Piotr Rybka

MATEMATYKA
dla studentów chemii

skrypt do wykładu Matematyka B

Warszawa 2002

Nakładem Wydziału Chemii Uniwersytetu Warszawskiego, Warszawa 2002
Druk:

Wstęp

Niniejszy skrypt powstał na podstawie notatek do wykładu jaki prowadziłem w latach 1999 – 2002 na wydziale Chemii UW. Był to tzw. wykład „B", oznaczało, to że jego materiał jest rozszerzony, po to aby przygotować studentów do wysłuchania wykładu z chemii kwantowej, termodynamiki, czy innych wykładów z fizyki wymagających solidnego matematycznego przygotowania. W istocie rzeczy na ten wykład uczęszczają nie tylko studenci wydziału chemii, ale także słuchacze Międzywydziałowych Indywidualnych Studów Matematyczno–Przyrodniczych. Dzięki nim wykład powinien się nazywać **Matematyka dla przyrodoznawców**. Jest to też dydaktyczne wyzwanie rzucone wykładowcy.

Od słuchacza – czytelnika oczekuje się zainteresowania przedmiotem i dobrego przygotowania na poziomie szkoły średniej. Tyko tyle jest niezbędne do zrozumienia rozdziału pierwszego, który jest poświęcony przedstawieniu języka teorii zbiorów, aksjomatyki liczb rzeczywistych, liczb naturalnych i zasady indukcji zupełnej a na koniec elementów kombinatoryki.

W rozdziale drugim przedstawiamy teorię przestrzeni i odwzorowań liniowych, której koronnym zastosowaniem jest badanie rozwiązań układów równań liniowych. Przy okazji poznajemy wyznaczniki, których geometryczne interpretacje przedstawiamy na końcu rozdziału. Będą one wielce przydatne później w rozdziałach szóstym i siódmym poświęconym całkowaniu.

Rozdział trzeci jest poświęcony zarysowi rachunku różniczkowego i całkowego funkcji jednej zmiennej. Jego kulminacją jest teoria całki Riemanna i definicja funkcji wykładniczej i funkcji trygomometrycznych za pomocą szeregów potęgowych.

Rozdział czwarty zaczyna się od nader zwięzłego wprowadzenia w topologię przestrzeni Euklidesowych. Jest to przygotowanie dla pozostałych rozdziałów. Następnie wykładamy podstawowe pojęcia rachunku różniczkowego funkcji wielu zmiennych. Pomijamy pochodne wyższych rzędów, a pochodną rzędu drugiego wprowadzamy tylko dla funkcji o wartościach rzeczywistych. Czynimy tak wiedzeni brakiem czasu i koniecznością.

Równaniom różniczkowym zwyczajnym jest poświęcony rozdział piąty. Przedstawiamy najważniejsze typy równań skupiając się na równaniach liniowych pierwszego i drugiego rzędu. Przedstawiamy też zarys teorii egzystencjalnej, jest to zastosowanie ogólnych pojęć metrycznych zaprezentowanych na początku poprzedniego rozdziału. Nie rozwijamy tematu „metody rozwiązywania równań", mniemając że współczesny przyrodoznawca częściej posługuje się komputerem, np. do rozwiązania równania, aniżeli sięga po tablice całek.

Rozdział szósty jest poświęcony całce wielokrotnej definiowanej jako iterowana całka Riemanna. To podejście jest poglądowe, wymaga też poświęcenia nieco uwagi zbiorom mierzalnym w sensie Jordana–Riemanna. Taki wykład całkowania jest dużo prostszy, niż ten posługu-

jący się teorią Lebesgue'a. Kończymy rozdział wzorem na zamianę zmiennej w całce. Jego wyprowadzenie ilustruje metodę patrzenia na całki jak na granice sum Riemannowskich.

W następnym rozdziale zajmujemy się całkowaniem na krzywych i na powierzchniach. Oczywiście wyjaśniamy czym dla nas są krzywe i powierzchnie. Stosujemy tu też metodę opowiadania o całkach przedstawioną uprzednio. Przedstawiając temat „praca sił pola wzdłuż krzywej" dochodzimy do nowych pojęć, takich jak 1-forma różniczkowa. Wyprowadzamy wzory Greena, Gaussa-Ostrogradskiego jako kolejne uogólnienia podstawowego rachunku różniczkowego. Kończymy na wzorze Stokesa.

Ostatni, ósmy rozdział jest poświęcony elementom teorii przestrzeni Hilberta. Jej kluczowym w zastosowaniach do mechaniki kwantowej przykładem jest przestrzeń $L^2(\mathbb{R}^d)$. Definiujemy ją jako zbiór funkcji, których kwadrat jest „niewłaściwie" całkowalny w niewłaściwym sensie Riemanna. Nadto przedstawiamy inne tematy jak szeregi Fouriera, czy metodę najmniejszych kwadratów.

Przygotowując wykład korzystałem z wielu źródeł, przy okazji warto wymienić pozycje będące literaturą uzupełniającą:

rozdział 1. – H.Rasiowa, *Wstęp do matematyki współczesnej*, PWN, Warszawa 1990;

rozdział 2. – J.Komorowski, *Od liczb zespolonych do tensorów, spinorów, algebr Liego i kwadryk*, PWN, Warszawa 1978;

rozdziały 3. i 4. – W. Rudin, *Podstawy analizy matematycznej*, PWN, Warszawa 1976;

rozdział 5. – W.I.Arnold, *Równania różniczkowe zwyczajne*, PWN, Warszawa 1975;

rozdziały 4., 6. i 7. – G.M.Fichtenholz, *Rachunek różniczkowy i całkowy* t. I-III, PWN, Warszawa 1978; J.Thorpe, *Elementary topics in Differential Geometry*, Springer, Nowy Jork, 1979;

rozdział 8. wspomniane wyżej książki Rudina i Komorowskiego nadto, J.Stoer, R.Bulirsch, *Wstęp do analizy numerycznej*, PWN, Warszawa 1987.

Trzeba też powiedzieć, że niniejsze opracowanie zawiera nieco więcej materiału, niż w istocie zostało wyłożone. Nie zawsze opierałem się pokusie dopisania pominiętych szczegółów czy dodatkowych wyjaśnień. Mimo tego, objętość Pracy zasadniczo nie wzrosła.

Moja przygoda z wykładem Matematyka B zaczęła się za sprawą mojego nauczyciela prof. Marka Burnata, który niejednokrotnie namawiał mnie do wzięcia tego wykładu i użyczył swoich notatek. Wywarły one ogromny wpływ na rozdziały poswięcone całkowaniu. Jestem wdzięczny memu serdecznemu koledze i współpracownikowi dr. Marcinowi Moszyńskiemu, który prowadził ćwiczenia do wykładu wnikliwie czytał maszynopis notatek. Mojej żonie Magdzie dziękuję za wyrozumiałość, wsparcie i cierpliwe wyłapywanie literówek w tekście. Wreszcie moje podziękowania są skierowane do dr. Leszka Stolarczyka i prof. Krystyny Jackowskiej za zachętę do napisania skryptu i wsparcie.

Na koniec objaśnimy oznaczenia stosowane później w tekście:

$p \Rightarrow q$ nazywa się implikacją i czytamy p pociąga q albo jeśli p, to q. Często strzałki $\Rightarrow$, $\Leftarrow$ oznaczają kierunek wykazywania twierdzenia.
$p \Leftrightarrow q$ oznacza wtedy i tylko wtedy, tj. zachodzą obie implikacje: $p \Rightarrow q$ i $q \Rightarrow p$.
a.a. (ad absurdum) oznacza zastosowanie w dowodzie metody sprowadzenia do niedorzeczności.

□ oznacza koniec dowodu.
Symbol $a := b$ należy rozumieć, że a jest równe b na mocy definicji.

W obrębie rozdziałów przyjmujemy ciągłą numerację przykładów, definicji i łącznie twierdzeń, stwierdzeń i lematów. Odwołanie w obrębie rozdziału odbywa się poprzez podanie numeru twierdzenia (przykładu) itp. Odwołanie do stwierdzenia (lematu itp.) z innego rozdziału ma postać podobną tyle, że numer stwierdzenia jest poprzedzony numerem rozdziału.

Pojęcia definiowane są pisane *kursywą*.

Spis treści

Rozdział 1

Wiadomości wstępne

1.1 Nasze cele

Naszym celem jest zapoznanie czytelnika z podstawami niezbędnymi do wysłuchania wykładów z mechaniki kwantowej i termodynamiki. Można go przedstawić obrazowo, porównując matematykę do samochodu:

1. Do jego prowadzenia potrzebne jest prawo jazdy, jego posiadacz ma wiedzieć do czego służy kierownica i jakie są podstawowe funkcje różnych dźwigni i przycisków, ma znać przepisy ruchu. W matematyce odpowiada to znajomości tabliczki mnożenia, której wyrafinowaną formą jest umiejętność prawidłowego wypełnienia PITu.

2. Doświadczony kierowca dodatkowo zna się na budowie samochodu, potrafi rozpoznać problem i powiedzieć mechanikowi w warsztacie co ma zrobić. Rozumie swój wóz i potrafi przeprowadzić drobne naprawy.

3. Mechanik samochodowy zna działanie mechanizmów i wie co jak jest zbudowane. Wiedza z następnego etapu jest przydatna, lecz przede wszystkim ma doświadczenie, którego nie można zdobyć szkoleniem teoretycznym.

4. Konstruktor zna ze szczegółami tajniki budowy i potrafi skonstruować nowy pojazd. Studenci matematyki są kształceni do takiego poziomu, aby móc być zatrudnionym w charakterze stażysty u boku konstruktora.

Nasz cel to osiągnięcie poziomu drugiego, co oznacza zrozumienie podstaw mechaniki kwantowej, tak by móc pogadać z fachowcem, a dzięki praktyce osiągnąć poziom czeladnika i majstra.

Aby przybliżyć nasz cel, będę opowiadał o matematyce, więcej i szerzej niż na wykładzie A, ale egzamin nie będzie trudniejszy. Położymy nacisk na wyjaśnienie związków pomiędzy faktami aniżeli na budowanie formalnej teorii. Nie będziemy zbyt głęboko wchodzić w szczegóły, dowody będą szkicowane lub pomijane. Jednak w całości przeprowadzane tylko te najbardziej typowe. Czytelnik ma nabyć wprawę w operowaniu pojęciami w praktyce, tj. w liczeniu. Jednocześnie czytelnik powiniem oswoić się z metodą dedukcyjną. Dobrym punktem wyjścia jest abstrakcyjna, acz łatwa teoria mnogości.

1.2 Zbiory i działania na nich

Zwykle pierwszy uniwersytecki wykład matematyki zaczyna się od stwierdzenia, że od słuchaczy nie jest wymagana żadna wiedza. Jest to oczywiście przesada, bo bagaż doświadczeń jest bardzo pomocny. Ale kryje się w tym stwierdzeniu ziarno prawdy: mianowicie musimy zacząć od uzgodnienia języka. Językiem matematyki jest aksjomatyczna teoria mnogości ze swym zespołem pewników i pojęć pierwotnych. Pojęcia pierwotne uznaje się za znane. Pewniki to twierdzenia uznane za prawdziwe bez dowodu. Nie jest naszym celem poprawne konstruowanie teorii mnogości, bo jej staranny wykład zwieńczony ścisłą definicją liczb rzeczywistych trwałby pół roku, co nie wchodzi w grę. Tym nie mniej musimy jej nieco liznąć. Zakładamy, że wszyscy wiedzą co to są zbiory (jest to właśnie owo niedefiniowane pojęcie pierwotne). Zbiór A można zdefiniować wyliczając jego elementy, np. $A = \{a, b, c\}$, jednak na ogół jest to niewykonalne jak w przypadku zbioru liczb rzeczywistych $\mathbb{R}$, zespolonych $\mathbb{C}$ — będą to główne obiekty naszego zainteresowania. Będziemy pisać $x \in A$ (x jest elementem A lub x należy do A). Uniwersalnym przykładem jest zbiór pusty $\emptyset$, który nie zawiera żadnego elementu (tj. następujące zdanie jest prawdziwe: dla każdego x, nieprawdą jest, że x należy do $\emptyset$), inny przykład $\emptyset \in \{\emptyset\}$.

1.2.1 Działania na zbiorach

Mając dwa zbiory A i B możemy utworzyć ich sumę $A \cup B$.

Definicja 1. Powiemy, że x należy do *sumy zbiorów* A i B, oznaczanej $A \cup B$, wtedy i tylko wtedy, gdy $x \in A$ lub $x \in B$. Piszemy też $x \in A \cup B \Leftrightarrow x \in A$ lub $x \in B$.

Możemy utworzyć przecięcie zbiorów A i B.

Definicja 2. *Iloczynem (przecięciem) zbiorów* A i B jest zbiór $A \cap B$ określony następująco: $x \in A \cap B \Leftrightarrow x$ należy do A i x należy do B.

Zilustrujmy te definicje prostymi przykładami: niech A oznacza zbiór liczb parzystych, zaś B zbiór liczb nieparzystych. Wtedy $A \cap B = \emptyset$ i $A \cup B$ jest zbiorem liczb całkowitych.

Inną ważną operacją jest różnica zbiorów, $A \setminus B$, określona poniżej.

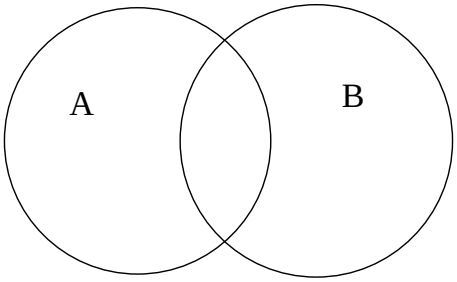
Definicja 3. *Różnicą zbiorów* A i B nazywamy zbiór tych elementów A, które nie należą do B, oznaczmy go symbolem $A \setminus B$.

Przykład 1. Jeśli A jest zbiorem liczb rzeczywistych, zaś B jest zbiorem dodatnich liczb rzeczywistych, to $A \setminus B$ jest zbiorem nieujemnych liczb rzeczywistych.

Zauważmy też, że zawsze jest prawdą, iż

$$(A \setminus B) \cap B = \emptyset.$$

Warto w tym momencie zwrócić uwagę na możliwość zobrazowania powyższych operacji za pomocą tzw. okręgów Eulera.



Rys. 1. okręgi Eulera.

Ważnymi pojęciami są: podzbiór i zawieranie się zbiorów.

Definicja 4. Powiemy, że A jest *podzbiorem* B (piszemy $A \subset B$) lub A zawiera się w B, wtedy i tylko wtedy, gdy $x \in A$, pociąga $x \in B$. Oczywiście, dla każdego zbioru A mamy, że

$$A \subset A; \qquad \emptyset \subset A.$$

Odnotujmy teraz prosty fakt.

Stwierdzenie 1. Jeśli $A \subset B$ i $B \subset A$, to $A = B$.

Dowód. Z założenia, jeśli $x \in A$, to $x \in B$, dodatkowo, jeśli $x \in B$, to $x \in A$, tj. $x \in A$ wtedy i tylko wtedy, gdy $x \in B$, czyli $A = B$. □

Potrzebne nam będą reguły tworzenia podzbiorów. A mianowicie podzbiory możemy określać następująco

$$B = \{x \in A : x \text{ ma właściwość } \Phi\}.$$

Trzeba tu wyraźnie powiedzieć, że B jest podzbiorem A tylko dla „rozsądnych" wyrażeń Φ i jego istnienie jest pewnikiem. Gdy A jest zbiorem liczb rzeczywistych, to zdanie „x jest liczbą dodatnią" jest rozsądne, a więc

$$R^+ = \{x \in A : \ x \text{ jest liczbą dodatnią}\}$$

jest zbiorem.

Inną ważną a prostą operacją jest *iloczyn kartezjański* (lub po prostu produkt) $A \times B$ zbiorów A i B. Do jego definicji jest potrzebne pojęcie *pary uporządkowanej* (a, b). Jest ono intuicyjnie jasne: pierwszym elementem pary jest a, drugim b. Ściśle rzecz ujmując

$$(a, b) := \{\{a\}, \{a, b\}\}.$$

Można wykazać że $(a, b) = (c, d)$ wtedy i tylko wtedy gdy $a = c$ i $b = d$. Prosty dowód zostawiamy Czytelnikowi jako ćwiczenie własne.

Teraz definiujemy $A \times B$ następująco:

Definicja 5. $A \times B$ jest zbiorem par uporządkowanych (a, b) takich, że $a \in A$ i $b \in B$.

Jeśli $A = B$ to zamiast $A \times A$ piszemy A^2. Możemy kartezjańsko mnożyć większą ilość zbiorów pisząc

$$A_1 \times A_2 \times A_3 \times .... \times A_n.$$

Powyższy zbiór jest złożony z n–tek uporządkowanych, których definicja jest dość oczywista. Odnotujemy tylko, że $(a, b, c) := ((a, b), c)$ itp.

Przykład 2. Niech A będzie zbiorem liczb rzeczywistych, wtedy
$A^2 = A \times A$ jest płaszczyzną Euklidesową, jej elementy (a, b), to punkty płaszczyzny;
$A^3 = A \times A \times A$ jest przestrzenią Euklidesową, jej elementy (a, b, c), to punkty.

1.3 Relacje

Pojęcie relacji jest wstępem do ścisłego ujęcia funkcji, która jest intuicyjnie dobrze znana: oznacza ona, że każdemu elementowi x zbioru X umiemy jednoznacznie przypisać element $f(x)$ należący do Y, (piszemy czasem $x \mapsto f(x)$). W niedalekiej przyszłości będziemy przeprowadzać operacje na funkcjach, tworzyć zbiory funkcji, dlatego chcielibyśmy więc mieć jasność co do natury tego obiektu. Nie jest to ćwiczenie czysto akademickie, bo mechanikę kwantową uprawia się w zbiorach, których elementami są właśnie funkcje. Zaczniemy od definicji relacji R.

Definicja 6. Niech będą dane dwa zbiory A i B. *Relacją* R nazywamy dowolny podzbiór $A \times B$. Jeśli $A = B$, to mówimy, że mamy *relację* R *w* A. Piszemy xRy na oznaczenie faktu, że x jest w relacji z y.

Przykład 3. Niech A będzie zbiorem liczb rzeczywistych, kładziemy wtedy $R = \{(x, y) \in A^2 : x \text{ jest mniejsze od } y\}$, wtedy R jest relacją nazywaną relacją mniejszości. Zamiast xRy piszemy zgodnie z tradycją $x < y$.

W dalszym ciągu będziemy opisywać relacje bardziej szczegółowo i rozróżniać je.

Definicja 7. Powiemy, że relacja R w A jest:
(a) *zwrotna*, jeśli $x \in A$ pociąga xRx;
(b) *symetryczna*, jeśli xRy pociąga yRx;
(c) *przechodnia*, jeśli xRy i yRz pociąga xRz.
(d) *relacją równoważności*, jeśli spełnia (a), (b) i (c).

Przykład 4. Niech A będzie zbiorem liczb naturalnych.
(a) Relację R w A definiujemy następująco: xRy wtedy i tylko wtedy, gdy x dzieli y. Wtedy R jest zwrotna, przechodnia, ale nie symetryczna.
(b) Niech p będzie liczbą naturalną większą od 1. Relację R_p w A definiujemy następująco: xR_py wtedy i tylko wtedy, gdy $x - y$ jest podzielne przez p. Łatwo sprawdzić, że R_p jest relacją równoważności. Pozostawiamy to Czytelnikowi do samodzielnego sprawdzenia.

Relacje równoważności mają ciekawą właściwość: Oznaczmy przez $[x]$ zbiór tych y z X, które pozostają w relacji z x, xRy, tj.

$$[x] = \{y \in X : xRy\}.$$

Zbiór $[x]$ nazywamy *klasą abstrakcji elementu* x. Zbiór klas abstrakcji oznaczamy następująco:

$$X/R.$$

Stwierdzenie 2. Niech R będzie relacją równoważności, wtedy dla dowolnych x, y należących do X mamy: $[x] = [y]$ albo $[x] \cap [y] = \emptyset$.

Dowód. Mamy dwie możliwości, albo przecięcie $[x] \cap [y]$ jest puste i wtedy nie mamy nic do roboty, albo $[x] \cap [y] \neq \emptyset$. Załóżmy więc, że z jest elementem $[x] \cap [y]$, wtedy zRx a także zRy. Z przechodniości relacji równoważności R wynika, że xRy, tj. x jest elementem $[y]$. Wynika, stąd, że $[x]$ zawiera się w $[y]$. Podobnie argumentujemy, że $[y]$ zawiera się w $[x]$. Zatem $[y] = [x]$. Co kończy dowód.

Wynika stąd prosty wniosek, że relacja równoważności w X wprowadza rozbicie X na rodzinę rozłącznych zbiorów, które w sumie dadzą X. Mamy mianowicie

$$X = \bigcup_{x \in X} [x] \quad \text{(suma zbiorów } [x] \text{ indeksowanych } x \text{ należącymi do } X).$$

Z drugiej strony przypuśćmy, że mamy rozbicie zbioru X :

$$X = \bigcup_{i \in I} A_i,$$

gdzie sumowanie przebiega po zbiorze wskaźników I. Wtedy takie rozbicie definiuje nam relację ρ w X, a mianowicie $x \rho y$ wtedy i tylko wtedy, gdy istnieje wskaźnik i, taki że x i y należą do A_i. Łatwo sprawdzić, że jest to relacja równoważności, co pozostawiamy Czytelnikowi jako ćwiczenie własne.

Przykład 5. Jeśli $\mathbb{N}$ jest zbiorem liczb naturalnych, zaś R_p była zdefiniowana powyżej, to $\mathbb{N}/R_p$ oznacza się przez $\mathbb{Z}_p$. Uwaga: można w nim w naturalny sposób wprowadzić działania arytmetyczne.

1.4 Funkcje

Intuicyjnie pojęcie funkcji jest jasne: jest to przyporządkowanie elementowi x zbioru X elementu $f(x)$ zbioru Y, często jest to wzór, np. $f(x) = x^2 + x + 1$. Jednak takie intuicyjne rozumienie funkcji jest niewystarczające do naszych celów. Z drugiej strony jakiekolwiek ścisłe ujęcie musi zgadzać się z intuicją.

Zacznijmy od tego, że para uporządkowana $(x, f(x))$ jest elementem iloczynu kartezjańskiego $X \times Y$. Podążymy tym tropem i będziemy traktować funkcję jako podzbiór $X \times Y$, tj relację. Zaczniemy od definicji:

Definicja 8. Powiemy, że relacja $f \subset X \times Y$ jest *prawostronnie jednoznaczna*, jeśli warunek xfy_1 i xfy_2, pociąga $y_1 = y_2$.

Jesteśmy gotowi do określenia funkcji. Niech będą dane zbiory X i Y (np. $X = Y$ jest zbiorem liczb rzeczywistych, będzie to nasz najważniejszy przykład).

Definicja 9. *Funkcją* f o *dziedzinie* X i *przeciwdziedzinie* Y, co oznaczamy symbolem $f : X \to Y$, nazywamy dowolną relację prawostronnie jednoznaczną $f \subset X \times Y$, taką, że dla każdego x istnieje $y \in Y$ pozostający w relacji f z x, tj. xfy, dla prostoty piszemy wtedy $y = f(x)$.

Uwaga. Właściwie, to na dobrą sprawę utożsamiamy funkcję z jej wykresem.

Zajmiemy się teraz bardziej szczegółowym opisem funkcji i ich właściwościami. Niech będzie dana funkcja $f : X \to Y$ i podzbiór A dziedziny X. *Obrazem* zbioru A, piszemy $f(A)$, jest zbiór

$$\{y \in Y : \text{ istnieje } x \in A, \text{ że } y = f(x)\}.$$

Przeciwobrazem zbioru $B \subset Y$ jest zbiór $f^{-1}(B)$ określony następująco:

$$f^{-1}(B) = \{x \in X : f(x) \in B\}.$$

Wykażemy teraz proste, acz ważne właściwości przeciwobrazu:

$$f^{-1}(B_1) \cap f^{-1}(B_2) = f^{-1}(B_1 \cap B_2) \qquad f^{-1}(B_1) \cup f^{-1}(B_2) = f^{-1}(B_1 \cup B_2)$$

Sprawdzimy tylko pierwszą równość, dowód drugiej jest zbliżony. Niech x należy do lewej strony równości. Jest to równoważne stwierdzeniu, że x należy do $f^{-1}(B_1)$ i $f^{-1}(B_2)$. Jest to z kolei równoważne, temu że $f(x) \in B_1$ i $f(x)$ jest w B_2, tzn. x należy do prawej strony.

Obraz zachowuje się podobnie, mianowicie mamy

$$f(A_1) \cup f(A_2) = f(A_1 \cup A_2), \qquad f(A_1 \cap A_2) \subset f(A_1) \cap f(A_2)$$

Udowodnimy tylko drugą inkluzję. Jeśli y należy do lewej strony to znaczy, że istnieje $x \in X$ należące do $A_1 \cap A_2$, takie że $f(x) = y$. To znaczy, że x jest w A_1 i jest w A_2 i $f(x) = y$. Zatem, y należy do $f(A_1)$ i $y \in f(A_2)$.

Chcemy podkreślić, że nie można zastąpić inkluzji równością. Pokazuje to prosty przykład, niech $f : \{-1, 1\} \to \{1\}$ będzie dana wzorem $f(x) = x^2$ i $A_1 = \{-1\}$, $A_2 = \{1\}$. Mamy wtedy, że $A_1 \cap A_2 = \emptyset$ a zatem $\emptyset = f(\emptyset) \subset f(A_1) \cap f(A_2) = \{1\}$, równości zbiorów, oczywiście nie ma.

Musimy rozróżniać funkcje, z tego powodu wprowadzamy więcej definicji.

Definicja 10. Powiemy, że funkcja $f : X \to Y$ jest *różnowartościowa* (jest *iniekcją*), jeśli z warunku $f(x_1) = f(x_2)$ wynika, że $x_1 = x_2$.

Definicja 11. Powiemy, że funkcja $f : X \to Y$ jest *„na”* (jest *suriekcją*), jeśli dla każdego $y \in Y$ istnieje $x \in X$ taki, że $f(x) = y$.

Definicja 12. Powiemy, że $f : X \to Y$, jest *wzajemnie jednoznaczna* (jest *bijekcją*), jeśli f jest różnowartościowa i „na”.

Istnienie bijekcji pomiędzy zbiorami A i B oznacza, że mają one równą ilość elementów. Czasem prowadzi to do wniosków sprzecznych ze zdrowym rozsądkiem. Za chwilę przedstawimy tego przykłady, ale najpierw zajmiemy się sprawami podstawowymi.

Podamy kilka definicji funkcji ilustrujących powyższe pojęcia.

Przykład 6. Niech $\mathbb{N}$ będzie zbiorem liczb naturalnych, zaś $\mathbb{Z}$ niech będzie zbiorem liczb całkowitych.

Funkcję $f : \mathbb{Z} \to \mathbb{N}$ określamy wzorem:

$$f(x) = x^2.$$

f nie jest ani „na", ani różnowartościowa.

Funkcja $g : \mathbb{N} \to \mathbb{N}$ jest dana wzorem, $g(x) = x^3$ jest ona różnowartościowa i „na", tj. jest wzajemnie jednoznaczna;

Funkcję $h : \mathbb{N} \to \mathbb{N}$ określamy wzorem $h(x) = x(x-1)(x-2)$, jest ona „na", ale nie różnowartościowa;

Funkcję $k : \mathbb{N} \to \mathbb{N}$ określamy wzorem $k(x) = x/(1+|x|)$. Jest różnowartościowa, ale nie „na".

Podamy teraz metodę tworzenia nowych funkcji z danych.

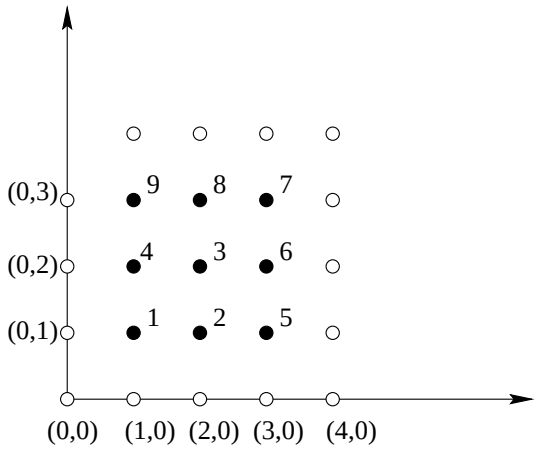
Definicja 13. Niech będą dane $f : X \to Y$, $g : Y \to Z$, funkcję $h : X \to Z$ daną wzorem

$$h(x) = g(f(x))$$

nazywamy *złożeniem funkcji* f i g, piszemy $h = g \circ f$.

Będziemy mieli bardzo często do czynienia ze złożeniem funkcji. Odnotujmy tutaj jedną jego właściwość. Niech $f : X \to Y$, $g : Y \to Z$, $h : Z \to W$, wtedy $(f \circ g) \circ h = f \circ (g \circ h)$.

Podamy teraz przykład, który może się wydawać zaskakujący. Określimy teraz funkcję $\phi : \mathbb{N} \to (\mathbb{N} \setminus \{0\}) \times (\mathbb{N} \setminus \{0\})$, która jest „na". Rysunek zwięźle podaje pomysł.



Rys. 2. Odwzorowanie $\mathbb{N}$ na $\mathbb{N} \times \mathbb{N}$.

Podamy też wzór. Zaczniemy od tego, że dowolna liczba $k \in \mathbb{N}$ jest postaci $n^2 < k \leq (n+1)^2$ dla pewnego $n \geq 0$. Wtedy kładziemy $\phi(k) = (m, p)$, gdzie

$$(m,p) = \begin{cases} (n+1, k-n^2), & \text{gdy } k - n^2 \leq n+1; \\ ((n+1)^2 - k, n+1), & \text{gdy } k - n^2 > n+1. \end{cases}$$

Sprawdzenie różnowartościowości jest łatwe i zostawimy to czytelnikowi. Zajmiemy się pokazaniem, że ϕ jest „na". Niech dane będzie (m, p), trzeba znaleźć k, takie że $\phi(k) = (m, p)$. Załóżmy, że $m \geq p$ wtedy istnieje n takie, że $m = n + 1$, co więcej możemy teraz położyć

$$k = n^2 + p.$$

Mamy, że $\phi(k) = (m, p)$.

Przypadek $m \geq p$ rozpatruje się podobnie i pozostawiamy go Czytelnikowi do zbadania. Pokazaliśmy więc, że ϕ jest wzajemnie jednoznaczna. □

Z punktu widzenia teorii mnogości można powiedzieć, że zbiory $\mathbb{N}$ i $(\mathbb{N} \setminus \{0\}) \times (\mathbb{N} \setminus \{0\})$ mają tyle samo elementów. Aby uściślić to pojęcie wprowadzimy nowe określenie.

Definicja 14. Powiemy, że zbiory A i B są *równoliczne*, jeśli istnieje funkcja. $f : A \to B$, która jest bijekcją.

Powyższy przykład pozwala nam sformułować ciekawy wniosek.

Wniosek 3. Liczb naturalnych jest tyle samo co par liczb naturalnych, tj. liczb wymiernych, $\frac{p}{q}$ jest tyle samo co liczb naturalnych!

Pokażemy teraz, że jednak istnieją teraz zbiory nierównoliczne. Wyjaśnimy to w następnym paragrafie.

1.4.1 Zbiór Potęgowy

Zdefiniujemy ważny zbiór, którego samo istnienie jest pewnikiem. Rozpatrzmy dowolny zbiór X, tworzymy nowy zbiór

$$P(X) := \{y \text{ jest podzbiorem } X\}.$$

Nazywamy go *zbiorem potęgowym.* Zauważmy, że podzbiory X można utożsamiać z funkcjami $f : X \to \{0, 1\}$, dlatego czasem piszemy też 2^X zamiast $P(X)$. Wspomniane utożsamienie jest następujące, jeśli $A \subset X$, to definiujemy następującą funkcję

$$\chi_A(x) = \begin{cases} 1, & \text{gdy } x \in A; \\ 0, & \text{w przeciwnym przypadku.} \end{cases}$$

Funkcję χ_A nazywamy *funkcją charakterystyczną zbioru* A.

Z drugiej strony, jeśli jest dana funkcja $f : X \to \{0, 1\}$, to kładziemy $A = f^{-1}(1)$.

Jednak zasadniczym faktem, o którym chcielibyśmy tu opowiedzieć jest poniższe twierdzenie.

Twierdzenie 4. Jeśli $X \neq \emptyset$, to X i $P(X)$ nie są równoliczne.
Dowód. Pokażemy, że równoliczność X i $P(X)$ prowadzi do niedorzeczności. Załóżmy, że istnieje $f : X \to P(X)$, która jest bijekcją, rozpatrzmy

$$Y = \{x \in X : x \notin f(x)\}.$$

Skoro f jest „na" to istnieje y takie, że $f(y) = Y$. Jednak ani y nie może być elementem Y, ani $y \notin Y$. Uzyskana sprzeczność dowodzi nasze twierdzenie. □

1.5 Liczby rzeczywiste i naturalne

Załóżmy, że zostały nam objawione liczby rzeczywiste, od tej pory będziemy oznaczać ich zbiór symbolem $\mathbb{R}$. Owo objawienie będziemy opisywali właściwościami liczb rzeczywistych, czyli pewnikami. Zaczniemy od działań $+$, $\cdot$, tj. zakładamy, że dane są funkcje

$$+ : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \quad \text{oraz} \quad \cdot : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$$

i wyróżnione elementy 0 i 1 o następujących właściwościach

$$0 \neq 1.$$

Dla prostoty będziemy pisali $a+b$ zamiast $+(a, b)$ itp. Pewniki opisujące działania podzielimy na kilka grup, aby ułatwić ich przyswojenie. Zaczniemy od pewników dotyczących pojedynczych działań $+$ (*dodawania*) i $\cdot$ (*mnożenia*).

Przyjmujemy, że

(G1) dla dowolnych liczb rzeczywistych a, b, c, $(a + b) + c = a + (b + c)$ (tj. działanie $+$ jest *łączne*);

(G2) dla dowolnego $a \in \mathbb{R}$, $a + 0 = 0 + a = a$, (tj. 0 jest *elementem obojętnym* dodawania);

(G3) dla dowolnej liczby rzeczywistej a istnieje $b \in \mathbb{R}$ taka, że $a+b = b+a = 0$, (tj. istnieje *element przeciwny* do a);

(G4) dla dowolnych $a, b \in \mathbb{R}$, mamy $a + b = b + a$ (tj. działanie $+$ jest *przemienne*).

Powyższe pewniki wprowadzają nowy obiekt.

Definicja 15. Niech będzie dany zbiór G z działaniem $+$ i wyróżnionym elementem 0. Jeśli trójka $(G, +, 0)$ spełnia pewniki (G1 - G3), to nazwiemy ją *grupą.* Jeśli dodatkowo grupa $(G, +, 0)$ spełnia (G4) to nazywamy ją *grupą przemienną* albo *Abelową*.

W myśl powyższej definicji $(\mathbb{R}, +, 0)$ jest grupą przemienną. Element b przeciwny do a oznaczamy w prosty sposób $b = -a$. Będziemy też pisali $a - b$ zamiast $a + (-b)$, gdy a i b są dowolnymi liczbami rzeczywistymi.

Musimy wypowiedzieć się na temat mnożenia. To co mamy na myśli można ująć zwięźle pisząc, że trójka $\mathbb{R} - \{0\}, \cdot, 1)$ jest grupą abelową. Dla porządku przepiszemy te aksjomaty:

(G5) dla a, b, $c \in \mathbb{R}$, mamy $(a \cdot b) \cdot c = a \cdot (b \cdot c)$;

(G6) dla $a \in \mathbb{R}$, $a \cdot 1 = 1 \cdot a = a$;

(G7) dla $a \in \mathbb{R}, a \neq 0$ istnieje $b \in \mathbb{R}$ taki, że $a \cdot b = b \cdot a = 1$;

(G8) dla $a, b \in \mathbb{R}, a \cdot b = b \cdot a$.

Aby uniknąć nieporozumień element b określony w (G7) nazywamy elementem odwrotnym do a i piszemy $b = a^{-1}$.

Zauważmy, że jeszcze nie powiązaliśmy dodawania i mnożenia. Zrobimy to teraz.

(C1) Dla dowolnych liczb rzeczywistych a, b, c, jest prawdą, że $a \cdot (b + c) = a \cdot b + a \cdot c$ (tj. mnożenie jest rozdzielne względem dodawania).

Wymienione wyżej pewniki można zebrać pod wspólną nazwą, zrobimy to wprowadzając nowe określenie.

Definicja 16. Niech będzie dany zbiór K z dwoma działaniami i wyróżnionymi elementami $0, 1$. Jeśli piątka $(K, +, \cdot, 0, 1)$ spełnia (G1 - G8) i (C1), to nazywamy ją *ciałem przemiennym*. Można opuścić żądanie (G8), wtedy dostaniemy np. nieprzemienne ciało kwaterionów, nie będziemy się tym zajmować.

Objawione liczby $\mathbb{R}$ mają właściwości (G1 - G8) i (C1). Co ciekawe, liczby wymierne $\mathbb{Q}$ tj. liczby postaci $\frac{p}{q}$, gdzie $a \neq 0$ i p, q są naturalne (jeszcze nie zdefiniowane) spełniają (G1 - G8) i C1). Co więcej, zbiory klas abstrakcji $\mathbb{Z}_p$ z naturalnie wprowadzonymi działaniami arytmetycznymi też są ciałami przemiennymi, gdy p jest liczbą pierwszą. Nie są ciałami, gdy p jest liczbą złożoną. Sprawdzenie faktów dotyczących $\mathbb{Z}_p$ polecamy Czytelnikowi jako ćwiczenie własne.

Przedstawimy teraz szereg prostych faktów dotyczących liczb rzeczywistych. Naszym celem jest zapoznanie czytelnika z rozwojem formalnej teorii.

Stwierdzenie 5. Element przeciwny jest wyznaczony jednoznacznie.

Dowód. Niech b i b' będą elementami przeciwnymi do a, wtedy

$$b = b + 0 = b + (a + b') = (b + a) + b' = 0 + b' = b'. \qquad \square$$

Stwierdzenie 6. $\alpha \cdot 0 = 0$

Dowód. Wykażmy najpierw, że dla dowolnych liczb rzeczywistych α, β, γ

$$\alpha \cdot (\beta - \gamma) = \alpha \cdot \beta - \alpha \cdot \gamma$$

Przekształcamy lewą stronę

$$\text{Lewa } = \alpha(\beta - \gamma) + \alpha \cdot \gamma - \alpha \cdot \gamma = \alpha(\beta - \gamma + \gamma) - \alpha\gamma = \alpha \cdot (\beta + 0) - \alpha\gamma = \text{ Prawa}$$

Wykorzystamy tę tożsamość położywszy $\beta = \gamma = 1$. Dostaniemy wtedy

$$\alpha \cdot 0 = \alpha \cdot (1 - 1) = \alpha \cdot 1 - \alpha \cdot 1 = \alpha - \alpha = 0. \qquad \square$$

Stwierdzenie 7. $(-1) \cdot \alpha = -\alpha$.

Dowód. Na mocy stwierdzenia 5 wystarczy wykazać, że $\alpha + (-1) \cdot \alpha = 0$. Mamy bowiem,

$$\alpha + (-1) \cdot \alpha = 1 \cdot \alpha + (-1) \cdot \alpha = (1 - 1) \cdot \alpha = 0 \cdot \alpha = 0. \qquad \square$$

1.6 Liczby naturalne

Wprowadziliśmy zbiór liczb rzeczywistych, poznaliśmy już część jego struktury. Teraz zajmiemy się jej składnikiem jakim jest zbiór liczb naturalnych. Jego ścisła definicja jest celem obecnego paragrafu. Chcemy, aby była ona zgodna z intuicją tj. że 1 jest liczbą naturalną i że zbiór liczb naturalnych jest wyczerpywany przez operację dodawania jedynki tj. są to liczby postaci $1 + \ldots + 1$. Umówimy się przy tym, że zero też jest liczbą naturalną.

Definicja 17. Powiemy, że podzbiór $I \subset \mathbb{R}$ jest *induktywny* wtedy i tylko wtedy, gdy
1. $0 \in I$;
2. jeśli $n \in I$, to $n + 1 \in I$.

Zbiory induktywne, to kandydaci na zbiór liczb naturalnych. Istotnie, mamy bowiem:

Twierdzenie 8. Istnieje najmniejszy induktywny podzbiór $\mathbb{R}$. Oznaczamy go przez $\mathbb{N}$ i nazywamy go zbiorem *liczb naturalnych*. Oznacza to, że każdy zbiór induktywny zawiera $\mathbb{N}$.

Dowód. Zaczniemy od definicji rodziny podzbiorów induktywnych $\mathbb{R}$

$$\Lambda = \{T \in P(\mathbb{R}) : \ T \text{ jest induktywny}\},$$

tj. Λ jest *rodziną* podzbiorów $\mathbb{R}$. Kładziemy teraz

$$\mathbb{N} := \bigcap_{I \in \Lambda} I$$

tj. $\mathbb{N}$ jest przecięciem wszystkich zbiorów należących do rodziny Λ.

Sprawdzamy, że $\mathbb{N}$ jest zbiorem induktywnym. Widzimy najpierw, że $0 \in \mathbb{N}$. Jest to prawda, bo dla każdego I z rodzimy Λ, $0 \in I$, zatem 0 należy do części wspólnej wszystkich I. Załóżmy, że $n \in \mathbb{N}$, to wtedy dla wszystkich $I \in \Lambda$, $n \in I$ a z induktywności I wynika, że $n + 1 \in I$, dla wszystkich I tj. $n + 1 \in \mathbb{N}$. Zatem $\mathbb{N}$ jest induktywny.

Sprawdzimy teraz, że $\mathbb{N}$ jest najmniejszym zbiorem induktywnym. Niech teraz J będzie dowolnym podzbiorem induktywnym $\mathbb{R}$. Wtedy $J \in \Lambda$ i mamy

$$\mathbb{N} = \bigcap_{I \in \Lambda} I \subset J,$$

co kończy dowód. □

Uwagi. $0 \in \mathbb{N}$ tj. jest liczbą naturalną z mocy definicji; $1 \in \mathbb{N}$, $1 + 1$ (piszemy 2) jest liczbą naturalną, $2 + 1$ (piszemy 3) należą do $\mathbb{N}$ itp.

Twierdzenie 9. (zasada indukcji zupełnej). Niech T będzie właściwością liczb naturalnych (tj. $T(n)$ może być zdaniem prawdziwym lub fałszywym). Tworzymy zbiór

$$I = \{n \in \mathbb{N} : \ \text{zdanie } T(n) \text{ jest prawdziwe}\}$$

Jeśli jest prawdą, że (a) $0 \in I$ i (b) prawdziwa jest implikacja: jeśli $n \in I$, to $n+1 \in I$, to wtedy $I = \mathbb{N}$

Uwaga. Nazwa zasady indukcji zupełnej jest tradycyjna, w istocie jest to twierdzenie podlegające dowodowi.

Dowód. Zauważmy, że zdefiniowany powyżej zbiór I jest induktywny, zatem z poprzedniego stwierdzenia wynika, że $\mathbb{N} \subset I$, ale z definicji $I \subset \mathbb{N}$, zatem $I = \mathbb{N}$. □

1.6.1 Zastosowania Zasady Indukcji Zupełnej

Wprowadzimy parę oznaczeń i nowych pojęć. Przyjmujemy, że $0^0 := 1$. Jeśli x jest dowolną liczbą rzeczywistą, to piszemy $x^0 := 1$, a jeśli n jest liczbą naturalną, to kładziemy $x^{n+1} := x^n \cdot x$. Symbol x^n nazywamy *n-ta potęgą* x.

Definicja 18. (a) *Silnią liczby naturalnej* n nazywamy liczbę naturalną określoną następująco:

$$0! := 1, \quad (n+1)! := n! \cdot (n+1).$$

Innymi słowy: $k! = 1 \cdot 2 \cdot \ldots \cdot k$.

(b) Jeśli $n \geq k$ są liczbami naturalnymi, to *symbolem Newtona* nazywamy liczbę

$$\binom{n}{k} := \frac{n!}{(n-k)!k!}.$$

Zauważmy od razu, że

$$\binom{n}{0} = 1, \quad \binom{n}{n} = 1.$$

Znaczenie owych określeń jest wyjaśnione w poniższym twierdzeniu.

Twierdzenie 10. (wzór Newtona) Załóżmy, że a, b są liczbami rzeczywistymi a n jest liczbą naturalną, wtedy

$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}, \tag{1}$$

gdzie przy okazji wprowadziliśmy wygodne oznaczenie,

$$\sum_{i=0}^{n} a_i \quad \text{oznacza } a_0 + a_1 + ... + a_n.$$

Dowód. Będzie on zastosowaniem zasady indukcji zupełnej. Zakładamy najpierw, że $n = 0$. Wtedy lewa strona równania (1) przyjmuje postać $(a+b)^0 = 1$. Zaś prawa:

$$\sum_{i=0}^{0} \binom{0}{i} a^i b^{0-i} = \binom{0}{0} a^0 b^0 = 1,$$

a więc obie strony równają się.

Zakładamy teraz prawdziwość naszego twierdzenia dla pewnego n i i wykazujemy ją dla $n+1$. L i P będą odpowiednio oznaczały lewą i prawą stronę.

$$\begin{aligned} L &= (a+b)^{n+1} = (a+b)^n(a+b) = \left(\sum_{i=0}^{n}\binom{n}{i}a^i b^{n-i}\right)(a+b) \\ &= \sum_{i=0}^{n}\binom{n}{i}a^{i+1}b^{n-1} + \sum_{i=0}^{n}\binom{n}{i}a^i b^{n+1-i} = \sum_{i=0}^{n}\binom{n}{i}a^{i+1}b^{n+1-i-1} + \sum_{j=0}^{n}\binom{n}{j}a^j b^{n+1-1}. \end{aligned}$$

Teraz w pierwszej sumie zmieniamy wskaźnik sumowania, przyjmujemy, że $j = i+1$, co prowadzi do drobnego uproszczenia sumy i zmiany granic sumowania. Dostaniemy, że

$$\begin{aligned} L = &\sum_{j=1}^{n+1}\binom{n}{j-1}a^j b^{n+1-j} + \sum_{j=0}^{n}\binom{n}{j}a^j b^{n+1-1} \\ = &\binom{n}{n}a^{n+1}b^0 + \sum_{j=1}^{n}a^j b^{n+1-j}\left(\binom{n}{j-1} + \binom{n}{j}\right) + \binom{n}{0}a^0 b^{n+1} \\ = &\binom{n+1}{n+1}a^{n+1}b^0 + \sum_{j=1}^{n}a^j b^{n+1-j}\binom{n+1}{j} + \binom{n+1}{0}a^0 b^{n+1} = P, \end{aligned}$$

gdzie wykorzystaliśmy następujący fakt

$$\begin{aligned} \binom{n}{j-1} + \binom{n}{j} &= \frac{n!}{(j-1)!(n-j+1)!} + \frac{n!}{j!(n-j)!} \\ &= \frac{n!(j+n-j+1)}{j!(n-j+1)!} = \frac{n!(n+1)}{j!(n-j+1)!} = \binom{n+1}{j}. \end{aligned}$$

Przekonaliśmy się, że założenia zasady indukcji są spełnione, wynika stąd prawdziwość wzoru dla wszystkich n.

1.7 Ciągi, kombinatoryka

Zaczniemy od definicji, która jest trochę na wyrost, bo ciągi będziemy później starannie badali w rozdziale 3. Teraz potrzebna nam będzie tylko terminologia.

Przymujemy, że X jest dowolnym zbiorem.

Definicja 19. *Ciągiem elementów z X* nazwiemy dowolną funkcję $a : \mathbb{N} \to X$. Zgodnie ze zwyczajem piszemy a_n zamiast $a(n)$. *Ciąg liczbowy* dostaniemy, gdy $X = \mathbb{R}$. Będą nam jeszcze potrzebne pojęcie *ciągu n elementowe z X*. Jest nim dowolna funkcja

$$a : I \to X$$

gdzie I jest równoliczne z $\{0, 1, ..., n-1\}$.

Wprowadzimy teraz podstawowe definicje kombinatoryczne.

Definicja 20. Niech X będzie dowolnym zbiorem skończonym. *Permutacją bez powtórzeń* elementów X nazwiemy dowolną funkcję $f : X \to X$, która jest wzajemnie jednoznaczna. Permutacje bez powtórzeń nazywamy też *przestawieniami.* Ilość przestawień zbioru n – elementowego oznaczamy symbolem P_n.

Stwierdzenie 11. $P_n = n!$.

Dowód. Nasze zadanie polega na policzeniu na ile sposobów możemy ustawić w ciąg elementy zbioru X. Na pierwszym miejscu możemy postawić jeden wybrany spośród n elementów. Na drugim miejscu możemy postawić jeden wybrany spośród już tylko $n-1$ elementów itp. Na ostatnim n-tym miejscu możemy wybrać element już tylko ze zbioru jednoelementowego. Dostaniemy zatem, że $P_n = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1 = n!$ □

Pora na kolejną definicję.

Definicja 21. Niech X będzie dowolnym zbiorem n – elementowym. *Kombinacją k – elementową* elementów zbioru X nazwiemy dowolny k– elementowy podzbiór X. Ilość kombinacji oznaczamy symbolem c_n^k.

Stwierdzenie 12. $c_n^k = \binom{n}{k}$.

Dowód. Ponownie będziemy ustawiali elementy X w ciągi, tym razem k–elementowe. Na pierwszym miejscu ciągu może być jeden z n elementów, na drugim miejscu już tylko jeden z $n-1$. Kontynuujemy ten proces dochodząc do k-tego miejsca. Na nim mamy wybór jednego z $(n-k+1)$ elementów. Tym samym dostaniemy, że k–elementowych ciągów ze zbioru n–elementowego jest

$$n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)(n-k)!}{(n-k)!} = \frac{n!}{(n-k)!}.$$

Ale kolejność ustawienia nie jest istotna, bo interesują nas podzbiory X zatem tę liczbę dzielimy przez $P_k = k!$, czyli ilość przestawień zbioru k–elementowego. W ostatecznym rachunku,

$$c_n^k = \frac{n!}{(n-k)!k!} = \binom{n}{k}.$$

□

Wykorzystamy ten fakt do policzenia ilości wszystkich podzbiorów skończonego zbioru. Mianowicie mamy.

Stwierdzenie 13. Ilość wszystkich podzbiorów zbioru n-elementowego równa się 2^n.

Dowód. Zauważmy, że szukana ilość podzbiorów to,

$$c_n^0 + c_n^1 + c_n^2 + \ldots + c_n^n = \sum_{k=0}^{n} c_n^k = \sum_{k=0}^{n} \binom{n}{k} 1^k \cdot 1^{n-k}$$

Na mocy wzoru Newtona (stwierdzenie 10.) powyższa suma równa się

$$(1+1)^n = 2^n.$$

□

1.8 Kresy zbiorów liczb rzeczywistych

Przyjęliśmy, że liczby rzeczywiste zostały nam objawione. Ich dotychczas wypowiedziane właściwości można ująć stwierdzeniem, że liczby rzeczywiste spełniają aksjomaty ciała przemiennego. Widzieliśmy też, że jest więcej przykładów ciał przemiennych. Poszukajmy więc dodatkowych struktur wyróżniających $\mathbb{R}$. Zauważmy, że w zbiorze liczb rzeczywistych dana jest *relacja niewiększości* $\leq$. Spełnia ona następujące warunki:
(P1) dla każdego $x \in \mathbb{R}, x \leq x$ (tj. relacja niewiększości jest zwrotna);
(P2) jeśli $x, y \in \mathbb{R}$ i $x \leq y$ oraz $y \leq x$, to $x = y$ (tj. relacja niewiększości jest antysymetryczna);
(P3) jeśli $x, y, z \in \mathbb{R}$ i $x \leq y$ oraz $y \leq z$, to $x \leq z$ (tj. relacja niewiększości jest przechodnia);
(P4) jeśli $x, y \in \mathbb{R}$, to $x \leq y$ lub $y \leq x$ (tj. relacja niewiększości jest *spójna*).

Warto w tym momencie zwrócić uwagę, że dowolna relacja $\leq$ w zbiorze X spełniająca (P1-P3) nazywa się *porządkiem częściowym*. Jeśli dodatkowo porządek częściowy w zbiorze X jest spójny (tj. (P4) jest spełnione), to nazywa się go *porządkiem liniowym.*

Przykład 7. Relacja zawierania się zbiorów jest porządkiem częściowym, ale nie liniowym, bo nie jest prawdą, że można porównać 2 dowolne zbiory.

Istotny jest związek porządku z działaniami arytmetycznymi. Określimy go za pomocą poniższych pewników
(P5) jeśli $z \leq y$ i $x \in \mathbb{R}$, to $x + z \leq x + y$;
(P6) jeśli $0 \leq x$ i $y \leq z$, to $xy \leq xz$.

Od razu powiedzmy, że ciało liczbowe K, spełniające dodatkowo (P1-P6) nazywa się *ciałem uporządkowanym*. Przykładem służy, oprócz $\mathbb{R}$, ciało liczb wymiernych $\mathbb{Q}$. Przykładami ciał, które nie są ciałami uporządkowanymi są $\mathbb{Z}_2, \mathbb{Z}_3, \mathbb{Z}_5$ itd. Można się o tym samemu przekonać badając tabelkę działań.

Kwestia, na czym polega różnica pomiędzy nimi pozostaje nierozstrzygnięta (aż do końca podrozdziału).

Aby uprościć nasze wypowiedzi, wprowadzimy definicje nowych relacji.

Definicja 22.
(a) $x \geq y$ wtedy i tylko wtedy, gdy $y \leq x$;
(b) $x < y$ wtedy i tylko wtedy, gdy $x \leq y$ i $x \neq y$;
(c) $x > y$ wtedy i tylko wtedy, gdy $x \geq y$ i $x \neq y$.

Chcemy teraz przedstawić wybrane właściwości liczb rzeczywistych związane z porządkiem i pokazać ich dowód z pewników. Zaczniemy od następującego faktu.

Stwierdzenie 14. Jeśli $x < 0$, to $-x > 0$ (a nadto, jeśli $x > 0$, to $-x < 0$).

Dowód. Z mocy (P5) możemy dodać $-x$ do obu stron nierówności $x \leq 0$. Dostaniemy wtedy $0 = -x + x \leq -x$ i skoro $x \leq 0$, to $0 < -x$. □

Mamy dwa cele na uwadze. Pierwszym jest sprawdzenie, że $x^2 \geq 0$ dla wszystkich liczb rzeczywistych. Drugim jest wykazanie odpowiednika (P6), gdy $x < 0$. Po drodze wykażemy parę niezbędnych faktów.

Stwierdzenie 15. Dla dowolnej liczby rzeczywistej x mamy $-(-x) = x$.

Dowód. $-(-x)$ jest elementem przeciwnym do x, ale $x + (-x) = 0$, bo $-x$ jest elementem przeciwnym do x. Zatem $x = -(-x)$, co wynika z jednoznaczności elementu przeciwnego (stwierdzenie 5). □

Stwierdzenie 16. $(-1)^2 = 1$

Dowód. $(-1)^2 - 1 = (-1)(-1) + (-1) \cdot 1 = (-1)(-1 + 1) = (-1) - 0 = 0$. □

Zdobyliśmy już dość wiedzy, aby osiągnąć zamierzone cele.

Stwierdzenie 17. Jeśli $x \in \mathbb{R}$, to $x^2 \geq 0$.

Dowód. Mamy dwa przypadki: (1) $x \geq 0$ i (2) $x < 0$. Jeśli $x \geq 0$, to z (P6) natychmiast dostaniemy, że $x \cdot x \geq x \cdot 0$, tj. $x^2 \geq 0$.

Jeśli zaś $x < 0$, to ze stwierdzenia 14. wynika, że $-x > 0$. Zatem z (P6) $(-x) \cdot (-x) \geq 0$. Z kolei lewa strona tej nierówności to $(-x) \cdot (-x) = (-1)x(-1)x = (-1)^2 \cdot x^2 = 1 \cdot x^2 = x^2$ □.

Obecnie nasz drugi cel jest na wyciągnięcie ręki:

Stwierdzenie 18. Jeśli $x < 0$ i $a \leq b$, to $ax \geq bx$.

Jego dowód pozostawiamy Czytelnikowi do samodzielnego przeprowadzenia.

Przypomnimy teraz znane definicje przedziałów:

Definicja 23.
Zbiór $[a, b] = \{x \in \mathbb{R} : x \leq b \text{ i } x \geq a\}$ nazywamy *przedziałem domkniętym*;
zbiór $(a, b) = \{x \in \mathbb{R} : x < b \text{ i } x > a\}$ nazywamy *przedziałem otwartym*;
zbiór $[a, b) = \{x \in \mathbb{R} : x < b \text{ i } x > a\}$ nazywamy *przedziałem lewostronnie domkniętym i prawostronnie otwartym*;
zbiór $(a, b] = \{x \in \mathbb{R} : x \leq b \text{ i } x > a\}$ nazywamy *przedziałem lewostronnie otwartym i prawostronnie domkniętym*;
kładziemy $(-\infty, a] = \{x \in \mathbb{R} : x \leq a\}$, podobnie definiujemy $(a, +\infty)$, $[a, +\infty)$, $(-\infty, a)$;
$\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$.

Trzeba przy tym podkreślić, że symbole $-\infty$, ∞ nie oznaczają żadnej liczby.

Naszym celem jest teraz zdefiniowanie pierwiastków liczb rzeczywistych. Po drodze okaże się, że samo ich istnienie wymaga dodatkowego pewnika. Stanie się też jasna różnica pomiędzy

$\mathbb{R}$ i $\mathbb{Q}$. Zaczniemy od określenia *wartości bezwzględnej* $|x|$ liczby rzeczywistej x i jej *znaku* $\operatorname{sgn} x$.

Definicja 24.

$$|x| = \begin{cases} x & \text{jeśli } x \geq 0 \\ -x & \text{jeśli } x < 0 \end{cases} \qquad \operatorname{sgn} x = \begin{cases} 1 & \text{jeśli } x > 0 \\ 0 & \text{jeśli } x = 0 \\ -1 & \text{jeśli } x < 0. \end{cases}$$

Najważniejszą właściwością wartości bezwzględnej jest *nierówność trójkąta*

$$|x + y| \leq |x| + |y| \tag{3}$$

Dowód jest łatwym zastosowaniem definicji $|x|$ i zostawiamy go Czytelnikowi. Prostym wnioskiem jest następna nierówność

$$||x| - |y|| \leq |x - y|. \tag{4}$$

Mianowicie, z nierówności trójkąta mamy,

$$|x| = |x - y + y| \leq |x - y| + |y|,$$

skąd wypływa (4).

Przystępujemy teraz do definiowania pierwiastów arytmetycznych.

Definicja 25. Niech $n \in \mathbb{N}$ i $n > 1$. Jeśli x będzie dodatnią liczbą rzeczywistą, to wtedy dodatnią liczbę rzeczywistą y nazywa się *pierwiastkiem arytmetycznym z* x stopnia n i piszemy, że $y = x^{1/n}$ lub $y = \sqrt[n]{x}$, jeśli $y^n = x$. Jeśli $x < 0$, to liczbę $y \in \mathbb{R}$ nazywa się *pierwiastkiem arytmetycznym z* x stopnia n, jeśli $y^n = x$.

Uwaga. Jeśli liczba jest n parzysta i $x < 0$, to **nie** istnieje y takie, że $y^n = x$, bo $x^2 > 0$ (patrz stwierdzenie 17.)

Zauważmy też, że pierwiastek arytmetyczny jeśli istnieje, to jest wyznaczany jednoznacznie. Jednak w tej chwili samo jego istnienie jest źródłem kłopotów.

Stwierdzenie 19. $\sqrt{2}$ nie jest liczbą wymierną.

Dowód. Załóżmy, że tak nie jest, tj. istnieje liczba wymierna $\frac{p}{q}$, taka, że $\frac{p^2}{q^2} = 2$, tj. $2q^2 = p^2$. Co prowadzi do sprzeczności z jednoznacznością rozkładu na czynniki pierwsze. Jest to fakt dość oczywisty, którego nie będziemy dowodzić. □

Widać wyraźnie, że istnienie pierwiastków jest kwestią delikatną. Do wykazania ich istnienia potrzebny jest dodatkowy *postulat zupełności.* Jego sformułowanie wymaga nowych definicji.

Definicja 26. (a) Niech $E \subset \mathbb{R}$. Powiemy, że E jest *ograniczony z góry*, jeśli istnieje $M \in \mathbb{R}$, taka że $M \geq x$ dla każdego $x \in E$.
(b) Niech $E \subset \mathbb{R}$. Powiemy, że E jest *ograniczony z dołu*, jeśli istnieje $m \in \mathbb{R}$, taka że $x \geq m$ dla każdego $x \in E$.

Zdefiniowawszy zbiory ograniczone możemy zająć się kresami.

Definicja 27. (a) Niech $E \subset \mathbb{R}$ będzie ograniczony z góry . Powiemy, że $b \in \mathbb{R}$ jest kresem górnym E i piszemy $b = \sup E$, jeśli dla każdego ograniczenia górnego M zbioru E jest prawdą że $M \geq b$.
(b) Niech $E \subset \mathbb{R}$ będzie ograniczony z dołu. Powiemy, że $b \in \mathbb{R}$ jest kresem dolnym i piszemy $a = \inf E$, jeśli dla każdego ograniczenia dolnego m zbioru E jest prawdą że $m \leq b$.

Można teraz wypowiedzieć ostatni z aksjomatów liczb rzeczywistych - *postulat zupełności*.

(Z) Każdy niepusty zbiór ograniczony z góry (odpowiednio: z dołu) ma kres górny b (odpowiednio: dolny a).

Uwaga. Kres górny zbioru E nie musi być jego elementem największym, np.

$$E = \{x \in \mathbb{R} : x^2 < 4\}$$

wtedy $\sup E = 2$, ale $2 \notin E$.

Istnienie kresów zapewni istnienie pierwiastków arytmetycznych. Mamy bowiem

Twierdzenie 20. Jeśli $0 < x \in \mathbb{R}$ i $n \in \mathbb{N}$ i $n > 1$, to istnieje $\sqrt[n]{x}$.

Dowód. Rozpatrzmy tylko przypadek $x > 1$, gdyż $x = 1$ nie wymaga pracy, zaś $x < 1$ sprowadza się do pierwszego poprzez podstawienie $y = x^{-1}$.

Określamy zbiór E w następujący sposób:

$$E = \{y \in \mathbb{R} : \ 0 < y, \ y^n < x\}.$$

Wtedy $\emptyset \neq E$, bo $1 \in E$. Jest to prawdą, bo $1^n = 1 < x$. Co więcej zbiór E jest ograniczony. Wystarczy sprawdzić, że dla każdego $y \in E$, $y < x$. Wymaga to sprawdzenia, że $x^n > x$. Zrobimy to z pomocą indukcji. Dla $n = 2$, mamy $x^2 > x$, bo jest to wynik mnożenia $x > 1$ obustronnie przez x. Dalej, z nierówności $x^n > x$ wynika, $x^{n+1} > x$. A mianowicie, $x^{n+1} = x^n \cdot x > x \cdot x > x$. Zatem na mocy zasady indukcji zupełnej nasza nierówność jest prawdziwa, dla $n > 1$. Tym samym $x \notin E$, a więc jedyna możliwość jaką mamy, to $y \leq x$ dla dowolnego $y \in E$. Zatem E jest niepusty i ograniczony i możemy zastosować (Z):

$$c := \sup E$$

Niech $\varepsilon > 0$ będzie dowolną liczbą rzeczywistą mniejszą od 1. Kładziemy $\delta := \frac{\varepsilon}{(1+c)^n}$, oczywiście $\delta < 1$, bo $1 + c > 1$. Skoro $\delta > 0$, to $c + \delta > c$, zatem $c + \delta$ jest ograniczeniem górnym E i

$$(c + \delta)^n \geq x \tag{5}$$

$c - \delta$ jest mniejsza od kresu górnego, a skoro tak, to

$$(c - \delta)^n < x \tag{6}$$

Oszacujemy $(c+\delta)^n$ z dwumianu Newtona

$$(c+\delta)^n = \sum_{i=0}^{n} \binom{n}{i} \delta^i c^{n-i} = c^n + \sum_{i=1}^{n} \binom{n}{i} \delta^i c^{n-i} = c^n + \delta \sum_{i=1}^{n} \binom{n}{i} \delta^{i-1} c^{n-1}.$$

Skoro $\delta < 1$, to dostaniemy $\delta^{i-1} \leq 1$, dla $i \geq 1$ i stąd

$$(c+\delta)^n \leq c^n + \delta \sum_{i=1}^{n} \binom{n}{i} 1^i c^{n-i} \leq c^n + \delta \sum_{i=0}^{n} \binom{n}{i} 1^i c^{n-i} = c^n + \delta(1+c)^n$$

i (5) dzięki definicji δ przyjmuje postać

$$x \leq c^n + \delta(1+c)^n = c^n + \varepsilon.$$

Podobnie postępujemy z $(c-\delta)^n$:

$$(c-\delta)^n = \sum_{i=0}^{n} \binom{n}{i} c^{n-i}(-\delta)^i = c^n - \delta \sum_{i=1}^{n} \binom{n}{i} c^{n-i}(-\delta)^{i-1}$$

teraz wykorzystamy fakt, iż $(-\delta)^{i-1} \leq 1$, dla $i \geq 1$. Zatem

$$(c-\delta)^n \geq c^n - \delta \sum_{i=1}^{n} \binom{n}{i} c^{n-i} = c^n - \delta(1+c)^n$$

tym samym (6) przyjmuje postać

$$x > c^n - \varepsilon.$$

Potrzebny jest nam teraz.

Lemat 21. Jeśli dla dowolnej rzeczywistej liczby dodatniej ε zachodzi

$$a - \varepsilon \leq b$$

to

$$a \leq b.$$

Dowód lematu. Załóżmy, że tak nie jest tj. $a > b$. Kładziemy $\varepsilon = \frac{a-b}{2}$. Wtedy dostaniemy

$$b \geq a - \varepsilon = a - \frac{a-b}{2} = \frac{a+b}{2}$$

tj. $b \geq a$ dostaniemy sprzeczność z założenia $a > b$. □

Dzięki powyższemu lematowi dostaniemy, że

$$x \geq c^n \text{ i } x \leq c^n.$$

A teraz z aksjomatu (P2) dostajemy, że $x = c^n$. □

W szczególności wnioskujemy, że $\sqrt{2}$ jest dobrze określoną liczbą rzeczywistą. Możemy też od razu podać nowy przykład ciała uporządkowanego. Kładziemy

$$\mathbb{Q}(\sqrt{2}) := \{x \in \mathbb{R} : x = a + \sqrt{2}b, \quad \text{gdzie } a, b \text{ są wymierne } \}.$$

Na koniec odnotujmy Zasadę Archimedesa, której dowód pomijamy.

Twierdzenie 22. Dla każdej liczby rzeczywistej x istnieje liczba naturalna n, taka że $n > x$.

Wypływa z niej gęstość liczb wymiernych.

Wniosek 23. Jeśli $x, y \in \mathbb{R}$ i $x < y$, to istnieje liczba wymierna r, taka, że $x < r < y$.

1.9 Liczby zespolone

Nasza metoda przedstawienia liczb rzeczywistych polegała na przedstawianiu kolejnych grup pewników ciała, ciała uporządkowanego, aby zakończyć na postulacie zupełności. Okazywało się, że prowadziło to do przykładów ciał o coraz lepszych właściwościach. Ciało $\mathbb{R}$ ma bardzo dobre właściwości analityczne, ale pewne wady algebraiczne, bo nie każde równanie wielomianowe ma pierwiastki rzeczywiste, np. $x^2 + 1 = 0$.

Ciało liczb zespolonych nie ma tej wady. Zacznijmy od definicji, kładziemy $\mathbb{C} = \mathbb{R} \times \mathbb{R}$.

Definicja 28. *Ciałem liczb zespolonych* nazywamy $(\mathbb{C}, +, (0,0), \cdot, (1,0))$ z następującymi działaniami

$$(a, b) + (a', b') = (a + a', b + b'),$$
$$(a, b) \cdot (a', b') = (aa' - bb', ab' + a'b).$$

Trzeba sprawdzić, że są spełnione aksjomaty ciała przemiennego. Jest to łatwe, sprawdzimy tylko istnienie elementu odwrotnego dla dowolnego $z = (a, b) \neq 0$. Kładziemy,

$$z^{-1} = (a, b)^{-1} = (a/(a^2 + b^2), -b/(a^2 + b^2)),$$

pozostawiając czytelnikowi sprawdzenie, że $z \cdot z^{-1} = 1$.

Wprowadźmy bardziej znaną notację, będziemy pisać $i = (0, 1)$ i będziemy utożsamiać liczby rzeczywiste z liczbami zespolonymi postaci $(a, 0)$. Od tej pory będziemy pisać $a + bi$ zamiast (a, b). Zauważmy teraz, że

$$i^2 = (0, 1) \cdot (0, 1) = (-1, 0) = -1.$$

Wprowadzamy nowe operacje dla liczb zespolonych $z = a + bi$, mianowicie

$$\bar{z} := a - bi,$$

liczbę $\bar{z}$ nazywamy *liczbą sprzężoną* do z. Dalej,

$$\mathrm{Re}z := (z + \bar{z})/2, \quad \mathrm{Im}z := (z - \bar{z})/2i$$

$\mathrm{Re}z$ nazywamy *częścią rzeczywistą* i $\mathrm{Im}z$ nazywamy *częścią urojoną* z. Co więcej

$$|z| := \sqrt{z\bar{z}},$$

nazywamy *wartością bezwzględną* z. Trzeba tylko sprawdzić, że ostatnia definicja jest poprawna, tj. że argument pierwiastka jest dodatni:

$$z\bar{z} = (a+bi)(a-bi) = a^2 - (bi)^2 = a^2 + b^2 \geq 0.$$

Wymieńmy najprostsze właściwości nowych obiektów, mamy

$$|\mathrm{Re}z| \leq |z|, \quad |\mathrm{Im}z| \leq |z|.$$

Jest to łatwe, bo $|\mathrm{Re}z| = |a| \leq \sqrt{a^2+b^2} = |z|$. Inną prostą, ale ważną właściwością jest *nierówność trójkąta*:

$$|z_1 + z_2| \leq |z_1| + |z_2|,$$

której sprawdzenie polecamy Czytelnikowi jako rachunkowe ćwiczenie własne.

Łatwe do sprawdzenia jest też, że

$$|z_1 z_2| = |z_1||z_2|,$$

albowiem,

$$Lewa = (aa' - bb')^2 + (ab' + a'b)^2 = Prawa.$$

Wprowadzimy jeszcze jedną funkcję mając świadomość, że jest ona nieuprawniona na obecnym etapie. Definiujemy *argument liczby zespolonej* z

$$\mathbb{C} \setminus \{0\} \ni z \mapsto \mathrm{arg}z \in [0, 2\pi)$$

(i nie wiemy jeszcze co to jest π). Mianowicie piszemy, że

$$\varphi = \mathrm{arg}z$$

φ gdy jest jedynym rozwiązaniem układu równań

$$\cos\varphi = \frac{\mathrm{Re}z}{|z|} \quad \sin\varphi = \frac{\mathrm{Im}z}{|z|}.$$

(Podkreślamy, że funkcje $\cos$ i $\sin$ nie zostały jeszcze zdefiniowane). Zatem,

$$z = |z|(\cos\mathrm{arg}z + i\sin\mathrm{arg}z), \tag{7}$$

co więcej owo przedstawienie jest jednoznaczne. Zauważmy, że wzór (7) pozwala na ciekawe zapisanie mnożenia liczb zespolonych. Jeśli $z_1 = r(\cos\phi + i\sin\phi)$ i $z_2 = R(\cos\psi + i\sin\psi)$, to wtedy korzystając ze wzorów na sinus i cosinus sumy kątów (patrz wzór (3.37)) dostaniemy, że

$$z_1 \cdot z_2 = rR(\cos(\phi+\psi) + i\sin(\phi+\psi)). \tag{8}$$

Z definicji argumentu liczby zespolonej wynika natychmiast, że

$$\arg z^{-1} = \arg \bar{z} = 2\pi - \arg z.$$

Łatwym wnioskiem ze wzoru (8) jest *wzór de Moivre'a* dla liczby zespolonej $z = |z|(\cos\phi + i\sin\phi)$ mamy

$$z^n = |z|^n(\cos n\phi + i\sin n\phi).$$

Ściśle rzecz ujmując należy zastosować indukcję ze względu na n. Zauważmy, że ten wzór pozwala obliczyć pierwiastki n-tego stopnia z dowolnej liczby zespolonej, tj. znaleźć n rozwiązań równania

$$y^n - x = 0.$$

Z (7) wynika, że $|y| = |x|^{1/n}$. Zatem wzór de Moivre'a daje, że

$$|x|(\cos\psi + i\sin\psi) = |y|^n(\cos n\phi + i\sin n\phi),$$

tj.

$$\psi = n\phi + 2k\pi,$$

albo

$$\phi = \frac{\psi}{n} + \frac{2k\pi}{n}, \quad k = 0, 1, \ldots, n-1.$$

Wykazaliśmy więc, że istnieje n różnych pierwiastków z dowolnej liczby zespolonej. Zadajmy pytanie: czy ten fakt można powiązać z rozwiązywaniem równań wielomianowych? Odpowiedź jest podana poniżej. Zaczynamy od określenia:

Definicja 29. Funkcję $f : \mathbb{K} \to \mathbb{K}$ postaci

$$f(z) = \sum_{i=0}^{n} a_i z^i,$$

gdzie $a_i \in \mathbb{K}$ i $a_n \neq 0$ nazywamy *wielomianem o współczynnikach z ciała* $\mathbb{K}$ *, stopnia* n.

Twierdzenie 24. (zasadnicze twierdzenie algebry) Każdy wielomian o współczynnikach zespolnych, różny od stałej, ma pierwiastek zespolony.

Jest to trudny fakt, nie będziemy go dowodzić. Natychmiast wynika z niego

Wniosek 25. Każdy wielomian P stopnia n o współczynnikach zespolonych rozkłada się na iloczyn czynników linowych, tj.

$$P(z) = a_n \prod_{i=1}^{n} (z - z_i).$$

Na koniec paragrafu o liczbach zespolonych wykażemy pewną ważną nierówność, której znajomość jest konieczna. Będzie ona pojawiała się w wielu późniejszych rozważaniach.

Twierdzenie 26. (nierówność Schwarza albo Cauchy'ego-Buniakowskiego-Schwarza). Załóżmy, że

$$a_1, \ldots, a_n, \quad b_1, \ldots, b_n$$

są liczbami zespolonymi. Wtedy

$$\left| \sum_{i=1}^{n} a_i \bar{b}_i \right|^2 \leq \sum_{i=1}^{n} |a_i|^2 \sum_{i=1}^{n} |b_i|^2 \tag{9}$$

Dowód. Niech

$$A = \sum_{i=1}^{n} |a_i|^2, \quad B = \sum_{i=1}^{n} |b_i|^2 \quad C = \sum_{i=1}^{n} a_i \bar{b}_i.$$

Jeśli $B = 0$, to nie pozostaje nam nic do zrobienia. Do pracy przystępujemy mając tylko na uwadze przypadek $B > 0$. Zauważmy jeszcze, że

$$\overline{z_1 z_2} = \bar{z}_1 \bar{z}_2.$$

Po tym następują rachunki, gdzie L jest dodatnią wielkością

$$L = \sum_{i=1}^{n} |Ba_i - Cb_i|^2 = \sum_{i=1}^{n} (Ba_i - Cb_i)(B\bar{a}_i - \bar{C}\bar{b}_i)$$

i dalej

$$L = B^2 \sum_{i=1}^{n} |a_i|^2 - B\bar{C} \sum_{i=1}^{n} a_i \bar{b}_i - BC \sum_{i=1}^{n} \bar{a}_i b_i + |C|^2 \sum_{i=1}^{n} |b_i|^2,$$

a więc

$$0 \leq L = B(AB - |C|^2).$$

Skoro $B > 0$, to wynika stąd, że $AB - |C|^2 \geq 0$, co należało wykazać. □

Rozdział 2

Przestrzenie liniowe i układy równań liniowych

Poznamy ogólne metody postępowania z układami równań liniowych. Przy okazji wyjdzie na jaw, że badanie układów równań prowadzi do bogatej teorii, mającej wielorakie zastosowania, wykraczające poza pierwotnie wyznaczony cel.

2.1 Wprowadzenie

Zaczniemy od przykładu. Rozpatrzmy trzy układy równań

$$\begin{cases} x_1 + 2x_2 = a \\ 2x_1 + 2x_2 = b \end{cases} \tag{1}$$

$$\begin{cases} 2x_1 + 2x_2 = a \\ 2x_1 + 2x_2 = b \end{cases} \tag{2}$$

gdzie $b \neq 2a$ i

$$\begin{cases} x_1 + 2x_2 = a \\ 2x_1 + 2x_2 = 2a \end{cases} \tag{3}$$

Łatwo sprawdzić, że układ (1) ma dokładnie jedno rozwiązanie: wystarczy od drugiego równania odjąć pierwsze, co nam daje, że $x_1 = b - a$, a stąd $x_2 = a - b/2$. Ten sam argument pokazuje, że układ (2) nie ma rozwiązań. Łatwo zauważyć, że (3) ma ich wiele, np. $x_1 = 3a$, $x_2 = -a$ lub $x_1 = a$, $x_2 = 0$. Zastanówmy się nad strukturą zbioru rozwiązań układu (3). Niech $z_0 = (x_1^0, x_2^0)$ będzie dowolnym rozwiązaniem, kładziemy $y_i = x_i - x_i^0$, tj. $x_i = x_i^0 + y_i$, zatem y_1, y_2 spełniają

$$\begin{cases} y_1 + 2y_2 = 0 \\ 2y_1 + 4y_2 = 0 \end{cases} \tag{4}$$

Niech $Z = \{y \in \mathbb{R}^2 : \text{para } (y_1, y_2) \text{ jest rozwiązaniem układu (4)}\}$. Zauważmy, że jeśli $y \in Z$ to para $(\lambda y_1, \lambda y_2)$ też jest rozwiązaniem. Podobnie jeśli $z_1, z_2 \in Z$, gdzie $z_1 = (a, b)$ i $z_2 = (x, y)$ to para $z_1 + z_2 := (a + x, b + y)$ też jest rozwiązaniem. Wskazaliśmy zatem pewną dodatkową

strukturę zbioru Z, mianowicie umiemy dodawać elementy zbioru Z i mnożyć je przez liczbę rzeczywistą.

Rozpatrzmy teraz ciężarek zawieszony na sprężynie. Siła wypadkowa działając na ciężarek jest sumą siły ciążenia i siły sprężystości i być może innych sił działających na układ laboratoryjny. Jeśli uwzględnimy tylko te dwie pierwsze, to warunek równowagi zapisujemy jako

$$F_s + F_c = 0.$$

Przyjrzyjmy się herbacie wirującej w szklance pod wpływem mieszania łyżeczką. W każdym punkcie cieczy można zaczepić wektor prędkości herbaty, ale dodawanie tych wektorów nie ma sensu. W tym miejscu należy się dodatkowy komentarz dotyczący wektorów. Niech $E = \mathbb{R}^2$, wtedy *wektorem związanym $\vec{xy}$ zaczepionym w punkcie x o końcu w y* nazywamy parę punktów $(x, y) \in E \times E$. Wektorów zaczepionych w różnych punktach na ogół nie dodajemy, ale odczuwamy potrzebę ich porównywania. Do tego służy pojęcie wektora swobodnego. W tym celu w zbiorze wektorów związanych wprowadzimy relację równoważności $\sim$. Poprzedzimy ją definicją dodawania i odejmowania punktów płaszczyzny E. Jeśli $x = (x_1, x_2)$, $y = (y_1, y_2)$, to piszemy

$$y - x = (y_1 - x_1, y_2 - x_2), \quad y + x = (y_1 + x_1, y_2 + x_2).$$

Definicja 1.

$$\vec{ab} \sim \vec{xy} \qquad \Leftrightarrow \qquad b - a = y - x,$$

Łatwo sprawdzić, że $\sim$ jest relacją równoważności i *wektory swobodne na płaszczyźnie* to elementy,

$$E \times E/\sim,$$

tj. klasy równoważności.

Wektory swobodne można interpretować jako wektory związane, zaczepione w punkcie $(0, 0)$. Jest teraz jasne, że wektory swobodne można dodawać do siebie i mnożyć przez liczbę rzeczywistą. Należy się spodziewać, że istnieje wspólna matematyczna struktura, łącząca powyższe przykłady. Istotnie, mamy bowiem:

Definicja 2. Załóżmy, że $\mathbb{K}$ jest ciałem. Powiemy, że $(V, +, 0; \mathbb{K}, \cdot)$ jest *przestrzenią wektorową* nad ciałem $\mathbb{K}$, jeśli:
(PW1) $(V, +, 0)$ jest grupą abelową.
Działanie $\cdot : \mathbb{K} \times V \to V$ jest nazywane mnożeniem wektora przez liczbę. Dla dowolnych α, $\beta \in \mathbb{K}$, $v, w \in V$ spełnia ono warunki:
(PW2) $\alpha \cdot (v + w) = \alpha \cdot v + \alpha \cdot w$; $\qquad (\alpha + \beta) \cdot v = \alpha \cdot v + \beta \cdot v$;
(PW3) $\alpha \cdot (\beta \cdot v) = (\alpha \cdot \beta) \cdot v$;
(PW4) $1 \cdot v = v$.
W skrócie będziemy pisać, że V jest p.w. Równoważnie mówimy też, że V jest przestrzenią liniową. Elementy p.w. V nazywamy wektorami. Definicja pozwala, aby ciało $\mathbb{K}$ było dowolne, ale najważniejszy przypadek, to $\mathbb{K} = \mathbb{R}$ lub $\mathbb{K} = \mathbb{C}$.

Następujące stwierdzenie jest dość oczywiste, łatwo je udowodnić posługując się metodami zastosowanymi do wykazania podobnych właściwości ciał przemiennych. Dlatego też jego dowód pomijamy.

Stwierdzenie 1. Niech V będzie p.w. nad $\mathbb{K}$, $v \in V$, $\alpha \in \mathbb{K}$, wtedy

$$0 \cdot v = 0, \quad \alpha \cdot 0 = 0, \quad (-\alpha) \cdot v = \alpha \cdot (-v) = -(\alpha \cdot v).$$

Okazuje się, że przestrzenie liniowe są dość powszechnymi obiektami. Już teraz możemy wskazać dość pokaźną ich liczbę.

Przykłady 1.

(1) Nasze rozważania z początku paragrafu można podsumować, mówiąc, że zbiór Z rozwiązań równania (4) jest p.w. nad $\mathbb{R}$.

(2) Zbiór wektorów swobodnych na płaszczyźnie jest p.w. nad $\mathbb{R}$.

(3) Jeśli $x, y \in \mathbb{R}^n$, tj. $x = (x_1, x_2, \ldots, x_n)$ i $y = (y_1, y_2, \ldots, y_n)$, $\alpha \in \mathbb{R}$, to kładziemy

$$x + y := (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n), \quad \alpha \cdot x := (\alpha x_1, \alpha x_2, \ldots, \alpha x_n).$$

Wtedy $\mathbb{R}^n$ z działaniami określonymi wyżej jest p.w. nad $\mathbb{R}$.

(4) Jeśli zdefiniujemy w $\mathbb{C}^n$ dodawanie punktów i mnożenie przez liczbę zespoloną w sposób taki jak wyżej, to wtedy $\mathbb{C}^n$ jest p.w. nad $\mathbb{C}$.

(5) Ciało liczb zespolonych jest p.w. nad $\mathbb{R}$.

(6) Niech W będzie zbiorem wielomianów o współczynnikach rzeczywistych. Niech $f, g \in W$, tj.

$$f(x) = \sum_{i=0}^{n} a_i x^i, \quad g(x) = \sum_{i=0}^{m} b_i x^i.$$

Sumę i iloczyn przez liczbę określamy następująco:

$$(f + g)(x) := \sum_{i=0}^{\max\{n,m\}} (a_i + b_i) x^i, \quad (\lambda \cdot f)(x) := \sum_{i=0}^{n} (\lambda a_i) x^i,$$

gdzie przyjmujemy, że jeśli $n > m$, to $b_i = 0$, dla $i = m+1, \ldots, n$ (podobnie postępujemy, gdy $m > n$). W z powyższymi działaniami jest p.w. nad $\mathbb{R}$.

(7) Niech X będzie dowolnym zbiorem, zbiór wszystkich funkcji $f : X \to \mathbb{R}$ oznaczamy symbolem $\mathbb{R}^X$. Wprowadzimy w nim działania

$$(f + g)(x) := f(x) + g(x), \qquad (\lambda \cdot f)(x) := \lambda f(x).$$

Wtedy $\mathbb{R}^X$ jest p.w. nad $\mathbb{R}$.

(8) $\mathbb{Q}(\sqrt{2})$ jest p.w. nad ciałem liczb wymiernych $\mathbb{Q}$.

(9) Rozpatrzmy układ równań

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= 0 \\ &\ldots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= 0. \end{aligned} \tag{5}$$

Kładziemy, $Z = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : (x_1, \ldots, x_n)$ jest rozwiązaniem układu (5)$\}$. Wtedy Z jest p.w. nad $\mathbb{R}$.

Sprawdzenie pewników (PW1-4) jest dość prostym ćwiczeniem, dajmy na to w ostatnim przykładzie przebiega ono tak, jak w przypadku układu rozpatrywanego na początku paragrafu.

Chcemy podkreślić, że p.w. z przykładu 1(9) jest naszą najważniejszą motywacją i nią będziemy się głównie zajmowali. Ważnym pytaniem jest czy w ogólności istnieją niezerowe rozwiązania (5) a jeśli tak, to czy można o zbiorze Z powiedzieć coś więcej oprócz tego, że jest nieskończony i $Z \neq \{(0, \ldots, 0)\}$. Pierwszym wynikiem w tym kierunku jest poniższe stwierdzenie.

Stwierdzenie 2. Załóżmy, że współczynniki a_{ij} w układzie (5) należą do ciała $\mathbb{K}$. Jeśli $n > m$ to istnieją niezerowe rozwiązania (5).
Dowód przeprowadzimy przez indukcję względem m. Wprowadzimy oznaczenie,

$$L_i = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{n1}x_2.$$

Jeżeli $m = 1$, to mamy dwa przypadki: (i) wszystkie współczynniki a_{1j} są równe zeru lub (ii) istnieje j_0 takie, że $a_{1j_0} \neq 0$, np. $j_0 = 1$. W pierwszym przypadku nie mamy nic do zrobienia: dowolny element $x \in \mathbb{R}^n$ spełnia równanie $L_1 = 0$. W drugim wystarczy przyjąć $x_2 = \ldots = x_{n-1} = 0$, $x_n = 1$, wtedy $x_1 = -a_{1n}/a_{11}$.

Przyjmijmy teraz, że nasze twierdzenie jest prawdziwe dla pewnego n. Wykażemy, że jest prawdziwe dla $n + 1$. Mamy znowu dwa przypadki: (i) wszystkie współczynniki $a_{ij} = 0$ lub (ii) istnieją i_0 i j_0 takie, że $a_{i_0j_0} \neq 0$. Można przyjąć, że $a_{11} \neq 0$. Wtedy z układu (5) można wyrugować niewiadomą x_1 i dostaniemy układ:

$$\begin{gathered} L_2 - \tfrac{a_{21}}{a_{11}} L_1 = 0 \\ \ldots \\ L_m - \tfrac{a_{m1}}{a_{11}} L_1 = 0. \end{gathered}$$

Powyższy układ ma $m - 1$ równań i $n - 1$ niewiadomych, można więc do niego stosować założenie indukcyjne, by dostać niezerowe rozwiązanie $x_2, \ldots x_n$. Niewiadomą x_1 wyznaczamy z $L_1 = 0$. □

Przedstawiona w dowodzie metoda jest podstawą *metody eliminacji Gaussa*, którą można stosować w praktyce do rowiązywania układów równań liniowych. Więcej na ten temat będzie w paragrafie 3.2.

Zwróciliśmy już wcześniej uwagę na kwestię: jak wielki jest zbiór Z rozwiązań układu (5)? Odpowiedź odłożymy na później, do czasu wprowadzenia stosownego aparatu. Będzie on wyłożony w następnym paragrafie.

2.2 Liniowa niezależność

Zaczniemy od definicji pomocniczej właściwości.

Definicja 3. Niech $v_1, \ldots, v_n$ będą wektorami p.w. V nad $\mathbb{K}$, zaś $\lambda_1, \ldots, \lambda_n \in \mathbb{K}$, to wektor

$$\sum_{i=1}^{n} \lambda_i v_i$$

nazywa się *kombinacją liniową* wektorów $v_1, \ldots, v_n$ o współczynnikach $\lambda_1, \ldots, \lambda_n \in \mathbb{K}$.

Mając ją, sformułujemy ważne, tytułowe pojęcie. Można ją wysłowić jak następuje:

Definicja 4. Układ wektorów $\{v_i\}_{i\in I}$ nazywa się *liniowo niezależnym* (piszemy lnz), jeśli dla każdej kombinacji liniowej $\sum_{i\in I} \lambda_i v_i$ mamy,

$$\text{jeśli} \quad \sum_{i\in I} \lambda_i v_i = 0, \quad \text{to} \quad \lambda_i = 0 \text{ dla } i \in I.$$

Gdy nie wypowiadamy się na temat charakteru zbioru I i dopuszczamy, aby był on nieskończony, to powyższa suma oznacza, że tylko skończenie wiele współczynników λ_i jest różnych od zera. Podkreślamy też, że liniowa niezależność jest cechą **zbioru**.

Mówimy, że układ wektorów B jest *liniowo zależny* (piszemy: lz), jeśli nie jest prawdą, że jest lnz. Zanotujmy oczywiste właściwości, których dowód pomijamy.

Stwierdzenie 3. Niech V będzie p.w. i $v, v_1, \ldots, v_n \in V$. Wtedy

(a) $\{0\}$ jest układem wektorów zależnych;

(b) jeśli $v \neq 0$, to $\{v\}$ jest układem wektorów liniowo niezależnych;

(c) układ wektorów $v_1, \ldots, v_n$ jest lz, wtedy i tylko wtedy, gdy jeden z nich jest kombinacją liniową pozostałych.

Będą nas interesowały operacje na zbiorze $\mathcal{B}$, które nie zmieniają ilości wektorów liniowo niezależnych. Łatwo jest sprawdzić, że jest prawdziwy następujący fakt, który pozostawiamy Czytelnikowi do samodzielnego wykazania.

Stwierdzenie 4. Jeśli w zbiorze wektorów $\{v_1, \ldots, v_n\}$ znajduje się dokładnie $k \leq n$ wektorów lnz, to w zbiorach:

(a) $\{v_1, \ldots, v_{n-1}, \lambda v_n\}$, $\lambda \neq 0$;

(b) $\{v_1, \ldots, v_{n-1}, v_{n-1} + v_n\}$

jest dokładnie k wektorów lnz. □

Często zapisujemy wektory jako kombinacje liniowe z pewnego ustalonego zbioru wektorów. W tym momencie jest ważnym, aby wiedzieć kiedy taki zapis jest jednoznaczny. W badaniu tego problemu pomocne będzie pojęcie wprowadzone poniżej.

Definicja 5. Zbiór $\mathcal{B}$ nazywa się *bazą* p.w. V, jeśli każdy wektor z V daje się zapisać **jednoznacznie** w postaci kombinacji liniowej wektorów z $\mathcal{B}$. Jeśli $\mathcal{B}$ jest bazą, to dla każdego $v = \sum_{i\in I} \alpha_i v_i$ *współczynnikami* v w bazie $\mathcal{B}$ nazywamy liczby α_i.

Jeśli $\mathcal{O} \subset V$ i każdy wektor z V jest kombinacją liniową wektorów z $\mathcal{O}$, to mówimy, że $\mathcal{O}$ *rozpina* p.w. V i piszemy

$$V = \text{span } \mathcal{O} \qquad \text{albo} \qquad V = \text{lin } \mathcal{O}.$$

O użyteczności pojęcia bazy decyduje poniższy fakt, będący charakteryzacją bazy.

Stwierdzenie 5. $\mathcal{B}$ jest bazą p.w. V wtedy i tylko wtedy, gdy $\mathcal{B}$ jest zbiorem wektorów lnz i rozpinających V.

Dowód. $\Rightarrow$ Niech $\mathcal{B} = \{v_i\}_{i\in I}$. Gdyby wektory z $\mathcal{B}$ były l.z. to jeden z nich np. v_{i_0} byłby kombinacją liniową pozostałych,

$$v_{i_0} = \sum_{j\in I\setminus\{i_0\}} \alpha_i v_i.$$

Jednocześnie v_{i_0} jest kombinacją liniową wektorów z $\{v_{i_0}\} \subset \mathcal{B}$. Dostaniemy sprzeczność z jednoznacznością przedstawienia dowolonego wektora z V jako kombinacji liniowej wektorów z $\mathcal{B}$.

$\Leftarrow$ Niech $\mathcal{B}$ będzie zbiorem wektorów rozpinających V. Załóżmy, że wektor w ma dwa przedstawienia

$$\sum_{i\in I} \alpha_i v_i = w = \sum_{i\in I} \beta_i v_i.$$

Jeśliby dla pewnego $j \in I$, $\alpha_j \neq \beta_j$, to dostalibyśmy, że

$$0 = \sum_{i\in I} (\alpha_i - \beta_i) v_i$$

i nie wszystkie współczynnki są równe zero. Jest to sprzeczne z liniową niezależnością wektorów z $\mathcal{B}$. Co kończy dowód. □

Przykłady 2. Rozważmy $\mathbb{R}^2$. Niech $\mathbf{e}_1 = (1,0)$, $\mathbf{e}_2 = (0,1)$. Łatwo się przekonać, że wektory $\mathbf{e}_1$ i $\mathbf{e}_2$ tworzą bazę w $\mathbb{R}^2$. Tak samo łatwo jest się przekonać, że $\mathbf{d}_1 = (1,0)$ i $\mathbf{d}_2 = (1,1)$ stanowią bazę w $\mathbb{R}^2$. Tym samym wybór bazy jest niejednoznaczny. Zauważmy, że obie bazy mają tę samą ilość elementów. Można zapytać, czy jest to fakt ogólny. Odpowiedź jest poniżej

Stwierdzenie 6. Jeśli pewna baza p.w. V ma n elementów, to każda inna ma ich n.

Nim wykażemy to stwierdzenie, zadajmy sobie pytanie: Co by było, gdyby było nam dane l wektorów, podczas gdy pewna baza ma ich n i $l > n$? Musiałyby być l.z. Istotnie, mamy bowiem

Lemat 7. Jeśli $\mathcal{B} = \{v_1, \ldots, v_n\}$ jest bazą p.w. V i $l > n$, to jakikolwiek zbiór wektorów $\{w_1, \ldots, w_l\} \subset V$ jest l.z.

Dowód lematu. Wykażemy, że równanie

$$\lambda_1 w_1 + \ldots + \lambda_l w_l = 0$$

ma niezerowe rozwiązanie. Ponieważ $\mathcal{B}$ jest bazą, to istnieją liczby a_{ij}, $i = 1, \ldots, l$, $j = 1, \ldots, n$ takie, że $w_i = \sum_{j=1}^n a_{ij} v_j$. Zatem,

$$\lambda_1 \sum_{j=1}^n a_{1j} v_j + \ldots + \lambda_l \sum_{j=1}^n a_{lj} v_j = 0,$$

czyli po przegrupowaniu wyrazów:

$$\left(\sum_{i=1}^{l} \lambda_i a_{i1}\right) v_1 + \ldots + \left(\sum_{i=1}^{l} \lambda_i a_{in}\right) v_n = 0.$$

Z drugiej strony, wektory z $\mathcal{B}$ są lnz, zatem współczynniki w powyższej kombinacji liniowej znikają, tzn.

$$\begin{gathered} \lambda_1 a_{11} + \ldots + \lambda_l a_{l1} = 0 \\ \ldots \\ \lambda_1 a_{1n} + \ldots + \lambda_l a_{ln} = 0 \end{gathered}$$

Zauważmy, że niewiadomych jest więcej niż równań, a zatem na mocy stwierdzenia 2 dostaniemy istnienie niezerowych rozwiązań powyższego układu. Co kończy dowód lematu. □

Możemy zatem zająć się **dowodem stwierdzenia.** Powiedzmy, że dane są dwie bazy $\mathcal{B}$ i $\mathcal{B}'$. Niech $\mathcal{B}$ ma n elementów a $\mathcal{B}'$ ma ich m. Jeśli uznamy $\mathcal{B}$ za bazę a $\mathcal{B}'$ za dowolny zbiór, to z lematu dostaniemy, że m nie może być większe niż n, tj. $m \leq n$. Zamieniając rolami $\mathcal{B}$ i $\mathcal{B}'$ dostaniemy, że $n \leq m$. Stąd wynika, że $m = n$, co należało wykazać. □

Podaliśmy przykłady baz w $\mathbb{R}^2$. Warto zastanowić się, czy w każdej przestrzeni liniowej można znaleźć bazę. Najprostszy sposób postępowania jaki się narzuca, to wybrać w p.w. V wektor niezerowy v_1 i zadać pytanie czy zbiór $\{v_1\}$ rozpina V. Jeśli odpowiedź jest twierdząca, to kończymy postępowanie, jeśli nie, to znaczy, że istnieje taki wektor v_2, że wektory $\{v_1, v_2\}$ są lnz. Znowu możemy zadać pytanie, czy $\{v_1, v_2\}$ rozpinają V. Są dwie możliwe odpowiedzi i dwie drogi postępowania. Nie ma problemu, jeśli proces kończy się w skończonej ilości kroków. Gorzej, jeśli nie ma to miejsca. Powraca pytanie zasugerowane wcześniej: co to jest kombinacja liniowa nieskończenie wielu wektorów? Już wcześniej pisaliśmy

$$\sum_{i \in I} \lambda_i v_i, \tag{6}$$

niejako ukrywając charakter zbioru I. Przypominamy za definicją 3, że jeśli zbiór I jest nieskończony, to przyjmujemy, że w (6) wszystkie współczynniki, z wyjątkiem skończenie wielu, są zerowe, tj. tak naprawdę sumowanie jest skończone.

Nasze naiwne powyższe rozumowanie można uściślić i wykazać dwa poniższe twierdzenia, jednak nie mamy narzędzi, aby to zrobić.

Twierdzenie 8. Jeżeli V jest p.w. i $V \neq \{0\}$, to V ma bazę.

Ten wynik nie jest dla nas zaskoczeniem. Podobnie jak następne twierdzenie.

Twierdzenie 9. Jeżeli V jest p.w. i $\mathcal{B}$ jest zbiorem wektorów lnz, to $\mathcal{B}$ można rozszerzyć do bazy.

Możemy teraz podać charakteryzację zbioru Z z przykładu 1 (9). Poprawność poniższej definicji zapewnia stwierdzenie 6.

Definicja 6. Jeśli V jest p.w. i ma bazę złożoną z n elementów, to powiemy, że ma wymiar n i piszemy

$$\dim V = n.$$

W przeciwnym przypadku powiemy, że V ma wymiar nieskończony i piszemy

$$\dim V = \infty.$$

Przykłady 3. Policzymy teraz wymiary przestrzeni liniowych zdefiniowanych wcześniej.
(1) $\dim \mathbb{R}^n = n$.
(2) $\dim (\mathbb{C}, \mathbb{C}) = 1$, podkreślamy, że chodzi o wymiar przestrzeni wektorowej $\mathbb{C}$ nad $\mathbb{C}$.
(3) $\dim (\mathbb{C}, \mathbb{R}) = 2$, podkreślamy, że chodzi o wymiar przestrzeni wektorowej $\mathbb{C}$ nad $\mathbb{R}$.
(4) $\dim (\mathbb{C}^n, \mathbb{C}) = n$.
(5) $\dim (\mathbb{C}^n, \mathbb{R}) = 2n$.
(6) $\dim (\mathbf{Q}(\sqrt{2}), \mathbf{Q}) = 2$, wynika to wprost z faktu, iż $\sqrt{2}$ jest liczbą niewymierną.
(7) Niech W będzi przestrzenią wektorową wielomianów o współczynnikach rzeczywistych. Trzeba przypomnieć sobie (6) i co oznacza pozornie nieskończona suma. Zauważmy teraz, że wektory

$$\mathcal{B} = \{x^i\}_{i=0}^{\infty},$$

są lnz, bo jeśli $f \in W$, to oczywiście f jest kombinacją liniową wektorów z $\mathcal{B}$. Jeśli

$$f(x) = \sum_{i=0}^{n} a_i x^i = 0$$

dla wszystkich $x \in \mathbb{R}$, to znaczy, że f jest wielomianem stałym, równym zero, tj. $\mathcal{B}$ jest zbiorem wektorów lnz. Wynika stąd, że

$$\dim W = \infty.$$

2.2.1 Sumy proste

Potrzebne nam będą dodatkowe pojęcia pozwalające na rozkładanie przestrzeni liniowych na prostsze składniki.

Definicja 7. Niech V_i, $i \in I$ będą podprzestrzeniami p.w. V *Suma algebraiczna* $\sum_{i\in I} V_i$ jest złożona z wektorów postaci

$$v = \sum_{i\in I} v_i, \tag{7}$$

gdzie $v_i \in V_i$. Sumę algebraiczną nazywa się *sumą prostą*, jeśli każdy element v tej sumy algebraicznej można przedstawić jednoznacznie w postaci (7). Piszemy $\bigoplus_{i\in I} V_i$.

Potrzebna nam będzie charakteryzacja, kiedy suma algebraiczna jest prosta.

Stwierdzenie 10. Załóżmy, że V_i są podprzestrzeniami p.w. V. Wtedy, $\sum_{i\in I} V_i$ jest sumą prostą wtedy i tylko wtedy, gdy dla każdego $j \in I$, $V_j \cap \sum_{i\in I\setminus\{j\}} V_i = \{0\}$

Dowód. $\Rightarrow$ a.a. tj. zastosujemy metodę sprowadzenia do niedorzeczności. Załóżmy więc, że istnieje j takie, że

$$0 \neq w \in V_j \cap \sum_{i \in I \setminus \{j\}} V_i.$$

Zatem $w = \sum_{i \in I \setminus \{j\}} w_i$, gdzie $w_i \in V_i$. Wtedy kładziemy:

$$v_i = \begin{cases} w_i, & \text{gdy } i \neq j \\ 0, & \text{gdy } i = j \end{cases} \qquad v_i' = \begin{cases} 0 & \text{gdy } i \neq j \\ w, & \text{gdy } i = j. \end{cases}$$

Zatem w ma dwa różne przedstawienia:

$$\sum_{i \in I} v_i = w = \sum_{i \in I} v_i',$$

co daje żądaną sprzeczność.

$\Leftarrow$ a.a. Jeśli $\sum_{i \in I} V_i$ nie jest sumą prostą, to istnieje wektor w i takie układy wektorów $\{v_i\}_{\{i \in I\}}$ i $\{v_i'\}_{\{i \in I\}}$, że v_i i v_i' należą do V_i i

$$\sum_{i \in I} v_i = w = \sum_{i \in I} v_i',$$

nadto dla pewnego k, $v_k \neq v_k'$. Zatem

$$0 \neq v_k - v_k' = \sum_{i \in I \setminus \{k\}} (v_i' - v_i),$$

to znaczy, że istnieje niezerowy element przecięcia V_k i $\sum_{i \in I\{k\}} V_i$. Otrzymana sprzeczność kończy dowód. □

Sytuacja, gdy pewna przestrzeń liniowa jest sumą prostą **dwu** podprzestrzeni jest dość szczególna i warta odnotowania.

Definicja 8. Niech V_1, V_2 będą podprzestrzeniami p.w. V oraz $V = V_1 \oplus V_2$, to wtedy powiemy, że V_2 (odpowiednio V_1) jest *dopełnieniem* V_1 (odpowiednio V_2). *Współwymiarem* (kowymiarem) V_1 nazywamy wymiar podprzestrzeni do niej dopełniajacej, piszemy $\operatorname{codim} V_1 = \dim V_2$.

Przykład 4. Niech V_1 będzie prostą w $\mathbb{R}^3$ przechodzącą przez początek układu współrzędnych a V_2 niech będzie płaszczyzną, która nie zawiera V_1. Wtedy $V_1 \oplus V_2$ i $\operatorname{codim} V_1 = 2$ i $\operatorname{codim} V_2 = 1$.

Definicja 9. Niech V będzie p.w. a W jej podprzestrzenią wektorową. Wprowadzamy relację równoważności wzorem $v \,\rho\, w$ wtedy i tylko wtedy, gdy $v - w \in W$. Zamiast pisać V/ρ będziemy używać oznaczenia: V/W.

Przykład 5. Przy oznaczeniach powyższego przykładu niech $V = \mathbb{R}^3$, $W = V_1$, wtedy łatwo się przekonać, że V/W można utożsamiać z V_2. Uwaga: to nie jest ten sam obiekt.

Obiektem podobnym do sumy prostej i który może być z nią mylony jest iloczyn kartezjański p.w. Niech będą dane p.w. nad $\mathbb{K}$, V_i $i = 1, \ldots, n$. W zbiorze $W := V_1 \times V_2 \times \ldots \times V_n$ wprowadzamy działania w następujący sposób. Jeśli $v, w \in W$ i

$$v = (v_1, \ldots, v_n), \quad w = (w_1, \ldots, w_n),$$

to kładziemy

$$v + w = u, \quad \lambda v = t,$$

gdzie

$$u = (u_1, \ldots, u_n) \quad t = (t_1, \ldots, t_n) \qquad \text{i} \qquad u_i = v_i + w_i, \quad t_i = \lambda v_i.$$

W nazywamy *iloczynem kartezjańskim* pw V_i, $i = 1, \ldots, n$. Fakt, że z tak określonymi działaniami W jest p.w. nad $\mathbb{K}$ jest na tyle jasny, że jego dowód pomijamy.

2.3 Przestrzeń wektorowa macierzy

Możemy teraz omówić ważny przykład przestrzeni wektorowej, który okaże się bardzo pomocny w badaniu równań liniowych. Poświęcimy mu niniejszy podrozdział. Zaczniemy od określenia głównego obiektu naszego zainteresowania. Zakładamy w całym podrozdziale, że $\mathbb{K}$ jest dowolnym ciałem, aczkolwiek nasza uwaga jest skupiona na przypadkach $\mathbb{K} = \mathbb{R}$ lub $\mathbb{K} = \mathbb{C}$.

Definicja 10. *Macierzą* nad ciałem $\mathbb{K}$ o wymiarach $m \times n$ nazywamy dowolną funkcję $A : \{1, \ldots, m\} \times \{1, \ldots, n\} \to \mathbb{K}$. Piszemy $A = \{a_{ij}\}_{i=1,j=1}^{m,n}$ lub po prostu $\{a_{ij}\}$, jeśli zakres zmienności wskaźników jest znany. Zbiór macierzy $m \times n$ oznaczamy przez $M_{m\times n}(\mathbb{K})$.

Zauważmy jeszcze, że w $M_{m\times n}(\mathbb{K})$ można w naturalny sposób wprowadzić działania dodawania i mnożenia przez liczbę: Jeśli A, B są macierzami, to

$$A + B =: C, \quad \lambda A =: D,$$

gdzie

$$C = \{c_{ij}\}_{i=1,j=1}^{m,n}, \quad c_{ij} = a_{ij} + b_{ij}, \qquad D = \{d_{ij}\}_{i=1,j=1}^{m,n}, \quad d_{ij} = \lambda a_{ij}.$$

Łatwo się przekonać, że $M_{m\times n}(\mathbb{K})$ z tak określonymi działaniami jest p.w. nad $\mathbb{K}$.

Macierz nazywamy *kwadratową* jeśli $m = n$. Określimy teraz szczególną macierz kwadratową , $Id = \{a_{ij}\}_{i,j=1}^{n}$ nazywaną *macierzą jednostkową*, a mianowicie

$$a_{ij} = \begin{cases} 1, & \text{gdy } i = j \\ 0, & \text{w przeciwnym przypadku.} \end{cases}$$

Piszemy też I i wymiennie mówimy o I lub Id, że jest macierzą *tożsamościową* lub *identycznościową*. Później się przekonamy, że nazwa „macierzy tożsamościowej" jest w pełni uzasadniona.

Niech będzie dana macierz $A = \{a_{ij}\}_{i=1,j=1}^{m,n} \in M_{m\times n}(\mathbb{K})$, wtedy macierze $C_j \in M_{m\times 1}(\mathbb{K})$, $j = 1, \ldots, n$ i $R_i \in M_{1\times n}(\mathbb{K})$, $i = 1, \ldots, m$, dane wzorami

$$C_j = \begin{bmatrix} a_{1j} \\ \vdots \\ a_{1m} \end{bmatrix}, \quad R_i = [a_{i1} \ldots a_{in}]$$

nazywamy odpowiednio *j–tą kolumną* i *i–tym wierszem* macierzy A.

Wprowadzimy teraz szereg operacji na macierzach, zaczynając od operacji transpozycji, $T : M_{m\times n}(\mathbb{K}) \to M_{n\times m}(\mathbb{K})$, której wartość na macierzy A utarło się oznaczać A^T. Jeśli

$A = \{a_{ij}\}_{i=1,j=1}^{m,n}$, to kładziemy $A^T = B$, gdzie $B = \{b_{ij}\}_{i=1,j=1}^{n,m}$ i $b_{ij} = a_{ji}$. Macierz A^T nazywamy *macierzą transponowaną* macierzy A. Zauważmy przy okazji, że $(A^T)^T = A$.

Macierze mają nam ułatwić badanie rozwiązalności układów równań liniowych. Do tego potrzebne nam będzie kolejne ważne pojęcie:

Definicja 11. *Rzędem kolumnowym* (odpowiednio: *wierszowym*) macierzy A nazywa się ilość jej lnz kolumn (odpowiednio: wierszy).

Będziemy badać (liczyć) rzędy macierzy. Do tego celu przydatne jest poniższe stwierdzenie.

Stwierdzenie 11. Niech $C_1, \ldots, C_n$ (odpowiednio: $R_1, \ldots, R_m$) będą kolumnami (odpowiednio: wierszami) macierzy A. Wtedy rząd kolumnowy (odpowiednio: wierszowy) macierzy jest równy kowymiarowi w K^n (odpowiednio: w K^m) podprzestrzeni złożonej z rozwiązań

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \left(\text{odpowiednio: } \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \right)$$

równania

$$\sum_{i=1}^{n} x_i C_i = 0 \quad \left(\text{odpowiednio: } \sum_{i=1}^{m} x_i R_i = 0 \right). \tag{8}$$

Dowód przeprowadzimy dla wersji kolumnowej. Wersję wierszową uzyskamy zastępując macierz macierzą transponowaną.

Wiemy, że rozwiązania (8) tworzą podprzestrzeń wektorową. Niech rząd kolumnowy A wynosi k i kolumny $\{C_{n-k+1}, \ldots, C_n\}$ są lnz. Możemy więc przepisać (8) w postaci

$$-\sum_{i=1}^{n-k} x_i C_i = \sum_{i=n-k+1}^{n} x_i C_i. \tag{9}$$

Dla ustalonego $j \in \{1, \ldots, n-k\}$ kładziemy $x_j = 1$ i $x_i = 0$ dla $i \neq$ i $1 \leq i \leq n-k$, tj. (9) przyjmuje postać

$$-C_j = \sum_{i=n-k+1}^{n} x_i C_i.$$

Na mocy liniowej niezależności C_j, $n-k+1 \leq j \leq n$ liczby x_i w powyższym równaniu są wyznaczone jednoznacznie i oznaczymy je $x_{n-k+1}^j, \ldots, x_n^j$, tj. dla każdego j mamy rozwiązanie v_j równania (9):

$$v_j = (0, \ldots, 0, 1, 0, \ldots, 0, x_{n-k+1}^j, \ldots, x_n^j)^T,$$

gdzie jedynka jest na j–tym miejscu. Od razu też widać, że zbiór wektorów v_j, $1 \leq j \leq n-k$ jest lnz. Za chwilę przekonamy się, że jeśli $v = (\alpha_1, \ldots, \alpha_n)$ jest rozwiązaniem, to

$$v = \sum_{i=1}^{n-k} \alpha_i v_i.$$

W tym celu przyjrzyjmy się różnicy wektorów

$$w = v - \sum_{i=1}^{n-k} \alpha_i v_i.$$

Widać że spełnia ona (8) i ma postać

$$(0, \ldots, 0, \beta_{n-k+1}, \ldots, \beta_n).$$

Skoro spełnia (8), to z liniowej niezależności C_j dostajemy, że $\beta_k = 0$, $n-k+1 \leq k \leq n$. Tym samym wykazaliśmy, że dowolne rozwiązanie układu (8) ma jednoznaczne przedstawienie jako kombinacja liniowa wektorów $\{v_i\}_{i=1}^{n-k}$, tzn. tworzą one bazę podprzestrzeni rozwiązań (8), jej kowymiar jest równy k. □

Wynika stąd prosty fakt pomocny przy badaniu przestrzeni rozwiązań równań liniowych.

Stwierdzenie 12. Rząd wierszowy macierzy A jest równy jej rzędowi kolumnowemu.

Dowód. Niech $A = \{a_{ij}\}_{i,j=1}^{m,n}$ i $R_1, \ldots, R_m$ będą wierszami zaś $C_1, \ldots, C_n$ kolumnami i rząd wierszowy to r i rząd kolumnowy to c. Niech $\{R_{i_1}, \ldots, R_{i_r}\}$ będą liniowo niezależnymi wierszami. Wtedy rzędy macierzowe macierzy

$$A = \begin{bmatrix} R_1 \\ \vdots \\ R_m \end{bmatrix}, \quad B = \begin{bmatrix} R_{i_1} \\ \vdots \\ R_{i_r} \end{bmatrix}$$

są równe. W myśl poprzedniego stwierdzenia rząd kolumnowy A wyznaczony jest przez rozwiązania układu

$$\begin{gathered} a_{11}x_1 + \ldots + a_{1n}x_n = 0 \\ \ldots \\ a_{m1}x_1 + \ldots + a_{mn}x_n = 0 \end{gathered}$$

a rząd kolumnowy B wyznaczony jest przez rozwiązania układu

$$\begin{gathered} a_{i_1 1}x_1 + \ldots + a_{i_1 n}x_n = 0 \\ \ldots \\ a_{i_r 1}x_1 + \ldots + a_{i_r n}x_n = 0 \end{gathered}$$

Jest oczywiste, że drugi układ powstaje z wykreślenia równań, które są liniowo zależne. A skoro tak, to oba zbiory rozwiązań są równe. Wiedząc, że rzędy kolumnowe A i B są równe, zauważmy, że rząd kolumnowy B wynosi najwyżej r. Wynika to stąd, że kolumny B są elementami p.w. $\mathbb{K}^r$ i $\dim \mathbb{K}^r = r$, zatem $c \leq r$. Ten sam argument zastosowany do A^T daje $r \leq c$ i ostatecznie $r = c$, co należało wykazać. □

Dzięki temu stwierdzeniu następujące określenie jest poprawne:

Definicja 12. *Rzędem macierzy* nazywamy jej rząd kolumnowy lub rzędowy.

Ważnym zadaniem jest ustalenie jakie operacje na macierzach nie zmieniają ich rzędu. Odpowiedź jest zawarta poniżej.

Stwierdzenie 13. Rząd macierzy nie zmieni się, jeśli:

(a) do wiersza (odpowiednio: kolumny) dodamy inny wiersz (odpowiednio: kolumnę);

(b) wiersz (odpowiednio: kolumnę) pomnożymy przez liczbę różną od zera;

(c) przestawimy wiersze (odpowiednio: kolumny).

Dowód polega na zastosowaniu stwierdzenia 2. □

Wprowadzimy teraz operację na macierzach, której uzasadnienie musi nieco poczekać. Mamy na myśli mnożenie macierzy.

Definicja 13. Niech $A \in M_{m\times n}(\mathbb{K})$, $A = \{a_{ik}\}$, $B \in M_{n\times r}(\mathbb{K})$, $B = \{b_{kj}\}$. *Iloczyn macierzy* $A \cdot B$ to macierz $\{\gamma_{ij}\} \in M_{m\times r}(\mathbb{K})$ dana wzorem

$$\gamma_{ij} := \sum_{k=1}^{n} a_{ik}b_{kj}, \quad i = 1, \ldots, m,\ j = 1, \ldots, r.$$

Łatwo się pokazuje, że mnożenie macierzy jest łączne i rozdzielne względem dodawania. Powierzamy to zadanie Czytelnikowi do samodzielnego wykonania. Natomiast mnożenie nie jest przemienne, wynik istotnie zależy od kolejności np. niech $A \in M_{1\times m}(\mathbb{R})$, $B \in M_{m\times 1}(\mathbb{R})$, to $A \cdot B \in \mathbb{R}$, ale $B \cdot A \in M_{m\times m}(\mathbb{R})$. Nawet jeśli ograniczymy się do macierzy kwadratowych, to wynik zależy od kolejności działań, np.

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

Zauważmy, że zgodnie z nazwą macierzy tożsamościowej dla każdej macierzy $A \in M_{n\times n}(\mathbb{K})$ mamy, iż

$$A \cdot I = I \cdot A = A.$$

W tym momencie można zadać pytanie, czy dla każdej macierzy $A \in M_{n\times n}(\mathbb{K})$ istnieje macierz B odwrotna do niej tj. taka, że $A \cdot B = B \cdot A = I$. Odpowiedź jest przecząca, np. niech

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \text{wtedy} \quad A \cdot A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = 0.$$

Gdyby istniała macierz odwrotna A^{-1}, to mielibyśmy

$$A = A \cdot I = A \cdot (A \cdot A^{-1}) = (A \cdot A)A^{-1} = 0.$$

Co nie jest prawdą.

Nasze dotychczasowe rozważania dotyczące macierzy kwadratowych muszą zostać uzupełnione o fakty niezbędne do definicji wyznacznika. Temu celowi służy następny paragraf.

2.3.1 Dygresja na temat permutacji

W podrozdziale o kombinatoryce (patrz definicja 20 w §1.7) wprowadziliśmy pojęcie permutacji zbioru n-elementowego. Umówimy się oznaczać ich zbiór symbolem $\Pi(n)$. Obecnie spojrzymy na permutacje z algebraicznego punktu widzenia. Jeśli $\sigma \in \Pi(n)$, to działanie σ na elementach zbioru $\{1, \ldots, n\}$ opisujemy następująco $\begin{pmatrix} 1 & 2 & \ldots & n \\ \sigma(1) & \sigma(2) & \ldots & \sigma(n) \end{pmatrix}$, gdzie w dolnym wierszu piszemy $\sigma(i)$ pod i by oznaczyć wartość σ na liczbie i. Przykładowo wypiszemy wszystkie permutacje ze zbioru $\Pi(3)$

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}.$$

Przypominamy, że permutacje są funkcjami, można je składać i złożenie też jest permutacją, co więcej jeśli $\sigma, \tau, \omega \in \Pi(n)$, to $(\sigma \circ \tau) \circ \omega = \sigma \circ (\tau \circ \omega)$. Ponieważ każdy element $\sigma \in \Pi(n)$ jest funkcją wzajemnie jednoznaczną, to istnieje funkcja odwrotna σ^{-1} taka, że $\sigma \circ \sigma^{-1} = \sigma^{-1} \circ \sigma = id$, gdzie id jest permutacją tożsamościową, tj. $id(i) = i$. Podsumowując stwierdzamy, że $\Pi(n)$ z wyróżnioną permutacją id jest grupą. Co więcej, jest to naturalny przykład grupy nieabelowej.

Wróćmy do wypisanych wyżej permutacji z $\Pi(3)$: druga, trzecia i czwarta są szczególne, są bowiem przestawieniami sąsiednich elementów:

$$\tau_{ij}(k) = \begin{cases} k & k \neq i, j \\ j & k = i \\ i & k = j \end{cases}.$$

Permutację τ_{ij} nazywamy *transpozycją*. Rolę transpozycji wyjaśnia następujące stwierdzenie, którego dowód pozostawiamy Czytelnikowi do samodzielnego przemyślenia.

Stwierdzenie 14. Każda permutacja $\sigma \in \Pi(n)$ jest złożeniem pewnej ilości transpozycji, tj.

$$\sigma = \tau_{i_1 j_1} \circ \tau_{i_2 j_2} \circ \ldots \circ \tau_{i_r j_r} \tag{10}$$

Oczywiście owo przedstawienie nie jest jednoznaczne, ale następujący fakt jest prawdziwy i łatwy do okazania.

Stwierdzenie 15. Jeśli $\sigma \in \Pi(n)$, to liczba

$$(-1)^r,$$

gdzie r jest ilością transpozycji we wzorze (10), nie zależy od reprezentacji (10).

Dzięki temu stwierdzeniu następująca definicja jest poprawna.

Definicja 14. Jeśli $\sigma \in \Pi(n)$ to powiemy, że σ jest *parzysta* (odpowiednio: *nieparzysta*), jeśli $(-1)^r = 1$ (odpowiednio: $(-1)^r = -1$). Piszemy

$$\operatorname{sgn} \sigma = (-1)^r$$

Przykład 6. Spośród wypisanych permutacji z $\Pi(3)$ parzyste są: pierwsza, piąta, szósta, pozostałe są nieparzyste.

2.3.2 Wyznacznik macierzy kwadratowej

Tak jak zapowiadaliśmy, pokażemy w jaki sposób powyższe uwagi stosują się w definicji wyznacznika macierzy. Najpierw podamy ogólne określenie, potem objaśnimy je na przykładach i podamy zasadnicze właściwości. Wprawdzie zasadnicze przypadki wymiarów, tj. 2 i 3 nie wymagają rozbudowanej teorii, lecz spójność wykładu i potrzeba pełnej teorii skłaniają nas ku ogólnej definicji.

Definicja 15. Niech $B \in M_{n\times n}(\mathbb{K}), B = \{b_{ij}\}$, liczbę

$$\det B := \sum_{\sigma\in\Pi(n)} \operatorname{sgn}\sigma \;\; b_{\sigma(1)1}, b_{\sigma(2)2} \ldots . b_{\sigma(n)n}$$

nazywamy *wyznacznikiem* macierzy B.

Przykłady 7. Powyższa definicja w wymiarach 2 i 3 przyjmuje następującą postać:

$$\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

$$\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{21}a_{32}a_{13} - a_{31}a_{22}a_{13} - a_{32}a_{23}a_{11} - a_{12}a_{21}a_{33}$$

Zajmiemy się teraz badaniem właściwości wyznacznika. W tym celu zapiszemy macierz A w postaci kolumnowej tj. $A = (C_1, C_2, \ldots C_n)$. Mamy wtedy,

Stwierdzenie 16. Jeśli $A \in M_{n\times n}(\mathbb{K})$, $C_k^1 \in \mathbb{K}^n$ i $\alpha \in \mathbb{K}$, to wtedy

(a_1) $\det(C_1, \ldots, C_k + C_k^1, \ldots, C_n) = \det(C_1, \ldots, C_k, \ldots, C_n) + \det(C_1, \ldots, C_k^1, \ldots, C_n)$;

(a_2) $$\det(C_1, \ldots, \alpha C_k, \ldots, C_n) = \alpha \det(C_1, \ldots, C_k, \ldots, C_n).$$

(b) Przestawienie 2 kolumn prowadzi do zmiany znaku, tj.

$$\det(C_1, \ldots, C_i, \ldots, C_j, \ldots, C_n) = -\det(C_1, \ldots, C_j, \ldots, C_i, \ldots, C_n).$$

(c) Jeśli $C_i = C_j$ to $\det A = 0$.
(d) $\det I = 1$.
(e) $\det 0 = 0$.
(f) Jeśli $A, B \in M_{n\times n}(\mathbb{K})$ i $A = (C_1, \ldots, C_n)$, $B = \{\beta_{ij}\}$, to

$$\det(A \cdot B) = \det A \sum_{\sigma\in\Pi(n)} \operatorname{sgn}\sigma \beta_{\sigma(1)1} \ldots \beta_{\sigma(n)n}$$

Dowód. (a) Żądane właściwości wynikają wprost z definicji wyznacznika.

(b) Definicja zastosowana do prawej strony pokazuje, że wszystkie występujące tam permutacje zostały dodatkowo złożone z transpozycją elementów i-tego i j-tego.

(c) Z (b) wynika, że $\det(C_1, \ldots, C_i, \ldots, C_i, \ldots, C_n) = -\det(C_1, \ldots, C_i, \ldots, C_i, \ldots, C_n)$. Zatem, $\det A = -\det A$, a więc $\det A = 0$.

(d) i (e) wynikają wprost z definicji.

(f) Niech $A \cdot B = (C'_1, \ldots, C'_n)$, z definicji mnożenia macierzy wynika, że $C'_k = \sum_{i=1}^{n} \beta_{ik}\, C_i$, tj. z (a) i (c) dostaniemy

$$\begin{aligned}
\det(C'_1, \ldots, C'_n) &= \sum_{i_1,\ldots,i_n=1}^{n} \det(\beta_{i1} C_{i1}, \ldots, \beta_{i_n n} C_{i_n}) \\
&= \sum_{i_1,\ldots,i_n=1}^{n} \beta_{i_1 1} \ldots \beta_{i_n n} \det(C_{i_1}, \ldots, C_{i_n}) \\
&= \sum_{\sigma \in \Pi(n)} \beta_{\sigma(1)1} \ldots \beta_{\sigma(n)n} \det(C_{\sigma(1)}, \ldots, C_{\sigma(n)}) \\
&= \sum_{\sigma \in \Pi(n)} \beta_{\sigma(1)1} \ldots \beta_{\sigma(n)n} \operatorname{sgn}\sigma \det(C_1, \ldots, C_n).
\end{aligned}$$

□

Dowód części (f) prowadzi do ciekawego wniosku, a mianowicie:

Stwierdzenie 17. Jeśli $F : M_{n\times n}(\mathbb{K}) \to \mathbb{K}$ spełnia (a), (b) i (d), to

$$F(A) = \det A$$

Dowód. Dowód części (f) pokazuje, że (a) i (b) prowadzą do równości

$$F(A \cdot B) = F(A) \sum_{\sigma \in \Pi(n)} \beta_{\sigma(1)1} \ldots \beta_{\sigma(n)n} \operatorname{sgn}\sigma = F(A) \det B.$$

Zatem $F(I \cdot A) = F(I) \sum_{\sigma\in\Pi(n)} \operatorname{sgn}\sigma a_{\sigma(1)1}, \ldots a_{\sigma(n)n} = 1 \cdot \det A$ co kończy dowód. □

Zauważmy teraz, że dla dowolnej macierzy kwadratowej mamy

Stwierdzenie 18. $\det A = \det A^T$.

Dowód. Zauważmy, że $\operatorname{sgn}\sigma = \operatorname{sgn}\sigma^{-1}$. Wtedy

$$\det A^T = \sum_{\sigma\in\Pi(n)} \operatorname{sgn}\sigma\; a_{1\sigma(1)} \ldots a_{n\sigma(n)} = \sum_{\sigma\in\Pi(n)} \operatorname{sgn}\sigma a_{\sigma^{-1}\sigma(1)\sigma(1)} \ldots a_{\sigma^{-1}\sigma(n)\sigma(n)}.$$

Po zamianie kolejności mnożenia dostaniemy

$$\det A^T = \sum_{\sigma^{-1}\in\Pi(n)} \operatorname{sgn}\sigma^{-1} a_{\sigma^{-1}(1)1} \ldots a_{\sigma^{-1}(n)n},$$

ale permutacje odwrotne wyczerpują zbiór $\Pi(n)$. Zatem,

$$\det A^T = \sum_{\sigma\in\Pi(n)} \operatorname{sgn}\sigma a_{\sigma(1)1}\dots a_{\sigma(n)n} = \det A. \qquad \square$$

Podamy teraz ważny fakt łączący liniową zależność z zerowaniem się wyznacznika układu wektorów. Będziemy z tego faktu często korzystać.

Stwierdzenie 19. Jeśli $A \in M_{n\times n}(\mathbb{K})$, to wtedy $\det A = 0 \Leftrightarrow$ kolumny macierzy A są liniowo zależne.

Dowód. $\Leftarrow$ z liniowej zależności wynika, że pewna kolumna C_k jest kombinacją liniową pozostałych, tj.

$$C_k = \sum_{i\neq k} \lambda_i C_i$$

Z tej równości i stwierdzenia 16 wynika, że

$$\det A = \det(C_1,\dots,C_n) = \sum_{i\neq k} \det(C_1,\dots,\lambda_i C_i,\dots,C_n).$$

Każda z macierzy występujących po prawej stronie ma to do siebie, że na pozycji k-tej pojawia się kolumna i-ta. Zatem ze stwierdzenia 16 (c) wynika, że wszystkie wyznaczniki są równe zeru, tj. $\det A = 0$.

$\Rightarrow$ Zastosujemy metodę sprowadzenia do niedorzeczności. Oznacza to, że zakładamy iż $C_1,\dots,C_n$ stanowią bazę w $\mathbb{K}^n$. Wektory $\mathbf{e}_1,\dots,\mathbf{e}_n$, gdzie $\mathbf{e}_i$ jest wektorem, który na i-tej współrzędnej ma 1, na pozostałych 0, też stanowią bazę $\mathbb{K}^n$. Zatem istnieją β_{ik} takie, że

$$\mathbf{e}_k = \sum_{i=1}^{n} \beta_{ik}\, C_i$$

i mamy równość macierzy

$$(\mathbf{e}_1,\dots,\mathbf{e}_n) = (C_1,\dots,C_n)\cdot B,$$

gdzie $B = \{\beta_i\}$. Wyznacznik lewej strony jest równy 1 (na mocy stwierdzenia 16 (d)) zaś prawej z (f) i z założenia jest równy

$$\det(C_1,\dots,C_n)\det B = 0.$$

Jest to sprzeczne z założeniem. $\square$

Ponieważ wykazaliśmy wcześniej, że $\det A = \det A^T$, to możemy stwierdzić, że każda właściwość wyznacznika wyrażona w terminach kolumn jest prawdziwa, jeśli wyrazimy ją w terminach wierszy. Będzie to wielce przydatne.

Podamy teraz praktyczny sposób obliczania niedużych wyznaczników. Poprzedzimy go nowym określeniem.

Definicja 16. Załóżmy, że $A \in M_{n\times n}(\mathbb{K})$ i $n \geq 2$. Z macierzy A skreślamy wiersz i kolumnę zawierające wyraz a_{ij}. Wyznacznik uzyskanej macierzy $(n-1)$ na $(n-1)$ pomnożony przez $(-1)^{i+j}$ nazywamy *dopełnieniem algebraicznym* elementu a_{ij} i oznaczamy A_{ij}.

Przedstawimy teraz wspomnianą wyżej metodę.

Twierdzenie 20. (rozwinięcie Laplace'a) Niech $A = \{a_{ij}\} \in M_{n\times n}(\mathbb{K})$ i $n \geq 2$. Wtedy

$$\det A = \sum_{i=1}^{n} a_{1i} \cdot A_{1i}.$$

Dowód polega na sprawdzeniu, iż prawa strona spełnia założenia stwierdzenia 17 o jednoznaczności wyznacznika. Szczegóły rachunkowe pozostawiamy zainteresowanemu czytelnikowi. □

Rozwinięcie Laplace'a można nieco poprawić, mamy bowiem

Wniosek 21. Niech $A = (a_{ij}) \in M_{n\times n}(\mathbb{K})$, $n \geq 2$, zaś i jest dowolnym wskaźnikiem, $i = 1, \ldots, n$. Wtedy,

$$\det A = \sum_{k=1}^{n} a_{ki}A_{ki} = \sum_{k=1}^{n} a_{ik}A_{ik}$$ □

Z uwagi na znaczenie zbioru macierzy odwracalnych wymiaru n oznaczamy ich zbiór osobnym symbolem, $GL(n, \mathbb{K})$. Podsumowując poznane właściwości macierzy odwracalnych jest teraz oczywistym, że

Stwierdzenie 22. $GL(n, \mathbb{K})$ jest grupą. □

Mając na uwadze zastosowania macierzy odwracalnych, ważnym jest umiejętność scharakteryzowania elementów $GL(n, \mathbb{K})$. Jest to treścią następującego twierdzenia.

Twierdzenie 23. Macierz $A \in M_{n\times n}(\mathbb{K})$ jest odwracalna wtedy i tylko wtedy, gdy $\det A \neq 0$.
Dowód. $\Rightarrow$ Skoro A jest odwracalna, to $A \cdot A^{-1} = I$. A zatem,

$$1 = \det I = \det(A \cdot A^{-1}) = \det A \cdot \det A^{-1},$$

czyli $\det A \neq 0$.

$\Leftarrow$ Jeśli $\det A \neq 0$, to definiujemy $B = \{\beta_{ij}\}$ następująco

$$\beta_{jk} = \frac{1}{\det A} A_{kj}.$$

Sprawdzimy, że $A \cdot B = I = B \cdot A$. Jeśli $A \cdot B = \{\gamma_{ik}\}$, to z definicji mnożenia macierzowego dostaniemy, że

$$\gamma_{ik} = \sum_{j=1}^{n} \alpha_{ij}\beta_{jk} = \frac{1}{\det A} \sum_{i} \alpha_{ij}A_{ki}.$$

Jeśli $k = i$, to $\gamma_{ii} = 1$, dzięki rozwinięciu Laplace'a wyznacznika macierzy A. Jeśli $k \neq i$, z tegoż samego rozwinięcia dostaniemy, że γ_{ik} jest wyznacznikiem macierzy takiej, że jej kolumny i-ta i k-ta są równe zatem $\gamma_{ik} = 0$ dla $k \neq i$ tym samym $\{\gamma_{ik}\} = Id$. □

Łatwo jest też przekonać się o prawdziwości następującego faktu

Wniosek 24. $\det A^{-1} = \det A$.

Dowód. Mamy bowiem $1 = \det I = \det(AA^{-1}) = \det A \cdot \det A^{-1}$. □

Będzimy musieli liczyć rzędy macierzy, niekoniecznie kwadratowych, do tego celu pomocnym jest następujące pojęcie.

Definicja 17. Niech $A \in M_{m\times n}(\mathbb{K})$ i $k \leq m$. Macierz $B \in M_{k\times k}(\mathbb{K})$ powstałą poprzez wykreślenie z A dowolnych $m-k$ wierszy i $n-k$ kolumn nazywa się *minorem stopnia* k macierzy A.

Powiążemy teraz rząd macierzy z wyznacznikami jej minorów. Okaże się to później przydatne przy badaniu równań liniowych.

Stwierdzenie 25. Jeśli $A \in M_{m\times n}(\mathbb{K})$, to wtedy

(a) $\operatorname{rz} A \geq k \Leftrightarrow$ istnieje minor B stopnia k taki, że $\det B \neq 0$;

(b) każdy minor o niezerowym wyznaczniku ma wymiar nie większy niż $k \Leftrightarrow \operatorname{rz} A \leq k$.

Dowód. (a) $\Rightarrow$ Niech kolumny $C_{i_1}, \ldots, C_{i_k}$ będą lnz. Wtedy rząd kolumnowy macierzy $D = (C_{i_1}, \ldots, C_{i_k})$ jest równy k. Skoro jest on równy rzędowi wierszowemu, to istnieje k wierszy macierzy D, które są lnz. Tworzą one żądany minor B o niezerowym wyznaczniku.

$\Leftarrow$ Z drugiej strony, jeśli minor B stopnia k ma niezerowy wyznacznik, to znaczy, że jego kolumny są lnz. Powstały one ze skreślenia pewnej ilości wierszy z kolumn $C_{i_1}, \ldots, C_{i_k}$ macierzy A, zatem kolumny $C_{i_1}, \ldots, C_{i_k}$ są lnz i $\operatorname{rz} A \geq k$.

(b) Łatwy dowód przez sprowadzenie do niedorzeczności (w każdą) ze stron pozostawiamy czytelnikowi. Należy skorzystać z części (a). □

Powyższe stwierdzenie prowadzi do następującego wniosku.

Wniosek 26. Macierz $A \in M_{n\times n}(\mathbb{K})$ ma rząd $k \Leftrightarrow$ istnieje minor stopnia k o niezerowym wyznaczniku i każdy minor stopnia większego niż k znika.

O rozwiązywalności równań będzie nam się wygodniej mówiło mając dodatkową podbudowę teoretyczną uzyskaną w następnym paragrafie.

2.4 Odwzorowania liniowe

Nasz cel to teoria układów równań. Pragnienie uczynienia naszego zapisu zwięzłym prowadzi do wniosku, że badamy czy dany punkt leży w obrazie pewnego szczególnego przekształcenia, którego definicję zaraz podamy.

Definicja 18. Niech V i W będą p.w. nad $\mathbb{K}$. Powiemy, że funkcja $F : V \to W$ jest *odwzorowaniem liniowym* (albo po prostu jest liniowa, albo jest *homomorfizmem* przestrzeni wektorowych) jeśli dla dowolnych $w, u \in V$ i $\alpha \in K$ jest prawdą, że

(a) $F(w+u) = F(w) + F(u)$;

(b) $F(\alpha w) = \alpha F(w)$.

Zbiór odwzorowań liniowych oznaczamy przez Hom (V, W). Jeśli $F \in$ Hom (V, W) jest różnowartościowe i „na", to powiemy, że F jest *izomorfizmem* liniowym, ich zbiór to Iso(V, W). Jeśli zbiór Iso(V, W) jest niepusty, to powiemy, że p.w. V i W są *izomorficzne*.

Wprowadzimy teraz dodatkowe oznaczenia. Jeśli $F \in$ Hom (V, W), to będziemy pisać

$$\ker F = F^{-1}(\{0\}), \qquad \mathrm{Im}F = F(V).$$

i mówić, że $\ker F$ jest *jądrem* F a $\mathrm{Im}F$, to *obraz* funkcji F.

Zauważmy od razu, że

Stwierdzenie 27. $\dim V = \dim \ker F + \dim \mathrm{Im}F$.

Dowód. Niech B_1 będzie bazą $\ker F$ (albo zbiorem pustym jeśli $\ker F = (\{0\})$, zaś B_2 jest takim zbiorem wektorów V, że $B_1 \cup B_2$ stanowi bazę V. Wtedy oczywiście obraz F jest rozpinany przez wektory z B_2, co więcej $F(B_2)$ stanowi bazę w $\mathrm{Im}F$. □

Natychmiast wynika stąd następujący fakt.

Wniosek 28. $\dim \mathrm{Im}F \leq \dim V$.

Odnotujemy jeszcze istotne spostrzeżenie.

Stwierdzenie 29. $F \in$ Hom (V, W), wtedy różnowartościowość F jest równoważna temu, że $\dim \ker F = 0$.

Dowód. $\Rightarrow$ Jeśli F jest różnowartościowe, to jedynym wektorem v, takim, że $F(v) = 0$ jest $v = 0$.

$\Leftarrow$ Niech $F(v_1) = F(v_2)$, z liniowości dostaniemy, że $0 = F(v_1 - v_2)$, a skoro $\dim \ker F = 0$, to $\ker F = \{0\}$ i $v_1 - v_2 = 0$, tj. i F jest różnowartościowe. □

Wcześniej używaliśmy specjalnej bazy w $\mathbb{K}^r$, $\mathbf{e}_k$, $k = 1, \ldots, r$, która jest wygodna w użyciu. Przypominamy, że wektory $\mathbf{e}_k$ mają jedynkę na k-tej współrzędnej, na pozostałych zera. Tę bazę nazwiemy *bazą standardową* w $\mathbb{K}^r$. Odegra ona istotną rolę w następnej konstrukcji.

Skonstruujemy teraz ważny przykład izomorfizmu p.w. I : Hom $(\mathbb{K}^n, \mathbb{K}^m) \to M_{m\times n}(\mathbb{K})$. Jeśli F jest odwzorowaniem liniowym, to $I(F)$ będzie nazywać się macierzą F (w bazie standardowej). Kolumny macierzy $I(F)$ to wektory $F\mathbf{e}_k$ zapisane w bazie standardowej $\mathbb{K}^m$. Jeśli $I(F) = \{a_{ij}\}$ to dostaniemy, że

$$F\mathbf{e}_i = I(F)\mathbf{e}_i = \sum_{j=1}^{m} a_{ji}\mathbf{f}_j,$$

gdzie $\mathbf{f}_j$ jest bazą standardową w $\mathbb{K}^m$.

Jest rzeczą jasną, że I jest homomorfizmem. Co więcej $\ker I = 0$, bo jedynym odwzorowaniem liniowym, którego macierz jest zerowa, jest odwzorowanie zerowe. Zauważmy też, że jeśli $A \in M_{m\times n}(\mathbb{K})$, to istnieje odwzorowanie F takie, że $I(F) = A$. Mianowicie definiujemy je wzorem

$$F(\sum_{i=1}^{n} \lambda_i \mathbf{e}_i) = \sum_{j=1}^{m}\sum_{i=1}^{n} a_{ji}\lambda_i \mathbf{f}_j.$$

Zauważmy na koniec, że $I(F \circ G) = I(F) \cdot I(G)$, bo jeśli $F : \mathbb{K}^n \to \mathbb{K}^m$ oraz $G : \mathbb{K}^r \to \mathbb{K}^n$ i $\mathbf{e}_i$, $\mathbf{f}_j$, $\mathbf{g}_k$ są bazami standardowymi, to mamy że

$$(F \circ G)\mathbf{e}_i = \sum_{k=1}^{m} I(F \circ G)_{ki}\mathbf{g}_k$$

jednocześnie

$$\begin{aligned} F \circ G\mathbf{e}_i &= F(G\mathbf{e}_i) = F\sum_{j=1}^{n} I(G)_{ji}\mathbf{f}_j = \sum_{j=1}^{n} I(G)_{ji}F\mathbf{f}_j \\ = &\sum_{k=1}^{m}\sum_{j=1}^{n} I(G)_{ji}I(F)_{kj}\mathbf{g}_k = \sum_{k=1}^{m}\sum_{j=1}^{n} I(F)_{kj}I(G)_{ji}\mathbf{g}_k \end{aligned}$$

tj. $I(F \circ G)_{ki} = \sum_j I(F)_{kj}I(G)_{ji}$. □

Izomorfizm I gra ogromną rolę, dzięki niemu możemy utożsamiać odwzorowania liniowe z odpowiadającymi im macierzami w bazach standardowych. W dalszym ciągu wykładu będziemy sobie pozwalali nawet na domyślne takie utożsamienia.

Wyrazimy teraz różnowartościowość F w terminach wyznaczników.

Stwierdzenie 30. Jeśli $F \in \mathrm{Hom}\,(\mathbb{K}^n, \mathbb{K}^n)$, to

$$\ker F = \{0\} \Leftrightarrow \det I(F) \neq 0.$$

Dowód. $\Rightarrow$ Wektory $F\mathbf{e}_i$ rozpinają $\mathrm{Im}F$ i są lnz., zatem $\det I(F) \neq 0$.
$\Leftarrow$ Wektory $F\mathbf{e}_i$ rozpinają $\mathrm{Im}F$ i z założenia wynika że są lnz. Zatem $\sum_{i=1}^{n} \lambda_i F\mathbf{e}_i = 0$ pociąga, że $\lambda_i = 0$, dla $i = 1, \ldots, n$. □

2.4.1 Problemy liniowe

Chcemy zająć się zasadniczym tematem rozdziału, któremu poświęcamy niniejszy podrozdział.

Niech $F : \mathbb{K}^n \to \mathbb{K}^m$ i $b \in \mathbb{K}^m$, załóżmy, że $I(F) = A = \{a_{ij}\}_{i,j=1}^{m,n}$ tytułowe zagadnienie liniowe można sformułować na 3 sposoby

$$(I) \qquad \begin{array}{l} a_{11}x_1 + \quad \ldots \quad + a_{1n}x_n = b_1 \\ \quad \vdots \\ a_{m1}x_1 + \quad \ldots \quad + a_{mn}x_n = b_m \end{array}$$

$$(II) \qquad x_1 \begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix} + \ldots + x_n \begin{bmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$$

$$(III) \qquad Fx = b.$$

Zauważmy, że powyższe zagadnienia są równoważne. Jeśli jest oczywistym, że (I) i (II) są tym

samym. Równoważność (II) i (III) wynika stąd, że

$$Fx = F(\sum_{i=1}^{n} x_i \mathbf{e}_i) = \sum_{i=1}^{n} x_i F\mathbf{e}_i = x_1 \begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix} + \ldots + x_n \begin{bmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{bmatrix} = b$$

Możemy teraz sformułować zasadnicze wyniki. Poprzedzimy je definicją. Powiemy, że macierz współczynników $A = \{a_{ij}\}$ nazywamy *macierzą układu (I)*. Zaś *macierz rozszerzona* to

$$\begin{bmatrix} a_{11} & \ldots & a_{1n} & b_1 \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & \ldots & a_{mn} & b_n \end{bmatrix},$$

piszemy też (A, b).

Twierdzenie 31. (Kronekera - Capellego). Zagadnienia (I), (II), (III) mają rozwiązanie, wtedy i tylko wtedy, gdy rząd macierzy układu jest równy rzędowi macierzy rozszerzonej.

Dowód. Ponieważ wszystkie zagadnienia są równoważne, to zajmiemy się jednym z nich, np. (II). Rozwiązalność układu (II) oznacza, że ostatnia kolumna macierzy rozszerzonej jest kombinacją liniową kolumn macierzy układu. Jest to równoważne stwierdzeniu, że rzędy macierzy układu i rozszerzonej macierzy układu są równe. □

W szczególnym przypadku, gdy macierz układu A jest kwadratowa, możemy powiedzieć więcej.

Twierdzenie 32. (wzory Cramera) Zagadnienie (I) ma dokładnie jedno rozwiązanie wtedy i tylko wtedy, gdy $\det A \neq 0$, gdzie A jest macierzą układu (I). Nadto, jedyne rozwiązanie wyraża się wzorami

$$x_i = \frac{\det A_i}{\det A}, \quad i = 1, \ldots, n$$

gdzie A_i jest macierzą powstałą z A poprzez zastąpienie i-tej kolumny wektorem b.

Dowód. Możemy równoważnie rozpatrywać (III), tj. $Fx = b$. Jeśli (III) ma dokładnie jedno rozwiązanie, to oznacza to, że $\ker F = \{0\}$. Jest to równoważne temu, że $\mathrm{Im}F = \mathbb{K}^m$, co z kolei odpowiada temu, że $\det I(F) \neq 0$. Wyznaczymy owo rozwiązanie mnożąc lewostronnie równanie $Fx = b$ przez F^{-1}. Dostaniemy wtedy $x = F^{-1}b$. Ze wzoru na macierz odwrotną mamy, że

$$x_i = (F^{-1}b)_i = \frac{1}{\det A} \sum_{k=1}^{n} b_k A_{ki} = \frac{1}{\det A} \det A_i$$

co należało wykazać. □

Uwaga. W związku z dużą złożonością obliczeniową powyższe wzory nie są stosowane w praktyce do rozwiązywania dużych układów równań, tj. gdy $n \geq 3$.

Przeformułujemy teraz znane fakty o zbiorze rozwiązań układu (I). Oznaczymy go symbolem $R(A,b) \subset \mathbb{K}^n$. Mamy wtedy

Stwierdzenie 33. (a) $R(A,b) = \{x_0\} + R(A,0)$, gdzie x_0 jest dowolnym elementem $R(A,b)$.
(b) $\dim R(A,0) = n - \operatorname{rz} A$.

Dowód. (a) Łatwe do sprawdzenia, co zalecamy czytelnikowi. (b) Jest to przeformułowanie części (b) stwierdzenia 27. □

2.4.2 Metoda eliminacji Gaussa

Zajmiemy się teraz praktycznymi aspektami obliczania rzędu macierzy i rozwiązywania układów równań liniowych. Użyjemy do tego celu metodę eliminacji Gaussa, która ma tę zaletę, że jest ogólna i prosta w zastosowaniu. Dla naszej wygody sformułujemy ją dla macierzy rzeczywistych, co nie stanowi jednak ograniczenia. Zaczniemy od wprowadzenia pomocniczych pojęć.

Definicja 19. Niech $A \in M_{n\times m}(\mathbb{R})$.
(a) Powiemy, że $A = \{a_{ij}\}$ jest *macierzą trójkątną*, jeśli

$$a_{ij} = 0, \quad \text{gdy } i > j.$$

(b) Powiemy, że macierz trójkątna A jest *zredukowana*, jeśli $a_{ii} = 0$, to $a_{kl} = 0$ dla $k, l > i$.

Przekonamy się, że dwa podstawowe zadania obliczeniowe, a mianowicie
(a) obliczenie rzędu macierzy A;
(b) rozwiązanie układu równań $Ax = b$
bardzo się upraszczają, gdy A jest macierzą trójkątną. Zauważmy najpierw, że obliczenie wyznacznika kwadratowej trójkątnej jest proste.

Stwierdzenie 34. Załóżmy, że $A \in M_{n\times n}(\mathbb{R})$ jest macierzą trójkątną. Wtedy

$$\det A = \prod_{i=1}^{n} a_{ii}.$$

Dowód. Ten fakt wynika z kolejnego stosowania rozwinięcia Laplace'a (twierdzenie 20). □

Podobnie ma się sprawa obliczenia rzędu macierzy.

Stwierdzenie 35. Załóżmy, że $A \in M_{n\times m}(\mathbb{R})$ jest macierzą trójkątną zredukowaną. Wtedy

$$\operatorname{rz} A = \max\{j : \ a_{jj} \neq 0\}.$$

Dowód. Połóżmy, $k = \max\{j : \ a_{jj} \neq 0\}$. Z definicji macierzy trójkątnej zredukowanej wynika, że wiersze o numerach większych niż k są wypełnione zerami. Wynika stąd, że $\operatorname{rz} A \leq k$. Zdefiniujemy nową macierz $B \in M_{k\times k}(\mathbb{R})$. A mianowicie, B powstaje przez odrzucenie

kolumn o numerach $k+1, \ldots, m$ i rzędów o numerach $k+1, \ldots, n$, które są wypełnione zerami. Jest teraz oczywistym, że $\operatorname{rz} A = \operatorname{rz} B$. Zauważmy, że dzięki stwierdzeniu 34 powyżej $\det B = \prod_{i=1}^{k} a_{ii}$. Co więcej, dzięki założeniu, iż A jest macierzą trójkątną zredukowaną żadna z liczb a_{ii} nie może być zerem, zatem $\det B \neq 0$ i $\operatorname{rz} A = k$. □

Tym samym mamy odpowiedź na sformułowane wyżej pytanie (a). Zajmiemy się teraz drugą kwestią.

Twierdzenie 36. Niech $A \in M_{n\times n}(\mathbb{R})$ będzie macierzą trójkątną. Jeśli $\det A \neq 0$ i $b \in \mathbb{R}^n$, to układ $Ax = b$ ma dokładnie jedno rozwiązanie, kóre jest dane wzorami

$$x_i = (b_i - \sum_{k=i+1}^{n} a_{ik}x_k)/a_{ii}, \quad i = 1, \ldots, n. \tag{11}$$

Dowód. Z założenia, że $\det A \neq 0$ i stwierdzenia 34 wynika, że $a_{ii} \neq 0$ dla wszystkich $i = 1, \ldots, n$. Przepiszemy układ $Ax = b$ w postaci

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{n-1\,n-1}x_{n-1} + a_{1n}x_n &= b_1 \\ a_{22}x_2 + \ldots + a_{n-1\,n-1}x_{n-1} + a_{2n}x_n &= b_2 \\ \vdots &= \vdots \\ a_{n-1\,n-1}x_{n-1} + a_{n-1\,n}x_n &= b_{n-1} \\ a_n x_n &= b_n \end{aligned}$$

Ostatnie równanie jest łatwe do rozwiązania:

$$x_n = b_n/a_n.$$

Wstawiamy wynik do przedostatniego równania, dostaniemy wtedy

$$x_{n-1} = (b_{n-1} - a_{n-1\,n}x_n)/a_{n-1\,n-1}.$$

Kontynuując ten proces dostaniemy wzór (11). □

Pozostaje teraz wykazać, że każdą macierz można sprowadzić do postaci trójkątnej zredukowanej. Osiągniemy ten cel za pomocą trzech zasadniczych operacji:

(I) mnożenia wiersza przez liczbę różną od zera;

(II) dodawania dwu wierszy;

(III) zamiany wierszy i-tego i k-tego;

i jednej pomocniczej:

(IV) zamiany kolumn i-tej i k-tej.

Ostatnia operacja jest tylko przenumerowaniem zmiennych w równaniu. Zauważmy, że działania (I-III) nie zmieniają zbioru rozwiązań $R(A, b)$.

Stwierdzenie 37. Załóżmy, że dane są układy równań $Ax = b$ i $A'x = b'$, gdzie macierz rozszerzona układu (A', b') została uzyskana z macierzy rozszerzonej (A, b) za pomocą ciągu operacji (I-III). Wtedy zbiory rozwiązań obu układów są równe. Jeśli dodatkowo wykonywana była operacja (IV) na macierzy A, to aby otrzymać zbiór rozwiązań układu $A'x = b'$, koniecznym jest przenumerowanie współrzędnych odpowiadających operacjom (IV).

Dowód. Jest to łatwe ćwiczenie, które pozostawiamy Czytelnikowi. □

Możemy teraz przedstawić zasadniczy wynik, którego metoda dowodowa pozwala na praktyczne znajdowanie rozwiązań układu równań liniowych.

Twierdzenie 38. Załóżmy, że dana jest macierz $A \in M_{n\times m}(\mathbb{R})$. Wtedy istnieje ciąg operacji (I-IV) sprowadzających ją do postaci trójkątnej zredukowanej.

Dowód. Załóżmy, że $A \neq 0$. Inaczej, osiągnęliśmy cel, bo macierz $A \equiv 0$ jest trójkątna zredukowana. Załóżmy następnie, że $a_{i_0 j_0} \neq 0$ dla pewnych i_0, j_0. W razie potrzeby zamieniamy wiersz i_0-y z pierwszym (operacja (III)) i j_0-ą kolumną z pierwszą (operacja (IV)), aby dostać $a_{11} \neq 0$. Następnie mnożymy pierwszy wiersz przez $-a_{1j}/a_{11}$ dodajemy wynik do j-tego wiersza, $j = 2, \ldots, n$, (operacje (I) i (II)). Jeśli $a_{1j} = 0$, to nie wykonujemy żadnej operacji na j-tym wierszu. Nowa macierz A' ma tę cechę, że $a'_{1j} = 0$, dla $j = 2, \ldots, n$. Jeśli A' jest macierzą trójkątną zredukowaną, to kończymy pracę, jeśli nie, to powtarzamy argument rozważając tylko $i, j > 1$. □

Ciąg operacji, o którym jest mowa w twierdzeniu nie jest wyznaczony jednoznacznie.

W praktyce operacja (IV), która jest wygodna z teoretycznego punktu widzenia, prowadzi do niepotrzebnych komplikacji, na szczęście nie musi być wykonywana, ale uzyskana macierz zredukowana A' nie ma już tak eleganckiej struktury. Szczegóły pozostawiamy Czytelnikowi jako ćwiczenie rachunkowe.

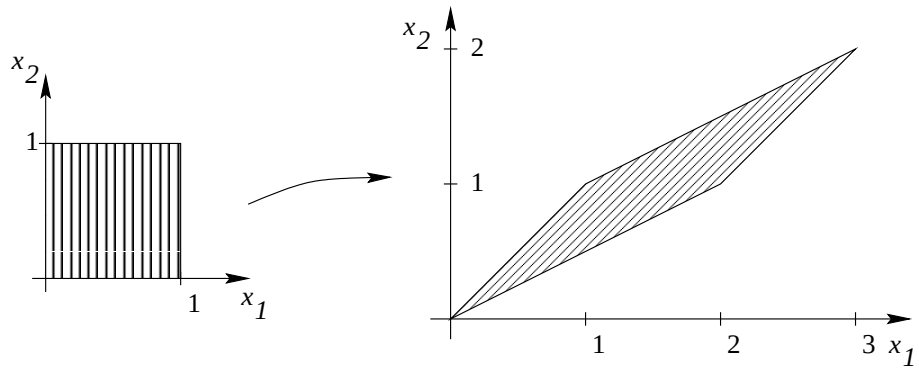
Rozwiązując układ $Ax = b$ operacje opisane w dowodzie twierdzenia 38 wykonujemy nie tylko na A, ale i na macierzy rozszerzonej układu (A, b). Robimy tak, po to aby znaleźć się w sytuacji opisanej w stwierdzeniu 37.

2.5 Interpretacja i zastosowania geometryczne wyznaczników

Będziemy rozważali całą serię przykładów rachunkowych. Będą one dotyczyły głównie płaszczyzny.

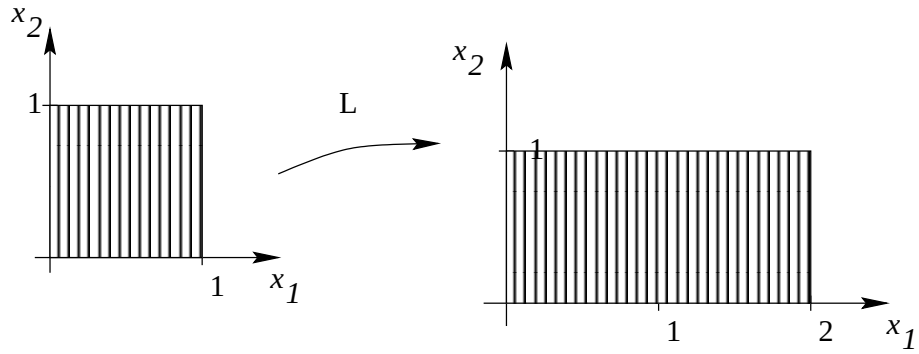
2.5.1 Przykłady przekształceń płaszczyzny

Przykład 8. Przekształcenie liniowe $F : \mathbb{R}^2 \to \mathbb{R}^2$ zadajemy podając jego macierz $\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$, tj. $F\mathbf{e}_1 = 2\mathbf{e}_1 + \mathbf{e}_2$ i $F\mathbf{e}_2 = \mathbf{e}_1 + \mathbf{e}_2$. Wtedy obrazem kwadratu jednostkowego jest równoległobok.



Rys. 1. Obraz kwadratu jednostkowego.

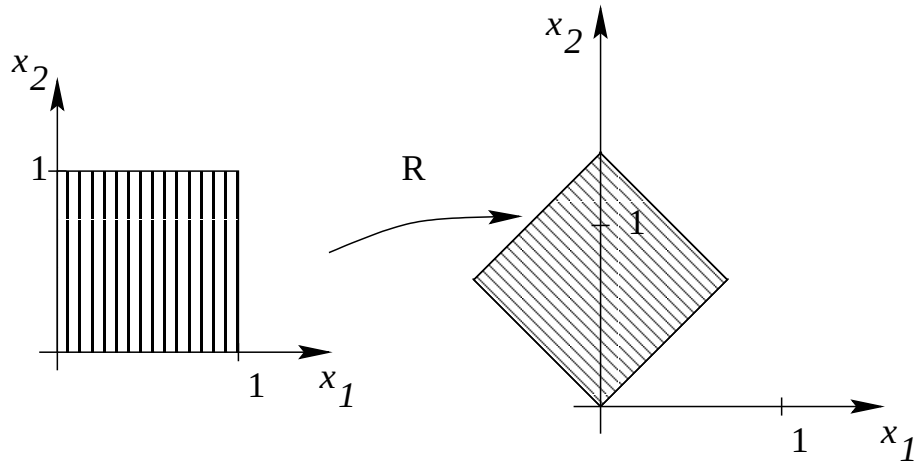
Przykład 9. Określimy $L, R, Q : \mathbb{R}^2 \to \mathbb{R}^2$, podając ich macierze. L ma macierz $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$. Wynika stąd $L\mathbf{e}_1 = 2\mathbf{e}_1$ i $L\mathbf{e}_2 = \mathbf{e}_2$, a zatem L rozciąga w kierunku $\mathbf{e}_1$ o czynnik 2,



Rys. 2. Wynik działania L na kwadrat jednostkowy.

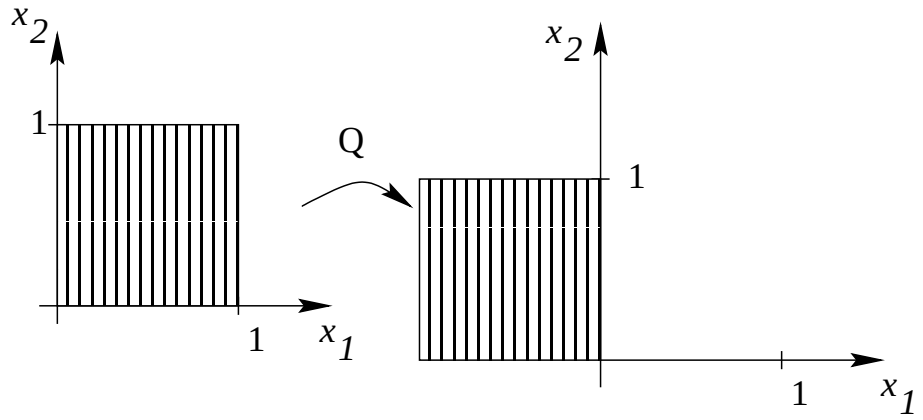
zaś L^{-1} ściska w tym kierunku o czynnik $0{,}5$.

$R = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$ i rysunek pokazuje, że R jest obrotem o 45^o:



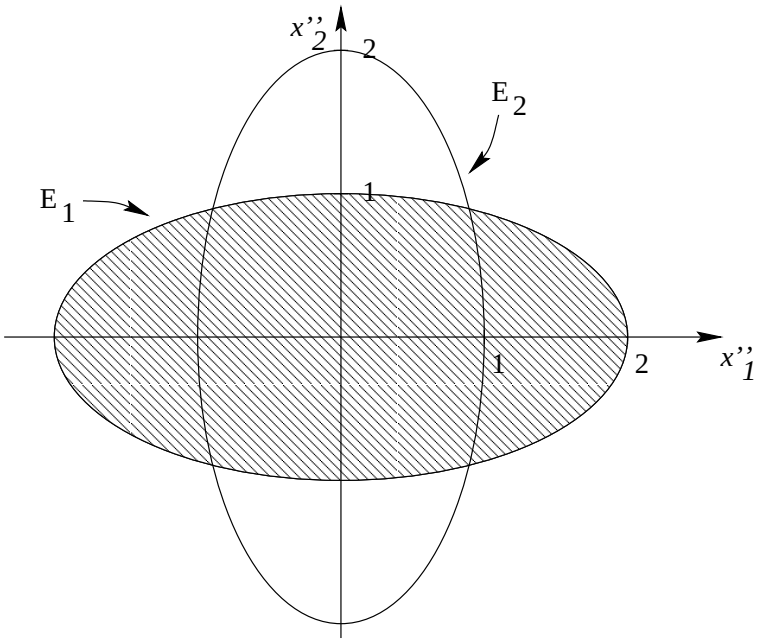
Rys. 3. Wynik działania R na kwadrat jednostkowy.

$Q = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, widać, że Q jest obrotem o 90^o:



Rys. 4. Wynik działania Q na kwadrat jednostkowy.

Zastanówmy się co się dzieje z okręgiem $S = \{(x_1, x_2) \in \mathbb{R}^2; x_1^2 + x_2^2 = 1\}$ pod działaniem LQ i QL. Z rysunków wynika, że $LQ(S) = E_1$ i $QL(S) = E_2$, gdzie E_1, E_2 są elipsami jak niżej.



Rys. 5. Elipsy E_1 i E_2.

Aby znaleźć równania na $LQ(S)$ i $QL(S)$ postępujemy następująco: wyznaczamy $\binom{x_1}{x_2}$ z równania $L\binom{x_1}{x_2} = \binom{x_1^1}{x_2^1}$ tj. $\binom{x_1}{x_2} = L^{-1}\binom{x_1^1}{x_2^1}$, podobnie rozwiązujemy $\binom{x_1^{11}}{x_2^{11}} = Q\binom{x_1^1}{x_2^1}$ tj. $\binom{x_1^1}{x_2^1} = Q^{-1}\binom{x_1^{11}}{x_2^{11}}$.

Po przeprowadzeniu powyższych rachunków i wstawieniu wyników do równania okręgu $x_1^2 + x_2^2 = 1$ dostaniemy, że zbiory $LQ(S) = E_1$, $QL(S) = E_2$ są opisywane odpowiednio

równaniami

$$\frac{1}{4}(x_1^{11})^2 + (x_2^{11})^2 = 1 \quad \text{i} \quad (x_1^{11})^2 + \frac{(x_2^{11})^2}{4} = 1.$$

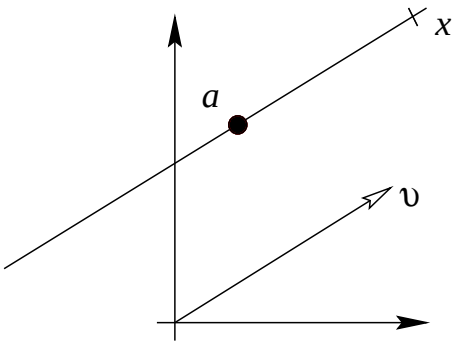
Przy wykonywaniu wskazanych wyżej obliczeń zauważamy, że

$$R^{-1} = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} \text{ i } Q^{-1} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

co więcej $R^{-1} = R^T$ i $Q^{-1} = Q^T$. Macierze A o wyrazach rzeczywistych takie, że $A^T = A^{-1}$ nazywamy *ortogonalnymi*.

2.5.2 Prosta na płaszczyźnie

Prostą na płaszczyźnie przechodzącą przez punkt a i równoległą do wektora swobodnego v nazywamy zbiór punktów $x \in \mathbb{R}^2$, takich że wektor związany $\vec{ax}$ jest równoległy do wektora $\vec{0v}$



Rys. 6.

tj.

$$x = a + tv, \qquad t \in \mathbb{R}. \tag{12}$$

We współrzędnych dostaniemy:

$$x_1 = a_1 + tv_1$$
$$x_2 = a_2 + tv_2$$

równoważnie,

$$x_1 - a_1 = tv_1$$
$$x_2 - a_2 = tv_2.$$

Załóżmy, że $v_1 \neq 0$ i $v_2 \neq 0$. (Czytelnik jest proszony o samodzielne zbadanie przypadku $v_1 = 0$ lub $v_2 = 0$). Wtedy ostatni układ równań jest równoważny pojedynczemu równaniu

$$v_2(x_1 - a_1) - (x_2 - a_2)v_1 = 0. \tag{13}$$

Zauważmy, że (13) kojarzy się z wyznacznikiem, a mianowicie lewa strona tej równości może być przepisana jako

$$\det \begin{bmatrix} x_1 - a_1 & v_1 \\ x_2 - a_2 & v_2 \end{bmatrix} = 0. \tag{14}$$

Skoro wyznacznik wektorów jest równy zero, to znaczy, że wektory $x - a$ i v są liniowo zależne. Jeśli wektor v jest wyznaczony przez wektor związany $\vec{ab}$, to dostaniemy z (14), że

$$\det \begin{bmatrix} x_1 - a_1 & b_1 - a_1 \\ x_2 - a_2 & b_2 - a_2 \end{bmatrix} = 0. \tag{15}$$

Jest to równanie prostej przechodzącej przez punkty a i b o współrzędnych $a = (a_1, a_2), b = (b_1, b_2)$.

Zauważmy, że macierz w (15) wygląda tak, jakby od kolumn pierwszej i drugiej odjęto wektor $\binom{a_1}{a_2}$, który możemy uważać za trzecią kolumnę. Nie ma w tym nic dziwnego, bo dzięki rozwinięciu Laplace'a, widzimy, że

$$0 = \det \begin{bmatrix} x_1 - a_1 & b_1 - a_1 \\ x_2 - a_2 & b_2 - a_2 \end{bmatrix} = \det \begin{bmatrix} x_1 - a_1 & b_1 - a_1 & a_1 \\ x_2 - a_2 & b_2 - a_2 & a_2 \\ 0 & 0 & 1 \end{bmatrix} = \det \begin{bmatrix} x_1 & b_1 & a_1 \\ x_2 & b_2 & a_2 \\ 1 & 1 & 1 \end{bmatrix} = 0 \tag{16}$$

Przedstawiliśmy kolejną postać równania prostej przechodzącej przez zadane punkty. Zauważmy też, że (16) jest warunkiem koniecznym i dostatecznym współliniowości 3 punktów $x = (x_1, x_2)$, $a = (a_1, a_2)$, $b = (b_1, b_2)$.

Zadajmy teraz pytanie: czym dla dowolnych punktów płaszczyzny x, a, b jest liczba

$$\det \begin{bmatrix} x_1 - a_1 & b_1 - a_1 \\ x_2 - a_2 & b_2 - a_2 \end{bmatrix} ?$$

Zdefiniujmy *długość wektora* $v \in \mathbb{R}^n$ wzorem

$$|v| := \sqrt{\sum_{i=1}^{n} v_i^2} \tag{17}$$

tj. gdy $n = 2$ dostaniemy $|v| = \sqrt{v_1^2 + v_2^2}$. Niech φ będzie miarą kąta jaki tworzy oś $0x_1$ z wektorem $v \in \mathbb{R}^2$, wtedy

$$v_1 = |v| \cos \varphi, \quad v_2 = |v| \sin \varphi \tag{18}$$

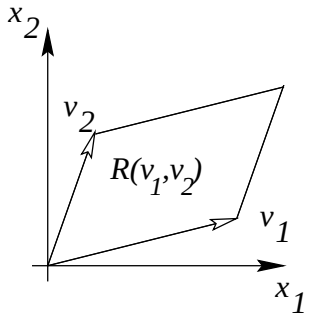
Załóżmy, że mamy dwa wektory na płaszczyźnie $v_1 = (v_1^1, v_2^1)$ i $v_2 = (v_1^2, v_2^2)$. Sprawdzimy czym jest $\det[v_1, v_2] \equiv \det \begin{bmatrix} v_1^1 & v_1^2 \\ v_2^1 & v_2^2 \end{bmatrix} =: D$. Wykorzystując wzory (18) dostaniemy, że

$$\begin{aligned} D &= |v_1||v_2|(\cos\varphi_1 \ \sin\varphi_2 - \cos\varphi_2 \ \sin\varphi_1) = \\ &= |v_1||v_2| \sin(\varphi_2 - \varphi_1) = |v_1||v_2| \sin\alpha, \end{aligned}$$

gdzie α jest kątem pomiędzy v_1 i v_2, tj. $\alpha = \varphi_1 - \varphi_2$. Jeśli zdefiniujemy równoległobok $R(v_1, v_2)$ jako zbiór

$$R(v_1, v_2) = \{x \in \mathbb{R}^2 : \ x = v_1 t_1 + v_2 t_2; \quad t_1, t_2 \in [0, 1]\},$$

(patrz rysunek 7), to wtedy $|D|$ jest jego polem.



Rys. 7.

Dzięki tej uwadze zauważamy, że pole obrazu kwadratu $F([0, 1]^2)$, gdzie F jest jak w przykładzie 8 jest równe 1, bo $\det \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} = 1$.

Chcemy teraz podać jeszcze jedną interpretację wzoru (13). Do tego celu potrzebujemy definicji:

Definicja 20. Niech $v, w \in \mathbb{R}^n$. *Iloczynem skalarnym* wektorów v, w nazywamy liczbę

$$(v, w) := v^T \cdot w$$

gdzie po prawej stronie wektory v, w traktujemy jako macierze $1 \times n$ zaś $\cdot$ oznacza mnożenie macierzy, jeśli $v = (v_1, \ldots, v_n)$, $w = (w_1, \ldots, w_n)$, to

$$(v, w) = \sum_{i=1}^{n} v_i w_i.$$

Z tego wzoru wynika, że $|v| = \sqrt{(v, v)}$.

Z drugiej strony dla $v, w \in \mathbb{R}^2$, rachunki takie jak przy wyznaczaniu D, prowadzą do wniosku, że

$$(v, w) = |v| \cdot |w| \cos \alpha,$$

gdzie α jest kątem pomiędzy v i w.

Spójrzmy z jeszcze jednej strony na (13); jeśli wprowadzimy wektor $w = (v_2, -v_1)$ to (13) jest równoważne

$$(x - a, w) = 0$$

dla punktów x z prostej.

Jeśli umówimy się nazywać wektory v i w takie, że $(w, v) = 0$ *prostopadłymi*, to dostaniemy, że (13) oznacza prostopadłość wektorów $x - a$ i w. Albo inaczej: prosta przechodząca przez

punkt a, to zbiór punktów x płaszczyzny takich, że $x - a$ wektor jest prostopadły do zadanego wektora w.

Zwróćmy uwagę, że prostą opisywaliśmy analitycznie na 2 sposoby:

(a) parametrycznie: patrz równanie (12);

(b) jako poziomicę funkcji $f : \mathbb{R}^2 \to \mathbb{R}$, tj. przeciwobraz liczby.

W naszym przypadku $f(x) = (x - a, w)$ i prostą jest zbiór $f^{-1}(\{0\})$.

Każdy ze sposobów ma swoje wady i zalety. Zauważmy, że okrąg łatwo opisuje się równaniem np. $x_1^2 + x_2^2 = 1$, ale przedstawienie **całego** okręgu parametrycznie nie jest możliwe.

2.5.3 Prosta i płaszczyna w $\mathbb{R}^3$

Będziemy się teraz zajmować przykładami głównie w $\mathbb{R}^3$. Zauważmy, że jeśli dwie płaszczyzny V_1, V_2 w $\mathbb{R}^3$, nie pokrywające się, mają punkt wspólny, to przecinają się wzdłuż prostej. Jest to wniosek z następującego rachunku.

$$3 = \dim \mathbb{R}^3 = \dim (V_1 + V_2) = \dim V_1 + \dim V_2 - \dim V_1 \cap V_2 = 2 + 2 - \dim V_1 \cap V_2$$

zatem $\dim V_1 \cap V_2 = 1$. Wynika stąd, że jeśli chcemy opisać prostą w $\mathbb{R}^3$, to musimy umieć opisać płaszczyznę w $\mathbb{R}^3$. Jak to zrobić? Chcemy, by płaszczyzna była rozpinana przez liniowo niezależne wektory v i w i przechodziła przez punkt a, zatem punkt x płaszczyny musi mieć tę właściwość, że wektor $x - a$ jest kombinacją liniową v i w, tj.

$$x - a = tv + sw \tag{19}$$

albo we współrzędnych

$$\begin{aligned} x_1 &= a_1 + tv_1 + sw_1 \\ x_2 &= a_2 + tv_2 + sw_2 \\ x_3 &= a_3 + tv_3 + sw_3. \end{aligned}$$

Jednak zapis (19) jest wygodniejszy, bo automatycznie dostaniemy, że

$$\det \begin{bmatrix} x_1 - a_1 & v_1 & w_1 \\ x_2 - a_2 & v_2 & w_2 \\ x_3 - a_3 & v_3 & w_3 \end{bmatrix} = 0 \tag{20}$$

a jeśli wektor v jest wektorem swobodnym wyznaczonym przez wektor związany $\vec{ab}$ i podobnie w jest wyznaczony przez $\vec{ac}$, to dostaniemy:

$$\det \begin{bmatrix} x_1 - a_1 & b_1 - a_1 & c_1 - a_1 \\ x_2 - a_2 & b_2 - a_2 & c_2 - a_2 \\ x_3 - a_3 & b_2 - a_3 & c_3 - a_3 \end{bmatrix} = 0.$$

Jest to analitycznie zapisany warunek współpłaszczyznowości 4 punktów $(x_1, x_2, x_3), (a_1, a_2, a_3), (b_1, b_2, b_3), (c_1, c_2, c_3)$.

Wprowadzimy dodatkowe oznaczenie: $y = x-a$. Wtedy korzystając z rozwinięcia Laplace'a wyznacznika macierzy 3×3 we wzorze (20) dostaniemy:

$$0 = \det(y, v, w) = y_1 \ \det \begin{bmatrix} v_2 & w_2 \\ v_3 & w_3 \end{bmatrix} - y_2 \ \det \begin{bmatrix} v_1 & w_1 \\ v_3 & w_3 \end{bmatrix} + y_3 \ \det \begin{bmatrix} v_1 & w_1 \\ v_2 & w_2 \end{bmatrix}$$

prawą stronę możemy zinterpretować jako iloczyn skalarny wektora y z pewnym wektorem. Ów wektor zależy od v i w; oznaczamy go symbolem

$$v \times w := \left(\det \begin{bmatrix} v_2 & w_2 \\ v_3 & w_3 \end{bmatrix}, -\det \begin{bmatrix} v_1 & w_1 \\ v_3 & w_3 \end{bmatrix}, \det \begin{bmatrix} v_1 & w_1 \\ v_2 & w_2 \end{bmatrix} \right)$$

i nazywamy *iloczynem wektorowym* wektorów v i w. Z definicji natychmiast dostaniemy,

$$\det(y, v, w) = (y, v \times w) \tag{21}$$

i równanie płaszczyzny rozpinanej przez v i w, przechodzącej przez punkt a przyjmie postać

$$(x - a, v \times w) = 0$$

tj. jest przeciwobrazem zera funkcji, $F : \mathbb{R}^3 \to \mathbb{R}$, danej wzorem $F(x) = (x - a, v \times w)$. Pamiętając, że prosta jest przecięciem dwu płaszczyzn można ją opisać układem dwu równań

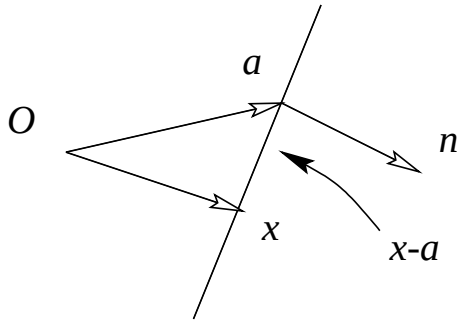
$$\begin{cases} ((x-a), \ \ v_1 \times w_1) = 0 \\ ((x-a), \ \ v_2 \times w_2) = 0 \end{cases} \tag{22}$$

Dostaniemy wtedy prostą przechodzącą przez punkt a. Pamiętać przy tym trzeba, by $\mathbb{R}^3 = \text{span}\{v_1, v_2, w_1, w_2\}$, bo tylko wtedy rozważania z początku wykładu będą miały zastosowanie.

Układ (22) można skomentować jeszcze inaczej. Mianowicie, że prosta w $\mathbb{R}^3$ jest przeciwobrazem punktu $(0, 0)$ odwzorowania $G : \mathbb{R}^3 \to \mathbb{R}^2$ danego wzorem

$$G(x) = \begin{pmatrix} (x - a_1, v_1 \times w_1) \\ (x - a_1, v_2 \times w_2) \end{pmatrix}$$

W praktyce, płaszczyznę w $\mathbb{R}^3$ łatwiej jest opisać zadając wektor $\mathbf{e}$ do niej prostopadły tj. taki, że $(\mathbf{e}, x - a) = 0$, patrz rysunek poniżej



Rys. 8.

2.5.4 Właściwości iloczynu wektorowego

Z definicji iloczynu wektorowego łatwo się przekonać, że jeśli wektory $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ stanowią bazę standardową w $\mathbb{R}^3$, to

$$(a) \qquad \mathbf{e}_1 \times \mathbf{e}_2 = \mathbf{e}_3, \quad \mathbf{e}_2 \times \mathbf{e}_3 = \mathbf{e}_1, \quad \mathbf{e}_3 \times \mathbf{e}_1 = \mathbf{e}_2.$$

Ponadto,

$$(b) \qquad u \times w = -w \times u,$$

u i w są prostopadłe do $w \times u$, tj.

$$(c) \qquad (u, w \times u) = 0 = (w, w \times u).$$

Chcielibyśmy teraz policzyć długość wektora $u \times w$. Oznaczmy przez A_i i-tą współrzędną wektora z definicji $u \times w$. Z naszej wiedzy o wyznacznikach 2×2 wynika, że $|A_i|$ jest równe polu powierzchni rzutu równoległoboku $R(u, w)$ na płaszczyznę rozpiętą przez wektory $\mathbf{e}_j, \mathbf{e}_k$, $j, k \neq i$. Z geometrycznych rozważań wynika, że $|A_i|$ = pole $R(u, w) \cdot \cos \alpha_i$, gdzie α_i jest kątem jaki tworzy wektor prostopadły do $R(u, w)$ z osią $0x_i$. (Proszę samemu zrobić odpowiedni rysunek). Dostaniemy zatem

$$|u \times w| = \sqrt{A_1^2 + A_2^2 + A_3^2} = \text{ pole } R(u, w) \sqrt{\sum_{i=1}^{3} \cos^2 \alpha_i},$$

ale wiemy, że wektorem prostopadłym do $R(u, w)$ jest $u \times w$.

Wykazaliśmy, że w przypadku wektorów na płaszczyźnie a i b ich iloczyn skalarny (a, b) wyraża się wzorem

$$(a, b) = |a| \cdot |b| \cos \varphi, \tag{23}$$

gdzie φ jest miarą kąta pomiędzy a i b.

Ten wzór jest prawdziwy także dla dwóch wektorów w $\mathbb{R}^3$, bo rozpinają one płaszczyznę, gdzie (23) jest spełniony. (Pomijamy bardziej ścisłe uzasadnienie wzoru (23)). Tym samym możemy dokończyć rachunki, mamy bowiem, że

$$\cos \alpha_i = \frac{(u \times w, \mathbf{e}_i)}{|u \times w|}$$

a stąd

$$\sum_{i=1}^{3} \cos^2 \alpha_i = \sum_{i=1}^{3} \frac{(u \times w, \mathbf{e}_i)^2}{|u \times w|^2} = \frac{\sum_{i=1}^{3} A_i^2}{\sum_{i=1}^{3} A_i^2} = 1.$$

Inny ogólny dowód tego faktu będzie podany w rozdziale 8, w paragrafie poświęconym szeregom Fouriera, (patrz tożsamość (8.2)).

Możemy zatem napisać

$$|u \times w| = \text{ pole } R(u, w). \tag{24}$$

Pozostaje nam teraz ustalić czym jest

$$\det(u, v, w).$$

Na mocy wzoru (21) dostaniemy

$$|\det(u, v, w)| = |(u, v \times w)| = D,$$

to jest

$$D = |u| \cdot |v \times w||\cos\varphi|$$

gdzie φ jest kątem pomiędzy u i $v \times w$, tj. wektorem prostopadłym do $R(v, w)$. Zatem ze wzoru (24) dostaniemy

$$D = \text{ pole } R(v, w) \cdot h$$

gdzie h jest wysokością równoległościanu

$$R(u, v, w) = \{x \in \mathbb{R}^3; x = su + tv + \tau w, \quad s, t, \tau \in [0, 1]\}$$

spuszczoną na płaszczyznę span (v, w), tj.

$$|\det(u, v, w)| = \text{ objętość } R(u, v, w)$$

W przestrzeniach $\mathbb{R}^n$, $n > 3$ gdzie nasza intuicja może nas opuścić, musimy **definiować** *równoległościan* jako zbiór

$$R(u_1, \ldots, u_n) = \{x \in \mathbb{R}^n : \sum_{i=1}^{n} t_i u_i, \quad t_i \in [0, 1]\}$$

a jego *objętość* w następujący sposób

$$\text{vol } R(u_1, \ldots, u_n) = |\det(u_1, \ldots, u_n)|.$$

2.5.5 Stożkowe

Przypominamy, że *elipsa*, tj. zbiór punktów, których suma odległości od ognisk jest równa $2a$ jest opisywana równaniem

$$\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} = 1,$$

gdzie $b^2 = a^2 - c^2$ i $2c$ jest odległością ognisk.

Podobnie *hiperbola*, tj. zbiór punktów, których różnica odległości od ognisk jest równa $2a$, jest opisywana równaniem

$$\frac{x_1^2}{a^2} - \frac{x_2^2}{b^2} = 1,$$

gdzie $b^2 = a^2 + c^2$ i $2c$ jest jak wyżej.

Parabola jest opisywana równaniem $x_2 = x_1^2$.

Wiemy już jak znaleźć równanie elipsy (hiperboli i paraboli) w nowych współrzędnych $x^1 = (x_1^1, x_2^1)$, gdzie nowe x^1 i stare współrzędne $x = (x_1, x_2)$ są powiązane przekształceniem liniowym $x^1 = Lx$. Trzeba wtedy znaleźć $x = L^{-1}x^1$ i wstawić do równań. W ogólności dostaniemy wtedy

$$ax_1^2 + 2bx_1x_2 + cx_2^2 + Ax_1 + Bx_2 = C \tag{25}$$

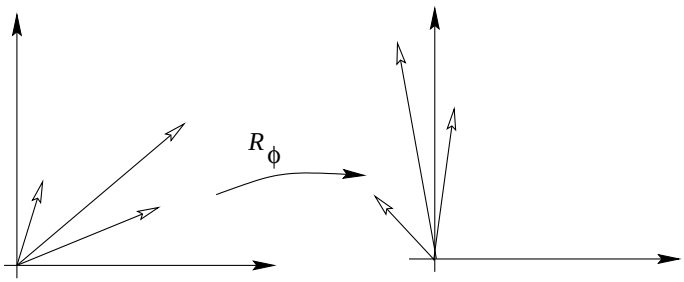
Powstaje przy tym naturalne pytanie odwrotne. Czym jest zbiór rozwiązań (x_1, x_2) równania (25)? tj. czy jest możliwym, by istniały inne rozwiązania niż te do tej pory poznane tj. elipsa, parabola, hiperbola, prosta? Odpowiedź jest twierdząca, np. stożek, czyli na płaszczyźnie para prostych, jest opisywany postaci jak wyżej,

$$(x_1 - a_1)(x_2 - a_2) = 0$$

przechodzących przez punkt (a_1, a_2). W drugiej części skryptu poznamy sposoby badania równań, takich jak (25).

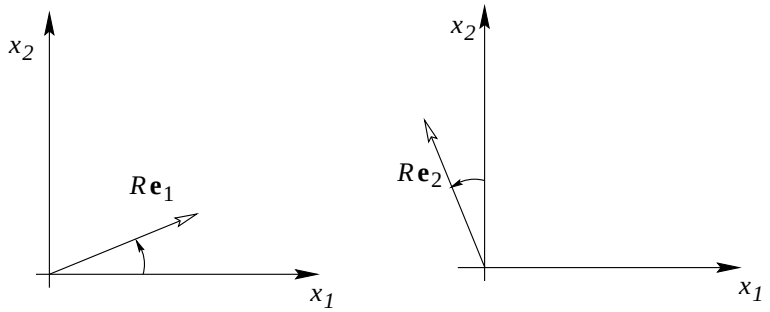
2.5.6 Obrót

Na koniec rozdziału znajdziemy macierz obrotu o kąt φ. Łatwo widać z rysunku, że obrót płaszczyzny wokół początku układu współrzędnych jest przekształceniem liniowym.



Rys. 9. Obrót o kąt ϕ.

Aby wyznaczyć macierz obrotu R_φ wystarczy znaleźć $R_\varphi \mathbf{e}_1$ i $R_\varphi \mathbf{e}_2$, będą one pierwszą i drugą kolumną macierzy. Z rysunków poniżej



Rys. 10.

wynika, że

$$R_\varphi \mathbf{e}_1 = \begin{pmatrix} \cos\varphi \\ \sin\varphi \end{pmatrix} \qquad R_\varphi \mathbf{e}_2 = \begin{pmatrix} -\sin\varphi \\ \cos\varphi \end{pmatrix},$$

zatem

$$R_\varphi = \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix}.$$

Rozdział 3

Rachunek różniczkowy i całkowy jednej zmiennej

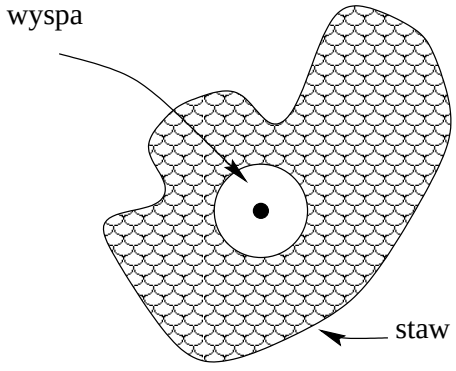
Jest to pierwszy typowo analityczny rozdział. Do uprawiania analizy będą nam potrzebne tylko dwa ciała liczbowe: $\mathbb{R}$ i $\mathbb{C}$. Pewne fakty dopuszczają wspólne sformułowania, wtedy będziemy pisali $\mathbb{K}$ na oznaczenia $\mathbb{R}$ lub $\mathbb{C}$. Wprawdzie analiza zmiennej rzeczywistej jest bardziej przejrzysta, to mając na uwadze późniejsze zastosowania w §3.8 będziemy eksponowali aspekt ogólny wspólny dla $\mathbb{R}$ i $\mathbb{C}$ tam, gdzie konieczne.

Naszym celem jest badanie ciągłości i różniczkowalności funkcji jednej zmiennej i pokażemy zastosowania do obliczeń przybliżonych i znajdowania wartości największej i najmniejszej funkcji w zbiorze. Potem zajmiemy się liczeniem pola pod wykresem, do czego będzie potrzebna całka Riemanna. Na koniec zajmiemy się szeregami potęgowymi, które pozwolą nam na ścisłe zdefiniowanie funkcji elementarnych takich, jak e^x, $\sin x$, $\cos x$.

Zaczniemy od pojęcia granicy.

3.1 Ciąg i jego granica

Rozpatrzmy dwa przykłady.
1. Punktowa żaba siedzi w środku kolistej wyspy na stawie.



Rys. 1. Wyspa z żabą na stawie.

Promień wyspy wynosi $\frac{3}{2}$ jednostek długości. W pewnej chwili żaba wykonuje skoki wzdłuż promienia koła w kierunku wody, pierwszy skok ma długość 1. Każdy następny skok żabę męczy, więc następny skok ma zawsze długość równą $\frac{1}{3}$ skoku poprzedniego. Kiedy żaba dotrze do wody ? Policzymy: po n skokach żaba przebyła

$$1 + \frac{1}{3} + \frac{1}{3^2} + ... + \frac{1}{3^n} = S_n.$$

Aby policzyć S_n zauważmy, że

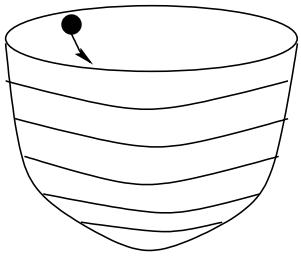
$$(1 + x + x^2 + ... + x^n)(1 - x) = 1 + x + x^2 + ... + x^n - x - x^2 - ... - x^{n+1} = 1 - x^{n+1},$$

zatem

$$S_n = \frac{1 - \frac{1}{3^{n+1}}}{1 - \frac{1}{3}} = \frac{3}{2}\left(1 - \frac{1}{3^{n+1}}\right).$$

Widać więc, że żaba nigdy nie dojdzie do wody, ale w **granicy** osiągnie brzeg wyspy, bo $\frac{1}{3^{n+1}}$ jest coraz to mniejsze.

2. Wyobraźmy sobie stalową misę, do której wrzuciliśmy szklaną kulę.



Rys. 2. Kulka w misie.

Kula będzie się przez jakiś czas toczyć kreśląc ładne wzory na powierzchni, aż wreszcie tarcie ją wyhamuje i jej ruch przestanie być widoczny. W praktyce zatrzyma się, bo nie obserwujemy w makroskopowej skali ruchu o podatomowej wielkości. Ale teoretycznie dopiero w **granicy**, po nieskończonym czasie kulka osiągnie spoczynek.

Daliśmy tym do zrozumienia, co będzie przedmiotem naszego zainteresowania, na początek będą to granice ciągów. Do tego potrzebne jest nam uściślenie pojęcia odległości, chcemy przy tym by nowe pojęcie było dostatecznie pojemne, tj. obejmowało odległości punktów w $\mathbb{C}^n$. W tym celu znane pojęcie długości wektora $x \in \mathbb{R}^n$ (patrz wzór 2.17) rozszerzymy w naturalny sposób na $z \in \mathbb{C}^n$.

Definicja 1. Jeśli $z \in \mathbb{C}^n$, tj. $z = (z_1, \ldots, z_n)^T$, to możemy utożsamiać z z macierzą o jednej kolumnie i n wierszach. Liczbę rzeczywistą

$$|z| := \sqrt{z^T \cdot \bar{z}} \quad \equiv \quad \sqrt{\left(\sum_{i=1}^{n} |z_i|^2\right)},$$

gdzie kropka oznacza mnożenie macierzowe, nazywamy *długością wektora* z (albo inaczej *normą wektora* z).

Pragniemy podkreślić, że w §3.8 będziemy istotnie wykorzystywali zbieżność w $\mathbb{C}$. Przypominamy, że $\mathbb{C} = \mathbb{R} \times \mathbb{R}$ i długość wektora $z = x + iy \in \mathbb{C}$ jest w istocie długością wektora $(x, y) = \mathbb{R}^2$. Dlatego zgrabniej jest od razu przedstawić ogólną definicję normy wektora w $\mathbb{R}^n$.

Wykażemy teraz

Stwierdzenie 1. (nierówność trójkąta). Jeśli $x, y \in \mathbb{K}^n$, to

$$|x + y| \leq |x| + |y|.$$

Dowód. Liczymy

$$\begin{aligned} |x+y|^2 &= \sum_{i=1}^{n} |x_i + y_i|^2 = \sum_{i=1}^{n} (|x_i|^2 + x_i\bar{y}_i + y_i\bar{x}_i + |y_i|^2) \\ &= |x|^2 + |y|^2 + 2\mathrm{Re}\sum_{i=1}^{n} x_i\bar{y}_i \end{aligned}$$

Uprzednio (patrz twierdzenie 1.26) wykazaliśmy nierówność Schwarza. Dzięki niej dostaniemy

$$\left|\mathrm{Re}\sum_{i=1}^{n} x_i\bar{y}_i\right| \leq |x||y|.$$

A skoro $|\mathrm{Re}w| \leq |w|$, to otrzymamy

$$|x + y|^2 \leq |x|^2 + |y|^2 + 2|x| \cdot |y| = (|x| + |y|)^2$$

Co należało wykazać. □

Wprowadzimy teraz odległość punktów opierając się na długości wektora.

Definicja 2. *Odległością punktów* $a, b \in \mathbb{K}^n$ nazwiemy liczbę $d(a, b) := |a - b|$.
Warto podkreślić, że z definicji natychmiast wynika, iż

$$d(x, y) = d(y, x).$$

Natychmiast też dostaniemy następujący wniosek, znowu nazywany *nierównością trójkąta*.

Wniosek 2. Dla dowolnych $x, y, z \in \mathbb{K}^n$ mamy, że

$$d(x, y) \leq d(x, z) + d(z, y)$$

Dowód.

$$d(x, y) = |x - y| = |(x - z) + (z - y)| \leq |x - z| + |z - y| = d(x, z) + d(z, y).$$ □

Jesteśmy teraz gotowi wypowiedzieć definicję granicy ciągu punktów w $\mathbb{K}^k$.

Definicja 3. Powiemy, że wektor $g \in \mathbb{K}^k$ ($\mathbb{K} = \mathbb{R}$ lub $\mathbb{K} = \mathbb{C}$) jest *granicą* ciągu $\{a_n\}_{n=1}^{\infty}$ punktów z $\mathbb{K}^k$, jeśli dla dowolnego $\varepsilon > 0$ istnieje liczba naturalna N_ε, taka że dla dowolnej liczby naturalnej n spełniającej $n > N_\varepsilon$ jest prawdą, że

$$d(a_n, g) < \varepsilon.$$

Piszemy wtedy

$$\lim_{n\to\infty} a_n = g.$$

Innymi słowy: w dowolnie małym otoczeniu punktu g znajdują się wszystkie, z wyjątkiem skończenie wielu, wyrazy ciągu. Zamiennie będziemy też pisali $a_n \to g$ na oznaczenie faktu iż g jest granicą ciągu $\{a_n\}_{n=1}^{\infty}$.

Wprawdzie mówiliśmy wyżej o granicy ciągu elementów $\mathbb{R}^k$ (lub $\mathbb{C}^k$), ale w istocie wystarczy badać ciągi liczbowe. Mamy bowiem,

Stwierdzenie 3. Załóżmy, że $a_n = (a_n^1, \ldots, a_n^k)$, $n = 1, \ldots$ jest ciągiem elementów $\mathbb{R}^k$. Ciąg $\{a_n\}_{n=1}^{\infty}$ jest zbieżny do $g = (g^1, \ldots, g^k)$, wtedy i tylko wtedy, gdy każdy ciąg $\{a_n^i\}_{n=1}^{\infty}$ jest zbieżny do g^i, $i = 1, \ldots, k$.

Dowód. $\Rightarrow$ Z definicji, dla dowolnego $\varepsilon > 0$ istnieje N_ε, takie że dla $n > N_\varepsilon$ jest prawdą, że

$$\varepsilon > d(a_n, g) = |a_n - g| \geq |a_n^i - g^i|$$

dla wszystkich $i = 1, \ldots k$, tj.

$$\lim_{n\to\infty} a_n^i = g^i. \tag{1}$$

$\Leftarrow$ Z drugiej strony, jeśli zachodzi (1), to dla dowolnego $\varepsilon > 0$ istnieją N_ε^i, takie że dla $n > N_\varepsilon^i$, $|a_n^i - g^i| < \frac{\varepsilon}{\sqrt{k}}$ dla $i = 1, \ldots, k$. Połóżmy $N_\varepsilon = \max_i N_\varepsilon^i$. Wtedy

$$d(a_n, g) = \sqrt{\sum_{i=1}^{k} |a_n^i - g^i|^2} \leq \sqrt{\sum_{i=1}^{k} \frac{\varepsilon^2}{k}} = \sqrt{\frac{k\varepsilon^2}{k}} = \varepsilon$$

dla $n > N_\varepsilon$. Zatem $\lim_{n\to\infty} a_n = g$. □

Od tej chwili w bieżącym rozdziale będziemy zajmowali się głównie ciągami liczbowymi. Nim rozpatrzymy serię przykładów wprowadzimy pomocne oznaczenia. Jeśli $x \in \mathbb{R}$, to piszemy

$$[x] = \max\{n \in \mathbb{Z} : \ n \leq x\}$$

i mówimy, że $[x]$ jest *częścią całkowitą* liczby x.

Przykład 1.
(a) $S_n = \frac{1}{n}$. Jeśli $\varepsilon > 0$ jest nam dane, to bierzemy $N_\varepsilon = [\frac{1}{\varepsilon}] + 1$. Wtedy mamy, że

$$0 < \frac{1}{n} < \frac{1}{N_\varepsilon} < \varepsilon \quad \text{dla } n > N_\varepsilon$$

i dlatego $\lim_{n\to\infty} S_n = 0$.
(b) $S_n = 1 - \frac{(-1)^n}{n}$. Bierzemy $g = 1$ i argumentujemy podobnie jak wyżej: dla $\varepsilon > 0$ wybieramy $N_\varepsilon = [\frac{1}{\varepsilon}] + 1$, wtedy mamy

$$d(S_n, g) = |S_n - 1| = \frac{1}{n} < \frac{1}{N_\varepsilon} < \varepsilon \quad \text{dla } n > N_\varepsilon.$$

(c) $S_n = (-1)^n$ nie ma żadnej granicy, bo dla wyrazów parzystych $S_n = 1$, dla nieparzystych $S_n = -1$.
(d) $S_n = n$, też nie ma granicy.

W przykładach (a) i (b) znaleźliśmy granice ciągów, rodzi się naturalne pytanie, czy są one wyznaczone w sposób jednoznaczny. Odpowiedź jest zawarta poniżej.

Stwierdzenie 4. Załóżmy, że ciąg liczb rzeczywistych $\{a_n\}_{n=1}^{\infty}$ jest zbieżny. Wtedy
(a) granica ciągu a_n jest wyznaczona jednoznacznie;
(b) ciąg a_n jest *ograniczony*, tj. istnieje $M > 0$, takie że $|a_n| < M$ dla $n = 1, 2, \ldots$

Dowód. (a) Jeśliby $\lim_{n\to\infty} a_n = g$ i $\lim_{n\to\infty} a_n = g^1$ oraz $g \neq g^1$ to wystarczy wziąć $\varepsilon = \frac{1}{2}|g - g^1|$, aby dostać sprzeczność, bo dla $n > N$ wszystkie wyrazy mają spełniać $|a_n - g| < \varepsilon$ i $|a_n - g^1| < \varepsilon$, do tego $2\varepsilon = |g - g^1| = |g - a_n + a_n - g^1| < \varepsilon + \varepsilon$, sprzeczność.
(b)Weźmy $\varepsilon = 1$, wtedy dla $n > N_1$ mamy

$$|a_n| = |a_n - g + g| \leq |a_n - g| + |g| = 1 + |g|.$$

Zatem możemy położyć $M = \max\{|a_1|, \ldots, |a_{N_1}|, 1 + |g|\}$. □

Uwaga. Powyższe stwierdzenie jest prawdziwe także i dla ciągów punktów z $\mathbb{R}^k$.

Przyglądając się przykładowi 1(d) widzimy, że wyrazy ciągu rosną nieograniczenie. Chciałoby się powiedzieć, że S_n ma granicę nieskończoną. W tym celu przyjmiemy następujące określenie.

Definicja 4. Niech $\{a_n\}_{n=1}^{\infty}$ będzie ciągiem liczb rzeczywistych. Powiemy, że $\{a_n\}_{n=1}^{\infty}$ zbiega do *plus nieskończoności* i piszemy

$$\lim_{n\to\infty} a_n = +\infty$$

(odpowiednio, *minus nieskończoności* i $\lim_{n\to\infty} a_n = -\infty$) jeśli dla dowolnej liczby rzeczywistej M istnieje $N_M \in \mathbb{N}$, taka, że dla $n > N_\varepsilon$ mamy, że $a_n > M$ (odpowiednio, $a_n < M$).

Odnotujmy teraz właściwości działań arytmetycznych na granicach.

Stwierdzenie 5. Załóżmy, że ciągi $\{a_n\}$ i $\{b_m\}$ liczb rzeczywistych są zbieżne do granic skończonych, $\lim_{n\to\infty} a_n = a$ i $\lim_{n\to\infty} b_n = b$. Wtedy
(a) $\lim_{n\to\infty}(a_n + b_n) = a + b$;
(b) jeśli c jest liczbą rzeczywistą, to $\lim_{n\to\infty} ca_n = ca$;
(c) $\lim_{n\to\infty} a_n b_n = ab$;
(d) jeśli b_n i b są różne od zera, to $\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{a}{b}$

Dowód. Wykażmy tylko punkt (a), bo dowody pozostałych są podobne. Dla dowolnego $\varepsilon > 0$ istnieją N_ε^a i N_ε^b takie, że $|a_n - a| < \frac{\varepsilon}{2}$ i $|b_n - b| < \frac{\varepsilon}{2}$ dla $n > N_\varepsilon^a$, $n > N_\varepsilon^b$. Zatem dla $n > N = \max\{N_\varepsilon^a, N_\varepsilon^b\}$ mamy

$$|a_n + b_n - (a + b)| \leq |a_n - a| + |b_n - b| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2}. \qquad \square$$

Uwagi. Z faktu istnienia $\lim_{n\to\infty}(a_n + b_n)$ nie można wnosić istnienia granic ciągów a_n i b_n. Na przykład $a_n = (-1)^n n$, $b_n = -(-1)^n n$, nie są zbieżne, ale $a_n + b_n = 0$ jest oczywiście zbieżny.

Trzeba wyłączyć z rozważań granice nieskończone, symbole $\frac{\infty}{\infty}$, $\frac{0}{0}$, $\infty - \infty$ są nieoznaczone, to znaczy, że można zbudować ciągi a_n i b_n takie, że $a_n \to \infty, b_n \to \infty$, ale zachowanie $a_n + b_n, a_n/b_n$ jest **dowolne**. Podobnie, jeśli $a_n \to 0, b_n \to 0$ ($a_n \to 0, b_n \to \infty$), to $\frac{a_n}{b_n}$ (odpowiednio, $a_n \cdot b_n$) może się dowolnie zachowywać. Z drugiej strony można wykazać odpowiedniki punktów (a) i (b) jeśli a albo b są nieskończone.

Przykład 2. Czasem łatwo sobie poradzić z przypadkiem $\frac{\infty}{\infty}$, np. $a_n = 7n + 6$, $b_n = 6n - 3$, wtedy

$$\frac{a_n}{b_n} = \frac{7n + 6}{6n - 3}.$$

Po podzieleniu licznika i mianownika przez n dostaniemy dzięki poprzedniemu stwierdzeniu i przykładowi 1 (a)

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \lim_{n\to\infty} \frac{7 + 6/n}{6 - 3/n} = \frac{7}{6}$$

Pojawia się naturalny problem: Czy można scharakteryzować ciągi zbieżne? Zauważmy najpierw pewien prosty fakt: wyrazy ciągu zbieżnego zbliżają się do siebie, ściślej dla dowolnego $\varepsilon > 0$ istnieje N_ε, takie że dla $m, n > N_\varepsilon$ mamy, że

$$d(a_n, a_m) \leq d(a_n, g) + d(g, a_m) < 2\varepsilon,$$

gdzie po drodze zastosowaliśmy nierówność trójkąta. Przepiszmy powyższą obserwację porządnie:

$$\text{dla dowolnego } \varepsilon > 0 \text{ istnieje } N_\varepsilon, \text{ takie że mamy } d(a_n, a_m) < \varepsilon \text{ dla } n, m > N_\varepsilon. \qquad (2)$$

Ciąg, który spełnia (2) nazywa się *ciągiem Cauchy'ego*. O znaczeniu ciągów Cauchy'ego przekona nas poniższy ogólny fakt.

Twierdzenie 6. Ciąg $\{a_n\}_{n=1}^\infty$ punktów przestrzeni $\mathbb{K}^k$ jest zbieżny wtedy i tylko wtedy, gdy jest on ciągiem Cauchy'ego.

Jest to oczekiwana charakteryzacja ciągów zbieżnych, którą jednak pozostawimy bez dowodu. Jest ona o tyle ważna, że często łatwiej będzie wykazać zbieżność ciągu niż wskazać jego granicę. Przykład podamy za chwilę, najpierw definicja.

Definicja 5.

(a) Ciąg liczb rzeczywistych $\{a_n\}_{n=1}^{\infty}$ nazywamy *rosnącym* (odpowiednio, *ściśle rosnącym*), jeśli $a_{n+1} \geq a_n$ (odpowiednio, $a_{n+1} > a_n$) dla wszystkich $n \in \mathbb{N}$.

(b) Ciąg liczb rzeczywistych $\{a_n\}_{n=1}^{\infty}$ nazywamy *malejącym* (odpowiednio, *ściśle malejącym*), jeśli $a_{n+1} \leq a_n$ (odpowiednio $a_{n+1} < a_n$) dla wszystkich $n \in \mathbb{N}$.

(c) Ciąg liczb rzeczywistych nazywamy *monotonicznym*, jeśli jest on rosnący albo malejący.

Ciągi monotoniczne mają pewną cenną właściwość.

Stwierdzenie 7. Załóżmy, że $\{a_n\}_{n=1}^{\infty}$ jest monotoniczny i ograniczony, wtedy jest on zbieżny.

Dowód. Załóżmy, że a_n jest rosnący i niech $a = \sup_{n\geq 1} a_n$. Z definicji kresu górnego wynika, że dla dowolnego $\varepsilon > 0$ można wskazać, a_{n_0} takie że $a_{n_0} > a - \varepsilon$. Z monotoniczności ciągu dostaniemy ponadto, że $a_k \geq a_{n_0} > a - \varepsilon$ dla $k \geq n_0$. Zatem $a = \lim_{n\to\infty} a_n$. □

Rozważania poniżej są zastosowaniem nowego stwierdzenia.

Przykład 3. Niech $a_n = \sum_{k=0}^{n} \frac{1}{k!}$, oczywiście $a_{n+1} > a_n$. Co więcej ciąg $\{a_n\}_{n=1}^{\infty}$ jest ograniczony, bo dla $k > 2$ mamy, że $\frac{1}{k!} < \frac{1}{2^{k-1}}$, zatem

$$0 \leq a_n < 1 + \sum_{k=0}^{n-1} \frac{1}{2^k} \leq 3.$$

Tym samym, ciąg a_n jest zbieżny, jego granicę oznaczamy literą e:

$$\lim_{n\to\infty} \sum_{k=0}^{n} \frac{1}{k!} = e$$

Podamy jeszcze jedno wygodne narzędzie badania ciągów, które zastosujemy w przykładach poniżej.

Twierdzenie 8. (o trzech ciągach). Niech $\{a_n\}_{n=1}^{\infty}$, $\{b_n\}_{n=1}^{\infty}$ i $\{c_n\}_{n=1}^{\infty}$ będą takimi ciągami liczb rzeczywistych, że $a_n \leq b_n \leq c_n$, nadto ciągi $\{a_n\}_{n=1}^{\infty}$ i $\{c_n\}_{n=1}^{\infty}$ są zbieżne do wspólnej granicy g. Wtedy ciąg $\{b_n\}$ też jest zbieżny do g.

Dowód. Z definicji granicy dla dowolnego $\varepsilon > 0$ istnieje takie N_ε, że $|a_n - g| < \varepsilon$ i $|c_n - g| < \varepsilon$ dla $n > N_\varepsilon$. Zatem $g - \varepsilon < a_n \leq b_n \leq c_n < g + \varepsilon$. Oznacza to, że $g - \varepsilon < b_n < g + \varepsilon$ albo inaczej $|g - b_n| < \varepsilon$, gdy $n > N_\varepsilon$. □

Przykład 4.
(a) Niech $p > 0$, wtedy $\lim_{n\to\infty} \frac{1}{n^p} = 0$. Jest tak, bo dla $\varepsilon > 0$ wystarczy przyjąć $N_\varepsilon = [(\frac{1}{\varepsilon})^{1/p}] + 1$.
(b) Niech $p > 0$, wtedy $\lim_{n\to\infty} \sqrt[n]{p} = 1$. Wystarczy rozpatrzyć $p > 1$, bo przypadek $p = 1$ jest nieciekawy, a $p < 1$ sprowadzamy do pierwszego podstawiając $q = \frac{1}{p}$, bo $\sqrt[n]{p} = 1/\sqrt[n]{q}$. Kładziemy $x_n = \sqrt[n]{p} - 1$, oczywiście $x_n > 0$, co więcej

$$p = (1 + x_n)^n = \sum_{k=0}^{n} \binom{n}{k} x_n^k \geq 1 + \binom{n}{1} x_n = 1 + nx_n$$

tj. $0 \leq x_n \leq \frac{p-1}{n}$, a skoro $\frac{p-1}{n} \to 0$, to twierdzenie o trzech ciągach, daje że $\lim_{n\to\infty} x_n = 0$, tj.

$$\lim_{n\to\infty} \sqrt[n]{p} = 1$$

(c) $\lim_{n\to\infty} \sqrt[n]{n} = 1$. Będziemy postępowali podobnie, kładziemy $x_n = \sqrt[n]{n} - 1 \geq 0$. Zauważmy, że

$$n = (1+x_n)^n = \sum_{k=0}^{n} \binom{n}{k} x_n^k \geq 1 + \binom{n}{1} x_n + \binom{n}{2} x_n^2 \geq 1 + \frac{n(n-1)}{2} x_n^2$$

tym samym

$$\frac{2(n-1)}{n(n-1)} \geq x_n^2 \quad \text{a równoważnie } x_n \leq \sqrt{\frac{2}{n}}$$

a skoro $\frac{1}{\sqrt{n}} \to 0$, to $\lim_{n\to\infty} x_n = 0$ i $\lim_{n\to\infty} \sqrt[n]{n} = 1$.
(d) Jeśli $|x| < 1$, to $\lim_{n\to\infty} x^n = 0$. Niech $\varepsilon > 0$ będzie dowolne. Na mocy (b) $\sqrt[n]{\varepsilon} \to 1$. Istnieje zatem takie N, że $\sqrt[n]{\varepsilon} > |x|$, dla $n > N$, tj. $\varepsilon > |x|^n$.

3.1.1 Podciągi i Twierdzenie Bolzano–Weierstrassa

Zastanówmy się nad ciągiem z przykładu 17 (c), $S_n = (-1)^n$. Można o nim powiedzieć, że jedna połowa wyrazów dąży do 1, a druga do -1. Aby ściśle opisać tę sytuację wprowadzimy nowe pojęcie.

Definicja 6. Niech będzie dany ciąg $\{a_n\}_{n=1}^{\infty}$ elementów X a $\{n_k\}_{k=1}^{\infty}$ jest śiście rosnącym ciągiem liczb naturalnych. Wtedy ciąg $\{a_{n_k}\}_{k=1}^{\infty}$ nazywamy podciągiem ciągu $\{a_n\}_{n=1}^{\infty}$.

Wtedy opisana sytuacja jest szczególnym przypadkiem ogólniejszego twierdzenia.

Twierdzenie 9. (Bolzano–Weierstrassa) Niech $\{a_n\}_{n=1}^{\infty}$ będzie ograniczonym ciągiem liczb rzeczywistych. Wtedy istnieje zbieżny podciąg ciagu $\{a_n\}_{n=1}^{\infty}$.

Dowód. Podamy dowód, który łatwo przenieść na przypadek przestrzeni $\mathbb{R}^n$.

Skoro $\{a_n\}_{n=1}^{\infty}$ jest ograniczony, to istnieje taka liczba $M > 0$, że $a_n \in [-M, M] =: I_0$. Dzielimy przedział I_0 na dwie części, mające tylko jeden punkt wspólny, o równej długości I_{00} i I_{01}, tj. $I_0 = I_{00} \cup I_{01}$. Przynajmniej jeden z nich (oznaczmy go symbolem I_1) zawiera nieskończenie wiele elementów a_n. Niech n_1 będzie najmniejszą taką liczbą, że $a_{n_1} \in I_1$. Następnie dzielimy I_1 na dwie części, mające tylko jeden punkt wspólny, o równej długości I_{10} i I_{11}, tj. $I_1 = I_{10} \cup I_{11}$ i ich długość jest równa $M/2$. Przynajmniej jeden z nich (oznaczmy go symbolem I_2) zawiera nieskończenie wiele elementów a_n. Niech n_2 będzie najmniejszą taką liczbą, że $a_{n_2} \in I_2$. W dalszym ciągu postępujemy wg. przedstawionego schematu, w k-tym kroku uzyskujemy $a_{n_k} \in I_k$ i długość I_k jest równa $M/2^{k-1}$. Dlatego ciąg $\{a_{n_k}\}_{k=1}^{\infty}$ (tj. podciąg $\{a_n\}_{n=1}^{\infty}$) jest ciągem Cauchy'ego, czyli jest zbieżny. □

3.2 Szeregi

Zajmiemy się teraz szczególnymi ciągami liczbowymi.

Definicja 7. Dla danego ciągu $\{a_n\}_{n=1}^{\infty}$ liczb rzeczywistych lub zespolonych tworzymy nowy ciąg, kładąc $S_n = \sum_{k=1}^{n} a_n$. Symbol

$$a_1 + a_2 + a_3 + \ldots \text{ lub } \quad \sum_{n=1}^{\infty} a_n$$

będziemy nazywali *szeregiem*. Ciąg $\{S_n\}_{n=1}^{\infty}$ nazywamy jego *ciągiem sum częściowych szeregu*. Jeśli ciąg $\{S_n\}_{n=1}^{\infty}$ jest zbieżny do liczby rzeczywistej S to mówimy, że *szereg jest zbieżny* do S.

Zapytamy teraz jak wygląda warunek Cauchy'ego dla szeregów. W odpowiedzi dostaniemy użyteczny fakt.

Stwierdzenie 10. Szereg $\sum_{n=1}^{\infty} a_n$ jest zbieżny wtedy i tylko wtedy, gdy dla każdego $\varepsilon > 0$ istnieje N_ε, takie, że dla $n > k \geq N_\varepsilon$ mamy

$$|\sum_{i=k}^{n} a_i| < \varepsilon.$$

Dowód. Z definicji zbieżności szeregu wynika, że jego zbieżność jest równoważna zbieżności ciągu jego sum częściowych S_n. Różnica $S_n - S_k$ przyjmuje postać

$$S_n - S_k = \sum_{i=k}^{n} a_i,$$

skąd wynika prawdziwość naszego twierdzenia. □

Zauważmy, że jeśli przyjmiemy $n = k + 1$, to dostaniemy

$$|\sum_{i=k}^{k+1} a_i| = |a_k| < \varepsilon.$$

tj. jeśli szereg jest zbieżny, to koniecznie $a_k \to 0$, gdy $k \to \infty$.

Twierdzenie o ciągu monotonicznym da nam inny wynik.

Wniosek 11. Jeśli $a_n \geq 0$, to szereg $\sum_{n=1}^{\infty}$ jest zbieżny wtedy i tylko wtedy, gdy ciąg sum częściowych jest ograniczony.

Dowód. $\Rightarrow$ Jeśli ciąg S_n jest zbieżny, to jest ograniczony, patrz stwierdzenie 4.

$\Leftarrow$ Na mocy założenia o a_n, dostaniemy $S_{n+1} \geq S_n$. Zatem ograniczoność $\{S_n\}_{n=1}^{\infty}$ pociąga jego zbieżność na mocy stwierdzenia 7. □

Obliczmy jedną prostą sumę szeregu, a mianowicie szeregu geometrycznego.

Stwierdzenie 12. Jeśli $0 \leq |x| < 1$, to $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$.

Dowód. Jak zauważyliśmy wcześniej w przykładzie 1

$$S_n = 1 + x + x^2 + \ldots + x^n = \frac{(1 + x + x^2 + \ldots + x^n)(1 - x)}{1 - x} = \frac{1 + x^{n+1}}{1 - x}$$

Wcześniej wykazaliśmy (patrz przykład 4(d)) też, że $x^n \to 0$, zatem $\lim_{n\to\infty} S_n = \frac{1}{1-x}$. □

Chcemy sformułować teraz kilka kryteriów zbieżności szeregów. Zaczniemy od najprostszych. Poniższe twierdzenie nosi nazwę kryterium porównawczego.

Twierdzenie 13. Dane są szeregi liczb rzeczywistych $\sum_{n=1}^{\infty} a_n$, $\sum_{n=1}^{\infty} b_n$ i $\sum_{n=1}^{\infty} c_n$.
(a) jeśli $|a_n| \leq c_n$ i jest $\sum_{n=1}^{\infty} c_n$ zbieżny, to $\sum_{n=1}^{\infty} a_n$ też jest zbieżny.
(b) jeśli $a_n \geq b_n > 0$ i $\sum_{n=1}^{\infty} b_n = \infty$, to $\sum_{n=1}^{\infty} a_n = \infty$.

Dowód. (a) skoro $\sum_{n=1}^{\infty} c_n$ jest zbieżny, to spełnia warunek Cauchy'ego: dla $\varepsilon > 0$ istnieje takie N, że $\varepsilon > \sum_{k=n}^{m} c_k$ dla $m \geq n > N$. Z założenia i nierówności trójkąta dostaniemy, $\varepsilon \geq \sum_{k=n}^{m} |a_k| \geq |\sum_{k=n}^{m} a_k|$. Zatem ciąg sum częściowych szeregu $\sum_{n=1}^{\infty} a_n$ też spełnia warunek Cauchy'ego, tj. szereg $\sum_{n=1}^{\infty} a_n$ jest zbieżny.
(b) jest oczywiste, szczegóły pozostawiamy Czytelnikowi. □

Kryterium porównawcze pozwala wyprowadzić następujący wniosek.

Twierdzenie 14. (kryterium Cauchy'ego zagęszczające) Załóżmy, że ciąg dodatnich liczb rzeczywistych $\{a_n\}_{n=1}^{\infty}$ jest monotonicznie zbieżny do zera. Wtedy szereg $\sum_{n=1}^{\infty} a_n$ jest zbieżny wtedy i tylko wtedy, gdy szereg $\sum_{k=0}^{\infty} 2^k a_{2^k}$ jest zbieżny.

Dowód. $\Rightarrow$ Niech $S_n = \sum_{k=1}^{n} a_n$, wtedy

$$S_{2^n} = a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7 + a_8 + \ldots + a_{2^n}.$$

Dzięki monotoniczności wyrazów szeregu dostaniemy

$$\begin{aligned} S_{2^n} &\geq a_1 + a_2 + a_4 + a_4 + a_8 + a_8 + a_8 + a_8 + a_{16} + \ldots + a_{2^n} \\ &= a_1 + a_2 + 2a_4 + 4a_8 + 8a_{16} + \ldots + 2^{n-1} a_{2^n} \\ &= \frac{1}{2} \sum_{k=0}^{n} 2^k a_{2^k} = S_n^1 \end{aligned}$$

Zatem ciąg sum częściowych S_n^1 jest monotoniczny i ograniczony (przez $\sum_{k=1}^{\infty} a_k$). Tym samym szereg $\sum_{k=1}^{\infty} 2^k a_{2^k}$ jest zbieżny.
$\Leftarrow$ Przyjmijmy

$$\begin{aligned} S_n^1 &= a_1 + 2a_2 + 4a_4 + 8a_8 + \ldots + 2^n a_{2^n} \\ &= a_1 + a_2 + a_2 + \\ &\quad + a_4 + a_4 + a_4 + a_4 \\ &\quad + a_8 + a_8 + a_8 + a_8 + a_8 + a_8 + a_8 + a_8 + \ldots + a_{2^n} \end{aligned}$$

dzięki monotoniczności ciągu $\{a_n\}_{n=1}^{\infty}$ mamy

$$S_n^1 \geq a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7 + a_8 + a_9 + \ldots + a_{2^n} = S_n \equiv \sum_{k=1}^{n} a_k.$$

Skoro ciąg S_n^1 jest zbieżny, a więc i ograniczony, to ciąg S_n jest ograniczony dzięki powyższej nierówności. Wiemy jeszcze, że S_n jest ciągiem monotonicznym, więc S_n jest zbieżny, tj. szereg $\sum_{n=1}^{\infty} a_n$ jest zbieżny, co należało wykazać. □

Kryterium Cauchy'ego pozwala nam zbadać nowe przykłady szeregów.

Przykład 5. Szereg harmoniczny $\sum_{n=1}^{\infty} \frac{1}{n^p}$ jest zbieżny wtedy i tylko wtedy, gdy $p > 1$. Zauważmy na wstępie, że przypadek $p \leq 0$ jest nieciekawy, bo wtedy $\frac{1}{n^p} = n^{-p} \geq 1$ i ciąg wyrazów szeregu **nie** zbiega do zera. Natomiast w przypadku, gdy $p > 0$ zagadnienie zbieżności szeregu, dzięki kryterium Cauchy'ego zostaje zredukowane do badania zbieżności szeregu

$$\sum_{k=0}^{\infty} 2^k \frac{1}{2^{pk}} = \sum_{k=0}^{\infty} 2^{(1-p)k} = \sum_{k=0}^{\infty} x^k,$$

gdzie $x = 2^{1-p}$. Jeśli $p > 1$, to dostaniemy szereg geometryczny zbieżny (co było treścią wcześniejszego stwierdzenia). Jeśli $p \leq 1$, to $x^k \geq 1$ i x^k nie zbiega do zera, tj. szereg $\sum_{k=0}^{\infty} x^k$ nie jest zbieżny (do sumy skończonej).

Zastosowanie szeregu geometrycznego w kryterium porównawczym prowadzi do ciekawych wniosków. Jednym jest kolejne kryterium zbieżności.

Twierdzenie 15. (kryterium Cauchy'ego) Rozpatrzmy szereg liczbowy $\sum_{n=1}^{\infty} a_n$. Załóżmy, że istnieje granica $\lim_{n\to\infty} \sqrt[n]{|a_n|} = \alpha$. Wtedy,

(a) jeśli $\alpha < 1$, to szereg $\sum_{n=1}^{\infty} a_n$ jest zbieżny;

(b) jeśli $\alpha > 1$ to szereg $\sum_{n=1}^{\infty} a_n$ jest rozbieżny;

(c) jeśli $\alpha = 1$ to kryterium nie rozstrzyga zbieżności.

Dowód. (a) Skoro $\alpha < 1$, to istnieje $\varepsilon > 0$, takie że $\alpha + \varepsilon < 1$. Dla owego ε istnieje takie N_ε, że

$$|\sqrt[n]{|a_n|} - \alpha| < \varepsilon \quad \text{tj.} \quad \sqrt[n]{|a_n|} < \alpha + \varepsilon < 1, \text{ gdy } n > N_\varepsilon. \tag{3}$$

Mamy zatem

$$|a_n| = (\sqrt[n]{|a_n|})^n < (\alpha + \varepsilon)^n, \text{ gdy } n > N_\varepsilon$$

Możemy zastosować kryterium porównawcze, punkt (a), gdzie $c_n = (\alpha+\varepsilon)^n$, aby wywnioskować zbieżność.

(b) Skoro $\alpha > 1$, to istnieje takie $\varepsilon > 0$, że $\alpha - \varepsilon > 1$. Zatem dla tego ε istnieje takie N_ε, że

$$\sqrt[n]{|a_n|} > \alpha - \varepsilon > 1 \quad \text{gdy} \quad n > N_\varepsilon.$$

Tym samym dostaniemy

$$|a_n| = (\sqrt[n]{|a_n|})^n > (\alpha - \varepsilon)^n > 1 \quad \text{gdy} \quad n > N_\varepsilon \tag{4}$$

i ciąg a_n nie zbiega do zera, więc szereg jest rozbieżny.
(c) Dla szeregu harmonicznego mamy, że

$$a_n = \frac{1}{n^p} \quad \text{i} \quad \sqrt[n]{a_n} = \frac{1}{(\sqrt[n]{n^p})}.$$

Obliczyliśmy wcześniej, że $\sqrt[n]{n} \to 1$, więc $\sqrt[n]{n^p} \to 1$ tj. $\alpha = 1$, ale dla $p > 1$ szereg jest zbieżny, a dla $p < 1$ rozbieżny.

Uwaga. Ściśle rzecz ujmując **nie** wykorzystywaliśmy istnienia granicy $\lim_{n\to\infty} \sqrt[n]{|a_n|}$ tylko **słabsze** właściwości (3) i (4).

Łatwiejszym w użyciu, bo wymagającym wykonania prostszych operacji jest poniższe kryterium.

Twierdzenie 16. (kryterium d'Alemberta). Rozpatrzmy szereg liczbowy $\sum_{n=1}^{\infty} a_n$. Załóżmy, że istnieje granica

$$\lim_{n\to\infty} |a_{n+1}|/|a_n| = \alpha.$$

Wtedy,
(a) jeśli $\alpha < 1$, to szereg $\sum_{n=1}^{\infty} a_n$ jest zbieżny
(b) jeśli $\alpha > 1$, to szereg $\sum_{n=1}^{\infty} a_n$ jest rozbieżny
(c) jeśli $\alpha = 1$, to brak jest rozstrzygnięcia.

Dowód. (a) Postępujemy podobnie jak w poprzednim dowodzie. Skoro $\alpha < 1$, to istnieje $\varepsilon > 0$, spełniające $\alpha + \varepsilon < 1$. Dla owego $\varepsilon > 0$ istnieje takie $N_\varepsilon \in \mathbb{N}$, że $|\frac{a_{n+1}}{a_n}| < \alpha + \varepsilon$, dla $n > N_\varepsilon$. Skoro tak, to

$$|a_n| = |\frac{a_n}{a_{n-1}}||a_{n-1}| = \ldots = |\frac{a_n}{a_{n-1}}||\frac{a_{n-1}}{a_{n-2}}|\ldots|\frac{a_{N_\varepsilon - 1}}{a_{N_\varepsilon}}||a_{N_\varepsilon}| \leq (\alpha+\varepsilon)^{n-N_\varepsilon}|a_{N_\varepsilon}| \text{ gdy } n > N_\varepsilon$$

Teraz korzystamy z kryterium porównawczego punktu (a) dla $c_n = (\alpha+\varepsilon)^{n-N_\varepsilon}|a_{N_\varepsilon}|$

(b) Rozumujemy podobnie. Skoro $\alpha > 1$, to istnieje takie $\varepsilon > 0$, że $\alpha - \varepsilon > 1$ i dla tego ε istnieje $N_\varepsilon \in \mathbb{N}$, że $|\frac{a_{n+1}}{a_n}| > \alpha - \varepsilon > 1$, gdy $n > N_\varepsilon$. Stąd

$$|a_n| = |\frac{a_n}{a_{n-1}}| \cdot |\frac{a_{n-1}}{a_{n-2}}|\ldots|\frac{a_{N_\varepsilon - 1}}{a_{N_\varepsilon}}||a_{N_\varepsilon}| > (\alpha-\varepsilon)^{n-N}|a_{N_\varepsilon}|$$

i okazuje się, że $|a_n|$ nie zbiega do zera.

(c) Wystarczy, tak jak poprzednio rozpatrzyć przykład szeregów harmonicznych, by przekonać się o braku rozstrzygnięcia, gdy $\alpha = 1$.

Uwaga. Kryterium d'Alemberta jest łatwiejsze w użyciu, ale jest słabsze niż kryterium Cauchy'ego, bo istnieje przykład szeregu dla którego kryterium d'Alemberta nie daje rozstrzygnięcia podczas gdy kryterium Cauchy'ego rozstrzyga kwestię. Pozostawimy tę kwestię bez dowodu.

Zajmiemy się teraz krótko szeregami, których wyrazy zmieniają znak. Sformułujemy jedno zasadnicze twierdzenie, którego dowód pominiemy.

Twierdzenie 17. Załóżmy, że mamy dane 2 ciągi liczb rzeczywistych $\{a_n\}$ i $\{b_n\}$. Połóżmy $A_n = \sum_{k=1}^{n} a_k$. Załóżmy, że
(a) A_n jest ciągiem ograniczonym;
(b) b_n jest ciągiem monotonicznie malejącym;
(c) $b_n \to 0$.
Wtedy szereg $\sum_{n=1}^{\infty} a_n b_n$ jest zbieżny. □

Z pomocą tego twierdzenia wykażemy zbieżność szeregu anharmonicznego

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}. \tag{5}$$

Kładziemy mianowicie $a_n = (-1)^{n-1}$ i $b_n = \frac{1}{n}$, oczywiście

$$A_n = \begin{cases} 1 & \text{dla } n \text{ nieparzystych} \\ 0 & \text{dla } n \text{ parzystych.} \end{cases}$$

i b_n monotonicznie maleje do zera. Wnosimy zatem, że szereg (5) jest zbieżny.

Zadajmy sobie pytanie, co by się stało gdybyśmy w szeregu (5) zmienili porządek sumowania i zamiast

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \dots$$

napisali

$$1 + \frac{1}{3} - \frac{1}{2} + \frac{1}{5} + \frac{1}{7} - \frac{1}{4} + \frac{1}{9} + \frac{1}{11} \dots$$

Można wykazać, że nowy szereg jest zbieżny, ale do innej sumy! Należy przypuszczać, że ma to związek z faktem, że szereg wartości bezwzględnych szeregu (5), tj. szereg harmoniczny $\sum_{n=1}^{\infty} \frac{1}{n}$, nie ma granicy skończonej. By móc opisać powyższe zjawisko wprowadzimy nowe pojęcia.

Definicja 8. Powiemy, że szereg $\sum_{n=1}^{\infty} a_n$ jest zbieżny *bezwzględnie*, jeśli szereg $\sum_{n=1}^{\infty} |a_n|$ jest zbieżny.

Np. szereg anharmoniczny nie jest zbieżny bezwzględnie, ale szeregi harmoniczne dla $p > 1$ są zbieżne bezwzględnie.

Definicja 9. Szereg $\sum_{n=1}^{\infty} a_n$ jest zbieżny *bezwarunkowo*, jeśli każda zmiana porządku sumowania daje szereg zbieżny (do tej samej sumy).

Odpowiedź na pytanie, które szeregi są bezwarunkowo zbieżne jest zawarta w twierdzeniu niżej, które przytaczamy bez dowodu.

Twierdzenie 18. Szereg $\sum_{n=1}^{\infty} a_n$ jest zbieżny bezwarunkowo wtedy i tylko wtedy, gdy szereg $\sum_{n=1}^{\infty} a_n$ jest zbieżny bezwzględnie. □

3.3 Granica i ciągłość funkcji jednej zmiennej

Pojęcie granicy funkcji w punkcie jest jednym z podstawowych pojęć analizy. Jest ono uściśleniem zdania: 'niezależnie od sposobu, w jaki argumenty x przyliżają się do punktu x_0, ale być może bez osiągania go, wartości funkcji f, tj. $f(x)$ nieograniczenie zbliżają się do punktu g, także, być może bez osiągania go'. Chcemy podkreślić, że dopuszczamy sytuację, kiedy punkt x_0, w którym badamy f nie należy do jej dziedziny. Zanim sformułujemy naszą definicję wprowadzimy dodatkowe oznaczenie. Dla $a, b \in \mathbb{R}$ i takich, że $a < b$ symbolem

$$|a, b|$$

oznaczamy jeden z podziałów $[a, b]$, (a, b), $[a, b)$, $(a, b]$.

Wysłowimy definicję granicy funkcji w punkcie ogólnie dopuszczając funkcje o wartościach zespolonych, mając na uwadze późniejsze zastosowania.

Definicja 10. Niech $f : |a, b| \to \mathbb{K}$ powiemy, że funkcja f ma w punkcie $x_0 \in [a, b]$ granicę g i zapiszemy

$$\lim_{x \to x_0} f(x) = g,$$

jeśli dla dowolnego $\varepsilon > 0$ istnieje $\delta > 0$ taka, że dla dowolnego $x \in |a, b|$ spełniającego $0 < |x - x_0| < \delta$ mamy, że

$$|f(x) - g| < \varepsilon.$$

Można zapytać jaki jest związek granicy funkcji i granicy ciągu. O tym jak bliskie są te pojęcia przekonuje nas poniższa ciągowa charakteryzacja granicy funkcji w punkcie.

Twierdzenie 19. Załóżmy, że $f : |a, b| \to \mathbb{K}$ i $x_0 \in [a, b]$. Wtedy następujące warunki są równoważne

(a) $\lim_{x \to x_0} f(x) = q$;

(b) dla każdego ciągu $\{p_n\}_{n=1}^{\infty} \subset |a, b|$ zbieżnego do x_0 i takiego, że $p_n \neq x_0$ mamy, że

$$\lim_{n \to \infty} f(p_n) = q.$$

Dowód. (a)⇒(b) Skoro $q = \lim_{x \to x_0} f(x)$, to z mocy definicji dla dowolnego $\varepsilon > 0$ istnieje $\delta > 0$, takie że mamy $|f(x) - f(x_0)| < \varepsilon$ dla $x \in |a, b|$ spełniających $0 < |x - x_0| < \delta$. Jeśli zatem $\{p_n\}_{n=1}^{\infty}$ jest dowolnym ciągiem zbieżnym do x_0 i $p_n \neq x_0$, to dla pewnego N i każdego $n > N$ mamy $|p_n - x_0| < \delta$. A zatem $|f(p_n) - x_0| < \varepsilon$ tj. ciąg $f(p_n)$ zbiega do q.

(b)⇐(a) a.a. Zaprzeczenie istnienia granicy oznacza, że dla pewnego $\varepsilon > 0$ i dla wszystkich $\delta > 0$ istnieje takie $x_\delta \in |a, b|$, że $0 < |x_\delta - x_0| < \delta$ i $|f(x_\delta) - q| \geq \varepsilon$. Skoro δ jest dowolne, to weźmy teraz $\delta_n = \frac{1}{n}$. Dostaniemy wtedy ciąg $\{x_n\}_{n=1}^{\infty}$ zbieżny do x_0 i $x_n \neq x_0$. Przede wszystkim jednak $|f(x_n) - q| \geq \varepsilon > 0$ tj. nieprawdą jest, że $\lim_{x \to x_0} f(x) = q$.

Wniosek 20. Granica funkcji w punkcie jest wyznaczona jednoznacznie.

Dowód. Wynika to z faktu, że granica ciągu jest wyznaczona jednoznacznie. □

Podamy teraz podstawowe właściwości granicy funkcji w punkcie.

Twierdzenie 21. Załóżmy, że $f, g : |a, b| \to \mathbb{C}$ i istnieją granice funkcji f i g w punkcie x_0, $\lim_{x\to x_0} f(x) = A$ i $\lim_{x\to g(x)} = B$. Wtedy,
(a) $\lim_{x\to x_0}(f(x) + g(x)) = A + B$;
(b) $\lim_{x\to x_0}(f(x) \cdot g(x)) = A \cdot B$;
(c) $\lim_{x\to x_0} \frac{f(x)}{g(x)} = \frac{A}{B}$, jeśli tylko $B \neq 0$.

Dowód. Powyższe twierdzenie wynika z analogicznego faktu dla granic ciągów, szczegółowy dowód pomijamy. □

Chcielibyśmy teraz uściślić pojęcie ciągłości funkcji. Ciągłą funkcją wydaje się nam funkcja prędkości wody w rzece czy funkcja przypisująca każdemu punktowi na mapie pogodowej temperaturę i ciśnienię powietrza. Oznacza to, że jeśli zmienimy troszkę nasze stanowisko obserwacyjne (lub punkt na mapie) to interesujące nas wielkości zmienią się tylko trochę, tj. nie doznają gwałtownych zmian w postaci skoku. Z drugiej strony gęstość materii w sali wykładowej jawi się wielkością nieciągłą, bo na granicy ławki i powietrza mamy gwałtowny skokowy wzrost (lub spadek gęstości). Podejrzewamy, też że ciągłość powinna mieć też związek z granicą funkcji w punkcie.

Definicja 11. Załóżmy, że $f : |a, b| \to \mathbb{K}$.

(a) O funkcji f powiemy, że jest *ciągła* w punkcie $x_0 \in |a, b|$, jeśli dla dowolnego $\varepsilon > 0$ istnieje takie $\delta > 0$, że mamy $|f(x) - f(x_0)| < \varepsilon$ dla wszystkich $x \in |a, b|$ spełniających $|x - x_0| < \delta$.

(b) O funkcji f powiemy, że jest *ciągła* w przedziale $|a, b|$ wtedy i tylko wtedy, gdy jest ciągła w każdym punkcie przedziału $|a, b|$. Zbiór funkcji ciągłych w przedziale $[a, b]$ oznaczamy symbolem $C([a, b])$.

Uwaga. Z definicji wynika, że ciągłość funkcji f rozpatrujemy wyłącznie w punktach należących do dziedziny f!

Z powyższego określenia granicy funkcji w punkcie wynika natychmiast następujący wynik.

Stwierdzenie 22. Załóżmy, że $f : |a, b| \to \mathbb{K}$. Funkcja f jest ciągła w punkcie x_0 wtedy i tylko wtedy, gdy $\lim_{x\to x_0} f(x) = f(x_0)$.

Nieco bardziej złożonym i do tego ważnym faktem jest następujący wynik, który sformułujemy wyłącznie dla funkcji o wartościach rzeczywistych.

Twierdzenie 23. (o ciągłości funkcji złożonej) Załóżmy, że $f : |a, b| \to \mathbb{R}$, $g : |c, d| \to \mathbb{R}$ i $f(|a, b|) \subset |c, d|$. Zakładamy, że funkcja f jest ciągła w x_0, zaś g ciągła w punkcie $y = f(x_0)$. Wtedy funkcja $h(x) = g \circ f(x)$ jest ciągła w x_0.

Dowód. Skoro g jest ciągła w $f(x_0)$, to dla $\varepsilon > 0$ istnieje takie $\eta > 0$, że $|g(y) - g(f(x_0))| < \varepsilon$ dla $y \in |c, d|$ i $|y - f(x_0)| < \eta$. Skoro f jest ciągła w x_0, to istnieje takie $\delta > 0$, że mamy $|f(x_0) - f(x)| < \eta$ dla $x \in |a, b|$ i $|x - x_0| < \delta$ tj. wynika stąd $|h(x) - h(x_0)| = |g(f(x)) - g(f(x_0))| < \varepsilon$ dla $|x - x_0| < \delta$. □

Do wypowiedzenia następnych właściwości funkcji ciągłych potrzebne będzie dodatkowe pojęcie.

Definicja 12. Zbiór $E \subset \mathbb{R}^n$ (lub $E \subset \mathbb{C}^n$) nazywamy *ograniczonym*, jeśli istnieje $\mu > 0$, takie, że dla wszystkich $x \in E$ jest prawdą, że $|x| < \mu$.

Możemy teraz przedstawić ważne twierdzenie, którego główną część zostawiamy bez dowodu z braku narzędzi pozwalających na jego sprawne przeprowadzenie, natomiast przedstawimy jego szkic. Będzimy się też odwoływać do intuicji i rysunku niżej.

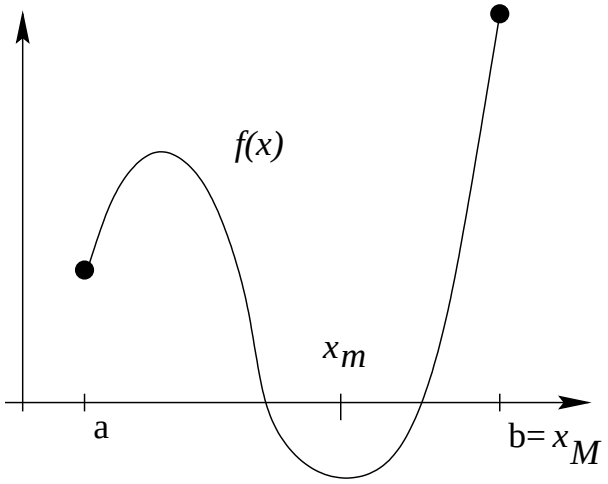
Twierdzenie 24. Załóżmy, że funkcja f o wartościach rzeczywistych jest ciągła w przedziale domkniętym $[a, b]$. Wtedy,

(a) Obraz przedziału $[a, b]$, tj. $f([a, b])$ jest zbiorem ograniczonym.

(b) Istnieją liczby x_M i x_m należące do $[a, b]$ takie, że

$$f(x_M) = \sup_{x\in[a,b]} f(x) \text{ i } f(x_m) = \inf_{x\in[a,b]} f(x).$$

(c) obraz przedziału $[a, b]$ tj. $f([a, b])$ jest przedziałem.



Rys. 3. Przebieg funkcji.

Uwagi na temat dowodu. Rzut oka na rys. 1 przekonuje nas o prawdziwości części (a) powyższego twierdzenia. Zaś szkic ścisłego argumentu dotyczący (b) jest następujący. Niech $q = \sup_{x\in[a,b]} f(x)$, gdybyśmy dla wszystkich $x \in [a, b]$ mieli, że $f(x) < q$, to funkcja $1/q - f(x)$ byłaby dobrze określona i ciągła w $[a, b]$, bo funkcja $x \to 1/x$ jest ciągła, gdy $x \neq 0$. Zatem na mocy punktu (a) funkcja $g(x) = 1/q - f(x)$ byłaby ograniczona. Z drugiej strony z definicji kresu istnieje ciąg $x_n \in [a, b]$, taki, że $f(x_n) \to q$. Co więcej, dzięki twierdzeniu Bolzano-Wierestrassa ten ciąg można tak wybrać, aby miał granicę, $\lim_{n\to\infty} x_n = x^\infty$. Wniosek stąd taki, że $g(x)$ nie może być ograniczona. Uzyskana sprzeczność dowodzi istnienia takiego x_M, że $f(x_M) = \sup_{x\in[a,b]} f(x)$. Podobny argument stosuje się by wykazać istnienie x_m.

(c) Obrazowy argument na podstawie rys. 1 jest taki, że gdyby $f([a, b])$ nie był przedziałem, to byłaby w nim dziura. Zatem, żeby przejść z jednego do drugiego kawałka zbioru $f([a, b])$, funkcja musiałaby wykonać skok, co nie jest możliwe dla funkcji ciągłej.

Uwagi dotyczące twierdzenia. Punkty (a) i (b) twierdzenia są nieprawdziwe, jeśli pominiemy domkniętość przedziału np. $g(x) = \frac{1}{x}$, dla $x \in (0,1)$. Wtedy ani obraz przedziału $(0,1)$ nie jest ograniczony, ani g nie osiąga swoich kresów.

Można podać prosty choć dłuższy niż w (b) ścisły argument używający podobnych metod jak w punkcie (b), aby wykazać (c).

Wnioskiem z punktu (c) jest tzw. *własność Darboux*.

Twierdzenie 25. Jeśli $f : [a,b] \to \mathbb{R}$ jest funkcją ciągłą i liczba c jest pomiędzy $f(a)$ i $f(b)$, to istnieje $x_0 \in [a,b]$, takie że $f(x_0) = c$.

Osobnym tematem jest badanie różnowartościowych funkcji ciągłych, można powiedzieć o nich coś więcej. Na przedziałach domkniętych mają one jeszcze jedną ważną właściwość.

Twierdzenie 26. Niech funkcja $f : [a,b] \to \mathbb{R}$ będzie ciągła i różnowartościowa. Zatem funkcja odwrotna f^{-1} określona na przedziale $f([a,b])$, tj. $f^{-1} : f([a,b]) \to [a,b]$, jest ciągła.

Szkic dowodu jest następujący. Przypuśćmy, że f^{-1} nie jest ciągła. Wtedy istnieje taki ciąg $f(x_n)$, że $f(x_n) \to f(x_0)$ i $f^{-1}(f(x_n))$ nie zbiega do $f^{-1}(f(x_0)) = x_0$. Można tak wybrać ciąg x_n, aby $x_n \to x_1 \neq x_0$, ale wtedy z ciągłości funkcji f mamy $f(x_n) \to f(x_1)$. Skoro $x_1 \neq x_0$, to $f(x_1) \neq f(x_0)$, co przeczy ciągłości f w x_0. Wykazuje to nasze twierdzenie. □

Objaśnimy teraz kilka podstawowych przykładów.

Przykład 6.
(a) Niech $\chi_{\mathbb{Q}}$ będzie funkcją charakterystyczną zbioru liczb wymiernych (patrz §1.4.1). Ponieważ pomiędzy dowolnymi dwoma liczbami rzeczywistymi a i b można znaleźć liczbę wymierną q, tj. $a < q < b$, to funkcja $\chi_{\mathbb{Q}}$ nigdzie nie ma granicy.
(b) Kładziemy

$$g(x) = \begin{cases} x & x \in \mathbb{Q} \\ 0 & x \in \mathbb{R}\backslash\mathbb{Q} \end{cases}$$

Z powodów j.w. funkcja g nie ma granicy w żadnym punkcie różnym od zera, ale dla $x = 0$ mamy nierówność

$$|g(0) - g(x)| = |0 - g(x)| \leq |x|.$$

Tym samym w definicji ciągłości funkcji w punkcie $x = 0$ wystarczy dla zadanego ε przyjąć $\delta = \varepsilon$, aby otrzymać ciągłość funkcji g.
(c) Przyjmijmy, że funkcja $x \to \sin x$ jest ciągła, wtedy funkcja dana wzorem

$$h(x) = \begin{cases} \sin\frac{1}{x} & x \neq 0 \\ 0 & x = 0 \end{cases}$$

jest ciągła w punktach $x \neq 0$. Zbadajmy istnienie granicy w punkcie $x = 0$. Niech $x_n = \frac{1}{2\pi n}$ i $y_n = \frac{1}{2\pi n + \frac{\pi}{2}}$, wtedy $x_n \to 0$ i $y_n \to 0$, ale

$$h(x_n) = \sin(2\pi n) = 0, \quad h(y_n) = \sin(2\pi n + \pi/2) = 1$$

zatem na mocy ciągowej charakteryzacji granicy, granica $\lim_{x\to 0} h(x)$ nie istnieje.
(d) Niech

$$k(x) = \begin{cases} x\sin\frac{1}{x} & x \neq 0 \\ 0 & x = 0 \end{cases}$$

Na mocy (c) funkcja $k(x)$ jest ciągła w punktach $x \neq 0$. Badamy istnienie granicy w punkcie $x = 0$,

$$|k(0) - k(x)| = |0 - x \sin \frac{1}{x}| = |x| \cdot |\sin \frac{1}{x} \leq |x|.$$

Zatem dla zadanego $\varepsilon > 0$ wystarczy przyjąć $\delta = \varepsilon$ w definicji granicy, by otrzymać, że

$$\lim_{x \to 0} k(x) = 0.$$

3.3.1 Funkcje monotoniczne

Wyróżnimy teraz klasę funkcji o ważnych właściwościach.

Definicja 13. Niech będzie dana funkcja $f : [a, b] \to \mathbb{R}$.

(a) Powiemy, że funkcja f jest *rosnąca* (odpowiednio, *ściśle rosnąca*), jeśli dla dowolnych $x, y \in [a, b]$ i takich, że $x < y$ mamy $f(x) \leq f(y)$ (odpowiednio, $f(x) < f(y)$).

(b) Powiemy, że funkcja f jest *malejąca* (odpowiednio, *ściśle malejąca*), jeśli dla dowolnych $x, y \in [a, b]$ i takich, że $x < y$ mamy $f(x) \geq f(y)$ (odpowiednio, $f(x) > f(y)$).

(c) O funkcji, która jest albo rosnąca, albo malejąca powiemy, że jest *monotoniczna*.

Pamiętamy, że ograniczone ciągi monotoniczne są zbieżne. Podejrzewamy więc, że funkcje monotoniczne będą miały granice 'w wielu punktach', zaś nieciągłości będą skokami. Do wysłowienia tych przypuszczeń przydadzą się nowe pojęcia.

Definicja 14. Niech będzie dana funkcja $f : [a, b] \to \mathbb{R}$ i $D \subset [a, b]$. Funkcję $f|_D : D \to \mathbb{R}$ daną wzorem

$$f|_D(x) := f(x)$$

nazywamy *obcięciem* f do D.

Możemy teraz zdefiniować granice jednostronne funkcji w punkcie.

Definicja 15. Niech będzie dana funkcja $f : [a, b] \to \mathbb{R}$ i $x_0 \in (a, b)$. Połóżmy,

$$g_l := f|_{[a,x_0)} \qquad g_p := f|_{(x_0,b]}.$$

(a) Jeśli granica $\lim_{x \to x_0} g_l(x) = q_l$ istnieje, to powiemy, że funkcja f ma *granicę lewostronną* w x_0 i piszemy

$$\lim_{x \to x_0^-} f(x) = q_l.$$

(b) Jeśli granica $\lim_{x \to x_0} g_p(x) = q_p$ istnieje, to powiemy, że funkcja f ma *granicę prawostronną* w x_0 i piszemy

$$\lim_{x \to x_0^+} f(x) = q_p.$$

Teraz ciągowa charakteryzacja granicy funkcji w punkcie łatwo nas przekonuje o prawdziwości następującego faktu.

Twierdzenie 27. Niech dana funkcja $f : [a, b] \to \mathbb{R}$ będzie monotoniczna, wtedy w każdym punkcie $x_0 \in (a, b)$ istnieje granica lewo- i prawostronna funkcji f w punkcie x_0. Istnieją też granice prawostronne w a i lewostronna w b.

Zauważmy, że jeśli granice lewostronna i prawostronna funkcji monotonicznej g w x_0 są równe, to w tym punkcie koniecznie funkcja g jest ciągła. Jeśli są różne, to w obrazie funkcji g jest dziura. Pytanie: ile może być takich dziur? Okazuje się, że najwyżej przeliczalnie wiele, bo prawdziwym jest następujące twierdzenie.

Twierdzenie 28. Niech funkcja $g : [a, b] \to \mathbb{R}$ będzie monotoniczna, wtedy funkcja g jest ciągła we wszystkich punktach z wyłączeniem co najwyżej przeliczalnie wielu, tj. punkty nieciągłości można ustawić w ciąg. (Bez dowodu).

Dla porządku wprowadzimy jeszcze pojęcie granicy nieskończonej funkcji w punkcie i granicy w nieskończoności.

Definicja 16. (a) Niech $f : (a, b) \to \mathbb{R}$ i $x_0 \in [a, b]$. Powiemy, że funkcja f ma w punkcie x_0 granicę $+\infty$ (odpowiednio, $-\infty$), jeśli dla każdego $M \in \mathbb{R}$ istnieje takie $\delta > 0$, że dla każdego $x \in [a, b]$ spełniającego $0 < |x - x_0| < \delta$ mamy $f(x) > M$ (odpowiednio, $f(x) < M$) piszemy

$$\lim_{x\to x_0} f(x) = +\infty \qquad \text{(odpowiednio,} \qquad \lim_{x\to x_0} f(x) = -\infty).$$

(b) Niech $f : (a, +\infty) \to \mathbb{R}$ powiemy, że g jest granicą funkcji w $+\infty$, jeśli dla każdego $\varepsilon > 0$ istnieje $k > a$, takie że dla $x > k$ mamy $|g - f(x)| < \varepsilon$. Piszemy wtedy

$$\lim_{x\to+\infty} f(x) = g.$$

Analogicznie definiujemy granicę w $-\infty$ i granice nieskończone w nieskończoności.

3.3.2 Klasyfikacja punktów nieciągłości

Załóżmy, że funkcja $g : [a, b] \to \mathbb{R}$ jest nieciągła w punkcie $x_0 \in (a, b)$, istnieją wtedy dwie możliwości:

(a) Obie granice jednostronne w p. x_0 funkcji g istnieją. Mówimy, że wtedy g ma *nieciągłość pierwszego rodzaju*. Przykładowo funkcja $\mathbb{R} \ni x \mapsto [x] \in \mathbb{Z}$ ma nieciągłość pierwszego rodzaju w punktach całkowitych.

(b) Przynajmiej jedna z granic jednostronnych funkcji g w x_0 nie istnieje. Mówimy, że wtedy g ma *nieciągłość drugiego rodzaju*. Np. funkcja h z przykładu 6(c).

3.4 Różniczkowanie

Rozważmy kilka sytuacji:

(1) Gdy mamy daną krzywą (np. płaską), to często chcielibyśmy narysować prostą styczną. Chcielibyśmy też w ogóle ją zdefiniować, bo takie proste stwierdzenie: ‘prosta, która ma z daną

krzywą tylko jeden punkt wspólny, jest do niej styczna', które jest prawdziwe dla okręgu, w ogólności jest fałszywe. Natomiast umiemy rysować sieczne, tj. proste przechodzące przez 2 punkty prostej p_1 i p_2. Możemy próbować szukać granicy owych siecznych, gdy p_1 zbliża się do p_2.

(2) Chcielibyśmy roztropnie przybliżać daną krzywą prostymi. Jakimi?

(3) Wiele praw fizyki jest formułowanych w następujący sposób, 'szybkość zmiany wielkości w jest równa funkcji, której argumentem jest czas, położenie w przestrzeni i być może inne wielkości'. Przykładem będzie prawo Newtona: szybkość zmiany pędu = siła.

Uwagi (1) i (3) sugerują wykonanie pewnego przejścia granicznego. Natomiast odpowiedź na pytanie w p. (2) wymaga zastosowania nowego, jeszcze nieznanego narzędzia.

Spróbujmy wprost napisać co to jest styczna do krzywej, która jest wykresem funkcji i zbadać zachowanie współczynnika kierunkowego stycznej. Podobny wynik uzyskamy pisząc, że szybkość to iloraz drogi przez czas, w którym była ona przebyta.

Definicja 17. (pochodnej funkcji). Niech $f : (a, b) \to \mathbb{R}$ i $x_0 \in (a, b)$. Wyrażenie

$$\frac{f(x_0 + h) - f(x_0)}{h},$$

nazywamy *ilorazem różnicowym*, granicę (jeśli istnieje)

$$\lim_{h\to 0} \frac{f(x_0 + h) - f(x)}{h} = g$$

nazywamy *pochodną* funkcji f w punkcie x_0 i piszemy

$$f'(x_0) = g \qquad \text{lub równoważnie} \qquad \frac{df}{dx}(x_0) = g.$$

(Czasem dla funkcji $x \mapsto y(x)$ stosujemy zapis $\dot{y}$.) O funkcji f mówimy wtedy, że jest *różniczkowalna* w x_0.

Ciekawy jest związek pomiędzy różniczkowalnością a ciągłością. Mamy mianowicie,

Twierdzenie 29. Jeśli funkcja $f : (a, b) \to \mathbb{R}$, jest różniczkowalna w punkcie $x_0 \in (a, b)$, to jest ona ciągła w punkcie x_0.

Dowód. Badamy różnicę

$$f(x) - f(x_0) = \frac{f(x) - f(x_0)}{(x - x_0)} \cdot (x - x_0).$$

Połóżmy $h = x - x_0$, wtedy

$$\begin{aligned} |f(x) - f(x_0)| &= \left|\frac{f(x_0 + h) - f(x_0)}{h}\right| |x - x_0| \\ &= \left|\frac{f(x_0 + h) - f(x_0)}{h} - f'(x_0) + f'(x_0)\right| |x - x_0| \\ &\leq \left|\frac{f(x_0 + h) - f(x_0)}{h} - f'(x_0)\right| |x - x_0| + |f'(x_0)||x - x_0|. \end{aligned}$$

Dla danego $\varepsilon > 0$ weźmy $\delta < \varepsilon$, takie że $|f'(x_0)|\delta < \frac{\varepsilon}{2}$ i $|\frac{f(x_0+h)-f(x_0)}{h} - f'(x_0)| < \frac{1}{2}$ dla wszystkich $|h| < \delta$. Zatem,

$$|f(x) - f(x_0)| \leq \frac{1}{2}\delta + |f'(x_0)|\delta < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

co było do wykazania. □

Zanim przystąpimy do obliczania przykładowych pochodnych zapoznamy się z podstawowymi właściwościami różniczkowania.

Twierdzenie 30. Załóżmy, że funkcje f i g są określone na przedziale (a, b) i są różniczkowalne w punkcie $x_0 \in (a, b)$. Wtedy,
(a) $f + g$, $f \cdot g$ są różniczkowalne w x_0 i

$$(f + g)'(x_0) = f'(x_0) + g'(x_0), \quad (fg)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0)$$

(b) jeśli dodatkowo $g(x_0) \neq 0$, to f/g jest różniczkowalna w x_0 i

$$\frac{d}{dx}\left(\frac{f}{g}\right)(x_0) = \frac{f(x_0)g(x_0) - g'(x_0)f(x_0)}{g^2(x_0)}$$

Dowód. Zajmiemy się wyłącznie przypadkiem iloczynu

$$\begin{aligned} & \lim_{h\to 0} \frac{f(x_0+h)g(x_0+h) - f(x_0)g(x_0)}{h} \\ = & \lim_{h\to 0} \frac{f(x_0+h)(g(x_0+h) - g(x_0)) + g(x_0)(f(x_0+h) - f(x_0))}{h} \\ = & f(x_0)g'(x_0) + g(x_0)f'(x_0) \end{aligned}$$

Po drodze skorzystaliśmy z twierdzenia o granicy iloczynu i sumy. Obecnie pozostałe przypadki nie przedstawiają problemu. □

Twierdzenie 31. (o pochodnej funkcji złożonej.) Załóżmy, że $f : (a, b) \to \mathbb{R}$, $g : (c, d) \to \mathbb{R}$, $f((a, b)) \subset (c, d)$ i funkcja f jest ciągła w (a, b) i różniczkowalna w x_0, zaś g jest różniczkowalna w punkcie $y = f(x_0)$. Wtedy funkcja $k(x) = g(f(x))$ jest różniczkowalna w x_0 i $k'(x_0) = g'(f(x_0))f'(x_0)$.
Dowód. Tworzymy iloraz różnicowy dla k,

$$k'(x_0) = \lim_{h\to 0} \frac{k(x_0+h) - k(x_0)}{h} = \lim_{h\to 0} \frac{g(f(x_0)+h)) - g(f(x_0))}{h}$$

$$= \lim_{h\to 0} \frac{g(f(x_0+h)) - g(f(x_0))}{f(x_0+h) - f(x_0)} \cdot \frac{f(x_0+h) - f(x_0)}{h}$$

z twierdzenia o granicy iloczynu mamy, że

$$k'(x_0) = \lim_{h\to 0} \frac{g(f(x_0+h)) - g(f(x_0))}{f(x_0+h) - f(x_0)} \cdot \lim_{h\to 0} \frac{f(x_0+h) - f(x_0)}{h} = (*)$$

z ciągłości funkcji f wynika, że $\lim_{h\to 0} f(x_0+h) = f(x_0)$. Zatem, jeśli napiszemy $H := f(x_0+h) - f(x_0)$, to $H \to 0$, gdy $h \to 0$. Tym samym dostaniemy

$$(*) = \lim_{H\to 0} \frac{g(f(x_0)+H) - g(x_0)}{H} \cdot f'(x_0) = g'(f(x_0)) \cdot f'(x_0). \qquad \square$$

Przedstawimy teraz pewną liczbę przykładów rachunkowych.

Przykład 7.
(a) Niech $c : (a,b) \to \mathbb{R}$ będzie funkcją stałą, $c(x) = c$. Wtedy $\lim_{h\to 0} \frac{c-c}{h} = 0$ tj. $\frac{dc}{dx} = 0$.
(b) $(x^n)' = nx^{n-1}$, gdy $n > 0$. Liczymy,

$$\begin{aligned}\lim_{h\to 0} \frac{(x+h)^n - x^n}{h} &= \lim_{h\to 0} \frac{x^n + h\cdot nx^{n-1} + \binom{n}{2}h^2x^{n-2} + \ldots + h^n - x^n}{h} \\ &= \lim_{h\to 0} \left[nx^{n-1} + h\left(\binom{n}{2}x^{n-2} + \ldots + \binom{n}{0}h^{n-2} \right) \right] = nx^{n-1}.\end{aligned}$$

(c) Wykażemy, że poprzedni wzór jest prawdziwy, też dla $n < 0$. Zastosowanie wzoru na pochodną ilorazu daje nam,

$$\frac{d}{dx}x^n = \frac{d}{dx}\left(\frac{1}{x^{-n}}\right) = -\frac{-nx^{-n-1}}{x^{-2n}} = nx^{n-1}.$$

(d) Przyjmijmy na chwilę, że $\frac{d}{dx}(\sin x) = \cos x$. Funkcję $f : \mathbb{R} \to \mathbb{R}$ zadajemy wzorem,

$$f(x) = \begin{cases} 0, & \text{gdy } x = 0 \\ x\sin\frac{1}{x}, & \text{gdy } x \neq 0 \end{cases}$$

wtedy dla $x \neq 0$ $f'(x) = \sin\frac{1}{x} - \frac{1}{x}\cos\frac{1}{x}$ na mocy twierdzenia o funkcji złożonej i punktu (c) dla $n = -1$. Badamy różniczkowalność w punkcie $x = 0$. Tworzymy iloraz różnicowy i dostaniemy

$$\frac{f(h) - f(0)}{h} = \sin(1/h).$$

Na mocy przykładu 6(c) to wyrażenie nie ma granicy, gdy $h \to 0$.
(e) Dla $n > 1$ kładziemy,

$$g_n(x) = \begin{cases} 0 & x = 0 \\ x^n \sin\frac{1}{x} & x \neq 0. \end{cases}$$

Wtedy dla $x \neq 0$

$$g_n(x) = nx^{n-1}\sin\frac{1}{x} - x^{n-2}\cos\frac{1}{x}.$$

Dla $x = 0$ mamy

$$\lim_{h\to 0} \frac{g_n(0+h) - g_n(0)}{h} = \lim_{h\to 0} \frac{h^n}{h} \sin\frac{1}{h}$$

skoro $n > 1$, to $|\frac{h^n}{h} \sin\frac{1}{h}| \leq h^{n-1} \to 0$ tj. $g_n'(0) = 0$.
(f) Kładziemy $f(x) = |x|$. Wtedy f nie jest różniczkowalna w punkcie $x = 0$, bo nie istnieje granica

$$\lim_{h\to 0} \frac{|0+k| - |h|}{h} = \lim_{h\to 0} \frac{|h|}{h}.$$

Jednym z ważniejszych zadań analizy jest poszukiwanie najmniejszej i największej wartości funkcji w zbiorze. Musimy uściślić takie zadanie, okaże się przy tym, że jest ono szersze niż mogłoby nam się to początkowo wydawać. Określenie poniższe jest nieco na wyrost, bo chwilowo zajmujemy się funkcjami tylko jednej zmiennej.

Definicja 18. Załóżmy, że $f : D \to \mathbb{R}$, i D jest podzbiorem $\mathbb{R}^n$.

(a) Powiemy, że funkcja f ma w punkcie $x_0 \in D$ *maksimum lokalne*, jeśli istnieje takie $\delta > 0$, że jeśli $x \in D$ spełnia $d(x, x_0) < \delta$, to $f(x_0) \geq f(x)$.

(b) Powiemy, że funkcja f ma w punkcie $x_0 \in D$ *ścisłe maksimum lokalne*, jeśli istnieje takie $\delta > 0$, że jeśli $x \in D$ spełnia $0 < d(x, x_0) < \delta$, to $f(x_0) > f(x)$.

(c) Powiemy, że funkcja f ma w punkcie $x_0 \in D$ *maksimum globalne*, jeśli dla wszystkich $x \in D$, mamy $f(x_0) \geq f(x)$.

(d) Powiemy, że funkcja f ma w punkcie $x_0 \in D$ *ścisłe maksimum globalne*, jeśli dla każdego $x \in D$ różnego od x_0 jest prawdą, że $f(x_0) > f(x)$.

(e) Powiemy, że funkcja f ma w punkcie $x_0 \in D$ *minimum lokalne* (odpowiednio, *ścisłe minimum lokalne, minimum globalne, ścisłe minimum globalne*) jeśli funkcja $-f(x)$ ma w punkcie x_0 maksimum lokalne (odpowiednio, ścisłe maksimum lokalne, maksimum globalne, ścisłe maksimum globalne).

(f) Powiemy, że w punkcie x_0 funkcja f ma ekstremum jeśli ma w nim maksimum lub minimum lokalne.

Wprawdzie podana definicja jest ogólna, ale na razie badamy funkcje jednej zmiennej. Wtedy warunek konieczny istnienia ekstremum w punkcie x_0 dla funkcji różniczkowalnej jest podany niżej.

Twierdzenie 32. Niech $f : (a, b) \to \mathbb{R}$ będzie funkcją różniczkowalną w $x_0 \in (a, b)$, w którym przyjmuje maksimum (odpowiednio, minimum) lokalne. Wtedy $f'(x_0) = 0$.

Dowód. Rozpatrzymy najpierw przypadek maksimum lokalnego. Załóżmy, że $h > 0$, wtedy

$$\frac{f(x_0+h) - f(x_0)}{h} \geq 0 \qquad \text{a stąd} \qquad \lim_{h\to 0^+} \frac{f(x_0+h) - f(x_0)}{h} \geq 0.$$

Dla $h < 0$ mamy

$$\lim_{h\to 0^-} \frac{f(x_0+h) - f(x_0)}{h} \leq 0.$$

Ponieważ funkcja f jest różniczkowalna w x_0, to powyższe granice jednostronne są równe pochodnej. Skoro tak, to mamy

$$0 \leq \lim_{h\to 0^+} \frac{f(x_0+h)-f(x_0)}{h} = \lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h} = \lim_{h\to 0^-} \frac{f(x_0+h)-f(x_0)}{h} \leq 0.$$

Tym samym $0 \leq f'(x_0) \leq 0$, czyli $f'(x_0) = 0$.

Aby uzyskać nasz wynik dla minimum stosujemy część udowodnioną do $-f$. □

3.4.1 Twierdzenia o wartości średniej

Z powyższego prostego twierdzenia wypływa wiele ciekawych wniosków. Pierwszymi z serii będą twierdzenia o wartości średniej.

Twierdzenie 33. (uogólnione o wartości średniej). Załóżmy, że funkcje $f, g : [a,b] \to \mathbb{R}$ są ciągłe w $[a,b]$ i różniczkowalne w (a,b). Wtedy istnieje $c \in (a,b)$, takie że

$$(f(b)-f(a))g'(c) = (g(b)-g(a))f'(c).$$

Dowód. Tworzymy funkcję pomocniczą $h : [a,b] \to \mathbb{R}$,

$$h(t) = (f(b)-f(a))g(t) - (g(b)-g(a))f(t)$$

funkcja h jest ciągła w $[a,b]$ i różniczkowalna w (a,b). Co więcej, łatwo sprawdzić, że $h(a) = h(b)$. Wynika stąd, że istnieje punkt $c \in (a,b)$, w kórym funkcja $h(t)$ ma maksimum lub minimum. Zatem z twierdzenia 32 $h'(c) = 0$, tj.

$$0 = (f(b)-f(a))g'(c) - (g(b)-g(a))f'(c),$$

czyli dostaliśmy żądany wynik. □

Szczególnym przypadkiem jest twierdzenie Lagrange'a o wartości średniej.

Twierdzenie 34. Załóżmy, że $f : [a,b] \to \mathbb{R}$ jest ciągła w $[a,b]$ i różniczkowalna w (a,b). Wtedy istnieje takie $c \in (a,b)$, że

$$\frac{f(b)-f(a)}{b-a} = f'(c)$$

Dowód. W poprzednim twierdzeniu wystarczy przyjąć $g(x) = x$. □

Przedstawimy teraz kilka zasadniczych zastosowań twierdzeń o wartości średniej. Zaczniemy od badania przebiegu funkcji.

Twierdzenie 35. Załóżmy, że funkcja $f : (a,b) \to \mathbb{R}$ jest różniczkowalna w (a,b). Wtedy
(a) f jest rosnąca w $(a,b) \Leftrightarrow$ dla każdego $x \in (a,b)$, mamy $f(x) \geq 0$;
(b) f jest malejąca w $(a,b) \Leftrightarrow$ dla każdego $x \in (a,b)$, mamy $f(x) \leq 0$;
(c) f jest stała w $(a,b) \Leftrightarrow$ dla każdego $x \in (a,b)$ mamy $f'(x) = 0$.

Dowód. (a) $\Rightarrow$ Skoro dla $x_1 < x_2$ mamy, że $f(x_1) \leq f(x_2)$, to dla $x_2 = x_1 + h$ otrzymamy, dzięki różniczkowalności f w x_1, że

$$0 \leq \lim_{h\to 0^+} \frac{f(x_1+h)-f(x_2)}{h} = \lim_{h\to 0} \frac{f(x_1+h)-f(x_2)}{h} = f'(x_1).$$

$\Leftarrow$ Niech $x_1 < x_2$, wtedy na mocy twierdzenia o wartości średniej dostaniemy, że

$$f(x_2) - f(x_1) = (x_2 - x_1)f'(c),$$

dla pewnego $c \in (x_1, x_2)$. Zatem skoro $f'(x) \geq 0$, dla $x \in (a, b)$, to

$$f(x_2) - f(x_1) \geq 0.$$

(b) Ta część wynika z punktu (a) zastosowanego do $-f$.

(c) $\Leftarrow$ skoro $f(x_2) - f(x_1) = f'(c)(x_2 - x_1)$, to $f'(c) = 0$ pociąga $f(x_2) - f(x_1) = 0$.
$\Rightarrow$ było treścią przykładu 7(a). □

Innym ważkim i często nadużywanym wnioskiem jest narzędzie do obliczania granic.

Twierdzenie 36. (Reguła de l'Hospitala) Niech f i g będą funkcjami ciągłymi w przedziale (a, b) (nie wykluczamy możliwości, że $b = \infty$, $a = -\infty$) i różniczkowalnymi w (a, b). Jeśli

$$\lim_{x\to a} \frac{f'(x)}{g'(x)} = A$$

i

$$\lim_{x\to a} f(x) = 0 = \lim_{x\to a} g(x), \tag{6}$$

to

$$\lim_{x\to a} \frac{f(x)}{g(x)} = A.$$

Otrzymamy tę samą tezę, jeśli zamiast (6) przyjmiemy

$$\lim_{x\to a} f(x) = \infty = \lim_{x\to a} g(x).$$

Dowód. Rozpatrzymy wyłącznie przypadek (6), gdy $a \in \mathbb{R}$ i $A \in \mathbb{R}$. Skoro $f'(x)/g'(x) \to A$, to dla dowolnego $q > A$ istnieje r, takie że $A < r < q$. Z definicji granicy istnieje takie ${}_1c > a$, że $f'(t)/g'(t) < r$ dla $a < t < c_1$. Zatem z twierdzenia 33 dostaniemy, dla dowolnych $x, y \in \mathbb{R}$, takich że $a < x$, $y < c_1$

$$\frac{f(x)-f(y)}{g(x)-g(y)} = \frac{f'(t)}{g'(t)} < r \tag{7}$$

Skoro $\lim_{x\to a} f(x) = 0 = \lim_{x\to b} g(x)$, to możemy w (7) przejść do granicy z y. Dostaniemy wtedy,

$$\frac{f(x)}{g(x)} = \lim_{y\to a} \frac{f(x)-f(y)}{g(x)-g(y)} \leq r < q \text{ dla } \; x \in (a_1, c_1)$$

podobny argument doprowadzi nas do wniosku, że dla dowolnego $p < A$ dostaniemy, że

$$p < \frac{f(x)}{g(x)} < q \text{ gdy } x \in (a, c_2),$$

dla pewnego $c_2 > a$; tj. $\lim_{x\to a} \frac{f(x)}{g(x)} = A$. □

Ostrzeżenie: reguły de l'Hospitala nie należy nadużywać, np. aby obliczyć

$$\lim_{x\to 0} \frac{\sin x}{x}$$

wystarczy zauważyć, że mamy do czynienia z definicją pochodnej funkcji $\sin x$ w $x = 0$, tj. $\lim_{x\to 0} \frac{\sin x}{x} = \cos 0 = 1$.

Zajmiemy się teraz różniczkowaniem funkcji odwrotnej.

Twierdzenie 37. Załóżmy, że funkcja $f : (a, b) \to \mathbb{R}$ jest różnowartościowa i na zbiór $E = f(a, b)$. Załóżmy, że f jest różniczkowalna w punkcie $x_0 \in (a, b)$ zaś funkcja odwrotna do f jest ciągła w punkcie $y_0 = f(x_0)$. Wtedy funkcja odwrotna f^{-1} jest różniczkowalna w punkcie $f(x_0)$:

$$\frac{d}{dy} f^{-1}(f(y_0)) = \frac{1}{\frac{d}{dx} f(f^{-1}(y_0))}$$

Dowód. Korzystamy z definicji pochodnej

$$\lim_{h\to 0} \frac{f^{-1}(y_0 + h) - f^{-1}(y_0)}{h},$$

ale $y_0 = f(x_0)$ i możemy napisać $h = f(x_0+H)-f(x_0)$. Skoro f^{-1} jest ciągła, to ze zbieżności h do zera wynika, że H dąży do zera. Mamy teraz

$$\begin{aligned}\lim_{h\to 0} \frac{f^{-1}(y_0 + h) - f^{-1}(y_0)}{h} &= \lim_{H\to 0} \frac{f^{-1}(f(x_0 + H)) - f^{-1}(f(x_0))}{f(x_0 + H) - f(x_0)} \\ &= \lim_{H\to 0} \frac{x_0 + H - x_0}{f(x_0 + H) - f(x_0)} = \frac{1}{f'(x_0)} = \frac{1}{f'(f^{-1}(y_0))}.\end{aligned}$$

□

Poniżej przekonamy się, jak bardzo to twierdzenie upraszcza nam rachunki.

Przykład 8. Niech $f : (0, \infty) \to (0, \infty)$, będzie dana wzorem $f(x) = x^n$, $n \geq 1$, $n \in \mathbb{N}$, wtedy $f^{-1}(y) = y^{\frac{1}{n}}$. Poprzednie twierdzenie daje nam,

$$\frac{d}{dy} f^{-1}(y) = \frac{1}{\frac{d}{dx} f(f^{-1}(y))} = \frac{1}{n(f^{-1}(y))^{n-1}} = \frac{1}{n} y^{\frac{1}{n}-1}.$$

3.5 Twierdzenie Taylora i pochodne wyższych rzędów

Przypuśćmy, że funkcja f określona na (a,b) o wartościach rzeczywistych jest różniczkowalna w (a,b). Możemy zastanowić się czy funkcja pochodna f' jest różniczkowalna w (a,b). Jeśli tak, to pochodna pochodnej $(f')'$ nazywa się *drugą pochodną* i piszemy

$$(f')' \equiv f''$$

Możemy więc indukcyjnie określić pochodną rzędu $(n+1)$-szego jako pochodną n-tej pochodnej. Piszemy więc

$$f^{(n+1)}(x) = (f^{(n)})'(x)$$

tj. umiemy określić pochodną trzecią f''', czwartą $f^{(iv)}$ itp. Inny, równoważny zapis to

$$f^{(n)}(x) \equiv \frac{d^n f}{dx^n}(x)$$

.

3.5.1 Interpretacje fizyczne wyższych pochodnych

Przekonajmy się, że wyższe pochodne, np. drugiego rzędu mogą mieć przejrzyste interpretacje. Wiadomo, że siła równa się pochodnej pędu po czasie

$$F = \frac{dp}{dt}.$$

Jeśli założymy, że interesuje nas punkt materialny, którego masa nie zmienia się w czasie, to dostaniemy

$$\frac{dp}{dt} = \frac{d}{dt}(mv) = m\frac{dv}{dt}$$

a skoro $v = \frac{ds}{dt}$, gdzie s oznacza drogę, to

$$F = m\frac{ds^2}{dt^2}.$$

3.5.2 Interpretacje geometryczne

Powiemy, że funkcja $f : (a,b) \to \mathbb{R}$ jest *wypukła* (odpowiednio, *wklęsła*), jeśli dla dowolnych $x, y \in (a,b)$ i $t \in [0,1]$ jest prawdą, że

$$f(tx+(1-t)y) \geq tf(x)+(1-t)f(y) \quad \text{(odpowiednio, } f(tx+(1-t)y) \leq tf(x)+(1-t)f(y)),$$

tj. cięciwa leży nad (odpowiednio, pod) wykresem.

Geometryczną właściwość, jaką jest wypukłość można scharakteryzować różniczkowo.

Twierdzenie 38. Niech funkcja $f : (a,b) \to \mathbb{R}$ będzie dwukrotnie różniczkowalna w (a,b). Wtedy f jest wypukła w (a,b) (odpowiednio, wklęsła) wtedy i tylko wtedy, gdy $f''(x) \geq 0$ dla wszystkich $x \in (a,b)$ (odpowiednio, $f''(x) \leq 0$ dla wszystkich $x \in (a,b)$). □

Powyższy fakt pozostawiamy bez dowodu. Jest on użyteczny przy badaniu przebiegu funkcji.

3.5.3 Twierdzenie Taylora

Przedstawimy tytułowe twierdzenie a potem niektóre jego zastosowania. Załóżmy, że $f : (a, b) \to \mathbb{R}$ jest n-krotnie różniczkowalna w (a, b) i $y, x \in (a, b)$. Kładziemy

$$P_n(y) = \sum_{k=0}^{n} f^{(k)}(x)\frac{(y-x)^k}{k!}.$$

Uwaga. Piszemy $f^{(0)}(x)$ zamiast $f(x)$.

Twierdzenie 39. (Taylora) Załóżmy, że $f : (a, b) \to \mathbb{R}$ jest $(n + 1)$ - krotnie różniczkowalna w (a, b). Wtedy dla dowolnych x i y z przedziału (a, b) istnieje c pomiędzy x i y takie, że

$$f(y) = P_n(y) + \frac{f^{(n+1)}(c)}{(n+1)!}(y-x)^{n+1}$$

Uwaga. Jeśli $n = 0$, to jest to znane twierdzenie o wartości średniej.

Dowód. Dla zadanych x i y dobieramy liczbę M taką, że

$$f(y) = P_n(y) + M(y-x)^{n+1}.$$

Kładziemy

$$g(t) = f(t) - P_n(t) - M(t-x)^{(n+1)}.$$

Trzeba wykazać, że $M = \frac{f^{(n+1)}}{(n+1)!}(c)$ dla pewnego c. Ponieważ $g^{(n+1)}(t) = f^{(n+1)}(t) - (n+1)!M$ to znaczy, że c ma spełniać, $g^{(n+1)}(c) = 0$. Wiemy, że $P_n^{(k)}(x) = f^{(k)}(x)$ dla $k = 0, 1, \ldots, n$, zatem

$$g(x) = g'(x) = g''(x) = \ldots = g^{(n)}(x) = 0.$$

Ponieważ $0 = g(x) = g(y)$, to z twierdzenia 34 wynika istnienie x_1 pomiędzy x i y takiego, że

$$g(y) - g(x) = g'(x_1)(y-x) \text{ tj. } g'(x_1) = 0$$

Teraz $g'(x) = 0 = g'(x_1)$ zatem stosując ponownie twierdzenie 33 znajdujemy x_2 pomiędzy x i x_1, takie że

$$g'(x_1) - g'(x) = g''(x_2)(x_1 - x) \quad \text{tj.} \quad g'(x_2) = 0$$

Postępując w ten sposób, po n krokach znajdziemy x_n, takie że

$$g^{(n)}(x_n) = 0.$$

A skoro $g^{(n)}(x) = 0$, to

$$g^{(n)}(x_n) - g^{(n)}(x) = g^{(n+1)}(c)(x_n - x)$$

dla pewnego c pomiędzy x i x_n, tj.

$$g^{(n+1)}(c) = 0$$

co należało wykazać. □

Jednak w zastosowaniach bywa tak, że brakuje nam wiedzy o wyższej różniczkowalności. Tym nie mniej prawdziwe jest odpowiednie twierdzenie Taylora, które wykorzystamy do badania ekstremów lokalnych. Mając na uwadze to zastosowanie sformułujemy nasz wynik tylko w szczególnej postaci.

Stwierdzenie 40. Załóżmy, że $f : (a, b) \to \mathbb{R}$ jest dwukrotnie różniczkowalna w (a, b) i $f'(x_0) = f''(x_0) = 0$. Wtedy,

$$\lim_{x\to x_0} \frac{f(x) - f(x_0)}{(x - x_0)^2} = 0.$$

Dowód. Dwukrotne zastosowanie reguły de l'Hospitala (twierdzenie 35) daje nam

$$\lim_{x\to x_0} \frac{f(x) - f(x_0)}{(x - x_0)^2} = \lim_{x\to x_0} \frac{f'(x)}{2(x - x_0)} = f''(x_0) = 0.$$ □

Wypływa stąd prosty wiosek

Wniosek 41. (Taylora) Załóżmy, że $f : (a, b) \to \mathbb{R}$ jest dwukrotnie różniczkowalna w (a, b). Wtedy,

$$f(x) - f(x_0) = f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + r(x),$$

gdzie

$$\lim_{x\to x_0} \frac{r(x)}{(x - x_0)^2} = 0.$$

Dowód. Wprowadzamy funkcję pomocniczą g, daną wzorem:

$$g(x) = f(x) - f(x_0) - f'(x_0)(x - x_0) - \frac{1}{2}f''(x_0)(x - x_0)^2.$$

Wtedy g spełnia założenia poprzedniego stwierdzenia, bo

$$g'(x_0) = f'(x_0) - f'(x_0) = 0, \quad g''(x_0) = \frac{1}{2}f''(x_0) - \frac{1}{2}f''(x_0) = 0.$$

Zatem

$$\lim_{x\to x_0} \frac{g(x) - g(x_0)}{(x - x_0)^2} = 0,$$

a to oznacza prawdziwość naszej tezy. □

3.5.4 Zastosowania do obliczeń przybliżonych

Chcemy obliczyć wartość $\sin 0{,}1$ z dokładnością 0,001. Definicja pochodnej podsuwa nam, że

$$\sin 0{,}1 = 0{,}1 + \text{ błąd.}$$

Chcemy ustalić jaki jest ów „błąd". Do tego celu posłuży nam twierdzenie 39. Jeśli przyjmiemy, że $(\sin x)' = \cos x$ i $(\cos x)' = -\sin x$ to dostaniemy, że

$$\sin x = x + \sin c \cdot \frac{x^3}{3!}$$

gdzie $x = 0,1$ i $c \in (0,\ 0{,}1)$, bo

$$\frac{d}{dx}\sin x|_{x=0} = \cos 0 = 1, \qquad \frac{d^2}{dx^2}\sin x|_{x=0} = -\sin 0 = 0.$$

Skoro $|\sin c| \leq 1$, to $\frac{\sin c(0,1)^3}{3!} \leq \frac{0{,}001}{6}$, więc $\sin x = x$ z dokładnością lepszą niż 0,001.

3.5.5 Różniczkowa charakteryzacja ekstremów lokalnych

W praktyce często interesuje nas znalezienie wartości największej i najmniejszej funkcji różniczkowalnej $f : [a, b] \to \mathbb{R}$. Często wystarczy znaleźć miejsca zerowe pochodnej, np. jest to zbiór $\{x_1, \ldots, x_m\}$. Wtedy zagadnienie znalezienia maksimum upraszcza się, bo na mocy twierdzenia 32

$$\max_{x\in[a,b]} f(x) = \max_{x\in\{x_1,\ldots,x_m,a,b\}} f(x).$$

Dodaliśmy końce przedziałów, bo twierdzenie 32 nie jest w nich spełnione a funkcja może w nich osiągać maksimum, bądź minimum, (patrz rys. 3). Jednak sama wiedza, że $f'(x_0) = 0$ nie wystarcza do rozstrzygnięcia czy jest to ekstremum i jaki jest jego charakter. Do tego celu posłużymy się drugą wersją twierdzenia Taylora.

Twierdzenie 42. Niech funkcja. $f : (a, b) \to \mathbb{R}$ będzie dwukrotnie różniczkowalna w (a, b) i $x_0 \in (a, b)$.

(a) (warunek konieczny ekstremum) Jeśli f ma maksimum (odpowiednio: minimum) lokalne w punkcie x_0, to $f'(x_0) = 0$ i $f''(x_0) \leq 0$ (odpowiednio: $f''(x_0) \geq 0$).

(b) (warunek dostateczny ekstremum) Jeśli $f'(x_0) = 0$ i $f''(x_0) < 0$ (odpowiednio: $f''(x_0) > 0$), to f ma w punkcie x_0 maksimum (odpowiednio: minimum) lokalne.

Dowód. (a) Rozpatrzymy wyłącznie przypadek maksimum, bo przypadek minimum uzyskujemy zamianą f na $-f$.

Wykażemy, że $f''(x_0) \leq 0$. Z wniosku 41 i istnienia maksimum w x_0 mamy

$$0 \geq f(x) - f(x_0) = f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + r(x).$$

Po podzieleniu przez $(x-x_0)^2$ dostaniemy:

$$0 \geq \frac{f(x)-f(x_0)}{(x-x_0)^2} = \frac{1}{2}f''(x_0) + \frac{r(x)}{(x-x_0)^2}.$$

Z twierdzenia 39 ostatni wyraz dąży do zera, gdy $x \to x_0$, zatem $0 \geq f''(x_0)$.

(b) Ponownie z twierdzenia 39 dostaniemy:

$$f(x)-f(x_0) = 0 + \frac{1}{2}f''(x_0)(x-x_0)^2 + r(x) = (x-x_0)^2\left(\frac{1}{2}f''(x_0) + r(x)/(x-x_0)^2\right),$$

a skoro $r(x)/(x-x_0)^2$ dąży do zera, gdy $x \to x_0$, to dla dostatecznie małych $|x-x_0|$ dostaniemy, że $|r(x)/(x-x_0)^2| < |f''(x_0)|/4$, tj.

$$f(x)-f(x_0) \leq \frac{(x-x_0)^2}{4}(2f''(x_0) + |f''(x_0)|) < 0.$$

Zatem w punkcie x_0 funkcja f ma maksimum. □

Uwaga. Nie można poprawić założeń, a mianowicie:

a) $f(x) = 1-x^4$ ma maksimum w punkcie $x=0$, ale $f''(0)=0$;

b) $g(x) = x^3$, to $g''(0)=0$, ale g nie ma ekstremum w punkcie $x=0$.

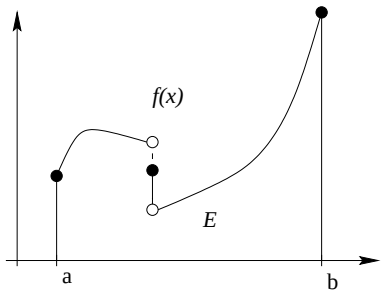
3.6 Całka Riemanna

Chcemy nauczyć się obliczać pole powierzchni złożonych figur płaskich. Punktem wyjścia jest umiejętność obliczenia pola prostokąta. Wiemy już też, że pole równoległoboku $R(a,b)$ wyraża się wzorem

$$\text{pole}\,(R(a,b)) = |\det(a,b)|.$$

Równie dobrze moglibyśmy przyjąć powyższą równość jako definicję i przekonać się, że jeśli wektory a i b są prostopadłe, tj. $R(a,b)$ jest prostokątem, to

$$\text{pole}\,(R(a,b)) = |a| \cdot |b|.$$

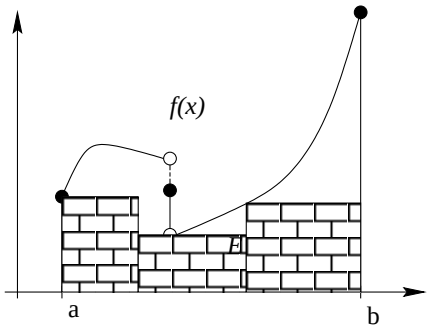


Rys. 4. Podwykres E.

Zajmiemy się znalezieniem pola podwykresu. Jeśli dana jest funkcja $f : [a, b] \to \mathbb{R}$, to kładziemy $E = \{(x, y) \in \mathbb{R}^2;\ y \leq f(x), x \in [a, b]\}$ i zbiór E nazywamy *podwykresem* f.

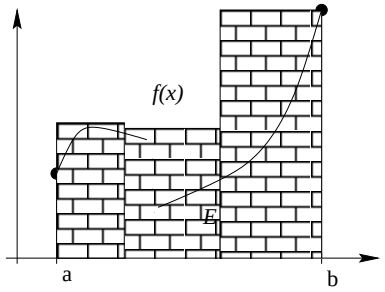
Będziemy przybliżali pole zbioru E. Można robić to na 3 sposoby, które od razu zilustrujemy:

(a) nie doceniając go (choć to określenie ma zastosowanie tylko, gdy $f \geq 0$); przybliżamy zbiór E prostokątami, które całkowicie mieszczą się **pod** wykresem funkcji f;



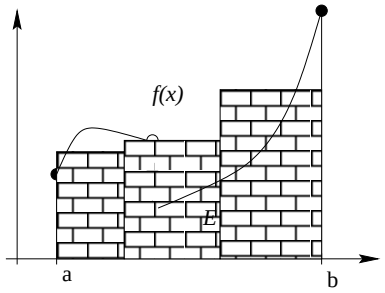
Rys. 5 a. Sposób (a).

(b) przeceniając je (z zastrzeżeniem j.w.); przybliżamy zbiór E prostokątami, które całkowicie mieszczą **w sobie** zbiór E;



Rys. 5 b. Sposób (b).

(c) niechlujnie; jest to wariant pośredni, bezplanowy.



Rys. 5 c. Sposób (c).

Spodziewamy się, że zmniejszając podstawy prostokątów uzyskamy coraz to lepsze przybliżenia, zaś dla ‘dobrych’ funkcji wynik nie powinien zależeć od sposobu przybliżania, tj. trzeba ustalić czym są owe ‘dobre’ funkcje.

Aby opisać starannie powyższe sposoby przybliżania zbioru E potrzebne są nam definicje.

Definicja 19. Niech będzie dany przedział domknięty $[a, b]$. *Podziałem* P przedziału $[a, b]$ nazywamy skończony zbiór punktów tego przedziału

$$a = x_0 \leq x_1 \leq x_2 \ldots \leq x_n = b.$$

Piszemy też $\Delta x_i = x_i - x_{i-1}$, $i = 1, \ldots, n$. *Średnicą podziału* nazywamy liczbę

$$\delta(P) := \max_{i=1,\ldots,n} \Delta x_i.$$

Kładziemy

$$M_i = \sup_{x\in[x_{i-1},x_i]} f(x), \quad m_i = \inf_{x\in[x_{i-1},x_i]} f(x)$$

i

$$L(P, f) := \sum_{n=1}^{n} m_i \Delta x_i, \quad U(P, f) := \sum_{i=i}^{n} M_i \Delta x_i.$$

Zauważmy, że $L(P, f)$ odpowiada sposobowi (a), zaś $U(P, f)$ odpowiada sposobowi (b). Sposobem (c) zajmiemy się później.

Jest rzeczą oczywistą, wynikającą bezpośrednio z definicji, że

$$L(P, f) \leq U(P, f).$$

Szukając najlepszego oszacowania z dołu odpowiadającego sposobowi (a) i najlepszego oszacowania z góry odpowiadającego sposobowi (b), wprowadzamy kolejne definicje.

Definicja 20.

$$\overline{\int_a^b} f = \inf_P U(P, f), \qquad \underline{\int_a^b} f = \sup_P L(P, f)$$

gdzie kresy brane są po wszystkich możliwych podziałach P przedziału $[a, b]$. Liczbę $\underline{\int_a^b} f$ nazywamy *dolną całką Riemanna*, zaś $\overline{\int_a^b} f$, to *górna całka Riemanna*.

Podstawowe pytanie teraz brzmi,

$$\text{czy} \quad \underline{\int_a^b} f = \overline{\int_a^b} f \ ? \tag{P}$$

Aby je zbadać rozpatrzmy przykład.

Przykład 9. Załóżmy, że funkcja $g : [0, 1] \to \mathbb{R}$ dana jest wzorem $g(x) = \chi_{\mathbb{Q}\cap[0,1]}$. Wtedy dla dowolnego podziału P odcinka $[0, 1]$ mamy, że

$$L(P, g) = 0 \quad \text{zaś} \quad U(P, g) = 1.$$

Skoro tak łatwo jest podać przykład funkcji takiej, że

$$\underline{\int_a^b} g < \overline{\int_a^b} g,$$

to nasze pytanie (P) nabiera ostrości:

Kiedy dolna całka Riemanna równa się górnej całce Riemanna ?

Udzielimy dwu odpowiedzi na to pytanie:
(i) łatwej, ale niepełnej;
(ii) trudnej, za to pełnej.
Zaczniemy od (i) wprowadzając definicję.

Definicja 21. Niech $f : [a, b] \to \mathbb{R}$, powiemy, że funkcja f spełnia *warunek Lipschitza ze stałą* K, jeśli

$$|f(x) - f(y)| \leq K|x - y|, \text{ dla wszystkich } x, y \in [a, b].$$

Okazuje się, że mamy:

Stwierdzenie 43. Załóżmy, że $f : [a, b] \to \mathbb{R}$ spełnia warunek Lipschitza ze stałą K. Wtedy f jest ciągła w $[a, b]$.

Dowód. Niech $x \in [a, b]$ i $\varepsilon > 0$ będzie dowolne, trzeba dobrać $\delta > 0$ z definicji ciągłości w punkcie. Przyjmijmy $\delta = \frac{\varepsilon}{K}$. Mamy wtedy z założenia, że

$$|f(x) - f(y)| \leq K|x - y| < K \cdot \delta = \frac{K\varepsilon}{K} = \varepsilon,$$

dla $y \in [a, b]$ spełniających $|x - y| < \delta$. Co należało wykazać. □

Wprowadzimy teraz ważne określenie.

Definicja 22. Funkcja $f : [a, b] \to \mathbb{R}$ nazywa się *całkowalną w sensie Riemanna*, jeśli $\overline{\int_a^b} f = \underline{\int_a^b} f$. Piszemy wtedy $\int_a^b f(x)dx := \overline{\int_a^b} f \equiv \underline{\int_a^b} f$. Zbiór funkcji całkowalnych w sensie Riemanna określonych na odcinku $[a, b]$ oznacza się symbolem $\mathcal{R}(a, b)$.

Teraz łatwa odpowiedź na pytanie (P) jest następująca.

Twierdzenie 44. Jeśli $f : [a, b] \to \mathbb{R}$ spełnia warunek Lipschitza ze stałą K, to f jest całkowalna w sensie Riemanna. Dodatkowo, jeśli P_n jest takim ciągiem podziałów, że $\delta(P_n) \to 0$, to

$$\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{i=1}^{k_n} f(\xi_i^n)\Delta x_i^n \tag{8}$$

gdzie $P_n = \{x_0^n, x_1^n, \ldots, x_{k_n}^n\}$, $\xi_i^n \in [x_{i-1}^n, x_i^n]$ i ξ_i^n są wybrane dowolnie.

Uwaga. W przypadku funkcji spełniającej warunek Lipschitza wszystkie 3 sposoby przybliżania pola podwykresu E są równoważne.

Aby ułatwić wysłowienie dowodu powyższego twierdzenie wprowadzimy definicję.

Definicja 23. Powiemy, że podział P^* przedziału jest rozdrobnieniem podziału P, jeśli $P^* \supset P$.

Dowód twierdzenia. Niech $\varepsilon > 0$ będzie dowolne. Wybieramy takie podziały P_1, P_2, że

$$\left|\overline{\int_a^b} f - U(P_1, f)\right| < \frac{\varepsilon}{2} \quad \text{oraz} \quad \left|\underline{\int_a^b} f - L(P_2, f)\right| < \frac{\varepsilon}{2}.$$

Zauważmy, że jeśli P jest dowolnym wspólnym rozdrobnieniem podziałów P_1 i P_2, to

$$U(P,f) \leq U(P_1,f) \quad \text{oraz} \quad L(P,f) \geq L(P_2,f).$$

Przyjmujemy $P = P_1 \cup P_2$, wprowadzamy dodatkowy podpodział P^* podziału, tak aby $\delta(P^*) \leq \varepsilon/K(b-a)$, gdzie K jest stałą występującą w warunku Lipschitza.

Teraz nasze oszacowanie różnicy $U(P^*,f) - L(P^*,f)$ przebiega następująco. Liczymy

$$U(P^*,f) - L(P^*,f) = \sum_{i=1}^{k} \Delta x_i (M_i - m_i) = I,$$

gdzie dzięki ciągłości funkcji f, $M_i = f(\zeta_i)$, $m_i = f(\eta_i)$ dla pewnych $\zeta_i, \eta_i \in [x_{i-1}, x_i]$. Tym samym,

$$I = \sum_{i=1}^{k} \Delta x_i (f(\zeta_i) - f(\eta_i)).$$

Dalej, warunek Lipschitza pociąga za sobą

$$|I| \leq \sum_{i=1}^{k} \Delta x_i K(\zeta_i - \eta_i) \leq \sum_{i=1}^{k} \Delta x_i K \Delta x_i$$

$$\leq \sum_{i=1}^{k} \Delta x_i K \delta(P^*) = K(b-a)\delta(P^*) \leq \varepsilon, \tag{9}$$

Z definicji kresów mamy też

$$L(P^*,f) \leq \underline{\int_a^b} f \leq \overline{\int_a^b} f \leq U(P^*,f). \tag{10}$$

Zatem na mocy (9)

$$\overline{\int_a^b} f - \underline{\int_a^b} f \leq U(P^*,f) - L(P^*,f) \leq \varepsilon$$

tj.

$$\overline{\int_a^b} f \leq \underline{\int_a^b} f + \varepsilon,$$

a skoro $\varepsilon > 0$ było dowolne, to wynika stąd, że

$$\overline{\int_a^b} f \leq \underline{\int_a^b} f.$$

Zatem biorąc pod uwagę (10) dostaniemy, iż całka górna jest równa dolnej, tj. f jest całkowalna w sensie Riemanna.

Pozostaje teraz udowodnić, że $\int_a^b f(x)dx$ jest granicą sum. Niech P_n będzie dowolnym ciągiem podziałów takim, że $\delta(P_n) \to 0$ i $\zeta_i^n \in [x_{i-1}^n, x_i^n]$, $i = 1, \ldots, k_n$. Niech

$$S_n = \sum_{i=1}^{k_n} \Delta x_i f(\zeta_i^n)$$

i $\varepsilon > 0$ będzie dowolne. Wtedy istnieje N takie, że dla $n \geq N$, mamy $\delta(P_n) < \frac{\varepsilon}{K(b-a)}$. Pokażemy, że $|\int_a^b f - S_n| \leq 2\varepsilon$. Oczywiście

$$L(P_n, f) \leq S_n \leq U(P_n, f)$$

zaś (9) daje, że

$$U(P_n, f) - S_n \leq U(P_n, f) - L(P_n, f) \leq \varepsilon,$$

tj.

$$\begin{aligned} |\int_a^b f(x) - S_n| &= |\int_a^b f(x)dx - U(P_n, f) + U(P_n, f) - S_n| \\ &\leq |\int_a^b f(x)dx - U(P_n, f)| + U(P_n, f) - S_n \\ &= U(P_n, f) - \underline{\int_a^b} f + U(P_n, f) - S_n \\ &\leq U(P_n, f) - L(P_n, f) + U(P_n, f) - S_n \\ &\leq 2(U(P_n, f) - L(P_n, f)) \leq 2\varepsilon. \end{aligned}$$

Wynika stąd, że istotnie

$$\int_a^b f(x)dx = \lim_{n\to\infty} S_n \equiv \lim_{n\to\infty} \sum_{i=1}^{k_n} \Delta x_i f(\zeta_i^n),$$

co należało wykazać. □

Możemy teraz zasygnalizować trudną odpowiedź. Zaczniemy od definicji.

Definicja 24. Załóżmy, że $N \subset \mathbb{R}$. Powiemy, że *zbiór N jest miary zero*, jeśli dla dowolnego $\varepsilon > 0$ istnieje taki ciąg otwartych przedziałów $I_i = (a_i, b_i)$, $i = 1, \ldots$, że

$$N \subset \bigcup_{i=1}^{\infty} I_i \quad \text{i} \quad \sum_{i=1}^{\infty} (b_i - a_i) < \varepsilon. \tag{11}$$

Piszemy wtedy $\mu(N) = 0$.

Przykład 10. Zbiór liczb wymiernych ma miarę zero. Wszystkie liczby wymierne można ustawić w ciąg $\{r_i\}_{i=1}^{\infty}$. Niech $\varepsilon > 0$ będzie dowolne. Wtedy kładziemy,

$$I_i = (r_i - \frac{\varepsilon}{2^{i+1}},\ r_i + \frac{\varepsilon}{2^{i+1}}).$$

Pierwsza część warunku (11) jest oczywiście spełniona. Zauważmy, że każdy przedział I_i ma długość $\frac{\varepsilon}{2^i}$ a zatem $\sum_{i=1}^{\infty} \frac{\varepsilon}{2^i} = \varepsilon$, tj. warunek (11_2) też jest spełniony.

Uwaga. Istnieją bardziej złożone przykłady zbiorów miary zero.

Możemy teraz sformułować pełną charakteryzację funkcji całkowalnych w sensie Riemanna.

Twierdzenie 45. Funkcja $f : [a,b] \to \mathbb{R}$ jest całkowalna w sensie Riemanna wtedy i tylko wtedy, gdy

(1) funkcja f jest ograniczona;

(2) funkcja f jest ciągła w każdym punkcie zbioru $[a,b]\backslash N$, gdzie $\mu(N) = 0$ (tj. N jest miary zero).

Pozostawimy to trudne twierdzenie bez dowodu. □

Wniosek 46. (1) funkcje ciągłe są całkowalne, bo $N = \emptyset$;
(2) funkcje monotoniczne i ograniczone są ciągłe, bo zbiór N jest co najwyżej przeliczalny tj. ma miarę zero (patrz twierdzenie 28);
(3) funkcja $g : [0,1] \to \mathbb{R}$, $g(x) = \chi_{\mathbb{Q}\cap[0,1]}(x)$ nie jest całkowalna w sensie Riemanna.

Mamy więc już elegancką charakteryzację funkcji całkowalnych w sensie Riemanna. Jednak jest ona dość trudna i nie całkiem poręczna. Dlatego dowody twierdzeń będziemy przedstawiali dla szczególnych funkcji całkowalnych, a mianowicie takich, że

$$f : [a,b] \to \mathbb{R} \text{ jest ograniczona i ma najwyżej skończenie wiele punktów nieciągłości.} \quad (U)$$

Wtedy wykład staje się prostszy. Zaczniemy od podstawowych właściwości elementów zbioru $\mathcal{R}(a,b)$, tj. funkcji całkowalnych w sensie Riemanna na odcinku $[a,b]$.

Twierdzenie 47. Niech $f, g \in \mathcal{R}(a,b)$ i $\alpha \in \mathbb{R}$. Wtedy,
(a) $f + g \in \mathcal{R}(a,b)$, $\alpha f \in \mathcal{R}(a,b)$, ponadto

$$\int_a^b (f(x) + g(x))\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx \quad \text{i} \quad \int_a^b \alpha f(x)\,dx = \alpha \int_a^b f(x)\,dx$$

(b) jeśli $f \le g$, to $\int_a^b f(x)\,dx \le \int_a^b g(x)\,dx$;
(c) jeśli $c \in (a,b)$, to

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$$

(d) $\int_a^b 1\,dx = b - a$.

Dowód przeprowadzimy dla funkcji f i g spełniających (U). Oczywiście, jeśli f i g są ograniczone i ciągłe z wyjątkiem skończenie wielu punktów, to podobnie $f + g$ spełnia (U), tak samo jest z αf. Wykażemy pierwszy wzór z (a). Przyjmiemy oznaczenia jak we wzorze (8) i dostaniemy z niego i z właściwości granicy, że

$$\begin{aligned}\int_a^b (f(x) + g(x))\,dx &= \lim_{n\to\infty} \sum_{i=1}^{k_n} (f(\xi_i) + g(\xi_i))\Delta x_i \\ &= \lim_{n\to\infty} \sum_{i=1}^{k_n} f(\xi_i)\Delta x_i + \lim_{n\to\infty} \sum_{i=1}^{k_n} g(\xi_i)\Delta x_i \\ &= \int_a^b f(x)\,dx + \int_a^b g(x)\,dx.\end{aligned}$$

Punkty (b) i (d) zostawiamy Czytelnikowi.

Do wykazania (c) definiujemy

$$f_1(x) = \begin{cases} f(x) & \text{dla } x \in [a,c] \\ 0 & \text{dla } x \in (c,b] \end{cases}, \quad f_2(x) = \begin{cases} 0 & \text{dla } x \in [a,c] \\ f(x) & \text{dla } x \in (c,b] \end{cases}.$$

Wtedy $f_1, f_2 \in \mathcal{R}(a,b)$, ponadto $f = f_1 + f_2$ i stosujemy punkt (a), tj.

$$\int_a^b f(x)\,dx = \int_a^b f_1(x)\,dx + \int_a^b f_2(x)\,dx = \int_a^c f_1(x)\,dx + \int_c^b f_2(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

□

Następna właściwość jest uogólnieniem nierówności trójkąta.

Twierdzenie 48. Załóżmy, że $f \in \mathcal{R}(a,b)$, wtedy $|f| \in \mathcal{R}(a,b)$ i

$$\left| \int_a^b f(x)\,dx \right| \leq \int_a^b |f(x)|\,dx.$$

Dowód. Zauważmy, że skoro f spełnia (U), to $|f|$ też ma tę właściwość. Następnie korzystamy ze wzoru (8) i z nierówności trójkąta:

$$\begin{aligned} \left|\int_a^b f(x)\,dx\right| &= \lim_{n\to\infty} \left| \sum_{i=1}^{k_n} f(\xi_i)\Delta x_i \right| \\ &\leq \lim_{n\to\infty} \sum_{i=1}^{k_n} |f(\xi_i)|\Delta x_i \\ &= \int_a^b |f(x)|\,dx. \end{aligned}$$

□

Nim sformułujemy następny fakt wprowadzimy dogodną konwencję, pisząc

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx.$$

Twierdzenie 49. Załóżmy, że $f \in \mathcal{R}(a,b)$, wtedy funkcja $F : [a,b] \to \mathbb{R}$ nazywana *funkcją górnej granicy całkowania* i dana wzorem

$$F(x) = \int_a^x f(t)\,dt$$

jest ciągła w $[a,b]$, a nawet spełnia warunek Lipschitza ze stałą $M = \sup_{x\in[a,b]} f(x)$. Jeśli dodatkowo funkcja f jest ciągła w $x_0 \in (a,b)$, to F jest różniczkowalna w x_0 i mamy:

$$\frac{d}{dx}F(x) = f(x).$$

Dowód. Niech $x \in [a,b]$ będzie dowolnym punktem. Szacujemy różnicę $F(x) - F(y)$:

$$\begin{aligned}|F(x)-F(y)| &= \left|\int_a^x f(t)\,dt - \int_a^y f(t)\,dt\right| = \left|\int_a^y f(t)\,dt + \int_y^x f(t)\,dt - \int_a^y f(t)\,dt\right| \\ &= \left|\int_y^x f(t)\,dt\right|.\end{aligned}$$

Dzięki twierdzeniu 47 dostaniemy, że

$$|F(x)-F(y)| \le \int_{\min\{x,y\}}^{\max\{x,y\}} |f(t)|\,dt \le |y-x|M.$$

co oznacza, że F spełnia warunek Lipschitza ze stałą M, co na mocy twierdzenia 44 pociąga ciągłość.

Wykażemy różniczkowalność w x_0. Mamy, także dzięki twierdzeniu 47

$$\begin{aligned}\lim_{h\to 0}\frac{F(x_0+h)-F(x_0)}{h} = &\lim_{h\to 0}\frac{1}{h}\left(\int_a^{x_0+h} f(t)\,dt - \int_a^{x_0} f(t)\,dt\right) \\ &= \lim_{h\to 0}\left(\frac{1}{h}\left(\int_{x_0}^{x_0+h} f(t)\,dt - f(x_0) + f(x_0)\right)\right) \\ &= \lim_{h\to 0}\left(\frac{1}{h}\left(\int_{x_0}^{x_0+h} f(t)\,dt - hf(x_0)\right) + f(x_0)\right) \\ &= \lim_{h\to 0}\frac{1}{h}\int_{x_0}^{x_0+h}(f(t)-f(x_0))\,dt + f(x_0).\end{aligned}$$

Ponieważ f jest ciągła w x_0, to dla dowolnego $\varepsilon > 0$ istnieje takie $\delta > 0$, że mamy $|f(x) - f(x_0)| < \varepsilon$ dla $0 < |x - x_0| < \delta$. Zatem dla $0 < h < \delta$ dostaniemy

$$\frac{1}{h}\left|\int_{x_0}^{x_0+h}(f(t)-f(t_0))\,dt\right| \le \frac{h}{h}\varepsilon = \varepsilon.$$

Tym samym dostajemy tezę. □

Powyższe twierdzenie pozwala sformułować podstawowe narzędzie obliczania całek.

Twierdzenie 50. (Podstawowe twierdzenie rachunku różniczkowego i całkowego) Załóżmy, że $f : [a, b] \to \mathbb{R}$ jest ciągła. Wtedy istnieje funkcja F taka, że $F'(x) = f(x)$ i

$$\int_a^b f(x)\,dx = F(b) - F(a)(=: F|_a^b).$$

Funkcję F nazywamy *funkcją pierwotną* i często piszemy $F(x) = \int f(x)dx$.

Dowód. Istnienie F jest zagwarantowane dzięki ciągłości f i poprzedniemu twierdzeniu. Zauważmy, że

$$F(x) - \int_a^x f(t)\,dt = c.$$

Aby się o tym przekonać policzymy pochodną lewej strony:

$$\frac{d}{dx}\left(F(x) - \int_a^x f(t)\,dt\right) = f(x) - f(x) = 0.$$

Zatem pozostaje wyznaczenie stałej c. Mamy

$$F(x) - c = \int_a^x f(t)\,dt.$$

Lewa strona dla $x = a$ równa się $F(a) - c$, zaś prawa zero. Zatem $c = F(a)$, skąd wynika nasz wzór. □

Przykład obliczania funkcji pierwotnych

$$\int x^n\,dx = \frac{1}{n+1}x^{n+1} + C.$$

Jednak w praktyce potrzebne są bardziej wyszukane metody.

Wprowadzimy teraz bardzo ważne narzędzie obliczania całek.

Twierdzenie 51. (o całkowaniu przez części) Załóżmy, że f, g są ciągłe i różniczkowalne, zaś pochodne f', g' są całkowalne w sensie Riemanna. Wtedy $f'g, fg' \in \mathcal{R}(a,b)$ i mamy

$$\int_a^b f'(x)g(x)\,dx = -\int_a^b f(x)g'(x)\,dx + (fg)|_a^b.$$

Dowód. Całkowalność $f'g$ i fg' jest oczywista, a stąd wynika $(fg)' \in \mathcal{R}(a,b)$. Mamy wtedy

$$\int_a^b (fg)'(x)\,dx = fg|_a^b.$$

Zaś lewa strona to

$$\int_a^b (f'(x)g(x) + f(x)g'(x))\,dx,$$

skąd wynika żądany wzór. □

Przykład zastosowania twierdzenia 51 do obliczania całek oznaczonych. Zakładamy, że $\cos x' = -\sin x$, wtedy

$$\begin{aligned}\int_0^{\pi/2} x\sin x\,dx &= \int_0^{\pi/2} x\left(-\frac{d}{dx}\cos x\right)\,dx \\ &= \int_0^{\pi/2} \frac{dx}{dx}\cos x\,dx - x\cos x|_0^{\pi/2} \\ &= \int_0^{\pi/2} \cos x\,dx + 0 = \sin x|_0^{\pi/2} = 1.\end{aligned}$$

Innym podstawowym narzędziem obliczania całek jest następujący wynik.

Twierdzenie 52. (o całkowaniu przez podstawienie) Załóżmy, że funkcja $f : [c,d] \to \mathbb{R}$ jest ciągła i $\phi : [a,b] \to [c,d]$ jest ściśle rosnąca, „na" i jest różniczkowalna w (a,b) zaś ϕ' jest ograniczona. Wtedy

$$\int_c^d f(y)\,dy = \int_a^b f(\phi(x))\phi'(x)\,dx. \tag{12}$$

Dowód. Na mocy założeń istnieje F funkcja pierwotna funkcji f. Lewa strona przyjmuje wtedy postać

$$F(d) - F(c) = F(\phi(b)) - F(\phi(a)).$$

Zauważmy, że funkcja złożona $F(\phi(x))$ jest funkcją pierwotną $f(\phi(x))\phi'(x)$:

$$\frac{d}{dx}F(\phi(x)) = \frac{dF}{dy}|_{y=\phi(x)}\frac{d\phi(x)}{dx} = f(\phi(x))\phi'(x).$$

Zatem, prawa strona (12), to

$$F(\phi(b)) - F(\phi(a)),$$

czyli prawa strona jest równa lewej. □

Wykorzystajmy natychmiast nowo zdobytą wiedzę do obliczeń.

Przykład 11. Obliczymy wartość całki oznaczonej, sprawdzimy, że

$$\int_0^1 \sqrt{1-x^2}\,dx = \frac{\pi}{4}.$$

Podstawmy $x = \sin y$. Zauważmy, że $\sin' y = \cos y$ i $0 = \sin 0$, $1 = \sin \pi/2$, wtedy

$$\int_0^1 \sqrt{1-y^2}\,dy = \int_0^{\pi/2} \sqrt{1-\sin^2 y}\cos y\,dy = \int_0^{\pi/2} \cos^2 y\,dy.$$

Liczymy ostatnią całkę, podstawiamy $y = \pi/2 - t$ i mamy $\frac{dy}{dt} = -1$, zatem

$$\int_0^{\pi/2} \cos^2 y\,dy = -\int_{\pi/2}^0 \cos^2(\pi/2 - t)\,dt = \int_0^{\pi/2} \sin^2 t\,dt.$$

Dalej

$$\begin{aligned}\textstyle\int_0^{\pi/2} \cos^2 y\,dy &= \frac{1}{2}\left(\int_0^{\pi/2} \cos^2 y\,dy + \int_0^{\pi/2} \sin^2 y\,dy\right)\\ &= \frac{1}{2}\int_0^{\pi/2} (\cos^2 y + \sin^2 y)\,dy\\ &= \frac{1}{2}\int_0^{\pi/2} 1\,dy = \pi/4\end{aligned}$$

Paragraf o właściwościach całkowania zakończymy twierdzeniem, które jest wnioskiem z twierdzenia Lagrange'a.

Twierdzenie 53. (o wartości średniej) Niech $f : [a,b] \to \mathbb{R}$ będzie ciągła, wtedy istnieje takie $c \in (a,b)$, że

$$\int_a^b f(x)\,dx = f(c)(b-a)$$

Dowód. Istnieje funkcja pierwotna F, zatem z twierdzenia 34 zastosowanego do F wynika, że

$$\int_a^b f(x)\,dx = F(b) - F(a) = F'(c)(b-a) = f(c)(b-a).$$ □

3.7 Ciągi i szeregi funkcyjne

Przypominamy, że ciąg był definiowany jako funkcja

$$a : \mathbb{N} \to X,$$

gdzie zbiór X jest dowolny. Jeśli $X = \mathbb{R}$ lub $\mathbb{C}$, to mamy do czynienia z ciągiem liczbowym. Nic nie stoi na przeszkodzie by przyjąć, że $X = \mathcal{R}(a, b)$ lub $C[a, b]$ (jest to zbiór funkcji ciągłych na przedziale $[a, b]$ o wartościach rzeczywistych) lub ogólniej, że X jest pewnym zbiorem funkcji. W tych przypadkach mamy do czynienia z *ciągami funkcyjnymi*. Zastanówmy się nad zbieżnością ciągów funkcyjnych, co to w ogóle miałoby znaczyć? Rozpatrzmy przykład:

$$f_n(x) = x^n \text{ dla } x \in [0, 1]. \tag{13}$$

Dla każdego x z osobna powyższa definicja określa ciąg liczbowy, którego granicę łatwo zbadać:

$$\lim_{n\to\infty} f_n(x) = \begin{cases} 0, & \text{gdy } x \in [0, 1) \\ 1, & \text{gdy } x = 1. \end{cases}$$

Widzimy rzecz zaskakującą: ciąg funkcji ciągłych różniczkowalnych, a nawet nieskończenie różniczkowalnych, jest zbieżny w każdym punkcie przedziału $[0, 1]$, ale granica nie jest ciągła! Przyjrzyjmy się więc zastosowanemu pojęciu granicy.

3.7.1 Zbieżność jednostajna

Wyjaśnimy, że ciągi funkcyjne wymagają nowego pojęcia granicy. Dla naszej wygody od razu nadamy mu nazwę.

Definicja 25. Powiemy, że ciąg funkcji $f_n : D \to \mathbb{K}$, $D \subset \mathbb{K}^l$, jest zbieżny *punktowo* do funkcji $f : D \to \mathbb{K}$, jeśli:

• **dla każdego** $x \in D$, dla każdego $\varepsilon > 0$ istnieje takie N, że dla $n > N$ mamy

$$|f_n(x) - f(x)| < \varepsilon.$$

Zmieńmy nieco to określenie: przesuńmy wytłuszczone wyrażenie na koniec i zobaczymy co wyjdzie:

Definicja 26. Powiemy, że ciąg funkcji $f_n : D \to \mathbb{K}$, $D \subset \mathbb{K}^l$, jest zbieżny *jednostajnie* do funkcji $f : D \to \mathbb{K}$, jeśli:

• dla każdego $\varepsilon > 0$ istnieje takie N, że dla $n > N$ i dla **każdego** $x \in D$ mamy

$$|f_n(x) - f(x)| < \varepsilon.$$

Różnica jest taka, że w ostatnim określeniu dobieramy N, tak aby było tak samo dobre dla wszystkich $x \in D$, w poprzedniej N może zależeć od x! Zauważmy, że ciąg określony wzorem (13) jest zbieżny jednostajnie na $[0, a]$ dla dowolnego $a < 1$:

$$|f_n(x) - 0| = x^n \leq a^n < \varepsilon$$

dla $n > N$, gdzie N spełnia $a^N \leq \varepsilon$. Natomiast takie oszacowanie nie jest możliwe dla $x \leq 1$. Zauważmy, że teraz granica jest ciągła. W istocie rzeczy nie jest trudno wykazać:

Twierdzenie 54. Załóżmy, że funkcje $f_n : [a, b] \to \mathbb{K}$ są ciągłe i ciąg $\{f_n\}$ jest zbieżny jednostajnie do funkcji f. Wtedy funkcja graniczna f jest ciągła.

Dowód. Niech punkt $x \in [a, b]$ i $\varepsilon > 0$ będą dowolne. Badamy różnicę $|f(x) - f(y)|$, mamy

$$\begin{aligned} |f(x) - f(y)| &= |f(x) - f_n(x) + f_n(x) - f_n(y) + f_n(y) - f(y)| \\ &\leq |f(x) - f_n(x)| + |f_n(x) - f_n(y)| + |f_n(y) - f(y)|. \end{aligned}$$

Ze zbieżności jednostajnej ciągu f_n wynika istnienie takiego N, że mamy $|f(t) - f_n(t)| < \varepsilon/3$ dla $n > N$ i każdego $t \in [a, b]$. Ustalmy zatem $n = N+1$, wtedy powyższa nierówność przyjmie postać

$$|f(x) - f(y)| < \varepsilon/3 + |f_{N+1}(x) - f_{N+1}(y)| + \varepsilon/3.$$

Teraz korzystamy z ciągłości funkcji f_{N+1} w punkcie x i dla zadanego $\varepsilon/3$ dobieramy $\delta > 0$ takie, że $|f_{N+1}(x) - f_{N+1}(y)| < \varepsilon/3$ dla $y \in [a, b]$ spełniających $|x - y| < \delta$. Ostatecznie:

$$|f(x) - f(y)| < \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon,$$

dla $|x - y| < \delta$. □

Zapytajmy teraz o różniczkowalność granicy jednostajnie zbieżnego ciągu funkcyjnego. Zaczniemy od rozważenia konkretnej sytuacji.

Przykład 12. Niech ciąg f_n będzie dany wzorem:

$$f_n(x) = \frac{1}{\sqrt{n}} \sin nx, \qquad x \in \mathbb{R}.$$

Wtedy f_n zbiega jednostajnie do zera, bo jeśli $\varepsilon > 0$ jest dowolne, to

$$|f_n(x)| \leq \frac{1}{\sqrt{n}} < \varepsilon$$

dla $n > N = [1/\varepsilon^2] + 1$ dla wszystkich $x \in \mathbb{R}$. Z drugiej strony dla dowolnego $x \in \mathbb{R}$

$$f_n'(x) = \sqrt{n} \cos nx \not\to 0.$$

Widać, że potrzebne są dodatkowe założenia, aby wykazać różniczkowalność granicy. Istotnie, mamy bowiem następujący wynik.

Twierdzenie 55. Załóżmy, że $f_n,\ f,\ g : [a, b] \to \mathbb{R}$ i funkcje f_n są różniczkowalne w (a, b). Zakładamy, że

(a) istnieje $x_0 \in [a, b]$ takie, że ciąg liczbowy $f_n(x_0)$ jest zbieżny do $f(x_0)$;

(b) ciąg funkcyjny f_n' jest zbieżny jednostajnie do funkcji g.
Wtedy,

$$\lim_{n\to\infty} f_n(x) = f(x)$$

i zbieżność jest jednostajna. Co więcej

$$f'(x) = g(x).$$

(Bez dowodu). □

Zastanówmy się na koniec nad całkowalnością granicy ciągu funkcji całkowalnych w sensie Riemanna. Zaczniemy od negatywnego wyniku.

Przykład 13. Niech $\{r_n\}$ będzie ciągiem wszystkich liczb wymiernych z przedziału $[0,1]$. Kładziemy dla $x \in [0,1]$:

$$f_n(x) = \begin{cases} 1 & \text{dla } x = r_n \\ 0 & \text{w przeciwnym przypadku} \end{cases} \quad \text{i} \quad S_n(x) = \sum_{k=1}^{n} f_k(x).$$

Wtedy funkcje S_n są ciągłe poza skończoną ilością punktów i ograniczone, zatem $S_n \in \mathcal{R}(0,1)$. Ponadto, ciąg S_n jest zbieżny punktowo. Mianowicie łatwo się przekonać, że

$$\lim_{n\to\infty} S_n(x) = \chi_{\mathbb{Q}\cap[0,1]}(x).$$

Wiemy już, że $\chi_{\mathbb{Q}\cap[0,1]}$ nie jest całkowalne w sensie Riemanna!

Okazuje się, że sprawę ratuje zbieżność jednostajna. Mamy bowiem:

Twierdzenie 56. Załóżmy, że $f_n \in \mathcal{R}(a,b)$ i ciąg f_n jest zbieżny jednostajnie do f, wtedy $f \in \mathcal{R}(a,b)$ i

$$\lim_{n\to\infty} \int_a^b f_n(x)\,dx = \int_a^b f(x)\,dx.$$

Podkreślamy, że całkowalność granicy jest częścią tezy.

Dowód. Przeprowadzimy go dla najprostszego przypadku, tj. dla f_n będących funkcjami ciągłymi. Wtedy na mocy Twierdzenia 54 granica f jest ciągła, a zatem jest całkowalna. Wykażemy teraz powyższy wzór. Niech $\varepsilon > 0$ będzie dowolne, zaś N takie, że dla $n > N$ i dla wszystkich $x \in [a,b]$ mamy

$$|f_n(x) - f(x)| < \varepsilon/(b-a).$$

Następnie, korzystając z twierdzenia 48 dostaniemy,

$$\begin{aligned} \left| \int_a^b f_n(x)\,dx - \int_a^b f(x)\,dx \right| &\le \int_a^b |f_n(x) - f(x)|\,dx \\ &\le \int_a^b \varepsilon/(b-a)\,dx = \varepsilon. \end{aligned}$$

□

Zajmijmy się teraz szczególnym przypadkiem ciągów funkcyjnych, tj. szeregami funkcyjnymi.

Definicja 27. Załóżmy, że $f_n : D \subset \mathbb{K} \to \mathbb{K}$, $n = 1, \ldots$ *Szeregiem funkcyjnym* nazywamy napis

$$\sum_{n=0}^{\infty} f_n(x), \tag{14}$$

zaś $S_n(x) = \sum_{k=0}^{n} f_k(x)$ jest jego *ciągiem sum częściowych*. Powiemy, że szereg (14) jest zbieżny:

(a) *punktowo*, jeśli ciąg $S_n(x)$ jest zbieżny punktowo;

(b) *jednostajnie*, jeśli ciąg $S_n(x)$ jest zbieżny jednostajnie.

Badanie zbieżności jednostajnej ciągów i szeregów wydaje się skomplikowanym zadaniem. Czasem jednak jest to zadanie dość proste.

Twierdzenie 57. Załóżmy, że dany jest nam szereg (14), zaś funkcje $f_n : D \to \mathbb{K}$, gdzie $D \subset \mathbb{K}$ i spełniają one warunek $|f_n(x)| \leq M_n$ dla $x \in D$. Dodatkowo, szereg

$$\sum_{n=0}^{\infty} M_n < \infty. \tag{15}$$

Wtedy, szereg funkcyjny (14) jest jednostajnie zbieżny.

Dowód. Zauważmy, że z twierdzenia 11 (a) wynika, że dla każdego $x \in D$ szereg

$$\sum_{n=0}^{\infty} f_n(x)$$

jest zbieżny. Sumę oznaczymy jako $f(x)$. Wykażemy jednostajność zbieżności:

$$\begin{aligned} \left| f(x) - \sum_{k=0}^{n} f_n(x) \right| &= \left| \sum_{k=n+1}^{\infty} f_n(x) \right| \leq \sum_{k=n+1}^{\infty} |f_n(x)| \\ &\leq \sum_{k=n+1}^{\infty} M_n = R_n. \end{aligned}$$

Z uwagi na zbieżność szeregu liczbowego (15) dla dowolnego $\varepsilon > 0$ istnieje takie N, że mamy $R_n < \varepsilon$ dla $n > N$ a to oznacza jednostajną zbieżność szeregu funkcyjnego. □

3.7.2 Szeregi potęgowe

Wspomnieliśmy na początku podrozdziału, że interesują nas ciągi funkcji argumentu zespolonego. Istotnie, dopiero teraz tym się zajmiemy wprowadzając bardziej szczegółowe pojęcie.

Definicja 28. *Szeregiem potęgowym* nazwiemy szereg funkcyjny postaci

$$\sum_{n=0}^{\infty} a_n z^n, \quad z \in \mathbb{K}. \tag{16}$$

Wypada skomentować zbieżność szeregu liczb zespolonych. Zauważmy, że z definicji zbiór $\mathbb{C}$, to $\mathbb{R} \times \mathbb{R}$. Ze stwierdzenia 3 wynika, że szereg (16) zbiega wtedy i tylko wtedy, gdy zbiegają szeregi liczb rzeczywistych $\sum_{n=0}^{\infty} \mathrm{Re}(a_n z^n)$ i $\sum_{n=0}^{\infty} \mathrm{Im}(a_n z^n)$.

Dla szeregów potęgowych wprowadzimy pojęcie promienia zbieżności.

Definicja 29. Załóżmy, że dla ciągu $\{a_n\}_{n=0}^{\infty}$ istnieje granica

$$\lim_{n\to\infty} (|a_n|)^{1/n} = \alpha.$$

Wtedy liczbę $R = 1/\alpha$ nazwiemy *promieniem zbieżności* szeregu (16).

Znaczenie definicji wyjaśnią poniższe twierdzenia.

Twierdzenie 58. Niech $R > 0$ będzie promieniem zbieżności szeregu potęgowego (16). Wtedy szereg (16) jest zbieżny dla każdego z, takiego że $|z| < R$.

Dowód Polega on na zastosowaniu kryterium Cauchy'ego do szeregu (16). Dostaniemy wtedy:

$$\lim_{n\to\infty} (|a_n||z|^n)^{1/n} = \lim_{n\to\infty} (|a_n|)^{1/n}|z| = \alpha|z|.$$

Granica $\alpha|z|$ jest mniejsza od jedności, jeśli $|z| < 1/\alpha = R$. Zauważmy, że wtedy $|\mathrm{Re}(a_n z^n)|^{1/n}$, $|\mathrm{Im}(a_n z^n)|^{1/n} \leq |z||a_n|^{1/n}$ i możemy zastosować twierdzenie 15 (albo jeśli trzeba uwagę poniżej twierdzenia 15) aby wywnioskować zbieżność (16), gdy $|z| < R$. □

Będzie nas interesować kwestia zbieżności jednostajnej. W tej sprawie wypowiadamy się poniżej.

Twierdzenie 59. Niech $R > 0$ będzie promieniem zbieżności szeregu potęgowego

$$\sum_{n=0}^{\infty} a_n x^n, \quad x \in \mathbb{R}. \tag{17}$$

Wtedy,

(a) szereg (17) jest zbieżny jednostajnie dla $|z| < R - \varepsilon$, gdzie $\varepsilon \in (0, R)$ jest dowolne;

(b) szereg pochodnych jest zbieżny jednostajnie i promień zbieżności jest równy R. Co więcej,

$$\frac{d}{dx}\left(\sum_{n=0}^{\infty} a_n x^n\right) = \sum_{n=1}^{\infty} n a_n x^{n-1} \tag{18}$$

Uwaga. Część (a) jest prawdziwa dla $x \in \mathbb{C}$ i dowód nie ulega zmianie.

Dowód. Części (a) polega na sprowadzeniu do sytuacji, w której można stosować twierdzenie 57. Z definicji promienia zbieżności wynika istnienie takiego N, że mamy

$$|a_n z^n| = ((|a_n|)^{1/n}|z|)^n \leq ((\alpha + \alpha^2\varepsilon)(R - \varepsilon))^n = (1 - \varepsilon^2\alpha^2)^n,$$

dla $n > N$. Zatem dla $n > N$ możemy przyjąć $M_n = (1 - \varepsilon^2\alpha^2)^n$ (N pierwszych wyrazów szeregu nie ma wpływu na zbieżność szeregu!) i zastosować twierdzenie 57.

(b) Policzmy najpierw promień zbieżności szeregu pochodnych:

$$\lim_{n\to\infty}(|a_n|n)^{1/n} = \lim_{n\to\infty}(|a_n|)^{1/n}\lim_{n\to\infty} n^{1/n} = \alpha\cdot 1,$$

gdzie skorzystaliśmy z przykładu 4(c). Zatem promienie zbieżności obu szeregów są równe. Twierdzenie 55. o różniczkowaniu granicy ciągu funkcyjnego daje wzór (18). □

Z powyższego twierdzenia płynie ważny praktyczny wniosek: szeregi potęgowe można bezstrosko różniczkować i całkować w kole zbieżności $|x| < R$.

Na koniec rozpatrzmy zasadnicze zastosowanie rozwijanej do tej pory teorii do funkcji zadawanych szeregami potęgowymi,

$$f(x) = \sum_{n=0}^{\infty} a_n x^n. \tag{19}$$

Załóżmy, że jego promień zbieżności jest równy $R > 0$. Policzmy k-tą pochodną f w $x = 0$. Skoro $(x^i)^{(k)} \equiv 0$, dla $i < k$, $(x^i)^{(k)}|_{x=0} = 0$ dla $i > k$ i $(x^k)^{(k)} = k!$, to dostaniemy, że

$$f^{(n)}(0) = a_n n! \quad \text{albo} \quad a_n = f^{(n)}(0)/n!. \tag{20}$$

Jednak w praktycznych obliczeniach okazuje się, że wygodniej jest rozwijać funkcje f w innym punkcie, np. $x = a$

Twierdzenie 60. (Taylora) Załóżmy, że funkcja f jest zadana szeregiem (19), którego promień zbieżności wynosi $R > 0$. Wtedy dla a, takiego że $|a| < R$ mamy

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \tag{21}$$

dla x spełniających $|x - a| + |a| < R$.

Szkic dowodu: Ze wzoru (19) i dwumianu Newtona (twierdzenie 1.10) dostaniemy, że

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty} a_n((x-a)+a)^n \\ &= \sum_{n=0}^{\infty} a_n \sum_{k=0}^{n} \binom{n}{k} a^{n-k}(x-a)^k \\ &= \sum_{k=0}^{\infty}\sum_{n=k}^{\infty} a_n \binom{n}{k} a^{n-k}(x-a)^k. \end{aligned}$$

Można pokazać, że otrzymany szereg jest zbieżny i wzór (20) daje żądane rozwinięcie, tj. (21). □

3.8 Funkcja wykładnicza i funkcje trygonometryczne

Niniejszy podrozdział jest zwieńczeniem naszej dotychczasowej pracy. Pokażemy w nim jak można za pomocą szeregów potęgowych wprowadzić funkcję wykładniczą logarytmiczną a następnie funkcje $\sin$ i $\cos$.

3.8.1 Funkcje wykładnicza i logarytmiczna

Zaczniemy od definicji naszego obiektu zainteresowania.

Definicja 30. Dla $z \in \mathbb{C}$ kładziemy

$$E(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!} \tag{22}$$

Stwierdzenie 61. Promień zbieżności szeregu (22) jest równy ∞, tj. (22) jest zbieżny dla dowolnego $z \in \mathbb{C}$.

Dowód. Wystarczy wykazać, że $\lim_{n\to\infty} 1/(n!)^{1/n} = 0$. Niech $\varepsilon > 0$ będzie dowolne. Zauważmy, że dla pewnego $M \in \mathbb{N}$ i $n > M$, na mocy przykładu 4(b) mamy

$$(n!)^{1/n} \geq ((M-1)!)^{1/n} M^{(n-M)/n} \geq \sqrt{M},$$

dla $n > 2M$. Tym samym dla $M > 1/\varepsilon^2$ mamy

$$\frac{1}{(n!)} \leq \frac{1}{\sqrt{M}} < \varepsilon \quad \text{dla } n > 2[1/\varepsilon^2],$$

bo wtedy $(n-M)/M > 1/2$. Co należało wykazać. □

Zajmiemy się teraz wykazaniem podstawowej właściwości funkcji E. Zauważmy, że:

$$\begin{aligned} E(z)E(w) &= \lim_{n\to\infty} \sum_{k=0}^{n} \frac{z^k}{k!} \lim_{n\to\infty} \sum_{l=0}^{n} \frac{w^l}{l!} \\ &= \lim_{n\to\infty} \sum_{k=0}^{n} \sum_{l=0}^{n} \frac{z^k}{k!} \frac{w^l}{l!} = \lim_{n\to\infty} \sum_{k=0}^{2n} \sum_{i=0}^{k} \frac{z^{k-i} w^i}{(k-i)!\, i!} \\ &= \lim_{n\to\infty} \sum_{k=0}^{2n} \frac{1}{k!} \sum_{i=0}^{k} \binom{k}{i} = \lim_{n\to\infty} \sum_{k=0}^{2n} \frac{(z+w)^k}{k!} \\ &= E(z+w), \end{aligned}$$

gdzie wykorzystaliśmy dwumian Newtona (patrz twierdzenie 1.10). Zatem

$$E(z)E(w) = E(z+w). \tag{23}$$

Możemy teraz wyciągnąć serię wniosków.

Niech $z \in \mathbb{C}$, to wtedy

$$E(z)E(-z) = E(0) = 1. \tag{24}$$

O ostatniej równości przekonujemy się bezpośrednio z (22).

Niech teraz $x \in \mathbb{R}$ i $x > 0$. Z równania (22) natychmiast się przekonujemy, że

$$E(x) > 0. \tag{25}$$

Zatem wzór (24) daje

$$E(-x) > 0.$$

Jeśli $0 < x < y$, to ze wzoru (22) wynika, że

$$E(x) < E(y).$$

Wzór (24) dodatkowo daje

$$E(-y) < E(-x).$$

Możemy to ująć w formie stwierdzenia.

Stwierdzenie 62. Funkcja $\mathbb{R} \ni x \mapsto E(x) \in (0, \infty)$ jest ściśle rosnąca. □

Policzmy teraz granice tej funkcji. Skoro dla $x > 0$ mamy $E(x) > x$, to

$$\lim_{x\to\infty} E(x) = +\infty. \tag{26}$$

Zatem dzięki (24) dostaniemy jeszcze, że

$$\lim_{x\to-\infty} E(x) = 0. \tag{27}$$

Niech $h \in \mathbb{R}$, wtedy

$$\lim_{h\to 0} \frac{E(h) - 1}{h} = 1. \tag{28}$$

Jest tak, bo

$$\frac{1}{h}(E(h) - 1) = \frac{1}{h} \sum_{n=1}^{\infty} \frac{h^n}{n!} = 1 + h \sum_{n=2}^{\infty} \frac{h^{n-2}}{n!}$$

Zaś dla $|h| \leq 1$,

$$\left| \sum_{n=2}^{\infty} \frac{h^{n-2}}{n!} \right| \leq \sum_{n=2}^{\infty} \frac{|h|^{n-2}}{n!} \leq \sum_{n=2}^{\infty} \frac{1}{n!} < e$$

Z równości (23) i (28) wynika, że dla dowolnych $x \in \mathbb{R}$

$$\lim_{h\to 0} \frac{E(x + h) - E(x)}{h} = E(x). \tag{29}$$

Zajmijmy się algebraicznymi właściwościami funkcji E. Wzór (24) daje dla dowolnych $z \in \mathbb{C}$ i $n \in \mathbb{N}$

$$E(nz) = E(z)^n. \tag{30}$$

W szczególności, gdy $z = 1$ dostaniemy

$$E(n) = e^n.$$

Niech teraz p i $q > 0$ będą całkowite. Dzięki (30) dostaniemy, że

$$E(p/q)^q = E(pq/q) = E(p) = e^p,$$

a zatem po wyciągnięciu pierwiastków:

$$E(p/q) = e^{p/q}.$$

Do tej pory nie zdefiniowaliśmy dowolnej potęgi x liczby e (czy ogólniej: dowolnej liczby rzeczywistej a). Powyższa równość sugeruje, że możemy to łatwo zrobić, tak aby nowa definicja była zgodna ze starą, gdy x jest liczbą wymierną:

Definicja 31. Dla $x \in \mathbb{R}$ kładziemy

$$e^x = E(x),$$

funkcję $x \mapsto e^x$ nazywamy funkcją *wykładniczą*.

Zbierzmy właściwości funkcji wykładniczej.

Twierdzenie 63. Funkcja wykładnicza ma następujące właściwości:
(a) jest ciągła;
(b) jest różniczkowalna i $(e^x)' = e^x$;
(c) jest ściśle rosnąca i $e^x > 0$ dla $x \in \mathbb{R}$;
(d) spełnia $e^{x+y} = e^x e^y$;
(e) $\lim_{x\to\infty} e^x = \infty$, $\lim_{x\to-\infty} e^x = 0$;
(f) dla dowolnego $n \in \mathbb{N}$ mamy $\lim_{x\to\infty} x^n e^{-x} = 0$;

Dowód. (a) wynika z twierdzenia 59. Część (b) w całości wynika z równania (29). Stwierdzenie 62 i wzór (22) dają (c). Ze wzoru (24) wynika (d). Wzory (26) i (27) pociągają (e). Zajmiemy się punktem (f). Z definicji $E(x)$ wynika, że $x^{n+1}/(n+1)! < e^x$. Zatem $x^n e^{-x} < (n+1)!/x$, co pociąga tezę. □

Zauważmy teraz, że dodatniość funkcji wykładniczej i punkt (e) sprawiają, iż funkcja e^x jest **na** zbiór $(0, \infty)$. A zatem istnieje funkcja odwrotna $L : (0, \infty) \to \mathbb{R}$, która spełnia:

$$L(E(x)) = x, \quad x \in \mathbb{R} \quad \text{i} \quad E(L(y)) = y, \quad y > 0. \tag{31}$$

Zastosowanie twierdzenia o pochodnej funkcji odwrotnej daje nam teraz, że

$$\frac{dL}{dy}(y) = \frac{1}{e^x}|_{x=L(y)} = \frac{1}{e^{L(y)}} = \frac{1}{y}.$$

Wzory (31) prowadzą do wniosku, że $L(1) = 0$. Z twierdzenia 49 wynika, że

$$\int_1^y \frac{1}{y}\, dy = \int_1^y \frac{dL}{dy}(y)\, dy = L(y) - L(1) = L(y).$$

Zbierzmy dodatkowe właściwości funkcji $L(y)$:

Jeśli $u = E(x)$ i $v = E(y)$, to wtedy wzory (31) dają:

$$L(uv) = L(E(x)E(y)) = L(E(x+y)) = x + y = L(u) + L(v). \tag{32}$$

Zauważmy, że twierdzenie 63 (e) daje, że

$$\lim_{y\to\infty} L(y) = \infty \text{ i } \lim_{y\to 0^+} L(y) = 0. \tag{33}$$

Nie musimy już dalej kryć czym naprawdę jest funkcja L:

Definicja 32. Dla $y > 0$ piszemy $\ln y := L(y)$ i nazywamy *logarytmem naturalnym*.

Możemy już zdefiniować *potęgę dwu liczb rzeczywistych dodatnich*, tak by nowa definicja pokrywała się ze starą.

Definicja 33. Jeśli $a, b > 0$, to kładziemy:

$$a^b := e^{b \ln a}.$$

łatwo jest sprawdzić (jest to ćwiczenie dla Czytelnika), że gdy $b = p/q$, to prawa strona przyjmuje postać $a^{p/q}$.

3.8.2 Funkcje trygonometryczne

Wskażemy na bardzo bliski związek funkcji wykładniczej i funkcji trygonometrycznych.

Dla $x \in \mathbb{R}$ kładziemy:

$$C(x) = \frac{1}{2}(E(ix) + E(-ix)), \qquad S(x) = \frac{1}{2i}(E(ix) - E(-ix)). \tag{34}$$

Ponieważ dla $z \in \mathbb{C}$ równanie (22) daje, że $E(\bar{z}) = \overline{E(z)}$, to dostaniemy stąd, że

$$\overline{E(ix)} = E(\overline{(ix)}) = E(-ix).$$

Tym samym $C(x), S(x) \in \mathbb{R}$ i

$$E(ix) = C(x) + iS(x).$$

Równość (24) prowadzi do nowych, ciekawych wniosków:

$$|E(ix)|^2 = E(ix)E(-ix) = E(0) = 1.$$

Tym samym dostaniemy *jedynkę trygonometryczną*:

$$C^2 + S^2 = 1. \tag{35}$$

Zauważmy jeszcze, że

$$C(0) = 1, \qquad S(0) = 0.$$

Definicje (22) i (34) prowadzą po prostych działaniach na liczbach zespolonych do przedstawienia C i S w postaci następujących szeregów:

$$C(x) = 1 - \tfrac{x^2}{2!} + \tfrac{x^4}{4!} - \tfrac{x^6}{4!} + \tfrac{x^8}{8!} - \dots$$
$$S(x) = x - \tfrac{x^3}{3!} + \tfrac{x^5}{5!} - \tfrac{x^7}{7!} + \tfrac{x^9}{9} - \dots$$

Dostaniemy stąd też, że

$$C'(x) = -S(x), \qquad S'(x) = C(x).$$

Udowodnimy teraz ważny fakt, a mianowicie istnienie liczby π i jej właściwości.

Stwierdzenie 64. Istnieje $x > 0$ takie, że $C(x) = 0$.

Dowód. A.a. Gdyby taka liczba x nie istniała, to z $C(0) = 1$ wynikałoby, że $C(x) > 0$ dla wszystkich $x > 0$. Co więcej, skoro $S'(x) = C(x)$, to funkcja $S(x)$ byłaby ściśle rosnąca dla $x > 0$. Niech teraz $0 < x < y$, dzięki monotoniczności $S(x)$ mamy

$$S(x)(y - x) < \int_x^y S(t)\,dt = C(x) - C(y) \leq 2, \tag{36}$$

gdzie ostatnia nierówność wynika z (35). Pamiętamy, że $S(x) > 0$ dla $x > 0$, zatem dla dużych y nierówność (36) nie może być prawdziwa. Dowodzi to naszej tezy. □

Kładziemy

$$x_0 = \inf\{0 < x : \ C(x) = 0\}.$$

Z ciągłości funkcji C wynika, że $C(x_0) = 0$ a zatem $x_0 > 0$. Teraz wprowadzimy ważną definicję w „łatwy sposób".

Definicja 34.

$$\pi = 2x_0.$$

Z samej definicji wynika, że

$$C(\pi/2) = 0, \qquad S(\pi/2) = \pm 1.$$

Ponieważ, $S'(x) = C(x) > 0$ dla $x \in (0, \pi/2)$, to $S(\pi/2) = 1$. Tym samym

$$E(i\pi/2) = i,$$

dalej

$$E(i\pi) = (E(i\pi/2))^2 = -1, \qquad E(2i\pi) = (E(i\pi))^2 = 1.$$

Ostatecznie, dla dowolnego $z \in \mathbb{C}$ mamy $E(z + 2\pi i) = E(z)$.

Podsumujmy właściwości funkcji C i S:

Twierdzenie 65. (a) Funkcja E jest okresowa o okresie $2\pi i$, tj. dla dowolnych $z \in \mathbb{C}$ mamy $E(z + 2\pi i) = E(z)$.

(b) Funkcje C i S są ciągłe, różniczkowalne i okresowe o okresie 2π, tj. dla dowolnych $x \in \mathbb{R}$ mamy $C(x + 2\pi) = C(x)$, $S(x + 2\pi) = S(x)$.

(c) Jeśli $0 < t < 2\pi$, to $E(it) \neq 1$.

(d) Jeśli $z \in \mathbb{C}$ i $|z| = 1$, to istnieje dokładnie jedno $t \in [0, 2\pi)$, spełniające $z = E(ti)$.

Dowód. Punkt (a) został już wykazany. Ciągłość i różniczkowalność funkcji C i S wynika z faktu, iż są one przedstawialne zbieżnymi szeregami potęgowymi. Sprawdzamy okresowość:

$$\begin{aligned} C(x+2\pi) &= \frac{1}{2}(E(ix+2\pi i)+E(-ix-2\pi i)) = \frac{1}{2}(E(ix)+E(-ix)/E(2\pi i)) \\ &= \frac{1}{2}(E(ix)+E(-ix)) = C(x). \end{aligned}$$

Pomijamy analogiczne rachunki dla $S(x)$.

(c) Niech na początek $0 < t < \pi/2$, zatem $E(it) = x + iy$, gdzie $x, y \in (0,1)$. Policzymy $E(4ti)$. Mamy

$$E(4ti) = (x+iy)^4 = x^4 - 6x^2y^2 + y^4 = 4ixy(x^2-y^2).$$

Tym samym $E(it) \in \mathbb{R}$ wtedy i tylko wtedy, gdy $x^2 - y^2 = 0$, ale $x^2 + y^2 = 1$, tj. $x^2 = y^2 = 1/2$ i $E(it) = -1$. Wynika stąd prawdziwość (c).

Część (d) zostawiamy bez dowodu. □

Rozdział zakończymy definicją:

Definicja 35.

$$\sin x := S(x), \qquad \cos x := C(x).$$

Właściwości tych funkcji trygonometrycznych zostały podsumowane w twierdzeniu 64, aczkolwiek nie wszystkie.

Twierdzenie 66. Dla dowolnych liczb $x, y \in \mathbb{R}$ mamy

$$\sin(x+y) = \sin x \cos y + \cos x \sin y \quad \cos(x+y) = \cos x \cos - \sin x \sin y. \tag{37}$$

Dowód. Wynika on ze wzoru 34 i tożsamości (23). Przeprowadzenie rachunków powierzamy Czytelnikowi.

Rozdział 4

Rachunek różniczkowy funkcji wielu zmiennych

Jest to dość krótki rozdział, w którym przedstawiamy zarys rachunku różniczkowego funkcji wielu zmiennych. Wymaga to rozszerzenia naszej wiedzy o zbiorach w $\mathbb{R}^n$ i wprowadzenia nowych struktur w przestrzeniach liniowych. Jest to konieczne do wprowadzenia pojęcia pochodnej funkcji wielu zmiennych. Rachunek różniczkowy znajdzie główne zastosowanie w badaniu lokalnych ekstremów funkcji i obliczeniach przybliżonych za pomocą twierdzenia Taylora. Podobnie jak w poprzednim rozdziale symbol $\mathbb{K}$ oznacza ciało liczb rzeczywistych $\mathbb{R}$ albo ciało liczb zespolonych $\mathbb{C}$.

4.1 Przestrzenie unormowane i metryczne

Mówienie o funkcjach wielu zmiennych, ich ciągłości i różniczkowalności wymaga lepszej znajomości struktury zbiorów w $\mathbb{R}^n$. Mianowicie, nie w każdym punkcie można sensownie badać istnienie granicy, czy nie w każdym zbiorze określoności funkcji można ją różniczkować. Musimy wprowadzić szereg dodatkowych nowych pojęć. Szczęśliwie nasza dotychczasowa wiedza pozwala na umotywowanie ich pokaźną liczbą przykładów.

4.1.1 Definicje i przykłady

Oto pierwsze z nowych pojęć.

Definicja 1. Niech V będzie p.w. nad $\mathbb{K}$, funkcję $\|\cdot\| : V \to \mathbb{R}$ nazywamy *normą*, jeśli spełnia poniższe warunki:

(N1) dla dowolnego $v \in V$, $\|v\| \geq 0$, nadto $\|v\| = 0$ pociąga $v = 0$;

(N2) dla dowolnych $\lambda \in \mathbb{K}$ i $v \in V$, $\|\lambda v\| = |\lambda|\|v\|$;

(N3) (nierówność trójkąta) dla dowolnych $v, w \in V$, mamy $\|v + w\| \leq \|v\| + \|w\|$.

Parę $(V, \|\cdot\|)$ nazywamy *przestrzenią unormowaną*.

Przykłady 1. Niech $V = \mathbb{R}^n$, pokażemy, że normę można wprowadzić na kilka sposobów. Kładziemy:

(1) $\|v\|_2 = (v,v)^{1/2} \equiv \sqrt{\sum_{i=1}^n v_i^2}$. Warunki (N1) i (N2) definicji są spełnione w sposób oczywisty. Sprawdzimy nierówność trójkąta:

$$\|v+w\|_2^2 = \|v\|_2^2 + \|w\|_2^2 + 2(v,w)$$

Nierówność Schwarza (twierdzenie 1.26) daje

$$\|v+w\|_2^2 \leq \|v\|_2^2 + \|w\|_2^2 + 2\|v\|_2\|w\|_2 = (\|v\|_2 + \|w\|_2)^2.$$

(2) $\|v\|_\infty = \max_i |v_i|$. Prawdziwość warunków (N1) i (N2) jest oczywista. Warunek trójkąta za chwilę wykażemy w ogólniejszej sytuacji.

(3) $\|v\|_1 = \sum_{i=1}^n |v_i|$. Prawdziwość warunków (N1) i (N2) jest znów oczywista. Warunek (N3) jest łatwy do sprawdzenia, wynika on z nierówności trójkąta dla wartości bezwzględnej i pozostawiamy to zadanie czytelnikowi.

(4) Niech $V = C([0,T])$, gdzie symbol $C([0,T])$ oznacza p.w. funkcji ciągłych o wartościach rzeczywistych określonych na przedziale $[0,T]$. Kładziemy:

$$\|f\|_\infty = \max_{x\in[0,T]} |f(x)|.$$

Zauważmy, że dzięki właściwościom funkcji ciągłych określonych na przedziałach domkniętych i ograniczonych prawa strona jest dobrze określona. Ponownie sprawdzenie warunku (N1) i (N2) nie przedstawia kłopotu. Zajmiemy się nierównością trójkąta. Niech $x \in [0,T]$ będzie dowolny, wtedy mamy

$$|f(x)+g(x)| \leq |f(x)| + |g(x)| \leq \max_{y\in[0,T]} |f(y)| + \max_{y\in[0,T]} |g(y)| = \|f\|_\infty + \|g\|_\infty$$

Ponieważ x po lewej stronie jest dowolny, to można zań przyjąć ten punkt, w którym jest osiągane $\max_{y\in[0,T]} |f(y)+g(y)|$. Zatem,

$$\|f+g\|_\infty = \max_{y\in[0,T]} |f(y)+g(y)| \leq \|f\|_\infty + \|g\|_\infty.$$

Zauważmy, że ten sam argument dowodzi prawdziwości (N3) w części (2).

Nie każdy interesujący zbiór ma strukturę p.w., dlatego potrzebne jest nowe pojęcie.

Definicja 2. Niech X będzie dowolnym zbiorem, funkcję $d : X \times X \to \mathbb{R}$ nazwiemy *metryką (odległością),* jeśli spełnia:

(D1) $d(x,y) > 0$, jeśli $x \neq y$ i $d(x,x) = 0$;

(D2) $d(x,y) = d(y,x)$;

(D3) (nierówność trójkąta) dla dowolnego $x,y,z \in X$, $d(x,y) \leq d(x,z) + d(z,y)$.

Parę (X,d) nazywamy *przestrzenią metryczną.*

Przykłady 2.

(1) Niech $X = V$ i V jest p.w. nad $\mathbb{R}$, zaś $\|\cdot\|$ jest normą w V. Kładziemy

$$d(x,y) = \|x-y\|.$$

Łatwo sprawdzić, że warunki (D1)-(D2) są spełnione, sprawdzamy (D3):

$$d(x,y) = \|x-y\| = \|(x-z)+(z-y)\| \leq \|x-z\| + \|z-y\| = d(x,z) + d(z,y).$$

(2) Niech $X = \mathbb{R}^n$, kładziemy

$$d(x,y) = \begin{cases} 0 & \text{gdy } x = y \\ 1 & \text{gdy } x \neq y. \end{cases}$$

Metrykę tę nazywa się *dyskretną*. Warunki (D1)-(D2) są łatwe do sprawdzenia, zajmiemy się (D3):

(a) jeśli $x = y$, to lewa strona nierówności trójkąta jest zerem a prawa jest nieujemna. Nierówność jest prawdziwa.

(b) jeśli $x \neq y$, to lewa strona nierówności trójkąta jest równa 1, zaś jakie by nie było z, to $z \neq x$ lub $z \neq y$, więc prawa strona jest co najmniej 1, więc nierówność jest spełniona.

4.1.2 Zbiory w przestrzeniach metrycznych

Wprowadzimy teraz szereg nowych pojęć geometrycznych:

Definicja 3. Niech X będzie przestrzenią metryczną.
(a) *kulą otwarta* o środku x i promieniu $r > 0$ nazywamy zbiór

$$B(x,r) = \{y \in Y : d(x,y) < r\}.$$

(a) *kulą domkniętą* o środku x i promieniu $r > 0$ nazywamy zbiór

$$\bar{B}(x,r) = \{y \in Y : d(x,y) \leq r\}.$$

Przykłady 3. Niech $X = \mathbb{R}^2$:
(1) $d_2(x,y) = \|x-y\|_2$, to $B(0,1)$ jest zwykłym kołem o środku w punkcie 0 i promieniu 1 (bez okręgu!).
(2) $d_\infty(x,y) = \|x-y\|_\infty$, to $B(0,1)$ jest kwadratem o wierzchołkach: (1,1), (-1,1), (1,-1) i (-1,-1).
(3) $d_1(x,y) = \|x-y\|_1$, to $B(0,1)$ jest kwadratem o wierzchołkach: (1,0), (0,1), (-1,0) i (0,-1).
(4) Jeśli d jest metryką dyskretną, to $B(0,1) = \{0\}$ i $\bar{B}(0,1) = \mathbb{R}^2$.

Definicja kuli otwartej jest podstawową dla dalszego rozwoju teorii zbiorów w przestrzeniach metrycznych. Wyjaśnią to poniższe definicje:

Definicja 4. Niech X będzie przestrzenią metryczną.
(1) Powiemy, że zbiór $U \subset X$ jest *otwarty*, jeśli dla każdego $x \in X$ istnieje takie $r > 0$, że $B(x,r) \subset U$.
(2) Powiemy, że zbiór $F \subset X$ jest *domknięty,* jeśli zbiór $X \setminus F$ jest otwarty.

Odnotujmy od razu podstawowe wnioski z definicji. Mianowicie, kule otwarte są zbiorami otwartymi a kule domknięte są zbiorami domkniętymi. Nadto, zbiory $\emptyset$ i X - cała przestrzeń metryczna są zbiorami jednocześnie otwartymi i domkniętymi.

Proste właściwości zbiorów otwartych i domkniętych sformułujemy poniżej.

Stwierdzenie 1.
(1) Niech $\{U_i\}_{i\in I}$ będzie rodziną zbiorów otwartych, wtedy
(a) zbiór $\bigcup_{i\in I} U_i$ jest otwarty;
(b) jeśli dodatkowo zbiór I jest skończony, to zbiór $\bigcap_{i\in I} U_i$ jest otwarty.
(2) Niech $\{F_i\}_{i\in I}$ będzie rodziną zbiorów domkniętych, wtedy
(a) zbiór $\bigcap_{i\in I} F_i$ jest domknięty;
(b) jeśli dodatkowo zbiór I jest skończony, to zbiór $\bigcup_{i\in I} F_i$ jest domknięty.

Dowód. (1a) Jeśli $x \in \bigcup_{i\in I} U_i$, to jest to równoważne temu, że istnieje taki wskaźnik $j \in I$, że, $x \in U_j$. Skoro U_j jest otwarty, to istnieje kula otwarta $B(x, r)$ zawarta w nim, zatem jest ona zawarta w sumie zbiorów.

(1b) Jeśli $x \in \bigcap_{i=1}^{N} U_i$, to jest to równoważne temu, że $x \in U_j$ dla wszystkich wskaźników $j = 1, \ldots, N$. Skoro U_j są otwarte, to istnieją kule otwarte $B(x, r_j)$ zawarte w U_j, $j = 1, \ldots, N$. Zatem, kula $B(x, r)$, gdzie $r = \min\{r_1, \ldots, r_N\} > 0$ jest zawarta w przecięciu zbiorów.

Część (2) wynika z (1) zastosowanej do uzupełnień zbiorów F_i. □

Podamy metodę budowania nowych zbiorów otwartych.

Definicja 5. Niech (X, d) będzie przestrzenią metryczną i $G \subset X$. Zbiór

$$\overset{o}{G} = \{x \in G : \text{ istnieje takie } \varepsilon > 0, \text{ że } B(x, \varepsilon) \subset G\}$$

nazywamy *wnętrzem.*

Zgodnie z zapowiedzią mamy.

Stwierdzenie 2. $\overset{o}{G}$ jest zbiorem otwartym.

Dowód. Niech $x \in G$, zatem $B(x, \varepsilon) \subset G$ dla pewnego ε. Jeśli teraz $y \in B(x, \varepsilon)$, to

$$B(y, \varepsilon - d(x, y)) \subset B(x, \varepsilon) \subset G.$$

Oznacza to, że $y \in \overset{o}{G}$, dla wszystkich $y \in B(x, \varepsilon)$. Tym samym $\overset{o}{G}$ jest zbiorem otwartym. □

Sformułujemy teraz podstawową właściwość zbiorów domkniętych, którą poprzedzimy nowym określeniem, które odegra ważną rolę.

Definicja 6. Niech (X, d) będzie przestrzenią metryczną i $D \subset X$. Punkt $x_0 \in X$ nazwiemy *punktem skupienia* zbioru D, jeśli dla dowolnego $r > 0$

$$(B(x_0, r) \setminus \{x_0\}) \cap D \neq \emptyset,$$

tj. dowolna kula o środku w x_0 zawiera punkty z D różne od x_0.

Twierdzenie 3. Zbiór F jest domknięty wtedy i tylko wtedy, gdy zawiera wszystkie swoje punkty skupienia.

Dowód. $\Rightarrow$ Niech F będzie domknięty, wtedy $X \setminus F =: U$ jest otwarty. Zatem dla dowolnego $x \in U$ istnieje kula otwarta $B(x,r)$ zawarta w U. Zatem x nie jest punktem skupienia F, bo $B(x,r) \cap F = \emptyset$.
$\Leftarrow$ Niech teraz F zawiera wszystkie swoje punkty skupienia i $x \notin F$. Ponieważ x nie jest punktem skupienia F, to istnieje $B(x,r) \cap F = \emptyset$. Oznacza to, że $B(x,r) \subset X \setminus F$. Skoro x był dowolny, to oznacza to, że $X \setminus F$ jest otwarty.

Przykład 4. Kładziemy $E = \{1/n : \ n \in \mathbb{N}, \ n \geq 1\}$. Wtedy 0 jest jedynym punktem skupienia zbioru E i 0 nie należy do E. Tym samym E nie jest domknięty.

Odnotujmy jeszcze prosty fakt, poprzedzony definicją.

Definicja 7. *Otoczeniem* punktu $x_0 \in X$, gdzie (X,d) jest przestrzenią metryczną, nazywamy dowolny zbiór otwarty zawierający x_0.

Twierdzenie 4. Jeśli p jest punktem skupienia zbioru E, to w każdym otoczeniu punktu p istnieje nieskończenie wiele punktów należących do E.

Dowód. a.a. Załóżmy, że w otoczeniu $B(x,r)$ punktu x jest tylko skończenie wiele punktów z E różnych od x i są to $p_1, \ldots, p_n$. Wtedy kładziemy $\delta = \min\{d(x,p_1), \ldots, d(x,p_n)\}$ i widzimy, że w otoczeniu $B(x,\delta)$ nie ma punktów z E, wbrew założeniu, że x jest punktem skupienia E. □

Wprowadzimy jeszcze jedno określenie.

Definicja 8. Niech $D \subset X$, gdzie (X,d) jest przestrzenią metryczną. Punkt x_0 nazwiemy *punktem brzegowym* zbioru D, jeśli w każdym otoczeniu x_0 znajdują się punkty należące i nie należące do D. Zbiór punktów brzegowych nazwiemy *brzegiem* zbioru D i oznaczymy symbolem ∂D.

Odnotujmy podstawową właściwość brzegu.

Stwierdzenie 5. Zbiór ∂D jest domknięty.

Dowód. Zauważmy, że jeśli $x \notin \partial D$, to istnieje takie otoczenie U punktu x, że $U \cap D = \emptyset$. Tym samym $X \setminus \partial D$ jest zbiorem otwartym. □

4.2 Granica i ciągłość funkcji

4.2.1 Granica ciągu

W rozdziale 3. pojawiło się pojęcie zbieżności ciągu punktów w $\mathbb{R}^n$ wyrażone w terminach odległości. Korzystając z okazji, że wprowadziliśmy pojęcie przestrzeni metrycznej, której przykładem jest $\mathbb{R}^n$, możemy przedstawić zbieżność dość ogólnie:

Definicja 9. Niech (X,d) będzie przestrzenią metryczną i $\{a_n\}_{n=0}^{\infty}$ będzie ciągiem punktów X. Powiemy, że ciąg $\{a_n\}_{n=0}^{\infty}$ jest *zbieżny* i ma *granicę* równą g, jeśli dla dowolnego $\varepsilon > 0$ istnieje

taka liczba naturalna N, że

$$d(a_n, g) < \varepsilon \quad \text{dla } n > N.$$

Ta ogólna definicja wkrótce nam się przyda, lecz chcemy podkreślić dwa fakty:

(•) pracujemy przede wszystkim z ciągami punktów z $\mathbb{R}^k$, $k \geq 1$;

(•) bywa, że pojęcie zbieżności zależy od wyboru metryki, np. jeśli d jest metryką dyskretną w $\mathbb{R}^2$, to jedynymi ciągami zbieżnumi są ciągi, które są stałe od pewnego numeru k_0.

Chcemy w tym miejscu przypomnieć, że jeśli $a_n = (a_n^1, \ldots, a_n^k)$, to zbieżność a_n do $g = (g^1, \ldots, g^k)$ jest równoważna temu, iż $\lim_{n\to\infty} a_n^i = g^i$, dla wszystkich $i = 1, \ldots, k$, (patrz stwierdzenie 3.3). Dzięki temu czytelnik sam może sprawdzić następujący fakt.

Twierdzenie 6. Niech $\{a_n\}_{n=1}^{\infty}$, $\{b_n\}_{n=1}^{\infty}$ będą zbieżnymi ciągami punktów w $\mathbb{R}^k$, $\lim_{n\to\infty} a_n = a$, $\lim_{n\to\infty} b_n = b$; zaś $\{c_n\}_{n=1}^{\infty}$ jest zbieżnym ciągiem liczbowym, $\lim_{n\to\infty} c_n = c$. Wtedy,

(a) $$\lim_{n\to\infty} (a_n + b_n) = a + b;$$

(b) $$\lim_{n\to\infty} (a_n, b_n) = (a, b),$$

gdzie symbol $(\cdot, \cdot)$ oznacza tu iloczyn skalarny w $\mathbb{R}^k$;

(c) $$\lim_{n\to\infty} a_n c_n = ac;$$

(d) jeśli dodatkowo c, $c_n \neq 0$, to

$$\lim_{n\to\infty} a_n / c_n = a/c.$$

4.2.2 Podciągi i Twierdzenie Bolzano–Weierstrassa

W przypadku ciągów punktów z $\mathbb{R}^n$ można wykazać wielowymiarowy odpowiednik twierdzenia Bolzano–Weierstrassa.

Twierdzenie 7. Niech $\{a_k\}_{k=1}^{\infty}$ będzie ograniczonym ciągiem punktów z $\mathbb{R}^n$. Wtedy istnieje zbieżny podciąg ciągu $\{a_k\}_{k=1}^{\infty}$.

Dowód. Jeden z dowodów polega na powieleniu dowodu jednowymiarowego twierdzenia, zastępując przedziału kostkami $[-M, M]^n$. Inny polega na zastosowaniu twierdzenia z jednego wymiaru do każdej ze współrzędnej. Dla uproszczenia przyjmiemy $n = 2$. Niech $a_k = (a_k^1, a_k^2)$. Wtedy stosujemy znane twierdzenie do $\{a_k^1\}_{k=1}^{\infty}$. Dostajemy istnienie podciągu zbieżnego $\{a_{k_l}^1\}_{l=1}^{\infty}$. Rozpatrujemy następnie ciąg $\{a_{k_l}^2\}_{l=1}^{\infty}$. Ten sam argument daje istnienie zbieżnego podciągu $\{a_{k_{l_r}}^2\}_{r=1}^{\infty}$. Oczywiście podciąg $\{a_{k_{l_r}}^1\}_{r=1}^{\infty}$ też jest zbieżny. □

4.2.3 Granica funkcji w punkcie

W dalszym ciągu zajmiemy się funkcjami wielu zmiennych o wartościach wektorowych. Przykładem takiej funkcji jest przypisanie każdemu punktowi w rurze prędkości cieczy przepływającej przez rurę. Opisywanie właściwości takich funkcji poprzedzimy określeniem.

Definicja 10. Załóżmy, że (X, d) i (Y, ρ) są przestrzeniami metrycznymi. Niech $E \subset X$, $f : E \to Y$ i p będzie punktem skupienia zbioru E. Powiemy, że funkcja f ma granicę w punkcie p równą $g \in Y$, jeśli dla dowolnego $\varepsilon > 0$ istnieje takie $\delta > 0$, że

$$\rho(f(x), g) < \varepsilon$$

dla wszystkich $x \in E$, takich że $0 < d(x, p) < \delta$. Piszemy wtedy

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

Uwaga. Punkt p nie musi należeć do E.

Tak jak w przypadku funkcji jednej zmiennej o wartościach liczbowych prawdziwa jest ciągowa charakteryzacja granicy funkcji punkcie:

Stwierdzenie 8. Niech X, Y, E, f, p będą takie jak w definicji 10. Wówczas

$$\lim_{x \to p} f(x) = g$$

wtedy i tylko wtedy, gdy dla każdego ciągu $\{x_n\}_{n=0}^{\infty} \subset X$ zbieżnego do p i takiego, że $x_n \neq p$ mamy

$$\lim_{n \to \infty} f(x_n) = p.$$

Dowód przebiega tak samo, jak w przypadku twierdzenia 3.19. Wystarczy zamienić tylko odległość liczb x, y, tj. $|x - y|$ na odległość punktów $d(x, y)$ w przestrzeni metrycznej. Jest tak dlatego, że wykorzystywaliśmy jedynie właściwości (D1-D3) odległosci $|x-y|$ w $\mathbb{R}$. Szczegóły pozostawiamy Czytelnikowi. □

Tak samo, jak w przypadku liczbowym wykazujemy, że granica ciągu (funkcji w punkcie) jeśli istnieje, to jest wyznaczona jednoznacznie.

Sformułujemy jeszcze podstawowe właściwości granicy funkcji w punkcie. Przedtem wygodnie nam będzie, wzorem przykładu 2.1(7), wprowadzić strukturę przestrzeni wektorowej w zbiorze funkcji $f : X \to \mathbb{K}^n$, gdzie X jest dowolnym zbiorem. Jeśli $f, g : X \to \mathbb{K}^n$ i $\lambda \in \mathbb{K}$, $h : X \to \mathbb{K}$, to w naturalny sposób wprowadzamy dodawanie i mnożenie funkcji przez liczbę, a mianowicie:

$$(f + g)(x) := f(x) + g(x), \quad (\lambda \cdot f)(x) := \lambda f(x).$$

Nadto, jeśli $h : X \to \mathbb{K}$, to wprowadzimy mnożenie dwóch funkcji,

$$(h \cdot f)(x) := h(x) \cdot f(x).$$

Twierdzenie 9. Niech X będzie przestrzenią metryczną, $E \subset X$ i p jest punktem skupienia E, $f, g : E \to \mathbb{K}^m$ i

$$\lim_{x\to p} f(x) = A, \quad \lim_{x\to p} g(x) = B.$$

Wtedy

(a) $\lim_{x\to p}(f+g)(x) = A + B$;

(b) jeśli $m = 1$, to $\lim_{x\to p}(fg)(x) = AB$;

(c) jeśli $m = 1$, $B \neq 0$ i $g \neq 0$ w pewnym otoczeniu punktu p, to $\lim_{x\to p}(f/g)(x) = A/B$.

Dowód. Wynika z analogicznego twierdzenia dla ciągów.

4.2.4 Ciągłość funkcji

Zajmiemy się teraz jednym z najważniejszych pojęć analizy, tj. ciągłością funkcji w punkcie.

Definicja 11. Załóżmy, że (X, d) i (Y, ρ) są przestrzeniami metrycznymi. Niech $E \subset X$, $f : E \to Y$ i $p \in E$. Powiemy, że funkcja f jest *ciągła w punkcie* p, jeśli dla dowolnego $\varepsilon > 0$ istnieje takie $\delta > 0$, że

$$\rho(f(x), f(p)) < \varepsilon$$

dla wszystkich $x \in E$, takich że $d(x, p) < \delta$.

Zauważmy, że ciągłość w punkcie oznacza, że

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

Zauważmy też, że poznane właściwości liczbowych funkcji ciągłych przenoszą się na przypadek ogólny, a mianowicie mamy:

Twierdzenie 10. (o ciągłości funkcji złożonej) Załóżmy, że X, Y, Z są przestrzeniami metrycznymi, $E \subset X$. Niech funkcja $f : E \to Y$ będzie ciągła w punkcie $x = p$, zaś funkcja $g : f(E) \to Z$ będzie ciągła w punkcie $y = f(p)$, wtedy funkcja $h : E \to Z$ dana wzorem $h(x) = g(f(x))$ jest ciągła w punkcie $x = p$.

Dowód. Postępujemy tak samo jak w dowodzie twierdzenia 3.23, zamieniając odległość liczb na odległość punktów przestrzeni metrycznych. □

Definicja 12. Niech X, Y będą przestrzeniami metrycznymi, $E \subset X$, $f : E \to Y$. Powiemy, że funkcja f jest *ciągła na zbiorze* E, jeśli f jest ciągła w każdym punkcie zbioru E.

Możliwa jest następująca elegancka charakteryzacja funkcji ciągłych na E.

Twierdzenie 11. Niech (X, d) i (Y, ρ) będą przestrzeniami metrycznymi. Funkcja $f : X \to Y$ jest ciągła na X wtedy i tylko wtedy, gdy dla każdego zbioru otwartego $V \subset Y$ jego przeciwobraz $f^{-1}(V)$ jest zbiorem otwartym.

Dowód. $\Rightarrow$ Załóżmy, że V jest otwarty i dla pewnego $p \in X$ mamy $f(p) \in V$. Z otwartości V wynika istnienie takiego $\varepsilon > 0$, że $B(f(p), \varepsilon) \subset V$. Z ciągłości zaś mamy istnienie takiego $\delta > 0$, że jeśli $d(x, p) < \delta$, to $d(f(p), f(x)) < \varepsilon$, tj. $B(p, \delta) \subset f^{-1}(V)$. Tym samym $f^{-1}(V)$ jest zbiorem otwartym.

$\Leftarrow$ Jeśli dla otwartego V wiemy, że $f^{-1}(V)$ jest otwarty, to oznacza, że $f^{-1}(B(f(p), \varepsilon))$ też jest otwarty. Tym samym istnieje takie $\delta > 0$, że $B(p, \delta) \subset f^{-1}(B(f(p), \varepsilon))$ a oznacza to ciągłość w punkcie p.

Z tego twierdzenia i definicji zbiorów domkniętych wypływa prosty wniosek.

Stwierdzenie 12. Niech (X, d) i (Y, ρ) będą przestrzeniami metrycznymi. Jeśli funkcja $f : X \to Y$ jest ciągła na X, to dla dowolnego zbioru domkniętego $F \subset Y$, jego przeciwobraz $f^{-1}(F)$ jest zbiorem domkniętym.

Dowód. Skoro $Y \setminus F$ jest otwarty i $f^{-1}(Y \setminus F) = X \setminus f^{-1}(F)$, to nasza teza wynika natychmiast z twierdzenia 11. □

Przykłady 5. Jeśli $x \in \mathbb{R}^n$, to piszemy $x = (x_1, \ldots, x_n)$.

(a) Niech $f : \mathbb{R}^n \to \mathbb{R}$ będzie dana wzorem: $f(x) = x_i$, dla pewnego i, $1 \leq i \leq n$. Wtedy f jest ciągła. Wynika to z nierówności $|f(x) - f(y)| = |x_i - y_i| \leq |x - y| = d(x, y)$. Zatem możemy przyjąć $\delta := \varepsilon$ w definicji ciągłości.

(b) Na mocy definicji, ciągłość funkcji f w punkcie x_0 oznacza, że $\lim_{x \to x_0} f(x) = f(x_0)$. Nadto, w przypadku funkcji wektorowych konieczna i dostateczna jest ciągłość każdej ze składowych.

Niech teraz $g : \mathbb{R}^2 \to \mathbb{R}^2$ będzie dana wzorem: $g(x_1, x_2) = (e^{x_1} \cos x_2, e^{x_1} \sin x_2)^T$. Wtedy na mocy powyższej uwagi g jest ciągła.

(c) Niech funkcja $h : \mathbb{R}^2 \to \mathbb{R}$ będzie dana wzorem

$$h(x_1, x_2) = \begin{cases} \frac{x_1 x_2}{x_1^2 + x_2^2} & \text{gdy } (x_1, x_2) \neq (0, 0) \\ 0 & \text{gdy } (x_1, x_2) = (0, 0). \end{cases}$$

Wtedy funkcja h jest ciągła w każdym punkcie poza $(0, 0)$, co jest oczywiste. Nie jest zaś ciągła w punkcie $(0, 0)$. Aby to wykazać wystarczy odwołać się do ciągowej charakteryzacji granicy w punkcie i rozpatrzeć dwa ciągi

$$p_n = (1/n, 1/n), \qquad q_n = (1/n, 2/n).$$

Wtedy

$$\lim_{n \to \infty} h(p_n) = \frac{1}{2} \neq \frac{2}{5} = \lim_{n \to \infty} h(q_n).$$

(d) Niech $A \in M_{n \times m}(\mathbb{R})$, $A = \{a_{ij}\}$, wtedy funkcja $f : \mathbb{R}^n \to \mathbb{R}^m$ dana wzorem $f(x) = Ax$ jest ciągła. W tym celu wystarczy sprawdzić, że każda składowa $f(x)$ jest ciągła,

$$(Ax)_i = \sum_{j=1}^{n} a_{ij} x_j.$$

Teraz nasze stwierdzenie jest oczywiste.

(e) Niech $A \in M_{n\times n}(\mathbb{R})$, $A = \{a_{ij}\}$, wtedy funkcja $f : \mathbb{R}^n \to \mathbb{R}$ dana wzorem $f(x) = (Ax, x)$ jest ciągła. W tym celu wystarczy sprawdzić, że

$$f(x) = \sum_{i,j=1}^{n} a_{ij}x_i x_j,$$

bo teraz ciągłość f wynika z ciągłości iloczynów $x_i x_j$.

Sformułujemy teraz najważniejszą właściwość funkcji ciągłych. Jest ona związana z szukaniem największej i najmniejszej wartości funkcji.

Twierdzenie 13. Załóżmy, że zbiór $D \subset \mathbb{R}^n$ jest domknięty i ograniczony. Nadto funkcja $f : D \to \mathbb{R}$ jest ciągła. Wtedy istnieją takie punkty x_m i x_M należące do D, że

$$f(x_m) = \inf_{x\in D} f(x), \quad f(x_M) = \sup_{x\in D} f(x).$$

Wprawdzie pojęciowo dowód nie jest trudny, ale zabrałby sporo miejsca, dlatego zainteresowanego czytelnika odsyłamy do literatury, np. do książki Fichtenholza. □

Trzeba jeszcze podkreślić, że powyższe twierdzenie nie jest prawdziwe, gdy opuścimy którekolwiek z założeń:

- domkniętość D,
- ograniczoność D,
- ciągłość f.

Osobnym zagadnieniem jest znalezienie kandydatów na x_m i x_M. Do tego celu potrzebna jest nam nowa maszyneria, będąca odpowiednikiem rachunku różniczkowego funkcji jednej zmiennej.

4.3 Różniczkowanie funkcji wielu zmiennych

Zaczniemy od definicji podstawowych pojęć. Przyjmujemy, że przestrzeń $\mathbb{R}^k$ jest przestrzenią unormowaną z normą $\|\cdot\|_2$ (patrz przykład 1 (1)), jednak dla prostoty zapisu będziemy pisali $\|\cdot\|$ lub $|\cdot|$, tak jak dotąd.

Definicja 13. Niech E będzie otwartym podzbiorem $\mathbb{R}^n$, $x \in E$ i $f : E \to \mathbb{R}^m$. Jeśli istnieje odwzorowanie $A \in \mathrm{Hom}\,(\mathbb{R}^n, \mathbb{R}^m)$, takie że

$$\lim_{h\to 0} \frac{\|f(x+h) - f(x) - Ah\|}{\|h\|} = 0,$$

to mówimy, że funkcja f jest *różniczkowalna w punkcie* x. Nazywamy A *pochodną* f w punkcie x i piszemy $Df(x) = A$. Jeśli funkcja f jest różniczkowalna w każdym punkcie zbioru E, to mówimy, że f jest *różniczkowalna w* E.

Uwaga. Z definicji pochodnej wynika, że wymaga ona koniecznie, aby ów zbiór E był otwarty. Z ogólnych właściwości granicy wynika też, że pochodna jest wyznaczona jednoznacznie.

Zauważmy, że $x \mapsto Df(x) \in \mathrm{Hom}\,(\mathbb{R}^n, \mathbb{R}^m)$, tj. pochodna jest macierzą (jeśli utożsamimy najpierw przekształcenia liniowe i macierze, patrz §2.3). Tym samym, jeśli $n = 1 = m$, to $\mathrm{Hom}\,(\mathbb{R}, \mathbb{R}) = \mathbb{R}$ i nowa definicja pokrywa się ze starą, bo wtedy A jest liczbą.

Z definicji wynika, że

$$f(x+h) = f(x) + Df(x)h + r(h),$$

gdzie $\|r(h)\|/\|h\| \to 0$. Tym samym funkcja różniczkowalna w punkcie może być przybliżana funkcjami liniowymi (plus wartość stała). Prowadzi to do następującego wniosku.

Stwierdzenie 14. Niech E będzie otwartym podzbiorem $\mathbb{R}^n$, $x_0 \in E$ i funkcja $f : E \to \mathbb{R}^m$ jest różniczkowalna w x_0. Wtedy, f jest ciągła w x_0.

Dowód. Rozważmy różnicę $f(x_0 + h) - f(x_0)$, gdy $x_0 + h \in E$. Dzięki różniczkowalności dostaniemy

$$\|f(x_0+h) - f(x_0)\| \le \|Df(x_0)h\| + \|r(x,h)\|,$$

gdzie $\|r(x,h)\|/h$ dąży do zera, gdy $h \to 0$. Na mocy przykładu 5(d) wyraz $\|Df(x_0)h\|$ zmierza do zera, a to oznacza ciągłość f w x_0. □

Chcieliśmy naUczyć się praktycznie liczyć pochodne. W tym celu wprowadźmy prostsze, poręczniejsze pojęcie pochodnej cząstkowej, jako szczególny przypadek pochodnej kierunkowej.

Definicja 14. Niech $E \subset \mathbb{R}^n$ będzie zbiorem otwartym, $x \in E$ i $f : E \to \mathbb{R}^m$, tj. $f(x) = (f_1(x), \ldots, f_m(x))$.
(a) Jeśli $0 \neq v \in \mathbb{R}^n$, to wektor o współrzędnych

$$\left(\frac{\partial f}{\partial v}\right)_i (x) = \lim_{t\to 0} \frac{f_i(x+tv) - f_i(x)}{t} \quad i = 1, \ldots, n,$$

nazywamy *pochodną kierunkową* w kierunku wektora v w punkcie x.
(b) Jeśli $v = \mathbf{e}_j$ jest wektorem z bazy standardowej w $\mathbb{R}^n$, to $\frac{\partial f}{\partial \mathbf{e}_j}$ nazywamy *pochodną cząstkową* i piszemy wtedy $\frac{\partial f}{\partial x_j}(x)$ zamiast $\frac{\partial f}{\partial \mathbf{e}_j}(x)$.

Z definicji wynika, że $\frac{\partial f_i}{\partial x_j}(x)$ jest pochodną funkcji jednej zmiennej:

$$\zeta \mapsto f_i(x_1, \ldots, x_{j-1}, \zeta, x_{j+1}, \ldots, x_n),$$

gdzie na miejscach poza j-tym mamy stałe.

Korzyść z nowego pojęcia będzie łatwo widoczna. Zauważmy, że jeśli funkcja f jest różniczkowalna w punkcie x, to można położyć $h = t\mathbf{e}_i$ w definicji 13. Wtedy dostaniemy

$$\lim_{h\to 0} \frac{\|f(x+h) - f(x) - Df(x)h\|}{\|h\|} = \lim_{t\to 0} \frac{\|f(x+t\mathbf{e}_i) - f(x) - Df(x)t\mathbf{e}_i\|}{t} = 0,$$

z jednoznaczności granicy.

Jeśli utożsamimy odwzorowanie $Df(x)$ z jego macierzą (patrz §2.4), to jej elementy łatwo już odczytać

$$\frac{\partial f_i(x)}{\partial x_j} = (Df(x)\mathbf{e}_j)_i = (Df(x))_{ij}, \tag{1}$$

tym samym wykazaliśmy następujący fakt.

Wniosek 15. Istnienie pochodnej w punkcie x pociąga istnienie pochodnych cząstkowych w punkcie x. □

Powyższy wzór pozwala na praktyczne obliczenie $Df(x)h$.

Wniosek 16. Jeśli funkcja f jest różniczkowalna w x, to

$$Df(x)h = \sum_{i=1}^{n} h_i \frac{\partial f}{\partial x_i}.$$

Dowód. Z (1) wynika, że

$$\begin{aligned}(Df(x)h)_j &= (Df(x)\sum_{i=1}^{n} h_i\mathbf{e}_i)_j = \left(\sum_{i=1}^{n} h_i Df(x)\mathbf{e}_i\right)_j \\ &= \left(\sum_{i=1}^{n} h_i \frac{\partial f}{\partial x_i}(x)\right)_j = \sum_{i=1}^{n} h_i \frac{\partial f_j}{\partial x_i}(x)\end{aligned}$$

□

Zastosujmy ten wniosek do praktyki.

Przykład 6. Niech g będzie funkcją z przykładu 5(b). Obliczymy jej pochodną

$$Dg(x_1, x_2) = \begin{bmatrix} e^{x_1}\cos x_2 & -e^{x_1}\sin x_2 \\ e^{x_1}\sin x_2 & e^{x_1}\cos x_2 \end{bmatrix}.$$

Narzuca się teraz nowe pytanie: Czy może prawdziwe jest twierdzenie odwrotne, tj. czy z istnienia pochodnych cząstkowych wynika istnienie pochodnej. Na ogół odpowiedź jest przecząca. Wystarczy rozpatrzeć funkcję h z przykładu 5(c). Wtedy $\frac{\partial h}{\partial x_i}(0,0) = 0$, $i = 1, 2$, ale funkcja jest nieciągła w $(0, 0)$, więc nieróżniczkowalna na mocy stwierdzenia 14.

W wielu interesujących przypadkach odpowiedź jest na szczęście twierdząca.

Twierdzenie 17. Niech E będzie otwartym podzbiorem $\mathbb{R}^n$ i $f : E \to \mathbb{R}^m$. Jeśli $\frac{\partial f_i}{\partial x_j}(x)$ są funkcjami ciągłymi dla $i = 1, \ldots, m$, $j = 1, \ldots, n$, to funkcja f jest różniczkowalna w E.

Pozostawimy ten fakt bez dowodu.

Poprzednie twierdzenie uzasadnia następującą definicję.

Definicja 15. Powiemy, że funkcja $f : E \to \mathbb{R}^m$ jest klasy C^1, jeśli wszystkie pochodne cząstkowe istnieją i są ciągłe.

Na koniec podrozdziału sformułujemy twierdzenie o pochodnej funkcji złożonej.

Twierdzenie 18. Załóżmy, że $E \subset \mathbb{R}^n$, $U \subset \mathbb{R}^m$ są otwarte i $f : E \to \mathbb{R}^m$, zaś $g : U \to \mathbb{R}^k$, nadto $f(E) \subset U$. Załóżmy też, że f jest różniczkowalna w punkcie x_0, zaś g jest różniczkowalna w punkcie $f(x_0)$. Wtedy funkcja $F : E \to \mathbb{R}^k$ określona wzorem

$$F(x) = g(f(x))$$

jest różniczkowalna w punkcie x_0 oraz

$$DF(x_0) = Dg(y)|_{y=f(x_0)} \cdot Df(x_0).$$

Podkreślamy, że kropka oznacza tutaj mnożenie macierzy. Pozostawiamy je bez dowodu, z uwagi na złożoność pojęciową, mimo iż na poziomie rachunków dowód wygląda podobnie jak dla funkcji liczbowych, (patrz twierdzenie 3.31).

Wskażemy teraz na związek funkcji klasy C^1 i spełniających warunek Lipschitza.

Stwierdzenie 19. Załóżmy, że $\Phi : B(0,r) \subset \mathbb{R}^m \to \mathbb{R}^n$ jest klasy C^1 i

$$M = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} \max_{x\in B(0,r)} \left(\frac{\partial \Phi_i(x)}{\partial x_j}\right)^2} < \infty.$$

Wtedy Φ spełnia warunek Lipschitza ze stałą M.

Dowód. Ustalmy współrędną i, $1 \le i \le n$ i dwa dowolne punkty $x, y \in B(0,r)$. Wprowadzamy funkcję pomocniczą

$$g_i(t) = \Phi_i(yt + (1-t)x).$$

Wtedy $g_i(1) = \Phi(y)$, $g_i(0) = \Phi(x)$ i

$$g_i'(t) = D\Phi_i(yt + (1-t)x)(y-x) = \sum_{j=1}^{m} \frac{\partial \Phi_i}{\partial x_j}(yt + (1-t)x)(y_j - x_j).$$

Z twierdzenia o wartości średniej dostaniemy

$$g_i(1) - g_i(0) = g_i'(c)(1-0) = \sum_{j=1}^{m} \frac{\partial \Phi_i}{\partial x_j}(yc + (1-c)x)(y_j - x_j),$$

gdzie $c \in [0,1]$. Z nierówności Schwarza mamy

$$|\Phi_i(y) - \Phi_i(x)| \le \sqrt{\sum_{j=1}^{m} \left(\frac{\partial \Phi_i}{\partial x_j}(yc + (1-c)x)\right)^2}\, \|y - x\| \le \max_{\xi\in B(0,r)} \sqrt{\sum_{j=1}^{m} \left(\frac{\partial \Phi_i}{\partial x_j}(\xi)\right)^2}\, \|y - x\|.$$

Stąd wynika

$$\|\Phi(y) - \Phi(x)\| \le M\|y - x\|. \qquad \square$$

Uwaga. Powyższy fakt jest prawdziwy jeśli zastąpić kulę dowolnym zbiorem otwartym G, na którym M zdefiniowane w podobny sposób jest wielkością skończoną. Szczegóły pomijamy.

4.4 Ekstrema lokalne

Zajmiemy się w tym podrozdziale badaniem ekstremów lokalnych. Temat ten w przypadku funkcji wielu zmiennych okaże się bardziej złożony, niż jego jednowymiarowy odpowiednik. Samo pojęcie drugiej pochodnej okaże się na tyle głębokie, że rozważymy tylko przypadek pochodnej drugiego rzędu funkcji o wartościach rzeczywistych, całkowicie pomijając funkcje wektorowe. Będziemy też musieli zająć się odpowiednią wersją twierdzenia Taylora, które sformułujemy tylko dla funkcji dwukrotnie różniczkowalnych.

Wykażemy najpierw warunek konieczny ekstremów lokalnych, taki jak dla funkcji jednej zmiennej.

Twierdzenie 20. Niech E będzie otwartym podzbiorem $\mathbb{R}^n$ $f : E \to \mathbb{R}$ i załóżmy, że $x_0 \in E$ jest lokalnym minimum (odpowiednio, lokalnym maksimum) i funkcja f jest różniczkowalna w x_0. Wtedy $Df(x_0) = 0$.

Dowód. Rozpatrzymy tylko przypadek minimum. W drugim przypadku wystarczy zamienić funkcję f na $-f$, aby maksimum sprowadzić do minimum.

Wprowadźmy funkcje pomocnicze g_i:

$$g_i(t) = f(x_0 + t\mathbf{e}_i),$$

gdzie $\mathbf{e}_i$, $i = 1, \ldots, n$ są wektorami bazy standardowej. Skoro E jest otwarty, to istnieje kula $B(x_0, r)$ zawarta w E. Zatem funkcje g_i są dobrze określone dla $t \in (-r, r)$. Co więcej, są one różniczkowalne w punkcie $t = 0$. Zatem z twierdzenia 3.32 mamy, że

$$\frac{dg_i}{dt}(0) = 0.$$

Z drugiej strony, twierdzenie 18 o pochodnej funkcji złożonej daje:

$$\frac{dg_i}{dt}(0) = Df(x_0)\frac{d(t\mathbf{e}_i)}{dt} = Df(x_0)\mathbf{e}_i = \frac{\partial f}{\partial x_i}(x_0) = 0.$$

Skąd wynika teza. □

Z powyższego twierdzenia wynika, że znalezienie największej lub najmniejszej wartości bądź ekstremów lokalnych wymaga znalezienia *punktów krytycznych* funkcji f, tj. miejsc zerowania się pochodnej. Jeśli szukamy maksimum/minimum w domknięciu zbioru otwartego, to znalezienie punktów krytycznych nie wystarcza, widzieliśmy to już w rozdziale 3. Trzeba osobno zbadać funkcję na brzegu obszaru, lecz ogólne potraktowanie tego tematu wykracza poza ramy niniejszej książki.

Przykład 7. Niech funkcja $f : \mathbb{R}^2 \to \mathbb{R}$ będzie dana wzorem:

$$f(x, y) = x^2 + xy^3 + y^4.$$

Liczymy pochodne cząstkowe:

$$\frac{\partial f}{\partial x} = 2x + y^3, \quad \frac{\partial f}{\partial y} = 3xy^2 + 4y^3,$$

tj. $Df(x,y) = (2x + y^3, 3y^2 + 4y^3)$. Szukamy miejsc zerowych $Df(x,y)$, dostaniemy układ

$$\begin{cases} 2x + y^3 = 0 \\ 3y^2 + 4y^3 = 0, \end{cases}$$

którego rozwiązaniami są pary $(0,0)$, $(\mp 8\sqrt{6}/9, \pm 2\sqrt{6}/9)$.

Zbadamy teraz tylko punkt (0,0). Podejrzewamy, że jest to minimum. Istotnie, na mocy nierówności $2xy^3 \leq x^2 + y^6$ mamy:

$$x^2 + xy^3 + y^4 \geq x^2 - \frac{1}{2}x^2 - \frac{1}{2}y^6 + y^4 = \frac{x^2}{2} + y^4\left(1 - \frac{y^2}{2}\right) \geq \frac{1}{2}(x^2 + y^4).$$

Ostatnia nierówność jest prawdziwa dla dostatecznie małych y.

W podrozdziale 3.5.5 widzieliśmy, jak badać charakter punktów krytycznych za pomocą drugiej pochodnej. W związku z tym wprowadzimy to pojęcie także dla wielu zmiennych.

4.5 Druga pochodna funkcji o wartościach rzeczywistych

Niech E będzie otwartym podzbiorem $\mathbb{R}^n$, zaś $f : E \to \mathbb{R}$ będzie różniczkowalna w E. Z definicji $Df(x)$ jest elementem $\mathrm{Hom}\,(\mathbb{R}^n, \mathbb{R})$, tj. przekształceniem liniowym. Jak wiemy, możemy je utożsamiać z macierzą o jednym wierszu i n kolumnach. Możemy je z kolei utożsamiać z wektorami, tj. macierzami o jednej kolumnie i n wierszach. Dlatego funkcji różniczkowalnej można przypisać wektor grad f nazywany *gradientem*,

$$\mathrm{grad}\, f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x)\right)^T.$$

Jeśli odwzorowanie $E \ni x \mapsto Df(x) \in \mathbb{R}^n$ jest różniczkowalne w punkcie x, to powiemy wtedy, że funkcja f jest *dwukrotnie różniczkowalna w punkcie* x. Piszemy wtedy $D^2f(x)$ na oznaczenie drugiej pochodnej. Zauważmy, że $D^2f(x) \in \mathrm{Hom}\,(\mathbb{R}^n, \mathbb{R}^n)$, tj. $D^2f(x)$ może być utożsamiona z macierzą kwadratową $n \times n$. Chcemy teraz wyznaczyć elementy macierzy $D^2f(x)$. Skoro $Df(x) \in \mathrm{Hom}\,(\mathbb{R}^n, \mathbb{R})$, to dla dowolnego ustalonego wektora $v \in \mathbb{R}^n$ funkcja

$$x \mapsto Df(x)v$$

ma wartości liczbowe. Przyjrzyjmy się jej pochodnej w punkcie x na wektorze w. Będzie to

$$D(Df(x)v)w$$

Połóżmy więc

$$D^2f(x)(w,v) := D(Df(x)v)w.$$

Gdy $v = (v_1, \ldots, v_n)$, $w = (w_1, \ldots, w_n)$, to mamy

$$Df(x)v = \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i} v_i$$

i dalej,

$$D(Df(x)v)w = \sum_{j=1}^{n} \frac{\partial}{\partial x_j} \sum_{i=1}^{n} \left(\frac{\partial f(x)}{\partial x_i} v_i \right) w_j.$$

Powyższy wzór nieco się uprości, jeśli przyjąć zapis

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(x) := \frac{\partial}{\partial x_j} \left(\frac{\partial f}{\partial x_i} \right)(x),$$

tj. ostatecznie mamy

$$D^2 f(x)(w, v) = \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\partial^2 f(x)}{\partial x_j \partial x_i} v_i w_j. \tag{2}$$

Jeśli wyżej przyjąć, że $v = \mathbf{e}_i$, $w = \mathbf{e}_j$ i jak zwykle $\mathbf{e}_k$ oznacza wektor bazy standardowej, to na mocy (2) wyznaczymy elementy $(D^2 f(x))_{ij}$. Mianowicie

$$(D^2 f(x))_{ij} \equiv D^2 f(x)(\mathbf{e}_i, \mathbf{e}_j) = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}.$$

Można teraz zapytać czy kolejność różniczkowania ma znaczenie. Odpowiedź jest podana niżej.

Twierdzenie 21. Niech E będzie otwartym podzbiorem $\mathbb{R}^n$. Załóżmy, że $f : E \to \mathbb{R}$ ma drugą pochodną $D^2 f(x)$ w punkcie $x \in E$. Wtedy

$$\frac{\partial^2 f(x)}{\partial x_j \partial x_i} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}.$$

Innymi słowy $D^2 f(x) = (D^2 f(x))^T$. Ten fakt pozostawimy bez dowodu.

Pozostaje problem sprawdzenia, kiedy istnieje druga pochodna funkcji w punkcie $x_0 \in E$. Otóż na mocy twierdzenia 17 ciągłość pochodnych cząstkowych

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(x)$$

w punkcie x_0 pociąga istnienie $D^2 f(x_0)$.

Przykład 6, cd. Obliczymy $D^2 f(x, y)$. Dostaniemy,

$$D^2 f(x, y) = \begin{bmatrix} 2 & 3y^2 \\ 3y^2 & 6yx + 12y^2 \end{bmatrix}.$$

Właściwie, to nic nie stoi na przeszkodzie, aby liczyć pochodne cząstkowe wyższych rzędów.

Definicja 16. Niech E będzie otwartym podzbiorem $\mathbb{R}^n$, $f : E \to \mathbb{R}$.
(a) Gdy $m \geq 0$, to piszemy

$$\frac{\partial^{m+1} f}{\partial x_{j_{m+1}} \partial x_{j_m} \dots \partial x_{j_1}}(x) := \frac{\partial f}{\partial x_{j_{m+1}}} \left(\frac{\partial^m f}{\partial \partial x_{j_m} \dots \partial x_{j_1}} \right)(x)$$

i nazywamy *pochodną cząstkową rzędu* $m + 1$. (b) Powiemy, że f jest klasy C^k jeśli pochodne cząstkowe k-tego rzędu istnieją i są ciągłe.

W warunkach poprzedniej definicji, jeśli f jest klasy C^2, to oznacza to, że f ma drugą pochodną w każdym punkcie zbioru E.

4.5.1 Twierdzenie Taylora

Możemy teraz sformułować wielowymiarową wersję twierdzenia Taylora

Twierdzenie 22. Niech E będzie otwartym podzbiorem $\mathbb{R}^n$ i $f : E \to \mathbb{R}$ będzie klasy C^2 i $x_0 \in E$. Wtedy

$$f(x) - f(x_0) = Df(x)(x - x_0) + \frac{1}{2}D^2f(x)(x - x_0, x - x_0) + R(x - x_0),$$

gdzie

$$\lim_{x \to x_0} \frac{\|R(x - x_0)\|}{\|x - x_0\|^2} = 0.$$

Dowód. Pokażemy tylko jego ideę. Dla pewnego $r > 0$ mamy $B(x_0, r) \subset E$. Niech teraz $x \in B(x_0, r)$ i kładziemy

$$v = \frac{x - x_0}{\|x - x_0\|},$$

wtedy $v \in \mathbb{R}^n$ jest wektorem o długości 1. Jednocześnie definiujemy funkcję pomocniczą $g : (-r, r) \to \mathbb{R}$, wzorem

$$g(t) = f(x_0 + tv).$$

Wtedy $g(0) = f(x_0)$, $g(r) = f(x)$. Z §3.5.3 wynika, że

$$g(t) - g(0) = g'(0)t + \frac{1}{2}g''(0)t^2 + \rho(t), \tag{3}$$

gdzie $|\rho(t)|/t^2 \to 0$, gdy $t \to 0$. Zauważmy, że dla $t = \|x - x_0\|$ dostaniemy

$$g'(0)t = Df(x_0)v, \quad g''(0)t^2 = D^2f(x_0)(v, v)$$

Zatem po wstawieniu do (3) dostalibyśmy tezę, ale ρ w (3) zależy od v i $\rho(t)/t^2 \to 0$ tylko wzdłuż v. Jednak nasze obawy są płonne i w istocie

$$|R(x - x_0)|/\|x - x_0\|^2 \to 0,$$

aczkolwiek nie udowodnimy tego. □

4.6 Warunki konieczne i dostateczne ekstremów lokalnych

Podamy teraz różniczkową charakteryzację ekstremów lokalnych. Jak się przekonamy nie jest ona pełna i nie mamy szans na taką.

Twierdzenie 23. Niech E będzie otwartym podzbiorem $\mathbb{R}^n$, załóżmy, że $f : E \to \mathbb{R}$, $x_0 \in E$ i f jest dwukrotnie różniczkowalna w punkcie x_0. Wtedy,

(a) (warunek konieczny) jeśli x_0 jest lokalnym maksimum (odpowiednio, minimum), to $Df(x_0) = 0$ i dla dowolnego $v \in \mathbb{R}^n$ mamy $Df^2(x_0)(v,v) \leq 0$ (odpowiednio, $Df^2(x_0)(v,v) \geq 0$);

(b) (warunek dostateczny) jeśli $Df(x_0) = 0$ i istnieje taka stała $\lambda > 0$, że dla każdego $v \in \mathbb{R}^n$ $Df^2(x_0)(v,v) < -\lambda\|v\|^2$ (odpowiednio, $Df^2(x_0)(v,v) > \lambda\|v\|^2$), to f ma w x_0 lokalne maksimum (odpowiednio, lokalne minimum).

Dowód. (a) Wiemy już, że $Df(x_0) = 0$. Rozpatrzymy tylko przypadek maksimum. Drugi przypadek sprowadzamy do pierwszego zamianą funkcji f na $-f$.

Do wykazania nierówności zastosujemy wzór Taylora dla $x = x_0 + vt \in E$, gdzie $v \in \mathbb{R}^n$ i $t \in \mathbb{R}$:

$$0 \geq f(x_0 + vt) - f(x_0) = tDf(x_0)v + \frac{1}{2}D^2f(x_0)(v,v)t^2 + R(tv).$$

tym samym,

$$0 \geq f(x_0 + vt) - f(x_0) = \frac{t^2}{2}\left(D^2f(x_0)(v,v) + 2R(tv)/t^2\right),$$

ale skoro $R(tv)/t^2$ dąży do zera, gdy $t \to 0$, to dostaniemy

$$D^2f(x_0)(v,v) \leq 0,$$

dla dowolnego $v \in \mathbb{R}^n$.

(b) Rozpatrzymy ponownie tylko przypadek maksimum. Znów użyjemy wzoru Taylora dla $x = x_0 + v$:

$$f(x) - f(x_0) = Df(x_0)v + \frac{1}{2}D^2f(x_0)(v,v) + R(v) = \frac{\|v\|^2}{2}\left(D^2f(x_0)(h,h) + \frac{2R(v)}{\|v\|^2}\right).$$

gdzie $h = v/\|v\|$. Z twierdzenia Taylora istnieje takie $\delta > 0$, że dla $\|v\| < \delta$ dostaniemy

$$|R(v)/\|v\|^2 < \lambda/4.$$

Tym samym uzyskamy,

$$f(x) - f(x_0) \leq \frac{\|v\|^2}{2}\left(-\lambda\|h\|^2 + 2\lambda/2\right) = -\frac{\lambda}{4}\|v\|^2,$$

bo $\|h\| = 1$. Zatem f przyjmuje w punkcie x_0 lokalne maksimum. □

Doświadczenie z funkcjami jednej zmiennej podpowiada, że nie możemy nic poprawić w powyższym twierdzeniu, tj. nie możemy zamienić nierówności nieostrej na ostrą w (a) czy ostrej na nieostrą w (b), (patrz §3.5.5).

Natomiast powinniśmy zająć się objaśnieniem nierówności występujących w tezie twierdzenia. Mają one nawet własne nazwy.

4.6.1 Macierze dodatnio i ujemnie określone

Definicja 17. Powiemy, że macierz kwadratowa $A \in M_{n\times n}(\mathbb{R})$ jest *dodatnio* (odpowiednio, *ujemnie*) określona, jeśli dla dowolnego $v \in \mathbb{R}^n$

$$(Av, v) > 0 \qquad \text{(odpowiednio, } (Av, v) < 0).$$

Niestety, jest więcej możliwości niż dwie w/w, bo macierz może jeszcze spełniać nierówność nieostrą lub być *nieokreślona*, wtedy gdy dla pewnych wektorów v, w mamy $(Av, v) > 0$ oraz $(Aw, w) < 0$, np. $A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$ ma to do siebie, że $(A\mathbf{e}_1, \mathbf{e}_1) = -1$, zaś $(A\mathbf{e}_2, \mathbf{e}_2) = 1$. Tym niemniej, rozpoznawanie określoności macierzy jest dość proste rachunkowo. Mamy bowiem:

Twierdzenie 24. (Sylvestra, I) Załóżmy, że $A \in M_{n\times n}(\mathbb{R})$ i $A = A^T$. Wtedy następujące warunki (a) i (b) są równoważne:

(a) A jest dodatnio określona;

(b) $\det A_i > 0$, $i = 1, \ldots, n$, gdzie $A_i \in M_{i\times i}(\mathbb{R})$ jest macierzą, która powstaje poprzez wykreślenie z A wierszy i kolumn o wyrazach o numerach większych niż i. Uwaga, piszemy $A_n = A$.

Charakteryzację ujemności uzyskuje się zastępując macierz A macierzą $-A$. Zastosowanie powyższego twierdzenia daje następujący wynik:

Twierdzenie 25. (Sylvestra, II) Załóżmy, że $A \in M_{n\times n}(\mathbb{R})$ i $A = A^T$. Wtedy warunki (a) i (b) są równoważne:

(a) A jest ujemnie określona;

(b) $(-1)^i \det A_i > 0$, $i = 1, \ldots, n$, gdzie A_i są j.w.

Sformułowane wyżej kryterium Sylvestra, wydaje się odnosić do słabszego faktu niż ten wskazany w twierdzeniu 23. Tam żądamy, by $(Av, v) \geq \lambda\|v\|^2$, tu zaś tylko $(Av, v) > 0$. W istocie obie nierówności są równoważne.

Stwierdzenie 26. Załóżmy, że $A \in M_{n\times n}(\mathbb{R})$ i $A = A^T$. Jeśli dla wszystkich $0 \neq v \in \mathbb{R}^n$ mamy $(Av, v) > 0$, to istnieje takie $\lambda > 0$, że

$$(Av, v) \geq \lambda\|v\|^2.$$

Dowód. Wykazaliśmy w przykładzie 5(e), że funkcja $\mathbb{R}^n \ni v \to f(v) = (Av, v) \in \mathbb{R}$ jest ciągła. Tym samym na mocy twierdzenia 13 f przyjmuje kresy na zbiorach domkniętych i ograniczonych. Takim zbiorem jest sfera $S = \{x \in \mathbb{R}^n : \ 1 = \sum_{i=1}^n x_i^2\}$, bo jest przeciwobrazem zbioru domkniętego (punktu) przy odzwzorowaniu ciągłym, (patrz stwierdzenie 12). Niech zatem

$$\lambda = \min_{v\in S} f(v) = f(v_0) > 0.$$

Jeśli teraz $v \in \mathbb{R}^n$ jest dowolnym wektorem, to

$$f(v) = (Av, v) = \|v\|^2 \left(A\frac{v}{\|v\|}, \frac{v}{\|v\|} \right) \geq \lambda\|v\|^2.$$

Co należało wykazać. □

W przypadku macierzy 2 na 2 możliwy jest pełniejszy opis sytuacji. Co więcej dowód jest bardzo prosty. Będzie to jedyne kryterium, które wykażemy.

Twierdzenie 27. (kryterium Sylvestra, III) Załóżmy, że $A \in M_{2\times 2}(\mathbb{R})$ i A jest symetryczną macierzą. Wtedy,

(a) jeśli $\det A < 0$, to A jest nieokreślona;

(b) warunki $\det A > 0$ i $a_{11} > 0$ są równoważne dodatniości A;

(b) warunki $\det A > 0$ i $a_{11} < 0$ są równoważne ujemności A.

Dowód. Weźmy dowolny wektor $v \in \mathbb{R}^2$, $v = (x, y)$, wtedy

$$\begin{aligned}(Av, v) &= a_{11}x^2 + a_{12}xy + a_{21}xy + a_{22}y^2 \\ &= a_{11}x^2 + 2a_{12}xy + a_{22}y^2 =: Q_y(x).\end{aligned}$$

Nierówność $(Av, v) > 0$ możemy potraktować jako nierówność za względu na x, która ma być spełniona dla wszystkich x. Jest rzeczą oczywistą, że

$$a_{11}x^2 + 2a_{12}xy + a_{22}y^2 > 0 \quad \text{dla wszystkich } x \in \mathbb{R} \tag{4}$$

jest równoważne warunkom

$$a_{11} > 0 \quad \text{i} \quad 0 > \Delta = 4(a_{12}^2y^2 - a_{11}a_{22}y^2) = -4y^2 \det A,$$

o ile $y \neq 0$. A jeśli $y = 0$, to (4) redukuje się do

$$a_{11}x^2 > 0.$$

Powyższe rachunki pokazują, że

$$(Av, v) < 0$$

jest równoważne warunkom

$$\det A > 0 \quad \text{i} \quad a_{11} < 0.$$

Nadto, $\det A < 0$ jest równoważne temu, że $\Delta > 0$, czyli funkcja $x \mapsto Q_y(x)$ zmienia znak, o ile $y \neq 0$. □

Powyższe twierdzenie ma natychmiastowe zastosowanie w badaniu funkcji.

Wniosek 28. Jeśli $E \subset \mathbb{R}^2$ jest otwarty, $a \in E$, $f : E \to \mathbb{R}$, $Df(a) = 0$ i $D^2f(a)$ istnieje, i $\det D^2f(a) < 0$, to funkcja f nie ma ekstremum w punkcie a. Jest tam punkt określany jako *siodłowy*.

Przykładu 6 ciąg dalszy. Przypominamy $D^2f(x, y)$:

$$D^2f(x, y) = \begin{bmatrix} 2 & 3y^2 \\ 3y^2 & 6yx + 12y^2 \end{bmatrix},$$

Zatem dla punktu krytycznego (0,0) mamy

$$D^2 f(0,0) = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}.$$

Wtedy dla dowolnego $v = (v_1, v_2)$ dostaniemy

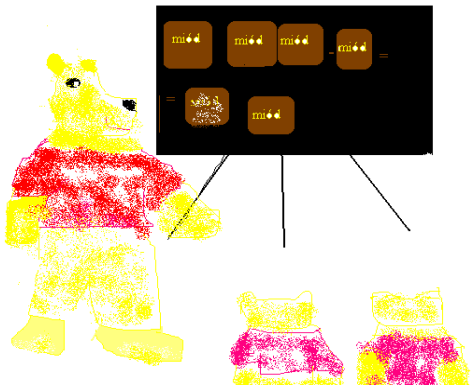
$$D^2 f(0,0)(v,v) = 2v_1^2 \geq 0,$$

zgodnie z twierdzeniem 17, ale $D^2 f(0,0)$ nie jest dodatnio określona!

Zauważmy, dla pozostałych punktów krytycznych (x_i, y_i), $i = 1, 2$, funkcji f mamy

$$D^2 f(x_i, y_i) = \begin{bmatrix} 2 & 8 \\ 8 & 32/3 \end{bmatrix}.$$

Nadto, $\det D^2 f(x_i, y_i) = 64(2/3 - 1) < 0$, tj. nie ma ekstremum w (x_i, y_i), $i = 1, 2$.



Rozdział 5

Równania różniczkowe zwyczajne

5.1 Wprowadzenie

Przedstawimy teraz przykłady paru zjawisk z dziedziny fizyki i biologii.

Przykład 1. Rozpad promieniotwórczy opisuje prawo mówiące, że ubytek substancji jest proporcjonalny do jej ilości:

$$\frac{dN}{dt} = -kN,$$

gdzie $N(t)$ jest masą substancji w chwili t, zaś $k > 0$ jest stałą proporcjonalności.

Przykład 2a. Wzrost populacji bakterii w warunkach dostatku pożywienia jest opisywany podobnie: przyrost liczby bakterii jest proporcjonalny do ich ilości:

$$\frac{dN}{dt} = kN,$$

gdzie $N(t)$ oznacza liczbę bakterii w chwili t, zaś $k > 0$ jest stałą proporcjonalności. Zauważmy, że dokonaliśmy w obu przypadkach koniecznego uproszczenia modelowego przyjmując, że tak masa substancji, jak ilość bakterii są ciągłymi funkcjami czasu. W rzeczywistości tak nie jest, bo obie wielkości zmieniają się skokowo. Jednak z uwagi na to, że owe skoki są malutkie w porównaniu do całej populacji cząstek czy bakterii, to owa nieścisłość nie ma większego praktycznego znaczenia.

Podkreślamy, że dobór modelu zjawiska jest sprawą niezależną od matematyki, co za chwilę zobaczymy.

Przykład 2b. Wzrost populacji bakterii w warunkach dostatku pożywienia. Można przyjąć, że jest on proporcjonalny do liczby par:

$$\frac{dN}{dt} = kN^2,$$

gdzie jak poprzednio $N(t)$ oznacza liczbę bakterii w chwili t.

Możliwe są dalsze uściślenia powyższego przykładu uwzględniające śmiertelność.

Przykład 2c. Wzrost populacji bakterii uwzględniający prawdopodobieństwo śmierci:

$$\frac{dN}{dt} = kN^2 - rN,$$

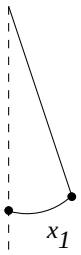
gdzie $r > 0$ jest stałą charakteryzująca śmiertelność.

Rozpatrzmy teraz przykład mechaniczny. Jeśli teraz x oznacza drogę przebytą przez cząstkę pod działaniem siły F, to wtedy dx/dt jest prędkością, d^2x/dt^2 zaś przyspieszeniem. Mechanika Newtona mówi, że

$$mx'' = F,$$

gdzie napis x'' oznacza do samo co, d^2x/dt^2, czy $\frac{d^2x}{dt^2}$; m to masa. Siła F może zależeć od czasu, położenia i prędkości, tj. $F = F(t, x, x')$. Szczególnym przypadkiem układu mechanicznego jest wahadło matematyczne.

Przykład 3. Ruch punktu materialnego o masie m na sztywnej nieważkiej nici o długości l w polu grawitacyjnym ziemi, (wahadło matematyczne).



Rys. 1. Wahadło matematyczne

Kąt wychylenia od pionu oznaczamy przez x_1, wtedy z bilansu sił wynika, że jeśli x_2 oznacza prędkość kątową wahadła, to

$$\begin{aligned} x_1' &= x_2 \\ mlx_2' &= -mg \sin x_1, \end{aligned} \tag{1}$$

a po uproszczeniach układ przyjmuje postać

$$\frac{d^2x_1}{dt^2} = -k \sin x_1, \tag{2}$$

gdzie $k = g/l$.

Przykład 4. Równanie oscylatora harmonicznego dostajemy zakładając, że wychylenia od położenia równowagi są małe, przez co można z dobrą dokładnością przybliżać funkcję sinus funkcją liniową (patrz §3.5.4):

$$\frac{d^2x}{dt^2} = -kx.$$

Jest jasnym, że wszystkie powyższe równania trzeba uzupełnić o opis tego, co dzieje się w chwili początkowej, jeśli chcemy wyznaczyć ilościowy opis procesu: w przykładach 1 i 2 jest to początkowa ilość bakterii:

$$N(t_0) = N_0.$$

Dla zagadnień mechanicznych są to: początkowe wychylenie i początkowa prędkość,

$$x(t_0) = x_0, \quad x'(t_0) = v_0.$$

Naszym celem jest poznanie metod rozwiązywania najprostszych typów równań i wskazanie metod badania rozwiązań, gdy znalezienie wzoru na rozwiązanie jest niemożliwe lub niepraktyczne.

Przyjrzyjmy się równaniu w przykładzie 2c. Niech $f(N)$ oznacza prawą jego stronę. Zauważmy że:

a) jeśli $N < 0$ lub $N > r/k$, to $f(N) > 0$;

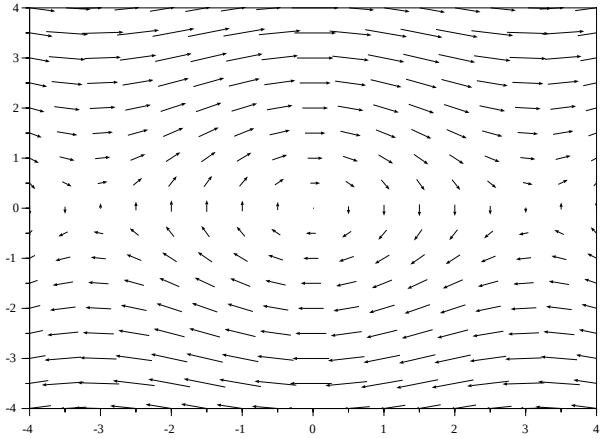
b) jeśli $N \in (0, r/k)$, to $f(N) < 0$;

c) jeśli $N = 0$ lub $N = r/k$, to $f(N) = 0$.

Jeśli teraz $0 < N_0$ spełnia nierówność w a), to wielkość $N(t)$ będzie rosła i jeśli $N_0 < 0$, to wielkość $N(t)$ będzie malała dla wszystkich $t > t_0$, nb. założenie $N_0 < 0$ jest niefizyczne.

Jeśli N_0 spełnia b), to wielkość $N(t)$ będzie malała i zawsze pozostanie w przedziale $(0, r/k)$. Co więcej, funkcja N jest malejąca aż do wymarcia populacji.

Jeśli zachodzi trzecia możliwość, tj. $N = 0$ lub $N = r/k$, to funkcja stała $N(t) = N_0$ jest rozwiązaniem równania. Zauważmy, że zdobywaliśmy wiedzę o rozwiązanich bez rozwiązywania równania.

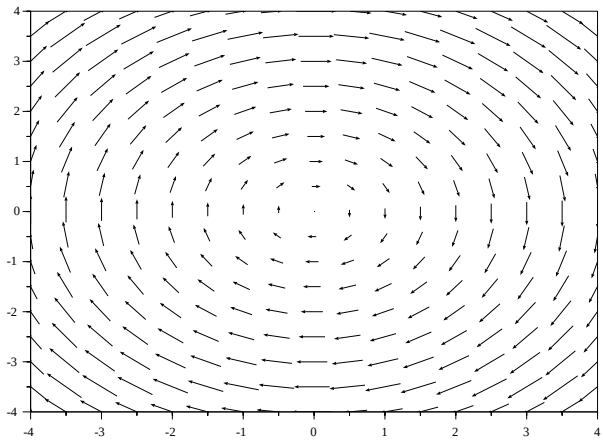


Rys. 2. Pole wektorowe układu (1)

Badanie układu (1) wymaga robienia rysunków pól wektorowych. Mianowicie, *polem wektorowym* nazywamy dowolne odwzorowanie $v : \mathbb{R}^n \to \mathbb{R}^n$. W naszym przypadku $n = 2$. Zróbmy rysunki pól wektorowych zadawanych układem (1) (patrz rys. 2) i równaniem (2)

rozpisanym jako układ za pomocą wprowadzenia zmiennej $y = x'$, (patrz rys. 3). Strzałki pokazują jak się będzie poruszać cząstka pod wpływem pola.

Znów warto podkreślić, że zdobyliśmy wiedzę o rozwiązaniu bez znajdowania go. Często nie można go zapisać wzorem albo znany wzór jest nieczytelny. Dlatego trzeba się posługiwać innymi metodami (np. sporządzaniem rysunków) do badania rozwiązań.



Rys. 3. Pole wektorowe układu (2)

Powyższe szkice pól wektorowych zostały uzyskane z pomocą pakietu do obliczeń numerycznych `scilab`. Tenże pakiet pozwala na numeryczne rozwiązywanie równań.

5.2 Najprostsze typy równań i ich rozwiązywanie

Równaniem różniczkowym zwyczajnym nazywamy równanie postaci

$$\frac{dy}{dt} = f(y,t),$$

gdzie $y \in \mathbb{R}^n$ i $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$. W istocie jest to układ, gdy $n > 1$. Przekonamy się w §5.4, że wszystkie dotychczasowe przykłady mogą być tak zapisane. Równania różniczkowe zwyczajne są zazwyczaj uzupełniane warunkiem początkowym

$$y(t_0) = y_0.$$

Rozpatrzmy najprostszy przykład, w którym funkcja $f : \mathbb{R} \to \mathbb{R}$ jest ciągła:

$$\frac{dy}{dt} = f(t), \quad y(t_0) = y_0.$$

Zadanie sprowadza się do znalezienia funkcji pierwotnej f. Bo jeśli F jest funkcją pierwotną i do tego taką, że $F(t_0) = y_0$, to jest ona rozwiązaniem powyższego zagadnienia.

Następny przykład jest nieco bardziej złożony, ale $g : \mathbb{R} \to \mathbb{R}$ też jest ciągła:

$$\frac{dy}{dt} = g(y), \quad y(t_0) = y_0. \tag{3}$$

Jeśli $g(y_0) = 0$, to szukanym rozwiązaniem jest funkcja stała $y(t) = y_0$. Jak później zobaczymy przy odpowiednich założeniach jest to jedyne rozwiązanie. Niech teraz $g(y_0) \neq 0$. Możemy podzielić obie strony (3) przez $g(y)$,

$$\frac{1}{g(y)}\frac{dy}{dt} = 1$$

a wynik scałkować na $[t_0, t]$,

$$\int_{t_0}^{t} \frac{1}{g(y)}\frac{dy}{ds}\, ds = t - t_0.$$

Twierdzenie o całkowaniu przez podstawienie zastosowane do lewej strony daje nam

$$\int_{y(t_0)}^{y(t)} \frac{1}{g(y)}\, dy = t - t_0.$$

Niech G będzie taką funkcją pierwotną funkcji $1/g(y)$, że $G(y_0) = t_0$. Zauważmy, że G jest monotoniczna w pewnym otoczeniu punktu y_0 a zatem funkcja odwrotna G^{-1} istnieje. Wtedy funkcja $y(t) = G^{-1}(t)$ jest rozwiązaniem. Sprawdzamy, że

$$\frac{dG^{-1}}{dt}(t) = \frac{1}{G'(y)} = g(y).$$

Ponadto na mocy definicji $y(t_0) = y_0$.

Przykład 5. Jako zastosowanie powyższych rozważań przedstawimy rozwiązanie równania z przykładu 1:

$$y' = -ky, \quad y(t_0) = y_0.$$

Po podzieleniu obu stron przez y dostaniemy:

$$\frac{y'}{y} = -k.$$

Teraz całkujemy względem t na przedziale $[t_0, t]$,

$$\int_{t_0}^{t} \frac{1}{y}\frac{dy}{dt}\, ds = \int_{t_0}^{t} -k\, dt = -k(t - t_0). \tag{4}$$

Twierdzenie o całkowaniu przez podstawienie zastosowane do lewej strony daje nam

$$\int_{t_0}^{t} \frac{1}{y}\frac{dy}{dt}\, ds = \int_{y(t_0)}^{y(t)} \frac{1}{y}\, dy = \ln|y| - \ln|y_0|. \tag{5}$$

Wyznaczamy y z (4) i (5):

$$|y| = |y_0| e^{-k(t-t_0)}$$

i dalej,

$$y(t) = y_0 e^{-k(t-t_0)}. \tag{6}$$

Wynika stąd, że ilość materiału radioaktywnego maleje wykładniczo w czasie.

Następny przykład łączy oba poprzednie.

$$\frac{dy}{dt} = g(y)f(t), \quad y(t_0) = y_0. \tag{7}$$

Równanie (7) nazywa się *równaniem o zmiennych rozdzielonych*. Postępujemy jak poprzednio. Niech $g(y_0) \neq 0$. Możemy podzielić obie strony (7) przez $g(y)$, dostaniemy,

$$\frac{1}{g(y)}\frac{dy}{dt} = f(t).$$

Niech G będzie dowolną funkcją pierwotną funkcji $1/g(y)$, a F funkcją pierwotną f. Wtedy jeśli scałkujemy obie strony (7) w granicach od t_0 do t, to dostaniemy,

$$\int_{t_0}^{t} \frac{1}{g(y)}\frac{dy}{dt}\, ds = \int_{t_0}^{t} f(s)\, ds = F(t) - F(t_0).$$

Lewą stronę potraktujemy podobnie jak wyżej

$$\int_{t_0}^{t} \frac{1}{g(y)}\frac{dy}{dt}\, ds = \int_{y(t_0)}^{y(t)} \frac{1}{g(y)}\, dy = G(y(t)) - G(y(t_0))$$

i dalej

$$y(t) := G^{-1}(F(t) - F(t_0) + G(y_0)). \tag{8}$$

Sprawdzamy, tak zdefiniowane y jest rozwiązaniem:

$$\frac{dy}{dt} = \frac{1}{G'(y)}\frac{dF(t)}{dt} = g(y)f(t).$$

Oczywiście $y(t_0) = G^{-1}(G(y_0)) = y_0$.

Przykład 6. W praktyce wygląda to tak. Mamy do rozwiązania

$$y' = 2y/t, \quad y(1) = 1.$$

Dzielimy obie strony przez y a wynik całkujemy

$$\int \frac{dy}{y} = 2\int \frac{dt}{t}.$$

Skąd mamy, że $\ln|y| = 2\ln|t| + C$, tj. $|y| = C|t|^2$. Dzięki warunkowi początkowemu dostaniemy: $y = t^2$.

Jednak sama postać wzoru (8) podpowiada, że jawna postać rozwiązania może być trudna do uzyskania, bądź niepraktyczna w obliczeniach.

5.3 Równania liniowe

5.3.1 Równania liniowe pierwszego rzędu

Podamy teraz metody rozwiązywania liniowych równań pierwszego rzędu. Pierwszym z nich jest równanie z przykładu 1, tj.

$$x'(t) = kx(t), \quad x(0) = C$$

Jego rozwiązanie podane jest wzorem (6), tj.

$$x(t) = Ce^{kt}. \tag{9}$$

Rozpatrzmy równanie liniowe z członem źródłowym,

$$x'(t) = kx(t) + f(t), \qquad x(0) = x_0. \tag{10}$$

W ogólności nie jest to równanie o zmiennych rozdzielonych. Zastosujemy tutaj *metodę uzmienniania stałej*. Polega ona na tym, że stałą C występującą we wzorze (9) traktujemy jako funkcję zmiennej t. Po wstawieniu do (10) dostaniemy

$$C'e^{kt} + kCe^{kt} = kCe^{kt} + f(t),$$

a stąd $C'e^{kt} = f(t)$, które jest równaniem na C pierwszego rozpatrywanego typu. Proste rachunki dają

$$x(t) = x_0e^{kt} + e^{kt}\int_0^t e^{-ks}f(s)\,ds.$$

W szczególnych przypadkach można łatwo zgadnąć postać rozwiązań równania (10). Jednym z tych przypadków jest

$$f(t) = (P_1(t)\cos\alpha t + P_2(t)\sin\alpha t)e^{\beta t}, \tag{11}$$

gdzie P_1 i P_2 są wielomianami zmiennej t, zaś $\alpha, \beta \in \mathbb{R}$. Powiemy wtedy, że f jest *quasi-wielomianem*.

Okazuje się, że jeśli f jest quasi-wielomianem, takim jak w (11), to rozwiązania równania (10) przyjmują następującą postać

$$y(t) = x_0e^{tk} + (Q_1(t)\cos\alpha t + Q_2(t)\sin\alpha t)e^{\beta t} \tag{12}$$

gdzie x_0 jest stałą zaś Q_1, Q_2 są wielomianami, których współczynniki trzeba dopiero określić (stąd nazwa: *metoda współczynników nieoznaczonych*) poprzez wstawienie (12) do (10). Widać od razu, że stopień Q_i nie może być niższy niż stopień P_i, $i = 1, 2$.

Przykład 7. Załóżmy, że $k \neq 1$. Rozpatrzmy

$$y' = ky + te^t \tag{13}$$

Wtedy $P_1 = t$, $P_2 = 0$, $\alpha = 0$, $\beta = 1$. Bierzemy $Q_1 = c_0 + c_1 t$ i $Q_2 = 0$. Zatem

$$y(t) = x_0 e^{kt} + (c_0 + c_1 t)e^t; \quad y' = e^{kt} k x_0 + e^t(c_0 + c_1 + c_1 t).$$

Stąd dostaniemy po wstawieniu do równania (13)

$$kx_0 e^{kt} + e^t(c_0 + c_1 + c_1 t) = kx_0 e^{kt} + ke^t(c_0 + c_1 t) + te^t$$

Zatem

$$c_0 + c_1 + c_1 t = kc_0 + kc_1 t + t$$

skąd wynika układ

$$\begin{aligned} c_0 + c_1 &= kc_0 \\ c_1 &= kc_1 + 1. \end{aligned}$$

Jego rozwiązanie to $c_0 = -1/(1-k)^2$, $c_1 = 1/(1-k)$.

Uwaga. Wzory przestają być prawdziwe gdy $k = 1$. Należy wtedy szukać wielomianu kwadratowego Q, a nie liniowego jak wyżej.

5.3.2 Równania liniowe drugiego rzędu

Równanie oscylatora harmonicznego, tłumionego oscylatora i prądu w obwodzie z kondensatorem i cewką ma wspólną postać

$$a_2 x'' + a_1 x' + a_0 x = 0 \quad \text{z warunkami } x'(0) = v, \quad x(0) = x_0. \tag{14}$$

Dotychczasowe doświadczenie z równaniami liniowymi podpowiada nam, że możemy szukać rozwiązań w postaci

$$x(t) = e^{rt},$$

gdzie r jest jeszcze nie ustalone. Wstawmy tę funkcję do (14). Po zróżniczkowaniu i skróceniu przez e^{rt} dostaniemy, że

$$a_2 r^2 + a_1 r + a_0 = 0. \tag{15}$$

Równanie (15) nazywa się *równaniem charakterystycznym* równania (14).

Mamy 3 przypadki do zbadania w zależności od znaku wyróżnika $\Delta = a_1^2 - 4a_2 a_0$ równania (15).

(1) $\Delta > 0$. Mamy wtedy 2 różne pierwiastki rzeczywiste równania (15) r_1 i r_2 a tym samym 2 rozwiązania równania (14)

$$e^{r_1 t} \quad \text{i} \quad e^{r_2 t}.$$

Funkcje te traktowane jako wektory w przestrzeni wektorowej $C([0,T])$ są liniowo niezależne, czyli jest szansa, że istnieje kombinacja liniowa $C_1 e^{r_1 t} + C_2 e^{r_2 t}$ która spełnia warunki początkowe. Zauważmy, że dzięki liniowości (15) każda kombinacja liniowa rozwiązań jest rozwiązaniem.

(2) $\Delta = 0$. Istnieje tylko 1 pierwiastek rzeczywisty r_0 i mamy rozwiązanie $e^{r_0 t}$. Podejrzewamy, że to za mało rozwiązań! Istotnie, mamy jeszcze jedno $te^{r_0 t}$, co łatwo sprawdzić

$$\begin{aligned} a_2\tfrac{d^2}{dt^2}(te^{r_0t}) + a_1\tfrac{d}{dt}(te^{r_0t}) + a_0 &= (a_2r_0^2 + a_1r_0 + a_0)te^{r_0t} + (2a_2r_0 + a_1)e^{r_0t} \\ &= 0 + \frac{d}{dr}D(r)_{|r=r_0}e^{r_0t} \end{aligned}$$

gdzie $D(r) = a_2r^2 + a_1r + a_0$. Skoro $D(r)$ ma z założenia pierwiastek podwójny w $r = r_0$ to

$$D(r) = a_2(r - r_0)^2 \equiv a_2r^2 + a_1r + a_0 \quad \text{i} \quad \frac{dD}{dr}(r) = 2a_2r + a_1 = 2a_2(r - r_0).$$

Zatem,

$$\frac{dD}{dr}(r_0) = 0.$$

(3) $\Delta < 0$. Mamy wtedy 2 pierwiastki zespolone r_1, r_2 sprzężone tj. $r_1 = \overline{r_2}$. Funkcje o wartościach zespolonych $t \to e^{r_j t}$, $j = 1, 2$ są rozwiązaniami. Przypominamy, że liczby zespolone można utożsamiać z $\mathbb{R}^2$. Różniczkowanie funkcji o wartościach wektorowych (np. w $\mathbb{C}$ czyli $\mathbb{R}^2$) polega na różniczkowaniu każdej współrzędnej z osobna. Dzięki temu przekonamy się, że

$$\frac{d}{dt}e^{(\alpha+i\beta)t} = (\alpha + i\beta)e^{(\alpha+i\beta t)}$$

Chcemy koniecznie dostać rozwiązania rzeczywiste dla rzeczywistych danych początkowych oraz $a_0, a_1, a_2 \in \mathbb{R}$. Zauważmy, że dla $r_1 = \alpha + i\beta$, $r_2 = \alpha - i\beta$, mamy

$$\begin{aligned} y_1(t) &= e^{r_1t} + e^{r_2t} = e^{r_1t} + e^{\overline{r_1t}} = 2\mathrm{Re}(e^{r_1t}) = 2e^{\alpha t}\cos(\beta t); \\ y_2(t) &= \frac{1}{i}(e^{r_1t} - e^{r_2t}) = \frac{1}{i}(e^{r_1t} - e^{\overline{r_1t}}) = 2\mathrm{Im}(e^{r_1t}) = 2e^{\alpha t}\sin(\beta t). \end{aligned}$$

Dzięki liniowości równania dostaliśmy dwa różne rozwiązania.

Przykład 8. W przypadku oscylatora harmonicznego

$$x'' + kx = 0, \quad k > 0$$

dostaniemy równanie charakterystyczne

$$r^2 + k = 0$$

i stąd $r_{1,2} = \pm i\sqrt{k}$, zatem

$$x_1(t) = \cos t\sqrt{k}, \qquad x_2(t) = \sin t\sqrt{k}$$

zgodnie z oczekiwaniami.

5.4 Teoria rozwiązalności

Zajmiemy się teraz odpowiedzią na pytanie, czy każde równanie ma rozwiązanie. Okazuje się, że odpowiedź jest twierdząca, przy rozsądnych założeniach. Wynik sformułujemy dla układów równań tak, aby obejmował równania wyższych rzędów. Jeśli mamy równanie

$$\frac{d^n y}{dt} = f\left(t, y, \frac{dy}{dt}, \ldots, \frac{d^{n-1}y}{dt^{n-1}}\right)$$

to podstawienia $y_1 = \frac{dy}{dt}, \ldots, y_{n-1} = \frac{d^{n-1}y}{dt^{n-1}}$ dadzą układ

$$\begin{cases} \frac{dy}{dt} = y_1 \\ \vdots \\ \frac{dy_{n-1}}{dt} = f(t, y, y_1, ..., y_{n-1}). \end{cases}$$

Twierdzenie 1. Załóżmy, że $f : U \subset \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, gdzie U jest otwartym zbiorem, spełnia warunki:

(i) f jest funkcją ciągłą;

(ii) istnieje takie $K > 0$, że dla dowolnych $(t, y_1), (t, y_2) \in U$ mamy

$$||f(t, y_1) - f(t, y_2)|| \leq K\|y_1 - y_2\|.$$

Wtedy, jeśli $(t_0, y_0) \in U$, to istnieje $T > t_0$ i dokładnie jedna funkcja $y : [t_0, T] \to \mathbb{R}^n$ klasy C^1 spełniająca układ równań

$$\begin{cases} \frac{dy}{dt} = f(t, y) \\ y(t_0) = y_0 \end{cases} . \tag{16}$$

Dowód. Dla prostoty dowód przeprowadzimy w przypadku pojednynczego równania.

Podamy jedynie dowód istnienia, który jest ilustracją metody kolejnych przybliżeń, która ma wiele zastosowań. Pomijamy dowód jednoznaczności.

Niech $T > t_0$ będzie takie, że $q := (T - t_0)K < 1$ zauważmy, że jeśli $y(t)$ jest rozwiązaniem równania (16), to po jego scałkowaniu dostaniemy

$$y(t) - y_0 = \int_{t_0}^{t} \frac{dy}{ds} ds = \int_{t_0}^{t} f(s, y(s)) ds$$

Ta równość podpowiada nam następującą definicję

$$\begin{aligned} &y_{-1}(t) \equiv 0, \quad y_0(t) \equiv y_0, \quad \text{na} \quad [t_0, T] \\ &y_{k+1}(t) = y_0 + \int_{t_0}^{t} f(s, y_k(s)) ds \quad \text{na} \quad [t_0, T] \end{aligned}$$

Pokażemy, że szereg funkcyjny

$$\sum_{n=0}^{\infty} (y_n(t) - y_{n-1}(t)) \tag{17}$$

jest jednostajnie zbieżny. Jego sumy częściowe S_n to

$$\sum_{k=0}^{n}(y_k - y_{k-1}) = y_n.$$

Wiemy (patrz twierdzenie 3.57), że do wykazania jednostajnej zbieżności szeregu (17) wystarczy sprawdzić, że

$$|y_k(t) - y_{k-1}(t)| \leq a_k \quad \text{dla} \quad t \in [t_0, T], \tag{18}$$

gdzie szereg liczbowy $\sum_{k=0}^{\infty} a_n$ jest zbieżny.

Zauważmy, że w myśl definicji normy $||\cdot||_\infty$ w przestrzeni $C([0,T])$ nierówność (18) oznacza, iż

$$||y_k - y_{k-1}||_\infty \leq a_k.$$

Sprawdzamy (18)

$$y_{k+1}(t) - y_k(t) = \int_{t_0}^{t} [f(s, y_k(s)) - f(s, y_{k-1}(s))]\, ds$$

Zatem z właściwości całki Riemanna i założenia (ii)

$$\begin{aligned}
|y_{k+1}(t) - y_k(t)| &\leq \int_{t_0}^{t} |f(s, y_k(s)) - f(s, y_{k-1}(s))| ds \\
&\leq \int_{t_0}^{t} K|y_k(s) - y_{k-1}(s)| ds \\
&\leq \int_{t_0}^{t} K \max_{s\in[t_0,T]} |y_k(s) - y_{k-1}(s)| ds \\
&= \int_{t_0}^{t} K||y_k - y_{k-1}||_\infty ds = K(t-t_0)||y_k - y_{k-1}||_\infty \\
&\leq q||y_k - y_{k-1}||_\infty \leq q^{k+1}||y_0 - y_{-1}|| = q^{k+1}||y_0||,
\end{aligned}$$

gdzie $\|v\|_\infty$ oznacza normę z przykładu 4.1(4) w p.w. $C[t_0, T])$. Skoro $q < 1$, to $a_{k+1} := q^{k+1}||y_0||$ jest wyrazem zbieżnego szeregu geometrycznego. Zatem dzięki twierdzeniu 3.57 szereg $\sum_{k=0}^{\infty}[y_k - y_{k-1}]$ jest zbieżny jednostajnie do granicy, którą oznaczymy symbolem y^∞. Dzięki jednostajnej zbieżności mamy, że y^∞ jest funkcją ciągłą.

Nadto, twierdzenie 3.56 gwarantuje, że

$$\lim_{n\to\infty} \int_{t_0}^{t} f(s, y_n(s)) ds = \int_{t_0}^{t} f(s, y^\infty(s)) ds.$$

Zatem $y^\infty(t)$ spełnia

$$y^\infty(t) = y_0 + \int_{t_0}^{t} f(s, y^\infty(s)) ds. \tag{19}$$

Co więcej prawa strona jest różniczkowalna dzięki podstawowemu twierdzeniu rachunku różniczkowego i całkowego. Zatem (19) pociąga

$$\frac{dy^\infty}{dt}(t) = f(t, y^\infty(t)) \quad \text{i} \quad y^\infty(t_0) = y_0,$$

co należało wykazać. □

5.4.1 Uwagi na temat jakościowej teorii równań

Chcielibyśmy orzekać o właściwościach rozwiązań równania wahadła

$$\begin{cases} x_1' = x_2 \\ x_2' = -k \ \sin x_1, \end{cases} \tag{20}$$

bez konieczności rozwiązywania tego układu. W tym celu zdefiniujemy nową funkcję daną wzorem, $H(x_1, x_2) = \frac{1}{2}x_2^2 - k\cos x_1$. Zauważmy, że H ma tę właściwość, że jeśli $t \mapsto (x_1(t), x_2(t))$ jest rozwiązaniem układu (20), to

$$\frac{d}{dt}H(x_1(t), x_2(t)) = \frac{2}{2}x_2(t)x_2'(t) + k \ \sin x_1(t)x_1'(t) = x_2(t)x_2'(t) - x_2'(t)x_2(t) = 0. \tag{21}$$

Oznacza, to że rozwiązanie $t \mapsto (x_1(t), x_2(t))$ jest krzywą zawartą w poziomicy funkcji H. Może ono być krzywą zamknięta albo nie w zależności od danych początkowych. Jeśli wartość $H(x_1(0), x_2(0))$ jest mała, to istotnie można pokazać, że rozwiązania są krzywymi zamkniętymi.

Okazuje się, że dzięki funkcji H równanie (20) można zredukować do zagadnienia pierwszego rzędu, mianowicie dostaniemy:

$$x_2' = \sqrt{2}\sqrt{H(x_1(0), x_2(0)) + k\cos x_1}.$$

Uwaga. Układ (20) jest szczególnym przypadkiem *układu Hamiltona*, tj. układu postaci

$$\begin{cases} p' = \frac{\partial H}{\partial q}(p, q), \\ y' = -\frac{\partial H}{\partial p}(p, q) \\ p(0) = p_0 \quad q(0) = q_0, \end{cases} \tag{22}$$

gdzie $p, q \in \mathbb{R}^n$ i funkcję $H(p, q)$ nazywa się *Hamiltomianem* ($2n$ zmiennych!). Okazuje się, że jeśli $p(t)$, $q(t)$ jest rozwiązaniem (22), to

$$\frac{d}{dt}H(p(t), q(t)) \equiv 0.$$

Dowód. Jest to łatwe ćwiczenie, które pozostawiamy Czytelnikowi. □

Chcielibyśmy też przekonać się, ile jest prawdy w stwierdzeniu, że równanie oscylatora harmonicznego

$$\begin{cases} x_1' = x_2 \\ x_2 = -kx_1 \end{cases} \tag{23}$$

jest przybliżeniem równania wahadła (20) w przypadku małej amplitudy drgań. Zapiszmy (20) i (23) w równoważnej postaci równań drugiego rzędu:

$$x'' = -k\sin x, \quad y'' = -ky$$

i załóżmy, że wychylenie początkowe jest $0{,}25$ tj. $x(0) = y(0) = 0{,}25$, prędkość początkowa $x'(0) = y'(0) = 0$ i $0 < k \leq 1$ oraz interesuje nas przedział czasu od 0 do 1. Badamy różnicę $x - y$ za pomocą twierdzenia Taylora

$$x(t) - y(t) = x(0) - y(0) + (x' - y')(0)t + \frac{1}{2}t^2(x'' - y'')(c),$$

gdzie $c \in [0, t]$. Dzięki założeniom o x'' i y''

$$x(t) - y(t) = \frac{1}{2}t^2(-k \sin x + ky) = \frac{k}{2}t^2[(y - x)(c) + r(x(c))]$$

gdzie $|r(\eta)| \leq \frac{1}{6}|\eta|^3$ (dzięki twierdzeniu Taylora). Obliczenie maksimum prawej strony dla $t \in [0, 1]$ daje

$$|(x - y)(t)| \leq \frac{k}{2} \max_{s\in[0,t]} |x(s) - y(s)| + \frac{1}{2}t^2 k \frac{(0, 25)^3}{6}.$$

Zatem

$$\max_{s\in[0,t]} |(x - y)(s)| \leq \frac{k}{2} \max_{s\in[0,t]} |x(s) - y(s)| + \frac{t^2 k}{3 \cdot 2562}$$

i dalej

$$\max_{s\in[0,t]} |(x - y)(s)| \leq \frac{t^2 k}{768 \cdot (1 - \frac{k}{2})},$$

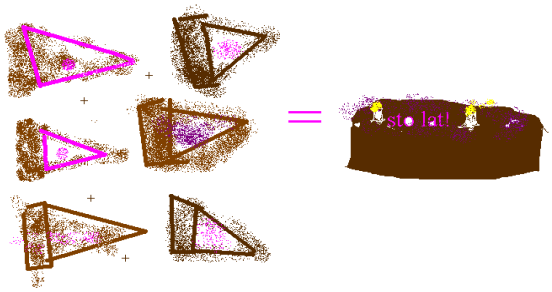
co dla $k = 1$ daje niezły wynik.

Można zadać ogólne pytanie, kiedy układ $y' = f(y)$, gdy $y \in \mathbb{R}^n$ można przybliżyć układem $\dot{x} = Df(0)x$ (gdzie $x = y - x_0$) w okolicy punktu stacjonarnego x_0. W tym miejscu wspomnimy, że punkt $x_0 \in \mathbb{R}^n$ nazywa się *stacjonarnym* układu

$$y' = f(y),$$

jeśli $f(x_0) = 0$.

Okazuje się, że odpowiedź twierdzącą na nasze pytania można uzyskać, gdy macierz $Df(0)$ jest „porządna" (objaśnimy dużo później w rozdziale 8). Teraz nadmienimy, że dla $f : \mathbb{R} \to \mathbb{R}$, oznacza to, że $f'(0) \neq 0$. Czytelnika zainteresowanego szczegółami odsyłamy do książek poświęconych teorii równań różniczkowych zwyczajnych.


sto lat!

Rozdział 6

Całki Iterowane i Wielokrotne

W rozdziale trzecim przedstawiliśmy teorię całki Riemanna funkcji jednej zmiennej. Jej natychmiastowym wielowymiarowym uogólnieniem jest całka iterowana. Przyjmiemy ją za definicję całki wielokrotnej na prostokącie (prostopadłościanie). Lecz próba określenia całki na bardziej złożonych zbiorach natychmiast prowadzi do trudności: musimy ustalić jaki funkcje można całkować, po to aby wyjaśnić na jakiego rodzaju zbiorach całkowanie jest możliwe.

Zaczniemy nasze rozważania od całek iterowanych.

6.1 Całka Iterowana

Zaczniemy od pomocniczej definicji. Niech $Q_k \subset \mathbb{R}^k$ oznacza *prostokąt*, (gdy $k = 2$), *prostopadłościan* (gdy $k = 3$) lub *uogólniony prostopadłościan*, (gdy $k > 3$):

$$Q_k = [a_1, b_1] \times [a_2, b_2] \times ... \times [a_k, b_k]$$

gdzie $a_i < b_i$. Wprowadzana nowa notacja dopuszcza przypadek wielowymiarowy, ale najważniejszy dla nas to ten, gdy $k = 2$ lub $k = 3$.

Jeśli $f : Q_k \to \mathbb{R}$ jest funkcją ciągłą, to liczbę

$$I(f, Q_k) := \int_{a_k}^{b_k} (\int_{a_{k-1}}^{b_{k-1}} ...(\int_{a_1}^{b_1} f(x_1, x_2, ..., x_k)dx_1)dx_2, ...)dx_k$$

nazywamy *całką iterowaną* funkcji f na prostopadłościanie Q_k. Jeśli nie będzie prowadzić to do niejasności, to będziemy pomijali Q_k w oznaczeniu.

Pojawiają się jednak natychmiast pytania:

1. Czy $I(f, Q_k)$ zależy od kolejności całkowania?

2. Dla jakiej szerszej klasy funkcji liczba $I(f, Q_k)$ jest dobrze określona?

3. Czy możliwe są uogólnienia, dopuszczające innego rodzaju zbiory aniżeli prostopadłościany?

W §6.2 uzasadnimy, że liczba $I(f, Q_k)$ oznacza miarę *podwykresu* nieujemnej funkcji f, tj. zbioru

$$E(f) = \{(x_1, \ldots, x_k, x_{k+1}) \in Q_k \times \mathbb{R} : 0 \leq x_{k+1} \leq f(x_1, ..., x_k)\}.$$

Przykład 1. Przyjmując geometryczną interpretację $I(f, Q_2)$ i $Q_2 = [0,1]^2$ obliczmy dla wprawy objętość ostrosłupa o podstawie kwadratu, mającego wierzchołki w $(\pm 1, 0), (0, \pm 1)$ i o wysokości 1. Dla wygody rachunkowej zajmiemy się jego częścią leżącą w pierwszej ćwiartce układu $X0Y$. Zauważmy, że ściana boczna jest wykresem funkcji

$$g(x,y) = \begin{cases} 1 - (|x| + |y|), & \text{gdy } 0 \leq x, y \text{ oraz } |x| + |y| \leq 1 \\ 0, & \text{gdy } 0 \leq x, y < 1 \text{ oraz } |x| + |y| > 1. \end{cases}$$

Mamy wtedy

$$\begin{aligned} I(g, Q_2) &= \int_0^1 (\int_0^1 g(x,y)dy)dx = \int_0^1 \int_0^{1-x} (1 - x - y)dydx \\ &= \int_0^1 ((1-x)^2 - \frac{y^2}{2}\Big|_0^{1-x})dx = \frac{1}{2}\int_0^1 (1-x)^2 = \frac{1}{6}. \end{aligned}$$

Odpowiemy teraz na pierwsze z zadanych pytań.

Twierdzenie 1. Jeśli $f : Q_k \to \mathbb{R}$ jest ciągła, to $I(f, Q_k)$ nie zależy od kolejności całkowania.

Dowód. Rozpatrzymy f szczególnej postaci,

$$f(x) = g_1(x_1) \cdot \ldots \cdot g_k(x_k). \tag{1}$$

Wtedy

$$\begin{aligned} I(f, Q_k) &= \int_{a_k}^{b_k} (\int_{a_{k-1}}^{b_{k-1}} ... \int_{a_1}^{b_1} g_1(x_1) \cdot g_2(x_2) \cdot ... \cdot g_k(x_k)dx_1)...dx_k \\ &= \left(\int_{a_k}^{b_k} g_k(x_k)dx_k\right) ... \left(\int_{a_2}^{b_2} g_2(x_2)dx_2\right)\left(\int_{a_1}^{b_1} g_1(x_1)dx_1\right) \\ &= \int_{a_{\tau(k)}}^{b_{\tau(k)}} (\int_{a_{\tau(k-1)}}^{b_{\tau(k-1)}} (\int_{a_{\tau(1)}}^{b_{\tau(1)}} g_i(x_1)...g_k(x_k)dx_{\tau(1)})...)dx_{\tau(k)} \end{aligned}$$

gdzie τ jest dowolną permutacją zbioru $\{1, .., k\}$. Tym samym nasze twierdzenie jest prawdziwe dla funkcji postaci (1). Zauważmy, że nasze twierdzenie jest prawdziwe dla sum funkcji postaci (1). W dalszym ciągu potrzebny nam będzie fakt, który pozostawimy bez dowodu.

Twierdzenie 2. Dowolną funkcję ciągłą $f : Q_k \to \mathbb{R}$ można przybliżać funkcjami postaci

$$h = \sum_{i=1}^{n} f_i \tag{2}$$

gdzie f_i są postaci (1), tj. dla dowolnej funkcji ciągłej f i dowolnego $\varepsilon > 0$ istnieje taka funkcja h_ε dana wzorem postaci (2), która spełnia

$$||f - h_\varepsilon||_\infty \leq \varepsilon. \tag{3}$$

Dzięki powyższemu twierdzeniu dostaniemy

$$|I(f,Q_k)-I(h,Q_k)|=|I(f-h,Q_k)|=\left|\int_{a_k}^{b_k}...\int_{a_1}^{b_1}(f-h)(x)dx_1\dots dx_k\right|.$$

Z własności całki Riemanna (twierdzenie 3.48) mamy,

$$|I(f,Q_k)-I(h,Q_k)|\leq\int_{a_k}^{b_k}\dots\int_{a_1}^{b_1}|(f-h)(x)|dx_1...dx_k\leq\int_{a_k}^{b_k}\dots\int_{a_1}^{b_1}\varepsilon dx_1\dots dx_k=\varepsilon\text{vol}\,(Q_k) \tag{4}$$

gdzie $\text{vol}\,Q_k=(b_1-a_1)\cdot\ldots\cdot(b_k-a_k)$, (patrz też 2.5.4).

Skoro $I(h,Q_k)$ nie zależy od kolejności całkowania, to i $I(f,Q_k)$ nie zależy od kolejności całkowania. Aby się o tym przekonać położymy

$$I_1(f,Q_k)=\int_{a_{\tau(k)}}^{b_{\tau(k)}}\dots\int_{a_{\tau(1)}}^{b_{\tau(1)}}f(x_1,\dots,x_k)dx_{\tau(1)}\dots dx_{\tau(k)}$$

Wtedy dla h takiego, jak w (2), spełniającego (3)

$$\begin{aligned}|I_1(f,Q_k)-I(f,Q_k)| &= |I_1(f,Q_k)-I_1(h,Q_k)+I(h,Q_k)-I(f,Q_k)|\\ &\leq |I_1(f-h,Q_k)|+|I(f-h,Q_k)|.\end{aligned}$$

Na mocy nierónności (4), która jest niezależna od kolejności całkowania, dostaniemy

$$|I_1(f,Q_k)-I(f,Q_k)|\leq 2\varepsilon\text{vol}\,(Q_k)$$

dla dowolnego ε, stąd $I_1(f,Q_k)=I(f,Q_k)$. □

Wróćmy do przykładu 1. Zauważmy, że w istocie rzeczy policzyliśmy

$$\int_\Delta g,\quad \text{gdzie } \Delta=\{(x,y)\in\mathbb{R}^2:\ x,y\geq 0,\quad x+y\leq 1\},$$

bo funkcja g była równa 0 poza Δ.

Za pomocą całki iterowanej możemy określić $\int_G f(x)\,dx$ dla funkcji ciągłej f wzorem

$$\int_G f(x)\,dx=I(f,Q_k)\equiv\int_{a_k}^{b_k}\dots\int_{a_1}^{b_1}f(x_1,\dots,x_k)dx_1\dots dx_k,$$

jeśli tylko f na brzegu G jest równa zero, bo wtedy f można łatwo przedłużyć na dowolny prostopadłościan Q_k zawierający G, a mianowicie

$$\tilde{f}(x)=\begin{cases}f(x), & \text{gdy } x\in G\\ 0, & \text{gdy } x\in Q_k\setminus G\end{cases} \tag{5}$$

Wtedy $\tilde{f}:Q_k\to$ jest ciągła i $I(\tilde{f},Q_k)$ jest dobrze określona.

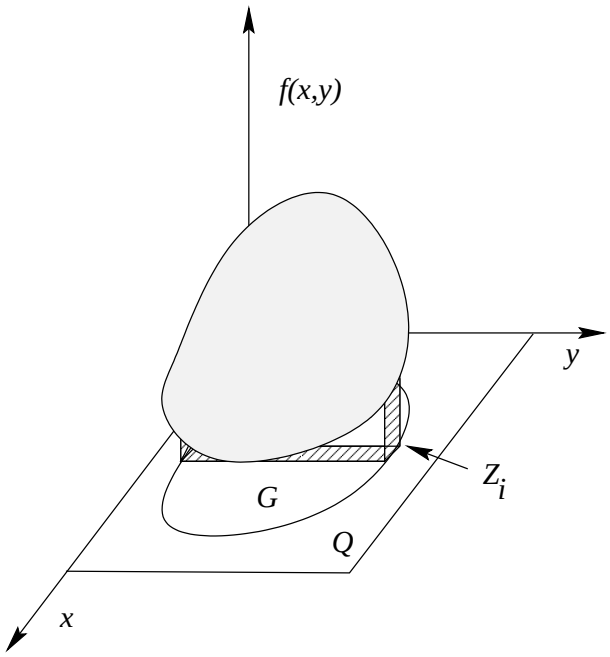
Jednak w ogólności funkcja $\tilde{f}$ dana wzorem (5) nie jest ciągła w Q_k. Można ten problem próbować obchodzić dla zbiorów G specjalnej postaci, np. gdy

$$G=\{(x_1,x_2):a\leq x_1\leq b\quad\text{i}\quad\varphi(x_1)\leq x_2\leq\psi(x_1)\},$$

to wtedy możemy położyć

$$\int_G f(x_1, x_2)\, dx_1 dx_2 = \int_a^b \left(\int_{\varphi(x_1)}^{\psi(x_1)} f(x_1, x_2) dx_2 \right) dx_1$$

Co robić w ogólności? W odpowiedzi uważniej przyjrzymy się całce iterowanej. Dla prostoty przyjmiemy $k = 2$. Niech $G \subset \mathbb{R}^2$ będzie (dość) dowolnym zbiorem ograniczonym i $f : G \to \mathbb{R}$ niech będzie funkcją ciągłą. Dzięki temu, że zbiór G jest ograniczony, istnieje taki prostokąt $Q = [a_1, b_1] \times [a_2, b_2]$, że $G \subset Q$. Określamy $\tilde{f} : Q \to \mathbb{R}$ wzorem (5). Chcemy obliczyć objętość zbioru $Z = E(\tilde{f}) \subset \mathbb{R}^3$. W tym celu wprowadzamy podział odcinka $[a_1, b_1]$, $a_1 = x_0 < x_1 < ... < x_n = b_1$ i kładziemy $\Delta x_i = x_{i+1} - x_i$. Dzielimy Z na cienkie plasterki Z_i o grubości Δx_i (jak na rysunku 1).



Rys. 1. Plasterek o grubości Δx_i.

Ich przybliżona objętość, to wysokość razy pole powierzchni. Wysokość równa się Δx_i, pole powierzchni, to

$$\int_{a_2}^{b_2} \tilde{f}(\xi_i, y) dy \tag{6}$$

gdzie $\xi_i \in [x_i, x_{i+1}]$. Zauważmy też, że w (6) całkujemy koniecznie funkcję nieciągłą, nawet jeśli f była funkcją ciągłą. Tak więc całkowanie funkcji nieciągłych okazało się zwykłą życiową koniecznością! Oznaczmy objętość plasterka Z_i przez $\mu(Z_i)$ mamy więc

$$\mu(Z_i) = \Delta x_i \int_{a_2}^{b_2} \tilde{f}(\xi_i, y) dy.$$

Zauważmy, że prawa strona ma postać

$$\Delta x_i \cdot \varphi(\xi_i),$$

gdzie

$$\varphi(\xi_i) = \int_{a_2}^{b_2} \tilde{f}(\xi_i, y)dy$$

i $\xi_i \in [x_i, x_{i+1}]$. Zatem suma

$$\sum_{i=1}^{n} \mu(Z_i) = \sum_{i=1}^{n} \Delta x_i \int_{a_2}^{b_2} \tilde{f}(\xi_i, y)dy = \sum_{i=1}^{n} \Delta x_i \varphi(\xi_i)$$

jest sumą Riemanowską. Aż prosi się, aby napisać, że objętość podwykresu Z, to

$$\mu(Z) = \int_{a_1}^{b_1} \varphi(x)dx \equiv \int_{a_1}^{b_1} (\int_{a_2}^{b_2} \tilde{f}(x, y)dy)dx = I(\tilde{f}, Q)$$

Tyle że nie wiemy wiele o ciągłości funkcji φ albo raczej o zbiorze jej punktów nieciągłości.

Wyrażenie $\int_{a_1}^{b_1} \int_{a_2}^{b_2} \tilde{f}(x, y)dydx$ jest iterowaną całką Riemanna funkcji $\tilde{f}$, która jest rozszerzeniem zerem funkcji f. Napiszemy zatem

$$\int_G f(x, y)\, dxdy = I(\tilde{f}, Q),$$

gdzie prostokąt $Q \supset G$ jest dowolny i nazwiemy *całką wielokrotną* funkcji f na zbiorze G.

Zajmiemy się teraz ustaleniem kiedy iterowana całka Riemanna istnieje i nie zależy od kolejności całkowania. Spodziewać się należy, że odpowiedź zależy i od funkcji f, i od zbioru G.

6.2 Miara zbiorów w $\mathbb{R}^n$

Podejrzewamy, że miarę, będącą uogólnieniem długości odcinka, pola powierzchni czy objętości, będziemy przypisywali, być może nie wszystkim, ale tylko „dobrym" zbiorom. Nie jest wszak naszym celem wchodzenie w subtelności ogólnej teorii. O zbiorach, które mają miarę (być może nieskończoną) będziemy mówili, że należą do Dobrej Klasy Zbiorów, tj. DKZ. Poniżej opiszemy postulaty, których wypełnienia oczekujemy od DKZ:

(DKZ1) $\emptyset, \mathbb{R}^n \in DKZ$;

(DKZ2) jeśli $Z_1, Z_2 \in DKZ$, to $Z_1 \cup Z_2, Z_1 \cap Z_2$ i $Z_1 \backslash Z_2$ są w DKZ;

(DKZ3) jeśli Q jest dowolnym prostopadłościanem, to Q i $\overset{o}{Q}$ są elementami DKZ.

Trzeba w tym miejscu przypomnieć, że $\overset{o}{Q}$ oznacza wnętrze zbioru G (patrz §4.1.2. definicja 5).

Oczekujemy, że *miara* spełnia następujące, dość naturalne, warunki:

(M1) dla dowolnego $Z \in DKZ$ mamy $\mu(Z) \geq 0$ i $\mu(\emptyset) = 0$;

(M2) jeśli A i $B \in DKZ$, nadto $A \cap B = \emptyset$, to $\mu(A \cup B) = \mu(A) + \mu(B)$;

(M3) jeśli Q jest uogólnionym prostopadłościanem, to $\mu(Q) = \text{vol}\, Q$.

Wyciągniemy stąd kilka prostych wniosków.

Stwierdzenie 3. Jeśli $A, B \in DKZ$, to

$$\mu(A \cup B) \leq \mu(A) + \mu(B). \tag{7}$$

Dowód. Z założenia wynika, że $A \setminus B$, $B \setminus A$, $B \cap A \in DKZ$. Nadto, $A \cup B = (A \setminus B) \cup (B \setminus A) \cup (B \cap A)$, skąd (M2) daje

$$\mu(A \cup B) = \mu((A \setminus B) \cup B) = \mu(A \setminus B) + \mu(B).$$

Co więcej

$$\mu(A) = \mu((A \setminus B) \cup (B \cap A)) = \mu(A \setminus B) + \mu(B \cap A) \geq \mu(A \setminus B),$$

skąd wypływa teza. □

Powyższe rozumowanie prowadzi do dalszych wniosków.

Wniosek 4. Jeśli $A, B \in DKZ$ i $A \subset B$, to $\mu(A) \leq \mu(B)$. □

Wniosek 5. Jeśli $A \subset B$ i dodatkowo $\mu(A) < \infty$, to $\mu(A - B) = \mu(A) - \mu(B)$. □

Aby uniknąć nieporozumień będziemy pisać μ_n dla oznaczenia miary zbiorów w $\mathbb{R}^n$.

Chcielibyśmy mieć praktyczny sposób obliczania $\mu_n(Z)$, gdy $Z \in DKZ$ jest zbiorem ograniczonym. Niech Q jest dowolnym prostokątem zawierającym Z, zaś χ_Z jest funkcją charakterystyczną zbioru Z. Wtedy zgodnie z §6.2 mamy,

$$\mu_n(Z) = \mu_n(Z) \cdot 1 = \mu_{n+1}(E(\chi_Z)) = I(\chi_Z, Q) = \int_Q \chi_Z(x)dx.$$

W tym momencie jest jasnym, że nie unikniemy dokładnej charakteryzacji warunków istnienia całek iterowanych.

Zaczniemy od definicji. Dla $x_0 \in \mathbb{R}^n$ i $r > 0$ zbiór

$$Q(x_0, r) = \{x \in \mathbb{R}^n; |x_i - x_{0i}| < \frac{r}{2},\ i = 1, \ldots, n\}$$

nazwiemy *kostką otwartą o środku w punkcie* x_0 *i krawędzi* r. *Wierzchołkami* kostki nazwiemy punkty spełniające $|x_i - x_{0i}| = \frac{r}{2}$, $i = 1, \ldots, n$.

Definicja 1. Powiemy, że zbiór $E \subset \mathbb{R}^n$ ma *miarę zero*, jeśli dla dowolnego $\varepsilon > 0$ istnieje taka rodzina kostek $\{Q(x_i, r_i)\}_{i=1}^{\infty}$, że $E \subset \cup_{i=1}^{\infty} Q(x_i, r_i)$ i

$$\sum_{i=1}^{\infty} \text{vol}\,(Q(x_i, r_i)) = \sum_{i=1}^{\infty} r_i^n < \varepsilon.$$

Przykład 2. Jeśli $E = [a, b] \times \{c\} \subset \mathbb{R}^2$, to wykażemy, że $\mu_2(E) = 0$. Niech $a = x_0 < x_1 < \ldots < x_n = b$ będzie rozbiciem przedziału $[a, b]$ zadanym wzorami $x_i = a + \frac{b-a}{n} i$. Wtedy kostki $Q(x_i, \frac{2}{n})$, $i = 0, \ldots, n$ pokrywają E, $\text{vol}\, Q(x_i, \frac{2}{n}) = \frac{4}{n^2}$, nadto

$$\sum_{i=0}^{n} \text{vol}\, Q(x_i, \frac{2}{n}) = \frac{4(n+1)}{n^2}$$

zbiega do 0, gdy $n \to \infty$. Innymi słowy dostaliśmy:

Wniosek 6. Brzeg kwadratu ma miarę zero.

Dowód. Niech ∂Q oznacza brzeg kwadratu, jest on sumą czterech odcinków b_1, b_2, b_3, b_4. Zatem z wniosku 4 wynika, że $\mu_2(\partial Q) \leq \mu_2(b_1)+\mu_2(b_2)+\mu_2(b_3)+\mu_2(b_4)$. Zaś z poprzedniego przykładu wynika, że $\mu_2(b_i) = 0$, $i = 1, 2, 3, 4$. Stąd wynika teza. □

Przykład 3. Niech $E = \{(x, y) \in \mathbb{R}^2;\ \ x^2 + y^2 = 1\}$. Wykażemy, że $\mu_2(E) = 0$. Kładziemy

$$(x_i^n, y_i^n) = (\cos\frac{2\pi i}{n}, \sin\frac{2\pi i}{n}) \quad i = 0, 1, ..., n-1, \quad r_n = 2 \cdot \frac{4\pi}{n}.$$

Z twierdzenia cosinusów i ze wzoru Taylora nietrudno sprawdzić, że dla dostatecznie dużych n, mamy że $\bigcup_{i=0}^{n-1} Q((x_i^n, y_i^n), r_n) \supset E$. Dalej $\operatorname{vol} Q((x_i^n, y_i^n), r_n) = \frac{64\pi^2}{n^2}$, zatem

$$\sum_{i=0}^{n-1} \operatorname{vol} Q((x_i, y_i), r_i) = \frac{64\pi^2}{n} \to 0,$$

gdy $n \to \infty$. Szczegółowe rachunki pozostawiamy czytelnikowi.

Przykład 4. Niech $E = [0, 1]^2 \times \{0\} \subset \mathbb{R}^3$, wtedy $\mu_3(E) = 0$. Połóżmy

$$z_{ij} = (\frac{i}{n}, \frac{j}{n}, 0) \quad j, i = 0, 1, ..., n.$$

Wtedy

$$\sum_{ij=0}^{n} \operatorname{vol} Q(z_{ij}, \frac{2}{n}) = (n+1)^2 \frac{8}{n^3} \to 0, \qquad \text{gdy } n \to \infty.$$

Możemy teraz wysłowić warunki istnienia całki iterowanej.

Twierdzenie 7. Niech $G \subset \mathbb{R}^n$ będzie zbiorem ograniczonym, $f : G \to \mathbb{R}$ i Q jest dowolnym prostopadłościanem zawierającym G. Kładziemy

$$\tilde{f}(x) = \begin{cases} f(x), & \text{gdy } x \in G \\ 0, & \text{gdy } x \in Q \setminus G. \end{cases}$$

Iterowana całka Riemanna $I(\tilde{f}, Q)$ istnieje i nie zależy od kolejności całkowania wtedy i tylko wtedy, gdy funkcja $\tilde{f}$ jest ograniczona i zbiór

$$N = \{x \in Q :\ \tilde{f} \text{ nie jest ciągła w punkcie } x\} \tag{8}$$

ma miarę 0. (Bez dowodu).

Od tego momentu, jeśli f i G spełniają założenia 7, to będziemy pisać $\int_G f(x)dx$ zamiast $I(\tilde{f}, Q)$ i będziemy ją nazywać *całką wielokrotną*.

Zastosujmy przedstawioną wyżej charakteryzację do miary zbioru $Z \subset Q$,

$$\mu_n(Z) = \int_Q \chi_Z(x)\,dx \tag{9}$$

Wynika z niego, że funkcja χ_Z musi mieć zbiór nieciągłości N, spełniający $\mu(N) = 0$. Zauważmy, że w przypadku funkcji charakterystycznej zbioru Z mamy $N = \partial Z$. Uzasadnia to następujące określenie.

Definicja 2. Powiemy, że zbiór $G \subset \mathbb{R}^n$ jest *mierzalny w sensie Jordana-Riemanna*, jeśli ∂G ma miarę zero.

Uwaga. Od tej chwili uznajemy, że zbiory mierzalne w sensie Jordana-Riemanna należą do DKZ. Aczkolwiek, być może nie wyczerpują one tej klasy zbiorów. Nie będziemy zgłębiać tego tematu.

Widać zatem, że możemy całkować tylko po zbiorach mierzalnych w sensie Jordana-Riemanna. Chcielibyśmy mieć warunek istnienia $\int_G f(x)\,dx$ nie w terminach $\tilde{f}$, ale samej funkcji f. Jeśli zbiór N jest określony tak, jak w (8), to od razu widać, że

$$N = N_0 \cup F$$

gdzie

$$N_0 = \{x \in G :\ f \text{ nie jest ciągła w punkcie } x\}$$

i $F \subset \partial G$. Zakładając mierzalność G dostaniemy, że $\mu(F) \leq \mu(\partial G) = 0$. Zatem z wniosku 4

$$0 = \mu(N) = \mu(N_0 \cup F) \geq \mu(N_0) = 0.$$

Zatem z powyższego twierdzenia dostajemy następujący wniosek.

Wniosek 8. Załóżmy, że $G \subset \mathbb{R}^n$ jest ograniczony i mierzalny w sensie Jordana-Riemanna. Wtedy, istnienie

$$\int_G f(x)\,dx$$

jest równoważne temu, że zbiór N (taki jak w (8)) ma miarę 0 i funkcja f jest ograniczona.

6.3 Właściwości całek i miary

Sformułujemy teraz szereg właściwości całek będących uogólnieniami znanych cech całek Riemanna na przedziałach. Zakładamy poniżej, że zbiór $G \subset \mathbb{R}^n$ jest ograniczony i mierzalny w sensie Jordana-Riemanna a funkcje $f, g : G \to \mathbb{R}$ są *całkowalne* tj. całki $\int_G f, \int_G g$ są dobrze określone.

Stwierdzenie 9. Niech $\alpha, \beta \in \mathbb{R}$, wtedy

$$\int_G (\alpha f(x) + \beta g(x))\,dx = \alpha \int_G f(x)\,dx + \beta \int_G g(x)\,dx.$$

Dowód. Niech Q będzie taką kostką, że $Q \supset G$ i $\tilde{f}, \tilde{g}$ są dane wzorem (5), wtedy

$$L = \int_Q (\alpha\tilde{f}(x) + \beta\tilde{g}(x))dx = \int_{a_n}^{b_n} ... \int_{a_1}^{b_1} (\alpha\tilde{f}(x) + \beta\tilde{g}(x))dx_1...dx_n = (J).$$

Dzięki właściwościom całek Riemanna na przedziałach mamy

$$\begin{aligned}(J) &= \int_{a_n}^{b_n} ... (\alpha \int_{a_1}^{b_1} \tilde{f}(x)dx_1 + \beta \int_{a_1}^{b_1} \tilde{g}(x)dx_1)dx_2 \ ... \ dx_n = \\ &= \alpha \int_{a_n}^{b_n} ... \int_{a_1}^{b_1} \tilde{f}(x)dx_1 \dots dx_n + \beta \int_{a_n}^{b_n} ... \int_{a_1}^{b_1} \tilde{g}(x)dx_1 \dots dx_n = \\ &= \alpha \int_G f(x)\,dx + \beta \int_G g(x)\,dx = P\end{aligned}$$

□

Stwierdzenie 10. Niech $G = G_1 \cup G_2$, gdzie $G_1 \cap G_2 = \emptyset$ i G_1, G_2 są mierzalne. Jeśli $f : G \to \mathbb{R}$ jest całkowalna, to

$$\int_G f(x)\,dx = \int_{G_1} f(x)\,dx + \int_{G_2} f(x)\,dx.$$

Dowód. Mierzalność zbiorów G_1 i G_2 zapewnia, że funkcje $g_i = f\chi_{G_i}$, $i = 1, 2$ są całkowalne. Dzięki stwierdzeniu 9 mamy

$$\int_G f(x)\,dx = \int_G (g_1 + g_2)(x)\,dx = \int_G g_1(x)\,dx + \int_G g_2(x)\,dx = \int_{G_1} f(x)\,dx + \int_{G_2} f(x)\,dx. \ \square$$

Stwierdzenie 11. Jeśli funkcje $f, g : G \to \mathbb{R}$ są całkowalne i $f(x) \leq g(x)$ dla wszystkich $x \in G$, to wtedy

$$\int_G f(x)dx \leq \int_G g(x)dx.$$

Dowód. Wynika on z analogicznej właściwości dla całek po przedziałach zastosowanej do schematu dowodu stwierdzenia 9. Mianowicie, z założenia wynika $\tilde{f} \leq \tilde{g}$, a następnie

$$\int_G f(x)dx = I(\tilde{f}, Q) \leq I(\tilde{g}, Q) = \int_G g(x)dx. \qquad \square$$

Stwierdzenie 12. Niech $m \leq f(x) \leq M$ dla wszystkich $x \in G$. Wtedy

$$m\mu(G) \leq \int_G f(x)dx \leq M\mu(G).$$

Dowód. Wynika on bezpośrednio ze stwierdzenia 11 i z faktu, że $\int_G f(x)\,dx = \mu(G)$. □

Stwierdzenie 13. Jeśli $\mu(G) = 0$ i funkcja f jest ograniczona, to $\int_G f(x)\,dx = 0$.

Dowód. Skoro $-M \leq f(x) \leq M$, to ze stwierdzenia 12 mamy $-\mu(G)M \leq \int_G f(x)\,dx \leq M\mu(G)$. Stąd i z założenia $\mu(G) = 0$, wypływa teza. □

Twierdzenie 14. (o wartości średniej). Załóżmy, że f jest całkowalna, nadto $f(x) \in [m, M]$ dla pewnych $m, M \in \mathbb{R}$ i wszystkich $x \in G$. Wtedy istnieje takie $\xi \in [m, M]$, że

$$\int_G f(x)\,dx = \xi\,\mu(G).$$

Dowód. Ze stwierdzenia 12 mamy, że $\int_G f(x)\,dx \in [m\mu(G), M\mu(G)]$. Teraz istnienie żądanej liczby jest oczywiste. □

Stwierdzenie 15. Jeśli $f, g : G \to \mathbb{R}$ są całkowalne i zbiór $Z = \{x \in G : f(x) \neq g(x)\}$ jest mierzalny i ma miarę zero, to wtedy $\int_G f(x)\,dx = \int_G g(x)\,dx$.

Dowód. Kładziemy $h(x) = f(x) - g(x)$. Ze stwierdzenia 10 mamy,

$$\int_G h(x)dx = \int_{G\setminus Z} h(x)dx + \int_Z h(x)dx = 0 + \int_Z h(x)dx.$$

Ze stwierdzenia 13 dostaniemy $\int_Z h(x)dx = 0$, zatem w myśl stwierdzenia 9

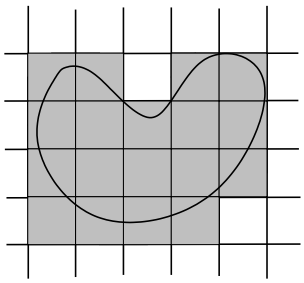
$$0 = \int_G h(x)dx = \int_G (f(x) - g(x))dx = \int_G f(x)dx - \int_G g(x)dx,$$

stąd wypływa teza. □

6.4 Interpretacja geometryczna mierzalności Jordana-Riemanna

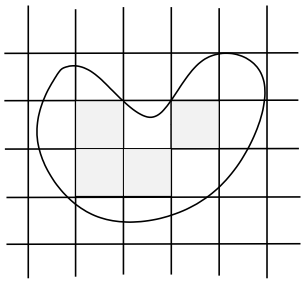
Wprowadzimy w $\mathbb{R}^d$ siatki zbudowane z punktów o wszystkich współrzędnych będących wielokrotnościami $\delta > 0$, tj. z punktów, które są wierzchołkami kostek o boku δ,

$$S(\delta) = \{x \in \mathbb{R}^d : x = \delta\zeta, \quad \zeta \in \mathbb{Z}^d\}.$$



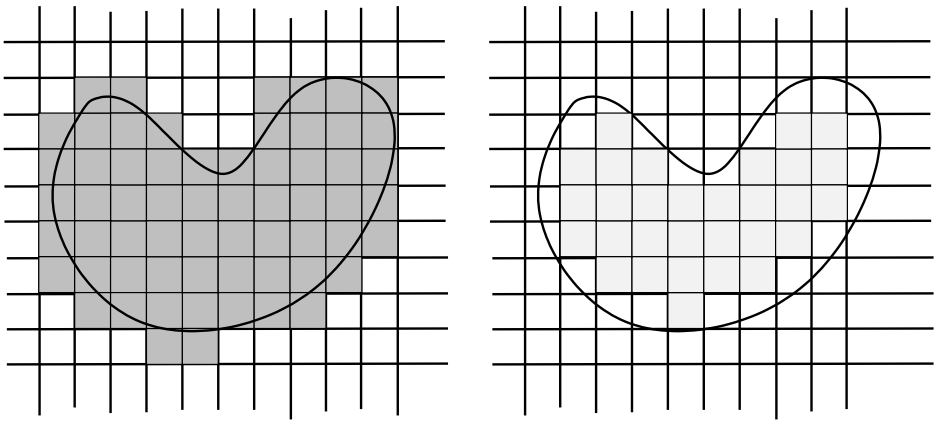
Rys. 2. Zbiór $G^*(\delta)$

Badamy zbiór $G \subset \mathbb{R}^d$. Tworzymy zbiory:
$G^*(\delta)$ zawierający G a będący sumą kostek o boku δ i o wierzchołkach z siatki $S(\delta)$;
$G_*(\delta)$ będący sumą kostek o boku δ i o wierzchołkach z siatki $S(\delta)$ zawartych w G.



Rys. 3. Zbiór $G_*(\delta)$

Na rysunku 2. w $G^*(\delta)$ jest 18 kostek; rysunek 3. pokazuje w $G_*(\delta)$ 4 kostki, w $G^*(\frac{\delta}{2})$ jest ich 62, zaś w $G_*(\frac{\delta}{2})$ jest ich 30.



Rys. 4. Zbiory $G^*(\frac{\delta}{2})$ oraz $G_*(\frac{\delta}{2})$

Jeśli mamy

$$G^*(\delta) = \bigcup_{i=1}^{N^*} Q(x_i, \delta), \quad G_*(\delta) = \bigcup_{i=1}^{N_*} Q(x_i, \delta),$$

to

$$\mu_d(G^*(\delta)) = N^* \text{vol}\, Q(x_i, \delta) = N^* \delta^d \geq N_* = \delta^d \mu_d(G_*(\delta)).$$

Jest dość oczywistym, że funkcja $\delta \to \mu_d(G^*(\delta))$ jest malejąca a funkcja $\delta \to \mu_d(G_*(\delta))$ jest rosnąca. Jest zatem zawsze jest prawdą, że

$$\sup_{\delta>0} \mu(G_*(\delta)) \leq \inf_{\delta>0} \mu(G^*(\delta)).$$

Natomiast równość nie zawsze zachodzi. Mamy:

Twierdzenie 16. Zbiór ograniczony G jest mierzalny w sensie Jordana-Riemana wtedy i tylko wtedy, gdy

$$\sup_{\delta>0} \mu(G_*(\delta)) = \inf_{\delta>0} \mu(G^*(\delta)).$$

Pozostawimy ten fakt bez dowodu.

Takie pojęcie mierzalności jak w/g Jordana-Riemanna jest poglądowe, lecz ma wady, np. zbiór $G = \mathbb{Q}^2 \cap [0,1]^2$, tj. zbiór punktów z kwadratu o obu współrzędnych wymiernych nie jest mierzalny, bo dla dowolnego $\delta > 0$

$$\mu_2(G^*(\delta)) \geq 1, \quad \text{zaś} \quad \mu_2(G_*(\delta)) = 0.$$

Podana wyżej charakteryzacja zbiorów mierzalnych ma następujące zastosowania do teorii całkowania. Można się przekonać, że można przybliżać zbiór, po którym całkujemy zbiorami prostszymi. Mamy mianowicie.

Twierdzenie 17. Niech $G \subset \mathbb{R}^d$ będzie mierzalny w/g Jordana-Riemanna a funkcja $f : G \to \mathbb{R}$ całkowalna. Załóżmy, że $\{P_n\}_{n=1}^{\infty}$ jest takim ciągiem rodzin kostek

$$P_n = \{Q(x_i^n, \delta_n)\}_{i=1}^{k_n},$$

że

$$\lim_{n\to\infty} \delta_n = 0, \quad \text{i} \quad \bigcup_{i=1}^{k_n} Q(x_i^n, \delta_n) \subset G \quad \text{oraz } \lim_{n\to\infty} k_n \delta^d = \mu_d(G).$$

Wtedy

$$\int_G f(x)dx = \lim_{n\to\infty} \sum_{i=1}^{k_n} \int_{Q(x_i^n,\delta_n)} f(x)dx$$

Dowód. Załóżmy, że $\max_G |f| \leq M$. Weźmy dowolne $\varepsilon > 0$. Zbadajmy różnicę

$$J_n = \int_G f(x)dx - \sum_{i=1}^{k_n} \int_{Q(x_i^n,\delta_n)} f(x)dx = \int_{G\setminus\bigcup_{i=1}^{k_n} Q(x_i^n,\delta_n)} f(x)dx$$

Możemy dzięki mierzalności G tak dobrać δ_n, aby

$$\mu_d(G\setminus \bigcup_{i=1}^{k_n} Q(x_i^n, \delta_n)) \leq \frac{\varepsilon}{M}.$$

Wtedy $|J_n| \leq M \cdot \frac{\varepsilon}{M} = \varepsilon$. □

Najwygodniej by było, gdyby można było także przybliżać funkcję podcałkową funkcjami, które są stałe na kostkach $Q(x_i^n, r_n)$. Istotnie, prawdziwy jest następujący fakt, który pozostawimy bez dowodu.

Twierdzenie 18. Jeśli G jest mierzalny, $\{P_n\}_{n=1}^{\infty}$ jest ciągiem kostek takim jak wyżej i punkty $\xi_i^n \in Q(x_i^n, \delta_n)$ są dowolne, to wtedy dla całkowalnej funkcji $f : G \to \mathbb{R}$ mamy

$$\int_G f(x)dx = \lim_{n\to\infty} \sum_{i=1}^{k_n} \text{vol}\, Q(x_i^n, \delta_n) f(\xi_i^n).$$

Ten fakt leży u podstaw teorii całki rozwijanej w niniejszym opracowaniu: każda całka jest sumą nieskończenie wielu przyczynków, tj. jest granicą sum Riemannowskich. Są one postaci: stała razy miara kwadratu (lub równoległoboku, która jest obrazem kwadratu).

6.5 Miara zbiorów nieograniczonych i całki niewłaściwe

Chcemy odpowiedzieć na pytanie: czy zbiór nieograniczony może mieć dobrze określone pole (objętość itd)? Odpowiedź jest twierdząca, aczkolwiek szczegółowo zbadamy wyłącznie przypadek zbiorów nieograniczonych, które są podwykresami funkcji. Do wyłożenia myśli przewodniej przyda się nowy język: powiemy, że rodzina prostopadłościanów $\{Q_k\}_{k=1}^{\infty}$ *wyczerpuje* $\mathbb{R}^n$, jeśli dla każdego k mamy $Q_{k+1} \supset Q_k$ i $\bigcup_{i=1}^{\infty} Q_i = \mathbb{R}^n$.

Załóżmy teraz, że $E \subset \mathbb{R}^n$ jest nieograniczony, ale taki, że

> dla dowolnej rodziny prostopadłościanów $\{Q_k\}_{k=1}^{\infty}$ wyczerpującej $\mathbb{R}^n$, przecięcie $Q_k \cap E$ jest mierzalne w sensie Jordana-Riemanna. (O)

Wtedy kładziemy

$$\mu_n(E) = \lim_{k\to\infty} \mu_n(E \cap Q_k).$$

Musimy zastanowić się, czy jest to dobrze określona wielkość, tj. czy zależy od wyboru ciągu $\{Q_k\}_{k=1}^{\infty}$. Dopuszczamy, aby $\mu_n(E) = \infty$.

Stwierdzenie 19. Przy dotychczasowych założeniach na E liczba $\mu_n(E)$ jest dobrze określona.

Dowód. Zajmiemy się tylko przypadkiem $\mu_n(E) < \infty$, gdy $\mu_n(E) = \infty$ rozumowanie wymaga jedynie kosmetycznych zmian, których dokonanie pozostawiamy Czytelnikowi. Niech będą dane 2 ciągi $\{Q^1{}_k\}_{k=1}^{\infty}$ i $\{Q^2{}_k\}_{k=1}^{\infty}$ spełniające (O). Zauważmy, że ciągi liczbowe $\{\mu_n(E \cap Q_k^1)\}_{k=1}^{\infty}$, $\{\mu_n(E \cap Q_k^2)\}_{k=1}^{\infty}$ są rosnące, a więc zbieżne. Trzeba pokazać, że granice

$$\lim_{k\to\infty} \mu_n(E \cap Q_k^1) =: \mu^1, \qquad \lim_{k\to\infty} \mu_n(E \cap Q_k^2) =: \mu^2$$

są równe. Wykażemy, że

$$\mu^1 \le \mu^2. \tag{10}$$

Z definicji granicy wynika, że dla dowolnego $\varepsilon > 0$ istnieje takie N_ε, że

$$\mu^1 - \varepsilon \le \mu_n(E \cap Q_k^1)$$

dla $k \ge N_\varepsilon$. Nadto, $\mu_n(E \cap Q_k^1) \le \mu^1$, co wynika z monotoniczności ciągu. Ponieważ oba ciągi prostopadłościanów $\{Q^1{}_k\}_{k=1}^{\infty}$ i $\{Q^2{}_k\}_{k=1}^{\infty}$ wyczerpują $\mathbb{R}^n$, to dla każdego k istnieje takie l_k, że $Q_k^1 \subset Q_{l_k}^2$. Dlatego

$$\mu^1 - \varepsilon \le \mu_n(E \cap Q_k^1) \le \mu_n(E \cap Q_l^2) \le \mu^2,$$

gdy $k \ge N_\varepsilon$. Ponieważ ε jest dowolne, to wnosimy stąd nierówność (10).

Ten sam argument daje nierówność przeciwną $\mu^2 \le \mu^1$, a stąd wynika równość $\mu^2 = \mu^1$. □

Uzyskany wynik posłuży do rozszerzenia pojęcia całki Riemanna na przypadek funkcji lub zbiorów nieograniczonych.

Definicja 3. Niech $D \subset \mathbb{R}^n$ będzie takim zbiorem, że: (a) D jest nieograniczony i spełnia (O); albo (b) D jest ograniczony i mierzalny w sensie Jordana–Riemanna. Załóżmy, że g :

$D \to \mathbb{R}$ jest funkcją nieujemną, której zbiór punktów nieciągłości ma miarę zero. Powiemy, że g jest *całkowalna w niewłaściwym sensie Riemanna*, jeśli $\mu_{n+1}(E(g)) < \infty$, gdzie $E(g)$ jest podwykresem g. Wtedy położymy

$$\int_D g(x)\,dx := \mu_{n+1}(E(g))$$

i nazwiemy *całką w niewłaściwym sensie Riemanna* funkcji g na zbiorze D.

W praktyce mamy do czynienia nie tylko z funkcjami nieujemnymi, dlatego musimy rozszerzyć powyższą definicję, tak aby obejmowała funkcje o zmiennym znaku.

Definicja 4. Załóżmy, że zbiór $D \subset \mathbb{R}^n$ jest taki, jak poprzednio, zaś funkcja $f : D \to \mathbb{R}$ jest ograniczona i jej zbiór nieciągłości ma miarę zero. Powiemy, że f jest *całkowalna w niewłaściwym sensie Riemanna*, jeśli

$$\int_D |f(x)|\,dx < \infty.$$

Jeśli położymy, $f^+(x) = \max\{f(x), 0\}$, $f^-(x) = \max\{0, -f(x)\}$, to *całkę w niewłaściwym sensie Riemanna* funkcji f na D określamy wzorem

$$\int_D f(x)\,dx = \int_D f^+(x)\,dx - \int_D f^-(x)\,dx.$$

Łatwo jest zauważyć, że $f(x) = f^+(x) - f^-(x)$. Nadto, $0 \le f^+(x), f^-(x) \le |f(x)|$. Czyli powyższa definicja jest poprawna.

Skupimy się teraz na objaśnieniu powyższych definicji w dwu szczególnych przypadkach: (i) D jest nieograniczony i funkcje $f, g : D \to \mathbb{R}$ są ograniczone; albo (ii) D jest ograniczony i funkcje $f, g : D \to \mathbb{R}$ są nieograniczona. (Zakładamy, że g jest nieujemna).

Zaczniemy od rozpatrzenia pierwszej sytuacji. Z mocy stwierdzenia 19, jeśli $\{Q_k\}_{k=1}^{\infty}$ jest dowolnym ciągiem prostopadłościanów wyczerpujących $\mathbb{R}^n \times \mathbb{R}$, to

$$\mu_{n+1}(E(g)) = \lim_{k\to\infty} \mu_{n+1}(E(g) \cap Q_k).$$

Zauważmy, że gdy $Q_k = Q'_k \times [a_k, b_k]$ i $f(x) \in [a_k, b_k]$ dla $k \ge 1$, to mamy

$$\mu_{n+1}(E(g) \cap Q_k) = \int_{D\cap Q'_k} g(x)\,dx.$$

A więc

$$\int_D g(x)\,dx \equiv \mu_{n+1}(E(g)) = \lim_{k\to\infty} \mu_{n+1}(E(g) \cap Q_k) = \lim_{k\to\infty} \int_{D\cap Q'_k} g(x)\,dx.$$

Natomiast, jeśli f jest funkcją o zmiennym znaku, to dostaniemy

$$\begin{aligned}
\int_D f(x)\,dx &= \int_D f^+(x)\,dx - \int_D f^-(x)\,dx \\
&= \lim_{k\to\infty} \int_{D\cap Q_k} f^+(x)\,dx - \lim_{k\to\infty} \int_{D\cap Q_k} f^-(x)\,dx \\
&= \lim_{k\to\infty} \int_{D\cap Q_k} (f^+(x) - f^-(x))\,dx = \\
&= \lim_{k\to\infty} \int_{D\cap Q_k} f(x)\,dx.
\end{aligned}$$

Trzeba tu podkreślić, że jeśli $f : D \to \mathbb{R}$ jest dowolną funkcją ograniczoną i zmiennego znaku, której zbiór punktów nieciągłości ma miarę zero, to granica po prawej stronie może w ogóle nie istnieć. Może też się zdarzyć, że granica istnieje, mimo iż funkcja f nie jest całkowalna w niewłaściwym sensie Riemanna. Powiemy, że istnieje wtedy *niewłaściwa całka Riemanna* funkcji f na D. Przykładem takiej sytuacji są całki Fresnela:

Przykład 5.

$$\int_0^\infty \frac{\sin x}{x}\,dx := \lim_{R\to\infty}\int_0^R \frac{\sin x}{x}\,dx = \frac{\pi}{2},$$

ale funkcja $\frac{\sin x}{x}$ nie jest całkowalna w niewłaściwym sensie Riemanna!

Przykład 6. Sprawdzimy teraz kiedy funkcje $f(x) = x^{-\alpha}$ na przedziale $D = [1, \infty)$ są całkowalne w niewłaściwym sensie Riemanna. Zakładamy, że $\alpha > 0$. Weźmy $Q_R = [-R, R]$, wtedy $D \cap Q_R = [0, R]$ i dostaniemy,

$$\begin{cases} \int_1^R x^{-\alpha}\,dx = \dfrac{-1}{\alpha-1}x^{1-\alpha}|_0^R, & \text{gdy } \alpha \neq 1; \\ \int_1^R x^{-1}\,dx = \ln R, & \text{gdy } \alpha = 1. \end{cases}$$

Granica $\lim\limits_{R\to\infty}\int_1^R x^{-\alpha}\,dx$ jest więc skończona wtedy i tylko wtedy, gdy $\alpha > 1$.

W podobny sposób postępujemy z funkcjami nieograniczonymi na zbiorach ograniczonych. Aby rozszyfrować znaczenie wprowadzonego określenia rozpatrzmy ciąg wyczerpujący $Q_k = Q'_k \times [-k, k]$, gdzie $Q = Q'_1$ jest dowolnym prostopadłościanem zawierającym D. Wtedy,

$$\mu_{n+1}(E(g)) = \lim_{k\to\infty}\mu_{n+1}(E(g)\cap Q_k) = \lim_{k\to\infty}\int_D g_k(x)\,dx,$$

gdzie g_k są zdefiniowane następująco

$$g_k(x) = \begin{cases} g(x), & \text{gdy } g(x) \leq k; \\ k, & \text{gdy } g(x) > k. \end{cases}$$

W przypadku funkcji o zmiennym znaku dostaniemy do obliczenia następującą granicę,

$$\int_D f(x)\,dx = \int_D f^+(x)\,dx - \int_D f^-(x)\,dx = \int_D f(x)\,dx = \lim_{k\to\infty}\int_D f_k(x)\,dx, \qquad (11)$$

gdzie

$$f_k(x) = \begin{cases} f(x), & \text{gdy } |f(x)| \leq k; \\ k, & \text{gdy } f(x) > k; \\ -k, & \text{gdy } f(x) < -k. \end{cases}$$

Jednak w praktyce (11) nie jest dogodne w obliczeniach. Lepiej jest mieć do czynienia z granicą

$$\lim_{k\to\infty}\int_D \chi_{|f|\leq k} f(x)\,dx.$$

Trzeba sprawdzić, czy obie granice są równe. Istotnie, mamy

Twierdzenie 20. Przy założeniach na D i f takich jak wyżej, jeśli f jest całkowalna w niewłaściwym sensie Riemanna, to

$$\int_D f(x)\,dx = \lim_{k\to\infty} \int_D \chi_{|f|\le k} f(x)\,dx = \lim_{k\to\infty} \int_D f_k(x)\,dx.$$

Pozostawimy ten fakt bez dowodu. Zajmiemy się zaś prostymi zastosowaniami. Przydatne wnioski z tego faktu ujrzymy w rozdziale 7.

Przykład 7. Zbadajmy całkowalność $\int_0^1 x^{-\alpha}\,dx$, $\alpha \in \mathbb{R}$ posługując się poprzednim twierdzeniem. Liczymy dla $k = 1/\varepsilon$,

$$\begin{cases} \displaystyle\int_\varepsilon^1 x^{-\alpha}\,dx = \frac{1}{\alpha-1}(1-\varepsilon^{1-\alpha}), & \text{gdy } \alpha \neq 1, \\ \displaystyle\int_\varepsilon^1 x^{-1}\,dx = -\ln\varepsilon, & \text{gdy } \alpha = 1. \end{cases}$$

Granica $\lim_{\varepsilon\to 0} \int_\varepsilon^1 x^{-\alpha}\,dx$ jest skończona wtedy i tylko wtedy, gdy $\alpha < 1$.

6.6 Zamiana zmiennych w całce wielokrotnej

Będziemy zajmowali się wyłącznie przypadkiem dwuwymiarowym, choć wyniki są prawdziwe w wielu wymiarach.

6.6.1 Całka na równoległoboku

Niech $R(a,b) \subset \mathbb{R}^2$ będzie równoległobokiem, patrz §2.5.2. Chcemy obliczyć $\int_{R(a,b)} f(x)dx$. Podejrzewamy, że prościej by było liczyć tę całkę na kwadracie $Q = Q((\frac{1}{2},\frac{1}{2}),1)$. Zauważmy, że $R(a,b) = AQ$, gdzie A jest przekształceniem liniowym, takim mianowicie, że $A\binom{1}{0} = a$ i $A\binom{0}{1} = b$. Załóżmy na początek, że funkcja f jest stała i równa c. Wtedy korzystając ze wzoru na pole równoległoboku dostaniemy

$$\int_R c dx = \mu(R)\cdot c = c|\det(a,b)| = c|\det A| = \int_Q c|\det A| dy.$$

Zanim opowiemy o przypadku ogólnym przedstawimy pewien pomocniczy fakt.

Twierdzenie 21. Niech zbiór $G \subset \mathbb{R}^n$ będzie domknięty i ograniczony oraz $f : G \to \mathbb{R}$. Wtedy poniższe warunki są równoważne.

(i) f jest funkcją ciągłą;

(ii) dla każdego $\varepsilon > 0$ istnieje takie $\delta > 0$, że dla wszystkich $x, y \in G$ spełniających $d(x,y) < \delta$ mamy $d(f(x), f(y)) < \varepsilon$.

Jeśli funkcja f spełnia (ii), to mówi się, że jest *jednostajnie ciągła*.

Uwagi. (1) Funkcja $x \to \frac{1}{x}$, $x \in (0,1)$ nie jest jednostajnie ciągła, tj. istotna jest domkniętość D. Funkcja $x \to x^2$ dla $x \in [1, +\infty)$ nie jest jednostajnie ciągła, tj. istotna jest ograniczoność D.

(2) Jeśli funkcja $f : G \to \mathbb{R}$ spełnia warunek Lipschitza ze stałą L, to jest automatycznie jednostajnie ciągła.

Wracamy do całkowania. Niech $Q_{ij} = Q(x_{ij}^n, \frac{1}{n})$, gdzie $x_{ij}^n = (\frac{i+\frac{1}{2}}{n}, \frac{j+\frac{1}{2}}{n})$, $\quad i, j = 0, ..., n-1$

$$\begin{aligned}\int_R f(y)dy &= \int_{AQ} f(y)dy = \sum_{ij=0}^{n-1} \int_{AQ_{ij}} f(y)dy \\ &= \sum_{ij=0}^{n-1} (\int_{AQ_{ij}} f(Ax_{ij}^n)dy + \int_{AQ_{ij}} (f(y) - f(Ax_{ij}^n)dy) = (*)\end{aligned}$$

Niech teraz ε będzie dowolną liczbą dodatnią. Dobieramy takie N, aby $|f(y) - f(Ax_{ij}^n)| < \varepsilon$ dla dowolnego $y \in Q_{ij}$ i $n > N$. Można to osiągnąć dzięki twierdzeniu 21. Wtedy

$$(*) = \sum_{i,j=0}^{n-1} f(Ax_{ij}^n) \int_{Q_{ij}} |\det A| dx + \text{ błąd}_1,$$

gdzie pierwszy składnik jest sumą Riemannowską, zaś $|\text{błąd}_1| < \sum_{ij=0}^{n-1} \varepsilon \cdot \text{vol}\, Q_{ij}$. Zatem, jeśli n dąży do nieskończoności, to dzięki twierdzeniu 18 i dowolności ε dostaniemy

$$\int_R f(y)dy = \int_Q f(Ax)|\det A| dx. \tag{12}$$

Jest to *wzór na zamianę zmiennych w całce wielokrotnej.*

Zastosujmy zdobytą wiedzę.

Przykład 8. Obliczmy

$$\int_R (y_1 - y_2)^2 \, dy_1 dy_2,$$

gdzie $R = \{(y_1, y_2) : \ 0 \le y_1 \le 1, \ y_1 \le y_2 \le y_1 + 1\}$. Zauważmy, że $R = AQ$, $Q = [0,1]^2$ oraz

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

tj. $A\binom{x_1}{x_2} = \binom{x_1+x_2}{x_2} = \binom{y_1}{y_2}$, nadto $\det A = 1$. Stosujemy wzór (12),

$$\begin{aligned}\int_R (y_1 - y_2)^2 dy_1 dy_2 &= \int_Q (x_1 + x_2 - x_2)^2 |\det A| dx_1 dx_2 \\ &= \int_Q x_1^2 dx_1 dx_2 = \int_0^1 x_1^2 dx_1 \int_0^1 dx_2 = \frac{1}{3}.\end{aligned}$$

6.6.2 Mierzalność obrazu zbioru

Zanim zajmiemy się ogólnym wzorem na zamianę zmiennych musimy zastanowić się, czy obraz zbioru mierzalnego jest mierzalny. W tym celu wykażemy następujący fakt.

Lemat 22. Załóżmy, że $\phi : Q(x_0, \delta) \subset \mathbb{R}^n \to \mathbb{R}^d$ jest funkcją spełniającą warunek Lipschitza ze stałą L. Wtedy

$$\phi(Q(x_0, \delta)) \subset \bar{B}(\phi(x_0), L\delta)) \subset Q(\phi(x_0), L\delta)$$

Dowód. Druga inkluzja jest oczywistym faktem geometrycznym, pozostaje nam wykazać pierwsze zawieranie. Niech $y \in Q(x_0, \delta)$, wtedy z założenia

$$\|\phi(x_0) - \phi(y)\| \leq L\|x_0 - y\| \leq L\delta.$$

Oznacza to, że $\phi(y) \in \bar{B}(\phi(x_0), L\delta)$. □

Możemy teraz sformułować zasadniczy wynik.

Stwierdzenie 23. Załóżmy, że zbiór $D \subset \mathbb{R}^n$ jest mierzalny w sensie Jordana-Riemanna. Nadto, $\phi : \mathbb{R}^n \to \mathbb{R}^n$ jest różnowartościową funkcją spełniającą warunek Lipschitza ze stałą L. Wtedy $\phi(D)$ jest zbiorem mierzalnym w sensie Jordana-Riemanna.

Dowód. Niech $\varepsilon > 0$ będzie dowolne i $\{Q(x_k, \delta_k)\}_{k=1}^{\infty}$ jest ciągiem kostek pokrywających ∂D takim, że $\sum_{k=1}^{\infty} \text{vol}\,(Q(x_k, \delta_k)) < \varepsilon$. Zauważmy, że wtedy na mocy poprzedniego lematu rodzina $\{Q(\phi(x_k), L\delta_k)\}_{k=1}^{\infty}$ pokrywa $\phi(\partial D)$, co więcej,

$$\sum_{k=1}^{\infty} \text{vol}\,(Q\phi(x_k), L\delta_k) \leq \sum_{k=1}^{\infty} \text{vol}\,(Q(x_k, \delta_k))L^n < \varepsilon L^n.$$

A więc zbiór $\phi(\partial D)$ jest mierzalny w sensie Jordana-Riemanna. Trzeba nam jeszcze wiedzieć, że $\phi(\partial D) = \partial\phi(D)$. Jest to fakt ogólno–geometrycznej natury, korzystający z różnowartościowości ϕ. Jego dowód pomijamy. □

Wykażemy teraz odpowiednik twierdzenia 18.

Twierdzenie 24. Niech $G \subset \mathbb{R}^d$ będzie mierzalny w/g Jordana-Riemanna a funkcja $f : G \to \mathbb{R}$ będzie całkowalna. Załóżmy, że $\{P_n\}_{n=1}^{\infty}$ jest takim ciągiem rodzin kostek

$$P_n = \{Q(x_i^n, \delta_n)\}_{i=1}^{k_n},$$

że

$$\lim_{n\to\infty} \delta_n = 0 \quad \text{i} \quad \bigcup_{i=1}^{k_n} Q(x_i^n, \delta_n) = G_*(\delta_n) \subset G \quad \text{oraz} \quad \lim_{n\to\infty} \mu_d(G_*(\delta_n)) = \lim_{n\to\infty} k_n\delta^d = \mu_d(G).$$

Jeśli $\phi : \mathbb{R}^d \to \mathbb{R}^d$ jest różnowartościową funkcją klasy C^1, to wtedy

$$\int_{\phi(G)} f(y)\,dy = \lim_{n\to\infty} \sum_{i=1}^{k_n} \int_{\phi(Q_i^n)} f(y)\,dy.$$

Dowód. Załóżmy, że $\max_{\phi(G)} |f| \leq M$. Z przyjętych założeń wynika, że ϕ rozpatrywana na kuli $B(0,R)$ zawierającej G spełnia warunek Lipschitza z pewną stałą L.

Badamy różnicę

$$J_n = \int_{\phi(G)} f(x)\,dx - \int_{\phi(G_*(\delta_n))} f(x)\,dx.$$

Odnotujmy $\phi(G) \setminus \phi(G_*(\delta_n)) = \phi(G \setminus G_*(\delta_n))$, gdzie wykorzystujemy różnowartościowość ϕ. Rozpatrzmy też rodzinę kostek $\{Q(x_j^n, \delta_n)\}_{j=1}^{k^*}$, której suma daje $G^*(\delta_n)$. Wtedy,

$$\phi(G \setminus G_*(\delta_n)) = \phi(G \cap (G^*(\delta_n) \setminus G_*(\delta_n)) = \bigcup_{\{j: Q(x_j^n,\delta_n) \subset G^*(\delta_n),\ \text{ale } \not\subset G\}} \phi(Q_j^n \cap G).$$

Na mocy lematu 22 $\mu_d(\phi(Q(x_j^n,\delta_n)) \leq L^d \text{vol}\,(Q(x_j^n,\delta_n))$. Natomiast ilość kostek Q_j^n w $G^*(\delta_n)$ spełniających $Q_j^n \not\subset G$ jest równa $k^* - k_n$. Skoro G jest mierzalny, to $(k^* - k_n)\delta_n^d$ dąży do zera, gdy $n \to \infty$. Zatem

$$|J_n| \leq \int_{\phi(G \setminus G_*(\delta_n))} |f(x)|\,dx \leq (k^* - k_n)\delta_n^d M \to 0 \quad \text{gdy} \quad n \to \infty. \qquad \square$$

6.6.3 Wzór na zamianę zmiennych w całce

Przedstawimy teraz przypadek ogólny wzoru na zamianę zmiennych w całce wielokrotnej. Niech $G \subset \mathbb{R}^d$ będzie ograniczonym zbiorem mierzalnym i niech $\{P_n\}_{n=1}^{\infty}$ będzie rodziną kostek, $P_n = \{Q_i^n\}_{i=1}^{k_n}$, gdzie $Q_i^n = Q(x_i^n, \delta_n)$. O tej rodzinie zakładamy, że $\delta_n \to 0$ i dla dowolnego $\varepsilon > 0$ istnieje takie N, że

$$\mu(G \setminus \bigcup_{i=1}^{k_n} Q_i^n) < \varepsilon \qquad \text{dla } n > N.$$

Wiemy, że przy założeniach poprzedniego twierdzenia,

$$\int_{\phi(G)} f(y)\,dy = \lim_{n\to\infty} \sum_{i=1}^{k_n} \int_{\phi(Q_i^n)} f(y)\,dy.$$

Zajmiemy się teraz całką na obrazie kwadratu, $\phi(Q_i^n)$,

$$\int_{\phi(Q_i^n)} f(y)\,dy = \int_{R_i^n} f(y)\,dy + r_i(\delta_n),$$

gdzie $R_i^n = D\phi(x_i^n)Q_i^n + \phi(x_i^n)$. Nadto błąd

$$r_i(\delta_n) = \int_{\phi(Q_i^n)} f(y)\,dy - \int_{R_i^n} f(y)\,dy,$$

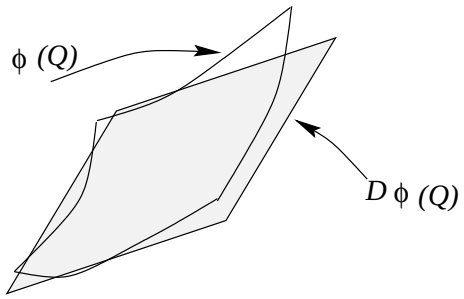
możemy oszacować i dostaniemy

$$|r_i(\delta_n)| \leq \max_{\phi(Q_i^n)} |f(y)| \cdot \mu(R_i^n \triangle \phi(Q_i^n)),$$

gdzie $A\triangle B$ oznacza *różnicę symetryczną*: $A\triangle B = (A\backslash B)\cup(B\backslash A)$. Miarę różnicy symetrycznej możemy oszacować dzięki temu, że dla dostatecznie małych δ_n zbiór $\phi(Q_i^n)$ ma postać

$$\phi(Q_i^n) = \{(x_1, x_2) :\ \psi_1(x_1) \le x_2 \le \psi_2(x_1)\},$$

gdzie funkcje ψ_1 i ψ_2 są ciągłe, (patrz też rysunek poniżej).



Rys. 5. Zdeformowany równoległobok.

Można sprawdzić (szczegóły pozostawiamy Czytelnikowi), że

$$\mu(R_i^n \triangle \phi(Q_i^n)) \le C_1(D\phi)\delta_n^3,$$

gdzie $C_1(D\phi)$ jest pewną stałą od pochodnej $D\phi$ i obszaru G.

Zatem

$$\int_{\phi(Q_i^n)} f(y)\,dy = \int_{Q_i^n} f(\phi(x_i^n))|D\phi(x_i^n)|\,dx + r_i(\delta_n) + r_i'(\delta_n),$$

gdzie $r_i'(\delta_n)$ jest błędem wynikłym z przybliżania $f(\phi(x))$ funkcją stałą $f(\phi(x_i^n))$ na Q_i^n, tym samym

$$|r_i'(\delta_n)| \le C_2(D\phi)\delta_n^3.$$

Po zsumowaniu względem i dostaniemy,

$$\sum_{i=1}^{k_n}\int_{\phi(Q_i^n)} f(y)\,dy = \sum_{i=1}^{k_n}\int_{Q_i^n} f(\phi(x_i^n))|D\phi(x_i^n)|\,dx + \sum_{i=1}^{k_n}(r_i(\delta_n) + r_i'(\delta_n)) = I_n + \text{błąd}_n.$$

Wiemy, (patrz twierdzenie 18), że

$$\lim_{n\to\infty} I_n = \int_G f(\phi(x))|D\phi(x)|\,dx,$$

zaś

$$|\text{błąd}_n| \le \sum_{i=1}^{k_n} C\delta_n^3 = Ck_n\delta_n^3.$$

Skoro $1 + \mu(G) \ge \mu(G^*(\delta_n)) = k_n^*\delta_n^2$, to dla dostatecznie dużych n dostaniemy oszacowanie

$$k_n \le k_n^* \le \frac{1+\mu(G)}{\delta_n^2},$$

tj.

$$|\text{błąd}_n| \leq C\frac{\delta_n^3}{\delta_n^2}(1+\mu(G)) = C\delta_n(1+\mu(G)) \to 0, \quad \text{gdy } n \to \infty.$$

Tym samym wykazaliśmy następujący, zasadniczy fakt.

Twierdzenie 25. Niech $G \subset \mathbb{R}^d$ będzie ograniczonym zbiorem mierzalnym i niech $f : G \to \mathbb{R}$ będzie funkcją ciągłą. Jeśli odwzorowanie $\phi : \mathbb{R}^d \to \mathbb{R}^d$ jest klasy C^1 i różnowartościowe, to mamy wtedy

$$\int_{\phi(G)} f(y)\,dy = \int_G f(\phi(x))|D\phi(x)|\,dx.$$

□

Te przydługie rachunki obrazują taktykę, którą będziemy stosować w następnych rozdziałach poświęconych całkowaniu w różnych sytuacjach. Można ją streścić następująco:

- badamy zachowanie się małego wycinka zbioru: odcinka, kwadratu, który przekształcamy w sposób liniowy. Kwadracikami możemy bowiem przybliżać badany zbiór mierzalny w sensie Jordana-Riemanna.
- następnie możemy ustalić co się dzieje z kwadracikiem, który jest przekształcany w sposób gładki. Przekształcenia różniczkowalne można dobrze przybliżać odwzorowaniami liniowymi.
- dostajemy oczekiwany wzór plus błąd i pokazujemy, że błąd dąży do zera.

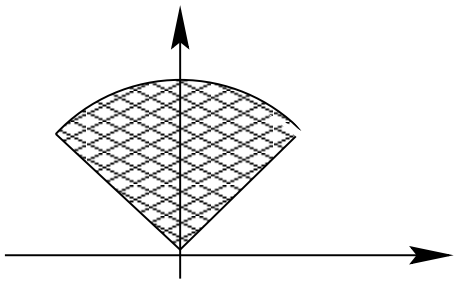
Zastosujmy w praktyce nowy wzór.

Przykład 9. Obliczymy całkę

$$I := \int_G \sqrt{2 - y_1^2 - y_2^2}\,dy_1 dy_2,$$

gdzie G jest wycinkiem kołowym (patrz rys. 6):

$$G = \{(y_1, y_2) : y_1^2 + y_2^2 \leq 2,\ y_2 \geq |y_1|\}.$$



Rys. 6. Wycinek kołowy G.

Wprowadzamy współrzędne biegunowe:

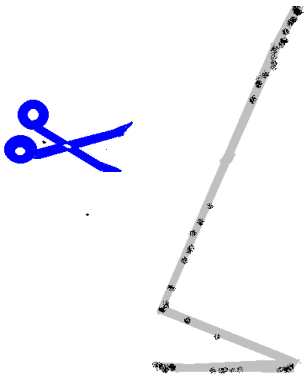
$$y_1 = r\cos\phi, \quad y_2 = r\sin\phi, \qquad r \in [0, \sqrt{2}], \quad \phi \in [\pi/4, 3\pi/4].$$

Piszemy $\Phi(r,\phi) = (y_1, y_2)$, wtedy

$$D\Phi(r,\phi) = \begin{bmatrix} \cos\phi & -r\sin\phi \\ \sin\phi & r\cos\phi \end{bmatrix},$$

i $\det D\Phi(r,\phi) = r$. Dostaniemy

$$I = \int_{\pi/4}^{3\pi/4} \int_0^{\sqrt{2}} \sqrt{2-r^2}\, r\, dr d\phi = \frac{-\pi}{2}\frac{2}{3}(2-r^2)^{3/2}|_0^{\sqrt{2}} = \frac{\pi}{3}2^{3/2}.$$



Rozdział 7

Całki na Krzywych i Powierzchniach

Nasz cel to nauczyć się obliczać długość krzywych, pól powierzchni w przestrzeni. Wymaga to odpowiedniego ich określenia. Obliczanie rozmaitych wielkości fizycznych prowadzi nas do całek krzywoliniowych i powierzchniowych. Osobną rolę gra tu fizyczne pojęcie pracy, które motywuje wprowadzenie orientacji krzywych i powierzchni.

Nasza zasadnicza metoda postępowania polega na powielaniu wyprowadzenia ostatniego twierdzenia poprzedniego rozdziału.

Przedstawimy zastosowania geometryczne i fizyczne, np. do praw zachowania i m.in. objaśnimy napis

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\mathbf{v}\rho) = 0$$

oznaczający prawo ciągłości przepływu cieczy tj. prawo zachowania masy. Symbol $\operatorname{div} X$ oznacza *dywergencję (źródłowość)* pola wektorowego $X : \mathbb{R}^n \to \mathbb{R}^n$,

$$\operatorname{div} X = \sum_{i=1}^{n} \frac{\partial X_i}{\partial x_i}.$$

Skomentujemy takie obiekty, jak ów magiczny operator $\operatorname{div} X$.

Zaczniemy od rzeczy prostszych.

7.1 Długość krzywej, całka krzywoliniowa

Trzeba zacząć od określenia krzywej. To czego od niej oczekujemy, to: (i) by dawała się narysować jednym pociągnięciem ołówka i (ii) by w każdym punkcie miała styczną, definiowaną jako granicę siecznych. Nie możemy pominąć drugiego warunku, bo jeśli to zrobimy, to musimy liczyć się z patologiami, takimi jak krzywa Peano wypełniająca cały kwadrat.

Nasze wymagania ujmiemy następująco.

Definicja 1. Podzbiór $\gamma^* \subset \mathbb{R}^n$ nazywamy *krzywą*, jeśli istnieje funkcja $\gamma : [a, b] \to \mathbb{R}^n$ nazywana *parametryzacją krzywej* γ^* spełniająca warunki:

(1) $\gamma([a, b]) = \gamma^*$;

(2) Funkcja $[a,b] \ni t \mapsto \gamma(t) \equiv (\gamma_1, \ldots, \gamma_n)(t)$ jest różnowartościowa i klasy C^1. Wektor $\frac{d\gamma}{dt}(t)$ nazywamy *wektorem stycznym* do γ^* w punkcie $\gamma(t)$.

(3) Dla wszystkich $t \in [a,b]$ mamy, $\frac{d\gamma}{dt}(t) \neq 0$.

Jeśli pominiemy warunek (3) to spotkamy się z osobliwościami: np. wykres funkcji $y = |x|$ okazałby się być krzywą, ale nie ma ona stycznej w punkcie $x = 0$! np. $t \to (t^3, |t|^3)$ jest funkcją klasy C^1, ale jej pochodna w punkcie $t = 0$ jest równa zero.

Zauważmy, że $\frac{d\gamma}{dt}(t)$ jest prędkością chwilową punktu na krzywej. Zmieniając parametryzację możemy dowolnie zmieniać długość i zwrot wektora stycznego.

Powyższa definicja nie obejmuje, niestety tak oczekiwanego przykładu jakim jest okrąg. Trzeba to nadrobić.

Definicja 2. Podzbiór $\gamma^* \subset \mathbb{R}^n$ nazywamy *krzywą zamkniętą*, jeśli istnieje funkcja $\gamma : [a,b] \to \mathbb{R}^n$ nazywana parametryzacją krzywej γ^* spełniająca warunki:

(1) $\gamma([a,b]) = \gamma^*$;

(2) funkcja $(a,b) \ni t \mapsto \gamma(t)$ jest różnowartościowa, $\gamma(a) = \gamma(b)$ i funkcja γ jest klasy C^1.

(3) dla wszystkich $t \in [a,b]$ mamy, $\frac{dg}{dt}(t) \neq 0$ oraz $\frac{dg}{dt}(a) = \frac{dg}{dt}(b)$.

Przykład 1.

(a) $[0, 2\pi] \ni t \mapsto (R\cos t, R\sin t)$ jest parametryzacją okręgu o promieniu R.

(b) $\mathbb{R} \ni t \mapsto (t^3, t^6)$ nie jest parametryzacją paraboli $y = x^2$, ale $\mathbb{R} \ni x \mapsto (x, x^2)$ już nią jest.

Niech teraz $0 \neq v \in \mathbb{R}^n$ i kładziemy $\gamma(t) = vt$, dla $t \in [a,b]$. Wtedy $\gamma^* = \gamma([a,b])$ jest odcinkiem prostej. Łatwo policzyć jego długość $d(\gamma^*)$:

$$d(\gamma^*) = \|\gamma(a) - \gamma(b)\| = (b-a)\|v\|.$$

Przyjmijmy teraz, że γ jest parametryzacją dowolnej krzywej γ^*. Tworzymy podział zbioru argumentów

$$a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b.$$

Długość łuku krzywej $\gamma([x_i, x_{i+1}])$, to w przybliżeniu odległość punktów $\gamma(x_i)$ i $\gamma(x_{i+1})$ tj.

$$\|\gamma(x_{i+1}) - \gamma(x_i)\|. \tag{1}$$

Dlatego przybliżona długość łuku krzywej to

$$\sum_{i=0}^{n-1} \|\gamma(x_{i+1}) - \gamma(x_i)\|. \tag{2}$$

Spodziewamy się, że wraz z rozdrabnianiem podziału odcinka $[a,b]$ (odpowiadającemu rozdrabnianiu podziału krzywej) uzyskujemy coraz lepsze przybliżenie. Dlatego liczbę

$$d(\gamma^*) = \sup_{n \in \mathbb{N}} \sup_{x_1 < \ldots < x_{n-1}} \sum_{i=0}^{n-1} \|\gamma(x_{i+1}) - \gamma(x_i)\|,$$

gdzie $a < x_1 < \ldots < x_{n-1} < b$ jest podziałem odcinka $[a, b]$ nazywamy *długością krzywej* γ^*. Wyprowadzimy teraz wygodny obliczeniowo wzór, zakładając, że γ^* ma parametryzację γ, która jest klasy C^2. W tym celu oszacujemy (1). Z twierdzenia Taylora 3.39 dostaniemy dla każdej ze składowych $\gamma = (\gamma_1, \ldots, \gamma_n)$,

$$\gamma_k(x_{i+1}) - \gamma_k(x_i) = \gamma_k'(x_i)(x_{i+1} - x_i) + r_k^i(x_{i+1} - x_i),$$

gdzie $|r_k^i(x_{i+1} - x_i)| \leq C|x_{i+1} - x_i|^2$ i C zależy od γ''. Zatem

$$\|\gamma(x_{i+1}) - \gamma(x_i)\| = \|\gamma'(x_i)\|(x_i - x_{i+1}) + r^i(x_{i+1} - x_i), \tag{3}$$

gdzie $|r^i(x_{i+1} - x_i)| \leq C|x_{i+1} - x_i|^2$. Wtedy (2) przyjmie postać

$$\sum_{i=0}^{n-1} \|\gamma'(x_i)\|(x_{i+1} - x_i) + r(\delta), \tag{4}$$

gdzie $\delta = \max_{i=0,\ldots,n-1} |x_{i+1} - x_i|$ jest średnicą podziału $[a, b]$ oraz

$$|r(\delta)| \leq \sum_{i=0}^{n-1} |r^i(x_{i+1} - x_i)| \leq C \sum_{i=0}^{n-1} |x_{i+1} - x_i|^2 \leq C\delta^2 n.$$

Jeśli $\delta \leq C_1/n$, dla pewnej stałej C_1, to dostaniemy

$$|r(\delta)| \leq CC_1\delta \leq C_1^2 C/n. \tag{5}$$

Niech teraz P_n będzie takim ciągiem podziałów, których średnice dążą do zera i takim, że

$$\lim_{n\to\infty} \sum_{i=0}^{k_n-1} \|\gamma(x_{i+1}) - \gamma(x_i)\| = d(\gamma^*)$$

i δ_n jest odpowiadającym ciągiem średnic. Możemy założyć, że $\delta_n = 5/k_n$. Zatem z (4) i (5) wynika, że

$$d(\gamma^*) = \lim_{n\to\infty} \sum_{i=0}^{k_n-1} \|\gamma'(x_i)\||x_i - x_{i+1}|.$$

Dzięki ciągłości funkcji $t \to \|\gamma'(t)\|$ prawa strona jest całkowalna w sensie Riemanna, tj. wykazaliśmy:

Twierdzenie 1. Jeśli γ jest parametryzacją klasy C^2 krzywej γ^*, to

$$d(\gamma^*) = \int_a^b \|\gamma'(t)\|\, dt \equiv \int_a^b \sqrt{\sum_{i=1}^n \left(\frac{d\gamma_i}{dt}(t)\right)^2}\, dt. \tag{6}$$

Uwaga. Powyższy wzór (6) nie zależy od wyboru parametryzacji γ. Co więcej, w istocie wystarczy zakładać, że γ jest klasy C^1.

Przykład 2. Obliczmy długość okręgu $S(0,R)$ o promieniu R i środku 0. Wykorzystamy jego parametryzację z poprzedniego przykładu:

$$d(S(0,R)) = \int_0^{2\pi} \sqrt{\left(\frac{dx}{dt}\right)^2}\, dt = \int_0^{2\pi} \sqrt{R^2\sin^2 t + R^2\cos^2 t}\, dt = 2\pi R.$$

Rozpatrzmy inną sytuację.

Przykład 3. Obliczyć energię kinetyczną półokręgu o liniowej gęstości masy ρ, wirującego ze stałą prędkością kątową ω wokół osi ℓ przebiegającej przez końce półokręgu.

Spróbujmy znaleźć wynik przybliżony. W tym celu możemy podzielić krzywą (tj. okrąg) na drobne kawałki Δl_i i przyjąć, że odległość każdego punktu $p \in \Delta l_i$ od osi obrotu jest stała. Wtedy energia kinetyczną fragmentu Δl_i to w przybliżeniu

$$\Delta e_i = \frac{1}{2}\rho\omega^2 d(p_i,\ell)^2|\Delta l_i| + \text{błąd},$$

gdzie $|\Delta l_i|$ jest długością fragmentu Δl_i a $d(p,\ell)$ jest odległością punktu p od osi obrotu ℓ. Ze wzoru (3) dostaniemy

$$\Delta e_i = \frac{1}{2}\rho\omega^2 d(\gamma(t_i),\ell)^2|\gamma'(t_i)|\Delta t_i + \text{błąd},$$

gdzie γ jest parametryzacją okręgu i $p_i = \gamma(t_i)$ Sumując te przyczynki dostaniemy

$$E_k \approx \sum_{i=0}^{k} \frac{1}{2}\rho\omega^2 d(\gamma(t_i),\ell)^2|\gamma'(t_i)|\Delta t_i.$$

Ten wynik przypomina sumę Riemannowską. Możemy więc wprowadzić nowe pojęcie. Jeśli przejdziemy do zera z rozdrobnieniem przedziału zbioru parametrów γ, to dostaniemy całkę Riemanna z $\frac{1}{2}\rho\omega^2 d(\gamma(t))^2|\gamma'(t)|$ po $[a,b]$ i dostaniemy nowe pojęcie.

Definicja 3. Załóżmy, że γ^* jest krzywą, $\gamma : [a,b] \to \mathbb{R}^n$ jest jej parametryzacją. Nadto, funkcja $f : \gamma^* \to \mathbb{R}$ jest ciągła. Wtedy *całką krzywolinią* funkcji f na krzywej γ^* nazywamy liczbę

$$\int_{\gamma^*} f(x)\, dl := \int_a^b f(\gamma(t))\|\gamma'(t)\|\, dt.$$

Nietrudno wykazać, że ta całka nie zależy od wyboru parametryzacji.

Możemy teraz dokończyć rozwiązywanie zadania. Zauważmy, że $d(\gamma(t),\ell) = R\cos t$, gdzie $x = R\cos t$, $y = R\sin t$, $t \in [-\pi/2, \pi/2]$, jest parametryzacją. Zatem $\|\gamma'\| = R$ i

$$E_k = \int_{\gamma^*} \frac{1}{2}\rho\omega^2 d(x,\ell)^2 dl = \int_{-\pi/2}^{\pi/2} \rho\omega^2 R^3\cos^2 t\, dt = \rho\omega^2 R^3\pi.$$

7.2 Powierzchnie

Przez powierzchnię będziemy rozumieli zbiór, który w otoczeniu każdego punktu można przedstawić jako zdeformowane koło. Oczywiście, będziemy musieli nadać temu pojęciu ścisłą postać, która będzie podobna do definicji krzywej i w miarę ogólna, choć najważniejszy przypadek to powierzchnia w $\mathbb{R}^3$.

Definicja 4. Niech $S \subset \mathbb{R}^m$ i istnieje takie $\phi : U \subset \mathbb{R}^n \to \mathbb{R}^m$, $m \geq n$, że
(P1) U jest otwartym podzbiorem $\mathbb{R}^n$ i $S = \phi(U)$;
(P2) ϕ jest klasy C^1, różnowartościowa i $\ker D\phi(x) = \{0\}$ dla wszystkich $x \in U$.
Wtedy powiemy wtedy, że ϕ jest *parametryzacją* zbioru S klasy C^1.

Uwaga. Warunek (P2) definicji oznacza, że wektory $\dfrac{\partial \phi}{\partial x_1}(x), \ldots, \dfrac{\partial \phi}{\partial x_n}(x)$ są lnz.

Zauważmy też, że jeśli $p = \phi(x_0) \in S$ i $0 \neq v \in \mathbb{R}^n$, to

$$t \mapsto \phi(x_0 + vt) = \gamma_v(t) \tag{7}$$

jest parametryzacją pewnej krzywej, bo dzięki warunkowi (P2) definicji mamy $\frac{d}{dt}\phi(x_0 + vt) = w \neq 0$, tj. w jest wektorem stycznym w punkcie p do krzywej $\gamma_v([-\varepsilon, \varepsilon])$ dla pewnego $\varepsilon > 0$. Możemy jednak wyobrazić sobie inne krzywe przechodzące przez p. Dlatego powiemy, że wektor w *jest styczny* do powierzchni S w punkcie p, jeśli istnieje krzywa γ^* przechodząca przez p i taka, że w jest styczny do γ^* w p. Wprowadzimy oznaczenie,

$$T_pS := \{v \in \mathbb{R}^m : \ v \text{ jest wektorem stycznym do } S \text{ w punkcie } p\}.$$

Zbiór T_pS nazwiemy *przestrzenią styczną* do S w punkcie p.

Zauważmy, że dzięki istnieniu parametryzacji ϕ wzór (7) zapewnia, iż T_pS jest przestrzenią wektorową, będącą podprzestrzenią $\mathbb{R}^m$. Nadto, jest ona rozpinana przez wektory $\dfrac{\partial \phi}{\partial x_1}(x), \ldots, \dfrac{\partial \phi}{\partial x_n}(x)$.

Dla porządku wprowadźmy dodatkowe oznaczenie. Mianowicie,

$$N_pS := \{v \in \mathbb{R}^m : \ v \text{ jest prostopadły do wszystkich wektorów z } T_pS\}.$$

Przykład 4. Spróbujmy teraz opisać parametrycznie sferę $S^2(0, R)$ w $\mathbb{R}^3$ o promieniu R i o środku w 0. Niech ϕ będzie szerokością geograficzną, tj. kąt ϕ liczymy od równika i $\phi \in (-\pi/2, \pi/2)$, wtedy $z = R\sin\phi$. Niech λ będzie długością geograficzną mierzoną od arbitralnie wybranego południka 0, np. przyjmijmy, że $(R, 0, 0)$ ma długość geograficzną 0. Wtedy równoleżniki są okręgami o średnicy $R\cos\phi$ i mamy ich parametryzację

$$x = R\cos\phi\cos\lambda, \quad y = R\cos\phi\sin\lambda, \qquad \lambda \in [0, 2\pi).$$

Zatem

$$\Phi(\phi,\lambda)=\begin{pmatrix} R\cos\phi\cos\lambda \\ R\cos\phi\sin\lambda \\ R\sin\phi \end{pmatrix}, \quad (\phi,\lambda)\in U:=(\pi/2,\pi/2)\times(0,2\pi)$$

jest parametryzacją sfery **bez** półokręgu $S^2(0,R)\cap\{y=0, x<0\}$. Odwzorowanie Φ jest różnowartościowe i ker $D\Phi=\{0\}$ dla (ϕ,λ) ze zbioru U, ale żadnego z przedziałów w definicji U nie można domknąć bez naruszania różnowartościowości.

Na marginesie zauważmy, że skoro każdy punkt $(x,y,z)\in\mathbb{R}^3$ leży na pewnej sferze o promieniu $R\geq 0$, to podanie R i w/w kątów jednoznacznie opisuje jego położenie. Wyjątkiem jest półpłaszczyzna $\{y=0, x<0\}$. Trójkę (R,ϕ,λ) nazywamy *współrzędnymi sferycznymi.*

Przykład próby parametryzacji sfery, którą chcielibyśmy nazywać powierzchnią, pokazuje dwie trudności:

(a) opis parametryczny jest kłopotliwy i może być mało przejrzysty;

(b) nawet prostych zbiorów nie daje się w całości opisać parametrycznie.

Kłopot (a) nie ma nic wspólnego z przyjętą definicją, gdy już raz uzgodnimy co to jest powierzchnia, to poszukamy alternatywnych sposobów opisu. Kłopot (b) oznacza tyle, że musimy ograniczyć nasze wymagania do tego, by opis parametyczny był możliwy przynajmniej lokalnie. Przyjmiemy więc nowe określenie.

Definicja 5. Podzbiór $S\subset\mathbb{R}^m$ nazywamy *powierzchnią n-wymiarową*, jeśli dla każdego punktu $x\in S$ istnieje jego otoczenie V, że zbiór $V\cap S$ ma parametryzację, której zbiór argumentów jest otwartym podzbiorem $\mathbb{R}^n$.

W myśl tej definicji sfera w $\mathbb{R}^3$ już będzie powierzchnią, aczkolwiek szczegóły poprawienia podanej wyżej parametryzacji zostawiamy Czytelnikowi.

Trzymając się naszego pierwszego przykładu przypominamy, że sfera o środku w punkcie 0 i promieniu R to

$$S^2(0,R)=\{(x,y,z):\ x^2+y^2+z^2=R^2\}.$$

Jest więc ona przeciwobrazem punktu R^2 funkcji $f:\mathbb{R}^3\to\mathbb{R}$ danej wzorem $f(x,y,z)=x^2+y^2+z^2$. Można się więc spodziewać, że jest to równoprawny sposób zadawania powierzchni. Istotnie tak jest, ale pod pewnymi warunkami.

Definicja 6. Niech będzie dana funkcja $f:\mathbb{R}^m\to\mathbb{R}$ klasy C^1 i niech $r\in\mathbb{R}$. Powiemy, że r jest *wartością regularną* funkcji f, jeśli $f^{-1}(r)\neq\emptyset$ i grad$f(x)\neq 0$ dla wszystkich $x\in f^{-1}(r)$. Dla dowolnego $c\in\mathbb{R}$ *poziomicą* nazwiemy zbiór $M=f^{-1}(c)$.

Okazuje się, że jest prawdziwy następujący fakt, który pozostawimy bez dowodu.

Twierdzenie 2. Niech funkcja $f:\mathbb{R}^{m+1}\to\mathbb{R}$ będzie klasy C^1. Jeśli $r\in\mathbb{R}$ jest wartością regularną funkcji f, to poziomica $M=f^{-1}(r)$ jest powierzchnią wymiaru m.

Przykład 5. Niech $g:\mathbb{R}^3\to\mathbb{R}$ będzie dana wzorem:

$$g(x,y,z)=x^2-y^2-z^2,$$

wtedy wszystkie jej wartości z wyjątkiem 0, są regularne i zbiory $g^{-1}(r)$ to hiperboloidy, zaś $g^{-1}(0)$ jest stożkiem, który nie jest powierzchnią, bo nie jest spełniona definicja powierzchni w okolicy punktu 0. Nie są tam też spełnione założenia twierdzenia 2. Lecz zbiór $g^{-1}(0)$ z usuniętym punktem 0 już jest powierzchnią.

Podamy jeszcze korzyść z zadawania powierzchni jako poziomic. Niech $M \subset \mathbb{R}^{m+1}$ będzie poziomicą funkcji f i niech γ będzie parametryzacją dowolnej krzywej γ^* zawartej w M. Wtedy,

$$\frac{df}{dt}(\gamma(t)) = \frac{dc}{dt} = 0,$$

a z drugiej strony

$$\frac{df}{dt}(\gamma(t)) = \left(\text{grad} f(\gamma(t)), \frac{d\gamma(t)}{dt}\right).$$

Zatem wektor $\text{grad} f(\gamma(t))$ jest prostopadły do wektora $\frac{d\gamma(t)}{dt}$. Skoro **wszystkie** wektory styczne do M są postaci $\frac{d\gamma(t)}{dt}$, to wykazaliśmy:

Stwierdzenie 3. $N_pM = \text{lin } \text{grad} f(p)$ i $T_pM = \{v \in \mathbb{R}^{m+1} : (v, \text{grad} f(p)) = 0\}$. □

Przykład 6. Niech $f(x, y, z) = x^2 + y^2 + z^2$ i $M = f^{-1}(1)$. Zauważmy, że dla wszystkich $(x, y, z) \neq 0$ mamy $\text{grad} f(x, y, z) \neq 0$. Rozpatrzmy punkt $p = (\sqrt{3}/3, \sqrt{3}/3, \sqrt{3}/3)$. Skoro

$$\text{grad} f(p) = \frac{2}{3}(\sqrt{3}, \sqrt{3}, \sqrt{3}),$$

to wektory $v = (v_1, v_2, v_3)$ styczne do M w punkcie p spełniają

$$v_1 + v_2 + v_3 = 0,$$

zaś punkty $x = (x_1, x_2, x_3)$ z płaszczyzny stycznej do M w p spełniają

$$(x_1 - \sqrt{3}/3, x_2 - \sqrt{3}/3, x_3 - \sqrt{3}/3) = 0.$$

Ustaliliśmy, kiedy poziomice są powierzchniami. Teraz możemy się zastanowić, czy są powierzchnie, które nimi nie są. Okazuje się, że mamy:

Twierdzenie 4. Jeśli S jest powierzchnią n–wymiarową w $\mathbb{R}^{n+1}$, to dla każdego punktu $x \in M$ istnieją: takie otoczenie $U \subset \mathbb{R}^{n+1}$ punktu x i funkcja $f : U \to \mathbb{R}$, że $S \cap U = f^{-1}(0)$, gdzie 0 jest wartością regularną f.

Uwaga. Nie można żądać, aby $f : \mathbb{R}^{n+1} \to \mathbb{R}$ i $S = f^{-1}(0)$, gdzie 0 jest wartością regularną f, tj. aby cała powierzchnia S była poziomicą. Przykładem jest wstęga Möbiusa powstała po skręceniu o 180 stopni i następnym sklejeniu paska papieru. Jej parametryzacja $\phi : [-1/4, 1/4] \times [0, 2\pi) \to \mathbb{R}^3$ jest dana wzorem:

$$\phi(t, \theta) = ((1 + t\cos\frac{\theta}{2})\cos\theta, (1 + t\cos\frac{\theta}{2})\sin\theta, t\sin\frac{\theta}{2}).$$

Próba pomalowania jej dwoma kolorami prowadzi do wniosku, że wstęga Möbiusa ma tylko jedną stronę! Natomiast poziomice mają wyróżnione strony: tam gdzie $f > 0$ i tę gdzie $f < 0$.

Pozostał nam jeszcze jeden ważny przykład do omówienia, który łączy w sobie oba podejścia do definiowania powierzchni. Niech $U \subset \mathbb{R}^n$ będzie otwarty i funkcja $f : U \to \mathbb{R}$ będzie klasy C^1. Zbiór $\Gamma(f) \subset \mathbb{R}^n \times \mathbb{R}$,

$$\Gamma(f) = \{(x, y) \in U \times \mathbb{R} : \ y = f(x)\}$$

nazywamy *wykresem* funkcji f.

Zauważamy, że $\Gamma(f)$ jest powierzchnią a funkcja $\phi : U \to \mathbb{R}^{n+1}$ dana wzorem $\phi(x) = (x, f(x))$ jest jej parametryzacją. Łatwo bowiem sprawdzić, że ϕ jest funkcją różnowartościową. Liczymy $D\phi(x)$:

$$D\phi(x) = [I, \text{grad} f(x)]^T,$$

tj. $D\phi(x)$ na n kolumn o długości $n+1$, $k-$ta kolumna ma 1 na miejscu k, na miejscu $n+1$ ma $\frac{\partial f(x)}{\partial x_k}$, zaś na pozostałych 0. Znowu łatwo sprawdzić, że $\ker D\phi(x) = \{0\}$. Co więcej okazuje się, że $\Gamma(f)$ jest poziomicą. Mianowicie kładziemy $\psi : U \times \mathbb{R} \to \mathbb{R}$, $\psi(x, y) = y - f(x)$ i oczywiście $\text{grad}\psi(x, y) \neq 0$ dla wszystkich (x, y) i $\Gamma(f) = \psi^{-1}(0)$.

Na koniec podrozdziału podamy bez dowodu ogólniejszy fakt dotyczący poziomic.

Twierdzenie 5. Niech $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, k$ i $M = f_1^{-1}(a_1) \cap \ldots \cap f_k^{-1}(a_k) \neq \emptyset$, gdzie $a_i \in \mathbb{R}$, $i = 1, \ldots, k$. Jeśli dla wszystkich $x \in M$ wektory

$$\text{grad}\, f_1(x), \ldots, \text{grad}\, f_k(x)$$

są lnz, to zbiór M jest $n - k$-wymiarową powierzchnią. Nadto, $N_p M = \text{lin}_{\,i=1,\ldots,k}\{\text{grad}\, f_i\}$.

Jest to bardzo naturalny fakt, np. przecięcie sfery $S^2(0, R)$ z dowolną płaszczyzną $a_1x_1 + a_2x_2 + a_3x_3 = 0$ jest okręgiem, czyli krzywą zamkniętą albo 1-wymiarową powierzchnią.

7.3 Pole powierzchni

Tytuł podrozdziału jednoznacznie określa nasze cele. Będziemy zakładali, że powierzchnia S jest sparametryzowana za pomocą $\phi : U \to \mathbb{R}^m$, gdzie $U \subset \mathbb{R}^n$. Rozpatrzmy najprostszy przypadek, $U = (0, 1)^2$, $m = 3$ i ϕ jest funkcją liniową, tj. $\phi(U)$ jest równoległobokiem $R = R(v, w)$ (patrz §2.5.2). Niech $\phi(\mathbf{e}_1) = v = (v_1, v_2, v_3)$, $\phi(\mathbf{e}_2) = w = (w_1, w_2, w_3)$, wtedy na mocy wzoru (24) z §2.5.4 i wzoru z §2.5.3 na współrzędne iloczynu wektorowego wektorów v i w:

$$\text{pole}\, R(v, w) = |v \times w| = \sqrt{\left(\det \begin{bmatrix} v_1 & w_1 \\ v_2 & w_2 \end{bmatrix}\right)^2 + \left(\det \begin{bmatrix} v_1 & w_1 \\ v_3 & w_3 \end{bmatrix}\right)^2 + \left(\det \begin{bmatrix} v_2 & w_2 \\ v_3 & w_3 \end{bmatrix}\right)^2}.$$

Wykażemy, że

$$\text{pole}\, R(v, w) = \sqrt{\det \begin{bmatrix} (v, v) & (v, w) \\ (v, w) & (w, w) \end{bmatrix}} \equiv \sqrt{\det A^T A}, \tag{8}$$

gdzie kolumnami macierzy A są wektory v i w. Mianowicie, jeśli α jest kątem pomiędzy wektorami v i w, to mamy

$$\begin{aligned}\det \begin{bmatrix} (v,v) & (v,w) \\ (v,w) & (w,w) \end{bmatrix} &= (v,v)(w,w) - (v,w)^2 = (v,v)(w,w)(1-\cos^2\alpha) \\ &= \|v\|^2 \cdot \|w\|^2 \sin^2\alpha = (\text{pole } R)^2.\end{aligned}$$

Co i należało wykazać.

Okazuje się, że powyższą uwagę można uogólnić. Jeśli $\phi : \mathbb{R}^n \to \mathbb{R}^m$ jest odwzorowaniem liniowym i $m \geq n$, to

$$\text{vol } (\phi([0,1]^n)) \equiv \text{vol } R(\phi(\mathbf{e}_1), \ldots, \phi(\mathbf{e}_n)) = \sqrt{\det G}, \tag{9}$$

gdzie, utożsamiwszy odwzorowanie ϕ z jego macierzą, napisaliśmy

$$G = \phi^T \cdot \phi$$

i G nazywaliśmy *macierzą Grama* układu wektorów $\phi(\mathbf{e}_1), \ldots, \phi(\mathbf{e}_n)$, tj. kolumn macierzy ϕ. Można wykazać, że

$$\det G = \sum_{j \in J} (\det A_j)^2, \tag{10}$$

gdzie $A_j \in M_{n\times n}(\mathbb{R})$ jest macierzą powstałą z macierzy ϕ poprzez wykreślenie $m-n$ wierszy ze zbioru j, zaś J jest zbiorem wszystkich podzbiorów n elementowych zbioru $\{1, \ldots, m\}$.

W praktyce najczęściej będziemy korzystali z (8), który jest odmianą (9) i (10).

Niech teraz ϕ będzie dowolną parametryzacją powierzchni S. Wtedy, dla y bliskiego x mamy $\phi(y) = \phi(x) + D\phi(x)(y-x)+$błąd, zatem dla kostki $Q_i = Q(x_i, \delta)$ mamy

$$\text{vol } (\phi(Q_i)) = \text{vol } (D\phi(x)Q_i) + r,$$

gdzie błąd r szacuje się przez stałą razy δ^{n+1}. Używamy do tego argumentacji podobnej do tej stosowanej w podrozdziale 6.6 o zamianie zmiennych. Niech teraz $\{\{Q_i^l(\delta_l)\}_{i=1}^{k_l}\}_{l=1}^{\infty}$ będzie rodziną rodzin kostek, spełniającą $\bigcup_{i=1}^{k_l} Q_i^l(\delta_l) = G_*(\delta_l)$. Wtedy

$$\sum_{i=1}^{k_l} \text{vol } (D\phi(x)Q_i^l) = \sum_{i=1}^{k_\delta} \int_{Q_i^l} \sqrt{\det[(D\phi(x_i))^T D\phi(x_i)]}\, dx$$

jest sumą Riemannowską, przybliżającą wielkość, którą można nazwać n-wymiarową miarą. Zauważmy, że w/w sumy Riemannowskie są zbieżne, gdy $\delta_l \to 0$. Możemy wtedy przyjąć określenie.

Definicja 7. Niech $\phi : U \to \mathbb{R}^m$ będzie dowolną parametryzacją n-wymiarowej powierzchni S. Kładziemy

$$|D\phi(x)| := \sqrt{\det(D\phi(x))^T D\phi(x)}$$

i nazywamy *modułem* ϕ, tj. $|D\phi(x)|^2$ jest wyznacznikiem macierzy Grama wektorów $D\phi(x_i)\mathrm{e}_1, \ldots, D\phi(x_i)\mathrm{e}_n$. Wreszcie przyjmujemy,

$$\mu_n(S) := \int_U |D\phi(x)|\, dx$$

i nazywamy *miarą* S. Jeśli $n = 2$, to $\mu_2(S)$ nazywamy *polem powierzchni* S.

Uwaga. Chcemy, aby $\mu_n(S)$ nie zależało od parametryzacji ϕ, w tym celu trzeba poprawić definicję 4. powierzchni n-wymiarowych **dodając** warunek:

(P3) jeśli $\phi : U \to S$, $\psi : V \to S$ i $\phi(U) \cap \psi(V) \neq \emptyset$ są parametryzacjami, to złożenia $\phi\psi^{-1}$ i $\psi\phi^{-1}$ są klasy C^1 (tam gdzie są określone). Można wtedy zastosować twierdzenie 6.25. o zamianie zmiennych w całce do tego, aby wywnioskować, że wyżej wprowadzona miara powierzchni $\mu(S)$ **nie zależy** od parametryzcji ϕ.

Przykład 7. Policzmy pole powierzchni S^+, tj. półsfery o promieniu 1 i środku w punkcie 0. Jest ona wykresem funkcji $g(x, y) = \sqrt{1 - x^2 - y^2}$. Dostaniemy zatem, że $\phi : B(0, 1) \to \mathbb{R}^3$ dana wzorem

$$\phi(x, y) = (x, y, g(x, y))$$

jest parametryzacją S. Liczymy,

$$D\phi(x, y) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{-x}{\sqrt{1-x^2-y^2}} & \frac{-y}{\sqrt{1-x^2-y^2}} \end{bmatrix},$$

i mamy

$$|D\phi(x, y)| = \frac{1}{\sqrt{1 - x^2 - y^2}}.$$

Wreszcie

$$\mu_2(S) = \int_{B(0,1)} |D\phi(x, y)|\, dxdy = \int_{B(0,1)} \frac{dxdy}{\sqrt{1 - x^2 - y^2}}.$$

Zauważamy, że w naturalny sposób pojawiła się całka niewłaściwa Riemanna, (patrz 6.5). Zgodnie z twierdzeniem 6.16 wystarczy obliczyć granicę

$$\lim_{\rho \to 1^-} \int_{B(0,1)} \chi_{B(0,\rho)} |D\phi(x, y)|\, dxdy.$$

Następnie przechodzimy do zmiennych biegunowych w powyższej całce i dostaniemy zwykłą całkę Riemanna na $B(0, \rho)$

$$\mu_2(S) = \lim_{\rho \to 1^-} \int_0^\rho \int_0^{2\pi} \frac{r}{\sqrt{1 - r^2}}\, drd\theta = 2\pi \lim_{\rho \to 1^-} \int_0^\rho \frac{r\, dr}{\sqrt{1 - r^2}} = -2\pi \lim_{\rho \to 1^-} \sqrt{1 - r^2}|_\rho^1 = 2\pi.$$

Rozpatrzmy nowe zadanie.

Przykład 8. Obliczyć masę półsfery S^+, jeśli jej gęstość jest równa stałej ρ_0 razy odległość punktu x od osi północ – południe.

Spróbujmy znaleźć wynik przybliżony, zakładając, że kawałek powierzchni ΔS_i, którego elementem jest ξ_i (będzie on wybrany później) znajduje się w odległości $d(\xi_i)$ od osi. Wtedy wynik przybliżony to

$$m = \sum_{i=1}^{k_\delta} \rho_0 d(\xi_i)\mu_2(\Delta S_i). \tag{11}$$

Widzimy, że prawa strona wygląda jak suma Riemannowska, spodziewamy się, że wynik będzie całką. Przyjrzyjmy się przyczynkom $d(\xi_i)\mu_2(\Delta S_i)$. Jeśli $\phi : U \to \mathbb{R}^3$ jest parametryzacją powierzchni S^+, to $\mu_2(\Delta S_i) = \int_{Q(y_i,\delta)} |D\phi(y)|\, dy$ i wzór (11) przyjmuje postać:

$$m = \sum_{i=1}^{k_\delta} \rho_0 d(\xi_i) \int_{Q(y_i,\delta)} |D\phi(y)|\, dy.$$

Teraz $\xi_i = \phi(x_i)$ i $x_i \in Q(y_i, \delta)$ jest tak wybrane dzięki twierdzeniu o wartości średniej, aby $\int_{Q(y_i,\delta)} |D\phi(y)|\, dy = |D\phi(x_i)|\text{vol}\,(Q(y_i, \delta))$. Zatem m jest sumą Riemannowską

$$m = \sum_{i=1}^{k_\delta} \rho_0 d(\phi(x_i))|D\phi(x_i)|\text{vol}\,(Q(y_i, \delta)).$$

Jest teraz jasnym, że możemy wykonać przejście graniczne $\delta \to 0$, gdzie $\delta = (\text{vol}\, Q(y_i, \delta))^{1/2}$.

Możemy przyjąć, następującą definicję

Definicja 8. Niech $S \subset \mathbb{R}^m$ będzie powierzchnią n-wymiarową, która ma parametryzację $\phi : U \to S$, U jest otwartym podzbiorem $\mathbb{R}^n$, dalej $f : S \to \mathbb{R}$ jest funkcją ciągłą, wtedy kładziemy

$$\int_S f(y)\, dS := \int_U f(\phi(x))|D\phi(x)|\, dx.$$

i nazywamy *całką powierzchniową*.

Uwagi. (1) Całka powierzchniowa nie zależy od wyboru parametryzacji. Wynika to z warunku (P3) definicji powierzchni (wprowadzonego tuż przed przykładem 7.) i twierdzenia 6.25.

(2) W przypadku, gdy nie można sparametryzować danej powierzchni S za pomocą jednej funkcji ϕ trzeba całkę powierzchniową $\int_S f(y)\, dS$ przedstawić w postaci sumy

$$\int_S f(y)\, dS = \sum_{i=1}^{k} \int_{S_k} f(y)\, dS,$$

gdzie każda powierzchnia S_i już ma parametryzcję $\phi_i : U_i \to S_i$. Musimy nadto założyć, że $S_i \cap S_j = \emptyset$, gdy $i \neq j$.

Przykład 8, cd. W warunkach naszego zadania mamy $f(x) = \rho_0\sqrt{x_1^2 + x_2^2}$ i przyjmujemy parametryzację

$$x_1 = R\cos\phi\cos\lambda, \quad x_2 = R\cos\phi\sin\lambda, \quad x_3 = R\sin\phi, \quad (\phi, \lambda) \in [0, 2\pi) \times (0, \pi/2).$$

Zatem będziemy pisać $\Phi(\phi, \lambda) = (x_1, x_2, x_3)$ i mamy

$$D\Phi(\phi, \lambda) = \begin{pmatrix} -R\sin\phi\cos\lambda & -R\cos\phi\sin\lambda \\ -R\sin\phi\sin\lambda & R\cos\phi\cos\lambda \\ R\cos\phi & 0 \end{pmatrix}.$$

Łatwo już dostaniemy, że $|D\Phi(\phi, \lambda)| = R^2|\cos\phi|$ i $f(x) = \rho_0\sqrt{R^2\cos^2\phi}$. Możemy dokończyć rachunki masy:

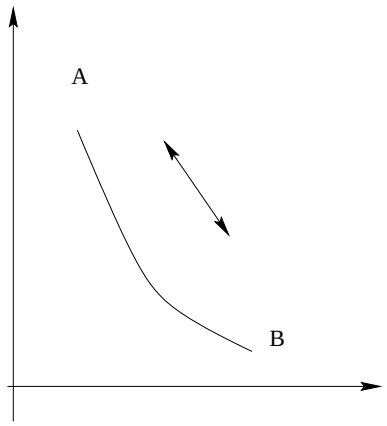
$$m = \int_0^{2\pi}\int_0^{\pi/2} \rho_0 R^3\cos^2\phi\, d\phi d\lambda = \frac{1}{2}\rho_0 R^3\pi^2.$$

7.4 Praca jako całka 1-formy

Wiemy, że praca jest równa sile razy przesunięcie. Z drugiej strony nie każda siła pracuje, np. siła odśrodkowa czy Lorentza nie wykonują pracy, ich wspólną cechą jest to, że są **prostopadłe do** kierunku ruchu ciała (cząstki itp.). Zatem naszą wstępną definicję należy poprawić:

$$W = F\Delta L\cos\alpha,$$

gdzie α jest kątem pomiędzy wektorem siły a wektorem w kierunku ruchu, a właściwie, stycznym do krzywej, wzdłuż której ruch się odbywa. Co więcej ruch, np. kulki po rynience, jak na rysunku 1, w ziemskim polu grawitacyjnym może się odbywać pod wpływem siły ciążenia albo przeciw niej.



Rys. 1. Praca pola sił wzdłuż krzywej.

Jeśli więc $\vec{F} = m\vec{g}$, gdzie $\vec{g} = (0, 0, -g)$, to praca sił pola ciężkości na odcinku o długości ΔL jest równa $m\Delta L(\vec{t}, \vec{g})$, gdzie $\vec{t}$ jest jednostkowym wektorem stycznym pokazującym kierunek ruchu. Naturalnie, praca **przeciwko** siłom pola ciężkości na tym samym odcinku, to $-m\Delta L(\vec{t}, \vec{g})$. Zatem praca pola po krzywej jak na rysunku 1. od punktu A do punktu B, to znowu suma drobnych przyczynków. Jest ona sumą Riemannowską całki krzywoliniowej

$$W = \int_{\vec{AB}} (\vec{F}, \vec{t})dl.$$

Niech teraz $\gamma : [a, b] \to \mathbb{R}^2$ będzie dowolną parametryzacją krzywej na rysunku. Wtedy wektor $\vec{t} = \gamma'(s)/\|\gamma(s)\|$ jest jednostkowym wektorem stycznym, pokazującym przebieg od A do B. Zatem

$$W = \int_a^b (\vec{F}(\gamma(s)), \gamma'(s)/\|\gamma'(s)\|)\|\gamma'(s)\|\, ds = \int_a^b (\vec{F}(\gamma(s)), \gamma'(s))\, ds,$$

co już nie jest zwykłą całką krzywoliniową.

W ogólności, jeśli chcemy znaleźć pracę pola wektorowego $\vec{F}$ w $\mathbb{R}^n$ wzdłuż krzywej γ^*, to dostaniemy:

$$W = \int_a^b (\vec{F}(\gamma(s)), \gamma'(s))\, ds \equiv \int_a^b \sum_{i=1}^{n} F_i(\gamma(s)) \frac{d\gamma_i}{dt}(s)\, ds, \tag{12}$$

gdzie γ jest parametryzacją γ^*. Jeśli praca byłaby wykonywana przeciwko siłom pola, to

$$-W = -\int_a^b (\vec{F}(\gamma(s)), \gamma'(s))\, ds = \int_b^a (\vec{F}(\gamma(s)), \gamma'(s))\, ds.$$

Podkreślamy, że ważny jest tu **kierunek**.

Zauważmy jeszcze jedną osobliwość wyniku: **liczba** (praca) jest wynikiem działania **pola** wektorowego wzdłuż **krzywej**, dokładniej funkcją podcałkową jest iloczyn skalarny, który możemy zapisać:

$$(\vec{F}, \vec{t}) \equiv \vec{F}^T \cdot \vec{t},$$

gdzie po prostu przypomnieliśmy definicję iloczynu skalarnego, jako iloczynu macierzy o jednym wierszu $\vec{F}^T$ i macierzy $\vec{t}$ o jednej kolumnie. Aby podkreślić znaczenie naszych rozważań wprowadzimy stosowne określenie.

Definicja 9. Przekształcenie liniowe $L : \mathbb{R}^n \to \mathbb{R}$, tj. macierz o jednym wierszu nazywamy *formą liniową (1-formą liniową)*.

Oczywiście, łatwo utożsamić 1-formy z wektorami:

$$\mathbb{R}^n \ni \begin{pmatrix} F_1 \\ \vdots \\ F_n \end{pmatrix} = F \mapsto F^T = F^* = (F_1, \ldots, F_n) \in \mathrm{Hom}\,(\mathbb{R}^n, \mathbb{R}).$$

W przypadku odwzorowań $\mathbb{R}^n \ni x \to F^*(x) \in \mathrm{Hom}\,(\mathbb{R}^n, \mathbb{R})$ będziemy też pisać: $F^*(x) = F_1(x)dx_1 + \ldots + F_n(x)dx_n$ i mówić, że $F^*(x)$ jest *formą różniczkową*.

Dlatego prawą stronę (12) można przepisać jako

$$\int_a^b F^*(\gamma(s)) \frac{d\gamma(s)}{dt}\, ds.$$

Podsumowaniem będą nowe określenia. Powiemy, że krzywa γ^* jest *zorientowana*, jeśli jest wyróżniony sposób jej obchodzenia, np. początek i koniec. Określenie sposobu obchodzenia krzywych nazywamy *orientacją krzywej*. W szczególnym przypadku płaszczyzny krzywe zamknięte można obchodzić zgodnie albo niezgodnie z ruchem wskazówek zegara, ten ostatni sposób obejścia nazywamy *orientacją naturalną*. Krzywą γ^* z wyróżnioną orientacją będziemy oznaczali symbolem $\vec{\gamma}^*$.

Po tych uwagach o orientacji możemy przejść do zasadniczego pojęcia.

Definicja 10. Niech będzie dana krzywa zorientowana $\vec{\gamma}^* \subset \mathbb{R}^n$ i 1-forma $F^*(x)$, zależąca w sposób ciągły od zmiennej x. *Całką z 1-formy różniczkowej $F^*(x)$ po krzywej zorientowanej $\vec{\gamma}^*$* nazwiemy liczbę

$$\int_{\vec{\gamma}^*} F^*(x) \equiv \int_{\vec{\gamma}^*} F_1(x)dx_1 + \ldots + F_n(x)dx_n := \int_a^b \sum_{i=1}^n F_i(\gamma(s))\gamma_i'(s)\, ds,$$

gdzie γ jest taką parametryzacją krzywej $\vec{\gamma}^*$, że parametr s rośnie, gdy przebiegamy krzywą zgodnie z jej orientacją.

Podkreślamy, że zmiana orientacji krzywej prowadzi do zmiany znaku całki.

Przykład 9. Obliczymy

$$\int_{\vec{0A}} 2xydx + x^2dy = I$$

gdzie punkty 0 i $A = (1,1)$ są połączone parabolą $y = x^2$. Parametryzacja tej krzywej to $\gamma(x) = (\gamma_1(x), \gamma_2(x)) = (x, x^2)$ i dalej $\gamma'(x) = (1, 2x)$. Zatem z definicji

$$I = \int_0^1 2y(x)\gamma_1'(x)dx + x^2\gamma_2'(x)dx = \int_0^1 (2x^3 + 2x^3)\, dx = 1.$$

Przykład 10. Praca w polu sił potencjalnych. Niech pole $\vec{F}$ będzie dane wzorem

$$\vec{F} = \text{grad } V,$$

gdzie V jest klasy C^1. Wtedy

$$F^* = \frac{\partial V}{\partial x_1}dx_1 + \ldots + \frac{\partial V}{\partial x_n}dx_n =: dV$$

gdzie prawą stronę nazywamy *różniczką funkcji* V. Niech γ^* będzie dowolną krzywą i $\gamma : [a,b] \to \mathbb{R}^n$ niech będzie jej parametryzacją. Wtedy

$$\int_{\vec{\gamma}^*} F^* \equiv \int_{\vec{\gamma}^*} dV = \sum_{i=1}^n \int_a^b \frac{\partial V}{\partial x_i}(\gamma(t))\frac{d\gamma_i}{dt}(t)\, dt = \int_a^b \frac{dV}{dt}(\gamma(t))\, dt = V(\gamma(b)) - V(\gamma(a)).$$

tj. wynik zależy tylko od początku i końca krzywej, ale nie od jej położenia. W szczególności wynik jest zerem, jeśli krzywa jest zamknięta.

Przykład 11. Inny ważny przykład to całka z 1-formy po zoriEntowanym wykresie funkcji jednej zmiennej o wartościach rzeczywistych. Niech $\gamma^* = \Gamma(f)$, wtedy jej parametryzacją jest $\gamma(t) = (t, f(t))$. Jeśli ω jest dowolną 1-formą różniczkową $\omega = Pdx + Qdy$, to wtedy

$$\int_{\Gamma(\vec{f})} \omega = \int_a^b (P(x, f(x))\, dx + Q(x, f(x))f'(x)\, dx)$$

bo $\gamma'(t) = (1, f'(t))$. W szczególności, jeśli $\omega = P(x,y)dx$, to

$$\int_{\Gamma(\vec{f})} P(x,y)dx = \int_a^b P(x, f(x))\, dx \tag{13}$$

7.5 Wzór Greena

Chcemy powiązać całkę z 1-formy z całką podwójną. Uzyskamy uogólnienie podstawowego twierdzenia rachunku różniczkowego i całkowego. Zaczniemy od określenia.

Definicja 11. Zbiór $L \subset \mathbb{R}^n$ nazywamy *konturem*, jeśli istnieje funkcja ciągła $\varphi : [a,b] \to \mathbb{R}^n$, która jest różnowartościowa na (a,b) i $\varphi(a) = \varphi(b)$ oraz istnieją takie punkty

$$a = x_0 < x_1 < ... < x_{k-1} < x_k = b,$$

że funkcje $\varphi|_{[x_i,x_{i+1}]}$ są parametryzacjami krzywych $\varphi([x_i, x_{i+1}])$.

W myśl tej definicji obwód prostokąta jest konturem, mimo że nie jest krzywą zamkniętą w naszym rozumieniu. Konturami są również zbiory będące sumą skończoną krzywych, jak na rysunkach 2 i 3.

Załóżmy, że krzywa łącząca A i B na rys.2. jest wykresem funkcji $y = y_0(x)$, zaś krzywa łącząca D i C to wykres funkcji $y = y_1(x)$. Dzięki wzorowi (13) możemy napisać

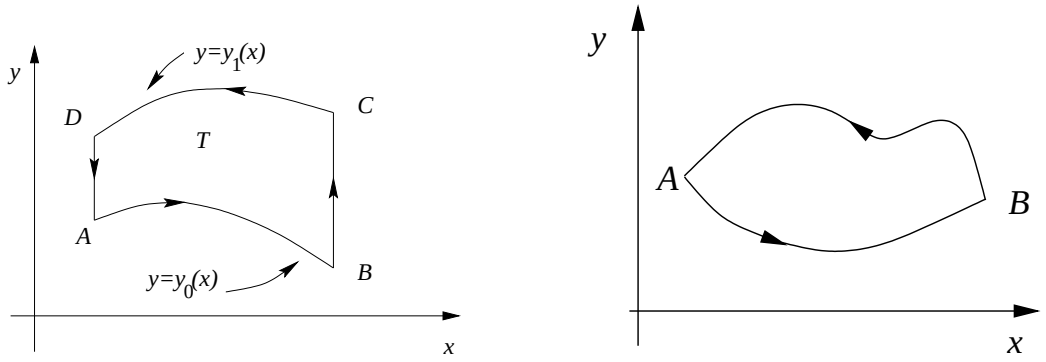
$$\int_{\overrightarrow{AB}} P dx = \int_a^b P(x, y_0(x))dx =: I_0,$$

$$\int_{\overrightarrow{DC}} P dx = \int_a^b P(x, y_1(x))dx =: I_1.$$

Dzięki znanej interpretacji całki Riemanna różnica

$$I = I_1 - I_0 = \int_a^b P(x, y_1(x))dx - \int_a^b P(x, y_0(x))dx$$

jest polem powierzchni obszaru pomiędzy wykresami funkcji $x \to P(x, y_1(x))$ i $x \to P(x, y_0(x))$.



Rys. 2 i 3. Przykłady konturów.

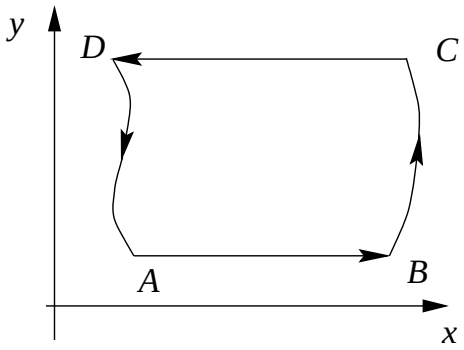
Zakładając, że P jest klasy C^1, podstawowe twierdzenie rachunku różniczkowego i całkowego pozwala nam napisać, że

$$I = \int_a^b \int_{y_0(x)}^{y_1(x)} \frac{\partial P}{\partial y}(x,y)dydx \equiv \int_T \frac{\partial P}{\partial y}(x,y)dydx,$$

gdzie po prawej stronie mamy całkę podwójną po obszarze ograniczonym „trapezem krzywoliniowym" T. Jego brzegiem jest kontur L złożony z krzywych $\vec{AB}$, $\vec{BC}$, $\vec{CD}$ i $\vec{DA}$ zorientowany tak, jak na rysunku 2, tj. w sposób naturalny. Uwzględniając to, że $\int_{CB} Pdx = 0 = \int_{AD} Pdx$ i stosując naturalna orientację konturu możemy napisać

$$-\int_L P(x,y)dx = \int_T \frac{\partial P}{\partial y} dxdy. \tag{14}$$

Dla odmiany rozważmy następujący kontur L, krzywa łącząca A i D jest wykresem funkcji $y \to x_0(y)$, zaś krzywa BC, to wykres funkcji $y \to x_1(y)$, wreszcie AB i DC są odcinkami, (patrz rys. 4).



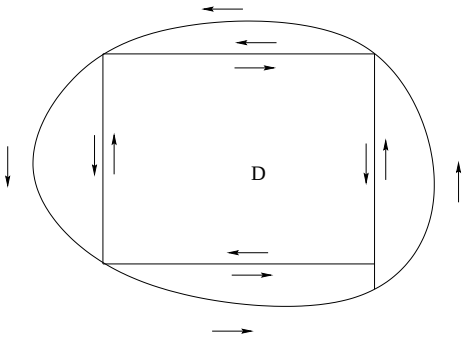
Rys. 4. Kontur L.

Rozumowanie takie jak wyżej, po zamianie rolami x i y da nam

$$\int_L Q(x,y)dy = \int_T \frac{\partial Q}{\partial x}(x,y)\, dxdy, \tag{15}$$

gdzie tym razem dostajemy znak $+$ po lewej stronie.

Przedstawione wyżej rachunki są przeprowadzone dla dość szczególnych konturów. Jednak wynik jest prawdziwy dla dowolnego konturu L, będącego brzegiem obszaru D.



Rys. 4a. Obszar podzielony konturami.

Osiągamy to dzieląc obszar D pomocniczymi konturami, na których możemy stosować wzory (14) i (15). Okazuje się, że wkład pochodzący od sztucznie wprowadzonych konturów jest zerowy. Po zsumowaniu (14) i (15) dostaniemy

$$\int_L P(x,y)dx + Q(x,y)dy = \int_T \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dxdy. \tag{16}$$

Tym samym wykazaliśmy:

Twierdzenie 6. Niech $L = \partial T$ będzie konturem zaś P, Q będą funkcjami klasy C^1 w zbiorze otwartym $\mathcal{U} \supset T$. Wtedy jest prawdziwy wzór (16). Nazywamy go *wzorem Greena*.

Wniosek 7. Jeśli $\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} = 1$, to prawa strona wzoru (16) jest równa polu powierzchni T, np. można przyjąć $P = -\frac{1}{2}y, \;\; Q = \frac{1}{2}x$.

Uwaga mnemotechniczna. 1-forma po lewej stronie (16) przechodzi na „pochodną" po prawej, gdzie różniczkujemy po „brakującej" zmiennej, tj. $Pdx \to \frac{\partial P}{\partial y}dydx, \;\; Qdy \to \frac{\partial Q}{\partial x}dxdy$. Pozostaje nam (jeszcze magiczny) wymóg ustalenia kolejności dx i dy prowadzacej do zmiany znaku przed $\frac{\partial P}{\partial y}$. W stosowanym przez nas zapisie $dxdy = -dydx$, będzimy to wyjaśniali w §7.7.

Mogłoby się wydawać , że gdy $\frac{\partial Q}{\partial x} = \frac{\partial P}{\partial y}$, to $\int_L Pdx + Qdy = 0$. Wymagana jednak jest ostrożność. Rozpatrzmy proste przykłady:

(a) $P = \frac{-y}{x^2+y^2}, \quad Q = \frac{x}{x^2+y^2} \qquad L = \partial B(0,1)$;

(b) $P = \frac{x}{x^2+y^2}, \quad Q = \frac{y}{x^2+y^2} \qquad L = \partial B(0,1)$.

W obu przypadkach $\frac{\partial Q}{\partial x} = \frac{\partial P}{\partial y}$, ale jeśli przyjmiemy parametryzację okręgu L, $x = \cos t, y = \sin t$, zorientowanego w sposób naturalny, to dostaniemy:

(a) $\int_L Pdx + Qdy = \int_D^{2\pi}(\sin^2 t + \cos^2 f)dt = 2\pi$;

(b) $\int_L Pdx + Qdy = \int_D^{2\pi}(-\sin t \cos t + \sin t \cos t)dt = 0$.

Wyjaśnienie tego pozornego paradoksu polega na tym, że w przypadku (a) nie można stosować wzoru Greena, bo P, Q **nie** są różniczkowalne w punkcie $(0,0) \in B(0,1)$!

7.5.1 Inna postać wzoru Greena

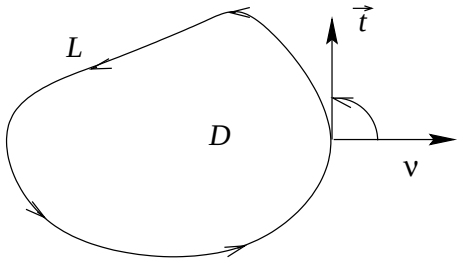
W myśl wzoru (16) mamy

$$\int_L Pdy - Qdx = \int_D \left(\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y}\right) dxdy \tag{17}$$

dla $\partial D = L$. W myśl definicji lewa strona to

$$\int_a^b ((-Q, P), \frac{\gamma'}{\|\gamma'\|})\|\gamma'\| \, ds = I.$$

Zauważmy, że wektor $(-Q, P)$ to skutek obrotu wektora (P, Q) o $\frac{\pi}{2}$, jeśli zatem $\overrightarrow{t}$ jest wektorem stycznym zgodnym z orientacją konturu L, to $((-Q, P), \overrightarrow{t}) = ((P, Q), \nu)$, gdzie wektor ν obrócony o $\frac{\pi}{2}$ przechodzi na $\overrightarrow{t}$. Otrzymana równość wynika z właściwości iloczynu skalarnego. Będziemy o tym jeszcze mówić.



Rys. 5. Wektory $\vec{t}$ i ν.

Zauważmy, że ν jest wektorem prostopadłym do L, pokazującym na zewnątrz obszaru D. Zatem

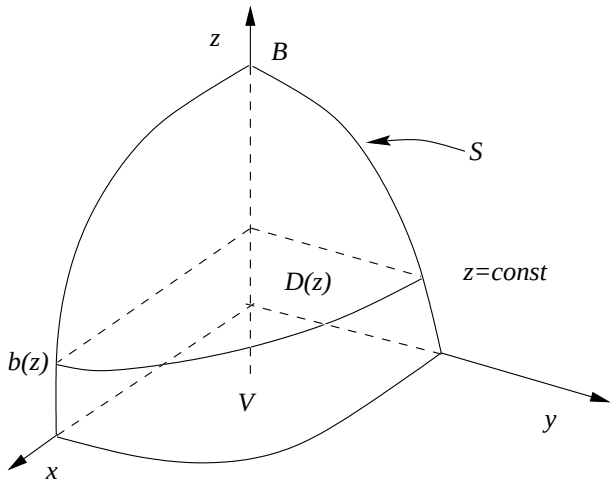
$$I = \int_a^b ((P,Q),\nu)\|\gamma'\|ds,$$

bo $\vec{t} = \gamma'/\|\gamma'\|$. Ostatecznie (17) przyjmuje postać

$$\int_L ((P,Q),\nu)dl = \int_D (\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y})\, dxdy. \tag{18}$$

7.6 Wzór Gaussa-Ostrogradskiego

Okazuje się, że wzór (18) ma postać dogodną do uogólnienia na przypadek trójwymiarowy.



Rys. 6. Obszar V.

Niech brzeg ∂V zbioru otwartego V będzie sumą $\partial V = S \cup D_x \cup D_y \cup D_z$, gdzie $D_x \subset \{x = 0\}, D_y \subset \{y = 0\}$ i $D_z \subset \{z = 0\}$, (patrz rys. 6). Zakładamy też, że S jest wykresem tj. $S = \Gamma(f)$. Niech będzie dane w V pole wektorowe (P,Q,R). Rozpatrzmy w płaszczyźnie $\Pi(z) = \{(x,y,z) : x,y \in \mathbb{R}\}$ zbiór $D(z) := V \cap \Pi(z)$, a w nim pole wektorowe (P,Q). Brzeg $D(z)$ jest konturem, można więc zastosować wzór (18). Dostaniemy wtedy

$$\int_{\partial D(z)} ((P,Q),\nu)dl = \int_{D(z)} (\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y})dxdy, \tag{19}$$

gdzie ν jest wektorem prostopadłym do $\partial D(z)$ wskazującym na zewnątrz $D(z)$. Możemy scałkować (19) względem z:

$$\int_0^B \int_{\partial D(z)} ((P,Q),\nu) dl dz = \int_D (\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y}) dx dy dz. \tag{20}$$

Prawa strona jest już w postaci ostatecznej, zajmiemy się tylko lewą stroną. Zauważmy, że dzięki naszym założeniom funkcja $x \to (x, \psi(x,z), z) =: \gamma(x)$ jest parametryzacją krzywej $S \cap \partial D(z)$. Liczymy, $\gamma'(x) = (1, \frac{\partial \psi}{\partial x}, 0)$. Gdy obrócimy wektor $\gamma'(x)$ o $\frac{\pi}{2}$ w płaszczyźnie $\Pi(z)$, to dostaniemy $(-\frac{\partial \psi}{\partial x}, 1, 0)$, dlatego

$$\nu = \left(-\frac{\partial \psi}{\partial x}, 1, 0\right) \Big/ \sqrt{1 + (\frac{\partial \psi}{\partial x})^2}$$

i jeśli $\partial D(z) \cap \{x \geq 0, y = 0\} = [0, b(z)]$, to mamy

$$\begin{aligned} \int_0^B \int_{S \cap \partial D(z)} ((P,Q),\nu) dl dz &= \int_0^B \int_0^{b(z)} ((P,Q), (-\frac{\partial \psi}{\partial x}, 1)) dx dz \\ &= \int_0^B \int_0^{b(z)} \frac{((P,Q), (-\frac{\partial \psi}{\partial x}, 1)) \sqrt{1 + (\frac{\partial \psi}{\partial x})^2 + (\frac{\partial \psi}{\partial z})^2}}{\sqrt{1 + (\frac{\partial \psi}{\partial x})^2 + (\frac{\partial \psi}{\partial z})^2}} dx dz \\ &=: J \end{aligned}$$

Pamiętamy, że skoro $\partial D(z) \cap S$ jest wykresem, to jest poziomicą tzn. $\partial D(z) \cap S = \varphi^{-1}(\{0\}) \cap \Pi(z)$ gdzie $\varphi(x,y,z) := y - \psi(x,z)$. W myśl tej definicji, jeśli punkt $p = (x,y,z) \in \Pi(z)$ spełnia $\varphi(p) > 0$, to leży poza $D(z)$, zaś $\varphi(p) < 0$ oznacza, że $p \in D(z)$, zatem wektor $\nabla\varphi(p)$ pokazuje na zewnątrz $D(z)$, gdy $p \in \partial D(z)$. Liczymy,

$$\nabla\varphi(p) = \left(-\frac{\partial \psi}{\partial x}, 1, -\frac{\partial \psi}{\partial z}\right).$$

Co więcej, $\mathbf{n} = \nabla\varphi / ||\nabla\varphi||$ jest wektorem prostopadłym do S wskazującym zewnętrzne V. Łatwo jest sprawdzić, że $|D\varphi| = ||\nabla\varphi||$. Konkretny przypadek był szczegółowo rozpatrzony w przykładzie 7 z §7.3. W myśl definicji całki powierzchniowej możemy zatem napisać,

$$J = \int_0^B \int_0^{b(z)} ((P,Q,0), \mathbf{n}) |D\varphi| dx dz = \int_S ((P,Q,0), \mathbf{n}) dS.$$

Na koniec zauważmy, że dla $p \in D_x \cup D_y$ wektor ν prostopadły do $\partial D(z)$ pokrywa się z prostopadłym do ∂V. Zaś dla $p \in D_z$ wektor $\mathbf{n}$ prostopadły do D_z jest postaci $(0,0,-1)$, tj. iloczyn skalarny $(P,Q,0)$ i $\mathbf{n}$ jest równy zero. Możemy zatem napisać, że lewa strona wzoru (20) to całka powierzchniowa z funkcji $((P,Q,0), \mathbf{n})$, gdzie $\mathbf{n}$ jest wektorem prostopadłym do ∂V pokazującym na zewnątrz V. Mamy więc

$$\int_S ((P,Q,0), \mathbf{n}) dS = \int_V (\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y}) dx dy dz. \tag{21}$$

Wprawdzie wykazaliśmy słuszność (21) dla szczególnej postaci zbioru V, ale można uzasadnić jego prawdziwość dla dowolnego zbioru V metodą rozcinania V płaszczyznami prostopadłymi do osi układu współrzędnych i sumując wynik.

Zauważmy jeszcze, że zamieniając zmienne x na y, y na z i z na x oraz P na Q, Q na R i R na P możemy przepisać (21) w następujących postaciach

$$\int_{\partial V}((P,0,R),\mathbf{n})dS = \int_V (\frac{\partial P}{\partial x} + \frac{\partial R}{\partial z})dxdydz, \tag{22}$$

$$\int_{\partial V}((0,Q,R),\mathbf{n})dS = \int_S (\frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z})dxdydz. \tag{23}$$

Sumując wzory (21), (22), (23) i skracając wynik przez 2 dostaniemy

$$\int_{\partial V}((P,Q,R),\mathbf{n})dS = \int_V (\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z})dxdydz. \tag{24}$$

Tym samym wykazaliśmy następujący fakt.

Twierdzenie 8. (wzór Gaussa-Ostrogradskiego). Załóżmy, że $V \subset \mathbb{R}^3$ jest takim zbiorem otwartym, że ∂V jest powierzchnią; X jest polem wektorowym klasy C^1 w otwartym zbiorze $\mathcal{U} \supset V \cup \partial V$, zaś $\mathbf{n}$ jest wektorem prostopadłym do ∂V (tj. $\mathbf{n}(x) \in N_x\partial V$) o długości 1, pokazującym na zewnątrz V. Wtedy

$$\int_{\partial V}(X,\mathbf{n})dS = \int_V \text{div}\, X dy_1 dy_2 dy_3,$$

gdzie $\text{div}\, X = \frac{\partial X_1}{\partial y_1} + \frac{\partial X_2}{\partial y_2} + \frac{\partial X_3}{\partial y_3}$ nazywa się *dywergencją (rozbieżnością)* pola wektorowego X.

Komentując wzór Gaussa warto wspomnieć, że dla danego pola wektorowego $\mathbf{v}$, wielkość $(\mathbf{v},\mathbf{n})$, gdzie $\mathbf{n}$ jest wektorem prostopadłym do powierzchni ∂V skierowanym na zewnątrz o długości 1 nazywa się *strumieniem pola wektorowego przez powierzchnię*. W szczególnym przypadku, gdy pole wektorowe $\mathbf{v}$ spełnia $\text{div}\, \mathbf{v} = 0$, nazwiemy je *bezźródłowym*.

Zwróćmy uwagę, że wzór Gaussa zależał od wyboru wektora $\mathbf{n}$ prostopadłego do powierzchni ∂V. Od razu umówmy się nazywać wektor prostopadły do powierzchni i o długości 1 wektorem *normalnym*. Owe wymagania prowadzą do określenia.

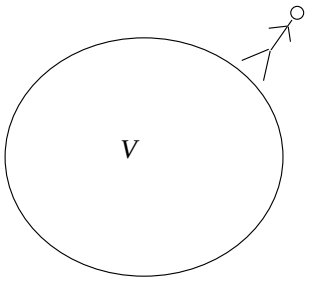
Definicja 12. Powierzchnię M wymiaru n w $\mathbb{R}^{n+1}$ nazywamy *orientowalną*, jeśli istnieje ciągłe pole wektorów normalnych, tj. jeśli istnieje ciągłe odwzorowanie $\mathbf{n} : M \to \mathbb{R}^{n+1}$, spełniające $||\mathbf{n}(x)|| = 1$ dla $x \in M$ i $\mathbf{n}(x) \in N_xM$.

Zauważmy, że jeśli M jest orientowalne, to istnieją dokładnie 2 pola, o których mowa w definicji. Wybór jednego z tych pól nazywamy *orientacją* powierzchni M.

Uwaga. Cylinder tj. $M = \{(x_1,x_2,x_3) \in \mathbb{R}^3 : x_3 \in (0,1), x_1^2 + x_2^2 = R^2\}$, sfera i ogólniej powierzchnie będące poziomicami są orientowalne, bo intuicyjnie rzecz ujmując są dwustronne: można pomalować jedną ich stronę np. na żółto a drugą na czerwono. Natomiast wstęga Möbiusa nie ma tej właściwości. Otóż próba pomalowania jednej strony wstęgi Möbiusa na

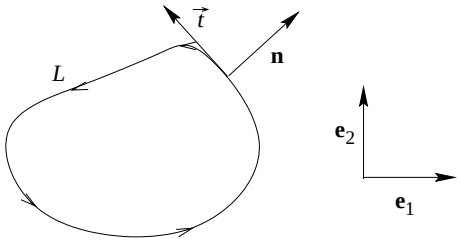
zielono kończy się tym, że cała wstęga jest zielona! Dlatego wstęga Möbiusa nie jest powierzchnią orientowalną. Wątpiących zachęcam do samodzielnej próby.

Jeśli orientowalna powierzchnia jest brzegiem zbioru ograniczonego V, to można wskazać orientację nazywaną *naturalną*. Mianowicie jest to pole wektorów normalnych wskazujący na zewnątrz V. Intuicyjnie wskazuje ją ludzik biegający po powierzchni ∂V, (patrz rys. 7).



Rys. 7. Orientacja naturalna brzegu ∂V.

Trzeba jeszcze sprawdzić, czy nowa orientacja naturalna krzywej zamkniętej (będącej powierzchnią wymiaru 1) pokrywa się ze starą definicją orientacji naturalnej krzywej zamkniętej. Otóż, załóżmy, że wektor styczny $\overrightarrow{t}$ mający długość 1 jest wybrany zgodnie ze starą definicją orientacji naturalnej. Zauważmy, że wektor normalny $\mathbf{n}$ wyznaczający nową orientację naturalną i wektor $\overrightarrow{t}$, tworzą bazę $(\mathbf{n}, \overrightarrow{t})$ w $\mathbb{R}^2$, która jest zgodna z bazą standardową $(\mathbf{e}_1, \mathbf{e}_2)$ tj. baza $(\mathbf{n}, \overrightarrow{t})$ powstaje z obrotu $(\mathbf{e}_1, \mathbf{e}_2)$ o pewien kąt.



Rys. 8. Orientacja naturalna krzywej L.

7.6.1 Przykład zastosowania wzoru Gaussa w fizyce

Niech $\varrho(x,t)$ będzie gęstością cieczy w naczyniu w punkcie x, w chwili t, zaś niech $\mathbf{v}(x,t)$ będzie jej prędkością w punkcie $x \in \mathbb{R}^3$. Chcemy ustalić, jaki jest związek pomiędzy $\mathbf{v}$ a ϱ, jeśli przyjmiemy prawo zachowania masy. Niech $V \subset \mathbb{R}^3$ będzie dowolnym zbiorem takim, że ∂V jest powierzchnią, zaś $\mathbf{n}$ jest wektorem normalnym zewnętrznym do ∂V. Zauważmy, że masa cieczy w V, to

$$m(t) = \int_V \varrho(x,t)dt.$$

Jej szybkość zmiany w jednostce czasu to $\frac{d}{dt}m(t)$. Można wykazać, że dla ϱ będącego klasy C^1 mamy

$$\frac{d}{dt}m(t) = \int_V \frac{\partial}{\partial t}\varrho(x,t)dt$$

Z drugiej strony zmiana $m(t)$ w jednostce czasu, to ilość cieczy przepływającej przez powierzchnię w jednostce czasu, albo minus strumień cieczy przez powierzchnię ∂V:

$$\frac{dm}{dt}(t) = -\int_{\partial V} (\mathbf{v}\varrho, \mathbf{n}) dS = -\int_V \text{div}\,(\mathbf{v}\varrho) dx,$$

gdzie równość wynika ze wzoru Gaussa. Łącząc obie równości dostaniemy

$$\int_V [\frac{\partial \varrho}{\partial t}(x,t) + \text{div}\,(\varrho \mathbf{v}(x,t))] dx = 0$$

Dzięki temu, że zbiór V był dowolny dostaniemy, że

$$\frac{\partial}{\partial t}\varrho + \text{div}\,(\varrho \mathbf{v}) = 0$$

w rozważanym naczyniu. Uzyskana równość nosi nazwę *prawa ciągłości*.

7.7 Wzór Stokesa

Tytułowy wzór jest jeszcze jednym uogólnieniem podstawowego wzoru rachunku różniczkowego i całkowego. Zajmiemy się nową interpretacją całki powierzchniowej jako *strumienia pola* przez powierzchnię S.

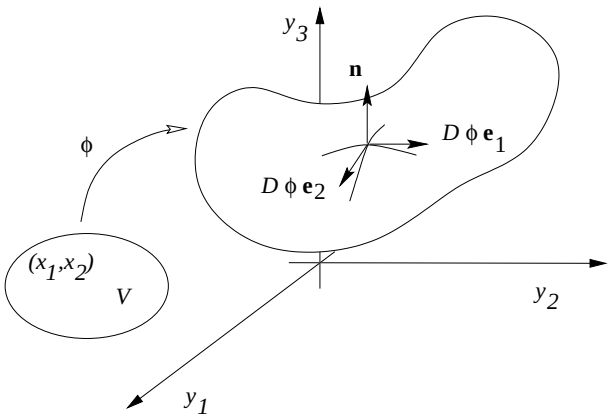
Do przedstawienia zasadniczego tematu potrzebna nam będzie następująca:
Dygresja na temat orientowania i orientacji p.w. Niech V będzie p.w. dajmy na to, $V = \mathbb{R}^k$. Jeżeli mamy pewną jej bazę $e_1, ..., e_k$, to możemy ustalić jej porządek $(e_1, e_2, ..., e_k)$. Powiemy, że baza $(e_1, e_2, ..., e_k)$ jest dodatnio zorientowana jeśli $\det(e_1, e_2, ..., e_k) > 0$. Powiemy, baza $(f_1, ..., f_2)$ jest zgodnie zorientowana z $(e_1, ..., e_k)$ jeśli macierz A przekształcenia liniowego $Ae_i = f_i$, $i = 1, \ldots, k$ ma dodatni wyznacznik. Zauważmy, że układ w $\mathbb{R}^3$ powstały przez cykliczną zamianę zmiennych, tj.

$$\mathbf{e}_1 \leftarrow \mathbf{e}_2, \quad \mathbf{e}_2 \leftarrow \mathbf{e}_3, \quad \mathbf{e}_3 \leftarrow \mathbf{e}_1$$

jest dodatnio zorientowany, bo

$$\det(\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_1) = -\det(\mathbf{e}_2, \mathbf{e}_1, \mathbf{e}_3) = \det(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) = 1.$$

Po tej dygresji przechodzimy do zasadniczego tematu. Będziemy zakładali, że powierzchnia dwuwymiarowa S ma parametryzację φ, tj. $\varphi : V \to \mathbb{R}^3$, $\varphi(V) = S$, $V \subset \mathbb{R}^2$. Na rysunku 9. $\mathbf{n}$ jest wektorem normalnym, wyznaczającym dodatnią orientację S, wektory $\frac{\partial \varphi}{\partial x_1}$, $\frac{\partial \varphi}{\partial x_2}$ są styczne do S.



Rys. 9. Powierzchnia S i jej parametryzacja ϕ.

Wektor $\frac{\partial\varphi}{\partial x_1} \times \frac{\partial\varphi}{\partial x_2}$ jest normalny do S, a więc jest on proporcjonalny do $\mathbf{n}$. Natomiast układ wektorów $\frac{\partial\varphi}{\partial x_1}$, $\frac{\partial\varphi}{\partial x_2}$, $\frac{\partial\varphi}{\partial x_1} \times \frac{\partial\varphi}{\partial x_2}$ jest zawsze dodatnio zorientowany. Wynika to ze wzoru (2.21), dla dowolnych v, w mamy

$$\det(v, w, v \times w) = \det(v \times w, v, w) = \|v \times w\|^2 > 0.$$

Załóżmy, że $\mathbf{n} = \frac{\partial\varphi}{\partial x_1} \times \frac{\partial\varphi}{\partial x_2} / \| \frac{\partial\varphi}{\partial x_1} \times \frac{\partial\varphi}{\partial x_2} \|$. Za chwilę to wykorzystamy. Na mocy definicji całki powierzchniowej mamy,

$$\int_S (\mathbf{v}, \mathbf{n})\, dS = \int_V (\mathbf{v}, \mathbf{n}) |D\varphi|\, dx_1 dx_2. \tag{25}$$

Chcemy uważnie przyjrzeć się prawej stronie tego wzoru. Wyznaczymy najpierw $\mathbf{n}$. Z definicji iloczynu wektorowego (patrz §2.5.3) $\frac{\partial\varphi}{\partial x_1} \times \frac{\partial\varphi}{\partial x_2} = (A, -B, C)$, gdzie

$$A = \det \begin{vmatrix} \frac{\partial\varphi_2}{\partial x_1} & \frac{\partial\varphi_3}{\partial x_1} \\ \frac{\partial\varphi_2}{\partial x_2} & \frac{\partial\varphi_3}{\partial x_2} \end{vmatrix}, \quad B = \det \begin{vmatrix} \frac{\partial\varphi_1}{\partial x_1} & \frac{\partial\varphi_3}{\partial x_1} \\ \frac{\partial\varphi_1}{\partial x_2} & \frac{\partial\varphi_3}{\partial x_2} \end{vmatrix}, \quad C = \det \begin{vmatrix} \frac{\partial\varphi_1}{\partial x_1} & \frac{\partial\varphi_2}{\partial x_1} \\ \frac{\partial\varphi_1}{\partial x_2} & \frac{\partial\varphi_2}{\partial x_2} \end{vmatrix}. \tag{26}$$

Z założenia o wybranej orientacji mamy, $\mathbf{n} = (A, -B, C)/\sqrt{A^2 + B^2 + C^2}$. Pamiętamy też, że $||(A, -B, C)|| = |D\varphi|$. Dostaniemy więc,

$$(\mathbf{v}, \mathbf{n})|D\varphi| = \left(\mathbf{v}, \frac{\partial\varphi}{\partial x_1} \times \frac{\partial\varphi}{\partial x_2}\right).$$

Tym samym prawa strona (25) przyjmie postać

$$\int_V (v_1 A - v_2 B + v_3 C) dx_1 dx_2 = \int_V \det\left(\mathbf{v}, \frac{\partial\varphi}{\partial x_1}, \frac{\partial\varphi}{\partial x_2}\right) dx_1 dx_2, \tag{27}$$

gdzie ponownie wykorzystaliśmy wzór (2.21).

Zauważmy też, że zgodnie z (26) $|A|$ jest polem rzutu równoległoboku $R(\frac{\partial\varphi}{\partial x_1}, \frac{\partial\varphi}{\partial x_2})$ na płaszczyznę y_2, y_3; $|B|$ jest polem rzutu tego równoległoboku na płaszczyznę y_1, y_3, zaś $|C|$ to pole

rzutu na płaszczyznę y_1, y_2. Musimy jeszcze rozeznać się w znakach. W tym celu pozwolimy sobie teraz na dygresję na temat znaków przed A, B, C.

Zauważmy, że zgodnie z wyborem $\mathbf{n}$ tak jak na rysunku 9. mamy $\mathbf{n} = (n_1, n_2, n_3)$ oraz $n_i > 0$, $i = 1, 2, 3$. Rozpatrzmy czworościan T ograniczony płaszczyznami $y_1 = 0$, $y_2 = 0$, $y_3 = 0$ oraz płaszczyzną Π prostopadłą do $\mathbf{n}$ przechodzącą przez punkt p powierzchni, wtedy $\mathbf{n} = \mathbf{n}(p)$. Można parametryzować $R \subset \Pi$ za pomocą zmiennych y_2 i y_3 z płaszczyzny $y_1 = 0$. Wtedy wektory $\mathbf{n}$, $\mathbf{e}_2$, $\mathbf{e}_3$ wyznaczają dodatnią orientaję T, dlatego przed A stoi znak $+$.

W przypadku rzutu R na płaszczyzne $y_2 = 0$ parametryzujemy R za pomocą y_1 i y_3. Wektory $\mathbf{n}$, $\mathbf{e}_1$, $\mathbf{e}_3$ są ujemnie zorientowane, dlatego przed B stoi znak $-$. Podobny argument prowadzi do znaku $+$ przed C.

Mając w pamięci subtelności związane ze znakami możemy powiedzieć, że

$$\int_V v_1 A dx_1 dx_2$$

jest całką po rzucie S na płaszczyznę $0y_2y_2$, tj. $y_1 = 0$, co uzasadnia nową definicję

$$\int_S v_1 dy_2 dy_3 := \int_S v_1 n_1 dS \equiv \int_V v_1 A dx_1 dx_2$$

gdzie $\mathbf{n} = (n_1, n_2, n_3)$ jest wektorem normalnym zewnętrznym do S.

Podobnie możemy położyć

$$\int_S v_2 dy_3 dy_1 \equiv -\int_S v_2 dy_1 dy_3 := \int_S v_2 n_2 dS \equiv \int_V -v_2 B dx_1 dx_2,$$

gdzie zmiana znaku wiąże się ze zmianą kolejności wektorów $\mathbf{e}_3$ i $\mathbf{e}_1$. Wreszcie,

$$\int_S v_3 dy_1 dy_2 := \int_S v_3 n_3 dS \equiv \int_V v_3 C dx_1 dx_2.$$

Po zsumowaniu dostaniemy definicję nowego obiektu.

Definicja 13. Załóżmy, że S jest powierzchnią orientowalną, $\mathbf{v}$ jest ciągłym polem wektorowym na S, zaś $\mathbf{n}$ jest ciągłym polem wektorów normalnych do S o długości 1, wyznaczającym orientację S. Wtedy całkę

$$\int_S v_1 dy_2 dy_3 + v_2 dy_3 dy_1 + v_3 dy_1 dy_2 := \int_S (\mathbf{v}, \mathbf{n}) dS \tag{28}$$

nazywamy *całką pola* $\mathbf{v}$ *po powierzchni zorientowanej* (albo *całką drugiego rodzaju*). Inna nazwa to *całką 2-formy po powierzchni*, gdzie *2-formą* nazwiemy wyrażenie

$$v_1 dy_2 dy_3 + v_2 dy_3 dy_1 + v_3 dy_1 dy_2.$$

Zauważmy, że wskaźniki po lewej stronie wzoru (28) są cyklicznie zamieniane: $1 \leftarrow 2 \leftarrow 3$.

Uwagi.

(1)Wzór (28) wymaga jedynie by S była powierzchnią orientowalną, S nie musi być brzegiem.

(2) Wzór (27) daje praktyczną metodę obliczania całek z 2-form.
(3) Wzór Gaussa można przepisać w następujący sposób:

$$\int_V \operatorname{div} \mathbf{v} = \int_{\partial V} v_1 dy_2 dy_3 + v_2 dy_3 dy + v_3 dy_1 dy_2.$$

Jednak nie jest naszym celem przepisywanie starych wzorów, lecz rozpatrywanie nowych sytuacji i idei. Nim to zrobimy policzymy 2 przykłady.

Przykład 12. Niech E będzie elipsoidą $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$ zorientowaną w sposób naturalny. V jest obszarem ograniczonym E, tj. $V = \{(x, y, z) : \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} \leq 1\}$. Obliczyć

$$(a) \int_E z dx dy, \qquad (b) \int_E z^2 dx dy.$$

W obu przypadkach można zastosować wzór Gaussa. Zrobimy to tylko w (a)

$$\int_E z dx dy = \int_V \frac{\partial z}{\partial z} dx dy dz = \mu_3(V).$$

Dla porządku obliczymy $\mu_3(V)$. Zauważmy, że $V = \varphi(B(0,1))$, gdzie $\varphi(\bar{x}, \bar{y}, \bar{z}) = (a\bar{x}, b\bar{y}, c\bar{z})$. Zatem

$$\mu_3(V) = \int_{B(0,1)} |D\varphi| d\bar{x} d\bar{y} d\bar{z} = abc\, \mu_3(B(0,1)) = \frac{4\pi}{3} abc.$$

Zajmijmy się punktem (b). Niech $\varphi : U \to E$ będzie parametryzacją zbioru $E^+ = E \cap \{z > 0\}$. Mamy wtedy, że $E^- = E \cap \{z < 0\}$ jest sparametryzowane za pomocą $\bar{\varphi} : U \to E$, gdzie $\bar{\varphi}$ jest odbiciem symetrycznym φ względem płaszczyzny $\{z = 0\}$, tj. $\bar{\varphi}(x, y) = -\varphi(x, y)$. Dostaniemy wtedy

$$\begin{aligned} \int_E z^2 dx dy &= \int_{E^+} z^2 dx dy + \int_{E^-} z^2 dx dy \\ &= \int_U z^2(\varphi) C(\varphi(u,v)) du dv - \int_U z^2(\bar{\varphi}) C(\bar{\varphi}(u,v)) du dv. \end{aligned}$$

Znak minus przed drugą całką bierze się stąd, że E^- jest odmienie zorientowane niż E^+. Dalej,

$$\int_E z^2 dx dy = \int_U z^2(\varphi)(C(\varphi(u,v)) - C(\varphi(u,v))) du dv = 0,$$

gdzie $C = C(\varphi(u,v))$ jest dane wzorem (26).

Wróćmy do nowych pomysłów związanych z całkami po powierzchniach zorientowanych. Chodzi nam o to, że taka np. półsfera $S^+ = \{(x_1, x_2, x_3) : x_1^2 + x_2^2 + x_3^2 = R^2, x_3 \geq 0\}$ **nie** jest powierzchnią w rozumieniu definicji 5 z §7.2, bo w żadnym otoczeniu punktu $(R, 0, 0)$ nie istnieje parametryzacja S^+. Przypominamy, że oczywiście $S^+ \cap \{x_3 > 0\}$ **jest** powierzchnią. Kłopoty łączą się z punktami, które chcielibyśmy nazwać brzegiem S^+. Dlatego wprowadzimy nowe pojęcie.

Definicja 14. Powiemy, że zbiór $S \subset \mathbb{R}^{n+1}$ jest *powierzchnią wymiaru n z brzegiem*, jeśli

$$S = f^{-1}(c) \cap g_1^{-1}((\infty, c_1]) \cap ... \cap g_k^{-1}((-\infty, c_k])$$

gdzie $f, g_i : \mathbb{R}^{n+1} \to \mathbb{R}$ są funkcjami klasy C^1, $i = 1, \ldots, k$, spełniającymi:
(i) c jest wartością regularną funkcji f;
(ii) $g_i^{-1}(c_i) \cap g_j^{-1}(c_j) \cap S = \emptyset$;
(iii) dla $p \in g_i^{-1}(c_i) \cap S$ wektory $\nabla f(p), \nabla g_i$ są lnz.
Brzegiem ∂S powierzchni S nazwiemy zbiór

$$\partial S = S \cap \bigcup_{i\equiv 1}^{k} g_i^{-1}(c_i).$$

Wnętrzem $\overset{o}{S}$ nazwiemy zbiór $S \backslash \partial S$.

Uwagi.
(a) Warunek (iii) automatycznie zapewnia (patrz twierdzenie 5), że brzeg ∂S jest powierzchnią wymiaru $n - 1$.
(b) Powyższa definicja jest zawężająca w tym sensie, że wstęga Möbiusa nie jest powierzchnią z brzegiem.

Przykład 13. Cylinder $C = S^1 \times [0, 1]$, gdzie $S^1 = \{(x_1, x_2) : x_1^2 + x_2^2 = 1\}$ jest powierzchnią z brzegiem. Mamy $C = \{(x_1, x_2, x_3) : x_1^2 + x_2^2 = 1, 0 \leq x_3 \leq 1\}$. Wtedy $\partial C = S^1 \times \{0\} \cup S^1 \times \{1\}$.

Zajmiemy się teraz przestrzeniami stycznymi do powierzchni z brzegiem S. Mamy kłopot tylko wtedy, gdy $p \in \partial S$, tj. $g_i(p) = c_i$ dla pewnego i. Wtedy kładziemy

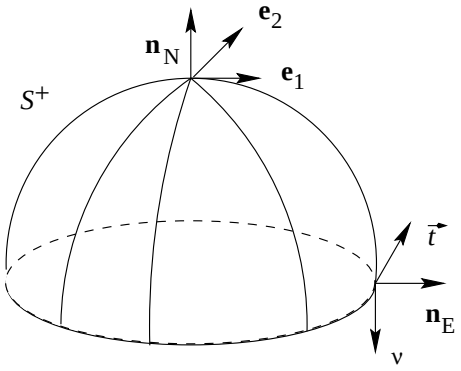
$$T_p S = \{v \in \mathbb{R}^{n+1} : (\nabla f(p); v) = 0\}.$$

Niech $v \in T_p S, p \in \partial S$, powiemy, że:

- v jest *skierowany na zewnątrz*, jeśli $(\nabla g_i(p), v) > 0$;
- v jest *skierowany do wewnątrz*, jeśli $(\nabla g_i(p), v) < 0$;
- v jest *styczny do brzegu*, jeśli $(\nabla g_i(p), v) = 0$, ich zbiór to $T_p(\partial S)$;
- v jest *normalny do brzegu*, jeśli $(v, w) = 0$, dla każdego wektora w stycznego do brzegu.

Zgodnie z naszą definicją powierzchni z brzegiem jest ona orientowalna tak jak i jej brzeg. Powstaje kwestia wyboru orientacji brzegu, przy zadanej orientacji powierzchni, tak aby były one „zgodne" . Zajmiemy się tylko przypadkiem $S \subset \mathbb{R}^3$ lub $S \subset \mathbb{R}^2$ i S ma wymiar 2. Tym samym ∂S ma wymiar 1 i trzeba wskazać kierunek obiegu ∂S, tj. wektor $\vec{t}(x) \in T_x(\partial S)$ o długości 1, gdy mamy zadany wektor $\mathbf{n}$ normalny do S. Mianowicie w punktach $p \in \partial S$ wybieramy wektor ν styczny do S i normalny do ∂S, tak aby wektory $(\mathbf{n}, \nu, \overrightarrow{t})$ były zgodnie zorientowane z $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, tj. $\det(\mathbf{n}, \nu, \overrightarrow{t}) > 0$.

Przykłady 14. **(a)**

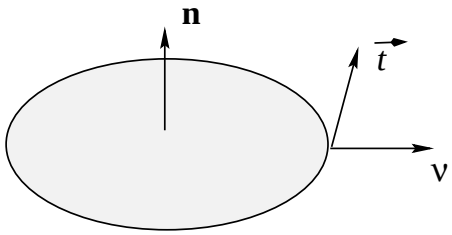


Rys. 10. Zgodnie zorientowane układy wektorów.

Układ $(\mathbf{n}_N, \mathbf{e}_1, \mathbf{e}_2)$ jest zorientowany zgodnie z $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, bo $\mathbf{n}_N = \mathbf{e}_3$ i $\det(\mathbf{e}_3, \mathbf{e}_1, \mathbf{e}_2) = 1 > 0$. Układ ten po obróceniu daje $(\mathbf{n}_E, \nu, \overrightarrow{t}) = (\mathbf{e}_1, -\mathbf{e}_3, \mathbf{e}_2)$ i $\det(\mathbf{e}_1, -\mathbf{e}_3, \mathbf{e}_1) = \det(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) > 0$.

Widzimy, więc że naturalne zorientowanie półsfery (odziedziczone po naturalnym zorientowaniu sfery) prowadzi do naturalnego zorientowania okręgu będącego brzegiem S^+.

(b)

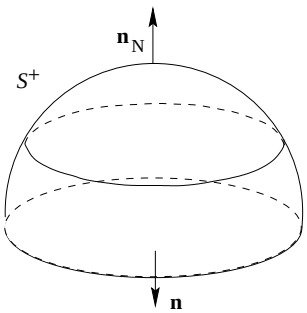


Rys. 11. Zgodne orientacje okręgu i koła.

Koło na płaszczyźnie orientujemy wybierając wektor $\mathbf{n}$ normalny (w $\mathbb{R}^3$) jak na rysunku 11.

Wybór $\overrightarrow{t}$ jak na rysunku prowadzi do zgodnej orientacji, bo $\mathbf{n} = \mathbf{e}_3, \nu = \mathbf{e}_1, \overrightarrow{t} = \mathbf{e}_2$ i $\det(\mathbf{n}, \nu, \overrightarrow{t}) = \det(\mathbf{e}_3, \mathbf{e}_1, \mathbf{e}_2) > 0$.

(c)



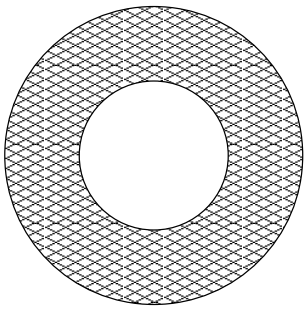
Rys. 12. Zgodna orientacja sumy powierzchni.

Niech $V = \{(x_1, x_2, x_3) \in \mathbb{R}^3,\ x_3 \geq 0,\ x_1^2 + x_2^2 + x_3^2 \leq R^2\}$, tj. V jest półkulą i $\partial V = S^+ \cup B(0, R)$. Jeśli przyjmiemy naturalną orientację ∂V, to okaże się, że orientacja spodniej części ∂V tj. koła $B(0, R)$ jest niezgodna z jego naturalną orientacją! Dlatego, mając na względzie

orientację możemy napisać $\partial V = S^+ - B(0,R)$ i jeśli ω jest 2-formą, to

$$\int_{\partial V} \omega = \int_{S^+} \omega - \int_{B(0,1)} \omega.$$

(d)



Rys. 13. Pierścienień V.

Niech $V = B(0,R)\setminus B(0,\varrho)$ będzie pierścieniem, ($R > \varrho$). Wtedy wektor normalny zewnętrzny do V w punkcie $x \in \partial B(0,\varrho)$ jest normalnym wewnętrznym do $B(0,\varrho)$, tj. $\partial V = S(0,R) - S(0,\varrho)$ i we wzorze Greena mamy

$$\int_V d\omega = \int_{S(0,R)} \omega - \int_{S(0,\varrho)} \omega,$$

gdzie ω jest 1-formą, zaś $d\omega$ jest oznaczeniem operacji na 1-formie ω postulowanej przez wzór Greena.

Niech teraz $S \subset \mathbb{R}^3$ będzie powierzchnią z brzegiem ∂S. Dodatkowo przyjmiemy, że istnieje $\varphi : \bar{U} \to \mathbb{R}^3, U \subset \mathbb{R}^2$ i U jest zbiorem otwartym. O φ zakładamy, że jest parametryzacją $S \setminus \partial S$ i $\varphi(\partial U) = \partial S$, nadto ∂U jest konturem. Powyższe założenia są z jednej strony ograniczające, bo żądamy by istniała parametryzacja całej powierzchni S. Z drugiej strony upraszczają się nasze wywody. Niech teraz $[a,b] \ni t \to \gamma(t) \in \partial U$ będzie parametryzacją konturu ∂U, która być może nie jest różniczkowalna w skończenie wielu punktach. Definicja konturu zezwala na to. Wtedy $t \to \varphi(\gamma(t))$ jest parametryzacją ∂S. Niech będzie dana 1-forma $Pdx + Qdy + Rdz$. Zajmijmy się $\int_{\partial S} Pdx$. Z definicji dostaniemy

$$\int_{\partial S} Pdx = \int_a^b P \frac{d}{dt} x(\gamma(t))dt = I$$

gdzie przyjmiemy, że $\varphi(u,v) = (x(u,v), y(u,v), z(u,v))$. Wtedy dostaniemy

$$\int_{\partial S} Pdx = \int_a^b P\left(\frac{\partial x}{\partial u}\frac{du}{dt} + \frac{\partial x}{\partial v}\frac{dv}{dt}\right)dt = \int_{\partial U} P\left(\frac{\partial x}{\partial u}du + \frac{\partial x}{\partial v}dv\right).$$

A ze wzoru Greena zastosowanego do prawej strony mamy

$$\begin{aligned}\int_{\partial S} Pdx &= \int_U \left[-\frac{\partial}{\partial v}(P\frac{\partial x}{\partial u}) + \frac{\partial}{\partial u}(P\frac{\partial x}{\partial v})\right]dudv \\ &= \int_U \left(-\frac{\partial P}{\partial v}\frac{\partial x}{\partial u} + \frac{\partial P}{\partial u}\frac{\partial x}{\partial v}\right)dudv.\end{aligned}$$

Zastosowanie wzoru na pochodną złożenia da nam

$$\int_{\partial S} Pdx = \int_U \left[-\frac{\partial x}{\partial u}\left(\frac{\partial P}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial P}{\partial y}\frac{\partial y}{\partial v} + \frac{\partial P}{\partial z}\frac{\partial z}{\partial v}\right) + \frac{\partial x}{\partial v}\left(\frac{\partial P}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial P}{\partial y}\frac{\partial y}{\partial u} + \frac{\partial P}{\partial z}\frac{\partial z}{\partial u}\right)\right] dudv.$$

Po redukcjach dostaniemy,

$$\begin{aligned}\int_{\partial S} Pdx &= \int_U \frac{\partial P}{\partial y}\left(\frac{\partial y}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial x}{\partial u}\frac{\partial y}{\partial v}\right) dudv + \int_U \frac{\partial P}{\partial z}\left(\frac{\partial z}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial x}{\partial u}\frac{\partial z}{\partial v}\right) dudv \\ &= \int_U \left[\frac{\partial P}{\partial y}\det\begin{vmatrix} \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \\ \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v}\end{vmatrix} + \frac{\partial P}{\partial z}\det\begin{vmatrix} \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} \\ \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v}\end{vmatrix}\right] dudv.\end{aligned}$$

Prawą stronę można zinterpretować jako całkę zorientowaną, zatem

$$\int_{\partial S} Pdx = \int_S -\frac{\partial P}{\partial y}dxdy + \frac{\partial P}{\partial z}dzdx.$$

Następnie cykliczna zamiana zmiennych $x \leftarrow y \leftarrow z$ prowadzi do 2 dalszych wzorów,

$$\int_{\partial S} Qdy = \int_S -\frac{\partial Q}{\partial z}dydz + \frac{\partial Q}{\partial x}dxdy, \qquad \int_{\partial S} Rdz = \int_S -\frac{\partial R}{\partial x}dzdx + \int_S \frac{\partial R}{\partial y}dydz.$$

Po zsumowaniu tych trzech wzorów dostaniemy:

Twierdzenie 9. (wzór Stokesa) Załóżmy, że S jest powierzchnią z brzegiem a P, Q, R są klasy C^1 w otwartym zbiorze $\mathcal{U} \supset S$, wtedy

$$\int_{\partial S} Pdx + Qdy + Rdz = \int_S (\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y})dxdy + (\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z})dydz + \int_S (\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x})dzdx. \quad (29)$$

Został on wyprowadzony dla pewnej szczególnej powierzchni S. Można uwolnić się od tego założenia odpowiednio rozcinając dowolną zadaną powierzchnię S płaszczynami prostopadłymi do osi układu, aby dostać kawałki do których można zastosować powyższą analizę. Otrzymane przyczynki sumujemy. Szczegóły w trzecim tomie książki Fichtenholza.

Uwaga mnemotechniczna. Uzyskany wzór jest uogólnieniem wzoru Greena. W przypadku wątpliwości związanych ze znakiem należy pamiętać, że nowy wzór ma się zgadzać ze starym, tj. pomijając jedną z liter P, Q, R i posługując się cykliczną zamianą zmiennych $x \to y \to z$ należy dostać właśnie wzór Greena.

7.7.1 Operacje analizy wektorowej

Jeśli lewa strona (29) to całka z 1-formy $\alpha = Pdx + Qdy + Rdz$, to zapiszmy (29) następująco

$$\int_{\partial S} \alpha = \int_S d\alpha$$

gdzie za definicję $d\alpha$ należy przyjąć prawą stronę wzoru Stokesa, tj.

$$d\alpha = (\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y})dxdy + (\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z})dydz + (\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x})dzdx.$$

Sytuacja jest następująca: jeśli mamy dane pole wektorowe $\overrightarrow{F}$, to możemy je utożsamiać z pewną 1-formą F^*. Dzięki wzorowi Stokesa możemy jej przypisać 2-formę dF^*. Z definicji całka z 2-formy jest całką powierzchniową strumienia przez powierzchnię pewnego pola wektorowego, które oznaczymy $\mathrm{rot}\, F$ i nazwiemy *rotacją pola* F. Mamy ostatecznie

$$\int_{\partial S} F^*\, dS = \int_S (\mathrm{rot}\, F, \mathbf{n})\, dS. \tag{30}$$

Porównanie prawych stron (29) i (30) prowadzi do następujących wzorów na współrzędne rotacj:

$$\mathrm{rot}\,(P, Q, R) = \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z},\ \frac{\partial P}{\partial z} - \frac{\partial R}{\partial x},\ \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right).$$

Zauważmy, że po lewej stronie (30) ∂S oznacza kontur. Całkę po lewej stronie (30) nazywa się czasami *krążeniem* albo *cyrkulacją pola*. Wzór Stokesa (30) mówi, że krążenie pola jest równa całce strumienia pola rotacji przez powierzchnię.

Aby przekonać się, że rotacja pola ma związek z obrotem rozpatrzmy przykład pola

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \to F = \begin{pmatrix} x_1 \cos\varphi - x_2 \sin\varphi \\ x_1 \sin\varphi + x_2 \cos\varphi \\ x_3 \end{pmatrix}$$

Widać, że punkt (x_1, x_2, x_3) jest obrócony o kąt φ wokół osi $0x_3$. Nadto mamy

$$\mathrm{rot}\, F = \begin{pmatrix} 0 \\ 0 \\ 2\sin\varphi \end{pmatrix}$$

tj. $\mathrm{rot}\, F$ pokazuje oś obrotu i sinus kąta obrotu.

Zbierzemy teraz poznane operacje. Jeśli f oznacza funkcję a F pole wektorowe, to mamy operacje:

$f \to \nabla f$ (gradient funkcji);

$F \to \mathrm{rot}\, F$ (rotacja pola); piszemy czasami $\mathrm{rot}\, F \equiv \nabla \times F$;

$F \to \mathrm{div}\, F$ (dywergencja pola).

Nową operacją jest laplasjan funkcji:

$\Delta f := \mathrm{div}\,(\nabla f) \equiv \sum_{i=1}^{3} \frac{\partial^2 f}{\partial x_1^2}$;

np. jeśli ϱ jest gęstością ładunków elektrycznych, to potencjał φ pola elektrycznego pochodzącego od ładunków ϱ spełnia równanie

$$\Delta\varphi = \varrho.$$

Rozdział 8

Przestrzenie Hilberta

Naszym celem jest przedstawienie podstawowych metod związanych z p.w. wyposażonych w iloczyn skalarny. Mamy na uwadze nie tylko zastosowania algebraiczne, np. do badania wartości i wektorów własnych, ale i analityczne jak przestrzenie $L^2(\mathbb{R}^n)$, które są przydatne m.in. do uprawiania mechaniki kwantowej.

Wstępny podrozdział ma charakter ogólny i niezależny od wymiaru przestrzeni. W §8.2 jesteśmy zainteresowani przestrzeniami nieskończenie wymiarowymi a od podrozdziału 8.3 skupiamy się ponownie na przypadku skończenie wymiarowym.

8.1 Przestrzenie unitarne

Z grubsza rzecz ujmując, tytułowe przestrzenie to są p.w. wyposażone w iloczyn skalarny.

Definicja 1. Niech V będzie p.w. nad $\mathbb{K}$. Powiemy, że V jest *przestrzenią unitarną* (p.u.) jeśli jest zadana 2-forma liniowa $(\cdot,\cdot) : V \times V \to \mathbb{K}$, tj. funkcja o następujących właściwościach:

(i) dla dowolnych $\alpha, \beta \in \mathbb{K}$ i $v, u, w \in V$ mamy

$$(\alpha v + \beta w, u) = \alpha(v, u) + \beta(w, u);$$

(ii) dla dowolnych $v, w \in V$ mamy $(v, w) = \overline{(w, v)}$;

(iii) dla każdego $v \in V$, $(v, v) \geq 0$, nadto $(v, v) = 0$ wtedy i tylko wtedy, gdy $v = 0$.

Przykłady 1.

(1) $\mathbb{R}^n$ ze zwykłym iloczynem skalarnym jest p.u.

(2) $V = \mathbb{R}^2$, dla $a, b \in V$ kładziemy

$$(a, b)_* := a_1 b_1 + k a_2 b_2,$$

gdzie $k > 0$. Wtedy $(\cdot,\cdot)_*$ jest iloczynem skalarnym.

(3) Niech $\tilde{L}^2(a, b)$, gdzie $a, b \in \mathbb{R}$, oznacza zbiór funkcji $f : (a, b) \to \mathbb{C}$ takich, że $|f|^2$ jest całkowalna w niewłaściwym sensie Riemanna na (a, b). Wtedy wzór

$$(f, g) = \int_a^b f(x)\bar{g}(x)\, dx.$$

definiuje iloczyn skalarny.

(4) Niech $\tilde{L^2}(\mathbb{R}^n)$, oznacza zbiór funkcji $f : \mathbb{R}^n \to \mathbb{R}$ takich, że $|f|^2$ jest całkowalna w niewłaściwym sensie Riemanna na $\mathbb{R}^n$. Wtedy wzór

$$(f, g) = \int_{\mathbb{R}^n} f(x)g(x)\, dx.$$

definiuje iloczyn skalarny.

(5) Niech $G \subset \mathbb{R}^n$ będzie otwarty, wtedy $\tilde{L^2}(G)$ oznacza zbiór funkcji $f : G \to \mathbb{K}$ takich, że $|f|^2$ jest całkowalna w niewłaściwym sensie Riemanna na G. $\tilde{L^2}(G)$ jest p.u. z iloczynem skalarnym zdefiniowanym podobnie jak wyżej.

Łatwe sprawdzenie, że podane wzory zadają iloczyny skalarne, pozostawiamy czytelnikowi.

Chcemy podkreślić, że przestrzenie $\tilde{L^2}(a, b)$, $\tilde{L^2}(\mathbb{R}^n)$ i $\tilde{L^2}(G)$ są uogólnieniami przestrzeni Euklidesowych $\mathbb{R}^n$.

Ważną rolę w dalszych rozważaniach odegra pojęcie prostopadłości. Niech V będzie p.u., powiemy, że $u, w \in V$ są *prostopadłe*, jeśli $(u, w) = 0$, piszemy $u \perp w$. O układzie wektorów $\{e_k\}_{k\in I}$ (zbiór wskaźników I może być nieskończowny) powiemy, że jest *układem ortonormalnym* (piszemy u.o.) jeśli dla dowolnych wektorów z układu mamy

$$(e_i, e_k) = \begin{cases} 1, & \text{jeśli } i = k, \\ 0, & \text{jeśli } i \neq k. \end{cases}$$

Jeśli baza $\{e_k\}_{k=1}^n$ n-wymiarowej p.u. V jest u.o., to znajdowanie przedstawień dowolnego $v \in V$ w tej bazie jest łatwe. Jeśli $v = \sum_{k=1}^n c_k e_k$, to mamy

$$v - \sum_{k=1}^{n} c_k e_k = 0.$$

Mnożąc skalarnie przez dowolny wektor bazy e_i dostaniemy

$$0 = (v - \sum_{k=1}^{n} c_k e_k, e_i) = (v, e_i) - \sum_{k=1}^{n} c_k (e_k, e_i) = (v, e_i) - c_i,$$

tj.

$$v = \sum_{k=1}^{n} (v, e_i) e_i.$$

Szczególnie interesujące będzie zastosowanie tego spostrzeżenia do $V = \tilde{L^2}(0, 2\pi)$. Nim to zrobimy sformułujemy wniosek wypływający wprost z właściwości iloczynu skalarnego, był on wykazany dla $V = \mathbb{C}^n$, (patrz twierdzenie 1.26), teraz zajmiemy się ogólnym przypadkiem. Dzięki swej prostocie i ogólności jest wart przytoczenia.

Stwierdzenie 1. (Nierówność Schwarza) Niech V będzie p.u., $v, w \in V$, wtedy

$$|(v, w)| \leq \|v\| \cdot \|w\|. \tag{1}$$

Dowód. Niech $(v, w) \neq 0$, bo inaczej nie ma czego dowodzić. Wtedy,

$$\begin{aligned} 0 \leq \left\| v - \tfrac{\|v\|^2}{(w,v)} w \right\|^2 &= \left(v - \frac{\|v\|^2}{(w,v)} w, v - \frac{\|v\|^2}{(w,v)} w \right) \\ &= \|v\|^2 - \frac{(v,w)}{\overline{(w,v)}} \|v\|^2 - \|v\|^2 + \frac{\|v\|^4}{|(w,v)|^2} \|w\|^2 \\ &= \frac{\|v\|^2}{|(w,v)|^2} (|(w,v)|^2 - \|v\|^2 \|w\|^2). \end{aligned}$$

□

Jeśli mamy daną p.u. V to możemy w niej zdefiniować normę wzorem

$$\|u\| = \sqrt{(u,u)}.$$

Dzięki nierówności Schwarza łatwo jest sprawdzić, że nowy obiekt spełnia warunki normy. Pozostawiamy to Czytelnikowi.

8.2 Szeregi Fouriera

Zajmiemy się układami u.o. w p.u. takich jak $\tilde{L^2}(0, 2\pi)$, zakładamy przy tym, że badane funkcje mają wartości rzeczywiste. Przedstawimy trygonometryczny u.o. Będzie nas interesować znajdowanie współczynników wektorów w bazie, tj. współczynników Fouriera. Intuicyjnie rzecz ujmując znajdowanie współczynników Fouriera oznacza szukanie składowych harmonicznych dźwięków.

Pozostawiamy czytelnikowi sprawdzenie następujących równości dla dowolnych liczb naturalnych n i k:

$$\int_0^{2\pi} \sin(nx)\sin(kx)\,dx = \begin{cases} 0 & \text{gdy } n \neq k; \\ \pi & \text{gdy } n = k > 0; \end{cases} \quad \int_0^{2\pi} \sin(nx)\cos(kx)\,dx = 0;$$

$$\int_0^{2\pi} \cos(nx)\cos(kx)\,dx = \begin{cases} 0 & \text{gdy } n \neq k; \\ \pi & \text{gdy } n = k > 0; \\ 2\pi & \text{gdy } n = k = 0. \end{cases}$$

Tym samym układ wektorów $\{e_k\}_{k=0}^{\infty}$, gdzie

$$e_k = \begin{cases} \frac{1}{\sqrt{2\pi}} & \text{gdy } k = 0; \\ \frac{1}{\sqrt{\pi}} \cos(nx) & \text{gdy } k = 2n; \\ \frac{1}{\sqrt{\pi}} \sin(nx) & \text{gdy } k = 2n+1. \end{cases}$$

jest u.o. i nazywamy go *trygonometrycznym układem ortonormalnym*. Spodziewamy się, że trygonometryczny u.o. jest bazą i dla dowolnej funkcji f należącej do $\tilde{L^2}(0, 2\pi)$ mamy

$$f = \sum_{k=0}^{\infty} (f, e_k) e_k = \sum_{k=0}^{\infty} c_k e_k.$$

Zajmijmy się wyznaczeniem $(f, e_k)e_k$. Niech $k = 0$, zauważmy, że

$$(f, e_0)e_0 = \frac{1}{2\pi}\int_0^{2\pi} f(x)\,dx =: a_0.$$

Dla $k = 2n$ mamy

$$(f, e_{2n})e_{2n} = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos(nx)\,dx = a_n\cos(nx),$$

dla $k = 2n + 1$

$$(f, e_{2n+1})e_{2n+1} = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin(nx)\,dx = b_n\sin(nx),$$

gdzie położyliśmy,

$$a_n = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos(nx)\,dx, \quad b_n = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin(nx)\,dx, \qquad n \geq 1.$$

Liczby a_n, b_n nazywamy *współczynnikami Fouriera funkcji* f. Powstaje naturalne pytanie: czy istotnie

$$f(x) = a_0 + \sum_{n=1}^{\infty}(a_n\cos(nx) + b_n\sin(nx)),$$

gdzie szereg po prawej stronie nazywamy *szeregiem Fouriera*? Odpowiedzią zajmiemy się nieco później. Przedtem przedstawimy pomocnicze rozważania.

Niech V będzie dowolną p.u. nad $\mathbb{C}$ i $\{e_k\}_{k=1}^{\infty}$ niech będzie dowolnym u.o. Badamy dla wektora $v \in V$ sumę

$$s_n = \sum_{k=0}^{n} c_k e_k,$$

gdzie $c_k = (v, e_k)$. Zauważmy, że dzięki właściwościom iloczynu skalarnego dostaniemy

$$\|s_n\|^2 = \left(\sum_{i=0}^{n} c_i e_i, \sum_{k=0}^{n} c_k e_k\right) = \sum_{i=0}^{n}\sum_{k=0}^{n} c_i\bar{c_k}(e_i, e_k) = \sum_{k=0}^{n}|c_k|^2. \tag{2}$$

Ponadto,

$$(v, s_n) = \left(v, \sum_{k=0}^{n} c_k e_k\right) = \sum_{k=0}^{n}\bar{c_k}(v, e_k) = \sum_{k=0}^{n}\bar{c_k}c_k = \|s_n\|^2.$$

Wynika stąd, że wektory $v - s_n$ i s_n są prostopadłe, mamy bowiem

$$(v - s_n, s_n) = (v, s_n) - (s_n, s_n) = \|s_n\|^2 - \|s_n\|^2 = 0.$$

Zatem nierówność Schwarza (1) prowadzi do następującego wniosku:

$$\|s_n\|^2 = |(v, s_n)| \leq \|v\|\|s_n\|$$

a stąd

$$\|s_n\|^2 \leq \|v\|^2.$$

W przypadku nieskończonego u.o $\{e_k\}$ po przejściu z n do nieskończoności dostaniemy

$$\sum_{k=0}^{\infty} |c_k|^2 \leq \|v\|^2 \tag{3}$$

zwaną *nierównością Bessla*.

Uwaga. Nierówność (3) staje się równością wtedy i tylko wtedy, gdy

$$\sum_{k=0}^{\infty} c_k e_k = v,$$

albowiem na mocy powyższych rachunków i właściwości iloczynu skalarnego mamy

$$\|v\|^2 = \|v - s_n + s_n\|^2 = \|v - s_n\|^2 + \|s_n\|^2 + 2\mathrm{Re}(v - s_n, s_n) = \|v - s_n\|^2 + \|s_n\|^2$$

i żądany fakt uzyskamy po przejściu z n do nieskończoności.

Powróćmy do zadanego wcześniej pytania o zbieżność szeregów Fouriera. Odpowiedź jest zawarta w twierdzeniu poniżej.

Twierdzenie 2. Jeśli funkcja $f : \mathbb{R} \to \mathbb{R}$ jest okresowa o okresie 2π, tj. dla dowolnej liczby $x \in \mathbb{R}$ mamy $f(x+2\pi) = f(x)$ i f spełnia warunek Lipschitza ze stałą K, to jej szereg Fouriera

$$a_0 + \sum_{n=1}^{\infty} (a_n \cos(nx) + b_n \sin(nx))$$

jest zbieżny jednostajnie i jego granicą jest funkcja f.

Uwagi. Przy naszych założeniach funkcja f ograniczona do przedziału $[0, 2\pi]$ jest elementem przestrzeni $\tilde{L}^2(0, 2\pi)$.

Sformułowane przed chwilą twierdzenie nie jest łatwe. Jeśli o funkcji f założyć jedynie ciągłość i okresowość, to zagadnienie staje się bardzo trudne. Dość powidzieć, że dopiero w latach 70-tych XX wieku wykazano, że szereg Fouriera funkcji ciągłej jest do niej zbieżny, ale być może poza zbiorem miary zero. Szeregi Fouriera są badane od początku XIX wieku.

Przykłady 2. Niech funkcje (a) $f(x) = \pi^2 - (x - \pi)^2$ i (b) $g(x) = x - \pi$ będą obcięciami do przedziału $[0, 2\pi]$ funkcji okresowych o okresie 2π, wtedy

$$f(x) = \frac{5\pi^2}{6} - \sum_{n=1}^{\infty} \frac{4}{n^2} \cos(nx), \qquad g(x) = -\sum_{n=1}^{\infty} \frac{2}{n} \sin(nx)$$

Zauważmy, że funkcja g **nie** jest ciągła, ale należy do $\tilde{L}^2(0, 2\pi)$, więc jest sens mówić o jej szeregu Fouriera. Od razu widać, że jej szereg Fouriera nie będzie zbieżny jednostajnie i nie będzie zbieżny w punktach $x = 0, x = 2\pi$.

8.2.1 Przestrzenie $L^2(G)$ i całka Lebesgue'a

Chcielibyśmy zwrócić uwagę na jeszcze jeden aspekt Twierdzenia 2. A mianowicie mówiliśmy w nim o zbieżności jednostajnej funkcji, które w naturalny sposób były elementami przestrzeni $\tilde{L^2}(0, 2\pi)$. Ta p.u. jest w naturalny sposób wyposażona w metrykę daną wzorem $d(x, y) = \|x - y\|$, ale w twierdzeniu nie ma mowy o takiej zbieżności, dlaczego? Po pierwsze, zbieżność jednostajna jest w miarę przejrzysta i wiemy, że granice jednostajnie zbieżnych ciągów funkcji ciągłych są ciągłe. Po drugie nie bardzo wiemy, co oznacza odległość w $\tilde{L^2}(0, 2\pi)$. Zajmijmy się tą sprawą. Podkreślamy, że w przestrzeniach unitarnych $\mathbb{R}^n$ ciągi Cauchy'ego są zbieżne.

Przykład 3. Zbadamy pewien przykład ciągu Cauchy'ego w metryce przestrzeni $\tilde{L^2}(0, 1)$, dla uproszczenia zastępując przedział $[0, 2\pi]$ przedziałem $[0, 1]$. Niech $\{r_n\}_{n=1}^{\infty}$ oznacza zbiór wszystkich liczb wymiernych z przedziału $[0, 1]$. Kładziemy

$$f_n(x) = \chi_{\{r_1,\ldots,r_n\}}(x) + \frac{1}{2^n}.$$

Oczywiście $f_n \in \tilde{L^2}(0, 1)$ i $\|f_n\| = 2^{-n}$. Co więcej ciąg $\{f_n\}_{n=1}^{\infty}$ jest ciągiem Cauchy'ego, bo

$$\|f_n - f_k\| = |2^{-n} - 2^{-k}| \to 0, \quad \text{gdy } n, k \to \infty.$$

Łatwo jest sprawdzić, że dla dowolnego $x \in [0, 1]$ mamy

$$\lim_{n\to\infty} f_n(x) = \chi_{\mathbb{Q}\cap[0,1]}.$$

Jak wiemy $\chi_{\mathbb{Q}\cap[0,1]}$ nie jest funkcją całkowalną w sensie Riemanna, nawet w sensie niewłaściwym, dlatego nie może być elementem $\tilde{L^2}(0, 1)$. Taką sytuację, jak powyżej uznajemy za patologiczną. Chcemy, by ciągi Cauchy'ego były zbieżne. Okazuje się, że mamy chody u dobrej wróżki:

Twierdzenie 3. Niech $\{f_n\}_{n=1}^{\infty} \subset \tilde{L^2}(G)$, gdzie G jest otwartym podzbiorem $\mathbb{R}^n$, będzie ciągiem Cauchy'ego w metryce wyznaczonej przez iloczyn skalarny w $\tilde{L^2}(G)$. Wtedy istnieje taka funkcja $f : G \to \mathbb{K}$, że

$$f = \lim_{n\to\infty} f_n,$$

w tym sensie, że dla dowolnego $\varepsilon > 0$ istnieje $N_\varepsilon \in \mathbb{N}$, że dla $n > N_\varepsilon$ mamy

$$\|f - f_n\| < \varepsilon.$$

Uwagi. Granica f może nie być elementem $\tilde{L^2}(G)$, (patrz poprzedni przykład), dlatego symbol $\|f\|$ oznacza rozszerzenie dotychczasowego znaczenia. Wkrótce to wyjaśnimy. Tym samym możemy powiedzieć, że dokonaliśmy kolejnego rozszerzenia pojęcia całki Riemanna. Zbiór granic ciągów Cauchy'ego gwarantowanych w poprzednim twierdzeniu będziemu oznaczali symblem

$$L^2(G)$$

o jego elementach f będziemy mówić, że $|f|^2$ są *całkowalne w sensie Lebesgue'a*, albo w skrócie, że są całkowalne z kwadratem. Tym samym $\chi_{\mathbb{Q}\cap[0,1]} \in L^2(0,1)$.

Co gorsza, można wskazać przykład ciągu f_n zbieżnego w $L^2(G)$, który nie jest zbieżny w żadnym punkcie $x \in G$.

Trzeba jeszcze określić całki z funkcji w $L^2(G)$ i jej normy. Zrobimy to teraz. Niech $f \in L^2(G)$, wtedy kładziemy

$$\int_G f(x)dx := \lim_{n\to\infty} \int_G f_n(x)dx,$$

gdzie $\{f_n\}_{n=1}^{\infty}$ jest ciągiem Cauchy'ego, którego granicą jest f. Trzeba sprawdzić, że w/w granica jest dobrze określona. Sprawdzamy mianowicie, że

$$a_n = \int_G f_n(x)\,dx,$$

jest ciągiem Cauchy'ego. Mamy bowiem

$$|a_n - a_m| \le \left|\int_G (f_n(x) - f_m(x))\,dx\right| \le \int_G |f_n(x) - f_m(x)|\,dx.$$

Zauważmy, że prawą stronę można zapisać jako iloczyn skalarny w $\tilde{L^2}(G)$ funkcji $|f_n - f_m|$ i funkcji tożsamościowo równej 1. Z nierówności Schwarza dostaniemy

$$|a_n - a_m| \le \|f_n - f_m\|_{\tilde{L^2}(G)} \|1\|_{\tilde{L^2}(G)}.$$

Zatem $\{a_n\}$ jest ciągiem Cauchy'ego i jego granica $\int_G f(x)dx$ jest dobrze określona i nazywamy ją *całką Lebesgue'a funkcji* f.

Niech $f, g \in L^2(G)$, wtedy kładziemy

$$(f,g)_{L^2(G)} := \lim_{n\to\infty} \int_G f_n(x)g_n(x)\,dx \equiv \lim_{n\to\infty} (f_n, g_n)_{\tilde{L^2}(G)}. \tag{4}$$

Sprawdzamy, że w/w granica jest dobrze określona. Mamy

$$\begin{aligned} |(f_n, g_n) - (f_m, g_m)| &= \left|\int_G (f_n(x)g_n(x) - f_m(x)g_m(x))\,dx\right| \\ &\le \int_G |f_n(x)g_n(x) - f_m(x)g_m(x)|\,dx = P \end{aligned}$$

i dalej dzięki nierówności Schwarza dostaniemy

$$\begin{aligned} P &\le \int_G |f_n(x) - f_m(x)||g_n(x)|\,dx + \int_G |g_n(x) - g_m(x)||f_m(x)|\,dx \\ &\le \|f_n - f_m\|_{\tilde{L^2}(G)} \|g_n\|_{\tilde{L^2}(G)} + \|g_n - g_m\|_{\tilde{L^2}(G)} \|f_m\|_{\tilde{L^2}(G)} \to 0 \text{ gdy } m, n \to \infty. \end{aligned}$$

Z powyższego rachunku, wynika iż $(f_n, g_n)_{\tilde{L^2}(G)}$ jest ciągiem Cauchy'ego, tym samym jego granica istnieje.

Powyższy rachunek pokazuje też, że definicja normy w $L^2(G)$ podana niżej jest poprawna:

$$\|f\|^2 := \lim_{n\to\infty} \int_G |f_n(x)|^2\,dx.$$

Podsumowując znaczenie ciągów Cauchy'ego wprowadzimy nową definicję.

Definicja 2. Powiemy, że p.u. V jest przestrzenią Hilberta, jeśli każdy ciąg Cauchy'ego ma granicę w V.

Ważnym faktem jest następujące stwierdzenie, którego dowód pomijamy.

Stwierdzenie 4. Przestrzenie $L^2(G)$ są przestrzeniami Hilberta.

Jak się Czytelnik być może przekona w przyszłości, przestrzenie Hilberta (a zwłaszcza $L^2(\mathbb{R}^3)$) są właściwymi przestrzeniami do uprawiania mechaniki kwantowej. Nie będziemy teraz rozwijali tego tematu.

Innym ważnym przykładem przestrzeni Hilberta są przestrzenie Euklidesowe $\mathbb{R}^n$.

8.3 Przekształcenia unitarne i ortogonalne

Będziemy osobno rozważali skończenie wymiarowe p.w. V nad $\mathbb{R}$ jak i nad $\mathbb{C}$. Zaczniemy od definicji przedmiotu naszego zainteresowania.

Definicja 3. Niech V i W będą p.u. i $F : V \to W$ będzie przekształceniem liniowym spełniającym warunek:

$$(Fv, Fw)_W = (v, w)_V \quad \text{dla wszystkich } v, w \in V. \tag{5}$$

Wtedy powiemy, że

(a) F jest unitarne, jeśli V jest p.w. nad $\mathbb{C}$;

(b) F jest ortogonalne, jeśli V jest p.w. nad $\mathbb{R}$.

Definicja ta oznacza, że F zachowuje iloczyn skalarny.

Podamy teraz serię przykładów zaczynając od najprostszych.

Przykład 4. Niech $V = W = \mathbb{C}$ i $Fv = \alpha v$, gdzie $|\alpha| = 1$. Przypominamy, że $(z, w) = z\bar{w}$. Widzimy, że $(Fz, Fw) = \alpha z \bar{\alpha} \bar{w} = (z, w)$, F jest unitarne.

Przykład 5. Załóżmy, że $V = W = \mathbb{R}^2$ jest wyposażone w naturalny iloczyn skalarny. Kładziemy $(F(x_1, x_2))^T := (-x_2, x_1)^T$. Sprawdzamy, że F jest ortogonalne:

$$(Fv, Fw) = ((-v_2, v_1)^T, (-w_2, w_1)^T) = ((v_1, v_2)^T, (w_1, w_2)^T) = (v, w).$$

Macierz przekształcenia F to

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Ogólniej przekonamy się, że odwzorowanie, $F : \mathbb{R}^2 \to \mathbb{R}^2$, którego macierz to

$$\begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$

jest ortogonalne.

Przykład 6. Załóżmy, że $V = W = \mathbb{R}^3$ jest wyposażone w naturalny iloczyn skalarny. Niech odwzorowanie $F : \mathbb{R}^3 \to \mathbb{R}^3$ będzie zadane macierzą

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Wtedy F jest ortogonalne.

Przykład 7. Załóżmy, że $V = W = \mathbb{C}^n$ i $\{e_1, \ldots, e_n\}$ niech będzie dowolną bazą ortonormalną. Jak wynika z §8.2 każdy wektor $v \in V$ można zapisać w postaci

$$v = \sum_{i=1}^{n} c_i e_i.$$

Kładziemy

$$Fv = \sum_{i=1}^{n} \alpha_i c_i e_i,$$

gdzie liczby $\alpha_i \in \mathbb{C}$ spełniają $|\alpha_i| = 1$, $i = 1, \ldots, n$. Odwzorowanie F jest unitarn, gdyż z właściwości iloczynu skalarnego i bazy ortonornormalnej $\{e_1, \ldots, e_n\}$ dostaniemy, że

$$(Fv, Fw) = \sum_{i,k=1}^{n} c_i \bar{c}'_k \alpha_i \bar{\alpha}_k (e_i, e_k) = \sum_{i=1}^{n} c_i \bar{c}'_i = (v, w).$$

Zauważmy, że odwzorowania unitarne i ortogonalne są odwracalne, gdy $\dim W < \infty$. Jest tak, bo $\ker F = \{0\}$. Mianowicie, jeśli $Fv = 0$, to

$$0 = \|Fv\|^2 = (Fv, Fv) = (v, v) = \|v\|^2.$$

Wnosimy stąd, że odwzorowanie odwrotne do F zawsze istnieje.

Wybór bazy, np. bazy ortonormalnej, pozwala nam na utożsamianie p.u. V z $\mathbb{C}^n$ (odpowiednio, z $\mathbb{R}^n$) dla odpowiedniego $n > 0$ a odwzorowań liniowych z macierzami. Możemy teraz zapytać jak scharakteryzować macierze unitarne (odpowiednio, ortogonalne). Dla ustalonej bazy ortonormalnej $\{e_1, \ldots, e_n\}$ i odwzorowania unitarnego F (odpowiednio, ortogonalnego) mamy

$$(e_i, e_j) = (Fe_i, Fe_j) \equiv e_i^T F^T \bar{F} \bar{e}_j = \begin{cases} 1 & \text{gdy } i = j; \\ 0 & \text{gdy } i \neq j. \end{cases}$$

Oznacza to, że element macierzy $F^T \bar{F}$ na przecięciu i-tego wiersza i j-tej kolumny jest równy 1, gdy $i = j$ albo 0, gdy $i \neq j$. Tym samym $F^T \bar{F}$ jest macierzą tożsamościową

$$F^T \bar{F} = Id. \tag{6}$$

Jeśli najpierw weźmiemy zespolone sprzężenie obu stron (6) a następnie pomnożymy to równanie z prawej strony przez F^{-1}, to dostaniemy

$$\bar{F}^T = F^{-1}. \tag{7}$$

Oznacza to, że znajdowanie macierzy odwrotnej do macierzy unitarnej (ortogonalnej) jest bardzo proste: wystaczy policzyć sprzężenie i transponować daną macierz. W przypadku rzeczywistym sprzężenie jest niepotrzebne. Równość (7) często jest używana jako definicja macierzy unitarnej (ortogonalnej). Zbiór macierzy unitarnych (odpowiednio, ortogonalnych) n na n oznacza się symbolem $U(n)$ (odpowiednio, $O(n)$).

Odnotujmy jeszcze dwie ciekawe właściwości. Zauważmy, że jeśli obliczymy wyznacznik obu stron (6), to dostaniemy

$$1 = \det Id = \det(F^T \bar{F}) = \det F^T \det \bar{F} = \det F \overline{\det F} = |\det F|^2.$$

A jeśli dodatkowo $F \in O(n)$, to mamy $\det F = \pm 1$. Płynie stąd ważki wniosek analityczny. Jeśli wprowadzimy nowe zmienne $y = Fx$, gdzie F jest przekształceniem ortogonalnym, to wzór na zamianę zmiennych w całce mówi, że miara zbiorów się zachowa, tj. jeśli A jest zbiorem mierzalnym w sensie Jordana-Riemanna, to

$$\mu_n(FA) = \mu_n(A).$$

Zauważmy jeszcze, że zbiory $U(n)$ i $O(n)$ są wyposażone w strukturę grupy: wiemy, że mnożenie macierzy jest łączne. Elementem obojętnym mnożenia jest macierz tożsamościowa. Widzimy też, że $U(n)$ i $O(n)$ są zamknięte na operację brania elementu odwrotnego tj. macierzy odwrotnej. Wynika to wprost z definicji i wzoru (7).

8.4 Formy dwuliniowe i kwadratowe

W tym paragrafie rozważamy dla naszej wygody wyłącznie p.w. V nad $\mathbb{R}$, teoria p.w. nad $\mathbb{C}$ wymaga pewnych zmian. Będziemy też zakładali dla uproszczenia, że V jest p.u. Tracimy przy tym nieco na ogólności, ale zyskujemy na przejrzystości. Nasz cel, to wprowadzenie naturalnego uogólnienia pojęcia iloczynu skalarnego.

Definicja 4. Niech V, W będą p.w. Powiemy, że funkcja

$$B : V \times V \to W$$

jest *przekształceniem dwuliniowym*, jeśli funkcje

$$V \ni v \mapsto A_1(v) := B(v, w) \in W \text{ przy ustalonym } w$$
$$V \ni w \mapsto A_2(w) := B(v, w) \in W \text{ przy ustalonym } v$$

są liniowe. Jeśli $W = \mathbb{R}$, to przekształcenie dwuliniowe B nazywamy *formą dwuliniową*.

Oczywistym przykładem formy dwuliniowej jest iloczyn skalarny. Inny, ogólniejszy, jest podany niżej. Niech $x, y \in \mathbb{R}^n$ i A będzie macierzą n na n, kładziemy

$$B(x, y) = x^T Ay \equiv (x, Ay), \tag{8}$$

gdzie $(\cdot,\cdot)$ jest zwykłym iloczynem skalarnym. Wtedy Czytelnik łatwo sprawdzi korzystając z właściwości iloczynu skalarnego, że istotnie jest to forma dwuliniowa.

O przekształceniu dwuliniowym, w szczególności o formie, powiemy, że jest *symetryczne*, jeśli $B(v,w) = B(w,v)$ dla wszystkich $v, w \in V$.

Możemy zadać naiwne pytanie, czy są inne przykłady form dwuliniowych niż te zadawane wzorem (8) dla pewnej macierzy A. Okazuje się, że prawdziwy jest następujący fakt.

Stwierdzenie 5. Każda forma dwuliniowa B jest postaci (8), tj. istnieje taka macierz A, że

$$B(v,w) = (v, Aw).$$

Jeśli dodatkowo B jest formą symetryczną, to A jest macierzą symetryczną, tj. $A = A^T$. Co więcej, A jest wyznaczona jednoznacznie.

Definicja 5. Niech V będzie p.w. Powiemy, że funkcja $Q : V \to \mathbb{R}$ jest *formą kwadratową*, jeśli
(a) dla dowolnego $\lambda \in \mathbb{R}$ i $v \in V$, mamy $Q(\lambda v) = \lambda^2 Q(v)$;
(b) funkcja $V \times V \ni (v,w) \mapsto B(v,w) := \frac{1}{2}(Q(v+w) - Q(v) - Q(w))$ jest formą dwuliniową nazywaną *formą stowarzyszoną*.

Zauważmy, że forma dwuliniowa stowarzyszona jest koniecznie symetryczna. Skoro wiemy, jak wygląda ogólna postać formy dwuliniowej, (patrz (8)), to automatycznie dostaniemy, że każda forma kwadratwa Q jest postaci

$$Q(v) = (v, Av).$$

Zastanówmy się, co się stanie z formą dwuliniową, gdy zmienimy bazę w p.u. V, tj. zmienimy układ współrzędnych. Niech v_s będzie dowolnym wektorem. W nowym układzie współrzędnych (w nowej bazie) zapiszemy go jako

$$v_s = F v_n, \tag{9}$$

gdzie F jest nieosobliwym przekształceniem liniowym, tj. $\ker F = \{0\}$. Dla $w_s = F w_n$ mamy

$$B(v_s, w_s) = v_s^T A w_s = (F v_n)^T A (F w_n) = v_n F^T A F w_n = v_n \tilde{A} w_n,$$

gdzie

$$\tilde{A} = F^T A F. \tag{10}$$

Dostaliśmy wzór opisujący zamianę macierzy formy wraz ze zmianą bazy. Wprawdzie został on wyprowadzony dla $V = \mathbb{R}^n$, lecz **(10) jest prawdziwy w dowolnej p.w.** V.

Można przy tym zapytać, czy istnieje taka zamiana bazy, że w nowej bazie macierz $\tilde{A}$ przyjmuje prostą formę. Odpowiedź jest następująca:

Twierdzenie 6. Niech $A \in M_{n\times n}(\mathbb{R})$ i $A = A^T$, wtedy istnieje taka zamiana bazy F w p.u. V, że w nowej bazie mamy

$$F^T A F = \text{diag}\,(\lambda_1, \ldots, \lambda_n), \tag{11}$$

tj. w macierzy $\tilde{A}$ tylko na przekątnej występują wyrazy niezerowe. Co więcej F można tak wybrać, aby

$$\tilde{A} = \text{diag}\,(1, \ldots, 1, -1, \ldots, -1, 0, \ldots, 0),$$

gdzie jedynka występuje $p \geq 0$ razy zaś -1 występuje $q \geq 0$ razy, $p + q \leq n$. Parę (p, q) nazywamy *sygnaturą formy* Q.

Czasem bardzo nam zależy na tym, aby zamiana zmiennych była prostą operacją. Mamy następujący pozostawiony bez dowodu fakt.

Twierdzenie 7. Jeśli A jest macierzą formy dwuliniowej B, to istnieje taka zamiana bazy p.u V, że macierz B w nowej bazie $\tilde{A}$ jest diagonalna, tj. $\tilde{A} = F^T AF = \text{diag}\,(\lambda_1, \ldots, \lambda_n)$. Nadto, macierz F można tak wybrać, aby $F \in O(n)$.

W tym miejscu zamieścimy też dodatkowe uwagi dotyczące iloczynu skalarnego w $\mathbb{R}^n$. Przypominamy, że dla dowolnego $v \neq 0$ mamy

$$0 < (v, v)$$

Patrząc na iloczyn skalarny jak na formę dwuliniową natychmiast zauważamy, że jego macierz A (przypominamy, że $(v, w) = v^T Aw$) jest dodatnio określona (patrz §4.6.1). Tym samym **każda macierz dodatnio określona zadaje iloczyn skalarny w $\mathbb{R}^n$ wzorem (8)! Co więcej, w pewnej bazie owa macierz jest macierzą przekątniową (taką jak (11)), a nawet można dostać, że dla pewnego F, mamy $F^T AF = Id$!**

Zajmiemy się teraz praktycznym sposobem znalezienia bazy ortonormalnej w p.u. V. A mianowicie, jeśli $v_1, \ldots, v_n$ jest dowolną bazą V, to przeprowadzimy *ortogonalizację Grama-Schmida* tego układu wektorów, którego wynikiem będzie baza ortonormalna V. Zakładamy przy tym, że $\dim V = n$. Kładziemy

$$e_1 := v_1/\|v_1\|,$$

oczywiście e_1 ma normę 1. Dalej,

$$w_2 := v_2 - (v_2, e_1)e_1, \quad e_2 := w_2/\|w_2\|,$$

zauważmy, że

$$(w_2, e_1) = (v_2, e_1) - (v_2, e_1)(e_1, e_1) = 0.$$

Tym samym wektory e_1 i e_2 są ortogonalne. Dalej postępujemy indukcyjnie kładąc,

$$w_k = v_k - \sum_{i=1}^{k-1} (v_k, e_i)e_i, \quad e_k = w_k/\|w_k\|.$$

Sprawdzamy, że w_k jest prostopadły do e_j, $j = 1, \ldots, k-1$:

$$(w_k, e_j) = (v_k, e_j) - \sum_{i=1}^{k-1} (v_k, e_i)(e_i, e_j) = (v_k, e_j) - (v_k, e_j) = 0.$$

Zauważmy, że operacje przejścia od układu $\{v_1, \ldots, v_n\}$ do układu $\{e_1, v_2, \ldots, v_n\}$ i dalej od $\{e_1, \ldots, e_i, \ldots, v_n\}$ do $\{e_1, \ldots, e_i, e_{i+1}, \ldots, v_n\}$ nie zmieniają ilości wektorów lnz. Tym samym nowy układ $\{e_1, \ldots, e_n\}$ jest bazą.

8.5 Metoda najmniejszych kwadratów

Będziemy zakładali, że H jest przestrzenią Hilberta, niekoniecznie nad $\mathbb{R}$. Zdefiniujemy pojęcie rzutu.

Niech L będzie domkniętą podprzestrzenia liniową H, założenie domkniętości jest uczynione na przypadek gdy, np. $H = L^2(G)$. W przypadku, gdy $\dim H < \infty$ jest ono automatycznie spełnione. Definiujemy odwzorowanie $P : H \to L$ w następujący sposób. Jeśli $x \in H$, to $Px = y$, gdzie y jest takie, że

$$\|x - y\| = \inf_{z \in H} \|x - z\|.$$

Musimy najpierw odpowiedzieć na pytanie, czy taki wektor $y \in L$ w ogóle istnieje. Z definicji kresu dostaniemy oczywiście istnienie takich wektorów z_n, że

$$\lim_{n\to\infty} \|x - z_n\| \to \inf_{z \in H} \|x - z\| =: d.$$

Wykażemy, że ciąg $\{z_n\}_{n=1}^{\infty}$ jest ciągiem Cauchy'ego. Otóż oprzemy się na *tożsamości równoległoboku*, która jest prawdziwa dla dowolnych a, b należących do p.u. H,

$$\|a + b\|^2 + \|a - b\|^2 = 2(\|a\|^2 + \|b\|^2).$$

Jej sprawdzenie pozostawiamy czytelnikowi. Zauważmy teraz, że

$$\begin{aligned} \|z_n - z_m\|^2 &= \|(x - z_m) - (x - z_n)\|^2 \\ &= 2(\|x - z_m\|^2 + \|x - z_n\|^2) - \|2x - z_m - z_n\|^2 \\ &= 2(\|x - z_m\|^2 + \|x - z_n\|^2) - 4\|x - (z_m + z_n/2)\|^2 \\ &\leq 2(\|x - z_m\|^2 + \|x - z_n\|^2) - 4d^2 \to 2(d^2 + d^2) - 4d^2 = 0. \end{aligned}$$

Zatem istotnie, $\{z_n\} \subset L$ jest ciągiem Cauchy'ego. Zakładaliśmy, że H jest przestrzenią Hilberta, zatem dostaniemy istnienie $y \in H$, że $y = \lim_{m\to\infty} z_m$. Skoro podprzestrzeń L jest domknięta, to mamy stąd, że $y \in L$.

Skonstruowane odwzorowanie P nazwiemy *rzutem ortogonalnym*. Odpowiedniość tej nazwy stanie się za chwilę jasna. Mamy bowiem:

Stwierdzenie 8. $Px = y$ wtedy i tylko wtedy, gdy $(x - y, z) = 0$ dla wszystkich $z \in L$.

Dowód. $\Rightarrow$ Niech $z \in L$ będzie dowolnym wektorem, wtedy $z = y + h$ dla pewnego $h \in L$ i mamy

$$\begin{aligned} \|x - y\|^2 &\leq \|x - z\|^2 = \|x - y - h\|^2 = (x - y - h, x - y - h) \\ &= \|x - y\|^2 + \|h\|^2 - \mathrm{Re}(x - y, h) \end{aligned}$$

Gdyby $(x - y, h) \neq 0$, to moglibyśmy zastąpić h przez ht, gdzie $t \in \mathbb{R}$. Mielibyśmy wtedy

$$\|x - y\|^2 \leq \|x - y\|^2 + t^2\|h\|^2 - t\mathrm{Re}(x - y, h) < \|x - y\|^2$$

dla dobranego odpowiednio t. Znalezienie przykładu odpowiedniego t pozostawiamy Czytelnikowi jako ćwiczenie własne. Uzyskana sprzeczność pokazuje, że $(x-y,z)=0$.

$\Leftarrow$ Dowolny element $z \in L$ możemy zapisać jako $z = y + h$, gdzie dzięki założeniom $x-y$ jest prostopadły do $h \in L$. Zatem

$$\|x-z\|^2 = \|x-y-h\|^2 = \|x-y\|^2 + \|h\|^2.$$

Co kończy dowód stwierdzenia. □

Zajmiemy się teraz sformułowaniem tytułowego zagadnienia. Przeprowadzając doświadczenie chcemy znaleźć badane wielkości $x_1, \ldots, x_n$. Przeważnie bezpośredni ich pomiar nie jest możliwy. Możliwy jest natomiast odczyt wielkości $y_1, \ldots, y_m$, które zależą w znany sposób od $x_1, \ldots, x_n$, a mianowicie

$$y_k = f(z_k; x_1, \ldots, x_n), \quad k = 1, \ldots, m,$$

gdzie z_k są parametrami doświadczenia. Przeprowadzamy $m \geq n$ doświadczeń. Powstaje zagadnienie znalezienia $x_1, \ldots, x_n$. Jednak jeśli $m > n$ to nie należy się spodziewać, że uzyskamy dokładne rozwiązanie powyższego równania, ale możemy szukać $x_1, \ldots, x_n$, które „najlepiej" przybliżają uzyskane wyniki doświadczalne $y_1, \ldots, y_m$. Jeśli uznamy, że

$$\mathbf{y} = (y_1, \ldots, y_m) \quad \text{i} \quad \mathbf{f}(\mathbf{x}) = (f(z_1; \mathbf{x}) \ldots, f(z_m; \mathbf{x}))$$

są wektorami z $\mathbb{R}^m$, to możemy chcieć znaleźć $\mathbf{x} \in \mathbb{R}^n$ minimalizujące odległość $\mathbf{y}$ i $\mathbf{f}$:

$$\|\mathbf{y} - \mathbf{f}\|^2 = \sum_{k=1}^{m} (y_k - f(z_k; x_1, \ldots, x_n))^2 \tag{12}$$

Załóżmy, że funkcja f jest klasy C^1, wtedy warunkiem koniecznym istnienia minimum jest

$$\frac{\partial}{\partial x_i} \sum_{k=1}^{m} (y_k - f(z_k; x_1, \ldots, x_n))^2 = 0, \quad i = 1, \ldots, n. \tag{13}$$

Powyższy układ równań nazywamy *układem równań normalnych*. Ważny przypadek szczególny dostaniemy, gdy funkcja f jest liniowa, tj.

$$\mathbf{f}(\mathbf{x}) = A\mathbf{x}$$

dla pewnej macierzy A o wymiarach m na n. Wtedy równanie normalne (13) przyjmie postać

$$\nabla[(\mathbf{y} - A\mathbf{x}, \mathbf{y} - A\mathbf{x})] = \nabla[(\mathbf{y} - A\mathbf{x})^T(\mathbf{y} - A\mathbf{x})] = 2A^TA\mathbf{x} - 2A^T\mathbf{y} = 0$$

albo

$$A^TA\mathbf{x} = A^T\mathbf{y}. \tag{14}$$

W najprostszych przypadkach można (14) łatwo rozwiązać, w ogólności trzeba posłużyć się komputerem i wyrafinowanymi metodami numerycznymi.

Przejdziemy teraz do wykazania, że nasze zagadnienie najmniejszych kwadratów ma rozwiązanie. Jednocześnie podamy jego interpretację geometryczną.

Twierdzenie 9. Liniowy problem najmniejszych kwadratów

$$\min_{x\in\mathbb{R}^n} \|y - Ax\| \tag{15}$$

ma co najmniej jedno rozwiązanie x_0. Jeśli x_1 jest innym rozwiązaniem, to $Ax_0 = Ax_1$. Co więcej, reszta

$$r := y - Ax_0$$

jest wyznaczona jednoznacznie i spełnia $A^T r = 0$. Każde rozwiązanie x_0 jest rozwiązaniem równania normalnego (14) i na odwrót: rozwiązania (14) są rozwiązaniami problemu minimalizacji (15).

Dowód. Zdefiniujemy $L := \operatorname{Im} A \subset \mathbb{R}^m$. Skoro L jest domkniętą podprzestrzenią, to istnieje rzut $P : \mathbb{R}^n \to L$. Z definicji rzutu istnieje dokładnie jedno $z_0 = Py \in L$ takie, że

$$\|y - Py\| = \min_{z\in L} \|y - z\|.$$

Z właściwości rzutu mamy, że $y - Py \perp z$ dla dowolnego $z \in L$. Skoro L jest obrazem $\mathbb{R}^n$, to istnieje taki $x_0 \in \mathbb{R}^n$, że $z_0 = Ax_0$. Wtedy $r = Ax_0 - y$ i $r \perp z$ dla dowolnego $z \in L$, tj.

$$0 = (r, z) = (r, Ax) = r^T Ax = (A^T r)^T x = (A^T r, x)$$

dla każdego $x \in \mathbb{R}^n$. W szczególności powyższa równość pociąga zerowanie się współczynników $A^T r$ w dowolnej bazie ortonormalnej. Zatem

$$0 = A^T r = A^T(Ax_0 - y)$$

i x_0 jest rozwiązaniem równania (14).

Opowiedzmy prosty przypadek.

Przykład 8. W wyniku przeprowadzenia serii doświadczeń dostajemy zbiór punktów na płaszczyźnie $\mathcal{I} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^m$. Zadanie polega na zastosowaniu metody najmniejszych kwadratów do wyznaczenia funkcji liniowej $y = ax + b$ najlepiej przybliżającego zbiór $\mathcal{I}$. Chcemy wyznaczyć a i b minimalizujące wyrażenie (12):

$$\sum_{k=1}^{m} (\tilde{y}_k - a\tilde{x}_k - b)^2.$$

Aby uzgodnić nowe i stare oznaczenia napiszemy $x_1 = a$, $x_2 = b$, $y_k = \tilde{y}_k$ i $A_{k1} = \tilde{x}_k$, $A_{k2} = 1$, $k = 1, \ldots, m$. Wtedy równanie (14) przyjmuje postać

$$A^T Ax = A^T y,$$

gdzie $A^T A$ jest macierzą 2×2 i $A^T y \in \mathbb{R}^2$. Istnienie rozwiązań jest zapewnione twierdzeniem 9.

8.6 Wektory i wartości własne

Będziemy się zajmowali strukturą odwzorowań liniowych $F : V \to V$, gdzie V jest dowolną p.w. nad $\mathbb{R}$ lub $\mathbb{C}$. Okaże się, że przypadek zespolony jest prostszy. W poprzednim wykładzie pokazaliśmy kilka prostych przykładów przekształceń. W przykładzie 7. można było wskazać wektory, których obrazem były one same pomnożone przez pewną liczbę, $Fv = \lambda v$. Prostota powyższej struktury zasługuje na podkreślenie:

Definicja 6. Niech V będzie p.w. nad $\mathbb{K}$ i $F : V \to V$ będzie przekształceniem liniowym. Jeśli niezerowy wektor $v \in V$ spełnia

$$Fv = \lambda v$$

dla pewnego $\lambda \in \mathbb{K}$, to powiemy, że v jest *wektorem własnym*, zaś λ jest *wartością własną* przekształcenia F. Zbiór wektorów własnych odpowiadających wartości własnej λ łącznie z wektorem zerowym jest podprzestrzenią liniową oznaczaną V_λ i nazywana *podprzestrzenią własną* odpowiadającą λ.

Określiliśmy pewien element strukturalny macierzy, ale nie mamy pewności, czy zawsze istnieje, ani jak go wyznaczać. Zauważmy, że jeśli v jest wektorem własnym a μ odpowiadającą wartość własną, to mamy

$$(F - \mu Id)v = 0, \tag{16}$$

tj. macierz $F - \mu Id$ jest osobliwa, czyli jej wyznacznik znika:

$$0 = \det(F - \mu Id). \tag{17}$$

Zauważmy teraz, że funkcja

$$p(\lambda) := \det(F - \lambda Id)$$

jest wielomianem zmiennej λ. Nazywamy go *wielomianem charakterystycznym*. Jeśli liczba μ (w ogólności zespolona) jest wartością własną, to μ jest pierwiastkiem wielomianu charakterystycznego. Na odwrót, jeśli liczba λ_0 jest jest pierwiastkiem $p(\lambda)$, to jest spełnione równanie (17), tj. λ_0 wartością własną F. Dostaliśmy tym samym praktyczne narzędzie wyznaczania wartości własnych.

Jeśli V jest p.w. nad $\mathbb{C}$, to dzięki zasadniczemu twierdzeniu algebry istnieje μ zespolony pierwiastek $p(\lambda)$. Lecz jeśli mamy do czynienia z macierzą rzeczywistą, to jej wielomian charakterystyczny może nie mieć rzeczywistych pierwiastków, np.

$$F = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \text{ i } \det(F - \lambda Id) = \lambda^2 + 1$$

wielomian charakterystyczny nie ma pierwiastków rzeczywistych.

Rozważmy teraz możliwość innej komplikacji struktury macierzy

Przykład 9. Niech $F : \mathbb{C}^n \to \mathbb{C}^n$ będzie dane macierzą

$$F = \begin{bmatrix} \mu & 1 & 0 & \dots & 0 \\ 0 & \mu & 1 & & \vdots \\ 0 & \dots & \ddots & & 1 \\ 0 & \dots & & & \mu \end{bmatrix}$$

w bazie $\{e_1, \dots e_n\}$. Łatwo się przekonać, $p(\lambda) = (\lambda - \mu)^n$, tj. $\lambda = \mu$ jest n-krotnym pierwiastkiem $p(\lambda)$. Przede wszystkim jednak odnotujmy, że

$$Fe_1 = \mu e_1, \qquad Fe_i = \mu e_i + e_{i-1} \quad i = 2, \dots, n.$$

Sprawdźmy, czy są inne wektory (oprócz ke_1, $k \in \mathbb{K}$) należące do V_μ. Gdybyśmy dla pewnego $w \in \mathbb{C}^n$, $w = \sum_{i=1}^n c_i e_i$, mieli

$$Fw = \mu w,$$

to

$$\begin{aligned} Lewa &= \sum_{i=1}^{n} c_i \mu e_i + \sum_{i=1}^{n-1} c_{i+1} e_i \\ Prawa &= \mu \sum_{i=1}^{n} c_i e_i. \end{aligned}$$

Wypływa skąd wniosek, że

$$\sum_{i=1}^{n-1} c_{i+1} e_i = 0$$

co jest równoważne temu, że $c_2 = \dots = c_n = 0$, tj. $v = c_1 e_1$.

Powinniśmy teraz zastanowić się, jak się zmienia macierz odwzorowania, gdy zmieniamy bazę w p.w. V. Niech wektor w starych współrzędnych to v^s, zaś w nowych to v^n, wtedy

$$v^s = Rv^n,$$

gdzie R nazywamy macierzą przejścia, (patrz (9)). Wtedy zagadnienie własne

$$Av^s = \lambda v^s$$

przyjmie postać

$$ARv^n = \lambda Rv^n.$$

Ponieważ chcemy znaleźć wektor v^n, to mnożymy powyższe równanie z lewej strony przez R^{-1}. Dostaniemy wtedy

$$R^{-1}ARv^n = \lambda v^n.$$

Możemy wtedy zapytać, jaka jest najprostsza struktura macierzy F po zamianie bazy i jak znaleźć macierz przejścia R. Oto odpowiedź na pierwsze z pytań.

Twierdzenie 10. (postać Jordana macierzy). Załóżmy, że $V = \mathbb{C}^n$ i $F : V \to V$, wtedy istnieje taka baza V, że F w nowej bazie ma postać

$$R^{-1}AR = \begin{bmatrix} J_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & J_k \end{bmatrix},$$

gdzie $k \leq n$, zaś J_i, $i = 1, \ldots, k$ nazywamy *klatkami Jordana* i mają one postać

$$J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \ldots & 0 \\ 0 & \lambda_i & 1 & & \vdots \\ 0 & \ldots & \ddots & & 1 \\ 0 & \ldots & & & \lambda_i \end{bmatrix}$$

gdzie λ_i jest wartością własną.

Przejdźmy teraz do pytania jak znaleźć macierz przejścia R. Ustalmy pierwszą wartość własną λ. Niech e_1 będzie wektorem własnym odpowiadającym wartości własnej λ i e_i $i = 2, \ldots, j$ będą *wektorami dołączonymi*, takimi, że $Fe_i = \lambda e_i + e_{i-1}$. Wtedy w pierwszej kolumnie macierzy R zapisujemy e_1, w drugiej e_2 itd aż do wyczerpania wektorów dołączonych do e_1. Podobnie postępujemy z pozostałymi wektorami własnymi odpowiadającymi wartości własnej λ i dalej aż do wyczerpania wszystkich wartości własnych.

Musimy więc umieć liczyć $\dim V_\lambda$ jak i znajdować wektory dołączone. Z równania (16) wynika natychmiast na podstawie stwierdzenia 2.33, że

$$\dim V_\lambda = n - \text{rząd macierzy } (F - \lambda Id), \tag{18}$$

gdzie n jest wymiarem macierzy F. Z kolei, to z samej definicji wektora dołączonego e_2 wiemy spełnia on

$$(F - \lambda Id)e_2 = e_1,$$

A zatem po wzięciu $F - \lambda Id$ od obu stron

$$(F - \lambda Id)^2 e_2 = 0.$$

Aby ułatwić wysłowienie się wprowadzimy pojęcie *krotności pierwiastka wielomianu* $W(x)$, jest to taka liczba naturalna k, że $W(x) = (x - \lambda)^k q(x)$, gdzie $q(x)$ jest wielomianem i $q(\lambda) \neq 0$. Używając nowego pojęcia, rozważmy przypadek, gdy $\dim V_\lambda =$ krotność λ pierwiastka wielomianu charaterystycznego $p(x)$. Wtedy proces badania kończy się, mamy tyle 1 na 1 klatek Jordana ile wynosi krotność λ. Innych sytuacji nie będziemy rozważali.

Zbadajmy konkretny

Przykład 10. Niech

$$A = \begin{bmatrix} 0 & 1 & 0 \\ -4 & 4 & 0 \\ -2 & 1 & 2 \end{bmatrix}$$

Wtedy

$$p(\lambda) = \det(A - \lambda Id) = -(\lambda - 2)^3,$$

tj. $\lambda = 2$ jest pierwiastkiem potrójnym wielomianu i bez dodatkowych informacji mamy 3 możliwości macierzy Jordana:

$$J_1 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad J_2 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix}, \quad J_3 = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix}.$$

Znajdujemy $A - 2Id$:

$$A - 2Id = \begin{bmatrix} -2 & 1 & 0 \\ -4 & 2 & 0 \\ -2 & 1 & 0 \end{bmatrix}.$$

Łatwo widzimy, że rząd tej macierzy jest równy jeden, więc z (18) wynika, że dim $V_2 = 2$ i postać Jordana macierzy to J_2. Znajdziemy teraz bazę, o której mowa w twierdzeniu 10. Przykładowe lnz elementy V_2 to

$$f_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad f_2 = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}.$$

Bardzo łatwo sprawdzić, że $(A - 2Id)^2 = 0$, więc dowolny wektor, który nie jest kombinacją liniową f_1 i f_2 jest kandydatem na dołączony. Weźmy $e = (1, 0, 0)^T$. Sprawdzamy, że $Ae = 2e - 2f_1 - 2f_2$. Widzimy, że wystarczy przyjąć $e_2 = -2f_1 - 2f_2$, bo jest on wektorem własnym i $e_3 = e$. Mamy wtedy $Ae_2 = 2e_2$, $Ae_3 = 2e_3 + e_2$. Na koniec kładziemy $e_1 = f_1$. Wtedy właściwą bazą jest (e_1, e_2, e_3), zaś macierz przejścia jest $R = [e_1, e_2, e_3]$ i sprowadza ona A do postaci J_2.

8.6.1 Układy liniowych równań różniczkowych

Pora na zastosowania. Niech A będzie rzeczywistą macierzą symetryczną 2 na 2. Zadanie polega na rozwiązaniu układ równań różniczkowych zwyczajnych

$$x' = Ax, \qquad x(0) = x_0. \tag{19}$$

Wiemy, że rozwiązanie $x(t)$ istnieje. Wiemy też, że istnieją dwie wartości własne λ_1, λ_2. Jeśli $\lambda_1 \neq \lambda_2$, to istnieją odpowiadające im lnz wektory własne v_1 i v_2. Jeśli $\lambda_1 = \lambda_2$, to chwilowo zakładamy istnienie lnz wektorów własnych v_1 i v_2. Wtedy nasze rozwiązanie zapisuje się następująco

$$x(t) = \alpha_1(t)v_1 + \alpha_2(t)v_2, \tag{20}$$

dla pewnych funkcji α_1 i α_2. Po wstawieniu do równania dostaniemy

$$\alpha_1'(t)v_1 + \alpha_2'(t)v_2 = \lambda_1\alpha_1(t)v_1 + \lambda_2\alpha_2(t)v_2.$$

Ponieważ wektory v_1 i v_2 są lnz, to dostaniemy, że

$$\begin{aligned} \alpha_1' &= \lambda_1\alpha_1 \\ \alpha_2' &= \lambda_2\alpha_2. \end{aligned}$$

jest to już naprawdę prosty układ równań różniczkowych zwyczajnych. Jego rozwiązanie to,

$$\begin{aligned} \alpha_1(t) &= e^{\lambda_1 t}\alpha_1(0) \\ \alpha_2(t) &= e^{\lambda_2 t}\alpha_2(0), \end{aligned}$$

gdzie liczby $\alpha_1(0)$ i $\alpha_2(0)$ wyznaczamy z warunków początkowych:

$$x_0 = \alpha_1(0)v_1 + \alpha_2(0)v_2,$$

co w zestawieniu z (20) daje nam pełne rozwiązanie zagadnienia.

Rozważmy teraz przykład układu równań (19) takiego, że rzeczywista macierz A ma wielomian chararakterystyczny $p(\lambda)$, taki, że $\mu = \alpha + \beta i$ jest jego pierwiastkiem. Ponieważ $p(\lambda)$ ma współczynniki rzeczywiste, to $0 = \overline{p(\mu)} = p(\overline{\mu})$, tj. $\bar{\mu}$ też jest wartością własną. Niech $v \in \mathbb{C}^2$ będzie wektorem własnym odpowiadającym μ. Wtedy z (16) wynika, że

$$0 = \overline{(A - \mu Id)v} = (A - \bar{\mu} Id)\bar{v}.$$

A zatem $\bar{v}$ jest wektorem własnym odpowiadającym $\bar{\mu}$. Wydaje się kłopotliwym, że wiemy o istnieniu rzeczywistych rozwiązań przy założeniu rzeczywistych danych, ale algebra liniowa nie wydaje się tego podpowiadać. Zauważmy, że funkcje $\exp(\lambda t)v$ i $\exp(\bar{\lambda})\bar{v}$ są rozwiązaniami układu (19). Podobnie ich kombinacje liniowe. Możemy więc zastosować technikę współczynników nieokreślonych do rozwiązania układ (19): kładziemy

$$x_1(t) = C_1 e^{\alpha t}\cos\beta t + C_2 e^{\alpha t}\sin\beta t, \quad x_2(t) = D_1 e^{\alpha t}\cos\beta t + D_2 e^{\alpha t}\sin\beta t.$$

Jeśli wstawimy te funkcje do układu (19), to dostaniemy układ 2 równań na 4 niewiadome. Jednak warunki początkowe od razu wyznaczają C_1 i D_2. Otrzymany układ jest już łatwo rozwiązać.

Nadto możemy uczynić obserwację natury algebraicznej: wektory

$$e_1 = v + \bar{v}, \quad e_2 = (v - \bar{v})/i$$

są rzeczywiste i rozpinają $\mathbb{C}^2$ (jak i $\mathbb{R}^2$). Zauważmy, że

$$\begin{aligned} Ae_1 &= \lambda v + \bar{\lambda}\bar{v} = \alpha e_1 - \beta e_2 \\ Ae_2 &= (\lambda v - \bar{\lambda}\bar{v})/i = \beta e_1 + \alpha e_2. \end{aligned}$$

Tym samym w bazie e_1 i e_2 macierz A ma postać

$$\begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix} = R^{-1}AR,$$

gdzie $R = [e_1, e_2]$.

Okazuje się, że jest to najprostsza postać macierzy rzeczywistej, której wielomian charakterystyczny ma pierwiastki zespolone.

Twierdzenie 11. (cd. postaci Jordana macierzy) Niech V będzie rzeczywistą p.w. wymiaru n oraz $F : V \to V$. Zakładamy, że $\mu \in \mathbb{C} \setminus \mathbb{R}$ jest pierwiastkiem wielomianu $p(\lambda)$ krotności k i $\dim V_\mu = k$. Wtedy istnieje taka baza w V, że F ma w niej postać

$$R^{-1}AR = \begin{bmatrix} J_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & J_l \end{bmatrix},$$

gdzie dla rzeczywistych wartości własnych klatki Jordana są takie jak poprzednio, natomiast dla zespolonych wartości własnych $\mu = \alpha + i\beta$ mamy

$$J_i = \begin{bmatrix} E & \dots & 0 \\ 0 & E & \dots \\ 0 & \dots & E \end{bmatrix}, \text{ gdzie } E = \begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix}.$$

Uwaga. Gdy krotność μ jest większa niż wymiar przestrzeni wektorów własnych odpowiadających μ, to postać klatek Jordana jest bardziej złożona. Pomijamy ten temat.

Na zakończenie opiszemy rozwiązywanie układu (19), gdy macierz A ma jedną podwójną wartość własną, ale jest tylko jeden wektor własny. Istnieje wtedy taka baza e_1, e_2, że $Ae_1 = \lambda e_1$ i $Ae_2 = \lambda e_2 + e_1$. Rozwiązanie $x(t)$ można przedstawić następująco

$$x(t) = \alpha_1(t)e_1 + \alpha_2(t)e_2. \tag{21}$$

Po wstawieniu do układu dostaniemy równania na α_1 i α_2:

$$\alpha_1'(t)e_1 + \alpha_2'(t)e_2 = \lambda\alpha_1(t)e_1 + \lambda\alpha_2(t)e_2 + \alpha_2(t)e_1.$$

Zatem

$$\alpha_1' = \lambda\alpha_1 + \alpha_2 \qquad \alpha_2' = \alpha_2.$$

Drugie równanie jest proste do rozwiązania: $\alpha_2(t) = \exp(\lambda t)\alpha_2(0)$. Po wstawieniu do pierwszego równania wyżej dostaniemy

$$\alpha_1' = \lambda\alpha_1 + \exp(\lambda t)\alpha_2(0).$$

Z tego co wiemy o równaniach liniowych dostaniemy, że

$$\alpha_1(t) = \exp(\lambda t)(\alpha_1(0) + \alpha_2(0)t),$$

gdzie $x(0) = \alpha_1(0)e_1 + \alpha_2(0)e_2$, co w połączeniu z (21) w pełni opisuje rozwiązanie.

KONIEC WYKŁADU