  $z \leftarrow Av$
  $t \leftarrow c/\langle v, z\rangle$
  $x \leftarrow x + tv$
  $r \leftarrow r - tz$
  $d \leftarrow \langle r, r\rangle$
  if $d < \varepsilon$ **then stop**
  $v \leftarrow r + (d/c)v$
  $c \leftarrow d$
  output k, x, r
end do

TWIERDZENIE 4.7.4. *Jeśli dla pewnego $m < n$ w metodzie sprzężonych gradientów wektory $v^{(0)}, v^{(1)}, \ldots, v^{(m)}$ są różne od 0, to $r^{(i)} = b - Ax^{(i)} \neq 0$ dla $0 \leqslant i \leqslant m$, a układ $\{r^{(0)}, r^{(1)}, \ldots, r^{(m)}\}$ jest ortogonalny.*

Dowód. Udowodnimy nieco więcej, a mianowicie, że:

1. $\langle r^{(m)}, v^{(i)}\rangle = 0$ dla $0 \leqslant i < m$.

2. $\langle r^{(i)}, r^{(i)}\rangle = \langle r^{(i)}, v^{(i)}\rangle$ dla $0 \leqslant i \leqslant m$.

3. $\langle v^{(m)}, Av^{(i)}\rangle = 0$ dla $0 \leqslant i < m$.

4. $r^{(i)} = b - Ax^{(i)}$ dla $0 \leqslant i \leqslant m$.

5. $\langle r^{(m)}, r^{(i)}\rangle = 0$ dla $0 \leqslant i < m$.

6. $r^{(i)} \neq 0$ dla $0 \leqslant i \leqslant m$.

Dowód jest indukcyjny względem m. Dla $m = 0$ przy założeniu, że $v^{(0)} \neq 0$, wystarczy wykazać prawdziwość zdań **2**, **4** i **6**. To wynika wprost z definicji metody, gdyż $r^{(0)} = b - Ax^{(0)} = v^{(0)} \neq 0$. Przypuśćmy, że dla pewnego $m \geqslant 0$ zdania **1**–**6** są prawdziwe. Założymy też, że wektory $v^{(0)}, v^{(1)}, \ldots, v^{(m+1)}$ są różne od 0. Trzeba więc dodatkowo udowodnić, że:

1′. $\langle r^{(m+1)}, v^{(i)}\rangle = 0$ dla $0 \leqslant i \leqslant m$.

2′. $\langle r^{(m+1)}, r^{(m+1)}\rangle = \langle r^{(m+1)}, v^{(m+1)}\rangle$.

3′. $\langle v^{(m+1)}, Av^{(i)}\rangle = 0$ dla $0 \leqslant i \leqslant m$.

4′. $r^{(m+1)} = b - Ax^{(m+1)}$.

5′. $\langle r^{(m+1)}, r^{(i)}\rangle = 0$ dla $0 \leqslant i \leqslant m$.

6′. $r^{(m+1)} \neq 0$.

Zdanie **1'**: Dla $i = m$ mamy, na mocy **2**,

$$\langle r^{(m+1)}, v^{(m)}\rangle = \langle r^{(m)} - t_m A v^{(m)}, v^{(m)}\rangle = \langle r^{(m)}, v^{(m)}\rangle - t_m \langle v^{(m)}, A v^{(m)}\rangle = \\ = \langle r^{(m)}, v^{(m)}\rangle - \langle r^{(m)}, r^{(m)}\rangle = 0.$$

Dla $0 \leqslant i < m$ na mocy **1** i **3**

$$\langle r^{(m+1)}, v^{(i)}\rangle = \langle r^{(m)}, v^{(i)}\rangle - t_m \langle v^{(m)}, A v^{(i)}\rangle = 0.$$

Zdanie **2'**: Z **1'** wynika, że

$$\langle r^{(m+1)}, v^{(m+1}\rangle = \langle r^{(m+1)}, r^{(m+1)} + s_m v^{(m)}\rangle = \langle r^{(m+1)}, r^{(m+1)}\rangle.$$

Zdanie **3'**: Niech będzie $s_{-1} := 0$ i $v^{(-1)} := 0$. Dla $0 \leqslant i \leqslant m$ mamy

$$\begin{aligned}\langle v^{(m+1)}, A v^{(i)}\rangle &= \langle r^{(m+1)} + s_m v^{(m)}, A v^{(i)}\rangle = \\ &= \langle r^{(m+1)}, A v^{(i)}\rangle + s_m \langle v^{(m)}, A v^{(i)}\rangle = \\ &= t_i^{-1} \langle r^{(m+1)}, r^{(i)} - r^{(i-1)}\rangle + s_m \langle v^{(m)}, A v^{(i)}\rangle = \\ &= t_i^{-1} \langle r^{(m+1)}, v^{(i)} - s_{i-1} v^{(i-1)} - v^{(i+1)} + s_i v^{(i)}\rangle + \\ &\quad + s_m \langle v^{(m)}, A v^{(i)}\rangle = \\ &= t_i^{-1} \big[\langle r^{(m+1)}, v^{(i)}\rangle - s_{i-1} \langle r^{(m+1)}, v^{(i-1)}\rangle - \\ &\quad - \langle r^{(m+1)}, v^{(i+1)}\rangle + s_i \langle r^{(m+1)}, v^{(i)}\rangle\big] + s_m \langle v^{(m)}, A v^{(i)}\rangle.\end{aligned}$$

Jeśli $i < m$, to wobec **1'** wektor $r^{(m+1)}$ jest ortogonalny względem $v^{(j)}$ dla $j = i-1, i, i+1$. Natomiast z **3** wynika, że $\langle v^{(m)}, A v^{(i)}\rangle = 0$. Dlatego $\langle v^{(m+1)}, A v^{(i)}\rangle = 0$. Przypadek $i = m$ trzeba zbadać oddzielnie. Z ostatniego ciągu równości wynika, że

$$\langle v^{(m+1)}, A v^{(m)}\rangle = \\ = t_m^{-1} \langle r^{(m+1)}, v^{(m)} - s_{m-1} v^{(m-1)} - v^{(m+1)} + s_m v^{(m)}\rangle + s_m \langle v^{(m)}, A v^{(m)}\rangle.$$

Na mocy **1'** wektor $r^{(m+1)}$ jest ortogonalny względem $v^{(m)}$ i $v^{(m-1)}$. Dlatego

$$\langle v^{(m+1)}, A v^{(m)}\rangle = -t_m^{-1} \langle r^{(m+1)}, v^{(m+1)}\rangle + s_m \langle r^{(m)}, A v^{(m)}\rangle = \\ = -\frac{\langle v^{(m)}, A v^{(m)}\rangle}{\langle r^{(m)}, r^{(m)}\rangle} \langle r^{(m+1)}, v^{(m+1)}\rangle + \frac{\langle r^{(m+1)}, r^{(m+1)}\rangle}{\langle r^{(m)}, r^{(m)}\rangle} \langle v^{(m)}, A v^{(m)}\rangle.$$

Wobec **2'** to wyrażenie znika.

Zdanie **4'**:

$$b - A x^{(m+1)} = b - A(x^{(m)} + t_m v^{(m)}) = b - A x^{(m)} - t_m A v^{(m)} = \\ = r^{(m)} - (r^{(m)} - r^{(m+1)}) = r^{(m+1)}.$$

Zdanie **5′**: Przy dodatkowych oznaczeniach jak w dowodzie **3′** ze zdania **1′** wynika, że

$$\langle r^{(m+1)}, r^{(i)} \rangle = \langle r^{(m+1)}, v^{(i)} - s_{i-1} v^{(i-1)} \rangle = \\ = \langle r^{(m+1)}, v^{(i)} \rangle - s_{i-1} \langle r^{(m+1)}, v^{(i-1)} \rangle = 0.$$

Zdanie **6′**: Z **3′** i dodatniej określoności macierzy A wynika, że

$$0 < \langle v^{(m+1)}, A v^{(m+1)} \rangle = \langle r^{(m+1)} + s_m v^{(m)}, A v^{(m+1)} \rangle = \\ = \langle r^{(m+1)}, A v^{(m+1)} \rangle + s_m \langle v^{(m)}, A v^{(m+1)} \rangle = \\ = \langle r^{(m+1)}, A v^{(m+1)} \rangle,$$

wobec czego $r^{(m+1)} \neq 0$. ■

To twierdzenie i jego dowód wzorowano na tekście Stoera i Bulirscha [1980].

Metoda sprzężonych gradientów ze wstępnym przekształceniem

Chcemy rozwiązać układ $Ax = b$ z macierzą symetryczną i dodatnio określoną, stosując pewien wariant metody sprzężonych gradientów. Jest celowe takie *wstępne przekształcenie* układu[19], aby nowy układ

$$\hat{A}\hat{x} = \hat{b} \tag{4.7.3}$$

był lepiej uwarunkowany. Ma on być taki, że

$$\hat{A} = S^{\mathsf{T}} A S, \quad \hat{x} = S^{-1} x, \quad \hat{b} = S^{\mathsf{T}} b$$

i $\varkappa(\hat{A}) < \varkappa(A)$. To przekształcenie może też spowodować szybszą zbieżność nowego układu.

Układ $\hat{A}\hat{x} = \hat{b}$ można oczywiście rozwiązywać, stosując pierwszy z programów podanych wcześniej dla metody sprzężonych gradientów; zamiast A, b, $x^{(k)}$, $r^{(k)}$, $v^{(k)}$, t_k, s_k w programie będą teraz występowały analogiczne wielkości, ale opatrzone znakiem ^.

Macierz S nie może być dowolna. Z powodów, które częściowo wyjaśnią się dalej, dla pewnej macierzy symetrycznej i dodatnio określonej Q określamy S tak, że

$$Q^{-1} = S S^{\mathsf{T}}.$$

[19] W oryginale występuje tu *preconditioning*. Kiełbasiński i Schwetlick [*1992] tłumaczą ten termin jako *prekondicjoning*, ale trudno uznać tę propozycję za udaną (*przyp. tłum.*).

W zastosowaniach macierz A bywa rzadka i przejście do $\hat{A}$ może tę istotną własność zniszczyć. Dlatego będziemy używać pierwotnego układu, ukrywając jego wstępne przekształcenie w algorytmie. Aby zrozumieć sens tego postępowania, przyjmijmy, że

$$\begin{aligned}
\hat{x}^{(k)} &:= S^{-1}x^{(k)},\\
\hat{v}^{(k)} &:= S^{-1}v^{(k)},\\
\hat{r}^{(k)} &:= \hat{b} - \hat{A}\hat{x}^{(k)} = S^{\mathsf{T}}b - (S^{\mathsf{T}}AS)(S^{-1}x^{(k)}) = S^{\mathsf{T}}r^{(k)},\\
\tilde{r}^{(k)} &:= Q^{-1}r^{(k)}.
\end{aligned}$$

Wielkości $\hat{t}_k$ i $\hat{s}_k$ zależą od określonych wyżej podobnie jak we wspomnianym już algorytmie.

Wielkości opatrzone znakiem $\hat{}$ wyrazimy przez analogiczne do nich, ale odnoszące się do pierwotnego układu (i przez macierz Q). Skorzystamy przy tym wielokrotnie ze znanej już tożsamości $\langle x, Cy\rangle = \langle C^{\mathsf{T}}x, y\rangle$. Mamy więc

$$\begin{aligned}
\hat{t}_k &= \langle \hat{r}^{(k)}, \hat{r}^{(k)}\rangle / \langle \hat{v}^{(k)}, \hat{A}\hat{v}^{(k)}\rangle =\\
&= \langle S^{\mathsf{T}}r^{(k)}, S^{\mathsf{T}}r^{(k)}\rangle / \langle S^{-1}v^{(k)}, (S^{\mathsf{T}}AS)(S^{-1}v^{(k)}\rangle =\\
&= \langle Q^{-1}r^{(k)}, r^{(k)}\rangle / \langle v^{(k)}, Av^{(k)}\rangle = \langle \tilde{r}^{(k)}, r^{(k)}\rangle / \langle v^{(k)}, Av^{(k)}\rangle.
\end{aligned}$$

W podobny sposób obliczamy

$$\begin{aligned}
\hat{s}_k &= \langle \hat{r}^{(k+1)}, \hat{r}^{(k+1)}\rangle / \langle \hat{r}^{(k)}, \hat{r}^{(k)}\rangle =\\
&= \langle S^{\mathsf{T}}r^{(k+1)}, S^{\mathsf{T}}r^{(k+1)}\rangle / \langle S^{\mathsf{T}}r^{(k)}, S^{\mathsf{T}}r^{(k)}\rangle =\\
&= \langle Q^{-1}r^{(k+1)}, r^{(k+1)}\rangle / \langle Q^{-1}r^{(k)}, r^{(k)}\rangle = \langle \tilde{r}^{(k+1)}, r^{(k+1)}\rangle / \langle \tilde{r}^{(k)}, r^{(k)}\rangle.
\end{aligned}$$

Prócz tego, mnożąc stronami równości

$$\hat{x}^{(k+1)} = \hat{x}^{(k)} + \hat{t}_k\hat{v}^{(k)}, \quad \hat{r}^{(k+1)} = \hat{r}^{(k)} - \hat{t}_k\hat{A}\hat{v}^{(k)}, \quad \hat{v}^{(k+1)} = \hat{r}^{(k+1)} + \hat{s}_k\hat{v}^{(k)}$$

odpowiednio przez S, $(S^{\mathsf{T}})^{-1}$ i S otrzymujemy

$$x^{(k+1)} = x^{(k)} + \hat{t}_k v^{(k)}, \quad r^{(k+1)} = r^{(k)} - \hat{t}_k Av^{(k)}, \quad v^{(k+1)} = \tilde{r}^{(k+1)} + \hat{s}_k v^{(k)}.$$

Stąd wynika pierwsza wersja algorytmu sprzężonych gradientów ze wstępnym przekształceniem. Wśród danych znajduje się macierz Q:

input $A, b, x^{(0)}, Q, M, \varepsilon$
$r^{(0)} \leftarrow b - Ax^{(0)}$
wyznaczenie $\tilde{r}^{(0)}$ z układu $Q\tilde{r}^{(0)} = r^{(0)}$
$v^{(0)} \leftarrow r^{(0)}$
output $0, x^{(0)}$

for $k = 0$ **to** $M - 1$ **do**
  if $v^{(k)} = 0$ **then stop**
  $\hat{t}_k \leftarrow \langle \tilde{r}^{(k)}, r^{(k)} \rangle / \langle v^{(k)}, Av^{(k)} \rangle$
  $x^{(k+1)} \leftarrow x^{(k)} + \hat{t}_k v^{(k)}$
  $r^{(k+1)} \leftarrow r^{(k)} - \hat{t}_k Av^{(k)}$
  wyznaczenie $\tilde{r}^{(k+1)}$ z układu $Q\tilde{r}^{(k+1)} = r^{(k+1)}$
  if $\langle \tilde{r}^{(k+1)}, r^{(k+1)} \rangle < \varepsilon$ **then**
    if $\langle r^{(k+1)}, r^{(k+1)} \rangle < \varepsilon$ **then stop**
  end if
  $\hat{s}_k \leftarrow \langle \tilde{r}^{(k+1)}, r^{(k+1)} \rangle / \langle \tilde{r}^{(k)}, r^{(k)} \rangle$
  $v^{(k+1)} \leftarrow \tilde{r}^{(k+1)} + \hat{s}_k v^{(k)}$
  output $k + 1, x^{(k+1)}, r^{(k+1)}$
end do

Dla $Q = I$ ten algorytm jest taki jak dla zwykłej metody sprzężonych gradientów. Jeśli $Q = A$ i $S = A^{-1/2}$, to przekształcony układ (4.7.3) redukuje się do postaci $\hat{x} = \hat{b}$, ale to nic nie daje, gdyż wyznaczenie wektora $\hat{b} = S^{\mathrm{T}} b$ jest tak samo trudne jak rozwiązanie układu $Ax = b$.

Ponieważ w każdym kroku powyższego algorytmu trzeba rozwiązać układ $Qx = y$, więc powinniśmy wybrać macierz Q tak, żeby to było łatwe. Mogłaby to zatem być macierz przekątniowa, ale w innych przypadkach zbieżność bywa szybsza. Im lepszym przybliżeniem macierzy A jest Q^{-1}, tym lepiej uwarunkowany jest układ (4.7.3) i tym mniej kroków procesu iteracyjnego trzeba wykonać. Z drugiej strony, rozwiązanie układu $Qx = y$ staje się wtedy trudniejsze. Ilustruje to typowy dylemat, jaką metodę iteracyjną wybrać: wymagającą niewiele kosztownych kroków, czy taką, w której te kroki są mało kosztowne, ale trzeba ich wykonać wiele.

Komentarza wymaga jeszcze moment zakończenia obliczeń. W przeciwieństwie do $\langle \tilde{r}^{(k+1)}, r^{(k+1)} \rangle$ iloczyn $\langle r^{(k+1)}, r^{(k+1)} \rangle$ nie jest w algorytmie potrzebny do innych celów. Dlatego ten drugi iloczyn znajdujemy tylko wtedy, gdy pierwszy jest dostatecznie mały.

Program komputerowy może bazować na następującym uproszczonym algorytmie:

input $A, b, x, Q, M, \varepsilon, \delta$
$r \leftarrow b - Ax$
wyznaczenie z z układu $Qz = r$
$v \leftarrow z$
$c \leftarrow \langle z, r \rangle$
for $k = 1$ **to** M **do**
if $\langle v, v \rangle < \delta$ **then stop**
  $z \leftarrow Av$
  $t \leftarrow c / \langle v, z \rangle$
  $x \leftarrow x + tv$

```
        r ← r − tz
        wyznaczenie z z układu Qz = r
        d ← ⟨z, r⟩
        if d < ε then
            if ⟨r, r⟩ < ε then stop
        end if
        v ← z + (d/c)v
        c ← d
    output k, x, r
    end do
```

Metoda sprzężonych gradientów ze wstępnym przekształceniem układu równań (tu opisana podobnie jak to zrobił Ortega [1988]) jest przedmiotem intensywnych badań. Proponuje się wiele sposobów wyboru macierzy Q, np. tak jak w metodzie nadrelaksacji. Dodatkowe informacje można znaleźć w wielu książkach; zob. np. Golub i van Loan [1989]. Opracowano też pakiety programów nie tylko dla tej, ale i dla innych metod iteracyjnych. Są to m.in. `ITPACKV 2D` (Kincaid, Oppe i Young [1989]), `NSPCG` (Oppe, Joubert i Kincaid [1988]), `PCGPAK2` [1990] i `PCG` (Joubert i in. [1995]).

Zadania 4.7

1. Udowodnić, że istnienie układu A-ortonormalnego $\{u^{(1)}, u^{(2)}, \ldots, u^{(n)}\}$, gdzie A jest macierzą stopnia n, implikuje jej symetrię i dodatnią określoność.
2. Wykazać, że jeśli $y := x + \hat{t}v$, gdzie $\hat{t}$ jest określone w (4.7.1), to wektory v i $b - Ay$ są ortogonalne.
3. Udowodnić, że gradient funkcji $q(x) := \langle x, Ax\rangle - 2\langle x, b\rangle$ dla macierzy symetrycznej A jest równy $2(Ax - b)$.
4. Udowodnić, że najmniejszą wartością funkcji q jest $-\langle b, A^{-1}b\rangle$.
5. Udowodnić, że jeśli macierz A jest symetryczna, $Ax = b$, a y jest dowolnym wektorem, to $\langle x-y, A(x-y)\rangle = \langle b, A^{-1}b\rangle + q(y)$. Stąd wynika, że minimalizacja funkcji q jest równoważna z tymże zadaniem dla $\langle x - y, A(x - y)\rangle$.
6. Udowodnić, że w metodzie najszybszego spadku jest

   $$q(x^{(k+1)}) = q(x^{(k)}) - \|r^{(k)}\|_2^4 / \langle r^{(k)}, Ar^{(k)}\rangle,$$

   gdzie $r^{(k)} := b - Ax^{(k)}$.
7. Wykazać, że w metodzie najszybszego spadku wektory $v^{(k)}$ i $v^{(k+1)}$ są ortogonalne.
8. Sprawdzić, czy jeśli układ ortonormalny $\{u^{(1)}, u^{(2)}, \ldots, u^{(n)}\}$ wyznacza kierunki poszukiwania minimum funkcji q, to będzie ono znalezione po n krokach.
9. Dla jakich wartości t_k metoda najszybszego spadku sprowadza się do metody Richardsona? Jakie warunki nałożone na t_k i A zapewniają identyczność pierwszej metody z metodą Jacobiego?

10. Dla układu równań

$$\begin{bmatrix} 2 & 0 & -1 \\ -2 & -10 & 0 \\ -1 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -12 \\ 2 \end{bmatrix}$$

i wektora początkowego $x^{(0)} := (0, 0, 0)$ wykonać dwa kroki metody iteracyjnej: (**a**) Jacobiego, (**b**) Gaussa-Seidela, (**c**) sprzężonych gradientów.

11. Udowodnić, że jeśli macierz A jest dodatnio określona, a b jest wektorem, to iloczyn skalarny $\langle b - Ax, A^{-1}b - x\rangle$ jest dodatni, jeśli tylko $Ax \neq b$.

12. Udowodnić, że w metodzie sprzężonych gradientów

$$t_k = \langle r^{(k)}, v^{(k)}\rangle / \langle v^{(k)}, Av^{(k)}\rangle, \qquad s_k = -\langle r^{(k+1)}, Av^{(k)}\rangle / \langle v^{(k)}, Av^{(k)}\rangle.$$

13. Udowodnić, że w metodzie sprzężonych gradientów z równości $v^{(k)} = 0$ wynika, iż $Ax^{(b)} = b$.

ZADANIA KOMPUTEROWE 4.7

K1. Zaprogramować metodę sprzężonych gradientów i sprawdzić ją dla

$$a_{ij} := (i + j + 1)^{-1} \quad (1 \leqslant i, j \leqslant n), \qquad b_i := \frac{1}{3}\sum_{j=1}^{n} a_{ij} \quad (1 \leqslant i \leqslant n)$$

(A jest tu macierzą Hilberta).

K2. Rozwiązać układ

$$\begin{bmatrix} 10 & 1 & 2 & 3 & 4 \\ 1 & 9 & -1 & 2 & -3 \\ 2 & -1 & 7 & 3 & -5 \\ 3 & 2 & 3 & 12 & -1 \\ 4 & -3 & -5 & -1 & 15 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 12 \\ -27 \\ 14 \\ -17 \\ 12 \end{bmatrix}$$

(z macierzą symetryczną i dodatnio określoną, ale nie dominującą przekątniowo) stosując dla $x^{(0)} := 0$ metodę: (**a**) Jacobiego, (**b**) Gaussa-Seidela, (**c**) sprzężonych gradientów.

4.8. Analiza błędów zaokrągleń w metodzie eliminacji Gaussa

W tym podrozdziale zbadamy skutki błędów zaokrągleń, występujących nieuchronnie, gdy rozwiązujemy układ $Ax = b$ równań liniowych o współczynnikach rzeczywistych. Analizę tych błędów dla metody eliminacji Gaussa

z wyborem wierszy głównych zawdzięczamy Wilkinsonowi. Jej wynikami są oszacowania a posteriori błędów [20].

Założymy, że niezbędne przestawienia wierszy wykonano na początku obliczeń, a więc elementy główne znajdują się już na właściwych pozycjach. Stosowana strategia upewnia nas, że wtedy w k-tym kroku algorytmu jest $|a_{kk}^{(k)}| \geqslant |a_{ik}^{(k)}|$ dla $i \geqslant k$ (oznaczenia są takie, jak w podrozdz. 4.3). Dzięki temu mnożniki stosowane w procesie eliminacji mają wartości bezwzględne nie większe od 1. Powyższe założenie nie ogranicza ogólności tw. 4.8.1, jedynie upraszcza oznaczenia.

W rozkładzie LU macierzy A stosujemy następujące wzory:

$$a_{ij}^{(k+1)} := \begin{cases} a_{ij}^{(k)} & (i \leqslant k) \\ a_{ij}^{(k)} - l_{ik}a_{kj}^{(k)} & (i > k,\ j > k) \\ 0 & (j \leqslant k < i), \end{cases} \qquad l_{ik} := \begin{cases} 0 & (i < k) \\ 1 & (i = k) \\ a_{ik}^{(k)}/a_{kk}^{(k)} & (i > k). \end{cases}$$

Na początku obliczeń jest $A^{(1)} = A$, a na końcu $U = A^{(n)}$. Macierz L jest trójkątna dolna, U – trójkątna górna.

Liczby występujące w komputerze będzie sygnalizował symbol ˜; np. $\tilde{l}_{ik}$ jest wielkością, która w obliczeniach reprezentuje l_{ik}. Jej wartość zależy od komputera (i stosowanego języka programowania). Dla uproszczenia zakładamy, że w obliczeniach nie występuje ani nadmiar, ani niedomiar.

Opisując obliczenia komputerowe, będziemy też używać symbolu fl określonego w podrozdz. 2.1. Przypomnijmy, że jeśli x i y są liczbami maszynowymi, a $\odot$ jest symbolem działania arytmetycznego, to komputer zamiast $x \odot y$ daje $\mathrm{fl}(x \odot y)$. Te liczby wiąże równość

$$\mathrm{fl}(x \odot y) = (x \odot y)(1 - \delta) = (x \odot y)/(1 - \delta'), \tag{4.8.1}$$

gdzie

$$|\delta|,\ |\delta'| \leqslant \varepsilon; \tag{4.8.2}$$

stała ε w tym podrozdziale oznacza precyzję arytmetyki w stosowanym komputerze (zob. podrozdz. 2.1).

Uwzględniając powyższe uwagi, wyrażamy w następujący sposób liczby tworzone w czasie eliminacji:

$$\tilde{a}_{ij}^{(k+1)} := \begin{cases} \tilde{a}_{ij}^{(k)} & (i \leqslant k) \\ \mathrm{fl}\big[\tilde{a}_{ij}^{(k)} - \mathrm{fl}(\tilde{l}_{ik}\tilde{a}_{kj}^{(k)})\big] & (i, j > k), \end{cases}$$

[20] Określenie *oszacowanie a posteriori* nie wyjaśnia istoty pomysłu Wilkinsona. Udane określenie zaproponował Kiełbasiński w przekładzie książki Wilkinsona [1963]: *analiza (pozornych) zaburzeń*; np. tw. 4.8.2 orzeka, że *obliczony* iloczyn skalarny jest *dokładnym* iloczynem zaburzonych wektorów (*przyp. tłum.*).

$$\tilde{l}_{ik} := \begin{cases} 1 & (i = k) \\ \mathrm{fl}(\tilde{a}_{ik}^{(k)}/\tilde{a}_{kk}^{(k)}) & (i > k). \end{cases}$$

W dalszym ciągu będzie potrzebna wielokrotnie wielkość

$$\rho := \max_{1 \leqslant i,j,k \leqslant n} |a_{ij}^{(k)}|.$$

Można ją wyznaczyć stosując metodę eliminacji Gaussa.

Twierdzenie 4.8.1. *Niech A będzie macierzą stopnia n, której elementy są liczbami maszynowymi w komputerze z precyzją arytmetyki ε. Metoda Gaussa z wyborem elementów głównych daje macierze $\tilde{L}$ i $\tilde{U}$ takie, że*

$$\tilde{L}\tilde{U} = A + E, \quad \textit{gdzie} \quad |e_{ij}| \leqslant 2n\varepsilon\rho.$$

Dowód. Używając wielkości δ_{ij} (nie należy ich mylić z deltą Kroneckera) i δ'_{ij} analogicznych do δ i δ' z (4.8.1), wyrażamy wielkości tworzone w komputerze jak następuje[21]:

$$\tilde{a}_{ij}^{(k+1)} = [\tilde{a}_{ij}^{(k)} - (1 - \delta_{ij})\tilde{l}_{ik}\tilde{a}_{kj}^{(k)}]/(1 - \delta'_{ij}) \qquad (i, j > k),$$
$$\tilde{l}_{ik} = (1 + \delta_{ik})\tilde{a}_{ik}^{(k)}/\tilde{a}_{kk}^{(k)} \qquad (i > k).$$

Pierwsze z tych równań można też napisać w postaci

$$\tilde{a}_{ij}^{(k+1)} = \tilde{a}_{ij}^{(k)} - \tilde{l}_{ik}\tilde{a}_{kj}^{(k)} + \delta_{ij}\tilde{l}_{ik}\tilde{a}_{kj}^{(k)} + \delta'_{ij}\tilde{a}_{ij}^{(k+1)} \qquad (i, j > k). \tag{4.8.3}$$

Macierz o elementach $\tilde{a}_{ij}^{(k)} - \tilde{l}_{ik}\tilde{a}_{kj}^{(k)}$ jest równa $\tilde{A}^{(k)} - \tilde{L}^{(k)}\tilde{A}^{(k)}$, gdzie $\tilde{L}^{(k)}$ ma niezerowe elementy tylko w k-tej kolumnie poniżej przekątnej:

$$\tilde{L}^{(k)} = \begin{bmatrix} 0 & & & & \\ & \ddots & & & \\ & & 0 & & \\ & & \tilde{l}_{k+1,k} & & \\ & & \vdots & \ddots & \\ & & \tilde{l}_{nk} & & 0 \end{bmatrix}.$$

[21] Łatwo zauważyć podwójną nieścisłość: te wielkości należałoby opatrzyć górnym wskaźnikiem k, a δ_{ik} w drugim wzorze nie ma nic wspólnego z analogiczną wielkością z pierwszego wzoru. Jest to jednak nieistotne, gdyż w dowodzie nie korzysta się z żadnych relacji wiążących wielkości δ i δ'. Ważne są tylko nierówności (4.8.2) (*przyp. tłum.*).

Wobec tego

$$\tilde{A}^{(k+1)} = \tilde{A}^{(k)} - \tilde{L}^{(k)}\tilde{A}^{(k)} + E^{(k)}, \tag{4.8.4}$$

gdzie $E^{(k)}$ jest macierzą o małych elementach zależnych od δ_{ij} i δ'_{ij} w sposób wynikający z (4.8.3). Zbadamy je teraz. Jeśli $i > k$, $j = k$, to

$$e_{ik}^{(k)} = \tilde{a}_{ik}^{(k+1)} - \tilde{a}_{ik}^{(k)} + \tilde{l}_{ik}^{(k)}\tilde{a}_{kk}^{(k)} = -\tilde{a}_{ik}^{(k)} + (\tilde{a}_{ik}^{(k)}/\tilde{a}_{kk}^{(k)})(1+\delta_{ik})\tilde{a}_{kk}^{(k)} = \delta_{ik}\tilde{a}_{ik}^{(k)}.$$

Natomiast dla $i, j > k$ mamy

$$e_{ij}^{(k)} = \tilde{a}_{ij}^{(k+1)} - \tilde{a}_{ij}^{(k)} + \tilde{l}_{ik}^{(k)}\tilde{a}_{kj}^{(k)} = \delta_{ij}\tilde{l}_{ik}\tilde{a}_{kj}^{(k)} + \delta'_{ij}\tilde{a}_{ij}^{(k+1)}.$$

Pozostałe elementy macierzy $E^{(k)}$ są zerami.

Sumując stronami równości (4.8.4) dla $k = 1, 2, \ldots, n-1$, otrzymujemy

$$\tilde{L}^{(1)}\tilde{A}^{(1)} + \ldots + \tilde{L}^{(n-1)}\tilde{A}^{(n-1)} + \tilde{A}^{(n)} = A^{(1)} + E^{(1)} + \ldots + E^{(n-1)}.$$

Wiersze macierzy $\tilde{L}^{(k)}\tilde{A}^{(k)}$ są wielokrotnościami wiersza $\tilde{a}_{k1}^{(k)}, \tilde{a}_{k2}^{(k)}, \ldots, \tilde{a}_{kn}^{(k)}$ identycznego z wierszem $\tilde{a}_{k1}^{(n)}, \tilde{a}_{k2}^{(n)}, \ldots, \tilde{a}_{kn}^{(n)}$. Dlatego $\tilde{L}^{(k)}\tilde{A}^{(k)} = \tilde{L}^{(k)}\tilde{A}^{(n)}$. Przypomnijmy jeszcze, że $A^{(1)} = A$ i że obliczonym czynnikiem trójkątnym górnym $\tilde{U}$ jest $\tilde{A}^{(n)}$. Przyjmując jeszcze, że $E := \sum_{k=1}^{n-1} E^{(k)}$, wyrażamy otrzymany wcześniej związek w postaci

$$(\tilde{L}^{(1)} + \ldots + \tilde{L}^{(n-1)} + I)\tilde{A}^{(n)} = A^{(1)} + E,$$

czyli $\tilde{L}\tilde{U} = A + E$.

Pozostaje oszacować elementy macierzy E. Ponieważ z założenia wierszami głównymi w procesie eliminacji są kolejne wiersze, więc wszystkie mnożniki spełniają warunek $|\tilde{l}_{ik}| \leqslant 1$. Dlatego z otrzymanych już wyrażeń dla $e_{ik}^{(k)}$ wynikają nierówności

$$|e_{ik}^{(k)}| = |\delta_{ik}|\,|\tilde{a}_{ik}^{(k)}| \leqslant \varepsilon\rho \qquad (i > k),$$
$$|e_{ij}^{(k)}| = |\delta_{ij}\tilde{l}_{ik}\tilde{a}_{kj}^{(k)} + \delta'_{ij}\tilde{a}_{ij}^{(k+1)}| \leqslant 2\varepsilon\rho \qquad (i, j > k).$$

Stąd i z definicji macierzy E wynika, że

$$|e_{ij}| = \Big|\sum_{k=1}^{n-1} e_{ij}^{(k)}\Big| \leqslant \sum_{k=1}^{n-1} |e_{ij}^{(k)}| \leqslant 2n\varepsilon\rho. \qquad \blacksquare$$

TWIERDZENIE 4.8.2. *Jeśli liczba naturalna n jest taka, że $n\varepsilon < \frac{1}{3}$ oraz jeśli x_i, y_i $(1 \leqslant i \leqslant n)$ są liczbami maszynowymi, to obliczona wartość iloczynu skalarnego $\sum_{i=1}^{n} x_i y_i$ jest równa $\sum_{i=1}^{n} x_i y_i(1+\delta_i)$, gdzie $|\delta_i| \leqslant \frac{6}{5}(n+1)\varepsilon$ dla każdego i.*

Dowód. Iloczyn skalarny obliczamy według wzorów

$$z_0 := 0, \qquad z_k := \mathrm{fl}[z_{k-1} + \mathrm{fl}(x_k y_k)] \quad (1 \leqslant k \leqslant n).$$

Udowodnimy przez indukcję względem n, że $|\delta_i| \leqslant (1+\varepsilon)^{n+2-i} - 1$. Jeśli $n = 1$, to

$$z_1 = \mathrm{fl}(x_1 y_1) = x_1 y_1 (1 + \delta_1), \quad |\delta_1| \leqslant \varepsilon.$$

Dla tego n twierdzenie jest prawdziwe, bo $\varepsilon \leqslant (1+\varepsilon)^2 - 1$. Przyjmując, że twierdzenie zachodzi dla $n = k-1$, sprawdzamy je dla $n = k$. W tym celu obliczamy

$$\begin{aligned} z_k &= [z_{k-1} + x_k y_k (1+\delta')](1+\delta) = \\ &= \Big[\sum_{i=1}^{k-1} x_i y_i (1+\delta_i) + x_k y_k (1+\delta')\Big](1+\delta) = \\ &= \sum_{i=1}^{k-1} x_i y_i (1 + \delta_i + \delta + \delta_i \delta) + x_k y_k (1 + \delta + \delta' + \delta\delta'). \end{aligned}$$

Wiedząc już, że

$$|\delta| \leqslant \varepsilon, \quad |\delta'| \leqslant \varepsilon, \quad |\delta_i| \leqslant (1+\varepsilon)^{k+1-i} - 1 \quad (1 \leqslant i \leqslant k-1),$$

mamy udowodnić, że

$$|\delta_i + \delta + \delta_i \delta| \leqslant (1+\varepsilon)^{k+2-i} - 1, \quad |\delta + \delta' + \delta\delta'| \leqslant (1+\varepsilon)^2 - 1.$$

Pierwszą z tych nierówności sprawdzamy tak:

$$\begin{aligned} |\delta_i + \delta + \delta_i \delta| &\leqslant |\delta_i| + |\delta|(1 + |\delta_i|) \leqslant \\ &\leqslant (1+\varepsilon)^{k+1-i} - 1 + \varepsilon(1+\varepsilon)^{k+1-i} = (1+\varepsilon)^{k+2-i} - 1. \end{aligned}$$

Dowód drugiej jest podobny. Trzeba jeszcze wykazać, że $|\delta_i|$ spełniają nierówność podaną w twierdzeniu. Dla $k \leqslant n+1$ jest

$$\begin{aligned} (1+\varepsilon)^k - 1 &= \left[1 + k\varepsilon + \tfrac{1}{2}k(k-1)\varepsilon^2 + \ldots + \varepsilon^k\right] - 1 = \\ &= k\varepsilon\left[1 + \tfrac{1}{2}(k-1)\varepsilon + \tfrac{1}{6}(k-1)k\varepsilon^2 + \ldots\right] \leqslant \\ &\leqslant k\varepsilon\left[1 + \tfrac{1}{2}k\varepsilon + \left(\tfrac{1}{2}k\varepsilon\right)^2 + \ldots\right] = k\varepsilon/(1 - \tfrac{1}{2}k\varepsilon) < \tfrac{6}{5}k\varepsilon. \end{aligned}$$

Ostatnia nierówność wynika z założenia, że $n\varepsilon < \frac{1}{3}$. ■

Twierdzenie 4.8.3. *Niech L będzie macierzą jedynkową trójkątną dolną stopnia n, której elementami są liczby maszynowe, a b – wektorem o n składowych będących takimi liczbami. Wtedy, jeśli $n\varepsilon < \frac{1}{3}$, to obliczone rozwiązanie $\tilde{y}$ układu $Ly = b$ jest dokładnym rozwiązaniem zaburzonego układu*

$$(L + \Delta)\tilde{y} = b, \quad \text{gdzie} \quad |(\Delta)_{ij}| \leqslant \tfrac{6}{5}(n+1)\varepsilon |l_{ij}|.$$

Dowód. Dokładne rozwiązanie układu $Ly = b$ jest obliczane ze wzoru

$$y_i := b_i - \sum_{j=1}^{i-1} l_{ij} y_j \qquad (1 \leqslant i \leqslant n).$$

Wobec tego obliczone składowe $\tilde{y}_i$ są takie, że

$$\tilde{y}_i = \Big[b_i - \sum_{j=1}^{i-1} l_{ij}\tilde{y}_j(1+\delta_{ij})\Big] \Big/ (1+\delta_{ii}), \tag{4.8.5}$$

gdzie na mocy tw. 2.1.3

$$|\delta_{ij}| \leqslant \tfrac{6}{5}(n+1)\varepsilon \qquad (1 \leqslant j \leqslant i \leqslant n).$$

Z (4.8.5) wynika, że

$$\sum_{j=1}^{i} l_{ij}\tilde{y}_j(1+\delta_{ij}) = b_i$$

(jest $l_{ii} = 1$). Mamy więc równanie macierzowe $(L + \Delta)\tilde{y} = b$, gdzie Δ jest macierzą trójkątną dolną o elementach $l_{ij}\delta_{ij}$, a to już daje tezę twierdzenia. ■

Twierdzenie 4.8.4. *Niech U będzie macierzą trójkątną górną stopnia n, której elementami są liczby maszynowe, a c – wektorem o n składowych będących takimi liczbami. Wtedy, jeśli $n\varepsilon < \frac{1}{3}$, to obliczone rozwiązanie $\tilde{y}$ układu $Uy = c$ jest dokładnym rozwiązaniem zaburzonego układu*

$$(U + \Delta)\tilde{y} = c, \quad \text{gdzie} \quad |(\Delta)_{ij}| \leqslant \tfrac{6}{5}(n+1)\varepsilon |u_{ij}|.$$

Dowód jest tematem zad. 3.

Twierdzenie 4.8.5. *Niech elementy macierzy A stopnia n i składowe wektora b będą liczbami maszynowymi. Jeśli $n\varepsilon < \frac{1}{3}$, to rozwiązanie $\tilde{x}$ układu $Ax = b$, obliczone metodą eliminacji Gaussa z wyborem wierszy głównych, jest dokładnym rozwiązaniem zaburzonego układu*

$$(A + F)\tilde{x} = b, \quad \text{gdzie} \quad |f_{ij}| \leqslant 10n^2\varepsilon\rho.$$

Dowód. Z twierdzeń 4.8.1, 4.8.3 i 4.8.4 wynika odpowiednio, że

$$\begin{aligned} A+E &= \tilde{L}\tilde{U}, & |e_{ij}| &\leqslant 2n\varepsilon\rho, \\ (\tilde{L}+\Delta)\tilde{y} &= b, & |(\Delta)_{ij}| &\leqslant \tfrac{6}{5}(n+1)\varepsilon|\tilde{l}_{ij}|, \\ (\tilde{U}+\Delta')\tilde{x} &= \tilde{y}, & |(\Delta')_{ij}| &\leqslant \tfrac{6}{5}(n+1)\varepsilon|\tilde{u}_{ij}| \end{aligned}$$

(oznaczenia tu nieco zmieniono, uwzględniając to, że układ rozwiązujemy w trzech etapach). Stąd wnioskujemy, że

$$\begin{aligned} b &= (\tilde{L}+\Delta)\tilde{y} = (\tilde{L}+\Delta)(\tilde{U}+\Delta')\tilde{x} = (\tilde{L}\tilde{U}+\Delta\tilde{U}+\tilde{L}\Delta'+\Delta\Delta')\tilde{x} = \\ &= (A+E+\Delta\tilde{U}+\tilde{L}\Delta'+\Delta\Delta')\tilde{x} = (A+F)\tilde{x}, \end{aligned}$$

gdzie $F := E + \Delta\tilde{U} + \tilde{L}\Delta' + \Delta\Delta'$. Aby oszacować elementy macierzy F, korzystamy z oszacowań podanych wyżej, a także z nierówności $|l_{ij}| \leqslant 1$, $|u_{ij}| = |a_{ij}^{(n)}| \leqslant \rho$:

$$\begin{aligned} |f_{ij}| &\leqslant |e_{ij}| + \sum_{\nu=1}^{n} \left[|(\Delta)_{ij}|\,|\tilde{u}_{\nu j}| + |\tilde{l}_{i\nu}|\,|(\Delta')_{\nu j}| + |(\Delta)_{i\nu}|\,|(\Delta')_{\nu j}| \right] \leqslant \\ &\leqslant 2n\varepsilon\rho + 2\cdot\tfrac{6}{5}n(n+1)\varepsilon\rho + \tfrac{36}{25}n(n+1)^2\varepsilon^2\rho = \\ &= n^2\varepsilon\rho\left[\frac{2}{n} + \frac{12}{5}\frac{n+1}{n} + \frac{36}{25}\varepsilon\left(\frac{n+1}{n}\right)^2\right]. \end{aligned}$$

Suma w nawiasach kwadratowych nie przekracza 10. ■

Za pomocą bardziej subtelnych rozumowań można otrzymać oszacowanie dla $\|F\|_\infty$ lepsze od tego, które wynika z tw. 4.8.5; zob. Forsythe i Moler [1967], Golub i van Loan [1989], Wilkinson [1965] oraz Isaacson i Keller [1966].

Wskaźnik wzrostu

Gdy rozwiązujemy numerycznie układ $Ax = b$ metodą Gaussa, to miarą jej stabilności jest *wskaźnik wzrostu*[22)]

$$g_n(A) := \frac{\max_{i,j,k}|a_{ij}^{(k)}|}{\max_{i,j}|a_{ij}|}.$$

Jak przedtem, $a_{ij}^{(k)}$ jest tu elementem macierzy wynikającej z A po k krokach eliminacji. Wielkość $g_n(A)$ pokazuje, jak duże mogą być te elementy w po-

[22)] W oryginale *growth factor*. Jak się wydaje, nie było wcześniej odpowiedników tego terminu w języku polskim (*przyp. tłum.*).

równaniu z elementami macierzy A. Eliminacja Gaussa jest stabilna, jeśli ta wielkość nie jest zbyt duża. Wynika to z oszacowań, w których występuje wskaźnik wzrostu. W szczególności Wilkinson [1965] udowodnił, że

$$\frac{\|\tilde{x} - x\|_\infty}{\|x\|_\infty} \leqslant 4n^2 g_n(A)\varkappa_\infty(A)\varepsilon,$$

gdzie x jest dokładnym rozwiązaniem układu $Ax = b$, a $\tilde{x}$ jego rozwiązaniem obliczonym w arytmetyce zmiennopozycyjnej metodą eliminacji Gaussa z częściowym wyborem elementów głównych, bez skalowania. Natomiast w nierówności

$$\frac{\|\tilde{A} - A\|_\infty}{\|A\|_\infty} < 8n^3 g_n(A)\varepsilon$$

(można ją porównać do tej z tw. 4.8.5) występuje oprócz A taka macierz $\tilde{A}$, że określony wyżej wektor $\tilde{x}$ jest dokładnym rozwiązaniem układu $\tilde{A}\tilde{x} = b$ (zob. np. Golub i van Loan [1989]).

Wilkinson [1965] udowodnił, że $g_n(A) \leqslant 2^{n-1}$ i znalazł przykłady macierzy, dla których to oszacowanie jest osiągnięte; podają je też Golub i van Loan [1989], Higham i Higham [1989] oraz Foster [1994]. Wilkinson wywnioskował jednak ze swych doświadczeń numerycznych, że duże wartości wskaźnika wzrostu są bardzo mało prawdopodobne nawet dla częściowego wyboru elementów głównych. Wilkinson zauważył, że trudno znaleźć *naturalną* macierz A, dla której $g_n(A) > n$; nie znalazł też podobnych przykładów, gdzie ten wskaźnik przekroczyłby 16. Jego poglądy sprowokowały innych badaczy do szukania macierzy z dużym wskaźnikiem wzrostu; m.in. Dongarra, Bunch, Moler i Stewart [1979] znaleźli macierz, dla której $g_n(A) = 23$. Później Trefethen i Schreiber [1990] wykazali, że dla macierzy o losowych elementach $g_n(A)$ nie rośnie wykładniczo. Natomiast Wright [1993] i Foster [1994] udowodnili, że rozwiązywanie numeryczne pewnych zagadanień brzegowych dla równań różniczkowych zwyczajnych i równań całkowych Volterry prowadzi do macierzy, których wskaźnik wzrostu rośnie wykładniczo, a to może powodować katastrofalny wzrost błędów. Mimo wszystko utrzymuje się opinia, że w zastosowaniach ten wskaźnik jest zwykle niezbyt duży i że eliminacja Gaussa z częściowym wyborem elementów głównych jest stabilna.

Dla eliminacji Gaussa z pełnym wyborem elementów głównych Wilkinson [1961] wykazał, że

$$g_n(A) \leqslant n^{1/2}[2 \cdot 3^{1/2} \cdot 4^{1/3} \ldots n^{1/(n-1)}]^{1/2}.$$

Prawa strona tej nierówności rośnie wraz z n, ale – jak zauważył Wilkinson – znacznie wolniej od 2^n. Cryer [1968] postawił hipotezę, że $g_n(A) \leqslant n$, ale

obalili ją Gould [1991] i Edelman [1992], znajdując przykłady odpowiednich macierzy. Decydując się na jedną z dwóch strategii wyboru elementów głównych, trzeba pamiętać, że pełny wybór znacznie wydłuża obliczenia.

W wielu przypadkach znalezienie macierzy o szczególnie dużym wskaźniku $g_n(A)$ wymagało użycia superkomputerów i wyspecjalizowanego oprogramowania. Niektórzy eksperci sądzą, że przeoczono możliwość dużego wzrostu błędów w eliminacji Gaussa. Będzie to na pewno tematem dalszych badań; zob. np. prace Edelmana [1992] i Fostera [1994].

ZADANIA 4.8

1. Wykazać, że w dowodzie tw. 4.8.1 jest

$$\|E^{(k)}\|_\infty \leqslant [1 + 2(n-k)]\varepsilon\rho, \quad \|E\|_\infty \leqslant n^2\varepsilon\rho.$$

2. Wykazać, że w tw. 4.8.3 jest $\|\Delta\|_\infty \leqslant \frac{3}{5}n(n+1)\varepsilon \max_{1\leqslant i,j\leqslant n}|l_{ij}|$.

3. Udowodnić tw. 4.8.4.

ROZDZIAŁ 5

Inne działy numerycznej algebry liniowej

5.0. Przegląd podstawowych pojęć

Tu, jak i we fragmencie podrozdz. 4.6 poświęconym metodzie nadrelaksacji, będziemy mieli do czynienia z macierzami o elementach zespolonych. Określono tam większość związanych z nimi pojęć. Tu wystarczy dodatkowo zdefiniować wielomian charakterystyczny macierzy.

Niech A będzie macierzą kwadratową stopnia n, o elementach zespolonych, a λ liczbą zespoloną. Jeśli równanie

$$Ax = \lambda x$$

jest spełnione przez pewien wektor x o n składowych zespolonych nie wszystkich równych 0, to – jak wiadomo – λ jest *wartością własną*, a x *wektorem własnym* macierzy A, odpowiadającym tej wartości własnej.

Istnienie nietrywialnego rozwiązania równania $Ax = \lambda x$ jest równoważne każdemu z trzech następujących warunków:

macierz $A - \lambda I$ przekształca pewien wektor niezerowy na 0,

macierz $A - \lambda I$ jest osobliwa,

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

Tak więc w zasadzie obliczanie wartości własnych można sprowadzić do rozwiązywania powyższego równania, które można napisać w bardziej konkretnej postaci:

$$\begin{vmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}-\lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn}-\lambda \end{vmatrix} = 0.$$

Jest to *równanie charakterystyczne* macierzy A. Jego lewa strona jest wielomianem stopnia n względem λ, zwanym *wielomianem charakterystycznym* tej macierzy. Stąd wniosek, że dowolna macierz stopnia n ma dokładnie n wartości własnych, gdy każdą z nich liczymy tyle razy, ile wynosi krotność pierwiastka równania charakterystycznego.

Przykład 5.0.1. Znaleźć wartości własne macierzy

$$A := \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 2 & 1 & 1 \end{bmatrix}.$$

Rozwiązanie. Równanie charakterystyczne macierzy A ma postać

$$\begin{vmatrix} 1-\lambda & 2 & 1 \\ 0 & 1-\lambda & 3 \\ 2 & 1 & 1-\lambda \end{vmatrix} = -\lambda^3+3\lambda^2+2\lambda+8 = -(\lambda-4)(\lambda^2+\lambda+2) = 0,$$

a jego pierwiastkami, czyli wartościami własnymi macierzy A, są liczby

$$\lambda_1 = 4, \quad \lambda_2 = -\tfrac{1}{2}+\tfrac{1}{2}\sqrt{7}i, \quad \lambda_3 = -\tfrac{1}{2}-\tfrac{1}{2}\sqrt{7}i.$$

Jak widać, wartości własne macierzy rzeczywistej nie muszą być rzeczywiste. ■

Wyżej zastosowano bezpośredni sposób obliczania wartości własnych. Jest on sensowny w obliczeniach ręcznych dla macierzy niskiego stopnia. W innych przypadkach stanowczo go odradzamy. Jednym z powodów jest to, że pierwiastki wielomianu mogą być bardzo czułe na zmiany jego współczynników, spowodowane np. błędami zaokrągleń. Ilustrujący to klasyczny przykład Wilkinsona podano w podrozdz. 2.3.

Kończąc te wstępne uwagi, warto jeszcze przypomnieć określone w podrozdz. 4.6 pojęcie podobieństwa macierzy: macierze A i B są z definicji

podobne, jeśli istnieje macierz nieosobliwa P taka, że $B = PAP^{-1}$ (czyli $A = QBQ^{-1}$, gdzie $Q := P^{-1}$). W cytowanym fragmencie zastosowano następującą fundamentalną własność tego pojęcia, która będzie często wykorzystywana i dalej:

Twierdzenie 5.0.2. *Macierze podobne mają te same zbiory wartości własnych.*

Dowód. Niech będzie $B = PAP^{-1}$. Wystarczy wykazać, że A i B mają wspólny wielomian charakterystyczny. Tak jest, gdyż

$$\det(B - \lambda I) = \det(PAP^{-1} - \lambda I) = \det[P(A - \lambda I)P^{-1}] =$$
$$= \det P \det(A - \lambda I) \det P^{-1} = \det(A - \lambda I).$$

Zastosowano tu dwa twierdzenia z teorii wyznaczników: wyznacznik iloczynu dwóch macierzy jest iloczynem ich wyznaczników, a wyznacznik odwrotności macierzy nieosobliwej jest odwrotnością jej wyznacznika. ■

W związku z tw. 5.0.2 warto zauważyć, że jeśli $P = PAP^{-1}$, a x jest wektorem własnym macierzy A odpowiadającym jej wartości własnej λ, to Px jest wektorem własnym macierzy B odpowiadającym tejże wartości λ.

5.1. Metoda potęgowa dla zadania własnego

Metoda potęgowa

Pierwszą metodą numeryczną, jaką opiszemy, jest *metoda potęgowa*. Pozwala ona obliczyć wartość własną o największym module i odpowiadający jej wektor własny, czyli rozwiązać pewne *zadanie własne*. Metoda działa bez zakłóceń, jeśli macierz ma następujące własności:

1. Tylko jedna jej wartość własna (rzeczywista lub zespolona) ma największy moduł.
2. Istnieje układ n wektorów własnych liniowo niezależnych czyli macierz ma *prostą strukturę*[1)].

   Dzięki własności **1** można tak ponumerować wartości własne λ_j, że

   $$|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|.$$

[1)] Macierz nie mająca własności **2** ma *złożoną strukturę*. W języku angielskim są używane odpowiednio terminy *nondefective* i *defective matrix*. Więcej informacji o tych dwóch klasach macierzy podaje np. Wilkinson [1965, rozdz. 1]; zob. też Kiełbasiński i Schwetlick [*1992] (*przyp. tłum.*).

Własność **2** gwarantuje, że wektory własne $u^{(j)}$ takie, iż

$$Au^{(j)} = \lambda_j u^{(j)} \quad (1 \leqslant j \leqslant n) \tag{5.1.1}$$

tworzą bazę przestrzeni $\mathbb{C}^n$. Niech $x^{(0)}$ będzie dowolną kombinacją liniową tych wektorów własnych, z niezerowym współczynnikiem przy $u^{(1)}$. Stąd, po ewentualnym pomnożeniu $u^{(1)}$ przez niezerową stałą, wynika, że

$$x^{(0)} = u^{(1)} + \sum_{j=2}^{n} a_j u^{(j)}.$$

Utwórzmy wektory

$$x^{(1)} := Ax^{(0)}, \quad x^{(2)} := Ax^{(1)}, \ldots$$

Ogólniej, dla $k \geqslant 0$ jest

$$x^{(k)} := A^k x^{(0)} = A^k u^{(1)} + \sum_{j=2}^{n} a_j A_k u^{(j)}.$$

Dzięki (5.1.1) wnioskujemy stąd, że

$$x^{(k)} = \lambda_1^k u^{(1)} + \sum_{j=2}^{n} a_j \lambda_j^k u^{(j)} = \lambda_1^k \left[u^{(1)} + \sum_{j=2}^{n} a_j \left(\frac{\lambda_j}{\lambda_1} \right)^k u^{(j)} \right].$$

Ponieważ $|\lambda_1| > |\lambda_j|$ dla $2 \leqslant j \leqslant n$, więc wyrażenia $(\lambda_j/\lambda_1)^k$ dążą do 0, gdy $x \to \infty$ i wektor w nawiasach kwadratowych dąży wtedy do $u^{(1)}$.

Aby uprościć oznaczenia, piszemy

$$x^{(k)} = \lambda_1^k (u^{(1)} + \varepsilon^{(k)}),$$

gdzie $\varepsilon^{(k)} \to 0$ dla $k \to \infty$. Niech φ będzie dowolnym funkcjonałem liniowym określonym na $\mathbb{C}^n$ i takim, że $\varphi(u^{(1)}) \neq 0$. Przypomnijmy, że jest to funkcjonał taki, że $\varphi(\alpha x + \beta y) = \alpha\varphi(x) + \beta\varphi(y)$ dla dowolnych liczb α, β i wektorów x, y. W szczególności wartością φ może być j-ta składowa wektora. Tak więc

$$\varphi(x^{(k)}) = \lambda_1^k [\varphi(u^{(1)}) + \varphi(\varepsilon^{(k)})].$$

Stąd i z relacji $\varphi(\varepsilon^{(k)}) \to 0$ wynika, że

$$r_k := \frac{\varphi(x^{(k+1)})}{\varphi(x^{(k)})} = \lambda_1 \frac{\varphi(u^{(1)}) + \varphi(\varepsilon^{(k+1)})}{\varphi(u^{(1)}) + \varphi(\varepsilon^{(k)})} \to \lambda_1.$$

Określiliśmy w ten sposób *metodę potęgową*; poleca ona obliczać wielkości

r_k, które są przybliżeniami wartości własnej λ_1 o największym module[2]. Ponieważ kierunek wektora $x^{(k)}$ zbliża się coraz bardziej do $u^{(1)}$, więc ta metoda daje zarazem przybliżenia wektora własnego $x^{(1)}$. W publikacjach można znaleźć wiele wariantów i udoskonaleń metody potęgowej.

Algorytm

Metodę potęgową można opisać najkrócej tak:

input n, A, x, M
output $0, x$
for $k = 1$ **to** M **do**
    $y \leftarrow Ax$
    $r \leftarrow \varphi(y)/\varphi(x)$
    $x \leftarrow y$
    output k, x, r
end do

W praktyce jest tu jednak konieczna pewna modyfikacja. Aby uniknąć zbieżności ciągu wektorów $x^{(k)}$ do 0 lub nieograniczonego wzrostu ich składowych, normalizujemy każdy wektor, czyli zmieniamy polecenie $x \leftarrow y$ na $x \leftarrow y/\|y\|$, co nie wpływa na obliczane ilorazy r. Można tu użyć dowolnej normy wektorowej, np. normy $\|x\|_\infty := \max_{1 \leqslant j \leqslant n} |x_j|$.

Przykład 5.1.1. Zastosować metodę potęgową dla macierzy

$$A := \begin{bmatrix} 6 & 5 & -5 \\ 2 & 6 & -2 \\ 2 & 5 & -1 \end{bmatrix},$$

wektora początkowego $x := (-1, 1, 1)$, funkcjonału φ określonego wzorem $\varphi(x) := x_2$ i normy $\|\cdot\|_\infty$.

Rozwiązanie. Niżej podano wektory unormowane $x^{(k)}$ i ilorazy r_k dla kilku wartości k:

[2] Wyżej założono, że wektor początkowy $x^{(0)}$ nie jest dowolny, bo w jego wyrażeniu przez wektory własne współczynnik przy $u^{(1)}$ nie znika. W praktyce numerycznej trudno sprawdzić, czy to założenie jest spełnione. Jeśli ten współczynnik jest równy 0, to i tak błędy zaokrągleń popełniane w obliczeniach powodują, że dla dużych k w analogicznym wyrażeniu wektora $x^{(k)}$ współczynnik przy $u^{(1)}$ już nie znika, co pozwala rozumować jak wyżej. Nietrafny wybór wektora $x^{(0)}$ może jednak znacznie wydłużyć obliczenia lub spowodować, że jedną z wartości $\lambda_2, \lambda_3, \ldots$ uznamy mylnie za wartość własną o największym module; por. zad. K2. Dodajmy, że w literaturze metodę potęgową wiąże się z konkretnym funkcjonałem φ, takim, że $r_k = (x_k^T x_{k+1})/(x_k^T x_k)$. Jest to tzw. *iloraz Rayleigha*; zob. Dahlquist i Björck [1974] oraz Kiełbasiński i Schwetlick [*1992] (*przyp. tłum.*).

$$
\begin{aligned}
x^{(0)} &= (-1.00000,\ \ 1.00000,\ \ 1.00000), \\
x^{(1)} &= (-1.00000,\ \ 0.33333,\ \ 0.33333), \qquad r_0 = \ 2.0, \\
x^{(2)} &= (-1.00000, -0.11111, -0.11111), \qquad r_1 = -2.0, \\
x^{(3)} &= (-1.00000, -0.40741, -0.40741), \qquad r_2 = 22.0, \\
x^{(4)} &= (-1.00000, -0.60494, -0.60494), \qquad r_3 = \ 8.9091, \\
x^{(6)} &= (-1.00000, -0.82442, -0.82442), \qquad r_5 = \ 6.71508, \\
x^{(28)} &= (-1.00000, -0.99998, -0.99998), \qquad r_{27} = \ 6.00007.
\end{aligned}
$$

Istotnie, największą co do modułu wartością własną macierzy A jest liczba 6, a odpowiednim wektorem własnym jest $(1, 1, 1)$ (i jego dowolny iloczyn przez niezerową stałą). ■

Metoda Aitkena

Ilorazy r_k są przybliżeniami wartości własnej λ_1. Może być interesujące oszacowanie błędów $|r_k - \lambda_1|$. Z zadań 15 i 16 wynika, że

$$r_{k+1} - \lambda_1 = (c + \delta_k)(r_k - \lambda_1),$$

gdzie $|c| < 1$ i $\{\delta_k\} \to 0$. Ciąg $\{r_k\}$ jest więc zbieżny do λ_1 liniowo (por. podrozdz. 1.2). To pozwala stosować *metodę Aitkena* przyspieszania zbieżności tego ciągu, polegającą na przekształceniu go na ciąg o elementach

$$s_n := \frac{r_n r_{n+2} - r_{n+1}^2}{r_{n+2} - 2r_{n+1} + r_n} \quad (n \geqslant 0). \tag{5.1.2}$$

Jest on szybciej zbieżny, co wynika z poniższego ogólnego twierdzenia.

Twierdzenie 5.1.2. *Niech $\{r_n\}$ będzie ciągiem liczbowym zbieżnym do granicy r. Jeśli $r_{n+1} - r = (c + \delta_n)(r_n - r)$, gdzie $|c| < 1$ i $\lim_{n\to\infty} \delta_n = 0$, to ciąg $\{s_k\}$ określony wzorem (5.1.2) jest zbieżny do r szybciej w tym sensie, że $(s_n - r)/(r_n - r) \to 0$ dla $n \to \infty$.*

Dowód. Niech będzie $h_n = r_n - r$. Łatwo sprawdzić, że

$$s_n = \frac{(r + h_n)(r + h_{n+2}) - (r + h_{n+1})^2}{(r + h_{n+2}) - 2(r + h_{n+1}) + (r + h_n)} = r + \frac{h_n h_{n+2} - h_{n+1}^2}{h_{n+2} - 2h_{n+1} + h_n}.$$

Ponieważ z założenia $h_{n+1} = (c + \delta_n)h_n$, więc $h_{n+2} = (c + \delta_{n+1})(c + \delta_n)h_n$ i

$$
\begin{aligned}
s_n - r &= \frac{h_n(c + \delta_{n+1})(c + \delta_n)h_n - (c + \delta_n)^2 h_n^2}{(c + \delta_{n+1})(c + \delta_n)h_n - 2(c + \delta_n)h_n + h_n} = \\
&= h_n \frac{(c + \delta_n)(\delta_{n+1} - \delta_n)}{(c + \delta_{n+1})(c + \delta_n) - 2(c + \delta_n) + 1}.
\end{aligned}
$$

To już pokazuje, że $\lim_{n\to\infty}(s_n - r)/h_n = 0$, gdyż w ostatnim ułamku licznik dąży do 0, a mianownik – do granicy $(c-1)^2$ różnej od 0[3]. ■

Jest ważne, aby obliczanie wielkości s_n przerwać, gdy tylko one się z grubsza ustabilizują, bo wtedy w dalszych obliczeniach odejmowanie bliskich wielkości daje bezwartościowe wyniki.

Odwrotna metoda potęgowa

Aby uzasadnić następny wariant metody potęgowej, stosujemy następującą elementarną własność wartości własnych:

Twierdzenie 5.1.3. *Jeśli λ jest wartością własną macierzy nieosobliwej A, to λ^{-1} jest wartością własną macierzy A^{-1}.*

Dowód. Jeśli $Ax = \lambda x$ i $x \neq 0$, to $x = A^{-1}(\lambda x) = \lambda A^{-1}x$. Wobec tego $A^{-1}x = \lambda^{-1}x$ i λ^{-1} jest wartością własną[4] macierzy A^{-1}. ■

Twierdzenie 5.1.3 sugeruje sposób obliczenia najmniejszej wartości własnej macierzy A. Załóżmy mianowicie, że

$$|\lambda_1| \geqslant |\lambda_2| \geqslant \ldots \geqslant |\lambda_{n-1}| > |\lambda_n| > 0.$$

Ponieważ 0 nie jest wartością własną, więc macierz A jest nieosobliwa.

Wartości własne macierzy A^{-1} równe λ_j^{-1} spełniają nierówność

$$|\lambda_n^{-1}| > |\lambda_{n-1}^{-1}| \geqslant \ldots \geqslant |\lambda_1^{-1}| > 0.$$

Dzięki temu możemy obliczyć λ_n^{-1}, stosując metodę potęgową do A^{-1}. Ponieważ jednak obliczanie tej odwrotności jest kosztowne, więc nie stosujemy wzoru $x^{(k+1)} = A^{-1}x^{(k)}$, ale obliczamy nowy wektor $x^{(k+1)}$, rozwiązując układ $Ax^{(k+1)} = x^{(k)}$. Można to skutecznie zrobić, stosując eliminację Gaussa. Ściślej, należy najpierw rozłożyć macierz A na czynniki trójkątne, a potem rozwiązywać układy z takimi macierzami (zob. podrozdz. 4.3). Tak właśnie działa *odwrotna metoda potęgowa.*

---

[3] Warto zauważyć, że przy założeniach twierdzenia ciąg $\{r_n\}$ jest zbieżny do r liniowo (podrozdz. 1.2) i że nawet dla $|c| > 1$ (gdy ten ciąg jest rozbieżny) ciąg $\{s_n\}$ jest zbieżny do r. Trzeba też podkreślić, że wzór z tw. 5.1.2 ma złe własności numeryczne; zamiast niego zaleca się stosowanie wzoru z zad. 14 (*przyp. tłum.*).

[4] Jest oczywiste, że tej ostatniej wartości odpowiada wektor własny x (*przyp. tłum.*).

Przykład 5.1.4. Zastosować odwrotną metodę potęgową do macierzy A z przykł. 5.1.1, przyjmując, że $\varphi(x) := x_1$.

Rozwiązanie. Rozkład LU macierzy A jest następujący:

$$\begin{bmatrix} 6 & 5 & -5 \\ 2 & 6 & -2 \\ 2 & 5 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 \\ \frac{1}{3} & \frac{10}{13} & 1 \end{bmatrix} \begin{bmatrix} 6 & 5 & -5 \\ 0 & \frac{13}{3} & -\frac{1}{3} \\ 0 & 0 & \frac{12}{13} \end{bmatrix}.$$

Zaczynając od wektora $x := (3, 7, -13)$, wykonujemy 25 kroków. W każdym z nich obliczamy $x^{(k+1)}$, rozwiązując układ $Ux^{(k+1)} = L^{-1}x^{(k)}$. Następnie obliczamy iloraz $r_k := x_1^{(k+1)}/x_1^{(k)}$ i normalizujemy wektor $x^{(k+1)}$, dzieląc go przez jego normę $\|\cdot\|_\infty$. Oto niektóre wyniki:

$$\begin{array}{lrrrlrl} x^{(0)} & = & 3.00000, & 7.00000, & -13.00000 & & \\ x^{(1)} & = & -0.80165, & -0.00826, & -1.00000 & r_0 & = -5.8889, \\ x^{(2)} & = & -0.95089, & -0.01774, & -1.00000 & r_1 & = 1.19759, \\ x^{(3)} & = & -0.98759, & -0.00712, & -1.00000 & r_2 & = 1.02750, \\ x^{(4)} & = & -0.99688, & -0.00223, & -1.00000 & r_3 & = 1.00446, \\ x^{(6)} & = & -0.99980, & -0.00017, & -1.00000 & r_5 & = 1.00012, \\ x^{(11)} & = & -1.00000, & 0.00000, & -1.00000 & r_{10} & = 1.00000. \end{array}$$

Dalsze iteracje nie zmieniają już tych wyników. Macierz A ma wartość własną 1 i związany z nią wektor własny $(1, 0, 1)$[5]. ■

Oprócz już omówionych wariantów metody potęgowej znane są jeszcze dwa inne, związane z *macierzą przesuniętą* $A - \mu I$. Operując taką macierzą, można znaleźć wartość własną macierzy A, najbliższą danej liczby μ. Przypuśćmy, że jedna z tych wartości, np. λ_k, spełnia nierówność $0 < |\lambda_k - \mu| < \varepsilon$, a wszystkie inne są takie, że $|\lambda_j - \mu| > \varepsilon$ dla $j \neq k$. Ponieważ macierz $A - \mu I$ ma wartości własne $\lambda_j - \mu$, więc stosując do niej odwrotną metodę potęgową, możemy znaleźć przybliżenie wielkości $(\lambda_k - \mu)^{-1}$, a więc i przybliżenie wartości własnej λ_k. Ciąg wektorów powstaje tu przez rozwiązywanie układów $(A - \mu I)x^{(k+1)} = x^{(k)}$, co wymaga wcześniejszego znalezienia rozkładu LU macierzy $A - \mu I$[6].

[5] To, że metoda potęgowa w tym przykładzie jest szybciej zbieżna niż w poprzednim, nie jest przypadkowe. Trzecią wartością własną macierzy A jest liczba 4; dlatego tu błędy przybliżeń wartości własnej są z grubsza proporcjonalne do $(1/4)^k$, a w przykładzie 5.1.1 do $(4/6)^k$ (*przyp. tłum.*).

[6] Ten wariant metody potęgowej (lub jego pewna istotna modyfikacja) jest też nazywany *metodą Wielandta*; zob. np. Dryja i Jankowscy [*1982] lub Stoer i Bulirsch [1980] (*przyp. tłum.*).

W podobny sposób, ale stosując zwykłą metodę potęgową do macierzy $A - \mu I$, znajdujemy wartość własną, np. λ_k, macierzy A, najdalszą od danego μ. Jeśli wszystkie wartości własne są rzeczywiste, to taka procedura daje – zależnie od μ – albo najmniejszą, albo największą wartość własną.

ZADANIA 5.1

1. Niech elementy k-tego wiersza macierzy A będą takie, że $a_{kj} = 0$ dla $k \neq j$. Wykazać, że a_{kk} jest wartością własną macierzy A i że jej pozostałe wartości własne są zarazem wartościami własnymi macierzy wynikającej z A przez wykreślenie k-tego wiersza i k-tej kolumny.
2. Udowodnić, że jeśli co najmniej jedna z macierzy A, B jest nieosobliwa, to macierze $I - AB$ i $I - BA$ mają te same układy wartości własnych.
3. Udowodnić, że jeśli macierz A jest nieosobliwa i wyrażona w postaci PQ, to A i QP mają te same układy wartości własnych.
4. Jak można znaleźć te wartości λ, dla których znika wyznacznik macierzy stopnia n, o elementach $a_{jk} = \alpha_{jk}\lambda + \beta_{jk}$? Ile jest na ogół tych wartości?
5. Wykazać, że macierz $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ ma złożoną strukturę.
6. Udowodnić, że macierz o różnych parami wartościach własnych ma prostą strukturę.
7. Pokazać, że macierz o prostej strukturze może mieć wielokrotne wartości własne.
8. Niech $\lambda_1, \lambda_2, \ldots, \lambda_n$ będą wartościami własnymi macierzy A o prostej strukturze i niech P będzie macierzą, której kolumnami są odpowiednie wektory własne $u^{(1)}, u^{(2)}, \ldots, u^{(n)}$. Jak się wyraża iloczyn $P^{-1}AP$?
9. Wielomian o współczynnikach rzeczywistych można rozłożyć na czynniki kwadratowe i liniowe o współczynnikach tegoż typu. Czynniki kwadratowe mogą mieć pierwiastki zespolone. Obliczyć prawdopodobieństwo tego, że wielomian $x^2 + ax + b$ ma pierwiastki zespolone, zakładając, że a i b są wartościami zmiennych losowych o rozkładzie jednostajnym w tym samym przedziale $[-\rho, \rho]$. Udowodnić, że dąży ono do 0 dla $\rho \to \infty$ i do $\frac{1}{2}$ dla $\rho \to 0$. Co stąd wynika dla wartości własnych macierzy rzeczywistych?
10. Znaleźć wielomian charakterystyczny macierzy

$$\begin{bmatrix} a_1 & a_2 & \ldots & a_{n-1} & a_n \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}.$$

11. Udowodnić, że dla dowolnej liczby $\lambda \in \mathbb{C}$ i dowolnej macierzy A stopnia n wymiar zbioru wektorów x takich, że $Ax = \lambda x$, jest równy $n - \operatorname{rank}(A - \lambda I)$ ($\operatorname{rank}(B)$ jest rzędem macierzy B).

12. Udowodnić, że w metodzie Aitkena $(s_n - r)/(r_{n+2} - r) \to 0$ $(n \to \infty)$, jeśli tylko $c \neq 0$.

13. Niech ciąg $\{r_n\}$ spełnia tylko założenie $|r_{n+1} - r| \leqslant c|r_n - r|$, gdzie $0 < c < 1$. Czy można udowodnić, że wtedy metoda Aitkena przyspiesza zbieżność tego ciągu?

14. Wykazać, że wzór opisujący metodę Aitkena można wyrazić w postaci

$$s_k := r_k - \frac{(\Delta r_k)^2}{\Delta^2 r_k},$$

gdzie $\Delta r_k := r_{k+1} - r_k$, $\Delta^2 r_k := \Delta r_{k+1} - \Delta r_k$. Ta postać jest zalecana, gdyż nie powoduje takich błędów zaokrągleń, jak wzór z tw. 5.1.2.

15. Udowodnić, że w metodzie potęgowej jest

$$\frac{r_k - \lambda_1}{\lambda_1} = \left(\frac{\lambda_2}{\lambda_1}\right)^k c_k,$$

gdzie ciąg $\{c_k\}$ jest ograniczony.

16. (cd.). Wykazać, że jeśli $|\lambda_2| > |\lambda_3|$, to $r_{k+1} - \lambda_1 = (c + \delta_k)(r_k - \lambda_1)$, gdzie $|c| < 1$ i $\{\delta_k\} \to 0$, co pozwala stosować metodę Aitkena.

17. Zaprojektować modyfikację metody potęgowej odpowiednią dla przypadku, gdy $\lambda_1 = -\lambda_2 > |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|$.

18. Niech wartości własne macierzy A będą rzeczywiste i parami różne. Jaki powinien być parametr β, żeby metoda potęgowa zastosowana do macierzy $A + \beta I$ była najszybciej zbieżna do $\lambda_1 + \beta$, gdzie λ_1 jest największą z tych wartości?

19. Jakie są skutki zastosowania metody potęgowej dla macierzy rzeczywistej i wektora początkowego rzeczywistego, gdy największy moduł mają wartości własne zespolone sprzężone?

20. Udowodnić, że jeśli w r_k definiującym metodę potęgową $\varphi(\cdot)$ jest normą wektora (to nie jest funkcjonał liniowy), to ciąg $\{r_k\}$ jest zbieżny do $|\lambda_1|$.

21. Wyznaczyć przybliżoną wartość promienia spektralnego $\rho(A)$ (określonego w podrozdz. 4.4) macierzy

$$A = \begin{bmatrix} 2 & 0 & -1 \\ -2 & -10 & 0 \\ -1 & -1 & 4 \end{bmatrix},$$

wykonując dwie iteracje metodą potęgową dla $\varphi(x) := \|x\|_\infty$. Wektorem początkowym ma być $(1, 1, 1)$.

22. *Ślad* macierzy A stopnia n, oznaczany symbolem $\operatorname{tr}(A)$, jest z definicji równy $\sum_{i=1}^n a_{ii}$. Udowodnić, że jest on równy sumie wartości własnych λ_i tej macierzy.

23. (cd.). Udowodnić, że $\operatorname{tr}(A^m) = \sum_{i=1}^n \lambda_i^m$ (m naturalne).

24. (cd.). Udowodnić, że jeśli $|\lambda_1| > |\lambda_i|$ dla $i > 1$, to

$$\lambda_1 = \lim_{m \to \infty} \operatorname{tr}(A^{m+1})/\operatorname{tr}(A^m).$$

ZADANIA KOMPUTEROWE 5.1

K1. (a) Napisać program pozwalający odtworzyć wyniki przykł. 5.1.1. Program powinien mieć strukturę modułową (por. zad. 4.1.K1). Można więc np. wyodrębnić procedury: (i) mnożenia macierzy przez wektor, (ii) obliczania iloczynu skalarnego, (iii) zastąpienia wektora przez inny wektor, (iv) obliczania normy wektora, (v) normalizacji wektora itd.

(b) Włączyć do programu metodę Aitkena i porównać wyniki obu wersji.

K2. Zastosować 100 kroków metody potęgowej do macierzy z przykł. 5.1.1 dla wektora początkowego $(1, 2, 3)$. Wytłumaczyć, dlaczego na początku obliczeń ciąg $\{r_k\}$ jest – jak się wydaje – zbieżny do pewnej granicy (czym ona jest dla macierzy?), a później zaczyna zbliżać się do innej wartości.

K3. Napisać program dla odwrotnej metody potęgowej (w dwóch wersjach, jak w zad. K1) oraz sprawdzić go dla macierzy z przykł. 5.1.1 i innych, dowolnie wybranych macierzy.

K4. Napisać program wykonujący M kroków metody potęgowej z normalizacją wektorów $x^{(k)}$ dla danej macierzy A stopnia n i danego wektora początkowego x. Włączyć metodę Aitkena. Program w każdym kroku ma wyświetlać wektor $x^{(k)}$, iloraz r_k i ulepszony iloraz s_{k-2}. Sprawdzić program dla macierzy $A - \mu I$, gdzie:

(a) $A = \begin{bmatrix} 5 & 4 & 1 & 1 \\ 4 & 5 & 1 & 1 \\ 1 & 1 & 4 & 2 \\ 1 & 1 & 2 & 4 \end{bmatrix}$, $\mu = 0, 3, 6, 11$

(b) $A = \begin{bmatrix} 2 & 3 & 4 \\ 7 & -1 & 3 \\ 1 & -1 & 5 \end{bmatrix}$, $\mu = 0, 5, 10$

K5. Napisać i sprawdzić dla macierzy z przykł. 5.1.1 program obliczający wartość własną najdalszą od danej liczby zespolonej.

K6. Zbudować przykład pokazujący, że metoda Aitkena daje bezsensowne wyniki, jeśli nie przerwie się jej we właściwym momencie.

5.2. Twierdzenia Schura i Gerszgorina

Faktoryzacja Schura

Twierdzenie 5.0.2 sugeruje pewną strategię obliczania wartości własnych macierzy A, mianowicie – przekształcenie jej przez podobieństwo na macierz $B := PAP^{-1}$, dla której te obliczenia byłyby prostsze. Przypuśćmy w szczególności, że B jest trójkątna. Jej wartości własne (a zarazem wartości własne macierzy A) są po prostu elementami przekątniowymi. Poniższe twierdzenie Schura (które będzie poprzedzone dwoma lematami) zapewnia, że przekształcenie na taką macierz B jest w teorii zawsze możliwe.

Przedtem trzeba podać kilka ważnych informacji. Po pierwsze, symbol A^{H} oznacza macierz *sprzężoną*[7] z A, tj. taką, że $(A^{\mathsf{H}})_{ij} = \overline{(A)_{ji}}$. Po drugie, macierz kwadratową U nazywamy *unitarną*, jeśli $UU^{\mathsf{H}} = I$ (albo, co jest równoważnym warunkiem, $U^{\mathsf{H}}U = I$). Po trzecie, macierze A i B są *unitarnie podobne*, jeśli $B = UAU^{\mathsf{H}}$ dla pewnej macierzy unitarnej U[8].

Lemat 5.2.1. *Macierz $I - vv^{\mathsf{H}}$ (v – wektor) jest unitarna wtedy i tylko wtedy, gdy $\|v\|_2^2 = 2$ lub $v = 0$.*

Dowód. Sprawdzamy, kiedy $UU^{\mathsf{H}} = I$, gdzie $U := I - vv^{\mathsf{H}}$. Jest

$$\begin{aligned} UU^{\mathsf{H}} &= (I - vv^{\mathsf{H}})(I - vv^{\mathsf{H}}) = I - 2vv^{\mathsf{H}} + vv^{\mathsf{H}}vv^{\mathsf{H}} = \\ &= I - 2vv^{\mathsf{H}} + (v^{\mathsf{H}}v)(vv^{\mathsf{H}}) = I - \left(2 - \|v\|_2^2\right) vv^{\mathsf{H}} \end{aligned}$$

(w iloczynie czterech czynników skalar $v^{\mathsf{H}}v$ można przesunąć na początek). Ostatnie wyrażenie jest równe I wtedy i tylko wtedy, gdy $\|v\|_2^2 = 2$ lub $vv^{\mathsf{H}} = 0$. ∎

Lemat 5.2.2. *Jeśli wektory x i y są takie, że $\|x\|_2 = \|y\|_2$ i xy^{H} jest rzeczywiste, to istnieje macierz unitarna $U = I - vv^{\mathsf{H}}$, dla której $Ux = y$.*

Dowód. Jeśli $x = y$, to dla $v = 0$ jest $U = I$ i $Ux = y$. Jeśli $x \neq y$, to przyjmujemy $v = \alpha(x - y)$, gdzie $\alpha := \sqrt{2}/\|x - y\|_2$; wtedy $\|v\|_2^2 = 2$. W obu przypadkach na mocy lematu 5.2.1 macierz U jest unitarna. Dla $x \neq y$ sprawdzamy jeszcze równość $Ux = y$. Jest wtedy

$$\begin{aligned} Ux - y &= (I - vv^{\mathsf{H}})x - y = x - vv^{\mathsf{H}}x - y = \\ &= x - y - \alpha^2(x - y)(x^{\mathsf{H}} - y^{\mathsf{H}})x = (x - y)[1 - \alpha^2(x^{\mathsf{H}}x - y^{\mathsf{H}}x)]. \end{aligned}$$

Wyrażenie w nawiasach kwadratowych znika. Istotnie, pierwsze założenie jest równoważne temu, że $x^{\mathsf{H}}x = y^{\mathsf{H}}y$. Ponieważ $y^{\mathsf{H}}x = \overline{x^{\mathsf{H}}y}$, więc drugie założenie gwarantuje, że $y^{\mathsf{H}}x = x^{\mathsf{H}}y$. Uwzględniając to, obliczamy

$$\begin{aligned} 1 - \alpha^2(x^{\mathsf{H}}x - y^{\mathsf{H}}x) &= 1 - \tfrac{1}{2}\alpha^2(x^{\mathsf{H}}x + x^{\mathsf{H}}x - y^{\mathsf{H}}x - y^{\mathsf{H}}x) = \\ &= 1 - \tfrac{1}{2}\alpha^2(x^{\mathsf{H}}x + y^{\mathsf{H}}y - y^{\mathsf{H}}x - x^{\mathsf{H}}y) = \\ &= 1 - \tfrac{1}{2}\alpha^2(x^{\mathsf{H}} - y^{\mathsf{H}})(x - y) = 1 - \tfrac{1}{2}\alpha^2\|x - y\|_2^2 = 0. \end{aligned}$$ ∎

[7] Spełnia ona tożsamości $(A^{\mathsf{H}})^{\mathsf{H}} = A$, $(AB)^{\mathsf{H}} = B^{\mathsf{H}}A^{\mathsf{H}}$.

[8] Dla macierzy unitarnej jest oczywiście $U^{-1} = U^{\mathsf{H}}$, czyli unitarne podobieństwo macierzy implikuje ich (zwykłe) podobieństwo (*przyp. tłum.*).

Twierdzenie 5.2.3 (Schur). *Każda macierz kwadratowa jest unitarnie podobna do macierzy trójkątnej górnej.*

Dowód. Dowód jest indukcyjny względem stopnia n macierzy. Twierdzenie jest oczywiste dla $n = 1$. Przypuśćmy, że jest ono prawdziwe dla wszystkich macierzy stopnia $n-1$ i rozważmy macierz A stopnia n. Niech λ będzie jej wartością własną, a $x = (x_1, x_2, \ldots, x_n)$ odpowiednim wektorem własnym. Nie tracąc ogólności, można założyć, że $\|x\|_2 = 1$. Niech będzie $\beta := \operatorname{sgn} x_1$ dla $x_1 \neq 0$ i $\beta := 1$ w przeciwnym razie. Wprowadzamy też wektor jednostkowy $e^{(1)} := (1, 0, \ldots, 0)$. Na mocy lematu 5.2.2 istnieje macierz unitarna U taka, że $Ux = \beta e^{(1)}$. Stąd $\beta^{-1}x = U^{\mathsf{H}}e^{(1)}$ i

$$UAU^{\mathsf{H}}e^{(1)} = UA\beta^{-1}x = \beta^{-1}\lambda Ux = \lambda e^{(1)}.$$

To dowodzi, że pierwszą kolumną iloczynu UAU^{H} jest $\lambda e^{(1)}$. Niech $\tilde{A}$ będzie macierzą powstałą z UAU^{H} przez skreślenie pierwszego wiersza i kolumny. Z założenia indukcyjnego wynika, że istnieje macierz unitarna Q stopnia $n-1$ taka, że iloczyn $Q\tilde{A}Q^{\mathsf{H}}$ jest trójkątny górny. Wprowadźmy macierz

$$V = \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix} U.$$

Jest ona unitarna jako iloczyn dwóch macierzy o tejże własności. Ponadto sprowadza ona A do postaci trójkątnej górnej. Istotnie,

$$VAV^{\mathsf{H}} = \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix} UAU^{\mathsf{H}} \begin{bmatrix} 1 & 0 \\ 0 & Q^{\mathsf{H}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix} \begin{bmatrix} \lambda & w \\ 0 & \tilde{A} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & Q^{\mathsf{H}} \end{bmatrix},$$

gdzie w środkowym czynniku w jest wektorem wierszowym o $n-1$ składowych, a 0 wektorem zerowym tego samego wymiaru. Tak więc

$$VAV^{\mathsf{H}} = \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix} \begin{bmatrix} \lambda & wQ^{\mathsf{H}} \\ 0 & \tilde{A}Q^{\mathsf{H}} \end{bmatrix} = \begin{bmatrix} \lambda & wQ^{\mathsf{H}} \\ 0 & Q\tilde{A}Q^{\mathsf{H}} \end{bmatrix}.$$

Otrzymana macierz jest trójkątna. ■

Warto zauważyć, że powyższy dowód nie daje metody konstrukcji macierzy trójkątnej unitarnie podobnej do danej macierzy.

Wniosek 5.2.4. *Każda macierz hermitowska jest unitarnie podobna do macierzy przekątniowej.*

Przypominamy, że z definicji macierz A jest hermitowska, jeśli $A^{\mathsf{H}} = A$ (podrozdz. 4.6).

Dowód. Niech U będzie macierzą unitarną taką, że iloczyn UAU^{H} jest macierzą trójkątną górną. Wtedy macierz $(UAU^{\mathsf{H}})^{\mathsf{H}}$ jest trójkątna dolna. Jest jednak $(UAU^{\mathsf{H}})^{\mathsf{H}} = (U^{\mathsf{H}})^{\mathsf{H}} A^{\mathsf{H}} U^{\mathsf{H}} = UAU^{\mathsf{H}}$, czyli ta ostatnia macierz – jako zarazem trójkątna górna i dolna – jest przekątniowa. ■

Przykład 5.2.5. Niżej podano przykład rozkładu Schura: $UAU^{\mathsf{H}} = T$, gdzie U jest unitarna, a T trójkątna górna.

$$\begin{bmatrix} 0.36 & 0.48 & 0.80 \\ 0.48 & 0.64 & -0.60 \\ 0.80 & -0.60 & 0.00 \end{bmatrix} \begin{bmatrix} 361 & 123 & -180 \\ 148 & 414 & -240 \\ -92 & 169 & 65 \end{bmatrix} \begin{bmatrix} 0.36 & 0.48 & 0.80 \\ 0.48 & 0.64 & -0.60 \\ 0.80 & -0.60 & 0.00 \end{bmatrix} =$$

$$= \begin{bmatrix} 125 & 380 & -125 \\ 0 & 465 & 1250 \\ 0 & 0 & 250 \end{bmatrix}.$$

■

Jeśli znamy wartość własną λ macierzy A stopnia n, to dowód twierdzenia Schura pokazuje, jak można utworzyć macierz $\tilde{A}$ stopnia $n-1$, której wartościami własnymi są pozostałe wartości macierzy A. Przejście od A do $\tilde{A}$ nazywamy *deflacją*.

Ten proces opisujemy ściśle tak:

1. Znaleźć wektor własny x, odpowiadający znanej wartości własnej λ.
2. Zdefiniować β wzorami $\beta := \operatorname{sgn} x_1$ dla $x_1 \neq 0$ i $\beta := 1$ w przeciwnym razie.
3. Znaleźć $\alpha := \sqrt{2}/\|x - \beta e^{(1)}\|_2$, $v := \alpha(x - \beta e^{(1)})$ i $U := I - vv^{\mathsf{H}}$.
4. Określić $\tilde{A}$ jako macierz wynikającą z UAU^{H} przez skreślenie pierwszego wiersza i tejże kolumny.

Warto wspomnieć, że deflację omawiano też w podrozdz. 3.5 w związku z obliczaniem pierwiastków wielomianu p. Po znalezieniu jednego z nich, np. ξ, można podzielić $p(x)$ przez $x - \xi$, co daje wielomian niższego stopnia o tych samych pierwiastkach, jednak z wyjątkiem ξ.

Większość metod numerycznych stosowanych do obliczania wartości własnych daje na raz tylko jedną taką wartość. Każdą taką metodę można połączyć z deflacją i dzięki temu obliczać tyle wartości własnych, ile nam potrzeba. W praktyce jednak trzeba zachować tu ostrożność, gdyż kolejne wartości własne mogą być coraz bardziej zaburzane przez błędy zaokrągleń.

Lokalizacja wartości własnych

Wiele twierdzeń daje informacje o położeniu wartości własnych na płaszczyźnie zespolonej. Najbardziej znane jest poniższe twierdzenie, w którym *widmo macierzy* oznacza zbiór jej wszystkich wartości własnych.

Twierdzenie 5.2.6 (Gerszgorin). *Widmo dowolnej macierzy A stopnia n zawiera się w sumie następujących kół na płaszczyźnie zespolonej:*

$$D_i = \Big\{z \in \mathbb{C}\colon\ |z - a_{ii}| \leqslant \sum_{j=1,\, j\neq i}^{n} |a_{ij}|\Big\} \quad (1 \leqslant i \leqslant n).$$

Dowód. Niech λ będzie wartością własną macierzy A i niech x będzie odpowiadającym jej wektorem własnym takim, że $\|x\|_\infty = 1$. Wobec tego dla pewnego i jest $|x_i| = 1$. Ponieważ $(Ax)_i = \lambda x_i$, więc

$$\lambda x_i = \sum_{j=1}^{n} a_{ij} x_j,$$

czyli

$$(\lambda - a_{ii}) x_i = \sum_{j=1,\, j\neq i}^{n} a_{ij} x_j.$$

Stąd, z nierówności trójkąta i nierówności $|x_j| \leqslant 1 = |x_i|$ wynika, że

$$|\lambda - a_{ii}| \leqslant \sum_{j=1,\, j\neq i}^{n} |a_{ij}|\,|x_j| \leqslant \sum_{j=1,\, j\neq i}^{n} |a_{ij}|.$$

Wobec tego $\lambda \in D_i$. ■

Przykład 5.2.7. Na rysunku 5.1 pokazano koła Gerszgorina dla macierzy

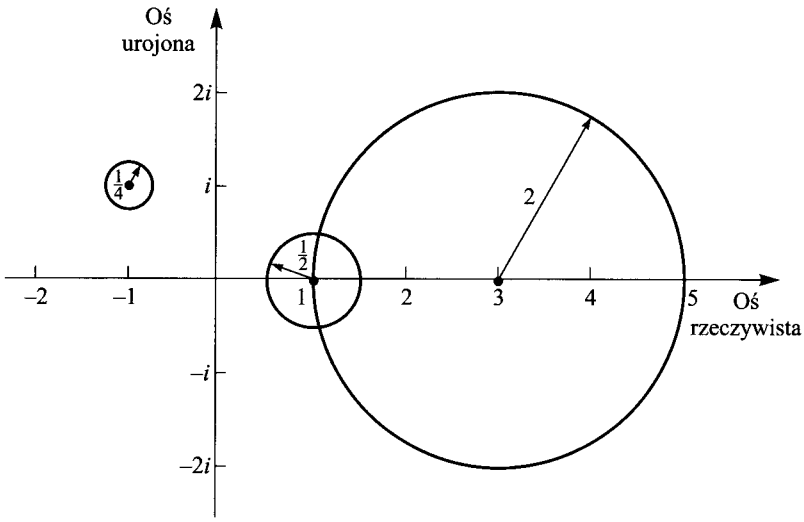
$$A := \begin{bmatrix} -1+i & 0 & \frac{1}{4} \\ \frac{1}{4} & 1 & \frac{1}{4} \\ 1 & 1 & 3 \end{bmatrix}.$$

Wszystkie jej wartości własne spełniają zatem nierówność $\frac{1}{2} \leqslant |\lambda| \leqslant 5$. ■

Twierdzenie 5.2.8. *Jeśli macierz A ma wartości własne $\lambda_1, \lambda_2, \ldots, \lambda_n$ i jest podobna do macierzy przekątniowej $P^{-1}AP$, a B jest dowolną macierzą, to każda wartość własna sumy $A + B$ leży w jednym z kół*

$$\{\lambda \in \mathbb{C}\colon\ |\lambda - \lambda_i| \leqslant \varkappa_\infty(P)\|B\|_\infty\},$$

gdzie $\varkappa_\infty(P) := \|P\|_\infty \|P^{-1}\|_\infty$.



RYS. 5.1. Koła Gerszgorina

Wielkość $\varkappa$ jest wskaźnikiem uwarunkowania określonym w podrozdz. 4.4 dla dowolnej normy macierzy; wyżej występuje konkretna norma $\|\cdot\|_\infty$.

Dowód. Jeśli $P^{-1}AP = D$, to macierz przekątniowa D ma na głównej przekątnej elementy $\lambda_1, \lambda_2, \ldots, \lambda_n$. Widma macierzy podobnych $A + B$ i $P^{-1}(A+B)P = D + C$, gdzie $C := P^{-1}BP$, są identyczne. Z twierdzenia Gerszgorina dla C wynika, że widmo macierzy $A + B$ składa się z kół

$$\Big\{\lambda \in \mathbb{C}\colon\ |\lambda - \lambda_i - c_{ii}| \leqslant \sum_{j=1,\, j\neq i}^{n} |d_{ij} + c_{ij}| = \sum_{j=1,\, j\neq i}^{n} |c_{ij}|\Big\}.$$

Stosujemy nierówność trójkąta i wyrażenie dla $\|C\|_\infty$ (zob. tw. 4.4.3):

$$|\lambda - \lambda_i| \leqslant |\lambda - \lambda_i - c_{ii}| + |c_{ii}| \leqslant \sum_{j=1}^{n} |c_{ij}| \leqslant$$
$$\leqslant \|C\|_\infty \leqslant \|P^{-1}\|_\infty \|B\|_\infty \|P\|_\infty = \varkappa_\infty(P)\|B\|_\infty. \qquad \blacksquare$$

Sens tw. 5.2.8 jest następujący: jeśli macierz A jest zaburzona o składnik B, to jej wartości własne są zaburzone co najwyżej o $\varkappa(P)\|B\|_\infty$.

Dla macierzy hermitowskiej (tzn. takiej, że $A^{\mathsf{H}} = A$) macierz P w tw. 5.2.8 może być unitarna, co wynika z wniosku 5.2.4. Wtedy wiersze macierzy P są wektorami o normie $\|\cdot\|_2$ równej 1. Stąd $\|P\|_\infty \leqslant \sqrt{n}$.

To samo jest prawdziwe dla P^{-1}, więc $\varkappa_\infty(P) \leqslant n$. Dlatego dla dowolnej macierzy B wartości własne sumy $A+B$ leżą w kołach

$$\{\lambda \in \mathbb{C}\colon\ |\lambda - \lambda_i| \leqslant n\|B\|_\infty\}.$$

ZADANIA 5.2

1. Udowodnić, że iloczyny AB i BA macierzy kwadratowych mają te same wartości własne.
2. Niech macierze A, B, C mają odpowiednio rozmiar: $n \times n$, $m \times m$ i $n \times m$, gdzie $n \geqslant m$. Wykazać, że jeśli rząd macierzy C jest równy m i $AC = CB$, to każda wartość własna macierzy B jest wartością własną macierzy A.
3. Udowodnić, że jeśli liczby $a, b, c, d \in \mathbb{R}$ są takie, że $a^2 + b^2 = c^2 + d^2 = 1$, to macierz

$$\begin{bmatrix} ad & ac & b \\ bd & bc & -a \\ c & -d & 0 \end{bmatrix}$$

jest unitarna.

4. Jaki jest wyznacznik macierzy unitarnej?
5. Czy macierz blokowa $\begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix}$ jest unitarna, jeśli U i V są unitarne?
6. Udowodnić, że zbiór wszystkich macierzy unitarnych ustalonego stopnia jest grupą ze względu na mnożenie i że wraz z A do tego zbioru należą A^{T}, A^{H} i $\bar{A}$.
7. Niech D będzie macierzą przekątniową, a A unitarną. Jakie dodatkowe warunki nałożone na D gwarantują, że iloczyn DU jest unitarny?
8. Co można powiedzieć o macierzy A, jeśli U i AU są unitarne?
9. Udowodnić, że jeśli Q jest macierzą unitarną, to dla wszystkich x i y

$$\|x\|_2 = \|Qx\|_2, \quad \langle x, y\rangle = \langle Qx, Qy\rangle,$$

czyli przekształcenie za pomocą Q zachowuje długości, odległości i kąty w przestrzeni euklidesowej. Obliczyć $\|Q\|_2$ (dla macierzy Q o elementach zespolonych jest $\|Q\|_2^2 = \rho(QQ^{\mathsf{H}})$).

10. Czy można twierdzić, że dla dowolnej macierzy A istnieje macierz unitarna i hermitowska U taka, że iloczyn UAU jest trójkątny?
11. Udowodnić, że dla macierzy Q unitarnej jest $\|A\|_2 = \|QA\|_2 = \|AQ\|_2$.
12. Udowodnić, że $\|A\|_2^2 = \|A^{\mathsf{H}}A\|_2$.
13. Niech A_j będzie j-tą kolumną macierzy A. Udowodnić, że dla dowolnej normy wektorowej i indukowanej przez nią normy macierzowej jest $\|A_j\| \leqslant \|A\|$.
14. Wykazać, że dla macierzy A przekątniowej jest $\|A\|_2 = \max_{1 \leqslant i \leqslant n} |a_{ii}|$.
15. Wykazać, że jeśli $x^{\mathsf{H}}x = 1$ i $U = I - 2xx^{\mathsf{H}}$, to $U^2 = I$.

16. Znaleźć $(I - xx^{\mathsf{H}})^{-1}$ wiedząc, że $x^{\mathsf{H}}x = 2$.

17. Udowodnić, że macierz $I - xx^{\mathsf{H}}$ jest osobliwa wtedy i tylko wtedy, gdy $x^{\mathsf{H}}x = 1$ i znaleźć jej odwrotność w pozostałych przypadkach.

18. Udowodnić, że jeśli dla pewnych wektorów v, x, y jest $(I - vv^{\mathsf{H}})x = y$, to iloczyn skalarny $\langle x, y\rangle$ jest rzeczywisty.

19. Znaleźć warunki konieczne i dostateczne na to, żeby macierz $I - uv^{\mathsf{H}}$ była unitarna.

20. Udowodnić, że jeśli wektory x, y mają tę samą normę euklidesową, to istnieje macierz unitarna U taka, że $Ux = y$.

21. Udowodnić, że jeśli wektor x jest taki, że $\|x\|_2 = 1$, to istnieje macierz unitarna, której wybrana kolumna jest równa x.

22. Udowodnić, że $\det(I + xx^{\mathsf{H}}) = 1 + x^{\mathsf{H}}x$. Wskazówka: Istnieje macierz unitarna odwzorowująca x na pewną wielokrotność wektora jednostkowego $e^{(1)} = (1, 0, 0, \ldots, 0)$.

23. Znaleźć rozkłady Schura dla macierzy

$$\begin{bmatrix} 3 & 8 \\ -2 & 3 \end{bmatrix}, \quad \begin{bmatrix} 4 & 7 \\ -1 & 12 \end{bmatrix}, \quad \begin{bmatrix} 2.888 & 0.984 & -1.440 \\ 1.184 & 3.312 & -1.920 \\ -0.160 & 2.120 & -0.200 \end{bmatrix}.$$

W ostatnim przypadku można użyć macierzy unitarnej z przykł. 5.2.5.

24. Czy w definicji kół D_i w tw. Gerszgorina 5.2.6 można zmienić $|a_{ij}|$ na $|a_{ji}|$?

25. Niech $D_1, D_2, \ldots, D_n$ będą kołami Gerszgorina dla macierzy A. Załóżmy, że jej wartość własna λ leży w D_k, ale nie w D_i dla $i \neq k$. Wykazać, że wektor własny $(x_1, x_2, \ldots, x_n)$, odpowiadający wartości λ, jest taki, że $|x_k| > |x_i|$ dla $i \neq k$.

26. Wiadomo, że jeśli $|a_{ii} - \lambda| > \sum_{j=1,\, j\neq i}^{n} |a_{ij}|$ dla $1 \leqslant i \leqslant n$, to $A - \lambda I$ jest macierzą przekątniowo dominującą. Jak stąd wynika tw. Gerszgorina 5.2.6?

27. Udowodnić, że wartości własne macierzy

$$\begin{bmatrix} 6 & 2 & 1 \\ 1 & -5 & 0 \\ 2 & 1 & 4 \end{bmatrix}$$

spełniają nierówność $1 \leqslant |\lambda| \leqslant 9$.

28. Wykazać, że część urojona każdej z wartości własnych macierzy

$$\begin{bmatrix} 3 & \frac{1}{3} & \frac{2}{3} \\ 1 & -4 & 0 \\ \frac{1}{2} & \frac{1}{2} & -1 \end{bmatrix}$$

leży w przedziale $[-1, 1]$.

29. Naszkicować koła Gerszgorina dla macierzy

$$A = \begin{bmatrix} 0 & 2 & -1 \\ -2 & -10 & 0 \\ -1 & -1 & 4 \end{bmatrix}$$

i oszacować promień spektralny $\rho(A)$.

30. Jak można oszacować promień spektralny $\rho(A)$ macierzy

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

posługując się: (**a**) normą $\|A\|_1$ (zob. podrozdz. 4.6), (**b**) tw. Gerszgorina 5.2.6?

5.3. Ortogonalizacja i zadanie najmniejszych kwadratów

Iloczyn skalarny $\langle \cdot, \cdot \rangle$ wektorów z $\mathbb{C}^n$ pozwala określić pojęcie *ortogonalności*. Układ wektorów $v_1, v_2, \ldots, v_k$ jest z definicji *ortogonalny*, jeśli $\langle v_i, v_j \rangle = 0$ dla $i \neq j$. Jeśli dodatkowo $\langle v_i, v_i \rangle$, czyli $\|v_i\|_2$, jest równe 1 dla $1 \leqslant i \leqslant k$, to ten układ jest *ortonormalny*. W tym drugim przypadku, jeśli wektory v_i są kolumnami macierzy A, to $A^{\mathsf{H}}A = I$ [9].

Podstawowe pojęcia

Załóżmy, że wektory $v_1, v_2, \ldots, v_n$ są bazą ortonormalną przestrzeni $\mathbb{C}^n$. Wtedy każdy wektor $x \in \mathbb{C}^n$ wyraża się tylko w jeden sposób jako kombinacja liniowa wektorów tej bazy:

$$x = \sum_{i=1}^{n} c_i v_i \quad (c_i \in \mathbb{C}).$$

Pomnóżmy skalarnie obie strony tej równości przez v_j:

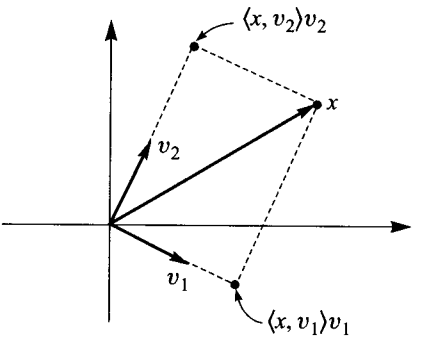
$$\langle x, v_j \rangle = \left\langle \sum_{i=1}^{n} c_i v_i, v_j \right\rangle = \sum_{i=1}^{n} c_i \langle v_i, v_j \rangle = c_j.$$

Stąd wynika, że dla każdego $x \in \mathbb{C}^n$

$$x = \sum_{i=1}^{n} \langle x, v_i \rangle v_i.$$

[9] Stąd wynika, że kolumny, a także wiersze, macierzy unitarnej (zob. podrozdz. 5.2) tworzą układ ortonormalny (*przyp. tłum.*).

i-ty składnik tej sumy nazywamy *składową* wektora x w kierunku v_i. Typową sytuację w $\mathbb{R}^2$ pokazuje rys. 5.2.



RYS. 5.2. Składowe ortogonalne wektora

Uogólnieniem przestrzeni $\mathbb{C}^n$ jest abstrakcyjna *przestrzeń unitarna* (czyli *przestrzeń z iloczynem skalarnym*). Jest to przestrzeń liniowa nad ciałem liczb zespolonych, w której określono iloczyn skalarny o niżej podanych własnościach. Elementy x, y,... przestrzeni nazywamy i tu wektorami.

1. $\langle x, x\rangle > 0$, jeśli $x \neq 0$.
2. $\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle$, gdzie α, $\beta \in \mathbb{C}$.
3. $\langle x, y\rangle = \overline{\langle y, x\rangle}$.

Stąd wynika, że $\langle x, y + z\rangle = \langle x, y\rangle + \langle x, z\rangle$ i $\langle x, \alpha y\rangle = \bar{\alpha}\langle x, y\rangle$.

Normę, ortogonalność i ortonormalność w tej przestrzeni określamy tak samo, jak w szczególnym przypadku $\mathbb{C}^n$. Przestrzeń unitarna może być nieskończeniewymiarowa. W każdej takiej przestrzeni zachodzi równość analogiczna do elementarnego twierdzenia Pitagorasa: jeśli $\|x\|_2 := (\langle x, x\rangle)^{1/2}$ i $\langle x, y\rangle = 0$, to

$$\|x + y\|_2^2 = \|x\|_2^2 + \|y\|_2^2. \tag{5.3.1}$$

Istotnie, wtedy

$$\|x+y\|_2^2 = \langle x+y, x+y\rangle = \langle x, x\rangle + \langle y, x\rangle + \langle x, y\rangle + \langle y, y\rangle = \|x\|_2^2 + \|y\|_2^2.$$

Ortogonalizacja Grama-Schmidta

Dla danego ciągu wektorów liniowo niezależnych $x_1, x_2, \ldots$ z przestrzeni unitarnej klasyczna procedura *ortogonalizacji Grama-Schmidta* pozwala zbudować ciąg ortonormalny $u_1, u_2, \ldots$, złożony z kombinacji liniowych tamtych wektorów. W twierdzeniu 5.3.1 dotyczącym tej procedury symbol $\operatorname{span}\{x, y, \ldots, z\}$ oznacza podprzestrzeń liniową złożoną ze wszystkich kombinacji liniowych wektorów x, y, $\ldots$, z, czyli *rozpiętą* na nich.

Twierdzenie 5.3.1. *Ciąg wektorów*

$$u_k := \Big\|x_k - \sum_{i<k}\langle x_k, u_i\rangle u_i\Big\|_2^{-1}\Big[x_k - \sum_{i<k}\langle x_k, u_i\rangle u_i\Big] \quad (k \geqslant 1) \tag{5.3.2}$$

jest taki, że dla każdego $k \geqslant 1$ układ $\{u_1, u_2, \ldots, u_k\}$ jest bazą ortonormalną podprzestrzeni $\operatorname{span}\{x_1, x_2, \ldots, x_k\}$.

Dowód. Dowód jest indukcyjny względem k. Dla $k = 1$ z (5.3.2) wynika, że $u_1 = \|x_1\|_2^{-1}x_1$ (norma każdego x_k jest dodatnia z założenia o niezależności liniowej), czyli u_1 ma normę równą 1. Podprzestrzenie liniowe $\operatorname{span}\{u_1\}$ i $\operatorname{span}\{x_1\}$ są identyczne.

Załóżmy, że układ $\{u_1, u_2, \ldots, u_{k-1}\}$ jest bazą ortonormalną dla podprzestrzeni $\operatorname{span}\{x_1, x_2, \ldots, x_{k-1}\}$. Niech będzie

$$v = x_k - \sum_{i<k}\langle x_k, u_i\rangle u_i. \tag{5.3.3}$$

Taki wektor v jest ortogonalny względem u_j dla $j < k$, gdyż

$$\begin{aligned}\langle v, u_j\rangle &= \langle x_k, u_j\rangle - \sum_{i<k}\langle x_k, u_i\rangle\langle u_i, u_j\rangle = \\ &= \langle x_k, u_j\rangle - \sum_{i<k}\langle x_k, u_i\rangle\delta_{ij} = \langle x_k, u_j\rangle - \langle x_k, u_j\rangle = 0.\end{aligned}$$

Sprawdzimy, że nie może być $v = 0$. Istotnie, wtedy z (5.3.3) wynikałoby, że $x_k \in \operatorname{span}\{u_1, u_2, \ldots, u_{k-1}\}$, a zatem $x_k \in \operatorname{span}\{x_1, x_2, \ldots, x_{k-1}\}$, wbrew niezależności liniowej wektorów $x_1, x_2, \ldots, x_k$. Ponieważ $v \neq 0$, więc można określić u_k jako $(\|v\|_2)^{-1}v$. Jest oczywiście $\|u_k\|_2 = 1$, czyli układ $\{u_1, u_2, \ldots, u_k\}$ jest ortonormalny. Z założenia indukcyjnego i (5.3.3) wynika, że $v \in \operatorname{span}\{x_1, x_2, \ldots, x_k\}$. Wiemy już zatem, że $\operatorname{span}\{u_1, u_2, \ldots, u_k\}$ zawiera się w $\operatorname{span}\{x_1, x_2, \ldots, x_k\}$. Ponieważ jednak układy $\{u_1, u_2, \ldots, u_k\}$ i $\{x_1, x_2, \ldots, x_k\}$ są liniowo niezależne, więc te podprzestrzenie są identyczne. ∎

Ortogonalizację Grama-Schmidta stosowaną do kolumn $A_1, A_2, \ldots, A_n$ macierzy A rozmiaru $m \times n$ możemy interpretować jako jej rozkład QR na czynniki B i T. W macierzy T występują m.in. iloczyny skalarne z (5.3.3). Natomiast kolumnami macierzy B rozmiaru $m \times n$ są wektory u_k. Całą procedurę można zwięźle opisać tak:

for $j = 1$ **to** n **do**
  for $i = 1$ **to** $j - 1$ **do**
    $t_{ij} \leftarrow \langle A_j, B_i\rangle$

end do
$C_j \leftarrow A_j - \sum_{i<j} t_{ij} B_i$
$t_{jj} \leftarrow \|C_j\|_2$
$B_j \leftarrow t_{jj}^{-1} C_j$
end do

TWIERDZENIE 5.3.2. *Ortogonalizacja Grama-Schmidta zastosowana do kolumn macierzy A rozmiaru $m \times n$ i rzędu n daje rozkład $A = BT$, w którym B jest macierzą tegoż rozmiaru o kolumnach ortonormalnych, a T jest macierzą trójkątną górną stopnia n o dodatnich elementach przekątniowych*[10].

Dowód. Z (5.3.2) wynika, że

$$x_j = \sum_{i \leqslant j} t_{ij} u_i,$$

gdzie

$$t_{ij} = \langle x_j, u_i \rangle \quad (i < j), \qquad t_{jj} = \Big\| x_j - \sum_{i<j} t_{ij} u_i \Big\|_2 > 0$$

(taki sam sens mają wielkości t_{ij} tworzone w ostatnim algorytmie). Jeśli x_j jest j-tą kolumną macierzy A, a u_i jest i-tą kolumną macierzy B oraz $t_{ij} = 0$ dla $i > j$, to

$$A_j = \sum_{i=1}^{n} t_{ij} B_i \quad (1 \leqslant j \leqslant n),$$

co jest równoważne relacji $A = BT$. ■

Zmodyfikowany algorytm Grama-Schmidta

Doświadczalnie stwierdzono (Rice [1966]), że pewna modyfikacja podanego wyżej algorytmu ma lepsze własności numeryczne[11]. W tej wersji składowe kolumny A_j dla $j > k$ w kierunku wektora bazowego A_k są zerowane, gdy tylko jest to możliwe. Dodatkowa zmiana usuwa pierwiastkowania konieczne, gdy obliczamy normy wektorów. To jednak powoduje, że rozkład macierzy A

[10] Warto przypomnieć, że założenie o rzędzie takiej macierzy A jest równoważne temu, że jej kolumny są niezależne liniowo (*przyp. tłum.*).

[11] Kiełbasiński i Schwetlick [*1992, s. 295], uzasadniają bardziej radykalną opinię: pierwotny algorytm jest „katastrofalnie numerycznie niestabilny" i „nie nadaje się do realizacji numerycznej" (*przyp. tłum.*).

jest teraz nieco inny: czynnik T ma na głównej przekątnej same jedynki, a kolumny macierzy B tworzą tylko układ ortogonalny, a nie ortonormalny.

Zmodyfikowany algorytm Grama-Schmidta jest następujący:

for $k = 1$ **to** n **do**
  $d_k \leftarrow \|A_k\|_2^2$
  $t_{kk} \leftarrow 1$
  for $j = k+1$ **to** n **do**
    $t_{kj} \leftarrow d_k^{-1}\langle A_j, A_k\rangle$
    $A_j \leftarrow A_j - t_{kj}A_k$
  end do
end do

Algorytm zapamiętuje kolumny macierzy B na miejscu danych kolumn A_j macierzy A.

Twierdzenie 5.3.3. *Zmodyfikowany algorytm Grama-Schmidta zastosowany do kolumn macierzy A rozmiaru $m \times n$ i rzędu n daje rozkład $A = BT$, w którym B jest macierzą tegoż rozmiaru o kolumnach ortogonalnych, a T jest macierzą trójkątną górną jedynkową stopnia n.*

Dowód. Aby ułatwić dowód, oznaczymy odrębnym symbolem $A_j^{(k)}$ każdy wektor występujący w powyższym algorytmie:

for $k = 1$ **to** n **do**
  $d_k \leftarrow \|A_k^{(k)}\|_2^2$
  $t_{kk} \leftarrow 1$
  for $j = k+1$ **to** n **do**
    $t_{kj} \leftarrow d_k^{-1}\langle A_j^{(k)}, A_k^{(k)}\rangle$
    $A_j^{(k+1)} \leftarrow A_j^{(k)} - t_{kj}A_k^{(k)}$
  end do
end do

Tak więc kolumnami A_j macierzy A i kolumnami B_k macierzy B są teraz odpowiednio wektory $A_j^{(1)}$ $(1 \leqslant j \leqslant n)$ i $A_k^{(k)}$ $(1 \leqslant k \leqslant n)$.

Stosując instrukcję z szóstego wiersza algorytmu wnioskujemy, że dla każdego $l = 1, 2, \ldots, k-1$ jest

$$A_k^{(k)} = A_k^{(l)} - \sum_{i=l}^{k-1} t_{ik}A_i^{(i)}. \tag{5.3.4}$$

Dla $l = 1$ wynika stąd, że

$$A_k^{(k)} = A_k^{(1)} - \sum_{i<k} t_{ik}A_i^{(i)},$$

czyli $A_k = \sum_{i \leqslant k} t_{ik} B_i$. Jeśli $t_{ik} = 0$ dla $i > k$ (macierz T jest trójkątna górna), to – jak w poprzednim dowodzie – jest $A = BT$.

Trzeba jeszcze wykazać, że kolumny macierzy B tworzą układ ortogonalny, tj. że $\langle A_k^{(k)}, A_l^{(l)} \rangle = 0$ dla $1 \leqslant k < l \leqslant n$. W tym celu w (5.3.4) uwzględniamy wartości podstawiane pod t_{ik}:

$$A_k^{(k)} = A_k^{(l)} - \sum_{i=l}^{k-1} \frac{\langle A_k^{(i)}, A_i^{(i)} \rangle}{\langle A_i^{(i)}, A_i^{(i)} \rangle} A_i^{(i)}.$$

Stąd

$$\langle A_k^{(k)}, A_l^{(l)} \rangle = \langle A_k^{(l)}, A_l^{(l)} \rangle - \sum_{i=l}^{k-1} \frac{\langle A_k^{(i)}, A_i^{(i)} \rangle}{\langle A_i^{(i)}, A_i^{(i)} \rangle} \langle A_i^{(i)}, A_l^{(l)} \rangle =$$
$$= - \sum_{i=l+1}^{k-1} \frac{\langle A_k^{(i)}, A_i^{(i)} \rangle}{\langle A_i^{(i)}, A_i^{(i)} \rangle} \langle A_i^{(i)}, A_l^{(l)} \rangle.$$

Udowodnimy indukcyjnie, przyjmując kolejno $l = k-1, k-2, \ldots, 1$, iż iloczyny skalarne po lewej stronie równości znikają. Jest to oczywiste dla $l = k-1$, bo wtedy suma po prawej stronie jest pusta. Jeśli $l < k-1$, to w sumie występują iloczyny skalarne $\langle A_i^{(i)}, A_l^{(l)} \rangle$ dla $1 \leqslant i - l \leqslant k - l - 1$; ich zerowanie się wykazano w poprzednich krokach rozumowania. ■

Zadanie najmniejszych kwadratów

Ważnym zastosowaniem rozkładu ortogonalnego (nazwa wiąże się z własnością macierzy B) jest *zadanie najmniejszych kwadratów* dla układu równań liniowych, które teraz zdefiniujemy. Rozważamy układ m równań z n niewiadomymi

$$Ax = b,$$

gdzie $m > n$; macierz A ma zatem rozmiar $m \times n$, $x \in \mathbb{C}^n$, $b \in \mathbb{C}^m$. Zakładamy, że rząd macierzy A jest równy n. Układ $Ax = b$ na ogół nie ma rozwiązania. W takich przypadkach szukamy często takiego x, które daje minimalną normę – będziemy tu używać normy euklidesowej $\|\cdot\|_2$ – wektora residualnego $b - Ax$. Jest to „rozwiązanie" układu $Ax = b$ w sensie *najmniejszych kwadratów*. Można udowodnić, że jeśli $\operatorname{rank} A = n$, to ten wektor x jest określony jednoznacznie.

Lemat 5.3.4. *Rozwiązaniem zadania najmniejszych kwadratów jest wektor x taki, że $A^{\mathsf{H}}(Ax - b) = 0$.*

Dowód. Niech y będzie wektorem z $\mathbb{C}^n$. Ponieważ $A^{\mathsf{H}}(Ax-b)=0$, więc wektor $b-Ax$ jest ortogonalny względem przestrzeni rozpiętej na kolumnach macierzy A. Wektor $A(x-y)$ do niej należy, więc $\langle b-Ax, A(x-y)\rangle = 0$ i na mocy (5.3.1) jest

$$\begin{aligned}\|b-Ay\|_2^2 &= \|b-Ax+A(x-y)\|_2^2 = \\ &= \|b-Ax\|_2^2 + \|A(x-y)\|_2^2 \geqslant \|b-Ax\|_2^2.\end{aligned}$$ ■

Jeśli macierz A rozłożono na czynniki B i T określone w tw. 5.3.3, to rozwiązanie x układu $Ax=b$ w sensie najmniejszych kwadratów jest dokładnym rozwiązaniem układu

$$Tx = (B^{\mathsf{H}}B)^{-1}B^{\mathsf{H}}b.$$

Wynika to z lematu 5.3.4:

$$A^{\mathsf{H}}Ax = (BT)^{\mathsf{H}}BTx = T^{\mathsf{H}}B^{\mathsf{H}}B(B^{\mathsf{H}}B)^{-1}B^{\mathsf{H}}b = T^{\mathsf{H}}B^{\mathsf{H}}b = A^{\mathsf{H}}b.$$

Jest też $(B^{\mathsf{H}}B)^{-1} = \operatorname{diag}(d_1^{-1}, d_2^{-1}, \ldots, d_n^{-1})$, gdzie liczby d_i są określone w zmodyfikowanym algorytmie Grama-Schmidta.

Innym podejściem do zadania najmniejszych kwadratów wiążącego się z układem $Ax=b$ jest bezpośrednie zastosowanie lematu 5.3.4. Wiemy, że norma $\|Ax-b\|_2$ jest najmniejsza, gdy

$$A^{\mathsf{H}}(Ax-b)=0.$$

Jeśli macierz A rozmiaru $m\times n$ jest rzędu n, to $A^{\mathsf{H}}A$ jest macierzą nieosobliwą stopnia n (zad. 14) i zadanie najmniejszych kwadratów ma dokładnie jedno rozwiązanie, będące rozwiązaniem tzw. *układu równań normalnych*

$$A^{\mathsf{H}}Ax = A^{\mathsf{H}}b.$$

Wiadomo też, że macierz $A^{\mathsf{H}}A$ jest hermitowska i dodatnio określona (zob. to samo zadanie). Dlatego do rozwiązania tego układu można stosować metodę Cholesky'ego. Jeśli $\operatorname{rank} A < n$, to układ jest niesprzeczny, ale ma wiele rozwiązań.

Bezpośrednie zastosowanie równań normalnych w zadaniu najmniejszych kwadratów wydaje się sensowne wobec pojęciowej prostoty tej metody. Nie jest ona jednak godna polecenia. Wynika to m.in. stąd, że macierz $A^{\mathsf{H}}A$ może być znacznie gorzej uwarunkowana niż A. Widać to na przykładzie:

$$A := \begin{bmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{bmatrix}, \quad A^{\mathsf{H}}A = \begin{bmatrix} 1+\varepsilon^2 & 1 & 1 \\ 1 & 1+\varepsilon^2 & 1 \\ 1 & 1 & 1+\varepsilon^2 \end{bmatrix}. \tag{5.3.5}$$

W macierzy A tylko niezerowe ε powoduje, że jej rząd nie jest równy 1. W $A^{\mathsf{H}}A$ tę samą rolę odgrywa ε^2. Dlatego nieosobliwość tego iloczynu jest szczególnie wątpliwa dla małych ε. W komputerze możemy mieć macierz A rzędu 3 i wynikającą z niej macierz $A^{\mathsf{H}}A$ rzędu 1.

Metoda Householdera

Przejdziemy teraz do najbardziej użytecznej metody rozkładu ortogonalnego, zaproponowanej przez Alstona Householdera i noszącej teraz jego imię. Celem metody jest rozkład

$$A = QR$$

macierzy A rozmiaru $m \times n$ (o jej rzędzie nic nie zakładamy), gdzie Q ma być macierzą unitarną stopnia m, a R macierzą trójkątną górną rozmiaru $m \times n$; stąd $(R)_{ij} = 0$ dla $i > j$. Ściślej, algorytm daje Q^{H} i R takie, że

$$Q^{\mathsf{H}}A = R,$$

a Q^{H} powstaje krok po kroku jako iloczyn macierzy unitarnych postaci

$$\begin{bmatrix} I_k & 0 \\ 0 & I_{m-k} - vv^{\mathsf{H}} \end{bmatrix}$$

(wskaźnik precyzuje wyżej stopień macierzy jednostkowej). Przekształcenie za pomocą takiej macierzy nazywamy *odbiciem* lub *przekształceniem Householdera.*

Metoda[12)] zaczyna się od wyboru takiego $v \in \mathbb{C}^m$, żeby macierz $I - vv^{\mathsf{H}}$ była unitarna i żeby pierwsza kolumna iloczynu $(I - vv^{\mathsf{H}})A$ miała, jak w R, postać $(\beta, 0, \ldots, 0)$. Niech A_1 będzie pierwszą kolumną macierzy A. Chcemy więc, żeby było $(I - vv^{\mathsf{H}})A_1 = \beta e^{(1)}$, gdzie $e^{(1)}$ jest pierwszym wektorem jednostkowym. Według dowodu lematu 5.2.2 można to osiągnąć, wybierając liczbę zespoloną β taką, że $|\beta| = \|A_1\|_2$ i że $\langle A_1, \beta e^{(1)} \rangle$ jest rzeczywiste oraz przyjmując $v := \alpha(A_1 - \beta e^{(1)})$, gdzie $\alpha := \sqrt{2}/\|A_1 - \beta e^{(1)}\|_2$. Parametr β ma dwie możliwe wartości. Aby uzasadnić wybór jednej z nich, napiszmy liczby zespolone β i a_{11} odpowiednio w postaci

$$\beta = \|A_1\|_2 e^{i\varphi}, \quad a_{11} = |a_{11}| e^{i\theta}.$$

[12)] Trzeba uprzedzić czytelników, że ciąg dalszy, włącznie z przykładem 5.3.5, nie wyjaśnia istotnych praktycznych szczegółów konstrukcji rozkładu $A = QR$ za pomocą przekształceń Householdera; zob. Dryja, Jankowscy [*1982, s. 54–55] (*przyp. tłum.*).

Mamy więc

$$\langle A_1, \beta e^{(1)} \rangle = a_{11}\bar{\beta} = |a_{11}| \, \|A_1\|_2 e^{i(\theta-\varphi)}.$$

Ta wielkość ma być rzeczywista, czyli $\theta - \varphi$ musi być równe 0 albo π. Dla $\theta - \varphi = \pi$ obliczanie pierwszej składowej wektora v jest bezpieczne, bo nie powoduje odejmowania wielkości tego samego znaku:

$$v_1 = \alpha(a_{11} - \beta) = \alpha(|a_{11}|e^{i\theta} - |\beta|e^{i(\theta-\pi)}) = \alpha(|a_{11}| + |\beta|)e^{i\theta}.$$

Dlatego definiujemy β wzorem

$$\beta := -\|A_1\|_2 e^{i\theta} = -\|A_1\|_2 a_{11}/|a_{11}|.$$

Ostatecznie daje to następujący algorytm konstrukcji macierzy unitarnej w pierwszym kroku:

$\beta \leftarrow -(a_{11}/|a_{11}|)\|A_1\|_2$
$y \leftarrow A_1 - \beta e^{(1)}$
$\alpha \leftarrow \sqrt{2}/\|y\|_2$
$v \leftarrow \alpha y$
$U \leftarrow I - vv^{\mathsf{H}}$

(nieco lepszy wariant tego algorytmu można znaleźć w zad. 8). Następne kroki rozkładu QR są podobne. Po k krokach macierz A jest pomnożona z lewej strony przez k macierzy unitarnych, a wynik tych mnożeń jest macierzą, w której k początkowych kolumn ma już poprawną postać, tj. zera poniżej głównej przekątnej. Można to wyrazić wzorem

$$U_k U_{k-1} \ldots U_1 A = \begin{bmatrix} J & H \\ 0 & W \end{bmatrix},$$

gdzie J jest macierzą trójkątną górną stopnia k, 0 jest macierzą zerową rozmiaru $(m-k)\times k$, a macierze H i W mają odpowiednio rozmiar $k\times(n-k)$ i $(m-k)\times(n-k)$. Jak już wiemy, istnieje wektor $v \in \mathbb{C}^{m-k}$ taki, że $I - vv^{\mathsf{H}}$ jest macierzą unitarną stopnia $m-k$, a iloczyn $(I - vv^{\mathsf{H}})W$ ma w pierwszej kolumnie zera począwszy od drugiego elementu. Zauważmy teraz, że

$$\begin{bmatrix} I & 0 \\ 0 & I - vv^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} J & H \\ 0 & W \end{bmatrix} = \begin{bmatrix} J & H \\ 0 & (I - vv^{\mathsf{H}})W \end{bmatrix}.$$

Pierwszy czynnik po lewej stronie jest unitarny i oznaczamy go symbolem U_{k+1}.

Proces się kończy po otrzymaniu macierzy R założonego wcześniej typu. Mamy wtedy równość $Q^{\mathsf{H}}A = R$, gdzie Q^{H} jest iloczynem wszystkich macierzy unitarnych U_k. Ponieważ Q jest unitarna, więc $A = QR$, co chcieliśmy otrzymać. Z równości $Q^{\mathsf{H}} = U_{n-1} \dots U_2 U_1$ wynika, że $Q = U_1^{\mathsf{H}} U_2^{\mathsf{H}} \dots U_{n-1}^{\mathsf{H}}$. Wobec wzoru

$$U_k = \begin{bmatrix} I_{k-1} & 0 \\ 0 & I_{n-k+1} - vv^{\mathsf{H}} \end{bmatrix}$$

macierz U_k jest hermitowska ($U_k^{\mathsf{H}} = U_k$), więc

$$Q = U_1 U_2 \dots U_{n-1}.$$

Przykład 5.3.5. Stosując przekształcenia Householdera, wyznaczyć rozkład QR macierzy

$$A := \begin{bmatrix} 63 & 41 & -88 \\ 42 & 60 & 51 \\ 0 & -28 & 56 \\ 126 & 82 & -71 \end{bmatrix}.$$

Rozwiązanie. W pierwszym kroku obliczamy β, równe $-\|A_1\|_2$, gdyż macierz A jest rzeczywista:

$$\beta = -\|A_1\|_2 = -\|(63,\, 42,\, 0,\, 126)\|_2 = -147.$$

Następnie obliczamy α:

$$\alpha = \sqrt{2}/\|A_1 - \beta e^{(1)}\|_2 = \sqrt{2}/\|(210,\, 42,\, 0,\, 126)\|_2 = 1/(21\sqrt{70}).$$

Wobec tego

$$v = \alpha(A_1 - \beta e^{(1)}) = (10,\, 2,\, 0,\, 6)/\sqrt{70},$$

a pierwszym czynnikiem unitarnym jest macierz

$$U_1 = I - vv^{\mathsf{H}} = \frac{1}{35} \begin{bmatrix} -15 & -10 & 0 & -30 \\ -10 & 33 & 0 & -6 \\ 0 & 0 & 35 & 0 \\ -30 & -6 & 0 & 17 \end{bmatrix}.$$

Obliczamy teraz iloczyn

$$U_1 A = \frac{1}{35} \begin{bmatrix} -5145 & -3675 & 2940 \\ 0 & 1078 & 2989 \\ 0 & -980 & 1960 \\ 0 & -196 & 1127 \end{bmatrix}.$$

W drugim kroku podobne obliczenia dają następujące wyniki:

$$\beta = -\|(30.8, -28, -5.6)\|_2 = -42,$$
$$\alpha = \sqrt{2}/\|(72.8, -28, -5.6)\|_2 = 0.018085,$$
$$v = \alpha(1.3166, -0.50637, -0.10127),$$
$$I - vv^{\mathsf{H}} = \begin{bmatrix} -0.73333 & 0.66667 & 0.13333 \\ 0.66667 & 0.74359 & -0.05128 \\ 0.13333 & -0.05128 & 0.98974 \end{bmatrix}.$$

Dlatego drugim czynnikiem unitarnym jest macierz

$$U_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -0.73333 & 0.66667 & 0.13333 \\ 0 & 0.66667 & 0.74359 & -0.05128 \\ 0 & 0.13333 & -0.05128 & 0.98974 \end{bmatrix},$$

a stąd

$$U_2U_1A = \begin{bmatrix} -147 & -105 & 84 \\ 0 & -42 & -21 \\ 0 & 0 & 96.9231 \\ 0 & 0 & 40.3846 \end{bmatrix}.$$

Ostatni krok daje następujące wyniki:

$$\beta = -\|(96.9231, 40.3846)\|_2 = -105,$$
$$\alpha = \sqrt{2}/\|(201.9231, 40.3846)\|_2 = 0.0068677,$$
$$v = (1.38675, 0.27735),$$
$$I - vv^{\mathsf{H}} = \begin{bmatrix} -0.92308 & -0.38462 \\ -0.38462 & 0.92308 \end{bmatrix}.$$

Wynika stąd trzeci czynnik unitarny:

$$U_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -0.92308 & -0.38462 \\ 0 & 0 & -0.38462 & 0.92308 \end{bmatrix}.$$

Macierz trójkątna górna R jest więc następująca:

$$R = \begin{bmatrix} -147 & -105 & 84 \\ 0 & -42 & -21 \\ 0 & 0 & -105 \\ 0 & 0 & 0 \end{bmatrix} = 21 \begin{bmatrix} -7 & -5 & 4 \\ 0 & -2 & -1 \\ 0 & 0 & -5 \\ 0 & 0 & 0 \end{bmatrix}.$$

Natomiast macierz Q^{H} jest równa

$$Q^{\mathsf{H}} = U_3U_2U_1 = \begin{bmatrix} -0.42857 & -0.28571 & 0 & -0.85714 \\ 0.09524 & -0.71429 & 0.66667 & 0.19048 \\ 0.47619 & -0.57143 & -0.66667 & -0.04762 \\ -0.76190 & -0.28571 & -0.33333 & 0.47619 \end{bmatrix} =$$

$$= \frac{1}{21}\begin{bmatrix} -9 & -6 & 0 & -18 \\ 2 & -15 & 14 & 4 \\ 10 & -12 & -14 & -1 \\ -16 & -6 & -7 & 10 \end{bmatrix}.$$

Możemy sprawdzić, że $A = QR$:

$$\begin{bmatrix} 63 & 41 & -88 \\ 42 & 60 & 51 \\ 0 & -28 & 56 \\ 126 & 82 & -71 \end{bmatrix} = \begin{bmatrix} -9 & 2 & 10 & -16 \\ -6 & -15 & -12 & -6 \\ 0 & 14 & -14 & -7 \\ -18 & 4 & -1 & 10 \end{bmatrix}\begin{bmatrix} -7 & -5 & 4 \\ 0 & -2 & -1 \\ 0 & 0 & -5 \\ 0 & 0 & 0 \end{bmatrix}. \quad \blacksquare$$

ZADANIA 5.3

1. Czemu jest równy wyznacznik macierzy kwadratowej, której kolumny tworzą układ ortogonalny?
2. Udowodnić, że wektory układu ortonormalnego są liniowo niezależne. Podać warunki konieczne i dostateczne na to, żeby tę własność miały wektory układu ortogonalnego.
3. Udowodnić, że macierz kwadratowa o elementach $\langle x_i, y_j\rangle$ jest unitarna, jeśli układy $\{x_1, x_2, \ldots, x_n\}$ i $\{y_1, y_2, \ldots, y_n\}$ są bazami ortonormalnymi przestrzeni $\mathbb{C}^n$.
4. Udowodnić, że macierz kwadratowa jest unitarna wtedy i tylko wtedy, gdy jej wiersze tworzą układ ortonormalny.
5. Niech układ $\{u_1, u_2, \ldots, u_n\}$ będzie bazą ortonormalną przestrzeni unitarnej X. Wykazać, że dla $x, y \in X$ jest:
   (**a**) $\|x\|_2^2 = \sum_{i=1}^n |\langle x, u_i\rangle|^2$, (**b**) $\langle x, y\rangle = \sum_{i=1}^n \langle x, u_i\rangle\overline{\langle y, u_i\rangle}$.
6. Dla bazy ortonormalnej $\{u_1, u_2, \ldots, u_n\}$ podprzestrzeni U przestrzeni unitarnej X określamy operator $P\colon X \to U$ wzorem $Px := \sum_{i=1}^n \langle x, u_i\rangle u_i$; jest to *rzut ortogonalny* z X na U. Wykazać, że: (**a**) operator P jest liniowy, (**b**) P jest idempotentny, tj. $P^2 = P$, (**c**) $Px = x$, jeśli $x \in U$, (**d**) $\|Px\|_2 \leqslant \|x\|_2$ dla każdego $x \in X$.
7. Podać przykład wektorów x i y takich, że $\|x + y\|_2^2 = \|x\|_2^2 + \|y\|_2^2$, chociaż $\langle x, y\rangle \neq 0$.
8. Udowodnić, że jeśli $x \neq y$ i $\langle x, y\rangle$ jest rzeczywiste, to macierz unitarna U taka, że $Ux = y$, jest równa $U := I - vu^{\mathsf{H}}$, gdzie $v := x - y$ i $u := 2v/\|v\|_2^2$. Dlaczego ta informacja daje lepszy sposób konstrukcji przekształceń Householdera? Założyć, że $\|x\|_2 = \|y\|_2$.

9. Macierz A taką, że $A^2 = I$, nazywamy *inwolucją*. Znaleźć warunki konieczne i dostateczne, które powinny spełniać wektory u i v, żeby macierz $I - uv^{\mathsf{H}}$ była inwolucją.

10. (cd.). Niech będzie $v^{\mathsf{H}}v = 2$. Wykazać, że inwolucją jest macierz blokowa

$$\begin{bmatrix} I & 0 \\ 0 & I - vv^{\mathsf{H}} \end{bmatrix}.$$

11. Czy iloczyn dwóch inwolucji jest inwolucją?

12. Zastosować algorytm Grama-Schmidta do ciągu wektorów: (3, 4, 0), (1, 1, 1), (1, 2, 0).

13. Poniższy przykład (Noble i Daniel [1988]) pokazuje wpływ błędów zaokrągleń na wyniki niezmodyfikowanego algorytmu Grama-Schmidta. Niech ε będzie na tyle małą liczbą dodatnią, że chociaż $1 + \varepsilon$ i $3 + 2\varepsilon$ są liczbami maszynowymi, to $3 + 2\varepsilon + \varepsilon^2$ jest pamiętane jako $3 + 2\varepsilon$. Wspomniany algorytm stosujemy do wektorów

$$x_1 := \mathrm{fl}(1 + \varepsilon, 1, 1), \quad x_2 := (1, 1 + \varepsilon, 1), \quad x_3 := \mathrm{fl}(1, 1, 1 + \varepsilon).$$

Sprawdzić, że obliczenia w arytmetyce zmiennopozycyjnej dają wektory

$$u_1 := \mathrm{fl}(1/\sqrt{3 + 2\varepsilon})x_1, \; u_2 := \mathrm{fl}(1/\sqrt{2})(-1, 1, 0), \; u_3 := \mathrm{fl}(1/\sqrt{2})(-1, 0, 1),$$

które powinny stanowić bazę ortonormalną. Jest jednak $\langle u_2, u_3 \rangle = \frac{1}{2}$.

14. Udowodnić, że jeśli A jest macierzą $m \times n$ i rzędu n, to iloczyn $A^{\mathsf{H}}A$ jest: (**a**) nieosobliwy, (**b**) hermitowski i dodatnio określony.

15. Jaka wartość parametru t daje minimum normy $\|u - tx\|_2$ (u, x – dane wektory)? Odpowiedź ma być poprawna w przypadku zespolonym.

16. Niech A będzie macierzą rozmiaru $m \times n$, a b wektorem o m składowych i niech $\alpha > 0$. Udowodnić, że funkcja $F(x) := \|Ax - b\|_2^2 + \alpha\|x\|_2^2$ osiąga minimum dla x spełniającego równanie $(A^{\mathsf{T}}A + \alpha I)x = A^{\mathsf{T}}b$ i że wtedy $F(x + h) = F(x) + (ah)^{\mathsf{T}}Ah + \alpha h^{\mathsf{T}}h$.

17. Udowodnić, że rozwiązując zadanie najmniejszych kwadratów dla układu $Ax = b$, można zastąpić równania normalne przez $CAx = Cb$, gdzie C jest dowolną macierzą rozmiaru $n \times m$ taką, że dla pewnej macierzy nieosobliwej F jest $C = FA^{\mathsf{T}}$.

18. Niech A będzie macierzą rozmiaru $m \times n$ i rzędu n i niech $b \in \mathbb{R}^m$. Pokazać, że dla każdego $\lambda \geqslant 0$ zbiór $K_\lambda = \{x \in \mathbb{R}^n \colon \|Ax - b\| \leqslant \lambda\}$ jest domknięty i ograniczony. Normy w $\mathbb{R}^m$ i $\mathbb{R}^n$ mogą tu być dowolne.

19. (cd.). Udowodnić, że jeśli $\lambda = 2\|b\|_2$, to

$$\inf_{x \in \mathbb{R}^n} \|Ax - b\|_2 = \inf_{x \in K_\lambda} \|Ax - b\|_2.$$

20. (cd.). Wykazać, że rozwiązanie w sensie najmniejszych kwadratów układu $Ax = b$ spełnia nierówność $\|x\|_2 \leqslant 2\|b\|_2\|B\|_2$, gdzie B jest dowolną lewą odwrotnością macierzy A.

21. Niech A będzie macierzą rozmiaru $m \times n$ i niech $b \in \mathbb{R}^m$. Udowodnić, że w $\mathbb{R}^n$ istnieje wektor x, dla którego $\|Ax - b\|$ (gdzie $\|\cdot\|$ jest dowolną normą w $\mathbb{R}^m$) osiąga minimum.

22. (cd.). Udowodnić, że równanie $A^T Ax = A^T b$ ma rozwiązanie.

23. Znaleźć rozwiązanie w sensie najmniejszych kwadratów układu

$$\begin{bmatrix} 3 & 2 \\ 2 & 3 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix}.$$

24. (cd.). Czy można, nie rozwiązując tego układu, potwierdzić przypuszczenie, że rozwiązaniem jest $(x, y) = (\frac{29}{21}, -\frac{2}{3})$?

25. Niech A będzie macierzą rozmiaru $(n+1) \times n$ i rzędu n, a z – wektorem ortogonalnym względem jej kolumn. Wykazać, że układ $Ax + \lambda z = b$ ma rozwiązanie x, λ oraz że to x jest rozwiązaniem w sensie najmniejszych kwadratów układu $Ax = b$.

26. Jakie operacje elementarne na wierszach w układzie $Ax = b$ nie naruszają wszystkich jego rozwiązań w sensie najmniejszych kwadratów?

27. Dla macierzy A z (5.3.5) znaleźć $\varkappa_\infty(A^H A)$. Co się dzieje, gdy $\varepsilon \to 0$?

28. Za pomocą algorytmu Householdera znaleźć rozkład QR macierzy

$$\begin{bmatrix} 0 & -4 \\ 0 & 0 \\ -5 & -2 \end{bmatrix}, \qquad \begin{bmatrix} 3 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}.$$

29. Niech A będzie macierzą rozmiaru $m \times n$, gdzie $m > n$, mającą rozkład $A = QR$. Udowodnić, że $A = Q'R'$, gdzie Q' powstaje z Q przez skreślenie $m - n$ ostatnich kolumn, a R' powstaje z R przez skreślenie tyluż ostatnich wierszy.

30. Niech rozkład macierzy A rozmiaru $m \times n$ za pomocą przekształceń Householdera daje iloczyn QR. Niech A_j będzie j-tą kolumną tej macierzy; podobny sens mają symbole Q_j i R_j. Sprawdzić, że $A_j = \sum_{k=1}^{j} r_{kj} Q_k$ i $r_{kj} = \langle A_j, Q_k \rangle$. Udowodnić, że stąd można otrzymać Q_1 i r_{11}. Pokazać, jak wyznacza się Q_j i R_j, gdy są już znane Q_k i R_k dla $k < j$.

ZADANIA KOMPUTEROWE 5.3

K1. Napisać procedury realizujące algorytm Grama-Schmidta – zwykły i zmodyfikowany. Obie dla danej macierzy A rozmiaru $m \times n$ i rzędu n mają dać macierze B (o kolumnach ortonormalnych) i T. Sprawdzić na przykładzie poniższych macierzy $A = (a_{ij})$ rozmiaru 20×10, który wariant daje iloczyn $B^T B$ bliższy macierzy jednostkowej. Dalsze informacje o takich testach podaje Rice [1966].

(**a**) Elementami macierzy są wartości zmiennej losowej o rozkładzie jednostajnym w przedziale $[0, 1]$.

(**b**) $a_{ij} := [(2i - 21)/19]^{j-1}$ $\quad (1 \leqslant i \leqslant 20,\ 1 \leqslant j \leqslant 10)$.

K2. Napisać program rozwiązujący układ równań $Ax = b$ w sensie najmniejszych kwadratów. Powinien on wywoływać procedurę realizującą zmodyfikowany algorytm Grama-Schmidta.

K3. Napisać i sprawdzić program rozkładu QR za pomocą przekształceń Householdera.

5.4. Rozkład względem wartości szczególnych i pseudoodwrotność

W wielu zastosowaniach występuje rozkład macierzy według wartości szczególnych. Zaczynamy od twierdzenia, które orzeka, że taki rozkład istnieje i określa jego postać.

Twierdzenie 5.4.1. *Dowolną macierz A rozmiaru $m \times n$, o elementach zespolonych, można wyrazić w postaci $A = PDQ$, gdzie P jest macierzą unitarną stopnia m, D jest macierzą przekątniową rozmiaru $m \times n$, a Q - macierzą unitarną stopnia n.*

Dowód. Iloczyn $A^{\mathsf{H}}A$ jest macierzą hermitowską stopnia n. Jest ona również dodatnio półokreślona (zob. podrozdz. 4.6), gdyż dla każdego $x \in \mathbb{C}^n$

$$x^{\mathsf{H}}(A^{\mathsf{H}}A)x = (Ax)^{\mathsf{H}}(Ax) \geqslant 0.$$

Stąd wynika, że wartości własne tej macierzy są nieujemne. Dowodzi się również, że dokładnie r z nich, gdzie $r := \operatorname{rank} A$, jest dodatnich (oczywiście $r \leqslant \min\{m, n\}$). Oznaczamy je $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, przy czym $\sigma_i^2 > 0$ dla $i \leqslant r$ i $\sigma_i^2 = 0$ dla $i > r$. Niech wektory własne u_i macierzy $A^{\mathsf{H}}A$, a więc takie, że $A^{\mathsf{H}}Au_i = \sigma_i^2 u_i$, tworzą układ ortonormalny. Wtedy

$$\|Au_i\|_2^2 = u_i^{\mathsf{H}}A^{\mathsf{H}}Au_i = u_i^{\mathsf{H}}\sigma_i^2 u_i = \sigma_i^2,$$

skąd wynika, że $Au_i = 0$ dla $i > r$.

Tworzymy macierz Q stopnia n o wierszach $u_1^{\mathsf{H}}, u_2^{\mathsf{H}}, \ldots, u_n^{\mathsf{H}}$. Definiujemy też wektory

$$v_i := \sigma_i^{-1}Au_i \quad (1 \leqslant i \leqslant r).$$

Ich układ jest ortonormalny, gdyż dla $1 \leqslant i, j \leqslant r$ jest

$$\begin{aligned} v_i^{\mathsf{H}}v_j &= \sigma_i^{-1}(Au_i)^{\mathsf{H}}\sigma_j^{-1}(Au_j) = \\ &= (\sigma_i\sigma_j)^{-1}(u_i^{\mathsf{H}}A^{\mathsf{H}}Au_j) = (\sigma_i\sigma_j)^{-1}(u_i^{\mathsf{H}}\sigma_j^2 u_j) = \delta_{ij}. \end{aligned}$$

Jeśli $r < m$, to wybieramy jeszcze dodatkowe wektory v_i tak, żeby układ $\{v_1, v_2, \ldots, v_m\}$ był bazą ortonormalną przestrzeni $\mathbb{C}^m$. Niech P będzie macierzą stopnia m, której kolumnami są te wektory. Niech wreszcie D będzie macierzą rozmiaru $m \times n$, mającą liczby $\sigma_1, \sigma_2, \ldots, \sigma_r$ na początku głównej przekątnej i zera na innych pozycjach. Wtedy $A = PDQ$. Żeby to udowodnić, wykazujemy, że $P^{\mathsf{H}}AQ^{\mathsf{H}} = D$. Jest $(P^{\mathsf{H}}AQ^{\mathsf{H}})_{ij} = v_i^{\mathsf{H}}Au_j$. Ten iloczyn jest dla $j \leqslant r$ równy $v_i^{\mathsf{H}}\sigma_j v_j = \sigma_j \delta_{ij}$, a dla $j > r$ znika. ■

Określone w dowodzie liczby $\sigma_1, \sigma_2, \ldots, \sigma_n$, czyli pierwiastki kwadratowe (nieujemne) wartości własnych iloczynu $A^{\mathsf{H}}A$, nazywamy *wartościami szczególnymi* macierzy A. Rozkład $A = PDQ$ jest *rozkładem względem wartości szczególnych*. Zauważmy, że porządek wartości $\sigma_1, \sigma_2, \ldots, \sigma_r$ jest dowolny. Wektory $v_{r+1}, v_{r+2}, \ldots, v_m$ też nie są określone jednoznacznie; m.in. z tych powodów na ogół istnieje wiele określonych tu rozkładów.

Przykład 5.4.2. Znaleźć rozkład względem wartości szczególnych macierzy

$$A := \begin{bmatrix} 7 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Rozwiązanie. Ponieważ

$$A^{\mathsf{H}}A = \begin{bmatrix} 49 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

więc $\sigma_1 = 7$, $\sigma_2 = 3$ (można te liczby ustawić w odwrotnym porządku) i $\sigma_3 = \sigma_4 = 0$. Wzorując się na dowodzie tw. 5.4.1, tworzymy macierze

$$P := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad D := \begin{bmatrix} 7 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad Q := \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Tu jest oczywiste, że $A = PDQ$. Można sprawdzić, że inny taki rozkład jest następujący:

$$\begin{bmatrix} 7 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

■

Przykład 5.4.3. Znaleźć rozkład względem wartości szczególnych macierzy

$$A = \begin{bmatrix} 0 & -1.6 & 0.6 \\ 0 & 1.2 & 0.8 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Rozwiązanie. Także tutaj wzorujemy się na dowodzie tw. 5.4.1. Ponieważ

$$A^{\mathsf{H}}A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

więc $\sigma_1 = 1$, $\sigma_2 = 2$ i $\sigma_3 = 0$. Wybieramy wektory własne (nie są określone jednoznacznie) i tworzymy macierz

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

Wtedy

$$v_1 = Au_1 = (0.6, 0.8, 0, 0), \quad v_2 = \tfrac{1}{2}Au_2 = (0.8, -0.6, 0, 0).$$

Mamy też pewną swobodę w wyborze v_3 i v_4. Najprostszy wariant to

$$v_3 = (0, 0, 1, 0), \quad v_4 = (0, 0, 0, 1).$$

Stąd wynika jeden z możliwych rozkładów względem wartości szczególnych:

$$\begin{bmatrix} 0 & -1.6 & 0.6 \\ 0 & 1.2 & 0.8 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.6 & 0.8 & 0 & 0 \\ 0.8 & -0.6 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

■

Pseudoodwrotność

Dla macierzy D rozmiaru $m \times n$ takiej, że

$$(D)_{ij} := \begin{cases} \sigma_i & \text{dla } i = j \leqslant r \\ 0 & \text{w przeciwnym razie} \end{cases} \tag{5.4.1}$$

(gdzie każde σ_i jest dodatnie) definiujemy jej *pseudoodwrotność* jako macierz D^+ rozmiaru $n \times m$ taką, że

$$(D^+)_{ij} := \begin{cases} \sigma_i^{-1} & \text{dla } i = j \leqslant r \\ 0 & \text{w przeciwnym razie.} \end{cases}$$

Pseudoodwrotność A^+ dowolnej macierzy A wynika z jej rozkładu $A = PDQ$ względem wartości szczególnych; jest mianowicie

$$A^+ := Q^{\mathsf{H}} D^+ P^{\mathsf{H}}.$$

Przekonamy się, że jeśli nawet są możliwe różne takie rozkłady, to pseudoodwrotność jest określona jednoznacznie.

Przykład 5.4.4. Znaleźć pseudoodwrotność macierzy A z przykładu 5.4.2.

Rozwiązanie. Ponieważ (w pierwszym wariancie) macierze P i Q są jednostkowe, więc jest oczywiste, że

$$A^+ = \begin{bmatrix} 7^{-1} & 0 & 0 \\ 0 & 3^{-1} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

■

Przykład 5.4.5. Znaleźć pseudoodwrotność macierzy A z przykładu 5.4.3.

Rozwiązanie. Wyniki przykł. 5.4.3 dają następującą macierz A^+:

$$A^+ = \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.6 & 0.8 & 0 & 0 \\ 0.8 & -0.6 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} =$$

$$= \begin{bmatrix} 0 & 0 & 0 & 0 \\ -0.4 & 0.3 & 0 & 0 \\ 0.6 & 0.8 & 0 & 0 \end{bmatrix}.$$

■

Układy sprzeczne lub mające wiele rozwiązań

Główne zastosowanie pseudoodwrotności jest związane z układami równań, które są sprzeczne (nie mają rozwiązań) lub mają wiele rozwiązań. Rozważmy więc układ $Ax = b$, gdzie macierz A ma rozmiar $m \times n$, $x \in \mathbb{C}^n$ i $b \in \mathbb{C}^m$. *Rozwiązanie minimalne*[13] tego układu określamy tak:

1. jeśli układ jest niesprzeczny i ma jedyne rozwiązanie – jako to rozwiązanie;
2. jeśli układ jest niesprzeczny i ma wiele rozwiązań – jako to z nich, które ma najmniejszą normę euklidesową;
3. jeśli układ jest sprzeczny i ma jedyne rozwiązanie w sensie najmniejszych kwadratów – jako to rozwiązanie;
4. jeśli układ jest sprzeczny i ma wiele rozwiązań w sensie najmniejszych kwadratów – jako to z nich, które ma najmniejszą normę euklidesową.

Te cztery przypadki można uwzględnić razem przyjmując, że dla

$$\rho := \min_{x \in \mathbb{C}^n} \|Ax - b\|_2$$

rozwiązaniem minimalnym jest wektor x ze zbioru $K = \{x\colon \|Ax - b\|_2 = \rho\}$ mający najmniejszą normę euklidesową. W przypadkach **1** i **2** jest $\rho = 0$, w pozostałych $\rho > 0$.

Twierdzenie 5.4.6. *Rozwiązanie minimalne układu $Ax = b$ jest równe*

$$x = A^+ b.$$

Dowód. Niech $A = PDQ$ określa rozkład macierzy A względem jej wartości szczególnych. Niech będzie

$$c = P^{\mathsf{H}} b, \quad y = Qx.$$

Przekształcenie $y = Qx$ jest suriektywne, tj. przekształca $\mathbb{C}^n$ na $\mathbb{C}^n$. Dlatego

$$\rho = \min_x \|Ax - b\|_2 = \min_x \|PDQx - b\|_2 = \min_x \|P^{\mathsf{H}}(PDQx - b)\|_2 =$$
$$= \min_x \|DQx - P^{\mathsf{H}} b\|_2 = \min_y \|Dy - c\|_2.$$

Dzięki specjalnej postaci macierzy D jest

$$\|Dy - c\|_2^2 = \sum_{i=1}^{r} (\sigma_i y_i - c_i)^2 + \sum_{i=r+1}^{m} c_i^2.$$

[13] Albo – *najkrótsze* lub *normalne* (*przyp. tłum.*).

Ta wielkość osiąga minimum, gdy $y_i = c_i/\sigma_i$ dla $1 \leqslant i \leqslant r$; pozostałe y_i mogą być dowolne. Wobec tego

$$\rho = \Big(\sum_{i=r+1}^{m} c_i^2 \Big)^{1/2}.$$

Wśród wszystkich wektorów y, dla których jest osiągnięte minimum ρ, najmniejszą normę ma taki, że $y_i = 0$ dla $i > r$. Jest on równy $y = D^+c$, a rozwiązaniem minimalnym jest wektor

$$x = Q^{\mathsf{H}}y = Q^{\mathsf{H}}D^+c = Q^{\mathsf{H}}D^+P^{\mathsf{H}}b = A^+b. \qquad \blacksquare$$

Pseudoodwrotność jest tym dla układów sprzecznych lub niedookreślonych, czym zwykła odwrotność dla układów z macierzą nieosobliwą. Dodajmy, że rozwiązanie minimalne dowolnego układu $Ax = b$ jest jedyne, gdyż w zbiorze wypukłym K tylko jeden element ma najmniejszą normę.

Przykład 5.4.7. Znaleźć rozwiązanie minimalne układu $Ax = b$ dla macierzy z przykładu 5.4.3 i wektora $b := (5, 7, 3, -2)$.

Rozwiązanie. Korzystając z pseudoodwrotności A^+ podanej w przykł. 5.4.5, stwierdzamy, że rozwiązaniem minimalnym jest wektor

$$A^+b = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -0.4 & 0.3 & 0 & 0 \\ 0.6 & 0.8 & 0 & 0 \end{bmatrix} \begin{bmatrix} 5 \\ 7 \\ 3 \\ -2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.1 \\ 8.6 \end{bmatrix}. \qquad \blacksquare$$

Własności Penrose'a

Pseudoodwrotność ma tylko niektóre z własności zwykłej odwrotności. Nie można m.in. twierdzić, że $A^+A = I$ dla $n > m$, gdyż rząd każdej z macierzy A^+, A, A^+A nie przewyższa m, a macierz jednostkowa I ma tu stopień i rząd n. Pewne równości, w tym $AA^+A = A$, są jednak ogólnie prawdziwe. W poniższym twierdzeniu podano cztery związki, wyrażające tzw. *własności Penrose'a*[14] dowolnej macierzy; zob. Penrose [1955].

Twierdzenie 5.4.8 (Penrose). *Dla dowolnej macierzy A istnieje co najwyżej jedna macierz X o następujących własnościach:*

1. $AXA = A$. **2.** $XAX = X$. **3.** $(AX)^{\mathsf{H}} = AX$. **4.** $(XA)^{\mathsf{H}} = XA$.

[14] Golub i van Loan [1989] oraz Kiełbasiński i Schwetlick [1992*] używają nazwy *własności Moore'a-Penrose'a* (*przyp. tłum.*).

Dowód. Przypuśćmy, że macierze X i Y mają powyższe własności. Stosując je systematycznie, otrzymujemy następujący ciąg przekształceń:

	Własność
$X = XAX =$	**2**
$= XAYAX =$	**1**
$= XAYAYAYAX =$	**1**
$= (XA)^{\mathsf{H}}(YA)^{\mathsf{H}}Y(AY)^{\mathsf{H}}(AX)^{\mathsf{H}} =$	**4, 3**
$= A^{\mathsf{H}}X^{\mathsf{H}}A^{\mathsf{H}}Y^{\mathsf{H}}YY^{\mathsf{H}}A^{\mathsf{H}}X^{\mathsf{H}}A^{\mathsf{H}} =$	
$= (AXA)^{\mathsf{H}}Y^{\mathsf{H}}YY^{\mathsf{H}}(AXA)^{\mathsf{H}} =$	
$= A^{\mathsf{H}}Y^{\mathsf{H}}YY^{\mathsf{H}}A^{\mathsf{H}} =$	**1**
$= (YA)^{\mathsf{H}}Y(AY)^{\mathsf{H}} =$	
$= YAYAY =$	**4, 3**
$= YAY =$	**2**
$= Y,$	**2**

czyli $X = Y$. ■

Twierdzenie 5.4.9. *Pseudoodwrotność dowolnej macierzy ma cztery własności Penrose'a, a zatem jest określona jednoznacznie.*

Dowód. Niech iloczyn PDQ będzie rozkładem macierzy A względem wartości szczególnych. Wtedy $A^+ = Q^{\mathsf{H}}D^+P^{\mathsf{H}}$. Macierz D ma postać określoną w (5.4.1), a stąd wynika, że $DD^+D = D$. Istotnie, zauważmy, że

$$(DD^+D)_{ij} = \sum_{\nu=1}^{n}(D)_{i\nu}\sum_{\mu=1}^{m}(D^+)_{\nu\mu}(D)_{\mu j}.$$

Prawa strona nie znika tylko dla $i \leqslant r$ i $j \leqslant r$, więc upraszcza się do postaci

$$\sum_{\nu=1}^{r}(D)_{i\nu}\sum_{\mu=1}^{r}(D^+)_{i\mu}(D)_{\mu j} = \sigma_i\sum_{\mu=1}^{r}(D^+)_{i\mu}(D)_{\mu j} = \sigma_i\sigma_i^{-1}(D)_{ij} = (D)_{ij}.$$

Podobnie rozumując, sprawdzamy, że macierz D^+ ma względem D trzy pozostałe własności Penrose'a. Stąd wynika prosty sposób ich sprawdzenia dla A^+. Tak np. sprawdzamy pierwszą własność:

$$AA^+A = PDQQ^{\mathsf{H}}D^+P^{\mathsf{H}}PDQ = PDD^+DQ = PDQ = A.$$

Pozostałe własności są tematem zad. 22. ■

Twierdzenie 5.4.10. *Niech dla macierzy A obowiązują oznaczenia z dowodu tw. 5.4.1. Wtedy:*

1. $\operatorname{rank} A = r$.
2. *Układ* $\{u_{r+1}, u_{r+2}, \ldots, u_n\}$ *jest bazą ortonormalną jądra macierzy* A.
3. *Układ* $\{v_1, v_2, \ldots, v_r\}$ *jest bazą ortonormalną przestrzeni wartości macierzy* A.
4. $\|A\|_2 = \max_{1 \leqslant i \leqslant n} |\sigma_i|$.

Dowód. Macierze P i Q są nieosobliwe, więc $\operatorname{rank} A = \operatorname{rank} D = r$. Z dowodu tw. 5.4.1 wynika, że dla $r < i \leqslant n$ jest $Au_i = 0$. Jądro macierzy A jest $(n-r)$-wymiarowe, a jego bazą jest układ $\{u_{r+1}, u_{r+2}, \ldots, u_n\}$. Natomiast przestrzeń wartości tej macierzy jest r-wymiarowa, a jego bazą ortonormalną jest układ $\{v_1, v_2, \ldots, v_r\}$, gdzie $v_i := \sigma_i^{-1} A u_i$[15]. Przekształcenia przestrzeni $\mathbb{C}^m$ i $\mathbb{C}^n$ odpowiednio za pomocą macierzy unitarnych P i Q zachowują normę (zad. 5.2.9), więc

$$\begin{aligned}
\|A\|_2 &= \sup\{\|PDQx\|_2 \colon \|x\|_2 = 1\} = \sup\{\|Dy\|_2 : \|y\|_2 = 1\} = \\
&= \sup\{\sqrt{\sigma_1^2 y_1^2 + \ldots + \sigma_n^2 y_n^2} \colon \|y\|_2 = 1\} = \\
&= \Big[\sup\Big\{\sigma_1^2 y_1^2 + \ldots + \sigma_n^2 y_n^2 \colon \sum_{i=1}^{n} y_i^2 = 1\Big\}\Big]^{1/2} = \\
&= \big[\max_{1 \leqslant i \leqslant n} \sigma_i^2\big]^{1/2} = \max_{1 \leqslant i \leqslant n} |\sigma_i|. \qquad \blacksquare
\end{aligned}$$

Poznamy teraz oszczędny wariant rozkładu względem wartości szczególnych:

Twierdzenie 5.4.11. *Jeśli* A *jest macierzą* $m \times n$ *rzędu* r, *gdzie* $m \geqslant n \geqslant r$, *to można ją wyrazić w postaci* $A = VSU$, *gdzie* V *jest macierzą* $m \times r$ *o kolumnach ortonormalnych,* S *– macierzą przekątniową nieosobliwą stopnia* r, *a* U *– macierzą* $r \times n$ *o wierszach ortonormalnych.*

Twierdzenie 5.4.12. *Jeśli* L *jest przekształceniem liniowym z* $\mathbb{C}^m$ *w* $\mathbb{C}^n$, *to istnieją bazy ortonormalne* $\{u_1, u_2, \ldots, u_m\}$ *dla* $\mathbb{C}^m$ *i* $\{v_1, v_2, \ldots, v_n\}$ *dla* $\mathbb{C}^n$ *takie, że*

$$Lu_i = \begin{cases} \sigma_i v_i, & \text{jeśli } 1 \leqslant i \leqslant \min\{m,n\} \\ 0, & \text{jeśli } \min\{m,n\} < i \leqslant m. \end{cases}$$

[15] Jądro i przestrzeń wartości są określone w przypisie nr 7 w podrozdz. 4.1 (*przyp. tłum.*).

Dowód. Oznaczmy bazy złożone z wektorów jednostkowych symbolami $\{e_1, e_2, \ldots, e_m\}$ dla $\mathbb{C}^m$ i $\{e'_1, e'_2, \ldots, e'_n\}$ dla $\mathbb{C}^n$. Niech $A := (a_{ij})$ będzie macierzą $m \times n$ określoną wzorem $Le_i = \sum_{j=1}^{n} a_{ij} e'_j$ $(1 \leqslant i \leqslant m)$, a PDQ jej rozkładem względem wartości szczególnych. Oznaczmy wiersze macierzy P^{H} symbolami $u_1, u_2, \ldots, u_m$, a kolumny Q^{H} – symbolami $v_1, v_2, \ldots, v_m$. Sprawdzimy, że Lu_i wyraża się tak, jak podano w twierdzeniu. W tym celu dla $k = 1, 2, \ldots, m$ obliczamy

$$\begin{aligned} Lu_k &= L\Big(\sum_{i=1}^{m} \langle u_k, e_i\rangle\Big) = \sum_{i=1}^{m} \langle u_k, e_i\rangle Le_i = \\ &= \sum_{i=1}^{m} (P^{\mathsf{H}})_{ki} \sum_{j=1}^{n} a_{ij} e'_j = \sum_{j=1}^{n} \sum_{i=1}^{m} (P^{\mathsf{H}})_{ki} a_{ij} e'_j = \\ &= \sum_{j=1}^{n} (P^{\mathsf{H}} A)_{kj} e'_j = \sum_{j=1}^{n} (P^{\mathsf{H}} A)_{kj} \sum_{s=1}^{n} \langle e'_j, v_s\rangle v_s = \\ &= \sum_{s=1}^{n} \sum_{j=1}^{n} (P^{\mathsf{H}} A)_{kj} (Q^{\mathsf{H}})_{js} v_s = \sum_{s=1}^{n} (P^{\mathsf{H}} A Q^{\mathsf{H}})_{ks} v_s. \end{aligned}$$

Ponieważ $A = PDQ$, więc $P^{\mathsf{H}} A Q^{\mathsf{H}} = D$. Wystarczy zauważyć, że D jest macierzą $m \times n$ taką, że $(D)_{ij} = \sigma_i$ dla $i = j \leqslant \min\{m, n\}$ i $(D)_{ij} = 0$ w przeciwnym razie. ■

Zadania 5.4

1. Niech A będzie macierzą $m \times n$ rzędu r, gdzie $m \geqslant n \geqslant r$, o rozkładzie $A = PDQ$ względem wartości szczególnych. Udowodnić, że układ $Ax = b$ jest niesprzeczny wtedy i tylko wtedy, gdy $(P^{\mathsf{H}} b)_i = 0$ dla $r < i \leqslant m$.
2. Macierze A i B są *unitarnie równoważne*, gdy istnieją takie macierze unitarne U i V, że $A = UBV$. Udowodnić, że takie A i B mają identyczne układy wartości szczególnych.
3. Udowodnić, że jeśli macierz stopnia n ma wartości szczególne $\sigma_1, \sigma_2, \ldots, \sigma_n$, to jej wyznacznik jest równy $\pm\sigma_1\sigma_2 \ldots \sigma_n$.
4. Udowodnić, że jeśli macierz kwadratowa A ma rozkład $A = PDQ$ względem wartości szczególnych, to jej wielomian charakterystyczny jest równy $\pm \det(D - \lambda P^{\mathsf{H}} Q^{\mathsf{H}})$.
5. Udowodnić, że suma kwadratów elementów iloczynu $v_i u_i^{\mathsf{H}}$ jest równa 1 (oznaczenia jak w dowodzie tw. 5.4.1).
6. Odwołując się do dowodu tw. 5.4.1, wykazać, że

$$A = \sum_{j=1}^{r} \sigma_j v_j u_j^{\mathsf{H}}, \qquad A^{+} = \sum_{j=1}^{r} \sigma_j^{-1} u_j v_j^{\mathsf{H}}.$$

7. Korzystając z wyniku poprzedniego zadania, wykazać, że jeśli znamy rozkład macierzy A względem wartości szczególnych, to Ax można obliczyć kosztem $(n+m+1)r$ mnożeń i $(n+m-1)r-m$ dodawań. Porównać to z kosztem bezpośredniego mnożenia (nm mnożeń i $(n-1)m$ dodawań).

8. Obcięcie pierwszej sumy z zad. 6 do k składników daje pewne przybliżenie macierzy A. Wykazać, że $\left\|A-\sum_{j=1}^{k}\sigma_j v_j u_j^{\mathsf{H}}\right\|_2=\sigma_{k+1}$.

9. Zakładając, że $A=UDV$, gdzie U jest macierzą unitarną stopnia m, V macierzą unitarną stopnia n, a D macierzą przekątniową $m\times n$, wykazać, że liczby $|(D)_{ii}|^2$ $(1\leqslant i\leqslant n)$ są wartościami własnymi iloczynu $A^{\mathsf{H}}A$.

10. Znaleźć rozkład względem wartości szczególnych macierzy

$$\begin{bmatrix}4&0&0\\0&0&0\\0&0&7\\0&0&0\end{bmatrix},\qquad \begin{bmatrix}2&1\end{bmatrix},\qquad \begin{bmatrix}5\\-4\end{bmatrix}.$$

11. Znaleźć pseudoodwrotność A^+ zakładając, że: (**a**) macierz AA^{H} jest nieosobliwa, (**b**) $A^{\mathsf{H}}A=I$, (**c**) A jest hermitowska i taka, że $A^2=A$.

12. Udowodnić, że jeśli macierz jest hermitowska, to jej pseudoodwrotność ma tę samą własność.

13. Udowodnić następujące własności pseudoodwrotności: (**a**) $(A^+)^+=A$, (**b**) $(A^+)^{\mathsf{H}}=(A^{\mathsf{H}})^+$, (**c**) $(AA^{\mathsf{H}})^+=(A^+)^{\mathsf{H}}A^+$, (**d**) $A^+=A^{\mathsf{H}}(AA^{\mathsf{H}})^+$.

14. Udowodnić, że $A^+=(A^{\mathsf{H}}A)^{-1}A^{\mathsf{H}}$ dla macierzy A rozmiaru $m\times n$ i rzędu n.

15. Wykazać, że jeśli macierze A i B są pełnego rzędu, to $(AB)^+=B^+A^+$ (macierz $m\times n$ tak się nazywa, gdy jej rząd jest równy $\min\{m,n\}$).

16. Znaleźć przykład świadczący o tym, że na ogół $(AB)^+\neq B^+A^+$.

17. Udowodnić, że pseudoodwrotność macierzy przekątniowej $m\times n$ jest macierzą przekątniową $n\times m$.

18. Wykazać, że z symetrii macierzy A wynika taka sama własność dla A^+.

19. Znaleźć pseudoodwrotność: (**a**) dowolnej macierzy $m\times 1$, (**b**) dowolnej macierzy $1\times n$, (**c**) iloczynu uv^{H}, gdzie $u,v\in\mathbb{C}^n$, (**d**) macierzy $m\times n$, której wszystkie elementy są równe 1.

20. Korzystając z tw. 5.4.8, udowodnić, że jeśli B jest macierzą $m\times r$, C jest macierzą $r\times n$ i jeśli rząd każdej z macierzy B, C i BC jest równy r, to $(BC)^+=C^{\mathsf{H}}(CC^{\mathsf{H}})^{-1}(B^{\mathsf{H}}B)^{-1}B^{\mathsf{H}}$.

21. Znaleźć rozwiązania minimalne następujących układów równań:

(**a**) $x_1+x_2=b_1$

(**b**) $x_1=b_1$, $x_1=b_2$, $x_1=b_3$

(**c**) $4x_1=b_1$, $0x_1=b_2$, $7x_3=b_3$, $0x_2=b_4$

22. Udowodnić, że pseudoodwrotności D^+ i A^+ mają własności **2–4** Penrose'a (zob. tw. 5.4.9).

23. Niech A będzie macierzą $m\times n$, a X macierzą $n\times m$ mającą własności Penrose'a względem A. Udowodnić, że rozwiązaniem minimalnym układu $Ax=b$ jest Xb.

24. Znaleźć zależności wiążące wartości własne i wartości szczególne macierzy hermitowskiej.

25. Udowodnić, że dla macierzy hermitowskiej i dodatnio półokreślonej wartości szczególne są identyczne z wartościami własnymi.

26. Jaki jest rozkład macierzy hermitowskiej i dodatnio określonej względem wartości szczególnych?

27. Udowodnić, że pseudoodwrotność jest funkcją nieciągłą elementów macierzy. Wskazówka: Zbadać macierz

$$\begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \\ 0 & 0 \end{bmatrix}.$$

28. Udowodnić tw. 5.4.11.

ZADANIA KOMPUTEROWE 5.4

K1. Napisać program, który korzystając z rozkładu względem wartości szczególnych i pseudoodwrotności, oblicza rozwiązanie minimalne układu $Ax = b$.

5.5. Metoda QR obliczania wartości własnych

Twierdzenie Schura 5.2.3 gwarantuje, że dowolna macierz kwadratowa jest unitarnie podobna do macierzy trójkątnej:

$$UAU^{\mathsf{H}} = T \tag{5.5.1}$$

(U jest unitarna, T – trójkątna). Istotne jest to, że wartości własne macierzy A i T są identyczne i że w T są one po prostu elementami jej przekątnej. Istnienie rozkładu (5.5.1) nie oznacza jednak, że łatwo go znaleźć. Konstrukcja macierzy U musi być tak trudna, jak obliczenie wszystkich (na ogół zespolonych) pierwiastków wielomianu, bo przecież każdy wielomian jest, z dokładnością do czynnika stałego, wielomianem charakterystycznym pewnej macierzy.

Rozkład QR

Metoda QR (Francis, [1961])[16] jest procedurą iteracyjną, prowadzącą do macierzy T z (5.5.1). Jak sugeruje jego nazwa, algorytm rozkłada daną macierz na czynniki Q i R.

[16] W tymże roku, niezależnie, ten algorytm zaproponowała Kublanowska; zob. Kiełbasiński, Schwetlick [*1992, s. 455] (*przyp. tłum.*).

W podrozdziale 5.3 podano algorytm rozkładu

$$A = QR \tag{5.5.2}$$

na czynniki: unitarny Q i trójkątny górny R. Teraz dodatkowo chcemy, aby elementy przekątniowe macierzy R były nieujemne. Nie jest to trudne. Istotnie, jeśli rozkład (5.5.2) nie ma tej własności, to określamy macierz unitarną przekątniową D tak, że $d_{ii} := r_{ii}/|r_{ii}|$ dla $r_{ii} \neq 0$ i $d_{ii} := 1$ w przeciwnym razie i zamiast (5.5.2) stosujemy rozkład

$$A = (QD)(D^{\mathsf{H}}R) = \hat{Q}\hat{R}.$$

Macierz $\hat{R} = D^{\mathsf{H}}R$ ma już na przekątnej liczby nieujemne.

Podstawowa postać metody QR jest następująca:

$A_1 \leftarrow A$
for $k = 1$ **to** M **do**
  rozkład $A_k = Q_k R_k$
    (Q_k unitarna, R_k trójkątna górna z nieujemną przekątną)
  $A_{k+1} \leftarrow R_k Q_k$
end do

Przy pewnych założeniach elementy przekątniowe macierzy A_k dążą, gdy $k \to \infty$, do wartości własnych macierzy A.

W praktyce ten podstawowy wariant łączymy z dodatkowymi procedurami, które skracają obliczenia i przyspieszają zbieżność. Rozważymy je teraz. Przedtem zauważmy, że – po pierwsze – wszystkie macierze A_k są unitarnie podobne do A, gdyż

$$A_k = Q_k R_k = (Q_k R_k)(Q_k Q_k^{\mathsf{H}}) = Q_k A_{k+1} Q_k^{\mathsf{H}}.$$

Po drugie, jeśli macierz A jest rzeczywista, to i następne macierze A_k będą takie. Skoro tak, to dla macierzy mającej wartości własne zespolone możemy w najlepszym razie oczekiwać zbieżności do macierzy blokowej trójkątnej, z blokami 2×2 wzdłuż przekątnej.

Redukcja do macierzy górnej Hessenberga

Aby zmniejszyć koszt obliczeń iteracyjnych, poprzedzamy je sprowadzeniem macierzy A do unitarnie podobnej macierzy *górnej Hessenberga* H. Nazywamy tak macierz, w której $h_{ij} = 0$ dla $i > j + 1$. Inaczej mówiąc, w takiej macierzy nie znikają co najwyżej elementy leżące nad przekątną, na niej lub tuż pod nią. Macierz A redukujemy do H, stosując algorytm Householdera. W jego k-tym kroku doprowadzamy do właściwej postaci k-tą kolumnę macierzy, nie naruszając zer w poprzednich kolumnach. Napiszmy macierz otrzymaną w $(k-1)$-szym kroku w postaci blokowej

$$\begin{bmatrix} B_{k\times k} & C \\ D & E_{(n-k)\times(n-k)} \end{bmatrix}.$$

Wskaźniki oznaczają rozmiar bloków. B jest macierzą górną Hessenberga stopnia k. Macierz D rozmiaru $(n-k)\times k$ ma wszędzie zera poza ostatnią kolumną. C jest macierzą $k\times(n-k)$. Macierze C i E nie mają jakiejś szczególnej postaci.

Jeśli U jest macierzą unitarną stopnia $n-k$, to

$$\begin{bmatrix} I & 0 \\ 0 & U \end{bmatrix}\begin{bmatrix} B & C \\ D & E \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & U^{\mathsf{H}} \end{bmatrix} = \begin{bmatrix} B & CU^{\mathsf{H}} \\ UD & UEU^{\mathsf{H}} \end{bmatrix}. \tag{5.5.3}$$

Wybieramy macierz U tak, aby iloczyn UD miał w k-tej kolumnie elementy $\beta, 0, 0, \ldots, 0$. Przypomnijmy, że

$$D = \begin{bmatrix} 0 & \ldots & 0 & d_1 \\ 0 & \ldots & 0 & d_2 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & d_{n-k} \end{bmatrix}.$$

Wobec tego macierz U ma być taka, że

$$Ud = \beta e^{(1)}, \quad \text{gdzie} \quad d := \begin{bmatrix} d_1 \\ d_2 \\ \ldots \\ d_{n-k} \end{bmatrix}.$$

W iloczynie UD kolumny do $(k-1)$-szej zerują się automatycznie, a k-ta kolumna ma zera poniżej elementu β.

Zgodnie z lematem 5.2.2 wybieramy β tak, żeby iloczyn $\langle d, \beta e^{(1)}\rangle$ był rzeczywisty i żeby było $\|\beta e^{(1)}\|_2 = \|d\|_2$. Przyjmujemy

$$U := I - vv^{\mathsf{H}}, \quad \text{gdzie} \quad v := \alpha(d - \beta e^{(1)}) \tag{5.5.4}$$

oraz $\beta := -(d_1/|d_1|)\|d\|_2$ i $\alpha := \sqrt{2}/\|d - \beta e^{(1)}\|_2$ (te szczegóły wynikają z dowodu wspomnianego lematu i komentarzy do rozkładu QR Hessenberga w podrozdz. 5.3).

Przykład 5.5.1. Zredukować macierz

$$A = \left[\begin{array}{c|ccc} 1 & 2 & 3 & 4 \\ \hline 4 & 5 & 6 & 7 \\ 2 & 1 & 5 & 0 \\ 4 & 2 & 1 & 0 \end{array}\right]$$

do unitarnie podobnej macierzy górnej Hessenberga.

Rozwiązanie. Wyżej podzielono macierz A na takie bloki, z jakich korzysta się w pierwszym kroku naszkicowanej już procedury. Jest wtedy $d = (4, 2, 4)$. Pierwsza składowa jest rzeczywista, więc wystarczy przyjąć, że $\beta = -\|d\|_2 = -6$ i $\alpha = 1/\sqrt{60}$. Zgodnie z (5.5.4) obliczamy

$$v = \frac{1}{\sqrt{60}}(10, 2, 4)$$

i pierwszą macierzą unitarną jest

$$U = I - vv^{\mathsf{H}} = \frac{1}{15}\begin{bmatrix} -10 & -5 & -10 \\ -5 & 14 & -2 \\ -10 & -2 & 11 \end{bmatrix}.$$

Aby zakończyć pierwszy krok, wykonujemy mnożenia wskazane w (5.5.3):

$$UAU^{\mathsf{H}} = \left[\begin{array}{cc|cc} 1 & -5 & \frac{8}{5} & \frac{6}{5} \\ -6 & \frac{77}{9} & -\frac{163}{45} & \frac{34}{45} \\ \hline 0 & \frac{62}{45} & \frac{677}{225} & -\frac{311}{225} \\ 0 & \frac{259}{45} & -\frac{536}{225} & -\frac{352}{225} \end{array}\right] =$$

$$= \left[\begin{array}{cc|cc} 1 & -5.0000 & 1.6000 & 1.2000 \\ -6 & 8.5556 & -3.6222 & 0.75556 \\ \hline 0 & 1.3778 & 3.0089 & -1.3822 \\ 0 & 5.7556 & -2.3822 & -1.5644 \end{array}\right].$$

Tę częściowo zredukowaną macierz podzielono na bloki, przygotowując ją do drugiego kroku, który daje następujące wyniki:

$$d = (1.3778, 5.7556), \quad \beta = -5.9182, \quad \alpha = 0.15218,$$

$$v = (1.1103, 0.87590), \quad U = \begin{bmatrix} -0.23280 & -0.97252 \\ -0.97252 & 0.23280 \end{bmatrix}.$$

Wykonujemy znów obliczenia zgodne z (5.5.3), co daje ostateczną macierz górną Hessenberga:

$$H = \begin{bmatrix} 1 & -5 & -1.5395 & -1.2767 \\ -6 & 8.5556 & 0.10848 & 3.6986 \\ 0 & -5.9182 & -2.1689 & -1.1428 \\ 0 & 0 & -0.14276 & 3.6133 \end{bmatrix}.$$

Obliczenia wykonywano z siedmioma cyframi znaczącymi, a podane wyżej wyniki są zaokrąglone. ■

Metoda QR z przesunięciami

Aby polepszyć własności podstawowej metody QR, uzupełniamy ją powtarzanymi przesunięciami. Przykład pozwoli zrozumieć ich sens.

Przykład 5.5.2. Zastosować metodę QR do macierzy H z przykł. 5.5.1, tj. generować ciąg macierzy A_k według wzorów

$$A_k = Q_k R_k, \quad A_{k+1} = R_k Q_k.$$

Rozwiązanie. Niżej podano wybrane wyniki zaokrąglone do pięciu cyfr znaczących: $A_1 = H$,

$$A_2 = \begin{bmatrix} 10.135 & 1.9821 & -0.75082 & 5.5290 \\ 6.7949 & -2.8402 & 0.52664 & 1.2616 \\ 0 & 0.19692 & 1.5057 & 1.7031 \\ 0 & 0 & 1.7508 & 2.1994 \end{bmatrix}, \ldots,$$

$$A_{10} = \begin{bmatrix} 11.105 & -4.7599 & 3.8826 & -4.0296 \\ -0.00045570 & -3.8487 & -0.72647 & 1.2553 \\ 0 & -0.068658 & 3.5669 & 0.16324 \\ 0 & 0 & 0 & 0.17645 \end{bmatrix}, \ldots,$$

$$A_{20} = \begin{bmatrix} 11.106 & -4.7403 & 3.9060 & -4.0296 \\ 0 & -3.8526 & -0.68985 & 1.2559 \\ 0 & -0.032156 & 3.5706 & 0.15706 \\ 0 & 0 & 0 & 0.17645 \end{bmatrix}.$$

Widzimy, że nie osiągnęliśmy jeszcze celu obliczeń: macierz A_{20} nie jest trójkątna górna, bo jej element $(A_{20})_{32}$ jest daleki od 0. Wolna zbieżność do macierzy trójkątnej górnej jest kłopotliwa, chociaż jedną z wartości własnych, mianowicie 0.17645, daje już wystarczająco dokładnie dziesiąty krok obliczeń. Inną wartość własną (11.106) otrzymaliśmy też dostatecznie szybko, ale dokładność dwóch pozostałych wynosi chyba tylko dwie lub trzy cyfry znaczące. ■

Wolną zbieżność podstawowego algorytmu można przyspieszyć, stosując *przesunięcia* kolejnych macierzy; rozumiemy przez to zamianę A na $A - zI$. Daje to następującą *metodę QR z przesunięciami*:

$A_1 \leftarrow$ macierz górna Hessenberga dla A
for $k = 1$ **to** M **do**
    rozkład $A_k - z_k I = Q_k R_k$
    $A_{k+1} \leftarrow R_k Q_k + z_k I$
end do

Liczbę z_k w tym algorytmie określa się jako ostatni element przekątniowy macierzy A_k. Wtedy iteracje powinny szybko dać ostatni wiersz postaci $0, 0, \ldots, 0, \alpha$. Liczba α jest więc wartością własną macierzy A. W tym momencie obliczeń jest wskazana deflacja macierzy, tj. usunięcie jej ostatniego wiersza i kolumny w sposób opisany w podrozdz. 5.2. Dalsze obliczenia wykonujemy już dla macierzy niższego stopnia (co jest sensowne wobec poniższego lematu). Początkowa redukcja do macierzy górnej Hessenberga jest wskazana dla dużych macierzy, gdyż zmniejsza koszt obliczeń. Istotnie, taką samą postać mają macierze A_k konstruowane według podanej wyżej metody.

Lemat 5.5.3. *Jeśli macierz A wyraża się w postaci blokowej*

$$A = \begin{bmatrix} B & C \\ 0 & E \end{bmatrix},$$

gdzie B i E są macierzami kwadratowymi, to dowolna liczba λ jest wartością własną tej macierzy wtedy i tylko wtedy, gdy jest wartością własną bloku B lub bloku E.

Dowód. Równanie $Ax = \lambda x$, czyli

$$\begin{bmatrix} B & C \\ 0 & E \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \lambda \begin{bmatrix} u \\ v \end{bmatrix},$$

wyrażamy ostatecznie w postaci

$$Bu + Cv = \lambda u, \quad Ev = \lambda v. \tag{5.5.5}$$

Jeśli λ jest wartością własną macierzy A, to ten układ ma nietrywialne rozwiązanie (u, v). Jeśli $v \neq 0$, to λ jest wartością własną macierzy E. W przeciwnym razie jest $u \neq 0$ i λ jest wartością własną macierzy B.

Przypuśćmy teraz, że λ jest wartością własną macierzy B, a u związanym z nią wektorem własnym. Wtedy układ $(u, 0)$ spełnia warunki (5.5.5). Jeśli λ jest wartością własną macierzy E (a więc istnieje wektor $v \neq 0$ taki, że $Ev = \lambda v$), ale nie jest taką wartością dla B, to równanie $(B - \lambda I)u = -Cv$ względem u ma rozwiązanie, czyli układ (u, v) spełnia (5.5.5). ■

Przykład 5.5.4. Zastosować algorytm QR z przesunięciami do macierzy H z przykładu 5.5.1.

Rozwiązanie. Niżej podano wyniki pięciu iteracji tego algorytmu i zastosowania deflacji:

$$A \to H,$$

$$H \to A_5 = \begin{bmatrix} 2.6141 & -10.087 & -2.4480 & -2.4727 \\ -5.5345 & 4.6668 & 3.5719 & 2.8753 \\ 0 & -0.28730 & 0.14546 & 0.10900 \\ 0 & 0 & 0 & 3.5736 \end{bmatrix},$$

$$\text{deflacja } A_5 \to \tilde{A}_5 = \begin{bmatrix} 11.001 & -5.0329 & -4.1730 \\ -0.30955 & -3.7507 & 1.3719 \\ 0 & 0 & 0.17645 \end{bmatrix},$$

$$\text{deflacja } \tilde{A}_5 \to \hat{A}_5 = \begin{bmatrix} 11.106 & -4.7234 \\ 0 & -3.8556 \end{bmatrix}.$$

Obliczonymi wartościami własnymi są: 3.5736, 0.17645, 11.106 i −3.8556. Wszystkie podane cyfry są poprawne. ■

Operacje elementarne na wierszach i kolumnach

Inna, nieco prostsza, metoda redukcji macierzy do postaci Hessenberga polega na zastosowaniu operacji elementarnych (zob. podrozdz. 4.1) na wierszach i kolumnach. W każdym kroku przekształcamy jeden wiersz i jedną kolumnę, mnożąc macierz A z lewej przez E_i i z prawej przez E_i^{-1}; tę odwrotność łatwo obliczyć (zob. zad. 4.1.8). Poniższy przykład wyjaśnia szczegóły postępowania.

Przykład 5.5.5. Zredukować do macierzy górnej Hessenberga macierz

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

Rozwiązanie. Gdybyśmy nie wybierali elementów głównych, to należałoby odjąć taką wielokrotność wiersza 2. od wiersza 3. i (inną) od 4., aby wyzerować odpowiednie elementy. Chcemy jednak, jak w eliminacji Gaussa, aby mnożniki były małe. Dlatego przestawiamy wiersze 2. i 4., a dla zachowania podobieństwa także te same kolumny (dla uproszczenia nie stosujemy tu tablicy permutacji p [por. podrozdz. 4.3], która pozwala uniknąć faktycznych przestawień):

$$A \to \begin{bmatrix} -3 & 3 & 7 & 2 \\ 4 & 2 & -2 & 4 \\ 2 & -1 & 0 & 3 \\ 1 & 2 & 3 & -5 \end{bmatrix} \to \begin{bmatrix} -3 & 2 & 7 & 3 \\ 4 & 4 & -2 & 2 \\ 2 & 3 & 0 & -1 \\ 1 & -5 & 3 & 2 \end{bmatrix}.$$

Odejmujemy teraz wiersz 2. pomnożony przez $\frac{1}{2}$ od wiersza 3. i pomnożony przez $\frac{1}{4}$ od wiersza 4. Następują po tym przeciwne operacje na kolumnach: dodajemy do kolumny 2. kolumnę 3. pomnożoną przez $\frac{1}{2}$ i kolumnę 4. pomnożoną przez $\frac{1}{4}$ (zob. zad. 4.1.8).

$$\begin{bmatrix} -3 & 2 & 7 & 3 \\ 4 & 4 & -2 & 2 \\ 0 & 1 & 1 & -2 \\ 0 & -6 & \frac{7}{2} & \frac{3}{2} \end{bmatrix} \to \begin{bmatrix} -3 & \frac{25}{4} & 7 & 3 \\ 4 & \frac{7}{2} & -2 & 2 \\ 0 & 1 & 1 & -2 \\ 0 & -\frac{31}{8} & \frac{7}{2} & \frac{3}{2} \end{bmatrix}.$$

W kolumnie 2. porównujemy moduły elementów trzeciego (1) i czwartego $(-\frac{31}{8})$. Większy jest ten ostatni, więc najpierw przestawiamy wiersze 3. i 4. oraz te same kolumny. Następnie wykonujemy operacje na wierszach:

$$\begin{bmatrix} -3 & \frac{25}{4} & 3 & 7 \\ 4 & \frac{7}{2} & 2 & -2 \\ 0 & -\frac{31}{8} & \frac{3}{2} & \frac{7}{2} \\ 0 & 1 & -2 & 1 \end{bmatrix} \to \begin{bmatrix} -3 & \frac{25}{4} & 3 & 7 \\ 4 & \frac{7}{2} & 2 & -2 \\ 0 & -\frac{31}{8} & \frac{3}{2} & \frac{7}{2} \\ 0 & 0 & -\frac{50}{31} & \frac{59}{31} \end{bmatrix}.$$

i przeciwne operacje na kolumnach:

$$\begin{bmatrix} -3 & \frac{25}{4} & \frac{37}{31} & 7 \\ 4 & \frac{7}{2} & \frac{78}{31} & -2 \\ 0 & -\frac{31}{8} & \frac{37}{62} & \frac{7}{2} \\ 0 & 0 & -\frac{2022}{961} & \frac{59}{31} \end{bmatrix}.$$

Końcowa macierz jest macierzą górną Hessenberga podobną do A. ■

ZADANIA 5.5

(Wszystkie macierze w poniższych zadaniach są kwadratowe).

1. Udowodnić, że wielomian charakterystyczny macierzy stopnia n

$$A := \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \dots & -a_{n-2} & -a_{n-1} \end{bmatrix}$$

jest równy $\pm p(\lambda) = \pm(a_0 + a_1\lambda + a_2\lambda^2 + \ldots + a_{n-1}\lambda^{n-1} + \lambda^n)$. Wskazówka: Wyznacznik $\det(A - \lambda I)$ rozwinąć względem ostatniej kolumny.

2. Dla macierzy A z poprzedniego zadania wykazać, że $p(A) = 0$.

3. Znaleźć wartości własne macierzy

$$\begin{bmatrix} -1 & -4 & 1 \\ -1 & -2 & -5 \\ 5 & 4 & 3 \end{bmatrix}.$$

4. Dla liczb całkowitych p i q takich, że $1 \leqslant p < q \leqslant n$ oraz liczb zespolonych α i β takich, że $|\alpha|^2 + |\beta|^2 = 1$ i $\alpha\bar{\beta}$ jest rzeczywiste, określamy macierz U stopnia n różniącą się od I tylko czterema elementami: $U_{pp} = U_{qq} = \alpha$, $U_{pq} = -U_{qp} = \beta$. Udowodnić, że U jest unitarna.

5. Wykazać, że w algorytmie QR jest $A_{k+1} = Q_k^{\mathrm{H}} A_k Q_k$ i że wobec tego

$$A_k = (Q_1 Q_2 \ldots Q_k)(R_k R_{k-1} \ldots R_1).$$

6. Niech A będzie macierzą rzeczywistą o strukturze blokowej trójkątnej górnej:

$$A = \begin{bmatrix} A_{11} & A_{12} & \ldots & A_{1n} \\ 0 & A_{22} & \ldots & A_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & A_{nn} \end{bmatrix},$$

gdzie wszystkie bloki są kwadratowe stopnia drugiego. Podać i uzasadnić prostą procedurę obliczania wartości własnych tej macierzy.

7. Udowodnić, że macierz A z poprzedniego zadania jest osobliwa wtedy i tylko wtedy, gdy taki jest co najmniej jeden z bloków A_{ii}.

8. Sprawdzić, czy stąd, że macierz U jest unitarna, R trójkątna górna, a UR jest macierzą górną Hessenberga, wynika, że i U jest tej ostatniej postaci.

9. Wykazać, że jeśli macierz T jest trójkątna górna, a A jest macierzą górną Hessenberga, to AT i TA są tej drugiej postaci.

10. Wykazać, że jeśli macierz T jest trójkątna górna, a AT jest macierzą górną Hessenberga, to TA jest tej drugiej postaci. Nie zakładać, że A i T są nieosobliwe.

11. Ilu mnożeń wymaga w przybliżeniu sprowadzenie (jak w przykł. 5.5.5) macierzy stopnia n przez podobieństwo do macierzy górnej Hessenberga?

12. Udowodnić, że w algorytmie QR z przesunięciami macierze A_k i A_{k+1} są unitarnie podobne.

13. *Uogólnione zadanie własne* polega na wyznaczeniu wartości λ, dla których układ $Ax = \lambda Bx$ ma nietrywialne rozwiązanie x. Wykazać, że jeśli macierz B jest nieosobliwa, to zadanie redukuje się do zwykłego zadania własnego.

14. (cd.). Wykazać, że jeśli znamy uogólnioną wartość własną μ dla zadania $Ax = \mu(B + tA)x$, to dla $\mu t \neq 1$ można łatwo znaleźć wartość własną dla zadania $Ax = \lambda Bx$.

ZADANIA KOMPUTEROWE 5.5

K1. Napisać procedurę redukującą macierz do postaci górnej Hessenberga zgodnie z (5.5.3) i zastosować ją do macierzy z przykł. 5.5.1.

K2. (cd.). Dołączyć procedurę realizującą podstawowy algorytm QR i odtworzyć wyniki przykł. 5.5.2.

K3. (cd.). Zmodyfikować tę procedurę tak, aby realizowała algorytm QR z przesunięciami i odtworzyć wyniki przykł. 5.5.4. Za pomocą tejże procedury sprawdzić, czy macierz

$$\begin{bmatrix} 190 & 66 & -84 & 30 \\ 66 & 303 & 42 & -36 \\ 336 & -168 & 147 & -112 \\ 30 & -36 & 28 & 291 \end{bmatrix}$$

ma wartości własne 343, 294 i $147 \pm 196i$.

K4. Napisać procedurę redukującą macierz do postaci górnej Hessenberga za pomocą przekształceń przez podobieństwo użytych w przykł. 5.5.5. Sprawdzić procedurę dla macierzy z tego przykładu oraz dla macierzy z przykł. 5.5.1. Porównać wyniki z tymi, które tam podano.

ROZDZIAŁ 6

Aproksymacja funkcji

6.0. Wstęp

W tym rozdziale omawiamy sposoby wyrażania funkcji stosowane w obliczeniach komputerowych. Metody są bardzo rozmaite; zależą od własności funkcji i od tego, czy jej wartości znamy tylko w niewielu punktach, czy wszędzie. Istotne jest również to, jak funkcję chcemy wyrazić – jako wielomian, funkcję sklejaną, ułamek łańcuchowy, czy jeszcze inaczej. Zaczynamy od aproksymacji wielomianowej, która jest najstarsza i najprostsza.

6.1. Interpolacja wielomianowa

Rozwiązujemy następujące zadanie: znaleźć wielomian p możliwie najniższego stopnia taki, że dla danych $n+1$ punktów (x_i, y_i) jest

$$p(x_i) = y_i \qquad (0 \leqslant i \leqslant n).$$

Mówimy, że ten wielomian *interpoluje* wartości y_k w *węzłach* x_k. Jeśli są to wartości pewnej funkcji f, to mówimy też, że p interpoluje f. Omawiając zadania interpolacji wielomianowej, będziemy oznaczać symbolem Π_n zbiór wszystkich wielomianów co najwyżej n-tego stopnia, czyli wielomianów postaci $a_0 + a_1 x + \ldots + a_n x^n$.

Twierdzenie 6.1.1. *Jeśli liczby $x_0, x_1, \ldots, x_n$ są parami różne, to istnieje dokładnie jeden wielomian $p_n \in \Pi_n$ taki, że*

$$p_n(x_i) = y_i \qquad (0 \leqslant i \leqslant n).$$

Dowód. Udowodnimy najpierw jednoznaczność tego wielomianu. Przypuśćmy, że dwa wielomiany, p_n i q_n, z klasy Π_n spełniają powyższe warunki. Wtedy różnica $p_n - q_n$ znika w $n+1$ różnych punktach x_i. Ponieważ jednak wielomian z Π_n nie znikający tożsamościowo ma co najwyżej n zer, więc $p_n \equiv q_n$.

Istnienie wielomianu sprawdzamy przez indukcję. Dla $n = 0$ jest oczywiste, że wielomian stały $p_0(x) = y_0$ spełnia jedyny warunek interpolacyjny. Wiedząc już, że dla pewnego k naturalnego istnieje wielomian $p_{k-1} \in \Pi_{k-1}$ taki, że $p_{k-1}(x_i) = y_i$ dla $0 \leqslant i \leqslant k-1$, próbujemy wyrazić p_k w postaci

$$p_k(x) = p_{k-1}(x) + c(x - x_0)(x - x_1)\ldots(x - x_{k-1}).$$

Dla każdej stałej c stopień tego wielomianu nie przewyższa k. Jest też oczywiste, że p_k spełnia warunki interpolacyjne dla $0 \leqslant i \leqslant k-1$. Pozostaje wyznaczyć c tak, żeby było również $p_k(x_k) = y_k$. Tak jest, gdy

$$p_{k-1}(x_k) + c(x_k - x_0)(x_k - x_1)\ldots(x_k - x_{k-1}) = y_k, \tag{6.1.1}$$

a stąd można obliczyć c. ■

Wzór interpolacyjny Newtona

Dowód tw. 6.1.1 zawiera już w istocie metodę konstrukcji wielomianu interpolacyjnego. Każdy z wielomianów $p_1, p_2, \ldots, p_n$ powstaje z poprzedniego przez dodanie jednego składnika:

$$p_k(x) = c_0 + c_1(x - x_0) + \ldots + c_k(x - x_0)(x - x_1)\ldots(x - x_{k-1}). \tag{6.1.2}$$

Wielomiany $p_0, p_1, \ldots, p_{k-1}$ są krótszymi sumami początkowych składników z (6.1.2). Bardziej zwarta postać tego wzoru jest następująca:

$$p_k(x) = \sum_{i=0}^{k} c_i \prod_{j=0}^{i-1}(x - x_j). \tag{6.1.3}$$

Jest to *wzór interpolacyjny Newtona* (poznamy też inne możliwe wyrażenia wielomianu interpolacyjnego). Z (6.1.1) wynika sposób obliczania współczynników c_k tego wzoru. Stosując go, można np. sprawdzić, że dla $n = 3$ i punktów

$$\begin{array}{c|cccc} x & 5 & -7 & -6 & 0 \\ \hline y & 1 & -23 & -54 & -954 \end{array} \tag{6.1.4}$$

jest

$$\begin{aligned} p(x) &= 1 + 2(x-5) + 3(x-5)(x+7) + 4(x-5)(x+7)(x+6) = \\ &= 4x^3 + 35x^2 - 84x - 954. \end{aligned} \tag{6.1.5}$$

Nie warto jednak stosować tej metody, bo jest zbyt kosztowna (i może powodować nadmiernie duże błędy zaokrągleń) w porównaniu z algorytmem konstrukcji tablicy ilorazów różnicowych (zob. podrozdz. 6.2), do której liczby c_k należą.

Wzór interpolacyjny Lagrange'a

Poznamy inną postać wielomianu interpolacyjnego. Trzeba tu przypomnieć, że przy założeniu jak w tw. 6.1.1 jest on określony jednoznacznie. To jednak nie wyklucza istnienia jego różnych postaci i różnych algorytmów ich konstrukcji. Zauważmy, że inne uporządkowanie punktów (x_i, y_i) zupełnie zmienia składniki wzoru interpolacyjnego Newtona, choć nie narusza całej sumy.

Wyrażamy wielomian p jako sumę

$$p(x) = \sum_{k=0}^{n} y_k l_k(x), \tag{6.1.6}$$

w której l_k są wielomianami zależnymi od węzłów $x_0, x_1, \ldots, x_n$, ale nie od wartości $y_0, y_1, \ldots, y_n$. Gdyby jedna z nich, mianowicie y_i, była równa 1, a pozostałe by znikały, to mielibyśmy równość

$$\delta_{ij} = p_n(x_j) = \sum_{k=0}^{n} y_k l_k(x_j) = \sum_{k=0}^{n} \delta_{ki} l_k(x_j) = l_i(x_j) \quad (0 \leqslant j \leqslant n)$$

(δ_{ij} jest tu deltą Kroneckera). Łatwo znaleźć wielomian l_i o tej własności. Jego zerami są wszystkie węzły oprócz x_i, czyli dla pewnej stałej c jest

$$l_i(x) = c(x - x_0) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n),$$

a c wynika z warunku $l_i(x_i) = 1$:

$$l_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} \quad (0 \leqslant i \leqslant n). \tag{6.1.7}$$

Wzór (6.1.6) z takimi l_i jest *wzorem interpolacyjnym Lagrange'a.*

Przykład 6.1.2. Znaleźć wielomiany l_i i wzór Lagrange'a dla $n = 3$ i punktów (6.1.4).

Rozwiązanie. Wielomiany l_i wyrażają się przez węzły tak:

$$l_0(x) = \frac{(x+7)(x+6)x}{(5+7)(5+6)\cdot 5} = \frac{1}{660}(x+7)(x+6)x,$$
$$l_1(x) = \frac{(x-5)(x+6)x}{(-7-5)(-7+6)(-7)} = -\frac{1}{84}(x-5)(x+6)x,$$
$$l_2(x) = \frac{(x-5)(x+7)x}{(-6-5)(-6+7)(-6)} = \frac{1}{66}(x-5)(x+7)x,$$
$$l_3(x) = \frac{(x-5)(x+7)(x+6)}{(0-5)(0+7)(0+6)} = -\frac{1}{210}(x-5)(x+7)(x+6).$$

Stąd wynika, że

$$p(x) = l_0(x) - 23l_1(x) - 54l_2(x) - 954l_3(x). \qquad \blacksquare$$

Jeszcze inne wyrażenia wielomianu interpolacyjnego mają też swoje zalety i wady. Możemy np. szukać współczynników wielomianu przy potęgach zmiennej:

$$p(x) = a_0 + a_1x + \ldots + a_nx^n.$$

Warunki interpolacyjne prowadzą do układu $n+1$ równań liniowych względem tych współczynników:

$$\begin{bmatrix} 1 & x_0 & \ldots & x_0^n \\ 1 & x_1 & \ldots & x_1^n \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & \ldots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \ldots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \ldots \\ y_n \end{bmatrix}. \tag{6.1.8}$$

Macierz tego układu (nieosobliwa, gdyż na mocy tw. 6.1.1 układ ma jednoznaczne rozwiązanie) nazywamy *macierzą Vandermonde'a*. W zadaniu 13 podano wyrażenie dla jej wyznacznika. Ta macierz bywa źle uwarunkowana (Gautschi [1984]) i dlatego nie zaleca się stosowania układu (6.1.8). Zresztą koszt obliczeń byłby tu nadmiernie duży.

W obliczeniach numerycznych najbardziej użyteczny jest wzór interpolacyjny Newtona w wersji (6.2.1), zawierającej ilorazy różnicowe. Jego zaletą jest to, że dołączenie dodatkowych punktów (x_i, y_i) nie narusza obliczonych wcześniej współczynników c_j. Wartość tak wyrażonego wielomianu można łatwo obliczyć, stosując wariant schematu Hornera. Zaletą wzoru Lagrange'a jest natomiast niezależność wielomianów l_i od rzędnych y_j, co przydaje

się w rozważaniach analitycznych, np. w konstrukcji kwadratur (rozdz. 7). Ten sam wzór może być też wygodny wtedy, gdy dla ustalonych węzłów trzeba uwzględnić różne układy wielkości y_j, wynikających np. z pomiarów. Opracowano algorytmy efektywnego obliczania wartości wielomianu ze wzoru Lagrange'a (Werner [1984] i prace tam cytowane).

Błąd interpolacji wielomianowej

Poznamy teraz twierdzenie pozwalające oszacować odchylenie wielomianu interpolacyjnego od funkcji interpolowanej, jeśli jest ona dostatecznie regularna.

Twierdzenie 6.1.3. *Jeśli $f \in C^{n+1}[a,b]$, a wielomian $p \in \Pi_n$ interpoluje wartości funkcji f w $n+1$ różnych punktach $x_0, x_1, \ldots, x_n$ przedziału $[a,b]$, to dla każdego $x \in [a,b]$ istnieje takie $\xi_x \in (a,b)$, że*

$$f(x) - p(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi_x) \prod_{i=0}^{n} (x - x_i). \tag{6.1.9}$$

Dowód. Twierdzenie wystarczy udowodnić dla x różnego od wszystkich węzłów, bo w przeciwnym razie obie strony powyższej równości znikają. Niech będzie

$$w(t) := \prod_{i=0}^{n} (t - x_i), \quad \varphi := f - p - \lambda w,$$

gdzie λ jest liczbą rzeczywistą, dla której $\varphi(x) = 0$ (x jest ustalone). Stąd

$$\lambda = \frac{f(x) - p(x)}{w(x)}.$$

Wtedy φ jest funkcją klasy $C^{n+1}[a,b]$ znikającą w $n+2$ punktach $x, x_0, x_1, \ldots, x_n$. Stosując twierdzenie Rolle'a, wnioskujemy, że funkcja φ' ma co najmniej $n+1$ różnych zer w (a,b), funkcja φ'' ma tam co najmniej n różnych zer itd., a funkcja $\varphi^{(n+1)}$ ma tam co najmniej jedno zero. Oznaczmy je symbolem ξ_x. Ponieważ

$$\varphi^{(n+1)} = f^{(n+1)} - p^{(n+1)} - \lambda w^{(n+1)} = f^{(n+1)} - (n+1)!\lambda,$$

więc

$$0 = \varphi^{(n+1)}(\xi_x) = f^{(n+1)}(\xi_x) - (n+1)!\lambda =$$
$$= f^{(n+1)}(\xi_x) - (n+1)! \frac{f(x) - p(x)}{w(x)},$$

a to daje wzór (6.1.9). ■

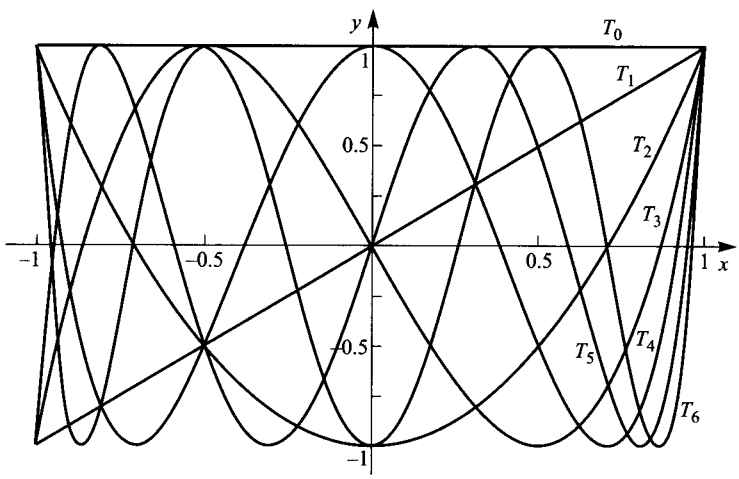
Przykład 6.1.4. Jaki jest błąd przybliżenia funkcji $f(x) = \sin x$ wielomianem interpolacyjnym stopnia 9 w przedziale $[0,1]$, do którego należą węzły?

Rozwiązanie. Ponieważ $|f^{(10)}(\xi_x)| \leqslant 1$ i $\prod_{i=0}^{9}|x - x_i| \leqslant 1$, więc z (6.1.9) wynika, że dla każdego $x \in [0,1]$ jest

$$|\sin x - p(x)| \leqslant \frac{1}{10!} < 2.8_{10}{-}7. \qquad \blacksquare$$

Wielomiany Czebyszewa

W oszacowaniu różnicy $f(x) - p(x)$ wynikającym z tw. 6.1.3, ostatni czynnik można zoptymalizować, wybierając węzły w szczególny sposób. To zadanie zbadał wielki matematyk rosyjski Czebyszew (1821–1894), a rozwiązanie wyraża się przez wielomiany noszące jego imię. Zaczniemy od ich definicji i podstawowych własności.



RYS. 6.1. Wielomiany Czebyszewa T_n

Wielomiany Czebyszewa I rodzaju można określić wzorem rekurencyjnym

$$T_0(x) := 1, \quad T_1(x) := x,$$
$$T_n(x) := 2xT_{n-1}(x) - T_{n-2}(x) \quad (n \geqslant 2). \tag{6.1.10}$$

Stąd wynika, że w szczególności

$$T_2(x) = 2x^2 - 1, \quad T_3(x) = 4x^3 - 3x, \quad T_4(x) = 8x^4 - 8x^2 + 1,$$
$$T_5(x) = 16x^5 - 20x^3 + 5x, \quad T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1.$$

Rysunek 6.1 pokazuje wykresy tych wielomianów w przedziale $[-1,1]$.

Czebyszew otrzymał te wielomiany, rozwiązując zadanie optymalizacji mechanizmu przenoszenia ruchu posuwistego tłoka w parowozie na ruch obrotowy kół. Od tamtych czasów wielomiany Czebyszewa, które mają wiele fascynujących własności, są ważnym narzędziem matematyki stosowanej; zob. Fox i Parker [*1968], Mason i Handscomb [*2002], Paszkowski [*1975] i Rivlin [1990].

Twierdzenie 6.1.5. *W przedziale* $[-1,1]$ *wielomiany Czebyszewa wyrażają się wzorem*

$$T_n(x) = \cos(n \arccos x) \quad (n \geqslant 0). \tag{6.1.11}$$

Dowód. Ze znanej tożsamości trygonometrycznej

$$\cos(A+B) = \cos A \cos B - \sin A \sin B$$

wynika, że

$$\cos(n \pm 1)\theta = \cos\theta \cos n\theta \mp \sin\theta \sin n\theta.$$

Sumujemy te dwie tożsamości stronami:

$$\cos(n+1)\theta = 2\cos\theta \cos n\theta - \cos(n-1)\theta.$$

Dla $\theta = \arccos x$, czyli $x = \cos\theta$, daje to wzór rekurencyjny z (6.1.10). Dla $n = 0$ i $n = 1$ z (6.1.11) wynika, że $T_0(x) = 1$, $T_1(x) = x$. ■

Z (6.1.11) wynika wiele własności wielomianów Czebyszewa, w tym następujące:

$$|T_n(x)| \leqslant 1 \quad (-1 \leqslant x \leqslant 1),$$

$$T_n\left(\cos\frac{j\pi}{n}\right) = (-1)^j \quad (0 \leqslant j \leqslant n),$$

$$T_n\left(\cos\frac{(2j-1)\pi}{2n}\right) = 0 \quad (1 \leqslant j \leqslant n).$$

Tu i w paru innych rozdziałach wielomian $p(x)$ stopnia n nazywamy *standardowym*[1)], jeśli jego współczynnik przy x^n jest równy 1. Takim wielomianem jest np. $2^{1-n}T_n$, co łatwo wynika z (6.1.10).

[1)] W angielskiej terminologii matematycznej: *monic polynomial*. Poza nią ten przymiotnik nie jest chyba notowany w słownikach. Powyższa propozycja przekładu nie jest zapewne najlepsza, ale np. nazwa *wielomian unormowany* lub *znormalizowany* byłaby zbyt wieloznaczna (*przyp. tłum.*).

W dalszym ciągu używamy też następującej normy funkcji f ciągłej w przedziale $[a, b]$:

$$\|f\|_{[a,b]} := \max_{a \leqslant x \leqslant b} |f(x)|.$$

Twierdzenie 6.1.6. *Jeśli p jest wielomianem standardowym stopnia $n > 0$, to*

$$\|p\|_{[-1,1]} \geqslant 2^{1-n},$$

a znak równości jest tu osiągnięty dla $p = 2^{1-n}T_n$.

Dowód. Przypuśćmy, że $|p(x)| < 2^{1-n}$ dla $|x| \leqslant 1$. Określamy też wielomian standardowy $q(x) := 2^{1-n}T_n(x)$. Jeśli $x_i = \cos(i\pi/n)$, to

$$(-1)^i p(x_i) \leqslant |p(x_i)| < 2^{1-n} = (-1)^i q(x_i),$$

czyli

$$(-1)^i [q(x_i) - p(x_i)] > 0 \quad (0 \leqslant i \leqslant n).$$

Stąd wynika, że różnica $q - p$ w przedziale $[-1, 1]$ zmienia znak n razy, czyli ma tam n zer. To jednak jest niemożliwe, gdyż stopień tego wielomianu jest mniejszy od n. ■

Optymalne węzły interpolacji

Niech w tw. 6.1.3 będzie $[a, b] = [-1, 1]$. Wtedy z (6.1.9) wynika, że

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

Wobec tw. 6.1.6 dla dowolnego układu węzłów jest

$$\left\| \prod_{i=0}^{n} (x - x_i) \right\|_{[-1,1]} \geqslant 2^{-n}$$

i ta norma jest najmniejsza, gdy pod jej znakiem występuje $2^{-n}T_{n+1}$, tzn. gdy węzłami interpolacji są punkty

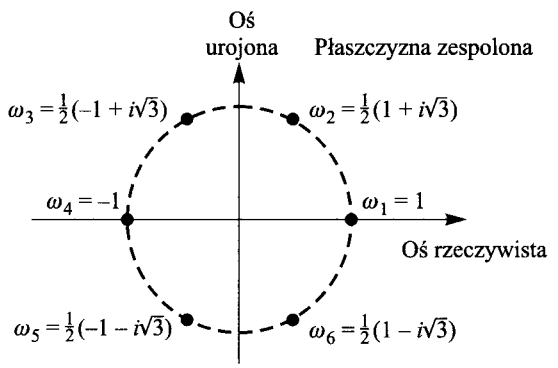
$$x_i = \cos \frac{(2i+1)\pi}{2n+2} \quad (0 \leqslant i \leqslant n).$$

Twierdzenie 6.1.7. *Jeśli węzłami x_i są zera wielomianu Czebyszewa T_{n+1}, to dla $|x| \leqslant 1$ jest*

$$|f(x) - p(x)| \leqslant \frac{1}{2^n (n+1)!} \|f^{(n+1)}\|_{[-1,1]}.$$

Zbieżność wielomianów interpolacyjnych

Dla funkcji f ciągłej w przedziale $[a, b]$ i dla n naturalnych utwórzmy wielomiany interpolacyjne p_n, każdy z równomiernie rozmieszczonymi węzłami. Można by przypuścić, że ciąg $\{p_n\}$ jest zbieżny *jednostajnie* do f, tzn. że ciąg norm $\|f - p_n\|_{[a,b]}$ dąży do 0, gdy $n \to \infty$. Z przykładu 6.1.4 wynika, że tak rzeczywiście jest dla $f(x) := \sin x$, ale to nie jest typowa funkcja ciągła. Istotnie, należy ona do klasy C^∞ na całej prostej rzeczywistej, a jako funkcja zmiennej zespolonej jest funkcją *całkowitą*, czyli nie mającą żadnych osobliwości na płaszczyźnie zespolonej.



RYS. 6.2. Sześć pierwiastków stopnia szóstego z 1

Zaskakujące jest to, że dla większości funkcji ciągłych powyższy ciąg norm nie dąży do 0. Pierwszy przykład tego typu podał Meray w 1884 r. Dotyczy on funkcji $f(z) = 1/z$ na okręgu jednostkowym $|z| = 1$. Niech, dla ustalonego n, liczby $\omega_1, \omega_2, \ldots, \omega_n$ będą n-tymi pierwiastkami z liczby 1. Są one równomiernie rozmieszczone na tym okręgu, co dla $n = 6$ pokazuje rys. 6.2. Dla tej funkcji f wielomian interpolacyjny p_{n-1} z węzłami ω_j jest równy po prostu z^{n-1}. Istotnie,

$$p_{n-1}(\omega_j) = \omega_j^{n-1} = \frac{1}{\omega_j}\omega_j^n = \frac{1}{\omega_j} = f(\omega_j) \quad (1 \leqslant j \leqslant n).$$

Norma różnicy $f - p_{n-1}$ na kole jednostkowym jest równa

$$\max_{|z|=1} |f(z) - p_{n-1}(z)| = \max_{|z|=1} |z^{-1} - z^{n-1}| = \max_{|z|=1} \frac{1}{|z|}|1 - z^n| = 2,$$

bo taką wartość otrzymuje się wtedy, gdy $z^n = -1$.

Inny przykład, odnoszący się do dziedziny rzeczywistej, podał Runge w 1901 r. Wykazał on, że dla funkcji $f(x) = (x^2 + 1)^{-1}$ rozpatrywanej

w przedziale $[-5,5]$ ciąg wielomianów interpolacyjnych z węzłami rozmieszczonymi tam w równych odstępach nie dąży do f. Nawet dla niewielkich wartości n (np. dla $n = 15$) wielomian p_n „szaleje". Jest to *zjawisko Rungego*, opisywane w wielu książkach i artykułach; zob. np. Epperson [1987]. Jego doświadczalne potwierdzenie jest tematem zad. 6.2.K1.

Dowód tego, że zjawisko Rungego występuje, można znaleźć w książce Steffensena [1950] (I wyd. opublikowane w 1927 r.). Książka Davisa [1982] daje ujęcie tego tematu związane z dziedziną zespoloną. Funkcja Rungego, pozornie „niewinna", ma jednak osobliwość na osi urojonej w pobliżu przedziału $[-5,5]$ i to właśnie jest przyczyną kłopotów. Uzasadnienie tego wykracza jednak poza zakres niniejszej książki. Niżej podano dwa twierdzenia – tw. 6.1.9 pokazuje, że pod pewnymi warunkami ciąg $\{p_n\}$ jest zbieżny do f, a tw. 6.1.8 świadczy o tym, że nie zawsze tak jest. Różnica kryje się w kwantyfikatorach!

W 1914 roku Faber udowodnił następujące bardzo ogólne twierdzenie:

Twierdzenie 6.1.8 (Faber). *Dla dowolnego ciągu układów węzłów*

$$a \leqslant x_0^{(n)} < x_1^{(n)} < \ldots < x_n^{(n)} \leqslant b \quad (n \geqslant 0) \tag{6.1.12}$$

istnieje taka funkcja ciągła w $[a,b]$, że ciąg wielomianów interpolacyjnych zbudowanych dla tych węzłów nie jest do niej zbieżny.

Jest to twierdzenie bardziej subtelne niż mogłoby się wydawać na pierwszy rzut oka, skoro zachodzi też takie twierdzenie o zbieżności:

Twierdzenie 6.1.9. *Jeśli f jest funkcją ciągłą w $[a,b]$, to istnieje taki ciąg układów węzłów* (6.1.12)*, że zbudowane dla nich wielomiany interpolacyjne tworzą ciąg zbieżny do f.*

To twierdzenie wynika z zestawienia dwóch bardzo ważnych twierdzeń, odpowiednio – autorstwa Weierstrassa (o aproksymacji) i Czebyszewa (o alternansie). To drugie znajduje się w podrozdz. 6.9, dowód pierwszego (tw. 6.1.11) opiera się na pomocniczym tw. 6.1.10 i własnościach wielomianów, które w 1912 r. zdefiniował i zbadał S. Bernstein.

Poniższe twierdzenie dotyczy operatorów *liniowych* i *dodatnich* przekształcających przestrzeń funkcji ciągłych $C[a,b]$ w nią samą. Operator L ma te własności, jeśli odpowiednio

$$L(\varphi f + \psi g) = \varphi Lf + \psi Lg \quad (\varphi, \psi \in \mathbb{R},\ f, g \in C[a,b]),$$
$$Lf \geqslant 0, \quad \text{gdy} \quad f \geqslant 0.$$

Okazuje się, że są to własności niezmiernie istotne dla teorii aproksymacji.

Twierdzenie 6.1.10 (Bohman-Korowkin). *Niech $\{L_n\}$ $(n \geqslant 1)$ będzie ciągiem operatorów liniowych dodatnich, przekształcających przestrzeń $C[a,b]$ w nią samą. Jeśli dla $f(x) := 1, x, x^2$ jest $\{\|L_n f - f\|_{[a,b]}\} \to 0$, to jest tak również dla każdej innej funkcji $f \in C[a,b]$.*

Dowód. Jeśli L jest operatorem liniowym i dodatnim, to zachodzą następujące implikacje:

$$f \geqslant g \Rightarrow f - g \geqslant 0 \Rightarrow L(f-g) \geqslant 0 \Rightarrow Lf - Lg \geqslant 0 \Rightarrow Lf \geqslant Lg.$$

Prócz tego jest $|f| \geqslant f$ i $|f| \geqslant -f$, a zatem $L(|f|) \geqslant Lf$ i $L(|f|) \geqslant -Lf$, czyli $L(|f|) \geqslant |Lf|$.

Niech teraz będzie $h_k(x) = x^k$ dla $k = 0, 1, 2$. Zdefiniujmy też funkcje:

$$\alpha_n = L_n h_0 - h_0, \quad \beta_n = L_n h_1 - h_1, \quad \gamma_n = L_n h_2 - h_2.$$

Z założenia wynika, że

$$\{\|\alpha_n\|\} \to 0, \quad \{\|\beta_n\|\} \to 0, \quad \{\|\gamma_n\|\} \to 0$$

(dla uproszczenia tu i dalej pomijamy w symbolu normy przedział $[a,b]$).

Rozważmy teraz dowolną funkcję f z $C[a,b]$. Jest to więc funkcja jednostajnie ciągła, tzn. dla każdego $\varepsilon > 0$ istnieje takie $\delta > 0$, że jeśli $x, y \in [a,b]$ i $|x-y| < \delta$, to $|f(x) - f(y)| < \varepsilon$. Stąd wynika, że dla $c = 2\|f\|/\delta^2$ jest

$$|x-y| \geqslant \delta \Rightarrow |f(x) - f(y)| \leqslant 2\|f\| \leqslant 2\|f\| \frac{(x-y)^2}{\delta^2} = c(x-y)^2.$$

Dlatego dla wszystkich x i y z $[a,b]$ jest $|f(x) - f(y)| \leqslant \varepsilon + c(x-y)^2$. Używając funkcji $h_k(x)$, wyrażamy to tak:

$$|f - f(y)h_0| \leqslant \varepsilon h_0 + c(h_2 - 2yh_1 + y^2 h_0).$$

Uwzględniając uwagi z początku dowodu, wnioskujemy stąd, że

$$|L_n f - f(y) L_n h_0| \leqslant \varepsilon L_n h_0 + c(L_n h_2 - 2y L_n h_1 + y^2 L_n h_0).$$

Jest to nierówność wiążąca funkcje zmiennej x. Podstawmy w jej miejsce y:

$$\begin{aligned}
&|(L_n f)(y) - f(y)(L_n h_0)(y)| \leqslant \\
&\quad \leqslant \varepsilon (L_n h_0)(y) + c[(L_n h_2)(y) - 2y(L_n h_1)(y) + y^2 (L_n h_0)(y)] = \\
&\quad = \varepsilon[1 + \alpha_n(y)] + c[y^2 + \gamma_n(y) - 2y(y + \beta_n(y)) + y^2(1 + \alpha_n(y))] = \\
&\quad = \varepsilon + \varepsilon\alpha_n(y) + c\gamma_n(y) - 2cy\beta_n(y) + cy^2\alpha_n(y) \leqslant \\
&\quad \leqslant \varepsilon + \varepsilon\|\alpha_n\| + c\|\gamma_n\| + 2c\|h_1\|\,\|\beta_n\| + c\|h_2\|\,\|\alpha_n\|.
\end{aligned}$$

Zgodnie z założeniem istnieje takie m, że dla wszystkich $n \geqslant m$ prawa strona tej nierówności nie przewyższa 2ε. Wtedy

$$\|L_n - fL_nh_0\| \leqslant 2\varepsilon.$$

Kończąc dowód, zauważmy, że

$$\|L_nf - f\| \leqslant \|L_nf - fL_nh_0\| + \|fL_nh_0 - fh_0\| \leqslant 2\varepsilon + \|f\|\,\|\alpha_n\|.$$

W razie potrzeby zwiększamy m tak, żeby dla $n \geqslant m$ było $\|f\|\,\|\alpha_n\| < \varepsilon$. Dla tych n zachodzi więc nierówność

$$\|L_nf - f\| < 3\varepsilon.$$

ε było dowolne, a zatem twierdzenie zostało udowodnione. ■

Zapowiedziane już *wielomiany Bernsteina* są zdefiniowane, dla dowolnej funkcji $f \in C[0,1]$, wzorem

$$(B_nf)(x) := \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k}x^k(1-x)^{n-k}.$$

Użyta symbolika podkreśla to, że wielomian Bernsteina jest w istocie operatorem liniowym; przekształca on funkcję ciągłą w $[0,1]$ na wielomian n-tego stopnia. Ponieważ w tym przedziale jest $x^k(1-x)^{n-k} \geqslant 0$, więc jest to operator dodatni. Stosujemy te wielomiany w elementarnym dowodzie[2)] twierdzenia Weierstrassa, ale używa się ich także w projektowaniu wspomaganym komputerowo, choć tu wypierają je tzw. funkcje B-sklejane (zob. podrozdz. 6.5 i 6.6).

Twierdzenie 6.1.11 (Weierstrass). *Jeśli funkcja f jest ciągła w przedziale $[a,b]$, to dla każdego $\varepsilon > 0$ istnieje taki wielomian p, że $\|f-p\|_{[a,b]} \leqslant \varepsilon$.*

Dowód. Możemy ograniczyć się do przedziału $[0,1]$, ponieważ przekształcenie liniowe zmiennej $x = a + t(b-a)$ zmienia $[0,1]$ na $[a,b]$, a wielomian p na inny wielomian tego samego stopnia.

Zgodnie z tw. 6.1.10 pozostaje wykazać, że dla $h_k(x) := x^k$ i $k = 0,1,2$ jest $\{B_nh_k\} \to h_k$. Dla $k = 0$ mamy równość

$$(B_nh_0)(x) = \sum_{k=0}^{n}\binom{n}{k}x^k(1-x)^{n-k} = [x + (1-x)]^n = 1.$$

[2)] Dowód odwołuje się jednak do raczej wyrafinowanego tw. 6.1.10. Pierwotny dowód Bernsteina jest chyba bardziej naturalny; zob. Achiezer [*1957]. Istnieje zresztą wiele innych jeszcze dowodów (*przyp. tłum.*).

Dla $k = 1$ korzystamy z zad. 22:

$$(B_n h_1)(x) = \sum_{k=0}^{n} \frac{k}{n} \binom{n}{k} x^k (1-x)^{n-k} = \sum_{k=1}^{n} \binom{n-1}{k-1} x^k (1-x)^{n-k} =$$
$$= x \sum_{k=0}^{n-1} \binom{n-1}{k} x^k (1-x)^{n-1-k} = x.$$

Dla $k = 2$ stosujemy to zadanie dwukrotnie, zakładając, że $n > 1$:

$$(B_n h_2)(x) = \sum_{k=0}^{n} \left(\frac{k}{n}\right)^2 \binom{n}{k} x^k (1-x)^{n-k} = \sum_{k=1}^{n} \frac{k}{n} \binom{n-1}{k-1} x^k (1-x)^{n-k} =$$
$$= \sum_{k=1}^{n} \left(\frac{n-1}{n} \frac{k-1}{n-1} + \frac{1}{n}\right) \binom{n-1}{k-1} x^k (1-x)^{n-k} =$$
$$= \frac{n-1}{n} x^2 \sum_{k=2}^{n} \binom{n-2}{k-2} x^{k-2} (1-x)^{n-k} + \frac{x}{n} =$$
$$= \frac{n-1}{n} x^2 + \frac{x}{n} \to x^2. \qquad \blacksquare$$

ZADANIA 6.1

1. Udowodnić, że jeśli g interpoluje funkcję f w węzłach $x_0, x_1, \ldots, x_{n-1}$, a h interpoluje f w węzłach $x_1, x_2, \ldots, x_n$, to funkcja

$$g(x) + \frac{x_0 - x}{x_n - x_0}[g(x) - h(x)]$$

interpoluje f we wszystkich węzłach $x_0, x_1, \ldots, x_n$ (g i h nie muszą być wielomianami).

2. Udowodnić, że jeśli funkcja g (wielomian lub nie) interpoluje funkcję f w węzłach $x_0, x_1, \ldots, x_{n-1}$, a h jest funkcją taką, że $h(x_i) = \delta_{in}$ $(0 \leqslant i \leqslant n)$, to dla pewnego c funkcja $g + ch$ interpoluje f w $x_0, x_1, \ldots, x_n$.

3. Niech E będzie $(n+1)$-wymiarową przestrzenią wektorową funkcji określonych w obszarze D. Niech $x_0, x_1, \ldots, x_n$ będą różnymi punktami z D. Udowodnić, że dla $f \in E$ zadanie interpolacyjne

$$f(x_i) = y_i \quad (0 \leqslant i \leqslant n)$$

ma dla dowolnych rzędnych y_i jednoznaczne rozwiązanie wtedy i tylko wtedy, gdy jedyną funkcją z E znikającą we wszystkich x_i jest funkcja tożsamościowo równa 0.

4. Znaleźć wielomian możliwie niskiego stopnia, interpolujący poniższe dane:

(a)

x	3	7
y	5	−1

(b)

x	7	1	2
y	146	2	1

(c)

x	3	7	1	2
y	10	146	2	1

(d)

x	3	7	1	2
y	12	146	2	1

(e)

x	1	2	0	3
y	3	2	−4	5

(f)

x	1	3	2	6
y	−2	−22	−1	−37

(g)

x	1.5	2.7	3.1	−2.1	−6.6	11.0
y	0	0	0	1	0	0

5. Zastosować wzory interpolacyjne Lagrange'a i Newtona dla następujących danych:

(a)

x	2	0	3
y	11	7	28

(b)

x	−2	0	1
y	0	1	−1

W obu przypadkach wyrażając wielomian w postaci $a+bx+cx^2$, upewnić się, że wyniki są identyczne.

6. Niech będzie $w_i = \prod_{j=0,\, j\neq i}^{n}(x_i - x_j)^{-1}$. Wykazać, że jeśli x nie jest węzłem, to wzór interpolacyjny Lagrange'a wyraża się w postaci *barycentrycznej*

$$p(x) = \Big[\sum_{i=0}^{n} y_i w_i (x-x_i)^{-1}\Big] \Big/ \Big[\sum_{i=0}^{n} w_i (x-x_i)^{-1}\Big].$$

7. (cd.). Udowodnić, że niedokładne wyznaczenie wielkości w_i nie zakłóca własności interpolacyjnej wielomianu p, a ściślej, że $\lim_{x\to x_k} p(x) = y_k$ $(0 \leqslant k \leqslant n)$.

8. Wykazać, że odwzorowanie L funkcji f na wielomian interpolacyjny $p \in \Pi_n$ wyznaczony z warunków $p(x_i) = f(x_i)$ $(0 \leqslant i \leqslant n)$ dla ustalonych węzłów x_i jest liniowe, tj. $L(\alpha f + \beta g) = \alpha Lf + \beta Lg$.

9. (cd.). Udowodnić, że $Lq = q$ dla każdego wielomianu $q \in \Pi_n$.

10. Korzystając ze wzoru Lagrange'a, udowodnić, że współczynnik przy x^n w wielomianie $p(x)$ jest równy

$$\sum_{i=0}^{n} y_i \prod_{j=0,\, j\neq i}^{n} (x_i - x_j)^{-1}.$$

11. Udowodnić, że dla każdego wielomianu $q \in \Pi_{n-1}$

$$\sum_{i=0}^{n} q(x_i) \prod_{j=0,\, j\neq i}^{n} (x_i - x_j)^{-1} = 0.$$

12. Niech odwzorowanie G będzie określone wzorem $Gf = \sum_{i=0}^{n} f(x_i) l_i^2$, gdzie wielomiany l_i są określone jak we wzorze Lagrange'a. Udowodnić, że Gf należy do Π_{2n} i interpoluje f w węzłach oraz że $Gf \geqslant 0$, jeśli $f \geqslant 0$.

13. Udowodnić, że

$$\begin{vmatrix} 1 & x_0 & \dots & x_0^n \\ 1 & x_1 & \dots & x_1^n \\ \dots & \dots & \dots & \dots \\ 1 & x_n & \dots & x_n^n \end{vmatrix} = \prod_{0 \leqslant j < k \leqslant n} (x_k - x_j).$$

14. Udowodnić, że dla dowolnego układu 23 węzłów z przedziału $[-1, 1]$ wielomian interpolacyjny $p \in \Pi_{22}$ dla funkcji $f(x) := \cosh x$ jest taki, że w tym przedziale $|[p(x) - f(x)]/f(x)| \leqslant 5_{10}{-16}$.

15. Oszacować błąd $\|p - f\|_{[-1,1]}$, gdzie $f(x) := e^{x-1}$, a p jest dla f wielomianem interpolacyjnym stopnia 12 z węzłami należącymi do $[-1, 1]$.

16. Udowodnić, że jeśli wśród n węzłów z przedziału $[-1, 1]$ znajduje się liczba 0, to wielomian $p \in \Pi_{n-1}$ interpolujący w tych węzłach funkcję $f(x) := \sinh x$, spełnia nierówność

$$|p(x) - f(x)| \leqslant \frac{2^n}{n!} |f(x)| \quad (|x| \leqslant 1).$$

17. Równanie $x - 9^{-x} = 0$ ma pierwiastek w przedziale $[0, 1]$. Znaleźć wielomian interpolujący lewą stronę równania w punktach 0, 0.5, 1. Przyrównując ten wielomian do 0, znaleźć przybliżenie tego pierwiastka.

18. Zaprojektować metodę rozwiązywania równania $f(x) = 0$, dającą w $n + 1$ krokach dokładną wartość pierwiastka, gdy funkcja odwrotna f^{-1} jest w jego otoczeniu wielomianem stopnia n.

19. Znaleźć współczynniki wielomianu $T_n(x)$ przy x^{n-1} i x^{n-2}.

20. Wzorując się na dowodzie tw. 6.1.5, wykazać, że dla $|x| \geqslant 1$ jest $T_n(x) = \cosh(n \operatorname{arcosh} x)$.

21. Sprawdzić, czy jeśli naturalne n jest dzielnikiem naturalnego m, to każde zero wielomianu T_n jest zerem wielomianu T_m.

22. Udowodnić, że

$$\frac{k}{n}\binom{n}{k} = \binom{n-1}{k-1} \qquad (n \geqslant k \geqslant 1).$$

6.2. Ilorazy różnicowe

Wracamy do zadania interpolacyjnego rozważanego w poprzednim podrozdziale. Wiemy już, że wielomian $p \in \Pi_n$ spełniający dla danej funkcji f warunki interpolacyjne

$$p(x_i) = f(x_i) \quad (0 \leqslant i \leqslant n)$$

w parami różnych węzłach x_i można wyrazić za pomocą wzoru Newtona:

$$p(x) = \sum_{k=0}^{n} c_k q_k(x),$$

gdzie

$$q_k(x) := \prod_{j=0}^{k-1} (x - x_j).$$

Z dowodu tw. 6.1.1 wynika natychmiast, że c_k zależy tylko od węzłów $x_0, x_1, \ldots, x_k$ i wartości funkcji w tych węzłach. Tę zależność przyjęto oznaczać następująco:

$$c_k = f[x_0, x_1, \ldots, x_k].$$

Jest to *iloraz różnicowy rzędu k* dla funkcji f i wymienionych wyżej węzłów.

Używając ilorazów różnicowych, wyrażamy wzór interpolacyjny Newtona w ogólnie przyjęty sposób, pokazujący zależność wielomianu interpolacyjnego p od funkcji interpolowanej f:

$$p(x) = \sum_{k=0}^{n} f[x_0, x_1, \ldots, x_k] \prod_{j=0}^{k-1} (x - x_j). \tag{6.2.1}$$

Jawne wyrażenia ilorazów rzędu zerowego i pierwszego są oczywiste:

$$f[x_0] = f(x_0), \qquad f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Obliczanie ilorazów różnicowych wyższych rzędów

Ilorazy wyższych rzędów oblicza się rekurencyjnie stosując wzór podany w poniższym twierdzeniu, a podobny do wyrażenia dla $f[x_0, x_1]$:

Twierdzenie 6.2.1. *Ilorazy różnicowe spełniają zależność*

$$f[x_0, x_1, \ldots, x_k] = \frac{f[x_1, x_2, \ldots, x_k] - f[x_0, x_1, \ldots, x_{k-1}]}{x_k - x_0}. \tag{6.2.2}$$

Dowód. Niech p_k będzie wielomianem klasy Π_k, interpolującym f w węzłach $x_0, x_1, \ldots, x_k$. Będzie on nam potrzebny dla $k = n - 1$ i $k = n$. Prócz tego niech $q \in \Pi_{n-1}$ interpoluje f w węzłach $x_1, x_2, \ldots, x_n$. Zachodzi równość

$$p_n(x) = q(x) + \frac{x - x_n}{x_n - x_0}[q(x) - p_{n-1}(x)]$$

(por. zad. 6.1.1). Porównanie współczynników obu stron przy x^n daje tożsamość (6.2.2). ■

Oznaczenia węzłów użyte w (6.2.2) oczywiście nie są istotne. Ich zmiana daje ogólny wzór

$$f[x_i, x_{i+1}, \ldots, x_{i+j}] = \\ = \frac{f[x_{i+1}, x_{i+2}, \ldots, x_{i+j}] - f[x_i, x_{i+1}, \ldots, x_{i+j-1}]}{x_{i+j} - x_i}. \tag{6.2.3}$$

Znając węzły x_i i wartości funkcji $f(x_i)$, czyli ilorazy $f[x_i]$ zerowego rzędu można za pomocą tego wzoru tworzyć tablicę ilorazów różnicowych wyższych rzędów. Oto przykładowa tablica dla przypadku, gdy znamy cztery węzły:

$$\begin{array}{cc|ccc} x_0 & f[x_0] & f[x_0,x_1] & f[x_0,x_1,x_2] & f[x_0,x_1,x_2,x_3] \\ x_1 & f[x_1] & f[x_1,x_2] & f[x_1,x_2,x_3] & \\ x_2 & f[x_2] & f[x_2,x_3] & & \\ x_3 & f[x_3] & & & \end{array}$$

Tablica jest trójkątna. Pionowa kreska oddziela wielkości dane od obliczanych. Pierwszy wiersz tablicy zawiera te ilorazy, które występują we wzorze interpolacyjnym Newtona (6.2.1).

Przykład 6.2.2. Utworzyć tablicę ilorazów różnicowych dla danych

x	3	1	5	6
$f(x)$	1	−3	2	4

i podać postać wzoru interpolacyjnego Newtona.

Rozwiązanie. Tworzymy tablicę na wzór podanej wyżej:

$$\begin{array}{cc|ccc} 3 & 1 & 2 & -\frac{3}{8} & \frac{7}{40} \\ 1 & -3 & \frac{5}{4} & \frac{3}{20} & \\ 5 & 2 & 2 & & \\ 6 & 4 & & & \end{array}$$

Stąd wynika, że

$$p(x) = 1 + 2(x-3) - \tfrac{3}{8}(x-3)(x-1) + \tfrac{7}{40}(x-3)(x-1)(x-5). \quad ■$$

Algorytm

Ogólniej, przyjmując, że

$$c_{ij} = f[x_i, x_{i+1}, \ldots, x_{i+j}],$$

wyrażamy tablicę ilorazów różnicowych tak:

x_0	c_{00}	c_{01}	c_{02}	$\ldots$	$c_{0,n-1}$	c_{0n}
x_1	c_{10}	c_{11}	c_{12}	$\ldots$	$c_{1,n-1}$	
$\ldots$	$\ldots$	$\ldots$	$\ldots$			
x_{n-1}	$c_{n-1,0}$	$c_{n-1,1}$				
x_n	c_{n0}					

Przy tych samych oznaczeniach wzór rekurencyjny (6.2.3) przybiera postać

$$c_{ij} = \frac{c_{i+1,j-1} - c_{i,j-1}}{x_{i+j} - x_i}.$$

Algorytm konstrukcji powyższej tablicy wynikający wprost z tego wzoru jest w tym sensie nieefektywny, że wystarczy użyć tablicy d zmiennych z jednym wskaźnikiem. Początkową wartością zmiennej d_i jest $c_{i0} = f(x_i)$ z drugiej kolumny trójkątnej tablicy ilorazów, następnymi wartościami – wielkości $c_{i-1,1}, \ldots, c_{1,i-1}, c_{0i}$; ostatnia z nich znajduje się w pierwszym wierszu tejże tablicy i we wzorze Newtona. Tablicę tworzymy kolumnami, a w każdej kolumnie – z dołu do góry. Dzięki takiej kolejności obliczeń tablica d zawiera w każdej chwili ilorazy, które będą później potrzebne[3]:

for $i = 0$ **to** n **do**
    $d_i \leftarrow f(x_i)$
end do
for $j = 1$ **to** n **do**
    for $i = n$ **to** j **step** -1 **do**
        $d_i \leftarrow (d_i - d_{i-1})/(x_i - x_{i-j})$
    end do
end do

W zadaniu 13 ocenia się koszt obliczenia ilorazów różnicowych według tego algorytmu.

[3] Stosując podany tu algorytm, warto zwrócić uwagę na pominięty przez autorów aspekt. Jeśli chcemy obliczyć możliwie dokładnie ilorazy różnicowe, to zaleca się porządkowanie danych węzłów od najmniejszego do największego (albo odwrotnie), bo wtedy – jak udowodnił Kiełbasiński – algorytm jest numerycznie poprawny; zob. Jankowscy [*1981, w szczególności tw. 2.8]. Jeśli natomiast naszym celem są jak najdokładniejsze wartości wielomianu interpolacyjnego (6.2.1), to istotny jest tzw. porządek Leji węzłów (Higham [2002, s. 100]); można go uzyskać kosztem ok. n^2 operacji (*przyp. tłum.*).

Własności ilorazów różnicowych

Podrozdział 6.2 kończymy omówieniem pewnych własności ilorazów różnicowych.

TWIERDZENIE 6.2.3. *Iloraz różnicowy nie zależy od porządku jego argumentów.*

Zamiast dowodu wystarczy przypomnieć, że iloraz $f[x_0, x_1, \ldots, x_n]$ jest współczynnikiem przy x^n w wielomianie interpolacyjnym $p \in \Pi_n$ określonym warunkami $p(x_i) = f(x_i)$ $(0 \leqslant i \leqslant n)$, a porządek tych warunków nie wpływa na p.

TWIERDZENIE 6.2.4. *Jeśli wielomian $p \in \Pi_n$ interpoluje funkcję f w węzłach $x_0, x_1, \ldots, x_n$ parami różnych, to dla każdego t różnego od nich jest*

$$f(t) - p(t) = f[x_0, x_1, \ldots, x_n, t] \prod_{j=0}^{n} (t - x_j).$$

Dowód. Niech q będzie wielomianem interpolacyjnym klasy Π_{n+1} określonym dla f i węzłów $x_0, x_1, \ldots, x_n, t$. Wiadomo, że q powstaje z p przez dodanie jednego składnika:

$$q(x) = p(x) + f[x_0, x_1, \ldots, x_n, t] \prod_{j=0}^{n} (x - x_j).$$

Teza twierdzenia wynika stąd i z równości $q(t) = f(t)$. ■

TWIERDZENIE 6.2.5. *Jeśli funkcja f ma n-tą pochodną ciągłą w przedziale $[a, b]$ zawierającym punkty $x_0, x_1, \ldots, x_n$ parami różne, to istnieje $\xi \in (a, b)$ takie, że*

$$f[x_0, x_1, \ldots, x_n] = \frac{1}{n!} f^{(n)}(\xi).$$

Dowód. Niech wielomian $p \in \Pi_{n-1}$ interpoluje f w węzłach $x_0, x_1, \ldots, x_{n-1}$. Z twierdzenia 6.1.3 wynika istnienie takiego $\xi \in (a, b)$, że

$$f(x_n) - p(x_n) = \frac{1}{n!} f^{(n)}(\xi) \prod_{j=0}^{n-1} (x_n - x_j).$$

Natomiast na mocy tw. 6.2.4

$$f(x_n) - p(x_n) = f[x_0, x_1, \ldots, x_n] \prod_{j=0}^{n-1} (x_n - x_j).$$

Porównując te dwa związki, otrzymujemy tezę twierdzenia. ■

W wielu sytuacjach jest potrzebny *wzór Hermite'a-Genocchiego*, podany w poniższym twierdzeniu. Wyraża on iloraz różnicowy n-tego rzędu przez pewną całkę; obszarem całkowania jest *sympleks*

$$S_n := \Big\{u = (u_0, u_1, \ldots, u_n) \in \mathbb{R}^{n+1} \colon\ u_i \geqslant 0,\ \sum_{i=0}^{n} u_i = 1\Big\}.$$

Twierdzenie 6.2.6. *Dla dowolnych liczb rzeczywistych $x_0, x_1, \ldots, x_n$ jest*

$$f[x_0, x_1, \ldots, x_n] = \int_{S_n} f^{(n)}(u_0 x_0 + u_1 x_1 + \ldots + u_n x_n)\, du. \tag{6.2.4}$$

Twierdzenie udowodnimy, zakładając, że punkty x_i są parami różne, ale pozostaje ono prawdziwe bez tego założenia; odpowiednie uogólnienie ilorazów różnicowych będzie podane w podrozdz. 6.3.

Dowód. Załóżmy najpierw, że $n = 1$. Ponieważ

$$\begin{aligned} S_1 &= \{u = (u_0, u_1) \in \mathbb{R}^2 \colon\ u_0 \geqslant 0,\ u_1 \geqslant 0,\ u_0 + u_1 = 1\} = \\ &= \{(1 - u_1, u_1) \colon\ 0 \leqslant u_1 \leqslant 1\}, \end{aligned}$$

więc całka z twierdzenia jest wtedy równa

$$\begin{aligned} \int_{S_1} f'(u_0 x_0 + u_1 x_1)\, du &= \int_0^1 f'((1 - u_1)x_0 + u_1 x_1)\, du_1 = \\ &= \int_0^1 f'(x_0 + u_1(x_1 - x_0))\, du_1 = \\ &= \int_0^1 \frac{d}{du_1} f(x_0 + u_1(x_1 - x_0))\, \frac{du_1}{x_1 - x_0} = \\ &= \frac{1}{x_1 - x_0} f(x_0 + u_1(x_1 - x_0))\,\Big|_{u_1=0}^{u_1=1} = \\ &= \frac{1}{x_1 - x_0}[f(x_1) - f(x_0)] = f[x_0, x_1]. \end{aligned}$$

Kontynuujemy dowód przez indukcję. Jeśli $n > 1$, to $u_0 = 1 - \sum_{i=1}^{n} u_i$, a całka w (6.2.4) (oznaczmy ją $I(x_0, x_1, \ldots, x_n)$) wyraża się tak:

$$I(x_0, x_1, \ldots, x_n) = \int_{S_n} f^{(n)}(x_0 + u_1(x_1 - x_0) + \ldots + u_n(x_n - x_0))\, du =$$
$$= \int_0^1 \int_0^{1-u_1} \ldots \int_0^{1-u_1-\ldots-u_{n-1}} f^{(n)}(x_0 + u_1(x_1 - x_0) + \ldots$$
$$\ldots + u_n(x_n - x_0))\, du_n \ldots du_2\, du_1.$$

Całkę względem u_n obliczamy tak samo jak dla $n = 1$:

$$\frac{1}{x_n - x_0}\Big[f^{(n-1)}\Big(x_n + \sum_{i=1}^{n-1} u_i(x_i - x_n)\Big) - f^{(n-1)}\Big(x_0 + \sum_{i=1}^{n-1} u_i(x_i - x_0)\Big)\Big].$$

Wobec tego

$$I(x_0, x_1, \ldots, x_n) =$$
$$= \frac{1}{x_n - x_0} \int_0^1 \int_0^{1-u_1} \ldots \int_0^{1-u_1-\ldots-u_{n-2}} \Big[f^{(n-1)}\Big(x_n + \sum_{i=1}^{n-1} u_i(x_i - x_n)\Big) -$$
$$- f^{(n-1)}\Big(x_0 + \sum_{i=1}^{n-1} u_i(x_i - x_0)\Big)\Big] du_{n-1} \ldots du_2\, du_1 =$$
$$= \frac{1}{x_n - x_0}[I(x_n, x_1, \ldots, x_{n-1}) - I(x_0, x_1, \ldots, x_{n-1})].$$

Z założenia indukcyjnego wynika, że to wyrażenie jest równe

$$\frac{1}{x_n - x_0}\{f[x_n, x_1, \ldots, x_{n-1}] - f[x_0, x_1, \ldots, x_{n-1}]\}.$$

Na mocy tw. 6.2.1 i 6.2.3 jest to iloraz $f[x_0, x_1, \ldots, x_n]$. ■

ZADANIA 6.2

1. Znaleźć wielomian interpolacyjny ze wzoru Newtona dla następujących danych:

(a)

x	4	2	0	3
y	63	11	7	28

(b)

x	0	1	2	7
y	51	3	1	201

(c)

x	1	3/2	0	2
y	3	13/4	3	5/3

2. Wielomian $p(x) = 2 - (x+1) + x(x+1) - 2x(x+1)(x-1)$ interpoluje cztery początkowe punkty z tablicy

x	−1	0	1	2	3
y	2	1	2	−7	10

Dodając do p jeden składnik, wyznaczyć wielomian interpolujący wszystkie dane.

3. Udowodnić, że jeśli $f \in \Pi_k$, to dla $n > k$ jest $f[x_0, x_1, \ldots, x_n] = 0$.

4. Odwołując się do wzoru Lagrange'a, wykazać, że

$$f[x_0, x_1, \ldots, x_n] = \sum_{i=0}^{n} \alpha_i f(x_i), \quad \text{gdzie} \quad \alpha_i = \prod_{j=0,\, j\neq i}^{n} (x_i - x_j)^{-1}.$$

Udowodnić, że jeśli $x_0 < x_1 < \ldots < x_n$, to liczby α_i są na przemian dodatnie i ujemne.

5. (cd.). Udowodnić, że

$$\sum_{i=0}^{n} \alpha_i x_i^n = 1, \quad \sum_{i=0}^{n} \alpha_i = \begin{cases} 1 & (n=0) \\ 0 & (n>0). \end{cases}$$

6. Udowodnić, że

$$m!\, f[0, 1, \ldots, m] = \sum_{j=0}^{m} (-1)^{m-j} \binom{m}{j} f(j).$$

7. Korzystając ze wzoru Cramera, wykazać, że

$$f[x_0, x_1, \ldots, x_n] = \begin{vmatrix} 1 & x_0 & \ldots & x_0^{n-1} & f(x_0) \\ 1 & x_1 & \ldots & x_1^{n-1} & f(x_1) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & \ldots & x_n^{n-1} & f(x_n) \end{vmatrix} \Big/ \begin{vmatrix} 1 & x_0 & \ldots & x_0^{n-1} & x_0^n \\ 1 & x_1 & \ldots & x_1^{n-1} & x_1^n \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & \ldots & x_n^{n-1} & x_n^n \end{vmatrix}.$$

8. Udowodnić, że dla $f(x) = x^m$ (m naturalne) iloraz $f[x_0, x_1, \ldots, x_n]$ jest równy 1 dla $n = m$ i równy 0 dla $n > m$.

9. Udowodnić, że jeśli $f(x) = 1/x$, to $f[x_0, x_1, \ldots, x_n] = (-1)^n \prod_{i=0}^{n} x_i^{-1}$.

10. Udowodnić, że

$$(\alpha f + \beta g)[x_0, x_1, \ldots, x_n] = \alpha f[x_0, x_1, \ldots, x_n] + \beta g[x_0, x_1, \ldots, x_n].$$

11. Udowodnić *wzór Leibniza*

$$(fg)[x_0, x_1, \ldots, x_n] = \sum_{k=0}^{n} f[x_0, x_1, \ldots, x_k] g[x_k, x_{k+1}, \ldots, x_n]$$

(warto dostrzec jego podobieństwo do wzoru na pochodną iloczynu dwóch funkcji).

12. Udowodnić, że jeśli f jest wielomianem, to $f[x_0, x_1, \ldots, x_n]$ jest wielomianem względem argumentów tego ilorazu.

13. Sprawdzić, jaki jest koszt obliczenia tablicy trójkątnej ilorazów różnicowych aż do rzędu n.

14. Udowodnić, że jeśli $h > 0$, to dla pewnego $\xi \in (x, x+2h)$ jest

$$f(x+2h) - 2f(x+h) + f(x) = h^2 f''(\xi).$$

15. Udowodnić, że jeśli funkcja f jest ciągła, to iloraz $f[x_0, x_1, \ldots, x_n]$ jest funkcją ciągłą swoich argumentów w $(n+1)$-wymiarowym zbiorze otwartym, w którym są one różne.

16. Udowodnić, że jeśli $f \in C^n$, to iloraz $f[x_0, x_1, \ldots, x_n]$ jest wszędzie ciągły.

17. Czy ξ w tw. 6.2.5 zależy w sposób ciągły od argumentów $x_0, x_1, \ldots, x_n$? A wartość $f^{(n)}(\xi)$?

18. Rozważmy tablicę

$$\begin{array}{cc|ccc} x_0 & y_0 & a_0 & b_0 & c_0 \\ x_1 & y_1 & a_1 & b_1 & \\ x_2 & y_2 & a_2 & & \\ x_3 & y_3 & & & \end{array}$$

w której dla pewnego $x \neq x_0, x_1, x_2, x_3$ wielkości a_i, b_i, c_i oblicza się ze wzorów

$$a_i = [(x_{i+1} - x)y_i + (x - x_i)y_{i+1}]/(x_{i+1} - x_i),$$
$$b_i = [(x_{i+2} - x)a_i + (x - x_i)a_{i+1}]/(x_{i+2} - x_i),$$
$$c_i = [(x_{i+3} - x)b_i + (x - x_i)b_{i+1}]/(x_{i+3} - x_i).$$

Korzystając z wyniku zad. 6.1.1, udowodnić, że c_0 jest wartością w x wielomianu interpolacyjnego stopnia 3 dla punktów (x_i, y_i).

19. (cd.). Uogólniając wzory z poprzedniego zadania, otrzymać *algorytm Neville'a* obliczania $p_n(x)$ dla dowolnego n.

Zadania komputerowe 6.2

K1. Dla funkcji $f(x) := 1/(1+x^2)$ i $n = 5$, 10, 15 znaleźć wielomian interpolacyjny p_n z $n+1$ węzłami równoodległymi w przedziale $[-5, 5]$. Obliczyć $f(x) - p_n(x)$ w 30 równoodległych punktach tego przedziału. Czy otrzymane wyniki potwierdzają rozbieżność ciągu $\{p_n(x)\}$?

K2. Zaprogramować i sprawdzić następującą metodę rozwiązywania równania $f(x) = 0$: dla danych punktów x_0, x_1, x_2 i $n \geqslant 2$ tworzymy wielomian q_2 stopnia drugiego, interpolujący f w węzłach x_{n-2}, x_{n-1}, x_n, i określamy x_{n+1} jako pierwiastek tego wielomianu bliższy punktu x_n.

6.3. Interpolacja Hermite'a

Interpolacja Hermite'a, czyli *interpolacja z węzłami wielokrotnymi*, polega na poszukiwaniu wielomianu, który w węzłach ma dane nie tylko wartości, ale i wartości pochodnych (ścisła definicja będzie podana nieco dalej). W tym sensie przeciwstawiamy ją *interpolacji Lagrange'a* z węzłami pojedynczymi, omawianej w podrozdz. 6.1 i 6.2.

Prostym, ale pouczającym przykładem interpolacji Hermite'a jest następujące zadanie: szukamy wielomianu p możliwie niskiego stopnia, który

w dwóch różnych punktach x_0 i x_1 interpoluje funkcję f i jej pochodną f'. Rozumiemy to tak, że mają być spełnione cztery warunki:

$$p(x_i) = f(x_i), \quad p'(x_i) = f'(x_i) \quad (i = 0, 1). \tag{6.3.1}$$

Dlatego wydaje się rozsądne szukać rozwiązania w klasie Π_3. Przy tym, zamiast obliczać współczynniki przy 1, x, x^2, x^3, szukamy wielkości a, b, c, d takich, że

$$p(x) = a + b(x - x_0) + c(x - x_0)^2 + d(x - x_0)^2(x - x_1)$$

(przypomina to nieco wzór interpolacyjny Newtona). Stąd

$$p'(x) = b + 2c(x - x_0) + 2d(x - x_0)(x - x_1) + d(x - x_0)^2.$$

Warunki nałożone na p prowadzą do równań

$$f(x_0) = a, \qquad f'(x_0) = b,$$
$$f(x_1) = a + bh + ch^2 \quad (h = x_1 - x_0), \qquad f'(x_1) = b + 2ch + dh^2.$$

Stąd od razu wynikają współczynniki a i b. Znając je, wyznaczamy c z trzeciego, a potem d z czwartego równania. Tak więc zadanie można rozwiązać, i to jednoznacznie, dla dowolnych danych.

Ogólnie rzecz biorąc, jeśli szukany wielomian ma dane wartości w pewnych punktach i dane wartości jakichś pochodnych w tych lub innych punktach, to układ równań liniowych względem współczynników może mieć macierz osobliwą. Przekonuje nas o tym prosty przykład.

Przykład 6.3.1. Znaleźć wielomian p taki, że $p(0) = 0$, $p(1) = 1$ i $p'(\frac{1}{2}) = 2$.

Rozwiązanie. Mając trzy warunki, szukamy wielomianu $p(x) = a + bx + cx^2$. Pierwszy warunek daje $a = 0$, a z dwóch pozostałych wynika, że ma być

$$1 = p(1) = b + c, \quad 2 = p'(\tfrac{1}{2}) = b + c.$$

Te równania są jednak sprzeczne, czyli wielomian p założonej postaci nie istnieje. Natomiast w klasie wielomianów $p(x) = a + bx + cx^2 + dx^3$ zadanie ma nieskończenie wiele rozwiązań, bo warunki interpolacyjne prowadzą do równań $a = 0$ i

$$1 = b + c + d, \quad 2 = b + c + \tfrac{3}{4}d.$$

Jest tak, jeśli $d = -4$ i $b + c = 5$. ■

W powyższym przykładzie mamy do czynienia z *interpolacją Birkhoffa*. Ta klasa zadań, których rozwiązywanie sprawia intrygujące kłopoty, była intensywnie badana. Wyniki tych dociekań omawiają w swej monografii Lorentz, Jetter i Riemenschneider [1983].

Aby wykluczyć takie osobliwe przypadki, wystarczy założyć, że jeśli w pewnym węźle wielomian p ma daną wartość j-tej pochodnej, to dane są tam również wartości $p, p', \ldots, p^{(j-1)}$. Tak jest właśnie w interpolacji Hermite'a. Określamy ją ściśle w następujący sposób: dla danych węzłów x_i $(0 \leqslant i \leqslant m)$ parami różnych, danych liczb naturalnych k_i i danych wartości c_{ij} szukamy wielomianu p takiego, że

$$p^{(j)}(x_i) = c_{ij} \quad (0 \leqslant j \leqslant k_i - 1,\ 0 \leqslant i \leqslant m). \tag{6.3.2}$$

Łączną liczbę warunków oznaczamy $n+1$, jest więc

$$n + 1 = k_0 + k_1 + \ldots + k_m$$

i wydaje się sensowne szukanie p w klasie Π_n.

Twierdzenie 6.3.2. *W klasie Π_n istnieje dokładnie jeden wielomian p spełniający warunki* (6.3.2).

Dowód. Z założenia liczba warunków interpolacyjnych jest równa liczbie współczynników szukanego wielomianu. Chcemy upewnić się, że macierz A układu wynikającego z tych warunków jest nieosobliwa. W tym celu wystarczy udowodnić, że układ jednorodny $Au = 0$ ma tylko zerowe rozwiązanie. Taki układ odpowiada zadaniu

$$p^{(j)}(x_i) = 0 \quad (0 \leqslant j \leqslant k_i - 1,\ 0 \leqslant i \leqslant m).$$

Jeśli tak, to dla $0 \leqslant i \leqslant m$ węzeł x_i jest k_i-krotnym zerem wielomianu p i ten musi się dzielić przez $q(x) = \prod_{i=0}^{m}(x - x_i)^{k_i}$. Ponieważ jednak stopniem wielomianu q jest $n+1$, więc q, a tym bardziej p, znika tożsamościowo. ■

Przykład 6.3.3. Co daje interpolacja Hermite'a w przypadku tylko jednego węzła?

Rozwiązanie. W tym przypadku wielomian $p \in \Pi_n$ ma spełniać warunki

$$p^{(j)}(x_0) = c_{0j} \quad (0 \leqslant j \leqslant n).$$

Jego postać wynika ze wzoru Taylora (podrozdz. 1.1):

$$p(x) = c_{00} + c_{01}(x - x_0) + \frac{1}{2!}c_{02}(x - x_0)^2 + \ldots + \frac{1}{n!}c_{0n}(x - x_0)^n. \quad ■$$

Uogólnienie wzoru Newtona

Aby zrozumieć podaną dalej definicję ilorazów różnicowych odnoszącą się do interpolacji z węzłami wielokrotnymi, rozważmy kilka prostych przykładów. Szukamy najpierw wielomianu $p \in \Pi_2$ takiego, że

$$p(x_0) = c_{00}, \quad p'(x_0) = c_{01}, \quad p(x_1) = c_{10}. \tag{6.3.3}$$

Tablica ilorazów różnicowych wygląda tak:

$$\begin{array}{cc|cc} x_0 & c_{00} & c_{01} & ? \\ x_0 & c_{00} & ? & \\ x_1 & c_{10} & & \end{array}$$

Znaki zapytania w tablicy sygnalizują nieznane pozycje. Zauważmy, że w jej pierwszej kolumnie x_0 występuje dwa razy – tyle, ile warunków w tym węźle nałożono na p. Zauważmy też, że daną wartość $p'(x_0)$ umieszczono w kolumnie przeznaczonej dla ilorazów różnicowych pierwszego rzędu. Jest to zgodne z równością

$$\lim_{x \to x_0} f[x_0, x] = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0),$$

która usprawiedliwia definicję $f[x_0, x_0] = f'(x_0)$. Pozostałe pozycje w tablicy obliczamy, stosując formalnie tw. 6.2.1:

$$p[x_0, x_1] = \frac{p(x_1) - p(x_0)}{x_1 - x_0} = \frac{c_{10} - c_{00}}{x_1 - x_0}, \tag{6.3.4}$$

$$p[x_0, x_0, x_1] = \frac{p[x_0, x_1] - p[x_0, x_0]}{x_1 - x_0} = \frac{c_{10} - c_{00}}{(x_1 - x_0)^2} - \frac{c_{01}}{x_1 - x_0}. \tag{6.3.5}$$

Również formalnie stosujemy wzór (6.2.1):

$$p(x) = p(x_0) + p[x_0, x_0](x - x_0) + p[x_0, x_0, x_1](x - x_0)^2. \tag{6.3.6}$$

W zadaniu 2 sprawdza się, że to postępowanie daje poprawny wynik.

W podobny sposób możemy wyrazić wielomian p spełniający warunki (6.3.1):

$$\begin{aligned} p(x) = {} & f(x_0) + f'(x_0)(x - x_0) + f[x_0, x_0, x_1](x - x_0)^2 + \\ & + f[x_0, x_0, x_1, x_1](x - x_0)^2(x - x_1). \end{aligned}$$

Jego współczynniki znajdują się w pierwszym wierszu tablicy

$$\begin{array}{cc|ccc} x_0 & f(x_0) & f'(x_0) & f[x_0, x_0, x_1] & f[x_0, x_0, x_1, x_1] \\ x_0 & f(x_0) & f[x_0, x_1] & f[x_0, x_1, x_1] & \\ x_1 & f(x_1) & f'(x_1) & & \\ x_1 & f(x_1) & & & \end{array}$$

Ilorazy różnicowe, których wszystkie argumenty są identyczne, określa się zgodnie z tw. 6.2.5. Dzięki niemu wiadomo, że jeśli pochodna $f^{(k)}$ jest ciągła w najmniejszym przedziale zawierającym węzły $x_0, x_1, \ldots, x_k$, to w tymże przedziale istnieje ξ takie, że

$$f[x_0, x_1, \ldots, x_k] = \frac{1}{k!} f^{(k)}(\xi).$$

Gdy długość przedziału dąży do 0, otrzymujemy w granicy równość

$$f[x_0, x_0, \ldots, x_0] = \frac{1}{k!} f^{(k)}(x_0). \tag{6.3.7}$$

Trzeba pamiętać o czynniku $1/k!$, istotnym dla $k > 1$.

Przykład 6.3.4. Korzystając z podanych już informacji, znaleźć wielomian p klasy Π_4 spełniający warunki

$$p(1) = 2, \quad p'(1) = 3, \quad p(2) = 6, \quad p'(2) = 7, \quad p''(2) = 8.$$

Rozwiązanie. Niżej, w tablicy po lewej stronie, rozmieszczono wszystkie wielkości dane i wynikające z (6.3.7); znak zapytania sygnalizuje te, które trzeba obliczyć stosując wzór (6.2.3). Prawa tablica jest już kompletna:

1	2	3	?	?	?	1	2	3	1	2	−1
1	2	?	?	?		1	2	4	3	1	
2	6	7	4			2	6	7	4		
2	6					2	6				
2	6					2	6				

Pierwszy wiersz daje współczynniki wielomianu p:

$$p(x) = 2 + 3(x-1) + (x-1)^2 + 2(x-1)^2(x-2) - (x-1)^2(x-2)^2.$$

Można sprawdzić, że spełnia on wszystkie warunki podane wyżej. ■

Definicja ilorazów różnicowych

Ścisłą definicję ilorazów różnicowych, którą sugerują podane już przykłady, możemy wprowadzić na różne (ale równoważne) sposoby. Można je określać rekurencyjnie (Braess [1984]), za pomocą wyznaczników (Schumaker [1981]), albo po prostu przenosząc na dyskutowaną teraz interpolację Hermite'a definicję z podrozdz. 6.2 (Conte i de Boor [1980]). Wybieramy ten ostatni wariant. Przypomnijmy więc, że dla węzłów $x_0, x_1, \ldots, x_n$ parami

różnych iloraz różnicowy $f[x_0, x_1, \ldots, x_n]$ był z definicji współczynnikiem przy x^n wielomianu $p(x) \in \Pi_n$ interpolującego f w tych węzłach. Tak samo w istocie zrobimy teraz. Będzie jednak wygodniej zmienić nieco oznaczenia i w pewien sposób uporządkować węzły (co nie ogranicza ogólności rozważań). Przyjmiemy mianowicie, że w zadaniu interpolacji Hermite'a dostatecznie regularnej funkcji f za pomocą wielomianu $p \in \Pi_n$:

(i) węzłami są liczby $x_0, x_1, \ldots, x_n$ uporządkowane tak, że

$$x_0 \leqslant x_1 \leqslant \ldots \leqslant x_n;$$

(ii) jeśli pewien punkt x_i występuje na liście węzłów k_i razy, to mają być spełnione warunki

$$p^{(j)}(x_i) = f^{(j)}(x_i) \quad (0 \leqslant j \leqslant k_i - 1).$$

Tak więc w przykł. 6.3.4 należy przyjąć $x_0 = x_1 = 1$ i $x_2 = x_3 = x_4 = 2$.

Ponieważ istota interpolacji Hermite'a nie ulega zmianie, więc na mocy tw. 6.3.2 wielomian p jest określony jednoznacznie. *Iloraz różnicowy n-tego rzędu* $f[x_0, x_1, \ldots, x_n]$ jest z definicji równy współczynnikowi przy x^n w $p(x)$. Na przypadek interpolacji Hermite'a można uogólnić wzór interpolacyjny Newtona (6.2.1) i tw. 6.2.1 opisujące sposób rekurencyjnego obliczania ilorazów różnicowych.

Twierdzenie 6.3.5. *Jeśli węzły* $x_0, x_1, \ldots, x_n$ *i warunki interpolacyjne spełniają warunki* (i), (ii), *to*

$$p(x) = \sum_{j=0}^{n} f[x_0, x_1, \ldots, x_j] \prod_{i=0}^{j-1} (x - x_i). \tag{6.3.8}$$

Dowód. Dla $n = 0$ wzór (6.3.8) jest oczywiście poprawny. Jeśli $n > 0$, wielomian

$$q(x) := \sum_{j=0}^{n-1} f[x_0, x_1, \ldots, x_j] \prod_{i=0}^{j-1} (x - x_i)$$

spełnia warunki interpolacyjne w węzłach $x_0, x_1, \ldots, x_{n-1}$, a wielomian p z klasy Π_n dodatkowo spełnia takiż warunek w x_n, to z definicji ilorazu różnicowego wynika, że różnica

$$p(x) - f[x_0, x_1, \ldots, x_n] \prod_{i=0}^{n-1} (x - x_i)$$

jest wielomianem klasy Π_{n-1}. Interpoluje on f w węzłach $x_0, x_1, \ldots, x_{n-1}$ (zob. zad. 5 i 6). Z jednoznaczności wielomianu interpolacyjnego wynika, że

$$p(x) - f[x_0, x_1, \ldots, x_n] \prod_{i=0}^{n-1} (x - x_i) = q(x),$$

a to już daje szukany wzór (6.3.8). ■

TWIERDZENIE 6.3.6. *Niech będzie $x_0 \leqslant x_1 \leqslant \ldots \leqslant x_n$. Jeśli $x_0 = x_n$, to*

$$f[x_0, x_0, \ldots, x_0] = \frac{1}{n!} f^{(n)}(x_0). \tag{6.3.9}$$

Jeśli $x_0 < x_n$, to

$$f[x_0, x_1, \ldots, x_n] = \frac{f[x_1, x_2, \ldots, x_n] - f[x_0, x_1, \ldots, x_{n-1}]}{x_n - x_0}. \tag{6.3.10}$$

Dowód. Dla $n = 1$ wielomian interpolacyjny p wyraża się łatwym do sprawdzenia wzorem

$$p(x) = \begin{cases} f(x_0) + \dfrac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0) & (x_0 \neq x_1) \\ f(x_0) + f'(x_0)(x - x_0) & (x_0 = x_1). \end{cases}$$

W obu przypadkach współczynnik przy x jest równy ilorazowi $f[x_0, x_1]$, określonemu odpowiednio zgodnie z (6.3.10) albo (6.3.9).

Dla $n > 1$ i $x_0 < x_n$ rozumujemy tak jak w dowodzie tw. 6.2.1. Jeśli natomiast $x_0 = x_n$, to ze wzoru Taylora wynika, że

$$p(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(x_0)(x - x_0)^k$$

i współczynnik przy x^n jest zgodny z (6.3.9). ■

Uogólnienie wzoru Lagrange'a

W zasadzie wzór interpolacyjny Lagrange'a można uogólnić na dowolne zadanie interpolacji Hermite'a. Tu ograniczymy się jednak do ważnego szczególnego przypadku; założymy mianowicie, że w każdym węźle są dane wartości funkcji i jej pierwszej pochodnej:

$$p(x_i) = c_{i0}, \quad p'(x_i) = c_{i1} \quad (0 \leqslant i \leqslant n).$$

Przez analogię do wzoru Lagrange'a wyrażamy p w postaci

$$p(x) = \sum_{i=0}^{n} c_{i0} A_i(x) + \sum_{i=0}^{n} c_{i1} B_i(x). \tag{6.3.11}$$

Chwila zastanowienia wystarczy, by się upewnić, że wielomiany A_i, B_i powinny spełniać następujące warunki:

$$A_i(x_j) = \delta_{ij}, \qquad A_i'(x_j) = 0,$$
$$B_i(x_j) = 0, \qquad B_i'(x_j) = \delta_{ij}.$$

Można sprawdzić, że jeśli

$$l_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} \quad (0 \leqslant i \leqslant n)$$

są wielomianami występującymi we wzorze Lagrange'a, to

$$A_i(x) = [1 - 2(x - x_i) l_i'(x_i)] l_i^2(x) \qquad (0 \leqslant i \leqslant n),$$
$$B_i(x) = (x - x_i) l_i^2(x) \qquad (0 \leqslant i \leqslant n).$$

Ponieważ stopień każdego wielomianu l_i wynosi n, więc wielomiany A_i i B_i są stopnia $2n + 1$. Tak powinno być wobec liczby warunków nałożonych na wielomian p.

Dla $n = 1$, jak w (6.3.1), to uogólnienie wzoru Lagrange'a daje wzór

$$p(x) = f(x_0)A_0(x) + f(x_1)A_1(x) + f'(x_0)B_1(x) + f'(x_1)B_1(x),$$

gdzie A_i, B_i wyrażają się jak wyżej przez wielomiany l_i i ich pochodne, tu szczególnie proste:

$$l_0(x) = \frac{x - x_1}{x_0 - x_1}, \qquad l_0'(x) = \frac{1}{x_0 - x_1},$$

$$l_1(x) = \frac{x - x_0}{x_1 - x_0}, \qquad l_1'(x) = \frac{1}{x_1 - x_0}.$$

Dla opisanego tu wariantu interpolacji Hermite'a można oszacować jej błąd:

Twierdzenie 6.3.7. *Jeśli $f \in C^{2n+2}[a, b]$, a węzły $x_0, x_1, \ldots, x_n$ parami różne należą do $[a, b]$, to dla wielomianu $p \in \Pi_{2n+1}$ określonego warunkami*

$$p(x_i) = f(x_i), \quad p'(x_i) = f'(x_i) \quad (0 \leqslant i \leqslant n)$$

i dla każdego $x \in [a,b]$ *istnieje punkt* $\xi \in (a,b)$ *taki, że*

$$f(x) - p(x) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \prod_{i=0}^{n} (x - x_i)^2.$$

Dowód jest taki jak dla tw. 6.1.3, z tą oczywistą różnicą, że teraz pomocniczy wielomian w jest równy $\prod_{i=0}^{n}(t - x_i)^2$. Nietrudno zrozumieć, jak powinno wyglądać twierdzenie podobne do powyższego dla ogólnego zadania interpolacji Hermite'a.

ZADANIA 6.3

1. Znaleźć takie węzły x_0, x_1, że dla dowolnych stałych c_{i0}, c_{i2} istnieje wielomian $p \in \Pi_3$ spełniający warunki

$$p(x_i) = c_{i0}, \quad p''(x_i) = c_{i2} \quad (i = 0, 1).$$

2. Wykazać, że wielomian (6.3.6), w którym ilorazy różnicowe obliczono ze wzorów (6.3.4) i (6.3.5), spełnia warunki (6.3.3).

3. Za pomocą uogólnionego wzoru Newtona znaleźć wielomian $p \in \Pi_4$ określony warunkami

x	0	1	2
$p(x)$	2	−4	44
$p'(x)$	−9	4	

4. (cd.). Znaleźć wielomian $p \in \Pi_5$ spełniający warunki jak w poprzednim zadaniu i dodatkowy warunek $p(3) = 2$.

5. Przyjmując konwencję podaną przed tw. 6.3.5, wykazać, że wielomian interpoluje funkcję równą 0 w węzłach $x_0, x_1, \ldots, x_n$ wtedy i tylko wtedy, gdy dzieli się przez $\prod_{j=0}^{n}(x - x_j)$.

6. (cd.). Wykazać, że jeśli w węzłach $x_0, x_1, \ldots, x_n$ funkcja f interpoluje g, a funkcja h interpoluje 0, to $f + ch$ (c – stała) interpoluje tamże g.

7. Udowodnić, że jeśli f ma pochodną ciągłą i $x_0 < x_1 < \ldots < x_n$, to

$$\frac{\partial}{\partial x_i} f[x_0, x_1, \ldots, x_n] = f[x_0, x_1, \ldots, x_i, x_i, x_{i+1}, \ldots, x_n].$$

8. Niech funkcja f ma m-krotne zero α i k-krotne zero β, gdzie $\alpha < \beta$. Udowodnić, że jeśli $f \in C^n[\alpha, \beta]$, gdzie $n = m + k - 1$, to $f^{(n)}$ ma co najmniej jedno zero w (α, β). Zastosować tw. Rolle'a.

9. Obliczyć $l_i'(x)$.

10. Sprawdzić, że wielomiany A_i, B_i mają własności potrzebne w (6.3.11).

11. Znaleźć wielomian p jak najniższego stopnia taki, że

$$p(x_i) = y_i, \quad p'(x_i) = 0 \quad (0 \leqslant i \leqslant n).$$

12. Udowodnić, że jeśli

$$p(t) = b - (b-a)\left[3\left(\frac{b-t}{b-a}\right)^2 - 2\left(\frac{b-t}{b-a}\right)^3\right],$$

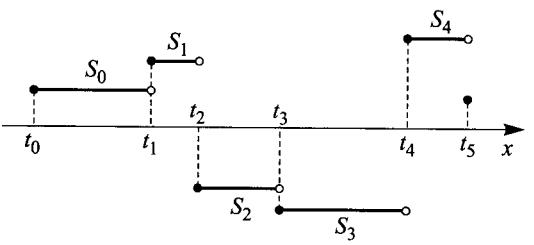
to $|p'(t)| \leqslant p'((a+b)/2) = 3/2$, $p(a) = a$, $p(b) = b$, $p'(a) = 0$ i $p'(b) = 0$.

13. Dla $n = 2$ uprościć wyrażenia dla A_i, B_i.

6.4. Interpolujące funkcje sklejane

Określając te funkcje, ustalamy przede wszystkim $n+1$ *węzłów* $t_0, t_1, \ldots, t_n$ takich, że $t_0 < t_1 < \ldots < t_n$. Dla danej liczby całkowitej nieujemnej k *funkcją sklejaną stopnia* k nazywamy taką funkcję S, która:

1. W każdym z przedziałów $[t_i, t_{i+1})$ $(0 \leqslant i \leqslant n-1)$ jest wielomianem klasy Π_k.
2. Ma ciągłą $(k-1)$-szą pochodną w przedziale $[t_0, t_n]$.

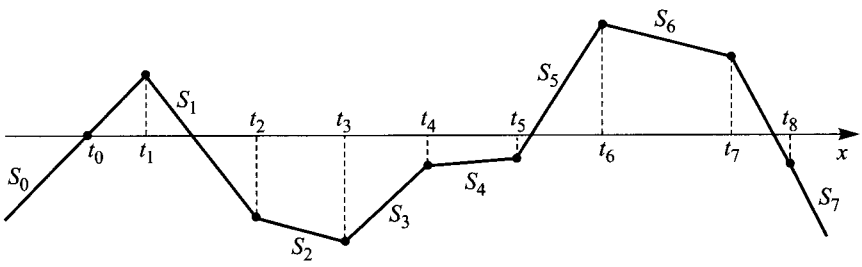


RYS. 6.3. Funkcja sklejana stopnia 0

Funkcja sklejana stopnia 0 (dla niej warunek **2** nic nie wnosi) jest przedziałami stała. Przykład takiej funkcji dla $n = 5$ pokazano na rys. 6.3. Natomiast na rys. 6.4 mamy funkcję sklejaną stopnia 1 dla $n = 8$. Dla $x \in [t_i, t_{i+1})$ jest $S(x) = a_i x + b_i$. Wygodnie jest rozszerzyć to określenie dla $i = 0$ na półprostą $(-\infty, t_1)$, a dla $i = n-1$ na półprostą $[t_{n-1}, \infty)$. Zauważmy, że warunek **2**, czyli żądanie ciągłości funkcji S, wiąże współczynniki a_i, b_i z a_{i+1}, b_{i+1}.

Funkcje sklejane stopnia trzeciego

Opiszemy teraz bardziej szczegółowo konstrukcję i własności funkcji sklejanych stopnia trzeciego (*sześciennych*). Są one często używane w praktyce. Szukamy takiej funkcji S, która w danych węzłach t_i ma dane wartości y_i



RYS. 6.4. Funkcja sklejana stopnia 1

i która w każdym z przedziałów $[t_i, t_{i+1})$ jest wielomianem S_i klasy Π_3. Warunek

$$S_{i-1}(t_i) = y_i = S_i(t_i) \quad (1 \leqslant i \leqslant n-1) \tag{6.4.1}$$

zapewnia ciągłość funkcji S. Ponadto żądamy ciągłości pochodnych S' i S''; to powinno określić pozostałe parametry funkcji sklejanej. Czy tak jednak jest? Wszystkie wielomiany S_i mają łącznie $4n$ współczynników. Z (6.4.1) wynika $2n$ warunków. W każdym węźle wewnętrznym ciągłość pochodnej S' daje jeden warunek:

$$S'_{i-1}(t_i) = S'_i(t_i) \quad (1 \leqslant i \leqslant n-1); \tag{6.4.2}$$

jest ich razem $n-1$. Tyleż warunków wynika z ciągłości drugiej pochodnej. Mamy więc w sumie $4n-2$ warunki. Zostają nam dwa *stopnie swobody*, które można wyzyskać na różne sposoby.

Znajdziemy teraz wzór na $S_i(x)$ w przedziale $[t_i, t_{i+1})$. Wprowadzamy pomocniczą wielkość $z_i = S''(t_i)$. Ponieważ $S_i \in \Pi_3$, więc funkcja S''_i jest liniowa. Z równości $S''_i(t_i) = z_i$, $S''_i(t_{i+1}) = z_{i+1}$ wynika zatem, że

$$S''_i(x) = \frac{z_i}{h_i}(t_{i+1} - x) + \frac{z_{i+1}}{h_i}(x - t_i),$$

gdzie $h_i = t_{i+1} - t_i$. Całkując dwukrotnie obie strony tej równości, otrzymujemy wielomian S_i:

$$S_i(x) = \frac{z_i}{6h_i}(t_{i+1} - x)^3 + \frac{z_{i+1}}{6h_i}(x - t_i)^3 + C(x - t_i) + D(t_{i+1} - x). \tag{6.4.3}$$

C i D są stałymi całkowania, które wynikają z warunków interpolacyjnych $S_i(t_i) = y_i$, $S_i(t_{i+1}) = y_{i+1}$. Łatwo sprawdzić, że ostatecznie

$$\begin{aligned} S_i(x) = {} & \frac{z_i}{6h_i}(t_{i+1} - x)^3 + \frac{z_{i+1}}{6h_i}(x - t_i)^3 + \\ & + \left(\frac{y_{i+1}}{h_i} - \frac{z_{i+1}h_i}{6}\right)(x - t_i) + \left(\frac{y_i}{h_i} - \frac{z_i h_i}{6}\right)(t_{i+1} - x). \end{aligned} \tag{6.4.4}$$

Aby wyznaczyć wielkości z_i, korzystamy z warunków (6.4.2). Różniczkując wielomian (6.4.4) i podstawiając $x = t_i$, otrzymujemy równość

$$S_i'(t_i) = -\frac{h_i}{3}z_i - \frac{h_i}{6}z_{i+1} - \frac{y_i}{h_i} + \frac{y_{i+1}}{h_i}.$$

Natomiast zmiana w obliczonej pochodnej wskaźnika i na $i-1$ i podstawienie $x = t_i$ daje równość

$$S_{i-1}'(t_i) = \frac{h_{i-1}}{6}z_{i-1} + \frac{h_{i-1}}{3}z_i - \frac{y_{i-1}}{h_{i-1}} + \frac{y_i}{h_{i-1}}. \tag{6.4.5}$$

Przyrównujemy do siebie dwa otrzymane wyrażenia:

$$\begin{aligned} &h_{i-1}z_{i-1} + 2(h_{i-1} + h_i)z_i + h_i z_{i+1} = \\ &\quad = \frac{6}{h_i}(y_{i+1} - y_i) - \frac{6}{h_{i-1}}(y_i - y_{i-1}) \quad (1 \leqslant i \leqslant n-1). \end{aligned} \tag{6.4.6}$$

Mamy więc układ $n-1$ równań z $n+1$ niewiadomymi $z_0, z_1, \ldots, z_n$. Każdy wybór z_0 i z_n pozwala wyznaczyć pozostałe niewiadome. Najprościej przyjąć, że $z_0 = z_n = 0$. Daje to tzw. *naturalną* funkcję sklejaną. Jest to rzeczywiście sensowny wybór, o czym świadczy tw. 6.4.1.

Niech będzie

$$h_i = t_{i+1} - t_i, \quad u_i = 2(h_{i-1} + h_i), \quad b_i = \frac{6}{h_i}(y_{i+1} - y_i), \quad v_i = b_i - b_{i-1}.$$

Wtedy układ (6.4.6) wyraża się tak:

$$\begin{bmatrix} u_1 & h_1 & & & \\ h_1 & u_2 & h_2 & & \\ & \ddots & \ddots & \ddots & \\ & & h_{n-3} & u_{n-2} & h_{n-2} \\ & & & h_{n-2} & u_{n-1} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_{n-2} \\ z_{n-1} \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-2} \\ v_{n-1} \end{bmatrix}.$$

Do szczególnej postaci macierzy układu, która jest symetryczna, trójprzekątniowa i przekątniowo dominująca, dopasowujemy pewien wariant metody eliminacji Gaussa; wybór elementów głównych nie jest tu potrzebny:

input $n, (t_i), (y_i)$
for $i = 0$ **to** $n-1$ **do**
    $h_i \leftarrow t_{i+1} - t_i$
    $b_i \leftarrow 6(y_{i+1} - y_i)/h_i$
end do
$u_1 \leftarrow 2(h_0 + h_1)$

$v_1 \leftarrow b_1 - b_0$
for $i = 2$ **to** $n-1$ **do**
    $u_i \leftarrow 2(h_{i-1} + h_i) - h_{i-1}^2/u_{i-1}$
    $v_i \leftarrow b_i - b_{i-1} - h_{i-1}v_{i-1}/u_{i-1}$
end do
$z_n \leftarrow 0$
for $i = n-1$ **to** 1 **step** -1 **do**
    $z_i \leftarrow (v_i - h_i z_{i+1})/u_i$
end do
$z_0 \leftarrow 0$
output (z_i)

Wartości u_i, v_i obliczane dla $i > 1$ nie są identyczne z tymi, które występują w układzie równań. Ponieważ w algorytmie występuje dzielenie przez u_i, trzeba wykazać, że $u_i \neq 0$. Indukcyjnie udowodnimy więcej – mianowicie, że $u_i > h_i$, a wiemy, że $h_i > 0$. Dla $i = 1$ jest $u_1 = 2(h_0 + h_1) > h_1$. Jeśli $u_{i-1} > h_{i-1}$, to

$$u_i = 2(h_{i-1} + h_i) - \frac{h_{i-1}^2}{u_{i-1}} > 2(h_{i-1} + h_i) - h_{i-1} > h_i.$$

Po obliczeniu wielkości $z_0, z_1, \ldots, z_n$ można znaleźć wartość funkcji sklejanej S w dowolnym punkcie x. Najpierw należy ustalić, do którego z przedziałów

$$(-\infty, t_1), \quad [t_1, t_2), \quad \ldots, \quad [t_{n-2}, t_{n-1}), \quad [t_{n-1}, \infty)$$

x należy. W tym celu badamy kolejno różnice

$$x - t_{n-1}, \quad x - t_{n-2}, \quad \ldots, \quad x - t_1.$$

Jeśli znajdziemy pierwszą z nich, np. $x - t_i$, która jest nieujemna, to $x \in [t_i, t_{i+1})$. Jeśli wszystkie te liczby są ujemne, to $x \in (-\infty, t_1)$ i przyjmujemy $i = 0$. W ten sposób znajdujemy wskaźnik właściwego wielomianu S_i i obliczamy jego wartość $S_i(x)$ [4]. Przy tym zamiast (6.4.4) stosujemy prostszy wzór

$$S_i(x) = y_i + (x - t_i)\{C_i + (x - t_i)[B_i + (x - t_i)A_i]\}, \tag{6.4.7}$$

gdzie

$$A_i := \frac{1}{6h_i}(z_{i+1} - z_i), \quad B_i := \frac{z_i}{2}, \quad C_i := -\frac{h_i}{6}(z_{i+1} + 2z_i) + \frac{1}{h_i}(y_{i+1} - y_i).$$

[4] Taki sposób wyznaczania i jest sensowny tylko dla niewielkich n, bo wymaga zbadania średnio około $n/2$ różnic $x - t_j$. Godny polecenia jest inny, znany sposób: jeśli np. $n = 17$, to najpierw porównujemy x z t_9, później – zależnie od wyniku – z t_5 albo z t_{13} itd. Daje to wskaźnik i najwyżej po czterech porównaniach (*przyp. tłum.*).

W zadaniu 12 sprawdza się poprawność tego wzoru.

Niżej podano wyniki prostego testu, w którym zastosowano zaproponowany algorytm. Dla funkcji $f(x) = \sqrt{x}$ i 10 węzłów $0.25i$ $(0 \leqslant i \leqslant 9)$ wyznaczono funkcję sklejaną sześcienną S i obliczono różnice $E(x) = S(x) - f(x)$ w 37 punktach $0.0625j$ $(0 \leqslant j \leqslant 36)$:

x	$\lvert E(x)\rvert$
0	0
0.0625	$1.07321_{10}-1$
0.125	$7.52666_{10}-2$
0.1875	$3.32617_{10}-2$
0.25	0
.....	
1.75	0
1.8125	$3.64780_{10}-5$
1.875	$6.36578_{10}-5$
1.9375	$5.85318_{10}-5$
2	0
2.0625	$1.14083_{10}-4$
2.125	$2.12312_{10}-4$
2.1875	$2.04682_{10}-4$
2.25	0

Maksymalna wartość $|E(x)|$ równa 0.11 jest osiągnięta w przedziale $(0, 0.25)$ (warto zastanowić się, dlaczego właśnie tam).

Poznamy teraz twierdzenie, z którego wynika, że – w pewnym sensie – naturalna funkcja sklejana sześcienna jest *najgładszą* funkcją interpolującą.

Twierdzenie 6.4.1. *Jeśli $f \in C^2[a,b]$, $a = t_0 < t_1 < \ldots < t_n = b$, a S jest naturalną funkcją sklejaną sześcienną, interpolującą f w węzłach t_i dla $0 \leqslant i \leqslant n$, to*

$$\int_a^b [S''(x)]^2\,dx \leqslant \int_a^b [f''(x)]^2\,dx. \tag{6.4.8}$$

Dowód. Niech będzie $g := f - S$. Stąd

$$\int_a^b (f'')^2\,dx = \int_a^b (S'')^2\,dx + \int_a^b (g'')^2\,dx + 2\int_a^b S''g''\,dx.$$

Wystarczy udowodnić, że ostatnia całka po prawej stronie znika. Całkując przez części uwzględniamy, że: $S''(t_0) = S''(t_n) = 0$, $g(t_i) = 0$ dla $0 \leqslant i \leqslant n$ oraz to, że S''' jest przedziałami stała, np. równa c_i w $[t_{i-1}, t_i)$:

$$\int_a^b S''g''\,dx = \sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} S''g''\,dx =$$

$$= \sum_{i=1}^{n} \Big[(S''g')(t_i) - (S''g')(t_{i-1}) - \int_{t_{i-1}}^{t_i} S'''g'\,dx\Big] =$$

$$= -\sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} S'''g'\,dx = -\sum_{i=1}^{n} c_i \int_{t_{i-1}}^{t_i} g'\,dx =$$

$$= -\sum_{i=1}^{n} c_i[g(t_i) - g(t_{i-1})] = 0. \qquad \blacksquare$$

Przypomnijmy, że *krzywizna* krzywej o równaniu $y = f(x)$ jest w punkcie x równa

$$|f''(x)|\{1 + [f'(x)]^2\}^{-3/2}.$$

Jeśli pierwsza pochodna funkcji f jest niezbyt duża, to $|f''(x)|$ dość dobrze aproksymuje krzywiznę. Wtedy zgodnie z tw. 6.4.1 funkcja sklejana S ma w przybliżeniu najmniejszą średnio krzywiznę w przedziale $[a, b]$.

Zauważmy też, że obliczając ostatnią całkę, otrzymujemy m.in. sumę

$$\sum_{i=1}^{n}[(S''g')(t_i) - (S''g')(t_{i-1})] = (S''g')(b) - (S''g')(a).$$

Dlatego nierówność (6.4.8) pozostaje prawdziwa, gdy ta różnica jest nieujemna, tzn. gdy

$$S''(b)[f'(b) - S'(b)] \geqslant S''(a)[f'(a) - S'(a)].$$

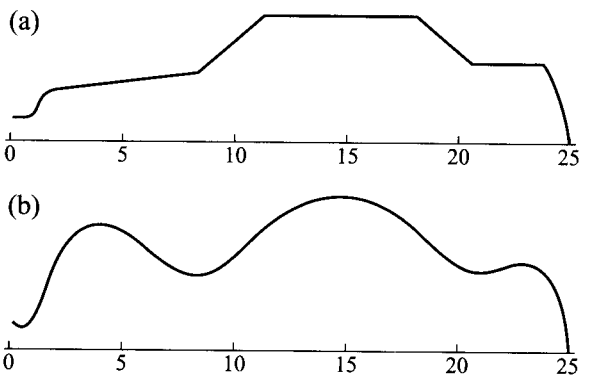
Tak jest np. wtedy, gdy warunki dające naturalną funkcję sklejaną, zastąpimy warunkami brzegowymi $S'(a) = f'(a)$ i $S'(b) = f'(b)$ [5].

Funkcje sklejane hiperboliczne

W pewnych zadaniach wygładzania danych jest celowe wprowadzenie parametru τ zwanego *naprężeniem*. Im τ jest większe, tym bardziej naprężona jest krzywa, która przechodzi przez dane punkty, a więc tym bardziej

[5] Ta zmiana daje inne twierdzenie: z tw. 6.4.1 wynika, że całka $\int_a^b [f''(x)]^2\,dx$, gdzie funkcja f należy do $C^2[a, b]$ i ma ustalone wartości w węzłach, jest minimalna, gdy f jest naturalną funkcją sklejaną interpolującą, natomiast przyjęcie powyższych warunków brzegowych oznacza, że nierówność dotyczy podklasy funkcji f z dodatkowo ustalonymi wartościami $f'(a)$ i $f'(b)$ i funkcji sklejanej spełniającej te same warunki. W tej podklasie minimum wspomnianej całki jest na ogół większe niż w tw. 6.4.1 (*przyp. tłum.*).

przypomina ona łamaną, czyli wykres funkcji sklejanej stopnia pierwszego. Natomiast dla małych τ jest ona bardziej podobna do wykresu podobnej funkcji, ale sześciennej. Ilustrują to rys. 6.5(a) i 6.5(b).



RYS. 6.5. (a) Krzywa sklejana mocno naprężona ($\tau = 10$). (b) Krzywa sklejana mało naprężona ($\tau = 0.1$)

Dla danego τ i danych węzłów takich, że

$$t_0 < t_1 < \ldots < t_n,$$

funkcję sklejaną hiperboliczną[6] f określamy przyjmując, że:

1. W przedziale $[t_0, t_n]$ istnieje czwarta pochodna $f^{(4)}$.
2. Są spełnione warunki interpolacyjne $f(t_i) = y_i$ dla $0 \leqslant i \leqslant n$.
3. W każdym z przedziałów otwartych (t_{i-1}, t_i) jest $f^{(4)} - \tau^2 f'' = 0$.

Oczywiście dla $\tau = 0$ te żądania określają zwykłą funkcję sklejaną sześcienną, bo równanie $f^{(4)} = 0$ jest spełnione przez każdy wielomian klasy Π_3.

Aby znaleźć funkcję f, postępujemy podobnie jak przedtem. Przyjmujemy więc $z_i = f''(t_i)$ i formułujemy warunki, jakie mają być spełnione w przedziale $[t_i, t_{i+1}]$:

$$f^{(4)} - \tau^2 f'' = 0,$$
$$f(t_i) = y_i, \quad f''(t_i) = z_i, \quad f(t_{i+1}) = y_{i+1}, \quad f''(t_{i+1}) = z_{i+1}.$$

Można sprawdzić, że to zadanie brzegowe ma następujące rozwiązanie:

$$f(x) = \{z_i \sinh[\tau(t_{i+1} - x)] + z_{i+1} \sinh[\tau(x - t_i)]\}/[\tau^2 \sinh(\tau h_i)] +$$
$$+ (y_i - z_i/\tau^2)(t_{i+1} - x)/h_i + (y_{i+1} - z_{i+1}/\tau^2)(x - t_i)/h_i. \quad (6.4.9)$$

[6] W oryginale *tension spline*. Dosłowny przekład wydaje się niezgrabny, natomiast tu użyty wynika oczywiście z (6.4.9) (*przyp. tłum.*).

Jak przedtem, $h_i = t_{i+1} - t_i$. I tym razem trzeba jeszcze wyznaczyć wielkości z_i. Ciągłość pierwszej pochodnej wynikająca z warunku **2** pociąga za sobą równości

$$\lim_{x \nearrow x_i} f'(x) = \lim_{x \searrow x_i} f'(x) \quad (1 \leqslant i \leqslant n-1).$$

Pominiemy tu szczegóły dowodu, że wynika stąd układ równań

$$\alpha_{i-1} z_{i-1} + (\beta_{i-1} + \beta_i) z_i + \alpha_i z_{i+1} = \gamma_i - \gamma_{i-1} \quad (1 \leqslant i \leqslant n-1), \tag{6.4.10}$$

gdzie

$$\alpha_i := 1/h_i - \tau/\sinh(\tau h_i), \qquad \beta_i := \tau \operatorname{ctgh}(\tau h_i) - 1/h_i,$$
$$\gamma_i := \tau^2 (y_{i+1} - y_i)/h_i.$$

Jak dla naturalnych funkcji sklejanych, musimy narzucić jeszcze dwa warunki na parametry z_i; możemy, jak przedtem, przyjąć, że $z_0 = z_n = 0$.

Znalezienie funkcji sklejanej hiperbolicznej dla danych punktów (t_i, y_i) można opisać tak:

1. Upewnić się, że $t_0 < t_1 < \ldots < t_n$.
2. Obliczyć h_i, α_i, β_i, γ_i dla $0 \leqslant i \leqslant n-1$.
3. Rozwiązać układ trójprzekątniowy (6.4.10) przyjmując, że $z_0 = z_n = 0$.

Wartości funkcji oblicza się ze wzoru (6.4.9) dla odpowiedniego i.

Stosując się do powyższych wskazówek, można testować skutki wyboru różnych wartości parametru τ. Tak właśnie, dla $n = 25$ i $t_i = i$, otrzymano rys. 6.5(a) i 6.5(b).

Funkcje sklejane hiperboliczne wprowadził Schweikert [1966]; zob. też Cline [1974a, b] i Pruess [1976, 1978]. Cline opracował programy konstrukcji krzywych i powierzchni za pomocą takich funkcji. Alternatywnym narzędziem są funkcje sklejane *napięte* (*taut splines*) de Boora [1984]. Są to zwykłe funkcje sklejane sześcienne z dodatkowymi punktami interpolacji wybieranymi tam, gdzie krzywa ma zmieniać gwałtownie kierunek. Zaletą tego wariantu jest to, że nie wymaga on odrębnego oprogramowania i pozwala uniknąć kłopotliwego obliczania wartości funkcji hiperbolicznych.

Naturalne funkcje sklejane wyższych stopni

Naturalne funkcje sklejane można określić także dla wyższych stopni, ale tylko nieparzystych. Niech $2m+1$ będzie tym stopniem. Przypadek $m = 1$ już opisano. Ogólnie, dla danego układu węzłów takich, że

$$t_0 < t_1 < \ldots < t_n,$$

naturalną funkcją sklejaną stopnia $2m+1$ jest funkcja $S \in C^{2m}(\mathbb{R})$, która w każdym z przedziałów $[t_{i-1}, t_i]$ $(1 \leqslant i \leqslant n)$ jest wielomianem klasy Π_{2m+1}, natomiast na półprostych $(-\infty, t_0)$ i (t_n, ∞) jest wielomianem klasy Π_m (a więc funkcją liniową, gdy $m = 1$). Przestrzeń tak określonych funkcji oznaczamy symbolem $\mathcal{N}_n^{2m+1}$.

W twierdzeniach dotyczących tych funkcji używamy symbolu x_+^n do oznaczenia *obciętej funkcji potęgowej*. Z definicji jest ona równa x^n dla $x \geqslant 0$ i równa 0 dla $x < 0$. Funkcja należy do klasy C^{n-1}.

Twierdzenie 6.4.2. *Każda funkcja z przestrzeni* $\mathcal{N}_n^{2m+1}$ *wyraża się w postaci*

$$S(x) = \sum_{i=0}^{m} a_i x^i + \sum_{j=0}^{n} b_j (x - t_j)_+^{2m+1},$$

gdzie b_j *są takie, że* $\sum_{j=0}^{n} b_j t_j^i = 0$ *dla* $0 \leqslant i \leqslant m$.

Dowód. Na półprostej $(-\infty, t_0)$ funkcja S jest z definicji pewnym wielomianem p_0 klasy Π_m. Jest nim suma $\sum_{i=0}^{m} a_i x^i$. W przedziale (t_0, t_1) funkcja S ma być wielomianem $p_1 \in \Pi_{2m+1}$. Z warunków ciągłości w t_0 wynika, że

$$p_0^{(i)}(t_0) = p_1^{(i)}(t_0) \quad (0 \leqslant i \leqslant 2m).$$

Korzystając ze wzoru Taylora, wyrażamy wielomian p_1 w postaci

$$p_1(x) = \sum_{i=0}^{2m+1} \frac{1}{i!} p_1^{(i)}(t_0)(x - t_0)^i =$$

$$= \sum_{i=0}^{2m} \frac{1}{i!} p_0^{(i)}(t_0)(x - t_0)^i + b_0(x - t_0)^{2m+1} = p_0(x) + b_0(x - t_0)^{2m+1}.$$

Dlatego na półprostej $(-\infty, t_1)$ jest

$$S(x) = p_0(x) + b_0(x - t_0)_+^{2m+1}$$

(zauważmy przy okazji, że $S^{(i)}(t_0) = 0$ dla $m < i \leqslant 2m$). Rozumując tak samo dalej, otrzymujemy pozostałe składniki $b_j(x - t_j)_+^{2m+1}$. Na półprostej (t_n, ∞) funkcja S musi być wielomianem klasy Π_m, a więc musi tam znikać tożsamościowo jej $(m+1)$-sza pochodna:

$$S^{(m+1)}(x) = \sum_{j=0}^{n} (2m+1)(2m)\ldots(m+1) b_j (x - t_j)^m \equiv 0.$$

Wiadomo, że

$$\sum_{j=0}^{n} b_j (x - t_j)^m = \sum_{j=0}^{n} b_j \sum_{i=0}^{m} \binom{m}{i} (-t_j)^i x^{m-i}.$$

Przyrównując do 0 współczynniki tego wielomianu, otrzymujemy układ równań podany w twierdzeniu. ■

Twierdzenie 6.4.3. *Jeśli $0 \leqslant m \leqslant n$, to istnieje dokładnie jedna naturalna funkcja sklejana $S \in \mathcal{N}_n^{2m+1}$, mająca w danych węzłach t_i dane wartości λ_i.*

Dowód. Na mocy tw. 6.4.2 parametry a_i, b_j funkcji S powinny spełniać następujące warunki:

$$S(t_k) = \sum_{i=0}^{m} a_i t_k^i + \sum_{j=0}^{n} b_j (t_k - t_j)_+^{2m+1} = \lambda_k \quad (0 \leqslant k \leqslant n),$$

$$\sum_{j=0}^{n} b_j t_j^i = 0 \quad (0 \leqslant i \leqslant m).$$

Jest to układ $m+n+2$ równań z tyluż niewiadomymi. Aby sprawdzić, że jego macierz jest nieosobliwa, dowodzimy, iż odpowiedni układ jednorodny ma tylko zerowe rozwiązanie. Przyjmijmy zatem, że $S(t_k) = 0$ dla $0 \leqslant k \leqslant n$. Wykażemy, że

$$I := \int_a^b [S^{(m+1)}(x)]^2 \, dx = 0, \tag{6.4.11}$$

gdzie $a = t_0$, $b = t_n$. Całkowanie przez części daje

$$I = S^{(m+1)}(x) S^{(m)}(x) \Big|_a^b - \int_a^b S^{(m)}(x) S^{(m+2)}(x) \, dx =$$

$$= - \int_a^b S^{(m)}(x) S^{(m+2)}(x) \, dx.$$

Wykorzystano tu równości $S^{(m+1)}(a) = S^{(m+1)}(b) = 0$ wynikające stąd, że poza przedziałem (a, b) funkcja S jest wielomianem klasy Π_m. Powtarzając całkowanie przez części, wnioskujemy, że

$$I = (-1)^m \int_a^b S'(x) S^{(2m+1)}(x) \, dx.$$

Ponieważ pochodna $S^{(2m+1)}$ jest przedziałami stała, więc

$$I = (-1)^m \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} c_j S'(x)\,dx = (-1)^m \sum_{j=1}^{n} c_j [S(t_j) - S(t_{j-1})] = 0.$$

Stąd wynika (6.4.11), a więc i to, że $S^{(m+1)} \equiv 0$. Dlatego S jest wielomianem klasy Π_m. Wiemy też jednak, że S ma zera $t_0, t_1, \ldots, t_n$. Ponieważ $n+1 > m$, więc $S \equiv 0$. ■

Twierdzenie 6.4.1 uogólnia się na funkcje sklejane dowolnych stopni nieparzystych:

Twierdzenie 6.4.4. *Jeśli $f \in C^{m+1}[a,b]$ i $a = t_0 < t_1 < \ldots < t_n = b$, gdzie $m \leqslant n$, a S jest naturalną funkcją sklejaną klasy $\mathcal{N}_n^{2m+1}$, interpolującą f w węzłach t_i dla $0 \leqslant i \leqslant n$, to*

$$\int_a^b [S^{(m+1)}(x)]^2\,dx \leqslant \int_a^b [f^{(m+1)}(x)]^2\,dx.$$

Dowód jest taki sam jak dla tw. 6.4.1.

Zadania 6.4

1. Znaleźć $\int_0^1 S(x)\,dx$, gdzie S jest funkcją sklejaną stopnia pierwszego, interpolującą f w węzłach takich, że $0 = t_0 < t_1 < \ldots < t_n = 1$.
2. Wykazać, że jeśli węzły są takie, że $t_0 < t_1 < \ldots < t_n$, to funkcja sklejana stopnia pierwszego wyraża się wzorem $S(x) = ax + b + \sum_{i=1}^{n-1} c_i|x - t_i|$.
3. Sprawdzić, czy wzór

$$f(x) := \begin{cases} x, & x \in (-\infty, 1] \\ -\frac{1}{2}(2-x)^2 + \frac{3}{2}, & x \in [1,2] \\ \frac{3}{2}, & x \in [2,\infty) \end{cases}$$

określa funkcję sklejaną stopnia drugiego.

4. Wzorując się na rozważaniach dotyczących funkcji sklejanych sześciennych, rozważyć konstrukcję takiej funkcji Q stopnia drugiego dla danych (t_i, y_i) $(0 \leqslant i \leqslant n)$, gdzie $t_0 < t_1 < \ldots < t_n$. Znaleźć równania spełnione przez wielkości $z_i := Q'(t_i)$. Powinno się okazać, że jedną z nich można wybrać dowolnie.
5. Sprawdzić, czy naturalna funkcja sklejana sześcienna, interpolująca punkty

x	0	1	2	3
y	1	1	0	10

jest określona wzorem

$$f(x) := \begin{cases} 1 + x - x^3, & x \in [0,1] \\ 1 - 2(x-1) - 3(x-1)^2 + 4(x-1)^3, & x \in [1,2] \\ 4(x-2) + 9(x-2)^2 - 3(x-2)^3, & x \in [2,3]. \end{cases}$$

6. Znaleźć naturalną funkcję sklejaną sześcienną dla następujących układów punktów (x_i, y_i):

(a)
x	0	1	4
y	26	7	25

(b)
x	-1	0	1
y	5	7	9

(c)
x	-1	0	1
y	13	7	9

7. Sprawdzić, czy następujące wzory określają naturalne funkcje sklejane sześcienne:

(a) $$f(x) := \begin{cases} 2(x+1)+(x+1)^3, & x \in [-1,0] \\ 3+5x+3x^2, & x \in [0,1] \\ 11+11(x-1)+3(x-1)^2-(x-1)^3, & x \in [1,2] \end{cases}$$

(b) $$f(x) := \begin{cases} x+1+(x+1)^3, & x \in [-1,0] \\ 4+(x-1)+(x-1)^3, & x \in [0,1] \end{cases}$$

8. Sprawdzić dla: (a) $t_1 = \frac{1}{2}$, (b) $t_1 = 0$, czy funkcja

$$S(x) := \begin{cases} x^3-1, & x \in [-1,t_1] \\ 3x^3-1, & x \in [t_1,1] \end{cases}$$

jest naturalną funkcją sklejaną sześcienną.

9. Sprawdzić, dla jakich wartości parametrów występujących w definicji poniższych funkcji te ostatnie są funkcjami sklejanymi sześciennymi:

(a) $$f(x) := \begin{cases} a(x-2)^2+b(x-1)^3, & x \in (-\infty,1] \\ c(x-2)^2, & x \in [1,3] \\ d(x-2)^2+e(x-3)^3, & x \in [3,\infty) \end{cases}$$

(b) $$f(x) := \begin{cases} 1-2x, & x \in (-\infty,-3] \\ a+bx+cx^2+dx^3, & x \in [-3,4] \\ 157-32x, & x \in [4,\infty) \end{cases}$$

(c) $$f(x) := \begin{cases} (x-2)^3+a(x-1)^2, & x \in (-\infty,2] \\ (x-2)^3-(x-3)^2, & x \in [2,3] \\ (x-3)^3+b(x-2)^2, & x \in [3,\infty) \end{cases}$$

10. Sprawdzić, dla jakich wartości parametrów występujących w definicji poniższych funkcji te ostatnie są (naturalnymi) funkcjami sklejanymi sześciennymi dla trzech węzłów, które są końcami podanych przedziałów.

(a) $$f(x) := \begin{cases} 3+x-9x^2, & x \in [0,1] \\ a+b(x-1)+c(x-1)^2+d(x-1)^3, & x \in [1,2] \end{cases}$$

(b) $$f(x) := \begin{cases} x^3, & x \in [-1,0] \\ a+bx+cx^2+dx^3, & x \in [0,1] \end{cases}$$

(c) $$f(x) := \begin{cases} x^3, & x \in [0,1] \\ \frac{1}{2}(x-1)^3+a(x-1)^2+b(x-1)+c, & x \in [1,3] \end{cases}$$

11. Sprawdzić poprawność wzorów (6.4.4), (6.4.5) i (6.4.6).

12. Zastępując w (6.4.3) t_{i+1} przez $t_i + h_i$, przekształcając $(x - t_i - h_i)^3$ i stosując właściwe wartości dla C i D, udowodnić wzór (6.4.7).

13. Wielkości z_i obliczane za pomocą algorytmu podanego w tekście są takie, że $E_i := u_i z_i + h_i z_{i+1} - v_i = 0$. Sprawdzić poprawność algorytmu wykazując, że $(h_i/u_i)E_i + E_{i+1} = 0$ i że to równanie można sprowadzić do (6.4.6).

14. Sprawdzić, że wielkości u_i tworzone w algorytmie są takie, że $u_i > h_{i-1} + 2h_i$ dla $1 \leqslant i \leqslant n-1$.

15. Uprościć algorytm podany w tekście dla przypadku, gdy węzły są równoodległe.

16. Zmodyfikować podany w tekście algorytm, zastępując warunek $S''(t_0) = S''(t_n) = 0$ ustaleniem wartości $S'(t_0)$ i $S'(t_n)$.

17. Udowodnić, że funkcja (6.4.9) jest rozwiązaniem podanego wcześniej zadania brzegowego.

18. Wykazać, że układ (6.4.10) wynika z żądania ciągłości pierwszej pochodnej funkcji f.

19. Udowodnić, że macierz układu (6.4.10) jest dominująca przekątniowo.

Zadania komputerowe 6.4

K1. W zadaniu 4 wspomniano, że jeden z parametrów z_i można wybrać dowolnie. Znaleźć taki sposób ich określania, aby suma $\sum_{i=0}^{n} z_i^2$ była najmniejsza. Zastosować wynik w programie i sprawdzić jego skutki.

K2. Udowodnić, że dla wielomianu S_i z definicji funkcji sklejanej sześciennej jest

$$\int_{t_i}^{t_{i+1}} S_i(x)\,dx = \tfrac{1}{2}h_i(y_i + y_{i+1}) - \tfrac{1}{24}h_i^3(z_i + z_{i+1}).$$

Napisać i sprawdzić program obliczający całkę $\int_{t_0}^{t_n} S(x)\,dx$.

K3. (a) Narysować na papierze milimetrowym krzywą, np. owal lub spiralę. Wybrać na niej w regularnych odstępach punkty i przypisać im wartości $t_i = i$. Niech współrzędnymi tych punktów będą x_i, y_i. Interpolować układy punktów (t_i, x_i) i (t_i, y_i) odpowiednio za pomocą funkcji sklejanych sześciennych $S_x(t)$ i $S_y(t)$. Wzory $x = S_x(t)$ i $y = S_y(t)$ dają przybliżenie parametryczne danej krzywej. Sprawdzić skutki takiego postępowania.

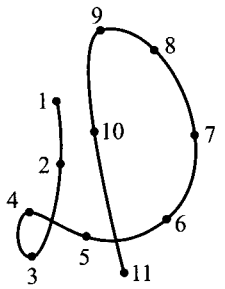
(b) Tę samą metodę zastosować do fantazyjnej litery z rys. 6.6 z zaznaczonymi na niej 11 węzłami. Knuth [1979] wyjaśnia, jak w projektowaniu różnych krojów pisma stosuje się krzywe sklejane.

K4. Rozważyć skutki następujących eksperymentów numerycznych:

(a) Niech $p \in \Pi_{20}$ interpoluje funkcję $f(x) = (1 + 6x^2)^{-1}$ w 21 punktach $0.1i$ $(-10 \leqslant i \leqslant 10)$. Obliczyć i wydrukować wartości $f(x)$, $p(x)$, $f(x) - p(x)$ w punktach $0.05j$ $(-20 \leqslant j \leqslant 20)$.

(b) Zrobić to samo, ale dla węzłów $x_i = \cos(\pi i/20)$ $(0 \leqslant i \leqslant 20)$.

(c) Dla funkcji i węzłów jak w (a) znaleźć interpolującą funkcję sklejaną sześcienną.



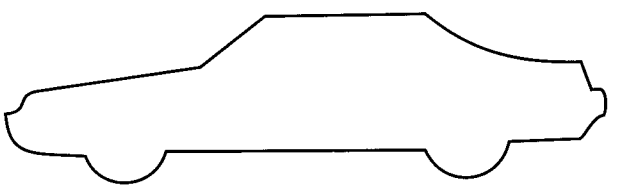
RYS. 6.6. Litera i węzły

K5. Napisać i sprawdzić program obliczania funkcji sklejanej sześciennej S, która dla danego układu węzłów takich, że $t_0 < t_1 < \ldots < t_n$, spełnia warunki

$$S(t_i) = y_i \quad (0 \leqslant i \leqslant n), \quad S''(t_0) = \alpha, \quad S''(t_n) = \beta.$$

K6. Napisać program obliczający funkcję sklejaną hiperboliczną i sprawdzić jego działanie dla wielu wartości parametru τ.

K7. Przygotować tablicę współrzędnych od 10 do 20 punktów na konturze samochodu (rys. 6.7). Dla wartości 0.25, 4, 10 parametru τ obliczyć i narysować krzywe interpolujące ten kontur, a korzystające z funkcji sklejanych hiperbolicznych. Sprawdzić, jakie τ daje najlepsze wyniki.



RYS. 6.7. Kontur samochodu

6.5. Podstawy teorii funkcji B-sklejanych

W tym podrozdziale zajmujemy się układem bazowych funkcji sklejanych, tj. takich, że każda funkcja sklejana (z pewnej przestrzeni) jest ich kombinacją liniową. Stąd nazwa: *funkcje B-sklejane*. Dla danego układu węzłów można je łatwo tworzyć za pomocą związków rekurencyjnych. Teoria tych funkcji jest elegancka, a własności numeryczne są wzorowe. Można też te funkcje uogólniać.

Z teoretycznego punktu widzenia będzie wygodnie założyć, że na prostej rzeczywistej mamy układ nieskończony węzłów:

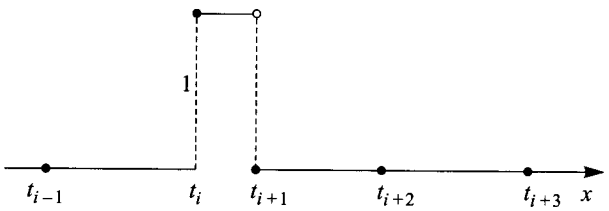
$$\ldots < t_{-2} < t_{-1} < t_0 < t_1 < t_2 < \ldots,$$

taki, że $\lim_{i\to-\infty} t_i = -\infty$, $\lim_{i\to\infty} t_i = \infty$. Ten układ jest ustalony w całym podrozdziale i konstruowane funkcje sklejane bazują na nim.

Funkcje B-sklejane stopnia 0

Funkcje B-sklejane stopnia 0 oznaczamy symbolem B_i^0, gdzie wskaźnik i przebiega zbiór $\mathbb{Z}$ wszystkich liczb całkowitych. Przykładową funkcję tego typu pokazuje rys. 6.8. Kropki • oznaczają tam, że w szczególności przyjmujemy $B_i^0(t_i) = 1$ i $B_i^0(t_{i+1}) = 0$. Pełna definicja jest następująca:

$$B_i^0(x) := \begin{cases} 1 & (t_i \leqslant x < t_{i+1}) \\ 0 & (x < t_i \text{ lub } x \geqslant t_{i+1}). \end{cases}$$



RYS. 6.8. Funkcja B-sklejana B_i^0

Oto kilka istotnych własności tych funkcji:

1. *Nośnik* funkcji B_i^0, czyli zbiór tych x, gdzie $B_i^0(x) \neq 0$, jest przedziałem $[t_i, t_{i+1})$.
2. $B_i^0(x) \geqslant 0$ dla wszystkich i, x.
3. B_i^0 jest ciągła prawostronnie na całej prostej rzeczywistej.
4. $\sum_{i=-\infty}^{\infty} B_i^0(x) = 1$ dla każdego x.

Trzeba też podkreślić, że funkcje B_i^0 są bazą wszystkich funkcji sklejanych stopnia 0 (związanych z ustalonymi węzłami), jeśli tylko przyjmiemy, że i one są ciągłe prawostronnie. Istotnie, jeśli S jest taką funkcją, czyli jest przedziałami stała:

$$S(x) = c_i \quad \text{dla} \quad t_i \leqslant x < t_{i+1} \quad (i \in \mathbb{Z}),$$

to $S(x) = \sum_{i=-\infty}^{\infty} c_i B_i^0(x)$. Funkcje B_i^0 tworzą więc bazę w sensie Schaudera: każda funkcja sklejana ma jednoznaczne wyrażenie przez taki szereg nieskończony.

Funkcje B_i^0 stanowią punkt wyjścia do rekurencyjnego określania funkcji B-sklejanych wyższych stopni. Służy do tego wzór

$$B_i^k(x) := \frac{x - t_i}{t_{i+k} - t_i} B_i^{k-1}(x) + \frac{t_{i+k+1} - x}{t_{i+k+1} - t_{i+1}} B_{i+1}^{k-1}(x) \quad (k \geqslant 1). \quad (6.5.1)$$

Wprowadźmy jeszcze pomocnicze funkcje liniowe

$$V_i^k(x) := \frac{x - t_i}{t_{i+k} - t_i}.$$

Daje to bardziej elegancki wzór rekurencyjny:

$$B_i^k = V_i^k B_i^{k-1} + (1 - V_{i+1}^k) B_{i+1}^{k-1}. \tag{6.5.2}$$

Z tego wzoru wynikają wszystkie własności funkcji B-sklejanych wyższych stopni. Zauważmy, że ponieważ B_i^0 jest funkcją przedziałami stałą, a V_i^k – funkcją liniową, więc indukcyjnie dowodzi się, że B_i^k jest w każdym z tych przedziałów wielomianem klasy Π_k.

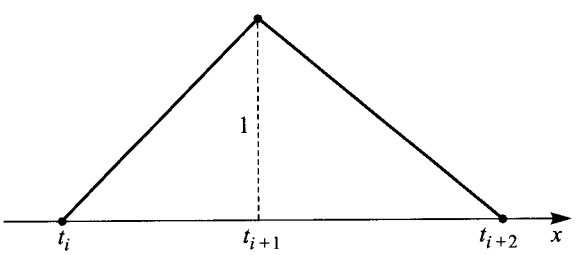
Funkcje B-sklejane stopnia 1

Ze wzoru (6.5.1) wynika, że

$$B_i^1(x) = \frac{x - t_i}{t_{i+1} - t_i} B_i^0(x) + \frac{t_{i+2} - x}{t_{i+2} - t_{i+1}} B_{i+1}^0(x) =$$

$$= \begin{cases} \dfrac{x - t_i}{t_{i+1} - t_i} & (t_i \leqslant x < t_{i+1}) \\ \dfrac{t_{i+2} - x}{t_{i+2} - t_{i+1}} & (t_{i+1} \leqslant x < t_{i+2}) \\ 0 & (x < t_i \text{ lub } x \geqslant t_{i+2}). \end{cases}$$

Wykres funkcji B_i^1 pokazano na rys. 6.9.



RYS. 6.9. Funkcja B-sklejana B_i^1

Pewne własności tych funkcji są niemal oczywiste:

1. Nośnikiem funkcji B_i^1 jest przedział (t_i, t_{i+2}).
2. $B_i^1(x) \geqslant 0$ dla wszystkich i, x.
3. B_i^1 jest ciągła, a różniczkowalna wszędzie poza punktami t_i, t_{i+1}, t_{i+2}.
4. $\sum_{i=-\infty}^{\infty} B_i^1(x) = 1$ dla każdego x.

Tylko własność 4 wymaga komentarza. Rozważmy dowolne x rzeczywiste. Z założenia o węzłach wynika, że istnieje takie j, iż $t_j \leqslant x < t_{j+1}$. Wtedy co najwyżej wartości $B_{j-1}(x)$ i $B_j(x)$ nie znikają. Dlatego

$$\sum_{i=-\infty}^{\infty} B_i^1(x) = B_{j-1}^1(x) + B_j^1(x) = \frac{t_{j+1} - x}{t_{j+1} - t_j} + \frac{x - t_j}{t_{j+1} - t_j} = 1.$$

Własności funkcji B-sklejanych

Poznamy teraz kilka lematów, z których wynikają ważne własności funkcji B_i^k dla $i \in \mathbb{Z}$ i $k \in \mathbb{N}$.

Lemat 6.5.1. *Jeśli $k \geqslant 1$ i $x \notin (t_i, t_{i+k+1})$, to $B_i^k(x) = 0$.*

Oczywisty dowód indukcyjny wynika z (6.5.2).

Lemat 6.5.2. *Jeśli $k \geqslant 0$ i $x \in (t_i, t_{i+k+1})$, to $B_i^k(x) > 0$.*

Dowód. Wiemy już, że lemat jest prawdziwy dla $k = 0$ i $k = 1$. Kontynuując dowód indukcyjny, przyjmujemy, że dla pewnego $k \geqslant 2$ oraz wszystkich i, x jest $B_i^{k-1}(x) \geqslant 0$; wynika to także z lem. 6.5.1. Niech będzie $t_i < x < t_{i+k+1}$. Wtedy oba ilorazy w (6.5.1) są dodatnie. Wobec założenia indukcyjnego jest $B_i^{k-1}(x) > 0$ w (t_i, t_{i+k}) i $B_{i+1}^{k-1}(x) > 0$ w (t_{i+1}, t_{i+k+1}). Dla $k \geqslant 2$ te dwa przedziały zachodzą na siebie, a w ich sumie $B_i^k(x) > 0$. ∎

Mamy nadzieję, że funkcje B_i^k tworzą bazę przestrzeni wszystkich funkcji sklejanych stopnia k. Dlatego interesują nas szeregi postaci $\sum_{i=-\infty}^{\infty} c_i B_i^k(x)$.

Lemat 6.5.3. *Dla każdego k zachodzi równość*

$$\sum_{i=-\infty}^{\infty} c_i B_i^k = \sum_{i=-\infty}^{\infty} [c_i V_i^k + c_{i-1}(1 - V_i^k)] B_i^{k-1}.$$

Dowód opiera się na tożsamości (6.5.2) i wymaga tylko elementarnego przekształcenia szeregu.

Procedura numeryczna

W lemacie 6.5.3 współczynniki c_i mogą zależeć od x. Lemat daje więc sposób obliczania wartości funkcji wyrażonej w postaci

$$f(x) = \sum_{i=-\infty}^{\infty} C_i^k(x) B_i^k(x).$$

Zakładamy, że funkcje C_i^k są dane. Jeśli

$$C_i^{k-1}(x) := C_i^k(x) V_i^k(x) + C_{i-1}^k(x)[1 - V_i^k(x)], \tag{6.5.3}$$

to ze wspomnianego lematu wynika, że

$$\sum_{i=-\infty}^{\infty} C_i^k(x) B_i^k(x) = \sum_{i=-\infty}^{\infty} C_i^{k-1}(x) B_i^{k-1}(x).$$

Iterując tę równość, wnioskujemy ostatecznie, że

$$\sum_{i=-\infty}^{\infty} C_i^k(x) B_i^k(x) = \sum_{i=-\infty}^{\infty} C_i^0(x) B_i^0(x).$$

Wiemy, że dla $t_j \leqslant x < t_{j+1}$ sumą szeregu po prawej stronie jest po prostu $C_j^0(x)$. Zauważmy jeszcze, że z (6.5.3) i z definicji funkcji V_i^k wynika równość

$$C_i^{j-1}(x) = \frac{(x - t_i) C_i^j(x) + (t_{i+j} - x) C_{i-1}^j(x)}{t_{i+j} - t_i}. \tag{6.5.4}$$

Te wszystkie uwagi uzasadniają następującą procedurę:

Algorytm 6.5.4. *Dla danych C_i^k oraz x takiego, że $t_m \leqslant x < t_{m+1}$, wartość funkcji sklejanej $S(x) := \sum_{i=-\infty}^{\infty} C_i^k B_i^k(x)$ jest równa wielkości C_m^0 obliczonej za pomocą* (6.5.4) *jako ostatni element tablicy trójkątnej*

$$\begin{array}{llll} C_m^k & C_m^{k-1} & \ldots\ C_m^1 & C_m^0 \\ C_{m-1}^k & C_{m-1}^{k-1} & \ldots\ C_{m-1}^1 & \\ \ldots\ldots\ldots\ldots & & & \\ C_{m-k+1}^k & C_{m-k+1}^{k-1} & & \\ C_{m-k}^k & & & \end{array}$$

Lemat 6.5.5. *Dla wszystkich k, x zachodzi równość*

$$\sum_{i=-\infty}^{\infty} B_i^k(x) = 1.$$

Dowód. Zaczynamy od szeregu $\sum_{i=-\infty}^{\infty} C_i^k B_i^k(x)$, w którym wszystkie C_i^k są równe 1. Dla ustalonego x z (6.5.3) wynika, że

$$C_i^{k-1} = C_i^k V_i^k + C_{i-1}^k (1 - V_i^k) = V_i^k + 1 - V_i^k = 1.$$

Rozumując tak dalej, wnioskujemy, że $C_i^0 = 1$ dla każdego i. Lemat wynika więc z własności 4 funkcji B_i^0. ■

Pochodne i całki funkcji B-sklejanych

Niech będzie

$$\alpha_i^k := \frac{1}{t_{i+k} - t_i}.$$

Jest więc niemal oczywiste, że

$$\frac{d}{dx} V_i^k(x) = \alpha_i^k, \tag{6.5.5}$$

$$\alpha_i^k V_i^{k+1} = \alpha_i^{k+1} V_i^k, \tag{6.5.6}$$

$$\alpha_{i+1}^k (1 - V_i^{k+1}) = \alpha_i^{k+1} (1 - V_{i+1}^k). \tag{6.5.7}$$

Lemat 6.5.6. *Jeśli $k \geqslant 2$, to*

$$\frac{d}{dx} B_i^k(x) = k[\alpha_i^k B_i^{k-1}(x) - \alpha_{i+1}^k B_{i+1}^{k-1}(x)]. \tag{6.5.8}$$

Dla $k = 1$ ten wzór jest prawdziwy dla każdego x różnego od t_i, t_{i+1}, t_{i+2}.

Dowód. Sprawdzenie (6.5.8) dla $k = 1, 2$ pozostawiamy czytelnikom. W dalszym ciągu dowodu indukcyjnego zakładamy prawdziwość tego wzoru dla pewnego k. Ze wzoru rekurencyjnego (6.5.2) i z tożsamości (6.5.5) wynika, że

$$\frac{d}{dx} B_i^{k+1} = \frac{d}{dx} [V_i^{k+1} B_i^k + (1 - V_{i+1}^{k+1}) B_{i+1}^k] =$$

$$= V_i^{k+1} \frac{d}{dx} B_i^k + \alpha_i^{k+1} B_i^k + (1 - V_{i+1}^{k+1}) \frac{d}{dx} B_{i+1}^k - \alpha_{i+1}^{k+1} B_{i+1}^k.$$

Pochodne wyrażamy zgodnie z założeniem indukcyjnym (6.5.8):

$$\frac{d}{dx} B_i^{k+1} = kV_i^{k+1}(\alpha_i^k B_i^{k-1} - \alpha_{i+1}^k B_{i+1}^{k-1}) + \alpha_i^{k+1} B_i^k +$$

$$+ k(1 - V_{i+1}^{k+1})(\alpha_{i+1}^k B_{i+1}^{k-1} - \alpha_{i+2}^k B_{i+2}^{k-1}) - \alpha_{i+1}^{k+1} B_{i+1}^k =$$

$$= \alpha_i^{k+1} B_i^k + k\alpha_i^k V_i^{k+1} B_i^{k-1} - \alpha_{i+1}^{k+1} B_{i+1}^k -$$
$$- k\alpha_{i+2}^k (1 - V_{i+1}^{k+1}) B_{i+2}^{k-1} - k\alpha_{i+1}^k V_i^{k+1} B_{i+1}^{k-1} +$$
$$+ k\alpha_{i+1}^k (1 - V_{i+1}^{k+1}) B_{i+1}^{k-1}.$$

Korzystając z (6.5.6) i (6.5.7), przekształcamy pewne wyrażenia występujące wyżej:

$$\alpha_i^k V_i^{k+1} B_i^{k-1} = \alpha_i^{k+1} V_i^k B_i^{k-1},$$
$$-\alpha_{i+2}^k (1 - V_{i+1}^{k+1}) B_{i+2}^{k-1} =$$
$$= -\alpha_{i+1}^{k+1} (1 - V_{i+2}^k) B_{i+2}^{k-1} - \alpha_{i+1}^k V_i^{k+1} B_{i+1}^{k-1} + \alpha_{i+1}^k (1 - V_{i+1}^{k+1}) B_{i+1}^{k-1} =$$
$$= \alpha_{i+1}^k (1 - V_i^{k+1}) B_{i+1}^{k-1} - \alpha_{i+1}^k V_{i+1}^{k+1} B_{i+1}^{k-1} =$$
$$= \alpha_i^{k+1} (1 - V_{i+1}^k) B_{i+1}^{k-1} - \alpha_{i+1}^{k+1} V_{i+1}^k B_{i+1}^{k-1}.$$

Wobec tego pochodna funkcji B_i^{k+1} wyraża się tak:

$$\frac{d}{dx} B_i^{k+1} = \alpha_i^{k+1} B_i^k + k[\alpha_i^{k+1} V_i^k B_i^{k-1} + \alpha_i^{k+1} (1 - V_{i+1}^k) B_{i+1}^{k-1}] -$$
$$- \alpha_{i+1}^{k+1} B_{i+1}^k - k[\alpha_{i+1}^{k+1} V_{i+1}^k B_{i+1}^{k-1} + \alpha_{i+1}^{k+1} (1 - V_{i+2}^k) B_{i+2}^{k-1}].$$

Sumy w nawiasach kwadratowych upraszczamy, korzystając z (6.5.2):

$$\frac{d}{dx} B_i^{k+1} = \alpha_i^{k+1} B_i^k + k\alpha_i^{k+1} B_i^k - \alpha_{i+1}^{k+1} B_{i+1}^k - k\alpha_{i+1}^{k+1} B_{i+1}^k =$$
$$= (k+1)\alpha_i^{k+1} B_i^k - (k+1)\alpha_{i+1}^{k+1} B_{i+1}^k,$$

a to należało udowodnić. ■

Lemat 6.5.7. *Dla $k \geqslant 1$ funkcje B_i^k należą do klasy $C^{k-1}(\mathbb{R})$.*

Dowód. Funkcje B_i^1 są ciągłe, czyli $B_i^1 \in C^0(\mathbb{R})$. Jeśli $B_i^k \in C^{k-1}(\mathbb{R})$, to na mocy lem. 6.5.6 do tej samej klasy należy $(d/dx)B_i^{k-1}$, a stąd już wynika, że $B_i^{k+1} \in C^k(\mathbb{R})$. ■

Korzystając z lem. 6.5.6 otrzymujemy wzór

$$\frac{d}{dx} \sum_{i=-\infty}^{\infty} c_i B_i^k(x) = k \sum_{i=-\infty}^{\infty} \frac{c_i - c_{i-1}}{t_{i+k} - t_i} B_i^{k-1}(x) \quad (k \geqslant 2) \tag{6.5.9}$$

(dla $k = 1$ jest on poprawny poza węzłami), który można stosować w różniczkowaniu numerycznym. Trzeba jednak się zastrzec, że różniczkowanie funkcji o wartościach zakłóconych przez błędy pomiaru jest bardzo ryzykowne.

Lemat 6.5.8. *Całka z funkcji B-sklejanej wyraża się wzorem*

$$\int_{-\infty}^{x} B_i^k(s)\,ds = \frac{t_{i+k+1} - t_i}{k+1} \sum_{j=i}^{\infty} B_j^{k+1}(x). \tag{6.5.10}$$

Dowód. Sprawdzamy najpierw, że pochodne obu stron tej równości są identyczne. W tym celu stosujemy (6.5.9) ze zmianą k na $k+1$, dla $c_j = 0$ dla $j < i$ i $c_j = 1$ w przeciwnym razie. Wtedy $c_j - c_{j-1}$ jest równe 1 dla $j = i$ i równe 0 dla $j \neq i$, a zatem

$$\frac{d}{dx} \sum_{j=i}^{\infty} B_i^{k+1}(x) = \frac{k+1}{t_{i+k+1} - t_i} B_i^k(x).$$

To samo wynika z dowodzonej równości (6.5.10) przez jej zróżniczkowanie stronami. Pozostaje zauważyć, że na mocy lem. 6.5.1 obie strony tej równości znikają dla $x = t_i$. ■

Dodatkowe własności

Niech $f \mid K$ będzie *zwężeniem* funkcji f do podzbioru K jej dziedziny:

$$(f \mid K)(x) = f(x) \quad (x \in K).$$

To pojęcie przydaje się w rozważaniach dotyczących funkcji sklejanych, bo każda funkcja $B_i^k \mid (t_j, t_{j+1})$ jest wielomianem. Gdy mówimy, że układ funkcji f_i jest *niezależny liniowo na* K, mamy na myśli taką niezależność, w zwykłym sensie, układu zwężeń $f_i \mid K$.

Rozważmy funkcje B-sklejane B_0^k, $B_1^k, \ldots, B_k^k$. Ich zwężenia do dowolnego przedziału $(t_\nu, t_{\nu+1})$ tworzą układ wielomianów, a dla $\nu = k$ ten układ jest bazą przestrzeni wielomianów Π_k.

Lemat 6.5.9. *Układ* $\{B_j^k, B_{j+1}^k, \ldots, B_{j+k}^k\}$ *jest niezależny liniowo w przedziale* (t_{k+j}, t_{k+j+1}).

Dowód. Dla $k = 0$ lemat jest oczywisty, bo wielomian B_j^0 jest liniowo niezależny w (t_j, t_{j+1}), gdzie nie znika. Niech będzie $k \geqslant 1$. Załóżmy, że lemat jest prawdziwy po zmianie k na $k-1$. Niech będzie $S := \sum_{i=0}^{k} c_{j+i} B_{j+i}^k$. Przypuśćmy, że $S \mid (t_{k+j}, t_{k+j+1}) = 0$. Stąd i z (6.5.9) wynika, że

$$0 = S' \mid (t_{k+j}, t_{k+j+1}) = k \sum_{i=1}^{k} \frac{c_{j+i} - c_{j+i-1}}{t_{j+i+k} - t_{j+i}} B_{j+i}^{k-1} \mid (t_{k+j}, t_{k+j+1}). \tag{6.5.11}$$

Wykorzystano tu równości $B^{k-1}_{j+k+1} = B^{k-1}_j = 0$ w przedziale (t_{k+j}, t_{k+j+1}). Z założenia indukcyjnego wynika, że układ $\{B^{k-1}_{j+1}, B^{k-1}_{j+2}, \ldots, B^{k-1}_{j+k}\}$ jest niezależny liniowo w tym samym przedziale. Dlatego w sumie po prawej stronie (6.5.11) wszystkie współczynniki $c_{j+i} - c_{j+i-1}$ znikają, czyli $c_j = c_{j+1} = \ldots = c_{j+k}$. Jeśli λ jest wspólną wartością tych liczb, to na mocy lem. 6.5.5 jest $S(x) = \lambda$ dla $x \in (t_{k+j}, t_{k+j+1})$ (w szeregu z tego lematu w przedziale (t_{k+j}, t_{k+j+1}) nie znikają tylko składniki od B^k_j do B^k_{j+k}). Ponieważ jednak założono, że S znika w tym przedziale, więc $\lambda = 0$. ■

Lemat 6.5.10. *Układ funkcji $\{B^k_{-k}, B^k_{-k+1}, \ldots, B^k_{n-1}\}$ jest niezależny liniowo w przedziale (t_0, t_n).*

Dowód. Niech będzie $S := \sum_{i=-k}^{n-1} c_i B^k_i$. Przypuśćmy, że $S \mid (t_0, t_n) = 0$. W przedziale (t_0, t_1) nie znikają tylko funkcje $B^k_{-k}, B^k_{-k+1}, \ldots, B^k_0$, więc

$$0 = S \mid (t_0, t_1) = \sum_{i=-k}^{0} c_i B^k_i \mid (t_0, t_1). \tag{6.5.12}$$

Na mocy lem. 6.5.9 układ $\{B^k_{-k}, B^k_{-k+1}, \ldots, B^k_0\}$ jest niezależny liniowo w (t_0, t_1). Dlatego wszystkie c_i występujące w (6.5.12) znikają. Jeśli jest też $c_1 = \ldots = c_{n-1} = 0$, to lemat jest udowodniony. W przeciwnym razie niech j będzie najmniejszym wskaźnikiem dodatnim, dla którego $c_j \neq 0$. Wtedy $(t_j, t_{j+1}) \subset (t_0, t_n)$. Dla dowolnego $x \in (t_j, t_{j+1})$ otrzymaliśmy sprzeczność:

$$0 = S(x) = \sum_{i=j}^{n-1} c_i B^k_i(x) = c_j B^k_j(x) \neq 0.$$

Dlatego wszystkie c_i znikają. ■

Zadania 6.5

1. Udowodnić, że wzór (6.5.8) jest poprawny dla $k = 1$ poza węzłami t_i, t_{i+1}, t_{i+2}.

2. Udowodnić, że

$$B^2_i(x) = \begin{cases} V^2_i V^1_1, & x \in [t_i, t_{i+1}) \\ V^2_1 - V^1_{i+1}(V^2_i + V^2_{i+1} - 1), & x \in [t_{i+1}, t_{i+2}) \\ (1 - V^2_{i+1})(1 - V^1_{i+2}), & x \in [t_{i+2}, t_{i+3}) \\ 0 & \text{w przeciwnym razie.} \end{cases}$$

3. Sprawdzić, że

$$B^2_j(t_i) = \frac{t_i - t_{i-1}}{t_{i+1} - t_{i-1}} \delta_{i,j+1} + \frac{t_{i+1} - t_i}{t_{i+1} - t_{i-1}} \delta_{i,j+2}.$$

4. Wykazać, że jeśli dla $h_i := t_{i+1} - t_i$ jest

$$c_{i-1}h_{i-1} + c_{i-2}h_i = y_i(h_{i-1} + h_i) \quad (i \in \mathbb{Z}),$$

to funkcja sklejana $S := \sum_{i=-\infty}^{\infty} c_i B_i^2$ spełnia warunki $S(t_j) = y_j$ $(j \in \mathbb{Z})$.

5. Znaleźć część wspólną nośników funkcji B_0^k, $B_1^k, \ldots, B_r^k$.

6. Udowodnić, że $\sum_{i=0}^{n} B_i^k(x) = 1$ dla $t_k \leqslant x \leqslant t_{k+n}$.

7. Udowodnić, że $\sum_{i=0}^{n} B_i^k(x) > 0$ dla $t_1 < x < t_{n+k+1}$.

8. Wykazać, że dla $t_m \leqslant x < t_{m+1}$ jest

$$\sum_{i=-\infty}^{\infty} c_i B_i^k(x) = \sum_{i=m-k}^{m} c_i B_i^k(x).$$

9. Wykazać, że jeśli $a_i := k^{-1}(t_{i+1} + t_{i+2} + \ldots + t_{i+k})$, to

$$\sum_{i=-\infty}^{\infty} a_i B_i^k(x) = x \quad (k \geqslant 1).$$

10. Udowodnić, że jeśli $t_i = i$ $(i \in \mathbb{Z})$, to $B_i^k(x) = B_0^k(x - t_i)$.

11. Niech będzie $U_i^0(s) := 1$, $U_i^k(s) := (t_{i+1} - s)(t_{i+2} - s) \ldots (t_{i+k} - s)$ dla $k > 0$. Udowodnić, że

$$U_i^k(s)V_i^k(x) + U_{i-1}^k(s)[1 - V_i^k(x)] = (x - s)U_i^{k-1}(s).$$

12. (cd.). Udowodnić, że

$$\sum_{i=-\infty}^{\infty} U_i^k(s)B_i^k(x) = (x - s) \sum_{i=-\infty}^{\infty} U_i^{k-1}(s)B_i^{k-1}(x).$$

13. (cd.). Udowodnić *tożsamość Marsdena*

$$\sum_{i=-\infty}^{\infty} U_i^k(s)B_i^k(x) = (x - s)^k.$$

14. (cd.). Wykazać, że każdy wielomian klasy Π_k można wyrazić w postaci $\sum_{i=-\infty}^{\infty} c_i B_i^k$.

15. Podać przykład funkcji sklejanej $\sum_{i=-\infty}^{\infty} c_i B_i^k$, która nie jest tożsamościowo równa 0, chociaż znika we wszystkich węzłach.

16. Wykazać, że jeśli w ciągu stałych c_{m-k}, $c_{m-k+1}, \ldots, c_m$ dwa sąsiednie elementy są identyczne, to funkcja sklejana $\sum_{i=-\infty}^{\infty} c_i B_i^k(x)$ jest w przedziale (t_m, t_{m+1}) wielomianem klasy Π_{k-1}.

17. Zastosować alg. 6.5.4 do dowodu, że $\sum_{i=-\infty}^{\infty} c_i B_i^k$ jest wielomianem klasy Π_k między kolejnymi węzłami.

18. Za pomocą alg. 6.5.4 podać nowe dowody lematów 6.5.1 i 6.5.2.

19. Udowodnić, że jeśli $\sum_{i=-\infty}^{\infty} c_i B_i^k(x) = 0$ dla każdego x, to wszystkie c_i znikają.

20. Udowodnić wzór (6.5.9).

21. Wykazać, że dla $k \geqslant 3$ jest

$$\frac{d^2}{dx^2} \sum_{i=-\infty}^{\infty} c_i B_i^k(x) = \\ = k(k-1) \sum_{i=-\infty}^{\infty} [(c_i - c_{i-1})\alpha_i^k - (c_{i-1} - c_{i-2})\alpha_{i-1}^k]\alpha_i^{k-1} B_i^{k-2}(x).$$

22. Udowodnić, że

$$\int_{-\infty}^{\infty} B_i^k(x)\, dx = \frac{t_{i+k+1} - t_i}{k+1}.$$

23. Udowodnić, że

$$\sup_{x \in \mathbb{R}} \Big| \sum_{i=-\infty}^{\infty} c_i B_i^k(x) \Big| \leqslant \sup_{i \in \mathbb{Z}} |c_i|.$$

24. Udowodnić, że jeśli $\sup_i |t_{i+1} - t_i| \leqslant m$, to

$$\int_{-\infty}^{\infty} \Big| \sum_{i=-\infty}^{\infty} c_i B_i^k(x) \Big| dx \leqslant m \sum_{i=-\infty}^{\infty} |c_i|.$$

25. Oszacować z góry wielkość

$$\int_{-\infty}^{\infty} \Big[\sum_{i=-\infty}^{\infty} c_i B_i^k(x) \Big]^2 dx.$$

Zadania komputerowe 6.5

K1. Napisać i sprawdzić program obliczający, dla danej funkcji f, przybliżone wartości funkcji $g(x) := \int_a^x f(t)\, dt$ dla $x \in [a, b]$ (oddzielna procedura oblicza wartości $f(x)$). Program ma najpierw znaleźć dla f naturalną funkcję sklejaną sześcienną S dla układu n węzłów równoodległych z $[a, b]$, a potem obliczać $\int_a^x S(t)\, dt$. Parametry a, b, n określa użytkownik programu.

K2. Napisać procedurę, która dla danych n, k, węzłów $t_1, t_2, \ldots, t_{n+k+1}$, współczynników $c_1, c_2, \ldots, c_n$ i danego x oblicza $f(x) := \sum_{i=1}^{n} c_i B_i^k(x)$.

K3. Przyjmując, że $t_i = i$ $(i \in \mathbb{Z})$, napisać program obliczający normy $\|B_0^k\|_\infty$ dla $1 \leqslant k \leqslant 100$. Do kontroli mogą posłużyć wartości tych norm, wynoszące 1, 0.55 i 0.410963 odpowiednio dla $k = 1, 5, 10$.

6.6. Zastosowania funkcji B-sklejanych

Zachowujemy tu oznaczenia przyjęte w poprzednim podrozdziale. Określone tam funkcje B-sklejane B_i^k chcemy teraz powiązać z funkcjami sklejanymi wprowadzonymi w podrozdz. 6.4. Każda taka funkcja z definicji należy do klasy C^{k-1}, a w poszczególnych przedziałach $[t_0, t_1]$, $[t_1, t_2], \ldots, [t_{n-1}, t_n]$, gdzie $n \geqslant 1$, jest wielomianem stopnia $\leqslant k$. Zbiór tych funkcji będziemy oznaczać symbolem $\mathcal{S}_n^k$; nie precyzujemy tu węzłów, przyjmując, że są ustalone. Zakładamy też, że dziedziną każdej funkcji z tej klasy jest przedział $[t_0, t_n]$. Do tegoż przedziału ograniczamy się, rozpatrując funkcje B_i^k. Używamy więc symbolu $B_i^k \mid [t_0, t_n]$, także wprowadzonego w podrozdz. 6.5.

Baza przestrzeni $\mathcal{S}_n^k$

Twierdzenie 6.6.1. *Układ funkcji*

$$\{B_i^k \mid [t_0, t_n]\colon \ -k \leqslant i \leqslant n-1\} \tag{6.6.1}$$

jest bazą przestrzeni $\mathcal{S}_n^k$. Jej wymiar jest równy $k+n$.

Dowód. Jest oczywiste, że funkcje z układu (6.6.1) należą do $\mathcal{S}_n^k$, bo każda z nich jest funkcją sklejaną stopnia k dla danego układu węzłów. Wymiar przestrzeni $\mathcal{S}_n^k$ nie przekracza $k+n$, gdyż każdy jej element jest kombinacją liniową

$$S(x) = \sum_{i=0}^{k} a_i x^i + \sum_{j=1}^{n-1} b_j (x - t_j)_+^k$$

$k+n$ funkcji

$$1, x, \ldots, x^k, (x-t_1)_+^k, (x-t_2)_+^k, \ldots, (x-t_{n-1})_+^k. \tag{6.6.2}$$

Wynika to z tw. 6.4.2 i jego dowodu. Wystarczy teraz przypomnieć, że na mocy lem. 6.5.10 funkcje w (6.6.1) są liniowo niezależne. ∎

W związku z powyższym dowodem trzeba podkreślić, że baza (6.6.2) jest bardzo źle uwarunkowana i w obliczeniach należy ją zastąpić bazą (6.6.1).

Macierz interpolacyjna

Gdy stosujemy funkcje sklejane w interpolacji, to jej węzły $x_1, x_2, \ldots, x_n$ (gdzie $x_1 < x_2 < \ldots < x_n$) mogą być różne od punktów t_i[7]. Interpolująca

[7] W oryginale nie ma konfliktu terminów: w interpolacji występują *nodes* x_j, a w definicji funkcji sklejanych *knots* t_i. Natomiast w języku polskim sensownym odpowiednikiem

funkcja sklejana ma mieć postać $\sum_{j=1}^{n} c_j B_j^k$. Jest to sensowne, jeśli *macierz interpolacyjna* A o elementach

$$(A)_{ij} := B_j^k(x_i) \quad (1 \leqslant i, j \leqslant n) \tag{6.6.3}$$

jest nieosobliwa. Na mocy pięknego tw. 6.6.4 Schoenberga i Whitneya jest tak wtedy i tylko wtedy, gdy wszystkie elementy przekątniowe tej macierzy są różne od 0. Rozbity na kilka części dowód jest wzorowany na pracy de Boora [1976] i uzyskanych od niego informacjach. Przedtem jednak przypomnijmy, że warunek

$$B_i^k(x_i) \neq 0 \quad (1 \leqslant i \leqslant n)$$

jest na mocy lematów 6.5.1 i 6.5.2 równoważny temu, że

$$x_i \in (t_i, t_{i+k+i}) \quad (1 \leqslant i \leqslant n), \tag{6.6.4}$$

a więc oznacza, iż węzły interpolacji x_i muszą być rozmieszczone w określony sposób wśród punktów t_i definiujących funkcje B-sklejane.

Lemat 6.6.2. *Jeśli macierz A określona wzorem* (6.6.3) *jest nieosobliwa, to $(A)_{ii} \neq 0$ dla $1 \leqslant i \leqslant n$.*

Dowód. Niech dla pewnego $r \leqslant n$ będzie $(A)_{rr} = 0$, czyli $B_r^k(x_r) = 0$. Stąd $x_r \notin (t_r, t_{r+k+1})$. Niech będzie najpierw $x_r \leqslant t_r$. Jeśli $i \leqslant r \leqslant j$, to $x_i \leqslant x_r \leqslant t_r \leqslant t_j$ i x_i nie należy do nośnika funkcji B_j^k. Dlatego $(A)_{ij} = B_j^k(x_i) = 0$ dla $j = r, r+1, \ldots, n$. Tak więc w każdym z r początkowych wierszy macierzy A nie znika co najwyżej $r-1$ początkowych elementów. Można je interpretować jako składowe wektorów z $\mathbb{R}^{r-1}$. Skoro tych wektorów jest r, więc muszą one być zależne liniowo i macierz A jest osobliwa.

Jeśli natomiast $x_r \geqslant t_{r+k+1}$ oraz $i \geqslant r \geqslant j$, to $x_i \geqslant x_r \geqslant t_{r+k+1} \geqslant t_{j+k+1}$ i $(A)_{ij} = B_j^k(x_i) = 0$. Wtedy kolumny od pierwszej do r-tej są liniowo zależne, bo w każdej z nich znikają elementy od r-tego do n-tego. Stąd znów wynika, że A jest osobliwa. ■

Lemat 6.6.3. *Jeśli $k = 1$ i $t_i < x_i < t_{i+2}$ dla $1 \leqslant i \leqslant n$, to macierz A jest nieosobliwa.*

Dowód. Dowód jest indukcyjny. Lemat jest oczywisty dla $n = 1$, bo wtedy A jest macierzą o jednym elemencie $B_1^1(x_1) \neq 0$.

Niech teraz będzie $n > 1$. Zakładamy, że lemat jest prawdziwy dla macierzy A stopnia $< n$.

obu słów są chyba tylko *węzły*. Aby uniknąć nieporozumień, będziemy teraz tak nazywać jedynie punkty x_j (*przyp. tłum.*).

Przypadek 1. Jeśli istnieje takie r, że $1 \leqslant r < n$ i $x_r \leqslant t_{r+1}$, to dla każdej pary (i,j) takiej, że $i \leqslant r < j$ jest $x_i \leqslant x_r \leqslant t_{r+1} \leqslant t_j$ i $(A)_{ij} = B_j^1(x_i) = 0$. Dlatego

$$A = \begin{bmatrix} C & 0 \\ E & D \end{bmatrix},$$

gdzie C jest macierzą stopnia r, a D macierzą stopnia $n-r$. Macierz A jest nieosobliwa wtedy i tylko wtedy, gdy takie są C i D (zob. zad. 7). Te dwie macierze mają taką strukturę jak A, ale mniejszy stopień, więc z założenia indukcyjnego są nieosobliwe.

Przypadek 2. Podobnie rozumujemy gdy $x_r \geqslant t_{r+1}$ dla r takiego, że $1 < r \leqslant n$. Wtedy $A_{ij} = 0$ dla $j < r \leqslant i$, czyli

$$A = \begin{bmatrix} C & E \\ 0 & D \end{bmatrix},$$

gdzie C i D są macierzami odpowiednio stopnia $r-1$ i $n-r+1$. Jak poprzednio stwierdzamy, że A jest nieosobliwa, bo takie są C i D.

Przypadek 3. Jeśli nie są spełnione warunki określające przypadki 1 i 2, to $x_i > t_{i+1}$ dla $1 \leqslant i < n$ i jednocześnie $x_i < t_{i+1}$ dla $1 < i \leqslant n$. Tak może być tylko dla $n = 1$ (macierz A tego stopnia już rozważono) albo $n = 2$. Wtedy zakładamy, że $x_1 > t_2$ i $x_2 < t_3$, czyli $t_2 < x_1 < x_2 < t_3$. W przedziale (t_2, t_3) jest $B_1^1(x) + B_2^1(x) = 1$, czyli

$$A = \begin{bmatrix} B_1^1(x_1) & B_2^1(x_1) \\ B_1^1(x_2) & B_2^1(x_2) \end{bmatrix} = \begin{bmatrix} \lambda & 1-\lambda \\ \mu & 1-\mu \end{bmatrix}.$$

Wyznacznik tej macierzy jest równy $\lambda - \mu = B_1^1(x_1) - B_1^1(x_2) > 0$ (zob. rys. 6.9 dla $i = 1$). ■

Twierdzenie 6.6.4 (Schoenberg-Whitney). *Macierz A określona za pomocą wzoru (6.6.3) jest nieosobliwa wtedy i tylko wtedy, gdy jej wszystkie elementy przekątniowe są różne od 0.*

Dowód. Warunek konieczny sprawdzono w lem. 6.6.2. Warunek dostateczny jest trywialny dla $k = 0$, bo równość $B_i^0(x_i) \neq 0$ jest równoważna temu, że $t_i \leqslant x_i < t_{i+1}$. Nośniki funkcji B_j^0 są rozłączne, więc $B_j^0(x_i) = \delta_{ij}$.

Dla $k = 1$ warunek dostateczny wynika z lem. 6.6.3. Dalej rozumujemy przez indukcję względem k. Dla ustalonego $k \geqslant 2$ zastosujemy też indukcję względem n, ale już w sposób mniej formalny. Jak w dowodzie lem. 6.6.3 wnioskujemy, że jeśli $x_r \leqslant t_{r+1}$ dla r takiego, że $1 \leqslant r < n$ albo jeśli

$x_r \geqslant t_{r+k}$ dla r takiego, że $1 < r \leqslant n$, to macierz A dzieli się na bloki tam określone, wobec czego jej nieosobliwość wynika z tejże własności bloków C i D. Możemy więc założyć, że

$$t_{i+1} < x_i < x_{i+1} < t_{i+k+1} \quad (1 \leqslant i < n).$$

Przypuśćmy, że – wbrew temu, co chcemy udowodnić – macierz A jest osobliwa. Wtedy $Au = 0$ dla pewnego $u \neq 0$. Niech będzie $f = \sum_{j=1}^{n} u_j B_j^k$. Ta funkcja znika w $n+2$ punktach $t_1, x_1, x_2, \ldots, x_n, t_{n+k+1}$ (tworzą one ciąg rosnący; dwa skrajne nie należą do nośnika żadnej z funkcji B_j^k obecnych w f). Ponieważ $k \geqslant 2$, więc pochodna f' istnieje i jest ciągła. Z twierdzenia Rolle'a zastosowanego do f wynika zatem istnienie $n+1$ zer ξ_i tej pochodnej, takich, że

$$t_1 < \xi_1 < x_1 < \xi_2 < x_2 < \ldots < x_{n-1} < \xi_n < x_n < \xi_{n+1} < t_{n+k+1}. \tag{6.6.5}$$

Z (6.5.9) wynika, że

$$f' = \sum_{j=1}^{n+1} v_j B_j^{k-1}, \quad \text{gdzie} \quad v_j := \frac{k(u_j - u_{j-1})}{t_{j+k} - t_j}$$

(trzeba tu przyjąć, że $u_0 = u_{n+1} = 0$). Z definicji punktów ξ_i wynika, że

$$\sum_{j=1}^{n+1} v_j B_j^{k-1}(\xi_i) = 0 \quad (1 \leqslant i \leqslant n+1), \tag{6.6.6}$$

a z (6.6.5), że ξ_i dla $2 \leqslant i \leqslant n$ należy do nośnika funkcji B_i^{k-1}.

Rozróżnimy teraz cztery przypadki, uwzględniając nierówności $\xi_1 < t_{k+1}$ i $\xi_{n+1} > t_{n+1}$. Jeśli $\xi_1 < t_{k+1}$, to ξ_1 należy do nośnika funkcji B_1^{k-1}; w przeciwnym razie żaden z punktów ξ_i tam się nie znajduje, czyli $B_1^{k-1}(\xi_i) = 0$ dla $1 \leqslant i \leqslant n+1$. Podobnie, jeśli $\xi_{n+1} > t_{n+1}$, to ξ_{n+1} należy do nośnika funkcji B_{n+1}^{k-1}, a w przeciwnym razie $B_{n+1}^{k-1}(\xi_i) = 0$ dla $1 \leqslant i \leqslant n+1$.

Przypadek 1. $\xi_1 < t_{k+1}$ i $\xi_{n+1} > t_{n+1}$. Wtedy każdy z punktów ξ_i należy do nośnika funkcji B_i^{k-1} i zgodnie z założeniem indukcyjnym macierz układu (6.6.6) jest nieosobliwa, a zatem wszystkie współczynniki v_j znikają. Z ich definicji wynika, że $u_1 - u_0 = \ldots = u_{n+1} - u_n = 0$, a ponieważ $u_0 = u_{n+1} = 0$, więc i pozostałe u_i znikają, co przeczy założeniu.

Przypadek 2. $\xi_1 \geqslant t_{k+1}$ i $\xi_{n+1} > t_{n+1}$. Wtedy układ (6.6.6) upraszcza się nieco:

$$\sum_{j=2}^{n+1} v_j B_j^{k-1}(\xi_i) = 0 \quad (2 \leqslant i \leqslant n+1).$$

Ponieważ dla $2 \leqslant i \leqslant n+1$ punkt ξ_i należy do nośnika funkcji B_i^{k-1}, więc z założenia indukcyjnego wynika, że wszystkie v_j w nowym układzie znikają, a to wraz z równością $u_{n+1} = 0$ pozwala wnioskować, że $u_i = 0$ dla $1 \leqslant i \leqslant n$, znów wbrew założeniu indukcyjnemu.

Przypadek 3. $\xi_1 < t_{k+1}$ i $\xi_{n+1} \leqslant t_{n+1}$. Rozumowanie jak w poprzednim przypadku daje sprzeczność z założeniem indukcyjnym.

Przypadek 4. $\xi_1 \geqslant t_{k+1}$ i $\xi_{n+1} \leqslant t_{n+1}$. Jest $B_1^{k-1}(\xi_i) = 0$ i $B_{n+1}^{k-1}(\xi_i) = 0$ dla $1 \leqslant i \leqslant n+1$. Układ (6.6.6) redukuje się do postaci

$$\sum_{j=2}^{n} v_j B_j^{k-1}(\xi_i) = 0 \quad (2 \leqslant i \leqslant n).$$

Jak w przyp. 2 wnioskujemy, że $u_2 - u_1 = \ldots = u_n - u_{n-1} = 0$. Niech λ oznacza wspólną wartość tych u_i. Jest więc $f = \lambda \sum_{j=1}^{n} B_j^k$. Ponieważ $B_j^k \geqslant 0$ i $B_1^k(x_1) > 0$, więc z równości $f(x_1) = 0$ wynika, że $\lambda = 0$. ■

Przykład 6.6.5. Niech będzie $t_i = i$ dla $i \in \mathbb{Z}$. Czy punkty x_i $(1 \leqslant i \leqslant 5)$ równe odpowiednio 3.1, 3.5, 3.6, 6.1, 6.6 mogą być węzłami interpolacji dla $k = 2$?

Rozwiązanie. Odpowiedź jest twierdząca, bo warunek $i < x_i < i+3$ $(1 \leqslant i \leqslant 5)$ wynikający z tw. 6.6.4, jest spełniony. ■

Istnienie

Z twierdzenia 6.6.1 wynika, że przestrzeń $\mathcal{S}_n^k$ funkcji sklejanych ma wymiar $n+k$. Można przyjąć, że rozpatrujemy je w przedziale $[t_0, t_n]$. Stawiamy następujące pytanie: Jakie warunki powinny spełniać węzły $x_1, x_2, \ldots, x_{n+k}$ z tego przedziału, aby interpolacja za pomocą funkcji z $\mathcal{S}_n^k$ była możliwa?

Twierdzenie 6.6.6. *Jeśli węzły $x_1, x_2, \ldots, x_{n+k}$ należą do przedziału $[t_0, t_n]$ i są takie, że*

$$t_{i-k-1} < x_i < t_i \quad (1 \leqslant i \leqslant n+k),$$

to dowolne wartości w tych węzłach można interpolować za pomocą funkcji z przestrzeni $\mathcal{S}_n^k$.

Dowód. Na mocy tw. 6.6.1 bazą przestrzeni $\mathcal{S}_n^k$ są funkcje $B_j^k \mid [t_0, t_n]$ dla $-k \leqslant j \leqslant n-1$. Zmieńmy numerację węzłów: niech będzie $y_i = x_{i+k+1}$ dla $-k \leqslant i \leqslant n-1$. Wtedy y_i należy do nośnika funkcji B_i^k i tw. 6.6.4 gwarantuje, że macierz o elementach $B_j^k(y_i)$ jest nieosobliwa. ■

Wracamy teraz do zadania interpolacji z n węzłami $x_1, x_2, \ldots, x_n$ za pomocą funkcji

$$S(x) = \sum_{j=1}^{n} c_j B_j^k(x).$$

Wynika stąd układ równań

$$\sum_{j=1}^{n} B_j^k(x_i) c_j = f(x_i) \quad (1 \leqslant i \leqslant n).$$

Zgodnie z tw. 6.6.4 jego macierz jest nieosobliwa, jeśli są spełnione nierówności (6.6.4).

Załóżmy teraz, że $x_i = t_i$ (co jest sprzeczne z (6.6.4)) i rozważmy zadanie interpolacyjne prowadzące do układu

$$\sum_{j=-\infty}^{\infty} c_j B_j^k(t_i) = f(t_i) \quad (1 \leqslant i \leqslant n) \tag{6.6.7}$$

(mamy tu w istocie sumę skończoną $\sum_j$, bo nośniki funkcji B_j^k są przedziałami skończonymi). W najprostszym przypadku, dla $k = 0$, mamy oczywiste rozwiązanie $c_j = f(t_j)$, gdyż $B_j^0(t_i) = \delta_{ij}$. Dla $k = 1$ jest $c_{i-1} = f(t_i)$, gdyż $B_j^1(t_i) = \delta_{i-1,j}$.

Aby rozwiązać to samo zadanie dla $k = 2$, korzystamy z wyrażenia dla $B_j^2(t_i)$ podanego w zad. 6.5.3. Każde równanie układu (6.6.7) zawiera tylko dwa szukane współczynniki:

$$\frac{t_{i+1} - t_i}{t_{i+1} - t_{i-1}} c_{i-2} + \frac{t_i - t_{i-1}}{t_{i+1} - t_{i-1}} c_{i-1} = f(t_i) \quad (1 \leqslant i \leqslant n).$$

Tu mamy układ n równań z $n + 1$ niewiadomymi. Można go rozwiązać, wybierając dowolne c_{-1} i obliczając z tych równań kolejno $c_0, c_1, \ldots, c_{n-1}$. Zadanie interpolacyjne ma zatem nieskończenie wiele rozwiązań.

Dla $k = 3$ z (6.6.7) wynika w taki sam sposób układ n równań z $n + 2$ niewiadomymi:

$$B_{i-3}^3(t_i) c_{i-3} + B_{i-2}^3(t_i) c_{i-2} + B_{i-1}^3(t_i) c_{i-1} = f(t_i) \quad (1 \leqslant i \leqslant n). \tag{6.6.8}$$

Współczynniki tego układu można obliczyć, stosując wzór (6.5.1):

$$B_{i-3}^3(t_i) = \frac{(t_{i+1} - t_i)^2}{(t_{i+1} - t_{i-2})(t_{i+1} - t_{i-1})},$$

$$B_{i-2}^3(t_i) = \frac{(t_{i+2} - t_i)(t_i - t_{i-1})}{(t_{i+2} - t_{i-1})(t_{i+1} - t_{i-1})} + \frac{(t_i - t_{i-2})(t_{i+1} - t_i)}{(t_{i+1} - t_{i-2})(t_{i+1} - t_{i-1})},$$

$$B^3_{i-1}(t_i) = \frac{(t_i - t_{i-1})^2}{(t_{i+2} - t_{i-1})(t_{i+1} - t_{i-1})}.$$

Podobnie jak dla $k = 2$, można tu wybrać dowolne wartości dla c_{-2} i c_{-1}, a potem za pomocą równań (6.6.8) obliczać kolejno $c_0, c_1, \ldots, c_{n-1}$. Nie jest to jednak zalecany sposób. Zazwyczaj narzuca się dodatkowe warunki na funkcję sklejaną w skrajnych punktach t_1 i t_n. W szczególności warunki $S''(t_1) = S''(t_n) = 0$ dają *naturalną* funkcję sklejaną. Ten wariant jest tematem zadań 4–6. Rady dotyczące interpolacji za pomocą funkcji sklejanych daje de Boor [1984] (w jego książce B^k_j oznacza taką funkcję stopnia $k-1$).

Nieinterpolacyjne metody aproksymacji

Aby zilustrować takie metody, wspomnimy o eleganckiej procedurze Schoenberga [1967]. Dla danej funkcji f określamy funkcję sklejaną Sf wzorem

$$Sf := \sum_{i=-\infty}^{\infty} f(x_i) B^k_i, \quad \text{gdzie} \quad x_i := \frac{1}{k}(t_{i+1} + \ldots + t_{i+k}). \tag{6.6.9}$$

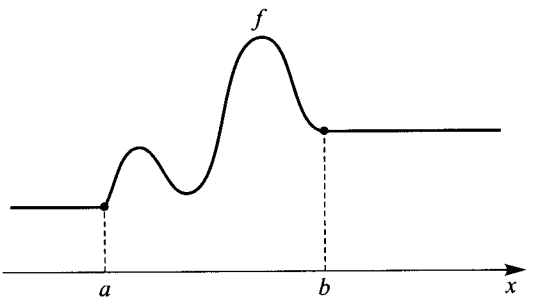
Przypadki, gdy $k = 0$ (wtedy przyjmuje się, że $x_i = t_i$) i $k = 1$ nie są ciekawe, bo redukują się do już opisanej interpolacji. Natomiast dla $k > 1$ powyższy wzór określa funkcję Sf, która nie interpoluje funkcji f w jakichś z góry ustalonych węzłach. Mamy tu do czynienia z operatorem *quasi-interpolacyjnym*. Jego istotne własności są następujące:

1. Jeśli f jest funkcją liniową, to $Sf = f$.
2. Dla dowolnej funkcji liniowej l różnica $Sf - l$ zmienia znak nie więcej razy niż różnica $f - l$.
3. Jeśli $f \geqslant 0$, to $Sf \geqslant 0$.
4. Jeśli $|f| \leqslant M$, to $|Sf| \leqslant M$.
5. S jest operatorem liniowym, tzn. $S(\alpha f + \beta g) = \alpha Sf + \beta Sg$.

Czytelników zainteresowanych tym tematem odsyłamy do prac Marsdena [1970] i Schoenberga [1967].

Naszym następnym celem jest zbadanie, czy funkcje ciągłe można przybliżać z dowolną dokładnością funkcjami sklejanymi. Ustalamy przy tym stopień k tych funkcji, natomiast poprawę dokładności chcemy uzyskać, zwiększając liczbę węzłów.

Jak dotąd, punkty t_i określa się dla wszystkich $i \in \mathbb{Z}$, ale przyjmuje się, że aproksymowana funkcja ciągła f jest dana tylko w $[a, b] := [t_0, t_n]$. Jej określenie rozszerza się na całą prostą rzeczywistą tak, aby zachować ciągłość funkcji; zob. rys. 6.10.



RYS. 6.10. Rozszerzenie funkcji f

Dla dowolnej funkcji f (ciągłej lub nie) określamy jej *moduł ciągłości* wzorem

$$\omega(f;\delta) := \max_{|s-t|\leqslant\delta} |f(s)-f(t)|.$$

Jeśli f jest ciągła w przedziale domkniętym $[a,b]$, to jest tam jednostajnie ciągła, czyli każdemu $\varepsilon > 0$ odpowiada takie $\delta > 0$, że dla wszystkich s i t z $[a,b]$ jest $|s-t| < \delta \Rightarrow |f(s)-f(t)| < \varepsilon$. Wtedy $\omega(f;\delta) \leqslant \varepsilon$. Inaczej mówiąc, moduł ciągłości takiej funkcji dąży do 0 wraz z δ.

Jeśli pochodna f' istnieje, jest ciągła i taka, że $|f'(x)| \leqslant M$, to wobec twierdzenia o wartości średniej $|f(s)-f(t)| = |f'(\xi)|\,|s-t| \leqslant M|s-t|$, a zatem $\omega(f;\delta) \leqslant M\delta$.

Poniższe twierdzenie świadczy o tym, że funkcja sklejana

$$g := \sum_{i=-\infty}^{\infty} f(t_{i+2})B_i^k \tag{6.6.10}$$

może być sensownym przybliżeniem dla f.

Twierdzenie 6.6.7. *Jeśli funkcja f jest określona w przedziale $[a,b]$, jeśli $k \geqslant 2$, a g wyraża się wzorem* (6.6.10), *to*

$$\max_{a\leqslant x\leqslant b} |f(x)-g(x)| \leqslant k\omega(f;\delta), \quad \textit{gdzie} \quad \delta := \max_{-k\leqslant i\leqslant n+1} (t_i - t_{i-1}).$$

Dowód. Przypominamy, że $B_i^k \geqslant 0$ i $\sum_{i=-\infty}^{\infty} B_i^k = 1$. Dlatego

$$|g(x)-f(x)| = \Big|\sum_{i=-\infty}^{\infty} f(t_{i+2})B_i^k(x) - f(x)\sum_{i=-\infty}^{\infty} B_i^k(x)\Big| =$$

$$= \Big|\sum_{i=-\infty}^{\infty} [f(t_{i+2})-f(x)]B_i^k(x)\Big| \leqslant \sum_{i=-\infty}^{\infty} |f(t_{i+2})-f(x)|B_i^k(x).$$

Niech będzie $x \in [t_j, t_{j+1}] \subset [a,b]$ (czyli $0 \leqslant j \leqslant n-1$). W przedziale $[t_j, t_{j+1}]$ nie znikają co najwyżej funkcje $B^k_{j-k}, B^k_{j-k+1}, \ldots, B^k_j$. Wobec tego

$$|g(x) - f(x)| \leqslant \sum_{i=j-k}^{j} |f(t_{i+2}) - f(x)| B^k_i(x) \leqslant \max_{j-k \leqslant i \leqslant j} |f(t_{i+2}) - f(x)|.$$

Dla i takich, że $j-k \leqslant i \leqslant j$, jest

$$t_{i+2} - x \leqslant t_{j+2} - t_j \leqslant 2\delta, \quad x - t_{i+2} \leqslant t_{j+1} - t_{j-k+2} \leqslant k\delta.$$

Stąd i z własności modułu ciągłości podanej w zad. 14 wynika, że

$$|f(t_{i+2}) - f(x)| \leqslant \omega(f; k\delta) \leqslant k\omega(f; \delta).$$ ■

Odległość funkcji od przestrzeni funkcji sklejanych

Ostatnie twierdzenie można wyrazić inaczej – za pomocą odległości funkcji od przestrzeni $\mathcal{S}^k_n$. Ogólniej, odległość punktu f przestrzeni unormowanej od jej podprzestrzeni G określamy jako

$$\operatorname{dist}(f, G) := \inf_{g \in G} \|f - g\|.$$

Tu, ponieważ funkcje sklejane są (dla $k > 0$) ciągłe, używamy normy

$$\|f\|_\infty := \max_{a \leqslant x \leqslant b} |f(x)|.$$

Twierdzenie 6.6.7 oznacza, że jeśli $k \geqslant 2$, to

$$\operatorname{dist}(f, \mathcal{S}^k_n) \leqslant \omega(f; \delta).$$

Jeśli funkcja f jest ciągła, to $\lim_{\delta \to 0} \omega(f; \delta) = 0$, czyli ta odległość dąży do 0, gdy gęstość punktów t_i rośnie.

Dla funkcji mających pochodne ciągłe można o wspomnianej odległości powiedzieć coś więcej:

Twierdzenie 6.6.8. *Jeśli r jest liczbą naturalną, $r < k < n$, $f \in C^r[a,b]$, to dla δ określonego w tw.* 6.6.7 *jest*

$$\operatorname{dist}(f, \mathcal{S}^k_n) \leqslant k^r \delta^r \|f^{(r)}\|_\infty.$$

Dowód. Niech g będzie dowolną funkcją z $\mathcal{S}^k_n$. Na mocy tw. 6.6.7 jest

$$\operatorname{dist}(f, \mathcal{S}^k_n) = \operatorname{dist}(f - g, \mathcal{S}^k_n) \leqslant k\omega(f - g; \delta) \leqslant k\delta \|f' - g'\|_\infty.$$

Jeśli g przebiega przestrzeń $\mathcal{S}_n^k$, to g' przebiega $\mathcal{S}_n^{k-1}$ (zob. zad. 15). Dlatego biorąc wyżej kres dolny względem g otrzymujemy nierówność

$$\operatorname{dist}(f, \mathcal{S}_n^k) \leqslant k\delta \operatorname{dist}(f', \mathcal{S}_n^{k-1}).$$

Iterowanie tej nierówności $r-2$ razy daje następujący wynik:

$$\begin{aligned}\operatorname{dist}(f, \mathcal{S}_n^k) &\leqslant k^{r-1}\delta^{r-1} \operatorname{dist}(f^{(r-1)}, \mathcal{S}_n^{k-r+1}) \leqslant \\ &\leqslant k^r \delta^{r-1} \omega(f^{(r-1)}; \delta) \leqslant k^r \delta^r \|f^{(r)}\|_\infty. \end{aligned}$$ ■

ZADANIA 6.6

1. Udowodnić, że jeśli $x_1 < x_2 < \ldots < x_n$ i $B_j^k(x_j) \neq 0$ dla $1 \leqslant j \leqslant n$, to macierz A określona przed lem. 6.6.2 zawiera w każdym wierszu i każdej kolumnie najwyżej $2k+1$ niezerowych elementów.

2. Niech będzie $K \subset \mathbb{R}$. Udowodnić, że układ $\{B_1^k, \ldots, B_n^k\}$ jest niezależny liniowo na K wtedy i tylko wtedy, gdy wszystkie zbiory $K \cap (t_i, t_{i+k+1})$ $(1 \leqslant i \leqslant n)$ są niepuste.

3. Udowodnić, że $x^2 = \sum_{i=-\infty}^{\infty} t_{i+1}t_{i+2}B_i^2(x)$.

4. Udowodnić, że jeśli $S = \sum_{j=-\infty}^{\infty} c_j B_j^3$, to $S'' = \sum_{j=-\infty}^{\infty} e_j B_j^1$, gdzie

$$e_j := \frac{6}{t_{j+2} - t_j}\left(\frac{c_j - c_{j-1}}{t_{j+3} - t_j} - \frac{c_{j-1} - c_{j-2}}{t_{j+2} - t_{j-1}}\right).$$

5. (cd.). Udowodnić, że $S''(t_i) = e_{i-1}$.

6. (cd.). Udowodnić, że suma $\sum_{j=-2}^{n-1} c_j B_j^3$ o współczynnikach c_j spełniających równania (6.6.8) i warunki

$$(t_{i+2} - t_{i-1})c_{i-3} - (t_{i+2} + t_{i+1} - t_{i-1} - t_{i-2})c_{i-2} + (t_{i+1} - t_{i-2})c_{i-1} = 0$$

dla $i = 1, n$ jest naturalną funkcją sklejaną stopnia 3, interpolującą f.

7. Udowodnić, że macierz kwadratowa $\begin{bmatrix} C & 0 \\ E & D \end{bmatrix}$, gdzie bloki C i D też są kwadratowe, jest nieosobliwa wtedy i tylko wtedy, gdy te bloki mają taką samą własność.

8. Sprawdzić, czy tw. 6.6.4 można zmodyfikować w następujący sposób: Dla dowolnie uporządkowanych węzłów x_i nieosobliwość macierzy o elementach $B_j^k(x_i)$ jest równoważna temu, że każdy przedział (t_i, t_{i+k+1}) zawiera co najmniej jeden węzeł.

9. Udowodnić, że z tw. 6.6.4 wynika: (**a**) lem. 6.5.9, (**b**) lem. 6.5.10.

10. Niech będzie $k = 2$. Udowodnić, że jeśli: (**a**) $f(x) := 1$, (**b**) $f(x) := x$, to w (6.6.9) jest $Sf = f$.

11. Udowodnić własności **1**, **3**, **4** procedury Schoenberga (zob. (6.6.9)). W związku z **1** wystarczy sprawdzić, że $Sf = f$ dla $f(x) := 1$ i $f(x) := x$. Dla drugiej funkcji wykazać za pomocą (6.5.9), że $(Sf') = 1$.

12. Wzorując się na dowodzie tw. 6.6.7, otrzymać podobny wynik dla procedury Schoenberga.

13. (cd.). Poprawić wynik poprzedniego zadania, zakładając, że punkty t_i są równoodległe.

14. Udowodnić, że $\omega(f; k\delta) \leqslant k\omega(f; \delta)$.

15. Wykazać, że różniczkowanie funkcji jest operacją suriektywną z $\mathcal{S}_n^k$ na $\mathcal{S}_n^{k-1}$.

6.7. Szeregi potęgowe

Zaczynając ten krótki podrozdział przypomnijmy wzór Taylora z podrozdz. 1.1, dający użyteczne przybliżenia funkcji, które mają dostatecznie wiele ciągłych pochodnych. Jeśli funkcja f ma w przedziale $[c-\delta, c+\delta]$ ciągłą $(n+1)$-szą pochodną, to dla każdego x z tego przedziału

$$f(x) = p_n(x) + E_n(x),$$

gdzie p_n jest wielomianem klasy Π_n, a E_n – resztą:

$$p_n(x) := \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(c)(x-c)^k,$$

$$E_n(x) := \frac{1}{(n+1)!} f^{(n+1)}(\xi_x)(x-c)^{n+1}, \quad \text{gdzie} \quad |\xi_x - c| < \delta.$$

Jeśli ta funkcja ma we wspomnianym przedziale wszystkie pochodne, to przynajmniej formalnie możemy napisać, że

$$f(x) = \sum_{k=0}^{\infty} \frac{1}{k!} f^{(k)}(c)(x-c)^k.$$

Jest to *szereg Taylora*; dla $c = 0$ nazywamy go *szeregiem Maclaurina*. Ścisłe uzasadnienie jego poprawności wymaga dowodu, że $E_n(x)$ dąży do 0 dla $n \to \infty$. Taki dowód prowadzi m.in. do wniosku, że

$$\cos x = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k} \qquad (-\infty < x < \infty), \tag{6.7.1}$$

$$\frac{1}{x} = \sum_{k=0}^{\infty} (-1)^k (x-1)^k \qquad (0 < x < 2). \tag{6.7.2}$$

Wprowadzone nazwy szeregów kojarzą się z ich konstrukcją za pomocą wzoru Taylora (Maclaurina). Często jest wygodniej abstrahować od związku

współczynników szeregu z pochodnymi. Piszemy go więc w postaci

$$\sum_{k=0}^{\infty} a_k(x-c)^k \tag{6.7.3}$$

i nazywamy *szeregiem potęgowym.*

Twierdzenie 6.7.1. *Dla każdego szeregu potęgowego* (6.7.3) *istnieje liczba* $r \in [0,\infty]$ *taka, że jest on zbieżny dla* $|x-c| < r$ *i rozbieżny dla* $|x-c| > r$.

Liczbę r nazywamy *promieniem zbieżności* szeregu (6.7.3). Do jego obliczenia służą znane kryteria zbieżności; jedno z nich (kryterium d'Alemberta) stosujemy w zad. 1. Promień zbieżności szeregu (6.7.1) dla cosinusa jest nieskończony, szereg (6.7.2) ma promień zbieżności 1. Szeregi o promieniu zbieżności $r = 0$ nie mają praktycznego znaczenia.

W zastosowaniach ważne jest także następne twierdzenie:

Twierdzenie 6.7.2. *Niech* r *będzie promieniem zbieżności szeregu* (6.7.3). *Wtedy jego sumą dla każdego* $x \in (c-r, c+r)$ *jest funkcja* $f(x)$, *która ma tam ciągłą pochodną wyrażającą się wzorem*

$$f'(x) = \sum_{k=0}^{\infty} k a_k (x-c)^{k-1},$$

gdzie szereg ma także promień zbieżności r. *Ponadto, jeśli* $b, x \in (c-r, c+r)$, *to całkę*

$$\int_b^x f(t)\,dt$$

można otrzymać, całkując szereg (6.7.3) *wyraz po wyrazie; daje to szereg o promieniu zbieżności* r.

Na mocy tego twierdzenia szereg dla $f'(x)$ otrzymujemy z szeregu dla $f(x)$, różniczkując jego poszczególne składniki.

Aby zilustrować zastosowania tego twierdzenia, rozważmy jedną z funkcji specjalnych, a mianowicie *sinus całkowy*

$$\operatorname{Si}(x) := \int_0^x \frac{\sin t}{t}\,dt.$$

Ta całka nie wyraża się w skończonej postaci przez funkcje elementarne, jednak z szeregu potęgowego dla sinusa wynika, że

$$\frac{\sin t}{t} = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} t^{2k}.$$

Całkując składniki tego szeregu, otrzymujemy szukane wyrażenie:

$$\mathrm{Si}(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} \int_0^x t^{2k}\,dt = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)(2k+1)!}\, x^{2k+1}.$$

Ten szereg jest szybko zbieżny dla małych $|x|$; np. jego dziesięć początkowych składników daje 20 cyfr znaczących wartości Si(1) (zob. zad. 22). W podobny sposób otrzymuje się szereg

$$\int_0^x e^{t^2}\,dt = \sum_{k=0}^{\infty} \frac{1}{(2k+1)k!}\, x^{2k+1}. \tag{6.7.4}$$

Zadania 6.7

1. *Kryterium d'Alemberta* mówi, że jeśli granica $\lim_{n\to\infty}|A_{n+1}/A_n|$ jest mniejsza (odpowiednio, większa) od 1, to szereg $\sum_{k=0}^{\infty} A_k$ jest zbieżny (odpowiednio, rozbieżny). Stosując to kryterium: (**a**) wykazać, że szereg (6.7.1) jest zbieżny dla każdego x, (**b**) znaleźć promień zbieżności szeregu (6.7.2).
2. Znaleźć promień zbieżności szeregu $\sum_{k=0}^{\infty} k!x^k$.
3. Znaleźć promień zbieżności szeregu $\sum_{k=0}^{\infty} a_k x^k$ takiego, że:
   (**a**) $a_k = 2^k k$ dla $k \geqslant 0$,
   (**b**) $a_0 = 1$ i $a_k = a_{k-1}/[2(k+1)]$ dla $k > 0$,
   (**c**) jest to pochodna szeregu z poprzedniego zadania,
   (**d**) $a_0 = 1$ i $a_k = [2 + (-1)^k]a_{k-1}$ dla $k > 0$,
   (**e**) $a_0 = 1$ i $a_k = (k/3)a_{k-1}$ dla $k > 0$,
   (**f**) $a_0 = 1$ i $a_k = [3k/(k+4)]a_{k-1}$ dla $k > 0$.
4. Jeśli szeregi $\sum_{k=0}^{\infty} a_k x^k$ i $\sum_{k=0}^{\infty} b_k x^k$ mają odpowiednio promień zbieżności r i r', to co można powiedzieć o promieniu zbieżności szeregu $\sum_{k=0}^{\infty}(a_k + b_k)x^k$?
5. Jeśli szereg $f(x) := \sum_{k=0}^{\infty} a_k(x-c)^k$ ma promień zbieżności r, to f ma wszystkie pochodne w przedziale $(c-r, c+r)$ i w tymże przedziale

   $$f^{(n)}(x) = \sum_{k=n}^{\infty} \frac{k!a_k}{(k-n)!}\,(x-c)^{k-n}.$$

   Udowodnić to, korzystając z tw. 6.7.2.
6. Udowodnić, że

   $$\frac{1+x}{1-x} = 1 + 2\sum_{k=1}^{\infty} x^k \quad (|x| < 1).$$

7. Udowodnić równość (6.7.4).
8. Całkując szereg z (6.7.2), otrzymać szereg dla $\log x$.

9. Pokazać, jak szereg dla $\sin x$ wynika przez różniczkowanie bądź całkowanie szeregu z (6.7.1).

10. Rozwinąć w szereg potęgowy funkcję $\operatorname{arctg} x$, całkując szereg

$$(1+x^2)^{-1} = \sum_{k=0}^{\infty} (-x^2)^k.$$

Porównać ten sposób z konstrukcją wzoru Taylora, wymagającą obliczania pochodnych funkcji $\operatorname{arctg} x$.

11. Rozwinąć w szereg potęgowy *funkcję błędu* (stosowaną w statystyce)

$$\operatorname{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt.$$

12. Rozwinąć w szereg potęgowy funkcję

$$f(x) := \int_0^x \frac{e^t - 1}{t}\, dt.$$

Za pomocą tego szeregu obliczyć $f(1)$ z trzema cyframi znaczącymi. Funkcja jest ważna w zastosowaniach; zob. Abramowitz i Stegun [1964, rozdz. 5].

13. Funkcja związana z tzw. *dilogarytmem* jest określona wzorem

$$f(x) := -\int_0^x \frac{\log(1-t)}{t}\, dt \quad (-\infty < x < 1).$$

Znaleźć jej szereg Maclaurina i jego promień zbieżności. Czy ten szereg pozwala obliczyć $f(-2)$? A $f(0.001)$?

14. Znaleźć szereg Maclaurina dla *cosinusa całkowego*

$$\operatorname{Ci}(x) = \int_0^x \frac{1-\cos t}{t}\, dt.$$

15. Znaleźć szereg Maclaurina dla *całki Fresnela*

$$S(x) := \int_0^x \sin t^2\, dt.$$

16. Jeśli szereg $f(x) = \sum_{k=0}^{\infty} a_k x^k$, gdzie $a_0 = 1$, ma dodatni promień zbieżności, to w pewnym otoczeniu punktu 0 funkcja $1/f(x)$ jest poprawnie określona. Znaleźć wzór rekurencyjny pozwalający wyznaczać współczynniki jej rozwinięcia w szereg $\sum_{k=0}^{\infty} b_k x^k$.

17. (cd.). Znaleźć początkowe współczynniki b_k szeregu Maclaurina dla funkcji $x/(e^x - 1)$.

18. (cd.). Wykazać, że współczynniki b_k z poprzedniego zadania są takie, iż $b_3 = b_5 = b_7 = \ldots = 0$. **Uwaga**: Iloczyny $B_k := k!b_k$ są nazywane *liczbami Bernoulliego*.

19. (cd.). Udowodnić, że liczby Bernoulliego spełniają tożsamość

$$\sum_{k=0}^{n-1} \binom{n}{k} B_k = 0 \quad (n > 1).$$

Wykazać, że $B_0 = 1$ i że ta tożsamość pozwala obliczać $B_1, B_2, \ldots$, co daje $B_1 = -\frac{1}{2}$, $B_2 = \frac{1}{6}$, $B_3 = 0$ i $B_4 = -\frac{1}{30}$.

20. Znaleźć początkowe współczynniki szeregu Maclaurina dla funkcji (**a**) $1/\sin x - 1/x$, (**b**) $x/[(1-x)\log(1-x)]$.

21. Znaleźć trzy początkowe składniki szeregu Maclaurina dla funkcji $e^{\cos x}$. Wskazówka: Użyć zmiennej $z := 1 - \cos x$.

22. Wykazać, że 20 cyfr znaczących wartości Si(1) sinusa całkowego wynika z 10 początkowych składników szeregu potęgowego. Ile tych składników trzeba wziąć, aby otrzymać 25 cyfr znaczących?

ZADANIA KOMPUTEROWE 6.7

K1. Napisać procedurę obliczającą sinus całkowy z dokładnością do 10 cyfr znaczących dla każdego $x \in [-1, 1]$. Liczba użytych składników powinna zależeć od x. Procedura ma sygnalizować wartość x wykraczającą poza podany przedział.

6.8. Aproksymacja średniokwadratowa

Ogólne zadanie optymalnej aproksymacji formułujemy tak: Niech E będzie przestrzenią unormowaną, a G jej podprzestrzenią. Dla każdego $f \in E$ określamy jego *odległość* od G jako wielkość

$$\operatorname{dist}(f, G) = \inf_{g \in G} \|f - g\|.$$

Jeśli pewien element $g \in G$ jest taki, że

$$\|f - g\| = \operatorname{dist}(f, G),$$

to mówimy, że g *najlepiej aproksymuje* dany punkt f, czyli jest *punktem optymalnym* dla f. Sens tej *najlepszej* (*optymalnej*) *aproksymacji* zależy od wyboru normy.

W klasycznym zadaniu tzw. *aproksymacji jednostajnej* przestrzenią E jest zbiór $C[a, b]$ wszystkich funkcji ciągłych w $[a, b]$, norma jest określona wzorem

$$\|f\| := \max_{a \leqslant x \leqslant b} |f(x)|, \tag{6.8.1}$$

a podprzestrzenią G jest zbiór Π_n wszystkich wielomianów stopnia co najwyżej n. Znalezienie wielomianu $g \in \Pi_n$ najlepiej aproksymującego daną funkcję $f \in C[a,b]$ jest istotnie różne – i znacznie trudniejsze – od konstrukcji wielomianu interpolacyjnego dla ustalonego układu węzłów i od wyznaczania wielomianu ze wzoru Taylora. Ten klasyczny przypadek (i jego uogólnienia) jest tematem następnego podrozdziału.

Istnienie punktu optymalnego

W ogólnym zadaniu sformułowanym wyżej nasuwa się od razu ważne pytanie, czy punkt optymalny istnieje.

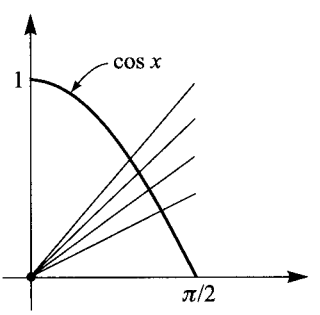
Twierdzenie 6.8.1. *Jeśli G jest podprzestrzenią skończeniewymiarową przestrzeni unormowanej E, to dla każdego punktu z E istnieje punkt optymalny w G.*

Dowód. Niech będzie $f \in E$. Wśród kandydatów na punkt optymalny jest punkt 0 należący do G. Wobec tego $\|f-g\| \leqslant \|f-0\| = \|f\|$ i szukanie punktu optymalnego możemy ograniczyć do zbioru

$$K := \{g \in G\colon \|g-f\| \leqslant \|f\|\}.$$

Jest on domknięty i ograniczony, a także – dzięki założeniu o G – zwarty. Ponieważ funkcja $g \mapsto \|f-g\|$ jest ciągła, więc wystarczy powołać się na twierdzenie o tym, że funkcja ciągła o wartościach rzeczywistych osiąga minimum na dowolnym zbiorze zwartym. ■

Element optymalny na ogół nie jest określony jednoznacznie. Świadczy o tym przykład aproksymacji funkcji $f(x) := \cos x$ w przedziale $[0, \pi/2]$ za pomocą rodziny funkcji $g(x) := \lambda x$ z parametrem λ. Na rysunku 6.11 pokazano, że jeśli stosujemy normę (6.8.1), to istnieje nieskończenie wiele funkcji optymalnych; dla każdej z nich $\|f-g\| = 1$.



RYS. 6.11. Aproksymacja funkcji $\cos x$ za pomocą λx

Przestrzeń unitarna

Jak już wspomniano, dla dowolnych E, G i f znalezienie elementu optymalnego jest trudne. Wymaga ono rozwiązania układu nieliniowego równań. Jest on jednak liniowy, jeśli E jest *przestrzenią unitarną*, czyli przestrzenią liniową, w której określono iloczyn skalarny i w której norma wyraża się przez ten iloczyn. W podrozdziale 4.7 zdefiniowano iloczyn skalarny wektorów z przestrzeni $\mathbb{R}^n$ i podano jego własności, które wynikają z tej definicji. Tu natomiast te własności (oczywiście z wyjątkiem ostatniej, gdzie występuje iloczyn macierzy przez wektor) stanowią aksjomatyczne określenie iloczynu skalarnego. Przyjmujemy mianowicie, że *iloczyn skalarny* $\langle f, g\rangle$ jest taką funkcją punktów (elementów) f, $g \in E$, która ma wartości rzeczywiste i następujące własności:

a. $\langle f, g\rangle = \langle g, f\rangle$,

b. $\langle f, \alpha g + \beta h\rangle = \alpha\langle f, g\rangle + \beta\langle f, h\rangle$ dla dowolnych rzeczywistych α i β,

c. $\langle f, f\rangle \geqslant 0$ i $\langle f, f\rangle = 0$ wtedy i tylko wtedy, gdy $f = 0$.

(Częściej przyjmuje się, jak w podrozdz. 4.6, że iloczyn skalarny ma wartości zespolone. Wtedy własność **a** należy zmienić na $\langle f, g\rangle = \overline{\langle g, f\rangle}$).

Przez iloczyn skalarny wyraża się *norma* elementu $f \in E$:

$$\|f\| := \sqrt{\langle f, f\rangle}. \tag{6.8.2}$$

Wobec tego norma jest zawsze liczbą nieujemną.

Wspomniano już przypadek $E = \mathbb{R}^n$. Inną ważną przestrzenią unitarną jest przestrzeń $C_w[a, b]$ funkcji ciągłych w $[a, b]$, w której dla ustalonej *funkcji wagowej* w o wartościach dodatnich jest

$$\langle f, g\rangle := \int_a^b w(x)f(x)g(x)\,dx.$$

W dowolnej przestrzeni unitarnej używamy jeszcze symbolu $\perp$: piszemy $f \perp g$, jeśli $\langle f, g\rangle = 0$ i $f \perp G$, jeśli $f \perp g$ dla każdego $g \in G$.

Lemat 6.8.2. *Dowolna przestrzeń unitarna ma następujące własności:*

1. $\langle \sum_{i=1}^n a_i f_i, g\rangle = \sum_{i=1}^n a_i\langle f_i, g\rangle$.

2. $\|f + g\|^2 = \|f\|^2 + 2\langle f, g\rangle + \|g\|^2$.

3. *Jeśli* $f \perp g$, *to* $\|f + g\|^2 = \|f\|^2 + \|g\|^2$ (twierdzenie Pitagorasa).

4. $|\langle f, g\rangle| \leqslant \|f\|\,\|g\|$ (nierówność Schwarza).

5. $\|f + g\| \leqslant \|f\| + \|g\|$ (nierówność trójkąta).

6. $\|f + g\|^2 + \|f - g\|^2 = 2\|f\|^2 + \|g\|^2$.

Dowód. Część **1** lematu jest oczywistym uogólnieniem własności **a**. Część **2** wynika z **a** i **b**:

$$\|f+g\|^2 = \langle f+g, f+g\rangle = \langle f,f\rangle + \langle f,g\rangle + \langle g,f\rangle + \langle g,g\rangle = \\ = \|f\|^2 + 2\langle f,g\rangle + \|g\|^2.$$

Części **3** i **6** wynikają wprost z **2**. Aby udowodnić **4**, przypuśćmy, że dla pewnych f i g ta nierówność nie zachodzi:

$$|\langle f,g\rangle| > \|f\|\,\|g\|.$$

Ponieważ $\langle f,0\rangle = 0$, więc musi być $g \neq 0$. Wobec jednorodności warunku możemy założyć, że $\|g\| = 1$, wobec czego $|\langle f,g\rangle| > \|f\|$. Z tegoż powodu wynika, że można przyjąć, iż $\langle f,g\rangle = 1$, a więc $\|f\| < 1$. To jednak prowadzi do sprzeczności:

$$0 \leqslant \|f-g\|^2 = \|f\|^2 - 2\langle f,g\rangle + \|g\|^2 = \|f\|^2 - 1.$$

Z nierówności Schwarza i z **2** wynika część **5** lematu:

$$\|f+g\|^2 \leqslant \|f\|^2 + 2\|f\|\,\|g\| + \|g\|^2 = (\|f\| + \|g\|)^2. \qquad \blacksquare$$

Twierdzenie 6.8.3. *Niech G będzie podprzestrzenią przestrzeni unitarnej E. Punkt $g \in G$ jest optymalny dla $f \in F$ wtedy i tylko wtedy, gdy $f-g \perp G$. Punkt optymalny dla każdego f jest określony jednoznacznie.*

Dowód. Jeśli $f-g \perp G$, to z twierdzenia Pitagorasa wynika, że dla każdego $h \in G$

$$\|f-h\|^2 = \|(f-g)+(g-h)\|^2 = \|f-g\|^2 + \|g-h\|^2 \geqslant \|f-g\|^2,$$

czyli punkt g jest optymalny. Przypuśćmy teraz, że g ma tę własność. Niech będzie też $h \in G$ i $\lambda > 0$. Wtedy

$$0 \leqslant \|f-g+\lambda h\|^2 - \|f-g\|^2 = \\ = \|f-g\|^2 + 2\lambda\langle f-g,h\rangle + \lambda^2\|h\|^2 - \|f-g\|^2 = \\ = \lambda[2\langle f-g,h\rangle + \lambda\|h\|^2].$$

Dla dostatecznie małego λ wynika stąd, że $\langle f-g,h\rangle \geqslant 0$. Tak samo rozumując, ale ze zmianą h na $-h$, dochodzimy do nierówności $\langle f-g,-h\rangle \geqslant 0$. Dlatego $\langle f-g,h\rangle = 0$. Ponieważ h jest dowolnym punktem podprzestrzeni G, więc $f-g \perp G$.

Jeśli g i g' byłyby różnymi punktami optymalnymi, to mielibyśmy relację $g-g' \perp G$. To jest jednak niemożliwe, bo $\langle g-g', g-g'\rangle > 0$. $\blacksquare$

Równania normalne. Macierz Grama

Niech $\{u_1, u_2, \ldots, u_n\}$ będzie bazą podprzestrzeni G. Zgodnie z tw. 6.8.3 element $g \in G$ jest optymalny dla danego $f \in E$ wtedy i tylko wtedy, gdy $g - f \perp G$, czyli $\langle g - f, u_i \rangle = 0$ dla $1 \leqslant i \leqslant n$. Ponieważ $g = \sum_{j=1}^{n} c_j u_j$, więc ma być spełniony następujący układ *równań normalnych*:

$$\sum_{j=1}^{n} c_j \langle u_j, u_i \rangle = \langle f, u_i \rangle \quad (1 \leqslant i \leqslant n). \tag{6.8.3}$$

Macierz tego układu, o elementach $\langle u_j, u_i \rangle$, nazywamy *macierzą Grama*.

Lemat 6.8.4. *Jeśli układ $\{u_1, u_2, \ldots, u_n\}$ jest niezależny liniowo, to macierz Grama jest nieosobliwa.*

Zamiast dowodu wystarczy przypomnieć, że na mocy tw. 6.8.3 element optymalny g jest określony jednoznacznie, wobec czego układ równań normalnych ma dokładnie jedno rozwiązanie.

Przykład 6.8.5. Stosując tw. 6.8.3, wyznaczyć wielomian

$$g(x) := c_1 x + c_3 x^3 + c_5 x^5$$

najlepiej aproksymujący funkcję $f(x) := \sin x$ w przedziale $[-1, 1]$ według normy

$$\|f\| := \left\{ \int_{-1}^{1} [f(x)]^2 \, dx \right\}^{1/2}.$$

Rozwiązanie. Funkcji g szukamy w podprzestrzeni $G := \operatorname{span}\{u_1, u_2, u_3\}$, gdzie $u_i(x) = x^{2i-1}$ dla $i = 1, 2, 3$. Układ (6.8.3) równań normalnych jest tu następujący:

$$c_1 \int_{-1}^{1} x^{2i} \, dx + c_2 \int_{-1}^{1} x^{2i+2} \, dx + c_3 \int_{-1}^{1} x^{2i+4} \, dx = \int_{-1}^{1} x^{2i-1} \sin x \, dx$$

dla $i = 1, 2, 3$. Obliczywszy te wszystkie całki, wyrażamy go w postaci

$$\begin{bmatrix} \frac{1}{3} & \frac{1}{5} & \frac{1}{7} \\ \frac{1}{5} & \frac{1}{7} & \frac{1}{9} \\ \frac{1}{7} & \frac{1}{9} & \frac{1}{11} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} \alpha - \beta \\ -3\alpha + 5\beta \\ 65\alpha - 101\beta \end{bmatrix},$$

gdzie $\alpha := \sin 1 \approx 0.84147\,0984$ i $\beta := \cos 1 \approx 0.54030\,2305$. Rozwiązując ten układ numerycznie, używamy przybliżonych wartości α i β. Daje to

współczynniki: $c_1 \approx 0.99998$, $c_2 \approx -0.16652$ i $c_3 \approx 0.00802$[8]. Powyższy układ jest jednak źle uwarunkowany (jego macierz jest podobna do macierzy Hilberta wspomnianej w podrozdz. 2.3), co oznacza, że przyjęta baza podprzestrzeni G nie jest dobra. Lepszą jest baza ortonormalna, dla której macierz Grama jest macierzą jednostkową. ■

Układy ortonormalne

Właściwym podejściem do rozważanego teraz zadania aproksymacyjnego jest zastosowanie układu ortonormalnego. Układ skończony lub nieskończony $f_1, f_2, \ldots$ elementów przestrzeni E nazywamy *ortogonalnym*, jeśli

$$\langle f_i, f_j \rangle = 0 \quad (i \neq j), \quad \langle f_i, f_i \rangle > 0$$

(mówiąc o ortogonalności dwóch elementów f i g żądamy tylko, żeby było $\langle f, g \rangle = 0$). Ten układ nazywamy *ortonormalnym*, jeśli

$$\langle f_i, f_j \rangle = \delta_{ij}.$$

Znaczenie takich układów wynika z tw. 6.8.7, które poprzedzimy dowodem *uogólnionego twierdzenia Pitagorasa.*

Lemat 6.8.6. *Jeśli układ* $\{g_1, g_2, \ldots, g_n\}$ *jest ortogonalny, to*

$$\Big\| \sum_{i=1}^{n} a_i g_i \Big\|^2 = \sum_{i=1}^{n} a_i^2 \|g_i\|^2.$$

Dowód. Tożsamość już udowodniono dla $n = 2$. Zakładamy jej poprawność dla pewnego $n \geqslant 2$. Dzięki ortogonalności elementu g_{n+1} względem $\sum_{i=1}^{n} a_i g_i$ możemy skorzystać z własności **3** z lem. 6.8.2:

$$\Big\| \sum_{i=1}^{n+1} a_i g_i \Big\|^2 = \Big\| \sum_{i=1}^{n} a_i g_i \Big\|^2 + \|a_{n+1} g_{n+1}\|^2 = \sum_{i=1}^{n+1} a_i^2 \|g_i\|^2. \qquad \blacksquare$$

Twierdzenie 6.8.7. *Jeśli układ* $\{g_1, g_2, \ldots, g_n\}$ *jest ortonormalny w przestrzeni unitarnej* E*, to dla dowolnego* $f \in E$ *elementem optymalnym w podprzestrzeni* $G := \operatorname{span}\{g_1, g_2, \ldots, g_n\}$ *jest*

$$g := \sum_{i=1}^{n} \langle f, g_i \rangle g_i, \tag{6.8.4}$$

[8] Te wielkości są bliskie początkowych współczynników: $1, -\frac{1}{6}$ i $\frac{1}{120}$ rozwinięcia funkcji $\cos x$ w szereg potęgowy (*przyp. tłum.*).

a błąd aproksymacji optymalnej jest równy

$$\|f-g\| = \Big[\|f\|^2 - \sum_{i=1}^{n}\langle f, g_i\rangle^2\Big]^{1/2}. \tag{6.8.5}$$

Dowód. Każdy punkt podprzestrzeni G jest sumą $\sum_{i=1}^{n} c_i g_i$. Jeśli jest ona optymalna dla danego f, to na mocy tw. 6.8.3 jest

$$f - \sum_{i=1}^{n} c_i g_i \perp G.$$

Ten warunek jest spełniony wtedy i tylko wtedy, gdy powyższa różnica jest ortogonalna względem każdego g_j, tj. gdy dla $1 \leqslant j \leqslant n$ jest

$$\Big\langle f - \sum_{i=1}^{n} c_i g_i, g_j\Big\rangle = \langle f, g_j\rangle - \sum_{i=1}^{n} c_i\langle g_i, g_j\rangle = \langle f, g_j\rangle - c_j = 0.$$

Ponieważ $f - g \perp G$, więc z lem. 6.8.6 wynika, że

$$\|f\|^2 = \|f-g\|^2 + \|g\|^2 = \|f-g\|^2 + \sum_{i=1}^{n}\langle f, g_i\rangle^2. \qquad \blacksquare$$

Z (6.8.5) wynika ważna *nierówność Bessela*: dla układu ortonormalnego $\{g_1, g_2, \ldots, g_n\}$ i dowolnego $f \in E$ jest

$$\sum_{i=1}^{n}\langle f, g_i\rangle^2 \leqslant \|f\|^2.$$

Z powyższego twierdzenia wynika też następująca sugestia: Jeśli chcemy aproksymować punkty przestrzeni E za pomocą punktów jej skończeniewymiarowej podprzestrzeni G, to najpierw znajdujemy bazę ortonormalną $\{g_1, g_2, \ldots, g_n\}$ tej ostatniej. Wtedy dla każdego $f \in E$ element optymalny jest sumą (6.8.4).

Aby zilustrować to postępowanie, wróćmy do przykł. 6.8.5. Bazę ortonormalną tworzą tu wielomiany Legendre'a:

$$g_1(x) := \sqrt{\frac{3}{2}}x,$$

$$g_2(x) := \sqrt{\frac{7}{8}}(5x^3 - 3x),$$

$$g_3(x) := \sqrt{\frac{11}{128}}(63x^5 - 70x^3 + 15x).$$

Rozwiązaniem zadania jest wielomian $\sum_{i=1}^{3} c_i g_i$, gdzie $c_i = \langle f, g_i \rangle$:

$$c_1 = \sqrt{\frac{3}{2}} \int_{-1}^{1} x \sin x \, dx = \sqrt{6}(\alpha - \beta),$$

$$c_2 = \sqrt{\frac{7}{8}} \int_{-1}^{1} (5x^3 - 3x) \sin x \, dx = \sqrt{\frac{7}{2}}(-18\alpha + 28\beta),$$

$$c_3 = \sqrt{\frac{11}{128}} \int_{-1}^{1} (63x^5 - 70x^3 + 15x) \sin x \, dx = \sqrt{\frac{11}{32}}(4320\alpha - 6728\beta),$$

gdzie, jak przedtem, $\alpha := \sin 1$ i $\beta := \cos 1$. Stąd $c_1 \approx 0.738$, $c_2 \approx -3.37_{10}{-2}$ i $c_3 \approx 4.34_{10}{-4}$. Ponieważ znamy bazę ortonormalną, więc koszt obliczeń tych wielkości jest mniejszy od kosztu rozwiązywania układu równań podanego w przykładzie. Ta baza jest dobrze uwarunkowana [9].

Metoda Grama-Schmidta

Zajmiemy się teraz konstrukcją bazy ortonormalnej. Służy do tego *metoda Grama-Schmidta*, znana już z podrozdz. 5.3, gdzie dotyczyła przestrzeni $\mathbb{R}^n$. Podstawowy wzór przenosi się bez zmian na dowolną przestrzeń unitarną, co pozwala pominąć dowód poniższego twierdzenia.

Twierdzenie 6.8.8. *Jeśli układ $\{v_1, v_2, \ldots, v_n\}$ jest bazą podprzestrzeni U przestrzeni unitarnej, to wzór*

$$u_i = \Big\| v_i - \sum_{j=1}^{i-1} \langle v_i, u_j \rangle u_j \Big\|^{-1} \Big[v_i - \sum_{j=1}^{i-1} \langle v_i, u_j \rangle u_j \Big] \quad (i = 1, 2, \ldots, n)$$

daje bazę ortonormalną $\{u_1, u_2, \ldots, u_n\}$ tej podprzestrzeni.

Ta metoda upraszcza się w znaczący sposób, gdy stosujemy ją do ciągu jednomianów $1, x, x^2, \ldots$. W następnym twierdzeniu może wystąpić każdy iloczyn skalarny, który ma tę własność, że dla dowolnych trzech funkcji jest $\langle fg, h \rangle = \langle f, gh \rangle$. Tak oczywiście jest, gdy

$$\langle f, g \rangle := \int_a^b f(x) g(x) w(x) \, dx. \tag{6.8.6}$$

O występującej tu *funkcji wagowej* zakładamy, że jest dodatnia w przedziale $[a, b]$ (ściślej, może ona znikać np. w końcach tego przedziału; tak jest dla funkcji $\sqrt{1 - x^2}$ w $[-1, 1]$).

[9] Warto jednak zwrócić uwagę na to, że w obu wariantach występują kombinacje liniowe o dość dużych współczynnikach wielkości α i β, które to kombinacje mają bardzo małe wartości; np. $4320\alpha - 6728\beta \approx 7.4_{10}{-4}$, chociaż odjemna i odjemnik są równe ok. 3635. Może to powodować istotną redukcję cyfr znaczących (*przyp. tłum.*).

Z tak określonym iloczynem skalarnym wiąże się ważne, nie tylko w analizie numerycznej, pojęcie: wielomiany p_n $(n = 0, 1, \ldots)$ są *ortogonalne w przedziale* $[a, b]$ *z wagą* w, jeśli stopień n-tego z nich jest równy n i jeśli

$$\langle p_m, p_n \rangle = 0 \quad \text{dla} \quad m \neq n.$$

Każdy wielomian p_n jest określony z dokładnością do czynnika stałego. Te czynniki można wybrać np. tak, że $\langle p_n, p_n \rangle = 1$ (wtedy mamy ciąg ortonormalny), albo tak, że p_n jest wielomianem standardowym (czyli współczynnik w $p_n(x)$ przy x^n jest równy 1). Poniższe twierdzenie dotyczy drugiego wariantu.

Twierdzenie 6.8.9. *Ciąg* $\{p_n\}$ *wielomianów ortogonalnych standardowych w przedziale* $[a, b]$ *z wagą* w *można wyznaczyć ze wzorów*

$$p_0(x) = 1, \quad p_1(x) = x - a_1,$$
$$p_n(x) = (x - a_n)p_{n-1}(x) - b_n p_{n-2}(x) \quad (n \geqslant 2),$$

gdzie dla iloczynu skalarnego (6.8.6) *jest*

$$a_n := \frac{\langle x p_{n-1}, p_{n-1} \rangle}{\langle p_{n-1}, p_{n-1} \rangle}, \qquad b_n := \frac{\langle x p_{n-1}, p_{n-2} \rangle}{\langle p_{n-2}, p_{n-2} \rangle}.$$

Dowód. Ze wzorów podanych w twierdzeniu wynika, że $p_n(x)$ jest wielomianem standardowym, a więc $p_n \not\equiv 0$. Wobec tego mianowniki wyrażeń dla a_n i b_n nie znikają. Udowodnimy przez indukcję względem n, że $\langle p_n, p_i \rangle = 0$ dla $i < n$. Zaczynamy od $n = 1$. Z definicji stałej a_1 wynika, że

$$\langle p_1, p_0 \rangle = \langle (x - a_1)p_0, p_0 \rangle = \langle x p_0, p_0 \rangle - a_1 \langle p_0, p_0 \rangle = 0.$$

Jeśli $n \geqslant 2$ i $\langle p_{n-1}, p_i \rangle = 0$ dla $i < n - 1$, to z definicji a_n i b_n wynika, że

$$\langle p_n, p_{n-1} \rangle = \langle x p_{n-1}, p_{n-1} \rangle - a_n \langle p_{n-1}, p_{n-1} \rangle - b_n \langle p_{n-2}, p_{n-1} \rangle = 0,$$
$$\langle p_n, p_{n-2} \rangle = \langle x p_{n-1}, p_{n-2} \rangle - a_n \langle p_{n-1}, p_{n-2} \rangle - b_n \langle p_{n-2}, p_{n-2} \rangle = 0.$$

Natomiast dla $i < n - 2$ jest

$$\begin{aligned} \langle p_n, p_i \rangle &= \langle x p_{n-1}, p_i \rangle - a_n \langle p_{n-1}, p_i \rangle - b_n \langle p_{n-2}, p_{n-1} \rangle = \\ &= \langle p_{n-1}, x p_i \rangle = \langle p_{n-1}, p_{i+1} + a_{i+1} p_i + b_{i+1} p_{i-1} \rangle = 0 \end{aligned}$$

(dla $i = 0$ drugim czynnikiem w ostatnim iloczynie skalarnym powinna być suma $p_1 + a_1 p_0$). ■

Przykład 6.8.10. Znaleźć początkowe *wielomiany Legendre'a*, z definicji ortogonalne w przedziale $[-1, 1]$ z wagą 1.

Rozwiązanie. Początek obliczeń jest następujący:

$$p_0(x) = 1, \quad a_1 = \frac{\langle xp_0, p_0\rangle}{\langle p_0, p_0\rangle} = 0,$$

$$p_1(x) = x, \quad a_2 = \frac{\langle xp_1, p_1\rangle}{\langle p_1, p_1\rangle} = 0, \quad b_2 = \frac{\langle xp_1, p_0\rangle}{\langle p_0, p_0\rangle} = \tfrac{1}{3},$$

$$p_2(x) = x^2 - \tfrac{1}{3}.$$

Trzy następne wielomiany Legendre'a obliczamy w ten sam sposób:

$$p_3(x) = x^3 - \tfrac{3}{5}x, \quad p_4(x) = x^4 - \tfrac{6}{7}x^2 + \tfrac{3}{35}, \quad p_5(x) = x^5 - \tfrac{10}{9}x^3 + \tfrac{5}{21}x.$$

■

Przykład 6.8.11. Wykazać, że wielomiany Czebyszewa (zob. podrozdz. 6.1) są ortogonalne w przedziale $[-1, 1]$ z wagą $(1 - x^2)^{-1/2}$, tj. dla iloczynu skalarnego

$$\langle f, g\rangle := \int_{-1}^{1} (1 - x^2)^{-1/2} f(x)g(x)\, dx.$$

Rozwiązanie. Podstawienie $x = \cos\theta$ upraszcza powyższą całkę:

$$\langle f, g\rangle = \int_0^{\pi} f(\cos\theta)g(\cos\theta)\, d\theta.$$

Ponieważ $T_n(x) = \cos(n \arccos x)$ (tw. 6.1.5), więc dla $n \neq m$

$$\begin{aligned}\langle T_n, T_m\rangle &= \int_0^{\pi} \cos n\theta \cos m\theta\, d\theta = \\ &= \frac{1}{2}\int_0^{\pi} [\cos(n+m)\theta + \cos(n-m)\theta]\, d\theta = \\ &= \frac{1}{2}\left[\frac{\sin(n+m)\theta}{n+m} + \frac{\sin(n-m)\theta}{n-m}\right]_0^{\pi} = 0.\end{aligned}$$

■

Algorytm

Jeśli wielomian u jest sumą $\sum_{k=0}^{n} c_k p_k$, gdzie wielomiany p_k wyznaczono za pomocą tw. 6.8.9, to wartość $u(x)$ można obliczyć, nie przekształcając tej sumy na kombinację potęg zmiennej (nie byłoby to nawet wskazane). Służy do tego następujący algorytm:

$d_{n+2} \leftarrow 0$; $d_{n+1} \leftarrow 0$
for $k = n$ **to** 0 **step** -1 **do**
    $d_k \leftarrow c_k + (x - a_{k+1})d_{k+1} - b_{k+2}d_{k+2}$
end do

Wartością $u(x)$ jest d_0.

Dowód poprawności algorytmu jest prosty:

$$u(x) = \sum_{k=0}^{n} c_k p_k(x) = \sum_{k=0}^{n} [d_k - (x - a_{k+1})d_{k+1} + b_{k+2}d_{k+2}]p_k(x) =$$
$$= d_0 p_0(x) + d_1[p_1(x) - (x - a_1)p_0(x)] +$$
$$+ \sum_{k=2}^{n} d_k[p_k(x) - (x - a_k)p_{k-1}(x) + b_k p_{k-2}(x)] = d_0.$$

TWIERDZENIE 6.8.12. *Wśród wszystkich wielomianów standardowych stopnia n najmniejszą normę* $\|\cdot\|$, *wyrażającą się wzorem* (6.8.2) *przez iloczyn skalarny, ma wielomian* p_n *z tw.* 6.8.9.

Dowód. Każdy wielomian, którego dotyczy twierdzenie, można wyrazić w postaci $p_n - \sum_{i=0}^{n-1} c_i p_i$. Na mocy tw. 6.8.3 taka różnica ma najmniejszą normę, jeśli

$$p_n - \sum_{i=0}^{n-1} c_i p_i \perp \Pi_{n-1},$$

a ta relacja ortogonalności zachodzi, gdy wszystkie c_i znikają. ■

ZADANIA 6.8

1. Znaleźć najlepsze, w sensie normy $\{\int_0^{\pi/2}[f(x)]^2\,dx\}^{1/2}$, przybliżenie λx dla funkcji $\sin x$.
2. Udowodnić, że układ ortogonalny jest niezależny liniowo.
3. Udowodnić, że jeśli f, g należą do podprzestrzeni rozpiętej na elementach układu ortonormalnego $\{u_1, u_2, \ldots, u_n\}$, to zachodzi *tożsamość Parsevala*

$$\langle f, g\rangle = \sum_{i=1}^{n} \langle f, u_i\rangle\langle g, u_i\rangle.$$

4. Udowodnić, że jeśli przybliżamy funkcję parzystą wielomianem klasy Π_n w sensie normy $\{\int_{-1}^{1}[f(x)]^2\,dx\}^{1/2}$, to i wielomian optymalny jest parzysty. Uogólnić.
5. Udowodnić, że jeśli $\{u_1, u_2, \ldots\}$ jest ciągiem nieskończonym ortonormalnym w przestrzeni unitarnej, to dla jej dowolnego elementu F jest

$$\sum_{n=1}^{\infty} \langle f, u_n\rangle^2 < \infty.$$

6. Niech będzie $\langle u, v\rangle := \int_{-1}^{1} u(x)v(x)\,dx$. Zastosować metodę Grama-Schmidta do funkcji $v_k(x) := (x^2 - 1)x^k$ $(k \geqslant 0)$. Udowodnić, że jeśli otrzymane wielomiany przenormujemy na wielomiany standardowe q_n, to spełniają one wzór rekurencyjny $q_{n+1}(x) = xq_n(x) - b_n q_{n-1}(x)$. Znaleźć wzór dla b_n. Wyznaczyć trzy początkowe wielomiany q_n.

7. Udowodnić, że w tw. 6.8.9 jest $b_n = \|p_{n-1}\|^2/\|p_{n-2}\|^2$, skąd wynika, że b_n jest dodatnie.

8. Jak zmieni się wzór rekurencyjny z tw. 6.8.9, jeśli zażądamy, żeby dawał układ ortonormalny wielomianów?

9. Niech w definicji $\langle f, g\rangle := \int_{-a}^{a} w(x)f(x)g(x)\,dx$ funkcja wagowa w będzie parzysta. Udowodnić, że wtedy we wzorze z tw. 6.8.9 jest $a_n = 0$ dla $n \geqslant 1$ i że p_n jest wielomianem tej samej parzystości co n.

10. Znaleźć wzór rekurencyjny wiążący standardowe wielokrotności $\tilde{T}_n$ wielomianów Czebyszewa T_n.

11. Sprawdzić podane w przykł. 6.8.10 wyrażenia wielomianów Legendre'a p_3, p_4 i p_5.

12. Udowodnić, że we wzorze rekurencyjnym dla wielomianów Legendre'a jest $a_n = 0$ i $b_n = (n-1)^2/[(2n-1)(2n-3)]$.

13. Stosując tw. 6.8.9 dla $[a, b] = [0, 1]$ i $w(x) = 1$, znaleźć p_0, p_1, p_2, p_3.

14. (cd.). Znaleźć wielomian $x^3 + Bx^2 + Cx + D$ ortogonalny względem Π_2. Porównać wynik z p_3 z poprzedniego zadania.

15. Opracować algorytm obliczania wartości w danym punkcie wszystkich sum $\sum_{i=0}^{k} c_i p_i$ $(k = 0, 1, \ldots, n)$. Wielomiany p_i są wyznaczane zgodnie z tw. 6.8.9. Założyć, że wielkości a_k i b_k są znane.

16. Niech A będzie przekształceniem liniowym w przestrzeni unitarnej. Zakładając, że A jest *samosprzężone*, tj. że $\langle Af, g\rangle = \langle f, Ag\rangle$ dla dowolnych f i g, udowodnić, że rozwiązania równania $Af = \lambda f$ odpowiadające różnym λ, są ortogonalne.

6.9. Aproksymacja jednostajna

W tym podrozdziale zajmiemy się aproksymacją funkcji z przestrzeni $C(X)$, gdzie X jest przestrzenią zwartą Hausdorffa[10]. Z założenia te funkcje mają wartości rzeczywiste. Czytelnicy nie chcący wgłębiać się w zagadnienia topologiczne, mogą przyjąć, że X jest podzbiorem domkniętym i ograniczonym przestrzeni $\mathbb{R}^k$; w najprostszym przypadku jest to przedział domknięty $[a, b]$ na osi rzeczywistej.

W $C(X)$ wprowadzamy normę

$$\|f\| := \max_{x \in X} |f(x)|.$$

Ta przestrzeń staje się dzięki temu przestrzenią Banacha.

[10] Pojęcia występujące tu i dalej oraz ich własności omawia np. Rudin [*2002] (*przyp. tłum.*).

Dla ustalonej podprzestrzeni skończeniewymiarowej $G \subset C(X)$ problem aproksymacji optymalnej funkcji $f \in C(X)$ jest następujący: jak w podrozdz. 6.8 określamy odległość f od G wzorem

$$\operatorname{dist}(f, G) := \inf_{g \in G} \|f - g\|$$

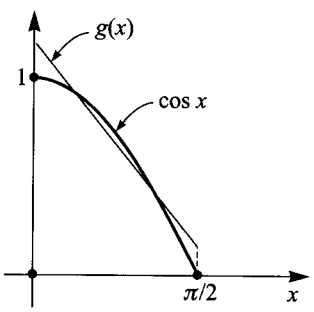
i chcemy znaleźć *element optymalny* $g \in G$, dla którego

$$\|f - g\| = \operatorname{dist}(f, G).$$

Na mocy tw. 6.8.1 ten element istnieje. Zajmiemy się więc metodami jego konstrukcji. Poniższy przykład daje pewne pojęcie o tym, jakich trudności można się tu spodziewać.

Przykład 6.9.1. Podać warunki, jakie powinien spełniać wielomian z Π_1 najlepiej przybliżający funkcję $f(x) := \cos x$ w przedziale $[0, \pi/2]$.

Rozwiązanie. Rysunek 6.12 pokazuje wykresy funkcji $\cos x$ i niezbyt odległego od niej wielomianu klasy Π_1.



RYS. 6.12. Aproksymacja funkcji $\cos x$ w przedziale $[0, \pi/2]$

Czy pokazana tam funkcja liniowa jest optymalna? Na pewno nie, bo obniżając nieco jej wykres, zmniejszamy błąd przybliżenia. Można by też nieco poprawić nachylenie prostej. Wnioskujemy stąd, że dla optymalnej funkcji liniowej g różnica $f - g$ w trzech punktach osiąga wartości ekstremalne $\pm\delta$, gdzie $\delta := \|f - g\|$; tymi punktami powinny być końce 0 i $\pi/2$ przedziału oraz jakiś punkt wewnętrzny ξ. Daje to równania

$$f(0) - g(0) = -\delta, \quad f(\xi) - g(\xi) = \delta,$$
$$f(\pi/2) - g(\pi/2) = -\delta, \quad f'(\xi) - g'(\xi) = 0.$$

Służą one do wyznaczenia czterech wielkości: δ, ξ i dwóch współczynników wielomianu g. Równania są nieliniowe względem ξ i niełatwo je rozwiązać. ■

Charakteryzacja elementu optymalnego

Lemat 6.9.2. *Element $g \in G$ jest optymalny dla f wtedy i tylko wtedy, gdy element $0 \in G$ jest optymalny dla $f - g$.*

Dowód. Jeśli g jest optymalny dla f, to $\|f - g - 0\| \leqslant \|f - g - h\|$ dla każdego $h \in G$, czyli 0 jest elementem optymalnym dla $f - g$. Równie oczywiste jest wynikanie w drugą stronę. ■

Uwzględniając powyższy lemat, chcemy zrozumieć, dla jakich elementów f przestrzeni $C(X)$ element 0 jest optymalny, tzn. kiedy

$$\|f\| = \operatorname{dist}(f, G). \tag{6.9.1}$$

Aby sformułować odpowiedź na to pytanie, definiujemy dla $f \in C(X)$ *zbiór krytyczny*:

$$\operatorname{crit}(f) := \{x \in X\colon\ |f(x)| = \|f\|\}.$$

Twierdzenie 6.9.3 (Kołmogorow). *Dla dowolnego f równość (6.9.1) zachodzi wtedy i tylko wtedy, gdy żaden element podprzestrzeni G nie ma w zbiorze $\operatorname{crit}(f)$ takich samych znaków jak f.*

Dowód. Przypuśćmy najpierw, że równość podana w twierdzeniu nie zachodzi. Wtedy istnieje takie $g \in G$, że $\|f - g\| < \|f\|$. Dla $x \in \operatorname{crit}(f)$ przyjmujemy $\sigma(x) := \operatorname{sgn} f(x)$. Jest

$$\sigma(x)[f(x)-g(x)] \leqslant |f(x)-g(x)| \leqslant \|f-g\| < \|f\| = |f(x)| = \sigma(x)f(x).$$

Tak więc $\sigma(x)g(x) > 0$, czyli znaki funkcji f i g w zbiorze $\operatorname{crit}(f)$ są identyczne.

Przypuśćmy teraz, że dla pewnej funkcji g jest $g(x)f(x) > 0$ w tymże zbiorze. Nie tracąc ogólności, można przyjąć, że $\|g\| = 1$. Ponieważ zbiór $\operatorname{crit}(f)$ jest zwarty, a iloczyn gf jest funkcją ciągłą, więc istnieje takie $\varepsilon > 0$, że $g(x)f(x) > \varepsilon$ w zbiorze $\operatorname{crit}(f)$. Niech będzie

$$\mathcal{O} := \{x \in X\colon\ g(x)f(x) > \varepsilon\}.$$

Ten zbiór jest otwarty i zawiera $\operatorname{crit}(f)$. Jego dopełnieniem jest zbiór zwarty rozłączny z $\operatorname{crit}(f)$. Dlatego

$$\rho := \max\{|f(x)|\colon\ x \in X \setminus \mathcal{O}\} < \|f\|.$$

Aproksymujmy teraz funkcję f za pomocą iloczynu λg; stałą λ odpowiednio dobierzemy. W zbiorze $\mathcal{O}$ jest

$$(f - \lambda g)^2 = f^2 - 2\lambda fg + \lambda^2 g^2 \leqslant \|f\|^2 - 2\lambda\varepsilon + \lambda^2 = \|f\|^2 - \lambda(2\varepsilon - \lambda).$$

Nierówność $(f-\lambda g)^2 < \|f\|^2$ jest więc tam prawdziwa, jeśli $0 < \lambda < 2\varepsilon$. Dla $\lambda > 0$ w zbiorze $X \setminus \mathcal{O}$ jest

$$|f - \lambda g| \leqslant |f| + \lambda |g| \leqslant \rho + \lambda.$$

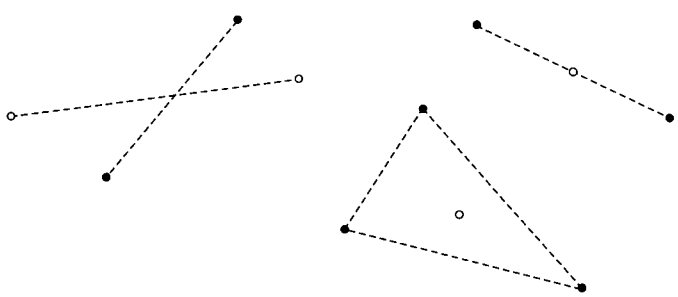
Chcemy, żeby tam było $|f-\lambda g| < \|f\|$. Tak jest, jeśli tylko $0 < \lambda < \|f\| - \rho$. Istnieje zatem takie λ, że $\|f - \lambda g\| < \|f\|$, czyli $\|f\| > \operatorname{dist}(f, G)$. ■

Z lematu 6.9.2 i tw. 6.9.3 wynika

Wniosek 6.9.4. *Funkcja $g^* \in G$ najlepiej przybliża funkcję $f \in C(X)$ wtedy i tylko wtedy, gdy nie istnieje $g \in G$, dla którego $g(x)[f(x) - g^*(x)] > 0$ na zbiorze $\{x\colon |f(x) - g^*(x)| = \|f - g^*\|\}$.*

Wniosek 6.9.5. *Wielomian $g \in \Pi_1$ jest optymalnym przybliżeniem funkcji $f \in C[a,b]$ wtedy i tylko wtedy, gdy w $[a,b]$ istnieją trzy punkty, w których różnica $f - g$ jest równa $\pm\|f - g\|$, z naprzemiennymi znakami.*

Dowód. Na mocy tw. 6.9.3 zastosowanego do $f - g$ żaden wielomian z Π_1 nie może mieć znaku zgodnego z $f - g$ na zbiorze $A := \operatorname{crit}(f - g)$. Gdyby w tym zbiorze nie było trzech punktów, w których różnica $f - g$ jest na przemian dodatnia i ujemna, to w A po lewej stronie pewnego ξ ta różnica byłaby ujemna, a po drugiej dodatnia. Wtedy dla pewnych c funkcja $c(x - \xi)$ miałaby w A znak zgodny z $f - g$. ■



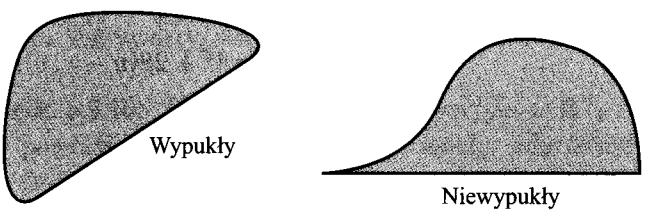
RYS. 6.13. Punkty krytyczne funkcji liniowej optymalnej

Wniosek 6.9.6. *Niech X będzie zbiorem domkniętym i ograniczonym w $\mathbb{R}^2$ i niech G będzie zbiorem funkcji liniowych $g(x,y) = a + bx + cy$. Funkcja $g \in G$ jest optymalna dla $f \in C(X)$ wtedy i tylko wtedy, gdy zbiór $\operatorname{crit}(f - g)$ zawiera punkty tworzące jeden z układów pokazanych na rys. 6.13.*

(Na tym rysunku w punktach oznaczonych ∘ i • różnica $f-g$ ma różne znaki).

Wypukłość

Zbiór K w przestrzeni liniowej jest *wypukły*, jeśli wraz ze swymi dwoma punktami zawiera odcinek, który je łączy. Inaczej mówiąc, jeśli $u, v \in K$, to dla każdego $\theta \in [0,1]$ jest $\theta u + (1-\theta)v \in K$. Przykładowe zbiory – wypukły i niewypukły – pokazano na rys. 6.14.



RYS. 6.14. Zbiór wypukły i zbiór niewypukły

Kombinacja liniowa $\sum_{i=1}^{k} \theta_i u_i$ jest *wypukła*, jeśli $\sum_{i=1}^{k} \theta_i = 1$ i $\theta_i \geqslant 0$. Zbiór wszystkich kombinacji wypukłych punktów z danego zbioru S, czyli

$$\Big\{\sum_{i=1}^{k} \theta_i u_i \colon\ k \in \mathbb{N},\ u_i \in S,\ \sum_{i=1}^{k} \theta_i = 1,\ \theta_i \geqslant 0\Big\},$$

nazywamy jego *powłoką wypukłą*.

Lemat 6.9.7. *Każdy zbiór domknięty wypukły w $\mathbb{R}^n$ (lub w dowolnej przestrzeni Hilberta) zawiera dokładnie jeden punkt o minimalnej normie. Ponadto $0 \notin K$ wtedy i tylko wtedy, gdy istnieje takie v, że $\langle v, u\rangle > 0$ dla każdego $u \in K$.*

Dowód. Niech będzie $\rho := \inf_{u\in K}\|u\|$ i niech ciąg punktów $u_j \in K$ będzie taki, że $\{\|u_j\|\} \to \rho$. Jest to ciąg Cauchy'ego. Istotnie, skorzystajmy z własności **5** z lem. 6.8.2 i z wypukłości zbioru K:

$$\begin{aligned}\|u_i - u_j\|^2 &= 2\|u_i\| + 2\|u_j\|^2 - \|u_i + u_j\|^2 = \\ &= 2\|u_i\|^2 + 2\|u_j\|^2 - 4\|(u_i + u_j)/2\|^2 \leqslant \\ &\leqslant 2\|u_i\|^2 + 2\|u_j\|^2 - 4\rho^2 \to 2\rho^2 + 2\rho^2 - 4\rho^2 = 0.\end{aligned}$$

Ponieważ przestrzeń jest zupełna, więc ciąg $\{u_j\}$ jest zbieżny. Jego granica u należy do K, bo ten zbiór jest domknięty. Jest też $\|u\| = \rho$. Jednoznaczność

punktu o minimalnej normie wynika z już zastosowanej własności: jeśli u i u' mają normę ρ, to

$$\|u-u'\|^2 \leqslant 2\|u\|^2 + 2\|u'\|^2 - 4\rho^2 = 0.$$

Jeśli istnieje v takie, że $\langle v,u\rangle > 0$ dla wszystkich $u \in K$, to oczywiście $0 \notin K$. Przypuśćmy, że punkt 0 ma tę własność. Jeśli zatem $v \in K$ jest punktem o najmniejszej normie, to $\|v\| > 0$. Wtedy dla dowolnego $u \in K$ i $\theta \in (0,1)$ jest

$$0 \leqslant \|\theta u + (1-\theta)v\|^2 - \|v\|^2 = \|\theta(u-v)+v\|^2 - \|v\|^2 = \\ = \theta^2\|u-v\|^2 + 2\theta\langle u-v, v\rangle,$$

skąd

$$0 \leqslant \theta\|u-v\|^2 + 2\langle u-v, v\rangle.$$

θ może być dowolnie małe, więc $\langle u-v, v\rangle \geqslant 0$, czyli $\langle v,u\rangle \geqslant \langle v,v\rangle > 0$. ■

Twierdzenie 6.9.8 (Carathéodory). *Jeśli S jest podzbiorem n-wymiarowej przestrzeni liniowej, to każdy punkt powłoki wypukłej zbioru S jest kombinacją wypukłą co najwyżej $n+1$ jego punktów.*

Dowód. Niech p będzie punktem powłoki wypukłej zbioru S. Przesunięcie tego zbioru polegające na zamianie każdego $s \in S$ na $s-p$ powoduje, że kombinacje wypukłe i powłoka wypukła przesuwają się w ten sam sposób. Dlatego nie tracąc ogólności możemy przyjąć, że badanym punktem tej powłoki jest 0. Z jej definicji wynika, że $0 = \sum_{i=1}^{k} \theta_i u_i$, gdzie $u_i \in S$, $\sum_{i=1}^{k} \theta_i = 1$ i $\theta_i \geqslant 0$. Wybierzmy tę kombinację o wartości 0, w której k jest najmniejsze; wtedy wszystkie θ_i są dodatnie. Jeśli $k \leqslant n+1$, to twierdzenie jest udowodnione. W przeciwnym razie z założenia o wymiarze przestrzeni wynika, że zachodzi nietrywialna tożsamość $\sum_{i=2}^{k} \lambda_i u_i = 0$. Niech też będzie $\lambda_1 := 0$. Dla dowolnego t rzeczywistego $0 = \sum_{i=1}^{k} (\theta_i + t\lambda_i)u_i$. Funkcja

$$\varphi(t) := \min_{1\leqslant i\leqslant k} (\theta_i + t\lambda_i)$$

jest ciągła i taka, że $\varphi(0) > 0$. Natomiast dla pewnych t jest $\varphi(t) < 0$. Istnieje więc takie t_0, że $\varphi(t_0) = 0$. Niech będzie $\theta_i' := \theta_i + t_0\lambda_i$. Stąd $\theta_i' \geqslant 0$, $\min \theta_i' = 0$ i $\theta_1' = \theta_1 > 0$. Dlatego w kombinacji liniowej $\sum_{i=1}^{k} \theta_i' u_i$ równej 0 pewien składnik znika. Dzieląc ją przez $\sum \theta_i'$ wyrażamy 0 jako kombinację wypukłą $k-1$ elementów zbioru S. To jednak przeczy założeniu, że k jest minimalne. ■

Lemat 6.9.9. *Powłoka wypukła zbioru zwartego w skończeniewymiarowej przestrzeni unormowanej jest zwarta.*

Dowód. Niech S będzie zbiorem zwartym w przestrzeni n-wymiarowej. Definiujemy zbiór V wszystkich układów $v := (\theta_0, \theta_1, \ldots, \theta_n, u_0, u_1, \ldots, u_n)$, gdzie $u_i \in S$, $\sum_{i=0}^{n} \theta_i = 1$ i $\theta_i \geqslant 0$. Są to punkty iloczynu kartezjańskiego $\mathbb{R}^{n+1} \times X^{n+1}$. Wymiar tej przestrzeni wynosi $n + 1 + (n+1)n = (n+1)^2$. Zbiór V jest domknięty i ograniczony, a więc zwarty. Funkcja f taka, że $f(v) := \sum_{i=0}^{n} \theta_i u_i$, odwzorowuje V na powłokę wypukłą zbioru S. Na mocy tw. 6.9.8 funkcja f jest suriektywna; jest też ciągła. Ponieważ obraz ciągły zbioru zwartego jest zwarty, więc ta powłoka jest zwarta. ■

Twierdzenie 6.9.10. *Jeśli S jest zbiorem zwartym w $\mathbb{R}^n$, to 0 nie należy do jego powłoki wypukłej wtedy i tylko wtedy, gdy istnieje takie v, że $\langle v, u\rangle > 0$ dla każdego $u \in S$.*

Dowód. Jeśli 0 należy do powłoki wypukłej C zbioru S, to dla pewnych $u_i \in S$ i $\theta_i > 0$ jest $0 = \sum_{i=1}^{k} \theta_i u_i$. Dla dowolnego $v \in \mathbb{R}^n$

$$0 = \langle v, 0\rangle = \sum_{i=1}^{k} \theta_i \langle v, u_i\rangle.$$

Wobec tego nie wszystkie liczby $\langle v, u_i\rangle$ są dodatnie. Jeśli zaś $0 \notin C$, to ponieważ na mocy lem. 6.9.9 zbiór C jest zwarty, więc z lem. 6.9.7 wynika istnienie takiego v, że $\langle v, u\rangle > 0$ dla każdego $u \in S$. ■

Rozwiązywanie układów liniowych w sensie Czebyszewa

Wniosek 6.9.11. *Niech A będzie macierzą $m \times n$, $b \in \mathbb{R}^m$, $x \in \mathbb{R}^n$,*

$$\sigma_i := \operatorname{sgn}(Ax - b)_i, \quad I := \{i\colon\ |(Ax - b)_i| = \|Ax - b\|_\infty\}.$$

Norma $\|Ax - b\|_\infty$ jest najmniejsza wtedy i tylko wtedy, gdy 0 należy do powłoki wypukłej zbioru $\{\sigma_i A_i\colon\ i \in I\}$, gdzie A_i jest i-tym wierszem macierzy A.

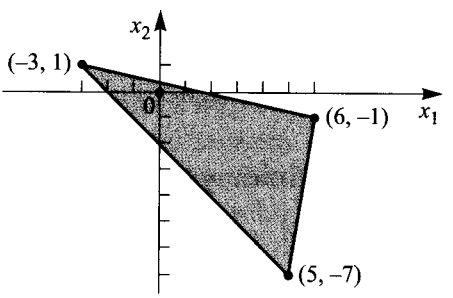
Dowód. Wyznaczanie minimum normy $\|Ax - b\|_\infty$ polega na szukaniu takiej kombinacji liniowej kolumn macierzy A, która byłaby najbliższa wektora b. Interpretujemy wektory kolumnowe jako funkcje wskaźnika o wartościach $1, 2, \ldots, m$. Elementem przybliżanym jest b, a podprzestrzeń G przybliżeń jest rozpięta na kolumnach macierzy A. Zgodnie z tw. 6.9.3 rozwiązanie x charakteryzuje się tym, że w żadnym wektorze z G składowe

o wskaźnikach z I nie mają znaków σ_i. Inaczej mówiąc, układ $\sigma_i(Av)_i > 0$ $(i \in I)$ jest sprzeczny. Ponieważ te nierówności można wyrazić w postaci $\langle \sigma_i A_i, v\rangle > 0$ $(i \in I)$, więc na mocy tw. 6.9.10 równoważnym warunkiem jest ten, podany we wniosku, który dotyczy wektora 0. ■

Przykład 6.9.12. Stosując wniosek 6.9.11 sprawdzić, czy wektor $x = (2, 3)$ jest rozwiązaniem w sensie Czebyszewa układu

$$\begin{aligned} 5x_1 - 7x_2 &= -14,\\ 3x_1 + x_2 &= 8,\\ x_1 - 9x_2 &= -23,\\ 3x_1 - x_2 &= 6,\\ 6x_1 - x_2 &= 6. \end{aligned}$$

Rozwiązanie. Reszty $r_i := \langle A_i, x\rangle - b_i$ są równe 3, 1, −2, −3, 3, a zatem σ_i są równe 1, 1, −1, −1, 1. Zbiorem krytycznym jest $I = \{1, 4, 5\}$. Z wniosku 6.9.11 wynika, że x jest szukanym rozwiązaniem, jeśli wektor 0 należy do powłoki wypukłej zbioru $\{\sigma_1 A_1, \sigma_4 A_4, \sigma_5 A_5\}$. Na rysunku 6.15 zaznaczono te trzy punkty i określony przez nie zbiór wypukły; punkt 0 doń należy. ■



RYS. 6.15. Zero leży w powłoce wypukłej trzech punktów

Inne twierdzenie o charakteryzacji

W następnych twierdzeniach G oznacza podprzestrzeń n-wymiarową przestrzeni $C(X)$, mającą bazę $\{g_1, g_2, \ldots, g_n\}$. Wtedy

$$\vec{g}(x) := (g_1(x), g_2(x), \ldots, g_n(x)) \quad (x \in X).$$

Jest to punkt w $\mathbb{R}^n$.

Dla każdego $x \in X$ definiujemy funkcjonał liniowy $\hat{x}$ na $C(X)$ wzorem

$$\hat{x}(f) := f(x) \quad (f \in C(X)).$$

Twierdzenie 6.9.13. *Jeśli $f \in C(X)$, to następujące zdania są równoważne:*

1. $\|f\| = \operatorname{dist}(f, G)$.
2. *Żaden element z G nie ma znaków takich jak $f(x)$ na zbiorze* $\operatorname{crit}(f)$.
3. 0 *leży w powłoce wypukłej zbioru* $\{f(x)\vec{g}(x)\colon x \in \operatorname{crit}(f)\}$.
4. *Istnieje funkcjonał* $\sum_{i=1}^{k} \lambda_i \hat{x}_i$ *znikający na G i spełniający warunki* $x_i \in \operatorname{crit}(f)$, $\lambda_i f(x_i) > 0$ *i* $k \leqslant n+1$.

Dowód. Implikacja $\mathbf{1} \Rightarrow \mathbf{2}$ jest częścią tw. 6.9.3.

Aby sprawdzić implikację $\mathbf{2} \Rightarrow \mathbf{3}$, zakładamy, że pierwsze zdanie jest prawdziwe. Wtedy nie można znaleźć liczb c_i, dla których

$$\sum_{i=1}^{n} c_i f(x) g_i(x) > 0 \quad (x \in \operatorname{crit}(f)).$$

Ponieważ tę nierówność można napisać w postaci $\langle c, f(x)\vec{g}(x)\rangle > 0$, więc na mocy tw. 6.9.10 punkt 0 leży w powłoce wypukłej zbioru

$$\{f(x)\vec{g}(x)\colon x \in \operatorname{crit}(f)\}.$$

Sprawdzamy implikację $\mathbf{3} \Rightarrow \mathbf{4}$. Z założenia **3** i tw. 6.9.8 wynika, że

$$0 = \sum_{i=1}^{k} \theta_i f(x_i)\vec{g}(x_i),$$

gdzie $k \leqslant n+1$, $\theta_i > 0$ i $x_i \in \operatorname{crit}(f)$. Dla $\lambda_i := \theta_i f(x_i)$ jest $\lambda_i f(x_i) > 0$ i $0 = \sum_{i=1}^{k} \lambda_i \vec{g}(x_i)$, czyli – wobec definicji $\vec{g}$ –

$$0 = \sum_{i=1}^{k} \lambda_i g_j(x_i) = \Big(\sum_{i=1}^{k} \lambda_i \hat{x}_i\Big)(g_j) \quad (j = 1, 2, \ldots, n).$$

Ponieważ podprzestrzeń G jest rozpięta na funkcjach g_i, więc wykazaliśmy, że funkcjonał $\sum_{i=1}^{k} \lambda_i \hat{x}_i$ zeruje się na G.

Na koniec sprawdzamy implikację $\mathbf{4} \Rightarrow \mathbf{1}$. Z założenia **4** wynika, że jeśli $h \in G$, to

$$\|f\| \sum_{i=1}^{k} |\lambda_i| = \sum_{i=1}^{k} \lambda_i f(x_i) = \sum_{i=1}^{k} \lambda_i [f(x_i) - h(x_i)] \leqslant \|f - h\| \sum_{i=1}^{k} |\lambda_i|.$$

Dlatego $\|f\| \leqslant \|f - h\|$ i zdanie **1** jest prawdziwe. ■

Przykład 6.9.14. Niech będzie $X := [0,1]$ i niech podprzestrzeń G będzie rozpięta na funkcjach $g_1(x) := 1 - 2x^2$ i $g_2(x) := x - x^2$. Wykazać, że każda funkcja $f \in C[0,1]$, dla której $f(0) = f(1) = \|f\|$, jest taka, że $\|f\| = \operatorname{dist}(f, G)$.

Rozwiązanie. Sprawdzamy własność **4** z tw. 6.9.13. Dla $g \in G$ jest $g(0) + g(1) = 0$. Możemy więc tam przyjąć, że $x_1 = 0$, $x_2 = 1$, $\lambda_1 = \lambda_2 = 0$. ■

Podprzestrzeń Haara

Podprzestrzeń n-wymiarową G przestrzeni $C(X)$ nazywamy *podprzestrzenią Haara*, jeśli każda nieznikająca tożsamościowo funkcja z G ma co najwyżej $n-1$ zer w X. W szczególności zbiór Π_n wielomianów jest taką przestrzenią $(n+1)$-wymiarową.

Poniższy lemat świadczy o tym, że to pojęcie jest idealnie dobrane do zadań interpolacyjnych.

Lemat 6.9.15. *Podprzestrzeń n-wymiarowa G przestrzeni $C(X)$ jest podprzestrzenią Haara wtedy i tylko wtedy, gdy dla dowolnych liczb rzeczywistych $\lambda_1, \lambda_2, \ldots, \lambda_n$ i dowolnych różnych punktów $x_1, x_2, \ldots, x_n$ z X istnieje dokładnie jedna funkcja $g \in G$ taka, że $g(x_i) = \lambda_i$ dla $i = 1, 2, \ldots, n$.*

Dowód. Niech $\{g_1, g_2, \ldots, g_n\}$ będzie bazą zbioru G. Zadanie interpolacyjne polega na znalezieniu takich $c_1, c_2, \ldots, c_n$, że

$$\sum_{j=1}^{n} c_j g_j(x_i) = \lambda_i \quad (1 \leqslant i \leqslant n).$$

Jeśli ten układ możemy rozwiązać dla dowolnych λ_i, to macierz o elementach $g_j(x_i)$ jest nieosobliwa, a podobne zadanie jednorodne (ze wszystkimi $\lambda_i = 0$) ma tylko zerowe rozwiązanie. Tak więc żadna kombinacja liniowa funkcji g_j nie może zerować się we wszystkich x_i. To właśnie charakteryzuje podprzestrzeń Haara. ■

Twierdzenie 6.9.16. *Jeśli $G \subset C(X)$ jest podprzestrzenią n-wymiarową Haara, a $f \in C(X)$, to $\|f\| = \operatorname{dist}(f, G)$ wtedy i tylko wtedy, gdy istnieje funkcjonał postaci $\sum_{i=1}^{n+1} \lambda_i \widehat{x}_i$, zerujący się na G i taki, że $x_i \in \operatorname{crit}(f)$ i $\lambda_i f(x_i) > 0$.*

Dowód. Twierdzenie 6.9.13 zawiera ten sam warunek konieczny i dostateczny ze zmianą $n+1$ na k. Równanie $\sum_{i=1}^{k} \lambda_i g(x_i) = 0$ może być

nietrywialnie spełnione tylko dla $k \geqslant n+1$. Istotnie, niech będzie $k \leqslant n$. Korzystając z lem. 6.9.15, wybieramy funkcję $g \in G$ taką, że $g(x_i) = \lambda_i$. Wtedy $0 = \sum_{i=1}^{k} \lambda_i g(x_i) = \sum_{i=1}^{k} \lambda_i^2$. ■

Jednoznaczność aproksymacji optymalnej

Twierdzenie 6.9.17. *Jeśli G jest podprzestrzenią skończeniewymiarową Haara w $C(X)$, a f takim elementem przestrzeni $C(X)$, że $\|f\| = \operatorname{dist}(f, G)$, to istnieje stała dodatnia γ (zależna od f) taka, że $\|f+g\| \geqslant \|f\| + \gamma\|g\|$.*

Dowód. Niech n będzie wymiarem zbioru G. Na mocy tw. 6.9.16 istnieją punkty $x_0, x_1, \ldots, x_n$ w zbiorze $\operatorname{crit}(f)$ i liczby dodatnie θ_i takie, że $\sum_{i=0}^{n} \theta_i \sigma_i g(x_i) = 0$ dla wszystkich $g \in G$, gdzie $\sigma_i := \operatorname{sgn} f(x_i)$. Niech $h \in G$ ma normę 1. Z ostatniej równości dla $g = h$ wynika, że co najmniej jedna z liczb $\sigma_i h(x_i)$ jest dodatnia. Tym bardziej jest $\max_i \sigma_i h(x_i) > 0$. Ponieważ lewa strona tej nierówności zależy w sposób ciągły od h, a zbiór wszystkich funkcji $h \in G$ o normie 1 jest zwarty, więc

$$\gamma := \inf_h \max_i \sigma_i h(x_i) > 0.$$

Jeśli $g \in G$ i $g \neq 0$, to dla pewnego i zachodzi nierówność $\sigma_i g(x_i)/\|g\| \geqslant \gamma$. Dlatego

$$\|f+g\| \geqslant \sigma_i f(x_i) + \sigma_i g(x_i) \geqslant \|f\| + \gamma\|g\|. \quad ■$$

Wniosek 6.9.18. *Jeśli G jest podprzestrzenią skończeniewymiarową Haara w $C(X)$, to każda funkcja z $C(X)$ ma dokładnie jedno najlepsze przybliżenie w G.*

Dowód. Jeśli $g \in G$ najlepiej przybliża funkcję f, to dla różnicy $f-g$ najlepszym przybliżeniem w G jest 0. Niech będzie $h \in G$ i $h \neq 0$. Wtedy na mocy tw. 6.9.17

$$\|f-g+h\| \geqslant \|f-g\| + \gamma\|h\| > \|f-g\|. \quad ■$$

Twierdzenie 6.9.19. *Jeśli G jest podprzestrzenią skończeniewymiarową Haara w $C(X)$ i jeśli A jest odwzorowaniem funkcji z $C(X)$ na jej najlepsze przybliżenie w G, to dla każdej funkcji $f \in C(X)$ istnieje liczba dodatnia $\lambda(f)$ taka, że*

$$\|Af - Ah\| \leqslant \lambda(f)\|f-h\| \quad (h \in C(X)).$$

Na mocy tego twierdzenia najlepsze przybliżenie funkcji zależy od niej w sposób ciągły.

Dowód. Zgodnie z tw. 6.9.17 istnieje liczba $\gamma(f) > 0$ taka, że dla każdego $g \in G$

$$\|f - g\| \geqslant \|f - Af\| + \gamma(f)\|Af - g\|.$$

Stąd dla $g := Ah$ wynika, że

$$\begin{aligned}\gamma(f)\|Af - Ah\| &\leqslant \|f - Ah\| - \|f - Af\| \leqslant \\ &\leqslant \|f - h\| + \|h - Ah\| - \|f - Af\| \leqslant \\ &\leqslant \|f - h\| + \|h - Af\| - \|f - Af\| \leqslant \\ &\leqslant \|f - h\| + \|h - f\| + \|f - Af\| - \|f - Af\| = \\ &= 2\|f - h\|.\end{aligned}$$

■

Alternans

Lemat 6.9.20. *Niech G będzie n-wymiarową podprzestrzenią Haara przestrzeni $C[a,b]$. Jeżeli istnieją punkty x_i takie, że $a \leqslant x_1 < x_2 < \ldots < x_n \leqslant b$ i że $\sum_{i=0}^{n} \lambda_i g(x_i) = 0$ dla każdego $g \in G$, gdzie $\sum_{i=0}^{n} |\lambda_i| > 0$, to wielkości λ_i są na przemian dodatnie i ujemne.*

Dowód. Dla każdego $j = 1, 2, \ldots, n$ rozumujemy jak następuje: na mocy lem. 6.9.15 istnieje dokładnie jedna funkcja $g_j \in G$ taka, że $g_j(x_i) = \delta_{ji}$ dla $0 \leqslant i \leqslant n$, $i \neq j - 1$. Wtedy

$$0 = \sum_{i=0}^{n} \lambda_i g_j(x_i) = \lambda_{j-1} g_j(x_{j-1}) + \lambda_j. \tag{6.9.2}$$

Musi tu być $g_j(x_{j-1}) > 0$, bo w przeciwnym razie funkcja g_j, nieznikająca tożsamościowo, miałaby n zer wbrew definicji podprzestrzeni Haara. Z (6.9.2) wynika więc, że gdyby którakolwiek z liczb λ_j była równa 0, to i pozostałe by znikały. To jednak przeczy założeniu. Dlatego wobec (6.9.2) liczby λ_{j-1} i λ_j mają przeciwne znaki. ■

Twierdzenie 6.9.21. *Niech G będzie n-wymiarową podprzestrzenią Haara przestrzeni $C[a,b]$, a X – podzbiorem domkniętym przedziału $[a,b]$. Dla każdej funkcji $f \in C(X)$ równość $\|f\| = \operatorname{dist}(f, G)$ zachodzi wtedy i tylko wtedy, gdy w X istnieją punkty $x_0 < x_1 < \ldots < x_n$ takie, że $f(x_{i-1})f(x_i) = -\|f\|^2$ dla $1 \leqslant i \leqslant n$.*

Dowód. Na mocy tw. 6.9.13 równość $\|f\| = \operatorname{dist}(f, G)$ jest równoważna istnieniu punktów $x_i \in \operatorname{crit}(f)$ i współczynników λ_i takich, że

$$x_0 < x_1 < \ldots < x_n, \quad \sum_{i=0}^{n} \lambda_i g(x_i) = 0 \quad \text{dla każdego } g \in G, \ \lambda_i f(x_i) > 0.$$

Jeśli zaś tak jest, to liczby λ_i, a zatem również $f(x_i)$, mają naprzemienne znaki. To dowodzi wynikania w jedną stronę.

Jeśli istnieją punkty x_i o własnościach podanych w twierdzeniu, to należą one do zbioru $\operatorname{crit}(f)$, a wartości $f(x_i)$ mają naprzemienne znaki. Niech będzie np. $\operatorname{sgn} f(x_i) = (-1)^i$. Przypuśćmy, że wbrew twierdzeniu istnieje taka funkcja $g \in G$, że $\|f - g\| < \|f\|$. Wtedy

$$(-1)^i[f(x_i) - g(x_i)] \leqslant \|f - g\| < \|f\| = (-1)^i f(x_i).$$

Dlatego $(-1)^i g(x_i) > 0$ i funkcja g ma co najmniej n zer, co jest sprzeczne z definicją przestrzeni Haara. Zauważmy, że te zera leżą w przedziale $[a, b]$, ale niekoniecznie w X. Dlatego właśnie było konieczne założenie, że G jest podprzestrzenią Haara przestrzeni $C[a, b]$. ■

Wniosek 6.9.22. *Jeśli $G \subset C[a, b]$ jest n-wymiarową podprzestrzenią Haara, to funkcja $g \in G$ jest najlepszym przybliżeniem funkcji $f \in C[a, b]$ wtedy i tylko wtedy, gdy w $[a, b]$ istnieją punkty $x_0 < x_1 < \ldots < x_n$ takie, że*

$$f(x_i) - g(x_i) = (-1)^i \sigma \|f - g\| \quad (0 \leqslant i \leqslant n, \quad |\sigma| = 1).$$

Każdy układ $\{x_0, x_1, \ldots, x_n\}$ punktów o własnościach wymienionych we wniosku nazywamy *alternansem*; nie zawsze jest on określony jednoznacznie.

Algorytmy

Ostatni fragment tego podrozdziału dotyczy zagadnień numerycznych aproksymacji jednostajnej. Naszkicujemy tu niektóre znane algorytmy. Podstawowym zadaniem jest minimalizacja wyrażenia

$$\Delta(c) := \Big\| f - \sum_{i=1}^{n} c_i g_i \Big\|_\infty = \sup_{x \in X} \Big| f(x) - \sum_{i=1}^{n} c_i g_i(x) \Big|,$$

gdzie funkcje $f, g_1, g_2, \ldots, g_n$ są dane, a szukany jest wektor c.

To zadanie możemy rozwiązywać, stosując tzw. *pierwszy algorytm Remeza*, który w każdej z kolejnych iteracji wymaga rozwiązania podobnego, ale bardziej elementarnego zadania; opiszemy go teraz.

Niech będzie $G := \operatorname{span}\{g_1, g_2, \ldots, g_n\}$. W k-tym kroku algorytmu jest dany podzbiór skończony X_k zbioru X. Wiąże się z nim *półnorma* w $C(X)$ określona wzorem

$$\|f\|_k := \max_{x \in X_k} |f(x)|$$

(półnorma nie ma tylko jednej istotnej własności normy: może znikać dla niezerowej funkcji i tutaj może się to oczywiście zdarzyć). Stosując metodę opisaną dalej, wyznaczamy funkcję $h_k \in G$, dla której półnorma $\|f - h_k\|_k$ jest najmniejsza. Mamy tu znów zadanie aproksymacji jednostajnej, ale na zbiorze skończonym X_k. Następną czynnością jest znalezienie w X takiego punktu x_k, że

$$|f(x_k) - h_k(x_k)| = \|f - h_k\|.$$

Zbiór X_{k+1} powstaje z X_k przez dołączenie punktu x_{k+1}. Następna iteracja przebiega tak samo. Przy tym zwykle funkcja h_k wyznaczona w k-tym kroku jest dobrym przybliżeniem początkowym, gdy wyznaczamy h_{k+1}.

Twierdzenie 6.9.23. *Jeśli początkowa półnorma $\|\cdot\|_1$ jest równa normie w $C(X)$, to $\lim_{k\to\infty}\|f - h_k\| = \operatorname{dist}(f, G)$. Ciąg $\{h_k\}$ ma punkty skupienia, a każdy z nich jest najlepszym przybliżeniem dla f.*

Dowód. Wprost z definicji półnorm i zbiorów X_k wynika, że dla $g \in G$ i $1 \leqslant k \leqslant i$ zachodzi nierówność

$$\|f - g\|_1 \leqslant \|f - g\|_k \leqslant \|f - g\|_i \leqslant \|f - g\|,$$

wobec czego

$$\|f - h_k\|_1 \leqslant \|f - h_k\|_k \leqslant \|f - h_i\|_i \leqslant \operatorname{dist}(f, G).$$

Ciąg $\{\|h_k\|_1\}$ jest zatem ograniczony. Ponieważ w przestrzeni skończeniewymiarowej wszystkie normy są równoważne, więc także ciąg $\{\|h_k\|\}$ jest ograniczony. Dlatego pewien podciąg ciągu $\{h_k\}$ jest zbieżny do h^*. Dla danego $\varepsilon > 0$ wybierzmy k tak, żeby było $\|h_k - h^*\| < \varepsilon$. Niech $i > k$ będzie takie, że $\|h_i - h^*\| < \varepsilon$. Wtedy

$$\begin{aligned}\operatorname{dist}(f, G) &\leqslant \|f - h^*\| \leqslant \|f - h_k\| + \|h_k - h^*\| \leqslant \\ &\leqslant |f(x_k) - h_k(x_k)| + \varepsilon \leqslant \|f - h_k\|_i + \varepsilon \leqslant \\ &\leqslant \|f - h_i\|_i + \|h_i - h^*\|_i + \|h^* - h_k\|_i + \varepsilon \leqslant \operatorname{dist}(f, G) + 3\varepsilon.\end{aligned}$$

Wobec dowolności ε wnioskujemy stąd, że $\|f - h^*\| = \operatorname{dist}(f, G)$.

Pozostaje wykazać, że wielkości $d_k := \|f - h_k\|$ dążą do $\operatorname{dist}(f, G)$. Ich ciąg jest ograniczony. Istnieje zatem podciąg $\{d_{k_i}\}$ zbieżny, np. do d^*. Niech h' będzie punktem skupienia podciągu $\{h_{k_i}\}$. Z pierwszej części dowodu wynika, że $d^* = \|f - h'\| = \operatorname{dist}(f, G)$. Dlatego ciąg $\{d_k\}$ ma tylko jeden punkt skupienia, mianowicie $\operatorname{dist}(f, G)$. Daje to zbieżność, którą należało udowodnić. ■

Warto zauważyć, że w każdym kroku pierwszego algorytmu Remeza można łatwo oszacować z obu stron szukaną wielkość $d^* = \operatorname{dist}(f, G)$. Istotnie,

$$\|f - h_k\|_k \leqslant d^* \leqslant \min_{1 \leqslant i \leqslant k} \|f - h_i\|.$$

Te oszacowania z dołu i z góry dążą monotonicznie do d^* gdy $k \to \infty$; pierwsze rośnie, a drugie maleje. Najprostsze oszacowanie z góry, mianowicie $\|f - h_k\|$, też dąży do d^*, ale nie zawsze monotonicznie.

W pewnym wariancie tego algorytmu, zwanym *metodą wymiany*, wszystkie zbiory X_k składają się z $n+1$ punktów, gdyż przejście do X_{k+1} polega na wymianie jednego z punktów poprzedniego zbioru na nowy. W praktyce stosuje się zwykle ten właśnie wariant, chociaż jego poprawność wymaga dodatkowych założeń o funkcjach bazowych $g_1, g_2, \ldots, g_n$[11)].

Z twierdzenia 6.9.13 wynika, że w k-tym kroku wektor $0 \in \mathbb{R}^n$ leży w powłoce wypukłej $n+1$ wektorów

$$[f(x) - h_k(x)]\,(g_1(x), g_2(x), \ldots, g_n(x)) \quad (x \in X_k). \tag{6.9.3}$$

Znając h_k, znajdujemy punkt x_k określony przed tw. 6.9.23; jest to punkt krytyczny funkcji $f - h_k$. To on właśnie zastępuje jeden z punktów zbioru X_k. Usuwany punkt wybieramy tak, aby wektor 0 leżał w powłoce wypukłej układu analogicznego do (6.9.3), ale dla $x \in X_{k+1}$. Możliwość takiej wymiany wynika z poniższego twierdzenia.

Twierdzenie 6.9.24. *Niech $u_0, u_1, \ldots, u_{n+1}$ będą różnymi punktami przestrzeni $\mathbb{R}^n$. Jeśli 0 leży w powłoce wypukłej układu $\{u_0, u_1, \ldots, u_n\}$, to jest tak również po zmianie jednego z punktów u_k dla $k \leqslant n$ na u_{n+1}.*

Dowód. Przypuśćmy najpierw, że

$$0 = \sum_{i=0}^{n} \theta_i u_i, \tag{6.9.4}$$

[11)] Można też wymieniać – według określonych reguł – wszystkie punkty zbioru X_k na nowe. Jest to tzw. *drugi algorytm Remeza*, którego zbieżność np. w aproksymacji wielomianowej jest bardzo szybka; zob. Meinardus [*1968] (*przyp. tłum.*).

gdzie $\sum_{i=0}^{n}\theta_i = 1$, $\theta_i \geqslant 0$ i dla pewnego k jest $\theta_k = 0$. Wtedy pierwsza równość pozostaje prawdziwa po zmianie u_k na u_{k+1}.

Przypuśćmy teraz, że w (6.9.4) wszystkie θ_i są dodatnie. Z twierdzenia 6.9.8 wynika, że na wektorach $u_0, u_1, \ldots, u_n$ można rozpiąć przestrzeń n-wymiarową; oczywiście jest nią $\mathbb{R}^n$. Wobec tego istnieją takie λ_i, że $u_{n+1} = \sum_{i=0}^{n}\lambda_i u_i$. Niech iloraz λ_i/θ_i będzie największy dla $i = k$. Niech będzie $\theta'_i := \lambda_k\theta_i - \lambda_i\theta_k$ dla $0 \leqslant i \leqslant n$ oraz $\theta'_{n+1} := \theta_k$. Oczywiście $\theta'_k = 0$, a ogólniej $\theta'_i \geqslant 0$, gdyż $\theta'_{n+1} = \theta_k$, a dla $i \leqslant n$

$$\theta'_i = (\lambda_k/\theta_k - \lambda_i/\theta_i)\theta_i\theta_k.$$

Można też sprawdzić, że

$$\begin{aligned}\sum_{i=0}^{n+1}\theta'_i u_i &= \sum_{i=0}^{n}(\lambda_k\theta_i - \lambda_i\theta_k)u_i + \theta_k u_{n+1} = \\ &= \lambda_k \sum_{i=0,\, i\neq k}^{n} \theta_i u_i - \theta_k \sum_{i=0,\, i\neq k}^{n} \lambda_i u_i + \theta_k u_{n+1} = \\ &= \lambda_k(-\theta_k u_k) - \theta_k(u_{n+1} - \lambda_k u_k) + \theta_k u_{n+1} = 0.\end{aligned}$$

Dzieląc pierwszą sumę przez $\sum_{i=0}^{n+1}\theta'_i$, wyrażamy punkt 0 jako kombinację wypukłą punktów $u_0, \ldots, u_{k-1}, u_{k+1}, \ldots, u_{n+1}$. ■

Zadania 6.9

1. Wyrazić wielomian klasy Π_0 (czyli stałą) najlepiej przybliżający funkcję ciągłą f w przedziale domkniętym I przez wielkości

$$m(f) := \min_{x\in I} f(x), \quad M(f) := \max_{x\in I} f(x).$$

2. Rozwiązać układ równań z przykł. 6.9.1.
3. Znaleźć wielomian klasy Π_1 najlepiej przybliżający funkcję $\sqrt{x}$ w przedziale $[0,1]$.
4. Udowodnić, że wielomian klasy Π_2 najlepiej przybliżający funkcję $\cosh x$ w przedziale $[-1,1]$ jest równy $a + bx^2$, gdzie $b = \cosh 1 - 1$ i a spełnia wraz z pewnym t układ równań $2a = 1 + \cosh t - bt^2$, $\sinh t = 2bt$.
5. Niech będzie $f := a_0T_0 + a_1T_1 + \ldots + a_{n+1}T_{n+1}$, gdzie T_k jest k-tym wielomianem Czebyszewa. Udowodnić, że w klasie Π_n najlepszym przybliżeniem dla f w $[-1,1]$ jest suma $a_0T_0 + a_1T_1 + \ldots + a_nT_n$.
6. Niech f będzie funkcją zmiennych x i y ciągłą w kwadracie $[0,1]\times[0,1]$. Opisać jej najlepsze przybliżenie za pomocą funkcji zależnej tylko od x.
7. Udowodnić, że powłoka wypukła dowolnego zbioru S jest najmniejszym zbiorem wypukłym zawierającym S.

8. Udowodnić, że w przestrzeni unormowanej każdy zbiór $\{f\colon \|f-g\| \leqslant r\}$ jest wypukły.

9. Podać przykład zbioru wypukłego, którego dopełnienie jest ograniczone.

10. Udowodnić, że zbiory rozpięte na poniższych układach funkcji są podprzestrzeniami Haara przestrzeni $C[0,1]$.

(a) $\{1, x^2, x^3\}$, (b) $\{1, e^x, e^{2x}\}$, (c) $\{(x+2)^{-1}, (x+3)^{-1}, (x+4)^{-1}\}$.

11. Udowodnić, że zbiory rozpięte na poniższych układach funkcji nie są podprzestrzeniami Haara przestrzeni $C[-1,1]$.

(a) $\{1, x^2, x^3\}$, (b) $\{|x|, |x-1|\}$, (c) $\{e^x, x+1\}$.

12. Udowodnić, że jeśli $f^{(n)}(x) > 0$ w $[a,b]$, to $\operatorname{span}\{1, x, x^2, \ldots, x^{n-1}, f\}$ jest podprzestrzenią Haara w tym przedziale.

13. Czy $\operatorname{span}\{1, x, y\}$ jest przestrzenią Haara w $C(\mathbb{R}^2)$?

14. Dla danej prostej o równaniu $ax+by+c=0$ $(a^2+b^2>0)$ opisać ściśle zbiór jej punktów leżących najbliżej punktu $(0,0)$; odległość punktów ma być określona za pomocą normy l_∞.

15. Niech A będzie macierzą $n \times (n+2)$. Wykazać, że jeśli układ

$$Ax = 0, \quad x_i \geqslant 0 \quad (i \leqslant n+1), \qquad x_{n+2} = 0, \quad x \neq 0 \quad (x \in \mathbb{R}^{n+2})$$

jest niesprzeczny, to taką samą własność ma dla pewnego $k < n+2$ układ

$$Ax = 0, \quad x_i \geqslant 0 \quad (i \neq k), \quad x_k = 0, \quad x \neq 0 \quad (x \in \mathbb{R}^{n+2}).$$

6.10. Interpolacja funkcji wielu zmiennych

Konstrukcja gładkich funkcji interpolujących wielu zmiennych jest trudnym zadaniem, które od dawna przyciąga uwagę matematyków. Już dla dwóch zmiennych ma ono takie aspekty, których brakuje w przypadku jednej zmiennej. Dlatego niewiele tracimy, ograniczając się w tym podrozdziale niemal wyłącznie do przypadku dwóch zmiennych niezależnych.

Zadanie interpolacyjne

Główne zadanie, którym się zajmiemy, jest następujące: Dane są parami różne punkty, zwane *węzłami*, na płaszczyźnie xy. Oznaczamy je chwilowo $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Z i-tym punktem jest związana liczba rzeczywista c_i. Naszym celem jest znalezienie takiej funkcji F gładkiej i łatwo obliczalnej (w intuicyjnym rozumieniu tych słów), że

$$F(x_i, y_i) = c_i \quad (1 \leqslant i \leqslant n).$$

Przyjmuje się, że ta funkcja ma być określona na całej przestrzeni $\mathbb{R}^2$ lub co najmniej w pewnym dużym obszarze zawierającym węzły.

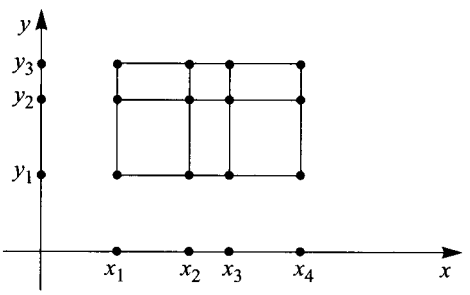
Iloczyn kartezjański i siatka

Sformułowanym już zadaniem zajmiemy się najpierw w przypadku, gdy można je sprowadzić do interpolacji funkcji jednej zmiennej. Zbiór $\mathcal{N}$ jest więc teraz *iloczynem kartezjańskim*:

$$\mathcal{N} := \{x_1, x_2, \ldots, x_p\} \times \{y_1, y_2, \ldots, y_q\}. \tag{6.10.1}$$

Inaczej mówiąc, jest to zbiór wszystkich par (x_i, y_j), gdzie x_i wybieramy z pierwszego układu, a y_j z drugiego. Te oznaczenia są w rozważanym przypadku wygodniejsze.

Zbiór $\mathcal{N}$ nazywamy *siatką kartezjańską*. Dla $p = 4$ i $q = 3$ pokazuje ją rys. 6.16. Zazwyczaj, jak tam, punkty x_i i y_j porządkujemy od najmniejszego do największego.



RYS. 6.16. Siatka kartezjańska węzłów

Przypuśćmy, że znamy pewną metodę liniową interpolacji funkcji jednej zmiennej. Ta metoda jest określona przez operator liniowy P o wartościach

$$(Pf)(x) = \sum_{i=1}^{p} f(x_i)u_i(x).$$

Funkcje u_i są takie, że

$$u_i(x_j) = \delta_{ij} \quad (1 \leqslant i, j \leqslant p). \tag{6.10.2}$$

Jak wiadomo z podrozdz. 6.1, w zwykłej interpolacji wielomianowej wyrażają się one wzorem

$$u_i(x) = \prod_{j=1,\, j\neq i}^{p} \frac{x - x_j}{x_i - x_j} \quad (1 \leqslant i \leqslant p). \tag{6.10.3}$$

Zauważmy, że operator P można w oczywisty sposób zastosować do funkcji wielu zmiennych. Istotnie, jeśli np. f zależy od zmiennych x i y, to przyjmujemy, że

$$(\bar{P}f)(x,y) = \sum_{i=1}^{p} f(x_i,y)u_i(x).$$

Wtedy $\bar{P}f$ jest funkcją dwóch zmiennych, interpolującą f na liniach pionowych

$$L_i := \{(x_i,y):\ -\infty < y < \infty\} \quad (1 \leqslant i \leqslant p).$$

Niech teraz pewien operator liniowy Q opisuje interpolację funkcji jednej zmiennej w węzłach $y_1, y_2, \ldots, y_q$:

$$(Qf)(y) = \sum_{i=1}^{q} f(y_i)v_i(y),$$

gdzie

$$v_i(y_j) = \delta_{ij} \quad (1 \leqslant i, j \leqslant q). \tag{6.10.4}$$

Także ten operator można uogólnić na przypadek dwóch zmiennych, przyjmując, że

$$(\bar{Q}f)(x,y) = \sum_{i=1}^{q} f(x,y_i)v_i(y).$$

Funkcja $\bar{Q}f$ interpoluje f na liniach poziomych

$$L^i := \{(x,y_i):\ -\infty < x < \infty\} \quad (1 \leqslant i \leqslant q).$$

Suma boolowska

Za pomocą $\bar{P}$ i $\bar{Q}$ tworzymy dwa operatory stosowane w interpolacji funkcji dwóch zmiennych. Są to *iloczyn* $\overline{PQ}$ i *suma boolowska* $\bar{P} \oplus \bar{Q}$, określona wzorem

$$\bar{P} \oplus \bar{Q} := \bar{P} + \bar{Q} - \overline{PQ}.$$

Bardziej konkretne wzory wynikają z definicji operatorów $\bar{P}$ i $\bar{Q}$. Po pierwsze,

$$(\overline{PQ}f)(x,y) := \bar{P}(\bar{Q}f)(x,y) = \sum_{i=1}^{p}(\bar{Q}f)(x_i,y)u_i(x) =$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q} f(x_i,y_j)v_j(y)u_i(x).$$

Ponieważ $v_j(y_k)u_i(x_l) = \delta_{jk}\delta_{il}$, więc – jak łatwo zauważyć – funkcja $\overline{PQ}f$ interpoluje f na całej siatce $\mathcal{N}$. Zamiast $\overline{PQ}$ stosuje się też oznaczenie $P \otimes Q$ przyjęte dla iloczynu tensorowego. Po drugie,

$$[(\bar{P} \oplus \bar{Q})f](x,y) := (\bar{P}f)(x,y) + (\bar{Q}f)(x,y) - (\overline{PQ}f)(x,y) = \\ = \sum_{i=1}^{p} f(x_i,y)u_i(x) + \sum_{j=1}^{q} f(x,y_j)v_j(y) - \sum_{i=1}^{p}\sum_{j=1}^{q} f(x_i,y_j)v_j(y)u_i(x). \tag{6.10.5}$$

Tematem zad. 2 jest sprawdzenie, że funkcja $(\bar{P} \oplus \bar{Q})f$ interpoluje f na wszystkich liniach pionowych L_i $(1 \leqslant i \leqslant p)$ i poziomych L^j $(1 \leqslant j \leqslant q)$. Ta funkcja może być użyteczna wtedy, gdy znamy f nie tylko w węzłach, ale i na tych liniach, a chcemy ją przybliżać poza nimi.

Przykład 6.10.1. Znaleźć wielomian dwóch zmiennych, który na pewnej siatce kartezjańskiej interpoluje funkcję o wartościach zawartych w tablicy

y \ x	1	2	4	5
1	1.7	−4.1	−3.2	4.9
3	6.1	−4.2	2.3	7.5
4	−5.9	3.8	−1.7	2.5

Rozwiązanie. Funkcje u_i i v_i wyrażamy zgodnie z (6.10.3):

$$u_1(x) = -\tfrac{1}{12}(x-2)(x-4)(x-5),$$
$$u_2(x) = \tfrac{1}{6}(x-1)(x-4)(x-5),$$
$$u_3(x) = -\tfrac{1}{6}(x-1)(x-2)(x-5),$$
$$u_4(x) = \tfrac{1}{12}(x-1)(x-2)(x-4),$$
$$v_1(y) = \tfrac{1}{6}(y-3)(y-4),$$
$$v_2(y) = -\tfrac{1}{2}(y-1)(y-4),$$
$$v_3(y) = \tfrac{1}{3}(y-1)(y-3).$$

Szukany wielomian jest następujący:

$$F(x,y) = u_1(x)[1.7v_1(y) + 6.1v_2(y) - 5.9v_3(y)] + \\ + u_2(x)[-4.1v_1(y) - 4.2v_2(y) + 3.8v_3(y)] + \\ + u_3(x)[-3.2v_1(y) + 2.3v_2(y) - 1.7v_3(y)] + \\ + u_4(x)[4.9v_1(y) + 7.5v_2(y) + 2.5v_3(y)].$$

■

Iloczyn tensorowy

Funkcja F z przykł. 6.10.1 jest kombinacją liniową następujących dwunastu składników $x^i y^j$:

$$1,\ x,\ x^2,\ x^3,\ y,\ xy,\ x^2y,\ x^3y,\ y^2,\ xy^2,\ x^2y^2,\ x^3y^2. \tag{6.10.6}$$

Tak więc mamy tu interpolację za pomocą 12-wymiarowej podprzestrzeni wielomianów dwóch zmiennych. Właściwym dla niej symbolem jest $\Pi_3 \otimes \Pi_2$. Jest to *iloczyn tensorowy* dwóch przestrzeni, złożony ze wszystkich funkcji postaci

$$(x,y) \mapsto \sum_{i=1}^{m} a_i(x) b_i(y),$$

gdzie $a_i \in \Pi_3$ i $b_i \in \Pi_2$. Bazą tej przestrzeni jest oczywiście układ (6.10.6).

Trzeba podkreślić, że naszkicowana już teoria stosuje się do dowolnych funkcji u_i, v_i, nie tylko do wielomianów. Ważne są tylko własności (6.10.2) i (6.10.4). (W abstrakcyjnej teorii występują bezpośrednio operatory P i Q; ich konkretna struktura nie jest istotna).

W interpolacji wielomianowej stosujemy iloczyn tensorowy $\Pi_{p-1} \otimes \Pi_{q-1}$, gdzie p i q występują w definicji siatki kartezjańskiej $\mathcal{N}$. Bazą tej przestrzeni jest układ funkcji

$$(x,y) \mapsto x^i y^j \quad (0 \leqslant i \leqslant p-1,\ 0 \leqslant j \leqslant q-1),$$

a każdy jej element ma postać

$$(x,y) \mapsto \sum_{i=0}^{p-1} \sum_{j=0}^{q-1} c_{ij} x^i y^j.$$

Stopniem składnika $x^i y^j$ jest $i+j$. Baza przestrzeni $\Pi_{p-1} \otimes \Pi_{q-1}$ zawiera tylko jeden element stopnia $p+q-2$, mianowicie $x^{p-1}y^{q-1}$. *Stopniem wielomianu* zmiennych x, y nazywamy maksymalny stopień jego składników. Przestrzeń wszystkich wielomianów dwóch zmiennych, stopnia niewiększego od k, będzie oznaczana symbolem $\Pi_k(\mathbb{R}^2)$. Każdy jej element ma postać

$$(x,y) \mapsto \sum_{i=0}^{k} \sum_{j=0}^{k-i} c_{ij} x^i y^j = \sum_{0 \leqslant i+j \leqslant k} c_{ij} x^i y^j.$$

Twierdzenie 6.10.2. *Bazą przestrzeni* $\Pi_k(\mathbb{R}^2)$ *jest układ funkcji*

$$(x,y) \mapsto x^i y^j \quad (0 \leqslant i+j \leqslant k).$$

Dowód. Ponieważ oczywiście na tym układzie można rozpiąć przestrzeń $\Pi_k(\mathbb{R}^2)$, więc wystarczy udowodnić jego liniową niezależność. Przypuśćmy zatem, że pewna funkcja z tej przestrzeni jest równa tożsamościowo 0. Jeśli przyjmiemy, że $y = y_0$, to otrzymamy równość

$$\sum_{i=0}^{k}\Big(\sum_{j=0}^{k-i} c_{ij}y_0^j\Big)x^i \equiv 0,$$

która wobec niezależności liniowej funkcji $x \mapsto x^i$ pociąga za sobą układ równań

$$\sum_{j=0}^{k-1} c_{ij}y_0^j = 0 \quad (0 \leqslant i \leqslant k).$$

y_0 jest tu dowolne. Dlatego, rozumując jak wyżej, stwierdzamy, że wszystkie c_{ij} są równe 0. ■

Wniosek 6.10.3. *Wymiar przestrzeni* $\Pi_k(\mathbb{R}^2)$ *jest równy* $\frac{1}{2}(k+1)(k+2)$.

Dowód można pominąć – wymiar przestrzeni jest równy liczbie elementów w jej bazie.

Przypomnijmy, że w interpolacji funkcji jednej zmiennej za pomocą wielomianów z Π_k można użyć dowolnego układu $k+1$ węzłów. Można by zatem spodziewać się, że w przypadku dwóch zmiennych i klasy wielomianów określonej wyżej każdy układ $n := \frac{1}{2}(k+1)(k+2)$ węzłów jest sensowny. Tak jednak nie jest, o czym można się przekonać już dla $k = 1$, gdy $n = 3$. Istotnie, każdy wielomian klasy $\Pi_1(\mathbb{R}^2)$ ma postać $c_0 + c_1x + c_2y$. Próba rozwiązania zadania interpolacji z trzema węzłami (x_i, y_i) prowadzi do układu równań liniowych, którego macierz ma wyznacznik

$$\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}.$$

Z geometrii analitycznej wiadomo, że ten wyznacznik jest równy z dokładnością do znaku podwojonemu polu trójkąta o wierzchołkach (x_i, y_i). Jeśli zatem te punkty leżą na jednej prostej, to wyznacznik jest równy zeru i zadanie interpolacji na ogół nie ma rozwiązania.

Geometria węzłów

Powyższy przykład pokazuje, że o możliwości rozwiązania zadania interpolacyjnego w klasie $\Pi_k(\mathbb{R}^2)$ decyduje rozmieszczenie węzłów na płaszczyźnie.

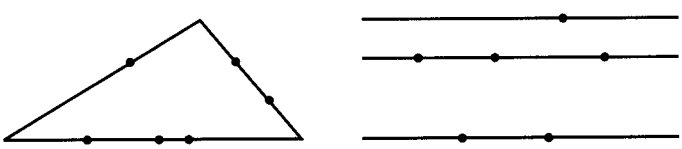
Oczywiście ich liczba ma być równa $\frac{1}{2}(k+1)(k+2)$. Niżej podano pewne twierdzenia ilustrujące to co wiadomo o tym rozmieszczeniu. Aby uniknąć nieporozumień, podkreślamy, że teraz zbiór $\mathcal{N}$ węzłów jest dowolny, czyli definicja (6.10.1) nie obowiązuje.

Pierwsze twierdzenie udowodnili Gasca i Maeztu [1982]:

Twierdzenie 6.10.4 (Gasca i Maeztu). *Jeśli dla układu $\frac{1}{2}(k+1)(k+2)$ węzłów istnieją proste $L_0, L_1, \ldots, L_k$ takie, że na L_i leży dokładnie $i+1$ węzłów, to dowolne wartości w węzłach można interpolować wielomianami z $\Pi_k(\mathbb{R}^2)$.*

Dowód. Założenie można napisać w postaci $\#(\mathcal{N} \cap L_i) = i+1$, gdzie $\#$ oznacza liczbę elementów zbioru. Zbiory $\mathcal{N} \cap L_i$ są parami rozłączne. Ponieważ liczba węzłów jest równa wymiarowi przestrzeni $\Pi_k(\mathbb{R}^2)$, więc wystarczy udowodnić, że jednorodne zadanie interpolacyjne ma tylko zerowe rozwiązanie. Niech zatem będzie $p \in \Pi_k(\mathbb{R}^2)$ i $p(x,y) = 0$ w każdym węźle. Niech $l_i(x,y) = 0$ będzie równaniem prostej L_i. Wielomian $p^2 + l_k^2$ ma co najmniej $k+1$ zer, którymi są punkty zbioru $\mathcal{N} \cap L_k$. Można tu zastosować *twierdzenie Bézouta*: jeśli $p \in \Pi_k(\mathbb{R}^2)$, $q \in \Pi_m(\mathbb{R}^2)$ i jeśli $p^2 + q^2$ ma więcej niż km zer, to p i q mają wspólny czynnik stopnia co najmniej pierwszego. Wynika z niego, że l_k jest dzielnikiem wielomianu p. Rozumowanie można iterować. Istotnie, wielomian $(p/l_k)^2 + l_{k-1}^2$ ma co najmniej k zer, czyli l_{k-1} jest dzielnikiem wielomianu p/l_k. Po k takich krokach dochodzimy do wniosku, że wielomian p dzieli się przez $l_1 l_2 \ldots l_k$. Iloraz jest stałą, gdyż p ma stopień k. Ponieważ p znika na zbiorze $\mathcal{N} \cap L_0$, a iloczyn $l_1 l_2 \ldots l_k$ tam nie znika, więc musi być $p \equiv 0$. ■

Dowód algorytmiczny tw. 6.10.4 nieodwołujący się do tw. Bézouta będzie podany później.



RYS. 6.17. Układy węzłów dopuszczalne w interpolacji za pomocą $\Pi_2(\mathbb{R}^2)$

Z twierdzenia 6.10.4 wynika, że oba układy węzłów z rys. 6.17 są sensowne w interpolacji za pomocą wielomianów z $\Pi_2(\mathbb{R}^2)$.

Poniższe twierdzenie Chunga i Yao [1977] dotyczy interpolacji funkcji d zmiennych, gdzie $d \geqslant 2$. Korzystamy tu z zad. 8, w którym dowodzi się,

że wymiar przestrzeni $\Pi_k(\mathbb{R}^d)$ jest równy

$$n(k) := \binom{d+k}{k}.$$

Twierdzenie 6.10.5 (Chung i Yao). *Jeśli węzły $z_1, z_2, \ldots, z_{n(k)}$ leżą w $\mathbb{R}^d$ na takich hiperpłaszczyznach H_{ij} $(1 \leqslant i \leqslant n(k),\ 1 \leqslant j \leqslant k)$, że*

$$z_j \in \bigcup_{\nu=1}^{k} H_{i\nu} \iff j \neq i \quad (1 \leqslant i, j \leqslant n(k)), \tag{6.10.7}$$

to dowolne wartości w węzłach można interpolować wielomianami z $\Pi_k(\mathbb{R}^d)$.

Dowód. Niech $l_{ij}(z) = 0$ $(z \in \mathbb{R}^d)$ będzie równaniem hiperpłaszczyzny H_{ij} i niech

$$q_i(z) := \prod_{j=1}^{k} l_{ij}(z) \quad (1 \leqslant i \leqslant n(k)).$$

Na mocy (6.10.7) z_i nie należy do żadnej z hiperpłaszczyzn $H_{i1}, H_{i2}, \ldots, H_{ik}$, wobec czego $l_{ij}(x_i) \neq 0$ dla $1 \leqslant j \leqslant k$. Stąd $q_i(z_i) \neq 0$. Na mocy tegoż założenia, jeśli $j \neq i$, to dla pewnego ν jest $z_j \in H_{i\nu}$, a zatem $l_{i\nu}(z_j)$ i $q_i(z_j)$ znikają. W ten sposób wykazaliśmy, że wielomian $p_i(z) := q_i(z)/q_i(z_i)$ jest taki, że $p_i(z_j) = \delta_{ij}$. Ponieważ $p_i \in \Pi_k(\mathbb{R}^{n(k)})$, więc daje to wzór interpolacyjny analogiczny do wzoru Lagrange'a:

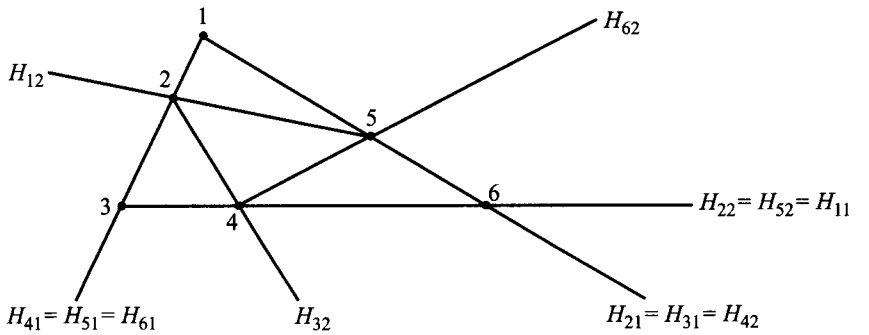
$$P(z) = \sum_{i=1}^{n(k)} f(z_i) p_i(z). \qquad \blacksquare$$

Rysunek 6.18 pokazuje układ węzłów spełniający założenia tw. 6.10.5 dla $d = k = 2$; jest $n(2) = 6$.

Ostatnie twierdzenie świadczy o tym, że pewną liczbę węzłów można wybrać dowolnie.

Twierdzenie 6.10.6. *Dla dowolnego układu $k+1$ różnych węzłów w $\mathbb{R}^2$ i dowolnych wartości w węzłach istnieje w $\Pi_k(\mathbb{R}^2)$ rozwiązanie zadania interpolacyjnego.*

Dowód. Niech węzłami będą punkty (x_i, y_i) dla $0 \leqslant i \leqslant k$. Wybieramy funkcję liniową $l(x, y) := ax + by + c$ taką, że wartości $t_i := l(x_i, y_i)$ są parami różne (zob. zad. 10). Dla funkcji f, którą chcemy interpolować,



RYS. 6.18. Ilustracja twierdzenia 6.10.5

istnieje wielomian jednej zmiennej $p \in \Pi_k$ taki, że $p(t_i) = f(x_i, y_i)$. Wtedy $p \circ l \in \Pi_k(\mathbb{R}^2)$ i

$$(p \circ l)(x_i, y_i) = p(l(x_i, y_i)) = p(t_i) = f(x_i, y_i). \qquad \blacksquare$$

Twierdzenie 6.10.6 można w oczywisty sposób uogólnić na przestrzeń $\mathbb{R}^d$.

Wzór Newtona

W praktycznych zastosowaniach metod interpolacyjnych wygodnie jest mieć algorytm podobny do tego, który daje wzór Newtona w przypadku jednej zmiennej. Przypomnijmy istotną zaletę tego wzoru. Znając wielomian p interpolujący funkcję f w węzłach $x_1, x_2, \ldots, x_n$, możemy łatwo otrzymać wielomian p^* interpolujący ją także w dodatkowym węźle x_{n+1} – wystarczy do p dodać jeden składnik:

$$p^*(x) = p(x) + cq(x),$$

gdzie

$$q(x) := (x - x_1)(x - x_2)\ldots(x - x_n), \quad c := \frac{f(x_{n+1}) - p(x_{n+1})}{q(x_{n+1})}.$$

Dzięki temu wielomian interpolacyjny można tworzyć iteracyjnie, dołączając w każdym kroku jeden węzeł i dodając jeden składnik do p.

To postępowanie można uogólnić. Niech będą dane funkcja f o wartościach rzeczywistych, określona na zbiorze X i układ węzłów $\mathcal{N}$. Jeśli funkcja p interpoluje f na $\mathcal{N}$, a funkcja q znika na tym zbiorze, to funkcję p^* interpolującą f na zbiorze $\mathcal{N} \cup \{\xi\}$ można wyrazić w postaci $p^* = p + cq$, jeśli tylko $q(\xi) \neq 0$. W jeszcze ogólniejszym wariancie operujemy na zbiorach węzłów. Niech q będzie funkcją o wartościach rzeczywistych, określoną

na X, a Z zbiorem jej zer. Jeśli p interpoluje f na $\mathcal{N} \cap Z$, a r interpoluje $(f-p)/q$ na $\mathcal{N} \setminus Z$, to $p+qr$ interpoluje f na $\mathcal{N}$.

Naszkicowana procedura daje algorytmiczny dowód tw. 6.10.4. Zaczynamy od wyboru wielomianu $p_k \in \Pi_k(\mathbb{R}^2)$ interpolującego f na $\mathcal{N} \cap L_k$ (korzystamy tu z tw. 6.10.6). Stosując indukcję wstecz, zakładamy, że jest już znany wielomian $p_i \in \Pi(\mathbb{R}^2)$ interpolujący f na zbiorze $L_k \cup L_{k-1} \cup \ldots \cup L_1$. Wzorując się na wzorze Newtona, próbujemy wyznaczyć p_{i-1} w postaci

$$p_{i-1} = p_i + r l_k l_{k-1} \ldots l_i.$$

Jest oczywiste, że takie p_{i-1} interpoluje f na tymże zbiorze, gdyż składnik dodany do p_i tam znika. Chcemy, żeby wielomian p_i interpolował f także na L_{i-1}. Niech będzie

$$f(x) = p_i(x) + r(x)(l_k l_{k-1} \ldots l_i)(x) \quad \text{dla} \quad x \in \mathcal{N} \cap L_{i-1}.$$

Wnioskujemy stąd, że funkcja r powinna interpolować $(f-p_i)/(l_k l_{k-1} \ldots l_i)$ na $\mathcal{N} \cap L_{i-1}$. Na mocy tw. 6.10.6 taka funkcja r istnieje w klasie $\Pi_{i-1}(\mathbb{R}^2)$. Zauważmy też, że $p_{i-1} \in \Pi_k(\mathbb{R}^2)$, gdyż r jest stopnia $i-1$, a $(l_k l_{k-1} \ldots l_i)$ jest stopnia $k-i+1$. Taki algorytm podał Micchelli [1986a].

Jest interesujące, że żadna podprzestrzeń n-wymiarowa w $C(\mathbb{R}^2)$ nie pozwala na interpolację na dowolnym zbiorze n węzłów (wyjątkiem jest tylko banalny przypadek $n=1$). Jako pierwszy zauważył to chyba Haar w 1918 r. Jego rozumowanie było z grubsza takie: Niech będą dane funkcje $u_1, u_2, \ldots, u_n \in C(\mathbb{R}^2)$ i węzły $p_i = (x_i, y_i)$ $(1 \leqslant i \leqslant n)$. Chcąc interpolować jakąś funkcję w tych węzłach, używając funkcji bazowych u_i, musimy rozwiązać układ równań liniowych, którego macierz ma wyznacznik

$$D = \begin{vmatrix} u_1(p_1) & u_2(p_1) & \ldots & u_n(p_1) \\ u_1(p_2) & u_2(p_2) & \ldots & u_n(p_2) \\ \ldots & \ldots & \ldots & \ldots \\ u_1(p_n) & u_2(p_n) & \ldots & u_n(p_n) \end{vmatrix}.$$

Przypuśćmy, że jest on różny od 0 dla pewnego układu węzłów. Przesuwajmy węzły p_1 i p_2 w sposób ciągły tak, aby cały czas były różne od siebie i od pozostałych węzłów i aby ostatecznie zamieniły się miejscami. Wtedy końcowy wyznacznik D ma tylko przestawione dwa początkowe wiersze, a więc różni się tylko znakiem od wartości początkowej. Wobec ciągłej zależności wyznacznika od węzłów będzie on znikał w pewnym momencie ruchu. Tak więc D bywa równy 0 nawet dla różnych węzłów. Zauważmy, że to rozumowanie jest poprawne dla $\mathbb{R}^d$, gdy $d > 1$, ale nie dla $d = 1$. To właśnie tłumaczy, dlaczego do interpolacji funkcji wielu zmiennych trzeba podchodzić inaczej. Zazwyczaj, mając ustalony układ węzłów, szukamy sensownej podprzestrzeni funkcji interpolujących.

Interpolacja Sheparda

Bardzo ogólną metodę interpolacji wspomnianego wyżej typu (z podprzestrzenią dobieraną do węzłów) zaproponował Shepard [1968]. W *interpolacji Sheparda* jest dany układ n różnych parami węzłów

$$p_i := (x_i, y_i) \quad (1 \leqslant i \leqslant n).$$

Wybieramy funkcję φ o wartościach rzeczywistych, określoną na $\mathbb{R}^2 \times \mathbb{R}^2$ taką, że dla $p, q \in \mathbb{R}^2$

$$\varphi(p, q) = 0 \quad \text{wtedy i tylko wtedy, gdy} \quad p = q. \tag{6.10.8}$$

Taką własność mają np. funkcje $\|p - q\|$ i $\|p - q\|^2$. Wzorując się na wzorze interpolacyjnym Lagrange'a, definiujemy funkcje

$$u_i(p) := \prod_{j=1,\, j\neq i}^{n} \frac{\varphi(p, p_j)}{\varphi(p_i, p_j)} \quad (1 \leqslant i \leqslant n).$$

Wobec (6.10.8) są one poprawnie określone i takie, że

$$u_i(p_j) = \delta_{ij} \quad (1 \leqslant i, j \leqslant n).$$

Jest więc oczywiste, że funkcja

$$F := \sum_{i=1}^{n} f(p_i) u_i$$

interpoluje f w danych węzłach.

Przykład 6.10.7. Niech będzie $\varphi(p, q) := \|p - q\|^2$. Jaki jest wtedy wzór w interpolacji Sheparda?

Rozwiązanie. Jeśli $p_i = (x_i, y_i)$, $p = (x, y)$, to

$$F(x, y) = \sum_{i=1}^{n} f(x_i, y_i) \prod_{j=1,\, j\neq i}^{n} \frac{(x - x_j)^2 + (y - y_j)^2}{(x_i - x_j)^2 + (y_i - y_j)^2}. \qquad \blacksquare$$

W innej wersji metody Sheparda przyjmuje się dodatkowe założenie, że wartości funkcji φ są nieujemne. Określamy też funkcje

$$v_i(p) := \prod_{j=1,\, j\neq i}^{n} \varphi(p, p_j), \quad v(p) := \sum_{i=1}^{n} v_i(p), \quad w_i(p) := \frac{v_i(p)}{v(p)}.$$

Z założenia o φ wynika, że $v_i(p_j) = 0$ dla $i \neq j$ i $v_i(p) > 0$, jeśli tylko $p \neq p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_n$. Dlatego $v(p) > 0$ i funkcje w_i są poprawnie określone. Prócz tego

$$w_i(p_j) = \delta_{ij}, \quad 0 \leqslant w_i(p) \leqslant 1 \tag{6.10.9}$$

i $\sum_{i=1}^{n} w_i(p) = 1$. Wzór interpolacyjny ma tu postać

$$F = \sum_{i=1}^{n} f(p_i) w_i = \sum_{i=1}^{n} f(p_i) v_i / v.$$

Ta wersja interpolacji Sheparda ma dwie dodatkowe zalety: jeśli wartości $f(p_i)$ są nieujemne, to F jest wszędzie nieujemna, a jeśli f jest stała, to $F = f$. Tak więc funkcja interpolująca F zachowuje pewne własności funkcji interpolowanej. Z drugiej strony, jeśli φ jest różniczkowalna, to w każdym węźle funkcja F jest „płaska". Istotnie, wobec (6.10.9) każda funkcja w_i ma ekstrema w węzłach, czyli znikają tam jej pierwsze pochodne cząstkowe i tę samą własność ma F.

Ważny przypadek szczególny interpolacji Sheparda otrzymujemy, przyjmując, że dla $x, y \in \mathbb{R}^d$

$$\varphi := \|x - y\|^{\mu} \quad (\mu > 0).$$

Łatwo sprawdzić, że ta funkcja jest różniczkowalna dla $\mu > 1$, ale nie dla $0 < \mu \leqslant 1$. Wzór dla w_i można wyrazić na dwa sposoby:

$$w_i(x) := \prod_{j=1,\, j\neq i}^{n} \|x - x_j\|^{\mu} \Big/ \sum_{k=1}^{n} \prod_{j=1,\, j\neq k}^{n} \|x - x_j\|^{\mu},$$

$$w_i(x) := \|x - x_i\|^{-\mu} \Big/ \sum_{j=1}^{n} \|x - x_j\|^{-\mu}.$$

Drugi wzór jest prostszy, ale trzeba go stosować ostrożnie, bo w x_i mamy iloraz ∞/∞.

Metoda Franke'a i Little'a interpolacji funkcji wielu zmiennych jest lokalna w tym sensie, że wartość dana w węźle nieznacznie wpływa na funkcję interpolującą daleko od tego punktu. Dla danych węzłów (x_i, y_i) $(1 \leqslant i \leqslant n)$ wprowadzamy funkcje

$$g_i(x, y) := \left[1 - r_i^{-1} \sqrt{(x - x_i)^2 + (y - y_i)^2}\right]_{+}^{\mu},$$

gdzie dolny wskaźnik $+$ sygnalizuje, że jeśli wyrażenie w nawiasach kwadratowych jest ujemne, to zastępujemy je zerem. Tak jest, gdy punkt (x, y)

leży daleko od (x_i, y_i). Parametr μ wpływa na gładkość funkcji, a r_i steruje nośnikiem funkcji g_i; jest $g_i(x, y) = 0$, jeśli odległość (x, y) od (x_i, y_i) przekracza r_i.

Wygodnie jest przyjąć, że r_i jest odległością punktu (x_i, y_i) od najbliższego węzła, bo wtedy $g_i(x_j, y_j) = \delta_{ij}$ i funkcja interpolująca ma standardową postać

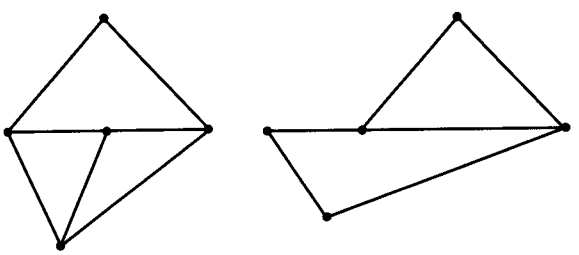
$$\sum_{i=1}^{n} f(x_i, y_i) g_i(x, y).$$

Triangulacja

Inna ogólna strategia wyboru funkcji interpolujących w $\mathbb{R}^2$ zaczyna się od *triangulacji* płaszczyzny. Jest to podział jej części, otrzymany dzięki połączeniu węzłów odcinkami tak, aby dało to rodzinę trójkątów $T_1, T_2, \ldots, T_m$ o następujących własnościach:

1. Każdy węzeł jest wierzchołkiem jakiegoś trójkąta T_i.
2. Każdy wierzchołek każdego trójkąta jest węzłem.
3. Jeśli węzeł należy do pewnego trójkąta, to jest jego wierzchołkiem.

Reguła **3** wyklucza przypadki pokazane na rys. 6.19.

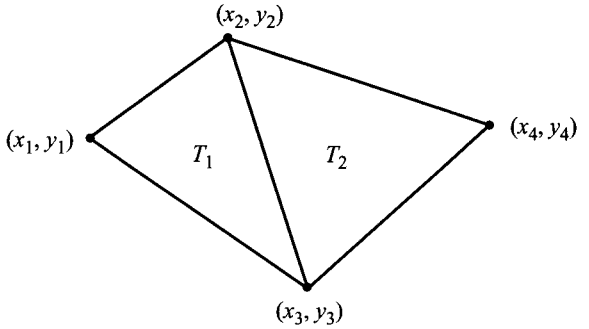


RYS. 6.19. Błędne triangulacje

Najprostszym typem funkcji interpolującej na triangulacji jest funkcja liniowa w każdym trójkącie T_i:

$$l_i(x, y) := a_i x + b_i y + c_i \quad \text{dla} \quad (x, y) \in T_i.$$

Jej współczynniki są jednoznacznie określone przez dane wartości funkcji interpolowanej w wierzchołkach trójkąta T_i. Wynika to z tw. 6.10.4, w którym jako L_1 wybieramy jeden bok trójkąta, a jako L_0 – prostą równoległą do L_1 przechodzącą przez jego trzeci wierzchołek.



RYS. 6.20. Triangulacja

Rozważmy sytuację z rys. 6.20. Odcinek o końcach (x_2, y_2) i (x_3, y_3) jest wspólny dla dwóch trójkątów. Można go opisać tak:

$$\{t(x_2, y_2) + (1-t)(x_3, y_3) \colon 0 \leqslant t \leqslant 1\}.$$

Funkcja l_1 na odcinku jest liniowa względem t:

$$\begin{aligned} &a_1(tx_2 + (1-t)x_3) + b_1(ty_2 + (1-t)y_3) + c_1 = \\ &\quad = [a_1(x_2 - x_3) + b_1(y_2 - y_3)]t + a_1x_3 + b_1y_3 + c_1. \end{aligned}$$

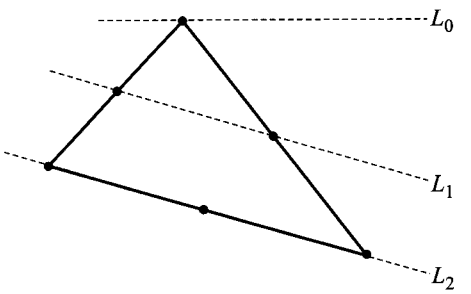
Warunki interpolacyjne w punktach (x_2, y_2) i (x_3, y_3) określają ją jednoznacznie. To samo dotyczy funkcji l_2. Wobec tego l_1 i l_2 na tym odcinku są identyczne, czyli funkcja trójkątami liniowa jest w $T_1 \cup T_2$ ciągła. Wynika stąd następujące twierdzenie:

Twierdzenie 6.10.8. *Jeśli $\{T_1, T_2, \ldots, T_m\}$ jest triangulacją na płaszczyźnie, to funkcja trójkątami liniowa o danych wartościach we wszystkich wierzchołkach trójkątów T_i jest ciągła.*

Rozważmy teraz funkcję, która w trójkącie T_i należy do klasy $\Pi_2(\mathbb{R}^2)$:

$$q_i(x, y) := a_1x^2 + a_2xy + a_3y^2 + a_4x + a_5y + a_6.$$

Potrzebujemy sześciu warunków, żeby wyznaczyć sześć współczynników takiej funkcji. Możemy np. ustalić wartości funkcji w wierzchołkach trójkąta i środkach jego boków. I w tym przypadku tw. 6.10.4 zapewnia istnienie funkcji interpolującej. Jak na rys. 6.21, możemy określić L_2 jako jeden z boków trójkąta, L_1 jako prostą przechodzącą przez środki pozostałych boków, a L_0 jako prostą przechodzącą przez ich wspólny wierzchołek, np. równoległą do L_2. Rozumując jak poprzednio, dowodzimy, że lokalne określenia funkcji q_i dają funkcję ciągłą w zbiorze $T_1 \cup T_2 \cup \ldots \cup T_m$.



RYS. 6.21. Zastosowanie twierdzenia 6.10.4 w interpolacji kwadratowej

Metoda ruchomych najmniejszych kwadratów

Inną uniwersalną metodą interpolacji za pomocą funkcji gładkich jest metoda *ruchomych najmniejszych kwadratów*. Wyjaśnijmy najpierw ogólnie, na czym ona polega.

Dany jest obszar X, który nas interesuje; może to być np. przestrzeń $\mathbb{R}$, $\mathbb{R}^2$ lub podzbiór którejś z nich. Dany jest też układ $\{x_1, x_2, \ldots, x_n\}$ węzłów, w których znamy wartości rzeczywiste $f(x_i)$ pewnej funkcji f. Wybieramy też funkcje $u_1, u_2, \ldots, u_m$ o wartościach rzeczywistych, określone na X. Ich liczba m jest zwykle mała w porównaniu z n.

W tradycyjnej metodzie najmniejszych kwadratów ustalamy jeszcze wagi $w_i > 0$. Próbujemy znaleźć współczynniki $c_1, c_2, \ldots, c_m$, dla których suma ważona kwadratów reszt (odchyleń)

$$\sum_{i=1}^{n} w_i \Big[f(x_i) - \sum_{j=1}^{m} c_j u_j(x_i) \Big]^2$$

jest najmniejsza. Definiujemy iloczyn skalarny i normę wzorami

$$\langle f, g \rangle := \sum_{i=1}^{n} w_i f(x_i) g(x_i), \quad \|f\| := \sqrt{\langle f, f \rangle},$$

co pozwala stosować teorię aproksymacji w przestrzeni unitarnej (podrozdz. 6.8). Wiemy więc, że rozwiązanie postawionego zadania spełnia warunek ortogonalności

$$f - \sum_{j=1}^{m} c_j u_j \perp u_i \quad (1 \leqslant i \leqslant m),$$

który daje układ równań normalnych

$$\sum_{j=1}^{m} c_j \langle u_j, u_i \rangle = \langle f, u_i \rangle \quad (1 \leqslant i \leqslant m).$$

Metoda ruchomych najmniejszych kwadratów od opisanej wyżej różni się tym, że wagi w_i mogą być dodatkowo funkcjami zmiennej x. Iloczyn skalarny, układ równań normalnych i funkcja aproksymująca wyrażają się teraz odpowiednio tak:

$$\langle f, g\rangle_x := \sum_{i=1}^{n} w_i(x) f(x_i) g(x_i),$$

$$\sum_{j=1}^{m} c_j(x)\langle u_j, u_i\rangle_x = \langle f, u_i\rangle_x \quad (1 \leqslant i \leqslant m),$$

$$g(x) := \sum_{j=1}^{m} c_j(x) u_j(x).$$

Obliczenia potrzebne do zbudowania tej funkcji są kosztowne dla dużych m, gdyż układy równań normalnych trzeba rozwiązywać oddzielnie dla każdego x. Z tego powodu przyjmuje się na ogół, że $m \leqslant 10$.

Funkcje wagowe pozwalają uzyskać różne pożądane wyniki. Jeśli $w_i(x)$ jest duże dla $x = x_i$, to funkcja g prawie interpoluje f w tym punkcie. W granicy, dla $w_i(x_i) = \infty$, jest $g(x_i) = f(x_i)$. Jeśli $w_i(x)$ szybko maleje do 0, gdy x oddala się od x_i, to węzły dalekie od x_i mają mały wpływ na $g(x_i)$. Te dwa cele osiągamy w przestrzeni $\mathbb{R}^d$, przyjmując, że

$$w_i(x) := \|x - x_i\|^{-2},$$

gdzie normę można wybrać dowolnie, choć zwykle decydujemy się na normę euklidesową.

Jeśli $m = 1$, $u_1(x) \equiv 1$, a funkcja wagowa jest określona jak wyżej, to metoda ruchomych najmniejszych kwadratów sprowadza się do metody Sheparda. Żeby się w tym upewnić, napiszmy dla tego przypadku równanie normalne przyjmując $c_1(x) = c(x)$, $u_1(x) = u(x) = 1$:

$$c(x)\langle u, u\rangle_x = \langle f, u\rangle_x.$$

Wtedy

$$g(x) = c(x)u(x) = c(x) = \frac{\langle f, u\rangle_x}{\langle u, u\rangle_x} = \frac{\sum_{i=1}^{n} f(x_i) w_i(x)}{\sum_{j=1}^{n} w_j(x)}.$$

Jeśli $w_i(x) = \|x - x_i\|^{-2}$, to usunięcie osobliwości daje wyrażenie $w_i / \sum_{j=1}^{n} w_j$ równe 1 w x_i i 0 w pozostałych węzłach.

Interpolacja za pomocą multikwadryk

Jeszcze inna metoda interpolacji funkcji wielu zmiennych, którą zaproponował Hardy [1971], opiera się na funkcjach bazowych zwanych *multikwadrykami*:

$$z_i(p) := (\|p - p_i\|^2 + c^2)^{1/2} \quad (1 \leqslant i \leqslant n).$$

Występuje tu norma euklidesowa. Hardy sugeruje, żeby parametr c był równy iloczynowi 0.8 przez średnią odległość między węzłami. Stosując te funkcje w interpolacji, musimy wiedzieć, że macierz o elementach $z_i(p_j)$ jest nieosobliwa. Udowodnił to Micchelli [1986b].

Oto inne prace dotyczące interpolacji funkcji wielu zmiennych: Chui [1988], Hartley [1976], Micchelli [1986a], Franke [1982] oraz Lancaster i Salkauskas [1986].

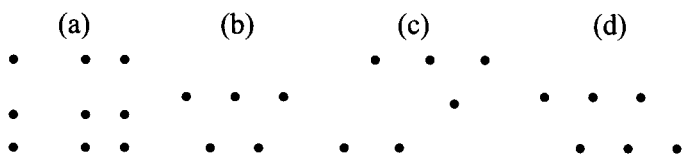
Informacje o metodzie Sheparda podają: Shepard [1968], Gordon i Wixom [1978], Newman i Rivlin [1983], Barnhill, Dube i Little [1983], Farwig [1986] oraz Cheney i Light [1999].

Zadania 6.10

1. Podać algorytm, który sprawdza, czy dany zbiór węzłów jest iloczynem kartezjańskim (6.10.1) i daje, gdy odpowiedź jest twierdząca, czynniki tego iloczynu.
2. Udowodnić własność sumy boolowskiej $\bar{P} \oplus \bar{Q}$ z (6.10.5) podaną po tym wzorze.
3. Sprawdzić poprawność poniższych relacji:
   (a) $\Pi_n \otimes \Pi_m \subset \Pi_{n+m}(\mathbb{R}^2)$.
   (b) $\Pi_k(\mathbb{R}^2) \subset \Pi_n \otimes \Pi_m$, gdzie $k := \max\{n, m\}$.
   (c) $\Pi_k(\mathbb{R}^2) \subset \Pi_n \otimes \Pi_m$, gdzie $k := \min\{n, m\}$.
4. Niech U i V będą przestrzeniami wektorowymi funkcji określonych odpowiednio na X i Y. Wtedy $U \otimes V$ składa się ze wszystkich funkcji postaci

   $$\sum_i u_i(x)v_i(y) \quad (u_i \in U,\ v_i \in V).$$

   Udowodnić, że $\dim(U \otimes V) = \dim U \cdot \dim V$, jeśli tylko przestrzenie U i V są skończeniewymiarowe.
5. Wykazać, że interpolacja za pomocą $\Pi_2(\mathbb{R}^2)$ na ogół nie jest możliwa, gdy sześć węzłów leży na elipsie, paraboli, hiperboli lub na dwóch prostych.
6. Wykazać, że jeśli cztery węzły są wierzchołkami prostokąta o bokach równoległych do osi współrzędnych, to dla dowolnych danych w tych węzłach jest możliwa interpolacja za pomocą wielomianu $p(x, y) = a + bx + cy + dxy$.
7. Jaki rodzaj wielomianu dwóch zmiennych jest odpowiedni dla interpolacji, gdy układy węzłów są takie jak na rys. 6.22?



RYS. 6.22

8. Wykazać, że wymiar zbioru $\Pi_k(\mathbb{R}^d)$ jest równy $\binom{d+k}{k}$.

9. Niech będzie $f \in \Pi_k(\mathbb{R})$ i $l \in \Pi_1(\mathbb{R}^2)$. Udowodnić, że $f \circ l \in \Pi_k(\mathbb{R}^2)$. Czy każdy element klasy $\Pi_k(\mathbb{R}^2)$ można otrzymać w ten sposób?

10. Udowodnić, że dla danych n punktów x_i parami różnych w $\mathbb{R}^d$ istnieje takie $b \in \mathbb{R}^d$, że liczby $\langle x_i, b\rangle$ są parami różne (własność potrzebna w dowodzie tw. 6.10.6).

11. Rozważyć interpolację Sheparda dla dwóch zmiennych, n węzłów i funkcji $\varphi(p,q) := \|p-q\|^2$. Jaka jest przestrzeń $\Pi_k(\mathbb{R}^2)$, zawierająca wszystkie funkcje bazowe? (Podać najmniejsze k).

12. Dla $\varphi(p,q) := \|p-q\|_\infty^2$ (gdzie $\|(x,y)\|_\infty := \max\{|x|,|y|\}$) podać wzory opisujące drugą wersję interpolacji Sheparda.

13. Czy reguły określające triangulację można zwięźle wyrazić, mówiąc, że zbiór węzłów jest identyczny ze zbiorem wierzchołków trójkątów?

14. Udowodnić, że dla operatora L określonego wzorem $Lf := \sum_{i=1}^n w_i f(x_i)$ następujące własności są równoważne:
(i) Dla każdego f jest $\min f(x_i) \leqslant Lf \leqslant \max f(x_i)$.
(ii) $w_i \geqslant 0$ dla $1 \leqslant i \leqslant n$ oraz $\sum_{i=1}^n w_i = 1$.

ZADANIA KOMPUTEROWE 6.10

K1. Zaprogramować w efektywny sposób metodę Sheparda. Danymi mają być węzły (x_i, y_i), wartości c_i w węzłach i lista punktów, w których należy obliczyć wartości funkcji interpolującej.

6.11. Aproksymacja wymierna

Tematem kilku poprzednich podrozdziałów były różne rodzaje aproksymacji wielomianowej – interpolacja Lagrange'a i Hermite'a, aproksymacja średniokwadratowa i jednostajna, rozwinięcia w szeregi potęgowe. Ta aproksymacja ma wiele zalet, ale i ograniczenia wynikające z własności wielomianów. Dlatego właśnie stosuje się także funkcje sklejane, budowane zresztą w znany już sposób z wielomianów. Inny sposób przełamania wspomnianych ograniczeń polega na zastosowaniu w aproksymacji funkcji wymiernych.

Funkcja wymierna r jest z definicji ilorazem dwóch wielomianów p i q; drugi z nich nie jest równy tożsamościowo zeru. Aby uniknąć kłopotliwej

niejednoznaczności, będziemy zakładać, że te wielomiany nie mają wspólnego czynnika stopnia dodatniego. Wspólny czynnik stały może być dowolny (byle różny od 0).

Oczywiście, każdy wielomian jest funkcją wymierną, ale nie odwrotnie. Istotną cechą funkcji wymiernej $r = p/q$ jest to, że jeśli q nie jest stałą, to r ma *bieguny*. Nazywamy tak każde zero mianownika q. Tak więc funkcja $(x+1)/[(x-2)(x^2+1)]$ ma bieguny 2, i, $-i$. W biegunach ta funkcja nie ma określonej wartości, a gdy argument x dąży do bieguna, to moduł wartości funkcji rośnie nieograniczenie. Zauważmy też, że ta sama przykładowa funkcja dąży do 0, gdy $x \to \infty$; żaden wielomian (z wyjątkiem stałej 0) nie ma tej własności.

Istnieją różne metody tworzenia funkcji wymiernych, a także sposoby ich wyrażania. Zajmiemy się tu kolejno interpolacją wymierną, aproksymacją Padégo i ułamkami łańcuchowymi.

Interpolacja wymierna

Ograniczymy się do interpolacji typu Lagrange'a, tj. do następującego zadania: dla danych liczb całkowitych nieujemnych m i n, danych węzłów $x_0, x_1, \ldots, x_{m+n}$ parami różnych i danych wartości $y_0, y_1, \ldots, y_{m+n}$ znaleźć współczynniki takiej funkcji wymiernej $r = p/q$, gdzie $p \in \Pi_m$ i $q \in \Pi_n$, że

$$r(x_j) = y_j \quad (0 \leqslant j \leqslant m+n). \tag{6.11.1}$$

Liczba $m+n+1$ warunków nie jest wybrana przypadkowo. Istotnie, wielomiany p i q mają łącznie $m+n+2$ współczynniki, ale ten z nich, który jest różny od 0, możemy wybrać dowolnie.

Podobne zadanie interpolacji wielomianowej ma, jak wiemy, jedno i tylko jedno rozwiązanie. Tu jest inaczej. Zauważmy, że dla

$$l = 0, \quad m = 1, \quad x_0 = y_0 = 0, \quad x_1 = y_1 = 1 \tag{6.11.2}$$

zadanie (6.11.1) nie ma rozwiązania. Istotnie, z pierwszego warunku wynika, że licznik (stała) jest równy 0, ale wtedy na pewno $r(1) \neq 1$. Druga istotna własność interpolacji wielomianowej zachowuje się i tutaj:

Twierdzenie 6.11.1. *Jeśli funkcja r spełniająca warunki* (6.11.1) *istnieje, to jest określona jednoznacznie.*

Dowód. Przypuśćmy, że dwie funkcje wymierne, p/q i $\tilde{p}/\tilde{q}$, spełniają te warunki. Wtedy wielomian $p\tilde{q} - \tilde{p}q$ należący do Π_{m+n} ma $m+n+1$ różnych zer x_i, a więc znika tożsamościowo, czyli $p/q = \tilde{p}/\tilde{q}$. ■

Z równań (6.11.1) wynika układ liniowy jednorodny względem współczynników licznika i mianownika:

$$p(x_j) = y_j q(x_j) \quad (0 \leqslant j \leqslant m+n).$$

Ten układ jest jednak źle uwarunkowany (zob. podrozdz. 4.4). Ponadto jego rozwiązanie może być takie, że $q(x_j) = 0$ dla pewnego j. Niżej opisano *metodę Wernera*, która polega na iterowanej redukcji zadania (6.11.1) do podobnego, na ogół prostszego zadania z innymi parametrami. Metoda pozwala też wykryć nieistnienie funkcji r. Aby uniknąć nieporozumień, oznaczymy ją teraz r_{mn}; jej licznik należy do Π_m, a mianownik do Π_n.

TWIERDZENIE 6.11.2 (WERNER)

I. *Jeśli* $y_0 = y_1 = \ldots = y_{m+n}$, *to funkcja* r_{mn} *istnieje i jest równa tożsamościowo* y_0.

II. *Jeśli* $m = 0$, *istnieje takie* k, *że* $y_k = 0$ *i nie wszystkie* y_j *są równe* 0, *to funkcja* r_{mn} *nie istnieje.*

III. *Jeśli* $m > 0$, *istnieje takie* k, *że* $y_k = 0$, *nie wszystkie* y_j *są równe* 0, *a funkcja* r_{mn} *istnieje, to istnieje funkcja wymierna*

$$r_{m-1,n}(x) := \frac{r_{mn}(x)}{x - x_k},$$

spełniająca warunki interpolacyjne

$$r_{m-1,n}(x_j) = \frac{y_j}{x_j - x_k} \quad (0 \leqslant j \leqslant m+n,\ j \neq k).$$

IV. *Jeśli* $m \leqslant n$, *wszystkie* y_j *są różne od* 0, *a funkcja* r_{mn} *istnieje, to dla dowolnego* $k = 0, 1, \ldots, m+n$ *istnieje funkcja wymierna*

$$r_{m-1,n}(x) := \frac{r_{mn}(x) - y_k}{x - x_k},$$

spełniająca warunki interpolacyjne

$$r_{m-1,n}(x_j) = \frac{y_j - y_k}{x_j - x_k} \quad (0 \leqslant j \leqslant m+n,\ j \neq k).$$

V. *Jeśli* $m < n$, *wszystkie* y_j *są różne od* 0, *a funkcja* r_{mn} *istnieje, to istnieje funkcja wymierna*

$$r_{nm} := \frac{1}{r_{mn}},$$

spełniająca warunki interpolacyjne

$$r_{nm}(x_j) = \frac{1}{y_j} \quad (0 \leqslant j \leqslant m+n).$$

Twierdzenie jest niemal oczywiste. Jego część II wykazaliśmy już na przykładzie (6.11.2). Jeśli $r_{mn} = p_{mn}/q_{mn}$, to w przypadkach III i IV

$$p_{m-1,n}(x) = \frac{p_{mn}(x) - y_k q_{mn}(x)}{x - x_k}, \quad q_{m-1,n} = q_{mn}.$$

$p_{m-1,n}$ należy do Π_{m-1}, gdyż x_k jest zerem wielomianu $p_{mn}(x) - y_k q_{mn}(x)$, który wobec tego dzieli się przez $x - x_k$. Wartości nowej funkcji wymiernej są ilorazami różnicowymi wartości funkcji r_{mn}.

Zwróćmy jeszcze uwagę na ostrożne sformułowania w przypadkach III–V: *jeśli* istnieje r_{mn}, to istnieje...; nie wiemy bowiem, redukując pierwotne zadanie do coraz prostszych, czy ma ono rozwiązanie.

Stosując tw. 6.11.1, badamy kolejno, czy są spełnione założenia przypadków I, II, ..., V, dotyczące danych m, n, y_j.

Przykład 6.11.3. Niech będzie $m = 1$, $n = 2$. Znaleźć funkcję wymierną r_{12} spełniającą warunki interpolacyjne

x	0	2	3	4
y	$\frac{1}{2}$	$\frac{3}{8}$	$\frac{2}{7}$	$\frac{5}{22}$

(zapisano je tak jak dla interpolacji wielomianowej).

Rozwiązanie. Niżej dla każdego etapu obliczeń podano numer przypadku, związek nowej funkcji wymiernej z poprzednią i tablicę warunków, które ta nowa funkcja ma spełniać:

V. $r_{21}(x) := \dfrac{1}{r_{12}(x)}$

x	0	2	3	4
y	2	$\frac{8}{3}$	$\frac{7}{2}$	$\frac{22}{5}$

IV. $r_{11}(x) := \dfrac{r_{21}(x) - 2}{x}$

x	2	3	4
y	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{3}{5}$

IV. $r_{01}(x) := \dfrac{r_{11}(x) - \frac{1}{3}}{x - 2}$

x	3	4
y	$\frac{1}{6}$	$\frac{2}{15}$

V. $r_{10}(x) := \dfrac{1}{r_{01}(x)}$

x	3	4
y	6	$\frac{15}{2}$

IV. $r_{00}(x) := \dfrac{r_{10}(x) - 6}{x - 3}$

x	4
y	$\frac{3}{2}$

I. $r_{00}(x) \equiv \dfrac{3}{2}$

Aby znaleźć r_{12}, stosujemy teraz od końca relacje wiążące kolejne funkcje wymierne:

$$r_{10}(x) = 6 + (x - 3)r_{00}(x) = \frac{3}{2}x + \frac{3}{2},$$

$$r_{01}(x) = \frac{1}{r_{10}(x)} = \frac{2}{3x+3},$$
$$r_{11}(x) = \frac{1}{3} + (x-2)r_{01}(x) = \frac{x-1}{x+1},$$
$$r_{21}(x) = 2 + xr_{11}(x) = \frac{x^2+x+2}{x+1},$$
$$r_{12}(x) = \frac{1}{r_{21}(x)} = \frac{x+1}{x^2+x+2}.$$

Łatwo sprawdzić, że ta funkcja spełnia założone warunki interpolacyjne. ■

Według wzorów podanych na końcu przykładu można by – nie znajdując współczynników funkcji wymiernych – obliczyć wartość funkcji r_{12} w dowolnym punkcie. Trzeba tu jednak się zastrzec, że nawet wtedy, gdy szukana funkcja r_{mn} nie istnieje, algorytm oparty na tw. 6.11.1 i wyjaśniony na przykładzie może działać pozornie poprawnie aż do końca (czyli dojścia do r_{00}). Wtedy jednak funkcja r_{mn} spełnia nie wszystkie warunki interpolacyjne. Można to wykryć tylko obliczając współczynniki licznika i mianownika każdej z funkcji $r_{00}, \ldots, r_{mn}$ i badając je dokładniej.

Wyjaśnijmy teraz dokładniej, kiedy r_{mn} nie istnieje, choć nie wykryliśmy tego, stosując tw. 6.11.1. Załóżmy, że zachodzi przypadek IV lub V i że funkcja $r_{m-1,n}$ ma biegun x_k; nie jest to sprzeczne z warunkami interpolacyjnymi dla tej funkcji, bo nie dotyczą one tego punktu. Wtedy istnieją takie wielomiany $p \in \Pi_{m-1}$ i $q \in \Pi_{n-1}$, że

$$r_{m-1,n} = \frac{p(x)}{(x-x_k)q(x)}, \quad \text{gdzie} \quad p(x_k) \neq 0.$$

Ze wzorów odnoszących się do przypadków IV i V wynika zatem, że

$$r_{mn}(x) = \frac{p(x)}{q(x)} + y_k,$$

a wobec tego $r_{mn}(x_k) \neq y_k$ wbrew warunkowi interpolacyjnemu nałożonemu na r_{mn}. Oczywiście, podobnie może się zdarzyć na jednym z dalszych etapów obliczeń, przy przejściu od $r_{m'n'}$ do $r_{m'-1,n'}$; wtedy też r_{mn} spełnia nie wszystkie warunki typu (6.11.1). Warto zauważyć, że taka „ułomna" funkcja wymierna na pewno ma licznik stopnia $< m$ i mianownik stopnia $< n$ (te nierówności mogą jednak być spełnione i wtedy, gdy nasze zadanie ma rozwiązanie).

Przykład 6.11.4. Udowodnić, że funkcja wymierna r_{22} spełniająca warunki

x	0	1	2	4	6
y	-1	1	2	3	3

nie istnieje.

Rozwiązanie. Wybierając w określony sposób punkty x_k w przypadku IV, otrzymujemy, jak można sprawdzić, następujący ciąg równości:

$$r_{12}(x) = \frac{r_{22}(x)-3}{x-6}, \quad r_{02}(x) = \frac{r_{12}(x)}{x-4}, \quad r_{20}(x) = \frac{1}{r_{02}(x)},$$
$$r_{10}(x) = \frac{r_{20}+6}{x}, \quad r_{00}(x) = \frac{r_{10}(x)+\frac{3}{2}}{x-1}, \quad r_{00}(x) \equiv \frac{1}{2}.$$

Stąd m.in. $r_{12}(x) = (x-4)/[(\frac{1}{2}x+1)(x-6)]$. Ta funkcja ma biegun 6, który znika przy obliczeniu $r_{22}(x) = 3 + (x-6)r_{12}(x)$. Dlatego wynikająca stąd funkcja $r_{22}(x) = (5x-2)/(x+2)$ nie spełnia warunku $r_{22}(6) = 3$. ■

W obu przykładach pominięto aspekty czysto numeryczne, a są one ważne. Ograniczymy się do uwagi, że najstaranniej napisany program wzorowany na metodzie Wernera (lub jakiejkolwiek innej) może dać w wyniku błędów zaokrągleń fałszywe jakościowo informacje o zadaniu (6.11.1): że funkcja r_{mn} istnieje, chociaż tak nie jest, lub odwrotnie.

Warto również wiedzieć, że: (i) metodę Wernera można uogólnić na przypadek interpolacji Hermite'a (z węzłami wielokrotnymi), (ii) zupełnie inna metoda służy do obliczania wartości funkcji wymiernych w ustalonym punkcie (zob. np. Stoer i Bulirsch [1980]). Wiele innych wiadomości o interpolacji wymiernej podają Cuyt i Wuytack [*1987].

Aproksymacja Padégo

Załóżymy teraz, że dla funkcji aproksymowanej f znamy rozwinięcie w szereg potęgowy:

$$f(x) = \sum_{k=0}^{\infty} a_k x^k. \tag{6.11.3}$$

Zarówno współczynniki a_k, jak i zmienna x mogą być zespolone, chociaż często ograniczamy się do przypadku czysto rzeczywistego. Jeśli ten szereg ma skończony promień zbieżności ρ, to jest zbieżny w kole $|x| < \rho$ (w przypadku rzeczywistym dla $-\rho < x < \rho$), rozbieżny dla $|x| > \rho$, natomiast jego

własności na okręgu $|x| = \rho$ (dla $x = \pm\rho$) zależą od dodatkowych informacji o współczynnikach a_k.

Jeśli $|x|$ jest niewiele mniejsze od ρ, to szereg jest wolno zbieżny, a więc niezbyt użyteczny w obliczeniach. Dla $|x| > \rho$ z tego szeregu nie można korzystać nawet wtedy, gdy wiemy, że funkcja $f(x)$ określona w inny sposób (ale niesprzeczny z (6.11.3)) jest tam bardzo regularna. Tak jest np. dla logarytmu. Szereg

$$\frac{1}{x}\log(1+x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k+1} x^k \tag{6.11.4}$$

jest w dziedzinie rzeczywistej zbieżny tylko dla $-1 < x \leqslant 1$ (bardzo wolno dla $x = 1$), chociaż funkcja $\log(1+x)$ jest określona także dla $x > 1$.

Zobaczymy teraz, że znając szereg (6.11.3) można zbudować przybliżenia wymierne funkcji f, sensowne także poza kołem jego zbieżności.

Przybliżeniem Padégo $[m/n]_f$ (m, n – liczby całkowite nieujemne) szeregu (6.11.3) nazywamy funkcję wymierną $r = p/q$, gdzie $p \in \Pi_m$, $q \in \Pi_n$ i $q(0) \neq 0$ taką, że szereg potęgowy dla różnicy $f(x)q(x) - p(x)$ nie zawiera składników z $x^0, x^1, \ldots, x^{m+n}$. Ponieważ z założenia $q(0) \neq 0$, więc tę samą własność ma szereg potęgowy dla $f(x) - r(x)$. Inaczej mówiąc, chcemy aby rozwinięcie funkcji wymiernej r w szereg potęgowy miało $m+n+1$ początkowych składników wspólnych z szeregiem (6.11.3). Można to wyrazić jeszcze inaczej:

$$r^{(k)}(0) = f^{(k)}(0) \quad (0 \leqslant k \leqslant m+n).$$

Jest to więc specyficzne zadanie interpolacji wymiernej: dane są wartości funkcji r i jej $m+n$ początkowych pochodnych w jednym węźle 0. Dlatego można by sądzić, że funkcja wymierna $[m/n]_f$, jak i suma częściowa szeregu Taylora dla f, dobrze przybliża tę funkcję tylko w pewnym otoczeniu punktu 0. Okazuje się jednak, że obszar użyteczności przybliżeń Padégo może być znacznie szerszy; zob. uwagi do funkcji (6.11.8). To zdanie jest ogólnikowe, bo nie ma tu miejsca na prezentację teorii aproksymacji Padégo. Można ją znaleźć w kilku monografiach tej dziedziny (Baker [*1975], Baker i Graves-Morris [*1981]), a także w cytowanych dalej książkach poświęconych ułamkom łańcuchowym. Trzeba też jednak pamiętać, że konkretne przybliżenie $[m/n]_f$ zależy tylko od m, n i $a_0, a_1, \ldots, a_{m+n}$, a więc (jak zresztą każda funkcja interpolująca) od cząstkowej informacji o funkcji f. Nie powinno nas zatem dziwić, że taka informacja daje czasem przybliżenia zgoła fałszywe (por. zad. 15). Jeśli jednak ciąg współczynników a_k jest dostatecznie regularny, to porównanie kilku przybliżeń Padégo pozwala nam ocenić ich dokładność.

Niech będzie

$$p(x) = b_0 + b_1 x + \ldots + b_m x^m, \quad q(x) = c_0 + c_1 x + \ldots + c_n x^n \quad (c_0 \neq 0).$$

Szereg potęgowy dla różnicy $f(x)q(x) - p(x)$ otrzymujemy, wykonując występujące tu działania i grupując składniki z x w tej samej potędze. Łatwo sprawdzić, że składniki z $x^{m+1}, \ldots, x^{m+n}$ w tym szeregu znikają, jeśli

$$\sum_{j=0}^{\min\{k,n\}} a_{k-j} c_j = 0 \quad (m+1 \leqslant k \leqslant m+n). \tag{6.11.5}$$

Ponieważ z założenia $c_0 \neq 0$, więc można też przyjąć, że $c_0 = 1$, bo to wpływa tylko na unormowanie współczynników wielomianów p i q. Ostatecznie mamy układ n równań liniowych (6.11.5) względem tyluż szukanych współczynników c_j mianownika q. Jeśli ten układ ma rozwiązanie $c_0 = 1, c_1, \ldots, c_n$, to licznik p wyznaczamy ze wzorów

$$b_k = \sum_{j=0}^{\min\{k,n\}} a_{k-j} c_j \quad (0 \leqslant k \leqslant m). \tag{6.11.6}$$

Otrzymujemy je, przyrównując do 0 współczynniki różnicy $f(x)p(x) - q(x)$ przy $x^0, x^1, \ldots, x^m$.

Jeśli natomiast układ wraz z warunkiem $c_0 = 1$ nie ma rozwiązania, to przybliżenie $[m/n]_f$ nie istnieje. W szczególności tak jest dla $m = n = 1$ i $f(x) := 1 + x^2 + x^4 + \ldots$, bo wtedy układ (6.11.5) zawiera tylko równanie $a_2 c_0 + a_1 c_1 \equiv c_0 = 0$.

Rozwiązywanie układu (6.11.5) i zastosowanie wzoru (6.11.6) jest sensowną metodą znalezienia jednego przybliżenia Padégo dla niezbyt dużych m i n. W monografiach aproksymacji Padégo można znaleźć bardziej oszczędne metody, które tworzą rekurencyjnie cały ciąg tych przybliżeń, np. taki:

$$[0/0],\ [0/1],\ [1/1],\ [1/2],\ [2/2], \ldots \tag{6.11.7}$$

Tu jednak ograniczymy się do pokazania na przykładach, że przybliżenia Padégo mogą dobrze aproksymować funkcję, dla której je utworzono, a także do zwrócenia uwagi na pewne osobliwości teorii.

Przykład 6.11.5. Znaleźć przybliżenia Padégo $[n/n]$ $(n = 1, 2, 3)$ funkcji (6.11.4), obliczyć wynikające stąd wartości przybliżone dla $\log 1.5$, $\log 2$ i $\log 3$ oraz ocenić ich dokładność.

Rozwiązanie. Ponieważ $a_k = (-1)^k/(k+1)$ dla $k \geqslant 0$, więc dla $m = n = 1$ układ (6.11.5) redukuje się do jednego równania $\frac{1}{3}c_0 - \frac{1}{2}c_1 = 0$. Dla $c_0 = 1$ mamy $c_1 = \frac{2}{3}$. Z (6.11.6) wynika, że $b_0 = 1$, $b_1 = \frac{1}{6}$. Dla uproszczenia mnożymy te wszystkie współczynniki przez 6, co daje wzór

$$[1/1](x) = \frac{6+x}{6+4x}.$$

W podobny sposób, ale po rozwiązaniu układu dwóch lub trzech równań (6.11.5), otrzymujemy pozostałe szukane przybliżenia:

$$[2/2](x) = \frac{30+21x+x^2}{30+36x+9x^2}, \quad [3/3](x) = \frac{420+510x+140x^2+3x^3}{420+720x+360x^2+48x^3}.$$

Obliczenie wartości tych przybliżeń nawet na kalkulatorze jest bardzo łatwe:

	$x = 0.5$	$x = 1$	$x = 2$
$\log(1+x)$:	0.40546 51081	0.69314 71806	1.09861 2289
$x[1/1](x)$:	0.40625 00000	0.70000 00000	1.14285 7143
$x[2/2](x)$:	0.40547 26368	0.69333 33333	1.10144 9275
$x[3/3](x)$:	0.40546 51826	0.69315 24548	1.09880 5646

Wyżej podkreślono te początkowe cyfry w wartościach przybliżonych, które są takie jak w wartościach dokładnych. Jak widać, dokładność przybliżeń $[n/n](x)$ rośnie wraz z n, nawet wtedy, gdy argument x leży poza obszarem zbieżności szeregu potęgowego (ale w tym przypadku wolniej). ■

Przybliżenia Padégo mogą być pożyteczne nawet wtedy, gdy szereg potęgowy (6.11.3) jest rozbieżny dla każdego $x \neq 0$. Klasycznym przykładem jest szereg Eulera

$$\sum_{k=0}^{\infty} (-1)^k k! x^k. \tag{6.11.8}$$

Wynika on z całki niewłaściwej

$$\int_0^{\infty} \frac{e^{-t}}{1+xt}\, dt \tag{6.11.9}$$

przez rozwinięcie funkcji $1/(1+xt)$ w szereg potęgowy $1 - xt + x^2t^2 - \ldots$ i zamianę porządku całkowania i sumowania. To postępowanie jest czysto formalne, bo wspomniany szereg jest zbieżny tylko dla $|xt| < 1$. Okazuje się jednak, że ciąg przybliżeń $[n/n](x)$ jest zbieżny do wartości powyższej całki dla każdego x rzeczywistego lub zespolonego z wyjątkiem półosi rzeczywistej

$(-\infty, 0)$. Zbieżność jest dość szybka dla małych x dodatnich, a wolna np. dla x zespolonych leżących w pobliżu tej półosi (zob. zad. 13, 14).

Kończąc te wstępne informacje o aproksymacji Padégo, zauważmy jeszcze, że obliczywszy przybliżenie $[m/n]_f$ dowiadujemy się czasem, że pewne inne przybliżenia są z tym identyczne, a jeszcze inne nie istnieją; w obu przypadkach nie warto ich szukać. Przypomnijmy, że z definicji jest

$$f(x) - [m/n]_f(x) = d_{m+n+1}x^{m+n+1} + d_{m+n+2}x^{m+n+2} + \dots$$

Niech po prawej stronie pierwszym niezerowym współczynnikiem będzie d_l dla pewnego $l \geqslant m+n+1$. Wtedy:

(i) wszystkie przybliżenia $[m'/n']_f$ takie, że

$$m' \geqslant m, \quad n' \geqslant n, \quad m'+n' < l,$$

istnieją i są identyczne z $[m/n]_f$;

(ii) żadne z przybliżeń $[m'/n']_f$ takich, że

$$m < m' \leqslant l-n, \quad n < n' \leqslant l-m, \quad m'+n' \geqslant l,$$

nie istnieje.

W dwuwymiarowej *tablicy Padégo* na pozycji (μ, ν) umieszczamy przybliżenie $[\mu/\nu]_f$ (jeśli ono istnieje). W tej tablicy pozycje (m', n') wymienione wyżej w (i), (ii) tworzą kwadrat o rozmiarach $(l-m-n)\times(l-m-n)$, zwany *blokiem*. Jego lewa górna część (nieco większa od połowy) zawiera „kopie" przybliżenia $[m/n]$, a dolna część jest pusta (nie ma jej dla $l = m+n+1$, gdy blok jest jednoelementowy). Poza tym blokiem przybliżenia są różne od $[m/n]_f$. Można tu doszukać się analogii do funkcji wymiernych z zadania interpolacji rozważanego wcześniej – i tam, jak już wiemy, zadanie nie zawsze ma rozwiązanie. Szereg potęgowy jest *normalny*, jeśli cała tablica Padégo dzieli się na bloki jednoelementowe, tzn. wszystkie przybliżenia istnieją i są parami różne. Jest tak wtedy, gdy wszystkie wyznaczniki pewnego typu, utworzone ze współczynników a_k są różne od 0.

Na zakończenie tego bardzo pobieżnego przeglądu faktów związanych z aproksymacją Padégo trzeba podkreślić, że tzw. *algorytm* ε Wynna pozwala tworzyć tablicę wartości przybliżeń Padégo w ustalonym punkcie x (np. dla $x = 1$) bez obliczania ich współczynników. Ten algorytm można zatem interpretować jako metodę *przyspieszania zbieżności* szeregu $\sum_{k=0}^{\infty} a_k$. Uzasadnienie algorytmu i jego własności podają np. Baker i Graves-Morris [*1981] oraz Brezinski i Redivo Zaglia [*1991]. Przyspieszanie zbieżności jest ważną dziedziną analizy numerycznej, której dotyczy niemal w całości druga z tych książek; zob. też Weniger [*1989] (szczególnie aspekty praktyczne) i Sidi [*2003] (teoria). Metoda Aitkena (podrozdz. 5.1) i ekstrapolacja Richardsona (podrozdz. 7.1) to również metody z tej dziedziny.

Ułamki łańcuchowe

Ułamkiem łańcuchowym nazywamy szczególny rodzaj ułamka piętrowego:

$$b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \dots}}}.$$

Zapisujemy go w bardziej zwarty sposób:

$$b_0 + \frac{a_1\Big|}{\Big|b_1} + \frac{a_2\Big|}{\Big|b_2} + \frac{a_3\Big|}{\Big|b_3} + \dots$$

(w publikacjach można spotkać także nieco inne sposoby zapisu ułamków). Dla ułamków łańcuchowych nieskończonych – a te są najważniejsze – używamy także symbolu

$$b_0 + \mathop{\mathbf{K}}_{k=1}^{\infty} \frac{a_k\Big|}{\Big|b_k}, \tag{6.11.10}$$

podobnego do symbolu sumy nieskończonej. Jak w jej przypadku, tak i tutaj, trzeba oczywiście wyjaśnić, co znaczy taki ułamek, a ściślej, jaką wartość mu przypisujemy. W tym celu zdefiniujmy *n-ty redukt* ułamka (6.11.10); jest to ułamek skończony

$$b_0 + \mathop{\mathbf{K}}_{k=1}^{n} \frac{a_k\Big|}{\Big|b_k}, \tag{6.11.11}$$

powstały przez odrzucenie z (6.11.10) *n-tego ogona*, czyli ułamka nieskończonego

$$\mathop{\mathbf{K}}_{k=n+1}^{\infty} \frac{a_k\Big|}{\Big|b_k}. \tag{6.11.12}$$

Redukt i ogon są więc odpowiednikami sumy częściowej i reszty szeregu nieskończonego. Interpretacja reduktu jest oczywista. Jego wartość obliczamy, wykonując działania od końca, tzn. obliczając najpierw iloraz a_n/b_n, dodając doń b_{n-1}, dzieląc a_{n-1} przez tę sumę itd. Inny sposób, wygodniejszy, jeśli chcemy znać wszystkie redukty od zerowego do n-tego, polega na zastosowaniu następującego twierdzenia:

Twierdzenie 6.11.6. *Jeśli ciągi* $\{A_n\}$ *i* $\{B_n\}$ *są określone wzorami*

$$A_{-1} := 1, \quad B_{-1} := 0, \quad A_0 := b_0, \quad B_0 := 1,$$
$$A_n := a_n A_{n-2} + b_n A_{n-1}, \quad B_n := a_n B_{n-2} + b_n B_{n-1} \quad (n \geqslant 1),$$

to n-ty redukt (6.11.11) *ułamka łańcuchowego* (6.11.10) *jest równy* A_n/B_n $(n \geqslant 0)$.

Dowód. Redukty zerowy i pierwszy są odpowiednio równe b_0 i $b_0 + a_1/b_1$, czyli $(b_0b_1 + a_1)/b_1$. Dlatego twierdzenie jest poprawne dla $n = 0, 1$. Niech będzie $n > 1$. Załóżmy, że $(n-1)$-szy redukt wyraża się zgodnie z twierdzeniem. Zmieniając w tym ułamku ostatni mianownik, czyli b_{n-1}, na $b_{n-1} + a_n/b_n$ otrzymalibyśmy n-ty redukt ułamka (6.11.10). Można go zatem obliczyć, zmieniając wyrażenia dla A_{n-1} i B_{n-1} odpowiednio na

$$a_{n-1}A_{n-3} + \left(b_{n-1} + \frac{a_n}{b_n}\right) A_{n-2} = A_{n-1} + \frac{a_n}{b_n}A_{n-2} = \frac{A_n}{b_n}$$

i B_n/b_n. Iloraz tych liczb jest równy A_n/B_n. ■

Można już teraz sformułować następującą definicję: jeśli ciąg A_n/B_n reduktów ułamka łańcuchowego (6.11.10) ma granicę v (być może, nieskończoną), to ten ułamek jest *zbieżny*, a jego *wartością* jest v. Wtedy są zbieżne również wszystkie ogony (6.11.12). I tu mamy oczywiste analogie do szeregów nieskończonych.

Gdy dla jakiejś stałej matematycznej (jak π lub e) albo funkcji elementarnej lub specjalnej umiemy znaleźć ułamek łańcuchowy zbieżny do wartości takiej stałej lub funkcji, to mamy do dyspozycji jego redukty, które są coraz lepszymi przybliżeniami tej wartości. Oczywiście, o przydatności ułamka decyduje szybkość zbieżności ciągu reduktów. W wielu przypadkach ułamki łańcuchowe górują nad innymi wyrażeniami stałych lub funkcji, np. szeregami nieskończonymi. Nic więc dziwnego, że te ułamki są od setek lat obiektem badań i że mają rozliczne zastosowania.

Na paru przykładach pokażemy teraz, jak powstają ułamki łańcuchowe ważne w praktyce numerycznej (i jakie mają zalety). Pierwszy sposób można zastosować, gdy interesujące nas wielkości f_n (na ogół zależne od jakichś parametrów) spełniają równanie różnicowe jednorodne drugiego rzędu:

$$\alpha_n f_{n-1} - \beta_n f_n - f_{n+1} = 0 \quad (n = 1, 2, \ldots). \tag{6.11.13}$$

Wtedy

$$\alpha_n \frac{f_{n-1}}{f_n} - \beta_n - \frac{f_{n+1}}{f_n} = 0,$$

a stąd wynika, że

$$\frac{f_n}{f_{n-1}} = \frac{\alpha_n}{\beta_n + \dfrac{f_{n+1}}{f_n}}.$$

Iterując tę równość, otrzymujemy ułamek łańcuchowy nieskończony:

$$\frac{f_1}{f_0} = \mathop{\mathbf{K}}_{n=1}^{\infty} \frac{\alpha_n}{\beta_n}. \tag{6.11.14}$$

Z tych przekształceń nie wynika, czy otrzymany ułamek jest zbieżny, a jeśli tak, to czy jego wartością jest f_1/f_0. Trzeba pamiętać, że równanie (6.11.13) ma dwuparametrową rodzinę rozwiązań $\{f_n\}$, określonych przez dowolny wybór właśnie f_0 i f_1. Istotne jest to, czy w tej rodzinie istnieje rozwiązanie *minimalne* $\{f_n\}$, tj. takie, że nie wszystkie jego elementy znikają i że dla pewnego innego rozwiązania $\{f'_n\}$ jest $\lim_{n\to\infty} f_n/f'_n = 0$. Oczywiście każde rozwiązanie $\{cf_n\}$ $(c \neq 0)$ też jest minimalne, ale iloraz f_1/f_0 dla tych rozwiązań nie zależy od c.

Twierdzenie 6.11.7 (Pincherle). *Jeśli równanie różnicowe* (6.11.13) *ma rozwiązanie minimalne* $\{f_n\}$, *to ułamek łańcuchowy z* (6.11.14) *jest zbieżny do* f_1/f_0.

W podrozdziałach 1.3 i 2.3 wspomniano o funkcjach Bessela $J_n(x)$ i $I_n(x)$ spełniających to samo równanie różnicowe, które napiszemy teraz w postaci zgodnej z (6.11.13):

$$-f_{n-1} + 2nx^{-1}f_n - f_{n+1} = 0.$$

Z teorii tych funkcji wynika, że rozwiązaniem minimalnym jest $\{J_n(x)\}$ i dlatego na mocy tw. 6.11.7 dla każdego x zespolonego jest

$$\frac{J_1(x)}{J_0(x)} = \mathop{\mathrm{K}}_{n=1}^{\infty} \frac{-1}{2nx^{-1}} = \frac{x}{2} - \frac{x^2}{4} - \frac{x^2}{6} + \dots$$

(fragment $-\dfrac{\alpha}{\beta}$ ułamka łańcuchowego jest identyczny z $+\dfrac{-\alpha}{\beta}$; drugą równość uzasadnia zad. 16).

Inną metodę stosujemy, chcąc przekształcić szereg potęgowy na ułamek łańcuchowy szczególnego typu. Rozważymy zresztą nieco ogólniejsze zadanie, a mianowicie przekształcenie ilorazu

$$\frac{\sum_{k=0}^{\infty} a_{0k}x^k}{\sum_{k=0}^{\infty} a_{1k}x^k} \tag{6.11.15}$$

takich szeregów (oczywiście jeden z nich może być stałą). Służy do tego *metoda Wiskowatowa* oparta na poniższym twierdzeniu, które zarazem pokazuje związek dwóch pojęć: przybliżeń Padégo i ułamków łańcuchowych.

Twierdzenie 6.11.8 (Wiskowatow). *Jeśli wszystkie liczby* a_{n0} $(n \geqslant 1)$ *obliczane według wzorów*

$$\gamma_n := \frac{a_{n0}}{a_{n+1,0}}, \quad a_{n+2,k} := a_{n,k+1} - \gamma_n a_{n+1,k+1} \quad (k \geqslant 0)$$

stosowanych dla $n = 0, 1, \ldots$, *są różne od zera, to redukty ułamka łańcuchowego*

$$\gamma_0 + \mathop{\mathbf{K}}_{n=1}^{\infty} \frac{x}{\gamma_n} \tag{6.11.16}$$

są przybliżeniami Padégo

$$[m/m](x), \quad [m+1/m](x) \quad (m \geqslant 0)$$

ilorazu (6.11.15).

Zmodyfikowana nieco metoda Wiskowatowa daje taki ułamek łańcuchowy, którego redukty tworzą inny ciąg przybliżeń, również tworzący w tablicy Padégo linię schodkową. Szczegóły pomijamy.

Dowód. W pierwszym kroku metody Wiskowatowa dla $\gamma_0 = a_{00}/a_{10}$ przekształcamy iloraz (6.11.15) w następujący sposób:

$$\gamma_0 + \frac{\sum_{k=0}^{\infty} a_{0k}x^k - \gamma_0 \sum_{k=0}^{\infty} a_{1k}x^k}{\sum_{k=0}^{\infty} a_{1k}x^k} = \gamma_0 + \frac{\sum_{k=1}^{\infty}(a_{0k} - \gamma_0 a_{1k})x^k}{\sum_{k=0}^{\infty} a_{1k}x^k} =$$

$$= \gamma_0 + \frac{x}{\dfrac{\sum_{k=0}^{\infty} a_{1k}x^k}{\sum_{k=0}^{\infty} a_{2k}x^k}}.$$

Postępując podobnie dalej, przekształcamy formalnie iloraz (6.11.15) na ułamek (6.11.16). Dowód podanej w twierdzeniu własności reduktów tego ułamka jest bardziej złożony; zob. np. Lorentzen i Waadeland [*1992, s. 253]. ■

Przykład 6.11.9. Stosując metodę Wiskowatowa, znaleźć czwarty redukt ułamka (6.11.16) dla ilorazu szeregów potęgowych o współczynnikach

$$a_{0k} := \frac{(-1)^k}{(2k+1)!}, \quad a_{1k} := \frac{(-1)^k}{(2k)!} \quad (k \geqslant 0),$$

tj. dla funkcji $(1/\sqrt{x})\operatorname{tg}\sqrt{x}$ (dzięki przekształceniu zmiennej nie będą powstawały szeregi z zerowym początkowym współczynnikiem).

Rozwiązanie. Łatwo sprawdzić, że wystarczy tu uwzględnić współczynniki a_{0k}, a_{1k} dla $k \leqslant 4$. Wyniki obliczeń według tw. 6.11.8 są następujące:

n	γ_n	a_{n0}	a_{n1}	a_{n2}	a_{n3}	a_{n4}
0	1	1	$-\frac{1}{6}$	$\frac{1}{120}$	$-\frac{1}{5040}$	$\frac{1}{362880}$
1	3	1	$-\frac{1}{2}$	$\frac{1}{24}$	$-\frac{1}{720}$	$\frac{1}{40320}$
2	$-\frac{5}{6}$	$\frac{1}{3}$	$-\frac{1}{30}$	$\frac{1}{840}$	$-\frac{1}{45360}$	
3	252	$-\frac{2}{5}$	$\frac{4}{105}$	$-\frac{1}{756}$		
4	$-\frac{1}{10}$	$-\frac{1}{630}$	$\frac{1}{11340}$			
5		$\frac{1}{63}$				

Szukany redukt jest więc równy

$$1+\cfrac{x}{3}+\cfrac{x}{-\frac{5}{6}}+\cfrac{x}{252}+\cfrac{x}{-\frac{1}{10}}. \tag{6.11.17}$$

Stąd oraz z tw. 6.11.8 i 6.11.6 wynika, że dla rozważanego ilorazu szeregów potęgowych jest m.in.

$$[1/1](x)=\frac{5-\frac{1}{3}x}{5-2x},\quad [2/2](x)=\frac{63-7x+\frac{1}{15}x^2}{63-28x+x^2}$$

(zad. 20). Zmieniając tu x na x^2 i mnożąc ilorazy przez x, otrzymujemy przybliżenia wymierne funkcji $\operatorname{tg} x$. Zamiast porównywać ich wartości, zwróćmy tu uwagę na inną cenną cechę tych przybliżeń. Otóż najmniejszy dodatni biegun funkcji $(1/\sqrt{x})\operatorname{tg}\sqrt{x}$ jest równy $\frac{1}{4}\pi^2\approx 2.467401$, natomiast jedyny biegun pierwszego przybliżenia i mniejszy drugiego są odpowiednio równe 2.5 i $14-\sqrt{133}\approx 2.46744$. Podobnie, najmniejsze dodatnie zero tejże funkcji jest równe $\pi^2\approx 9.87$, a przybliżenie $[2/2](x)$ ma najmniejsze zero 9.94. Jak widać, przybliżenia $[n/n]$ nawet dla małych n dobrze odtwarzają tak istotne punkty szeregu potęgowego, jakimi są zera i bieguny. W pewnych zastosowaniach obliczanie dalekich współczynników szeregu potęgowego jest kosztowne lub wręcz niemożliwe. Znając kilka początkowych współczynników, można jednak dość dokładnie zlokalizować np. najmniejszy biegun.

Dodajmy jeszcze, że najprostszy cytowany ułamek łańcuchowy dla tangensa ma postać

$$\operatorname{tg} x=\cfrac{x}{1}-\cfrac{x^2}{3}-\cfrac{x^2}{5}-\cfrac{x^2}{7}-\dots \tag{6.11.18}$$

i jest zbieżny na całej płaszczyźnie zespolonej (a szereg potęgowy tylko dla $|x|<\pi/2$). ■

W związku z przykł. 6.11.5 warto wspomnieć, że na całej płaszczyźnie zespolonej z wyjątkiem półprostej rzeczywistej $(-\infty,-1]$ jest

$$\log(1+x)=\cfrac{x}{1}+\cfrac{x}{2}+\cfrac{x}{3}+\cfrac{2x}{2}+\cfrac{2x}{5}+\cfrac{3x}{2}+\cfrac{3x}{7}+\dots \tag{6.11.19}$$

Zbieżność tego ułamka jest na ogół tym gorsza, im x jest bliższe wspomnianej półprostej. Ogólniej, jeśli ułamek jest bardzo wolno zbieżny, to często jest możliwe i celowe zastosowanie specyficznych metod przyspieszania zbieżności. Opierają się one na ogół na tym, że dla wielu ułamków można stosunkowo łatwo – uwzględniając ich własności – znaleźć dość dobre

przybliżenie $\tilde{t}_n$ wartości n-tego ogona (6.11.12). *Dokładna* wartość t_n tego ogona pozwoliłaby wyrazić wartość v ułamka nieskończonego (6.11.10) w skończonej postaci:

$$v = b_0 + \frac{a_1}{\left|b_1\right.} + \ldots + \frac{a_{n-1}}{\left|b_{n-1}\right.} + \frac{a_n}{\left|b_n + t_n\right.}.$$

Skoro t_n nie znamy, to jest naturalne użyć tu jego przybliżenia $\tilde{t}_n$. Daje to *n-ty redukt zmodyfikowany*, którego wartość $\tilde{v}$ przybliża szukane v. Można łatwo udowodnić, że

$$\tilde{v} = \frac{A_{n-1} + \tilde{t}_n A_n}{B_{n-1} + \tilde{t}_n B_n} \tag{6.11.20}$$

(zad. 24).

Na zakończenie warto podkreślić, że ułamki (6.11.18) i (6.11.19) (a także wiele innych ważnych w zastosowaniach) wynikają z twierdzeń dotyczących szeregów hipergeometrycznych. Ułamki podobne do zacytowanych są przybliżeniami Padégo odpowiednich funkcji; zob. zad. 22.

Literatura dotycząca ułamków łańcuchowych jest bardzo obfita; p. np. Perron [1929], Khowanskii [1963], Jones i Thron [*1980] oraz Lorentzen i Waadeland [*1992].

ZADANIA 6.11

1. Znaleźć funkcję wymierną r (lub udowodnić, że ona nie istnieje) określoną warunkami (6.11.1), gdzie:

(a) $m = 1, n = 2,$

x	0	1	2	3
y	3	2	0	4

(b) $m = n = 2,$

x	0	2	3	5	6
y	−4	1	8	11	12

2. Udowodnić, że jeśli funkcja wymierna $r = p/q$ spełniająca warunki (6.11.1), gdzie $m > 0$, $n > 0$, istnieje i jest taka, że $p \in \Pi_{m-1}$, $q \in \Pi_{n-1}$, to nie istnieje funkcja $r^* = p^*/q^*$, dla której $p^* \in \Pi_m$, $q^* \in \Pi_n$ i która spełnia te same warunki z wyjątkiem jednego: $r^*(x_j) = y_j$ dla $j \neq k$ i $r^*(x_k) = y_k^* \neq y_k$.

3. (cd.). Korzystając z poprzedniego zadania, znaleźć takie przykładowe warunki interpolacyjne (6.11.1), które nie mogą być spełnione.

4. (cd.). Sprawdzić na przykładzie, że nieistnienie funkcji wymiernej spełniającej ustalone warunki (6.11.1) może się ujawnić – zależnie od wyboru wskaźnika k w przypadku III — na dwa różne sposoby: albo w toku obliczeń wystąpi przypadek II, albo zdarzy się sytuacja opisana w przykł. 6.11.4.

5. Metodę Wernera można uogólnić na interpolację z węzłami wielokrotnymi. Jak warunki interpolacyjne $r_{mn}^{(i)}(x_j) = y_{ji}$ $(i = 0, 1, \ldots, d)$ przenoszą się na funkcję wymierną $r_{m-1,n}$ (przypadek IV w tw. 6.11.1) lub r_{nm} (przypadek V)?

6. Dla dowolnego szeregu (6.11.3) jest $[m/0](x) = a_0 + a_1 x + \ldots + a_m x^m$ (sprawdzić). Jak dla $a_0 \neq 0$ wyraża się $[0/m]$?

7. Jaki fragment tablicy Padégo można znaleźć, znając współczynniki a_k szeregu tylko dla $k \leqslant 6$?

8. Niech f będzie funkcją parzystą: $f(x) = a_0 + a_2 x^2 + a_4 x^4 + \ldots$. Czy możemy stąd coś wywnioskować, nie wykonując konkretnych obliczeń, o wielkości i rozmieszczeniu bloków w tablicy Padégo? Odpowiedzieć na to samo pytanie dla funkcji nieparzystej.

9. Niech będzie $f(x) := 1 - x^4 + x^5 - x^6 + \ldots$. Znaleźć wszystkie przybliżenia Padégo $[m/n]_f$ zależne tylko od współczynników a_k tego szeregu dla $k \leqslant 6$. Jak można wykorzystać informacje o blokach w tablicy Padégo? Czy zawiera ona bloki rozmiaru 2×2 lub większe?

10. Niech będzie $[m/n] = p_{mn}/q_{mn}$ i $d_{mn} := f q_{mn} - p_{mn}$. Załóżmy, że dla szeregu normalnego znamy przybliżenia $[m/n]$ i $[m/n+1]$ oraz rozwinięcia w szereg różnic d_{mn} i $d_{m,n+1}$. Udowodnić, że dla pewnego φ_{mn} jest

$$s_{m+1,n+1}(x) = x s_{mn}(x) + \varphi_{mn} s_{m,n+1}(x) \quad (s \equiv p, q, d).$$

Jak znaleźć tę stałą? Podobnie, wykazać, że

$$s_{m+1,n+1}(x) = x s_{mn}(x) + \psi_{mn} s_{m+1,n}(x).$$

(Oba wzory, wraz z rozwiązaniem zad. 6, pozwalają tworzyć kolejne przybliżenia Padégo z linii schodkowej podobnej do (6.11.7)).

11. Udowodnić, że jeśli szereg potęgowy jest normalny, to istnieją stałe χ_{mn} i ω_{mn} takie, że

$$s_{m,n+1}(x) = x s_{mn}(x) + \chi_{mn} s_{m+1,n}(x),$$
$$s_{m+1,n}(x) = x s_{mn}(x) + \omega_{mn} s_{m,n+1}(x)$$

(oznaczenia jak w poprzednim zadaniu). Czy stosowanie tych wzorów wymaga posługiwania się także szeregami d_{mn}?

12. (cd.). Przybliżenie Padégo $[m/-1](x)$ można sztucznie określić jako iloraz x^m przez 0. Jaka własność to usprawiedliwia? Korzystając z tej definicji i zad. 11 i 6, znaleźć dla funkcji e^x wszystkie przybliżenia Padégo $[m/n]$ takie, że $m + n = 4$.

13. Udowodnić, że mianownik przybliżenia Padégo $[n/n](x)$ szeregu (6.11.8) jest równy

$$\sum_{k=0}^{n} \frac{(n-k+1)_k (n-k+2)_k}{k!} x^k,$$

gdzie $(a)_k := a(a+1)\ldots(a+k-1)$ (jest to *symbol Pochhammera*). Wskazówka: Sprawdzając, że jest spełniony układ (6.11.5) dla $m = n$, skorzystać z tego, że $\Delta^n \varphi(j) = 0$, jeśli φ jest wielomianem stopnia $< n$.

14. (cd.). Znaleźć mianowniki i liczniki przybliżeń Padégo z poprzedniego zadania dla $n = 1, 2, 3, 4$ oraz ich wartości dla $x = 0.5, 1, 2$; zob. też zad. K2.

15. Niech pewna funkcja f ciągła co najmniej w przedziale $[-1, 1]$ rozwija się w szereg $1+\varepsilon x+x^2+\ldots$, gdzie ε jest małą liczbą dodatnią. Znaleźć przybliżenie $[1/1]_f$ i narysować jego wykres. Wnioski powinny skłaniać do ostrożności przy stosowaniu konkretnych przybliżeń Padégo.

16. Udowodnić, że jeśli liczby ρ_n $(n \geqslant 1)$ są różne od zera, to ułamek

$$b_0 + \frac{a_1\rho_1\rfloor}{\lceil b_1\rho_1} + \frac{a_2\rho_1\rho_2\rfloor}{\lceil b_2\rho_2} + \frac{a_3\rho_2\rho_3\rfloor}{\lceil b_3\rho_3} + \ldots$$

jest *równoważny* ułamkowi (6.11.10), tj. że redukty o tym samym wskaźniku obu ułamków są sobie równe (a te ułamki mają identyczne wartości).

17. Udowodnić, że redukty ułamka łańcuchowego (6.11.10) spełniają równość

$$\frac{A_{n+1}}{B_{n+1}} - \frac{A_n}{B_n} = \frac{(-1)^n a_1 a_2 \ldots a_{n+1}}{B_n B_{n+1}} \qquad (n \geqslant 0).$$

18. (cd.). Udowodnić, że jeśli parametry a_k, b_k zbieżnego ułamka łańcuchowego są rzeczywiste i dodatnie, to jego redukty są na przemian mniejsze i większe od wartości ułamka.

19. (cd.). Jak zachowują się redukty względem wartości ułamka, jeśli założenie $a_k > 0$ z poprzedniego zadania zmienimy na $(-1)^{k-1}a_k > 0$?

20. Znaleźć A_n, B_n dla początkowych reduktów ułamka (6.11.17) i sprawdzić poprawność przybliżeń Padégo z przykł. 6.11.9.

21. (cd.). Znaleźć A_n, B_n $(n \leqslant 5)$ dla ułamka (6.11.18). Czy są one w jakiś sposób związane z reduktami ułamka (6.11.17)?

22. Znaleźć początkowe redukty ułamka łańcuchowego (6.11.19) i ich związki z przybliżeniami Padégo z przykł. 6.11.5.

23. Wiadomo, że

$$e^x = \frac{1\rfloor}{\lceil 1} - \frac{x\rfloor}{\lceil 1} + \frac{x\rfloor}{\lceil 2} - \frac{x\rfloor}{\lceil 3} + \frac{x\rfloor}{\lceil 2} - \frac{x\rfloor}{\lceil 5} + \frac{x\rfloor}{\lceil 2} - \frac{x\rfloor}{\lceil 7} + \ldots$$

Udowodnić, że stąd wynika ułamek łańcuchowy

$$e^x = 1 + \frac{x\rfloor}{\lceil 1} - \frac{x\rfloor}{\lceil 2} + \frac{x\rfloor}{\lceil 3} - \frac{x\rfloor}{\lceil 2} + \frac{x\rfloor}{\lceil 5} - \frac{x\rfloor}{\lceil 2} + \frac{x\rfloor}{\lceil 7} - \ldots$$

Obliczyć kilka początkowych reduktów tych ułamków dla $x = 2$. Korzystając z zad. 19, oszacować z obu stron liczbę e^2.

24. Udowodnić wyrażenie (6.11.20) dla reduktu zmodyfikowanego (por. dowód tw. 6.11.6).

ZADANIA KOMPUTEROWE 6.11

K1. Zaprogramować metodę Wernera (tw. 6.11.2) i zbadać jej skutki w przypadkach, gdy funkcja r_{mn} nie istnieje.

K2. Dla kontroli jakości przybliżeń Padégo z zad. 14 obliczyć dostatecznie dokładne wartości całki (6.11.9). W tym celu: (i) podzielić przedział całkowania $[0,\infty)$ na części $[0,K]$ i $[K,\infty)$ dla K tak dużego, aby wartość całki w drugim przedziale była zaniedbywalnie mała, (ii) całkę w $[0,K]$ obliczyć, stosując np. metodę adaptacyjną z podrozdz. 7.6.

K3. Zaprogramować obliczanie: (**a**) ciągu reduktów według tw. 6.11.6, (**b**) jednego reduktu z definicji (6.11.11).

K4. Zaprogramować metodę Wiskowatowa (tw. 6.11.8).

6.12. Interpolacja trygonometryczna

Jak wiadomo, dla dowolnych $n+1$ różnych liczb rzeczywistych (węzłów) $x_0, x_1, \ldots, x_n$ i dowolnych wartości $y_0, y_1, \ldots, y_n$ istnieje dokładnie jeden wielomian p klasy Π_n, czyli stopnia co najwyżej n, spełniający warunki interpolacyjne

$$p(x_j) = y_j \quad (0 \leqslant j \leqslant n).$$

Szeregi Fouriera

Wielomiany algebraiczne nie nadają się do opisu zjawisk okresowych. Znacznie bardziej odpowiednie są tu funkcje trygonometryczne. Ściślej, trzeba je dobrać do okresu interesującej nas funkcji. Załóżmy dla prostoty, że funkcja, którą chcielibyśmy interpolować, ma okres 2π. Wtedy funkcjami bazowymi mogą być $1, \cos x, \cos 2x, \ldots, \sin x, \sin 2x, \ldots$. Jedno z podstawowych twierdzeń analizy Fouriera głosi, że jeśli funkcja f o okresie 2π ma pierwszą pochodną ciągłą, to jej szereg Fouriera

$$\frac{1}{2}a_0 + \sum_{k=1}^{\infty}(a_k \cos kx + b_k \sin kx), \tag{6.12.1}$$

o współczynnikach

$$a_k := \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos kt\,dt, \qquad b_k := \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin kt\,dt, \tag{6.12.2}$$

jest zbieżny jednostajnie do f. Uzasadnia to sensowność aproksymacji funkcji o okresie 2π za pomocą kombinacji liniowych wyliczonych wyżej funkcji trygonometrycznych.

Zespolone szeregi Fouriera

Wiele wyników analizy Fouriera przybiera bardziej elegancką postać, gdy wyrażamy je, korzystając ze wzoru Eulera

$$e^{i\theta} = \cos\theta + i\sin\theta,$$

gdzie $i^2 = -1$. Wtedy szereg Fouriera funkcji okresowej f o wartościach zespolonych jest dany wzorem

$$f(x) \sim \sum_{k=-\infty}^{\infty} \hat{f}(k)e^{ikx}, \tag{6.12.3}$$

gdzie

$$\hat{f}(k) := \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)e^{-ikt}\,dt.$$

Jeśli f jest funkcją rzeczywistą, to jej szereg Fouriera (6.12.1) jest częścią rzeczywistą szeregu zespolonego (6.12.3). Istotnie,

$$\hat{f}(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)[\cos kt - i\sin kt]\,dt = \frac{1}{2}(a_k - ib_k) \quad (k \geqslant 0)$$

i podstawiając to wyrażenie do (6.12.3), otrzymujemy (6.12.1).

Twierdzenie 6.12.1. *Jeśli dla danych ciągów rzeczywistych* $\{a_k\}_0^\infty$ *i* $\{b_k\}_1^\infty$

$$b_0 := 0, \quad a_{-k} := a_k, \quad b_{-k} := -b_k, \quad c_k = \tfrac{1}{2}(a_k - ib_k) \quad (k > 0),$$

to

$$\frac{1}{2}a_0 + \sum_{k=1}^{n}(a_k\cos kx + b_k\sin kx) = \sum_{k=-n}^{n} c_k e^{ikx}.$$

Dowód. Suma po prawej stronie jest równa

$$\frac{1}{2}\sum_{k=-n}^{n}(a_k - ib_k)(\cos kx + i\sin kx).$$

Jej część rzeczywista jest równa

$$\frac{1}{2}\sum_{k=-n}^{n}(a_k\cos kx+b_k\sin kx)=$$
$$=\frac{1}{2}\sum_{k=1}^{n}[a_{-k}\cos(-kx)+b_{-k}\sin(-kx)]+\frac{1}{2}a_0+$$
$$+\frac{1}{2}\sum_{k=1}^{n}(a_k\cos kx+b_k\sin kx)=\frac{1}{2}a_0+\sum_{k=1}^{n}(a_k\cos kx+b_k\sin kx),$$

a część urojona jest równa 0, gdyż

$$\frac{1}{2}\sum_{k=-n}^{n}(a_k\sin kx-b_k\cos kx)=$$
$$=\frac{1}{2}\sum_{k=1}^{n}[a_{-k}\sin(-kx)-b_{-k}\cos(-kx)]-\frac{1}{2}b_0+$$
$$+\frac{1}{2}\sum_{k=1}^{n}(a_k\sin kx-b_k\cos kx)=$$
$$=\frac{1}{2}\sum_{k=1}^{n}(-a_k\sin kx+b_k\cos kx)+\frac{1}{2}\sum_{k=1}^{n}(a_k\sin kx-b_k\cos kx)=0. \quad \blacksquare$$

Iloczyn skalarny i pojęcia pochodne

Funkcje E_k takie, że

$$E_k(x):=e^{ikx}\quad(k=0,\pm1,\pm2,\ldots),$$

tworzą układ ortonormalny w przestrzeni Hilberta $L_2[-\pi,\pi]$ funkcji o wartościach zespolonych, w której iloczyn skalarny jest określony wzorem

$$\langle f,g\rangle:=\frac{1}{2\pi}\int_{-\pi}^{\pi}f(x)\overline{g(x)}\,dx.$$

Istotnie, $\langle E_k,E_m\rangle=0$ dla $k\neq m$ i $\langle E_k,E_k\rangle=1$. Sprawdzimy pierwszą równość:

$$\langle E_k,E_m\rangle=\frac{1}{2\pi}\int_{-\pi}^{\pi}E_k(x)\overline{E_m(x)}\,dx=\frac{1}{2\pi}\int_{-\pi}^{\pi}e^{i(k-m)x}\,dx=$$
$$=\frac{1}{2\pi}\left.\frac{e^{i(k-m)x}}{i(k-m)}\right|_{-\pi}^{\pi}=0.$$

Będzie też nam potrzebny tzw. *pseudoiloczyn skalarny*

$$\langle f,g\rangle_N:=\frac{1}{N}\sum_{j=0}^{N-1}f(2\pi j/N)\overline{g(2\pi j/N)}. \tag{6.12.4}$$

Nazwę uzasadnia to, że ta wielkość nie spełnia tylko jednego warunku określającego iloczyn skalarny (o wartościach zespolonych): jeśli $\langle f, f\rangle_N = 0$, to wiemy tylko, że f znika we wszystkich punktach $2\pi j/N$, a nie, że jest równa wszędzie zeru. Jest natomiast prawdą, że:

1. $\langle f, f\rangle_N \geqslant 0$.
2. $\langle f, g\rangle_N = \overline{\langle g, f\rangle_N}$.
3. $\langle \alpha f + \beta g, h\rangle_N = \alpha\langle f, h\rangle_N + \beta\langle g, h\rangle_N$.

Pseudoiloczyn skalarny indukuje *pseudonormę*:

$$\|f\|_N := \sqrt{\langle f, f\rangle_N}.$$

Jest ona zerem wtedy i tylko wtedy, gdy $f(2\pi j/N) = 0$ dla $0 \leqslant j \leqslant N-1$.

W interpolacji trygonometrycznej następujące twierdzenie jest szczególnie ważne:

TWIERDZENIE 6.12.2. *Dla każdego* $N \geqslant 1$

$$\langle E_k, E_m\rangle_N = \begin{cases} 1, & \text{jeśli } k-m \text{ dzieli się przez } N \\ 0 & \text{w przeciwnym razie.} \end{cases}$$

Dowód. Powyższy pseudoiloczyn skalarny jest równy

$$\frac{1}{N}\sum_{j=0}^{N-1} E_k\left(\frac{2\pi j}{N}\right)\overline{E_m\left(\frac{2\pi j}{N}\right)} = \frac{1}{N}\sum_{j=0}^{N-1}(e^{2\pi i(k-m)/N})^j.$$

Jeśli $k-m$ dzieli się przez N, czyli iloraz $(k-m)/N$ jest liczbą całkowitą, to $e^{2\pi i(k-m)/N} = 1$ i cała suma też jest równa 1. W przeciwnym razie $e^{2\pi i(k-m)/N} \neq 1$ i można zastosować wzór na sumę wyrazów ciągu geometrycznego:

$$\sum_{j=0}^{N-1}\lambda^j = \frac{\lambda^N - 1}{\lambda - 1} \quad (\lambda \neq 1).$$

Interesująca nas suma jest zatem równa

$$\frac{e^{2\pi i(k-m)} - 1}{e^{2\pi i(k-m)/N} - 1} = 0. \qquad \blacksquare$$

Wielomiany wykładnicze

Wielomianem wykładniczym stopnia co najwyżej n-tego nazywamy funkcję

$$P(x) = \sum_{k=0}^{n} c_k e^{ikx} = \sum_{k=0}^{n} c_k E_k(x) = \sum_{k=0}^{n} c_k (e^{ix})^k,$$

czyli zwykły wielomian względem zmiennej e^{ix}. Pewne fakty dotyczące interpolacji za pomocą wielomianów wykładniczych podano niżej.

Twierdzenie 6.12.3. *Układ* $\{E_0, E_1, \ldots, E_{N-1}\}$ *jest ortonormalny względem pseudoiloczynu skalarnego* $\langle \cdot, \cdot \rangle_N$.

Wniosek 6.12.4. *Wielomian wykładniczy* P *interpolujący funkcję* f *w węzłach równoodległych* $x_j := 2\pi j/N$ *wyraża się wzorem*

$$P = \sum_{k=0}^{N-1} c_k E_k, \quad \text{gdzie} \quad c_k := \langle f, E_k \rangle_N. \tag{6.12.5}$$

Dowód. Z definicji c_k i z tw. 6.12.3 wynika, że

$$\begin{aligned}
\sum_{k=0}^{N-1} c_k E_k(x_\nu) &= \sum_{k=0}^{N-1} \langle f, E_k \rangle_N E_k(x_\nu) = \\
&= \sum_{k=0}^{N-1} \frac{1}{N} \sum_{j=0}^{N-1} f(x_j) \overline{E_k(x_j)} E_k(x_\nu) = \\
&= \sum_{j=0}^{N-1} f(x_j) \frac{1}{N} \sum_{k=0}^{N-1} \overline{E_j(x_k)} E_\nu(x_k) = \\
&= \sum_{j=0}^{N-1} f(x_j) \langle E_\nu, E_j \rangle_N = f(x_\nu).
\end{aligned}$$

■

Ponieważ, jak już wspomniano, wielomiany wykładnicze różnią się od zwykłych tylko wyborem zmiennej, więc z tw. 6.1.1 wynika, że wielomian P spełniający warunki interpolacyjne podane w powyższym wniosku jest określony jednoznacznie.

Przykład 6.12.5. Stosując wniosek 6.12.4, wyrazić funkcję interpolującą dla $N = 2$.

Rozwiązanie. W tym przypadku mamy wielomian wykładniczy stopnia co najwyżej pierwszego, interpolujący f w węzłach 0 i π. Z (6.12.5) wynika, że

$$P(x) = \tfrac{1}{2}[f(0) + f(\pi)] + \tfrac{1}{2}[f(0) - f(\pi)]e^{ix}.$$

■

Z twierdzeń 6.8.7 i 6.12.3 wynika natychmiast

Wniosek 6.12.6. *Jeśli $n < N$, to wielomian wykładniczy $\sum_{k=0}^{n} c_k E_k$, którego współczynniki c_k są określone we wniosku 6.12.4, aproksymuje najlepiej funkcję f w sensie najmniejszych kwadratów na zbiorze skończonym*

$$x_j := 2\pi j/N \quad (0 \leqslant j \leqslant N-1).$$

ZADANIA 6.12

1. Sprawdzić, że $E_k E_m = E_{k+m}$ i $\overline{E_k} = E_{-k}$.
2. Udowodnić, że jeśli funkcje f i g są takie, że $f(x_j) = \langle g, E_j \rangle_n$ dla $x_j := 2\pi j/n$, to $g(x_j) = n\langle f, E_j \rangle_n$.
3. Sprawdzić, że pseudoiloczyn skalarny (6.12.4) ma własności **1–3** (zob. tekst).
4. Korzystając z informacji podanych w tekście, udowodnić, że

$$\frac{1}{n}\sum_{j=0}^{n-1} \cos\frac{2\pi jk}{n} = \begin{cases} 1, & \text{jeśli } k \text{ dzieli się przez } n \\ 0 & \text{w przeciwnym razie,} \end{cases}$$

$$\frac{1}{n}\sum_{j=0}^{n-1} \sin\frac{2\pi jk}{n} = 0.$$

6.13. Szybkie przekształcenie Fouriera

Przekształcenia (transformacje) Fouriera służą do rozkładu sygnału na częstotliwości składowe. Mamy tu podobieństwo do pryzmatu rozszczepiającego białe światło na składowe barwy. Inna analogia: okulary przeciwsłoneczne redukują oślepianie białym światłem, przepuszczając tylko łagodniejsze światło zielone, a przekształcenie Fouriera może zmodyfikować sygnał w celu otrzymania pożądanego efektu. Analizując częstotliwości składowe sygnału lub układu szeregi i przekształcenia Fouriera znajdują wielorakie zastosowania w sterowaniu samolotami i statkami kosmicznymi, przetwarzaniu sygnałów cyfrowych, obrazowaniu medycznym, rozpoznaniu złóż ropy lub gazu i rozwiązywaniu równań różniczkowych; zob. np. Briggs i Henson [1995] lub Walker [1992].

Ten podrozdział dotyczy obliczeniowych aspektów interpolacji trygonometrycznej. W szczególności jest tu opisany algorytm szybkiego przekształcenia Fouriera (angielski skrót: FFT), służący do efektywnego obliczania współczynników c_k z (6.12.5). Jego opis wzorujemy na opisie podanym

przez Stoera i Bulirscha [1980]. Wspomniane współczynniki można wyrazić w postaci

$$c_k = \frac{1}{N}\sum_{j=0}^{N-1} f(x_j)(\lambda^k)^j \quad (\lambda := e^{-2\pi i/N}).$$

Tak więc c_k wymaga obliczenia wartości pewnego wielomianu stopnia $N-1$ w punkcie λ^k. Koszt tych obliczeń to około N mnożeń i N dodawań. Ponieważ mamy obliczyć N współczynników c_k, więc ogólny koszt wyznaczenia wielomianu wykładniczego tą najprostszą metodą wynosi $\mathcal{O}(N^2)$ działań.

Szybkie przekształcenie Fouriera redukuje ten koszt do bardziej rozsądnej wielkości $N\log_2 N$. Poniższa tabelka pokazuje, co to znaczy dla dużych wartości N typowych w przetwarzaniu sygnałów:

N	N^2	$N\log_2 N$
1024	1048576	10240
4096	16777216	49152
16384	268435456	229376

Twierdzenie 6.13.1. *Jeśli dla danej funkcji f wielomiany wykładnicze p i q stopnia $\leqslant n-1$ są takie, że dla $x_j := \pi j/n$ jest*

$$p(x_{2j}) = f(x_{2j}), \quad q(x_{2j}) = f(x_{2j+1}) \quad (0 \leqslant j \leqslant n-1),$$

to wielomian wykładniczy P stopnia $\leqslant 2n-1$, interpolujący f w punktach $x_0, x_1, \ldots, x_{2n-1}$, wyraża się wzorem

$$P(x) = \tfrac{1}{2}(1+e^{inx})p(x) + \tfrac{1}{2}(1-e^{inx})q(x-\pi/n). \tag{6.13.1}$$

Dowód. Ponieważ z założenia stopień wielomianów p i q nie przekracza $n-1$, a e^{inx} jest stopnia n, więc stopień wielomianu P jest równy co najwyżej $2n-1$. Pozostaje sprawdzić warunki interpolacyjne. Dla $0 \leqslant j \leqslant 2n-1$ jest

$$P(x_j) = \tfrac{1}{2}[1+E_n(x_j)]p(x_j) + \tfrac{1}{2}[1-E_n(x_j)]q(x_j-\pi/n).$$

Zauważmy, że $E_n(x_j) = e^{\pi inj/n} = e^{\pi ij} = (-1)^j$. Dlatego dla j parzystych jest $P(x_j) = p(x_j) = f(x_j)$, a dla j nieparzystych

$$P(x_j) = q(x_j-\pi/n) = q(x_{j-1}) = f(x_j). \qquad \blacksquare$$

Twierdzenie 6.13.2. *Niech będzie, dla wielomianów z tw. 6.13.1,*

$$p = \sum_{j=0}^{n-1}\alpha_j E_j, \quad q = \sum_{j=0}^{n-1}\beta_j E_j, \quad P = \sum_{j=0}^{2n-1}\gamma_j E_j.$$

Wtedy dla $0 \leqslant j \leqslant n-1$ *jest*

$$\gamma_j = \tfrac{1}{2}\alpha_j + \tfrac{1}{2}e^{-\pi ij/n}\beta_j, \quad \gamma_{n+j} = \tfrac{1}{2}\alpha_j - \tfrac{1}{2}e^{-\pi ij/n}\beta_j. \tag{6.13.2}$$

Dowód. Ponieważ

$$q\left(x-\frac{\pi}{n}\right) = \sum_{j=0}^{n-1}\beta_j E_j\left(x-\frac{\pi}{n}\right) = \sum_{j=0}^{n-1}\beta_j e^{ij(x-\pi/n)} =$$
$$= \sum_{j=0}^{n-1}\beta_j e^{-\pi ij/n}E_j(x),$$

więc z (6.13.1) wynika, że

$$P(x) = \frac{1}{2}[1+E_n(x)]p(x) + \frac{1}{2}[1-E_n(x)]q\left(x-\frac{\pi}{n}\right),$$
$$P = \frac{1}{2}\sum_{j=0}^{n-1}[(1+E_n)\alpha_j E_j + (1-E_n)\beta_j e^{-\pi ij/n}E_j] =$$
$$= \frac{1}{2}\sum_{j=0}^{n-1}[(\alpha_j+\beta_j e^{-\pi ij/n})E_j + (\alpha_j-\beta_j e^{-\pi ij/n})E_{n+j}].$$

To daje wzory na współczynniki γ_j. ■

Przykład 6.13.3. Zastosować tw. 6.13.2 dla $n = 1$.

Rozwiązanie. Z (6.13.2) wynika, że $\gamma_0 = \frac{1}{2}(\alpha_0+\beta_0)$ i $\gamma_1 = \frac{1}{2}(\alpha_0-\beta_0)$. Dlatego $P = \gamma_0 E_0 + \gamma_1 E_1 = \frac{1}{2}(\alpha_0+\beta_0) + \frac{1}{2}(\alpha_0-\beta_0)E_1$. Podstawiając tu wartości $\alpha_0 = f(0)$ i $\beta_0 = f(\pi)$, otrzymujemy taki sam wynik, jak w przykł. 6.12.5. ■

Niech $R(n)$ oznacza minimalną liczbę mnożeń konieczną do obliczenia współczynników wielomianu wykładniczego, który interpoluje daną funkcję w węzłach $2\pi j/n$ dla $0 \leqslant j \leqslant n-1$.

Twierdzenie 6.13.4. *Funkcja* R *spełnia nierówność* $R(2^m) \leqslant 2^m m$.

Dowód. Udowodnimy najpierw nierówność

$$R(2n) \leqslant 2R(n) + 2n. \tag{6.13.3}$$

Na mocy tw. 6.13.2 współczynniki γ_j wielomianu P wynikają z α_j i β_j kosztem $2n$ mnożeń. Mamy tu na myśli mnożenie każdego α_j przez $\frac{1}{2}$ i każdego β_j przez $\frac{1}{2}e^{-\pi ij/n}$ (zakładamy, że te wielkości są znane). Natomiast obliczenie każdego z układów $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$ i $\beta_0, \beta_1, \ldots, \beta_{n-1}$ wymaga $R(n)$ mnożeń, a to już daje (6.13.3).

Reszta dowodu wynika przez iterowanie tej nierówności. Uwzględniamy również to, że dla $n = 1$ mnożenia nie są potrzebne, czyli $R(1) = 0$. Stąd $R(2) \leqslant 2$, $R(4) \leqslant 8$ itd. ■

Nierówność z tw. 6.13.4 można też, dla N będącego potęgą liczby 2, napisać w postaci $R(N) \leqslant N \log_2 N$. Algorytm *szybkiego przekształcenia Fouriera*, który polega na wielokrotnym stosowaniu tw. 6.13.2, wymaga wykonania właśnie $N \log_2 N$ mnożeń.

Sens tw. 6.13.1 można wyrazić w zwarty sposób, wprowadzając dwa operatory liniowe, L_n i T_h. Dla każdej funkcji f wartość $L_n f$ jest z definicji równa wielomianowi wykładniczemu stopnia $\leqslant n-1$ interpolującemu f w węzłach $2\pi j/n$ dla $0 \leqslant j \leqslant n-1$. Natomiast T_h jest *operatorem przesunięcia*:

$$(T_h f)(x) := f(x+h).$$

W twierdzeniu 6.13.1 jest $P = L_{2n} f$, $p = L_n f$ i $q = L_n T_{\pi/n} F$. Teza tego twierdzenia daje więc tożsamość operatorową

$$L_{2n} = \tfrac{1}{2}(1 + E_n) L_n + \tfrac{1}{2}(1 - E_n) T_{-\pi/n} L_n T_{\pi/n}. \tag{6.13.4}$$

Wprowadzamy jeszcze jeden symbol: niech dla danej funkcji f i ustalonego $N := 2^m$ będzie

$$P_k^{(n)} := L_{2^n} T_{2k\pi/N} f \qquad (0 \leqslant n \leqslant m, \quad 0 \leqslant k \leqslant 2^{m-n} - 1).$$

Jest to zatem wielomian wykładniczy stopnia $2^n - 1$, interpolujący funkcję f z przesuniętym argumentem:

$$P_k^{(n)}\left(\frac{2\pi j}{2^n}\right) = f\left(\frac{2\pi k}{N} + \frac{2\pi j}{2^n}\right) \quad (0 \leqslant j \leqslant 2^n - 1). \tag{6.13.5}$$

Tematem zad. 3 jest sprawdzenie, że zbiory argumentów funkcji f w (6.13.5) są dla różnych k rozłączne.

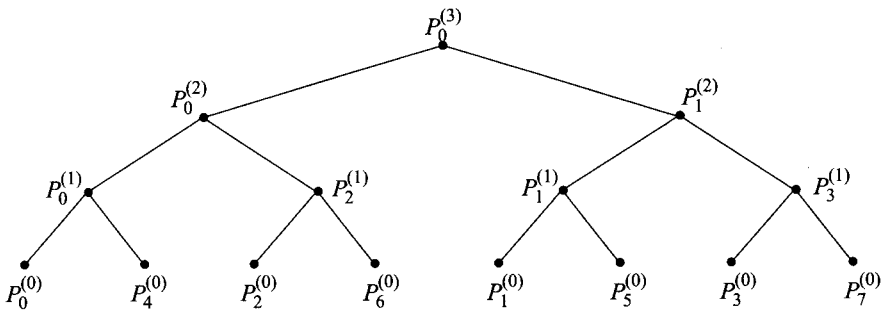
Wniosek 6.13.5. *Wielomiany $P_k^{(n)}$ spełniają tożsamość*

$$P_k^{(n+1)}(x) = \frac{1}{2}(1 + e^{2^n ix}) P_k^{(n)}(x) + \frac{1}{2}(1 - e^{2^n ix}) P_{k+2^{m-n-1}}^{(n)}\left(x - \frac{\pi}{2^n}\right). \tag{6.13.6}$$

Aby ją udowodnić, wystarczy złożyć obie strony (6.13.4) z operatorem $T_{2\pi k/N}$ i uwzględnić definicję wielomianów $P_k^{(n)}$.

Celem obliczeń jest znalezienie wielomianu interpolacyjnego L_N, czyli $P_0^{(m)}$. Współczynniki jego rozwinięcia względem $E_0, E_1, \ldots$ oblicza się za pomocą wzorów rekurencyjnych, które wynikają z (6.13.6). Obliczenia zaczynamy od wielomianów $P_k^{(0)}$ $(0 \leqslant k \leqslant 2^m - 1)$ stopnia zerowego. Wobec (6.13.5) jest $P_k^{(0)} \equiv f(2\pi k/N)$.

Drzewo na rys. 6.23 pokazuje, że dla $m = 3$ szukany wielomian $P_0^{(3)}$ wyraża się przez wielomiany $P_0^{(2)}$ i $P_1^{(2)}$, każdy z nich – przez dwa wielomiany niższego stopnia itd.



RYS. 6.23. Drzewo zależności wielomianów $P_k^{(n)}$

Algorytm

Niech będzie

$$P_k^{(n)}(x) = \sum_{j=0}^{2^n-1} A_{kj}^{(n)} e^{ijx},$$

gdzie $0 \leqslant n \leqslant m$, $0 \leqslant k \leqslant 2^{m-n} - 1$. Rozumując jak w dowodzie tw. 6.13.2, wnioskujemy z (6.13.6), że

$$A_{kj}^{n+1} = \tfrac{1}{2}\left[A_{kj}^{(n)} + e^{-\pi ij/2^n} A_{k+2^{m-n-1},j}^{(n)}\right],$$
$$A_{k,j+2^n}^{n+1} = \tfrac{1}{2}\left[A_{kj}^{(n)} - e^{-\pi ij/2^n} A_{k+2^{m-n-1},j}^{(n)}\right].$$

Dla każdego n wielkości $A_{kj}^{(n)}$ zapamiętujemy w tablicy N zmiennych, gdyż $0 \leqslant k \leqslant 2^{m-n} - 1$ i $0 \leqslant j \leqslant 2^n - 1$. Ściślej, przyjmujemy, że w pewnym momencie obliczeń wielkości $A_{kj}^{(n)}$ znajdują się w tablicy C, przy czym

$$C(2^n k + j) = A_{kj}^{(n)} \quad (0 \leqslant k \leqslant 2^{m-n} - 1,\ 0 \leqslant j \leqslant 2^n - 1),$$

a wynikające z nich $A_{kj}^{(n+1)}$ umieszczamy w tablicy D tak, że

$$D(2^{n+1}k+j) = A_{kj}^{(n+1)} \quad (0 \leqslant k \leqslant 2^{m-n-1}-1,\ 0 \leqslant j \leqslant 2^{n+1}-1).$$

Na początku oblicza się wielkości $Z(j) := e^{-2\pi ij/N}$. Później korzysta się z wynikającej stąd równości $e^{-\pi ij/2^n} = Z(2^{m-n-1}j)$. Kompletny algorytm szybkiego przekształcenia Fouriera jest następujący:

input m
$N \leftarrow 2^m$
$w \leftarrow e^{-2\pi i/N}$
$Z(0) \leftarrow 1$; $C(0) \leftarrow f(0)$
for $k = 1$ **to** $N-1$ **do**
    $Z(k) \leftarrow Z(k-1)w$
    $C(k) \leftarrow f(2\pi k/N)$
end do
for $n = 0$ **to** $m-1$ **do**
    for $k = 0$ **to** 2^{m-n-1} **do**
        for $j = 0$ **to** $2^n - 1$ **do**
            $u \leftarrow C(2^n k + j)$
            $v \leftarrow Z(2^{m-n-1}j)C(2^n k + 2^{m-1} + j)$
            $D(2^{n+1}k + j) \leftarrow (u+v)/2$
            $D(2^{n+1}k + j + 2^n) \leftarrow (u-v)/2$
        end do
    end do
    for $j = 0$ **to** $N-1$ **do**
        $C(j) \leftarrow D(j)$
    end do
end do
output $C(0), C(1), \ldots, C(N-1)$

Liczba mnożeń wymaganych w tym algorytmie wynosi $2^m m$; por. tw. 6.13.4. Jeśli jednak nie liczyć dzieleń przez 2, czyli mnożeń przez $\frac{1}{2}$ (w komputerze dwójkowym takie działanie sprowadza się do zmniejszenia o 1 cechy liczby zmiennopozycyjnej), to mnożeń mamy dwukrotnie mniej.

Przykład 6.13.6. Odtworzyć funkcję $f := \sum_{k=0}^{7}(k+1)E_k$ za pomocą szybkiego przekształcenia Fouriera, używając wartości $f(x)$ w ośmiu równoodległych punktach.

Rozwiązanie. Powyższy algorytm zastosowano dla $m = 3$ w komputerze z 32-bitowymi liczbami zmiennopozycyjnymi. Początkowe wartości zmiennych z tablicy C (zaokrąglone do 5 cyfr znaczących) podano niżej:

$$\begin{aligned} C(0) &= 36.000, & C(4) &= -4.0000, \\ C(1) &= -4.0000 - 9.6569i, & C(5) &= -4.0000 + 1.6568i, \\ C(2) &= -4.0000 - 4.0000i, & C(6) &= -4.0000 + 4.0000i, \\ C(3) &= -4.0000 - 1.6569i, & C(7) &= -4.0000 + 9.6569i. \end{aligned}$$

a wartości końcowe z tą samą dokładnością są równe $1, 2, \ldots, 8$. ■

Trzeba podkreślić, że podany algorytm wymaga w praktyce wielu ulepszeń. Oto dwa przykłady: dodatkowe instrukcje usuwają potrzebę użycia tablicy D, a iloczyn $Z(j)C(k)$ dla takich j, że $Z(j) = \pm 1$, należy obliczać jako $\pm C(k)$. Poza tym bardziej uniwersalny program powinien uwzględniać dowolne N, a nie tylko potęgi liczby 2.

Specjalne wersje programu warto opracować dla funkcji f rzeczywistych, a inne – dla rozwinięć tylko względem sinusów bądź cosinusów.

Funkcje nieodróżnialne. Częstotliwość Nyquista

Jest oczywiste, że za pomocą skończenie wielu wartości funkcji f nie można odtworzyć wszystkich informacji o niej. Aby zbadać to dokładniej, rozważmy rozwinięcie funkcji w szereg Fouriera:

$$f = \sum_{k=-\infty}^{\infty} \langle f, E_k \rangle E_k \tag{6.13.7}$$

(tu i w (6.13.8) stosujemy oznaczenia z podrozdz. 6.12). Natomiast wielomian wykładniczy stopnia $\leqslant N-1$ interpolujący f wyraża się wzorem

$$P = \sum_{k=0}^{N-1} \langle f, E_k \rangle_N E_k \tag{6.13.8}$$

(zob. wniosek 6.12.4). Porównajmy współczynniki obu rozwinięć:

$$\langle f, E_m \rangle_N = \Big\langle \sum_{k=-\infty}^{\infty} \langle f, E_k \rangle E_k, E_m \Big\rangle_N = \sum_{k=-\infty}^{\infty} \langle f, E_k \rangle \langle E_k, E_m \rangle_N .$$

Na mocy tw. 6.12.2 pseudoiloczyn $\langle E_k, E_m \rangle_N$ nie znika (i jest równy 1) tylko wtedy, gdy $k - m$ jest wielokrotnością liczby N. Ograniczamy się więc do wartości $k - m = \nu N$, gdzie $\nu = 0, \pm 1, \pm 2, \ldots$:

$$\langle f, E_m \rangle_N = \sum_{\nu=-\infty}^{\infty} \langle f, E_{m+\nu N} \rangle .$$

Uporządkujmy składniki tego szeregu według ich malejącego znaczenia:

$$\langle f, E_m\rangle_N = \langle f, E_m\rangle + \langle f, E_{m+N}\rangle + \langle f, E_{m-N}\rangle + \dots$$

Tak więc współczynnik przy E_m w wielomianie interpolacyjnym P zawiera nie tylko oczekiwany składnik $\langle f, E_m\rangle$ z (6.13.7), ale i składniki $\langle f, E_{m+\nu N}\rangle$ odpowiadające wyższym częstotliwościom. Są to składniki niepożądane, które zniekształcają odtwarzany sygnał. Pierwszym ze składników obecnych w szeregu dla f, ale nieobecnych w P jest $\langle f, E_N\rangle$. Odpowiada on tzw. *częstotliwości Nyquista.*

Obliczanie wartości wielomianu wykładniczego

Szybkie przekształcenie Fouriera stosujemy również, obliczając wartości wielomianu wykładniczego

$$p := \sum_{j=0}^{n-1} a_j E_j$$

na zbiorze punktów równoodległych

$$t - \frac{2k\pi}{n} \quad (0 \leqslant k \leqslant n-1).$$

Niech będzie $x_k := 2k\pi/n$. Wtedy

$$p(t-x_k) = \sum_{j=0}^{n-1} a_j E_j(t-x_k) = \sum_{j=0}^{n-1} a_j e^{ij(t-x_k)} =$$
$$= \sum_{j=0}^{n-1} a_j e^{ijt}\overline{E_k(x_j)} = n\langle g, E_k\rangle_n,$$

gdzie g jest funkcją taką, że

$$g(x_j) = a_j e^{ijt} \quad (0 \leqslant j \leqslant n-1).$$

Zastosowanie szybkiego przekształcenia Fouriera do tej funkcji daje współczynniki $\langle g, E_k\rangle_n$, a ich iloczyny przez n są szukanymi wartościami wielomianu p.

Źródła informacji o szybkim przekształceniu Fouriera: Davis i Rabinowitz [1984], Cooley, Lewis i Welch [1967], Bloomfield [1976], Briggs i Henson [1995], Brigham [1974], Conte i de Boor [1980], Kahaner [1970, 1978], Lanczos [1966], Kahaner, Moler i Nash [1989], Elliott i Rao [1982], Nussbaumer [1982] i Scheid [1988].

ZADANIA 6.13

1. Wykazać, że funkcje f i $f + \lambda E_k$ są nierozróżnialne w punktach $2\pi j/N$, jeśli k jest wielokrotnością liczby N.
2. Wykazać, że funkcje $f_\pm(x) := \cos[(n \mp \alpha)\pi x \pm \beta]$ są nierozróżnialne dla x całkowitych.
3. Wykazać, że w (6.13.5) zbiory punktów $\frac{2\pi k}{N} + \frac{2\pi j}{2^n}$ $(0 \leqslant j \leqslant 2^n - 1)$ dla różnych k są rozłączne.
4. Sprawdzić, że tw. 6.13.2 ma następującą równoważną postać:

   $$\langle f, E_j\rangle_{2n} = \tfrac{1}{2}(u_j + v_j), \quad \langle f, E_{j+n}\rangle_{2n} = \tfrac{1}{2}(u_j - v_j),$$

   gdzie $u_j := \langle f, E_j\rangle_n$, $v_j := e^{-\pi ij/n}\langle T_{\pi/n} f, E_j\rangle_n$.
5. Dla danych n i f niech u i v będą wektorami kolumnowymi o składowych

   $$u_j := f\left(\frac{2\pi j}{n}\right), \quad v_j := \langle f, E_j\rangle_n \quad (0 \leqslant j \leqslant n-1).$$

   Udowodnić, że $v = Au$, gdzie $(A)_{jk} = n^{-1}e^{-2\pi ijk/n}$ $(0 \leqslant j, k \leqslant n-1)$.
6. (cd.). Wykazać, że A jest macierzą symetryczną, która ma tylko n różnych elementów i że $n^{1/2}A$ jest macierzą unitarną.
7. (cd.). Wykazać, że $\|v\|_2 = n^{-1/2}\|u\|_2$.

ZADANIA KOMPUTEROWE 6.13

K1. Napisać i sprawdzić program obliczający za pomocą szybkiego przekształcenia Fouriera wartości wielomianu $p = \sum_{j=0}^{N-1} a_j E_j$ o danych współczynnikach a_j. Przyjąć, że $N = 2^m$.

6.14. Metody adaptacyjne

Wyróżniającą cechą *metod adaptacyjnych* aproksymacji jest to, że obszar, w którym funkcję przybliżamy, jest dzielony wielokrotnie na mniejsze podobszary aż do osiągnięcia żądanej dokładności. Ostateczne „globalne" przybliżenie powstaje zatem z połączenia przybliżeń lokalnych. Jest oczywiste, że funkcje sklejane ze swobodnie wybieranymi węzłami są dobrym narzędziem takiej aproksymacji.

Funkcje sklejane stopnia pierwszego

Naszkicowany wyżej pomysł wyjaśnimy na przykładzie algorytmu tworzącego dla danej funkcji ciągłej f w ustalonym przedziale $[a, b]$ funkcję sklejaną S stopnia pierwszego. Ustalamy też maksymalną wartość $\varepsilon > 0$ błędu bezwzględnego; ma zatem być

$$|f(x) - S(x)| \leqslant \varepsilon \quad (a \leqslant x \leqslant b). \tag{6.14.1}$$

Funkcja sklejana S jest przedziałami liniowa i interpoluje f w pewnych punktach. Wiadomo, że funkcja liniowa interpolująca f w punktach α i β wyraża się wzorem

$$l(\alpha,\beta;x) := \frac{f(\alpha)(\beta-x)+f(\beta)(x-\alpha)}{\beta-\alpha}. \tag{6.14.2}$$

Danemu układowi węzłów t_i takich, że

$$a = t_0 < t_1 < \ldots < t_{n-1} < t_n = b,$$

odpowiada funkcja sklejana S, która w każdym z przedziałów $[t_{i-1},t_i]$ jest identyczna z $l(t_{i-1},t_i;x)$. Jeśli spełnia ona warunek (6.14.1), to nasz cel jest osiągnięty. Jeśli natomiast

$$\|f-S\|_\infty := \max_{x\in[a,b]} |f(x)-S(x)| > \varepsilon,$$

to tamten warunek nie jest spełniony co najmniej w jednym punkcie przedziału. Niech $|f(x)-S(x)|$ będzie największe w punkcie ξ. Jeśli $\xi \in (t_{i-1},t_i)$, to do układu dotychczasowych węzłów dołączamy punkt ξ, który dzieli przedział $[t_{i-1},t_i]$ na dwie części. Stosownie do tego zamiast jednej funkcji liniowej $l(t_{i-1},t_i;x)$ będziemy teraz mieli dwie funkcje: $l(t_{i-1},\xi;x)$ w $[t_{i-1},\xi]$ i $l(\xi,t_i;x)$ w $[\xi,t_i]$. Powtarzając te podziały, dochodzimy ostatecznie do funkcji sklejanej przybliżającej f dostatecznie dokładnie.

Algorytm

Pisząc algorytm tego postępowania, kładziemy główny nacisk na jego efektywność. Dlatego przyjmujemy, że w każdym etapie mamy cztery tablice: tablicę węzłów $t = \{t_0,t_1,\ldots,t_n\}$, tablicę $y = \{y_0,y_1,\ldots,y_n\}$ wartości funkcji f w tych węzłach, tablicę $d = \{y_1,y_2,\ldots,y_n\}$ dokładności przybliżenia w podprzedziałach ($d_i = \max_{x\in[t_{i-1},t_i]}|f(x)-l(t_{i-1},t_i;x)|$) i tablicę $c = \{c_1,c_2,\ldots,c_n\}$ punktów, w których te maksymalne odchylenia są osiągnięte.

Znając tablice t, y, d, c, sprawdzamy, czy dany układ węzłów daje dostatecznie dokładną funkcję S, tj. czy $\max_i d_i \leqslant \varepsilon$. Jeśli jeszcze tak nie jest, to znajdujemy maksymalne d_i i dołączamy punkt c_i do tablicy węzłów. Wymaga to przesunięcia o jedną pozycję elementów o wskaźnikach $i, i+1,\ldots,n$ we wszystkich czterech tablicach. Obliczenia zaczynamy oczywiście od $n=1$, $t_0=a$ i $t_1=b$. W algorytmie korzystamy z procedury `Max`, której wywołanie `Max`(f,α,β,c,d) podstawia pod d maksimum wyrażenia $|f(x)-l(\alpha,\beta;x)|$ w $[\alpha,\beta]$, a pod c – punkt, gdzie to maksimum jest osiągnięte.

input a, b, ε, M
$t_0 \leftarrow a$; $t_1 \leftarrow b$; $y_0 \leftarrow f(t_0)$; $y_1 \leftarrow f(t_1)$
call Max(f, t_0, t_1, c_1, d_1)
for $n = 1$ **to** $M - 1$ **do**
    wybór i takiego, że $d_i = \max\{d_1, d_2, \ldots, d_n\}$
    if $d_i \leqslant \varepsilon$ **exit loop**
    for $j = n$ **to** $i + 1$ **step** -1 **do**
        $t_{j+1} \leftarrow t_j$; $y_{j+1} \leftarrow y_j$; $d_{j+1} \leftarrow d_j$; $c_{j+1} \leftarrow c_j$
    end do
    $t_{i+1} \leftarrow t_i$; $y_{i+1} \leftarrow y_i$; $t_i \leftarrow c_i$; $y_i \leftarrow f(c_i)$
    call Max$(f, t_{i-1}, t_i, c_i, d_i)$
    call Max$(f, t_i, t_{i+1}, c_{i+1}, d_{i+1})$
end do
output $n, (t_i), (y_i), (d_i)$

Wykonanie algorytmu kończy się, gdy wszystkie odchylenia d_i nie przewyższają ε, ale i wtedy, gdy liczba kroków osiągnie dane M.

Wielkości c i d, które oblicza procedura Max, nie muszą być wyznaczane z dużą dokładnością. Wystarczy zbadać wartości $|f(x) - l(\alpha, \beta; x)|$ w punktach podziału przedziału $[\alpha, \beta]$ np. na 10 równych części:

procedure Max(f, α, β, c, d)
$r \leftarrow f(\alpha)$; $s \leftarrow f(\beta)$
$k \leftarrow 10$; $h \leftarrow (\beta - \alpha)/k$
$d \leftarrow 0$
for $i = 1$ **to** $k - 1$ **do**
    $x \leftarrow \alpha + ih$; $z \leftarrow f(x)$
    $z \leftarrow |z - (is + (k - i)r)/k|$
    if $z > d$ **then**
        $d \leftarrow z$; $c \leftarrow x$
    end if
end do
return

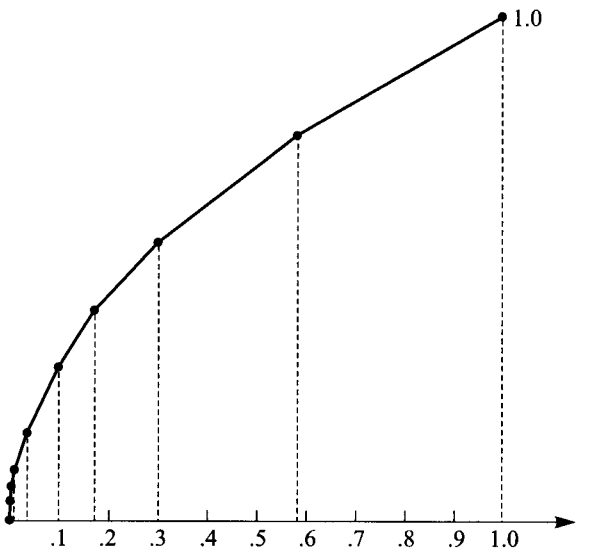
Te wartości są obliczane w siódmym wierszu procedury. Istotnie, z (6.14.2) wynika, że w oznaczeniach tu przyjętych jest $l(\alpha, \beta; \alpha + ih) = [is + (k - i)r]/k$.

Przykład 6.14.1. Opisaną wyżej metodę adaptacyjną zastosować do funkcji $f(x) := \sqrt{x}$ w przedziale $[0, 1]$ dla $\varepsilon = 10^{-2}$.

Rozwiązanie. Wyniki obliczeń są takie jak w poniższej tablicy, gdzie

$$d_i := \max_{x \in [t_{i-1}, t_i]} |f(x) - S(x)|.$$

Dokładność ε zapewnia więc funkcja sklejana S (rys. 6.24) o 10 węzłach skupionych w pobliżu punktu 0, gdzie f ma pochodną nieskończoną. Dla



RYS. 6.24. Funkcja sklejana S stopnia 1 otrzymana metodą adaptacyjną

maksymalnego błędu $\varepsilon = 10^{-3}$ ta sama metoda dałaby funkcję sklejaną o 32 węzłach.

t_i	$f(t_i)$	d_i
0	0	
0.000729	0.027	0.007
0.00243	0.049	0.002
0.0081	0.09	0.003
0.027	0.16	0.005
0.09	0.3	0.01
0.174	0.417	0.005
0.3	0.548	0.004
0.58	0.762	0.009
1	1	0.008

■

Pomysł zastosowany w aproksymacji za pomocą funkcji sklejanych stopnia pierwszego można uogólnić, zastępując funkcję liniową $l(\alpha, \beta; x)$ inną funkcją $A(\alpha, \beta; x)$, która – jak i tamta – aproksymuje daną funkcję f w przedziale $[\alpha, \beta]$. Wynikające stąd przybliżenie globalne, w całym przedziale $[a, b]$, nie musi być tam ciągłe. Stosuje się też inne zasady wyboru nowych punktów podziału; jeśli np. przybliżenie $A(\alpha, \beta; x)$ nie jest dostatecznie dokładne, to takim punktem może być $\frac{1}{2}(\alpha + \beta)$.

ZADANIA 6.14

1. Niech metoda adaptacyjna daje funkcję sklejaną stopnia zerowego (przedziałami stałą). Jak można oszacować z dołu konieczną liczbę podprzedziałów, jeśli funkcja f jest rosnąca? (Odpowiedź zależy od $f(b) - f(a)$ i ε).
2. Sprawdzić, że dla $f(x) := \sqrt{x}$ i $0 \leqslant \alpha < \beta$ funkcja $|f(x) - l(\alpha, \beta; x)|$ osiąga maksimum w punkcie $x := \frac{1}{4}(\sqrt{\alpha} + \sqrt{\beta})^2$.
3. Jakie są skutki zastosowania metody adaptacyjnej (dokładnie w wersji podanej w tekście) do funkcji $f(x) := \sin 10x$ w przedziale $[0, \pi]$?

ZADANIA KOMPUTEROWE 6.14

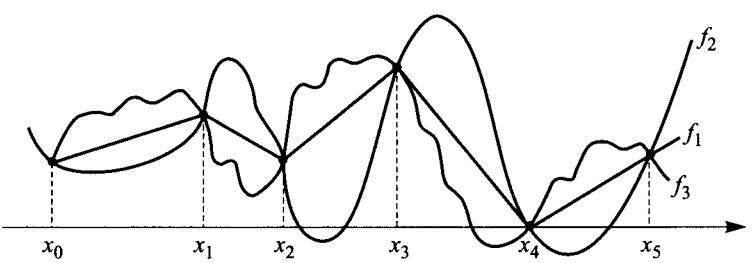
K1. Zaprogramować i sprawdzić metodę adaptacyjną dającą funkcję sklejaną: (**a**) stopnia zerowego, (**b**) stopnia pierwszego (opisaną w tekście), (**c**) stopnia trzeciego.

ROZDZIAŁ 7

Różniczkowanie i całkowanie numeryczne

7.1. Różniczkowanie numeryczne i ekstrapolacja Richardsona

Czy wartości funkcji f dane w wielu punktach, np. $x_0, x_1 \ldots, x_n$, mogą dać przybliżenia pochodnej $f'(c)$ albo całki $\int_a^b f(x)\,dx$? To pytanie, jak zobaczymy, wymaga uściślenia.



RYS. 7.1. Trzy funkcje ciągłe o wspólnych wartościach w sześciu punktach

Zauważmy najpierw, że same wartości $f(x_0), f(x_1), \ldots, f(x_n)$ mówią o f niewiele. Jeśli dodatkowo wiemy tylko, że jest to funkcja rzeczywista i ciągła, to – jak pokazuje rys. 7.1 – te wartości są niemal bezużyteczne.

Jeśli natomiast wiemy, że f jest wielomianem stopnia co najwyżej n, to $n+1$ wartości tej funkcji określa ją jednoznacznie (zob. podrozdz. 6.1). Wtedy wielkości $f'(c)$ i $\int_a^b f(x)\,dx$ możemy obliczyć dokładnie. Między tymi skrajnymi przypadkami znajdują się takie, z którymi mamy do czynienia w praktyce: informacje o funkcji f nie określają jej całkowicie, ale pozwalają oszacować błąd obliczonych przybliżonych wartości pochodnej lub całki.

Różniczkowanie numeryczne

Aby uzasadnić słuszność tych uwag, rozważmy wzór różniczkowania numerycznego wynikający wprost z definicji pochodnej:

$$f'(x) \approx \frac{1}{h}[f(x+h) - f(x)].$$

Dla funkcji liniowej ($f(x) = ax + b$) ten wzór jest dokładny dla każdego $h \neq 0$. Dla innych funkcji jest tak tylko wyjątkowo. Oszacujmy więc błąd tego przybliżenia. Pozwala na to wzór Taylora. Jeśli mianowicie pierwsza pochodna f' jest ciągła w przedziale domkniętym o końcach x i $x+h$, a druga pochodna f'' istnieje w przedziale otwartym o tychże końcach, to

$$f(x+h) = f(x) + hf'(x) + \frac{1}{2}h^2 f''(\xi),$$

gdzie ξ leży w tym ostatnim przedziale. Wynika stąd, że

$$f'(x) = \frac{1}{h}[f(x+h) - f(x)] - \frac{1}{2}hf''(\xi). \tag{7.1.1}$$

Pierwszy składnik prawej strony tej równości, tj. $[f(x+h) - f(x)]/h$, jest przybliżoną wartością pochodnej, a drugi, czyli $-\frac{1}{2}hf''(\xi)$, jest *błędem przybliżenia*, który można oszacować z góry, jeśli tylko funkcja należy do określonej wyżej klasy. Zauważmy, że błąd zawiera czynnik h, a więc dąży do 0 wraz z h. Tak jest i w wielu innych wzorach przybliżonych z tym jednak, że taki czynnik bywa wyższą potęgą parametru h. Pozostałe czynniki błędu określają, mówiąc z grubsza, jak regularna musi być funkcja f, aby całe wyrażenie miało sens.

Przykład 7.1.1. Stosując wzór (7.1.1) dla $h = 0.01$, znaleźć przybliżoną wartość pochodnej funkcji $f(x) := \cos x$ w punkcie $\pi/4$ i oszacować błąd tego przybliżenia.

Rozwiązanie. Wartością przybliżoną jest

$$\frac{1}{0.01}(0.70000\,0476 - 0.70710\,6781) = -0.71063\,051,$$

a jej błąd jest równy

$$|\tfrac{1}{2}hf''(\xi)| = 0.005|\cos\xi| \leqslant 0.005.$$

Ponieważ jednak $\xi \in (\pi/4, \pi/4 + h)$, więc $|\cos\xi| < 0.70710\,7$, co daje lepsze oszacowanie błędu, $0.00353\,55$. Prawdziwy błąd jest niewiele mniejszy:

$$-\sin(\pi/4) + 0.71063\,051 = 0.00352\,3729. \qquad \blacksquare$$

Składnik $-(h/2)f''(\xi)$ w (7.1.1) nazywamy *błędem obcięcia*; odpowiednia wartość przybliżona powstaje przez obcięcie szeregu Taylora do początkowych składników.

Na pierwszy rzut oka wzór (7.1.1) daje tym dokładniejszą wartość $f'(x)$, im mniejsze jest h. Sprawdźmy to przypuszczenie dla funkcji $f(x) := \operatorname{arctg} x$ i punktu $x := \sqrt{2}$. Ponieważ $f'(x) = (x^2+1)^{-1}$, więc $f'(\sqrt{2}) = \frac{1}{3}$. Oto kilka przybliżeń tej wartości otrzymanych dla różnych h na komputerze, w którym liczby zmiennopozycyjne są 32-bitowe:

h	$f(x+h)$	$f(x+h) - f(x)$	$[f(x+h) - f(x)]/h$
2^{-4}	0.97555 095	0.02023 435	0.32374 954
2^{-12}	0.95539 796	0.00008 136	0.33325 195
2^{-20}	0.95531 690	0.00000 030	0.31250 000
2^{-24}	0.95531 666	0.00000 006	1.00000 000
2^{-26}	0.95531 660	0.00000 000	0.00000 000

(we wszystkich różnicach $f(x+h) - f(x)$ jest $f(x) = \operatorname{arctg}\sqrt{2} = 0.95531\,660$).

Odejmowanie bliskich wielkości powoduje, że różnica $f(x+h) - f(x)$ ma tym mniej cyfr znaczących, im mniejsze jest h, a to pogarsza dokładność przybliżonej wartości pochodnej. Najlepszą (trzy cyfry dokładne) otrzymujemy dla $h = 2^{-12}$, gdy $f(x+h) - f(x)$ ma cztery cyfry znaczące. Tak więc redukcja liczby cyfr znaczących przy odejmowaniu uniemożliwia uzyskanie dobrych przybliżeń dla $f'(x)$ przy małych h. Wykonując obliczenia z wielokrotną dokładnością, możemy oczywiście otrzymać dokładniejsze przybliżenia, ale i wtedy wystąpi, chociaż dopiero dla mniejszych h, zasygnalizowane już zjawisko.

Wzory różniczkowania numerycznego stosujemy przede wszystkim rozwiązując numerycznie równania różniczkowe. Występujące w nich pochodne przybliżamy kombinacjami wartości funkcji, jak w (7.1.1). Dokładność przybliżeń oceniamy często według wykładnika p w czynniku h^p błędu; im większe jest p, tym lepiej. Wspomniany wzór jest dość kiepski, bo $p = 1$. Lepszy pod tym względem jest wzór przybliżony

$$f'(x) \approx \frac{1}{2h}[f(x+h) - f(x-h)].$$

Istotnie, ponieważ

$$f(x \pm h) = f(x) \pm hf'(x) + \frac{1}{2}h^2 f''(x) \pm \frac{1}{3!}h^3 f'''(\xi_\pm),$$

więc odejmując stronami dwa warianty tej równości, otrzymujemy wzór

$$f'(x) = \frac{1}{2h}[f(x+h) - f(x-h)] - \frac{1}{12}h^2[f'''(\xi_-) + f'''(\xi_+)],$$

w którym błąd zawiera czynnik h^2. Zauważmy jednak, że ten wzór możemy stosować tylko wtedy, gdy trzecia pochodna funkcji f istnieje.

Błąd w powyższym wzorze można uprościć przy założeniu, że trzecia pochodna f''' jest ciągła w przedziale $(x-h, x+h)$. Wtedy istnieje tam takie ξ, że $f'''(\xi) = \frac{1}{2}[f'''(\xi_-) + f'''(\xi_+)]$ i wyrażenie dla $f'(x)$ upraszcza się:

$$f'(x) = \frac{1}{2h}[f(x+h) - f(x-h)] - \frac{1}{6}h^2 f'''(\xi). \tag{7.1.2}$$

W podobny sposób, ale za pomocą wzoru Taylora z większą liczbą składników, otrzymujemy ważny wzór, dający przybliżenia drugiej pochodnej:

$$f''(x) = \frac{1}{h^2}[f(x+h) - 2f(x) + f(x-h)] - \frac{1}{12}h^2 f^{(4)}(\xi), \tag{7.1.3}$$

gdzie $\xi \in (x-h, x+h)$. Jest on często stosowany w rozwiązywaniu numerycznym równań różniczkowych rzędu drugiego.

Przykład 7.1.2. Powtórzyć obliczenia z przykł. 7.1.1, ale stosując wzór (7.1.2).

Rozwiązanie. Wybrane wyniki obliczeń podano w tabeli, w której nagłówku $D(h) := f(x+h) - f(x-h)$; ostatnia kolumna zawiera przybliżenia $D(h)/(2h)$ pochodnej:

h	$f(x+h)$	$f(x-h)$	$D(h)$	$D(h)/(2h)$
2^{-2}	1.02972 674	0.86112 982	0.16859 692	0.33719 385
2^{-10}	0.95564 199	0.95499 092	0.00065 106	0.33334 351
2^{-18}	0.95531 786	0.95531 535	0.00000 250	0.32812 500
2^{-26}	0.95531 660	0.95531 660	0.00000 000	0.00000 000

Jak widać, komentarze z przykł. 7.1.1 odnoszą się i do zastosowanego tu wzoru, choć najdokładniejsze przybliżenie wartości pochodnej jest teraz lepsze. Nie zawsze jednak tak jest, bo dokładność przybliżeń wynikających z (7.1.1) i (7.1.2) zależy także od wielkości pochodnych wyższych rzędów, które tam występują. ■

Różniczkowanie numeryczne funkcji empirycznych jest bardzo ryzykowne i lepiej go unikać, bo ta procedura potęguje skutki błędów pomiarów. Istotnie, gdy np. stosujemy wzór (7.1.2), to błąd wartości $f(x \pm h)$ jest mnożony przez wielkość $1/(2h)$ – dużą, gdy h jest małe. Warto podkreślić, że w całkowaniu numerycznym takiego niebezpieczeństwa nie ma.

Zastosowanie interpolacji wielomianowej

Ogólna metoda różniczkowania i całkowania numerycznego może bazować na interpolacji wielomianowej. Niech p będzie wielomianem interpolującym wartości funkcji f w $n+1$ punktach $x_0, x_1, \ldots, x_n$. Ze wzoru Lagrange'a (podrozdz. 6.1) i tw. 6.1.3 wynika, że

$$f(x) = \sum_{i=0}^{n} f(x_i) l_i(x) + \frac{1}{(n+1)!} f^{(n+1)}(\xi_x) w(x),$$

gdzie

$$l_i(x) := \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} \quad (0 \leqslant i \leqslant n), \quad w(x) := \prod_{i=0}^{n} (x - x_i).$$

Zróżniczkujmy stronami tę równość:

$$f'(x) = \sum_{i=0}^{n} f(x_i) l_i'(x) + \frac{1}{(n+1)!} \left[f^{(n+1)}(\xi_x) w'(x) + w(x) \frac{d}{dx} f^{(n+1)}(\xi_x) \right].$$

Wzór upraszcza się, gdy x jest jednym z węzłów, np. $x = x_m$, bo wtedy $w(x) = 0$. Prócz tego,

$$w'(x) = \sum_{i=0}^{n} \prod_{j=0, j\neq i}^{n} (x - x_j), \quad w'(x_m) = \prod_{j=0, j\neq m}^{n} (x_m - x_j),$$

a zatem

$$f'(x_m) = \sum_{i=0}^{n} f(x_i) l_i'(x_m) + \frac{1}{(n+1)!} f^{(n+1)}(\xi_{x_m}) \prod_{j=0, j\neq m}^{n} (x_m - x_j). \tag{7.1.4}$$

Ten wzór daje wartość przybliżoną pochodnej i jej błąd. Warto go stosować, gdy węzły nie są równoodległe.

Przykład 7.1.3. Podać jawną postać wzoru (7.1.4) dla $n = 2$ i $m = 1$.

Rozwiązanie. Z definicji wielomianów l_i wynika, że

$$l_0'(x) = \frac{2x - x_1 - x_2}{(x_0 - x_1)(x_0 - x_2)},$$
$$l_1'(x) = \frac{2x - x_0 - x_2}{(x_1 - x_0)(x_1 - x_2)},$$
$$l_2'(x) = \frac{2x - x_0 - x_1}{(x_2 - x_0)(x_2 - x_1)}.$$

Podstawiamy tu $x = x_1$ i wyrażamy (7.1.4) w postaci

$$f'(x_1) = f(x_0)\frac{x_1 - x_2}{(x_0 - x_1)(x_0 - x_2)} + \\ + f(x_1)\frac{2x_1 - x_0 - x_2}{(x_1 - x_0)(x_1 - x_2)} + \\ + f(x_2)\frac{x_1 - x_0}{(x_2 - x_0)(x_2 - x_1)} + \frac{1}{6}f'''(\xi_{x_1})(x_1 - x_0)(x_1 - x_2).$$

Jeśli $x_1 - x_0 = x_2 - x_1 = h$, to powyższy wzór jest identyczny (z dokładnością do oznaczeń) ze wzorem (7.1.2). ■

Ekstrapolacja Richardsona

Zobaczymy teraz, w jaki sposób można poprawić dokładność pewnych wzorów przybliżonych. W przypadku różniczkowania numerycznego punktem wyjścia jest szereg Taylora

$$f(x \pm h) = \sum_{k=0}^{\infty} \frac{(\pm 1)^k}{k!} h^k f^{(k)}(x).$$

Odejmując stronami te dwa równania, usuwamy wszystkie składniki z parzystymi k i wyrażamy pierwszą pochodną w następujący sposób:

$$f'(x) = \frac{1}{2h}[f(x + h) - f(x - h)] - \\ - \left[\frac{1}{3!}h^2 f^{(3)}(x) + \frac{1}{5!}h^4 f^{(5)}(x) + \frac{1}{7!}h^6 f^{(7)}(x) + \ldots\right].$$

Jest to równanie typu

$$L = \varphi(h) + a_2 h^2 + a_4 h^4 + a_6 h^6 + \ldots, \tag{7.1.5}$$

gdzie L oznacza pierwszą pochodną, a $\varphi(h)$ jej przybliżenie, takie jak w (7.1.2). Jest ono określone tylko dla $h \neq 0$. Suma $a_2 h^2 + a_4 h^4 + \ldots$ jest

jego błędem. Jeśli $a_2 \neq 0$, to dla dostatecznie małych h składnik a_2h^2 dominuje nad pozostałymi. Warto by go usunąć. W tym celu zmieńmy w (7.1.5) parametr h na $h/2$ i pomnóżmy obie strony przez 4:

$$4L = 4\varphi(h/2) + a_2h^2 + a_4h^4/4 + a_6h^6/16 + \dots$$

Odejmując stronami (7.1.5) od tego równania, otrzymujemy, po podzieleniu przez 3, nowe wyrażenie dla L:

$$L = \frac{4}{3}\varphi(h/2) - \frac{1}{3}\varphi(h) - a_4h^4/4 - 5a_6h^6/16 - \dots \tag{7.1.6}$$

Przejście od (7.1.5) do (7.1.6) jest pierwszym krokiem ekstrapolacji Richardsona. Widzimy, że łatwa do obliczenia kombinacja wartości $\varphi(h)$ i $\varphi(h/2)$ jest takim przybliżeniem dla L, którego błąd jest rzędu $\mathcal{O}(h^4)$, a nie $\mathcal{O}(h^2)$, jak to było w pierwotnym wzorze. Zauważmy też, że ponieważ nasze rozumowanie nie zależy od interpretacji wielkości L i $\varphi(h)$, więc stosuje się i do pewnych innych zadań numerycznych, np. całkowania za pomocą wzoru trapezów (podrozdz. 7.5).

Przykład 7.1.4. Stosując pierwszy krok ekstrapolacji Richardsona, obliczyć ponownie wartość pochodnej z przykł. 7.1.2.

Rozwiązanie. Niech będzie $f(x) := \operatorname{arctg} x$. Algorytm obliczania przybliżeń pochodnej $f'(\sqrt{2})$ jest następujący:

$s \leftarrow \sqrt{2}$; $h \leftarrow 1$; $M \leftarrow 30$
for $k = 0$ **to** M **do**
    $d_k \leftarrow [f(s+h) - f(s-h)]/(2h)$
    $h \leftarrow h/2$
end do
for $k = 1$ **to** M **do**
    $r_k \leftarrow d_k + (d_k - d_{k-1})/3$
    output k, d_k, r_k
end do

Algorytm daje m.in. takie wyniki:

k	d_k	r_k
2	0.33719 385	0.33333 480
4	0.33357 477	0.33333 364
8	0.33332 825	0.33332 571
16	0.33203 125	0.33138 022
26	0.00000 000	0.00000 000

Najlepsze przybliżenie ma sześć cyfr dokładnych, o dwie więcej niż w przykł. 7.1.2. ■

Powyższy przykład jest w pewnym sensie sztuczny – dokładną wartość pochodnej znamy i korzystamy z niej, oceniając dokładność przybliżeń. W realnych zastosowaniach powinniśmy uwzględniać liczbę cyfr znaczących, które tracimy wskutek odejmowania dającego d_k. Oczywiście r_k nie może mieć tych cyfr więcej niż ma d_k. Nie oczekujmy zatem, że nieograniczony wzrost k da coraz lepsze przybliżenia.

Czytelnicy chyba się domyślają, że rozumowanie zastosowane do (7.1.5) można powtórzyć, z małymi zmianami, dla (7.1.6). Żeby to zrobić przyjmijmy, że

$$\psi(h) := \frac{4}{3}\varphi(h/2) - \frac{1}{3}\varphi(h).$$

Mamy zatem

$$L = \psi(h) + b_4h^4 + b_6h^6 + \ldots,$$
$$L = \psi(h/2) + b_4h^4/16 + b_6h^6/64 + \ldots$$

Mnożąc drugie z tych równań stronami przez 16 i odejmując od wyniku pierwsze równanie, eliminujemy składnik z h^4. Daje to nowe wyrażenie dla L:

$$L = \frac{16}{15}\psi(h/2) - \frac{1}{15}\psi(h) - b_6h^6/20 - \ldots$$

Wobec tego dla

$$\theta(h) := \frac{16}{15}\psi(h/2) - \frac{1}{15}\psi(h)$$

mamy równość

$$L = \theta(h) + c_6h^6 + c_8h^8 + \ldots$$

Rozumując jak wyżej, wnioskujemy, że

$$L = \frac{64}{63}\theta(h/2) - \frac{1}{63}\theta(h) - \frac{1}{84}c_8h^8 - \ldots$$

Można już dać pełny opis metody zwanej *ekstrapolacją Richardsona*. Jeśli ograniczamy się do jej M początkowych kroków, to:

1. Wybieramy odpowiednie h (np. $h = 1$) i obliczamy wielkości

$$D(n,0) := \varphi(h/2^n) \quad (0 \leqslant n \leqslant M).$$

2. Dla $k = 1, 2, \ldots, M$ i $n = k, k+1, \ldots, M$ stosujemy wzór rekurencyjny

$$D(n,k) := D(n,k-1) + \frac{D(n,k-1) - D(n-1,k-1)}{4^k - 1}. \qquad (7.1.7)$$

Twierdzenie 7.1.5. *Przy założeniu* (7.1.5) *wyniki* $D(n,k)$ *ekstrapolacji Richardsona są takie, że*

$$D(n,k-1) = L + \sum_{j=k}^{\infty} A_{jk}(h/2^n)^{2j}. \tag{7.1.8}$$

Dowód. Zgodnie z definicją jest

$$D(n,0) = \varphi(h/2^n) = L - \sum_{j=1}^{\infty} a_{2j}(h/2^n)^{2j},$$

czyli (7.1.8) zachodzi dla $k=1$, przy czym $A_{j1} = -a_{2j}$. Zakładając, że ta równość jest prawdziwa dla pewnego $k \geqslant 1$, wnioskujemy z (7.1.7) i (7.1.8), że $D(n,k)$ jest równe

$$\frac{1}{4^k-1}\left\{4^k\left[L + \sum_{j=k}^{\infty} A_{jk}\left(\frac{h}{2^n}\right)^{2j}\right] - \left[L + \sum_{j=k}^{\infty} A_{jk}\left(\frac{h}{2^{n-1}}\right)^{2j}\right]\right\} =$$

$$= L + \sum_{j=k}^{\infty} \frac{4^k - 4^j}{4^k - 1} A_{jk}\left(\frac{h}{2^n}\right)^{2j}.$$

Wystarczy teraz przyjąć

$$A_{j,k+1} := \frac{4^k - 4^j}{4^k - 1} A_{jk}$$

i zauważyć, że $A_{k,k+1} = 0$. ■

Z opisu ekstrapolacji Richardsona wynika, że daje ona następującą trójkątną tablicę przybliżeń wielkości L:

$$\begin{array}{llll} D(0,0) & & & \\ D(1,0) & D(1,1) & & \\ D(2,0) & D(2,1) & D(2,2) & \\ \vdots & \vdots & \vdots & \ddots \\ D(M,0) & D(M,1) & D(M,2) & \ldots \quad D(M,M) \end{array}$$

W poniższym algorytmie konstrukcji tej tablicy zakładamy, że funkcja φ jest dana.

input h, M
for $n = 0$ **to** M **do**
    $D(n,0) \leftarrow \varphi(h/2^n)$
end do

```
for k = 1 to M do
    for n = k to M do
        D(n,k) ← D(n,k-1) + [D(n,k-1) - D(n-1,k-1)]/(4^k - 1)
        output D(n,k)
    end do
end do
```

Przykład 7.1.6. Zastosować ekstrapolację Richardsona dla danych z przykł. 7.1.2 i $h = 1$.

Rozwiązanie. Dla $M = 4$ wyniki są następujące:

n	$D(n,0)$	$D(n,1)$	$D(n,2)$	$D(n,3)$	$D(n,4)$
0	0.3926991				
1	0.3487710	0.3341283			
2	0.3371938	0.3333348	0.3332819		
3	0.3342981	0.3333329	0.3333328	0.3333336	
4	0.3335748	0.3333336	0.3333337	0.3333337	0.3333337

Są one nieco bardziej dokładne niż w przykł. 7.1.4. Wielkości $D(n,k)$ dla $n > 4$ mogą być gorsze lub nawet pozbawione sensu wskutek błędów wywoływanych odejmowaniem. ■

W ostatnich latach powstało oprogramowanie automatyzujące proces różniczkowania. Bischof, Carle, Khademi i Mauer [1994] zaprojektowali taki system o nazwie ADIFOR, który oblicza pochodne, stosując elementarne reguły różniczkowania. Stosowane metody opisują np. Griewank i Corliss [1991].

Zadania 7.1

1. Wykazać, że w (7.1.3) błąd ma postać $\sum_{n=1}^{\infty} a_{2n}h^{2n}$. Wyznaczyć współczynniki a_{2n}.
2. Znaleźć błąd wzorów przybliżonych

$$f'(x) \approx \frac{1}{2h}[-3f(x) + 4f(x+h) - f(x+2h)],$$

$$f''(x) \approx \frac{1}{h^2}[f(x) - 2f(x+h) + f(x+2h)].$$

3. Uzasadnić wzory przybliżone

$$f'(x) \approx \frac{1}{12h}[-f(x+2h) + 8f(x+h) - 8f(x-h) + f(x-2h)],$$

$$f''(x) \approx \frac{1}{12h^2}[-f(x+2h) + 16f(x+h) - 30f(x) + 16f(x-h) - f(x-2h)]$$

i wykazać, że w obu przypadkach błąd jest równy $\mathcal{O}(h^4)$.

4. Udowodnić, że

$$f'''(x) \approx \frac{1}{h^3}[f(x+3h) - 3f(x+2h) + 3f(x+h) - f(x)],$$
$$f'''(x) \approx \frac{1}{2h^3}[f(x+2h) - 2f(x+h) + 2(x-h) - f(x-2h)].$$

Znaleźć błędy tych przybliżeń. Które z nich jest dokładniejsze?

5. Powtórzyć poprzednie zadanie dla następujących przybliżeń czwartej pochodnej:

$$f^{(4)}(x) \approx \frac{1}{h^4}[f(x+4h) - 4f(x+3h) + 6f(x+2h) - 4f(x+h) + f(x)],$$
$$f^{(4)}(x) \approx \frac{1}{h^4}[f(x+2h) - 4f(x+h) + 6f(x) - 4f(x-h) + f(x-2h)].$$

6. Wyznaczyć sensowne współczynniki wzoru przybliżonego

$$f''(x) \approx \frac{1}{h^2}[Af(x+3h) + Bf(x+2h) + Cf(x+h) + Df(x)].$$

7. Niech będzie $L = \lim_{h\to 0} f(h)$ i $L - f(h) = c_6h^6 + c_9h^9 + \ldots$. Jaka kombinacja wartości $f(h)$ i $f(h/2)$ jest najlepszym przybliżeniem dla L?

8. Wyjaśnić, jak powinna działać ekstrapolacja Richardsona w przypadku, gdy: (**a**) $L = \varphi(h) + \sum_{j=1}^{\infty} a_j h^j$, (**b**) $L = \varphi(h) + \sum_{j=1}^{\infty} a_j h^{2j-1}$. Udowodnić w tych przypadkach odpowiednik tw. 7.1.5.

ZADANIA KOMPUTEROWE 7.1

K1. Gdy obliczamy $f'(x)$ ze wzoru przybliżonego (7.1.1), to błąd zaokrąglenia (wywołany głównie odejmowaniem bliskich wartości $f(x+h)$ i $f(x)$) jest z grubsza równy αh^{-1}. Sprawdzić to (i oszacować α) na kilku przykładach. Może być potrzebne wykonywanie obliczeń w podwójnej precyzji.

K2. Zaprogramować metodę ekstrapolacji Richardsona i sprawdzić ją, obliczając $f'(x)$ dla: (**a**) $f(x) := \log x$ i $x := 3$, (**b**) $f(x) := \operatorname{tg} x$ i $x := \arcsin 0.8$, (**c**) $f(x) := \sin(x^2 + \frac{1}{3}x)$ i $x := 0$.

K3. Napisać i sprawdzić program obliczający $f''(x)$, a wzorowany na ekstrapolacji Richardsona.

K4. Zaprojektować i sprawdzić algorytm przybliżonego obliczania f' i f'' za pomocą funkcji sklejanych stopnia trzeciego.

7.2. Interpolacja w całkowaniu numerycznym

Wiadomo, że całki nieoznaczone z wielu – nawet bardzo prostych – funkcji nie wyrażają się w skończonej postaci przez funkcje elementarne. Jest

tak m.in. dla $\int e^{-x^2}\,dx$. Stąd wynika, że bardzo często nie możemy znaleźć dokładnych wartości całek oznaczonych. Dotyczy to np. całek

$$\int_0^2 e^{-x^2}\,dx, \quad \int_0^\pi \cos(3\cos\theta)\,d\theta,$$
$$\int_0^1\int_0^1 \sin(xye^x)\,dx\,dy, \quad \int_0^1\int_{x^2}^x \operatorname{tg}(xy^2)\,dy\,dx.$$

Jeśli chcemy znaleźć przybliżoną wartość całki oznaczonej

$$\int_a^b f(x)\,dx,$$

to ogólny sposób postępowania polega na zastąpieniu funkcji f inną, bliską funkcją g, dla której całkę można łatwo obliczyć. Stosujemy więc wzór przybliżony

$$\int_a^b f(x)\,dx \approx \int_a^b g(x)\,dx.$$

Taką dobrą funkcją g może być wielomian (jego całkowanie jest łatwe), który interpoluje funkcję f w danych węzłach, co powinno zapewnić bliskość obu funkcji. Oczywiście, przybliżenie wielomianowe można otrzymać też inaczej, np. jako sumę częściową szeregu Taylora dla f. Oto przykład:

$$\int_0^1 e^{x^2}\,dx \approx \int_0^1 \sum_{k=0}^{n} \frac{1}{k!}\,x^{2k}\,dx = \sum_{k=0}^{n} \frac{1}{(2k+1)k!}. \tag{7.2.1}$$

Chcielibyśmy jednak mieć jakąś w miarę uniwersalną metodę wymagającą tylko obliczania wartości funkcji podcałkowej. Daje to interpolacja wielomianowa, ale i interpolacja za pomocą funkcji sklejanych, których całkowanie jest też łatwe.

Zastosowanie interpolacji wielomianowej

Wybierzmy w przedziale całkowania $[a,b]$ węzły $x_0, x_1, \ldots, x_n$ i zastosujmy wzór interpolacyjny Lagrange'a znany z podrozdz. 6.1:

$$p(x) = \sum_{i=0}^{n} f(x_i)l_i(x), \quad \text{gdzie} \quad l_i(x) := \prod_{j=0,\,j\neq i}^{n} \frac{x - x_j}{x_i - x_j}.$$

Wnioskujemy stąd, że

$$\int_a^b f(x)\,dx \approx \int_a^b p(x)\,dx = \sum_{i=0}^{n} f(x_i) \int_a^b l_i(x)\,dx.$$

Jest to szczególny wariant typowego wzoru całkowania numerycznego, zwanego też tradycyjnie *kwadraturą*:

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} A_i f(x_i). \tag{7.2.2}$$

W tym przypadku

$$A_i := \int_a^b l_i(x)\,dx \quad (0 \leqslant i \leqslant n). \tag{7.2.3}$$

Jeśli węzły interpolacji są takie, że $x_i = a + (b-a)i/n$ dla $0 \leqslant i \leqslant n$, to wzór (7.2.2) z powyższymi A_i nazywamy *wzorem Newtona-Cotesa.*

Wzór trapezów

Najprostszy, ale już użyteczny wzór otrzymujemy dla $n = 1$; wtedy $x_0 = a$ i $x_1 = b$. Ponieważ

$$l_0(x) = \frac{b-x}{b-a}, \quad l_1(x) = \frac{x-a}{b-a},$$

więc

$$A_0 = \int_a^b l_0(x)\,dx = \frac{1}{2}(b-a), \quad A_1 = \int_a^b l_1(x)\,dx = \frac{1}{2}(b-a),$$

co daje tzw. *wzór trapezów*

$$\int_a^b f(x)\,dx \approx \frac{1}{2}(b-a)[f(a)+f(b)].$$

Jest on dokładny, gdy $f \in \Pi_1$, czyli f jest wielomianem stopnia co najwyżej pierwszego, a w innych przypadkach jego błąd [1)] wynosi

$$-\frac{1}{12}(b-a)^3 f''(\xi), \quad \text{gdzie} \quad \xi \in (a,b)$$

Można to udowodnić, całkując błąd $f(x) - p_1(x) = \frac{1}{2}f''(\xi_x)(x-a)(x-b)$ interpolacji i korzystając z twierdzenia o wartości średniej dla całek. Wrócimy

[1)] Tu i dalej, dla innych wzorów całkowania numerycznego, *dokładna* wartość całki jest równa sumie wartości *przybliżonej* i błędu. Chociaż autorzy nie precyzują założeń o funkcji podcałkowej, to oczywiście każde wyrażenie dla błędu jest poprawne tylko wtedy, gdy występująca w nim pochodna istnieje w $[a,b]$, albo – gdy ponadto jest ciągła w (a,b) (*przyp. tłum.*).

do tego wzoru w podrozdz. 7.5, gdyż jest elementem opisanej tam metody Romberga. Tu zaś zauważmy, że można podzielić przedział $[a,b]$ punktami

$$a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$$

na podprzedziały i zastosować do każdego z nich wzór trapezów:

$$\int_a^b f(x)\,dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)\,dx \approx \frac{1}{2}\sum_{i=1}^{n}(x_i - x_{i-1})[f(x_{i-1}) + f(x_i)].$$

Jest to *złożony wzór trapezów* (oczywiście takie *złożenie* można zastosować do każdego innego wzoru całkowania numerycznego). Wzór jest dokładny, jeśli wykres funkcji jest linią łamaną, której wierzchołki mają odcięte x_i.

Złożony wzór trapezów jest najprostszy, gdy wszystkie podprzedziały są jednakowo długie, tj. dla $x_i = a + ih$, gdzie $h := (b-a)/n$:

$$\int_a^b f(x)\,dx \approx \frac{1}{2}h\Big[f(a) + 2\sum_{i=1}^{n-1} f(a+ih) + f(b)\Big],$$

czyli

$$\int_a^b f(x)\,dx \approx h\sum_{i=0}^{n}{}'' f(a+ih) \tag{7.2.4}$$

(symbol $\sum''$ oznacza, że pierwszy i ostatni składnik sumy należy podzielić przez 2). Jeśli druga pochodna funkcji f jest ciągła w (a,b), to błąd tego wzoru wynosi

$$-\frac{1}{12n^2}(b-a)^3 f''(\xi), \quad \text{gdzie} \quad \xi \in (a,b).$$

Żeby otrzymać to wyrażenie, sumujemy błędy całkowania w poszczególnych podprzedziałach. Trzeba też zauważyć, że istnieje takie ξ, iż średnia arytmetyczna $(1/n)\sum_{i=1}^{n} f''(\xi_i)$ jest równa $f''(\xi)$.

Przykład 7.2.1. Wykazać, że wzór Newtona-Cotesa dla $n = 2$ i $[a,b] = [0,1]$ jest następujący:

$$\int_0^1 f(x)\,dx \approx \frac{1}{6}[f(0) + 4f(\tfrac{1}{2}) + f(1)]. \tag{7.2.5}$$

Rozwiązanie. Wystarczy skorzystać ze wzorów

$$l_0(x) = 2(x - \tfrac{1}{2})(x-1), \quad l_1(x) = -4x(x-1), \quad l_2(x) = 2x(x - \tfrac{1}{2})$$

i obliczyć całki A_i. ■

Metoda nieoznaczonych współczynników

Z dowodu wzoru (7.2.2) wynika, że – dla określonych tam współczynników A_k – jest on dokładny, gdy $f \in \Pi_n$. Te wielkości spełniają więc układ równań

$$\int_a^b x^j\,dx = \sum_{i=0}^{n} A_i x_i^j \quad (0 \leqslant j \leqslant n)$$

i zamiast obliczać A_i z (7.2.3), wystarczy ten układ rozwiązać. Dla danych z przykł. 7.2.1 ma on postać

$$\begin{aligned} A_0 + A_1 + A_2 &= 1,\\ \tfrac{1}{2}A_1 + A_2 &= \tfrac{1}{2},\\ \tfrac{1}{4}A_1 + A_2 &= \tfrac{1}{3}, \end{aligned}$$

a jego rozwiązaniem są liczby: $A_0 = \frac{1}{6}$, $A_1 = \frac{2}{3}$, $A_2 = \frac{1}{6}$, oczywiście zgodne ze wspomnianym przykładem. Taki sposób wyznaczania liczb A_i nazywamy *metodą nieoznaczonych współczynników.*

Wzór Simpsona

Podobne obliczenia dla dowolnego przedziału $[a, b]$ dają znany *wzór Simpsona*

$$\int_a^b f(x)\,dx \approx \frac{1}{6}(b-a)\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]. \tag{7.2.6}$$

Wzór z założenia jest dokładny dla wszystkich wielomianów stopnia $\leqslant 2$, ale także – co jest zaskakujące – dla wielomianów stopnia trzeciego (zob. zad. 13). Błąd tego wzoru wynosi

$$-\frac{1}{2880}(b-a)^5 f^{(4)}(\xi), \quad \text{gdzie} \quad \xi \in (a, b). \tag{7.2.7}$$

Najłatwiej to udowodnić stosując tw. 7.7.2; zob. zad. 7.7.2.

Złożony wzór Simpsona powstaje z (7.2.6) przez zastosowanie tego ostatniego do n podprzedziałów jednakowej długości. Wprowadzamy oznaczenia

$$x_i = a + ih \quad (0 \leqslant i \leqslant 2n), \quad \text{gdzie} \quad h := (b-a)/(2n).$$

Ponieważ

$$\int_a^b f(x)\,dx = \sum_{i=1}^{n} \int_{x_{2i-2}}^{x_{2i}} f(x)\,dx,$$

więc z (7.2.6) wynika po oczywistych uproszczeniach, że

$$\int_a^b f(x)\,dx \approx \frac{1}{3}h\Big[f(x_0)+4\sum_{i=1}^{n} f(x_{2i-1})+2\sum_{i=1}^{n-1} f(x_{2i})+f(x_{2n})\Big]. \quad (7.2.8)$$

Błąd tego wzoru wynosi

$$-\frac{1}{2880n^4}(b-a)^5 f^{(4)}(\xi), \quad \text{gdzie} \quad \xi \in (a,b).$$

Całki z funkcją wagową

Metoda prowadząca do wzorów Newtona-Cotesa daje również ogólniejsze wzory przybliżone postaci

$$\int_a^b f(x)w(x)\,dx \approx \sum_{i=0}^{n} A_i f(x_i). \quad (7.2.9)$$

w jest tu *funkcją wagową* (krócej: *wagą*), z definicji nieujemną. Jedyna różnica polega na tym, że teraz

$$A_i := \int_a^b l_i(x)w(x)\,dx. \quad (7.2.10)$$

Przykład 7.2.2. Znaleźć wzór

$$\int_{-\pi}^{\pi} f(x)\cos x\,dx \approx A_0 f\left(-\frac{3}{4}\pi\right)+A_1 f\left(-\frac{1}{4}\pi\right)+A_2 f\left(\frac{1}{4}\pi\right)+A_3 f\left(\frac{3}{4}\pi\right),$$

dokładny[2)] dla każdej funkcji $f \in \Pi_3$.

Rozwiązanie. Stosujemy metodę nieoznaczonych współczynników, czyli wyznaczamy A_i, żądając, aby wzór był dokładny dla czterech jednomianów: $f(x) = 1$, x, x^2, x^3. Ponieważ cosinus jest funkcją parzystą, a węzły są rozmieszczone symetrycznie względem 0, więc dla $A_0 = A_3$ i $A_1 = A_2$ wzór jest na pewno dokładny dla każdej funkcji f nieparzystej, a pozostałe warunki się upraszczają:

$$0 = \int_{-\pi}^{\pi} \cos x\,dx = 2A_0 + 2A_1,$$

[2)] Ogólniej, dla A_i określonych w (7.2.10) wzór całkowania jest dokładny, jeśli f (a nie fw!) jest dowolnym wielomianem klasy Π_n. Rozkład funkcji podcałkowej na czynniki f i w może mieć różne powody. W szczególności funkcja wagowa – a więc i kompletna funkcja podcałkowa – może zawierać osobliwości, których nie ma żaden wielomian; np. dla $[a,b] = [-1,1]$ może być $w(x) = (1-x^2)^{-1/2}$ (*przyp. tłum.*).

$$-4\pi = \int_{-\pi}^{\pi} x^2 \cos x\, dx = 2A_0 \left(\frac{3}{4}\pi\right)^2 + 2A_1 \left(\frac{1}{4}\pi\right)^2.$$

Stąd $A_1 = A_2 = -A_0 = -A_3 = 4/\pi$ i szukany wzór jest następujący:

$$\int_{-\pi}^{\pi} f(x) \cos x\, dx \approx \frac{4}{\pi}\left[-f\left(-\frac{3}{4}\pi\right) + f\left(-\frac{1}{4}\pi\right) + f\left(\frac{1}{4}\pi\right) - f\left(\frac{3}{4}\pi\right)\right].$$

■

Zmiana przedziału całkowania

Znając wzór całkowania w pewnym przedziale, można go dostosować do innego przedziału przez przekształcenie liniowe zmiennej. Zachowuje się przy tym istotna własność wzoru, a mianowicie jego dokładność dla wielomianów pewnego stopnia. Niech pierwotny wzór ma postać

$$\int_c^d f(t)\, dt \approx \sum_{i=0}^{n} A_i f(t_i).$$

Jego pochodzenie nie jest istotne; zakładamy tylko, że jest on dokładny dla $f \in \Pi_m$. Aby otrzymać stąd wzór dla przedziału $[a, b]$, posługujemy się funkcją

$$\lambda(t) := \frac{(b-a)t + ad - bc}{d - c},$$

która przekształca przedział $[c, d]$ na $[a, b]$. Podstawienie $x = \lambda(t)$ w całce $\int_a^b f(x)\, dx$ daje równość

$$\int_a^b f(x)\, dx = \frac{b-a}{d-c} \int_c^d f(\lambda(t))\, dt \approx \frac{b-a}{d-c} \sum_{i=0}^{n} A_i f(\lambda(t_i)),$$

czyli ostatecznie

$$\int_a^b f(x)\, dx \approx \frac{b-a}{d-c} \sum_{i=0}^{n} A_i f\left(\frac{(b-a)t_i + ad - bc}{d-c}\right).$$

Jeśli $f(t)$ jest wielomianem, to $f(\lambda(t))$ jest nim również, a stopnie obu są identyczne. Dlatego nowy wzór jest dokładny, jeśli $f \in \Pi_m$. W opisany sposób można otrzymać ogólny wzór Simpsona (7.2.6) z (7.2.5).

Analiza błędu

Wzór przybliżony całkowania (7.2.2) wynika ze wzoru interpolacyjnego Lagrange'a. Jeśli $p \in \Pi_n$ interpoluje w węzłach $x_0, x_1, \ldots, x_n$ z przedziału $[a,b]$ funkcję f, która ma tam $(n+1)$-szą pochodną ciągłą, to na mocy tw. 6.1.5

$$f(x) - p(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi_x) \prod_{i=0}^{n} (x - x_i).$$

Stąd wynika, że

$$\int_a^b f(x)\,dx - \sum_{i=0}^{n} A_i f(x_i) = \frac{1}{(n+1)!} \int_a^b f^{(n+1)}(\xi_x) \prod_{i=0}^{n} (x - x_i)\,dx. \tag{7.2.11}$$

Jeśli w przedziale $[a,b]$ jest $|f^{(n+1)}(x)| \leqslant M$, to

$$\left| \int_a^b f(x)\,dx - \sum_{i=0}^{n} A_i f(x_i) \right| \leqslant \frac{M}{(n+1)!} \int_a^b \prod_{i=0}^{n} |x - x_i|\,dx. \tag{7.2.12}$$

Sprawdzając, kiedy to oszacowanie jest najlepsze (tzn. całka po prawej stronie jest najmniejsza), można ograniczyć się do przypadku $[a,b] = [-1,1]$. Jak się okazuje, rozwiązanie zadania wynika z własności *wielomianów Czebyszewa drugiego rodzaju* U_n. Można je zdefiniować rekurencyjnie wzorami

$$U_0(x) := 1, \quad U_1(x) := 2x, \quad U_n(x) := 2xU_{n-1}(x) - U_{n-2}(x) \quad (n \geqslant 2)$$

(ostatni jest taki, jak dla wielomianów pierwszego rodzaju T_n; zob. podrozdz. 6.1) albo jawnym wzorem

$$U_n(x) := \frac{\sin(n+1)\theta}{\sin\theta}, \quad \text{gdzie} \quad x := \cos\theta \tag{7.2.13}$$

(równoważność określeń wynika ze znanych wzorów trygonometrycznych). Z pierwszej definicji wynika, że $U_n(x)$ jest wielomianem stopnia n, ze współczynnikiem 2^n przy x^n, a z drugiej, że zerami U_n są punkty

$$x_i := \cos\frac{(i+1)\pi}{n+1} \quad (0 \leqslant i \leqslant n-1).$$

Twierdzenie 7.2.3. *W zbiorze wszystkich wielomianów $p(x) \in \Pi_n$ o współczynniku 1 przy x^n całka $\int_{-1}^{1} |p(x)|\,dx$ ma najmniejszą wartość, równą 2^{-n+1}, tylko dla $p = 2^{-n}U_n$.*

Dowód. Udowodnimy najpierw, że

$$I_m := \int_{-1}^{1} U_m(x) \operatorname{sgn}[U_n(x)]\, dx = 0 \quad (0 \leqslant m < n). \tag{7.2.14}$$

W tym celu podstawmy tu $x := \cos\theta$:

$$I_m = \int_0^{\pi} \frac{\sin(m+1)\theta}{\sin\theta} \operatorname{sgn}\left[\frac{\sin(n+1)\theta}{\sin\theta}\right] \sin\theta\, d\theta.$$

Stąd dla $\varphi := \pi/(n+1)$ wynika, że

$$I_m = \sum_{k=0}^{n} (-1)^k \int_{k\varphi}^{(k+1)\varphi} \sin(m+1)\theta\, d\theta =$$
$$= (m+1)^{-1} \sum_{k=0}^{n} (-1)^{k+1} [\cos(m+1)(k+1)\varphi - \cos(m+1)k\varphi].$$

Niech będzie $\alpha := (m+1)\varphi + \pi$; stąd i z założenia, że $m < n$, wynika nierówność $0 < \alpha < 2\pi$. Sprawdzamy, że

$$(m+1)I_m = \sum_{k=0}^{n} [\cos(k+1)\alpha + \cos k\alpha] =$$
$$= 2 \sum_{k=0}^{n+1}{}'' \cos k\alpha = \frac{\cos\frac{1}{2}\alpha \sin(n+1)\alpha}{\sin\frac{1}{2}\alpha}$$

(symbol $\sum''$ określono po wzorze (7.2.4)). Końcowa równość jest prawdziwa, jeśli tylko $\alpha \neq 0, \pm 2\pi, \pm 4\pi, \ldots$, co można udowodnić mnożąc jej obie strony przez $\sin\frac{1}{2}\alpha$. Ponieważ $(n+1)\alpha = (m+n+2)\pi$, więc $I_m = 0$ dla $m < n$. Natomiast dla $m = n$ jest $\alpha = 2\pi$, a wtedy przedostatnie wyrażenie w ciągu przekształceń jest równe $2n+2$, czyli

$$I_n = \int_{-1}^{1} |U_n(x)|\, dx = 2,$$

co należało wykazać.

Aby zakończyć dowód, zauważmy, że każdy wielomian p z twierdzenia można wyrazić w postaci

$$p = 2^{-n} U_n + a_{n-1} U_{n-1} + \ldots + a_0 U_0.$$

Stąd i z (7.2.14) wynika, że

$$\int_{-1}^{1} |p(x)|\,dx \geqslant \int_{-1}^{1} p(x)\,\mathrm{sgn}[U_n(x)]\,dx =$$

$$= 2^{-n}\int_{-1}^{1} U_n(x)\,\mathrm{sgn}[U_n(x)]\,dx = 2^{-n}\int_{-1}^{1} |U_n(x)|\,dx.$$

Ściślej, po pierwszej całce zamiast $\geqslant$ możemy napisać $=$ tylko wtedy, gdy $p = U_n$. Dlatego jedynie wielomian $2^{-n}U_n$ zapewnia minimum badanej całki. ■

Z twierdzenia 7.2.3 i poprzedzających je uwag wynika, że oszacowanie z góry (7.2.12) błędu wzoru całkowania (7.2.11) jest najlepsze wtedy, gdy węzły x_i powstają z zer wielomianu U_{n+1} przez odpowiednie przekształcenie liniowe, tj. gdy

$$x_i = \frac{1}{2}(a+b) + \frac{1}{2}(b-a)\cos\frac{(i+1)\pi}{n+2} \quad (0 \leqslant i \leqslant n).$$

Tylko dla takich x_i możemy twierdzić, że

$$\left|\int_a^b f(x)\,dx - \sum_{i=0}^{n} A_i f(x_i)\right| \leqslant \frac{M(b-a)^{n+2}}{2^{2n+2}(n+1)!}.$$

Więcej wiadomości o całkowaniu numerycznym można znaleźć w monografiach: Davis i Rabinowitz [1984], Ghizzetti i Ossiccini [1970] i Krylov [1962][3)].

ZADANIA 7.2

1. Korzystając z (7.2.1), znaleźć osiem cyfr znaczących całki $\int_0^1 e^{x^2}\,dx$.

2. Obliczyć siedem cyfr znaczących całki $\int_0^{0.01} x^{-1}\sin x\,dx$, korzystając z rozwinięcia funkcji podcałkowej w szereg potęgowy.

3. Znaleźć wzór Newtona-Cotesa dla całki $\int_0^1 f(x)\,dx$ i węzłów $0, \frac{1}{3}, \frac{2}{3}, 1$.

4. Które z niżej podanych wyrażeń równych dokładnie całce $\int_0^1 f(x)\,dx$ dla każdego $f \in \Pi_2$ jest lepsze (i w jakim sensie)?

$$af(0) + bf\left(\tfrac{1}{2}\right) + cf(1), \qquad \alpha f\left(\tfrac{1}{4}\right) + \beta f\left(\tfrac{1}{2}\right) + \gamma f\left(\tfrac{3}{4}\right).$$

5. Sprawdzić, że wzór

$$\int_0^1 f(x)\,dx \approx \frac{1}{90}\left[7f(0) + 32f\left(\frac{1}{4}\right) + 12f\left(\frac{1}{2}\right) + 32f\left(\frac{3}{4}\right) + 7f(1)\right]$$

jest dokładny dla $f \in \Pi_4$.

[3)] Krylov i Šulgina [*1966] oprócz informacji teoretycznych podają także tablice współczynników i węzłów dla wielu specjalnych wzorów przybliżonych całkowania (*przyp. tłum.*).

6. (cd.). Stosując wzór z poprzedniego zadania, obliczyć przybliżoną wartość całki $\int_0^1 (t+1)^{-1} dt$ i porównać ją z dokładną wartością, równą $\log 2$.

7. Znaleźć wzór przybliżony postaci $\int_0^1 f(x)\,dx \approx Af(\frac{1}{3}) + Bf(\frac{2}{3})$.

8. Dla: (**a**) $[a,b] = [0,2]$, (**b**) $[a,b] = [-1,3]$ znaleźć takie A, B, C, żeby wzór $\int_a^b xf(x)\,dx \approx Af(0)+Bf(1)+Cf(2)$ był dokładny dla wielomianów f możliwie wysokiego stopnia. Jaki jest ten maksymalny stopień?

9. Znaleźć, jeśli to możliwe, wzór $\int_0^1 f(x)\,dx \approx \alpha[f(x_0) + f(x_1)]$ dokładny dla każdego $f \in \Pi_2$.

10. Stosując wielomian stopnia $\leqslant 1$, interpolujący f w x_1 i x_2, otrzymać przybliżone wyrażenie dla całki $\int_{x_0}^{x_3} f(x)\,dx$. Założyć tylko, że $x_0 < x_1 < x_2 < x_3$.

11. Oszacować z góry prawą stronę nierówności (7.2.12), nie zakładając nic o rozmieszczeniu węzłów w przedziale $[a,b]$. Czy można znaleźć najlepsze takie oszacowanie?

12. Udowodnić, że jeśli wzór $\int_{-1}^1 f(x)\,dx \approx \sum_{i=0}^n A_i f(x_i)$, gdzie n jest parzyste, a węzły są rozmieszczone symetrycznie względem 0, jest dokładny dla wszystkich $f \in \Pi_n$, to ma również tę własność dla $f \in \Pi_{n+1}$.

13. Sprawdzić (nie odwołując się do wyrażenia dla błędu), że wzór Simpsona (7.2.6) jest dokładny dla wszystkich wielomianów stopnia trzeciego.

14. Znaleźć wzory przybliżone dla całki $\int_{-1}^1 x^m f(x)\,dx$ zależne tylko od $f(0)$, $f'(-1)$ i $f''(1)$ i dokładne dla każdego $f \in \Pi_2$. Przyjąć, że: (**a**) $m = 0$, (**b**) $m = 2$. Czy otrzymane wyrażenia są też dokładne dla $f \in \Pi_3$?

15. Podać wzory złożone całkowania numerycznego, oparte na następujących prostych wzorach:

(**a**) $\int_0^1 f(x)\,dx \approx f(1)$ (punkty podziału przedziału dowolne).

(**b**) $\int_{-1}^1 f(x)\,dx \approx 2f(0)$ (punkty podziału rozmieszczone w równych odstępach lub nie).

16. Jak, po zastosowaniu dla pewnego n wzoru złożonego Simpsona (7.2.8), obliczyć najmniejszym kosztem analogiczne wyrażenie, ale z $2n$ zamiast n?

17. Ile co najmniej podprzedziałów trzeba uwzględnić w złożonym wzorze trapezów, aby obliczyć całkę $\int_1^2 (x + e^{-x^2})\,dx$ z błędem mniejszym od $0.5_{10}{-}7$?

18. Udowodnić ściśle, jak wyraża się błąd: (**a**) prostego i złożonego wzoru trapezów, (**b**) prostego i złożonego wzoru Simpsona.

19. Udowodnić, że $T_n' = nU_{n-1}$.

20. Udowodnić, że $\int_{-1}^1 U_m(x)U_n(x)\sqrt{1-x^2}\,dx = (\pi/2)\delta_{mn}$.

ZADANIA KOMPUTEROWE 7.2

K1. Napisać program obliczający $\int_0^x e^{-t^2}\,dt$ przez sumowanie tych początkowych składników odpowiedniego szeregu potęgowego, których wartość bezwzględna przekracza 10^{-8}. Sprawdzić program dla $x = 0.1, 0.2, \ldots, 1.0$.

K2. Napisać program przybliżający całkę $\int_a^b f(x)\,dx$ za pomocą całki $\int_a^b S(x)\,dx$, gdzie S jest naturalną funkcją sklejaną sześcienną interpolującą f w węzłach $a+(b-a)i/n$ $(0 \leqslant i \leqslant n)$. Tę drugą całkę obliczyć, korzystając ze wzoru podanego w zad. 6.4.K2. Sprawdzić program dla całek:
(**a**) $\int_0^1 (1+x^2)^{-1}\,dx$, (**b**) $\int_1^3 x^{-1}\,dx$; w obu przypadkach dla $n = 4, 8, 16$.

7.3. Kwadratury Gaussa

W poprzednim podrozdziale dowiedzieliśmy się, jak tworzyć kwadratury czyli przybliżone wzory całkowania numerycznego

$$\int_a^b f(x)w(x)\,dx \approx \sum_{i=0}^{n} A_i f(x_i) \tag{7.3.1}$$

(w jest tu ustaloną funkcją wagową o wartościach dodatnich), dokładne dla każdego wielomianu f stopnia $\leqslant n$. Do tej pory układ węzłów x_i był dowolny. Określają one jednoznacznie współczynniki A_i.

Nasuwa się pytanie, czy pewne układy węzłów są w jakimś sensie lepsze od innych. Można np. uznać, że dobry byłby taki wzór całkowania, w którym wszystkie współczynniki A_i byłyby równe. Istotnie, stosowanie wzoru

$$\int_a^b f(x)w(x)\,dx \approx A\sum_{i=0}^{n} f(x_i) \tag{7.3.2}$$

wymaga tylko jednego mnożenia zamiast $n+1$. W najprostszym przypadku, gdy funkcja wagowa jest stała, taki wzór, zwany *kwadraturą Czebyszewa*, istnieje tylko dla $n = 0, 1, \ldots, 6$ i $n = 8$[4]. W szczególności dla $n = 4$ ma on postać

$$\int_{-1}^{1} f(x)\,dx \approx \frac{2}{5}[f(-\alpha) + f(-\beta) + f(0) + f(\beta) + f(\alpha)],$$

gdzie

$$\alpha := \sqrt{(5+\sqrt{11})/12} \approx 0.83249\,74870\,00982,$$
$$\beta := \sqrt{(5-\sqrt{11})/12} \approx 0.37454\,14095\,53581.$$

Węzły α i β wyznaczamy z żądania dokładności wzoru dla wszystkich $f \in \Pi_4$ (ale jest on dokładny nawet dla $f \in \Pi_5$; zob. zad. 7.2.12).

[4] Udowodnili to Czebyszew i S.N. Bernstein; zob. Natanson [*1949] (*przyp. tłum.*).

Dla funkcji wagowej $w(x) := (1-x^2)^{-1/2}$ wzór (7.3.2) istnieje dla dowolnego $n \geqslant 0$, ma postać

$$\int_{-1}^{1} f(x)(1-x^2)^{-1/2}\,dx \approx \frac{\pi}{n+1}\sum_{i=0}^{n} f\left(\cos\frac{(2i+1)\pi}{2(n+1)}\right) \tag{7.3.3}$$

i jest dokładny dla każdej funkcji $f \in \Pi_{2n+1}$; nazywamy go *kwadraturą Hermite'a.*

Kwadratury Gaussa

Zwróćmy uwagę na to, że wzór przybliżony (7.3.3) całkowania w przedziale $[-1,1]$, z funkcją wagową $(1-x^2)^{-1/2}$, jest dokładny w $(2n+2)$-wymiarowej przestrzeni Π_{2n+1}, chociaż węzłów jest tylko $n+1$. Ogólniej, niech będą dane przedział $[a,b]$ i funkcja wagowa w dodatnia w tym przedziale. Ze wzoru interpolacyjnego Lagrange'a wynika wzór przybliżony całkowania (7.3.1) o współczynnikach

$$A_i := \int_a^b w(x)\prod_{j=0,\,j\neq i}^{n}\frac{x-x_j}{x_i-x_j}\,dx \quad (0 \leqslant i \leqslant n). \tag{7.3.4}$$

Nazywamy go *kwadraturą Gaussa*, jeśli jest dokładny dla każdego $f \in \Pi_{2n+1}$. Właśnie Gauss (1777–1855) sformułował i rozwiązał zadanie poszukiwania takiego wzoru dla $w \equiv 1$.

Twierdzenie 7.3.1. *Jeśli węzły $x_0, x_1, \ldots, x_n$ są zerami $(n+1)$-szego wielomianu ortogonalnego p_{n+1} w przedziale $[a,b]$ z wagą w, to kwadratura* (7.3.1) *o współczynnikach* (7.3.4) *jest dokładna dla każdej funkcji $f \in \Pi_{2n+1}$.*

Dowód. Niech q będzie ilorazem, a r resztą z dzielenia wielomianu f klasy Π_{2n+1} przez p_{n+1}:

$$f = p_{n+1}q + r \quad (q, r \in \Pi_n).$$

Stąd $f(x_i) = r(x_i)$. Ponieważ wzór (7.3.1) jest z założenia dokładny dla wszystkich wielomianów z Π_n, a wielomian p_{n+1} jest ortogonalny względem tychże wielomianów, więc

$$\int_a^b f(x)w(x)\,dx = \int_a^b r(x)w(x)\,dx = \sum_{i=0}^{n} A_i r(x_i) = \sum_{i=0}^{n} A_i f(x_i). \qquad \blacksquare$$

Sensowność kwadratur Gaussa wynika m.in. stąd, że – jak wynika z następnego twierdzenia – każdy wielomian ortogonalny w danym przedziale ma tylko zera rzeczywiste, pojedyncze i leżące w jego wnętrzu.

Twierdzenie 7.3.2. *Jeśli funkcja $f \in C[a,b]$ jest ortogonalna w tym przedziale z wagą w względem wszystkich wielomianów klasy Π_n, to w (a,b) zmienia znak co najmniej $n+1$ razy.*

Dowód. Funkcja stała należy do zbioru Π_n, więc $\int_a^b f(x)w(x)\,dx = 0$, skąd wynika, że f zmienia znak co najmniej raz. Przypuśćmy, że f zmienia znak tylko r razy, gdzie $r \leqslant n$. Istnieją zatem punkty t_i takie, że

$$a = t_0 < t_1 < \ldots < t_r < t_{r+1} = b$$

i że w każdym z $r+1$ przedziałów (t_0,t_1), $(t_1,t_2),\ldots,(t_r,t_{r+1})$ funkcja f jest albo niedodatnia, albo nieujemna. Wielomian $p(x) := \prod_{i=1}^{r}(x-t_i)$ stopnia $r \leqslant n$ ma tę samą własność, czyli powinno być $\int_a^b f(x)p(x)w(x)\,dx \neq 0$, ale to przeczy założeniu. ■

Przypadek $[a,b] = [-1,1]$ i $w \equiv 1$ zbadał Gauss. Oto dwie kwadratury odnoszące się do tego przypadku, odpowiednio dla $n = 1$ i $n = 4$:

$$\int_{-1}^{1} f(x)\,dx \approx f\left(-1/\sqrt{3}\right) + f(\left(1/\sqrt{3}\right),$$

$$\int_{-1}^{1} f(x)\,dx \approx \sum_{i=0}^{4} A_i f(x_i),$$

gdzie

$$-x_0 = x_4 = \frac{1}{3}\sqrt{5+2\sqrt{10/7}} \approx 0.90617\,98459\,38664,$$

$$-x_1 = x_3 = \frac{1}{3}\sqrt{5-2\sqrt{10/7}} \approx 0.53846\,93101\,05683,$$

$$x_2 = 0,$$

$$A_0 = A_4 = \frac{1}{450}(161 - 65\sqrt{0.7}) \approx 0.23692\,68850\,56189,$$

$$A_1 = A_3 = \frac{1}{450}(161 + 65\sqrt{0.7}) \approx 0.47862\,86704\,99366,$$

$$A_2 = \frac{128}{225} \approx 0.56888\,88888\,88889.$$

Podane tu węzły są zerami wielomianów Legendre'a (podrozdz. 6.8), odpowiednio $p_2(x) = x^3 - \frac{1}{3}x$ i $p_5(x) = x^5 - \frac{10}{9}x^3 + \frac{5}{21}x$.

Węzły i współczynniki wielu wzorów całkowania przybliżonego podają Abramowitz i Stegun [1964] oraz Krylov i Šulgina [*1966]. Dla dostatecznie regularnych funkcji f występujących w kwadraturach Gaussa nawet nieduże n zapewnia rozsądną dokładność.

Zbieżność i błąd

Lemat 7.3.3. *Współczynniki dowolnej kwadratury Gaussa są dodatnie.*

Z tej własności kwadratur Gaussa korzystamy w dowodzie tw. 7.3.4, ale ma ona też praktyczne znaczenie: można się spodziewać, że błędy zaokrągleń mają mniejszy wpływ na obliczaną sumę $\sum_{i=0}^{n} A_i f(x_i)$ właśnie wtedy, gdy wszystkie A_i są dodatnie (dodajmy, że nie w każdym wzorze Newtona-Cotesa tak jest).

Dowód. Niech p_{n+1} będzie $(n+1)$-szym wielomianem ortogonalnym w $[a,b]$ z wagą w, a punkty $x_0, x_1, \ldots, x_n$ – jego zerami, czyli węzłami w (7.3.1). Niech dla ustalonego j będzie $q(x) := p_{n+1}(x)/(x-x_j)$. Ponieważ $q^2 \in \Pi_{2n}$, więc ta kwadratura dla q^2 jest dokładna:

$$0 < \int_a^b q^2(x)w(x)\,dx = \sum_{i=0}^{n} A_i q^2(x_i) = A_j q^2(x_j). \qquad \blacksquare$$

Stieltjes udowodnił poniższe twierdzenie o zbieżności ciągu kwadratur Gaussa. Węzły i współczynniki wzoru przybliżonego Gaussa (7.3.1) oznaczymy teraz odpowiednio x_{ni} i A_{ni}:

Twierdzenie 7.3.4 (Stieltjes). *Jeśli $f \in C[a,b]$, to*

$$\lim_{n\to\infty} \sum_{i=0}^{n} A_{ni} f(x_{ni}) = \int_a^b f(x)w(x)\,dx.$$

Dowód. Niech będzie $\varepsilon > 0$. Na mocy tw. Weierstrassa 6.1.11 istnieje taki wielomian p, że $|f(x) - p(x)| < \varepsilon$ dla $x \in [a,b]$. Dla dowolnego n większego od stopnia tego wielomianu wzór Gaussa jest dokładny. Wtedy

$$\begin{aligned}
&\left|\int_a^b f(x)w(x)\,dx - \sum_{i=0}^{n} A_{ni} f(x_{ni})\right| \leqslant \\
&\leqslant \left|\int_a^b f(x)w(x)\,dx - \int_a^b p(x)w(x)\,dx\right| + \left|\sum_{i=0}^{n} A_{ni}p(x_{ni}) - \sum_{i=0}^{n} A_{ni} f(x_{ni})\right| \leqslant \\
&\leqslant \int_a^b |f(x) - p(x)|w(x)\,dx + \sum_{i=0}^{n} A_{ni}|p(x_{ni}) - f(x_{ni})| \leqslant \\
&\leqslant \varepsilon \int_a^b w(x)\,dx + \varepsilon \sum_{i=0}^{n} A_{ni} = 2\varepsilon \int_a^b w(x)\,dx.
\end{aligned}$$

Ostatnia równość wynika stąd, że wszystkie wzory przybliżonego całkowania wynikające ze wzorów interpolacyjnych są dokładne dla każdej stałej. $\blacksquare$

Twierdzenie 7.3.5. *Jeśli $f \in C^{2n}[a,b]$, to kwadratura Gaussa z n węzłami ma tę własność, że*

$$\int_a^b f(x)w(x)\,dx = \sum_{i=0}^{n-1} A_i f(x_i) + \frac{f^{(2n)}(\xi)}{(2n)!}\int_a^b q^2(x)w(x)\,dx,$$

gdzie $\xi \in (a,b)$ i $q(x) = \prod_{i=0}^{n-1}(x - x_i)$.

Dowód. Jak wiadomo z teorii interpolacji Hermite'a (podrozdz. 6.3), istnieje wielomian $p \in \Pi_{2n-1}$ taki, że

$$p(x_i) = f(x_i), \quad p'(x_i) = f'(x_i) \quad (0 \leqslant i \leqslant n-1).$$

Zgodnie z tw. 6.3.7 i wobec założenia o f

$$f(x) - p(x) = \frac{f^{(2n)}(\xi_x)}{(2n)!}\,q^2(x).$$

Dlatego

$$\int_a^b f(x)w(x)\,dx - \int_a^b p(x)w(x)\,dx = \frac{1}{(2n)!}\int_a^b f^{(2n)}(\xi_x)q^2(x)w(x)\,dx.$$

Ponieważ $p \in \Pi_{2n-1}$, więc

$$\int_a^b p(x)w(x)\,dx = \sum_{i=0}^{n-1} A_i p(x_i) = \sum_{i=0}^{n-1} A_i f(x_i).$$

Natomiast twierdzenie o wartości średniej dla całek daje równość

$$\int_a^b f^{(2n)}(\xi_x)q^2(x)w(x)\,dx = f^{(2n)}(\xi)\int_a^b q^2(x)w(x)\,dx. \qquad \blacksquare$$

Więcej informacji o kwadraturach Gaussa podają Krylov [1962], Davis i Rabinowitz [1984], Stroud i Secrest [1966], Abramowitz i Stegun [1964] oraz Natanson [*1949].

Zadania 7.3

1. Znaleźć takie α, dla którego wzór przybliżony $\int_{-1}^{1} f(x)\,dx \approx f(-\alpha) + f(\alpha)$ jest dokładny dla każdego $f \in \Pi_2$. Czy jest on dokładny również dla wielomianów stopnia trzeciego?
2. Udowodnić, że dla przedziału $[-b,b]$ i funkcji wagowej parzystej węzły w kwadraturze Gaussa są rozmieszczone symetrycznie względem 0, a współczynniki są takie, że $A_i = A_{n-i}$ $(0 \leqslant i \leqslant n)$.

3. Udowodnić, że żadna kwadratura Gaussa z n węzłami nie może być dokładna w całej klasie Π_{2n}.

4. Jak wpłynęłoby na teorię kwadratur Gaussa odrzucenie założenia, że funkcja wagowa jest dodatnia?

5. Czy tw. 7.3.1 pozostanie prawdziwe, gdy zamiast p_{n+1} wystąpi w nim wielomian ortogonalny p_m dla $m > n+1$? (Węzły $x_0, x_1, \ldots, x_n$ mają być jego niektórymi zerami).

6. Udowodnić, że jeśli $\int_a^b f(x)w(x)\,dx = \sum_{i=0}^n A_i f(x_i)$ dla każdego $f \in \Pi_{2n+1}$, to wielomian $\prod_{i=0}^n (x - x_i)$ jest $(n+1)$-szym wielomianem ortogonalnym w $[a,b]$ z wagą w.

7. Udowodnić, że wzór $\int_{-1}^1 f(x)\,dx \approx \sum_{i=0}^3 A_i f(x_i)$, gdzie

$$-x_0 = x_3 = \sqrt{(3+4\sqrt{0.3})/7} \approx 0.86113\,63115\,94052,$$

$$-x_1 = x_2 = \sqrt{(3-4\sqrt{0.3})/7} \approx 0.33998\,10435\,84856,$$

$$A_0 = A_3 = \frac{1}{2}\left(1 - \frac{1}{6}\sqrt{10/3}\right) \approx 0.34785\,48451\,37454,$$

$$A_1 = A_2 = \frac{1}{2}\left(1 + \frac{1}{6}\sqrt{10/3}\right) \approx 0.65214\,51548\,62546,$$

jest kwadraturą Gaussa.

8. Znaleźć kwadraturę Gaussa $\int_0^1 xf(x)\,dx \approx \sum_{i=0}^n A_i f(x_i)$: (**a**) dla $n=1$, dokładną w Π_3, (**b**) dla $n=2$, dokładną w Π_5. (**c**) Dla $n=4$ znaleźć współczynniki wielomianu o zerach $x_0, x_1, \ldots, x_4$, które są węzłami takiej kwadratury.

9. Znaleźć takie A, B, C, żeby wzór całkowania numerycznego

$$\int_{-1}^1 xf(x)\,dx \approx Af(-1) + Bf(0) + Cf(1)$$

był dokładny dla wielomianów możliwie najwyższego stopnia. Jaki jest ten stopień?

10. Znaleźć wzór przybliżony $\int_{-1}^1 x^2 f(x)\,dx \approx \sum_{i=0}^n A_i f(x_i)$: (**a**) dla $n=1$, dokładny w Π_3, (**b**) dla $n=2$, dokładny w Π_5.

11. Znaleźć wielomian ortogonalny względem Π_2 w $[-1,1]$ z wagą $1+x^2$.

12. Znaleźć współczynniki i węzły kwadratury Gaussa

$$\int_a^b x^4 f(x)\,dx \approx A_0 f(x_0) + A_1 f(x_1)$$

dla: (**a**) $[a,b] = [0,1]$, (**b**) $[a,b] = [-1,1]$.

13. Znaleźć wielomian, którego zera x_0, x_1, x_2 są węzłami kwadratury Gaussa

$$\int_1^2 (x^4 - 1)f(x)\,dx \approx A_0 f(x_0) + A_1 f(x_1) + A_2 f(x_2).$$

14. Znaleźć kwadratury Gaussa dla przedziału $[-1,1]$ i funkcji wagowej $\sqrt{1-x^2}$. Wskazówka: Skorzystać z zad. 7.2.20.

15. Kwadraturę Gaussa

$$\int_{-1}^{1} f(x)\,dx \approx \frac{5}{9}f(-\sqrt{0.6}) + \frac{8}{9}f(0) + \frac{5}{9}f(\sqrt{0.6})$$

po odpowiednim przekształceniu przedziału całkowania zastosować do obliczenia całki $\int_0^4 t^{-1}\sin t\,dt$.

7.4. Wielomiany Bernoulliego i wzór Eulera-Maclaurina

Metoda Romberga całkowania opisana w podrozdz. 7.5 opiera się na klasycznym wzorze Eulera-Maclaurina, który tu będzie podany w tw. 7.4.3. Ten wzór z kolei wymaga zapoznania się z wielomianami Bernoulliego.

Wielomiany Bernoulliego

Wielomiany Bernoulliego B_n z definicji spełniają równanie

$$\sum_{k=0}^{n}\binom{n+1}{k}B_k(t) = (n+1)t^n \quad (n \geqslant 0). \tag{7.4.1}$$

Pozwala ono je wyznaczyć kolejno dla $n = 0, 1, \ldots$:

$$B_0(t) = 1, \qquad B_1(t) = t - \tfrac{1}{2},$$
$$B_2(t) = t^2 - t + \tfrac{1}{6}, \qquad B_3(t) = t^3 - \tfrac{3}{2}t^2 + \tfrac{1}{2}t, \ldots$$

TWIERDZENIE 7.4.1. *Wielomiany Bernoulliego mają następujące własności:*

1. $B_n' = nB_{n-1}$.

2. $B_n(t+1) - B_n(t) = nt^{n-1}$.

3. $B_n(t) = \sum_{k=0}^{n}\binom{n}{k}B_k(0)t^{n-k}$.

4. $B_n(1-t) = (-1)^n B_n(t)$.

Dowód. Własność **1** udowodnimy przez indukcję. Dla $n = 1$ jest ona oczywista. Jeśli $n > 1$ i $B_k' = kB_{k-1}$ dla $k < n$, to różniczkując stronami (7.4.1) wnioskujemy, że

$$\sum_{k=1}^{n}\binom{n+1}{k}B_k'(t) = n(n+1)t^{n-1} = (n+1)\sum_{k=0}^{n-1}\binom{n}{k}B_k(t).$$

Z założenia indukcyjnego wynika, że

$$(n+1)B_n' + \sum_{k=1}^{n-1} \binom{n+1}{k} kB_{k-1} = (n+1) \sum_{k=0}^{n-1} \binom{n}{k} B_k.$$

Ponieważ dla $k > 0$

$$\frac{k}{n+1} \binom{n+1}{k} = \binom{n}{k-1},$$

więc

$$B_n' + \sum_{k=1}^{n-1} \binom{n}{k-1} B_{k-1} = \sum_{k=0}^{n-1} \binom{n}{k} B_k,$$

co po uproszczeniu daje własność **1**.

Aby sprawdzić własność **2**, korzystamy wielokrotnie z tej, którą już udowodniliśmy. Dla $k = 1, 2, \ldots$ daje to równość

$$B_n^{(k)} = n(n-1)\ldots(n-k+1)B_{n-k}.$$

Stąd i ze wzoru Taylora dla B_n wynika, że

$$B_n(t+h) = \sum_{k=0}^{n} \frac{1}{k!} B_n^{(k)}(t)h^k = \sum_{k=0}^{n} \binom{n}{k} B_{n-k}(t)h^k. \tag{7.4.2}$$

Dla $h = 1$ z równości $\binom{n}{k} = \binom{n}{n-k}$ i z (7.4.1) wynika własność **2**:

$$B_n(t+1) = \sum_{k=0}^{n} \binom{n}{k} B_k(t) = B_n(t) + \sum_{k=0}^{n-1} \binom{n}{k} B_k(t) = B_n(t) + nt^{n-1}.$$

Własność **3** wynika z (7.4.2) po zmianie t i h odpowiednio na 0 i t.

Na koniec, aby udowodnić własność **4**, korzystamy z **2**, zmieniając tam t na $-t$:

$$B_n(1-t) - B_n(-t) = n(-1)^{n-1}t^{n-1} = (-1)^{n-1}[B_n(t+1) - B_n(t)].$$

Napiszmy tę równość w postaci

$$(-1)^n B_n(t+1) - B_n(-t) = (-1)^n B_n(t) - B_n(1-t).$$

Jest to równość typu $F(t+1) = F(t)$ oznaczająca, że F jest funkcją okresową o okresie 1. Tu jednak F jest wielomianem, czyli musi być stałą:

$$(-1)^n B_n(t) - B_n(1-t) = c_n.$$

Różniczkujemy tę równość stronami i korzystamy z własności **1**:

$$(-1)^n B_n'(t) + B_n'(1-t) = (-1)^n nB_{n-1}(t) + nB_{n-1}(1-t) = 0.$$

To już daje własność **4**. ■

Lemat 7.4.2. *Jeśli $n > 0$, to wielomian $G(t) := B_{2n}(t) - B_{2n}(0)$ nie ma zer w przedziale $(0,1)$.*

Dowód. Podstawienie $t = 0$ we własnościach **2** i **4** z tw. 7.4.1 daje dla $n > 1$ równość

$$B_n(0) = B_n(1) = (-1)^n B_n(0).$$

Stąd wynika, że

$$B_{2n-1}(0) = 0 \quad (n > 1). \tag{7.4.3}$$

Przypuśćmy, że G ma zero w $(0,1)$. Wtedy, ponieważ $G(0) = G(1) = 0$, funkcja G' ma na mocy twierdzenia Rolle'a co najmniej dwa zera w $(0,1)$. Wielomian B_{2n-1} różni się od G' tylko czynnikiem stałym, więc ma tę samą własność. Ten wielomian znika jednak także w punktach 0 i 1, czyli B'_{2n-1} (a więc i B_{2n-2}) ma trzy zera w $(0,1)$. Rozumując tak dalej, dochodzimy do wniosku, że każdy wielomian B_{2k-1} dla $k < n$ ma co najmniej dwa zera w $(0,1)$, a także zera 0 i 1. To jest jednak niemożliwe dla wielomianu B_3 stopnia trzeciego. ■

Udowodnione już tw. 7.4.1 i lemat 7.4.2 prowadzą do *wzoru Eulera-Maclaurina* podanego niżej:

Twierdzenie 7.4.3. *Jeśli $f \in C^{2n}[0,1]$, to*

$$\int_0^1 f(t)\,dt = \frac{1}{2}[f(0) + f(1)] - \sum_{k=1}^{n-1} \frac{B_{2k}(0)}{(2k)!}[f^{(2k-1)}(1) - f^{(2k-1)}(0)] - \\ - \frac{B_{2n}(0)}{(2n)!} f^{(2n)}(\xi) \qquad (0 < \xi < 1).$$

Dowód. Ponieważ $B_0(t) = 1$, a na mocy tw. 7.4.1 jest

$$B_n(t) = \frac{1}{n+1} B'_{n+1}(t),$$

więc różniczkując przez części, wnioskujemy, że

$$\int_0^1 f(t)\,dt = \int_0^1 f(t)B_0(t)\,dt = B_1(t)f(t)\Big|_0^1 - \int_0^1 B_1(t)f'(t)\,dt.$$

Wiemy też, że $B_1(1) = \frac{1}{2}$ i $B_1(0) = -\frac{1}{2}$. Dlatego

$$\int_0^1 f(t)\,dt = \frac{1}{2}[f(0) + f(1)] - \int_0^1 B_1(t)f'(t)\,dt.$$

Do całki po prawej stronie stosujemy znów całkowanie przez części:

$$\int_0^1 f(t)\,dt = \frac{1}{2}[f(0)+f(1)]-\frac{1}{2}B_2(0)[f'(1)-f'(0)]+\frac{1}{2}\int_0^1 B_2(t)f''(t)\,dt.$$

Postępujemy tak dalej, korzystając ze znanych już równości $B_n(0) = B_n(1)$ dla $n \geqslant 2$ i (7.4.3). Po $2n$ krokach otrzymujemy następujący wzór:

$$\int_0^1 f(t)\,dt = \frac{1}{2}[f(0)+f(1)] - \sum_{k=1}^{n}\frac{B_{2k}(0)}{(2k)!}[f^{(2k-1)}(1) - f^{(2k-1)}(0)]+ \\ +\frac{1}{(2n)!}\int_0^1 B_{2n}(t)f^{(2n)}(t)\,dt.$$

Po prawej stronie mamy już takie składniki jak w dowodzonym wzorze Eulera-Maclaurina. Pozostaje wykazać, że składnik dla $k = n$ w sumie i ostatni składnik prawej strony, równe łącznie

$$\frac{1}{(2n)!}\int_0^1 [B_{2n}(t) - B_{2n}(0)]f^{(2n)}(t)\,dt,$$

dają resztę tego wzoru. Z lematu 7.4.2 wynika, że różnica w nawiasach kwadratowych nie zmienia znaku w $(0,1)$. Możemy więc zastosować twierdzenie o wartości średniej dla całek i przekształcić otrzymane wyrażenie do postaci

$$\frac{1}{(2n)!}f^{(2n)}(\xi)\int_0^1 [B_{2n}(t) - B_{2n}(0)]\,dt.$$

Uwzględniając jeszcze, że $B_{2n} = B'_{2n+1}/(2n+1)$ i $B_{2n+1}(0) = B_{2n+1}(1) = 0$, otrzymujemy ostateczną postać reszty wzoru Eulera-Maclaurina. ■

Zadania 7.4

1. Udowodnić, że $B_n(t) = t^n - \frac{1}{2}nt^{n-1} + \ldots$ dla $n > 0$.
2. Korzystając z własności **2** z tw. 7.4.1, wykazać, że

$$1^p + 2^p + \ldots + n^p = [B_{p+1}(n+1) - B_{p+1}(0)]/(p+1).$$

3. Wykazać, że dla n nieparzystych wielomian $B_n(x) - B_n(0)$ ma pojedyncze zero $\frac{1}{2}$.
4. Udowodnić, że wartości $B_0(0), B_2(0), B_4(0), \ldots$ są na przemian dodatnie i ujemne.
5. Udowodnić, że wielomian B_n ma w przedziale $(0,1)$ co najmniej dwa zera, jeśli n jest parzyste, i co najmniej jedno zero w przeciwnym razie.

7.5. Metoda Romberga

Opiszemy teraz metodę całkowania numerycznego, której autorem jest Romberg. Daje ona tablicę coraz lepszych przybliżeń całki $\int_a^b f(x)\,dx$.

Rekurencyjne stosowanie wzoru trapezów

Punktem wyjścia do obliczeń jest złożony wzór trapezów (7.2.4). Stosujemy go, dzieląc przedział $[a,b]$ na 2^n podprzedziałów równej długości ($n \geqslant 0$). Niech będzie

$$R(n,0) := h_n \sum_{i=0}^{2^n}{}'' f(a+ih_n), \quad \text{gdzie} \quad h_n := \frac{b-a}{2^n}$$

(przypominamy, co oznacza symbol $\sum''$: skrajne składniki sumy należy podzielić przez 2). Najprostsze wyrażenie mamy dla $n=0$:

$$R(0,0) = \tfrac{1}{2}[f(a)+f(b)]. \tag{7.5.1}$$

Sumy $R(1,0)$, $R(2,0),\ldots$ obliczamy rekurencyjnie tak, aby uniknąć wielokrotnego obliczania wartości funkcji f w tych samych punktach. Jeśli $n>0$, to w $R(n,0)$ występują m.in. wartości funkcji f, które są potrzebne do obliczenia $R(n-1,0)$; teraz trzeba je podzielić dodatkowo przez 2. Prócz tego $R(n,0)$ zawiera wartości w punktach pośrednich $a+h_n$, $a+3h_n,\ldots$:

$$R(n,0) = \frac{1}{2}R(n-1,0) + h_n \sum_{i=1}^{2^{n-1}} f(a+(2i-1)h_n). \tag{7.5.2}$$

Poprawianie przybliżeń $R(n,0)$

Przybliżenia $R(n,0)$ tworzą pierwszą kolumnę tablicy przybliżeń całki. Następne kolumny obliczamy według wzoru

$$R(n,m) := R(n,m-1) + \frac{1}{4^m-1}[R(n,m-1)-R(n-1,m-1)] \tag{7.5.3}$$

($m>0$). Aby go uzasadnić, powołujemy się na wzór Eulera-Maclaurina (tw. 7.4.3): jeśli $f \in C^{2m}[0,1]$, to

$$\int_0^1 f(t)\,dt = \frac{1}{2}[f(0)+f(1)] + \\ + \sum_{k=1}^{m-1} A_{2k}[f^{(2k-1)}(0) - f^{(2k-1)}(1)] - A_{2m} f^{(2m)}(\xi_0),$$

gdzie $\xi_0 \in (0,1)$. Związek współczynników A_{2k} z liczbami Bernoulliego nie jest tu istotny. Po odpowiednim przekształceniu zmiennej wynika stąd, że

$$\int_{x_i}^{x_{i+1}} f(x)\,dx = \frac{1}{2}h_n[f(x_i)+f(x_{i+1})]+ \tag{7.5.4}$$
$$+\sum_{k=1}^{m-1} A_{2k}h_n^{2k}[f^{(2k-1)}(x_i)-f^{(2k-1)}(x_{i+1})] - A_{2m}h_n^{2m+1}f^{(2m)}(\xi_i),$$

gdzie $\xi_i \in (x_i, x_{i+1})$.

Te równości dla $i = 0, 1, \ldots, 2^n - 1$ sumujemy stronami:

$$\int_a^b f(x)\,dx = \frac{1}{2}h_n \sum_{i=0}^{2^n-1}[f(x_i)+f(x_{i+1}]+ \tag{7.5.5}$$
$$+\sum_{k=1}^{m-1} A_{2k}h_n^{2k}[f^{(2k-1)}(a)-f^{(2k-1)}(b)] - A_{2m}(b-a)h_n^{2m}f^{(2m)}(\xi),$$

gdzie $\xi \in (a,b)$. Tematem zad. 1 jest wykazanie, że błąd można wyrazić tak jak to zrobiono wyżej. Tak więc

$$\int_a^b f(x)\,dx = R(n,0)+c_2h_n^2+c_4h_n^4+\ldots+c_{2m-2}h_n^{2m-2}+c_{2m}h_n^{2m}f^{(2m)}(\xi).$$

Tego samego typu równość (7.1.5) była podstawą ekstrapolacji Richardsona. Możemy więc i tu zastosować wyniki podrozdz. 7.1. Wzór rekurencyjny (7.1.7) przybiera teraz postać (7.5.3).

Algorytm

Metoda Romberga wykorzystuje wzór złożony trapezów i ekstrapolację Richardsona, a dokładniej – wzory (7.5.1), (7.5.2) i (7.5.3). Pozwalają one obliczać, np. kolejnymi wierszami, elementy tablicy trójkątnej

$$\begin{array}{llll} R(0,0) & & & \\ R(1,0) & R(1,1) & & \\ R(2,0) & R(2,1) & R(2,2) & \\ \vdots & \vdots & \vdots & \ddots \\ R(M,0) & R(M,1) & R(M,2) & \ldots \quad R(M,M) \end{array} \tag{7.5.6}$$

Odpowiedni algorytm jest następujący:

input a, b, M
$h \leftarrow b - a$
$R(0,0) \leftarrow \frac{1}{2}(b-a)[f(a)+f(b)]$
for $n = 1$ **to** M **do**
  $h \leftarrow h/2$
  $R(n,0) \leftarrow \frac{1}{2}R(n-1,0) + h\sum_{i=1}^{2^{n-1}} f(a+(2i-1)h)$
  for $m = 1$ **to** n **do**
    $R(n,m) \leftarrow R(n,m-1) + [R(n,m-1) - R(n-1,m-1)]/(4^m - 1)$
    output $n, m, R(n,m)$
  end do
end do

Potrzebna wartość M jest zwykle niezbyt duża – pamiętajmy, że algorytm wymaga obliczenia $2^M + 1$ wartości funkcji. Bardziej wyrafinowany algorytm powinien uwzględniać możliwość przerwania obliczeń po spełnieniu odpowiednio dobranego kryterium dokładności. Nie jest też konieczne stosowanie tablicy R z dwoma wskaźnikami.

Zbieżność metody Romberga

Metoda Romberga jest szczególnie skuteczna, gdy funkcja f ma dostatecznie wiele pochodnych. Jeśli $f \in C^{2m}[a,b]$, to wielkości $R(n,m)$ przybliżają całkę z błędem rzędu $\mathcal{O}(h^{2m})$, gdzie $h := 2^{-n}(b-a)$. Jeśli jednak wiemy tylko, że f jest ciągła, to i wtedy ta metoda daje sensowne wyniki, bo każda kolumna w tablicy trójkątnej (7.5.6) jest zbieżna do całki z f:

Twierdzenie 7.5.1. *Jeśli $f \in C[a,b]$, to*

$$\lim_{n\to\infty} R(n,m) = \int_a^b f(x)\,dx.$$

Dowód. Pierwsza kolumna tej tablicy zawiera przybliżenia, które dla odpowiednich k można wyrazić w postaci

$$h{\sum_{i=0}^{k}}'' f(a+ih) = \frac{1}{2}h\sum_{i=0}^{k-1} f(a+ih) + \frac{1}{2}h\sum_{i=1}^{k} f(a+ih),$$

gdzie $h := (b-a)/k$. Po prawej stronie mamy tu średnią arytmetyczną dwóch sum riemannowskich dla rozważanej całki I. Z teorii całki Riemanna wynika, że każda z tych sum dąży do I, gdy $k \to \infty$. Druga kolumna zawiera wielkości

$$R(n,1) = \frac{4}{3}R(n,0) - \frac{1}{3}R(n-1,0),$$

więc $\lim_{n\to\infty} R(n,1) = \frac{4}{3}I - \frac{1}{3}I = I$. Tak samo rozumujemy dla następnych kolumn. ■

ZADANIA 7.5

1. Sprawdzić, że (7.5.5) wynika z (7.5.4).
2. Wykazać, że druga kolumna tablicy (7.5.6) zawiera liczby, które daje złożony wzór Simpsona i że $R(3,3)$ nie jest związane z tym wzorem (jest to tzw. *wzór Milne'a*).
3. Metodą Romberga obliczyć $R(2,2)$ dla $\int_1^3 x^{-1}\,dx$.

ZADANIA KOMPUTEROWE 7.5

K1. Napisać procedurę realizującą metodę Romberga dla danej funkcji f i przedziału $[a,b]$. Parametrem procedury powinna być też liczba wierszy tablicy (7.5.6), które mają być obliczone. Sprawdzić procedurę dla całek:
(**a**) $\int_0^1 x^{-1}\sin x\,dx$, (**b**) $\int_{-1}^1 (\cos x - e^x)/\sin x\,dx$, (**c**) $\int_1^\infty x^{-1}e^{-x}\,dx$.
Funkcję podcałkową należy obliczać tak, aby uniknąć utraty dokładności przy odejmowaniu bliskich liczb. W razie potrzeby trzeba skorzystać z relacji $f(x_0) = \lim_{x\to x_0} f(x)$. W przypadku (c) przez odpowiednie przekształcenie zmiennej, np. $x = 1/t$, otrzymać przedział skończony całkowania. Obliczyć siedem wierszy tablicy (7.5.6).

7.6. Metody adaptacyjne całkowania

Istotą metod adaptacyjnych całkowania jest to, że własności funkcji podcałkowej w różnych częściach przedziału są automatycznie uwzględniane w obliczeniach; podobne metody aproksymacji naszkicowano w podrozdz. 6.14. Stosując taką metodę do obliczania całki

$$\int_a^b f(x)\,dx$$

użytkownik podaje – w idealnym przypadku – tylko funkcję f, przedział całkowania $[a,b]$ i dopuszczalny błąd ε szukanej wartości. Natomiast sama metoda dzieli ten przedział na takie fragmenty, aby stosowanie w nich jakiejś klasycznej metody dało w sumie dostatecznie dokładny wynik.

Opiszemy teraz dokładniej typową metodę adaptacyjną. Załóżmy, że $f \in C^4[a,b]$, dzięki czemu można stosować wzór Simpsona (7.2.6) z błędem (7.2.7):

$$\int_u^v f(x)\,dx = S(u,v) - \frac{1}{2880}(v-u)^5 f^{(4)}(\xi), \tag{7.6.1}$$

gdzie

$$S(u,v) := \frac{1}{6}\left[f(u) + 4f\left(\frac{u+v}{2}\right) + f(v)\right], \quad \xi \in (u,v).$$

Jeśli w pewnym przedziale $[u, v]$ wzór Simpsona nie jest dostatecznie dokładny, to przedział dzielimy na połowy i tenże wzór stosujemy w każdej z nich. W razie potrzeby to postępowanie iterujemy aż do osiągnięcia odpowiedniej globalnej dokładności. Ostatecznie więc dochodzimy do równości

$$\int_a^b f(x)\,dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)\,dx = \sum_{i=1}^{n} S(x_{i-1}, x_i) + \sum_{i=1}^{n} e_i,$$

gdzie e_i jest błędem przybliżenia $S(x_{i-1}, x_i)$ całki w odpowiednim podprzedziale. Jeśli

$$|e_i| \leqslant \varepsilon \frac{x_i - x_{i-1}}{b-a}, \tag{7.6.2}$$

to globalny błąd można oszacować tak:

$$\Big|\sum_{i=1}^{n} e_i\Big| \leqslant \sum_{i=1}^{n} |e_i| \leqslant \frac{\varepsilon}{b-a} \sum_{i=1}^{n} (x_i - x_{i-1}) = \varepsilon.$$

Jeśli w pewnym przedziale $[u, v]$ wzór (7.6.1) jest za mało dokładny, to – jak już zapowiedziano – przedział dzielimy na połowy za pomocą punktu $w := (u+v)/2$ i w każdej z nich stosujemy ten sam wzór:

$$\begin{aligned} \int_u^v f(x)\,dx &= \int_u^w f(x)\,dx + \int_w^v f(x)\,dx = \\ &= S(u,w) + S(w,v) - \frac{1}{2880}(w-u)^5 f^{(4)}(\xi_1) - \frac{1}{2880}(v-w)^5 f^{(4)}(\xi_2) = \\ &= S(u,w) + S(w,v) - \frac{1}{46080}(v-u)^5 f^{(4)}(\xi'), \end{aligned} \tag{7.6.3}$$

gdzie $\xi' \in (u, v)$. Ostatnia równość jest prawdziwa dzięki założonej ciągłości pochodnej $f^{(4)}$. W metodzie adaptacyjnej przyjmujemy, że zmienia się ona na tyle wolno, iż można ją uznać za stałą w dostatecznie małych przedziałach. To pozwala wyeliminować składniki z tą pochodną występujące w (7.6.1) i (7.6.3). W tym celu mnożymy drugą równość przez 16 i odejmujemy stronami pierwszą:

$$\int_u^v f(x)\,dx \approx S(u,w) + S(w,v) + \frac{1}{15}[S(u,w) + S(w,v) - S(u,v)] \tag{7.6.4}$$

(takie postępowanie znamy już z ekstrapolacji Richardsona).

Mamy teraz trzy przybliżenia tej całki: zapewne najgorsze $S(u,v)$, lepsze $S(u,w) + S(w,v)$ i najlepsze, czyli całą prawą stronę równości (7.6.4). Jej ostatni składnik jest z grubsza błędem przybliżenia $S(u,w) + S(w,v)$,

ale stosujemy go jako (na ogół przesadnie pesymistyczne) oszacowanie błędu całej prawej strony. Dlatego, zgodnie z (7.6.2), uznajemy, że to najlepsze przybliżenie jest dostatecznie dokładne, jeśli

$$|S(u,w) + S(w,v) - S(u,v)| \leqslant 15(v-u)/(b-a). \tag{7.6.5}$$

Znamy już pomysł zastosowany w metodzie adaptacyjnej. Obliczenia zaczynają się od przedziału $[a,b]$. Dla niego (i dla wszystkich podprzedziałów rozważanych dalej) zapamiętujemy w tzw. *stosie*, na początku pustym, następujący wektor o sześciu składowych:

$$v := (a, h, f(a), f(a+h), f(a+2h), S(a,b)), \quad \text{gdzie} \quad h := \tfrac{1}{2}(b-a). \tag{7.6.6}$$

Przybliżona wartość $S(a,b)$ całki wyraża się przez składowe od trzeciej do piątej tego wektora. Następnie obliczamy $c := a+h$, $S(a,c)$ i $S(c,b)$. Aby ocenić, czy można już zakończyć obliczenia, sprawdzamy nierówność (7.6.5) z a, b, c zamiast u, v, w. Jeśli jest ona spełniona, to sumę podobną do prawej strony (7.6.4) uznajemy za dostatecznie dobre przybliżenie wartości całki i kończymy obliczenia. W przeciwnym razie wektor (7.6.6) usuwamy ze stosu, a dołączamy doń dwa nowe wektory:

$$(a, h/2, f(a), f(y), f(c), S(a,c)), \quad \text{gdzie} \quad y := a + h/2,$$
$$(c, h/2, f(c), f(z), f(b), S(c,b)), \quad \text{gdzie} \quad z := c + h/2.$$

Zauważmy, że wymaga to obliczenia tylko dwóch nowych wartości funkcji, w punktach y i z.

W dalszym ciągu postępujemy podobnie, wystarczy zatem dodać tylko kilka informacji:

- W każdym momencie obliczeń stos zawiera dane tylko o tych podprzedziałach, w których jeszcze nie znaleziono dostatecznie dobrych przybliżeń całki.
- Dopóki stos nie jest pusty, algorytm bada przybliżenia całki w przedziale opisanym za pomocą ostatniego wektora.
- Użytkownik ustala maksymalną liczbę n wektorów na stosie; jej przekroczenie świadczy o niemożności obliczenia całki z wymaganą dokładnością.
- W algorytmie występuje zmienna $\mathfrak{S}$, której początkową wartością jest 0 i do której dodajemy każde nowe, dostatecznie dobre przybliżenie całki w badanym podprzedziale. W razie pomyślnego zakończenia obliczeń wartość tej zmiennej jest szukanym przybliżeniem dla całki $\int_a^b f(x)\,dx$.

Niżej podano algorytm zgodny z już podanymi informacjami. Przyjęto w nim, że stos zawiera wektory $v^{(1)}, v^{(2)}, \ldots$; składowymi wektora $v^{(k)}$ są liczby $v_1^{(k)}, \ldots, v_6^{(k)}$.

input a, b, ε, n
$d \leftarrow b - a$; $\mathfrak{S} \leftarrow 0$; $h \leftarrow d/2$; $c \leftarrow (a+b)/2$; $k \leftarrow 1$
$A \leftarrow f(a)$; $B \leftarrow f(b)$; $C \leftarrow f(c)$
$S \leftarrow h(A + 4C + B)/3$
$v^{(1)} \leftarrow (a, h, A, C, B, S)$
while $1 \leqslant k \leqslant n$
    $h \leftarrow v_2^{(k)}/2$
    $Y \leftarrow f(v_1^{(k)} + h)$
    $S' \leftarrow h(v_3^{(k)} + 4Y + v_4^{(k)})/3$
    $Z \leftarrow f(v_1^{(k)} + 3h)$
    $S'' \leftarrow h(v_4^{(k)} + 4Z + v_5^{(k)})/3$
    $\delta \leftarrow S' + S'' - v_6^{(k)}$
    if $|\delta| < 60\varepsilon h/d$ **then**
        $\mathfrak{S} \leftarrow \mathfrak{S} + S' + S'' + \delta/15$
        $k \leftarrow k - 1$
        if $k = 0$ **then output** $\mathfrak{S}$; **exit**
    else
        if $k = n$ **then output** niepowodzenie; **exit**
        $V \leftarrow v_5^{(k)}$
        $v^{(k)} \leftarrow (v_1^{(k)}, h, v_3^{(k)}, Y, v_4^{(k)}, S')$
        $k \leftarrow k + 1$
        $v^{(k)} \leftarrow (v_1^{(k-1)} + 2h, h, v_5^{(k-1)}, Z, V, S'')$
    end if
end while

W tej wersji algorytmu stos występuje w jawnej postaci. W programie opartym na takiej wersji składowej $v_j^{(k)}$ odpowiada element dwuwymiarowej tablicy. W pewnych językach programowania jest istotne, żeby j było pierwszym wskaźnikiem, bo wtedy składowe ustalonego wektora są pamiętane w sąsiednich komórkach pamięci. Nie musimy jednak w ogóle korzystać z tej wersji, bo metoda adaptacyjna jest w istocie rekursywna, a w większości języków programowania procedura może wywoływać nią samą. Tu zaczynamy od przybliżenia całki w całym przedziale $[a, b]$. Jeśli to nie zapewnia zadowalającej dokładności, wywołujemy dwukrotnie tę samą procedurę, odpowiednio dla lewej i prawej połowy tego przedziału. Idea metody adaptacyjnej staje się wtedy bardziej widoczna.

ZADANIA 7.6

1. Kryterium (7.6.2) dotyczy lokalnego błędu bezwzględnego. Zaprojektować takie kryterium, które pozwoliłoby uzyskać zadowalająco mały błąd względny:

$$\left|\int_a^b f(x)\,dx - \sum_{i=1}^{n} S(x_{i-1}, x_i)\right| \leqslant \varepsilon \left|\int_a^b f(x)\,dx\right|.$$

2. Niech metoda adaptacyjna opiera się na wzorze trapezów:

$$\int_u^v f(x)\,dx = T(u,v) - \frac{1}{12}(v-u)^3 f''(\xi), \quad T(u,v) := \frac{1}{2}(v-u)[f(u)+f(v)].$$

Znaleźć dla tego przypadku odpowiednik wzoru (7.6.4).

3. Czy złożony wzór trapezów może być podstawą metody adaptacyjnej całkowania?

ZADANIA KOMPUTEROWE 7.6

K1. Zaprogramować podany w tekście algorytm metody adaptacyjnej i sprawdzić go dla całek
(**a**) $\int_0^1 x^{1/2}\,dx$, (**b**) $\int_0^1 (1-x)^{1/2}\,dx$, (**c**) $\int_0^1 (1-x)^{1/4}\,dx$.

K2. Zaprogramować metodę adaptacyjną jako procedurę rekursywną.

7.7. Teoria Sarda aproksymacji funkcjonałów

Funkcjonał liniowy w przestrzeni liniowej jest jej liniowym odwzorowaniem na przestrzeń skalarów, którą w tej książce jest zwykle $\mathbb{R}$. Ważnym przykładem funkcjonału liniowego w przestrzeni $C[a,b]$ jest całka z funkcji f:

$$\varphi(f) := \int_a^b f(x)\,dx \quad (f \in C[a,b]). \tag{7.7.1}$$

W praktyce numerycznej jedyne funkcjonały, które można bezpośrednio obliczać, odwzorowują funkcję f na kombinację liniową jej wartości w ustalonych punktach:

$$\psi = \sum_{i=0}^{n} c_i \widehat{x}_i, \quad \text{gdzie} \quad \hat{x}(f) := f(x).$$

Inne funkcjonały, np. (7.7.1), trzeba przybliżać za pomocą funkcjonałów podobnych do ψ. To właśnie się robi w metodach różniczkowania i całkowania numerycznego.

Spójną teorię aproksymacji funkcjonałów opracował Arthur Sard w latach 1940–1970. Jego teoria obejmuje w ciekawy sposób naturalne funkcje sklejane.

Funkcjonały, jakie będziemy aproksymować, są określone na przestrzeni $C^N[a,b]$ i mają postać

$$\varphi(f) := \sum_{i=0}^{N}\Big[\int_a^b \alpha_i(x)f^{(i)}(x)\,dx + \sum_{j=1}^{l}\beta_{ij}f^{(i)}(z_{ij})\Big]. \tag{7.7.2}$$

Funkcje α_i są z założenia przedziałami ciągłe w przedziale $[a,b]$, l jest dowolną liczbą całkowitą nieujemną, a punkty z_{ij} należą do tegoż przedziału.

Mówimy, że funkcjonał φ *anihiluje* przestrzeń W, jeśli $\varphi(f) = 0$ dla każdego $f \in W$. *Jądrem Peana* funkcjonału (7.7.2) jest każda z funkcji

$$K_m(t) := \frac{1}{m!}\varphi_x[(x-t)_+^m],$$

gdzie $m \geqslant N$; wskaźnik x oznacza, że funkcjonał działa na funkcję zmiennej x, a

$$(x-t)_+^m := \begin{cases}(x-t)^m & (x \geqslant t)\\ 0 & (x < t).\end{cases}$$

Przykład 7.7.1. Funkcjonał φ taki, że

$$\varphi(f) := \int_0^{\pi}(\cos x)f'(x)\,dx,$$

wynika z (7.7.2) dla $N = 1$, $\alpha_0(x) := 0$, $\alpha_1(x) := \cos x$ i $l = 0$. Jakie jest w tym przypadku jądro Peana K_1?

Rozwiązanie. Ponieważ

$$\frac{d}{dx}(x-t)_+^m = m(x-t)_+^{m-1} \quad (m \geqslant 1),$$

więc

$$K_1(t) = \varphi_x[(x-t)_+^1] = \int_0^{\pi}(\cos x)\frac{d}{dx}(x-t)_+^1\,dx =$$
$$= \int_0^{\pi}(\cos x)(x-t)_+^0\,dx = \int_t^{\pi}\cos x\,dx = -\sin t. \qquad \blacksquare$$

Poniższe twierdzenie udowodniono w 1905 r.

Twierdzenie 7.7.2. *Jeśli funkcjonał określony w* (7.7.2) *anihiluje* Π_m, *to dla każdej funkcji* $f \in C^{m+1}[a,b]$, *gdzie* $m \geqslant N$, *zachodzi równość*

$$\varphi(f) = \int_a^b K_m(t)f^{(m+1)}(t)\,dt.$$

Dowód. Z twierdzenia 1.1.7 wynika, że

$$f(x) = \sum_{k=0}^{m} \frac{1}{k!} f^{(k)}(a)(x-a)^k + r(x),$$

gdzie

$$r(x) := \frac{1}{m!} \int_a^x f^{(m+1)}(t)(x-t)^m \, dt = \frac{1}{m!} \int_a^b f^{(m+1)}(t)(x-t)_+^m \, dt.$$

Ponieważ φ anihiluje Π_m, więc $\varphi(f) = \varphi(r)$. Dla funkcjonałów φ określonych wzorem (7.7.2) można zastosować pewne twierdzenia analizy matematycznej, które pozwalają wprowadzić funkcjonał φ pod znak całki (tj. w szczególności przestawić kolejność całkowania):

$$\varphi(r) = \frac{1}{m!} \int_a^b f^{(m+1)}(t) \varphi_x[(x-t)_+^m] \, dt = \int_a^b K_m(t) f^{(m+1)}(t) \, dt. \qquad \blacksquare$$

Przykład 7.7.3. Dla funkcjonału

$$\varphi(f) := \int_0^1 x^{-1/2} f(x) \, dx$$

(Sard [1963, s. 31]) znaleźć jego przybliżenie

$$\psi(f) = c_1 f(0) + c_2 f(1)$$

dokładne w Π_1 i jego błąd dla innych f.

Rozwiązanie. Warunek nałożony na ψ oznacza, że różnica $\varphi - \psi$ anihiluje Π_1. Stosując metodę nieoznaczonych współczynników (podrozdz. 7.2), przyjmujemy, że $f(x) := 1$ i $f(x) := x$, co odpowiednio daje równania

$$0 = \varphi(f) - \psi(f) = \int_0^1 x^{-1/2} \, dx - (c_1 + c_2) = 2 - c_1 - c_2,$$
$$0 = \varphi(f) - \psi(f) = \int_0^1 x^{1/2} \, dx - c_2 = \frac{2}{3} - c_2.$$

Stąd $c_1 = \frac{4}{3}$, $c_2 = \frac{2}{3}$. Jądro Peana K_1 dla $\varphi - \psi$ jest równe

$$(\varphi_x - \psi_x)(x-t)_+^1 = \int_0^1 (x-t)_+ x^{-1/2} \, dx - \frac{4}{3}(0-t)_+ - \frac{2}{3}(1-t)_+ =$$
$$= \int_t^1 (x-t) x^{-1/2} \, dx - \frac{2}{3}(1-t) = \frac{4}{3} t(t^{1/2} - 1).$$

Dlatego, na mocy tw. 7.7.2 zastosowanego do funkcjonału $\varphi - \psi$,

$$\int_0^1 x^{-1/2} f(x)\,dx - \left[\frac{4}{3}f(0) + \frac{2}{3}f(1)\right] = \frac{4}{3}\int_0^1 t(t^{1/2}-1)f''(t)\,dt.$$

Otrzymaliśmy dokładne wyrażenie błędu aproksymacji wartości $\varphi(f)$ za pomocą $\psi(f)$, poprawne, gdy $f \in C^2[0,1]$. Ponieważ jądro jest niedodatnie w $[0,1]$, więc tw. 1.2.1 daje równość

$$\frac{4}{3}\int_0^1 t(t^{1/2}-1)f''(t)\,dt = \frac{4}{3}f''(\xi)\int_0^1 t(t^{1/2}-1)\,dt = -\frac{2}{15}f''(\xi). \qquad \blacksquare$$

Różnica $\varphi - \psi$ dwóch funkcjonałów identycznych na Π_m anihiluje tę przestrzeń. Jeśli K_m jest jądrem Peana tej różnicy, to z nierówności Cauchy'ego-Schwarza wynika, że

$$|\varphi(f) - \psi(f)| \leqslant \|K_m\|_2 \, \|f^{(m+1)}\|_2,$$

gdzie $\|g\|_2 := \left(\int_a^b g^2(x)\,dx\right)^{1/2}$. Jeśli anihilacja przestrzeni Π_m nie określa jednoznacznie parametrów funkcjonału ψ, to można je dobrać tak, aby norma $\|K_m\|_2$ była minimalna. Wtedy funkcjonał ψ najlepiej aproksymuje φ *w sensie Sarda*. Schoenberg udowodnił, że te najlepsze przybliżenia można otrzymać w dość prosty sposób opisany w poniższym twierdzeniu. Używamy w nim symbolu $\varphi \circ L$ dla złożenia funkcjonału φ z operatorem L: $(\varphi \circ L)(f) := \varphi(Lf)$.

TWIERDZENIE 7.7.4 (SCHOENBERG). *Niech φ będzie funkcjonałem liniowym postaci (7.7.2). Jeśli węzły t_i są takie, że $a = t_0 < t_1 < \ldots < t_{n-1} < t_n = b$, gdzie $n > N$, to w zbiorze wszystkich funkcjonałów $\sum_{i=0}^{n} c_i \widehat{t}_i$ identycznych z φ na Π_m najlepszym przybliżeniem dla φ w sensie Sarda jest funkcjonał $\varphi \circ L$, gdzie L jest operatorem liniowym, który przekształca daną funkcję na interpolującą ją w węzłach naturalną funkcję sklejaną stopnia $2m+1$.*

Dowód. Niech K_m będzie jądrem Peana funkcjonału $\varphi - \psi$ anihilującego Π_m. Ponieważ wielomiany tej klasy są zarazem naturalnymi funkcjami sklejanymi stopnia $2m+1$, więc $Lp = p$ dla $p \in \Pi_m$. Dlatego funkcjonał $\varphi - \varphi \circ L$ także anihiluje Π_m. Niech $\bar{K}_m$ będzie jego jądrem Peana. Mamy udowodnić, że

$$\int_a^b [\bar{K}_m(t)]^2\,dt \leqslant \int_a^b [K_m(t)]^2\,dt. \tag{7.7.3}$$

Jądrem Peana funkcjonału $\theta := \varphi \circ L - \psi = (\varphi - \psi) - (\varphi - \varphi \circ L)$ jest $\tilde{K}_m := K_m - \bar{K}_m$. Oczywiście

$$\tilde{K}_m(t) = \frac{1}{m!}\,\theta_x[(x-t)_+^m]. \tag{7.7.4}$$

Jeśli $\{s_0, s_1, \ldots, s_n\}$ jest taką bazą rozważanych tu naturalnych funkcji sklejanych, że $s_i(t_j) = \delta_{ij}$, to

$$Lf = \sum_{i=0}^{n} f(t_i)s_i.$$

Stąd wynika, że

$$\theta(f) = \varphi(Lf) - \psi(f) = \sum_{i=0}^{n} f(t_i)\varphi(s_i) - \sum_{i=0}^{n} c_i f(t_i) = \sum_{i=0}^{n} \gamma_i f(t_i)$$

i że wartości jądra $\tilde{K}_m$ wyrażają się wzorem

$$\tilde{K}_m(t) = \frac{1}{m!} \sum_{i=0}^{n} \gamma_i (t_i - t)_+^m. \tag{7.7.5}$$

Niech g będzie taką funkcją, że $g^{(m+1)} = \tilde{K}_m$. Wtedy $g^{(2m+1)} = \tilde{K}_m^{(m)}$; jest to funkcja przedziałami stała, czyli sklejana stopnia 0. Wobec tego g jest funkcją sklejaną stopnia $2m+1$, z węzłami t_i. Sprawdzimy, że jest to funkcja naturalna. W tym celu zauważmy najpierw, że wobec (7.7.5) jest $\tilde{K}_m(t) = 0$ dla $t \geqslant b$, a zatem $g^{(m+1)}(t) = 0$ dla tychże t. Jeśli natomiast $t \leqslant a \leqslant x$, to na mocy (7.7.4)

$$\tilde{K}_m(t) = \frac{1}{m!}\,\theta_x[(x-t)^m].$$

Ponieważ θ anihiluje Π_m, więc $\tilde{K}_m(t) = g^{(m+1)}(t) = 0$ dla $t \leqslant a$. Wiadomo już, że g jest naturalną funkcją sklejaną. Dlatego $Lg = g$ i

$$\int_a^b \bar{K}_m(t)\tilde{K}_m(t)\,dt = \int_a^b \bar{K}_m(t)g^{(m+1)}(t)\,dt = (\varphi - \varphi \circ L)(g) = 0.$$

Stąd już wynika (7.7.3), bo

$$\int_a^b K_m^2\,dt = \int_a^b (\bar{K}_m + \tilde{K}_m)^2\,dt = \int_a^b \bar{K}_m^2\,dt + \int_a^b \tilde{K}_m^2\,dt \geqslant \int_a^b \bar{K}_m^2\,dt.$$

■

Przykład 7.7.5. Niech będzie

$$\varphi(f) := \int_{-1}^{1} f(x)\,dx, \quad \psi(f) := c_0 f(-1) + c_1 f(0) + c_2 f(1).$$

W klasie funkcjonałów ψ identycznych z φ na Π_1 znaleźć najlepsze przybliżenie dla φ w sensie Sarda.

Rozwiązanie. Można sprawdzić, że w tym przypadku operator L dający naturalną funkcję sklejaną stopnia trzeciego jest taki, że

$$(Lf)(x) = a_0 + a_1 x + b_0(x+1)_+^3 + b_1(x)_+^3 + b_2(x-1)_+^3,$$

gdzie

$$a_0 = \frac{1}{4}[-f(-1) + 6f(0) - f(1)],$$
$$a_1 = \frac{1}{4}[-5f(-1) + 6f(0) - f(1)],$$
$$b_0 = b_2 = -\frac{1}{2}b_1 = \frac{1}{4}[f(-1) - 2f(0) + f(1)].$$

Dlatego najlepszy funkcjonał ψ jest taki, że

$$\psi(f) = \varphi(Lf) = \int_{-1}^{1} (Lf)(x)\,dx =$$
$$= 2a_0 + 4b_0 + \frac{1}{4}b_1 = \frac{1}{8}[3f(-1) + 10f(0) + 3f(1)]. \quad \blacksquare$$

Zadania 7.7

1. Niech x i $h > 0$ będą ustalone. Stosując tw. 7.7.2 do funkcjonału

$$\varphi(f) := f(x+h) - f(x) - \tfrac{1}{2}h[3f'(x) - f'(x-h)],$$

udowodnić, że $\varphi(f) = \frac{5}{12}h^3 f'''(\xi)$, gdzie $\xi \in (x-h, x+h)$.

2. Za pomocą tw. 7.7.2 sprawdzić wzór Simpsona z błędem:

$$\int_0^2 f(x)\,dx = \frac{1}{3}[f(0) + 4f(1) + f(2)] - \frac{1}{90}f^{(4)}(\xi).$$

3. Stosując tw. 7.7.4, udowodnić, że

$$\left| f(2) + \frac{1}{4}f(0) - \frac{7}{8}f(1) - \frac{3}{8}f(3) \right| \leqslant \sqrt{\frac{5}{48}}\,\|f''\|_2.$$

4. W przykładzie 7.7.5 znaleźć najmniejsze c takie, że $|\varphi(f) - \psi(f)| \leqslant c\|f''\|_2$.

5. Dla ustalonego $x \in (-1,1)$ znaleźć warunki konieczne i dostateczne na to, żeby wzór przybliżony $f'(x) \approx c_0 f(-1) + c_1 f(0) + c_2 f(1)$ był dokładny dla $f \in \Pi_1$. Znaleźć najlepszy taki wzór w sensie Sarda. Obliczyć minimalną wartość $\int_{-1}^{1} K_1^2(t)\,dt$.

6. Dla danych węzłów t_i i wartości λ_i $(0 \leqslant i \leqslant n)$ znaleźć tę funkcję f spełniającą warunki interpolacyjne $f(t_i) = \lambda_i$ i mającą drugą pochodną przedziałami ciągłą, dla której całka $\int_{t_0}^{t_n} [f''(t)]^2\,dt$ jest najmniejsza.

ROZDZIAŁ 8

Rozwiązywanie numeryczne równań różniczkowych zwyczajnych

8.0. Wstęp

Głównym zadaniem rozważanym w tym rozdziale jest rozwiązywanie numeryczne równań różniczkowych pierwszego rzędu z warunkiem początkowym. Dalsze podrozdziały dotyczą układów równań, równań wyższych rzędów i zadań brzegowych.

8.1. Istnienie i jednoznaczność rozwiązań

Typowe *zagadnienie początkowe* jest opisane równaniami

$$x' = f(t, x), \quad x(t_0) = x_0. \tag{8.1.1}$$

Mają być one spełnione przez funkcję x zmiennej t. Dla przykładowego zagadnienia,

$$x' = x\,\mathrm{tg}(t+3), \quad x(-3) = 1, \tag{8.1.2}$$

łatwo znaleźć dokładne rozwiązanie: jest mianowicie $x(t) = 1/\cos(t+3)$. Zauważmy jednak, że ponieważ $\cos t$ znika, gdy $t = \pm\pi/2$, więc to rozwiązanie jest poprawne tylko dla $-\pi/2 < t+3 < \pi/2$. Zagadnienie jest nietypowe o tyle, że na ogół nie znamy analitycznej postaci rozwiązania zagadnienia początkowego i musimy stosować metody numeryczne.

Istnienie

Nie każde zagadnienie początkowe (8.1.1) ma rozwiązanie. Żeby ono istniało, trzeba coś założyć o funkcji f.

Twierdzenie 8.1.1. *Jeśli dla pewnych α, $\beta > 0$ funkcja f jest ciągła w prostokącie*

$$R := \{(t,x)\colon\ |t-t_0| \leqslant \alpha,\ |x-x_0| \leqslant \beta\}, \tag{8.1.3}$$

to zagadnienie (8.1.1) *ma rozwiązanie $x(t)$ dla $|t-t_0| \leqslant \min\{\alpha, \beta/M\}$, gdzie $M := \max_R |f(t,x)|$.*

Przykład 8.1.2. Sprawdzić, gdzie istnieje rozwiązanie zagadnienia początkowego $x' = (t+\sin x)^2$, $x(0) = 3$.

Rozwiązanie. Funkcja $(t+\sin x)^2$ jest ciągła na całej płaszczyźnie (t,x), czyli α i β w definicji prostokąta R mogą być dowolne. Stała M w tw. 8.1.1 nie przewyższa $(\alpha+1)^2$. Jeśli $\beta := \alpha(\alpha+1)^2$, to $\min\{\alpha, \beta/M\} = \alpha$, więc rozwiązanie istnieje na całej prostej rzeczywistej. ■

Jednoznaczność

Jeśli nawet funkcja f jest ciągła, to zagadnienie (8.1.1) może mieć wiele rozwiązań. Dowodzi tego przykład $x' = x^{2/3}$, $x(0) = 0$. Tu rozwiązaniem jest funkcja $x \equiv 0$, ale także funkcja $x(t) = t^3/27$. Tak więc, aby zapewnić jednoznaczność rozwiązania, trzeba założyć coś więcej o f. Oto typowe twierdzenie tego rodzaju:

Twierdzenie 8.1.3. *Jeśli funkcje f i $\partial f/\partial x$ są ciągłe w prostokącie* (8.1.3), *to dla $|t-t_0| < \min\{\alpha, \beta/M\}$ zagadnienie początkowe* (8.1.1) *ma jednoznaczne rozwiązanie.*

Zauważmy, że w obu powyższych twierdzeniach przedział zmiennej t, w którym rozwiązanie istnieje lub jest jedyne, może być krótszy od tego, który występuje w definicji prostokąta R. Innego typu twierdzenie zapewnia istnienie i jednoznaczność rozwiązania w danym przedziale $[a, b]$ (zob. Henrici [1962, s. 15]).

Twierdzenie 8.1.4. *Jeśli funkcja f jest ciągła dla $a \leqslant t \leqslant b$, $-\infty < x < \infty$ i jeśli istnieje stała L taka, że jest tam*

$$|f(t, x_1) - f(t, x_2)| \leqslant L|x_1 - x_2|, \tag{8.1.4}$$

to zagadnienie początkowe $x' = f(t, x)$, $x(a) = \alpha$ ma w przedziale $[a, b]$ jednoznaczne rozwiązanie.

Warunek (8.1.4) nazywamy *nierównością Lipschitza* (względem drugiej zmiennej). Taki warunek dla funkcji g jednej zmiennej, czyli

$$|g(x_1) - g(x_2)| \leqslant L|x_1 - x_2|, \tag{8.1.5}$$

pociąga za sobą ciągłość tej funkcji, ale może być spełniony nawet wtedy, gdy jej pochodna nie wszędzie istnieje. Oczywiście, jeśli g' istnieje i jeśli $|g'(x)| \leqslant L$, to na mocy twierdzenia o wartości średniej zachodzi (8.1.5).

Przykład 8.1.5. Sprawdzić, że funkcja $g(x) := \sum_{i=1}^{n} a_i|x - w_i|$ spełnia warunek Lipschitza ze stałą $L := \sum_{i=1}^{n}|a_i|$.

Rozwiązanie. Obliczamy

$$\begin{aligned} |g(x_1) - g(x_2)| &= \Big|\sum_{i=1}^{n} a_i|x_1 - w_i| - \sum_{i=1}^{n} a_i|x_2 - w_i|\Big| \leqslant \\ &\leqslant \sum_{i=1}^{n}|a_i|\,\big||x_1 - w_i| - |x_2 - w_i|\big| \leqslant \\ &\leqslant \sum_{i=1}^{n}|a_i|\,|x_1 - x_2| = L|x_1 - x_2|. \end{aligned}$$

■

Zadania 8.1

1. Stosując tw. 8.1.1, znaleźć najszerszy przedział, w którym istnieje rozwiązanie zagadnienia początkowego: (**a**) $x' = 1 + x^2$, $x(0) = 0$, (**b**) (8.1.2).
2. Wykazać, że następujące zagadnienia początkowe mają rozwiązanie:
   (**a**) $x' = \sqrt{|x|}$, $x(0) = 0$ – na całej prostej rzeczywistej,
   (**b**) $x' = \operatorname{tg} x$, $x(0) = 0$ – dla $|t| < \pi/4$,
   (**c**) $x' = 1 + x + x^2 \cos t$, $x(0) = 0$ – dla $|t| \leqslant 1/3$.

3. Wykazać, że zagadnienie początkowe $x' = tx^{2/3}$, $x(0) = 1$ ma rozwiązanie (czy tylko jedno?) dla $|t| \leqslant 2$.
4. Stosując twierdzenia z tego podrozdziału, sprawdzić, gdzie zagadnienie początkowe $x' = x^2$, $x(0) = 1$ ma rozwiązanie. Porównać wynik z rozwiązaniem analitycznym.
5. Dla funkcji f jednej zmiennej, ciągłej dla każdego $x \in \mathbb{R}$, niech będzie $M(r) := \max_{|x| \leqslant r} |f(x)|$. Udowodnić, że jeśli $M(r) = o(r)$ dla $r \to \infty$, to zagadnienie początkowe $x' = f(x)$, $x(0) = 0$ ma rozwiązanie na całej prostej rzeczywistej.
6. Wykazać, że jeśli funkcja $f(t, x)$ jest ciągła i ograniczona dla $a \leqslant t \leqslant b$, $-\infty < x < \infty$, to zagadnienie początkowe $x' = f(t, x)$, $x(a) = \alpha$ ma rozwiązanie w przedziale $[a, b]$.
7. Znaleźć dwa rozwiązania zagadnienia początkowego $x' = x^{1/3}$, $x(0) = 0$.
8. Pokazać, że zagadnienie początkowe $2x' = \sqrt{t^2 + 4x} - t$, $x(2) = -1$ ma dwa rozwiązania: $x = -t^2/4$ i $x = 1 - t$. Jakie założenie tw. 8.1.3 nie jest tu spełnione?
9. Wykazać, że zagadnienie początkowe z zad. 2a ma dwa rozwiązania. Jakie założenie tw. 8.1.3 nie jest tu spełnione?
10. Znaleźć przedział, w którym zagadnienie początkowe $x' = 1/\cos x$, $x(0) = 0$ ma jedoznaczne rozwiązanie.
11. Udowodnić, że zagadnienie początkowe $x' = t^2 + e^x$, $x(0) = 0$ ma jednoznaczne rozwiązanie dla $|t| \leqslant 0.351$.
12. Wykazać, że zagadnienie początkowe $x' = 2te^{-x}$, $x(0) = 0$ ma jednoznaczne rozwiązanie na całej prostej rzeczywistej. Jak dobrać α i β w stosowanych twierdzeniach, żeby dojść do tego wniosku?

8.2. Zastosowanie wzoru Taylora

Rozwiązując numerycznie równanie różniczkowe, otrzymujemy na ogół tylko tablicę przybliżeń x_i wartości dokładnego rozwiązania x w punktach t_i. Dopiero za pomocą tej tablicy można w razie potrzeby zbudować funkcję sklejaną lub inną, przybliżającą to rozwiązanie.

Przykład

Aby zastosować wzór Taylora, musimy wiedzieć, że pewne pochodne cząstkowe funkcji f istnieją. Szczegóły poznamy na przykładzie zagadnienia

$$x' = \cos t - \sin x + t^2, \quad x(-1) = 3. \tag{8.2.1}$$

Zastosujmy wzór Taylora

$$x(t+h) \approx x(t) + hx'(t) + \frac{1}{2!}h^2x''(t) + \frac{1}{3!}h^3x'''(t) + \frac{1}{4!}h^4x^{(4)}(t), \tag{8.2.2}$$

w którym odrzucono pochodne począwszy od piątej. Te, które wyżej występują, otrzymujemy różniczkując względem t obie strony równania różniczkowego (8.2.1):

$$x'' = -\sin t - x' \cos x + 2t,$$
$$x''' = -\cos t - x'' \cos x + (x')^2 \sin x + 2,$$
$$x^{(4)} = \sin t - x''' \cos x + 3x'x'' \sin x + (x')^3 \cos x.$$

Podane wyżej wzory składają się na metodę, która polega na tym, że znając wartość przybliżoną lub dokładną rozwiązania w punkcie t (na początku jest $t = t_0$), obliczamy z (8.2.2) wartość w punkcie $t+h$, korzystając z niej – wartość w $t + 2h$ itd. Ponieważ we wzorze Taylora zachowaliśmy składniki aż do $\mathcal{O}(h^4)$, mówimy, że jest to metoda *rzędu czwartego*. Jej konkretną realizacją jest algorytm, który dla zagadnienia (8.2.1) i $h = 0.01$ wykonuje 200 takich kroków i daje przybliżone wartości rozwiązania w punktach $-0.99, \ldots, 0.99$, 1:

input $M \leftarrow 200$; $h \leftarrow 0.01$; $t \leftarrow -1.0$; $x \leftarrow 3.0$
output 0, t, x
for $k = 1$ **to** M **do**
  $x' \leftarrow \cos t - \sin x + t^2$
  $x'' \leftarrow -\sin t - x' \cos x + 2t$
  $x''' \leftarrow -\cos t - x'' \cos x + (x')^2 \sin x + 2$
  $x^{(4)} \leftarrow \sin t + ((x')^3 - x''') \cos x + 3x'x'' \sin x$
  $x \leftarrow x + h(x' + \frac{1}{2}h(x'' + \frac{1}{3}h(x''' + \frac{1}{4}hx^{(4)})))$
  $t \leftarrow t + h$
  output k, t, x
end do

Zastosowanie powyższego algorytmu daje m.in. następujące wartości:

k	t	x
0	−1.00	3.00000
1	−0.99	3.01400
2	−0.98	3.02803
3	−0.97	3.04209
4	−0.96	3.05617
.....		
196	0.96	6.36566
197	0.97	6.37977
198	0.98	6.39386
199	0.99	6.40791
200	1.00	6.42194

Jakie są zalety i wady opisanej wyżej metody? Wadą jest to, że – jeśli tylko stosowany rząd metody jest większy od 1 – trzeba wielokrotnie różniczkować funkcję f. Jej odpowiednie pochodne cząstkowe muszą istnieć w obszarze, przez który przechodzi szukane rozwiązanie. Jest to warunek znacznie mocniejszy od tego, który zapewnia istnienie i jednoznaczność rozwiązania. Ponadto, ewentualny błąd, jaki popełnimy różniczkując analitycznie f, nie będzie wykryty w obliczeniach numerycznych. Natomiast zaletą metody jest jej prostota i możliwa do osiągnięcia wysoka dokładność. Jeśli pochodne funkcji f względem t aż do 20-tej dadzą się łatwo wyrazić analitycznie (np. za pomocą dostępnych już programów operacji na symbolach), to możemy zastosować metodę rzędu dwudziestego. Wtedy nawet dość duże h, np. $h = 0.2$, zapewnia założoną dokładność, a to zmniejsza liczbę kroków potrzebnych do przejścia do pewnego x (ale wydłuża obliczenia w każdym kroku).

Błędy

W opisanej już metodzie opartej na wzorze Taylora uwzględniamy składniki aż do zawierającego h^n. Odrzucona reszta wynosi

$$E_n := \frac{1}{(n+1)!} h^{n+1} x^{(n+1)}(t + \theta h) \quad (0 < \theta < 1).$$

Jest to *błąd lokalny metody*. Wprawdzie $(n+1)$-szej pochodnej nie znamy, ale najprostsze jej przybliżenie daje wzór

$$E_n \approx \frac{1}{(n+1)!} h^n [x^{(n)}(t+h) - x^{(n)}(t)]. \tag{8.2.3}$$

Podany wcześniej algorytm można niewielkim kosztem wzbogacić o obliczanie tej wielkości (dla $n = 4$). Okazuje się, że te przybliżone wartości błędu nie przekraczają co do modułu $3.42_{10}{-}11$; oczywiście tak jest, jeśli w obliczeniach używamy odpowiednio dużej precyzji.

Ogólniej, błędy występujące w każdej metodzie rozwiązywania numerycznego równań różniczkowych można sklasyfikować tak:

1. Błąd lokalny metody.
2. Błąd lokalny zaokrąglenia.
3. Błąd globalny metody.
4. Błąd globalny zaokrąglenia.
5. Błąd całkowity.

Pierwszy z nich, określony wyżej dla konkretnej metody, jest skutkiem obcięcia procesu (wyrażenia) nieskończonego do postaci skończonej. *Błąd lokalny* przenosi się na wartości obliczane później. Skutki wszystkich błędów lokalnych metody kumulują się i dają *błąd globalny* końcowej wartości (x_{200} w ostatnim przykładzie). Błędy lokalne $\mathcal{O}(h^{n+1})$ dają błąd globalny równy co najmniej $\mathcal{O}(h^n)$, gdyż liczba kroków konieczna dla przejścia od t_0 do dowolnego T wynosi $(T - t_0)/h$.

Błąd zaokrąglenia jest oczywiście spowodowany ograniczoną precyzją obliczeń, a jego wielkość zależy od typu liczb, na których komputer operuje. Również ten błąd lokalny wpływa na obliczane później wartości i determinuje błąd globalny zaokrąglenia końcowej wartości. Błąd całkowity jest sumą błędów globalnych metody i zaokrąglenia.

Metoda Eulera

Najprostszą metodą opartą na wzorze Taylora, a więc rzędu pierwszego, jest *metoda Eulera*. Opisuje ją wzór

$$x(t+h) \approx x(t) + hf(t,x). \tag{8.2.4}$$

Jego oczywistą zaletą jest to, że nie trzeba różniczkować funkcji f. Płacimy za to koniecznością wyboru bardzo małego h. Mamy tu jednak użyteczny przykład metod przybliżonych, ważny także teoretycznie, gdyż na metodzie Eulera bazuje jeden z dowodów istnienia rozwiązania; zob. Henrici [1962, s. 15–25].

Równanie z opóźnionym argumentem

W pewnych zastosowaniach występują równania różniczkowe *z opóźnionym argumentem*. Dotyczy to np. modeli zmienności populacji. Wtedy wartość $x'(t)$ zależy od wartości funkcji x dla wcześniejszych wartości zmiennej t, jak w równaniu

$$x'(t) = f(x(t-1)).$$

Żeby je rozwiązać począwszy od punktu $t = 0$, musimy znać „historię" funkcji $x(t)$ w przedziale $[-1, 0]$. To są warunki początkowe.

Oto konkretne, dobrze określone zagadnienie tego typu:

$$x'(t) = x(t-1) \quad (t \geqslant 0), \quad x(t) = t^2 \quad (-1 \leqslant t \leqslant 0).$$

Stąd dla $t \in [0,1]$ wynika, że

$$x'(t) = x(t-1) = (t-1)^2, \quad x(0) = 0.$$

To zagadnienie rozwiązujemy bez trudu dokładnie:

$$x(t) = \frac{1}{3}(t-1)^3 + \frac{1}{3} \quad (0 \leqslant t \leqslant 1).$$

W ten sam sposób otrzymujemy rozwiązanie w przedziale $[1,2]$. Ponieważ jest tam

$$x'(t) = x(t-1) = \frac{1}{3}(t-2)^3 + \frac{1}{3}, \quad x(1) = \frac{1}{3},$$

więc

$$x(t) = \frac{1}{12}(t-2)^4 + \frac{1}{3}t - \frac{1}{12} \quad (1 \leqslant t \leqslant 2).$$

Podobnie możemy postępować dalej.

Jeśli równanie różniczkowe jest bardziej złożone, np. takie:

$$x'(t) = \sin[x(t-1)^3] + \log[x(t) + t^5],$$

to musimy odwołać się do metod numerycznych. Jedna z nich opiera się na wzorze Taylora, którego zalety i wady już znamy. Rozważmy przykład

$$x'(t) = 2x(t-1) + x(t) \quad (t \geqslant 0), \qquad x(t) = t^2 \quad (-1 \leqslant t \leqslant 0). \tag{8.2.5}$$

Szukajmy rozwiązania w przedziale $[0,1]$, stosując wzór przybliżony

$$x(t+h) \approx x(t) + hx'(t) + \frac{1}{2}h^2x''(t) + \frac{1}{6}h^3x'''(t).$$

Niezbędne pochodne obliczamy ze wzorów:

$$\begin{aligned} x'(t) &= 2x(t-1) + x(t) = 2(t-1)^2 + x(t), \\ x''(t) &= 2x'(t-1) + x'(t) = 4(t-1) + x'(t), \\ x'''(t) &= 2x''(t-1) + x''(t) = 4 + x''(t). \end{aligned}$$

Algorytm daje przybliżenia wartości $x(t)$ na zbiorze dyskretnym punktów z przedziału $[0,1]$. Trzeba też zapamiętywać przybliżone wartości $x'(t)$, $x''(t)$ i $x'''(t)$ w tychże punktach. Będą one potrzebne, gdy będziemy rozwiązywać zagadnienie (8.2.5) w przedziale $[1,2]$ z tym samym h.

Teorię równań różniczkowych z opóźnionym argumentem opisują Driver [1977], Kuang [1993] oraz Diekmann, Van Gils, Verduyn Lunel i Walther [1995]. System DELSOL programów służących do rozwiązywania układów takich równań opracowali Willé i Baker [1992].

ZADANIA 8.2

1. Wykazać, że n-ta pochodna funkcji $\exp(-t^2)$ (por. zad. K6) jest równa $\exp(-t^2)P_n(t)$, gdzie $P_0 = 1$, $P_{n+1}(t) = P_n'(t) - 2tP_n(t)$ $(n \geqslant 0)$. Sprawdzić, że w szczególności $P_4(t) = 12 - 48t^2 + 16t^4$, $P_5(t) = -120t + 160t^3 - 32t^5$.

2. Wykazać, że pochodne funkcji $f(t) := \sin t^2$ (por. zad. K7) wyrażają się wzorem

$$f^{(n)}(t) = P_n(t)\sin t^2 + Q_n(t)\cos t^2,$$

gdzie $P_0 \equiv 1$, $Q_0 \equiv 0$,

$$P_{n+1}(t) = P_n'(t) - 2tQ_n(t), \quad Q_{n+1}(t) = Q_n'(t) + 2tP_n(t) \quad (n \geqslant 0).$$

Znaleźć wielomiany P_n, Q_n dla $n \leqslant 6$.

3. Stosując jeden krok metody rzędu n wynikającej ze wzoru Taylora, znaleźć wartość przybliżoną $x(t)$ rozwiązania zagadnienia początkowego:

(a) $x' = -tx^2$, $x(0) = 2$, $n = 2$, $t = 0.1$.

(b) $x' = x^2 + xe^t$, $x(0) = 1$, $n = 3$, $t = 0.01$.

(c) $5tx' + x^2 = 2$, $x(4) = 1$, $n = 2$, $t = 4.1$.

4. Znaleźć wyrażenia dla x'', x''' i $x^{(4)}$ wiedząc, że $x' = \cos(tx)$.

5. *Równanie całkowe* z definicji zawiera nieznaną funkcję pod znakiem całki. Przykładem *równania Volterry* jest równanie

$$x(t) = \int_0^t \cos\bigl(s + x(s)\bigr)\,ds + e^t.$$

Różniczkując jego obie strony, otrzymać równoważne zagadnienie początkowe.

ZADANIA KOMPUTEROWE 8.2

K1. Uzupełnić algorytm podany w tekście o obliczanie wielkości E_4 za pomocą wzoru (8.2.3).

K2. Napisać i sprawdzić program rozwiązujący podane niżej zagadnienie początkowe w przedziale $[a, b]$, stosując metodę rzędu n opartą na wzorze Taylora dla danego h.

(a) $x' = x + e^t + tx$, $x(1) = 2$, $[a, b] = [1, 3]$, $n = 5$, $h = 0.01$.

(b) $x' = 1 + x^2 - t^3$, $x(0) = -1$, $[a, b] = [0, 2]$, $n = 4$, $h \approx 0.01$ ma być liczbą o skończonym rozwinięciu dwójkowym.

(c) $x' = 1 + x^2$, $x(0) = 0$, $[a, b] = [0, 1.56]$, $n = 4$, $h = 0.01$ (następnie uznać obliczone $x(1.56)$ za wartość początkową i powtórzyć obliczenia wstecz aż do punktu $t = 0$; porównać i skomentować wyniki).

(d) $x' = -(3t^2 + x^2)/(2t^3 + 3tx)$, $x(1) = -2$, $[a, b] = [0, 1]$, $n = 2$ oraz $h = -0.01$ (sprawdzić, że $t^3x^2 + tx^3 + 4 = 0$ i wykorzystać to do kontroli rozwiązania przybliżonego).

(e) $x' = (t + x)^2$, $x(0) = 7$, $[a, b] = [-2, 2]$, $n = 3$, $h = \pm 0.01$.

K3. Równanie $\operatorname{arctg}(x/t) = \log\sqrt{x^2+t^2}$ określa x jako funkcję uwikłaną zmiennej t. Wykazać, że $x' = (t+x)/(t-x)$, $x(1) = 0$. Dla tego zagadnienia początkowego oraz $[a,b] = [0,2]$, $n = 4$, $h = \pm 0.01$ wykonać obliczenia jak w zad. K2.

K4. Udowodnić, że jeśli $x' = 1 - xt^{-1}$, to $x'' = (1-2x')t^{-1}$ i $x^{(n)} = -nx^{(n-1)}t^{-1}$ dla $n \geqslant 3$. Znaleźć rozwiązanie dokładne tego równania różniczkowego z warunkiem początkowym $x(2) = 2$. Wykonać obliczenia jak w zad. K2, przyjmując, że $[a,b] = [2,20]$, $n = 10$, $h = 1$. Porównać wyniki z rozwiązaniem dokładnym (Conte i de Boor [1980]).

K5. Rozwiązać numerycznie zagadnienie początkowe $(x')^2 - 2tx' - x\cos t = 0$, $x(0) = 0$ w przedziale $[0,1]$ (rozwiązanie równe tożsamościowo 0 nie jest tym, które należy znaleźć). Jakie skutki ma zmiana $(x')^2$ na $(x')^3$?

K6. Metody numeryczne stosowane w zagadnieniach początkowych mogą również służyć do obliczania całek oznaczonych lub nieoznaczonych. Przykład: całkę $\int_0^2 \exp(-s^2)\,ds$ można obliczyć, rozwiązując zagadnienie $x' = \exp(-t^2)$, $x(0) = 0$ w przedziale $[0,2]$. Zrobić to, stosując wzór Taylora rzędu: (**a**) $n = 4$, (**b**) $n = 6$ dla $h = 0.01$. Wykorzystać zad. 1. Do kontroli użyć wartości funkcji błędu $\operatorname{erf}(t) := (2/\sqrt{\pi})\int_0^t \exp(-s^2)\,ds$. Powinno być $x(2) \approx 0.88208\,13907$ (Abramowitz i Stegun [1964, s. 311]).

K7. Stablicować wartości *całki Fresnela* $\varphi(t) := \int_0^t \sin s^2\,ds$ w przedziale $[0,10]$, korzystając z zad. 2 i postępując jak w zad. K6 dla: (**a**) $n = 5$, $h = 0.1$; (**b**) $n = 7$, $h = 0.1$; (**c**) $n = 7$, $h = 0.2$.
Porównać wyniki tych trzech wariantów.

K8. Postępując jak w zad. K6 stablicować:

(**a**) w przedziale $[0, \pi/2]$ wartości *całki eliptycznej drugiego rodzaju*

$$x(t) := \int_0^t (1 - k^2\sin^2\theta)^{1/2}\,d\theta$$

dla $k = 1/2$, $n = 3$ i $h = 0.01$,

(**b**) w przedziale $[0,5]$, dla $n = 3$ i $h = 1/64$, wartości funkcji

$$f(x) := \int_0^x \sqrt{1+t^3}\,dt$$

(ta całka, jak i poprzednia, nie wyraża się w skończonej postaci przez funkcje elementarne),

(**c**) w przedziale $[-2,0]$, dla wybranych n i h, *dilogarytm*

$$f(x) := -\int_0^x t^{-1}\log(1-t)\,dt,$$

(**d**) w przedziale $[2,5]$ wartości całki $x(t) := \int_2^t \sin u^3\,du$, przyjmując, że $n = 3$ i $h = 0.05$,

(**e**) w przedziale $[0,3]$ rozwiązanie równania całkowego z zad. 5 dla $n = 3$ i $h = 0.01$.

K9. Metodą podaną w tekście rozwiązać numerycznie w przedziale $[0,1]$ równanie (8.2.5), przyjmując, że $n = 3$, $h = 0.01$.

8.3. Metody Rungego-Kutty

Metoda n-tego rzędu opisana w podrozdz. 8.2 dla zagadnienia początkowego

$$x' = f(t,x), \quad x(t_0) = x_0$$

wymaga, jak już wiemy, znalezienia wyrażeń dla pochodnych funkcji f względem t do n-tej włącznie; potem ich obliczanie trzeba zaprogramować. Metody Rungego-Kutty nie mają tej wady. Zamiast tych pochodnych trzeba tu obliczać dobrane w szczególny sposób kombinacje wartości funkcji f. Ideę całej klasy metod poznamy na prostym przykładzie metody rzędu drugiego.

Metoda Rungego-Kutty rzędu drugiego

Z równania różniczkowego wynika, że

$$\begin{aligned} x'(t) &= f, \\ x''(t) &= f_t + f_x x' = f_t + f f_x, \\ x'''(t) &= f_{tt} + f_{tx} f + (f_t + f f_x) f_x + f(f_{xt} + f f_{xx}), \ldots \end{aligned}$$

(wskaźniki oznaczają tu pochodne cząstkowe względem odpowiednich zmiennych). Te wyrażenia podstawiamy do wzoru Taylora:

$$\begin{aligned} x(t+h) &= x + fh + \tfrac{1}{2}h^2(f_t + f f_x) + \mathcal{O}(h^3) = \\ &= x + \tfrac{1}{2}hf + \tfrac{1}{2}h(f + hf_t + hf f_x) + \mathcal{O}(h^3). \end{aligned} \tag{8.3.1}$$

Argumentem funkcji x jest tu t, f oznacza $f(t,x)$ itd. Pochodne cząstkowe można stąd wyeliminować stosując wzór Taylora dla funkcji dwóch zmiennych. Istotnie,

$$f(t+h, x+hf) = f + hf_t + hf f_x + \mathcal{O}(h^2),$$

czyli

$$x(t+h) = x + \tfrac{1}{2}hf + \tfrac{1}{2}hf(t+h, x+hf) + \mathcal{O}(h^3).$$

Wobec tego przybliżone wyrażenie dla $x(t+h)$ jest następujące:

$$x(t+h) \approx x(t) + \tfrac{1}{2}(F_1 + F_2), \tag{8.3.2}$$

gdzie

$$F_1 := hf(t,x), \quad F_2 := hf(t+h, x+F_1).$$

Wzór (8.3.2) określa *metodę Rungego-Kutty rzędu drugiego*, znaną też jako *metoda Heuna*.

Ogólniej, każda metoda Rungego-Kutty rzędu drugiego wynika ze wzoru

$$x(t+h) = x + w_1 hf + w_2 hf(t+\alpha h, x+\beta hf) + \mathcal{O}(h^3)$$

z parametrami w_1, w_2, α, β. Ta równość zachodzi, jeśli

$$x(t+h) = x + w_1 hf + w_2 h(f + \alpha h f_t + \beta h f f_x) + \mathcal{O}(h^3).$$

Porównując tę zależność z (8.3.1), wnioskujemy, że powinno być

$$w_1 + w_2 = 1, \quad w_2\alpha = \tfrac{1}{2}, \quad w_2\beta = \tfrac{1}{2}. \tag{8.3.3}$$

Jedno z rozwiązań tego układu ($w_1 = w_2 = \frac{1}{2}$, $\alpha = \beta = 1$) daje metodę Heuna, ale można też przyjąć np., że $w_1 = 0$, $w_2 = 1$ i $\alpha = \beta = \frac{1}{2}$. Wtedy otrzymujemy *zmodyfikowaną metodę Eulera* opisaną wzorem

$$x(t+h) \approx x(t) + F_2,$$

gdzie

$$F_1 := hf(t,x), \quad F_2 := hf(t+\tfrac{1}{2}h, x+\tfrac{1}{2}F_1).$$

Warto ją porównać ze standardową metodą Eulera (8.2.4).

Metoda Rungego-Kutty rzędu czwartego

Konstrukcja metod Rungego-Kutty wyższych rzędów jest znacznie bardziej kłopotliwa i nie będziemy nią się zajmować. Końcowe wzory są jednak raczej eleganckie i łatwe do zaprogramowania. Klasyczna *metoda Rungego-Kutty rzędu czwartego* polega na zastosowaniu wzoru

$$x(t+h) \approx x(t) + \tfrac{1}{6}(F_1 + 2F_2 + 2F_3 + F_4), \tag{8.3.4}$$

gdzie

$$F_1 := hf(t,x), \qquad F_2 := hf(t+\tfrac{1}{2}h, x+\tfrac{1}{2}F_1),$$
$$F_3 := hf(t+\tfrac{1}{2}h, x+\tfrac{1}{2}F_2), \qquad F_4 := hf(t+h, x+F_3).$$

Rząd tej metody jest równy 4, ponieważ błąd wzoru przybliżonego (8.3.4) wynosi $\mathcal{O}(h^5)$; wyrażenie składnika błędu z h^5 jest znane.

Przykład 8.3.1. Zastosować metodę Rungego-Kutty rzędu czwartego do zagadnienia początkowego

$$x' = t^{-2}(tx - x^2), \quad x(1) = 2$$

w przedziale $[1,3]$ dla $h = 1/128$.

Rozwiązanie. To zagadnienie ma rozwiązanie $x(t) := t\left(\frac{1}{2} + \log t\right)^{-1}$, można więc sprawdzić, jakimi błędami jest obarczone rozwiązanie przybliżone. Poniższy algorytm wyznacza dla każdego $k = 0, 1, \ldots, 256$ wartość $t = 1 + k/128$, przybliżoną wartość rozwiązania x w tym punkcie i jego błąd bezwzględny:

$M \leftarrow 256$; $t \leftarrow 1.0$; $x \leftarrow 2.0$; $h \leftarrow 1/128$
define $f(t, x) = (tx - x^2)/t^2$
define $u(t) = t/(1/2 + \log t)$
output 0, t, x, $|u(t) - x|$
for $k = 1$ **to** M **do**
  $F_1 \leftarrow hf(t, x)$
  $F_2 \leftarrow hf\left(t + \frac{1}{2}h, x + \frac{1}{2}F_1\right)$
  $F_3 \leftarrow hf\left(t + \frac{1}{2}h, x + \frac{1}{2}F_2\right)$
  $F_4 \leftarrow hf(t + h, x + F_3)$
  $x \leftarrow x + (F_1 + 2F_2 + 2F_3 + F_4)/6$
  $t \leftarrow t + h$
  output k, t, x, $|u(t) - x|$
end do

Oto niektóre wyniki, które ten algorytm daje:

k	t	x	$\lvert u(t) - x\rvert$
0	1.00000	2.00000	0
1	1.00781	1.98473	$1.2_{10}-7$
2	1.01563	1.97016	0
3	1.02344	1.95623	0
4	1.03125	1.94293	$1.2_{10}-7$
5	1.03906	1.93020	$1.2_{10}-7$
6	1.04688	1.91802	0
7	1.05469	1.90637	$1.2_{10}-7$
8	1.06250	1.89521	$1.2_{10}-7$
9	1.07031	1.88452	0
.....			
249	2.94531	1.86387	$7.2_{10}-7$
250	2.95313	1.86569	$6.0_{10}-7$
251	2.96094	1.86750	$7.2_{10}-7$
252	2.96875	1.86932	$6.0_{10}-7$
253	2.97656	1.87115	$4.8_{10}-7$
254	2.98438	1.87297	$6.0_{10}-7$
255	2.99219	1.87480	$6.0_{10}-7$
256	3.00000	1.87663	$6.0_{10}-7$

■

Błąd lokalny metody

Pierwszy krok metody Rungego-Kutty (8.3.4) daje przybliżone rozwiązanie $\tilde{x}(t_0+h)$. Błąd lokalny metody wynosi więc $x(t_0+h)-\tilde{x}(t_0+h)$. Z teorii metody wynika, że dla małych h ten błąd zachowuje się tak jak Ch^5, gdzie nieznane C zależy od t_0 i rozwiązania x (a nie zależy od h). Aby oszacować ten iloczyn, załóżmy jednak, że C lokalnie się nie zmienia. Niech $\hat{x}(t_0+h)$ będzie rozwiązaniem przybliżonym otrzymanym z $x(t_0)$ za pomocą dwóch kroków metody, z $h/2$ zamiast h. Stąd

$$x(t_0+h)=\tilde{x}(t_0+h)+Ch^5, \quad x(t_0+h)=\hat{x}(t_0+h)+2C(h/2)^5.$$

Jeśli tak, to lokalny błąd metody jest równy

$$Ch^5=\tfrac{15}{16}[\hat{x}(t_0+h)-\tilde{x}(t_0+h)]\approx\hat{x}(t_0+h)-\tilde{x}(t_0+h).$$

Programując metodę Rungego-Kutty, możemy łatwo sterować wielkością błędu, obliczając $|\hat{x}(t_0+h)-\tilde{x}(t_0+h)|$. Jeśli jest on zbyt duży, zmniejszamy (zwykle połowimy) h, a jeśli jest znacznie mniejszy od dopuszczalnego, to h możemy podwoić.

Metoda adaptacyjna Rungego-Kutty-Fehlberga

Udowodniono, że jeśli liczba wartości funkcji f obliczanych w jednym kroku metody Rungego-Kutty jest ustalona, to jej rząd nie przekracza pewnej wielkości:

Liczba wartości funkcji:	1	2	3	4	5	6	7	8
Maksymalny rząd metody R.-K.:	1	2	3	4	4	5	6	6

Jak widać, metody wysokiego rzędu mogą być mniej atrakcyjne, gdyż każdy krok wymaga obliczania wielu wartości funkcji f. Fehlberg [1969] zaproponował więc stosowanie metody Rungego-Kutty rzędu czwartego, wymagającej obliczania pięciu wartości funkcji, i metody rzędu piątego, w której trzeba obliczać sześć takich wartości. Na pierwszy rzut oka wydaje się to bardzo nieekonomiczne. Fehlberg wykorzystał jednak możliwość wyboru pewnych parametrów metody Rungego-Kutty (wspomnianą wcześniej w związku z metodami rzędu drugiego) i dobrał je tak, aby w obu metodach występowały te same wartości. Dlatego wystarczy sześć takich wartości. Daje to ostatecznie *metodę Rungego-Kutty-Fehlberga* rzędu piątego, która daje dwa przybliżenia wartości $x(t+h)$:

$$\hat{x}(t+h):=x(t)+\sum_{i=1}^{6}a_iF_i, \quad \tilde{x}(t+h):=x(t)+\sum_{i=1}^{6}b_iF_i. \tag{8.3.5}$$

Wielkości F_i obliczamy ze wzoru

$$F_i := hf\Big(t + c_i h, x + \sum_{j=1}^{i-1} d_{ij} F_j\Big) \quad (1 \leqslant i \leqslant 6). \tag{8.3.6}$$

Wszystkie potrzebne stałe podano niżej w tablicy:

i	a_i	$a_i - b_i$	c_i	d_{i1}	d_{i2}	d_{i3}	d_{i4}	d_{i5}
1	$\frac{16}{135}$	$\frac{1}{360}$	0					
2	0	0	$\frac{1}{4}$	$\frac{1}{4}$				
3	$\frac{6656}{12825}$	$-\frac{128}{4275}$	$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$			
4	$\frac{28561}{56430}$	$-\frac{2197}{75240}$	$\frac{12}{13}$	$\frac{1932}{2197}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$		
5	$-\frac{9}{50}$	$\frac{1}{50}$	1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$	
6	$\frac{2}{55}$	$\frac{2}{55}$	$\frac{1}{2}$	$-\frac{8}{27}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$

Pierwszy wzór (8.3.5) jest rzędu piątego, drugi – rzędu czwartego. Różnica

$$e := \hat{x}(t+h) - \tilde{x}(t+h) = \sum_{i=1}^{6} (a_i - b_i) F_i \tag{8.3.7}$$

jest z grubsza błędem drugiego, mniej dokładnego wzoru i możemy jej używać do sterowania wielkością h. Natomiast pierwszy wzór, bardziej dokładny, daje (dla właściwego h) ostateczne przybliżenie wartości $x(t+h)$. Można więc sądzić, że jego rzeczywisty błąd jest znacznie mniejszy od e.

Metodę Rungego-Kutty-Fehlberga zastosowano w podanym niżej *algorytmie adaptacyjnym*. Rozwiązuje on zagadnienie początkowe $x' = f(t,x)$, $x(a) = \alpha$ w danym przedziale $[a,b]$. Ustalamy także początkowe $h > 0$, wielkość $\delta > 0$, której $|e|$ nie powinno przekroczyć, i maksymalną liczbę M kroków. Ponieważ e jest w przybliżeniu równe Ch^5, więc wydaje się rozsądne podwajać h, gdy $|e| < \delta/128$, bo wtedy nowe e zapewne nie przekroczy $\delta/4$.

input a, α, b, h, δ, M
$t \leftarrow a$; $x \leftarrow \alpha$
$k \leftarrow 0$; `iflag` $\leftarrow 1$
while $k < M$ **do**
    $d \leftarrow b - t$
    if $|d| \leqslant h$ **then**
        `iflag` $\leftarrow 0$
        $h \leftarrow d$
    end if

```
        y ← x
        obliczyć F_1, F_2, ..., F_6 z (8.3.6)
        obliczyć x z I wzoru (8.3.5)
        obliczyć e z (8.3.7)
        if |e| ⩾ δ then
            h ← h/2; x ← y
        else
            if |e| < δ/128 then
                h ← 2h; x ← y
            else
                t ← t + h; k ← k + 1
                output k, t, x, e
                if iflag = 0 then stop
            end if
        end if
    end while
```

Inny sposób sterowania wielkością h polega na podstawieniu

$$h \leftarrow 0.9h\,[\delta/|e|]^{1/(1+p)},$$

gdzie p jest rzędem pierwszego ze wzorów takiej pary jak w (8.3.5); zob. np. Hull, Enright, Fellen i Sedgwick [1972] oraz Shampine, Watts i Davenport [1976].

W ostatnich latach znaleziono wiele innych par wzorów Rungego-Kutty, które korzystają z tych samych wartości funkcji f i które mają różny rząd (zwykle p i $p+1$). Są one na ogół efektywne, jeśli równanie nie jest sztywne (por. podrozdz. 8.12). Kilka przykładów znajdziemy niżej w zadaniach komputerowych. Klasyczna metoda rzędu czwartego jest jednak dalej bardzo popularna, jeśli nie korzysta się ze schematów adaptacyjnych. Dodatkowe informacje o metodach Rungego-Kutty można znaleźć w wielu pracach; p. np. Butcher [1987], Fehlberg [1969], Gear [1971], Jackson, Enright i Hull [1978], Prince i Dormand [1981], Shampine i Gordon [1975], Thomas [1986] oraz Verner [1978].

Zadania 8.3

1. Jaką postać ma metoda Rungego-Kutty rzędu drugiego dla $\alpha = 2/3$?
2. Stosując ekstrapolację Richardsona do metody Eulera, uzasadnić zmodyfikowaną metodę Eulera.
3. Udowodnić, że wzór

   $$x(t+h) \approx x(t) + \tfrac{1}{9}(2F_1 + 3F_2 + 4F_3),$$

   gdzie

   $$F_1 := hf(t,x), \quad F_2 := hf\big(t + \tfrac{1}{2}h, x + \tfrac{1}{2}F_1\big), \quad F_3 := hf\big(t + \tfrac{3}{4}h, x + \tfrac{3}{4}F_2\big),$$

określa metodę Rungego-Kutty rzędu trzeciego. Wykazać, że dla równania $x' = x + t$ jest ona równoważna zastosowaniu wzoru Taylora tegoż rzędu.

4. Niech $f(t,x)$ nie zależy od x. Udowodnić, że rząd metody Rungego-Kutty (8.3.4) jest równy 4 i że jest ona równoważna zastosowaniu wzoru Simpsona z podrozdz. 7.2.

5. Udowodnić, że dla równania $x' = \lambda x$ wzór (8.3.4) można uprościć do postaci

$$x(t+h) \approx \left[1 + h\lambda + \tfrac{1}{2}(h\lambda)^2 + \tfrac{1}{6}(h\lambda)^3 + \tfrac{1}{24}(h\lambda)^4\right]x(t).$$

ZADANIA KOMPUTEROWE 8.3

K1. Zaprogramować metodę Rungego-Kutty rzędu czwartego (tak, aby działała poprawnie także dla $h < 0$) i sprawdzić ją w następujących zagadnieniach początkowych:

(**a**) $(e^t + 1)x' + (e^t - 1)x = 0$, $x(0) = 3$, przedział $[-2, 0]$, $h = -0.01$. Znaleźć dokładne rozwiązanie i porównać z nim wyniki obliczeń.

(**b**) $x' = \lambda x + \cos t - \lambda \sin t$ (gdzie $\lambda = 5, -5, -10$), $x(0) = 0$, przedział $[0, 5]$, $h = 0.01$. Sprawdzić, jak λ wpływa na dokładność przybliżeń.

(**c**) $x' = e^{xt} + \cos(x - t)$, $x(1) = 3$, $h = 0.01$. Obliczenia powinny być przerwane *przed* wystąpieniem nadmiaru.

K2. Zaprogramować metodę adaptacyjną Rungego-Kutty-Fehlberga i sprawdzić ją dla zagadnienia początkowego $x' = x^2$, $x(0) = 1$ w przedziale $[0, 2]$. Rozwiązanie dokładne jest równe $x(t) = 1/(1-t)$. Jak działa program w pobliżu punktu nieciągłości rozwiązania?

K3. Porównać na przykładach klasyczną metodę Rungego-Kutty (8.3.4) z następującą *metodą Rungego-Kutty-Gilla* rzędu czwartego:

$$x(t+h) \approx x(t) + \tfrac{1}{6}\left[F_1 + (2 - \sqrt{2})F_2 + (2 + \sqrt{2})F_3 + F_4\right],$$

gdzie:

$$F_1 := hf(t, x),$$

$$F_2 := hf\left(t + \tfrac{1}{2}h, x + \tfrac{1}{2}F_1\right),$$

$$F_3 := hf\left(t + \tfrac{1}{2}h, x + \tfrac{1}{2}(\sqrt{2} - 1)F_1 + \tfrac{1}{2}(2 - \sqrt{2})F_2\right),$$

$$F_4 := hf\left(t + h, x - \tfrac{1}{2}\sqrt{2}F_2 + \tfrac{1}{2}(2 + \sqrt{2})F_3\right).$$

K4. Porównać na przykładzie zagadnienia o znanym dokładnym rozwiązaniu klasyczną metodę Rungego-Kutty (8.3.4) z następującą *metodą Rungego-Kutty* rzędu piątego:

$$x(t+h) \approx x(t) + \tfrac{1}{24}F_1 + \tfrac{5}{48}F_4 + \tfrac{27}{56}F_5 + \tfrac{125}{336}F_6,$$

gdzie F_1 i F_2 są określone jak w pierwszej metodzie i

$$F_3 := hf\left(t + \tfrac{1}{2}h, x + \tfrac{1}{4}F_1 + \tfrac{1}{4}F_2\right),$$

$$F_4 := hf(t + h, x - F_2 + 2F_3),$$

$$F_5 := hf\left(t + \tfrac{2}{3}h, x + \tfrac{7}{27}F_1 + \tfrac{10}{27}F_2 + \tfrac{1}{27}F_4\right),$$

$$F_6 := hf\left(t + \tfrac{1}{5}h, x + \tfrac{28}{625}F_1 - \tfrac{1}{5}F_2 + \tfrac{546}{625}F_3 + \tfrac{54}{625}F_4 - \tfrac{378}{625}F_5\right).$$

K5. Zaprogramować i sprawdzić metodę adaptacyjną określoną (jak metoda Rungego-Kutty-Fehlberga w tekście) za pomocą poniższej tablicy stałych:

(**a**) (*metoda Rungego-Kutty-Mersona* rzędu piątego)

i	a_i	$a_i - b_i$	c_i	d_{i1}	d_{i2}	d_{i3}	d_{i4}
1	$\frac{1}{6}$	$\frac{1}{15}$	0				
2	0	0	$\frac{1}{3}$	$\frac{1}{3}$			
3	0	$-\frac{3}{10}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$		
4	$\frac{2}{3}$	$\frac{4}{15}$	$\frac{1}{2}$	$\frac{1}{8}$	0	$\frac{3}{8}$	
5	$\frac{1}{6}$	$-\frac{1}{30}$	1	$\frac{1}{2}$	0	$-\frac{3}{2}$	2

(**b**) (*metoda Rungego-Kutty-Vernera* rzędu piątego; zob. Verner [1978])

i	a_i	$a_i - b_i$	c_i	d_{i1}	d_{i2}	d_{i3}	d_{i4}	d_{i5}	d_{i6}	d_{i7}
1	$\frac{3}{80}$	$\frac{33}{640}$	0							
2	0	0	$\frac{1}{18}$	$\frac{1}{18}$						
3	$\frac{4}{25}$	$-\frac{132}{325}$	$\frac{1}{6}$	$-\frac{1}{12}$	$\frac{1}{4}$					
4	$\frac{243}{1120}$	$\frac{891}{2240}$	$\frac{2}{9}$	$-\frac{2}{81}$	$\frac{4}{27}$	$\frac{8}{81}$				
5	$\frac{77}{160}$	$-\frac{33}{320}$	$\frac{2}{3}$	$\frac{40}{33}$	$-\frac{4}{11}$	$-\frac{56}{11}$	$\frac{54}{11}$			
6	$\frac{73}{700}$	$-\frac{73}{700}$	1	$-\frac{369}{73}$	$\frac{72}{73}$	$\frac{5380}{219}$	$-\frac{12285}{584}$	$\frac{2695}{1752}$		
7	0	$\frac{891}{8320}$	$\frac{8}{9}$	$-\frac{8716}{891}$	$\frac{656}{297}$	$\frac{39520}{891}$	$-\frac{416}{11}$	$\frac{52}{27}$	0	
8	0	$\frac{2}{35}$	1	$\frac{3015}{256}$	$-\frac{9}{4}$	$-\frac{4219}{78}$	$\frac{5985}{128}$	$-\frac{539}{384}$	0	$\frac{693}{3328}$

(**c**) (para metod Rungego-Kutty, rzędu szóstego i piątego; zob. Prince i Dormand [1981])

i	a_i	b_i	c_i
1	$\frac{821}{10800}$	$\frac{61}{864}$	0
2	0	0	$\frac{1}{10}$
3	$\frac{19683}{71825}$	$\frac{98415}{321776}$	$\frac{2}{9}$
4	$\frac{175273}{912600}$	$\frac{16807}{146016}$	$\frac{3}{7}$
5	$\frac{395}{3672}$	$\frac{1375}{7344}$	$\frac{3}{5}$
6	$\frac{785}{2704}$	$\frac{1375}{5408}$	$\frac{4}{5}$
7	$\frac{3}{50}$	$-\frac{37}{1120}$	1
8	0	$\frac{1}{10}$	1

i	d_{i1}	d_{i2}	d_{i3}	d_{i4}	d_{i5}	d_{i6}	d_{i7}
2	$\frac{1}{10}$						
3	$-\frac{2}{81}$	$\frac{20}{81}$					
4	$\frac{615}{1372}$	$-\frac{270}{343}$	$\frac{1053}{1372}$				
5	$\frac{3243}{5500}$	$-\frac{54}{55}$	$\frac{50949}{71500}$	$\frac{4998}{17875}$			
6	$-\frac{26492}{37125}$	$\frac{72}{55}$	$\frac{2808}{23375}$	$-\frac{24206}{37125}$	$\frac{338}{459}$		
7	$\frac{5561}{2376}$	$-\frac{35}{11}$	$-\frac{24117}{31603}$	$\frac{899983}{200772}$	$-\frac{5225}{1836}$	$\frac{3925}{4056}$	
8	$\frac{465467}{266112}$	$-\frac{2945}{1232}$	$-\frac{5610201}{14158144}$	$\frac{10513573}{3212352}$	$-\frac{424325}{205632}$	$\frac{376225}{454272}$	0

K6. Rozważyć zagadnienie początkowe $x' = -kx$, $x(0) = 1$ (mające rozwiązanie dokładne $x(t) = e^{-kt}$) w przedziale $[0, 1]$ dla różnych wartości parametru k. Dla $k = 5$ i kilku wartości h, np. od 0.2 co 0.2 do 1, porównać zachowanie metod Eulera i Rungego-Kutty-Gilla z zad. K3. Obliczenia powtórzyć dla $k = 25$. Pierwsza metoda działa poprawnie, jeśli $0 \leqslant hk \leqslant 2$, a druga, jeśli $0 \leqslant hk \leqslant 2.8$. Są to *obszary stabilności* obu metod dla tego zagadnienia. Dla danego zagadnienia mamy do wyboru: albo wybrać h z takiego obszaru, albo zmienić metodę na taką, której obszar stabilności jest większy. Jak pokazał Thomas [1986], oba sposoby zawodzą dla dużych k.

8.4. Metody wielokrokowe

Metody z dwóch poprzednich podrozdziałów są *jednokrokowe* w tym sensie, że przechodząc od punktu t do $t + h$, musimy znać tylko jedną wartość rozwiązania, mianowicie $x(t)$. W bardziej efektywnych procedurach wartość funkcji x w nowym punkcie wyraża się przez jej wartości w kilku punktach; stąd nazwa *metody wielokrokowe*. Wyjaśnijmy najpierw ogólnie, jak one powstają. Szukamy rozwiązania zagadnienia początkowego

$$x' = f(t, x), \quad x(t_0) = x_0$$

w punktach $t_1, t_2, \ldots$, niekoniecznie równoodległych. Całkowanie obu stron równania różniczkowego daje równość

$$x(t_{n+1}) = x(t_n) + \int_{t_n}^{t_{n+1}} f(t, x(t))\, dt. \tag{8.4.1}$$

Całkę po prawej stronie można obliczać numerycznie, stosując kwadraturę z węzłami t_i.

Wzór Adamsa-Bashfortha

Niech będzie $f_i := f(t_i, x_i)$, gdzie x_i jest przybliżoną wartością rozwiązania zagadnienia początkowego w punkcie t_i. Wynikające z (8.4.1) wzory

$$x_{n+1} = x_n + af_n + bf_{n-1} + cf_{n-2} + \dots \tag{8.4.2}$$

nazywamy *wzorami Adamsa-Bashfortha*. Jednym z nich jest wzór rzędu piątego, poprawny przy założeniu, że $t_i = t_0 + ih$ $(i \geqslant 1)$:

$$x_{n+1} = x_n + \tfrac{1}{720}h(1901f_n - 2774f_{n-1} + 2616f_{n-2} - \\ - 1274f_{n-3} + 251f_{n-4}). \tag{8.4.3}$$

Łatwo się domyślić, że wynika on z następującego przybliżenia całki z (8.4.1):

$$\int_{t_n}^{t_{n+1}} f(t, x(t))\,dt \approx h(Af_n + Bf_{n-1} + Cf_{n-2} + Df_{n-3} + Ef_{n-4}).$$

Współczynniki $A, B, \dots, E$ tej kombinacji liniowej określamy tak, aby powyższa równość przybliżona była dokładna, gdy tylko $f(t, x(t))$ jest wielomianem klasy Π_4. Można przy tym ograniczyć się do przypadku, gdy $t_n = 0$ i $h = 1$ (zob. zad. 1). Wystarczy też oczywiście zażądać spełnienia tej równości dla wielomianów tworzących bazę przestrzeni Π_4, a więc np. dla

$$p_n(t) := \prod_{j=0}^{n-1}(t + j) \quad (0 \leqslant n \leqslant 4).$$

Warunki

$$\int_0^1 p_n(t)\,dt = Ap_n(0) + Bp_n(-1) + Cp_n(-2) + Dp_n(-3) + Ep_n(-4)$$

dla tychże n dają układ równań

$$\begin{aligned} A + B + C + D + E &= 1, \\ -B - 2C - 3D - 4E &= \tfrac{1}{2}, \\ 2C + 6D + 12E &= \tfrac{5}{6}, \\ -6D - 24E &= \tfrac{9}{4}, \\ 24E &= \tfrac{251}{30}. \end{aligned}$$

Z tych równań, począwszy od ostatniego, wyznaczamy kolejno takie E, D, $\dots$, A, jakie podano w (8.4.3). Zastosowano tu, jak widać, metodę nieoznaczonych współczynników, znaną z podrozdz. 7.2.

Wzory Adamsa-Moultona

W praktyce numerycznej wzory Adamsa-Bashfortha są rzadko używane samodzielnie. Aby polepszyć dokładność, stosuje się je wraz z pewnymi innymi wzorami. Otrzymujemy je, przyjmując, że wartość przybliżona całki z (8.4.1) może zależeć także od f_{n+1}. To daje, zamiast (8.4.2), *wzory Adamsa-Moultona*

$$x_{n+1} = x_n + af_{n+1} + bf_n + cf_{n-1} + \dots$$

Do tej klasy należy wzór rzędu piątego poprawny, jak (8.4.3), dla $t_i = t_0 + ih$:

$$x_{n+1} = x_n + \tfrac{1}{720}h(251f_{n+1} + 646f_n - 264f_{n-1} + 106f_{n-2} - 19f_{n-3}). \tag{8.4.4}$$

Dowód może być taki sam jak dla wzoru poprzedniego typu.

Zauważmy, że w (8.4.4) szukane x_{n+1} występuje także po prawej stronie, w f_{n+1}. Dlatego stosujemy kombinację dwóch wzorów: wzór Adamsa-Bashfortha (8.4.3) daje wstępną wartość, np. x^*_{n+1}, szukanego x_{n+1}, a wzór Adamsa-Moultona z $f(t_{n+1}, x^*_{n+1})$ po prawej stronie daje poprawioną wartość. Jest to metoda typu *predyktor-korektor*, czyli *metoda ekstrapolacyjno-interpolacyjna*.

Szczególny sposób postępowania jest potrzebny na początku obliczeń, gdy znamy tylko x_0. Wartości x_1, x_2, x_3, x_4 możemy znaleźć metodą Rungego-Kutty. Na ogół wszystkie stosowane wzory powinny mieć ten sam rząd. Dlatego wzory (8.4.3) i (8.4.4) zaleca się poprzedzić jednym ze wzorów Rungego-Kutty piątego rzędu, określonych w zad. 8.3.K5.

Wstępną wartość x_{n+1} niezbędną w (8.4.4) można znaleźć jeszcze inaczej. Jest to przecież punkt stały odwzorowania

$$\varphi(z) := \frac{251}{720}hf(t_{n+1}, z) + C,$$

gdzie C jest sumą pozostałych składników prawej strony tego wzoru. Można zatem zastosować metodę iteracyjną opisaną wzorem

$$z_{k+1} := \varphi(z_k) \quad (k \geqslant 0) \tag{8.4.5}$$

(podrozdz. 3.4) i zbieżną do punktu stałego ξ, gdy $|\varphi'(\xi)| < 1$, a przybliżenie początkowe jest dostatecznie dobre. W rozważanym tu przypadku jest

$$\varphi'(z) = \frac{251}{720}h\frac{\partial f(t_{n+1}, z)}{\partial z}$$

i wybór odpowiednio małego h gwarantuje zbieżność metody. W praktyce jeden lub dwa kroki metody (8.4.5) dają już x_{n+1} z dobrą dokładnością.

Analiza liniowych metod wielokrokowych

Rozważymy teraz teorię metod opisanych ogólnym wzorem

$$a_k x_n + a_{k-1}x_{n-1} + \ldots + a_0 x_{n-k} = h(b_k f_n + b_{k-1}f_{n-1} + \ldots + b_0 f_{n-k}). \tag{8.4.6}$$

Taki wzór, o danych współczynnikach a_i, b_i, gdzie $a_k \neq 0$, opisuje metodę *k-krokową*. Założymy znów, że $t_i = t_0 + ih$; x_i jest przybliżoną wartością rozwiązania w t_i, a $f_i := f(t_i, x_i)$. Za pomocą (8.4.6) chcemy obliczyć x_n, znając $x_0, x_1, \ldots, x_{n-1}$. Jeśli $b_k = 0$ (jak we wzorze Adamsa-Bashfortha), to metoda jest *jawna* i za pomocą (8.4.6) można obliczyć x_n. W przeciwnym razie (jak dla wzoru Adamsa-Moultona) metoda jest *niejawna*, bo w (8.4.6) x_n występuje po obu stronach.

W związku z (8.4.6) określamy funkcjonał liniowy L wzorem

$$Lx := \sum_{i=0}^{k} [a_i x(ih) - hb_i x'(ih)]. \tag{8.4.7}$$

Aby uprościć oznaczenia, przyjęliśmy tu, że $n = k$ i $t_0 = 0$. Funkcjonał L jest określony dla dowolnej funkcji różniczkowalnej. W dalszym ciągu przyjmiemy jednak mocniejsze założenie: funkcja x rozwija się w szereg Taylora w punkcie 0. Wtedy Lx można wyrazić w postaci

$$Lx = d_0 x(0) + d_1 h x'(0) + d_2 h^2 x''(0) + \ldots \tag{8.4.8}$$

Współczynniki d_i wynikają stąd, że

$$x(ih) = \sum_{j=0}^{\infty} \frac{(ih)^j}{j!} x^{(j)}(0), \qquad x'(ih) = \sum_{j=0}^{\infty} \frac{(ih)^j}{j!} x^{(j+1)}(0).$$

Podstawienie tych wyrażeń do (8.4.7) daje po uporządkowaniu względem potęg parametru h wzory

$$d_0 = \sum_{i=0}^{k} a_i, \quad d_j = \sum_{i=0}^{k} \left(\frac{i^j}{j!} a_i - \frac{i^{j-1}}{(j-1)!} b_i \right) \qquad (j \geqslant 1). \tag{8.4.9}$$

Twierdzenie 8.4.1. *Następujące własności metody wielokrokowej* (8.4.6) *są równoważne:*

1. $d_0 = d_1 = \ldots = d_m = 0$.
2. $Lp = 0$ *dla każdego wielomianu* $p \in \Pi_m$.
3. $Lx = \mathcal{O}(h^{m+1})$ *dla każdej funkcji* $x \in C^{m+1}$.

Dowód. Jeśli L ma własność **1**, to

$$Lx = d_{m+1}h^{m+1}x^{(m+1)}(0) + \dots \tag{8.4.10}$$

i prawa strona tej równości znika dla każdego $x \in \Pi_m$ (własność **2**).

Jeśli L ma własność **2**, to ze wzoru Taylora dla funkcji $x \in C^{m+1}$ wynika, że $x = p + r$, gdzie $p \in \Pi_m$ i $r(0) = r'(0) = \dots = r^{(m)}(0) = 0$. Ponieważ $Lp = 0$, więc $Lx = Lr = d_{m+1}h^{m+1}r^{(m+1)}(0) + \dots = \mathcal{O}(h^{m+1})$ (własność **3**). Wreszcie, jeśli L ma własność **3**, to w (8.4.8) jest $d_0 = d_1 = \dots = d_m = 0$ (własność **1**). ■

Rząd metody wielokrokowej (8.4.6) jest liczbą naturalną m taką, że w (8.4.8) jest $d_0 = d_1 = \dots = d_m = 0$ i $d_{m+1} \neq 0$.

Przykład 8.4.2. Znaleźć rząd metody opisanej wzorem

$$x_n - x_{n-2} = \tfrac{1}{3}h(f_n + 4f_{n-1} + f_{n-2}).$$

Rozwiązanie. W tym przypadku

$$(a_0, a_1, a_2) = (-1, 0, 1), \quad (b_0, b_1, b_2) = (\tfrac{1}{3}, \tfrac{4}{3}, \tfrac{1}{3}),$$

wobec czego:

$$\begin{aligned}
d_0 &= a_0 + a_1 + a_2 = 0,\\
d_1 &= -b_0 + (a_1 - b_1) + (2a_2 - b_2) = 0,\\
d_2 &= (\tfrac{1}{2}a_1 - b_1) + (2a_2 - 2b_2) = 0,\\
d_3 &= (\tfrac{1}{6}a_1 - \tfrac{1}{2}b_1) + (\tfrac{4}{3}a_2 - 2b_2) = 0,\\
d_4 &= (\tfrac{1}{24}a_1 - \tfrac{1}{6}b_1) + (\tfrac{2}{3}a_2 - \tfrac{4}{3}b_2) = 0,\\
d_5 &= (\tfrac{1}{120}a_1 - \tfrac{1}{24}b_1) + (\tfrac{4}{15}a_2 - \tfrac{2}{3}b_2) = -\tfrac{1}{90}.
\end{aligned}$$

Rząd metody jest równy 4. ■

Na ogół preferujemy metody wysokiego rzędu. Wśród metod postaci (8.4.6) istnieje metoda rzędu $2k$, tj. taka, że $d_0 = d_1 = \dots = d_{2k} = 0$. Ten warunek wobec (8.4.9) daje układ $2k + 1$ równań liniowych jednorodnych względem $2k + 2$ niewiadomych a_i, b_i. Taki układ ma nietrywialne rozwiązanie. Dahlquist [1956] wykazał, że istnieje takie rozwiązanie, iż $a_k \neq 0$ (co założyliśmy). W podrozdziale 8.5 przekonamy się jednak, że oprócz rzędu metody wielokrokowej trzeba brać pod uwagę także inne jej cechy, a przede wszystkim stabilność. Również Dahlquist udowodnił, że stabilna metoda k-krokowa nie może mieć rzędu większego od $k + 2$.

ZADANIA 8.4

1. Wykazać, że jeśli wzór

$$\int_0^1 f(x)\,dx \approx \sum_{i=-n}^{0} A_i f(i)$$

jest dokładny dla każdego $f \in \Pi_m$, to tę samą własność ma wzór

$$\int_{t_0}^{t_0+h} f(x)\,dx \approx h \sum_{i=-n}^{0} A_i f(t_0 + ih).$$

2. Znaleźć niejawną metodę wielokrokową wynikającą ze wzoru Simpsona.

3. Stosując metodę nieoznaczonych współczynników, udowodnić wzór (8.4.4).

4. Znaleźć jednokrokowy wzór Adamsa-Moultona i sprawdzić, że jest on równoważny wzorowi trapezów (podrozdz. 7.2).

5. Sprawdzić następujące wzory:

(**a**) wzór Adamsa-Bashfortha rzędu drugiego

$$x_{n+1} = x_n + \tfrac{1}{2}h(3f_n - f_{n-1}),$$

(**b**) wzór Adamsa-Bashfortha rzędu czwartego

$$x_{n+1} = x_n + \tfrac{1}{24}h(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}),$$

(**c**) wzór Adamsa-Moultona rzędu czwartego

$$x_{n+1} = x_n + \tfrac{1}{24}h(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}).$$

6. Wiadomo, że wzór

$$x_{n+1} = (1-A)x_n + Ax_{n-1} + \tfrac{1}{12}h[(5-A)x'_{n+1} + 8(1+A)x'_n + (5A-1)x'_{n-1}]$$

z dowolnym A jest dokładny dla wszystkich wielomianów klasy Π_m. Znaleźć to m, jak również takie A, żeby wzór był dokładny w Π_{m+1}.

7. Znaleźć współczynniki metody wielokrokowej typu

$$x_{n+1} = x_n + h(Af_n + Bf_{n-2} + Cf_{n-4}),$$

która byłaby dokładna dla wszystkich funkcji $f(t,x)$ postaci $a + bt + ct^2$.

8. Stosując metodę nieoznaczonych współczynników, znaleźć następujące wzory:

(**a**) wzór Adamsa-Moultona $x_{n+1} = x_n + h(Af_{n+1} + Bf_n)$,

(**b**) $x_{n+1} = x_n + h(Af_n + Bf_{n-1} + Cf_{n-2})$,

(**c**) $x_{n+1} = x_n + h(Af_{n+1} + Bf_n + Cf_{n-1})$.

9. Udowodnić, że dla metody Eulera $d_0 = d_1 = 0$ i $d_j = 1/j!$ $(j \geqslant 2)$.

10. Obliczając $d_0, d_1, \ldots$, znaleźć rząd metody opisanej wzorem:

(**a**) $x_n = x_{n-2} + 2hf_{n-1}$,

(**b**) $x_n = x_{n-3} + \frac{3}{8}h(f_n + 3f_{n-1} + 3f_{n-2} + f_{n-3})$.

11. Niech dane $a_0, a_1, \ldots, a_k$ będą takie, że ich suma jest równa 0. Czy istnieją takie $b_0, b_1, \ldots, b_k$, że rząd metody wielokrokowej (8.4.6) wynosi co najmniej k?

12. Znaleźć taki wzór (8.4.6), gdzie $k = 2$, którego rząd wynosi 4.

ZADANIA KOMPUTEROWE 8.4

K1. Zaprogramować metodę opartą na wzorach rzędu czwartego z zad. 5(b) i 5(c) oraz na metodzie Rungego-Kutty tegoż rzędu. Sprawdzić ją dla zagadnienia początkowego $y' = -2xy^2$, $y(0) = 1$, przyjmując $h = 0.25$. Obliczyć wartości rozwiązania w punktach $0.25j$ dla $1 \leqslant j \leqslant 4$ i porównać z dokładnym rozwiązaniem $y = 1/(1 + x^2)$.

K2. Napisać procedurę stosującą wzory Adamsa-Bashfortha i Adamsa-Moultona rzędu piątego oraz metodę Rungego-Kutty tegoż rzędu z zad. 8.3.4. W pamięci powinno się zawsze znajdować tylko pięć ostatnich par (t_i, x_i). Sprawdzić tę procedurę dla zagadnienia początkowego

$$x' = (t - e^{-t})/(x + e^x), \quad x(0) = 0.$$

Rozwiązać je w przedziale $[-1, 1]$ przyjmując $h = 2^{-8}$. Wykazać analitycznie, że dokładne rozwiązanie jest takie, iż $x^2 - t^2 + 2(e^x - e^{-t}) = 0$ i wykorzystać to w programie do kontroli wartości przybliżonych.

8.5. Błędy lokalne i globalne. Stabilność

Zbieżność

Jak już wiemy, zagadnienie początkowe

$$x' = f(t, x), \quad x(t_0) = x_0$$

można rozwiązywać metodami wielokrokowymi. Każdą taką metodę opisuje równość

$$a_k x_n + a_{k-1} x_{n-1} + \ldots + a_0 x_{n-k} = h(b_k f_n + b_{k-1} f_{n-1} + \ldots + b_0 f_{n-k}), \tag{8.5.1}$$

w której $a_k \neq 0$ i $f_i := f(t_i, x_i)$. Wiążą się z nią wielomiany

$$p(z) := a_k z^k + a_{k-1} z^{k-1} + \ldots + a_0,$$
$$q(z) := b_k z^k + b_{k-1} z^{k-1} + \ldots + b_0.$$

Okazuje się, że pewne pożądane własności metody zależą od rozmieszczenia zer tych dwóch wielomianów. Jest oczywiste, że metoda powinna być przede wszystkim zbieżna. Aby zdefiniować tę własność, przyjmijmy, że metodę (8.5.1) stosujemy dla wielu wartości h. Niech $x(h, t)$ będzie przybliżeniem rozwiązania x w punkcie t otrzymanym dla ustalonego h. Metoda jest *zbieżna*, gdy dla każdego t z pewnego przedziału $[t_0, t_m]$

$$\lim_{h \to 0} x(h, t) = x(t), \tag{8.5.2}$$

jeśli tylko wartości początkowe niezbędne w metodzie są takie, że

$$\lim_{h\to 0} x(h, t_0 + nh) = x_0 \quad (0 \leqslant n < k), \tag{8.5.3}$$

a funkcja f spełnia założenia tw. 8.1.4.

Stabilność i zgodność

Określimy jeszcze dwie inne własności metody wielokrokowej (8.5.1). Jest ona *stabilna*, jeśli wszystkie zera wielomianu p leżą na płaszczyźnie zespolonej w kole $|z| \leqslant 1$, a każde zero o module 1 jest pojedyncze. Metoda jest *zgodna*, jeśli $p(1) = 0$ i $p'(1) = q(1)$. Warto przypomnieć, że w podrozdz. 8.4 określono funkcjonał L związany z metodą wielokrokową. Wobec (8.4.9) warunek zgodności oznacza, że w rozwinięciu (8.4.8) tego funkcjonału jest $d_0 = d_1 = 0$, czyli rząd metody zgodnej jest nie mniejszy od 1 (tw. 8.4.1).

Poniższe twierdzenie pokazuje związek określonych już własności metody wielokrokowej.

Twierdzenie 8.5.1. *Metoda wielokrokowa* (8.5.1) *jest zbieżna wtedy i tylko wtedy, gdy jest stabilna i zgodna.*

Dowód. Ograniczymy się do wykazania, że stabilność i zgodność są konieczne na to, żeby metoda była zbieżna; dowód dostateczności tych warunków jest znacznie trudniejszy; zob. Henrici [1962, podrozdz. 5.3].

I. Przypuśćmy, że metoda nie jest stabilna. Wtedy pewne zero λ wielomianu p jest takie, że $|\lambda| > 1$ lub takie, że $|\lambda| = 1$ i $p'(\lambda) = 0$. W obu przypadkach rozważmy proste zagadnienie początkowe

$$x' = 0, \quad x(0) = 0,$$

którego rozwiązaniem jest funkcja x równa tożsamościowo 0. Równość (8.5.1) upraszcza się tu do postaci

$$a_k x_n + a_{k-1} x_{n-1} + \ldots + a_0 x_{n-k} = 0. \tag{8.5.4}$$

Jest to równanie różnicowe liniowe o stałych współczynnikach. Jednym z jego rozwiązań jest ciąg o elementach $x_n = h\lambda^n$. Jeśli $|\lambda| > 1$, to dla $0 \leqslant n < k$ jest

$$|x(h, nh)| = h|\lambda^n| < h|\lambda|^k \to 0 \quad \text{dla } h \to 0.$$

Stąd wynika (8.5.3). Natomiast równość (8.5.2) nie zachodzi, bo jeśli $t = nh$, to $h = tn^{-1}$ i $|x(h,t)| = |x(h,nh)| = tn^{-1}|\lambda|^n \to \infty$.

Jeśli $|\lambda| = 1$ i $p'(\lambda) = 0$, to rozwiązaniem równania (8.5.4) jest ciąg o elementach $x_n = hn\lambda^n$. I tym razem zachodzi (8.5.2), gdyż

$$|x(h, nh)| = hn|\lambda|^n = hn < hk \to 0 \quad \text{dla } h \to 0,$$

a nie zachodzi (8.5.3), gdyż $|x(h,t)| = (tn^{-1})n|\lambda|^n = t \neq 0$.

II. Załóżmy, że metoda (8.5.1) jest zbieżna. Rozważmy najpierw zagadnienie początkowe

$$x' = 0, \quad x(0) = 1,$$

którego rozwiązaniem jest funkcja x równa tożsamościowo 1. Równość (8.5.1) ma znów prostą postać (8.5.4). Pewne rozwiązanie tego równania różnicowego możemy otrzymać, przyjmując, że $x_0 = x_1 = \ldots = x_{k-1} = 1$ i obliczając kolejno $x_k, x_{k+1}, \ldots$ (co jest możliwe, gdyż $a_k \neq 0$). Skoro metoda jest zbieżna, to $\lim_{n\to\infty} x_n = 1$. Stąd i z (8.5.4) wynika, że $a_k + a_{k-1} + \ldots + a_0 = 0$, czyli $p(1) = 0$.

Rozważmy z kolei zagadnienie początkowe

$$x' = 1, \quad x(0) = 0$$

mające rozwiązanie $x(t) = t$. Równanie (8.5.1) jest teraz następujące:

$$a_k x_n + a_{k-1}x_{n-1} + \ldots + a_0 x_{n-k} = h(b_k + b_{k-1} + \ldots + b_0).$$

Na mocy I części dowodu metoda jest stabilna, czyli $p'(1) \neq 0$; wcześniej wykazano, że $p(1) = 0$. Pewne rozwiązanie powyższego równania różnicowego wyraża się wzorem $x_n = (n+k)h\gamma$, gdzie $\gamma := q(1)/p'(1)$. Istotnie, podstawienie takich x_n do lewej strony równania daje

$$\begin{aligned} &h\gamma[a_k(n+k) + a_{k-1}(n+k-1) + \ldots + a_0 n] = \\ &\quad = nh\gamma(a_k + a_{k-1} + \ldots + a_0) + h\gamma[ka_k + (k-1)a_{k-1} + \ldots + a_1] = \\ &\quad = nh\gamma p(1) + h\gamma p'(1) = h\gamma p'(1) = hq(1) = h(b_k + b_{k-1} + \ldots + b_0). \end{aligned}$$

Zauważmy, że $\lim_{h\to 0}(n+k)h\gamma = 0$ dla $n = 0, 1, \ldots, k-1$. Natomiast założenie zbieżności znaczy tyle, że $\lim_{n\to\infty} x_n = t$, gdy $nh = t$. Tak więc $\lim_{n\to\infty}(n+k)h\gamma = t$. Stąd $\gamma = 1$, czyli $p'(1) = q(1)$. ■

Metoda Milne'a

Zastosujmy tw. 8.5.1 do *metody Milne'a* określonej równością

$$x_n - x_{n-2} = \tfrac{1}{3}h(f_n + 4f_{n-1} + f_{n-2}). \tag{8.5.5}$$

Dla tej metody niejawnej jest

$$p(z) = z^2 - 1, \quad q(z) = \tfrac{1}{3}(z^2 + 4z + 1).$$

Wielomian p ma zera pojedyncze: -1 i 1. Prócz tego $p'(z) = 2z$, $p'(1) = 2$ i $q(1) = 2$. Metoda jest zgodna i stabilna, a zatem – na mocy tego twierdzenia – zbieżna.

Błąd lokalny metody

Zbadajmy teraz, jaki jest błąd lokalny metody wielokrokowej. Z równania (8.5.1) chcemy wyznaczyć x_n, zakładając, że wszystkie poprzednie wartości rozwiązania są dokładne, tj. że $x_i = x(t_i)$ dla $i < n$. *Błąd lokalny* jest z definicji równy $x(t_n) - x_n$. Wynika on z dyskretyzacji zagadnienia, która polega na zastąpieniu równania różniczkowego równaniem różnicowym. Błędów zaokrągleń nie bierzemy pod uwagę, przyjmując, że wszystkie działania wynikające z (8.5.1) są wykonywane dokładnie. Chcemy udowodnić, że jeśli metoda jest rzędu m (w sensie określonym w podrozdz. 8.4), to błąd lokalny metody wynosi $\mathcal{O}(h^{m+1})$. Wymaga to jednak założenia pewnej gładkości funkcji f i rozwiązania x.

Twierdzenie 8.5.2. *Jeśli metoda wielokrokowa* (8.5.1) *jest rzędu m, jeśli $x \in C^{m+2}$, a funkcja $\partial f/\partial x$ jest ciągła, to*

$$x(t_n) - x_n = \frac{d_{m+1}}{a_k} h^{m+1} x^{(m+1)}(t_{n-k}) + \mathcal{O}(h^{m+2}), \tag{8.5.6}$$

gdzie współczynnik d_{m+1} jest określony wzorem (8.4.9).

Dowód. Dowód wystarczy przeprowadzić dla $n = k$. Używamy funkcjonału liniowego L zdefiniowanego w (8.4.7):

$$Lx = \sum_{i=0}^{k} [a_i x(t_i) - h b_i x'(t_i)] = \sum_{i=0}^{k} [a_i x(t_i) - h b_i f(t_i, x(t_i))].$$

Z drugiej strony, rozwiązanie numeryczne jest takie, że

$$0 = \sum_{i=0}^{k} [a_i x_i - h b_i f(t_i, x_i)].$$

Ponieważ założyliśmy, że $x_i = x(t_i)$ dla $i < k$, więc odejmując stronami dwa ostatnie równania, otrzymujemy następujący wynik:

$$Lx = a_k [x(t_k) - x_k] - h b_k [f(t_k, x(t_k)) - f(t_k, x_k)].$$

Do drugiej różnicy w nawiasach kwadratowych stosujemy twierdzenie o wartości średniej:

$$Lx = a_k[x(t_k)-x_k]-hb_k\frac{\partial f}{\partial x}(t_k,\xi)[x(t_k)-x_k] = (a_k-hb_kF)[x(t_k)-x_k],$$

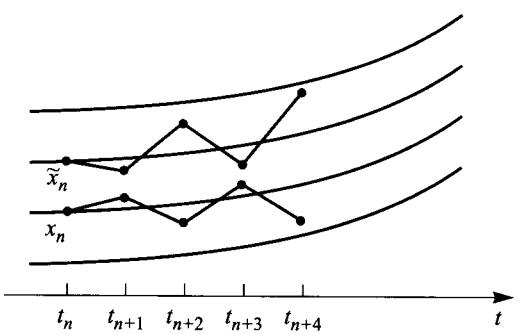
gdzie F oznacza występującą tu pochodną cząstkową. Przypomnijmy jeszcze, że wobec (8.4.10) dla metody rzędu m jest

$$Lx = d_{m+1}h^{m+1}x^{(m+1)}(t_0) + \mathcal{O}(h^{m+2}).$$

Ta równość wraz z poprzednią, w której można odrzucić hb_kF (zob. zad. 4), daje tezę twierdzenia. ■

Błąd globalny metody

Chcemy teraz oszacować *błąd globalny* $x(t_n)-x_n$ popełniany, gdy rozwiązujemy numerycznie zagadnienie początkowe. Jak poprzednio, zakładamy, że wszystkie obliczenia są wykonywane dokładnie, czyli nie ma błędów zaokrągleń. Błąd globalny nie jest po prostu sumą błędów lokalnych związanych z poszczególnymi krokami metody. Istotnie, trzeba pamiętać, że obliczając x_n, korzystamy z wartości rozwiązania znalezionych wcześniej. Jeśli one są przybliżone, to ich błędy przenoszą się na x_n. Dlatego zaczynamy od zbadania, jak różne wartości początkowe wpływają na kształt rozwiązań. Rysunek 8.1 pokazuje możliwą sytuację.



RYS. 8.1. Skutki zmiany warunku początkowego

Rozważamy zagadnienie początkowe

$$x' = f(t,x), \quad x(0) = s \tag{8.5.7}$$

zakładając, że pochodna cząstkowa f_x jest ciągła i że $f_x(t,x) \leqslant \lambda$ w obszarze określonym warunkami $0 \leqslant t \leqslant T$ i $-\infty < x < \infty$. Aby podkreślić

zależność rozwiązania zagadnienia (8.5.7) od s, oznaczamy je $x(t;s)$. Niech będzie $u(t) := \partial x(t;s)/\partial s$. Ta funkcja spełnia *równanie wariacyjne*, które otrzymujemy różniczkując stronami względem s równanie różniczkowe z (8.5.7). To samo robimy z warunkiem początkowym i ostatecznie otrzymujemy zagadnienie

$$u'(t) = f_x(t,x)u, \quad u(0) = 1. \tag{8.5.8}$$

Twierdzenie 8.5.3. *Jeśli $f_x \leqslant \lambda$, to rozwiązanie zagadnienia* (8.5.8) *spełnia nierówność*

$$|u(t)| \leqslant e^{\lambda t} \quad (t \geqslant 0).$$

Dowód. Z (8.5.8) wynika, że $u'/u = f_x = \lambda - \alpha(t)$, gdzie $\alpha(t) \geqslant 0$. Całkując to równanie, otrzymujemy równość

$$\log|u| = \lambda t - A(t), \quad \text{gdzie} \quad A(t) := \int_0^t \alpha(\tau)\,d\tau.$$

Ponieważ $t \geqslant 0$ i $A(t) \geqslant 0$, więc $\log|u| \leqslant \lambda t$, skąd wynika teza twierdzenia. ■

Twierdzenie 8.5.4. *Rozwiązania w punkcie $t \geqslant 0$ zagadnienia* (8.5.7) *z wartościami początkowymi s i $s+\delta$ różnią się co najwyżej o $|\delta|e^{\lambda t}$.*

Dowód. Z definicji funkcji u, twierdzenia o wartości średniej i tw. 8.5.3 wynika, że

$$|x(t;s) - x(t;s+\delta)| = \left|\frac{\partial}{\partial s}x(t,s+\theta\delta)\right||\delta| = |u(t)|\,|\delta| \leqslant |\delta|e^{\lambda t}. \qquad ■$$

Twierdzenie 8.5.5. *Jeśli błąd lokalny metody w punktach $t_1, t_2, \ldots, t_n$ nie przewyższa co do modułu δ, to wartość bezwzględna błędu globalnego w t_n jest nie większa od $\delta(e^{n\lambda h} - 1)/(e^{\lambda h} - 1)$.*

Dowód. Niech błędy lokalne wymienione w twierdzeniu będą równe $\delta_1, \delta_2, \ldots, \delta_n$. Gdy obliczamy x_2, wartość początkowa jest obarczona błędem δ_1. Na mocy tw. 8.5.4 ten błąd zniekształca x_2 co najwyżej o $|\delta_1|e^{\lambda h}$. Do tego trzeba dołączyć błąd lokalny δ_2:

$$|\delta_1|e^{\lambda h} + |\delta_2|.$$

Ta suma zniekształca x_3 co najwyżej o

$$(|\delta_1|e^{\lambda h} + |\delta_2|)e^{\lambda h}.$$

Dochodzi do tego błąd lokalny δ_3. Kontynuując to rozumowanie, wnioskujemy, że błąd globalny przybliżenia x_n nie przewyższa sumy

$$\sum_{k=1}^{n} |\delta_k| e^{(n-k)\lambda h} \leqslant \delta \sum_{k=0}^{n-1} e^{k\lambda h},$$

co daje tezę twierdzenia. ■

Twierdzenie 8.5.6. *Jeśli błędy lokalne metody są rzędu $\mathcal{O}(h^{m+1})$, to błąd globalny jest równy $\mathcal{O}(h^m)$.*

Dowód. Niech w tw. 8.5.5 będzie $\delta = \mathcal{O}(h^{m+1})$. Ponieważ dla małych z jest $e^z - 1 = \mathcal{O}(z)$, a $nh = t$, więc wyrażenie $e^{\lambda h} - 1$ w mianowniku podanego tam wyrażenia pogarsza błąd i daje $\mathcal{O}(h^m)$. ■

ZADANIA 8.5

1. Zbadać, czy poniższe metody wielokrokowe mają własności wymienione w tw. 8.5.1.
   (a) $x_n - x_{n-2} = 2hf_{n-1}$
   (b) $x_n - x_{n-2} = \frac{1}{3}h(7f_{n-1} - 2f_{n-2} + f_{n-3})$
   (c) Metoda Adamsa-Moultona rzędu czwartego z zad. 5(c)
   (d) $x_n + 4x_{n-1} - 5x_{n-2} = h(4f_{n-1} + 2f_{n-2})$
2. Czy istnieją powody do nieufności wobec metody
   $$x_n - 3x_{n-1} + 2x_{n-2} = h(f_n + 2f_{n-1} + f_{n-2} - 2f_{n-3})?$$
3. Które z poniższych metod wielokrokowych są zbieżne?
   (a) $x_n - x_{n-2} = h(f_n - 3f_{n-1} + 4f_{n-2})$
   (b) $x_n - 2x_{n-1} + x_{n-2} = h(f_n - f_{n-1})$
   (c) $x_n - x_{n-1} - x_{n-2} = h(f_n - f_{n-1})$
   (d) $x_n - x_{n-2} = h(f_n - 3f_{n-1} + 2f_{n-2})$
4. Udowodnić, że jeśli $B \neq 0$, to
   $$\frac{Ah^{m+1} + \mathcal{O}(h^{m+2})}{B + Ch} = \frac{A}{B} h^{m+1} + \mathcal{O}(h^{m+2}).$$

8.6. Układy równań. Równania wyższego rzędu

Układ równań różniczkowych rzędu pierwszego ma standardową postać

$$x_i' = f_i(t, x_1, x_2, \ldots, x_n) \quad (1 \leqslant i \leqslant n). \tag{8.6.1}$$

Niewiadomymi są funkcje $x_1, x_2, \ldots, x_n$ zmiennej t. Przykładowy układ

$$x' = x + 4y - e^t, \qquad y' = x + y + 2e^t$$

(w którym użyto prostszych oznaczeń) ma *ogólne* rozwiązanie

$$x = 2ae^{3t} - 2be^{-t} - 2e^t, \qquad y = ae^{3t} + be^{-t} + \tfrac{1}{4}e^t,$$

gdzie a i b są dowolnymi stałymi. Można to łatwo sprawdzić. Układ jest oczywiście liniowy względem x i y.

W dobrze sformułowanym problemie, który – jak się przypuszcza – ma jednoznaczne rozwiązanie, układowi równań różniczkowych powinny towarzyszyć dodatkowe warunki, które by pozwoliły wyznaczyć takie stałe, jak a i b wyżej. Mogą to być warunki początkowe, czyli określenie wartości funkcji x_i w pewnym punkcie t_0. W przykładowym układzie warunki

$$x(0) = 4, \quad y(0) = \tfrac{5}{4}$$

określają już rozwiązanie jednoznacznie:

$$x = 4e^{3t} + 2e^{-t} - 2e^t, \qquad y = 2e^{3t} - e^{-t} + \tfrac{1}{4}e^t.$$

Symbolika wektorowa

Układ (8.6.1) można wyrazić prościej używając symboliki wektorowej. Niech X będzie wektorem kolumnowym o składowych $x_1, x_2, \ldots, x_n$ zależnych od t. Wobec tego X odwzorowuje $\mathbb{R}$ (lub przedział z $\mathbb{R}$) w $\mathbb{R}^n$. Podobnie niech F będzie wektorem kolumnowym o składowych $f_1, f_2, \ldots, f_n$. Każda z nich jest określona na przestrzeni $\mathbb{R}^{n+1}$ (lub jej podzbiorze), czyli F odwzorowuje $\mathbb{R}^{n+1}$ w $\mathbb{R}^n$. Układ (8.6.1) przybiera postać $X' = F(t, X)$. Zagadnienie początkowe jest określone przez ten układ i wartość $X(t_0)$:

$$X' = F(t, X), \quad X(t_0) = X_0. \tag{8.6.2}$$

Równanie różniczkowe

$$y^{(n)} = f(t, y, y', \ldots, y^{(n-1)})$$

rzędu większego od 1 można przekształcić na układ równań różniczkowych rzędu pierwszego, wprowadzając nowe zmienne równe pochodnym funkcji y:

$$x_k' = x_{k+1} \quad (1 \leqslant k \leqslant n-1),$$
$$x_n' = f(t, x_1, x_2, \ldots, x_n).$$

Aby rozwiązać równanie różniczkowe lub układ takich równań stosując standardowe oprogramowanie, trzeba na ogół wyrazić zagadnienie w postaci (8.6.2). Pokażemy na przykładach, jak się to robi.

Przykład 8.6.1. Doprowadzić zagadnienie początkowe

$$(\sin t)y''' + \cos(ty) + \sin(t^2 + y'') + (y')^3 = \log t,$$
$$y(2) = 7, \quad y'(2) = 3, \quad y''(2) = -4$$

do postaci (8.6.2).

Rozwiązanie. Wprowadzamy nowe zmienne jak podano wyżej:

$$x_1' = x_2, \quad x_2' = x_3, \quad x_3' = \frac{\log t - \cos(tx_1) - \sin(t^2 + x_3) - x_2^3)}{\sin t},$$
$$x_1(2) = 7, \quad x_2(2) = 3, \quad x_3(2) = -4. \qquad \blacksquare$$

W taki sam sposób można zmieniać postać układu równań wyższych rzędów:

Przykład 8.6.2. Przekształcić układ

$$(x'')^2 + te^y + y' = x' - x,$$
$$y'y'' - \cos(xy) + \sin(tx'y) = x$$

na układ równań rzędu pierwszego.

Rozwiązanie. Wprowadzenie zmiennych $x_1 = x$, $x_2 = x'$, $x_3 = y$, $x_4 = y'$ daje układ

$$x_1' = x_2, \qquad x_2' = \pm[x_2 - x_1 - t\exp(x_3) - x_4]^{1/2},$$
$$x_3' = x_4, \qquad x_4' = [x_1 + \cos(x_1x_3) - \sin(tx_2x_3)]/x_4$$

(znak prawej strony drugiego równania można wybrać dowolnie; otrzymujemy więc w istocie dwa różne układy równań różniczkowych). $\blacksquare$

Zastosowanie wzoru Taylora dla układu równań

Metodę poznaną w podrozdz. 8.2 można zastosować bez istotnych zmian do układu równań różniczkowych rzędu pierwszego. Wzór Taylora w symbolice wektorowej daje równość przybliżoną

$$X(t+h) \approx X(t) + hX'(t) + \frac{1}{2!}h^2X''(t) + \ldots + \frac{1}{n!}h^nX^{(n)}(t).$$

Występujące tu pochodne trzeba wyrazić analitycznie, posługując się danym układem.

Przykład 8.6.3. Dla zagadnienia początkowego

$$x' = x + y^2 - t^3, \qquad x(1) = 3,$$
$$y' = y + x^3 + \cos t, \quad y(1) = 1$$

podać wzory wynikające ze wzoru Taylora dla $n = 3$. Znaleźć przybliżone rozwiązanie w przedziale $[-2, 1]$, przyjmując, że $h = -0.1$.

Rozwiązanie. Obliczamy potrzebne pochodne wyższych rzędów:

$$x'' = x' + 2yy' - 3t^2,$$
$$y'' = y' + 3x^2x' - \sin t,$$
$$x''' = x'' + 2yy'' + 2(y')^2 - 6t,$$
$$y''' = y'' + 6x(x')^2 + 3x^2x'' - \cos t.$$

Obliczenia można wykonywać według następującego algorytmu:

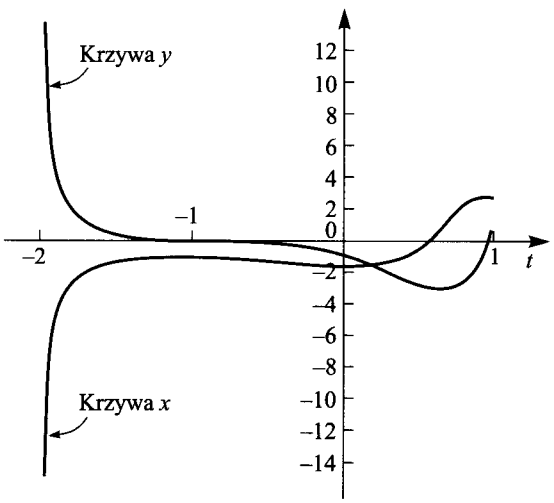
$t \leftarrow 1$; $x \leftarrow 3$; $y \leftarrow 1$; $h \leftarrow -0.1$; $M \leftarrow 30$
output $0, t, x, y$
for $k = 1$ **to** M **do**
  $x' \leftarrow x + y^2 - t^3$
  $y' \leftarrow y + x^3 + \cos t$
  $x'' \leftarrow x' + 2yy' - 3t^2$
  $y'' \leftarrow y' + 3x^2x' - \sin t$
  $x''' \leftarrow x'' + 2yy'' + 2(y')^2 - 6t$
  $y''' \leftarrow y'' + 6x(x')^2 + 3x^2x'' - \cos t$
  $x \leftarrow x + h(x' + \frac{1}{2}h(x'' + \frac{1}{3}hx'''))$
  $y \leftarrow y + h(y' + \frac{1}{2}h(y'' + \frac{1}{3}hy'''))$
  $t \leftarrow t + h$
  output k, t, x, y
end do

Zamiast tablicy wartości pokazano na rys. 8.2 wykresy rozwiązań x i y. ■

Dodajmy jeszcze, że nie tracąc ogólności, można założyć, iż w równaniach (8.6.1) zmienna t jawnie nie występuje. Istotnie, jeśli przyjmiemy, że $x_0 := t$, to $x_0' = 1$, a te równania zmieniamy na $x_i' = f_i(x_0, x_1, \ldots, x_n)$. Daje to układ *autonomiczny*

$$X' = F(X), \tag{8.6.3}$$

w którym $X := (x_0, x_1, \ldots, x_n)$.



RYS. 8.2. Rozwiązanie przykładu 8.6.3

Inne metody

Metody Rungego-Kutty dla układu równań różniczkowych rzędu pierwszego mają najprostszą postać, gdy ten układ jest autonomiczny. Klasyczna metoda rzędu czwartego jest opisana wzorem

$$X(t+h) \approx X(t) + \tfrac{1}{6}(F_1 + 2F_2 + 2F_3 + F_4),$$

gdzie

$$F_1 := hF(X), \qquad F_2 := hF(X + \tfrac{1}{2}F_1),$$
$$F_3 := hF(X + \tfrac{1}{2}F_2), \quad F_4 := hF(X + F_3).$$

W podobny sposób można wyrazić np. metodę Rungego-Kutty-Fehlberga (podrozdz. 8.3).

Na układy równań przenoszą się także metody wielokrokowe. Przykładem jest kombinacja metod Adamsa-Bashfortha (8.4.3) i Adamsa-Moultona (8.4.4):

$$X^*_{n+1} := X_n + \tfrac{1}{720}h(1901F_n - 2774F_{n-1} + 2616F_{n-2} - \\ - 1274F_{n-3} + 251F_{n-4}),$$
$$X_{n+1} := X_n + \tfrac{1}{720}h[251F(X^*_{n+1}) + 646F_n - \\ - 264F_{n-1} + 106F_{n-2} - 19F_{n-3}],$$

gdzie $F_m := F(X_m)$. Jak dla pojedynczych równań, na początku obliczeń musimy za pomocą jakiejś metody jednokrokowej, np. metody Rungego-

-Kutty rzędu piątego (zob. zad. 8.3.4), obliczyć początkowe wartości X_1, X_2, X_3, X_4.

ZADANIA 8.6

1. Znaleźć ogólne rozwiązanie układu

$$x' = 3x - 4y + e^t,$$
$$y' = x - y - e^t.$$

Wskazówka: Wypróbować funkcje postaci e^t, te^t, t^2e^t.

2. Przekształcić zagadnienie początkowe:

(a) $x'' = x\cos t + e^t x' + 3t^2 + 7$, $x(1) = 5$, $x'(1) = 9$,

(b) $y'' + yz = 0$, $z' + 2yz = 4$, $y(0) = 1$, $y'(0) = 0$, $z(0) = 3$

na analogiczne zadanie z równaniami rzędu pierwszego. Użyć symboliki wektorowej.

3. Podane niżej równanie różniczkowe lub układ takich równań przekształcić na układ autonomiczny równań rzędu pierwszego.

(a) $x''' + 2x'' - x' - 2x = e^t$

(b) $x''' - \sin x'' + e^t x' + 2t\cos x = 25$

(c) $x''' - (\sin x'' + e^t x')^2 + \cos x = 0$

(d) $x'' - x'y = 3xy' \log t$, $y'' - 2xy' = 5x'y\sin t$

(e) $x''' - 5tx''y'' + (\log x')z = 0$, $y'' - \sin(ty) + 7tx'' = 0$, $z' + 16ty' - e^t x'z = 0$

ZADANIA KOMPUTEROWE 8.6

K1. Napisać i sprawdzić program rozwiązujący w przedziale $[-1, 1]$ zagadnienie początkowe

$$x_1' = t + x_1^2 + x_2, \qquad x_1(-1) = 0.43,$$
$$x_2' = t^2 - x_1 + x_2^2, \qquad x_2(-1) = -0.69,$$

metodą opartą na wzorze Taylora rzędu trzeciego dla $h = 0.01$.

K2. Wykonać czynności jak w zad. K1 dla zagadnienia początkowego

$$x_1' = \sin x_1 + \cos(tx_2), \qquad x_1(-1) = 2.37,$$
$$x_2' = t^{-1}\sin(tx_1), \qquad x_2(-1) = -3.48.$$

Uwaga: Obliczanie wartości x_2', x_2'', x_2''' trzeba zaprogramować szczególnie starannie ze względu na pozorną osobliwość dla $t = 0$.

K3. Napisać procedurę wykonującą dla danego h jeden krok metody Rungego-Kutty rzędu czwartego dla układu n równań różniczkowych rzędu pierwszego. Sprawdzić ją, rozwiązując w przedziale $[1, 2]$, dla $h = -0.01$, zagadnienie początkowe

$$x' = x^{-2} + \log y + t^2, \qquad x(2) = -2,$$
$$y' = e^y - \cos x + x\sin t - (xy)^{-3}, \qquad y(2) = 1.$$

K4. Napisać i sprawdzić procedurę realizującą metodę adaptacyjną Rungego-Kutty-Fehlberga (podrozdz. 8.3) dla układu równań różniczkowych rzędu pierwszego.

K5. Napisać i sprawdzić procedurę realizującą zestaw metod podany na końcu tego podrozdziału.

K6. Rozwiązać numerycznie w przedziale $[0, 10]$ zagadnienie początkowe

$$x_1' = (12sc - 9c^2 - 1)x_1 + (12c^2 + 9sc)x_2, \qquad x_1(0) = -12,$$
$$x_2' = (9s - 12s^2)x_1 - (12sc + 9s^2 + 1)x_2, \qquad x_2(0) = 6,$$

gdzie $c := \cos 6t$ i $s := \sin 6t$ (Lambert [1973]). Przyjąć $h = 0.01$. Sprawdzić, że rozwiązaniem dokładnym jest $x_1 = e^{-13t}(s - 2c)$, $x_2 = e^{-13t}(2s + c)$. Czy rozwiązanie przybliżone jest dostatecznie dokładne?

K7. Rozwiązać numerycznie w przedziale $[0, 5]$ zagadnienie początkowe

$$x'' + 192x = 0, \quad x(0) = \tfrac{1}{6}, \quad x'(0) = 0.$$

Narysować wykres rozwiązania.

8.7. Zagadnienia brzegowe

W poprzednich podrozdziałach podano metody rozwiązywania zagadnień początkowych. Należy do nich m.in. zagadnienie

$$x'' = f(t, x, x'), \quad x(a) = \alpha, \quad x'(a) = \beta.$$

Inne metody stosujemy w przypadku *zagadnień brzegowych*, w których na rozwiązanie równania różniczkowego rzędu drugiego nakłada się warunki w dwóch punktach. Typowy przykład takiego zagadnienia jest następujący:

$$x'' = f(t, x, x'), \quad x(a) = \alpha, \quad x(b) = \beta. \tag{8.7.1}$$

Najprostsze zagadnienia brzegowe można rozwiązać analitycznie. Tak np. dla zagadnienia

$$x'' = -x, \quad x(0) = 3, \quad x(\pi/2) = 7 \tag{8.7.2}$$

możemy łatwo znaleźć ogólne rozwiązanie równania różniczkowego:

$$x(t) = A \sin t + B \cos t.$$

Pozostaje wyznaczyć stałe A i B z warunków brzegowych:

$$3 = x(0) = A \sin 0 + B \cos 0 = B,$$
$$7 = x(\pi/2) = A \sin(\pi/2) + B \cos(\pi/2) = A,$$

co daje rozwiązanie

$$x(t) = 7 \sin t + 3 \cos t.$$

Istnienie rozwiązania

Na ogół nie wiemy, jak się dokładnie wyraża rozwiązanie ogólne równania różniczkowego. Opis metod numerycznych, które można stosować do zagadnień brzegowych, poprzedzimy pewnymi informacjami teoretycznymi. Nasuwa się od razu pytanie, czy każde zagadnienie (8.7.1) ma rozwiązanie. Prosty przykład pokazuje, że nie zawsze tak jest. Istotnie, zmieńmy w (8.7.2) drugi warunek brzegowy na $x(\pi) = 7$. Wtedy stałe A i B powinny spełniać równania $3 = B$ i $7 = -B$, co oczywiście jest niemożliwe.

Twierdzenia o istnieniu rozwiązań zagadnienia brzegowego (8.7.1) są raczej skomplikowane; zob. Stoer i Bulirsch [1980] lub Keller [1968]. Przytoczymy tylko jedno eleganckie twierdzenie (Keller [1968, s. 108]), dotyczące przypadku, gdy w równaniu różniczkowym nie występuje pierwsza pochodna funkcji x.

Twierdzenie 8.7.1. *Jeśli funkcja $\partial f/\partial x$ jest ciągła, nieujemna i ograniczona w obszarze określonym nierównościami $0 \leqslant t \leqslant 1$, $-\infty < x < \infty$, to zagadnienie brzegowe*

$$x'' = f(t, x), \quad x(0) = x(1) = 0 \tag{8.7.3}$$

ma w przedziale $[0, 1]$ jednoznaczne rozwiązanie.

Przykład 8.7.2. Zastosować tw. 8.7.1 do zagadnienia

$$x'' = (5x + \sin 3x)e^t, \quad x(0) = x(1) = 0.$$

Rozwiązanie. Funkcja

$$\frac{\partial f}{\partial x} = (5 + 3\cos 3x)e^t$$

jest ciągła dla $0 \leqslant t \leqslant 1$, $-\infty < x < \infty$, nieujemna $(5 + 3\cos 3x \geqslant 2)$ i ograniczona z góry przez $8e$. Twierdzenie zapewnia więc, że rozwiązanie zagadnienia brzegowego istnieje i jest określone jednoznacznie. ■

Uproszczenie warunków brzegowych

Rozważmy zagadnienie brzegowe

$$x'' = f(t, x), \quad x(a) = \alpha, \quad x(b) = \beta, \tag{8.7.4}$$

w którym – jak w tw. 8.7.1 – nie występuje pierwsza pochodna x', ale warunki brzegowe są zupełnie ogólne. Takie zagadnienie można zredukować

w prosty sposób do (8.7.3). W tym celu najpierw przekształcamy przedział $[a,b]$ zmiennej t na przedział $[0,1]$ zmiennej s takiej, że $t=(b-a)s+a$. Wprowadzamy też funkcję $y(s):=x(\lambda s+a)$, gdzie $\lambda:=b-a$. Stąd $y'(s)=\lambda x'(\lambda s+a)$ i $y''(s)=\lambda^2 x''(\lambda s+a)$. Wobec tego funkcja x jest rozwiązaniem zagadnienia (8.7.4) wtedy i tylko wtedy, gdy funkcja y jest rozwiązaniem zagadnienia

$$y''(s)=\lambda^2 f(\lambda s+a,y(s)),\quad y(0)=\alpha,\quad y(1)=\beta.$$

Niech $g(s,y)$ oznacza prawą stronę powyższego równania różniczkowego. Mamy więc teraz zagadnienie

$$y''=g(s,y),\quad y(0)=\alpha,\quad y(1)=\beta.$$

Funkcja y jest jego rozwiązaniem wtedy i tylko wtedy, gdy funkcja

$$z(s):=y(s)-[(\beta-\alpha)s+\alpha]$$

jest rozwiązaniem zagadnienia brzegowego

$$z''=h(s,z):=g(s,z+(\beta-\alpha)s+\alpha),\quad z(0)=z(1)=0.$$

Przykład 8.7.3. Przekształcić zagadnienie brzegowe

$$x''=x^2+tx-t^2+3,\quad x(3)=7,\quad x(5)=9 \tag{8.7.5}$$

do standardowej postaci (8.7.3).

Rozwiązanie. Dzięki przekształceniu zmiennej t na s, gdzie $t=2s+3$, warunki brzegowe dotyczą punktów 0 i 1. Prawa strona równania różniczkowego względem funkcji $y(s)=x(2s+3)$ jest równa

$$4[y^2+(2s+3)y-(2s+3)^2+3],$$

czyli nowe zagadnienie brzegowe jest następujące:

$$y''=g(s,y):=4[y^2+(2s+3)y-4s^2-12s-6],\quad y(0)=7,\quad y(1)=9.$$

Drugie przekształcenie wprowadza funkcję niewiadomą

$$z(s)=y(s)-2s-7,$$

która spełnia równanie różniczkowe z prawą stroną równą

$$g(s,z+2s+7)=4[(z+2s+7)^2+(2s+3)(z+2s+7)-4s^2-12s-6]$$

(można ją oczywiście uprościć) i warunki brzegowe $z(0)=z(1)=0$. Rozwiązanie x pierwotnego zagadnienia (8.7.5) wyraża się przez funkcję z wzorem

$$x(t)=z\left(\frac{t-3}{2}\right)+t+4.$$

■

Twierdzenie 8.7.4. *Jeśli funkcja* $f(t,x)$ *jest dla* $0 \leqslant t \leqslant 1$ *i* $-\infty < x < \infty$ *ciągła i taka, że*

$$|f(t,x_1) - f(t,x_2)| \leqslant k|x_1 - x_2| \quad (k < 8), \tag{8.7.6}$$

to zagadnienie brzegowe (8.7.3) *ma w przedziale* $[0,1]$ *jednoznaczne rozwiązanie ciągłe.*

Szkic dowodu. W zadaniu 10 dowodzi się, że to zagadnienie jest równoważne równaniu całkowemu

$$x(t) = \int_0^1 G(t,s)f(s,x(s))\,ds, \tag{8.7.7}$$

gdzie G jest funkcją Greena określoną w zad. 9. To równanie zaś ma postać $x = F(x)$, gdzie F jest operatorem wyrażającym się przez tę całkę. Na mocy twierdzenia Banacha o punkcie stałym (Kantorovič i Akilov [*1977, s. 605]) operator F ma przy założeniu (8.7.6) jedyny punkt stały. ■

Przykład 8.7.5. Wykazać, że zagadnienie

$$x'' = 2\exp(t\cos x), \quad x(0) = x(1) = 0$$

ma jednoznaczne rozwiązanie.

Rozwiązanie. Zastosujmy do funkcji $f(t,x) := 2\exp(t\cos x)$ twierdzenie o wartości średniej. Ponieważ $\partial f/\partial x = -2t\sin x\exp(t\cos x)$, więc

$$|f(t,x_1) - f(t,x_2)| = \left|\frac{\partial f}{\partial x}(t,\xi)\right| |x_1 - x_2| \leqslant 2e|x_1 - x_2| < 8|x_1 - x_2|$$

i tw. 8.7.4 gwarantuje jednoznaczność rozwiązania zagadnienia brzegowego. ■

Zadania 8.7

1. Znaleźć wszystkie pary (α,β), dla których zagadnienie brzegowe:
   (a) $x'' = x$, $x(0) = \alpha$, $x(1) = \beta$,
   (b) $x'' = -x$, $x(0) = \alpha$, $x(\pi) = \beta$,
   ma rozwiązanie.
2. Podać przykład zagadnienia brzegowego, mającego niejednoznaczne rozwiązanie. Wskazówka: Skorzystać z zad. 1a.
3. Rozwiązać zagadnienie brzegowe $x'' - 2x' + x = 0$, $x(0) = \alpha$, $x(1) = \beta$. Czy dla pewnych α, β to zagadnienie nie ma rozwiązania?
4. Rozwiązać następujące zagadnienia brzegowe:

(a) $x'' = x^2$, $x(0) = \frac{2}{3}$, $x(1) = \frac{3}{8}$

(b) $x'' = x^2$, $x(0) = 0$, $x(1) = 1$ (Davis [1962])

(c) $x'' = x^3$, $x(0) = 1$, $x(1) = 2 - \sqrt{2}$

5. Sprawdzić, czy zagadnienie brzegowe:

(a) $x'' = -4x$, $x(0) = 1$, $x(\pi/2) = -1$,

(b) $x'' = -4x$, $x(0) = 1$, $x(\pi/2) = 2$

(i) nie ma rozwiązania, (ii) ma dokładnie jedno rozwiązanie, (iii) ma dokładnie dwa rozwiązania, (iv) ma więcej niż jedno rozwiązanie, (v) nie ma rozwiązania wyrażającego się przez funkcje elementarne.

6. Sprowadzić zagadnienie brzegowe $x'' = \cos(t^2x^2)$, $x(3) = 5$, $x(7) = 12$ do analogicznego zagadnienia z warunkami $z(0) = z(1) = 0$.

7. Niech w zagadnieniu brzegowym występuje równanie $x'' = f(t, x, x')$. Uogólnić na ten przypadek sposób sprowadzenia warunków brzegowych do punktów 0 i 1.

8. Stosując wynik zad. 7 do zagadnienia $x'' = x^2 - 3x' + t$, $x(3) = \alpha$, $x(7) = \beta$, sprowadzić warunki brzegowe do punktów 0 i 1.

9. Wykazać, że rozwiązanie zagadnienia brzegowego $x'' = f(t)$, $x(0) = x(1) = 0$ wyraża się wzorem

$$x(t) = \int_0^1 G(t,s)f(s)\,ds,$$

gdzie

$$G(t,s) := \begin{cases} s(t-1) & (0 \leqslant s \leqslant t \leqslant 1) \\ t(s-1) & (0 \leqslant t \leqslant s \leqslant 1) \end{cases}$$

(jest to *funkcja Greena* dla tego zadania).

10. Wykazać, że zagadnienie brzegowe (8.7.3) jest równoważne równaniu całkowemu (8.7.7) z funkcją G określoną w zad. 9.

11. Udowodnić, że poniższe zagadnienia brzegowe mają jednoznaczne rozwiązania.

(a) $x'' = \cos(tx)$, $x(0) = 1$, $x(1) = 4$

(b) $x'' = (t^3 + 5)x + \sin t$, $x(0) = x(1) = 0$

(c) $x'' = \operatorname{arctg} x + 2x + \cos t$, $x(0) = x(1) = 0$

(d) $x'' = \frac{1}{2}\exp\left[\frac{1}{2}(t+1)\cos(x - 3t + 7)\right]$, $x(-1) = -10$, $x(1) = -4$

12. Dla jakich wyliczonych niżej funkcji $f(t,x)$ można być pewnym, że zagadnienie brzegowe (8.7.3) ma jednoznaczne rozwiązanie?

(a) $x^{1/3}$, (b) t/x, (c) $|tx|$, (d) t^2x^3, (e) $tx^{4/3}$, (f) $t^2(1+x^2)^{-1}$, (g) $\log(1+x^2)$, (h) $t\sin x$, (i) $(\operatorname{tg} t)\operatorname{tg} x$, (j) $t\operatorname{arctg} x$.

8.8. Zagadnienia brzegowe: metody strzału

Jedną z naturalnych metod rozwiązywania zagadnienia brzegowego

$$x'' = f(t, x, x'), \quad x(a) = \alpha, \quad x(b) = \beta \tag{8.8.1}$$

jest zamiana go na zagadnienie początkowe z sensownie dobraną wartością początkową $x'(a)$. Rozwiązujemy to nowe zagadnienie, co daje jakąś wartość $x(b)$. Jeśli wybór $x'(a)$ był trafny, to $x(b) = \beta$ i obliczenia są zakończone. W przeciwnym razie wybieramy inne $x'(a)$ i znów rozwiązujemy zagadnienie początkowe. Na tym – z grubsza – polegają *metody strzału*.

Niech z będzie próbną wartością $x'(a)$, a x_z – rozwiązaniem zagadnienia początkowego

$$x'' = f(t, x, x'), \quad x(a) = \alpha, \quad x'(a) = z. \tag{8.8.2}$$

Naszym celem jest taki wybór parametru z, żeby było $x_z(b) = \beta$. Inaczej mówiąc, chcemy rozwiązać równanie $\varphi(z) = 0$, gdzie

$$\varphi(z) := x_z(b) - \beta.$$

W rozdziale 3 opisano wiele metod, które można tu zastosować: metodę bisekcji, siecznych, Newtona itd. Wybierając jedną z nich, trzeba pamiętać, że obliczenie każdej wartości funkcji φ jest kosztowne, bo wymaga rozwiązania pewnego zagadnienia początkowego.

Metoda siecznych

Przypomnijmy, na czym ta metoda polega. Gdy znamy już dwie wartości, np. $\varphi(z_1)$ i $\varphi(z_2)$, to zakładamy, że funkcja φ jest liniowa i wyznaczamy z_3 jako punkt, w którym ona znika. W następnym kroku postępujemy podobnie, posługując się wartościami $\varphi(z_2)$ i $\varphi(z_3)$. Ogólnie, dla danych z_1 i z_2 określamy następne przybliżenia pierwiastka, stosując rekurencyjnie wzór

$$z_n := z_{n-1} - \frac{z_{n-1} - z_{n-2}}{\varphi(z_{n-1}) - \varphi(z_{n-2})}\,\varphi(z_{n-1}) \quad (n > 2).$$

Gdy w ten sposób otrzymamy wiele wartości funkcji φ dostatecznie bliskich 0, możemy zmienić metodę na lepszą. Niech wszystkie $\varphi(z_i)$ dla $1 \leqslant i \leqslant n$ będą małe. Wtedy stosujemy *interpolację odwrotną*, tj. znajdujemy wielomian $p \in \Pi_{n-1}$ taki, że $p(\varphi(z_i)) = z_i$ $(1 \leqslant i \leqslant n)$ i przyjmujemy $z_{n+1} := p(0)$. Powodzenie tej metody zależy od tego, czy funkcja φ ma różniczkowalną funkcję odwrotną w otoczeniu pierwiastka. To z kolei wymaga, aby ten pierwiastek był pojedynczy.

Zagadnienie liniowe

Naszkicowana już metoda strzału jest kosztowna i wszystkie sposoby skrócenia obliczeń są bardzo ważne. W szczególności warto wykorzystać każdą informację o poprawnej wartości $x'(a)$. Można też na początku obliczeń rozwiązywać zagadnienia początkowe z dużym h i zmniejszać h dopiero po otrzymaniu małych wartości funkcji φ.

Jeśli funkcja φ jest liniowa, to metoda siecznych daje dokładne rozwiązanie w jednym kroku. Jest tak, gdy równanie różniczkowe jest liniowe:

$$x'' = u + vx + wx'. \tag{8.8.3}$$

Warto przypomnieć dwa ważne i potrzebne tu twierdzenia dotyczące równań różniczkowych liniowych rzędu drugiego (zob. Coddington i Levinson [1955]).

Twierdzenie 8.8.1. *Jeśli funkcje u, v, w zmiennej t są ciągłe w przedziale $[a, b]$, to dla dowolnych liczb rzeczywistych α, α' zagadnienie początkowe*

$$x'' = u + vx + wx', \quad x(a) = \alpha, \quad x'(a) = \alpha'$$

ma w tym przedziale jednoznaczne rozwiązanie.

Twierdzenie 8.8.2. *Każde rozwiązanie równania niejednorodnego* (8.8.3) *jest sumą $x_0 + c_1x_1 + c_2x_2$, gdzie x_0 jest jego dowolnym szczególnym rozwiązaniem, a x_1 i x_2 są niezależnymi liniowo rozwiązaniami równania jednorodnego $x'' = vx + wx'$.*

Rozważmy teraz zagadnienie brzegowe

$$x'' = u(t) + v(t)x + w(t)x', \quad x(a) = \alpha, \quad x(b) = \beta, \tag{8.8.4}$$

gdzie – jak w tw. 8.8.1 – funkcje u, v, w są ciągłe w przedziale $[a, b]$. Przyjmijmy, że powyższe równanie różniczkowe rozwiązano dwukrotnie, z różnymi próbnymi wartościami $x'(a)$. Ściślej, niech te dwa rozwiązania, x_1 i x_2, będą takie, że

$$x_1(a) = x_2(a) = \alpha, \quad x_1'(a) = z_1, \quad x_2'(a) = z_2. \tag{8.8.5}$$

Ponieważ równanie różniczkowe jest liniowe, więc dla każdego λ kombinacja liniowa

$$y := \lambda x_1 + (1 - \lambda)x_2 \tag{8.8.6}$$

tych rozwiązań też spełnia to równanie; jest również $y(a) = \alpha$. Wybieramy λ tak, żeby było $y(b) = \beta$:

$$\beta = \lambda x_1(b) + (1 - \lambda)x_2(b),$$

czyli

$$\lambda = \frac{\beta - x_2(b)}{x_1(b) - x_2(b)}. \tag{8.8.7}$$

Realizując ten pomysł w praktyce dla zadania liniowego (8.8.4), obliczamy funkcje x_1 i x_2 jednocześnie. Obie mają spełniać równanie różniczkowe i mieć wartość α dla $t = a$. Prócz tego można przyjąć, że pierwsza z nich spełnia warunek $x_1'(a) = 0$, a druga – warunek $x_2'(a) = 1$.

Twierdzenie 8.8.3. *Jeśli liniowe zagadnienie brzegowe* (8.8.4) *ma rozwiązanie, to albo jest nim funkcja* x_1, *albo* $x_1(b) \neq x_2(b)$ *i rozwiązaniem jest funkcja* (8.8.6) *dla* λ *określonego w* (8.8.7).

Dowód. Niech funkcje y_0, y_1, y_2 będą rozwiązaniami następujących zagadnień początkowych:

$$\begin{aligned} y_0'' &= u + vy_0 + wy_0', & y_0(a) &= \alpha, & y_0'(a) &= 0, \\ y_1'' &= vy_1 + wy_1', & y_1(a) &= 1, & y_1'(a) &= 0, \\ y_2'' &= vy_2 + wy_2', & y_2(a) &= 0, & y_2'(a) &= 1. \end{aligned}$$

Z twierdzenia 8.8.2 wynika, że ogólne rozwiązanie równania różniczkowego w (8.8.4) jest równe $y_0 + c_1y_1 + c_2y_2$, gdzie c_1 i c_2 są dowolnymi stałymi. Wobec tego funkcje x_1 i x_2 spełniające warunki początkowe (8.8.5) można wyrazić tak:

$$x_1 = y_0 + z_1y_2, \quad x_2 = y_0 + z_2y_2. \tag{8.8.8}$$

Ponieważ założyliśmy, że zagadnienie (8.8.4) ma rozwiązanie, więc istnieją stałe c_1 i c_2 takie, że

$$\begin{aligned} \alpha &= y_0(a) + c_1y_1(a) + c_2y_2(a), \\ \beta &= y_0(b) + c_1y_1(b) + c_2y_2(b). \end{aligned}$$

Pierwsze równanie upraszcza się do równości $c_1 = 0$, musi zatem istnieć c_2 takie, że

$$\beta = y_0(b) + c_2y_2(b). \tag{8.8.9}$$

Jeśli $x_1(b) \neq x_2(b)$, to funkcja y określona wzorami (8.8.6) i (8.8.7) jest rozwiązaniem zagadnienia (8.8.4). W przeciwnym razie z (8.8.8) wynika, że $y_2(b) = 0$, a z (8.8.9) – że $y_0(b) = \beta$, czyli x_1 jest rozwiązaniem tego zagadnienia. ∎

Metoda Newtona

Powróćmy teraz do ogólnego (nieliniowego) zagadnienia brzegowego (8.8.1) i zastanówmy się, jak można tu zastosować metodę Newtona. Niech funkcja x_z będzie, jak przedtem, rozwiązaniem zagadnienia początkowego (8.8.2). Parametr z ma być pierwiastkiem równania $\varphi(z) := x_z(b) - \beta = 0$. *Metoda Newtona* wyraża się wzorem

$$z_{n+1} := z_n - \frac{\varphi(z_n)}{\varphi'(z_n)}.$$

Aby wyznaczyć pochodną φ', zróżniczkujmy stronami względem z wszystkie trzy równości (8.8.2):

$$\frac{\partial x''}{\partial z} = \frac{\partial f}{\partial t}\frac{\partial t}{\partial z} + \frac{\partial f}{\partial x}\frac{\partial x}{\partial z} + \frac{\partial f}{\partial x'}\frac{\partial x'}{\partial z}, \quad \frac{\partial x(a)}{\partial z} = 0, \quad \frac{\partial x'(a)}{\partial z} = 1.$$

Łatwe uproszczenia i wprowadzenie funkcji $v := \partial x/\partial z$ dają związki

$$v'' = f_x(t, x, x')v + f_{x'}(t, x, x')v', \quad v(a) = 0, \quad v'(a) = 1$$

opisujące łącznie pewne zagadnienie początkowe. Równanie różniczkowe, które w nim występuje, nazywamy *pierwszym równaniem wariacyjnym*. Rozwiązujemy to zagadnienie wraz z (8.8.2) i dostajemy na końcu wartość $v(b)$ równą z definicji $\partial x_z(b)/\partial z = \varphi'(z)$ (trzeba pamiętać, że równości (8.8.2) i wnioski z nich dotyczą funkcji x_z). To już pozwala zastosować metodę Newtona.

Metoda wielocelowa strzałów

Metoda wielocelowa strzałów polega na tym, że dzielimy przedział $[a, b]$ na podprzedziały i w każdym z nich rozwiązujemy pewne zagadnienie początkowe. Przyjmijmy najpierw, że podprzedziały są tylko dwa: $[a, c]$ i $[c, b]$. Wtedy zagadnienie brzegowe (8.8.1) generuje dwa zagadnienia początkowe:

$$x_1'' = f(t, x_1, x_1'), \quad x_1(a) = \alpha, \quad x_1'(a) = z_1 \quad (a \leqslant t \leqslant c),$$
$$x_2'' = f(t, x_2, x_2'), \quad x_2(b) = \beta, \quad x_2'(b) = z_2 \quad (c \leqslant t \leqslant b).$$

Drugie z nich rozwiązujemy w kierunku malejącego t.

Parametry z_1 i z_2, które mamy do dyspozycji, wybieramy tak, aby rozwiązanie

$$x(t) := \begin{cases} x_1(t) & (a \leqslant t \leqslant c) \\ x_2(t) & (c \leqslant t \leqslant b) \end{cases}$$

było ciągłe i miało ciągłą pochodną w c. Ma zatem być $x_1(c) = x_2(c)$ i $x_1'(c) = x_2'(c)$. Ten układ dwóch równań rozwiązujemy zwykle metodą Newtona (podrozdz. 3.2).

Gdy przedział $[a, b]$ dzielimy na k podprzedziałów, daje to tyleż zagadnień początkowych i ich rozwiązań określonych w tych podprzedziałach. Warunki ciągłości w punktach podziału dają $2k-2$ równań. Taka sama jest liczba nieznanych parametrów, występujących w warunkach początkowych. I w tym przypadku układ równań można rozwiązywać metodą Newtona.

Większość dostępnych programów, których możemy użyć rozwiązując zagadnienia brzegowe, odnosi się do układu równań różniczkowych zwyczajnych rzędu pierwszego

$$X' = F(t, X),$$

gdzie $X := (x_1, x_2, \ldots, x_n)$ i $F := (f_1, f_2, \ldots, f_n)$. W wielu takich programach warunki brzegowe mogą być bardzo ogólne, np. wyrażone równością

$$G(X(a), X(b)) = 0,$$

gdzie $G := (g_1, g_2, \ldots, g_n)$. W pewnych programach trzeba podać macierz Jacobiego J stopnia n, o elementach $(J)_{ij} := \partial f_i / \partial x_j$.

Przykład 8.8.4. Podać F, G i J dla zagadnienia brzegowego

$$x'' = \cos x' + tx, \quad x(3) + x'(5) = 7, \quad [x'(3)]^2 x(5) = 10.$$

Rozwiązanie. Jeśli $x_1 := x$, $x_2 := x'$, to zagadnienie wyraża się tak:

$$\begin{aligned} x_1' &= x_2, & x_1(3) + x_2(5) - 7 &= 0, \\ x_2' &= \cos x_2 + tx_1, & [x_2(3)]^2 x_1(5) - 10 &= 0. \end{aligned}$$

Można stąd odczytać funkcje F i G. Jakobian jest równy

$$J(t, X) = \begin{bmatrix} 0 & 1 \\ t & -\sin x_2 \end{bmatrix}.$$

■

Zadania 8.8

1. Znaleźć funkcję φ dla następujących zagadnień brzegowych:
   (a) $x'' = x$, $x(0) = 0$, $x(1) = 17$
   (b) $x'' = -x$, $x(0) = 1$, $x(\pi/2) = 3$
   (c) $x'' = -2t(x')^2$, $x(0) = 1$, $x(1) = \pi/4 + 1$
   (d) $x'' = -(x')^2/x$, $x(1) = 3$, $x(2) = 5$

2. Rozwiązać zagadnienie brzegowe $x'' = -9x$, $x(0) = 1$, $x(\pi/6) = 5$, znajdując najpierw rozwiązanie x_z zagadnienia początkowego $x'' = -9x$, $x(0) = 1$, $x'(0) = z$, a następnie poprawiając z tak, aby było $x_z(\pi/6) = 5$. Jak na wynik wpływa zmiana drugiego warunku brzegowego na $x(\pi/3) = 5$?

3. Znaleźć $x_z(1)$, gdzie x_z jest rozwiązaniem zagadnienia początkowego $x'' = x$, $x(0) = 0$, $x'(0) = z$.

4. Jeśli x_z jest rozwiązaniem zagadnienia początkowego $x'' = -x$, $x(0) = 5$, $x'(0) = z$ i $\varphi(z) := x_z(\pi/2) - 3$, to jak się wyraża $\varphi'(z)$?

5. Udowodnić, że jeśli zagadnienie brzegowe liniowe (8.8.4) ma rozwiązanie, to można je znaleźć, rozwiązując zagadnienia początkowe

$$x_1'' = u + vx_1 + wx_1', \qquad x_1(a) = \alpha, \qquad x_1'(a) = 0,$$
$$x_2'' = vx_2 + wx_2', \qquad x_2(a) = 0, \qquad x_2'(a) = 1$$

i że tym rozwiązaniem jest $x_1 + [\beta - x_1(b)]/x_2(b)\, x_2$.

6. Stosując do zagadnienia brzegowego $x'' = 37t^2x' + 95$, $x(6) = 1$, $x(12) = 2$ metodę strzału opartą na metodzie siecznych, otrzymujemy dwie pary liczb $(z_i, x_{z_i}(b))$ $(i = 1, 2)$, np. $(4, 5)$ i $(2, 9)$. Jakie zagadnienie początkowe będzie rozwiązywane w następnym kroku?

7. Wyjaśnić, jak metodę strzału można zastosować do zagadnienia brzegowego

$$x'' = u + vx + wx', \quad c_{11}x(a) + c_{12}x'(a) = \alpha, \quad c_{21}x(b) + c_{22}x'(b) = \beta,$$

gdzie stałe α, β, c_{ij} są dane. **Wskazówka**: Niech x_1 spełnia to równanie różniczkowe z warunkami początkowymi takimi, że $c_{11}x_1(a) + c_{12}x_1'(a) = \alpha$, a x_2 – to samo równanie z warunkami $x_2(a) = -c_{12}$ i $x_2'(a) = c_{11}$. Rozważyć kombinację $x_1 + \lambda x_2$.

8. Znaleźć pierwsze równanie wariacyjne dla równania różniczkowego

(a) $x'' = a(t) + b(t)x + c(t)x'$

(b) $x'' = \cos(tx) + \sin(t^2x')$ (czy to równanie wariacyjne można rozwiązywać oddzielnie, czy tylko wraz z danym równaniem różniczkowym?).

9. Wykazać, że jeśli funkcje x_1 i x_2 spełniają to samo równanie różniczkowe liniowe rzędu drugiego, a także warunki początkowe $x_1(a) = x_2(a) = \alpha$, $x_1'(a) = 0$ i $x_2'(a) = 1$, to różnica $x_2 - x_1$ jest rozwiązaniem pierwszego równania wariacyjnego.

10. Wykazać, że jeśli zagadnienie brzegowe jest liniowe i jeśli do równania $\varphi = 0$ stosujemy metodę Newtona, przy czym φ' znajdujemy z pierwszego równania wariacyjnego, to wynik jest zgodny z (8.8.6) i (8.8.7).

ZADANIA KOMPUTEROWE 8.8

K1. Zaprogramować opisaną w tekście metodę rozwiązywania zagadnienia brzegowego liniowego (8.8.4). Odpowiednie zagadnienia początkowe rozwiązywać np. metodą Rungego-Kutty rzędu czwartego. Program sprawdzić dla zagadnień:

(a) $x'' = e^t + x\cos t - (t+1)x'$, $x(0) = 1$, $x(1) = 3$

(b) $x'' = e^{t-3} + (t^2+2)x + (\sin t)x'$, $x(2.6) = 7$, $x(5.1) = -3$

8.9. Zagadnienia brzegowe: różnice skończone

Inne podejście do zagadnień brzegowych polega na zastąpieniu pochodnych wyrażeniami przybliżonymi, które zastosowano już w podrozdz. 7.1. Szczególnie użyteczne będą teraz dwa wzory (w których podano też błędy takich wyrażeń):

$$x'(t) = \frac{1}{2h}[x(t+h) - x(t-h)] - \frac{1}{6}h^2 x'''(\xi), \tag{8.9.1}$$

$$x''(t) = \frac{1}{h^2}[x(t+h) - 2x(t) + x(t-h)] - \frac{1}{12}h^2 x^{(4)}(\tau). \tag{8.9.2}$$

Jak przedtem, zajmujemy się zagadnieniem brzegowym

$$x'' = f(t, x, x'), \quad x(a) = \alpha, \quad x(b) = \beta. \tag{8.9.3}$$

Podzielmy przedział $[a, b]$ punktami $a = t_0 < t_1 < \ldots < t_n < t_{n+1} = b$. W praktyce przyjmuje się, że są one równoodległe:

$$t_i = a + ih \quad (0 \leqslant i \leqslant n+1), \quad \text{gdzie } h := (b-a)/(n+1)$$

(odrzucając to założenie, musielibyśmy zastąpić wzory (8.9.1) i (8.9.2) bardziej skomplikowanymi). Niech y_i oznacza przybliżoną wartość $x(t_i)$. Wersja dyskretna zadania oparta na wspomnianych wzorach jest następująca:

$$\begin{aligned} &y_0 = \alpha, \\ &h^{-2}(y_{i-1} - 2y_i + y_{i+1}) = f\left(t_i, y_i, (2h)^{-1}(y_{i+1} - y_{i-1})\right) \quad (1 \leqslant i \leqslant n), \\ &y_{n+1} = \beta. \end{aligned}$$

Zagadnienie liniowe

Powyższy układ n równań z tyluż niewiadomymi $y_1, y_2, \ldots, y_n$ jest nieliniowy, jeśli taka jest zależność funkcji f od drugiego lub trzeciego argumentu. Wtedy rozwiązywanie układu jest kłopotliwe. Znacznie łatwiej jest go rozwiązać w przypadku liniowym, rozważanym także w poprzednim podrozdziale. Jeśli mianowicie

$$f(t, x, x') = u(t) + v(t)x + w(t)x',$$

to mamy układ

$$\begin{aligned} &y_0 = \alpha, \\ &(1 + \tfrac{1}{2}hw_i)y_{i-1} - (2 + h^2 v_i)y_i + (1 - \tfrac{1}{2}hw_i)y_{i+1} = h^2 u_i \quad (1 \leqslant i \leqslant n), \\ &y_{n+1} = \beta, \end{aligned} \tag{8.9.4}$$

gdzie $u_i := u(t_i)$ itd. Wprowadzamy oznaczenia

$$a_i := 1 + \tfrac{1}{2}hw_{i+1}, \quad d_i := -(2 + h^2 v_i), \qquad c_i := 1 - \tfrac{1}{2}hw_i, \quad b_i := h^2 u_i$$

i wyrażamy ten układ w postaci

$$\begin{bmatrix} d_1 & c_1 & & & \\ a_1 & d_2 & c_2 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ & & a_{n-2} & d_{n-1} & c_{n-1} \\ & & & a_{n-1} & d_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} b_1 - a_0\alpha \\ b_2 \\ \cdots \\ b_{n-1} \\ b_n - c_n\beta \end{bmatrix}. \tag{8.9.5}$$

Macierz układu jest trójprzekątniowa, co pozwala rozwiązywać go za pomocą specjalnej wersji eliminacji Gaussa. Zauważmy też, że jeśli $v(t) > 0$, a h jest na tyle małe, że $|hw_i| \leqslant 2$, to ta macierz jest przekątniowo dominująca. Istotnie, wtedy $1 \pm \frac{1}{2}hw_i \geqslant 0$ i

$$|2 + h^2 v_i| > |1 + \tfrac{1}{2}hw_i| + |1 - \tfrac{1}{2}hw_i| = 2.$$

W dalszym ciągu skorzystamy z równości

$$|d_i| - |c_i| - |a_{i-1}| = h^2 v_i \tag{8.9.6}$$

poprawnej przy powyższych założeniach o v i h.

Zbieżność

Wykażemy, że dla $h \to 0$ rozwiązanie zadania dyskretnego jest zbieżne do rozwiązania zagadnienia brzegowego

$$x'' = u + vx + wx', \quad x(a) = \alpha, \quad x(b) = \beta. \tag{8.9.7}$$

Istnienie i jednoznaczność rozwiązania tego zagadnienia wynika z następującego twierdzenia (Keller [1968, s. 9]):

Twierdzenie 8.9.1. *Zagadnienie brzegowe*

$$x'' = f(t, x, x'), \quad c_{11}x(a) + c_{12}x'(a) = c_{13}, \quad c_{21}x(b) + c_{22}x'(b) = c_{23}$$

ma jednoznaczne rozwiązanie w przedziale $[a, b]$, *jeśli:*

1. *Funkcja* f *i jej pierwsze pochodne cząstkowe* f_t, f_x, $f_{x'}$ *są ciągłe w obszarze* $D := [a, b] \times \mathbb{R}^2$.
2. $f_x > 0$, $f_x \leqslant M$ *i* $|f_{x'}| \leqslant M$ *w* D.
3. $|c_{11}| + |c_{12}| > 0$, $|c_{21}| + |c_{22}| > 0$, $|c_{11}| + |c_{21}| > 0$ *i* $c_{11}c_{12} \leqslant 0 \leqslant c_{21}c_{22}$.

Wracamy do zagadnienia (8.9.7). Niech $x(t)$ będzie jego rozwiązaniem, a wektor $(y_1, y_2, \dots, y_n)$ – rozwiązaniem układu (8.9.5), oczywiście zależnym od h. Udowodnimy, że $|x_i - y_i|$ (gdzie $x_i := x(t_i)$) dąży do 0, gdy $h \to 0$.

Z (8.9.1) i (8.9.2) wynika, że

$$h^{-2}(x_{i-1} - 2x_i + x_{i+1}) - \tfrac{1}{12}h^2 x^{(4)}(\tau_i) = \\ = u_i + v_i x_i + w_i\left[(2h)^{-1}(x_{i+1} - x_{i-1}) - \tfrac{1}{6}h^2 x'''(\xi_i)\right] \quad (1 \leqslant i \leqslant n).$$

Natomiast rozwiązanie zadania dyskretnego spełnia równania wynikające z poprzednich przez odrzucenie składników z x''' i $x^{(4)}$:

$$h^{-2}(y_{i-1} - 2y_i + y_{i+1}) = u_i + v_i y_i + w_i(2h)^{-1}(y_{i+1} - y_{i-1}) \quad (1 \leqslant i \leqslant n).$$

Odejmujemy stronami odpowiednie równania z obu układów, przyjmując oznaczenie $e_i := x_i - y_i$:

$$h^{-2}(e_{i-1} - 2e_i + e_{i+1}) = v_i e_i + w_i(2h)^{-1}(e_{i+1} - e_{i-1}) + h^2 g_i, \quad (8.9.8)$$

gdzie

$$g_i := \tfrac{1}{12}x^{(4)}(\tau_i) - \tfrac{1}{6}x'''(\xi_i). \quad (8.9.9)$$

Po oczywistych przekształceniach otrzymujemy układ podobny do (8.9.4):

$$(1 + \tfrac{1}{2}hw_i)e_{i-1} - (2 + h^2 v_i)e_i + (1 - \tfrac{1}{2}hw_i)e_{i+1} = h^4 g_i \quad (1 \leqslant i \leqslant n),$$

czyli układ

$$a_{i-1}e_{i-1} + d_i e_i + c_i e_{i+1} = h^4 g_i \quad (1 \leqslant i \leqslant n).$$

Stąd $|d_i|\,|e_i| \leqslant h^4|g_i| + |c_i|\,|e_{i+1}| + |a_{i-1}|\,|e_{i-1}|$. Niech $\lambda := \|e\|_\infty$ i niech i będzie takie, że $|e_i| = \lambda$. Dla tego i zachodzi nierówność

$$|d_i|\lambda \leqslant h^4\|g\|_\infty + |c_i|\lambda + |a_{i-1}|\lambda,$$

czyli $\lambda(|d_i| - |c_i| - |a_{i-1}|) \leqslant h^4\|g\|_\infty$. Stąd i z (8.9.6) wynika, że

$$h^2 v_i \lambda \leqslant h^4\|g\|_\infty, \quad \|e\|_\infty \leqslant h^2\|g\|_\infty / \inf v(t).$$

Wobec (8.9.9) jest $\|g\|_\infty \leqslant \|x^{(4)}\|_\infty/12 + \|x'''\|_\infty/6$. Ta suma, jak i $\inf v(t)|$ nie zależy od h, więc $\|e\|_\infty = \mathcal{O}(h^2)$ dla $h \to 0$.

Zadania 8.9

1. Rozwiązać zagadnienie brzegowe $x'' + 2x' + 10t = 0$, $x(0) = 1$, $x(1) = 2$, stosując metodę opisaną w tekście dla $h = 0.5$.
2. Jak należy zmodyfikować układ (8.9.5), gdy warunki brzegowe są takie, jak w tw. 8.9.1?

3. Stosując tw. 8.9.1, udowodnić, że zagadnienie brzegowe (8.9.7) ma jednoznaczne rozwiązanie, gdy funkcje u, v, w są ciągłe i $v > 0$.

ZADANIA KOMPUTEROWE 8.9

K1. Napisać program rozwiązujący zagadnienie brzegowe (8.9.7) metodą podaną w tekście dla danych u, v, w, a, α, b, β, n. Sprawdzić dokładność wyników na następujących przykładach:

(**a**) $x'' = -x$, $x(0) = 3$, $x(\pi/2) = 7$ (rozwiązanie: $7\sin t + 3\cos t$).

(**b**) $x'' = 2e^t - x$, $x(0) = 2$, $x(1) = e + \cos 1$ (rozwiązanie: $e^t + \cos t$).

8.10. Zagadnienia brzegowe: kollokacja

Kollokacja jest strategią możliwą do zastosowania w wielu zagadnieniach matematyki stosowanej. Zaczynamy od jej ogólnego opisu. Przypuśćmy, że dla danego operatora liniowego L (może to być operator całkowy lub różniczkowy) chcemy rozwiązać równanie

$$Lu = w, \tag{8.10.1}$$

w którym w jest dane, a u szukane. Liczne metody stosowane w tym przypadku zaczynają się od wyboru układu $\{v_1, v_2, \ldots, v_n\}$ wektorów bazowych. Próbujemy znaleźć rozwiązanie postaci

$$u = c_1 v_1 + c_2 v_2 + \ldots + c_n v_n.$$

Dzięki liniowości operatora L stąd i z (8.10.1) wynika, że

$$\sum_{j=1}^{n} c_j L v_j = w. \tag{8.10.2}$$

Na ogół nie potrafimy stąd wyznaczyć współczynników c_j. Można to jednak zrobić, gdy zaakceptujemy rozwiązanie przybliżone. W metodzie *kollokacji* u, v_j, w są funkcjami określonymi w pewnym wspólnym obszarze, a polega ona na żądaniu, aby obie strony (8.10.2) były zgodne w n ustalonych punktach t_i:

$$\sum_{j=1}^{n} c_j (Lv_j)(t_i) = w(t_i) \quad (1 \leqslant i \leqslant n). \tag{8.10.3}$$

Jest to układ n równań liniowych względem tyluż niewiadomych c_j. Można go rozwiązać, jeśli tylko macierz o elementach $(Lv_j)(t_i)$ jest nieosobliwa.

Zagadnienie Sturma-Liouville'a

Pokażemy teraz, jak metoda kollokacji działa w rozważanym już wcześniej zagadnieniu brzegowym Sturma-Liouville'a:

$$u'' + pu' + qu = w, \quad u(0) = 0, \quad u(1) = 0.$$

Dane funkcje p, q, w są ciągłe w przedziale $[0,1]$. Aby zastosować metodę kollokacji, określamy operator L wzorem

$$Lu := u'' + pu' + qu.$$

Rozwiązania u szukamy w przestrzeni wektorowej

$$V := \{u \in C^2[0,1]\colon\ u(0) = u(1) = 1\}.$$

Jeśli funkcje bazowe $v_1, v_2, \ldots, v_n$ do niej należą, to warunki brzegowe jednorodne są spełnione automatycznie. Można np. wybrać układ funkcji (z dwoma wskaźnikami)

$$v_{jk}(t) := t^j(1-t)^k \quad (j,\ k \geqslant 1).$$

Ponieważ

$$v'_{jk} = jv_{j-1,k} - kv_{j,k-1},$$
$$v''_{jk} = j(j-1)v_{j-2,k} - 2jkv_{j-1,k-1} + k(k-1)v_{j,k-2}$$

(powyższą definicję rozszerzamy na przypadek, gdy $j = 0$ lub $k = 0$), więc obliczanie wartości funkcji Lv_{jk} jest łatwe. Wybrawszy sensownie n funkcji v_{jk} i tyleż punktów t_i z przedziału $[0,1]$, możemy rozwiązywać układ (8.10.3), co da przybliżone rozwiązanie zagadnienia brzegowego.

Funkcje B-sklejane sześcienne

Lepszym zapewne układem funkcji bazowych jest układ funkcji B-sklejanych (podrozdz. 6.5). Możemy tu rozważyć nieco ogólniejsze zagadnienie brzegowe:

$$u'' + pu' + qu = w, \quad u(a) = \alpha, \quad u(b) = \beta. \tag{8.10.4}$$

Ponieważ funkcje bazowe powinny mieć drugą pochodną ciągłą, wybieramy funkcje B_i^k dla $k \geqslant 3$, a ściślej – dla prostoty – funkcje B_i^3. Przyjmiemy też, że węzły używane w określeniu tych funkcji, a zarazem punkty stosowane w metodzie kollokacji, są równoodległe. Niech $n > 3$ będzie liczbą funkcji

bazowych (i współczynników do obliczenia). Wobec tego oprócz warunków brzegowych

$$\sum_{j=1}^{n} c_j v_j(a) = \alpha, \quad \sum_{j=1}^{n} c_j v_j(b) = \beta \tag{8.10.5}$$

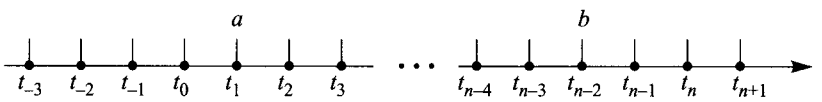
powinniśmy mieć $n-2$ warunki wynikające z metody kollokacji:

$$\sum_{j=1}^{n} c_j (Lv_j)(t_i) = w(t_i) \quad (1 \leqslant i \leqslant n-2). \tag{8.10.6}$$

To już narzuca definicję punktów t_i:

$$t_i = a + (i-1)h \quad (i = 0, \pm 1, \pm 2, \ldots), \quad \text{gdzie} \quad h := \frac{b-a}{n-3}. \tag{8.10.7}$$

W przedziale $[a, b]$ leżą punkty $t_1 = a, t_2, \ldots, t_{n-2} = b$. Występują one w warunkach (8.10.6). Dodatkowe punkty t_i spoza tego przedziału są potrzebne do określenia funkcji B_j^3. Rozkład węzłów pokazuje rys. 8.3.



RYS. 8.3. Rozkład węzłów

Potrzebne nam są te funkcje B_i^3, które nie znikają tożsamościowo w $[a, b]$. Tak jest dla $-2 \leqslant i \leqslant n-3$. Niech więc będzie $v_j := B_{j-3}^3$ dla $1 \leqslant j \leqslant n$.

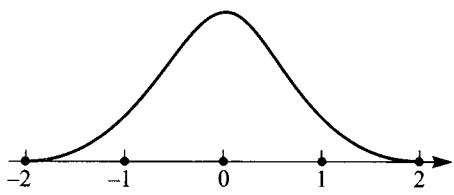
Dzięki równomiernemu rozmieszczeniu węzłów wszystkie funkcje v_j powstają z tej samej funkcji B-sklejanej określonej wzorem

$$B(t) := \begin{cases} \frac{1}{6}(t+2)^3 & (t \in [-2, -1]) \\ \frac{1}{6}(-3t^3 - 6t^2 + 4) & (t \in [-1, 0]) \\ B(-t) & (t \in [0, 2]) \\ 0 & (t \notin [-2, 2]). \end{cases}$$

Łatwo sprawdzić, że jest to rzeczywiście funkcja sklejana sześcienna. Jej wykres pokazano na rys. 8.4. Funkcje bazowe v_j wyrażają się przez nią wzorem

$$v_j(t) = B\left(\frac{t-a}{h} - j + 2\right) \quad (1 \leqslant j \leqslant n). \tag{8.10.8}$$

Aby obliczyć $(Lv_j)(t_i)$, musimy znać pierwszą i drugą pochodną funkcji v_j. Można je wyrazić przez odpowiednie pochodne funkcji B. Macierz o elementach $(Lv_j)(t_i)$ jest taśmowa, gdyż każda z funkcji v_j ma krótki nośnik. Trzeba to wykorzystać w programie.



RYS. 8.4. Funkcja sklejana sześcienna B

ZADANIA KOMPUTEROWE 8.10

K1. Napisać program rozwiązujący zagadnienie brzegowe (8.10.4) metodą kollokacji. Użytkownik programu powinien określić: (i) funkcje p, q, w, (ii) liczby rzeczywiste a, α, b, β, gdzie $a < b$, (iii) liczbę n funkcji bazowych, (iv) funkcje bazowe v_i oraz ich pierwsze i drugie pochodne. Program ma wygenerować punkty (8.10.7) i współczynniki układu liniowego (8.10.5), (8.10.6), a następnie wywołać jakąś procedurę rozwiązywania takiego układu. Program ma również sprawdzić dokładność przybliżonego rozwiązania przez obliczenie reszty $Lu - w$ w $2n - 5$ punktach równomiernie rozmieszczonych w $[a, b]$ (co drugi z nich jest równy t_i; tam reszta powinna znikać).

K2. (cd.). Sprawdzić program z poprzedniego zadania, rozwiązując zagadnienie brzegowe

$$u'' + (\sin t)u' + (t^2 + 2)u = e^{t-3}, \quad u(2.6) = 7, \quad u(5.1) = -3.$$

Funkcjami bazowymi v_j mają być funkcje B-sklejane (8.10.8).

8.11. Układy równań różniczkowych liniowych

Tematem tego podrozdziału są układy autonomiczne równań różniczkowych liniowych o stałych współczynnikach:

$$x_i' = \sum_{j=1}^{n} a_{ij} x_j \quad (1 \leqslant i \leqslant n).$$

W symbolice macierzowej mamy tu układ

$$X' = AX, \tag{8.11.1}$$

gdzie $A = (a_{ij})$, $X := (x_1, x_2, \ldots, x_n)$ i $X' := (x_1', x_2', \ldots, x_n')$.

Wartości i wektory własne

Próbujemy znaleźć rozwiązanie tego układu, mające postać $X(t) = e^{\lambda t} V$, gdzie V jest stałym wektorem. Podstawienie tego próbnego rozwiązania do (8.11.1) daje równość $\lambda e^{\lambda t} V = e^{\lambda t} AV$, czyli $AV = \lambda V$. Stąd wynika

Twierdzenie 8.11.1. *Jeśli λ jest wartością własną macierzy A, a V odpowiadającym jej wektorem własnym, to rozwiązaniem układu* (8.11.1) *jest $X(t) = e^{\lambda t}V$.*

Z twierdzenia wynika, że chcąc rozwiązać układ (8.11.1), musimy znać wartości i wektory własne macierzy A.

Poniższe twierdzenie dotyczy najprostszego przypadku.

Twierdzenie 8.11.2. *Jeśli macierz A stopnia n ma układ n niezależnych liniowo wektorów własnych V_i takich, że $AV_i = \lambda_i V_i$, to przestrzeń rozwiązań układu $X' = AX$ ma bazę $X_i = \exp(\lambda_i t)V_i$ $(1 \leqslant i \leqslant n)$.*

Dowód. Układ $\{X_1, X_2, \ldots, X_n\}$ jest niezależny liniowo, gdyż z równości $\sum_{i=1}^n c_i X_i = 0$ wynika, że $\sum_{i=1}^n c_i \exp(\lambda_i t)V_i = 0$, a wtedy z założenia o V_i wynika, że wszystkie c_i znikają.

Niech X będzie dowolnym rozwiązaniem układu $X' = AX$. Wektor wartości początkowych $X(0)$ (należący do $\mathbb{R}^n$ lub $\mathbb{C}^n$) jest kombinacją liniową wektorów $V_1, V_2, \ldots, V_n$:

$$X(0) = \sum_{i=1}^n c_i V_i.$$

Niech będzie $Y := \sum_{i=1}^n c_i X_i$. Wtedy

$$Y' = \sum_{i=1}^n c_i X_i' = \sum_{i=1}^n c_i \lambda_i \exp(\lambda_i t)V_i =$$
$$= \sum_{i=1}^n c_i \exp(\lambda_i t)AV_i = A\Big(\sum_{i=1}^n c_i X_i\Big) = AY.$$

Wobec tego Y jest rozwiązaniem danego układu równań różniczkowych. Jest też $X(0) = Y(0)$ (dlaczego?). Z twierdzenia o jednoznaczności zagadnienia początkowego wynika, że $X = Y = \sum_{i=1}^n c_i X_i$. ■

Jeśli macierz A ma własność podaną w tw. 8.11.2, to macierz P o kolumnach $V_1, V_2, \ldots, V_n$ jest nieosobliwa. Równania $AV_i = \lambda_i V_i$ dają łącznie równość macierzową

$$AP = P\Lambda,$$

gdzie $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Jeśli $X = PY$, to

$$Y' = P^{-1}X' = P^{-1}AX = P^{-1}APY = \Lambda Y. \tag{8.11.2}$$

Dlatego układ dla Y jest znacznie prostszy: każde równanie można rozwiązać niezależnie od pozostałych.

Przykład 8.11.3. Rozwiązać zagadnienie początkowe

$$X' = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} X, \quad X(0) = \begin{bmatrix} 5 \\ 7 \\ 6 \end{bmatrix}.$$

Rozwiązanie. Wartości własne powyższej macierzy A są zerami jej wielomianu charakterystycznego, równego

$$\det(A - \lambda I) = \begin{vmatrix} 1-\lambda & 0 & 1 \\ 0 & -\lambda & 0 \\ 0 & 0 & -1-\lambda \end{vmatrix} = (1-\lambda)(-\lambda)(-1-\lambda),$$

czyli $\lambda_1 = 1$, $\lambda_2 = 0$, $\lambda_3 = -1$. Dla każdej z tych wartości znajdujemy wektor własny, rozwiązując równanie $AV_i = \lambda_i V_i$. Daje to kolumny macierzy P:

$$P = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix}.$$

Znajdujemy macierz odwrotną:

$$P^{-1} = \begin{bmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{2} \end{bmatrix}.$$

Układ równań różniczkowych dla Y, czyli $Y' = \Lambda Y$, jest tu szczególnie prosty: $y_1' = y_1$, $y_2' = 0$, $y_3' = -y_3$. Wiążą się z nim warunki początkowe $Y(0) = P^{-1}X(0) = (8, 7, -3)$, więc $y_1 = 8e^t$, $y_2 = 7$, $y_3 = -3e^{-t}$. Ponieważ $X = PY$, więc

$$x_1 = 8e^t - 3e^{-t}, \quad x_2 = 7, \quad x_3 = 6e^{-t}. \qquad \blacksquare$$

Funkcja wykładnicza z argumentem macierzowym

Istnieje elegancka formalna metoda rozwiązywania układu $X' = AX$, która nie odwołuje się do wartości własnych macierzy A, chyba że jego rozwiązanie chcemy znaleźć numerycznie. Zaczynamy od definicji funkcji wykładniczej z argumentem macierzowym. Przyjmujemy mianowicie, że

$$e^A := I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \ldots \tag{8.11.3}$$

Definicja jest analogiczna do rozwinięcia w szereg potęgowy zwykłej funkcji wykładniczej:

$$e^z = 1 + z + \frac{1}{2!}z^2 + \frac{1}{3!}z^3 + \dots$$

Aby udowodnić, że szereg w (8.11.3) jest zbieżny, wybieramy dowolną normę w $\mathbb{C}^n$ i indukowaną przez nią normę macierzy n-tego stopnia. Szacujemy normę reszty tego szeregu:

$$\Big\| \sum_{k=m}^{\infty} \frac{1}{k!} A^k \Big\| \leqslant \sum_{k=m}^{\infty} \frac{1}{k!} \|A^k\| \leqslant \sum_{k=m}^{\infty} \frac{1}{k!} \|A\|^k. \tag{8.11.4}$$

Ostatnie wyrażenie jest resztą zwykłego szeregu wykładniczego dla $z = \|A\|$, a ten szereg jest zbieżny dla dowolnego z. Wobec tego ciąg reszt szeregu definiującego e^A dąży do 0, gdy $m \to \infty$. (W tym rozumowaniu przyjmuje się, że jest już znana zupełność przestrzeni macierzy dla danej normy).

Jeśli $t \in \mathbb{R}$, to $tA = At$ i z przyjętej definicji wynika, że

$$e^{At} = \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k.$$

Różniczkowanie względem t prowadzi do wniosku, że

$$\frac{d}{dt} e^{At} = Ae^{At}.$$

Twierdzenie 8.11.4. *Rozwiązaniem zagadnienia początkowego*

$$X' = AX, \quad X(0) = X_0 \tag{8.11.5}$$

jest wektor $X(t) = e^{At}X_0$.

Dowód. Z równości $X = e^{At}X_0$ wynika, że $X' = Ae^{At}X_0 = AX$. Jest też $X(0) = e^{A0}X_0 = X_0$. ■

Macierze przekątniowe i podobne do nich

Aby zastosować w praktyce tw. 8.11.4, musimy wiedzieć, jak się w sensowny sposób oblicza e^A. Zaczynamy od przypadku, gdy macierz A jest przekątniowa. Jeśli $A = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$, to $A^k = \operatorname{diag}(\lambda_1^k, \lambda_2^k, \dots, \lambda_n^k)$. Dzięki temu dla takiej macierzy A

$$e^{At} = \sum_{k=0}^{\infty} \frac{t^k}{k!} \operatorname{diag}(\lambda_1^k, \lambda_2^k, \dots, \lambda_n^k) = \operatorname{diag}(e^{\lambda_1 t}, e^{\lambda_2 t}, \dots, e^{\lambda_n t})$$

i składowe rozwiązania zagadnienia początkowego (8.11.5) są równe

$$x_i(t) = e^{\lambda_i t} x_i(0) \quad (1 \leqslant i \leqslant n).$$

To rozumowanie przenosi się od razu na przypadek, gdy A jest macierzą *prostej struktury* (zob. podrozdz. 5.1). Jest ona podobna do macierzy przekątniowej: $P^{-1}AP = \Lambda$, gdzie P jest macierzą nieosobliwą, a Λ macierzą przekątniową. Wtedy podstawienie $X = PY$ zmienia układ $X' = AX$ na $Y' = \Lambda Y$ (zob. (8.11.2)). Z warunku początkowego dla X wynika, że $Y_0 = P^{-1}X_0$, więc

$$X = PY = P(e^{\Lambda t}P^{-1}X_0) = P \operatorname{diag}(e^{\lambda_1 t}, e^{\lambda_2 t}, \ldots, e^{\lambda_n t})P^{-1}X_0.$$

Bloki jordanowskie

Pominęliśmy aż dotąd analizę przypadku, gdy A nie jest podobna do macierzy przekątniowej. W takim przypadku układ wektorów własnych macierzy A nie jest bazą przestrzeni $\mathbb{C}^n$. Taka sytuacja dotyczy np. macierzy

$$J(\lambda, 2) := \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}, \quad J(\lambda, 3) := \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}.$$

Rozważmy szczegółowo macierz $J(\lambda, 3)$. Jest ona trójkątna górna, więc jej wartościami własnymi są elementy na przekątnej, wszystkie równe λ. Wektor własny X spełnia równanie $J(\lambda, 3)X = \lambda X$, czyli układ

$$\begin{aligned} \lambda x_1 + x_2 &= \lambda x_1, \\ \lambda x_2 + x_3 &= \lambda x_2, \\ \lambda x_3 &= \lambda x_3. \end{aligned}$$

Stąd oczywiście wynika, że $x_2 = x_3 = 0$, czyli $X = (\beta, 0, 0)$. Zbiór tych wektorów własnych tworzy w $\mathbb{C}^3$ podprzestrzeń zaledwie jednowymiarową. Tak jest, ogólniej, dla każdej macierzy $J(\lambda, k)$ $(k \geqslant 2)$ zbudowanej podobnie jak $J(\lambda, 2)$ i $J(\lambda, 3)$. Są to tzw. *bloki jordanowskie*. Najprostszy taki blok, $J(\lambda, 1)$, ma jedyny element równy λ.

TWIERDZENIE 8.11.5. *Każda macierz kwadratowa jest podobna do macierzy blokowej, która na przekątnej ma bloki jordanowskie, a poza nią bloki zerowe.*

Macierz blokową określoną w tym twierdzeniu nazywamy *postacią kanoniczną Jordana* danej macierzy. Oto kilka macierzy mających taką postać:

$$\begin{bmatrix} 7 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad \begin{bmatrix} 5 & 1 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad \begin{bmatrix} 5 & 0 & 0 \\ 0 & 5 & 1 \\ 0 & 0 & 5 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 5 & 1 & 0 \\ 0 & 5 & 1 \\ 0 & 0 & 5 \end{bmatrix}.$$

W pierwszej z nich mamy trzy bloki jordanowskie, w drugiej, trzeciej i czwartej – dwa bloki, a piąta macierz jest cała takim blokiem.

Każdy blok jordanowski można wyrazić w postaci

$$J(\lambda, k) = \lambda I_k + H_k,$$

gdzie I_k jest macierzą jednostkową stopnia k, a macierz H_k tegoż stopnia ma tuż nad przekątną jedynki, a gdzie indziej zera. Iloczynem tej macierzy przez wektor $(\xi_1, \xi_2, \ldots, \xi_k)$ jest wektor $(\xi_2, \xi_3, \ldots, \xi_k, 0)$. Jest więc oczywiste, że $H_k^k V = 0$, gdyż każde mnożenie przez H_k usuwa jedną składową wektora. Jeśli tak, to $H_k^k = 0$. Wykorzystamy to, obliczając e^{At}, gdzie A jest blokiem jordanowskim. Mamy mianowicie

$$e^{(\lambda I_k + H_k)t} = e^{t\lambda I_k} e^{H_k t} = \sum_{j=0}^{\infty} \frac{(\lambda t I_k)^j}{j!} \sum_{j=0}^{\infty} \frac{(tH_k)^j}{j!} =$$

$$= e^{\lambda t} \left[I_k + tH_k + \frac{t^2}{2!} H_k^2 + \ldots + \frac{t^{k-1}}{(k-1)!} H_k^{k-1} \right],$$

gdyż drugi szereg redukuje się do sumy skończonej. Zauważmy też, że zastosowano tu równość $e^{A+B} = e^A e^B$ prawdziwą tylko przy pewnych założeniach o A i B (zob. zad. 14).

Możemy teraz rozwiązać układ równań różniczkowych $X' = AX$ przy założeniu, że A jest blokiem jordanowskim, tj. $A = \lambda I_k + H_k$. Ponieważ na mocy tw. 8.11.4 $X(t) = e^{At} X(0)$, więc z powyższej tożsamości wynika, że

$$X(t) = e^{\lambda t} \left[I_k + tH_k + \frac{t^2}{2!} H_k^2 + \ldots + \frac{t^{k-1}}{(k-1)!} H_k^{k-1} \right] X(0). \qquad (8.11.6)$$

Przykład 8.11.6. Rozwiązać zagadnienie początkowe $X' = AX$ dla

$$A = \begin{bmatrix} 3 & 1 & 0 & 0 \\ 0 & 3 & 1 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 3 \end{bmatrix}, \quad X(0) = \begin{bmatrix} 7 \\ 5 \\ 3 \\ 9 \end{bmatrix}.$$

Rozwiązanie. Ponieważ $A = 3I_4 + H_4$, więc z (8.11.6) wynika, że

$$X(t) = e^{3t}(I_4 + tH_4 + \tfrac{1}{2}t^2H_4^2 + \tfrac{1}{6}t^3H_4^3)X(0) =$$

$$= e^{3t}\begin{bmatrix}7\\5\\3\\9\end{bmatrix} + te^{3t}\begin{bmatrix}5\\3\\9\\0\end{bmatrix} + \frac{1}{2}t^2e^{3t}\begin{bmatrix}3\\9\\0\\0\end{bmatrix} + \frac{1}{6}t^3e^{3t}\begin{bmatrix}9\\0\\0\\0\end{bmatrix} =$$

$$= \begin{bmatrix}7 + 5t + 1.5t^2 + 1.5t^3\\5 + 3t + 4.5t^2\\3 + 9t\\9\end{bmatrix} e^{3t}.$$

■

Rozwiązanie ogólne

Można teraz opisać już całkowicie ogólnie sposób rozwiązania układu równań różniczkowych $X' = AX$. Niech będzie

$$P^{-1}AP = C,$$

gdzie C jest postacią kanoniczną Jordana macierzy A. Wiemy już (zob. (8.11.2)), że przekształcenie $X = PY$ prowadzi do układu równań $Y' = CY$. Ten układ można podzielić na niezależne podukłady związane z poszczególnymi blokami jordanowskimi. Aby to pokazać, rozważmy macierz

$$C := \begin{bmatrix}5 & 1 & 0 & 0 & 0\\0 & 5 & 1 & 0 & 0\\0 & 0 & 5 & 0 & 0\\0 & 0 & 0 & 7 & 1\\0 & 0 & 0 & 0 & 7\end{bmatrix}.$$

Układ $C' = CY$ składa się z równań

$$y_1' = 5y_1 + y_2, \qquad y_2' = 5y_2 + y_3, \qquad y_3' = 5y_3,$$
$$y_4' = 7y_4 + y_5, \qquad y_5' = 7y_5.$$

Trzy początkowe równania (związane z blokiem $J(5,3)$) tworzą podukład, który można rozwiązać niezależnie od podukładu złożonego z dwóch ostatnich równań (blok $J(7,2)$). Tak samo jest dla innych macierzy C.

Przykład 8.11.7. Dla powyższej macierzy C rozwiązać zagadnienie początkowe: $Y' = CY$, $Y(0) = (3,2,8,4,1)$.

Rozwiązanie. Podane już informacje pozwalają rozbić to zagadnienie na dwie niezależne części:

$$\begin{bmatrix} y_1' \\ y_2' \\ y_3' \end{bmatrix} = \begin{bmatrix} 5 & 1 & 0 \\ 0 & 5 & 1 \\ 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, \quad \text{wartość początkowa} \quad \begin{bmatrix} 3 \\ 2 \\ 8 \end{bmatrix},$$

$$\begin{bmatrix} y_4' \\ y_5' \end{bmatrix} = \begin{bmatrix} 7 & 1 \\ 0 & 7 \end{bmatrix} \begin{bmatrix} y_4 \\ y_5 \end{bmatrix}, \quad \text{wartość początkowa} \quad \begin{bmatrix} 4 \\ 1 \end{bmatrix}.$$

Rozwiązujemy je metodą pokazaną w przykł. 8.11.6.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = e^{5t}\left(I_3 + tH_3 + \frac{1}{2}t^2H_3^2\right) = e^{5t}\begin{bmatrix} 3 \\ 2 \\ 8 \end{bmatrix} + te^{5t}\begin{bmatrix} 2 \\ 8 \\ 0 \end{bmatrix} + \frac{1}{2}t^2e^{5t}\begin{bmatrix} 8 \\ 0 \\ 0 \end{bmatrix} =$$

$$= \begin{bmatrix} 3 + 2t + 4t^2 \\ 2 + 8t \\ 8 \end{bmatrix} e^{5t},$$

$$\begin{bmatrix} y_4 \\ y_5 \end{bmatrix} = e^{7t}(I_2 + tH_2)\begin{bmatrix} 4 \\ 1 \end{bmatrix} = e^{7t}\begin{bmatrix} 4 \\ 1 \end{bmatrix} + te^{7t}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 + t \\ 1 \end{bmatrix} e^{7t}. \quad \blacksquare$$

Wiemy już, że funkcja wykładnicza e^{At} jest kluczem do rozwiązywania układów $X' = AX$ z warunkami początkowymi. Znając postać kanoniczną Jordana C macierzy A, a także macierz P taką, że $P^{-1}AP = C$, możemy wykonać przekształcenie $X = PY$, rozwiązać układ $Y' = CY$ i wrócić do X, co daje ostatecznie równość

$$X = PY = Pe^{Ct}Y(0) = Pe^{Ct}P^{-1}X(0).$$

Z drugiej strony wiemy, że $X = e^{At}X(0)$. Porównując tę równość z poprzednią, wnioskujemy, że

$$e^{At} = Pe^{Ct}P^{-1}.$$

Stąd wynika sposób obliczenia macierzy e^{At}.

Zagadnienie niejednorodne

Podane już wiadomości można teraz zastosować do zagadnienia *niejednorodnego*

$$X' = AX + W, \tag{8.11.7}$$

gdzie W jest funkcją wektorową zmiennej t. Załóżmy najpierw, że A jest macierzą prostej struktury, tzn. jest podobna do macierzy przekątniowej: $P^{-1}AP = \Lambda$. Przekształcenie $X = PY$ zastosowane do (8.11.7) daje układ

$$Y' = \Lambda Y + P^{-1}W,$$

który rozpada się na niezależne równania postaci

$$\eta'(t) = \lambda\eta(t) + g(t).$$

Rozwiązanie każdego z nich wyraża się wzorem

$$\eta(t) = e^{\lambda t}\Big[\eta(0) + \int_0^t e^{-\lambda s} g(s)\, ds\Big]. \tag{8.11.8}$$

Przykład 8.11.8. Rozwiązać układ $X' = AX + W$ dla

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \qquad W = \begin{bmatrix} t^2 \\ t \\ \sin t \end{bmatrix}, \qquad X(0) = \begin{bmatrix} 5 \\ 7 \\ 9 \end{bmatrix}.$$

Rozwiązanie. Układ rozpada się na niezależne równania różniczkowe

$$x_1' = t^2, \qquad x_2' = x_2 + t, \qquad x_3' = 2x_3 + \sin t,$$

które zgodnie z (8.11.8) mają odpowiednio rozwiązania

$$x_1(t) = 5 + \int_0^t s^2\, ds = 5 + \frac{1}{3}t^3,$$

$$x_2(t) = e^t\left[7 + \int_0^t e^{-s} s\, ds\right] = 8e^t - t - 1,$$

$$x_3(t) = e^{2t}\left[9 + \int_0^t e^{-2s}\sin s\, ds\right] = \frac{46}{5}e^{2t} - \frac{2}{5}\sin t - \frac{1}{5}\cos t. \qquad \blacksquare$$

Jeśli A nie jest macierzą prostej struktury, to aby otrzymać oddzielne równania jak wyżej, używamy postaci kanonicznej Jordana i odpowiedniego przekształcenia $X = PY$. Wystarczy wytłumaczyć to na przykładzie pojedynczego bloku jordanowskiego.

Przykład 8.11.9. Rozwiązać układ $X' = AX + W$ dla

$$A = \begin{bmatrix} 5 & 1 & 0 \\ 0 & 5 & 1 \\ 0 & 0 & 5 \end{bmatrix}, \quad W = \begin{bmatrix} t^2 \\ t \\ \sin t \end{bmatrix}, \quad X(0) = \begin{bmatrix} 5 \\ 7 \\ 9 \end{bmatrix}.$$

Rozwiązanie. W układzie mamy równania

$$x_1' = 5x_1 + x_2 + t^2, \quad x_2' = 5x_2 + x_3 + t, \quad x_3' = 5x_3 + \sin t.$$

Postępowanie zalecane w takich przypadkach polega na kolejnym rozwiązywaniu równań począwszy od ostatniego:

$$x_3(t) = e^{5t}\left[9 + \int_0^t e^{-5s}\sin s\,ds\right] = \left(9 + \frac{1}{26}\right)e^{5t} - \frac{5}{26}\sin t - \frac{1}{26}\cos t,$$

$$x_2(t) = e^{5t}\left\{7 + \int_0^t e^{-5s}[x_3(s) + s]\,ds\right\} =$$

$$= \left(7 + \frac{1}{25} - \frac{10}{26^2}\right)e^{5t} + \frac{1}{2}\left(9 + \frac{1}{26}\right)te^{5t} +$$

$$+ \frac{2}{26^2}(12\sin t + 5\cos t) - \frac{1}{5}t - \frac{1}{25},$$

$$x_1(t) = e^{5t}\left\{5 + \int_0^1 e^{-5s}[x_2(s) + s^2]\,ds\right\} =$$

$$= \left(5 + \frac{74}{26^3}\right)e^{5t} + \left(7 + \frac{1}{25} - \frac{10}{26^2}\right)te^{5t} + \frac{1}{4}\left(9 + \frac{1}{26}\right)t^2e^{5t} -$$

$$- \frac{2}{26^3}(55\sin t + 37\cos t) - \frac{1}{5}t^2 - \frac{1}{25}t. \qquad \blacksquare$$

Obliczanie numeryczne wartości e^A może nastręczać pewne kłopoty. Polecamy czytelnikom lekturę artykułu Molera i van Loana [1978]. Ich badania dowodzą, że następująca czteroetapowa procedura działa dobrze w większości przypadków:

1. Wybieramy najmniejszą liczbę naturalną j, dla której $\|A\|/2^j \leqslant 1/2$.
2. Dla danego dopuszczalnego błędu ε wybieramy najmniejszą liczbę naturalną p, dla której $2^{p-3}(p+1) \geqslant 1/\varepsilon$.
3. Obliczamy $e^{A/2^j}$ za pomocą sumy częściowej $\sum_{k=0}^{p} z^k/k!$ szeregu potęgowego dla e^z.
4. Podnosimy j razy do kwadratu macierz otrzymaną w kroku **3**, co daje $(e^{A/2^j})^{2^j} = e^A$.

To postępowanie daje e^{A+E}, gdzie macierz E spełnia warunek $\|E\|/\|A\| \leqslant \varepsilon$; wynika to z wniosku 1 zawartego w cytowanym artykule.

ZADANIA 8.11

1. Uzasadnić szczegółowo, jak dowodzi się zbieżności szeregu (8.11.3), korzystając z (8.11.4) i z zupełności przestrzeni macierzy.
2. Stosując indukcję względem k, udowodnić, że jeśli wektory własne $V_1, V_2, \ldots, V_k$ pewnej macierzy odpowiadają różnym wartościom własnym, to są niezależne liniowo.

3. Niech B będzie macierzą stopnia n, a $V_1, V_2, \ldots, V_k$ – wektorami z $\mathbb{R}^n$ takimi, że $V_1 \neq 0$, $BV_1 = 0$, $BV_2 = V_1, \ldots, BV_k = V_{k-1}$. Stosując indukcję względem k, wykazać, że te wektory są niezależne liniowo. Jak duże może być k?

4. Znaleźć ogólne rozwiązanie układu $x_1' = 3x_1 - 5x_2$, $x_2' = 2x_1 + x_2$.

5. Znaleźć ogólne rozwiązanie układu $X' = AX$ dla

$$A = \begin{bmatrix} 1 & 0 & 3 \\ -1 & 1 & -1 \\ 3 & 0 & 1 \end{bmatrix}$$

i rozwiązanie spełniające warunek $X(0) = (-1, 4, 7)$.

6. Rozwiązać zagadnienie początkowe $X' = AX$, $X(0) = X_0$ dla

$$A = \begin{bmatrix} 5 & 4 & 3 \\ -1 & 0 & -3 \\ 1 & -2 & 1 \end{bmatrix}, \quad X_0 = \begin{bmatrix} 1 \\ 5 \\ 3 \end{bmatrix}.$$

7. Wykazać, że jeśli

$$A = \begin{bmatrix} -1 & 6 \\ 1 & -2 \end{bmatrix},$$

to

$$e^{At} = \frac{1}{5} \begin{bmatrix} 2e^{-4t} + 3e^t & -6e^{-4t} + 6e^t \\ -e^{-4t} + e^t & 3e^{-4t} + 2e^t \end{bmatrix}.$$

8. Udowodnić, że każda macierz złożonej struktury jest granicą ciągu macierzy prostej struktury.

9. (cd.). Niech w zagadnieniu początkowym $X' = AX$, $X(0) = V$ występuje macierz A złożonej struktury. Niech B będzie macierzą prostej struktury, bliską A. Co można powiedzieć o rozwiązaniu zagadnienia $Y' = BY$, $Y(0) = V$?

10. Udowodnić, że $\det e^A = e^{\mathrm{tr}A}$, gdzie $\mathrm{tr}A$ jest *śladem* macierzy A, tj. sumą jej elementów przekątniowych.

11. Udowodnić, że dla dowolnej macierzy kwadratowej A macierz e^A jest nieosobliwa.

12. Dla macierzy $A = (a_{ij})$ stopnia n niech będzie $|A| := \sum_{i=1}^n \sum_{j=1}^n |a_{ij}|$. Udowodnić, że $|e^A| \leqslant n - 1 + e^{|A|}$.

13. Udowodnić, że j-ta kolumna macierzy e^{At} jest rozwiązaniem zagadnienia początkowego $X' = AX$, $X(0) = U_j$, gdzie U_j jest j-tym wektorem jednostkowym.

14. Udowodnić, że $e^{A+B} = e^A e^B$ wtedy i tylko wtedy, gdy $AB = BA$.

15. Znaleźć macierz odwrotną względem e^A.

16. Wykazać, że (chociaż na ogół $e^{A+B} \neq e^A e^B$) jest $e^{A+B} = \lim_{k\to\infty}(e^{A/k} e^{B/k})^k$.

17. Udowodnić, że jeśli $PAP^{-1} = \Lambda$ i $PBP^{-1} = \Lambda^*$, gdzie Λ i Λ^* są macierzami przekątniowymi, to $AB = BA$.

18. Wykazać, że jeśli $A = P^{-1}BP$, to $e^A = P^{-1}e^BP$.

19. Znaleźć wszystkie macierze stopnia 2, dla których postać kanoniczna Jordana zawiera tylko jeden blok jordanowski.

20. Jak rozwiązać układ $X' = AX$ z warunkiem początkowym w punkcie $t_0 \neq 0$?

21. Udowodnić, że rozwiązanie zagadnienia początkowego $X' = AX + V(t)$, $X(0) = W$ wyraża się wzorem

$$X(t) = e^{At}W + e^{At}\int_0^t e^{-As}V(s)\,ds.$$

Jak należy rozumieć tu całkę?

22. Uzasadnić wszystkie przekształcenia wykonywane w przykł. 8.11.9.

8.12. Równania sztywne

Sztywność układu równań różniczkowych oznacza szybkie zanikanie rozwiązań w miarę wzrostu t. Pewne metody numeryczne, na ogół bardzo skuteczne, zawodzą dla równań sztywnych. Zdarza się tak, gdy stabilność rozwiązania można uzyskać tylko dla bardzo małych h.

Sztywne równania różniczkowe występują w wielu zastosowaniach. Tak np. w sterowaniu pojazdem kosmicznym trajektoria lotu powinna być gładka, ale każde odchylenie od niej może wymagać bardzo szybkich korekt. Innym źródłem takich zadań jest monitorowanie procesów chemicznych, gdyż skale czasu dla poszczególnych zjawisk fizycznych lub chemicznych mogą być bardzo różne. W teorii obwodów elektrycznych równania sztywne pojawiają się wtedy, gdy przebiegi nieustalone rzędu mikrosekund nakładają się na ogólnie gładkie zachowanie obwodu.

Metoda Eulera

Kłopoty, jakie nastręcza rozwiązywanie równań sztywnych, można zilustrować na przykładzie zastosowania metody Eulera do prostego równania. *Metoda Eulera* dla zagadnienia początkowego

$$x' = f(t,x), \quad x(t_0) = x_0$$

polega na stosowaniu wzoru

$$x_{n+1} := x_n + hf(t_n, x_n) \quad (n \geqslant 0),$$

gdzie $t_n := t_0 + nh$. Zobaczmy jak ta metoda działa dla prostego zagadnienia

$$x' = \lambda x, \quad x(0) = 1. \tag{8.12.1}$$

Jest oczywiste, że wtedy $x_{n+1} = x_n + h\lambda x_n = (1 + h\lambda)x_n$, czyli

$$x_n = (1 + h\lambda)^n. \tag{8.12.2}$$

Natomiast dokładne rozwiązanie zagadnienia (8.12.1) jest równe $x(t) = e^{\lambda t}$. Dla $\lambda < 0$ i $t \to \infty$ dąży ono wykładniczo do 0. Tę samą własność ma rozwiązanie numeryczne (8.12.2), ale tylko wtedy, gdy $|1 + h\lambda| < 1$. To zmusza nas do wybrania takiego h, żeby było $1 + h\lambda > -1$, czyli $h < -2/\lambda$. Jeśli np. $\lambda = -20$, to musi być $h < 0.1$, chociaż szukane rozwiązanie jest skrajnie płaskie (i praktycznie równe 0) niedaleko od punktu początkowego 0, gdzie $x = 1$. Istotnie, $x(t) = e^{-20t} \leqslant 2.1_{10}{-}9$ dla $t \geqslant 1$. Tak więc w obliczeniach musimy stosować *małe* kroki tam, gdzie charakter rozwiązania skłania do stosowania *dużych* kroków. Jest to pewną cechą równań sztywnych. O funkcji takiej, jak e^{-20t}, bardzo szybko zanikającej do 0 (której wykres ma silne ujemne nachylenie), mówimy, że ma przebieg *przejściowy* (*nieustalony*), gdyż jej fizyczny efekt jest krótkotrwały. Chcąc wiernie odtworzyć przebieg takiej funkcji, w obliczeniach numerycznych musimy stosować mały krok aż do momentu, gdy ten efekt stanie się pomijalny. Potem dobra metoda numeryczna powinna już pozwolić na zwiększenie kroku. Metoda Eulera do tego się nie nadaje.

Modyfikacja metody Eulera

Zupełnie inne są własności *metody niejawnej Eulera*, którą określimy wzorem

$$x_{n+1} := x_n + hf(t_{n+1}, x_{n+1}) \quad (n \geqslant 0). \tag{8.12.3}$$

Dla zagadnienia testowego (8.12.1) wynika stąd, że $x_{n+1} = x_n + h\lambda x_{n+1}$, czyli $x_{n+1} = (1 - h\lambda)^{-1}x_n$ i ostatecznie

$$x_n = (1 - h\lambda)^{-n}.$$

Dla ujemnych λ te wielkości naśladują zachowanie dokładnego rozwiązania, gdy tylko

$$|1 - h\lambda|^{-1} < 1, \tag{8.12.4}$$

a tak jest dla każdego $h > 0$.

Układy równań różniczkowych

Podobne rozważania odnoszą się do *układów* równań różniczkowych. I tym razem stosujemy zasadę, że dobra metoda numeryczna powinna dobrze działać dla prostych układów równań liniowych. (Warto przypomnieć, że tak rozumowano w podrozdz. 8.5, aby dojść do wniosku, że stabilność i zgodność

są istotnymi atrybutami akceptowalnych metod wielokrokowych). Takim układem testowym jest np. następujący układ dwóch równań różniczkowych z warunkami początkowymi:

$$\begin{aligned} x' &= \alpha x + \beta y, \qquad x(0) = 2, \\ y' &= \beta x + \alpha y, \qquad y(0) = 0. \end{aligned}$$

Jego rozwiązaniem są funkcje

$$x(t) = e^{(\alpha+\beta)t} + e^{(\alpha-\beta)t}, \qquad y(t) = e^{(\alpha+\beta)t} - e^{(\alpha-\beta)t}. \tag{8.12.5}$$

Metoda Eulera zastosowana do tego zagadnienia poleca stosować wzory

$$\begin{aligned} x_0 &:= 2, \quad x_{n+1} := x_n + h(\alpha x_n + \beta y_n), \\ y_0 &:= 0, \quad y_{n+1} := y_n + h(\beta x_n + \alpha y_n). \end{aligned} \tag{8.12.6}$$

Stąd wynika, że

$$\begin{aligned} x_n &= (1 + \alpha h + \beta h)^n + (1 + \alpha h - \beta h)^n, \\ y_n &= (1 + \alpha h + \beta h)^n - (1 + \alpha h - \beta h)^n. \end{aligned} \tag{8.12.7}$$

W interesującym nas przypadku jest $\alpha < \beta < 0$. Wtedy rozwiązania dokładne (8.12.5) dążą wykładniczo do 0. Rozwiązania przybliżone zachowują się podobnie, jeśli

$$|1 + \alpha h + \beta h| < 1, \quad |1 + \alpha h - \beta h| < 1. \tag{8.12.8}$$

Dla $\alpha < \beta < 0$ te dwie nierówności są równoważne warunkowi

$$0 < h < -2/(\alpha + \beta).$$

Jeśli np. $\alpha = -20$ i $\beta = -19$, to rozwiązania dokładne są sumą i różnicą funkcji e^{-39t} i e^{-t}. Pierwsza z nich jest funkcją przejściową; już dla niewielkich t jest ona pomijalna w porównaniu z e^{-t}, ale właśnie ona określa sensowną długość kroku.

Ogólna metoda wielokrokowa

Zbadajmy teraz *ogólną metodę wielokrokową*, opisaną równością

$$a_k x_n + a_{k-1} x_{n-1} + \ldots + a_0 x_{n-k} = h(b_k f_n + b_{k-1} f_{n-1} + \ldots + b_0 f_{n-k})$$

(podrozdz. 8.5), aby sprawdzić, jakie dodatkowe własności zapewniają jej skuteczność w zagadnieniu testowym (8.12.1). W tym przypadku relacja upraszcza się do postaci

$$a_k x_n + a_{k-1} x_{n-1} + \ldots + a_0 x_{n-k} = h\lambda(b_k x_n + b_{k-1} x_{n-1} + \ldots + b_0 x_{n-k}).$$

Mamy tu zatem równanie różnicowe liniowe jednorodne

$$(a_k - h\lambda b_k)x_n + (a_{k-1} - h\lambda b_{k-1})x_{n-1} + \ldots + (a_0 - h\lambda b_0)x_{n-k} = 0. \quad (8.12.9)$$

Każde jego rozwiązanie jest kombinacją liniową rozwiązań typu $x_n = r^n$, gdzie r jest zerem wielomianu $\varphi := p - \lambda h q$, a p i q są wielomianami określonymi w cytowanym podrozdziale:

$$p(z) := a_k z^k + a_{k-1} z^{k-1} + \ldots + a_0,$$
$$q(z) := b_k z^k + b_{k-1} z^{k-1} + \ldots + b_0$$

(ściślej, rozwiązania są takie, jeśli wielomian φ ma tylko zera pojedyncze).

A-stabilność

Niech będzie $\lambda < 0$. Rozwiązanie równania różnicowego (8.12.9) maleje do 0 podobnie jak rozwiązanie dokładne zagadnienia (8.12.1), jeśli wszystkie zera wielomianu φ leżą w kole $|z| < 1$. Dla $\lambda = \mu + i\nu$ zespolonego to rozwiązanie dokładne jest równe

$$x(t) = e^{\lambda t} = e^{\mu t}(\cos \nu t + i \sin \nu t)$$

i dąży wykładniczo do 0, gdy $\mu < 0$. Metoda wielokrokowa w takim przypadku jest sensowna, jeśli zera wielomianu φ leżą w kole $|z| < 1$ dla $h > 0$ i $\Re\lambda < 0$. Taką metodę nazywamy *A-stabilną*.

Niejawna metoda Eulera (8.12.3) jest A-stabilna, co wynika z (8.12.4). *Niejawna metoda trapezów* określona wzorem

$$x_n - x_{n-1} = \tfrac{1}{2}h(f_n + f_{n-1}) \quad (8.12.10)$$

jest także A-stabilna, gdyż dla niej

$$\varphi(z) = z - 1 - \tfrac{1}{2}\lambda h(z + 1),$$

a zerem tego wielomianu jest $z = (2 + \lambda h)/(2 - \lambda h)$. Można łatwo sprawdzić, że jeśli $h > 0$ i $\Re\lambda < 0$, to $|z| < 1$.

Ważne twierdzenie (Dahlquist [1963]) orzeka, że metoda wielokrokowa A-stabilna musi być niejawna, a jej rząd nie przekracza 2. Jest to bardzo mocne ograniczenie klasy metod A-stabilnych. Należąca do niej metoda (8.12.10) jest często używana dla równań sztywnych, gdyż jej błąd jest najmniejszy w tej klasie.

Obszar stabilności bezwzględnej

Każda metoda wielokrokowa ma pewien *obszar stabilności bezwzględnej* (*absolutnej*). Jest to zbiór tych liczb zespolonych ω, dla których wszystkie zera wielomianu $\varphi = p - \omega q$ leżą w kole $|z| < 1$. Z wcześniejszych rozważań wynika, że metoda dobrze działa dla równania testowego $x' = \lambda x$, jeśli λh należy do tego obszaru. Zauważmy także, że metoda jest A-stabilna, jeśli jej obszar stabilności bezwzględnej zawiera półpłaszczyznę $\Re z < 0$.

Przykład 8.12.1. Znaleźć obszar stabilności bezwzględnej metody Eulera.

Rozwiązanie. Ponieważ metodę Eulera określa wzór

$$x_n - x_{n-1} = hf_{n-1},$$

więc $p(z) = z - 1$, $q(z) = 1$ i $\varphi(z) = z - 1 - \omega$, gdzie $\omega := \lambda h$. Zerem wielomianu φ jest $z = 1 + \omega$. Obszarem bezwzględnej stabilności jest więc koło $|1 + \omega| < 1$. ■

Jak już wiemy, żądanie A-stabilności bardzo ogranicza wybór użytecznych metod. Dlatego, rozwiązując zadania sztywne, stosujemy także metody, które nie mają tej własności. W takich przypadkach spodziewamy się, że $\omega = \lambda h$ leży w obszarze bezwzględnej stabilności metody. Wielkość λ odnosi się do rozwiązywanego równania różniczkowego. Dla równania liniowego jest to współczynnik przy x. Dla równania nieliniowego λ może być analogiczną stałą, związaną z lokalnym przybliżeniem tego równania równaniem liniowym. Dla układu $X' = AX$ równań liniowych λ może być dowolną wartością własną macierzy A. Dlatego iloczyn każdej takiej wartości przez h powinien leżeć w obszarze bezwzględnej stabilności wybranej metody. Dla układu nieliniowego można by stosować lokalne przybliżenie liniowe, a następnie rozumować jak wyżej. W praktyce jednak na ogół jest to niewykonalne i jedyną bezpieczną strategią w przypadku trudnych zagadnień sztywnych jest powrót do metody trapezów. Najlepsze programy dla zagadnień początkowych wykrywają sztywność w toku obliczeń i podejmują odpowiednie czynności zaradcze. Szczegóły omawiają Gear [1971], Shampine i Allen [1973] oraz Shampine i Gordon [1975]. Na ogół metody wyższego rzędu mają mały obszar bezwzględnej stabilności i wspomniane programy po wykryciu sztywności odwołują się do metod niższego rzędu z szerszym takim obszarem; zob. Byrne i Hindmarsh [1987] lub Shampine i Gear [1979].

Równania nieliniowe

Aby pokazać związek powyższych rozważań z układami równań *nieliniowych*, rozważmy typowy układ autonomiczny

$$X' = F(X),$$

w którym $X = (x_1, x_2, \ldots, x_n)$; prawa strona jest funkcją $F\colon \mathbb{R}^n \to \mathbb{R}^n$ o składowych f_i. W punkcie X_0 układ $X' = F(X)$ zachowuje się podobnie jak układ liniowy

$$X' = F(X_0) + J(X_0)(X - X_0),$$

gdzie J jest jakobianem funkcji F, czyli macierzą kwadratową J o elementach $\partial f_i/\partial x_j$. W badaniu metod numerycznych istotne są wartości własne λ_i macierzy $J(X_0)$. Pierwotny układ jest sztywny w otoczeniu punktu X_0, jeśli spełniają one nierówności

$$\Re\lambda_1 \leqslant \Re\lambda_2 \leqslant \ldots \leqslant \Re\lambda_n < 0$$

i jeśli $\Re\lambda_1$ jest znacznie mniejsze od $\Re\lambda_n$.

Metoda wielokrokowa dla takiego układu nieliniowego może poprawnie działać tylko wtedy, gdy wszystkie $\lambda_i h$ leżą w obszarze jej bezwzględnej stabilności. Jeśli wartości własne λ_i są znane, to można użyć tej informacji wybierając odpowiednie h.

Zadania 8.12

1. Sprawdzić, czy zagadnienie początkowe

$$x'' = -20x' - 19x, \quad x(0) = 2, \quad x'(0) = -20$$

jest sztywne.

2. Sprawdzić, czy równanie $x'' = (\sin t + 57)x$ jest sztywne.

3. Pewną reakcję chemiczną opisuje układ sztywny równań różniczkowych:

$$x_1' = -1000x_1 + x_2, \quad x_1(0) = 1, \qquad x_2' = 999x_1 - 2x_2, \quad x_2(0) = 0.$$

Wykazać, że x_1 dąży do 0 szybko, a x_2 wolno.

4. Wykazać, że wielkości określone w (8.12.7) spełniają równania (8.12.6).

5. Wykazać, że dla $\alpha < \beta < 0$ warunki (8.12.8) są równoważne nierówności $0 < h < -2/(\alpha + \beta)$.

6. Znaleźć takie α, dla których metoda wielokrokowa

$$x_n + \alpha x_{n-1} - (1 + \alpha)x_{n-2} = \tfrac{1}{2}h[-\alpha f_n + (4 + 3\alpha)f_{n-1}]$$

jest stabilna, zgodna, zbieżna, A-stabilna i rzędu drugiego (zob. podrozdziały 8.4 i 8.5).

7. Znaleźć obszar bezwzględnej stabilności dla:
(**a**) niejawnej metody Eulera (8.12.3),
(**b**) niejawnej metody trapezów (8.12.10).

ZADANIA KOMPUTEROWE 8.12

K1. Przetestować dla równania $x' = \lambda x$, gdzie $\lambda < 0$, *niejawną metodę punktu środkowego*

$$x_n - x_{n-1} = hf\big(t_{n-1} + \tfrac{1}{2}, \tfrac{1}{2}(x_n + x_{n-1})\big).$$

Czy można ją stosować do zagadnień sztywnych?

K2. Stosując przekształcenie Riccatiego $y = x'/x$, przekształcić równanie różniczkowe z zad. 2 z warunkami początkowymi $x(0) = 1$, $x'(0) = -\sqrt{57}$ na zagadnienie początkowe

$$x' = xy, \quad x(0) = 1, \quad y' = 57 + \sin t - y^2, \quad y(0) = -\sqrt{57}$$

i rozwiązać je, stosując dostępne programy.

ROZDZIAŁ 9

Rozwiązywanie numeryczne równań różniczkowych cząstkowych

9.0. Wstęp

Tematem tego rozdziału jest rozwiązywanie numeryczne równań różniczkowych cząstkowych. Naszkicowane tu typowe zadania i metody ich rozwiązywania pozwolą zrozumieć, że jest do tego niezbędne zastosowanie największych i najszybszych komputerów. Nic więc dziwnego, że ta dziedzina analizy numerycznej w dalszym ciągu intensywnie się rozwija.

9.1. Równania paraboliczne: metody jawne

Równanie ciepła

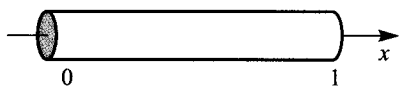
Zaczynamy od przykładowego równania *typu parabolicznego*, a mianowicie od *równania przewodnictwa cieplnego*. Wybierając w sensowny sposób jednostki miary dla wielkości fizycznych, wyrażamy to równanie tak:

$$u_{xx} + u_{yy} + u_{zz} = u_t. \tag{9.1.1}$$

Nieznana funkcja u zależy od zmiennych przestrzennych x, y, z i czasu t. Jej pochodne cząstkowe $\partial u/\partial t$, $\partial^2 u/\partial x^2, \ldots$ oznaczamy tu $u_t, u_{xx}, \ldots$ Równanie (9.1.1) opisuje zależność temperatury u od czasu t i punktu (x, y, z) pewnego ciała trójwymiarowego.

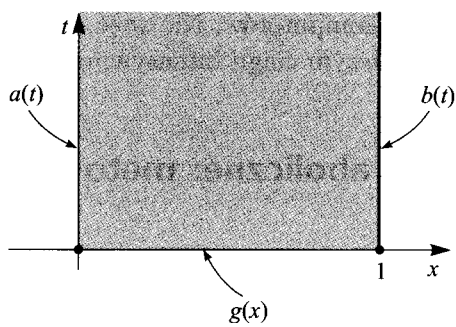
Jak w przypadku równań różniczkowych zwyczajnych, tak i tutaj samo równanie różniczkowe nie określa jednoznacznie rozwiązania sensownie postawionego zadania fizycznego. Potrzebne są dodatkowe *warunki graniczne* w dostatecznej liczbie. W prostym modelowym zadaniu typu (9.1.1), w którym ograniczamy się do jednej zmiennej przestrzennej, równanie wraz z niezbędnymi warunkami początkowymi i brzegowymi ma następującą postać:

$$\begin{aligned} &u_{xx} = u_t && (t \geqslant 0,\ \ 0 \leqslant x \leqslant 1),\\ &u(x,0) = g(x) && (0 \leqslant x \leqslant 1),\\ &u(0,t) = a(t) && (t \geqslant 0),\\ &u(1,t) = b(t) && (t \geqslant 0). \end{aligned} \tag{9.1.2}$$



RYS. 9.1. Pręt o długości 1

Jest to model matematyczny rozkładu temperatury w pręcie o długości 1, na którego końcach jest podtrzymywana temperatura $a(t)$ i $b(t)$ (rys. 9.1); funkcje a i b są dane. Ponadto przyjmuje się, że dana jest temperatura początkowa $g(x)$ na całej długości pręta. Funkcja u zależy od x i t. Rysunek 9.2 pokazuje obszar, w którym jej szukamy.



RYS. 9.2. Obszar, w którym ma być określone rozwiązanie

Metoda różnic skończonych

Jedną z głównych metod rozwiązywania zadań typu (9.1.2) jest *metoda różnic skończonych*. We wspomnianym wyżej obszarze szukamy przybliżonych wartości funkcji u na *siatce* punktów (x_i, t_j) o współrzędnych

$$t_j := jk \quad (j \geqslant 0), \qquad x_i := ih \quad (0 \leqslant i \leqslant n+1).$$

Długość kroku dla zmiennych t i x (odpowiednio k i h) może być różna. Ponieważ x przebiega przedział $[0,1]$, więc $h = 1/(n+1)$.

Drugim istotnym elementem metody jest zastąpienie pochodnych występujących w równaniu różniczkowym wyrażeniami przybliżonymi. Najprostsze z nich, ale nie jedyne, już znamy:

$$f'(x) \approx \frac{1}{h}[f(x+h) - f(x)],$$
$$f'(x) \approx \frac{1}{2h}[f(x+h) - f(x-h)],$$
$$f''(x) \approx \frac{1}{h^2}[f(x+h) - 2f(x) + f(x-h)]$$

(zob. podrozdz. 7.1). Jeśli zastosujemy pierwszy i trzeci wzór, to z równania $u_{xx} = u_t$ wyniknie równość

$$\frac{1}{h^2}[v(x+h,t) - 2v(x,t) + v(x-h,t)] = \frac{1}{k}[v(x,t+k) - v(x,t)].$$

Użyto tu litery v zamiast u, aby podkreślić, że wykonana *dyskretyzacja* zadania zmienia jego rozwiązanie. Tę równość dla punktów siatki wyrażamy prościej:

$$\frac{1}{h^2}(v_{i+1,j} - 2v_{ij} + v_{i-1,j}) = \frac{1}{k}(v_{i,j+1} - v_{ij}), \tag{9.1.3}$$

gdzie $v_{ij} := v(x_i, t_j)$. Dla $j = 0$ wszystkie występujące tu wielkości z wyjątkiem v_{i1} są znane, gdyż początkowy rozkład temperatur określa wartości (dokładne)

$$v_{i0} = u(x_i, 0) = g(x_i).$$

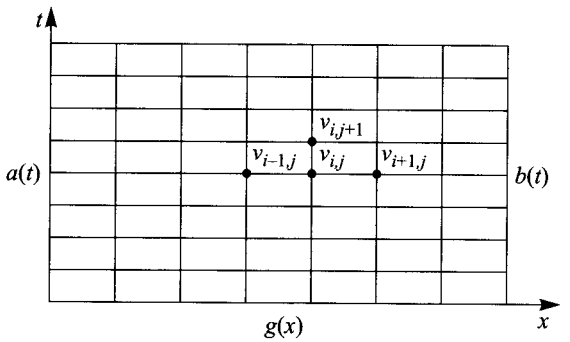
Dlatego korzystając z (9.1.3), można obliczyć v_{i1}. Ogólniej, przekształćmy te równości do postaci

$$v_{i,j+1} = \frac{k}{h^2}(v_{i+1,j} - 2v_{ij} + v_{i-1,j}) + v_{ij},$$

czyli

$$v_{i,j+1} = sv_{i-1,j} + (1-2s)v_{ij} + sv_{i+1,j}, \quad \text{gdzie} \quad s := \frac{k}{h^2}. \tag{9.1.4}$$

Posługując się wzorem (9.1.4) dla $j = 0, 1, \ldots$ i $1 \leqslant i \leqslant n$, można obliczać rozwiązanie przybliżone dla coraz większych t. Rysunek 9.3 pokazuje rozmieszczenie czterech punktów siatki, których ten wzór dotyczy. Dla skrajnych wartości wskaźnika i korzystamy przy tym z warunków brzegowych: $v_{0j} = a(jk)$, $v_{n+1,j} = b(jk)$. Ponieważ wzór (9.1.4) wyraża nowe wartości $v_{i,j+1}$ tylko przez wartości już znane, mamy tu *metodę jawną*.



RYS. 9.3. Typowa siatka

Podkreślamy, że jeśli nawet wszystkie działania w (9.1.4) wykonujemy dokładnie, to wartości v_{ij} nie są równe odpowiednim wartościom rozwiązania u zagadnienia (9.1.2). Przeciwnie, można się spodziewać dość dużych błędów, gdyż użyto mało dokładnych wyrażeń przybliżonych dla pochodnych.

Algorytm

W poniższym algorytmie zachowano sens symboli n i k, natomiast M oznacza liczbę dodatnich wartości zmiennej t, dla których chcemy znaleźć rozwiązanie.

input n, k, M
$h \leftarrow 1/(n+1)$; $s \leftarrow k/h^2$
for $i = 0$ **to** $n+1$ **do**
    $w_i \leftarrow g(ih)$
end do
$t \leftarrow 0$
output 0, t, $(w_0, w_1, \ldots, w_{n+1})$
for $j = 1$ **to** M **do**

```
        t ← jk
        v_0 ← a(t); v_{n+1} ← b(t)
        for i = 1 to n do
            v_i ← s w_{i-1} + (1 − 2s) w_i + s w_{i+1}
        end do
        output j, t, (v_0, v_1, ..., v_{n+1})
        for i = 0 to n + 1 do
            w_i ← v_i
        end do
    end do
```

Zachęcamy czytelników do zaprogramowania tego algorytmu i sprawdzenia go na przykładzie z zad. K1. Takie testy prowadzą do wniosku, że nie wszystkie pary (h, k) można zaakceptować. Powody wyjaśniamy niżej dla szczególnego przypadku wartości brzegowych: $a(t) = b(t) = 0$.

Analiza stabilności

Wzór (9.1.4) można wyrazić w bardziej zwarty sposób, używając symboliki macierzowej. Niech V_j będzie wektorem przybliżonych wartości rozwiązania w czasie $t = jk$; tak więc $V_j = (v_{1j}, v_{2j}, \ldots, v_{nj})$ (zerowe wartości brzegowe pomijamy). Wtedy

$$V_{j+1} = AV_j, \tag{9.1.5}$$

gdzie A jest następującą macierzą stopnia n:

$$A = \begin{bmatrix} 1-2s & s & 0 & \ldots & 0 \\ s & 1-2s & s & \ldots & 0 \\ 0 & s & 1-2s & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1-2s \end{bmatrix}.$$

Można teraz rozumować dwojako. Po pierwsze, uzwględniając interpretację fizyczną zadania zauważamy, że dla $t \to \infty$ temperatura w pręcie dąży do 0, bo na jego końcach jest taka. Wobec tego rozwiązanie numeryczne też się powinno tak zachowywać. Z (9.1.5) wynika, że $V_j = A^j V_0$. Na mocy tw. 4.6.7 ciąg $\{A^j V\}$ dąży do 0 dla każdego V wtedy i tylko wtedy, gdy promień spektralny $\rho(A)$ macierzy A jest mniejszy od 1. Dlatego parametr $s = k/h^2$ trzeba wybrać tak, aby ta nierówność była spełniona.

Do tego samego wniosku można dojść nie odwołując się do fizycznej natury zadania, natomiast analizując wpływ błędów zaokrągleń na wyniki obliczeń. Przypuśćmy, że w pewnym ich etapie (możemy przyjąć, że pierwszym) dane są obarczone takimi błędami. Inaczej mówiąc, zamiast dokładnego V_0 mamy wektor zaburzony $\tilde{V}_0$. Wtedy metoda daje wektory $\tilde{V}_j := A^j \tilde{V}_0$

i błąd j-tego etapu wynosi

$$V_j - \tilde{V}_j = A^j(V_0 - \tilde{V}_0).$$

Aby być pewnym, że ten błąd zanika dla $j \to \infty$, musimy znów założyć, że $\rho(A) < 1$.

Obliczając $\rho(A)$, zauważmy, że $A = I - sB$, gdzie

$$B := \begin{bmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ 0 & -1 & 2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 2 \end{bmatrix}. \tag{9.1.6}$$

Wartości własne λ_j macierzy A wyrażają się przez takież wartości μ_j macierzy B wzorem $\lambda_j = 1 - s\mu_j$. Z poniższego lem. 9.1.1 wynika więc, że

$$\lambda_j := 1 - 2s(1 - \cos\theta_j), \quad \text{gdzie} \quad \theta_j := \frac{j\pi}{n+1} \quad (1 \leqslant j \leqslant n).$$

Nierówność $\rho(A) < 1$ zachodzi, jeśli dla każdego j jest

$$-1 < 1 - 2s(1 - \cos\theta_j) < 1,$$

czyli $s < (1 - \cos\theta_j)^{-1}$. Ponieważ s jest dodatnie, więc ta nierówność tym bardziej ogranicza s, im $\cos\theta_j$ jest bliższe -1. Dla $j = n$ ta wielkość jest prawie równa -1 i żądamy spełnienia nierówności $s \leqslant 1/2$.

Ostateczny wniosek jest więc taki: *warunkiem koniecznym stabilności* metody opisanej wzorem (9.1.4) jest nierówność $s = k/h^2 \leqslant 1/2$. Jest ona bardzo kłopotliwa, gdyż zmusza do długich obliczeń. Jeśli np. $h = 0.01$, to powinno być $k \leqslant 5_{10}{-}5$. Chcąc znaleźć rozwiązanie dla $t \leqslant 10$, musimy przejść przez 200 000 wartości zmiennej t, a liczba punktów siatki przekroczy 20 milionów! Tak więc elegancji i prostocie metody towarzyszy jej skrajna nieefektywność.

Lemat 9.1.1. *Wartości własne macierzy* (9.1.6) *stopnia n są równe*

$$2 - 2\cos\theta_j, \quad \textit{gdzie} \quad \theta_j := j\pi/(n+1) \quad (1 \leqslant j \leqslant n).$$

Dowód. Wartości własne macierzy B_n (wskaźnik oznacza jej stopień) są zerami jej wielomianu charakterystycznego $D_n(\lambda) := \det(B_n - \lambda I_n)$. Ponieważ B_n jest macierzą trójprzekątniową, więc rozwijając ten wyznacznik względem pierwszej kolumny, otrzymujemy zależność rekurencyjną

$$D_n(\lambda) = (2 - \lambda)D_{n-1}(\lambda) - D_{n-2}(\lambda) \quad (n \geqslant 2),$$

gdzie $D_0(\lambda) := 1$, $D_1(\lambda) = 2 - \lambda$. Natomiast wielomiany Czebyszewa drugiego rodzaju (podrozdz. 7.2) U_n są takie, że

$$U_0(x) = 1, \quad U_1(x) = 2x, \quad U_n(x) = 2xU_{n-1}(x) - U_{n-2}(x) \quad (n \geqslant 2).$$

Jest więc oczywiste, że $D_n(\lambda) = U_n(1 - \lambda/2)$. Tamże podano, że zera wielomianu U_n są równe $\cos\theta_j$, skąd wynika teza lematu. ■

Analiza stabilności – metoda Fouriera

Kwestia stabilności jest istotna niemal dla wszystkich równań różniczkowych cząstkowych, w których czas jest jedną ze zmiennych niezależnych. Jest to naturalne, gdyż mogą być nam potrzebne rozwiązania w dużym przedziale czasowym. Wyżej stabilność badano *metodą macierzową*. Inna metoda, przypisywana von Neumannowi, nosi nazwę *metody Fouriera*. Polega ona na tym, że szukamy rozwiązania równań różnicowych mającego postać

$$v_{jn} = e^{ij\beta h}e^{n\lambda k} \quad (\beta \text{ rzeczywiste}, \ i := \sqrt{-1}). \tag{9.1.7}$$

Jeśli to się udało dzięki dobraniu odpowiedniego λ, to badamy własności rozwiązania dla $t \to \infty$, czyli $n \to \infty$. Oczywiście zależą one od czynnika $(e^{\lambda k})^n$. Jeśli $|e^{\lambda k}| > 1$, to rozwiązanie staje się nieograniczone. Każde takie rozwiązanie będzie dominować nad rozwiązaniem dokładnym, które dąży do 0.

Dlaczego jednak rozważamy właśnie rozwiązania (9.1.7)? Wynika to stąd, że jednym z rozwiązań równania przewodnictwa cieplnego w (9.1.2) jest funkcja $u(x,t) = \exp(-\pi^2 t)\sin\pi x$, a rozwiązanie równania różnicowego powinno mieć podobną postać.

Zbadajmy w proponowany tu sposób metodę jawną (9.1.4). Podstawiamy próbne rozwiązanie (9.1.7) do tego wzoru (ze zmienionymi odpowiednio wskaźnikami):

$$e^{ij\beta h}e^{(n+1)\lambda k} = se^{i(j-1)\beta h}e^{n\lambda k} + (1-2s)e^{ij\beta h}e^{n\lambda k} + se^{i(j+1)\beta h}e^{n\lambda k},$$

co po uproszczeniu daje

$$e^{\lambda k} = se^{-i\beta h} + 1 - 2s + se^{i\beta h} = 2s\cos\beta h + 1 - 2s = 1 - 4s\sin^2(\beta h/2).$$

Ta wielkość jest rzeczywista, a dla metody stabilnej powinna należeć do przedziału $[-1,1]$. Jest tak, jeśli $s\sin^2(\beta h/2) \leqslant 1/2$. Ponieważ $\sin^2(\beta h/2)$ może być bliskie 1, więc otrzymujemy ostatecznie znany już warunek stabilności: $s = k/h^2 \leqslant 1/2$.

ZADANIA 9.1

1. Wykazać, że funkcja

$$u(x,t) := \sum_{n=1}^{N} c_n \exp(-n^2\pi^2 t)\sin n\pi x$$

spełnia równanie przewodnictwa cieplnego $u_{xx} = u_t$ i warunki

$$u(x,0) = \sum_{n=1}^{N} c_n \sin n\pi x, \quad u(0,t) = u(1,t) = 0.$$

ZADANIA KOMPUTEROWE 9.1

K1. Korzystając z algorytmu w tekście, znaleźć rozwiązanie numeryczne zadania przewodnictwa cieplnego

$$u_{xx} = u_t, \quad u(x,0) = \sin \pi x, \quad u(0,t) = u(1,t) = 0.$$

Przyjąć, że $h = 0.1$, $k = 0.005125$ i $M = 200$. Porównać wyniki z dokładnym rozwiązaniem $u(x,t) = \exp(-\pi^2 t)\sin(\pi x)$. Następnie powtórzyć obliczenia dla $k = 0.006$ i $M = 171$.

9.2. Równania paraboliczne: metody niejawne

Będziemy w dalszym ciągu badać zadanie przewodnictwa cieplnego z zerowymi warunkami brzegowymi:

$$\begin{aligned} &u_{xx} = u_t && (t \geqslant 0,\ 0 \leqslant x \leqslant 1),\\ &u(x,0) = g(x) && (0 \leqslant x \leqslant 1),\\ &u(0,t) = u(1,t) = 0 && (t \geqslant 0). \end{aligned} \tag{9.2.1}$$

Podobnie jak w poprzednim podrozdziale zastępujemy pochodne ich przybliżeniami różnicowymi, ale tym razem zastosujemy inne przybliżenie pierwszej pochodnej:

$$\frac{1}{h^2}[v(x+h,t) - 2v(x,t) + v(x-h,t)] = \frac{1}{k}[v(x,t) - v(x,t-k)].$$

Używając zwykłych oznaczeń dla wartości funkcji w punktach siatki, wyrażamy to prościej:

$$\frac{1}{h^2}(v_{i+1,j} - 2v_{ij} + v_{i-1,j}) = \frac{1}{k}(v_{ij} - v_{i,j-1}). \tag{9.2.2}$$

Pozornie drobna zmiana w stosunku do relacji (9.1.3) daje inny algorytm. Istotnie, w (9.2.2) występują trzy wartości v z drugim wskaźnikiem j i tylko jedna z tym wskaźnikiem równym $j-1$. Jeśli dla ustalonego j znamy wszystkie $v_{i,j-1}$, to znalezienie wszystkich v_{ij} wymaga rozwiązania układu równań. Przekształćmy (9.2.2) do postaci

$$-sv_{i-1,j} + (1+2s)v_{ij} - sv_{i+1,j} = v_{i,j-1} \quad (1 \leqslant i \leqslant n), \tag{9.2.3}$$

gdzie $s := k/h^2$. Jeśli, jak poprzednio, $V_j := (v_{1j}, v_{2j}, \ldots, v_{nj})$, to mamy układ

$$AV_j = V_{j-1}, \tag{9.2.4}$$

gdzie

$$A := \begin{bmatrix} 1+2s & -s & 0 & \ldots & 0 \\ -s & 1+2s & -s & \ldots & 0 \\ 0 & -s & 1+2s & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1+2s \end{bmatrix}$$

(wykorzystano tu zerowe warunki brzegowe). Stąd dla danego V_{j-1} mamy wyznaczyć V_j. Formalnie rzecz biorąc, $V_j = A^{-1}V_{j-1}$, czyli $V_j = A^{-j}V_0$. Wektor V_0 jest znany, bo jego składowymi są wartości początkowe $g(ih)$. Rozumując jak w podrozdz. 9.1, wnioskujemy, że metoda jest stabilna, jeśli $\rho(A^{-1}) < 1$. Ponieważ $A = I + sB$, gdzie macierz B jest określona w (9.1.6), więc z lem. 9.1.1 wynika, że wartościami własnymi macierzy A są liczby

$$\lambda_j := 1 + 2s(1 - \cos\theta_j), \qquad \text{gdzie} \quad \theta_j := \frac{j\pi}{n+1} \quad (1 \leqslant j \leqslant n).$$

Wszystkie one są większe od 1, czyli wartości własne λ_j^{-1} macierzy A^{-1} leżą w przedziale $(0,1)$ i zdefiniowana teraz *metoda niejawna* jest stabilna.

Algorytm

Algorytm, który realizuje opisaną metodę niejawną, korzysta z procedury `tri` rozwiązującej układ równań o macierzy trójprzekątniowej (zob. koniec podrozdz. 4.3). W tej procedurze danymi są oprócz n elementy przekątniowe macierzy $d_1, d_2, \ldots, d_n$, elementy nad przekątną i pod nią, odpowiednio $c_1, c_2, \ldots, c_{n-1}$ i $a_1, a_2, \ldots, a_{n-1}$, a także prawe strony równań $b_1, b_2, \ldots, b_n$. Procedura zapisuje rozwiązanie w $x_1, x_2, \ldots, x_n$. Jej parametrami są więc n, a, d, c, b, x. W rozważanym tu przypadku wszystkie elementy tablic a, d, c są stałe. Ponieważ procedura `tri` zmienia tablicę d, więc trzeba ją odnowić przed każdym wywołaniem.

```
input n, k, M
h ← 1/(n + 1); s ← k/h^2
for i = 1 to n do
    v_i ← g(ih)
end do
t ← 0
output 0, t, (v_1, v_2, ..., v_n)
for i = 1 to n − 1 do
    c_i ← −s; a_i ← −s
end do
for j = 1 to M do
    for i = 1 to n do
        d_i ← 1 + 2s
    end do
    call tri(n, a, d, c, v, v)
    t ← jk
output j, t, (v_1, v_2, ..., v_n)
end do
```

Metoda Cranka-Nicolson

Metody jawną i niejawną można w pewnym sensie połączyć i otrzymać ogólniejszą metodę zależną od parametru θ:

$$\frac{\theta}{h^2}(v_{i+1,j} - 2v_{ij} + v_{i-1,j}) + \frac{1-\theta}{h^2}(v_{i+1,j-1} - 2v_{i,j-1} + v_{i-1,j-1}) = \\ = \frac{1}{k}(v_{ij} - v_{i,j-1}). \tag{9.2.5}$$

Dla $\theta = 0$ jest to równość (9.1.3), a dla $\theta = 1$ – równość (9.2.2). Natomiast dla $\theta = \frac{1}{2}$ mamy metodę zwaną od nazwisk jej twórców, Johna Cranka i Phyllis Nicolson [*1947], *metodą Cranka-Nicolson*[1)].

Zbadajmy ją bardziej szczegółowo. Powyższą równość przekształcamy umieszczając, jak zwykle, nowe punkty (z drugim wskaźnikiem j) po lewej stronie, a wcześniejsze (z $j-1$) po prawej:

$$-sv_{i-1,j} + (2+2s)v_{ij} - sv_{i+1,j} = sv_{i-1,j-1} + (2-2s)v_{i,j-1} + sv_{i+1,j-1},$$

gdzie $s := k/h^2$. W symbolice macierzowej ta równość przybiera postać

$$(2I + sB)V_j = (2I - sB)V_{j-1},$$

[1)] W polskim przekładzie książki Dahlquista i Björcka [1974] wspomniałem o metodzie Cranka-Nicholsona, co było chyba podwójnym błędem. Kincaid i Cheney podając imiona autorów, chcieli zapewne podkreślić, że drugi z nich jest *panią* Nicolson (*przyp. tłum.*).

gdzie V_j jest wektorem o składowych v_{ij} dla $1 \leqslant i \leqslant n$, a B – macierzą (9.1.6). Rozumując w znany już sposób, stwierdzamy, że metoda jest stabilna, jeśli

$$\rho((2I + sB)^{-1}(2I - sB)) < 1,$$

tzn. jeśli wszystkie wartości własne μ_i macierzy B spełniają nierówność

$$|(2 + s\mu_i)^{-1}(2 - s\mu_i)| < 1.$$

Ponieważ $\mu_i = 2(1 - \cos\theta_i)$, więc $0 < \mu_i < 4$ i ta nierówność jest spełniona. Dlatego metoda Cranka-Nicolson jest stabilna dla dowolnych wartości ilorazu $s = k/h^2$.

Stabilność nie jest oczywiście jedynym kryterium wyboru parametrów h i k. Na ogół, im one są mniejsze, tym rozwiązanie numeryczne $v(x,t)$ lepiej naśladuje rozwiązanie dokładne $u(x,t)$ równania różniczkowego. Potrzebne jest nam twierdzenie zapewniające zbieżność, dla $h \to 0$ i $k \to 0$, funkcji $v(x,t)$ do $u(x,t)$. Zbadamy teraz tę kwestię dla metody jawnej (9.1.4).

Zbieżność

Niech e_{ij} oznacza błąd rozwiązania przybliżonego w punkcie siatki:

$$e_{ij} := u(x_i, t_j) - v(x_i, t_j) \equiv u_{ij} - v_{ij}.$$

W metodzie jawnej jest

$$v_{i,j+1} = s(v_{i-1,j} - 2v_{ij} + v_{i+1,j}) + v_{ij}.$$

Stąd i z definicji błędu e_{ij} wynika, że

$$\begin{aligned} u_{i,j+1} - e_{i,j+1} = {} & s(u_{i-1,j} - 2u_{ij} + u_{i+1,j}) + u_{ij} - \\ & - s(e_{i-1,j} - 2e_{ij} + e_{i+1,j}) - e_{ij}, \end{aligned}$$

czyli

$$\begin{aligned} e_{i,j+1} = {} & se_{i-1,j} + (1 - 2s)e_{ij} + se_{i+1,j} - \\ & - s(u_{i-1,j} - 2u_{ij} + u_{i+1,j}) + (u_{i,j+1} - u_{ij}). \end{aligned} \tag{9.2.6}$$

Aby uprościć wyrażenia zależne od funkcji u, odwołujemy się do znanych wzorów różniczkowania numerycznego. Wiadomo mianowicie, że

$$f''(x) = \frac{1}{h^2}[f(x+h) - 2f(x) + f(x-h)] - \frac{1}{12}h^2 f^{(4)}(\xi),$$

$$g'(t) = \frac{1}{k}[g(t+k) - g(t)] - \frac{1}{2}kg''(\tau).$$

Stosujemy te wzory w (9.2.6):

$$e_{i,j+1} = se_{i-1,j} + (1-2s)e_{ij} + se_{i+1,j} - \\ - s\Big[h^2 u_{xx}(x_i,t_j) + \frac{1}{12}h^4 u_{xxxx}(\xi_i,t_j)\Big] + ku_t(x_i,t_j) + \frac{1}{2}k^2 u_{tt}(x_i,\tau_j).$$

Ponieważ $sh^2 = k$ i $u_{xx} = u_t$, więc

$$e_{i,j+1} = se_{i-1,j} + (1-2s)e_{ij} + se_{i+1,j} - \\ - kh^2\Big[\frac{1}{12}u_{xxxx}(\xi_i,t_j) - \frac{1}{2}su_{tt}(x_i,\tau_j)\Big]. \tag{9.2.7}$$

Niech (x,t) należy do zbioru domkniętego

$$S := \{(x,t)\colon\ 0 \leqslant x \leqslant 1,\ 0 \leqslant t \leqslant T\}.$$

Zakładając, że funkcje u_{xxxx} i u_{tt} są ciągłe w S, definiujemy

$$M := \frac{1}{12}\max_S |u_{xxxx}(x,t)| + \frac{1}{2}s\max_S |u_{tt}(x,t)|.$$

Wprowadzamy też wektor błędu $E_j := (e_{1j}, e_{2j}, \ldots, e_{nj})$. Jego normą niech będzie $\|E_j\|_\infty := \max_{1\leqslant i\leqslant n}|e_{ij}|$. Na koniec załóżmy, że $1-2s \geqslant 0$. Wtedy z (9.2.7) wynika, że

$$|e_{i,j+1}| \leqslant s|e_{i-1,j}| + (1-2s)|e_{ij}| + s|e_{i+1,j}| + kh^2M \leqslant \|E_j\|_\infty + kh^2M.$$

Ponieważ prawa strona tej nierówności nie zależy od i, więc

$$\|E_{j+1}\|_\infty \leqslant \|E_j\|_\infty + kh^2M \leqslant \|E_{j-1}\|_\infty + 2kh^2M \leqslant \ldots \\ \ldots \leqslant \|E_0\|_\infty + (j+1)kh^2M.$$

Początkowe wartości rozwiązania przybliżonego $v(x,t)$ są dokładne, czyli $E_0 = 0$ i ostatecznie

$$\|E_j\|_\infty \leqslant jkh^2M.$$

Niech będzie $jk = t$, tak że $\|E_j\|_\infty$ jest maksymalnym błędem rozwiązania dla ustalonego t. Ponieważ $t \leqslant T$, więc

$$\|E_j\|_\infty \leqslant Th^2M = \mathcal{O}(h^2).$$

Jeśli zatem $s = k/h^2 \leqslant 1/2$, a funkcje u_{xxxx} i u_{tt} są ciągłe, to dla $h \to 0$ błąd rozwiązania dla dowolnego t dąży do 0 co najmniej tak szybko jak h^2.

Podsumowanie

Opisano trzy metody przybliżonego rozwiązywania zadania przewodnictwa cieplnego:

Metoda	Równanie macierzowe
Jawna	$V_j = (I - sB)V_{j-1}$
Niejawna	$(I + sB)V_j = V_{j-1}$
Cranka-Nicolson	$(2I + sB)V_j = (2I - sB)V_{j-1}$

Odwołują się one do mnożenia macierzy trójprzekątniowej przez wektor oraz rozwiązywania układu równań liniowych z taką macierzą.

ZADANIA 9.2

1. Wykazać, że dla $r > 0$ największa wartość własna macierzy $(I+rB)^{-1}(I-rB)$ jest równa $(1-q)/(1+q)$, gdzie $q := 4r\sin^2 \pi/(2n+2)$.
2. Udowodnić, że warunkiem stabilności metody (9.2.5) dla $0 \leqslant \theta < \frac{1}{2}$ jest nierówność $x \leqslant (2-4\theta)^{-1}$ i że dla $\frac{1}{2} \leqslant \theta \leqslant 1$ żadnych ograniczeń na s nie ma.
3. Zbadać zbieżność metody niejawnej (9.2.4). Wskazówka: W pewnym etapie analizy zbieżności wystąpi równanie wektorowe $(I+sB)E_j = E_{j-1} - C_j$, gdzie C_j jest wektorem o składowych

$$\tfrac{1}{2}k^2 u_{tt}(x_i, t_j + \varphi_j k) + \tfrac{1}{12}sh^4 u_{xxxx}(x_i + \psi_j h, t_j).$$

4. Uogólnić metody z tego podrozdziału na zadanie (9.1.2) (z niezerowymi warunkami brzegowymi).

ZADANIA KOMPUTEROWE 9.2

K1. Za pomocą metody niejawnej rozwiązać następujące zadanie przewodnictwa cieplnego w kwadracie jednostkowym:

$$u_{xx} = u_t, \quad u(x,0) = (x - x^2)e^x, \quad u(0,t) = u(1,t) = 0.$$

Sugerowane wartości parametrów: $n = 20$, $M = 50$ i $k = 0.05$.

K2. Napisać i sprawdzić program według metody Cranka-Nicolson dla zadania

$$\begin{aligned} &u_{xx} - \alpha u_t = f(x) && (0 \leqslant x \leqslant L, \ t \geqslant 0), \\ &u(x,0) = g(x) && (0 \leqslant x \leqslant L), \\ &u(0,t) = u(L,t) = 0 && (t \geqslant 0). \end{aligned}$$

K3. Porównać metody jawną, niejawną i Cranka-Nicolson na przykładzie z zad. 9.1.K1.

9.3. Zadania niezależne od czasu: różnice skończone

Typowym równaniem różniczkowym cząstkowym, w którym czas nie występuje, jest *równanie Laplace'a*

$$\nabla^2 u := u_{xx} + u_{yy} = 0.$$

u jest w nim funkcją zmiennych x i y, a $\nabla^2 u$ jej *laplasjanem*. W konkretnym zadaniu fizycznym prowadzącym do takiego równania trzeba określić również obszar (zbiór otwarty) Ω płaszczyzny xy, w którym szukamy rozwiązania i warunki brzegowe. Te ostatnie wyznaczają wartości funkcji u lub jej pochodnej normalnej na brzegu $\partial\Omega$ tego obszaru. Jego domknięcie oznaczamy $\bar{\Omega}$. Jest więc $\bar{\Omega} = \Omega \cup \partial\Omega$.

Zagadnienie Dirichleta

Zagadnienie Dirichleta występuje w badaniach różnych zjawisk fizycznych. W wersji dwuwymiarowej jest ono opisane równaniem Laplace'a w Ω i danymi wartościami funkcji u na brzegu tego obszaru:

$$\begin{aligned} &u_{xx} + u_{yy} = 0 \quad \text{w } \Omega, \\ &u(x,y) = g(x,y) \quad \text{na } \partial\Omega, \\ &u \text{ ciągła w } \bar{\Omega}. \end{aligned} \tag{9.3.1}$$

Jeśli obszar Ω spełnia pewne słabe warunki, a funkcja g jest ciągła, to można udowodnić, że zagadnienie Dirichleta ma jednoznaczne rozwiązanie.

Aby zilustrować pewne stosowane tu techniki numeryczne, rozważmy to zagadnienie w kwadracie

$$\Omega := \{(x,y)\colon\ 0 < x < 1,\ 0 < y < 1\}.$$

Dla tak prostego obszaru zagadnienie można rozwiązać analitycznie, stosując metodę rozdzielenia zmiennych i szeregi Fouriera. Na ogół jednak jest konieczne odwołanie się do metod numerycznych opisanych w tym i następnym podrozdziale. Bywają one użyteczne nawet wtedy, gdy można znaleźć dokładne rozwiązanie w postaci szeregu nieskończonego.

Różnice skończone

Jeden ze sposobów przybliżonego rozwiązywania zagadnienia (9.3.1) polega na zastąpieniu pochodnych wyrażeniami różnicowymi. Można użyć cytowanego wielokrotnie wzoru

$$f''(x) = \frac{1}{h^2}[f(x+h) - 2f(x) + f(x-h)] + \mathcal{O}(h^2).$$

Stosujemy go w punktach siatki pokrywającej równomiernie kwadrat $\bar{\Omega}$; dla $h = 1/(n+1)$ są to punkty

$$(x_i, y_j) := (ih, jh) \quad (0 \leqslant i, j \leqslant n+1).$$

Powyższy wzór daje różnicowy odpowiednik równania Laplace'a w punktach siatki:

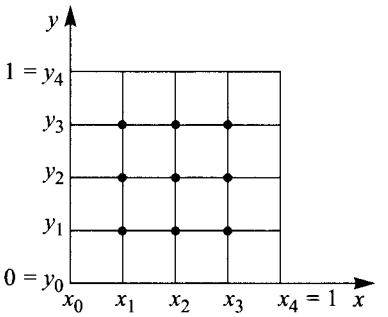
$$\frac{1}{h^2}(v_{i-1,j} - 2v_{ij} + v_{i+1,j}) + \frac{1}{h^2}(v_{i,j-1} - 2v_{ij} + v_{i,j+1}) = 0,$$

czyli

$$4v_{ij} - v_{i-1,j} - v_{i+1,j} - v_{i,j-1} - v_{i,j+1} = 0 \quad (1 \leqslant i, j \leqslant n), \tag{9.3.2}$$

gdzie v_{ij} ma być przybliżoną wartością $u(x_i, y_j)$. Jak w poprzednim podrozdziale litera v odróżnia rozwiązanie zadania *dyskretnego* od rozwiązania u pierwotnego równania różniczkowego.

Wartości v_{ij} są znane, gdy jeden ze wskaźników jest równy 0 lub $n+1$. Są to wartości danej funkcji g. Pozostałe v_{ij} występujące w równaniach (9.3.2) są nieznane. Mamy więc łącznie układ n^2 równań niejednorodnych. Każde z nich odpowiada jednemu z punktów wewnętrznych siatki zaznaczonych na rys. 9.4 kropkami •.



RYS. 9.4. Siatka na kwadracie jednostkowym ($n = 3$)

Układ (9.3.2) ma szczególną strukturę. Jeśli składowymi wektora v są nieznane wartości ustawione w naturalnej kolejności

$$v_{11}, v_{21}, \ldots, v_{n1}, v_{12}, v_{22}, \ldots, v_{n2}, \ldots, v_{1n}, v_{2n}, \ldots, v_{nn}$$

i tak samo są uporządkowane równania (według środkowych wartości v_{ij}), to układ ma postać $Av = b$, gdzie macierz A ma pewną szczególną postać. Staje się ona jasna, gdy użyjemy podziału macierzy na bloki. Dla $n = 3$ jest

$$A = \begin{bmatrix} T & -I & 0 \\ -I & T & -I \\ 0 & -I & T \end{bmatrix}, \quad \text{gdzie} \quad T := \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}.$$

Algorytm

Układ równań (9.3.2) ma macierz *rzadką*, gdyż każde równanie zawiera co najwyżej pięć niewiadomych. Można sprawdzić, że wśród n^4 elementów macierzy tylko $(5n-4)n$ różni się od 0 (ściślej, są one równe 4 lub -1). Taki układ można rozwiązywać jedną z metod iteracyjnych opisanych w podrozdz. 4.6, np. metodą Gaussa-Seidela. W algorytmie wszystkie wielkości v_{ij} pamiętamy w tablicy o $(n+2)^2$ elementach. Wartości brzegowe są dane: jeśli i lub j jest równe 0 lub $n+1$, to $v_{ij} = g(x_i, y_j)$. Z każdym punktem wewnętrznym siatki wiąże się równanie (9.3.2). Piszemy je teraz w postaci

$$v_{ij} = \frac{1}{4}(v_{i-1,j} + v_{i+1,j} + v_{i,j-1} + v_{i,j+1}),$$

bo z niego obliczamy nowe przybliżenie wartości v_{ij}.

Szkic algorytmu jest następujący:

1. Obliczenie wartości brzegowych.
2. Ustalenie początkowych przybliżeń wartości funkcji v w punktach wewnętrznych siatki.
3. M-krotne zastosowanie metody Gaussa-Seidela.

Pierwszy etap można zrealizować tak:

for $i = 0$ **to** $n+1$ **do**
    $v_{i0} \leftarrow g(x_i, 0)$; $v_{i,n+1} \leftarrow g(x_i, 1)$
    $v_{0i} \leftarrow g(0, y_i)$; $v_{n+1,i} \leftarrow g(1, y_i)$
end do

Trzeci etap wyraża się równie prosto:

for $k = 1$ **to** M **do**
    for $j = 1$ **to** n **do**
        for $i = 1$ **to** n **do**
            $v_{ij} \leftarrow (v_{i-1,j} + v_{i+1,j} + v_{i,j-1} + v_{i,j+1})/4$
        end do
    end do
end do

Obliczenia według takiego algorytmu wykonano, przyjmując, że na brzegu kwadratu

$$g(x, y) := 10^{-4} \sin(3\pi x) \sin(3\pi y).$$

Ta funkcja jest harmoniczna (tzn. spełnia równanie Laplace'a), co pozwala porównać rozwiązanie przybliżone v_{ij} z dokładnym. Dla $n = 18$ po 200 iteracjach (wykonanych w podwójnej precyzji) błąd spełnia nierówność

$$|v_{ij} - u(x_i, y_j)| < 0.345_{10}{-}7.$$

ZADANIA 9.3

1. Znaleźć dokładne rozwiązanie zagadnienia Dirichleta w kwadracie jednostkowym, przyjmując, że $u(x,0) = \sin 4\pi x$ i że na pozostałych bokach kwadratu $u = 0$. Zastosować rozdzielenie zmiennych i szeregi Fouriera.
2. Niech laplasjan *dyskretny* będzie określony wzorem

$$\delta f(x,y) := f(x+1,y) - 2f(x,y) + f(x-1,y) + \\ + f(x,y+1) - 2f(x,y) + f(x,y-1).$$

Sprawdzić, czy dla każdego punktu (x,y) istnieje takie (ξ,η), że

$$\delta f = \nabla^2 f(\xi,\eta).$$

Jaka byłaby wersja jednowymiarowa tego zadania?

ZADANIA KOMPUTEROWE 9.3

K1. Zaprogramować procedurę naszkicowaną w tekście dla zagadnienia Dirichleta w kwadracie. Parametrami powinny być liczba n, liczba iteracji i częstotliwość drukowania wyników. Sprawdzić program dla:
(**a**) $g(x,y) = 4xy(x^2 - y^2)$, (**b**) funkcji g podanej na końcu podrozdziału.
Porównać wyniki z rozwiązaniem dokładnym.

K2. Zmodyfikować program z poprzedniego zadania, zastępując metodę Gaussa-Seidela metodą SOR (podrozdz. 4.6).

9.4. Zadania niezależne od czasu: metody Galerkina

Metody Galerkina są szeroko stosowane w zadaniach, w których trzeba znaleźć pewną funkcję. Oczywiście równania różniczkowe i całkowe należą do tej kategorii. Zaczynamy od ogólnego sformułowania zasad stosujących się do dowolnego zadania liniowego. Następnie zilustrujemy je na przykładzie rozwiązywania numerycznego zagadnienia Dirichleta w prostokącie.

Metoda Galerkina

Niech zadaniem, o którym mowa, będzie wyznaczanie funkcji u z równania

$$Lu = f.$$

L jest tu operatorem liniowym, a f daną funkcją. *Metoda Galerkina* polega na tym, że wybieramy funkcje *bazowe* (*testowe*) $u_1, u_2, \ldots, u_n$ i próbujemy znaleźć rozwiązanie w postaci ich kombinacji liniowej:

$$u = \sum_{j=1}^{n} c_j u_j.$$

Ponieważ operator L jest liniowy, więc ma być

$$\sum_{j=1}^{n} c_j L u_j = f. \tag{9.4.1}$$

To równanie jest na ogół sprzeczne, gdyż f może nie należeć do przestrzeni rozpiętej na funkcjach Lu_j. Możemy jednak szukać przybliżonego rozwiązania tego równania. Kryteria, jakie to rozwiązanie ma spełniać, mogą być różne; każde z nich daje inne rozwiązanie przybliżone. Najbardziej naturalne jest żądanie, aby współczynniki c_j dawały minimum jakiejś normy

$$\Big\| \sum_{j=1}^{n} c_j L u_j - f \Big\|.$$

Jest to zatem zadanie aproksymacji optymalnej: szukamy dla f najbliższego elementu w podprzestrzeni rozpiętej na funkcjach Lu_j. Jest to stosunkowo łatwe w przestrzeni unitarnej, gdyż wtedy można stosować technikę rzutu ortogonalnego.

Aby znaleźć rozwiązanie przybliżone równania (9.4.1), można postąpić tak: wybieramy funkcjonały liniowe $\varphi_1, \varphi_2, \ldots, \varphi_n$ i żądamy, żeby było

$$\varphi_i\Big(\sum_{j=1}^{n} c_j L u_j - f\Big) = 0 \quad (1 \leqslant i \leqslant n),$$

czyli

$$\sum_{j=1}^{n} \varphi_i(Lu_j) c_j = \varphi_i(f) \quad (1 \leqslant i \leqslant n). \tag{9.4.2}$$

Jest to układ n równań liniowych z tyluż niewiadomymi c_j. W szczególnym przypadku, gdy funkcjonały są wartościami funkcji, na które działają:

$$\varphi(v) := v(x_i) \quad (1 \leqslant i \leqslant n),$$

wynika stąd *metoda kollokacji* opisana już w podrozdz. 8.10. Natomiast w klasycznej metodzie Galerkina przyjmujemy, że funkcje należą do przestrzeni Hilberta i że $\varphi(v) := \langle u_i, v \rangle$. Wtedy układ (9.4.2) jest następujący:

$$\sum_{j=1}^{n} c_j \langle u_i, Lu_j \rangle = \langle u_i, f \rangle \quad (1 \leqslant i \leqslant n).$$

Zagadnienie Dirichleta

Zastosujemy teraz metody Galerkina do zagadnienia Dirichleta w prostokącie

$$\Omega := \{(x,y)\colon\ |x| < 1,\ |y| < 2\}. \tag{9.4.3}$$

Szukamy funkcji u ciągłej w $\bar{\Omega}$ i spełniającej warunki

$$\nabla^2 u = 0 \quad \text{w } \Omega, \quad u(x,y) = x^2 + y^2 \quad \text{na } \partial\Omega. \tag{9.4.4}$$

Jest to zagadnienie typu $Lu = f$. Istotnie, wystarczy przyjąć, że

$$Lu := \begin{bmatrix} \nabla^2 u \\ u \mid \partial\Omega \end{bmatrix}, \quad f(x,y) := \begin{bmatrix} 0 \\ x^2 + y^2 \end{bmatrix},$$

gdzie $u \mid S$ oznacza zwężenie funkcji u do zbioru S. Jest to operacja liniowa w tym sensie, że $(\alpha u + \beta v) \mid S = \alpha(u \mid S) + \beta(v \mid S)$.

Aby zastosować metodę Galerkina, musimy wybrać odpowiedni układ funkcji bazowych. W rozważanym teraz zagadnieniu określamy je tak, aby wszystkie spełniały równanie różniczkowe Laplace'a, tj. żeby były harmoniczne. Szeroki repertuar takich funkcji wynika z twierdzenia, że część rzeczywista i część urojona dowolnej funkcji analitycznej zmiennej zespolonej $z = x + iy$ są harmoniczne. W szczególności możemy posłużyć się potęgami z^k:

$$\begin{aligned}
z^0 &= 1,\\
z^1 &= x + iy,\\
z^2 &= (x^2 - y^2) + (2xy)i,\\
z^3 &= (x^3 - 3xy^2) + (3x^2y - y^3)i,\\
z^4 &= (x^4 - 6x^2y^2 + y^4) + (4x^3y - 4xy^3)i,\\
z^5 &= (x^5 - 10x^3y^2 + 5xy^4) + (5x^4y - 10x^2y^3 + y^5)i,\\
z^6 &= (x^6 - 15x^4y^2 + 15x^2y^4 - y^6) + (6x^5y - 20x^3y^3 + 6xy^5)i, \ldots
\end{aligned}$$

Ze względu na symetrię naszego zadania względem obu osi współrzędnych wybieramy potęgi z wykładnikiem parzystym, bo ich części rzeczywiste i urojone też są symetryczne. Dla $n = 4$ funkcjami bazowymi mogą być następujące części rzeczywiste:

$$\begin{aligned}
u_1(x,y) &:= 1,\\
u_2(x,y) &:= x^2 - y^2,
\end{aligned}$$

$$u_3(x,y) := x^4 - 6x^2y^2 + y^4,$$
$$u_4(x,y) := x^6 - 15x^4y^2 + 15x^2y^4 - y^6.$$

Wtedy oczywiście każda funkcja $u = \sum_{j=1}^{4} c_j u_j$ jest rozwiązaniem równania różniczkowego $\nabla^2 u = 0$. Współczynniki c_j wybieramy tak, aby spełnić w przybliżeniu warunek brzegowy:

$$\sum_{j=1}^{4} c_j u_j(x,y) \approx x^2 + y^2 \quad \text{na } \partial\Omega. \tag{9.4.5}$$

Dzięki symetrii wystarczy brać pod uwagę tę część brzegu $\partial\Omega$, która leży w pierwszej ćwiartce.

Stosując metodę kollokacji, wybieramy na tym brzegu cztery punkty, np. $(0,2)$, $(1,0)$, $(1,1)$ i $(1,2)$. W tych punktach równości (9.4.5) mają być dokładnie spełnione:

$$\begin{bmatrix} 1 & -4 & 16 & -64 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & -4 & 0 \\ 1 & -3 & -7 & 117 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} 4 \\ 1 \\ 2 \\ 5 \end{bmatrix}.$$

Ten układ ma rozwiązanie

$$c = (1.8261,\ -0.7870,\ -0.04348,\ 0.004348).$$

Największe odchylenie sumy $\sum_{j=1}^{4} c_j u_j$ od funkcji $x^2 + y^2$ na brzegu prostokąta wynosi 0.074.

Zastosujmy teraz inny sposób rozwiązywania układu (9.4.5). Wybieramy mianowicie m punktów (x_i, y_i) na brzegu $\partial\Omega$ i szukamy minimum wyrażenia

$$\sum_{i=1}^{m} \Big[\sum_{j=1}^{4} c_j u_j(x_i, y_i) - (x_i^2 + y_i^2) \Big]^2. \tag{9.4.6}$$

Dla $m = 4$ mielibyśmy tu metodę kollokacji zastosowaną wyżej. Dla większych m to minimum wyznaczamy, przyrównując do 0 pochodne sumy (9.4.6) względem c_j. Daje to układ czterech równań liniowych. Rozwiązano go, wybierając 76 punktów (x_i, y_i) rozmieszczonych w równych odstępach na czwartej części brzegu: $(1, i/25)$ dla $0 \leqslant i \leqslant 50$ i $(i/25, 2)$ dla $0 \leqslant i \leqslant 24$. Dało to wynik

$$c = (1.8216,\ -0.7811,\ -0.04458,\ 0.004052).$$

Maksymalny błąd spełnienia warunku brzegowego wynosi 0.049.

Jeszcze inny wariant definicji współczynników c_j polega na tym, że szukamy minimum wyrażenia

$$\max_{1\leqslant i\leqslant m}\left|\sum_{j=1}^{4} c_j u_j(x_i,y_i) - (x_i^2+y_i^2)\right|,$$

czyli rozwiązujemy układ nadokreślony m równań z czterema niewiadomymi. Dla dużych m zadanie jest podobne do minimalizacji normy jednostajnej

$$\max_{(x,y)\in\partial\Omega}\left|\sum_{j=1}^{4} c_j u_j(x,y) - (x^2+y^2)\right|.$$

Oba te zadania można rozwiązać metodą Remeza (podrozdz. 6.9). Dla określonych wyżej 76 punktów daje to wynik

$$c = (1.8072,\ -0.7950,\ -0.0400,\ 0.003692)$$

i maksymalny błąd na brzegu 0.033. Jest on tylko nieco mniejszy niż w poprzednich przypadkach, a koszt obliczeń jest tu znacznie większy. Bardziej celowe jest zwiększenie liczby funkcji bazowych i zastosowanie prostszej metody.

Równanie Poissona

Często trzeba rozwiązać zagadnienie brzegowe z *równaniem Poissona*, mające np. postać

$$\nabla^2 w = f \quad \text{w } \Omega, \quad w = g \quad \text{na } \partial\Omega. \tag{9.4.7}$$

Jednym ze sposobów, jakie można tu zastosować, jest rozkład zagadnienia na dwa prostsze:

$$\begin{aligned} &\nabla^2 v = 0 \quad \text{w } \Omega, \quad v = g \quad \text{na } \partial\Omega,\\ &\nabla^2 u = f \quad \text{w } \Omega, \quad u = 0 \quad \text{na } \partial\Omega. \end{aligned} \tag{9.4.8}$$

Suma $u+v$ rozwiązań tych zagadnień jest szukaną funkcją w. Stosując metodę Galerkina do zagadnień składowych, korzystamy z tego, że w każdym z nich jedno równanie jest jednorodne. W szczególności, szukając funkcji u, wybieramy funkcje bazowe tak, aby znikały na brzegu $\partial\Omega$. Ich dowolna kombinacja liniowa $u := \sum_{j=1}^{n} c_j u_j$ ma tę samą własność. Trzeba więc tylko spróbować rozwiązać równanie $\nabla^2 u = f$, czyli równanie

$$\sum_{j=1}^{n} c_j \nabla^2 u_j = f.$$

Na ogół można je spełnić tylko w przybliżeniu.

Metoda Rayleigha-Ritza

Innym sposobem rozwiązywania drugiego zagadnienia (9.4.8) jest metoda Rayleigha-Ritza. Aby ją zdefiniować, wprowadzamy przestrzeń unitarną V, której elementami są funkcje u mające pochodne u_{xx} i u_{yy} ciągłe w Ω i znikające na brzegu tego obszaru. Zakładamy, że funkcja f jest ciągła w Ω. Rozwiązanie zadania powinno więc należeć do V. Iloczyn skalarny w tej przestrzeni definiujemy wzorem

$$\langle u, v\rangle := \iint_\Omega u(x,y)v(x,y)\,dx\,dy.$$

Twierdzenie 9.4.1. *Operator $-\nabla^2$ jest samosprzężony i dodatnio określony w przestrzeni V.*

Dowód. Samosprzężenie wyraża się równością $\langle -\nabla^2 u, v\rangle = \langle u, -\nabla^2 v\rangle$. Żeby ją sprawdzić, korzystamy z twierdzenia Greena, które orzeka, że dla dostatecznie „porządnych" funkcji i obszarów zachodzi równość

$$\iint_\Omega (P_x + Q_y)\,dx\,dy = \int_{\partial\Omega} (P\,dy - Q\,dx)$$

(zob. np. Leja [*1956, s. 338]). Ponieważ funkcje u, $v \in V$ znikają na brzegu $\partial\Omega$, więc z twierdzenia Greena wynika, że

$$\begin{aligned}\langle \nabla^2 u, v\rangle &= \iint_\Omega (u_{xx} + u_{yy})v\,dx\,dy = \\ &= \iint_\Omega [(u_x v)_x + (u_y v)_y - u_x v_x - u_y v_y]\,dx\,dy = \\ &= \int_{\partial\Omega} (u_x v\,dy - u_y v\,dx) - \iint_\Omega (u_x v_x + u_y v_y)\,dx\,dy = \\ &= -\iint_\Omega (u_x v_x + u_y v_y)\,dx\,dy.\end{aligned}$$

Ostatnia całka jest symetryczna względem u i v, czyli $\langle \nabla^2 u, v\rangle = \langle u, \nabla^2 v\rangle$, co należało wykazać.

Mamy również udowodnić, że jeśli $u \in V$ i $u \neq 0$, to $\langle -\nabla^2 u, u\rangle > 0$. Z wykonanych już obliczeń wynika, że

$$\langle -\nabla^2 u, u\rangle = \iint_\Omega [(u_x)^2 + (u_y)^2]\,dx\,dy.$$

Ta całka jest oczywiście nieujemna, a równa 0 tylko wtedy, gdy $u_x = u_y = 0$ w Ω. Tak więc u jest funkcją tylko zmiennej y i zarazem funkcją tylko zmiennej x. To jest możliwe jedynie wtedy, gdy ta funkcja jest stała. Ponieważ jednak $u \in V$, więc $u = 0$ na $\partial\Omega$, czyli także w Ω. ■

Na mocy tw. 9.4.1 iloczyn skalarny w przestrzeni V można też określić wzorem

$$[u, v] := \langle -\nabla^2 u, v \rangle = \iint_\Omega (u_x v_x + u_y v_y)\, dx\, dy.$$

Stąd wynika określenie normy: $\|u\| := [u, u]^{1/2}$. Można już teraz zdefiniować *metodę Rayleigha-Ritza*. Polega ona na tym, że dla wybranych funkcji bazowych $u_1, u_2, \ldots, u_n \in V$ wyznaczamy współczynniki $c_1, c_2, \ldots, c_n$ tak, aby w sensie normy $\|\cdot\|$ suma $\sum_{j=1}^n c_j u_j$ była najbliższa dokładnego rozwiązania u. W przestrzeni unitarnej tak jest, jeśli są spełnione warunki ortogonalności

$$u - \sum_{j=1}^n c_j u_j \perp u_i \quad (1 \leqslant i \leqslant n).$$

Wynika z nich od razu, że

$$\sum_{j=1}^n c_j [u_j, u_i] = [u, u_i] \quad (1 \leqslant i \leqslant n).$$

Ponieważ

$$[u, u_i] = \langle -\nabla^2 u, u_i \rangle = \langle -f, u_i \rangle \quad (1 \leqslant i \leqslant n),$$

więc końcowy układ równań normalnych jest następujący:

$$\sum_{j=1}^n c_j [u_j, u_i] = -\langle f, u_i \rangle \quad (1 \leqslant i \leqslant n).$$

Można stąd obliczyć współczynniki c_j, bo układ nie zawiera już nieznanej funkcji u.

Metoda elementu skończonego

Jeśli w metodzie Galerkina funkcje bazowe są wielomianami w podobszarach, na które podzielono Ω, to mówimy o *metodzie elementu skończonego*. Zilustrujemy ją zakładając, że ten obszar jest wielokątem i że podzielono go na trójkąty, czyli wykonano jego *triangulację*. W podrozdziale 6.10 wyjaśniono dokładnie, na czym ona polega i jak buduje się funkcję ciągłą, która w każdym trójkącie jest liniowa, czyli ma postać $ax + by + c$.

Teoria i zastosowania metody elementu skończonego rozwinęły się tak bardzo, że radzimy czytelnikom zapoznać się z obszerną literaturą tej dziedziny; zob. np. Becker, Carey i Oden [1981], Mitchell i Wait [1977], Oden [1972], Oden i Reddy [1976], Strang i Fix [1973], Vichnevetsky [1981], Wait i Mitchell [1986] oraz Zienkiewicz i Morgan [1983].

ZADANIA 9.4

1. Udowodnić, że jeśli $w \equiv u+iv$ jest funkcją analityczną zmiennej $z \equiv x+iy$, to funkcje u i v są harmoniczne. Wskazówka: Zastosować równania Cauchy'ego-Riemanna $\partial u/\partial x = \partial v/\partial y$, $\partial u/\partial y = -\partial v/\partial x$.
2. Niech będzie $z^n = u_n+iv_n$. Udowodnić, że u_n i v_n można obliczać rekurencyjnie ze wzorów $u_0 = 1$, $v_0 = 0$, $u_{n+1} = xu_n - yv_n$, $v_{n+1} = xv_n + yu_n$.
3. (cd.). Wykazać, że dla parzystego n funkcja u_n jest parzysta, a v_n nieparzysta względem każdej ze zmiennych x, y.
4. Udowodnić, że jeśli funkcja $u_0 + iv_0$ jest analityczna, to tę samą własność ma $u_n + iv_n$, gdzie u_n i v_n spełnia wzory rekurencyjne z zad. 2.
5. Wykazać, że nie istnieje funkcja u, która by spełniała równanie $u_{xy} = 1$ w kwadracie jednostkowym oraz warunki brzegowe $u(x,0) = x$, $u(0,y) = y$ i $u(1,y) = 1$.

ZADANIA KOMPUTEROWE 9.4

K1. Zaprogramować podane w tekście trzy metody wyznaczania współczynników c_j dla zagadnienia (9.4.4).

9.5. Równania rzędu pierwszego: charakterystyki

Jak równania różniczkowe zwyczajne, tak i tutaj równania wyższych rzędów można przekształcić na układy równań rzędu pierwszego. Poniższy przykład pokazuje, jak się to robi.

Układy równań rzędu pierwszego

PRZYKŁAD 9.5.1. Przekształcić równanie przewodnictwa cieplnego $u_{xx} = u_t$ na układ równań rzędu pierwszego.

Rozwiązanie. Wprowadzenie nowej funkcji $v := u_x$ daje układ

$$v_x - u_t = 0, \quad u_x - v = 0. \qquad \blacksquare$$

PRZYKŁAD 9.5.2. Wykazać, że równanie przewodnictwa cieplnego w trzech wymiarach

$$u_{xx} + u_{yy} + u_{zz} = u_t$$

można przekształcić na układ jak w przykł. 9.5.1.

Rozwiązanie. Wprowadzamy funkcje: $u^{(1)} := u$, $u^{(2)} := u_x$, $u^{(3)} := u_y$, $u^{(4)} := u_z$, co daje układ

$$u_x^{(2)} + u_y^{(3)} + u_z^{(4)} - u_t^{(1)} = 0,$$
$$u_x^{(1)} = u^{(2)}, \quad u_y^{(1)} = u^{(3)}, \quad u_z^{(1)} = u^{(4)}. \qquad \blacksquare$$

Przykład 9.5.3. Przekształcić jak w poprzednich przykładach ogólne równanie różniczkowe cząstkowe rzędu drugiego z dwiema zmiennymi niezależnymi:

$$F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}) = 0.$$

Rozwiązanie. Równanie jest równoważne układowi

$$F(x, y, u, v, w, v_x, v_y, w_y) = 0, \quad u_x = v, \quad u_y = w. \qquad \blacksquare$$

Charakterystyki

Zajmiemy się teraz pojęciem *charakterystyki równania różniczkowego cząstkowego*. W najprostszym przypadku jest to krzywa, na której rozwiązanie równania jest stałe. Rozważmy równanie rzędu pierwszego

$$u_x + cu_y = 0. \tag{9.5.1}$$

Na płaszczyźnie xy wykreślamy krzywą o równaniu $y = y(x)$. Wzdłuż tej krzywej wartości rozwiązania u są równe $u(x, y(x))$, a więc zależą tylko od x. Niech krzywa będzie taka, aby $u(x, y(x))$ było stałe. Wtedy

$$0 = \frac{d}{dx} u(x, y(x)) = u_x + u_y \frac{dy}{dx}.$$

Ta krzywa spełnia zatem równanie różniczkowe zwyczajne

$$\frac{dy}{dx} = -\frac{u_x}{u_y}. \tag{9.5.2}$$

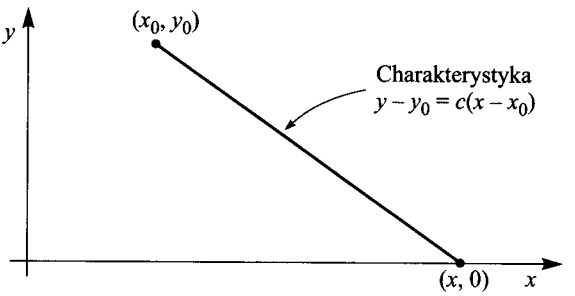
W przypadku (9.5.1) to równanie upraszcza się do postaci $dy/dx = c$ i jego rozwiązaniem jest każda prosta o równaniu $y = cx + d$. Przez każdy punkt (x_0, y_0) na płaszczyźnie przechodzi dokładnie jedna taka *charakterystyka* o równaniu

$$y - y_0 = c(x - x_0) \tag{9.5.3}$$

(rys. 9.5).

Jaki jest pożytek z takich charakterystyk? Przypuśćmy, że mamy rozwiązać nie tyle samo równanie (9.5.1), co zadanie z dodatkowym warunkiem, np.

$$u_x + cu_y = 0, \quad u(x, 0) = f(x), \tag{9.5.4}$$



RYS. 9.5. Najprostsza charakterystyka

gdzie f jest daną funkcją. Aby znaleźć rozwiązanie w punkcie (x_0, y_0), wyznaczamy charakterystykę przechodzącą przez ten punkt i przesuwamy się po niej do punktu $(x, 0)$. Wiemy, że w tym punkcie rozwiązanie u ma wartość $f(x)$. Jest to więc zarazem wartość $u(x_0, y_0)$. Ponieważ $(x, 0)$ leży na charakterystyce (9.5.3), więc $0 - y_0 = c(x - x_0)$, czyli $x = x_0 - c^{-1}y_0$ i $u(x_0, y_0) = f(x_0 - c^{-1}y_0)$. Odrzucamy zbędny wskaźnik i stwierdzamy, że rozwiązaniem zadania (9.5.4) jest funkcja $u(x, y) := f(x - c^{-1}y)$.

Rozważymy jeszcze jeden, nieco bardziej skomplikowany przykład.

PRZYKŁAD 9.5.4. Stosując charakterystyki, rozwiązać zadanie

$$u_x + yu_y = 0, \quad u(0, y) = f(y). \tag{9.5.5}$$

Rozwiązanie. Równanie różniczkowe zwyczajne (9.5.2) ma tu postać

$$\frac{dy}{dx} = -\frac{u_x}{u_y} = y.$$

Jego rozwiązaniem przechodzącym przez punkt (x_0, y_0) jest $y = y_0 e^{x - x_0}$. Ta charakterystyka przecina oś y, na której jest dany warunek $u(0, y) = f(y)$, dla $y = y_0 e^{-x_0}$. Stąd po uproszczeniu oznaczeń wynika rozwiązanie zadania (9.5.5):

$$u(x, y) = f(ye^{-x}). \qquad \blacksquare$$

Zachęcamy czytelników do rozwiązywania podobnych problemów, np. tych z zad. 4(a)–(c).

Ogólna teoria charakterystyk

Ogólna teoria charakterystyk dotyczy szerszej klasy równań różniczkowych. Nie wymagamy też, żeby rozwiązanie było stałe wzdłuż takiej krzywej; wystarczy, że spełnia tam ono pewne równanie różniczkowe zwyczajne.

Rozważmy najpierw *równanie quasi-liniowe* rzędu pierwszego

$$au_x + bu_y = c, \tag{9.5.6}$$

gdzie a, b i c mogą zależeć od x, y i u. Załóżmy, że wartość rozwiązania w punkcie (x_0, y_0) jest znana. Może ona wynikać z warunku brzegowego bądź z jakichś wcześniejszych obliczeń. Przekonamy się teraz, jak – rozwiązując pewne równania różniczkowe zwyczajne – można znaleźć wartości funkcji u w innych punktach. Rozważmy mianowicie układ trzech takich równań z warunkami początkowymi:

$$\begin{aligned}
\frac{dx}{ds} &= a, \qquad x(0) = x_0, \\
\frac{dy}{ds} &= b, \qquad y(0) = y_0, \\
\frac{du}{ds} &= c, \qquad u(0) = u_0 := u(x_0, y_0)
\end{aligned}$$

(układ jest autonomiczny, tj. zmienna s nie występuje explicite po prawej stronie żadnego z równań). Rozwiązując go numerycznie, można otrzymać tablicę wartości $x(s)$, $y(s)$ i $u(s)$. Łatwo sprawdzić, że to daje rozwiązanie równania cząstkowego (9.5.6) wzdłuż krzywej $(x(s), y(s))$, którą nazywamy *charakterystyką* tego równania. Istotnie,

$$\begin{aligned}
&a(x(s), y(s), u(s))u_x(x(s), y(s)) + b(x(s), y(s), u(s))u_y(x(s), y(s)) = \\
&\quad = \frac{dx(s)}{ds} u_x(x(s), y(s)) + \frac{dy(s)}{ds} u_y(x(s), y(s)) = \\
&\quad = \frac{du(x(s), y(s))}{ds} = c(x(s), y(s), u(s)).
\end{aligned}$$

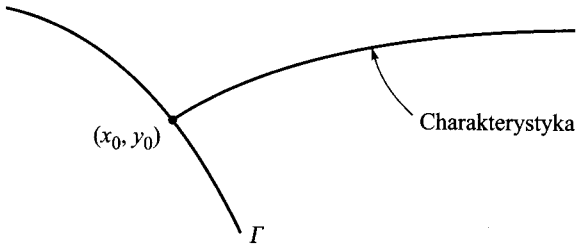
Jeśli znamy wartości rozwiązania $u(x, y)$ na pewnej krzywej Γ, która *nie* jest charakterystyką, to w zasadzie możemy – zaczynając od dowolnego punktu (x_0, y_0) na Γ i całkując powyższy układ równań różniczkowych zwyczajnych – otrzymać wartości tego rozwiązania wzdłuż charakterystyki; zob. rys. 9.6.

Przykład 9.5.5. Podać równania różniczkowe zwyczajne, określające charakterystyki dla równania cząstkowego

$$\sin(x^2 + y^2)u_x + (3x + y^2)u_y = e^{xy}.$$

Rozwiązanie. Układ składa się z równań

$$x' = \sin(x^2 + y^2), \quad y' = 3x + y^2, \quad u' = e^{xy}$$



RYS. 9.6. Charakterystyka krzywoliniowa

(znak $'$ oznacza tu i w następnych przykładach różniczkowanie względem s). Ponieważ w tym przykładzie a i b nie zależą od u, więc charakterystyki otrzymujemy, całkując tylko dwa początkowe równania. ■

Przykład 9.5.6. Niech dla równania różniczkowego cząstkowego

$$6u_x + xu_y = y$$

będzie znana wartość $u(3,3) = 4$ jego rozwiązania. Czy stosując metodę charakterystyk, można znaleźć wartość $u(15,21)$? Jeśli tak, to jaka ona jest?

Rozwiązanie. Mamy tu następujące zagadnienie początkowe:

$$x' = 6, \quad x(0) = 3, \quad y' = x, \quad y(0) = 3, \quad u' = y, \quad u(0) = 4.$$

Ma ono rozwiązanie

$$x = 6s + 3, \quad y = 3s^2 + 3s + 3, \quad u = s^3 + \tfrac{3}{2}s^2 + 3s + 4.$$

Charakterystyka przechodząca przez punkt $(3,3)$, przechodzi również przez punkt $(15,21)$; w tym punkcie, czyli dla $s = 2$, jest $u = 24$. ■

Przykład 9.5.7. Metodą charakterystyk rozwiązać zagadnienie brzegowe

$$6u_x + xu_y = y, \quad u(x,x) = 4.$$

Rozwiązanie. Równanie cząstkowe jest takie jak w poprzednim przykładzie, a więc i równania zwyczajne się nie zmieniają. Warunki początkowe mają teraz postać: $x(0) = r$, $y(0) = r$ i $u(0) = 4$. Rozwiązanie układu jest następujące:

$$x = 6s + r, \quad y = 3s^2 + rs + r, \quad u = s^3 + \tfrac{1}{2}rs^2 + rs + 4. \tag{9.5.7}$$

Dla danego punktu (x, y) wyznaczamy takie r i s, żeby były spełnione dwie początkowe równości (9.5.7). Jeśli założymy, że $r \leqslant 6$ (nie jest to konieczne; zob. zad. 7), to

$$r = 6 \pm \sqrt{(x-6)^2 + 12(x-y)}, \quad s = \tfrac{1}{6}(x-r).$$

Dla punktu (x, y), w którym chcemy znaleźć rozwiązanie, obliczamy stąd r i s. Ostatnia równość (9.5.7) daje $u(x, y)$. Jeśli np. $(x, y) = (15, 21)$, to możemy przyjąć, że $r = 3$, $s = 2$ i $u = 24$, jak w przykł. 9.5.6. Zauważmy jednak, że ta procedura jest skuteczna tylko wtedy, gdy w wyrażeniu dla r suma pod pierwiastkiem jest nieujemna, tzn. gdy $y \leqslant \frac{1}{12}x^2 + 3$. ■

Przykład 9.5.8. Metodą charakterystyk rozwiązać zagadnienie brzegowe

$$xu_x + yuu_y = xy, \quad u(x, 3/x) = 6.$$

Rozwiązanie. Charakterystyki są opisane warunkami

$$x' = x, \quad x(0) = x_0, \quad y' = uy, \quad y(0) = 3/x_0, \quad u' = xy, \quad u(0) = 6.$$

Ponieważ

$$(xy)' = x'y + xy' = xy + xuy = u'(1+u) = \tfrac{1}{2}[(1+u)^2]',$$

więc

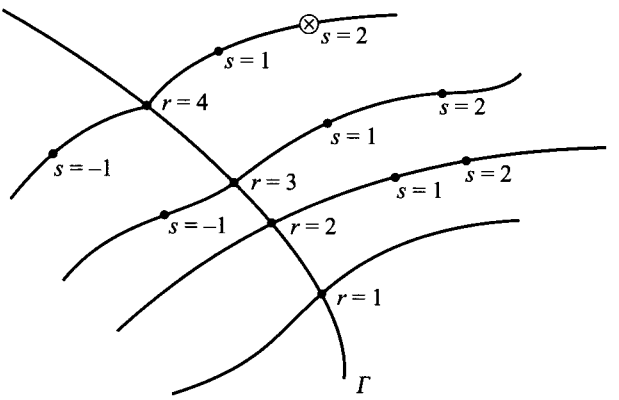
$$xy = \tfrac{1}{2}u^2 + u + c. \tag{9.5.8}$$

Dla $xy = 3$ jest $u = 6$, więc $c = -21$. Rozwiązując równanie (9.5.8) względem u, otrzymujemy

$$u(x, y) = -1 + \sqrt{2xy + 43},$$

oczywiście tylko dla $2xy + 43 \geqslant 0$. ■

Niech w przykładach jak wyżej r będzie parametrem punktu na krzywej Γ, na której są dane wartości funkcji u, a s – parametrem punktu na dowolnej charakterystyce. Wartość $s = 0$ odpowiada punktowi przecięcia charakterystyki z krzywą Γ. Rysunek 9.7 pokazuje tę krzywą i kilka charakterystyk. W punkcie oznaczonym $\otimes$ jest $r = 4$ i $s = 2$. Trzeba podkreślić, że nie każdemu punktowi płaszczyzny można przypisać współrzędne (r, s). Może się też zdarzyć, że przez jeden punkt przechodzą dwie charakterystyki;



RYS. 9.7. Krzywa Γ i charakterystyki

wtedy te współrzędne nie są określone jednoznacznie. Jeśli jednak Γ nie jest charakterystyką, a funkcje a, b, c są gładkie, to każdemu punktowi w pobliżu Γ odpowiada tylko jedna para (r, s).

Interpretując r i s jak wyżej, możemy tak opisać czynności wykonywane dla danego równania cząstkowego (9.5.6): Rozwiązujemy układ równań różniczkowych zwyczajnych z warunkami początkowymi

$$\frac{\partial x}{\partial s} = a, \quad x(r,0) = f(r),$$
$$\frac{\partial y}{\partial s} = b, \quad y(r,0) = g(r),$$
$$\frac{\partial u}{\partial s} = c, \quad u(r,0) = h(r).$$

Przyjęto tu, że na krzywej Γ określonej równaniami $x = f(r)$, $y = g(r)$ rozwiązanie u równania cząstkowego ma mieć dane wartości $h(r)$. Rozwiązujemy ten układ i otrzymujemy funkcje

$$x = x(r,s), \quad y = y(r,s), \quad u = u(r,s),$$

które w sposób parametryczny opisują rozwiązanie u danego zagadnienia brzegowego.

ZADANIA 9.5

1. Dla równania różniczkowego cząstkowego z przykł. 9.5.4 znaleźć $u(17,3)$, wiedząc, że $u(18,3e) = k\pi/2$.
2. Dla równania różniczkowego $au_x + bu_y = 0$, w którym a i b zależą tylko od x, znaleźć równanie krzywej przechodzącej przez punkt (x_0, y_0), na której rozwiązanie u jest stałe.

3. Sprawdzić, że funkcja u otrzymana w przykł. 9.5.8 jest rozwiązaniem postawionego tam zagadnienia.
4. Metodą charakterystyk rozwiązać następujące zagadnienia brzegowe:
   (a) $u_x + xu_y = 0$, $u(0,y) = f(y)$.
   (b) $u_x + 2uu_y = 0$, $u(0,y) = f(y)$.
   (c) $xu_x + 2yu_y = 0$, $u(1,y) = f(y)$.
   (d) $u_x + u_y = u^2$, $u(x,y) = y$ na prostej $x + y = 0$.
   (e) $u_x + 2u_y = u$, $y(x,2x) = 1$.
   (f) $uu_x + u_y = 1$, $u = r$ na krzywej $x = r^2$, $y = 2r$.
   (g) $u_x + 2u_y = y$, $u = r^2$ na okręgu $x = \cos r$, $y = \sin r$.
5. Odwołując się do ogólnej teorii charakterystyk, wyjaśnić, dlaczego rozwiązania równań (9.5.1) i (9.5.5) są na takich krzywych stałe.
6. Rozwiązać równanie różniczkowe z przykł. 9.5.6 z warunkiem brzegowym $u = e^x \sin y$ na krzywej $y = x^3$. Znaleźć $u(7,5)$, wiedząc, że punkt $(7,5)$ leży na charakterystyce przechodzącej przez $(1,1)$.
7. Pokazać, że rozwiązanie zagadnienia z przykł. 9.5.7 nie jest jednoznaczne.

9.6. Równania quasi-liniowe rzędu drugiego: charakterystyki

Tematem tego podrozdziału są *równania quasi-liniowe rzędu drugiego*

$$au_{xx} + bu_{xy} + cu_{yy} + e = 0, \tag{9.6.1}$$

w których a, b, c i e mogą zależeć od x, y, u, u_x i u_y. Jak poprzednio, zbadamy własności rozwiązania wzdłuż pewnych krzywych na płaszczyźnie xy.

Charakterystyki

Niech C będzie krzywą określoną wzorami

$$x = x(s), \quad y = y(s) \quad (s \in \mathbb{R}).$$

Przyjmiemy upraszczające oznaczenia $p = u_x$, $q = u_y$. Te wielkości zależą pośrednio od s. Zróżniczkujmy je względem tego parametru:

$$\frac{dp}{ds} = \frac{\partial p}{\partial x}\frac{dx}{ds} + \frac{\partial p}{\partial y}\frac{dy}{ds} = u_{xx}x' + u_{xy}y',$$
$$\frac{dq}{ds} = \frac{\partial q}{\partial x}\frac{dx}{ds} + \frac{\partial q}{\partial y}\frac{dy}{ds} = u_{xy}x' + u_{yy}y'.$$

Z pierwszego równania wyznaczamy u_{xx}:

$$u_{xx} = \left(\frac{dp}{ds} - u_{xy}y'\right)/x' = \frac{dp}{ds}\frac{ds}{dx} - u_{xy}\frac{dy}{ds}\frac{ds}{dx} = \frac{dp}{dx} - u_{xy}\frac{dy}{dx},$$

a z drugiego u_{yy}:

$$u_{yy} = \left(\frac{dq}{ds} - u_{xy}x'\right)/y' = \frac{dq}{ds}\frac{ds}{dy} - u_{xy}\frac{dx}{ds}\frac{ds}{dy} = \frac{dq}{dy} - u_{xy}\frac{dx}{dy}.$$

Te dwa związki są spełnione na krzywej C. Otrzymane wyrażenia dla u_{xx} i u_{yy} podstawiamy do równania różniczkowego (9.6.1):

$$a\left(\frac{dp}{dx} - u_{xy}\frac{dy}{dx}\right) + bu_{xy} + c\left(\frac{dq}{dy} - u_{xy}\frac{dx}{dy}\right) + e = 0.$$

To równanie mnożymy stronami przez dy/dx:

$$a\left[\frac{dp}{dx}\frac{dy}{dx} - u_{xy}\left(\frac{dy}{dx}\right)^2\right] + bu_{xy}\frac{dy}{dx} + c\left(\frac{dq}{dx} - u_{xy}\right) + e\frac{dy}{dx} = 0,$$

czyli

$$-u_{xy}\left[a\left(\frac{dy}{dx}\right)^2 - b\frac{dy}{dx} + c\right] + a\frac{dp}{dx}\frac{dy}{dx} + c\frac{dq}{dx} + e\frac{dy}{dx} = 0. \tag{9.6.2}$$

Krzywa C dotąd była dowolna. Określamy ją teraz tak, aby na niej znikał składnik z u_{xy}:

$$a\left(\frac{dy}{dx}\right)^2 - b\frac{dy}{dx} + c = 0. \tag{9.6.3}$$

Krzywą o takim równaniu nazywamy *charakterystyką* równania (9.6.1).

Klasyfikacja równań różniczkowych cząstkowych

Ponieważ równanie (9.6.3) jest kwadratowe względem dy/dx, więc rodzaj charakterystyk zależy od wyróżnika $\Delta := b^2 - 4ac$. Jego znak określa zarazem typ równania różniczkowego (9.6.1). Jeśli dla pewnych x, y, u jest $\Delta > 0$, to mówimy, że jest ono *hiperboliczne*. Dla $\Delta = 0$ równanie jest *parabo*liczne, a dla $\Delta < 0$ – *eliptyczne*. Ściślej, ta klasyfikacja może się zmieniać od punktu do punktu na płaszczyźnie xy, a nawet może zależeć od u, gdy a, b i c zależą od rozwiązania u lub jego pierwszych pochodnych. W przypadku *liniowym*, tj. gdy te współczynniki zależą tylko od x i y, klasyfikacja jest prostsza.

Klasyfikację równań zilustrujemy na przykładzie trzech najprostszych i dobrze znanych równań:

Nazwa	Równanie	a	b	c	Δ	Typ
Przewodnictwo cieplne	$u_{xx} - u_y = 0$	1	0	0	0	paraboliczny
Ruch falowy	$u_{xx} - u_{yy} = 0$	1	0	-1	4	hiperboliczny
Laplace	$u_{xx} + u_{yy} = 0$	1	0	1	-4	eliptyczny

Przykład 9.6.1. Określić typ równania różniczkowego

$$(x+y)u_{xx} + (1+x^2)u_{yy} = 0.$$

Rozwiązanie. Wyróżnikiem tego równania jest $\Delta(x,y) = -4(x+y)(1+x^2)$. Równanie jest więc eliptyczne dla $x+y>0$, paraboliczne na prostej $x+y=0$ i hiperboliczne dla $x+y<0$. ■

Przykład 9.6.2. Znaleźć charakterystyki równania

$$yu_{xx} + (x+y^2)u_{xy} + xyu_{yy} = 0. \tag{9.6.4}$$

Rozwiązanie. Równanie charakterystyk (9.6.3) ma tu postać

$$y\left(\frac{dy}{dx}\right)^2 - (x+y^2)\frac{dy}{dx} + xy = 0.$$

Wyróżnik jest równy $\Delta = (x+y^2)^2 - 4xy^2 = (x-y^2)^2$. Poza parabolą $x=y^2$ jest on dodatni, czyli równanie (9.6.4) jest hiperboliczne. Z równania kwadratowego wynika, że charakterystyki spełniają jedno z dwóch równań różniczkowych zwyczajnych

$$\frac{dy}{dx} = \frac{x+y^2 \pm |x-y^2|}{2y}.$$

Jeśli $x>y^2$, to dla pierwszego równania, odpowiadającego znakowi $+$, czyli

$$\frac{dy}{dx} = \frac{x}{y},$$

rozwiązaniami jest rodzina hiperbol $y^2 - x^2 = \alpha$. Dla drugiego równania (ze znakiem $-$), czyli

$$\frac{dy}{dx} = y,$$

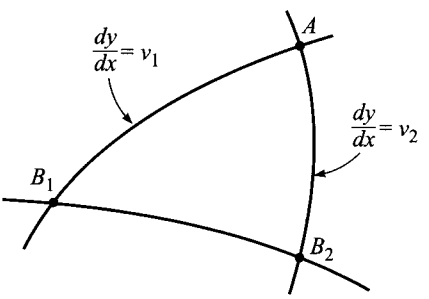
rozwiązaniami jest rodzina krzywych wykładniczych $y = \beta e^x$. Te same rozwiązania, tylko przestawione, są poprawne dla $x<y^2$. Na paraboli $x=y^2$ równanie (9.6.4) jest paraboliczne i istnieje tylko jedna rodzina charakterystyk o równaniu $dy/dx = y$. ■

Algorytm dla równań hiperbolicznych

Powróćmy do równania (9.6.2), podtrzymując założenie, że krzywa C jest charakterystyką, czyli dy/dx spełnia warunek (9.6.3). To równanie upraszcza się wtedy do postaci

$$a\frac{dp}{dx}\frac{dy}{dx} + c\frac{dq}{dx} + e\frac{dy}{dx} = 0. \tag{9.6.5}$$

Zobaczymy teraz, jak można to wykorzystać, szukając numerycznie rozwiązania równania cząstkowego (9.6.1), jeśli jest ono hiperboliczne. Wtedy przez dany punkt A przechodzą dwie charakterystyki (rys. 9.8).



RYS. 9.8. Charakterystyki AB_1 i AB_2

Oznaczmy symbolami v_1 i v_2 rozwiązania równania kwadratowego (9.6.3):

$$\frac{dy}{dx} = v_1 := \frac{b+\sqrt{\Delta}}{2a}, \quad \frac{dy}{dx} = v_2 := \frac{b-\sqrt{\Delta}}{2a}.$$

Oczywiste związki $v_1 + v_2 = b/a$ i $v_1 v_2 = c/a$ pozwalają jeszcze bardziej uprościć równanie (9.6.5):

$$\frac{dp}{dx} + v_1\frac{dq}{dx} = -\frac{e}{a}, \quad \text{gdy} \quad \frac{dy}{dx} = v_2, \tag{9.6.6}$$

$$\frac{dp}{dx} + v_2\frac{dq}{dx} = -\frac{e}{a}, \quad \text{gdy} \quad \frac{dy}{dx} = v_1. \tag{9.6.7}$$

Te dwa równania można rozwiązywać metodą różnic skończonych. Przyjmując, że punkty A, B_1 i B_2 (rys. 9.8) leżą blisko siebie, otrzymujemy przybliżoną (dyskretną) postać równań charakterystyk:

$$\frac{y(A)-y(B_1)}{x(A)-x(B_1)} = \frac{1}{2}[v_1(A) + v_1(B_1)], \tag{9.6.8}$$

$$\frac{y(A)-y(B_2)}{x(A)-x(B_2)} = \frac{1}{2}[v_2(A) + v_2(B_2)], \tag{9.6.9}$$

a także równań, które otrzymaliśmy z (9.6.5):

$$\frac{p(A)-p(B_2)}{x(A)-x(B_2)}+\frac{1}{2}[v_1(A)+v_1(B_2)]\frac{q(A)-q(B_2)}{x(A)-x(B_2)}= \tag{9.6.10}$$
$$=-\frac{1}{2}[(e/a)(A)+(e/a)(B_2)],$$
$$\frac{p(A)-p(B_1)}{x(A)-x(B_1)}+\frac{1}{2}[v_2(A)+v_2(B_1)]\frac{q(A)-q(B_1)}{x(A)-x(B_1)}= \tag{9.6.11}$$
$$=-\frac{1}{2}[(e/a)(A)+(e/a)(B_1)].$$

Odpowiednikiem związku

$$du=u_x\,dx+u_y\,dy=p\,dx+q\,dy$$

jest wzór, z którego możemy obliczać przybliżoną wartość $u(A)$:

$$u(A)=u(B_1)+\frac{1}{2}[p(A)+p(B_1)][x(A)-x(B_1)]+ \tag{9.6.12}$$
$$+\frac{1}{2}[q(A)+q(B_1)][y(A)-y(B_1)].$$

Zauważmy, że w (9.6.8)-(9.6.12) stosujemy średnie wartości funkcji na poszczególnych krzywych.

Możemy już naszkicować sposób wykorzystania otrzymanych równań. Jeśli znamy x, y, u, p i q w punktach B_1 i B_2, to z tych równań można wyznaczyć takie same wielkości w A. Jest to o tyle trudne, że równania są nieliniowe i trzeba rozwiązywać je iteracyjnie. Obliczenia przebiegają tak:

1. Wybieramy próbne wartości $x(A)$, $y(A)$, $u(A)$, $p(A)$ i $q(A)$; można je np. obliczyć w sposób podany niżej.
2. Obliczamy $v_1(A)$, $v_2(A)$ i $(e/a)(A)$.
3. Za pomocą (9.6.8) i (9.6.9) obliczamy nowe wartości $x(A)$ i $y(A)$, a za pomocą (9.6.10) i (9.6.11) – nowe wartości $p(A)$ i $q(A)$ (odpowiednie równania po przekształceniu są w obu przypadkach liniowe). Na koniec ze wzoru (9.6.12) obliczamy nowe $u(A)$. Wracamy do kroku **2**, jeśli te nowe wielkości różnią się istotnie od poprzednich.

Próbne wartości ustalane w kroku **1** można łatwo znaleźć, używając równań (9.6.8)-(9.6.11) w mniej dokładnej wersji. Polega ona na tym, że średnie wartości funkcji zmieniamy na wartości w B_1 lub B_2, co daje łatwo rozwiązywalne równania liniowe:

$$x(A)=\frac{y(B_2)-y(B_1)+x(B_1)v_1(B_1)-x(B_2)v_2(B_2)}{v_1(B_1)-v_2(B_2)},$$

$$y(A) = y(B_1) + v_1(B_1)[x(A) - x(B_1)],$$
$$R := p(B_2) - p(B_1) + v_1(B_2)q(B_2) - v_2(B_1)q(B_1),$$
$$S := (e/a)(B_2)[x(A) - x(B_2)] - (e/a)(B_1)[x(A) - x(B_1)],$$
$$q(A) = \frac{R - S}{v_1(B_2) - v_2(B_1)},$$
$$p(A) = p(B_2) - (e/a)(B_2)[x(A) - x(B_2)] - v_1(B_2)[q(A) - q(B_2)].$$

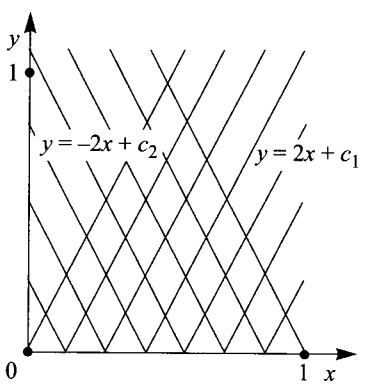
Zauważmy jeszcze, że jeśli równanie (9.6.1) jest liniowe, czyli a, b, c i e, a zatem również Δ, v_1 i v_2, zależą tylko od x i y, to $x(A)$ i $y(A)$ wyznaczamy z układu (9.6.8), (9.6.9), a $p(A)$ i $q(A)$ – z układu (9.6.10), (9.6.11).

Przykład 9.6.3. Naszkicować program rozwiązywania metodą charakterystyk zagadnienia brzegowego

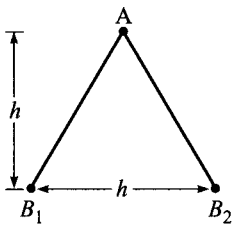
$$u_{xx} - 4u_{yy} - u_y = 0, \quad u(x,0) = f(x), \quad u_y(x,0) = g(x) \quad (0 \leqslant x \leqslant 1).$$

Rozwiązanie. W przedziale $[0,1]$ wybieramy n punktów równoodległych x_j i obliczamy przybliżone rozwiązanie w punktach przecięcia charakterystyk przechodzących przez x_j. Ponieważ $\Delta = 16$, więc $v_1 = 2$ i $v_2 = -2$, czyli charakterystyki spełniają równania różniczkowe $dy/dx = \pm 2$. Są to więc linie proste (rys. 9.9, gdzie $n = 8$). Równania (9.6.6) i (9.6.7) są następujące:

$$\frac{dp}{dx} + 2\frac{dq}{dx} = q, \quad \text{gdy} \quad \frac{dy}{dx} = -2,$$
$$\frac{dp}{dx} - 2\frac{dq}{dx} = q, \quad \text{gdy} \quad \frac{dy}{dx} = 2.$$



RYS. 9.9. Charakterystyki w przykładzie 9.6.3



RYS. 9.10. Układ punktów A, B_1 i B_2 w przykładzie 9.6.3

Niech $h = 1/(n-1)$ oznacza odstęp między punktami x_j w przedziale $[0,1]$. Jeśli przyjmiemy, że na rys. 9.10 jest $B_1 = (x,y)$, to $B_2 = (x+h, y)$ i $A = (x + h/2, y + h)$. Równania (9.6.8) i (9.6.9) są tu zbędne. Ponieważ $e/a = -q$, więc równania (9.6.10) i (9.6.11) upraszczają się do postaci

$$p(A) + (2 + h/4)q(A) = p(B_2) + (2 - h/4)q(B_2),$$
$$p(A) - (2 + h/4)q(A) = p(B_1) - (2 - h/4)q(B_1).$$

Stąd wynika, że

$$p(A) = \frac{1}{2}[p(B_1) + p(B_2)] + (1 - h/8)[q(B_2) - q(B_1)],$$
$$q(A) = \frac{p(B_2) - p(B_1) + (2 - h/4)[q(B_2) + q(B_1)]}{4 + h/2}.$$

Upraszcza się również wzór (9.6.12):

$$u(A) = u(B_1) + \frac{1}{4}h[p(A) + p(B_1)] + \frac{1}{2}h[q(A) + q(B_1)].$$

Z tych wzorów wynika następujący algorytm:

input n
$h \leftarrow 1/(n-1)$
for $j = 1$ **to** n **do**
    $x_j \leftarrow (j-1)h$
    $u_j \leftarrow f(x_j)$; $p_j \leftarrow f'(x_j)$; $q_j \leftarrow g(x_j)$
    output x_j, 0, p_j, q_j, u_j
end do
for $i = 1$ **to** n **do**
    for $j = 1$ **to** $n - i$ **do**
        $x_j \leftarrow x_j + h/2$
        $\tilde{p} \leftarrow (p_j + p_{j+1})/2 + (1 - h/8)(q_{j+1} - q_j)$
        $\tilde{q} \leftarrow [p_{j+1} - p_j + (2 - h/4)(q_{j+1} + q_j)]/(4 + h/2)$
        $u_j \leftarrow u_j + h(\tilde{p} + p_j)/4 + h(\tilde{q} + q_j)/2$
        $p_j \leftarrow \tilde{p}$; $q_j \leftarrow \tilde{q}$

output x_j, ih, p_j, q_j, u_j
end do
end do

Zauważmy, że dane zagadnienie brzegowe możemy rozwiązać tylko w trójkącie o wierzchołkach $(0,0)$, $(1,0)$ i $(0.5,1)$. ■

Inne podejście do charakterystyk

Sprawdzimy najpierw, czy jakaś zmiana zmiennych, z (x,y) na (ξ,η), może uprościć równanie różniczkowe (9.6.1). Jeśli

$$\xi = \xi(x,y), \quad \eta = \eta(x,y), \tag{9.6.13}$$

to żeby tę zmianę wprowadzić, obliczamy najpierw

$$\begin{aligned}
u_x &= u_\xi \xi_x + u_\eta \eta_x,\\
u_y &= u_\xi \xi_y + u_\eta \eta_y,\\
u_{xx} &= u_\xi \xi_{xx} + u_{\xi\xi}\xi_x^2 + u_{\xi\eta}\eta_x\xi_x + u_\eta \eta_{xx} + u_{\eta\xi}\xi_x\eta_x + u_{\eta\eta}\eta_x^2,\\
u_{xy} &= u_\xi \xi_{xy} + u_{\xi\xi}\xi_y\xi_x + u_{\xi\eta}\eta_y\xi_x + u_\eta \eta_{xy} + u_{\eta\xi}\xi_y\eta_x + u_{\eta\eta}\eta_y\eta_x,\\
u_{yy} &= u_\xi \xi_{yy} + u_{\xi\xi}\xi_y^2 + u_{\xi\eta}\eta_y\xi_y + u_\eta \eta_{yy} + u_{\eta\xi}\xi_y\eta_y + u_{\eta\eta}\eta_y^2.
\end{aligned}$$

Podstawiamy te wyrażenia do (9.6.1):

$$\begin{aligned}
&u_{\xi\xi}(a\xi_x^2 + b\xi_x\xi_y + c\xi_y^2) + u_{\xi\eta}[2a\xi_x\eta_x + b(\xi_x\eta_y + \xi_y\eta_x) + 2c\xi_y\eta_y] +\\
&\quad + u_{\eta\eta}(a\eta_x^2 + b\eta_x\eta_y + c\eta_y^2) + f = 0.
\end{aligned} \tag{9.6.14}$$

Podobnie trzeba przekształcić a, b, c i e, ale nie wnosi to nowych składników z drugimi pochodnymi cząstkowymi. Postać f jest tematem zad. 7. Zajmujemy się teraz równaniami hiperbolicznymi, więc równanie kwadratowe

$$a\lambda^2 + b\lambda + c = 0$$

ma z założenia dodatni wyróżnik $\Delta = b^2 - 4ac$ i dwa różne pierwiastki $(-b \pm \sqrt{\Delta})/(2a)$. Jeśli zmienne przekształcamy tak, żeby było

$$\frac{\xi_x}{\xi_y} = \frac{-b+\sqrt{\Delta}}{2a}, \qquad \frac{\eta_x}{\eta_y} = \frac{-b-\sqrt{\Delta}}{2a}, \tag{9.6.15}$$

to równanie różniczkowe (9.6.14) upraszcza się do postaci

$$u_{\xi\eta}[2a\xi_x\eta_x + b(\xi_x\eta_y + \xi_y\eta_x) + 2c\xi_y\eta_y] + f = 0.$$

Jest to *postać kanoniczna* równania hiperbolicznego. Krzywe, na których

$$\xi(x,y) = \text{const}, \qquad \eta(x,y) = \text{const},$$

są jego charakterystykami, jest bowiem niemal oczywiste, że wtedy zachodzi (9.6.3).

Przykład 9.6.4. Znaleźć obszar, w którym równanie

$$u_{xx} - yu_{yy} = 0$$

jest hiperboliczne i wyznaczyć przekształcenie zmiennych, prowadzące do postaci kanonicznej.

Rozwiązanie. Wyróżnik równania jest równy $4y$, czyli jest ono hiperboliczne w górnej półpłaszczyźnie. Równania (9.6.15), czyli

$$\xi_x/\xi_y = y^{1/2}, \qquad \eta_x/\eta_y = -y^{1/2},$$

są spełnione, jeśli

$$\xi = x + 2y^{1/2}, \qquad \eta = x - 2y^{1/2}.$$

Po przejściu do nowych zmiennych dane równanie uzyskuje postać

$$4u_{\xi\eta} + \frac{2}{\xi - \eta}(u_\xi - u_\eta) = 0. \qquad \blacksquare$$

Zadania 9.6

1. Sprawdzić typ każdego z poniższych równań cząstkowych:
   (a) $yu_{xx} + xu_{xy} + u_{yy} + u_x + u = 0$
   (b) $xyu_{xy} + e^x u_x + yu_y = 0$
   (c) $3u_{xx} + u_{xy} + u_{yy} + 2yu + 7 = 0$
2. Zakładając, że w równaniu hiperbolicznym jest $c = -a$, wykazać, że przez każdy punkt płaszczyzny xy przechodzą dwie charakterystyki wzajemnie prostopadłe.
3. Znaleźć równanie różniczkowe cząstkowe rzędu drugiego, którego charakterystyki spełniają równania: $x\cos\alpha + y\sin\alpha = 0$ i $x^2 + y^2 = \beta$.
4. Sprawdzić obliczenia poprzedzające przykł. 9.6.3.
5. Sprawdzić obliczenia wykonane w przykł. 9.6.3.
6. Wykazać, że równanie różniczkowe z przykł. 9.6.3 ma postać kanoniczną

$$16u_{\xi\eta} + u_\xi + u_\eta = 0.$$

7. Sprawdzić, że w (9.6.14) jest

$$f = e + a(u_\xi \xi_{xx} + u_\eta \eta_{xx}) + b(u_\xi \xi_{xy} + u_\eta \eta_{xy}) + c(u_\xi \xi_{yy} + u_\eta \eta_{yy}).$$

8. Wyznaczyć obszar, w którym równanie $xy^2 u_{xx} = u_{yy}$ jest hiperboliczne. Znaleźć jego postać kanoniczną i przekształcenie zmiennych, które do niej prowadzi.

9. Przekształcenie (9.6.13) zmiennych daje równanie (9.6.14) postaci

$$\alpha u_{\xi\xi} + \beta u_{\xi\eta} + \gamma u_{\eta\eta} + e = 0.$$

Udowodnić, że $\beta^2 - 4\alpha\gamma = (b^2 - 4ac)J^2$, gdzie J jest jakobianem przekształcenia:

$$J := \begin{vmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{vmatrix}.$$

10. Sprawdzić, że krzywa, na której $\xi(x, y)$ lub $\eta(x, y)$ jest stałe, jest charakterystyką.

11. Sprawdzić końcowe równanie w przykł. 9.6.4.

Zadania komputerowe 9.6

K1. Napisać program rozwiązujący zagadnienie brzegowe

$$au_{xx} + bu_{yy} + cu_x + eu_y = 0,$$
$$u(0, y) = f(y), \quad u_x(0, y) = g(y) \quad (0 \leqslant y \leqslant 1),$$

gdzie a, b, c i e są stałymi i $ab < 0$ (równanie typu hiperbolicznego). Przedział $[0, 1]$ zmiennej y ma być podzielony na równe podprzedziały.

9.7. Inne metody dla zagadnień hiperbolicznych

Opiszemy teraz pewne metody różnicowe dla układów hiperbolicznych równań różniczkowych cząstkowych rzędu pierwszego. Zaczynamy od jednego równania z warunkiem początkowym:

$$u_t = \alpha u_x, \quad u(x, 0) = f(x) \quad (-\infty < x < \infty). \tag{9.7.1}$$

α jest tu stałą rzeczywistą. Ponieważ rozwiązanie jest dane na prostej $t = 0$, więc naturalną wydaje się metoda przejścia kolejno na proste $t = k$, $t = 2k$ itd., gdzie k jest długością kroku wybranego dla zmiennej t.

Metoda Laxa-Wendroffa

Załóżmy, że rozwiązanie u ma wszystkie pochodne ciągłe względem obu zmiennych. Ze wzoru Taylora wynika, że

$$u(x,t+k) = u + ku_t + \frac{1}{2!}k^2u_{tt} + \frac{1}{3!}k^3u_{ttt} + \dots,$$

gdzie wszystkie funkcje po prawej stronie mają argumenty x i t. Szereg można obciąć w dowolnym miejscu, dołączając odpowiednią resztę.

Funkcja u spełnia równanie różniczkowe $u_t = \alpha u_x$. Stąd wynika, że $u_{tt} = \alpha^2 u_{xx}$ itd. (zob. zad. 1). Dlatego powyższą równość można napisać tak:

$$u(x,t+k) = u + k\alpha u_x + \frac{1}{2!}(k\alpha)^2u_{xx} + \frac{1}{3!}(k\alpha)^3u_{xxx} + \dots \tag{9.7.2}$$

Chcąc otrzymać algorytm o dokładności rzędu drugiego względem k, ograniczymy ten szereg do trzech początkowych składników. Prócz tego pochodne zastąpimy odpowiednimi wyrażeniami różnicowymi, w których h jest długością kroku dla zmiennej x:

$$v(x,t+k) = v(x,t) + k\alpha\frac{v(x+h,t)-v(x-h,t)}{2h} + \\ + \frac{1}{2}(k\alpha)^2\frac{v(x+h,t)-2v(x,t)+v(x-h,t)}{h^2},$$

czyli

$$v(x,t+k) = (2s^2+s)v(x+h,t) + (1-4s^2)v(x,t) + (2s^2-s)v(x-h,t),$$

gdzie $s := (\alpha k)/(2h)$. Symbol v użyty zamiast u przypomina, że mamy tu na ogół dwie różne funkcje.

Metoda oparta na powyższym wzorze nosi nazwę *metody Laxa-Wendroffa*. W taki sam sposób można oczywiście zbudować metody wyższych rzędów; zob. zad. 2.

Analiza stabilności

W prostym zagadnieniu (9.7.1) rozwiązanie dokładne łatwo odgadnąć:

$$u(x,t) = f(x+\alpha t).$$

Porównamy z nim rozwiązanie przybliżone v. Dla siatki punktów (jh, nk) niech będzie $v_{jn} := v(jh, nk)$. Metodę Laxa-Wendroffa opisuje zatem wzór

$$v_{j,n+1} = (2s^2+s)v_{j+1,n} + (1-4s^2)v_{jn} + (2s^2-s)v_{j-1,n}. \tag{9.7.3}$$

Jego stabilność zbadamy metodą Fouriera, opisaną w podrozdz. 9.1. Szukamy więc wielkości

$$v_{jn} = e^{ij\beta h} e^{n\lambda k} \quad (i := \sqrt{-1}), \tag{9.7.4}$$

spełniających ten wzór. Ich podstawienie do (9.7.3) daje po uproszczeniach równość

$$\begin{aligned} e^{\lambda k} &= 1 - 4s^2 + (2s^2 + s)e^{i\beta h} + (2s^2 - s)e^{-i\beta h} = \\ &= 1 - 4s^2 + s(e^{i\beta h} - e^{-i\beta h}) + 2s^2(e^{i\beta h} + e^{-i\beta h}) = \\ &= 1 - 4s^2 + 2is \sin \beta h + 4s^2 \cos \beta h. \end{aligned}$$

Metoda jest stabilna, jeśli $|e^{\lambda k}| \leqslant 1$. Ponieważ

$$|e^{\lambda k}|^2 = 1 - 16s^2 \sin^2 \theta (1 - 4s^2 \sin^2 \theta + \cos^2 \theta), \tag{9.7.5}$$

gdzie $\theta := \beta h/2$, więc warunkiem stabilności jest nierówność

$$1 - 4s^2 \sin^2 \theta + \cos^2 \theta \geqslant 0. \tag{9.7.6}$$

Wyrażenie po lewej stronie jest najmniejsze dla $\theta = \pi/4$, a to minimum jest nieujemne, jeśli $|s| \leqslant 1/2$. Tak więc warunkiem koniecznym stabilności jest nierówność $k|\alpha| \leqslant h$.

Układy równań

Rozważmy teraz układ równań

$$U_t = AU_x,$$

gdzie U jest wektorem o m składowych, które są funkcjami zmiennych x i t. Układ jest z definicji *hiperboliczny*, jeśli macierz A stopnia m ma m różnych wartości własnych rzeczywistych. Wektorowy wariant metody Laxa-Wendroffa jest analogiczny do wzoru dla jednego równania. Ponieważ

$$\begin{aligned} V(x, t + k) = V(x, t) &+ \frac{k}{2h} A[V(x + h, t) - V(x - h, t)] + \\ &+ \frac{k^2}{2h^2} A^2[V(x + h, t) - 2V(x, t) + V(x - h, t)], \end{aligned}$$

więc

$$\begin{aligned} V(x, t + h) = \tau A(2\tau A + I)V(x + h, t) &+ [I - (2\tau A)^2]V(x, t) + \\ &+ \tau A(2\tau A - I)V(x - h, t), \end{aligned}$$

gdzie $\tau := k/(2h)$. Nie wchodząc w szczegóły, podajemy tylko, że ta metoda numeryczna jest stabilna wtedy i tylko wtedy, gdy wszystkie wartości własne macierzy A leżą w przedziale $[-h/k, h/k]$.

Metoda niejawna Wendroffa

Rozważaliśmy dotąd równanie różniczkowe wraz z warunkiem początkowym. Jeśli on dotyczy tylko przedziału, np. $[0,1]$, to w poprawnie postawionym zagadnieniu wartości funkcji u muszą być dane na całej granicy obszaru określonego nierównościami $0 \leqslant x \leqslant 1$ i $t \geqslant 0$. W takim przypadku równanie różniczkowe $u_t = \alpha u_x$ można rozwiązywać metodami niejawnymi. Należy do nich *metoda Wendroffa*. Napiszemy ją najpierw w postaci sugerującej jej pochodzenie:

$$\frac{1}{2}\left(\frac{v_{j,n+1}-v_{jn}}{k}+\frac{v_{j+1,n+1}-v_{j+1,n}}{k}\right)= \\ =\frac{\alpha}{2}\left(\frac{v_{j+1,n}-v_{jn}}{h}+\frac{v_{j+1,n+1}-v_{j,n+1}}{h}\right). \tag{9.7.7}$$

Jak widać, pochodne są tu zastąpione średnimi różnic.

Analiza błędu

Sprawdzimy, że błąd metody Wendroffa wynosi $\mathcal{O}(h^2+k^2)$. To twierdzenie wymaga dodatkowych wyjaśnień. Metoda dotyczy równania $Lu = 0$, gdzie $Lu := u_t - \alpha u_x$. Operator L zastąpiliśmy operatorem różnicowym

$$\tfrac{1}{2}(A+B)-\tfrac{1}{2}\alpha(C+D),$$

gdzie

$$(Au)(x,t) := \frac{u(x,t+k)-u(x,t)}{k} = u_t\left(x,t+\tfrac{1}{2}h\right)+\mathcal{O}(k^2),$$
$$(Bu)(x,t) := \frac{u(x+h,t+k)-u(x+h,t)}{k} = u_t\left(x+h,t+\tfrac{1}{2}k\right)+\mathcal{O}(k^2),$$
$$(Cu)(x,t) := \frac{u(x+h,t)-u(x,t)}{h} = u_x\left(x+\tfrac{1}{2}h,t\right)+\mathcal{O}(h^2),$$
$$(Du)(x,t) := \frac{u(x+h,t+k)-u(x,t+k)}{h} = u_x\left(x+\tfrac{1}{2}h,t+k\right)+\mathcal{O}(h^2).$$

Nie twierdzimy, że dla dowolnej dostatecznie gładkiej funkcji u jest

$$[L-\tfrac{1}{2}(A+B)+\tfrac{1}{2}\alpha(C+D)]u = \mathcal{O}(h^2+k^2),$$

ale że tak jest, gdy spełnia ona równanie $Lu = 0$. Mamy więc wykazać, że wtedy

$$(A+B)u-\alpha(C+D)u = \mathcal{O}(h^2+k^2).$$

Twierdzenie 9.7.1. *Błąd metody Wendroffa jest równy* $\mathcal{O}(h^2+k^2)$.

Dowód. Ze wzoru Taylora i równania różniczkowego wynika, że

$$Au = u_t(x,t)+\tfrac{1}{2}ku_{tt}(x,t)+\mathcal{O}(k^2) = \alpha u_x(x,t)+\tfrac{1}{2}\alpha^2 h u_{xx}(x,t)+\mathcal{O}(k^2).$$

Zmieniając tu x na $x+h$ wnioskujemy, że

$$\begin{aligned} Bu &= \alpha u_x(x+h,t)+\tfrac{1}{2}\alpha^2 h u_{xx}(x+h,t)+\mathcal{O}(k^2) = \\ &= \alpha u_x(x,t)+\alpha h u_{xx}(x,t)+\mathcal{O}(h^2)+\tfrac{1}{2}\alpha^2 k u_{xx}(x,t)+\mathcal{O}(hk)+\mathcal{O}(k^2). \end{aligned}$$

W podobny sposób otrzymujemy równości

$$\begin{aligned} Cu &= u_x(x,t)+\tfrac{1}{2}hu_{xx}(x,t)+\mathcal{O}(h^2), \\ Du &= u_x(x,t+k)+\tfrac{1}{2}hu_{xx}(x,t+k)+\mathcal{O}(h^2) = \\ &= u_x(x,t)+ku_{xt}(x,t)+\mathcal{O}(k^2)+\tfrac{1}{2}hu_{xx}(x,t)+\mathcal{O}(hk)+\mathcal{O}(h^2) = \\ &= u_x(x,t)+\alpha k u_{xx}(x,t)+\mathcal{O}(k^2)+\tfrac{1}{2}hu_{xx}(x,t)+\mathcal{O}(hk)+\mathcal{O}(h^2). \end{aligned}$$

Dlatego $[A+B-\alpha(C+D)]u = \mathcal{O}(k^2)+\mathcal{O}(hk)+\mathcal{O}(h^2) = \mathcal{O}(h^2+k^2)$. ■

Analiza stabilności

Aby zbadać, czy metoda Wendroffa zdefiniowana równaniem (9.7.7) jest stabilna, wyrażamy je w postaci

$$(1+r)v_{j,n+1}+(1-r)v_{j+1,n+1} = (1-r)v_{jn}+(1+r)v_{j+1,n},$$

gdzie $r := \alpha k/h$. Jak dla metody Laxa-Wendroffa, podstawiamy tu wielkości v_{jn} typu (9.7.4). Po uproszczeniach otrzymujemy równość

$$(1+r)e^{\lambda k}+(1-r)e^{\lambda k}e^{i\beta h} = 1-r+(1+r)e^{i\beta h},$$

czyli

$$e^{\lambda k} = \frac{e^{i\beta h}+1+r(e^{i\beta h}-1)}{e^{i\beta h}+1-r(e^{i\beta h}-1)}.$$

Nierówność $|e^{\lambda k}| \leqslant 1$ jest spełniona wtedy i tylko wtedy, gdy moduł licznika po prawej stronie jest nie większy od modułu mianownika. Z zadania 6 wynika równoważna nierówność:

$$(1+\cos\beta h)(\cos\beta h-1)+\sin^2\beta h \leqslant 0. \tag{9.7.8}$$

Jej lewa strona jest równa 0, metoda jest zatem stabilna dla dowolnych α, h i k.

Metody Galerkina

Do zagadnień hiperbolicznych można również zastosować metody Galerkina. Aby uzasadnić jedną z nich, rozważmy równanie rzędu pierwszego z warunkami granicznymi:

$$\begin{aligned} &u_t = \alpha u_x, \\ &u(x,0) = g(x) \qquad (0 \leqslant x \leqslant 1), \\ &u(0,t) = u(1,t) = 0 \quad (t > 0). \end{aligned}$$

Wybieramy najpierw funkcje bazowe $w_1, w_2, \ldots, w_n$ zmiennej x. Próbujemy znaleźć rozwiązanie powyższego zagadnienia, mające postać

$$u(x,t) = \sum_{j=1}^{n} v_j(t) w_j(x). \tag{9.7.9}$$

Podstawiamy je do równania różniczkowego:

$$\sum_{j=1}^{n} [v_j'(t) w_j(x) - \alpha v_j(t) w_j'(x)] = 0. \tag{9.7.10}$$

Jak zwykle w takich przypadkach, nie możemy spodziewać się, aby ten układ był niesprzeczny. W metodzie Galerkina (i innych podobnych) szukamy więc rozwiązania przybliżonego. Trzeba przy tym pamiętać o warunkach granicznych. Jeśli przyjmiemy, że każda z funkcji w_j znika w punktach 0 i 1, to warunek brzegowy będzie automatycznie spełniony.

Określmy iloczyn skalarny funkcji wzorem $\langle f, g \rangle := \int_0^1 f(x) g(x)\, dx$. Mnożąc skalarnie obie strony (9.7.10) przez w_i, otrzymujemy układ równań

$$\sum_{j=1}^{n} [v_j'(t) \langle w_j, w_i \rangle - \alpha v_j(t) \langle w_j', w_i \rangle] = 0 \quad (1 \leqslant i \leqslant n). \tag{9.7.11}$$

Jest to układ n równań różniczkowych liniowych jednorodnych względem tyluż nieznanych funkcji v_j. Zauważamy, że warto wybrać funkcje w_j tak, żeby tworzyły układ ortonormalny w przedziale $[0,1]$. Wtedy układ (9.7.11) przybiera postać

$$V' = AV, \tag{9.7.12}$$

gdzie $V := (v_1, v_2, \ldots, v_n)$, a macierz A ma elementy $a_{ij} := \alpha \langle w_j', w_i \rangle$.

Warunek początkowy w badanym zagadnieniu wyraża się równaniem

$$\sum_{j=1}^{n} v_j(0) w_j(x) = g(x),$$

którego na ogół też nie można spełnić dokładnie. Możemy natomiast wybrać $v_j(0)$ tak, żeby różnica obu stron tego równania miała najmniejszą normę w L^2. Dzięki ortonormalności układu $\{w_j\}$ jest tak, jeśli

$$v_j(0) = \langle g, w_j \rangle.$$

Są to warunki początkowe dla układu (9.7.12). Z podrozdziału 8.11 wynika, że

$$V(t) = e^{tA}V(0).$$

Wygodny układ ortonormalny składa się z funkcji

$$w_j(x) := 2^{-1/2} \sin \pi j x,$$

które znikają w punktach 0 i 1. Inny taki układ można otrzymać, stosując metodę Grama-Schmidta do funkcji $x \mapsto (x-1)x^j$ $(1 \leqslant j \leqslant n)$.

Zadania 9.7

1. Udowodnić (także w przypadku wektorowym), że jeśli $u_t = \alpha u_x$, to

$$\frac{\partial^n u}{\partial t^n} = \alpha^n \frac{\partial^n u}{\partial x^n} \quad (n \geqslant 0).$$

2. Posługując się rozwinięciem (9.7.2), znaleźć metodę przybliżoną rzędu trzeciego.

3. Sprawdzić szczegóły obliczeń prowadzących do relacji (9.7.5) i (9.7.6) oraz warunku stabilności $k|\alpha| \leqslant h$.

4. Dla równania $u_t = \alpha u_x$ zbadać stabilność:

(**a**) metody opisanej równaniem

$$\frac{1}{2k}(v_{j,n+1} - v_{j,n-1}) = \frac{\alpha}{2h}(v_{j+1,n} - v_{j-1,n}),$$

(**b**) metody Eulera

$$v_{j,n+1} = v_{jn} + \frac{\alpha k}{2h}(v_{j+1,n} - v_{j-1,n}).$$

5. Wykazać, że dla metody Wendroffa główny składnik błędu jest równy

$$\tfrac{1}{12}(\alpha h^2 - \alpha^3 k^2) u_{xxx}(x,t).$$

6. Niech $u = x + iy$ i $v = a + ib$ będą dwiema liczbami zespolonymi. Wykazać, że nierówności $|u+v| \leqslant |u-v|$ i $xa + yb \leqslant 0$ są równoważne. Sprawdzić, że zachodzi (9.7.8).

ZADANIA KOMPUTEROWE 9.7

K1. Zaprogramować metodę Laxa-Wendroffa, przyjmując, że rozwiązanie numeryczne ma być znalezione dla $t = T$ i $a \leqslant x \leqslant b$, gdzie a, b i T (a także f, α, h i k) są dane.

K2. Sprawdzić program z zad. K1 dla zagadnienia $u_t = 2u_x$, $u(x,0) = (1-x)^2$. Przyjąć, że $h = 0.02$, $k = 0.01$, $a = 1$, $b = 2$ i $T = 1$. Porównać wyniki z dokładnym rozwiązaniem.

9.8. Metody wielosiatkowe

W *metodach wielosiatkowych*, jak w wielu innych, pochodne przybliża się wyrażeniami różnicowymi. Istotą tych metod jest to, że w czasie obliczeń stosuje się kolejno wiele siatek, od rzadkich do gęstych. Na rzadkiej siatce obliczenia są szybkie, ale ich wyniki są mało dokładne. Mogą one jednak być dobrym punktem wyjścia do obliczeń na gęstszej siatce. Jest to tylko jeden z aspektów strategii wielu siatek, ale przed poznaniem innych warto prześledzić działanie takiej metody na bardzo prostym przykładzie. Będzie to zagadnienie brzegowe z równaniem różniczkowym zwyczajnym

$$u''(x) = f(x), \quad u(0) = u(1) = 0, \tag{9.8.1}$$

chociaż zalety metod wielosiatkowych ujawniają się w pełni dopiero w zastosowaniach do równań różniczkowych cząstkowych. W dyskretnej wersji tego zadania dla danego n naturalnego uwzględniamy tylko punkty

$$x_j := jh, \quad \text{gdzie} \quad h := \frac{1}{n+1}.$$

Stosując standardowe przybliżenie drugiej pochodnej, otrzymujemy równania

$$h^{-2}(v_{j-1} - 2v_j + v_{j+1}) = f_j \quad (1 \leqslant j \leqslant n), \tag{9.8.2}$$

gdzie $v_j \approx u(x_j)$, $v_0 = v_{n+1} = 0$, $f_j := f(x_j)$. Ten układ można łatwo rozwiązać bezpośrednio, ale mając na uwadze równania różniczkowe cząstkowe, zastosujemy tu metodę iteracyjną, a konkretniej metodę Gaussa-Seidela. Taka metoda daje oczywiście zadowalające rozwiązanie tym szybciej, im lepsze jest jego przybliżenie początkowe. Można je otrzymać, rozwiązując analogiczny układ dla rzadszej siatki. Każdy taki układ ma postać

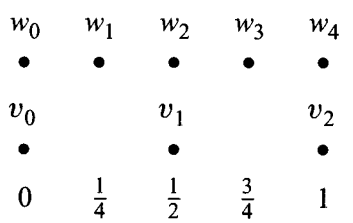
$$\begin{bmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ 0 & -1 & 2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \dots \\ v_n \end{bmatrix} = \begin{bmatrix} -h^2 f_1 \\ -h^2 f_2 \\ -h^2 f_3 \\ \dots \\ -h^2 f_n \end{bmatrix},$$

różne są tylko n i $h = 1/(n+1)$. Narzuca się myśl, aby zacząć od najrzadszej siatki, dla której powyższy układ ma sens, tj. od $n = 1$. Jego rozwiązaniem (wraz z wartościami brzegowymi) są liczby

$$v_0 = 0, \quad v_1 = -\tfrac{1}{8} f\left(\tfrac{1}{2}\right), \quad v_2 = 0.$$

Możemy teraz przejść do gęstszej siatki. Dzielimy więc h przez 2, zmieniamy n na $2n + 1$ i tworzymy nowy układ równań, w którym $n = 3$ i $h = 1/4$ (rys. 9.11). Jego rozwiązaniem niech będzie chwilowo wektor w. Początkowe wartości jego składowych tworzymy za pomocą wektora v. Nazwiemy to *rozszerzeniem* wektora v. Można to robić na wiele sposobów. Najprostszy polega na tym, że w punktach należących do poprzedniej siatki kopiujemy odpowiednie v_i, a w pozostałych punktach bierzemy średnie arytmetyczne wartości z jej sąsiednich punktów. Na początku zatem przyjmujemy, że

$$w_0 := v_0, \quad w_2 := v_1, \quad w_4 := v_2, \quad w_1 := \tfrac{1}{2}(v_0+v_1), \quad w_3 := \tfrac{1}{2}(v_1+v_2).$$



RYS. 9.11. Zagęszczenie siatki

Następnie wpisujemy wektor w na miejsce v i wykonujemy pewną liczbę iteracji metodą Gaussa-Seidela. Podobnie postępujemy dalej. Daje to następujący algorytm, w którym m jest liczbą stosowanych siatek, a k – liczbą iteracji dla każdej siatki:

input m, k
$h \leftarrow 1/2$; $n \leftarrow 1$
$v_0 \leftarrow 0$; $v_2 \leftarrow 0$; $v_1 \leftarrow -f(h)/8$
for $i = 2$ **to** m **do**
    $h \leftarrow h/2$; $n \leftarrow 2n + 1$
    for $j = 0$ **to** $(n + 1)/2$ **do**
        $w_{2j} \leftarrow v_j$
    end do
    for $j = 1$ **to** $(n + 1)/2$ **do**
        $w_{2j-1} \leftarrow (v_{j-1} + v_j)/2$
    end do
    for $j = 0$ **to** $n + 1$ **do**

```
        v_j ← w_j
    end do
    for p = 1 to k do
        for j = 1 to n do
            v_j ← [v_{j-1} + v_{j+1} − h²f(jh)]/2
        end do
    end do
    output (v_i)
    end do
```

Zagęszczenie siatki za pomocą interpolacji można znacznie uprościć; zob. zad. K1. Podany program służy tylko celom dydaktycznym, ale możemy go stosować w testach, szczególnie dla zagadnień, których dokładne rozwiązanie jest znane, bo wtedy można obliczyć odchylenie rozwiązania numerycznego od dokładnego. Zrobiono tak dla $f(x) = \cos x$, gdy rozwiązaniem zagadnienia brzegowego jest funkcja

$$u(x) = -\cos x + x(\cos 1 - 1) + 1. \tag{9.8.3}$$

Dla $m = 6$ (wtedy na końcu obliczeń jest $h = 1/64$, czyli $h^2 \approx 2_{10}{-}4$) i $k = 3$ to odchylenie nie przekracza 10^{-3}.

Tłumienie błędów

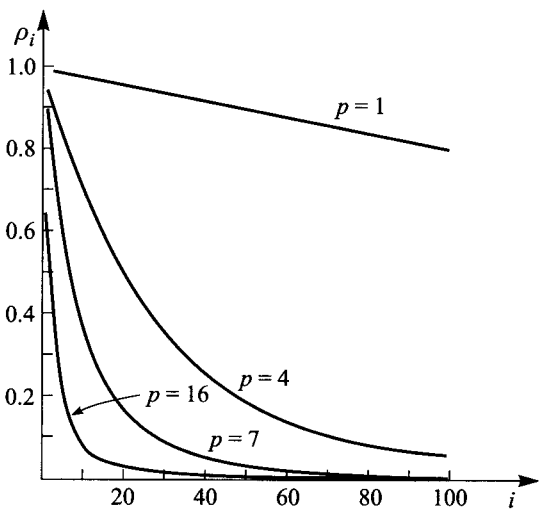
Inną ważną częścią składową metod wielosiatkowych jest systematyczne przechodzenie od siatek gęstych do rzadszych, a więc w odwrotnym kierunku do opisanego wcześniej. Tak postępujemy w drugiej fazie obliczeń, gdyż – jak się okazuje – błędy o niskiej częstotliwości można skutecznie wytłumić, operując na rzadkiej siatce. Na odwrót, błędy o wysokiej częstotliwości są dobrze tłumione w obliczeniach na gęstej siatce. Wykorzystanie tych ważnych własności w metodach wielosiatkowych w dużej mierze przesądza o ich efektywności.

Prosty test numeryczny związany z zagadnieniem (9.8.1) pozwala zrozumieć sens powyższych uwag. Niech będzie $f(x) \equiv 0$, czyli zagadnienie jest jednorodne i jego rozwiązanie u również znika tożsamościowo. Zaczynamy jednak, dla ustalonego n, od wartości

$$v_j := \sin \frac{jp\pi}{n+1} \quad (0 \leqslant j \leqslant n+1).$$

Wektor v można tu uznać za modelowy błąd rozwiązania. Parametr p steruje *częstotliwością* błędu. Niżej podano algorytm, który dla czterech wybranych wartości p wykonuje k iteracji metodą Gaussa-Seidela. Po każdej iteracji program oblicza normę aktualnego wektora v.

```
input n, k, p1, p2, p3, p4
for p = p1, p2, p3, p4 do
    for j = 0 to n + 1 do
        v_j ← sin(jpπ/(n + 1))
    end do
    for i = 1 to k do
        for j = 1 to n do
            v_j ← (v_{j-1} + v_{j+1})/2
        end do
        ρ_i ← ||v||_∞ := max_{1≤j≤n} |v_j|
        output p, i, ρ_i
    end do
end do
```



RYS. 9.12. Redukcja błędu w kolejnych iteracjach

Rysunek 9.12 pokazuje, jak dla $n = 63$, $k = 100$ i $p = 1, 4, 7, 16$ maleje ρ_i wraz ze wzrostem i. Jest oczywiste, że tylko błędy o wysokiej częstotliwości są szybko tłumione w kolejnych iteracjach.

Analiza tłumienia błędu

Badając wpływ kolejnych iteracji na tłumienie błędu, wybierzemy teraz jako przykład metodę Jacobiego. Dla układu $Ax = b$ polega ona na stosowaniu wzoru

$$x^{(k+1)} := (I - D^{-1}A)x^{(k)} + D^{-1}b, \tag{9.8.4}$$

gdzie $D = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$. W przykładzie z tego podrozdziału jest

$$A = \begin{bmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ldots & 0 \\ 0 & -1 & 2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 2 \end{bmatrix},$$

a wektor b ma składowe $h^2 f_j$. Dlatego

$$G := I - D^{-1}A = I - \frac{1}{2}A = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \ldots & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \ldots & 0 \\ 0 & \frac{1}{2} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix}.$$

Z lematu 9.1.1 wynika, że macierz A ma wartości własne

$$\mu_j = 2 - 2\cos\frac{j\pi}{n+1} \quad (1 \leqslant j \leqslant n).$$

Wobec tego wartościami własnymi macierzy G są liczby

$$\lambda_j = 1 - \frac{1}{2}\mu_j = \cos\frac{j\pi}{n+1} \quad (1 \leqslant j \leqslant n),$$

a jej promień spektralny jest równy

$$\rho(G) = \cos\frac{\pi}{n+1} \approx 1 - \frac{1}{2}\left(\frac{\pi}{n+1}\right)^2 = 1 - \frac{1}{2}\pi^2 h^2.$$

Z powyższych rozważań wynikają już pewne wnioski. Po pierwsze, ponieważ $\rho(G) < 1$, więc metoda Jacobiego jest zbieżna (tw. 4.6.7). Po drugie, dla małych h promień spektralny jest bliski 1, czyli dla gęstych siatek jej zbieżność jest wolna. Po trzecie, wektory własne macierzy G są bazą przestrzeni $\mathbb{R}^n$. Każdy błąd jest ich kombinacją liniową. Jeśli $v^{(j)}$ jest wektorem odpowiadającym wartości własnej λ_j, to po k iteracjach (9.8.4) daje on wektor $\lambda_j^k v^{(j)}$, który dla $k \to \infty$ dąży do 0. Ten efekt *tłumienia błędu* jest tym wyraźniejszy, im $|\lambda_j|$ jest mniejsze. Te wielkości są małe dla j ze środka przedziału $[1, n]$, natomiast dla skrajnych j jest $|\lambda_j| \approx 1$.

Rozrzedzanie siatki

Całość obliczeń z użyciem wielu siatek dzieli się na etapy, w których wyniki otrzymane dla jednej siatki przenosi się na inną. Wiemy już, jak się przenosi

wyniki z siatki rzadkiej na gęstszą. Rozważmy teraz odwrotne przejście. Niech v^i oznacza przybliżone rozwiązanie układu (9.8.2) (do końca tego podrozdziału górne wskaźniki wektorów i macierzy, mające inny sens niż przedtem, nie są ujmowane w nawiasy; pamiętajmy, że ten wskaźnik nie oznacza tu wykładnika potęgi). Chcemy ulepszyć v^i, dodając odpowiednią poprawkę e^i. Niech A^i będzie macierzą układu, a f^i – wektorem wartości funkcji f na danej siatce. Powinno więc być $A^i(v^i + e^i) = f^i$, czyli

$$A^i e^i = f^i - A^i v^i = r^i,$$

gdzie r^i jest wektorem residualnym odpowiadającym przybliżeniu v^i. Zamiast rozwiązywać ten układ, przechodzimy do rzadszej siatki, której odpowiada prostszy układ

$$A^{i-1} e^{i-1} = r^{i-1} \tag{9.8.5}$$

i dla niego wykonujemy kilka iteracji metodą Gaussa-Seidela, skracając w ten sposób obliczenia. Wektor r^{i-1} wynika z r^i przez naturalne przeniesienie informacji z gęstszej siatki. Najprościej jest przyjąć, że

$$r^{i-1}_j := r^i_{2j}. \tag{9.8.6}$$

Bardziej wyrafinowany sposób daje średnie ważone

$$r^{i-1}_j := \tfrac{1}{4} r^i_{2j-1} + \tfrac{1}{2} r^i_{2j} + \tfrac{1}{4} r^i_{2j+1}.$$

Zawsze takie przejście nazywamy *ograniczeniem*.

Gdy już znamy rozwiązanie e^{i-1} układu (9.8.5), możemy przejść do e^i, stosując w znany sposób interpolację i poprawić wektor v^i przez dodanie e^i.

Cykl obliczeń

Naszym ostatecznym celem jest opis części składowej metod wielosiatkowych, zwanej *cyklem obliczeń*. Obliczenia zaczynają się od najgęstszej siatki. Po kilku iteracjach otrzymujemy residuum r i przechodzimy do rzadszej siatki, dla której rozważamy układ $Az = r$. Stosowane wielokrotnie takie przejścia „w dół" doprowadzają nas do najrzadszej siatki i układu łatwego do rozwiązania, bo zawierającego np. tylko jedno równanie. Druga część cyklu składa się z przejść „w górę", do coraz gęstszych siatek. Stosujemy przy tym interpolację i dodatkowe iteracje poprawiające dokładność.

Opisując formalnie cykl, oznaczamy symbolami $\mathcal{G}^1, \mathcal{G}^2, \ldots, \mathcal{G}^m$ stosowane siatki, od najrzadszej do najgęstszej. Siatce $\mathcal{G}^i$ odpowiada macierz A^i. Ostatecznym celem obliczeń jest rozwiązanie układu $A^m v^m = f^m$, który jest

dyskretną wersją danego zagadnienia brzegowego na najgęstszej siatce. f^m jest więc danym wektorem, natomiast $f^{m-1}, f^{m-2}, \ldots, f^1$ trzeba obliczyć. Cykl składa się z następujących czynności:

1. Przyjąć $i := m$, utworzyć wektor f^m i możliwie dobre wstępne przybliżenie dla v^m.
2. Dla układu $A^i v^i = f^i$ wykonać k iteracji wybraną metodą, zaczynając od danego v^i. Obliczyć residuum: $r^i \leftarrow f^i - A^i v^i$. Zastosować operator ograniczenia R_i: $f^{i-1} \leftarrow R_i r^i$. Zmniejszyć i o 1.
3. Jeśli $i = 1$, przejść do etapu **4**, a w przeciwnym razie do **2**.
4. Rozwiązać dokładnie układ $A^1 v^1 = f^1$.
5. Zwiększyć i o 1. Stosując operator rozszerzenia E_i, poprawić wektor v_i, dodając doń $E_i v^{i-1}$. Dla układu $A^i v^i = f^i$ wykonać k iteracji wybraną metodą, zaczynając od danego v^i.
6. Jeśli $i < m$, przejść do etapu **5**, a w przeciwnym razie uznać v^m za wynik obliczeń i zakończyć je.

Algorytm wzorujący się na tym opisie podano niżej. Dla oszczędności miejsca niektóre pętle są wyrażone w uproszczony sposób. Na siatce $\mathcal{G}^i$ długość kroku h wynosi 2^{-i}. Tablica w służy do czasowego przechowywania potrzebnych wielkości. Operator ograniczenia występuje w najprostszej wersji (9.8.6). Rozwiązania poprawia się metodą Gaussa-Seidela. Inne szczegóły algorytmu są zgodne z powyższym opisem.

input m, k
$n \leftarrow 2^m - 1$
$h \leftarrow 1/(n+1)$
$v_j^i \leftarrow 0$; $f_j^i \leftarrow 0 \quad (1 \leqslant i \leqslant m,\ 0 \leqslant j \leqslant n+1)$
$f_j^m \leftarrow f(jh) \quad (1 \leqslant j \leqslant n)$
for $i = m$ **to** 2 **step** -1 **do**
    for $p = 1$ **to** k **do**
        $v_j^i \leftarrow (v_{j-1}^i + v_{j+1}^i - h^2 f_j^i)/2 \quad (1 \leqslant j \leqslant n)$
    end do
    $w_j \leftarrow f_j^i - (v_{j-1} - 2v_j^i + v_{j+1}^i)/h^2 \quad (1 \leqslant j \leqslant n)$
    $n \leftarrow (n-1)/2$
    $f_j^{i-1} \leftarrow w_{2j} \quad (1 \leqslant j \leqslant n)$
    $h \rightarrow 2h$
end do
$v_1^1 \leftarrow -f(1/2)/8$
for $i = 2$ **to** m **do**
    $h \leftarrow h/2$
    $w_{2j} \leftarrow v_j^{i-1} \quad (0 \leqslant j \leqslant n+1)$
    $w_{2j-1} \leftarrow (v_{j-1}^{i-1} + v_j^{i-1})/2 \quad (1 \leqslant j \leqslant n+1)$
    $n \leftarrow 2n+1$
    $v_j^i \leftarrow v_j^i + w_j \quad (0 \leqslant j \leqslant n+1)$

for $p = 1$ **to** k **do**
$v_j^i \leftarrow (v_{j-1}^i + v_{j+1}^i - h^2 f_j^i)/2 \quad (1 \leqslant j \leqslant n)$
end do
end do
output $v_j^m \quad (0 \leqslant j \leqslant n+1)$

Zaprogramowany algorytm został użyty w różnych testach numerycznych dla zagadnienia

$$u'' = \cos x, \quad u(0) = u(1) = 0, \tag{9.8.7}$$

którego rozwiązaniem jest funkcja (9.8.3). W jednym z nich było $m = 7$, czyli dla najgęstszej siatki $h = 1/128$. Nadając różne wartości parametrowi k, sprawdzano, jak wpływa on na błąd $\max_{0 \leqslant j \leqslant n+1} |u(jh) - v_j^m|$. Dla $3 \leqslant k \leqslant 8$ zwiększenie k o 1 powodowało pomnożenie poprzedniego błędu przez ok. 0.4. Później błąd malał już coraz wolniej; dla $9 \leqslant k \leqslant 12$ analogiczny czynnik wynosił odpowiednio 0.6, 0.7, 0.8, 0.9.

Koszt obliczeń

Zbadajmy jeszcze koszt wykonania cyklu. Weźmy najpierw pod uwagę fazę przechodzenia przez m coraz rzadszych siatek. Dla siatki $\mathcal{G}^i$ mamy 2^i punktów, tyleż niewiadomych i równań. Każda aktualizacja niewiadomej w metodzie Gaussa-Seidela wymaga czterech działań, razem jest ich więc $4k \cdot 2^i$. Obliczenie residuum kosztuje $5 \cdot 2^i$ dodatkowych działań. Dla m siatek trzeba wykonać

$$\sum_{i=1}^{m} (4k+5)2^i \approx (8k+10)2^m$$

działań. Dla fazy zagęszczania siatek podobny rachunek daje ok. $(8k+4)2^m$ działań. Tak więc cały cykl wymaga wykonania ok. $16(k+1)2^m$ działań.

Jakie byłyby odpowiednie liczby dla zagadnienia dwuwymiarowego, np. z równaniem $\nabla^2 u = f(x,y)$ w kwadracie jednostkowym? Najprostsza dyskretyzacja daje równania zawierające po pięć niewiadomych, a zatem aktualizacja wartości każdej z nich wymaga ok. sześciu działań. W i-tej siatce mamy teraz $(2^i)^2$ punktów. Wobec tego najbardziej istotna w naszych obliczeniach jest zmiana 2^m na 4^m. Drugi czynnik będzie nieco większy od $16(k+1)$, ale jednak liniowy względem k. Wynika stąd, że koszt obliczeń metodą wielosiatkową zależy wykładniczo od liczby m siatek i liniowo od liczby k wykonywanych iteracji. Jest tak niezależnie od wymiaru zadania.

ZADANIA KOMPUTEROWE 9.8

K1. Zaprogramować zagęszczenie siatki za pomocą interpolacji bez użycia tablicy w.

K2. Pierwszy algorytm poprawić tak, aby usunąć zbędne obliczanie wielkości v_0 i v_{n+1}.

K3. Testując pierwszy algorytm na różnych przykładach, sprawdzić, czy jakaś liczba k iteracji metodą Seidela jest optymalna.

K4. Powtórzyć obliczenia opisane w tekście dla zagadnienia (9.8.1) i $f(x) = \cos x$.

K5. Powtórzyć eksperyment numeryczny pokazujący tłumienie błędów o różnych częstotliwościach.

K6. Powtórzyć obliczenia opisane w tekście dla zagadnienia (9.8.7).

K7. Uogólnić algorytmy podane w tekście na zagadnienie $u''(x) = f(x)$, $u(a) = \alpha$, $u(b) = \beta$.

K8. Zaprogramować obliczenia w cyklu dla zagadnienia dwuwymiarowego

$$u_{xx} + u_{yy} = f(x,y) \quad (0 < x < 1,\ 0 < y < 1), \quad u(x,y) = 0 \quad \text{na brzegu.}$$

9.9. Szybkie metody dla równania Poissona

Równanie Poissona dla dwóch zmiennych ma postać

$$u_{xx} + u_{yy} = f(x,y).$$

W typowych zagadnieniach fizycznych związanych z tym równaniem szukamy funkcji u, która je spełnia w pewnym zbiorze otwartym Ω, a także spełnia pewne warunki na brzegu $\partial\Omega$ tego zbioru. Funkcja f ma być określona w Ω.

W ostatnich latach zastosowano w takich zagadnieniach analizę Fouriera. Umożliwia to stosowanie szybkiego przekształcenia Fouriera. Na prostym przykładzie zobaczymy, na czym polegają te nowe algorytmy.

Zagadnienie modelowe

Zagadnienie jest następujące: dla

$$\Omega := \{(x,y):\ 0 < x < 1,\ 0 < y < 1\}$$

i danej funkcji f znaleźć funkcję u taką, że

$$u_{xx} + u_{yy} = f(x,y) \quad \text{w } \Omega, \quad u(x,y) = 0 \quad \text{na } \partial\Omega. \tag{9.9.1}$$

Dyskretyzując zagadnienie, przyjmujemy, że

$$h := \frac{1}{n+1}, \quad x_i := ih, \quad y_j := jh \quad (0 \leqslant i, j \leqslant n+1),$$

a także, jak zwykle,

$$v_{ij} \approx u(x_i, y_j), \quad f_{ij} := f(x_i, y_j).$$

Wersja dyskretna równania różniczkowego jest następująca:

$$h^{-2}(v_{i+1,j} - 2v_{ij} + v_{i-1,j}) + h^{-2}(v_{i,j+1} - 2v_{ij} + v_{i,j-1}) = f_{ij}. \quad (9.9.2)$$

Wskaźniki i, j przebiegają wartości $1, 2, \ldots, n$. Spośród wielkości v_{ij} są znane tylko te, które wynikają z warunku brzegowego:

$$v_{0j} = v_{n+1,j} = v_{i0} = v_{i,n+1} = 0.$$

Tradycyjny sposób rozwiązywania układu n^2 równań (9.9.2) polega na użyciu jednej z metod iteracyjnych. W przypadku metody nadrelaksacji koszt obliczeń jest rzędu $\mathcal{O}(n^3 \log n)$. Alternatywne podejście, tj. zastosowanie szybkich przekształceń zmniejsza ten koszt do $\mathcal{O}(n^2 \log n)$.

Zastosowanie szybkiego sinusowego przekształcenia Fouriera

Będziemy szukać rozwiązania układu (9.9.2) w postaci

$$v_{ij} = \sum_{k=1}^{n} \hat{v}_{kj} \sin ik\varphi \quad (0 \leqslant i,\, j \leqslant n+1), \quad (9.9.3)$$

gdzie $\varphi := \pi/(n+1)$. Trzeba wobec tego znaleźć niewiadome $\hat{v}_{kj}$. Gdy to zrobimy, będzie można obliczyć wszystkie v_{ij}, stosując *szybkie sinusowe przekształcenie Fouriera* wspomniane w podrozdz. 6.13.

Podstawiamy sumy (9.9.3) do równań (9.9.2):

$$\sum_{k=1}^{n} \hat{v}_{kj}[\sin(i+1)k\varphi - 2\sin ik\varphi + \sin(i-1)k\varphi] + \\ + \sum_{k=1}^{n} (\hat{v}_{k,j+1} - 2\hat{v}_{kj} + \hat{v}_{k,j-1}) \sin ik\varphi = h^2 f_{ij}.$$

Wyrażenie w nawiasach kwadratowych po lewej stronie upraszczamy, korzystając z tożsamości

$$\sin(A+B) - 2\sin A + \sin(A-B) = -4\sin A \sin^2 \frac{B}{2}.$$

Jednocześnie wyrażamy f_{ij} podobnie jak v_{ij}:

$$f_{ij} = \sum_{k=1}^{n} \hat{f}_{kj} \sin ik\varphi.$$

Daje to następujące równania:

$$\sum_{k=1}^{n} \hat{v}_{kj}(-4\sin ik\varphi)\sin^2\frac{k}{2}\varphi + \\ + \sum_{k=1}^{n}(\hat{v}_{k,j+1} - 2\hat{v}_{kj} + \hat{v}_{k,j-1})\sin ik\varphi = h^2\sum_{k=1}^{n}\hat{f}_{kj}\sin ik\varphi.$$

Na końcu tego podrozdziału w lem. 9.9.1 udowodniono, że macierz o elementach $\sin ij\varphi$ jest nieosobliwa. Dlatego z powyższych równań można wywnioskować, że

$$-4\left(\sin^2\frac{k\varphi}{2}\right)\hat{v}_{kj} + \hat{v}_{k,j+1} - 2\hat{v}_{kj} + \hat{v}_{k,j-1} = h^2\hat{f}_{kj}. \tag{9.9.4}$$

Na pierwszy rzut oka wydaje się, że otrzymaliśmy układ równań niewiele różniący się od (9.9.2). Zauważmy jednak, że dla ustalonego k mamy tu układ n równań zawierający tylko niewiadome $\hat{v}_{k1}, \hat{v}_{k2}, \ldots, \hat{v}_{kn}$. Macierz tego układu jest trójprzekątniowa, dzięki czemu można go łatwo rozwiązać kosztem $\mathcal{O}(n)$ (ściślej mniej niż $10n$) działań. Takich układów jest n. Prócz tego szybkie sinusowe przekształcenie Fouriera wektora o n składowych wymaga wykonania $\mathcal{O}(n\log n)$ działań. Uzasadnia to podaną wcześniej informację o koszcie rozwiązania zagadnienia dyskretnego (9.9.2).

Dodatkowe szczegóły

Wyjaśnimy tu jeszcze kilka szczegółów. Zauważmy najpierw, że warunki brzegowe

$$v_{0j} = v_{n+1,j} = 0 \quad (0 \leqslant j \leqslant n+1)$$

wynikają z (9.9.3) dla dowolnych $\hat{v}_{kj}$. Natomiast pozostałe takie warunki, tj.

$$v_{i0} = v_{i,n+1} = 0,$$

wymagają dołączenia dwóch dodatkowych równań:

$$\sum_{k=1}^{n}\hat{v}_{k0}\sin ik\varphi = \sum_{k=1}^{n}\hat{v}_{k,n+1}\sin ik\varphi = 0 \quad (0 \leqslant i \leqslant n+1).$$

Tak więc wektory

$$(\hat{v}_{10}, \hat{v}_{20}, \ldots, \hat{v}_{n0}), \quad (\hat{v}_{1,n+1}, \hat{v}_{2,n+1}, \ldots, \hat{v}_{n,n+1})$$

mają być ortogonalne względem wszystkich wierszy macierzy A z poniższego lematu. Ponieważ jest ona nieosobliwa, więc jej wiersze są niezależne liniowo i musi być

$$\hat{v}_{k0} = \hat{v}_{k,n+1} = 0 \quad (1 \leqslant k \leqslant n).$$

Lemat 9.9.1. *Macierz A stopnia n o elementach*

$$a_{kj} := \left(\frac{2}{n+1}\right)^{1/2} \sin\frac{kj\pi}{n+1} \quad (1 \leqslant k, j \leqslant n)$$

jest symetryczna i ortogonalna, czyli $A^2 = I$.

Dowód. Symetria macierzy A jest oczywista. Obliczmy dowolny element macierzy A^2:

$$\begin{aligned}(A^2)_{kj} &= \sum_{\nu=1}^{n} a_{k\nu}a_{\nu j} = \frac{2}{n+1}\sum_{\nu=1}^{n} \sin\frac{k\nu\pi}{n+1}\sin\frac{j\nu\pi}{n+1} = \\ &= \frac{1}{n+1}\sum_{\nu=1}^{n}\left[\cos\frac{\nu(k-j)\pi}{n+1} - \cos\frac{\nu(k+j)\pi}{n+1}\right] = \\ &= \frac{1}{n+1}\,\Re\sum_{\nu=0}^{n}(e^{i\nu\varphi} - e^{i\nu\theta}),\end{aligned}$$

gdzie $\varphi := (k-j)\pi/(n+1)$ i $\theta := (k+j)\pi/(n+1)$.

Jeśli $k = j$, to $\varphi = 0$ i $\theta = 2k\pi/(n+1)$. Ta ostatnia wielkość nie jest wielokrotnością liczby 2π, gdyż $1 \leqslant k \leqslant n$. Dlatego

$$(A^2)_{kk} = \frac{1}{n+1}\,\Re\left[n+1-\frac{e^{i(n+1)\theta}-1}{e^{i\theta}-1}\right] = 1.$$

Istotnie, $e^{i(n+1)\theta} = e^{2ik\pi} = 1$.

Jeśli $k \neq j$, to ani φ, ani θ nie jest wielokrotnością liczby 2π i obie sumy w ostatnim wyrażeniu dla $(A^2)_{kj}$ można wyrazić w znany sposób:

$$(A^2)_{kj} = \frac{1}{n+1}\,\Re\left[\frac{e^{i(n+1)\varphi}-1}{e^{i\varphi}-1} - \frac{e^{i(n+1)\theta}-1}{e^{i\theta}-1}\right].$$

Jeśli różnica $k-j$ jest parzysta, to suma $k+j$ też jest taka. Wtedy $e^{i(n+1)\varphi} = e^{i(n+1)\theta} = 1$ i $(A^2)_{kj} = 0$. W przeciwnym razie liczby $k-j$ i $k+j$ są nieparzyste, a wtedy $e^{i(n+1)\varphi} = e^{i(n+1)\theta} = -1$. Trzeba zatem wykazać, że

$$\Re\left[-\frac{2}{e^{i\varphi}-1} + \frac{2}{e^{i\theta}-1}\right] = 0.$$

W tym celu zauważmy najpierw, że jeśli $z \neq 0$, to

$$\Re \frac{1}{z} = \Re \frac{\bar{z}}{z\bar{z}} = \frac{\Re z}{|z|^2}.$$

Stąd wynika, że

$$\Re \frac{1}{e^{i\varphi} - 1} = \frac{\cos\varphi - 1}{(\cos\varphi - 1)^2 + \sin^2\varphi} = \frac{\cos\varphi - 1}{2 - 2\cos\varphi} = -\frac{1}{2}.$$

To samo dotyczy podobnego wyrażenia z θ zamiast φ. ■

ROZDZIAŁ 10

Programowanie liniowe i pokrewne zagadnienia

10.1. Wypukłość i nierówności liniowe

Podstawowe pojęcia

Wszystkie wektory i macierze, z którymi będziemy mieli tu do czynienia, są rzeczywiste (a nie zespolone), gdyż teoria nierówności wykorzystuje w istotny sposób uporządkowanie zbioru liczb rzeczywistych. Dla dwóch punktów (wektorów) $x, y \in \mathbb{R}^n$ piszemy

$$x \geqslant y \quad \text{wtedy i tylko wtedy, gdy} \quad x_i \geqslant y_i \quad (1 \leqslant i \leqslant n).$$

Podobnie definiujemy nierówności $x \leqslant y$, $x > y$ i $x < y$. Warto zwrócić uwagę, że nierówność $x > y$ nie jest równoważna temu, że $x \geqslant y$ i $x \neq y$, gdyż ta ostatnia relacja oznacza tylko, że $x_i \neq y_i$ dla pewnego i.

Układ m nieostrych nierówności liniowych z n zmiennymi,

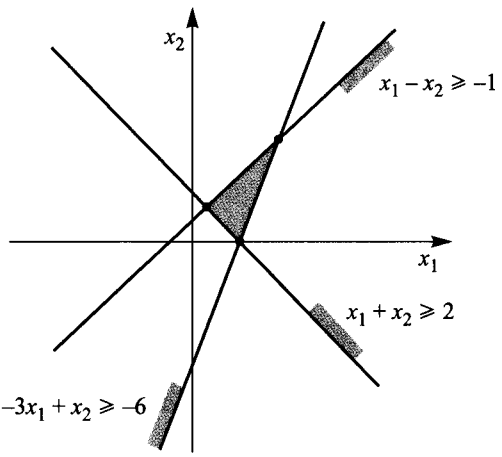
$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n \geqslant b_i \quad (1 \leqslant i \leqslant m),$$

można wyrazić w postaci

$$Ax \geqslant b, \tag{10.1.1}$$

gdzie A jest macierzą $m \times n$, $x \in \mathbb{R}^n$ i $b \in \mathbb{R}^m$. Oto podstawowe pytanie dotyczące takiego układu: czy jest on *niesprzeczny*? Inaczej mówiąc, czy

istnieje punkt x taki, że $Ax \geqslant b$? Jeśli tak, to chcielibyśmy mieć algorytmy wyznaczania go. Można też oczywiście rozważać układy, w których niektóre (lub wszystkie) nierówności są ostre, ale dla prostoty zajmujemy się dalej tylko układami (10.1.1).



RYS. 10.1. Zbiór rozwiązań układu (10.1.2)

Aby zorientować się, czego można tu się spodziewać, rozważmy prosty układ dla $n = 2$ i $m = 3$:

$$\begin{aligned} x_1 + x_2 &\geqslant 2, \\ x_1 - x_2 &\geqslant -1, \\ -3x_1 + x_2 &\geqslant -6, \end{aligned} \tag{10.1.2}$$

czyli

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \geqslant \begin{bmatrix} 2 \\ -1 \\ -6 \end{bmatrix}.$$

Punkty spełniające nierówność $x_1 + x_2 \geqslant 2$ leżą po jednej stronie prostej o równaniu $x_1 + x_2 = 2$ (lub na niej samej). Podobnie interpretujemy dwie pozostałe nierówności. Część wspólna trzech półpłaszczyzn jest trójkątem pokazanym na rys. 10.1. Układ (10.1.2) jest więc niesprzeczny. Łatwo natomiast wywnioskować z tego rysunku, że po zmianie wszystkich nierówności $\geqslant$ na $\leqslant$ układ stałby się *sprzeczny*. Można też zmodyfikować układ tak, żeby zbiór rozwiązań był nieograniczony; wystarczy np. zmienić w trzeciej nierówności -3 na -1. Rysunek pokazuje wreszcie, że przykładowy zbiór rozwiązań jest wypukły. To pojęcie i pewne jego własności podano już w podrozdz. 6.9. Tu przypomnimy niezbędne fakty.

Zbiory wypukłe. Powłoka wypukła

Zbiór K w przestrzeni liniowej jest *wypukły*, jeśli wraz ze swymi dwoma punktami zawiera odcinek, który je łączy. Inaczej mówiąc, jeśli $u, v \in K$, to dla każdego $\theta \in [0,1]$ jest $\theta u + (1-\theta)v \in K$. Kombinacja liniowa $\sum_{i=1}^{k} \theta_i u^{(i)}$ punktów $u^{(i)}$ jest *wypukła*, jeśli jej współczynniki są nieujemne, a ich sumą jest 1. *Powłoka wypukła* zbioru S, oznaczana symbolem $\mathrm{co}(S)$, jest zbiorem wszystkich kombinacji liniowych wypukłych punktów tego zbioru.

Twierdzenie 10.1.1. *Jeśli zbiór K jest wypukły, to każda kombinacja wypukła jego punktów należy do K.*

Dowód. Dowodzimy przez indukcję względem liczby m punktów w kombinacji. Przypadek $m = 1$ jest trywialny, a dla $m = 2$ twierdzenie wynika wprost z definicji zbioru wypukłego. Jeśli $m > 2$, to kombinację wypukłą z ostatnim współczynnikiem dodatnim możemy wyrazić tak:

$$\sum_{i=1}^{m} \lambda_i x^{(i)} = \lambda_m x^{(m)} + (1-\lambda_m) \sum_{i=1}^{m-1} \frac{\lambda_i}{1-\lambda_m} x^{(i)}.$$

Jeśli $\lambda_i \geqslant 0$, $\sum_{i=1}^{m} \lambda_i = 1$, to również współczynniki $\lambda_i/(1-\lambda_m)$ są nieujemne, a ich sumą jest 1. ■

Twierdzenie 10.1.2. *Iloczyn rodziny zbiorów wypukłych jest wypukły.*

Twierdzenie 10.1.3. *Zbiór rozwiązań układu nierówności liniowych jest wypukły.*

Dowód. Zbiór rozwiązań pojedynczej nierówności liniowej $a^{\mathsf{T}}x \geqslant \beta$ jest wypukły, gdyż dla $0 \leqslant \theta \leqslant 1$ jest

$$a^{\mathsf{T}}(\theta x + (1-\theta)y) = \theta a^{\mathsf{T}}x + (1-\theta)a^{\mathsf{T}}y \geqslant \theta\beta + (1-\theta)\beta = \beta.$$

Zbiór rozwiązań układu nierówności liniowych jest iloczynem takich zbiorów dla poszczególnych nierówności i teza wynika z tw. 10.1.2. ■

Twierdzenie 10.1.4. *Powłoka wypukła zbioru S jest najmniejszym zbiorem wypukłym zawierającym S.*

Dowód. Niech K będzie tą powłoką wypukłą, a T dowolnym zbiorem wypukłym, zawierającym S. Każdy element x zbioru K wyraża się w postaci $x = \sum_{i=1}^{n} \lambda_i x^{(i)}$, gdzie $x^{(i)} \in S$, $\lambda_i \geqslant 0$ i $\sum_{i=1}^{n} \lambda_i = 1$. Oczywiście, punkty $x^{(i)}$ należą także do T. Ponieważ ten zbiór jest wypukły, więc $x \in T$. Dlatego $K \subseteq T$. ■

Twierdzenie 10.1.5. *Jeśli X jest zbiorem domkniętym wypukłym w $\mathbb{R}^n$, a $p \notin X$, to dla pewnego $v \neq 0$ zachodzi nierówność*

$$\langle v, p\rangle < \inf_{x \in X} \langle v, x\rangle.$$

Dowód. Niech S będzie sferą domkniętą o środku w p i promieniu ρ tak dużym, aby iloczyn $K := S \cap X$ był niepusty. Zbiór K jest zwarty i wypukły. Funkcja $x \mapsto \|x - p\|$ jest ciągła, wobec czego osiąga w pewnym punkcie $\xi \in K$ minimum. Z założenia wynika, że $p \notin K$, czyli $v := \xi - p \neq 0$. Jeśli $x \in X$ i $0 < \theta < 1$, to $y := \theta x + (1-\theta)\xi \in X$. Zachodzi nierówność $\|y - p\| \geqslant \|v\|$, gdyż jeśli $y \in K$, to jest tak z definicji ξ, a dla $y \notin S$ jest $\|y - p\| > \rho \geqslant \|\xi - p\|$. Wobec tej nierówności

$$\begin{aligned}\|v\|^2 &\leqslant \|\theta x + (1-\theta)\xi - p\|^2 = \|v + \theta(x - \xi)\|^2 = \\ &= \|v\|^2 + 2\theta\langle v, x - \xi\rangle + \theta^2\|x - \xi\|^2.\end{aligned}$$

Stąd wynika, że $0 \leqslant 2\langle v, x - \xi\rangle + \theta\|x - \xi\|^2$. Dla θ dążącego do 0 daje to nierówność $0 \leqslant \langle v, x - \xi\rangle$, czyli

$$0 \leqslant \langle v, x - p + p - \xi\rangle = \langle v, x - p - v\rangle = \langle v, x - p\rangle - \|v\|^2$$

i ostatecznie

$$0 < \|v\|^2 \leqslant \langle v, x\rangle - \langle v, p\rangle,$$

a to jest nawet więcej niż należało udowodnić. ■

Twierdzenie 10.1.5 jest prawdziwe w przestrzeniach Hilberta, a po odpowiednich modyfikacjach także w przestrzeniach jeszcze ogólniejszych.

Pewne zastosowanie geometryczne tego twierdzenia wiąże się z pojęciem *półprzestrzeni domkniętej* w $\mathbb{R}^n$, tj. zbioru

$$\{x\colon \langle a, x\rangle \geqslant \lambda\},$$

gdzie $a \in \mathbb{R}^n$, $a \neq 0$ i $\lambda \in \mathbb{R}$.

Twierdzenie 10.1.6. *Każdy zbiór domknięty wypukły w $\mathbb{R}^n$ jest iloczynem wszystkich zawierających go półprzestrzeni domkniętych.*

Dowód. Jest oczywiste, że zbiór domknięty wypukły X jest zawarty w iloczynie zawierających go półprzestrzeni wypukłych. Z drugiej strony, niech p będzie punktem spoza X. Na mocy tw. 10.1.5 istnieje wektor $v \neq 0$ taki, że $\langle v, p\rangle < \lambda := \inf_{x \in X}\langle v, x\rangle$. Stąd wynika, że $X \subseteq \{x\colon \langle v, x\rangle \geqslant \lambda\}$, ale ta półprzestrzeń nie zawiera punktu p. ■

Twierdzenie 10.1.7. *Jeśli X jest zbiorem domkniętym wypukłym, a Y zbiorem zwartym wypukłym w $\mathbb{R}^n$ i jeśli te zbiory są rozłączne, to istnieje punkt $v \in \mathbb{R}^n$ taki, że*

$$\inf_{x\in X}\langle v,x\rangle > \sup_{y\in Y}\langle v,y\rangle.$$

Dowód. Wykażemy najpierw, że zbiór $Z := X - Y$ (złożony z różnic $x - y$, gdzie $x \in X$, $y \in Y$) jest domknięty. Niech będzie $z_k := x_k - y_k \in Z$ i $\{z_k\} \to z$. Dzięki zwartości zbioru Y pewien podciąg $\{y_{k_i}\}$ jest zbieżny do punktu $y \in Y$. Wtedy $\{z_{k_i}\} \to z$ i $\{x_{k_i}\} \to z + y$. Ponieważ zbiór X jest domknięty, więc $z+y \in X$ i $z \in X-Y$. Zauważmy następnie, że $0 \notin X-Y$, gdyż w przeciwnym razie zbiory X i Y miałyby wspólny punkt. Z wypukłości zbiorów X i Y wynika, że Z też jest taki. Stosując do Z i punktu $p := 0$ tw. 10.1.5, wnioskujemy, że istnieje v takie, iż

$$\langle v,0\rangle < \inf_{x\in X,\, y\in Y}\langle v,x-y\rangle.$$

Jeśli ε oznacza prawą stronę tej nierówności, to $\langle v, x - y\rangle \geqslant \varepsilon > 0$ dla dowolnych $x \in X$ i $y \in Y$. Stąd już wynika teza twierdzenia. ■

Punkty ekstremalne

Punkt x zbioru wypukłego K jest jego *punktem ekstremalnym*, jeśli nie można go przedstawić w postaci $x = \theta y+(1-\theta)z$, gdzie $0 < \theta < 1$ i y, $z \in K$. Innymi słowy, nie jest to punkt wewnętrzny żadnego odcinka zawartego w K. Mówiąc jeszcze inaczej, x jest punktem ekstremalnym zbioru K, jeśli usunięcie zeń tego punktu daje również zbiór wypukły. W szczególności jedynymi punktami ekstremalnymi sześcianu są jego wierzchołki, natomiast dla sfery są to wszystkie punkty jej powierzchni.

Następne twierdzenie jest wariantem *twierdzenia Kreina-Milmana*, dostosowanym do przestrzeni skończeniewymiarowych (zob. np. Royden [1968] lub Rudin [*2002]).

Twierdzenie 10.1.8 (Krein-Milman). *Zbiór zwarty wypukły w przestrzeni n-wymiarowej jest domknięciem powłoki wypukłej zbioru jego punktów ekstremalnych.*

Dowód. Jeśli $n = 1$, to zbiór zwarty wypukły jest przedziałem domkniętym i ograniczonym. Jego punktami ekstremalnymi są końce przedziału. Twierdzenie jest w tym przypadku prawdziwe. Załóżmy teraz, że jest tak również dla przestrzeni wymiaru $< n$. Niech K będzie zbiorem zwartym wypukłym w przestrzeni n-wymiarowej, E – zbiorem jego punktów ekstremalnych, a H – powłoką wypukłą zbioru E. Trzeba udowodnić, że $\bar{H} = K$.

Ponieważ K jest wypukły i $E \subseteq K$, więc $H \subseteq K$. Ponieważ K jest domknięty, więc $\bar{H} \subseteq K$. Pozostaje wykazać, że $\bar{H}$ nie jest podzbiorem właściwym zbioru K. Przypuśćmy, że istnieje $p \in K \setminus \bar{H}$. Dopuszczalne jest przesunięcie zbioru K, więc można założyć, że $p = 0$. Wtedy $0 \notin \bar{H}$ i na mocy tw. 6.9.10 istnieje wektor v taki, że $\langle v, u\rangle < 0$ dla każdego $u \in \bar{H}$. Niech będzie

$$c := \sup_{x \in K} \langle v, x\rangle.$$

Ponieważ $0 \in K$, więc $c \geqslant 0$. Ten kres górny jest osiągnięty, gdyż zbiór K jest zwarty. Wobec tego zbiór

$$K' := \{x \in K\colon\ \langle v, x\rangle = c\}$$

jest niepusty. Jest on również zwarty i wypukły, a jego wymiar nie przewyższa $n-1$, gdyż zawiera się w hiperpłaszczyźnie. Z założenia indukcyjnego wynika, że K' ma co najmniej jeden punkt ekstremalny z. Wykażemy, że $z \in E$. Przypuśćmy, że tak nie jest: $z = \theta z_1 + (1-\theta)z_2$, gdzie $0 < \theta < 1$ i $z_1, z_2 \in K$. Wtedy

$$c = \langle v, z\rangle = \theta\langle v, z_1\rangle + (1-\theta)\langle v, z_2\rangle \leqslant \theta c + (1-\theta)c = c.$$

Stąd $\langle v, z_1\rangle = \langle v, z_2\rangle = c$ i $z_1, z_2 \in K'$. Ponieważ jednak z jest punktem ekstremalnym zbioru K', więc $z_1 = z_2$ i, wbrew przypuszczeniu, $z \in E$. ■

Znaczenie punktów ekstremalnych w zagadnieniach optymalizacji wynika stąd, że szukając minimum funkcji liniowej na zbiorze zwartym wypukłym, można ograniczyć się do tych jego punktów.

Twierdzenie 10.1.9. *Jeśli K jest zbiorem zwartym wypukłym w $\mathbb{R}^n$, a f funkcjonałem liniowym na tej przestrzeni, to jego minimum i maksimum w K są osiągnięte w punktach ekstremalnych tego zbioru.*

Dowód. Niech będzie $c := \sup_{x \in K} f(x)$. Ponieważ funkcjonał f jest ciągły, a zbiór K zwarty, więc zbiór $K' := \{x \in K\colon\ f(x) = c\}$ jest niepusty. Jest on także zwarty i wypukły. Na mocy tw. 10.1.8 ma on punkt ekstremalny z. Rozumując podobnie jak w dowodzie tamtego twierdzenia wykazujemy, że jest to również punkt ekstremalny zbioru K. Część twierdzenia dotycząca minimum funkcjonału wynika z poprzedniej przez zbadanie maksimum funkcjonału $-f$. ■

Zadania 10.1

1. Niech U będzie zbiorem zwartym w $\mathbb{R}^n$. Udowodnić, że jeśli układ nierówności liniowych $\langle u, x\rangle > 0$ $(u \in U)$ jest sprzeczny, to istnieje jego sprzeczny podukład zawierający co najwyżej $n+1$ nierówności.

2. Wykazać, że jeśli zbiory S i T są wypukłe, to tę samą własność mają zbiory λS, $S+T$ i $S-T$.

3. Udowodnić, że jeśli zbiór K jest wypukły, a L jest odwzorowaniem liniowym, to zbiór $\{L(x)\colon\ x \in K\}$ jest wypukły.

4. Czy zbiór ograniczony może mieć dopełnienie wypukłe?

5. Wykazać, że domknięcie zbioru wypukłego jest wypukłe.

6. Niech K będzie zbiorem wypukłym, p jego punktem wewnętrznym, a q jego dowolnym punktem. Wykazać, że jeśli $0 < \theta < 1$, to $\theta p + (1-\theta)q$ jest punktem wewnętrznym zbioru K.

7. Udowodnić, że dowolny zbiór domknięty wypukły jest iloczynem wszystkich zawierających go półprzestrzeni otwartych, czyli zbiorów $\{x\colon\ \langle a, x\rangle > \lambda\}$.

8. Udowodnić, że dowolny zbiór wypukły na płaszczyźnie jest sumą wszystkich trójkątów, których wierzchołki leżą w tym zbiorze.

9. Niech X będzie zbiorem domkniętym wypukłym w przestrzeni Hilberta. Udowodnić, że jeśli $p \notin X$, $\xi \in X$ i $\|p-\xi\| = \operatorname{dist}(p, X)$, to $\langle p-\xi, x-\xi\rangle \leqslant 0$ dla każdego $x \in X$.

10. Udowodnić, że jeśli K jest zbiorem domkniętym wypukłym w przestrzeni Hilberta, to dla każdego punktu $p \notin K$ istnieje w K dokładnie jeden najbliższy punkt k.

11. (cd.). Udowodnić, że odwzorowanie $p \mapsto k$ określone w poprzednim zadaniu jest takie, że jeśli $p_1 \mapsto k_1$ i $p_2 \mapsto k_2$, to $\|k_1 - k_2\| \leqslant \|p_1 - p_2\|$.

12. Udowodnić, że $p \in \operatorname{co}(X)$ wtedy i tylko wtedy, gdy $0 \in \operatorname{co}(X-p)$.

13. Wykazać, że $\operatorname{co}(\lambda S) = \lambda \operatorname{co}(S)$ i $\operatorname{co}(S+T) = \operatorname{co}(S) + \operatorname{co}(T)$.

14. Wykazać, że powłoka wypukła zbioru otwartego w $\mathbb{R}^n$ jest otwarta.

15. Pokazać na przykładzie, że powłoka wypukła zbioru domkniętego nie musi być domknięta.

16. Udowodnić, że jeśli zbiory X_i są wypukłe, to $\operatorname{co}(X_1 \cup \ldots \cup X_k)$ jest zbiorem wszystkich kombinacji wypukłych $\sum_{i=1}^{k} \theta_i x_i$, gdzie $x_i \in X_i$.

17. Wykazać, że jeśli zbiory S i T są zwarte i wypukłe, to te same własności ma zbiór $\operatorname{co}(S \cup T)$.

18. Wykazać, że jeśli X jest zbiorem ograniczonym w $\mathbb{R}^n$, to dla dowolnego p

$$\sup_{x \in X} \|p-x\| = \sup_{x \in \operatorname{co}(X)} \|p-x\|.$$

19. Niech zbiór $X \subset \mathbb{R}^n$ skończony zawiera co najmniej $n+2$ punkty. Wykazać, że można go wyrazić w postaci $X = X_1 \cup X_2$, gdzie $\operatorname{co}(X_1) \cap \operatorname{co}(X_2) = \emptyset$.

20. Niech $\overline{xy}$ oznacza odcinek łączący punkty $x, y \in \mathbb{R}^n$. Dla danego zbioru X_0 niech będzie

$$X_{k+1} := \bigcup \{\overline{xy}\colon\ x, y \in X_k\} \quad (k \geqslant 0).$$

Udowodnić, że $X_{2^n+1} = \operatorname{co}(X_0)$.

21. Udowodnić, że jeśli układ nierówności $\langle u, x\rangle > 0$ dla $u \in U$ jest niesprzeczny, to tę samą własność ma układ $\langle u, x\rangle > 0$ dla $u \in \operatorname{co}(U)$.

22. Wykazać, że jeśli U jest zbiorem zwartym wypukłym, a układ nierówności $\langle u, x\rangle > 0$ $(u \in U)$ jest sprzeczny, to tę samą własność ma układ tych nierówności ograniczony do punktów ekstremalnych u zbioru U.

10.2. Nierówności liniowe

Niech X będzie przestrzenią wektorową nad ciałem rzeczywistym $\mathbb{R}$ [1]. Będziemy się tu zajmować układami, skończonymi lub nie, nierówności liniowych $f_i(x) \geqslant \alpha_i$, gdzie każde f_i jest funkcjonałem liniowym, odwzorowującym X w $\mathbb{R}$.

Układy równań jednorodnych

Przypomnimy najpierw twierdzenie z algebry liniowej, dotyczące układów równań liniowych jednorodnych; potem podamy jego odpowiednik dla nierówności. Symbolem $\mathcal{N}(f)$ będziemy oznaczać *jądro* funkcjonału f, tj. zbiór

$$\{x \in X\colon\ f(x) = 0\}.$$

Znany już symbol $\operatorname{span}\{f_1, f_2, \ldots, f_n\}$ oznacza zbiór wszystkich kombinacji liniowych elementów $f_1, f_2, \ldots, f_n$.

Twierdzenie 10.2.1. *Następujące własności funkcjonałów liniowych są równoważne:*

1. $\mathcal{N}(f) \supset \bigcap_{i=1}^{m} \mathcal{N}(f_i)$,

2. $f \in \operatorname{span}\{f_1, f_2, \ldots, f_m\}$.

Dowód. Implikacja **2⇒1** jest oczywista. Istotnie, jeśli $f = \sum_{i=1}^{m} \lambda_i f_i$ i $x \in \bigcap_{i=1}^{m} \mathcal{N}(f_i)$, to $f_i(x) = 0$ dla każdego i, a zatem $f(x) = 0$.

Sprawdźmy teraz indukcyjnie względem m implikację **1⇒2**. Niech będzie $m = 1$ i $\mathcal{N}(f) \supset \mathcal{N}(f_1)$. Jeśli $f_1 \equiv 0$, to $\mathcal{N}(f_1) = X$, a zatem $\mathcal{N}(f) = X$ i $f = 0$. Wtedy $f \in \operatorname{span}\{f_1\}$. Jeśli $f_1 \not\equiv 0$, to wybieramy punkt y taki, że $f_1(y) = 1$. Dla $x \in X$ jest $f_1(x - f_1(x)y) = f_1(x) - f_1(x)f_1(y) = 0$, więc $x - f_1(x)y \in \mathcal{N}(f_1)$. Stąd i z założenia wynika, że $x - f_1(x)y \in \mathcal{N}(f)$. Dlatego $f(x) - f_1(x)f(y) = 0$, czyli $f = f(y)f_1$, a zatem $f \in \operatorname{span}\{f_1\}$.

Załóżmy teraz, że twierdzenie jest prawdziwe dla pewnego m naturalnego. Niech będzie

$$\mathcal{N}(f) \supset \bigcap_{i=1}^{m+1} \mathcal{N}(f_i)$$

[1] W dalszym ciągu będzie potrzebna tylko przestrzeń $X = \mathbb{R}^n$ (*przyp. tłum.*).

i $Y := \mathcal{N}(f_{m+1})$. Jest to pewna podprzestrzeń przestrzeni X. Jeśli $f \mid Y$ oznacza zwężenie funkcjonału f do zbioru Y (zob. podrozdz. 6.5), to

$$\mathcal{N}(f \mid Y) \supset \bigcap_{i=1}^{m} \mathcal{N}(f_i \mid Y).$$

Z założenia indukcyjnego

$$f \mid Y = \sum_{i=1}^{m} \lambda_i f_i \mid Y$$

dla odpowiednich λ_i. Mamy teraz dwie równoważne relacje:

$$\Big(f - \sum_{i=1}^{m} \lambda_i f_i\Big) \mid Y = 0, \quad \mathcal{N}\Big(f - \sum_{i=1}^{m} \lambda_i f_i\Big) \supset \mathcal{N}(f_{m+1}).$$

Korzystając z pierwszej części dowodu (dla $m = 1$), wnioskujemy, że dla pewnego λ_{m+1} jest

$$f - \sum_{i=1}^{m} \lambda_i f_i = \lambda_{m+1} f_{m+1}. \qquad \blacksquare$$

Układy nierówności liniowych

Aby otrzymać odpowiednik tw. 10.2.1 dla nierówności liniowych, trzeba zastąpić jądra funkcjonałów *półprzestrzeniami*

$$f^+ := \{x \in X\colon\ f(x) \geqslant 0\},$$

a podprzestrzeń $\operatorname{span}\{f_1, f_2, \ldots, f_m\}$ – *stożkiem*

$$\mathcal{C}\{f_1, f_2, \ldots, f_m\} := \Big\{\sum_{i=1}^{m} \lambda_i f_i\colon\ \lambda_i \geqslant 0\Big\}.$$

Twierdzenie 10.2.2 (Farkas). *Następujące własności funkcjonałów liniowych są równoważne:*

1. $f^+ \supset \bigcap_{i=1}^{m} f_i^+$,
2. $f \in \mathcal{C}\{f_1, f_2, \ldots, f_m\}$.

Dowód. Aby sprawdzić implikację $\mathbf{2} \Rightarrow \mathbf{1}$, załóżmy, że $f = \sum_{i=1}^{m} \lambda_i f_i$, gdzie $\lambda_i \geqslant 0$. Jeśli $x \in \bigcap_{i=1}^{m} f_i^+$, to $f_i(x) \geqslant 0$ dla wszystkich i, a wtedy oczywiście także $f(x) \geqslant 0$.

Udowodnimy teraz, że jeśli nie zachodzi **2**, to nie zachodzi także **1**; ograniczymy się przy tym do przypadku, gdy $X = \mathbb{R}^n$. Wtedy każdy funk-

cjonał liniowy wyraża się przez iloczyn skalarny. Niech będzie zatem $f(x) := \langle x, w\rangle$ i $f_i(x) := \langle x, w_i\rangle$ (x, w, $w_i \in \mathbb{R}^n$). Stożkowi $C := \mathcal{C}\{f_1, f_2, \ldots, f_m\}$ odpowiada zbiór wypukły

$$W := \Big\{\sum_{i=1}^{m} \lambda_i w_i\colon\ \lambda_i \geqslant 0\Big\} \subset \mathbb{R}^n.$$

Załóżmy, że $f \notin C$. Wtedy $w \notin W$. Z twierdzenia 10.1.5 (ze zmianą X na W i p na w) wynika istnienie takiego wektora x, że

$$\langle x, w\rangle < \inf_{z\in W} \langle x, z\rangle.$$

Niech k będzie tym kresem dolnym. Ponieważ $0 \in W$, więc $k \leqslant 0$. Jeśli $k < 0$, to istnieje takie z, że $k \leqslant \langle x, z\rangle < 0$. Dla każdego $t > 0$ jest $tz \in W$, ale dla dostatecznie dużych t mamy nierówność $\langle x, tz\rangle < k$, co przeczy definicji kresu k. Tak więc $k = 0$ i dla każdego $z \in W$ jest

$$\langle x, w\rangle < 0 \leqslant \langle x, z\rangle.$$

Dla odpowiednich z wynika stąd, że $f(x) < 0 \leqslant f_i(x)$, czyli własność **1** jest fałszywa. ■

Dla $X = \mathbb{R}^n$, dzięki postaci funkcjonałów liniowych w tej przestrzeni, można sformułować twierdzenie Farkasa w języku macierzy:[2)]

Twierdzenie 10.2.3. *Następujące własności macierzy kwadratowej A i wektora c są równoważne:*

1. *Dla każdego x z nierówności $Ax \geqslant 0$ wynika, że $c^{\mathsf{T}}x \geqslant 0$.*
2. *Dla pewnego wektora $y \geqslant 0$ jest $c = A^{\mathsf{T}}y$.*

Układy sprzeczne i niesprzeczne

Zajmiemy się teraz układami nierówności niejednorodnych. Będziemy stosować następujące określenie: układ nierówności $f_i(x) \geqslant \alpha_i$ jest *konsekwencją* układu $g_i(x) \geqslant \beta_i$, jeśli każde rozwiązanie drugiego układu spełnia także pierwszy, tzn.

$$\{x\colon\ g_i(x) \geqslant \beta_i \text{ dla każdego } i\} \subseteq \{x\colon\ f_i(x) \geqslant \alpha_i \text{ dla każdego } i\}$$

(warto porównać z tą relacją własność **1** z tw. 10.2.2).

[2)] Ściślej, poniższe twierdzenie udowodnił wcześniej Minkowski w 1896 r. (*przyp. tłum.*).

Twierdzenie 10.2.4 (Farkas). *Jeśli nierówność liniowa* $f(x) \geqslant \alpha$ *jest konsekwencją niesprzecznego układu nierówności liniowych*

$$g_i(x) \geqslant \beta_i \quad (1 \leqslant i \leqslant n),$$

to dla pewnych $\theta_i \geqslant 0$ *jest*

$$f = \sum_{i=1}^{n} \theta_i g_i, \quad \sum_{i=1}^{n} \theta_i \beta_i \geqslant \alpha.$$

Dowód. Rozważmy układ nierówności

$$g_i(x) - \lambda\beta_i \geqslant 0 \quad (1 \leqslant i \leqslant n), \quad \lambda > 0. \tag{10.2.1}$$

Jeśli para (x, λ) je spełnia, to $g_i(x/\lambda) \geqslant \beta_i$ i wobec założenia jest $f(x/\lambda) \geqslant \alpha$, czyli

$$f(x) - \lambda\alpha \geqslant 0 \tag{10.2.2}$$

i (10.2.2) jest konsekwencją układu (10.2.1). Zmodyfikujmy teraz ostatni warunek w (10.2.1):

$$g_i(x) - \lambda\beta_i \geqslant 0 \quad (1 \leqslant i \leqslant n), \quad \lambda \geqslant 0. \tag{10.2.3}$$

Jeśli nierówność (10.2.2) jest konsekwencją układu (10.2.3), to możemy zastosować tw. 10.2.2 do par $(f; -\lambda)$ interpretowanych jako wektory. Ponieważ zachodzi tam **1**, więc na mocy równoważnej własności **2**

$$(f; -\alpha) = \sum_{i=1}^{n} \theta_i (g_i; -\beta_i) + \theta_0 (0; 1), \quad \theta_i \geqslant 0 \quad (1 \leqslant i \leqslant n).$$

Stąd

$$f = \sum_{i=1}^{n} \theta_i g_i, \quad \alpha = \sum_{i=1}^{n} \theta_i \beta_i - \theta_0 \leqslant \sum_{i=1}^{n} \theta_i \beta_i.$$

To właśnie należało wykazać. Dowód nie jest jednak jeszcze skończony, gdyż – być może – (10.2.2) nie jest konsekwencją (10.2.3), choć wykazaliśmy, że jest konsekwencją (10.2.1). Wtedy istniałoby takie rozwiązanie $(u; \lambda)$ układu (10.2.3), które nie spełnia ani (10.2.1), ani (10.2.2). Musi zatem być $\lambda = 0$, a więc $g_i(u) \geqslant 0 > f(u)$. Z założeń twierdzenia wynika, że dla pewnego v jest $g_i(v) \geqslant \beta_i$. Wybierzmy liczbę $\lambda > 0$ tak, żeby było $f(u + \lambda v) < \lambda\alpha$. To jest możliwe, gdyż dla $\lambda \to 0$ prawa strona nierówności dąży do 0, a lewa do liczby ujemnej $f(u)$. Jest to sprzeczne z założeniami, gdyż $f(u/\lambda + v) < \alpha$, a jednocześnie $g_i(u\lambda + v) \geqslant g_i(v) \geqslant \beta_i$. ■

Warianty macierzowe twierdzeń

Z twierdzenia 10.2.4 wynika

Twierdzenie 10.2.5. *Jeśli układ* $Ax \geqslant b$ *jest niesprzeczny, a układ* $Ax \geqslant b$, $c^{\mathsf{T}}x < \alpha$ *jest sprzeczny, to układ*

$$A^{\mathsf{T}}y = c, \quad y^{\mathsf{T}}b \geqslant \alpha, \quad y \geqslant 0$$

jest niesprzeczny.

Twierdzenie 10.2.6. *Jeśli układ*

$$Ax = b, \quad x \geqslant 0 \tag{10.2.4}$$

jest sprzeczny, to układ

$$A^{\mathsf{T}}y \geqslant 0, \quad b^{\mathsf{T}}y < 0 \tag{10.2.5}$$

jest niesprzeczny.

Dowód Jeśli układ (10.2.4) jest sprzeczny, to $b \notin K := \{Ax\colon\ x \geqslant 0\}$. Do zbioru K, który jest domknięty i wypukły, stosuje się tw. 10.1.5. Istnieje zatem wektor y taki, że

$$\langle y, b\rangle < \inf_{x \geqslant 0} \langle y, Ax\rangle. \tag{10.2.6}$$

Ponieważ po prawej stronie może być $x = 0$, więc $\langle y, b\rangle < 0$. Pozostaje sprawdzić, że $A^{\mathsf{T}}y \geqslant 0$. W przeciwnym razie pewna składowa, np. i-ta, wektora $A^{\mathsf{T}}y$ byłaby ujemna. Niech wektor x ma składowe $x_j := \lambda\delta_{ij}$. Wtedy

$$\langle y, Ax\rangle = \sum_{j=1}^{n} (A^{\mathsf{T}}y)_j x_j = \lambda (A^{\mathsf{T}}y)_i \to -\infty \quad \text{dla} \quad \lambda \to +\infty.$$

Dlatego dla pewnych λ mielibyśmy nierówność $\langle y, Ax\rangle < \langle y, b\rangle$ sprzeczną z (10.2.6). Wykazaliśmy więc, że wektor y spełnia układ (10.2.5). ■

Twierdzenie 10.2.7. *Jeśli układ*

$$Ax \leqslant b, \quad x \geqslant 0 \tag{10.2.7}$$

jest sprzeczny, to układ

$$A^{\mathsf{T}}y \geqslant 0, \quad b^{\mathsf{T}}y < 0, \quad y \geqslant 0$$

jest niesprzeczny.

Dowód. Jeśli układ (10.2.7) jest sprzeczny, to tę samą własność ma układ

$$Ax + z = b, \quad x \geqslant 0, \quad z \geqslant 0,$$

czyli układ

$$\begin{bmatrix} A & I \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} = b, \quad \begin{bmatrix} x \\ z \end{bmatrix} \geqslant 0.$$

Na mocy tw. 10.2.6 układ

$$\begin{bmatrix} A^{\mathsf{T}} \\ I \end{bmatrix} y \geqslant 0, \quad b^{\mathsf{T}} y < 0$$

jest niesprzeczny, a to właśnie należało wykazać. ■

ZADANIA 10.2

1. Udowodnić, że w poniższych parach układów (oddzielonych średnikiem) tylko jeden jest niesprzeczny.

(a) $Ax = 0,\ x \geqslant 0,\ x \neq 0; \quad A^{\mathsf{T}} y > 0.$

(b) $Ax \leqslant 0,\ x \geqslant 0,\ x \neq 0; \quad A^{\mathsf{T}} y > 0,\ y > 0$ (tw. Ville'a, 1938).

(c) $Ax = b; \quad A^{\mathsf{T}} y = 0,\ b^{\mathsf{T}} y = 1.$

2. Udowodnić, że jeśli układ $Ax \geqslant 0,\ x \geqslant 0,\ x \neq 0$ jest sprzeczny, to układ $A^{\mathsf{T}} y < 0,\ y \geqslant 0$ jest niesprzeczny.

3. Udowodnić, że jeśli układ $Ax = 0,\ x > 0$ jest sprzeczny, to układ $Ay \leqslant 0$ jest niesprzeczny.

4. Udowodnić, że dla macierzy A rozmiaru $m \times n$ układ $Ax > 0$ jest niesprzeczny wtedy i tylko wtedy, gdy układ $A^{\mathsf{T}} y = 0,\ y > 0$ jest sprzeczny.

5. Wykazać, że dla każdej macierzy A rozmiaru $n \times (n+1)$ układ $Ax \geqslant 0,\ x \neq 0$ jest niesprzeczny.

6. Udowodnić, że jeśli nierówność $Ax \geqslant b$ nie ma rozwiązań, to dla pewnego $y \geqslant 0$ jest $A^{\mathsf{T}} y = 0$ i $b^{\mathsf{T}} y = 1$.

7. Udowodnić, że jeśli układ $Ax \geqslant b,\ x \geqslant 0$ nie ma rozwiązań, to układ $A^{\mathsf{T}} y \leqslant 0,\ b^{\mathsf{T}} y > 0,\ y \geqslant 0$ ma rozwiązania.

8. Niech będzie $P_n := \{x \in \mathbb{R}^n\colon x \geqslant 0,\ \sum_{i=1}^n x_i = 1\}$. Udowodnić, że dla dowolnej macierzy A rozmiaru $m \times n$ jest $Ax \geqslant 0$ dla pewnego $x \in P_n$ lub $A^{\mathsf{T}} y \leqslant 0$ dla pewnego $y \in P_m$.

9. (cd.). Udowodnić, że

$$\max_{x \in P_n} \min_{y \in P_m} y^{\mathsf{T}} A x \leqslant \min_{y \in P_m} \max_{x \in P_n} y^{\mathsf{T}} A x.$$

Wskazówka: Zacząć od nierówności $\min_y y^{\mathsf{T}} A x \leqslant \max_x y^{\mathsf{T}} A x$.

10. Udowodnić, że jeśli wszystkie elementy macierzy U rozmiaru $m \times n$ są równe 1, to dla wszystkich $x \in P_n$ i $y \in P_m$ jest $y^{\mathsf{T}}(A - \lambda U)x = y^{\mathsf{T}}Ax - \lambda$.

11. (*Twierdzenie minimaksowe z teorii gier*). Udowodnić, że

$$\max_{x \in P_n} \min_{y \in P_m} y^{\mathsf{T}}Ax = \min_{y \in P_m} \max_{x \in P_n} y^{\mathsf{T}}Ax.$$

Wskazówka: Jeśli nierówność z zad. 9 jest ostra, to wybrać liczbę λ leżącą między jej dwiema stronami, skorzystać z zad. 10 i zastosować zad. 8 do macierzy $A - \lambda U$.

12. Znaleźć warunek konieczny i dostateczny na ograniczoność zbioru

$$K := \{x \in \mathbb{R}^n \colon Ax \leqslant b\}.$$

10.3. Programowanie liniowe

Programowanie liniowe jest działem matematyki, który zajmuje się następującym zagadnieniem: znaleźć maksimum funkcji liniowej n zmiennych rzeczywistych w wielościanie wypukłym w przestrzeni $\mathbb{R}^n$. Przyjmujemy więc taką standardową postać zagadnień tego typu:

I postać standardowa zagadnienia programowania liniowego

Niech będzie $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ i $A \in \mathbb{R}^{m \times n}$. Znaleźć maksimum wyrażenia $c^{\mathsf{T}}x$ przy warunkach $x \in \mathbb{R}^n$, $Ax \leqslant b$ i $x \geqslant 0$.

Przypominamy, że nierówność $x \geqslant 0$ jest równoważna temu, że wszystkie składowe wektora x są nieujemne. Podobnie, nierówność $Ax \leqslant b$ oznacza, że

$$\sum_{j=1}^{n} a_{ij}x_j \leqslant b_i \quad (1 \leqslant i \leqslant m). \tag{10.3.1}$$

Zastosujemy tu kilka nowych pojęć. *Zbiorem dopuszczalnym* wektorów jest

$$K := \{x : \mathbb{R}^n \colon Ax \leqslant b,\ x \geqslant 0\}.$$

Szukamy zatem *wartości rozwiązania*, równej z definicji

$$v := \sup_{x \in K} c^{\mathsf{T}}x.$$

Punktem dopuszczalnym jest dowolny element zbioru K. *Rozwiązaniem* (*punktem optymalnym*) jest każdy punkt $x \in K$ taki, że $c^{\mathsf{T}}x = v$. Funkcja $x \mapsto c^{\mathsf{T}}x = \sum_{j=1}^{n} c_j x_j$ jest *funkcją celu*. Zagadnienie jest w pełni określone przez dane A, b i c, więc mówimy o *zagadnieniu* (A, b, c) *programowania liniowego*.

Redukcja zagadnień do postaci standardowej

Prawie każde zadanie szukania ekstremum funkcji liniowej ze zmiennymi podlegającymi jakimś ograniczeniom można sprowadzić do standardowej postaci podanej wyżej. Są tu przydatne następujące wskazówki:

1. Szukanie minimum wyrażenia $c^{\mathsf{T}}x$ sprowadza się do szukania maksimum wyrażenia $-c^{\mathsf{T}}x$.
2. Warunek $a^{\mathsf{T}}x \geqslant \beta$ jest równoważny warunkowi $-a^{\mathsf{T}}x \leqslant -\beta$.
3. Warunek $a^{\mathsf{T}}x = \beta$ jest równoważny warunkom $a^{\mathsf{T}}x \leqslant \beta$ i $-a^{\mathsf{T}}x \leqslant -\beta$.
4. Warunek $|a^{\mathsf{T}}x| \leqslant \beta$ jest równoważny warunkom $a^{\mathsf{T}}x \leqslant \beta$ i $-a^{\mathsf{T}}x \leqslant \beta$.
5. Dodanie stałej do funkcji celu nie zmienia rozwiązania: suma $c^{\mathsf{T}}x + \beta$ osiąga maksimum w tym samym punkcie x co $c^{\mathsf{T}}x$.
6. Jeśli w zagadnieniu nie żąda się, aby zmienna x_j była nieujemna, to można ją zastąpić różnicą $u_j - x_j$ zmiennych, które mają być nieujemne.

Przykład 10.3.1. Sprowadzić do standardowej postaci następujące zagadnienie: obliczyć minimum sumy

$$7x_1 - x_2 + x_3 - 4$$

przy warunkach

$$x_1 + x_2 - x_3 \geqslant 2, \quad 3x_1 + 4x_2 + x_3 = 6, \quad |x_1 - 2x_2 + 3x_3| \leqslant 5,$$
$$x_1 \geqslant 0, \quad x_2 \leqslant 0.$$

Rozwiązanie. Zgodnie z wcześniejszymi sugestiami wprowadzamy nowe zmienne: $u_1 = x_1$, $u_2 = -x_2$ i $u_3 - u_4 = x_3$. Wtedy zagadnienie polega na znalezieniu maksimum sumy

$$-7u_1 - u_2 - u_3 + u_4$$

przy warunkach

$$-u_1 + u_2 + u_3 - u_4 \leqslant -2, \quad 3u_1 - 4u_2 + u_3 - u_4 \leqslant 6,$$
$$-3u_1 + 4u_2 - u_3 + u_4 \leqslant -6, \quad u_1 + 2u_2 + 3u_3 - 3u_4 \leqslant 5,$$
$$-u_1 - 2u_2 - 3u_3 + 3u_4 \leqslant 5, \quad u_1 \geqslant 0,\ u_2 \geqslant 0,\ u_3 \geqslant 0,\ u_4 \geqslant 0.$$

(Zauważmy, że drugi pierwotny warunek pozwala wyrazić zmienną x_3, której znak jest dowolny, przez x_1 i x_2. Daje to prostsze zagadnienie standardowe, tylko z dwiema zmiennymi i trzema warunkami typu (10.3.1), ale wymaga wykonania pewnych wstępnych obliczeń). ■

Zagadnienie (A, b, c) może nie mieć rozwiązania. Istotnie, po pierwsze, zbiór dopuszczalny K może być pusty. Po drugie, jeśli jest on niepusty i nieograniczony, to funkcja celu może nie mieć kresu górnego skończonego. Tylko wtedy, gdy zbiór K jest niepusty i ograniczony, istnieje co najmniej jedno rozwiązanie. Wynika to stąd, że wtedy zbiór K jest zwarty, a więc funkcja celu (ciągła względem swych zmiennych) osiąga na nim kres górny.

Zagadnienie dualne

Z każdym zagadnieniem (A, b, c) programowania liniowego można związać *zagadnienie dualne* $(-A^{\mathsf{T}}, -c, -b)$. Tak na przykład obliczenie maksimum sumy

$$3x_1 - 2x_2$$

przy warunkach

$$7x_1 + x_2 \leqslant 18, \quad -3x_1 + 5x_2 \leqslant 25,$$
$$6x_1 - x_2 \leqslant 13, \quad x_1 \geqslant 0, \; x_2 \geqslant 0$$

ma następujące zagadnienie dualne: znaleźć maksimum sumy

$$-18y_1 - 25y_2 - 13y_3$$

przy warunkach

$$-7y_1 + 3y_2 - 6y_3 \leqslant -3, \quad -y_1 - 5y_2 + y_3 \leqslant 2,$$
$$y_1 \geqslant 0, \quad y_2 \geqslant 0, \quad y_3 \geqslant 0.$$

Relacje między zagadnieniem pierwotnym i dualnym są przedmiotem *teorii dualności*. Poznamy teraz jej pewne wyniki.

Twierdzenie 10.3.2. *Jeśli x jest punktem dopuszczalnym dla zagadnienia (A, b, c), a y takim punktem dla zagadnienia dualnego $(-A^{\mathsf{T}}, -c, -b)$, to*

$$c^{\mathsf{T}}x \leqslant y^{\mathsf{T}}Ax \leqslant b^{\mathsf{T}}y.$$

Jeśli występują tu znaki równości, to x i y są rozwiązaniami odpowiednich zagadnień.

Dowód. Z definicji zagadnień wynika, że

$$x \geqslant 0, \quad Ax \leqslant b, \quad y \geqslant 0, \quad -A^{\mathsf{T}}y \leqslant -c,$$

a stąd wnioskujemy, że

$$c^{\mathsf{T}}x \leqslant (A^{\mathsf{T}}y)^{\mathsf{T}}x = y^{\mathsf{T}}Ax \leqslant y^{\mathsf{T}}b = b^{\mathsf{T}}y.$$

Wartości v_1 i v_2 dla tych zagadnień spełniają zatem nierówności

$$c^{\mathsf{T}}x \leqslant v_1 \leqslant b^{\mathsf{T}}y, \quad -b^{\mathsf{T}}y \leqslant v_2 \leqslant -c^{\mathsf{T}}x. \tag{10.3.2}$$

Jeśli $c^{\mathsf{T}}x = b^{\mathsf{T}}y$, to oczywiście $c^{\mathsf{T}}x = v_1 = b^{\mathsf{T}}y = -v_2$. ■

Pierwsza z nierówności (10.3.2) daje obustronne oszacowanie wartości v_1 zagadnienia (A, b, c).

Twierdzenie 10.3.3. *Jeśli dane zagadnienie programowania liniowego i dualne do niego mają punkty dopuszczalne, to oba zagadnienia mają rozwiązania, a wartości różnią się tylko znakiem.*

Dowód. Wobec tw. 10.3.2 wystarczy udowodnić istnienie takich x i y, że

$$x \geqslant 0, \quad Ax \leqslant b, \quad y \geqslant 0, \quad -A^{\mathsf{T}}y \leqslant -c, \quad c^{\mathsf{T}}x \geqslant b^{\mathsf{T}}y.$$

Trzeba więc wykazać, że następujący układ nierówności liniowych jest niesprzeczny:

$$\begin{bmatrix} A & 0 \\ 0 & -A^{\mathsf{T}} \\ -c^{\mathsf{T}} & b^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \leqslant \begin{bmatrix} b \\ -c \\ 0 \end{bmatrix}, \quad \begin{bmatrix} x \\ y \end{bmatrix} \geqslant \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Przypuśćmy, że – przeciwnie – jest on sprzeczny. Wtedy na mocy tw. 10.2.7 następujący układ jest niesprzeczny:

$$\begin{bmatrix} A^{\mathsf{T}} & 0 & -c \\ 0 & -A & b \end{bmatrix} \begin{bmatrix} u \\ v \\ \lambda \end{bmatrix} \geqslant \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} b^{\mathsf{T}} & -c^{\mathsf{T}} & 0 \end{bmatrix} \begin{bmatrix} u \\ v \\ \lambda \end{bmatrix} < 0, \quad \begin{bmatrix} u \\ v \\ \lambda \end{bmatrix} \geqslant 0;$$

u i v są tu wektorami, a λ liczbą. Jeśli (u, v, λ) spełnia ten układ, to

$$\begin{aligned} A^{\mathsf{T}}u - \lambda c \geqslant 0, \quad -Av + \lambda b \geqslant 0, \quad b^{\mathsf{T}}u - c^{\mathsf{T}}v < 0, \\ u \geqslant 0, \qquad v \geqslant 0, \qquad \lambda \geqslant 0. \end{aligned}$$

Niech będzie najpierw $\lambda > 0$. Wtedy $\lambda^{-1}v$ jest punktem dopuszczalnym dla zagadnienia (A, b, c), a $\lambda^{-1}u$ – takim punktem dla zagadnienia dualnego $(-A^{\mathsf{T}}, -c, -b)$. Wtedy z tw. 10.3.2 wynika, że $c^{\mathsf{T}}(\lambda^{-1}v) \leqslant b^{\mathsf{T}}(\lambda^{-1}u)$, czyli $b^{\mathsf{T}}u - c^{\mathsf{T}}v \geqslant 0$, wbrew definicji u i v.

Jeśli $\lambda = 0$, to $a^{\mathsf{T}}u \geqslant 0 \geqslant Av$. Niech x i y będą punktami dopuszczalnymi odpowiednio dla zagadnień pierwotnego i dualnego. I tym razem otrzymujemy nierówność sprzeczną z określeniem u i v:

$$c^{\mathsf{T}}v \leqslant (A^{\mathsf{T}}y)^{\mathsf{T}}v = y^{\mathsf{T}}Av \leqslant 0 \leqslant (A^{\mathsf{T}}u)^{\mathsf{T}}x = u^{\mathsf{T}}Ax \leqslant u^{\mathsf{T}}b = b^{\mathsf{T}}u. \quad \blacksquare$$

TWIERDZENIE 10.3.4. *Jeśli jedno z dwóch zagadnień programowania liniowego – pierwotne lub dualne – ma rozwiązanie, to drugie również je ma.*

Dowód. Ponieważ zagadnienie dualne do dualnego jest identyczne z pierwotnym, więc wystarczy wykazać, że jeśli zagadnienie $(-A^{\mathsf{T}}, -c, -b)$ ma rozwiązanie, np. y_0, to istnieje rozwiązanie zagadnienia pierwotnego. Przyjmujemy więc, że układ nierówności

$$-A^{\mathsf{T}}y \leqslant -c, \quad y \geqslant 0, \quad -b^{\mathsf{T}}y > -b^{\mathsf{T}}y_0$$

jest sprzeczny. Oczywiście ten układ bez ostatniej nierówności ma rozwiązanie. Pełny układ piszemy w postaci

$$\begin{bmatrix} A^{\mathsf{T}} \\ I \end{bmatrix} y \geqslant \begin{bmatrix} c \\ 0 \end{bmatrix}, \quad b^{\mathsf{T}}y < b^{\mathsf{T}}y_0.$$

Z twierdzenia 10.2.5 wynika niesprzeczność układu

$$\begin{bmatrix} A & I \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = b, \quad \begin{bmatrix} x^{\mathsf{T}} & u^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} c \\ 0 \end{bmatrix} \geqslant b^{\mathsf{T}}y_0, \quad \begin{bmatrix} x \\ u \end{bmatrix} \geqslant \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

czyli układu $Ax + u = b$, $x^{\mathsf{T}}c \geqslant b^{\mathsf{T}}y_0$, $x \geqslant 0$, $u \geqslant 0$. Stąd wynika, że

$$Ax \leqslant b, \quad c^{\mathsf{T}}x \geqslant b^{\mathsf{T}}y_0, \quad x \geqslant 0,$$

czyli, na mocy tw. 10.3.2, x jest rozwiązaniem pierwotnego zagadnienia. $\blacksquare$

TWIERDZENIE 10.3.5. *Niech x i y będą punktami dopuszczalnymi odpowiednio dla zagadnień pierwotnego i dualnego. Te punkty są rozwiązaniami odpowiednich układów wtedy i tylko wtedy, gdy $(Ax)_i = b_i$ dla każdego wskaźnika i takiego, że $y_i > 0$, a $(A^{\mathsf{T}}y)_i = c_i$ dla każdego i takiego, że $x_i > 0$.*

Dowód. Jeśli x i y są rozwiązaniami, to wobec tw. 10.3.2 i 10.3.3 jest

$$y^{\mathsf{T}}b = b^{\mathsf{T}}y = y^{\mathsf{T}}Ax = c^{\mathsf{T}}x = x^{\mathsf{T}}c.$$

Stąd $y^{\mathsf{T}}(b-Ax)=0$. Ponieważ $y \geqslant 0$ i $b-Ax \geqslant 0$, więc $y_i(b_i-(Ax)_i)=0$ dla każdego i. Dlatego $(Ax)_i = b_i$, jeśli tylko $y_i > 0$. Drugi warunek sprawdza się tak samo. Na odwrót, załóżmy teraz, że dla każdego i jest

$$y_i(b_i-(Ax)_i)=0, \quad x_i(c_i-(A^{\mathsf{T}}y)_i)=0.$$

Wtedy $b^{\mathsf{T}}y = y^{\mathsf{T}}b = y^{\mathsf{T}}Ax = x^{\mathsf{T}}A^{\mathsf{T}}y = x^{\mathsf{T}}c = c^{\mathsf{T}}x$ i na mocy tw. 10.3.2 x i y są rozwiązaniami odpowiednich zagadnień. ■

ZADANIA 10.3

1. Każde z poniższych zagadnień wyrazić w standardowej postaci, a następnie podać zagadnienie dualne.

   (a) Znaleźć minimum sumy $3x_1 + x_2 - 5x_3 + 2$ przy warunkach

   $$x_1 \geqslant x_2, \quad x_2 \leqslant 0, \quad -x_1 + 4x_3 \geqslant 0, \quad x_1 + x_2 + x_3 = 0.$$

   (b) Znaleźć minimum wyrażenia $|x_1 + x_2 + x_3|$ przy warunkach

   $$x_1 - x_2 = 5, \quad x_2 - x_3 = 7, \quad x_1 \leqslant 0, \quad x_3 \geqslant 2.$$

   (c) Znaleźć minimum różnicy $|x_1| - |x_2|$ przy warunkach

   $$x_1 + x_2 = 5, \quad 2x_1 + 3x_2 - x_3 \leqslant 0, \quad x_3 \geqslant 4.$$

2. Co można udowodnić o zagadnieniu programowania liniowego (A, b, c), jeśli każdy punkt dopuszczalny jest rozwiązaniem?

10.4. Algorytm sympleks

Zagadnienie programowania liniowego sformułowane w podrozdz. 10.3 zawierało ograniczenia postaci $Ax \leqslant b$. Wprowadziwszy wektor $u \geqslant 0$ (jego składowe nazywamy zmiennymi *uzupełniającymi*), możemy te ograniczenia wyrazić jako równości

$$Ax + u = b.$$

Dzięki temu każde zagadnienie programowania liniowego sprowadza się do następującego:

II postać standardowa zagadnienia programowania liniowego

Znaleźć maksimum wyrażenia $c^{\mathsf{T}}x$ *przy warunkach* $x \in \mathbb{R}^n$, $Ax = b$ *i* $x \geqslant 0$, *gdzie* $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$ *i* $x \in \mathbb{R}^n$.

Dla tak sformułowanego zagadnienia zbiorem dopuszczalnym jest

$$K := \{x \in \mathbb{R}^n\colon\ Ax = b,\ x \geqslant 0\}.$$

Dla każdego $x \in K$ określamy zbiór wskaźników, dla których składowe x_i tego wektora są dodatnie:

$$I(x) := \{i\colon\ 1 \leqslant i \leqslant n,\ x_i > 0\}.$$

Niech $A_1, A_2, \ldots, A_n$ będą kolumnami macierzy A. Równanie $Ax = b$ ma zatem równoważną postać

$$\sum_{i=1}^{n} x_i A_i = b.$$

Można już sformułować ważne twierdzenie (występują w nim punkty ekstremalne, które zdefiniowano w podrozdz. 10.1):

Twierdzenie 10.4.1. *Poniższe własności punktu $x \in K$ są równoważne:*

1. *x jest punktem ekstremalnym zbioru K.*

2. *Układ kolumn $\{A_i\colon\ i \in I(x)\}$ jest niezależny liniowo.*

Dowód. Załóżmy, że x można wyrazić w postaci $x = \theta u + (1-\theta)v$, gdzie $u, v \in K$ i $0 < \theta < 1$. Dla każdego $i \notin I(x)$ jest

$$0 = x_i = \theta u_i + (1-\theta)v_i.$$

Ponieważ $u_i \geqslant 0$ i $v_i \geqslant 0$, więc $u_i = v_i = 0$ i

$$0 = Au - Av = \sum_{i=1}^{n}(u_i - v_i)A_i = \sum_{i \in I(x)} (u_i - v_i)A_i.$$

Jeśli x ma własność **2**, to $u_i = v_i$ dla każdego i, czyli $u = v$ i x jest punktem ekstremalnym dla K.

Niech teraz x ma własność **1**. Jeśli

$$\sum_{i \in I(x)} w_i A_i = 0$$

i jeśli $w_i := 0$ dla $i \notin I(x)$, to dla dowolnego λ

$$\sum_{i \in I(x)} (x_i \pm \lambda w_i)A_i = b.$$

Ponieważ dla $i \in I(x)$ jest $x_i > 0$, więc istnieje $\lambda \neq 0$ takie, że obie liczby $x_i \pm \lambda w_i$ są dla tych i dodatnie. Dlatego $u := x + \lambda w$ i $v := x - \lambda w$ są punktami dopuszczalnymi. Ponieważ $x = \frac{1}{2}(u+v)$ jest punktem ekstremalnym dla K, więc $u = v$ i $w = 0$, to zaś znaczy, że x ma własność **2**. ■

Wniosek 10.4.2. *Zbiór dopuszczalny K może mieć tylko skończenie wiele punktów ekstremalnych.*

Dowód. Niech E będzie zbiorem punktów ekstremalnych dla K. Niech x i y będą dwoma różnymi punktami zbioru E. Wtedy

$$b = Ax = \sum_{i=1}^{n} x_i A_i = \sum_{i \in I(x)} x_i A_i,$$

$$b = Ay = \sum_{i=1}^{n} y_i A_i = \sum_{i \in I(y)} y_i A_i.$$

Gdyby było $I(x) = I(y)$, to te dwie równości przeczyłyby niezależności liniowej kolumn A_i dla $i \in I(x)$. Jest więc $I(x) \neq I(y)$. Zauważmy jednak, że każdy zbiór $I(x)$ jest podzbiorem zbioru skończonego $\{1, 2, \ldots, n\}$, a takich różnych podzbiorów jest 2^n. Dlatego najwyżej tyle jest punktów ekstremalnych w E. ■

Algorytm sympleks Dantziga

Algorytm sympleks Dantziga [1948] pozwala rozwiązać zagadnienie programowania liniowego w II postaci standardowej. Algorytm dzieli się na dwa etapy. W pierwszym znajdujemy jakiś punkt ekstremalny $x^{(1)}$ zbioru K. W drugim – algorytm, zaczynając od tego punktu, tworzy taki ciąg skończony punktów ekstremalnych $x^{(j)}$, że w każdym następnym punkcie funkcja celu ma większą wartość: $c^{\mathsf{T}}x^{(1)} < c^{\mathsf{T}}x^{(2)} < \ldots$. Jeśli zagadnienie nie ma rozwiązania, to ujawnia się to w toku obliczeń. W przeciwnym razie algorytm daje punkt $x^{(k)}$, w którym funkcja celu osiąga maksimum.

Opiszemy teraz drugi etap algorytmu. Przyjmiemy przy tym następujące założenie:

Każdy punkt ekstremalny zbioru dopuszczalnego K ma dokładnie m składowych dodatnich.

Niech x będzie punktem ekstremalnym zbioru K. Zgodnie z tw. 10.4.1 układ $\{A_i\colon\ i \in I(x)\}$ jest niezależny liniowo, a zatem – wobec powyższego założenia – jest bazą przestrzeni $\mathbb{R}^m$ (dlatego w publikacjach takie x jest nazywane *bazowym punktem dopuszczalnym*). Istnieją współczynniki D_{ij} takie, że

$$A_j = \sum_{i \in I(x)} D_{ij} A_i \quad (1 \leqslant j \leqslant n).$$

Przyjmujemy też, że dla $i \notin I(x)$ jest $D_{ij} = 0$. Stąd

$$A_j = \sum_{i=1}^{n} D_{ij} A_i,$$

czyli $A = AD$. Niech będzie $d := D^{\mathsf{T}}c$, gdzie c jest wektorem z funkcji celu. Zauważmy, że D i d zależą od x i będą się zmieniać w czasie obliczeń.

Dla dowolnego wskaźnika $q \notin I(x)$ i dowolnego $\lambda \in \mathbb{R}$

$$b = Ax = \sum_{i=1}^{n} x_i A_i + \lambda A_q - \lambda \sum_{i=1}^{n} D_{iq} A_i = \\ = \sum_{i=1}^{n} (x_i - \lambda D_{iq} + \lambda \delta_{iq}) A_i = \sum_{i=1}^{n} y_i A_i,$$

czyli

$$Ay = b,$$

gdzie $y := x - \lambda D_q + \lambda e_q$ (D_q jest tu q-tą kolumną macierzy D, a e_q jest q-tym wektorem jednostkowym).

Naszym celem jest teraz tak wybrać q i λ, żeby y było punktem ekstremalnym zbioru K (ma zatem być $y \geqslant 0$) i żeby $c^{\mathsf{T}}y > c^{\mathsf{T}}x$.

Ponieważ $q \notin I(x)$, więc $D_{qq} = 0$ i $x_q = 0$. Stąd $y_q = \lambda$. Punkt y ma być dopuszczalny, więc $\lambda \geqslant 0$. Obliczamy

$$c^{\mathsf{T}}y = \sum_{i=1}^{n} c_i x_i - \lambda \sum_{i=1}^{n} c_i D_{iq} + \lambda \sum_{i=1}^{n} c_i \delta_{iq} = \\ = c^{\mathsf{T}}x - \lambda c^{\mathsf{T}} D_q + \lambda c_q = c^{\mathsf{T}}x + \lambda(c_q - d_q). \tag{10.4.1}$$

Aby zwiększyć wartość funkcji celu, wybieramy q tak, żeby było $c_q > d_q$. Jeśli $c \leqslant d$, to takiego q nie ma i obliczenia się kończą, a x jest rozwiązaniem. W przeciwnym razie na ogół wybieramy q tak, żeby różnica $c_q - d_q$ była jak największa.

Dla ustalonego już q wybieramy λ dodatnie tak, żeby zbiór $I(y)$ miał co najwyżej m elementów. Z definicji punktu y wynika, że $I(y) \subseteq I(x) \cup \{q\}$. Ponieważ w $I(x)$ jest dokładnie m wskaźników, więc wybieramy λ tak, żeby jedna z liczb $x_i - \lambda D_{iq}$ była zerem, a inne były nieujemne. Zauważmy jednak, że jeśli $D_{iq} \leqslant 0$ dla $1 \leqslant i \leqslant n$, to $y \in K$ dla każdego $\lambda > 0$. Z (10.4.1) wynika, że wtedy $\lim_{\lambda \to \infty} c^{\mathsf{T}}y = +\infty$, to zaś znaczy, że rozwiązanie nie istnieje, bo funkcja celu jest nieograniczona na K. Jeśli $D_{iq} > 0$ dla pewnego i, to zwiększamy λ począwszy od 0. Na początku $x_i - \lambda D_{iq} > 0$ dla $i \in I(x)$. Pewne takie składniki maleją; gdy pierwszy z nich staje się zerem, mamy szukane λ. Inaczej mówiąc,

$$\lambda = \min \left\{ \frac{x_i}{D_{iq}} : \; x_i > 0, \; D_{iq} > 0 \right\}.$$

Punkt y jest już całkowicie określony. Sprawdzimy teraz, że jest on ekstremalny. Niech będzie $\lambda = x_p/D_{pq}$; oczywiście $x_p > 0$ i $D_{pq} > 0$. Jest więc

$$I(y) \subseteq I(x) \cup \{q\} \setminus \{p\}.$$

Wobec tw. 10.4.1 wystarczy udowodnić, że układ kolumn

$$\{A_i\colon\ i \in I(x) \cup \{q\} \setminus \{p\}\}$$

jest niezależny liniowo. Przypuśćmy, że

$$\sum_{i \in I(x)} \beta_i A_i + \beta_q A_q = 0,$$

gdzie $\beta_p := 0$. Jeśli $\beta_q = 0$, to równość upraszcza się do postaci

$$\sum_{i \in I(x)} \beta_i A_i = 0,$$

co dzięki niezależności liniowej występujących tu kolumn A_i znaczyłoby, że wszystkie β_i znikają. Trzeba więc jeszcze rozważyć przypadek, gdy $\beta_q \neq 0$. Można przyjąć, że $\beta_q = -1$. Wtedy

$$A_q = \sum_{i \in I(x)} \beta_i A_i \quad (\beta_p = 0).$$

Jest też

$$A_q = \sum_{i \in I(x)} D_{iq} A_i.$$

Ze wspomnianej już niezależności liniowej wynika, że $\beta_i = D_{iq}$. To jednak jest niemożliwe, bo $\beta_p = 0$, a $D_{pq} > 0$. Tak więc $\beta_q = 0$, a y należy do K i jest punktem ekstremalnym.

Pozostaje udowodnić, że jeśli $c \leqslant d$, to x jest rozwiązaniem. Niech u będzie dowolnym punktem dopuszczalnym. Wtedy $Ax = b = Au = A(Du)$. Stąd wynika, że $x = Du$, gdyż Ax i ADu są kombinacjami liniowymi kolumn A_i dla $i \in I(x)$. Dlatego $c^{\mathsf{T}}u \leqslant d^{\mathsf{T}}u = c^{\mathsf{T}}Du = c^{\mathsf{T}}x$, co należało wykazać.

Przykład 10.4.3. Działanie algorytmu sympleks poznamy na następującym przykładzie: znaleźć maksimum sumy

$$F(x) := x_1 + 2x_2 + x_3$$

przy warunkach:

$$x_1 + x_2 + x_5 = 1, \quad x_1 + x_3 + x_4 + x_5 = 1, \quad x_i \geqslant 0 \quad (1 \leqslant i \leqslant 5).$$

Rozwiązanie. Z tych danych wynika, że

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad c^{\mathsf{T}} = \begin{bmatrix} 1 & 2 & 1 & 0 & 0 \end{bmatrix}.$$

Zaczynamy od punktu $x = (0, 1, 0, 1, 0)$, dla którego $I(x) = \{2, 4\}$. Zauważmy, że kolumny A_2 i A_4 stanowią bazę w $\mathbb{R}^2$. Dlatego na mocy tw. 10.4.1 x jest punktem ekstremalnym zbioru dopuszczalnego, czyli bazowym punktem dopuszczalnym.

Każda kolumna macierzy A jest kombinacją liniową kolumn A_2 i A_4:

$$A_1 = A_2 + A_4, \quad A_2 = A_2, \quad A_3 = A_4, \quad A_4 = A_4, \quad A_5 = A_2 + A_4.$$

Stąd wynika macierz D:

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

Wektor d^{T} jest kombinacją liniową wierszy D^i tej macierzy:

$$d^{\mathsf{T}} = c^{\mathsf{T}} D = \sum_{i=1}^{n} c_i D^i = D^1 + 2D^2 + D^3 = (2, 2, 0, 0, 2).$$

Wektor $c - d = (-1, 0, 1, 0, -2)$ ma tylko jedną, trzecią, składową dodatnią. Stąd $q = 3$. Tworzymy wektor $x - \lambda D_q$:

$$x - \lambda D_q = (0, 1, 0, 1, 0) - \lambda(0, 0, 0, 1, 0) = (0, 1, 0, 1 - \lambda, 0).$$

Dla $\lambda = 1$ jest $y = x - \lambda D_q + \lambda e_q = (0, 1, 1, 0, 0)$. Podobnie postępujemy w następnym kroku, zmieniając x na y. Ten krok daje $d = (3, 2, 1, 1, 3)$. Ponieważ $c \leqslant d$, więc y jest rozwiązaniem i $\max F = F(y) = 3$. ■

Organizacja prostych obliczeń

W obliczeniach „na papierze" wygodnie jest umieszczać dane i wyniki każdego etapu obliczeń w pewnej tablicy. Pokażemy to na prostym przykładzie szukania maksimum sumy

$$F(x) := 6x_1 + 14x_2$$

przy warunkach:

$$2x_1 + x_2 \leqslant 12, \quad 2x_1 + 3x_2 \leqslant 15,$$
$$x_1 + 7x_2 \leqslant 21, \quad x_1 \geqslant 0, \; x_2 \geqslant 0.$$

Przygotowując się do właściwych obliczeń, wprowadzamy zmienne uzupełniające. Szukamy więc maksimum sumy

$$F(x) := 6x_1 + 14x_2 + 0x_3 + 0x_4 + 0x_5$$

przy warunkach:

$$2x_1 + x_2 + x_3 = 12, \quad 2x_1 + 3x_2 + x_4 = 15,$$
$$x_1 + 7x_2 + x_5 = 21, \quad x_i \geqslant 0 \quad (1 \leqslant i \leqslant 5).$$

Pierwszym wektorem x jest (0, 0, 12, 15, 21). Wszystkie te dane rozmieszczamy w pierwszej tablicy:

6	14	0	0	0	0
2	1	1	0	0	12
2	3	0	1	0	15
1	7	0	0	1	21
0	0	12	15	21	

Ogólniej, taka tablica ma postać

c^T		0	$F(x)$
A		I	b
x (niebazowe)		x (bazowe)	

Tablicę tego typu tworzymy na początku każdego etapu obliczeń. Jej górny wiersz zawiera składowe wektora c określającego funkcję celu $F(x) = c^T x$, a w prawym rogu – jej bieżącą wartość. W następnych m wierszach znajdują się współczynniki układu równań $Ax = b$. Warto przypomnieć, że operacje elementarne wykonywane na wierszach nie zmieniają zbioru rozwiązań. Ostatni wiersz zawiera bieżący wektor x. Wartość $F(x)$ wynika łatwo ze skrajnych wierszy tablicy.

Reguły postępowania

W każdym etapie obliczeń wielkości pamiętane w tablicy podlegają następującym regułom:

1. Wektor x spełnia warunki $Ax = b$.
2. Wektor x spełnia nierówność $x \geqslant 0$.

3. n składowych tego wektora znika (są to zmienne niebazowe); m pozostałych składowych (zmienne bazowe) na ogół nie znika. (Parametry n i m odnoszą się do pierwotnego zagadnienia, jeszcze bez zmiennych uzupełniających).
4. Każda zmienna bazowa występuje tylko w jednym wierszu macierzy określającej warunki.
5. Funkcja celu F wyraża się tylko przez zmienne niebazowe.

Ciąg dalszy przykładu

W pierwszej tablicy z ostatniego przykładu zmiennymi bazowymi są x_3, x_4, x_5, a niebazowymi x_1, x_2. Powyższe reguły są tu zachowane.

W każdym etapie badamy bieżącą tablicę, aby sprawdzić, czy można zwiększyć wartość $F(x)$, zaliczając pewną zmienną niebazową do bazowych. W naszym przykładzie widzimy, że ta wartość się zwiększy, gdy wzrośnie x_1 lub x_2 (co trzeba zrekompensować poprawiając odpowiednie pozostałe składowe). Ponieważ $14 > 6$, więc jednostkowy wzrost składowej x_2 zwiększa szybciej wartość $F(x)$. Dlatego pozostawiamy wartość $x_1 = 0$ i zwiększamy x_2 tak bardzo, jak to jest możliwe. Z warunków zagadnienia wynika, że musi być

$$0 \leqslant x_3 = 12 - x_2, \quad 0 \leqslant x_4 = 15 - 3x_2, \quad 0 \leqslant x_5 = 21 - 7x_2,$$

co odpowiednio daje nierówności $x_2 \leqslant 12$, $x_2 \leqslant 5$ i $x_2 \leqslant 3$. Wobec tego x_2 może wzrosnąć tylko do 3. Stąd i z powyższych równości wynikają nowe wartości x_3, x_4, x_5 i cały wektor x:

$$x = (0, 3, 9, 6, 0).$$

Nowymi zmiennymi bazowymi są teraz x_2, x_3 i x_4. Uwzględniamy to, budując taką nową tablicę, która będzie spełniała pięć podanych wyżej reguł. Ostatnia z nich wymaga wyrażenia $F(x)$ przez x_1 i x_5. Ponieważ $x_2 = (21 - x_5)/7$, więc

$$F(x) = 6x_1 + 14x_2 = 6x_1 + 2(21 - x_5) = 6x_1 - 2x_5 + 42.$$

Nową zmienną bazową jest x_2. Ma ona występować tylko w jednym warunku (reguła **4**). Żeby to uzyskać, eliminujemy x_2 z warunków pierwszego i drugiego posługując się trzecim z nich. Są to czynności jak w metodzie eliminacji Gaussa; elementem głównym jest tu 7. Daje to następującą tablicę:

6	0	0	0	-2	42
$\frac{13}{7}$	0	1	0	$-\frac{1}{7}$	9
$\frac{11}{7}$	0	0	1	$-\frac{3}{7}$	6
1	7	0	0	1	21
0	3	9	6	0	

Mamy teraz podobną sytuację jak na początku obliczeń, zmieniła się tylko rola pewnych zmiennych. Niebazowymi zmiennymi są x_1 i x_5. Zwiększenie wartości drugiej z nich zmniejsza $F(x)$, więc nową zmienną bazową musi stać się x_1. Jej nowa, większa wartość musi być taka, że

$$0 \leqslant x_3 = 9 - \tfrac{13}{7}x_1, \quad 0 \leqslant x_4 = 6 - \tfrac{11}{7}x_1, \quad 0 \leqslant 7x_2 = 21 - x_1.$$

Stąd wynika, że odpowiednio $x_1 \leqslant \frac{63}{13}$, $x_1 \leqslant \frac{42}{11}$ i $x_1 \leqslant 21$. Dlatego musi być $x_1 = \frac{42}{11}$, a to wraz z powyższymi warunkami daje wartości x_3, x_4 i x_2 oraz wektor

$$x = (\tfrac{42}{11}, \tfrac{27}{11}, \tfrac{21}{11}, 0, 0).$$

Zmiennymi niebazowymi są teraz x_4 i x_5. Zgodnie z regułą **5** wyrażamy przez nie $F(x)$. Ponieważ $x_1 = \frac{7}{11}(6 - x_4)$, więc

$$F(x) = 6x_1 - 2x_5 + 42 = \tfrac{42}{11}(6 - x_4) - 2x_5 + 42 = -\tfrac{42}{11}x_4 - 2x_5 + \tfrac{714}{11}.$$

Wypełnianie trzeciej tablicy nie jest już potrzebne. Istotnie, współczynniki przy obu zmiennych są ujemne, a to znaczy, że ani x_4, ani x_5 nie może stać się zmienną bazową, bo to by zmniejszyło $F(x)$. Dlatego obliczone x jest rozwiązaniem, a w pierwotnym zagadnieniu szukane maksimum jest równe $F(\frac{42}{11}, \frac{27}{11}) = \frac{630}{11}$.

Podsumowanie

Czynności wykonywane dla każdej tablicy można streścić tak:

1. Jeśli wszystkie współczynniki przy zmiennych w F są niedodatnie, to bieżące x jest rozwiązaniem.
2. W przeciwnym razie wybieramy tę zmienną niebazową x_j, której w F odpowiada największy współczynnik. Staje się ona nową zmienną bazową.
3. Dzielimy każde b_i przez współczynnik a_{ij} przy tej zmiennej. Jej nową wartością staje się najmniejszy z ilorazów b_i/a_{ij}. Niech to będzie iloraz dla $i = k$.
4. Używając elementu głównego a_{kj} w eliminacji Gaussa, zerujemy elementy j-tej kolumny macierzy A (z wyjątkiem k-tego).

Oszacowanie kosztu obliczeń

Liczba wykonywanych kroków w metodzie sympleks może w teorii wynosić $\binom{n}{m}$. Nawet w zagadnieniach umiarkowanie dużych ta liczba jest astronomicznie wielka. Istotnie, np. dla $n = 300$ i $m = 100$ ze wzoru Stirlinga

$n! \approx \sqrt{2\pi n}(n/e)^n$ wynika, że $\binom{300}{100} \approx 4_{10}81$. W praktyce jednak trzeba na ogół wykonać najwyżej $2m$ kroków.

Inne algorytmy

Karmarkar [1984] zaproponował nowy algorytm programowania liniowego, który ma być lepszy od algorytmu sympleks, gdy liczba zmiennych przekracza 15 000. Koszt algorytmu Karmarkara jest funkcją wielomianową (a nie wykładniczą, jak w metodzie sympleks) wielkości zagadnienia, ale ostatecznie o jego wyższości świadczy wieloletnie doświadczenie rozwiązywania licznych skrajnie wielkich zagadnień. Wcześniejszy algorytm Khachiana ma też koszt wielomianowy, ale nie może konkurować z metodą sympleks, gdyż w jego kolejnych etapach potrzebna jest coraz większa dokładność.

ZADANIA 10.4

1. W związku z opisem algorytmu sympleks udowodnić, że jeśli x jest rozwiązaniem, to $c \leqslant d$.

2. Dokończyć rozwiązywanie zagadnienia z przykł. 10.4.3.

3. Rozwiązać następujące zagadnienia programowania liniowego:

(a) Znaleźć maksimum sumy $F(x) := 2x_1 - 3x_2$ przy warunkach

$$2x_1 + 5x_2 \geqslant 10, \quad x_1 + 8x_2 \leqslant 24, \quad x_1 \geqslant 0, \quad x_2 \geqslant 0.$$

(b) Znaleźć maksimum sumy $F(x) := 6x_1 + 14x_2$ przy warunkach

$$x_1 + x_2 \leqslant 12, \quad 2x_1 + 3x_2 \leqslant 15, \quad x_1 + 7x_2 \leqslant 21, \quad x_1 \geqslant 0, \quad x_2 \geqslant 0.$$

Powtórzyć obliczenia po zmianie drugiego warunku na $2x_1 + 3x_2 \geqslant 15$.

ROZDZIAŁ 11

Optymalizacja

11.0. Wstęp

Przez *optymalizację* rozumiemy wyznaczanie minimum funkcji rzeczywistej wielu zmiennych w danym obszarze. Ściślej, na ogół chcemy znaleźć zarówno wartość minimalną, jak i punkt, w którym jest ona osiągnięta. Oczywiście, poprzedni rozdział dotyczy pewnego ważnego typu optymalizacji: funkcja jest tam liniowa, a obszar określony za pomocą układu nierówności liniowych. W takim szczególnym przypadku można stosować równie specjalne metody. Ich efektywność pozwala rozwiązywać zadania z tysiącami niewiadomych i tyluż warunkami.

Poza komfortową dziedziną funkcji liniowych panuje chaos. Nawet szukanie minimum funkcji jednej zmiennej może być ambitnym zadaniem. Istotnie, sprawia ona kłopoty, gdy ma wiele minimów lokalnych, a tym, co chcemy znaleźć, jest minimum globalne. Jeszcze większych trudności można się spodziewać w przypadku funkcji wielu zmiennych z dodatkowymi warunkami. Niedawno wydany podręcznik Nocedala i Wrighta [1999] obejmuje całość

zagadnień optymalizacyjnych. Inne książki wymieniamy w dalszym ciągu tego rozdziału.

Teoria optymalizacji dzieli się w naturalny sposób na wiele działów, stosownie do struktury badanych zagadnień. Przypomniano już jeden z nich, mianowicie programowanie liniowe. Innym jest analiza wypukła dotycząca zadań, w których funkcja lub nakładane na nią ograniczenia są wypukłe; zob. niedawną książkę Borweina i Lewisa [2000]. Co najmniej od 40 lat są badane szczególne zagadnienia, w których minimum szuka się tylko na danej siatce punktów, np. w punktach w $\mathbb{R}^n$ o współrzędnych całkowitych. Książka Cornuejols [2000] zawiera nowe metody stosowane w takich zagadnieniach. Programy komputerowe dotyczące optymalizacji można znaleźć w wielu książkach; zob. np. Bhatti [2000].

Internet jest obfitym źródłem informacji i programów komputerowych związanych z optymalizacją. I tak na przykład strona[1)]

`http://plato.la.asu.edu/guide.html`

daje drzewo decyzyjne, ułatwiające wybór oprogramowania. Strona

`http://www.aic.nrl.navy.mil:80/galist/`

jest poświęcona algorytmom genetycznym. Najnowsze informacje podaje NA-Digest pod adresem

`http://www.netlib.org/na-digest/`

Ponieważ zawartość stron i ich adresy mogą się zmieniać, więc w razie potrzeby warto korzystać z wyszukiwarek internetowych.

Terminologia stosowana w tym rozdziale jest standardowa. Zadania dotyczą ustalonej funkcji rzeczywistej f zależnej od n zmiennych. Szukamy jej *minimum globalnego*, tj. takiego p, że $f(p) \leqslant f(x)$ dla każdego x. Natomiast *minimum lokalnym* jest takie q, że nierówność $f(q) \leqslant f(x)$ zachodzi dla każdego x z pewnego zbioru otwartego, do którego należy q. Oczywiście minimum globalne jest zarazem lokalne. Jest też oczywiste, że łatwiej jest znaleźć minimum lokalne, natomiast bynajmniej nie jest proste ustalenie, czy któreś ze znanych już minimów lokalnych jest zarazem minimum globalnym.

Wektor (kolumnowy) w $\mathbb{R}^n$ oznaczamy zwykle jedną literą, np. x, a jeśli są potrzebne jego składowe, to piszemy $x = (x_1, x_2, \ldots, x_n)$. Elementy ciągu wektorów (punktów) wyróżniamy za pomocą górnego wskaźnika; piszemy więc $x^{(1)}, x^{(2)}$ itd.

Algorytmy optymalizacji dla funkcji wielu zmiennych wymagają często jej badania wzdłuż pewnej prostej. Dlatego tematem następnego podrozdziału są funkcje jednej zmiennej.

[1)] Podane adresy stron internetowych dotyczą oryginału książki, tzn. wydania w jęz. ang. (*przyp. red. WNT*).

11.1. Przypadek jednej zmiennej

Prostą w przestrzeni wektorowej V jest zbiór

$$\{u + tv\colon\ t \in \mathbb{R}\},$$

gdzie u i v są ustalonymi wektorami z V (i $v \neq 0$), a t parametrem rzeczywistym. Jeśli funkcja F o wartościach rzeczywistych jest określona na V, to $f(t) := F(u + tv)$ jest funkcją jednej zmiennej rzeczywistej. Jeśli prosta przechodzi blisko minimum funkcji F, to możemy zacząć jego poszukiwanie od wyznaczenia minimum prostszej funkcji f. W dalszym ciągu podrozdziału związek f z F nie jest istotny. Rozważamy po prostu obliczanie minimum dowolnej funkcji $f\colon\ \mathbb{R} \to \mathbb{R}$.

Jeśli nasza funkcja f jest wszędzie różniczkowalna, to w punkcie q lokalnego minimum jest na pewno $f'(q) = 0$. Jest więc rozsądnie szukać wszystkich punktów, w których pochodna funkcji f znika. Dla każdego z nich trzeba sprawdzić, czy jest to lokalne maksimum, lokalne minimum, czy też *punkt siodłowy*. W tym ostatnim przypadku funkcja po jednej stronie punktu q ma wartości mniejsze od $f(q)$, a po drugiej stronie większe.

Są nam potrzebne dwa typy algorytmów. Pierwszy używa pochodnej (jeśli ona istnieje) funkcji f. Drugi obywa się bez niej (jeśli nawet ona istnieje). Gdy mamy do czynienia z funkcjami wielu zmiennych, to rozróżnienie jest jeszcze ważniejsze, gdyż funkcja n zmiennych ma tyleż pochodnych cząstkowych pierwszego rzędu i każda z nich powinna znikać w lokalnym minimum. Programowanie obliczania ich wartości może być zbyt kosztowne dla człowieka, a same obliczenia na komputerze mogą być zbyt czasochłonne.

Jeśli z jakiegokolwiek powodu pochodna funkcji f nie jest dostępna, to możemy odwołać się do prostej procedury przeszukiwania. Można np. wybrać pewną długość kroku h (uwzględniając jakieś wiadomości o tym, jak zachowuje się funkcja). Obliczamy wartości $f(kh)$ dla $k = 0, \pm 1, \pm 2, \ldots$, co daje nam informacje o zmienności funkcji i o tym, gdzie może kryć się jej minimum. Dla każdego h łatwo podać funkcję, dla której badanie tych wartości nic nie da. Tak jest zresztą dla prawie każdego algorytmu, jeśli funkcja nie podlega żadnym ograniczeniom. Nie wystarczy nawet zażądać jej ciągłości. Bardziej użyteczne mogą być oszacowania pochodnej, np. założenie, że $|f'(x)| \leqslant M$ dla każdego x. Jeśli znamy M, a także wartości funkcji w punktach a i b, gdzie $a < b$, to wiemy, że w przedziale $[a, b]$ jest

$$f(x) \geqslant \min\{f(a),\ f(b)\} - \tfrac{1}{2}(b - a)M.$$

Wynika to z twierdzenia o wartości średniej. Istotnie, jeśli x leży w lewej połowie tego przedziału, to

$$f(x) - f(a) = f'(\xi)(x - a) \geqslant -M(x - a) \geqslant -\tfrac{1}{2}(b - a)M.$$

Tak samo rozumujemy dla pozostałych $x \in [a, b]$. Jeśli więc taką funkcję znamy w punktach $x^{(k)} := kh$, to

$$\min_k f(x^{(k)}) \geqslant \inf_x f(x) \geqslant \min_k f(x^{(k)}) - \tfrac{1}{2}hM.$$

Wobec tego, szukając minimum, możemy ograniczyć się do tych przedziałów $[x^{(j)}, x^{(j+1)}]$, dla których

$$\min\{f(x^{(j)}),\ f(x^{(j+1)})\} \leqslant \min_k f(x^{(k)}) + \tfrac{1}{2}hM.$$

Istotnie, jeśli wspomniany przedział narusza ten warunek, to

$$\inf_{x \in [x^{(j)}, x^{(j+1)}]} f(x) \geqslant \min\{f(x^{(j)}),\ f(x^{(j+1)})\} - \tfrac{1}{2}M > \min_k f(x^{(k)}),$$

czyli minimum funkcji f w tym przedziale jest większe od wartości funkcji w jednym z punktów $x^{(k)}$.

Jeśli nic pożytecznego o pochodnej funkcji f nie wiemy, to można posłużyć się algorytmem opartym na słabszej własności tej funkcji. Załóżmy mianowicie, że w przedziale $[a, b]$, w którym szukamy minimum, jest ona ciągła i *unimodalna*, tj. że ma ona tam jedno minimum lokalne, np. x^*. Jeśli tak jest, to f maleje w $[a, x^*]$ i rośnie w $[x^*, b]$.

Dla funkcji ciągłej i unimodalnej można zastosować metodę *złotego podziału odcinka*. Występuje w niej liczba

$$r := \tfrac{1}{2}(\sqrt{5} - 1) = 0.61803\,39887\ldots,$$

czyli pierwiastek równania $r^2 + r - 1 = 0$. Starożytni Grecy uważali ją za optymalny z estetycznego punktu widzenia stosunek długości dwóch części odcinka.

W każdym kroku tej metody znamy przedział $[a, b]$ zawierający szukane minimum. Przyjmujemy $x := a + r(b - a)$ i $y := a + r^2(b - a)$. Będą nam potrzebne wartości $u := f(x)$ i $v := f(y)$, ale – jak się okaże – po starcie do obliczeń wystarczy w każdym nowym przedziale obliczyć jedną wartość funkcji.

Jeśli $u > v$, to dzięki unimodalności funkcji jej minimum leży w $[a, x]$. Tego przedziału używamy w następnym kroku. Zauważmy, że znamy już wartość funkcji w jego punkcie wewnętrznym y, czyli w $a + r(x - a)$. W następnym kroku y odgrywa rolę dawnego x, a wartość funkcji trzeba obliczyć w $a + r^2(x - a)$. Wykonujemy więc kolejno następujące czynności:

$b \leftarrow x$; $x \leftarrow y$; $u \leftarrow v$
$y \leftarrow a + r^2(b - a)$; $v \leftarrow f(y)$

Jeśli natomiast $u \leqslant v$, to minimum leży w przedziale $[y, b]$. Zauważamy, że $x = y + r^2(b - y)$. Podstawienia podobne do poprzednich są następujące:

$$a \leftarrow y;\ y \leftarrow x;\ v \leftarrow u$$
$$x \leftarrow a + r^2(b - a);\ u \leftarrow f(x)$$

Metoda przypomina algorytm bisekcji stosowany do obliczania zera funkcji, w którym długość nowego przedziału jest długością poprzedniego, pomnożoną przez 0.5. W metodzie złotego podziału odcinka ta długość maleje wolniej, bo zamiast 0.5 mamy czynnik równy w przybliżeniu 0.62 [2)].

Mając do dyspozycji wartości pochodnej, szukamy najpierw punktu, gdzie $f'(x) = 0$. W tym celu możemy (jeśli tylko i druga pochodna istnieje) zastosować metodę Newtona, czyli wzór

$$x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})}. \tag{11.1.1}$$

Można wykazać (zob. zad. 3), że ta metoda jest równoważna temu, iż dla funkcji f wyznaczamy wielomian Q stopnia $\leqslant 2$ spełniający warunki interpolacyjne Hermite'a:

$$Q(x^{(k)}) = f(x^{(k)}), \quad Q'(x^{(k)}) = f'(x^{(k)}), \quad Q''(x^{(k)}) = f''(x^{(k)})$$

i definiujemy $x^{(k+1)}$ jako punkt, w którym znika funkcja liniowa Q'. Zamiast metody Newtona możemy stosować metodę siecznych (wtedy unikamy obliczania wartości f''), metodę bisekcji lub jeszcze inną.

Skoro już wiemy, że można użyć wielomianu interpolacyjnego do przybliżania funkcji f, to nasuwa się myśl, aby posłużyć się zwykłą interpolacją Lagrange'a. W takim przypadku budujemy wielomian Q stopnia drugiego, interpolujący funkcję f w trzech ostatnich punktach, np. $x^{(k)}$, $x^{(k-1)}$ i $x^{(k-2)}$. Przypuszczając, że lokalnie funkcje Q i f są podobne, definiujemy $x^{(k+1)}$ jako punkt, w którym Q osiąga minimum. Trzeba jednak zawczasu przewidzieć, co robimy, jeśli ten punkt nie poprawia poprzednich; zob. Powell

[2)] To porównanie nie świadczy o wyższości metody bisekcji. Trzeba pamiętać, że gdy szukamy zera funkcji w $[a, b]$, to obliczywszy jej wartość w jednym punkcie wewnętrznym ξ wiemy, czy leży ono w $[a, \xi]$, czy w $[\xi, b]$. W zadaniu tu rozważanym dopiero *dwie* wartości $f(y)$ i $f(x)$ (gdzie $y < x$) pozwalają ustalić, w którym z *zachodzących* na siebie przedziałów $[a, x]$ i $[y, b]$ leży minimum. Długość właściwego przedziału może przekraczać $\frac{1}{2}(b - a)$. Jeśli np. $a = 0$, $y = 0.45$, $x = 0.55$ i $b = 1$, to pierwszy etap obliczeń daje przedział o długości $0.55 < r$, np. $[0, 0.55]$, ale wtedy po drugim ta długość może wynosić nawet $0.45 > r^2$. Można udowodnić, że złoty podział odcinka jest optymalny. Warto też zauważyć, że wbrew temu, co sugerują wzory, konstrukcja punktów wewnętrznych wymaga *łącznie* tylko jednego mnożenia przez r lub $r^2 = \frac{1}{2}(3 - \sqrt{5})$. Dowód tego faktu może być tematem dodatkowego zadania (*przyp. tłum.*).

[1964]. Także Walsh [1975] rozważa ten algorytm; zob. również Dahlquist i Björck [1974]. Są też znane inne algorytmy oparte na interpolacji, np. metoda Davidona, w której na każdym etapie stosuje się wielomian stopnia trzeciego. Metodę opisują Walsh [1975] oraz Buchanan i Turner [1992].

ZADANIA 11.1

1. Udowodnić, że w metodzie złotego podziału odcinka długości kolejnych przedziałów maleją w stosunku $\frac{1}{2}(\sqrt{5}-1)$.
2. Rozważyć algorytm minimalizacji funkcji f przez obliczanie zera pochodnej f' metodą bisekcji. Jak wytłumaczyć to, że ten prosty algorytm jest szybszy od metody złotego podziału odcinka?
3. Wykazać, że metoda Newtona (11.1.1) jest identyczna z opisaną w tekście metodą opartą na interpolacji Hermite'a.

ZADANIA KOMPUTEROWE 11.1

K1. Napisać program lokalizujący minimum funkcji f w przedziale $[a,b]$. Założyć, że znamy oszacowanie z góry M dla $|f'(x)|$ w tym przedziale. Dla liczby naturalnej N podanej przez użytkownika program oblicza wartości funkcji w punktach $x^{(j)} := a + jh$ $(0 \leqslant j \leqslant N)$, gdzie $h := (b-a)/N$. Program ma podać wszystkie podprzedziały $[x^{(j)}, x^{(j+1)}]$, w których – według informacji podanych w tekście – może się znajdować minimum. Bardziej ambitnym celem byłby program iterujący to postępowanie przez dokładniejsze badanie wybranych podprzedziałów.

11.2. Metody spadku

Po pokazaniu kilku algorytmów minimalizacji funkcji na prostej przejdziemy do podobnego, ale trudniejszego zadania – minimalizacji funkcji f o wartościach rzeczywistych, zależnej od n zmiennych rzeczywistych. *Gradientem* takiej funkcji w punkcie x jest wektor G o składowych

$$G_i \equiv G_i(x) := \frac{\partial f(x)}{\partial x_i} \quad (1 \leqslant i \leqslant n).$$

Jest on też oznaczany symbolem $\nabla f(x)$ albo po prostu $f'(x)$. To ostatnie wyrażenie jest standardowym oznaczeniem dla *pochodnej Frécheta*. Intrepretujemy ją jako odwzorowanie liniowe, którego wartością w u jest $u^{\mathsf{T}} f'(x)$.

Hesjan funkcji f w x jest macierzą H o elementach

$$H_{ij} \equiv H_{ij}(x) := \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \quad (1 \leqslant i, j \leqslant n).$$

On też może być interpretowany jako pochodna Frécheta, skąd wynika symbol $f''(x)$. Gradient i hesjan dają nam początkowe składniki wzoru Taylora dla f:

$$f(x+h) = f(x) + G(x)^{\mathsf{T}}h + \frac{1}{2}h^{\mathsf{T}}H(x)h + \dots$$

Gradient wskazuje kierunek najszybszego wzrostu funkcji f; w przeciwnym kierunku funkcja ma *najszybszy spadek*. Sprawdzamy to, rozważając dowolny wektor u o długości 1 i badając, jak ta funkcja zmienia się lokalnie w x w kierunku u. Łatwo to obliczyć za pomocą pochodnej kierunkowej

$$\frac{d}{dt} f(x+tu)\Big|_{t=0}.$$

Zmieniając we wzorze Taylora h na tu, otrzymujemy równość

$$f(x+tu) = f(x) + tG(x)^{\mathsf{T}}u + \frac{1}{2}t^2u^{\mathsf{T}}H(x)u + \dots,$$

skąd

$$\frac{d}{dt} f(x+tu) = G(x)^{\mathsf{T}}u + tu^{\mathsf{T}}H(x)u + \dots$$

Można też zastosować wzór na różniczkowanie funkcji złożonej:

$$\frac{d}{dt} f(x+tu) = u^{\mathsf{T}}f'(x+tu) = u^{\mathsf{T}}G(x+tu) = u^{\mathsf{T}}\nabla f(x+tu). \quad (11.2.1)$$

Dla $t = 0$ otrzymujemy $G(x)^{\mathsf{T}}u$ jako miarę szybkości zmiany funkcji f w punkcie x i kierunku u. Na mocy nierówności Cauchy'ego-Schwarza ta wielkość nie przekracza $\|G(x)\|\,\|u\| = \|G(x)\|$. Z drugiej strony, można osiągnąć to oszacowanie, przyjmując, że wektor u jest współliniowy z $G(x)$.

Oczywista strategia szukania minimum funkcji f polega na tym, że startując z dowolnego punktu x, posuwamy się w kierunku $-G(x)$, w którym ona lokalnie najszybciej maleje. Na promieniu $\{x - tG(x)\colon t \geqslant 0\}$ znajdujemy minimum funkcji f, z tego punktu posuwamy się w kierunku najszybszego spadku itd.

Mimo uzasadnienia teoretycznego *metoda najszybszego spadku* często działa źle, gdyż dużo czasu tracimy na zygzakowanie w kierunkach nie prowadzących do minimum globalnego. Jest tak nawet dla funkcji kwadratowych, co pokazuje rys. 4.3. Łatwo udowodnić, że kolejne kierunki są względem siebie ortogonalne (zob. poniższy lemat), tak że wspomniane zygzakowanie jest nieuchronne. Byłoby znacznie lepiej posuwać się wzdłuż prostej mającej jakiś pośredni kierunek albo też przemyśleć od początku wybór właściwych kierunków. Ten temat wypłynie w dalszych podrozdziałach.

Lemat 11.2.1. *Jeśli na prostej $\{w+tu\colon t \in \mathbb{R}\}$ funkcja f ma minimum w x, to wektor u jest ortogonalny względem jej gradientu w tym punkcie.*

Dowód. Niech będzie $g(t) := f(x+tu)$. Punkty $x+tu$ leżą na powyższej prostej, a funkcja g ma minimum w 0, zatem $g'(0) = 0$. Wobec (11.2.1) jest $g'(t) = u^{\mathsf{T}}G(x+tu)$, skąd $0 = g'(0) = u^{\mathsf{T}}G(x)$. ■

Zaproponowano wiele funkcji, których można używać, testując programy optymalizacyjne. Zmienność argumentów tych funkcji wpływa na nie w różny sposób albo też do minimum prowadzą „doliny" z licznymi załomami. Scales [1985] podaje takie funkcje, które sprawiają kłopoty nawet najlepszym algorytmom. Oto pięć funkcji tego typu wraz z kłopotliwymi punktami początkowymi:

Funkcja Scalesa:

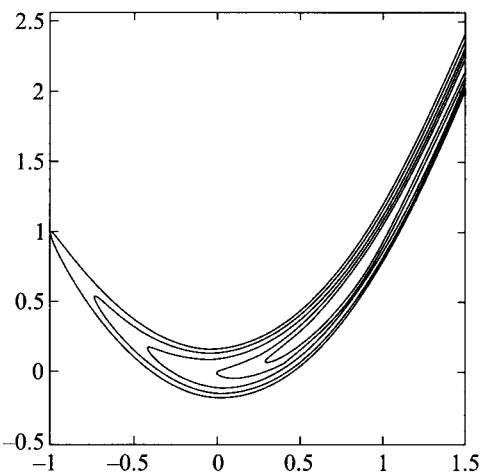
$$e^x + 0.01/x,$$

punkt początkowy 1.

Funkcja Rosenbrocka (rys. 11.1)**:**

$$100(x_1^2 - x_2)^2 + (1 - x_1)^2,$$

punkt początkowy $(-1.2,\ 1)$.



RYS. 11.1. Poziomice dla funkcji Rosenbrocka

Funkcja śrubowa Fletchera i Powella:

$$100[(x_3 - 10\theta)^2 + (r-1)^2] + x_3^2,$$

gdzie

$$r := \sqrt{x_1^2 + x_2^2}, \quad 2\pi\theta = \begin{cases} \operatorname{arctg}(x_2/x_1) & (x_1 \geqslant 0) \\ \operatorname{arctg}(x_2/x_1) + \pi & (x_1 < 0). \end{cases}$$

Punkt początkowy $(-1,\ 0,\ 0)$.

Funkcja Wooda:

$$100(x_2 - x_1^2)^2 + (1 - x_1)^2 + 90(x_4 - x_3^2)^2 + \\ + (1 - x_3)^2 + 10.1[(x_2 - x_1)^2 + (x_4 - 1)^2] + 19.8(x_2 - 1)(x_4 - 1),$$

punkt początkowy $(-3, -1, -3, -1)$.

Funkcja szczególna Powella:

$$(x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4,$$

punkt początkowy $(3, -1, 0, 1)$.

ZADANIA 11.2

1. Dla każdej z funkcji testowych znaleźć gradient i hesjan.
2. Udowodnić, że hesjan funkcji f jest równy jakobianowi dla ∇f.

11.3. Analiza funkcji kwadratowych celu

Zwykłą strategią w zagadnieniach optymalizacji jest założenie, że naszą funkcję można lokalnie przybliżyć wielomianem kwadratowym, tj. – dla n zmiennych – funkcją postaci

$$f(x) = a - b^{\mathrm{T}}x + \frac{1}{2}x^{\mathrm{T}}Ax,$$

gdzie a jest skalarem, b wektorem o n składowych, natomiast A jest macierzą kwadratową symetryczną stopnia n. Znak minus przed składnikiem liniowym $b^{\mathrm{T}}x$ upodabnia f do wyrażeń z podrozdz. 4.7, poświęconego metodom sprzężonych kierunków. Funkcja f ma minimum (a nie maksimum albo punkt siodłowy), gdy macierz A jest nieujemnie określona. Wynika to ze standardowego testu dotyczącego drugich pochodnych funkcji wielu zmiennych.

Funkcja f zależy od $\frac{1}{2}(n+1)(n+2)$ parametrów (jest to wymiar przestrzeni wielomianów kwadratowych n zmiennych):

$$f(x) = a - \sum_{i=1}^{n} b_i x_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_i x_j.$$

Dzięki symetrii macierzy A gradient G funkcji f jest wektorem o składowych

$$\frac{\partial f}{\partial x_k} = -b_k + \sum_{i=1}^{n} A_{ik}x_i,$$

czyli

$$G(x) = \nabla f(x) = Ax - b.$$

Ta pochodna znika, gdy $Ax = b$, czyli $x = A^{-1}b$. Ostatnie przejście zakłada, że macierz A jest dodatnio określona.

Będzie nam potrzebna znajomość minimum funkcji f na dowolnej prostej $\{w + th\colon\ t \in \mathbb{R}\}$. Ponieważ $w^{\mathsf{T}}Ah = h^{\mathsf{T}}Aw$, więc

$$\begin{aligned} f(w+th) &= a - b^{\mathsf{T}}(w+th) + \tfrac{1}{2}(w+th)^{\mathsf{T}}A(w+th) = \\ &= a - b^{\mathsf{T}}w - tb^{\mathsf{T}}h + \tfrac{1}{2}w^{\mathsf{T}}Aw + tw^{\mathsf{T}}Ah + \tfrac{1}{2}t^2h^{\mathsf{T}}Ah, \end{aligned}$$

a stąd wynika, że

$$\frac{d}{dt} f(w+th) = -b^{\mathsf{T}}h + w^{\mathsf{T}}Ah + th^{\mathsf{T}}Ah.$$

Ta pochodna zeruje się w punkcie

$$t^* := \frac{(b-Aw)^{\mathsf{T}}h}{h^{\mathsf{T}}Ah},$$

czyli na danej prostej funkcja f osiąga minimum w punkcie

$$w + t^*h = w + \frac{(b-Aw)^{\mathsf{T}}h}{h^{\mathsf{T}}Ah}\, h.$$

W szczególnym przypadku, gdy $h := b - Aw$, tj. gdy h wyznacza kierunek przeciwny do gradientu w punkcie w, mamy wzór

$$u := w + t^*h = w + \frac{(b-Aw)^{\mathsf{T}}(b-Aw)}{(b-Aw)^{\mathsf{T}}A(b-Aw)}\,(b-Aw).$$

Jeden krok metody najszybszego spadku polega na przejściu z danego punktu w do punktu u określonego powyższym wzorem. Oczywiście jest to zbędne dla funkcji kwadratowej, z danymi A i b, bo minimum wynika wprost ze wzoru $x = A^{-1}b$. Jeśli nie znamy A i b, to możemy znaleźć funkcję kwadratową interpolującą f na zbiorze $\frac{1}{2}(n+1)(n+2)$ punktów. Muszą one jednak być rozmieszczone tak, żeby zadanie interpolacyjne miało jednoznaczne rozwiązanie. Inaczej mówiąc, nie może istnieć funkcja kwadratowa Q znikająca w tych wszystkich punktach, ale nierówna tożsamościowo 0. Dla $n = 2$ sensownymi sześcioma punktami są np.: 0, $e^{(1)}$, $e^{(2)}$, $2e^{(1)}$, $2e^{(2)}$ i $e^{(1)}+e^{(2)}$, gdzie $e^{(1)} := (1,0)$ i $e^{(2)} := (0,1)$ (por. tw. 6.10.4 i rys. 6.17). Taki układ można też przesuwać i zmieniać jego skalę. Znając wartości funkcji f

w tych sześciu punktach, możemy obliczyć przybliżenia wszystkich pięciu pochodnych cząstkowych rzędu pierwszego i drugiego.

Jeśli funkcja f nie jest kwadratowa, ale ma ciągłe drugie pochodne, to można obliczyć gradient $G(x)$ i hesjan $H(x)$. Może to zająć nam lub komputerowi sporo czasu, ale znajomość G i H daje oczywiste przybliżenie minimum, wynikające z metody Newtona:

$$x^{(k+1)} = x^{(k)} - H(x^{(k)})^{-1}G(x^{(k)}).$$

Wiadomo jednak, że działa ona zadowalająco dopiero wtedy, gdy punkt $x^{(k)}$ znajduje się w dostatecznie małym otoczeniu rozwiązania. Dlatego w algorytmie Biggsa i Dixona (zob. Dahlquist i Björck [1974], w pol. wyd. s. 426) punkt $v^{(k)} := H(x^{(k)})^{-1}G(x^{(k)})$ służy tylko do wyznaczenia kierunku poszukiwań; będą one prowadzone wzdłuż prostej $\{x^{(k)} + tv^{(k)} : t \in \mathbb{R}\}$. Wyznaczywszy na niej nowe przybliżenie $x^{(k+1)}$, postępujemy jak poprzednio. W wersji *spowolnionej* wektor kierunkowy $v^{(k)}$ zmieniamy na $u^{(k)} := G(x^{(k)})$ (jak w metodzie najszybszego spadku), gdy macierz $H(x^{(k)})$ jest osobliwa lub gdy

$$G(x^{(k)})^{\mathsf{T}}v^{(k)} < \max\{0, \|G(x^{(k)})\|^4/[G(x^{(k)})^{\mathsf{T}}H(x^{(k)})G(x^{(k)})]\}.$$

11.4. Algorytmy aproksymacji kwadratowej

W metodach z tej szerokiej kategorii na każdym etapie obliczeń aktualizujemy funkcję kwadratową n zmiennych, modyfikując przybliżenie hesjanu tej funkcji lub jego odwrotności. Przyjmuje się standardowo, że – w przeciwieństwie do hesjanu – postać analityczna gradientu jest znana. Oczywistym żądaniem jest, aby metoda szybko lokalizowała minimum funkcji kwadratowej.

Poznamy teraz sławny *algorytm Davidona, Fletchera i Powella*. Jak dotąd, $G(x)$ oznacza gradient funkcji f w punkcie x, a więc pewien wektor kolumnowy. Prócz tego J będzie oznaczać przybliżenie odwrotności hesjanu $H(x)$. Na początku, dla danego $x^{(1)}$, przyjmujemy, że J jest macierzą jednostkową I stopnia n. k-ty etap algorytmu Davidona, Fletchera i Powella ($k \geqslant 1$) jest następujący:

ALGORYTM 11.4.1

1. $u \leftarrow -JG(x^{(k)})$.
2. *Wybrać t dające przybliżone minimum funkcji f na promieniu $x^{(k)} + tu$.*
3. $v \leftarrow tu$.

4. $x^{(k+1)} \leftarrow x^{(k)} + v$.

5. $y \leftarrow G(x^{(k+1)}) - G(x^{(k)})$.

6. $J \leftarrow J + (vv^{\mathsf{T}})/(y^{\mathsf{T}}v) - (Jy)(Jy)^{\mathsf{T}}/(y^{\mathsf{T}}Jy)$.

Mamy tu metodę sprzężonych gradientów podobną do tych, które opisano w podrozdz. 4.7. Niech będzie

$$f(x) := a - b^{\mathsf{T}}x + \tfrac{1}{2}x^{\mathsf{T}}Ax.$$

Jeśli algorytm spadku jest określony poprzez układ kierunków poszukiwania $v^{(1)}, v^{(2)}, \ldots$, to dla dowolnego $x^{(1)}$ i dla $k \geqslant 1$ definiujemy $x^{(k+1)}$ jako punkt minimum funkcji f na promieniu $x^{(k)} + tv^{(k)}$. Na mocy tw. 4.7.3, jeśli kierunki $v^{(k)}$ tworzą układ A-ortogonalny (tj. $(v^{(i)})^{\mathsf{T}}Av^{(j)} = 0$ dla $i \neq j$), to minimum funkcji f otrzymujemy nie później niż w $(n+1)$-szym kroku. Omawiana teraz metoda buduje taki właśnie układ z wektorów u, obliczanych zgodnie z instrukcją **1**.

Prace oryginalne opisujące metodę: Davidon [1959] oraz Fletcher i Powell [1963]. W podręczniku Luenbergera [1973] można znaleźć istotne twierdzenia dotyczące tej metody.

ZADANIA 11.4

1. Udowodnić, że jeśli każdy z wektorów $v^{(k)}$, o których mowa na końcu podrozdziału, jest ortogonalny względem $G(x^{(j-1)}) - G(x^{(j)})$ dla $1 \leqslant j < k$, to ich układ jest A-ortogonalny.

11.5. Algorytm Neldera-Meada

Tytułowy algorytm służący do wyznaczania minimum funkcji $f\colon \mathbb{R}^n \to \mathbb{R}$ jest metodą bezpośrednią w tym sensie, że obywa się bez badania tej funkcji na prostych, a także bez obliczania jej pochodnych.

Na początku obliczeń użytkownik nadaje wartości trzem parametrom: $\alpha > 0$, $\beta \in (0,1)$ i $\gamma > 0$. Można np. przyjąć, że te wartości są odpowiednio równe 1, $\frac{1}{2}$ i 1. W każdym kroku obliczeń jest dany układ $n+1$ punktów z $\mathbb{R}^n$:

$$\{x^{(0)}, x^{(1)}, \ldots, x^{(n)}\}. \tag{11.5.1}$$

Powinny one być takie, żeby punkty $x^{(i)} - x^{(0)}$ $(1 \leqslant i \leqslant n)$ były niezależne liniowo. Wtedy powłoka wypukła układu (11.5.1) jest n-sympleksem. W szczególności 2-sympleks jest trójkątem w $\mathbb{R}^2$, a 3-sympleks jest czworościanem w $\mathbb{R}^3$. Aby uprościć opis algorytmu, przyjmujemy, że w każdym etapie obliczeń punkty $x^{(i)}$ są uporządkowane tak, że

$$f(x^{(0)}) \geqslant f(x^{(1)}) \geqslant \ldots \geqslant f(x^{(n)}) \tag{11.5.2}$$

(ich przestawiania można uniknąć operując odpowiednio wskaźnikami, przy czym – jak się okaże – istotne są tylko wskaźniki największego $f(x^{(i)})$, drugiego z kolei i najmniejszego).

Ponieważ naszym celem jest minimalizacja funkcji f, więc punkt $x^{(0)}$ jest najgorszy z $n+1$ danych. Dlatego na ogół skutkiem kolejnego etapu obliczeń jest zamiana tego właśnie punktu na lepszy.

Na początku każdego etapu tworzymy trzy punkty:

$$u := \frac{1}{n}\sum_{i=1}^{n} x^{(i)}, \quad v := (1+\alpha)u - \alpha x^{(0)}, \quad w := (1+\gamma)v - \gamma u$$

(u jest środkiem ściany sympleksu przeciwległej w stosunku do najgorszego punktu). Konstrukcja nowego sympleksu zależy od wartości funkcji w nowych punktach. Rozróżniamy trzy przypadki.

(i) $f(v) < f(x^{(n)})$. Jeśli $f(w) < f(x^{(n)})$, to zmieniamy $x^{(0)}$ na w, a w przeciwnym razie na v.

(ii) $f(v) \geqslant f(x^{(n)})$ i $f(v) \leqslant f(x^{(1)})$. Zmieniamy $x^{(0)}$ na v.

(iii) $f(v) > f(x^{(1)})$. Jeśli $f(v) \leqslant f(x^{(0)})$, to zmieniamy $x^{(0)}$ na v. Definiujemy na nowo punkt w: $w := \beta x^{(0)} + (1-\beta)u$. Jeśli $f(w) \leqslant f(x^{(0)})$, to zmieniamy $x^{(0)}$ na w, a w przeciwnym razie dla $0 \leqslant i \leqslant n-1$ zmieniamy $x^{(i)}$ na $\frac{1}{2}(x^{(i)} + x^{(n)})$ (ściągnięcie sympleksu).

Po uporządkowaniu punktów według (11.5.2) obliczamy wartość ilorazu

$$\frac{f(x^{(0)}) - f(x^{(n)})}{|f(x^{(0)})| + |f(x^{(n)})|}.$$

Jeśli jest on dostatecznie mały, to uznajemy, że minimum obliczono. W przeciwnym razie zaczynamy nowy etap od wyznaczenia u, v i w.

Powyższy opis jest wzorowany na oryginalnej pracy Neldera i Meada [1965]; zob. też Dennis i Woods [1987] oraz Torczon [1997]. Różni autorzy modyfikują pierwotny algorytm. Udoskonaloną wersję, zwaną *wzmocnionym spadkiem*, podał Tseng [1998]; zob. też Nazareth i Tseng [1998].

11.6. Wyżarzanie symulowane

Ta metoda okazała się efektywna dla „trudnych" funkcji, w szczególności takich, które mają wiele minimów lokalnych. Niech funkcja f ma minimum globalne w x^*. Metoda daje takie punkty $x^{(1)}, x^{(2)}, \ldots$, że – jak można się spodziewać – ciąg liczb $\min_{j<k} f(x^{(j)})$ dąży do $f(x^*)$, gdy $k \to \infty$.

Wystarczy wyjaśnić, jakie obliczenia prowadzą do $x^{(k+1)}$, gdy dane jest $x^{(k)}$. W dużym otoczeniu tego ostatniego punktu generujemy umiarkowanie dużą liczbę punktów losowych $u^{(1)}, u^{(2)}, \ldots, u^{(m)}$ i obliczamy w nich wartości funkcji f. Niech $f(u^{(j)})$ będzie najmniejszą z nich. Jeśli $f(u^{(j)}) < f(x^{(k)})$, to przyjmujemy $x^{(k+1)} := u^{(j)}$. W przeciwnym razie każdemu $u^{(i)}$ przypisujemy prawdopodobieństwo

$$p_i := \exp\{\alpha[f(x^{(k)}) - f(u^{(i)})]\} \quad (1 \leqslant i \leqslant m).$$

α jest tu stałą dodatnią, wybraną przez użytkownika metody. Normalizujemy wielkości p_i, dzieląc je przez ich sumę S: $p_i \leftarrow p_i/S$. Następnie za punkt $x^{(k+1)}$ uznajemy pewien punkt $u^{(i)}$ wybrany losowo z uwzględnieniem tych prawdopodobieństw. Najprostszy sposób takiego wyboru polega na tym, że dla danej liczby losowej ξ z przedziału $(0,1)$ określamy i jako najmniejszą liczbę całkowitą, dla której $\xi \leqslant p_1 + p_2 + \ldots + p_i$.

Powyższy wzór dla p_i wywodzi się z zasad termodynamiki. Jego uzasadnienie można znaleźć w oryginalnej pracy Metropolisa i in. [1953] lub w książce Ottena i van Ginnekena [1989]. Możliwe są zapewne i inne określenia tych wielkości.

Celem skomplikowanej definicji nowego punktu $x^{(i)}$ jest unikanie minimów lokalnych. Dlatego algorytm musi niekiedy wybrać punkt, w którym wartość funkcji jest większa niż w bieżącym punkcie. Wtedy jest szansa na to, że kolejne punkty będą zmierzać do innego minimum lokalnego.

Po nieznacznych modyfikacjach metoda nadaje się do szukania minimum funkcji $f\colon X \to \mathbb{R}$, gdzie X jest dowolnym zbiorem. Tak np. w zagadnieniu komiwojażera X jest zbiorem wszystkich permutacji układu liczb całkowitych $1, 2, \ldots, N$. Wtedy wystarczy dysponować procedurą, która generuje losowo permutacje i procedurą obliczania wartości funkcji f.

Programy w Fortranie realizujące ten algorytm można znaleźć na stronie[3)]

`http://www.netlib.org`

11.7. Algorytmy genetyczne

Aktualne badania w zakresie teorii optymalizacji mają na celu zapobieżenie takiej sytuacji, że obliczenia kończą się znalezieniem lokalnego minimum, które nie jest minimum globalnym. Te badania doprowadziły do klasy metod zwanych *algorytmami genetycznymi.* Mają one swój początek w naukach

[3)] Podany adres strony internetowej dotyczy oryginału książki, tzn. wydania w jęz. ang. (*przyp. red. WNT*).

biologicznych. Algorytmy zawierają element losowy, wymuszający szukanie minimów z dala od aktualnego minimum lokalnego.

Tę klasę algorytmów opisał Holland [1989], a nieco później ukazała się książka Lawrence'a [1991]. Algorytmy posługują się tylko wartościami funkcji (w typowym przypadku określonych na $\mathbb{R}^n$); pochodne nie są potrzebne. Algorytm genetyczny dla zagadnienia komiwojażera podał Karloy [1993].

Szkicując działanie takich algorytmów, załóżmy, że funkcja f określona na $\mathbb{R}^n$ ma wartości nieujemne. Minimum jest nieznane, ale na pewno nieujemne. Nazwijmy więc $f(x)$ *ułomnością* punktu x. Szukamy najmniej ułomnego punktu. Kombinując pary punktów, tworzymy nowe. Ich konstrukcja jest bardziej prawdopodobna, gdy ułomność punktów łączonych w parę jest niska. Sposób tworzenia nowych punktów zależy od specyfiki zagadnienia. Algorytm zaczyna się od utworzenia losowego układu punktów za pomocą odpowiednich generatorów liczb losowych. Dodatkowa wiedza o zagadnieniu może wpłynąć na rozmieszczenie tych punktów. Kombinowanie par punktów może przebiegać równolegle.

11.8. Programowanie wypukłe

Tą nazwą obejmujemy zagadnienia optymalizacji, w których funkcja jest wypukła i taki jest też obszar, gdzie szukamy minimum. Zbiory wypukłe określono w podrozdz. 6.9. Funkcję f nazywamy *wypukłą*, jeśli dla $\alpha \geqslant 0$, $\beta \geqslant 0$, $\alpha + \beta = 1$ i dowolnych x i y z obszaru, w którym funkcja jest określona,

$$f(\alpha x + \beta y) \leqslant \alpha f(x) + \beta f(y).$$

Funkcję wypukłą określoną na podzbiorze wypukłym w $\mathbb{R}^n$ można przybliżać funkcjami postaci

$$F(x) := \max_{1 \leqslant i \leqslant k} \Big[\sum_{j=1}^{n} a_{ij} x_j - b_i \Big].$$

Analogicznie, obszar wypukły można aproksymować zbiorem

$$K := \{x\colon\ G(x) \leqslant 0\},$$

gdzie G jest funkcją wypukłą podobną do F; niech będzie

$$G(x) := \max_{k+1 \leqslant i \leqslant m} \Big[\sum_{j=1}^{n} a_{ij} x_j - b_i \Big].$$

Zgodnie z tym rozważmy zagadnienie opisane przez parę (F, G), a polegające na minimalizacji funkcji F pod warunkiem, że $G(x) \leqslant 0$. Inaczej mówiąc, pełny zestaw danych o zagadnieniu obejmuje macierz A, wektor b i wskaźnik k. Założymy, że A spełnia *warunek Haara*, tzn. że każdy układ jej n wierszy jest niezależny liniowo. Założymy także, że zbiór K jest niepusty. Prócz tego niech będzie

$$A_i := (a_{i1}, a_{i2}, \ldots, a_{in}) \quad (1 \leqslant i \leqslant m),$$
$$r_i(x) := \sum_{j=1}^{n} a_{ij}x_j - b_i = \langle A_i, x\rangle - b_i \quad (x \in \mathbb{R}^n,\ 1 \leqslant i \leqslant m).$$

W każdym kroku algorytmu występuje zbiór wskaźników $J \subset \{1, 2, \ldots, m\}$. Zakłada się, że ten zbiór ma dokładnie $n+1$ elementów, co najmniej jeden z nich jest nie większy od k, a punkt 0 należy do powłoki wypukłej zbioru $\{A_i\colon\ i \in J\}$. W bieżącym kroku znajdujemy najpierw wektor $x \in \mathbb{R}^n$ i liczbę rzeczywistą λ, dla których $r_i(x) = \lambda$, gdy $i \in J$ i $i \leqslant k$, ale $r_i(x) = 0$, gdy $i \in J$ i $i > k$. Wymaga to rozwiązania układu $n+1$ równań liniowych z tyluż niewiadomymi (są nimi składowe wektora x i liczba λ). Znając już x i λ, obliczamy $F(x)$ i $G(x)$. Jeśli $G(x) \leqslant 0$ i $F(x) = \lambda$, to obliczenia się kończą, a x jest punktem, w którym F osiąga minimum na K. Jeśli $G(x) > 0$, to $x \notin K$ i wybieramy wskaźnik α taki, że $\alpha > k$ i $r_\alpha(x) > 0$. Jeśli $G(x) \leqslant 0$ i $F(x) > \lambda$, to wybieramy α tak, żeby było $\alpha \leqslant k$ i $r_\alpha(x) > \lambda$. Na mocy tw. 6.9.24 możemy znaleźć taki wskaźnik $\beta \in J$, że punkt 0 należy do powłoki wypukłej zbioru $\{A_i\colon\ i \in J'\}$, gdzie $J' := J \cup \{\alpha\} \setminus \{\beta\}$. W następnym kroku zamiast J używamy zbioru J'.

Twierdzenie 11.8.1. *Po skończonej liczbie kroków powyższy algorytm daje punkt $x \in K$, w którym funkcja F osiąga minimum.*

Dowód tego twierdzenia podaje Goldstein [1967].

11.9. Minimalizacja z warunkami

W poprzednim podrozdziale mieliśmy już do czynienia z minimalizacją funkcji pod pewnymi warunkami. Jeśli ani funkcja, ani obszar nie są wypukłe, to pojęciowo prosta metoda odwołuje się do pojęcia *funkcji kary*. Jest to funkcja o bardzo dużych wartościach tam, gdzie warunki nie są spełnione. Po dodaniu funkcji kary do funkcji celu zagadnienie sprowadza się do takiego, w którym dodatkowych warunków brak i które można rozwiązywać metodami opisanymi wcześniej.

Ściślej mówiąc, załóżmy, że szukamy minimum funkcji celu f na zbiorze $K := \{x\colon\ g(x) \leqslant 0\}$. Wtedy równoważnym zagadnieniem bez warunków jest minimalizacja sumy $f + g^*$, gdzie $g^*(x) := \infty$ dla $g(x) > 0$ i $g^*(x) := 0$ dla $g(x) \leqslant 0$. Ta sztuczna funkcja g^* jest jednak zbyt prymitywna w procedurach numerycznych. Zamiast niej trzeba raczej stosować funkcję, która zmienia się szybko ale gładko, gdy wartości $g(x)$ przechodzą przez zero. Możemy np. przyjąć, że $g^*(x) := \exp g(x) - 1$. Inną sensowną funkcją kary jest $g^*(x) := \max\{0, \alpha g(x)\}$, gdzie α jest parametrem dodatnim dobranym przez użytkownika metody.

11.10. Optymalizacja Pareto

Rozważmy układ skończony funkcji $f_1, f_2, \ldots, f_n$ o wartościach rzeczywistych. *Optymalizacja Pareto* polega na poszukiwaniu punktu x^* takiego, że żadne $y \neq x^*$ nie spełnia nierówności $f_i(y) < f_i(x^*)$ dla $1 \leqslant i \leqslant n$. Takie x^* nazywamy *punktem optymalnym Pareto* dla danego układu funkcji.

Są tu znów potrzebne pojęcia wypukłości zbiorów i wypukłości funkcji (podrozdz. 11.8). Następujące twierdzenia, zaczerpnięte z książki Aubina [1998], dają łącznie charakteryzację punktu optymalnego Pareto.

Twierdzenie 11.10.1. *Niech $f_1, f_2, \ldots, f_n$ będą funkcjami wypukłymi o wartościach rzeczywistych, określonymi na zbiorze wypukłym w pewnej przestrzeni wektorowej. Jeśli x^* jest punktem optymalnym Pareto tego układu funkcji, to istnieją liczby nieujemne $h_1, h_2, \ldots, h_n$ nie wszystkie równe 0, dla których suma $\sum_{i=1}^n h_i f_i(x)$ osiąga minimum w x^*.*

Dowód. Niech dziedziną wypukłą rozważanych funkcji będzie X. Funkcję $f\colon\ X \to \mathbb{R}^n$ określamy wzorem

$$f(x) := (f_1(x), f_2(x), \ldots, f_n(x)).$$

W $\mathbb{R}^n$ relacje $u \leqslant v$ i $u < v$ oznaczają, że odpowiednio $u_i \leqslant v_i$ i $u_i < v_i$ dla każdego i. Definiujemy zbiór

$$K := \{u \in \mathbb{R}^n\colon\ u > f(x) \quad \text{dla pewnego } x \in X\}.$$

Wykażemy, że jest on wypukły. Niech będzie $u, v \in K$. Wtedy istnieją takie x i y, że $u > f(x)$ i $v > f(y)$. Jeśli $0 \leqslant \lambda \leqslant 1$ i $\theta = 1 - \lambda$, to z wypukłości każdej z funkcji f_i wynika, że

$$\lambda u_i + \theta v_i > \lambda f_i(x) + \theta f_i(y) \geqslant f_i(\lambda x + \theta y) = f_i(w),$$

gdzie $w := \lambda x + \theta y$. Dzięki wypukłości zbioru X ten punkt doń należy. Powyższe nierówności dają relację $\lambda u + \theta v > f(w)$. Wobec tego $\lambda u + \theta v \in K$, czyli zbiór K jest wypukły.

Niech x^* będzie punktem optymalnym Pareto. Jest $f(x^*) \notin K$. Istotnie, w przeciwnym razie zachodziłaby nierówność $f(x^*) > f(x)$ dla pewnego $x \in X$, a to byłoby sprzeczne z definicją punktu x^*. Na mocy tw. 10.1.5 istnieje wektor niezerowy $h \in \mathbb{R}^n$ taki, że $\langle h, f(x^*)\rangle \leqslant \langle h, u\rangle$ dla każdego $u \in K$. Ponieważ K zawiera wszystkie wektory postaci $f(x) + p$ dla $x \in X$, $p \in \mathbb{R}^n$ i $p > 0$, więc

$$\langle h, f(x^*)\rangle \leqslant \langle h, f(x)\rangle + \langle h, p\rangle.$$

Dla p o dużych składowych wynika stąd, że $h \geqslant 0$. Biorąc kres dolny względem p, wnioskujemy, że

$$\langle h, f(x^*)\rangle \leqslant \langle h, f(x)\rangle,$$

a to znaczy, że suma $\sum_{i=1}^n h_i f_i(x)$ osiąga minimum dla $x = x^*$. ■

Twierdzenie 11.10.2. *Jeśli w x^* jedno z wyrażeń $\max_i f_i(x)$, $\sum_{i=1}^n \theta_i f_i(x)$, gdzie $\theta_i \geqslant 0$ i $\sum_{i=1}^n \theta_i > 0$, osiąga minimum, to x^* jest punktem optymalnym Pareto dla układu funkcji $f_1, f_2, \ldots, f_n$.*

Dowód. Jeśli x^* nie jest punktem optymalnym Pareto, to istnieje takie y, że $f_i(y) < f_i(x^*)$ dla każdego i. Wtedy $\max_i f_i(y) < \max_i f_i(x^*)$, czyli $\max_i f_i(x)$ nie jest najmniejsze dla $x = x^*$. Podobny dowód dotyczy drugiego z wyrażeń w twierdzeniu. ■

Przykład 11.10.3. Niech dla $x \in \mathbb{R}$ będzie $f_1(x) := x^2$ i $f_2(x) = (1-x)^2$. Punktem optymalnym Pareto układu $\{f_1, f_2\}$ jest $x^* = \frac{1}{2}$. Istotnie, jeśli $f_i(y) < f_i(x^*)$, to $|y| < \frac{1}{2}$ i $|y-1| < \frac{1}{2}$, a te nierówności są sprzeczne. Zbiór K z dowodu tw. 11.10.1 jest tu równy

$$K = \{u \in \mathbb{R}^2 \colon u_1 > x^2 \text{ i } u_2 > (x-1)^2 \text{ dla pewnego } x\}.$$

Hiperpłaszczyzna o równaniu $u_1 + u_2 = \frac{1}{2}$ oddziela K od $f(x^*) = (\frac{1}{2}, \frac{1}{2})$. Dlatego, zgodnie z dowodem tw. 11.10.2, suma $f_1 + f_2$ powinna osiągać minimum w x^* i tak rzeczywiście jest. To samo dotyczy wyrażenia $\max_i f_i$. ■

Zadania 11.10

1. Udowodnić, że jeśli funkcje f_i są wypukłe, to tę samą własność mają funkcje $\max_i f_i$ i $\sum_{i=1}^n \lambda_i f_i$, gdzie $\lambda_i \geqslant 0$.

Bibliografia

Uwaga: *Gwiazdka * przed rokiem wydania sygnalizuje pozycję bibliografii dodaną przez tłumacza.*

Skróty nazw:

ACM	Association for Computing Machinery
ACM-COM	ACM Communications
ACM-J	ACM Journal
ACM-TOMS	ACM Transactions on Mathematical Software
AMC	Applied Mathematics and Computation
AMM	American Mathematical Monthly
AMS	American Mathematical Society
AMS-B	AMS Bulletin
ANM	Applied Numerical Mathematics (IMACS)
ANSI	American National Standards Institute, Inc.
BIT	BIT Numerical Mathematics
CA	Constructive Approximation
CANM	Communications on Applied Numerical Methods
CJ	Computing Journal
CMP	Communications in Mathematical Physics
IEEE	Institute of Electrical and Electronic Engineers
IEEE-TAE	IEEE Transactions on Audio and Electroacoustics
IMA	Institute for Mathematics and Its Applications
IMA-JNA	IMA Journal of Numerical Analysis
IU-JM	Indiana University Journal of Mathematics
JAT	Journal of Approximation Theory
JCAM	Journal of Computational and Applied Mathematics
JCP	Journal of Computational Physics
JMP	Journal of Mathematical Physics

JR-NBS	Journal of Research in the National Bureau of Standards
LNCS	Lecture Notes in Computer Science
LNM	Lecture Notes in Mathematics
MAA	Mathematical Association of America
MI	Mathematics Intelligencer
MOC	Mathematics of Computation
NM	Numerische Mathematik
SA	Scientific American
SIAM	Society for Industrial and Applied Mathematics
SIAM-JO	SIAM Journal on Optimization
SIAM-MAA	SIAM Journal on Matrix Analysis and Applications
SIAM-NA	SIAM Journal on Numerical Analysis
SIAM-REV	SIAM Review
SIAM-SSC	SIAM Journal on Scientific and Statistical Computing
ZAMP	Zeitschrift für Angewandte Mathematik und Physik

ABRAMOWITZ M., STEGUN I. A. (eds.) 1964. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Nat. Bureau of Standards (przedruk: 1965, New York, Dover).

ACHIEZER N. I. *1957. *Teoria aproksymacji.* Warszawa, PWN.

AHLFORS L. V. 1966. *Complex Analysis.* New York, Mc Graw-Hill.

ALEFELD G., GRIGORIEFF R. (eds.) 1980. *Fundamentals of Numerical Computation.* Berlin, Springer.

ALEFELD G., HERZBERGER J. 1983. *Introduction to Interval Computations.* New York, Academic Press.

ALLGOWER E. L., GLASSHOFF K., PEITGEN H.-O. (eds.) 1981. *Numerical Solution of Nonlinear Equations. LNM* **878**. New York, Springer.

ANDERSON E., BAI Z., BISCHOF C., DEMMEL J., DONGARRA J., DU CROZ J., GREENBAUM A., HAMMARLING S., MCKENNEY A., OSTROUCHOV S., SORENSEN D. 1995. *LAPACK Users' Guide – Release 2.0.* Philadelphia, SIAM. Wersja online: `http://www.netlib.org/lapack/lug/lapack_lug.html`

ANSI/IEEE 1985. IEEE Standard for binary floating-point arithmetic. ANSI/IEEE Std. 754–1985. New York, IEEE.

ANSI/IEEE 1987. A radix-independent standard for floating-point arithmetic. ANSI/IEEE Std. 854-1987. New York, IEEE.

AUBIN J. P. 1998. *Optima and Equilibria: An Introduction to Nonlinear Analysis,* 2nd Ed. Berlin, Springer.

BAK J., NEWMAN D. J. 1982. *Complex Analysis.* New York, Springer.

BAKER G. A., JR. *1975. *Essentials of Padé Approximants.* London, Academic Press.

BAKER G. A., JR., GRAVES-MORRIS P. *1981. *Padé Approximants.* Part I: *Basic Theory.* Part II: *Extensions and Applications.* Reading, Addison-Wesley.

BARNHILL R., DUBE R. P., LITTLE F. F. 1983. Properties of Shepard's surfaces. *Rocky Mtn. J. Math.* **13**, s. 365–382.

BARNSLEY M. 1988. *Fractals Everywhere.* New York, Academic Press.

BARTLE R. G. 1976. *The Elements of Real Analysis*, 2nd ed. New York, Wiley.

BECKER E. B., CAREY G. F., ODEN J. T. 1981. *Finite Elements: An Introduction*, vol. 1. Englewood Cliffs, Prentice-Hall.

BHATTI M. A. 2000. *Practical Optimization Methods.* New York, Springer.

BIERNAT J. *2001. *Metody i układy arytmetyki komputerowej.* Wrocław, Oficyna Wyd. Polit. Wrocł.

BISCHOF C., CARLE A., KHADEMI P., MAUER A. 1994. The ADIFOR 2.0 system for the automatic differentiation of Fortran 77 programs. *Mathematics and Computer Sciences Rep.* ANL/MCS-P481-1194. Argonne, Argonne Nat. Laboratory.

BLOOMFIELD P. 1976. *Fourier Analysis of Time Series: An Introduction.* New York, Wiley.

BODEWIG E. 1946. Sur la méthode de Laguerre pour l'approximation des racines de certaines équations algébriques et sur la critique d'Hermite. *Nederl. Acad. Wetensch. Proc.* **49**, s. 911–921.

DE BOOR C. 1976. Total positivity of the spline collocation matrix. *IU-JM* **25**, s. 541–551.

DE BOOR C. 1984. *A Practical Guide to Splines*, 2nd ed. New York, Springer.

BORWEIN J., LEWIS A. S. 2000. *Convex Analysis and Nonlinear Optimization.* New York, Springer.

BRAESS D. 1984. *Nonlinear Approximation Theory.* New York, Springer.

BRENT R. P. 1973. *Algorithms for Minimization without Derivatives.* Englewood Cliffs, Prentice-Hall.

BREZINSKI C., REDIVO ZAGLIA M. *1991. *Extrapolation Methods. Theory and Practice.* Amsterdam, North Holland.

BRIGGS W. T., HENSON V. E. 1995. *The DFT: An Owner's Manual for the Discrete Fourier Transform.* Philadelphia, SIAM.

BRIGHAM E. O. 1974. *The Fast Fourier Transform.* Englewood Cliffs, Prentice-Hall.

BROPHY J. F., SMITH P. W. 1988. Prototyping Karmarkar's algorithm using `MATH/PROTAN`. *Directions* **5**, s. 2–3. Houston, Visual Numerics.

BUCHANAN J. L., TURNER P. R. 1992. *Numerical Methods and Analysis.* New York, McGraw-Hill.

BUTCHER J. C. 1987. *The Numerical Analysis of Ordinary Differential Equations: Runge-Kutta and General Linear Methods.* New York, Wiley.

BYRNE G. D., HALL C. A. (eds.) 1973. *Numerical Solution of Systems of Nonlinear Algebraic Equations.* New York, Academic Press.

BYRNE G., HINDMARSH A. 1987. Stiff ODE solvers: A review of current and coming attractions. *JCP* **70**, s. 1–62.

CASH J. R. 1979. *Stable Recursions.* New York, Academic Press.

CHENEY W., LIGHT W. 1999. *A Course in Approximation Theory.* Pacific Grove, Brooks/Cole.

CHUI C. K. 1988. Multivariate splines. *SIAM Regional Conf. Series in Mathematics*, Vol. 54.

CHUNG K. C., YAO T. H. 1977. On lattices admitting unique Lagrange interpolations. *SIAM-NA* **14**, s. 735–743.

CLINE A. K. 1974a. Scalar and planar valued curve-fitting using splines under tension. *ACM-COM* **17**, s. 218–220.

CLINE A. K. 1974b. Six subprograms for curve-fitting using splines under tension. *ACM-COM* **17**, s. 220–223.

CODDINGTON E. A., LEVINSON N. 1955. *Theory of Ordinary Differential Equations.* New York, McGraw-Hill.

CODY W. J. 1988. Floating-point standards – theory and practice. W: *Reliability in Computing*, New York, McGraw-Hill, s. 99–107.

CONTE S. D., DE BOOR C. 1980. *Elementary Numerical Analysis*, 3rd ed. New York, Mc Graw-Hill.

COOLEY J. W., LEWIS P. A., WELCH P. P. 1967. Historical notes on the fast Fourier transform. *Proc. IEEE* **55**, s. 1675–1677.

COONEN J. T. 1981. Underflow and the denormalized numbers. *Computer* **14**, s. 75–87.

CORNUEJOLS G. ET AL. (eds.) 2000. *Integer Programming and Combinatorial Optimization.* New York, Springer.

CRANK J., NICOLSON P. *1947. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Proc. Cambridge Phil. Soc.* **43**, s. 50–67.

CRYER C. W. 1968. Pivot size in Gaussian elimination. *NM* **12**, s. 335–345.

CURRY J. H., GARNETT L., SULLIVAN D. 1983. On the iteration of a rational function: Computer experiments with Newton's method. *CMP* **91**, s. 267–277.

CUYT, WUYTACK *1987. *Nonlinear Methods in Numerical Analysis.* Amsterdam, North-Holland.

DAHLQUIST G. 1956. Convergence and stability in the numerical integration of ordinary differential equations. *Math. Scand.* **4**, s. 33–35.

DAHLQUIST G. 1963. A special stability problem for linear multistep methods. *BIT* **3**, s. 27–43.

DAHLQUIST G., BJÖRCK A. 1974. *Numerical Methods.* Englewood Cliffs, Prentice-Hall. Przekład polski: Björck A., Dahlquist G., *Metody numeryczne*, wyd. 2, PWN, Warszawa 1987.

DANTZIG G. B. 1948. *Programming in a linear structure.* Washington, U.S. Air Force, Comptroller's Office.

DAVIDON W. C. 1959. Variable metric method for minimization. *Research and Development Rep.* ANL-5990 (Rev.), Atomic Energy Commission.

DAVIS P. J. 1982. *Interpolation and Approximation.* New York, Dover.

DAVIS P. J., RABINOWITZ P. 1956. Abscissas and weights for Gaussian quadratures of high order. *JR-NBS* **56**, s. 35-37.

DAVIS P. J., RABINOWITZ P. 1984. *Methods of Numerical Integration*, 2nd ed. New York, Academic Press.

DEKKER T. J. 1969. Finding a zero by means of successive linear interpolation. W: *Constructive Aspects of the Fundamental Theorem of Algebra*, Dejon B., Henrici P. (eds.). New York, Wiley.

DENNIS J. E., JR., WOODS D. J. 1987. Optimization on microcomputers: The Nelder-Mead simplex algorithm. W: *New Computing Environments*, A. Wouk (ed.). Philadelphia, SIAM.

DEWNDEY A. K. 1988. Computer recreations: Random walks that lead to fractal crowds. *SA*, Dec.

DIEKMANN O., VAN GILS S. A., VERDUYN LUNEL S. M., WALTHER H. O. 1995. *Delay Equations.* New York, Springer.

DIEUDONNÉ J. 1960. *Foundations of Modern Analysis.* New York, Academic Press.

DONGARRA J. J., BUNCH J. R., MOLER C. B., STEWART G. W. 1979. *LINPACK Users Guide.* Philadelphia, SIAM.

DONGARRA J. J., WALKER D. W. 1995. Software libraries for linear algebra computations on high performance computers. *SIAM-REV* **37**, s. 151–180.

DRIVER R. 1977. *Ordinary and Delay Differential Equations.* New York, Springer.

DRYJA M., JANKOWSCY, J. i M. *1982. *Przegląd metod i algorytmów numerycznych*, cz. II. Warszawa, WNT.

DURAND E. 1960. *Solutions Numériques des Équations Algébriques* (2 vols.). Paris, Mason.

EAVES B. C., GOULD F. J., PEITGEN H.-O., TODD M. J. (eds.) 1983. *Homotopy Methods and Global Convergence.* New York, Plenum.

EDELMAN A. 1992. *The complete pivoting conjecture for Gaussian elimination is false.* Dept. Math. Berkeley, Lawrence Berkeley National Laboratory and University of California.

EDELMAN A. 1994. *When is* $x * (1/x) \neq 1$*?* Dept. of Mathematics. Cambridge, MIT; praca dostępna na stronie internetowej
`http://www-math.mit.edu/~edelman/Edelman/publications.htm`

EGGERMONT P. P. B. 1988. Noncentral difference quotients and the derivative. *AMM* **95**, s. 551–553.

ELLIOTT D. F., RAO K. R. 1982. *Fast Transforms: Algorithms, Analyses, Applications.* New York, Academic Press.

EPPERSON J. F. 1987. On the Runge example. *AMM* **4**, s. 329–341.

FARWIG R. 1986. Rate of convergence of Shepard's global interpolation formula. *MOC* **46**, s. 577–590.

FEFFERMAN C. 1967. An easy proof of the fundamental theorem of algebra. *AMM* **74**, s. 854–855.

FEHLBERG E. 1969. Klassische Runge-Kutta Formeln fünfter und siebenter Ordnung mit Schrittweitenkontrolle. *Computing* **4**, s. 93–106.

FELDSTEIN A., TURNER P. 1986. Overflow, underflow, and severe loss of significance in floating-point addition and substraction. *IMA-JNA* **6**, s. 241–251.

FLETCHER R., POWELL M. J. D. 1963. A rapidly convergent descent method for minimization. *CJ* **6**, s. 163–168.

FORSYTHE G. E., MALCOLM M. A., MOLER C. B. 1977. *Computer Methods for Mathematical Computations.* Englewood Cliffs, Prentice-Hall.

FORSYTHE G. E., MOLER C. B. 1967. *Computer Solution of Linear Algebraic Systems.* Englewood Cliffs, Prentice-Hall.

FOSDICK L. D. 1993. *IEEE Arithmetic Short Reference.* High Performance Scientific Computing. Boulder, Univ. of Colorado at Boulder.

FOSTER L. V. 1981. Generalizations of Laguerre's method: Higher order methods. *SIAM-NA* **18**, s. 1004–1018.

FOSTER L. V. 1994. Gaussian elimination with partial pivoting can fail in practice. *SIAM-MAA* **15**, s. 1354–1362.

FOX L. 1987. *Biographical Memoirs of Fellows of the Royal Society: James Hardy Wilkinson 1919–1986* **33**. London, Royal Society.

FOX L., PARKER I. B. *1968. *Chebyshev Polynomials in Numerical Analysis.* London, Oxford Univ. Press.

FRANCIS J. G. F. 1961. The QR transformation: A unitary analogue to the LR transformation. *CJ* **4**, s. 265–272 i 332–345.

FRANKE R. 1982. Scattered data interpolation. Tests of some methods. *MOC* **38**, s. 181–200.

GALEONE L. 1977. Generalizzazione del methodo di Laguerre. *Calcolo* **14**, s. 121–131.

GARCIA C. B., ZANGWILL W. I. 1981. *Pathways to Solutions, Fixed Points, and Equilibria.* Englewood Cliffs, Prentice-Hall.

GASCA M., MAEZTU J. I. 1982. On Lagrange and Hermite interpolation in $\mathbb{R}^k$. *NM* **39**, s. 1–14.

GAUTSCHI W. 1961. Recursive computation of certain integrals. *ACM-J* **8**, s. 21–40.

GAUTSCHI W. 1967. Computational aspects of three-term recurrence relations. *SIAM-REV* **9**, s. 24–82.

GAUTSCHI W. 1975. Computational methods in special functions. W: *Theory and Applications of Special Functions*, Askey R. (ed.). New York, Academic Press, s. 1–98.

GAUTSCHI W. 1979. Families of algebraic test equations. *Calcolo* **16**, s. 383–398.

GAUTSCHI W. 1984. Questions of numerical condition related to polynomials. W: *Studies in Numerical Analysis*, Golub G.H. (ed.). Washington, MAA, s. 140–177.

GEAR C.W. 1971. *Numerical Initial Value Problems in Ordinary Differential Equations.* Englewood Cliffs, Prentice-Hall.

GHIZZETTI A., OSSICINI A. 1970. *Quadrature Formulae.* New York, Academic Press.

GLIECK J. 1987. *Chaos.* New York, Viking Press.

GOLDSTEIN A. A. 1967. *Constructive Real Analysis.* New York, Harper & Row.

GOLUB G. H., O'LEARY D. P. 1989. Some history of the conjugate gradient and Lanczos methods. *SIAM-REV* **31**, s. 50–102.

GOLUB G. H., VAN LOAN C. F. 1989. *Matrix Computations*, 2nd ed. Baltimore, Hopkins Univ. Press.

GORDON W. J., WIXOM J. A. 1978. Shepard's method of 'metric interpolation' to bivariate and multivariate interpolation. *MOC* **32**, s. 253–264.

GOULD N. 1991. On growth in Gaussian elimination with complete pivoting. *SIAM-MAA* **12**, s. 354–361.

GREGORY R. T. 1980. *Error-Free Computation.* Huntington, Krieger.

GRIEWANK A., CORLISS G. F. 1991. *Automatic Differentiation of Algorithms: Theory, Implementation, and Applications.* Philadelphia, SIAM.

HAGEMAN L. A., YOUNG D. M. 1981. *Applied Iterative Methods.* New York, Academic Press.

HARDY G. H. 1960. *A Course of Pure Mathematics*, 10th ed. New York, Cambridge Univ. Press.

HARDY R. L. 1971. Multiquadric equations of topography and other irregular surfaces. *J. Geophysical Research* **76**, s. 1905–1915.

HARTLEY P. H. 1976. Tensor product approximations to data defined on rectangle meshes in n-space. *CJ* **19**, s. 348–352.

HENRICI P. 1962. *Discrete Variable Methods in Ordinary Differential Equations.* New York, Wiley.

HENRICI P. 1964. *Elements of Numerical Analysis.* New York, Wiley.

HENRICI P. 1974. *Applied and Computational Complex Analysis* (3 vols.). New York, Wiley.

HESTENES M. R., STIEFEL E. 1952. Methods of conjugate gradient for solving linear systems. *JR-NBS* **45**, s. 409–436.

HESTENES M. R., TODD J. 1991. *Mathematicians Learning to Use Computers.* Special Publ. 730. Gaithersburg, Nat. Inst. of Standards and Technology.

HIGHAM N. J. 2002. *Accuracy and Stability of Numerical Algorithms*, 2nd ed. Philadelphia, SIAM.

HIGHAM N. J., HIGHAM D. J. 1989. Large growth factors in Gaussian elimination with pivoting. *SIAM-MAA* **10**, s. 155–164.

HOLLAND J. H. 1989. Searching nonlinear functions for high values. *AMC* **32**, s. 255–274.

HOUGH D. 1981. Applications of the proposed IEEE 754 standard for floating-point arithmetic. *Computer* **14**, s. 70–74.

HOUSEHOLDER A. S. 1970. *The Numerical Treatment of a Single Nonlinear Equation.* New York, McGraw-Hill.

HULL T. E., ENRIGHT W. H., FELLEN B. M., SEDGWICK A. E. 1972. Comparing numerical methods for ordinary differential equations. *SIAM-NA* **9**, s. 603–637.

IMSL 1995. *International Mathematical and Statistical Libraries Reference Manual.* Houston, Visual Numerics.

ISAACSON E., KELLER H. B. 1966. *Analysis of Numerical Methods.* New York, Wiley.

JACKSON K. R., ENRIGHT W. H., HULL T. E. 1978. A theoretical criterion for comparing Runge-Kutta formulas. *SIAM-NA* **15**, s. 618–641.

JANKOWSCY J. i M. *1981. *Przegląd metod i algorytmów numerycznych*, cz. I. WNT, Warszawa.

JENKINS M. A., TRAUB J. F. 1970a. A three-stage algorithm for real polynomials using quadratic iteration. *SIAM-NA* **7**, s. 545–566.

JENKINS M. A., TRAUB J. F. 1970b. A three-stage variable-shift iteration for polynomial zeros. *NM* **14**, s. 252–263.

JONES W. B., THRON W. J. *1980. *Continued Fractions. Analytic Theory and Applications.* Reading, Addison-Wesley.

JOUBERT W. D., CAREY G. F., BERNER N. A., KALHAN A., KOHLI H., LORBER A., MCLAY R. T., SHEN Y. 1995. PCG *Reference Manual.* Center for Numerical Analysis Rep. CNA-274. Austin, Univ. Texas at Austin.

KAHAN W. 1967. Laguerre's method and a circle which contains at least one zero of a polynomial. *SIAM-NA* **4**, s. 474–482.

KAHANER D. 1970. Matrix description of the fast Fourier transform. *IEEE-TAE* **AU-18**, s. 422–450.

KAHANER D. 1978. The fast Fourier transform by polynomial evaluation. *ZAMP* **29**, s. 387–394.

KAHANER D., MOLER C., NASH S. 1989. *Numerical Methods and Software.* Englewood Cliffs, Prentice-Hall.

KANTOROVIČ L. V., AKILOV G. P. *1977. *Funkcionalnyj analiz*, izd. 2, Moskva, Nauka.

KARLOY F. P. 1993. Genetic algorithms for the traveling salesman problem. W: *Biological Cybernetics.* Berlin, Springer, s. 539–546.

KARMARKAR N. 1984. A new polynomial-time algorithm for linear programming. *Combinatorica* **4**, s. 373–395.

KELLER H. B. 1968. *Numerical Methods for Two-Point Boundary-Value Problems.* Waltham, Blaisdel.

KHOVANSKII A. N. 1963. *The Application of Continued Fractions and Their Generalizations to Problems in Approximation Theory.* Groningen, Wolters-Noordhoff.

KIEŁBASIŃSKI A., SCHWETLICK H. *1992. *Numeryczna algebra liniowa. Wprowadzenie do obliczeń zautomatyzowanych*, wyd. 2. WNT, Warszawa.

KINCAID D. R., OPPE T. C. 1988. A parallel algorithm for the general *LU* factorization. *CANM* **4**, s. 349–359.

KINCAID D. R., OPPE T. C., YOUNG D. M. 1989. ITPACKV 2D *user's guide.* Center for Numerical Analysis Rep. CNA-232. Austin, Univ. Texas at Austin.

KINCAID D. R., YOUNG D. M. 1979. Survey of iterative methods. W: *Encyclopedia of Computer Science and Technology*, Belzer J., Holzman A. G., Kent A. (eds.). New York, Dekker, s. 354–391.

KLINE M. 1972. *Mathematical Thought from Ancient to Modern Times.* New York, Oxford Univ. Press.

KNUTH D. E. 1979. Mathematical typography. *AMS-B* **2**, s. 337–372.

KRYLOV V. I. 1962. *Approximate Calculation of Integrals.* New York, MacMillan. (2., uzup. wyd. oryginału: *Priближennoe vyčislenie integralov*, Moskva, Nauka, 1967).

KRYLOV V. I., ŠULGINA L. T. *1966. *Spravočnaja kniga po čislennomu integrirovaniju.* Moskva, Nauka.

KUANG Y. 1993. *Delay Differential Equations.* New York, Academic Press.

KUDREWICZ J. *1993. *Fraktale i chaos*, wyd. 2. Warszawa, WNT.

KULISCH U., MIRANKER W. 1981. *Computer Arithmetic in Theory and Practice.* New York, Academic Press.

LAMBERT J. 1973. *Computational Methods in Ordinary Differential Equations.* New York, Wiley.

LANCASTER P., SALKAUSKAS K. 1986. *Curve and Surface Fitting.* New York, Academic Press.

LANCZOS C. 1966. *Discourse on Fourier Series.* Edinburgh, Oliver and Boyd.

LAWRENCE D. 1991. *Handbook of Genetic Algorithms.* New York, Reinhold.

LE D. 1985. An efficient derivative-free method for solving nonlinear equations. *ACM-TOMS* **11**, s. 250–262.

LEJA F. *1956. *Rachunek różniczkowy i całkowy ze wstępem do równań różniczkowych*, wyd. IV popr. Warszawa, PWN.

LORENTZ G. G., JETTER K., RIEMENSCHNEIDER S. D. 1983. *Birkhoff Interpolation.* Reading, Addison-Wesley.

LORENTZEN L., WAADELAND H. *1992. *Continued Fractions with Applications.* Amsterdam, North Holland.

LUENBERGER D. G. 1973. *Introduction to Linear and Nonlinear Programming.* Reading, Addison-Wesley.

MANDELBROT B. 1982. *The Fractal Geometry of Nature.* New York, Freeman.

MARDEN M. 1966. *Geometry of Polynomials.* Providence, AMS.

MARSDEN M. J. 1970. An identity for spline functions with applications to variation-diminishing spline approximation. *JAT* **3**, s. 7–49.

MASON J. C., HANDSCOMB D. C. *2002. *Chebyshev Polynomials.* Chapman & Hall/CRC Press.

MEINARDUS G. *1968. *Aproksymacja funkcji i jej metody numeryczne*, Warszawa, PWN.

METROPOLIS N., ROSENBLUTH A., ROSENBLUTH M., TELLER A., TELLER E. 1953. Equations of state calculation by fast computing machines. *J. Chem. Physics* **21**, s. 1087–1092.

MICCHELLI C. A. 1986a. Algebraic aspects of interpolation. W: *Approximation Theory*, C. de Boor (ed.), Proc. Symp. in Applied Mathematics **36**. Providence, AMS, s. 81–102.

MICCHELLI C. A. 1986b. Interpolation of scattered data. Distance matrices and conditionally positive definite functions. *CA* **2**, s. 11–22.

MITCHELL A. R., WAIT R. 1977. *The Finite Element Method in Partial Differential Equations.* New York, Wiley.

MOLER C. B., VAN LOAN C. F. 1978. Nineteen dubious ways to compute the exponential of a matrix. *SIAM-REV* **20**, s. 801–836.

MOORE R. 1966. *Interval Analysis.* Englewood Cliffs, Prentice-Hall.

MOORE R. E. 1979. *Methods and Applications of Interval Analysis.* Philadelphia, SIAM.

NAG 1995. *NAG Fortran Library Manual.* Downers Grove, NAG.

NATANSON I. P. *1949. *Konstruktivnaja teorija funkcij.* Moskva, GITTL.

NAZARETH L., TSENG P. 1998. *Gilding the lily: A variant of the Nelder-Mead algorithm.* Preprint.

NELDER J. A., MEAD R. 1965. A simplex method for function minimization. *CJ* **7**, s. 308–313.

NERINCKX D., HAEGEMANS A. 1976. A comparison of nonlinear equations solvers. *JCAM* **2**, s. 145–148.

NEWMAN D. J., RIVLIN T. J. 1983. Optimal universally stable interpolation. *Analysis* **3**, s. 355–367.

NOBLE B., DANIEL J. W. 1988. *Applied Linear Algebra*, 3rd ed. Englewood Cliffs, Prentice-Hall.

NOCEDAL J., WRIGHT S. 1999. *Numerical Optimization.* New York, Springer.

NUSSBAUMER H. J. 1982. *Fast Fourier Transform and Convolution Algorithms.* New York, Springer.

ODEN J. T. 1972. *Finite Elements of Nonlinear Continua.* New York, McGraw-Hill.

ODEN J. T., REDDY J. N. 1976. *An Introduction to the Mathematical Theory of Finite Elements.* New York, Wiley.

OPPE T. C., JOUBERT W. D., KINCAID D. R. 1988. NSPCG *user's guide*, version 1.0, package for solving large sparse linear systems by various iterative methods. Center for Numerical Analysis Rep. CNA-216. Austin, Univ. Texas at Austin.

OPPE T. C., KINCAID D. R. 1988. Parallel *LU*-factorization algorithms for dense matrices. W: *Supercomputing*, Houstis E. N., Papatheodorou T. S., Polychronopoulos C. D. (eds.), *LNCS* **297**. New York, Springer, s. 576–594.

ORTEGA J. M. 1988. *Introduction to Parallel and Vector Solution of Linear Systems.* New York, Plenum.

ORTEGA J. M., RHEINBOLDT W. C. 1970. *Iterative solution of Nonlinear Equations in Several Variables.* New York, Academic Press.

OSTROWSKI A. M. 1966. *Solution of Equations and Systems of Equations*, 2nd ed. New York, Academic Press.

OTTEN R. H. J. M., VAN GINNEKEN L. P. P. 1989. *The Annealing Algorithm.* Dordrecht, Kluwer.

OVERTON M. 2001. *Numerical Computing with IEEE Floating Point Arithmetic.* Philadelphia, SIAM.

PARLETT B. N. 1964. Laguerre's method applied to the matrix eigenvalue problem. *MOC* **18**, s. 464–485.

PASZKOWSKI S. *1975. *Zastosowania numeryczne wielomianów i szeregów Czebyszewa.* Warszawa, PWN.

PCGPAK2 1990. PCGPAK2 *user's guide.* New Haven, Scientific Computing Associates.

PEITGEN H.-O., JÜRGENS H., SAUPE D. *1995-6. *Granice chaosu. Fraktale.* Warszawa, PWN (cz. 1: 1995, cz. 2: 1996).

PEITGEN H., RICHTER P. 1986. *The Beauty of Fractals.* New York, Springer.

PEITGEN H.-O., SAUPE D., HAESELER F. V. 1984. Cayley's problem and Julia sets. *MI* **6**, s. 11–20.

PENROSE R. 1955. A generalized inverse for matrices. *Proc. Cambridge Phil. Soc.* **51**, s. 406–413.

PERRON O. 1929. *Die Lehre von Kettenbrüchen.* Leipzig, Teubner. (Dritte, verbesserte und erweiterte Aufl.: 1957, Stuttgart, Teubner; przedruk: New York, Chelsea).

PETERS G., WILKINSON J. H. 1971. Practical problems arising in the solution of polynomial equations. *IMA-JNA* **8**, s. 16–35.

PICKOWER C. A. 1988. A note on chaos and Halley's method. *ACM-COM* (11) **31**, s. 11.

POWELL M. J. D. 1964. An efficient method for finding stationary values of a function of several variables. *CJ* **7**, s. 155–162.

PRINCE P. J., DORMAND J. R. 1981. High order embedded Runge-Kutta formulae. *JCAM* **1**, s. 67–75.

PRUESS S. 1976. Properties of splines in tension. *JAT* **17**, s. 86–96.

PRUESS S. 1978. An algorithm for computing smoothing splines in tension. *Computing* **19**, s. 365–373.

RAIMI R. A. 1969. On the distribution of first significant figures. *AMM* **76**, s. 342–347.

RALL L. B. 1965. *Error in Digital Computation.* New York, Wiley.

RALSTON A. *1971. *Wstęp do analizy numerycznej.* Warszawa, PWN.

RALSTON A., RABINOWITZ P. 1978. *A First Course in Numerical Analysis.* New York, McGraw-Hill.

REDISH K. A. 1974. On Laguerre's method. *Int. J. Math. Educ. Sci. Technol.* **5**, s. 91–102.

RHEINBOLDT W. C. 1974. *Methods for Solving Systems of Nonlinear Equations.* CBMS Series in Applied Math. 14. Philadelphia, SIAM.

RICE J. R. 1966. Experiments on Gram-Schmidt orthogonalization. *MOC* **20**, s. 325–328.

RICE J. R. 1992. *Numerical Methods, Software, and Analysis*, 2nd ed. New York, Academic Press.

RIVLIN T. J. 1990. *The Chebyshev Polynomials*, 2nd ed. New York, Wiley.

ROYDEN H. L. 1968. *Real Analysis*, 2nd ed. New York, Macmillan.

RUDIN W. *1998. *Analiza rzeczywista i zespolona*, wyd. 2. Warszawa, PWN.

RUDIN W. *2002. *Analiza funkcjonalna*, wyd. 2. Warszawa, PWN.

SANDER L. M. 1987. Fractal growth. *SA* **256**, Jan., s. 94–100.

SARD A. 1963. *Linear Approximation.* Math. Surveys, No. 9, Providence, AMS.

SCALES L. E. 1985. *Introduction to Non-Linear Optimization.* New York, Macmillan.

SCHEID F. 1988. *Numerical Analysis.* New York, McGraw-Hill.

SCHNABEL R. B., FRANK P. D. 1984. Tensor methods for nonlinear equations. *SIAM-NA* **21**, s. 815–843.

SCHOENBERG I. J. 1967. On spline functions. W: *Inequalities*, O. Shisha (ed.). New York, Academic Press, s. 255–291.

SCHUMAKER L. L. 1981. *Spline Functions.* New York, Wiley.

SCHWEIKERT D. G. 1966. An interpolation curve using splines in tension, *JMP* **45**, s. 312–317.

SCOTT N. R. 1985. *Computer Number Systems and Arithmetic.* Englewood Cliffs, Prentice-Hall.

SHAMPINE L. F., ALLEN R. C. 1973. *Numerical Computing: An Introduction.* Philadelphia, Saunders.

SHAMPINE L. F., GEAR C. W. 1979. A user's view of solving stiff ordinary differential equations. *SIAM-REV* **21**, s. 1–17.

SHAMPINE L. F., GORDON M. K. 1975. *Computer Solution of Ordinary Differential Equations: The Initial Value Problem.* San Francisco, Freeman.

SHAMPINE L. F., WATTS H. A., DAVENPORT S. M. 1976. Solving nonstiff ordinary differential equations – The state of the art. *SIAM-REV* **18**, s. 376–411.

SHEPARD D. 1968. A two-dimensional interpolation function for irregularly spaced data. *Proc. 23rd Nat. Conf. ACM*, s. 517–524.

SIDI A. *2003. *Practical Extrapolation Methods. Theory and Applications.* Cambridge, Cambridge Univ. Press.

SMALE S. 1981. The fundamental theorem of algebra and complexity theory. *AMS-B* **4**, s. 1–36.

SMITH K. T. 1971. *Primer of Modern Analysis.* New York, Springer.

STEFFENSEN J. F. 1950. *Interpolation*, 2nd ed. New York, Chelsea.

STERNBENZ P. H. 1974. *Floating-Point Computations.* Englewood Cliffs, Prentice-Hall.

STOER J., BULIRSCH R. 1980. *Introduction to Numerical Analysis.* New York, Springer. (Polski przekład wcześniejszego wyd. niemieckiego: *Wstęp do metod numerycznych*, Warszawa, PWN, 1979 [t. I] i 1980 [t. II]).

STRANG G., FIX G. 1973. *An Analysis of the Finite Element Method.* Englewood Cliffs, Prentice-Hall.

STROUD A. H., SECREST D. 1966. *Gaussian Quadrature Formulas.* Englewood Cliffs, Prentice-Hall.

THOMAS B. 1986. The Runge-Kutta methods. *Byte*, April, s. 191–210.

TORCZON V. 1997. On the convergence of pattern search methods. *SIAM-JO* **7**, s. 1–25.

TRAUB J. F. 1964. *Iterative Methods for the Solution of Equations.* Englewood Cliffs, Prentice-Hall.

TREFETHEN L. N., SCHREIBER R. S. 1990. Average-case stability of Gaussian elimination. *SIAM-MAA* **11**, s. 335–360.

TSENG P. 1998. *Fortified-descent simplicial search method: a general approach.* Preprint.

TUROWICZ A. *1967. *Geometria zer wielomianów.* Warszawa, PWN.

VAN DER CORPUT J. G. 1946. Sur l'approximation de Laguerre des racines d'une équation qui a toutes ses racines réelles. *Nederl. Acad. Wetensch. Proc.* **49**, s. 922–929.

VARGA R. S. 1962. *Matrix Iterative Analysis.* Englewood Cliffs, Prentice-Hall. Wyd. 2. rozszerzone: New York, Springer 2000.

VERNER J. H. 1978. Explicit Runge-Kutta methods with estimates of the local truncation error. *SIAM-NA* **15**, s. 772–790.

VICHNEVETSKY R. 1981. *Computer Methods for Partial Differential Equations.* Vol. 1: *Elliptic Equations and the Finite Element Methods.* Vol. 2: *Initial Value Problems.* Englewood Cliffs, Prentice-Hall.

WACHSPRESS E. L. 1966. *Iterative Solution of Elliptic Systems.* Englewood Cliffs, Prentice-Hall.

WAIT R., MITCHELL A. R. 1986. *Finite Element Analysis and Applications.* New York, Wiley.

WALKER J. S. 1992. *Fast Fourier Transforms.* Boca Raton, CRC Press.

WALSH G. R. 1975. *Methods of Optimization.* New York, Wiley.

WASER S., FLYNN M. J. 1982. *Introduction to Arithmetic for Digital Systems Designers.* New York, Holt, Reinhart & Winston.

WENIGER E. J. *1989. Nonlinear sequence transformations for the acceleration of convergence and the summation of divergent series. *Computer Physics Reports* **10**, s. 189–371.

WERNER W. 1984. Polynomial interpolation: Lagrange versus Newton. *MOC* **43**, s. 205–217.

WILKINSON J. H. 1961. Error analysis of direct methods of matrix inversion. *ACM-J* **8**, s. 281–330.

WILKINSON J. H. 1963. *Rounding errors in Algebraic Processes.* Englewood Cliffs, Prentice-Hall. Polski przekład: *Błędy zaokrągleń w procesach algebraicznych.* PWN, Warszawa 1967.

WILKINSON J. H. 1965. *The Algebraic Eigenvalue Problem.* New York, Oxford Univ. Press.

WILKINSON J. H. 1984. The perfidious polynomial. W: *Studies in Numerical Analysis*, G. H. Golub (ed.). Washington, MAA, s. 1–28.

WILLÉ D. R., BAKER C. T. H. 1992. DELSOL: A numerical code for the solution of systems of delay-differential equations. *ANM* **9**, s. 223–234.

WIMP J. 1984. *Computation with Recurrence Relations.* Boston, Pitman.

WRIGHT S. J. 1993. A collection of problems for which Gaussian elimination with partial pivoting is unstable. *SIAM-SSC* **14**, s. 231–238.

YOUNG D. M. 1971. *Iterative Solution of Large Linear Systems.* New York, Academic Press.

YOUNG D. M., GREGORY R. T. 1972. *A Survey of Numerical Mathematics* (2 vols.). Reading, Addison-Wesley (przedruk: 1988, New York, Dover).

ZIENKIEWICZ O. C., MORGAN K. 1983. *Finite Elements and Approximation.* New York, Wiley.

Skorowidz