Lokalne ektrema, formy kwadratowe

Ostatnio poprawiłem 6 grudnia 2014 r.

Wypada raz jeszcze wrócić do ekstremów warunkowych. W przypadku ekstremów funkcji rozpatrywanych na zbiorach otwartych podaliśmy warunek wystarczający na to, aby funkcja miała w pewnym punkcie ekstremum lokalne. Zrobimy teraz to samo w przypadku funkcji rozpatrywanej na zbiorze zdefiniowanym za pomocą równań, określonej na większym zbiorze otwartym, czyli podamy warunek wystarczający na to, by funkcja miała w pewnym punkcie lokalne ekstremum związane (warunkowe).

Twierdzenie 6.1 (o lokalnych ekstremach warunkowych, war. dostateczny)
Niech $F\colon G \longrightarrow \mathbb{R}^l$ będzie odwzorowaniem klasy C^2 ze zbioru G otwartego w $\mathbb{R}^{k+l}$ zaś $f\colon G \longrightarrow \mathbb{R}$ — funkcją klasy C^2. Załóżmy, że $\mathbf{0} \in \mathbb{R}^l$ jest wartością regularną odwzorowania F, tzn. że jeżeli $F(\mathbf{x}) = \mathbf{0}$, to $DF(\mathbf{x})\colon \mathbb{R}^{k+l} \longrightarrow \mathbb{R}^l$ jest epimorfizmem. Niech $\mathbf{p} \in M = F^{-1}(\mathbf{0})$ będzie takim punktem, że istnieją takie liczby $\lambda_1, \lambda_2, \ldots, \lambda_l$, że $\operatorname{grad} f(\mathbf{p}) = \sum_j \lambda_j \operatorname{grad} F_j(\mathbf{p})$. Niech $L(\mathbf{x}) = f(\mathbf{x}) - \sum_j \lambda_j F_j(\mathbf{x})$ (L nazywana jest funkcją Lagrange'a). W tej sytuacji $DL(\mathbf{p}) = \mathbf{0}$ oraz

- **a.** jeżeli $D^2L(\mathbf{p})\mathbf{v}^2 > 0$ dla każdego wektora $\mathbf{v} \in T_{\mathbf{p}}M \setminus \{\mathbf{0}\}$ ($D^2L(\mathbf{p})$ jest dodatnio określona na przestrzeni stycznej w punkcie $\mathbf{p}$ do zbioru M), to funkcja $f_{|M}$ ma lokalne minimum właściwe w punkcie $\mathbf{p}$;
- **b.** jeżeli $D^2L(\mathbf{p})\mathbf{v}^2 < 0$ dla każdego wektora $\mathbf{v} \in T_{\mathbf{p}}M \setminus \{\mathbf{0}\}$ ($D^2L(\mathbf{p})$ jest ujemnie określona na przestrzeni stycznej w punkcie $\mathbf{p}$ do zbioru M), to funkcja $f_{|M}$ ma lokalne maksimum właściwe w punkcie $\mathbf{p}$;
- **c.** jeżeli $D^2L(\mathbf{p})\mathbf{v}^2 > 0 > D^2L(\mathbf{p})\mathbf{w}^2$ dla pewnych wektorów $\mathbf{v}, \mathbf{w} \in T_{\mathbf{p}}M$, to funkcja $f_{|M}$ nie ma lokalnego ekstremum w punkcie $\mathbf{p}$.

Dowód. Zauważmy, że $L_{|M} = f_{|M}$. Wobec tego możemy zajmować się w dalszym ciągu funkcją L. Z twierdzenia o funkcji uwikłanej wynika, że istnieją: takie k–wymiarowe otoczenie U punktu $\mathbf{0} \in \mathbb{R}^k$ oraz $k+l$–wymiarowe otoczenie V punktu $\mathbf{p} \in \mathbb{R}^{k+l}$ i taki homeomorfizm φ zbioru U na zbiór $V \cap M$, że dla każdego $\mathbf{x} \in U$ różniczka $D\varphi(\mathbf{x})$ jest włożeniem (monomorfizmem) oraz $\varphi(\mathbf{0}) = \mathbf{p}$. Mamy $D(L \circ \varphi) = DL \circ \varphi \cdot D\varphi$ i wobec tego

$$D^2(L \circ \varphi)(\mathbf{x})\mathbf{v}^2 = D^2L(\varphi(\mathbf{x}))\big(D\varphi(\mathbf{x})\mathbf{v}\big)^2 + DL(\varphi(\mathbf{x}))\big(D^2\varphi(\mathbf{x})\mathbf{v}^2\big).$$

Dla $\mathbf{x} = \mathbf{0}$ mamy więc

$$\begin{aligned} D^2(L \circ \varphi)(\mathbf{0})\mathbf{v}^2 &= D^2L(\varphi(\mathbf{0}))\big(D\varphi(\mathbf{0})\mathbf{v}\big)^2 + DL(\varphi(\mathbf{0}))\big(D^2\varphi(\mathbf{0})\mathbf{v}^2\big) = \\ &= D^2L(\mathbf{p})\big(D\varphi(\mathbf{0})\mathbf{v}\big)^2 + DL(\mathbf{p})\big(D^2\varphi(\mathbf{0})\mathbf{v}^2\big) = D^2L(\mathbf{p})\big(D\varphi(\mathbf{0})\mathbf{v}\big)^2, \end{aligned}$$

bo L zdefiniowaliśmy tak, by $DL(\mathbf{p}) = \mathbf{0}$. Teza wynika teraz od razu z twierdzenia o lokalnych ekstremach zastosowanego do funkcji $L \circ \varphi$ określonej na zbiorze U otwartym w $\mathbb{R}^k$. ■

Uwaga 6.2
W twierdzeniu o lokalnych ekstremach warunkowych trzeba koniecznie rozpatrywać funkcję Lagrange'a L zamiast funkcji f, chociaż te dwie funkcje pokrywają się na

zbiorze M. Podamy przykład. Niech $F\binom{x}{y} = -x + \left(\frac{x}{2} - 2y\right)^2$, $f\binom{x}{y} = x - y^2$. Mamy $F\binom{0}{0} = 0$, $\operatorname{grad} F\binom{0}{0} = \binom{-1}{0}$, $\operatorname{grad} f\binom{0}{0} = \binom{1}{0}$, zatem w punkcie $\binom{0}{0}$ jest spełniony warunek Lagrange'a dla funkcji f na zbiorze $M = F^{-1}(0)$.

$$T_{\mathbf{0}}M = \ker\left(DF\binom{0}{0}\right) = \left\{\binom{x}{y}: \quad x = 0\right\}.$$

Mamy więc $D^2 f\binom{0}{0}\binom{0}{y}^2 = -2y^2$, co sugeruje, że funkcja $f_{|M}$ ma w punkcie $\mathbf{0}$ lokalne maksimum. To jednak nie jest prawda. Niech $\varphi(t) = \binom{4t^2}{t+t^2}$. Mamy więc, $F(\varphi(t)) = = -4t^2 + \left(2t^2 - 2t - 2t^2\right)^2 = 0$ i $D\varphi(\mathbf{0}) = \binom{0}{1}$, zatem φ parametryzuje pewne otoczenie punktu $\mathbf{0}$ w M. Mamy również $f(\varphi(t)) = 4t^2 - \left(t + t^2\right)^2 = 3t^2 - 2t^3 - t^4$. Jasne jest więc, że funkcja $f \circ \varphi$ ma w punkcie $\mathbf{0}$ lokalne minimum właściwe, więc również funkcja $f_{|M}$ ma w punkcie $\mathbf{0}$ lokalne minimum właściwe. Przyczyną tego pozornego paradoksu jest to, że wektory postaci $D^2\varphi(\mathbf{0})\mathbf{v}^2$ nie musi nie muszą być styczne do M w punkcie $\mathbf{p}$, więc ich obrazy przy $Df(\mathbf{0})$ nie muszą być zerowe. W przypadku funkcji Lagrange'a ta kwestia nie występuje, bo jej różniczka w punkcie $\mathbf{p}$ jest przekształceniem zerowym, funkcja Lagrange'a jest tak właśnie dobrana! ■

Zadanie 6.1 Znaleźć lokalne ekstrema oraz oba kresy funkcji $x^2 + y^2 + z^2$ na zbiorze zdefiniowanym równaniem $x^4 + \frac{1}{16}y^4 + \frac{1}{81}z^4 = 1$ (H.Cartan).

Twierdzenie 6.3 (o niemal jednostajnej zbieżności)
Załóżmy, że zbiór G jest otwarty i spójny. Niech (f_n) będzie ciągiem funkcji klasy C^1 określonych na G. Załóżmy, że ciąg (Df_n) jest zbieżny jednostajnie na każdym zbiorze zwartym $C \subset G$ do pewnej funkcji g oraz że istnieje punkt $\mathbf{p} \in G$ taki, że ciąg $(f_n(\mathbf{p}))$ jest zbieżny. Wtedy ciąg (f_n) jest zbieżny jednostajnie na każdym zbiorze zwartym do pewnej funkcji $f \in C^1(G)$ i zachodzi równość $Df = g$.

Dowód. (szkic) Jeśli C jest zbiorem zwartym wypukłym, to dowód tego twierdzenia jest powtórzeniem dowodu podanego w zeszłym roku w przypadku funkcji jednej zmiennej określonych na przedziale z jedną drobną różnicą: teraz twierdzenie o wartości średniej to nierówność, więc trzeba dokonać kosmetycznych zmian w oszacowaniach, by pasowały do wielowymiarowej wersji twierdzenia o wartości średniej. Następnie należy skorzystać z tego, że każde dwa punkty zbioru otwartego i spójnego można połączyć łamaną w nim zawartą, taką łamaną można pokryć skończoną liczbą kul otwartych, których domknięcia są zawarte w zbiorze G, ponumerować je tak, by pierwsza zawierała punkt $\mathbf{p}$, druga – przecinała pierwszą, trzecia – drugą itd. Następnie z tego, że twierdzenie jest prawdziwe w przypadku zbioru zwartego wypukłego wywnioskować tezę dla dowolnej łamanej zawartej w G zaczynającej się w punkcie $\mathbf{p}$, a stąd już bez trudności da się uzyskać tezę twierdzenia. ■

Jest jasne, że jeśli założymy, że funkcje $f_1, f_2, \ldots$ są klasy C^m oraz ciąg $(D^m f_n)$ jest jednostajnie zbieżny na każdym zbiorze zwartym zawartym w zbiorze G oraz że dla $j = 0, 1, \ldots, m-1$ ciąg $(D^j f_n)(\mathbf{p})$ jest zbieżny w pewnym punkcie $\mathbf{p} \in G$, to okaże się, że ciąg (f_n) jest jednostajnie zbieżny na każdym zbiorze zwartym zawartym w G oraz że funkcja graniczna f jest klasy C^m i $\lim\limits_{n\to\infty} D^j f_n = D^j f$ przy czym zbieżność jest jednostajna na zwartych podzbiorach zbioru G.

Przypomnijmy, że na analizie I wykazaliśmy, ze funkcja α zdefiniowana wzorami $\alpha(t) = 0$ dla $t \leqslant 0$ i $\alpha(t) = e^{-1/t}$ dla $t > 0$ jest funkcją klasy C^∞ na całej prostej. Wynika stąd, że funkcja β zdefiniowana wzorem $\alpha(1 - \|\mathbf{x}\|^2)$ jest klasy C^∞ na całej przestrzeni przy czym na kuli otwartej $B(\mathbf{0}, 1)$ przyjmuje wartości dodatnie a poza kulą otwartą $B(\mathbf{0}, 1)$ jest równa 0. Załóżmy, że $C \subseteq \mathbb{R}^k$ jest zbiorem domkniętym. Niech $G = \mathbb{R}^k \setminus C$. Zbiór G jest otwarty, więc jest sumą kul otwartych. Z tej rodziny kul można wybrać rodzinę przeliczalną $\{B(\mathbf{p}_n, r_n)\}$, której suma równa jest G. Definiujemy funkcję $f(\mathbf{x}) = \sum_n \varepsilon_n \beta\big(\frac{\mathbf{x}-\mathbf{p}_n}{r_n}\big)$, przy czym liczby dodatnie $\varepsilon_1, \varepsilon_2, \ldots$ są tak małe, że $\varepsilon_n \sup\limits_{j,\mathbf{x}} \|D^j(\mathbf{x})\| \leqslant 2^{-n}$ dla $j = 0, 1, \ldots, n$. Oczywiście oznacza to nałożenie na każdą z liczb $\varepsilon_1, \varepsilon_2, \ldots$ skończenie wielu warunków, zatem można je tak dobrać, że postulowane nierówności będą zachodzić w całej przestrzeni. Oznacza to, że szeregi $\sum_n \varepsilon_n D^j \beta\big(\frac{\mathbf{x}-\mathbf{p}_n}{r_n}\big)$ są zbieżne jednostajnie w całej przestrzeni $\mathbb{R}^k$ dla $j = 0, 1, 2, \ldots$. Wobec tego funkcja f jest klasy C^∞. Jest ona dodatnia poza zbiorem domkniętym C, a na zbiorze C jest tożsamościowo równa 0. Wykazaliśmy więc

Twierdzenie 6.4 (o najpaskudniejszej poziomicy)
Dla każdego zbioru domkniętego C istnieje funkcja f klasy C^∞ taka, że $C = f^{-1}(0)$. ■

Dodajmy, że wielu matematyków usiłuje opisać poziomice „typowych” funkcji klasy C^∞. Wiele przypadków już opisano, ale jest wysoce prawdopodobne, że badania te jeszcze przez wiele lat będą dostarczać rozrywki matematykom. Tematyka jest ważna również dzięki temu, że osiągnięte wyniki zazwyczaj znajdują zastosowanie również poza matematyką.

Uwaga 6.5 (o lokalnych ekstremach warunkowych)
Ten temat interesuje z różnych przyczyn ekonomistów. Omówimy teraz twierdzenie, które pojawia się w książce „Foundation of Economics Analysis”, 1947, P.A.Samuelsona (nagroda Nobla z ekonomii, 1970) z błędem poprawionym w 1952 w pracy G.Debreu (nagroda Nobla z ekonomii, 1983). Twierdzenie nie jest specjalnie trudne, a informacje historyczne służą jedynie podkreśleniu jego wagi w ekonomii, na której autor tego tekstu zna się tak jak wszyscy w RP (z wyjątkiem ekonomistów z prawdziwego zdarzenia). Ten fragment tekstu oparty jest na pracy G.Debreu.

Przypomnimy najpierw najbardziej podstawowych własności form kwadratowych. Niech $A = (a_{i,j})$ będzie macierzą symetryczną wymiaru k, tzn. $a_{i,j} = a_{j,i}$. Wtedy funkcja Q zdefiniowana wzorem $Q(\mathbf{x}) = A\mathbf{x} \cdot \mathbf{x}$ nazywana jest formą kwadratową. Niech $\mathbf{x} = D\mathbf{y}$ dla pewnej macierzy nieosobliwej D (D jest macierzą izomorfizmu). Wtedy $Q(D\mathbf{y}) = AD\mathbf{y} \cdot D\mathbf{y} = D^T AD\mathbf{y} \cdot \mathbf{y} = \tilde{Q}(\mathbf{y})$ też jest formą kwadratową, ale zmiennej $\mathbf{y}$. Formy Q i $\tilde{Q}$ są równoważne — to definicja.

W dalszym ciągu macierz A jest symetryczna. Funkcja Q na sferze jednostkowej osiąga swe kresy. Niech $m = \inf_{|\mathbf{x}|=1} Q(\mathbf{x})$. Istnieje punkt $\mathbf{p}$ taki, że $m = Q(\mathbf{p})$ i $\|\mathbf{p}\| = 1$. Na mocy twierdzenia Lagrange'a o ekstremach warunkowych istnieje taka liczba λ, że grad $Q(\mathbf{x}) = \lambda \operatorname{grad}(\|\mathbf{x}\|^2) = 2\lambda\mathbf{x}$. Dzięki symetrii macierzy A mamy również grad $Q(\mathbf{x}) = 2A\mathbf{x}$. Wobec tego $A\mathbf{p} = \lambda\mathbf{p}$. Stąd wynika, że $m = Q(\mathbf{p}) = A\mathbf{p} \cdot \mathbf{p} = \lambda$. Wykazaliśmy więc, że macierz A ma co najmniej jedną wartość własną rzeczywistą oraz że najmniejsza wartość formy kwadratowej Q przyjmowana w punktach sfery jednostkowej o środku w punkcie $\mathbf{0}$ jest jej wartością własną.

Załóżmy teraz, że λ_1 jest wartością własną macierzy A a $\mathbf{v}_1$ jest odpowiadającym jej wektorem własnym, tzn. $A\mathbf{v}_1 = \lambda_1\mathbf{v}_1$, $\mathbf{v}_1 \neq \mathbf{0}$. Jeśli $\mathbf{w}$ jest wektorem prostopadłym do wektora $\mathbf{v}_1$, to zachodzą równości $A\mathbf{w}\cdot\mathbf{v}_1 = \mathbf{w}\cdot A\mathbf{v}_1 = \mathbf{w}\cdot(\lambda_1\mathbf{v}_1) = \lambda_1\mathbf{w}\cdot\mathbf{v}_1 = 0$, zatem również wektor $A\mathbf{w}$ jest prostopadły do wektora $\mathbf{v}_1$. Niech V oznacza zbiór wszystkich wektorów prostopadłych do wektora $\mathbf{v}_1$. V jest podprzestrzenią liniową wymiaru $k-1$, niezmienniczą ze względu na A: $\mathbf{w} \in V \Rightarrow A\mathbf{w} \in V$. Rozumując dokładnie tak jak w przypadku całej przestrzeni przekonujemy się, że przekształcenie liniowe $A_{|V}$ ma rzeczywistą wartość własną λ_2, odpowiadający jej wektor własny $\mathbf{v}_2 \in V$ jest oczywiście prostopadły do wektora $\mathbf{v}_1$.

Teraz można zastosować to samo rozumowanie do zbioru złożonego ze wszystkich wektorów prostopadłych do obu wektorów $\mathbf{v}_1, \mathbf{v}_2$. Otrzymamy trzeci wektor własny prostopadły do $\mathbf{v}_1$ i do $\mathbf{v}_2$. Prowadzi do do bazy złożonej z wzajemnie prostopadłych wektorów własnych. Wykazaliśmy więc, że wartości własne macierzy symetrycznej są rzeczywiste i że istnieje baza złożona z wzajemnie prostopadłych wektorów własnych, w szczególności macierz symetryczna jest diagonalizowalna. Niech V_+, V_0 i V_- oznaczają podprzestrzenie liniowe niezmiennicze odpowiadające wartościom dodatnim własnym macierzy A, – zerowej wartości własnej macierzy A i wartościom ujemnym. Na $V_+\setminus\{\mathbf{0}\}$ forma Q przyjmuje wartości dodatnie, na V_0 jest tożsamościowo równa 0, na $V_-\setminus\{\mathbf{0}\}$ – wartości ujemne.

W szczególności: *macierz symetryczna A jest dodatnio określona wtedy i tylko wtedy, gdy jej wszystkie jej wartości własne są dodatnie.*

Zastępując formę Q wyznaczoną przez macierz A równoważną formą $\tilde{Q}$ wyznaczoną przez macierz D^TAD, D — macierz o wyznaczniku $\neq 0$, stwierdzamy, że wymiary analogicznie zdefiniowanych podprzestrzeni $\tilde{V}_+$, $\tilde{V}_0$ i $\tilde{V}_-$ są takie same jak w przypadku macierzy A, chociaż wartości własne mogą być inne — wynika to stąd, że V_+ jest maksymalną podprzestrzenią liniową, na której forma Q jest dodatnio określona, V_- to maksymalna podprzestrzeń liniowa, na której forma Q jest ujemnie określona, zaś V_0 to maksymalna podprzestrzeń liniowa, na której forma Q jest zerowa. Zastępując macierz A macierzą D^TAD mamy odpowiedni rozkład na podprzestrzenie $D^{-1}V_+$, $D^{-1}V_-$ i $D^{-1}V_0$, które mogą nie być niezmiennicze dla przekształcenia li11niowego zdefiniowanego za pomocą macierzy D^TAD. Oznacza to, że rozkład $\mathbb{R}^k$ na sumę prostą podprzestrzeni $\tilde{V}_+$, $\tilde{V}_0$ i $\tilde{V}_-$ określonych jako podprzestrzenie, na których forma kwadratowa jest dodatnio określona, zerowa, ujemnie określona nie jest jednoznaczny (konkretne przykłady ci studenci, którzy nie zdają sobie sprawy z tego powinni wymyśleć sami, bo to bardzo proste).

Niech $B = (b_{i,j})$ będzie macierzą kwadratową wymiaru k. Niech B_r oznacza dla $r = 1, 2, \ldots, k$ macierz wymiaru r znajdującą się w lewym górnym rogu macierzy B, np. $B_2 = \begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{pmatrix}$, $|B|$ oznacza wyznacznik macierzy kwadratowej B, $|B_0| = 1$. Przez y_r oznaczamy funkcję liniową zmiennych $x_r, x_{r+1}, \ldots, x_k$ postaci $x_r + d_{r+1}x_{r+1} + \cdots + d_kx_k$.

Twierdzenie 6.6 (o postaci kanonicznej niektórych form kwadratowych) Niech $A = (a_{i,j})$ będzie macierzą symetryczną wymiaru k, tzn. $a_{i,j} = a_{j,i}$. W tej sytuacji wzór $A\mathbf{x}\cdot\mathbf{x} = \sum_{r=1}^{k} c_r y_r^2$, $c_r \neq 0$ zachodzi wtedy i tylko wtedy, gdy $|A_r| \neq 0$ dla $r = 1, 2, \ldots, k$. Mamy wtedy $c_r = \frac{|A_r|}{|A_{r-1}|}$.[1]

Dowód. Jeśli formę kwadratową $Q(\mathbf{x}) := A\mathbf{x}\cdot\mathbf{x}$ można zapisać w postaci $\sum_{r=1}^{k} c_r y_r^2$, $c_r \neq 0$, to oczywiście $a_{1,1} \neq 0$, bo $c_1 \neq 0$ a zmienna x_1 występuje tylko w y_1. Jeśli $|A_1| = a_{1,1} \neq 0$, to możemy napisać

$$Q(\mathbf{x}) = \sum a_{i,j}x_ix_j = a_{1,1}\big(x_1 + \tfrac{a_{1,2}}{a_{1,1}}x_2 + \tfrac{a_{1,3}}{a_{1,1}}x_3 + \cdots + \tfrac{a_{1,k}}{a_{1,1}}x_k\big)^2 + Q_2(x_2, x_3, \ldots, x_k),$$

gdzie przez $Q_2(\mathbf{x})$ oznaczyliśmy odpowiednią formę kwadratową $k-1$ zmiennych $x_2, x_3, \ldots, x_k$. Spróbujmy przekształcić nasza formę raz jeszcze, by zapisać ją w postaci $c_1y_1^2 + c_2y_2^2 + Q_3(x_3, \ldots, x_k)$. Zróżniczkujmy stronami równość

$$Q(\mathbf{x}) = \sum a_{i,j}x_ix_j = c_1y_1^2 + c_2y_2^2 + Q_3(x_3, \ldots, x_k)$$

względem x_1 i względem x_2. Otrzymujemy równości $\frac{1}{2}\frac{\partial Q}{\partial x_1}(\mathbf{x}) = \sum_{j=1}^{k} a_{1,j}x_j = c_1y_1\frac{\partial y_1}{\partial x_1}$ oraz $\frac{1}{2}\frac{\partial Q}{\partial x_2}(\mathbf{x}) = \sum_{j=1}^{k} a_{2,j}x_j = c_1y_1\frac{\partial y_1}{\partial x_2} + c_2y_2\frac{\partial y_2}{\partial x_2}$. Z wzoru $y_2 = x_2 + d_3x_3 + \cdots + d_kx_k$ wynika, że $\frac{\partial y_2}{\partial x_2} = 1$. Niech $y_1 = 0$ i $x_3 = x_4 = \cdots = x_k = 0$. Otrzymujemy więc równania $a_{1,1}x_1 + a_{1,2}x_2 = 0$ i $a_{2,1}x_1 + a_{2,2}x_2 = c_2x_2$. Podprzestrzeń opisana równaniami $y_1 = 0$ i $x_3 = x_4 = \cdots = x_k = 0$ ma oczywiście wymiar 1. Wobec tego układ dwóch równań $a_{1,1}x_1 + a_{1,2}x_2 = 0$, $a_{2,1}x_1 + (a_{2,2} - c_2)x_2 = 0$ ma niezerowe rozwiązanie, zatem jego wyznacznik jest równy 0, czyli

$$0 = \begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} - c_2 \end{vmatrix} = \begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{vmatrix} - \begin{vmatrix} a_{1,1} & 0 \\ a_{2,1} & c_2 \end{vmatrix} = |A_2| - c_2|A_1|.$$

Stąd wynika, że $c_2 = \frac{|A_2|}{|A_1|}$, zatem przy założeniu, że $|A_1| \neq 0$ stwierdzamy, że $c_2 = 0 \Leftrightarrow |A_2| = 0$. Kolej na c_3. Chcemy, aby była spełniona równość

$$Q(\mathbf{x}) = \sum a_{i,j}x_ix_j = c_1y_1^2 + c_2y_2^2 + c_3y_3^2 + Q_4(x_4, x_5, \ldots, x_k).$$

Różniczkując ją stronami względem x_1, x_2, x_3 otrzymujemy równości

$$\begin{aligned}
\tfrac{1}{2}\tfrac{\partial Q}{\partial x_1} &= \textstyle\sum_j a_{1,j}x_j = c_1y_1\tfrac{\partial y_1}{\partial x_1}\\
\tfrac{1}{2}\tfrac{\partial Q}{\partial x_2} &= \textstyle\sum_j a_{2,j}x_j = c_1y_1\tfrac{\partial y_1}{\partial x_2} + c_2y_2\tfrac{\partial y_2}{\partial x_2}\\
\tfrac{1}{2}\tfrac{\partial Q}{\partial x_3} &= \textstyle\sum_j a_{3,j}x_j = c_1y_1\tfrac{\partial y_1}{\partial x_3} + c_2y_2\tfrac{\partial y_2}{\partial x_3} + c_3y_3\tfrac{\partial y_3}{\partial x_3}
\end{aligned}$$

Przyjmijmy teraz $y_1 = y_2 = 0$, $x_4 = x_5 = \cdots = x_k = 0$. Te równania definiują jednowymiarową podprzestrzeń liniową w $\mathbb{R}^k$, zatem poniższy układ równań (*wiemy, że* $\frac{\partial y_1}{\partial x_3} = \frac{\partial y_2}{\partial x_3} = 0$, $\frac{\partial y_3}{\partial x_3} = 1$)

$$\begin{aligned}
a_{1,1}x_1 + a_{1,2}x_2 + a_{1,3}x_3 &= 0\\
a_{2,1}x_1 + a_{2,2}x_2 + a_{2,3}x_3 &= 0\\
a_{3,1}x_1 + a_{3,2}x_2 + a_{3,3}x_3 &= c_3x_3
\end{aligned}$$

[1] To nie jest ogólne twierdzenie o postaci kanonicznej, bo przekształcenie, za pomocą którego sprowadzamy formę kwadratową do postaci kanonicznej, ma szczególną postać, jasne jest też, że mowa jest jedynie o formach kwadratowych niezdegenerowanych

ma niezerowe rozwiązanie. Wobec tego jego wyznacznik równy jest 0, czyli

$$0 = \begin{vmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} - c_3 \end{vmatrix} = \begin{vmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{vmatrix} - \begin{vmatrix} a_{1,1} & a_{1,2} & 0 \\ a_{2,1} & a_{2,2} & 0 \\ a_{3,1} & a_{3,2} & c_3 \end{vmatrix} = |A_3| - c_3|A_2|\,.$$

Podobnie jak poprzednio jest oczywiste, że $c_3 = 0 \Leftrightarrow |A_3| = 0$. Tę procedurę można kontynuować. Dowód został zakończony. ■

Z twierdzenia tego wynika twierdzenie Sylvestera: macierz A jest dodatnio określona wtedy i tylko wtedy, gdy wszystkie wyznaczniki $|A_r|$ są dodatnie. Jasne jest też, że jest ujemnie określona wtedy i tylko wtedy, gdy $c_1, c_2, \ldots, c_k < 0$, czyli gdy $|A_1| < 0$, $|A_2| > 0$, $|A_3| < 0$, $|A_4| > 0$, ...

Zbliżamy się do głównej części tej opowieści. A jest w dalszym ciągu macierzą symetryczną, ale od tej pory wymiaru $k+l$. Zakładamy też, że B jest macierzą o l wierszach i $k+l$ kolumnach. Zajmować się będziemy dodatnią określonością formy kwadratowej Q, $Q(\mathbf{x}) = A\mathbf{x}\cdot\mathbf{x}$ ale na podprzestrzeni M zdefiniowanej równaniem $B\mathbf{x} = \mathbf{0}$, czyli układem l równań liniowych z k niewiadomymi. Chodzi o to, by warunek typu Sylvestera wyrazić w terminach macierzy A i B.

Lemat 6.7
Forma kwadratowa Q jest dodatnio określona na podprzestrzeni M wtedy i tylko wtedy, gdy istnieje liczba λ taka, że dla każdego $\mathbf{x} \neq \mathbf{0}$ zachodzi $Q(\mathbf{x}) + \lambda B\mathbf{x} \cdot B\mathbf{x} > 0$.
Dowód. Warunek wystarcza, bo jeśli $\mathbf{x} \in M$, to $0 < Q(\mathbf{x}) + \lambda B\mathbf{x} \cdot B\mathbf{x} = Q(\mathbf{x})$. Wykażemy, że jest również konieczny. Załóżmy więc, że $Q(\mathbf{x}) > 0$ dla każdego $\mathbf{x} \in M$. Ponieważ $B\mathbf{x} = \mathbf{0}$ i Q jest funkcją ciągłą, więc istnieje takie otoczenie U zbioru zwartego $\{\mathbf{x} \in \mathbb{R}^{k+l}\colon \quad \|x\| = 1, B\mathbf{x} = \mathbf{0}\}$, że jeśli $\mathbf{x} \in U$, to $Q(\mathbf{x}) > 0$. Funkcja ciągła określona zbiorze zwartym osiąga swój kres dolny, więc istnieje takie $\mathbf{x}_0$, że $\|\mathbf{x}_0\| = 1$ i $A\mathbf{x} \cdot \mathbf{x} \geqslant A\mathbf{x}_0 \cdot \mathbf{x}_0$ dla każdego $\mathbf{x}$, dla którego $\|\mathbf{x}\| = 1$. Z tego samego powodu istnieje taki punkt $\mathbf{x}_1$, że $\|\mathbf{x}_1\| = 1$ i $B\mathbf{x} \cdot B\mathbf{x} \geqslant B\mathbf{x}_1 \cdot B\mathbf{x}_1 > 0$ dla każdego $\mathbf{x} \notin U$, dla którego $\|\mathbf{x}\| = 1$. Teraz pozostaje wybrać $\lambda > 0$tak duże, by $\lambda B\mathbf{x}_1 \cdot B\mathbf{x}_1 + A\mathbf{x}_0 \cdot \mathbf{x}_0 > 0$, co oczywiście jest możliwe. Z określenia λ wynika od razu, że $Q(\mathbf{x}) + \lambda B\mathbf{x} \cdot B\mathbf{x} > 0$: w zbiorze U tak jest, bo pierwszy składnik jest dodatni, a drugi — nieujemny, poza U drugi składnik majoryzuje pierwszy. ■

Wniosek 6.8
Forma Q jest dodatnio określona na podprzestrzeni M, wtedy i tylko wtedy, gdy istnieje liczba λ^* taka, że dla każdej liczby $\lambda \geqslant \lambda^*$ i dla każdego $\mathbf{x} \neq \mathbf{0}$ zachodzi nierówność $Q(\mathbf{x}) + \lambda B\mathbf{x} \cdot B\mathbf{x} > 0$. ■

Lemat 6.9
$|A + \lambda B^T B|$ jest wielomianem zmiennej λ. którego współczynnik przy najwyższej potędze zmiennej (tzn. przy λ^l) równy jest $(-1)^l \begin{vmatrix} A & B^T \\ B & \mathbf{0}_l \end{vmatrix}$. (W tym miejscu $\mathbf{0}_l$ to macierz kwadratowa wymiaru l.)

Dowód. Z oczywistej równości

$$\begin{pmatrix} A & \lambda B^T \\ B & -I_l \end{pmatrix}\begin{pmatrix} I_{k+l} & \mathbf{0} \\ B & I_l \end{pmatrix} = \begin{pmatrix} A+\lambda B^T B & \lambda B^T \\ \mathbf{0} & -I_l \end{pmatrix}$$

wynika, że

$$\begin{vmatrix} A & \lambda B^T \\ B & -I_l \end{vmatrix} = (-1)^l \big|A+\lambda B^T B\big| .$$

Trzeba więc obliczyć współczynnik przy λ^l w wielomianie $\begin{vmatrix} A & \lambda B^T \\ B & -I_l \end{vmatrix}$. Ten współczynnik to wartość l–tej pochodnej tej funkcji podzielona przez $l!$. Pochodną wyznacznika możemy obliczyć np. zastępując jedną z $k+2l$ kolumn, kolumną złożoną z pochodnych funkcji występujących w tej kolumnie i sumując tych $k+2l$ składników; w rzeczywistości l składników, bo w $k+l$ kolumnach λ nie występuje. Różniczkując po raz drugi otrzymamy z każdego z l składników $l-1$ składników, bo teraz λ występuje tylko w $l-1$ kolumnach. W rezultacie po l różniczkowaniach otrzymamy $l!$ składników, każdy z nich równy $\begin{vmatrix} A & B^T \\ B & \mathbf{0} \end{vmatrix}$. Dowód został zakończony. ■

Twierdzenie 6.10 (o dodatniej określoności formy kwadratowej na podprzestrzeni)
Jeśli A jest macierzą symetryczną wymiaru $k+l$, B — macierzą o $k+l$ kolumnach i l wierszach, $\big|B_l\big| \neq 0$, to $A\mathbf{x}\cdot\mathbf{x} > 0$ dla każdego $\mathbf{x}$, dla którego $B\mathbf{x} = \mathbf{0}$, wtedy i tylko wtedy, gdy $(-1)^l\begin{vmatrix} A_r & (B_{l,r})^T \\ B_{l,r} & \mathbf{0} \end{vmatrix} > 0$ dla $l+1 \leqslant r \leqslant k+l$. $\big(B_{l,r} := (b_{i,j}\big)$, gdzie $1 \leqslant i \leqslant l$, $1 \leqslant j \leqslant r)$

Dowód. Wykażemy najpierw, że z dodatniej określoności na podprzestrzeni M złożonej z tych punktów $\mathbf{x}$, dla których $B\mathbf{x} = 0$ wynika, że k wyznaczników z tezy twierdzenia to liczby dodatnie.

Najpierw udowodnimy, że $\begin{vmatrix} A & B^T \\ B & \mathbf{0} \end{vmatrix} \neq 0$. Rozważmy dowolne punkty $\mathbf{x} \in \mathbb{R}^{k+l}$, $\mathbf{y} \in \mathbb{R}^l$ takie, że $A\mathbf{x} + B^T\mathbf{y} = \mathbf{0}$ i jednocześnie $B\mathbf{x} = \mathbf{0}$. Wtedy zachodzi równość

$$\mathbf{0} = A\mathbf{x}\cdot\mathbf{x} + B^T\mathbf{y}\cdot\mathbf{x} = A\mathbf{x}\cdot\mathbf{x} + \mathbf{y}\cdot B\mathbf{x} = A\mathbf{x}\cdot\mathbf{x}.$$

Wynika stąd, że $\mathbf{x} = \mathbf{0}$, zatem $B^T\mathbf{y} = \mathbf{0}$, co w świetle tego, że $|B_l| \neq 0$, oznacza, że $\mathbf{y} = \mathbf{0}$. Wobec tego jedynym rozwiązaniem układu $A\mathbf{x} + B^T\mathbf{y} = \mathbf{0}$, $B\mathbf{x} = \mathbf{0}$ jest rozwiązanie zerowe, a stąd wnioskujemy, że $\begin{vmatrix} A & B^T \\ B & \mathbf{0} \end{vmatrix} \neq 0$.

W taki sam sposób wykazujemy, że dla $r = l+1, l+2, \ldots, l+k-1$ zachodzi $\begin{vmatrix} A_r & (B_{l,r})^T \\ B_{l,r} & \mathbf{0} \end{vmatrix} \neq 0$ — rozpatrujemy po prostu takie wektory $\mathbf{x}$, że $0 = x_{r+1} = x_{r+2} = \ldots = x_{k+l}$.

Z wniosku 6.8 wynika, że dla dostatecznie dużych liczb λ macierz $A + \lambda\cdot B^T B$ jest dodatnio określona. Z twierdzenia Sylvestera wynika więc, że dla $r = 1, 2, \ldots, k+l$ wyznaczniki macierzy $\big(A+\lambda B^T B\big)_r = A_r + \lambda(B_{l,r})^T B_{l,r}$ są dodatnie. Wyznacznik $\big|A_r + \lambda\cdot(B_{l,r})^T B_{l,r}\big|$ jest dodatni dla $r = l+1, l+2, \ldots, l+k$ dla dostatecznie dużych liczb λ, a ponieważ jest to wielomian stopnia $r-l$, więc współczynnik przy λ^{r-l} jest dodatni. Stąd i z lematu 6.9 wynika więc, że $(-1)^l \begin{vmatrix} A_r & (B_{l,r})^T \\ B_{l,r} & \mathbf{0} \end{vmatrix} > 0$. Zakończyliśmy dowód pierwszej implikacji.

Teraz założymy, że wyznaczniki mają odpowiednie znaki i wykażemy, że forma zdefiniowana macierzą A jest dodatnio określona na podprzestrzeni zdefiniowanej równaniem $B\mathbf{x} = \mathbf{0}$. Wystarczy wykazać, że dla dostatecznie dużych λ macierz $A + \lambda B^T B$ jest dodatnio określona. Wykażemy, że dla dostatecznie dużych liczb λ wyznaczniki $\left|\big(A + \lambda B^T B\big)_r\right|$ są dodatnie dla $r = 1, 2, \ldots, k + l$.

Jest tak dla $r = 1, 2, \ldots, l$, bo wtedy macierz $\big(B^T B\big)_r$ jest dodatnio określona jako macierz Grama układu r liniowo niezależnych wektorów. Jej wyznacznik jest więc dodatni, zatem dla dostatecznie dużych $\lambda > 0$ wyznacznik macierzy $\big(A+\lambda B^T B\big)_r$ też jest dodatni (wyznacznik jest funkcją ciągłą macierzy). Dla $r > l$ jest to po prostu założenie. Stąd i z lematu 6.7 wynika, że macierz A jest dodatnio określona na podprzestrzeni M złożonej z tych wektorów $\mathbf{x}$, dla których $B\mathbf{x} = 0$. Dowód został zakończony. ■