Optymalizacja II

Jan Palczewski
poprawki Andrzej Palczewski
E-mail: J.Palczewski@mimuw.edu.pl
WWW: http://www.mimuw.edu.pl/~apalczew

10 marca 2018

Streszczenie. Wykład prezentuje teorię optymalizacji nieliniowej z ograniczeniami równościowymi i nierównościowymi. Uzupełniony jest z wprowadzeniem do metod numerycznych.

Tu będzie informacja o prawach autorskich i zasadach powielania

Spis treści

1 Wiadomości wstępne **8**

1.1 Problem optymalizacji . . . 8
1.2 Istnienie minimum funkcji ciągłej . . . 9
1.3 Minima lokalne funkcji jednej zmiennej . . . 10
1.4 Wzory Taylora . . . 11
1.5 Ekstrema globalne . . . 13
1.6 Zadania . . . 14

2 Ekstrema funkcji wielu zmiennych **15**

2.1 Notacja i twierdzenia Taylora w wielu wymiarach . . . 15
2.2 Znikanie gradientu . . . 17
2.3 Dodatnia i ujemna określoność macierzy . . . 17
2.4 Warunki II-go rzędu (kryterium drugiej różniczki) . . . 19
2.4.1 Ekstrema globalne i określoność drugiej różniczki . . . 19
2.5 Zadania . . . 20

3 Funkcje wypukłe **22**

3.1 Zbiory wypukłe i twierdzenia o oddzielaniu . . . 22
3.2 Definicja funkcji wypukłej . . . 25
3.3 Własności funkcji wypukłych . . . 26
3.4 Charakteryzacje funkcji wypukłej . . . 28
3.5 Subróżniczka . . . 30
3.6 Zadania . . . 36

4 Ekstrema funkcji wypukłej z ograniczeniami **39**

4.1 Problem minimalizacyjny . . . 39
4.2 Funkcje pseudowypukłe . . . 43
4.3 Maksymalizacja funkcji wypukłej . . . 45
4.4 Zadania . . . 49

5 Warunek konieczny I rzędu **52**

5.1 Stożek kierunków stycznych . . . 52

5.2 Ograniczenia nierównościowe . . . 55
5.3 Warunki konieczne Kuhna-Tuckera . . . 57
5.4 Zadania . . . 59

6 Warunki regularności i przykłady 61
6.1 Warunki regularności . . . 61
6.2 Przykłady . . . 64
6.3 Zadania . . . 66

7 Funkcje quasi-wypukłe i warunki dostateczne 68
7.1 Quasi-wypukłość . . . 68
7.2 Maksymalizacja funkcji quasi-wypukłej . . . 73
7.3 Warunki dostateczne . . . 74
7.4 Zadania . . . 75

8 Warunek konieczny dla ograniczeń mieszanych 79
8.1 Warunek konieczny pierwszego rzędu . . . 80
8.2 Warunki regularności . . . 81

9 Warunki drugiego rzędu 85
9.1 Warunki drugiego rzędu . . . 85
9.2 Podsumowanie . . . 88
9.3 Przykład . . . 89
9.4 Zadania . . . 91

10 Teoria dualności 92
10.1 Warunek dostateczny . . . 92
10.2 Warunek konieczny dla programowania wypukłego . . . 94
10.3 Zadanie pierwotne i dualne . . . 98
10.4 Zadania . . . 101

11 Teoria wrażliwości 104
11.1 Ograniczenia równościowe . . . 104
11.2 Ograniczenia nierównościowe . . . 106
11.3 Zadania . . . 109

12 Wprowadzenie do numerycznych metod optymalizacji 110
12.1 Własności algorytmów optymalizacyjnych . . . 111
12.2 Optymalizacja bez użycia pochodnych . . . 112
12.3 Zadania . . . 116

13 Metody spadkowe 118

13.1 Metody największego spadku 118
13.2 Metoda Newtona 126
13.3 Metoda kierunków i gradientów sprzężonych 128
13.4 Zadania 132

14 Metody optymalizacji z ograniczeniami **134**
14.1 Algorytm Zoutendijka dla ograniczeń afinicznych 134
14.2 Algorytm Zoutendijka dla ograniczeń nieliniowych 137
14.3 Modyfikacja Topkisa-Veinotta 141
14.4 Podsumowanie 142
14.5 Zadania 142

Wprowadzenie

„Optymalizacja" to poszukiwanie czegoś „najlepszego". Sprawdzenie, czy coś (np. decyzja) jest najlepsze, wymaga, przede wszystkim, określenia jakiejś miary, która pozwoli tą decyzję ocenić – *funkcji celu*. Konieczne jest także opisanie zbioru wszystkich dopuszczalnych decyzji – *zbioru punktów dopuszczalnych*. Matematyczne przedstawienie zadania optymalizacyjnego, które zaprezentujemy na tym wykładzie, może wydać się dużym uproszczeniem, lecz okazuje się ono być wystarczające w wielu praktycznych zastosowaniach. Będziemy opisywać zbiór punktów dopuszczalnych jako podzbiór wielowymiarowej przestrzeni rzeczywistej opisany przez układ równości i/lub nierówności. Funkcja celu będzie przyporządkowywała każdemu punktowi tego zbioru liczbę rzeczywistą mierzącą jego optymalność. W przypadku minimalizacji, funkcja celu często zwana jest funkcją kosztu, a celem optymalizacji jest wybranie spośród dozwolonych, opisanych przez ograniczenia, sposobów postępowania tych, które ten koszt uczynią najmniejszym.

Jeśli funkcja celu i wszystkie funkcje opisujące ograniczenia są liniowe, to mamy do czynienia z *programowaniem liniowym*. Istnieje wówczas algorytm (tzw. algorytm sympleks) pozwalający na szybkie i dokładne rozwiązywanie takich zagadnień (patrz monografie Bazaraa, Jarvis, Sherali [2] oraz Gass [8]). Sprawa znacznie się komplikuje, jeśli choć jedna z funkcji jest nieliniowa. Przenosimy się wówczas do świata *programowania nieliniowego*, który jest dużo bogatszy i trudniejszy. W ten właśnie świat mają wprowadzić czytelnika niniejsze notatki.

Okazuje się, że wiele zastosowań matematyki sprowadza się właśnie do problemów optymalizacyjnych. Zarządzanie procesami produkcyjnymi, logistyka, czy decyzje inwestycyjne to typowe problemy programowania nieliniowego. Nie dziwi zatem, że wielu ekonomistów jest ekspertami w tej dziedzinie. Co więcej, wiele teorii ekonomicznych opiera się na założeniu, że świat dąży lub oscyluje wokół punktu równowagi, czyli punktu będącego rozwiązaniem pewnego problemu optymalizacyjnego.

Równie ważne zastosowania ma programowanie nieliniowe w mechanice, elektronice, zarządzaniu zasobami wodnymi (tamy, irygacja itp.) oraz budownictwie. Można się pokusić o stwierdzenie, że to jedna z najczęściej stosowanych przez niematematyków dziedzin matematyki. Nie można też pominąć statystyki: np. metoda *najmniejszych* kwadratów, czy *największej* wiarogodności.

Zwykle, gdy teoria matematyczna zostaje użyta w praktyce, eleganckie metody analityczne oddają pola metodom numerycznym. Metody numeryczne mają na celu znalezienie przybliżenia rozwiązania zadania optymalizacyjnego, gdy staje się ono zbyt skomplikowane, by rozwiązać je w piękny analityczny sposób. Trzy ostatnie rozdziały tych notatek stanowią wprowadzenie do tej ważnej dziedziny. Zainteresowany czytelnik może poszerzyć swoją wiedzę czytając monografie Bertsekasa [4, 5], Luenbergera [9] lub Bazaraa, Sherali i Shetty [3].

Literatura

Monografia Bazaraa, Sherali, Shetty [3] prezentuje podejście bardziej inżynierskie, utylitarne. Prezentuje zarówno teorię, jak i metody numeryczne. Wszystkie twierdzenia poparte są dowodami i uzupełnione przykładami, tak więc stwierdzenie, że jest to pozycja inżynierska, tyczy się tego, że autorzy ilustrują matematyczne rozumowania intuicjami pochodzącymi z zastosowań.

Bertsekas jest ekspertem od metod numerycznych programowania nieliniowego. W jego książkach [4, 5] czytelnik może szukać zaawansowanych algorytmów. Teoria jest jednak zaprezentowana dość skrupulatnie, więc i tutaj można się całkiem dużo dowiedzieć o metodach analitycznych.

Książka Luenbergera [9] prezentuje zarówno teorię programowania liniowego jak i nieliniowego. Większy nacisk położony jest w niej na metody numeryczne niż na prezentację matematycznej teorii w pełnej ogólności.

Programowanie liniowe nie jest przedmiotem tych notatek, lecz co jakiś czas wspominane, szczególnie w przypadku metod numerycznych. Czytelnik chcący pogłębić wiedzę na jego temat odsyłany jest do książek Bazaraa, Jarvis, Shetty [2] i Gass [8].

Literatura w języku polskim nie jest zbyt obszerna. Warto tu wspomnieć monografie Zangwilla [13] i Canona, Culluma i Polaka [6].

Notacja

$\mathbb{R}$ – zbiór liczb rzeczywistych,

$\mathbf{x} = (x_1, \ldots, x_n)$ – wektor (pogrubiona litera)

$\mathbf{x} \leqslant \mathbf{y}$ – relacja pomiędzy dwoma wektorami; równoważna $x_i \leqslant y_i$ dla każdego i

$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$ – norma euklidesowa w $\mathbb{R}^n$,

$\operatorname{cl} W$ – domknięcie zbioru W w domyślnej normie (najczęściej euklidesowej)

$f'(x), f''(x)$ – pierwsza i druga pochodna funkcji jednej zmiennej

$Df(\mathbf{x})$ – pierwsza pochodna funkcji wielu zmiennych, wektor wierszowy

$D^2 f(\mathbf{x})$ – macierz drugich pochodnych funkcji wielu zmiennych

Podziękowania

Część wykładów i zadań bazuje na notatkach prof. Bronisława Jakubczyka. W największym stopniu dotyczy to wykładów 1, 2, 10, 11. Niektóre zadania pochodzą od dr. Wojciecha Kryńskiego. Ogromne wyrazy wdzięczności należą się Agnieszce Wiszniewskiej-Matyszkiel za liczne uwagi i komentarze dotyczące zarówno samego doboru materiału jak i jego przedstawienia.

Jan Palczewski, Warszawa 2012

Obecna wersja wykładu zawiera kilka modyfikacji w stosunku do wersji z roku 2012. Zasadnicza różnica polega na umieszczeniu nowej wersji wykładów 12 i 13, a także rozszerzeniu wykładów 10 i 14. Poza tym zostały skorygowane zauważone błędy lub niezręczności w dowodach i przykładach. Autor poprawek jest bardzo wdzięczny Przemysławowi Kiciakowi za wskazanie wielu z tych błędów, a także za uzupełnienie wykładów o rys. 4.2, 5.4 i 10.3.

Andrzej Palczewski, Warszawa 2016

Rozdział 1

Wiadomości wstępne

1.1 Problem optymalizacji

Niech $W \subset \mathbb{R}^n$ będzie niepustym zbiorem, zaś $f : W \to \mathbb{R}$ dowolną funkcją. Będziemy rozważać problem minimalizacji funkcji f na zbiorze W, przyjmując różne postaci W, w tym:

- $W = \mathbb{R}^n$ (problem optymalizacji bez ograniczeń),
- $W = \{x \in \mathbb{R}^n : g_1(x) = 0, \ldots, \ g_m(x) = 0\}$, gdzie $g_1, \ldots g_m : \mathbb{R}^n \to \mathbb{R}$ pewne funkcje (problem optymalizacji z ograniczeniami równościowymi),
- $W = \{x \in \mathbb{R}^n : g_1(x) \leqslant 0, \ldots, \ g_m(x) \leqslant 0\}$, gdzie $g_1, \ldots g_m : \mathbb{R}^n \to \mathbb{R}$ pewne funkcje (problem optymalizacji z ograniczeniami nierównościowymi).

Zbiór W nosi nazwę zbioru *punktów dopuszczalnych*.

Definicja 1.1. Punkt $x_0 \in W$ nazywamy *minimum globalnym* funkcji f na zbiorze W jeśli

$$f(x) \geqslant f(x_0) \quad \text{dla każdego} \ \ x \in W.$$

Definicja 1.2. Punkt $x_0 \in W$ nazywamy *minimum lokalnym* funkcji f jeśli istnieje $\varepsilon > 0$ takie, że dla kuli $B(x_0, \varepsilon)$ o środku w x_0 i promieniu ε zachodzi

$$f(x) \geqslant f(x_0) \quad \text{dla każdego} \ \ x \in B(x_0, \varepsilon) \cap W.$$

Oczywiście, jeśli x_0 jest minimum globalnym to jest minimum lokalnym. Minimum nazywamy *ścisłym*, jeśli w powyższych definicjach zachodzi $f(x) > f(x_0)$, dla $x \neq x_0$. Analogicznie definiujemy globalne i lokalne maksimum. Punkt x_0 nazywamy *ekstremum* (lokalnym, globalnym) jeśli jest on maksimum lub minimum (lokalnym, globalnym).

Minimum (globalne, lokalne) nie musi istnieć, tzn. może się okazać, że nie istnieje x_0 spełniające warunek z definicji 1.1 lub 1.2. W szczególności minimum globalne f na zbiorze W nie istnieje gdy:

(a) $\inf_{x \in W} f(x) = -\infty$, lub

(b) $\inf_{x \in W} f(x) = c \in \mathbb{R}$, ale $\nexists \ x \in U$ takie, że $f(x) = c$.

Przykład 1.1. Niech $W = \mathbb{R}$, $f(x) = x \sin x$. Dla tej funkcji $\inf_{x\in W} f(x) = -\infty$, zatem minimum globalne nie istnieje. Jeżeli natomiast ograniczymy się do przedziału $W = [a, b]$, to minimum globalne będzie istnieć. Funkcja ta osiąga minima lokalne w nieskończonej ilości punktów, dla $W = \mathbb{R}$. Jeżeli za W przyjmiemy odcinek otwarty, to minimum globalne istnieje lub nie istnieje, w zależności od tego odcinka. Ogólnie, funkcja ciągła może nie osiągać kresów na zbiorze niezwartym, w szczególności na podzbiorze otwartym $W \subset \mathbb{R}^n$.

1.2 Istnienie minimum funkcji ciągłej

Przypomnijmy, że podzbiór zwarty w $\mathbb{R}^n$ to podzbiór domknięty i ograniczony.

Twierdzenie 1.1. *Jeśli W jest zbiorem zwartym i $f : W \to \mathbb{R}$ jest funkcją ciągłą, to f osiąga kresy na W (dolny i górny). Oznacza to, że istnieją $\mathbf{x}_0$, $\mathbf{y}_0 \in W$ takie, że dla dowolnego $\mathbf{x} \in W$ zachodzi*

$$f(\mathbf{x}_0) \leqslant f(\mathbf{x}) \leqslant f(\mathbf{y}_0).$$

Będziemy oznaczali normę euklidesową w $\mathbb{R}^n$ przez

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}.$$

Warunek zwartości zbioru w powyższym twierdzeniu możemy osłabić do warunku domkniętości, jeśli funkcja jest koercywna. Koercywność funkcji definiujemy następująco:

Definicja 1.3. Funkcję $f : W \to \mathbb{R}$ na podzbiorze $W \subset \mathbb{R}^n$ nazywamy *koercywną*, jeśli $f(\mathbf{x}) \to \infty$ dla $\|\mathbf{x}\| \to \infty$. Można ten warunek zapisać równoważnie

$$\forall_{r>0}\ \exists_{s>0}\ \forall_{\mathbf{x}\in W}: \quad \|\mathbf{x}\| > s \quad \Longrightarrow \quad f(\mathbf{x}) > r.$$

W szczególności, jeśli W jest ograniczony, to f jest automatycznie koercywna na W.

Twierdzenie 1.2. *Jeśli zbiór W jest domknięty oraz funkcja $f : W \to \mathbb{R}$ jest ciągła i koercywna, to istnieje punkt $\mathbf{x}_0$ w którym funkcja f przyjmuje minimum, tzn. istnieje $\mathbf{x}_0 \in W$ taki, że $f(\mathbf{x}_0) = \inf_{\mathbf{x}\in W} f(\mathbf{x})$.*

Dowód. Niech $\mathbf{y}$ będzie ustalonym punktem w zbiorze $W \subset \mathbb{R}^n$. Rozpatrzmy zbiór $U = \{\mathbf{x} \in W : f(\mathbf{x}) \leqslant f(\mathbf{y})\}$. Zauważmy, że U jest zbiorem domkniętym w W, bo funkcja f jest ciągła, a nierówność w warunku nieostra. Z domkniętości W wynika, że U jest domknięty w $\mathbb{R}^n$. Jest on też ograniczony. Mianowicie, dla $r = f(\mathbf{y})$, z koercywności f istnieje $s > 0$ takie, że jeśli $\|\mathbf{x}\| > s$, to $f(\mathbf{x}) > r = f(\mathbf{y})$, skąd U jest zawarte w kuli $B(s) = \{\mathbf{x} : \|\mathbf{x}\| \leqslant s\}$. Zatem U jest zbiorem zwartym. Wówczas istnieje $\mathbf{x}_0 \in U$ – punkt minimum na zbiorze U. Dla $\mathbf{x} \notin U$ mamy $f(\mathbf{x}) > f(\mathbf{y}) \geqslant f(\mathbf{x}_0)$, więc $\mathbf{x}_0$ jest globalnym minimum na całym W. □

Domkniętość W nie jest potrzebna, jeśli f odpowiednio rośnie w pobliżu granicy ∂W.

Twierdzenie 1.3. *Niech $W \subset \mathbb{R}^n$ będzie dowolnym niepustym podzbiorem oraz $f : W \to \mathbb{R}$ – funkcją ciągłą. Jeśli dla pewnego ustalonego punktu $\mathbf{y} \in W$ oraz dowolnego ciągu $\mathbf{x}_n \in W$, takiego że*

$$\mathbf{x}_n \to \operatorname{cl} W \setminus W \qquad \textit{lub} \qquad \|\mathbf{x}_n\| \to \infty$$

zachodzi

$$\liminf_{n\to\infty} f(\mathbf{x}_n) > f(\mathbf{y}),$$

to istnieje punkt $\mathbf{x}_0$ w którym funkcja f przyjmuje minimum.

Dowód. Zbiór U definiujemy jak w poprzednim dowodzie, $U = \{\mathbf{x} \in W : f(\mathbf{x}) \leqslant f(\mathbf{y})\}$. Aby pokazać domkniętość U weźmy dowolny ciąg $(\mathbf{x}_n) \subset U$ zbieżny do $\tilde{\mathbf{x}}$. Pokażemy, że $\tilde{\mathbf{x}} \in U$. Z $\mathbf{x}_n \in U$ mamy $f(\mathbf{x}_n) \leqslant f(\mathbf{y})$ i jeśli $\tilde{\mathbf{x}} \notin W$, nierówność ta przeczy założeniu twierdzenia. Wynika stąd, że $\tilde{\mathbf{x}} \in W$. Korzystając teraz z ciągłości f na W dostajemy $f(\tilde{\mathbf{x}}) \leqslant f(\mathbf{y})$, skąd $\tilde{\mathbf{x}} \in U$. Ograniczoność zbioru U wynika z założonej w twierdzeniu implikacji $\|\mathbf{x}_n\| \to \infty \Rightarrow \liminf_{n\to\infty} f(\mathbf{x}_n) > f(\mathbf{y})$. Pozostała część dowodu jest identyczna jak w dowodzie poprzedniego twierdzenia. □

Przykład 1.2. Funkcja $f(x, y) = xy - \ln(xy)$ jest ciągła i spełnia założenia Twierdzenia 1.3 na zbiorze $W = \{(x, y) \in \mathbb{R}^n : x > 0, y > 0, x + y \leqslant 4\}$, osiąga więc minimum na W.

1.3 Minima lokalne funkcji jednej zmiennej

Niech $W \subset \mathbb{R}$ – podzbiór otwarty. Przypomnimy elementarne fakty.

Twierdzenie 1.4 (Warunek konieczny I rzędu)**.** *Jeśli $x_0 \in W$ jest punktem lokalnego minimum lub maksimum funkcji $f : W \to \mathbb{R}$ oraz f posiada pochodną w punkcie x_0, to*

$$f'(x_0) = 0.$$

Dowód twierdzenia 1.4. Niech x_0 - minimum lokalne. Dla dostatecznie małych h zachodzi $f(x_0+h) \geqslant f(x_0)$. Zatem dla $h > 0$ mamy

$$\frac{f(x_0+h)-f(x_0)}{h} \geqslant 0 \implies \lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h} \geqslant 0 \implies f'(x_0) \geqslant 0.$$

Dla $h < 0$

$$\frac{f(x_0+h)-f(x_0)}{h} \leqslant 0 \implies \lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h} \leqslant 0 \implies f'(x_0) \leqslant 0.$$

Stąd $f'(x_0) = 0$. □

Twierdzenie 1.5 (Warunek konieczny II rzędu)**.** *Jeśli $f : W \to \mathbb{R}$ jest klasy C^2 na zbiorze W i x_0 jest punktem lokalnego minimum, to*

$$f''(x_0) \geqslant 0.$$

Twierdzenie 1.6 (Warunek dostateczny II rzędu)**.** *Jeśli $f : W \to \mathbb{R}$ jest klasy C^2 na zbiorze W oraz $f'(x_0) = 0$, $f''(x_0) > 0$ dla pewnego $x_0 \in W$, to f ma ścisłe lokalne minimum w punkcie x_0.*

Uwaga 1.1. Twierdzenie 1.5 pozostaje prawdziwe przy zamianie lokalnego minimum na maksimum, jeśli znak znak drugiej pochodnej zmienimy na przeciwny.

Uwaga 1.2. Jeśli W nie jest otwarty, to Twierdzenia 1.4 i 1.5 nie są prawdziwe dla $x_0 \in \partial W$ (brzeg W), np. funkcja $f(x) = -x^2$ przyjmuje minimum na odcinku $[0, 2]$ w punkcie $x_0 = 2$, ale żaden z warunków koniecznych tych twierdzeń nie zachodzi. Natomiast Twierdzenie 1.6 zachodzi również dla W będącego odcinkiem domkniętym i $x_0 \in \partial W$.

Poniższe twierdzenie uogólnia warunek dostateczny II rzędu.

Twierdzenie 1.7. *Jeśli funkcja f jest klasy C^k na podzbiorze otwartym $W \subset \mathbb{R}$ i zachodzi $f'(x_0) = f''(x_0) = \cdots = f^{(k-1)}(x_0) = 0$ oraz $f^{(k)}(x_0) \neq 0$ w pewnym $x_0 \in W$, to:*

I) Jeśli k jest nieparzyste, to funkcja f nie posiada w punkcie x_0 ekstremum lokalnego.

II) Jeśli k jest parzyste oraz:

(a) $f^{(k)}(x_0) > 0$, to punkt x_0 jest ścisłym minimum lokalnym f,
(b) $f^{(k)}(x_0) < 0$, to punkt x_0 jest ścisłym maksimum lokalnym f.

1.4 Wzory Taylora

W tym podrozdziale przypomnimy wyniki, których będziemy używać w wielu dowodach w trakcie tego wykładu. Skorzystamy z nich również, aby przedstawić zwięzłe dowody twierdzeń 1.5-1.7.

Twierdzenie 1.8 (Twierdzenie o wartości średniej)**.** *Jeśli funkcja $f : [a, b] \to \mathbb{R}$ jest ciągła na $[a, b]$ i różniczkowalna na (a, b), to istnieje taki punkt $x \in (a, b)$, że*

$$f(b) - f(a) = f'(x)(b - a).$$

Zauważmy, że do prawdziwości powyższego twierdzenia nie jest konieczna ciągłość pierwszej pochodnej (w zadaniu 1.7 pokazujemy, że różniczkowalność nie musi pociągać ciągłości pochodnej).

Dowód twierdzenia 1.8. Niech $g(x) = [f(b) - f(a)]x - (b - a)f(x)$ dla $x \in [a, b]$. Wówczas g jest ciągła na $[a, b]$, różniczkowalna w (a, b) oraz

$$g(a) = f(b)a - f(a)b = g(b).$$

Pokażemy teraz, że istnieje punkt $x_0 \in (a, b)$, w którym pochodna g się zeruje. Jeśli g jest funkcją stałą, to dla dowolnego $x_0 \in (a, b)$ mamy $g'(x_0) = 0$. W przeciwnym przypadku, na mocy twierdzenia 1.1 funkcja g przyjmuje swoje kresy na $[a, b]$. Jeden z kresów jest różny od $g(a) = g(b)$. Zatem jest on przyjmowany w punkcie x_0 we wnętrzu przedziału $[a, b]$. Korzystając z twierdzenia 1.4 wnioskujemy, że $g'(x_0) = 0$. Różniczkując g otrzymujemy:

$$0 = g'(x_0) = [f(b) - f(a)] - (b - a)f'(x).$$

Po prostych przekształceniach otrzymujemy poszukiwany wzór. □

Twierdzenie 1.9 (Twierdzenie Taylora z resztą w postaci Peano)**.** *Niech $f : [a, b] \to \mathbb{R}$ będzie funkcją klasy C^1 na $[a, b]$ oraz dwukrotnie różniczkowalna w pewnym $x_0 \in (a, b)$. Wówczas dla $x \in [a, b]$ zachodzi następujący wzór:*

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2 + o((x - x_0)^2).$$

Uwaga 1.3. W sformułowaniu powyższego twierdzenia użyliśmy notacji *małe o*. Rozumieć ją należy następująco:

$$R(x) = f(x) - f(x_0) - f'(x_0)(x - x_0) - \frac{f''(x_0)}{2}(x - x_0)^2$$

jest rzędu mniejszego niż $(x - x_0)^2$, tzn.

$$\lim_{x \to x_0} \frac{R(x)}{(x - x_0)^2} = 0.$$

Przykład 1.3. Innym zastosowaniem powyższej notacji jest definicja pochodnej. Pochodną funkcji f w punkcie x_0 nazywamy taką liczbę $\alpha \in \mathbb{R}$, że

$$f(x) = f(x_0) + \alpha(x - x_0) + o(x - x_0).$$

Dowód twierdzenia 1.9. Bez straty ogólności możemy założyć $x_0 = 0$. Musimy wykazać, że $R(x) := f(x) - f(0) - f'(0)x - \frac{f''(0)}{2}x^2$ jest niższego rzędu niż x^2, tzn. $R(x) = o(x^2)$. Z ciągłości pierwszej pochodnej f dostajemy

$$f(x) - f(0) = \int_0^x f'(y)dy.$$

Wiemy, że f' jest różniczkowalna w 0. Zatem $f'(y) = f'(0) + f''(0)y + r(y)$, gdzie $r(y) = o(y)$. Oznacza to, że

$$\lim_{y\to 0} \frac{r(y)}{y} = 0.$$

Dla dowolnego $\varepsilon > 0$ istnieje zatem $\delta > 0$, taka że $|y| < \delta$ pociąga $|r(y)| < \varepsilon|y|$.

Ustalmy zatem dowolny $\varepsilon > 0$ i związaną z nim $\delta > 0$. Weźmy $|x| < \delta$ i scałkujmy pochodną $f'(y)$. Otrzymamy wówczas:

$$f(x) - f(0) = \int_0^x (f'(0) + f''(0)y + r(y))dy = f'(0)x + \frac{f''(0)}{2}x^2 + \int_0^x r(y)dy,$$

czyli $R(x) = \int_0^x r(y)dy$. Korzystając z oszacowania $|r(y)| < \varepsilon|y|$ dla $|y| < \delta$ dostajemy

$$|R(x)| \leqslant \int_0^x |r(y)|dy < \int_0^{|x|} \varepsilon|y|dy = \frac{\varepsilon x^2}{2}.$$

A zatem

$$\left|\frac{R(x)}{x^2}\right| < \frac{\varepsilon}{2}.$$

Z dowolności $\varepsilon > 0$ wynika, iż $R(x) = o(x^2)$. □

Uwaga 1.4. Twierdzenie 1.9 można uogólnić do dowolnie długiej aproksymacji Taylora. Dowód przebiega wówczas podobnie, lecz jest nieznacznie dłuższy.

Zakładając większą gładkość funkcji f możemy opisać dokładniej błąd aproksymacji we wzorze Taylora.

Twierdzenie 1.10 (Twierdzenie Taylora z resztą w postaci Lagrange'a)**.** *Niech $f : [a, b] \to \mathbb{R}$ będzie funkcją klasy C^{k-1} na $[a, b]$ oraz k-krotnie różniczkowalna na (a, b). Wtedy dla ustalonego $x_0 \in (a, b)$ i $x \in [a, b]$ zachodzi następujący wzór:*

$$f(x) = f(x_0) + \sum_{i=1}^{k-1} \frac{f^{(i)}(x_0)}{i!}(x - x_0)^i + \frac{f^{(k)}(\tilde{x})}{k!}(x - x_0)^k,$$

gdzie $\tilde{x}$ jest pewnym punktem pomiędzy x_0 i x.

W szczególności dla $k = 2$ dostajemy

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(\tilde{x})(x - x_0)^2.$$

Dowód twierdzenia 1.10. Niech liczba M spełnia równanie

$$f(x) = f(x_0) + \sum_{i=1}^{k-1} \frac{f^{(i)}(x_0)}{i!}(x - x_0)^i + M(x - x_0)^k.$$

Celem dowodu jest pokazanie, że $M = \frac{f^{(k)}(\tilde{x})}{k!}$ dla pewnego punktu $\tilde{x}$ pomiędzy x_0 i x. Określmy funkcję

$$g(y) = f(y) - \sum_{i=1}^{k-1} \frac{f^{(i)}(x_0)}{i!}(y - x_0)^i - M(y - x_0)^k, \qquad y \in [x_0, x].$$

Zauważmy, że

$$g(x_0) = g'(x_0) = \cdots = g^{(k-1)}(x_0) = 0.$$

Ponieważ $g(x) = 0$, to na podstawie twierdzenia 1.8 istnieje $x_1 \in (x_0, x)$, taki że $g'(x_1) = 0$. Stosując jeszcze raz tw. 1.8 do funkcji $g'(y)$ dla $y \in [x_0, x_1]$ dostajemy $x_2 \in (x_0, x_1)$, w którym $g''(x_2) = 0$. Postępując w ten sposób dostajemy ciąg punktów $x > x_1 > x_2 > \cdots > x_k > x_0$, takich że $g^{(j)}(x_j) = 0$, $j = 1, \ldots, k$. Z warunku dla punktu x_k otrzymujemy

$$0 = g^{(k)}(x_k) = f^{(k)}(x_k) - k!M.$$

Szukanym punktem $\tilde{x}$ w twierdzeniu jest więc x_k. □

1.5 Ekstrema globalne

Uzupełnimy jeszcze twierdzenie 1.6 o wynik dotyczący ekstremów globalnych.

Niech $I \subset \mathbb{R}$ będzie odcinkiem otwartym, domkniętym, lub jednostronnie domkniętym (być może nieograniczonym) i niech $f : I \to \mathbb{R}$ będzie funkcją klasy C^1 na I i klasy C^2 na $\operatorname{int} I$. Zachodzi następujące

Twierdzenie 1.11. *Przy powyższych założeniach, jeśli $f'(x_0) = 0$ oraz:*

(a) $f''(x) \geqslant 0 \quad \forall_{x \in I} \quad \Longrightarrow \quad x_0$ *jest globalnym minimum na* I,

(b) $f''(x) \leqslant 0 \quad \forall_{x \in I} \quad \Longrightarrow \quad x_0$ *jest globalnym maksimum na* I.

Jeśli założenia powyższe uzupełnimy o warunek $f''(x_0) > 0$ (odpowiednio $f''(x_0) < 0$), to x_0 będzie ścisłym globalnym minimum (maksimum).

Dowód. Wzór Taylora, tw. 1.10, daje

$$f(x) = f(x_0) + \frac{1}{2}f''(\tilde{x})(x - x_0)^2,$$

gdzie $\tilde{x}$ jest pewnym punktem pomiędzy x_0 i x. Stąd drugi człon wzoru Taylora decyduje o nierówności pomiędzy $f(x)$ a $f(x_0)$ i otrzymujemy obie implikacje w twierdzeniu dotyczące słabych ekstremów.

Załóżmy dodatkowo w pierwszym stwierdzeniu, że $f''(x_0) > 0$. Z założenia $f''(x) \geqslant 0$ i warunku $f'(x_0) = 0$ dostajemy

$$f'(x) = f'(x) - f'(x_0) = \int_{x_0}^{x} f''(y)dy \geqslant 0,$$

gdy $x > x_0$. Podobnie pokazujemy, że $f'(x) = -\int_x^{x_0} f''(y)dy \leqslant 0$, gdy $x < x_0$. Z faktu, że $f''(x_0) > 0$ i ciągłości drugiej pochodnej dostajemy dodatkowo, że ta pochodna jest ściśle

dodatnia w otoczeniu x_0. Zatem całki są dodatnie, co pociąga nierówności $f'(x) > 0$, gdy $x > x_0$, oraz $f'(x) < 0$, gdy $x < x_0$. Wynika stąd, że funkcja f jest ściśle rosnąca na prawo od x_0 i ściśle malejąca na lewo od x_0, a więc x_0 jest ścisłym minimum. Przypadek ścisłego maksimum dowodzimy analogicznie. □

1.6 Zadania

Ćwiczenie 1.1. Czy funkcja $f(x,y) = x^4 + x^2 - xy + 2y^2$ osiąga minimum na zbiorze $\{(x,y) \in \mathbb{R}^2 : \ x \geqslant -1\}$.

Ćwiczenie 1.2. Znajdź minimum funkcji $f(x,y) = 2x + 3y$ na zbiorze $W = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$.

Ćwiczenie 1.3. Znajdź maksimum funkcji $f(x,y) = x^2 - 4y^2$ na zbiorze $W = \{(x,y) \in \mathbb{R}^2 : 2x^2 + |y| = 1\}$.

Ćwiczenie 1.4. Znajdź minimum funkcji $f(x,y) = e^{x^2-y^2}$ na zbiorze $W = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$.

Ćwiczenie 1.5. Rozważmy następujący nieliniowy problem optymalizacyjny:

$$\begin{cases} (x_1-4)^2 + (x_2-2)^2 \to \min, \\ 4x_1^2 + 9x_2^2 \leqslant 36, \\ x_1^2 + 4x_2 = 4, \\ 2x_1 + 3 \geqslant 0. \end{cases}$$

1. Naszkicuj zbiór punktów dopuszczalnych, czyli punktów spełniających wszystkie ograniczenia.
2. Znajdź graficznie rozwiązanie powyższego problemu optymalizacyjnego.
3. Znajdź następnie rozwiązanie w przypadku, gdy minimalizacja zamieniona zostanie na maksymalizację.

Ćwiczenie 1.6. Niech g będzie funkcją spełniającą: $0 \leqslant g(y) \leqslant 1$, $y \in [0,1]$, oraz $\int_0^1 g(y)dy = 1$. Znajdź $x \in [0,1]$, dla którego następująca całka jest minimalna:

$$\int_0^1 (x-y)^2 g(y) dy.$$

Ćwiczenie 1.7. Wykaż, że funkcja

$$f(x) = \begin{cases} x^2 \sin\left(\frac{1}{x}\right), & x \neq 0, \\ 0, & x = 0, \end{cases}$$

jest różniczkowalna w $\mathbb{R}$, lecz jej pochodna nie jest ciągła.

Ćwiczenie 1.8. Udowodnij, że poniższe definicje pochodnej funkcji $f : [a,b] \to \mathbb{R}$ w punkcie $x_0 \in (a,b)$ są równoważne:

(a) $f'(x_0) = \lim_{h \to 0} \dfrac{f(x_0+h) - f(x_0)}{h}$,

(b) $f(x) = f(x_0) + \alpha(x - x_0) + o(x - x_0)$ dla $x \in [a,b]$ i $\alpha \in \mathbb{R}$ niezależnego od x.

Przez równoważność rozumiemy to, że jeśli granica w (a) istnieje, to zależność (b) jest spełniona z $\alpha = f'(x_0)$; i odwrotnie, jeśli (b) zachodzi dla pewnego α, to granica w (a) istnieje i jest równa α.

Rozdział 2

Ekstrema funkcji wielu zmiennych

2.1 Notacja i twierdzenia Taylora w wielu wymiarach

W tym podrozdziale przypomnimy krótko twierdzenia Taylora dla funkcji wielu zmiennych. Wprowadźmy najpierw niezbędną notację.
Niech $f : W \to \mathbb{R}$, gdzie $W \subset \mathbb{R}^n$ jest zbiorem otwartym. Przyjmiemy następujące oznaczenia:

- $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ – wektor kolumnowy,
- $f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n)$,
- $Df(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n}\right)$ – gradient funkcji f,
- $D^2 f(\mathbf{x})$ – Hesjan funkcji f:

$$D^2 f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}.$$

Definicja 2.1. Funkcja f jest *różniczkowalna* w punkcie $\mathbf{x}_0 \in W$, jeśli istnieje wektor $\alpha \in \mathbb{R}^n$, taki że

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \alpha^T(\mathbf{x} - \mathbf{x}_0) + o(\|\mathbf{x} - \mathbf{x}_0\|)$$

dla $\mathbf{x} \in W$.

Funkcja f jest *dwukrotnie różniczkowalna* w punkcie $\mathbf{x}_0 \in W$, jeśli istnieje wektor $\alpha \in \mathbb{R}^n$ oraz macierz $H \in \mathbb{R}^{n \times n}$, takie że

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \alpha^T(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T H(\mathbf{x} - \mathbf{x}_0) + o(\|\mathbf{x} - \mathbf{x}_0\|^2)$$

dla $\mathbf{x} \in W$.

Uwaga 2.1. Możemy założyć, że macierz H w powyższej definicji jest symetryczna. Wystarczy zauważyć, że

$$(\mathbf{x} - \mathbf{x}_0)^T H(\mathbf{x} - \mathbf{x}_0) = (\mathbf{x} - \mathbf{x}_0)^T \frac{H + H^T}{2}(\mathbf{x} - \mathbf{x}_0).$$

Twierdzenie 2.1.

I) Jeśli funkcja f jest różniczkowalna w $\mathbf{x}_0$, to $Df(\mathbf{x}_0)$ istnieje i $\alpha = Df(\mathbf{x}_0)^T$. Odwrotnie, jeśli $Df(\mathbf{x})$ istnieje w pewnym otoczeniu $\mathbf{x}_0$ i jest ciągłe w $\mathbf{x}_0$, to f jest różniczkowalna w $\mathbf{x}_0$.

II) Jeśli hesjan $D^2f(\mathbf{x})$ istnieje w pewnym otoczeniu $\mathbf{x}_0$ i jest ciągły w $\mathbf{x}_0$, to f jest dwukrotnie różniczkowalna w $\mathbf{x}_0$, $D^2f(\mathbf{x}_0)$ jest macierzą symetryczną oraz $H = D^2f(\mathbf{x}_0)$.

Dowód powyższego twierdzenia pomijamy. Zainteresowany czytelnik znajdzie go w podręcznikach analizy wielowymiarowej.

Uwaga 2.2. Ilekroć będziemy chcieli wykorzystać drugą pochodną funkcji wielowymiarowej, będziemy musieli zakładać, że hesjan D^2f jest funkcją ciągłą. Jeśli nie poczynimy takiego założenia, nie będziemy mieli dobrego sposobu na policzenie drugiej pochodnej, a zatem taki rezultat będzie miał małą wartość praktyczną.

Uwaga 2.3. Dla funkcji $f : W \to \mathbb{R}$ określonej na zbiorze otwartym $W \subset \mathbb{R}^n$ mówimy, że f jest klasy C^1 (odpowiednio, klasy C^2) i piszemy $f \in C^1$ ($f \in C^2$), gdy f jest ciągła na W oraz $\frac{\partial f}{\partial x_i}$ (odpowiednio, $\frac{\partial f}{\partial x_i}$ i $\frac{\partial^2 f}{\partial x_i \partial x_j}$) istnieją i są ciągłe na W. Gdy rozważany zbiór $W \subset \mathbb{R}^n$ nie jest otwarty, mówimy że f jest klasy C^1 (odpowiednio, klasy C^2) na W, jeśli istnieje rozszerzenie $\tilde{f}$ funkcji f do zbioru otwartego $\tilde{W}$ zawierającego W takie, że $\tilde{f}$ jest klasy C^1 (odpowiednio, klasy C^2) na $\tilde{W}$. W tym wypadku można więc mówić o pochodnych cząstkowych funkcji f również w punktach brzegowych zbioru W. Pochodne te są jednoznacznie określone przez wartości funkcji na $\operatorname{int} W$, jeśli zachodzi $W \subset \operatorname{cl}(\operatorname{int} W)$ (wynika to z ciągłości tych pochodnych).

Zapiszemy teraz rozwinięcie Taylora rzędu 2.

Lemat 2.1. *Niech $W \subset \mathbb{R}^n$ otwarty. Dla funkcji $f : W \to \mathbb{R}$ klasy C^2 i punktów $\mathbf{x}, \mathbf{x}_0 \in W$ takich, że odcinek łączący $\mathbf{x}_0$ z $\mathbf{x}$ leży w W zachodzi*

$$f(\mathbf{x}) = f(\mathbf{x}_0) + Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T D^2 f(\tilde{\mathbf{x}})(\mathbf{x} - \mathbf{x}_0),$$

gdzie $\tilde{\mathbf{x}}$ jest pewnym punktem wewnątrz odcinka łączącego $\mathbf{x}_0$ z $\mathbf{x}$.

Dowód. Dowód wynika z zastosowania twierdzenia 1.10 do funkcji $g(t) = f(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0))$, $t \in [0, 1]$. □

Definicja 2.2. Podzbiór $W \subset \mathbb{R}^n$ jest wypukły, jeśli

$$\lambda x + (1 - \lambda)y \in W$$

dla każdych $x, y \in W$ i każdego $\lambda \in [0, 1]$.

Wniosek 2.1. *Niech $W \subset \mathbb{R}^n$ zbiór otwarty, wypukły oraz $f : W \to \mathbb{R}$ klasy C^2. Wówczas dla dowolnych $\mathbf{x}_0, \mathbf{x} \in W$ mamy*

$$f(\mathbf{x}) = f(\mathbf{x}_0) + Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T D^2 f(\tilde{\mathbf{x}})(\mathbf{x} - \mathbf{x}_0),$$

gdzie $\tilde{\mathbf{x}}$ należy do wnętrza odcinka łączącego $\mathbf{x}_0$ i $\mathbf{x}$, tzn. istnieje $\lambda \in (0, 1)$, taka że $\tilde{\mathbf{x}} = \lambda \mathbf{x}_0 + (1 - \lambda)\mathbf{x}$.

Dowód. Z wypukłości W wynika, że dla każdego $\mathbf{x}_0, \mathbf{x} \in W$ odcinek łączący te punkty zawarty jest w W. Teza wynika teraz z lematu 2.1. □

2.2 Znikanie gradientu

Będziemy rozważać funkcję $f : W \to \mathbb{R}$, gdzie W jest podzbiorem w $\mathbb{R}^n$ mającym niepuste wnętrze $\operatorname{int} W$.

Twierdzenie 2.2 (Warunek konieczny I rzędu)**.** *Jeśli funkcja $f : W \to \mathbb{R}$ jest różniczkowalna w punkcie $\mathbf{x}_0$ należącym do wnętrza zbioru W oraz $\mathbf{x}_0$ jest lokalnym minimum (maksimum) funkcji f to*

$$Df(\mathbf{x}_0) = 0.$$

Dowód. Z faktu, że $\mathbf{x}_0 \in \operatorname{int} W$ wynika, że funkcja $g(t) = f(\mathbf{x}_0 + t\mathbf{e}_i)$, gdzie $\mathbf{e}_i$ jest i-tym wersorem (tj. $\mathbf{e}_i$ ma jedynkę na i-tej współrzędnej i zera poza nią), jest dobrze określona na otoczeniu 0. Ma ona również lokalne ekstremum w punkcie 0. Na mocy tw. 1.4 mamy $g'(0) = 0$. W terminach funkcji f oznacza to, że $\frac{\partial f}{\partial x_i}(\mathbf{x}_0) = 0$. Przeprowadzając to rozumowanie dla $i = 1, 2, \ldots, n$ dostajemy tezę.. □

Warunek znikania gradientu będzie często używany, zatem użyteczna będzie

Definicja 2.3. Punkt $\mathbf{x}_0 \in \operatorname{int} W$ nazywamy *punktem krytycznym* funkcji $f : W \to \mathbb{R}$, jeśli f jest różniczkowalna w $\mathbf{x}_0$ oraz $Df(\mathbf{x}_0) = 0$.

Oczywiście, warunek znikania gradientu $Df(\mathbf{x}_0)$ nie jest wystarczający na to, by w $\mathbf{x}_0$ znajdowało się lokalne minimum lub maksimum. Do rozstrzygnięcia tego jest potrzebny analog warunku o znaku drugiej pochodnej (tw. 1.6). W przypadku wielowymiarowym ten warunek definiuje się jako dodatnią (ujemną) określoność macierzy drugich pochodnych.

2.3 Dodatnia i ujemna określoność macierzy

Niech $A = \{a_{ij}\}_{i,j=1}^n$ będzie macierzą symetryczną, tzn. $a_{ij} = a_{ji}$. Rozważmy formę kwadratową

$$\mathbf{x}^T A\mathbf{x} = \sum_{i,j=1}^{n} a_{ij} x_i x_j.$$

Definicja 2.4. Określoność macierzy A lub formy kwadratowej $\mathbf{x}^T A\mathbf{x}$ definiujemy następująco:

- A jest *nieujemnie określona*, co oznaczamy $A \geqslant 0$, jeśli
$$\mathbf{x}^T A\mathbf{x} \geqslant 0 \qquad \forall\, \mathbf{x} \in \mathbb{R}^n.$$
- A jest *dodatnio określona*, co oznaczamy $A > 0$, jeśli
$$\mathbf{x}^T A\mathbf{x} > 0 \qquad \forall\, \mathbf{x} \in \mathbb{R}^n \setminus \{0\}.$$

Odwracając nierówności definiujemy *niedodatnią określoność* i *ujemną określoność*.

- Macierz A nazywamy *nieokreśloną*, jeśli istnieją wektory $\mathbf{x},\ \tilde{\mathbf{x}} \in \mathbb{R}^n$ takie, że
$$\mathbf{x}^T A\mathbf{x} > 0, \qquad \tilde{\mathbf{x}}^T A\tilde{\mathbf{x}} < 0.$$

Zauważmy, że z definicji określoności macierzy, wyliczając wyrażenie $\mathbf{e}_i^T A\mathbf{e}_i = a_{ii}$ na wersorze $\mathbf{e}_i = (0, \ldots, 1, \ldots, 0)^T$, z jedynką na i-tym miejscu, wynikają następujące warunki konieczne odpowiedniej określoności macierzy A:

- Jeśli A jest dodatnio określona, to $a_{11} > 0, \ldots, a_{nn} > 0$.
- Jeśli A jest nieujemnie określona, to $a_{11} \geqslant 0, \ldots, a_{nn} \geqslant 0$.
- Jeśli A jest ujemnie określona, to $a_{11} < 0, \ldots, a_{nn} < 0$.
- Jeśli A jest niedodatnio określona, to $a_{11} \leqslant 0, \ldots, a_{nn} \leqslant 0$.
- Jeśli $a_{ii} > 0$ i $a_{jj} < 0$, dla pewnych i, j, to A jest nieokreślona.

Warunki konieczne i dostateczne podane są w poniższym twierdzeniu, którego dowód pomijamy.

Twierdzenie 2.3 (Kryterium Sylvestera)**.**

I. Forma kwadratowa $\mathbf{x}^T A\mathbf{x}$ jest dodatnio określona wtedy i tylko wtedy, gdy zachodzi:

$$D_1 > 0,\ D_2 > 0, \ldots\ldots, D_n > 0,$$

gdzie przez $D_1, \ldots, D_n$ oznaczamy minory główne macierzy A:

$$D_1 = \det(a_{11}),\ D_2 = \det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \ldots, D_n = \det\begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \ldots & a_{nn} \end{pmatrix}.$$

Forma kwadratowa $\mathbf{x}^T A\mathbf{x}$ jest ujemnie określona wtedy i tylko wtedy, gdy $\mathbf{x}^T(-A)\mathbf{x}$ jest dodatnio określona, co przekłada się na ciąg warunków:

$$-D_1 > 0,\ D_2 > 0, \ldots\ldots, (-1)^n D_n > 0.$$

II. Forma kwadratowa $x^T Ax$ jest nieujemnie określona wtedy i tylko wtedy, gdy dla dowolnych $1 \leqslant k \leqslant n$ oraz $1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n$ zachodzi

$$\det\begin{pmatrix} a_{i_1 i_1} & a_{i_1 i_2} & \ldots & a_{i_1 i_k} \\ a_{i_2 i_1} & a_{i_2 i_2} & \ldots & a_{i_2 i_k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i_k i_1} & a_{i_k i_2} & \ldots & a_{i_k i_k} \end{pmatrix} \geqslant 0$$

(jest to minor rzędu k złożony z kolumn $i_1, \ldots, i_k$ i rzędów $i_1, \ldots, i_k$).

Określoność macierzy symetrycznej jest niezależna od bazy, w której jest reprezentowana. W bazie własnej macierz A jest diagonalna z wartościami własnymi na diagonali. Dostajemy zatem następujące warunki równoważne określoności:

- Macierz A jest dodatnio określona wtw, gdy wszystkie jej wartości własne są dodatnie.
- Macierz A jest nieujemnie określona wtw, gdy wszystkie jej wartości własne są nieujemne.
- Macierz A jest ujemnie określona wtw, gdy wszystkie jej wartości własne są ujemne.
- Macierz A jest niedodatnio określona wtw, gdy wszystkie jej wartości własne są niedodatnie.

2.4 Warunki II-go rzędu (kryterium drugiej różniczki)

Twierdzenie 2.4 (Warunek konieczny II rzędu)**.** *Jeśli f jest klasy C^2 na zbiorze otwartym $W \subset \mathbb{R}^n$ i $\mathbf{x}_0 \in W$ jest minimum lokalnym, to macierz $D^2f(\mathbf{x}_0)$ jest nieujemnie określona. Podobnie, jeśli $\mathbf{x}_0$ jest lokalnym maksimum, to $D^2f(\mathbf{x}_0)$ jest niedodatnio określona.*

Twierdzenie 2.5 (Warunek dostateczny II rzędu)**.** *Jeśli f jest klasy C^2 na zbiorze otwartym $W \subset \mathbb{R}^n$, $Df(\mathbf{x}_0) = 0$ oraz $D^2f(\mathbf{x}_0)$ jest dodatnio określona (ujemnie określona) to f ma ścisłe lokalne minimum (lokalne maksimum) w $\mathbf{x}_0$.*

Dowód twierdzenia 2.4. Niech $\mathbf{x}_0 \in W$ będzie minimum lokalnym f. Ustalmy niezerowy wektor $\mathbf{h} \in \mathbb{R}^n$ i funkcję

$$g(t) = f(\mathbf{x}_0 + t\mathbf{h}),$$

gdzie $t \in \mathbb{R}$ jest z dostatecznie małego otoczenia zera, aby $\mathbf{x}_0 + t\mathbf{h} \in W$. Wtedy funkcja g ma lokalne minimum w punkcie $t = 0$. Ponieważ f jest klasy C^2, funkcja g również jest klasy C^2. Z Twierdzenia 1.5 dla przypadku skalarnego wiemy, że skoro $t = 0$ jest lokalnym minimum, to $g''(0) \geqslant 0$. Ze wzorów na pochodną funkcji złożonej mamy

$$g''(0) = \mathbf{h}^T D^2f(\mathbf{x}_0)\mathbf{h}.$$

Z dowolności wektora $\mathbf{h}$ wynika nieujemna określoność macierzy $D^2f(\mathbf{x}_0)$. □

Dowód twierdzenia 2.5. Załóżmy najpierw, że $D^2f(\mathbf{x}_0) > 0$. Określmy funkcję $\alpha : W \to \mathbb{R}$ wzorem

$$\alpha(\mathbf{x}) = \inf_{\|\mathbf{h}\|=1} \mathbf{h}^T D^2f(\mathbf{x})\mathbf{h}.$$

Funkcja ta jest ciągła na mocy ciągłości hesjanu f oraz ćwiczenia 2.2. Istnieje zatem kula $B(\mathbf{x}_0, \varepsilon)$, taka że $\alpha(\mathbf{x}) > 0$ dla $\mathbf{x} \in B(\mathbf{x}_0, \varepsilon)$.

Ustalmy dowolny $\mathbf{x} \in B(\mathbf{x}_0, \varepsilon)$. Na mocy wzoru Taylora, lemat 2.1, mamy

$$f(\mathbf{x}) = f(\mathbf{x}_0) + Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T D^2f(\tilde{\mathbf{x}})(\mathbf{x} - \mathbf{x}_0),$$

dla pewnego punktu $\tilde{\mathbf{x}}$ leżącego na odcinku łączącym $\mathbf{x}_0$ i $\mathbf{x}$, a zatem i należącego do kuli $B(\mathbf{x}_0, \varepsilon)$. Pierwsza pochodna f znika w punkcie $\mathbf{x}_0$, zaś

$$(\mathbf{x} - \mathbf{x}_0)^T D^2f(\tilde{\mathbf{x}})(\mathbf{x} - \mathbf{x}_0) = \|\mathbf{x} - \mathbf{x}_0\|^2 \frac{(\mathbf{x} - \mathbf{x}_0)^T}{\|\mathbf{x} - \mathbf{x}_0\|} D^2f(\tilde{\mathbf{x}}) \frac{(\mathbf{x} - \mathbf{x}_0)}{\|\mathbf{x} - \mathbf{x}_0\|} \geqslant \|\mathbf{x} - \mathbf{x}_0\|^2 \alpha(\tilde{\mathbf{x}}).$$

Mamy zatem

$$f(\mathbf{x}) - f(\mathbf{x}_0) \geqslant \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|^2 \alpha(\tilde{\mathbf{x}}) > 0,$$

gdyż funkcja α jest dodatnia na kuli $B(\mathbf{x}_0, \varepsilon)$. Wnioskujemy więc, że $\mathbf{x}_0$ jest ścisłym minimum lokalnym.

Dowód przypadku $D^2f(\mathbf{x}_0) < 0$ jest analogiczny. □

2.4.1 Ekstrema globalne i określoność drugiej różniczki

Niech teraz $f : W \to \mathbb{R}$ będzie funkcją klasy C^1 na zbiorze wypukłym $W \in \mathbb{R}^n$, oraz klasy C^2 na $\operatorname{int} W$.

Twierdzenie 2.6. *Jeśli* $\mathbf{x}_0 \in \operatorname{int} W$ *jest punktem krytycznym* f*, to:*

I) $D^2 f(\mathbf{x}) \geqslant 0 \quad \forall_{\mathbf{x}\in \operatorname{int} W} \quad \Longrightarrow \quad \mathbf{x}_0$ *jest globalnym minimum,*

II) $D^2 f(\mathbf{x}) \leqslant 0 \quad \forall_{\mathbf{x}\in \operatorname{int} W} \quad \Longrightarrow \quad \mathbf{x}_0$ *jest globalnym maksimum.*

Jeśli dodatkowo $D^2 f(\mathbf{x}_0) > 0$ *w pierwszym stwierdzeniu (*$D^2 f(\mathbf{x}_0) < 0$ *w drugim stwierdzeniu), to* $\mathbf{x}_0$ *jest ścisłym globalnym minimum (maksimum).*

Dowód. Jeśli $\mathbf{x} \in W$, to z wypukłości W cały odcinek łączący $\mathbf{x}_0$ z $\mathbf{x}$ (poza punktem $\mathbf{x}$) leży w $\operatorname{int} W$ i możemy zastosować wzór Taylora, lemat 2.1, który daje

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T D^2 f(\tilde{\mathbf{x}})(\mathbf{x}-\mathbf{x}_0),$$

gdzie $\tilde{\mathbf{x}}$ jest pewnym punktem z odcinka łączącego $\mathbf{x}_0$ z $\mathbf{x}$. Nierówność $D^2 f(\tilde{\mathbf{x}}) \geqslant 0$ (odpowiednio, $D^2 f(\tilde{\mathbf{x}}) \leqslant 0$) oznacza, że drugi człon w powyższym wzorze jest nieujemny (niedodatni), co pociąga obie implikacje w twierdzeniu.

W przypadku, gdy w (I) mamy dodatkowo $D^2 f(\mathbf{x}_0) > 0$, odwołamy się do używanej już funkcji $g(t) = f(\mathbf{x}_0 + t(\mathbf{x}-\mathbf{x}_0))$, $t \in [0,1]$. Z wypukłości W wynika, że g jest dobrze określona, tzn. $\mathbf{x}_0 + t(\mathbf{x}-\mathbf{x}_0) \in W$ dla $t \in [0,1]$. Nasze założenia implikują, że $g'(0) = 0$, $g''(0) > 0$ oraz $g''(t) \geqslant 0$. Możemy skorzystać z tw. 1.11, które stwierdza, że g ma ścisłe globalne minimum w $t = 0$. Zatem $g(1) > g(0)$, czyli $f(\mathbf{x}) > f(\mathbf{x}_0)$. Z dowolności $\mathbf{x}$ wynika, iż $\mathbf{x}_0$ jest ścisłym minimum globalnym.

Przypadek $D^2 f(\mathbf{x}_0) < 0$ w stwierdzeniu (II) dowodzimy analogicznie. □

2.5 Zadania

Ćwiczenie 2.1. Wykaż, że hesjan funkcji

$$f(x_1,x_2) = \begin{cases} 0, & x_1 = x_2 = 0, \\ \frac{x_1 x_2 (x_1^2 - x_2^2)}{x_1^2 + x_2^2}, & \text{w p.p.}, \end{cases}$$

nie jest symetryczny w punkcie $(0,0)$.

Ćwiczenie 2.2. Niech $W \subset \mathbb{R}^k$, $A \subset \mathbb{R}^n$ zwarty oraz $f : W \times A \to \mathbb{R}$ ciągła. Udowodnij, że funkcja $g : W \to \mathbb{R}$ zadana wzorem

$$g(\mathbf{x}) = \inf_{\mathbf{y}\in A} f(\mathbf{x},\mathbf{y})$$

jest ciągła.

Ćwiczenie 2.3. Pochodną kierunkową funkcji f w punkcie $\bar{\mathbf{x}}$ i kierunku $\mathbf{d}$ nazywamy granicę

$$D_{\mathbf{d}} f(\bar{\mathbf{x}}) = \lim_{h\to 0} \frac{f(\bar{\mathbf{x}} + h\mathbf{d}) - f(\bar{\mathbf{x}})}{h}.$$

Udowodnij, że $\max_{\|\mathbf{d}\|=1} \|D_{\mathbf{d}} f(\bar{\mathbf{x}})\|$ jest przyjmowane dla $\mathbf{d} = Df(\bar{\mathbf{x}})/\|Df(\bar{\mathbf{x}})\|$.

Ćwiczenie 2.4. Rozważmy następującą funkcję (czasami zwaną funkcją Peano):

$$f(x_1,x_2) = (x_2^2 - x_1)(x_2^2 - 2x_1).$$

1. Udowodnij, że funkcja f ograniczona do każdej prostej przechodzącej przez $\mathbf{0}$ ma w tym punkcie minimum lokalne.
2. Wykaż, że f jako funkcja wielu zmiennych nie ma ekstremum lokalnego w $\mathbf{0}$.
3. Znajdź wartości własne macierzy drugiej pochodnej f. Co możesz z nich wywnioskować? Czy tłumaczą one zachowanie funkcji f w $\mathbf{0}$?

Ćwiczenie 2.5. Rozważmy funkcję kwadratową wielu zmiennych:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x} + c,$$

gdzie A jest macierzą kwadratową, niekoniecznie symetryczną, $\mathbf{b}$ jest wektorem, zaś c stałą. Wyznacz gradient i hesjan (macierz drugiej pochodnej) funkcji f.

Wskazówka. Załóż najpierw, że A jest symetryczna. Udowodnij później, że dla każdej macierzy kwadratowej A istnieje macierz symetryczna $\hat{A}$, taka że $\mathbf{x}^T\hat{A}\mathbf{x} = \mathbf{x}^T A\mathbf{x}$ dla każdego $\mathbf{x}$.

Ćwiczenie 2.6. Zbadaj określoność następujących macierzy i porównaj wyniki z ich formą zdiagonalizowaną:

$$\begin{bmatrix} -3 & 1 \\ 1 & -2 \end{bmatrix}, \qquad \begin{bmatrix} 3 & 1 \\ 1 & -2 \end{bmatrix}, \qquad \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}, \qquad \begin{bmatrix} 2 & -2 & 0 \\ -2 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

Ćwiczenie 2.7. Znajdź ekstrema globalne funkcji

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \begin{bmatrix} -3 & 1 \\ 1 & -2 \end{bmatrix} \mathbf{x} + [2, 1]\mathbf{x} + 17.$$

Ćwiczenie 2.8. Niech $(\Omega, \mathcal{F}, \mathbb{P})$ będzie przestrzenią probabilistyczną, co między innymi oznacza, że $\mathbb{P}(\Omega) = 1$. Dana jest zmienna losowa $\eta \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, tzn. funkcja mierzalna $\eta : \Omega \to \mathbb{R}^n$ o tej własności, że $\mathbb{E}\|\eta\|^2 < \infty$. Znajdź wektor $\bar{\mathbf{x}} \in \mathbb{R}^n$, taki że $\mathbb{E}\|\eta - \mathbf{x}\|^2$ jest najmniejsza.

Wskazówka. Zapisz $\mathbb{E}\|\eta - \mathbf{x}\|^2$ jako funkcję kwadratową.

Ćwiczenie 2.9. Niech $f : \mathbb{R}^n \to \mathbb{R}$ i $\bar{\mathbf{x}} \in \mathbb{R}^n$. Załóżmy, że f jest klasy C^2 na otoczeniu $\bar{\mathbf{x}}$ oraz $Df(\bar{\mathbf{x}}) = \mathbf{0}^T$. Udowodnij, że jeśli macierz $D^2 f(\bar{\mathbf{x}})$ jest nieokreślona, to f nie ma ekstremum lokalnego w $\bar{\mathbf{x}}$.

Ćwiczenie 2.10. Udowodnij nierówność średnich rozwiązując zadanie optymalizacyjne:

$$\begin{cases} xytz \to \max, \\ x + y + t + z = 4c, \\ x, y, t, z \in [0, \infty). \end{cases}$$

Ćwiczenie 2.11. Znajdź minima lokalne funkcji

$$f(x, y) = \frac{1}{4}x^4 + \frac{1}{3}x^3 - 2xy + y^2 + 2x - 2y + 1.$$

Rozdział 3

Funkcje wypukłe

3.1 Zbiory wypukłe i twierdzenia o oddzielaniu

Przypomnijmy, za def. 2.2, definicję zbioru wypukłego: zbiór $W \subset \mathbb{R}^n$ jest wypukły, jeśli

$$\lambda \mathbf{x} + (1-\lambda)\mathbf{y} \in W$$

dla każdych $\mathbf{x}, \mathbf{y} \in W$ i każdego $\lambda \in [0,1]$. Równoważnie można wypukłość zdefiniować za pomocą m-tek punktów:

Lemat 3.1. *Zbiór $W \subset \mathbb{R}^n$ jest wypukły wtw, gdy dla dowolnego $m \geqslant 2$, punktów $\mathbf{x}_1, \ldots, \mathbf{x}_m \in W$ oraz liczb $a_1, \ldots, a_m \geqslant 0$, takich że $a_1 + \ldots + a_m = 1$ mamy*

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \ldots + a_m\mathbf{x}_m \in W.$$

Dowód tego lematu pozostawiamy jako ćwiczenie (patrz ćw. 3.3). Udowodnimy natomiast geometryczną własność zbiorów wypukłych, która przyda nam się wielokrotnie.

Lemat 3.2. *Niech $W \subset \mathbb{R}^n$ będzie zbiorem wypukłym o niepustym wnętrzu. Wówczas:*

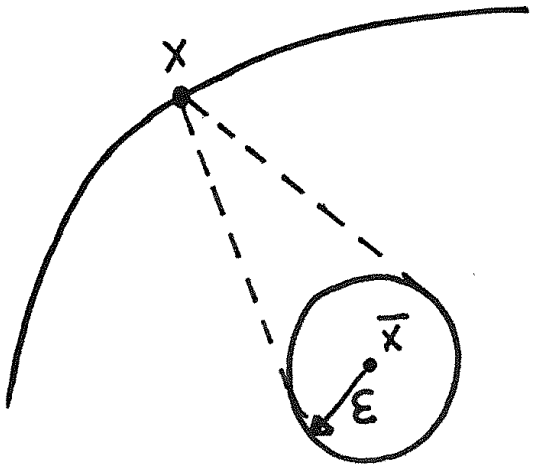
I) Dla dowolnego $\mathbf{x} \in W$ oraz $\mathbf{x}_0 \in \operatorname{int} W$ odcinek łączący $\mathbf{x}_0$ z $\mathbf{x}$, z pominięciem punktu $\mathbf{x}$, należy do wnętrza W:

$$\lambda \mathbf{x}_0 + (1-\lambda)\mathbf{x} \in \operatorname{int} W, \qquad \forall\ 0 < \lambda \leqslant 1.$$

II) $W \subset \operatorname{cl}(\operatorname{int} W)$.

Dowód. Weźmy punkty $\mathbf{x}_0$ i $\mathbf{x}$ jak w założeniach lematu. Z otwartości $\operatorname{int} W$ wynika, że istnieje kula $B(\mathbf{x}_0, \varepsilon) \subset \operatorname{int} W$. Połączmy punkty tej kuli z punktem $\mathbf{x}$. Dostaniemy „stożek" o wierzchołku $\mathbf{x}$ i podstawie $B(\mathbf{x}_0, \varepsilon)$ (patrz rys. 3.1). Stożek ten leży w całości w W. Jego wnętrze zawiera odcinek od $\mathbf{x}_0$ do $\mathbf{x}$ bez końca $\mathbf{x}$. Kończy to zatem dowód (I). Dowód (II) wynika natychmiast z (I). □

Przypomnijmy twierdzenia o oddzielaniu zbiorów wypukłych w przestrzeniach $\mathbb{R}^n$ (dowody tych twierdzeń można znaleźć np. w rozdziale 2.4 monografii Bazaraa, Sherali, Shetty [3] lub w rozdziale 11 monografii Rockafellara [11]).



Rysunek 3.1: Stożek o wierzchołku $\mathbf{x}$ i podstawie $B(\bar{\mathbf{x}}, \varepsilon)$.

Twierdzenie 3.1 (Twierdzenie o oddzielaniu)**.** *Niech $U, V \subset \mathbb{R}^n$ będą niepustymi zbiorami wypukłymi, takimi że $U \cap V = \emptyset$. Wówczas istnieje hiperpłaszczyzna rozdzielająca U od V, tzn. istnieje niezerowy wektor $\mathbf{a} \in \mathbb{R}^n$ spełniający*

$$\mathbf{a}^T\mathbf{x} \leqslant \mathbf{a}^T\mathbf{y}, \qquad \mathbf{x} \in U,\ \mathbf{y} \in V.$$

Korzystając z ciągłości odwzorowania liniowego $\mathbf{x} \mapsto \mathbf{a}^T\mathbf{x}$ dostajemy bardzo przydatny wniosek, który również będziemy nazywać twierdzeniem o oddzielaniu.

Wniosek 3.1. *Niech $U, V \subset \mathbb{R}^n$ będą niepustymi zbiorami wypukłymi, takimi że $\operatorname{int} U \neq \emptyset$ i $(\operatorname{int} U) \cap V = \emptyset$. Wówczas istnieje hiperpłaszczyzna rozdzielająca U od V, tzn. istnieje niezerowy wektor $\mathbf{a} \in \mathbb{R}^n$ spełniający*

$$\mathbf{a}^T\mathbf{x} \leqslant \mathbf{a}^T\mathbf{y}, \qquad \mathbf{x} \in U,\ \mathbf{y} \in V.$$

Twierdzenie 3.2 (Twierdzenie o ostrym oddzielaniu)**.** *Niech $U, V \subset \mathbb{R}^n$ będą niepustymi zbiorami wypukłymi domkniętymi, U zwarty i $U \cap V = \emptyset$. Wówczas istnieje hiperpłaszczyzna ściśle rozdzielająca U od V, tzn. istnieje niezerowy wektor $\mathbf{a} \in \mathbb{R}^n$ spełniający*

$$\sup_{\mathbf{x} \in U} \mathbf{a}^T\mathbf{x} < \inf_{\mathbf{y} \in V} \mathbf{a}^T\mathbf{y}.$$

Uwaga 3.1. Powyższe twierdzenia mają intuicyjną geometryczną interpretację. Dwa rozłączne zbiory wypukłe mogą być rozdzielone hiperpłaszczyzną w ten sposób, iż jeden z nich leży w domkniętej półprzestrzeni po jednej stronie hiperpłaszczyzny, podczas gdy drugi leży po przeciwnej stronie tejże hiperpłaszczyzny. W twierdzeniu o oddzielaniu nie możemy zagwarantować, iż któryś ze zbiorów leży we wnętrzu półprzestrzeni, tzn. ma puste przecięcie z hiperpłaszczyzną. Wersja o ostrym oddzielaniu właśnie to gwarantuje.

Hiperpłaszczyzna, o której mowa w twierdzeniach zadana jest wzorem:

$$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^T\mathbf{x} = \alpha\},$$

gdzie $\alpha = \sup_{\mathbf{x} \in U} \mathbf{a}^T\mathbf{x}$. Nie musi to być jedyna hiperpłaszczyzna spełniająca warunki rozdzielania lub ostrego rozdzielenia.

Twierdzenia o oddzielaniu będziemy dowodzić w przeciwnej kolejności niż są podane. Okazuje się bowiem, że łatwiej udowodnić twierdzenie o ostrym oddzielaniu. Później w dowodzie twierdzenia o oddzielaniu skorzystamy z ostrego oddzielania zbiorów wypukłych.

Dowód twierdzenia 3.2. Określmy funkcję $d : U \times V \to \mathbb{R}$ wzorem $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$. Z ograniczoności zbioru U wynika, że d jest funkcją koercywną; zbieżność do nieskończoności może się tylko odbywać po argumencie ze zbioru V. Na mocy tw. 1.2 funkcja d jako ciągła i koercywna określona na zbiorze domkniętym $U \times V$ osiąga swoje minimum w pewnym punkcie $(\mathbf{x}_0, \mathbf{y}_0) \in U \times V$. Z faktu, że $U \cap V = \emptyset$ wynika, że $\mathbf{x}_0 \neq \mathbf{y}_0$. Połóżmy $\mathbf{a} = \mathbf{y}_0 - \mathbf{x}_0$. Pokażemy, że jest to szukany wektor z tezy twierdzenia.

Udowodnimy, że $\mathbf{a}^T\mathbf{y} \geqslant \mathbf{a}^T\mathbf{y}_0$. Niech $\mathbf{y} \in V$, $\mathbf{y} \neq \mathbf{y}_0$. Zdefiniujmy funkcję

$$g(t) = \Big(d(\mathbf{x}_0, \mathbf{y}_0 + t(\mathbf{y} - \mathbf{y}_0))\Big)^2, \qquad t \in \mathbb{R}.$$

Rozwijając dostajemy

$$g(t) = \|\mathbf{y}_0 - \mathbf{x}_0\|^2 + 2t(\mathbf{y}_0 - \mathbf{x}_0)^T(\mathbf{y} - \mathbf{y}_0) + t^2(\mathbf{y} - \mathbf{y}_0)^T(\mathbf{y} - \mathbf{y}_0).$$

Funkcja ta jest różniczkowalna dla $t \in \mathbb{R}$ i $g(0) \leqslant g(t)$ dla $t \in [0, 1]$ (na mocy wypukłości V i definicji $\mathbf{y}_0$). Zatem $g'(0) \geqslant 0$, czyli

$$(\mathbf{y}_0 - \mathbf{x}_0)^T(\mathbf{y} - \mathbf{y}_0) \geqslant 0.$$

Nierówność ta jest równoważna $\mathbf{a}^T\mathbf{y} \geqslant \mathbf{a}^T\mathbf{y}_0$.

Podobnie pokazujemy, że $\mathbf{a}^T\mathbf{x} \leqslant \mathbf{a}^T\mathbf{x}_0$ dla $\mathbf{x} \in U$. Do zakończenia dowodu wystarczy sprawdzić, że $\mathbf{a}^T\mathbf{y}_0 > \mathbf{a}^T\mathbf{x}_0$. Ta nierówność jest równoważna $\|\mathbf{a}\|^2 > 0$, co zachodzi, gdyż $\mathbf{x}_0 \neq \mathbf{y}_0$. □

Dowód twierdzenia 3.1. Rozważmy zbiór $C = V - U = \{\mathbf{y} - \mathbf{x} : \ \mathbf{x} \in U, \ \mathbf{y} \in V\}$. Zbiór ten jest wypukły i $\mathbf{0} \notin C$. Równoważne tezie twierdzenia jest znalezienie wektora niezerowego $\mathbf{a} \in \mathbb{R}^n$ oddzielającego C od $\{\mathbf{0}\}$, tzn. takiego że $\mathbf{a}^T\mathbf{x} \geqslant 0$ dla $\mathbf{x} \in C$.

Zdefiniujmy zbiory

$$A_{\mathbf{x}} = \{\mathbf{a} \in \mathbb{R}^n : \ \|\mathbf{a}\| = 1, \ \mathbf{a}^T\mathbf{x} \geqslant 0\}.$$

Wystarczy pokazać, że $\bigcap_{\mathbf{x} \in C} A_{\mathbf{x}} \neq \emptyset$. Będziemy rozumować przez sprzeczność. Załóżmy więc, że $\bigcap_{\mathbf{x} \in C} A_{\mathbf{x}} = \emptyset$. Niech $B_{\mathbf{x}} = S \setminus A_{\mathbf{x}}$, gdzie S jest sferą jednostkową $S = \{\mathbf{a} \in \mathbb{R}^n : \ \|\mathbf{a}\| = 1\}$. Zbiory $B_{\mathbf{x}}$, $\mathbf{x} \in C$, są otwartymi podzbiorami zbioru zwartego S. Z założenia, że przecięcie ich dopełnień w S jest puste, wynika, że rodzina $\{B_{\mathbf{x}}\}_{\mathbf{x} \in C}$ jest pokryciem otwartym S. Na mocy zwartości S istnieje podpokrycie skończone $B_{\mathbf{x}_1}, \ldots, B_{\mathbf{x}_k}$, czyli

$$A_{\mathbf{x}_1} \cap \cdots \cap A_{\mathbf{x}_k} = \emptyset.$$

Połóżmy

$$\hat{C} = \operatorname{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_k\} = \{\sum_{i=1}^k \lambda_i \mathbf{x}_i : \ \lambda_1, \ldots, \lambda_k \geqslant 0, \ \sum_{i=1}^k \lambda_i = 1\}.$$

Zbiór $\hat{C}$ jest wypukły i domknięty oraz $\hat{C} \subset C$. Stąd $\mathbf{0} \notin \hat{C}$ i na mocy twierdzenia o ostrym oddzielaniu zastosowanego do $\hat{C}$ i $\{\mathbf{0}\}$ istnieje wektor $\mathbf{a} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, taki że

$$\mathbf{a}^T\mathbf{x} > 0, \qquad \mathbf{x} \in \hat{C}.$$

W szczególności, $\mathbf{a}^T\mathbf{x}_1 > 0, \ldots, \mathbf{a}^T\mathbf{x}_k > 0$, czyli $\frac{\mathbf{a}}{\|\mathbf{a}\|} \in A_{\mathbf{x}_i}$, $i = 1, \ldots, k$, co przeczy założeniu, że przecięcie $\bigcap_{i=1}^k A_{\mathbf{x}_i} = \emptyset$. □

3.2 Definicja funkcji wypukłej

Definicja 3.1. Funkcję $f : W \to \mathbb{R}$, gdzie $W \subset \mathbb{R}^n$ wypukły, nazwiemy:

- *wypukłą*, jeśli dla każdego $\mathbf{x}, \mathbf{y} \in W$ i $\lambda \in (0,1)$ zachodzi
$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leqslant \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}).$$
- *ściśle wypukłą*, jeśli dla każdego $\mathbf{x}, \mathbf{y} \in W$, $\mathbf{x} \neq \mathbf{y}$ i $\lambda \in (0,1)$ zachodzi
$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}).$$

Funkcja f jest (ściśle) wklęsła, jeśli $(-f)$ jest (ściśle) wypukła.

Okazuje się, że wystarczy rozważać $\lambda = 1/2$ w definicji wypukłości. Zacytujemy najpierw silny wynik noszący nazwisko Sierpińskiego, a później łatwiejszy, który dowiedziemy:

Twierdzenie 3.3. *Jeśli funkcja $f : W \to \mathbb{R}$, gdzie $W \subset \mathbb{R}^n$ wypukły, jest mierzalna w sensie Lebesgue'a oraz spełnia*
$$f\Big(\frac{\mathbf{x}+\mathbf{y}}{2}\Big) \leqslant \frac{f(\mathbf{x}) + f(\mathbf{y})}{2},$$
to jest wypukła.

Dowód. Patrz dowód tw. Sierpińskiego w [7, str. 12]. □

My udowodnimy słabszy wynik; założymy mianowicie ciągłość f.

Twierdzenie 3.4. *Jeśli funkcja $f : W \to \mathbb{R}$, gdzie $W \subset \mathbb{R}^n$ wypukły, jest ciągła oraz spełnia*
$$f\Big(\frac{\mathbf{x}+\mathbf{y}}{2}\Big) \leqslant \frac{f(\mathbf{x}) + f(\mathbf{y})}{2},$$
to jest wypukła.

Dowód. Pokażemy najpierw, przez indukcję po k, że nierówność wypukłości zachodzi dla λ postaci $p/2^k$, $p = 0, 1, \ldots, 2^k$. Własność ta jest spełniona dla $k = 1$ z założenia twierdzenia. Przeprowadźmy teraz krok indukcyjny. Załóżmy, że nierówność jest prawdziwa dla k. Weźmy $p, q > 0$ całkowite, o sumie $p+q = 2^{k+1}$. Załóżmy $p \leqslant q$. Wówczas $p \leqslant 2^k \leqslant q$ i możemy napisać:
$$\mathbf{z} = \frac{p}{2^{k+1}}\mathbf{x} + \frac{q}{2^{k+1}}\mathbf{y} = \frac{1}{2}\Big(\frac{p}{2^k}\mathbf{x} + \frac{q-2^k}{2^k}\mathbf{y} + \mathbf{y}\Big).$$
Mamy zatem
$$f(\mathbf{z}) \leqslant \frac{1}{2} f\Big(\frac{p}{2^k}\mathbf{x} + \frac{q-2^k}{2^k}\mathbf{y}\Big) + \frac{1}{2} f(\mathbf{y}) \leqslant \frac{1}{2}\frac{p}{2^k} f(\mathbf{x}) + \frac{1}{2}\frac{q-2^k}{2^k} f(\mathbf{y}) + \frac{1}{2} f(\mathbf{y}) = \frac{p}{2^{k+1}} f(\mathbf{x}) + \frac{q}{2^{k+1}} f(\mathbf{y}).$$
Pierwsza nierówność wynika z założenia twierdzenia, zaś druga – z założenia indukcyjnego. Dowód w przypadku $p \geqslant q$ jest symetryczny, ze zmienioną rolą $\mathbf{x}$ i $\mathbf{y}$.

Zbiór punktów postaci $p/2^k$, $k = 1, 2, \ldots$, jest gęsty w odcinku $(0,1)$. Z ciągłości funkcji f otrzymujemy nierówność wypukłości dla dowolnego $\lambda \in (0,1)$. □

Przykład 3.1.

- Funkcja afiniczna $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$ jest wypukła i wklęsła.
- Norma w $\mathbb{R}^n$ jest funkcją wypukłą. Wystarczy skorzystać z nierówności trójkąta.
- Odległość punktu od zbioru definiujmy następująco $d(\mathbf{x}, W) = \inf_{\mathbf{y}\in W} \|\mathbf{x}-\mathbf{y}\|$. Odległość punktu od zbioru wypukłego jest funkcją wypukłą: $f(\mathbf{x}) = d(\mathbf{x}, W)$ dla pewnego zbioru wypukłego $W \subset \mathbb{R}^n$.

3.3 Własności funkcji wypukłych

W tym rozdziale zakładamy, iż funkcja $f : W \to R$ jest określona na niepustym wypukłym podzbiorze $W \subset \mathbb{R}^n$.

Definicja 3.2. *Epigrafem* funkcji $f : W \to \mathbb{R}$ nazywamy zbiór

$$\operatorname{epi}(f) = \{(\mathbf{x}, z) \in W \times \mathbb{R} : z \geqslant f(\mathbf{x})\}.$$

Definicja 3.3. *Zbiorem poziomicowym* funkcji $f : W \to \mathbb{R}$ nazywamy

$$W_\alpha(f) = \{\mathbf{x} \in W : f(\mathbf{x}) \leqslant \alpha\}, \qquad \alpha \in \mathbb{R}.$$

Twierdzenie 3.5 (Twierdzenie o epigrafie)**.** *Funkcja f jest wypukła wtedy i tylko wtedy, gdy jej epigraf* $\operatorname{epi}(f)$ *jest wypukłym podzbiorem $\mathbb{R}^{n+1}$.*

Twierdzenie 3.6. *Jeśli funkcja f jest wypukła, to zbiór poziomicowy $W_\alpha(f)$ jest wypukły dla dowolnego α.*

Dowód powyższych twierdzeń pozostawiamy czytelnikowi.

Uwaga 3.2. Twierdzenie odwrotne do tw. 3.6 nie jest prawdziwe. Połóżmy $W = \mathbb{R}^n$ i weźmy dowolny niepusty wypukły zbiór $A \subset \mathbb{R}^n$. Rozważmy funkcję

$$f(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in A, \\ 1, & \mathbf{x} \notin A. \end{cases}$$

Każdy zbiór poziomicowy tej funkcji jest wypukły ($W_\alpha(f) = A$ dla $\alpha < 1$ i $W_\alpha(f) = \mathbb{R}^n$ dla $\alpha \geqslant 1$), zaś funkcja nie jest wypukła.

Okazuje się, że wypukłość gwarantuje ciągłość we wnętrzu $\operatorname{int} W$. Rezultat ten nie może być rozszerzony na brzeg W.

Twierdzenie 3.7. *Jeśli funkcja f jest wypukła, to jest również ciągła na* $\operatorname{int} W$.

Dowód powyższego twierdzenia wykracza poza ramy tego wykładu. Zainteresowany czytelnik może go znaleźć w monografii [10] w rozdziale IV.41.

Twierdzenie 3.8 (Twierdzenie o hiperpłaszczyźnie podpierającej)**.** *Jeśli f jest wypukła, to w każdym punkcie $\bar{\mathbf{x}} \in \operatorname{int} W$ istnieje hiperpłaszczyzna podpierająca, tzn. istnieje $\xi \in \mathbb{R}^n$ takie że*

$$f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}}) + \xi^T(\mathbf{x} - \bar{\mathbf{x}}), \qquad \forall\, \mathbf{x} \in W.$$

Jeśli f jest ściśle wypukła, to

$$f(\mathbf{x}) > f(\bar{\mathbf{x}}) + \xi^T(\mathbf{x} - \bar{\mathbf{x}}), \qquad \forall\, \mathbf{x} \in W \setminus \{\bar{\mathbf{x}}\}.$$

Jeśli f jest różniczkowalna w $\bar{\mathbf{x}}$, to w obu powyższych nierównościach możemy przyjąć $\xi = Df(\bar{\mathbf{x}})^T$.

Dowód tw. 3.8. Na mocy twierdzenia 3.5, epigraf $\operatorname{epi}(f)$ jest zbiorem wypukłym. Zastosujemy twierdzenie o oddzielaniu do zbiorów $U = \operatorname{int}\operatorname{epi}(f)$ i $V = \{(\bar{\mathbf{x}}, f(\bar{\mathbf{x}}))\}$. Istnieje niezerowy wektor $\mathbf{a} = (\xi, \alpha) \in \mathbb{R}^n \times \mathbb{R}$, takie że

$$\xi^T \mathbf{x} + \alpha y \leqslant \xi^T \bar{\mathbf{x}} + \alpha f(\bar{\mathbf{x}}) \tag{3.1}$$

dla $(\mathbf{x}, y) \in \operatorname{epi}(f)$. Nierówność ta musi być prawdziwa dla wszystkich $y \geqslant f(\mathbf{x})$. Zatem $\alpha \leqslant 0$. Okazuje się, że $\alpha \neq 0$. Dowiedziemy tego przez sprzeczność: załóżmy $\alpha = 0$. Wówczas dla dowolnego $\mathbf{x} \in W$ mamy $\xi^T(\mathbf{x} - \bar{\mathbf{x}}) \leqslant 0$. Korzystając z tego, że $\bar{\mathbf{x}}$ jest we wnętrzu W, wiemy, że $\bar{\mathbf{x}} + \varepsilon\xi \in W$ dla dostatecznie małego $\varepsilon > 0$. Połóżmy $\mathbf{x} = \bar{\mathbf{x}} + \varepsilon\xi$. Wtedy $0 \geqslant \xi^T(\mathbf{x} - \bar{\mathbf{x}}) = \varepsilon\xi^T\xi = \varepsilon\|\xi\|^2$, a zatem $\xi = \mathbf{0}$. Przeczy to niezerowości wektora (ξ, α). Wnioskujemy więc, że $\alpha < 0$.

Możemy założyć, że $\alpha = -1$ w nierówności (3.1) (wystarczy podzielić obie strony przez $|\alpha|$). Dla dowolnego $\mathbf{x} \in W$ dostajemy zatem

$$\xi^T \mathbf{x} - f(\mathbf{x}) \leqslant \xi^T \bar{\mathbf{x}} - f(\bar{\mathbf{x}}),$$

co przepisujemy następująco

$$f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}}) + \xi^T(\mathbf{x} - \bar{\mathbf{x}}).$$

Kończy to dowód pierwszej części twierdzenia.

Dowód drugiej części twierdzenia. Przypuśćmy, że f jest ściśle wypukła. Ustalmy $\bar{\mathbf{x}} \in \operatorname{int} W$. Na mocy pierwszej części twierdzenia $f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}}) + \xi^T(\mathbf{x} - \bar{\mathbf{x}})$ dla dowolnego $\mathbf{x} \in W$. Przypuśćmy, że dla pewnego $\mathbf{x} \in W$, $\mathbf{x} \neq \bar{\mathbf{x}}$, ta nierówność nie jest ścisła, tj. $f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \xi^T(\mathbf{x} - \bar{\mathbf{x}})$. Ze ścisłej wypukłości dostajemy

$$f\Big(\frac{\bar{\mathbf{x}} + \mathbf{x}}{2}\Big) < \frac{1}{2}f(\mathbf{x}) + \frac{1}{2}f(\bar{\mathbf{x}}) = f(\bar{\mathbf{x}}) + \frac{1}{2}\xi^T(\mathbf{x} - \bar{\mathbf{x}}). \tag{3.2}$$

Z drugiej strony, z istnienia płaszczyzny podpierającej w $\bar{\mathbf{x}}$ mamy

$$f\Big(\frac{\bar{\mathbf{x}} + \mathbf{x}}{2}\Big) \geqslant f(\bar{\mathbf{x}}) + \xi^T\Big(\frac{\mathbf{x} + \bar{\mathbf{x}}}{2} - \bar{\mathbf{x}}\Big) = f(\bar{\mathbf{x}}) + \xi^T\frac{\mathbf{x} - \bar{\mathbf{x}}}{2}.$$

Dostajemy sprzeczność z nierównością (3.2). Pokazuje to, że dla funkcji ściśle wypukłej f musi być spełniona ścisła nierówność $f(\mathbf{x}) > f(\bar{\mathbf{x}}) + \xi^T(\mathbf{x} - \bar{\mathbf{x}})$ jeśli tylko $\mathbf{x} \neq \bar{\mathbf{x}}$. Kończy to dowód drugiej części twierdzenia.

Załóżmy teraz, że f jest różniczkowalna w $\bar{\mathbf{x}}$. Dla dowolnego $\mathbf{x} \in W$, $\mathbf{x} \neq \bar{\mathbf{x}}$ i $\lambda \in (0, 1)$, z wypukłości mamy

$$\begin{aligned} f(\mathbf{x}) - f(\bar{\mathbf{x}}) &= \frac{(1-\lambda)f(\bar{\mathbf{x}}) + \lambda f(\mathbf{x}) - f(\bar{\mathbf{x}})}{\lambda} \\ &\geqslant \frac{f((1-\lambda)\bar{\mathbf{x}} + \lambda\mathbf{x}) - f(\bar{\mathbf{x}})}{\lambda} \\ &= \frac{f(\bar{\mathbf{x}} + \lambda(\mathbf{x} - \bar{\mathbf{x}})) - f(\bar{\mathbf{x}})}{\lambda}. \end{aligned} \tag{3.3}$$

Policzmy granicę, gdy λ dąży do zera:

$$f(\mathbf{x}) - f(\bar{\mathbf{x}}) \geqslant \lim_{\lambda \downarrow 0} \frac{f(\bar{\mathbf{x}} + \lambda(\mathbf{x} - \bar{\mathbf{x}})) - f(\bar{\mathbf{x}})}{\lambda} = Df(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}),$$

gdzie istnienie granicy i ostatnia równość wynika z różniczkowalności f w punkcie $\bar{\mathbf{x}}$. Jeśli f jest ściśle wypukła, to powtarzamy dowód drugiej części twierdzenia korzystając z ostatniej nierówności, czyli zastępując ξ^T przez $Df(\bar{\mathbf{x}})$. Dostajemy wówczas $f(\mathbf{x}) - f(\bar{\mathbf{x}}) > Df(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})$. □

Prawdziwe jest twierdzenie odwrotne do twierdzenia 3.8 (patrz tw. 3.9).

Wniosek 3.2. *Jeśli f wypukła i różniczkowalna w $\bar{\mathbf{x}}$, to w $\bar{\mathbf{x}} \in \operatorname{int} W$ jest minimum globalne wtw, gdy $Df(\bar{\mathbf{x}}) = 0$.*

Dowód. Implikacja w prawą stronę wynika z tw. 1.4. Aby dowieść implikacji przeciwnej, załóżmy $Df(\bar{\mathbf{x}}) = 0$ dla pewnego $\bar{\mathbf{x}} = 0$. Na mocy tw. 3.8 dla dowolnego $\mathbf{x} \in W$ mamy

$$f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}}) + Df(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) = f(\bar{\mathbf{x}}),$$

co dowodzi, że w punkcie $\bar{\mathbf{x}}$ jest minimum globalne. □

Poniżej wymieniamy pozostałe własności funkcji wypukłych. Ich dowody pozostawione są jako ćwiczenia.

- Niech $W \subset \mathbb{R}^n$, wypukły, zaś I dowolny zbiór. Jeśli funkcje $f_i : W \to \mathbb{R}$, $i \in I$, są wypukłe, to $f(\mathbf{x}) = \sup_{i \in I} f_i(\mathbf{x})$ jest wypukła ze zbiorem wartości $\mathbb{R} \cup \{\infty\}$.
- Niech $W \subset \mathbb{R}^n$, wypukły. Jeśli funkcje $f_i : W \to \mathbb{R}$, $i = 1, 2, \ldots, m$, są wypukłe i $\alpha_i > 0$ dla $i = 1, 2, \ldots, m$, to funkcja $f = \sum_{i=1}^m \alpha_i f_i$ jest wypukła. Jeśli jedna z funkcji f_i jest ściśle wypukła, to f jest również ściśle wypukła.
- Niech $W \subset \mathbb{R}^n$, $A \subset \mathbb{R}^m$ wypukłe zbiory. Jeśli funkcja $h : W \times A \to \mathbb{R}$ jest wypukła i ograniczona z dołu, to
$$f(\mathbf{x}) = \inf_{\mathbf{a} \in A} h(\mathbf{x}, \mathbf{a}), \qquad \mathbf{x} \in W,$$
jest wypukła.
- Niech $f : \mathbb{R}^n \to \mathbb{R}$ wypukła. Jeśli $h : \mathbb{R}^m \to \mathbb{R}^n$ afiniczna, to złożenie $f \circ h$ jest funkcją wypukłą.
- Niech $W \subset \mathbb{R}^n$ wypukły. Jeśli $f : W \to \mathbb{R}$ jest funkcją wypukłą i $g : \mathbb{R} \to \mathbb{R}$ jest wypukła i niemalejąca, to $g \circ f$ jest funkcją wypukłą. Jeśli dodatkowo g jest rosnąca, zaś f ściśle wypukła, to $g \circ f$ jest ściśle wypukła.
- Niech $W \subset \mathbb{R}^n$ wypukły. Jeśli $f : W \to \mathbb{R}$ jest funkcją (ściśle) wklęsłą i $f > 0$, to funkcja $1/f$ jest (ściśle) wypukła.

3.4 Charakteryzacje funkcji wypukłej

Twierdzenie 3.9. *Niech $W \subset \mathbb{R}^n$ wypukły o niepustym wnętrzu. Jeśli w każdym punkcie $\bar{\mathbf{x}} \in \operatorname{int} W$ istnieje wektor $\xi \in \mathbb{R}^n$, taki że*

$$f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}}) + \xi^T(\mathbf{x} - \bar{\mathbf{x}}), \qquad \forall\, \mathbf{x} \in W,$$

to funkcja f jest wypukła. Jeśli nierówność jest ostra dla $\mathbf{x} \neq \bar{\mathbf{x}}$, to f jest ściśle wypukła.

Dowód. Weźmy $\mathbf{x} \in \operatorname{int} W$, $\mathbf{y} \in W$ i $\lambda \in (0, 1)$. Oznaczmy $\mathbf{x}_\lambda = \lambda\mathbf{x} + (1-\lambda)\mathbf{y}$. Chcemy wykazać, że $f(\mathbf{x}_\lambda) \leqslant \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$. Na mocy Lematu 3.2 punkt $\mathbf{x}_\lambda$ należy do wnętrza W. Stosując założenie do tego punktu dostajemy $\xi \in \mathbb{R}^n$, takie że

$$\begin{aligned} f(\mathbf{x}) &\geqslant f(\mathbf{x}_\lambda) + \xi^T(\mathbf{x} - \mathbf{x}_\lambda), \\ f(\mathbf{y}) &\geqslant f(\mathbf{x}_\lambda) + \xi^T(\mathbf{y} - \mathbf{x}_\lambda). \end{aligned} \tag{3.4}$$

Stąd

$$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \geqslant f(\mathbf{x}_\lambda) + \xi^T \left[\lambda(\mathbf{x}-\mathbf{x}_\lambda) + (1-\lambda)(\mathbf{y}-\mathbf{x}_\lambda)\right] = f(\mathbf{x}_\lambda),$$

gdyż wielkości w nawiasie kwadratowym zerują się. Kończy to dowód twierdzenia. Jeśli nierówność w założeniu jest ostra i $\mathbf{x} \neq \mathbf{y}$, to nierówności (3.4) są ostre i dostajemy warunek ścisłej wypukłości. □

Twierdzenie 3.10. *Niech $W \subset \mathbb{R}^n$ niepusty, otwarty i wypukły, zaś $f : W \to \mathbb{R}$ dwukrotnie różniczkowalna. Wówczas:*

I) f jest wypukła wtw, gdy hesjan $D^2 f(\mathbf{x})$ jest nieujemnie określony dla każdego $\mathbf{x} \in W$.

II) Jeśli hesjan $D^2 f(\mathbf{x})$ jest dodatnio określony dla każdego $\mathbf{x} \in W$, to f jest ściśle wypukła.

Analogiczna teza zachodzi w przypadku wklęsłości.

Dowód. Załóżmy najpierw, że hesjan jest nieujemnie określony dla każdego $\mathbf{x} \in W$. Wówczas, na mocy wniosku 2.1, dla dowolnych $\bar{\mathbf{x}}, \mathbf{x} \in W$ mamy

$$f(\mathbf{x}) = f(\bar{\mathbf{x}}) + Df(\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}}) + \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})^T D^2 f(\tilde{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}}),$$

gdzie $\tilde{\mathbf{x}}$ jest punktem leżącym na odcinku łączącym $\bar{\mathbf{x}}$ z $\mathbf{x}$. Założenie o nieujemnej określoności hesjanu pociąga nieujemność ostatniego składnika powyższej sumy. Stąd

$$f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}}) + Df(\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}}). \tag{3.5}$$

Ponieważ powyższa nierówność zachodzi dla dowolnych $\bar{\mathbf{x}}, \mathbf{x} \in W$, to f jest wypukła na mocy twierdzenia 3.9.

Jeśli założymy, że hesjan jest dodatnio określony dla każdego punktu W, to nierówność (3.5) jest ostra i twierdzenie 3.9 implikuje ścisłą wypukłość f.

Przejdźmy teraz do dowodu implikacji przeciwnej. Załóżmy, że funkcja f jest wypukła. Ustalmy $\bar{\mathbf{x}} \in W$ i $\mathbf{h} \in \mathbb{R}^n$, $\mathbf{h} \neq 0$. Z otwartości W wynika, że istnieje $\delta > 0$, dla której $\bar{\mathbf{x}} + t\mathbf{h} \in W$, $t \in (-\delta, \delta)$. Zdefiniujmy funkcję $g(t) = f(\bar{\mathbf{x}} + t\mathbf{h})$, $t \in (-\delta, \delta)$. Jest to funkcja jednej zmiennej, dwukrotnie różniczkowalna oraz wypukła. Stosując twierdzenie 3.8 dostajemy

$$g(t) \geqslant g(0) + g'(0)t, \qquad t \in (-\delta, \delta).$$

Na mocy twierdzenie Taylora z resztą w postaci Peano, tw. 1.9, mamy

$$g(t) = g(0) + g'(0)t + \frac{1}{2}g''(0)t^2 + o(t^2), \qquad t \in (-\delta, \delta).$$

Powyższa nierówność i wzór Taylora dają następujące oszacowanie na drugą pochodną

$$\frac{1}{2}g''(0)t^2 + o(t^2) \geqslant 0.$$

Dzielimy obie strony przez t^2:

$$\frac{1}{2}g''(0) + \frac{o(t^2)}{t^2} \geqslant 0.$$

W granicy przy t dążącym do zera, drugi składnik zanika i pozostaje $g''(0) \geqslant 0$. Co ta nierówność znaczy w terminach funkcji f? Liczymy:

$$g'(t) = Df(\bar{\mathbf{x}} + t\mathbf{h})\mathbf{h}, \qquad g''(t) = \mathbf{h}^T D^2 f(\bar{\mathbf{x}} + t\mathbf{h})\mathbf{h}.$$

Stąd $g''(0) = \mathbf{h}^T D^2 f(\bar{\mathbf{x}})\mathbf{h}$.

Powyższe rozumowanie możemy przeprowadzić dla dowolnego $\mathbf{h} \in \mathbb{R}^n$. Wykazaliśmy więc nieujemną określoność $D^2 f(\bar{\mathbf{x}})$. □

3.5 Subróżniczka

Zajmiemy się teraz uogólnieniem pojęcia pochodnej na nieróżniczkowalne funkcje wypukłe. Niech $W \subseteq \mathbb{R}^n$ będzie zbiorem wypukłym, zaś $f : W \to \mathbb{R}$ funkcją wypukłą.

Definicja 3.4. Wektor $\xi \in \mathbb{R}^n$ nazywamy *subgradientem* funkcji f w punkcie $\mathbf{x}_0 \in W$, jeśli

$$f(\mathbf{x}) \geqslant f(\mathbf{x}_0) + \xi^T(\mathbf{x} - \mathbf{x}_0), \qquad \mathbf{x} \in W.$$

Zbiór wszystkich subgradientów f w punkcie $\mathbf{x}_0$ nazywamy *subróżniczką* i oznaczamy $\partial f(\mathbf{x}_0)$.

Wniosek 3.3. *Jeśli $W \subset \mathbb{R}^n$ jest zbiorem wypukłym o niepustym wnętrzu, to $f : W \to \mathbb{R}$ jest wypukła wtw, gdy w każdym punkcie zbioru $\operatorname{int} W$ istnieje subgradient:*

$$\partial f(\mathbf{x}) \neq \emptyset \qquad \forall\ \mathbf{x} \in \operatorname{int} W.$$

Dowód. Implikacja w prawą stronę wynika z twierdzenia 3.8, zaś implikacja w stronę lewą – z tw. 3.9. □

Lemat 3.3. *Niech $W \subset \mathbb{R}^n$ będzie zbiorem wypukłym, zaś $f : W \to \mathbb{R}$ funkcją wypukłą. Wówczas subróżniczka $\partial f(\mathbf{x})$ jest zbiorem wypukłym i domkniętym. Jeśli $\mathbf{x}$ jest wewnętrznym punktem W, $\mathbf{x} \in \operatorname{int} W$, to zbiór $\partial f(\mathbf{x})$ jest ograniczony więc zwarty.*

Dowód. Dowód wypukłości i domkniętości pozostawiamy jako ćwiczenie (patrz ćw. 3.24). Ustalmy $\bar{\mathbf{x}} \in \operatorname{int} W$. Wówczas istnieje $\varepsilon > 0$, taki że kula domknięta $B(\bar{\mathbf{x}}, \varepsilon)$ jest zawarta w $\operatorname{int} W$. Dla dowolnego $\xi \in \partial f(\bar{\mathbf{x}})$

$$f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}}) + \xi^T(\mathbf{x} - \bar{\mathbf{x}}), \qquad \mathbf{x} \in W.$$

A zatem

$$\sup_{\mathbf{x} \in B(\bar{\mathbf{x}}, \varepsilon)} f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}}) + \sup_{\mathbf{x} \in B(\bar{\mathbf{x}}, \varepsilon)} \xi^T(\mathbf{x} - \bar{\mathbf{x}}).$$

Lewa strona jest niezależna od ξ oraz, z ciągłości f na $\operatorname{int} W$, skończona. Supremum po prawej stronie jest przyjmowane dla $\mathbf{x} = \bar{\mathbf{x}} + \varepsilon\xi/\|\xi\|$ i wynosi $\varepsilon\|\xi\|$. Dostajemy zatem

$$\varepsilon\|\xi\| \leqslant \sup_{\mathbf{x} \in B(\bar{\mathbf{x}}, \varepsilon)} f(\mathbf{x}) - f(\bar{\mathbf{x}}),$$

co dowodzi ograniczoności zbioru $\partial f(\bar{\mathbf{x}})$. □

Pokażemy teraz związek subróżniczki z pochodnymi kierunkowymi funkcji. Związek ten przyda nam się w dalszych dowodach.

Definicja 3.5. *Pochodną kierunkową* funkcji f w punkcie $\bar{\mathbf{x}}$ w kierunku $\mathbf{d}$ nazywamy granicę

$$f'(\bar{\mathbf{x}}; \mathbf{d}) = \lim_{\lambda \downarrow 0} \frac{f(\bar{\mathbf{x}} + \lambda\mathbf{d}) - f(\bar{\mathbf{x}})}{\lambda}.$$

Z własności pochodnych jednostronnych dla skalarnych funkcji wypukłych wynikają następujące własności pochodnych kierunkowych:

Lemat 3.4. *Niech $W \subset \mathbb{R}^n$ będzie zbiorem wypukłym otwartym, zaś $f : W \to \mathbb{R}$ funkcją wypukłą. Wówczas dla każdego $\mathbf{d} \in \mathbb{R}^n$ i $\mathbf{x} \in W$*

I) *istnieje pochodna kierunkowa* $f'(\mathbf{x};\mathbf{d})$.

II) $f'(\mathbf{x};\mathbf{d}) = \inf_{\lambda>0} \frac{f(\bar{\mathbf{x}}+\lambda\mathbf{d})-f(\bar{\mathbf{x}})}{\lambda}$.

III) $f'(\mathbf{x};\mathbf{d}) \geqslant -f'(\mathbf{x};-\mathbf{d})$.

Dowód. Zdefiniujmy funkcję skalarną $g(t) = f(\mathbf{x}+t\mathbf{d})$ dla t, takich że $\mathbf{x}+t\mathbf{d} \in W$. Z otwartości W wynika, że g jest określona na pewnym otoczeniu zera $(-\delta,\delta)$. Jest ona także wypukła. Wówczas iloraz różnicowy jest monotoniczny, tzn. dla $-\delta < t_1 < t_2 < \delta$ mamy

$$\frac{g(t_1)-g(0)}{t_1} \leqslant \frac{g(t_2)-g(0)}{t_2}. \tag{3.6}$$

Z monotoniczności ilorazu różnicowego wynika, że istnieją pochodne lewostronna $g'(0-)$ i prawostronna $g'(0+)$, $g'(0-) \leqslant g'(0+)$ oraz

$$g'(0+) = \inf_{t>0} \frac{g(t)-g(0)}{t}.$$

Wystarczy teraz zauważyć, że $f'(\mathbf{x};\mathbf{d}) = g'(0+)$, zaś $f'(\mathbf{x};-\mathbf{d}) = -g'(0-)$.

Pozostał nam jeszcze dowód monotoniczności ilorazu różnicowego. Weźmy najpierw $0 < t_1 < t_2 < \delta$ i zauważmy, że nierówność (3.6) jest równoważna

$$g(t_1) \leqslant \lambda g(t_2) + (1-\lambda)g(0),$$

gdzie $\lambda = t_1/t_2 \in (0,1)$ oraz $\lambda t_2 + (1-\lambda)0 = t_1$. Prawdziwość ostatniej nierówności wynika z wypukłości funkcji g. Przypadek $-\delta < t_1 < t_2 < 0$ dowodzimy analogicznie. Dla $-\delta < t_1 < 0 < t_2 < \delta$ nierówność (3.6) jest równoważna

$$g(0) \leqslant \frac{1}{1+\lambda} g(t_1) + \frac{\lambda}{1+\lambda} g(t_2), \qquad \lambda = -\frac{t_1}{t_2},$$

która wynika z wypukłości g, gdyż $1/(1+\lambda) \in (0,1)$ oraz

$$\frac{1}{1+\lambda} t_1 + \frac{\lambda}{1+\lambda} t_2 = 0.$$

□

Pochodne kierunkowe pozwalają na nową charakteryzację subróżniczki.

Lemat 3.5. *Niech* $W \subset \mathbb{R}^n$ *będzie zbiorem wypukłym otwartym, zaś* $f : W \to \mathbb{R}$ *funkcją wypukłą. Prawdziwa jest następująca równoważność:* $\xi \in \partial f(\bar{\mathbf{x}})$ *wtw, gdy*

$$f'(\bar{\mathbf{x}};\mathbf{d}) \geqslant \xi^T\mathbf{d}, \qquad \forall\ \mathbf{d} \in \mathbb{R}^n.$$

Dowód. Ustalmy $\bar{\mathbf{x}} \in W$ i $\xi \in \partial f(\bar{\mathbf{x}})$. Wówczas dla $\lambda > 0$ i $\mathbf{d} \in \mathbb{R}^n$ (oczywiście takich, że $\bar{\mathbf{x}} + \lambda\mathbf{d} \in W$) mamy

$$f(\bar{\mathbf{x}} + \lambda\mathbf{d}) \geqslant f(\bar{\mathbf{x}}) + \lambda\xi^T\mathbf{d}.$$

Zatem

$$\frac{f(\bar{\mathbf{x}} + \lambda\mathbf{d}) - f(\bar{\mathbf{x}})}{\lambda} \geqslant \xi^T\mathbf{d},$$

co implikuje $f'(\bar{\mathbf{x}};\mathbf{d}) \geqslant \xi^T\mathbf{d}$.

Weźmy teraz wektor $\xi \in \mathbb{R}^n$ spełniający warunek $f'(\bar{\mathbf{x}};\mathbf{d}) \geqslant \xi^T\mathbf{d}$ dla każdego $\mathbf{d} \in \mathbb{R}^n$. Na mocy lematu 3.4(II) dla $\lambda > 0$ mamy

$$f'(\bar{\mathbf{x}};\mathbf{d}) \leqslant \frac{f(\bar{\mathbf{x}}+\lambda\mathbf{d}) - f(\bar{\mathbf{x}})}{\lambda}.$$

A zatem

$$f(\bar{\mathbf{x}}+\lambda\mathbf{d}) \geqslant f(\bar{\mathbf{x}}) + \xi^T(\lambda\mathbf{d}).$$

Z dowolności λ i $\mathbf{d}$ wynika, iż ξ jest subgradientem. □

Poniższe twierdzenie precyzuje związek pomiędzy subróżniczką i pochodną kierunkową funkcji. Zwróć uwagę na wykorzystanie twierdzenia o oddzielaniu w dowodzie. Metoda polegająca na sprytnym dobraniu rozłącznych zbiorów wypukłych, które można rozdzielić hiperpowierzchnią, będzie wielokrotnie wykorzystywana w dowodzeniu twierdzeń dotyczących funkcji wypukłych.

Twierdzenie 3.11. *Niech $f : W \to \mathbb{R}$ będzie funkcją wypukłą zadaną na wypukłym otwartym zbiorze $W \subset \mathbb{R}^n$. Dla dowolnego $\bar{\mathbf{x}} \in W$ i $\mathbf{d} \in \mathbb{R}^n$ zachodzi*

$$f'(\bar{\mathbf{x}};\mathbf{d}) = \max_{\xi\in\partial f(\bar{\mathbf{x}})} \xi^T\mathbf{d}.$$

Ponadto, funkcja f jest różniczkowalna w $\bar{\mathbf{x}}$ wtw, gdy subróżniczka $\partial f(\bar{\mathbf{x}})$ składa się z jednego subgradientu. Tym subgradientem jest wówczas $Df(\bar{\mathbf{x}})^T$.

Dowód. Z lematu 3.5 wynika, że $f'(\bar{\mathbf{x}};\mathbf{d}) \geqslant \xi^T\mathbf{d}$ dla $\xi \in \partial f(\bar{\mathbf{x}})$. Stąd

$$f'(\bar{\mathbf{x}};\mathbf{d}) \geqslant \max_{\xi\in\partial f(\bar{\mathbf{x}})} \xi^T\mathbf{d}.$$

Udowodnienie przeciwnej nierówności wymaga skorzystania z twierdzenia o oddzielaniu. Zdefiniujmy dwa zbiory:

$$C_1 = \{(\mathbf{x},z) \in W\times\mathbb{R} :\ z > f(\mathbf{x})\},$$
$$C_2 = \{(\mathbf{x},z) \in W\times\mathbb{R} :\ \mathbf{x} = \bar{\mathbf{x}}+\lambda\mathbf{d},\ z = f(\bar{\mathbf{x}}) + \lambda f'(\bar{\mathbf{x}};\mathbf{d}),\quad \lambda \geqslant 0\}.$$

Zauważmy, że C_1 różni się od epigrafu f tylko brzegiem $\{(\mathbf{x},z) \in W\times\mathbb{R} :\ z = f(\mathbf{x})\}$. Możemy zatem zastosować podobne rozumowanie jak w dowodzie tw. 3.5, aby wykazać, że C_1 jest zbiorem wypukłym. Zbiór C_2 jest półprostą o początku w punkcie $(\bar{\mathbf{x}}, f(\bar{\mathbf{x}}))$ i o kierunku $(\mathbf{d}, f'(\bar{\mathbf{x}};\mathbf{d}))$, więc jest wypukły. Można zauważyć, że jest on wykresem liniowego przybliżenia funkcji f wokół punktu $\bar{\mathbf{x}}$ wzdłuż odcinka $\{\bar{\mathbf{x}}+\lambda\mathbf{d} : \lambda \geqslant 0\} \cap W$.

Na mocy lematu 3.4(II) mamy $f'(\mathbf{x};\mathbf{d}) \leqslant \frac{f(\bar{\mathbf{x}}+\lambda\mathbf{d})-f(\bar{\mathbf{x}})}{\lambda}$, czyli

$$f(\bar{\mathbf{x}}+\lambda\mathbf{d}) \geqslant f(\bar{\mathbf{x}}) + \lambda f'(\mathbf{x};\mathbf{d}).$$

Wnioskujemy stąd, że zbiory C_1 i C_2 są rozłączne. Stosujemy twierdzenie o oddzielaniu, tw. 3.1: istnieje niezerowy wektor $(\mu,\gamma) \in \mathbb{R}^n \times \mathbb{R}$, taki że

$$\mu^T\mathbf{x} + \gamma z \geqslant \mu^T(\bar{\mathbf{x}}+\lambda\mathbf{d}) + \gamma\big(f(\bar{\mathbf{x}}) + \lambda f'(\bar{\mathbf{x}};\mathbf{d})\big), \qquad (\mathbf{x},z) \in C_1,\ \lambda \in [0,L), \tag{3.7}$$

gdzie $L = \sup\{\lambda \geqslant 0 :\ \bar{\mathbf{x}}+\lambda\mathbf{d} \in W\}$. Zauważmy, że γ nie może być ujemna, gdyż wtedy lewa strona mogłaby być dowolnie mała (z może być dowolnie duże). Nie może także być $\gamma = 0$, gdyż wówczas dla każdego $\mathbf{x} \in W$ musiałoby zachodzić $\mu^T(\mathbf{x}-\bar{\mathbf{x}}) \geqslant \lambda\mu^T\mathbf{d}$. Jest to możliwe tylko, gdy $\mu = \mathbf{0}$ (korzystamy tutaj z faktu, że $(\mathbf{x}-\bar{\mathbf{x}})$ może być dowolnie małe bo W jest otwarty).

A to przeczy niezerowości wektora (μ, γ). Dowiedliśmy zatem, że $\gamma > 0$. Bez straty ogólności można przyjąć, że $\gamma = 1$ dokonując przeskalowania μ w nierówności (3.7), czyli

$$\mu^T \mathbf{x} + z \geqslant \mu^T(\bar{\mathbf{x}} + \lambda \mathbf{d}) + f(\bar{\mathbf{x}}) + \lambda f'(\bar{\mathbf{x}}; \mathbf{d}), \qquad (\mathbf{x}, z) \in C_1,\ \lambda \in [0, L).$$

Zbiegając z z do $f(\mathbf{x})$ otrzymujemy nierówność, która zachodzi dla wszystkich $\mathbf{x} \in W$ oraz $\lambda \in [0, L)$

$$\mu^T \mathbf{x} + f(\mathbf{x}) \geqslant \mu^T(\bar{\mathbf{x}} + \lambda \mathbf{d}) + f(\bar{\mathbf{x}}) + \lambda f'(\bar{\mathbf{x}}; \mathbf{d}), \qquad \mathbf{x} \in W,\ \lambda \in [0, L). \tag{3.8}$$

Kładąc $\lambda = 0$ dostajemy

$$\mu^T(\mathbf{x} - \bar{\mathbf{x}}) + f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}}),$$

co po przekształceniu daje

$$f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}}) - \mu^T(\mathbf{x} - \bar{\mathbf{x}}).$$

A zatem $-\mu \in \partial f(\bar{\mathbf{x}})$. Biorąc w nierówności (3.8) $\lambda > 0$ i $\mathbf{x} = \bar{\mathbf{x}}$ dostajemy

$$-\mu^T(\lambda \mathbf{d}) \geqslant \lambda f'(\bar{\mathbf{x}}; \mathbf{d}),$$

czyli

$$\sup_{\xi \in \partial f(\bar{\mathbf{x}})} \xi^T \mathbf{d} \geqslant f'(\bar{\mathbf{x}}; \mathbf{d}).$$

To kończy dowód pierwszej części twierdzenia.

Dowód drugiej części twierdzenia wynika z następujących obserwacji.

1. Funkcja f jest różniczkowalna w punkcie $\bar{\mathbf{x}}$ wtw, gdy $f'(\bar{\mathbf{x}}; \mathbf{d}) = \alpha^T \mathbf{d}$ dla każdego $\mathbf{d} \in \mathbb{R}^n$ i pewnego wektora $\alpha \in \mathbb{R}^n$.

Równoważność ta wynika wprost z definicji pochodnej (patrz roz. 2). Jeśli f jest różniczkowalna w $\bar{\mathbf{x}}$, jedynym wektorem spełniającym $f'(\bar{\mathbf{x}}; \mathbf{d}) = \alpha^T \mathbf{d}$ jest $\alpha = Df(\bar{\mathbf{x}})^T$. Wynika stąd, że jeśli $\partial f(\bar{\mathbf{x}})$ jest zbiorem jednoelementowym to f jest różniczkowalna w $\bar{\mathbf{x}}$.

2. Załóżmy teraz, że f jest różniczkowalna w $\bar{\mathbf{x}}$. Z definicji różniczkowalności wynika, że dla dostatecznie małego $\lambda > 0$ mamy (bez zmniejszania ogólności możemy założyć, że $\|\mathbf{d}\| = 1$)

$$f(\bar{\mathbf{x}} + \lambda \mathbf{d}) = f(\bar{\mathbf{x}}) + \lambda Df(\bar{\mathbf{x}})\mathbf{d} + o(\lambda).$$

Z drugiej strony korzystając z definicji subgradientu mamy

$$f(\bar{\mathbf{x}} + \lambda \mathbf{d}) \geqslant f(\bar{\mathbf{x}}) + \lambda \xi^T \mathbf{d},$$

gdzie ξ jest subgradientem f w $\bar{\mathbf{x}}$.

Odejmując stronami dostajemy

$$\lambda(\xi^T - Df(\bar{\mathbf{x}}))\mathbf{d} - o(\lambda) \leqslant 0.$$

Dzieląc tę nierówność przez λ i przechodząc z λ do zera dostajemy $(\xi^T - Df(\bar{\mathbf{x}}))\mathbf{d} \leqslant 0$. Biorąc teraz $\mathbf{d} = (\xi^T - Df(\bar{\mathbf{x}}))/\|\xi^T - Df(\bar{\mathbf{x}})\|$ dostajemy równość $\xi^T = Df(\bar{\mathbf{x}})$, czyli subróżniczka $\partial f(\bar{\mathbf{x}})$ jest zbiorem jednoelementowym.

□

Twierdzenie 3.11 wykorzystamy kilkukrotnie w dowodzie poniższego twierdzenia, które będzie użyteczne w znajdowaniu subróżniczek.

Twierdzenie 3.12. *Niech $W \subset \mathbb{R}^n$ będzie zbiorem wypukłym otwartym, zaś $f_1, f_2 : W \to \mathbb{R}$ funkcjami wypukłymi.*

I) Niech $f = f_1 + f_2$. Wówczas $\partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}) = \partial f(\mathbf{x})$, gdzie

$$\partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}) = \{\xi_1 + \xi_2 : \ \xi_1 \in \partial f_1(\mathbf{x}), \ \xi_2 \in \partial f_2(\mathbf{x})\}.$$

II) Niech $f = \max(f_1, f_2)$. Wówczas

$$\partial f(\mathbf{x}) = \begin{cases} \partial f_1(\mathbf{x}), & f_1(\mathbf{x}) > f_2(\mathbf{x}), \\ \operatorname{conv}\left(\partial f_1(\mathbf{x}) \cup \partial f_2(\mathbf{x})\right), & f_1(\mathbf{x}) = f_2(\mathbf{x}), \\ \partial f_2(\mathbf{x}), & f_1(\mathbf{x}) < f_2(\mathbf{x}), \end{cases}$$

gdzie $\operatorname{conv}(\partial f_1(\mathbf{x}) \cup f_2(\mathbf{x}))$ jest zbiorem kombinacji wypukłych elementów zbiorów $\partial f_1(\mathbf{x})$ i $\partial f_2(\mathbf{x})$.

Dowód. Zaczniemy od dowodu (I). Ustalmy $\bar{\mathbf{x}} \in W$. Niech $\xi_1 \in \partial f_1(\bar{\mathbf{x}})$ i $\xi_2 \in \partial f_2(\bar{\mathbf{x}})$. Wówczas dla $\mathbf{x} \in W$ zachodzi

$$f_1(\mathbf{x}) \geqslant f_1(\bar{\mathbf{x}}) + \xi_1^T(\mathbf{x} - \bar{\mathbf{x}}),$$
$$f_2(\mathbf{x}) \geqslant f_2(\bar{\mathbf{x}}) + \xi_2^T(\mathbf{x} - \bar{\mathbf{x}}).$$

Dodając te nierówności stronami otrzymujemy

$$f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}}) + (\xi_1 + \xi_2)^T(\mathbf{x} - \bar{\mathbf{x}}),$$

czyli $\xi_1 + \xi_2 \in \partial f(\bar{\mathbf{x}})$. Dowiedliśmy zawierania $\partial f_1(\bar{\mathbf{x}}) + \partial f_2(\bar{\mathbf{x}}) \subset \partial f(\bar{\mathbf{x}})$. Przypuśćmy, że inkluzja ta jest ostra, tzn. istnieje $\xi \in \partial f(\bar{\mathbf{x}})$, taki że $\xi \notin \partial f_1(\bar{\mathbf{x}}) + \partial f_2(\bar{\mathbf{x}})$. Na mocy lematu 3.3 subróżniczki $\partial f_1(\bar{\mathbf{x}})$ i $\partial f_2(\bar{\mathbf{x}})$ są zwartymi zbiorami wypukłymi. A zatem ich suma algebraiczna jest również zbiorem zwartym i wypukłym (patrz ćw. 3.25). Stosujemy twierdzenie o ostrym oddzielaniu do zbiorów $\{\xi\}$ oraz $\partial f_1(\bar{\mathbf{x}}) + \partial f_2(\bar{\mathbf{x}})$. Dostajemy $\mu \in \mathbb{R}^n$, takie że

$$\mu^T\xi_1 + \mu^T\xi_2 < \mu^T\xi, \qquad \forall\ \xi_1 \in \partial f_1(\bar{\mathbf{x}}), \ \xi_2 \in \partial f_2(\bar{\mathbf{x}}).$$

Bierzemy maksimum po ξ_1, ξ_2 po lewej stronie i stosujemy twierdzenie 3.11:

$$f_1'(\bar{\mathbf{x}}; \mu) + f_2'(\bar{\mathbf{x}}; \mu) < \xi^T\mu \leqslant f'(\bar{\mathbf{x}}; \mu).$$

Z drugiej strony, z własności pochodnych kierunkowych mamy

$$f_1'(\bar{\mathbf{x}}; \mu) + f_2'(\bar{\mathbf{x}}; \mu) = f'(\bar{\mathbf{x}}; \mu).$$

Doprowadziliśmy do sprzeczności: ξ o żądanych własnościach nie może istnieć. Kończy to dowód (I).

Dowód (II). Postać subróżniczki ∂f na zbiorach $\{\mathbf{x} \in W : \ f_1(\mathbf{x}) > f_2(\mathbf{x})\}$ i $\{\mathbf{x} \in W : \ f_1(\mathbf{x}) < f_2(\mathbf{x})\}$ jest oczywista. Jedynie przypadek $\{\mathbf{x} \in W : \ f_1(\mathbf{x}) = f_2(\mathbf{x})\}$ wymaga dokładnego dowodu. Ustalmy $\bar{\mathbf{x}} \in W$, dla którego $f_1(\bar{\mathbf{x}}) = f_2(\bar{\mathbf{x}})$. Oznaczmy $A = \operatorname{conv}\left(\partial f_1(\bar{\mathbf{x}}) \cup \partial f_2(\bar{\mathbf{x}})\right)$. Dla $i = 1, 2$ oraz $\mathbf{x} \in W$ mamy

$$f(\mathbf{x}) - f(\bar{\mathbf{x}}) \geqslant f_i(\mathbf{x}) - f(\bar{\mathbf{x}}) = f_i(\mathbf{x}) - f_i(\bar{\mathbf{x}}) \geqslant \xi_i^T(\mathbf{x} - \bar{\mathbf{x}}), \qquad \forall\ \xi_i \in \partial f_i(\bar{\mathbf{x}}).$$

Stąd dostajemy $\partial f_1(\bar{\mathbf{x}}) \cup \partial f_1(\bar{\mathbf{x}}) \subset \partial f(\bar{\mathbf{x}})$. Z wypukłości subróżniczki (patrz lemat 3.3) wynika, że $A \subset \partial f(\bar{\mathbf{x}})$. Załóżmy teraz, że istnieje $\xi \in \partial f(\bar{\mathbf{x}}) \setminus A$. Zbiór A jest wypukły i zwarty. Stosujemy twierdzenie o ostrym oddzielaniu do zbiorów A i $\{\xi\}$. Dostajemy $\mu \in \mathbb{R}^n$ i stałą $b \in \mathbb{R}$, takie że

$$\mu^T \tilde{\xi} < b < \mu^T \xi, \qquad \forall\, \tilde{\xi} \in A.$$

W szczególności $\mu^T \xi_i < b$ dla $\xi_i \in \partial f_i(\bar{\mathbf{x}})$, $i = 1, 2$, czyli, na mocy tw. 3.11,

$$\max\left(f_1'(\bar{\mathbf{x}}; \mu), f_2'(\bar{\mathbf{x}}; \mu)\right) \leqslant b.$$

Podobnie, $b < \xi^T \mu \leqslant f'(\bar{\mathbf{x}}; \mu)$. Podsumowując:

$$\max\left(f_1'(\bar{\mathbf{x}}; \mu), f_2'(\bar{\mathbf{x}}; \mu)\right) < f'(\bar{\mathbf{x}}; \mu). \tag{3.9}$$

Z drugiej strony, definicja pochodnej kierunkowej daje nam następującą równość (przypomnijmy, że $f(\bar{\mathbf{x}}) = f_1(\bar{\mathbf{x}}) = f_2(\bar{\mathbf{x}})$):

$$\frac{f(\bar{\mathbf{x}} + \lambda \mathbf{d}) - f(\bar{\mathbf{x}})}{\lambda} = \max\left(\frac{f_1(\bar{\mathbf{x}} + \lambda \mathbf{d}) - f(\bar{\mathbf{x}})}{\lambda}, \frac{f_2(\bar{\mathbf{x}} + \lambda \mathbf{d}) - f(\bar{\mathbf{x}})}{\lambda}\right), \qquad \lambda > 0.$$

Przechodząc z λ do zera dostajemy

$$f'(\bar{\mathbf{x}}; \mathbf{d}) = \max\left(f_1'(\bar{\mathbf{x}}; \mathbf{d}), f_2'(\bar{\mathbf{x}}; \mathbf{d})\right).$$

Biorąc $\mathbf{d} = \mu$ dostajemy sprzeczność z (3.9), a więc nie może istnieć $\xi \in \partial f(\bar{\mathbf{x}}) \setminus A$. □

Twierdzenie 3.13. *Niech $W \subset \mathbb{R}^n$ będzie zbiorem wypukłym otwartym, $f : W \to \mathbb{R}$ funkcją wypukłą, zaś A będzie macierzą $n \times m$. Zdefiniujmy $\tilde{W} = \{\mathbf{x} \in \mathbb{R}^m : A\mathbf{x} \in W\}$. Wówczas $\tilde{W}$ jest zbiorem wypukłym otwartym oraz funkcja $F : \tilde{W} \to \mathbb{R}$ zadana wzorem $F(\mathbf{x}) = f(A\mathbf{x})$ ma w punkcie $\mathbf{x} \in \tilde{W}$ subróżniczkę zadaną wzorem*

$$\partial F(\mathbf{x}) = A^T \partial f(A\mathbf{x}).$$

Dowód. Dowód będzie przebiegał podobnie jak dowód poprzedniego twierdzenia. Ustalmy $\bar{\mathbf{x}} \in \tilde{W}$. Weźmy $\xi \in \partial f(A\bar{\mathbf{x}})$. Wówczas

$$f(A\mathbf{x}) \geqslant f(A\bar{\mathbf{x}}) + \xi^T(A\mathbf{x} - A\bar{\mathbf{x}}) = f(A\bar{\mathbf{x}}) + (A^T\xi)^T(\mathbf{x} - \bar{\mathbf{x}}),$$

czyli $A^T\xi \in \partial F(\bar{\mathbf{x}})$. Stąd mamy zawieranie $A^T \partial f(A\bar{\mathbf{x}}) \subset \partial F(\bar{\mathbf{x}})$. Równości tych dwóch zbiorów dowiedziemy przez sprzeczność. Załóżmy, że istnieje $\xi \in \partial F(\bar{\mathbf{x}}) \setminus A^T \partial f(A\bar{\mathbf{x}})$. Zbiór $A^T \partial f(A\bar{\mathbf{x}})$ jest wypukły i domknięty, jako liniowy obraz zbioru wypukłego i domkniętego. Zastosujemy twierdzenie o ostrym oddzielaniu, aby oddzielić go od zbioru jednoelementowego $\{\xi\}$. Dostajemy $\mu \in \mathbb{R}^m$ oraz $b \in \mathbb{R}$, takie że

$$\mu^T A^T \tilde{\xi} < b < \mu^T \xi, \qquad \forall\, \tilde{\xi} \in \partial f(A\bar{\mathbf{x}}).$$

Biorąc supremum po $\tilde{\xi} \in \partial f(A\bar{\mathbf{x}})$ po lewej stronie i stosując tw. 3.11 otrzymujemy $f'(A\bar{\mathbf{x}}; A\mu) \leqslant b$. Prawą stronę możemy również oszacować przez pochodną kierunkową: $\mu^T \xi \leqslant F'(\bar{\mathbf{x}}; \mu)$. Podsumowując:

$$f'(A\bar{\mathbf{x}}; A\mu) < F'(\bar{\mathbf{x}}; \mu).$$

Jednak z własności pochodnych kierunkowych natychmiast wnioskujemy, że $F'(\bar{\mathbf{x}}; \mathbf{d}) = f'(A\bar{\mathbf{x}}; A\mathbf{d})$ dla dowolnego $\mathbf{d} \in \mathbb{R}^m$. Mamy zatem sprzeczność. Dowodzi to, iż zbiór $\partial F(\bar{\mathbf{x}}) \setminus A^T \partial f(A\bar{\mathbf{x}})$ jest pusty. □

3.6 Zadania

Ćwiczenie 3.1. Niech $U, V \subset \mathbb{R}^n$ będą zbiorami wypukłymi. Wykaż, że

1. $U \cap V$ jest zbiorem wypukłym.
2. $U \cup V$ może nie być zbiorem wypukłym.

Ćwiczenie 3.2. Czy zbiór $\{(x_1, x_2, x_3) \in \mathbb{R}^3 : \ x_1^6 + x_1^2 + x_2 x_3 + x_3^2 \leqslant 1\}$ jest zbiorem wypukłym?

Ćwiczenie 3.3. Udowodnij lemat 3.1.

Ćwiczenie 3.4. Niech

$$W_1 = \{(x_1, x_2) : \ x_2 \geqslant e^{-x_1}\}, \qquad W_2 = \{(x_1, x_2) : \ x_2 \leqslant -e^{-x_1}\}.$$

Wykaż, że zbiory W_1, W_2 są wypukłe i rozłączne. Znajdź prostą (hiperpłaszczyznę) je dzielącą. Czy istnieje hiperpłaszczyzna dzieląca ściśle te zbiory (w sensie twierdzenia o ostrym oddzielaniu)?

Ćwiczenie 3.5. Niech $W_1, W_2 \subset \mathbb{R}^n$ będą zbiorami wypukłymi. Udowodnij, że $\inf\{\|\mathbf{x} - \mathbf{y}\| : \mathbf{x} \in W_1, \quad \mathbf{y} \in W_2\} > 0$ wtw, gdy istnieje hiperpłaszczyzna ściśle (w sensie tw. o ostrym oddzielaniu) oddzielająca te zbiory.

Ćwiczenie 3.6. Niech $\mathbb{X} \subset \mathbb{R}^n$ będzie dowolnym zbiorem. Zdefiniujmy zbiór $\tilde{\mathbb{X}}$ jako zbiór takich punktów $\mathbf{x} \in \mathbb{R}^n$, dla których istnieje liczba naturalna $m \in \mathbb{N}$, zależna od punktu, wektor $[\lambda_1, \ldots, \lambda_m] \in [0,1]^m$ o tej własności, że $\sum_{i=1}^m \lambda_i = 1$, oraz $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{X}$ i

$$\mathbf{x} = \sum_{i=1}^m \lambda_i \mathbf{x}_i.$$

Udowodnij, że

1. $\tilde{\mathbb{X}}$ jest zbiorem wypukłym.
2. $\tilde{\mathbb{X}}$ jest najmniejszym zbiorem wypukłym zawierającym $\mathbb{X}$.
3. W definicji zbioru $\tilde{\mathbb{X}}$ można założyć, że $m \leqslant n+1$, gdzie n jest wymiarem przestrzeni (Tw. Caratheodory'ego).

Zbiór $\tilde{X}$ nazywa się otoczką wypukłą zbioru $\mathbb{X}$ i oznaczane jest $\mathrm{conv}(\mathbb{X})$.

Ćwiczenie 3.7. Udowodnij twierdzenie 3.5.

Ćwiczenie 3.8. Udowodnij twierdzenie 3.6.

Ćwiczenie 3.9. Znajdź przykład zbioru $W \subseteq \mathbb{R}^n$ i funkcji wypukłej $f : W \to \mathbb{R}$, która nie jest ciągła na całym zbiorze W.

Wskazówka. Twierdzenie 3.7 pomoże w wyborze zbioru W.

Ćwiczenie 3.10. Niech $W \subseteq \mathbb{R}^n$ będzie zbiorem wypukłym i otwartym, zaś $f : W \to \mathbb{R}$ funkcją różniczkowalną. Udowodnij następującą równoważność: f jest wypukła wtw, gdy

$$\big(Df(\mathbf{y}) - Df(\mathbf{x})\big)(\mathbf{y} - \mathbf{x}) \geqslant 0, \qquad \forall\, \mathbf{x}, \mathbf{y} \in W.$$

Ćwiczenie 3.11. Udowodnij: Niech $W \subset \mathbb{R}^n$, wypukły, zaś I dowolny zbiór. Jeśli funkcje $f_i : W \to \mathbb{R}$, $i \in I$, są wypukłe, to $f(\mathbf{x}) = \sup_{i\in I} f_i(\mathbf{x})$ jest wypukła ze zbiorem wartości $\mathbb{R} \cup \{\infty\}$.

Ćwiczenie 3.12. Udowodnij: Niech $W \subset \mathbb{R}^n$, wypukły. Jeśli funkcje $f_i : W \to \mathbb{R}$, $i = 1, 2, \ldots, m$, są wypukłe i $\alpha_i > 0$ dla $i = 1, 2, \ldots, m$, to funkcja $f = \sum_{i=1}^m \alpha_i f_i$ jest wypukła. Jeśli jedna z funkcji f_i jest ściśle wypukła, to f jest również ściśle wypukła.

Ćwiczenie 3.13. Udowodnij: Niech $W \subset \mathbb{R}^n$, wypukły, zaś $A \subset \mathbb{R}^m$ wypukły, zwarty. Jeśli funkcja $h : W \times A \to \mathbb{R}$ jest wypukła i ciągła, to

$$f(\mathbf{x}) = \inf_{\mathbf{a}\in A} h(\mathbf{x}, \mathbf{a}), \qquad \mathbf{x} \in W,$$

jest wypukła.

Wskazówka. Korzystając ze zwartości A i ciągłości h, dla każdego $\mathbf{x}$ istnieje $\mathbf{a}(\mathbf{x}) \in A$ realizujący infimum. Dowodzimy teraz z definicji funkcji wypukłej.

Ćwiczenie 3.14. Udowodnij uogólnienie twierdzenia z ćwiczenia 3.13: opuścimy zwartość A i ciągłość h. Niech $W \subset \mathbb{R}^n$, $A \subset \mathbb{R}^m$ wypukłe zbiory. Jeśli funkcja $h : W \times A \to \mathbb{R}$ jest wypukła i ograniczona z dołu, to

$$f(\mathbf{x}) = \inf_{\mathbf{a}\in A} h(\mathbf{x}, \mathbf{a}), \qquad \mathbf{x} \in W,$$

jest wypukła.

Ćwiczenie 3.15. Skonstruuj przykład, który pokaże, że założenie o ograniczoności z dołu nie może być pominięte w twierdzeniu z ćw. 3.14.

Ćwiczenie 3.16. Udowodnij, że funkcja odległości od zbioru wypukłego jest funkcją wypukłą. Podaj przykład wskazujący na konieczność wypukłości zbioru.

Ćwiczenie 3.17. Udowodnij: Niech $f : \mathbb{R}^n \to \mathbb{R}$ wypukła. Jeśli $h : \mathbb{R}^m \to \mathbb{R}^n$ afiniczna, to złożenie $f \circ h$ jest funkcją wypukłą.

Ćwiczenie 3.18. Udowodnij: Niech $W \subset \mathbb{R}^n$, wypukły. Jeśli $f : W \to \mathbb{R}$ jest funkcją wypukłą i $g : \mathbb{R} \to \mathbb{R}$ jest wypukła i niemalejąca, to $g \circ f$ jest funkcją wypukłą. Kiedy $g \circ f$ jest ściśle wypukła?

Stwórz analog powyższego twierdzenia dla funkcji wklęsłych.

Ćwiczenie 3.19. Udowodnij: Niech $W \subset \mathbb{R}^n$ wypukły. Jeśli $f : W \to \mathbb{R}$ jest funkcją (ściśle) wklęsłą i $f > 0$, to funkcja $1/f$ jest (ściśle) wypukła.

Ćwiczenie 3.20. Czy jeśli $f : \mathbb{R}^n \to \mathbb{R}$ i $g : \mathbb{R} \to \mathbb{R}$ wypukłe, to $g \circ f : \mathbb{R}^n \to \mathbb{R}$ jest wypukła?

Ćwiczenie 3.21. Udowodnij, że funkcja $f(x_1, x_2) = e^{x_1^2 - x_2}$ jest wypukła.

Ćwiczenie 3.22. Niech $W \subset \mathbb{R}^n$ i funkcja $f : W \to \mathbb{R}$ klasy C^2. Załóżmy ponadto, że $D^2 f(\mathbf{x}) \geqslant 0$ dla każdego $\mathbf{x} \in W$ (tzn. Hesjan jest nieujemnie określony). Rozstrzygnij, który z następujących warunków jest wystarczający, by zdanie

$$Df(\bar{\mathbf{x}}) = \mathbf{0} \qquad \Longrightarrow \qquad \text{w } \bar{\mathbf{x}} \text{ jest globalne minimum}$$

było prawdziwe:

- $W = \mathbb{R}^n$,

- W jest wypukły i otwarty,
- W jest wypukły,
- W jest otwarty.

Ćwiczenie 3.23. Znajdź przykład zbioru $\mathbb{X} \subset \mathbb{R}^n$ oraz funkcji $f : \mathbb{X} \to \mathbb{R}$ klasy C^2, takiej że

1. $D^2 f(\mathbf{x}) \geqslant 0$ (tzn. hesjan jest nieujemnie określony) dla każdego $\mathbf{x} \in \mathbb{X}$,
2. istnieje punkt $\bar{\mathbf{x}} \in \mathbb{X}$, w którym jest minimum lokalne funkcji f, ale nie jest ono globalne.

Ćwiczenie 3.24. Niech $W \subset \mathbb{R}^n$ będzie zbiorem wypukłym, zaś $f : W \to \mathbb{R}$ funkcją wypukłą. Udowodnij, że $\partial f(\mathbf{x})$ jest zbiorem wypukłym i domkniętym.

Ćwiczenie 3.25. Niech $A, B \subset \mathbb{R}^n$ będą zbiorami wypukłymi zwartymi. Udowodnij, że suma algebraiczna tych zbiorów

$$A + B = \{\mathbf{a} + \mathbf{b} : \ \mathbf{a} \in A, \ \mathbf{b} \in B\}$$

jest zbiorem zwartym i wypukłym.

Ćwiczenie 3.26. Niech $W \subset \mathbb{R}^n$ wypukły otwarty i $f : W \to \mathbb{R}$ wypukła. Wykaż, że $\xi \in \mathbb{R}^n$ jest subgradientem f w $\bar{\mathbf{x}} \in W$ wtw, gdy odwzorowanie $\mathbf{x} \mapsto \xi^T \mathbf{x} - f(\mathbf{x})$ osiąga swoje minimum w $\bar{\mathbf{x}}$.

Ćwiczenie 3.27. Znajdź subróżniczkę funkcji $\mathbb{R}^n \ni \mathbf{x} \mapsto \|\mathbf{x}\|$.

Ćwiczenie 3.28. Niech A będzie macierzą $n \times n$ symetryczną i nieujemnie określoną. Udowodnij, że $f(\mathbf{x}) = \sqrt{\mathbf{x}^T A \mathbf{x}}$, $\mathbf{x} \in \mathbb{R}^n$, jest funkcją wypukłą i znajdź jej subróżniczkę.

Ćwiczenie 3.29. Wykaż, że dla skalarnej funkcji wypukłej $f : (a, b) \to \mathbb{R}$ mamy

$$\partial f(x) = [f'(x-), f'(x+)],$$

gdzie $f'(x\pm)$ oznaczają prawo- i lewo-stronne pochodne w punkcie $x \in (a, b)$.

Ćwiczenie 3.30. Niech $f : \mathbb{R}^n \to \mathbb{R}$ wypukła, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ ustalone wektory. Zdefiniujmy $g(t) = f(t\mathbf{x} + (1-t)\mathbf{y})$, $t \in \mathbb{R}$. Wykaż, że g jest wypukła oraz

$$\partial g(t) = (\mathbf{x} - \mathbf{y})^T \partial f(t\mathbf{x} + (1-t)\mathbf{y}).$$

Ćwiczenie 3.31. Wykaż, że funkcja $f(x) = \max(-2x + 5, 3x^2 - 1)$ jest wypukła na $\mathbb{R}$ i znajdź jej subróżniczkę.

Ćwiczenie 3.32. Niech $f_W(\mathbf{x}) = \inf_{\mathbf{y} \in W} \|x - y\|$, $\mathbf{x} \in \mathbb{R}^n$, gdzie $W \subset \mathbb{R}^n$ jest zbiorem wypukłym. Znajdź subróżniczkę w następujących przypadkach:

- $W = \{\bar{\mathbf{x}}\}$ dla pewnego $\bar{\mathbf{x}} \in \mathbb{R}^n$,
- $W = \mathrm{conv}(\{(0,0), (1,0)\}) \subset \mathbb{R}^2$, tzn. W jest odcinkiem,
- $W = \mathrm{conv}(\{(0,0), (1,0), (0,1), (1,1)\}) \subset \mathbb{R}^2$, tzn. W jest kwadratem.

Rozdział 4

Ekstrema funkcji wypukłej z ograniczeniami

4.1 Problem minimalizacyjny

Niech W będzie niepustym wypukłym podzbiorem $\mathbb{R}^n$, zaś $f : W \to \mathbb{R}$ będzie funkcją wypukłą. Rozważmy problem *minimalizacji* funkcji f na zbiorze W:

$$\begin{cases} f(\mathbf{x}) \longrightarrow \min, \\ \mathbf{x} \in W. \end{cases} \tag{4.1}$$

Przypomnijmy, że *punktami dopuszczalnymi* nazywamy elementy zbioru W. *Rozwiązaniem globalnym* nazywamy taki punkt $\bar{\mathbf{x}} \in W$, że $f(\bar{\mathbf{x}}) \leqslant f(\mathbf{x})$ dla każdego $\mathbf{x} \in W$. *Rozwiązaniem lokalnym* jest taki punkt $\bar{\mathbf{x}} \in W$, że istnieje $\varepsilon > 0$, dla którego $f(\bar{\mathbf{x}}) \leqslant f(\mathbf{x})$, jeśli $\mathbf{x} \in W \cap B(\bar{\mathbf{x}}, \varepsilon)$. Rozwiązanie jest *ścisłe*, jeśli $f(\bar{\mathbf{x}}) < f(\mathbf{x})$ dla $\mathbf{x} \in W \cap B(\bar{\mathbf{x}}, \varepsilon)$ i $\mathbf{x} \neq \bar{\mathbf{x}}$. Innymi słowy, $\bar{\mathbf{x}}$ jest rozwiązaniem lokalnym, jeśli $\bar{\mathbf{x}}$ jest minimum funkcji f na pewnym otoczeniu $\bar{\mathbf{x}}$.

Twierdzenie 4.1. *Niech $W \subset R^n$ wypukły, $f : W \to \mathbb{R}$ wypukła. Jeśli $\bar{\mathbf{x}} \in W$ będzie rozwiązaniem lokalnym (4.1), to:*

I) $\bar{\mathbf{x}}$ jest rozwiązaniem globalnym,

II) Zbiór rozwiązań globalnych jest wypukły.

III) Jeśli f jest ściśle wypukła, to $\bar{\mathbf{x}}$ jest ścisłym rozwiązaniem lokalnym.

IV) Jeśli $\bar{\mathbf{x}}$ jest ścisłym rozwiązaniem lokalnym, to $\bar{\mathbf{x}}$ jest jedynym rozwiązaniem globalnym.

Zauważmy, że w powyższym twierdzeniu nie zakładamy różniczkowalności funkcji f.

Dowód twierdzenia 4.1. (I): Dowód przez sprzeczność. Przypuśćmy, że istnieje $\mathbf{x}^* \in W$ takie że $f(\mathbf{x}^*) < f(\bar{\mathbf{x}})$. Ponieważ $\bar{\mathbf{x}}$ jest rozwiązaniem lokalnym, to $f(\bar{\mathbf{x}}) \leqslant f(\mathbf{x})$ dla $\mathbf{x} \in W \cap B(\bar{\mathbf{x}}, \varepsilon)$ i $\varepsilon > 0$. Z wypukłości zbioru W wynika, iż odcinek łączący $\bar{\mathbf{x}}$ i $\mathbf{x}^*$ znajduje się w zbiorze W. Ma on więc niepuste przecięcie z kulą $B(\bar{\mathbf{x}}, \varepsilon)$: dla pewnego $\lambda \in (0, 1)$ mamy $\lambda\bar{\mathbf{x}} + (1-\lambda)\mathbf{x}^* \in B(\bar{\mathbf{x}}, \varepsilon)$. Z wypukłości f dostajemy

$$f(\lambda\bar{\mathbf{x}} + (1-\lambda)\mathbf{x}^*) \leqslant \lambda f(\bar{\mathbf{x}}) + (1-\lambda) f(\mathbf{x}^*) < f(\bar{\mathbf{x}}),$$

co przeczy lokalnej optymalności $\bar{\mathbf{x}}$.

(II) Pozostawione jako ćwiczenie.

(III) Wynika wprost z definicji ścisłej wypukłości.

(IV) Pozostawione jako ćwiczenie. □

Dotychczas pokazaliśmy, że warunkiem koniecznym i dostatecznym minimum różniczkowalnej funkcji wypukłej na zbiorze wypukłym i *otwartym* jest zerowanie się pochodnej/gradientu. Uogólnimy teraz te wyniki na przypadek dowolnych zbiorów wypukłych.

Twierdzenie 4.2. *Niech $W \subset \mathbb{R}^n$ wypukły, $f : W \to \mathbb{R}$ wypukła. Jeśli f jest różniczkowalna w punkcie $\bar{\mathbf{x}} \in W$, to mamy następującą równoważność: $\bar{\mathbf{x}}$ jest rozwiązaniem (4.1) wtw, gdy $Df(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) \geqslant 0$ dla każdego $\mathbf{x} \in W$.*

Uwaga 4.1. W sformułowaniu powyższego twierdzenia, jak i w wielu miejscach w dalszej części tych notatek, zastosowany jest następujący skrót myślowy. Aby mówić o różniczkowalności funkcji f w punkcie $\bar{\mathbf{x}} \in W$, musi być ona określona w pewnym otoczeniu $\bar{\mathbf{x}}$, czyli w kuli $B(\bar{\mathbf{x}}, \varepsilon)$ dla pewnego $\varepsilon > 0$. Jeśli $\bar{\mathbf{x}}$ jest na brzegu W, to zakładać będziemy, że f jest określona na $W \cup B(\bar{\mathbf{x}}, \varepsilon)$ mimo, że jest to pominięte, dla prostoty notacji, w założeniach twierdzenia.

Uwaga 4.2. Jeśli $\bar{\mathbf{x}} \in \operatorname{int} W$, to warunek powyższy sprowadza się do warunku zerowania się pochodnej $Df(\bar{\mathbf{x}}) = 0$.

Dowód tw. 4.2. Niech $Df(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) \geqslant 0$ dla każdego $\mathbf{x} \in W$. Załóżmy, że w punkcie $\bar{\mathbf{x}}$ funkcja f nie osiąga minimum. Istnieje wtedy punkt $\mathbf{x}' \in W$, w którym $f(\mathbf{x}') < f(\bar{\mathbf{x}})$. Tworzymy ciąg $\mathbf{x}_k = (1 - \frac{1}{k})\bar{\mathbf{x}} + \frac{1}{k}\mathbf{x}'$. Z wypukłości zbioru W wynika, że $\mathbf{x}_k \in W$. Rozważmy pochodną kierunkową funkcji f w punkcie $\bar{\mathbf{x}}$ w kierunku wektora $(\mathbf{x}' - \bar{\mathbf{x}})$

$$\begin{aligned} f'(\bar{\mathbf{x}}; \mathbf{x}' - \bar{\mathbf{x}}) &= \lim_{k\to\infty} \frac{f(\bar{\mathbf{x}} + \frac{1}{k}(\mathbf{x}' - \bar{\mathbf{x}})) - f(\bar{\mathbf{x}})}{1/k} = \lim_{k\to\infty} \frac{f(\mathbf{x}_k) - f(\bar{\mathbf{x}})}{1/k} \\ &\leqslant \lim_{k\to\infty} \frac{(1 - 1/k)f(\bar{\mathbf{x}}) + 1/k f(\mathbf{x}') - f(\bar{\mathbf{x}})}{1/k} = f(\mathbf{x}') - f(\bar{\mathbf{x}}) < 0. \end{aligned}$$

Z założenia mamy

$$f'(\bar{\mathbf{x}}; \mathbf{x}' - \bar{\mathbf{x}}) = Df(\bar{\mathbf{x}})(\mathbf{x}' - \bar{\mathbf{x}}) \geqslant 0.$$

Otrzymaliśmy więc sprzeczność, co dowodzi, że $\bar{\mathbf{x}}$ jest minimum funkcji f w zbiorze W, czyli rozwiązaniem (4.1).

Załóżmy teraz, że $\bar{\mathbf{x}}$ jest rozwiązaniem (4.1). Ustalmy $\mathbf{x} \in W$. Zauważmy, że wypukłość W implikuje $\bar{\mathbf{x}} + \lambda(\mathbf{x} - \bar{\mathbf{x}}) = (1 - \lambda)\bar{\mathbf{x}} + \lambda\mathbf{x} \in W$ dla $\lambda \in [0, 1]$. Z definicji pochodnej mamy

$$Df(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) = \lim_{\lambda \downarrow 0,\ \lambda<1} \frac{f(\bar{\mathbf{x}} + \lambda(\mathbf{x} - \bar{\mathbf{x}})) - f(\bar{\mathbf{x}})}{\lambda}.$$

Ponieważ w punkcie $\bar{\mathbf{x}}$ jest minimum, to $f(\bar{\mathbf{x}} + \lambda(\mathbf{x} - \bar{\mathbf{x}})) \geqslant f(\bar{\mathbf{x}})$. Stąd $Df(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) \geqslant 0$. □

Z dowodu powyższego twierdzenia dostajemy użyteczny wniosek.

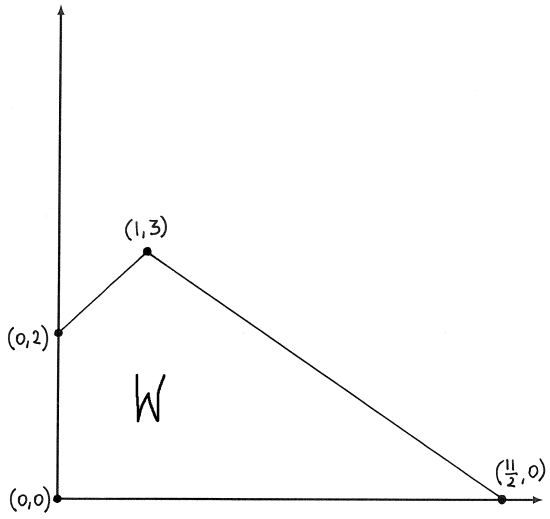
Wniosek 4.1. *Jeśli $\bar{\mathbf{x}} \in W$, gdzie $W \subset \mathbb{R}^n$ jest wypukły, jest rozwiązaniem lokalnym (4.1) dla funkcji $f : W \to \mathbb{R}$ (niekoniecznie wypukłej) różniczkowalnej w $\bar{\mathbf{x}}$, to $Df(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) \geqslant 0$ dla każdego $\mathbf{x} \in W$.*

Przykład 4.1. Rozważmy problem minimalizacyjny:

$$\begin{cases} \left(x_1 - \frac{3}{2}\right)^2 + (x_2 - 5)^2 \longrightarrow \min, \\ -x_1 + x_2 \leqslant 2, \\ 2x_1 + 3x_2 \leqslant 11, \\ x_1, x_2 \geqslant 0, \end{cases}$$

Zbiór W zadany jest przez ograniczenia liniowe (patrz rysunek 4.1):

$$W = \{(x_1, x_2) \in \mathbb{R}^2 : x_1, x_2 \geqslant 0,\ -x_1 + x_2 \leqslant 2,\ 2x_1 + 3x_2 \leqslant 11\}.$$



Rysunek 4.1: Zbiór punktów dopuszczalnych.

Łatwo sprawdzić, że jest on wypukły. Funkcja $f(x_1, x_2) = (x_1 - \frac{3}{2})^2 + (x_2 - 5)^2$ jest ściśle wypukła: jej hesjan wynosi

$$D^2 f(x_1, x_2) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$$

Na mocy tw. 4.1 minimum jest jednoznaczne. Policzmy gradient f:

$$Df(x_1, x_2) = (2x_1 - 3, 2x_2 - 10).$$

Gradient zeruje się w punkcie $(\frac{3}{2}, 5)$, który jest poza zbiorem W. Minimum należy zatem szukać na brzegu W. Nie mamy jeszcze narzędzi ułatwiających znalezienie tego punktu. Zgadujemy... Sprawdźmy wierzchołek $\bar{\mathbf{x}} = (1, 3)$. Gradient w tym punkcie wynosi $(-1, -4)$. Zauważmy, że wektory $\mathbf{x} - \bar{\mathbf{x}}$ są kombinacjami liniowymi dodatnimi wektorów $\mathbf{x}_1 = (0, 2) - (1, 3) = (-1, -1)$ i $\mathbf{x}_2 = (\frac{11}{2}, 0) - (1, 3) = (\frac{9}{2}, -3)$, tzn. $\mathbf{x} - \bar{\mathbf{x}} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2$ dla pewnych $a_1, a_2 \geqslant 0$. Wystarczy zatem sprawdzić, że $Df(\bar{\mathbf{x}})(\mathbf{x}_1) \geqslant 0$ i $Df(\bar{\mathbf{x}})\mathbf{x}_2 \geqslant 0$:

$$Df(\bar{\mathbf{x}})(\mathbf{x}_1) = (-1, -4)\begin{pmatrix} -1 \\ -1 \end{pmatrix} = 5,$$

$$Df(\bar{\mathbf{x}})(\mathbf{x}_2) = (-1, -4)\begin{pmatrix} 9/2 \\ -3 \end{pmatrix} = 16\frac{1}{2}.$$

Na mocy tw. 4.2 minimum jest rzeczywiście w punkcie $(1, 3)$.

Rozszerzymy teraz twierdzenie 4.2 na klasę funkcji wypukłych nieróżniczkowalnych – skorzystamy z wprowadzonego w poprzednim rozdziale pojęcia subróżniczki.

Twierdzenie 4.3. *Niech $\mathbb{X} \subset \mathbb{R}^n$ wypukły otwarty, $f : \mathbb{X} \to \mathbb{R}$ wypukła. Załóżmy, że zbiór punktów dopuszczalnych W jest wypukłym podzbiorem $\mathbb{X}$. Mamy następującą równoważność: $\bar{\mathbf{x}}$ jest rozwiązaniem (4.1) wtw, gdy istnieje $\xi \in \partial f(\bar{\mathbf{x}})$, takie że $\xi^T(\mathbf{x}-\bar{\mathbf{x}}) \geqslant 0$ dla każdego $\mathbf{x} \in W$.*

Wniosek 4.2. *Jeśli $\bar{\mathbf{x}} \in \operatorname{int} W$, to f ma w $\bar{\mathbf{x}} \in W$ minimum globalne wtw, gdy $\mathbf{0} \in \partial f(\bar{\mathbf{x}})$. Teza ta jest w szczególności prawdziwa, gdy $W \subset \mathbb{R}^n$ jest wypukły otwarty a $f : \mathbb{X} \to \mathbb{R}$ wypukła.*

Dowód twierdzenia 4.3. Zacznijmy od łatwiejszej implikacji. Załóżmy, że istnieje $\xi \in \partial f(\bar{\mathbf{x}})$, takie że $\xi^T(\mathbf{x}-\bar{\mathbf{x}}) \geqslant 0$ dla każdego $\mathbf{x} \in W$. Z faktu, że ξ jest subgradientem wynika, że

$$f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}}) + \xi^T(\mathbf{x}-\bar{\mathbf{x}}), \qquad \mathbf{x} \in W.$$

Wystarczy teraz skorzystać z założenia, żeby zauważyć, że $f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}})$ dla $\mathbf{x} \in W$, czyli $\bar{\mathbf{x}}$ jest rozwiązaniem (4.1).

Załóżmy teraz, że $\bar{\mathbf{x}} \in W$ jest rozwiązaniem (4.1). Zdefiniujmy dwa zbiory

$$\begin{aligned} C_1 &= \{(\mathbf{x}, z) \in \mathbb{R}^n \times \mathbb{R} : \ \mathbf{x} \in \mathbb{X}, \ z > f(\mathbf{x}) - f(\bar{\mathbf{x}})\}, \\ C_2 &= \{(\mathbf{x}, z) \in \mathbb{R}^n \times \mathbb{R} : \ \mathbf{x} \in W, \ z \leqslant 0\}. \end{aligned}$$

Oba zbiory są wypukłe, C_1 ma niepuste wnętrze (zbiór C_2 ma puste wnętrze, jeśli W ma puste wnętrze). Z faktu, że punkt $\bar{\mathbf{x}}$ jest rozwiązaniem (4.1) wynika, że $C_1 \cap C_2 = \emptyset$. Stosujemy twierdzenie o oddzielaniu: istnieje niezerowy wektor $(\mu, \gamma) \in \mathbb{R}^n \times \mathbb{R}$ i stała $b \in \mathbb{R}$, takie że

$$\begin{aligned} &\mu^T\mathbf{x} + \gamma z \leqslant b, && \forall\ \mathbf{x} \in \mathbb{X}, \ z > f(\mathbf{x}) - f(\bar{\mathbf{x}}) \\ &\mu^T\mathbf{x} + \gamma z \geqslant b, && \forall\ \mathbf{x} \in W, \ z \leqslant 0. \end{aligned} \tag{4.2}$$

Zanim przejdziemy do analitycznych rozważań popatrzmy na geometryczny obraz (patrz Rys. 4.2). Zauważmy, że zbiory C_1 i C_2 „stykają" się w punkcie $(\bar{\mathbf{x}}, 0)$. Hiperpłaszczyzna oddzielająca te zbiory musi zatem przechodzić przez ten punkt. Jest ona styczna do wykresu funkcji $\mathbf{x} \mapsto f(\mathbf{x}) - f(\bar{\mathbf{x}})$ – pierwsza grupa współrzędnych μ wektora (μ, γ) do niej normalnego, z dokładnością do długości i zwrotu, wyznacza subgradient tego odwzorowania w punkcie $\bar{\mathbf{x}}$ (a zatem także subgradient f). Z faktu, że ta hiperpłaszczyzna jest również styczna do C_2 dostajemy $\mu^T(\mathbf{x} - \bar{\mathbf{x}}) \geqslant 0$.

Udowodnijmy to teraz analitycznie. Odejmijmy od obu stron nierówności (4.2) $\mu^T\bar{\mathbf{x}}$:

$$\mu^T(\mathbf{x}-\bar{\mathbf{x}}) + \gamma z \leqslant \tilde{b}, \qquad \forall\ \mathbf{x} \in \mathbb{X}, \ z > f(\mathbf{x}) - f(\bar{\mathbf{x}}) \tag{4.3}$$

$$\mu^T(\mathbf{x}-\bar{\mathbf{x}}) + \gamma z \geqslant \tilde{b}, \qquad \forall\ \mathbf{x} \in W, \ z \leqslant 0, \tag{4.4}$$

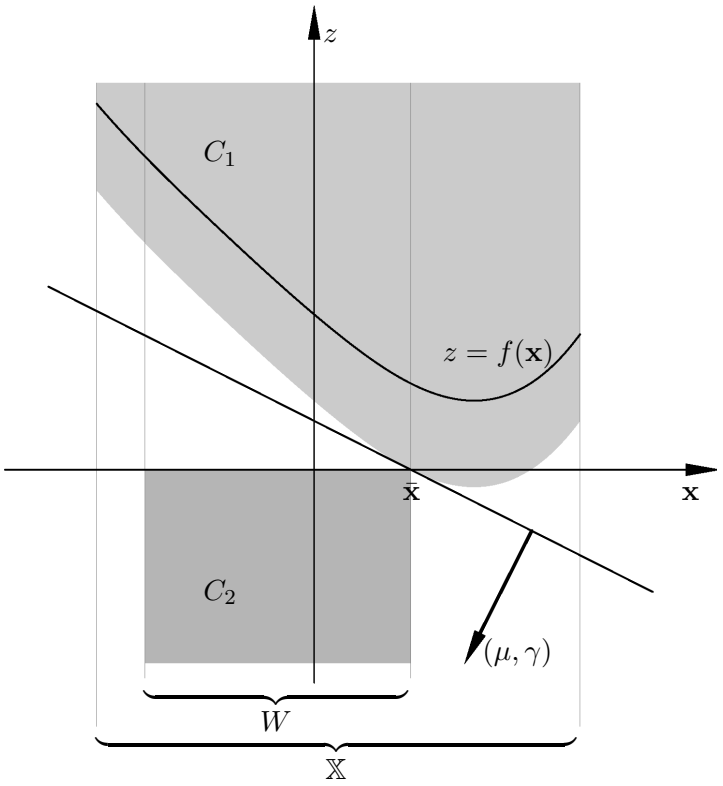
gdzie $\tilde{b} = b - \mu^T\bar{\mathbf{x}}$. Zauważmy najpierw, że γ nie może być większa od zera, bo wykorzystując dowolność $z \leqslant 0$, dostajemy sprzeczność z (4.4).

Kładąc w (4.4) $\mathbf{x} = \bar{\mathbf{x}}$ i $z = 0$, otrzymujemy $\tilde{b} \leqslant 0$. Biorąc z kolei $\mathbf{x} = \bar{\mathbf{x}}$, nierówność (4.3) upraszcza się do $\gamma z \leqslant \tilde{b}$ dla $z > 0$. Stąd $\tilde{b} \geqslant 0$. Podsumowując:

$$\tilde{b} = 0.$$

Korzystając z faktu, że $\tilde{b} = 0$ pokażemy niemożliwość spełnienia warunku $\gamma = 0$. Z nierówności (4.3) (pamiętając o otwartości $\mathbb{X}$) dostalibyśmy bowiem $\mu = \mathbf{0}$, co przeczyłoby niezerowości wektora (μ, γ). Wykazaliśmy więc, że

$$\gamma < 0.$$



Rysunek 4.2: Zbiory C_1 i C_2 z dowodu twierdzenia 4.3.

Przechodząc w (4.3) z z do $f(\mathbf{x}) - f(\bar{\mathbf{x}})$ mamy

$$\mu^T(\mathbf{x} - \bar{\mathbf{x}}) + \gamma(f(\mathbf{x}) - f(\bar{\mathbf{x}})) \leqslant 0.$$

Dzieląc obie strony przez γ i pamiętając, że $\gamma < 0$ dostajemy

$$\frac{\mu^T}{\gamma}(\mathbf{x} - \bar{\mathbf{x}}) + f(\mathbf{x}) - f(\bar{\mathbf{x}}) \geqslant 0,$$

co dowodzi, że $-\frac{\mu}{\gamma} \in \partial f(\bar{\mathbf{x}})$. Kładąc $z = 0$ w (4.4) otrzymujemy $\mu^T(\mathbf{x} - \bar{\mathbf{x}}) \geqslant 0$. Dzielimy obie strony przez $-\gamma > 0$:

$$-\frac{\mu^T}{\gamma}(\mathbf{x} - \bar{\mathbf{x}}) \geqslant 0,$$

co kończy dowód. □

4.2 Funkcje pseudowypukłe

W tym podrozdziale wprowadzimy rodzinę funkcji, dla której spełniony jest warunek:

$$Df(\bar{\mathbf{x}}) = 0 \quad \Longleftrightarrow \quad \text{w } \bar{\mathbf{x}} \text{ jest minimum globalne.}$$

Okazuje się, że rodzina ta obejmuje nie tylko funkcje wypukłe.

Definicja 4.1. Niech $W \subset \mathbb{R}^n$ będzie wypukły, otwarty i niepusty, zaś $f : W \to \mathbb{R}$. Funkcja f jest *pseudowypukła* w zbiorze W, jeśli f jest różniczkowalna w W oraz

$$\forall\, \mathbf{x}, \mathbf{y} \in W : \qquad Df(\mathbf{x})(\mathbf{y} - \mathbf{x}) \geqslant 0 \implies f(\mathbf{y}) \geqslant f(\mathbf{x}).$$

Funkcja f jest *ściśle pseudowypukła* w zbiorze W, jeśli

$$\forall\, \mathbf{x},\mathbf{y} \in W: \qquad Df(\mathbf{x})(\mathbf{y}-\mathbf{x}) \geqslant 0,\ \mathbf{x}\neq\mathbf{y} \implies f(\mathbf{y}) > f(\mathbf{x}).$$

Wprowadzimy także funkcje *pseudowypukłe w punkcie*. Funkcja f jest *pseudowypukła w punkcie* $\bar{\mathbf{x}}$ zbioru W, jeśli f jest różniczkowalna w $\bar{\mathbf{x}}$ oraz

$$\forall\, \mathbf{y} \in W: \qquad Df(\bar{\mathbf{x}})(\mathbf{y}-\bar{\mathbf{x}}) \geqslant 0 \implies f(\mathbf{y}) \geqslant f(\bar{\mathbf{x}}).$$

Analogicznie definiuje się funkcję *ściśle pseudowypukłą w punkcie*.

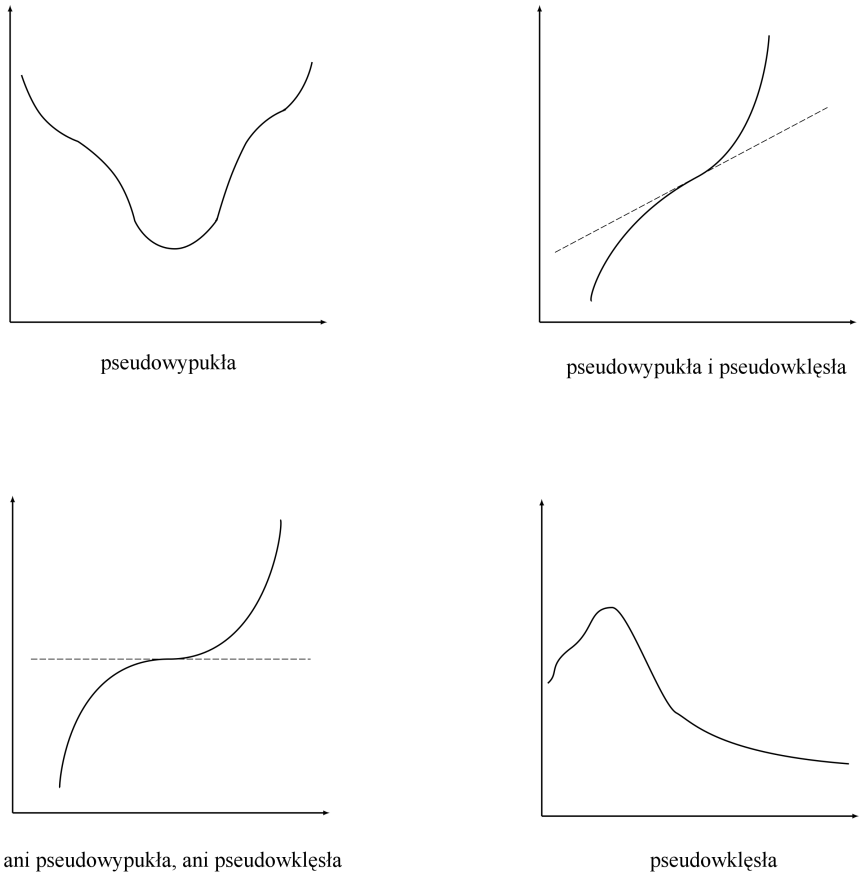
Pseudowypukłość w punkcie jest więc własnością funkcji f odniesioną do całego zbioru W (podobnie jak wypukłość), ale definiowaną jedynie w punktach różniczkowalności funkcji f.

Jeśli funkcja f jest (ściśle) pseudowypukła w każdym punkcie $\bar{\mathbf{x}} \in W$, to jest ona (ściśle) pseudowypukła w zbiorze W,

Funkcja f jest *(ściśle) pseudowklęsła*, jeśli $(-f)$ jest (ściśle) pseudowypukła.

Uwaga 4.3. Implikacja w definicji pseudowypukłości ma równoważną postać:

$$\forall\, \mathbf{x},\mathbf{y} \in W: \qquad f(\mathbf{y}) < f(\mathbf{x}) \implies Df(\mathbf{x})(\mathbf{y}-\mathbf{x}) < 0.$$



Rysunek 4.3: Przykłady funkcji pseudowypukłych i pseudowklęsłych.

Na rysunku 4.3 znajdują się przykłady jednowymiarowych funkcji pseudowypukłych i pseudowklęsłych.

Lemat 4.1. *Niech $f : W \to \mathbb{R}$, gdzie $W \subset \mathbb{R}^n$ wypukły, otwarty i niepusty. Jeśli f jest (ściśle) wypukła i różniczkowalna w W, to f jest (ściśle) pseudowypukła.*

Dowód. Załóżmy, że f wypukła. Na mocy tw. 3.8 dla dowolnych $\bar{\mathbf{x}}, \mathbf{x} \in W$ mamy

$$f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}}) + Df(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}).$$

Jeśli zatem $Df(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) \geqslant 0$, to $f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}})$ i f jest pseudowypukła. W analogiczny sposób dowodzimy, że ścisła wypukłość pociąga ścisłą pseudowypukłość. □

Lemat 4.2. *Niech $f : W \to \mathbb{R}$, gdzie $W \subset \mathbb{R}^n$ wypukły, otwarty i niepusty. Jeśli f jest funkcją pseudowypukłą w $\bar{\mathbf{x}} \in W$, to $\bar{\mathbf{x}}$ jest minimum globalnym wtw, gdy $Df(\bar{\mathbf{x}}) = 0$.*

Dowód. Identyczny jak dowód wniosku 3.2. □

Dowód poniższego lematu jest identyczny jak dowód twierdzenia 4.2.

Lemat 4.3. *Niech $W \subset \mathbb{R}^n$ wypukły, $f : W \to \mathbb{R}$ pseudowypukła. Mamy następującą równoważność: $\bar{\mathbf{x}}$ jest rozwiązaniem (4.1) wtw, gdy $Df(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) \geqslant 0$ dla każdego $\mathbf{x} \in W$.*

4.3 Maksymalizacja funkcji wypukłej

Definicja 4.2. *Punktem ekstremalnym* zbioru wypukłego $W \subset \mathbb{R}^n$ nazwiemy taki punkt $\bar{\mathbf{x}} \in W$, który nie jest punktem wewnętrznym żadnego odcinka zawartego w W, tj. jeśli $\bar{\mathbf{x}} = \lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2$ dla pewnych $\lambda \in (0, 1)$ i $\mathbf{x}_1, \mathbf{x}_2 \in W$, to $\mathbf{x}_1 = \mathbf{x}_2 = \bar{\mathbf{x}}$.

Definicja 4.3. *Otoczką wypukłą* punktów $(\mathbf{x}_i)_{i \in I}$ nazywamy zbiór punktów będących kombinacją wypukłą skończonej liczby spośród punktów $(\mathbf{x}_i)$.

Otoczkę wypukłą można równoważnie definiować jako najmniejszy zbiór wypukły zawierający punkty $(\mathbf{x}_i)_{i \in I}$.
Przedstawimy teraz prostą wersję twierdzenia Kreina-Milmana.

Twierdzenie 4.4. *Niech $U \subset \mathbb{R}^n$ będzie zbiorem wypukłym i zwartym. Jest on wówczas otoczką wypukłą swoich punktów ekstremalnych.*

Przed przejściem do dowodu powyższego twierdzenia wprowadzimy niezbędne pojęcia. Przypomnijmy, że przestrzenią afiniczną nazywamy zbiór A, taki że $\sum_{i=1}^k \lambda_i \mathbf{x}_i \in A$ dla dowolnych $(\mathbf{x}_i) \subset A$ i $(\lambda_i) \subset \mathbb{R}$ takich że $\sum_{i=1}^k \lambda_i = 1$. Każda podprzestrzeń afiniczna $\mathbb{R}^n$ może zostać przesunięta tak, aby zawierała $\mathbf{0}$. Staje się ona wówczas podprzestrzenią liniową. Wymiarem podprzestrzeni afinicznej nazywamy wymiar tej podprzestrzeni liniowej.

Definicja 4.4. *Wymiarem* zbioru wypukłego $U \subset \mathbb{R}^n$ nazywamy wymiar otoczki afinicznej U, tzn. podprzestrzeni afinicznej generowanej przez U:

$$\operatorname{aff} U = \{\sum_{i=1}^k \lambda_i \mathbf{x}_i : \ \mathbf{x}_1, \ldots, \mathbf{x}_k \in U, \ \sum_{i=1}^k \lambda_i = 1\}.$$

Zapiszmy łatwe wnioski z powyższej definicji i rozważań ją poprzedzających:

Wniosek 4.3.

1. *Zbiór wypukły o wymiarze $m < n$ możemy traktować jako podzbiór przestrzeni $\mathbb{R}^m$.*
2. *Zbiór wypukły w przestrzeni $\mathbb{R}^n$ ma niepuste wnętrze wtw, gdy jego wymiar wynosi n.*

Dotychczas rozważaliśmy hiperpłaszczyzny podpierające epigraf funkcji wypukłej. Teraz uogólnimy to pojęcie na dowolny zbiór wypukły.

Lemat 4.4. *Niech $U \subset \mathbb{R}^n$ będzie zbiorem wypukłym o niepustym wnętrzu. Przez punkt brzegowy $\bar{\mathbf{x}} \in U$ przechodzi wówczas hiperpłaszczyzna, taka że zbiór U leży w jednej z wyznaczonych przez nią półprzestrzeni. Hiperpłaszczyznę tę nazywamy* hiperpłaszczyzną podpierającą *zbiór U w punkcie $\bar{\mathbf{x}}$.*

Dowód. Na mocy twierdzenia o oddzielaniu zastosowanego do $\operatorname{int} U$ i $V = \{\bar{\mathbf{x}}\}$ (zauważmy, że $(\operatorname{int} U) \cap V = \emptyset$) istnieje $\mathbf{a} \in \mathbb{R}^n$, takie że $\mathbf{a}^T\mathbf{x} \leqslant \mathbf{a}^T\bar{\mathbf{x}}$ dla każdego $\mathbf{x} \in U$. Szukaną hiperpłaszczyzną jest

$$H = \{\mathbf{x} \in \mathbb{R}^n : \ \mathbf{a}^T\mathbf{x} = \mathbf{a}^T\bar{\mathbf{x}}\}.$$

Zbiór U zawarty jest w półprzestrzeni $\{\mathbf{x} \in \mathbb{R}^n : \ \mathbf{a}^T\mathbf{x} \leqslant \mathbf{a}^T\bar{\mathbf{x}}\}$. □

Dowód twierdzenia 4.4. Przeprowadzimy indukcję po wymiarze m zbioru zwartego i wypukłego U. Przypadki $m = 0$ (U jest punktem) i $m = 1$ (U jest odcinkiem) są trywialne. Czas na krok indukcyjny. Załóżmy, że każdy zbiór wypukły i zwarty o wymiarze nie większym od m jest otoczką wypukłą swoich punktów ekstremalnych. Weźmy zbiór wypukły i zwarty U o wymiarze $m + 1$. Zbiór ten traktujemy jako podzbiór $\mathbb{R}^{m+1}$. Ma on wówczas niepuste wnętrze. Niech $\bar{\mathbf{x}} \in U$.

Przypadek (I): $\bar{\mathbf{x}}$ leży na brzegu U. Na mocy lematu 4.4 istnieje hiperpłaszczyzna podpierająca H przechodząca przez $\bar{\mathbf{x}}$. Zbiór $U_{\bar{\mathbf{x}}} = U \cap H$ jest wypukły, zwarty i o wymiarze co najwyżej m. Na mocy założenia indukcyjnego punkt $\bar{\mathbf{x}}$ jest kombinacją wypukłą punktów ekstremalnych $U_{\bar{\mathbf{x}}}$. Pozostaje już tylko wykazać, że punkty ekstremalne $U_{\bar{\mathbf{x}}}$ są punktami ekstremalnymi U. Wynika to stąd, że żaden punkt $\mathbf{x} \in U_{\bar{\mathbf{x}}}$ nie może być przedstawiony jako kombinacja wypukła punktów, z których jeden lub oba nie należą do $U_{\bar{\mathbf{x}}}$.

Przypadek (II): $\bar{\mathbf{x}}$ leży we wnętrzu U. Przeprowadźmy przez $\bar{\mathbf{x}}$ dowolną prostą. Przecina ona brzeg U w dwóch punktach $\mathbf{x}_1, \mathbf{x}_2$. Z przypadku (I) wiemy, że każdy z punktów $\mathbf{x}_1, \mathbf{x}_2$ może zostać przedstawiony jako kombinacja wypukła skończonej liczby punktów ekstremalnych U. Punkt $\bar{\mathbf{x}}$ może być zapisany jako kombinacja wypukła $\mathbf{x}_1, \mathbf{x}_2$, a więc należy do otoczki wypukłej punktów ekstremalnych U. □

Poniżej prezentujemy inny dowód twierdzenia 4.4 bazujący częściowo na pomysłach wykorzystywanych w dowodzie ogólnej wersji twierdzenia Kreina-Milmana. W przeciwieństwie do powyższego rozumowania, dowód ten nie jest konstruktywny.

Alternatywny dowód twierdzenia 4.4. Jeśli zbiór U zawarty jest w $\mathbb{R}^1$, to teza twierdzenia jest trywialna. Załóżmy więc, że każdy wypukły zwarty podzbiór $\mathbb{R}^m$ jest otoczką wypukłą swoich punktów ekstremalnych. Udowodnimy prawdziwość tego stwierdzenia dla $U \subset \mathbb{R}^{m+1}$. Niech W będzie otoczką wypukłą punktów ekstremalnych U. Oczywiście $W \subset U$. Przypuśćmy, że istnieje $\bar{\mathbf{x}} \in U \setminus W$. Wówczas możemy znaleźć kulę o środku w $\bar{\mathbf{x}}$ i dostatecznie małym promieniu, która nie przecina W. Na mocy twierdzenia o ostrym oddzielaniu, tw. 3.2, istnieje wektor $\mathbf{a} \in \mathbb{R}^{m+1}$, taki że $\mathbf{a}^T\mathbf{x} \leqslant \alpha$ dla $\mathbf{x} \in W$ i $\mathbf{a}^T\bar{\mathbf{x}} > \alpha$. Niech $\beta = \sup_{\mathbf{x}\in U} \mathbf{a}^T\mathbf{x}$. Supremum to jest po całym zbiorze U i jest skończone ze zwartości U. Hiperpłaszczyzna $P = \{\mathbf{x} \in \mathbb{R}^{m+1} : \mathbf{a}^T\mathbf{x} = \beta\}$ nie przecina W (bo $\alpha < \mathbf{a}^T\bar{\mathbf{x}} \leqslant \beta$), ale ma punkt wspólny z U. Rzeczywiście, $P_U := P \cap U$ jest niepusty, gdyż ze zwartości U wynika, że supremum definiujące β jest osiągane,

czyli istnieje $\mathbf{x} \in U$, dla którego $\mathbf{a}^T\mathbf{x} = \beta$. Pokażemy, że P_U zawiera punkt ekstremalny U, co będzie sprzeczne z definicją zbioru W. Zbiór P_U jest niepustym, zwartym zbiorem wypukłym w przestrzeni afinicznej o wymiarze m. Zbiór P_U możemy traktować jako podzbiór $\mathbb{R}^m$, czyli na mocy założenia indukcyjnego jest on otoczką wypukłą swoich punktów ekstremalnych. Niech $\bar{\mathbf{y}}$ będzie jednym z punktów ekstremalnych P_U i załóżmy, że jest on kombinacją wypukłą $\mathbf{y}_1, \mathbf{y}_2 \in U$, $\bar{\mathbf{y}} = \lambda\mathbf{y}_1 + (1-\lambda)\mathbf{y}_2$ dla $\lambda \in (0,1)$. Wówczas $\beta = \mathbf{a}^T\bar{\mathbf{y}} = \lambda\mathbf{a}^T\mathbf{y}_1 + (1-\lambda)\mathbf{a}^T\mathbf{y}_2$. Z konstrukcji β wynika, że zarówno $\mathbf{a}^T\mathbf{y}_1$ jak i $\mathbf{a}^T\mathbf{y}_2$ muszą być równe β, czyli $\mathbf{y}_1, \mathbf{y}_2 \in P_U$. Z faktu, że $\bar{\mathbf{y}}$ jest punktem ekstremalnym P_U wnioskujemy, że $\mathbf{y}_1 = \mathbf{y}_2 = \bar{\mathbf{y}}$, czyli $\bar{\mathbf{y}}$ jest punktem ekstremalnym U. □

Twierdzenie 4.5. *Niech $f : W \to \mathbb{R}$ wypukła, ciągła, określona na wypukłym i zwartym zbiorze $W \subset \mathbb{R}^n$. Wówczas punkt ekstremalny zbioru W jest jednym z rozwiązań globalnych problemu*

$$\begin{cases} f(\mathbf{x}) \to \max, \\ \mathbf{x} \in W. \end{cases}$$

Dowód. Funkcja ciągła osiąga swoje kresy na zbiorze zwartym. Powyższy problem maksymalizacyjny ma zatem rozwiązanie $\bar{\mathbf{x}} \in W$. Na mocy tw. 4.4 punkt $\bar{\mathbf{x}}$ jest kombinacją wypukłą skończonej liczby punktów ekstremalnych, $\mathbf{x}_1, \ldots, \mathbf{x}_m$, zbioru W, tzn.

$$\bar{\mathbf{x}} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \ldots + a_m\mathbf{x}_m$$

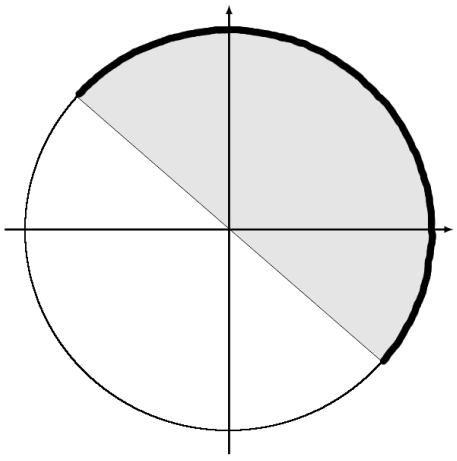
dla liczb $a_1, \ldots, a_m > 0$ takich że $a_1 + \ldots + a_m = 1$. Z wypukłości f dostajemy

$$f(\bar{\mathbf{x}}) \leqslant a_1 f(\mathbf{x}_1) + \ldots + a_m f(\mathbf{x}_m).$$

Z faktu, że $\bar{\mathbf{x}}$ jest maksimum f na zbiorze W wynika, że $f(\mathbf{x}_1) = \ldots = f(\mathbf{x}_m) = f(\bar{\mathbf{x}})$. □

Przykład 4.2. Rozważmy następujące zadanie optymalizacyjne:

$$\begin{cases} x_1 + x_2^2 \to \max, \\ x_1^2 + x_2^2 \leqslant 1, \\ x_1 + x_2 \geqslant 0. \end{cases}$$



Rysunek 4.4: Zbiór punktów dopuszczalnych. Punkty ekstremalne zaznaczone pogrubioną linią.

Funkcja celu jest wypukła, więc na mocy twierdzenia 4.5 punkt ekstremalny jest rozwiązaniem. Zbiór punktów dopuszczalnych jest kołem przeciętym z półprzestrzenią, czyli zbiorem wypukłym. Zbiór punktów ekstremalnych zaznaczony jest pogrubioną linią na rysunku 4.4. Jest to fragment okręgu. Zmaksymalizujmy więc funkcję celu na całym okręgu i zobaczmy, czy rozwiązanie należy do tego fragmentu okręgu. Podstawiając $x_2^2 = 1 - x_1^2$ do funkcji celu dostajemy

$$x_1 + 1 - x_1^2 \to \max.$$

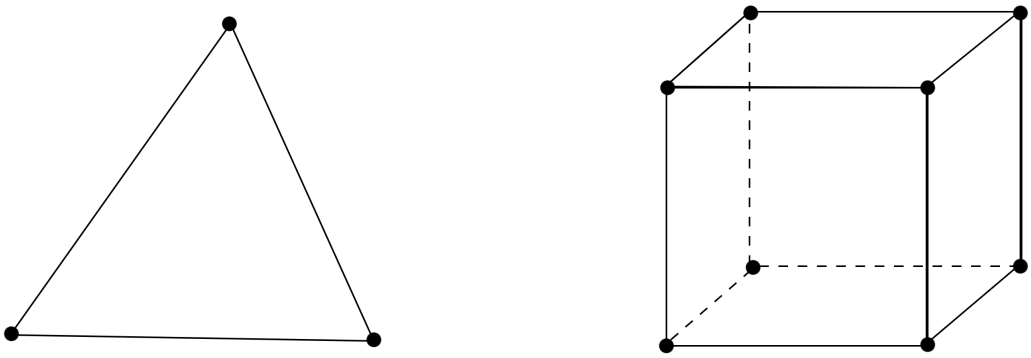
Rozwiązaniem tego problemu jest $x_1 = 1/2$. Stąd $x_2 = \pm\sqrt{3}/2$. Para $(1/2, \sqrt{3}/2)$ należy do półokręgu punktów ekstremalnych, więc jest rozwiązaniem, lecz niekoniecznie jedynym. Para $(1/2, -\sqrt{3}/2)$ nie należy nawet do zbioru punktów dopuszczalnych.

Twierdzenie 4.5 jest szczególnie użyteczne przy maksymalizacji funkcji wypukłej na zbiorach wielościennych:

Definicja 4.5. Zbiór $W \subset \mathbb{R}^n$ nazwiemy *zbiorem wielościennym*, jeśli jest przecięciem skończonej rodziny półprzestrzeni, tzn.

$$W = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{p}_i^T \mathbf{x} \leqslant \alpha_i,\ i = 1, \ldots, m\},$$

gdzie $\mathbf{p}_i \in \mathbb{R}^n$, $\mathbf{p}_i \neq 0$ i $\alpha_i \in \mathbb{R}$.



Rysunek 4.5: Zbiory wielościenne. Ciemnymi kropami zaznaczone są punkty ekstremalne.

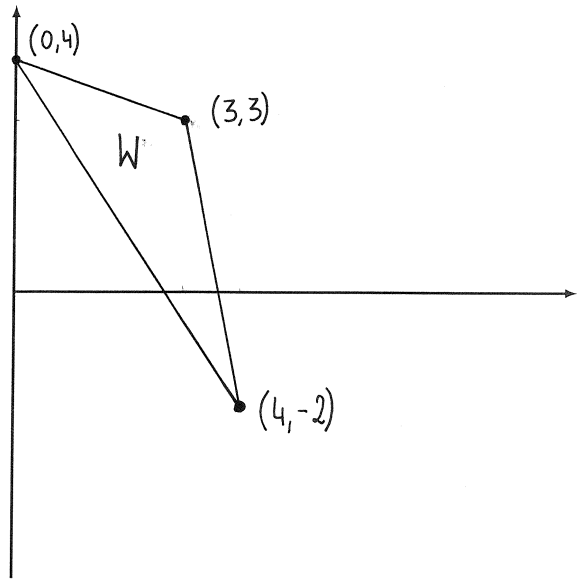
Łatwo można zauważyć, że punktami ekstremalnymi ograniczonych zbiorów wielościennych są „wierzchołki" (patrz rys. 4.5).

Lemat 4.5. *Zbiór wielościenny jest domknięty i wypukły.*

Dowód. Zbiór wielościenny jest domknięty i wypukły jako przecięcie rodziny zbiorów domkniętych i wypukłych. □

Przykład 4.3. Rozważmy problem optymalizacyjny

$$\begin{cases} x_1^2 + x_2^2 \to \max, \\ x_1 + 3x_2 \leqslant 12, \\ 5x_1 + x_2 \leqslant 18, \\ 3x_1 + 2x_2 \geqslant 8. \end{cases}$$



Rysunek 4.6: Zbiór punktów dopuszczalnych. Punkty ekstremalne zaznaczone dużymi kropkami.

Funkcja celu jest wypukła, zaś zbiór punktów dopuszczalnych jest wielościenny. Jego punkty ekstremalne to $(0,4)^T$, $(3,3)^T$ i $(4,-2)^T$, patrz rysunek 4.6. Wartość funkcji celu w tych punktach wynosi odpowiednio: 16, 18 i 20. Rozwiązaniem jest zatem punkt $(4,-2)^T$. Można łatwo dowieść, że jest to jedyne rozwiązanie tego problemu.

4.4 Zadania

Ćwiczenie 4.1. Udowodnij, że zbiór rozwiązań globalnych zagadnienia (4.1) jest wypukły dla wypukłego zbioru $W \subset \mathbb{R}^n$ i wypukłej funkcji $f : W \to \mathbb{R}$.

Ćwiczenie 4.2. Rozważmy zagadnienie (4.1) dla wypukłego zbioru $W \subset \mathbb{R}^n$ i wypukłej funkcji $f : W \to \mathbb{R}$. Wykaż, że ścisłe rozwiązanie lokalne jest jedynym rozwiązaniem globalnym.

Ćwiczenie 4.3. Udowodnij: Niech $W \subset \mathbb{R}^n$ wypukły oraz $g, h : W \to \mathbb{R}$ różniczkowalne. Jeśli zachodzi jeden z poniższych warunków:

- g jest wypukła, $g \geqslant 0$, oraz h jest wklęsła, $h > 0$,
- g jest wypukła, $g \leqslant 0$, oraz h jest wypukła, $h > 0$,

to $f = g/h$ jest pseudowypukła. Podaj przykład, że f nie jest wypukła.

Ćwiczenie 4.4. Udowodnij: Niech $W \subset \mathbb{R}^n$ wypukły oraz $g, h : W \to \mathbb{R}$ różniczkowalne. Jeśli g jest wypukła i $g \leqslant 0$ oraz h jest wklęsła, $h > 0$, to $f = gh$ jest pseudowypukła.

Ćwiczenie 4.5. Udowodnij, że zbiór minimów funkcji pseudowypukłej $f : W \to \mathbb{R}$, gdzie $W \subset \mathbb{R}^n$ wypukły, jest zbiorem wypukłym.

Ćwiczenie 4.6. Odpowiedz na następujące pytania:

1. Czy suma funkcji pseudowypukłych jest pseudowypukła?
2. Czy jeśli funkcje g_1, g_2 są pseudowypukłe w punkcie $\mathbf{x}$ oraz $Dg_1(\mathbf{x}) = Dg_2(\mathbf{x}) = \mathbf{0}^T$, to funkcja $g_1 + g_2$ jest pseudowypukła w $\mathbf{x}$?
3. Czy suma funkcji pseudowypukłej i wypukłej jest pseudowypukła?

Odpowiedzi uzasadnij kontrprzykładami lub dowodami.

Ćwiczenie 4.7. Korzystając z tw. 4.2 rozwiąż zadanie

$$\begin{cases} 2x + 3y \to \min, \\ x^2 + 2y^2 \leqslant 1. \end{cases}$$

Ćwiczenie 4.8. Znaleźć rozwiązania zadania

$$\begin{cases} \log(x_1 + 4) + x_2 \to \max, \\ x_2 \geqslant 2|x_1|, \\ x_2 \leqslant 4. \end{cases}$$

Wskazówka. Skorzystaj z tw. 4.2.

Ćwiczenie 4.9. Rozwiąż zadanie

$$\begin{cases} \frac{e^{(x_1-3)^2+x_2}}{\log(x_2)} \to \min, \\ x_2 > 1, \\ x_1 \in [1, 100]. \end{cases}$$

Ćwiczenie 4.10. Udowodnij, że $\bar{\mathbf{x}} \in W$ jest punktem ekstremalnym zbioru wypukłego $W \subset \mathbb{R}^n$ wtw, gdy $W \setminus \bar{\mathbf{x}}$ jest zbiorem wypukłym.

Ćwiczenie 4.11. Wykaż, że jeśli w punkcie brzegowym $\bar{\mathbf{x}}$ zbioru wypukłego U istnieje hiperpłaszczyzna podpierająca H, taka że $H \cap U = \{\bar{\mathbf{x}}\}$, to punkt $\bar{\mathbf{x}}$ jest punktem ekstremalnym U.

Ćwiczenie 4.12. Zbadaj związek pomiędzy subróżniczką funkcji wypukłej w punkcie $\bar{\mathbf{x}}$ a hiperpłaszczyznami podpierającymi epigraf tej funkcji w tym punkcie.

Ćwiczenie 4.13. Znaleźć rozwiązanie zadania

$$\begin{cases} \log\left(\frac{1}{x_1+4}\right) + x_2 \to \max, \\ x_2 \geqslant 2|x_1|, \\ x_2 \leqslant 4. \end{cases}$$

Ćwiczenie 4.14. Rozwiąż problem programowania nieliniowego z ograniczeniami:

$$\begin{cases} \log_2(x_1+x_2) - 2|x_2-2x_1| - x_2^2 + 2x_2 \to \min, \\ x_1 \geqslant 0, \qquad x_2 \geqslant 0, \qquad x_1 + x_2 \geqslant 1, \\ |x_1 - x_2| + |x_1| \leqslant 4. \end{cases}$$

Ćwiczenie 4.15. ([3, Zadania 3.28, 3.29]) Zdefiniujmy $f : [0,\infty)^n \to \mathbb{R}$ następująco:

$$f(\mathbf{x}) = \min\{\mathbf{c}^T\mathbf{y} + \mathbf{x}^T(A\mathbf{y} - \mathbf{b}) : \ \mathbf{y} \in W\},$$

gdzie W jest zbiorem wielościennym, $\mathbf{b}, \mathbf{c} \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$.

1. Wykaż, że f jest wklęsła.
2. Scharakteryzuj subróżniczkę f.
3. Znajdź subróżniczkę w punktach $\mathbf{x} \geqslant \mathbf{0}$ dla

$$A = \begin{pmatrix} 3 & 2 \\ -1 & 2 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 6 \\ 4 \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} -1 \\ -2 \end{pmatrix}.$$

Ćwiczenie 4.16. Korzystając z tw. 4.3 rozwiąż następujący problem optymalizacyjny:

$$\begin{cases} x^2 - 6x + y^2 + 2y \to \min, \\ x + 2y - 10 = 0, \\ 25 - x^2 - y^2 \geqslant 0. \end{cases}$$

Rozdział 5

Warunek konieczny I rzędu

5.1 Stożek kierunków stycznych

Rozważmy problem optymalizacyjny

$$\begin{cases} f(\mathbf{x}) \to \min, \\ \mathbf{x} \in W, \end{cases} \tag{5.1}$$

gdzie $W \subset \mathbb{R}^n$ i $f : W \to \mathbb{R}$. Niech $\bar{\mathbf{x}}$ będzie rozwiązaniem lokalnym. Będziemy chcieli powiązać geometrię lokalną zbioru W w punkcie $\bar{\mathbf{x}}$ z zachowaniem funkcji f, czyli kierunkami spadku jej wartości. Przez lokalną geometrię W rozumiemy zbiór kierunków, w których możemy się poruszyć z punktu $\bar{\mathbf{x}}$ nie opuszczając W.

Definicja 5.1. *Stożkiem kierunków stycznych* $T(\bar{\mathbf{x}})$ do W w punkcie $\bar{\mathbf{x}} \in \operatorname{cl} W$ nazywamy zbiór wektorów $\mathbf{d} \in \mathbb{R}^n$ takich że

$$\mathbf{d} = \lim_{k\to\infty} \lambda_k(\mathbf{x}_k - \bar{\mathbf{x}})$$

dla pewnych $\lambda_k > 0$, $\mathbf{x}_k \in W$ i $\mathbf{x}_k \to \bar{\mathbf{x}}$.

Powyższa definicja mówi, iż kierunek $\mathbf{d}$ należy do stożka kierunków stycznych $T(\bar{\mathbf{x}})$, jeśli jest on granicą kierunków wyznaczonych przez ciąg punktów dopuszczalnych $(\mathbf{x}_k)$ zmierzających to $\bar{\mathbf{x}}$. Zapisać możemy to formalnie w następujący sposób:

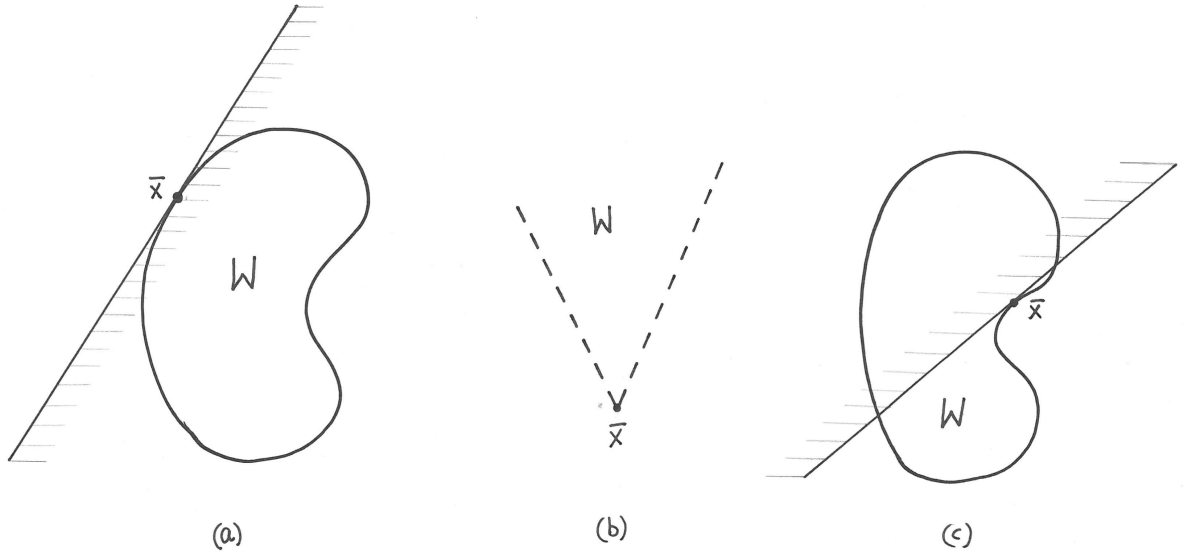
$$T(\bar{\mathbf{x}}) = \left\{\mathbf{d} \in \mathbb{R}^n : \ \mathbf{d} = \lambda \lim_{k\to\infty} \frac{\mathbf{x}_k - \bar{\mathbf{x}}}{\|\mathbf{x}_k - \bar{\mathbf{x}}\|} \text{ dla pewnych } \lambda \geqslant 0,\ (\mathbf{x}_k) \subset W \text{ oraz } \mathbf{x}_k \to \bar{\mathbf{x}},\ \mathbf{x}_k \neq \bar{\mathbf{x}}\right\}. \tag{5.2}$$

Dowód tej tożsamości oraz poniższego lematu pozostawiamy jako ćwiczenie.

Lemat 5.1.

1. *Zbiór* $T(\bar{\mathbf{x}})$ *jest stożkiem, tzn.* $\lambda\mathbf{d} \in T(\bar{\mathbf{x}})$ *dla dowolnych* $\mathbf{d} \in T(\bar{\mathbf{x}})$ *i* $\lambda \geqslant 0$. *W szczególności,* $\mathbf{0} \in T(\bar{\mathbf{x}})$.
2. *Jeśli* $\bar{\mathbf{x}}$ *jest punktem wewnętrznym zbioru* W, *to* $T(\bar{\mathbf{x}}) = \mathbb{R}^n$.
3. *Stożek* $T(\bar{\mathbf{x}})$ *jest domknięty.*

Przykład 5.1. Na rysunku 5.1 znajdują się trzy przykłady zbiorów i stożków do nich stycznych w punkcie $\bar{\mathbf{x}}$. Stożki te są przesunięte o wektor $\bar{\mathbf{x}}$, by pokazać ich zależność od kształtu zbioru. W przykładzie (a) i (c) zakładamy, że brzeg zbioru W jest gładki, więc stożki te są półprzestrzeniami ograniczonymi przez styczną w $\bar{\mathbf{x}}$. W przykładzie (b) zbiór W to wnętrze narysowanego stożka (bez brzegu). Wówczas stożek kierunków stycznych jest domknięciem zbioru W.



Rysunek 5.1: Stożki styczne zaczepione w punkcie styczności.

Jak już wspomnieliśmy, stożek kierunków stycznych jest ściśle związany z rozwiązaniem zagadnienia (5.1). Jeśli w punkcie $\bar{\mathbf{x}} \in W$ funkcja f ma minimum lokalne na W, to wówczas kierunki spadku wartości funkcji f nie mogą należeć do zbioru kierunków stycznych w punkcie $\bar{\mathbf{x}}$. Gdyby tak nie było, to poruszając się w kierunku spadku funkcji f zmniejszalibyśmy jej wartość jednocześnie pozostając w zbiorze W. Intuicje te formalizujemy poniżej.

Definicja 5.2. Niech $f : \mathbb{X} \to \mathbb{R}$ będzie różniczkowalna w $\bar{\mathbf{x}} \in \mathbb{X}$. Zbiorem *kierunków spadku* funkcji f w punkcie $\bar{\mathbf{x}}$ nazywamy

$$D(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^n : Df(\bar{\mathbf{x}})\,\mathbf{d} < 0\}.$$

Twierdzenie 5.1. *Niech $\bar{\mathbf{x}}$ będzie rozwiązaniem lokalnym problemu (5.1). Jeśli f jest różniczkowalna w $\bar{\mathbf{x}}$, to*

$$D(\bar{\mathbf{x}}) \cap T(\bar{\mathbf{x}}) = \emptyset.$$

Dowód. Weźmy $\mathbf{d} \in T(\bar{\mathbf{x}})$. Wówczas $\mathbf{d} = \lim_{k\to\infty} \lambda_k(\mathbf{x}_k - \bar{\mathbf{x}})$ dla pewnego ciągu punktów $(\mathbf{x}_k) \subset W$ zbieżnego do $\bar{\mathbf{x}}$ oraz ciągu liczb $(\lambda_k) \subset (0, \infty)$. Z definicji różniczkowalności f w $\bar{\mathbf{x}}$ mamy

$$f(\mathbf{x}_k) = f(\bar{\mathbf{x}}) + Df(\bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}}) + o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|).$$

Z faktu, że $\bar{\mathbf{x}}$ jest rozwiązaniem lokalnym $f(\mathbf{x}_k) \geqslant f(\bar{\mathbf{x}})$ dla dostatecznie dużych k. W połączeniu z powyższym wzorem daje to następujące oszacowanie:

$$0 \leqslant f(\mathbf{x}_k) - f(\bar{\mathbf{x}}) = Df(\bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}}) + o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|).$$

Mnożąc obie strony powyższej nierówności przez λ_k dostajemy

$$0 \leqslant Df(\bar{\mathbf{x}})(\lambda_k(\mathbf{x}_k - \bar{\mathbf{x}})) + \lambda_k o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|).$$

Zrobimy teraz sztuczkę, aby rozwiązać problem z $o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|)$ i przejdziemy z k to nieskończoności:

$$0 \leqslant Df(\bar{\mathbf{x}})\underbrace{(\lambda_k(\mathbf{x}_k - \bar{\mathbf{x}}))}_{\to \mathbf{d}} + \underbrace{\lambda_k\|\mathbf{x}_k - \bar{\mathbf{x}}\|}_{\to \|\mathbf{d}\|}\underbrace{\frac{o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|)}{\|\mathbf{x}_k - \bar{\mathbf{x}}\|}}_{\to 0}.$$

Udowodniliśmy zatem, że $Df(\bar{\mathbf{x}})\mathbf{d} \geqslant 0$, czyli $\mathbf{d} \notin D(\bar{\mathbf{x}})$. Kończy to dowód twierdzenia. □

Przykład 5.2. Rozważmy następujący problem optymalizacyjny:

$$\begin{cases} x_1^2 + x_2^2 \to \min, \\ x_1 + x_2 \geqslant 1. \end{cases}$$

Oznaczmy $f(x_1, x_2) = x_1^2 + x_2^2$ i $W = \{\mathbf{x} \in \mathbb{R}^2 : \ x_1 + x_2 \geqslant 1\}$. Zbadamy zbiory $T(\bar{\mathbf{x}})$ i $D(\bar{\mathbf{x}})$ w następujących punktach: $[1,1]^T, [1,0]^T, [\frac{1}{2}, \frac{1}{2}]^T$.

- $\bar{\mathbf{x}} = [1,1]^T$. Punkt ten leży wewnątrz zbioru W, czyli $T(\bar{\mathbf{x}}) = \mathbb{R}^2$. Zbiór kierunków spadku funkcji f dany jest następująco:

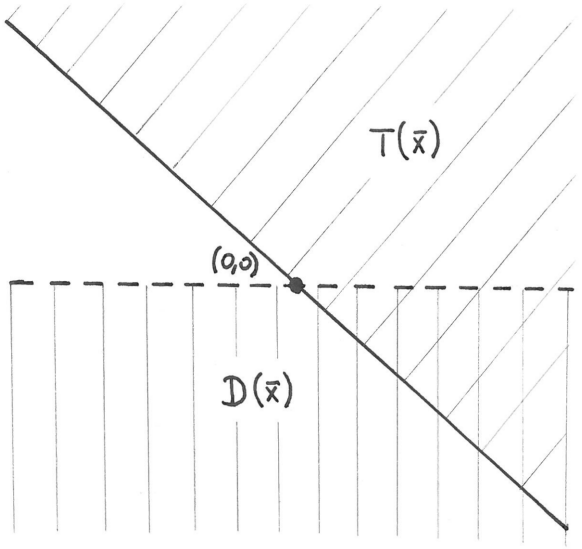
  $$D(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^2 : \ Df(\bar{\mathbf{x}})\mathbf{d} < 0\} = \{\mathbf{d} \in \mathbb{R}^2 : \ [2,2]\mathbf{d} < 0\} = \{\mathbf{d} \in \mathbb{R}^2 : \ d_1 + d_2 < 0\}.$$

  W oczywisty sposób część wspólna powyższych zbiorów nie jest pusta, czyli w punkcie $[1,1]^T$ nie ma minimum.

- $\bar{\mathbf{x}} = [1,0]^T$. Punkt ten leży na brzegu zbioru W. Łatwo można zauważyć, że

  $$T(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^2 : d_1 + d_2 \geqslant 0\}, \qquad D(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^2 : d_1 < 0\}.$$

  Zbiory te mają niepuste przecięcie (patrz podwójna kratka na rys. 5.2), więc w punkcie



Rysunek 5.2: Stożek kierunków stycznych i stożek kierunków spadku w punkcie $\bar{\mathbf{x}} = [1,0]^T$.

  $[1,0]^T$ nie ma minimum.

- $\bar{\mathbf{x}} = [\frac{1}{2}, \frac{1}{2}]^T$. Zauważmy, że

  $$T(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^2 : d_1 + d_2 \geqslant 0\}, \qquad D(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^2 : \ [1,1]\mathbf{d} < 0\} = \{\mathbf{d} \in \mathbb{R}^2 : d_1 + d_2 < 0\}.$$

  Zbiory te mają zatem puste przecięcie, więc w punkcie $[\frac{1}{2}, \frac{1}{2}]^T$ **może** być minimum.

Bezpośrednie wykorzystanie twierdzenia 5.1 do szukania kandydatów na rozwiązania zadań z ograniczeniami nie wygląda zachęcająco. Dlatego postaramy się opisać prościej zbiór $T(\bar{\mathbf{x}})$ oraz warunek $T(\bar{\mathbf{x}}) \cap D(\bar{\mathbf{x}}) = \emptyset$.

5.2 Ograniczenia nierównościowe

Zajmiemy się problemem optymalizacyjnym w następującej formie:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ g_i(\mathbf{x}) \leqslant 0, \quad i = 1, \dots, m, \\ \mathbf{x} \in \mathbb{X}, \end{cases} \tag{5.3}$$

gdzie $\mathbb{X} \subset \mathbb{R}^n$ jest zbiorem otwartym i $f, g_1, \dots, g_m : \mathbb{X} \to \mathbb{R}$. A zatem

$$W = \{\mathbf{x} \in \mathbb{X} : g_1(\mathbf{x}) \leqslant 0, \dots, g_m(\mathbf{x}) \leqslant 0\}. \tag{5.4}$$

Funkcje g_i nazywane są *ograniczeniami nierównościowymi*, zaś cały problem (5.3) zadaniem optymalizacyjnym z ograniczeniami nierównościowymi.

Ustalmy $\bar{\mathbf{x}} \in W$ i załóżmy, że funkcje g_i są ciągłe. Wówczas ruch wokół $\bar{\mathbf{x}}$ ograniczają lokalnie tylko te warunki, dla których $g_i(\bar{\mathbf{x}}) = 0$. W przypadku pozostałych, z ciągłości g_i wynika, iż istnieje pewne otoczenie $\bar{\mathbf{x}}$, na którym mamy $g_i < 0$. Okazuje się, że ta obserwacja będzie pełnić ważną rolę w procesie optymalizacji z ograniczeniami nierównościowymi.

Definicja 5.3. Zbiorem *ograniczeń aktywnych* w punkcie $\bar{\mathbf{x}} \in W$ nazywamy zbiór

$$I(\bar{\mathbf{x}}) = \{i \in \{1, 2, \dots, m\} : \quad g_i(\bar{\mathbf{x}}) = 0\}.$$

Głównym wynikiem tego rozdziału będzie powiązanie własności ograniczeń aktywnych w danym punkcie $\bar{\mathbf{x}} \in W$ z lokalną geometrią tego zbioru wokół $\bar{\mathbf{x}}$. W tym celu wprowadźmy następującą definicję.

Definicja 5.4. Niech $\bar{\mathbf{x}} \in W$ i g_i różniczkowalne w $\bar{\mathbf{x}}$ dla ograniczeń aktywnych $i \in I(\bar{\mathbf{x}})$. *Stożkiem kierunków stycznych dla ograniczeń zlinearyzowanych* nazywamy zbiór

$$T_{lin}(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^n : \ \forall i \in I(\bar{\mathbf{x}}) \quad Dg_i(\bar{\mathbf{x}})\mathbf{d} \leqslant 0\}.$$

Stożek kierunków stycznych dla ograniczeń zlinearyzowanych jest zbiorem wielościennym, a zatem wypukłym i domkniętym.

Lemat 5.2. *Jeśli $\bar{\mathbf{x}} \in W$, to*

$$T(\bar{\mathbf{x}}) \subset T_{lin}(\bar{\mathbf{x}}).$$

Dowód. Dowód przebiega bardzo podobnie do dowodu twierdzenia 5.1. Weźmy $\mathbf{d} \in T(\bar{\mathbf{x}})$. Wówczas $\mathbf{d} = \lim_{k\to\infty} \lambda_k(\mathbf{x}_k - \bar{\mathbf{x}})$ dla pewnego ciągu punktów $(\mathbf{x}_k) \subset W$ zbieżnego do $\bar{\mathbf{x}}$ oraz ciągu liczb dodatnich (λ_k). Ustalmy $i \in I(\bar{\mathbf{x}})$. Z definicji różniczkowalności g_i w $\bar{\mathbf{x}}$ mamy

$$g_i(\mathbf{x}_k) = g_i(\bar{\mathbf{x}}) + Dg_i(\bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}}) + o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|).$$

Ograniczenie i-te jest aktywne w $\bar{\mathbf{x}}$. Zatem $g_i(\bar{\mathbf{x}}) = 0$. Oczywiście, $g_i(\mathbf{x}_k) \leqslant 0$, ponieważ $\mathbf{x}_k \in W$. W połączeniu w powyższym wzorem daje to następujące oszacowanie:

$$0 \geqslant g_i(\mathbf{x}_k) - g_i(\bar{\mathbf{x}}) = Dg_i(\bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}}) + o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|).$$

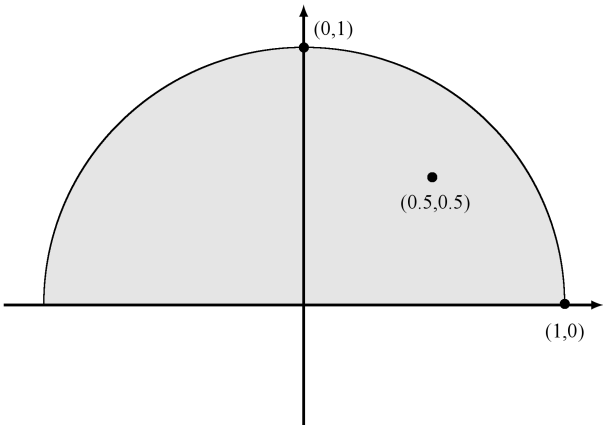
Mnożąc obie strony powyższej nierówności przez λ_k dostajemy

$$0 \geqslant Dg_i(\bar{\mathbf{x}})(\lambda_k(\mathbf{x}_k - \bar{\mathbf{x}})) + \lambda_k o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|).$$

Zrobimy teraz sztuczkę, aby rozwiązać problem $o(\|\mathbf{x}_k-\bar{\mathbf{x}}\|)$ i przejdziemy z k to nieskończoności:

$$0 \geqslant Dg_i(\bar{\mathbf{x}})\underbrace{(\lambda_k(\mathbf{x}_k-\bar{\mathbf{x}}))}_{\to\mathbf{d}} + \underbrace{\lambda_k\|\mathbf{x}_k-\bar{\mathbf{x}}\|}_{\to\|\mathbf{d}\|}\underbrace{\frac{o(\|\mathbf{x}_k-\bar{\mathbf{x}}\|)}{\|\mathbf{x}_k-\bar{\mathbf{x}}\|}}_{\to 0}.$$

Udowodniliśmy zatem, że $Dg_i(\bar{\mathbf{x}})\mathbf{d} \leqslant 0$. Analogicznie wynik otrzymujemy dla każdego ograniczenia aktywnego $i \in I(\bar{\mathbf{x}})$. A zatem $\mathbf{d} \in T_{lin}(\bar{\mathbf{x}})$. □



Rysunek 5.3: Zbiór punktów dopuszczalnych (zaznaczony na szaro) z przykładu 5.3.

Przykład 5.3. Rozważmy zbiór $W = \{\mathbf{x} \in \mathbb{R}^2 : \ x_1^2 + x_2^2 \leqslant 1, \ x_2 \geqslant 0\}$, patrz rysunek 5.3. Zapiszmy go w kanonicznej formie (5.4):

$$\mathbb{X} = \mathbb{R}^2, \quad g_1(x_1,x_2) = x_1^2 + x_2^2 - 1, \quad g_2(x_1,x_2) = -x_2.$$

Zbadajmy trzy punktu tego zbioru $(\frac{1}{2},\frac{1}{2})$, $(0,1)$ i $(1,0)$.

- $\bar{\mathbf{x}} = (\frac{1}{2},\frac{1}{2})$: $I(\bar{\mathbf{x}}) = \emptyset$ i $T(\bar{\mathbf{x}}) = T_{lin} = \mathbb{R}^2$.
- $\bar{\mathbf{x}} = (0,1)$: $I(\bar{\mathbf{x}}) = \{1\}$, $T(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^2 : \ d_2 \leqslant 0\}$,

$$T_{lin}(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^2 : \ Dg_1(0,1)\mathbf{d} \leqslant 0\} = \{\mathbf{d} \in \mathbb{R}^2 : \ [0,2]\mathbf{d} \leqslant 0\} = T(\bar{\mathbf{x}}).$$

- $\bar{\mathbf{x}} = (1,0)$: $I(\bar{\mathbf{x}}) = \{1,2\}$, $T(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^2 : \ d_1 \leqslant 0, \ d_2 \geqslant 0\}$,

$$\begin{aligned} T_{lin}(\bar{\mathbf{x}}) &= \{\mathbf{d} \in \mathbb{R}^2 : \ Dg_1(1,0)\mathbf{d} \leqslant 0, \ Dg_2(1,0)\mathbf{d} \leqslant 0\} \\ &= \{\mathbf{d} \in \mathbb{R}^2 : \ [2,0]\mathbf{d} \leqslant 0, \ [0,-1]\mathbf{d} \leqslant 0\} = T(\bar{\mathbf{x}}). \end{aligned}$$

Przykład 5.4. Rozważmy ten sam zbiór W co w powyższym przykładzie, lecz zapiszmy go nieco inaczej:

$$W = \{\mathbf{x} \in \mathbb{R}^2 : \ x_1^2 + x_2^2 \leqslant 1, \ x_2^3 \geqslant 0\}.$$

Zmianie uległo drugie ograniczenie: z $x_2 \geqslant 0$ na $x_2^3 \geqslant 0$. Nowy opis zbioru W odpowiada ograniczeniom $g_1(x_1,x_2) = x_1^2+x_2^2$, $g_2(x_1,x_2) = -x_2^3$. Rozważmy zbiory $T(\bar{\mathbf{x}})$ i $T_{lin}(\bar{\mathbf{x}})$ w punkcie $\bar{\mathbf{x}} = (1,0)$. Stożek kierunków stycznych jest identyczny, gdyż zbiór się nie zmienił:

$$T(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^2 : \ d_1 \leqslant 0, \ d_2 \geqslant 0\}.$$

Natomiast stożek kierunków stycznych dla ograniczeń zlinearyzowanych jest następujący:

$$\begin{aligned} T_{lin}(\bar{\mathbf{x}}) &= \{\mathbf{d} \in \mathbb{R}^2 : \ Dg_1(1,0)\mathbf{d} \leqslant 0, \ Dg_2(1,0)\mathbf{d} \leqslant 0\} \\ &= \{\mathbf{d} \in \mathbb{R}^2 : \ [2,0]\mathbf{d} \leqslant 0, \ [0,0]\mathbf{d} \leqslant 0\} \\ &= \{\mathbf{d} \in \mathbb{R}^2 : \ d_1 \leqslant 0\}. \end{aligned}$$

Widzimy zatem, że nie zawsze zachodzi równość pomiędzy $T(\bar{\mathbf{x}})$ i $T_{lin}(\bar{\mathbf{x}})$.

5.3 Warunki konieczne Kuhna-Tuckera

W lemacie 5.2 wykazaliśmy, że $T(\bar{\mathbf{x}}) \subset T_{lin}(\bar{\mathbf{x}})$. Pokazaliśmy w przykładach, że dość często mamy równość tych dwóch zbiorów. Okazuje się, że to bardzo ważna własność, która będzie punktem wyjścia dla całej teorii optymalizacji nieliniowej Kuhna-Tuckera. Zanim jednak przejdziemy do głównego twierdzenia dowiedziemy pomocniczy, lecz bardzo ważny lemat.

Lemat 5.3 (Farkas (1901))**.** *Niech A będzie macierzą $m \times n$ i $\mathbf{d} \in \mathbb{R}^n$. Wówczas dokładnie jeden z układów ma rozwiązanie:*

$$(1) \begin{cases} A\mathbf{x} \leqslant \mathbf{0}, \\ \mathbf{d}^T\mathbf{x} > 0, \\ \mathbf{x} \in \mathbb{R}^n, \end{cases} \qquad (2) \begin{cases} A^T\mathbf{y} = \mathbf{d}, \\ \mathbf{y} \geqslant \mathbf{0}, \\ \mathbf{y} \in \mathbb{R}^m. \end{cases}$$

Dowód. Pokażemy najpierw, że jeśli układ (2) ma rozwiązanie, to układ (1) go nie ma. Weźmy zatem $\mathbf{y}$ spełniające (2). Zatem $\mathbf{d} = A^T\mathbf{y}$. Wstawiamy to do układu (1) i otrzymujemy

$$\begin{cases} A\mathbf{x} \leqslant \mathbf{0}, \\ \mathbf{y}^T A\mathbf{x} > 0. \end{cases}$$

Pierwsza nierówność oznacza, że każda współrzędna wektora $A\mathbf{x}$ jest niedodatnia. Ponieważ współrzędne $\mathbf{y}$ są nieujemne, to iloczyn skalarny $\mathbf{y}^T(A\mathbf{x})$ jest niedodatni. Przeczy to drugiej nierówności, a zatem dowodzi, że (1) nie ma rozwiązania.

Załóżmy teraz, że układ (2) nie ma rozwiązania. Zdefiniujmy zbiór

$$V = \{\mathbf{x} \in \mathbb{R}^n : \ \mathbf{x} = A^T\mathbf{y} \text{ dla pewnego } \mathbf{y} \in \mathbb{R}^m, \ \mathbf{y} \geqslant \mathbf{0}\}.$$

Łatwo sprawdzić, że jest to zbiór wypukły i domknięty. Z faktu, że układ (2) nie ma rozwiązania wynika, że $\mathbf{d} \notin V$. Zastosujmy więc twierdzenia o ostrym oddzielaniu, tw. 3.2 do zbiorów V i $U = \{\mathbf{d}\}$. Istnieje więc wektor $\mathbf{a} \in \mathbb{R}^n$ taki że

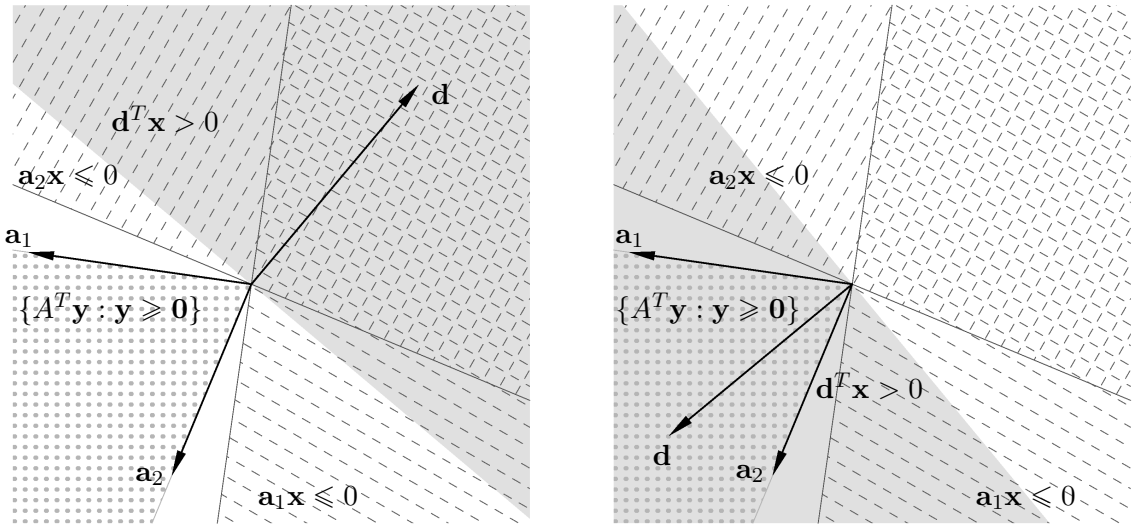
$$\mathbf{a}^T\mathbf{d} > \sup_{\mathbf{x} \in V} \mathbf{a}^T\mathbf{x}.$$

Pokażemy, że $\bar{\mathbf{x}} = \mathbf{a}$ jest rozwiązaniem układu (1). Oznaczmy $\alpha = \sup_{\mathbf{x} \in V} \bar{\mathbf{x}}^T\mathbf{x}$. Ponieważ $\mathbf{0} \in V$, to $\alpha \geqslant 0$, co pociąga $\mathbf{d}^T\bar{\mathbf{x}} > 0$ a także $\alpha < +\infty$. Pozostaje tylko udowodnić, że $A\bar{\mathbf{x}} \leqslant \mathbf{0}$. Przypuśćmy, że i-ta współrzędna $A\bar{\mathbf{x}}$ jest dodatnia. Z definicji zbioru V wynika, że dla dowolnego $\mathbf{y} \geqslant 0$ zachodzi $\alpha \geqslant \bar{\mathbf{x}}^T A^T\mathbf{y} = \mathbf{y}^T A\bar{\mathbf{x}}$. Zdefiniujmy ciąg $\mathbf{y}_k = [0, \ldots, 0, k, 0, \ldots, 0]^T$, gdzie k jest na i-tej współrzędnej. Na mocy założenia o dodatniości $(A\bar{\mathbf{x}})_i$ dostajemy

$$\lim_{k \to \infty} \mathbf{y}_k^T A\bar{\mathbf{x}} = \lim_{k \to \infty} k(A\bar{\mathbf{x}})_i = \infty,$$

co przeczy temu, że $\mathbf{y}_k^T A\bar{\mathbf{x}} \leqslant \alpha$. Sprzeczność ta dowodzi, że $A\bar{\mathbf{x}} \leqslant \mathbf{0}$. □

Rysunek 5.4 ilustruje lemat Farkasa w przypadku dwuwymiarowym. Macierz A jest wymiaru 2×2 a jej wierszami są wektory $\mathbf{a}_1$ i $\mathbf{a}_2$ (wektory te traktujemy jako wektory "wierszowe", tj. inaczej niż pozostałe wektory, które są wektorami "kolumnowymi"). Dla układu (1) z lematu, warunek $A\mathbf{x}\leqslant\mathbf{0}$ oznacza jednoczesne spełnienie dwóch nierówności $\mathbf{a}_1\mathbf{x}\leqslant 0$ i $\mathbf{a}_2\mathbf{x}\leqslant 0$. Zbiór wartości $\mathbf{x}$, dla których każda z tych nierówności jest spełniona zaznaczono przerywanymi liniami. Warunek $A\mathbf{x}\leqslant\mathbf{0}$ jest spełniony w części wspólnej tych dwóch zbiorów. Dla danego wektora $\mathbf{d}$, warunek $\mathbf{d}^T\mathbf{x}>0$ jest spełniony na zbiorze zaznaczonym szarym kolorem. Jest więc jasne, że układ (1) posiada rozwiązanie, jeśli obszar szary i obszar z podwójnym kreskowaniem posiadają część wspólną. Ten przypadek jest zilustrowany lewym obrazkiem na rysunku. Na prawym obrazku część wspólna jest zbiorem pustym, dlatego prawy obrazek odpowiada przypadkowi, gdy układ (1) nie posiada rozwiązania. Aby geometrycznie zinterpretować układ (2) musimy zobaczyć jak wygląda zbiór $V=\{A^T\mathbf{y}:\mathbf{y}\geqslant\mathbf{0}\}$. Wektory $A^T\mathbf{y}$ można zapisać jako kombinację liniową $y_1\mathbf{a}_1^T+y_2\mathbf{a}_2^T$, gdzie $\mathbf{y}=(y_1,y_2)^T$. Łatwo zauważyć, że dla $y_1\geqslant 0, y_2\geqslant 0$ jest to stożek rozpięty przez wektory $\mathbf{a}_1^T$ i $\mathbf{a}_2^T$. Stożek ten został zakropkowany na rysunku 5.4. Równanie $A^T\mathbf{y}=\mathbf{d}$ jest spełnione, jeśli wektor $\mathbf{d}$ leży w zakropkowanym stożku. Taki przypadek ilustruje prawy obrazek. Na obrazku lewym wektor $\mathbf{d}$ leży na zewnątrz zakropkowanego stożka, dlatego układ (2) nie posiada rozwiązania dla tego przypadku.



Rysunek 5.4: Ilustracja lematu Farkasa (lemat 5.3). Na rysunku lewym układ (1) posiada rozwiązanie a układ (2) nie posiada rozwiązania. Na rysunku prawym jest odwrotnie.

Twierdzenie 5.2 (Twierdzenie Kuhna-Tuckera)**.** *Niech $\bar{\mathbf{x}}$ będzie rozwiązaniem lokalnym (5.3). Jeśli funkcje f oraz g_i, $i\in I(\bar{\mathbf{x}})$, są różniczkowalne w $\bar{\mathbf{x}}$ oraz $T(\bar{\mathbf{x}})=T_{lin}(\bar{\mathbf{x}})$, to istnieje $\mu\in[0,\infty)^m$, takie że*

$$\begin{cases} Df(\bar{\mathbf{x}})+\sum_{i\in I(\bar{\mathbf{x}})}\mu_i Dg_i(\bar{\mathbf{x}})=\mathbf{0}^T,\\ \mu_i g_i(\bar{\mathbf{x}})=0,\quad i=1,2,\dots,m.\end{cases}\tag{5.5}$$

Drugi z tych warunków nazywa się po angielsku "complementary slackness condition" i nie ma polskiego odpowiednika.

Często tezę powyższego twierdzenia zapisuje się biorąc sumę po wszystkich $i=1,\dots,m$:

$$\begin{cases} Df(\bar{\mathbf{x}})+\sum_{i=1}^m\mu_i Dg_i(\bar{\mathbf{x}})=\mathbf{0}^T,\\ \mu_i g_i(\bar{\mathbf{x}})=0,\quad i=1,2,\dots,m.\end{cases}$$

Jest to naginanie notacji, gdyż funkcje opisujące ograniczenia nieaktywne nie muszą być różniczkowalne w punkcie $\bar{\mathbf{x}}$. Z drugiej strony są one mnożone przez zerowe współczynniki μ_i. Jest to pewne usprawiedliwienie powyższej notacji, którą należy rozumieć tak, jak zapisane zostało w twierdzeniu 5.2.

Dowód twierdzenia 5.2. Na mocy twierdzenia 5.1 mamy $D(\bar{\mathbf{x}}) \cap T(\bar{\mathbf{x}}) = \emptyset$. Dalej, korzystając z założenia, dostajemy $D(\bar{\mathbf{x}}) \cap T_{lin}(\bar{\mathbf{x}}) = \emptyset$, co innymi słowy oznacza, że nie istnieje rozwiązanie $\mathbf{z} \in \mathbb{R}^n$ układu

$$\begin{cases} Df(\bar{\mathbf{x}})\mathbf{z} < 0, \\ Dg_i(\bar{\mathbf{x}})\mathbf{z} \leqslant 0, \quad i \in I(\bar{\mathbf{x}}). \end{cases}$$

Stosujemy lemat Farkasa z $\mathbf{d} = -\big(Df(\bar{\mathbf{x}})\big)^T$ i macierzą A złożoną wierszowo z gradientów ograniczeń aktywnych $Dg_i(\bar{\mathbf{x}})$, $i \in I(\bar{\mathbf{x}})$. Istnieje zatem $\mathbf{y} \in [0,\infty)^{|I(\bar{\mathbf{x}})|}$ takie że $\mathbf{y}^T A = -Df(\bar{\mathbf{x}})$ lub inaczej

$$Df(\bar{\mathbf{x}}) + \mathbf{y}^T A = \mathbf{0}^T.$$

Zdefiniujmy $\mu \in [0,\infty)^m$ następująco: $(\mu_i)_{i\in I(\bar{\mathbf{x}})} = \mathbf{y}$ i $(\mu_i)_{i\notin I(\bar{\mathbf{x}})} = \mathbf{0}$. Wówczas powyższa równość jest równoważna następującej

$$Df(\bar{\mathbf{x}}) + \sum_{i\in I(\bar{\mathbf{x}})} \mu_i Dg_i(\bar{\mathbf{x}}) = \mathbf{0}^T.$$

Z definicji μ_i oczywiste jest, że $\mu_i g_i(\bar{\mathbf{x}}) = 0$. □

Uwaga 5.1. Założenia twierdzenia Kuhna-Tuckera są trywialnie spełnione, gdy $\bar{\mathbf{x}} \in \operatorname{int} W$, tzn. gdy $I(\bar{\mathbf{x}}) = \emptyset$. Wówczas warunki (5.5) sprowadzają się do

$$Df(\bar{\mathbf{x}}) = \mathbf{0}^T, \qquad \mu = \mathbf{0}.$$

Na zakończenie zwróćmy jeszcze raz uwagę na tezę twierdzenia Kuhna-Tuckera. Warunki (5.5) nazywane są *warunkami koniecznymi pierwszego rzędu*. Wektor μ będzie pojawiał się jeszcze wiele razy na tym wykładzie. Nadajmy mu zatem nazwę:

Definicja 5.5. Wektor μ spełniający (5.5) nazywa się wektorem *mnożników Lagrange'a* w punkcie $\bar{\mathbf{x}}$.

5.4 Zadania

Ćwiczenie 5.1. Wykaż, że zbiór $T(\bar{\mathbf{x}})$ dla $\bar{\mathbf{x}} \in \operatorname{cl} W$ jest stożkiem.

Ćwiczenie 5.2. Udowodnij, że stożek kierunków stycznych $T(\bar{\mathbf{x}})$ jest zbiorem domkniętym.

Ćwiczenie 5.3. Udowodnij tożsamość (5.2).

Ćwiczenie 5.4. Znajdź stożek kierunków stycznych do zbioru W w punkcie $\bar{\mathbf{x}} = \mathbf{0}$, gdy

1. $W = \{(x_1, x_2) \in \mathbb{R}^2 : \ x_2 \geqslant -x_1^3\}$,
2. $W = \{(x_1, x_2) \in \mathbb{R}^2 : \ x_1 \in \mathbb{Z}, \ x_2 = 0\}$,
3. $W = \{(x_1, x_2) \in \mathbb{R}^2 : \ x_1 \in \mathbb{Q}, \ x_2 = 0\}$.

Ćwiczenie 5.5. Udowodnić, że dla zadania

$$\begin{cases} f(\mathbf{x}) \to \min, \\ g_i(\mathbf{x}) \leqslant 0, \qquad i = 1, \dots, m, \\ x_i \geqslant 0, \qquad i = 1, \dots, n, \end{cases}$$

warunek konieczny pierwszego rzędu przyjmuje postać:

$$\begin{cases} Df(\mathbf{x}) + \sum_{i \in I(\mathbf{x})} \mu_i Dg_i(\mathbf{x}) \geqslant \mathbf{0}^T, \\ \left(Df(\mathbf{x}) + \sum_{i \in I(\mathbf{x})} \mu_i Dg_i(\mathbf{x}) \right) \mathbf{x} = 0, \\ \mu_i g_i(\mathbf{x}) = 0, \qquad i = 1, \dots, m, \\ \mu_i \geqslant 0, \qquad i = 1, \dots, m. \end{cases}$$

Nierówność dla wektorów oznacza nierówność po współrzędnych.

Rozdział 6

Warunki regularności i przykłady

W tym rozdziale podamy warunki dostateczne równości stożka kierunków stycznych $T(\bar{\mathbf{x}})$ i stożka kierunków stycznych dla ograniczeń zlinearyzowanych $T_{lin}(\bar{\mathbf{x}})$. Przypomnijmy, że jest to główne założenie twierdzenia Kuhna-Tuckera, tw. 5.2, opisującego warunek konieczny pierwszego rzędu dla lokalnego rozwiązania problemu optymalizacyjnego z ograniczeniami nierównościowymi (5.3).

6.1 Warunki regularności

Sformułujemy teraz trzy warunki dostateczne równości $T(\bar{\mathbf{x}}) = T_{lin}(\bar{\mathbf{x}})$, zwane *warunkami regularności*. Dowody dostateczności podamy w kolejnych twierdzeniach.

Definicja 6.1. W punkcie $\bar{\mathbf{x}} \in W$ spełniony jest:

- *warunek liniowej niezależności*, jeśli funkcje g_i, $i \notin I(\bar{\mathbf{x}})$, są ciągłe w $\bar{\mathbf{x}}$ oraz wektory $Dg_i(\bar{\mathbf{x}})$, dla $i \in I(\bar{\mathbf{x}})$, są liniowo niezależne,
- *warunek afiniczności*, jeśli funkcje g_i, $i \in I(\bar{\mathbf{x}})$, są afiniczne oraz funkcje g_i, $i \notin I(\bar{\mathbf{x}})$, są ciągłe w $\bar{\mathbf{x}}$,
- *warunek Slatera*, jeśli funkcje g_i, $i \in I(\bar{\mathbf{x}})$ są pseudowypukłe w $\bar{\mathbf{x}}$ $(Df(\bar{\mathbf{x}})(\mathbf{y} - \bar{\mathbf{x}}) \geqslant 0 \implies f(\mathbf{y}) \geqslant f(\bar{\mathbf{x}}))$, funkcje g_i, $i \notin I(\bar{\mathbf{x}})$, są ciągłe w $\bar{\mathbf{x}}$ oraz istnieje $\mathbf{x} \in \mathbb{X}$, dla którego $g_i(\mathbf{x}) < 0$ dla $i \in I(\bar{\mathbf{x}})$.

Zauważmy, że w warunku Slatera nie wymagamy, aby punkt $\mathbf{x}$ spełniał warunki ograniczeń nieaktywnych, tzn. $\mathbf{x}$ nie musi być w zbiorze W.

Twierdzenie 6.1. *Jeśli w punkcie $\bar{\mathbf{x}} \in W$ spełniony jest warunek afiniczności, to $T(\bar{\mathbf{x}}) = T_{lin}(\bar{\mathbf{x}})$.*

Dowód. Zawieranie $T(\bar{\mathbf{x}}) \subset T_{lin}(\bar{\mathbf{x}})$ wynika z lematu 5.2. Wystarczy zatem wykazać zawieranie w drugą stronę.

Niech $\mathbf{d} \in T_{lin}(\bar{\mathbf{x}})$. Udowodnimy, że istnieje $\lambda^* > 0$, taka że cały odcinek $\bar{\mathbf{x}} + \lambda\mathbf{d}$, $\lambda \in [0, \lambda^*]$, zawarty jest w W:

$$\{\bar{\mathbf{x}} + \lambda\mathbf{d} : \lambda \in [0, \lambda^*]\} \subset W. \tag{6.1}$$

Zauważmy, że $g_i(\bar{\mathbf{x}}) < 0$ dla $i \notin I(\bar{\mathbf{x}})$. Z ciągłości tych funkcji w $\bar{\mathbf{x}}$ wynika, że istnieje $\lambda^* > 0$, dla której $g_i(\bar{\mathbf{x}} + \lambda\mathbf{d}) \leqslant 0$, $i \notin I(\bar{\mathbf{x}})$, $\lambda \in [0, \lambda^*]$. Pozostaje jeszcze wykazać, że nierówność taka

zachodzi dla ograniczeń aktywnych. Ustalmy $i \in I(\bar{\mathbf{x}})$. Z definicji $\mathbf{d}$ wiemy, że $Dg_i(\bar{\mathbf{x}})\mathbf{d} \leqslant 0$. Ograniczenie g_i jest afiniczne, czyli postaci $g_i(\mathbf{x}) = \mathbf{a}_i^T\mathbf{x} + b_i$, gdzie $\mathbf{a}_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$. Zatem $Dg_i(\bar{\mathbf{x}})d = \mathbf{a}_i^T\mathbf{d}$. Z faktu, że g_i jest aktywne w $\bar{\mathbf{x}}$ dostajemy również $0 = g_i(\bar{\mathbf{x}}) = \mathbf{a}_i^T\bar{\mathbf{x}} + b_i$. Stąd dla dowolnego $\lambda \geqslant 0$ mamy

$$0 \geqslant \lambda\mathbf{a}_i^T\mathbf{d} = \lambda\mathbf{a}_i^T\mathbf{d} + \underbrace{\mathbf{a}_i^T\bar{\mathbf{x}} + b_i}_{0} = \mathbf{a}_i^T(\bar{\mathbf{x}} + \lambda\mathbf{d}) + b_i = g_i(\bar{\mathbf{x}} + \lambda\mathbf{d}).$$

Dowodzi to (6.1).

Pozostaje już tylko skonstruować ciąg $(\mathbf{x}_k) \subset W$, $\mathbf{x}_k \to \bar{\mathbf{x}}$ i $(\lambda_k) \subset (0, \infty)$. Kładziemy

$$\mathbf{x}_k = \bar{\mathbf{x}} + \frac{\lambda^*}{k}\mathbf{d}, \qquad \lambda_k = \frac{k}{\lambda^*}.$$

Wówczas $\mathbf{x}_k \in W$, $\mathbf{x}_k \to \bar{\mathbf{x}}$ oraz $\lambda_k(\mathbf{x}_k - \bar{\mathbf{x}}) = \mathbf{d}$, czyli trywialnie

$$\lim_{k\to\infty} \lambda_k(\mathbf{x}_k - \bar{\mathbf{x}}) = \mathbf{d},$$

a zatem $\mathbf{d} \in T(\bar{\mathbf{x}})$. □

Przed przystąpieniem do dowodu analogicznych twierdzeń dla pozostałych warunków regularności sformułujemy pomocniczy lemat. Wprowadźmy zbiór

$$T_{int}(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^n : \ \forall\, i \in I(\bar{\mathbf{x}}) \quad Dg_i(\bar{\mathbf{x}})\mathbf{d} < 0\}.$$

Lemat 6.1. *Jeśli w punkcie $\bar{\mathbf{x}} \in W$ funkcje g_i, $i \in I(\bar{\mathbf{x}})$, są różniczkowalne, zaś funkcje g_i, $i \notin I(\bar{\mathbf{x}})$, są ciągłe, to $\mathbf{d} \in T_{int}(\bar{\mathbf{x}})$ oznacza, że $\bar{\mathbf{x}} + \lambda\mathbf{d} \in \operatorname{int} W$ dla dostatecznie małych $\lambda > 0$.*

Dowód. Dla funkcji g_i, $i \notin I(\bar{\mathbf{x}})$, mamy $g_i(\bar{\mathbf{x}}) < 0$. Z ciągłości g_i wynika, że dla dostatecznie małych $\lambda > 0$ $g_i(\bar{\mathbf{x}} + \lambda\mathbf{d}) < 0$. Dla funkcji g_i, $i \in I(\bar{\mathbf{x}})$, różniczkowalnych w $\bar{\mathbf{x}}$ mamy

$$\lim_{\lambda\downarrow 0} \frac{g_i(\bar{\mathbf{x}} + \lambda\mathbf{d}) - g_i(\bar{\mathbf{x}})}{\lambda} = Dg_i(\bar{\mathbf{x}})\mathbf{d} < 0,$$

bo $\mathbf{d} \in T_{int}(\bar{\mathbf{x}})$. Nierówność się zachowuje dla dostatecznie małych $\lambda > 0$. Mamy więc

$$\frac{g_i(\bar{\mathbf{x}} + \lambda\mathbf{d}) - g_i(\bar{\mathbf{x}})}{\lambda} < 0,$$

czyli $g_i(\bar{\mathbf{x}} + \lambda\mathbf{d}) - g_i(\bar{\mathbf{x}}) < 0$, a ponieważ $g_i(\bar{\mathbf{x}}) = 0$ więc $g_i(\bar{\mathbf{x}} + \lambda\mathbf{d}) < 0$. □

Lemat 6.2. *Niech $\bar{\mathbf{x}} \in W$ oraz funkcje g_i, $i \in I(\bar{\mathbf{x}})$, są różniczkowalne w $\bar{\mathbf{x}}$, zaś funkcje g_i, $i \notin I(\bar{\mathbf{x}})$ są ciągłe w $\bar{\mathbf{x}}$. Wówczas*

I) $T_{int}(\bar{\mathbf{x}}) \subset T(\bar{\mathbf{x}})$,

II) Jeśli $T_{int}(\bar{\mathbf{x}}) \neq \emptyset$, to $\operatorname{cl}(T_{int}(\bar{\mathbf{x}})) = T_{lin}(\bar{\mathbf{x}})$.

Dowód. Dowód (I) wynika z Lematu 6.1. Do dowodu (II) zauważmy, że $T_{int}(\bar{\mathbf{x}})$ jest wnętrzem zbioru $T_{lin}(\bar{\mathbf{x}})$. Ponadto $T_{lin}(\bar{\mathbf{x}})$ jest zbiorem wielościennym, a zatem wypukłym i domkniętym (patrz lemat 4.5). Zastosowanie lematu 3.2 kończy dowód. □

Twierdzenie 6.2. *Jeśli w punkcie $\bar{\mathbf{x}} \in W$ spełniony jest warunek Slatera, to $T(\bar{\mathbf{x}}) = T_{lin}(\bar{\mathbf{x}})$.*

Dowód. Pokażemy najpierw, że $T_{int}(\bar{\mathbf{x}}) \neq \emptyset$. Niech $\mathbf{x} \in \mathbb{X}$ taki że $g_i(\mathbf{x}) < 0$ dla $i \in I(\bar{\mathbf{x}})$. Z pseudowypukłości g_i w punkcie $\bar{\mathbf{x}}$ dostajemy $Dg_i(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) < 0$ dla $i \in I(\bar{\mathbf{x}})$, czyli $(\mathbf{x} - \bar{\mathbf{x}}) \in T_{int}(\bar{\mathbf{x}})$.

Na mocy lematu 6.2 mamy $T_{lin}(\bar{\mathbf{x}}) = \mathrm{cl}(T_{int}(\bar{\mathbf{x}}))$. Udowodniliśmy także, że $T_{int}(\bar{\mathbf{x}}) \subset T(\bar{\mathbf{x}}) \subset T_{lin}(\bar{\mathbf{x}})$ oraz, że $T(\bar{\mathbf{x}})$ jest domknięty. A zatem $T_{lin}(\bar{\mathbf{x}}) = T(\bar{\mathbf{x}})$. □

Dowód analogicznego twierdzenia dla warunku liniowej niezależności wymaga pomocniczego lematu w stylu lematu Farkasa. Wynik ten jest jednak wcześniejszy i został uzyskany przez Paula Gordana w 1873 roku.

Lemat 6.3 (Gordan, 1873)**.** *Niech A będzie macierzą $m \times n$. Wówczas dokładnie jeden z układów ma rozwiązanie:*

$$(1)\ \begin{cases} A\mathbf{x} < \mathbf{0}, \\ \mathbf{x} \in \mathbb{R}^n, \end{cases} \qquad (2)\ \begin{cases} A^T\mathbf{y} = \mathbf{0}, \\ \mathbf{y} \geqslant \mathbf{0},\ \mathbf{y} \neq \mathbf{0} \\ \mathbf{y} \in \mathbb{R}^m. \end{cases}$$

Dowód. Dowód rozpoczniemy od uzasadnienia, że oba te układy nie mogą mieć rozwiązania jednocześnie. Załóżmy więc przez sprzeczność, że istnieją $\mathbf{x}, \mathbf{y}$ spełniające (1)-(2). Z faktu, że $\mathbf{y}$ rozwiązuje (2), dostajemy, że $\mathbf{y}^T A\mathbf{x} = 0$. Z drugiej strony $\mathbf{x}$ rozwiązuje (1), czyli $(Ax)_i < 0$ dla każdego $i = 1, \ldots, m$. Pamiętając, że $y_j \geqslant 0$, $j = 1, \ldots, m$, i $\mathbf{y} \neq \mathbf{0}$ mamy $\mathbf{y}^T A\mathbf{x} < 0$. Dostaliśmy więc sprzeczność, co dowodzi, że nie może istnieć jednocześnie rozwiązanie (1) i (2).

Następnym krokiem dowodu będzie wykazanie, że zawsze któryś z układów ma rozwiązanie. Udowodnimy, że jeśli układ (1) nie ma rozwiązania, to układ (2) ma rozwiązanie. W tym celu zdefiniujmy następujące zbiory wypukłe:

$$U = (-\infty, 0)^m, \qquad V = \{\mathbf{z} \in \mathbb{R}^m :\ \mathbf{z} = A\mathbf{x} \text{ dla pewnego } \mathbf{x} \in \mathbb{R}^n\}.$$

Z faktu, że układ (1) nie ma rozwiązania, wynika, iż zbiory te są rozłączne. Twierdzenie o oddzielaniu implikuje istnienie wektora $\mathbf{a} \in \mathbb{R}^m$, $\mathbf{a} \neq \mathbf{0}$, takiego że

$$\sup_{\mathbf{z} \in U} \mathbf{a}^T\mathbf{z} \leqslant \inf_{\mathbf{z} \in V} \mathbf{a}^T\mathbf{z}.$$

Wykażemy, że $\mathbf{a} \geqslant \mathbf{0}$. Przypuśćmy przeciwnie, że istnieje współrzędna $a_i < 0$. Rozważmy ciąg $\mathbf{z}_k = [-\frac{1}{k}, -\frac{1}{k}, \ldots, -k, \ldots, -\frac{1}{k}]^T$, gdzie $(-k)$ jest na i-tej pozycji. Wówczas $\mathbf{z}_k \in U$ oraz $\lim_{k\to\infty} \mathbf{a}^T\mathbf{z}_k = \infty$, a zatem dostaliśmy sprzeczność, gdyż $\sup_{\mathbf{z}\in U} \mathbf{a}^T\mathbf{z}$ jest skończone ($\mathbf{0} \in V$ więc $\inf_{\mathbf{z}\in V} \mathbf{a}^T\mathbf{z} \leqslant 0$).

Wiemy zatem, że wektor $\mathbf{a}$ ma wszystkie współrzędne nieujemne. Pociąga to $\sup_{\mathbf{z}\in U} \mathbf{a}^T\mathbf{z} = 0$. Weźmy $\mathbf{z} = A(-A^T\mathbf{a})$. Wówczas $\mathbf{z} \in V$, czyli $\mathbf{a}^T\mathbf{z} \geqslant 0$. A zatem

$$0 \leqslant \mathbf{a}^T\mathbf{z} = \mathbf{a}^T A(-A^T\mathbf{a}) = -\|A^T\mathbf{a}\|^2,$$

co implikuje, że $\|A^T\mathbf{a}\| = 0$, czyli $A^T\mathbf{a} = \mathbf{0}$. Rozwiązaniem układu (2) jest zatem $\mathbf{y} = \mathbf{a}$. □

Twierdzenie 6.3. *Jeśli w punkcie $\bar{\mathbf{x}} \in W$ spełniony jest warunek liniowej niezależności, to $T(\bar{\mathbf{x}}) = T_{lin}(\bar{\mathbf{x}})$.*

Dowód. Identycznie jak w dowodzie tw. 6.2 wystarczy pokazać, że $T_{int}(\bar{\mathbf{x}}) \neq \emptyset$. Niech A będzie macierzą składającą się z gradientów $Dg_i(\bar{\mathbf{x}})$ ograniczeń aktywnych (jako wierszy). Na mocy liniowej niezależności nie istnieje wektor $\mu \in \mathbb{R}^{|I(\bar{\mathbf{x}})|}$, $\mu \neq \mathbf{0}$, o tej własności, że $A^T\mu = \mathbf{0}$. Nie

istnieje więc rozwiązanie układu (2) w lemacie 6.3. A zatem ma rozwiązanie układ (1), czyli istnieje $\mathbf{d} \in \mathbb{R}^n$, takie że $A\mathbf{d} < \mathbf{0}$:

$$Dg_i(\bar{\mathbf{x}})\mathbf{d} < 0, \qquad \forall\, i \in I(\bar{\mathbf{x}}).$$

Stąd $\mathbf{d} \in T_{int}(\bar{\mathbf{x}})$. □

6.2 Przykłady

Przykład 6.1. Rozważmy problem optymalizacji na zbiorze:

$$W = \{\mathbf{x} \in \mathbb{R}^2 :\ x_1^2 + x_2^2 \leqslant 1, \quad x_1 + 2x_2 \leqslant 1, \quad x_1 - 3x_2 \leqslant 1\}.$$

Zbiór $\mathbb{X} = \mathbb{R}^2$. Ograniczenia są opisane przez trójkę funkcji

$$g_1(x_1, x_2) = x_1^2 + x_2^2 - 1, \qquad g_2(x_1, x_2) = x_1 + 2x_2 - 1, \qquad g_3(x_1, x_2) = x_1 - 3x_2 - 1.$$

W punkcie $\bar{\mathbf{x}} = [1, 0]^T$ wszystkie ograniczenia są aktywne i mamy

$$Dg_1(\bar{\mathbf{x}}) = [2, 0], \qquad Dg_2(\bar{\mathbf{x}}) = [1, 2], \qquad Dg_3(\bar{\mathbf{x}}) = [1, -3].$$

Warunek liniowej niezależności ograniczeń nie jest spełniony w $\bar{\mathbf{x}}$. Ograniczenia aktywne nie są również afiniczne. Pozostaje zatem sprawdzić warunek Slatera. Wszystkie funkcje ograniczeń są wypukłe, a więc także pseudowypukłe. W punkcie $\mathbf{x} = [0, 0]^T$ mamy $g_i(\mathbf{x}) = -1 < 0$ dla $i = 1, 2, 3$. Zachodzi więc warunek Slatera.

Przykład 6.2 (Kuhn, Tucker (1951))**.** Rozważmy zadanie optymalizacyjne:

$$\begin{cases} x_1 \to \min, \\ x_2 \leqslant x_1^3, \\ x_2 \geqslant 0. \end{cases}$$

Zbiór punktów dopuszczalnych W naszkicowany jest na rysunku 6.1. Zapiszmy funkcje opisujące ograniczenia:

$$g_1(x_1, x_2) = -x_1^3 + x_2, \qquad g_2(x_1, x_2) = -x_2.$$

Zauważmy, że w każdym punkcie dopuszczalnym, poza $[0, 0]^T$, spełniony jest warunek liniowej niezależności. Widać jednak, że rozwiązaniem powyższego problemu optymalizacyjnego jest $\bar{\mathbf{x}} = [0, 0]^T$. Niestety w tym punkcie $T(\bar{\mathbf{x}}) \neq T_{lin}(\bar{\mathbf{x}})$:

$$T(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^2 :\ d_1 \geqslant 0, \quad d_2 = 0\}, \qquad T_{lin}(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^2 :\ d_2 = 0\},$$

bo $Dg_1 = [0, 1]$, $Dg_2 = [0, -1]$ wiec $Dg_1\mathbf{d} \leqslant 0 \implies d_2 \leqslant 0$ oraz $Dg_2\mathbf{d} \leqslant 0 \implies d_2 \geqslant 0$. Z drugiej strony

$$D(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^2 :\ d_1 < 0\},$$

bo $Df = [1, 0]$ więc $Df\mathbf{d} < 0 \implies d_1 < 0$.

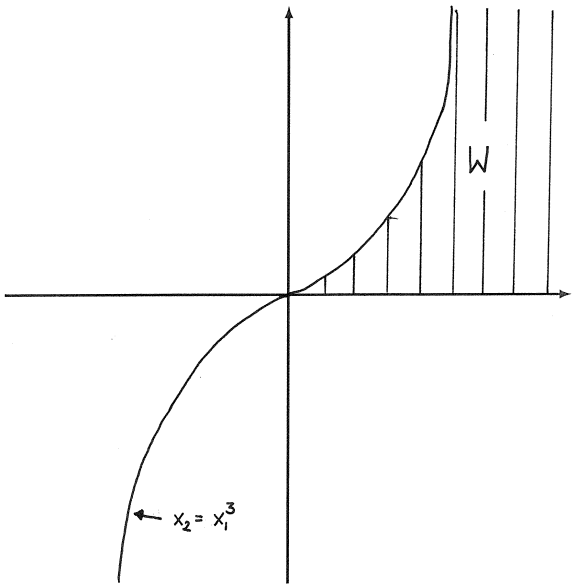
Dla zadania optymalizacyjnego

$$\begin{cases} x_2 \to \min, \\ x_2 \leqslant x_1^3, \\ x_2 \geqslant 0, \end{cases}$$

mamy dalej $T(\bar{\mathbf{x}}) \neq T_{lin}(\bar{\mathbf{x}})$. Ale dla tego nowego zadania $D(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^2 :\ d_2 < 0\}$. Wynika stąd równość

$$T(\bar{\mathbf{x}}) \cap D(\bar{\mathbf{x}}) = T_{lin}(\bar{\mathbf{x}}) \cap D(\bar{\mathbf{x}}) = \emptyset.$$

Ten ostatni warunek wystarczy, aby prawdziwa była teza twierdzenia Kuhna-Tuckera 5.2.



Rysunek 6.1: Zbiór punktów dopuszczalnych dla przykładu Kuhna-Tuckera.

Przykład 6.3. Niech A będzie macierzą symetryczną $n \times n$. Rozważmy zadanie optymalizacyjne:

$$\begin{cases} \mathbf{x}^T A\mathbf{x} \to \max, \\ \|\mathbf{x}\| \leqslant 1, \\ \mathbf{x} \in \mathbb{R}^n. \end{cases}$$

Zapiszmy najpierw problem optymalizacyjny w kanonicznej formie. Zauważmy przy tym, że warunek $\|\mathbf{x}\| \leqslant 1$ jest równoważny $\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} \leqslant 1$.

$$\begin{cases} -\mathbf{x}^T A\mathbf{x} \to \min, \\ \mathbf{x}^T\mathbf{x} - 1 \leqslant 0, \\ \mathbf{x} \in \mathbb{R}^n. \end{cases}$$

Oznaczmy przez W zbiór punktów dopuszczalnych. Zauważmy, że w każdym punkcie dopuszczalnym spełnione są warunki liniowej niezależności i Slatera. Wynika stąd, że rozwiązanie powyższego zadania spełnia warunki konieczne pierwszego rzędu (warunki Kuhna-Tuckera (5.5)). Znajdziemy teraz wszystkie punkty „podejrzane".

Warunki Kuhna-Tuckera mają następującą postać:

$$\begin{cases} -2\mathbf{x}^T A + 2\mu\mathbf{x}^T = \mathbf{0}^T, \\ \mu(\mathbf{x}^T\mathbf{x} - 1) = 0, \\ \mu \geqslant 0, \quad \mathbf{x} \in \mathbb{R}^n. \end{cases}$$

Przypadek $\mathbf{x}^T\mathbf{x} - 1 < 0$. Wówczas $\mu = 0$ (z powodu drugiego równania) i pierwsze równanie przyjmuje postać $\mathbf{x}^T A = \mathbf{0}^T$, którą poprzez transponowanie obu stron sprowadzamy do

$$A\mathbf{x} = \mathbf{0}.$$

Równanie to spełnione jest przez wszystkie $\mathbf{x} \in \ker A$, $\|\mathbf{x}\| < 1$. W szczególności ma ono co najmniej jedno rozwiązanie ($\mathbf{x} = \mathbf{0}$).

Przypadek $\mathbf{x}^T\mathbf{x} - 1 = 0$. Teraz nic nie możemy powiedzieć o μ. Może być zerowe lub dodatnie. Pierwsze równanie przyjmuje jednak postać $\mu\mathbf{x}^T = \mathbf{x}^T A$, które po transponowaniu wygląda następująco:

$$\mu\mathbf{x} = A\mathbf{x}.$$

Rozwiązaniami są zatem wektory własne ($\mathbf{x}$) odpowiadające nieujemnym wartościom własnym (μ). Zbiór rozwiązań może być pusty, jeśli wszystkie wartości własne są ujemne.

Podsumowując, zbiór punktów podejrzanych, czyli punktów, w których spełniony jest warunek konieczny pierwszego rzędu, ma postać:

$$\{\mathbf{x} \in \ker A : \ \|\mathbf{x}\| < 1\}$$
$$\cup \{\mathbf{x} \in \mathbb{R}^2 : \ \|\mathbf{x}\| = 1 \text{ oraz } \mathbf{x} \text{ jest wektorem własnym } A \text{ dla nieujemnej wartości własnej}\}.$$

Na obecnym etapie nie mamy żadnej techniki pozwalającej na znalezienie rozwiązania. Możemy tylko posłużyć się zdrowym rozsądkiem. Otóż, jeśli $\mathbf{x} \in \ker A$, to wartość funkcji celu $\mathbf{x}^T A\mathbf{x}$ jest zerowa. Dla dowolnego elementu drugiego zbioru, wartość funkcji celu jest równa wartości własnej. Możemy zatem wyciągnąć następujący wniosek: jeśli maksymalna wartość własna jest dodatnia, to każdy wektor jej odpowiadający jest rozwiązaniem globalnym. Jeśli macierz A nie ma wartości własnych większych od zera, to rozwiązaniem jest dowolny punkt z jądra A o normie nie większej niż 1.

Przykład 6.4. Rozważmy zadanie optymalizacyjne:

$$\begin{cases} x_1 + x_2 \to \min, \\ x_2 \geqslant x_1^2, \\ x_2 \leqslant 0. \end{cases}$$

Łatwo zauważyć, że rozwiązaniem tego zadania jest punkt $\bar{\mathbf{x}} = (0, 0)$. Z drugiej strony mamy $Df = [1, 1]$, $Dg_1 = [2x_1, -1]$ oraz $Dg_2 = [0, 1]$. Widać więc, że dla dowolnych stałych μ_1 i μ_2

$$Df(\bar{\mathbf{x}}) + \mu_1 Dg_1(\bar{\mathbf{x}}) + \mu_2 Dg_2(\bar{\mathbf{x}}) \neq \mathbf{0}.$$

Czyli nie zachodzi teza twierdzenia Kuhna-Tuckera. Ponieważ $T(\bar{\mathbf{x}}) = \{\mathbf{0}\}$ a $T_{lin}(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^2 : \ d_2 = 0\}$, więc $T(\bar{\mathbf{x}}) \neq T_{lin}(\bar{\mathbf{x}})$. Mamy także $\emptyset = T(\bar{\mathbf{x}}) \cap D(\bar{\mathbf{x}}) \neq T_{lin}(\bar{\mathbf{x}}) \cap D(\bar{\mathbf{x}})$, bo $D(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^2 : \ d_1 + d_2 < 0\}$. Wynika stąd, że w punkcie $\bar{\mathbf{x}}$ nie są spełnione założenia twierdzenia Kuhna-Tuckera 5.2.

6.3 Zadania

Ćwiczenie 6.1. Przez "punkty podejrzane" problemu optymalizacyjnego rozumiemy takie punkty dopuszczalne, w których nie są spełnione warunki regularności albo spełnione są warunki konieczne pierwszego rzędu. Znajdź wszystkie punkty podejrzane dla następującego problemu optymalizacyjnego:

$$\begin{cases} x^2 - 6x + y^2 + 2y \to \min, \\ x + 2y - 10 = 0, \\ 25 - x^2 - y^2 \geqslant 0. \end{cases}$$

Czy w którymś z nich jest rozwiązanie? Jeśli tak, to w którym? Odpowiedź uzasadnij.

Ćwiczenie 6.2. Znajdź zbiór rozwiązań problemu optymalizacyjnego

$$\begin{cases} \sum_{i=1}^{n} c_i x_i^2 \to \min, \\ \sum_{i=1}^{n} x_i = b, \end{cases}$$

gdzie $b, (c_i) > 0$. Uzasadnij, że są to wszystkie rozwiązania.

Ćwiczenie 6.3. Znajdź rozwiązania globalne problemu optymalizacyjnego

$$\begin{cases} x_1^2 + 3x_2^2 - x_1 \to \min, \\ x_1^2 - x_2 \leqslant 1, \\ x_1 + x_2 \geqslant 1. \end{cases}$$

Ćwiczenie 6.4. Znajdź minima i maksima globalne funkcji

$$f(x_1, x_2) = 4x_1^2 + 2x_2^2 - 6x_1x_2 + x_1$$

na zbiorze

$$W = \{(x_1, x_2) \in \mathbb{R}^2 : \ -2x_1 + 2x_2 \geqslant 1, \ 2x_1 - x_2 \leqslant 0, \ x_1 \leqslant 0, \ x_2 \geqslant 0\}.$$

Rozdział 7

Funkcje quasi-wypukłe i warunki dostateczne

Mogliśmy już zaobserwować na kilku przykładach, że wypukłość znacznie ułatwia rozwiązywanie problemów optymalizacyjnych. W rozdziale 3 zauważyliśmy, że warunkiem koniecznym i dostatecznym minimum funkcji wypukłej jest zerowanie się pochodnej. Później wykazaliśmy, że identyczny warunek zachodzi dla większej rodziny funkcji: funkcji pseudowypukłych. W rozdziale 4 zajmowaliśmy się maksymalizacją funkcji wypukłej na zbiorze wypukłym i zwartym. Udowodniliśmy, że maksimum jest przyjmowane w jednym z punktów ekstremalnych tego zbioru. W tym rozdziale rozszerzymy rodzinę funkcji, dla których jest to prawdą; wprowadzimy własność quasi-wypukłości.

Patrząc na powyższą listę można się domyślać, że wypukłość może również pomagać przy rozwiązywaniu zadań z ograniczeniami nierównościowymi. Jeśli założymy, że funkcje g_i, opisujące ograniczenia nierównościowe, są wypukłe lub, ogólniej, quasi-wypukłe oraz funkcja f jest pseudowypukła, to spełnienie warunku pierwszego rzędu w pewnym punkcie $\bar{\mathbf{x}}$ jest wystarczające, by stwierdzić jego optymalność. Dowód powyższego faktu przytoczymy pod koniec tego rozdziału.

7.1 Quasi-wypukłość

W tym podrozdziale rozszerzymy rodzinę funkcji, dla których maksimum znajduje się w punktach ekstremalnych dziedziny.

Definicja 7.1. Niech $W \subset \mathbb{R}^n$ będzie zbiorem wypukłym, zaś $f : W \to \mathbb{R}$.

Funkcję f nazywamy *quasi-wypukłą*, jeśli dla dowolnych $\mathbf{x}, \mathbf{y} \in W$ i $\lambda \in [0, 1]$ mamy

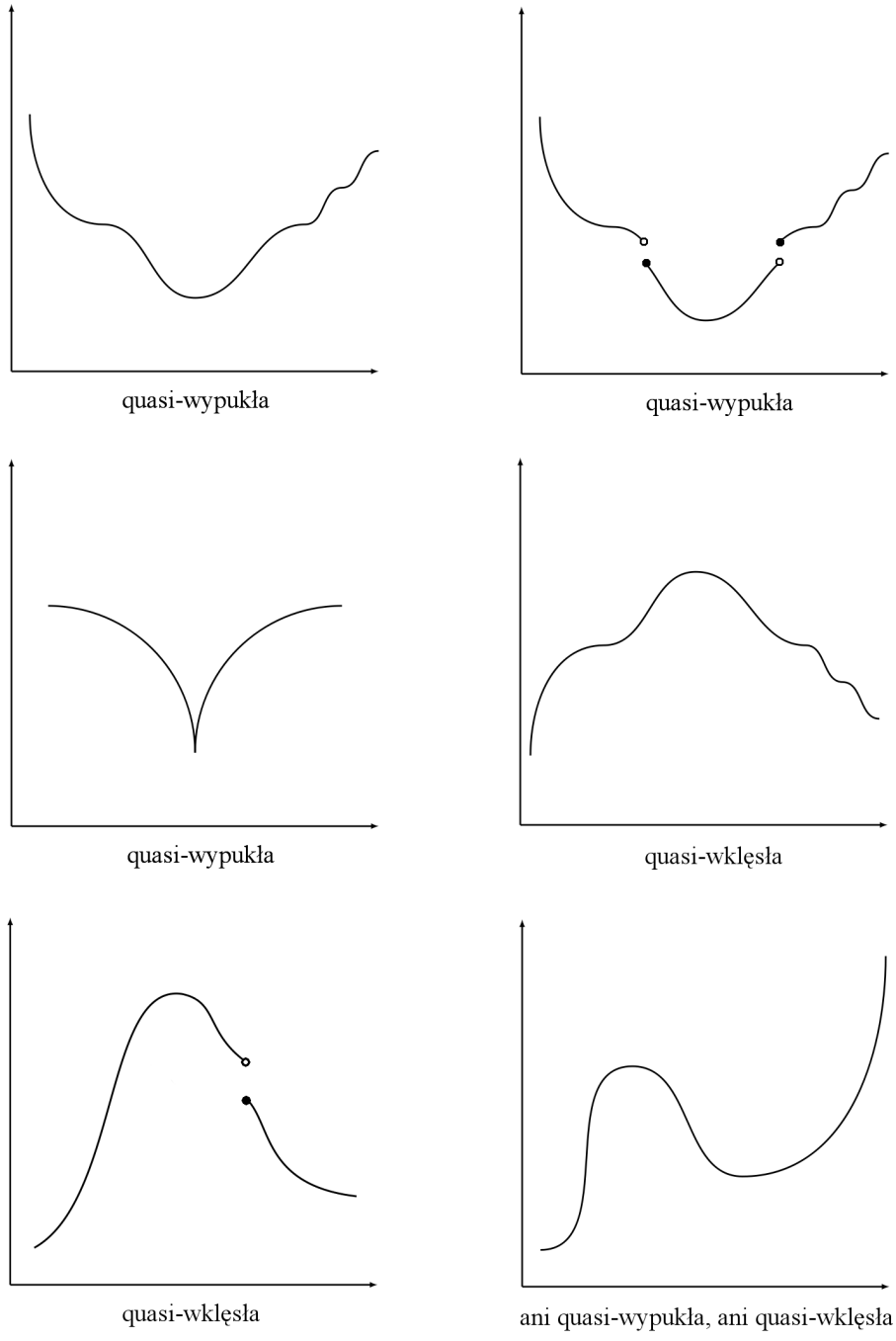
$$f\big(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\big) \leqslant \max\big(f(\mathbf{x}), f(\mathbf{y})\big).$$

Funkcję f nazywamy *quasi-wklęsłą*, jeśli funkcja $(-f)$ jest quasi-wypukła, tzn. dla dowolnych $\mathbf{x}, \mathbf{y} \in W$ i $\lambda \in [0, 1]$ zachodzi

$$f\big(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\big) \geqslant \min\big(f(\mathbf{x}), f(\mathbf{y})\big).$$

Funkcję f nazywamy *quasi-liniową*, jeśli jest ona jednocześnie quasi-wypukła i quasi-wklęsła.

Na rys. 7.1 pokazane są przykłady jednowymiarowych funkcji quasi-wypukłych i quasi-wklęsłych. Zwróćmy uwagę na to, że funkcje takie nie muszą być ciągłe, nie mówiąc już o różniczkowalności. Ciekawym jest również uogólnienie rodziny funkcji afinicznych do funkcji quasi-liniowych, patrz rys. 7.2.

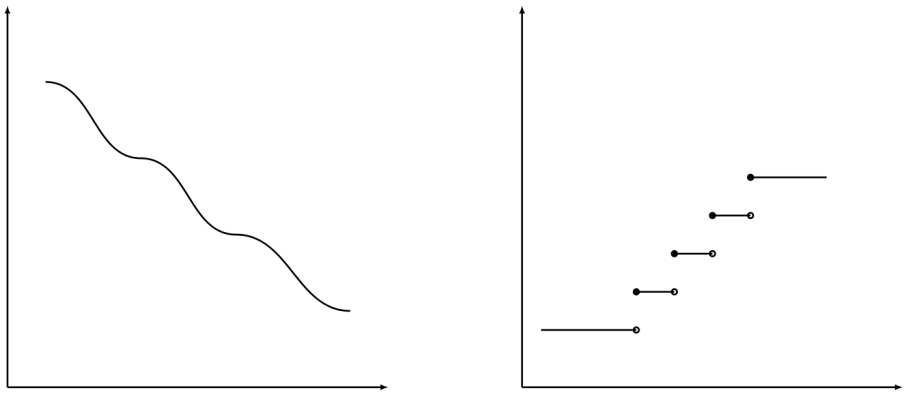


Rysunek 7.1: Przykłady funkcji quasi-wypukłych i quasi-wklęsłych.

Przykład 7.1. Funkcja afiniczna jest quasi-liniowa. Rzeczywiście, niech $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$ dla $\mathbf{a} \in \mathbb{R}^n$ i $b \in \mathbb{R}$. Wówczas dla dowolnych $\mathbf{x}, \mathbf{y} \in W$ i $\lambda \in [0, 1]$ mamy

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) = \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}).$$

Prawa strona jest oczywiście nie mniejsza od minimum z $f(\mathbf{x})$ i $f(\mathbf{y})$ i nie większa od maksimum tych liczb, czyli f jest zarówno quasi-wypukła jak i quasi-wklęsła.



Rysunek 7.2: Przykłady funkcji quasi-liniowych.

Przypomnijmy, że zbiorem poziomicowym funkcji $f : W \to \mathbb{R}$ nazywamy zbiór

$$W_\alpha(f) = \{\mathbf{x} \in W : f(\mathbf{x}) \leqslant \alpha\}, \qquad \alpha \in \mathbb{R}.$$

Twierdzenie 7.1. *Niech $f : W \to \mathbb{R}$, gdzie $W \subset \mathbb{R}^n$ wypukły. Wówczas funkcja f jest quasi-wypukła wtw, gdy zbiór poziomicowy $W_\alpha(f)$ jest wypukły dla każdego $\alpha \in \mathbb{R}$.*

Dowód. Załóżmy, że funkcja f jest quasi-wypukła i ustalmy $\alpha \in \mathbb{R}$. Niech $\mathbf{x}, \mathbf{y} \in W_\alpha(f)$. Wówczas $f(\mathbf{x}) \leqslant \alpha$ i $f(\mathbf{y}) \leqslant \alpha$. Dla dowolnego $\lambda \in (0,1)$ dostajemy

$$f\big(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\big) \leqslant \max\big(f(\mathbf{x}), f(\mathbf{y})\big) \leqslant \alpha.$$

Wnioskujemy stąd, że $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in W_\alpha(f)$, czyli $W_\alpha(f)$ jest zbiorem wypukłym.

Załóżmy teraz, że $W_\alpha(f)$ jest wypukły dla każdego $\alpha \in \mathbb{R}$. Ustalmy $\mathbf{x}, \mathbf{y} \in W$ oraz $\lambda \in (0,1)$. Na mocy założenia zbiór $W_\alpha(f)$ jest wypukły dla $\alpha = \max\big(f(\mathbf{x}), f(\mathbf{y})\big)$. Wynika stąd, że $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in W_\alpha(f)$, czyli

$$f\big(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\big) \leqslant \alpha = \max\big(f(\mathbf{x}), f(\mathbf{y})\big).$$

□

Uwaga 7.1. Analogiczne twierdzenie dla funkcji wypukłej, tw. 3.5, brzmiało: funkcja f jest wypukła wtw, gdy jej epigraf jest zbiorem wypukłym.

Wniosek 7.1. *Funkcja wypukła jest quasi-wypukła.*

Dowód. Wynika to wprost z twierdzeń 3.6 i 7.1. □

Twierdzenie przeciwne nie jest prawdziwe. Poniżej podajemy przykład funkcji quasi-wypukłej, która nie jest wypukła.

Przykład 7.2. Funkcja $f(x) = -e^x$ jest quasi-wypukła choć jest ściśle wklęsła. Dla $\alpha \geqslant 0$ zbiór $W_\alpha(f) = \mathbb{R}$, zaś dla $\alpha < 0$ mamy $W_\alpha(f) = [\log(-\alpha), \infty)$. Wszystkie te zbiory są wypukłe, więc na mocy twierdzenia 7.1 funkcja f jest quasi-wypukła. W podobny sposób możemy także pokazać, że funkcja f jest quasi-wklęsła, a zatem quasi-liniowa.

Przykład 7.3. Funkcja $f(x) = x^2$ jest quasi-wypukła, lecz nie jest quasi-wklęsła. Quasi-wypukłość wynika z wypukłości f. Quasi-wklęsłość badamy rozpatrując zbiory poziomicowe funkcji $(-f)$. Dla $\alpha < 0$ mamy $W_\alpha(-f) = (-\infty, -\sqrt{-\alpha}] \cup [\sqrt{-\alpha}, \infty)$. Nie jest to zbiór wypukły, czyli funkcja $(-f)$ nie jest quasi-wypukła.

Dowód poniższego lematu pozostawiamy jako zadanie.

Lemat 7.1. *Funkcja $f : W \to \mathbb{R}$, $W \subset \mathbb{R}^n$ wypukły, jest quasi-liniowa wtw, gdy jej obcięcie do dowolnego odcinka jest funkcją monotoniczną.*

Na mocy tego lematu możemy od razu zauważyć, że funkcja z przykładu 7.2 jest quasi-liniowa. Okazuje się, że istnieją funkcje quasi-wypukłe jednej zmiennej, które nie są ani wypukłe ani monotoniczne.

Przykład 7.4. Funkcja $f(x) = -e^{-x^2}$ jest quasi-wypukła. Mamy następujące zbiory poziomicowe w zależności od α:

$$\begin{aligned} &W_\alpha(f) = \emptyset, && \alpha \leqslant -1, \\ &W_\alpha(f) = \left[-\sqrt{-\log(-\alpha)}, \sqrt{-\log(-\alpha)}\right], && \alpha \in (-1, 0], \\ &W_\alpha(f) = \mathbb{R}, && \alpha > 0. \end{aligned}$$

Wszystkie te zbiory są wypukłe, czyli na mocy tw. 7.1 funkcja f jest quasi-wypukła.

Na koniec podamy przykład funkcji quasi-wypukłej wielu zmiennych, która nie jest wypukła.

Przykład 7.5. Funkcja $f : [0, \infty)^2 \to \mathbb{R}$ zadana wzorem $f(x_1, x_2) = -x_1 x_2$ jest quasi-wypukła. Jej zbiory poziomicowe dla $\alpha \geqslant 0$ są trywialne: $W_\alpha(f) = [0, \infty)^2$, zaś dla $\alpha < 0$ mają formę przedstawioną na rysunku 7.3. Funkcja f nie jest ani wypukła ani wklęsła, gdyż jej hesjan ma wartości własne -1 i 1.

Przykład 7.6. Niech $\mathbf{a}, \mathbf{c} \in \mathbb{R}^n$ i $b, d \in \mathbb{R}$. Połóżmy $D = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{c}^T\mathbf{x} + d > 0\}$. Wówczas funkcja wymierna $f : D \to \mathbb{R}$

$$f(\mathbf{x}) = \frac{\mathbf{a}^T\mathbf{x} + b}{\mathbf{c}^T\mathbf{x} + d}$$

jest quasi-liniowa. Dowód pozostawiamy jako ćwiczenie.

Zbadamy teraz własności różniczkowalnych funkcji quasi-wypukłych.

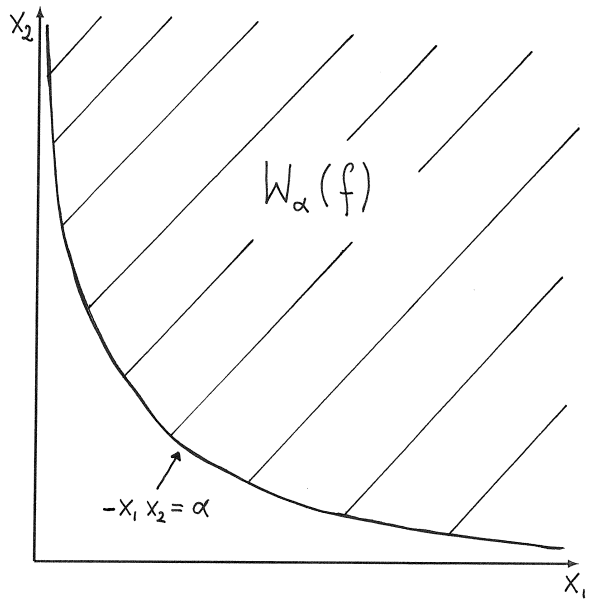
Twierdzenie 7.2. *Niech $f : W \to \mathbb{R}$ dla wypukłego zbioru $W \subset \mathbb{R}^n$.*

I) Jeśli funkcja f jest quasi-wypukła i różniczkowalna w $\mathbf{y} \in W$, to

$$\forall\, \mathbf{x} \in W \quad f(\mathbf{x}) \leqslant f(\mathbf{y}) \implies Df(\mathbf{y})(\mathbf{x} - \mathbf{y}) \leqslant 0.$$

II) Załóżmy, że funkcja f jest różniczkowalna w każdym punkcie W. Wówczas f jest quasi-wypukła wtw, gdy zachodzi następujący warunek:

$$\forall\, \mathbf{x}, \mathbf{y} \in W \quad f(\mathbf{x}) \leqslant f(\mathbf{y}) \implies Df(\mathbf{y})(\mathbf{x} - \mathbf{y}) \leqslant 0.$$



Rysunek 7.3: Zbiór poziomicowy dla funkcji z przykładu 7.5 ($\alpha < 0$).

Uwaga 7.2.

1. Implikacja $f(\mathbf{x}) \leqslant f(\mathbf{y}) \implies Df(\mathbf{x})(\mathbf{x}-\mathbf{y}) \leqslant 0$ ma równoważną postać:
$$Df(\mathbf{y})(\mathbf{x}-\mathbf{y}) > 0 \implies f(\mathbf{x}) > f(\mathbf{y}).$$
2. Jeśli funkcja f jest quasi-liniowa i $f(\mathbf{x}) = f(\mathbf{y})$, to $Df(\mathbf{y})(\mathbf{x}-\mathbf{y}) = 0$.

Dowód tw. 7.2. (I): Ustalmy $\mathbf{x}, \mathbf{y} \in W$, dla których zachodzi warunek $f(\mathbf{x}) \leqslant f(\mathbf{y})$. Dla każdego $\lambda \in (0,1)$ mamy

$$f(\mathbf{y} + \lambda(\mathbf{x}-\mathbf{y})) = f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leqslant \max\left(f(\mathbf{x}), f(\mathbf{y})\right) = f(\mathbf{y}).$$

Wynika stąd, że

$$\frac{f(\mathbf{y} + \lambda(\mathbf{x}-\mathbf{y})) - f(\mathbf{y})}{\lambda} \leqslant 0.$$

Z definicji pochodnej kierunkowej dostajemy

$$Df(\mathbf{y})(\mathbf{x}-\mathbf{y}) = \lim_{\lambda \downarrow 0} \frac{f(\mathbf{y} + \lambda(\mathbf{x}-\mathbf{y})) - f(\mathbf{y})}{\lambda} \leqslant 0.$$

Dowód (II) pozostawiamy jako niełatwe ćwiczenie. □

Wiemy już, że funkcja wypukła jest quasi-wypukła. Okazuje się, że również funkcja pseudowypukła jest quasi-wypukła.

Twierdzenie 7.3. *Jeśli $f : W \to \mathbb{R}$ określona na zbiorze wypukłym $W \subset \mathbb{R}^n$ jest pseudowypukła, to f jest quasi-wypukła.*

Dowód. Zakładając, że funkcja f nie jest quasi-wypukła doprowadzimy do sprzeczności z pseudowypukłością. Weźmy więc punkty $\mathbf{x}, \mathbf{y} \in W$ oraz $\lambda \in (0,1)$ spełniające

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) > \max\left(f(\mathbf{x}), f(\mathbf{y})\right).$$

Oznaczmy $\mathbf{z} = \lambda\mathbf{x} + (1-\lambda)\mathbf{y}$. Na mocy pseudowypukłości (wykorzystujemy tu warunek pseudowypukłości zapisany w uwadze 4.3) dostajemy:

$$\begin{aligned} f(\mathbf{x}) < f(\mathbf{z}) &\implies Df(\mathbf{z})(\mathbf{x}-\mathbf{z}) < 0, \\ f(\mathbf{y}) < f(\mathbf{z}) &\implies Df(\mathbf{z})(\mathbf{y}-\mathbf{z}) < 0. \end{aligned}$$

Wektory $(\mathbf{x}-\mathbf{z})$ i $(\mathbf{y}-\mathbf{z})$ mają ten sam kierunek lecz przeciwne zwroty. Pochodna f w punkcie $\mathbf{z}$ nie może być ujemna w obu kierunkach. Sprzeczność. □

Twierdzenie odwrotne nie jest prawdziwe. Możemy jednak podać warunek dostateczny, przy którym funkcja quasi-wypukła jest pseudowypukła.

Twierdzenie 7.4. *Niech $f : W \to \mathbb{R}$ określona na zbiorze wypukłym* otwartym *$W \subset \mathbb{R}^n$ będzie quasi-wypukła i ciągła. Jeśli f jest różniczkowalna w $\bar{\mathbf{x}} \in W$ oraz $Df(\bar{\mathbf{x}}) \neq \mathbf{0}^T$, to f jest pseudowypukła w $\bar{\mathbf{x}}$.*

Dowód. Musimy pokazać, że dla każdego $\mathbf{y} \in W$ warunek $Df(\bar{\mathbf{x}})(\mathbf{y}-\bar{\mathbf{x}}) \geqslant 0$ pociąga $f(\mathbf{y}) \geqslant f(\bar{\mathbf{x}})$. Oznaczmy przez A przestrzeń afiniczną prostopadłą do $Df(\bar{\mathbf{x}})$ i przechodzącą przez $\bar{\mathbf{x}}$:

$$A = \{\mathbf{x} \in \mathbb{R}^n : \ Df(\bar{\mathbf{x}})\ (\mathbf{x}-\bar{\mathbf{x}}) = 0\}.$$

Z warunku, że $Df(\bar{\mathbf{x}}) \neq \mathbf{0}^T$ wynika, że przestrzeń A ma wymiar $n-1$, czyli jest właściwą hiperpłaszczyzną w $\mathbb{R}^n$.

Zauważmy najpierw, że jeśli $\mathbf{y} \in W \setminus A$ i $Df(\bar{\mathbf{x}})(\mathbf{y}-\bar{\mathbf{x}}) \geqslant 0$, to pochodna kierunkowa jest ściśle dodatnia: $Df(\bar{\mathbf{x}})(\mathbf{y}-\bar{\mathbf{x}}) > 0$. Na mocy uwagi 7.2 wnioskujemy, że $f(\mathbf{y}) > f(\bar{\mathbf{x}})$, czyli to, co mieliśmy wykazać. Ustalmy teraz punkt $\mathbf{y} \in W \cap A$. Z otwartości W i z tego, że A jest hiperpłaszczyzną wynika, że istnieje ciąg punktów $(\mathbf{y}_k) \subset W \setminus A$ zmierzający do $\mathbf{y}$ i taki że $Df(\bar{\mathbf{x}})(\mathbf{y}_k - \bar{\mathbf{x}}) > 0$. Zatem $f(\mathbf{y}_k) > f(\bar{\mathbf{x}})$. Korzystając z ciągłości funkcji f dostajemy $f(\mathbf{y}) \geqslant f(\bar{\mathbf{x}})$. □

7.2 Maksymalizacja funkcji quasi-wypukłej

Jak zostało zasygnalizowane wcześniej, funkcja quasi-wypukła zachowuje się podobnie jak funkcja wypukła przy maksymalizacji na zbiorze wypukłym zwartym. Dla pełności przytoczymy dowód, który jest prawie identyczny jak dowód twierdzenia 4.5.

Twierdzenie 7.5. *Niech $f : W \to \mathbb{R}$ quasi-wypukła, ciągła, określona na wypukłym i zwartym zbiorze $W \subset \mathbb{R}^n$. Wówczas jednym z rozwiązań globalnych problemu*

$$\begin{cases} f(\mathbf{x}) \to \max, \\ \mathbf{x} \in W \end{cases}$$

jest pewien punkt ekstremalny zbioru W.

Dowód. Funkcja ciągła osiąga swoje kresy na zbiorze zwartym. Powyższy problem maksymalizacyjny ma zatem rozwiązanie $\bar{\mathbf{x}} \in W$. Na mocy tw. 4.4 punkt $\bar{\mathbf{x}}$ jest kombinacją wypukłą skończonej liczby punktów ekstremalnych, $\mathbf{x}_1, \ldots, \mathbf{x}_m$, zbioru W, tzn.

$$\bar{\mathbf{x}} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \ldots + a_m\mathbf{x}_m$$

dla liczb $a_1, \ldots, a_m > 0$ takich że $a_1 + \ldots + a_m = 1$. Z quasi-wypukłości f dostajemy

$$f(\bar{\mathbf{x}}) \leqslant \max\big(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_m)\big).$$

Ponieważ w punkcie $\bar{\mathbf{x}}$ jest maksimum f na zbiorze W, to dla któregoś z punktów x_i zachodzi równość $f(\mathbf{x}_i) = f(\bar{\mathbf{x}})$. □

7.3 Warunki dostateczne

Zajmiemy się problemem optymalizacyjnym, w którym występują zarówno ograniczenia nierównościowe jak i równościowe:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ g_i(\mathbf{x}) \leqslant 0, \quad i = 1, \ldots, m, \\ h_j(\mathbf{x}) = 0, \quad j = 1, \ldots, l, \\ \mathbf{x} \in \mathbb{X}, \end{cases} \tag{7.1}$$

gdzie $\mathbb{X} \subset \mathbb{R}^n$ jest zbiorem otwartym i $f, g_1, \ldots, g_m, h_1, \ldots, h_l : \mathbb{X} \to \mathbb{R}$. A zatem zbiór punktów dopuszczalnych zadany jest następująco:

$$W = \big\{\mathbf{x} \in \mathbb{X} : g_1(\mathbf{x}) \leqslant 0, \ldots, g_m(\mathbf{x}) \leqslant 0,\ h_1(\mathbf{x}) = 0, \ldots, h_l(\mathbf{x}) = 0\big\}. \tag{7.2}$$

Funkcje g_i nazywane są *ograniczeniami nierównościowymi*, funkcje h_j są *ograniczeniami równościowymi*, zaś cały problem (7.1) nazywa się zadaniem optymalizacyjnym z *ograniczeniami mieszanymi.*

Podamy teraz warunki dostateczne, by punkt spełniający warunki pierwszego rzędu był rozwiązaniem globalnym.

Twierdzenie 7.6. *Rozważmy problem optymalizacyjny w kanonicznej formie (7.1) i punkt $\bar{\mathbf{x}} \in W$. Załóżmy, że*

- *funkcje g_i, $i \notin I(\bar{\mathbf{x}})$ są ciągłe w $\bar{\mathbf{x}}$, funkcje g_i, $i \in I(\bar{\mathbf{x}})$ są różniczkowalne w $\bar{\mathbf{x}}$ i quasi-wypukłe,*
- *funkcje h_j, $j = 1, \ldots, l$, są quasi-liniowe i różniczkowalne w $\bar{\mathbf{x}}$,*
- *funkcja f jest pseudowypukła w $\bar{\mathbf{x}}$.*

Jeśli istnieją stałe $\mu \in [0, \infty)^m$ oraz $\lambda \in \mathbb{R}^l$ spełniające warunek pierwszego rzędu:

$$\begin{cases} Df(\bar{\mathbf{x}}) + \sum_{i \in I(\bar{\mathbf{x}})} \mu_i Dg_i(\bar{\mathbf{x}}) + \sum_{j=1}^{l} \lambda_j Dh_j(\bar{\mathbf{x}}) = \mathbf{0}^T, \\ \mu_i g_i(\bar{\mathbf{x}}) = 0, \quad i = 1, \ldots, m, \end{cases} \tag{7.3}$$

to punkt $\bar{\mathbf{x}}$ jest rozwiązaniem globalnym.

Dowód. Ustalmy dowolny punkt dopuszczalny $\mathbf{x} \in W$ i pomnóżmy obie strony pierwszego równania w warunku (7.3) przez $(\mathbf{x} - \bar{\mathbf{x}})$:

$$Df(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) + \sum_{i \in I(\bar{\mathbf{x}})} \mu_i Dg_i(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) + \sum_{j=1}^{l} \lambda_j Dh_j(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) = 0.$$

Na mocy tw. 7.2 mamy $Dh_j(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) = 0$ dla każdego j, bo $h_j(\mathbf{x}) = h_j(\bar{\mathbf{x}}) = 0$. To samo twierdzenie implikuje, że $Dg_i(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) \leqslant 0$ dla $i \in I(\bar{\mathbf{x}})$, ponieważ $0 = g_i(\bar{\mathbf{x}}) \geqslant g_i(\mathbf{x})$. Korzystając z tych obserwacji wnioskujemy z powyższego równania, że

$$Df(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) \geqslant 0.$$

Z definicji funkcji pseudowypukłej, $f(\mathbf{x}) \geqslant f(\bar{\mathbf{x}})$. Punkt $\mathbf{x}$ wybraliśmy dowolnie spośród punktów dopuszczalnych, a zatem $\bar{\mathbf{x}}$ jest rozwiązaniem globalnym. □

Uwaga 7.3. Jeśli założenia twierdzenia 7.6 są spełnione lokalnie, na pewnym otoczeniu $\bar{\mathbf{x}}$, to $\bar{\mathbf{x}}$ jest rozwiązaniem lokalnym.

Uwaga 7.4. Na mocy twierdzenia 7.4 zamiast zakładać pseudowypukłość funkcji f w punkcie $\bar{\mathbf{x}}$ możemy założyć jej ciągłość na $\mathbb{X}$, quasi-wypukłość oraz warunek $Df(\bar{\mathbf{x}}) \neq \mathbf{0}^T$. Jest to jedna z form warunku koniecznego zaprezentowana w pracy Arrowa i Enthovena z 1961 roku [1].[1]

7.4 Zadania

Ćwiczenie 7.1. Udowodnij, że funkcja $f : [a, b] \to \mathbb{R}$ jest quasi-liniowa wtw, gdy jest monotoniczna.

Ćwiczenie 7.2. Niech $f : W \to \mathbb{R}$, $W \subset \mathbb{R}^n$ wypukły. Udowodnij następujący fakt: funkcja f jest quasi-liniowa wtw, gdy jej obcięcie do każdego odcinka zawartego w W jest funkcją monotoniczną.

Ćwiczenie 7.3. Wykaż, że jeśli funkcja $f : W \to \mathbb{R}$, $W \subset \mathbb{R}^n$ wypukły, jest quasi-wypukła oraz $g : \mathbb{R} \to \mathbb{R}$ jest niemalejąca, to funkcja $g \circ f$ jest quasi-wypukła. Jeśli natomiast funkcja g jest nierosnąca, to $g \circ f$ jest quasi-wklęsła.

Ćwiczenie 7.4. Niech $f : (a, b) \to \mathbb{R}$ będzie funkcją jednej zmiennej. Wykaż, że f jest quasi-wypukła wtw, gdy zachodzi jeden z warunków:

- f jest monotoniczna,
- istnieje $\bar{x} \in (a, b)$ taki że f jest nierosnąca dla $x < \bar{x}$ oraz niemalejąca dla $x > \bar{x}$.

Ćwiczenie 7.5. Dla jakich wartości parametrów $a, b, c, d \in \mathbb{R}$ funkcja $f(x) = ax^3 + bx^2 + cx + d$ jest quasi-wypukła?

Ćwiczenie 7.6. Sprawdź, że funkcja $f : [0, \infty)^2 \to \mathbb{R}$ zadana wzorem $f(x_1, x_2) = e^{-x_1} x_2$ jest quasi-wklęsła.

Ćwiczenie 7.7. Znajdź przykład pokazujący, że suma funkcji quasi-wypukłych nie musi być quasi-wypukła.

[1] Kenneth Joseph Arrow – amerykański ekonomista. Wspólnie z Johnem Hicksem otrzymał nagrodę Nobla z ekonomii w 1972 roku.

Ćwiczenie 7.8. Niech $\mathbf{a}, \mathbf{c} \in \mathbb{R}^n$ i $b, d \in \mathbb{R}$. Wykaż, że funkcja wymierna

$$f(\mathbf{x}) = \frac{\mathbf{a}^T\mathbf{x} + b}{\mathbf{c}^T\mathbf{x} + d}$$

jest quasi-liniowa na swojej dziedzinie $D = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{c}^T\mathbf{x} + d > 0\}$.

Ćwiczenie 7.9. Niech $h : \mathbb{R}^n \to \mathbb{R}$ będzie funkcją liniową i $W = \{\mathbf{x} \in \mathbb{R}^n : \ h(\mathbf{x}) > 0\}$. Wykaż, że jeśli $f : W \to \mathbb{R}$ jest wypukła, to funkcja $g(\mathbf{x}) = f(\mathbf{x})/h(\mathbf{x})$ dla $\mathbf{x} \in W$ jest quasi-wypukła.

Ćwiczenie 7.10. Udowodnij, że jeśli $(g_\alpha)_{\alpha\in I}$ jest rodziną funkcji quasi-wypukłych, $w : I \to \mathbb{R}$ jest funkcją nieujemną, to $h(\mathbf{x}) = \sup_{\alpha\in I} w(\alpha)g_\alpha(\mathbf{x})$ jest quasi-wypukła, o ile jest skończona dla każdego $\mathbf{x}$.

Ćwiczenie 7.11. Niech $A \subset \mathbb{R}^n$, $C \subset \mathbb{R}^m$ będą zbiorami wypukłymi, zaś $f : A \times C \to \mathbb{R}$ będzie quasi-wypukła. Wykaż, że $h(\mathbf{x}) = \inf_{c\in C} f(\mathbf{x}, c)$ jest quasi-wypukła.

Ćwiczenie 7.12. Niech $W \subset \mathbb{R}^n$ zbiór wypukły, $f : W \to \mathbb{R}$. Udowodnij, że jeśli f quasi-liniowa, to zbiór $\{\mathbf{x} \in W : f(\mathbf{x}) = \alpha\}$ jest wypukły dla dowolnego $\alpha \in \mathbb{R}$.

Ćwiczenie 7.13. Niech $f : \mathbb{R}^n \to \mathbb{R}$, ciągła. Udowodnij następującą równoważność: f jest quasi-liniowa wtw, gdy $f(\mathbf{x}) = g(\mathbf{a}^T\mathbf{x})$ dla funkcji monotonicznej, ciągłej $g : \mathbb{R} \to \mathbb{R}$ oraz wektora $\mathbf{a} \in \mathbb{R}^n$.

Wykaż, że powyższa równoważność nie musi być prawdziwa, gdy funkcję f rozważamy na wypukłym podzbiorze właściwym $\mathbb{R}^n$.

Ćwiczenie 7.14. Przeprowadź dowód punktu (II) twierdzenia 7.2.

Wskazówka. Dla dowolnych punktów $\mathbf{x}, \mathbf{y} \in W$ uporządkowanych tak, że $f(\mathbf{x}) \leqslant f(\mathbf{y})$ rozważ funkcję

$$g(\lambda) = f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}), \qquad \lambda \in [0,1]$$

oraz zbiór

$$A = \{\lambda \in [0,1] : \ g(\lambda) > g(0)\}.$$

Pokaż, że jeśli zbiór ten jest niepusty, to prowadzi to do sprzeczności z ciągłością funkcji f. Oznacz przez $\hat{A}$ spójną składową A, tzn. przedział.

1. *Udowodnij, że $\forall\ \lambda \in A \quad g'(\lambda) = 0$.*
2. *Wykaż, że $\hat{A}$ ma niepuste wnętrze.*
3. *Wykaż, że funkcja f jest stała na $\hat{A}$ oraz ściśle większa od $g(0)$.*
4. *Wykaż, że istnieje przedział otwarty $I \subset [0,1]$ taki że $I \cap A = \hat{A}$.*
5. *Zauważ sprzeczność z ciągłością funkcji f, bo $f|_{\hat{A}} > f|_{I\setminus\hat{A}}$.*

Ćwiczenie 7.15. Wykaż, że funkcja quasi-wypukła (niekoniecznie ciągła) określona na zbiorze wielościennym zwartym przyjmuje swoje maksimum globalne w jednym z punktów ekstremalnych.

Ćwiczenie 7.16. Rozważmy problem optymalizacyjny (7.1). Niech $\bar{\mathbf{x}}$ będzie punktem dopuszczalnym, w którym spełnione są warunki pierwszego rzędu z mnożnikami Lagrange'a μ i λ.

1. Załóżmy, że f jest pseudowypukła w $\bar{\mathbf{x}}$, zaś funkcja

$$\phi(\mathbf{x}) = \sum_{i=1}^{m} \mu_i g_i(\mathbf{x}) + \sum_{j=1}^{l} \lambda_j h_j(\mathbf{x})$$

   jest quasi-wypukła w $\bar{\mathbf{x}}$. Udowodnij, że $\bar{\mathbf{x}}$ jest rozwiązaniem globalnym.

2. Przypomnijmy, że $L(\mathbf{x}, \mu, \lambda)$ jest funkcją Lagrange'a. Udowodnij, że jeśli $\mathbf{x} \mapsto L(\mathbf{x}, \mu, \lambda)$ jest pseudowypukła w $\bar{\mathbf{x}}$, to $\bar{\mathbf{x}}$ jest rozwiązaniem globalnym.

3. Wykaż, że warunki zawarte w powyższych punktach nie są równoważne. Znajdź zależności pomiędzy nimi a założeniami twierdzenia 7.6.

Ćwiczenie 7.17. Niech $\bar{\mathbf{x}}$ będzie rozwiązaniem globalnym problemu optymalizacyjnego (7.1). Załóżmy, że $g_k(\bar{\mathbf{x}}) < 0$ dla pewnego $k \in \{1, \ldots, m\}$. Wykaż, że jeśli ograniczenie $g_k(\mathbf{x}) \leqslant 0$ zostanie usunięte, to $\bar{\mathbf{x}}$ może nie być nawet lokalnym rozwiązaniem otrzymanego problemu. Udowodnij natomiast, że jeśli funkcja g_k jest ciągła w $\bar{\mathbf{x}}$, to po usunięciu tego ograniczenia $\bar{\mathbf{x}}$ pozostaje rozwiązaniem lokalnym.

Ćwiczenie 7.18. Dla jakich wartości parametru $\alpha \in \mathbb{R}$ problem

$$\begin{cases} x_1 + x_2 + \alpha(x_1 - 1)^2 \to \min, \\ x_1 - x_2 \leqslant 0, \\ x_1 \geqslant 0 \end{cases}$$

ma rozwiązanie? Jak zależy ono od wartości α?

Ćwiczenie 7.19. Rozwiąż zadanie optymalizacyjne:

$$\begin{cases} \mathbf{b}^T\mathbf{x} \to \min, \\ \|\mathbf{x}\|_p \leqslant 1, \end{cases}$$

gdzie $\mathbf{b} \in \mathbb{R}^n$, zaś $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$ i $p > 1$.

Ćwiczenie 7.20. Niech $u : [0, \infty)^n \to \mathbb{R}$ będzie funkcją quasi-wklęsłą o pochodnej $Du(\mathbf{x}) > \mathbf{0}$ dla każdego $\mathbf{x}$. Ustalmy liczbę $w > 0$ i wektor $\mathbf{p} \in (0, \infty)^n$. Rozważmy problem optymalizacyjny

$$\begin{cases} u(\mathbf{x}) \to \max, \\ \sum_{i=1}^{n} p_i x_i \leqslant w, \\ \mathbf{x} \geqslant \mathbf{0}. \end{cases}$$

Wektor $\mathbf{p}$ pełni rolę cen produktów, $\mathbf{x}$ ich ilości, w jest wielkością budżetu, zaś funkcja u ocenia satysfakcję z decyzji zakupowej $\mathbf{x}$.

Zapisz warunki Kuhna-Tuckera dla powyższego problemu.

1. Załóżmy, że $\bar{\mathbf{x}}$ jest rozwiązaniem. Czy będzie wówczas istniał wektor mnożników Lagrange'a, dla którego warunki Kuhna-Tuckera są spełnione w $\bar{\mathbf{x}}$?

2. Załóżmy, że warunki Kuhna-Tuckera są spełnione w $\bar{\mathbf{x}}$. Czy $\bar{\mathbf{x}}$ jest rozwiązaniem problemu optymalizacyjnego?

3. Niech $\bar{\mathbf{x}}$ będzie rozwiązaniem. Co możesz powiedzieć o związku pomiędzy p_i/p_j a $u_i'(\bar{\mathbf{x}})/u_j'(\bar{\mathbf{x}})$, gdy

- $\bar{x}_i > 0$ i $\bar{x}_j > 0$?
- $\bar{x}_i = 0$ i $\bar{x}_j > 0$?
- $\bar{x}_i = 0$ i $\bar{x}_j = 0$?

Ćwiczenie 7.21. Rozwiąż problem optymalizacyjny

$$\begin{cases} \sqrt{x} + y \to \max, \\ px + y \leqslant I, \\ x, y \geqslant 0, \end{cases}$$

gdzie $p, I > 0$ są parametrami.

Ćwiczenie 7.22. Niech A będzie macierzą symetryczną $n \times n$ dodatnio określoną, zaś $b \in (0, \infty)$. Rozwiąż problem optymalizacyjny:

$$\begin{cases} -\log\det X \to \min, \\ tr(AX) \leqslant b, \end{cases}$$

w zbiorze macierzy $X \in \mathbb{R}^{n\times n}$ symetrycznych i dodatnio określonych.

Rozdział 8

Warunek konieczny dla ograniczeń mieszanych

W tym rozdziale wyprowadzimy warunek konieczny pierwszego rzędu dla problemu optymalizacyjnego w następującej formie:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ g_i(\mathbf{x}) \leqslant 0, \quad i = 1, \ldots, m, \\ h_j(\mathbf{x}) = 0, \quad j = 1, \ldots, l, \\ \mathbf{x} \in \mathbb{X}, \end{cases} \tag{8.1}$$

gdzie $\mathbb{X} \subset \mathbb{R}^n$ jest zbiorem otwartym i $f, g_1, \ldots, g_m, h_1, \ldots, h_l : \mathbb{X} \to \mathbb{R}$. Zbiór punktów dopuszczalnych zadany jest następująco:

$$W = \{\mathbf{x} \in \mathbb{X} : g_1(\mathbf{x}) \leqslant 0, \ldots, g_m(\mathbf{x}) \leqslant 0,\ h_1(\mathbf{x}) = 0, \ldots, h_l(\mathbf{x}) = 0\}. \tag{8.2}$$

Przypomnijmy, że funkcje g_i nazywane są *ograniczeniami nierównościowymi*, funkcje h_j są *ograniczeniami równościowymi*, zaś cały problem (8.1) nazywa się zadaniem optymalizacyjnym z *ograniczeniami mieszanymi.*

Przykład 8.1. Rozważmy następujący problem optymalizacyjny:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ \mathbf{a}^T\mathbf{x} + b = 0, \\ \mathbf{x} \in \mathbb{R}^n, \end{cases}$$

dla pewnego $\mathbf{a} \in \mathbb{R}^n$ i $b \in \mathbb{R}$. Ograniczenie równościowe możemy zamienić na dwa ograniczenia nierównościowe:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ \mathbf{a}^T\mathbf{x} + b \leqslant 0, \\ -\mathbf{a}^T\mathbf{x} - b \leqslant 0, \\ \mathbf{x} \in \mathbb{R}^n. \end{cases}$$

Ograniczenia są afiniczne, czyli w każdym punkcie spełniony jest warunek afiniczności. Jeśli $\bar{\mathbf{x}}$ jest rozwiązaniem lokalnym, to istnieje wektor mnożników Lagrange'a $\mu = [\mu_1, \mu_2]^T$ i spełnione

są warunki Kuhna-Tuckera (5.5):

$$\begin{cases} Df(\bar{\mathbf{x}}) + \mu_1 \mathbf{a}^T + \mu_2(-\mathbf{a}^T) = \mathbf{0}^T, \\ \mu_1(\mathbf{a}^T\mathbf{x} + b) = 0, \\ \mu_2(-\mathbf{a}^T\mathbf{x} - b) = 0, \\ \mu_1, \mu_2 \geqslant 0. \end{cases}$$

Punkt $\bar{\mathbf{x}}$ jest dopuszczalny (jako że jest rozwiązaniem), czyli spełnia ograniczenia: $\mathbf{a}^T\mathbf{x} + b = 0$. Stąd trywialnie spełnione są druga i trzecia równość. Możemy zatem powyższe warunki równoważnie zapisać jako:

$$\begin{cases} Df(\bar{\mathbf{x}}) + (\mu_1 - \mu_2)\mathbf{a}^T = \mathbf{0}^T, \\ \mu_1, \mu_2 \geqslant 0. \end{cases}$$

Oznaczmy $\lambda = \mu_1 - \mu_2$. Warunki nieujemności μ_1, μ_2 implikują, że $\lambda \in \mathbb{R}$. Dostajemy więc finalnie:

$$Df(\bar{\mathbf{x}}) + \lambda \mathbf{a}^T = \mathbf{0}^T, \qquad \lambda \in \mathbb{R}.$$

Jest to warunek Kuhna-Tuckera dla ograniczeń równościowych.

Powyższy przykład sugerowałby, że teoria dla problemów z ograniczeniami nierównościowymi, zbudowana w poprzednich rozdziałach, pozwala poradzić sobie z ograniczeniami równościowymi. Niestety nie jest to prawda. Ograniczenia afiniczne są szczególnym przypadkiem. Jeśli któreś z ograniczeń równościowych nie jest afiniczne i rozbijemy je na dwie nierówności, jak powyżej, to w żadnym punkcie zbioru W nie jest spełniony ani warunek liniowej zależności ograniczeń ani warunek Slatera.

8.1 Warunek konieczny pierwszego rzędu

Teoria wprowadzana w tym podrozdziale jest prostym rozszerzeniem tego, co już zrobiliśmy dla problemu optymalizacyjnego z ograniczeniami nierównościowymi. Rozpoczniemy od rozszerzenia T_{lin}:

Definicja 8.1. Niech $\bar{\mathbf{x}} \in W$, g_i różniczkowalne w $\bar{\mathbf{x}}$ dla ograniczeń aktywnych $i \in I(\bar{\mathbf{x}})$ oraz h_j są różniczkowalne w $\bar{\mathbf{x}}$ dla $j = 1, \dots, l$. *Stożkiem kierunków stycznych dla ograniczeń zlinearyzowanych* nazywamy zbiór

$$T_{lin}(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^n : \ \forall i \in I(\bar{\mathbf{x}}) \quad Dg_i(\bar{\mathbf{x}})\mathbf{d} \leqslant 0, \quad \forall j = 1, \dots, l \quad Dh_j(\bar{\mathbf{x}})\mathbf{d} = 0\}.$$

Podobnie jak poprzednio zauważmy, że stożek kierunków stycznych dla ograniczeń zlinearyzowanych jest zbiorem wielościennym, a zatem wypukłym i domkniętym. Jeśli jest choć jedno ograniczenie równościowe, to ma on puste wnętrze.

Warunek konieczny istnienia rozwiązania lokalnego problemu z ograniczeniami mieszanymi jest sformułowany poniżej. Identycznie jak w twierdzeniu 5.2 zakładamy równość stożka kierunków stycznych dla ograniczeń oryginalnych i zlinearyzowanych. Później uogólnimy warunki regularności, które będą taką równość pociągały.

Twierdzenie 8.1 (Twierdzenia Kuhna-Tuckera)**.** *Niech $\bar{\mathbf{x}}$ będzie rozwiązaniem lokalnym (8.1). Jeśli funkcje f, g_i, $i \in I(\bar{\mathbf{x}})$, oraz h_j, $j = 1, \dots, l$, są różniczkowalne w $\bar{\mathbf{x}}$ oraz $T(\bar{\mathbf{x}}) = T_{lin}(\bar{\mathbf{x}})$, to istnieją $\mu \in [0, \infty)^m$ oraz $\lambda \in \mathbb{R}^l$ takie że*

$$\begin{cases} Df(\bar{\mathbf{x}}) + \sum_{i \in I(\bar{\mathbf{x}})} \mu_i Dg_i(\bar{\mathbf{x}}) + \sum_{j=1}^{l} \lambda_j Dh_j(\bar{\mathbf{x}}) = \mathbf{0}^T, \\ \mu_i g_i(\bar{\mathbf{x}}) = 0, \quad i = 1, 2, \dots, m. \end{cases} \tag{8.3}$$

Dowód. Na mocy twierdzenia 5.1 mamy $D(\bar{\mathbf{x}}) \cap T(\bar{\mathbf{x}}) = \emptyset$. Dalej, korzystając z założenia, dostajemy $D(\bar{\mathbf{x}}) \cap T_{lin}(\bar{\mathbf{x}}) = \emptyset$, co innymi słowy oznacza, że nie istnieje rozwiązanie $\mathbf{z} \in \mathbb{R}^n$ układu

$$\begin{cases} Df(\bar{\mathbf{x}})\mathbf{z} < 0, \\ Dg_i(\bar{\mathbf{x}})\mathbf{z} \leqslant 0, \quad i \in I(\bar{\mathbf{x}}). \\ Dh_j(\bar{\mathbf{x}})\mathbf{z} \leqslant 0, \quad j = 1, \ldots, l, \\ -Dh_j(\bar{\mathbf{x}})\mathbf{z} \leqslant 0, \quad j = 1, \ldots, l. \end{cases} \tag{8.4}$$

Stosujemy lemat Farkasa, lemat 5.3, z $\mathbf{d} = -Df(\bar{\mathbf{x}})$ i macierzą A następującej postaci:

$$A = \begin{bmatrix} Dh_j(\bar{\mathbf{x}}), & j = 1, \ldots, l \\ -Dh_j(\bar{\mathbf{x}}), & j = 1, \ldots, l \\ Dg_i(\bar{\mathbf{x}}), & i \in I(\bar{\mathbf{x}}) \end{bmatrix}$$

Istnieje zatem $\mathbf{y} \in [0, \infty)^{|I(\bar{\mathbf{x}})|+2l}$ takie że $\mathbf{y}^T A = -Df(\bar{\mathbf{x}})$ lub inaczej

$$Df(\bar{\mathbf{x}}) + \mathbf{y}^T A = \mathbf{0}^T. \tag{8.5}$$

Zdefiniujmy $\lambda_j = y_j - y_{l+j}$, $j = 1, \ldots, l$. Przypiszmy współrzędnym μ odpowiadającym ograniczeniom aktywnym, $i \in I(\bar{\mathbf{x}})$, ostatnie $|I(\bar{\mathbf{x}})|$ wartości wektora y. Na pozostałych współrzędnych połóżmy zera. Wówczas równość (8.5) jest równoważna następującej

$$Df(\bar{\mathbf{x}}) + \sum_{i \in I(\bar{\mathbf{x}})}^{m} \mu_i Dg_i(\bar{\mathbf{x}}) + \sum_{j=1}^{l} \lambda_j Dh_j(\bar{\mathbf{x}}) = \mathbf{0}^T.$$

Z definicji μ_i oczywiste jest, że $\mu_i g_i(\bar{\mathbf{x}}) = 0$. □

8.2 Warunki regularności

Sformułujemy teraz trzy warunki dostateczne równości $T(\bar{\mathbf{x}}) = T_{lin}(\bar{\mathbf{x}})$, zwane *warunkami regularności*.

Definicja 8.2. W punkcie $\bar{\mathbf{x}} \in W$ spełniony jest:

- *warunek liniowej niezależności*, jeśli funkcje g_i, $i \notin I(\bar{\mathbf{x}})$, są ciągłe w $\bar{\mathbf{x}}$, pozostałe ograniczenia nierównościowe i wszystkie równościowe są klasy C^1 na otoczeniu $\bar{\mathbf{x}}$ oraz wektory $Dg_i(\bar{\mathbf{x}})$ dla $i \in I(\bar{\mathbf{x}})$ i $Dh_j(\bar{\mathbf{x}})$ dla $j = 1, \ldots, l$ są liniowo niezależne,
- *warunek afiniczności*, jeśli funkcje g_i, $i \in I(\bar{\mathbf{x}})$, oraz h_j, $j = 1, \ldots, l$, są afiniczne, a g_i, $i \notin I(\bar{\mathbf{x}})$, są ciągłe w $\bar{\mathbf{x}}$,
- *warunek Slatera*, jeśli
  - funkcje g_i, $i \in I(\bar{\mathbf{x}})$ są pseudowypukłe w $\bar{\mathbf{x}}$, funkcje g_i, $i \notin I(\bar{\mathbf{x}})$, są ciągłe w $\bar{\mathbf{x}}$,
  - funkcje h_j, $j = 1, \ldots, l$, są afiniczne,
  - istnieje $\mathbf{x} \in \mathbb{X}$, dla którego $g_i(\mathbf{x}) < 0$ dla $i \in I(\bar{\mathbf{x}})$ oraz $h_j(\mathbf{x}) = 0$ dla $j = 1, \ldots, l$.

Zaczniemy od najprostszego przypadku.

Twierdzenie 8.2. *Jeśli w punkcie $\bar{\mathbf{x}} \in W$ spełniony jest warunek afiniczności, to zachodzi równość $T(\bar{\mathbf{x}}) = T_{lin}(\bar{\mathbf{x}})$.*

Dowód. Postępując jak w przykładzie 8.1 zamieniamy ograniczenia afiniczne równościowe na ograniczenia afiniczne nierównościowe. Teza wynika z twierdzenia 6.1. □

Twierdzenie 8.3. *Jeśli w punkcie $\bar{\mathbf{x}} \in W$ spełniony jest warunek Slatera, to zachodzi równość $T(\bar{\mathbf{x}}) = T_{lin}(\bar{\mathbf{x}})$.*

Dowód. Zapiszmy najpierw funkcje h_j dla $j = 1, \dots, l$:

$$h_j(\mathbf{y}) = \mathbf{a}_j^T \mathbf{y} + b_j, \qquad \mathbf{a}_j \in \mathbb{R}^n, \quad b_j \in \mathbb{R}.$$

Wprowadźmy uogólnienie zbioru $T_{int}(\bar{\mathbf{x}})$ do przypadku ograniczeń mieszanych:

$$T_{int}(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^n : \ \forall i \in I(\bar{\mathbf{x}}) \quad Dg_i(\bar{\mathbf{x}})\mathbf{d} < 0, \quad \forall j = 1, \dots, l \quad Dh_j(\bar{\mathbf{x}})\mathbf{d} = 0\}.$$

(1) $T_{int}(\bar{\mathbf{x}}) \neq \emptyset$. Weźmy punkt $\mathbf{x}$ z warunku Slatera. Na mocy pseudowypukłości, patrz uwaga 4.3, mamy

$$Dg_i(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) < 0, \qquad \forall\, i \in I(\bar{\mathbf{x}}).$$

Dla każdego j mamy także

$$\mathbf{a}_j^T(\mathbf{x} - \bar{\mathbf{x}}) = \mathbf{a}_j^T \mathbf{x} + b_j - \mathbf{a}_j^T \bar{\mathbf{x}} - b_j = h_j(\mathbf{x}) - h_j(\bar{\mathbf{x}}) = 0.$$

Wnioskujemy więc, że wektor $(\mathbf{x} - \bar{\mathbf{x}}) \in T_{int}(\bar{\mathbf{x}})$.

(2) $T_{int}(\bar{\mathbf{x}}) \subset T(\bar{\mathbf{x}})$. W tym celu weźmy dowolny $\mathbf{d} \in T_{int}(\bar{\mathbf{x}})$. Wystarczy pokazać, że pewien odcinek o końcu $\bar{\mathbf{x}}$ i kierunku $\mathbf{d}$ zawiera się w całości w zbiorze W. Rozważmy w tym celu funkcję $\mathbf{y}(\lambda) = \mathbf{x} + \lambda\mathbf{d}$. Na mocy ciągłości funkcji opisujących ograniczenia nieaktywne istnieje $\varepsilon > 0$ taki że $g_i(\mathbf{y}(\lambda)) \leqslant 0$ dla $\lambda \in [0, \varepsilon]$ oraz $i \notin I(\bar{\mathbf{x}})$. Z faktu, że $\mathbf{d} \in T_{int}(\bar{\mathbf{x}})$ dostajemy również, że $h_j(\mathbf{y}(\lambda)) = 0$ dla $j = 1, \dots, l$ i dowolnego λ (bo h_j są afiniczne). Pozostaje tylko zająć się ograniczeniami aktywnymi. Z faktu, że g_i są różniczkowalne w $\bar{\mathbf{x}}$ dla $i \in I(\bar{\mathbf{x}})$ mamy

$$\lim_{\lambda \downarrow 0} \frac{g_i(\mathbf{y}(\lambda)) - g_i(\bar{\mathbf{x}})}{\lambda} = Dg_i(\bar{\mathbf{x}})\mathbf{d} < 0,$$

gdzie ostatnia nierówność wynika z tego, że $\mathbf{d} \in T_{int}(\bar{\mathbf{x}})$. A zatem $g_i(\mathbf{y}(\lambda)) - g_i(\bar{\mathbf{x}}) < 0$ dla dostatecznie małych λ.

(3) $\operatorname{cl} T_{int}(\bar{\mathbf{x}}) = T_{lin}(\bar{\mathbf{x}})$. Zbiory $T_{int}(\bar{\mathbf{x}})$ i $T_{lin}(\bar{\mathbf{x}})$ leżą w podprzestrzeni H wyznaczonej przez afiniczne ograniczenia równościowe. Możemy zatem znaleźć przekształcenie afiniczne P o pełnym rzędzie przekształcające tę podprzestrzeń w przestrzeń $\mathbb{R}^{n'}$, gdzie n' jest wymiarem H (jeśli funkcje h_j są parami różne, to $n' = n - l$). Przekształcenie to jest wzajemnie jednoznaczne rozpatrywane jako funkcja określona na H. A zatem topologie w $\mathbb{R}^{n'}$ i na H są identyczne. Wystarczy więc udowodnić tezę tego podpunktu na obrazach $T'_{int}(\bar{\mathbf{x}})$ i $T'_{lin}(\bar{\mathbf{x}})$ zbiorów $T_{int}(\bar{\mathbf{x}})$ i $T_{lin}(\bar{\mathbf{x}})$. Zauważmy, że zbiór $T'_{int}(\bar{\mathbf{x}})$ jest otwarty. Wykazaliśmy, że jest niepusty. Jest również wnętrzem zbioru $T'_{lin}(\bar{\mathbf{x}})$. Na mocy lematu 6.2 mamy $\operatorname{cl} T'_{int}(\bar{\mathbf{x}}) = T'_{lin}(\bar{\mathbf{x}})$.

(4) $T(\bar{\mathbf{x}}) \subset T_{lin}(\bar{\mathbf{x}})$. Identycznie jak dowód lematu 5.2.

Pozostaje już tylko przypomnieć, że $T(\bar{\mathbf{x}})$ jest zbiorem domkniętym. A zatem

$$\operatorname{cl} T_{int}(\bar{\mathbf{x}}) \subset T(\bar{\mathbf{x}}) \subset T_{lin}(\bar{\mathbf{x}}) = \operatorname{cl} T_{int}(\bar{\mathbf{x}}).$$

□

Zanim przejdziemy do rozważań nad trzecim warunkiem regularności, warunkiem liniowej niezależności, przypomnijmy twierdzenie o funkcji uwikłanej, by, korzystając z niego, podać opis stożka kierunków stycznych do powierzchni zadanej przez ograniczenia równościowe.

Twierdzenie 8.4 (Twierdzenie o funkcji uwikłanej)**.** *Niech $f : \mathbb{X} \to \mathbb{R}^n$, gdzie $\mathbb{X} \subset \mathbb{R}^{n+m}$ otwarty, będzie odwzorowaniem klasy C^k, $k \geqslant 1$. Załóżmy, że $f(\mathbf{a},\mathbf{b}) = \mathbf{0}$, gdzie $(\mathbf{a},\mathbf{b}) \in \mathbb{X}$. Przyjmujemy tutaj notację, że $\mathbf{a} \in \mathbb{R}^n$, zaś $\mathbf{b} \in \mathbb{R}^m$. Oznaczmy przez $A_{\mathbf{x}}$ macierz pochodnych cząstkowych, w punkcie $(\mathbf{a},\mathbf{b})$, względem pierwszych n zmiennych: $A_{\mathbf{x}} \in \mathbb{R}^{n\times n}$ zadana jest wzorem $(A_{\mathbf{x}})_{ij} = \frac{\partial f_i}{\partial u_j}(\mathbf{a},\mathbf{b})$.*

Jeśli macierz $A_{\mathbf{x}}$ jest odwracalna, to istnieje zbiór otwarty $W \subset \mathbb{R}^m$ zawierający $\mathbf{b}$ oraz funkcja $g : W \to \mathbb{R}^n$ klasy C^k, taka że $(g(\mathbf{y}),\mathbf{y}) \in \mathbb{X}$ dla $\mathbf{y} \in W$, $g(\mathbf{b}) = \mathbf{a}$ oraz $f(g(\mathbf{y}),\mathbf{y}) = \mathbf{0}$ dla $\mathbf{y} \in W$. Ponadto, $Dg(\mathbf{b}) = -(A_{\mathbf{x}})^{-1}A_{\mathbf{y}}$, gdzie $A_{\mathbf{y}}$ jest pochodną f w punkcie $(\mathbf{a},\mathbf{b})$ względem ostatnich m zmiennych: $A_{\mathbf{y}} \in \mathbb{R}^{n\times m}$ zadana jest wzorem $(A_{\mathbf{y}})_{ij} = \frac{\partial f_i}{\partial u_{n+j}}(\mathbf{a},\mathbf{b})$.

Rozważmy powierzchnię opisaną przez układ m^* równań:

$$S = \{\mathbf{x} \in \mathbb{X} :\ c_i(\mathbf{x}) = 0, \quad i = 1,\dots,m^*\},$$

gdzie $\mathbb{X} \subset \mathbb{R}^n$ otwarty. Przez $T^S(\bar{\mathbf{x}})$ oznaczmy stożek kierunków stycznych do S punkcie $\bar{\mathbf{x}} \in S$.

Twierdzenie 8.5. *Załóżmy, że funkcje c_i, $i = 1,\dots,m^*$, są klasy C^k, $k \geqslant 1$, na otoczeniu $\bar{\mathbf{x}}$ oraz gradienty $Dc_i(\bar{\mathbf{x}})$, $i = 1,\dots,m^*$, są liniowo niezależne. Wówczas*

$$T^S(\bar{\mathbf{x}}) = T^S_{lin}(\bar{\mathbf{x}}) := \{\mathbf{d} \in \mathbb{R}^n :\ Dc_i(\bar{\mathbf{x}})\mathbf{d} = 0, \quad i = 1,\dots,m^*\}.$$

Ponadto, dla każdego $\mathbf{d} \in T^S(\bar{\mathbf{x}})$ istnieje $\varepsilon > 0$ i krzywa $\mathbf{y} : (-\varepsilon,\varepsilon) \to S$ klasy C^k o tej własności, że $\mathbf{y}(0) = \bar{\mathbf{x}}$ oraz $\mathbf{y}'(0) = \mathbf{d}$.

Dowód. Pokażemy najpierw, że $T^S(\bar{\mathbf{x}}) \subset T^S_{lin}(\bar{\mathbf{x}})$. Niech $\mathbf{d} \in T^S(\bar{\mathbf{x}})$. Wówczas $\mathbf{d} = \lim_{k\to\infty} \lambda_k(\mathbf{x}_k - \bar{\mathbf{x}})$ dla $(\mathbf{x}_k) \subset S$, $\mathbf{x}_k \neq \mathbf{x}$. Z definicji pochodnej dostajemy dla każdego $i = 1,\dots,m^*$:

$$\underbrace{c_i(\mathbf{x}_k)}_{=0} = \underbrace{c_i(\bar{\mathbf{x}})}_{=0} + Dc_i(\bar{\mathbf{x}})\underbrace{\lambda_k(\mathbf{x}_k - \bar{\mathbf{x}})}_{\to\mathbf{d}} + \underbrace{\lambda_k\|\mathbf{x}_k - \bar{\mathbf{x}}\|}_{\to\|\mathbf{d}\|}\underbrace{\frac{o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|)}{\|\mathbf{x}_k - \bar{\mathbf{x}}\|}}_{\to 0},$$

czyli $Dc_i(\bar{\mathbf{x}})\mathbf{d} = 0$. Stąd wynika, że $\mathbf{d} \in T^S_{lin}(\bar{\mathbf{x}})$.

Pozostało jeszcze zawieranie w drugą stronę. Dowód tej części będzie zdecydowanie trudniejszy. Ustalmy $\mathbf{d} \in T^S_{lin}(\bar{\mathbf{x}})$. Skonstruujemy krzywą przechodzącą przez $\bar{\mathbf{x}}$ i zawartą w S, której pochodna w punkcie $\bar{\mathbf{x}}$ jest równa $\mathbf{d}$. Oznaczmy $\mathbf{c}(\mathbf{x}) = (c_1(\mathbf{x}),\dots,c_{m^*}(\mathbf{x}))$ i zdefiniujmy funkcję $\Phi : \mathbb{R}^{m^*} \times \mathbb{R} \to \mathbb{R}^{m^*}$ wzorem

$$\Phi(\mathbf{u},t) = \mathbf{c}(\bar{\mathbf{x}} + t\mathbf{d} + (D\mathbf{c}(\bar{\mathbf{x}}))^T\mathbf{u}).$$

Zauważmy, że $\Phi(\mathbf{0},0) = \mathbf{0}$. Oznaczmy przez $D_{\mathbf{u}}\Phi$ macierz pochodnych cząstkowych względem zmiennych wektora $\mathbf{u}$: $D_{\mathbf{u}}\Phi = (\frac{\partial\Phi_i}{\partial u_j})_{i,j=1}^{m^*}$. W $(\mathbf{0},0)$ mamy $D_{\mathbf{u}}\Phi(\mathbf{0},0) = D\mathbf{c}(\bar{\mathbf{x}})\big(D\mathbf{c}(\bar{\mathbf{x}})\big)^T$. Przypomnijmy, że zgodnie z założeniem macierz $D\mathbf{c}(\bar{\mathbf{x}})$ ma maksymalny rząd (równy m^*), czyli $D_{\mathbf{u}}\Phi(\mathbf{0},0)$ jest odwracalna. Na mocy twierdzenia o funkcji uwikłanej istnieje zatem $\varepsilon > 0$ oraz funkcja $\mathbf{u} : (-\varepsilon,\varepsilon) \to \mathbb{R}^{m^*}$ klasy C^k, taka że $\Phi(\mathbf{u}(t),t) = \mathbf{0}$. Połóżmy

$$\mathbf{y}(t) = \bar{\mathbf{x}} + t\mathbf{d} + (D\mathbf{c}(\bar{\mathbf{x}}))^T\mathbf{u}(t).$$

Krzywa ta, zgodnie z konstrukcją, leży na powierzchni S, bo $\mathbf{c}(\mathbf{y}(t)) = \Phi(\mathbf{u}(t),t) = \mathbf{0}$ dla $t \in (-\varepsilon,\varepsilon)$ oraz $\mathbf{y}(0) = \bar{\mathbf{x}}$. Różniczkując złożenie $\mathbf{c}(\mathbf{y}(t))$ dostajemy

$$\frac{d}{dt}\mathbf{c}(\mathbf{y}(t)) = D\mathbf{c}(\mathbf{y}(t))\big(\mathbf{d} + (D\mathbf{c}(\bar{\mathbf{x}}))^T\mathbf{u}'(t)\big).$$

czyli w $t = 0$ mamy

$$\frac{d}{dt}\mathbf{c}(\mathbf{y}(t))\big|_{t=0} = D\mathbf{c}(\bar{\mathbf{x}})(\mathbf{d} + (D\mathbf{c}(\bar{\mathbf{x}}))^T\mathbf{u}'(0)).$$

Z drugiej strony wiemy, że $\mathbf{c}(\mathbf{y}(t)) = \mathbf{0}$, czyli powyższa pochodna jest równa zero: $D\mathbf{c}(\bar{\mathbf{x}})(\mathbf{d} + (D\mathbf{c}(\bar{\mathbf{x}}))^T\mathbf{u}'(0)) = \mathbf{0}$. Przypomnijmy, że $\mathbf{d} \in T^S_{lin}(\bar{\mathbf{x}})$, co w naszym zapisie oznacza $D\mathbf{c}(\bar{\mathbf{x}})\mathbf{d} = \mathbf{0}$. Wynika stąd, że $D\mathbf{c}(\bar{\mathbf{x}})(D\mathbf{c}(\bar{\mathbf{x}}))^T\mathbf{u}'(0) = \mathbf{0}$. Korzystając z faktu, że $D\mathbf{c}(\bar{\mathbf{x}})$ ma rząd m^* dostajemy $\mathbf{u}'(0) = \mathbf{0}$. Jesteśmy już teraz gotowi, aby dokończyć dowód. Różniczkując funkcję $\mathbf{y}$ dostajemy

$$\mathbf{y}'(t) = \mathbf{d} + (D\mathbf{c}(\bar{\mathbf{x}}))^T\mathbf{u}'(t),$$

co w $t = 0$ daje $\mathbf{y}'(0) = \mathbf{d}$. Możemy stąd już łatwo wywnioskować, że $\mathbf{d} \in T^S(\bar{\mathbf{x}})$. □

Uwaga 8.1. Powyższe twierdzenie dowodzi powszechnie znanego faktu dotyczącego przestrzeni stycznej do rozmaitości. Otóż, z założeń wynika, że S jest lokalnie wokół punktu $\bar{\mathbf{x}}$ rozmaitością różniczkową klasy C^k. Przestrzeń styczna do rozmaitości w punkcie $\bar{\mathbf{x}}$ definiowana jest jako zbiór wektorów, które są pochodnymi (w punkcie $\bar{\mathbf{x}}$) krzywych leżących na tej rozmaitości i przechodzących przez $\bar{\mathbf{x}}$ (jest to równoważne definicji $T(\bar{\mathbf{x}})$). Równość $T_{lin}(\bar{\mathbf{x}}) = T(\bar{\mathbf{x}})$ oznacza, że przestrzeń styczna jest jądrem przekształcenia liniowego $D\mathbf{c}(\bar{\mathbf{x}})$.

Z powyższego twierdzenia będziemy wielokrotnie korzystać w następnych rozdziałach. Będzie ono głównym narzędziem przy dowodzeniu warunku koniecznego drugiego rzędu. W tym rozdziale pozwoli łatwo wykazać równość $T(\bar{\mathbf{x}}) = T_{lin}(\bar{\mathbf{x}})$ przy założeniu warunku liniowej niezależności.

Twierdzenie 8.6. *Jeśli w punkcie $\bar{\mathbf{x}} \in W$ spełniony jest warunek liniowej niezależności, to zachodzi równość $T(\bar{\mathbf{x}}) = T_{lin}(\bar{\mathbf{x}})$.*

Dowód. Ustalmy $\mathbf{d} \in T_{lin}(\bar{\mathbf{x}})$. Niech $\hat{I}(\bar{\mathbf{x}}) = \{i \in I(\bar{\mathbf{x}}) : \ Dg_i(\bar{\mathbf{x}})\mathbf{d} = 0\}$. Zdefiniujmy powierzchnię

$$S = \{\mathbf{x} \in \mathbb{X} : c_k(\mathbf{x}) = 0, \qquad \text{gdzie } c_k = g_i, i \in \hat{I}(\bar{\mathbf{x}}) \text{ lub } c_k = h_j, j = 1, \dots, l\}.$$

Wtedy

$$T^S(\bar{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^n : Dg_i(\bar{\mathbf{x}})\mathbf{d} = 0, i \in \hat{I}(\bar{\mathbf{x}}), Dh_j(\bar{\mathbf{x}})\mathbf{d} = 0, j = 1, \dots, l\}.$$

Na mocy twierdzenia 8.5 istnieje krzywa $\mathbf{y} : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$, taka że $\mathbf{y}(0) = \bar{\mathbf{x}}$, $\mathbf{y}'(0) = \mathbf{d}$ oraz $g_i(\mathbf{y}(t)) = 0$, $i \in \hat{I}(\bar{\mathbf{x}})$, i $h_j(\mathbf{y}(t)) = 0$, $j = 1, \dots, l$. Dla $i \in I(\bar{\mathbf{x}}) \setminus \hat{I}(\bar{\mathbf{x}})$ połóżmy $\hat{g}_i(t) = g_i(\mathbf{y}(t))$, $t \in (-\varepsilon, \varepsilon)$. Wówczas $\hat{g}_i'(0) = Dg_i(\bar{\mathbf{x}})\mathbf{d} < 0$, czyli istnieje $\varepsilon_i > 0$, takie że $\hat{g}_i(t) < 0$ dla $t \in [0, \varepsilon_i)$. Dla $i \notin I(\bar{\mathbf{x}})$ z ciągłości g_i wynika, że $g_i(\mathbf{y}(t)) < 0$ na pewnym otoczeniu 0. Podsumowując, istnieje $\bar{\varepsilon} > 0$, takie że $\mathbf{y}(t) \in W$ dla $t \in [0, \bar{\varepsilon})$. Stąd trywialnie $\mathbf{d} \in T(\bar{\mathbf{x}})$.

Dowód zawierania $T(\bar{\mathbf{x}}) \subset T_{lin}(\bar{\mathbf{x}})$ jest identyczny do dowodu lematu 5.2. □

Rozdział 9

Warunki drugiego rzędu

W tym rozdziale przedstawimy warunki konieczne i dostateczne drugiego rzędu, tzn. warunki sformułowane w języku drugich pochodnych oraz podsumujemy dotychczasową wiedzę dotyczącą rozwiązywania problemów optymalizacyjnych z ograniczeniami mieszanymi.

9.1 Warunki drugiego rzędu

Rozważmy problem optymalizacyjny z ograniczeniami mieszanymi (8.1). Załóżmy, że pewnym punkcie $\bar{\mathbf{x}} \in W$ spełniony jest warunek pierwszego rzędu, tzn. istnieją wektory $\mu \in [0,\infty)^m$ i $\lambda \in \mathbb{R}^l$, takie że

$$\begin{cases} Df(\bar{\mathbf{x}}) + \sum_{i\in I(\bar{\mathbf{x}})} \mu_i Dg_i(\bar{\mathbf{x}}) + \sum_{j=1}^{l} \lambda_j Dh_j(\bar{\mathbf{x}}) = \mathbf{0}^T, \\ \mu_i g_i(\bar{\mathbf{x}}) = 0, \quad i = 1,2,\ldots,m. \end{cases}$$

Definicja 9.1. *Funkcją Lagrange'a* nazywamy funkcję

$$L(\mathbf{x},\mu,\lambda) = f(\mathbf{x}) + \sum_{i=1}^{m} \mu_i g_i(\mathbf{x}) + \sum_{j=1}^{l} \lambda_j h_j(\mathbf{x}).$$

Możemy teraz zapisać warunek pierwszego rzędu w skróconej formie:

$$D_{\mathbf{x}} L(\bar{\mathbf{x}},\mu,\lambda) = \mathbf{0}^T, \qquad \text{oraz} \qquad \mu_i g_i(\bar{\mathbf{x}}) = 0, \quad i = 1,2,\ldots,m,$$

gdzie $D_{\mathbf{x}}$ oznacza różniczkowanie względem zmiennej $\mathbf{x} \in \mathbb{R}^n$.

Definicja 9.2. Zbiór

$$I^*(\bar{\mathbf{x}}) = \{i \in I(\bar{\mathbf{x}}) : \ \mu_i > 0\}$$

nazywamy zbiorem *ograniczeń nierównościowych mocno aktywnych.*

Zbiór

$$I^0(\bar{\mathbf{x}}) = I(\bar{\mathbf{x}}) \setminus I^*(\bar{\mathbf{x}})$$

nazywamy zbiorem *ograniczeń nierównościowych słabo aktywnych.*

Twierdzenie 9.1 (Warunek konieczny drugiego rzędu)**.** *Załóżmy, że punkt $\bar{\mathbf{x}}$ jest lokalnym rozwiązaniem problemu z ograniczeniami mieszanymi i spełniony jest w nim warunek liniowej niezależności. Niech μ, λ będą mnożnikami Lagrange'a z warunku pierwszego rzędu. Jeśli f, g_i, $i \in I(\bar{\mathbf{x}})$ oraz h_j, $j = 1,\ldots,l$, są klasy C^2 w otoczeniu $\bar{\mathbf{x}}$, to*

$$\mathbf{d}^T D^2_{\mathbf{x}} L(\bar{\mathbf{x}},\mu,\lambda)\mathbf{d} \geqslant 0$$

dla każdego $\mathbf{d} \in \mathbb{R}^n$ *spełniającego*

$$\begin{aligned} Dg_i(\bar{\mathbf{x}})\mathbf{d} &= 0, \quad i \in I(\bar{\mathbf{x}}), \\ Dh_j(\bar{\mathbf{x}})\mathbf{d} &= 0, \quad j = 1, \dots, l. \end{aligned}$$

Dowód. Ustalmy $\mathbf{d} \in \mathbb{R}^n$ jak w warunkach twierdzenia. Na mocy twierdzenia 8.5 istnieje $\varepsilon > 0$ i krzywa $\mathbf{y} : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ klasy C^2 o następujących własnościach: $\mathbf{y}(0) = \bar{\mathbf{x}}$, $\mathbf{y}'(0) = \mathbf{d}$, oraz dla $t \in (-\varepsilon, \varepsilon)$ mamy $h_j(\mathbf{y}(t)) = 0$, $j = 1, \dots, l$, i $g_i(\mathbf{y}(t)) = 0$, $i \in I(\bar{\mathbf{x}})$. Wynika stąd, że funkcja $F(t) := L(\mathbf{y}(t), \mu, \lambda)$ równa jest $f(\mathbf{y}(t))$ dla t w otoczeniu 0. Z ciągłości ograniczeń nieaktywnych wnioskujemy, że $\mathbf{y}(t) \in W$ dla t w otoczeniu 0. Zatem F ma minimum lokalne w 0, gdyż $\mathbf{y}(0)$ jest minimum lokalnym f na W. Na mocy założeń, F jest klasy C^2. Istnienie minimum w zerze implikuje, że $F''(0) \geqslant 0$, a więc

$$0 \leqslant \mathbf{d}^T D^2_{\mathbf{x}} L(\bar{\mathbf{x}}, \mu, \lambda)\mathbf{d} + D_{\mathbf{x}} L(\bar{\mathbf{x}}, \mu, \lambda)\mathbf{y}''(0).$$

To kończy dowód, gdyż w punkcie $\bar{\mathbf{x}}$ spełnione są warunki konieczne pierwszego rzędu

$$D_{\mathbf{x}} L(\bar{\mathbf{x}}, \mu, \lambda) = \mathbf{0}^T.$$

□

Bez dowodu pozostawiamy następujące uogólnienie powyższego twierdzenia:

Twierdzenie 9.2. *Przy założeniach tw. 9.1 następująca nierówność*

$$\mathbf{d}^T D^2_{\mathbf{x}} L(\bar{\mathbf{x}}, \mu, \lambda)\mathbf{d} \geqslant 0$$

zachodzi dla każdego $\mathbf{d} \in \mathbb{R}^n$ *spełniającego*

$$\begin{aligned} Dg_i(\bar{\mathbf{x}})\mathbf{d} &= 0, \quad i \in I^*(\bar{\mathbf{x}}), \\ Dg_i(\bar{\mathbf{x}})\mathbf{d} &\leqslant 0, \quad i \in I^0(\bar{\mathbf{x}}), \\ Dh_j(\bar{\mathbf{x}})\mathbf{d} &= 0, \quad j = 1, \dots, l. \end{aligned}$$

Podamy teraz warunek dostateczny istnienia rozwiązania lokalnego. Zwróćmy uwagę na to, że warunek ten implikuje, iż rozwiązanie jest ścisłe. Pozostaje więc szara strefa, gdzie jest spełniony warunek konieczny drugiego rzędu, lecz nie zachodzi warunek dostateczny drugiego rzędu. Podobnie jak w przypadku optymalizacji bez ograniczeń, na to nie ma niestety rady.

Twierdzenie 9.3 (Warunek dostateczny drugiego rzędu)**.** *Załóżmy, że w punkcie* $\bar{\mathbf{x}} \in W$ *spełniony jest warunek pierwszego rzędu, tzn. istnieją wektory* $\mu \in [0, \infty)^m$ *i* $\lambda \in \mathbb{R}^l$*, takie że*

$$\begin{cases} Df(\bar{\mathbf{x}}) + \sum_{i \in I(\bar{\mathbf{x}})} \mu_i Dg_i(\bar{\mathbf{x}}) + \sum_{j=1}^{l} \lambda_j Dh_j(\bar{\mathbf{x}}) = \mathbf{0}^T, \\ \mu_i g_i(\bar{\mathbf{x}}) = 0, \quad i = 1, 2, \dots, m. \end{cases} \tag{9.1}$$

Załóżmy ponadto, że funkcje g_i, $i \in I^*(\bar{\mathbf{x}})$*, oraz* h_j, $j = 1, \dots, l$*, są dwukrotnie różniczkowalne w* $\bar{\mathbf{x}}$*. Jeśli*

$$\mathbf{d}^T D^2_{\mathbf{x}} L(\bar{\mathbf{x}}, \mu, \lambda)\mathbf{d} > 0$$

dla każdego $\mathbf{d} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ *spełniającego*

$$\begin{aligned} Dg_i(\bar{\mathbf{x}})\mathbf{d} &= 0, \quad i \in I^*(\bar{\mathbf{x}}), \\ Dg_i(\bar{\mathbf{x}})\mathbf{d} &\leqslant 0, \quad i \in I^0(\bar{\mathbf{x}}), \\ Dh_j(\bar{\mathbf{x}})\mathbf{d} &= 0, \quad j = 1, \dots, l, \end{aligned} \tag{9.2}$$

to $\bar{\mathbf{x}}$ *jest ścisłym rozwiązaniem lokalnym.*

Zanim przejdziemy do dowodu podkreślmy, iż w powyższym twierdzeniu nie zakłada się regularności punktu ograniczeń.

Dowód twierdzenia 9.3. Przeprowadzimy dowód przez zaprzeczenie. Załóżmy mianowicie, że $\bar{\mathbf{x}}$ nie jest ścisłym rozwiązaniem lokalnym. Istnieje zatem zbieżny do $\bar{\mathbf{x}}$ ciąg punktów dopuszczalnych $\mathbf{x}_k \in W$, takich że $\mathbf{x}_k \neq \bar{\mathbf{x}}$ oraz $f(\mathbf{x}_k) \leqslant f(\bar{\mathbf{x}})$. Zdefiniujmy

$$\mathbf{d}_k = \frac{\mathbf{x}_k - \bar{\mathbf{x}}}{\|\mathbf{x}_k - \bar{\mathbf{x}}\|}, \qquad \text{oraz} \qquad \xi_k = \|\mathbf{x}_k - \bar{\mathbf{x}}\|.$$

Wówczas $\mathbf{x}_k = \bar{\mathbf{x}} + \xi_k \mathbf{d}_k$. Z faktu $\|\mathbf{d}_k\| = 1$ wynika, że istnieje podciąg $\mathbf{d}_{k_n}$ zbieżny do pewnego $\mathbf{d}$ o normie jednostkowej. Aby uprościć notację zakładamy, że od początku $\mathbf{d}_k$ był zbieżny do $\mathbf{d}$. Z definicji ξ_k wynika, że $\lim_{k\to\infty} \xi_k = 0$.

W dalszej części dowodu wykażemy dwie własności $\mathbf{d}$: (a) $\mathbf{d}^T D^2_{\mathbf{x}} L(\bar{\mathbf{x}}, \mu, \lambda)\mathbf{d} \leqslant 0$ oraz (b) $\mathbf{d}$ spełnia warunki (9.2). A to przeczy założeniom twierdzenia.

Rozpocznijmy od dowodu własności (a). Z definicji drugiej pochodnej

$$L(\mathbf{x}, \mu, \lambda) = L(\bar{\mathbf{x}}, \mu, \lambda) + D_{\mathbf{x}} L(\bar{\mathbf{x}}, \mu, \lambda)(\mathbf{x} - \bar{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^T D^2_{\mathbf{x}} L(\bar{\mathbf{x}}, \mu, \lambda)(\mathbf{x} - \bar{\mathbf{x}}) + o(\|\mathbf{x} - \bar{\mathbf{x}}\|^2).$$

Przypomnijmy, że $L(\mathbf{x}_k, \mu, \lambda) \leqslant L(\bar{\mathbf{x}}, \mu, \lambda)$, bo $f(\mathbf{x}_k) \leqslant f(\bar{\mathbf{x}})$ oraz $g_i(\mathbf{x}_k) \leqslant g_i(\bar{\mathbf{x}})$ dla $i \in I(\bar{\mathbf{x}})$ a $h(\mathbf{x}_k) = h(\bar{\mathbf{x}}) = 0$. Z warunku pierwszego rzędu (9.1) wynika również, że $D_{\mathbf{x}} L(\bar{\mathbf{x}}, \mu, \lambda) = \mathbf{0}^T$. Zatem

$$(\mathbf{x}_k - \bar{\mathbf{x}})^T D^2_{\mathbf{x}} L(\bar{\mathbf{x}}, \mu, \lambda)(\mathbf{x}_k - \bar{\mathbf{x}}) + o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|^2) \leqslant 0.$$

Ponieważ $\mathbf{x}_k = \bar{\mathbf{x}} + \xi_k \mathbf{d}_k$, to wstawiając tę reprezentację do powyższego wzoru dostajemy

$$\xi_k^2 \mathbf{d}_k^T D^2_{\mathbf{x}} L(\bar{\mathbf{x}}, \mu, \lambda)\mathbf{d}_k + o(\xi_k^2 \|\mathbf{d}_k\|^2) \leqslant 0.$$

Dzielimy obie strony przez ξ_k^2

$$\mathbf{d}_k^T D^2_{\mathbf{x}} L(\bar{\mathbf{x}}, \mu, \lambda)\mathbf{d}_k + \frac{o(\xi_k^2 \|\mathbf{d}_k\|^2)}{\xi_k^2} \leqslant 0.$$

Przypomnijmy, że $\|\mathbf{d}_k\| = 1$. A zatem w granicy, gdy $k \to \infty$, drugi składnik powyższej sumy dąży do 0. Korzystając z faktu, że $\lim_{k\to\infty} \mathbf{d}_k = \mathbf{d}$ dostajemy

$$\mathbf{d}^T D^2_{\mathbf{x}} L(\bar{\mathbf{x}}, \mu, \lambda)\mathbf{d} \leqslant 0,$$

co kończy dowód faktu (a).

W dowodzie własności (b) zastosujemy podobne podejście jak powyżej. Z różniczkowalności funkcji f w punkcie $\bar{\mathbf{x}}$ mamy

$$f(\mathbf{x}_k) = f(\bar{\mathbf{x}}) + Df(\bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}}) + o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|).$$

Przypomnijmy, że $f(\mathbf{x}_k) \leqslant f(\bar{\mathbf{x}})$, co implikuje

$$Df(\bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}}) + o(\|\mathbf{x}_k - \bar{\mathbf{x}}\|) \leqslant 0.$$

Korzystając znów z reprezentacji $\mathbf{x}_k = \bar{\mathbf{x}} + \xi_k \mathbf{d}_k$ dostajemy

$$Df(\bar{\mathbf{x}})\mathbf{d}_k + \frac{o(\xi_k \|\mathbf{d}_k\|)}{\xi_k} \leqslant 0.$$

Zatem w granicy, przy $k \to \infty$, otrzymujemy $Df(\bar{\mathbf{x}})\mathbf{d} \leqslant 0$. Zauważając, że $g_i(\mathbf{x}_k) \leqslant g_i(\bar{\mathbf{x}}) = 0$ dla $i \in I(\bar{\mathbf{x}})$ i postępując podobnie jak powyżej dostajemy $Dg_i(\bar{\mathbf{x}})\mathbf{d} \leqslant 0$, $i \in I(\bar{\mathbf{x}})$. Analogicznie również dowodzimy, że $Dh_j(\bar{\mathbf{x}})\mathbf{d} = 0$ dla $j = 1, \dots, l$.

Pomnóżmy obie strony pierwszej równości w (9.1) przez $\mathbf{d}$:

$$Df(\bar{\mathbf{x}})\mathbf{d} + \sum_{i \in I(\bar{\mathbf{x}})} \mu_i Dg_i(\bar{\mathbf{x}})\mathbf{d} + \sum_{j=1}^{l} \lambda_j Dh_j(\bar{\mathbf{x}})\mathbf{d} = 0.$$

Suma ta składa się z wyrazów zerowych lub niedodatnich. Ponieważ sumują się one do zera, to wszystkie muszą być zerowe. W szczególności

$$Df(\bar{\mathbf{x}})\mathbf{d} = 0, \qquad \text{oraz} \qquad Dg_i(\bar{\mathbf{x}})\mathbf{d} = 0, \quad i \in I^*(\bar{\mathbf{x}}).$$

Dowiedliśmy zatem, że $\mathbf{d}$ spełnia warunki (9.2). □

9.2 Podsumowanie

Opiszemy teraz ogólny algorytm postępowania w przypadku rozwiązywania problemów optymalizacyjnych z ograniczeniami mieszanymi.

Krok 1. Szukamy punktów podejrzanych:

$$\begin{aligned} A_1 &= \{\mathbf{x} \in \mathbb{X} : \text{w punkcie } \mathbf{x} \textbf{ nie} \text{ zachodzi warunek regularności}\},\\ A_2 &= \{\mathbf{x} \in \mathbb{X} : \text{w punkcie } \mathbf{x} \text{ zachodzi warunek regularności}\\ &\qquad\qquad \text{i spełnione są warunki pierwszego rzędu}\}. \end{aligned}$$

Krok 2. Sprawdzamy, czy w punktach ze zbiorów A_1, A_2 spełnione są założenia tw. 7.6, tzw. warunki dostateczne pierwszego rzędu. Jeśli tak, to w punktach tych są rozwiązania globalne. Pozostałe kroki podejmujemy, jeśli

- nie znaleźliśmy żadnego rozwiązania globalnego, lub
- chcemy znaleźć wszystkie rozwiązania globalne, lub
- chcemy znaleźć wszystkie rozwiązania lokalne.

Usuwamy ze zbiorów A_1, A_2 punkty, które są rozwiązaniami globalnymi. Oznaczmy nowe zbiory A_1', A_2'.

Krok 3. Eliminujemy ze zbioru A_2' te punkty, gdzie nie zachodzi warunek konieczny drugiego rzędu. Pozostałe punkty oznaczamy A_2''.

Krok 4. W każdym punkcie ze zbioru $A_1' \cup A_2''$ sprawdzamy warunek dostateczny drugiego rzędu. Punkty, w których jest spełniony, są rozwiązaniami lokalnymi.

Krok 5. Optymalność punktów ze zbioru $A_1' \cup A_2''$, w których **nie** jest spełniony warunek dostateczny drugiego rzędu sprawdzamy innymi metodami.

9.3 Przykład

W tym podrozdziale opiszemy bardzo dokładnie rozwiązanie następującego problemu optymalizacyjnego:

$$\begin{cases} (x_1-1)^2+x_2^2 \to \min, \\ 2kx_1 - x_2^2 \leqslant 0, \\ \mathbf{x} = (x_1, x_2) \in \mathbb{R}^2, \end{cases}$$

gdzie $k > 0$ jest parametrem.

Zauważmy, że w każdym punkcie, gdzie aktywne jest ograniczenie, spełniony jest warunek liniowej niezależności ograniczeń: pierwsza pochodna ograniczenia wynosi $[2k, -2x_2] \neq \mathbf{0}^T$. Wynika stąd, że $A_1 = \emptyset$.

Funkcja Lagrange'a dla powyższego problemu ma postać:

$$L(x_1, x_2; \mu) = (x_1-1)^2 + x_2^2 + \mu(2kx_1 - x_2^2).$$

Zapiszmy warunek pierwszego rzędu:

$$[2(x_1-1), 2x_2] + \mu[2k, -2x_2] = \mathbf{0}^T, \tag{9.3}$$

$$\mu(2kx_1 - x_2^2) = 0, \tag{9.4}$$

$$\mu \geqslant 0.$$

Sprawdźmy, czy możliwe jest $\mu = 0$. Wówczas z równania (9.3) dostajemy

$$2(x_1-1) = 0, \qquad 2x_2 = 0,$$

co pociąga $x_1 = 1$, $x_2 = 0$. Punkt ten nie jest dopuszczalny: nie spełnia warunku nierównościowego dla żadnego $k > 0$. Wnioskujemy zatem, że $\mu > 0$ i warunek konieczny pierwszego rzędu może być spełniony tylko w punktach, w których ograniczenie nierównościowe jest aktywne:

$$2kx_1 - x_2^2 = 0. \tag{9.5}$$

Rozpiszmy równanie (9.3):

$$\begin{cases} x_1 - 1 + \mu k = 0, \\ x_2 - \mu x_2 = 0. \end{cases}$$

Z drugiego równania wynika, że albo $\mu = 1$ albo $x_2 = 0$. Jeśli $x_2 = 0$, to z (9.5) dostajemy $x_1 = 0$. Punkt $(0,0)$ wraz z mnożnikiem Lagrange'a $\mu = 1/k$ spełnia warunek pierwszego rzędu.

Rozważmy teraz przypadek $\mu = 1$. Wówczas z równania $x_1 - 1 + \mu k = 0$ dostajemy $x_1 = 1 - k$. Jeśli $k > 1$, to $x_1 < 0$ i równanie (9.5) nie ma rozwiązania. Dla $k = 1$ dostajemy punkt $(0,0)$, zaś dla $k \in (0,1)$ dwa punkty

$$x_1 = 1-k, \qquad x_2 = \pm\sqrt{2k(1-k)}.$$

Podsumowując: $A_1 = \emptyset$ oraz

$$A_2 = \{(0,0)\} \qquad \text{dla} \quad k \geqslant 1,$$

$$A_2 = \{(0,0), (1-k, \sqrt{2k(1-k)}), (1-k, -\sqrt{2k(1-k)})\} \qquad \text{dla} \quad 0 < k < 1.$$

Funkcja $g_1(x_1, x_2) = 2kx_1 - x_2^2$ nie jest quasi-wypukła, wiec nie możemy skorzystać z twierdzenia 7.6 stwierdzającego dostateczność warunku pierwszego rzędu. Zatem $A_2' = A_2$.

Przechodzimy do kroku 3 i sprawdzamy warunek konieczny drugiego rzędu dla punktów z A_2'. Rozważmy najpierw punkt $(0,0)$. Odpowiada mu mnożnik Lagrange'a $\mu = 1/k$. Pierwsza pochodna funkcji Lagrange'a ma postać:

$$D_{\mathbf{x}}L(x_1, x_2; 1/k) = [2(x_1 - 1) + 2, 2x_2(1 - 1/k)].$$

Macierz drugich pochodnych to

$$D^2_{\mathbf{x}}L(x_1, x_2; 1/k) = \begin{bmatrix} 2, & 0 \\ 0, & 2(1 - 1/k) \end{bmatrix}.$$

Na mocy twierdzenia wystarczy sprawdzić, że

$$\mathbf{d}^T \begin{bmatrix} 2, & 0 \\ 0, & 2(1 - 1/k) \end{bmatrix} \mathbf{d} \geqslant 0$$

dla wektorów $\mathbf{d} \in \mathbb{R}^2 \setminus \{\mathbf{0}\}$ spełniających $Dg_1(0,0)\mathbf{d} = [2k, 0]\mathbf{d} = 0$, czyli takich że $d_1 = 0$. Wstawiając do powyższej nierówności dostajemy $(1 - \frac{1}{k})d_2^2 \geqslant 0$. Nierówność ta zachodzi dla $k \geqslant 1$: w punkcie $(0,0)$ *może być* rozwiązanie lokalne. Dla $k < 1$ nierówność nie zachodzi, więc w $(0,0)$ *nie ma* rozwiązania lokalnego.

Załóżmy teraz $0 < k < 1$ i rozważmy dwa pozostałe punkty. W obu przypadkach mnożnik Lagrange'a $\mu = 1$. Mamy zatem

$$\begin{aligned} D_{\mathbf{x}}L(x_1, x_2; 1) &= [2(x_1 - 1) + 2k, 0], \\ D^2_{\mathbf{x}}L(x_1, x_2; 1) &= \begin{bmatrix} 2, & 0 \\ 0, & 0 \end{bmatrix}. \end{aligned}$$

Macierz drugich pochodnych jest nieujemnie określona, więc warunek konieczny drugiego rzędu jest spełniony w obu punktach. Podsumowując:

$$\begin{aligned} &A_2'' = \{(0,0)\} \qquad \text{dla} \quad k \geqslant 1, \\ &A_2'' = \{(1-k, \sqrt{2k(1-k)}), (1-k, -\sqrt{2k(1-k)})\} \qquad \text{dla} \quad 0 < k < 1. \end{aligned}$$

Przechodzimy do sprawdzenia warunku dostatecznego drugiego rzędu. W przypadku $k > 1$ macierz drugich pochodnych funkcji Lagrange'a jest dodatnio określona, więc na mocy tw. 9.3 w punkcie $(0,0)$ jest ścisłe rozwiązanie lokalne. Niestety nie możemy nic powiedzieć o punkcie $(0,0)$, gdy $k = 1$.

Rozważmy przypadek $0 < k < 1$. Zajmijmy się punktem $\bar{\mathbf{x}} = (1-k, \sqrt{2k(1-k)})$. Musimy sprawdzić warunki tw. 9.3:

$$\mathbf{d}^T \begin{bmatrix} 2, & 0 \\ 0, & 0 \end{bmatrix} \mathbf{d} = 2d_1^2 > 0$$

dla $\mathbf{d} \in \mathbb{R}^2 \setminus \{\mathbf{0}\}$ spełniającego $[2k, -2\sqrt{2k(1-k)}]\mathbf{d} = 0$, czyli $d_1 = \frac{1}{k}\sqrt{2k(1-k)}d_2$. A zatem jeśli $\mathbf{d} \neq \mathbf{0}$, to $d_1 \neq 0$ i w punkcie $\bar{\mathbf{x}}$ spełniony jest warunek dostateczny drugiego rzędu: $\bar{\mathbf{x}}$ jest ścisłym rozwiązaniem lokalnym. Analogicznie dowodzimy, że punkt $(1-k, -\sqrt{2k(1-k)})$ jest również ścisłym rozwiązaniem lokalnym.

Pozostaje jeszcze przypadek $k = 1$. Choć $(0,0)$ jest rozwiązaniem globalnym, to nie udało nam się tego pokazać korzystając z teorii Kuhna-Tuckera. Łatwo można to jednak udowodnić bezpośrednio.

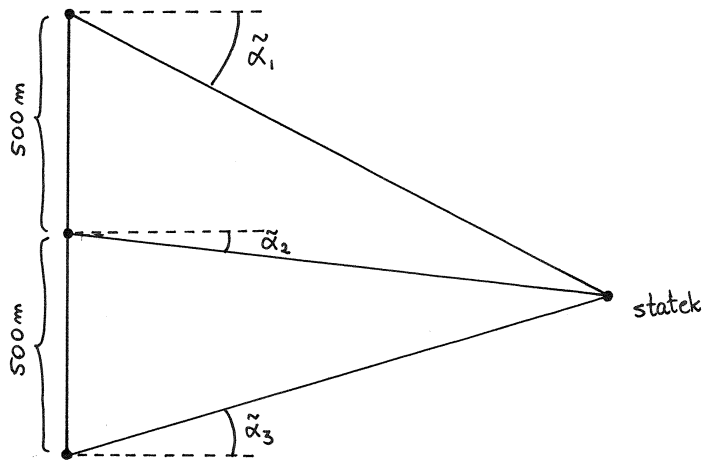
9.4 Zadania

Ćwiczenie 9.1. Rozwiąż geometrycznie i analitycznie zadanie minimalizacji x_2 na zbiorze $W = \{(x_1, x_2) : \ x_1 + x_2 \geqslant 2, \ x_1^2 + x_2^2 \leqslant 2\}$.

Ćwiczenie 9.2. Niech A będzie macierzą $m \times n$ o pełnym rzędzie (tzn. o rzędzie równym $\min(m, n)$). Udowodnij, że rzut na jądro $\ker A$ jest przekształceniem liniowym zadanym wzorem

$$I - A^T(AA^T)^{-1}A.$$

W dowodzie wykorzystaj fakt, że rzeczony rzut, to taki punkt zbioru $\ker A$, który jest najmniej oddalony od rzutowanego punktu.



Rysunek 9.1: Położenie punktów pomiarowych.

Ćwiczenie 9.3. Na brzegu będącym prostym odcinkiem co 500 metrów stoją trzy stacje pomiarowe, patrz rys. 9.1. Każda z nich mierzy kąt między prostą prostopadłą do brzegu a prostą przechodzącą przez punkt pomiarowy i statek. Wyniki obserwacji są następujące:

$$\tilde{\alpha}_1 = 0.083, \qquad \tilde{\alpha}_2 = 0.024, \qquad \tilde{\alpha}_3 = -0.017.$$

Pomiary te są obarczone małymi błędami. Skoryguj je tak, aby były zgodne tzn. wskazywały na ten sam punkt na płaszczyźnie i korekta była jak najmniejsza w sensie średniokwadratowym (suma kwadratów różnic pomiędzy zmierzonymi kątami i skorygowanymi kątami). Znajdź położenie statku względem punktów pomiarowych. W obliczeniach przyjmij, że $\operatorname{tg}(\alpha) \approx \alpha$ dla małych kątów α.

Ćwiczenie 9.4. Metalowa belka ma przekrój trapezoidalny. Pole tego przekroju ze względów wytrzymałościowych musi wynosić S. Górna i boczna powierzchnia musi zostać zabezpieczona antykorozyjnie. Ustal tak kształt przekroju, aby powierzchnia do zabezpieczenia antykorozyjnego była jak najmniejsza.

Rozdział 10

Teoria dualności

W tym rozdziale omówimy elementy teorii dualności, tzn. innej charakteryzacji optymalności rozwiązania zadania optymalizacyjnego z ograniczeniami nierównościowymi. Teoria ta odróżnia się od poprzednio opisywanego podejścia Kuhna-Tuckera tym, że nie wymagamy różniczkowalności funkcji celu f i funkcji ograniczeń g_i. Ponadto, przy odpowiednich założeniach, rozwiązanie pierwotnego zadania optymalizacyjnego możemy łatwo uzyskać z rozwiązania tzw. zadania do niego dualnego. Niezależnie od zadania pierwotnego, zadanie dualne polega na maksymalizacji wklęsłej funkcji celu po nieujemnym oktancie. Jak zobaczymy w następnych rozdziałach, wklęsłość jest cechą gwarantującą dobrą zbieżność metod numerycznych. Prosty zbiór punktów dopuszczalnych dodatkowo przyspiesza działanie i ułatwia implementację algorytmów numerycznych. Nie należy zapominać także o tym, że zadanie dualne jest czasami łatwiejsze do rozwiązania metodami analitycznymi, czego przykłady zobaczymy w zadaniach na końcu niniejszego rozdziału.

10.1 Warunek dostateczny

Definicja 10.1. Niech A, B będą dowolnymi zbiorami, zaś $h : A \times B \to \mathbb{R}$ funkcją. Punkt $(\bar{\mathbf{x}}, \bar{\mu}) \in A \times B$ nazywamy *punktem siodłowym* funkcji h, jeśli

$$h(\bar{\mathbf{x}}, \mu) \leqslant h(\bar{\mathbf{x}}, \bar{\mu}) \leqslant h(\mathbf{x}, \bar{\mu}), \qquad \forall\ \mathbf{x} \in A,\ \mu \in B.$$

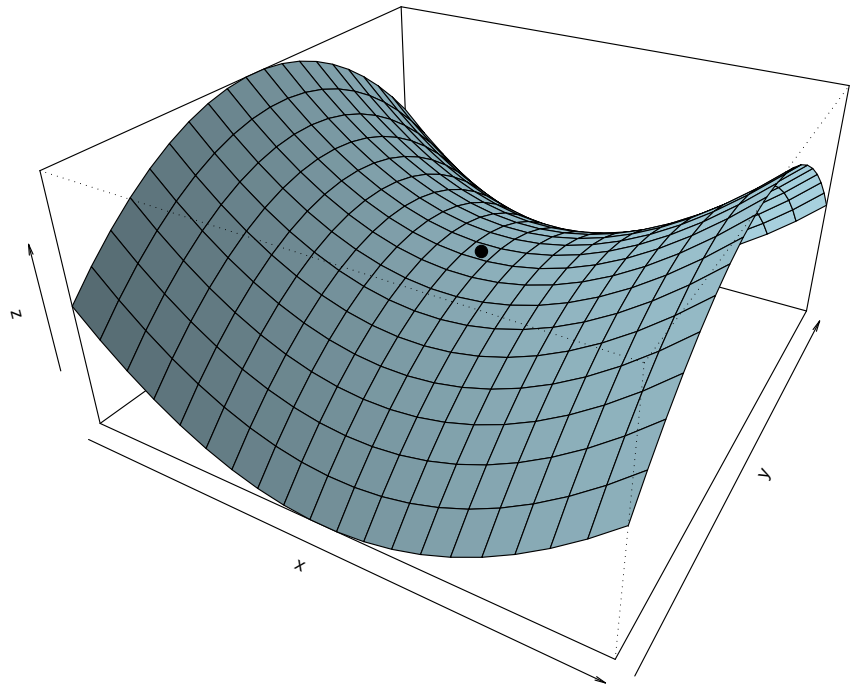
Przykład 10.1. Najprostszym przykładem punktu siodłowego jest "środek siodła" (patrz rys. 10.1): $A, B = \mathbb{R}$, $h(x, \mu) = x^2 - \mu^2$ ma punkt siodłowy w $(0, 0)$. Funkcja h ma minimum w $(0, 0)$ ze względu na zmienną x i maksimum ze względu na μ.

Okazuje się, że punkt siodłowy funkcji Lagrange'a jest związany z rozwiązaniem globalnym problemu optymalizacji z ograniczeniami nierównościowymi:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ g_i(\mathbf{x}) \leqslant 0, \quad i = 1, \ldots, m, \\ \mathbf{x} \in \mathbb{X}. \end{cases} \tag{10.1}$$

Przypomnijmy, że przez W oznaczamy zbiór punktów dopuszczalnych, tj.

$$W = \{\mathbf{x} \in \mathbb{X} : \quad g_1(\mathbf{x}) \leqslant 0, \ldots, g_m(\mathbf{x}) \leqslant 0\}.$$



Rysunek 10.1: Wykres funkcji $(x, y) \mapsto x^2 - y^2$.

Twierdzenie 10.1. *Jeśli* $(\bar{\mathbf{x}}, \bar{\mu}) \in W \times [0, \infty)^m$ *jest punktem siodłowym funkcji Lagrange'a na* $W \times [0, \infty)^m$

$$L(\mathbf{x}, \mu) = f(\mathbf{x}) + \sum_{i=1}^{m} \mu_i g_i(\mathbf{x}),$$

tzn.

$$L(\bar{\mathbf{x}}, \mu) \leqslant L(\bar{\mathbf{x}}, \bar{\mu}) \leqslant L(\mathbf{x}, \bar{\mu}), \qquad \forall\ \mathbf{x} \in W,\ \mu \in [0, \infty)^m,$$

to $\bar{\mathbf{x}}$ *jest rozwiązaniem globalnym problemu* (10.1) *oraz* $\bar{\mu}_i g_i(\bar{\mathbf{x}}) = 0$ *dla* $i = 1, \ldots, m$.

Dowód. Udowodnimy najpierw, że $\bar{\mu}_i g_i(\bar{\mathbf{x}}) = 0$ dla $i = 1, \ldots, m$. Nierówność $L(\bar{\mathbf{x}}, \mu) \leqslant L(\bar{\mathbf{x}}, \bar{\mu})$

możemy rozwinąć następująco:

$$f(\bar{\mathbf{x}}) + \sum_{i=1}^{m} \mu_i g_i(\bar{\mathbf{x}}) \leqslant f(\bar{\mathbf{x}}) + \sum_{i=1}^{m} \bar{\mu}_i g_i(\bar{\mathbf{x}}).$$

Zatem dla każdego $\mu \in [0,\infty)^m$ mamy

$$\sum_{i=1}^{m} \mu_i g_i(\bar{\mathbf{x}}) \leqslant \sum_{i=1}^{m} \bar{\mu}_i g_i(\bar{\mathbf{x}}).$$

W szczególności, podstawiając $\mu = \bar{\mu}/2$ dostajemy

$$\sum_{i=1}^{m} \bar{\mu}_i g_i(\bar{\mathbf{x}}) \geqslant 0.$$

Wiemy, że $\bar{\mathbf{x}}$ jest punktem dopuszczalnym ($\bar{\mathbf{x}} \in W$), czyli $g_i(\bar{\mathbf{x}}) \leqslant 0$ dla $i = 1,\dots,m$. Pamiętając, że $\bar{\mu}$ ma wszystkie współrzędne nieujemne wnioskujemy, iż $\sum_{i=1}^{m} \bar{\mu}_i g_i(\bar{\mathbf{x}}) = 0$ oraz każdy wyraz jest niedodatni. Stąd już wynika, że $\bar{\mu}_i g_i(\bar{\mathbf{x}}) = 0$ dla $i = 1,\dots,m$.

Skorzystamy teraz z drugiej nierówności $L(\bar{\mathbf{x}},\bar{\mu}) \leqslant L(\mathbf{x},\bar{\mu})$ dla $\mathbf{x} \in W$, aby wykazać globalną optymalność $\bar{\mathbf{x}}$. Nierówność tą rozpisujemy następująco:

$$f(\bar{\mathbf{x}}) + \sum_{i=1}^{m} \bar{\mu}_i g_i(\bar{\mathbf{x}}) \leqslant f(\mathbf{x}) + \sum_{i=1}^{m} \bar{\mu}_i g_i(\mathbf{x}), \qquad \mathbf{x} \in W.$$

Z pierwszej części dowodu mamy $\sum_{i=1}^{m} \bar{\mu}_i g_i(\bar{\mathbf{x}}) = 0$. Z faktu, że $\mathbf{x} \in W$ dostajemy $\sum_{i=1}^{m} \bar{\mu}_i g_i(\mathbf{x}) \leqslant 0$. Czyli

$$f(\bar{\mathbf{x}}) \leqslant f(\mathbf{x}), \qquad \mathbf{x} \in W.$$

□

Uwaga 10.1.

1. W powyższym twierdzeniu nie zakładamy otwartości $\mathbb{X}$ ani ciągłości funkcji f i g_i.
2. Zbiór punktów dopuszczalnych W nie musi być wypukły.
3. Nie ma żadnych warunków regularności.
4. Tw. 10.1 nie podaje sposobu szukania punktu siodłowego. Można go znaleźć np. przy pomocy warunków koniecznych pierwszego rzędu, a twierdzenie 10.1 używać jako warunek dostateczny.
5. Tw. 10.1 pełni ważną rolę teoretyczną (podejście dualne) i służy do budowy algorytmów numerycznych rozwiązujących zadanie (10.1).

10.2 Warunek konieczny dla programowania wypukłego

W tym podrozdziale zakładamy, że w problemie (10.1) zbiór $\mathbb{X} \subset \mathbb{R}^n$ jest wypukły oraz funkcje $f, g_i : \mathbb{X} \to \mathbb{R}$, $i = 1,\dots,m$, są wypukłe. Dla takiego zadania optymalizacyjnego warunek punktu siodłowego funkcji Lagrange'a jest warunkiem koniecznym dla rozwiązania globalnego. Zaczniemy od prostszego przypadku, gdy wszystkie funkcje są różniczkowalne, by przejść później do twierdzenia nie wymagającego różniczkowalności. Jak wspomnieliśmy wcześniej, brak wymagania różniczkowalności odróżnia metodę punktu siodłowego od opisanej wcześniej metody Kuhna-Tuckera.

Lemat 10.1. *Załóżmy, że zbiór $\mathbb{X}$ w problemie programowania wypukłego jest otwarty oraz funkcje f i g_i, $i = 1, \ldots, m$, są różniczkowalne w punkcie $\bar{\mathbf{x}}$. Jeśli $\bar{\mathbf{x}}$ jest rozwiązaniem lokalnym (10.1) i spełniony jest jeden z warunków regularności: liniowej niezależności, afiniczności lub Slatera, to istnieje $\bar{\mu} \in [0, \infty)^m$, taki że $(\bar{\mathbf{x}}, \bar{\mu})$ jest punktem siodłowym funkcji Lagrange'a na przestrzeni $\mathbb{X} \times [0, \infty)^m$.*

Dowód. Na mocy twierdzenia 5.2 istnieje wektor mnożników Lagrange'a $\bar{\mu} \in [0, \infty)^m$, dla których spełniony jest warunek optymalności pierwszego rzędu (spełnienie założeń tego twierdzenia wynika z regularności punktu $\bar{\mathbf{x}}$ oraz rozważań rozdziału 6). Ponieważ

$$L(\mathbf{x}, \mu) = f(\mathbf{x}) + \sum_{i=1}^{m} \mu_i g_i(\mathbf{x})$$

jest funkcja wypukłą jako kombinacja funkcji wypukłych z nieujemnymi współczynnikami, to mamy

$$L(\mathbf{x}, \bar{\mu}) \geqslant L(\bar{\mathbf{x}}, \bar{\mu}) + D_{\mathbf{x}} L(\bar{\mathbf{x}}, \bar{\mu})(\mathbf{x} - \bar{\mathbf{x}}).$$

Z twierdzenia Kuhna-Tuckera mamy

$$D_{\mathbf{x}} L(\bar{\mathbf{x}}, \bar{\mu}) = Df(\bar{\mathbf{x}}) + \sum_{i=1}^{m} \bar{\mu}_i Dg_i(\bar{\mathbf{x}}) = \mathbf{0}^T,$$

czyli $D_{\mathbf{x}} L(\bar{\mathbf{x}}, \bar{\mu})(\mathbf{x} - \bar{\mathbf{x}}) = 0$. Stąd $L(\mathbf{x}, \bar{\mu}) \geqslant L(\bar{\mathbf{x}}, \bar{\mu})$.
Aby udowodnić nierówność $L(\bar{\mathbf{x}}, \bar{\mu}) \geqslant L(\bar{\mathbf{x}}, \mu)$ zauważmy, że

$$\sum_{i=1}^{m} \mu_i g_i(\bar{\mathbf{x}}) \leqslant 0 = \sum_{i=1}^{m} \bar{\mu}_i g_i(\bar{\mathbf{x}}),$$

bo $\mu_i \geqslant 0$ a $g_i(\bar{\mathbf{x}}) \leqslant 0$ z faktu, że $\bar{\mathbf{x}}$ jako rozwiązanie lokalne jest puntem dopuszczalnym. Ostatnia równość jest zawarta w tezie twierdzenia 5.2. □

Uwaga 10.2. Na mocy twierdzenia 7.6 każdy punkt spełniający warunki pierwszego rzędu jest rozwiązaniem globalnym zadania programowania wypukłego. Nie jest zatem ważne, czy wymagać będziemy w powyższym lemacie, aby $\bar{\mathbf{x}}$ był rozwiązaniem lokalnym czy globalnym.

Przechodzimy teraz do głównego twierdzenia.

Twierdzenie 10.2. *Niech $\bar{\mathbf{x}} \in \mathbb{X}$ będzie rozwiązaniem globalnym problemu programowania wypukłego (10.1) oraz istnieje $\mathbf{x}^* \in \mathbb{X}$, taki że $g_i(\mathbf{x}^*) < 0$ dla $i = 1, \ldots, m$. Wówczas istnieje $\bar{\mu} \in [0, \infty)^m$ o tej własności, że $(\bar{\mathbf{x}}, \bar{\mu})$ jest punktem siodłowym funkcji Lagrange'a na przestrzeni $\mathbb{X} \times [0, \infty)^m$, tzn.*

$$L(\bar{\mathbf{x}}, \mu) \leqslant L(\bar{\mathbf{x}}, \bar{\mu}) \leqslant L(\mathbf{x}, \bar{\mu}), \qquad \forall\, \mathbf{x} \in \mathbb{X},\ \mu \in [0, \infty)^m.$$

Ponadto, $\bar{\mu}_i g_i(\bar{\mathbf{x}}) = 0$ dla $i = 1, \ldots, m$.

Uwaga 10.3. Punkt siodłowy funkcji Lagrange'a jest rozpatrywany na różnych przestrzeniach w tw. 10.1 i 10.2. W drugim ze wspomnianych twierdzeń przestrzeń jest większa, gdyż pierwsza zmienna przebiega cały zbiór $\mathbb{X}$, a nie tylko zbiór punktów dopuszczalnych W. W sumie, dostajemy równoważność istnienia punktu siodłowego funkcji Lagrange'a i rozwiązania globalnego zadania optymalizacji wypukłej.

Dowód tw. 10.2. Podobnie jak w dowodzie warunku koniecznego pierwszego rzędu, tw. 5.2, główną rolę będzie tutaj odgrywać twierdzenie o oddzielaniu zbiorów wypukłych. Wskaże nam ono wektor mnożników Lagrange'a $\bar{\mu}$.

Oznaczmy $\mathbf{g}(\mathbf{x}) = \big(g_1(\mathbf{x}), \dots, g_m(\mathbf{x})\big)^T$. Zdefiniujmy następujące podzbiory $\mathbb{R}^{m+1}$:

$$\begin{aligned} A &= \{\bar{\mathbf{y}} = (y_0, \mathbf{y}) \in \mathbb{R} \times \mathbb{R}^m : \quad y_0 \geqslant f(\mathbf{x}),\ \mathbf{y} \geqslant \mathbf{g}(\mathbf{x}) \text{ dla pewnego } \mathbf{x} \in \mathbb{X}\}, \\ B &= \{\bar{\mathbf{y}} = (y_0, \mathbf{y}) \in \mathbb{R} \times \mathbb{R}^m : \quad y_0 = f(\mathbf{x}),\ \mathbf{y} = \mathbf{g}(\mathbf{x}) \text{ dla pewnego } \mathbf{x} \in \mathbb{X}\}, \\ C &= \{\bar{\mathbf{y}} = (y_0, \mathbf{y}) \in \mathbb{R} \times \mathbb{R}^m : \quad y_0 < f(\bar{\mathbf{x}}),\ \mathbf{y} < \mathbf{0}\}. \end{aligned}$$

W powyższych definicjach użyty został uproszczony zapis "nierówności między dwoma wektorami" (taka uproszczająca zapis konwencja została wyjaśniona na początku skryptu).
Zauważmy, że zbiór C jest "oszukany": jest on produktem półprostej kończącej się w minimum $f(\bar{\mathbf{x}})$ i stożka $\{\mathbf{y} < \mathbf{0}\}$. Łatwo widzimy, że jest on wypukły. Z optymalności $\bar{\mathbf{x}}$ dostajemy, że $B \cap C = \emptyset$. Zbiór B nie jest jednak wypukły, więc nie możemy stosować twierdzeń o oddzielaniu. Radą na to jest spostrzeżenie, że zamiast B można brać zbiór wypukły A, który ma również puste przecięcie z C. Przypuśćmy przeciwnie: niech $\bar{\mathbf{y}} = (y_0, \mathbf{y}) \in A \cap C$. Mamy zatem dla pewnego $\mathbf{x}' \in \mathbb{X}$ następujące nierówności:

$$y_0 \geqslant f(\mathbf{x}'), \quad \mathbf{y} \geqslant \mathbf{g}(\mathbf{x}'), \quad y_0 < f(\bar{\mathbf{x}}), \quad \mathbf{y} < \mathbf{0}.$$

Wnioskujemy z nich, że $f(\mathbf{x}') < f(\bar{\mathbf{x}})$ oraz $\mathbf{g}(\mathbf{x}') < \mathbf{0}$. A zatem punkt $\mathbf{x}'$ jest dopuszczalny, zaś funkcja f przyjmuje w nim wartość mniejszą od $f(\bar{\mathbf{x}})$. Przeczy to optymalności $\bar{\mathbf{x}}$.

Przykład 10.2. Przed przystąpieniem do dalszej części dowodu popatrzmy na zbiory A, B, C dla następującego problemu optymalizacyjnego:

$$\begin{cases} -x \to \min, \\ x^2 - 1 \leqslant 0, \\ x \in \mathbb{X} = \mathbb{R}. \end{cases}$$

Rozwiązaniem tego zagadnienia jest $\bar{x} = 1$. Mamy jedno ograniczenie, więc szukane zbiory leżą w przestrzeni $\mathbb{R}^2$. Na rysunku 10.2 znajduje się ich szkic. Zwróćmy uwagę na zależność między zbiorami A i B. Zbiór B jest brzegiem A dla $y_0 \leqslant 0$, lecz znajduje się w jego wnętrzu dla $y_0 > 0$.

Powróćmy do dowodu. Wypukłość zbioru C już została uzasadniona. Wypukłość A dowodzimy bezpośrednio. Weźmy dwa punkty $\bar{\mathbf{y}}', \bar{\mathbf{y}}'' \in A$ oraz $\lambda \in (0, 1)$. Istnieją wówczas punkty $\mathbf{x}', \mathbf{x}'' \in \mathbb{X}$ o następującej własności:

$$\begin{aligned} y_0' &\geqslant f(\mathbf{x}'), \quad \mathbf{y}' \geqslant \mathbf{g}(\mathbf{x}'), \\ y_0'' &\geqslant f(\mathbf{x}''), \quad \mathbf{y}'' \geqslant \mathbf{g}(\mathbf{x}''). \end{aligned}$$

Zdefiniujmy $\mathbf{x} = \lambda\mathbf{x}' + (1-\lambda)\mathbf{x}''$. Z wypukłości $\mathbb{X}$ wynika, że $\mathbf{x} \in \mathbb{X}$. Mamy również

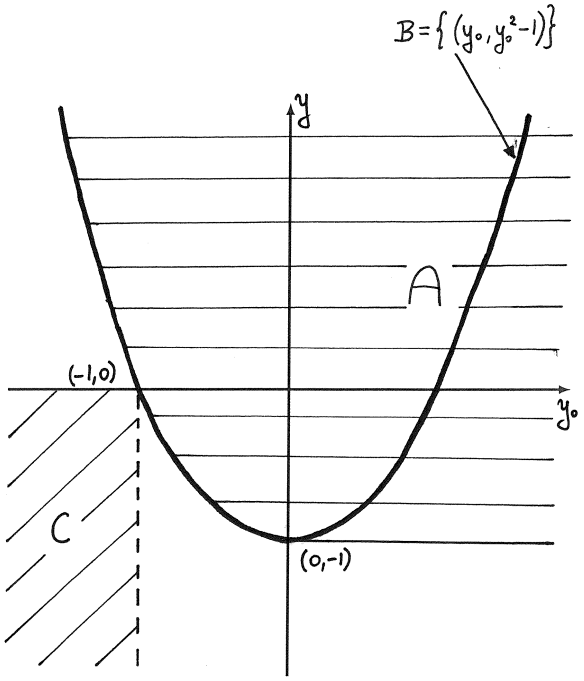
$$\lambda y_0' + (1-\lambda) y_0'' \geqslant \lambda f(\mathbf{x}') + (1-\lambda) f(\mathbf{x}'') \geqslant f(\lambda\mathbf{x}' + (1-\lambda)\mathbf{x}'') = f(\mathbf{x}),$$

gdzie pierwsza nierówność wynika z powyższych własności $\mathbf{x}'$ i $\mathbf{x}''$, zaś druga nierówność z wypukłości f. Podobnie, korzystając z wypukłości składowych wektora $\mathbf{g}$, pokazujemy, że

$$\lambda\mathbf{y}' + (1-\lambda)\mathbf{y}'' \geqslant \mathbf{g}(\mathbf{x}).$$

Stąd $\bar{\mathbf{y}} = \lambda\bar{\mathbf{y}}' + (1-\lambda)\bar{\mathbf{y}}'' \in A$, ponieważ

$$y_0 \geqslant f(\mathbf{x}), \qquad \mathbf{y} \geqslant \mathbf{g}(\mathbf{x})$$



Rysunek 10.2: Szkic zbiorów A, B, C zdefiniowanych w dowodzie twierdzenia 10.2.

dla zdefiniowanego powyżej punktu $\mathbf{x}$. Kończy to dowód wypukłości zbioru A.
Na mocy twierdzenia o oddzielaniu, tw. 3.1, istnieje $\tilde{\mu} \in \mathbb{R}^{m+1}$, $\tilde{\mu} \neq \mathbf{0}$ i takie że

$$\tilde{\mu}^T \bar{\mathbf{y}} \geqslant \tilde{\mu}^T \bar{\mathbf{z}}, \qquad \forall\, \bar{\mathbf{y}} \in A,\ \bar{\mathbf{z}} \in C.$$

Z faktu, że $\sup_{\bar{\mathbf{z}} \in C} \tilde{\mu}^T \bar{\mathbf{z}} < \infty$ wynika, że $\tilde{\mu} \geqslant 0$. Z ciągłości funkcji liniowej wnioskujemy, że $\bar{\mathbf{z}}$ można brać z domknięcia C:

$$\tilde{\mu}^T \bar{\mathbf{y}} \geqslant \tilde{\mu}^T \bar{\mathbf{z}}, \qquad \forall\, \bar{\mathbf{y}} \in A,\ \bar{\mathbf{z}} \in \operatorname{cl} C.$$

Zatem dla $\bar{\mathbf{z}} = (f(\bar{\mathbf{x}}), \mathbf{0})$ mamy

$$\tilde{\mu}_0 y_0 + \sum_{i=1}^{m} \tilde{\mu}_i y_i \geqslant \tilde{\mu}_0 f(\bar{\mathbf{x}}), \qquad \forall\, (y_0, \mathbf{y}) \in A.$$

W szczególności powyższa nierówność zachodzi dla $y_0 = f(\mathbf{x})$ i $\mathbf{y} = \mathbf{g}(\mathbf{x})$ dla $\mathbf{x} \in \mathbb{X}$:

$$\tilde{\mu}_0 f(\mathbf{x}) + \sum_{i=1}^{m} \tilde{\mu}_i g_i(\mathbf{x}) \geqslant \tilde{\mu}_0 f(\bar{\mathbf{x}}). \tag{10.2}$$

Wykażemy teraz, że $\tilde{\mu}_0 \neq 0$, co razem z obserwacją $\tilde{\mu} \geqslant \mathbf{0}$ będzie implikować $\tilde{\mu}_0 > 0$. Dowód przeprowadzimy przez zaprzeczenie: załóżmy $\tilde{\mu}_0 = 0$. Wówczas z nierówności (10.2) wynika, że

$$\sum_{i=1}^{m} \tilde{\mu}_i g_i(\mathbf{x}) \geqslant 0, \qquad \forall\, \mathbf{x} \in \mathbb{X}.$$

W szczególności zachodzi to dla punktu $\mathbf{x}^*$ z założenia twierdzenia. W tym punkcie mamy jednak $g_i(\mathbf{x}^*) < 0$ dla każdego $i = 1, \ldots, m$. To, w połączeniu z faktem, iż $\tilde{\mu} \geqslant \mathbf{0}$ pociąga

$\tilde{\mu}_1 = \ldots = \tilde{\mu}_m = 0$. Przypomnijmy, że $\tilde{\mu}_0 = 0$, czyli $\tilde{\mu} = \mathbf{0}$, a to przeczy wyborowi $\tilde{\mu}$ z twierdzenia o oddzielaniu.

Wiemy zatem, że $\tilde{\mu}_0 > 0$. Zdefiniujmy

$$\bar{\mu} = \Big(\frac{\tilde{\mu}_1}{\tilde{\mu}_0}, \ldots, \frac{\tilde{\mu}_m}{\tilde{\mu}_0}\Big)^T.$$

Oczywiście $\bar{\mu} \in [0,\infty)^m$. Ponieważ $\bar{\mathbf{x}}$, jako rozwiązanie, jest punktem dopuszczalnym, to $g_i(\bar{\mathbf{x}}) \leqslant 0$, $i = 1,\ldots,m$, i $\sum_{i=1}^m \bar{\mu}_i g_i(\bar{\mathbf{x}}) \leqslant 0$. Dodajemy tę sumę do prawej strony nierówności (10.2) podzielonej przez $\tilde{\mu}_0$:

$$f(\mathbf{x}) + \sum_{i=1}^m \bar{\mu}_i g_i(\mathbf{x}) \geqslant f(\bar{\mathbf{x}}) + \sum_{i=1}^m \bar{\mu}_i g_i(\bar{\mathbf{x}}), \qquad \forall\, \mathbf{x} \in \mathbb{X}.$$

Inaczej,

$$L(\mathbf{x}, \bar{\mu}) \geqslant L(\bar{\mathbf{x}}, \bar{\mu}), \qquad \forall\, \mathbf{x} \in \mathbb{X}.$$

Pozostaje jeszcze wykazanie drugiej nierówności punktu siodłowego. Biorąc $\mathbf{x} = \bar{\mathbf{x}}$ i dzieląc obie strony nierówności (10.2) przez $\tilde{\mu}_0$ dostajemy $\sum_{i=1}^m \bar{\mu}_i g_i(\bar{\mathbf{x}}) \geqslant 0$. Z drugiej strony punkt $\bar{\mathbf{x}}$ jest dopuszczalny, czyli $g_i(\bar{\mathbf{x}}) \leqslant 0$. Pamiętając, że $\bar{\mu} \geqslant 0$ wnioskujemy, że każdy składnik tej sumy jest niedodatni. Stąd już mamy

$$\bar{\mu}_i g_i(\bar{\mathbf{x}}) = 0, \qquad i = 1,\ldots,m.$$

Dla dowolnego innego $\mu \in [0,\infty)^m$ mamy $\sum_{i=1}^m \mu_i g_i(\bar{\mathbf{x}}) \leqslant 0$, czyli

$$\sum_{i=1}^m \mu_i g_i(\bar{\mathbf{x}}) \leqslant \sum_{i=1}^m \bar{\mu}_i g_i(\bar{\mathbf{x}}), \qquad \forall \mu \in [0,\infty)^m.$$

Ta nierówność jest równoważna

$$L(\bar{\mathbf{x}}, \mu) \leqslant L(\bar{\mathbf{x}}, \bar{\mu}), \qquad \forall\, \mu \in [0,\infty)^m.$$

□

10.3 Zadanie pierwotne i dualne

Z teorią puntów siodłowych związane są pojęcia zadania pierwotnego i dualnego. Rozważmy zadanie optymalizacyjne (10.1) i związaną z nim funkcję Lagrange'a $L(\mathbf{x},\mu)$. Zdefiniujmy funkcję $L_P : \mathbb{X} \to (-\infty,\infty]$

$$L_P(\mathbf{x}) = \sup_{\mu \in [0,\infty)^m} L(\mathbf{x},\mu).$$

Zauważmy, że

$$L_P(\mathbf{x}) = \begin{cases} f(\mathbf{x}), & \mathbf{g}(\mathbf{x}) \leqslant \mathbf{0}, \\ \infty, & \text{w przeciwnym przypadku.} \end{cases}$$

A zatem zadanie (10.1) można zapisać w wydawałoby się prostszej postaci

$$L_P(\mathbf{x}) \to \min, \qquad \mathbf{x} \in \mathbb{X}.$$

Niestety powyższe przeformułowanie sprowadza się do rozwiązania oryginalnego zadania, a więc nie zawiera żadnej „wartości dodanej"; ale tylko do czasu. Zanim zdradzimy jego zastosowanie, zdefiniujmy kolejną funkcję $L_D : [0,\infty)^m \to [-\infty,\infty)$

$$L_D(\mu) = \inf_{\mathbf{x} \in \mathbb{X}} L(\mathbf{x},\mu).$$

Uwaga 10.4.

1. Dla dowolnego $\mathbf{x} \in \mathbb{X}$ i $\mu \in [0,\infty)^m$ mamy $L_P(\mathbf{x}) \geqslant L(\mathbf{x},\mu) \geqslant L_D(\mu)$.
2. Jeśli $(\bar{\mathbf{x}},\bar{\mu})$ jest punktem siodłowym funkcji Lagrange'a na $\mathbb{X}\times[0,\infty)^m$, to $L_P(\bar{\mathbf{x}}) = L_D(\bar{\mu})$.

Jeśli $(\bar{\mathbf{x}},\bar{\mu})$ jest punktem siodłowym to mamy nierówność $L(\bar{\mathbf{x}},\bar{\mu}) \leqslant L(\mathbf{x},\bar{\mu})$. W połączeniu z punktem 1. daje to równość $L(\bar{\mathbf{x}},\bar{\mu}) = L_D(\bar{\mu})$. Podobnie dowodzi się równości $L(\bar{\mathbf{x}},\bar{\mu}) = L_P(\bar{\mathbf{x}})$. Powyższe spostrzeżenia kierują nas we właściwą stronę. Będziemy wykorzystywać funkcje L_P i L_D do znajdowania punktów siodłowych.

Definicja 10.2. *Zadaniem pierwotnym* nazywamy problem optymalizacyjny

$$L_P(\mathbf{x}) \to \min, \qquad \mathbf{x} \in \mathbb{X}.$$

Zadaniem dualnym do niego jest problem optymalizacyjny

$$L_D(\mu) \to \max, \qquad \mu \in [0,\infty)^m.$$

Z własności wspomnianych w uwadze 10.4 wynika, że wartość rozwiązania zadania pierwotnego jest nie mniejsza niż wartość rozwiązania zadania dualnego

$$\inf_{\mathbf{x}\in\mathbb{X}} L_P(\mathbf{x}) \geqslant \sup_{\mu\in[0,\infty)^m} L_D(\mu).$$

Co więcej, rozwiązanie zadania dualnego daje dolne oszacowanie na wartość funkcji f.

Lemat 10.2 (Słabe twierdzenie o dualności)**.** *Dla dowolnego punktu dopuszczalnego $\mathbf{x} \in W$ oraz dowolnego $\mu \in [0,\infty)^m$ mamy*

$$f(\mathbf{x}) \geqslant L_D(\mu).$$

A zatem

$$f(\mathbf{x}) \geqslant \sup_{\mu\in[0,\infty)^m} L_D(\mu).$$

Dowód. Mamy

$$f(\mathbf{x}) \geqslant L(\mathbf{x},\mu) \geqslant L_D(\mu).$$

Pierwsza z tych nierówności wynika z założenia $\mathbf{x} \in W$ bo wtedy $g_i(\mathbf{x}) \leqslant 0$. Druga nierówność wynika z punktu 1. uwagi 10.4. □

Definicja 10.3. *Luką dualności* nazwiemy różnicę między wartością rozwiązania zadania pierwotnego i dualnego:

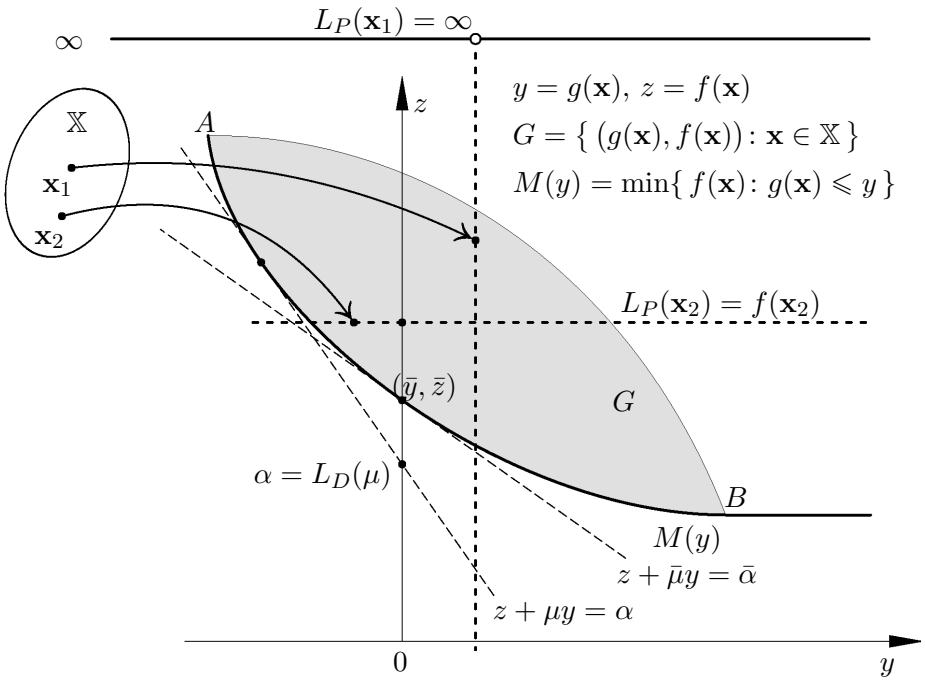
$$\inf_{\mathbf{x}\in\mathbb{X}} L_P(\mathbf{x}) - \sup_{\mu\in[0,\infty)^m} L_D(\mu).$$

Zapiszmy w języku funkcji pierwotnej i dualnej warunek punktu siodłowego: $(\bar{\mathbf{x}},\bar{\mu})$ jest punktem siodłowym, jeśli

$$L_P(\bar{\mathbf{x}}) = L(\bar{\mathbf{x}},\bar{\mu}) = L_D(\bar{\mu}).$$

Innymi słowy, jeśli funkcja Lagrange'a posiada punkt siodłowy, to luka dualności jest zerowa. Ma to miejsce, na przykład, jeśli spełnione są założenia tw. 10.2.

Rysunek 10.3 ilustruje rozwiązanie zadania pierwotnego i dualnego, gdy problem optymalizacyjny zawiera tylko jedno ograniczenie nierównościowe ($m = 1$). Zbiór G jest zbiorem wartości funkcji $f(\mathbf{x})$ i $g(\mathbf{x})$ dla $\mathbf{x} \in \mathbb{X}$. Proste $z + \mu y = \alpha$ są zbiorami wartości funkcji Lagrange'a



Rysunek 10.3: Szkic rozwiązania problemu pierwotnego i dualnego.

$L(\mathbf{x}, \mu) = f(\mathbf{x}) + \mu g(\mathbf{x})$. Z definicji funkcji pierwotnej $L_P(\mathbf{x})$ wiemy, że $L_P(\mathbf{x}) = f(\mathbf{x})$, jeśli $g(\mathbf{x}) \leqslant 0$, bo $\sup_{\mu \geqslant 0}\{\alpha : \alpha = z + \mu y\}$ jest osiągane dla $\mu = 0$. Ten przypadek ilustruje obraz punktu $\mathbf{x}_2$. Gdy $g(\mathbf{x}) > 0$, to $L_P(\mathbf{x}) = +\infty$. Ten przypadek ilustruje obraz punktu $\mathbf{x}_1$. Funkcję dualną $L_D(\mu)$ znajdujemy rozpatrując proste $z + \mu y = \alpha$ przy ustalonym μ poszukując $\inf_{\mathbf{x} \in \mathbb{X}} \alpha$. Z rysunku widać, że minimum jest osiągane na prostej stycznej do obszaru G, a wartość funkcji $L_D(\mu)$ to punkt przecięcia tej prostej z osią $0z$. Z rysunku widać, że prostą dającą największą wartość $L_D(\mu)$ jest prosta $z + \bar{\mu} y = \bar{\alpha}$ o nachyleniu $-\bar{\mu}$, która jest styczna do obszaru G w punkcie $(\bar{y}, \bar{z})$. Punkt ten jest też rozwiązaniem zadania pierwotnego, bo $\bar{z} = \inf_{\mathbf{x} \in \mathbb{X}, g(\mathbf{x}) \leqslant 0} f(\mathbf{x})$.

Twierdzenie 10.3 (Mocne twierdzenie o dualności)**.** *Niech $\mathbb{X}$ będzie niepustym wypukłym podzbiorem $\mathbb{R}^n$ a funkcje f oraz g_i, $i = 1 \ldots, m$, będą wypukłe na $\mathbb{X}$. Załóżmy ponadto, że istnieje punkt $\mathbf{x}^* \in \mathbb{X}$, w którym $g_i(\mathbf{x}^*) < 0$, $i = 1 \ldots, m$. Wtedy*

$$\inf_{\mathbf{x} \in \mathbb{X}} L_P(\mathbf{x}) = \sup_{\mu \in [0,\infty)^m} L_D(\mu).$$

Jeśli $\inf_{\mathbf{x} \in \mathbb{X}} L_P(\mathbf{x})$ jest skończone, to $\sup_{\mu \in [0,\infty)^m} L_D(\mu)$ jest osiągane w punkcie $\bar{\mu}$ takim że $\bar{\mu} \geqslant 0$. Jeśli $\inf_{\mathbf{x} \in \mathbb{X}} L_P(\mathbf{x})$ jest osiągane w punkcie $\bar{\mathbf{x}}$, to $\bar{\mu}_i g_i(\bar{\mathbf{x}}) = 0$, $i = 1 \ldots, m$.

Dowód. Niech $\gamma = \inf_{\mathbf{x} \in \mathbb{X}} L_P(\mathbf{x})$. Jeśli $\gamma = -\infty$, to $\sup_{\mu \in [0,\infty)^m} L_D(\mu) = -\infty$ (z lematu 10.2), czyli twierdzenie jest prawdziwe. Załóżmy teraz, że $\gamma > -\infty$.

Z dowodu twierdzenia 10.2 zastosowanego do zbiorów

$$\begin{aligned} A &= \{\bar{\mathbf{y}} = (y_0, \mathbf{y}) \in \mathbb{R} \times \mathbb{R}^m : \quad y_0 \geqslant f(\mathbf{x}),\ \mathbf{y} \geqslant \mathbf{g}(\mathbf{x}) \text{ dla pewnego } \mathbf{x} \in \mathbb{X}\}, \\ C &= \{\bar{\mathbf{y}} = (y_0, \mathbf{y}) \in \mathbb{R} \times \mathbb{R}^m : \quad y_0 < \gamma,\ \mathbf{y} < \mathbf{0}\}, \end{aligned}$$

otrzymamy

$$f(\mathbf{x}) + \sum_{i=1}^{m} \bar{\mu}_i g_i(\mathbf{x}) \geqslant \gamma, \qquad \forall\, \mathbf{x} \in \mathbb{X}. \tag{10.3}$$

Oznacza to, że

$$L_D(\bar{\mu}) = \inf_{\mathbf{x}\in\mathbb{X}}\Big(f(\mathbf{x}) + \sum_{i=1}^{m} \bar{\mu}_i g_i(\mathbf{x})\Big) \geqslant \gamma.$$

Z uwagi 10.4 wynika, że $\gamma = \inf_{\mathbf{x}\in\mathbb{X}} L_P(\mathbf{x}) \geqslant \sup_{\mu\in[0,\infty)^m} L_D(\mu) \geqslant L_D(\bar{\mu}) \geqslant \gamma$. Stąd $L_D(\bar{\mu}) = \gamma$ a $\bar{\mu}$ jest rozwiązaniem problemu dualnego.

Jeśli $\inf_{\mathbf{x}\in\mathbb{X}} L_P(\mathbf{x})$ jest osiągane w punkcie $\bar{\mathbf{x}}$, to $L_P(\bar{\mathbf{x}}) = f(\bar{\mathbf{x}})$ z definicji funkcji $L_P(\mathbf{x})$ oraz $\bar{\mathbf{x}}$ jest rozwiązaniem problemu pierwotnego i $\bar{\mathbf{x}} \in \mathbb{X}$, $g_i(\bar{\mathbf{x}}) \leqslant 0$, $i = 1\ldots, m$, oraz $f(\bar{\mathbf{x}}) = \gamma$. Kładąc w nierówności (10.3) $\mathbf{x} = \bar{\mathbf{x}}$ dostajemy $\sum_i \bar{\mu}_i g_i(\bar{\mathbf{x}}) \geqslant 0$. Ponieważ $\bar{\mu}_i \geqslant 0$ a $g_i(\bar{\mathbf{x}}) \leqslant 0$, to wynika stąd równość $\bar{\mu}_i g_i(\bar{\mathbf{x}}) = 0$ dla $i = 1\ldots, m$. □

Możemy teraz zaproponować algorytm rozwiązywania zagadnienia (10.1) przy pomocy metod dualnych.

1. Rozwiąż zadanie dualne. Jego wartość daje dolne ograniczenie na wartość rozwiązania problemu pierwotnego na mocy lematu 10.2.

2. Załóżmy, że istnieje rozwiązanie skończone $\bar{\mu} \in [0,\infty)^m$ zadania dualnego oraz taki punkt $\bar{\mathbf{x}} \in \mathbb{X}$, że $L_D(\bar{\mu}) = L(\bar{\mathbf{x}}, \bar{\mu})$. Jeśli $\bar{\mathbf{x}}$ jest dopuszczalny oraz $f(\bar{\mathbf{x}}) = L_D(\bar{\mu})$, to $(\bar{\mathbf{x}}, \bar{\mu})$ jest punktem siodłowym funkcji Lagrange'a i twierdzenie 10.1 implikuje, że $\bar{\mathbf{x}}$ jest rozwiązaniem zadania (10.1).

Wyjaśnijmy warunki punktu drugiego. Z faktu $L_D(\bar{\mu}) = L(\bar{\mathbf{x}}, \bar{\mu})$ wynika, że $L(\bar{\mathbf{x}}, \bar{\mu}) \leqslant L(\mathbf{x}, \bar{\mu})$ dla dowolnego $\mathbf{x} \in \mathbb{X}$. Zatem mamy prawą nierówność warunku punktu siodłowego. Pozostaje jeszcze nierówność lewa. Przypomnijmy, że $L_P(\mathbf{x}) = f(\mathbf{x})$ dla dopuszczalnego punktu $\mathbf{x}$ i $\inf_{\mathbf{x}\in\mathbb{X}} L_P(\mathbf{x}) \geqslant L_D(\mu)$ dla dowolnego $\mu \in [0,\infty)^m$. W punkcie drugim zakładamy, że $f(\bar{\mathbf{x}}) = L_D(\bar{\mu})$, co pociąga

$$L_P(\bar{\mathbf{x}}) = f(\bar{\mathbf{x}}) = L_D(\bar{\mu}),$$

a zatem $(\bar{\mathbf{x}}, \bar{\mu})$ jest punktem siodłowym.

10.4 Zadania

Ćwiczenie 10.1. Udowodnij, że jeśli w problemie optymalizacyjnym (10.1) funkcje f i g_i, $i = 1,\ldots,m$, są wypukłe, to punkt spełniający warunek konieczny pierwszego rzędu jest punktem siodłowym funkcji Lagrange'a na przestrzeni $\mathbb{X} \times [0,\infty)^m$.

Ćwiczenie 10.2. Uzasadnij, że $L_P(\mathbf{x}) \geqslant L(\mathbf{x}, \mu) \geqslant L_D(\mu)$ dla dowolnego $\mathbf{x} \in \mathbb{X}$ i $\mu \in [0,\infty)^m$.

Ćwiczenie 10.3. Uzasadnij nierówność:

$$\inf_{\mathbf{x}\in\mathbb{X}} L_P(\mathbf{x}) \geqslant \sup_{\mu\in[0,\infty)^m} L_D(\mu).$$

Ćwiczenie 10.4. Udowodnij, że jeśli $(\bar{\mathbf{x}}, \bar{\mu})$ jest punktem siodłowym funkcji Lagrange'a, to $L_P(\bar{\mathbf{x}}) = L_D(\bar{\mu})$ lub, innymi słowy, luka dualności jest zerowa.

Ćwiczenie 10.5. Udowodnij lemat 10.2.

Ćwiczenie 10.6. Wykaż, że funkcja dualna L_D jest wklęsła.

Ćwiczenie 10.7. Podaj przykład problemu optymalizacyjnego, dla którego luka dualności jest dodatnia.

Ćwiczenie 10.8. Rozwiąż metodą dualną zadanie

$$\begin{cases} x_1 \to \min, \\ x_1^2 + x_2^2 \leqslant 2, \\ (x_1, x_2) \in \mathbb{X}, \end{cases}$$

gdzie $\mathbb{X} = \{\mathbf{x} \in \mathbb{R}^2 : \ x_1 \geqslant 1\}$. Zwróć uwagę na umieszczenie jednego z ograniczeń w zbiorze $\mathbb{X}$.

Ćwiczenie 10.9. Rozwiąż metodą dualną zadanie

$$\begin{cases} \frac{1}{2}\sum_{i=1}^n x_i^2 \to \min, \\ \sum_{i=1}^n x_i = 1, \\ 0 \leqslant x_i \leqslant u_i, \quad i = 1, \ldots, n, \\ \mathbf{x} \in \mathbb{R}^n, \end{cases}$$

gdzie $0 \leqslant u_1 \leqslant \ldots \leqslant u_n$ oraz $\sum_{i=1}^n u_i \geqslant 1$.

Wskazówka. Rozważ zbiór $\mathbb{X} = \{\mathbf{x} \in \mathbb{R}^n : \ 0 \leqslant x_i \leqslant u_i \ \textit{dla}\ i = 1, \ldots, n\}$.

Ćwiczenie 10.10. Znajdź zadanie dualne (czyli formę zadania $\sup_{\mu \in [0,\infty)^m} L_D(\mu)$) dla zadania optymalizacji liniowej

$$\begin{cases} \mathbf{d}^T\mathbf{x} \to \min, \\ A\mathbf{x} \leqslant \mathbf{b}, \\ \mathbf{x} \in \mathbb{R}^n, \end{cases}$$

gdzie $\mathbf{d} \in \mathbb{R}^n$, A jest macierzą $m \times n$ i $\mathbf{b} \in \mathbb{R}^m$.

Ćwiczenie 10.11. Znajdź zadanie dualne do zadania programowania kwadratowego

$$\begin{cases} \frac{1}{2}\mathbf{x}^T H\mathbf{x} + \mathbf{d}^T\mathbf{x} \to \min, \\ A\mathbf{x} \leqslant \mathbf{b}, \\ \mathbf{x} \in \mathbb{R}^n, \end{cases}$$

gdzie H jest macierzą symetryczną dodatnio określoną, $\mathbf{d} \in \mathbb{R}^n$, A jest macierzą $m \times n$ i $\mathbf{b} \in \mathbb{R}^m$.

Definicja 10.4. Niech $\mathbb{X} \subset \mathbb{R}^n$. Transformatą Legendre'a-Fenchela funkcji $f : \mathbb{X} \to \mathbb{R}$ nazywamy funkcję $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ daną wzorem

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in \mathbb{X}} (\mathbf{y}^T\mathbf{x} - f(\mathbf{x})).$$

Ćwiczenie 10.12. Rozważmy problem optymalizacyjny:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ A\mathbf{x} \leqslant \mathbf{b}, \\ C\mathbf{x} = \mathbf{d}, \\ \mathbf{x} \in \mathbb{X}, \end{cases}$$

gdzie $\mathbb{X} \subset \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{d} \in \mathbb{R}^l$, zaś A, C są dowolnymi macierzami o odpowiednich wymiarach. Udowodnij, że problem do niego dualny ma następującą postać:

$$\begin{cases} -\mathbf{b}^T\mu - \mathbf{d}^T\lambda - f^*(-A^T\mu - C^T\lambda) \to \max, \\ \mu \in [0,\infty)^m, \quad \lambda \in \mathbb{R}^l. \end{cases}$$

Wskazówka. Rozbij ograniczenie równościowe na dwa ograniczenia nierównościowe.

Ćwiczenie 10.13. Znajdź transformatę Legendre'a-Fenchela następujących funkcji:

$$f(x) = \frac{1}{2}x^2, \qquad x \in \mathbb{X} = \mathbb{R},$$
$$f(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{n} x_i^2, \qquad \mathbf{x} \in \mathbb{X} = \mathbb{R}^n,$$
$$f(x) = e^x, \qquad x \in \mathbb{X} = \mathbb{R},$$
$$f(\mathbf{x}) = \|\mathbf{x}\|_p, \qquad \mathbf{x} \in \mathbb{X} = \mathbb{R}^n, \quad p > 1,$$
$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T H \mathbf{x}, \qquad \mathbf{x} \in \mathbb{X} = \mathbb{R}^n, \quad H \text{ - macierz symetryczna, nieosobliwa.}$$

Ćwiczenie 10.14. Udowodnij, że transformata Legendre'a-Fenchela f^* jest wypukła dla dowolnej funkcji f.

Ćwiczenie 10.15. Wykaż równoważność następujących dwóch zadań optymalizacyjnych:

$$\log\left(\sum_{i=1}^{m} e^{\mathbf{a}_i^T\mathbf{x}+b_i}\right) \to \min$$

oraz

$$\begin{cases} \log\left(\sum_{i=1}^{m} e^{y_i}\right) \to \min, \\ A\mathbf{x} + \mathbf{b} = \mathbf{y}, \\ \mathbf{x} \in \mathbb{R}^n, \quad \mathbf{y} \in \mathbb{R}^m, \end{cases} \tag{10.4}$$

gdzie przez $\mathbf{a}_1, \ldots, \mathbf{a}_m$ oznaczamy wiersze macierzy A. Udowodnij następnie, że zadaniem dualnym do (10.4) jest

$$\begin{cases} \mathbf{b}^T\nu - \sum_{i=1}^{m} \nu_i \log \nu_i \to \max, \\ \sum_{i=1}^{m} \nu_i = 1, \\ A^T\nu = \mathbf{0}, \\ \nu \in [0,\infty)^m. \end{cases}$$

Rozdział 11

Teoria wrażliwości

Dotychczas mnożniki Lagrange'a wydawały się techniczną sztuczką służącą do znajdowania rozwiązania problemu optymalizacyjnego z ograniczeniami. W tym rozdziale pokażemy, że pełnią one rolę kosztu związanego ze zmianą ograniczeń. Badanie wpływu zmiany ograniczeń nazywane jest *teorią wrażliwości*. Oddzielnie zajmiemy się ograniczeniami równościowymi i nierównościowymi.

11.1 Ograniczenia równościowe

Rozważmy problem optymalizacyjny z ograniczeniami równościowymi:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ h_j(\mathbf{x}) = 0, \quad j = 1, \dots, l, \\ \mathbf{x} \in \mathbb{X}, \end{cases} \tag{11.1}$$

gdzie $\mathbb{X} \subset \mathbb{R}^n$ jest zbiorem otwartym i $f, h_1, \dots, h_l : \mathbb{X} \to \mathbb{R}$. Dla uproszczenia notacji połóżmy $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), \dots, h_l(\mathbf{x})]^T$. Rozważmy problem zaburzony, tzn.

$$\begin{cases} f(\mathbf{x}) \to \min, \\ \mathbf{h}(\mathbf{x}) = \mathbf{z}, \\ \mathbf{x} \in \mathbb{X}, \end{cases} \tag{11.2}$$

gdzie $\mathbf{z} \in \mathbb{R}^l$.

Twierdzenie 11.1. *Niech $\bar{\mathbf{x}}$ będzie rozwiązaniem lokalnym problemu (11.1), zaś $\bar{\lambda}$ wektorem jego mnożników Lagrange'a. Załóżmy, że funkcje $f, h_1, \dots, h_l$ są klasy C^2 na otoczeniu $\bar{\mathbf{x}}$, gradienty ograniczeń są liniowo niezależne (spełniony jest warunek liniowej niezależności) oraz*

$$\mathbf{d}^T D^2_{\mathbf{x}} L(\bar{\mathbf{x}}, \bar{\lambda}) \mathbf{d} > 0 \tag{11.3}$$

dla $\mathbf{d} \in \mathbb{R}^n \setminus \mathbf{0}$ spełniających $Dh_j(\bar{\mathbf{x}})\mathbf{d} = 0$, $j = 1, \dots, l$. Wówczas istnieje otoczenie $\tilde{O}$ punktu $\mathbf{0} \in \mathbb{R}^l$ oraz funkcja $\mathbf{x} : \tilde{O} \to \mathbb{X}$ klasy C^1, taka że $\mathbf{x}(\mathbf{0}) = \bar{\mathbf{x}}$ oraz $\mathbf{x}(\mathbf{z})$ jest ścisłym rozwiązaniem lokalnym problemu zaburzonego (11.2). Ponadto,

$$D(f \circ \mathbf{x})(\mathbf{0}) = -\bar{\lambda}^T.$$

Dowód. Na mocy tw. 8.1 punkt $\bar{\mathbf{x}}$ rozwiązuje układ równań:

$$\begin{cases} D_{\mathbf{x}}L(\bar{\mathbf{x}}, \bar{\lambda}) = \mathbf{0}^T, \\ \mathbf{h}(\bar{\mathbf{x}}) = \mathbf{0}, \end{cases}$$

gdzie

$$D_{\mathbf{x}}L(\bar{\mathbf{x}}, \bar{\lambda}) = Df(\bar{\mathbf{x}}) + \bar{\lambda}^T D\mathbf{h}(\bar{\mathbf{x}}) = Df(\bar{\mathbf{x}}) + \sum_{j=1}^{l} \bar{\lambda}_j Dh_j(\bar{\mathbf{x}}).$$

Zaburzając prawą stronę drugiej równości przez $\mathbf{z}$ będziemy chcieli pokazać, że istnieje rozwiązanie, które jest funkcją klasy C^1 zaburzenia $\mathbf{z}$. Rozważmy więc układ:

$$D_{\mathbf{x}}L(\mathbf{x}, \lambda) = \mathbf{0}^T, \qquad \mathbf{h}(\mathbf{x}) = \mathbf{z},$$

gdzie niewiadomymi są λ oraz $\mathbf{x}$. Zdefiniujmy funkcję $G : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^l \to \mathbb{R}^n \times \mathbb{R}^l$ wzorem

$$G(\mathbf{x}, \lambda, \mathbf{z}) = \begin{bmatrix} (D_{\mathbf{x}}L(\mathbf{x}, \lambda))^T \\ \mathbf{h}(\mathbf{x}) - \mathbf{z} \end{bmatrix}.$$

Zaburzony układ możemy wówczas zapisać jako

$$G(\mathbf{x}, \lambda, \mathbf{z}) = \mathbf{0}.$$

Wiemy, że $G(\bar{\mathbf{x}}, \bar{\lambda}, \mathbf{0}) = \mathbf{0}$. Skorzystamy z twierdzenia o funkcji uwikłanej, aby rozwikłać pierwsze dwie zmienne w zależności od trzeciej. W tym celu rozważamy macierz pochodnych $DG(\bar{\mathbf{x}}, \bar{\lambda}, \mathbf{0})$:

$$DG(\bar{\mathbf{x}}, \bar{\lambda}, \mathbf{0}) = \begin{bmatrix} D^2_{\mathbf{x}}L(\bar{\mathbf{x}}, \bar{\lambda}), & (D\mathbf{h}(\bar{\mathbf{x}}))^T, & \mathbf{0} \\ D\mathbf{h}(\bar{\mathbf{x}}), & \mathbf{0}, & -I \end{bmatrix}.$$

Warunek liniowej niezależności gradientów ograniczeń implikuje nieosobliwość podmacierzy (patrz zadanie 11.1)

$$\begin{bmatrix} D^2_{\mathbf{x}}L(\bar{\mathbf{x}}, \bar{\lambda}), & (D\mathbf{h}(\bar{\mathbf{x}}))^T \\ D\mathbf{h}(\bar{\mathbf{x}}), & \mathbf{0} \end{bmatrix}.$$

Spełnione są zatem założenia twierdzenia 8.4 i istnieje otoczenie O punktu $\mathbf{0} \in \mathbb{R}^l$ oraz funkcje $\mathbf{x} : O \to \mathbb{X}$ i $\lambda : O \to \mathbb{R}^l$ klasy C^1, takie że dla $\mathbf{z} \in O$ zachodzi $G(\mathbf{x}(\mathbf{z}), \lambda(\mathbf{z}), \mathbf{z}) = \mathbf{0}$, czyli

$$D_{\mathbf{x}}L(\mathbf{x}(\mathbf{z}), \lambda(\mathbf{z})) = \mathbf{0}^T, \qquad \mathbf{h}(\mathbf{x}(\mathbf{z})) = \mathbf{z}.$$

Korzystając z faktu, iż funkcje $D^2_{\mathbf{x}}L, D\mathbf{h}, \mathbf{x}, \lambda$ są ciągłe oraz nierówność (11.3) spełniona jest dla niezaburzonego problemu, wnioskujemy, że istnieje być może mniejsze otoczenie $\tilde{O}$ punktu $\mathbf{0} \in \mathbb{R}^l$, takie że

$$\mathbf{d}^T D^2_{\mathbf{x}}L(\mathbf{x}(\mathbf{z}), \lambda(\mathbf{z}))\mathbf{d} > 0$$

dla $\mathbf{z} \in \tilde{O}$ i $\mathbf{d} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ spełniających $Dh_j(\mathbf{x}(\mathbf{z}))\mathbf{d} = 0$, $j = 1, \ldots, l$. Kluczowa dla tego wyniku jest ostra nierówność w warunku (11.3). Na mocy tw. 9.3 punkt $\mathbf{x}(\mathbf{z})$ jest zatem ścisłym rozwiązaniem lokalnym problemu zaburzonego (11.2). Przypomnijmy, że funkcja $\mathbf{x}$ jest klasy C^1. Możemy zatem zdefiniować pochodną złożenia

$$D(f \circ \mathbf{x})(\mathbf{0}) = Df(\bar{\mathbf{x}})D\mathbf{x}(\mathbf{0}).$$

Od zakończenia dowodu dzielą nas dwa spostrzeżenia. Po pierwsze, mnożąc obie strony warunku koniecznego pierwszego rzędu dla problemu niezaburzonego

$$Df(\bar{\mathbf{x}}) + \bar{\lambda}^T D\mathbf{h}(\bar{\mathbf{x}}) = \mathbf{0}^T$$

przez $D\mathbf{x}(\mathbf{0})$ dostajemy

$$Df(\bar{\mathbf{x}})D\mathbf{x}(\mathbf{0}) + \bar{\lambda}^T D\mathbf{h}(\bar{\mathbf{x}})D\mathbf{x}(\mathbf{0}) = \mathbf{0}^T.$$

Po drugie, różniczkując $\mathbf{h}(\mathbf{x}(\mathbf{z})) = \mathbf{z}$ po $\mathbf{z}$ otrzymujemy w punkcie $\mathbf{z} = \mathbf{0}$ następującą pochodną $D(\mathbf{h} \circ \mathbf{x})(\mathbf{0}) = I$, czyli $D\mathbf{h}(\bar{\mathbf{x}})D\mathbf{x}(\mathbf{0}) = I$. Upraszcza to powyższe równanie do

$$Df(\bar{\mathbf{x}})D\mathbf{x}(\mathbf{0}) + \bar{\lambda}^T = \mathbf{0}^T.$$

Stąd już teza wynika natychmiast. □

Twierdzenie 11.1 można rozumieć następująco: nieznaczna zmiana j-tego ograniczenia z zera do ε prowadzi do zmiany optymalnej wartości funkcji f o $-\bar{\lambda}_j\varepsilon + o(\varepsilon)$, tzn. dla małych ε zmiana ta jest w przybliżeniu równa $-\bar{\lambda}_j\varepsilon$.

11.2 Ograniczenia nierównościowe

W przypadku ograniczeń nierównościowych zastosujemy inne podejście. Skoncentrujemy się na zadaniu optymalizacji wypukłej:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ g_i(\mathbf{x}) \leqslant 0, \quad i = 1, \dots, m, \\ \mathbf{x} \in \mathbb{X}, \end{cases} \tag{11.4}$$

gdzie $\mathbb{X} \subset \mathbb{R}^n$ jest wypukły, $f, g_1, \dots, g_m : \mathbb{X} \to \mathbb{R}$ są wypukłe. Dla uproszczenia notacji będziemy pisać $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), \dots, g_m(\mathbf{x})]^T$. Problem (11.4) zapisujemy więc jako

$$\begin{cases} f(\mathbf{x}) \to \min, \\ \mathbf{g}(\mathbf{x}) \leqslant \mathbf{0}, \\ \mathbf{x} \in \mathbb{X}. \end{cases} \tag{11.5}$$

Rozważmy zadanie zaburzone: dla $\mathbf{z} \in \mathbb{R}^m$

$$\begin{cases} f(\mathbf{x}) \to \min, \\ \mathbf{g}(\mathbf{x}) \leqslant \mathbf{z}, \\ \mathbf{x} \in \mathbb{X}. \end{cases} \tag{11.6}$$

Definicja 11.1. Niech D_M oznacza zbiór takich $\mathbf{z} \in \mathbb{R}^m$, dla których *zbiór punktów dopuszczalnych* $W_{\mathbf{z}} = \{\mathbf{x} \in \mathbb{X} : \ \mathbf{g}(\mathbf{x}) \leqslant \mathbf{z}\}$ jest niepusty. Funkcję

$$M(\mathbf{z}) = \inf_{\mathbf{x}\in\mathbb{X},\ \mathbf{g}(\mathbf{x})\leqslant\mathbf{z}} f(\mathbf{x})$$

zdefiniowaną dla $\mathbf{z} \in D_M$ nazywamy *funkcją perturbacji* a zbiór D_M *dziedziną* funkcji perturbacji.

Zauważmy, że $M(\mathbf{z}) < \infty$ dla $\mathbf{z} \in D_M$, ale może się zdarzyć, że $M(\mathbf{z}) = -\infty$.

Na rysunku 10.3 wykresem funkcji perturbacji jest krzywa $M(y)$ (z rysunku widać, że jest to wykres funkcji wypukłej). Zauważmy, że funkcja perturbacji między punktami A i B jest dobrze określona, bo wtedy $\exists_{\mathbf{x}\in\mathbb{X}}(g(\mathbf{x}), f(\mathbf{x})) \in G$, co oznacza, ze $y = g(\mathbf{x})$ należy do dziedziny funkcji $M(y)$. Jeśli $A = (y_A, z_A)$, to dla $y < y_A$ zbiór punktów dopuszczalnych jest pusty i punkty $y < y_A$ nie należą do dziedziny funkcji perturbacji. Na prawo od punktu B funkcja perturbacji jest stała. Jeśli $B = (y_B, z_B)$, to $\min\{f(\mathbf{x}) : g(\mathbf{x}) \leqslant y_B\} = \min\{f(\mathbf{x}) : g(\mathbf{x}) \leqslant y\}$, dla $y > y_B$.

Twierdzenie 11.2.

1. *Zbiór D_M jest wypukły.*
2. *Funkcja $M : D_M \to \mathbb{R} \cup \{-\infty\}$ jest wypukła.*
3. *Jeśli istnieje punkt $\mathbf{x}^* \in \mathbb{X}$, taki że $g_i(\mathbf{x}^*) < 0$, $i = 1, \ldots, m$, to $\operatorname{int} D_M \neq \emptyset$ i $\mathbf{0} \in \operatorname{int} D_M$.*

Dowód. Z wypukłości każdej składowej wektora $\mathbf{g}$ wynika następująca implikacja (zapis nierówności dla wektorów oznacza, że zachodzi ona dla wszystkich odpowiednich współrzędnych wektorów):

$$\mathbf{g}(\mathbf{x}^1) \leqslant \mathbf{z}^1, \quad \mathbf{g}(\mathbf{x}^2) \leqslant \mathbf{z}^2 \quad \Longrightarrow \quad \forall\, \lambda \in [0,1] \quad \mathbf{g}(\lambda\mathbf{x}^1 + (1-\lambda)\mathbf{x}^2) \leqslant \lambda\mathbf{z}^1 + (1-\lambda)\mathbf{z}^2. \tag{11.7}$$

Będziemy korzystać z tego spostrzeżenia wielokrotnie w dowodzie.
(1) Niech $\mathbf{z}^1, \mathbf{z}^2 \in D_M$ i $\lambda \in (0,1)$. Wówczas istnieją $\mathbf{x}^1, \mathbf{x}^2 \in \mathbb{X}$ takie że $\mathbf{g}(\mathbf{x}^1) \leqslant \mathbf{z}^1$ i $\mathbf{g}(\mathbf{x}^2) \leqslant \mathbf{z}^2$. Z wzoru (11.7) dostajemy $\mathbf{g}(\lambda\mathbf{x}^1+(1-\lambda)\mathbf{x}^2) \leqslant \lambda\mathbf{z}^1+(1-\lambda)\mathbf{z}^2$, skąd wynika $\lambda\mathbf{z}^1+(1-\lambda)\mathbf{z}^2 \in D_M$.
(2) Niech $\mathbf{z}^1, \mathbf{z}^2 \in D_M$ i $\lambda \in (0,1)$. Wówczas

$$\begin{aligned}
\lambda M(\mathbf{z}^1) + (1-\lambda)M(\mathbf{z}^2) &= \inf_{\mathbf{x}^1 \in \mathbb{X},\ \mathbf{g}(\mathbf{x}^1) \leqslant \mathbf{z}^1} (\lambda f(\mathbf{x}^1)) + \inf_{\mathbf{x}^2 \in \mathbb{X},\ \mathbf{g}(\mathbf{x}^2) \leqslant \mathbf{z}^2} ((1-\lambda) f(\mathbf{x}^2)) \\
&= \inf_{\mathbf{x}^1 \in \mathbb{X},\ \mathbf{g}(\mathbf{x}^1) \leqslant \mathbf{z}^1,\ \mathbf{x}^2 \in \mathbb{X},\ \mathbf{g}(\mathbf{x}^2) \leqslant \mathbf{z}^2} (\lambda f(\mathbf{x}^1) + (1-\lambda) f(\mathbf{x}^2)) \\
&\geqslant \inf_{\mathbf{x}^1 \in \mathbb{X},\ \mathbf{g}(\mathbf{x}^1) \leqslant \mathbf{z}^1,\ \mathbf{x}^2 \in \mathbb{X},\ \mathbf{g}(\mathbf{x}^2) \leqslant \mathbf{z}^2} f(\lambda\mathbf{x}^1 + (1-\lambda)\mathbf{x}^2) \\
&\geqslant \inf_{\mathbf{x} \in \mathbb{X},\ \mathbf{g}(\mathbf{x}) \leqslant \lambda\mathbf{z}^1 + (1-\lambda)\mathbf{z}^2} f(\mathbf{x}),
\end{aligned}$$

gdzie pierwsza nierówność wynika z wypukłości f, zaś ostatnia – z własności (11.7):

$$\{\lambda\mathbf{x}^1+(1-\lambda)\mathbf{x}^2 :\ \mathbf{x}^1 \in \mathbb{X},\ \mathbf{g}(\mathbf{x}^1) \leqslant \mathbf{z}^1,\ \mathbf{x}^2 \in \mathbb{X},\ \mathbf{g}(\mathbf{x}^2) \leqslant \mathbf{z}^2\} \subseteq \{\mathbf{x} \in \mathbb{X} :\ \mathbf{g}(\mathbf{x}) \leqslant \lambda\mathbf{z}^1+(1-\lambda)\mathbf{z}^2\}.$$

(3) Musimy udowodnić, że zbiór dopuszczalny $W_{\mathbf{z}}$ jest niepusty dla $\mathbf{z}$ z pewnego otoczenia $\mathbf{0} \in \mathbb{R}^m$. Wiemy, że istnieje punkt $\mathbf{x}^* \in \mathbb{X}$, taki że $g_i(\mathbf{x}^*) < 0$, $i = 1, \ldots, m$. Weźmy $a = \min\{-g_i(\mathbf{x}^*) :\ i = 1, \ldots, m\}$. Wówczas dla każdego $\mathbf{z} \in [-a, a]^m$ mamy $\mathbf{x}^* \in W_{\mathbf{z}}$. Zatem $[-a, a]^m \subset D_M$. □

Uwaga 11.1.

1. Jeśli $M(\bar{\mathbf{z}}) = -\infty$ dla pewnego $\bar{\mathbf{z}} \in D_M$, to z wypukłości M mamy $\lambda\bar{\mathbf{z}} + (1-\lambda)\mathbf{z} = -\infty$ dla dowolnego $\mathbf{z} \in D_M$ i $\lambda \in (0,1)$.
2. Jeśli $M(\bar{\mathbf{z}}) = -\infty$ dla pewnego $\bar{\mathbf{z}} \in D_M$, to $M(\mathbf{z}) = -\infty$ dla $\mathbf{z} \in \operatorname{int} D_M$. Wynika to wprost z powyższej uwagi.
3. Jeśli istnieje $\bar{\mathbf{z}} \in \operatorname{int} D_M$ taki że $M(\bar{\mathbf{z}}) > -\infty$, to $M(\mathbf{z}) > -\infty$ dla każdego $\mathbf{z} \in D_M$. Inaczej mielibyśmy sprzeczność z punktem (2).

Twierdzenie 11.3. *Jeśli w problemie optymalizacji wypukłej istnieje punkt $\mathbf{x}^* \in \mathbb{X}$, taki że $g_i(\mathbf{x}^*) < 0$, $i = 1, \ldots, m$, oraz $M(\mathbf{0}) > -\infty$, to $M(\mathbf{z}) > -\infty$ dla każdego $\mathbf{z} \in D_M$ oraz istnieje wektor $\mu \in [0, \infty)^m$ wyznaczający hiperpłaszczyznę podpierającą M:*

$$M(\mathbf{z}) \geqslant M(\mathbf{0}) - \mu^T\mathbf{z}, \qquad \mathbf{z} \in D_M.$$

Dowód. Z tw. 11.2 wynika, że M jest funkcją wypukłą i $\mathbf{0} \in \operatorname{int} D_M$. Zatem na mocy ostatniej z powyższych uwag mamy $M(\mathbf{z}) > -\infty$ dla $\mathbf{z} \in D_M$. Istnienie płaszczyzny podpierającej wynika z tw. 3.8:

$$M(\mathbf{z}) \geqslant M(\mathbf{0}) - \mu^T \mathbf{z}, \qquad \mathbf{z} \in D_M,$$

dla pewnego $\mu \in \mathbb{R}^m$. Udowodnimy teraz, że wszystkie współrzędne μ muszą być nieujemne. Przypuśćmy przeciwnie, tzn. $\mu_i < 0$ dla pewnego $i \in \{1, \dots, m\}$. Ponieważ $\mathbf{0} \in \operatorname{int} D_M$, to dla dostatecznie małego $a > 0$ punkt $\bar{\mathbf{z}} = [0, \dots, 0, a, 0, \dots, 0]^T$, gdzie a jest na i-tej pozycji, należy do D_M. Korzystając z ujemności μ_i mamy

$$M(\bar{\mathbf{z}}) \geqslant M(\mathbf{0}) - \mu_i a > M(\mathbf{0}).$$

Z drugiej strony $W_{\mathbf{0}} \subseteq W_{\bar{\mathbf{z}}}$ (bo $\bar{\mathbf{z}} \geqslant 0$), czyli $M(\bar{\mathbf{z}}) \leqslant M(\mathbf{0})$. To daje sprzeczność, czyli dowiedliśmy, że $\mu \in [0, \infty)^m$. □

Wektor μ nazywamy *wektorem wrażliwości* dla problemu (11.4). Z tw. 3.8 wynika, że jeśli funkcja M jest różniczkowalna w punkcie $\mathbf{0}$, to $\mu = -\big(DM(\mathbf{0})\big)^T$. Zatem μ oznacza szybkość i kierunek zmian wartości minimalnej f przy zmianie ograniczeń, podobnie jak w przypadku ograniczeń równościowych omawianych wcześniej w tym rozdziale.

Zbadajmy teraz relacje pomiędzy wektorem wrażliwości a punktem siodłowym i warunkiem pierwszego rzędu. Zauważmy, że powiązanie punktu siodłowego z wektorem wrażliwości nie wymaga wypukłości problemu optymalizacyjnego.

Twierdzenie 11.4.

1. *Jeśli $(\bar{\mathbf{x}}, \bar{\mu})$ jest punktem siodłowym funkcji Lagrange'a na zbiorze $\mathbb{X} \times [0, \infty)^m$, to $\bar{\mu}$ jest wektorem wrażliwości (tzn. wyznacza płaszczyznę podpierającą). Teza ta nie wymaga założenia o wypukłości problemu optymalizacyjnego.*
2. *Załóżmy, że funkcje $f, g_1, \dots, g_m$ są różniczkowalne w $\bar{\mathbf{x}}$, i wypukłe. Jeśli w $\bar{\mathbf{x}}$ spełniony jest warunek pierwszego rzędu z mnożnikami Lagrange'a $\bar{\mu} \in [0, \infty)^m$, to $\bar{\mu}$ jest wektorem wrażliwości.*

Dowód. (1) Oznaczmy przez $L_{\mathbf{z}}(\mathbf{x}, \mu)$ funkcję Lagrange'a dla problemu zaburzonego. Wówczas

$$L_{\mathbf{z}}(\mathbf{x}, \mu) = f(\mathbf{x}) + \sum_{i=1}^{m} \mu_i (g_i(\mathbf{x}) - z_i) = L(\mathbf{x}, \mu) - \mu^T \mathbf{z}.$$

Z faktu, że $(\bar{\mathbf{x}}, \bar{\mu})$ jest punktem siodłowym wynika, że

$$M(\mathbf{0}) = L(\bar{\mathbf{x}}, \bar{\mu}) = \inf_{\mathbf{x} \in \mathbb{X}} L(\mathbf{x}, \bar{\mu}).$$

Zatem

$$M(\mathbf{0}) = \inf_{\mathbf{x} \in \mathbb{X}} L(\mathbf{x}, \bar{\mu}) = \inf_{\mathbf{x} \in \mathbb{X}} \big(L_{\mathbf{z}}(\mathbf{x}, \bar{\mu}) + \bar{\mu}^T \mathbf{z}\big) = \inf_{\mathbf{x} \in \mathbb{X}} L_{\mathbf{z}}(\mathbf{x}, \bar{\mu}) + \bar{\mu}^T \mathbf{z}. \tag{11.8}$$

Zauważmy, że dla dowolnego $\mathbf{x} \in W_{\mathbf{z}}$ i $\mu \in [0, \infty)^m$ mamy $f(\mathbf{x}) \geqslant L_{\mathbf{z}}(\mathbf{x}, \mu)$, czyli, w szczególności,

$$M(\mathbf{z}) = \inf_{\mathbf{x} \in W_{\mathbf{z}}} f(\mathbf{x}) \geqslant \inf_{\mathbf{x} \in W_{\mathbf{z}}} L_{\mathbf{z}}(\mathbf{x}, \bar{\mu}) \geqslant \inf_{\mathbf{x} \in \mathbb{X}} L_{\mathbf{z}}(\mathbf{x}, \bar{\mu}).$$

Wstawiając tą zależność do (11.8) otrzymujemy

$$M(\mathbf{0}) \leqslant M(\mathbf{z}) + \bar{\mu}^T \mathbf{z}.$$

(2) Z lematu 10.1 wynika, iż punkt $(\bar{\mathbf{x}}, \bar{\mu})$ jest punktem siodłowym funkcji Lagrange'a. Tezę dostajemy z pierwszej części niniejszego twierdzenia. □

11.3 Zadania

Ćwiczenie 11.1. Udowodnij, że przy założeniach tw. 11.1 macierz

$$\begin{bmatrix} D^2_{\mathbf{x}} L(\bar{\mathbf{x}}, \bar{\lambda}), & (D\mathbf{h}(\bar{\mathbf{x}}))^T \\ D\mathbf{h}(\bar{\mathbf{x}}), & \mathbf{0} \end{bmatrix}$$

jest nieosobliwa.

Ćwiczenie 11.2. Dla problemu

$$\begin{cases} x_1^2 + 2x_2^2 \to \min, \\ x_1^2 + x_2^2 \leqslant 0, \\ x_1 \leqslant 0. \end{cases}$$

1. Znaleźć D_M.
2. Znaleźć wektor wrażliwości.
3. Znaleźć funkcję perturbacji.

Wskazówka. Umieszczenie ograniczenia $x_1^2 + x_2^2 \leqslant 0$ było intencją autora zadania.

Ćwiczenie 11.3. Znajdź funkcję perturbacji i wektor wrażliwości dla problemu

$$\begin{cases} x_1^2 + x_2^2 \to \min, \\ x_1 + x_2 \leqslant 0. \end{cases}$$

Ćwiczenie 11.4. Załóżmy, że w zadaniu optymalizacyjnym (11.4) funkcje f i g_i są klasy C^1 oraz $\mathbb{X} = \mathbb{R}^n$. Czy funkcja perturbacji musi być wówczas klasy C^1? Udowodnij lub podaj kontrprzykład.

Ćwiczenie 11.5. Załóżmy, że w zadaniu optymalizacyjnym (11.4) funkcje f i g_i są ciągłe oraz $\mathbb{X} = \mathbb{R}^n$. Czy funkcja perturbacji musi być wówczas ciągła? Udowodnij lub podaj kontrprzykład.

Ćwiczenie 11.6. Rozważmy problem producenta. Dysponuje on budżetem $b > 0$, który może spożytkować na wytworzenie dwóch rodzajów towarów. Pierwszy z towarów sprzedaje po cenie $p_1 > 0$, zaś drugi – po cenie $p_2 > 0$. Cena produkcji opisana jest funkcją $c(x_1, x_2)$, gdzie wektor $\mathbf{x}$ opisuje wielkość produkcji każdego z towarów. Celem producenta jest maksymalizacja przychodów ze sprzedaży bez przekroczenia budżetu produkcyjnego:

$$\begin{cases} p_1 x_1 + p_2 x_2 \to \max, \\ c(x_1, x_2) \leqslant b. \end{cases}$$

1. Podaj warunki konieczne istnienia rozwiązania powyższego problemu.
2. Załóżmy, że c jest funkcją wypukłą. Jak należy zmodyfikować wielkość produkcji, jeśli budżet produkcyjny b wrośnie nieznacznie?

Ćwiczenie 11.7. Rozważmy funkcję $f : [0, \infty) \to \mathbb{R}$ zadaną wzorem

$$f(t) = \min_{x,y \in \mathbb{R}} \{e^{x^2 - y} + y^2 - x : \ x^2 + x^4 - 2xy + 3y^2 \leqslant t\}.$$

Uzasadnij, że funkcja f jest dobrze określona i wypukła.

Rozdział 12

Wprowadzenie do numerycznych metod optymalizacji

W poprzednich rozdziałach budowaliśmy teorię służącą rozwiązywaniu problemów optymalizacyjnych. By być bardziej precyzyjnym, dowiedliśmy twierdzeń, które *ułatwiają* znalezienie kandydatów na rozwiązania oraz stwierdzenie, czy któryś z nich jest rozwiązaniem. Niestety w obu tych krokach musimy rozwiązywać układy równań nieliniowych. W praktycznych zadaniach optymalizacyjnych układy te mogą nie mieć "ładnych" rozwiązań lub po prostu nie jesteśmy w stanie ich znaleźć w sposób analityczny. Pozostaje wówczas zastosowanie metod numerycznych, które pozwolą na *przybliżenie* rozwiązania. W tym i kolejnych rozdziałach opiszemy kilka takich algorytmów numerycznych. Algorytmy te będą raczej atakowały problem optymalizacyjny bezpośrednio (tzn. nie będziemy się starali znaleźć zbioru punktów spełniających warunek konieczny pierwszego rzędu). Będą one krok po kroku tworzyły ciąg punktów zbiegający do rozwiązania. Nie oznacza to jednak, że mnożniki Lagrange'a i metody dualne są nieprzydatne przy tworzeniu algorytmów numerycznych. Wręcz przeciwnie. W przypadku ograniczeń równościowych klasyczne podejście wiedzie właśnie poprzez mnożniki Lagrange'a (tzw. metoda mnożników Lagrange'a), patrz monografie Bazaraa, Sherali, Shetty [3] oraz Bertsekas [4, 5]. My jednak pójdziemy inną drogą, gdyż możemy poświęcić optymalizacji z ograniczeniami zaledwie jeden rozdział. Na kolejnych stronach będziemy opisywać algorytmy optymalizacyjne oraz analizować ich działanie.

Rozpoczniemy od ogólnych definicji.

Definicja 12.1. *Procesem iteracyjnym* nazywamy czwórkę (Q, I, Ω, h), gdzie Q jest pewnym zbiorem, $h : Q \to Q$ odwzorowaniem w tym zbiorze, $I \subset Q$, $\Omega \subset Q$, przy czym odwzorowanie h jest identycznością na Ω. Ta czwórka reprezentuje proces obliczeniowy: I jest zbiorem danych początkowych, Ω jest zbiorem wyników a h opisuje proces prowadzenia obliczeń, tj. startując ze stanu początkowego $x \in I$ generuje ciąg

$$x_1 = x, \quad x_{k+1} = h(x_k), \quad k = 1, 2, \ldots$$

Mówimy, że proces iteracyjny kończy się w m krokach, jeśli $x_m \in \Omega$ (zgodnie z definicja odwzorowania h, jeśli $x_m \in \Omega$, to x_{m+1} też jest w Ω). *Algorytmem* nazywamy proces iteracyjny, który kończy się w skończonej liczbie kroków.

W zadaniach optymalizacyjnych interesować nas będzie stosowanie algorytmów numerycznych do zadań, które posiadają rozwiązania. Algorytmy dla takich problemów powinny spełniać następujące warunki:

1. Poprawność algorytmu, czyli warunek, że dla każdych dopuszczalnych danych początkowych $x \in I$ dostaniemy poprawną odpowiedź. Dla naszych problemów będziemy ten warunek formułowali jako warunek zbieżności do rozwiązania.
2. Warunek końca, czyli kryterium rozstrzygające, kiedy kolejne przybliżenie jest już w zbiorze wyników Ω, tzn. kiedy możemy przerwać wykonywanie algorytmu, ponieważ znaleziony punkt jest dostatecznie blisko rozwiązania.
3. Efektywność algorytmu. W naszych problemach będzie to pytanie o szybkość zbieżności algorytmu do rozwiązania.

W tym rozdziale opiszemy ogólne własności metod optymalizacyjnych. Będą one wykorzystane przy implementacji algorytmów optymalizacyjnych, którymi zajmiemy się w roz. 13 dla zadania bez ograniczeń oraz w roz. 14 dla zadania z ograniczeniami.

12.1 Własności algorytmów optymalizacyjnych

Zanim przejdziemy do dyskusji właściwych algorytmów, sformułujmy ściśle problem optymalizacyjny:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ \mathbf{x} \in \mathbb{X} \subset \mathbb{R}^n. \end{cases}$$

Uwaga 12.1. W zadaniach optymalizacyjnych do poprawnego działania algorytmu potrzebne są zwykle procedury pozwalające obliczać w punktach $\mathbf{x}_k$ wartości funkcji $f(\mathbf{x}_k)$ oraz jej pochodnych $Df(\mathbf{x}_k)$, $D^2f(\mathbf{x}_k)$, itd. Obliczanie tych wartości nie jest częścią algorytmu. Zakłada się zwykle, że są to zewnętrzne procedury dostarczające wymienione wartości z dowolną wymaganą dokładnością (mówi się o nich, że są dostarczane przez "wyrocznię" *(oracle)*).

Obecnie zdefiniujemy jakie mogą być warunki zatrzymania algorytmu (warunki końca).

Definicja 12.2. Niech $\mathbf{x}^*$ będzie rozwiązaniem zadania optymalizacyjnego a $f^* = f(\mathbf{x}^*)$. Kryteria zatrzymania algorytmu przy warunku bezwzględnej tolerancji optymalności przy poziomie tolerancji $\varepsilon > 0$:

i) $|f(\mathbf{x}_k) - f^*| \leqslant \varepsilon$;

ii) $\|\mathbf{x}_k - \mathbf{x}^*\| \leqslant \varepsilon$;

iii) $\|Df(\mathbf{x}_k)\| \leqslant \varepsilon$;

iv) $|f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)| \leqslant \varepsilon$;

v) $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leqslant \varepsilon$.

Analogicznie można sformułować kryteria zatrzymania dla względnej tolerancji optymalności:

i) $|f(\mathbf{x}_k) - f^*|/|f^*| \leqslant \varepsilon$;

ii) $\|\mathbf{x}_k - \mathbf{x}^*\|/\|\mathbf{x}^*\| \leqslant \varepsilon$;

iii) $|f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)|/|f(\mathbf{x}_k)| \leqslant \varepsilon$;

iv) $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|/\|\mathbf{x}_k\| \leqslant \varepsilon$.

Zauważmy, że podane w obu powyższych przypadkach kryteria i) oraz ii), mimo że są najbardziej poprawnymi kryteriami zatrzymania, mają małą przydatność praktyczną (nie znamy przecież punktu $\mathbf{x}^*$ a tym samym także wartości f^*). W praktyce obliczeniowej stosujemy raczej pozostałe kryteria, chociaż nie gwarantują one zatrzymania algorytmu dostatecznie blisko rzeczywistego rozwiązania problemu.

Na koniec zajmijmy się problemem szybkości zbieżności algorytmu.

Definicja 12.3. Jeśli

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leqslant c\|\mathbf{x}_k - \mathbf{x}^*\|^p,$$

gdzie p jest największą liczbą, przy której $c > 0$, a oszacowanie zachodzi dla każdego k większego od pewnego $k_0 > 1$, to mówimy, że algorytm posiada *rząd zbieżności* p.

Uwaga 12.2. Z tej definicji także trudno skorzystać w praktyce, bo nie znamy $\mathbf{x}^*$. Zwykle rząd zbieżności wyznacza się z następującego kryterium przybliżonego

$$\limsup_k \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|}{\|\mathbf{x}_k - \mathbf{x}_{k-1}\|^p} = c > 0,$$

jeśli taka granica istnieje (choćby na podciągu).

12.2 Optymalizacja bez użycia pochodnych

Zanim przejdziemy do dyskusji właściwego algorytmu, wprowadzimy pojęcie funkcji ściśle quasi-wypukłej i jej własności.

Definicja 12.4. Niech $W \subset \mathbb{R}^n$ będzie zbiorem wypukłym. Funkcję $f : W \to \mathbb{R}$ nazywamy *ściśle quasi-wypukłą*, jeśli dla dowolnych $\mathbf{x}, \mathbf{y} \in W$, $\mathbf{x} \neq \mathbf{y}$ oraz $\lambda \in (0, 1)$ zachodzi

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) < \max\left(f(\mathbf{x}), f(\mathbf{y})\right).$$

Dowód poniższych własności funkcji quasi-wypukłych pozostawiamy jako ćwiczenie.

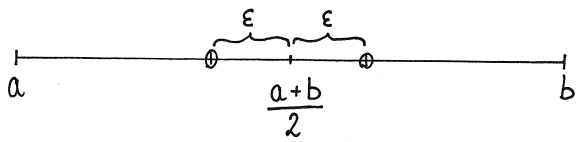
Lemat 12.1.

1. *Funkcja ściśle quasi-wypukła ma co najwyżej jedno minimum (lokalne i globalne).*
2. *Funkcja ściśle wypukła jest ściśle quasi-wypukła.*
3. *Funkcja określona na prostej lub* otwartym *przedziale jest ściśle quasi-wypukła jeśli jest ściśle malejąca, ściśle rosnąca lub istnieje punkt $\bar{x}$ w jej dziedzinie o tej własności, że funkcja jest ściśle malejąca dla $x \leqslant \bar{x}$ i ściśle rosnąca dla $x \geqslant \bar{x}$.*

Okazuje się, że własność ścisłej quasi-wypukłości pozwala na skonstruowanie algorytmu znajdowania minimum funkcji na domkniętym odcinku bez wykorzystania pochodnych. Główna obserwacja znajduje się w poniższym lemacie:

Lemat 12.2. *Niech $f : [a, b] \to \mathbb{R}$ będzie funkcją ściśle quasi-wypukłą i $a \leqslant x \leqslant y \leqslant b$.*

1. *Jeśli $f(x) \geqslant f(y)$, to $f(z) > f(y)$ dla każdego $z \in [a, x)$.*
2. *Jeśli $f(x) \leqslant f(y)$, to $f(z) > f(x)$ dla każdego $z \in (y, b]$.*



Rysunek 12.1: Podział odcinka.

Dowód. Udowodnimy punkt (1) przez sprzeczność. Przypuśćmy, że istnieje $z \in [a, x)$ o tej własności, że $f(z) \leqslant f(y)$. Wówczas z quasi-wypukłości f wynika, że $f(x) < \max\big(f(z), f(y)\big) = f(y)$, co przeczy założeniu, że $f(x) \geqslant f(y)$. Dowód punktu (2) jest analogiczny. □

Bazując na lemacie 12.2 można skonstruować wiele algorytmów minimalizacji funkcji ściśle quasi-wypukłej. Rozpoczniemy od opisu algorytmu *podziału dychotomicznego*. Idea algorytmu jest bardzo prosta. Szukając minimum funkcji ściśle quasi-wypukłej $f : [a, b] \to \mathbb{R}$ wybieramy we wnętrzu przedziału $[a, b]$ dwa punkty λ i μ, takie że $\lambda < \mu$. Korzystając z lematu 12.2 zauważamy, że jeśli $f(\lambda) < f(\mu)$, to przedział w którym funkcja f ma minimum można ograniczyć do $[a, \mu]$, a kiedy $f(\lambda) > f(\mu)$, to do przedziału $[\lambda, b]$. Optymalna strategia polega oczywiście na wybraniu tak punktów λ i μ, aby otrzymany przedział był jak najmniejszy. Ponieważ *a priori* nie wiemy, w którym punkcie wartość funkcji f jest większa, powinniśmy uwzględnić najgorszy przypadek, czyli zminimalizować maksimum z dwóch liczb $(\mu - a)$ oraz $(b - \lambda)$. To minimum jest osiągane dla $\mu = \lambda = \frac{a+b}{2}$. Ponieważ takie rozwiązanie nie daje dwóch różnych punktów, dlatego wybieramy mały $\varepsilon > 0$ i porównujemy wartości funkcji w punktach położonych symetrycznie względem środka przedziału $\frac{a+b}{2} - \varepsilon$ i $\frac{a+b}{2} + \varepsilon$ (patrz rys. 12.1). Jeśli $f\big(\frac{a+b}{2} - \varepsilon\big) \geqslant f\big(\frac{a+b}{2} + \varepsilon\big)$, to minimum znajduje się w przedziale $\big[\frac{a+b}{2} - \varepsilon, b\big]$. W przeciwnym przypadku jest ono w przedziale $\big[a, \frac{a+b}{2} + \varepsilon\big]$. Algorytm wygląda następująco:

Inicjalizacja: Wybierz $0 < \varepsilon_1 < \frac{a+b}{2}$. Połóż $x_1 = a$, $y_1 = b$.

Krok k-ty:

1. Obliczamy $\lambda_k = \frac{x_k + y_k}{2} - \varepsilon_k$ oraz $\mu_k = \frac{x_k + y_k}{2} + \varepsilon_k$
2. Jeśli $f(\lambda_k) < f(\mu_k)$, to $x_{k+1} = x_k$ i $y_{k+1} = \mu_k$.
3. Jeśli $f(\lambda_k) \geqslant f(\mu_k)$, to $x_{k+1} = \lambda_k$ oraz $y_{k+1} = y_k$.
4. $\varepsilon_{k+1} = \varepsilon_k / 2$.

Koniec: $y_{k+1} - x_{k+1} \leqslant \eta$, gdzie η jest założoną dokładnością algorytmu.

Załóżmy, że istnieje minimum $\bar{x}$ funkcji f na przedziale $[a, b]$ (tak być nie musi, patrz zadanie 12.2). Na mocy lematu 12.2 minimum to leży w przedziale $[x_k, y_k]$ dla każdego k. Długość tego przedziału dąży do zera, co pociąga zbieżność zarówno (x_k) jak i (y_k) do rozwiązania $\bar{x}$. Dowiedliśmy zatem poprawności algorytmu.

Jeśli zatrzymamy się w kroku k-tym, to dokładność wyznaczenia punktu minimum zadana jest przez długość przedziału (x_{k+1}, y_{k+1}). Uzasadnia to wybór warunku końca.

W zadaniu 12.3 pokażemy, że długość przedziału (x_k, y_k) zmniejsza się wykładniczo szybko, tzn. istnieje stała $c \in (0, 1)$, taka że

$$y_{k+1} - x_{k+1} \leqslant (b - a)c^k.$$

Pamiętając, że rozwiązanie $\bar{x}$ leży w tym przedziale dla każdego k, wnioskujemy, iż zbieżność x_k i y_k do rozwiązania $\bar{x}$ jest wykładnicza.

Algorytm podziału dychotomicznego wymaga policzenia w każdym kroku iteracyjnym wartości funkcji f w dwóch punktach λ_k oraz μ_k. Znacznie efektywniejszy jest algorytm *złotego podziału*, który wymaga w każdym kroku policzenia tylko jednej wartości funkcji. Idea algorytmu jest bardzo podobna do poprzedniego. W kolejnych krokach iteracyjnych wyznaczamy punkty λ_k oraz μ_k a następnie zmniejszamy przedział $[x_k, y_k]$ do przedziału $[x_k, \mu_k]$ lub $[\lambda_k, y_k]$ w zależności od wartości funkcji f w punktach λ_k i μ_k. Przy wyborze punktów λ_k oraz μ_k kierujemy się jednak innymi regułami niż dla algorytmu podziału dychotomicznego:

1. Długość przedziału $[x_{k+1}, y_{k+1}]$ jest niezależna od wyniku poprzedniej iteracji, czyli relacji między $f(\lambda_k)$ a $f(\mu_k)$. Wynika stąd, że $y_k - \lambda_k = \mu_k - x_k$. Jeśli λ_k przedstawimy w postaci

$$\lambda_k = x_k + (1-\tau)(y_k - x_k), \tag{12.1}$$

   gdzie $\tau \in (0,1)$, to μ_k dana jest wyrażeniem

$$\mu_k = x_k + \tau(y_k - x_k). \tag{12.2}$$

   Wynika stąd, że

$$y_{k+1} - x_{k+1} = \tau(y_k - x_k). \tag{12.3}$$

2. Punkty λ_{k+1} oraz μ_{k+1} wybieramy w taki sposób, aby $\lambda_{k+1} = \mu_k$ albo $\mu_{k+1} = \lambda_k$. Taki wybór tych punktów gwarantuje, że w kolejnym kroku iteracyjnym oblicza się tylko jedną wartość funkcji f.

Rozważmy jakie są konsekwencje tych reguł. Jeśli $f(\lambda_k) \geqslant f(\mu_k)$, to $x_{k+1} = \lambda_k$, $y_{k+1} = y_k$ oraz $\lambda_{k+1} = \mu_k$. Korzystając z równości (12.1) mamy

$$\mu_k = \lambda_{k+1} = x_{k+1} + (1-\tau)(y_{k+1} - x_{k+1}) = \lambda_k + (1-\tau)(y_k - \lambda_k).$$

Podstawiając λ_k i μ_k z wyrażeń (12.1) i (12.2) otrzymujemy równanie

$$\tau^2 + \tau - 1 = 0. \tag{12.4}$$

Jeśli $f(\lambda_k) < f(\mu_k)$, to analogiczne rachunki prowadzą także do równania (12.4). Równanie (12.4) ma jeden pierwiastek dodatni $\tau = \frac{\sqrt{5}-1}{2}$, którego wartość jest stosunkiem złotego podziału (stąd nazwa algorytmu).

Algorytm złotego podziału składa się z następujących kroków:

Inicjalizacja: Weź $\tau = \frac{\sqrt{5}-1}{2}$. Połóż $x_1 = a$, $y_1 = b$. Oblicz $\lambda_1 = x_1 + (1-\tau)(y_1 - x_1)$ i $\mu_1 = x_1 + \tau(y_1 - x_1)$.

Krok k-ty:

1. Jeśli $f(\lambda_k) < f(\mu_k)$, to $x_{k+1} = x_k$ i $y_{k+1} = \mu_k$. Ponadto przyjmij $\mu_{k+1} = \lambda_k$ i $\lambda_{k+1} = x_{k+1} + (1-\tau)(y_{k+1} - x_{k+1})$ oraz oblicz $f(\lambda_{k+1})$.
2. Jeśli $f(\lambda_k) \geqslant f(\mu_k)$, to $x_{k+1} = \lambda_k$ oraz $y_{k+1} = y_k$. Ponadto przyjmij $\lambda_{k+1} = \mu_k$ i $\mu_{k+1} = x_{k+1} + \tau(y_{k+1} - x_{k+1})$ oraz oblicz $f(\mu_{k+1})$.

Koniec: $y_{k+1} - x_{k+1} \leqslant \eta$, gdzie η jest założoną dokładnością algorytmu.

Nieco bardziej skomplikowany jest *algorytm Fibonacciego*. Algorytm ten wykorzystuje ciąg Fibonacciego zdefiniowany rekurencyjnie

$$\begin{aligned} F_{k+1} &= F_k + F_{k-1}, \\ F_0 &= F_1 = 1. \end{aligned} \tag{12.5}$$

W algorytmie opartym o ciąg Fibonacciego ustalamy na początku docelową liczbę iteracji n. Następnie, wykorzystując liczby Fibonacciego, definiujemy iteracyjnie punkty λ_k oraz μ_k:

$$\lambda_k = x_k + \frac{F_{n-k-1}}{F_{n-k+1}}(y_k - x_k), \quad k = 1, \ldots, n-1, \tag{12.6}$$

$$\mu_k = x_k + \frac{F_{n-k}}{F_{n-k+1}}(y_k - x_k), \quad k = 1, \ldots, n-1. \tag{12.7}$$

Jeśli $f(\lambda_k) \geqslant f(\mu_k)$, to jako przedział $[x_{k+1}, y_{k+1}]$ przyjmujemy przedział $[\lambda_k, y_k]$. Wtedy

$$\begin{aligned} y_{k+1} - x_{k+1} =& y_k - \lambda_k = y_k - x_k - \frac{F_{n-k-1}}{F_{n-k+1}}(y_k - x_k) \\ =& \frac{F_{n-k}}{F_{n-k+1}}(y_k - x_k). \end{aligned} \tag{12.8}$$

Analogicznie, jeśli $f(\lambda_k) < f(\mu_k)$, to jako przedział $[x_{k+1}, y_{k+1}]$ przyjmujemy przedział $[x_k, \mu_k]$ i otrzymujemy

$$y_{k+1} - x_{k+1} = \mu_k - x_k = \frac{F_{n-k}}{F_{n-k+1}}(y_k - x_k). \tag{12.9}$$

Pokażemy teraz, że dla iteracji algorytmu Fibonacciego zachodzi jedna z dwóch równości $\lambda_{k+1} = \mu_k$ lub $\mu_{k+1} = \lambda_k$. Niech $f(\lambda_k) \geqslant f(\mu_k)$. Wtedy $x_{k+1} = \lambda_k$ oraz $y_{k+1} = y_k$. Wykorzystując równość (12.6) otrzymujemy

$$\lambda_{k+1} = x_{k+1} + \frac{F_{n-k-2}}{F_{n-k}}(y_{k+1} - x_{k+1}) = \lambda_k + \frac{F_{n-k-2}}{F_{n-k}}(y_k - \lambda_k).$$

Podstawiając do tej równości λ_k ze wzoru (12.6) mamy

$$\lambda_{k+1} = x_k + \frac{F_{n-k-1}}{F_{n-k+1}}(y_k - x_k) + \frac{F_{n-k-2}}{F_{n-k}}\left(1 - \frac{F_{n-k-1}}{F_{n-k+1}}\right)(y_k - x_k).$$

Ponieważ z definicji ciągu Fibonacciego wynika tożsamość

$$1 - \frac{F_{n-k-1}}{F_{n-k+1}} = \frac{F_{n-k}}{F_{n-k+1}},$$

więc

$$\lambda_{k+1} = x_k + \frac{F_{n-k-1} - F_{n-k-2}}{F_{n-k+1}}(y_k - x_k).$$

Korzystając ponownie z definicji ciągu Fibonacciego oraz równości (12.7) otrzymujemy

$$\lambda_{k+1} = x_k + \frac{F_{n-k}}{F_{n-k+1}}(y_k - x_k) = \mu_k.$$

Podobnie dla $f(\lambda_k) < f(\mu_k)$ dostajemy

$$\mu_{k+1} = \lambda_k.$$

Zauważmy, że z równań (12.6) i (12.7) wynika dla $k = n-1$

$$\lambda_{n-1} = \mu_{n-1} = \frac{x_{n-1} + y_{n-1}}{2}.$$

Oznacza to niemożliwość wykonania kolejnego kroku iteracyjnego. Aby dokonać jeszcze jednej iteracji należy przesunąć λ_{n-1} w lewo lub μ_{n-1} w prawo o $\varepsilon > 0$. Takie postępowanie pozwala zmniejszyć długość końcowego przedziału, w którym znajduje się minimum funkcji f do wielkości $(y_{n-1} - x_{n-1})/2$.

Ostatecznie algorytm Fibonacciego składa się z następujących kroków:

Inicjalizacja: Wybierz liczbę iteracji n oraz $\varepsilon > 0$. Połóż $x_1 = a$, $y_1 = b$. Oblicz $\lambda_1 = x_1 + \frac{F_{n-2}}{F_n}(y_1 - x_1)$ i $\mu_1 = x_1 + \frac{F_{n-1}}{F_n}(y_1 - x_1)$.

Krok k-ty:

1. Jeśli $f(\lambda_k) < f(\mu_k)$, to $x_{k+1} = x_k$ i $y_{k+1} = \mu_k$. Ponadto przyjmij $\mu_{k+1} = \lambda_k$ a $\lambda_{k+1} = x_{k+1} + \frac{F_{n-k-2}}{F_{n-k}}(y_{k+1} - x_{k+1})$ oraz oblicz $f(\lambda_{k+1})$. Jeśli $k = n-2$ idź następnie do punktu 3.
2. Jeśli $f(\lambda_k) \geqslant f(\mu_k)$, to $x_{k+1} = \lambda_k$ oraz $y_{k+1} = y_k$. Ponadto przyjmij $\lambda_{k+1} = \mu_k$ a $\mu_{k+1} = x_{k+1} + \frac{F_{n-k-1}}{F_{n-k}}(y_{k+1} - x_{k+1})$ oraz oblicz $f(\mu_{k+1})$. Jeśli $k = n-2$ idź następnie do punktu 3.
3. Przyjmij $\lambda_n = \lambda_{n-1}$ oraz $\mu_n = \lambda_{n-1} + \varepsilon$. Jeśli $f(\lambda_k) \geqslant f(\mu_k)$, to $x_n = \lambda_n$ a $y_n = y_{n-1}$. Jeśli $f(\lambda_k) < f(\mu_k)$, to $x_n = x_{n-1}$ a $y_n = \lambda_n$.

Koniec: Gdy $k = n$. Wtedy minimum funkcji f leży w przedziale $[x_n, y_n]$.

Trudność inicjalizacji tego algorytmu związana jest z wyborem właściwej wartości stałej ε. Musi być ona dostatecznie mała, aby punkt $\lambda_{n-1} + \varepsilon$ leżał wewnątrz przedziału $[x_{n-1}, y_{n-1}]$. Wybór właściwej stałej jest związany liczbą kroków algorytmu n (patrz zadanie 12.4).

12.3 Zadania

Ćwiczenie 12.1. Udowodnij lemat 12.1.

Ćwiczenie 12.2. Znajdź przykład funkcji ściśle quasi-wypukłej określonej na przedziale domkniętym, która nie przyjmuje swojego infimum, czyli zadanie minimalizacyjne nie ma rozwiązania.

Ćwiczenie 12.3. Wykaż, że w metodzie optymalizacji bez użycia pochodnych długość przedziału $[x_k, y_k]$ zmniejsza się w sposób wykładniczy, tzn. istnieje stała $c \in (0,1)$, taka że

$$y_{k+1} - x_{k+1} \leqslant (b-a)c^k.$$

Oszacuj stałą c z góry możliwie najlepiej, tzn. znajdź najmniejszą wartość górnego oszacowania.

Ćwiczenie 12.4. Wykaż, że w algorytmie poszukiwania minimum w oparciu o ciąg Fibonacciego długość przedziału, w którym znajduje się rozwiązanie dana jest wzorem

$$y_n - x_n = \frac{b-a}{F_n}.$$

Korzystając z tego związku znajdź górne oszacowanie na stałą ε gwarantujące poprawność algorytmu.

Ćwiczenie 12.5. Korzystając z wyniku zadania 12.3, wzoru (12.3) oraz wyniku zadania 12.4 oceń, który z omawianych algorytmów: podziału dychotomicznego, złotego podziału oraz Fibonacciego zbiega najszybciej, tj. w ustalonej liczbie iteracji daje przedział, w którym istnieje minimum funkcji f, o najmniejszej długości.

Rozdział 13

Metody spadkowe

W tym rozdziale przyjrzymy się wielowymiarowym metodom optymalizacji bez ograniczeń. Zakładamy, że $f : \mathbb{R}^n \to \mathbb{R}$ jest funkcją klasy C^1. Naszym celem jest znalezienie jej minimum. Wszystkie opisywane metody zbudowane są w oparciu o następujący ogólny schemat. Startujemy z punktu początkowego $\mathbf{x}_1$, który w naszym przekonaniu znajduje się blisko minimum. W kolejnych krokach generujemy $\mathbf{x}_2, \mathbf{x}_3, \ldots$ w ten sposób, aby $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$. Spodziewamy się, że w ten sposób dojdziemy aż do minimum f. Może się jednak okazać, że punkty skupienia ciągu $(\mathbf{x}_k)$ nie będą rozwiązaniami. Na kolejnych stronach tego rozdziału zajmiemy się opisem różnych metod konstrukcji ciągu $(\mathbf{x}_k)$ i ich zbieżnością do rozwiązania. Wszystkie wyniki tego rozdziału w oczywisty sposób przenoszą się na przypadek funkcji określonych na zbiorach otwartych w $\mathbb{R}^n$. Metody spadkowe będą również grały pierwsze skrzypce w następnym rozdziale, przy optymalizacji numerycznej z ograniczeniami.

Sformułujmy ściśle problem optymalizacyjny:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ \mathbf{x} \in \mathbb{R}^n. \end{cases}$$

Metodami spadkowymi nazywamy takie algorytmy, w których kolejny punkt $\mathbf{x}_{k+1}$ zadany jest wzorem

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k,$$

gdzie $\alpha_k > 0$ oraz $\mathbf{d}_k$ jest kierunkiem spadku, tzn.

$$\begin{aligned} Df(\mathbf{x}_k)\mathbf{d}_k < 0, &\quad \text{jeśli } Df(\mathbf{x}_k) \neq \mathbf{0}, \\ \mathbf{d}_k = 0, &\quad \text{jeśli } Df(\mathbf{x}_k) = \mathbf{0}. \end{aligned}$$

Zauważmy, że dla dostatecznie małych α_k mamy $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ (patrz zadanie 13.1). Możemy więc liczyć, że przy dobrym doborze α_k ciąg punktów $(\mathbf{x}_k)$ będzie zbiegał do minimum. Niestety bez dodatkowych informacji o zadaniu nie możemy zagwarantować, że będzie to minimum globalne, ale będziemy starać się znaleźć sposób na zapewnienie zbieżności do minimum lokalnego.

13.1 Metody największego spadku

W tym podrozdziale skupimy się na metodach największego spadku, czyli takich że $\mathbf{d}_k$ jest równoległe do $\left(Df(\mathbf{x}_k)\right)^T$.

Algorytm opisujący metody największego spadku jest dość prosty:

Inicjalizacja: Wybierz punkt początkowy $\mathbf{x}_1$.

Krok k-ty:

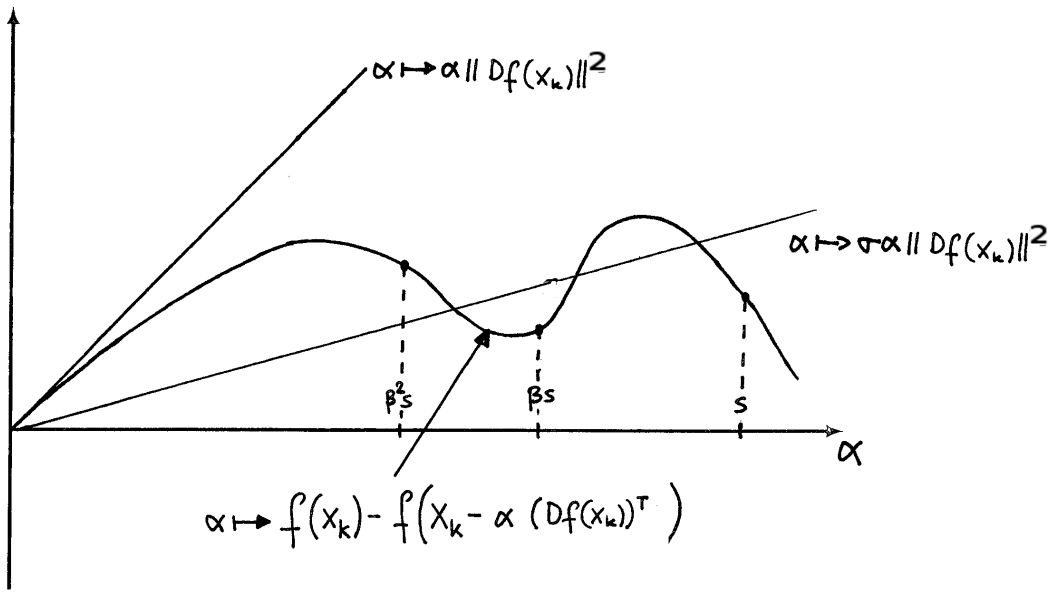
1. Wybierz krok α_k.
2. Połóż $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k (Df(\mathbf{x}_k))^T$.

Koniec: gdy $\|Df(\mathbf{x}_{k+1})\| \leqslant \varepsilon$.

Pozostaje doprecyzowanie reguł znajdowania kroku α_k. Reguły te są stosowane dla dowolnych kierunków spadku $\mathbf{d}_k$ i w takiej formie je prezentujemy. Czytelnik powinien pamiętać, że w przypadku metod największego spadku $\mathbf{d}_k = -(Df(\mathbf{x}_k))^T$.

- *reguła (dokładnej) minimalizacji*: wybierz α_k takie że
$$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) = \min_{\alpha \geqslant 0} f(\mathbf{x}_k + \alpha \mathbf{d}_k).$$
- *reguła ograniczonej minimalizacji*: ustalmy $A > 0$. Wybierz α_k takie że
$$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) = \min_{\alpha \in [0,A]} f(\mathbf{x}_k + \alpha \mathbf{d}_k).$$
- *reguła Armijo*: ustalmy $s > 0$ i $\beta, \sigma \in (0,1)$. Połóżmy $\alpha_k = \beta^{m_k} s$, gdzie m_k jest najmniejszą liczbą całkowitą m spełniającą nierówność
$$f(\mathbf{x}_k) - f(\mathbf{x}_k + \beta^m s\, \mathbf{d}_k) \geqslant -\sigma \beta^m s\; Df(\mathbf{x}_k)\mathbf{d}_k.$$
Oznacza to, że dla $m_k - 1$ zachodzi już nierówność przeciwna
$$f(\mathbf{x}_k) - f(\mathbf{x}_k + \beta^{m_k - 1} s\, \mathbf{d}_k) < -\sigma \beta^{m_k - 1} s\; Df(\mathbf{x}_k)\mathbf{d}_k.$$
Stała $s > 0$ nazywana jest krokiem, β reguluje szybkość zmniejszania kroku (czym mniejsza wartość, tym szybciej krok się zmniejsza), zaś σ odpowiedzialna jest za ostrość warunku: mniejsza wartość osłabia warunek. Dla metody największego spadku, gdzie $\mathbf{d}_k = -(Df(\mathbf{x}_k))^T$, warunek wyznaczania m_k upraszcza się nieznacznie, ponieważ $Df(\mathbf{x}_k)\mathbf{d}_k = -\|Df(\mathbf{x}_k)\|^2$:
$$f(\mathbf{x}_k) - f\left(\mathbf{x}_k - \beta^m s \left(Df(\mathbf{x}_k)\right)^T\right) \geqslant \sigma \beta^m s \|Df(\mathbf{x}_k)\|^2. \tag{13.1}$$
Ilustracja działania metody Armijo w tej wersji znajduje się na rysunku 13.1. Jej implementacja jest bardzo łatwa. Startujemy z punktu $\mathbf{x}_k - s\left(Df(\mathbf{x}_k)\right)^T$ i sprawdzamy, czy spełniony jest w nim warunek (13.1). Jeśli nie, to rozważamy punkty $\mathbf{x}_k - \beta s \left(Df(\mathbf{x}_k)\right)^T$, $\mathbf{x}_k - \beta^2 s \left(Df(\mathbf{x}_k)\right)^T$, itd. Jeśli warunek (13.1) jest spełniony w punkcie $\mathbf{x}_k - s\left(Df(\mathbf{x}_k)\right)^T$, sprawdzamy czy pozostaje on spełniony w punktach $\mathbf{x}_k - \beta^{-1} s \left(Df(\mathbf{x}_k)\right)^T$, $\mathbf{x}_k - \beta^{-2} s \left(Df(\mathbf{x}_k)\right)^T$, itd. W zadaniu 13.2 pokażemy, że w skończonej liczbie kroków znajdziemy punkt spełniający warunek (13.1) dla m i nie spełniający go dla $m-1$, jeśli $\mathbf{d}_k \neq \mathbf{0}$.

Reguła dokładnej minimalizacji jest dobrze określona, jeśli minimum po prawej stronie istnieje. Ze względu na nieograniczoność przedziału, na którym poszukujemy minimum, reguła dokładnej minimalizacji nie zawsze musi być dobrze określona. W dalszym ciągu podamy założenia, które będą gwarantowały poprawność działania tej reguły. Ograniczenie przedziału poszukiwań kroku α_k do $[0, A]$ ma dwie zalety. Po pierwsze zadanie to ma zawsze rozwiązanie, gdyż minimalizujemy



Rysunek 13.1: Ilustracja działania metody Armijo. Dla kroków $\alpha = s$ i $\alpha = \beta s$ warunek (13.1) nie jest spełniony. Krok $\alpha = \beta^2 s$ jest pierwszym, dla którego (13.1) zachodzi, więc $\alpha_k = \beta^2 s$.

funkcję ciągłą na zbiorze zwartym. Po drugie, możemy zastosować szybsze metody poszukiwania minimum.

W zadaniu 13.1 dowodzimy, że obie reguły minimalizacyjne wyznaczają krok o tej własności, że $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$. Z nierówności (13.1) wynika, że także dla reguły Armijo $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$. Pozostają jednak dwa pytania: czy warunek końca jest poprawny i czy ciąg $(\mathbf{x}_k)$ zbiega do minimum.

Poniżej dowodzimy, że każdy punkt skupienia ciągu $(\mathbf{x}_k)$ utworzonego metodą największego spadku jest punktem krytycznym, tzn. zeruje się w nim pochodna. Nie jest jednak w ogólności prawdą, że musi w nim być minimum lokalne. Dopiero założenie, że funkcja f jest pseudowypukła daje nam pewność, że w znalezionym punkcie jest minimum, które dodatkowo jest globalne.

Twierdzenie 13.1. *Niech $(\mathbf{x}_k)$ będzie ciągiem skonstruowanych przy pomocy metody największego spadku z regułą minimalizacji, regułą minimalizacji ograniczonej lub regułą Armijo. Wówczas każdy punkt skupienia tego ciągu jest punktem krytycznym.*

Dowód. Niech $\bar{\mathbf{x}}$ będzie punktem skupienia, zaś $(\mathbf{x}_{k_n})$ podciągiem do niego zbieżnym. Dowód przeprowadzimy przez sprzeczność zakładając, że $Df(\bar{\mathbf{x}}) \neq \mathbf{0}^T$.

Dla reguły minimalizacji i minimalizacji ograniczonej głównym pomysłem dowodu będzie pokazanie, że jeśli $Df(\bar{\mathbf{x}}) \neq \mathbf{0}^T$, to istnieje stała $\gamma > 0$, taka że dla dostatecznie dużych n, tzn. dla $\mathbf{x}_{k_n}$ dostatecznie bliskich $\bar{\mathbf{x}}$, zachodzi nierówność

$$f(\mathbf{x}_{k_n+1}) \leqslant f(\mathbf{x}_{k_n}) - \gamma,$$

czyli możliwe jest zmniejszenie wartości funkcji f w kroku k_n o co najmniej γ. Pamiętając, że funkcja f maleje wzdłuż ciągu $(\mathbf{x}_k)$ mamy

$$f(\mathbf{x}_{k_{n+1}}) \leqslant f(\mathbf{x}_{k_n+1}) \leqslant f(\mathbf{x}_{k_n}) - \gamma.$$

Przechodząc do granicy z $n \to \infty$ i korzystając z faktu, że $\bar{\mathbf{x}}$ jest punktem skupienia oraz ciągłości f dostajemy sprzeczność z dodatniością γ.

Do zakończenia dowodu pozostaje nam udowodnienie istnienia stałej dodatniej γ. Na mocy ciągłości pochodnej (przypomnijmy, że w tym rozdziale zakładamy, że f jest klasy C^1) istnieje otoczenie V punktu $\bar{\mathbf{x}}$, takie że

$$\frac{\|Df(\mathbf{x}) - Df(\mathbf{y})\|}{\|Df(\mathbf{x})\|} \leqslant \frac{1}{2}, \qquad \forall\, \mathbf{x}, \mathbf{y} \in V. \tag{13.2}$$

Głównym spostrzeżeniem pozwalającym na uzasadnienie powyższej nierówności jest to, że na pewnym otoczeniu $\bar{\mathbf{x}}$ norma pochodnej f jest ściśle oddzielona od zera.

Weźmy $\delta > 0$, takie że $B(\bar{\mathbf{x}}, 2\delta) \subset V$ oraz $\delta \leqslant A \inf_{\mathbf{x}\in V} \|Df(\mathbf{x})\|$, gdzie A jest stałą z definicji reguły minimalizacji ograniczonej (zauważmy, że $\inf_{\mathbf{x}\in V} \|Df(\mathbf{x})\| > 0$ na mocy (13.2)). Dla $\mathbf{x} \in B(\bar{\mathbf{x}}, \delta)$ punkt $\mathbf{x} - \dfrac{\delta}{\|Df(\mathbf{x})\|}(Df(\mathbf{x}))^T$ należy do $B(\mathbf{x}, 2\delta)$, a więc również do V. Ponadto, w przypadku reguły minimalizacji ograniczonej

$$\frac{\delta}{\|Df(\mathbf{x})\|} \leqslant \frac{A \inf_{\mathbf{x}\in V} \|Df(\mathbf{x})\|}{\|Df(\mathbf{x})\|} \leqslant A,$$

czyli $\frac{\delta}{\|Df(\mathbf{x})\|} \in [0, A]$. Z twierdzenia o wartości średniej dostajemy

$$f(\mathbf{x}) - f\Big(\mathbf{x} - \frac{\delta}{\|Df(\mathbf{x})\|}(Df(\mathbf{x}))^T\Big) = Df(\theta)\Big(\frac{\delta}{\|Df(\mathbf{x})\|}(Df(\mathbf{x}))^T\Big) = \frac{\delta}{\|Df(\mathbf{x})\|} Df(\theta)(Df(\mathbf{x}))^T,$$

gdzie θ jest punktem pośrednim, a więc należy do V. Zajmijmy się teraz produktem pochodnych:

$$\begin{aligned} Df(\theta)(Df(\mathbf{x}))^T &= \Big(Df(\mathbf{x}) + Df(\theta) - Df(\mathbf{x})\Big)(Df(\mathbf{x}))^T \\ &= \|Df(\mathbf{x})\|^2 + (Df(\theta) - Df(\mathbf{x}))(Df(\mathbf{x}))^T \\ &\geqslant \|Df(\mathbf{x})\|^2 - \|Df(\theta) - Df(\mathbf{x})\|\|Df(\mathbf{x})\|. \end{aligned}$$

Wstawiamy to oszacowanie do poprzedniego wzoru:

$$\begin{aligned} f(\mathbf{x}) - f\Big(\mathbf{x} - \frac{\delta}{\|Df(\mathbf{x})\|}(Df(\mathbf{x}))^T\Big) &\geqslant \delta\|Df(\mathbf{x})\| - \delta\|Df(\theta) - Df(\mathbf{x})\| \\ &= \delta\|Df(\mathbf{x})\|\Big(1 - \frac{\|Df(\theta) - Df(\mathbf{x})\|}{\|Df(\mathbf{x})\|}\Big) \\ &\geqslant \frac{\delta}{2}\|Df(\mathbf{x})\|, \end{aligned}$$

gdzie ostatnia nierówność wynika z (13.2). Powyższe oszacowanie jest prawdziwe dla dowolnego $\mathbf{x} \in B(\bar{\mathbf{x}}, \delta)$, a więc dla $\mathbf{x}_{k_n}$ dla dostatecznie dużych n

$$f(\mathbf{x}_{k_n}) - f(\mathbf{x}_{k_n+1}) \geqslant f(\mathbf{x}_{k_n}) - f\Big(\mathbf{x}_{k_n} - \frac{\delta}{\|Df(\mathbf{x}_{k_n})\|}(Df(\mathbf{x}_{k_n}))^T\Big) \geqslant \frac{\delta}{2}\|Df(\mathbf{x}_{k_n})\|.$$

Możemy zatem przyjąć

$$\gamma = \frac{\delta}{2} \inf_{\mathbf{x}\in B(\bar{\mathbf{x}},\delta)} \|Df(\mathbf{x})\|.$$

Dla reguły Armijo zastosowanej do metody największego spadku mamy nierówność

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geqslant \sigma\alpha_k\|Df(\mathbf{x}_k)\|^2,$$

z której wynika, że ciąg $(f(\mathbf{x}_k))$ jest monotonicznie malejący, czyli albo zbieżny albo rozbieżny do $-\infty$. Ponieważ $f(\mathbf{x}_{k_n}) \to f(\bar{\mathbf{x}})$, więc $f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \to 0$. Wynika stąd, że także $\alpha_k \|Df(\mathbf{x}_k)\|^2 \to 0$. Ponieważ $Df(\mathbf{x}_{k_n}) \to Df(\bar{\mathbf{x}}) \neq \mathbf{0}^T$, więc $\alpha_{k_n} \to 0$.
Z drugiej strony wartości α_k w regule Armijo są dobierane optymalnie, tzn.

$$f(\mathbf{x}_k) - f(\mathbf{x}_k - \alpha_k/\beta(Df(\mathbf{x}_k))^T) < \sigma\alpha_k/\beta\|Df(\mathbf{x}_k)\|^2.$$

Stosując twierdzenie o wartości średniej do lewej strony tej nierówności dostajemy

$$f(\mathbf{x}_k) - f(\mathbf{x}_k - \alpha_k/\beta(Df(\mathbf{x}_k))^T) = Df(\mathbf{x}_k - \tilde{\alpha}_k/\beta(Df(\mathbf{x}_k))^T)\alpha_k/\beta(Df(\mathbf{x}_k))^T,$$

więc nierówność przyjmuje postać

$$Df(\mathbf{x}_k - \tilde{\alpha}_k/\beta(Df(\mathbf{x}_k))^T)(Df(\mathbf{x}_k))^T < \sigma\|Df(\mathbf{x}_k)\|^2.$$

Przechodząc do podciągu (k_n) mamy

$$Df(\mathbf{x}_{k_n} - \tilde{\alpha}_{k_n}/\beta(Df(\mathbf{x}_{k_n}))^T)(Df(\mathbf{x}_{k_n}))^T < \sigma\|Df(\mathbf{x}_{k_n})\|^2.$$

Jeśli teraz w ostatniej nierówności przejdziemy do granicy, to zauważmy, że $\tilde{\alpha}_{k_n} \in [0, \alpha_{k_n}]$, czyli $\tilde{\alpha}_{k_n} \to 0$ bo $\alpha_{k_n} \to 0$. W granicy ostatnia nierówność ma postać $\|Df(\bar{\mathbf{x}})\|^2 \leqslant \sigma\|Df(\bar{\mathbf{x}})\|^2$ czyli $(1-\sigma)\|Df(\bar{\mathbf{x}})\|^2 \leqslant 0$. Ponieważ $(1-\sigma) > 0$ prowadzi to do sprzeczności z założeniem, że $Df(\bar{\mathbf{x}}) \neq \mathbf{0}^T$. □

Okazuje się, że przy nieco wzmocnionych założeniach $Df(\mathbf{x}_k) \to 0$ dla całego ciągu wygenerowanego algorytmem największego spadku, a nie tylko na podciągu posiadającym punkt skupienia.
Rozpocznijmy od następującego pomocniczego lematu.

Lemat 13.1. *Niech f będzie funkcją klasy C^1 ograniczoną z dołu. Niech $(\mathbf{x}_k)$ będzie ciągiem skonstruowanych przy pomocy metody największego spadku. Jeśli istnieje stała $c > 0$ niezależna od k, taka że*

$$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) < f(\mathbf{x}_k) - c\|Df(\mathbf{x}_k)\|^2, \quad \text{dla } k = 1, \ldots. \tag{13.3}$$

Wówczas albo istnieje K, takie że $Df(\mathbf{x}_K) = \mathbf{0}^T$, albo ciąg $(Df(\mathbf{x}_k))$ zbiega do zera.

Dowód. Przypadek stopu po skończonej liczbie kroków jest oczywisty. Przyjmijmy więc, że istnieje nieskończony ciąg $(\mathbf{x}_k)$ skonstruowany przy pomocy metody największego spadku. Z nierówności (13.3) wynika, że ciąg $(f(\mathbf{x}_k))$ monotonicznie maleje. Ponieważ jest on ograniczony z dołu, więc jest zbieżny, czyli $f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \to 0$. Ponieważ z nierówności (13.3) wynika

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) > c\|Df(\mathbf{x}_k)\|^2,$$

więc $\|Df(\mathbf{x}_k)\| \to 0$. □

Twierdzenie 13.2. *Niech funkcja f będzie ograniczona z dołu a gradient funkcji f będzie funkcją spełniającą warunek Lipschitza ze stałą L na zbiorze poziomicowym $S = W_{f(\mathbf{x}_1)}(f)$. Niech $(\mathbf{x}_k)$ będzie ciągiem skonstruowanych przy pomocy metody największego spadku z regułą Armijo, regułą dokładnej minimalizacji, która jest dobrze określona w każdym kroku (poprawność dokładnej reguły minimalizacji w każdym kroku gwarantuje w szczególności założenie zwartości zbioru S) lub minimalizacji ograniczonej, dla której $A > \frac{1}{2L}$. Wówczas albo istnieje K, takie że $Df(\mathbf{x}_K) = 0$, albo ciąg $(Df(\mathbf{x}_k))$ zbiega do zera.*

Dowód. Dowód przeprowadzimy przez wykazanie, że przy założeniach twierdzenia spełnione są założenia lematu 13.1.

Rozpoczniemy od dowodu dla metody Armijo. Niech $\alpha_k = s\beta^{m_k}$ będzie krokiem wyznaczonym w metodzie Armijo. Wtedy

$$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) \leqslant f(\mathbf{x}_k) + \sigma\alpha_k Df(\mathbf{x}_k)\mathbf{d}_k. \tag{13.4}$$

Ponieważ α_k zostało wyznaczone optymalnie, czyli wybrano najmniejszą potęgę β, przy której zachodzi powyższa nierówność, to dla niższej potęgi β zachodzi nierówność odwrotna

$$f(\mathbf{x}_k + \beta^{-1}\alpha_k \mathbf{d}_k) > f(\mathbf{x}_k) + \sigma\beta^{-1}\alpha_k Df(\mathbf{x}_k)\mathbf{d}_k. \tag{13.5}$$

Ponieważ $Df(\mathbf{x}_k)\mathbf{d}_k < 0$, to z faktu, że $\mathbf{x}_k \in S$ oraz nierówności (13.4) wynika $\mathbf{x}_{k+1} \in S$. Z twierdzenia o wartości średniej dostajemy

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) = f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) - f(\mathbf{x}_k) = \alpha_k Df(\tilde{\mathbf{x}})\mathbf{d}_k,$$

gdzie $\tilde{\mathbf{x}}$ jest wypukłą kombinacją $\mathbf{x}_k$ i $\mathbf{x}_{k+1}$ oraz $\tilde{\mathbf{x}} \in S$ (jeśli $\mathbf{x}_k \in S$ i $\mathbf{x}_{k+1} \in S$, to w S są także wszystkie punkty na odcinku między tymi dwoma punktami, co wynika z algorytmu minimalizacji w kierunkach spadkowych). Wykorzystując fakt, że kierunek $\mathbf{d}_k$ jest kierunkiem największego spadku, czyli $\mathbf{d}_k = -\big(Df(\mathbf{x}_k)\big)^T$, dostajemy

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) =& \alpha_k Df(\tilde{\mathbf{x}})\mathbf{d}_k = -\alpha_k \big(Df(\mathbf{x}_k) - Df(\mathbf{x}_k) + Df(\tilde{\mathbf{x}})\big)\big(Df(\mathbf{x}_k)\big)^T = \\ & - \alpha_k \|Df(\mathbf{x}_k)\|^2 + \alpha_k \big(Df(\mathbf{x}_k) - Df(\tilde{\mathbf{x}})\big)\big(Df(\mathbf{x}_k)\big)^T \leqslant \\ & - \alpha_k \|Df(\mathbf{x}_k)\|^2 + \alpha_k \|Df(\mathbf{x}_k) - Df(\tilde{\mathbf{x}})\| \|Df(\mathbf{x}_k)\|. \end{aligned}$$

Ponieważ $\mathbf{x}_k \in S$, dla każdego $k > 1$, a tym samym także każdy $\tilde{\mathbf{x}} \in S$, to z warunku Lipschitza dla gradientu funkcji f na S mamy oszacowanie

$$\|Df(\mathbf{x}_k) - Df(\tilde{\mathbf{x}})\| \leqslant L\,\|\mathbf{x}_k - \tilde{\mathbf{x}}\| \leqslant L\,\|\mathbf{x}_k - \mathbf{x}_{k+1}\| = L\,\alpha_k \|Df(\mathbf{x}_k)\|.$$

Wstawiając to oszacowanie do wcześniejszej nierówności dostajemy

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leqslant -\alpha_k \|Df(\mathbf{x}_k)\|^2 (1 - \alpha_k L). \tag{13.6}$$

Z oszacowania (13.6) wynika, że (13.4) zachodzi jeśli $1 - \alpha_k L \geqslant \sigma$. Niech m_σ będzie taką potęgą β, że $1 - s\beta^{m_\sigma} L \geqslant \sigma$ a dla potęgi β mniejszej niż m_σ zachodzi nierówność przeciwna $1 - s\beta^{m_\sigma - 1} L < \sigma$. Wynika stąd, że

$$\sigma s \beta^{m_\sigma} > \frac{\sigma(1-\sigma)\beta}{L}.$$

Z drugiej strony m_σ nie musi być potęgą optymalną, dla której są spełnione nierówności (13.4) i (13.5). Reguła Armijo gwarantuje jednak, że $m_k \leqslant m_\sigma$, bo m_k jest najmniejszą potęgą β przy której jest spełniona nierówność (13.4). Stąd

$$\sigma s \beta^{m_k} \geqslant \sigma s \beta^{m_\sigma} > \frac{\sigma(1-\sigma)\beta}{L}.$$

Wstawiając powyższe oszacowania do nierówności (13.4) oraz wykorzystując fakt, że $\mathbf{d}_k$ jest kierunkiem największego spadku dostajemy nierówność

$$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) - f(\mathbf{x}_k) \leqslant -\sigma s \beta^{m_k} \|Df(\mathbf{x}_k)\|^2 \leqslant -\sigma s \beta^{m_\sigma} \|Df(\mathbf{x}_k)\|^2 < -\frac{\sigma(1-\sigma)\beta}{L} \|Df(\mathbf{x}_k)\|^2,$$

czyli nierówność (13.3) z lematu 13.1.

W przypadku reguły dokładnej minimalizacji zakładamy, że w każdym kroku reguła ta jest dobrze określona, czyli istnieje zawsze skończone α_k, które minimalizuje $f(\mathbf{x}_k+\alpha\mathbf{d}_k)$ dla $\alpha \geqslant 0$.

Postępując podobnie jak w dowodzie dla reguły Armijo dostajemy oszacowanie

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leqslant -\alpha_k \|Df(\mathbf{x}_k)\|^2(1-\alpha_k L). \tag{13.7}$$

Prawa strona tego oszacowania osiąga najmniejszą wartość dla $\alpha_k = \frac{1}{2L}$. Wynika stąd oszacowanie

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leqslant -\frac{1}{4L}\|Df(\mathbf{x}_k)\|^2,$$

czyli nierówność (13.3) z lematu 13.1.

W przypadku reguły ograniczonej minimalizacji dowód jest analogiczny jak dla reguły dokładnej minimalizacji. Musimy jedynie założyć, że przedział $[0, A]$ na którym poszukujemy α_k zawiera punkt $\frac{1}{2L}$ minimalizujący prawa stronę nierówności (13.7). □

Powyższe twierdzenie dowodzi poprawności metod największego spadku, lecz nie mówi nic o szybkości zbieżności i o warunku końca. Wzmacniając założenia będziemy mogli dowieść, że zbieżność jest liniowa oraz warunek końca jest poprawny.

Zdefiniujmy zbiór $S = \{\mathbf{x} \in \mathbb{R}^n : \ f(\mathbf{x}) \leqslant f(\mathbf{x}_1)\}$, gdzie $\mathbf{x}_1$ jest punktem początkowym. Oznaczmy przez $m(\mathbf{x})$ najmniejszą wartość własną macierzy drugich pochodnych $D^2f(\mathbf{x})$, zaś przez $M(\mathbf{x})$ – jej największą wartość własną.

Lemat 13.2. *Załóżmy, że zbiór S jest wypukły i zwarty a funkcja f jest klasy C^2 na S oraz $m = \inf_{\mathbf{x}\in S} m(\mathbf{x}) > 0$. Wówczas punkt $\bar{\mathbf{x}}$ będący granicą ciągu $(\mathbf{x}_k)$ wyznaczonego metodą największego spadku należy do S i jest minimum funkcji f na S oraz dla dowolnego $\mathbf{x} \in S$*

$$\|\mathbf{x}-\bar{\mathbf{x}}\| \leqslant \frac{1}{m}\|Df(\mathbf{x})\|, \qquad f(\mathbf{x}) - f(\bar{\mathbf{x}}) \leqslant \frac{1}{m}\|Df(\mathbf{x})\|^2.$$

Dowód. Ponieważ $m = \inf_{\mathbf{x}\in S} m(\mathbf{x}) > 0$ więc macierz $D^2f(\mathbf{x})$ jest dodatnio określona na S. Z wypukłości S wynika, że $f(\mathbf{x})$ jest funkcją wypukłą na S.

Ponieważ zbiór S jest zwarty, to funkcja f jest ograniczona z dołu i jej gradient spełnia warunek Lipschitza. Z twierdzenia 13.2 wynika, że $Df(\mathbf{x}_k) \to \mathbf{0}$, czyli $\bar{\mathbf{x}} \in S$. Ponieważ funkcja f jest wypukła na S, to punkt krytyczny jest punktem minimum.

Na zbiorze S mamy ze wzoru Taylora

$$\begin{aligned} f(\mathbf{x}) &= f(\bar{\mathbf{x}}) + \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})^T D^2f(\tilde{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}}),\\ f(\bar{\mathbf{x}}) &= f(\mathbf{x}) + Df(\mathbf{x})(\bar{\mathbf{x}}-\mathbf{x}) + \frac{1}{2}(\bar{\mathbf{x}}-\mathbf{x})^T D^2f(\hat{\mathbf{x}})(\bar{\mathbf{x}}-\mathbf{x}). \end{aligned}$$

Wstawiając pierwsze równanie do drugiego otrzymujemy

$$Df(\mathbf{x})(\bar{\mathbf{x}}-\mathbf{x}) + (\mathbf{x}-\bar{\mathbf{x}})^T \frac{D^2f(\hat{\mathbf{x}}) + D^2f(\tilde{\mathbf{x}})}{2}(\mathbf{x}-\bar{\mathbf{x}}) = 0.$$

Wynika stąd oszacowanie

$$Df(\mathbf{x})(\mathbf{x}-\bar{\mathbf{x}}) = (\mathbf{x}-\bar{\mathbf{x}})^T \frac{D^2f(\hat{\mathbf{x}}) + D^2f(\tilde{\mathbf{x}})}{2}(\mathbf{x}-\bar{\mathbf{x}}) \geqslant \|\mathbf{x}-\bar{\mathbf{x}}\|^2 m.$$

Co daje

$$\|Df(\mathbf{x})\|\|\mathbf{x}-\bar{\mathbf{x}}\| \geqslant \|\mathbf{x}-\bar{\mathbf{x}}\|^2 m.$$

Dzieląc ostatnią nierówność stronami przez $m\|\mathbf{x}-\bar{\mathbf{x}}\|$ dostajemy

$$\|\mathbf{x}-\bar{\mathbf{x}}\| \leqslant \frac{1}{m}\|Df(\mathbf{x})\|.$$

Wykorzystując ponownie wzór Taylora oraz wypukłość funkcji f na S otrzymujemy oszacowanie

$$f(\bar{\mathbf{x}}) = f(\mathbf{x}) + Df(\mathbf{x})(\bar{\mathbf{x}}-\mathbf{x}) + \frac{1}{2}(\bar{\mathbf{x}}-\mathbf{x})^T D^2 f(\hat{\mathbf{x}})(\bar{\mathbf{x}}-\mathbf{x}) \geqslant f(\mathbf{x}) + Df(\mathbf{x})(\mathbf{x}-\bar{\mathbf{x}}).$$

Stąd

$$f(\mathbf{x}) - f(\bar{\mathbf{x}}) \leqslant Df(\mathbf{x})(\mathbf{x}-\bar{\mathbf{x}}),$$

czyli

$$f(\mathbf{x}) - f(\bar{\mathbf{x}}) = |f(\mathbf{x}) - f(\bar{\mathbf{x}})| \leqslant \|Df(\mathbf{x})\|\|\mathbf{x}-\bar{\mathbf{x}}\| \leqslant \frac{1}{m}\|Df(\mathbf{x})\|^2.$$

□

Lemat 13.3. *Załóżmy, że zbiór S jest wypukły i zwarty a funkcja f jest klasy C^2 na S oraz $m = \inf_{\mathbf{x}\in S} m(\mathbf{x}) > 0$. Oznaczmy $M = \sup_{\mathbf{x}\in S} M(\mathbf{x})$. Wówczas $M < +\infty$ a dla ciągu $(\mathbf{x}_k)$ wygenerowanego przy pomocy metody największego spadku z regułą dokładnej minimalizacji mamy*

$$f(\mathbf{x}_{k+1}) - f(\bar{\mathbf{x}}) \leqslant \Big(1 - \frac{m}{2M}\Big)(f(\mathbf{x}_k) - f(\bar{\mathbf{x}})),$$

zaś dla reguły minimalizacji ograniczonej

$$f(\mathbf{x}_{k+1}) - f(\bar{\mathbf{x}}) \leqslant \Big(1 - m\gamma + \frac{mM\gamma^2}{2}\Big)(f(\mathbf{x}_k) - f(\bar{\mathbf{x}})),$$

gdzie

$$\gamma = \min\left(\frac{1}{M}, A\right).$$

Dowód. Rozważmy przypadek reguły minimalizacji bez ograniczeń. Na mocy wzoru Taylora, dla $\delta \geqslant 0$, mamy

$$\begin{aligned} f\big(\mathbf{x}_k - \delta(Df(\mathbf{x}_k))^T\big) &\leqslant f(\mathbf{x}_k) + Df(\mathbf{x}_k)\big(-\delta(Df(\mathbf{x}_k))^T\big) + \delta^2\frac{M}{2}\|Df(\mathbf{x}_k)\|^2 \\ &= f(\mathbf{x}_k) - \delta\|Df(\mathbf{x}_k)\|^2 + \delta^2\frac{M}{2}\|Df(\mathbf{x}_k)\|^2. \end{aligned} \tag{13.8}$$

Minimum po prawej stronie przyjmowane jest dla $\delta = 1/M$, patrz rys. 13.2. Przypomnijmy również, że $\mathbf{x}_{k+1}$ realizuje minimum po $\alpha \geqslant 0$

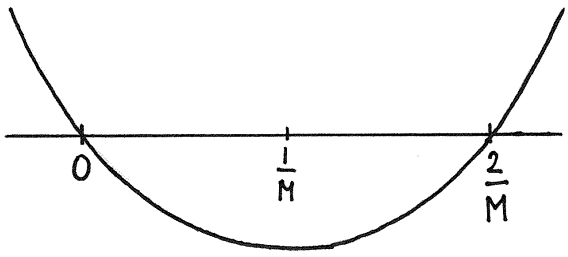
$$f(\mathbf{x}_{k+1}) = \inf_{\alpha\geqslant 0} f\big(\mathbf{x}_k - \alpha(Df(\mathbf{x}_k))^T\big).$$

Zatem

$$f(\mathbf{x}_{k+1}) \leqslant f\Big(\mathbf{x}_k - \frac{1}{M}(Df(\mathbf{x}_k))^T\Big) \leqslant f(\mathbf{x}_k) - \frac{1}{2M}\|Df(\mathbf{x}_k)\|^2.$$

Odejmijmy od obu stron $f(\bar{\mathbf{x}})$ i zastosujmy nierówność $\|Df(\mathbf{x}_k)\|^2 \geqslant m(f(\mathbf{x}_k) - f(\bar{\mathbf{x}}))$ wynikającą z lematu 13.2:

$$f(\mathbf{x}_{k+1}) - f(\bar{\mathbf{x}}) \leqslant f(\mathbf{x}_k) - f(\bar{\mathbf{x}}) - \frac{m}{2M}(f(\mathbf{x}_k) - f(\bar{\mathbf{x}})).$$



Rysunek 13.2: Wykres funkcji $\delta \mapsto \frac{M}{2}\delta^2 - \delta$.

Po prostym przekształceniu dostajemy tezę.

Zajmijmy się regułą minimalizacji ograniczonej. Wówczas $\mathbf{x}_{k+1}$ realizuje minimum dla $\alpha \in [0, A)$. Minimum po prawej stronie w (13.8) przy ograniczeniu $\delta \in [0, A)$ realizowane jest przez $\delta = \gamma$. Postępując teraz identycznie jak powyżej dostajemy tezę. □

Podsumujmy. Na mocy lematu 13.2 wiemy, że warunek końca sformułowany jako ograniczenie na normę gradientu funkcji f jest poprawny i daje ograniczenia zarówno na odległość przybliżenia $\mathbf{x}_{k+1}$ od punktu minimum, jak i na dokładność wyznaczenia wartości minimalnej funkcji. Zwróćmy uwagę, że czym "bardziej" ściśle wypukła jest funkcja na otoczeniu $\bar{\mathbf{x}}$, tj. czym większe m, tym ostrzejsza zależność między normą gradientu a odległością od punktu $\bar{\mathbf{x}}$. Lemat 13.3 sugeruje, że zbieżność wartości funkcji jest najszybsza, jeśli funkcja podobnie zachowuje się we wszystkich kierunkach, czyli wartości własne macierzy drugich pochodnych leżą dość blisko siebie. Wówczas iloraz m/M jest największy, co korzystnie wpływa na współczynnik kontrakcji $1 - m/(2M)$. A zatem algorytm największego spadku, podobnie jak wszystkie inne metody spadkowe, będzie najlepiej pracował na takich funkcjach, które mają stosunkowo duży iloraz m/M, lub inaczej, m jest tego samego rzędu co M.

Dla reguły Armijo poprawność warunku stopu algorytmu pozostaje w mocy, gdyż Lemat 13.2 nie zależy od reguły wyboru długości kroku. Teza lematu 13.3 wymaga pewnych zmian, pozostawiając jednak wykładniczą zbieżność (patrz zadanie 13.3).

Jakie zatem są zalety reguły Armijo? Otóż, jak wspomniane zostało wyżej, jest ona prosta w implementacji i nie wymaga stosowania metod optymalizacji funkcji jednej zmiennej. Koszt tego uproszczenia nie jest zwykle również duży, lecz zależy od parametrów s, β, σ. Nie ma niestety reguł doboru tych parametrów – pozostawia się to doświadczeniu i intuicji użytkownika.

13.2 Metoda Newtona

Metody spadkowe, które opisaliśmy dotychczas, poszukiwały kierunku kolejnego przybliżenia minimum funkcji f bazując na rozwinięciu pierwszego rzędu

$$f(\mathbf{x} + \mathbf{d}) \approx f(\mathbf{x}) + Df(\mathbf{x})\mathbf{d}.$$

W metodzie Newtona kierunek poszukiwań wybieramy w oparciu o przybliżenie drugiego rzędu w rozwinięciu Taylora funkcji f. Załóżmy, że funkcja $f : \mathbb{R}^n \to \mathbb{R}$ jest dwukrotnie różniczkowalna. Przybliżamy ją wokół punktu $\mathbf{x}$ za pomocą wielomianu Taylora drugiego stopnia

$$f(\mathbf{x} + \mathbf{d}) \approx f(\mathbf{x}) + Df(\mathbf{x})\mathbf{d} + \frac{1}{2}\mathbf{d}^T D^2 f(\mathbf{x})\mathbf{d}.$$

Zamiast minimalizować funkcję f szukamy minimum jej przybliżenia, czyli prawej strony powyższego wyrażenia. Aby miało to sens musimy założyć, że hesjan $D^2f(\mathbf{x})$ jest dodatnio określony. Ponieważ $f(\mathbf{x})$ jest ustalone, problem sprowadza się do następującego zadania:

$$\begin{cases} h(\mathbf{d}) = \frac{1}{2}\mathbf{d}^T D^2 f(\mathbf{x})\mathbf{d} + Df(\mathbf{x})\mathbf{d} \to \min, \\ \mathbf{d} \in \mathbb{R}^n. \end{cases}$$

Ponieważ założyliśmy, że hesjan $D^2f(\mathbf{x})$ jest dodatnio określony, to zadanie to posiada rozwiązanie dane wzorem

$$\mathbf{d} = -\left(D^2 f(\mathbf{x})\right)^{-1}\left(Df(\mathbf{x})\right)^T.$$

Zauważmy, że jeśli w punkcie $\mathbf{x}$ zeruje się pochodna $Df(\mathbf{x})$, to $\mathbf{d} = \mathbf{0}$, czyli jesteśmy w punkcie krytycznym funkcji f. Algorytm metody Newtona wygląda następująco:

Inicjalizacja: Wybierz punkt początkowy $\mathbf{x}_1$ i dokładność $\varepsilon > 0$.

Krok k**-ty**: $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$, gdzie $\mathbf{d}_k = -\left(D^2 f(\mathbf{x}_k)\right)^{-1}\left(Df(\mathbf{x}_k)\right)^T$.

Koniec: gdy $\|Df(\mathbf{x}_{k+1})\| \leqslant \varepsilon$.

Powyższy algorytm niesie ze sobą wiele wątpliwości. Po pierwsze, nie jest wcale jasne, czy ciąg $(\mathbf{x}_k)$ ma granicę. Co więcej, łatwo znaleźć niezdegenerowany przykład, żeby był on rozbieżny. Rozumiemy przez to, że funkcja przyjmuje minimum globalne w pewnym punkcie, lecz nieodpowiedni wybór punktu początkowego $\mathbf{x}_1$ może spowodować, że ciąg $(\mathbf{x}_k)$ zbiega do nieskończoności. W poniższym twierdzeniu pokazujemy warunek dostateczny, aby taki przypadek nie miał miejsca.

Twierdzenie 13.3. *Załóżmy, że f jest funkcją klasy C^3 na otoczeniu minimum lokalnego $\bar{\mathbf{x}}$ oraz hesjan $D^2f(\bar{\mathbf{x}})$ jest dodatnio określony. Wówczas istnieją stałe c i $\delta > 0$, takie że dla $\mathbf{x}_k \in B(\bar{\mathbf{x}}, \delta)$ mamy*

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}\| \leqslant c\|\mathbf{x}_k - \bar{\mathbf{x}}\|^2.$$

Dowód. Z ciągłości D^2f na otoczeniu $\bar{\mathbf{x}}$ oraz z dodatniej określoności $D^2f(\bar{\mathbf{x}})$ wynika istnienie $\delta > 0$, takiej że dla $\mathbf{x} \in B(\bar{\mathbf{x}}, \delta)$ normy $\|D^2f(\mathbf{x})\|$ oraz $\|\left(D^2f(\mathbf{x})\right)^{-1}\|$ są ograniczone oraz większe niż $r > 0$.

Rozwijając gradient funkcji f we wzór Taylora w otoczeniu punktu $\mathbf{x}_k$ mamy

$$\left(Df(\mathbf{x}_k + \mathbf{h})\right)^T = \left(Df(\mathbf{x}_k)\right)^T + D^2f(\mathbf{x}_k)\mathbf{h} + O(\|\mathbf{h}\|^2).$$

Biorąc w powyższej równości $\mathbf{h} = -\mathbf{h}_k$, gdzie $\mathbf{h}_k = (\mathbf{x}_k - \bar{\mathbf{x}})$, otrzymamy

$$\left(Df(\mathbf{x}_k)\right)^T - D^2f(\mathbf{x}_k)\mathbf{h}_k + O(\|\mathbf{h}_k\|^2) = \left(Df(\bar{\mathbf{x}})\right)^T = \mathbf{0}.$$

Niech $\mathbf{x}_k \in B(\bar{\mathbf{x}}, \delta)$. Mnożąc powyższe równanie przez $\left(D^2f(\mathbf{x}_k)\right)^{-1}$ dostajemy

$$\mathbf{0} = \left(D^2 f(\mathbf{x}_k)\right)^{-1}\left(Df(\mathbf{x}_k)\right)^T - \mathbf{h}_k + O(\|\mathbf{h}_k\|^2) = -\mathbf{d}_k - \mathbf{h}_k + O(\|\mathbf{h}_k\|^2) = -\mathbf{h}_{k+1} + O(\|\mathbf{h}_k\|^2). \tag{13.9}$$

Ostatnia równość wynika z następującego ciągu tożsamości

$$-\mathbf{d}_k - \mathbf{h}_k = \mathbf{x}_k - \mathbf{x}_{k+1} - (\mathbf{x}_k - \bar{\mathbf{x}}) = -(\mathbf{x}_{k+1} - \bar{\mathbf{x}}) = -\mathbf{h}_{k+1}.$$

Z równania (13.9) wynika istnienie stałej $c > 0$, dla której

$$\|\mathbf{h}_{k+1}\| \leqslant c\|\mathbf{h}_k\|^2,$$

czyli kwadratowa szybkość zbieżności algorytmu.

Jeśli $\mathbf{x}_k \in B(\bar{\mathbf{x}}, \frac{\alpha}{c})$ dla $\alpha \in (0,1)$, takiego że $\frac{\alpha}{c} \leqslant \delta$, to z ostatniego oszacowania dostajemy

$$\|\mathbf{h}_{k+1}\| \leqslant c\frac{\alpha}{c}\|\mathbf{h}_k\| = \alpha\|\mathbf{h}_k\|,$$

co oznacza geometryczną zbieżność ciągu $\mathbf{h}_k$ do zera. □

Uwaga 13.1.

1. Zauważmy, że jeśli punkt $\mathbf{x}_k$ jest daleko od rozwiązania $\bar{\mathbf{x}}$ to hesjan w punkcie $\mathbf{x}_k$ może nie być dodatni.
2. Jeśli hesjan jest dodatnio określony w punkcie $\mathbf{x}_k$, to wyznaczony w tym punkcie kierunek $\mathbf{d}_k$ jest kierunkiem spadku funkcji f, tzn. $Df(\mathbf{x}_k)\mathbf{d}_k < 0$, bo $\mathbf{d}_k = -\left(D^2 f(\mathbf{x}_k)\right)^{-1}\left(Df(\mathbf{x}_k)\right)^T$, czyli
$$Df(\mathbf{x}_k)\mathbf{d}_k = -Df(\mathbf{x}_k)\left(D^2 f(\mathbf{x}_k)\right)^{-1}\left(Df(\mathbf{x}_k)\right)^T < 0.$$
3. Nawet kiedy hesjan jest dodatnio określony w punkcie $\mathbf{x}_k$, to nie ma pewności, że $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$, gdyż w metodzie Newtona nie ma minimalizacji kierunkowej. Oznacza to, że poruszamy się w kierunku spadku, ale wybrany krok $\mathbf{d}_k$ może być za długi i wtedy może być $f(\mathbf{x}_{k+1}) = f(\mathbf{x}_k + \mathbf{d}_k) > f(\mathbf{x}_k)$.

Bolączką metody Newtona jest to, iż zbiega ona do punktu krytycznego, który może również wyznaczać maksimum lokalne lub punkt przegięcia. Założenie pseudowypukłości funkcji f gwarantuje, że punkt krytyczny jest minimum. W pozostałych przypadkach należy zbadać zachowanie funkcji w otoczeniu znalezionego numerycznie punktu krytycznego i na tej podstawie ustalić, czy jest w nim minimum, maksimum lub punkt przegięcia.

Zwróćmy uwagę, że wszystkie matematyczne wyniki dla metody Newtona zakładają dodatniość hesjanu, a zatem ścisłą wypukłość. W pozostałych przypadkach ocena wyników i dobór parametrów $\mathbf{x}_1, \varepsilon$ pozostaje powierzyć intuicji i doświadczeniom. Nie przeszkadza to wszakże w powszechnym stosowaniu metody Newtona i jej różnych modyfikacji. Zainteresowany czytelnik znajdzie dużo więcej informacji w monografii D. Bertsekasa [4].

13.3 Metoda kierunków i gradientów sprzężonych

Definicja 13.1. Niech dana będzie dodatnio określona symetryczna macierz H wymiaru $n \times n$. Niezerowe wektory $\mathbf{d}_1, \ldots, \mathbf{d}_n$ nazywamy sprzężonymi względem macierzy H, jeśli

$$\mathbf{d}_i^T H \mathbf{d}_j = 0, \quad \text{dla } i \neq j, \quad i, j = 1, \ldots, n.$$

Łatwo jest zauważyć, że jeśli wektory $\mathbf{d}_1, \ldots, \mathbf{d}_n$ są sprzężone względem macierzy H, to są one liniowo niezależne.

Rozważmy teraz funkcję kwadratową $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T H\mathbf{x} + b^T\mathbf{x} + c$ z symetryczną dodatnio określoną macierzą H. Znalezienie minimum tej funkcji metodą *kierunków sprzężonych* polega na znajdowaniu kolejnych przybliżeń rozwiązania metodą dokładnej minimalizacji w kierunkach będących kierunkami sprzężonymi względem macierzy H. Niech $\mathbf{d}_1, \ldots, \mathbf{d}_n$ będą tymi kierunkami sprzężonymi, wtedy algorytm metody kierunków sprzężonych przebiega następująco:

Inicjalizacja: Wybierz punkt początkowy $\mathbf{x}_1$.

Krok k-ty:

1. Wybierz kierunek $\mathbf{d}_k$.
2. Połóż $\mathbf{x}_{k+1} = \mathbf{x}_k + t_k\mathbf{d}_k$, gdzie t_k znajdowany jest z warunku dokładnej minimalizacji $t_k = \operatorname{Argmin}\{f(\mathbf{x}_k + t\mathbf{d}_k) : t \geqslant 0\}$.

Koniec: gdy $\|Df(\mathbf{x}_{k+1})\| = 0$.

Twierdzenie 13.4. *Metoda kierunków sprzężonych z dokładną minimalizacją zastosowana do funkcji kwadratowej $f : \mathbb{R}^n \to \mathbb{R}$ z dodatnio określonym hesjanem H wyznacza minimum po co najwyżej n iteracjach.*

Dowód. Zauważmy, że funkcja kwadratowa ma postać $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T H\mathbf{x} + b^T\mathbf{x} + c$.
Wprowadźmy oznaczenie $\mathbf{g}_k = Df(\mathbf{x}_k)$. Łatwo zauważyć następującą tożsamość

$$\mathbf{g}_{k+1}\mathbf{d}_k = 0, \quad k = 1, \ldots, n-1. \tag{13.10}$$

Wynika ona z rozważenia funkcji $h(t) = f(\mathbf{x}_k + t\mathbf{d}_k)$. Funkcja h osiąga minimum w punkcie t_k (warunek dokładnej minimalizacji), więc $h'(t_k) = 0$. Z drugiej strony $h'(t) = Df(\mathbf{x}_k + t\mathbf{d}_k)\mathbf{d}_k$, czyli $h'(t_k) = \mathbf{g}_{k+1}\mathbf{d}_k$.
Z postaci funkcji kwadratowej f dostajemy związek

$$\mathbf{g}_{k+1} - \mathbf{g}_k = Df(\mathbf{x}_{k+1}) - Df(\mathbf{x}_k) = (H\mathbf{x}_{k+1})^T - (H\mathbf{x}_k)^T = t_k\mathbf{d}_k^T H. \tag{13.11}$$

Wykorzystując ten związek udowodnimy indukcyjnie równość

$$\mathbf{g}_{k+1}\mathbf{d}_j = 0, \quad \text{dla } k = 1, \ldots, n-1,\ j = 1, \ldots, k. \tag{13.12}$$

Dla $k = 1$ równość ta została już udowodniona (jako równanie (13.10)). Załóżmy teraz, że zachodzą równości $\mathbf{g}_k\mathbf{d}_j = 0$ dla $j = 1, \ldots, k-1$. Wtedy mamy

$$\mathbf{g}_{k+1}\mathbf{d}_j = (\mathbf{g}_k + t_k\mathbf{d}_k^T H)\mathbf{d}_j = \mathbf{g}_k\mathbf{d}_j + t_k\mathbf{d}_k^T H\mathbf{d}_j = 0,$$

dla $j = 1, \ldots, k-1$, bo $\mathbf{g}_k\mathbf{d}_j = 0$ z założenia indukcyjnego, a $\mathbf{d}_k^T H\mathbf{d}_j = 0$ bo są to kierunki sprzężone. Ponieważ $\mathbf{g}_{k+1}\mathbf{d}_k = 0$ z równania (13.10), więc równość (13.12) została udowodniona.
Z równości (13.12) wynika, że funkcja f ma w punkcie $\mathbf{x}_{k+1}$ zerowe pochodne kierunkowe w kierunkach $\mathbf{d}_1, \ldots, \mathbf{d}_k$

$$Df(\mathbf{x}_{k+1})\mathbf{d} = 0, \quad \mathbf{d} \in \operatorname{Lin}(\mathbf{d}_1, \ldots, \mathbf{d}_k).$$

Niech

$$K_k = \mathbf{x}_{k+1} + \operatorname{Lin}(\mathbf{d}_1, \ldots, \mathbf{d}_k) = \mathbf{x}_1 + \operatorname{Lin}(\mathbf{d}_1, \ldots, \mathbf{d}_k).$$

Niech $F = f|_{K_k}$. Funkcja F jest wypukła, bo funkcja f jest wypukła oraz zbiór K_k jest wypukły. Ponieważ zerowanie się wszystkich pochodnych kierunkowych jest warunkiem wystarczającym dla istnienia minimum funkcji wypukłej to

$$\mathbf{x}_{k+1} = \operatorname{Argmin}\{F(\mathbf{x}) : \mathbf{x} \in K_k\} = \operatorname{Argmin}\{f(\mathbf{x}) : \mathbf{x} \in \mathbf{x}_1 + \operatorname{Lin}(\mathbf{d}_1, \ldots, \mathbf{d}_k)\}.$$

□

Opisana w poprzednim twierdzeniu metoda znajdowania minimum funkcji kwadratowej jest bardzo efektywna (znajdujemy minimum co najwyżej w n krokach). Jej słabością jest jednak konieczność wyznaczenia na początku całego zbioru kierunków sprzężonych (można to oczywiście zrobić znajdując wektory własne hesjanu H). Opisana poniżej metoda *gradientów sprzężonych* pozwala wyznaczać kierunki sprzężone sukcesywnie w trakcie obliczania kolejnych kroków algorytmu iteracyjnego. Metoda ta nazywa się *metodą gradientów sprzężonych Fletchera-Reevesa.*

Algorytm gradientów sprzężonych Fletchera-Reevesa:

Inicjalizacja: Wybieramy punkt początkowy $\mathbf{x}_1$.

Krok pierwszy: $\mathbf{d}_1 = -\mathbf{g}_1^T$, czyli wybieramy kierunek największego spadku;

Krok k-ty:

1. Znamy już kierunki $\mathbf{d}_1, \dots, \mathbf{d}_{k-1}$;
2. Wybieramy kierunek $\mathbf{d}_k = -\mathbf{g}_k^T + \beta_{k-1}\mathbf{d}_{k-1}$;
3. Dobieramy $\beta_{k-1} \in \mathbb{R}$, tak aby kierunek $\mathbf{d}_k$ był sprzężony do $\mathbf{d}_1, \dots, \mathbf{d}_{k-1}$,
$$\beta_{k-1} = \frac{\mathbf{g}_k \mathbf{g}_k^T}{\mathbf{g}_{k-1}\mathbf{g}_{k-1}^T}. \tag{13.13}$$
4. Obliczamy punkt kolejnej iteracji $\mathbf{x}_{k+1} = \mathbf{x}_k + t_k\mathbf{d}_k$, gdzie $t_k = \operatorname{Argmin}\{f(\mathbf{x}_k + t\mathbf{d}_k) : t \geqslant 0\}$.

Koniec: gdy $\|Df(\mathbf{x}_{k+1})\| = 0$.

Twierdzenie 13.5. *Metoda gradientów sprzężonych Fletchera-Reevesa z dokładną minimalizacją zastosowana do funkcji kwadratowej $f : \mathbb{R}^n \to \mathbb{R}$ z dodatnio określonym hesjanem H wyznacza kierunki sprzężone względem H*

$$\mathbf{d}_i^T H \mathbf{d}_j = 0, \quad i = 1, \dots, n, \ j = 1, \dots, i-1. \tag{13.14}$$

Ponadto dla $i \leqslant m = \max\{i : \mathbf{g}_i \neq \mathbf{0}^T\}$ zachodzą równości

$$\mathbf{g}_i \mathbf{g}_j^T = 0, \quad j = 1, \dots, i-1 \tag{13.15}$$

oraz

$$\mathbf{g}_i \mathbf{d}_i = -\mathbf{g}_i \mathbf{g}_i^T, \quad i = 1, \dots, n. \tag{13.16}$$

Dowód. Jeśli $m = 0$, to punkt startowy $\mathbf{x}_1$ jest rozwiązaniem zadania minimalizacji. Niech więc $m \geqslant 1$. Dowód przeprowadzimy przez indukcję względem i.

Dla $i = 1$ tylko równość (13.16) wymaga dowodu. Jest ona oczywista z uwagi na inicjalizację algorytmu kierunkiem najszybszego spadku.

Przypuśćmy, że równości w twierdzeniu zachodzą dla pewnego $i < m$. Pokażemy, że zachodzą one dla $i + 1$.

Z dowodu twierdzenia 13.4 wiemy, że dla funkcji kwadratowej f zachodzą równości

$$\mathbf{g}_{i+1} - \mathbf{g}_i = Df(\mathbf{x}_{i+1}) - Df(\mathbf{x}_i) = \big(H(\mathbf{x}_{i+1} - \mathbf{x}_i)\big)^T = t_i \mathbf{d}_i^T H. \tag{13.17}$$

Wykorzystując powyższą tożsamość oraz równość (13.10) dostajemy

$$0 = \mathbf{g}_{i+1}\mathbf{d}_i = \mathbf{g}_i\mathbf{d}_i + t_i \mathbf{d}_i^T H \mathbf{d}_i,$$

co pozwala wyznaczyć t_i

$$t_i = -\frac{\mathbf{g}_i\mathbf{d}_i}{\mathbf{d}_i^T H\mathbf{d}_i} = \frac{\mathbf{g}_i\mathbf{g}_i^T}{\mathbf{d}_i^T H\mathbf{d}_i}, \tag{13.18}$$

gdzie w ostatniej równości skorzystaliśmy z założenia indukcyjnego i równości (13.16).

Zatem dla $j < i$ mamy na mocy założenia indukcyjnego i kroku 2 algorytmu

$$\begin{aligned}\mathbf{g}_{i+1}\mathbf{g}_j^T =& \mathbf{g}_i\mathbf{g}_j^T + t_i\mathbf{d}_i^T H\mathbf{g}_j^T = \mathbf{g}_i\mathbf{g}_j^T - t_i\mathbf{d}_i^T H(\mathbf{d}_j - \beta_{j-1}\mathbf{d}_{j-1}) = \\ & \mathbf{g}_i\mathbf{g}_j^T - t_i\mathbf{d}_i^T H\mathbf{d}_j + t_i\beta_{j-1}\mathbf{d}_i^T H\mathbf{d}_{j-1} = 0,\end{aligned}$$

bo pierwszy składnik ostatniej sumy jest równy zero z założenia indukcyjnego i równości (13.15) a drugi i trzeci składnik sumy, z założenia indukcyjnego i równości (13.14).

Dla $j = i$ mamy po podobnych przekształceniach, korzystając z założenia indukcyjnego oraz równości (13.14) i (13.18)

$$\begin{aligned}\mathbf{g}_{i+1}\mathbf{g}_j^T = \mathbf{g}_{i+1}\mathbf{g}_i^T =& \mathbf{g}_i\mathbf{g}_i^T - t_i\mathbf{d}_i^T H\mathbf{d}_i + t_i\beta_{i-1}\mathbf{d}_i^T H\mathbf{d}_{i-1} = \\ & \mathbf{g}_i\mathbf{g}_i^T - \frac{\mathbf{g}_i\mathbf{g}_i^T}{\mathbf{d}_i^T H\mathbf{d}_i}\mathbf{d}_i^T H\mathbf{d}_i = 0.\end{aligned}$$

Udowodniliśmy więc, że

$$\mathbf{g}_{i+1}\mathbf{g}_j^T = 0, \quad \text{dla } j = 1, \ldots, i, \tag{13.19}$$

co jest krokiem indukcyjnym dowodu równości (13.15).

Obecnie przejdziemy do dowodu równości (13.14) czyli pokazania, że wyznaczone przez algorytm kierunki są kierunkami sprzężonymi. Korzystając z równości (13.17) oraz algorytmu wyznaczania $\mathbf{d}_{i+1}$ dostajemy

$$\begin{aligned}\mathbf{d}_{i+1}^T H\mathbf{d}_j =& -\mathbf{g}_{i+1}H\mathbf{d}_j + \beta_i\mathbf{d}_i^T H\mathbf{d}_j = \frac{1}{t_j}\mathbf{g}_{i+1}(\mathbf{g}_j - \mathbf{g}_{j+1})^T + \beta_i\mathbf{d}_i^T H\mathbf{d}_j = \\ & -\frac{1}{t_j}\mathbf{g}_{i+1}\mathbf{g}_{j+1}^T + \beta_i\mathbf{d}_i^T H\mathbf{d}_j.\end{aligned} \tag{13.20}$$

Dla $j < i$ dostajemy natychmiast

$$\mathbf{d}_{i+1}^T H\mathbf{d}_j = 0,$$

bo $\mathbf{g}_{i+1}\mathbf{g}_{j+1}^T = 0$ z równości (13.19) a $\mathbf{d}_i^T H\mathbf{d}_j = 0$ z założenia indukcyjnego i równości (13.14).

Dla $j = i$, korzystając z wzoru na t_i oraz wzoru (13.13) na β_i, dostajemy z równania (13.20)

$$\begin{aligned}\mathbf{d}_{i+1}^T H\mathbf{d}_i =& -\frac{1}{t_i}\mathbf{g}_{i+1}\mathbf{g}_{i+1}^T + \frac{\mathbf{g}_{i+1}\mathbf{g}_{i+1}^T}{\mathbf{g}_i\mathbf{g}_i^T}\mathbf{d}_i^T H\mathbf{d}_i = \\ & -\frac{1}{t_i}\mathbf{g}_{i+1}\mathbf{g}_{i+1}^T + \frac{\mathbf{g}_{i+1}\mathbf{g}_{i+1}^T}{\mathbf{g}_i\mathbf{g}_i^T}\frac{\mathbf{g}_i\mathbf{g}_i^T}{t_i} = 0.\end{aligned}$$

Otrzymaliśmy więc dowód, że

$$\mathbf{d}_{i+1}^T H\mathbf{d}_j = 0, \quad \text{dla } j = 1, \ldots, i.$$

Pozostaje nam jeszcze dowód równości (13.16). Korzystając z algorytmu wyznaczania $\mathbf{d}_{i+1}$ oraz równości (13.10) wynikającej ze stosowania dokładnej minimalizacji dostajemy

$$\mathbf{g}_{i+1}\mathbf{d}_{i+1} = \mathbf{g}_{i+1}(-\mathbf{g}_{i+1}^T + \beta_i\mathbf{d}_i) = -\mathbf{g}_{i+1}\mathbf{g}_{i+1}^T + \beta_i\mathbf{g}_{i+1}\mathbf{d}_i = -\mathbf{g}_{i+1}\mathbf{g}_{i+1}^T.$$

Tym samym udowodniliśmy, że

$$\mathbf{g}_{i+1}\mathbf{d}_{i+1} = -\mathbf{g}_{i+1}\mathbf{g}_{i+1}^T,$$

czyli krok indukcyjny w dowodzie równości (13.16). □

Kiedy funkcja f nie jest funkcją kwadratową, metoda gradientów sprzężonych Fletchera-Reevesa musi ulec pewnej modyfikacji:

Inicjalizacja: Wybieramy punkt początkowy $\mathbf{x}_1$ oraz dokładność $\varepsilon > 0$.

Krok pierwszy: $\mathbf{d}_1 = -\mathbf{g}_1^T$, czyli wybieramy kierunek największego spadku;

Krok k-ty:

1. Znamy już kierunki $\mathbf{d}_1, \dots, \mathbf{d}_{k-1}$;
2. Wybieramy kierunek $\mathbf{d}_k = -\mathbf{g}_k^T + \beta_{k-1}\mathbf{d}_{k-1}$;
3. Dobieramy $\beta_{k-1} \in \mathbb{R}$, tak aby kierunek $\mathbf{d}_k$ był sprzężony do $\mathbf{d}_1, \dots, \mathbf{d}_{k-1}$,
$$\beta_{k-1} = \frac{\mathbf{g}_k \mathbf{g}_k^T}{\mathbf{g}_{k-1}\mathbf{g}_{k-1}^T}. \tag{13.21}$$
4. Obliczamy punkt kolejnej iteracji $\mathbf{x}_{k+1} = \mathbf{x}_k + t_k\mathbf{d}_k$, gdzie $t_k = \operatorname{Argmin}\{f(\mathbf{x}_k + t\mathbf{d}_k) : t \geqslant 0\}$.

Koniec: gdy $\|Df(\mathbf{x}_{k+1})\| \leqslant \varepsilon$.

Ponieważ wyznaczanie β_i ze wzoru (13.21) nie gwarantuje, że otrzymane kierunki będą kierunkami spadku, stosuje się więc różne modyfikacje tego wzoru. Zalecana w literaturze modyfikacja dająca lepsze wyniki niż oryginalna metoda Fletchera-Reevesa wygląda następująco

$$\beta_{k-1} = \frac{\mathbf{g}_k(\mathbf{g}_k - \mathbf{g}_{k-1})^T}{\mathbf{g}_{k-1}\mathbf{g}_{k-1}^T}.$$

13.4 Zadania

Ćwiczenie 13.1. Udowodnij, że jeśli $\mathbf{d}$ jest kierunkiem spadku w punkcie $\mathbf{x}$, to istnieje $\delta > 0$, taka że dla $\alpha \in (0, \delta)$ mamy
$$f(\mathbf{x} + \alpha\mathbf{d}) < f(\mathbf{x}).$$

Ćwiczenie 13.2. Wykaż, że jeśli $\mathbf{d}_k \neq 0$, to wartość α_k z reguły Armijo można wyznaczyć w skończonej liczbie kroków (choć różnej dla każdego $\mathbf{x}_k$).

Wskazówka. Skorzystaj z definicji pochodnej na początku rozdziału 2. Zastanów się, dlaczego teza nie jest poprawna dla $\sigma \geqslant 1$.

Ćwiczenie 13.3. Zmodyfikuj tezę i dowód lematu 13.3 tak, aby stosował się on do metody Armijo.

Ćwiczenie 13.4. Udowodnij, że w metodzie największego spadku z regułą minimalizacji (bez ograniczenia) kolejne kierunki $\mathbf{d}_k$ są do siebie prostopadłe.

Ćwiczenie 13.5. Udowodnij, że teza lematu 13.2 może zostać wzmocniona bez zmiany założeń:
$$f(\mathbf{x}) - f(\bar{\mathbf{x}}) \leqslant \frac{1}{2m}\|Df(\mathbf{x})\|^2.$$

Wskazówka. Przeprowadź dowód bez użycia oszacowania na $|\mathbf{x}-\bar{\mathbf{x}}|$. Skorzystaj oczywiście z twierdzenia Taylora.

Ćwiczenie 13.6. Opracuj szybki algorytm minimalizowania funkcji wypukłej $f : \mathbb{R} \to \mathbb{R}$ klasy C^1.

Ćwiczenie 13.7. Podaj przykład funkcji $f : \mathbb{R} \to \mathbb{R}$ oraz punktu startowego x_0, takich żeby ciąg generowany przez metodę Newtona nie miał granicy. Dodatkowo wymagamy, aby funkcja f miała minimum globalne w 0.

Wskazówka. Rozważ funkcję malejącą przy x dążącym do nieskończoności.

Rozdział 14

Metody optymalizacji z ograniczeniami

W tym rozdziale skupimy się na metodach numerycznych rozwiązywania problemów optymalizacyjnych z ograniczeniami nierównościowymi. Pokażemy problemy z zastosowaniem metody największego spadku i jej naiwnych modyfikacji oraz zaproponujemy skuteczne, aczkolwiek bardziej skomplikowane podejście.

Problem optymalizacyjny na następującą postać:

$$\begin{cases} f(\mathbf{x}) \to \min, \\ g_i(\mathbf{x}) \leqslant 0, \quad i = 1, \ldots, m, \\ \mathbf{x} \in \mathbb{R}^n, \end{cases} \tag{14.1}$$

gdzie $f, g_1, \ldots, g_m : \mathbb{R}^n \to \mathbb{R}$. A zatem zbiór punktów dopuszczalnych jest zadany przez

$$W = \{\mathbf{x} \in \mathbb{R}^n : g_1(\mathbf{x}) \leqslant 0, \ldots, g_m(\mathbf{x}) \leqslant 0\}. \tag{14.2}$$

W rozważaniach tego rozdziału będziemy odwoływać się często do pojęcia zbioru *kierunków dopuszczalnych* zdefiniowanego następująco

$$F(\mathbf{x}) = \{\mathbf{d} \in \mathbb{R}^n : \ \mathbf{d} \neq \mathbf{0} \text{ oraz istnieje } \lambda^* > 0 \text{ taka że } \mathbf{x} + \lambda\mathbf{d} \in W \ \ \forall\, \lambda \in [0, \lambda^*]\}.$$

14.1 Algorytm Zoutendijka dla ograniczeń afinicznych

Rozważmy bardzo prostą modyfikację algorytmu największego spadku. Otóż jeśli punkt znajduje się wewnątrz zbioru W, to istnieje możliwość poruszania się wzdłuż kierunku największego spadku aż do uderzenia w brzeg. Jeśli punkt już jest na brzegu, to naturalne jest wybrać taki kierunek ruchu, by pozwalał on na jak największy spadek wartości funkcji celu, a jednocześnie na ruch w jego kierunku. Kierunek taki nazywamy *dopuszczalnym kierunkiem spadku* w punkcie $\mathbf{x}$ i definiujemy jako kierunek dopuszczalny $\mathbf{d}$, taki że $Df(\mathbf{x})\mathbf{d} < 0$.

Okazuje się, że pomysł ten działa całkiem dobrze, jeśli ograniczenia są liniowe i nosi nazwę algorytm Zoutendijka. Przypomnijmy, że liniowość ograniczeń znacznie upraszcza problem, co pozwoli nam na rozszerzenie analizy w tym podrozdziale do zagadnień z ograniczeniami nierów-

nościowymi i równościowymi, tzn.

$$\begin{cases} f(\mathbf{x}) \to \min, \\ A\mathbf{x} \leqslant \mathbf{b}, \\ Q\mathbf{x} = \mathbf{a}, \\ \mathbf{x} \in \mathbb{R}^n. \end{cases} \tag{14.3}$$

Tutaj A jest macierzą $m \times n$, Q jest macierzą $l \times n$, zaś $\mathbf{b} \in \mathbb{R}^m$ i $\mathbf{a} \in \mathbb{R}^l$. Poniższy lemat charakteryzuje zbiór dopuszczalnych kierunków spadku. Jego dowód pozostawiamy jako ćwiczenie.

Lemat 14.1. *Niech* $\mathbf{x}$ *będzie punktem dopuszczalnym dla zagadnienia* (14.3) *i załóżmy, że macierz* A *i wektor* $\mathbf{b}$ *mogą być podzielone w zależności od aktywności ograniczeń na* A_1, A_2 *i* $\mathbf{b}_1, \mathbf{b}_2$ *(z dokładnością do przenumerowania ograniczeń), tzn.* $A_1\mathbf{x} = \mathbf{b}_1$ *oraz* $A_2\mathbf{x} < \mathbf{b}_2$. *Wektor* $\mathbf{d} \in \mathbb{R}^n$ *jest kierunkiem dopuszczalnym w* $\mathbf{x}$*, jeśli* $A_1\mathbf{d} \leqslant \mathbf{0}$ *oraz* $Q\mathbf{d} = \mathbf{0}$. *Jeśli, ponadto,* $Df(\mathbf{x})\mathbf{d} < 0$*, to* $\mathbf{d}$ *jest dopuszczalnym kierunkiem spadku.*

Jak wybrać najlepszy dopuszczalny kierunek spadku w punkcie $\mathbf{x}$? Najprościej byłoby rozwiązać zagadnienie

$$Df(\mathbf{x})\mathbf{d} \to \min, \qquad \mathbf{d} \in F(\mathbf{x}), \quad \|\mathbf{d}\| \leqslant 1. \tag{14.4}$$

Ograniczenie na normę wektora $\mathbf{d}$ jest konieczne. Jeśli byśmy je opuścili, to dla dowolnego dopuszczalnego kierunku spadku $\mathbf{d}$ jego wielokrotność $\lambda\mathbf{d}$, $\lambda > 0$, jest również kierunkiem spadku. Co więcej, $Df(\mathbf{x})\mathbf{d} < 0$, czyli $\lim_{\lambda\to\infty} Df(\mathbf{x})\lambda\mathbf{d} = -\infty$ i problem powyższy nie ma jednoznacznego rozwiązania.

Korzystając z rozkładu macierzy A w lemacie 14.1 zagadnienie (14.4) można zapisać jako

$$\begin{cases} Df(\mathbf{x})\mathbf{d} \to \min, \\ A_1\mathbf{d} \leqslant \mathbf{0}, \\ Q\mathbf{d} = \mathbf{0}, \\ \mathbf{d}^T\mathbf{d} \leqslant 1. \end{cases}$$

Zauważmy, że jedyna nieliniowość związana jest z ograniczeniem na normę wektora $\mathbf{d}$. W praktyce, bez większych strat dla jakości algorytmu, zamienia się ją na ograniczenia liniowe, które pozwalają skorzystać z szybkich metod optymalizacji liniowej (np. algorytmu sympleks – patrz monografie Bazaraa, Jarvisa, Shetty [2], Gassa [8] lub Luenbergera [9]). Najpopularniejsze są następujące dwa zamienniki normy euklidesowej $\|\mathbf{d}\|$:

- norma l^∞, tzn. $\sup_j |d_j| \leqslant 1$, co zapisuje się jako
$$-1 \leqslant d_j \leqslant 1, \qquad j = 1, \ldots, n.$$
- norma l^1, tzn. $\sum_j |d_j| \leqslant 1$, co zapisuje się jako
$$\begin{cases} \sum_{j=1}^n \eta_j \leqslant 1, \\ -\eta_j \leqslant d_j \leqslant \eta_j, \qquad j = 1, \ldots, n, \end{cases}$$
gdzie $\eta_1, \ldots, \eta_n$ są nowymi zmiennymi (pomocniczymi).

W dalszej części wykładu rozpatrywać będziemy problem z wykorzystaniem normy l^∞

$$\begin{cases} Df(\mathbf{x})\mathbf{d} \to \min, \\ A_1\mathbf{d} \leqslant \mathbf{0}, \\ Q\mathbf{d} = \mathbf{0}, \\ -1 \leqslant d_j \leqslant 1, \quad j = 1, \ldots, n. \end{cases} \tag{14.5}$$

Pełny algorytm ma wtedy następującą postać:

Inicjalizacja: Wybierz punkt początkowy $\mathbf{x}_1$.

Krok k-ty:

1. Mając dany punkt $\mathbf{x}_k$ dokonaj rozkładu macierzy A na macierze A_1 i A_2 oraz wektora $\mathbf{b}$ na wektory $\mathbf{b}_1$ i $\mathbf{b}_2$, tak aby $A_1\mathbf{x}_k = \mathbf{b}_1$ i $A_2\mathbf{x}_k < \mathbf{b}_2$ (podobnie jak w lemacie 14.1).
2. Wybierz kierunek ruchu $\mathbf{d}_k$ jako rozwiązanie problemu optymalizacyjnego:
$$\begin{cases} Df(\mathbf{x}_k)\mathbf{d} \to \min, \\ A_1\mathbf{d} \leqslant \mathbf{0}, \\ Q\mathbf{d} = \mathbf{0}, \\ -1 \leqslant d_j \leqslant 1, \quad j = 1, \dots, n. \end{cases} \tag{14.6}$$
3. Jeśli $Df(\mathbf{x}_k)\mathbf{d}_k = 0$, to zakończ działanie algorytmu. Punkt $\mathbf{x}_k$ spełnia warunek konieczny pierwszego rzędu. W przeciwnym przypadku kontynuuj.
4. Połóż $\alpha_k = \operatorname{argmin}_{\alpha\in[0,A_k]} f(\mathbf{x}_k+\alpha\mathbf{d}_k)$, gdzie A_k jest największą liczbą o tej własności, że odcinek łączący $\mathbf{x}_k$ i $\mathbf{x}_k + A_k\mathbf{d}_k$ zawarty jest w W.
5. Połóż $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k$.

Przyjrzyjmy się dokładniej wyborowi kierunku $\mathbf{d}_k$. Wektor $\mathbf{d} = \mathbf{0}$ spełnia wszystkie ograniczenia, a zatem optymalna wartość funkcji celu $Df(\mathbf{x}_k)\mathbf{d}_k$ jest co najwyżej równa zeru. Wówczas punkt $\mathbf{x}_k$ spełnia warunek konieczny pierwszego rzędu, czego dowodzimy w następującym lemacie.

Lemat 14.2. *W $\mathbf{x}_k$ spełniony jest warunek konieczny pierwszego rzędu dla problemu (14.3) wtw, gdy rozwiązanie problemu (14.6) spełnia warunek $Df(\mathbf{x}_k)\mathbf{d}_k = 0$.*

Dowód. Przypomnijmy, że w punkcie $\mathbf{x}_k$ jest spełniony warunek konieczny pierwszego rzędu wtw, gdy istnieją wektory $\mu \in [0,\infty)^{m_1}$ i $\lambda \in \mathbb{R}^l$, takie że

$$Df(\mathbf{x}_k) + \mu^T A_1 + \lambda^T Q = \mathbf{0}^T,$$

gdzie macierz A_1 ma wymiar $m_1 \times n$ i odpowiada ograniczeniom aktywnym w punkcie $\mathbf{x}_k$ dla problemu (14.3).

Z lematu Farkasa 5.3 wynika, że jeśli powyższy układ równań posiada rozwiązanie, to układ

$$\begin{cases} Df(\mathbf{x}_k)\mathbf{d} < 0, \\ A_1\mathbf{d} \leqslant \mathbf{0}, \\ Q\mathbf{d} = \mathbf{0}, \end{cases} \tag{14.7}$$

nie posiada rozwiązań. Ponieważ zauważyliśmy wcześniej, że $\mathbf{d} = \mathbf{0}$ jest rozwiązaniem problemu jaki otrzymamy z układu (14.7) zastępując nierówność $Df(\mathbf{x}_k)\mathbf{d} < 0$ równością $Df(\mathbf{x}_k)\mathbf{d} = 0$, to dowodzi to implikacji w prawo.

Aby udowodnić implikację w lewo zauważmy, że jeśli $Df(\mathbf{x}_k)\mathbf{d}_k = 0$, to $\mathbf{d}_k$ nie jest rozwiązaniem układu (14.7). Stosując ponownie lemat Farkasa zauważamy, że spełniony jest wtedy opisany wcześniej warunek konieczny pierwszego rzędu. □

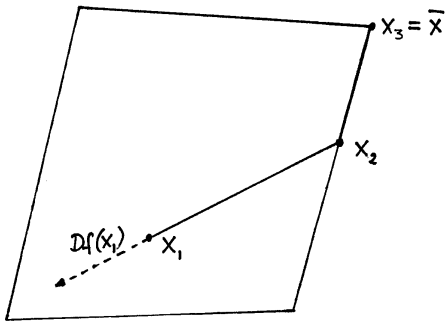
14.2 Algorytm Zoutendijka dla ograniczeń nieliniowych

Zastanówmy się, czy algorytm Zoutendijka działa równie dobrze dla ograniczeń nieliniowych:

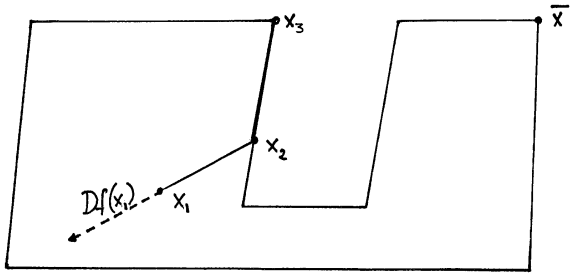
Inicjalizacja: Wybierz punkt początkowy $\mathbf{x}_1$.

Krok k**-ty**:

1. Dany jest punkt $\mathbf{x}_k$.
2. Wybierz kierunek ruchu $\mathbf{d}_k$ jako rozwiązanie problemu optymalizacyjnego:
$$Df(\mathbf{x}_k)\mathbf{d} \to \min, \qquad \mathbf{d} \in F(\mathbf{x}_k), \quad \|\mathbf{d}\| \leqslant 1.$$
3. Jeśli $Df(\mathbf{x}_k)\mathbf{d}_k = 0$, to zakończ działanie algorytmu. Punkt $\mathbf{x}_k$ spełnia warunek konieczny pierwszego rzędu. W przeciwnym przypadku kontynuuj.
4. Połóż $\alpha_k = \operatorname{argmin}_{\alpha\in[0,A_k]} f(\mathbf{x}_k+\alpha\mathbf{d}_k)$, gdzie A_k jest największą liczbą o tej własności, że odcinek łączący $\mathbf{x}_k$ i $\mathbf{x}_k + A_k\mathbf{d}_k$ zawarty jest w W.
5. Połóż $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k$.

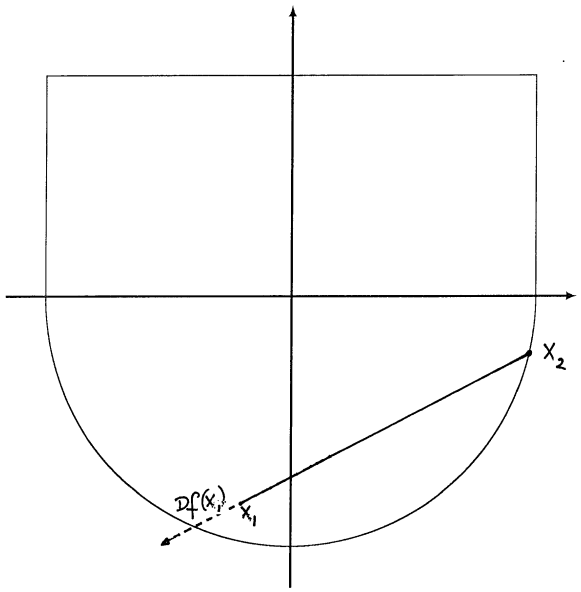


Rysunek 14.1: Ilustracja szybkiej zbieżności metody największego spadku.



Rysunek 14.2: Ilustracja zakleszczenia metody największego spadku dla zbioru niewypukłego.

W poniższych przykładach rozważamy minimalizację funkcji $f(x_1, x_2) = -2x_1 - x_2$ na różnych zbiorach. Punkt minimum oznaczamy zawsze przez $\bar{\mathbf{x}}$. Na rysunku 14.1 widać, że już w kroku drugim osiągamy minimum. Jeśli zbiór W nie jest wypukły algorytm prowadzi nas w kozi róg, z którego już nie możemy się uwolnić, patrz rysunek 14.2. Jest to niestety cecha wszystkich algorytmów tego typu, więc musimy zawsze wymagać, by zbiór punktów dopuszczalnych był wypukły. Czy to już wystarczy? Niestety nie. Na rysunku 14.3 możemy zobaczyć, że nawet



Rysunek 14.3: Ilustracja problemu ze znalezieniem kierunku dopuszczalnego największego spadku.

w przypadku zbioru wypukłego algorytm nie działa. Problem wyboru kierunku $\mathbf{d}_1$ nie ma rozwiązania, gdyż zbiór $F(\mathbf{x}_1)$ nie jest domknięty. Intuicyjnie łatwo jest podać rozwiązanie tego problemu. Należy wybrać taki kierunek $\mathbf{d}_k$, aby nie tylko spadek był jak największy, ale również, żeby dość duży fragment półprostej poprowadzonej w tym kierunku zawierał się w zbiorze W. Co więcej, zależy nam na prostocie, czytaj liniowości, zagadnienia optymalizacyjnego wyznaczającego kierunek $\mathbf{d}_k$. Rozwiązanie podpowiada następujący lemat.

Lemat 14.3. *Niech* $\mathbf{x}$ *będzie punktem dopuszczalnym. Jeśli* f, g_i, $i \in I(\mathbf{x})$ *są różniczkowalne w* $\mathbf{x}$ *a* g_i, $i \notin I(\mathbf{x})$ *są ciągłe w* $\mathbf{x}$, *to kierunek* $\mathbf{d} \in \mathbb{R}^n$ *spełniający* $Df(\mathbf{x})\mathbf{d} < 0$ *i* $Dg_i(\mathbf{x})\mathbf{d} < 0$, $i \in I(\mathbf{x})$, *jest dopuszczalnym kierunkiem spadku.*

Dowód. W pierwszym kroku dowodu pokażemy, że kierunek $\mathbf{d}$ jest kierunkiem dopuszczalnym, tzn. dla dostatecznie małego $\lambda > 0$ punkt $\mathbf{x} + \lambda\mathbf{d} \in W$. Dla ograniczeń nieaktywnych wynika to z ciągłości funkcji g_i, $i \notin I(\mathbf{x})$, bo jeśli $g_i(\mathbf{x}) < 0$, to dla małego λ także $g_i(\mathbf{x} + \lambda\mathbf{d}) < 0$.

Dla ograniczeń aktywnych $i \in I(\mathbf{x})$ mamy

$$g_i(\mathbf{x} + \lambda\mathbf{d}) = g_i(\mathbf{x}) + \lambda Dg_i(\mathbf{x})\mathbf{d} + o(\lambda\mathbf{d}).$$

Ponieważ $g_i(\mathbf{x}) = 0$, $Dg_i(\mathbf{x})\mathbf{d} < 0$ a $o(\lambda\mathbf{d}) \to 0$ dla $\lambda \to 0$, więc $g_i(\mathbf{x} + \lambda\mathbf{d}) < 0$ dla dostatecznie małego λ.

Pokazaliśmy więc, że kierunek $\mathbf{d}$ jest dopuszczalny. Z założeń lematu wynika teraz, że jest to także dopuszczalny kierunek spadku. Zauważmy ponadto, że

$$f(\mathbf{x} + \lambda\mathbf{d}) = f(\mathbf{x}) + \lambda Df(\mathbf{x})\mathbf{d} + o(\lambda\mathbf{d}).$$

Z założeń lematu wynika więc, że $f(\mathbf{x} + \lambda\mathbf{d}) < f(\mathbf{x})$ dla dostatecznie małych λ, czyli wartość funkcji celu w kierunku $\mathbf{d}$ zmniejsza się. □

Lemat 14.3 podaje tylko warunek *dostateczny*. Łatwo znaleźć przykład zagadnienia optymalizacyjnego z ograniczeniami nierównościowymi, dla którego jeden z dopuszczalnych kierunków spadku nie spełnia założeń lematu (ćwiczenie 14.4).

Znalezienie wektora $\mathbf{d} \in \mathbb{R}^n$ spełniającego $Df(\mathbf{x})\mathbf{d} < 0$ i $Dg_i(\mathbf{x})\mathbf{d} < 0$, $i \in I(\mathbf{x})$ sprowadza się do rozwiązania zagadnienia

$$\begin{cases} \max\{Df(\mathbf{x})\mathbf{d},\ Dg_i(\mathbf{x})\mathbf{d},\ i \in I(\mathbf{x})\} \to \min, \\ -1 \leqslant d_j \leqslant 1, \quad j = 1, \ldots, n. \end{cases}$$

Trudną w implementacji funkcję celu można sprowadzić do znacznie prostszej formy problemu optymalizacji liniowej:

$$\begin{cases} \eta \to \min, \\ Df(\mathbf{x})\mathbf{d} \leqslant \eta, \\ Dg_i(\mathbf{x})\mathbf{d} \leqslant \eta, \quad i \in I(\mathbf{x}), \\ -1 \leqslant d_j \leqslant 1, \quad j = 1, \ldots, n. \end{cases} \tag{14.8}$$

Optymalizacji dokonuje się tutaj względem dwóch zmiennych: $\mathbf{d} \in \mathbb{R}^n$ oraz $\eta \in \mathbb{R}$. Zauważmy, że $\eta \leqslant 0$, gdyż para $\mathbf{d} = \mathbf{0}$, $\eta = 0$ rozwiązuje powyższy układ. Jeśli wartość funkcji celu jest mniejsza od zera, to na mocy lematu 14.1 rozwiązanie jest dopuszczalnym kierunkiem spadku. Jeśli $\eta = 0$ jest rozwiązaniem oraz w punkcie $\mathbf{x}$ spełniony jest warunek liniowej niezależności ograniczeń, to w $\mathbf{x}$ zachodzi warunek konieczny pierwszego rzędu. Prawdziwa jest również odwrotna implikacja.

Lemat 14.4. *Jeśli w punkcie dopuszczalnym* $\mathbf{x}$ *spełniony jest warunek liniowej niezależności ograniczeń i rozwiązaniem problemu* (14.8) *jest* $\eta = 0$*, to w* $\mathbf{x}$ *zachodzi warunek konieczny pierwszego rzędu. I odwrotnie, jeśli w* $\mathbf{x}$ *spełniony jest warunek konieczny pierwszego rzędu, to rozwiązaniem* (14.8) *jest* $\eta = 0$ *(nie jest tu konieczne założenie o regularności punktu* $\mathbf{x}$*).*

Dowód. Jeśli $\eta = 0$ jest rozwiązaniem, to układ $A\mathbf{d} < \mathbf{0}$, gdzie A składa się wierszowo z gradientów $Df(\mathbf{x})$ i $Dg_i(\mathbf{x})$, $i \in I(\mathbf{x})$, nie ma rozwiązania. Na mocy lematu 6.3 istnieje $\mathbf{y} \geqslant \mathbf{0}$, $\mathbf{y} \neq \mathbf{0}$, dla którego $A^T\mathbf{y} = \mathbf{0}$. Połóżmy $(\hat{\mu}_0,\ \hat{\mu}_i, i \in I(\mathbf{x})) = \mathbf{y}$ i $\hat{\mu}_i = 0$, $i \notin I(\mathbf{x})$. Równość $A^T\mathbf{y} = \mathbf{0}$ zapisać można w następujący sposób:

$$\hat{\mu}_0 Df(\mathbf{x}) + \sum_{i \in I(\mathbf{x})} \hat{\mu}_i Dg_i(\mathbf{x}) = \mathbf{0}^T.$$

Z założenia o liniowej niezależności gradientów ograniczeń aktywnych wnioskujemy, że $\hat{\mu}_0 \neq 0$. Kładąc $\mu_i = \hat{\mu}_i/\hat{\mu}_0$, $i = 1, \ldots, m$, dostajemy wektor mnożników Lagrange'a z warunku koniecznego pierwszego rzędu.

Aby dowieść implikacji odwrotnej, zauważmy, że jeśli w $\mathbf{x}$ spełniony jest warunek konieczny pierwszego rzędu, to $\mathbf{y} = (1, \mu_i,\ i \in I(\mathbf{x}))$ spełnia następujące warunki: $\mathbf{y} \geqslant \mathbf{0}$, $\mathbf{y} \neq \mathbf{0}$ i $A^T\mathbf{y} = \mathbf{0}$. Na mocy lematu 6.3 nie istnieje $\mathbf{d} \in \mathbb{R}^n$, dla którego $A\mathbf{d} < \mathbf{0}$. A zatem rozwiązaniem 14.8 jest $\eta = 0$. □

Zapiszmy w pełni algorytm zaproponowany przez Zoutendijka dla problemów z nieliniowymi ograniczeniami nierównościowymi:

Inicjalizacja: Wybierz punkt początkowy $\mathbf{x}_1$.

Krok k**-ty**:

1. Dany jest punkt $\mathbf{x}_k$.

2. Wybierz kierunek ruchu $\mathbf{d}_k$ jako rozwiązanie problemu optymalizacyjnego

$$\begin{cases} \eta \to \min, \\ Df(\mathbf{x}_k)\mathbf{d} \leqslant \eta, \\ Dg_i(\mathbf{x}_k)\mathbf{d} \leqslant \eta, \quad i \in I(\mathbf{x}_k), \\ -1 \leqslant d_j \leqslant 1, \quad j = 1, \dots, n. \end{cases}$$

3. Jeśli $\eta = 0$, to zakończ działanie algorytmu. Punkt $\mathbf{x}_k$ spełnia warunek konieczny pierwszego rzędu. W przeciwnym przypadku kontynuuj.
4. Połóż $\alpha_k = \operatorname{argmin}_{\alpha \in [0, A_k]} f(\mathbf{x}_k + \alpha \mathbf{d}_k)$, gdzie A_k jest największą liczbą o tej własności, że odcinek łączący $\mathbf{x}_k$ i $\mathbf{x}_k + A_k \mathbf{d}_k$ zawarty jest w W.
5. Połóż $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$.

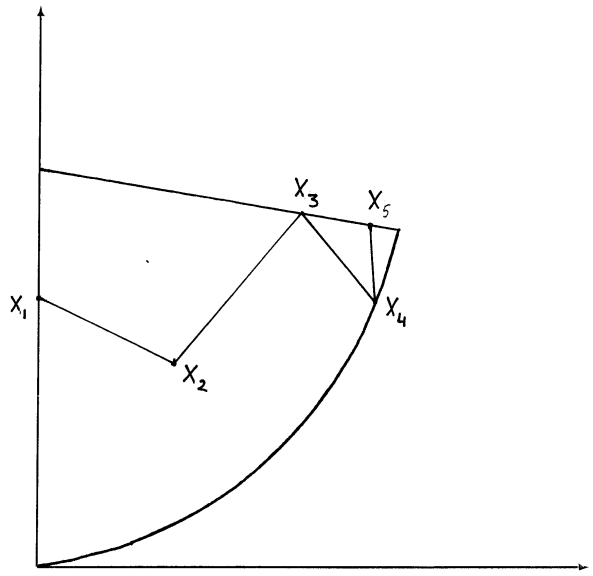
Przykład 14.1. Rozważmy problem optymalizacji z ograniczeniami nieliniowymi:

$$\begin{cases} 2x_1^2 + 2x_2^2 - 2x_1x_2 - 4x_1 - 6x_2 \to \min, \\ x_1 + 5x_2 \leqslant 5, \\ 2x_1^2 - x_2 \leqslant 0, \\ x_1 \geqslant 0, \quad x_2 \geqslant 0. \end{cases}$$

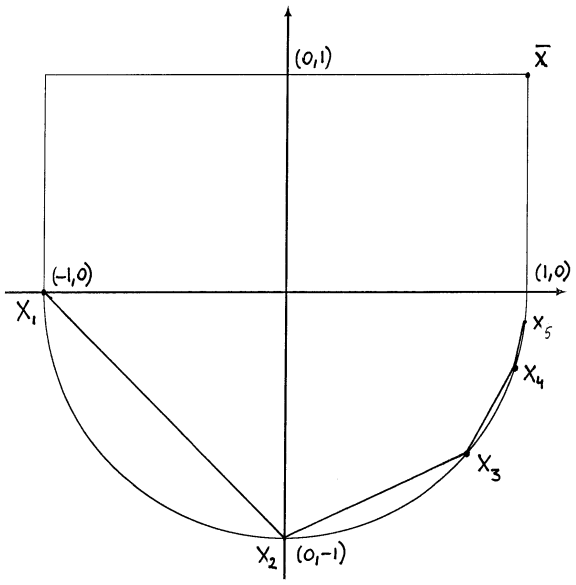
Pokażemy, jak generowane są iteracje metody Zoutendijka. Weźmy punkt startowy $\mathbf{x}_1 = (0, 0.75)^T$. Dostajemy następujący ciąg punktów:

$$\mathbf{x}_2 = (0.2083, 0.5477)^T, \quad \mathbf{x}_3 = (0.5555, 0.8889)^T,$$
$$\mathbf{x}_4 = (0.6479, 0.8397)^T, \quad \mathbf{x}_5 = (0.6302, 0.8740)^T.$$

Widzimy, że ciąg ten dość znacznie oscyluje w zbiorze punktów dopuszczalnych, patrz rys. 14.4. Jest to charakterystyczne zachowanie metody kierunków spadku dla problemów z ograniczeniami.



Rysunek 14.4: Oscylacja ciągu generowanego przez metodę Zoutendijka.



Rysunek 14.5: Ilustracja zbieżności algorytmu Zoutendijka do punktu nie będącego rozwiązaniem.

Algorytm Zoutendijka może polec nawet na dość prostych problemach optymalizacyjnych. Rozważmy minimalizację funkcji liniowej $f(\mathbf{x}) = -2x_1 - x_2$ na zbiorze zaznaczonym na rysunku 14.5. Minimum znajduje się w punkcie $\bar{\mathbf{x}} = (1,1)^T$. Rozpoczynając od punktu $\mathbf{x}_1 = (-1,0)^T$, kolejne iteracje algorytmu Zoutendijka wygenerują ciąg punktów zbiegający do $(1,0)^T$. Algorytm ten nie pozwoli nam na przybliżenie właściwego rozwiązania $\bar{\mathbf{x}}$. Co więcej, wartość funkcji celu w punkcie $(1,0)^T$ wynosi -2, w porównaniu do $f(\bar{\mathbf{x}}) = -3$. W następnym podrozdziale pokażemy jak mała modyfikacja pozwoli poprawić algorytm Zoutendijka.

14.3 Modyfikacja Topkisa-Veinotta

Topkis i Veinott zaproponowali w roku 1967 niewielką modyfikację wyznaczania kierunku $\mathbf{d}_k$ w algorytmie Zoutendijka:

$$\begin{cases} \eta \to \min, \\ Df(\mathbf{x})\mathbf{d} \leqslant \eta, \\ Dg_i(\mathbf{x})\mathbf{d} \leqslant \eta - g_i(\mathbf{x}), \quad i = 1, \ldots, m, \\ -1 \leqslant d_j \leqslant 1, \quad j = 1, \ldots, n. \end{cases} \tag{14.9}$$

Nierówność dotycząca warunku na gradienty ograniczeń obejmuje teraz wszystkie ograniczenia. Dla ograniczeń aktywnych, $i \in I(\mathbf{x})$, mamy $g_i(\mathbf{x}) = 0$, a więc warunki te są identyczne jak w (14.8). W przypadku ograniczeń nieaktywnych $g_i(\mathbf{x}) < 0$, czyli prawa strona jest większa niż η. O ile dla dużych wartości $g_i(\mathbf{x})$ ograniczenie takie jest prawie niezauważalne, to dla ograniczeń, które są "prawie aktywne", odgrywa znaczną rolę. Poza tym, z punktu widzenia implementacji, rozwiązuje to kwestię znajdowania zbioru ograniczeń aktywnych (ze względu na niedokładności zapisu liczb, prawie nigdy nie będzie spełniony warunek $g_i(\mathbf{x}) = 0$).

O skuteczności modyfikacji (14.9) świadczy następujące twierdzenie, które podajemy bez dowodu:

Twierdzenie 14.1. *Załóżmy, że f, g_i, $i = 1, \dots, m$, są klasy C^1. Jeśli ciąg $(\mathbf{x}_k)$ wygenerowany przez algorytm Zoutendijka z modyfikacją Topkisa-Veinotta posiada punkt skupienia, w którym spełniony jest warunek liniowej niezależności, to zachodzi w nim warunek konieczny pierwszego rzędu.*

14.4 Podsumowanie

Metody numeryczne opisane w tym rozdziale pozwalają na znalezienie aproksymacji punktów, w których spełniony jest warunek konieczny pierwszego rzędu. Dopiero twierdzenie 7.6 zagwarantuje optymalność tych punktów. W szczególności, jeśli ograniczenia są liniowe, to wystarczy założyć pseudowypukłość funkcji f. Zwróćmy uwagę, że wymagaliśmy podobnych założeń w poprzednim rozdziale, dla optymalizacji bez ograniczeń. Wymóg wypukłości okazuje się bardzo naturalnym i, co więcej, koniecznym dla sprawnego działania tych metod.

14.5 Zadania

Ćwiczenie 14.1. Znajdź graficznie zbiór dopuszczalnych kierunków spadku w punkcie $\mathbf{x} = (2,3)^T$ dla zagadnienia

$$\begin{cases} (x_1-6)^2 + (x_2-2)^2 \to \min, \\ -x_1 + 2x_2 \leqslant 4, \\ 3x_1 + 2x_2 \leqslant 12, \\ x_1 \geqslant 0, \quad x_2 \geqslant 0. \end{cases}$$

Ćwiczenie 14.2. Udowodnij lemat 14.1.

Ćwiczenie 14.3. Rozwiąż następujące zagadnienie optymalizacyjne z ograniczeniami liniowymi:

$$\begin{cases} 2x_1^2 + 2x_2^2 - 2x_1x_2 - 4x_1 - 6x_2 \to \min, \\ x_1 + x_2 \leqslant 2, \\ x_1 + 5x_2 \leqslant 5, \\ x_1 \geqslant 0, \quad x_2 \geqslant 0. \end{cases}$$

Zastosuj algorytm Zoutendijka i weź jako punkt początkowy $\mathbf{x}_1 = \mathbf{0}$.

Wskazówka. Algorytm kończy się w trzeciej iteracji (optymalna wartość funkcji celu w zagadnieniu poszukiwania kierunku $\mathbf{d}_k$ wynosi wówczas 0).

Ćwiczenie 14.4. Podaj przykład problemu optymalizacyjnego z ograniczeniami nierównościowymi, dla którego istnieje dopuszczalny kierunek spadku, który nie spełnia założeń lematu 14.3.

Bibliografia

[1] K.J. Arrow, A.C. Enthoven. Quasi-concave programming. *Econometrica*, 29:779–800, 1961.

[2] M. Bazaraa, J. Jarvis, H. Sherali. *Linear Programming and Network Flows.* John Wiley and Sons, 1990.

[3] M. Bazaraa, H. Sherali, C. Shetty. *Nonlinear programming.* John Wiley and Sons, wydanie 3, 2006.

[4] D.P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods (Optimization and Neural Computation Series).* Athena Scientific, wydanie 1, 1996.

[5] D.P. Bertsekas. *Nonlinear programming.* Athena Scientific, wydanie 2, 1999.

[6] M.D. Canon, C.D. Cullum, E. Polak. *Sterowanie optymalne i programowanie matematyczne.* WNT, 1975.

[7] W.F. Donoghue. *Distributions and Fourier transforms.* Academic Press, New York, 1969.

[8] S. Gass. *Programowanie liniowe.* PWN, 1980.

[9] D.G. Luenberger. *Linear and Nonlinear Programming.* Addison-Wesley, wydanie 2, 1984.

[10] A.W. Roberts, D. Varberg. *Convex functions.* Academic Press, New York, 1973.

[11] R.T. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

[12] L. Schwartz. *Kurs analizy matematycznej. Tom 1.* PWN, 1979.

[13] W.I. Zangwill. *Programowanie nieliniowe.* WNT, 1974.