ROZDZIAŁ 1

Wiadomości wstępne

Najbardziej praktyczną rzeczą na świecie jest dobra teoria

Hermann von Helmholtz

1.1 Czym jest optymalizacja?

Optymalizacja (ang. *optimization*) jest dziedziną nauki na pograniczu matematyki, fizyki, techniki i ekonomii. Zadaniem optymalizacji jest poszukiwanie najlepszych, pod względem pewnych kryteriów, rozwiązań problemów pochodzących z tych dziedzin, dających się przedstawić w postaci modeli matematycznych. Niektóre z tych modeli da się rozwiązać analitycznie, tj. przedstawić rozwiązanie za pomocą wzoru. Natomiast do rozwiązania większości z nich używa się odpowiednich procedur (metod) iteracyjnych generujących ciągi kolejnych przybliżeń rozwiązań. Budując model matematyczny problemu należy określić przestrzeń $\mathbb{X}$ i jej podzbiór X „możliwych" rozwiązań zwanych *rozwiązaniami dopuszczalnymi* oraz pewien funkcjonał $f : \mathbb{X} \to \mathbb{R}$, określony na tej przestrzeni. Funkcjonał ten stanowi kryterium umożliwiające porównanie między sobą możliwych rozwiązań. Za najlepsze rozwiązanie można wówczas uznać to, dla którego funkcjonał osiąga minimum względnie maksimum na zbiorze X. Ponieważ problem maksymalizacji można sprowadzić do problemu minimalizacji, w dalszych rozważaniach ograniczymy się to tego ostatniego. Jesteśmy oczywiście zainteresowani wyznaczeniem minimum globalnego tego funkcjonału na zbiorze X. Często jednak musimy zadowolić się minimum lokalnym. W tym przypadku przestrzeń $\mathbb{X}$ musi być wyposażona w pewną metrykę d pozwalającą na zdefiniowanie kuli $B(\bar{x}, r) = \{x \in \mathbb{X} : d(x, \bar{x}) \leq r\}$, gdzie $\bar{x} \in \mathbb{X}$ i $r > 0$. Przypomnijmy, że funkcjonał $f : \mathbb{X} \supseteq X \longrightarrow \mathbb{R}$ osiąga w punkcie $\bar{x}$ *minimum lokalne* (ang. *local minimum*), jeśli

$$\exists_{r>0} \forall_{x \in B(\bar{x}, r) \cap X} \quad f(x) \geq f(\bar{x}).$$

Punkt $\bar{x}$ nazywa się wówczas *minimizerem* (ang. *minimizer*) funkcji f. Jeśli nierówność powyższa jest ostra dla $x \neq \bar{x}$, to mówimy, że funkcja f osiąga w punkcie $\bar{x}$ *minimum lokalne izolowane* (ang. *local isolated minimum*).

Można rozważać również problemy optymalizacyjne, w których jest więcej niż jedno kryterium porównujące między sobą możliwe rozwiązania. Problemami tego rodzaju zajmuje się *optymalizacja wielokryterialna*.

Przestrzeń $\mathbb{X}$ na której opisany jest funkcjonał f jest wynikiem przyjętego modelu matematycznego danego zagadnienia optymalizacji. Może to być na przykład pewien zbiór skończony, zbiór przeliczalny, przestrzeń euklidesowa $\mathbb{R}^n$ lub jej podzbiór, albo przestrzenie nieskończenie wymiarowe, takie jak na przykład przestrzeń Hilberta czy też przestrzeń Banacha lub bardziej ogólne przestrzenie metryczne. W trzech ostatnich przypadkach $\mathbb{X}$ jest najczęściej przestrzenią

funkcyjną, czyli jej elementy są funkcjami. Przestrzeń ta jest dobrana specjalnie do rozpatrywanego zagadnienia. Należy tu jednak podkreślić, że przybliżone rozwiązanie zagadnień w przestrzeniach nieskończenie wymiarowych uzyskuje się często porzez zrzutowanie na podporzestrzeń skończenie wymiarową. Tak jest na przykład w tzw. metodzie elementów skończonych.

W dalszych rozważaniach będziemy ograniczać się w zasadzie do zagadnień optymalizacji, których modele matematyczne określone są na przestrzeni euklidesowej $\mathbb{R}^n$ (lub czasem na jej podzbiorze $\mathbb{X}$). Niech więc dane będą: funkcja $f : \mathbb{R}^n \to \mathbb{R}$ oraz pewien podzbiór $X \subseteq \mathbb{R}^n$. Podzbiór ten może być podany w postaci abstrakcyjnej, ale najczęściej podawany jest w postaci

$$X = \{x \in \mathbb{X} : c_i(x) = 0 \text{ dla } i \in E \text{ oraz } c_i(x) \leq 0 \text{ dla } i \in I\},$$

gdzie $E = \{1, ..., p\}$, $I = \{p+1, ..., m\}$, $c_i : \mathbb{R}^n \to \mathbb{R}$, $i \in E \cup I$. Funkcja f nazywa się *funkcją celu* (ang. *objective*), zaś funkcje c_i, $i \in E \cup I$, nazywają się *funkcjami ograniczeń* lub *ograniczeniami* (ang. *constraints*). Ograniczenia ponumerowane wskaźnikami $i \in E$ nazywamy *ograniczeniami równościowymi* (ang. *equality constraints*), zaś ponumerowane wskaźnikami $i \in I$ – *ograniczeniami nierównościowymi* (ang. *inequality constraints*). Podzbiór $\mathbb{X}$ najczęściej podawany jest w postaci: $\mathbb{X} = \mathbb{R}^n$, $\mathbb{X} = \mathbb{R}^n_+$, $\mathbb{X} = \mathbb{Z}^n$, bądź $\mathbb{X} = \mathbb{Z}^n_+$. W ostatnich dwóch przypadkach mówimy o tzw. *programowaniu całkowitoliczbowym* (ang. *integer programming*) lub inaczej o tzw. *programowaniu dyskretnym* (ang. *discrete programming*). Dla zadań ostatniego typu stosuje się specjalne metody, które są przedmiotem odrębnego wykładu.

Przedmiotem dalszych rozważań będzie w szczególności zadanie *minimalizacji bez ograniczeń* (ang. *unconstrained minimization*)

$$\begin{array}{ll} \text{minimalizować} & f(x) \\ \text{względem} & x \in \mathbb{R}^n \end{array}$$

i zadanie *minimalizacji z ograniczeniami* (ang. *constrained minimization)* przedstawione w postaci abstrakcyjnej

$$\begin{array}{ll} \text{minimalizować} & f(x) \\ \text{względem} & x \in X \subseteq \mathbb{R}^n \end{array}$$

bądź w postaci

$$\begin{array}{lll} \text{minimalizować} & f(x) & \\ \text{względem} & x \in \mathbb{R}^n & \\ \text{przy ograniczeniach} & c_i(x) = 0, & i \in E = \{1, ..., m\}, \\ & c_i(x) \leq 0, & i \in I = \{m+1, ..., p\}. \end{array}$$

Elementy $x = (x_1, x_2, ..., x_n)$ przestrzeni $\mathbb{R}^n$ nazywają się *zmiennymi zadania*, a liczba n nazywa się *wymiarem zadania*.

Dokładniej rzecz ujmując, optymalizacja zajmuje się:

(a) warunkami istnienia rozwiązań dopuszczalnych (warunkami niesprzeczności problemu),

(b) warunkami koniecznymi i warunkami wystarczającymi istnienia minimum (lub przynajmniej skończonego kresu dolnego)

(c) metodami wyznaczenia tego minimum i punktu x^* realizującego to minimum (w sposób dokładny lub przybliżony),

(d) badaniem szybkości zbieżności ciągu kolejnych przybliżeń do rozwiązania zadania.

Czasem do zadań optymalizacji zalicza się również:

(e) zbudowanie – dla konkretnego zagadnienia praktycznego – odpowiedniego modelu matematycznego w postaci zadania minimalizacji funkcji wielu zmiennych,

(f) interpretację rozwiązania takiego zadania.

Dział optymalizacji, którego głównym celem jest konstrukcja metod iteracyjnych służących rozwiązaniu zadania optymalizacji i badaniem zbieżności tych metod nazywa się *programowaniem matematycznym* (ang. *mathematical programming*). Czasem będziemy zamiennie używać nazwy optymalizacja i programowanie matematyczne.

Na ogół zakłada się, że funkcje ograniczeń c_i, $i \in E \cup I$ są ciągłe. W konsekwencji zbiór ograniczeń X jest domknięty. Jeśli ponadto jest on ograniczony i funkcja f jest również ciągła, to na mocy twierdzenia Weierstrassa osiąga ona minimum na X.

W zadaniu minimalizacji z ograniczeniami:

(a) ograniczenie równościowe $c_i(x) = 0$ można zastąpić dwoma ograniczeniami nierównościowymi $c_i(x) \leq 0$ i $-c_i(x) \leq 0$,

(b) ograniczenie nierównościowe $c_i(x) \leq 0$ można zastąpić ograniczeniem równościowym $c_i(x) + u_i = 0$, wprowadzając tak zwaną *zmienną uzupełniającą* (ang. *slack variable*) $u_i \geq 0$,

Można mówić również o zadaniu maksymalizacji. Ponieważ

$$\max_{x \in X} f(x) = -\min_{x \in X} -f(x),$$

więc zadanie maksymalizacji można sprowadzić do zadania minimalizacji i odwrotnie.

Niech $f : X \to \mathbb{R}$ przyjmuje wartości w zbiorze $Y \subseteq \mathbb{R}$. Wówczas dla dowolnej funkcji rosnącej $g : Y \to \mathbb{R}$, funkcja f osiąga minimum (maksimum) w punkcie $x^* \in X$ wtedy i tylko wtedy, gdy funkcja $g \circ f$ osiąga minimum (maksimum) w punkcie x^* równe $g(f(x^*))$.

Przykład 1.1.1 Dla funkcji $F : \mathbb{R}^n \to \mathbb{R}^m$ funkcja f określona wzorem $f(x) = \|F(x)\|$ osiąga minimum (maksimum) w punkcie $x^* \in \mathbb{R}^n$ wtedy i tylko wtedy, gdy funkcja h określona wzorem $h(x) = \|F(x)\|^2$ osiąga minimum (maksimum) w punkcie x^*. Zaletą minimalizacji funkcji h w stosunku do minimalizacji funkcji f jest to, że funkcja h jest różniczkowalna, jeśli tylko funkcja F jest różniczkowalna, natomiast nie musi tak być dla funkcji f.

Przykład 1.1.2 Problem maksymalizacji funkcji $f : \mathbb{R}^n \to (0, +\infty)$ jest równoważny maksymalizacji funkcji $h : \mathbb{R}^n \to \mathbb{R}$ określonej wzorem $h(x) = \ln(f(x))$. Jeśli funkcja f jest iloczynem innych funkcji, $f(x) = f_1(x)f_2(x)...f_m(x)$, to logarytm zamienia ją na sumę logarytmów tych funkcji, czyli $h(x) = \ln f_1(x) + \ln f_2(x) + ... + \ln f_m(x)$, co jest wygodne przy różniczkowaniu. Fakt ten jest wykorzystywany w statystyce matematycznej, sieciach neuronowych, czy uczeniu maszynowym, gdzie maksymalizuje się tzw. funkcję wiarygodności.

Zadania minimalizacji dzielimy na:

(a) *minimalizację różniczkowalną* (ang. *differentialble minimization*), gdy wszystkie funkcje f, c_i, $i \in E \cup I$, są różniczkowalne,

(b) *minimalizację nieróżniczkowalną* (ang. *nondifferentialble minimization*), gdy przynajmniej jedna z funkcji f, c_i, $i \in E \cup I$, nie jest różniczkowalna.

Ponadto, jeśli:

- wszystkie funkcje f, c_i, $i \in E \cup I$, są różniczkowalne w sposób ciągły, to mówimy wówczas o zadaniu *minimalizacji gładkiej* (ang. *smooth minimization*),
- przynajmniej jedna z funkcji f, c_i, $i \in E \cup I$, nie jest różniczkowalna w sposób ciągły, to mówimy wówczas o zadaniu *minimalizacji niegładkiej* (ang. *nonsmooth minimization*).

Wśród zadań minimalizacji z ograniczeniami wyróżniamy:

(a) *zadanie programowania liniowego* (ZPL) (ang. *linear programming problem*), gdy wszystkie funkcje f, c_i, $i \in E \cup I$, są afiniczne,

(b) zadanie programowania nieliniowego (ang. *nonlinear programming problem*), gdy przynajmniej jedna z funkcji f, c_i, $i \in E \cup I$, nie jest afiniczna

(c) *zadanie programowania kwadratowego* (ang. *quadratic programming problem*), gdy funkcja f jest kwadratowa, zaś wszystkie funkcje c_i, $i \in E \cup I$, są liniowe,

(d) *zadanie minimalizacji wypukłej* (ang. *convex minimization problem*), gdy $E = \emptyset$ i wszystkie funkcje f, c_i, $i \in I$, są wypukłe. Z zadaniem minimalizacji wypukłej mamy do czynienia również wówczas, gdy założenie o braku ograniczeń równościowych zastąpimy założeniem, że wszystkie ograniczenia równościowe są liniowe. Każde ograniczenie równościowe liniowe możemy bowiem zastąpić dwoma ograniczeniami nierównościowymi liniowymi.

1.2 Przykłady

Programowanie matematyczne znajduje liczne zastosowania w wielu dziedzinach między innymi w ekonomii, w finansach, czy też w technice. Podamy teraz kilka najprostszych przykładów zagadnień praktycznych prowadzących do zadań programowania matematycznego.

Przykład 1.2.1 (*zagadnienie analizy działalności gospodarczej*) Producent wytwarza n towarów P_1,...,P_n wykorzystując m surowców lub ogólniej, czynników produkcyjnych R_1,...,R_m. Producent dysponuje b_i jednostkami surowca R_i, $i = 1, ..., m$. Do wyprodukowania jednostki towaru P_j potrzeba a_{ij} jednostek surowca R_i, $i = 1, ..., m$, $j = 1, ..., n$. Jednostka wyprodukowanego towaru P_j przynosi zysk c_j jednostek pieniężnych, $j = 1, ..., n$. Ile jednostek każdego z towarów powinien wytwarzać producent, aby zapewnić sobie maksymalny zysk?

Oznaczając przez x_j ilość jednostek wyprodukowanego towaru P_j, $j = 1, ..., n$, możemy powyższe zagadnienie sformułować następująco:

$$\begin{array}{lll} \text{maksymalizować} & c_1x_1 + c_2x_2 + ... + c_nx_n & \\ \text{przy ograniczeniach} & a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n \leq b_1, & \\ & a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n \leq b_2, & \\ & ... & \\ & a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n \leq b_m, & \\ & x_1, ..., x_n \geq 0. & \end{array}$$

Przykład 1.2.2 (*zadanie transportowe*) W sieci m magazynów A_1,..., A_m składuje się pewien towar. Należy go dostarczyć do sieci n sklepów B_1,..., B_n. Zapas magazynu A_i wynosi a_i jednostek towaru, $i = 1, ..., m$. Sklep B_j potrzebuje b_j jednostek towaru, $j = 1, ..., n$. Zakładamy, że

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j \tag{1.1}$$

(łączna podaż jest równa łącznemu popytowi). Koszty transportu jednostki towaru z magazynu A_i do sklepu B_j wynoszą c_{ij} jednostek pieniężnych, $i = 1, ..., m$, $j = 1, ..., n$. Należy określić plan transportowy o minimalnych kosztach zaspokajający zapotrzebowanie wszystkich sklepów (czyli, przy podanym założeniu, wyczerpujący łączne zapasy magazynów).

Jeśli przez x_{ij} oznaczymy ilość towaru transportowanego z magazynu A_i do sklepu B_j, $i = 1, ..., m$, $j = 1, ..., n$, to powyższe zagadnienie możemy sformułować następująco:

$$\begin{array}{rrcll} \text{minimalizować} & \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} & & & \\ \text{przy ograniczeniach} & \sum_{j=1}^{n} x_{ij} & = & a_i, & i = 1, ..., m \\ & \sum_{i=1}^{m} x_{ij} & = & b_j, & j = 1, ..., n \\ & x_{ij} & \geq & 0, & i = 1, ..., m, \\ & & & & j = 1, ..., n. \end{array} \tag{1.2}$$

Przykład 1.2.3 (*optymalny koszyk towarów*) Konsument dysponujący dochodem I może zakupić dowolne towary $P_1, P_2, ..., P_n$ znajdujące się na rynku. Niech p_j oznacza cenę jednostki j-tego towaru, zaś x_j oznacza ilość jednostek j-tego towaru zakupionego przez konsumenta, $j = 1, ..., n$. Określona jest tzw. *funkcja użyteczności* $u : \mathbb{R}^n_+ \to \mathbb{R}$, gdzie $u(x)$ wyraża stopień pożądania przez konsumenta koszyka towarów $x = (x_1, ..., x_n)$. Funkcja ta posiada pewne naturalne własności (np. monotoniczność, wklęsłość). Konsument ma za zadanie wybrać koszyk towarów, dla którego użyteczność jest największa. Zadanie to można sformułować następująco

$$\begin{array}{ll} \text{maksymalizować} & u(x) \\ \text{przy ograniczeniach} & \sum_{j=1}^{n} p_j x_j \leq I \\ & x_j \geq 0,\ j = 1, ..., n. \end{array}$$

Przykład 1.2.4 (*optymalny portfel akcji, ang. optimal portfolio*) Na giełdzie notowanych jest wiele akcji. Oznaczmy je kolejnymi liczbami naturalnymi $i = 1, 2, ..., n$. Każda akcja na giełdzie wiąże się z dochodem i z ryzykiem. Stopa zwrotu każdej akcji i, $i = 1, ..., n$, jest zmienną losową. Oznaczmy ją symbolem X_i. Zwykle analizie podlega jej wartość oczekiwana $R_i = EX_i$ (tzw. średnia stopa zwrotu) i jej odchylenie standardowe $s_i = \sqrt{E(X_i - EX_i)^2}$ (tzw. ryzyko). Przypuśćmy, że inwestor tworzy portfel składający się z m akcji $X_1, X_2, ..., X_m$ przez wybór wektora wag $w = (w_1, w_2, ..., w_m)$ takiego, że $w \geq 0$ oraz $\sum w_i = 1$. Wagi te określają procentowy udział pieniężny poszczególnych akcji w portfelu. Dla portfela akcji można określić zmienną losową postaci $X = \sum_{i=1}^{m} w_i X_i$ będącą stopą zwrotu tego portfela. Zgodnie z własnościami wartości oczekiwanej średnia stopa zwrotu portfela ma postać $R = EX = \sum w_i R_i$ Znając macierz kowariancji $S = [s_{ij}]$ zmiennych losowych losowych $X_1, ..., X_m$ można wyznaczyć ryzyko s tego portfela, które zgodnie z własnościami wariancji ma postać $s = (\sum_{i=1}^{m} \sum_{j=1}^{m} w_i w_j s_{ij})^{\frac{1}{2}}$. Należy wyznaczyć portfel maksymalizujący pewną funkcję użyteczności inwestora opisującą skłonność inwestora do ryzyka.

Przykład 1.2.5 (*zadanie najkrótszych dróg*) Należy znaleźć najkrótszą (ewentualnie najtańszą, najszybszą) trasę z miasta A do miasta B, przy czym możemy poruszać się wyłącznie drogami. Dla każdego odcinka drogi (łączącego dwa skrzyżowania) znana jest jego długość (ewentualnie koszt lub czas przejazdu). Sieć dróg można przedstawić w postaci pewnego grafu $G = (V, E)$, którego wierzchołki $v \in V$ są skrzyżowaniami, zaś krawędzie (łuki) $e \in E$ odcinkami dróg łączącymi te skrzyżowania. Graf ten jest zazwyczaj skierowany, ponieważ niektóre odcinki mogą być jednokierunkowe lub koszty (czas) mogą zależeć od kierunku przejazdu. Ponadto dana jest funkcja wag $c : E \to \mathbb{R}$ wyrażająca długości (koszty lub czas przejazdu) dla poszczególnych odcinków dróg. Zadanie sprowadza się więc do wyznaczenia drogi w grafie G minimalizującej długość (koszt lub czas przejazdu) spośród wszystkich możliwych dróg łączących miasta A i B.

Do zadania najkrótszych dróg można sprowadzić następujące zadanie:

Dany jest koszyk V składający się z n walut, zbiór par uporządkowanych $E \subseteq V \times V$. Ponadto dla dowolnej pary walut $(u, w) \in E$ dany jest kurs wymiany $r_{uw} > 0$. Wielkość r_{uw} wyraża ilość jednostek waluty w, które otrzymamy ze sprzedaży jednostki waluty u. Przypuśćmy, że dokonujemy ciągu wymian $v_1 \to v_2 \to ... \to v_k$. Wówczas za jednostkę waluty v_1 otrzymamy $r_{v_1 v_2} \cdot ... \cdot r_{v_{k-1} v_k}$ jednostek waluty v_k. Dla waluty v_1, poszukujemy najkorzystniejszego sposobu jej wymiany na walutę v_k.

Przykład 1.2.6 (*zagadnienie komiwojażera, ang. traveling salesman problem*) Handlowiec chce odwiedzić n miejscowości (klientów) startując ze swojej miejscowości i wracając tam po skończonej podróży. W jakiej kolejności powinien odwiedzać te miejscowości, aby łączna przebyta trasa (względnie łączny koszt podróży albo łączny czas podróży) była minimalna. Niech

$$x_{ij} = \begin{cases} 1 & \text{jeśli bezpośrednio po } i\text{-tej odwiedzana będzie } j\text{-ta miejscowość,} \\ 0 & \text{poza tym.} \end{cases}$$

Ponadto niech c_{ij} oznacza odległość (względnie koszt podróży albo czas podróży) między i-tą a j-tą miejscowością, $i, j = 1, ...n$. Zadanie można wówczas sformułować jako zadanie dyskretnego programowania liniowego w następujący sposób:

$$\begin{array}{rr} \text{minimalizować} & \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij} \\ \text{przy ograniczeniach} & \sum_{j=1}^{n} x_{ij} = 1, \quad i = 1, ..., n, \\ & \sum_{i=1}^{n} x_{ij} = 1, \quad j = 1, ..., n, \\ & x_{ij} \in \{0, 1\}, \quad i, j = 1, .., n \\ & \text{i „krótkie cykle” są zabronione.} \end{array}$$

Ostatnie ograniczenie gwarantuje, że trasa nie rozłoży się na kilka rozłącznych cykli.

Przykład 1.2.7 (*dyskretna aproksymacja Czebyszewa, ang. discrete Chebychev approximation*) Dane są wartości y_i pewnej funkcji $y : \mathbb{R} \to \mathbb{R}$ w punktach t_i, $i = 1, ..., m$. Należy znaleźć wielomian p stopnia co najwyżej n postaci

$$p(x,t) = \sum_{k=0}^{n} x_k t^k,$$

gdzie $x = (x_0, ..., x_n) \in \mathbb{R}^{n+1}$ jest wektorem współczynników tego wielomianu, minimalizujący funkcję $f : \mathbb{R}^n \to \mathbb{R}$ zadaną wzorem

$$f(x) = \max_{i=1,2,...,m} |y_i - p(x,t_i)|.$$

Ponieważ zadanie to nie jest różniczkowalne, zastępuje się je równoważnym zadaniem minimalizacji różniczkowalnej z ograniczeniami

$$\begin{array}{ll} \text{minimalizować} & \varepsilon \\ \text{względem} & (x,\varepsilon) \in \mathbb{R}^{n+1} \times \mathbb{R} \\ \text{przy ograniczeniach} & -\varepsilon \le y_i - p(x,t_i) \le \varepsilon,\ i = 1, ..., m. \end{array}$$

Omawiane zagadnienie nosi nazwę *dyskretnej aproksymacji Czebyszewa* (ang. *discrete Chebychev approximation*) i ma duże zastosowanie praktyczne, na przykład w sytuacji, gdy dla danych wyników pomiarów chcemy znaleźć wielomian określonego stopnia najbardziej "pasujący" do tych wyników.

Przykład 1.2.8 (*zagadnienie membrany z przeszkodą*) W obszarze $\Omega \subseteq \mathbb{R}^2$ rozciągnięta jest elastyczna membrana, która ugina się pod działaniem siły f. Ponadto ugięcie u jest ograniczone przeszkodą opisaną funkcją $h : \Omega \to \mathbb{R}$, czyli $u \ge h$. Należy wyznaczyć ugięcie u tej membrany. Ugięcie membrany jest rozwiązaniem pewnego równania różniczkowego cząstkowego lub rozwiązaniem związanego z nim zadania minimalizacji funkcjonału określonego na odpowiednio skonstruowanej przestrzeni Hilberta. W praktyce dokonuje się dyskretyzacji, co prowadzi do układu równań liniowych lub zadania minimalizacji kwadratowej dużej skali.

Przykład 1.2.9 (*rzut metryczny ang. metric projection*) Dany jest podzbiór wypukły domknięty $D \subseteq \mathbb{R}^n$ i pewien punkt $\bar{x} \notin D$. Należy w zbiorze D znaleźć punkt leżący najbliżej punktu $\bar{x}$ w sensie odległości euklidesowej. Zbiór D opisany jest często w postaci układu nierówności $c_i(x) \le 0$, $i = 1, ..., m$, gdzie $c_i : \mathbb{R}^n \to \mathbb{R}$ są funkcjami wypukłymi. Zadanie to można sformułować następująco

$$\begin{array}{ll} \text{minimalizować} & \|x - \bar{x}\| \\ \text{względem} & x \in \mathbb{R}^n \\ \text{przy ograniczeniu} & x \in D \end{array}$$

względnie w postaci

$$\begin{array}{ll} \text{minimalizować} & \frac{1}{2}\|x - \bar{x}\|^2 \\ \text{względem} & x \in \mathbb{R}^n \\ \text{przy ograniczeniach} & c_i(x) \le 0,\ i = 1, ..., m. \end{array}$$

Ostatnie zadanie jest zadaniem minimalizacji wypukłej z ograniczeniami. Jeśli dodatkowo założymy, że funkcje c_i, $i = 1, ..., m$, są różniczkowlane, to zadanie to jest zadaniem minimalizacji różniczkowalnej wypukłej z ograniczeniami.

Przykład 1.2.10 (*liniowe zadanie najmniejszych kwadratów, ang. linear least squares problem*) Dany jest układ równań $Ax = b$, gdzie A jest macierzą typu $m \times n$, $x \in \mathbb{R}^n$, zaś $b \in \mathbb{R}^m$. Należy wyznaczyć punkt $x^* \in \mathbb{R}^n$ minimalizujący funkcję $f(x) = \frac{1}{2}\|Ax - b\|^2$. Jeśli $Ax = b$ posiada rozwiązanie, to jest ono również minimizerem funkcji f. Jeśli natomiast układ nie posiada rozwiązania, to poszukujemy punktu x^*, dla którego funkcja f osiąga minimum, czyli lewe i prawe strony układu różnią się najmniej w sensie średnio-kwadratowym.

Przykład 1.2.11 (*dopasowanie krzywej do danych pomiarowych*) Pewne zjawisko fizyczne (np. długosć pręta w zależności od temperatury) opisane jest pewną nieznaną funkcją $f : \mathbb{R} \supseteq [a, b] \to \mathbb{R}$. Zadanie polega na znalezieniu przybliżonego kształtu krzywej opisanej tą funkcją. W tym celu wybieramy N punktów pomiarowych $x_i \in \mathbb{R}$, $i = 1, 2, ..., M$, i w punktach tych obserwujemy wartości t_i, $j = 1, 2, ..., M$, tej funkcji. Wartości te mogą być obarczone błędami wynikającymi z charakteru pomiarów. Dana jest pewna rodzina funkcji $Y := \{y(\cdot, w) : w \in \mathbb{R}^n\}$, gdzie w jest wektorem parametrów. Przykładowo, może to być rodzina wielomianów stopnia co najwyżej N, czyli $y(x, w) = w_N x^N + w_{N-1} x^{N-1} + ... + w_1 x + w_0$, $x \in \mathbb{R}$, wówczas $n = M + 1$. Należy dopasować wektor parametrów $w \in \mathbb{R}^n$ tak, aby zminimalizować tzw. błąd średnio-kwadratowy

$$E(w) := \frac{1}{2} \sum_{i=1}^{M} (y(x_i, w) - t_i)^2.$$

Przykład 1.2.12 (*zadanie dopuszczalności wypukłej, ang. convex feasibility problem*) Danych jest skończenie wiele zbiorów wypukłych, domkniętych $C_i \subseteq \mathbb{R}^n$, $i = 1, ..., m$. Należy wyznaczyć punkt x należący do przekroju tych zbiorów, o ile taki punkt istnieje. Zadanie to można zapisać w postaci

$$\begin{array}{ll} \text{minimalizować} & f(x) = \frac{1}{2} \sum_{i=1}^{m} d^2(x, C_i) \\ \text{względem} & x \in \mathbb{R}^n, \end{array}$$

gdzie $d(x, C) = \inf_{z \in C} \|x - z\|$. Jeśli $\bigcap_{i=1}^{m} C_i \neq \varnothing$, to minimalna wartość funkcji celu wynosi 0. Wówczas dowolne rozwiązanie tego zadania jest jednocześnie rozwiązaniem zadania dopuszczalności wypukłej, i odwrotnie.

Przykład 1.2.13 (*rozwiązanie układu równań lub nierówności*) Dany jest podzbiór $C \subseteq \mathbb{R}^n$. Należy wyznaczyć punkt $x \in C$ będący rozwiązaniem układu równań

$$f_i(x) = 0,\ i = 1, ..., m,$$

gdzie $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, 2, ..., m$. Zadanie to jest równoważne zadaniu minimalizacji różniczkowalnej z ograniczeniami

$$\begin{array}{ll} \text{minimalizować} & f(x) = \sum_{i=1}^{m} [f_i(x)]^2 \\ \text{względem} & x \in \mathbb{R}^n \\ \text{przy ograniczeniu} & x \in C. \end{array}$$

Jeśli wiadomo, że powyższy układ równań ma rozwiązanie, to minimalna wartość funkcji celu wynosi 0. Do podobnego zadania minimalizacji można również sprowadzić rozwiązanie układu nierówności

$$f_i(x) \leq 0,\ i = 1, ..., m.$$

Zadanie to jest bowiem równoważne zadaniu minimalizacji różniczkowalnej

$$\begin{array}{ll} \text{minimalizować} & f(x) = \sum_{i=1}^{m} [f_{i+}(x)]^2 \\ \text{względem} & x \in \mathbb{R}^n, \end{array}$$

gdzie $\alpha_+ := \max(\alpha, 0)$, lub, po wprowadzeniu zmiennych uzupełniających u_i, $i = 1, ..., m$, zadaniu

$$\begin{array}{ll} \text{minimalizować} & f(x, u) = \sum_{i=1}^{m} [f_i(x) + u_i]^2 \\ \text{względem} & (x, u) \in \mathbb{R}^n \times \mathbb{R}^m \\ \text{przy ograniczeniach} & u \geq 0. \end{array}$$

Przykład 1.2.14 (*zagadnienie tomografii komputerowej*) *Tomografia komputerowa* (skrótowo CT od ang. *computed tomography*) ma za zadanie wyznaczenie rozkładu gęstości (lub stopnia nieprzejrzystości) materii (tkanki) badanego obiektu (ciała ludzkiego) na podstawie obrazów rentgenowskich tego obiektu wykonanych w wielu różnych kierunkach. W tym znaczeniu CT stosowana jest jako jedno z podstawowych narzędzi w diagnostyce medycznej.

Rozważmy obiekt płaski (na przykład przekrój ciała ludzkiego w zadanej płaszczyźnie) i umieśćmy ten obiekt w układzie współrzędnych. Po odpowiedniej dyskretyzacji rozpatrujemy skończenie wiele punktów nazywanych *pikselami* (od ang. *picture element*). Każdemu pikselowi p_j, $j = 1, ..., n$, można przypisać gęstość materii (tkanki) lub stopień "nieprzejrzystości" materii (tkanki) w tym pikselu. Wielkości te są znane dla każdego rodzaju materii (tkanki). W celu wyznaczenia gęstości tkanki w każdym pikselu, obiekt zostaje prześwietlony promieniami rentgenowskimi l_i, $i = 1, ..., m$, wysyłanymi w różnych kierunkach z różnych pozycji źródła. Intensywność pojedynczego promienia rentgenowskiego przechodzącego przez obiekt zostaje osłabiona w każdym punkcie (pikselu) leżącym na drodze promienia proporcjonalnie do gęstości materii w tym punkcie (pikselu). Po przejściu przez obiekt intensywność promieniowania jest mierzona przez układ czujników. Zadanie wyznaczenia funkcji gęstości materii prowadzi do układu równań

$$\sum_{j=1}^{n} a_{ij} x_j = \beta_i$$

$i = 1, ..., m$, w którym j odpowiada numerowi piksela p_j, zaś i – numerowi promienia rentgenowskiego l_i. W układzie tym a_{ij} oznacza długość odcinka będącego częścią wspólną promienia l_i i piksela p_j, β_j odpowiada intensywności i-tego promienia zmierzonej przez detektor, zaś x_j jest poszukiwaną gęstością tkanki dla piksela p_j, $j = 1, ..., n$. W praktyce $m, n \simeq 10^5 \div 10^6$. Macierz tego układu jest tzw. *macierzą rzadką*, tzn. co najwyżej 1% jej elementów jest różnych od zera. Z uwagi na dyskretyzację oraz błędy pomiarów intensywności dokonanych przy pomocy detektorów, równości w tym układzie są przybliżone. Nie możemy się więc spodziewać, że posiada on rozwiązanie, ale możemy poszukiwać przybliżonego rozwiązania tego układu. Należy jednocześnie pamiętać o dodatkowym założeniu $x \geq 0$, gdyż gęstość materii (tkanki) jest nieujemna. CT jest więc problem dopuszczalności wypukłej.

Przykład 1.2.15 (*zagadnienie radioterapii*) Radioterapia z użyciem wiązek o modulowanej intensywności (IMRT od ang. *intensity modulated radiation therapy*) polega na ułożeniu planu naświetleń promieniami rentgenowskimi pewnego obszaru ciała ludzkiego zawierającego guz, mającego na celu zniszczenie tego guza zachowując jednocześnie zdrowe organy znajdujące się w jego pobliżu. Naświetlenia następują w kilku kierunkach i w każdym z nich wysyłana jest wiązka promieni o zmiennej intensywności zadanej odpowiednim rozkładem. Dla guza podana jest minimalna dawka promieniowania l_1, zaś dla każdego zdrowego organu – maksymalna dawka promieniowania u_k, $k = 2, ..., K$. Po odpowiedniej dyskretyzacji, "podziale" ciała na *voksele* (od ang. *volume element*) i rozpatrywaniu skończonej liczby promieni rentgenowskich, zadanie powyższe sprowadza się do rozwiązania pewnego układu nierówności liniowych określającego idealne ograniczenia dla dawek

$$\begin{array}{ll} d_i(x) \geq l_1 & \text{dla } i \in I_1 \\ d_i(x) \leq u_k & \text{dla } i \in I_k,\ k = 2, ..., K \\ x \geq 0 & \end{array}$$

gdzie $x = (x_1, ..., x_n)$ jest wektorem intensywności radiacji, $d_i(x) = \sum_{j=1}^{n} a_{ij} x_j$ jest całkowitą dawką zaabsorbowaną przez i-ty voksel wskutek radiacji pochodzącej ze wszystkich promieni. Ponieważ najczęściej ograniczenia są zbyt restrykcyjne (chcemy, aby maksymalne dawki dla

zdrowych organów były możliwie najmniejsze, zaś minimalne dawki dla guza – możliwie największe), powyższy układ najczęściej nie ma rozwiązań. Wprowadza się więc tzw. funkcję kary

$$f(x) = \underbrace{w_1}_{\substack{\text{waga}\\ \text{dla guza}}} \sum_{i\in I_1}[(\underbrace{l_1}_{\substack{\text{minimalna}\\ \text{dawka}\\ \text{dla guza}}} - \underbrace{d_i(x)}_{\substack{\text{dawka}\\ \text{dla } i\text{-tego}\\ \text{voksela guza}}})_+]^2$$

$$+\sum_{k=1}^{K} \underbrace{w_k}_{\substack{\text{waga dla}\\ \text{organu } k}} \sum_{i\in I_k}[(\underbrace{d_i(x)}_{\substack{\text{dawka}\\ \text{dla } i\text{-tego}\\ \text{voksela}\\ \text{organu } k}} - \underbrace{u_k}_{\substack{\text{maksymalna}\\ \text{dawka}\\ \text{dla organu } k}})_+]^2$$

gdzie $w = (w_1, ..., w_K) \geq 0$ jest tzw. wektorem wag dla poszczególnych organów. Funkcja kary mierzy sumaryczne ważone odstępstwa (średniokwadratowe) dawek od idealnych ograniczeń. Wyznaczenie optymalnego wektora intensywności radiacji polega na minimalizacji funkcji kary na zbiorze wszystkich nieujemnych wektorów radiacji. IMRT jest zadaniem dużej skali, ponieważ zarówno liczba vokseli, jak i liczba promieni jest rzędu $10^4 \div 10^6$.

1.3 Podstawowe oznaczenia

W dalszej części używać będziemy następujących oznaczeń i konwencji:

- $x = (x_1, ..., x_n) \in \mathbb{R}^n$, gdzie x_j, $j = 1, 2, ..., n$ są współrzędnymi punktu (wektora) x. Wektor x będziemy identyfikować z macierzą typu $n \times 1$ (wektor kolumnowy), tzn będziemy pisać $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ lub $x = [x_1, x_2, ..., x_n]^T$. Jeśli nie będzie prowadzić to do nieporozumień, dla punktów (wektorów) w $\mathbb{R}^2$ lub $\mathbb{R}^3$ będziemy zamiennie używać oznaczeń odpowiednio (x, y) bądź (x, y, z);
- x^k oznacza k-ty wyraz ciągu $\{x^k\}_{k=0}^{\infty}$ elementów przestrzeni $\mathbb{R}^n$;
- $x \geq 0$ oznacza, że wszystkie współrzędne wektora x są nieujemne;
- $\alpha_+ = \max\{0, \alpha\}$, $\alpha_- = \max\{0, -\alpha\}$ oznaczają odpowiednio cześć dodatnią i część ujemną liczby $\alpha \in \mathbb{R}$;
- $x_+ = ((x_1)_+, ..., (x_n)_+)$, czyli część dodatnią wektora $x \in \mathbb{R}^n$ otrzymujemy wyznaczając części dodatnie jego współrzędnych x_j, $j = 1, 2, ..., n$;
- $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x \geq 0\}$ oznacza *ortant nieujemny*;
- $x > 0$ oznacza, że wszystkie współrzędne wektora x są dodatnie;
- $\mathbb{R}^n_{++} = \{x \in \mathbb{R}^n : x > 0\}$ oznacza *ortant dodatni*;
- $B(x, r) = \{y \in \mathbb{R}^n : \|y - x\| \leq r\}$ oznacza *kulę domkniętą* o środku w punkcie x i promieniu r;
- $\operatorname{cl} D$ oznacza *domknięcie* niepustego zbioru $D \subseteq \mathbb{R}^n$ (ang. *closure*);
- $\operatorname{bd} D := \operatorname{cl} D \cap \operatorname{cl}(\mathbb{R}^n \setminus D)$ oznacza *brzeg* niepustego zbioru $D \subseteq \mathbb{R}^n$ (ang. *boundary*);
- $\#J$ oznacza *moc* zbioru J; w przypadku zbioru skończoego $\#J$ jest liczbą elementów tego zbioru.

Niech $f : X \to \mathbb{R}$, gdzie X jest pewnym podzbiorem $\mathbb{R}^n$. Wówczas:

- $\operatorname{Argmin}_{x \in X} f(x) = \{y \in X : \forall_{x \in X} \quad f(y) \leq f(x)\}$ oznacza zbiór, na którym funkcja f osiąga swoje minimum na zbiorze X;
- $\operatorname{argmin}_{x \in X} f(x)$ oznacza element zbioru $\operatorname{Argmin}_{x \in X} f(x)$, czyli minimizer funkcji f na zbiorze X;
- f_+ oznacza nieujemną część funkcji f, czyli $f_+(x) = \max\{0, f(x)\}$.

Definicja 1.3.1 *Poziomicą* (ang. *level set*) funkcji $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ dla poziomu $\alpha \in \mathbb{R}$ nazywamy zbiór

$$\{x \in \mathbb{R}^n : f(x) = \alpha\},$$

natomiast zbiór

$$S(f, \alpha) = \{x \in \mathbb{R}^n : f(x) \leq \alpha\})$$

nazywamy *podpoziomicą* (ang. *sublevel set*) tej funkcji dla poziomu $\alpha \in \mathbb{R}$.

Definicja 1.3.2 *Epigrafem* (ang. *epigraph*) funkcji $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ nazywamy zbiór

$$\operatorname{epi} f = \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq \alpha\}.$$

1.4 Elementy algebry liniowej

Pojęcia i fakty podane w tym ustępie są przypomnieniem lub uzupełnieniem odpowiednich definicji i twierdzeń z algebry liniowej.

- $\langle x, y\rangle$ oznacza *iloczyn skalarny* (ang. *inner product*) wektorów $x, y \in \mathbb{R}^n$,
- Iloczyn skalarny $\langle \cdot, \cdot \rangle : \mathbb{R}^n \to \mathbb{R}$ określony wzorem

$$\langle x, y\rangle := \sum_{j=1}^{n} x_j y_j,$$

gdzie $x = (x_1, ..., x_n)$, $y = (y_1, ..., y_n)$, lub zapisany w konwencji mnożenia macierzy

$$\langle x, y\rangle = x^T y$$

nazywamy *standardowym iloczynem skalarnym* w $\mathbb{R}^n$. Należy jednak mieć na uwadze, że iloczyn skalarny można wprowadzić na wiele sposobów, na przykład funkcja $\langle x, y\rangle_G = x^T G y$, gdzie G jest macierzą określoną dodatnio jest również iloczynem skalarnym w $\mathbb{R}^n$.

- $\|x\|$ oznacza *normę* (ang. *norm*) wektora $x \in \mathbb{R}^n$. Dowolny iloczyn skalarny $\langle \cdot, \cdot \rangle$ indukuje normę $\| \cdot \|$ wzorem $\|x\| := \sqrt{\langle x, x\rangle}$. Jeśli nie będzie powiedziane inaczej, symbolem $\|x\|$ oznaczać będziemy *normę euklidesową* (ang. *Euclidean norm*) wektora x, czyli normę indukowaną przez standardowy iloczyn skalarny, czyli $\|x\| = \sqrt{x^T x}$. Na marginesie przypomnijmy, że normę można również wprowadzić w przestrzeniach liniowych bez iloczynu skalarnego.
- Mówimy, że wektory $x, y \in \mathbb{R}^n$ są wzajemnie *ortogonalne* (ang. *orthogonal*), jeśli $\langle x, y\rangle = 0$.
- Dla dowolnego iloczynu skalarnego $\langle \cdot, \cdot \rangle$ i indukowanej przez niego normy $\| \cdot \|$ zachodzi *nierówność Cauchy'ego–Schwarza*

$$\langle x, y\rangle^2 \leq \|x\|^2 \cdot \|y\|^2$$

przy czym równość zachodzi wtedy i tylko wtedy, gdy wektory x, y są liniowo zależne. W szczególności,

  - $\langle x, y\rangle = \|x\| \cdot \|y\|$ wtedy i tylko wtedy, gdy x i y są dodatnio liniowo zależne (mają ten sam zwrot), czyli $x = \alpha y$ lub $y = \alpha x$ dla pewnego $\alpha > 0$.
  - $\langle x, y\rangle = -\|x\| \cdot \|y\|$ wtedy i tylko wtedy, gdy x i y są ujemnie liniowo zależne (mają przeciwny zwrot), czyli $x = \alpha y$ lub $y = \alpha x$ dla pewnego $\alpha < 0$.

- Dla normy $\| \cdot \|$ indukowanej przez iloczyn skalarny zachodzi *tożsamość równoległoboku*

$$2\|x\|^2 + 2\|y\|^2 = \|x + y\|^2 + \|x - y\|^2.$$

- $e_j = (0, ..., 0, 1, 0, ..., 0)$ oznacza j-ty *wersor*, tzn. element przestrzeni euklidesowej odpowiedniego wymiaru, którego j-ta współrzędna jest równa 1, zaś pozostałe są równe 0).
- $e = (1, ..., 1)$ oznacza wektor odpowiedniego wymiaru o wszystkich współrzędnych równych 1.

- Mówimy, że układ wektorów $\{a_1, a_2, ..., a_m\} \subseteq \mathbb{R}^n$ jest *liniowo niezależny* (ang. *linearly independent system*) lub w skrócie $a_1, a_2, ..., a_m$ są liniowo niezależne, jeśli dla dowolnych stałych $\alpha_1, \alpha_2, ..., \alpha_m \in \mathbb{R}$

$$\alpha_1 a_1 + \alpha_2 a_2 + ... + \alpha_m a_m = 0 \Longrightarrow a_1 = a_2 = ... = a_m = 0.$$

- Mówimy, że układ wektorów $\{a_1, a_2, ..., a_m\} \subseteq \mathbb{R}^n$ jest *liniowo zależny* (ang. *linearly dependent system*) lub w skrócie $a_1, a_2, ..., a_m$ są liniowo zależne, jeśli istnieją stałe $\alpha_1, \alpha_2, ..., \alpha_m \in \mathbb{R}$ takie, że

$$\alpha_1 a_1 + \alpha_2 a_2 + ... + \alpha_m a_m = 0 \text{ i przynajmniej jedna z tych stałych } \alpha_j \neq 0.$$

- Liczbę elementów maksymalnego układu liniowo niezależnego nazywamy *wymiarem* (ang. *dimension*) przestrzeni. Przestrzeń $\mathbb{R}^n$ ma wymiar n.

- Mówimy, że układ wektorów $\{a_1, a_2, ..., a_m\} \subseteq \mathbb{R}^n$ jest *ortogonalny* (ang. *orthogonal*) jeśli $\langle a_i, a_j \rangle = 0$ dla $i, j = 1, 2, ..., m$, $i \neq j$ i $\|a_i\| = 1$ dla $i = 1, 2, ..., m$. W niektórych podręcznikach taki układ nazywa się *ortonormalny*.

- Ortogonalny układ wektorów jest liniowo niezależny.

- I oznacza *macierz jednostkową* (ang. *identity matrix*) odpowiedniego wymiaru.

Niech A będzie macierzą typu $m \times n$. Wówczas:

- A^T oznacza *transpozycję* (ang. *transposition*) macierzy A (macierz transponowaną względem A).

- Macierz kwadratową A spełniającą równość $A^T = A$ nazywamy macierzą symetryczną (ang. *symetric*)

- A_j oznacza j-tą kolumnę macierzy A.

- A_J oznacza podmacierz utworzoną z kolumn A_j macierzy A, $j \in J \subseteq \{1, ..., n\}$.

- A^i oznacza i-ty wiersz macierzy A.

- A nazywa się *macierzą diagonalną* (ang. *diagonal matrix*), jeśli dla $i = 1, 2, ..., m$, $j = 1, 2, ..., n$, $i \neq j$, zachodzi $a_{ij} = 0$. Niech $d = (d_1, ..., d_n) \in \mathbb{R}^n$. Macierz diagonalna nie musi być kwadratowa. Macierz diagonalną kwadratową o elementach na głównej przekątnej równych d_j, $j = 1, 2, ..., n$, czyli macierz postaci

$$\begin{bmatrix} d_1 & 0 & ... & 0 \\ 0 & d_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & d_n \end{bmatrix}$$

  oznaczamy symbolem $\operatorname{diag} d$.

- *Minorem* macierzy A nazywamy wyznacznik jej podmacierzy kwadratowej powstałej przez skreślenie odpowiedniej liczby wierszy i kolumn macierzy A.

- Najwyższy stopień niezerowego minora macierzy A nazywamy się *rzędem* (ang. *rank*) macierzy A i oznaczamy symbolem $\mathrm{r}(A)$.

- Rząd macierzy A jest liczbą liniowo niezależnych wierszy bądź kolumn tej macierzy.

- Mówimy, że A ma *pełny rząd kolumnowy* (ang. *full column rank*) jeśli kolumny macierzy A są liniowo niezależne.

- Mówimy, że A ma *pełny rząd wierszowy* (ang. *full row rank*), jeśli wiersze macierzy A są liniowo niezależne.

- Ax jest kombinacją liniową kolumn macierzy A, gdzie $x \in \mathbb{R}^n$ jest wektorem współczynników tej kombinacji.

- $u^T A$ (lub $A^T u$) jest kombinacją liniową wierszy macierzy A, gdzie $u \in \mathbb{R}^m$ jest wektorem współczynników tej kombinacji.

- Podzbiór $X \subseteq \mathbb{R}^n$ nazywa się *podprzestrzenią liniową*, jeśli dla dowolnych $x, y \in X$ i dla dowolnych $\alpha, \beta \in \mathbb{R}$ zachodzi $\alpha x + \beta y \in X$.

- Zbiór $\ker A := \{x \in \mathbb{R}^n : Ax = 0\}$ nazywamy *jądrem* albo *przestrzenią zerową* (ang. *kernel* albo *nullspace*) macierzy A.

- Zbiór $\operatorname{im} A := \{y \in \mathbb{R}^m : \exists_{x \in \mathbb{R}^n} y = Ax\}$ nazywamy *obrazem* (ang: *image* albo *range*) macierzy A.

- Zarówno $\ker A$ jak i $\operatorname{im} A$ są podprzestrzeniami liniowymi.

- Podzbiór $X \subseteq \mathbb{R}^n$ nazywa się *podprzestrzenią afiniczną*, jeśli dla dowolnych $x, y \in X$ i dla dowolnego $\alpha \in \mathbb{R}$ zachodzi $(1 - \alpha)x + \alpha y \in X$. Przykładem podprzestrzeni afinicznej jest zbiór rozwiązań układu równań liniowych $Ax = b$, gdzie A jest macierzą typu $m \times n$, $x \in \mathbb{R}^n$ i $b \in \mathbb{R}^m$.

- Dla dowolnego $x \in \mathbb{R}^n$ istnieją wektory $u \in \ker A$ i $v \in \operatorname{im} A^T$ takie, że $x = u + v$. Rozkład ten jest jednoznaczny i $u^T v = 0$. Nazywamy go *rozkładem ortogonalnym* (ang. orthogonal decomposition). Fakt ten zapisujemy w postaci $\mathbb{R}^n = \ker A \oplus \operatorname{im} A^T$. Analogicznie $\mathbb{R}^m = \ker A^T \oplus \operatorname{im} A$. Te fakty możemy również zapisać w postaci $(\ker A)^\perp = \operatorname{im} A^T$ oraz $(\operatorname{im} A)^\perp = \ker A^T$, czyli dopełnieniem ortogonalnym jądra macierzy A jest obraz macierzy A^T oraz dopełnieniem ortogonalnym obrazu macierzy A jest jądro macierzy A^T.

- Dla macierzy kwadratowej A typu $n \times n$, jeśli $Ax = \lambda x$ dla $x \neq 0$, to $\lambda \in \mathbb{C}$ nazywa się *wartością własną* (ang. *eigenvalue*) macierzy A, zaś $x \in \mathbb{R}^n$ odpowiadającym jej *wektorem własnym* (ang. *eigenvector*). Wartość własna macierzy A może być liczbą zespoloną nawet jeśli A jest macierzą rzeczywistą.

- Wartości własne macierzy symetrycznej są liczbami rzeczywistymi.

- Macierz $G := A^T A$ nazywamy *macierzą Grama* (ang. *Gram matrix*) kolumn macierzy A. Elementami macierzy Grama są iloczyny skalarne kolumn macierzy A, czyli $g_{ij} = A_i^T A_j$. Jeśli A jest pełnego rzędu kolumnowego, to G jest nieosobliwa. Podobnie, macierz AA^T jest macierzą Grama wierszy macierzy A. Jeśli A jest pełnego rzędu wierszowego, to AA^T jest nieosobliwa.

- Macierz A^+ spełniającą warunki

(i) $AA^+A = A$,

(ii) $A^+AA^+ = A^+$,

(iii) $(AA^+)^T = AA^+$

(iv) $(A^+A)^T = A^+A$

nazywamy *macierzą pseudoodwrotną Moore'a–Penrose'a* macierzy A (ang. *Moore–Penrose pseudoinverse*). Jest to formalna definicja, natomiast nietrudno zauważyć, że jeśli A jest macierzą o pełnym rzędzie kolumnowym, to $A^+ = (A^TA)^{-1}A^T$, jeśli A jest macierzą o pełnym rzędzie wierszowym, to $A^+ = A^T(AA^T)^{-1}$, jeśli zaś A jest macierzą nieosobliwą, to $A^+ = A^{-1}$. W ogólnym przypadku, w celu wyznaczenia macierzy pseudoodwrotnej Moore'a–Penrose'a dla danej macierzy A używa się tzw. *rozkładu według wartości osobliwych* (ang. *singular value decomposition* – SVD): $A = U\Sigma V^T$, gdzie U jest macierzą ortogonalną typu $m \times m$, Σ jest macierzą diagonalną typu $m \times n$ o nieujemnych elementach σ_j na głównej przekątnej, $j = 1, 2, ..., p$, gdzie $p = \min\{m, n\}$, zwanych *wartościami osobliwymi* (ang. *singular values*), oraz V jest macierzą ortogonalną typu $n \times n$. Zazwyczaj rozkład ten przedstawia się w taki sposób, aby $\sigma_1 \geq ... \geq \sigma_r > \sigma_{r+1} = ... = \sigma_p = 0$. Macierz pseudoodwrotna Moore'a–Penrose'a macierzy A ma postać $A^+ = V\Sigma^+U^T$, gdzie Σ^+ jest macierzą pseudoodwrotną Moore'a–Penrose'a macierzy diagonalnej Σ powstałą z macierzy Σ przez odwrócenie niezerowych elementów na głównej przekątnej i transponowanie powstałej w ten sposód macierzy.

- Symbolem $\|A\|$ oznaczamy *normę macierzy* A, tzn.

$$\|A\| := \sup_{\|x\|=1} \|Ax\|. \tag{1.3}$$

Norma macierzy zależy od norm przyjętych w $\mathbb{R}^n$ i w $\mathbb{R}^m$. Dla macierzy A i B odpowiednich typów zachodzi $\|AB\| \leq \|A\| \cdot \|B\|$. Jeśli obie normy w (1.3) są normami euklidesowymi, to $\|A\|$ nazywamy *normą spektralną* (ang. *spectral norm*) macierzy A. Dla normy spektralnej macierzy A zachodzi równość $\|A\| = \sqrt{\lambda_{\max}(A^TA)}$, gdzie $\lambda_{\max}(A^TA)$ oznacza największą wartość własną (promień spektralny) macierzy A^TA.

- Jeśli A jest macierzą nieosobliwą, to wielkość $\operatorname{cond} A = \|A^{-1}\| \cdot \|A\|$ nazywamy *wskaźnikiem uwarunkowania* (ang. *condition number*) macierzy A. Jeśli $\operatorname{cond} A \gg 1$, to macierz A nazywamy *źle uwarunkowaną* (ang. *ill conditioned*). W takim przypadku, rozwiązując układ równań $Ax = b$ mała (względna) zmiana prawej strony powoduje dużą (względną) zmianę rozwiązania.

Niech A będzie macierzą kwadratową.

- $\lambda_{\max}(A)$ i $\lambda_{\min}(A)$ oznaczają odpowiednio największą i najmniejszą *wartość własną* (ang. *eigenvalue*) macierzy symetrycznej A.

- Wyznacznik macierzy powstałej z macierzy A przez skreślenie tych samych wierszy i kolumn nazywa się *minorem głównym* (ang. *main minor*). Minor główny stopnia i powstały przez skreślenie ostatnich $n-i$ wierszy i kolumn tej macierzy nazywamy *wiodącym minorem głównym* (ang. *leading main minor*) i oznaczamy symbolem Δ_i, $i = 0, 1, ..., n-1$.

- Funkcja $f : \mathbb{R}^n \to \mathbb{R}$ postaci $f(x) = a^Tx$, gdzie $a \in \mathbb{R}^n$, nazywa się funkcją liniową.

- Funkcja $f : \mathbb{R}^n \to \mathbb{R}$ postaci $f(x) = a^Tx + b$, gdzie $a \in \mathbb{R}^n$ i $c \in \mathbb{R}$, nazywa się funkcją afiniczną.

- Operator $A : \mathbb{R}^n \to \mathbb{R}^m$ nazywa się *operatorem afinicznym*, jeśli dla dowolnych $x, y \in \mathbb{R}^n$ i dla dowolnego $\alpha \in \mathbb{R}$ zachodzi $A((1-\alpha)x + \alpha y) = (1-\alpha)A(x) + \alpha A(y)$. Przykładem operatora afinicznego jest operator $T : \mathbb{R}^n \to \mathbb{R}^m$ określony równością $T(x) = Ax + b$, gdzie A jest macierzą typu $m \times n$, $x \in \mathbb{R}^n$ i $b \in \mathbb{R}^m$.

- Funkcja $f : \mathbb{R}^n \to \mathbb{R}$ postaci $f(x) = \frac{1}{2}x^T Ax + b^T x + c$, gdzie A jest macierzą symetryczną, $b \in \mathbb{R}^n$ i $c \in \mathbb{R}$ nazywa się *funkcją kwadratową* (ang. *quadratic function*).

Ćwiczenie 1.4.1 Sprowadzić funkcję kwadratową $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \frac{1}{2}x^T Ax$ do postaci symetrycznej, tzn. $f(x) = \frac{1}{2}x^T Gx$, gdzie G jest macierzą symetryczną.

W dalszych rozważaniach ograniczać się będziemy do macierzy rzeczywistych.
Będziemy również używać zapisu blokowego macierzy, np.

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \text{ lub } \begin{bmatrix} A_1 & A_2 & \cdots & A_n \end{bmatrix} \text{ lub } \begin{bmatrix} A^1 \\ A^2 \\ \vdots \\ A^m \end{bmatrix},$$

gdzie A, B, C, D, A_j, $j = 1, 2, ..., n$, i A^i, $i = 1, 2, ..., m$, są macierzami. W tym przypadku należy pamiętać o tym, aby macierze występujące w jednym wierszu macierzy blokowej miały tę samą liczbę wierszy i macierze wystepujące w tej samej kolumnie macierzy blokowej miały tę samą liczbę kolumn. Dla macierzy blokowych stosuje się reguły mnożenia takie same jak dla zwykłych macierzy, np.

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{bmatrix} E \\ F \end{bmatrix} = \begin{bmatrix} AE + BF \\ CE + DF \end{bmatrix}.$$

Należy przy tym zadbać o to, aby wszystkie działania po prawej strony powyższej równości były dobrze zdefiniowane.

Definicja 1.4.2 Macierz kwadratowa U nazywa się *macierzą ortogonalną* (ang. *orthogonal matrix*), jeśli

$$U^T U = I,$$

czyli kolumny macierzy U są wzajemnie ortogonalne i unormowane.

Twierdzenie 1.4.3 *Wiersze macierzy ortogonalnej są również wzajemnie ortogonalne i unormowane.*

Z powyższej definicji i z powyższego twierdzenia wynika, że aby odwrócić macierz ortogonalną wystarczy ją transponować, $U^{-1} = U^T$.

Definicja 1.4.4 Macierz L postaci

$$L = \begin{bmatrix} l_{11} & 0 & ... & 0 \\ l_{21} & l_{22} & ... & 0 \\ ... & ... & ... & ... \\ l_{m1} & l_{m2} & ... & l_{mm} \end{bmatrix}$$

nazywa się *macierzą dolną trójkątną* (ang. *lower triangular matrix*).

Definicja 1.4.5 Niech A będzie macierzą symetryczną typu $n \times n$. Mówimy, że A jest *określona nieujemnie* (ang. *nonnegative definite* albo *positive semidefinite)*, jeśli

$$\forall_{s \in \mathbb{R}^n} \quad s^T A s \geq 0.$$

Mówimy, że A jest *określona dodatnio* (ang. *positive definite*), jeśli

$$\forall_{s \in \mathbb{R}^n, s \neq 0} \quad s^T A s > 0.$$

Jeśli powyższe nierówności zachodzą w stronę przeciwną, to w pierwszym przypadku mówimy, że macierz A jest *określona niedodatnio* (ang. *nonpositive definite* albo *negative semidefinite*), zaś w drugim, że jest ona *określona ujemnie* (ang. *negative definite*). Macierz, która nie jest określona ani nieujemnie ani niedodatnio nazywa się macierzą *nieokreśloną*.

Przypomnimy teraz pewne fakty dotyczące macierzy dodatnio określonych. Dowody albo wynikają prosto z definicji albo można znaleźć w podręcznikach dotyczących algebry liniowej bądź numerycznej algebry liniowej.

Twierdzenie 1.4.6 *Macierz Grama $A^T A$ jest określona nieujemnie. Ponadto jest ona określona dodatnio wtedy i tylko wtedy, gdy macierz A jest pełnego rzędu kolumnowego (kolumny macierzy A są liniowo niezależne).*

Uwaga 1.4.7 Wróćmy do rozkładu ortogonalnego przestrzeni $\mathbb{R}^m$, $y = u + v$, gdzie $u \in \ker A^T$, $v \in \operatorname{im} A$ i $u^T v = 0$. W przypadku, gdy macierz A jest pełnego rzędu kolumnowego, nietrudno zauważyć, że równość ta jest spełniona dla

$$u = y - A(A^T A)^{-1} A^T y \quad \text{i} \quad v = A(A^T A)^{-1} A^T y. \tag{1.4}$$

Zgodnie z poprzednimi informacjami, w tym przypadku macierz $(A^T A)^{-1} A^T$ jest macierzą pseudoodwrotną Moore'a–Penrose'a macierzy A. Równości w (1.4) możemy napisać w postaci $u = y - AA^+ y$ i $v = AA^+ y$. Jeśli A nie jest pełnego rzędu kolumnowego, we wzorach na u i v wystarczy zamiast A wziąć dowolną macierz złożoną z liniowo niezależnych kolumn macierzy A rozpinających przestrzeń $\operatorname{im} A$. Podobnie, dla rozkładu ortogonalnego przestrzeni $\mathbb{R}^n$, $x = u + v$, gdzie $u \in \ker A$, $v \in \operatorname{im} A^T$ i $u^T v = 0$ oraz A jest pełnego rzędu wierszowego, mamy

$$u = x - A^T(AA^T)^{-1} A x \quad \text{i} \quad v = A^T(AA^T)^{-1} A x. \tag{1.5}$$

Zgodnie z poprzednimi informacjami, w tym przypadku macierz $A^T(AA^T)^{-1}$ jest macierzą pseudoodwrotną Moore'a–Penrose'a A^+. Równości w (1.5) możemy również napisać w postaci $u = x - A^+ A x$ i $v = A^+ A x$. Do wzorów (1.4)-(1.5) wrócimy w rozdziale 2.

Twierdzenie 1.4.8 *Dla macierzy symetrycznej A stopnia n następujące warunki są równoważne:*

(i) *A jest określona dodatnio;*

(ii) *wszystkie wartości własne macierzy A są dodatnie;*

(iii) *istnieje macierz ortogonalna U i macierz diagonalna D o dodatnich elementach na głównej przekątnej takie, że $A = UDU^T$ (tzw. rozkład według wartości własnych, ang. eigenvalue decomposition);*

(iv) *istnieje rozkład Cholesky'ego* (ang. *Cholesky factorization*) *macierzy A, postaci $A = LL^T$, gdzie L jest dolną macierzą trójkątną, taki, że wszystkie elementy diagonalne macierzy L są dodatnie: $l_{ii} > 0$ dla $i = 1, 2, ..., n$;*

(v) *wiodące minory główne* $\Delta_1, \Delta_2, ..., \Delta_n$ *macierzy* A *są dodatnie;*

(vi) *istnieje określona dodatnio macierz* G *taka, że* $A = G^2$.

Uwaga 1.4.9 Jeśli w powyższym twierdzeniu słowo „dodatnie" zastąpimy przez „nieujemne", to będą zachodzić związki:

$$\text{(i)} \Leftrightarrow \text{(ii)} \Leftrightarrow \text{(iii)} \Leftrightarrow \text{(vi)}, \text{(iv)} \Rightarrow \text{(i) oraz (iv)} \Rightarrow \text{(v)}$$

Nie będą natomiast prawdziwe implikacje (v)⇒(i) (patrz przykład 1.4.13) oraz (v)⇒(iv). Będą one jednak prawdziwe, jeśli w warunku (v) założymy, że wszystkie minory główne (a nie tylko wiodące minory główne $\Delta_1, \Delta_2, ..., \Delta_n$) macierzy A będą nieujemne. Sprawdzenie tego warunku dla macierzy większego wymiaru jest jednak kłopotliwe (wszystkich minorów głównych macierzy kwadratowej stopnia n jest $2^n - 1$).

Uwaga 1.4.10 Z uwagi na to, że dla macierzy ortogonalnej $U = [u_1, u_2, ...u_n]$ o kolumnach u_j, $j = 1, 2, ..., n$, zachodzi $U^T = U^{-1}$, rozkład na wartości własne możemy zapisać w postaci $AU = UD$, albo inaczej, $Au_j = d_j u_j$, $j = 1, 2, ..., n$. Widzimy więc, że d_j jest wartością własną macierzy A, zaś u_j odpowiadającym jej wektorem własnym, $j = 1, 2, ..., n$.

Uwaga 1.4.11 Istnieje ścisły związek między rozkładem według wartości osobliwych macierzy A typu $m \times n$, a rozkładem według wartości własnych macierzy A^TA i AA^T. Jeśli bowiem $A = U\Sigma V^T$, gdzie U i V są macierzami ortogonalnymi i Σ macierzą diagonalną o nierosnących elementach na głównej przekątnej σ_j, czyli spełniających $\sigma_1 \geq ... \geq \sigma_r > \sigma_{r+1} = ... = \sigma_p = 0$, to $A^TA = V\Sigma^T\Sigma V^T$. Nietrudno zauważyć, że $\Sigma^T\Sigma$ jest macierzą diagonalną typu $n \times n$ o elementach diagonalnych równych σ_j^2 dla $j = 1, 2, ..., r$ i $\sigma_j = 0$ dla $j = r+1, r+2, ..., n$. Zatem wielkości σ_j^2, $j = 1, 2, ..., n$, są wartościami własnymi macierzy A^TA, zaś kolumny macierzy V są odpowiadającymi im wektorami własnymi. Analogocznie, $AA^T = U\Sigma\Sigma^TU^T$, czyli wielkości σ_j^2,dla $j = 1, 2, ..., r$ i $\sigma_j = 0$ dla $j = r+1, r+2, ..., m$, są wartościami własnymi macierzy AA^T, zaś kolumny macierzy U są odpowiadającymi im wektorami własnymi.

Uwaga 1.4.12 Z równoważności (i)⇔(vi) w twierdzeniu 1.4.8 wynika, że macierz A określoną dodatnio (nieujemnie) można przedstawić jako macierz Grama kolumn pewnej symetrycznej macierzy G określonej dodatnio (nieujemnie). Dla macierzy A określonej dodatnio (nieujemnie) możemy więc formalnie zdefiniować $A^{\frac{1}{2}} := G$, gdzie G jest macierzą określoną dodatnio (nieujemnie) taką, że $A = G^2$.

Przykład 1.4.13 Dla macierzy symetrycznej

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

wiodące minory główne $\Delta_1, \Delta_2, \Delta_3$ są nieujemne (w istocie $\Delta_i = 0$, $i = 1, 2, 3$), natomiast macierz ta jest nieokreślona: dla $s = (0, 0, 1)$ mamy $s^TAs > 0$ zaś dla $s = (0, -1, 1)$ mamy $s^TAs < 0$. Zwróćmy uwagę na fakt, że minor główny powstały przez skreślenie pierwszego wiersza i pierwszej kolumny macierzy A jest ujemny.

Wniosek 1.4.14 *Jeśli macierz symetryczna* A *typu* $m \times m$ *jest określona nieujemnie, to dla dowolnego* $s \in \mathbb{R}^n$ *zachodzą nierówności*

$$\lambda_{\min}(A)\|s\|^2 \leq s^TAs \leq \lambda_{\max}(A)\|s\|^2, \tag{1.6}$$

gdzie $\lambda_{\min}(A)$ *i* $\lambda_{\max}(A)$ *są najmniejszą i największą wartością własną macierzy* A. *Ponadto, pierwsza bądź druga nierówność staje się równością dla* s *będącego wektorem własnym odpowiadającym wartości własnej* $\lambda_{\min}(A)$ *bądź* $\lambda_{\max}(A)$.

Dowód. Z twierdzenia 1.4.8 i z uwagi 1.4.9 wynika, że $A = UDU^T$, gdzie U jest macierzą ortogonalną, żaś $D = \mathrm{diag}(d)$, przy czym współrzędne d_i, $i = 1, 2, ..., m$, są nieujemne. Oczywiście $\lambda_{\max}(A) = \max_{i=1,2,...,m} d_i$ i $\lambda_{\min}(A) = \min_{i=1,2,...,m} d_i$. Oznaczając $u = U^T s$ mamy

$$s^T As = s^T UDU^T s = u^T Du = \sum_{i=1}^{m} d_i u_i^2.$$

Ponieważ U jest ortogonalna, więc $\|u\|^2 = \|s\|^2$, a więc

$$\lambda_{\min}(A)\|u\|^2 \leq s^T As \leq \lambda_{\max}(A)\|u\|^2,$$

co kończy dowód pierwszej części. Drugą część można sprawdzić bezpośrednio. ■

Stosując wniosek 1.4.14 do macierzy $A^T A$ otrzymujemy natychmiast:

Wniosek 1.4.15 *Dla normy spektralnej macierzy A zachodzi równość $\|A\|^2 = \lambda_{\max}(A^T A)$.*

Twierdzenie 1.4.16 *Jeśli G jest macierzą określoną dodatnio, to zachodzi równość* $\mathrm{cond}\, G = \frac{\lambda_{\max}(G)}{\lambda_{\min}(G)}$.

Dowód. Wystarczy zastosować równoważność (i)$\Longleftrightarrow$(iii) w twierdzeniu 1.4.8 oraz zauważyć że dla macierzy diagonalnej $D = \mathrm{diag}(d_1, ..., d_n)$, gdzie $d_i > 0$, $i = 1, ..., n$, mamy $\lambda_{\max}(D) = \max\{d_i : i = 1, ..., n\}$ oraz $\lambda_{\max}(D^{-1}) = (\lambda_{\min}(D))^{-1}$. Szczegóły pozostawiamy czytelnikowi. ■

1.5 Elementy różniczkowania funkcji wielu zmiennych

Pojęcia i fakty podane w tym ustępie są przypomnieniem lub uzupełnieniem odpowiednich definicji i twierdzeń z analizy matematycznej, dotyczących w szczególności różniczkowania funkcji wielu zmiennych.

1.5.1 Podstawowe oznaczenia, definicje i fakty

- Funkcja $q : \mathbb{R} \to \mathbb{R}$ jest wielkością rzędu $o(t)$, gdy $\lim_{t\to 0} \frac{q(t)}{t} = 0$.
- Funkcja $q : \mathbb{R} \to \mathbb{R}$ jest wielkością rzędu $O(t)$, gdy $|q(t)| \leq m\,|t|$ dla pewnej stałej $m > 0$.
- Funkcja $f : \mathbb{R}^n \to \mathbb{R}$ jest *koercytywna* (ang. *coercive*), jeśli $\lim_{\|x\|\to+\infty} f(x) = +\infty$.
- Operator $F : \mathbb{R}^n \to \mathbb{R}^m$ jest *ciągły w sensie Lipschitza* (ang. *Lipschitz continuous*) jeśli istnieje stała $L > 0$ taka, że

  $$\|F(x) - F(y)\| \leq L\|x - y\|.$$

  Mówimy również, że F jest operatorem L-lipschitzowskim.
- Jeśli istnieją wszystkie pochodne cząstkowe funkcji $f : \mathbb{R}^n \to \mathbb{R}$, to wektor

  $$g(x) = \nabla f(x) = (\frac{\partial f}{\partial x_1}(x), ..., \frac{\partial f}{\partial x_n}(x))$$

  nazywamy *gradientem* (ang. *gradient*) funkcji f w punkcie x (dla skrócenia zapisu będziemy używać konwencji $g = g(x)$, $g^* = g(x^*)$, $g' = g(x')$, $g_k = g(x_k)$, itd.). Zgodne z przyjętą konwencją, gradient $\nabla f(x)$ identyfikujemy z wektorem kolumnowym, tzn.

  $$\nabla f(x) = [\frac{\partial f}{\partial x_1}(x), ..., \frac{\partial f}{\partial x_n}(x)]^T.$$

- Jeśli istnieją wszystkie pochodne cząstkowe rzędu drugiego funkcji $f : \mathbb{R}^n \to \mathbb{R}$, to macierz

  $$G(x) = \nabla^2 f(x) = \nabla(\nabla f(x)^T) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & ... & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ ... & ... & ... \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & ... & \frac{\partial^2 f}{\partial x_n^2}(x) \end{bmatrix}$$

  nazywamy *macierzą Hessego* albo *hesjanem* (ang. *Hessian* lub *Hesse matrix*) funkcji $f : \mathbb{R}^n \to \mathbb{R}$ w punkcie x (dla skrócenia zapisu będziemy używać konwencji $G = G(x)$, $G^* = G(x^*)$, $G' = G(x')$, $G_k = G(x_k)$, itd.).
- Jeśli pochodne cząstkowe rzędu drugiego funkcji f są ciągłe w punkcie x, to hesjan $\nabla^2 f(x)$ jest macierzą symetryczną (twierdzenie Sylwestra).
- Niech $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, czyli $h = (h_1, ..., h_m)$, gdzie $h_j : \mathbb{R}^n \longrightarrow \mathbb{R}$, $j = 1, 2, ..., n$, są funkcjami posiadającymi wszystkie pochodne cząstkowe. Macierz

  $$J(x) := \begin{bmatrix} \frac{\partial h_1}{\partial x_1}(x) & ... & \frac{\partial h_1}{\partial x_n}(x) \\ ... & ... & ... \\ \frac{\partial h_m}{\partial x_1}(x) & ... & \frac{\partial h_m}{\partial x_n}(x) \end{bmatrix}$$

nazywamy *macierzą Jacobiego* odwzorowania h w punkcie $x \in \mathbb{R}^n$. Jej transpozycję oznaczamy symbolem ∇h. Zachodzą więc równości

$$\nabla h(x) = [\nabla h_1(x), ..., \nabla h_m(x)] = \begin{bmatrix} \frac{\partial h_1}{\partial x_1}(x) & ... & \frac{\partial h_m}{\partial x_1}(x) \\ ... & ... & ... \\ \frac{\partial h_1}{\partial x_n}(x) & ... & \frac{\partial h_m}{\partial x_n}(x) \end{bmatrix}.$$

- Wektor $\nabla_x r(\bar{x}, \bar{y}) = (\frac{\partial r}{\partial x_1}(\bar{x}, \bar{y}), ..., \frac{\partial r}{\partial x_n}(\bar{x}, \bar{y}))$ oznacza gradient funkcji $r : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ w punkcie $(\bar{x}, \bar{y})$ względem $x \in \mathbb{R}^n$,
- Macierz $\nabla_x p(\bar{x}, \bar{y}) = (\nabla_x p_1(\bar{x}, \bar{y}), ..., \nabla_x p_r(\bar{x}, \bar{y}))$ oznacza transpozycję macierzy Jacobiego odwzorowania $p : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^r$, $p = (p_1, ..., p_r)$ w punkcie $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ względem $x \in \mathbb{R}^n$,
- Macierz $\nabla^2_{xx} r(\bar{x}, \bar{y}) = \nabla_x(\nabla_x r(\bar{x}, \bar{y})^T)$ oznacza hesjan funkcji $r : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ w punkcie $(\bar{x}, \bar{y})$ względem $x \in \mathbb{R}^n$.

1.5.2 Funkcje różniczkowalne i ich własności

Definicja 1.5.1 Niech $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ i niech $\bar{x}, s \in \mathbb{R}^n$. Granicę

$$\lim_{t \downarrow 0} \frac{f(\bar{x} + ts) - f(\bar{x})}{t}$$

nazywamy *pochodną kierunkową* (ang. *directional derivative*) funkcji f w punkcie $\bar{x}$ w kierunku wektora s i oznaczamy ją symbolem $f'(\bar{x}, s)$. Podobnie definiujemy pochodną kierunkową odwzorowania $h : \mathbb{R}^n \to \mathbb{R}^m$.

Definicja 1.5.2 Funkcja $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ nazywa się *różniczkowalna* (*w sensie Frécheta*) (ang. *Fréchet differentiable*) w punkcie $x \in \mathbb{R}^n$, jeśli istnieje wektor $g \in \mathbb{R}^n$, taki że

$$f(x + d) = f(x) + g^T d + o(\|d\|), \tag{1.7}$$

gdzie $d \in \mathbb{R}^n$.

Twierdzenie 1.5.3 *Jeśli funkcja $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ jest różniczkowalna w punkcie x, to istnieje gradient $\nabla f(x)$ i zachodzi równość*

$$f(x + d) = f(x) + \nabla f(x)^T d + o(\|d\|),$$

gdzie $d \in \mathbb{R}^n$.

Definicja 1.5.4 Odwzorowanie $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ nazywa się *różniczkowalne* (*w sensie Frécheta*) w punkcie $x \in \mathbb{R}^n$, jeśli istnieje macierz J typu $m \times n$ taka, że

$$h(x + d) = h(x) + Jd + o(\|d\|), \tag{1.8}$$

gdzie $d \in \mathbb{R}^n$.

Twierdzenie 1.5.5 *Jeśli odwzorowanie $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ jest różniczkowalne w punkcie $x \in \mathbb{R}^n$, to istnieje macierz Jacobiego $\nabla h(x)^T$ i zachodzi równość*

$$h(x + d) = h(x) + \nabla h(x)^T d + o(\|d\|), \tag{1.9}$$

gdzie $d \in \mathbb{R}^n$.

Definicja 1.5.6 Funkcja $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ nazywa się *różniczkowalna w sensie Gâteaux* (ang. *Gâteaux differentiable*) w punkcie $x \in \mathbb{R}^n$, jeśli dla dowolnego wektora $d \in \mathbb{R}^n$ istnieje pochodna kierunkowa $f'(x,d)$ oraz $f'(x,\cdot)$ jest funkcjonałem liniowym, czyli $f'(x,d) = g^T d$ dla pewnego wektora $g \in \mathbb{R}^n$ i dla dowolnego wektora $d \in \mathbb{R}^n$.

Podobnie definiuje się różniczkowalność w sensie Gâteaux dla odwzorowania $h : \mathbb{R}^n \to \mathbb{R}^m$. Szczegóły pomijamy.

Uwaga 1.5.7 Jeśli funkcja $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ jest różniczkowalna w sensie Gâteaux w punkcie x, to istnieje jej gradient $\nabla f(x)$ oraz $f'(x,d) = \nabla f(x)^T d$. Mamy bowiem dla $d = e_j$, $f'(x,e_j) = \nabla f(x)^T e_j = \frac{\partial f(x)}{\partial x_j}$, $j = 1,2,...,n$. Jednak istnienie gradientu w punkcie x nie gwarantuje różniczkowalności w sensie Gâteaux w ym punkcie. Co więcej istnienie gradientu w punkcie x nie implikuje nawet ciągłości funkcji w tym punkcie.

Przykład 1.5.8 Funkcja $f : \mathbb{R}^2 \to \mathbb{R}$ zdefiniowana wzorem

$$f(x,y) = \begin{cases} 0 & \text{gdy } x = 0 \text{ lub } y = 0, \\ 1 & \text{poza tym} \end{cases}$$

posiada gradient w punkcie $(0,0)$, ale nie jest w tym punkcie ciągła.

Twierdzenie 1.5.9 *Jeżeli funkcja $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ jest różniczkowalna (w sensie Frécheta) w punkcie $x \in \mathbb{R}^n$, to jest ona różniczkowalna w tym punkcie w sensie Gâteaux, czyli w tym punkcie istnieją wszystkie pochodne kierunkowe $f'(x,d)$ i zachodzi równość*

$$f'(x,d) = d^T \nabla f(x).$$

Uwaga 1.5.10 Z różniczkowalności w sensie Gâteaux nie wynika różniczkowalność w sensie Frécheta.

Ćwiczenie 1.5.11 Sprawdzić, że funkcja $f : \mathbb{R}^2 \to \mathbb{R}$ określona wzorem

$$f(x,y) = \begin{cases} \frac{x^3}{x^2+y^2} & \text{jeśli } (x,y) \neq (0,0) \\ 0 & \text{jeśli } (x,y) = (0,0) \end{cases}$$

jest różniczkowalna w sensie Gâteaux, a nie jest różniczkowalna w sensie Frécheta.

Przy pewnych założeniach spełnionych w przypadku minimalizacji gładkiej, oba typy różniczkowalności w sensie Gâteaux i w sensie Frécheta są sobie równoważne.

Twierdzenie 1.5.12 *Jeśli funkcja $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ jest różniczkowalna w sensie Gâteaux w pewnym otoczeniu punktu $x \in \mathbb{R}^n$ i jej gradient jest ciągły w punkcie x, to f jest różniczkowalna w sensie Frécheta w punkcie x.*

Twierdzenie 1.5.13 (o różniczkowalności funkcji złożonej) *Jeśli funkcja $h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ jest różniczkowalna w punkcie $x \in \mathbb{R}^n$ i funkcja $f : \mathbb{R}^m \longrightarrow \mathbb{R}$ jest różniczkowalna w punkcie $h(x)$, to funkcja $f \circ h$ jest różniczkowalna w punkcie x i zachodzi równość*

$$\nabla(f \circ h)(x) = \nabla h(x) \cdot \nabla f(h(x)). \tag{1.10}$$

Po prawej stronie powyższego wzoru występuje mnożenie macierzy, więc ich kolejność jest ważna. W niektórych podręcznikach gradienty zapisuje się jako wektory wierszowe, co wraz z własnością transpozycji macierzy powoduje, że we wzorze (1.10) zamienia się kolejność macierzy występujących po prawej stronie. Wzór (1.10) jest prawdziwy również w przypadku, gdy $f : \mathbb{R}^m \to \mathbb{R}^p$. Wówczas $\nabla f(y)^T$ jest macierzą Jacobiego funkcji f w punkcie $y \in \mathbb{R}^m$.

Wniosek 1.5.14 *Jeśli* $u, v : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ *są różniczkowalne w punkcie* $x \in \mathbb{R}^n$*, to funkcja* $u^T v : \mathbb{R}^n \longrightarrow \mathbb{R}$ *jest różniczkowalna w punkcie* x *i zachodzi wzór:*

$$\nabla(u^T v)(x) = \nabla u(x) \cdot v(x) + \nabla v(x) \cdot u(x).$$

Jeśli ponadto u, v *są dwukrotnie różniczkowalne w punkcie* $x \in \mathbb{R}^n$*, to* $u^T v$ *jest dwukrotnie różniczkowalna w punkcie* x *i zachodzi wzór:*

$$\nabla^2(u^T v)(x) = \nabla v(x) \cdot \nabla u(x)^T + \nabla^2 u(x) \cdot v(x) + \nabla u(x) \cdot \nabla v(x)^T + \nabla^2 v(x) \cdot u(x).$$

Dowód. Wystarczy określić funkcję $h : \mathbb{R}^n \to \mathbb{R}^{2m}$, $h(x) = (u(x), v(x))$ i funkcję $f : \mathbb{R}^{2m} \to \mathbb{R}$, $f(y) = \eta_1\eta_{m+1} + ... + \eta_m\eta_{2m}$, gdzie $y = (\eta_1, ..., \eta_{2m})$ i skorzystać z twierdzenia 1.5.13. Szczegóły pozostawiamy czytelnikowi jako ćwiczenie. ■

Definicja 1.5.15 Funkcja $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ nazywa się różniczkowalna w obszarze $D \subseteq \mathbb{R}^n$, jeśli jest ona różniczkowalna w każdym punkcie tego obszaru. Funkcja f nazywa się funkcją klasy $C_1(x)$ ($C_1(D)$), jeśli gradient ∇f jest odwzorowaniem ciągłym (lub inaczej pochodne cząstkowe są ciągłe) w punkcie x (w obszarze D). Jeśli gradient ten jest odwzorowaniem różniczkowalnym w punkcie x (w obszarze D), to mówimy, że funkcja f jest dwukrotnie różniczkowalna w punkcie x (w obszarze D). Jeśli hesjan jest odwzorowaniem ciągłym (lub inaczej pochodne cząstkowe rzędu drugiego są ciągłe) w punkcie x (w obszarze D), to mówimy, że funkcja f jest klasy $C_2(x)$ ($C_2(D)$).

Podobnie definiuje się funkcje klasy $C_k(x)$ ($C_k(D)$). W definicji 1.5.15 ograniczyliśmy się do co najwyżej dwukrotnej różniczkowalności funkcji wielu zmiennych, gdyż w dalszych rozważaniach zazwyczaj używać będziemy funkcji co najwyżej klasy C_2.

Podamy teraz twierdzenie o funkcji uwikłanej w postaci przydatnej w dalszej cześci.

Twierdzenie 1.5.16 (o funkcji uwikłanej) *Dane jest odwzorowanie* $r : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ *i punkt* $(\bar{x}, \bar{t})$*, taki że* $r(\bar{x}, \bar{t}) = 0$ *i* r *jest klasy* C_1 *w pewnym otoczeniu tego punktu. Jeśli macierz Jacobiego* $\nabla_x r(\bar{x}, \bar{t})$ *jest nieosobliwa, to istnieje otoczenie* U *punktu* $\bar{t}$ *i dokładnie jedno odwzorowanie* $h : \mathbb{R} \to \mathbb{R}^n$ *klasy* $C_1(U)$ *takie, że* $r(h(t), t) = 0$ *dla* $t \in U$*.*

Ćwiczenie 1.5.17 Niech $r : \mathbb{R}^2 \to \mathbb{R}$ będzie dana wzorem $r(x, y) = x^2 + y^2 - 25$ i niech $\bar{y} = 3$. Dla jakich $\bar{x}$ istnieje dokładnie jedna funkcja $h : \mathbb{R} \to \mathbb{R}$, $x = h(y)$, taka, że $h(\bar{y}) = \bar{x}$ i $r(h(y), y) = 0$ w pewnym otoczeniu $\bar{y}$? Podać wzór funkcji h.

Ćwiczenie 1.5.18 Niech $r : \mathbb{R}^3 \to \mathbb{R}^2$ będzie dana wzorem $r(x, y, z) = (x^2 + y^2 + z^2 - 26, x + 2y + 3z - 14)$ i niech $(\bar{x}, \bar{y}, \bar{z}) = (3, 4, 1)$. Czy istnieje dokładnie jedna funkcja $h : \mathbb{R} \to \mathbb{R}^2$ taka, że $r(h_1(z), h_2(z), z) = 0$ w pewnym otoczeniu $\bar{z}$?

Poniżej przypominamy rozwinięcia Taylora dla funkcji jednej zmiennej. Postaci te wystarczą do przedstawienia odpowiednich rozwinięć Taylora dla funkcji wielu zmiennych.

Twierdzenie 1.5.19 *Jeśli funkcja* $q : \mathbb{R} \longrightarrow \mathbb{R}$ *jest* k*-krotnie różniczkowalna na przedziale* $[x, x + h]$*, to zachodzi równość*

$$q(x + h) = q(x) + q'(x)h + \frac{1}{2}q''(x)h^2 + ... + \frac{1}{k!}q^{(k)}(x)h^k + o(h^k) \tag{1.11}$$

Twierdzenie 1.5.20 *Jeśli funkcja* $q : \mathbb{R} \longrightarrow \mathbb{R}$ *jest* $(k+1)$*-krotnie różniczkowalna na przedziale* $[x, x+h]$*, to istnieje liczba* $\lambda \in (0,1)$ *taka, że*

$$q(x+h) = q(x) + q'(x)h + \frac{1}{2}q''(x)h^2 + ...\frac{1}{k!}q^{(k)}(x)h^k + \frac{1}{(k+1)!}q^{(k+1)}(x+\lambda h)h^{k+1} \qquad (1.12)$$

Równość (1.11) nosi nazwę rozwinięcia funkcji q we *wzór Taylora z resztą Peana*, zaś równość (1.12) nosi nazwę rozwinięcia funkcji q we *wzór Taylora z resztą Lagrange'a*.

Ćwiczenie 1.5.21 Niech f będzie funkcją określoną na przestrzeni $\mathbb{R}^n$. Pokazać, że:

(a) jeśli $f(x) = a^T x$, gdzie $a \in \mathbb{R}^n$, to $\nabla f(x) = a$ i $\nabla^2 f(x) = 0$;

(b) jeśli $f(x) = Ax$, gdzie A jest macierzą typu $m \times n$, to $\nabla f(x) = A^T$, czyli A jest macierzą Jacobiego funkcji f;

(c) jeśli $p(x) = s^T \nabla f(x)$, gdzie $s \in \mathbb{R}^n$ i f jest funkcją dwukrotnie różniczkowalną, to $\nabla p(x) = \nabla^2 f(x)s$;

(d) jeśli $f(x) = x^T Ax$, gdzie A jest macierzą typu $n \times n$, to $\nabla f(x) = (A + A^T)x$ i $\nabla^2 f(x) = A^T + A$;

(e) jeśli $f(x) = \frac{1}{2}x^T Ax + b^T x + c$, gdzie A jest macierzą symetryczną, $b \in \mathbb{R}^n$ i $c \in \mathbb{R}$ (czyli f jest funkcją kwadratową), to $\nabla f(x) = Ax + b$ i $\nabla^2 f(x) = A$;

(f) jeśli $f(x) = \frac{1}{2}\|Ax - b\|^2$, gdzie A jest macierzą typu $m \times n$, zaś $b \in \mathbb{R}^m$, to $\nabla f(x) = A^T(Ax - b)$.

Uwaga 1.5.22 Z postaci gradientu dla funkcji kwadratowej (ćwiczenie 1.5.21(e)) wynika, że

$$g(x) - g(y) = G(x)(x - y).$$

Innymi słowy, dla funkcji kwadratowej hesjan odwzorowuje przyrost argumentu w przyrost gradientu.

Niech $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ i niech $\bar{x} \in \mathbb{R}^n$. Dalej niech $h : \mathbb{R} \longrightarrow \mathbb{R}^n$ będzie odwzorowaniem określonym następująco $h(t) = \bar{x} + ts$ dla $s \in \mathbb{R}^n$ i $t \in \mathbb{R}$. Zdefiniujmy funkcję $q : \mathbb{R} \longrightarrow \mathbb{R}$ wzorem

$$q(t) = (f \circ h)(t) = f(\bar{x} + ts).$$

Funkcja q określa więc zachowanie funkcji f na prostej o równaniu $x = \bar{x} + ts$. Jeśli funkcja f jest różniczkowalna w punkcie x, to na mocy twierdzenia 1.5.13 mamy

$$q'(t) = \nabla h(t) \cdot \nabla f(h(t)) = s^T \nabla f(x) = \nabla f(x)^T s.$$

Jeśli ponadto f jest dwukrotnie różniczkowalna w punkcie x, to przez powtórne zastosowanie tego twierdzenia otrzymamy

$$q''(t) = (q'(t))' = s^T \nabla(\nabla f(x)^T s) = s^T \nabla^2 f(x)s.$$

Wielkość $q'(t)$ nazywamy *nachyleniem* funkcji f w punkcie x wzdłuż prostej $h(t) = \bar{x} + ts$, zaś wielkość $q''(t)$ – *krzywizną* funkcji f w punkcie x wzdłuż prostej $h(t) = \bar{x} + ts$. Oznaczmy teraz $d = ts$. Wektor d możemy więc traktować jako przyrost argumentu: $d = x - \bar{x}$. Rozwijając

funkcję q we wzór Taylora z resztą Peana (1.11) dla $k = 1$ w otoczeniu punktu 0, otrzymamy przy założeniu różniczkowalności funkcji q na odcinku $[0, t]$,

$$q(t) = q(0) + tq'(0) + o(t).$$

Jeśli więc funkcja f jest różniczkowalna w otoczeniu punktu $\bar{x}$, to otrzymamy następujące rozwinięcie Taylora z resztą Peana

$$f(\bar{x} + ts) = f(\bar{x}) + ts^T \nabla f(\bar{x}) + o(t)$$

lub inaczej

$$f(\bar{x} + d) = f(\bar{x}) + d^T \nabla f(\bar{x}) + o(\|d\|).$$

Z kolei rozwijając funkcję q we wzór Taylora z resztą Peana (1.11) dla $k = 2$ w otoczeniu punktu 0, otrzymamy, przy założeniu dwukrotnej różniczkowalności funkcji q na przedziale $[0, t]$:

$$q(t) = q(0) + tq'(0) + \frac{1}{2}t^2 q''(0) + o(t^2).$$

Jeśli więc funkcja f jest dwukrotnie różniczkowalna na odcinku $[\bar{x}, \bar{x} + ts]$, to otrzymamy następujące rozwinięcie Taylora z resztą Peana

$$f(\bar{x} + ts) = f(\bar{x}) + ts^T \nabla f(\bar{x}) + \frac{1}{2}t^2 s^T \nabla^2 f(\bar{x}) s + o(t^2)$$

lub inaczej

$$f(\bar{x} + d) = f(\bar{x}) + d^T \nabla f(\bar{x}) + \frac{1}{2} d^T \nabla^2 f(\bar{x}) d + o(\|d\|^2).$$

Podobnie, korzystając ze wzoru (1.12), otrzymuje się rozwinięcia funkcji q we wzór Taylora z resztą Lagrange'a:

$$f(x) = f(\bar{x}) + d^T \nabla f(\bar{x} + \lambda d)$$

i

$$f(\bar{x} + d) = f(\bar{x}) + d^T \nabla f(\bar{x}) + \frac{1}{2} d^T \nabla^2 f(\bar{x} + \lambda d) d$$

dla pewnego $\lambda \in (0, 1)$.

Definicja 1.5.23 Funkcję liniową $\bar{f} : \mathbb{R}^n \longrightarrow \mathbb{R}$ określoną wzorem

$$\bar{f}(x) = f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x})$$

nazywamy *linearyzacją* (ang. *linearization*) funkcji f w punkcie $\bar{x}$.

Wykresem linearyzacji jest hiperpłaszczyzna styczna do wykresu funkcji f w punkcie $\bar{x}$. Twierdzenie 1.5.3 mówi, że $f(x)$ różni się od $\bar{f}(x)$ o $o(\|x - \bar{x}\|)$.

Definicja 1.5.24 Funkcję kwadratową $\breve{f}_{\bar{x}} : \mathbb{R}^n \longrightarrow \mathbb{R}$ określoną wzorem

$$\breve{f}(x) = \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + f(\bar{x})$$

nazywamy *przybliżeniem kwadratowym* (ang. *quadratic approximation*) funkcji f w punkcie $\bar{x}$.

Podobnie możemy wyprowadzić rozwinięcie Taylora dla odwzorowania ∇f. Otrzymamy wówczas

$$\nabla f(\bar{x}+d) = \nabla f(\bar{x}) + \nabla^2 f(\bar{x})d + o(\|d\|)$$

lub inaczej

$$\nabla f(x) - \nabla f(\bar{x}) = \nabla^2 f(\bar{x})(x-\bar{x}) + o(\|x-\bar{x}\|),$$

czyli hesjan odwzorowuje „w przybliżeniu" różnicę argumentów w różnicę gradientów. Jak zauważyliśmy wcześniej (patrz uwaga 1.5.22), dla funkcji kwadratowej f odwzorowanie to jest dokładne.

Definicja 1.5.25 Kierunek $s \in \mathbb{R}^n$ nazywa się *kierunkiem spadku (*ang. *descent direction*) funkcji f w punkcie $\bar{x} \in \mathbb{R}^n$ jeśli

$$\exists_{\tau>0}\forall_{t\in(0,\tau)} \quad f(\bar{x}+ts) < f(\bar{x}).$$

Uwaga 1.5.26 Z definicji pochodnej kierunkowej wynika, że jeśli $f'(\bar{x}, s) < 0$, to s jest kierunkiem spadku funkcji f w punkcie $\bar{x}$. Zatem dla funkcji różniczkowalnej f zachodzi implikacja

$$s^T\nabla f(\bar{x}) < 0 \Longrightarrow s \text{ jest kierunkiem spadku funkcji } f \text{ punkcie } \bar{x}.$$

Odwrotna implikacja nie jest prawdziwa (wystarczy wziąć $f(x) = x^3$, $\bar{x} = 0$ i $s = -1$).

Definicja 1.5.27 Niech dla funkcji $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ i dla punktu $\bar{x} \in \mathbb{R}^n$ istnieją pochodne kierunkowe $f'(\bar{x}, s)$ dla wszystkich $s \in \mathbb{R}^n$. Kierunek $s \in \operatorname{Argmin}_{\|s\|=1} f'(\bar{x}, s)$ nazywa się *kierunkiem najszybszego spadku* zaś kierunek $s \in \operatorname{Argmax}_{\|s\|=1} f'(\bar{x}, s)$ – *kierunkiem najszybszego wzrostu* funkcji f w punkcie $\bar{x} \in \mathbb{R}^n$.

Definicja 1.5.28 Niech $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ będzie funkcją różniczkowalną w sensie Gâteaux. Punkt $\bar{x}$, dla którego $\nabla f(\bar{x}) = 0$ nazywa się *punktem stacjonarnym* (ang. *stationary point*) funkcji f.

Ćwiczenie 1.5.29 Niech $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ będzie funkcją różniczkowalną w sensie Gâteaux w punkcie $\bar{x}$ i $\nabla f(\bar{x}) \neq 0$. Kierunek najszybszego spadku funkcji f w punkcie $\bar{x}$ jest rozwiązaniem zadania

$$\begin{aligned} \text{minimalizować} \quad & s^T\nabla f(\bar{x}) \\ \text{względem} \quad & s \in \mathbb{R}^n \\ \text{przy ograniczeniu} \quad & \|s\| = 1. \end{aligned}$$

Korzystając z nierówności Cauchy'ego–Schwarza pokazać, że kierunkiem najszybszego spadku funkcji f w punkcie $\bar{x} \in \mathbb{R}^n$ jest wektor

$$s := -\frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|}$$

oraz kierunkiem najszybszego wzrostu funkcji f w punkcie $\bar{x} \in \mathbb{R}^n$ jest wektor

$$s := \frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|}.$$

Definicja 1.5.30 Niech $X \subseteq \mathbb{R}^n$ i niech $\bar{x} \in X$. Wektor $s \in \mathbb{R}^n$ nazywamy *kierunkiem stycznym* (ang. *tangent direction*) do zbioru X w punkcie $\bar{x}$ jeśli istnieje ciąg $x_k \in X$, $x_k \to \bar{x}$ i ciąg $t_k \in \mathbb{R}$, $t_k \downarrow 0$ takie, że

$$s = \lim_k \frac{x_k - \bar{x}}{t_k}.$$

W zadaniach minimalizacji z ograniczeniami, w przypadku, gdy X jest zbiorem rozwiązań dopuszczalnych, kierunek styczny nazywany jest również *kierunkiem dopuszczalnym* (ang. *feasible direction*). Zbiór wszystkich kierunków stycznych do zbioru X w punkcie $\bar{x}$ oznaczać będziemy symbolem $T_X(\bar{x})$.

Wyznaczenie zbioru kierunków stycznych z definicji jest zazwyczaj trudne. W dalszej części zobaczymy jak go wyznaczyć dla zbiorów X zadanych układem nierówności.

Czytelnikowi pozostawiamy dowód faktu, że zbiór kierunków stycznych jest domknięty.

Uwaga 1.5.31 Jeśli s jest kierunkiem stycznym (do zbioru X w punkcie $\bar{x}$), to jest nim również αs dla dowolnego $\alpha \geq 0$ (innymi słowy, zbiór kierunków stycznych jest stożkiem). Z tego względu można ograniczyć się do kierunków stycznych o ustalonej normie, na przykład równej 1, przyjmując w powyższej definicji $t_k = \|x_k - \bar{x}\|$. Nietrudno pokazać, że powstała w ten sposób definicja kierunku stycznego jest w pewnym sensie równoważna definicji 1.5.30.

Ćwiczenie 1.5.32 Niech $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ będzie funkcją różniczkowalną i niech $\bar{x} \in \mathbb{R}^n$ będzie punktem takim, że $\nabla f(\bar{x}) \neq 0$. Pokazać, że

$$T_{\{x \in \mathbb{R}^n : f(x) = f(\bar{x})\}}(\bar{x}) = \{s \in \mathbb{R}^n : s^T \nabla f(\bar{x}) = 0\}.$$

Równość powyższą można geometrycznie zinterpretować w ten sposób, że kierunkami stycznymi do poziomicy funkcji f w punkcie $\bar{x}$ są wszystkie wektory ortogonalne do gradientu $\nabla f(\bar{x})$. Stanowią one podprzestrzeń liniową – w tym przypadku hiperpłaszczyznę. Mówimy również, że gradient $\nabla f(\bar{x})$ jest ortogonalny do poziomicy $\{x \in \mathbb{R}^n : f(x) = f(\bar{x})\}$.

1.6 Zbiory wypukłe

1.6.1 Podstawowe pojęcia i fakty dotyczące zbiorów wypukłych

Definicja 1.6.1 Mówimy, że zbiór $K \subseteq \mathbb{R}^n$ jest *wypukły* (ang. *convex*), jeśli dla dowolnych $x, y \in K$ i dla każdego $\lambda \in [0,1]$ zachodzi $(1-\lambda)x+\lambda y \in K$. *Otoczką wypukłą* (ang. *convex hull*) zbioru $S \subseteq \mathbb{R}^n$ nazywamy najmniejszy zbiór wypukły zawierający S. Oznaczamy ją symbolem $\operatorname{conv} S$. *Kombinacją wypukłą* (ang. *convex combination*) elementów $x_1, ..., x_m \in \mathbb{R}^n$ nazywamy element $x \in \mathbb{R}^n$ postaci

$$x = \sum_{i=1}^{m} \lambda_i x_i,$$

gdzie $\lambda_i \geq 0$, $i = 1, ..., m$, $\sum_{i=1}^{m} \lambda_i = 1$, $m \geq 1$.

Ćwiczenie 1.6.2 Pokazać, że przekrój dowolnej rodziny zbiorów wypukłych jest zbiorem wypukłym.

Ćwiczenie 1.6.3 Pokazać, że otoczką wypukła zbioru $S \subseteq \mathbb{R}^n$ jest przekrój wszystkich zbiorów wypukłych zawierających S.

Czytelnikowi pozostawiamy dowód następującego lematu.

Lemat 1.6.4 *Podzbiór $C \subseteq \mathbb{R}^n$ jest wypukły wtedy i tylko wtedy, gdy do C należy dowolna kombinacja wypukła jego elementów.*

W istocie wystarczy pokazać konieczność warunku i w tym celu posłużyć się indukcją matematyczną.

Ćwiczenie 1.6.5 Pokazać, że poniższe zbiory są wypukłe:

(a) dowolna podprzestrzeń afiniczna, w szczególności dowolna *hiperpłaszczyzna* $\{x \in \mathbb{R}^n : a^T x = b\}$,

(b) dowolna *półprzestrzeń* $\{x \in \mathbb{R}^n : a^T x \leq b\}$,

(c) przekrój dowolnej rodziny podprzestrzeni afinicznych, w szczególności

$$\bigcap_{i \in I} \{x \in \mathbb{R}^n : a_i^T x = b_i\},$$

(d) przekrój dowolnej rodziny półprzestrzeni

$$\bigcap_{i \in I} \{x \in \mathbb{R}^n : a_i^T x \leq b_i\},$$

(e) *sympleks standardowy* (ang. *standard simplex*)

$$\Delta_m := \{w = (\lambda_1, ..., \lambda_m) \in \mathbb{R}^m : \lambda_i \geq 0,\ i = 1, ..., m,\ \sum_{i=1}^{m} \lambda_i = 1\},$$

(f) zbiór rozwiązań optymalnych zadania minimalizacji wypukłej,

(g) kula $B(\bar{x}, r) := \{x \in \mathbb{R}^n : \|x - \bar{x}\| \leq r\}$, gdzie $\|\cdot\|$ oznacza dowolną normę w $\mathbb{R}^n$, $\bar{x} \in \mathbb{R}^n$, $r \geq 0$,

(h) *elipsoida* (ang. *ellipsoid*) $J(D, \bar{x}, \rho) := \{x \in \mathbb{R}^n : (x - \bar{x})^\top D(x - \bar{x}) \leq \rho\}$, gdzie D jest macierzą określoną dodatnio, $\bar{x} \in \mathbb{R}^n$, $\rho > 0$,

(j) *stożek wypukły* (ang. *convex cone*), czyli podzbiór $C \subseteq \mathbb{R}^n$ spełniający warunki: (i) $x \in C$ i $\alpha > 0 \Longrightarrow \alpha x \in C$, (ii) $x, y \in C \Longrightarrow x + y \in C$.

Definicja 1.6.6 Przekrój skończenie wielu półprzestrzeni nazywamy *zbiorem wielościennym*. Ograniczony zbiór wielościenny nazywa się *wielościanem(ang. polyhedron)*.

Lemat 1.6.7 *Otoczka wypukła zbioru $S \subseteq \mathbb{R}^n$ jest postaci*

$$\operatorname{conv} S = \left\{ \sum_{i=1}^{m} \lambda_i x_i : x_i \in S,\ w = (\lambda_1, ..., \lambda_m) \in \Delta_m,\ m \geq 1 \right\}. \tag{1.13}$$

Dowód. „$\supseteq$" Na mocy definicji 1.6.1 $\operatorname{conv} S$ jest zbiorem wypukłym. Skoro $\operatorname{conv} S \supseteq S$, więc z lematu 1.6.4 wynika, że $\operatorname{conv} S$ zawiera również wszystkie kombinacje wypukłe elementów zbioru S.

„$\subseteq$" Nietrudno pokazać, że zbiór po prawej stronie równości (1.13) jest wypukły. Zawiera on S, więc zgodnie z definicją 1.6.1 zawiera on również $\operatorname{conv} S$. ■

W przestrzeni $\mathbb{R}^n$ liczbę elementów wchodzących do kombinacji wypukłych, o których mowa w równości (1.13) można ograniczyć do $n + 1$. Zachodzi bowiem następujące twierdzenie.

Twierdzenie 1.6.8 (Carathéodory) *Otoczka wypukła zbioru $S \subseteq \mathbb{R}^n$ składa się ze wszystkich kombinacji wypukłych co najwyżej $n + 1$ elementów zbioru S, czyli*

$$\operatorname{conv} S = \left\{ \sum_{i=1}^{m} \lambda_i x_i : x_i \in S,\ w = (\lambda_1, ..., \lambda_m) \in \Delta_m,\ m \leq n + 1 \right\}.$$

Dowód. Inkluzja $\supseteq$ wynika z lematu 1.6.7. Niech $x \in \operatorname{conv} S$. Zgodnie z równością (1.13)

$$x = \sum_{i=1}^{m} \lambda_i x_i \tag{1.14}$$

dla pewnych $x_1, ..., x_m \in S$, $w \in \Delta_m$ i $m \geq 1$. Spośród wszystkich przedstawień wektora x w postaci (1.14) wybierzmy to, dla którego liczba m jest najmniejsza. Jest jasne, że $\lambda_i > 0$ dla $i = 1, 2, ..., m$. Pozostaje pokazać, że $m \leq n + 1$. Przypuśćmy, że jest przeciwnie. Niech $x_i' = (x_i, 1) \in \mathbb{R}^n \times \mathbb{R}$. Ponieważ $m > n + 1$, więc wektory x_i', $i = 1, ..., m$, są liniowo zależne. Oznacza to, że istnieją takie liczby $\alpha_1, ..., \alpha_m$ nie wszystkie równe zeru, że $\sum_{i=1}^m \alpha_i x_i' = 0$, czyli $\sum_{i=1}^m \alpha_i x_i = 0$ i $\sum_{i=1}^m \alpha_i = 0$. Tak więc, wśród α_i, $i = 1, ..., m$, istnieją liczby dodatnie. Niech

$$\varepsilon_0 := \frac{\lambda_{i_0}}{\alpha_{i_0}} = \min \left\{ \frac{\lambda_i}{\alpha_i} : \alpha_i > 0,\ i = 1, ..., m \right\} \tag{1.15}$$

i niech

$$\bar{\lambda}_i = \lambda_i - \varepsilon_0 \alpha_i,\ i = 1, ..., m.$$

Z równości (1.15) wynika, że $\bar{\lambda}_i \geq 0$ dla $i = 1, ..., m$ i że $\bar{\lambda}_{i_0} = 0$. Zatem

$$\sum_{i=1}^{m} \bar{\lambda}_i x_i = \sum_{i=1}^{m} \lambda_i x_i - \varepsilon_0 \sum_{i=1}^{m} \alpha_i x_i = x$$

i

$$\sum_{i=1}^{m} \bar{\lambda}_i = \sum_{i=1}^{m} \lambda_i - \varepsilon_0 \sum_{i=1}^{m} \alpha_i = \sum_{i=1}^{m} \lambda_i = 1.$$

Ponieważ $\bar{\lambda}_{i_0} = 0$, więc widzimy, że x można przedstawić jako kombinację wypukłą $m-1$ elementów. Otrzymaliśmy sprzeczność z założeniem, że m jest najmniejsza liczbą dla której x daje się przedstawić w postaci (1.14). ■

Ćwiczenie 1.6.9 Wyznaczyć otoczki wypukłe następujących podzbiorów przestrzeni $\mathbb{R}^n$:

(a) $\{e_i : i = 1, 2, ..., n\}$, gdzie e_i jest wersorem jednostkowym w $\mathbb{R}^n$.

(b) $S(\bar{x}, r) = \{x \in \mathbb{R}^n : \|x - \bar{x}\| = r\}$, gdzie $\|\cdot\|$ oznacza dowolną normę w $\mathbb{R}^n$.

1.6.2 Punkty ekstremalne

Definicja 1.6.10 Punkt x należący do zbioru wypukłego $K \subseteq \mathbb{R}^n$ nazywa się *punktem ekstremalnym* tego zbioru, jeśli nie jest środkiem odcinka łączącego dwa różne punkty zbioru K, tzn.

$$x = \frac{1}{2}(x' + x'') \text{ i } x', x'' \in K \Rightarrow x' = x''.$$

Zbiór punktów ekstremalnych zbioru K oznaczamy symbolem $\operatorname{ext} K$. W przypadku, gdy K jest zbiorem wielościennym, jego punkty ekstremalne nazywają się *wierzchołkami*.

Ćwiczenie 1.6.11 Wyznaczyć zbiory punktów ekstremalnych dla następujących zbiorów wypukłych.

(a) $K = [a, b] \subseteq \mathbb{R}$, przy czym $a \leq b$,

(b) $K = \mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x \geq 0\}$,

(c) $K = \Delta_n \subseteq \mathbb{R}^n$,

(d) $K = \{x \in \mathbb{R}^n : Ax = b\}$, gdzie A jest macierzą typu $m \times n$, $b \in \mathbb{R}^m$.

1.6.3 Rzut metryczny

Definicja 1.6.12 Niech $C \subseteq \mathbb{R}^n$ i niech $x \in \mathbb{R}^n$. Punkt $y \in C$ nazywamy *rzutem metrycznym* (ang. *metric projection*) punktu x na zbiór C, jeśli

$$\|x - y\| \leq \|x - z\| \text{ dla dowolnego } z \in C,$$

i oznaczamy go symbolem $P_C(x)$.

Wprawdzie rzut metryczny można zdefiniować dla dowolnej normy, jednak będą nas interesować własności tego rzutu dla normy indukowanej przez iloczyn skalarny. W dalszym ciągu tego ustępu zakładamy więc, że $\|x\| = \sqrt{\langle x, x \rangle}$ dla pewnego iloczynu skalarnego $\langle \cdot, \cdot \rangle$, chyba że będzie powiedziane inaczej. Z definicji 1.6.12 nie wynika istnienie rzutu metrycznego, a jeśli nawet on istnieje, nie ma gwarancji jego jednoznaczności. Zachodzi natomiast poniższe twierdzenie.

Twierdzenie 1.6.13 *Niech $C \subseteq \mathbb{R}^n$ będzie zbiorem niepustym, domkniętym i wypukłym. Wówczas dla dowolnego $x \in \mathbb{R}^n$ istnieje dokładnie jeden jego rzut metryczny na C.*

Dowód. Twierdzenie pokażemy najpierw dla $x = 0$. Niech $d = \inf\{\|y\| : y \in C\}$ i niech ciąg $\{y_k\}_{k=1}^{\infty} \subseteq C$ będzie wybrany tak, by $\|y_k\| \to d$. Dowód rozbijemy na trzy części.

(a) Pokażemy, że $\{y_k\}_{k=1}^{\infty}$ jest ciągiem Cauchy'ego. Niech $\varepsilon > 0$ i niech $k_0 \geq 1$ będzie takie, że $\|y_k\|^2 \leq d^2 + \varepsilon/4$ dla $k \geq k_0$. Niech $k, l \geq k_0$. Oczywiście $\frac{1}{2}y_k + \frac{1}{2}y_l \in C$ ponieważ C jest wypukły. Stąd $\frac{1}{2}\|y_k + y_l\| \geq d$. Korzystając z tożsamości równoległoboku otrzymujemy w konsekwencji:

$$\|y_k - y_l\|^2 = 2\|y_k\|^2 + 2\|y_l\|^2 - \|y_k + y_l\|^2$$

$$\leq 2(d^2 + \varepsilon/4) + 2(d^2 + \varepsilon/4) - 4d^2 = \varepsilon,$$

tzn. $\{y_k\}_{k=1}^{\infty}$ jest ciągiem Cauchy'ego.

(b) Z (a) wynika, że y_k zbiega do pewnego $y \in \mathbb{R}^n$, gdyż $\mathbb{R}^n$ jest przestrzenią zupełną. Ponadto $y \in C$, ponieważ C jest domknięty. Stąd i z ciągłości normy wynika, że $\|y\| = d$. Oznacza to, że $y = P_C(0)$.

(c) Pokażemy teraz, że rzut metryczny określony jest jednoznacznie. Niech $y' \in C$ i niech $\|y'\| = d$. Z wypukłości C otrzymujemy $\frac{1}{2}y + \frac{1}{2}y' \in C$. Ponadto

$$d \leq \|\frac{1}{2}y + \frac{1}{2}y'\| \leq \frac{1}{2}\|y\| + \frac{1}{2}\|y'\| = d,$$

a więc $\|y + y'\| = 2d$. Korzystając powtórnie z tożsamości równoległoboku mamy:

$$\|y - y'\|^2 = 2\|y\|^2 + 2\|y'\|^2 - \|y + y'\|^2 = 2d^2 + 2d^2 - 4d^2 = 0,$$

czyli $y = y'$.

Niech teraz $x \in \mathbb{R}^n$ będzie dowolny. Z definicji rzutu metrycznego wynika, że $x + P_{C-x}(0)$ jest rzutem metrycznym punktu x na C, ponadto nietrudno zauważyć, że jest on określony jednoznacznie. Zatem twierdzenie jest prawdziwe dla dowolnego $x \in \mathbb{R}^n$. ■

Uwaga 1.6.14 Istnienie rzutu metrycznego na zbiór niepusty, domknięty i wypukły $C \subseteq \mathbb{R}^n$ można pokazać prościej korzystając z ciągłości normy i z twierdzenia Weierstrassa. Natomiast przeprowadzony powyżej dowód wskazuje, że twierdzenie 1.6.13 jest prawdziwe również dla dowolnej przestrzeni Hilberta.

Poniższe twierdzenie podaje charakteryzację rzutu metrycznego i jest przydatne przy wyznaczaniu rzutu metrycznego dla konkretnych zbiorów domkniętych wypukłych.

Twierdzenie 1.6.15 *Niech $x \in \mathbb{R}^n$, $C \subseteq \mathbb{R}^n$ będzie zbiorem niepustym, domkniętym i wypukłym i niech $y \in C$. Wówczas następujące warunki są równoważne*

(i) $y = P_C(x)$,

(ii) $\langle x - y, z - y\rangle \leq 0$ *dla dowolnego* $z \in C$.

Dowód. (i)⇒(ii). Niech $y = P_C(x)$ i niech $z \in C$. Ponadto, niech

$$z_\lambda = y + \lambda(z - y)$$

dla $\lambda \in (0, 1)$. Oczywiście $z_\lambda \in C$, ponieważ C jest wypukły. Z (i) i z własności iloczynu skalarnego mamy więc

$$\|x - y\|^2 \leq \|x - z_\lambda\|^2 = \|x - y - \lambda(z - y)\|^2$$

$$= \|x - y\|^2 - 2\lambda\langle x - y, z - y\rangle + \lambda^2\|z - y\|^2.$$

Skoro $\lambda > 0$, więc

$$\langle x - y, z - y\rangle \leq \frac{\lambda}{2}\|z - y\|^2,$$

a ponieważ λ jest dowolną liczbą z przedziału $(0, 1)$, więc musi zachodzić (ii).

(ii)⇒(i). Z własności iloczynu skalarnego oraz z (ii) mamy dla dowolnego $z \in C$

$$\|z - x\|^2 = \|z - y\|^2 + \|y - x\|^2 + 2\langle z - y, y - x\rangle \geq \|y - x\|^2,$$

co na mocy definicji rzutu metrycznego daje (i). ■

Ćwiczenie 1.6.16 Korzystając z charakteryzacji rzutu metrycznego sprawdzić słuszność wzorów na rzuty metryczne punktu $x \in \mathbb{R}$ dla podanych zbiorów domkniętych wypukłych $C \subseteq \mathbb{R}^n$:

(a) Jeśli C jest hiperpłaszczyzną, tzn. $C = H(a, \beta) := \{z \in \mathbb{R}^n : \langle a, z\rangle = \beta\}$, gdzie $a \in \mathbb{R}^n$, $\beta \in \mathbb{R}$, to zachodzi wzór

$$P_C(x) = x - \frac{\langle a, x\rangle - \beta}{\|a\|^2}a.$$

(b) Jeśli C jest półprzestrzenią, tzn. $C = H_-(a, \beta) := \{z \in \mathbb{R}^n : \langle a, z\rangle \leq \beta\}$, gdzie $a \in \mathbb{R}^n$ i $\beta \in \mathbb{R}$, to zachodzi wzór

$$P_C(x) = x - \frac{(\langle a, x\rangle - \beta)_+}{\|a\|^2}a.$$

(c) Niech $C \subseteq \mathbb{R}^n$ będzie zbiorem rozwiązań układu równań liniowych, tzn. $C = \{z \in \mathbb{R}^n : Az = b\}$, gdzie A jest macierzą typu $m \times n$ pełnego rzędu wierszowego i $b \in \mathbb{R}^m$. Wówczas zachodzi wzór

$$P_C(x) = x - A^\top(AA^\top)^{-1}(Ax - b).$$

(d) Jeśli $C \subseteq \mathbb{R}^n$ jest kulą, tzn. $C = B(\bar{x}, r) := \{z \in \mathbb{R}^n : \|z - \bar{x}\| \leq r\}$, gdzie $\|\cdot\|$ jest normą euklidesową, $\bar{x} \in \mathbb{R}^n$ i $r > 0$, to zachodzi wzór

$$P_C(x) = \begin{cases} x & \text{gdy } \|x - \bar{x}\| \leq r \\ \bar{x} + \frac{r}{\|x-\bar{x}\|}(x - \bar{x}) & \text{gdy } \|x - \bar{x}\| > r. \end{cases}$$

Czy wzór ten jest słuszny dla norm innych niż euklidesowa? Podać odpowiednie przykłady.

1.6.4 Twierdzenia o oddzielaniu i ich konsekwencje

Konsekwencją charakteryzacji rzutu metrycznego podanej w twierdzeniu 1.6.15 są tzw. twierdzenia o oddzielaniu.

Twierdzenie 1.6.17 (o ostrym oddzielaniu) *Niech $C \subseteq \mathbb{R}^n$ będzie zbiorem niepustym, domkniętym i wypukłym i niech $x \notin C$. Wówczas istnieje wektor $s \in \mathbb{R}^n$, taki że*

$$\langle s, x\rangle > \sup\{\langle s, y\rangle : y \in C\}.$$

Dowód. Niech $s = x - P_C(x)$. Wówczas z twierdzenia 1.6.15 wynika prosto, że dla każdego $y \in C$

$$\langle x - y, s\rangle \geq \|s\|^2,$$

czyli

$$\langle s, x\rangle \geq \langle s, y\rangle + \|s\|^2.$$

Zauważmy, że $s \neq 0$, bo C jest domknięty i $x \notin C$. Zatem

$$\langle s, x\rangle \geq \sup\{\langle s, y\rangle : y \in C\} + \|s\|^2 > \sup\{\langle s, y\rangle : y \in C\},$$

co kończy dowód. ■

Wniosek 1.6.18 *Niech $A, B \subseteq \mathbb{R}^n$ będą zbiorami wypukłymi i domkniętymi, przy czym jeden z nich jest zbiorem zwartym. Jeśli $A \cap B = \emptyset$, to istnieje wektor $s \in \mathbb{R}^n$, taki że*

$$\inf\{\langle s, u\rangle : u \in A\} > \sup\{\langle s, v\rangle : v \in B\}.$$

Dowód. Niech $C = B - A := \{z \in \mathbb{R}^n : z = v - u,\ v \in B,\ u \in A\}$. Nietrudno zauważyć,że C jest zbiorem wypukłym. Ponadto C jest zbiorem domkniętym, ponieważ A i B są domknięte, a jeden z nich zwarty. Istotnie, niech $\{z_k\}_{k=1}^{\infty} \subseteq C$ i niech $z_k \to z$. Wówczas $z_k = v_k - u_k$, gdzie $u_k \in A$ i $v_k \in B$. Przypuśćmy, że zbiór A jest zwarty. Wówczas istnieje podciąg zbieżny $\{u_{n_k}\}_{k=1}^{\infty}$ ciągu $\{u_n\}_{k=1}^{\infty}$. Niech $u = \lim_{k\to\infty} u_{n_k}$. Wówczas

$$v_{n_k} = z_{n_k} + u_{n_k} \to v = z + u.$$

Oczywiście $u \in A$ i $v \in B$, ponieważ A i B są zbiorami domkniętymi. Mamy więc $z = v - u \in B - A = C$. Oznacza to, że zbiór C jest domknięty. Ponieważ, $A \cap B = \emptyset$, więc $0 \notin C$. Na mocy twierdzenia o ostrym oddzielaniu istnieje wektor $s \in \mathbb{R}^n$, taki że

$$\begin{aligned} 0 &= \langle s, 0\rangle > \sup\{\langle s, y\rangle : y \in C\} \\ &= \sup\{\langle s, v - u\rangle : u \in A,\ v \in B\} \\ &= \sup\{\langle s, v\rangle : v \in B\} - \inf\{\langle s, u\rangle : u \in A\}, \end{aligned}$$

co kończy dowód. ■

Twierdzenie 1.6.19 (o słabym oddzielaniu) *Niech $C \subseteq \mathbb{R}^n$ będzie zbiorem niepustym i wypukłym i niech $x \notin C$. Wówczas istnieje $s \in \mathbb{R}^n$ takie, że*

$$\langle s, x\rangle \geq \sup\{\langle s, y\rangle : y \in C\}.$$

Dowód. Jeśli $x \notin \operatorname{cl} C$, to teza wynika z twierdzenia o ostrym oddzielaniu. Niech więc $x \in \operatorname{cl} C \setminus C$. Ponieważ ten ostatni zbiór jest zawarty w brzegu zbioru C, więc istnieje ciąg $x_k \to x$, $x_k \notin \operatorname{cl} C$. Z twierdzenia o ostrym oddzielaniu wynika, że istnieje ciąg $s_k \neq 0$, taki że dla każdego $y \in C$

$$\langle s_k, x_k\rangle > \langle s_k, y\rangle$$

lub inaczej

$$\langle \frac{s_k}{\|s_k\|}, x_k\rangle > \langle \frac{s_k}{\|s_k\|}, y\rangle.$$

Ciąg $\{s_k/\|s_k\|\}_{k=1}^{\infty}$ jest unormowany, posiada on więc podciąg zbieżny do pewnego wektora unormowanego, powiedzmy do s. Przechodząc do granicy i korzystając z ciągłości iloczynu skalarnego otrzymamy

$$\langle s, x\rangle \geq \langle s, y\rangle$$

dla każdego $y \in C$, co jest równoważne tezie twierdzenia. ■

Z dowodu twierdzenia 1.6.19 wynika, że twierdzenie to zachodzi również dla $x \in \operatorname{bd} C$.

Wniosek 1.6.20 *Niech $A, B \subseteq \mathbb{R}^n$ będą zbiorami wypukłymi i domkniętymi. Jeśli $A \cap B = \emptyset$, to istnieje wektor $s \in \mathbb{R}^n$, taki że*

$$\inf\{\langle s, u\rangle : u \in A\} \geq \sup\{\langle s, v\rangle : v \in B\}.$$

Dowód powyższego wniosku przeprowadza się podobnie do dowodu wniosku 1.6.18, przy czym należy w nim skorzystać z twierdzenia o słabym oddzielaniu.

Twierdzenie 1.6.21 (Minkowski) *Wielościan K jest otoczką wypukłą zbioru swoich punktów ekstremalnych:*

$$K = \operatorname{conv} \operatorname{ext} K.$$

Dowód indukcyjny twierdzenia Minkowskiego względem wymiaru przestrzeni oparty na twierdzeniu o oddzielaniu można znaleźć w podręcznikach do analizy wypukłej, np. w podręczniku J. B. Hiriarta-Urruty'ego i C. Lemaréchala [HL93, twierdzenie 2.3.4]. Podobne twierdzenie można sformułować dla zwartych i wypukłych podzbiorów dowolnej lokalnie wypukłej przestrzeni liniowo-topologicznej. Nosi ono nazwę twierdzenia Kreina–Milmana.

Uwaga 1.6.22 Prawdziwe jest również twierdzenie w pewnym sensie odwrotne do twierdzenia Minkowskiego: otoczka wypukła zbioru skończonego jest wielościanem. Zatem jest ona zbiorem domkniętym, wypukłym i ograniczonym.

1.6.5 Stożki wypukłe

Definicja 1.6.23 Zbiór $C \subseteq \mathbb{R}^n$ nazywa się *stożkiem wypukłym*, jeśli

(i) $x, y \in C \Rightarrow x + y \in C$,

(ii) $x \in C, \alpha \geq 0 \Rightarrow \alpha x \in C$.

Stożek wypukły jest zbiorem wypukłym (patrz ćwiczenie 1.6.5).

Ćwiczenie 1.6.24 Pokazać, że następujące podzbiory przestrzeni $\mathbb{R}^n$ są stożkami wypukłymi:

a) dowolna podprzestrzeń liniowa $V \subseteq \mathbb{R}^n$, w szczególności $\ker A := \{x \in \mathbb{R}^n : Ax = 0\}$ oraz $\operatorname{im} A^T := \{x \in \mathbb{R}^n : x = A^T y, y \in \mathbb{R}^m\}$ jako podprzestrzenie liniowe są zbiorami wypukłymi , gdzie A jest macierzą typu $m \times n$,

b) zbiór $\{x \in \mathbb{R}^n : Ax \leq 0\}$, gdzie A jest macierzą typu $m \times n$,

c) zbiór $C^* := \{x \in \mathbb{R}^n : \langle x, y\rangle \leq 0 \text{ dla każdego } y \in C\}$, gdzie $C \subseteq \mathbb{R}^n$ jest zbiorem niepustym,

d) *stożek generowanym* przez podzbiór $S \subseteq \mathbb{R}^n$ (zwany również *otoczką stożkową* zbioru S), czyli zbiór

$$\operatorname{cone} S := \{x \in \mathbb{R}^n : x = \sum_{i=1}^{p} \alpha_i s_i, s_i \in S, \alpha_i \geq 0, i = 1, ..., p, p \geq 1\}.$$

Które z tych stożków są domknięte?

Ćwiczenie 1.6.25 Niech $S \subseteq \mathbb{R}^n$. Pokazać, że

$$\operatorname{cone} \operatorname{conv} S = \operatorname{cone} S.$$

Czy $\operatorname{cone} S$ jest domknięty, jeśli S jest domknięty. Czy jest on domknięty, jeśli S jest zwarty?

Postępując podobnie jak w dowodzie twierdzenia Carathéodory'ego (tw. 1.6.8) można pokazać następujące twierdzenie.

Twierdzenie 1.6.26 Niech $S \subseteq \mathbb{R}^n$. Wówczas

$$\operatorname{cone} S = \{x \in \mathbb{R}^n : x = \sum_{i=1}^{p} \alpha_i s_i, s_i \in S, \alpha_i \geq 0, i = 1, ..., p, p \leq n\}.$$

Twierdzenie 1.6.27 *Otoczka stożkowa zbioru skończonego jest zbiorem domkniętym.*

Dowód. Niech $S = \{s_1, s_2, ..., s_m\} \subseteq \mathbb{R}^n$. Niech $\mathcal{A}$ będzie rodziną wszystkich liniowo niezależnych podzbiorów zbioru S. Z twierdzenia 1.6.26 wynika prosto, że

$$\operatorname{cone} S = \bigcup_{\{s_{i_1}, s_{i_2}, ..., s_{i_p}\} \in \mathcal{A}} \operatorname{cone}\{s_{i_1}, s_{i_2}, ..., s_{i_p}\}$$

Ponieważ $\mathcal{A}$ jest skończona, więc wystarczy pokazać, że dla dowolnego układu liniowo niezależnego $\{a_1, a_2, ..., a_p\} \subseteq \mathbb{R}^n$ zbiór $C := \operatorname{cone}\{a_1, a_2, ..., a_p\}$ jest domknięty. Niech więc ciąg $\{x_k\}_{k=1}^{\infty} \subseteq C$ będzie zbieżny do pewnego $x \in \mathbb{R}^n$. Jest jasne, że $x \in \operatorname{lin}\{a_1, a_2, ..., a_p\}$, ponieważ podprzestrzeń $\operatorname{lin}\{a_1, a_2, ..., a_p\}$ jest domknięta. Mamy więc $x_k = \sum_{i=1}^{p} \alpha_i^k a_i$, gdzie $\alpha_i^k \geq 0$, $i = 1, 2, ..., p$, $k \geq 1$ oraz $x = \sum_{i=1}^{p} \alpha_i a_i$, gdzie $\alpha_i \in \mathbb{R}$, $i = 1, 2, ..., p$, przy czym oba przedstawienia są jednoznaczne. Pozostawiamy czytelnikowi pokazanie, że $\lim_k \alpha_i^k = \alpha_i$, $i = 1, 2, ..., p$. Stąd już prosto wynika, że $\alpha_i \geq 0$, $i = 1, 2, ..., p$, czyli $x \in C$. ■

Twierdzenie 1.6.28 *Niech $C \subseteq \mathbb{R}^n$ będzie stożkiem domkniętym i niech $A \subseteq \mathbb{R}^n$ będzie zbiorem zwartym i wypukłym. Jeśli $A \cap C = \emptyset$, to istnieje wektor $w \in \mathbb{R}^n$, taki że $\sup_{v \in C}\langle w, v\rangle \leq 0$ i $\inf_{u \in A}\langle w, u\rangle > 0$.*

Dowód. W przypadku $C = \{0\}$ wystarczy wziąć $w = P_A 0$. Niech więc $C \neq \{0\}$. Z wniosku 1.6.18 wynika, że istnieje wektor w, taki że $\alpha = \inf_{u \in A}\langle w, u\rangle > \sup_{v \in C}\langle w, v\rangle = \gamma$. Pozostaje więc pokazać, że $\gamma = 0$. Przypuśćmy, że $\gamma > 0$. Niech wektor $z \in C$ będzie taki, że $\langle w, z\rangle > 0$. Wówczas $\beta z \in C$ dla każdego $\beta \geq 0$, gdyż C jest stożkiem. Otrzymamy wówczas

$$+\infty = \lim_{\beta \to +\infty} \langle w, \beta z\rangle \leq \sup_{v \in C}\langle w, v\rangle = \gamma,$$

czyli $\gamma = +\infty$. Równość ta stoi w sprzeczności z nierównością $\alpha > \gamma$. Przypuśćmy więc, że $\gamma < 0$. Otrzymamy wówczas

$$0 > \gamma = \sup_{v \in C}\langle w, v\rangle \geq \lim_{\beta \to 0^+} \langle w, \beta z\rangle = 0$$

dla pewnego $z \in C$. Otrzymana sprzeczność dowodzi, że $\gamma = 0$. ■

Definicja 1.6.29 Niech $C \subseteq \mathbb{R}^n$ będzie stożkiem. Podzbiór

$$C^* := \{y \in \mathbb{R}^n : \langle x, y\rangle \leq 0 \text{ dla dowolnego } x \in C\}$$

nazywa się *stożkiem sprzężonym* z C.

Poniższe twierdzenie odgrywa ważną rolę w optymalizacji. Wynika z niego w szczególności lemat Farkasa, którego konsekwencją są warunki konieczne minimalizacji różniczkowalnej z ograniczeniami (tzw. twierdzenie Kuhna–Tuckera).

Twierdzenie 1.6.30 *Niech $a_1, ..., a_m \in \mathbb{R}^n$. Stożki*

$$C := \{x \in \mathbb{R}^n : \langle a_i, x\rangle \leq 0, i = 1, ..., m\}$$

oraz

$$K := \{y \in \mathbb{R}^n : y = \sum_{i=1}^{m} w_i a_i,\ w_i \geq 0, i = 1, ..., m\}$$

są wzajemnie sprzężone, tzn. $C^* = K$ oraz $K^* = C$.

Dowód. (a) Pokażemy, że $C^* = K$. Inkluzja $K \subseteq C^*$ jest oczywista. Niech bowiem $y \in K$, czyli $y = \sum_{i=1}^m w_i a_i$ dla pewnego wektora $w = (w_1, ..., w_m) \geq 0$. Wówczas dla dowolnego $x \in C$ mamy

$$\langle y, x\rangle = \langle \sum_{i=1}^m w_i a_i, x\rangle = \sum_{i=1}^m w_i \langle a_i, x\rangle \leq 0,$$

czyli $y \in C^*$. Pokażemy, że $C^* \subseteq K$. Przypuśćmy bowiem, że istnieje wektor $y \in C^*$, taki że $y \notin K$. Ponieważ K jest stożkiem domkniętym (patrz twierdzenie 1.6.27) i wypukłym, więc na mocy twierdzenia 1.6.28 istnieje $u \in \mathbb{R}^n$, taki że

$$\langle u, z\rangle \leq 0 \text{ dla każdego } z \in K \text{ i } \langle u, y\rangle > 0.$$

Stąd, dla każdego $w \geq 0$ mamy

$$\sum_{i=1}^m w_i \langle u, a_i\rangle = \langle u, \sum_{i=1}^m w_i a_i\rangle \leq 0$$

Biorąc $w = e_j$ otrzymamy, że $\langle u, a_j\rangle \leq 0$, $j = 1, ..., m$. Zatem $u \in C$. Mamy więc $\langle u, y\rangle \leq 0$, bo $y \in C^*$. Uzyskaliśmy więc sprzeczność z nierównością $\langle u, y\rangle > 0$, co dowodzi, że $C^* \subseteq K$.

(b) Pokażemy równość $K^* = C$. Inkluzji $C \subseteq K^*$ dowodzi się podobnie, jak inkluzji $K \subseteq C^*$ w punkcie (a). Niech teraz $x \in K^*$, Wówczas dla dowolnego $w \geq 0$

$$\sum_{i=1}^m w_i \langle x, a_i\rangle = \langle x, \sum_{i=1}^m w_i a_i\rangle \leq 0.$$

Stąd już prosto wynika, że $\langle x, a_i\rangle \leq 0$ dla dowolnego i, $i = 1, ..., m$, a więc $x \in C$. ■

Definicja 1.6.31 Niech $X \subseteq \mathbb{R}^n$ i niech $x \in X$. Zbiór

$$N_X(x) := \{s \in \mathbb{R}^n : \langle s, y - x\rangle \leq 0 \text{ dla dowolnego } y \in X\}$$

nazywa się *stożkiem normalnym* do X w punkcie x.

Zauważmy, że $N_X(x)$ jest stożkiem wypukłym domkniętym. Z definicji stożka normalnego i z charakteryzacji rzutu metrycznego wynika natychmiast następujące twierdzenie.

Twierdzenie 1.6.32 *Niech $X \subseteq \mathbb{R}^n$ będzie zbiorem domkniętym i wypukłym i niech $x \in \mathbb{R}^n$. Wówczas $y = P_X(x)$ wtedy i tylko wtedy, gdy $x - y \in N_X(y)$.*

Dowód. Na mocy definicji stożka normalnego, inkluzję $x - y \in N_X(y)$ możemy zapisać w postaci $\langle x - y, z - y\rangle \leq 0$ dla dowolnego $z \in X$. Wobec tego twierdzenie wynika bezpośrednio z charakteryzacji rzutu metrycznego (twierdzenie 1.6.15). ■

Twierdzenie 1.6.30 jest szczególnym przypadkiem następującego twierdzenie, którego dowód pomijamy.

Twierdzenie 1.6.33 *Niech $X \subseteq \mathbb{R}^n$ będzie zbiorem wypukłym i niech $x \in X$. Wówczas $T_X(x)$ jest stożkiem domkniętym wypukłym oraz $(T_X(x))^* = N_X(x)$ i $(N_X(x))^* = T_X(x)$.*

1.7 Funkcje wypukłe

W kolejnych rozdziałach przekonamy się, że funkcje wypukłe odgrywają dużą rolę w optymalizacji. Dlatego teraz przedstawimy ważne własności tych funkcji, z których będziemy dalej korzystać.

Definicja 1.7.1 Niech $X \subseteq \mathbb{R}^n$ będzie podzbiorem wypukłym. Mówimy, że funkcja $f : X \longrightarrow \mathbb{R}$ jest *wypukła* (ang. *convex function*), jeśli

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y) \tag{1.16}$$

dla dowolnych $x, y \in X$ i $\lambda \in [0,1]$. Mówimy, że f jest *wklęsła* (ang. *concave function*), jeśli funkcja $-f$ jest wypukła. Jeśli nierówność w (1.16) jest ostra dla dlowolnych $x, y \in X$, $x \neq y$ i dla dowolnego $\lambda \in (0,1)$ to mówimy, że f jest *ściśle wypukła* (ang. *strictly convex function*). Jeśli istnieje stała $c > 0$, taka że

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y) - \frac{1}{2}c\lambda(1-\lambda)\|x-y\|^2 \tag{1.17}$$

dla dowolnych $x, y \in X$ i $\lambda \in [0,1]$, to mówimy, że f jest *mocno wypukła* (ang. *strongly convex function*). Stała c nazywa się *stałą mocnej wypukłości* lub *modułem* (mocnej wypukłości).

Nierówność (1.16) mówi, że odcinek łączący dwa dowolne punkty na wykresie funkcji leży nad tym wykresem. Z kolei nierówność (1.17) mówi, że różnica między punktem na tym odcinku, a wykresem funkcji jest co najmniej taka jak pewna funkcja kwadratowa znikająca na końcach tego odcinka.

Jeśli nie będzie powiedziane inaczej, w dalszej części będziemy rozważać głównie funkcje wypukłe określone na całej przestrzeni $\mathbb{R}^n$. Jednak nie wszystkie własności takich funkcji wypukłych przysługują funkcjom wypukłym zdefiniowanym na podzbiorze wypukłym $X \subseteq \mathbb{R}^n$.

1.7.1 Własności funkcji wypukłych

Poniższe cztery twierdzenia wynikają prosto z definicji funkcji wypukłej (ściśle wypukłej, mocno wypukłej) i ich dowody pozostawiamy czytelnikowi jako ćwiczenie.

Twierdzenie 1.7.2 *Jeśli $f_i : \mathbb{R}^n \longrightarrow \mathbb{R}, i = 1, ..., m$, są funkcjami wypukłymi i $\alpha_i > 0$ dla $i = 1, ..., m$, to funkcja $f := \sum_{i=1}^m \alpha_i f_i$ jest wypukła. Jeśli ponadto przynajmniej jedna z tych funkcji, powiedzmy f_j jest ściśle wypukła (mocno wypukła z modułem c), to funkcja f jest również ściśle wypukła (mocno wypukła z modułem $\alpha_j c$).*

Twierdzenie 1.7.3 *Jeśli $f_i : \mathbb{R}^n \longrightarrow \mathbb{R}, i \in I$, są funkcjami wypukłymi (mocno wypukłymi z modułem $c > 0$), to funkcja $f := \sup_{i \in I} f_i$ jest wypukła (mocno wypukła z modułem $c > 0$).*

Twierdzenie 1.7.4 *Jeśli $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ jest funkcją wypukłą i $A : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ jest odwzorowaniem afinicznym, to funkcja $h : \mathbb{R}^m \longrightarrow \mathbb{R}$, $h = f \circ A$ jest wypukła.*

Twierdzenie 1.7.5 *Jeśli $f : \mathbb{R}^n \longrightarrow Y$, gdzie $Y \subseteq \mathbb{R}$ jest zbiorem wypukłym, jest funkcją wypukłą i $g : Y \longrightarrow \mathbb{R}$ jest funkcją wypukłą niemalejącą, to funkcja $h : \mathbb{R}^n \longrightarrow \mathbb{R}$, $h = g \circ f$ jest wypukła. Jeśli ponadto g jest funkcją rosnącą i f jest funkcją ściśle wypukłą, to funkcja h jest ściśle wypukła.*

Ćwiczenie 1.7.6 Niech $f : \mathbb{R}^n \to \mathbb{R}$ będzie funkcją wypukłą (ściśle wypukłą), zaś $\|\cdot\|$ normą euklidesową w $\mathbb{R}^n$. Które z poniższych funkcji są wypukłe (ściśle wypukłe, mocno wypukłe)?

(a) $h(x) = \|x\|$,

(b) $h(x) = \|x\|^2$,

(c) $h(x) = \|f(x)\|$,

(d) $h(x) = \|f(x)\|^2$,

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

Które z tych własności zachodzą dla dowolnej normy?

Kres dolny funkcji wypukłych nie musi być funkcją wypukłą. Zachodzi natomiast następujące twierdzenie.

Twierdzenie 1.7.7 *Niech $C \subseteq \mathbb{R}^m$ będzie podzbiorem wypukłym. Jeśli $h : \mathbb{R}^n \times C \longrightarrow \mathbb{R}$ jest funkcją wypukłą ograniczoną z dołu, to funkcja $f : \mathbb{R}^n \longrightarrow \mathbb{R}$,*

$$f(x) = \inf\{h(x, y) : y \in C\},$$

jest wypukła.

Dowód. Oczywiście $f(x) > -\infty$, $x \in \mathbb{R}^n$, gdyż h jest ograniczona z dołu. Niech $x_1, x_2 \in \mathbb{R}^n$ i niech $\varepsilon > 0$. Dalej, niech $y_1, y_2 \in C$ będą takie, że $h(x_i, y_i) \leq f(x_i) + \varepsilon$, $i = 1, 2$. Wówczas dla $\lambda \in [0, 1]$ mamy

$$\begin{aligned} f((1-\lambda)x_1 + \lambda x_2) &\leq h((1-\lambda)x_1 + \lambda x_2, (1-\lambda)y_1 + \lambda y_2) \\ &\leq (1-\lambda)h(x_1, y_1) + \lambda h(x_2, y_2) \\ &\leq (1-\lambda)(f(x_1) + \varepsilon) + \lambda(f(x_2) + \varepsilon) \\ &= (1-\lambda)f(x_1) + \lambda f(x_2) + \varepsilon. \end{aligned}$$

Ponieważ $\varepsilon > 0$ jest dowolne, więc z powyższych nierówności otrzymujemy prosto tezę. ■

Ćwiczenie 1.7.8 *Niech $h(x, y) = x^2 - xy + y^2$. Wyznaczyć funkcję $f(x) = \inf\{h(x, y) : y \in \mathbb{R}\}$. Wykonać wykresy funkcji h i f.*

Lemat 1.7.9 *Funkcja wypukła $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ jest lokalnie ograniczona z góry, tzn. dla dowolnego $x \in \mathbb{R}^n$ istnieje $r > 0$ takie, że f jest ograniczona z góry na kuli $B(x, r)$.*

Dowód. Niech $x_0 = x - \frac{1}{2n}e$, $x_i = x_0 + e_i$, $i = 1, ..., n$, gdzie $e = (1, ..., 1)$) i niech

$$\Delta = \text{conv}\{x_i : i = 0, 1, ..., n\}.$$

Czytelnikowi pozostawiamy jako ćwiczenie dowód faktu, że Δ ma niepuste wnętrze (wystarczy pokazać na przykład, że $B(x, \frac{1}{2n}) \subseteq \Delta$). Niech c będzie maksymalną wartością funkcji f na zbiorze wierzchołków $x_0, x_1, ..., x_n$ zbioru Δ. Wówczas dla $r > 0$ takiego, że $B(x, r) \subseteq \Delta$ i dla dowolnego $y \in B(x, r)$ mamy $y = \sum_{i=0}^{n} \lambda_i x_i$, dla pewnego $w = (\lambda_1, ..., \lambda_m) \in \Delta_n$, i

$$f(y) = f(\sum_{i=0}^{n} \lambda_i x_i) \leq \sum_{i=0}^{n} \lambda_i f(x_i) \leq c \sum_{i=0}^{n} \lambda_i = c.$$

■

Funkcję wypukłą można zdefiniować na dowolnej przestrzeni liniowej, nawet nieskończenie wymiarowej. Należy natomiast zwrócić uwagę na to, że w dowodzie powyższego twierdzenia korzystamy z tego, że przestrzeń jest skończenie-wymiarowa. W dowodzie istotne jest również to, że f jest określona na całej przestrzeni $\mathbb{R}^n$.

Twierdzenie 1.7.10 *Funkcja wypukła $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ jest lokalnie lipschitzowska. W konsekwencji jest ona ciągła.*

Dowód. Niech $x \in \mathbb{R}^n$ i niech $h : \mathbb{R}^n \longrightarrow \mathbb{R}$, $h(z) = f(x+z) - f(x)$. Oczywiście h jest wypukła i $h(0) = 0$. Na mocy lematu 1.7.9 istnieją więc $r > 0$ i $c \in \mathbb{R}$ takie, że $h(z) \leq c$ dla każdego $z \in B(0,r)$. Niech $u \in B(x,r)$, $u \neq x$. Wówczas $u = x + z$ gdzie $z \in B(0,r)$, $z \neq 0$. Niech $y = \frac{r}{\|z\|}z$. Wówczas oczywiście $\|y\| = r$ i $z = \lambda y$ dla $\lambda = \frac{\|z\|}{r}$. Mamy więc

$$h(z) = h(\lambda y) = h(\lambda y + (1-\lambda)0) \leq \lambda h(y) + (1-\lambda)h(0) \leq \frac{\|z\|}{r}c.$$

Z drugiej strony

$$0 = h(0) = h(\frac{1}{2}z + \frac{1}{2}(-z)) \leq \frac{1}{2}h(z) + \frac{1}{2}h(-z),$$

czyli

$$h(z) \geq -h(-z) \geq -\frac{\|z\|}{r}c,$$

bo $-z \in B(0,r)$. Stąd $\mid h(z) \mid \leq \frac{\|z\|}{r}c$ i

$$\mid f(u) - f(x) \mid = \mid h(z) \mid \leq \frac{c}{r}\|z\| = \frac{c}{r}\|u - x\|,$$

a więc f jest lipschitzowska na $B(x,r)$. Jest więc ona ciągła w x. ■

Uwaga 1.7.11 Funkcja wypukła określona na podzbiorze wypukłym $X \subseteq \mathbb{R}^n$ nie musi być ciągła. Przykładem może być funkcja $f : [0,1] \to \mathbb{R}$ określona wzorem

$$f(x) = \begin{cases} 1 - \sqrt{1-x^2} & \text{dla } x \in [0,1) \\ 2 & \text{dla } x = 1. \end{cases}$$

Twierdzenie 1.7.12 *Dla dowolnej normy $\|\cdot\|$ w $\mathbb{R}^n$ i dla zbioru wypukłego $C \subseteq \mathbb{R}^n$ funkcja $d(\cdot, C) : \mathbb{R}^n \to \mathbb{R}$, $d(x,C) = \inf_{y \in C} \|x - y\|$ jest wypukła.*

Dowód. Niech $x, y \in \mathbb{R}^n$, $\lambda \in [0,1]$ i niech $z = (1-\lambda)x + \lambda y$. Bez szkody dla ogólności rozważań możemy założyć, że zbiór C jest domknięty (w razie potrzeby rozumowanie można przeprowadzić dla zbioru $\operatorname{cl} C$ i skorzystać z prostej do pokazania równości $d(x, \operatorname{cl} C) = d(x,C)$). Z wypukłości zbioru C i z w wypukłości normy $\|\cdot\|$, wynika, że

$$\begin{aligned} d((1-\lambda)x + \lambda y, C) &= d(z,C) \\ &= \|z - P_C(z)\| \\ &\leq \|(1-\lambda)x + \lambda y - (1-\lambda)P_C(x) - \lambda P_C(y)\| \\ &\leq (1-\lambda)\|x - P_C(x)\| + \lambda\|y - P_C(y)\| \\ &= (1-\lambda)d(x,C) + \lambda d(y,C). \end{aligned}$$

■

Wypukłość funkcji $d(\cdot, C)$ wynika również prosto z twierdzeń 1.7.4 i 1.7.7.

Ćwiczenie 1.7.13 Niech A będzie macierzą symetryczną stopnia n.

(a) Pokazać, że dla dowolnych $x, y \in \mathbb{R}^n$ i dla dowolnego $\lambda \in \mathbb{R}$ zachodzi równość

$$[(1-\lambda)x + \lambda y]^T A[(1-\lambda)x + \lambda] = (1-\lambda)x^T Ax + \lambda y^T Ay - (1-\lambda)\lambda(x-y)^T A(x-y). \quad (1.18)$$

Niech $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $f(x) = \frac{1}{2}x^T Ax + b^T x + c$ będzie funkcją kwadratową, gdzie $b \in \mathbb{R}^n$ i $c \in \mathbb{R}$.

(b) Korzystając z równości (1.18) pokazać, że f jest wypukła wtedy i tylko wtedy, gdy macierz A jest określona nieujemnie.

(c) Korzystając z równości (1.18) oraz z wniosku 1.4.14 pokazać, że f jest mocno wypukła wtedy i tylko wtedy, gdy macierz A jest określona dodatnio. Ponadto za moduł mocnej wypukłości można przyjąć $\lambda_{\min}(A)$.

(e) Pokazać, że

$$\|(1-\lambda)x+\lambda y\|^2 = (1-\lambda)\|x\|^2 + \lambda\|y\|^2 - (1-\lambda)\lambda\|x-y\|^2, \tag{1.19}$$

skąd wynika, że kwadrat normy euklidesowej jest funkcją mocno wypukłą z modułem mocnej wypukłości równym 2.

(f) Zauważyć, że dla macierzy dodatnio określonej A równość (1.18) możemy zapisać w postaci

$$\|(1-\lambda)x+\lambda y\|_A^2 = (1-\lambda)\|x\|_A^2 + \lambda\|y\|_A^2 - (1-\lambda)\lambda\|x-y\|_A^2, \tag{1.20}$$

gdzie $\|\cdot\|_A$ jest normą indukowaną przez iloczyn skalarny $\langle\cdot,\cdot\rangle_A$, $\langle x,y\rangle_A = x^TAy$, czyli $\|x\|_A = \sqrt{x^TAx}$.

Uwaga 1.7.14 Niech $h:\mathbb{R}^n\to\mathbb{R}$. Z twierdzenia 1.7.2 i z faktu podanego w ćwiczeniu 1.7.13(e) wynika, że jeśli funkcja h jest wypukła to funkcja $f = h+\frac{1}{2}c\|\cdot\|^2$, gdzie $c>0$, jest mocno wypukła z modułem c.

Zachodzi również własność odwrotna do podanej w uwadze 1.7.14.

Wniosek 1.7.15 *Funkcja mocno wypukła f z modułem $c>0$ jest sumą pewnej funkcji wypukłej h i funkcji kwadratowej $\frac{1}{2}c\|\cdot\|^2$*

Dowód. Z definicji funkcji mocno wypukłej i z równości (1.19) wynika prosto, że jeśli funkcja f jest mocno wypukła z modułem c, to funkcja $h := f - \frac{1}{2}c\|\cdot\|^2$ jest wypukła. Zatem,

$$f(x) = h(x) + \frac{1}{2}c\|x\|^2. \tag{1.21}$$

■

Uwaga 1.7.16 Okazuje się, że zachodzi nawet mocniejsza własność niż podana we wniosku1.7.15. Niech f będzie funkcją mocno wypukłą z modułem $c>0$ i niech $\bar{x}\in\mathbb{R}^n$. Wówczas istnieje nieujemna funkcja wypukła h osiągająca minimum w punkcie $\bar{x}$ równe 0, punkt $z\in\mathbb{R}^n$ oraz stała α takie, że

$$f(x) = h(x) + \frac{1}{2}c\|x-z\|^2 + \alpha. \tag{1.22}$$

W przypadku, gdy f jest funkcją różniczkowalną wystarczy bowiem przyjąć

$$z = \bar{x} - \frac{1}{c}\nabla f(\bar{x})$$

oraz

$$\alpha = f(\bar{x}) - \frac{1}{2}c\|\bar{x}-z\|^2. \tag{1.23}$$

Wówczas h jest wypukła, $h(\bar{x})=0$,

$$\nabla f(\bar{x}) = c(\bar{x}-z) \tag{1.24}$$

oraz

$$\nabla f(x) = \nabla h(x) + c(x - z),$$

a więc

$$\nabla f(\bar{x}) = \nabla h(\bar{x}) + c(\bar{x} - z) = \nabla h(\bar{x}) + \nabla f(\bar{x}),$$

czyli $\nabla h(\bar{x}) = 0$. Z przedstawienia (1.22) oraz z (1.23) i (1.24) wynika dla podanych wielkości z i α następujące oszacowanie

$$\begin{aligned} f^* &= \inf_x f(x) \geq \inf_x h(x) + \inf_x (\frac{1}{2}c\|x - z\|^2 + \alpha) = \alpha \\ &= f(\bar{x}) - \frac{1}{2}c\|\bar{x} - z\|^2 = f(\bar{x}) - \frac{1}{2c}\|\nabla f(\bar{x})\|^2. \end{aligned}$$

Zatem znając wartość i gradient funkcji mocno wypukłej f dla ustalonego punktu $\bar{x}$ oraz moduł mocnej wypukłości c funkcji f możemy oszacować z dołu jej kres dolny f^*:

$$f^* \geq f(\bar{x}) - \frac{1}{2c}\|\nabla f(\bar{x})\|^2. \tag{1.25}$$

Podobne oszacowanie przysługuje dowolnej funkcji mocno wypukłej (niekoniecznie różniczkowalnej). W nierówności (1.25) należy tylko zastąpić gradient przez tzw. subgradient. Szczegóły pomijamy. W kolejnym rozdziale podamy twierdzenie, z którego wynika, że dowolna funkcja mocno wypukła osiąga swoje minimum i minimizer jest określony jednoznacznie (twierdzenie 2.1.3 i wniosek 2.1.5). Podamy tam również oszacowanie odległości danego punktu od minimizera funkcji mocno wypukłej (uwaga 2.1.7).

Twierdzenie 1.7.17 *Funkcja $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ jest wypukła wtedy i tylko wtedy, gdy jej epigraf jest podzbiorem wypukłym.*

Dowód. Niech f będzie funkcją wypukłą i niech $(x, \alpha), (y, \beta) \in \operatorname{epi} f$. Wówczas dla $\lambda \in [0, 1]$ mamy

$$\begin{aligned} f((1 - \lambda)x + \lambda y) &\leq (1 - \lambda)f(x) + \lambda f(y) \\ &\leq (1 - \lambda)\alpha + \lambda\beta, \end{aligned}$$

co oznacza, że punkt $(1 - \lambda)(x, \alpha) + \lambda(y, \beta) = ((1 - \lambda)x + \lambda y, (1 - \lambda)\alpha + \lambda\beta)$ jest elementem epigrafu funkcji f. Odwrotnie, niech $\operatorname{epi} f$ będzie podzbiorem wypukłym. Niech $\lambda \in [0, 1]$. Ponieważ $(x, f(x)), (y, f(y)) \in \operatorname{epi} f$, więc $((1-\lambda)x+\lambda y, (1-\lambda)f(x)+\lambda f(y)) = (1-\lambda)(x, f(x))+\lambda(y, f(y)) \in \operatorname{epi} f$, czyli

$$f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y),$$

a więc funkcja f jest wypukła. ■

Wniosek 1.7.18 *Jeśli $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ jest funkcją wypukłą, to istnieje wektor $a \in \mathbb{R}^n$ i liczba $\alpha \in \mathbb{R}$ takie, że*

$$f(y) \geq a^T y + \alpha \tag{1.26}$$

dla dowolnego $y \in \mathbb{R}^n$.

Dowód. Niech $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ będzie funkcją wypukłą i niech $x \in \mathbb{R}^n$. Ponieważ $\operatorname{epi} f$ jest podzbiorem wypukłym i $(x, f(x)) \in \operatorname{bd} \operatorname{epi} f$, więc z twierdzenia o słabym oddzielaniu wynika, że istnieje para $(a, \gamma) \in \mathbb{R}^n \times \mathbb{R}$ taka, że

$$[a^T, \gamma] \begin{bmatrix} x \\ f(x) \end{bmatrix} \geq [a^T, \gamma] \begin{bmatrix} y \\ \beta \end{bmatrix}$$

dla dowolnej pary $(y, \beta) \in \operatorname{epi} f$. Zauważmy, że $\gamma < 0$. Wobec tego, bez szkody dla ogólności rozważań można przyjąć $\gamma = -1$. Dla $\beta = f(y)$ otrzymamy wówczas $f(y) \geq a^T y + \alpha$, gdzie $\alpha = -a^T x + f(x)$. ■

Wniosek 1.7.19 *Funkcja mocno wypukła jest koercytywna.*

Dowód. Niech f będzie funkcją mocno wypukłą z modułem $c > 0$. Na mocy wniosku 1.7.15 $f(x) = h(x) + \frac{1}{2}c\|x\|^2$, gdzie h jest funkcją wypukłą. Z wniosku 1.7.18 wynika, że $h(x) \geq a^T x + \alpha$ dla pewnego $a \in \mathbb{R}^n$ i stałej $\alpha \in \mathbb{R}$. Stąd i z nierówności Cauchy'ego–Schwarza otrzymujemy

$$f(x) \geq -\|a\| \cdot \|x\| + \alpha + \frac{1}{2}c\|x\|^2.$$

Ponieważ prawa strona dąży do $+\infty$, gdy $\|x\| \to +\infty$, więć $\lim_{\|x\|\to+\infty} f(x) = +\infty$. ■

Twierdzenie 1.7.20 *Jeśli funkcja $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ jest wypukła, to dowolna jej podpoziomica $S(f, \alpha)$, gdzie $\alpha \in \mathbb{R}$, jest zbiorem wypukłym.*

Dowód. Niech f będzie funkcją wypukłą i niech $x, y \in S(f, \alpha)$, gdzie $\alpha \in \mathbb{R}$ oraz niech $\lambda \in [0, 1]$. Mamy wówczas

$$\begin{aligned} f((1-\lambda)x + \lambda y) &\leq (1-\lambda)f(x) + \lambda f(y) \\ &\leq (1-\lambda)\alpha + \lambda\alpha = \alpha, \end{aligned}$$

czyli $(1-\lambda)x + \lambda y \in S(f, \alpha)$. ■

Uwaga 1.7.21 Twierdzenie odwrotne do powyższego nie jest prawdziwe. Aby to zauważyć wystarczy rozpatrzyć funkcję $\chi_C : \mathbb{R}^n \longrightarrow \mathbb{R}$,

$$\chi_C(x) = \begin{cases} 0 & \text{dla } x \in C \\ 1 & \text{dla } x \notin C, \end{cases}$$

gdzie $C \subseteq \mathbb{R}^n$ jest zbiorem wypukłym. Funkcja, której wszystkie podpoziomice są zbiorami wypukłymi nazywa się funkcją *quasi-wypukłą* (ang. *quasi-convex*).

Twierdzenie 1.7.22 *Niech $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ będzie funkcją wypukłą. Wówczas dla dowolnego punktu $x \in \mathbb{R}^n$ i dla dowolnego kierunku $s \in \mathbb{R}^n$ istnieje pochodna kierunkowa $f'(x, s)$ oraz zachodzi równość*

$$f'(x, s) = \inf_{t>0} \frac{f(x+ts) - f(x)}{t}.$$

Dowód. Pokażemy najpierw, że dla wypukłej funkcji f funkcja $h : \mathbb{R} \longrightarrow \mathbb{R}$, $h(t) = [f(x + ts) - f(x)]/t$ jest niemalejąca. Niech $0 < t_1 \leq t_2$. Wówczas $t_1/t_2 \in (0, 1]$ i

$$\begin{aligned} f(x + t_1 s) - f(x) &= f((1 - \frac{t_1}{t_2})x + \frac{t_1}{t_2}(x + t_2 s)) - f(x) \\ &\leq (1 - \frac{t_1}{t_2})f(x) + \frac{t_1}{t_2}f(x + t_2 s) - f(x) \\ &= \frac{t_1}{t_2}(f(x + t_2 s) - f(x)), \end{aligned}$$

tzn. h jest funkcją niemalejącą. Ponieważ każda funkcja niemalejąca posiada granice jednostronne, więc zachodzą równości

$$\inf_{t>0} h(t) = \lim_{t\downarrow 0} h(t) = f'(x, s).$$

■

Twierdzenie 1.7.23 *Jeśli $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ jest funkcją wklęsłą i $X \subseteq \mathbb{R}^n$ jest wielościanem, to istnieje punkt ekstremalny zbioru X, w którym f osiąga minimum na tym zbiorze.*

Dowód. Ponieważ funkcja wklęsła f jest ciągła zaś wielościan jest zbiorem zwartym, więc f osiąga minimum na zbiorze X na mocy twierdzenia Weierstrassa. Niech więc $x^* \in X$ będzie takie, że $f(x) \geq f(x^*) = f^*$ dla każdego $x \in X$. Niech $w_1, ..., w_p \in X$ będą punktami ekstremalnymi zbioru X takimi, że $x^* = \sum_{i=1}^{p} \lambda_i w_i$, dla pewnego wektora $u = (\lambda_1, ..., \lambda_p) \in \Delta_p$. Punkty takie istnieją na mocy twierdzenia Minkowskiego (tw. 1.6.21). Przypuśćmy, że $f(w_i) > f^*$ dla $i = 1, ..., p$. Wówczas na mocy wklęsłości funkcji f mamy

$$f^* = f(x^*) = f(\sum_{i=1}^{p} \lambda_i w_i) \geq \sum_{i=1}^{p} \lambda_i f(w_i) > \sum_{i=1}^{p} \lambda_i f^* = f^*.$$

Otrzymana sprzeczność dowodzi prawdziwości twierdzenia. ■

Poniższy wniosek z twierdzenia 1.7.23 ma zastosowanie na przykład w programowaniu liniowym.

Wniosek 1.7.24 *Funkcja liniowa określona na wielościanie osiąga swoje kresy w wierzchołkach tego wielościanu.*

1.7.2 Funkcja wypukła różniczkowalna

Twierdzenie 1.7.25 *Niech $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ będzie funkcją różniczkowalną. Wówczas:*

(i) *f jest wypukła wtedy i tylko wtedy, gdy*

$$\nabla f(x)^T (y - x) \leq f(y) - f(x), \tag{1.27}$$

dla dowolnych $x, y \in \mathbb{R}^n$;

(ii) *f jest ściśle wypukła wtedy i tylko wtedy, gdy*

$$\nabla f(x)^T (y - x) < f(y) - f(x), \tag{1.28}$$

dla dowolnych $x, y \in \mathbb{R}^n, x \neq y$;

(iii) *f jest mocno wypukła z modułem $c > 0$ wtedy i tylko wtedy, gdy*

$$\nabla f(x)^T (y - x) + \frac{1}{2} c \|x - y\|^2 \leq f(y) - f(x) \tag{1.29}$$

dla dowolnych $x, y \in \mathbb{R}^n$.

Dowód. Przeprowadzimy najpierw dowód części (iii). Przypuśćmy, że funkcja f jest mocno wypukła z modułem $c > 0$. Dla $x, y \in \mathbb{R}^n$ i dla $\lambda \in (0, 1)$ otrzymujemy wówczas, na mocy różniczkowalności funkcji f,

$$\begin{aligned} f(y) - f(x) &= \frac{(1-\lambda)f(x) + \lambda f(y) - f(x)}{\lambda} \\ &\geq \frac{f((1-\lambda)x + \lambda y) - f(x) + \frac{1}{2}c(1-\lambda)\lambda\|y-x\|^2}{\lambda} \\ &= \frac{f(x + \lambda(y-x)) - f(x) + \frac{1}{2}c(1-\lambda)\lambda\|y-x\|^2}{\lambda} = \\ &= (y-x)^T \nabla f(x) + \frac{1}{2}c(1-\lambda)\|y-x\|^2 + \frac{o(\lambda\|y-x\|)}{\lambda} \end{aligned}$$

W granicy przy $\lambda \downarrow 0$ otrzymamy nierówność (1.29). Przypuśćmy teraz, że dla dowolnych $x, y \in \mathbb{R}^n$ spełniona jest nierówność (1.29) dla pewnego $c > 0$. Niech $\lambda \in (0,1)$ i niech $z = (1-\lambda)x+\lambda y$. Wówczas oczywiście $z - x = \lambda(y - x)$ i $y - z = (1 - \lambda)(y - x)$. W konsekwencji otrzymujemy

$$\begin{aligned} \frac{f(x) - f(z)}{\lambda} &\geq \frac{(x - z)^T \nabla f(z) + \frac{1}{2}c\|x - z\|^2}{\lambda} \\ &= -(y - x)^T \nabla f(z) + \frac{1}{2}c\lambda\|y - x\|^2 \end{aligned}$$

i

$$\begin{aligned} \frac{f(y) - f(z)}{1 - \lambda} &\geq \frac{(y - z)^T \nabla f(z) + \frac{1}{2}c\|y - z\|^2}{1 - \lambda} \\ &= (y - x)^T \nabla f(z) + \frac{1}{2}c(1 - \lambda)\|y - x\|^2. \end{aligned}$$

Dodając powyższe nierówności stronami otrzymamy po prostych przekształceniach

$$f(z) \leq (1 - \lambda)f(x) + \lambda f(y) - \frac{1}{2}c\lambda(1 - \lambda)\|y - x\|^2.$$

Zatem f jest funkcją mocno wypukłą ze stała mocnej wypukłości c.

Przyjmując w powyższym dowodzie $c = 0$ otrzymamy część (i), natomiast zastępując dodatkowo nierówności słabe ostrymi i zakładając, że $x \neq y$ otrzymamy część (ii). ■

Uwaga 1.7.26 Część (i) twierdzenia 1.7.25 można również sformułować w następujący sposób: funkcja różniczkowalna f jest wypukła wtedy i tylko wtedy gdy jest ona niemniejsza od dowolnej swojej linearyzacji. Czytelnikowi pozostawiamy sformułowanie w języku linearyzacji pozostałych części tego twierdzenia.

Definicja 1.7.27 Odwzorowanie $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ nazywamy *monotonicznym* (ang. *monotone mapping*) jeśli

$$(F(y) - F(x))^T (y - x) \geq 0$$

dla dowolnych $x, y \in \mathbb{R}^n$. Jeśli nierówność powyższa jest ostra dla dowolnych $x, y \in \mathbb{R}^n$, $x \neq y$, to odwzorowanie F nazywa się *ściśle monotoniczne* (ang. *strictly monotone mapping*). Jeśli natomiast

$$(F(y) - F(x))^T (y - x) \geq c\|y - x\|^2,$$

dla pewnej stałej $c > 0$, to F nazywa się odwzorowaniem *mocno monotonicznym* (ang. *strongly monotone mapping*), zaś stała c nazywa się *modułem* mocnej monotoniczności.

Wniosek 1.7.28 *Niech $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ będzie funkcją różniczkowalną. Wówczas*

(i) *Jeśli f jest wypukła, to jej gradient $\nabla f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ jest odwzorowaniem monotonicznym, tzn.*

$$(\nabla f(y) - \nabla f(x))^T (y - x) \geq 0 \tag{1.30}$$

dla dowolnych $x, y \in \mathbb{R}^n$.

(ii) *Jeśli f jest ściśle wypukła, to jej gradient $\nabla f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ jest odwzorowaniem ściśle monotonicznym, tzn.*

$$(\nabla f(y) - \nabla f(x))^T (y - x) > 0 \tag{1.31}$$

dla dowolnych $x, y \in \mathbb{R}^n$, $x \neq y$.

(iii) *Jeśli f jest mocno wypukła (z modułem c), to jej gradient $\nabla f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ jest odwzorowaniem mocno monotonicznym (z modułem c), tzn.*

$$(\nabla f(y) - \nabla f(x))^T(y - x) \geq c\|y - x\|^2 \tag{1.32}$$

dla dowolnych $x, y \in \mathbb{R}^n$.

Dowód. Pokażemy tylko własność (iii). Pozostałe własności pokazuje się w podobny sposób. Przypuśćmy, że f jest funkcją mocno wypukłą. Wówczas na mocy twierdzenia 1.7.25(iii) dla dowolnych $x, y \in \mathbb{R}^n$ zachodzą nierówności

$$-\nabla f(x)^T(y - x) \geq f(x) - f(y) + \frac{1}{2}c\|x - y\|^2.$$

Zamieniając rolami x i y we wzorze (1.29) otrzymamy również

$$\nabla f(y)^T(y - x) \geq f(y) - f(x) + \frac{1}{2}c\|y - x\|^2.$$

Dodając te nierówności stronami otrzymamy tezę. ■

Uwaga 1.7.29 Można pokazać, że zachodzą również implikacje odwrotne do przedstawionych we wniosku 1.7.28. Szczegóły można znaleźć w podręczniku [HL93].

Twierdzenie 1.7.30 *Niech $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ będzie funkcją dwukrotnie różniczkowalną i niech $c > 0$. Wówczas:*

(i) *f jest wypukła wtedy i tylko wtedy, gdy dla dowolnego $x \in \mathbb{R}^n$ hesjan $\nabla^2 f(x)$ jest macierzą nieujemnie określoną.*

(ii) *f jest ściśle wypukła jeśli dla dowolnego $x \in \mathbb{R}^n$ hesjan $\nabla^2 f(x)$ jest macierzą dodatnio określoną.*

(iii) *f jest mocno wypukła ze stałą mocnej wypukłości c wtedy i tylko wtedy, gdy dla dowolnych $x, s \in \mathbb{R}^n$ zachodzi nierówność $s^T\nabla^2 f(x)s \geq c\|s\|^2$.*

Dowód. (i) Ponieważ f jest dwukrotnie różniczkowalna, więc dla dowolnych $x, s \in \mathbb{R}^n$ i dla $t > 0$ mamy

$$\nabla f(x + ts) - \nabla f(x) = t\nabla^2 f(x)s + \mathbf{o}(t),$$

gdzie $\mathbf{o}(t)$ jest odwzorowaniem, którego współrzędne są wielkościami typu $o(t)$. Przypuśćmy, że f jest funkcją wypukłą. Wówczas na mocy wniosku 1.7.28(i) mamy

$$0 \leq s^T(\nabla f(x + ts) - \nabla f(x)) = ts^T\nabla^2 f(x)s + s^T\mathbf{o}(t).$$

Dzieląc powyższą nierówność obustronnie przez $t > 0$ i przechodząc do granicy przy $t \downarrow 0$ otrzymamy $s^T\nabla^2 f(x)s \geq 0$. Zatem hesjan $\nabla^2 f(x)$ jest macierzą nieujemnie określoną. Przypuśćmy teraz, że dla dowolnego $z \in \mathbb{R}^n$ hesjan $\nabla^2 f(z)$ jest macierzą nieujemnie określoną. Niech $y \in \mathbb{R}^n$ będzie dowolny i niech $d = y - x$. Rozwijając funkcję f we wzór Taylora w otoczeniu punktu x, z resztą Lagrange'a otrzymujemy

$$f(x + d) = f(x) + d^T\nabla f(x) + \frac{1}{2}d^T\nabla^2 f(x + \lambda d)d$$

dla pewnego $\lambda \in (0, 1)$. Na mocy założenia, $d^T\nabla^2 f(x + \lambda d)d \geq 0$ i w konsekwencji

$$f(y) \geq f(x) + (y - x)^T\nabla f(x).$$

Nierówność ta zgodnie z twierdzeniem 1.7.25(i) oznacza, że funkcja f jest wypukła.

Części (ii) oraz (iii) dowodzi się podobnie. ■

Uwaga 1.7.31 Implikacja odwrotna do (ii) w powyższym twierdzeniu nie jest prawdziwa. Aby to zauważyć wystarczy rozpatrzyć funkcję $f : \mathbb{R} \longrightarrow \mathbb{R}$, $f(x) = x^4$.

Ćwiczenie 1.7.32 Pokazać wypukłość względnie wklęsłość następujących funkcji $f : X \to \mathbb{R}$:

(a) funkcja potęgowa, $f(x) = |x|^\alpha$, gdzie $\alpha \geq 1$, $X = \mathbb{R}$,

(b) funkcja wykładnicza $f : \mathbb{R} \to \mathbb{R}$, $f(x) = a^x$, gdzie $a \geq 0$, $X = \mathbb{R}$,

(c) funkcja logarytmiczna, $f(x) = \ln x$, $X = \mathbb{R}_{++}$,

(d) entropia, $f(x) = \sum_{i=1}^n x_i \ln x_i$, $X = \mathbb{R}^n_{++}$,

(e) średnia geometryczna, $f(x) = (x_1 \cdot ... \cdot x_n)^{\frac{1}{n}}$, $X = \mathbb{R}^n_{++}$,

Wskazówka: wyznaczyć hesjan funkcji $-f$ i pokazać jego nieujemną określoność ($s^\top \nabla^2 f(x) s \geq 0$ dla dowolnego $s = (\sigma_1, ..., \sigma_n)$) korzystając z nierówności

$$\left(\sum_{j=1}^n \frac{\sigma_j}{x_j} \right)^2 \leq n \sum_{j=1}^n \left(\frac{\sigma_j}{x_j} \right)^2$$

będącej szczególnym przypadkiem nierówności Cauchy'ego–Schwarza dla standardowego iloczynu skalarnego i indukowanej przez niego normy euklidesowej.

(f) $f(x) = \max\{|x_j| : j = 1, ..., n\}$, $X = \mathbb{R}^n$ (norma w przestrzeni l^∞)

(g) norma w przestrzeni l^p, $f(x) = (\sum_{j=1}^n |x_i|^p)^{\frac{1}{p}}$, $p \geq 1$, $X = \mathbb{R}^n$.

Wskazówka: skorzystać z nierówności Minkowskiego.

Ćwiczenie 1.7.33 Pokazać wklęsłość poniższych funkcji $f : \mathbb{R}^n_{++} \to \mathbb{R}$ występujących w ekonomii:

(a) funkcja użyteczności postaci $f(x) = \min\{\frac{x_j}{a_j} : j = 1, ..., n\}$, gdzie $a_j > 0$,

(b) funkcja produkcji Cobba–Douglasa $f(x) = x_1^{\beta_1} \cdot ... \cdot x_n^{\beta_n}$, gdzie $\beta_j > 0$, $\sum_{j=1}^n \beta_j \leq 1$,

Wskazówka: wyznaczyć hesjan funkcji $-f$ i pokazać jego nieujemną określoność ($s^\top \nabla^2 f(x) s \geq 0$ dla dowolnego $s = (\sigma_1, ..., \sigma_n)^\top$) korzystając z nierówności

$$\left(\sum_{j=1}^n \frac{\sigma_j \beta_j}{x_j} \right)^2 \leq \sum_{j=1}^n \left(\frac{\sigma_j \sqrt{\beta_j}}{x_j} \right)^2 \left(\sum_{j=1}^n \beta_j \right)$$

będącej szczególnym przypadkiem nierówności Cauchy'ego–Schwarza dla standardowego iloczynu skalarnego i indukowanej przez niego normy euklidesowej.

(Funkcja podana w ćwiczeniu 1.7.32(e) jest szczególnym przypadkiem funkcji Cobba–Douglasa.)

(c) funkcja CES (ang. *constant elasticity of substitution*) $f(x) = (\sum_{j=1}^n \alpha_j \xi_j^p)^{\frac{1}{p}}$, $\alpha_j > 0$, gdzie $\sum_{j=1}^n \alpha_j = 1$, $p \in (-\infty, 1) \backslash \{0\}$.

Wskazówka: wyznaczyć gradient funkcji $-f$ i pokazać jego monotoniczność. Czy funkcja CES jest również wklęsła bez założenia $\sum_{j=1}^n \alpha_j = 1$?

1.8 Iteracje punktu stałego

Niektóre metody w optymalizacji mają postać tzw. iteracji punktu stałego. W tym ustępie przedstawimy warunki zbieżności dla takich iteracji.

Definicja 1.8.1 Niech $X \subseteq \mathbb{R}^n$ i $T : X \to X$. Punkt $x \in X$ spełniający równanie $T(x) = x$ nazywamy *punktem stałym* (ang. *fixed point*) operatora T. Zbiór $\operatorname{Fix} T := \{x \in X : T(x) = x\}$ nazywamy *zbiorem punktów stałych* (ang. *fixed point set*) operatora T.

Definicja 1.8.2 Niech $X \subseteq \mathbb{R}^n$, $Y \subseteq \mathbb{R}^m$ i niech $c \in (0, 1)$. Operator $T : X \to Y$ nazywa się *c-kontrakcją* (lub po prostu kontrakcją, ang. *contraction*), jeśli dla dowolnych $x, y \in X$,

$$\|T(x) - T(y)\| \leq c\|x - y\|. \tag{1.33}$$

Operator T nazywa się c-kontrakcją względem punktu $x^* \in X$, jeśli dla dowolnego $x \in X$ zachodzi

$$\|T(x) - x^*\| \leq c\|x - x^*\|. \tag{1.34}$$

Zauważmy, że jeśli T jest c-kontrakcją względem punktu $x^* \in X$, to x^* jest punktem stałym odwzorowania T, nawet jeśli w (1.34) wziąć $c \in \mathbb{R}_+$.

Przykład 1.8.3 Punktem stałym odwzorowania $T : \mathbb{R} \to \mathbb{R}$ określonego równością $T(x) := \frac{1}{2}(x + \frac{1}{x})$ jest $\sqrt{2}$. Fakt ten jest podstawą algorytmu wyznaczania $\sqrt{2}$.

Przykład 1.8.4 Niech $f : \mathbb{R}^n \to \mathbb{R}$ będzie funkcją różniczkowalną, zaś $A(x)$ macierzą nieosobliwą typu $n \times n$ dla dowolnego $x \in \mathbb{R}^n$. Nietrudno zauważyć, że zbiorem punktów stałych operatora $T : \mathbb{R}^n \to \mathbb{R}^n$ określonego wzorem $T(x) := x - A(x)\nabla f(x)$ jest zbiór punktów stacjonarnych funkcji f. Własność ta jest o tyle ważna, że wiele metod minimalizacji różniczkowalnej (na przykład metoda Newtona) ma postać iteracji $x^{k+1} = x^k - A(x^k)\nabla f(x_k)$, gdzie $A(x^k)$ jest macierzą nieosobliwą.

Niech $T : X \to X$, gdzie $X \subseteq \mathbb{R}^n$, będzie operatorem posiadającym punkt stały $x^* \in X$. *Iteracja punktu stałego* operatora T ma postać

$$x_{k+1} = T(x_k),$$

$k \geq 0$, gdzie $x_0 \in X$. Zauważmy, że jeśli x_k jest punktem stałym operatora T, to x_{k+1} jest tym samym punktem stałym tego operatora. Zachodzi następujące twierdzenie.

Twierdzenie 1.8.5 *Niech $X \subseteq \mathbb{R}^n$ będzie podzbiorem domkniętym oraz $T : X \to X$ będzie c-kontrakcją względem punktu $x^* \in X$, gdzie $c \in (0, 1)$. Wówczas dla dowolnego $x_0 \in X$ ciąg x_k określony przez iterację $x_{k+1} = T(x_k)$ jest zbieżny geometrycznie do x^*, a dokładniej*

$$\|x_k - x^*\| \leq c^k\|x_0 - x^*\|.$$

Aby móc sprawdzić, czy spełniona jest nierówność (1.34), zazwyczaj musimy znać punkt stały x^*. Natomiast podobna własność przysługuje c-kontrakcji nawet jeśli nie znamy jej punktu stałego, co jest sformułowane w słynnym twierdzeniu Banacha o punkcie stałym. Poniżej podajemy jego szczególny przypadek dla przestrzeni $\mathbb{R}^n$.

Twierdzenie 1.8.6 (Banach, 1922) *Niech $X \subseteq \mathbb{R}^n$ będzie podzbiorem domkniętym oraz $T : X \to X$ będzie c-kontrakcją, gdzie $c \in (0,1)$. Wówczas operator T posiada dokładnie jeden punkt stały $x^* \in X$ oraz dla dowolnego $x_0 \in X$ ciąg x_k określony przez iterację $x_{k+1} = T(x_k)$ jest zbieżny geometrycznie do x^*, a dokładniej*

$$\|x_k - x^*\| \leq \frac{c^k}{1-c}\|x_0 - T(x_0)\|.$$

Twierdzenie Banacha jest dla nas ważne o tyle, że metody minimalizacji funkcji $f : \mathbb{R}^n \to \mathbb{R}$ podawane są często w postaci iteracji $x_{k+1} = T(x_k)$, gdzie punkt startowy $x_0 \in \mathbb{R}^n$, zaś $T : \mathbb{R}^n \to \mathbb{R}^n$ jest kontrakcją, dla której zbiór punktów stałych pokrywa się ze zbiorem punków stacjonarnych funkcji f. Poniższe twierdzenie, będące wnioskiem twierdzenia Banacha podaje warunki wystarczające zbieżności ciągu iteracji generowanych przez operator różniczowalny do punktu stałego tego operatora.

Twierdzenie 1.8.7 *Niech $X \subseteq \mathbb{R}^n$ będzie podzbiorem domkniętym i wypukłym, zaś $T : X \to \mathbb{R}^n$ będzie operatorem różniczkowalnym w sensie Gâteaux. Jeśli istnieje stała $c \in (0,1)$ taka, że $\|\nabla T(x)\| \leq c$ dla dowolnego $x \in X$, to operator T jest c-kontrakcją. W konsekwencji, jeśli $T(X) \subseteq X$, to T posiada dokładnie jeden punkt stały $x^* \in D$ oraz dla dowolnego punktu $x_0 \in X$ ciąg x_k określony przez iterację $x_{k+1} = T(x_k)$ jest zbieżny geometrycznie do x^*.*

Dowód. Niech $t \in [0,1]$ i $x, y \in X$. Wówczas $z := (1-t)x + ty \in X$, ponieważ X jest zbiorem wypukłym. Dla $u \in \mathbb{R}^n$, $\|u\| = 1$, zdefiniujmy

$$\varphi(t) := \langle T(z), u\rangle = \langle T((1-t)x + ty), u\rangle.$$

Oczywicie funkcja φ jest ciągła na odcinku $[0,1]$ jako złożenie funkcji ciągłych. Pokażemy, że φ jest różniczkowalna na $(0,1)$. Dla $t \in (0,1)$ i h takiego, że $t + h \in [0,1]$ mamy

$$\begin{aligned}
\varphi'(t) &= \lim_{h \downarrow 0} \frac{\varphi(t+h) - \varphi(t)}{h} = \lim_{h \downarrow 0} \frac{\langle T(z + h(y-x)) - T(z), u\rangle}{h} \\
&= \left\langle \lim_{h \downarrow 0} \frac{T(z + h(y-x)) - T(z)}{h}, u \right\rangle \\
&= \langle T'(z, y-x), u\rangle = \nabla T(z)^T (y-x))^T u.
\end{aligned}$$

Ostatnia z powyższych równości wynika z zastosowania różniczkowalności sensie Gâteaux do współrzędnych operatora T. Mamy więc

$$\varphi'(t) = \langle \nabla T(z)^T (y-x)), u\rangle. \tag{1.35}$$

Na mocy twierdzenia Lagrange'a o wartości średniej, istnieje $t^* \in (0,1)$ takie, że

$$\frac{\varphi(1) - \varphi(0)}{1-0} = \varphi'(t^*),$$

w konsekwencji,

$$|\varphi(1) - \varphi(0)| \leq \sup_{t \in [0,1]} |\varphi'(t)|.$$

Łącznie z równością (1.35), z nierównością Cauchy'ego–Schwarza i z definicją normy operatora, nierówność powyższa oznacza, że

$$\begin{aligned}
\|T(y) - T(x)\| &= |\varphi(1) - \varphi(0)| \leq \sup_{t \in [0,1]} \left|\langle \nabla T(z)^T (y-x), u\rangle\right| \\
&\leq \sup_{t \in [0,1]} \|\nabla T(z)\| \cdot \|y-x\| \cdot \|u\| = \sup_{t \in [0,1]} \|\nabla T(z)\| \cdot \|y-x\| \leq c\|y-x\|.
\end{aligned}$$

Zatem T jest kontrakcją. Dalsza częśc twierdzenia wynika z twierdzenia Banacha o punkcie stałym. ■