WYDAWNICTWA POLITECHNIKI WARSZAWSKIEJ

WŁADYSŁAW FINDEISEN
JACEK SZYMANOWSKI
ANDRZEJ WIERZBICKI

METODY OBLICZENIOWE OPTYMALIZACJI

WARSZAWA 1972

SPIS TREŚCI

WSTĘP

Skrypt niniejszy ma za zadanie przedstawić metody obliczeniowe optymalizacji, w odróżnieniu od podstaw teoretycznych. Uznano, że użytkownicy tego skryptu w zasadzie znają teorię optymalizacji, lecz brak im umiejętności rozwiązywania zadań praktycznych. W rozwiązywaniu tym, jak można przewidzieć, rolę podstawową odgrywa elektroniczna technika obliczeniowa. Tym niemniej, raczej dla wygody Czytelnika niż w celu wykładowym, w skrypcie przedstawione są w skrócie najważniejsze rezultaty współczesnej teorii optymalizacji; podana jest również odpowiednia źródłowa literatura.

Skrypt opracowany był z myślą o studentach specjalności automatyka; nie oznacza to oczywiście, że przedstawione tu metody obliczeniowe są specyficzne dla automatyki. Zadania optymalizacji występują w wielu dziedzinach, a po ich matematycznym sformułowaniu przestaje być istotne, z jakiego konkretnego problemu powstały. Specyfika automatyki odbiła się na skrypcie w ten sposób, że bardzo mało uwagi poświęcono zadaniom programowania liniowego. Zagadnienia optymalizacji procesów produkcyjnych prowadzą bowiem zazwyczaj do zadań programowania nieliniowego lub zadań optymalizacji dynamicznej. W związku z tym skrypt niniejszy poświęcony jest w zasadzie metodom, służącym do rozwiązywania zadań następujących:

a. Znaleźć

$$\max_{\underline{x}} \left[f(\underline{x}) = f(x_1, x_2, \ldots, x_n) \right], \tag{1}$$

gdzie $f(\underline{x})$ jest funkcją n zmiennych $x_1, x_2, \ldots x_n$

- przy warunkach ubocznych ("ograniczeniach")

$$g_i(\underline{x}) \left\{ \leqslant , = , \geqslant \right\} b_i , \qquad i = 1, \ldots m , \tag{2}$$

gdzie $g_i(\underline{x})$ są funkcjami zmiennych $x_1, x_2, \ldots x_n$, a b_i danymi liczbami; rozwiązania tego zadania są liczbami $\hat{x}_1, \hat{x}_2, \ldots \hat{x}_n$.

b. Znaleźć

$$\max_{u(t)} \left\{ Q\left[\underline{x}(t),\ \underline{u}(t),\ t\right] = Q\left[x_1(t), \ldots x_n(t),\ u_1(t), \ldots u_r(t),\ t\right]\right\} \quad (3)$$

gdzie $Q\left[\underline{x}(t),\ \underline{u}(t),\ t\right]$ jest funkcjonałem określonym na funkcjach zmiennej niezależnej t, $t \in \left[t_1, t_2\right]$
- przy warunkach ubocznych ("więzach")

$$\dot{\underline{x}} = \underline{f}(\underline{x},\ \underline{u},\ t), \quad (4)$$

gdzie (4) oznacza układ n równań różniczkowych pierwszego rzędu, tzn. $\dot{\underline{x}}$ oznacza wektor złożony z pochodnych $\frac{dx_1}{dt}$, $\frac{dx_2}{dt}$, ... $\frac{dx_n}{dt}$; rozwiązania tego zadania są funkcjami zmiennej niezależnej, $\hat{u}_1(t)$, $\hat{u}_2(t)$, ... $\hat{u}_r(t)$, określonymi na przedział sterowania $t \in \left[t_1, t_2\right]$.

Zadania typu (a) należą do dziedziny tzw. programowania nieliniowego; nazywa się je też niekiedy zadaniami optymalizacji statycznej. Poświęcona im jest część I skryptu. Zadania typu (b) należą do dziedziny optymalizacji dynamicznej; w skrypcie zajmuje się nimi część II.

Jeżeli, z innej strony patrząc, rozpatrywać zadania optymalizacji sterowania procesami produkcyjnymi, to łatwo jest podzielić je na dwie grupy.
- zadania optymalizacji stanu ustalonego,
- zadania optymalizacji dynamicznej.

W grupie pierwszej znajdą się przypadki wtedy, gdy proces produkcyjny ma charakter procesu ciągłego, jak to ma np. miejsce w przypadku wytwarzania kwasu siarkowego z siarki czy amoniaku z gazu ziemnego. Optimum średniej wydajności, średniego kosztu przy zadanej wydajności czy innego podobnego wskaźnika otrzymuje się wówczas najczęściej w sytuacji, gdy bieg procesu jest stały w czasie. Pozwala to przyjąć, jako element rozwiązania optymalnego, że pochodne względem czasu wszelkich zmiennych są równe zeru. Jeżeli obiekt ma stałe skupione, czyli jego równania stanu są np. postaci (4), to założenie stanu ustalonego oznacza, że opis ten przyjmuje postać równań typu (2)

$$\underline{f}(\underline{x},\ \underline{u}) = \underline{0}. \quad (5)$$

Pisząc (5) założono dla uproszczenia, że równania (4) nie zależą od czasu. Związki (5) pokazują, że zadanie optymalizacji stanu ustalonego sprowadzi się do zadania programowania nieliniowego, czyli zadania typu (a).

Zadania optymalizacji dynamicznej procesów produkcyjnych powstają przede wszystkim wówczas, gdy proces jest prowadzony

jako tzw. proces cykliczny. Przykładem może być proces wytopu w piecu stalowniczym: w czasie trwania procesu zmienia się temperatura, skład chemiczny itd., aby w chwili zakończenia procesu osiągnąć pewne zadane wartości. Istotne są zatem równania różniczkowe, opisujące obiekt, a zadanie matematyczne przyjmuje postać typu (b).

W niektórych przypadkach zadania optymalizacji stanu ustalonego mogą przybrać postać zadań optymalizacji dynamicznej, tj. zadań typu (b). Na przykład, gdy w obiekcie o stałych rozłożonych zmienne x_i są funkcjami czasu t oraz długości l, równania stanu mogą przyjąć postać

$$\frac{\partial \underline{x}}{\partial t} = \underline{f}\left(\underline{x}, \frac{\partial \underline{x}}{\partial l}, \underline{u}\right) \tag{6}$$

gdzie $\underline{x} = \underline{x}(l, t)$, $\underline{u} = \underline{u}(l, t)$. W stanie ustalonym procesu równania (6) sprowadzają się do

$$\underline{f}\left(\underline{x}, \frac{d\underline{x}}{dl}, \underline{u}\right) = \underline{0}. \tag{7}$$

Równania (7) nie są różniczkowe względem zmiennej t, lecz zawierają pochodne względem zmiennej l. Jest to zatem zadanie optymalizacji z więzami różniczkowymi; jest ono typu (b), a rozwiązanie będzie w postaci $\hat{\underline{u}}(l)$, $\hat{\underline{x}}(l)$, czyli w postaci funkcji zmiennej l.

Część I

PROGRAMOWANIE NIELINIOWE I LINIOWE

1. Podstawowe sformułowania i definicje

1.1. Zadania programowania nieliniowego i liniowego

Zadaniem programowania nazwany będzie problem następujący: Znaleźć wektor $\hat{\underline{x}}$, który minimalizuje bądź maksymalizuje skalarną funkcję

$$F = f(\underline{x}), \tag{1}$$

spełniając równocześnie układ równań lub nierówności o postaci (A)

$$g_i(\underline{x}) \left\{ \leqslant , = , \geqslant \right\} b_i, \qquad i = 1, \ldots m, \tag{2}$$

przy czym $\underline{x}$ jest n-wymiarowym wektorem kolumnowym, złożonym z elementów $x_1, x_2, \ldots x_n$. Zakłada się przy tym, że znane są postacie analityczne funkcji (1) oraz zależności (2), jak również wartości stałych b_i. W wielu przypadkach powyższe sformułowanie uzupełnione jest dodatkowymi wymaganiami; np. żąda się, aby niektóre bądź wszystkie zmienne niezależne x_j były nieujemne (tzn. $x_j \geqslant 0$ dla $j = 1, \ldots n$), lub też, aby przyjmowały tylko określone wartości dyskretne. Wartość $\hat{\underline{x}}$ nazwano rozwiązaniem zadania optymalizacji (zadania programowania). Wartość $f(\hat{\underline{x}})$ nazwano ekstremum warunkowym funkcji $f(\underline{x})$.

Funkcja (1) będzie dalej nazywana "funkcją celu", natomiast zbiór zależności (2) - "zbiorem ograniczeń". W zbiorze tym w każdym z ograniczeń może występować tylko jeden ze znaków wymienionych w zależności (2). Liczba ograniczeń m może być dowolna, to znaczy m może zarówno być większe, mniejsze jak i równe n. W przypadku gdy $m = 0$ rozpatrywane przez nas zadanie sprowadza się do problemu optymalizacji funkcji wielu zmiennych bez ograniczeń. Ten szczególny przypadek ma bardzo duże znaczenie przy poszukiwaniu ekstremum metodami iteracyjnymi i zostanie omówiony oddzielnie.

Zadania programowania (optymalizacji statycznej) dzielą się na dwie podstawowe grupy:
- zadania programowania liniowego,
- zadania programowania nieliniowego.

Jeżeli funkcja celu (1) i zbiór ograniczeń (2) są liniowe tzn. można je przedstawić w postaci:

$$F = f(x_1, \ldots, x_n) = \sum_{j=1}^{n} c_j x_j \tag{3}$$

oraz

$$g_i(x_1, \ldots, x_n) = \sum_{j=1}^{n} a_{ij} x_j, \quad i = 1, \ldots, m, \tag{4}$$

przy czym stałe współczynniki a_{ij} i c_j są znane, to zadanie należy do programowania liniowego. W przytoczonym sformułowaniu warunki (4) uzupełnia się zwykle żądaniem, aby wszystkie zmienne były nieujemne, tzn.

$$x_j \geq 0 \quad j = 1, \ldots, n, \tag{5}$$

co znacznie ułatwia numeryczne rozwiązanie zadania. Zwróćmy uwagę, że wymaganie to nie zmniejsza w jakimkolwiek stopniu ogólności rozważań, gdyż każde zadanie, w którym zmienne x_j są nieograniczonego znaku, można łatwo sprowadzić do postaci (5). Tak więc, zadanie programowania liniowego polega na znalezieniu nieujemnego wektora $\underline{\hat{x}}$, który ekstremalizuje liniową funkcję

$$F = \sum_{j=1}^{n} c_j x_j$$

oraz równocześnie spełnia zbiór ograniczeń (B)

$$\sum_{j=1}^{n} a_{ij} x_j \left\{ \leq , = , \geq \right\} b_i, \quad i = 1, \ldots, m.$$

Problem ten po raz pierwszy został rozwiązany w 1947 roku przez G. Dantzinga, który opracował dla niego metodę Simpleks stosowaną z różnymi modyfikacjami do dnia dzisiejszego.

Wszystkie pozostałe zadania optymalizacji typu (1) i (2), które nie mają postaci (3) (4), zalicza się do programowania nieliniowego, wyodrębniając przy tym szereg szczególnych przypadków, różniących się pod względem metod ich rozwiązywania:

1. Funkcja celu F jest nieliniowa, ograniczenia są liniowe.

Zadanie programowania nieliniowego sprowadza się tu do wyznaczenia nieujemnego wektora $\underline{\hat{x}}$, który minimalizuje lub maksymalizuje nieliniową funkcję

$$F = f(x_1, \ldots, x_n)$$

oraz równocześnie spełnia zbiór ograniczeń liniowych (C)

$$\sum_{j=1}^{n} a_{ij} x_j \left\{ \leqslant, =, \geqslant \right\} b_i, \qquad i = 1, \ldots, m,$$

przy

$$x_j \geqslant 0, \quad j = 1, \ldots, n.$$

Zadanie to posiada dwa interesujące przypadki, upraszczające rozwiązanie. W pierwszym funkcja celu ma postać sumy n składników, z których każdy jest funkcją tylko jednej zmiennej x_j. Tak więc

$$F = f(x_1, \ldots, x_n) = f_1(x_1) + f_2(x_2) + \ldots + f_n(x_n) \qquad (C1)$$

Funkcja celu posiadająca powyższą własność nazywana jest funkcją addytywną.

W drugim przypadku, funkcja celu może być przedstawiona jako suma formy liniowej i formy kwadratowej, a mianowicie:

$$\begin{aligned} F = f(x_1, \ldots, x_n) &= \sum_{j=1}^{n} c_j x_j + \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} x_i x_j = \\ &= c_1 x_1 + \ldots + c_n x_n + d_{11} x_1^2 + d_{12} x_1 x_2 + \ldots + \\ &+ d_{1n} x_1 x_n + \ldots + d_{nn} x_n^2 . \end{aligned} \qquad (C2)$$

Tego rodzaju problem optymalizacji nazywa się zadaniem programowania kwadratowego.

2. Funkcja celu F oraz ograniczenia są nieliniowe, ale zakładamy ich addytywność. Oznacza to, że w problemie (A) zbiór ograniczeń (2) można przedstawić w postaci

$$g_i(x_1, \ldots, x_n) = g_{i1}(x_1) + \ldots + g_{in}(x_n), \quad i = 1, \ldots, m \qquad (D)$$

3. Funkcja celu F oraz ograniczenia są nieliniowe, ale ponadto zadanie ma następujące cechy: ograniczenia są tylko równościowe, wszystkie zmienne x_j są nieograniczonego znaku, $m < n$, oraz zarówno $g_i(\underline{x})$ jak $f(\underline{x})$ są ciągłe i posiadają przynajmniej drugie pochodne. W tym przypadku zadanie ma zatem postać: znaleźć wektor $\hat{\underline{x}}$, który ekstremalizuje funkcję

$$F = f(\underline{x}), \qquad (E)$$

przy warunkach

$$g_i(\underline{x}) = b_i; \quad i = 1, \ldots, m.$$

Ten typ zadania nosi nazwę "klasycznego problemu optymalizacji"; jego analityczne rozwiązanie zostało po raz pierwszy podane przez Lagrange'a.

Na zakończenie listy poszczególnych przypadków programowania nieliniowego warto wspomnieć o jeszcze dwóch jego rodzajach. Pierwszy z nich nosi nazwę programowania liniowego w liczbach całkowitych i definiowany jest tak jak problem programowania liniowego (B), z tym jednak, że zmienne x_j mogą przyjmować tylko wartości całkowite. Drugi natomiast nazywany jest programowaniem stochastycznym; jego celem jest rozwiązanie ogólnego problemu (A), lub jego przypadków szczególnych, w warunkach gdy parametry zadania (np. b_i) są zmiennymi przypadkowymi.

Wśród metod stosowanych do rozwiązywania zadań optymalizacji statycznej wyróżnić należy metody analityczne oraz metody numeryczne. Nie istnieją jednak ani metody analityczne, ani numeryczne, pozwalające w sposób ogólny rozwiązać problem (A). Udaje się to tylko w szczególnych przypadkach, których lista prawie w całości została przedstawiona. I tak, metody analityczne dostarczają ogólnego narzędzia dla rozwiązywania "klasycznego problemu optymalizacji" (E) oraz takiej jego modyfikacji, gdy obok ograniczeń równościowych występują również ograniczenia nierównościowe. Ten ogólniejszy przypadek został w 1951 roku rozwiązany przez Kuhna i Tuckera, którzy podali warunki konieczne i wystarczające rozwiązania optymalnego. Nie trudno jednak się przekonać, że w przypadku wielowymiarowego problemu rozwiązanie analityczne staje się skomplikowane. Dlatego też, przy rozwiązywaniu konkretnych problemów posługujemy się z zasady metodami numerycznymi, natomiast teoria Lagrange'a oraz Kuhna i Tuckera służy głównie jako narzędzie teoretyczne. Od szeregu lat szuka się coraz efektywniejszych numerycznych metod optymalizacji, opartych o zastosowanie maszyn cyfrowych. Jako reprezentacyjne metody można wymienić:

- metodę Simpleks, stosowaną do rozwiązywania zadań programowania liniowego (B),
- metodę Wolfe'a, mającą zastosowanie do rozwiązywania zadań programowania kwadratowego (C2),
- metody bezgradientowe minimalizacji funkcji wielu zmiennych bez ograniczeń, takie jak: Rosenbrocka, Neldera i Meada, Powella i inne,
- metody gradientowe minimalizacji funkcji wielu zmiennych bez ograniczeń, a więc: największego spadku, gradientu sprzężonego, Davidona itp.

- metody poszukiwania ekstremum funkcji celu z ograniczeniami nierównościowymi typu (2) przez wprowadzenie funkcji kary, takie jak: Rosenbrocka, Carrolla, Powella itp.

Zwróćmy uwagę, że w powyższym wykazie nie sformułowano dodatkowych wymagań jakie musi spełniać funkcja $f(\underline{x})$ oraz zbiór ograniczeń, by dana metoda była zbieżna. Wymagania te zostaną szczegółowo omówione przy rozpatrywaniu odpowiednich algorytmów, dlatego na razie je pominięto. Natomiast nasuwa się pytanie dlaczego wymieniono wiele metod poszukiwania ekstremum bez ograniczeń, kiedy interesują nas w zasadzie zadania z ograniczeniami. Wynika to z tej przyczyny, że przy numerycznym rozwiązywaniu zadania programowania nieliniowego można adaptować metody ekstremalizacji funkcji wielu zmiennych bez ograniczeń do przypadków z nierozwikłanymi ograniczeniami równościowymi.

Bardzo efektywną metodą poszukiwania ekstremum jest metoda Simpleks i jej modyfikacje. Metodę tę stosuje się jednak jedynie do problemów liniowych. Stąd też, mając do rozwiązania nieliniowy problem optymalizacji często usiłuje się sprowadzić go drogą aproksymacji bądź odpowiednich podstawień do problemu liniowego. Jeśli się to powiedzie, posługujemy się następnie programowaniem liniowym. Tego typu postępowanie stosujemy przy rozwiązywaniu problemów (C1) i (D), dlatego nie będziemy się nimi zajmować oddzielnie. Zrozumiałą jest rzeczą, że jeśli zadanie nie da się sprowadzić do liniowego, sięgnąć trzeba do innych metod numerycznych. Wyliczono je już powyżej - należy zwrócić tylko uwagę, że nie wspomniano dotąd o metodzie programowania dynamicznego, którą również można stosować do rozwiązywania niektórych zadań programowania nieliniowego. Zrobiono to celowo, bowiem dokładniej jest ona rozpatrzona przy omawianiu optymalizacji dynamicznej.

1.2. Ekstrema bezwarunkowe i warunkowe

Dla dalszego tekstu tego skryptu użyteczne będzie przytoczenie definicji ekstremów funkcji. Ekstrema dzielą się na minima i maxima, ekstrema globalne i lokalne oraz warunkowe i bezwarunkowe. Podamy następujące definicje:

Globalne maksimum

Funkcja $f(\underline{x})$ określona w domkniętym zbiorze X w E^n przyjmuje globalne maksimum w punkcie $\hat{\underline{x}} \in X$,

$$\text{jeśli} \quad f(\underline{x}) \leq f(\hat{\underline{x}}) \quad \text{dla wszystkich punktów} \quad \underline{x} \in X. \tag{6}$$

Globalne maksimum $f(\underline{x})$ nosi również nazwę absolutnego maksimum. Jeśli domknięty zbiór X jest zbiorem ograniczonym, wówczas globalne maksimum $f(\underline{x})$ w X występuje w jednym lub

w większej ilości punktów należących do X, pod warunkiem, że $f(\underline{x})$ jest ciągła względem zbioru X. Własność ta jest znana pod nazwą twierdzenia Weierstrassa. Jeśli natomiast zbiór X nie jest zbiorem ograniczonym, wówczas globalne maksimum może nie występować w żadnym punkcie X przy skończonym $|\underline{x}|$, bądź też może istnieć ograniczona wartość $f(\underline{x})$, gdy $|\underline{x}| \to \infty$ w określony sposób.

Silne lokalne maksimum

Zakłada się, że $f(\underline{x})$ jest określona we wszystkich punktach należących do pewnego otoczenia δ punktu $\hat{\underline{x}}$ w E^n. Funkcja $f(\underline{x})$ posiada silne lokalne maksimum w $\hat{\underline{x}}$, jeśli istnieje takie ε, $0 < \varepsilon < \delta$, że dla wszystkich $\underline{x}$ spełniających zależność

$$0 < |\underline{x} - \hat{\underline{x}}| < \varepsilon, \quad \text{zachodzi} \quad f(x) < f(\hat{x}). \tag{7}$$

Inaczej mówiąc, funkcja $f(\underline{x})$ posiada silne lokalne maksimum w $\hat{\underline{x}}$, jeśli istnieje takie otoczenie ε punktu $\hat{\underline{x}}$, że dla wszystkich $\underline{x}$ z tego otoczenia oraz różnych od $\hat{\underline{x}}$, $(\underline{x})$ jest ściśle mniejsza od $f(\hat{\underline{x}})$.

Słabe lokalne maksimum

Zakłada się, że $f(\underline{x})$ jest określona we wszystkich punktach należących do pewnego otoczenia δ punktu $\hat{\underline{x}}$ w E^n. Funkcja $f(\underline{x})$ posiada słabe lokalne maksimum w $\hat{\underline{x}}$, jeśli nie osiąga silnego lokalnego maksimum w $\hat{\underline{x}}$ oraz istnieje takie ε, $0 < \varepsilon < \delta$, że dla wszystkich $\underline{x}$ spełniających zależność

$$0 < |\underline{x} - \hat{\underline{x}}| < \varepsilon \quad \text{zachodzi} \quad f(\underline{x}) \leqslant f(\hat{\underline{x}}). \tag{8}$$

W większości przypadków nie zachodzi potrzeba rozróżnienia silnego maksimum lokalnego od słabego.

Definicje globalnego oraz lokalnych minimów są analogiczne do podanych powyżej z tym, że w przypadku minimum musimy zmienić kierunek znaków nierówności.

Definicje (6), (7), (8) odnoszą się do ekstremów bezwarunkowych, tj. do przypadków gdy zadaniu (1) nie towarzyszą ograniczenia (2). Jeśli ograniczenia takie są sformułowane, jako równości lub nierówności, to mówimy wówczas o ekstremach warunkowych. Odpowiednie definicje są następujące:

Globalne maksimum warunkowe

Funkcja $f(\underline{x})$ określona w domkniętym zbiorze X w E^n przyjmuje globalne maksimum warunkowe w punkcie $\hat{\underline{x}} \in X$ spośród $\underline{x}$ spełniających $g_i(\underline{x}) \{ \leqslant, =, \geqslant \} b_i$, $i = 1, \ldots, m$, czyli $\underline{x} \in Y$, jeśli

dla wszystkich $\underline{x} \in X \wedge Y$ spełniona jest zależność $f(\underline{x}) \leqslant f(\hat{\underline{x}})$. (9)

Silne lokalne maksimum warunkowe

Funkcja $f(\underline{x})$ przyjmuje silne lokalne maksimum warunkowe w punkcie $\hat{\underline{x}}$ spośród $\underline{x}$ spełniających $g_i(\underline{x}) \{\leqslant, =, \geqslant\} b_i$ $i = 1,\ldots,m$, jeśli istnieje takie otoczenie ε punktu $\hat{\underline{x}} \in Y$, przy czym $\varepsilon > 0$, że

dla wszystkich $\underline{x} \neq \hat{\underline{x}}$ oraz $\underline{x} \in Y$ z tego otoczenia, zachodzi $f(\underline{x}) < f(\hat{\underline{x}})$. (10)

Słabe lokalne maksimum warunkowe

Funkcja $f(\underline{x})$ przyjmuje słabe lokalne maksimum warunkowe w punkcie $\hat{\underline{x}}$ spośród $\underline{x}$ spełniających $g_i(\underline{x}) \{\leqslant, =, \geqslant\} b_i$, $i = 1,\ldots,m$, czyli $\underline{x} \in Y$, jeśli nie osiąga silnego lokalnego maksimum warunkowego w $\hat{\underline{x}}$ oraz jeśli istnieje takie otoczenie ε punktu $\hat{\underline{x}} \in Y$, przy czym $\varepsilon > 0$, że

dla wszystkich $\underline{x} \in Y$ z tego otoczenia, zachodzi $f(\underline{x}) \leqslant f(\hat{\underline{x}})$. (11)

Odpowiednie modyfikacje powyższych definicji w przypadku globalnego oraz lokalnych minimów warunkowych są oczywiste, więc nie będą odrębnie podawane.

1.3. Funkcje wypukłe i wklęsłe

W teorii programowania nieliniowego istotną rolę odgrywają własności wypukłości funkcji. Okazuje się bowiem, np. w metodzie Lagrange'a lub w metodzie Kuhna-Tuckera, że pewne warunki konieczne rozwiązania stają się zarazem warunkami wystarczającymi, jeżeli funkcje $f(\underline{x})$ oraz $g_i(\underline{x})$ są, w odpowiednich przypadkach, wypukłe bądź wklęsłe. Dla wygody zatem czytelnika przypomnimy teraz definicje wypukłości i wklęsłości funkcji. Poprzedzić je musimy definicją wypukłości zbioru w przestrzeni E^n.

Zbiór wypukły

Zbiór X jest zbiorem wypukłym w E^n, jeśli każdy odcinek łączący dwa dowolne elementy $\underline{x}_1$ i $\underline{x}_2 \in X$:

$$\underline{x} = \lambda \underline{x}_2 + (1 - \lambda)\underline{x}_1, \text{ dla wszystkich } \lambda, \quad 0 \leqslant \lambda \leqslant 1 \tag{12}$$

należy również do tego zbioru, tzn. $\underline{x} \in X$.

Przykładami dwuwymiarowych figur wypukłych są m.in. koło, półkole, elipsa, trójkąt itp. W przestrzeni trójwymiarowej jako przykłady brył wypukłych mogą służyć: kula, równoległościan, graniastosłup itp.

Funkcja wypukła

Funkcja $f(\underline{x})$ określona w wypukłym zbiorze X w E^n jest wypukła, jeśli dla dowolnych dwóch punktów $\underline{x}_1$ i $\underline{x}_2 \in X$ oraz

dla wszystkich λ, $0 < \lambda < 1$ spełniona jest zależność

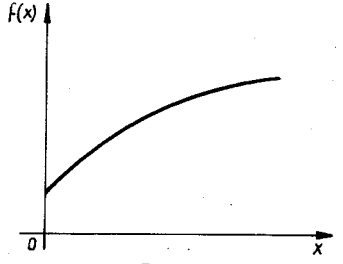
$$f\left[\lambda \underline{x}_2 + (1 - \lambda)\underline{x}_1\right] \leq \lambda f(\underline{x}_2) + (1 - \lambda) f(\underline{x}_1) \qquad (13)$$

Funkcja wklęsła

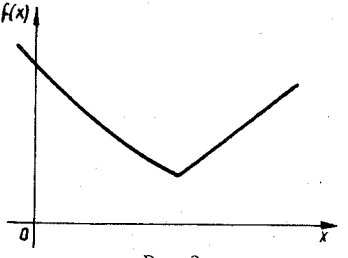
Funkcja $f(\underline{x})$ określona w wypukłym zbiorze X w E^n jest wklęsła, jeśli dla dowolnych dwóch punktów $\underline{x}_1$ i $\underline{x}_2 \in X$ oraz dla wszystkich λ, $0 \leq \lambda \leq 1$ spełniona jest zależność

$$f\left[\lambda \underline{x}_2 + (1 - \lambda)\underline{x}_1\right] \geq \lambda f(\underline{x}_2) + (1 - \lambda) f(\underline{x}_1). \qquad (14)$$

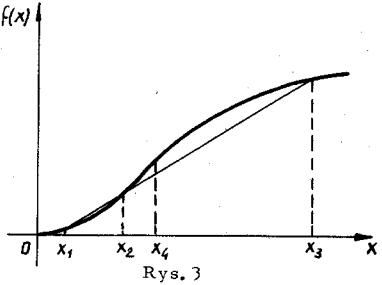
Inaczej mówiąc, hyperpowierzchnia $F = f(\underline{x})$ jest wklęsła, jeśli każdy odcinek łączący dwa dowolne punkty na powierzchni $[\underline{x}_1, F_1]$, $[\underline{x}_2, F_2]$ leży na lub poniżej tej powierzchni. Funkcję wklęsłą jednej zmiennej dla $x \geq 0$ przedstawiono na rys. 1. Podobnie możemy określić wypukłość funkcji, a więc: hyperpowierzchnia $F = f(\underline{x})$ jest wypukła, jeśli każdy odcinek łączący dwa dowolne punkty na powierzchni leży na lub powyżej tej powierzchni. Funkcję wypukłą jednej zmiennej przedstawiono na rys. 2.



Rys. 1



Rys. 2



Rys. 3

Zauważmy, że funkcja pokazana na rys. 3 nie jest ani wklęsła, ani wypukła, ponieważ odcinek łączący punkty $[x_1, f(x_1)]$ oraz $[x_3, f(x_3)]$ leży ponad $f(x)$ pomiędzy x_1 i x_2, natomiast poniżej $f(x)$ pomiędzy x_2 i x_3. Jednakże, funkcja ta jest wypukła w przedziale $0 \leq x \leq x_4$ oraz wklęsła w przedziale $x \geq x_4$.

Z przytoczonych definicji wynika, że jeśli $f(\underline{x})$ jest wypukła to $-f(\underline{x})$ jest wklęsła i odwrotnie. Jeśli w zależnościach (13) i (14) nierówności są ostre, przy czym $0 < \lambda < 1$, wówczas mówimy, że dana funkcja jest "ściśle" wypukła bądź wklęsła.

2. Metody analityczne programowania nieliniowego

Zadaniem niniejszego rozdziału jest przedstawienie podstaw i najważniejszych właściwości metod analitycznych, służących do rozwiązywania zadań programowania nieliniowego. Pomimo, że metody analityczne nie pozwalają na ogół na efektywne rozwiązanie zadań o dużej złożoności, to jednak dzięki opanowaniu tych metod i zbadaniu z ich pomocą różnych prostych przykładów, można potem - w zadaniach rzeczywistych o dużej złożoności - stosować odpowiednie metody numeryczne z większą świadomością.

W rozdziale niniejszym rozpatrywane będą kolejno: zadania bez ograniczeń, zadania z ograniczeniami równościowymi (rozwiązywane metodą Lagrange'a), zadania z ograniczeniami nierównościowymi (rozwiązywane metodą Kuhna-Tuckera) oraz rozszerzenia metody Kuhna-Tuckera na niektóre przypadki nieklasyczne. Zgodnie z charakterem tego skryptu, który dotyczy przede wszystkim metod obliczeniowych, nie będziemy w zasadzie podawać dowodów matematycznych, nastawiając się w większym stopniu na rozwiązywanie konkretnych zadań i problemów.

2.1. Zadanie bez ograniczeń

Przy braku ograniczeń zadanie programowania sprowadza się do poszukiwania ekstremum bezwarunkowego danej funkcji $f(\underline{x})$. Interesuje nas w zasadzie ekstremum (tj. maksimum lub minimum) globalne.

Klasyczna teoria optymalizacji nie dostarcza odpowiedniego narzędzia do określenia w sposób jednoznaczny czy dana funkcja posiada globalne-ekstremum, czy też nie. Formułuje ona jedynie warunki konieczne i dostateczne istnienia maksimum (minimum) lokalnego, zakładając przy tym, że badana funkcja jest klasy C^1 w całym interesującym nas obszarze.

Z teorii funkcji wielu zmiennych wiadomo, że warunkiem koniecznym istnienia maksimum (bądź minimum) lokalnego funkcji celu $F = f(\underline{x})$ w punkcie $\hat{\underline{x}}$ jest to, aby jej wszystkie pochodne cząstkowe w tym punkcie były równe zeru, a więc

$$\frac{\partial f(\hat{\underline{x}})}{\partial x_j} = 0; \qquad j = 1, \ldots, n. \tag{15}$$

Sformułowanie to jest równoważne stwierdzeniu, że:
1) gradient funkcji $f(\underline{x})$ w punkcie $\hat{\underline{x}}$ jest zerowy tzn.

$$\frac{\partial f(\underline{x})}{\partial \underline{x}} = \underline{0},$$

bądź też
2) płaszczyzna styczna do $F = f(\underline{x})$ w punkcie $\hat{\underline{x}}$ ma równanie

$$F_s(\underline{x}) = f(\underline{x}) = \text{const}.$$

Z powyższych równań wynika, że jeśli $f \in C^1$ oraz jeśli posiada ekstremum w punkcie $\hat{\underline{x}}$, to wówczas $\hat{\underline{x}}$ musi być rozwiązaniem układu n równań typu:

$$\frac{\partial f(\underline{x})}{\partial x_j} = 0, \quad j = 1, \ldots, n. \tag{16}$$

Podkreślić trzeba, że każdy punkt, który stanowi lokalne maksimum lub minimum funkcji $f(\underline{x})$, musi być rozwiązaniem tego układu równań. Jednakże nie każde rozwiązanie układu (16) będzie maksimum lub minimum lokalnym funkcji celu $f(\underline{x})$. Przykładem na to może być punkt przegięcia dla funkcji jednej zmiennej. Rozwiązania układu (16) nazywa się punktami stacjonarnymi funkcji $f(\underline{x})$. Tak więc, punkty w których funkcja $f(\underline{x})$ osiąga wartości ekstremalne są punktami stacjonarnymi $f(\underline{x})$, lecz odwrotne stwierdzenie nie jest prawdziwe.

Nie trudno wykazać, że punkt (lub punkty), w których $f(\underline{x})$ osiąga globalne maksimum muszą być również rozwiązaniami układu (16). Oznacza to, że jeśli znane są wszystkie rozwiązania układu (16), to wystarczy wybrać spośród nich takie rozwiązanie, dla którego funkcja celu przyjmuje wartość największą. Rozwiązanie to będzie szukanym globalnym ekstremum funkcji $f(\underline{x})$.

Niestety tego rodzaju tok postępowania w większości przypadków okazuje się zawodny, bowiem numeryczne wyznaczenie wszystkich rozwiązań układu (16) jest bardzo uciążliwe, szczególnie przy dużej wymiarowości problemu. Nie istnieje na razie ogólna procedura, która by umożliwiała w sposób jednoznaczny określenie wszystkich rozwiązań układu (16). Niekiedy nawet jeśli wiadomo, że istnieje tylko jedno rozwiązanie tego układu, to uzyskanie go nastręcza trudności. Stąd też, w praktycznych zastosowaniach staramy się unikać tego sposobu rozwiązywania zadań optymalizacji, pomimo że istnieje szereg efektywnych procedur iteracyjnych, które umożliwiają numeryczne rozwiązanie układu równań nieliniowych typu (16). Jako główne z tej grupy można wymienić: metodę Newtona-Raphsona i jej modyfikacje [48] oraz metodę Barnesa [2].

Metody te wymagają, aby badana funkcja celu posiadała drugie pochodne cząstkowe tzn. $f \in C^2$, a ponadto dosyć często nie zapewniają odpowiedniej zbieżności do szukanego rozwiązania układu (16). Metoda Newtona-Raphsona jest omówiona przy rozpatrywaniu optymalizacji dynamicznej, dlatego też należy tu poprzestać na tej krótkiej wzmiance.

W dotychczasowych rozważaniach opierano się jedynie na warunku koniecznym istnienia ekstremum lokalnego funkcji wielu zmiennych. Jak wspomniano, spełnienie tego warunku nie przesądza czy w ogóle istnieje ekstremum oraz czy jest nim maksimum, czy też minimum. Wiadomo, że ponadto sformułować można warunki dostateczne, które przesądzają wszystkie te wątpliwości. Warunki te mogą być podawane w różnej postaci. Jedna z nich może być następująca: warunkiem dostatecznym istnienia ekstremum funkcji wielu zmiennych jest to, aby - dla wartości zmiennych spełniających warunek konieczny istnienia ekstremum - macierz drugich pochodnych tej funkcji $\frac{\partial^2 f}{\partial \underline{x}^T \partial \underline{x}}$ była bądź dodatnio określona (w przypadku minimum), bądź ujemnie określona (w przypadku maksimum).

Można wykazać, że spełnienie warunków dostatecznych gwarantuje nam od razu, że rozwiązanie układu (16) jest jednocześnie silnym lokalnym minimum czy maksimum. Jednakże warunki te są interesujące głównie z punktu widzenia teoretycznego. Wynika to stąd, że kiedy problem jest wielowymiarowy, to dla rozwiązania układu (16) potrzebna jest już dość wielka liczba obliczeń numerycznych; gdyby jeszcze należało ją powiększyć przez sprawdzanie warunków dostatecznych, to zadanie stałoby się praktycznie nierozwiązalne.

Niekorzystne jest również to, że warunki dostateczne nie dostarczają informacji o tym, czy uzyskane rozwiązanie układu (16) stanowi ekstremum globalne, czy też lokalne.

Nieco inaczej przedstawia się sprawa wówczas, gdy funkcja $f(\underline{x})$ jest funkcją wypukłą (bądź wklęsłą). Mianowicie, przytoczymy bez dowodów (patrz [26]), że:

a. Jeśli funkcja $f(\underline{x})$ jest funkcją wypukłą określoną w domkniętym zbiorze wypukłym X w E^n, wtedy każde minimum lokalne $f(\underline{x})$ w X jest również globalnym minimum $f(\underline{x})$.

b. Zbiór punktów należących do X, w których $f(\underline{x})$ przyjmuje globalne minimum, jest zbiorem wypukłym. Stąd, jeśli globalne minimum występuje w dwóch różnych punktach, to znaczy, że występuje ono również w nieskończonej liczbie punktów. Ponadto, nie może być dwóch (lub więcej) punktów, w których $f(\underline{x})$ osiąga silne lokalne minimum.

c. Jeśli $f(\underline{x})$ jest funkcją ściśle wypukłą określoną w wypukłym zbiorze X, wtedy globalne minimum $f(\underline{x})$ występuje tylko w pojedynczym punkcie.

d. Jeśli $f(\underline{x})$ jest funkcją wypukłą określoną w zbiorze wypukłym X w E^n oraz $f \in C^1$, wtedy jeśli (16) jest spełnione w punkcie $\hat{\underline{x}}$ to punkt ten stanowi globalne minimum $f(\underline{x})$.

e. Jeśli zbiór X jest zbiorem domkniętym i ograniczonym oraz istnieje skończone globalne maksimum wypukłej funkcji $f(\underline{x})$ określonej w X, wtedy globalne maksimum $f(\underline{x})$ będzie występowało w punkcie (bądź punktach) na ograniczeniu zbioru X.

2.2. Zadanie z ograniczeniami równościowymi. Metoda Lagrange'a

Rozwiązanie klasycznego problemu optymalizacji polega na znalezieniu wektora (lub wektorów) $\hat{\underline{x}}$, który wyznacza globalne maksimum (minimum) funkcji celu

$$F = f(\underline{x}) = f(x_1, x_2, \ldots, x_n), \tag{17}$$

określonej w E^n, pod warunkiem spełnienia ograniczeń równościowych

$$g_i(\underline{x}) = b_i, \quad i = 1, \ldots, m, \text{ gdzie } m<n. \tag{18}$$

Zakłada się, przy tym, że $f \in C^1$ oraz wszystkie $g_i \in C^1$, $i = 1, \ldots, m$. Oznaczono przez Y zbiór punktów $\underline{x}$ spełniających równania (18). Zbiór ten nazywany jest niekiedy "zbiorem decyzji wewnętrznie zgodnych" lub "zbiorem decyzji dopuszczalnych". Zadanie polega więc na wybraniu decyzji optymalnej $\hat{\underline{x}}$ ze zbioru decyzji dopuszczalnych, $\underline{x} \in Y$, czyli na znalezieniu ekstremum warunkowego funkcji $f(\underline{x})$ (patrz definicje w punkcie 1.2).

W tzw. metodzie Lagrange'a formułuje się następujące warunki konieczne, by funkcja (17) osiągała ekstremum przy ograniczeniach (18):

$$\frac{\partial f(\hat{\underline{x}})}{\partial x_j} - \sum_{i=1}^{m} \hat{\lambda}_i \frac{\partial g_i(\hat{\underline{x}})}{\partial x_j} = 0, \quad j = 1, \ldots, n \tag{19}$$

$$g_i(\hat{\underline{x}}) = b_i, \quad i = 1, \ldots, m$$

Wprowadzając tzw. funkcję Lagrange'a

$$L(\underline{x}, \underline{\lambda}) = f(\underline{x}) + \underline{\lambda}^T \left[\underline{b} - \underline{g}(\underline{x})\right], \tag{20}$$

podane wyżej warunki można zapisać w postaci

$$\frac{\partial L(\hat{\underline{x}}, \hat{\underline{\lambda}})}{\partial x_j} = 0, \qquad j = 1, \ldots, n,$$

$$\frac{\partial L(\hat{\underline{x}}, \hat{\underline{\lambda}})}{\partial \hat{\lambda}_i} = 0, \qquad i = 1, \ldots, m. \tag{19'}$$

W równaniach (19) i (20) λ_i noszą nazwę mnożników Lagrange'a *).

Podkreślmy, że warunki (19) są warunkami koniecznymi: jeżeli $\hat{\underline{x}}$ jest rozwiązaniem zadania (17), (18), to jest ono również rozwiązaniem układu równań (19). Nie każde natomiast $\underline{x}$, które spełnia układ (19), stanowi rozwiązanie zadania (17), (18).

Ponadto, słuszność warunków koniecznych (19) jest zachowana oraz mnożniki $\hat{\lambda}_i$ są jednoznacznie określone, jeżeli rząd pewnej macierzy G w punkcie $\hat{\underline{x}}$ wynosi m, r(G) = m, przy czym macierz G utworzona jest w sposób następujący:

$$G = \left[\frac{\partial g_i(\hat{\underline{x}})}{\partial x_j}\right] = \begin{bmatrix} \frac{\partial g_1(\hat{\underline{x}})}{\partial x_1} & \cdots & \frac{\partial g_1(\hat{\underline{x}})}{\partial x_n} \\ \cdots & & \cdots \\ \frac{\partial g_m(\hat{\underline{x}})}{\partial x_1} & & \frac{\partial g_m(\hat{\underline{x}})}{\partial x_n} \end{bmatrix}$$

Jeżeli $r(G) \neq m$, rozwiązanie zadania może nie spełniać układu równań (19). Przypadki takie omawiamy dalej.

Sens wymagania $r[G] = m$ jest taki, że wówczas układ m równań $g_i(\underline{x}) = b_i$ wyznacza m spośród niewiadomych x_j jako jednoznaczne funkcje pozostałych n-m niewiadomych x_j. Pierwsze n równań w układzie (19) ma wówczas n-m niewiadomych x_j oraz m niewiadomych λ_i.

Praktyczne wykorzystanie warunków (19) polega na znalezieniu wszystkich $\hat{\underline{x}}$, które je spełniają, a następnie obliczeniu wartości funkcji celu $f(\underline{x})$ w tych punktach. Przeglądając te wartości znajduje się globalne maksimum bądź minimum, czyli właściwe rozwiązanie zadania (17), (18), jeżeli tylko to właściwe rozwiązanie $\hat{\underline{x}} \neq \infty$.

---

*) Można wykazać, że $\hat{\lambda}_i = \frac{\partial f(\hat{\underline{x}})}{\partial b_i}$, tzn. mnożnik Lagrange'a w punkcie spełniającym warunki (19) jest równy pochodnej funkcji celu względem parametru b_i w ograniczeniu. Ma to interpretację ekonomiczną.

Można określić przypadki, gdy warunki (19) są warunkami zarazem koniecznymi i wystarczającymi globalnego ekstremum warunkowego:

a) jeżeli $f(\underline{x})$ jest funkcją wypukłą, a ograniczenia $g_i(\underline{x})$ są liniowe, to $f(\underline{\hat{x}}) \leq f(\underline{x})$ dla wszystkich $\underline{x} \in Y$ (globalne minimum) wtedy i tylko wtedy, gdy $\underline{\hat{x}}$ spełnia układ (19);

b) jeżeli $f(\underline{x})$ jest funkcją wklęsłą, a ograniczenia $g_i(\underline{x})$ są liniowe, to $f(\underline{\hat{x}}) \geq f(\underline{x})$ dla wszystkich $\underline{x} \in Y$ (globalne maksimum) wtedy i tylko wtedy, gdy $\underline{\hat{x}}$ spełnia układ (19).

Ażeby sobie lepiej zdać sprawę z istoty metody mnożników Lagrange'a rozpatrzmy zdanie o dwóch zmiennych:

$$\text{znaleźć} \quad \max f(x_1, x_2),$$

przy warunku

$$g(x_1, x_2) = b,$$

przy czym zakłada się, że $f, g \in C^1$. Założono, że albo $\frac{\partial g(\underline{\hat{x}})}{\partial x_1}$ albo $\frac{\partial g(\underline{\hat{x}})}{\partial x_2}$ nie znika w punkcie $\left[\hat{x}_1, \hat{x}_2\right]$. Niech dla ustalenia uwagi, $\frac{\partial g(\underline{\hat{x}})}{\partial x_2} \neq 0$. Wówczas, na postawie tzw. twierdzenia o funkcjach niejawnych (patrz np. [26]) istnieje takie otoczenie ε punktu $\left[\hat{x}_1, \hat{x}_2\right]$, dla którego równanie

$$g(x_1, x_2) - b = 0,$$

określa zależność x_2 od x_1:

$$x_2 = \varphi(x_1).$$

Funkcja celu może być teraz zapisana jako zależna od jednej zmiennej

$$f(x_1, x_2) = f\left[x_1, \varphi(x_1)\right] = h(x_1),$$

bez dodatkowych ograniczeń, czyli rozwiązanie x_1 wyznacza się z warunku $\frac{dh(x_1)}{dx_1} = 0$.

Z twierdzenia o funkcjach niejawnych wynika, że jeżeli $g(x_1, x_2) \in C^1$, to także $\varphi(x_1) \in C^1$; ponieważ założono $f(x_1, x_2) \in C^1$, różniczkowalna jest również funkcja $h(x_1)$.

Interesująca nas jej pochodna będzie równa

$$\frac{d\,h}{d\,x_1} = \frac{\partial f}{\partial x_1} + \frac{\partial f}{\partial x_2} \cdot \frac{d\varphi}{dx_1}.$$

Własności funkcji niejawnych pozwalają napisać

$$\frac{d\varphi}{dx_1} = -\frac{\frac{\partial g}{\partial x_1}}{\frac{\partial g}{\partial x_2}} .$$

Zatem warunek konieczny rozwiązania $\hat{x}_1$ napiszemy jako

$$\frac{\partial f(\hat{\underline{x}})}{\partial x_1} - \frac{\frac{\partial f(\hat{\underline{x}})}{\partial x_2}}{\frac{\partial g(\hat{\underline{x}})}{\partial x_2}} \cdot \frac{\partial g(\hat{\underline{x}})}{\partial x_1} = 0. \tag{21}$$

Równanie (21) ma dwie niewiadome, $\hat{x}_1$ oraz $\hat{x}_2$. Ażeby je wyznaczyć, trzeba razem z (21) użyć równania ograniczeń z zadania

$$g(\hat{x}_1, \hat{x}_2) = b. \tag{22}$$

Jeżeli w równaniu (21) oznaczyć występujący tam iloraz pochodnych cząstkowych względem x_2 przez $\hat{\lambda}$

$$\frac{\frac{\partial f(\hat{\underline{x}})}{\partial x_2}}{\frac{\partial g(\hat{\underline{x}})}{\partial x_2}} = \hat{\lambda},$$

oraz zapisać to w postaci równania

$$\frac{\partial f(\hat{\underline{x}})}{\partial x_2} - \hat{\lambda} \frac{\partial g(\hat{\underline{x}})}{\partial x_2} = 0, \tag{23}$$

to otrzymany układ (21), (22), (23) będzie identyczny z układem równań metody mnożników Lagrange'a

$$\frac{\partial}{\partial x_1}\left[f(\hat{\underline{x}}) + \hat{\lambda}(b - g(\hat{\underline{x}}))\right] = 0,$$

$$\frac{\partial}{\partial x_2}\left[f(\hat{\underline{x}}) + \hat{\lambda}(b - g(\hat{\underline{x}}))\right] = 0,$$

$$\frac{\partial}{\partial \lambda}\left[f(\hat{\underline{x}}) + \hat{\lambda}(b - g(\hat{\underline{x}}))\right] = 0.$$

To, że warunki nasze są tylko konieczne wynika stąd, że $\frac{dh}{dx_1} = 0$ było tylko koniecznym warunkiem ekstremum funkcji jednej zmiennej $f[x_1, \varphi(x_1)] = h(x_1)$.

Rozpatrzmy teraz ogólniejsze sformułowanie warunków Lagrange'a (por. wzory 19), obejmujące przypadki gdy $r[G] \neq m$. Uogólnione twierdzenie o warunkach koniecznych rozwiązania zadania programowania z ograniczeniami równościowymi brzmi:

Jeśli $\hat{\underline{x}}$ jest warunkowym maksimum lub minimum funkcji $f(\underline{x})$ dla $\underline{x} \in Y$, to musi ono spełniać układ równań

$$\hat{\lambda}_o \frac{\partial f(\hat{\underline{x}})}{\partial x_j} - \sum_{i=1}^{m} \hat{\lambda}_i \frac{\partial g_i(\hat{\underline{x}})}{\partial x_j} = 0, \qquad j = 1, \ldots n \tag{24}$$

dla przynajmniej jednego zbioru $\hat{\lambda}_i$, $i = 0, \ldots, m$, w którym nie wszystkie $\hat{\lambda}_i = 0$, oraz musi spełniać równania

$$g_i(\hat{\underline{x}}) = b_i, \qquad i = 1, \ldots m . \tag{25}$$

Powyższe warunki konieczne nie mogą być spełnione, jeśli rząd macierzy współczynników równań (24), traktowanych jako układ równań z niewiadomymi $\hat{\lambda}_i$, wynosi $m + 1$, gdyż wówczas wszystkie $\hat{\lambda}_i = 0$. Rząd ten może być zatem co najwyżej równy m. W tej zaś sytuacji nie może być jednoznacznego rozwiązania na wszystkie $\hat{\lambda}_i$, $i = 0, \ldots m$.

Praktycznie zaleca się przyjąć $\hat{\lambda}_o = 1$; wówczas możliwe są następujące przypadki:

a) pozostałe $\hat{\lambda}_i$, $i = 1, \ldots m$ są określone jednoznacznie,

b) pozostałe $\hat{\lambda}_i$, $i = 1, \ldots m$ są określone niejednoznacznie,

c) dla pozostałych $\hat{\lambda}_i$, $i = 1, \ldots m$ układ nie ma rozwiązania.

Rozpatrzmy macierze złożone ze współczynników równań (24)

$$G = \begin{bmatrix} \frac{\partial g_1(\hat{\underline{x}})}{\partial x_1} & \cdots & \frac{\partial g_1(\hat{\underline{x}})}{x_n} \\ \cdots & & \cdots \\ \frac{\partial g_m(\hat{\underline{x}})}{\partial x_1} & \cdots & \frac{\partial g_m(\hat{\underline{x}})}{\partial x_n} \end{bmatrix}; \qquad G_f = \begin{bmatrix} \frac{\partial g_1(\hat{\underline{x}})}{x_1} & \cdots & \frac{\partial g_1(\hat{\underline{x}})}{x_n} \\ \cdots & & \cdots \\ \frac{\partial g_m(\hat{\underline{x}})}{x_1} & \cdots & \frac{\partial g_m(\hat{\underline{x}})}{x_n} \\ \frac{\partial f(\hat{\underline{x}})}{\partial x_1} & \cdots & \frac{\partial f(\hat{\underline{x}})}{\partial x_n} \end{bmatrix}$$

Wymieniony wyżej przypadek (a) wystąpi, gdy

$$r\left[G_f\right] = r\left[G\right] = m.$$

Przypadek (b) wystąpi, gdy

$$f\left[G_f\right] = r\left[G\right] < m.$$

Przypadek (c) wystąpi, gdy

$$r\left[G_f\right] > r\left[G\right].$$

W przypadku (c) nie można stwierdzić spełnienia warunków koniecznych, które mówią o istnieniu zbioru $\hat{\lambda}_i$, $i = 0,\ldots m$, w którym nie wszystkie $\hat{\lambda}_i$ są równe zeru. Stwierdziwszy sytuację (c), należy założyć $\hat{\lambda}_o = 0$ i badać te punkty $\hat{\underline{x}}$, w których układ równań (24) ma rozwiązanie, niejednoznaczne ale takie, że nie wszystkie pozostałe $\hat{\lambda}_i$ są równe zeru.

Trzeba zauważyć, że w przypadkach (a) oraz (b) - wobec $\hat{\lambda}_o \neq 0$, pozostałe $\hat{\lambda}_i$ mogą być równe zeru. Widać również, że przypadek (a) odpowiada podstawowemu (szczególnemu) sformułowaniu metody Lagrange'a, patrz wzory (19).

Widać, że w przypadku (c) musi być $r\left[G\right] < m$, gdyż $r\left[G_f\right] \leqslant m$.

Dla scharakteryzowania omawianych przypadków rozpatrzono trzy proste przykłady.

Przykład 1

Znaleźć $\min\left[f(\underline{x}) = x_1^2 + x_2^2 + x_3^2\right]$, przy

$$x_1 + x_2 + x_3 = 2,$$

$$x_1 - x_2 + 2x_3 = 3.$$

Funkcja Lagrange'a

$$L = x_1^2 + x_2^2 + x_3^2 + \lambda_1(2 - x_1 - x_2 - x_3) + \lambda_2(3 - x_1 + x_2 - 2x_3).$$

Równania warunków koniecznych, z założeniem $\hat{\lambda}_o = 1$:

$$2\hat{x}_1 - \hat{\lambda}_1 - \hat{\lambda}_2 = 0, \qquad (1)$$

$$2\hat{x}_2 - \hat{\lambda}_1 + \hat{\lambda}_2 = 0, \qquad (2)$$

$$2x_3 - \hat{\lambda}_1 - 2\hat{\lambda}_2 = 0, \qquad (3)$$

$$2 - \hat{x}_1 - \hat{x}_2 - \hat{x}_3 = 0, \quad (4),$$

$$3 - \hat{x}_1 + \hat{x}_2 - 2\hat{x}_3 = 0. \quad (5).$$

Równania (1), (2), (3) rozwiązywane względem $\hat{\lambda}_1, \hat{\lambda}_2$ dają

$$\hat{\lambda}_1 = \hat{x}_1 + \hat{x}_2,$$

$$\hat{\lambda}_2 = \hat{x}_3 - \frac{1}{2}(\hat{x}_1 + \hat{x}_2).$$

Rozwiązania te są jednoznaczne, mamy zatem do czynienia z przypadkiem a). Rozwiązując do końca układ równań warunków koniecznych znajduje się

$$\hat{x}_1 = \frac{11}{14}, \quad \hat{x}_2 = \frac{1}{14}, \quad \hat{x}_3 = \frac{8}{7}.$$

Przykład 2
Znaleźć $\min\left[f(\underline{x}) = x_1^2 + x_2^2 + x_3^2\right]$, przy

$$x_1 + x_2 + x_3 = 3,$$

$$3x_1 + 3x_2 + 3x_3 = 9.$$

Funkcja Lagrange'a

$$L = x_1^2 + x_2^2 + x_3^2 + \lambda_1(3 - x_1 - x_2 - x_3) + \lambda_2(9 - 3x_1 - 3x_2 - 3x_3).$$

Równania warunków koniecznych z założeniem $\hat{\lambda}_o = 1$:

$$2\hat{x}_1 - \hat{\lambda}_1 - 3\hat{\lambda}_2 = 0, \quad (1)$$

$$2\hat{x}_2 - \hat{\lambda}_1 - 3\hat{\lambda}_2 = 0, \quad (2)$$

$$2\hat{x}_3 - \hat{\lambda}_1 - 3\hat{\lambda}_2 = 0, \quad (3)$$

$$3 - \hat{x}_1 - \hat{x}_2 - \hat{x}_3 = 0, \quad (4)$$

$$9 - 3\hat{x}_1 - 3\hat{x}_2 - 3\hat{x}_3 = 0. \quad (5)$$

Równania (1), (2), (3) rozwiązywane względem $\hat{\lambda}_1, \hat{\lambda}_2$ dają warunek

$$\hat{\lambda}_1 + 3\hat{\lambda}_2 = 2\hat{x}_1 = 2\hat{x}_2 = 2\hat{x}_3.$$

Rozwiązanie to nie jest dla $\hat{\lambda}_1, \hat{\lambda}_2$ jednoznaczne, mamy zatem do czynienia z przypadkiem b). Nie przeszkadza to w dalszym rozwiązywaniu układu równań warunków koniecznych, co prowadzi do wyniku

$$\hat{x}_1 = 1, \quad \hat{x}_2 = 1, \quad \hat{x}_3 = 1.$$

Przykład 3

Znaleźć $[\max f(x) = x_1 + x_2]$

przy

$$x_1^2 + x_2^2 = 0.$$

Funkcja Lagrange'a

$$L = x_1 + x_2 + \lambda_1(x_1^2 + x_2^2).$$

Równania warunków koniecznych z założeniem $\hat{\lambda}_o = 1$:

$$1 + 2\hat{\lambda}_1 \hat{x}_1 = 0, \qquad (1)$$

$$1 + 2\hat{\lambda}_1 \hat{x}_2 = 0, \qquad (2)$$

$$\hat{x}_1^2 + \hat{x}_2^2 = 0. \qquad (3)$$

Powyższy układ równań nie ma rozwiązania na $\hat{\lambda}_1$; mianowicie równanie (3) wskazuje, że dopuszczalne jest tylko rozwiązanie $\hat{x}_1 = 0$, $\hat{x}_2 = 0$, a w tym punkcie równania (1), (2) nie posiadają rozwiązania na $\hat{\lambda}_1$. Jest to zatem przypadek $r[G_f] > r[G]$ (Czytelnik to łatwo sprawdzi) i należy równania warunków koniecznych przyjąć w postaci zakładającej $\hat{\lambda}_o = 0$:

$$2\hat{\lambda}_1 \hat{x}_1 = 0, \qquad (1)$$

$$2\hat{\lambda}_1 \hat{x}_2 = 0, \qquad (2)$$

$$\hat{x}_1^2 + \hat{x}_2^2 = 0. \qquad (3)$$

Ten układ równań ma rozwiązanie, a mianowicie $\hat{x}_1 = 0$, $\hat{x}_2 = 0$, $\hat{\lambda}_1$ dowolne. Warunki konieczne metody Lagrange'a (w jej postaci uogólnionej) są spełnione.

Rozpatrzone przykłady wskazują na słuszność kolejności postępowania polegającej na tym, by najpierw przyjmować $\hat{\lambda}_o = 1$ czyli postać warunków koniecznych według wzorów (19). Stwierdziwszy, że istnieje punkt lub punkty $\underline{\hat{x}}$, w których układ równań (19) nie

ma rozwiązań na mnożniki $\hat{\lambda}_i$, przejść należy do warunków w postaci uogólnionej (26), (27) z przyjęciem $\hat{\lambda}_o = 0$.

W każdym przypadku nie wolno zapomnieć o tym, że w zasadzie są do dyspozycji tylko warunki konieczne. Trzeba zatem, chcąc znaleźć potrzebne ekstremum, obliczyć wartość funkcji celu $f(\underline{x})$ w każdym znalezionym punkcie $\hat{\underline{x}}$.

Przedstawimy na zakończenie kilka dalszych przykładów.

Przykład 4

Znaleźć $\min \left[f(x_1, x_2) = x_1^2 + x_2^2 \right]$

przy ograniczeniu

$$g(x_1, x_2) = x_1 + x_2 = 1.$$

Funkcja Lagrange'a

$$L(x_1, x_2, \lambda) = x_1^2 + x_2^2 + \lambda(1 - x_1 - x_2).$$

Warunki konieczne, by $\hat{x}_1$, $\hat{x}_2$ było rozwiązaniem zadania:

$$\frac{\partial L}{\partial x_1} = 2\hat{x}_1 - \hat{\lambda} = 0,$$

$$\frac{\partial L}{\partial x_2} = 2\hat{x}_2 - \hat{\lambda} = 0,$$

$$\frac{\partial L}{\partial \lambda} = 1 - \hat{x}_1 - \hat{x}_2 = 0.$$

Rozwiązanie otrzymanego układu równań daje $\hat{\lambda} = 1$ oraz

$$\hat{x}_1 = \frac{1}{2}, \qquad \hat{x}_2 = \frac{1}{2}.$$

Należy zauważyć, że:

a) rozwiązanie $\hat{\underline{x}}$ jest jedno, zatem jeśli stanowi ono ekstremum (a nie tylko punkt stacjonarny), to jest to ekstremum globalne;

b) aby stwierdzić, czy $\hat{x}_1 = \frac{1}{2}$, $\hat{x}_2 = \frac{1}{2}$ istotnie stanowi szukane minimum warunkowe, trzebaby rozpatrzyć czy warunki konieczne są zarazem dostateczne (w tym przypadku tak jest). Można również zbadać otoczenie $\hat{x}_1$, $\hat{x}_2$ spełniające warunek pozostawiania w zbiorze dopuszczalnym, a więc w tym przykładzie rozpatrzyć $x_1 = \frac{1}{2} + \varepsilon$, $x_2 = \frac{1}{2} - \varepsilon$, oraz stwierdzić, że $f\left(\frac{1}{2} + \varepsilon, \frac{1}{2} - \varepsilon\right) \geq f\left(\frac{1}{2}, \frac{1}{2}\right)$.

Przykład 5

Znaleźć $\min \left[f(x_1 - x_2, x_3) = x_1^2 + x_2^2 + x_3^2 \right]$

przy ograniczeniach

$$g_1(x_1, x_2, x_3) = x_1 + x_2 = 1,$$

$$g_2(x_1, x_2, x_3) = x_2 + x_3 = 1.$$

Funkcja Lagrange'a

$$L(x_1, x_2, x_3, \lambda_1, \lambda_2) = x_1^2 + x_2^2 + x_3^2 + \lambda_1(1 - x_1 - x_2) + \lambda_2(1 - x_2 - x_3).$$

Warunki konieczne

$$\frac{\partial L}{\partial x_1} = 2\hat{x}_1 - \hat{\lambda}_1 = 0,$$

$$\frac{\partial L}{\partial x_2} = 2\hat{x}_2 - \hat{\lambda}_1 - \hat{\lambda}_2 = 0,$$

$$\frac{\partial L}{\partial x_3} = 2\hat{x}_3 - \hat{\lambda}_2 = 0,$$

$$\frac{\partial L}{\partial \lambda_1} = 1 - \hat{x}_1 - \hat{x}_2 = 0,$$

$$\frac{\partial L}{\partial \lambda_2} = 1 - \hat{x}_2 - \hat{x}_3 = 0.$$

Rozwiązanie otrzymanego układu równań jest tylko jedno (otrzymuje się je przy $\hat{\lambda}_1 = \frac{2}{3}$, $\hat{\lambda}_2 = \frac{2}{3}$);

$$\hat{x}_1 = 1/3, \quad \hat{x}_2 = 2/3, \quad \hat{x}_3 = 1/3.$$

Jest to szukane minimum, bowiem $f(\underline{x})$ było wypukłe, a ograniczenia - liniowe.

Przykład 6

Znaleźć $\min\left[f(x_1, x_2) = x_1^2 + x_2^2\right]$
przy ograniczeniu

$$g(x_1, x_2) = x_2 - x_1^2 = 1.$$

Funkcja Lagrange'a

$$L(x_1, x_2, \lambda) = x_1^2 + x_2^2 + \lambda(1 + x_1^2 - x_2).$$

Warunki konieczne

$$\frac{\partial L}{\partial x_1} = 2\,\hat{x}_1 + 2\,\hat{x}_1\hat{\lambda} = 0,$$

$$\frac{\partial L}{\partial x_2} = 2\,\hat{x}_2 - \hat{\lambda} = 0,$$

$$\frac{\partial L}{\partial \lambda} = 1 + \hat{x}_1^2 - \hat{x}_2 = 0.$$

Rozwiązanie układu równań

a) $\hat{x}_1 = 0, \quad \hat{x}_2 = 1, \quad$ przy $\quad \hat{\lambda} = 2,$

b) $\hat{x}_1 = \sqrt{-\frac{3}{2}}, \quad \hat{x}_2 = -\frac{1}{2}, \quad$ przy $\quad \hat{\lambda} = -1.$

Rozwiązaniem zadania może być tylko $\hat{x}_1 = 0, \quad \hat{x}_2 = 1,$ gdyż rozpatruje się zmienne rzeczywiste.

Badając otoczenie $[\hat{x}_1, \hat{x}_2]$ można sprawdzić, że jest to istotnie poszukiwane minimum.

Przykład 7

W obiekcie sterowania wielkość wyjściowa y (będąca natężeniem produkcji w t/godz) zależy od dwóch wielkości sterujących wg zależności

$$y = u_1 + 2\,u_1 u_2 = g(u_1, u_2).$$

Natężenie kosztu produkcji (funkcja celu) wynosi

$$f(u_1, u_2) = c_1 u_1 + c_2 u_2.$$

Należy zapewnić $\min f(u_1, u_2)$ przy zadanej produkcji $y = y_d$. Funkcja Lagrange'a:

$$L(u_1, u_2, \lambda) = c_1 u_1 + c_2 u_2 + \lambda(y_d - u_1 - 2\,u_1 u_2).$$

Warunki konieczne:

$$\frac{\partial L}{\partial u_1} = c_1 - \hat{\lambda}(1 + 2\,\hat{u}_2) = 0,$$

$$\frac{\partial L}{\partial u_2} = c_2 - 2\hat{\lambda}\hat{u}_1 = 0,$$

$$\frac{\partial L}{\partial \lambda} = y_d - \hat{u}_1 - 2\,\hat{u}_1\hat{u}_2 = 0.$$

Rozwiązania otrzymanego układu równań są następujące:

a) $\hat{u}_1 = \sqrt{\frac{y_d c_2}{2 c_1}}$, $\hat{u}_2 = \sqrt{\frac{y_d c_1}{2 c_2}} - \frac{1}{2}$, przy $\hat{\lambda} = \sqrt{\frac{c_1 c_2}{2 y_d}}$,

b) $\hat{u}_1 = -\sqrt{\frac{y_d c_2}{2 c_1}}$, $\hat{u}_2 = -\sqrt{\frac{y_d c_1}{2 c_2}} - \frac{1}{2}$, przy $\hat{\lambda} = -\sqrt{\frac{c_1 c_2}{2 y_d}}$.

Badając otoczenie rozwiązania (a) stwierdza się, że jest to minimum $f(u_1, u_2)$, czyli szukane rozwiązanie zadania. Punkt określony rozwiązaniem (b) przypada dla ujemnych u_1, u_2 oraz stanowi lokalne maksimum. Ujemne u_2 może wystąpić również w rozwiązaniu (a), zależnie od warunków zadania. Jeśli się tak zdarzy, a $u_2 < 0$ nie jest fizycznie dopuszczalne, zadanie trzeba rozwiązywać z warunkiem $u_2 \geqslant 0$. Do zadań tego typu służy omawiana dalej metoda Kuhna-Tuckera, gdyż metoda Lagrange'a przypadków z ograniczeniami nierównościowymi nie obejmuje.

2.3. Warunki konieczne i wystarczające punktu siodłowego

W omawianej w następnym podrozdziale teorii Kuhna i Tuckera, służącej do rozpatrywania zadań nieliniowych z ograniczeniami równościowymi i nierównościowymi, istotne miejsce zajmuje pojęcie punktu siodłowego funkcji wielu zmiennych. Tytułem zatem przygotowania do następnego podrozdziału przedstawimy warunki konieczne i wystarczające punktu siodłowego.

Punkt siodłowy definiuje się następująco:

Funkcja $L(\underline{x}, \underline{\lambda})$ ma punkt siodłowy w punkcie $[\hat{\underline{x}}, \hat{\underline{\lambda}}]$ jeśli istnieje takie otoczenie $\varepsilon > 0$, że dla wszystkich $\underline{x}$, $|\underline{x} - \hat{\underline{x}}| < \varepsilon$, oraz wszystkich λ, $|\lambda - \hat{\lambda}| < \varepsilon$ obowiązuje zależność

$$L(\underline{x}, \hat{\underline{\lambda}}) \leqslant L(\hat{\underline{x}}, \hat{\underline{\lambda}}) \leqslant L(\hat{\underline{x}}, \underline{\lambda}). \tag{26}$$

Definicja powyższa określiła lokalny punkt siodłowy; jeżeli układ nierówności (26) obowiązuje dla wszystkich $\underline{x}$, $\underline{\lambda}$ z pewnego zadanego obszaru, mamy globalny punkt siodłowy.

Jak wskazuje układ nierówności (26), można poszukiwanie punktu siodłowego uważać za poszukiwanie

$$\min_{\underline{\lambda}} \max_{\underline{x}} L(\underline{x}, \underline{\lambda}). \tag{27}$$

Związki (26), (27) odnoszą się do punktu siodłowego typu "maksimum względem $\underline{x}$, minimum względem $\underline{\lambda}$". Zależnie od zadania może nas oczywiście interesować sytuacja odwrotna.

Rozpatrzeć teraz trzeba punkty siodłowe w przypadku, gdy część spośród x_j oraz część spośród λ_i jest ograniczonego zna-

ku; zapisuje się to w postaci $\underline{x} \in W_1$, przy czym W_1 określony jest w sposób następujący:

$$x_j \geq 0, \qquad j = 1, \ldots s,$$

$$x_j \leq 0, \qquad j = s+1, \ldots t,$$

$$x_j \text{ nieograniczone } j = t+1, \ldots n$$

oraz $\underline{\lambda} \in W_2$, gdzie obszar W_2 określony jest jako

$$\lambda_i \geq 0, \qquad i = 1, \ldots u,$$

$$\lambda_i \leq 0, \qquad i = u+1, \ldots v,$$

$$\lambda_i \text{ nieograniczone}, \quad i = v+1, \ldots m.$$

Zakładając, że $L(\underline{x}, \underline{\lambda}) \in C^1$, można wyprowadzić następujące warunki konieczne by $L(\underline{x}, \underline{\lambda})$ miała punkt siodłowy w punkcie $[\hat{\underline{x}}, \hat{\underline{\lambda}}]$ dla $\underline{x} \in W_1$, $\underline{\lambda} \in W_2$:

$$\frac{\partial L(\hat{\underline{x}}, \hat{\underline{\lambda}})}{\partial x_j} \leq 0, \qquad j = 1, \ldots s,$$

$$\frac{\partial L(\hat{\underline{x}}, \hat{\underline{\lambda}})}{\partial x_j} \geq 0, \qquad j = s+1, \ldots t, \tag{28}$$

$$\frac{\partial L(\hat{\underline{x}}, \hat{\underline{\lambda}})}{\partial x_j} = 0, \qquad j = t+1, \ldots n,$$

$$\hat{x}_j \geq 0, \quad j = 1, \ldots s; \qquad \hat{x}_j \leq 0, \quad j = s+1, \ldots t; \tag{29}$$

$$\hat{x}_j \text{ nieograniczone}, \qquad j = t+1, \ldots n$$

$$\hat{x}_j \frac{\partial L(\hat{\underline{x}}, \hat{\underline{\lambda}})}{\partial x_j} = 0, \qquad j = 1, \ldots n, \tag{30}$$

$$\frac{\partial L(\hat{\underline{x}}, \hat{\underline{\lambda}})}{\partial \lambda_i} \geq 0, \qquad i = 1, \ldots u,$$

$$\frac{\partial L(\hat{\underline{x}}, \hat{\underline{\lambda}})}{\partial \lambda_i} \leq 0, \qquad i = u+1, \ldots v, \tag{31}$$

$$\frac{\partial L(\hat{\underline{x}}, \hat{\underline{\lambda}})}{\partial \lambda_i} = 0, \qquad i = v+1, \ldots m,$$

$\hat{\lambda}_i > 0, \qquad i = 1,\dots u; \qquad \hat{\lambda}_i \leqslant 0, \qquad i = u+1,\dots v; \qquad (32)$

$\hat{\lambda}_i$ nieograniczone, $\qquad i = v+1,\dots m$

$$\hat{\lambda}_i \frac{\partial L(\hat{\underline{x}}, \hat{\underline{\lambda}})}{\partial \hat{\lambda}_i} = 0, \qquad i = 1,\dots m. \qquad (33)$$

Powyższe warunki są warunkami koniecznymi i wystarczającymi punktu siodłowego w $[\hat{\underline{x}}, \hat{\underline{\lambda}}]$, jeżeli dla $\underline{x} \in W_1$ w otoczeniu ε punktu $\hat{\underline{x}}$ funkcja $L(\underline{x}, \hat{\underline{\lambda}})$ jest wklęsłą funkcją $\underline{x}$, oraz dla $\underline{\lambda} \in W_2$ w otoczeniu ε punktu $\hat{\underline{\lambda}}$ funkcja $L(\hat{\underline{x}}, \underline{\lambda})$ jest wypukłą funkcją $\underline{\lambda}$.

Jeżeli $L(\underline{x}, \hat{\underline{\lambda}})$ jest funkcją wklęsłą $\underline{x}$ dla wszystkich $\underline{x} \in W_1$ oraz $L(\hat{\underline{x}}, \underline{\lambda})$ jest funkcją wypukłą $\underline{\lambda}$ dla wszystkich $\underline{\lambda} \in W_2$, to warunki (28)...(33) są warunkami koniecznymi i wystarczającymi globalnego punktu siodłowego funkcji $L(\underline{x}, \underline{\lambda})$.

W przypadku, gdy szuka się minimum względem $\underline{x}$, maksimum względem $\underline{\lambda}$, należy w warunkach (28)...(33) zmienić znaki nierówności (28) oraz (31) na przeciwne.

2.4. Zadanie z ograniczeniami nierównościowymi. Metoda Kuhna-Tuckera

Rozpatrywane będzie teraz zadanie następujące:

$$\max f(\underline{x}), \qquad (34)$$

przy $\underline{x}$ spełniającym warunki

$$\begin{aligned} b_i - g_i(\underline{x}) &\geqslant 0, \qquad i = 1,\dots u, \\ b_i - g_i(\underline{x}) &\leqslant 0, \qquad i = u+1,\dots v, \\ b_i - g_i(\underline{x}) &= 0, \qquad i = v+1,\dots m \end{aligned} \qquad (35)$$

oraz dodatkowo

$$x_j \geqslant 0, \qquad j = 1,\dots s; \qquad x_j \leqslant 0, \qquad j = s+1,\dots t; \qquad (36)$$

x_j nieograniczonego znaku, $\qquad j = t+1,\dots n.$

Kuhn i Tucker [35], [26] wykazali, że jeżeli $f(\underline{x}) \in C^1$, $g_i(\underline{x}) \in C^1$ oraz $g_i(\underline{x})$ spełniają pewne warunki regularności w punkcie $\hat{\underline{x}}$, to warunkiem koniecznym istnienia ekstremum warunkowego $f(\underline{x})$ w punkcie $\hat{\underline{x}}$ jest spełnienie przez $[\hat{\underline{x}}, \hat{\underline{\lambda}}]$ warunków koniecznych (28)...(33) punktu siodłowego funkcji Lagrange'a:

$$L(\underline{x}, \underline{\lambda}) = f(\underline{x}) + \sum_{1}^{m} \lambda_i \left[b_i - g_i(\underline{x}) \right]. \qquad (37)$$

Dla określonych przypadków znane są warunki konieczne i wystarczające: jeżeli $f(\underline{x})$ jest wklęsłe względem $\underline{x} \in W_1$, a $g_i(\underline{x})$ są wypukłe względem $\underline{x}$ dla $\lambda_i > 0$, wklęsłe względem $\underline{x}$ dla $\lambda_i < 0$ oraz liniowe względem x dla λ_i nieograniczonego znaku, $i = 1, \ldots m$, to $\hat{\underline{x}}$ stanowi maksimum warunkowe globalne, $f(\hat{\underline{x}}) \geq f(\underline{x})$ dla $\underline{x} \in Y$, wtedy i tylko wtedy gdy spełnione są warunki (28)...(33) punktu siodłowego funkcji Lagrange'a (37).

W twierdzeniu powyższym $\underline{x} \in W_1$ oznacza spełnienie warunków (36), a $\underline{x} \in Y$ oznacza spełnienie warunków (35).

Ponieważ opisana w twierdzeniu sytuacja określa dokładnie, że $L(\underline{x}, \underline{\lambda})$ jest wklęsłe względem $\underline{x}$ oraz wypukłe względem $\underline{\lambda}$, spełnienie warunków (28)...(33) oznacza tu zarazem, że $L(\underline{x}, \underline{\lambda})$ ma w $[\hat{\underline{x}}, \hat{\underline{\lambda}}]$ globalny punkt siodłowy.

W myśl metody Kuhna i Tuckera, poszukiwanie rozwiązania zadania (34), (35), (36) polegać będzie na wyznaczeniu $[\hat{\underline{x}}, \hat{\underline{\lambda}}]$ spełniających warunki (28)...(33). Ponieważ nie może być rozwiązań optymalnych innych niż spełniające (28)...(33), wystarczy następnie dokonać przeglądu wartości $f(\underline{x})$ w wyznaczonych punktach $\hat{\underline{x}}$ i wybrać globalne maksimum. Przeglądu tego można oczywiście nie robić jeśli wiadomo, że w danym zadaniu warunki konieczne (28)...(33) były zarazem wystarczające.

Dla zadań, w których poszukuje się $\min f(\underline{x})$, można bądź przeformułować odpowiednie warunki konieczne punktu siodłowego, bądź też wykorzystać zależność

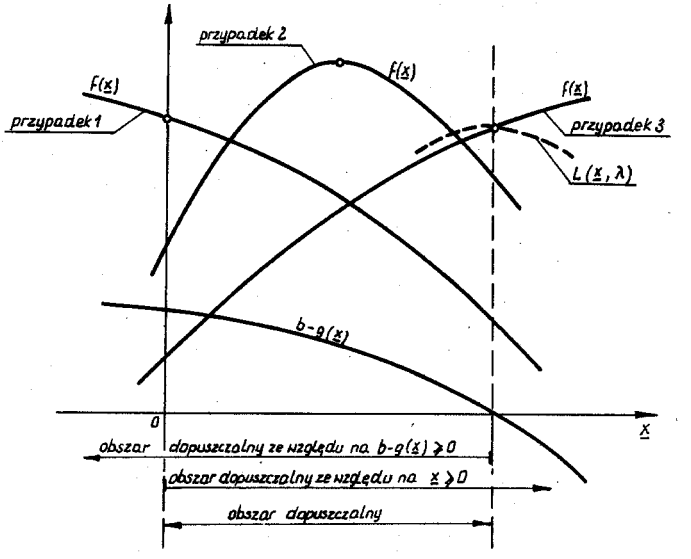
$$\max f(\underline{x}) = -\min \left[-f(\underline{x})\right] . \tag{38}$$

Warto jeszcze zwrócić uwagę, że wśród warunków (28)...(33) część pochodzi ze sformułowania zadania, patrz (34), (35), (36), a pozostała część stanowi właściwe postulaty warunków koniecznych rozwiązania. I tak, z zadania pochodzić muszą wymagania na nieujemność bądź niedodatniość zmiennych $\hat{x}_j$ czyli warunki (29), oraz z zadania pochodzą warunki (31), gdyż są one przepisaniem warunków (35).

Istotę warunków Kuhna-Tuckera uzmysłowić sobie można przy pomocy interpretacji graficznej następującego zadania: znaleźć $\max f(\underline{x})$, $x_j \geq 0$, przy ograniczeniu $b - g(\underline{x}) \geq 0$. Na rys. 4 przedstawiono umownie wklęsłą funkcję $f(\underline{x})$ oraz wartość ograniczenia $b - g(\underline{x})$, przy czym $g(\underline{x})$ jest wypukłe.

Możliwe są 3 przypadki położenia wklęsłej funkcji $f(\underline{x})$ na wykresie. W pierwszym (krzywa 1) właściwe rozwiązanie leży w punkcie $\hat{\underline{x}} = \underline{0}$, w drugim (krzywa 2) rozwiązanie $\hat{\underline{x}} > 0$ leży wewnątrz obszaru $b - g(\hat{\underline{x}}) > 0$, w trzecim przypadku (krzywa 3) rozwiązanie $\hat{\underline{x}} > 0$ leży na brzegu obszaru ograniczeń $b - g(\hat{\underline{x}}) = 0$.

Rozpatrzmy spełnienie warunków Kuhna-Tuckera w tych trzech przypadkach.



Rys. 4

Przypadek 1

Rozwiązanie optymalne $\hat{\underline{x}} = \underline{0}$ wynikło stąd, że jest $\frac{\partial f}{\partial \underline{x}} < \underline{0}$ w całym obszarze dopuszczalnym. Ograniczenie $b - g(\underline{x}) \geqslant 0$ jest w tym przypadku nieaktywne. Biorąc pod uwagę funkcję Lagrange'a

$$L(\underline{x}, \lambda) = f(\underline{x}) + \lambda \left[b - g(\underline{x}) \right] ,$$

widzimy, że przypadek rozwiązania $\hat{\underline{x}} = \underline{0}$ charakteryzują

$$\hat{\lambda} = 0, \qquad b - g(\hat{\underline{x}}) > 0, \qquad \frac{\partial L}{\partial \underline{x}} < \underline{0}, \qquad \hat{\underline{x}} = \underline{0} .$$

(Wartość $\hat{\lambda} = 0$ eliminuje z funkcji Lagrange'a nieaktywne ograniczenie).

Przypadek 2

Rozwiązanie optymalne $\hat{\underline{x}} > \underline{0}$ wynikło stąd, że w pewnym punkcie wewnątrz obszaru dopuszczalnego jest $\frac{\partial f}{\partial \underline{x}} = \underline{0}$. Ograniczenie $b - g(\underline{x}) \geqslant 0$ jest w tym przypadku nieaktywne, podobnie jak ograniczenie $\underline{x} \geqslant 0$. Biorąc pod uwagę funkcję Lagrange'a widzimy,

że ten przypadek charakteryzują:

$$\hat{\lambda} = 0, \quad b - g(\hat{\underline{x}}) > 0, \quad \frac{\partial L}{\partial \underline{x}} = \underline{0}, \quad \hat{\underline{x}} > \underline{0}.$$

Przypadek 3

Rozwiązanie optymalne $\hat{\underline{x}} > \underline{0}$, leżące na brzegu ograniczenia $b - g(\underline{x}) \geq 0$ wynikło stąd, że w całym obszarze jest $\frac{\partial f}{\partial \underline{x}} > \underline{0}$. Biorąc pod uwagę funkcję Legrange'a widzimy, że ten przypadek jest podobny do przypadku z ograniczeniem równościowym, jest tu bowiem $b - g(\hat{\underline{x}}) = 0$. Warunki konieczne rozwiązania zadania z ograniczeniem równościowym brzmią, jak wiadomo:

$$\frac{\partial L}{\partial \underline{x}} = \frac{\partial f}{\partial \underline{x}} - \hat{\lambda} \frac{\partial g}{\partial \underline{x}} = \underline{0},$$

$$\frac{\partial L}{\partial \lambda} = b - g(\hat{\underline{x}}) = 0.$$

W rozpatrywanym przez nas przypadku jest $\frac{\partial f}{\partial \underline{x}} > \underline{0}$; aby było $\frac{\partial L}{\partial \underline{x}} = \underline{0}$ musi zatem być $\hat{\lambda} \frac{\partial g}{\partial \underline{x}} > 0$. Charakter ograniczenia $b - g(\underline{x}) \geq 0$, patrz rysunek, określa że jest $\frac{\partial g}{\partial \underline{x}} > \underline{0}$. Musi zatem być $\hat{\lambda} > 0$ i ostatecznie rozwiązanie w przypadku trzecim charakteryzują dane

$$\hat{\lambda} > 0, \quad b - g(\hat{\underline{x}}) = 0, \quad \frac{\partial L}{\partial \underline{x}} = \underline{0}, \quad \hat{\underline{x}} > \underline{0}.$$

Zauważmy, że dzięki wypukłości $g(\underline{x})$ funkcja Lagrange'a ma ekstremum bezwarunkowe w punkcie optymalnym, będącym rozwiązaniem dla przypadku trzeciego.

Łącząc cechy wszystkich trzech przypadków widzimy, że mieszczą się one w sformułowaniu warunków Kuhna-Tuckera:

$$\frac{\partial L}{\partial \underline{x}} \leq \underline{0}, \qquad \hat{\underline{x}} \geq \underline{0},$$

$$\hat{x}_j \frac{\partial L}{\partial x_j} = 0, \quad j = 1, \ldots n,$$

$$\frac{\partial L}{\partial \lambda} \geq 0, \qquad \hat{\lambda} \geq 0,$$

$$\hat{\lambda} \frac{\partial L}{\partial \lambda} = 0.$$

Zanim przedstawimy kilka przykładów stosowania metody Kuhna-Tuckera, trzeba powrócić do warunków regularności, które mu-

szą spełniać funkcje $g_i(x)$, by teoria Kuhna-Tuckera była słuszna. W pełni ogólne omówienie tych warunków jest dość złożone; można je znaleźć w literaturze [26] [36]. W każdym razie należy zwrócić uwagę na to, że jeśli jakiekolwiek ograniczenia są nieaktywne lub którekolwiek elementy rozwiązania $\hat{x}_j$ różne od zera, to to musi istnieć pewien mały obszar w pobliżu $\hat{\underline{x}}$, w którym - dla każdego $\underline{x}$ leżącego w tym obszarze - dane ograniczenia są nadal nieaktywne lub dane x_j są nadal różne od zera. Przy rozważaniu warunków regularności analizować zatem należy tylko ograniczenia aktywne w punkcie $\hat{\underline{x}}$ oraz elementy $\hat{x}_j$ leżące na brzegu swych ograniczeń typu $x_j \geq 0$ lub $x_j \leq 0$.

Zauważmy, że punkt $\hat{\underline{x}}$ można uważać za rozwiązanie zadania z ograniczeniami tylko równościowymi, biorąc za te ograniczenia równościowe:

aktywne ograniczenia nierównościowe $g_i(\underline{x}) = b_i$, (39)

aktywne ograniczenia znaku $x_j = 0$. (40)

Przypomnijmy dla zadania z ograniczeniami równościowymi uogólnione warunki metody Lagrange'a: w punkcie $\hat{\underline{x}}$ spełniony ma być układ równań

$$\hat{\lambda}_o \frac{\partial f(\hat{\underline{x}})}{\partial x_j} - \sum_{i=1}^{m} \hat{\lambda}_i \frac{\partial g_i(\hat{\underline{x}})}{\partial x_j} = 0, \quad j = 1, \ldots n, \tag{41}$$

dla przynajmniej jednego zbioru $\hat{\lambda}_i$, $i = 0, \ldots m$, w którym nie wszystkie $\hat{\lambda}_i$ są zerami oraz, ponadto, równania ograniczeń czyli (39), (40).

Operując funkcją Lagrange'a, w której $\hat{\lambda}_o = 1$, teoria Kuhna-Tuckera zakłada w istocie, że mamy do czynienia z przypadkami, gdy

$$r\left[G_f\right] = r\left[G\right],$$

przy czym znaczenie macierzy G_f oraz G było omówione w rozdziale 2.2. Jeśli zatem istnieją takie punkty $\hat{\underline{x}}$, w których leży rozwiązanie zadania, lecz w których

$$r\left[G_f\right] > r\left[G\right],$$

to operując warunkami Kuhna-Tuckera punktów tych możemy nie wykryć. Praktycznie należy zatem w konkretnym zadaniu:

(1) znaleźć wszystkie $\hat{\underline{x}}$, w których spełnione są warunki Kuhna-Tuckera,

(2) znaleźć wszystkie punkty nieregularne rozpatrywanego zadania, tj. te, w których $r\left[G_f\right] > r\left[G\right]$, przy czym chodzi tu oczywiście o punkty równościowego spełnienia ograniczeń, czyli leżące na brzegach obszaru dopuszczalnego (ściślej, w pewnego ro-

dzaju "ostrzach" tego brzegu, gdyż tam zachodzić może $r[G] < m$, zatem może zaistnieć $r[G_f] > r[G]$),

(3) zbadać wartość $f(\underline{x})$ w punktach $\hat{\underline{x}}$ oraz w punktach nieregularności i wybrać szukane maksimum czy minimum.

Rozpatrzony będzie następujący przykład.

Znaleźć $\max[f(\underline{x}) = 10\,x_1 + x_2]$ przy warunkach

$$g(\underline{x}) = (3 - x_1)^3 - x_2 \geq 0; \qquad x_1 \geq 0, \quad x_2 \geq 0.$$

Tworzy się funkcję Lagrange'a

$$L(x_1, x_2, \lambda) = 10\,x_1 + x_2 + \lambda\left[(3 - x_1)^3 - x_2\right]$$

i pisze warunki Kuhna-Tuckera dla tego przypadku

$$\frac{\partial L}{\partial x_1} \leq 0, \quad \text{czyli} \quad 10 - 3\hat{\lambda}(3 - \hat{x}_1)^2 \leq 0, \tag{a}$$

$$\frac{\partial L}{\partial x_2} \leq 0, \quad \text{czyli} \quad 1 - \hat{\lambda} \leq 0, \tag{b}$$

$$\hat{x}_1 \frac{\partial L}{\partial x_1} = 0, \quad \text{czyli} \quad \hat{x}_1\left[10 - 3\hat{\lambda}(3 - \hat{x}_1)^2\right] = 0, \tag{c}$$

$$\hat{x}_2 \frac{\partial L}{\partial x_2} = 0 \qquad \hat{x}_2\,(1 - \hat{\lambda}) = 0, \tag{d}$$

$$\frac{\partial L}{\partial \lambda} \geq 0, \quad \text{czyli} \quad (3 - \hat{x}_1)^3 - \hat{x}_2 \geq 0, \tag{e}$$

$$\hat{\lambda} \geq 0,$$

$$\hat{\lambda} \frac{\partial L}{\partial \lambda} = 0, \quad \text{czyli} \quad \hat{\lambda}\left[(3 - \hat{x}_1)^3 - \hat{x}_2\right] = 0. \tag{f}$$

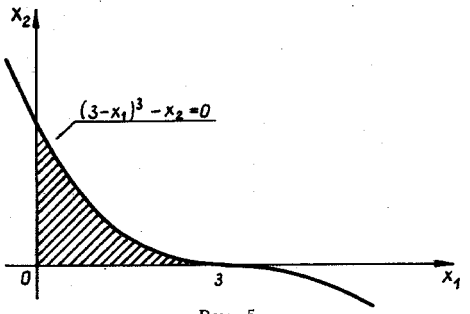
Z warunków (a) do (f) znaleźć można rozwiązanie:

$\hat{x}_1 = 0, \quad \hat{x}_2 = 27,$ przy czym $\hat{\lambda} = 1.$

Zwróćmy jednak uwagę na leżący na brzegu obszaru ograniczeń punkt $x_1 = 3, \quad x_2 = 0,$ zatem punkt, w którym spełnione są równościowo następujące ograniczenia (patrz rys. 5):

$$g(\underline{x}) = (3 - x_1)^3 - x_2 = 0,$$

$$x_2 = 0.$$



Rys. 5

Napiszmy macierze G oraz G_f w tym punkcie:

$$G = \begin{bmatrix} \frac{\partial g}{\partial x_1} & \frac{\partial g}{\partial x_2} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 0 & 1 \end{bmatrix}$$

$$G_f = \begin{bmatrix} \frac{\partial g}{\partial x_1} & \frac{\partial g}{\partial x_2} \\ 0 & 1 \\ \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 0 & 1 \\ 10 & 1 \end{bmatrix}$$

Łatwo zauważyć, że $r[G_f] = 2$, podczas gdy $r[G] = 1$. Jest zatem $r[G_f] > r[G]$ i punkt $[3, 0]$ jest punktem nieregularności (choć nie byłby takim punktem, gdyby funkcją celu było np. $f(\underline{x}) = x_2$). Sprawdzono wartość funkcji celu w punkcie nieregularności: $f(\underline{x}) = 30$, podczas gdy w punkcie poprzednio znalezionym z warunków Kuhna-Tuckera $f(\underline{x}) = 27$. Maksimum globalne leży zatem w punkcie nieregularności $[3, 0]$.

Powyższy przykład wskazuje na znaczenie sprawdzania punktów nieregularności ograniczeń. Podkreślmy jeszcze na praktyczny użytek, że jeśli ograniczenia $g(\underline{x})$ są liniowe, warunki regularności są zawsze spełnione.

Przedstawimy teraz kilka liczbowych przykładów stosowania metody Kuhna-Tuckera.

Przykład 1

Znaleźć $\max \left[f(\underline{x}) = -(x_1 - 2)^2 - (x_2 - 4)^2 \right]$ przy warunku $x_1 + x_2 \leqslant 4$, czyli $4 - x_1 - x_2 \geqslant 0$ bez ograniczenia znaku x_1, x_2.

Funkcja Lagrange'a:

$$L(x_1, x_2, \lambda) = -(x_1 - 2)^2 - (x_2 - 4)^2 + \lambda(4 - x_1 - x_2).$$

Warunki Kuhna-Tuckera:

$$\frac{\partial L}{\partial x_1} = 0, \quad \text{czyli} \quad -2(\hat{x}_1 - 2) - \hat{\lambda} = 0, \tag{a}$$

$$\frac{\partial L}{\partial x_2} = 0, \quad \text{czyli} \quad -2(\hat{x}_2 - 4) - \hat{\lambda} = 0, \tag{b}$$

$$\hat{x}_1 \frac{\partial L}{\partial x_1} = 0, \quad \text{wobec} \quad \frac{\partial L}{\partial x_1} = 0 \quad \text{warunek ten nic nie wnosi,}$$

$$\hat{x}_2 \frac{\partial L}{\partial x_2} = 0, \quad \text{wobec} \quad \frac{\partial L}{\partial x_2} = 0 \quad \text{warunek ten nic nie wnosi,}$$

$$\frac{\partial L}{\partial \lambda} \geqslant 0, \quad \text{czyli} \quad 4 - \hat{x}_1 - \hat{x}_2 \geqslant 0, \tag{c}$$

$$\hat{\lambda} \geqslant 0,$$

$$\hat{\lambda} \frac{\partial L}{\partial \lambda} = 0, \quad \text{czyli} \quad \hat{\lambda}(4 - \hat{x}_1 - \hat{x}_2) = 0. \tag{d}$$

Metodycznie można rozwiązać układ (a) (b) (c) (d) rozpoczynając od równań, tj. od (a) (b) (d) i akceptując następnie to rozwiązanie, które spełni nierówność (c). Sporządzono tabelkę rozwiązań układu równań (a) (b) (d):

	rozw. I	rozw. II
$\hat{x}_1$	2	1
$\hat{x}_2$	4	3
$\hat{\lambda}$	0	2
$\hat{\lambda} \geqslant 0$	tak	tak
$4 - \hat{x}_1 - \hat{x}_2 \geqslant 0$	nie	tak

Warunki K-T spełnia zatem $\hat{x}_1 = 1$, $\hat{x}_2 = 3$. W rozpatrywanym przykładzie $f(\underline{x})$ jest wklęsłe, $g(\underline{x})$ liniowe - zatem warunki K-T są konieczne i wystarczające. Wartości $\hat{x}_1 = 1$, $\hat{x}_2 = 3$ są wobec tego z pewnością rozwiązaniem zadania.

Zwróćmy jeszcze uwagę na wartości $x_1 = 2$, $x_2 = 4$, $\lambda = 0$, spełniające równania (a) (b) (d), lecz nie spełniające nierówności (c). Wartość $\lambda = 0$ oznacza, że w $L(\underline{x}, \lambda)$ ignoruje się ograniczenie (uważając je za "nieaktywne"), i rzeczywiście $x_1 = 2$, $x_2 = 4$ odpowiadają maksymalizacji $f(\underline{x})$ bez uwzględnienia ograniczenia.

Przykład 2

Znaleźć $\max\left[f(\underline{x}) = (x_1 - 2)^2 + (x_2 - 4)^2\right]$ przy warunkach $x_1 + x_2 \leqslant 8$ czyli $8 - x_1 - x_2 \geqslant 0$ oraz $x_1 \geqslant 0$, $x_2 \geqslant 0$.

Funkcja Lagrange'a

$$L(x_1, x_2, \lambda) = (x_1 - 2)^2 + (x_2 - 4)^2 + \lambda(8 - x_1 - x_2).$$

Warunki Kuhna-Tuckera

$$\frac{\partial L}{\partial x_1} \leqslant 0, \quad \text{czyli} \quad 2(\hat{x}_1 - 2) - \hat{\lambda} \leqslant 0, \tag{a}$$

$$\frac{\partial L}{\partial x_2} \leqslant 0, \quad \text{czyli} \quad 2(\hat{x}_2 - 4) - \hat{\lambda} \leqslant 0, \tag{b}$$

$$\hat{x}_1 \geqslant 0, \quad \hat{x}_2 \geqslant 0,$$

$$\hat{x}_1 \frac{\partial L}{\partial x_1} = 0, \quad \text{czyli} \quad \hat{x}_1(2\hat{x}_1 - 4 - \hat{\lambda}) = 0, \tag{c}$$

$$\hat{x}_2 \frac{\partial L}{\partial x_2} = 0, \quad \text{czyli} \quad \hat{x}_2(2\hat{x}_2 - 8 - \hat{\lambda}) = 0, \tag{d}$$

$$\frac{\partial L}{\partial \lambda} \geqslant 0, \quad \text{czyli} \quad 8 - \hat{x}_1 - \hat{x}_2 \geqslant 0, \tag{e}$$

$$\hat{\lambda} \geqslant 0,$$

$$\hat{\lambda} \frac{\partial L}{\partial \lambda} = 0, \quad \text{czyli} \quad \hat{\lambda}(8 - \hat{x}_1 - \hat{x}_2) = 0. \tag{f}$$

Rozwiązanie powstałego układu równań i nierówności dogodnie jest rozpocząć od sporządzenia tabelki rozwiązań równań (c), (d), (f), a następnie sprawdzić, które z rozwiązań spełniają nierówności (a), (b), (e) oraz warunki znaku $\hat{x}_1$, $\hat{x}_2$, $\hat{\lambda}$.

	rozw. I	rozw. II	rozw. III	rozw. IV	rozw. V	rozw. VI	rozw. VII
$\hat{x}_1$	3	2	0	2	0	0	8
$\hat{x}_2$	5	4	4	0	0	8	0
$\hat{\lambda}$	2	0	0	0	0	8	12
$\hat{x}_1 \geqslant 0,\ \hat{x}_2 \geqslant 0,\ \hat{\lambda} \geqslant 0$	tak	tak	tak	tak	tak	tak	tak
$\frac{\partial L}{\partial x_1} \leqslant 0$	tak	tak	tak	tak	tak	tak	tak
$\frac{\partial L}{\partial x_2} \leqslant 0$	tak	tak	tak	tak	tak	tak	tak
$\frac{\partial L}{\partial \lambda} \geqslant 0$	tak	tak	tak	tak	tak	tak	tak
wartość $f(\underline{\hat{x}})$	2	0	4	16	20	20	52

Obserwując w tabelce spełnienie warunków K-T widzimy, że są one spełnione przez 7 punktów $\underline{\hat{x}}$; przegląd wartości $f(\underline{x})$ w tych punktach pokazuje, że globalne maksimum wypada dla $\hat{x}_1 = 8$, $\hat{x}_2 = 0$. Czytelnik zauważy, że w tym przykładzie funkcja celu $f(\underline{x})$ nie była wklęsła.

Przykład 3

Znaleźć $\max\left[f(\underline{x}) = -10(x_1 - 3,5)^2 - 20(x_2 - 4)^2\right]$; przy warunkach

$$x_1 + x_2 \leqslant 6, \quad \text{czyli} \quad 6 - x_1 - x_2 \geqslant 0,$$

$$2x_1 + x_2 \geqslant 6, \quad \text{czyli} \quad 6 - 2x_1 - x_2 \leqslant 0$$

oraz $x_2 \geqslant 0$, x_1 nieograniczonego znaku.

Funkcja Lagrange'a:

$$L(x_1, x_2, \lambda_1, \lambda_2) = -10(x_1 - 3,5)^2 - 20(x_2 - 4)^2 +$$

$$+ \lambda_1(6 - x_1 - x_2) + \lambda_2(6 - 2x_1 - x_2).$$

Warunki Kuhna-Tuckera:

$$\frac{\partial L}{\partial x_1} = 0, \quad \text{czyli} \quad -20(\hat{x}_1 - 3,5) - \hat{\lambda}_1 - 2\hat{\lambda}_2 = 0, \qquad (a)$$

$$\frac{\partial L}{\partial x_2} \leqslant 0, \quad \text{czyli} \quad -40(\hat{x}_2 - 4) - \hat{\lambda}_1 - \hat{\lambda}_2 \leqslant 0, \tag{b}$$

$$\hat{x}_2 \geqslant 0,$$

$$\hat{x}_1 \frac{\partial L}{\partial x_1} = 0, \quad \text{wobec} \quad \frac{\partial L}{\partial x_1} = 0 \quad \text{warunek nic nie wnosi,}$$

$$\hat{x}_2 \frac{\partial L}{\partial x_2} = 0, \quad \text{czyli} \quad \hat{x}_2\left[-40(\hat{x}_2 - 4) - \hat{\lambda}_1 - \hat{\lambda}_2\right] = 0, \tag{c}$$

$$\frac{\partial L}{\partial \lambda_1} \geqslant 0, \quad \text{czyli} \quad 6 - \hat{x}_1 - \hat{x}_2 \geqslant 0, \tag{d}$$

$$\frac{\partial L}{\partial \lambda_2} \leqslant 0, \quad \text{czyli} \quad 6 - 2\hat{x}_1 - \hat{x}_2 \leqslant 0, \tag{e}$$

$$\hat{\lambda}_1 \geqslant 0, \quad \hat{\lambda}_2 \leqslant 0,$$

$$\hat{\lambda}_1 \frac{\partial L}{\partial \lambda_1} = 0, \quad \text{czyli} \quad \hat{\lambda}_1 (6 - \hat{x}_1 - \hat{x}_2) = 0, \tag{f}$$

$$\hat{\lambda}_2 \frac{\partial L}{\partial \lambda_2} = 0, \quad \text{czyli} \quad \hat{\lambda}_2 (6 - 2\hat{x}_1 - \hat{x}_2) = 0. \tag{g}$$

Sporządzono tabelkę rozwiązań równań (a), (c), (f), (g), oraz skontrolowano następnie spełnienie znaków zmiennych oraz spełnienie nierówności:

	rozw. I	rozw. II	rozw. III	rozw. IV	rozw. V	rozw. VI	rozw. VII
$\hat{x}_1$	3	23/18	6	5/2	7/2	7/2	0
$\hat{x}_2$	0	31/9	0	7/2	4	0	6
$\hat{\lambda}_1$	0	0	-50	20	0	0	-230
$\hat{\lambda}_2$	5	200/9	0	0	0	0	150
$\hat{x}_2 \geqslant 0$, $\hat{\lambda}_1 \geqslant 0$, $\hat{\lambda}_2 \leqslant 0$	nie	nie	nie	tak	tak	tak	nie
$\frac{\partial L}{\partial x_2} \leqslant 0$				tak	tak	nie	
$\frac{\partial L}{\partial \lambda_1} \geqslant 0$				tak	nie		
$\frac{\partial L}{\partial \lambda_2} \leqslant 0$				tak			

Tabelka wskazuje, że warunki konieczne optymalności spełnia tylko punkt $\hat{x}_1 = 5/2$, $\hat{x}_2 = 7/2$. Jest on rozwiązaniem zadania.

2.5. Zadania typu minimax

Jako przypadek nieklasyczny programowania nieliniowego będzie rozpatrzone poszukiwanie

$$\max_{\underline{x}} \min_{\underline{z}} (\underline{x}, \underline{z}) \tag{42}$$

przy ograniczeniach

$$g_i(\underline{x}, \underline{z}) \left\{ \leqslant , = , \geqslant \right\} b_i , \qquad i = 1, \ldots m \tag{43}$$

oraz ewentualnych ograniczeniach znaku x_j, z_k.

Zadanie typu (42) (43) może powstać w zagadnieniach optymalizacji w obecności zakłóceń; $\underline{z}$ byłyby zakłóceniami, $\underline{x}$ zmiennymi decyzyjnymi. Pokazano to dalej w przykładach.

W zasadzie, zadanie (42) (43) można rozwiązać traktując najpierw $\underline{x}$ jako parametr w zadaniu następującym:

$$\min_{\underline{z}} (\underline{x}, \underline{z}) \tag{44}$$

przy

$$g_i(\underline{x}, \underline{z}) \left\{ \leqslant , = , \geqslant \right\} b_i , \qquad i = 1, \ldots m \tag{45}$$

Rozwiązanie tego zadania będzie parametryczne względem $\underline{x}$, $\hat{\underline{z}}(\underline{x})$, pozwalając z kolei sformułować drugie i ostateczne zadanie

$$\max_{\underline{x}} f\left[\underline{x}, \hat{\underline{z}}(\underline{x})\right] \tag{46}$$

przy

$$g_i\left[\underline{x}, \hat{\underline{z}}(\underline{x})\right] \left\{ \leqslant , = , \geqslant \right\} b_i , \qquad i = 1, \ldots m \tag{47}$$

Znaczne uproszczenie rozwiązania otrzymamy wówczas, gdy ograniczenia (43) będą rozdzielne, tj. gdy będzie

$$g_i(\underline{z}) \left\{ \leqslant , = , \geqslant \right\} b_i , \qquad i = 1, \ldots t, \tag{48}$$

$$g_i(\underline{x}) \left\{ \leqslant , = , \geqslant \right\} b_i , \qquad i = t+1, \ldots m. \tag{49}$$

W tym przypadku zadanie minimalizacji względem $\underline{z}$ można będzie rozwiązywać tworząc funkcję Lagrange'a

$$L_1(\underline{z}, \underline{\lambda}) = f(\underline{x}, \underline{z}) + \sum_{1}^{t} \lambda_i \left[b_i - g_i(\underline{z}) \right]$$

Warunki konieczne rozwiązania $\hat{\underline{z}}, \hat{\underline{\lambda}}$ będą miały postać (28) ... (33), czyli postać warunków koniecznych następującego punktu siodłowego

$$\max_{\underline{\lambda}} \min_{\underline{z}} L_1(\underline{z}, \underline{\lambda}). \tag{50}$$

W punkcie $[\hat{\underline{z}}, \hat{\underline{\lambda}}]$ spełniającym te warunki $L_1(\hat{\underline{z}}, \hat{\underline{\lambda}}) = f(\underline{x}, \hat{\underline{z}})$. Można zatem zadanie maksymalizacji $f(\underline{x}, \hat{\underline{z}})$ względem $\underline{x}$ przy warunkach (49) rozwiązywać z użyciem funkcji Lagrange'a jak następuje:

$$L_2(\underline{x}, \underline{\mu}) = f(\underline{x}, \hat{\underline{z}}) + \sum_{i+1}^{m} \mu_i \left[b_i - g_i(\underline{x})\right] =$$

$$= L_1(\hat{\underline{z}}, \hat{\underline{\lambda}}) + \sum_{i+1}^{m} \mu_i \left[b_i - g_i(\underline{x})\right].$$

Warunki konieczne rozwiązania $\hat{\underline{x}}, \hat{\underline{\mu}}$ będą takie, jak warunki konieczne punktu siodłowego

$$\min_{\underline{\mu}} \max_{\underline{x}} \left\{ L_1(\hat{\underline{z}}, \hat{\underline{\lambda}}) + \sum_{i+1}^{m} \mu_i \left[b_i - g_i(x)\right] \right\}.$$

Wpisując tu z kolei, że $L_1(\hat{\underline{z}}, \hat{\underline{\lambda}})$ może być wyznaczone z warunków koniecznych punktu siodłowego (50), otrzymuje się, że rozwiązanie $\hat{\underline{x}}, \hat{\underline{\mu}}, \hat{\underline{z}}, \hat{\underline{\lambda}}$ może być znalezione z warunków koniecznych punktu siodłowego następującego

$$\min_{\underline{\mu}} \max_{\underline{x}} \left\{ \max_{\underline{\lambda}} \min_{\underline{z}} \left\{ f(\underline{x}, \underline{z}) + \sum_{i}^{t} \lambda_i \left[b_i - g_i(\underline{z})\right] + \right.\right.$$

$$\left.\left. + \sum_{i+1}^{m} \mu_i \left[b_i - g_i(\underline{x})\right] \right\} \right\} \tag{51}$$

Ponieważ wyrazy $\sum_{i+1}^{m} \mu_i \left[b_i - g_i(\underline{x})\right]$ nie zależą od $\underline{\lambda}$, $\underline{z}$ można wzór (51) zapisać w postaci

$$\min_{\underline{\mu}} \max_{\underline{x}} \max_{\underline{\lambda}} \min_{\underline{z}} \left\{ L(\underline{x}, \underline{z}, \underline{\mu}, \underline{\lambda}) = f(\underline{x}, \underline{z}) + \sum_{i}^{t} \lambda_i \left[b_i - g_i(\underline{z})\right] + \right.$$

$$\left. + \sum_{i+1}^{m} \mu_i \left[b_i - g_i(\underline{x})\right] \right\}. \tag{52}$$

Wynika stąd ostatecznie, że warunkami koniecznymi rozwiązania $\hat{\underline{x}}, \hat{\underline{z}}$ jest, by $\hat{\underline{x}}, \hat{\underline{z}}, \hat{\underline{\mu}}, \hat{\underline{\lambda}}$ spełniały warunki konieczne punk-

tu siodłowego funkcji $L(\underline{x}, \underline{z}, \underline{\mu}, \underline{\lambda})$, zapisanego wzorem (52). Należy zauważyć, że jest to punkt siodłowy typu max względem $\underline{x}$, min względem $\underline{z}$. Trzeba więc tu warunki (28)...(33) stosować w odpowiedniej postaci, tj. w szczególności

$$\frac{\partial L}{\partial x_j} \leqslant 0, \qquad \hat{x}_j \geqslant 0,$$

$$\frac{\partial L}{\partial x_j} \geqslant 0, \qquad \hat{x}_j \leqslant 0,$$

$$\frac{\partial L}{\partial z_j} \geqslant 0, \qquad \hat{z}_j \geqslant 0,$$

$$\frac{\partial L}{\partial z_j} \leqslant 0, \qquad \hat{z}_j \leqslant 0,$$

$$\frac{\partial L}{\partial \mu_i} \geqslant 0, \qquad \hat{\mu}_i \geqslant 0,$$

$$\frac{\partial L}{\partial \lambda_i} \leqslant 0, \qquad \hat{\mu}_i \leqslant 0,$$

$$\frac{\partial L}{\partial \lambda_i} \leqslant 0, \qquad \hat{\lambda}_i \geqslant 0,$$

$$\frac{\partial L}{\partial \lambda_i} \geqslant 0, \qquad \hat{\lambda}_i \leqslant 0.$$

Warunki konieczne punktu siodłowego (52) będą warunkami koniecznymi i wystarczającymi rozwiązania zadania (42) (48) (49), gdy $f(\underline{x}, \underline{z})$ jest wklęsła względem $\underline{x}$ i wypukła względem $\underline{z}$, $g_i(\underline{x})$ są wypukłe względem $\underline{x}$ dla $\mu_i \geqslant 0$, wklęsłe względem $\underline{x}$ dla $\mu_i \leqslant 0$ oraz liniowe względem $\underline{x}$ dla μ_i nieograniczonego znaku oraz $g_i(\underline{z})$ są wklęsłe względem $\underline{z}$ dla $\lambda_i \geqslant 0$, wypukłe względem $\underline{z}$ dla $\lambda_i \leqslant 0$ oraz liniowe względem $\underline{z}$ dla λ_i nieograniczonego znaku.

Rozpatrzone będą teraz dwa przykłady.

Przykład 1 [14]

Dane jest zadanie sterowania optymalno-wystarczającego: znaleźć $\hat{\underline{x}}$ zapewniające $\min f(\underline{x}, \underline{z}) \geqslant \alpha$ (a) dla wszystkich zakłóceń $\underline{z}$ należących do pewnego zbioru, $\underline{z} \in Z$, oraz równocześnie $\max f(\underline{x}, \underline{z}_0)$, (b) gdzie $\underline{z}_0$ jest pewną wybraną wartością zakłócenia, $\underline{z}_0 \in Z$, na przykład wartością najbardziej prawdopodobną.

Jeśli (a) potraktować za ograniczenie nierównościowe, to zadanie sprowadza się do warunkowej maksymalizacji (b) względem $\underline{x}$.

Rozwiązuje się to metodą poszukiwania warunków punktu siodłowego funkcji Lagrange'a

$$\min_{\mu}\max_{\underline{x}}\left\{f(\underline{x},\ \underline{z}_o) + \mu\left[\alpha - \min_{\underline{z}\in Z}(\underline{x},\ \underline{z})\right]\right\} \qquad (c)$$

Minimalizacja warunkowa względem $\underline{z}$ może być zapisana przed nawiasem, bowiem $f(\underline{x},\ \underline{z}_o)$ od $\underline{z}$ nie zależy

$$\min_{\mu}\max_{\underline{x}}\min_{\underline{z}\in Z}\left\{f(\underline{x},\ \underline{z}_o) + \mu\left[\alpha - f(\underline{x},\ \underline{z})\right]\right\} \qquad (d)$$

Jeżeli ograniczenie $\underline{z} \in Z$ jest na przykład wyrażone w postaci $g(\underline{z}) \geqslant 0$, to warunki konieczne minimum względem $\underline{z} \in Z$ zastąpić można warunkami koniecznymi punktu siodłowego funkcji Lagrange'a, powstałej przez dodanie $\lambda g(\underline{z})$ do wyrażenia $\{\cdot\}$. W rezultacie rozwiązanie $\hat{\underline{x}}$, $\hat{\underline{z}}$ znajdziemy wśród $\hat{\underline{x}}$, $\hat{\underline{z}}$, $\hat{\mu}$, $\hat{\lambda}$ spełniających warunki konieczne następującego punktu siodłowego

$$\min_{\mu}\max_{\underline{x}}\max_{\lambda}\min_{\underline{z}}\left\{f(\underline{x},\ \underline{z}_o) + \mu\left[\alpha - f(\underline{x},\ \underline{z})\right] + \lambda g(\underline{z})\right\} \qquad (e)$$

Przy $f(\underline{x},\ \underline{z})$ wklęsłym względem $\underline{x}$ oraz wypukłym względem $\underline{z}$, $g(\underline{z})$ wklęsłym względem $\underline{z}$, warunki konieczne punktu siodłowego (e) będą zarazem warunkami dostatecznymi rozwiązania, gdyż wobec wymaganych wówczas $\mu \leqslant 0$ oraz $\lambda \leqslant 0$ funkcja Lagrange'a (e) będzie wklęsła względem $\underline{x}$, wypukła względem $\underline{z}$.

Rozpatrzmy przykład liczbowy. Niech

$$f(x,\ z) = -\frac{1}{2}x^2 + xz + \frac{1}{2}(z-3)^2 \geqslant 2, \qquad 0 \leqslant z \leqslant 2$$

oraz żądajmy

$$\max f(x,\ z_o), \qquad z_o = \frac{1}{2}, \qquad x \geqslant 0.$$

Zapisując obszar Z jako $z \geqslant 0$, $g(z) = 2 - z \geqslant 0$, utworzymy funkcję Lagrange'a

$$L = -\frac{1}{2}x^2 + \frac{1}{2}x + \frac{25}{8} + \mu\left[2 + \frac{1}{2}x^2 - xz - \frac{1}{2}(z-3)^2\right] + \lambda(2-z).$$

Szukamy warunków punktu siodłowego następującego

$$\min_{\mu}\max_{x}\max_{\lambda}\min_{z} L(x,\ z,\ \mu,\ \lambda),$$

zatem warunki Kuhna-Tuckera będą

$$\frac{\partial L}{\partial x} \leqslant 0, \quad \hat{x} \geqslant 0, \qquad \text{czyli} \quad -\hat{x} + \frac{1}{2} + \hat{\mu}(\hat{x} - \hat{z}) \leqslant 0,$$

$$\frac{\partial L}{\partial \mu} \leqslant 0, \quad \hat{\mu} \leqslant 0, \qquad \text{czyli} \quad 2 + \frac{1}{2}\hat{x}^2 - \hat{x}\hat{z} - \frac{1}{2}(\hat{z} - 3)^2 \leqslant 0,$$

$$\frac{\partial L}{\partial z} \geqslant 0, \quad \hat{z} \geqslant 0, \qquad \text{czyli} \quad \hat{\mu}(-\hat{x} - \hat{z} + 3) - \hat{\lambda} \geqslant 0,$$

$$\frac{\partial L}{\partial \lambda} \geqslant 0, \quad \hat{\lambda} \leqslant 0, \qquad \text{czyli} \quad 2 - \hat{z} \geqslant 0,$$

$$\hat{x}\frac{\partial L}{\partial x} = 0, \qquad \text{czyli} \quad \hat{x}\left[-\hat{x} + \frac{1}{2} + \hat{\mu}(\hat{x} - \hat{z})\right] = 0,$$

$$\hat{z}\frac{\partial L}{\partial z} = 0, \qquad \text{czyli} \quad \hat{z}\left[\hat{\mu}(\hat{x} - \hat{z} + 3) - \hat{\lambda}\right] = 0,$$

$$\hat{\mu}\frac{\partial L}{\partial \mu} = 0, \qquad \text{czyli} \quad \hat{\mu}\left[2 + \frac{1}{2}\hat{x}^2 - \hat{x}\,\hat{z} - \frac{1}{2}(\hat{z} - 3)^2\right] = 0,$$

$$\hat{\lambda}\frac{\partial L}{\partial \lambda} = 0, \qquad \text{czyli} \quad \hat{\lambda}(2 - \hat{z}) = 0.$$

Rozwiązując to zadanie, znajdziemy $\hat{x} = 1$, $\hat{z} = 2$, $\hat{\mu} = -\frac{1}{2}$, $\hat{\lambda} = 0$. Ponieważ $f(x, z)$ jest tu wklęsłe względem x, liniowe (więc i wypukłe) względem z, $g(z)$ liniowe (więc i wklęsłe) względem z, warunki konieczne punktu siodłowego są zarazem warunkami dostatecznymi rozwiązania, którym jest zatem $\hat{x} = 1$ ("najlepsze sterowanie"), $\hat{z} = 2$ ("najmniej sprzyjające zakłócenie").

Przykład 2

Znaleźć $\hat{x}$, $\hat{z}$ przy których zachodzi

$$\max_{x}\min_{z}\left[f(x, z) = \frac{(z - 1)^2}{2} - \frac{(x - 2)^2}{4} + z\,x\right]$$

przy ograniczeniach

$$x + z \leqslant 6, \qquad x \geqslant 0, \qquad z \geqslant 0.$$

Ze względu na nierozdzielne ograniczenie, zadanie to trzeba rozwiązywać drogą optymalizacji parametrycznej. Poszukuje się najpierw minimum względem z metodą Kuhna-Tuckera.

Funkcja Lagrange'a

$$L_1(z, \lambda) = \frac{(z - 1)^2}{2} - \frac{(x - 2)^2}{4} + z\,x + \lambda(6 - x - z).$$

Warunki Kuhna-Tuckera:

$$\frac{\partial L_1}{\partial z} \geqslant 0, \qquad \text{czyli} \quad \hat{z} - 1 + x - \hat{\lambda} \geqslant 0, \tag{a}$$

$$\hat{z}\,\frac{\partial L_1}{\partial z} = 0, \qquad \text{czyli} \quad \hat{z}\left[\hat{z} - 1 + x - \hat{\lambda}\right] = 0, \tag{b}$$

$$\frac{\partial L_1}{\partial \lambda} \geqslant 0, \qquad \text{czyli} \quad 6 - x - \hat{z} \geqslant 0, \tag{c}$$

$$\hat{\lambda} \leqslant 0,$$

$$\hat{\lambda}\,\frac{\partial L_1}{\partial \lambda} = 0, \qquad \text{czyli} \quad \hat{\lambda}\left[6 - x - \hat{z}\right] = 0. \tag{d}$$

Sporządzono tabelkę rozwiązań równań (b), (d), a następnie sprawdzono spełnienie warunków znaków zmiennych i spełnienie nierówności.

	rozw. I	rozw. II	rozw. III
$\hat{z}$	0	$6 - x$	$1 - x$
$\hat{\lambda}$	0	5	0
$\hat{\lambda} \leqslant 0$	tak	nie	tak
$\hat{z} \geqslant 0$	tak		gdy $x \leqslant 1$
$\frac{\partial L_1}{\partial z} \geqslant 0$	gdy $x \geqslant 1$		tak
$\frac{\partial L_1}{\partial \lambda} \geqslant 0$	gdy $x \leqslant 6$		tak

Rozpatrywana funkcja celu $f(x, z)$ jest wypukła względem z, ograniczenie jest liniowe, tak że warunki K-T są dostateczne.

Tabelka wskazuje zatem, że poszukiwane rozwiązanie $\hat{z}(x)$ ma postać: $\hat{z} = 0$ w przedziale $1 \leqslant x \leqslant 6$,

$\hat{z} = 1 - x$ w przedziale $x \leqslant 1$.

Z kolei wykonać należy maksymalizację względem x. Tworzy się funkcję Lagrange'a:

$$L_2(x, \mu) = \frac{(\hat{z} - 1)^2}{2} - \frac{(x - 2)^2}{4} + \hat{z}\,x + \mu(6 - x - \hat{z})$$

oraz korzysta z warunków K-T:

$$\frac{\partial L_2}{\partial x} \leqslant 0 \qquad \text{czyli} \qquad -\frac{1}{2}(\hat{x} - 2) + \hat{z} - \hat{\mu} \leqslant 0 \tag{e}$$

$$\hat{x}\frac{\partial L_2}{\partial x} = 0, \qquad \hat{x}\left[-\frac{1}{2}(\hat{x} - 2) + \hat{z} - \hat{\mu}\right] = 0, \tag{f}$$

$$\frac{\partial L_2}{\partial \mu} \geqslant 0, \qquad 6 - \hat{x} - \hat{z} \quad 0, \tag{g}$$

$$\hat{\mu} \geqslant 0$$

$$\hat{\mu}\frac{\partial L_2}{\partial \mu} = 0 \qquad \hat{\mu}(6 - \hat{x} - \hat{z}) = 0. \tag{h}$$

Postępując jak poprzednio, sporządzono tabelkę

	rozw. I	rozw. II	rozw. III
$\hat{x}$	0	$2(1 + \hat{z})$	$6 - \hat{z}$
$\hat{\mu}$	0	0	$\frac{1}{2}(3\hat{z} - 4)$
$\hat{\mu} \geqslant 0$	tak	tak	nie, bo $\hat{z} < 1$
$\hat{x} \geqslant 0$	tak	tak	
$\frac{\partial L_2}{\partial x} \leqslant 0$	nie	tak	
$\frac{\partial L_2}{\partial \mu} \geqslant 0$		tak	

Tabelka wskazuje, że rozwiązaniem zadania parametrycznego jest $\hat{x} = 2(1 + \hat{z})$. Zauważmy, że funkcja celu jest wklęsła względem x, zatem warunki K-T są dostateczne.

Wykorzystać teraz należy wynik poprzedniego obliczenia, podającego ekstremalizującą wartość $\hat{z}$:

$$\hat{z} = 0 \qquad \text{dla} \qquad 1 \leqslant x \leqslant 6,$$

$$\hat{z} = 1 - x \qquad x \leqslant 1,$$

Przyjmując $\hat{z} = 1 - \hat{x}$ otrzymuje się

$$\hat{x} = 2(1 + 1 - \hat{x}), \quad \text{co daje} \quad \hat{x} = 4/3.$$

Wartość $\hat{x} = 4/3$ przypada poza zakresem obowiązywania $\hat{z} = 1 - x$. Poprawne może być zatem tylko rozwiązanie $\hat{z} = 0$. Daje ono wartość $\hat{x} = 2$, leżącą w zakresie obowiązywania $\hat{z} = 0$.

Stwierdźmy jeszcze, że wartość funkcji celu w punkcie $\hat{z} = 0$, $\hat{x} = 2$ wynosi $f(\hat{x}, \hat{z}) = 1/4$. Jest to rozwiązanie zadania.

3. Programowanie liniowe

3.1. Sformułowanie problemu. Twierdzenia podstawowe

Jak już wspomniano, ogólne zadanie programowania liniowego można sformułować następująco. Wyznaczyć wektor $\underline{\hat{x}}$, który ekstremalizuje liniową funkcję celu

$$F = \sum_{j=1}^{n} c_j x_j, \tag{53}$$

pod warunkiem spełnienia zbioru liniowych ograniczeń

$$\sum_{j=1}^{n} a_{ij} x_j \left\{ \leqslant, =, \geqslant \right\} b_i \qquad i = 1, \ldots m, \tag{54}$$

$$x_j \geqslant 0 \qquad j = 1, \ldots n, \tag{55}$$

Zakłada się przy tym, że wielkości a_{ij}, b_i oraz c_j są stałe zaś $m \leqslant n$. Przyjęto również, że wszystkie b_i są nieujemne tzn. $b_i \geqslant 0$, gdyż w przeciwnym przypadku odpowiednie równanie można pomnożyć przez -1.

Zwróćmy uwagę, że powyższe sformułowanie problemu programowania liniowego jest równoważne problemowi, w którym wszystkie nierówności w zbiorze ograniczeń (54) zastąpione są równaniami.

Jak wiadomo każdą nierówność

$$a_{i1} x_1 + a_{i2} x_2 + \ldots + a_{in} x_n \leqslant b_i,$$

można zastąpić równością

$$a_{i1} x_1 + a_{i2} x_2 + \ldots + a_{in} x_n + x_{n+i} = b_i,$$

po wprowadzeniu do niej nieujemnej zmiennej dopełniającej $x_{n+i} \geqslant 0$. W rezultacie więc, zadanie programowania liniowego można wypowiedzieć:

$$\text{znaleźć} \quad \max_{\underline{x}} F = \max_{\underline{x}} \sum_{j=1}^{n} c_j x_j, \tag{56}$$

przy warunkach

$$\sum_{j=1}^{n} a_{ij} x_j = b_i \qquad i = 1,\dots m, \tag{57}$$

$$x_j \geqslant 0 \qquad j = 1,\dots n. \tag{58}$$

Tego rodzaju postać zapisu problemu programowania liniowego nazywa się postacią kanoniczną, przy czym w notacji wektorowej przedstawia się ona następująco:

$$\text{znaleźć} \quad \max_{\underline{x}} F = \max_{\underline{x}} \underline{c}^T \underline{x}, \tag{59}$$

przy warunkach

$$A\,\underline{x} = \underline{b} \tag{60}$$

$$\text{oraz} \qquad \underline{x} \geqslant 0 \tag{61}$$

gdzie: $\underline{c} = \left[x_1, c_2, \dots, c_n\right]^T$;

$\underline{x} = \left[x_1, x_2, \dots, x_n\right]^T$;

$\underline{A} = \left[\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n\right]$; $\underline{a}_j = \left[a_{1j}, a_{2j}, \dots a_{mj}\right]^T$

$\underline{b} = \left[b_1, b_2, \dots, b_n\right]^T$; $j = 1,\dots n.$

Funkcję celu o postaci (53) bądź (59) nazwano formą liniową, wektory $\underline{a}_j$, $j = 1,\dots n$, wektorami warunków, natomiast wektor $\underline{b}$ - wektorem ograniczeń.

Nieujemne współrzędne wektora $\underline{x} = \left[x_1, x_2, \dots x_n\right]^T$ spełniające warunki (60) i (61) tworzą obszar rozwiązań dopuszczalnych zadania programowania liniowego. Trzeba zauważyć, że istnieje co najmniej jedno rozwiązanie równania (60) jeżeli rząd *) rozszerzonej macierzy $A_r = \left[A, \underline{b}\right]$ tzn. utworzonej z macierzy A, do której dołączono kolumnę $\underline{b}$, jest równy rzędowi A, czyli $r(A_r) = r(A)$. Jeśli tak nie jest, a więc $r(A_r) > r(A)$, wówczas wektor $\underline{b}$ nie może zostać wyrażony jako liniowa kombinacja kolumn macierzy A, a tym samym nie istnieje rozwiązanie równań (60).

W dalszych rozważaniach przyjęto więc następujące założenia: $r(A_r) = r(A)$ oraz $r(A) = m$. Ostatnie założenie oznacza, że mamy co najmniej tyle zmiennych, ile równań tzn. $n \geqslant m$. Jednakże

*) Przez "rząd macierzy" rozumie się maksymalną ilość liniowo niezależnych kolumn (czy też wierszy) tej macierzy.

przypadek $n = m$ nie jest interesujący, gdyż nie ma potrzeby rozwiązywania wówczas zadania optymalizacji ze względu na istnienie jednoznacznego rozwiązania układu (60). Tak więc, w praktyce będziemy rozpatrywać tylko zadania optymalizacji gdy $n > m$. Przy Przypadek $n < m$ nic nowego nie wnosi, bowiem niektóre z równań mogą być wtedy pominięte jako zbędne.

Rozwiązaniem bazowym równania (60) przy $r(A) = m$ oraz $n > m$, nazywamy rozwiązanie równania

$$B\,\underline{x}_B = \underline{b}, \tag{62}$$

w którym nieosobliwą macierz kwadratowa B o wymiarach $(m \times m)$ zwaną bazą *), jest utworzona z niezależnych liniowo kolumn macierzy A, przy czym $\underline{x}_B = B^{-1}\underline{b}$ nazywany jest wektorem bazowym. W przypadku gdy wszystkie zmienne $\underline{x}_B$ są nieujemne $x_B \geqslant 0$ to rozwiązanie bazowe staje się rozwiązaniem dopuszczalnym. Maksymalna ilość rozwiązań bazowych wynosi $n!/m!(n-m)!$. Geometrycznie rozwiązania bazowe odpowiadają wierzchołkom wielościanu warunków tworzącego obszar rozwiązań dopuszczalnych.

Zadanie programowania liniowego nazywa się niezdegenerowanym, jeśli każde jego rozwiązanie bazowe zawiera dokładnie m dodatnich składowych. W niniejszej pracy będą rozpatrywane jedynie zadania niezdegenerowane.

Rozwiązaniem optymalnym zadania programowania liniowego nazywa się wektor $\hat{\underline{x}} = \left[\hat{x}_1, \hat{x}_2, \ldots \hat{x}_n\right]^T$, spełniający warunki zadania i określający ekstremum formy liniowej. Inaczej mówiąc, rozwiązaniem optymalnym nazywa się rozwiązanie bazowe, przy którym forma liniowa (funkcja celu) osiąga ekstremum.

Przytoczymy teraz bez dowodów zestaw najważniejszych twierdzeń używanych na kolejnych etapach rozwiązania ogólnego problemu programowania liniowego. Odpowiednie dowody można znaleźć w pracach [13], [25]. Twierdzenia te brzmią następująco:

1. Zbiór wszystkich rozwiązań zagadnienia programowania liniowego jest zbiorem wypukłym K.

2. Jeśli K jest wielościanem wypukłym ograniczonym, to każdy punkt $\underline{x}$, będący kombinacją liniową wypukłą punktów wierzchołkowych zbioru K, należy do zbioru K.

3. Jeśli układ wektorów $\underline{a}_1, \underline{a}_2, \ldots \underline{a}_k$ jest liniowo niezależny tak, że

$$\underline{a}_1 x_1 + \underline{a}_2 x_2 + \ldots + \underline{a}_k x_k = \underline{b}; \qquad x_j \geqslant 0, \quad j = 1, \ldots k,$$

*) Bazą przestrzeni nazywamy taki zbiór liniowo niezależnych wektorów tej przestrzeni, że dowolny inny wektor stanowi kombinację liniową wektorów tego zbioru.

to punkt $\underline{x} = [x_1, x_2, \ldots x_k, \quad 0, \ldots 0]^T$ którego n-k współrzędne są równe zero) stanowi punkt wierzchołkowy zbioru wypukłego K rozwiązań dopuszczalnych układu (60) i (61).

4. Jeśli $\underline{x} = [x_1, x_2, \ldots x_n]^T$ jest punktem wierzchołkowym zbioru K, to wektory odpowiadające dodatnim x_j tworzą układ liniowo niezależny. Wynika stąd, że punkt wierzchołkowy ma nie więcej niż m dodatnich współrzędnych x_j. A więc, każdemu punktowi wierzchołkowemu ze zbioru K odpowiada m liniowo niezależnych wektorów z danego układu.

5. Forma liniowa zagadnienia programowania osiąga swoje maksimum (minimum) w punktach wierzchołkowych ograniczonego obszaru wypukłego K będącego zbiorem rozwiązań tego zagadnienia. Jeśli forma liniowa przyjmuje wartości maksymalne więcej niż w jednym punkcie wierzchołkowym, to osiąga ona te same wartości w dowolnym punkcie, stanowiącym kombinację liniową wypukłą tych punktów.

3.2. Interpretacja geometryczna zadania programowania liniowego

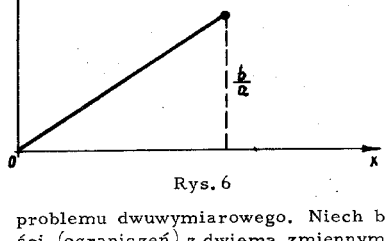
Rozpatrzmy na wstępie problem jednowymiarowy. Znaleźć maksimum formy liniowej o postaci

$$F = c\,x,$$

przy warunkach

$$a\,x \leqslant b, \qquad x \geqslant 0, \qquad (a,\ b > 0).$$

Przypadek ten ma prostą interpretację geometryczną przedstawioną na rys. 6.



Rys. 6

Jak wynika z postawionego zadania, część wspólna (przekrój) zbiorów określonych przez nierówności $a\,x \leqslant b$ i $x \geqslant 0$ stanowi odcinek na osi x-ów łącznie z punktami brzegowymi: $x = 0$ i $x = \frac{b}{a}$. Forma liniowa F osiąga swoje wartości ekstremalne na końcach odcinka, na którym, zgodnie z warunkami zadania jest ona określona.

Przejdziemy obecnie do problemu dwuwymiarowego. Niech będzie dany układ m nierówności (ograniczeń) z dwiema zmiennymi

$$a_{11} x_1 + a_{12} x_2 \leq b_1,$$
$$a_{21} x_1 + a_{22} x_2 \leq b_2, \qquad (63)$$
$$\cdots\cdots\cdots\cdots\cdots$$
$$a_{m1} x_1 + a_{m2} x_2 \leq b_m,$$

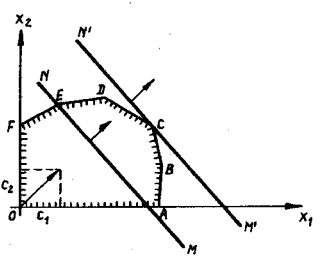
gdzie: $x_1 \geq 0, \quad x_2 \geq 0,$ (64)

szukamy maksimum

$$F_{max} = c_1 x_1 + c_2 x_2. \qquad (65)$$

Obszar zmienności formy liniowej (65) przedstawia wielokąt pokazany na rys. 7.

Proste tworzące wielokąt OABCDEFO na płaszczyźnie $x_1 0 x_2$ odpowiadają warunkom (63) i (64), w których nierówności zostały zastąpione równościami. Kreskowaniem odpowiedniej strony boków wielokąta pokazana jest ta część płaszczyzny, w której leżą punkty spełniające nierówności (63) i (64).



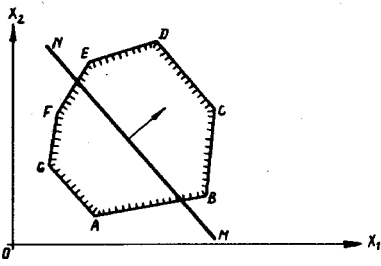
Rys. 7

Kierunek prostej MN określony jest przez wektor $[c_1, c_2]$ prostopadły do MN. Wektor ten wskazuje kierunek, w którym forma liniowa zwiększa swoje wartości.

Interpretacja geometryczna rozpatrywanego zagadnienia (dla n = 2) może być następująca. Obszar określenia formy liniowej zwany wielokątem warunków przetniemy prostą $F = c_1 x_1 + c_2 x_2$ tzn. prostą MN i będziemy ją przesuwać równolegle w kierunku wzrostu F (jeśli szukamy maksimum formy liniowej) lub w kierunku zmniejszania się F (gdy poszukujemy minimum). Istnieją przy tym różne możliwości.

W przypadku, przedstawionym na rys. 7, równoległe przesunięcie prostej MN doprowadzi ją do położenia M'N', w którym ma tylko jeden punkt wspólny z wielokątem warunków - punkt C. Punkt ten wyznacza jedyne rozwiązanie zagadnienia programowania liniowego. Jeśli natomiast prosta MN byłaby równoległa do jednego lub dwóch boków wielokąta, to ekstremum byłoby osiągane we wszystkich punktach odpowiedniego boku. Przypadek ten został pokazany na rys. 8, na którym we wszystkich punktach boku CD osiągane jest maksimum, zaś we wszystkich punktach boku AG

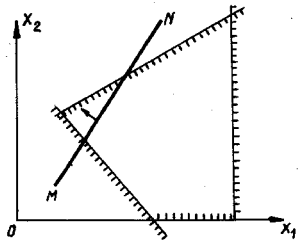
minimum formy liniowej. Wynika stąd, że zagadnienie programowania liniowego może mieć jedno lub też nieskończoną liczbę rozwiązań.



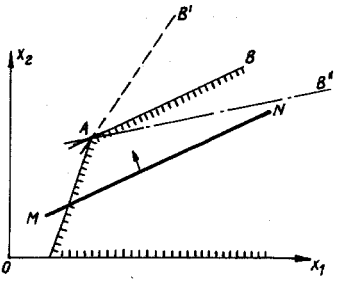
Rys. 8

Na rys. 9 przedstawiono przypadek gdy zadanie programowania liniowego jest nierozwiązalne w rezultacie sprzecznie sformułowanych warunków ograniczających. Natomiast na rys. 10 przypadek gdy obszar określenia formy liniowej jest nieograniczony.

W tym drugim przypadku, jeśli prosta AB || MN to forma liniowa osiąga skończone ekstremum we wszystkich punktach promienia AB. Jeśli zaś będziemy zmieniać obszar określenia formy liniowej, obracając np. promień AB dookoła punktu A, to mogą zaistnieć następujące dwie sytuacje.



Rys. 9



Rys. 10

W pierwszej forma liniowa staje się nieograniczona przy dopuszczalnych wartościach zmiennych, co odpowiada położeniu promienia AB'. W drugiej forma liniowa osiąga maksimum w jednym punkcie A - położenie promienia AB''.

Rozszerzymy teraz interpretację geometryczną Zadania Programowania Liniowego ZPL na przypadek dowolnej liczby zmiennych i nierówności. Podobnie jak to miało miejsce dla problemu dwuwymiarowego układ nierówności (54) i (55) określa zbiór rozwiązań ZPL, który w przestrzeni n-wymiarowej tworzy wielościan wypukły K. Wymiar tego wielościanu nie przewyższa n-m, gdyż

należy on do wspólnej części m hiperpłaszczyzn odpowiadających niezależnym liniowo równaniom układu. Forma liniowa F o postaci (53) określa w przestrzeni rodzinę równoległych hiperpłaszczyzn. Współczynniki tej formy wyznaczają wektor $\underline{c} = [c_1, c_2, \ldots, c_n]^T$ wskazujący kierunek wzrostu F. Wektor $\underline{c}$ jest prostopadły do rozważanej rodziny hiperpłaszczyzn. Wychodząc z pewnej hiperpłaszczyzny należącej do tej rodziny i mającej wspólne punkty z wielościanem K (wartości formy liniowej we wszystkich tych punktach są jednakowe), przy przesuwaniu jej równolegle w kierunku wzrostu (zmniejszenia) F, można dojść do takiego jej położenia, że przy dalszym jej przesuwaniu nie będzie ona miała punktów wspólnych z wielościanem K. Wielościan K położony jest wówczas z jednej strony otrzymanej granicznej phiperpłaszczyzny i ma z nią bądź to nieskończenie wiele punktów wspólnych, z których każdy nadaje formie liniowej F ekstremalną wartość, bądź też tylko jeden punkt wspólny. Punkt ten jest wtedy punktem wierzchołkowym wielościanu, w którym zagadnienie programowania liniowego posiada jedyne rozwiązanie.

3.3. Interpretacja ekonomiczna zadania programowania liniowego

Przy rozpatrywaniu zagadnień ekonomicznych stosuje się zazwyczaj następującą terminologię. Wektory

$$\underline{a}_j = [a_{1j}, a_{2j}, \ldots a_{mj}]^T \quad \text{oraz} \quad \underline{b} = [b_1, b_2, \ldots b_m]^T$$

$$j = 1, 2, \ldots n$$

zwane wektorami warunków i ograniczeń, nazywane są wektorami nakładów i zapasów odpowiednio. Współrzędne wektorów nakładów $\underline{a}_j$ określają zużycie poszczególnych środków produkcji na jednostkę czasu dla danego wyjściowego sposobu produkcji, a współrzędne wektora ograniczeń $\underline{b}$ - zapasy poszczególnych środków ograniczające ich zużycie.

Zagadnienia najlepszego wykorzystania zapasów w drodze optymalnego ich rozdziału według charakteru wykorzystania można podzielić na trzy grupy:

a) zagadnienia planowania produkcji (optymalne wykorzystanie mocy produkcyjnych),

b) zagadnienia sporządzania mieszanek; racjonalny podział materiałów (optymalne wykorzystanie surowców i materiałów),

c) zagadnienia transportowe (optymalny plan przewozów).

Do pierwszej grupy zagadnień zalicza się racjonalny rozdział obciążenia obrabiarek, optymalne rozdzielenie określonego programu na poszczególne warsztaty, ustalenie optymalnego obciążenia

sprzętu przy danym asortymencie, wyznaczanie optymalnego asortymentu produkcji, planowanie płodozmianu, rozmieszczanie maszyn wg rodzaju prac rolnych i inne.

Do zadań drugiej grupy zaliczamy racjonalne cięcie materiałów przemysłowych, określenie optymalnej mieszanki, stopu, diety, dobór najtańszych i najpożywniejszych racji żywnościowych dla zwierząt hodowlanych, wykorzystanie surowca i inne.

Do zagadnień transportowych zaliczamy: optymalne powiązanie punktów przeznaczenia z punktami odprawy przy przewozie ładunków, racjonalne rozdzielenie środków transportowych, wyznaczanie optymalnego planu przewozów itp.

Przytoczymy teraz szereg przykładów zaczerpniętych z książki R. Kulikowskiego [36].

1. Problem przydziału maszyn

Rozważmy zakład produkcyjny mający m maszyn i produkujący n wyrobów, przy czym wprowadźmy następujące oznaczenia:

a_{ij} - ilość czasu potrzebnego do produkcji jednostki produktu j na maszynie i,

x_{ij} - ilość produktu j wytwarzanego na maszynie i w danym okresie czasu,

a_i - dysponowany czas pracy maszyny i,

b_j - ilość produktu j, który powinien być wytworzony,

c_{ij} - koszt wytwarzania jednostki produktu j na maszynie i.

Problem przydziału maszyn polega na wyznaczeniu nieujemnych wartości x_{ij} spełniających warunki

$$\sum_{j=1}^{n} a_{ij} x_{ij} \leqslant a_i; \qquad i = 1, \ldots, m, \tag{66}$$

$$\sum_{j=1}^{n} x_{ij} = b_j; \qquad j = 1, \ldots, n \tag{67}$$

oraz minimalizujących wskaźnik jakości

$$F = \sum_{j=1}^{n} \sum_{i=1}^{n} c_{ij} x_{ij}. \tag{68}$$

2. Problem optymalnego mieszania surowców

Wprowadźmy oznaczenia:

a_{ij} - zawartość j-tego składnika w jednostce i-tego produktu,

b_j - wymagana zawartość j-tego składnika,

c_i - cena jednostki i-tego produktu.

Problem optymalnego mieszania surowców sprowadza się do wyznaczenia nieujemnych ilości produktów x_i, które minimalizują

całkowity koszt

$$F = \sum_{i=1}^{n} c_i x_i, \tag{69}$$

przy warunkach

$$\sum_{i=1}^{n} a_{ij} x_i \geqslant b_j; \qquad j = 1, \ldots, m. \tag{70}$$

3. Zagadnienie transportowe

Dany jest system m punktów nadawczych wysyłających jednolity produkt do n punktów odbiorczych.

Oznaczono:

x_{ij} - ilość produktu wysyłanego z punktu i-tego do punktu j-tego,
a_i - ilość produktu, którym dysponuje punkt i,
b_j - ilość produktu, którego wymaga punkt j,
c_{ij} - koszt przesłania jednostki produktu z punktu i do punktu j.

Zakładając, że całkowita ilość produktu w punktach nadawczych $\sum_{i=1}^{m} a_i$ jest nie mniejsza niż całkowita ilość produktu wymaganego w punktach odbiorczych $\sum_{j=1}^{n} b_j$ tzn.

$$\sum_{i=1}^{n} a_j \geqslant \sum_{j=1}^{m} b_j, \tag{71}$$

zagadnienie transportowe można formułować następująco: znaleźć takie wartości x_{ij}, które minimalizują całkowite koszty transportowe

$$F = \sum_{j=1}^{n} \sum_{i=1}^{m} c_{ij} x_{ij}, \tag{72}$$

pod warunkiem

$$\sum_{j=1}^{n} x_{ij} \leqslant a_i; \qquad i = 1, \ldots, m, \tag{73}$$

$$\sum_{i=1}^{m} x_{ij} = b_j; \qquad j = 1, \ldots, n, \tag{74}$$

$$x_{ij} \geqslant 0; \qquad i = 1, 2, \ldots, m; \quad j = 1, \ldots, n \tag{75}$$

3.4. Metody rozwiązywania zadania programowania liniowego

Istnieją dwie zasadnicze metody rozwiązania zadań programowania liniowego: pierwsza - graficzna, druga - algorytmiczna posługująca się algebrą liniową.

Metoda graficzna może być używana w praktyce tylko wtedy, gdy w zadaniu występują dwie lub ewentualnie trzy zmienne. Przy większej liczbie zmiennych rozwiązanie zadania tak się komplikuje, że w zasadzie staje się ono niewykonalne. Ze względu jednak na poglądowość tej metody zostanie ona omówiona oddzielnie w dalszej części pracy.

W konkretnych zastosowaniach posługujemy się więc metodami algorytmicznymi, których istnieje wielka różnorodność. Jako główne można wymienić metody Dantziga (simplex), Frischa, Dorfmana, Samuelsona itp. Wszystkie wymienione metody wiążą się bezpośrednio lub pośrednio z przedstawioną poprzednio interpretacją geometryczną problemu programowania liniowego. Polegają one na wykryciu najwyżej (lub najniżej) położonego wierzchołka obszaru (wielościanu) dopuszczalnych rozwiązań metodą kolejnych interacji tzn. przez stopniowe przechodzenie od niższych do wyższych wierzchołków tego obszaru.

W niniejszej pracy poprzestaniemy na omówieniu tylko jednej metody - klasycznej już dziś metody simpleks.

3.4.1. Metoda graficzna

Przy graficznym rozwiązywaniu zagadnień programowania liniowego najważniejszą sprawą jest wyznaczenie wielokąta rozwiązań dopuszczalnych. Następnie należy geometrycznie określić wartości formy liniowej w punktach wierzchołkowych. Mogą przy tym zaistnieć dwa warianty:

a) spośród otrzymanych liczb jedna z liczb okaże się większa od pozostałych,

b) spośród otrzymanych liczb dwie liczby okażą się równe sobie, a przy tym większe od pozostałych.

W przypadku pierwszym istnieje tylko jeden punkt optymalny, któremu odpowiada największa wartość. W przypadku drugim istnieje nieskończony zbiór punktów położonych na odcinku i wszystkim tym punktom odpowiada wartość optymalna.

Rozpatrzmy na przykładzie metodę graficznego rozwiązywania zadania programowania liniowego.

Przykład

Niech dany będzie układ nierówności

$$3x_1 + 5x_2 \leqslant 15,$$
$$5x_1 + 2x_2 \leqslant 10,$$

$$x_1 \geqslant 0, \quad x_2 \geqslant 0.$$

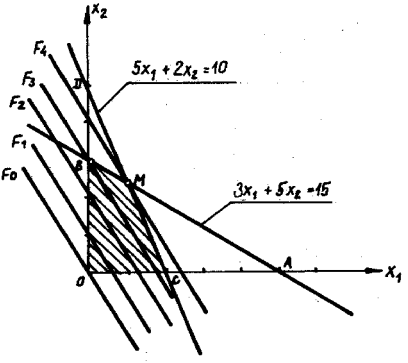
Znaleźć $F_{max} = 5x_1 + 3x_2$.

Sposób rozwiązania tego problemu przedstawiono graficznie na rys. 11.

Zakreskowane pole OCMB stanowi obszar rozwiązań dopuszczalnych. Przyjmując rozmaite wartości na F np. 0, 1, 3 itp. otrzymuje się rodzinę linii prostych równoległych. Jedna z tych prostych przechodząca przez wierzchołek M wielokąta wyznacza poszukiwane rozwiązanie. Współrzędne punktu M określa się rozwiązując układ równań

$$3x_1 + 5x_2 = 15,$$

$$5x_1 + 2x_2 = 10.$$



Rys. 11

W rezultacie mamy $\hat{x}_1 = 1,053$; $\hat{x}_2 = 2,368$ oraz $F_{max} = 12,37$.

3.4.2. Metoda Simpleks

Jak już wspomniano metoda simpleks jest metodą iteracyjną tzn. przez wykonanie szeregu kroków dochodzimy do rozwiązania optymalnego. W pierwszym kroku znajdujemy rozwiązanie bazowe, a następnie sprawdzamy czy nie zostało otrzymane rozwiązanie optymalne. Jeśli nie, to w drugim kroku usuwamy z bazy jeden z wektorów i na jego miejsce wprowadzamy inny. W ten sposób otrzymujemy nową bazę, dla której ponawiamy czynności kroku pierwszego. Jeśli otrzymane rozwiązanie nie jest optymalne, to dalej postępujemy podobnie jak w kroku drugim. Zadanie sprowadza się wobec tego do znalezienia dowolnego rozwiązania bazowego, a następnie ulepszenia go dotąd dopóki nie osiągnie się rozwiązania optymalnego. Rozwiązanie to znajdujemy po skończonej liczbie kroków, pod warunkiem, że ograniczenia nie są sprzeczne, bądź też wartość formy liniowej nie jest nieskończona.

Geometrycznie rozwiązanie zagadnienia programowania liniowego metodą simpleksów oznacza, że poczynając od określonego wierzchołka wielościanu w następnym kroku wybieramy wierzcho-

łek, który jest położony bliżej rozwiązania optymalnego. A więc stopniowo przybliżamy się do tego rozwiązania, aż zostanie ono osiągnięte.

Przed przystąpieniem do omawiania algorytmu metody simpleks przedstawimy sposób przechodzenia z jednej bazy do drugiej, sformułujemy kryterium zakończenia działania procedury oraz kryterium, według którego dokonuje się wyboru następnej ulepszonej bazy. Wybór ten sprowadza się do określenia nowego wektora bazowego, który należy wprowadzić do istniejącej bazy oraz do podjęcia decyzji, który z wektorów należy z niej usunąć.

Niech w danym zbiorze n wektorów $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_n$ znajduje się m wektorów jednostkowych, których współrzędne tworzą macierz jednostkową m-tego stopnia. Nie naruszając ogólności rozważań, można przyjąć, że takimi wektorami są pierwsze m wektory $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_m$, które tworzą bazę wyjściową $B = E = \underline{a}_1, \underline{a}_2, \ldots, \underline{a}_m$. Ponieważ $E^{-1} = E$, to wyjściowe rozwiązanie bazowe będzie miało postać:

$$\underline{x}_B = \underline{b},$$

gdzie

$$\underline{x}_B = \left[x_{B1}, x_{B2}, \ldots, x_{Bm}\right]^T; \quad x_{Bi} \geqslant 0,$$

natomiast wartość formy liniowej będzie wynosić

$$F_o = \sum_{i=1}^{m} c_i x_{Bi}. \tag{76}$$

Zauważmy ponadto, że dowolna kolumna $\underline{a}_j$ macierzy A może być wyrażona kombinacją liniową kolumn macierzy B, a więc

$$\underline{a}_j = B\underline{a}_j = \underline{a}_1 y_{1j} + \underline{a}_2 y_{2j} + \ldots + \underline{a}_m y_{mj}, \tag{77}$$

gdzie wektor kolumnowy $\underline{y}_j = \left[y_{1j}, y_{2j}, \ldots, y_{mj}\right]^T$, przy czym

$$\underline{y}_j = B^{-1} \underline{a}_j \quad \text{dla} \quad j = 1,2,\ldots,n. \tag{78}$$

Przypuśćmy, że został dokonany wybór nowego wektora bazowego $\underline{a}_k$, który następnie chcemy wprowadzić do bazy na miejsce wektora $\underline{a}_r$. Stąd, nowa bazowe rozwiązanie będzie posiadać bazę składającą się z wektorów $\underline{a}_1, \ldots, \underline{a}_{r-1}, \underline{a}_{r+1}, \ldots, \underline{a}_m, \underline{a}_k$. W celu wyznaczenia tego rozwiązania, musimy dokonać przejścia od jednej bazy do drugiej.

Zgodnie z przyjętym założeniem bazę wyjściową stanowi macierz

$$B = E = \left[\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_m\right],$$

a więc, możemy napisać

$$\underline{b} = \underline{a}_1 x_{B1} + \ldots + \underline{a}_r x_{Br} + \ldots + \underline{a}_m x_{Bm}, \tag{79}$$

$$\underline{a}_k = \underline{a}_1 y_{1k} + \ldots + \underline{a}_r y_{rk} + \ldots + \underline{a}_m y_{mk}, \tag{80}$$

oraz

$$\underline{a}_j = \underline{a}_1 y_{1j} + \ldots + \underline{a}_r y_{rj} + \ldots + \underline{a}_m y_{mj}, \tag{81}$$

ale z równania (80) mamy

$$\underline{a}_r = \frac{1}{y_{rk}} (\underline{a}_k - \underline{a}_1 y_{1k} \ldots - \underline{a}_m y_{mk}), \tag{82}$$

podstawiając wyrażenie (82) do wzoru (79) otrzymujemy

$$\underline{b} = \underline{a}_1 x_{B1} + \ldots + x_{Br} \frac{1}{y_{rk}} \left[\underline{a}_k - \underline{a}_1 y_{ik} - \ldots - \underline{a}_m y_{mk}\right] +$$

$$+ \ldots + \underline{a}_m x_{Bm},$$

a po uporządkowaniu

$$\underline{b} = \left(x_{B1} - \frac{x_{Br}}{y_{rk}} y_{1k}\right) \underline{a}_1 + \ldots + \frac{x_{Br}}{y_{rk}} \underline{a}_k + \ldots +$$

$$+ \left(x_{Bm} - \frac{x_{Br}}{y_{rk}} y_{mk}\right) \underline{a}_m. \tag{83}$$

Ze wzoru (83) wynika, że nowe bazowe rozwiązanie

$$\underline{x}'_B = \left[x'_{B1}, x'_{B2}, \ldots, x'_{Bk}, \ldots, x'_{Bm}\right]^T$$

$$x'_B \geqslant 0$$

układu równań

$$\underline{b} = \underline{a}_1 x'_{B1} + \ldots + \underline{a}_k x'_{Bk} + \ldots + \underline{a}_m x'_{Bm},$$

należy wyliczać ze wzorów

$$x'_{Bi} = x_{Bi} - y_{ik} \frac{x_{Br}}{y_{rk}} ,$$

dla $i = 1,2,\ldots,r-1, \quad r+1,\ldots,m$ (84)

oraz $x'_{Bk} = \frac{x_{Br}}{y_{rk}}$

Podstawiając (82) do (81) w podobny sposób otrzymujemy wzory na rozkład dowolnego wektora a_j według nowo znalezionej bazy. Tak więc

$$\underline{a}_j = \underline{a}_1 y_{ij} + \ldots + \underline{a}_k y_{kj} + \ldots + \underline{a}_m y_{mj},$$

gdzie

$$y'_{ij} = y_{ij} - y_{ik} \frac{y_{rj}}{y_{rk}} \quad \text{dla} \quad i \neq j , \qquad (85)$$

$$y'_{kj} = \frac{y_{rj}}{y_{rk}} .$$

Korzystając z wyprowadzonych wzorów zbadajmy teraz jak się zmieni wartość formy liniowej F_o przy przejściu z jednej bazy do drugiej.

W tym celu do wyrażenia

$$F'_o = c_1 x'_{B1} + \ldots + c_k x'_{Bk} + \ldots + c_m x'_{Bm}, \qquad (86)$$

reprezentującego wartość formy liniowej w nowo znalezionym rozwiązaniu bazowym podstawmy wzory (84), skąd po uporządkowaniu wyrazów mamy

$$F'_o = F_o - \frac{x_{Br}}{y_{rk}} \sum_{i=1}^{m} c_i y_{ik} + \frac{x_{Br}}{y_{rk}} c_k, \qquad (87)$$

oznaczając przez

$$F_k = \sum_{i=1}^{m} c_i y_{ik}, \qquad (88)$$

$$y_{ro} = x_{Br} \quad \text{oraz} \quad \varrho_{min} = \frac{y_{ro}}{y_{rk}},$$

ostatecznie otrzymujemy

$$F'_o = F_o - (F_k - c_k)\varrho_{min}, \qquad (89)$$

przy czym

$$\varrho_{min} \geq 0.$$

Z zależności (89) wynika bezpośrednio sens następujących trzech kryteriów, na których oparta jest metoda simpleks:

1. Warunkiem koniecznym i dostatecznym na to, aby rozwiązanie bazowe $\underline{x}_k$ było rozwiązaniem optymalnym jest spełnienie nierówności $F_j - c_j \geq 0$ dla wszystkich j, gdzie $F_j = \sum_{i=1}^{m} c_i y_{ij}$,
2. Jeśli w zbiorze $J_o = \left\{ j \mid F_j - c_j < 0; \quad \text{dla} \quad j = 1,2,\ldots,n \right\}$ istnieje takie $j = k$, że

$$F_k - c_k = \min \left\{ (F_j - c_j) \right\} \tag{90}$$

oraz $y_{ik} > 0$ przynajmniej dla jednego $i = 1,2,\ldots,m$, to przejście do nowej bazy spełniającej warunek $F_o' > F_o$ jest możliwe przez wprowadzenie do starej bazy wektora a_k na miejsce wektora $\underline{a}_r$, przy czym wskaźnik r określony jest z warunku

$$\varrho_r = \min \left\{ \varrho_i \right\} \quad \text{dla} \quad i \in I_o, \tag{91}$$

gdzie $\varrho_i = \dfrac{y_{io}}{y_{ik}}$ oraz $I_o = \left\{ i \mid y_{ik} > 0 \quad \text{dla} \quad i = 1,2,\ldots m \right\}$

3. Jeśli w zbiorze $J_o \left\{ j \mid F_j - c_j < 0; \quad \text{dla} \quad j = 1,2,\ldots n \right\}$ nie istnieje takie j, że $y_{ij} > 0$ przynajmniej dla jednego $i = 1,2,\ldots m$, to zadanie programowania liniowego nie posiada rozwiązania, gdyż wartość formy liniowej rośnie nieograniczenie.

Odpowiednie twierdzenia, z których wynikają przytoczone kryteria, wraz z dowodami można znaleźć w pracach [13], [25].

Przejdźmy teraz do omówienia algorytmu Simpleks.

Algorytm metody Simpleks

Rozważymy metodę simpleks dla niezdegenerowanego zagadnienia programowania liniowego o postaci:
znaleźć maksymalną wartość formy liniowej

$$F_{max} = \sum_{j=1}^{p} c_j \, x_j, \tag{92}$$

pod warunkiem spełnienia ograniczeń

$$\sum_{j=1}^{p} a_{ij} \, x_j \left\{ \begin{matrix} \leqq \\ \geqq \end{matrix} \right\} b_i \qquad i = 1,2,\ldots,m \tag{93}$$

$$x_j \geqslant 0 \qquad j = 1,2,\ldots,p \tag{94}$$

Algorytm metody Simpleks przebiega w następujący sposób:

a. W pierwszym kroku ogólne zagadnienie programowania liniowego sprowadzamy do postaci kanonicznej, przy czym sprawdzamy czy wszystkie b_i są nieujemne. A więc,
- po pierwsze wszystkie nierówności (93), w których występuje ujemne b_i mnożymy przez -1,
- po drugie przez wprowadzenie zmiennych dopełniających, każdą nierówność (93) przekształcamy w równanie, a mianowicie:

$$\sum_{j=1}^{p} a_{ij} x_j \leqslant b_i \quad \text{przechodzi w} \quad \sum_{j=1}^{p} a_{ij} x_j + x_{p+1} = b_i,$$

$$x_{p+i} \geqslant 0$$

zaś

$$\sum_{j=1}^{p} a_{ij} x_j \geqslant b_i \quad \text{przechodzi w} \quad \sum_{j=1}^{p} a_{ij} x_j - x_{p+i} = b_i.$$

$$x_{p+i} \geqslant 0$$

Zauważmy przy tym, że jeśli zmienną dopełniającą dodajemy do i-tej nierówności, to kolumna macierzy A, odpowiadająca zmiennej x_{p+i}, jest równa wektorowi jednostkowemu e_i, jeśli zaś zmienną dopełniającą odejmujemy od i-tej nierówności to kolumna macierzy A odpowiadająca x_{p+i} będzie równa $-e_i$.

b. W drugim kroku znajdujemy pierwsze rozwiązanie podstawowe. W tym celu, z macierzy A wybieramy macierz jednostkową, której wektory zapisujemy w kolumnie "wektory bazowe". Zwykle współczynniki przy zmiennych dopełniających x_{p+i} tworzą taką macierz jednostkową, której wyznacznik jest równy jedności. Przeto wektory $\underline{a}_{p+1}, \underline{a}_{p+2}, \ldots, \underline{a}_{p+m}$ są liniowo niezależne i tworzą bazę. Wówczas rozwiązanie podstawowe stanowi wektor

$$\underline{x}_B = \left[0, 0, \ldots, 0, \quad x_{p+1}, x_{p+2}, \ldots, x_{p+m}\right]^T.$$

W przypadku, gdy w macierzy A nie można otrzymać macierzy jednostkowej, wtedy dla jej utworzenia dodajemy niezbędną liczbę wektorów sztucznych q_i i sztucznych zmiennych x_{p+i}. Ponadto w formie liniowej jako wartość współczynników przy tych sztucznych zmiennych przyjmuje się dużą liczbę ujemną -M (przy poszukiwaniu maksimum) lub dodatnią +M (przy poszukiwaniu minimum).

c. W trzecim kroku tworzymy pierwszą tablicę simpleksów. Typową jej postać przedstawiono w tablicy 1. W tablicy tej nie naruszając ogólności rozważań przyjęto, że pierwsze m wektorów $\underline{a}_j$ tworzy bazę, przy czym w odpowiednie jej kratki wpisano elementy macierzy A oraz wektora $\underline{b}_i$ przyjmując oznaczenia

$$y_{io} = b_i,$$

$$y_{ij} = a_{ij}.$$

Elementy wiersza wskaźnikowego tzn. wiersza, w którym zapisujemy F_o oraz $F_j - c_j$, oblicza się z następujących wzorów:

$$F_o = \underline{c}_b^T \underline{b}, \qquad (95)$$

$$F_j - \underline{c}_j = \underline{c}_b^T \underline{a}_j - \underline{c}_j,$$

gdzie $\underline{c}_b$ oznacza wektor, którego współrzędne odpowiadają wektorom bazowym $\underline{a}_j$.

Tablica 1

Tablica Simpleksów

$\underline{c}_b$	Wektory bazowe	c_j	c_1	$c_2 \cdots$	$c_j \cdots$	$c_k \cdots$	c_n
		$\underline{b}$	$\underline{a}_1$	$\underline{a}_2 \cdots$	$\underline{a}_j \cdots$	$\underline{a}_k \cdots$	$\underline{a}_n$
c_1	$\underline{a}_1$	$x_1 = y_{10}$	y_{11}	$y_{12} \cdots$	$y_{1j} \cdots$	$y_{1k} \cdots$	y_m
c_2	$\underline{a}_2$	$x_2 = y_{20}$	y_{21}	$y_{22} \cdots$	$y_{2j} \cdots$	$y_{2j} \cdots$	y_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_i	$\underline{a}_i$	$x_i = y_{i0}$	y_{i1}	$y_{i2} \cdots$	$y_{ij} \cdots$	$y_{ik} \cdots$	y_{in}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_r	$\underline{a}_r$	$x_r = y_{r0}$	y_{r1}	$y_{r2} \cdots$	$y_{rj} \cdots$	$y_{rk} \cdots$	y_{rn}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_m	$\underline{a}_m$	$x_m = y_{m0}$	y_{m1}	$y_{m2} \cdots$	$y_{mj} \cdots$	$y_{mk} \cdots$	y_{mn}
Wiersz wskaźnikowy $F_j - c_j$		$F_0 = y_{m+1,0}$	$F_1 - c_1 = y_{m+1,1}$	$F_2 - c_2 = y_{m+2,2}$	$F_j - c_j = y_{m+1,j}$	$F_k - c_k = y_{m+1,k}$	$F_n - c_n = y_{m+1,n}$

d. W czwartym kroku badamy czy uzyskane rozwiązanie podstawowe jest optymalne, a jeśli nie to dokonujemy wyboru nowego wektora bazowego a_k, który następnie wprowadzamy do bazy.

Zgodnie z przytoczonymi kryteriami przy poszukiwaniu maksimum formy liniowej możliwe są następnie dwa przypadki:

1. Jeśli wszystkie elementy wiersza wskaźnikowego są nieujemne tzn. $F_j - c_j \geq 0$, $j = 1, 2, \ldots, n$, to rozwiązanie jest optymalne.
2. Jeśli jeden lub więcej elementów wiersza wskaźnikowego są ujemne tzn. $F_j - c_j < 0$ np. dla $j = k$, to uzyskane rozwiązanie bazowe nie jest optymalne i należy wprowadzić nowy wektor do bazy.

W celu dokonania wyboru nowego wektora bazowego stosujemy następujący tok postępowania:

- znajdujemy

$$F_k - c_k = \min(F_j - c_j) \quad \text{wśród} \quad F_j - c_j < 0,$$

- sprawdzamy czy przynajmniej dla jednego i

$$y_{ik} > 0 \quad \text{dla} \quad i = 1, 2, \ldots m.$$

Załóżmy, że powyższe dwa warunki są spełnione więc wektor a_k wprowadzamy do bazy. Należy teraz podjąć decyzję, który z wektorów należy z niej usunąć. Zgodnie z kryterium 2 dokonujemy tego w myśl następującej reguły:

- wyznaczamy kolejne ilorazy

$$\varphi_i = \frac{y_{io}}{y_{ik}} \quad \text{dla} \quad i = 1, 2, \ldots m,$$

- spośród zbioru φ_i wybieramy φ_{min} tzn.

$$\varphi_{min} = \min \frac{y_{io}}{y_{ik}}; \quad y_{ik} > 0. \tag{96}$$

jeśli φ_{min} występuje np. przy $i = r$, to wektor a_k włączamy do bazy na miejsce wektora a_r. W ten sposób otrzymujemy nową bazę, dla której rozwiązanie zadania będzie posiadać większą wartość formy liniowej, niż poprzednio.

Kolumnę k Tablicy Simpleksów nazywamy zwykle "kolumną kluczową", wiersz r - wierszem kluczowym, natomiast element znajdujący się na przecięciu kolumny kluczowej i wiersza kluczowego - "elementem rozwiązującym". W naszym przypadku elementem rozwiązującym jest y_{rk}.

e. W piątym kroku tworzymy następną tablicę simpleksów. W celu znalezienia jej elementów korzystamy z wyprowadzonych

uprzednio wzorów (85), których postać jest następująca:

$$y'_{rj} = \frac{y_{rj}}{y_{rk}}, \quad j = 0, 1, 2, \ldots, n, \tag{97}$$

$$y'_{ij} = y_{ij} - \frac{y_{ik}}{y_{rk}} y_{rj}, \tag{98}$$

$$i = 1,2,\ldots,m+1; \quad i \neq r; \quad j = 0, 1, 2, \ldots, n$$

oraz

$$b_i = y_{i0}; \quad F_0 = y_{m+1,0}; \quad F_j - c_j = y_{m+1,j} \tag{99}$$

Po wyznaczeniu elementów y'_{ij} należy następnie:

- w kolumnie c_b wartość c_r zastąpić przez c_k
- w kolumnie "wektory bazowe" wektor $\underline{a}_r$ zastąpić przez wektor $\underline{a}_k$.

f. W szóstym kroku powtarzamy czynności opisane w punkcie d niniejszego algorytmu.

Na zakończenie naszych rozważań rozpatrzymy przykład zastosowania metody Simpleks, przy czym posłużymy się tym samym problemem, który wcześniej rozwiązywaliśmy graficznie.

3.4.3. Przykład

Znaleźć maksymalną wartość formy liniowej

$$F_{max} = 5 x_1 + 3 x_2,$$

pod warunkiem spełnienia ograniczeń

$$\begin{aligned} 3 x_1 + 5 x_2 &\leqslant 15, \\ 5 x_1 + 2 x_2 &\leqslant 10, \\ x_1 \geqslant 0, \quad x_2 &\geqslant 0. \end{aligned} \tag{100}$$

Rozwiązania tego problemu dokonamy zgodnie z omówionym poprzednio algorytmem metody Simpleks.

Krok pierwszy - sprowadzamy problem do postaci kanonicznej.

A więc wprowadzimy do nierówności zmienne dopełniające x_3, $x_4 \geqslant 0$, w wyniku czego mamy

$$3x_1 + 5x_2 + x_3 = 15,$$
$$5x_1 + 2x_2 + x_4 = 10. \qquad (101)$$

Krok drugi - znajdujemy pierwsze rozwiązanie podstawowe. Układowi równań (101) odpowiada następujący układ wektorów:

$$\underline{a}_1 = \begin{bmatrix} 3 \\ 5 \end{bmatrix}, \quad \underline{a}_2 = \begin{bmatrix} 5 \\ 2 \end{bmatrix}, \quad \underline{a}_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \underline{a}_4 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \underline{b} = \begin{bmatrix} 15 \\ 10 \end{bmatrix}$$

czyli macierz A ma postać

$$A = \begin{bmatrix} 3, & 5, & 1, & 0 \\ 5, & 2, & 0, & 1 \end{bmatrix}$$

Jak nietrudno zauważyć, współczynniki przy zmiennych dopełniających x_3 i x_4 tworzą macierz jednostkową. Oznacza to, że wektory $\underline{a}_3$ i $\underline{a}_4$ wyznaczają bazę wyjściową oraz że istnieje pierwsze rozwiązanie podstawowe

$$\underline{x}_B = \left[0, \quad 0, \quad 15, \quad 10\right]^T.$$

Krok trzeci - tworzymy pierwszą tablicę simpleksów.

Na wstępie obliczamy elementy wiersza wskaźnikowego według (95). Tak więc

$$F_o = 5x_1 + 3x_2 + 0x_3 + 0x_4 = 0,$$

gdyż w formie liniowej zmiennym dopełniającym odpowiadają zerowe współczynniki c_3 i c_4.

Stąd wektor $\underline{c}_b = \left[0, \quad 0\right]^T$ oraz $F_j - c_j = -c_j$.

Wypełnimy teraz tablicę simpleksów zgodnie z tablicą 1.

Tablica 2

Pierwsza tablica simpleksów

$\underline{c}_b$	Wektory bazowe	c_j	5	3	0	0
		$\underline{b}$	$\underline{a}_1$	$\underline{a}_2$	$\underline{a}_3$	$\underline{a}_4$
0	$\underline{a}_3$	15	3	5	1	0
0	$\underline{a}_4$	10	5	2	0	1
Wiersz wskaźnikowy		0	-5	-3	0	0

Krok czwarty - badamy czy uzyskane rozwiązanie jest optymalne, a jeśli nie to dokonujemy wyboru nowego wektora bazowego, który następnie wprowadzamy do bazy.

Jak wynika z tablicy Simpleksów rozwiązanie wyjściowe nie jest optymalne, bowiem dwa elementy wiersza wskaźnikowego (-5 oraz -3) są ujemne. Mniejszą wartość ma różnica

$$F_1 - c_1 = -5.$$

Wobec tego wektor $\underline{a}_1$ będzie włączony w następnym kroku do kolumny "wektory bazowe". W tablicy 2 "kolumną kluczową" jest kolumna odpowiadająca wektorowi $\underline{a}_1$. Kolumna ta została ujęta w grubą ramkę.

Poszukamy teraz "wiersza kluczowego" odpowiadającego wektorowi wyłączanemu z bazy, którego miejsce zajmie $\underline{a}_1$. W tym celu skorzystamy ze wzoru (96), przy czym zauważmy, że elementy: $y_{11} = 3$ oraz $y_{21} = 5$ są dodatnie. Możemy więc, podzielić przez nie odpowiednie elementy kolumny wyrazów wolnych, a mianowicie:

$$\frac{y_{10}}{y_{11}} = \frac{15}{3} = 5, \qquad \frac{y_{20}}{y_{21}} = \frac{10}{5} = 2.$$

Wynika stąd, że minimalna wartość ϱ_{min} wynosi 2, a więc "wierszem kluczowym" jest wiersz, w którym występuje wektor $\underline{a}_4$. Przy tworzeniu nowej tablicy simpleksów zostanie on zastąpiony przez wektor $\underline{a}_1$. Zwróćmy uwagę, że elementem rozwiązującym jest

$$y_{rk} = y_{21} = 5.$$

Krok piąty - tworzymy następną tablicę simpleksów, przy czym odpowiednie jej elementy obliczamy ze wzorów (97), (98) i (99) w następujący sposób:

w pierwszej kolejności znajdujemy elementy przekształconego wiersza kluczowego zgodnie ze wzorem (97), a więc

$$y'_{20} = \frac{y_{20}}{y_{21}} = \frac{10}{5} = 2; \quad y'_{22} = \frac{2}{5} = 0,4; \quad y'_{24} = \frac{1}{5} = 0,2;$$

natomiast pozostałe dwa wiersze ze wzoru (98);

- dla wiersza odpowiadającego wektorowi $\underline{a}_3$ mamy

$$\frac{y_{1k}}{y_{rk}} = \frac{y_{11}}{y_{21}} = \frac{3}{5} = 0,6,$$

to znaczy, że $y'_{1j} = y_{1j} - 0,6\ y_{2j}$,
stąd

$$y'_{10} = 15 - 0,6 \cdot 10 = 9; \qquad y'_{12} = 5 - 0,6 \cdot 2 = 3,8;$$

$$y'_{14} = 0 - 0,6 \cdot 1 = -0,6;$$

- dla wiersza wskaźnikowego mamy

$$\frac{y_{3k}}{y_{2k}} = \frac{y_{31}}{y_{21}} = -1, \quad \text{a więc} \quad y'_{3j} = y_{3j} + y_{2j},$$

skąd

$$y'_{30} = 0 + 10 = 10; \qquad y'_{32} = 2 - 3 = -1; \qquad y'_{34} = 0 + 1 = 1.$$

Otrzymane dane wprowadzamy do drugiej tablicy simpleksów.

Tablica 3

Druga tablica simpleksów

c_b	Wektory bazowe	c_j	5	3	0	0
		$\underline{b}$	$\underline{a}_1$	$\underline{a}_2$	$\underline{a}_3$	$\underline{a}_4$
0	$\underline{a}_3$	9	0	3,8	1	-0,6
5	$\underline{a}_1$	2	1	0,4	0	0,2
	Wiersz wskaźnikowy	10	0	-1	0	1

Krok szósty - powtarzamy krok czwarty tzn. badamy czy uzyskane rozwiązanie jest optymalne. Z tablicy 3 wynika, że nowe rozwiązanie bazowe jest również nieoptymalne, gdyż

$$F_2 - c_2 = -1 < 0.$$

Postępując dalej analogicznie jak to przedstawiono w kroku czwartym i piątym dochodzimy do następnej tablicy simpleksów o postaci

Trzecia tablica simpleksów

$\underline{c}_b$	Wektory bazowe	c_j	5	3	0	0
		$\underline{b}$	$\underline{a}_1$	$\underline{a}_2$	$\underline{a}_3$	$\underline{a}_4$
3	$\underline{a}_2$	2,368	0	1	0,2632	-0,1579
5	$\underline{a}_1$	1,053	1	0	-0,1053	0,2632
	Wiersz wskaźnikowy	12,368	0	0	0,2632	0,8421

Analizując otrzymane wyniki widzimy, że w wierszu wskaźnikowym mamy

$$F_j - c_j \geqslant 0 \qquad \text{dla każdego } j.$$

Otrzymaliśmy przeto rozwiązanie optymalne, które wynosi:

$$\hat{x}_1 = 1,053; \quad \hat{x}_2 = 2,368 \quad \text{oraz} \quad F_{max} = 12,37.$$

3.5. Dualność w zadaniach programowania liniowego

Rozważmy następujące dwa problemy:

1) znaleźć wektor $\hat{\underline{x}}$, który maksymalizuje liniową funkcję celu o postaci:

$$\max_{\underline{x}} F = \max_{\underline{x}} \underline{c}^T \underline{x},$$

przy warunkach

$$A\,\underline{x} \leqslant \underline{b} \tag{102}$$

oraz

$$\underline{x} \geqslant 0;$$

2) znaleźć wektor $\hat{\underline{\lambda}}$, który minimalizuje liniową funkcję celu o postaci:

$$\min_{\underline{\lambda}} F_d = \min_{\underline{\lambda}} \underline{\lambda}^T \underline{b}$$

przy warunkach

$$A^T \lambda \geqslant c \tag{103}$$

oraz

$$\lambda \geqslant 0,$$

gdzie $\underline{x}$ i $\underline{c}$ są n-wymiarowymi wektorami, $\underline{\lambda}$ i $\underline{b}$ - wektorami m-wymiarowymi, a macierz A posiada wymiar $m \times n$.

Problemy te nazywamy dualnymi względem siebie, przy czym jeśli pierwszy jest problemem prymalnym to drugi jest dualnym i odwrotnie.

Jak można wykazać problem 2 można sprowadzić do problemu 1 przez zamianę znaków przy A, $\underline{b}$, $\underline{c}$ na przeciwne.

Utwórzmy najpierw funkcję Lagrange'a dla problemu pierwszego

$$L_1(\underline{x}, \underline{\lambda}) = \underline{c}^T \underline{x} + \sum_{i=1}^{m} \lambda_i (b_i - \underline{a}^i \underline{x}), \tag{104}$$

gdzie $\underline{a}^i$ oznacza i-ty wiersz macierzy A, natomiast

$\underline{\lambda} = [\lambda_1, \lambda_2, \ldots \lambda_m]^T$ - wektor mnożników Lagrange'a.

Zmieńmy teraz znaki A, $\underline{b}$ i $\underline{c}$ oraz zbudujmy funkcję Lagrange'a dla problemu drugiego. A więc

$$L_2(\underline{\lambda}, \underline{x}) = - \underline{\lambda}^T \underline{b} + \sum_{j=1}^{n} (\underline{a}_j^T \underline{\lambda} - c_j) x_j, \tag{105}$$

przy czym $\underline{a}_j$ oznacza j-tą kolumnę macierzy A, a ponadto $\underline{x} \geqslant 0$, $\underline{\lambda} \geqslant 0$.

Dokonując prostych przekształceń łatwo zauważyć, że

$$L_2(\underline{\lambda}, \underline{x}) = -L_1(\underline{x}, \underline{\lambda}), \tag{106}$$

skąd wynika równoważność obu rozpatrywanych problemów.

Związki pomiędzy rozwiązaniami jednego i drugiego problemu programowania liniowego określone są przez następujące dwa twierdzenia, które przytoczymy bez dowodów:

Twierdzenie 1. Aby wektor $\hat{\underline{x}}$ był rozwiązaniem problemu 1 programowania liniowego, trzeba i wystarcza, aby istniał taki wektor $\hat{\underline{\lambda}} \geqslant 0$, że $[\hat{\underline{x}}, \hat{\underline{\lambda}}]$ jest punktem siodłowym funkcji Lagrange'a $L_1(\underline{x}, \underline{\lambda})$.

Twierdzenie 2. Jeśli zadanie prymalne programowania liniowego (1) posiada rozwiązanie $\hat{\underline{x}}$, to zadanie dualne (2) ma rozwiązanie $\hat{\underline{\lambda}}$, przy czym

$$\underline{c}^T \hat{\underline{x}} = \hat{\underline{\lambda}}^T \underline{b} \tag{107}$$

oraz, jeżeli

$$\underline{a}^i \underline{x} < b_i \quad \text{to} \quad \hat{\lambda}_i = 0, \quad i = 1, 2, \ldots, m, \tag{108}$$

jeżeli

$$\underline{a}_j^T \underline{\lambda} > c_j \quad \text{to} \quad \hat{x}_j = 0, \quad j = 1, 2, \ldots, n, \tag{109}$$

jeżeli

$$\hat{\lambda}_i > 0 \qquad \text{to} \qquad \underline{a}^i \hat{\underline{x}} = b_i, \quad i = 1, 1, \ldots, m, \tag{110}$$

jeżeli

$$\hat{x}_i > 0 \qquad \text{to} \qquad \underline{a}_j^T \hat{\underline{\lambda}} = c_j, \quad j = 1, 2, \ldots, n. \tag{111}$$

Dowody wymienionych twierdzeń można znaleźć w pracach [25], [36].

Na zakończenie tych krótkich rozważań warto wspomnieć jakie korzyści wypływają z wprowadzenia zagadnień dualnych. W przypadku gdy w danym zadaniu programowania liniowego występuje dość znaczna liczba równań ograniczających przy niewielkiej liczbie zmiennych, to istnieje konieczność operowania bazą o dużej wymiarowości. Natomiast przy rozwiązywaniu zadania dualnego rozmiar bazy, równy liczbie zmiennych, odpowiednio maleje, co bezpośrednio wpływa na zmniejszenie się nakładu obliczeń. Można więc w takich przypadkach rozwiązywać zadanie dualne zamiast zadania prymalnego.

4. Programowanie kwadratowe

4.1. Sformułowanie problemu. Warunki Kuhna-Tuckera

Jak już wspomniano w punkcie 1, programowanie kwadratowe jest rodzajem programowania nieliniowego, w którym ograniczenia są liniowe, a funkcja celu stanowi sumę formy liniowej z formą kwadratową. Po wprowadzeniu do zbioru ograniczeń (C) punkt 1 współrzędnych dopełniających zgodnie z zasadą omówioną w punkcie 3, problem programowania kwadratowego (C), może być sformułowany następująco:

Znaleźć wektor $\hat{\underline{x}}$, który ekstremalizuje nieliniową funkcję celu

$$F = \underline{c}^T \underline{x} + \underline{x}^T D \underline{x}, \tag{112}$$

pod warunkiem spełnienia zbioru liniowych ograniczeń

$$A \underline{x} = \underline{b} \tag{113}$$

oraz

$$\underline{x} \geq 0, \tag{114}$$

przy czym $\underline{x}$ i $\underline{c}$ są wektorami n-wymiarowymi, $\underline{b}$ wektorem m-wymiarowym, macierz A posiada wymiar $m \times n$, natomiast macierz D - $n \times n$. Nie tracąc nic na ogólności przyjmujemy, że macierz D jest macierzą symetryczną.

W dalszym ciągu naszych rozważań, poprzestaniemy na rozpatrzeniu zagadnienia maksymalizacji wskaźnika (112). Jak wiadomo zagadnienie minimalizacji może być łatwo wprowadzone do maksymalizacji przez zmianę znaku funkcji celu

$$\min f(\underline{x}) = -\max(-f(\underline{x})).$$

Opracowanie ogólnej metody rozwiązania problemu programowania kwadratowego jest możliwe pod warunkiem, że każde znalezione optimum lokalne będzie zarazem optimum globalnym. Jak zostało to wykazane w punkcie 2.1 właściwość taką posiadają funkcje wypukłe bądź wklęsłe. Ponadto z własności tych funkcji wynika, że suma dwóch funkcji wklęsłych stanowi również funkcję wklęsłą.

W rozpatrywanym przez nas przypadku wskaźnik jakości (112) jest sumą formy liniowej (która jest funkcją wklęsłą) oraz formy kwadratowej $\underline{x}^T D \underline{x}$. Jak można wykazać forma kwadratowa $\underline{x}^T D \underline{x}$ jest funkcją wklęsłą jeśli

1) $\underline{x}^T D \underline{x}$ jest ujemnie określona tzn.

$$\underline{x}^T D \underline{x} < 0 \quad \text{dla każdego} \quad \underline{x} \quad \text{z wyjątkiem} \quad \underline{x} = 0$$

albo

2) $\underline{x}^T D \underline{x}$ jest ujemnie półokreślona tzn.

$$\underline{x}^T D \underline{x} \leqslant 0 \quad \text{dla każdego} \quad \underline{x},$$

przy czym istnieje takie $\underline{x} \neq 0$, dla którego $\underline{x}^T D \underline{x} = 0$.

Tak więc, jeśli forma kwadratowa będzie spełniać którykolwiek z powyższych warunków to wskaźnik jakości będzie funkcją wklęsłą. W dalszym ciągu będziemy zakładać, że warunki te są zawsze spełnione. Zauważmy przy tym, że jeśli mamy do czynienia tylko z pierwszym przypadkiem, to wówczas forma kwadratowa $\underline{x}^T D \underline{x}$ jest ściśle wklęsła, co oznacza, że problem posiada jedyne maksimum głobalne.

Skorzystamy teraz z twierdzenia Kuhna-Tuckera, które zostało omówione w punkcie 2.4. Ponieważ, zgodnie z naszymi założeniami, funkcja celu jest wklęsła oraz ograniczenia są wypukłe, a więc wystarczy sprawdzić warunki konieczne Kuhna-Tuckera, aby wykazać, że punkt $\hat{\underline{x}}$ jest szukanym rozwiązaniem.

W notacji stosowanej w punkcie 2.4 problem programowania kwadratowego zapiszemy

$$g_i(\underline{x}) = a^i \underline{x}; \qquad f(\underline{x}) = \underline{c}^T \underline{x} + \underline{x}^T D \underline{x}, \tag{115}$$

gdzie a^i jest i-tym wierszem macierzy A.

Znajdziemy teraz odpowiednie pochodne cząstkowe

$$\frac{\partial g_i}{\partial x_j} = a_{ij}; \qquad \nabla g_i(\underline{x}) = a^i, \tag{116}$$

$$\frac{\partial f}{\partial x_j} = c_j + 2 \sum_{i=1}^{n} x_i a_{ij}; \qquad \nabla f(\underline{x}) = \underline{c}^T + 2 \underline{x}^T D. \tag{117}$$

Z twierdzenia Kuhna-Tuckera wiemy, że jeśli $\hat{\underline{x}} \geqslant 0$ jest optymalnym rozwiązaniem problemu programowania kwadratowego, wtedy musi istnieć takie $\hat{\underline{\lambda}}$, że punkt $[\hat{\underline{x}}, \hat{\underline{\lambda}}]$ spełnia warunki konieczne (28) do (33).

Na wstępie rozważymy warunek (28) tzn.

$$\frac{\partial}{\partial x_j} L(\hat{\underline{x}}, \hat{\underline{\lambda}}) = \nabla f(\hat{\underline{x}}) - \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{\underline{x}}) \leqslant 0.$$

Po skorzystaniu ze wzorów (116) i (117), warunek ten przyjmie postać

$$\underline{c}^T + 2 \hat{\underline{x}}^T D - \hat{\underline{\lambda}}^T A \leqslant 0, \tag{118}$$

ale z założeń wynika, że $D^T = D$, to po wprowadzeniu n wymiarowego wektora dopełniającego

$$\hat{\underline{v}} = A^T \hat{\underline{\lambda}} - \underline{c} - 2 D \hat{\underline{x}} \geqslant 0, \tag{119}$$

nierówność (118) możemy zapisać

$$\underline{c} + 2 D \hat{\underline{x}} - A^T \hat{\underline{\lambda}} + \hat{\underline{v}} = 0. \tag{120}$$

Uwzględniając (119) warunek K-T (30)

$$\nabla_x L(\hat{\underline{x}}, \hat{\underline{\lambda}}) \hat{\underline{x}} = 0$$

wyrazi się następująco

$$\hat{\underline{x}}^T \hat{\underline{v}} = 0, \quad \text{lub} \quad x_j v_j = 0, \quad j = 1, \ldots, n, \tag{121}$$

natomiast warunek (31) równoważny jest równaniu

$$A \underline{x} = \underline{b}. \tag{122}$$

Pozostałe warunki Kuhna-Tuckera, tzn. (29), (32) i (33) są spełnione automatycznie dla każdego możliwego rozwiązania, a więc nie potrzeba się nimi oddzielnie zajmować.

Reasumując, jeśli punkt $\hat{\underline{x}} \geqslant 0$ jest optymalnym rozwiązaniem zagadnienia, to muszą istnieć takie $\hat{\underline{\lambda}}$ oraz $\hat{\underline{v}} \geqslant 0$, że zależności (120), (121) i (122) są spełnione. Zależności te reprezentują konieczne warunki Kuhna-Tuckera w zastosowaniu do problemu programowania kwadratowego.

Ponieważ zgodnie z przyjętymi założeniami funkcja celu jest wklęsła, a ograniczenia wypukłe, więc warunki konieczne stają się warunkami dostatecznymi. Inaczej mówiąc, jeśli znajdziemy takie $\underline{x} \geqslant 0$, $\underline{v} \geqslant 0$ oraz $\underline{\lambda}$, które spełniają układ równań

$$A\,\underline{x} = \underline{b},$$

$$2\,D\,\underline{x} - A^T\underline{\lambda} + \underline{v} = -\,\underline{c}, \qquad (123)$$

$$x_j v_j = 0, \quad j = 1, \ldots, n,$$

wtedy $\underline{x}$ jest optymalnym rozwiązaniem. W rezultacie więc, zadanie rozwiązania problemu programowania kwadratowego zostało sprowadzone do rozwiązania zagadnienia (123).

Istnieje szereg efektywnych metod iteracyjnych rozwiązujących ten problem. Do najważniejszych z nich można zaliczyć metody: Wolfa [61], Franka i Wolfa [23], Beale [3], Hildreth [29], Houthakkera [30] i innych. W niniejszej pracy poprzestaniemy na omówieniu algorytmu Wolfa.

4.2. Metoda Wolfa

Metodę tę stosuje się do rozwiązywania zagadnienia (123) przy założeniu, że forma kwadratowa $\underline{x}^T D\,\underline{x}$ jest ujemnie określona. Jedną z największych korzyści tej procedury w porównaniu z innymi metodami numerycznymi jest to, że przy rozwiązywaniu problemu optymalizacji posługuje się ona metodą Simpleks, stosowaną w programowaniu liniowym.

Metoda Wolfa oparta została na bardzo istotnym spostrzeżeniu. Mianowicie, jeśli $[\underline{x}, \underline{\lambda}, \underline{v}]$ jest rozwiązaniem $m + n$ równań

$$\begin{bmatrix} A & 0 & 0 \\ 2D & -A^T & I_n \end{bmatrix} \begin{bmatrix} \underline{x} \\ \underline{\lambda} \\ \underline{v} \end{bmatrix} = \begin{bmatrix} \underline{b} \\ -\underline{c} \end{bmatrix} \qquad (124)$$

takim, że

$$\underline{x} \geqslant 0, \qquad \underline{v} \geqslant 0, \qquad \underline{x}^T \underline{v} = 0,$$

wtedy nie więcej niż $m + n$ składników $[\underline{x}, \underline{\lambda}, \underline{v}]$ może być różne od zera (stwierdzenie to wynika bezpośrednio z warunku $\underline{x}^T \underline{v} = 0$). Stąd, jeśli wektor $[\underline{x}, \underline{\lambda}, \underline{v}]$ przy $\underline{x} \geqslant 0$, $\underline{v} \geqslant 0$ spełnia układ równań (123), to musi on być również bazowym rozwiązaniem (124). Wystarczy więc, badać bazowe rozwiązania (124), aby znaleźć rozwiązanie (123) pod warunkiem oczywiście, że ono istnieje. Rezultat ten był po raz pierwszy uzyskany przez Barankina i Dorfmana [26].

Algorytm obliczeniowy dla wyznaczenia bazowego rozwiązania układu (124) jest w istocie pewną modyfikacją techniki zmiennych sztucznych stosowanej w programowaniu liniowym. Pierwszym krokiem tej procedury jest określenie bazowego rozwiązania układu

równań $A\,\underline{x} = \underline{b}$, przy czym wykonuje się to w sposób analogiczny jak w metodzie Simpleks. Jeśli bazowe rozwiązanie układu $A\,\underline{x} = \underline{b}$ zostało już znalezione, to wówczas mamy pewność, że istnieje również bazowe rozwiązanie (124) dla $\underline{x} \geqslant 0$, $\underline{v} \geqslant 0$, $\underline{x}^T\,\underline{v} = 0$. Wyznaczamy go w drugim kroku. W tym celu po pierwsze powiększamy zestaw równań (124) przez dodatnie nieujemnych zmiennych sztucznych. Następnie maksymalizując ujemną sumę tych zmiennych przy warunku $\underline{x}^T\,\underline{v} = 0$, aż do momentu, gdy równać się ona będzie zero, otrzymujemy szukane bazowe rozwiązanie (124).

Rozpatrzmy teraz szczegółowo opisany powyżej algorytm. Załóżmy, że zostało już znalezione bazowe rozwiązanie układu $A\,\underline{x} = \underline{b}$. Oznaczmy je przez $\underline{x}_B$, natomiast odpowiednią macierz bazową przez B. Stąd

$$B\,\underline{x}_B = \underline{b}. \tag{125}$$

Zwróćmy uwagę, że bazowe rozwiązanie $\underline{x}_B$ spełnia m pierwszych ograniczeń (124). Dlatego też w układzie (124) tylko do ostatnich n ograniczeń wprowadzimy dodatkowe zmienne sztuczne. Tak więc, układ równań (124) przyjmie postać:

$$\begin{aligned} A\,\underline{x} \qquad\qquad\qquad &= \underline{b}\,, \\ 2D\,\underline{x} - A^T\underline{\lambda} + \underline{v} + E\,\underline{u} &= -\underline{c}\,, \end{aligned} \tag{126}$$

gdzie: $\underline{u} \geqslant 0$ jest n-składnikowym wektorem zmiennych sztucznych, $E = \| \Delta_j\, \delta_{ij} \|$ jest macierzą diagonalną, której elementy $\Delta_j = \pm 1$.

Rozważymy teraz sposób określenia znaku Δ_j. Oznaczmy przez D_B macierz zawierającą kolumny macierzy D odpowiadające kolumnom macierzy A w B oraz przez d^j_B j-ty wiersz macierzy D_B.

Wtedy

$$\Delta_j = \begin{cases} +1 & \text{jeśli } -c_j - 2d^j_B\,\underline{x}_B \geqslant 0 \\ -1 & \text{jeśli } -c_j - 2d^j_B\,\underline{x}_B < 0. \end{cases} \tag{127}$$

Z definicji tej wynika, że jeśli ustalimy

$$u_j = \left| -c_j - 2d^j_B\,\underline{x}_B \right| \geqslant 0, \quad j = 1, \ldots n; \quad \underline{\lambda} = 0; \quad \underline{v} = 0, \tag{128}$$

to otrzymujemy jedno z możliwych rozwiązań układu (126), które zawiera nie więcej niż $n + m$ niezerowych zmiennych.

Rozwiązanie to możemy zapisać

$$\begin{bmatrix} B & 0 \\ 2D_B & E \end{bmatrix} \begin{bmatrix} \underline{x}_B \\ \underline{u} \end{bmatrix} = \begin{bmatrix} \underline{b} \\ -\underline{c} \end{bmatrix} \tag{129}$$

Jak można wykazać rozwiązanie (129) stanowi bazowe rozwiązanie układu (126), a więc macierz

$$B_Q = \begin{bmatrix} B & 0 \\ 2D_B & E \end{bmatrix} \tag{130}$$

jest macierzą nieosobliwą, przy czym macierz odwrotna B_Q^{-1} ma postać

$$B_Q^{-1} = \begin{bmatrix} B^{-1} & 0 \\ -2ED_B B^{-1} & E \end{bmatrix} \tag{131}$$

Znając podstawowe rozwiązanie (126) przy $\underline{x}_B \geqslant 0$, $\underline{u} \geqslant 0$, w następnym kroku posługujemy się metodą Simpleks w celu zredukowania - $\sum_j u_j$ do zera, poprzez maksymalizację $F = -\sum_j u_j$. W metodzie tej należy wprowadzić jednak pewne modyfikacje, które umożliwiają spełnienie warunku $\underline{x}^T \underline{v} = 0$. Na ogół realizuje się to w ten sposób, że jeśli $x_j > 0$, to wówczas nie dopuszczamy do wprowadzenia do bazy v_j i vice versa. Ostatecznie więc problem programowania nieliniowego, który rozwiązujemy przy pomocy metod programowania liniowego sprowadził się nam do postaci:
znaleźć maksimum

$$F = -\sum_j u_j$$

pod warunkiem spełnienia zbioru ograniczeń

$$\begin{aligned} Q\,\underline{w} &= \underline{f}, \\ \underline{w} \geqslant 0; \quad \underline{x}^T \underline{v} &= 0, \end{aligned} \tag{132}$$

gdzie:

$$Q = \begin{bmatrix} A & 0 & 0 & 0 & 0 \\ & & & & \\ 2D & -A^T & A^T & I_n & E \end{bmatrix}; \quad f = \left[\underline{b}, \; -\underline{c}\right]^T$$

oraz $w = \left[\underline{x}, \underline{\jmath}, \underline{\xi}, \underline{v}, \underline{u}\right]^T$.

Zauważmy, że w powyższym problemie zamiast nieograniczonego wektora $\underline{\lambda}$ zostały wprowadzone dwa nieujemne wektory:

$$\underline{\jmath} \geqslant 0, \quad \underline{\xi} \geqslant 0,$$

na mocy następującego podstawienia

$$\underline{\lambda} = \underline{\jmath} - \underline{\xi}$$

Pozostało nam wyjaśnić sposób postępowania w przypadku, gdy forma kwadratowa $\underline{x}^T D \underline{x}$ jest ujemnie półokreślona tzn. $\underline{x}^T D \underline{x} \leqslant 0$.

W algorytmie obliczeniowym mogą zaistnieć wówczas dwie następujące sytuacje:

- pierwsza, kiedy istnieje możliwość, że problem programowania kwadratowego ma rozwiązanie nieograniczone,
- druga, gdy nie ma pewności czy procedura obliczeniowa jest zbieżna do żądanego rozwiązania.

W celu usunięcia wymienionych kłopotów Charnes [7] zaproponował modyfikację algorytmu Wolfa poprzez wprowadzenie do niego sekwencji umożliwiającej sprowadzenie problemu z formą kwadratową ujemnie półokresową do problemu z formą kwadratową ujemnie określoną. Jak można bowiem wykazać, jeśli $\underline{x}^T D \underline{x}$ jest ujemnie półokreśloną wtedy $\underline{x}^T (D + \varepsilon I) \underline{x}$ jest ujemnie określoną dla dowolnie małego $\varepsilon < 0$. Okazuje się jednak, że powoduje to szereg komplikacji w algorytmie, a ponieważ procedura Wolfa działa nawet gdy $\underline{x}^T D \underline{x} \leqslant 0$, więc modyfikacji tej w praktyce się nie stosuje.

Na zakończenie rozpatrzmy przykład zastosowania metody Wolfa.

4.3. Przykład

Znaleźć maksymalną wartość funkcji celu

$$F_{max} = 2 x_1 + x_2 - x_1^2, \tag{133}$$

pod warunkiem spełnienia ograniczeń

$$2 x_1 + 3 x_2 + x_3 = 6,$$

$$2x_1 + x_2 + x_4 = 4,$$

$$x_1,\ x_2,\ x_3,\ x_4 \geqslant 0.$$

Stosując poprzednio wprowadzone oznaczenia możemy napisać

$$\underline{x}^T D \underline{x} = -x_1^2,$$

$$A = \begin{bmatrix} 2 & 3 & 1 & 0 \\ 2 & 1 & 0 & 1 \end{bmatrix} \qquad D = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{134}$$

$$\underline{c} = \begin{bmatrix} 2, & 1, & 0, & 0 \end{bmatrix}^T; \qquad \underline{b} = \begin{bmatrix} 6, & 4 \end{bmatrix}^T.$$

W celu rozwiązania powyższego problemu posłużymy się procedurą Wolfa pomimo, że forma kwadratowa jest ujemnie półokreślona. Zgodnie z omówionym algorytmem wyznaczmy najpierw podstawowe rozwiązanie układu

$$A \underline{x} = \underline{b}.$$

Jak nietrudno zauważyć rozwiązanie to wynosi

$$\underline{x}_B = \begin{bmatrix} \frac{3}{2}, & 1, & 0, & 0 \end{bmatrix}^T, \tag{135}$$

przy czym

$$B = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}; \qquad B^{-1} = \begin{bmatrix} -\frac{1}{4} & \frac{3}{4} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix}; \qquad 2 D_B = \begin{bmatrix} -2 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Następnie według (132) sformułujemy zmodyfikowany zbiór ograniczeń

$$
\begin{aligned}
&2x_1 + 3x_2 + x_3 && = 6,\\
&2x_1 + x_2 + x_4 && = 4,\\
&-2x_1 - 2\zeta_1 + 2\xi_1 - 2\zeta_2 + 2\xi_2 + v_1 + \Delta_1 u_1 && = -2,\\
&- 3\zeta_1 + 3\xi_1 - \zeta_2 + \xi_2 + v_2 + \Delta_2 u_2 && = -1,\\
&- \zeta_1 + \xi_1 + v_3 + \Delta_3 u_3 && = 0,\\
&- \zeta_2 + \xi_2 + v_4 + \Delta_4 u_4 && = 0.
\end{aligned}
\tag{136}
$$

Określamy teraz znaki Δ_j oraz początkowe wartości zmiennych sztucznych u_j. Ustalając $\zeta = \xi = 0$ oraz $v = 0$, a ponadto wykorzystując rozwiązanie bazowe (135), w którym $x_1 = \frac{3}{2}$, z trzeciego równania (136) mamy

$$\Delta_1 u_1 = 1 \quad \text{skąd} \quad \Delta_1 = 1; \quad u_1 = 1,$$

z czwartego równania (136)

$$\Delta_2 u_2 = -1 \quad \text{skąd} \quad \Delta_2 = -1; \quad u_2 = 1,$$

wreszcie z piątego i szóstego równania (136) otrzymujemy

$$\Delta_3 u_3 = \Delta_4 u_4 = 0,$$

a więc możemy przyjąć

$$\Delta_3 = \Delta_4 = 1; \quad u_3 = u_4 = 0.$$

Utworzymy obecnie według (130) początkową macierz bazową B_Q, która ma postać

$$
B_Q = \begin{bmatrix}
2 & 3 & 0 & 0 & 0 & 0\\
2 & 1 & 0 & 0 & 0 & 0\\
-2 & 0 & 1 & 0 & 0 & 0\\
0 & 0 & 0 & -1 & 0 & 0\\
0 & 0 & 0 & 0 & 1 & 0\\
0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\tag{137}
$$

Skąd w oparciu o (131) otrzymujemy

$$B_Q^{-1} = \begin{bmatrix} -\frac{1}{4} & \frac{3}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ -\frac{1}{2} & \frac{3}{2} & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

W rezultacie uzyskujemy rozwiązanie bazowe układu (136), które wynosi: $x_1 = \frac{3}{2}$, $x_2 = 1$, $u_1 = 1$, $u_2 = 1$, $v_3 = v_4 = 0$ przy wszystkich pozostałych zmiennych równych zero. Ostatecznie problem (133) sprowadziliśmy do następującego zagadnienia: znaleźć maksymalną wartość formy liniowej

$$F = -u_1 - u_2, \tag{138}$$

pod warunkiem spełnienia ograniczeń (136) przy

$$\underline{x}, \underline{z}, \underline{\xi}, \underline{v}, \underline{u} \geqslant 0 \quad \text{oraz} \quad \underline{x}^T \underline{v} = 0,$$

mając dane pierwsze rozwiązanie bazowe układu (136).

Rozpatrywany przez nas problem, w postaci (138) stał się prbblemem z zakresu programowania liniowego, a więc na tym etapie możemy posłużyć się już metodą Simpleks. Dalszy tok postępowania wyjaśniono w punkcie 3, wobec tego na tym zakończymy rozwiązywanie zagadnienia (133).

5. Metody poszukiwania ekstremum bez ograniczeń

Jednym ze sposobów numerycznego rozwiązania problemu (A) pkt 1 jest sprowadzenie go najpierw do problemu bez ograniczeń, a następnie stosując którąś z metod iteracyjnych dla optymalizacji funkcji wielu zmiennych bez ograniczeń, wyznaczenie szukanego ekstremum. Wynika stąd, że o efektywności optymalizacji w dużym stopniu decyduje wówczas odpowiedni wybór metody iteracyjnej. W ciągu ostatnich ośmiu do dziesięciu lat można zauważyć bardzo duży nacisk położony na rozwój tych metod. W obecnej

chwili można by wymienić około 20 metod, którymi efektywnie posługują się różne zagraniczne ośrodki obliczeniowe. W niniejszej pracy ograniczymy się tylko do omówienia głównych reprezentacyjnych metod, natomiast w wykazie literatury zamieszczono prawie pełny wykaz materiałów źródłowych.

Metody numeryczne poszukiwania ekstremum funkcji wielu zmiennych $f(\underline{x})$ można podzielić na dwie zasadnicze grupy:

1) metody przypadkowe,
2) metody zdeterminowane.

W dalszej części metodami przypadkowymi zajmować się nie będziemy, natomiast szczegółowo rozpatrzymy metody drugiej grupy.

Wśród metod tych wyróżniamy następujące dwa rodzaje:

- metody bezgradientowe, które wymagają jedynie obliczania wartości samej funkcji, zwane często także metodami bezpośrednich poszukiwań ("direct search") oraz
- metody gradientowe, które wymagają zarówno obliczania wartości samej funkcji jak i jej pierwszych pochodnych (gradientu funkcji).

Za główne, reprezentacyjne metody bezgradientowe można uznać

1. HJ -- metodę Hooka i Jeevesa
2. R - metodę Rosenbrocka
3. N - metodę Simplexu Neldera i Meada
4. GA - metodę Gaussa-Seidela
5. DSC - metodę Daviesa, Swanna i Campeya
6. P - metodę Powella i jej modyfikacje
7. Z - metodę Zangwilla

Do metod gradientowych zaliczamy natomiast:

1. GP - metodę gradientu prostego
2. NS - metodę najszybszego spadku i jej modyfikacje
3. GS - gradientu sprzężonego wg Fletchera i Reevesa
4. D - metodę Davidona
5. PE - metody Pearsona
6. NR - metodę Newtona-Raphsona

Przed przystąpieniem do omawiania wymienionych metod zajmiemy się najpierw rozpatrzeniem algorytmów poszukiwania ekstremum w kierunku. Z algorytmów tych korzystają zarówno metody pierwszej jak i drugiej grupy, przy czym jak się o tym dalej przekonamy, oddziaływują one bardzo silnie na szybkość zbieżności poszczególnych metod. W celu ujednolicenia oznaczeń wprowadźmy następujące symbole:

$\underline{x}$ - wektor zmiennych niezależnych,
$\underline{x}_0$ - arbitralnie wybrany punkt startowy,
$\hat{\underline{x}}$ - punkt ekstremalny,

$\underline{x}_i$ - aktualny punkt bieżący,
$f(\underline{x})$ - funkcja celu,
$g(\underline{x})$ - gradient funkcji $f(\underline{x})$; $g(\underline{x}) = \nabla f(\underline{x})$,
$\underline{\xi}$ - wektor kierunku poszukiwań,
n - wymiarowość problemu,
A - macierz symetryczna dodatnio określona, której elementami są drugie pochodne cząstkowe $f(\underline{x})$.

Wszystkie wektory są n-wymiarowymi wektorami kolumnowymi. Wektory wierszowe oraz transpozycje macierzy oznaczać będziemy literą T, natomiast indeksem "i" wartości danej wielkości po (bądź w) i-tej iteracji.

5.1. Metody poszukiwania ekstremum w kierunku

Wśród dużego bogactwa metod stosowanych do rozwiązywania tego zagadnienia, najczęściej używane są metody:

1) złotego podziału,
2) interpolacji kwadratowej,
3) interpolacji sześciennej,

Pierwsze dwie metody wymagają bieżącej znajomości trzech kolejnych punktów $\underline{x}_i$ wzdłuż danego kierunku poszukiwań oraz wartości funkcji celu $f(\underline{x}_i)$ w tych punktach. W metodzie trzeciej natomiast wystarczy znajomość dwóch punktów $\underline{x}_i$, lecz w zamian za to muszą być wyliczane zarówno wartości funkcji celu jak i wartości pochodnych w tych punktach.

Rozpatrzymy je teraz po kolei.

5.1.1. Metoda złotego podziału

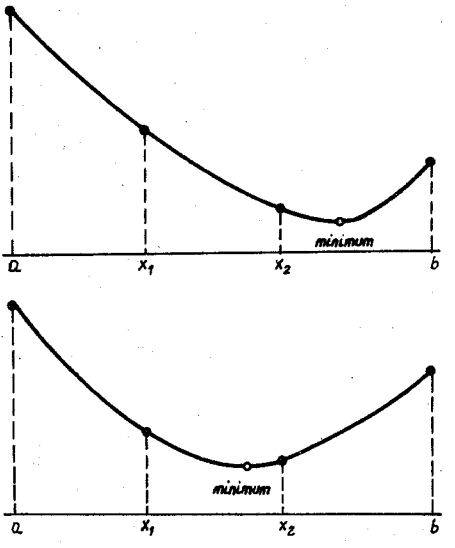
Definicja. Funkcję $f(x)$ nazywamy "unimodalną" w przedziale $a \leqslant x \leqslant b$ jeśli posiada ona pojedynczy punkt stacjonarny w tym przedziale.

Zakładając ciągłość i "unimodalność" funkcji $f(x)$ oraz przyjmując, że punkt stacjonarny stanowi poszukiwane minimum w kierunku można wypowiedzieć następujący lemat:

Lemat. Jeśli funkcja $f(x)$ jest unimodalna w przedziale $[a, b]$, wtedy dla zlokalizowania położenia punktu stacjonarnego w podprzedziale należącym do $[a, b]$ istnieje konieczność obliczenia wartości funkcji w dwóch punktach tego przedziału.

Dowód. Jeśli wartość funkcji obliczono by w jednym wewnętrznym punkcie x_1, to nie można by zdecydować po której stronie x_1 położone jest poszukiwane minimum. Wyliczając dopiero wartość funkcji w następnym punkcie x_2 (rys. 12), przy czym $a < x_1 < x_2 < b$, korzystając z nierówności $f(x_1) > f(x_2)$ można określić czy minimum położone jest w podprzedziale $[x_1, b]$, czy też $[a, x_2]$.

Na podstawie przytoczonego lematu można zbudować algorytm poszukiwania minimum w kierunku, przy czym minimum to może zostać określone z żądaną dokładnością. Istotną cechą tego algorytmu jest to, że jeden z dwóch wyznaczanych punktów x będzie



Rys. 12

zawsze znajdować się wewnątrz zredukowanego nowego podprzedziału, a tym samym w każdej następnej iteracji wystarczy wyliczać wartość funkcji tylko w jednym punkcie. Decydujący więc wpływ na efektywność tak skonstruowanego algorytmu będzie miał sposób doboru punktów wewnętrznych. Jednym ze stosowanych sposobów jest wprowadzenie zasady, że w każdej następnej iteracji bieżąco wyznaczane podprzedziały obejmujące poszukiwane minimum będą się zmniejszały o stały czynnik τ. Algorytm ten zwany algorytmem "złotego podziału" rozpatrzymy teraz bardziej szczegółowo.

Załóżmy, że w bieżącym przedziale $\left[a^{(i)}, b^{(i)}\right]$ zostały wyliczone dwie wartości funkcji celu w punktach $x_1^{(i)}$ oraz $x_2^{(i)}$ należących do tego przedziału, przy czym $x_1^{(i)} < x_2^{(i)}$.

Zgodnie z przyjętą zasadą działania procedury, punkty te muszą spełniać zależność:

$$\frac{x_2^{(i)} - a^{(i)}}{b^{(i)} - a^{(i)}} = \frac{b^{(i)} - x_1^{(i)}}{b^{(i)} - a^{(i)}} = \tau, \tag{139}$$

skąd

$$x_1^{(i)} - a^{(i)} = b^{(i)} - x_2^{(i)}. \tag{140}$$

Jeśli $f\left(x_2^{(i)}\right) > f\left(x_1^{(i)}\right)$, wówczas w następnej iteracji dokonujemy redukcji rozpatrywanego przedziału w myśl reguły:

$$\begin{aligned} b^{(i+1)} &= x_2^{(i)}, \\ a^{(i+1)} &= a^{(i)}, \\ x_2^{(i+1)} &= x_1^{(i)}. \end{aligned} \tag{141}$$

Spróbujmy określić teraz ile powinna wynosić wartość stałego czynnika τ, o który następuje zawężanie przedziału $\left[a^{(i)}, b^{(i)}\right]$.
Na podstawie (139), (140) i (141) można napisać:

$$\frac{x_2^{(i+1)} - a^{(i)}}{x_2^{(i)} - a^{(i)}} = \frac{x_1^{(i)} - a^{(i)}}{x_2^{(i)} - a^{(i)}} = \frac{x_2^{(i)} - a^{(i)}}{b^{(i)} - a^{(i)}} = \tau \tag{142}$$

ale z równania (140) wynika

$$x_1^{(i)} - a^{(i)} = b^{(i)} - a^{(i)} - \left(x_2^{(i)} - a^{(i)}\right), \tag{143}$$

tak więc

$$\frac{x_1^{(i)} - a^{(i)}}{x_2^{(i)} - a^{(i)}} = -1 + \frac{1}{\tau} = \tau. \tag{144}$$

Skąd

$$\tau^2 + \tau - 1 = 0. \tag{145}$$

Rozwiązując równanie (145) znajdujemy szukaną wartość

$$\tau = \frac{\sqrt{5} - 1}{2} \approx 0,618.$$

W rezultacie algorytm "złotego podziału" przebiega w następujący sposób:

(i) jeśli $f\left(x_2^{(i)}\right) > f\left(x_1^{(i)}\right)$, to

$$\begin{aligned} b^{(i+1)} &= x_2^{(i)}, \\ x_2^{(i+1)} &= x_1^{(i)}, \\ x_1^{(i+1)} &= a^{(i)} + (1 - \tau)\left(b^{(i+1)} - a^{(i)}\right), \end{aligned} \tag{146}$$

albo

(ii) jeśli $f\left(x_2^{(i)}\right) \leq f\left(x_1^{(i)}\right)$ to

$$\begin{aligned} a^{(i+1)} &= x_1^{(i)}, \\ x_1^{(i+1)} &= x_2^{(i)}, \\ x_2^{(i+1)} &= b^{(i)} - (1 - \tau)\left(b^{(i)} - a^{(i+1)}\right). \end{aligned} \tag{147}$$

Wynika stąd, że po n-krotnych obliczeniach funkcji $f(x)$ długość przedziału początkowego redukuje się do wielkości:

$$b^{(n)} - a^{(n)} = \left[b^{(1)} - a^{(1)}\right] \tau^{n-1}. \tag{148}$$

Oprócz przytoczonego algorytmu dosyć często stosuje się również algorytm oparty na liczbach Fibonacciego, jednakże jak wykazały badania [34] jest on mniej efektywny i bardziej skomplikowany od metody "złotego podziału".

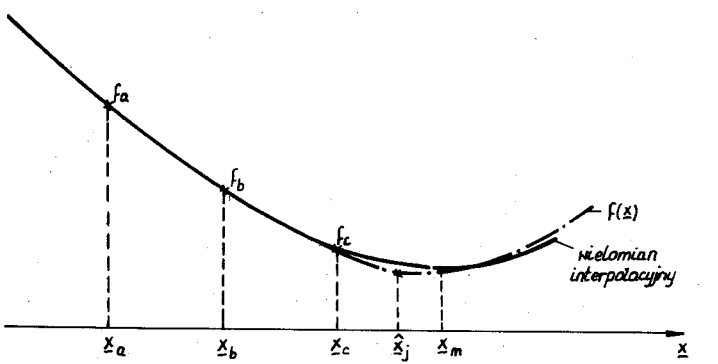
5.1.2. Metoda interpolacji kwadratowej

Zaletą poprzednio omówionej metody jest duża prostota oraz bardzo dobra zbieżność uzyskiwana w każdej sytuacji. Wadą jej natomiast jest niewielka szybkość zbieżności, która jest tym mniejsza im większą zakłada się dokładność obliczeń minimum

w kierunku. Dla przykładu, jeśli żąda się wyliczenia minimum z dokładnością $\leqslant 10^{-4}$, to ilość obliczeń wartości funkcji wynosi ok. 20. Zwróćmy jednak uwagę, że przy wyprowadzeniu metody "złotego podziału" przyjęto bardzo słabe założenia, w których wymagano tylko ciągłości funkcji oraz istnienia minimum w kierunku. W przypadkach gdy założenia te mogą zostać zaostrzone można wtedy posłużyć się bardziej efektywnym algorytmem opartym o interpolację kwadratową.

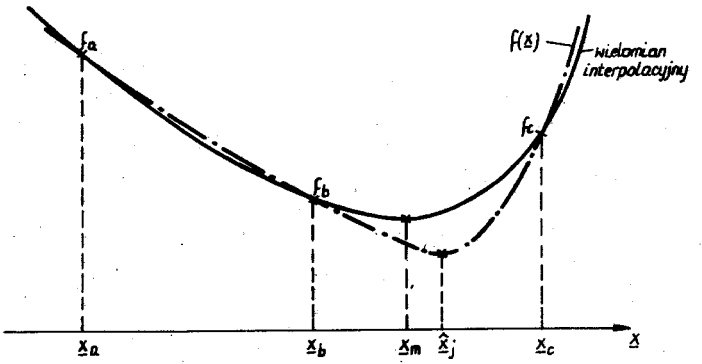
Zakładając, że w otoczeniu minimum w kierunku badaną funkcję można zastąpić wielomianem drugiego stopnia, to przy zastosowaniu interpolacji kwadratowej można zbudować dwa następujące algorytmy:

(i) wartość funkcji jest wyliczana w trzech kolejnych punktach, przez które zostaje poprowadzony wielomian interpolacyjny drugiego stopnia. Następnie dokonujemy predykcji poszukiwanego ekstremum określając punkt, w którym wartość tego wielomianu osiąga minimum (rys. 13). Z kolei punkt ten zostaje wprowadzony na miejsce jednego z punktów początkowych i opisana procedura jest powtarzana dotąd, aż zostanie spełnione odpowiednie kryterium zbieżności. Po raz pierwszy algorytm ten został zastosowany przez Powella [42]. Jednakże przy jego realizacji zalecana jest ostrożność, gdyż znaleziony punkt może okazać się maksimum w kierunku, a ponadto niewłaściwa zamiana punktów wyjściowych może doprowadzić do niezbieżności. Dlatego też w algorytmie tym należy wprowadzić szereg zabezpieczeń, przy czym niektóre z nich zostały wykorzystane w sieciach działań przedstawionych w punkcie 5.2;



Rys. 13

(ii) drugim sposobem, zaproponowanym przez Daviesa, Swanna i Campeya [34] jest wyznaczenie przedziału, w którym znajduje się minimum, a następnie dopiero zastosowanie interpolacji kwadratowej (rys. 14).



Rys. 14

W rozpatrywanych dalej bezgradientowych metodach poszukiwania ekstremum przy wyznaczaniu minimum w kierunku posłużono się algorytmem stanowiącym kompilację obu przytoczonych sposobów. Algorytm ten został zdefiniowany przy pomocy sieci działań podanych przy omawianiu poszczególnych metod optymalizacji. W algorytmie tym dla wyliczania kolejnej estymaty minimum w kierunku zastosowano wzór zaproponowany przez Powella [42], który wynika z wielomianu interpolacyjnego Lagrange'a w oparciu o następujące rozważania.

Załóżmy, że znamy wartości funkcji celu f_a, f_b i f_c w trzech kolejnych punktach x_a, x_b i x_c odpowiednio, przy czym $x_a < x_b < x_c$, to wykorzystując przyjęte na wstępie założenia, wielomian interpolacyjny Lagrange'a [31] będzie miał postać:

$$f(x) = f_a \frac{(x - x_b)(x - x_c)}{(x_a - x_b)(x_a - x_c)} + f_b \frac{(x - x_a)(x - x_c)}{(x_b - x_a)(x_b - x_c)} + \\ + f_c \frac{(x - x_a)(x - x_b)}{(x_c - x_a)(x_c - x_b)}. \qquad (149)$$

Jak wiadomo warunkiem koniecznym, a w naszym przypadku i dostatecznym istnienia w punkcie x_m ekstremum wyrażenia (149) jest warunek

$$\frac{\partial f(x)}{\partial x} = 0 \Big|_{\text{dla } x = x_m} \tag{150}$$

a więc

$$f_a \frac{2x_m - (x_b + x_c)}{(x_a - x_b)(x_a - x_c)} + f_b \frac{2x_m - (x_a + x_c)}{(x_b - x_a)(x_b - x_c)} +$$

$$+ f_c \frac{2x_m - (x_a + x_b)}{(x_c - x_a)(x_c - x_b)} = 0, \tag{151}$$

skąd

$$x_m = \frac{1}{2} \frac{(x_b^2 - x_c^2) f_a + (x_c^2 - x_a^2) f_b + (x_a^2 - x_b^2) f_c}{(x_b - x_c) f_a + (x_c - x_a) f_b + (x_a - x_b) f_c}. \tag{152}$$

Poza wyprowadzonym wzorem na x_m, można także stosować inną jego postać zaproponowaną przez Daviesa, Swanna i Campeya [34], a mianowicie

$$x_m = x_b - \frac{L}{2} \frac{f_c - f_a}{f_c - 2f_b + f_a}, \tag{153}$$

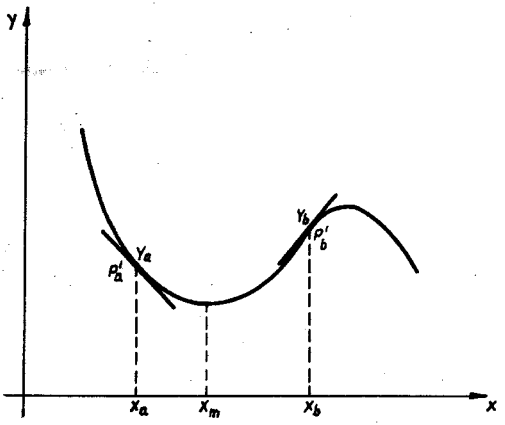
przy czym jest ona słuszna tylko w przypadku, gdy zachowane są równe odległości L pomiędzy bieżącymi punktami x_a, x_b i x_c w trakcie obliczeń.

5.1.3. Metoda interpolacji sześciennej

Metoda ta jest o wiele efektywniejsza od algorytmu omówionego w poprzednim punkcie, jednakże wymagana jest w niej znajomość wartości pochodnej kierunkowej. Dlatego też, stosowana jest ona prawie we wszystkich metodach gradientowych, przy czym po raz pierwszy posłużył się nią Davidon [11], a następnie Fletcher i Reeves [22] w metodzie gradientu sprzężonego. Metoda interpolacji sześciennej polega na tym, że przebieg wartości funkcji w kierunku poszukiwań (tzn. kontur przecięcia hiperpowierzchni funkcji przez hiperpłaszczyznę) aproksymujemy krzywą

trzeciego stopnia. Żąda się przy tym, aby przed przystąpieniem do wykonywania omawianej metody, został znaleziony przedział określony punktami x_a i x_b, w którym znajduje się poszukiwane minimum w kierunku.

Załóżmy, że dla dwóch punktów x_a i x_b mamy dane wartości funkcji Y_a i Y_b, wartości pochodnych P_a i P_b oraz, że w przedziale tym badaną funkcję celu $f(x)$ można zastąpić wielomianem trzeciego stopnia (rys. 15).



Rys. 15

W tej sytuacji, funkcja $f(x)$ wyrazi się przez

$$Y = f(x) = ax^3 + bx^2 + cx + r, \tag{154}$$

natomiast jej pochodna

$$P = 3ax^2 + 2bx + c. \tag{155}$$

Za punkt wyjścia można więc przyjąć następujący układ równań z czterema niewiadomymi:

$$\begin{aligned} Y_a &= a\,x_a^3 + b\,x_a^2 + c\,x_a + r, \\ Y_b &= a\,x_b^3 + b\,x_b^2 + c\,x_b + r, \end{aligned} \tag{156}$$

$$P_a = 3\,a\,x_b^2 + 2\,bx_a + c,$$
$$P_b = 3\,a\,x_b^2 + 2\,bx_b + c. \qquad (156)$$

Rozwiązując ten układ wyznaczymy wartości współczynników a, b i c, a następnie stosując do wyrażenia (153) warunek konieczny istnienia ekstremum (150) otrzymamy poszukiwaną wartość minimum x_m o postaci

$$x_m = \frac{-2b + \sqrt{4b^2 - 12\,ac}}{6\,a}, \qquad (157)$$

przy czym

$$4\,b^2 - 12\,ac \geqslant 0.$$

Zamiast rozwiązywać powyższe zadanie, w praktyce korzystamy ze wzorów podanych przez Davidona [11], które można przedstawić następująco:

$$e = x_b - x_a,$$
$$z = 3\,\frac{Y_a - Y_b}{e} + P_a + P_b,$$
$$w = \sqrt{z^2 - P_a \cdot P_b}, \qquad (158)$$
$$d = e\,\frac{P_b + w - z}{P_b - P_a + 2w},$$
$$x_m = x_b - d.$$

Zwróćmy przy tym uwagę, że podstawiając do tych wzorów na Y_a, Y_b, P_a oraz P_b związki (156) można przy pomocy prostych przekształceń wykazać równoważność warunków (158) i (157). W przytoczonych w niniejszej pracy algorytmach będziemy posługiwać się postacią (158).

5.2. Metody bezgradientowe poszukiwania ekstremum

Wśród metod bezgradientowych wyszczególnionych w punkcie 5 wyróżniamy dwa ich rodzaje:

- w pierwszym zakłada się jedynie wypukłość funkcji celu $f(\underline{x})$,
- w drugim natomiast żąda się ponadto, aby $f(\underline{x})$ była ograniczoną od dołu funkcją klasy C^2 oraz żeby można ją było dostatecznie dobrze aproksymować formą kwadratową o postaci

$$f(\underline{x}) = \underline{x}^T A \underline{x} + b \underline{x} + c. \qquad (159)$$

Do pierwszej grupy zaliczamy metody: Hooka i Jeevesa, Rosenbrocka oraz Neldera i Meada, do drugiej zaś wszystkie pozostałe. Omówimy je teraz po kolei.

5.2.1. Metoda Hooka i Jeevesa - HJ

Jak już wspomniano metoda ta zalicza się do metod iteracyjnych podzukiwania ekstremum co oznacza, że wyznacza ona punkt minimalny $\hat{\underline{x}}$ jako granicę ciągu $\underline{x}_0, \underline{x}_1, \underline{x}_2, \ldots$, gdzie $\underline{x}_0$ jest punktem początkowym. W metodzie tej w kolejnej iteracji występują dwa sposoby poruszania się: próbny oraz roboczy. Pierwszy sposób służy do zbadania lokalnego zachowania się funkcji w niewielkim wybranym obszarze, przez wykonywanie kroków próbnych wzdłuż wszystkich kierunków ortogonalnej bazy. Drugi - roboczy polega na przejściu w ściśle zdeterminowany sposób do następnego obszaru, w którym powtarzany jest pierwszy etap lecz tylko w tym przypadku, gdy przynajmniej jeden z wykonanych kroków próbnych był pomyślny *). W przeciwnym razie powracamy do poprzednio badanego obszaru i cykl przeszukiwania rozpoczynamy od nowa przy zmniejszonej długości kroku.

a. Informacje wejściowe

$\underline{x}_0$ - arbitralnie wybrany punkt startowy,

$\underline{\xi}_1, \underline{\xi}_2, \ldots, \underline{\xi}_n$ - baza wyjściowa utworzona z wzajemnie ortogonalnych wektorów,

e - początkowa długość kroku,

β - współczynnik korekcyjny zmniejszający e; $0 < \beta < 1$,

ε - wymagana dokładność obliczeń minimum,

n - liczba zmiennych niezależnych.

*) Krokiem pomyślnym nazywamy krok, w wyniku którego następuje zmniejszenie się wartości funkcji tzn. $f(\underline{x}) < f(\underline{x}_0)$ gdy $\underline{x} = \underline{x}_0 + e_j \underline{\xi}_j$, gdzie e oznacza długość kroku.

b. Algorytm obliczeń

ETAP PRÓBNY

(1) podstaw $j = 1$ oraz oblicz w punkcie $\underline{x}_o$ wartość funkcji $f(\underline{x}_o) \Rightarrow F_o$,

(2) wzdłuż kierunku $\underline{\xi}_j$ wykonaj krok próbny

$$\underline{x}_j = \underline{x}_{j-1} + e\,\underline{\xi}_j$$

oraz oblicz wartość funkcji w tym punkcie $f(\underline{x}_j) \Rightarrow F$,

(3) zbadaj czy krok był pomyślny tzn. czy $F < F_o$. Jeśli tak, to podstaw F w miejsce F_o oraz przejdź do wykonania punktu 6, natomiast jeśli nie, to

(4) wykonaj krok próbny w przeciwnym kierunku

$$\underline{x}_j = \underline{x}_j - 2\,e\,\underline{\xi}_j$$

oraz oblicz wartość funkcji w tym nowym punkcie $f(\underline{x}_j) \Rightarrow F$,

(5) zbadaj czy ten krok był pomyślny. Jeśli tak, to podstaw F w miejsce F_o oraz przejdź do wykonania punktu 6, natomiast w przeciwnym razie pozostaw bieżący punkt bez zmian tzn. podstaw $\underline{x}_{j-1}$ w miejsce $\underline{x}_j$,

(6) zbadaj czy wykonano kroki we wszystkich kierunkach ortogonalnej bazy tzn. czy $j = n$. Jeśli nie, to podstaw $j = j+1$ oraz powtórz czynności od punktu 2, natomiast jeśli tak, to

(7) zbadaj czy w wykonanym cyklu poszukiwania wystąpiły kroki pomyślne tzn. czy $f(\underline{x}^{Bo}) > f(\underline{x}_j)$, przy czym w pierwszej iteracji $\underline{x}^{Bo} = \underline{x}_o$. Jeśli tak, to podstaw $\underline{x}_j$ w miejsce $\underline{x}^B$, który nazywany jest punktem bazowym oraz przejdź do wykonania ETAPU ROBOCZEGO, w przeciwnym razie

(8) o ile nie zostało spełnione kryterium na minimum, zbadaj czy relizowana iteracja jest pierwszą. Jeśli tak, to zmień punkt startowy $\underline{x}_o$ i powtórz czynności od punktu 1, jeśli natomiast nie, to powróć do poprzednio przeszukiwanego obszaru w myśl zasady

$$\underline{x}_o = \underline{x}_o - \underline{x}^B,$$

zmniejsz długość kroku o β tzn. $e = \beta e$ oraz rozpocznij wykonywanie procedury od punktu 1.

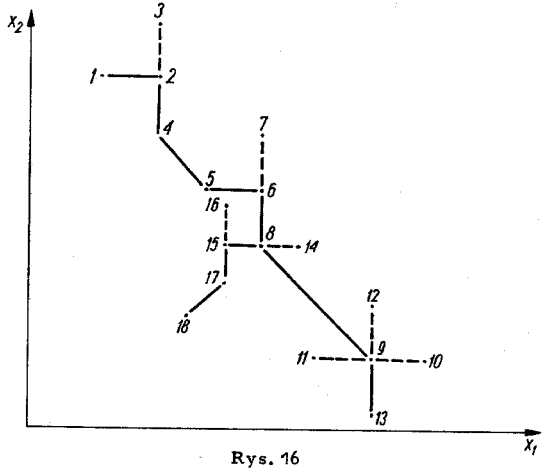
ETAP ROBOCZY

(1) wykonaj krok roboczy według reguły

$$\underline{x}_o = \underline{x}^B + (\underline{x}^B - \underline{x}^{Bo}) = 2\underline{x}^B - \underline{x}^{Bo},$$

(2) podstaw $\underline{x}^B$ w miejsce $\underline{x}^{Bo}$ oraz wróć do realizacji ETAPU PRÓBNEGO.

Rozpatrzmy działanie omówionego algorytmu na przykładzie funkcji dwuzmiennych (rys. 16).



Rys. 16

Na rysunku tym poszczególne punkty $\underline{x}_j$ oznaczono liczbami w kolejności w jakiej zostały wyliczane. Tak więc, startując z punktu $\underline{x}^1$ rozpoczynamy realizację ETAPU PROBNEGO, po zakończeniu którego otrzymujemy punkt $\underline{x}^4$. Punkt ten przyjmujemy jako punkt bazowy, bowiem w czasie przebiegu tego etapu miały miejsce dwa pomyślne kroki próbne $\underline{x}^2$ i $\underline{x}^4$ oraz tylko jeden niepomyślny $\underline{x}^3$, a przy tym $f(\underline{x}^4) < f(\underline{x}^1)$. Z kolei wykonujemy ETAP ROBOCZY w rezultacie czego uzyskujemy punkt $\underline{x}^5$. Z punktu tego rozpoczynamy następny cykl próbny kończący się znalezieniem nowego punktu bazowego $\underline{x}^8$, przy czym $f(\underline{x}^8) < f(\underline{x}^4)$. Postępując dalej w analogiczny sposób dochodzimy w trzeciej iteracji do sytuacji gdy warunek (7) nie zostaje spełniony, bowiem $f(\underline{x}^{13}) > f(\underline{x}^8)$. Wobec

tego powracamy do punktu bazowego $\underline{x}^8$, zmniejszamy długość kroku e i cykl poszukiwań rozpoczynamy od nowa.

c. Kryterium zbieżności

Idealnym kryterium zbieżności procedury iteracyjnej, decydującym o zakończeniu jej działania, byłoby przyjęcie warunku

$$\left| x_j - \hat{x}_j \right| \leqslant \varepsilon_j \quad \text{dla} \quad j = 1, 2, \ldots, n,$$

gdzie: x_j jest aktualną wartością składowej wektora $\underline{x}$ w j-tym kierunku,

$\hat{x}_j$ - składową wektora $\hat{\underline{x}}$ minimalizującego funkcję celu,

ε_j dowolnie mała założona liczba.

W rzeczywistości kryterium takie staje się nierealne, gdyż wektor $\hat{\underline{x}}$ jest nieznany, toteż w praktycznych zastosowaniach posługujemy się innymi kryteriami, będącymi kompromisem między liczbą iteracji jaką należy wykonać, a uzyskiwaną dokładnością. Hook i Jeeves w swojej pracy [34] zaproponowali jako kryterium zakończenia działania procedury przyjąć warunek, że aktualna długość kroku e będzie mniejsza od z góry założonej liczby ε.

5.2.2. Metoda Rosenbrocka - R

Metoda ta jest bardzo podobna do metody Hooke'a i Jeevesa w tym sensie, że ekstremum poszukujemy również w n wzajemnie ortogonalnych kierunkach. Istotna różnica tkwi w tym, że kierunki te nie pozostają stałe lecz w pewnych przypadkach w wyniku obrotu ulegają zmianom. Baza wyjściowa $\underline{\xi}_1^o, \underline{\xi}_2^o, \ldots \underline{\xi}_n^o$ tworzona jest zazwyczaj z wersorów układu współrzędnych kartezjańskich. Na wstępie w każdym z tych kierunków wykonujemy kolejno po jednym kroku o długości e. Jeśli eksperyment taki kończy się powodzeniem, to w następnym kroku w danym kierunku wartość e zostaje zwiększona α razy $(\alpha > 1)$, natomiast w przeciwnym razie zostaje pomnożona przez $-\beta$, przy czym $0 < \beta < 1$. Tego rodzaju tryb postępowania powtarzany jest aż do momentu, gdy wykonanie kroku we wszystkich n kierunkach daje niepomyślny wynik tzn. $\left(f(\underline{x}_{j-1}) < f(\underline{x}_j)\right)$.

W takiej sytuacji, jeśli jest spełnione kryterium na extremum, to procedura kończy swoje działanie, jeśli zaś nie - to wykonany zostaje Algorytm Obrotu Współrzędnych i działanie procedury rozpoczyna się od początku.

a. Informacje wejściowe

$\underline{x}_o$ - arbitralnie wybrany punkt startowy,

$\underline{\xi}^o_1, \underline{\xi}^o_2, \ldots, \underline{\xi}^o_n$ - baza wyjściowa utworzona z wzajemnie ortogonalnych wektorów,

$\underline{e}$ - n-wymiarowy wektor początkowy długości kroku,

α - współczynnik korekcyjny zwiększający $\underline{e}$; $\alpha > 1$,

β - współczynnik korekcyjny zmniejszający $\underline{e}$; $0 < \beta < 1$.

Wartości liczbowe współczynników α i β należy dobrać eksperymentalnie. W przypadkach rozpatrywanych przez Rosenbrocka, jako najoptymalniejsze przyjęto $\alpha = 3$, $\beta = 0,5$,

L - założona liczba iteracji,

n - liczba zmiennych niezależnych.

b. Algorytm obliczeń

(1) dla $j = 1, 2, \ldots, n$ oblicz $f(\underline{x}_{j-1} + e_j \underline{\xi}^o_j)$. Gdy $f(\underline{x}_{j-1} + e_j \underline{\xi}^o_j) < f(\underline{x}_{j-1})$, to oblicz sumę pomyślnych kroków $d_j = d_{j-1} + e_j$, przy czym $d_o = 0$, przesunięcie punktu $x_j = \underline{x}_{j-1} + e_j \underline{\xi}^o_j$ oraz podstaw $e_j = \alpha e_j$. W przeciwnym przypadku podstaw $\underline{x}_j = \underline{x}_{j-1}$ oraz $e_j = -\beta e_j$. Krok (1) powtarzaj do momentu, aż $d_j = 0$, dla $j = 1, 2, \ldots n$. Jeżeli będzie to miało miejsce po jednokrotnym wykonaniu (1), to:

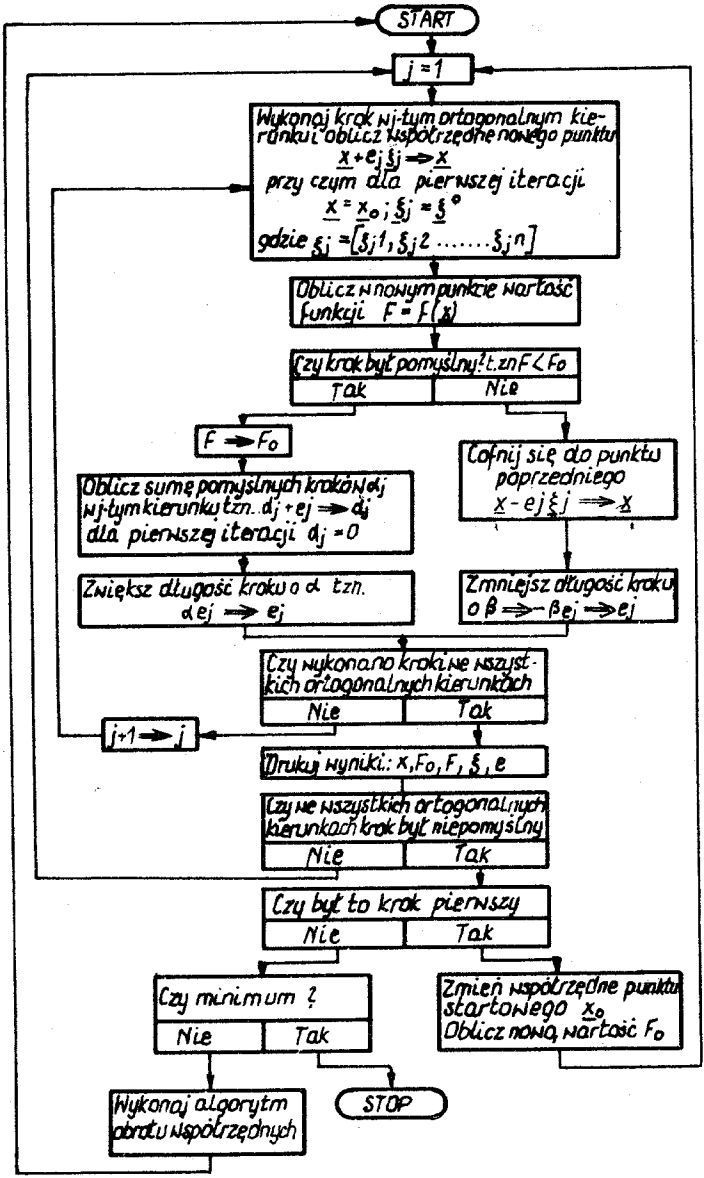
(2) zmień punkt startowy $\underline{x}_o$ i powtórz krok (1). W przeciwnym razie, o ile minimum nie jest jeszcze osiągnięte:

(3) dokonaj algorytmu obrotu ortogonalnego układu współrzędnych $\{\underline{\xi}\}$ i powróć do wykonywania kroku (1).

Sieć działań przytoczonego algorytmu przedstawiono na rys. 17, przy czym przed przystąpieniem do jego wykonywania, na wstępie wyliczamy w punkcie $\underline{x}_o$ wartość funkcji celu $F_o = f(\underline{x}_o)$.

W procedurze tej przez słowo "START" rozumianych jest szereg typowych czynności takich jak zerowanie miejsc roboczych i bloków, przepisywanie początkowych danych wejściowych itp. Czynności te celowo pominięto mając na uwadze większą przejrzystość schematu. Natomiast pod zapytaniem "Czy minimum" ukryte jest przyjęte przez Rosenbrocka kryterium zbieżności. Omówione ono zostanie w następnym podpunkcie.

Obecnie rozpatrzymy Algorytm Obrotu Współrzędnych, który wyznacza nowy układ wzajemnie ortogonalnych wektorów jednostkowych $\underline{\xi}$ tworzących nową bazę przestrzeni. Algorytm ten przebiega w następujący sposób:


START
j = 1
Wykonaj krok w j-tym ortogonalnym kierunku i oblicz współrzędne nowego punktu
x + e_j ξ_j ⇒ x
przy czym dla pierwszej iteracji
x = x_0 ; ξ_j = ξ°
gdzie ξ_j = [ξ_j1, ξ_j2 ..... ξ_jn]
Oblicz w nowym punkcie wartość funkcji F = f(x)
Czy krok był pomyślny? t.zn F < F_0
Tak
Nie
F ⇒ F_0
Cofnij się do punktu poprzedniego
x - e_j ξ_j ⇒ x
Oblicz sumę pomyślnych kroków d_j w j-tym kierunku t.zn. d_j + e_j ⇒ d_j
dla pierwszej iteracji d_j = 0
Zwiększ długość kroku o α t.zn. α e_j ⇒ e_j
Zmniejsz długość kroku o β ⇒ -β e_j ⇒ e_j
Czy wykonano kroki we wszystkich ortogonalnych kierunkach
Nie
Tak
j+1 ⇒ j
Drukuj wyniki: x, F_0, F, ξ, e
Czy we wszystkich ortogonalnych kierunkach krok był niepomyślny
Nie
Tak
Czy był to krok pierwszy
Nie
Tak
Czy minimum ?
Nie
Tak
Zmień współrzędne punktu startowego x_0
Oblicz nową wartość F_0
Wykonaj algorytm obrotu współrzędnych
STOP

Rys. 17

(1) wyznacz układ wektorów:

$$\begin{aligned} \underline{q}_1 &= d_1 \underline{\xi}_1^o + d_2 \underline{\xi}_2^o + \ldots + d_n \underline{\xi}_n^o , \\ \underline{q}_2 &= \qquad d_2 \underline{\xi}_2^o + \ldots + d_n \underline{\xi}_n^o , \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \underline{q}_n &= \qquad\qquad\qquad d_n \underline{\xi}_n^o , \end{aligned} \tag{160}$$

(2) dokonaj ortogonalizacji tego układu stosując procedurę Gram-Schmidta

$$\begin{aligned} \underline{v}_1 &= \underline{q}_1 , \\ \underline{\xi}_1 &= \frac{\underline{v}_1}{\|\underline{v}_1\|} , \\ \underline{v}_2 &= \underline{q}_2 - (\underline{q}_2, \underline{\xi}_1)\,\underline{\xi}_1 , \\ \underline{\xi}_2 &= \frac{\underline{v}_2}{\|\underline{v}_2\|} \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \underline{v}_n &= \underline{q}_n - \sum_{n=1}^{n-1} (\underline{q}_n, \underline{\xi}_j)\,\underline{\xi}_j , \\ \underline{\xi}_n &= \frac{\underline{v}_n}{\|\underline{v}_n\|} ; \end{aligned} \tag{161}$$

(3) prześlij znaleziony zbiór wektorów bazy $\{\underline{\xi}\}$ w miejsce pamięci $\{\underline{\xi}^o\}$. Warunkiem jednoznaczności przejścia od jednego układu do drugiego jest spełnienie nierówności $d_j \neq 0$ dla $j = 1, 2, \ldots, n$, gdzie d_j jest sumą pomyślnych kroków wykonanych w j-tym kierunku. W praktycznych realizacjach metody Rosenbrocka warunek ten jest zawsze spełniony przez odpowiedni sposób doboru długości kroku przedstawiony na rys. 17 oraz stosowane kryterium oceny "Czy minimum".

Zamiast omówionego algorytmu obrotu można również stosować inny algorytm zaproponowany przez Powella [44], w którym nie wymaga się spełnienia warunku $d_j \neq 0$.

Kryterium zbieżności

Rosenbrock w swojej pracy [49] zaproponował, uznawać bieżący punkt $\underline{x}$ za minimum, jeśli w pięciu kolejnych "obie-

gach" nie miał miejsca pomyślny krok (tzn. nie była spełniona zależność $F < F_0$). Przez "obieg" rozumie się przy tym, zespół czynności polegających na wykonaniu po jednym kroku we wszystkich ortogonalnych kierunkach aktualnie obowiązującej bazy $\{\xi\}$ zgodnie z algorytmem rys. 17. Oczywistą jest rzeczą, że kryterium takie znacznie przedłuża czas działania procedury.

W konkretnych realizacjach metody Rosenbrocka stosuje się zazwyczaj o wiele prostsze kryterium, a mianowicie narzuca się z góry ilość iteracji L, po której następuje automatyczne zatrzymanie procedury. Jeśli po przeanalizowaniu wydrukowanych danych okaże się, że uzyskany wynik jest zadowalający, to na tym się poprzestaje, w przeciwnym przypadku przyjmuje się nową wartość na L i startuje się z ostatnio wyliczonego punktu $\underline{x}$.

5.2.3. Metoda simplexu Neldera i Meada - N

Metoda ta jest dalszym rozwinięciem metody Spendleya, Hexta i Himswortha [28]. Polega ona na utworzeniu w przestrzeni E^{n+1} n-wymiarowego simplexu*) o $n+1$ wierzchołkach w taki sposób, że można go wpisać w powierzchnię reprezentującą badaną funkcję celu $f(\underline{x})$. Wylicza się więc na wstępie procedury współrzędne punktów wierzchołkowych simplexu P_i (dla $i = 1, 2, \ldots, n+1$), zakładając przy tym arbitralnie pewną odległość między tymi wierzchołkami, zwaną "krokiem". W następnych iteracjach dokonuje się przekształceń simplexu w sposób przedstawiony w poniższym algorytmie tak długo, dopóki odległość między jego wierzchołkami w pobliżu szukanego minimum będzie nie większa od założonej dokładności ε.

Wprowadźmy oznaczenia:

P_n - wybrany punkt wierzchołkowy simplexu spośród $n+1$ wierzchołków P_i, w którym wartość funkcji badanej osiąga maksimum,

---

*) N-wymiarowym simplexem o $n+1$ wierzchołkach nazywamy zbiór wszystkich punktów określonych przez wektory

$$\underline{x} = \sum_{i=1}^{n+1} x_i \underline{\xi}_i, \quad \text{przy czym} \quad \sum_{i=1}^{n+1} x_i = 1 \quad \text{oraz} \quad x_i \geq 0,$$

gdzie $\underline{\xi}_i$ oznaczają wektory, a x_i - współrzędne punktów simplexu. Inaczej mówiąc, n-wymiarowy simplex jest wielościanem o $n+1$ wierzchołkach rozpiętych na $n+1$ wektorach bazowych. Przykłady simplexów: jednowymiarowym jest odcinek o dwóch wierzchołkach, dwuwymiarowym jest trójkąt o trzech wierzchołkach itd.

P_l - wybrany punkt wierzchołkowy simplexu spośród n+1 wierzchołków P_i, w którym wartość funkcji badanej osiąga minimum,

$\bar{P}$ - środek symetrii simplexu wyłączając P_h, zdefiniowany jako

$$\bar{P} = \frac{\sum_{i=1}^{n+1} P_i}{n}, \quad \text{przy czym } i \neq h. \tag{162}$$

W algorytmie metody Simplex stosuje się następujące trzy operacje:

1) operacja "odbicia" punktu P_h względem $\bar{P}$ określona przez

$$P^* = (1 + \alpha)\bar{P} - \alpha P_h, \tag{163}$$

2) operacja "ekspansji" punktu P^{**} względem $\bar{P}$ określona przez

$$P^{**} = (1 - \gamma)P^* - \gamma\bar{P}, \tag{164}$$

3) operacja "kontrakcji" punktu P_h względem $\bar{P}$ określona przez

$$P^{***} = \beta P_h + (1 - \beta)\bar{P}. \tag{165}$$

a. Informacje wejściowe

$\underline{x}_o$ - arbitralnie wybrany punkt startowy,

d - początkowa odległość pomiędzy wierzchołkami wyjściowego simplexu,

α - współczynnik odbicia $\alpha > 0$,

β - współczynnik kontrakcji $0 < \beta < 1$,

γ - współczynnik ekspansji $\gamma > 1$.

Wartości liczbowe współczynników α, β i γ należy dobierać eksperymentalnie. W przykładach rozpatrywanych przez Neldera i Meada jako optymalną strategię przyjęto $\alpha = 1$, $\beta = 0,5$, $\gamma = 2$,

ε - wymagana dokładność,

n - liczba zmiennych niezależnych,

b. Algorytm obliczeń

(1) oblicz wartości funkcji celu w punktach wierzchołkowych simplexu $F_i = f(P_i)$ dla $i = 1, 2, \ldots, n+1$.
Wyznacz h i l takie, że $f(P_h) = \max$, $f(P_l) = \min$ spośród zbioru F_i.

(2) Oblicz środek symetrii simplexu $\bar{P} = \frac{\sum_{i=1}^{n+1} P_i}{n}$, $i \neq h$. Wykonaj odbicie P^* punktu P_h względem $\bar{P}$. Oblicz $f(\bar{P}) = F_s$ i $f(P^*) = F_o$.

Jeżeli $F_o < \min$, to:

(3) Oblicz $P^{**} = (1 - \gamma)P^* - \gamma\bar{P}^*$ i $f(P^{**}) = F_e$. Gdy $F_e < \max$, to podstaw $P^{**} \rightarrow P_h$, w innym przypadku podstaw $P^* \rightarrow P_h$.

(4) O ile nie spełnione jest kryterium na minimum, powtórz procedurę od kroku (1).

Jeżeli $F_o > \min$, to:

(3) jeżeli $F_o \geq f(P_i)$ dla $i = 1,2,\ldots,n+1$, $i \neq h$ i $F_o \geq \max$ przejdź do realizacji kroku następnego, natomiast jeśli $F_o < \max$, to podstaw uprzednio $P^* \rightarrow P_h$,

(4) wykonaj kontrakcję P^{***} punktu P_h względem $\bar{P}$ i oblicz $f(P^{***}) = F_k$; jeżeli $F_k \geq \max$ to wykonaj redukcję simplexu, tzn. oblicz $P_i = \frac{P_i + P_l}{2}$, $i = 1,2,\ldots n+1$, natomiast gdy $P_k < \max$ to podstaw $P^{***} \rightarrow P_h$, a następnie przejdź do realizacji kroku (6).

(5) Jeżeli $F_o < f(P_i)$ dla $i = 1,2,\ldots n+1$, $i \neq h$, podstaw $P^* \rightarrow P_h$,

(6) o ile minimum nie zostało osiągnięte, powtórz procedurę począwszy od kroku (1).

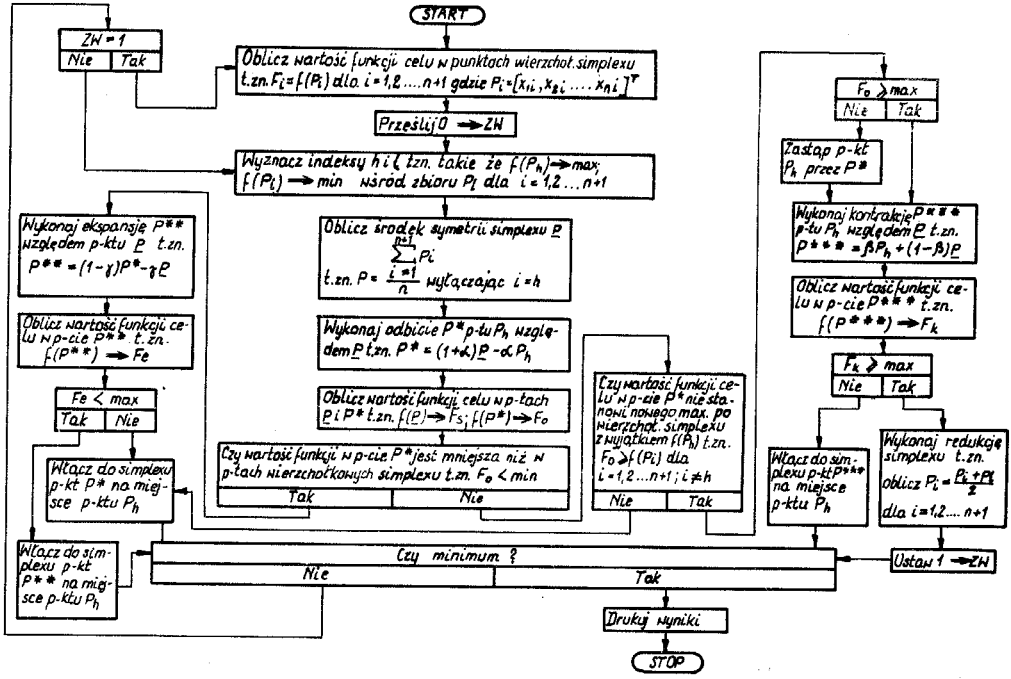
Sieć działań przytoczonego algorytmu przedstawiono na rys.18, przy czym przed przystąpieniem do jego wykonywania, na wstępie wyliczamy współrzędne punktów wierzchołkowych simplexu P_i (dla $i = 1,2,\ldots,n+1$).

c. Kryterium zbieżności

Jak już wspomniano, Nelder i Mead [39] jako kryterium zakończenia działania procedury iteracyjnej przyjęli warunek, że odległość pomiędzy punktami wierzchołkowymi simplexu będzie mniejsza od z góry założonej liczby ε.

5.2.4. Metoda Gaussa-Seidela - GA

Metoda ta zwana także "relaxacyjną" należy do drugiej grupy metod bezgradientowych, w których dla zapewnienia zbieżności stawia się ostrzejsze wymagania funkcji $f(\underline{x})$, niż w procedurach rozpatrywanych dotychczas. W metodach tych zakłada się, że $f(\underline{x})$ jest ograniczoną od dołu funkcją wypukłą klasy C^2 oraz, że w bliskim otoczeniu minimum można ją aproksymować formą kwadratową o postaci (159), przy czym macierz drugich pochodnych A jest dodatnio określona.


START
ZW = 1
Nie
Tak
Oblicz wartość funkcji celu w punktach wierzchoł. simplexu
t.zn. $F_i = f(P_i)$ dla $i = 1,2 \ldots n+1$ gdzie $P_i = [x_{1i}, x_{2i} \ldots x_{ni}]^T$
Prześlij 0 → ZW
Wyznacz indeksy h i ℓ tzn. takie że $f(P_h) \rightarrow max$; $f(P_\ell) \rightarrow min$ wśród zbioru P_i dla $i = 1,2 \ldots n+1$
Oblicz środek symetrii simplexu P
$P = \frac{\sum_{i=1}^{n+1} P_i}{n}$ wyłączając $i = h$
Wykonaj odbicie P* p-tu P_h względem P t.zn. $P^* = (1+\alpha)P - \alpha P_h$
Oblicz wartość funkcji celu w p-tach P i P* t.zn. $f(P) \rightarrow F_S$; $f(P^*) \rightarrow F_0$
Czy wartość funkcji w p-cie P* jest mniejsza niż w p-tach wierzchołkowych simplexu t.zn. $F_0 < min$
Tak
Nie
Czy wartość funkcji celu w p-cie P* nie stanowi największego max. po wierzchoł. simplexu z wyjątkiem $f(P_h)$ t.zn. $F_0 \geq f(P_i)$ dla $i = 1,2 \ldots n+1$; $i \neq h$
Nie
Tak
$F_0 \geq max$
Nie
Tak
Zastąp p-kt P_h przez P*
Wykonaj kontrakcję P*** p-tu P_h względem P t.zn. $P^{***} = \beta P_h + (1-\beta)P$
Oblicz wartość funkcji celu w p-cie P*** t.zn. $f(P^{***}) \rightarrow F_k$
$F_k \geq max$
Nie
Tak
Wykonaj redukcję simplexu t.zn. oblicz $P_i = \frac{P_i + P_\ell}{2}$ dla $i = 1,2 \ldots n+1$
Ustaw 1 → ZW
Wykonaj ekspansję P** względem p-ktu P t.zn. $P^{**} = (1-\gamma)P^* - \gamma P$
Oblicz wartość funkcji celu w p-cie P** t.zn. $f(P^{**}) \rightarrow Fe$
Fe < max
Tak
Nie
Włącz do simplexu p-kt P* na miejsce p-ktu P_h
Włącz do simplexu p-kt P** na miejsce p-ktu P_h
Włącz do simplexu p-kt P*** na miejsce p-ktu P_h
Czy minimum ?
Nie
Tak
Drukuj wyniki
STOP

Rys. 18

Istotą metody Gaussa-Seidela jest minimalizacja funkcji $f(\underline{x})$ wzdłuż kolejnych kierunków ortogonalnej bazy $\underline{\xi}_1, \underline{\xi}_2, \ldots, \underline{\xi}_n$, która utworzona jest z wersorów układu współrzędnych kartezjańskich i w trakcie obliczeń nie ulega zmianie. Tak więc, w przeciwieństwie do metod omawianych opoprzednio, w procedurze tej stosuje się minimalizację funkcji w kierunku, którą można zrealizować w sposób opisany w punkcie 5.1. Zbieżność metody GA oparta jest na następującym twierdzeniu.

Twierdzenie. Jeśli funkcja celu $f(\underline{x})$ ma postać $f(\underline{x}) = \frac{1}{2} \underline{x}^T A \underline{x}$, to metoda Gaussa-Seidela jest zbieżna do punktu ekstremalnego $\underline{\hat{x}}$, wtedy i tylko wtedy, gdy macierz A jest dodatnio określona.

Dowód. W rozpatrywanym przez nas przypadku gradient funkcji $f(\underline{x})$ wyrazi się przez

$$\nabla f = A \underline{x}, \tag{166}$$

stąd przesunięcie punktu w i-tej iteracji określone jest równaniem

$$\underline{\xi}_i^T A(\underline{x}_i + \lambda_i \underline{\xi}_i) = 0, \tag{167}$$

wynikającym z warunku koniecznego na istnienie ekstremum w kierunku.

Z równania tego otrzymujemy wartość λ_i minimalizującą funkcję $f(\underline{x})$ w kierunku $\underline{\xi}_i$

$$\lambda_i = -a^i \underline{x}_i / a_{ii} \tag{168}$$

gdzie: a^i oznacza i-ty wiersz macierzy A,
a_{ii} odpowiedni element tej macierzy.

Z drugiej strony wartość funkcji celu w punkcie $\underline{x}_{i+1} = \underline{x}_i + \lambda_i \underline{\xi}_i$ można przedstawić w następujący sposób:

$$f(\underline{x}_{i+1}) = f(\underline{x}_i) + \lambda_i \underline{\xi}_i^T A \underline{x}_i + \frac{1}{2} \lambda_i^2 a_{ii}, \tag{169}$$

skąd, po wykorzystaniu (168) uzyskujemy

$$f(\underline{x}_{i+1}) = f(\underline{x}_i) - \frac{1}{2} \lambda_i^2 a_{ii}. \tag{170}$$

Z zależności tej wynika, że w każdej następnej iteracji wartość funkcji celu będzie malała, bowiem elementy $a_{ii} > 0$, gdy A jest dodatnio określona, a ponieważ założ-

żono na wstępie, że funkcja $f(\underline{x})$ jest ograniczona od dołu więc tym samym wykazana została zbieżność algorytmu GA.

a. Informacje wejściowe

$\underline{x}_o$ - arbitralnie wybrany punkt startowy,

$\underline{\xi}_1, \underline{\xi}_2, \ldots, \underline{\xi}_n$ - baza wyjściowa utworzona z wzajemnie ortogonalnych wektorów,

e_o - początkowa długość kroku,

ε_j - wymagana dokładność obliczeń minimum w j-tym kierunku,

ε_o - wymagana dokładność obliczeń minimum globalnego,

n - liczba zmiennych niezależnych.

b. Algorytm obliczeń

(1) dla $j = 1, 2, \ldots, n$ oblicz λ_j minimalizujące

$$f(\underline{x}_{j-1} + \lambda_j \underline{\xi}_j)$$

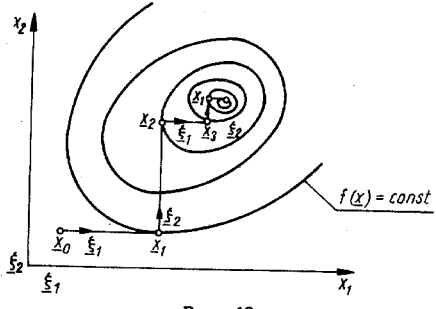
oraz współrzędne nowego punktu

$$\underline{x}_j = \underline{x}_{j-1} + \lambda_j \underline{\xi}_j ;$$

(2) zbadaj czy zostało spełnione kryterium zbieżności. Jeśli nie, to podstaw $\underline{x}_j$ w miejsce $\underline{x}_o$ i powtórz krok (1), w przeciwnym razie stop.

Rozpatrzmy działanie tej metody na przykładzie funkcji dwu zmiennych, której poziomice przedstawiono na rys. 19.

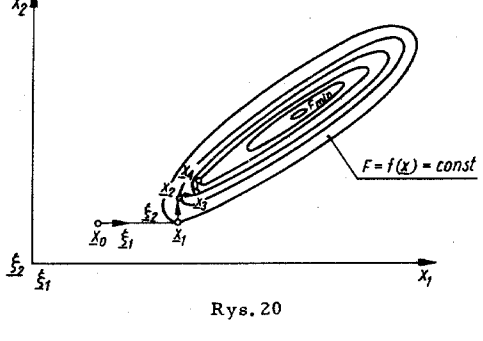
Przebieg powyższego algorytmu jest w tym przypadku następujący. Na wstępie zakładamy punkt startowy $\underline{x}_o$ oraz bazę wyjściową $\underline{\xi} = \{\underline{\xi}_1, \underline{\xi}_2\}$ pokrywającą się w naszym przykładzie z wersorami układu współrzędnych kartezjańskich. Następnie wykonujemy krok pierwszy tzn. startując z punktu $\underline{x}_o$ poszukujemy minimum funkcji w kierunku $\underline{\xi}_1$. Minimum to znajdujemy w punkcie $\underline{x}_1$. W drugim kroku powta-



Rys. 19

rzamy te same czynności z tym jednak, że startujemy teraz z punktu $\underline{x}_1$ oraz kierunkiem poszukiwań jest $\underline{\xi}_2$. Minimum znajdujemy w punkcie x_2. Jak wynika z rys. 19 procedurę tę powtarzamy tak długo, aż osiągniemy pożądane ekstremum.

Zauważmy jednak, jak mało efektywny staje się powyższy algorytm, gdy zamiast funkcji o regularnych eliptycznych poziomicach napotykamy na funkcję o charakterze wąskiej i długiej doliny. Przypadek ten przedstawiono na rys. 20.



Rys. 20

W celu uzyskania minimum F_{min} musimy teraz począwszy od punktu $\underline{x}_4$ poruszać się bardzo małą długością kroku. Widzimy stąd, jak bardzo zmalała szybkość zbieżności algorytmu, przy czym często się zdarza, że dla tego rodzaju powierzchni staje się on zawodny.

c. Kryterium zbieżności

W metodzie tej za minimum uznaje się punkt $\underline{x}$, jeżeli przesunięcie punktu d_j wzdłuż wszystkich kierunków ortogonalnej bazy będzie nie większe od założonej z góry dokładności ε tzn. $d_j \leq \underline{\xi}$, dla $j = 1, 2, \ldots, n$.

5.2.5. Metoda Daviesa, Swanna i Campeya - DSC

Dla usunięcia trudności omówionych w poprzednim punkcie została opracowana metoda DSC, będąca kompilacją metoda Gaussa-Seidela oraz metody Rosenbrocka. W metodzie DSC stosuje się więc, podobnie jak w procedurze GA, minimalizację wartości funkcji wzdłuż kolejnych ortogonalnych kierunków poszukiwań z tą różnicą, że w przypadku nie osiągnięcia minimum po cyklu złożonym z n takich kroków dokonuje się "obrotu współrzędnych" w myśl algorytmu Rosenbrocka. Jak wspomniano w punkcie 5.1, dla określ-

lenia ekstremum w kierunku skonstruowano procedurę charakteryzującą się następującymi dwiema właściwościami: 1) występowaniem zmiennej długości kroku, powiększanej stale 2-krotnie przy krokach pomyślnych (tzn. $f(\underline{x}_j) < f(\underline{x}_{j-1})$), aż do pierwszego kroku niepomyślnego $(f(\underline{x}_j) > f(\underline{x}_{j-1}))$ oraz 2) wyznaczaniem minimum w kierunku w oparciu o wzór interpolacyjny (152).

Przy tego rodzaju organizacji poszukiwań, czynność druga jest więc realizowana tylko wtedy, gdy wartości funkcji w trzech kolejnych punktach $\underline{x}_A$, $\underline{x}_B$, $\underline{x}_C$ spełniają układ nierówności: $F_B \leqslant F_C$ i $F_B \leqslant F_A$. Jak wykazały obliczenia tego typu algorytm cechuje dobra szybkość zbieżności dlatego też, został on zastosowany we wszystkich metodach bezgradientowych, w których dokonuje się minimalizacji w kierunku.

a. Informacje wejściowe

$\underline{x}_o$ - arbitralnie wybrany punkt startowy,

$\overset{o}{\underline{\xi}}_1, \overset{o}{\underline{\xi}}_2, \ldots, \overset{o}{\underline{\xi}}_n$ - baza wyjściowa utworzona z wzajemnie ortogonalnych wektorów,

$\underline{e}$ - wektor początkowej długości kroku,

β - współczynnik korekcyjny zmniejszający długość kroku e_j $0 < \beta < 1$,

ε - wymagana dokładność obliczeń minimum globalnego,

n - liczba zmiennych niezależnych.

b. Algorytm obliczeń

(1) dla $j = 1, 2, \ldots, n$ oblicz: λ_j minimalizujące $f(\underline{x}_{j-1} + \lambda_j \overset{o}{\underline{\xi}}_j)$ oraz zbiór $\underline{x}_j = \underline{x}_{j-1} + \lambda_j \overset{o}{\underline{\xi}}_j$,

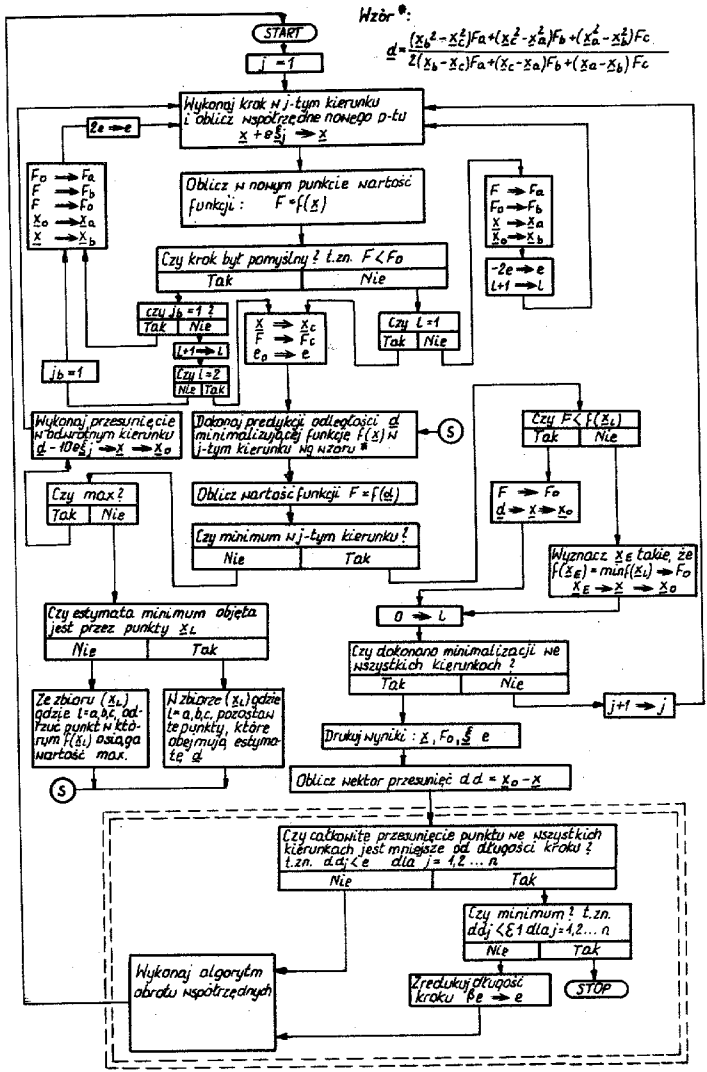
(2) oblicz wektor przesunięć $\underline{d} = \underline{x}_o - \underline{x}_n$,

(3) jeśli dla $j = 1, 2, \ldots, n$ $d_j > e_j$ lub $\varepsilon < d_j < e_j$ - dokonaj obrotu układu współrzędnych i powtórz procedurę od kroku (1), przy czym jeśli dla dowolnego $k \in [1, n]$ $d_k < \varepsilon$, podstaw uprzednio $d_k = 10\varepsilon$, natomiast gdy $\varepsilon < d_j < e_j$ podstaw $e_j = \beta d_j$.

Sieć działań przytoczonego algorytmu przedstawiono na rys 21, przy czym przed przystąpieniem do jego wykonywania, na wstępie wyliczamy w punkcie $\underline{x}_o$ wartość funkcji celu $F_o = f(\underline{x}_o)$.

c. Kryterium zbieżności

Analogiczne jak w metodzie Gaussa-Seidela pkt. 5.2.4c.


Wzór*:
$\underline{d} = \frac{(\underline{x}_b^2 - \underline{x}_c^2)F_a + (\underline{x}_c^2 - \underline{x}_a^2)F_b + (\underline{x}_a^2 - \underline{x}_b^2)F_c}{2(\underline{x}_b - \underline{x}_c)F_a + (\underline{x}_c - \underline{x}_a)F_b + (\underline{x}_a - \underline{x}_b)F_c}$
START
j = 1
Wykonaj krok w j-tym kierunku i oblicz współrzędne nowego p-tu
$\underline{x} + e\underline{s}_j \Rightarrow \underline{x}$
2e ⇒ e
Oblicz w nowym punkcie wartość funkcji : F = f(x)
Czy krok był pomyślny ? t.zn. F < Fo
Tak
Nie
$F_o \Rightarrow F_a$
$F \Rightarrow F_b$
$F \Rightarrow F_o$
$\underline{x}_o \Rightarrow \underline{x}_a$
$\underline{x} \Rightarrow \underline{x}_b$
$F \Rightarrow F_a$
$F_o \Rightarrow F_b$
$\underline{x} \Rightarrow \underline{x}_a$
$\underline{x}_o \Rightarrow \underline{x}_b$
-2e ⇒ e
l+1 ⇒ l
czy $j_b = 1$?
Tak
Nie
l+1 ⇒ l
Czy l = 2
Nie
Tak
$\underline{x} \Rightarrow \underline{x}_c$
$F \Rightarrow F_c$
$e_o \Rightarrow e$
Czy l = 1
Tak
Nie
$j_b = 1$
Wykonaj przesunięcie w odwrotnym kierunku
$\underline{d} - 10\underline{s}_j \Rightarrow \underline{x} \Rightarrow \underline{x}_o$
Dokonaj predykcji odległości $\underline{d}$ minimalizującej funkcję f(x) w j-tym kierunku wg wzoru *
S
Czy F < f(x_t)
Tak
Nie
Czy max?
Tak
Nie
Oblicz wartość funkcji F = f(x_t)
F ⇒ Fo
$\underline{d} \Rightarrow \underline{x} \Rightarrow \underline{x}_o$
Czy minimum w j-tym kierunku ?
Nie
Tak
Wyznacz $\underline{x}_E$ takie, że $f(\underline{x}_E) = \min\{f(\underline{x}_t)\} \Rightarrow F_o$
$\underline{x}_E \Rightarrow \underline{x} \Rightarrow \underline{x}_o$
Czy estymata minimum objęta jest przez punkty $\underline{x}_t$
Nie
Tak
0 ⇒ l
Czy dokonano minimalizacji we wszystkich kierunkach ?
Tak
Nie
j+1 ⇒ j
Ze zbioru $(\underline{x}_t)$ gdzie t = a,b,c, odrzuć punkt w którym f(x_t) osiąga wartość max.
W zbiorze $(\underline{x}_t)$ gdzie t = a,b,c, pozostałe punkty, które obejmują estymatę $\underline{d}$
Drukuj wyniki : $\underline{x}$, F_o, $\underline{s}$, e
S
Oblicz wektor przesunięć $\underline{d}\,\underline{d} = \underline{x}_o - \underline{x}$
Czy całkowite przesunięcie punktu we wszystkich kierunkach jest mniejsze od długości kroku ? t.zn. $\underline{d}\,d_j < e$ dla j = 1,2 ... n
Nie
Tak
Czy minimum ? t.zn. $\underline{d}\,d_j < \varepsilon_1$ dla j = 1,2 ... n
Nie
Tak
Wykonaj algorytm obrotu współrzędnych
Zredukuj długość kroku $\beta e \Rightarrow e$
STOP

Rys. 21

5.2.6. Metoda Powella i jej modyfikacje

Z rozważań nad metodą Gaussa-Seidela wynika, że w przypadku wystąpienia zadania optymalizacji, w którym poziomice funkcji celu mają kształt wąskich zakrzywionych dolin, metoda ta staje się bardzo wolno zbieżna. W tej sytuacji wyraźną poprawę szybkości zbieżności można uzyskać przez zastosowanie modyfikacji aktualnej bazy kierunków poszukiwań. Jednakże przy wykonywaniu tego rodzaju operacji należy pamiętać o zachowaniu niezależności liniowej kierunków w nowo utworzonej bazie, gdyż w przeciwnym razie może nastąpić redukcja jej wymiarowości, co w efekcie będzie prowadzić do niezbieżności metody. Jedno z rozwiązań tego problemu zostało przedstawione w omówionej już metodzie DSC (punkt 5.2.5), w której modyfikacji kierunków dokonuje się przez ortogonalny obrót układu współrzędnych. Zupełnie odmienną koncepcją posłużył się natomiast Powell w swoich dwóch procedurach [42]. Zaproponował on mianowicie, aby do istniejącej bazy, bądź co obieg *) (I wariant), bądź też po spełnieniu określonego warunku (II wariant), wprowadzać na miejsce jednego ze starych kierunków poszukiwań nowy kierunek, który byłby "sprzężony" do pozostałych.

Definicja. Dwa kierunki $\underline{\xi}_i$ oraz $\underline{\xi}_j$ są wzajemnie sprzężone względem dodatnio określonej macierzy A, jeśli

$$\underline{\xi}_i^T A \, \underline{\xi}_j = 0 \quad \text{dla} \quad i \neq j. \tag{171}$$

Jak zostało to wykazane [34], kierunki wzajemnie sprzężone są liniowo niezależne, a tym samym w I i II algorytmie Powella warunek jednoznaczności przekształcenia bazy kierunków poszukiwań jest zachowany.

I wariant metody Powella - P1

Koncepcja tej metody oparta została na następujących dwóch twierdzeniach:

Twierdzenie 1. Jeśli $\underline{\xi}_1, \underline{\xi}_2, \ldots, \underline{\xi}_n$ są wzajemnie sprzężonymi kierunkami względem dodatnio określonej macierzy A, a ponadto stanowią bazę rozpatrywanej przestrzeni, wtedy startując z dowolnie wybranego punktu $\underline{x}_0$, minimum formy kwadratowej

$$F = a + \underline{b}^T \underline{x} + \frac{1}{2} \underline{x}^T A \, \underline{x}, \tag{172}$$

może być wyznaczone w skończonej liczbie iteracji w wyniku minimalizacji funkcji $f(\underline{x})$ wzdłuż każdego z tych kierunków $\underline{\xi}_i$ tylko raz.

*) Przez obieg rozumie się dokonanie minimalizacji funkcji wzdłuż n kierunków obowiązującej bazy.

Dowód. Zwróćmy uwagę, że jeśli $\underline{\xi}_i$ są liniowo niezależne to dowolny wektor $\underline{\eta}$ może być wyrażony przez

$$\underline{\eta} = \sum_{i=1}^{n} \alpha_i \underline{\xi}_i, \tag{173}$$

gdzie

$$\alpha_i = \frac{\underline{\xi}_i^T A \underline{\eta}}{\underline{\xi}_i^T A \underline{\xi}_i}, \tag{174}$$

co bezpośrednio wynika z własności wzajemnie sprzężonych wektorów. Załóżmy, że po p iteracjach *) osiągnęliśmy punkt

$$\underline{x}_p = \underline{x}_o + \sum_{i=1}^{p} \lambda_i \underline{\xi}_i, \tag{175}$$

to zgodnie z (167), następną wartość λ_{p+1} minimalizującą funkcję $f(\underline{x})$ wzdłuż kierunku $\underline{\xi}_{p+1}$ można określić z równania

$$\underline{\xi}_{p+1}^T \nabla f(\underline{x}_{p+1}) = 0 \tag{176}$$

a więc

$$\underline{\xi}_{p+1}^T \left[A\left(\underline{x}_o + \sum_{i=1}^{p} \lambda_i \underline{\xi}_i + \lambda_{p+1} \underline{\xi}_{p+1} \right) + \underline{b} \right] = 0, \tag{177}$$

skąd

$$\lambda_{p+1} = - \frac{\underline{\xi}_{p+1}^T (A \underline{x}_o + \underline{b})}{\underline{\xi}_{p+1}^T A \underline{\xi}_{p+1}}, \tag{178}$$

gdyż wyrażenie $\underline{\xi}_{p+1}^T A \sum_{i=1}^{p} \lambda_i \underline{\xi}_i = 0$ z założenia.

Ze wzoru (178) wynika, że wartość λ_{p+1} zależy jedynie od położenia punktu startowego $\underline{x}_o$, natomiast sposób przejścia z punktu $\underline{x}_o$ do $\underline{x}_{p+1}$ nie ma na nią żadnego wpływu.

*) Przez iterację w tym przypadku rozumie się wyznaczenie minimum $f(\underline{x})$ w zadanym kierunku.

Po n iteracjach otrzymamy więc

$$\underline{x}_n = \underline{x}_o - \sum_{i=1}^{n} \frac{\underline{\xi}_i^T (A\,\underline{x}_o + \underline{b})\,\underline{\xi}_i}{\underline{\xi}_i^T A\,\underline{\xi}_i}, \qquad (179)$$

ale wykorzystując (173) i (174) widzimy, że (179) jest równoważne równaniu

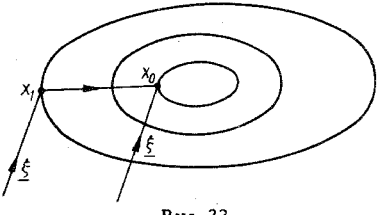
$$\underline{x}_n = \underline{x}_o - \underline{x}_o - A^{-1}\,\underline{b} = -A^{-1}\,\underline{b}, \qquad (180)$$

a to wskazuje, że zostało osiągnięte szukane ekstremum c.n.d.

Definicja. Mówimy, że procedura iteracyjna posiada zbieżność II rzędu, jeśli pozostaje w mocy twierdzenie 1.

Twierdzenie 2. Jeśli $\underline{x}_o$ jest minimum w kierunku $\underline{\xi}$ występującym w rozpatrywanej przestrzeni oraz jeśli $\underline{x}_1$ jest również minimum wzdłuż tego samego kierunku, to kierunek $(\underline{x}_1 - \underline{x}_o)$ łączący te dwa minima jest sprzężony z kierunkiem $\underline{\xi}$.

Interpretację geometryczną przytoczonego twierdzenia przedstawiono na rys. 22.



Rys. 22

Dowód. Z warunku koniecznego na ekstremum w kierunku mamy

$$\underline{\xi}^T (A\,\underline{x}_1 + \underline{b}) = 0, \qquad (181)$$

oraz

$$\underline{\xi}^T (A\,\underline{x}_o + \underline{b}) = 0, \qquad (182)$$

skąd po odjęciu stronami otrzymujemy

$$\underline{\xi}^T A (\underline{x}_1 - \underline{x}_o) = 0, \qquad (183)$$

a to przecież jest warunkiem wzajemnego sprzężenia kierunków c.n.d.

Biorąc pod uwagę obydwa podane twierdzenia można sformułować przebieg procedury P1. A więc na wstępie, podobnie jak to

miało miejsce w metodzie Gaussa i Seidela, dokonujemy minimalizacji wzdłuż n ortogonalnych kierunków $\underline{\xi}_1, \underline{\xi}_2, \ldots, \underline{\xi}_n$. Następnie po zakończeniu tego obiegu wyznaczamy nowy sprzężony kierunek $\underline{\xi}_{n+1}$, w myśl zasady

$$\underline{\xi}_{n+1} = \frac{\underline{x}_o - \underline{x}}{|\underline{x}_o - \underline{x}|} \tag{184}$$

oraz nowy punkt startowy $\underline{x}_o$, w rezultacie minimalizacji funkcji wzdłuż kierunku $\underline{\xi}_{n+1}$. Z kolei dokonujemy modyfikacji kierunków w ten sposób, że z bazy zostaje usunięty kierunek $\underline{\xi}_1$, a na jego miejsce zostaje włączony $\underline{\xi}_{n+1}$, przy jednoczesnej zmianie kolejności kierunków według reguły $\underline{\xi}_{k+1} \rightarrow \underline{\xi}_k$ dla $k = 1, \ldots n$. Opisany algorytm powtarzany jest tak długo, aż zostaje spełnione kryterium zbieżności procedury.

a. Informacje wejściowe

$\underline{x}_o$ - arbitralnie wybrany punkt startowy,

$\underline{\xi}_1, \underline{\xi}_2, \ldots, \underline{\xi}_n$ - baza wyjściowa utworzona z wzajemnie ortogonalnych wektorów,

e - początkowa długość kroku,

ε_j^o - wymagana dokładność obliczeń minimum w j-tym kierunku,

ε_o - wymagana dokładność obliczeń minimum globalnego,

n - liczba zmiennych niezależnych.

b. Algorytm obliczeń

(1) dla $j = 1, 2, \ldots, n$ oblicz λ_j minimalizujące $f(\underline{x}_{j-1} + \lambda_j \underline{\xi}_j)$ oraz współrzędne nowego punktu $\underline{x}_j = \underline{x}_{j-1} + \lambda_j \underline{\xi}_j$,

(2) wyznacz składowe kierunku sprzężonego w myśl wzoru

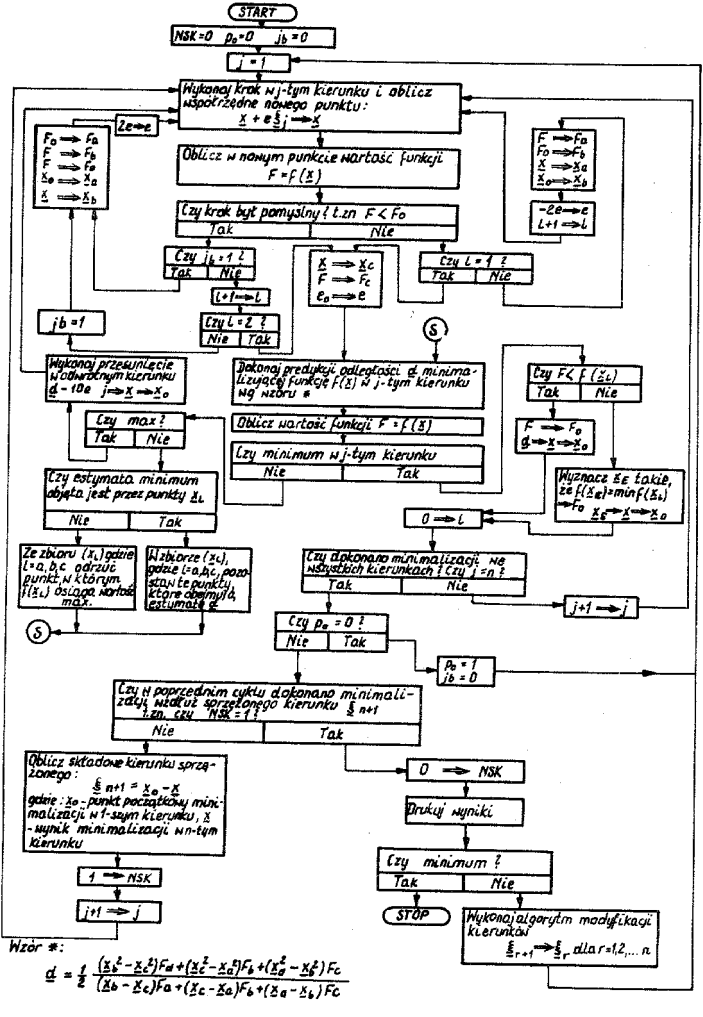
$$\underline{\xi}_{n+1} = \frac{\underline{x}_n - \underline{x}_o}{|\underline{x}_n - \underline{x}_o|},$$

(3) wzdłuż nowego kierunku $\underline{\xi}_{n+1}$ określ λ minimalizujące

$$f(\underline{x}_n + \lambda \underline{\xi}_{n+1})$$

oraz wyznacz współrzędne nowego punktu startowego

$$\underline{x}_{n+1} = \underline{x}_n + \lambda \underline{\xi}_{n+1} \rightarrow \underline{x}_o,$$


START
NSK=0 p_o=0 j_b=0
j=1
Wykonaj krok w j-tym kierunku i oblicz współrzędne nowego punktu: $\underline{x}+e\underline{s}_j=\underline{x}$
Ze=e
$F_o \Rightarrow F_a$
$F \Rightarrow F_b$
$F \Rightarrow F_o$
$\underline{x}_o \Rightarrow \underline{x}_a$
$\underline{x} \Rightarrow \underline{x}_b$
Oblicz w nowym punkcie wartość funkcji $F=f(\underline{x})$
Czy krok był pomyślny? t.zn. $F<F_o$
Tak
Nie
$F \Rightarrow F_a$
$F_o \Rightarrow F_b$
$\underline{x} \Rightarrow \underline{x}_a$
$\underline{x}_o \Rightarrow \underline{x}_b$
$-2e \Rightarrow e$
$L+1 \Rightarrow L$
Czy j_b=1 ?
Tak
Nie
$\underline{x} \Rightarrow \underline{x}_c$
$F \Rightarrow F_c$
$e_o \Rightarrow e$
Czy L=1 ?
Tak
Nie
L+1⇒L
j_b=1
Czy L=2 ?
Nie
Tak
Wykonaj przesunięcie w otrzymanym kierunku $\underline{d}=10e$ $\underline{x}\Rightarrow\underline{x}_o$
Czy max ?
Tak
Nie
Czy estymata minimum objęta jest przez punkty $\underline{x}_L$
Nie
Tak
Ze zbioru $(\underline{x}_L)$ gdzie L=a,b,c odrzuć punkt, w którym $f(\underline{x}_L)$ osiąga wartość max.
Wzbogać $(\underline{x}_L)$, gdzie L=a,b,c, pozostałe te punkty, które obejmują estymatę $\underline{x}_E$
S
Dokonaj predykcji odległości d, minimalizującej funkcję $F(\underline{x})$ w j-tym kierunku wg wzoru *
Oblicz wartość funkcji $F=f(\underline{x})$
Czy minimum w j-tym kierunku
Nie
Tak
S
Czy $F<f(\underline{x}_L)$
Tak
Nie
$F \Rightarrow F_o$
$\underline{d} \Rightarrow \underline{x} \Rightarrow \underline{x}_o$
Wyznacz $\underline{x}_E$ takie, że $f(\underline{x}_E)=\min f(\underline{x}_L)$
$F_o \Rightarrow \underline{x}_E \Rightarrow \underline{x} \Rightarrow \underline{x}_o$
0⇒L
Czy dokonano minimalizacji we wszystkich kierunkach ? Czy j=n ?
Tak
Nie
j+1⇒j
Czy p_o=0 ?
Nie
Tak
p_o=1
j_b=0
Czy w poprzednim cyklu dokonano minimalizacji wzdłuż sprzężonego kierunku $\underline{s}_{n+1}$ t.zn. czy NSK=1 ?
Nie
Tak
Oblicz składowe kierunku sprzężonego: $\underline{s}_{n+1}=\underline{x}_o-\underline{x}$ gdzie: $\underline{x}_o$ - punkt początkowy minimalizacji w 1-szym kierunku, $\underline{x}$ - wynik minimalizacji w n-tym kierunku
0⇒NSK
Drukuj wyniki
Czy minimum ?
Tak
Nie
STOP
Wykonaj algorytm modyfikacji kierunków $\underline{s}_{r+1}\Rightarrow\underline{s}_r$ dla r=1,2,... n
1⇒NSK
j+1⇒j
Wzór *:
$\underline{d}=\frac{1}{2}\frac{(\underline{x}_b^2-\underline{x}_c^2)F_a+(\underline{x}_c^2-\underline{x}_a^2)F_b+(\underline{x}_a^2-\underline{x}_b^2)F_c}{(\underline{x}_b-\underline{x}_c)F_a+(\underline{x}_c-\underline{x}_a)F_b+(\underline{x}_a-\underline{x}_b)F_c}$

Rys. 23

(4) dokonaj modyfikacji kierunków poszukiwań wg zasady

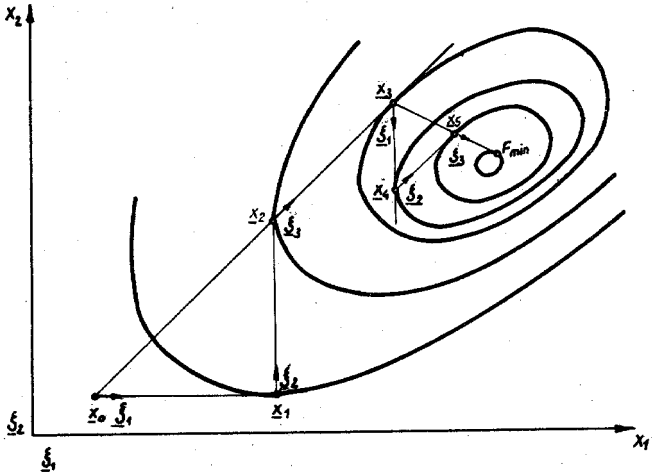
$$\underline{\xi}_{r+1} \Longrightarrow \underline{\xi}_{r} \quad \text{dla} \quad r = 1,2,\ldots,n$$

powtórz czynności od kroku (1), o ile nie spełnione jest kryterium na minimum.

Sieć działań przytoczonego algorytmu przedstawiono na rys. 23, przy czym przed przystąpieniem do jego wykonywania, na wstępie wyliczamy w punkcie $\underline{x}_o$ wartość funkcji celu $F_o = f(\underline{x}_o)$.

Z dotychczasowych rozważań wynikałoby, że procedura P1 powinna charakteryzować się zbieżnością II rzędu, jednakże jak wykazał to Zangwill [62], w niektórych przypadkach nie są spełnione założenia twierdzenia 2 i procedura Powella I traci tę właściwość. W celu zachowania zbieżności należy wtedy przed rozpoczęciem wykonywania "normalnej" procedury dokonać minimalizacji funkcji wzdłuż wszystkich kierunków ortogonalnej bazy początkowej. Spostrzeżenie to w postaci odpowiedniej zmiany zostało wprowadzone do sieci działań metody P1 (rys. 23).

Rozpatrzmy działanie omówionej procedury na przykładzie funkcji dwuzmiennych, której poziomice przedstawiono na rys.24



Rys. 24

Przebieg algorytmu w tym przypadku jest następujący. W pierwszym kroku, startując z punktu $\underline{x}_0$, dokonujemy minimalizacji F wzdłuż kierunku $\underline{\xi}_1$. Poszukiwanym punktem jest punkt $\underline{x}_1$. Następnie powtarzamy te same czynności wzdłuż kierunku $\underline{\xi}_2$ i w punkcie $\underline{x}_2$ osiągamy szukane minimum. W drugim kroku wyznaczamy składowe kierunku sprzężonego, przy czym w naszym przypadku $\underline{x}_n = \underline{x}_2$. Kierunek ten na rys. 24 oznaczono przez $\underline{\xi}_3$. W trzecim kroku wzdłuż tego nowego kierunku dokonujemy minimalizacji funkcji celu. Niech tym nowym punktem będzie punkt $\underline{x}_3$. Przeprowadzamy teraz modyfikację kierunku w myśl przyjętej zasady tzn. $\underline{\xi}_2 \Longrightarrow \underline{\xi}_1$ oraz $\underline{\xi}_3 \Longrightarrow \underline{\xi}_2$. Startując następnie z punktu $\underline{x}_3$ powtarzamy cały cykl od początku.

c. Kryterium zbieżności

W celu ustalenia warunków zakończenia działania procedury iteracyjnej Powell zaproponował następujący tok postępowania:

1) wykonywać "normalną" procedurę (zgodnie z punktem b) aż do momentu gdy w kolejnej iteracji przesunięcie punktu wzdłuż poszczególnych kierunków poszukiwań będzie mniejsze niż 0,1 wymaganej dokładności ε_0. Znaleziony punkt oznaczmy przez P_a,
2) obliczyć nowy punkt startowy procedury mnożąc współrzędne punktu P_a przez wielkość równą $10\,\varepsilon_0$.
3) powtórzyć czynności omówione w punkcie 1. Znaleziony punkt oznaczmy przez P_b.
4) znaleźć minimum funkcji wzdłuż linii przechodzącej przez punkty P_a i P_b. Znaleziony punkt oznaczmy przez P_c.
5) zakończyć działanie procedury jeśli $|P_a - P_c|$ oraz $|P_b - P_c|$ będą mniejsze od $0,1\,\varepsilon_0$, w przeciwnym przypadku wykonać punkt 6.
6) wyznaczyć nowy kierunek poszukiwań $\underline{\xi}_k$ równy

$$\underline{\xi}_k = \frac{P_a - P_c}{|P_a - P_c|},$$

a następnie włączyć go do bazy na miejsce $\underline{\xi}_1$.
7) przejść do ponownego wykonywania punktu 1.

Oczywiście, że powyżej opisane kryterium zbieżności jest bardzo ostre i wymagay dość dużego nakładu obliczeń, a tym samym przedłuża znacznie działanie procedury. Stąd też, w wielu przypadkach poprzestaje się tylko na spełnieniu warunku podanego w punkcie 1, bądź po prostu, jak to miało miejsce w procedurze Rosenbrocka, zakłada się z góry ilość iteracji L jaką należy wykonać. Następnie zaś w zależności od analizy

otrzymanych wyników podejmuje się dalsze decyzje. Tego rodzaju uproszczone kryteria muszą być jednak stosowane bardzo rozważnie, gdyż nie trudno o omyłkę. W przypadku bowiem funkcji silnie nieliniowej może się zdarzyć, że przez trzy lub więcej iteracji jest prawie niedostrzegalne przesunięcie punktu wzdłuż kierunków poszukiwań, chociaż obszar ten jest odległy od rzeczywistego minimum. Dlatego też jeśli procedura Powella ma być realizowana w postaci podprogramu lepiej stosować kryterium ogólniejsze przytoczone na wstępie.

II wariant metody Powella - P2

Drugi wariant tej metody jest bardzo podobny do poprzednio omówionego z tą jednak różnicą, że na innej zasadzie oparto w nim sposób tworzenia nowych kierunków poszukiwań. Jak można bowiem wykazać [62], w przypadkach wielowymiarowych istnieje możliwość powstawania w metodzie P1 kierunków zależnych liniowo. Stąd też, dla zabezpieczenia się przed tą ewentualnością w metodzie P2, zamiast dokonywać modyfikacji kierunków po każdym obiegu (jak w P1), zmiana kierunków następuje tylko przy spełnieniu warunku

$$\frac{\lambda_{max}\,\Delta}{\alpha} \geqslant 0{,}8, \tag{185}$$

gdzie: λ_{max} jest maksymalnym przesunięciem wzdłuż jednego z n kierunków poszukiwań (kierunek ten został oznaczony przez ξ_s),

Δ - wyznacznikiem macierzy utworzonej z n wektorów ξ_i,

α - całkowitym przesunięciem po dokonaniu kolejnego obiegu. Natomiast, w razie niespełnienia warunku (185) nowy obieg przebiega bez jakichkolwiek zmian tzn. obowiązują uprzednio przyjęte kierunki.

Uzasadnienie przedstawionego kryterium opiera się na następującym twierdzeniu:

Twierdzenie 3. Jeżeli kierunki $\xi_1, \ldots, \xi_n$ są tak wybrane, że

$$\xi_i\, A\, \xi_i = 1 \qquad i = 1, 2, \ldots, n. \tag{186}$$

to wyznacznik macierzy, której kolumnami są wektory ξ_i osiąga wartość maksymalną wtedy i tylko wtedy, gdy kierunki te są wzajemnie sprzężone.

Dowód. Niech będzie dany zbiór wzajemnie sprzężonych kierunków:

$$\eta_1, \eta_2, \ldots, \eta_n,$$

to z definicji będzie on miał własność

$$\underline{\eta}_i \; A \; \underline{\eta}_j = \delta_{ij}, \quad i = 1,2,\ldots,n; \quad j = 1,2,\ldots,n; \qquad (187)$$

gdzie: δ_{ij} jest deltą Kroneckera

$$\delta_{ij} = \begin{cases} 0 & \text{dla } i \neq j \\ 1 & \text{dla } i = j. \end{cases}$$

Załóżmy, że istnieje transformacja wiążąca $\{\underline{\xi}\}$ z $\{\underline{\eta}\}$

$$\underline{\xi}_i = \sum_{j=1}^{n} U_{ij} \, \underline{\eta}_j, \qquad (188)$$

to z warunku (187) otrzymujemy

$$\underline{\xi}_i \; A \; \underline{\xi}_j = \sum_{k=1}^{n} \sum_{l=1}^{n} U_{ik} \, U_{jl} \underline{\eta}_k \; A \; \underline{\eta}_l = \sum_{k=1}^{n} U_{ik} \, U_{jk}, \qquad (189)$$

a w szczególności

$$\sum_{k=1}^{n} U_{ik} \, U_{ik} = 1, \quad i = 1,2,\ldots,n. \qquad (190)$$

Oznacza to, że wyznacznik macierzy U nie przewyższa 1 i jest równy jedności tylko wówczas, gdy U jest macierzą ortogonalną. A więc

$$\underline{\xi}_i \; A \; \underline{\xi}_j = \delta_{ij}, \qquad (191)$$

a to wskazuje, że kierunki $\underline{\xi}_i$ są wzajemnie sprzężone c.n.d.

Wnioskiem z przytoczonego twierdzenia jest sposób dobierania kierunków $\underline{\xi}_1,\ldots,\underline{\xi}_n$ tak, aby wyznacznik macierzy był możliwie jak największy. Wynika stąd, że kryterium takie jest o wiele silniejsze niż stosowane w metodzie P1, gdyż wyklucza ono ewentualną zależność liniową pomiędzy nowo utworzonymi kierunkami. W celu pełnego wyjaśnienia zasady działania procedury P2, pozostało nam odpowiedzieć na pytanie, który z kierunków należy usunąć ze starej bazy, żeby na jego miejsce móc wprowadzić nowy, sprzężony kierunek. Okazuje się, że kierunkiem tym jest $\underline{\xi}_m$, wzdłuż którego następuje największe przesunięcie punktu. Rezultat ten wypływa z następującego rozumowania:

Załóżmy, że $\underline{x}_i$ jest punktem, w którym funkcja $f(\underline{x})$ osiąga minimum w kierunku $\underline{\xi}_i$, przy czym

$$\underline{\xi}_i \, A \, \underline{\xi}_i = 1, \tag{192}$$

to przesunięcie punktu wzdłuż tego kierunku można wyrazić

$$\left| \underline{x}_{i-1} - \underline{x}_i \right| \underline{\xi}_i = \alpha_i \, \underline{\xi}_i \, , \tag{193}$$

gdzie $\underline{x}_{i-1}$ jest punktem odpowiadającym minimum funkcji wzdłuż $\underline{\xi}_{i-1}$.

Natomiast po wykonaniu całego obiegu otrzymujemy

$$\underline{x}_n - \underline{x}_o = \alpha_1 \underline{\xi}_1 + \alpha_2 \underline{\xi}_2 + \ldots + \alpha_n \underline{\xi}_n, \tag{194}$$

gdzie: $\underline{x}_o$ oznacza punkt startowy,

$\underline{x}_n$ - punkt wynikowy po zakończeniu obiegu.

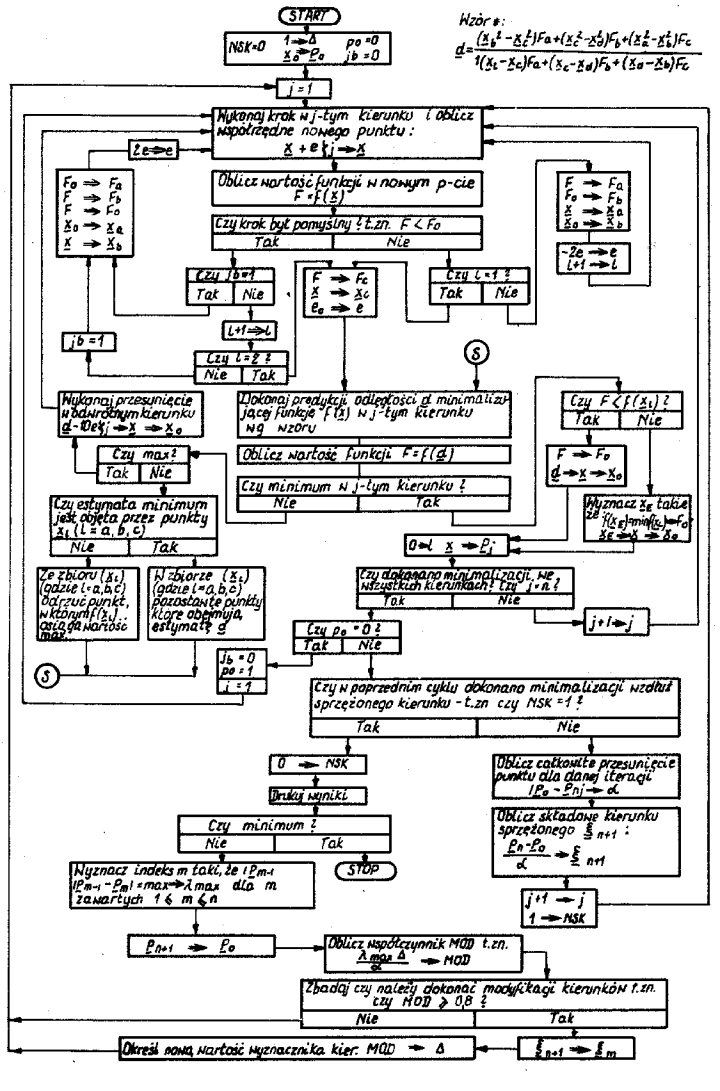
Z drugiej strony, zgodnie z zasadą tworzenia kierunku sprzężonego

$$\underline{x}_n - \underline{x}_o = \mu \underline{\xi}_{n+1} \quad \text{oraz} \quad \underline{\xi}_{n+1} \, A \, \underline{\xi}_{n+1} = 1, \tag{195}$$

a to wskazuje, że rezultatem zamiany jakiegokolwiek wektora kolumnowego $\underline{\xi}_i$ przez wektor sprzężony $\underline{\xi}_{n+1}$ jest pomnożenie wyznacznika macierzy kierunków przez czynnik $\frac{\alpha_i}{\mu}$. Stąd, najlepszy efekt z takiej zamiany uzyskamy wtedy, gdy kierunkiem tym będzie kierunek $\underline{\xi}_m$, dla którego przesunięcie α_i jest największe, przy czym zamiana ta będzie miała sens tylko wówczas, gdy $\alpha_m \geqslant \mu$. Tym więc, należy tłumaczyć istotę warunku (185), przy pomocy którego sprawdza się czy zastąpienie kierunku $\underline{\xi}_m$ przez $\underline{\xi}_{n+1}$ nie spowoduje, że nowy wyznacznik nie będzie mniejszy od z góry zadanej wartości (0,8), mającej zapewnić niezależność liniową nowo utworzonego zbioru kierunków. W przypadku, gdy warunek ten nie zostaje spełniony to w myśl procedury P2, ustalone w poprzednim "obiegu" kierunki pozostają bez zmian, co z kolei znacznie opóźnia otrzymanie n wzajemnie sprzężonych kierunków. Zwiększa się tym samym ilość obiegów i czas działania procedury, lecz teoretycznie niezawodność metody wydaje się być zagwarantowana.

a. Informacje wejściowe
Analogicznie jak w I wariancie metody Powella.

b. Algorytm obliczeń
Iterację k rozpoczynamy podstawiając k = 1:


START
NSK=0
1⇒Δ
x0⇒P0
po=0
jb=0
j=1
Wykonaj krok w j-tym kierunku i oblicz współrzędne nowego punktu:
x + e ⇒ x
Ze⇒e
Oblicz wartość funkcji w nowym p-cie
F=f(x)
Czy krok był pomyślny? t.zn. F < Fo
Tak
Nie
Fo ⇒ Fa
F ⇒ Fb
F ⇒ Fo
x0 ⇒ xa
x ⇒ xb
F ⇒ Fa
Fo ⇒ Fb
x ⇒ xa
x0 ⇒ xb
-Ze ⇒ e
L+1 ⇒ L
Czy jb=1
Tak
Nie
F ⇒ Fc
x ⇒ xc
e ⇒ e
Czy L=1?
Tak
Nie
jb=1
L+1⇒L
Czy L=2?
Nie
Tak
S
Wykonaj przesunięcie w odwrotnym kierunku
d-Dek ⇒ x ⇒ x0
Wykonaj predykcję odległości d minimalizującej funkcję f(x) w j-tym kierunku wg wzoru
Czy F < f(x1)?
Tak
Nie
Czy max?
Tak
Nie
Oblicz wartość funkcji F=f(d)
F ⇒ Fo
d ⇒ x ⇒ x0
Czy minimum w j-tym kierunku?
Nie
Tak
Czy estymata minimum jest objęta przez punkty x1 (i=a,b,c)
Nie
Tak
Wyznacz xe takie że f(xe)=min(fi) ⇒ Fo
xe ⇒ x ⇒ δe
0⇒i x ⇒ Pj
Ze zbioru (xi) (gdzie i=a,b,c) odrzuć punkt, w którym f(xi) osiąga wartość max.
W zbiorze (xi) (gdzie i=a,b,c) przesuwanie punkty które obejmują estymatę g
Czy dokonano minimalizacji, we wszystkich kierunkach? Czy j=n?
Tak
Nie
j+1⇒j
Czy po=0?
Tak
Nie
jb=0
po=1
i=1
S
Czy w poprzednim cyklu dokonano minimalizacji wzdłuż sprzężonego kierunku - t.zn czy NSK=1?
Tak
Nie
0 ⇒ NSK
Oblicz całkowite przesunięcie punktu dla danej iteracji
/P0 - Pn/ ⇒ α
Drukuj wyniki
Oblicz składowe kierunku sprzężonego ξ n+1:
Pn-P0 / α ⇒ ξ n+1
Czy minimum?
Nie
Tak
STOP
Wyznacz indeks m taki, że /Pm-1 - Pm/ = max ⇒ λmax dla m zawartych 1 ≤ m ≤ n
j+1 ⇒ j
1 ⇒ NSK
Pn+1 ⇒ P0
Oblicz współczynnik MOD t.zn. λmax Δ / α ⇒ MOD
Zbadaj czy należy dokonać modyfikacji kierunków t.zn. czy MOD ≥ 0,8?
Nie
Tak
Określ nową wartość wyznacznika kier. MOD ⇒ Δ
ξ n+1 ⇒ ξ m
Wzór ▪:
d = [(xb² - xc²)Fa + (xc² - xa²)Fb + (xa² - xb²)Fc] / [(xb - xc)Fa + (xc - xa)Fb + (xa - xb)Fc]

Rys. 25

(1) dla $j = 1,2,\ldots,n$ oblicz λ_j^k minimalizując

$f\left(\underline{x}_{j-1}^k + \lambda_j^k \underline{\xi}_j^k\right)$ i określ $\underline{x}_j^k = \underline{x}_{j-1}^k + \lambda_j^k \underline{\xi}_j^k$;

(2) określ $\alpha^k = \left|\underline{x}_n^k - \underline{x}_o^k\right|$ i $\underline{\xi}_{n+1}^k = \left(\underline{x}_n^k - \underline{x}_o^k\right)/\alpha^k$.
Oblicz λ_{n+1}^k minimalizujące $f\left(\underline{x}_n^k + \lambda_{n+1}^k \underline{\xi}_{n+1}^k\right)$ i określ zbiór $\underline{x}_o^{k+1} = \underline{x}_{n+1}^k = \underline{x}_n^k + \lambda_{n+1}^k \underline{\xi}_{n+1}^k$;

(3) znajdź $\lambda_s^k = \max \lambda_j^k$ dla $j = 1,2,\ldots,n$
przypadek (a): jeżeli $\lambda_s^k \Delta^k/\alpha^k \geq 0,8$, niech $\underline{\xi}_j^{k+1} = \underline{\xi}_j^k$
dla $j \neq s$, $\underline{\xi}_s^{k+1} = \underline{\xi}_{n+1}^{k+1}$ i niech $\Delta^{k+1} = \lambda_s^k \Delta^k/\alpha^k$;
przypadek (b): jeżeli $\lambda_s^k \Delta^k/\alpha^k < 0,8$, niech $\underline{\xi}_j^{k+1} = \underline{\xi}_j^k$, $j = 1,2,\ldots,n$ i $\Delta^{k+1} = \Delta^k$. Powtórz czynności od kroku (1) zwiększając k o 1 tzn. $k+1 \rightarrow k$.

Sieć działań przytoczonego algorytmu przedstawiono na rys. 25, przy czym przed przystąpieniem do jego wykonywania, na wstępie wyliczamy w punkcie $\underline{x}_o$ wartość funkcji celu $F_o = f(\underline{x}_o)$.

c. Kryterium zbieżności

Analogicznie jak w I wariancie metody Powella.

5.2.7. Metoda Zangwilla - Z

Metoda ta powstała jako dalsze rozwinięcie metod Powella. Zangwill w swojej pracy [62] wykazał, że w niektórych przypadkach pierwsza procedura Powella nie posiada zbieżności drugiego rzędu oraz przytoczył przykłady na potwierdzenie tego faktu. Zwrócił on przy tym uwagę, że dla zapewnienia zbieżności wystarczy wtedy przed przystąpieniem do wykonywania "normalnej" procedury Powella, dokonać minimalizacji funkcji wzdłuż wszystkich kierunków ortogonalnej bazy początkowej. Spostrzeżenie to posłużyło mu dalej do zbudowania własnej procedury będącej dalszą modyfikacją I wariantu metody Powella. Modyfikacja ta polega na wprowadzeniu dodatkowego zbioru ortogonalnych kierunków poszukiwań $\{\underline{c}\}$, który w trakcie przebiegu algorytmu nie ulega zmianom. Za każdym razem przed wykonaniem jednego "obiegu" procedury Powella dokonuje się minimalizacja funkcji wzdłuż ortogonalnego kierunku $\underline{c}_t$, począwszy od wartości $t = 1$ zmienianej cyklicznie o 1 co krok. Jeśli próba taka jest pomyślna tzn. prze-

sunięcie $\alpha_t \neq 0$, to następuje skok do procedury Powella i wykonany zostaje jej "obieg", jeśli natomiast $\alpha_t = 0$, to następuje minimalizacja funkcji wzdłuż następnego kierunku c_t (tzn. dla $t = t+1$) i badanie rozpoczyna się od początku. Po n kolejnych niepomyślnych próbach wzdłuż kierunków c znaleziony punkt uznaje się za minimum. Procedura Zangwilla posiada zbieżność II rzędu.

a. Informacje wejściowe

$\underline{x}_o$ - arbitralnie wybrany punkt startowy,

$\underline{\xi}_1^o, \underline{\xi}_2^o, \ldots, \underline{\xi}_n^o$ - baza wyjściowa dla pętli Powella,

$\underline{c}_1, \underline{c}_2, \ldots, \underline{c}_n$ - baza ortogonalna nie ulegająca zmianom w trakcie działania procedury, identyczna na wstępie ze zbiorem wektorów $\left\{\underline{\xi}^o\right\}$;

e_o - początkowa długość kroku,

ε_j - wymagana dokładność obliczeń w j-tym kierunku,

ε_o - wymagana dokładność obliczeń minimum globalnego,

n - liczba zmiennych niezależnych.

b. Algorytm obliczeń

Krok wstępny: oblicz λ_n^o minimalizujące $\left(f\ \underline{x}_n^o + \lambda_n^o\ \underline{\xi}_n^1\right)$ i określ $\underline{x}_{n+1}^o = \underline{x}_n^o + \lambda_n^o \underline{\xi}_n^1$.

Podstaw $t = 1$ i rozpocznij iterację k dla $k = 1$:

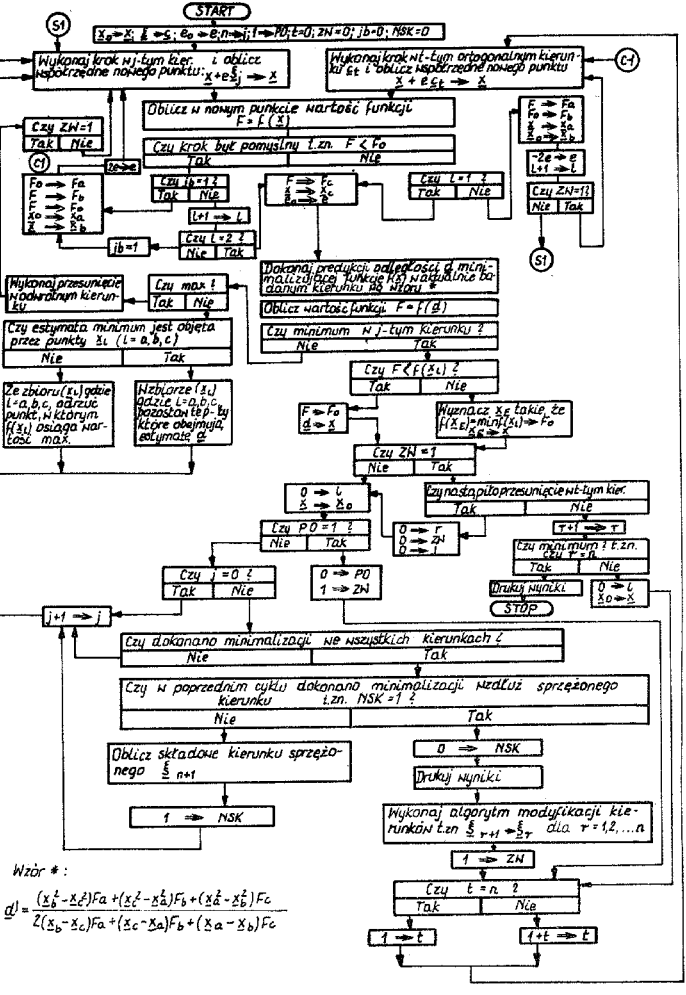
(1) oblicz α minimalizujące $f\left(\underline{x}_{n+1}^{k-1} + \alpha \underline{c}_t\right)$ i ustaw

$$t = \begin{cases} t+1 & \text{jeżeli } 1 \leqslant t < n \\ 1 & \text{jeżeli } t = n \end{cases}$$

Jeżeli $\alpha \neq 0$ oblicz $\underline{x}_o^k = \underline{x}_{n+1}^{k-1} + \alpha \underline{c}_t$ oraz przejdź do wykonania kroku (2), jeżeli $\alpha = 0$ powtórz krok (1) w przypadku, gdy krok (1) został powtórzony kolejno n razy to punkt $\underline{x}_{n+1}^{k-1}$ jest uznany za poszukiwane minimum;

(2) dla $j = 1,2,\ldots,n$ oblicz λ_j^k minimalizujące $f\left(\underline{x}_{j-1}^k + \lambda_j^k \underline{\xi}_j^k\right)$ i określ $\underline{x}_j^k = \underline{x}_{j-1}^k + \lambda_j^k \underline{\xi}_j^k$. Niech $\underline{\xi}_{n+1}^k = \left(\underline{x}_n^k - \underline{x}_{n+1}^{k-1}\right) \Big/ \left|\underline{x}_n^k - \underline{x}_{n+1}^{k-1}\right|$;

(3) oblicz λ_{n+1}^k minimalizujące $f\left(\underline{x}_n^k + \lambda_{n+1}^k \underline{\xi}_{n+1}^k\right)$ i określ $\underline{x}_{n+1}^k = \underline{x}_n^k + \lambda_{n+1}^k \underline{\xi}_{n+1}^k$. Dokonaj modyfikacji kierunków



Wzór *:

$$d^{l} = \frac{(x_b^t - x_c^t)Fa + (x_c^t - x_a^t)Fb + (x_a^t - x_b^t)Fc}{2(x_b - x_c)Fa + (x_c - x_a)Fb + (x_a - x_b)Fc}$$

Rys. 26

w myśl zasady:

$$\xi_j^{k+1} = \xi_{j+1}^{k} \qquad \text{dla} \quad j = 1, 2, \ldots, n.$$

Powtórz iterację k podstawiając k+1 w miejsce k.

Sieć działań przytoczonego algorytmu przedstawiono na rys. 26, przy czym przed przystąpieniem do jego wykonywania, na wstępie wyliczamy w punkcie $\underline{x}_o$ wartość funkcji celu $F_o = f(\underline{x}_o)$.

c. Kryterium zbieżności

Jako kryterium zbieżności Zangwill zaproponował podobne kryterium jak przyjęto w metodzie Gaussa-Seidela tzn. bieżący punkt $\underline{x}$ uznaje się za minimum, jeśli kolejne przesunięcie punktu α_t wzdłuż ortogonalnych kierunków $\underline{c}_t$ dla $t = 1, 2, \ldots, n$ są mniejsze od z góry założonej dokładności obliczeń, a więc

$$\alpha_t \leq \varepsilon_o, \qquad t = 1, 2, \ldots, n.$$

5.3. Metody gradientowe poszukiwania ekstremum

W odróżnieniu od rozważanych w poprzednim paragrafie metod bezgradientowych, metody gradientowe w czasie przebiegu algorytmu korzystają zarówno z informacji o wartości funkcji jak i o wartości jej gradientu. Metody te są na ogół o wiele szybciej zbieżne od metod bezgradientowych, lecz nie zawsze mogą być stosowane ze względu na brak znajomości gradientu funkcji. Wymagają one ponadto przynajmniej tak ostrych założeń jak druga grupa metod bezgradientowych. Tak więc, przy wykazywaniu ich zbieżności zakłada się, że funkcja celu $f(\underline{x})$ jest ograniczoną od dołu funkcją wypukłą klasy C^2, a przy tym taką, że można ją aproksymować formą kwadratową o postaci

$$f(\underline{x}) = a + \underline{b}^T \underline{x} + \frac{1}{2} \underline{x}^T A \underline{x}, \tag{196}$$

gdzie A oznacza macierz symetryczną dodatnio określoną, której elementami są drugie pochodne cząstkowe funkcji $f(\underline{x})$.

Wspólną cechą omawianych metod, z wyjątkiem metody gradientu prostego (GP), jest jednolity sposób realizacji kolejnej "iteracji" rozumianej jako zespół czynności, które należy wykonać dla przesunięcia bieżącego punktu $\underline{x}_i$ do $\underline{x}_{i+1}$. Na przesunięcie to składają się dwie następujące operacje:

1) wyznaczenie kierunku poszukiwań,
2) określenie minimum w tym kierunku.

Różnice występujące pomiędzy rozpatrywanymi metodami (poza GP) polegają jedynie na odmiennym sposobie wyznaczania kierunków poszukiwań $\underline{\xi}$, natomiast algorytm określania minimum w kierunku bywa na ogół ten sam. Dla zwiększenia efektywności metod korzysta się w nim zazwyczaj z interpolacji sześciennej dwupunktowej przedstawionej w punkcie 5.1.3.

Przejdźmy teraz do omówienia poszczególnych metod przy czym założymy, że będziemy poszukiwać tylko minimum funkcji $f(\underline{x})$.

5.3.1. Metoda Gradientu Prostego - GP

Metoda ta, zaliczana do klasycznych metod gradientowych rzadko już dziś jest stosowana, ze względu na małą jej efektywność. W metodzie GP realizacja kolejnej iteracji przebiega w sposób opisany powyżej z tą jednak różnicą, że nie dokonuje się w niej minimalizacji funkcji w kierunku. Polega ona więc na wyznaczeniu w bieżącym punkcie aktualnego kierunku poszukiwań, przez przypisanie mu wartości minus grandientu w tym punkcie, a następnie wzdłuż tego kierunku wykonaniu kroku o długości e. Procedurę tę powtarzamy tak długo, aż zostanie spełnione przyjęte kryterium "czy minimum". Dowód zbieżności metody GP można znaleźć w pracy [34].

a. Informacje wejściowe

$\underline{x}_o$ - arbitralnie wybrany punkt startowy,
e - początkowa długość kroku,
β - współczynnik korekcyjny zmniejszający e_o; $0 < \beta < 1$,
ε - wymagana dokładność obliczeń minimum,
n - liczba zmiennych niezależnych.

b. Algorytm obliczeń

(1) oblicz w punkcie startowym $\underline{x}_o$ wartość funkcji celu $F_o = f(\underline{x}_o)$ oraz jej gradientu $\underline{g}_o = g(\underline{x}_o)$;

(2) wyznacz kierunek poszukiwań

$$\underline{\xi} = -\underline{g}_o;$$

(3) wzdłuż kierunku $\underline{\xi}$ wykonaj krok o długości e oraz określ współrzędne nowego punktu tzn.

$$\underline{x}_{i+1} = \underline{x}_i + e\underline{\xi},$$

przy czym dla pierwszej iteracji $\underline{x}_i = \underline{x}_o$;

(4) oblicz w nowych punkcie wartość funkcji $F = f(\underline{x}_{i+1})$ oraz gradientu $\underline{g} = g(\underline{x}_{i+1})$.

Jeśli krok był pomyślny tzn.

$$F < F_o,$$

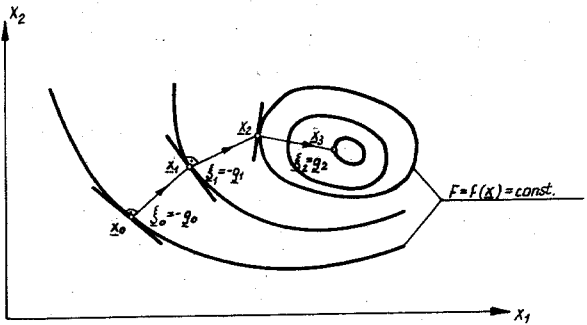
to powtórzyć cykl od punktu 2 przesyłając $g \Longrightarrow g_o$. W przypadku przeciwnym przejdź do wykonania punktu 5.

(5) zbadaj "czy minimum", jeśli nie, cofnij się do punktu poprzedniego tzn.

$$\underline{x}_i = \underline{x}_{i+1} - e\underline{\xi}$$

oraz zmniejsz krok o β. Powtórz cykl od punktu 3.

Przebieg powyższego algorytmu na przykładzie funkcji dwuzmiennych przedstawiono na rys. 27.



Rys. 27

c. Kryterium zbieżności

Analogiczne jak w metodzie gradientu sprzężonego punkt 5.3.3.

5.3.2. Metoda Najszybszego Spadku i jej modyfikacje - NS

Metoda ta stanowi dalsze rozwinięcie metody gradientu prostego, które polega na zastosowaniu minimalizacji funkcji wzdłuż wyznaczonego kierunku minus gradientu, zamiast wykonywania pojedynczego kroku jak w metodzie GP. Dowód zbieżności metody został podany przez Forsythe'a [34], który wykazał, że przy przyjętych powyżej założeniach metoda NS posiada zbieżność przynajmniej typu geometrycznego. Dowód ten został oparty na następującym rozumowaniu:

Załóżmy, że funkcje celu w otoczeniu ekstremum wyraża się przez

$$f(\underline{x}) = \frac{1}{2}\,\underline{x}^{T}\,A\,\underline{x}, \tag{197}$$

a gradient funkcji

$$\nabla f(\underline{x}) = A\,\underline{x}, \tag{198}$$

to zgodnie z opisanym algorytmem przesunięcie punktu $\underline{x}_i$ do $\underline{x}_{i+1}$ można przedstawić jako

$$\underline{x}_{i+1} = \underline{x}_i - \gamma A\,\underline{x}_i, \tag{199}$$

bądź

$$\underline{x}_{i+1} = \frac{P_1(A)}{P_1(0)}\,\underline{x}_i, \tag{200}$$

gdzie $P_1(\lambda) = 1 - \gamma\lambda$ przy czym λ oznacza wartość własną macierzy A.

Utwórzmy wielomian pierwszego stopnia macierzy A - $Q_1(\lambda)$, który zdefiniujemy następująco:

$$\begin{aligned} Q_1(\lambda_1) &= \alpha_o + \alpha_1\lambda_1 = -1, \\ Q_1(\lambda_n) &= \alpha_o + \alpha_1\lambda_n = 1, \end{aligned} \tag{201}$$

gdzie λ_1 i λ_n są najmniejszą i największą wartością własną macierzy A.

Ponieważ A z założenia jest dodatnio określona to wszystkie jej wartości własne są również dodatnie. Z zależności (201) po prostych przekształceniach otrzymujemy więc:

$$Q_1(\lambda) = \frac{2\lambda - (\lambda_1 + \lambda_n)}{\lambda_n - \lambda_1}. \tag{202}$$

Wobec tego możemy napisać

$$f(\underline{x}_{i+1}) = f\left(\frac{P_1(A)}{P_1(0)}\cdot\underline{x}_i\right) \leqslant f\left(\frac{Q_1(A)}{Q_1(0)}\cdot\underline{x}_i\right), \tag{203}$$

a wyrażając bieżący punkt $\underline{x}_i$ przez

$$\underline{x}_i = \sum_{j=1}^{n} a_{ij}\,\underline{v}_j, \tag{204}$$

gdzie $\underline{v}_j$ oznacza wektory własne odpowiadające wartościom własnym λ_j dla $j = 1, \ldots, n$.

Korzystając z założenia (197) nierówność (203) można sprowadzić do postaci

$$f(x_{i+1}) \leq f\left(\frac{Q_1(A)}{Q_1(0)}\ x_i\right) = \frac{1}{2}\left[\frac{Q_1(A)}{Q_1(0)}\ \underline{x}_i\right]^T A \frac{Q_1(A)}{Q_1(0)}\ \underline{x}_i, \quad (205)$$

skąd

$$f(x_{i+1}) \leq \frac{1}{2}\frac{1}{Q_1(0)^2} \sum_j a_{ij}^2\ Q_1(\lambda_j)^2\ \lambda_j \leq \frac{1}{2}\frac{1}{Q_1(0)^2} \sum_j a_{ij}^2\ \lambda_j, \quad (206)$$

bo przecież z definicji $\left|Q_1(\lambda_j)\right| \leq 1$, $j = 1, 2, \ldots, n$, a więc ostatecznie

$$f(\underline{x}_{i+1}) \leq \frac{1}{Q_1(0)^2}\ f(\underline{x}_i) \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 f(\underline{x}_i), \quad (207)$$

co należało wykazać.

a. Informacje wejściowe
   - $\underline{x}_o$ - arbitralnie wybrany punkt startowy,
   - e - początkowa długość kroku,
   - ε_j - wymagana dokładność obliczeń minimum w aktualnie występującym kierunku poszukiwań,
   - ε_o - wymagana dokładność obliczeń minimum globalnego,
   - n - liczba zmiennych niezależnych.

b. Algorytm obliczeń

   (1) oblicz w punkcie startowym $\underline{x}_o$ wartość funkcji celu $F_o = f(\underline{x}_o)$ oraz jej gradientu $\underline{g} = g(\underline{x}_o)$;

   (2) wyznacz kierunek poszukiwań

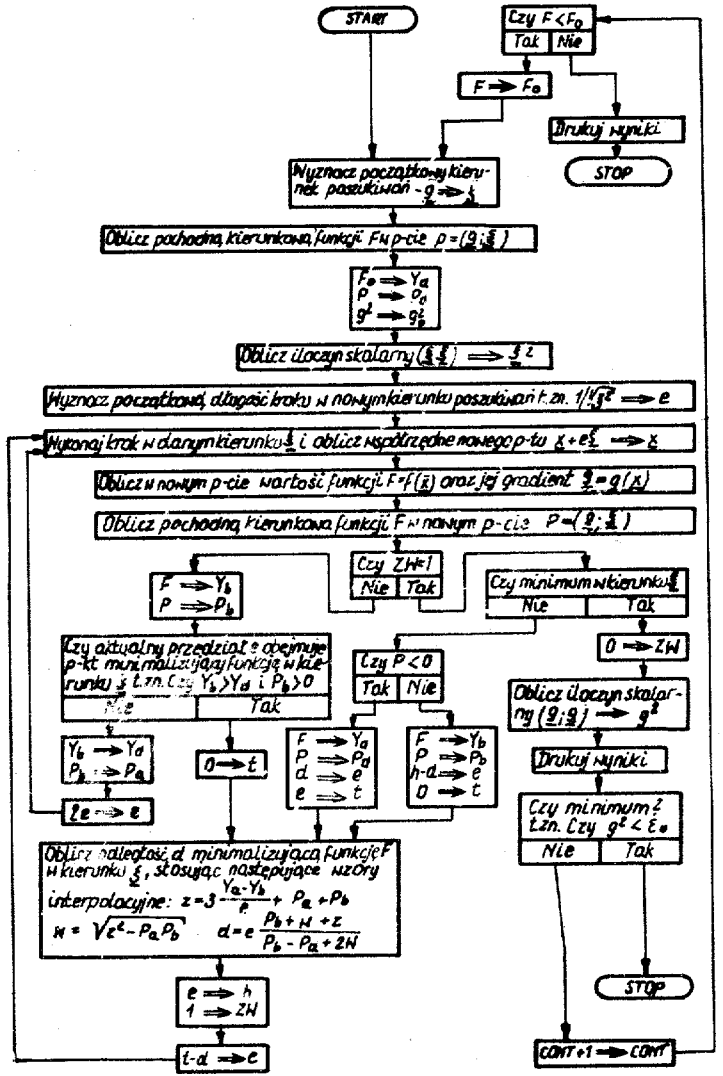
   $$\underline{\xi}_i = -\underline{g};$$

   (3) wzdłuż kierunku $\underline{\xi}_i$ określ λ_i minimalizujące $f(\underline{x}_{i-1} + \lambda_i \underline{\xi}_i)$ oraz współrzędne nowego punktu $\underline{x}_i = \underline{x}_{i-1} + \lambda_i \underline{\xi}_i$;

   (4) oblicz w punkcie $\underline{x}_i$ wartość gradientu $\underline{g} = g(\underline{x}_i)$;

   (5) zbadaj "czy minimum". Jeśli tak to stop, natomiast jeśli nie to powtórz czynności od kroku (2).

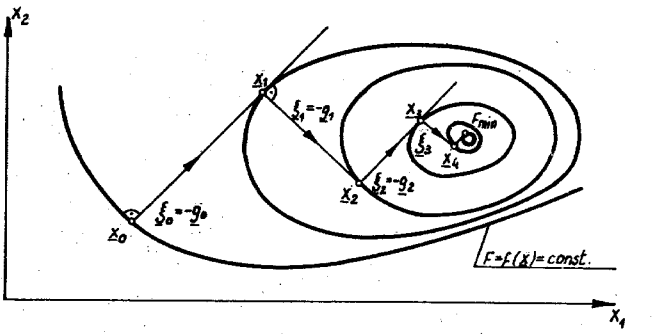
Sieć działań przytoczonego algorytmu przedstawiono na rys. 28, przy czym przed przystąpieniem do jego wykonywania,


START
Czy F<F0
Tak
Nie
F ⇒ F0
Drukuj wyniki
STOP
Wyznacz początkowy kierunek poszukiwań -g ⇒ ξ
Oblicz pochodną kierunkową funkcji F w p-cie p=(g;ξ)
F0 ⇒ Ya
P ⇒ Pa
g² ⇒ gba
Oblicz iloczyn skalarny (ξ;ξ) ⇒ ξ²
Wyznacz początkową długość kroku w nowym kierunku poszukiwań t.zn. 1/√ξ² ⇒ e
Wykonaj krok w danym kierunku ξ i oblicz współrzędne nowego p-tu x+eξ ⇒ x
Oblicz w nowym p-cie wartość funkcji F=f(x) oraz jej gradient g=g(x)
Oblicz pochodną kierunkową funkcji F w nowym p-cie P=(g;ξ)
Czy ZW=1
Nie
Tak
F ⇒ Yb
P ⇒ Pb
Czy minimum w kierunku ξ
Nie
Tak
0 ⇒ ZW
Czy aktualny przedział e obejmuje p-kt minimalizujący funkcję w kierunku ξ t.zn. Czy Yb>Ya i Pb>0
Nie
Tak
Czy P<0
Tak
Nie
Oblicz iloczyn skalarny (g;ξ) ⇒ g²
Drukuj wyniki
Yb ⇒ Ya
Pb ⇒ Pa
0 ⇒ t
F ⇒ Ya
P ⇒ Pa
d ⇒ e
e ⇒ t
F ⇒ Yb
P ⇒ Pb
h-d ⇒ e
0 ⇒ t
2e ⇒ e
Czy minimum? t.zn. Czy g² < E0
Nie
Tak
Oblicz odległość d minimalizującą funkcję F w kierunku ξ, stosując następujące wzory interpolacyjne: z=3(Ya-Yb)/e + Pa + Pb
w=√(z²-Pa·Pb)
d=e(Pb+w+z)/(Pb-Pa+2w)
STOP
e ⇒ h
1 ⇒ ZW
t-d ⇒ e
CONT+1 ⇒ CONT

Rys. 28

na wstępie wyliczamy w punkcie $\underline{x}_0$ wartość funkcji celu $F_0 = f(\underline{x}_0)$ oraz jej gradientu $\underline{g} = g(\underline{x}_0)$.

Rozpatrzmy działanie metody NS na przykładzie funkcji dwuzmiennych, której poziomice pokazano na rys. 29.

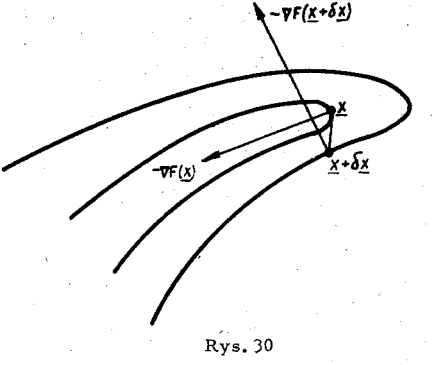


Rys. 29

Przebieg algorytmu jest w tym przypadku następujący. Na wstępie zakładamy punkt startowy $\underline{x}_0$ oraz znajdujemy w tym punkcie kierunek najstromszego zbocza (minus gradientu) $\underline{\xi}_0$. Wyznaczony w ten sposób kierunek poszukiwań $\underline{\xi}_0$ jest oczywiście normalny do poziomicy przechodzącej przez punkt $\underline{x}_0$. Następnie poszukujemy minimum funkcji celu wzdłuż tego kierunku. Minimum to występuje w punkcie $\underline{x}_1$. W drugim kroku ponawiamy te same czynności z tym jednak, że startujemy teraz z punktu $\underline{x}_1$, a nowy kierunek (minus gradientu) wynosi $\underline{\xi}_1$. Postępując w dalszym ciągu podobnie otrzymujemy szukane minimum.

Z rys. 29 wynika, że kierunek $\underline{\xi}^0$ jest "prawie" równoległy do $\underline{\xi}_2$, a kierunek $\underline{\xi}_1$ do $\underline{\xi}_3$. Widzimy więc, że w zasadzie operujemy tylko dwoma prostopadłymi wektorami o kierunku silnie zależnym od wyboru punktu początkowego. Podobny rezultat otrzymujemy również i w przypadku funkcji n wymiarowej. Jak wykazał bowiem Akaike [34] wyliczane w czasie przebiegu algorytmu kierunki poszukiwań dążą asymptotycznie tylko do dwóch kierunków tak, że w końcu minimum poszukujemy jedynie w dwuwymiarowej podprzestrzeni. Tego rodzaju zachowanie się metody NS stanowi jej podstawową wadę. Ponadto, szybkość zbieżności metody wyraźnie maleje z chwilą napotkania na wąską dolinę, w której ewentualnie znajduje się poszukiwane ekstremum. Powodem tego jest fakt, że w takich warunkach

występują znaczne trudności w dokładnym określeniu minimum w kierunku, a to z kolei pociąga za sobą powstawanie błędu przy wyznaczaniu właściwego kierunku minus gradientu. Przypadek ten został przedstawiony na rys. 30. Można więc, dopatrzyć się dużego podobieństwa omawianej metody do metody Gaussa-Seidela, która również charakteryzuje się tą niekorzystną właściwością. Podobieństwo to, jest szczególnie dobrze widoczne na rys. 29, gdzie wystarczy dokonać obrotu układu współrzędnych o kąt 90°, aby otrzymać przebieg algorytmu równoważny metodzie Gaussa-Seidela.



Rys. 30

W praktyce istnieje wiele modyfikacji metody NS. Jedną z nich stanowi następujący algorytm:

(1) wyznacz kierunek poszukiwań $\underline{\xi}$ jak poprzednio (punkt 1 i 2);
(2) określ minimum wzdłuż tego kierunku tzn. $F_{min} = f(\underline{x} + \hat{\lambda}\underline{\xi})$;
(3) cofnij się o $0,9\hat{\lambda}$ tzn. o 0,9 odległości pomiędzy punktem $\underline{x}$ a znalezionym punktem $\underline{x} + \hat{\lambda}\underline{\xi}$ w którym funkcja osiąga minimum w kierunku;
(4) jeśli nie "minimum", to powtórz procedurę od (1).

c. Kryterium zbieżności

Analogiczne jak w metodzie gradientu sprzężonego punkt 5.3.3c.

5.3.3. Metoda Gradientu Sprzężonego - GS

Metoda ta została opracowana przez Hestenesa i Stiefela (1952 r.) dla rozwiązywania układu równań liniowych. Do celów optymalizacji po raz pierwszy zastosowali ją Fletcher i Reeves [22]. Z rozważań nad metodami gradientu prostego oraz najszybszego spadku wynika, że sposób tworzenia kierunku poszukiwań, którym zawsze jest minus gradient, powoduje szereg niekorzystnych efektów takich jak: zmniejszenie wymiarowości bazy, trudności w określeniu właściwego kierunku gradientu itp. Metody te ponadto charakteryzują się tym, że każdą kolejną iteracją rozpoczyna się z zerową informacją o badanej powierzchni. Stąd, w celu przeciwdziałania tym wadliwym zjawiskom powstało szereg algo-

rytmów opartych o inną zasadę wyznaczania kierunków poszukiwań. Jednym z nich jest metoda gradientu sprzężonego, w której kierunki poszukiwań tworzone są tak, aby każdy następny był "sprzężony" do wszystkich poprzednio stosowanych kierunków. Definicję "sprzężenia" kierunków przytoczono w punkcie 5.2.6. Porównując tak sformułowany algorytm z omówioną poprzednio metodą Powella widzimy, że posiadają one dużo cech wspólnych oraz że twierdzenie 1 punkt 5.2.6 zachoduje w tym przypadku swoją ważność. Oznacza to, że metoda GS charakteryzuje się również zbieżnością drugiego rzędu.

Poza wspomnianym twierdzeniem 1 istotną rolę w metodzie gradientu sprzężonego odgrywa następujący lemat:

Lemat 1. Jeśli punkt $\underline{x}_{i+1}$ został osiągnięty w rezultacie i minimalizacji funkcji typu (196) wzdłuż $\underline{\xi}_1, \underline{\xi}_2, \ldots, \underline{\xi}_i$ kierunków, przy czym kierunki te są wzajemnie sprzężone względem macierzy A, to wówczas

$$\underline{\xi}_s^T \underline{g}_{i+1} = 0, \qquad s = 1, 2, \ldots, i, \tag{208}$$

gdzie

$$\underline{g}_{i+1} = \nabla f(\underline{x}_{i+1})$$

Dowód

Z założenia mamy

$$\begin{aligned} \underline{g}_{i+1} &= A\,\underline{x}_{i+1} + \underline{b} = \\ &= A\left(\underline{x}_{s+1} + \sum_{j=s+1}^{i} \lambda_j \underline{\xi}_j\right) + \underline{b} = \\ &= \underline{g}_{s+1} + \sum_{j=s+1}^{i} \lambda_j A\,\underline{\xi}_j \quad \text{dla} \quad s = 0, 1, 2, \ldots i-1 \end{aligned} \tag{209}$$

ale

$$\underline{\xi}_s^T \underline{g}_{i+1} = \underline{\xi}_s^T \underline{g}_{s+1} + \sum_{j=s+1}^{i} \lambda_j \underline{\xi}_s^T A\,\underline{\xi}_j = 0 \quad \text{c.n.d.} \tag{210}$$

gdyż

(i) $\underline{\xi}_s^T \underline{g}_{s+1} = 0$ z warunku minimalizacji funkcji wzdłuż kierunku $\underline{\xi}_s$,

(ii) $\sum_{j=s+1}^{i} \lambda_j \underline{\xi}_s^T A\,\underline{\xi}_j = 0$ z warunku sprzężenia kierunków.

Zwróćmy uwagę, że wykorzystując powyższy lemat można w łatwy sposób udowodnić twierdzenie 1, gdyż jak można zauważyć

po n iteracjach $\underline{g}_{n+1}$ będzie ortogonalny do n liniowo niezależnych wektorów, a tym samym musi się równać zero.

Na zakończenie tych rozważań, pozostało nam wyjaśnić zasadę tworzenia kierunków poszukiwań tak, aby spełniony był warunek ich wzajemnego sprzężenia.

Załóżmy, że pierwszy kierunek $\underline{\xi}_1$ jest kierunkiem minus gradientu. A więc

$$\underline{\xi}_1 = -\underline{g}_1 = -A\,\underline{x}_1 - \underline{b}, \tag{211}$$

dokonując minimalizacji $f(\underline{x})$ wzdłuż tego kierunku, z warunku koniecznego na istnienie ekstremum

$$\underline{\xi}_1^T \left[A\,(\underline{x}_1 + \lambda_1 \underline{\xi}_1) + \underline{b} \right] = 0, \tag{212}$$

znajdziemy λ_1, a następnie określimy punkt $\underline{x}_2$

$$\underline{x}_2 = \underline{x}_1 + \lambda_1 \underline{\xi}_1. \tag{213}$$

W drugim kroku przyjmujmy, że w punkcie $\underline{x}_2$ nowy kierunek wyznaczymy w myśl reguły

$$\underline{\xi}_2 = -\underline{g}_2 + \beta_2 \underline{\xi}_1, \tag{214}$$

z tym jednak, że współczynnik β_2 musi być tak dobrany, aby $\underline{\xi}_1$ i $\underline{\xi}_2$ były sprzężone, a mianowicie

$$\begin{aligned} 0 = \underline{\xi}_1^T A\,\underline{\xi}_2 &= -\underline{\xi}_1^T A\,\underline{g}_2 + \beta_2 \underline{\xi}_1^T A\,\underline{\xi}_1 = \\ &= -(\underline{x}_2 - \underline{x}_1)^T A(\underline{g}_2 - \beta_2 \underline{\xi}_1) = \\ &= -(\underline{g}_2 - \underline{g}_1)^T (\underline{g}_2 - \beta_2 \underline{\xi}_1), \end{aligned} \tag{215}$$

skąd

$$\beta_2 = \frac{\underline{g}_2^T \underline{g}_2}{\underline{g}_1^T \underline{g}_1}, \tag{216}$$

gdyż

$$\underline{g}_1^T \underline{g}_2 = 0 \quad \text{z założenia.} \tag{217}$$

Podobnie postępując można wykazać [34], że zasada (214) obowiązuje również i w i-tym kroku tzn.

$$\underline{\xi}_i = -\underline{g}_i + \beta_i \underline{\xi}_{i-1}, \tag{218}$$

przy czym

$$\beta_i = \frac{\underline{g}_i^T \underline{g}_i}{\underline{g}_{i-1}^T \underline{g}_{i-1}}. \tag{219}$$

Wystarczy w tym celu zauważyć:

1) $$\underline{g}_s^T \underline{g}_i = 0 \quad \text{dla} \quad s = 1, 2, \ldots, i-1, \tag{220}$$

gdyż

$$0 = \underline{\xi}_s^T \underline{g}_i = (-\underline{g}_s + \beta_s \underline{\xi}_{s-1})^T \underline{g}_i = -\underline{g}_s^T \underline{g}_i ; \tag{221}$$

2) $$-\underline{\xi}_s^T A \underline{\xi}_i = 0 \quad \text{dla} \quad s = 1, 2, \ldots, i-2, \tag{222}$$

gdyż

$$-\underline{\xi}_s^T A \underline{\xi}_i = \underline{\xi}_s^T A \underline{g}_i = \frac{1}{\lambda_s} (\underline{x}_{s+1} - \underline{x}_s)^T A \underline{g}_i = \frac{1}{\lambda_s} (\underline{g}_{s+1} - \underline{g}_s)^T \underline{g}_i = 0. \tag{223}$$

a. Informacje wejściowe

Analogicznie jak w metodzie najszybszego spadku punkt 5.3.2a.

b. Algorytm obliczeń

(1) oblicz w punkcie startowym $\underline{x}_o$ wartość funkcji celu $F_o = f(\underline{x}_o)$ oraz jej gradientu $\underline{g}_o = \underline{g}(\underline{x}_o)$,

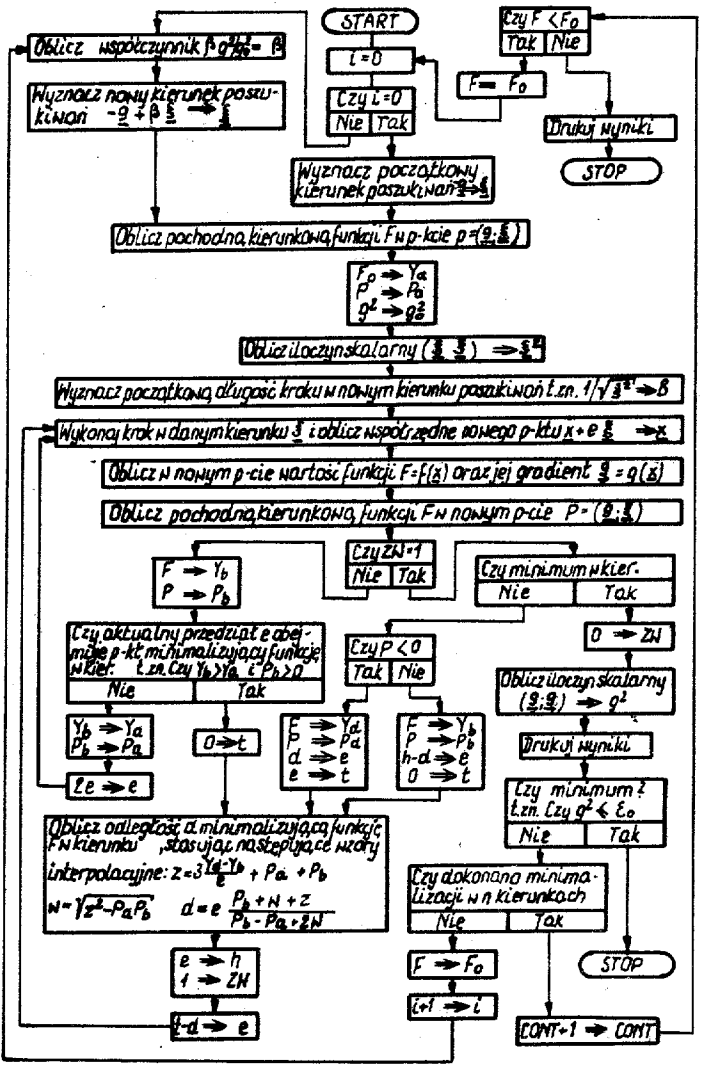
(2) podstaw $i = 1$ oraz wyznacz początkowy kierunek poszukiwań

$$\underline{\xi}_{i-1} = -\underline{g}.$$

(3) wzdłuż kierunku $\underline{\xi}_{i-1}$ określ λ_i minimalizujące $f(\underline{x}_{i-1} + \lambda_i \underline{\xi}_{i-1})$ oraz współrzędne nowego punktu $\underline{x}_i = \underline{x}_{i-1} + \lambda_i \underline{\xi}_{i-1}$;

(4) oblicz w punkcie $\underline{x}_i$ wartość gradientu $\underline{g}_i = \underline{g}(\underline{x}_i)$;

(5) zbadaj czy zostało spełnione kryterium na minimum. Jeśli tak to stop, natomiast jeśli nie, to wyznacz współczynnik β_i


START
Oblicz współczynnik β
Wyznacz nowy kierunek poszukiwań
Czy i=0
Nie
Tak
Wyznacz początkowy kierunek poszukiwań
Oblicz pochodną kierunkową funkcji F w p-kcie p
Oblicz iloczyn skalarny
Wyznacz początkową długość kroku w nowym kierunku poszukiwań t.zn.
Wykonaj krok w danym kierunku i oblicz współrzędne nowego p-ktu
Oblicz w nowym p-cie wartość funkcji
oraz jej gradient
Oblicz pochodną kierunkową funkcji F w nowym p-cie
Czy ZN=1
Czy minimum w kier.
Czy aktualny przedział zawiera minimalizujący funkcję w kier. t.zn. Czy
Czy P<0
Czy minimum ?
t.zn. Czy
Oblicz odległość d minimalizującą funkcję F w kierunku, stosując następujące wzory
interpolacyjne:
Czy dokonano minimalizacji w n kierunkach
Drukuj wyniki
STOP
i=i+1
CONT=1 → CONT

Rys. 31

$$\beta_i = \frac{\underline{g}_i^T \underline{g}_i}{\underline{g}_{i-1}^T \underline{g}_{i-1}}$$

oraz nowy sprzężony kierunek poszukiwań

$$\underline{\xi}_i = -\underline{g}_i + \beta_i \underline{\xi}_{i-1} ;$$

(6) podstaw $\underline{x}_i$ w miejsce $\underline{x}_{i-1}$ oraz zbadaj czy wykonano n iteracji tzn. czy i = n. Jeśli nie, to podstaw $\underline{\xi}_i$ w miejsce $\underline{\xi}_{i-1}$ oraz powtórz czynności od kroku (3) zwiększając i o 1. W przeciwnym razie podstaw $\underline{g}_i$ w miejsce $\underline{g}_o$ i powtórz procedurę od kroku (2).

Sieć działań przytoczonego algorytmu przedstawiono na rys. 31, przy czym przed przystąpieniem do jego wykonywania, na wstępie wyliczamy w punkcie $\underline{x}_o$ wartość funkcji celu $F_o = f(\underline{x}_o)$ oraz jej gradientu $\underline{g} = g(\underline{x}_o)$.

c. Kryterium zbieżności

Przy przyjętych powyżej założeniach, warunkiem koniecznym a zarazem dostatecznym na to, aby aktualnie wyliczony punkt $\underline{x}_T$ był szukanym ekstremum $\hat{\underline{x}}$ jest spełnienie kryterium $(\underline{g}^T \underline{g}) = 0$. Jednakże, ze względu na błędy zaokrągleń, tego rodzaju kryterium jest trudne do zrealizowania w maszynie cyfrowej i stąd często poprzestaje się na żądaniu, aby iloczyn skalarny $(\underline{g}^T \underline{g})$ był mniejszy od z góry założonej liczby ε_o. Oprócz powyższego kryterium można także stosować i inne, a mianowicie - jeśli w kolejnych n + 1 iteracjach przesunięcie punktu wzdłuż kierunków poszukiwań będzie mniejsze niż z góry zadana liczba ε, to wówczas należy zakończyć wykonywanie procedury.

5.3.4. Metoda Davidona - D

Metoda D została opracowana w 1959 r. przez Davidona [11], a następnie w 1963 r. ulepszona i zmodyfikowana przez Fletchera i Powella [21]. W metodzie tej posłużono się również koncepcją kierunków sprzężonych z tą jednak różnicą w porównaniu do metody GS, że kierunki te tworzone są w odmienny sposób. Zauważono bowiem, że jeśli funkcja celu jest o postaci (196) tzn.

$$f(\underline{x}) = a + \underline{b}^T \underline{x} + \frac{1}{2} \underline{x}^T A \underline{x}, \qquad (224)$$

a jej gradient wynosi

$$\underline{g}(x) = \underline{b} + A \underline{x}, \qquad (225)$$

to wówczas wystarczy znać tylko macierz odwrotną drugich pochodnych A^{-1}, aby w jednym kroku móc wyznaczyć poszukiwane ekstremum. Stwierdzenie to wynika z następującego prostego rozumowania:

jeśli $\hat{\underline{x}}$ jest minimum, to z równania (225) mamy

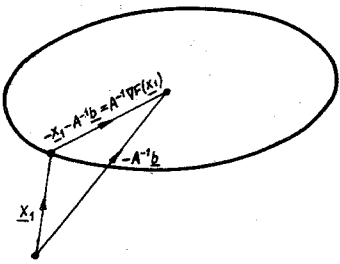
$$0 = \underline{b} + A\,\hat{\underline{x}}, \tag{226}$$

gdyż znikanie gradientu jest koniecznym, a w naszym przypadku - funkcji wypukłej, także i dostatecznym warunkiem istnienia ekstremum.

Odejmując stronami (225) od (226) ostatecznie otrzymujemy

$$\hat{\underline{x}} = \underline{x} - A^{-1}g(\underline{x}) = \underline{x} - A^{-1}\,\nabla f(\underline{x}) \tag{227}$$

Ilustracja graficzna powyższego stwierdzenia została przedstawiona na rys. 32.



Rys. 32

Koncepcja metody Davidona polega więc na takim sposobie tworzenia sprzężonych kierunków poszukiwań, aby począwszy od kierunku

$$\underline{\xi}_{i+1} = -\,H_i\,\nabla f(\underline{x}_{i+1}), \tag{228}$$

dowolnie wybrana macierz H_i w każdej następnej iteracji coraz to lepiej aproksymowała macierz A^{-1}. Z chwilą gdy macierz H_i zostaje przekształcona w macierz odwrotną drugich pochodnych A^{-1} procedura kończy swoje działanie i bieżący punkt $\underline{x}$ można wtedy uznać za poszukiwany punkt ekstremalny. Zwróćmy uwagę, że tego rodzaju algorytm charakteryzuje się zbieżnością drugiego rzędu, gdyż dzięki zastosowaniu poszukiwań jedynie wzdłuż kierunków sprzężonych przy przyjętych założeniach pozostają w mocy twierdzenie 1 punkt 5.2.6 oraz lemat 1 punkt 5.3.3.

W celu dokładniejszego wyjaśnienia zasady działania procedury Davidona, pozostało nam jeszcze odpowiedzieć na pytanie w jaki sposób należy dokonywać modyfikacji macierzy H_i, aby metoda D była zbieżna.

Utwórzmy z n kolejnych kierunków poszukiwań $\underline{\xi}_i$ macierz S o postaci

$$S = \left[\underline{\xi}_1, \underline{\xi}_2, \ldots, \underline{\xi}_n\right], \tag{229}$$

która z założenia będzie macierzą modalną, bowiem wszystkie kierunki $\underline{\xi}_i$ są niezależne liniowo. Skorzystamy teraz ze znanego w rachunku macierzowym faktu, że dowolną macierz kwadratową o pojedynczych wartościach własnych można poprzez tzw. przekształcenie podobieństwa sprowadzić do postaci diagonalnej.

A więc w naszym przypadku mamy

$$S^{-1} A S = \Lambda \tag{230}$$

gdzie Λ jest macierzą diagonalną o elementach $\underline{\xi}_i^T A \underline{\xi}_i$.

Ponieważ z założenia kierunki poszukiwań są wzajemnie sprzężone względem macierzy A, to związek (230) można wyrazić przez

$$S^T A S = \Lambda \tag{231}$$

skąd

$$A = (S^T)^{-1} \Lambda (S)^{-1} =$$

$$= (S \Lambda^{-1} S^T)^{-1} \tag{232}$$

oraz

$$A^{-1} = S \Lambda^{-1} S^T. \tag{233}$$

Korzystając z własności macierzy diagonalnej Λ związek (233) przekształcimy do trochę innej formy, a mianowicie

$$A^{-1} = \sum_{i=1}^{n} \left(\Lambda^{-1}\right)_{ii} \underline{\xi}_i \underline{\xi}_i^T, \tag{234}$$

bądź też

$$A^{-1} = \sum_{i=1}^{n} \frac{\underline{\xi}_i \underline{\xi}_i^T}{\underline{\xi}_i^T A \underline{\xi}_i}. \tag{235}$$

Otrzymany rezultat posłużył Davidonowi do skonstruowania algorytmu modyfikacji macierzy H_i, który polega na dodawaniu w każdej kolejnej iteracji do aktualnie obowiązującej macierzy H_i czynnika powodującego dążenie macierzy H_i do A^{-1} oraz pewnej korekcyjnej macierzy B_i. Tak więc, algorytm ten przedstawia się następująco:

$$\underline{\xi}_i = -H_{i-1} \underline{g}_i,$$

$$\underline{x}_{i+1} = \underline{x}_i + \lambda_i \underline{\xi}_i ,$$

$$H_i = H_{i-1} + \frac{\underline{\xi}_i \underline{\xi}_i^T}{\underline{\xi}_i^T A \underline{\xi}_i} + B_i , \tag{236}$$

przy czym macierz B_i musi być tak dobrana, aby zależność

$$\underline{\xi}_s^T A H_{i-1} = \underline{\xi}_s^T, \quad s = 1,2,\ldots, i-1, \tag{237}$$

gdzie $\underline{\xi}_1, \underline{\xi}_2, \ldots \underline{\xi}_{i-1}$ są kierunkami wzajemnie sprzężonymi, pozostawała również w mocy i dla kierunków $\underline{\xi}_1, \underline{\xi}_2, \ldots, \underline{\xi}_i$, tzn.

$$\underline{\xi}_s^T A H_i = \underline{\xi}_s^T, \quad s = 1,2,\ldots,i, \tag{238}$$

przy jednoczesnym spełnieniu związku

$$-\underline{\xi}_s^T A \underline{\xi}_i = \underline{\xi}_s^T A H_{i-1} \underline{g}_i = \underline{\xi}_s^T \underline{g}_i = 0 \tag{239}$$

wynikającym bezpośrednio z lematu 1 punkt 5.3.3.

Dla $s = i$ równanie (238) przyjmie więc postać

$$\underline{\xi}_i^T A \left(H_{i-1} + \frac{\underline{\xi}_i \underline{\xi}_i^T}{\underline{\xi}_i^T A \underline{\xi}_i} + B_i \right) = \underline{\xi}_i^T , \tag{240}$$

skąd

$$(\underline{g}_{i+1} - \underline{g}_i)^T (H_{i-1} + B_i) = 0, \tag{241}$$

co stanowi warunek, aby macierz B_i była dobrze dobrana.

Oznaczając przez

$$\underline{\gamma}_i = \underline{g}_{i+1} - \underline{g}_i, \tag{242}$$

z równania (241) otrzymujemy

$$B_i = - \frac{H_{i-1} \underline{\gamma}_i \underline{\gamma}_i^T H_{i-1}}{\underline{\gamma}_i^T H_{i-1} \underline{\gamma}_i} , \tag{243}$$

przy czym

$$\underline{\xi}_s^T A B_i = 0, \quad s = 1,2,\ldots,i-1, \tag{244}$$

z lematu 1 punkt 5.3.3.

W celu praktycznego zastosowania w procedurze Davidona wzoru (236) należy dokonać w nim jeszcze niewielkich modyfikacji, a mianowicie wyrażenie $\underline{\xi}_i^T A \underline{\xi}_i$ należy przedstawić w innej wygodniejszej postaci.

A więc

$$\underline{\xi}_i^T A \underline{\xi}_i = \frac{(\underline{x}_{i+1} - \underline{x}_i)^T}{\lambda_i} A \underline{\xi}_i =$$

$$= \frac{(\underline{g}_{i+1} - \underline{g}_i)^T}{\lambda_i} \underline{\xi}_i = \qquad (245)$$

$$= \frac{1}{\lambda_i} \underline{g}_i^T H_{i-1} \underline{g}_i .$$

Podstawiając teraz (243) oraz (245) do (236) ostatecznie mamy

$$H_i = H_{i-1} + \lambda_i \frac{\underline{\xi}_i \underline{\xi}_i^T}{\underline{g}_i^T H_{i-1} \underline{g}_i} - \frac{H_{i-1} \underline{\gamma}_i^T H_{i-1}}{\underline{\gamma}_i^T H_{i-1} \underline{\gamma}_i}, \qquad (246)$$

bądź też w równoważnej formie

$$H_i = H_{i-1} + \frac{\underline{\alpha}_i \underline{\alpha}_i^T}{\underline{\alpha}_i^T \underline{\gamma}_i} - \frac{H_{i-1} \underline{\gamma}_i \underline{\gamma}_i^T H_{i-1}}{\underline{\gamma}_i^T H_{i-1} \underline{\gamma}_i} , \qquad (247)$$

jeśli oznaczymy przesunięcie punktu przez

$$\underline{\alpha}_i = \underline{x}_{i+1} - \underline{x}_i . \qquad (248)$$

a. Informacje wejściowe

$\underline{x}_o$ - arbitralnie wybrany punkt startowy,
H_o - macierz wyjściowych kierunków poszukiwań przyjmowana zazwyczaj jako macierz jednostkowa,
e - początkowa długość kroku,
ε_j - wymagana dokładność obliczeń minimum w aktualnie występującym kierunku poszukiwań,
ε_o - wymagana dokładność obliczeń minimum globalnego,
n - liczba zmiennych niezależnych.

b. Algorytm obliczeń

(1) oblicz w punkcie startowym $\underline{x}_o$ wartość funkcji celu $F_o = f(\underline{x}_o)$ i jej gradientu $g_o = g(\underline{x}_o)$ oraz postaw $i = 1$;

(2) wyznacz kierunek poszukiwań

$$\underline{\xi}_{i-1} = -H_{i-1}\,\underline{g}_{i-1};$$

(3) wzdłuż kierunku $\underline{\xi}_{i-1}$ określ λ_i minimalizujące $f(\underline{x}_{i-1} + \lambda_i \underline{\xi}_{i-1})$, współrzędne nowego punktu $\underline{x}_i = \underline{x}_{i-1} + \lambda_i \underline{\xi}_{i-1}$ oraz podstaw $\lambda_i \underline{\xi}_{i-1}$ w miejsce $\underline{\alpha}_i$;

(4) oblicz w punkcie $\underline{x}_i$ wartość gradientu $\underline{g}_i = g(\underline{x}_i)$;

(5) zbadaj czy zostało spełnione kryterium na minimum. Jeśli tak to stop, w przeciwnym razie określ następujące wielkości

$$\underline{\gamma}_i = \underline{g}_i - \underline{g}_{i-1},$$

$$M_1 = \frac{\underline{\alpha}_i \underline{\alpha}_i^T}{\underline{\alpha}_i^T \underline{\gamma}_i},$$

$$M_2 = \frac{H_{i-1}\,\underline{\gamma}_i\,\underline{\gamma}_i^T\,H_{i-1}}{\underline{\gamma}_i^T\,H_{i-1}\,\underline{\gamma}_i};$$

(6) dokonaj modyfikacji macierzy kierunków

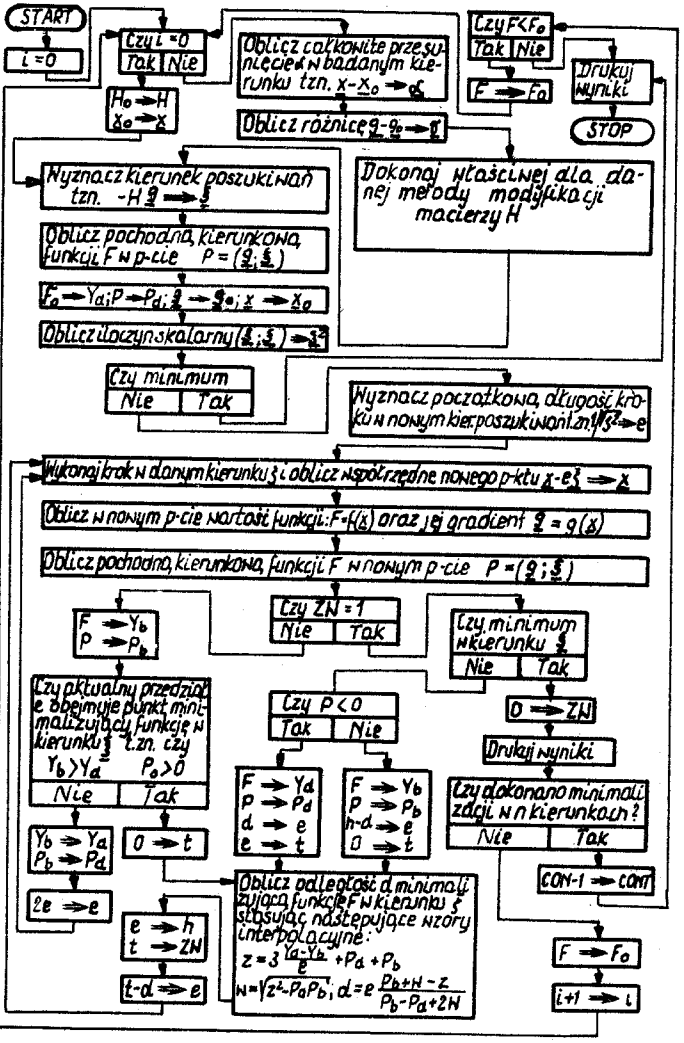
$$H_i = H_{i-1} + M_1 + M_2;$$

(7) podstaw $\underline{x}_i$ w miejsce $\underline{x}_{i-1}$ oraz zbadaj czy wykonano n iteracji tzn. czy i = n. Jeśli nie, to podstaw H_i w miejsce H_{i-1}, $\underline{g}_i$ w miejsce $\underline{g}_{i-1}$ oraz powtórz czynności od kroku (2) zwiększając i o 1. W przeciwnym przypadku podstaw H_o w miejsce H_{i-1} i powtórz procedurę od kroku (2) ustawiając i = 1.

Sieć działań przytoczonego algorytmu przedstawiono na na rys. 33, przy czym przed przystąpieniem do jego wykonywania, na wstępie wyliczamy w punkcie $\underline{x}_o$ wartość funkcji celu $F_o = f(\underline{x}_o)$ oraz jej gradientu $\underline{g} = g(\underline{x}_o)$.

c. Kryterium zbieżności

Jak już zostało wspomniane, bieżący punkt $\underline{x}$ można przyjąć za poszukiwane ekstremum, z chwilą gdy macierz H zostaje przekształcona w macierz A^{-1}. Jednakże, tak sformułowane kryterium ma znaczenie tylko teoretyczne, gdyż nie jest w rzeczywistości realizowane. Stąd, dla celów praktycznych Davidon zaproponował, aby bieżący punkt $\underline{x}$ uznać za dostatecznie dobrą estymatę $\underline{\hat{x}}$, jeśli bezwzględna odległość od minimum



Modyfikacje macierzy H

D: $H = H + \frac{(\alpha\,\alpha^T)}{(\alpha\,\gamma)} - \frac{(H\gamma)(H\gamma)^T}{\gamma^T H \gamma}$

P1: $H = H - \frac{(H\gamma)(H\gamma)^T}{\gamma^T H \gamma}$ P2: $H = H + \frac{(\alpha - H\gamma)\alpha^T}{\gamma^T \alpha}$

P3: $H = H + \frac{(\alpha - H\gamma)(H^T\gamma)^T}{\gamma^T H \gamma}$

Rys. 33

tzn. $\left(\underline{\xi}^T \underline{\xi}\right)^{\frac{1}{2}}$ będzie mniejsza od z góry założonej dowolnie małej liczby ε.

5.3.5. Metody Pearsona - PE

Metody te są bardzo podobne do omówionej poprzednio metody Davidona (punkt 5.3.4), a więc: informacje wejściowe, przebieg algorytmów obliczeń oraz kryteria zbieżności są w metodach tych takie same. Jedyną różnicą jaka występuje pomiędzy nimi jest odmienny sposób modyfikacji macierzy H, która w poszczególnych metodach Pearsona ma postać następującą:

- metoda I - PE 1 (zwana "projected gradient")

$$H_i = H_{i-1} - \frac{(H_{i-1}\underline{\gamma}_i)(H_{i-1}\underline{\gamma}_i)^T}{\underline{\gamma}_i^T H_{i-1}\underline{\gamma}_i} ; \qquad (249)$$

- metoda II - PE 2

$$H_i = H_{i-1} + \frac{(\underline{\alpha}_i - H_i\underline{\gamma}_i)\underline{\alpha}_i^T}{\underline{\gamma}_i^T \underline{\alpha}_i} ; \qquad (250)$$

- metoda III - PE 3

$$H_i = H_{i-1} + \frac{(\underline{\alpha}_i - H_{i-1}\underline{\gamma}_i)(H_{i-1}^T\underline{\gamma}_i)^T}{\underline{\gamma}_i^T H_{i-1}\underline{\gamma}_i} . \qquad (251)$$

Sieć działań metod Pearsona przedstawiono na rys. 33, przy czym przed przystąpieniem do wykonywania odpowiedniej procedury, na wstępie wyliczamy w punkcie $\underline{x}_o$ wartość funkcji celu $F_o = f(\underline{x}_o)$ oraz jej gradientu $\underline{g} = g(\underline{x}_o)$.

5.3.6. Metoda Newtona-Raphsona - NR

W odróżnieniu od metod gradientowych rozpatrzonych do tej pory, w metodzie NR poza znajomością gradientu funkcji celu wymagana jest również znajomość macierzy drugich pochodnych A, której elementy określone są przez

$$A_{jk} = \frac{\partial^2 f}{\partial x_j \partial x_k} . \qquad (252)$$

Istota metody NR polega więc na realizacji kolejnej iteracji w analogiczny sposób jak we wszystkich metodach gradientowych z tym jednak, że kierunek poszukiwań wyznaczany jest w myśl zasady

$$\underline{\xi}_i = -A_i^{-1}\, \underline{g}_i \,, \tag{253}$$

która bezpośrednio wynika z równania (227).

a. Informacje wejściowe

Analogiczne jak w metodzie Davidona punkt 5.3.4.

b. Algorytm obliczeń

(1) oblicz w punkcie startowym $\underline{x}_o$ wartość funkcji celu $F_o = f(\underline{x}_o)$ i jej gradientu $\underline{g}_o = g(\underline{x}_o)$ oraz podstaw $i = 1$;

(2) wyznacz kierunek poszukiwań

$$\underline{\xi}_{i-1} = - H_{i-1}\, \underline{g}_{i-1} \,;$$

(3) wzdłuż kierunku $\underline{\xi}_{i-1}$, określ λ_i minimalizujące $f(\underline{x}_{i-1} + \lambda_i \underline{\xi}_{i-1})$ oraz współrzędne nowego punktu $\underline{x}_i = \underline{x}_{i-1} + \lambda_i \underline{\xi}_{i-1}$;

(4) oblicz w punkcie $\underline{x}_i$ wartość gradientu $\underline{g}_i = g(\underline{x}_i)$;

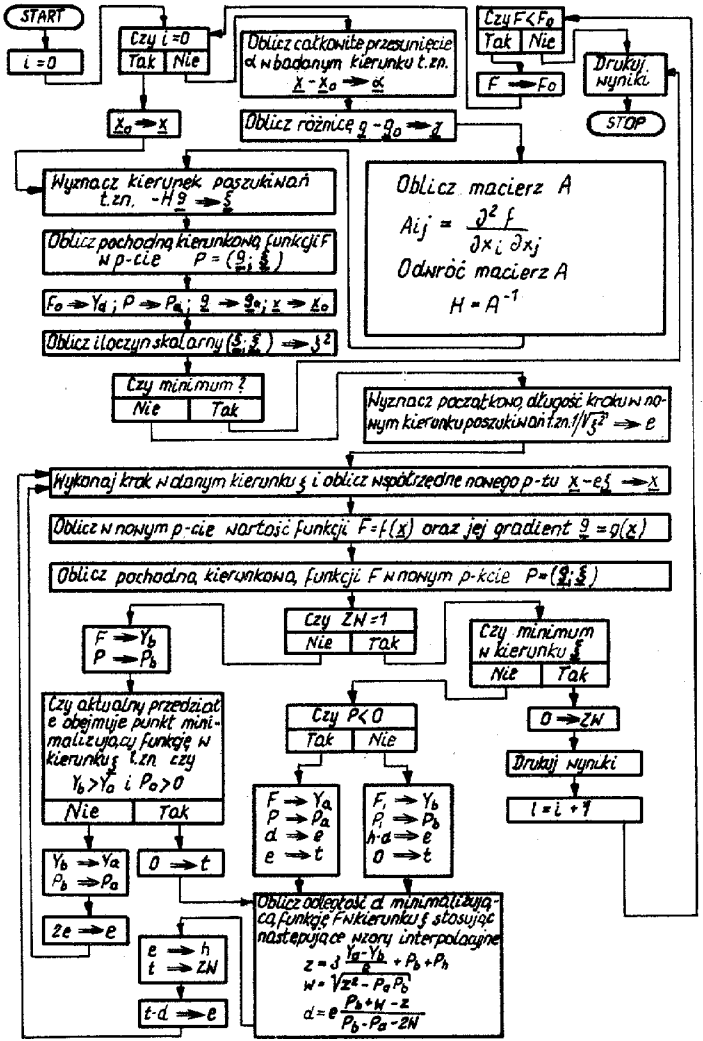
(5) zbadaj czy zostało spełnione kryterium na minimum. Jeśli tak to stop, w przeciwnym razie oblicz odwróconą macierz $A_i^{-1}(\underline{x}_i)$,

(6) podstaw $\underline{x}_i$ w miejsce $\underline{x}_{i-1}$ oraz zbadaj czy wykonano n iteracji tzn. czy $i = n$. Jeśli nie, to podstaw A_i^{-1} w miejsce H_{i-1}, $\underline{g}_i$ w miejsce $\underline{g}_{i-1}$ oraz powtórz czynności od kroku (2) zwiększając i o 1. Natomiast w przeciwnym przypadku podstaw H_o w miejsce H_{i-1} i powtórz procedurę od kroku (2) ustawiając $i = 1$.

Sieć działań przytoczonego algorytmu przedstawiono na rys. 34, przy czym przed przystąpieniem do jego wykonywania, na wstępie wyliczamy w punkcie $\underline{x}_o$ wartość funkcji celu $F_o = f(\underline{x}_o)$ oraz jej gradientu $\underline{g} = g(\underline{x}_o)$.

c. Kryterium zbieżności

Analogiczne jak w metodzie Davidona punkt 5.3.4.


START
i = 0
Czy i = 0
Tak
Nie
Oblicz całkowite przesunięcie α w badanym kierunku t.zn.
x - x0 = αξ
Czy F < F0
Tak
Nie
F → F0
Drukuj wyniki
STOP
x0 → x
Oblicz różnicę g - g0 → γ
Wyznacz kierunek poszukiwań t.zn. -H g → ξ
Oblicz macierz A
Aij = ∂²f / ∂xi ∂xj
Odwróć macierz A
H = A⁻¹
Oblicz pochodną kierunkową funkcji F w p-cie P = (g; ξ)
F0 → Y0 ; P → P0 ; g → g0 ; x → x0
Oblicz iloczyn skalarny (ξ, ξ) → ξ²
Czy minimum ?
Nie
Tak
Wyznacz początkową długość kroku w nowym kierunku poszukiwań t.zn. 1/√ξ² → e
Wykonaj krok w danym kierunku ξ i oblicz współrzędne nowego p-tu x - eξ → x
Oblicz w nowym p-cie wartość funkcji F = f(x) oraz jej gradient g = g(x)
Oblicz pochodną kierunkową funkcji F w nowym p-kcie P = (g; ξ)
Czy ZW = 1
Nie
Tak
Czy minimum w kierunku ξ
Nie
Tak
F → Yb
P → Pb
Czy aktualny przedział e obejmuje punkt minimalizującej funkcję w kierunku ξ t.zn. czy Yb > Y0 i P0 > 0
Nie
Tak
Czy P < 0
Tak
Nie
0 → ZW
Drukuj wyniki
l = i + 1
F → Ya
P → Pa
d → e
e → t
Fi → Yb
Pi → Pb
h·d → e
0 → t
Yb → Ya
Pb → Pa
2e → e
0 → t
e → h
t → ZW
t - d → e
Oblicz odległość d minimalizującą funkcję F w kierunku ξ stosując następujące wzory interpolacyjne
z = 3 (Ya - Yb)/e + Pb + Pa
w = √(z² - Pa Pb)
d = e (Pb + w - z) / (Pb - Pa - 2w)

Rys. 34

5.4. Porównanie metod

5.4.1. Wybór metod oraz kryteriów porównawczych

Wśród metod Poszukiwania Ekstremum Bez Ograniczeń PEBOG omówionych w poprzednich punktach, za główne i reprezentacyjne ze względu na zasadę działania można uznać metody: Rosenbrocka (punkt 5.2.2), Simplexu Neldera i Meada (punkt 5.2.3), Daviesa, Swanna i Campeya (punkt 5.2.5), metodę Powella (punkt 5.2.6), metodę Gradientu Sprzężonego (punkt 5.3.3) oraz metodę Davidona (punkt 5.3.4). Dla dokonania porównania rozważanych metod posłużono się więc wymienionymi sześcioma metodami, przy czym przyjęto oznaczenia wprowadzone w punkcie 5. Oceną i porównaniem pełnego zbioru metod, przedstawionych w punktach 5.2 i 5.3, zajmowano się w pracach [53], [55] oraz [56] z których wynika, że zdefiniowany ich podzbiór jest jak najbardziej reprezentatywny.

Przy porównywaniu metod optymalizacji można spotkać bardzo różnorodne kryteria oceny omawianych metod, a więc takie jak: szybkość zbieżności, wrażliwość na zakłócenia, zajętość pamięci maszyny cyfrowej przez daną procedurę itp. Wydaje się jednak, że kryterium zaproponowane przez Boxa [4] najlepiej nadaje się do ogólnej oceny metod optymalizacji i dlatego posłużono się nim w niniejszej pracy. Jako kryterium przyjęto więc za Boxem - liczbę obliczeń wartości funkcji celu, którą należy wykonać w trakcie działania procedury, aby uzyskać z góry założoną wartość szukanego minimum. Oczywiście, że w przypadku badania metod gradientowych należy do liczby obliczeń wartości samej funkcji dodać ponadto liczbę obliczeń wartości jej gradientu. Jak wykazało bowiem doświadczenie najbardziej pracochłonną czynnością w czasie wykonywania się procedury jest samo liczenie wartości badanej funkcji bądź jej gradientu. Natomiast czas przeznaczony na realizację tego czy innego algorytmu poszukiwania minimum jest tylko niewielkim ułamkiem czasu trwania tamtych czynności. W tej sytuacji o szybkości zbieżności danej procedury jak i czasie jej działania będzie można wnioskować bezpośrednio z uzyskanych danych.

5.4.2. Wybór przykładów oraz wyniki obliczeń

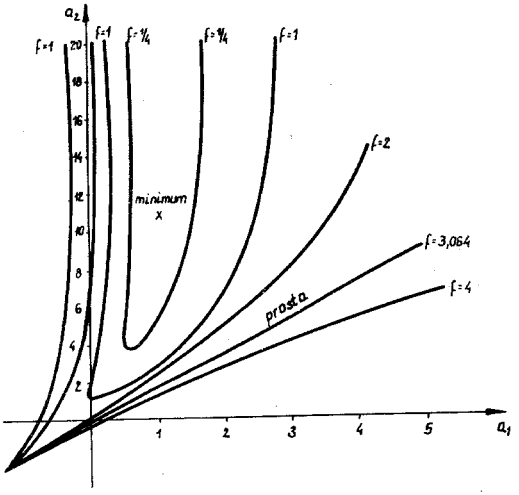
Jako funkcje wzorcowe stanowiące zadania optymalizacji przyjęto funkcje zaliczone przez Boxa [4] oraz Powella [41] do trudnych testów metod poszukiwania esktremum bez ograniczeń. Postać tych funkcji jest następująca:

1. Problem dwuwymiarowy

Znaleźć minimum funkcji

$$f(a_1, a_2) = \sum_x \left[(e^{-a_1 x} - e^{-a_2 x}) - (e^{-x} - e^{-10x}) \right]^2, \qquad (254)$$

przy czym sumowania po x należy dokonać poczynając od wartości $0,1$ co $0,1$ do 1. Wykres badanej funkcji w formie konturów $f = \text{const}$ przedstawiono na rys. 35.



Rys. 35

Jak nietrudno zauważyć szukane minimum funkcji o wartości $f = 0$ występuje przy $a_1 = 1$ oraz $a_2 = 10$. Obliczenia dokonano dla następujących punktów startowych

I	$a_1 = 0$;	$a_2 = 0$;	$f = 3,064$
II	$a_1 = 0$;	$a_2 = 20$;	$f = 2,087$
III	$a_1 = 5$;	$a_2 = 0$;	$f = 19,588$
IV	$a_1 = 5$;	$a_2 = 20$;	$f = 1,808$
V	$a_1 = 2,5$;	$a_2 = 10$;	$f = 0,808$

Porównanie wyników obliczeń dla problemu dwuwymiarowego

Metoda	Wartość funkcji	Punkt startowy				
		I	II	III	IV	V
R	1,0	15	4	13	12	0
	0,1	39	7	26	25	20
	0,01	52	45	73	77	89
	0,00001	96	68	103	103	109
DSC	1,0	23	3	15	7	0
	0,1	45	8	119	23	6
	0,01	55	22	203	28	6
	0,00001	78	57	231	52	6
N	1,0	6	2	5	5	0
	0,1	6	8	5	17	12
	0,01	19	20	19	26	13
	0,00001	41	43	39	49	40
P	1,0	33	2	26	15	0
	0,1	52	8	57	16	6
	0,01	56	30	69	37	7
	0,00001	64	51	84	77	23
GS	1,0	12	6	21	12	0
	0,1	45	12	75	12	12
	0,01	69	126	87	60	12
	0,00001	93	144	102	84	21
D	1,0	12	6	12	15	0
	0,1	33	9	18	18	6
	0,01	45	42	30	54	9
	0,00001	51	48	45	63	27

Wyniki minimalizacji przedstawiono w tablicy 5 w postaci równoważnej liczby obliczeń funkcji, którą należało wykonać dla zmniejszenia wartości funkcji do 1; 0,1; 0,01 oraz 0,00001.

W powyższej tablicy przez symbol P rozumiany jest pierwszy wariant metody Powella.

2. Problem trójwymiarowy

Znaleźć minimum funkcji

$$f(a_1, a_2, a_3) = \sum_x \left[(e^{-a_1 x} - e^{-a_2 x}) - a_3 (e^{-x} - e^{-10x}) \right]^2, \qquad (255)$$

przy czym sumowania po x należy dokonać poczynając od wartości 0,1 co 0,1 do 1.

Żądane minimum o wartości f = 0 występuje przy $a_1 = 1$, $a_2 = 10$ oraz $a_3 = 1$. Zwróćmy uwagę, że w rozpatrywanym przez nas przypadku istnieje możliwość uzyskania nieprawidłowej odpowiedzi, gdyż badana funkcja posiada ponadto ciągłe minimum f = 0 przy $a_3 = 0$ oraz $a_1 = a_2$. Niebezpieczeństwo to możemy wyeliminować jedynie przez alternatywny dobór punktów startowych oraz początkowej długości kroku.

Obliczenia dokonano dla następujących punktów startowych:

I	$a_1 = 0$;	$a_2 = 20$;	$a_3 = 1$;	f = 2,087
II	$a_1 = 2,5$;	$a_2 = 10$;	$a_3 = 10$;	f = 275,881
III	$a_1 = 0$;	$a_2 = 0$;	$a_3 = 10$;	f = 306,401
IV	$a_1 = 0$;	$a_2 = 10$;	$a_3 = 1$;	f = 1,885
V	$a_1 = 0$;	$a_2 = 10$;	$a_3 = 10$;	f = 213,673
VI	$a_1 = 0$;	$a_2 = 10$;	$a_3 = 20$;	f = 1 031,154
VII	$a_1 = 0$;	$a_2 = 20$;	$a_3 = 0$;	f = 9,706
VIII	$a_1 = 0$;	$a_2 = 20$;	$a_3 = 10$;	f = 209,280
IX	$a_1 = 0$;	$a_2 = 20$;	$a_3 = 20$;	f = 1 021,655

Wyniki minimalizacji przedstawiono w tablicy 6 w postaci równoważnej liczby obliczeń funkcji, którą należało wykonać dla zmniejszenia wartości funkcji do 1; 0,1; 0,01 oraz 0,0001.

W omawianej tablicy symbol F wskazuje, że metody DSC oraz P zawodzą przy tak wybranych punktach startowych. Jedynie co możemy uzyskać to następujące wartości:

$a_1 \cong 0,61$; $a_2 \longrightarrow \infty$; $a_3 = 1,32$ oraz $f \cong 0,076$.

Ponadto, w tablicy tej nie uwzględniono dodatkowych oddziaływań jakie należało wykonać dla otrzymania żądanego rozwiązania.

Tablica 6

Porównanie wyników obliczeń dla problemu trójwymiarowego

Metoda	Wartość funkcji	Punkt startowy								
		I	II	III	IV	V	VI	VII	VIII	IX
R	1,0	5	30	31	5	30	43	10	24	41
	0,1	7	57	31	18	30	50	25	55	61
	0,01	150	197	139	38	174	165	224	190	179
	0,00001	347	281	200	198	292	350	273	460	246
DSC	1,0	3	F	F	3	F	F	1	F	F
	0,1	8			4			6		
	0,01	20			6			8		
	0,00001	448			313			25		
N	1,0	2	11	6	2	18	27	11	18	27
	0,1	13	11	29	20	26	92	17	28	109
	0,01	48	42	46	28	70	198	47	127	210
	0,00001	119	128	112	73	110	307	79	164	315
P	1,0	2	F	F	2	F	F	3	F	F
	0,1	8			7			9		
	0,01	33			10			13		
	0,00001	78			56			19		
GS	1,0	8	28	44	8	36	40	8	36	40
	0,1	8	28	44	16	44	48	16	56	48
	0,01	116	40	48	28	68	48	116	208	124
	0,00001	564	112	104	56	92	92	344	608	188
D	1,0	8	8	44	8	52	52	8	52	52
	0,1	16	12	52	12	60	60	12	60	60
	0,01	72	16	128	20	128	112	76	120	116
	0,00001	92	68	144	24	148	140	96	148	140

3. Problem pięcio-, dziesięcio- oraz dwudziestowymiarowy

Znaleźć minimum funkcji

$$f = \sum_{i=1}^{n} \left[E_i - \sum_{j=1}^{n} (A_{ij} \sin x_j + B_{ij} \cos x_j) \right]^2, \qquad (256)$$

względem x_j, przy czym współczynniki A_{ij} oraz B_{ij} utworzono z liczb całkowitych pseudo-losowych w przedziale $[-100, 100]$, natomiast wolne wyrazy E_i przyjęto jako

$$E_i = \sum_{j=1}^{n} (A_{ij} \sin x_j + B_{ij} \cos x_j) \quad i = 1, \ldots, n, \qquad (257)$$

gdzie x_j (dla $j = 1, \ldots, n$) jest szukanym rozwiązaniem, które przed rozpoczęciem wyliczeń wygenerowano ze zbioru liczb rzeczywistych pseudo-losowych w przedziale $[-\pi, \pi]$. Zwróćmy uwagę, że tak sformułowane zadanie testowe umożliwia łatwe powiększenie wymiarowości problemu. Po raz pierwszy zostało ono opublikowane przez Fletchera i Powella (1963).

Obliczenia dokonano dla różnych punktów startowych tworzonych w następujący sposób

$$\underline{x}_o = \hat{\underline{x}} + \underline{\delta},$$

przy czym za każdym razem składowe wektora $\underline{\delta}$ generowano ze zbioru liczb rzeczywistych pseudo-losowych w przedziale $[-\pi/10, \pi/10]$.

Wyniki minimalizacji przedstawiono w tablicy 7 w postaci równoważnej liczby obliczeń funkcji, którą należało wykonać aby spełniony był warunek $|f(\underline{x}) - f(\hat{\underline{x}})| \leq 0{,}0001$.

W przeciwieństwie do poprzednich przykładów wyniki te nie są tak reprezentatywne. Powstało to stąd, że uzyskano je na różnego rodzaju maszynach cyfrowych oraz przy pomocy różnych translatorów. Dlatego też zalecana jest daleko idąca ostrożność przy formułowaniu wniosków.

Porównanie wyników obliczeń dla problemu wielowymiarowego

Metoda	Wymiarowość problemu		
	5	10	20
R	465	1 210	10 208
	465	1 258	4 681
	388	1 298	8 411
DSC	303	2 269	5 183
	281	938	5 924
	307	1 378	8 254
N	229	752	6 970
	195	962	12 100
	298	970	10 246
P	104	329	1 519
	103	369	2 206
GS	354	1 639	4 200
	288	2 860	7 854
	216	1 276	12 348
D	114	396	1 764
	138	349	1 428

5.4.3. Wnioski

Z przytoczonych w punkcie 5.4.2 rezultatów obliczeń wynika, że najefektywniejszymi metodami są: metoda Powella - P oraz metoda Davidona - D. Pierwszą z nich zaliczamy do metod "bezpośrednich poszukiwań", drugą zaś do metod "gradientowych", przy czym obydwie charakteryzują się zbieżnością drugiego rzędu. Warto jednak wspomnieć, że w niektórych przypadkach większą szybkość zbieżności niż metoda D posiada metoda Newtona-Raphsona NR, która jednak nie znalazła się w wykazie metod zakwalifikowa-

nych do analizy. Spowodowane to zostało tym, że jak wykazały badania przedstawione w [57] zakres stosowania metody NR jest ograniczony z dwóch przyczyn. Po pierwsze, w miarę zwiększania wymiarowości czasochłonność obliczeń szybko wzrasta, ze względu na konieczność obliczania $\frac{n(n+1)}{2}$ pochodnych drugiego rzędu oraz odwracania macierzy A. Po drugie, metoda ta jest bardzo czuła na dobór punktu startowego, co często powoduje niezbieżność metody. Zjawisko to wynika z błędu jaki powstaje przy pomijaniu form wyższych niż kwadratowe w rozwinięciu badanej funkcji na szereg Taylora.

Podobną czułością charakteryzuje się również metoda Powella - PI, której charakter zbieżności silnie zależy od doboru punktów startowych. W przypadku punktów mało korzystnych krzywa zbieżności tej metody zbliżona jest do krzywej kwadratowej, a więc w początkowym okresie zbieżność jest dosyć wolna, natomiast w otoczeniu minimum zaczyna szybko wzrastać. Tego rodzaju przebieg można wyjaśnić tym, że początkowo poszukiwanie odbywa się wzdłuż ortogonalnej bazy kierunków, w wyniku czego przesunięcie jest znikomo małe, dopiero utworzenie sprzężonych kierunków określa właściwy kierunek poszukiwania minimum i od tego momentu efektywność minimalizacji rośnie. W przypadku natomiast korzystnego doboru punktów startowych krzywa zbieżności metody PI ma charakter wykładniczy typu $1-e^{-y}$, a więc znacznie lepszy od liniowego, którym cechują się metody R i N.

Jak wykazały badania przedstawione w [56], podobne właściwości do metody PI posiada metoda Z, natomiast stosując metodę PII uzyskuje się znacznie gorsze rezultaty. Pogorszenie to wypływa bezpośrednio z przyjętego w niej sposobu modyfikacji kierunków poszukiwań, która wykonywana jest nie co kolejny "obieg" jak w metodzie PI, lecz tylko jeśli spełniony jest warunek (185). Zasada ta wprowadza pewnego rodzaju opóźnienie w tworzeniu sprzężonych kierunków poszukiwań, a tym samym powoduje, że ilość obiegów wzrasta.

Najmniej czułą na zmianę punktów startowych okazała się metoda Rosenbrocka R, która chociaż dosyć często jest znacznie wolniejsza od metody P czy NR, to jednak w każdej prawie sytuacji daje zadowalające wyniki. Dlatego też wydaje się, że w przypadkach gdy funkcja celu zawiera nieliniowości stopnia wyższego niż dwa warto posługiwać się tą metodą, jeśli oczywiście nie można zastosować metody gradientowej D.

Oprócz wspomnianych metod Davidona i Powella zbieżnością drugiego rzędu charakteryzuje się także metoda gradientu sprzężonego GS. Jednak jak wynika to z uzyskanych rezultatów metoda GS ustępuje im wyraźnie. Wytłumaczenia tego faktu można szukać w specyfice algorytmu Fletchera i Reevesa. Algorytm ten bowiem okresowo co n kroków (gdzie n oznacza wymiarowość

problemu) odrzuca całą nagromadzoną wiedzę o badanej powierzchni. W tych zwrotnych momentach proces szukania minimum nie rozpoczyna się wzdłuż kierunku sprzężonego, lecz wzdłuż aktualnie wyliczonego kierunku minus gradientu. Opisana sytuacja ulega radykalnej poprawie gdy badana funkcja celu jest wystarczająco dobrze aproksymowana formą kwadratową. W takich przypadkach efektywność metody sprzężonego gradientu szybko wzrasta i staje się ona porównywalna z efektywnością metod P i D. Istotną natomiast zaletą metody GS jest to, że nie wymaga ona tak dużej pamięci operacyjnej jak metoda Davidona. Powodem tego jest znacznie prostszy w stosunku od metody D algorytm tworzenia kierunków poszukiwań. Stąd też, niejednokrotnie opłaca się bardziej stosować metodę gradientu sprzężonego GS zamiast metody D pomimo pewnej straty na szybkości obliczeń. Ogólnie biorąc dla dowolnego typu problemu optymalizacji statycznej metody posiadające zbieżność drugiego rzędu charakteryzują się o wiele lepszą skutecznością obliczeń niż wszystkie pozostałe.

Z przedstawionych rezultatów wynika również, że efektywność metody simplexu maleje ze wzrostem wymiarowości problemu. Nelder i Mead w swojej publikacji [39] starali się udowodnić wyższość ich metody nad metodą Powella poprzez porównanie wyników uzyskanych z optymalizacji funkcji o wymiarowości nie większej niż cztery. Postępowanie takie doprowadziło ich do wyciągnięcia fałszywych wniosków, co wyraźnie widać z tablicy 7. Tak więc metodę tę można jedynie stosować do problemów o małej wymiarowości (<4). Posiada ona wówczas tę bardzo korzystną cechę, że dobór punktów startowych położonych w obszarach funkcji reprezentujących sobą strome zbocza "dolin" czy też wąskie "wąwozy" nie powoduje zmniejszenia szybkości zbieżności metody N. Fakt ten można wyjaśnić tym, że w trakcie posuwania się w kierunku minimum metoda ta automatycznie eliminuje silne nieliniowości. Wypływa to ze sposobu wyliczania środka symetrii simplexu $\bar{P}$, który jest obliczany z pominięciem wierzchołka F_{max}. Jak wynika z przeprowadzonych badań metoda N charakteryzuje się prawie liniową zbieżnością, a tym samym kolejne dokładności osiągane są z tą samą szybkością. Wadą jej natomiast jest dosyć duża czułość na ukształtowanie obszaru. Dla punktów startowych leżących na stosunkowo płaskich zboczach niewiele odległych od minimum, metoda N staje się wolniejsza w stosunku do innych metod. Wykonuje ona wtedy wiele niepotrzebnych obliczeń wartości funkcji w wierzchołkach simplexu, w rezultacie nieznacznych różnic występujących pomiędzy nimi.

Reasumując nasze rozważania można stwierdzić, że najbardziej zalecaną metodą optymalizacji jest metoda Davidona w wersji Fletchera i Powella.

6. Metody poszukiwania ekstremum z ograniczeniami

Wśród wielkiego bogactwa metod stosowanych dziś dla numerycznego rozwiązania zadania (A) można wyróżnić następujące cztery podejścia:

- pierwszym sposobem jest sprowadzenie problemu z ograniczeniami do problemu bez ograniczeń przez "transformację zmiennych niezależnych" w taki sposób, że funkcja celu pozostaje niezmieniona. Następnie poszukuje się ekstremum tej funkcji względem nowego układu współrzędnych jedną z metod iteracyjnych bez ograniczeń. Metoda ta została opublikowana przez Boxa [4].
- drugim sposobem jest modyfikacja funkcji celu przez wprowadzenie do niej wyrażenia reprezentującego karę za przekroczenie ograniczeń (tzw. funkcję kary). Następnie do tak zmienionej funkcji stosuje się jedną z metod poszukiwania ekstremum bez ograniczeń. W zależności od postaci oraz sposobu wprowadzenia funkcji kary można wymienić pięć podstawowych koncepcji:
  1) metoda Schmita i Foxa [50] będąca dalszym rozwinięciem metody Couranta [8],
  2) metoda Rosebrocka [48] ulepszona przez Boxa [5],
  3) metoda Carrolla - CRST (Created Response Surface, Technique) [6], matematycznie uzasadniona i rozwinięcia przez Fiacco i McCormicka [15], [16],
  4) metody SUMT (Sequential Unconstrained Minimization Technique) opracowane głównie przez Fiacco i McCormicka [17] oraz Huarda [32],
  5) metoda Powella [45], która następnie została uogólniona i matematycznie uzasadniona przez Wierzbickiego, Michalskiego i Szymanowskiego [55], [60].
- trzecim sposobem jest modyfikacja kierunku poszukiwań w otoczeniu ograniczeń bądź to przez "odbicie się" od ograniczeń, a następnie kontynuowanie poszukiwań w obszarze dopuszczalnym, bądź też przez rzutowanie kierunku gradientu na powierzchnię styczną do ograniczeń. Powstało wiele metod opartych na tych koncepcjach. Jako reprezentującą pierwszą grupę można wymienić metodę Klingmana i Himmelblaua [33], zwaną "Multiple Gradient Summation Technique", natomiast drugiej - metodę Rosena [47] i jej modyfikacje [9],
- czwartym sposobem jest tworzenie tzw. "ograniczonego simpleksu", który następnie krok za krokiem jest tak przekształcany, aż w bliskim otoczeniu poszukiwanego ekstremum warunkowego odległość pomiędzy jego wierzchołkami stanie się mniejsza od założonej dokładności obliczeń ε. Metoda ta została opublikowana przez Boxa [5] pod nazwą "Complex method".

Jak już zostało powiedziane, metody zaliczone do grupy pierwszej i drugiej sprowadzają zadanie optymalizacji z ograniczeniami (1), (2) do zadania bez ograniczeń. Wynika z tego, że w tych przypadkach o efektywności optymalizacji w dużym stopniu decydować będzie wybór odpowiedniej metody minimalizacji bądź maksymalizacji funkcji celu bez ograniczeń. Pełny opis tych metod wraz ze wskazówkami umożliwiającymi właściwy ich wybór przedstawiono w poprzednim rozdziale.

W naszych dalszych rozważaniach dla większej wygody zadanie optymalizacji (A) punkt 1 sformułujemy w trochę odmienny sposób, a mianowicie:

znaleźć wektor $\hat{\underline{x}}$, który ekstremalizuje skalarną funkcję

$$F = f(\underline{x}), \tag{258}$$

spełniając równocześnie układ równań lub nierówności o postaci

$$g_i(\underline{x}) \left\{ =, \geqslant \right\} 0, \qquad i = 1, \ldots, m, \tag{259}$$

przy czym przyjmujemy ponadto następujące założenia:

(A) wnętrze obszaru dopuszczalnego $R^o = \left\{ \underline{x} \middle| g_i(\underline{x}) > 0, \; i = 1, 2, \ldots m \right\}$ nie jest puste, a więc można znaleźć taki punkt $\underline{x}^o$, że $\underline{x}^o \in R^o$,

(B) funkcja celu $f(\underline{x})$ oraz funkcje ograniczeń $g_i(x)$ są klasy C^2,

(C) funkcja celu $f(\underline{x})$ jest ograniczona od dołu w obszarze dopuszczalnym $R = \left\{ \underline{x} \middle| g_i(\underline{x}) \geqslant 0, \; i = 1, 2, \ldots m \right\}$,

(D) zbiór $S = \left\{ \underline{x} \middle| f(\underline{x}) \leqslant K, \; \underline{x} \in R \right\}$ jest ograniczony dla każdej skończonej wartości K,

(E) funkcje $f(\underline{x})$ oraz $g_i^{-1}(\underline{x})$ są funkcjami wypukłymi w R^o, co zachodzi, gdy funkcje ograniczeń $g_i(\underline{x})$ są wklęsłe,

(F) przynajmniej jedno $f(\underline{x})$, $g_i^{-1}(\underline{x})$ jest ściśle wypukłe.

Tak sformułowane założenia, zapewniają nam, że w obszarze R istnieje jedno globalne ekstremum oraz, że omawiane w następnym punkcie algorytmy są zbieżne do tego eksstemum. W wielu przypadkach, założenia te mogą ulec znacznemu osłabieniu, przy jednoczesnym zachowaniu warunku zbieżności. Ciekawe potraktowanie tego tematu można znaleźć w [16], [18] i [60], natomiast w niniejszej pracy dalej się tym problemem nie będziemy zajmować.

Przejdźmy teraz do rozpatrzenia głównych, reprezentacyjnych metod poszukiwania ekstremum z ograniczeniami.

6.1. Transformacja zmiennych niezależnych

Metoda ta polega na takiej transformacji zmiennych niezależnych, że zadanie optymalizacji z ograniczeniami zostaje sprowadzone do zadania bez ograniczeń, natomiast funkcja celu pozostaje nie zmieniona. Zakres stosowania tej metody jest jednak ograniczony do wąskiej klasy zadań programowania nieliniowego. Głównie znajduje ona zastosowanie w przypadku gdy wszystkie bądź tylko niektóre zmienne niezależne x_i są ograniczone od dołu i od góry przez stałe wartości liczbowe l_i oraz u_i odpowiednio. Oznacza to, że postać tych ograniczeń jest następująca

$$l_i \leqslant x_i \leqslant u_i. \tag{260}$$

W zależności od konkretnych wartości liczbowych jakie mogą przyjmować l_i oraz u_i wyróżniamy następujące transformacje:

1) jeżeli $l_i = 0$, $u_i = +\infty$ to mamy do wyboru

$$x_i = \operatorname{abs}(\bar{x}_i), \qquad x_i = \bar{x}_i^2, \qquad x_i = e^{\bar{x}_i}, \tag{261}$$

gdzie przez $\bar{x}_i$ oznaczono i-tą zmienną niezależną szukanego układu współrzędnych;

2) jeżeli $l_i = 0$, $u_i = 1$ to stosujemy

$$x_i = \sin^2 \bar{x}_i \quad \text{lub} \quad x_i = \frac{e^{\bar{x}_i}}{e^{\bar{x}_i} + e^{-\bar{x}_i}}. \tag{262}$$

Natomiast w ogólnym przypadku ograniczeń typu (260) korzystamy z transformacji

$$x_i = l_i + (u_i - l_i) \sin^2 \bar{x}_i. \tag{263}$$

Zwróćmy uwagę na istotną cechę wszystkich tego rodzaju transformacji, a mianowicie na fakt, że nie mogą one wprowadzać żadnych dodatkowych ekstremów lokalnych.

6.2. Metody z zastosowaniem funkcji kary

Jak już wspomniano we wstępie istnieje pięć podstawowych koncepcji rozwiązywania zadania programowania nieliniowego przy pomocy funkcji kary.

Rozpatrzono je pokrótce.

6.2.1. Metoda Schmita i Foxa

Metoda ta powstała jako dalsze rozwinięcie metody Couranta [8], który po raz pierwszy wprowadził kwadratową funkcję kary o postaci

$$F(\underline{x},\ r) = f(\underline{x}) + r \sum_{l} g_i^2\ (\underline{x}), \tag{264}$$

dla rozwiązania ZPN z ograniczeniami równościowymi $g_i(\underline{x}) = 0$.

Courant intuicyjnie założył, że dla różnych wartości $r = r_k$ $(k = 1, 2, \ldots)$ ciąg kolejnych minimalizacji funkcji $F(\underline{x}, r_k)$ przy $r_k \to \infty$ będzie w granicy dążył do szukanego rozwiązania tzn.

$$\lim_{r_k \to \infty} \left[\min_{\underline{x}} F(\underline{x},\ r_k)\right] = \min_{\underline{x} \in R} f(\underline{x}) = f(\hat{\underline{x}}), \tag{265}$$

gdzie $\hat{\underline{x}}$ oznacza punkt ekstremalny, a
$R = \left\{ \underline{x} \mid g_i(\underline{x}) = 0, \quad i = 1, \ldots, m \right\}$.

Dowód zbieżności tej metody w kilkanaście lat później podał Moser [28].

Schmit i Fox [50] uogólnili omówioną metodę na przypadek ograniczeń nierównościowych tzn. rozpatrywali oni zadanie

$$\min f(\underline{x}),$$

pod warunkiem

$$\begin{aligned} g_i(\underline{x}) &\geqslant 0, \qquad i = 1, 2, \ldots, u, \\ g_i(\underline{x}) &= 0, \qquad i = u+1, \ldots, m. \end{aligned} \tag{266}$$

Dla rozwiązania tego zadania Schmit i Fox zaproponowali modyfikację funkcji celu w postaci

$$F_k(\underline{x}) = \left[f(\underline{x}) - f_k\right]^2 H\left(f_k - f(\underline{x})\right) + \sum_{i=1}^{u} g_i^2\ (\underline{x})\, H\left(g_i(\underline{x})\right) + \sum_{i=u+1}^{m} g_i^2 (\underline{x}), \tag{267}$$

gdzie f_k jest z góry zadaną wartością funkcji $f(\underline{x})$ dla k-tej iteracji spełniającą warunek $f_k < f(\underline{x})$, natomiast funkcja H posiada własność

$$H(z) = \begin{cases} 1 & \text{dla} \quad z < 0 \\ 0 & \text{dla} \quad z \geqslant 0 \end{cases} \tag{268}$$

Przebieg algorytmu jest następujący:

(1) Startując z dowolnie wybranego punktu $\underline{x}^0$ dokonaj minimalizacji funkcji

$$F_1(\underline{x}) = \left[f(\underline{x}) - f_1\right]^2 H\left(f_1 - f(\underline{x})\right) +$$

$$+ \sum_{i=1}^{u} g_i^2(\underline{x})\, H\left(g_i(\underline{x})\right) + \sum_{i=u+1}^{m} g_i^2(\underline{x}), \qquad (269)$$

gdzie wartość f_1 założono w ten sposób, aby

$$f_1 < f(\underline{x}^0) \qquad (270)$$

(2) Jeśli $\underline{x}^1$ jest rozwiązaniem zadania optymalizacji (269) takim, że

$$\min F_1(\underline{x}^1) = 0, \qquad (271)$$

to powtórz krok (1) algorytmu dla $F_2(\underline{x})$ przyjmując $\underline{x}^1$ jako nowy punkt startowy oraz zakładając $f_2 < f_1$. Zwróćmy przy tym uwagę, że dowolny punkt $\underline{x}$, w którym $F_k(\underline{x}) = 0$, spełnia następujące związki

$$f(\underline{x}) \leqslant f_k,$$

$$g_i(\underline{x}) \geqslant 0, \qquad i = 1, 2, \ldots u, \qquad (272)$$

$$g_i(\underline{x}) = 0, \qquad i = u+1, \ldots, m.$$

(3) Powtarzaj czynności opisane w punktach 1 i 2 tzn. minimalizuj $F_k(\underline{x})$ dla monotonicznie malejącego ciągu f_k, $k = 1, 2, \ldots$, aż do momentu gdy dla pewnego f_k uzyskamy

$$\min F_k(\underline{x}) > 0. \qquad (273)$$

Oznacza to, że

$$f_{k-1} \geqslant \min f(\underline{x}) > f_k. \qquad (274)$$

Jeśli otrzymamy punkt $\underline{x}$ spełnia kryterium na minimum to "stop", natomiast jeśli nie, to powtórz procedurę dla wartości f_k wybranej w przedziale (f_{k-1}, f_k).

W praktycznych zastosowaniach metody Schmita i Foxa do transformacji (267) wprowadza się dodatkowo współczynniki wagi λ_i określające ważność poszczególnych ograniczeń. Ponadto, zamiast ostrego kryterium na zakończenie minimalizacji w k-tym kroku o postaci $\min F_k(\underline{x}) = 0$, stosuje się zwykle kryterium łagodniejsze bardziej odpowiednie dla obliczeń numerycznych.

6.2.2. Metoda Rosenbrocka

Metoda ta jest nierozłącznie związana z metodą Rosenbrocka bez ograniczeń (punkt 5.2.2). Zakłada ona postać ograniczeń odmienną od ogólnie przyjętej postaci (259), a mianowicie

$$l_i(\underline{x}) \leqslant X_i(\underline{x}) \leqslant u_i(\underline{x}). \tag{275}$$

Zdefiniowany w ten sposób dopuszczalny obszar zmienności $\underline{x}$, Rosenbrock podzielił na trzy strefy: strefę dozwoloną oraz dwie strefy graniczne o szerokości $\alpha = 10^{-4}(u_i(\underline{x}) - l_i(\underline{x}))$ każda. Bieżący punkt $\underline{x}$ znajduje się w strefie granicznej, jeśli spełnione są następujące nierówności:

$$\begin{aligned} l_i(\underline{x}) \leqslant X_i(\underline{x}) < l_i(\underline{x}) + \alpha, \\ u_i(\underline{x}) \geqslant X_i(\underline{x}) > u_i(\underline{x}) - \alpha, \end{aligned} \tag{276}$$

natomiast w strefie dozwolonej gdy

$$l_i(\underline{x}) + \alpha \leqslant X_i(\underline{x}) \leqslant u_i(\underline{x}) - \alpha \tag{277}$$

Koncepcja algorytmu Rosenbrocka polega na zastosowaniu normalnej procedury bez ograniczeń do optymalizacji funkcji wielu zmiennych $F(\underline{x})$, która ulega następującym zmianom w procesie obliczeń:

- jeśli punkt $\underline{x}$ znajduje się w strefie dozwolonej wówczas funkcja celu pozostaje w swojej oryginalnej postaci (258), a więc

$$F(\underline{x}) = f(\underline{x}), \tag{278}$$

- jeśli natomiast punkt $\underline{x}$ wystąpi w strefie granicznej wtedy do funkcji celu (258) zostaje wprowadzone wyrażenie reprezentujące karę za zbliżanie się do ograniczeń. W tym przypadku funkcja $F(\underline{x})$ przyjmuje postać

$$F(\underline{x}) = f(\underline{x}) - (f(\underline{x}) - f^*)(3\eta - 4\eta^2 + 2\eta^3), \tag{279}$$

gdzie f^* jest ostatnio otrzymaną wartością $f(\underline{x})$ przed wejściem do strefy granicznej, a η wyraża się przez

$$\eta = \frac{l(\underline{x}) + \alpha - X(\underline{x})}{\alpha}$$

lub (280)

$$\eta = \frac{X(\underline{x}) - u(\underline{x}) + \alpha}{\alpha}$$

Zwróćmy uwagę, że dla $\eta = 0$, $F(\underline{x}) = f(\underline{x})$ oraz dla $\eta = 1$, $F(\underline{x}) = f^*$ przy czym $f^* \geqslant f(\underline{x})$. Stąd, w strefie granicznej, przy tego rodzaju modyfikacji, zostaje utworzone minimum funkcji występujące dla wartości η zawartych pomiędzy 0 a 1.

Przebieg algorytmu jest następujący:

(1) przyjmując $\underline{x}^o$ jako punkt startowy oblicz wartość funkcji celu $f(\underline{x}^o) \Rightarrow f^* \Rightarrow F0$, wartości ograniczeń (275) oraz szerokość strefy granicznej α,

(2) wykonaj krok o długości e_1 w pierwszym kierunku $\underline{\xi}_1$, należącym do bazy wyjściowej $\underline{\xi}_1, \underline{\xi}_2, \ldots, \underline{\xi}_n$ utworzonej z wzajemnie ortogonalnych wektorów jednostkowych,

(3) oblicz w tym nowym punkcie $\underline{x}$ wartość funkcji $f(\underline{x})$ oraz wartości ograniczeń (275),

(4) zbadaj czy nowy punkt $\underline{x}$ spełnia zbiór ograniczeń (275) jeśli tak, to przejdź do wykonania kroku 6, natomiast jeśli nie, to

(5) cofnij się do punktu poprzedniego tzn. oblicz $\underline{x} - e_j \underline{\xi}_j \Rightarrow \underline{x}$, oraz zmniejsz długość kroku o $-\beta\,(0 < \beta < 1)$ w myśl zasady $-\beta e_j \Rightarrow e_j$, a następnie przejdź do wykonania kroku 8,

(6) zbadaj czy krok był pomyślny tzn. czy $f(\underline{x}) < F0$; jeśli nie, to przejdź do wykonania kroku 5, natomiast w przeciwnym razie zarejestruj ten fakt przez powiększenie sumy pomyślnych kroków w danym kierunku d_j o e_j, a następnie zwiększ długość kroku o $\gamma\,(\gamma > 1)$ oraz podstaw $f(\underline{x}) \Rightarrow F0$,

(7) zbadaj czy punkt $\underline{x}$ znajduje się w strefie granicznej (276) jeśli tak, to dokonaj modyfikacji funkcji celu zgodnie z (279), jeśli nie - podstaw $f(\underline{x}) \Rightarrow f^*$,

(8) powtórz powyższą procedurę od kroku 2 dla wszystkich kierunków $\underline{\xi}_j$ ortogonalnej bazy, a następnie zbadaj czy wykonanie kroków wzdłuż n kierunków (w jednym "obiegu") dało niepomyślny wynik. Jeśli nie, to rozpocznij procedurę od (2) natomiast w przeciwnym przypadku:
- jeżeli nastąpiło to po pierwszym "obiegu" to zmień punkt startowy $\underline{x}^o$ i powtórz czynności od (1),
- jeżeli nie, to o ile znaleziony punkt nie spełnia kryterium na "minimum" dokonaj "obrotu współrzędnych" $\{\underline{\xi}\}$ zgodnie z algorytmem przedstawionym w punkcie 5.2.2 i powróć do wykonania kroku 2.

W algorytmie tym pominięto zagadnienie doboru punktu startowego $\underline{x}^o$, który musi znajdować się w obszarze dopuszczalnym. Problem ten zostanie poruszony w następnym punkcie przy rozpatrywaniu metody Carrolla. Szczegółowy opis omówionej procedury

wraz z sieciami działań można znaleźć w pracach [48], [54], natomiast jej program w ALGOLU opublikowano w [55].

6.2.3. Metoda Carrolla

Metoda ta zwana Created Response Surface Technique - CRST w pierwszej wersji została opracowana przez Carrolla [6], a następnie rozwinięta przez Fiacco i McCormicka [15], [16]. Nazwa jej pochodzi stąd, że istotą metody jest generowanie w obszarze dopuszczalnym ciągu nieograniczonych powierzchni, których kolejne ekstrema, wyznaczane jedną z metod optymalizacji bez ograniczeń, są zbieżne do szukanego rozwiązania ZPN o postaci (258), (259).

W przypadku gdy w zadaniu tym występują tylko ograniczenia nierównościowe typu $g_i(\underline{x}) \geq 0$, Carroll zaproponował modyfikować funkcję celu w następujący sposób

$$F(\underline{x},\ r_k) = f(\underline{x}) + r_k \sum_{i} g_i^{-1}(\underline{x}), \tag{281}$$

natomiast dla ograniczeń równościowych i nierównościowych Fiacco i McCormick uogólnili tę transformację do postaci

$$F(\underline{x},\ r_k) = f(\underline{x}) + r_k \sum_{i=1}^{u} g_i^{-1}(\underline{x}) + r_k^{-\frac{1}{2}} \sum_{i=u+1}^{m} g_i^{2}(\underline{x}) . \tag{282}$$

Jak już wspomniano koncepcja metody Carrolla polega na ciągu kolejnych minimalizacji funkcji (281) bądź (282) dla odpowiadających im dyskretnych, monotonicznie malejących wartości r_k, w rezultacie czego przy $r_k \to 0$ uzyskane zostaje ekstremum zadania z ograniczeniami (258), (259), a więc

$$\lim_{r_k \to 0} \left[\min_{\underline{x}} F(\underline{x},\ r_k) \right] = \min_{\underline{x} \in R} f(\underline{x}) = f(\hat{\underline{x}}), \tag{283}$$

gdzie: $\hat{\underline{x}}$ oznacza punkt ekstremalny,
R - obszar dopuszczalny zdefiniowany tak jak w (259).

Dowód zbieżności metody Carrolla został wykonany przez Fiacco i McCormicka [15].

Przebieg algorytmu jest następujący:

(1) dobierz punkt startowy $\underline{x}^0$ w taki sposób, aby znajdował się w obszarze dopuszczalnym tzn. $x^0 \in R$, gdzie

$$R = \left\{ \underline{x} \mid g_i(\underline{x}) > 0, \quad i = 1, 2, \ldots, m \right\},$$

(2) określ początkową wartość współczynnika przybliżeń r_1,

(3) dokonaj minimalizacji funkcji (281) oraz uzyskany punkt ekstremalny $\hat{\underline{x}}(r_k)$ podstaw w miejsce $\underline{x}^o$,

(4) zbadaj czy spełnione zostało kryterium na "minimum". Jeśli tak, to zakończ działanie procedury, natomiast w przeciwnym razie zmień wartość współczynnika r_k tak aby $r_k > r_{k+1} > 0$ dla $k = 1, 2, \ldots$ oraz przyjmując ostatnio wyliczony $\underline{x}^o$ jako nowy punkt startowy powtórz krok (3).

W omawianej procedurze dwie sprawy wymagają krótkiego komentarza.

Pierwszą - jest zagadnienie doboru punktu startowego $\underline{x}^o$, który musi należeć do obszaru dopuszczalnego R. W wielu zastosowaniach problem ten nie powoduje większych kłopotów, gdyż wartość $\underline{x}^o$ wynika bezpośrednio z właściwości danego procesu. Jednakże istnieją również przypadki gdy trafienie w obszar R nie jest od razu oczywiste. Dla rozwiązania tego zagadnienia Fiacco [15] zaproponował następującą procedurę:

(1) przyjmując dowolny punkt startowy $\underline{x}^o$ zbadaj czy w zbiorze ograniczeń $g_i(x) > 0$ dla $i = 1, 2, \ldots, m$ istnieje takie i dla którego ograniczenia te są niespełnione, tzn. czy $i \in I_2$ gdzie $I_2 = \left\{ i \mid g_i(\underline{x}^o) \leqslant 0 \right\}$,

(2) jeżeli miało to miejsce dla $i = p \in I_2$ to przy pomocy metody CRST dokonaj minimalizacji funkcji

$$T(\underline{x}, r_k) = -g_p(\underline{x}) + r_k \sum_{i \in I_1} g_i^{-1}(\underline{x}), \qquad (284)$$

gdzie $I_1 = \left\{ i \mid g_i(\underline{x}^o) > 0 \right\}$,
którą wykonuj aż do momentu gdy $g_p(\underline{x})$ stanie się dodatnie.

(3) zbadaj czy zbiór I_2 jest pusty. Jeśli tak, to ostatnio uzyskany punkt $\hat{\underline{x}}(r_k)$ przyjmij jako nowy punkt startowy $\underline{x}^o$, natomiast jeśli nie, to powtórz krok (2) dla następnego wybranego $i \in I_2$.

W przypadku, gdy omówiona procedura nie przyniesie spodziewanego rezultatu, tzn. po zakończeniu jej działania będą istniały takie $i = p \in I_2$, dla których $\min \left[-g_p(x) \right] > 0$, należy uznać, że zbiór ograniczeń $g_i(\underline{x})$ występujący w zadaniu optymalizacji jest sprzeczny.

Drugą sprawą, która wymaga wyjaśnienia, jest zagadnienie doboru początkowej wartości współczynnika przybliżeń r_k. Jak można bowiem wykazać nie powinno się go przyjmować dowolnie. Zbyt duża wartość r_k powoduje, że w procesie minimalizacji $F(\underline{x}, r_k)$ dominującą rolę odgrywa wyraz $\sum_i g_i^{-1}(\underline{x})$ i w wyniku tego otrzymany punkt $\hat{\underline{x}}(r_k)$ może być bardzo odległy od rzeczy-

wistego minimum. Natomiast w drugim przypadku, gdy r_k jest za małe, istnieje niebezpieczeństwo naruszenia ograniczeń w trakcie optymalizacji $F(\underline{x}, r_k)$ oraz utworzenie się bardzo wąskiej "doliny". Algorytm automatycznego doboru r_k został podany przez Fiacco i McCormicka [16], oraz także przez Boxa [34].

Przy rozpatrywaniu metody Carrolla nic dotychczas nie zostało powiedziane o wpływie metod optymalizacji bez ograniczeń na efektywność tej metody. Problem ten zostanie poruszony w dalszej części pracy w punkcie 6.5.

6.2.4. Metody SUMT

Istnieje szereg metod zaliczanych do tej grupy, których nazwa SUMT jest skrótem od "Sequential Unconstrained Minimization Technique". Jako głównych jej reprezentantów można wymienić: metodę "primal-dual" Fiacco i McCormicka [16], metodę Arrowa i Uzawy [1], korzystającą bezpośrednio w algorytmie z funkcji Lagrange'a i warunków Kuhna-Tuckera oraz metodę Huarda [32] zwaną także "method of centers".

Interesującą odmianą tej ostatniej metody jest metoda Fiacco i McCormicka opracowana dla przypadku wypukłego zadania programowania nieliniowego [17]. Zakłada ona transformację o postaci

$$F(\underline{x}, \underline{x}^{k-1}) = \left[f(\underline{x}^{k-1}) - f(\underline{x}) \right]^{-1} + \sum_i g_i^{-1}(\underline{x}), \qquad (285)$$

gdzie $f(\underline{x}^{k-1})$ jest wartością funkcji otrzymaną z poprzednich k iteracji.

Przebieg algorytmu jest następujący:

(1) dobierz punkt startowy $\underline{x}^0$ w taki sposób, aby $\underline{x}^0 \in R^0$ gdzie $R^0 = \{ \underline{x} \mid g_i(\underline{x}) > 0, \quad i = 1, \ldots m \}$ oraz podstaw $k = 1$,

(2) określ $R_1 = \{ \underline{x} \mid f(\underline{x}) \leqslant f(\underline{x}^0); \quad \underline{x} \in R \}$,
gdzie $R = \{ \underline{x} \mid g_i(\underline{x}) \geqslant 0, \quad i = 1, \ldots, m \}$,

(3) dokonaj minimalizacji funkcji (285) w R_k oraz uzyskany punkt ekstremalny $\underline{x}^k$ podstaw w miejsce $\underline{x}^{k-1}$,

(4) powtarzaj krok (3) dla $k = 2, 3, \ldots$ aż do momentu gdy zostanie spełnione kryterium na "minimum", przy czym za każdym razem bierz $R_k = \{ \underline{x} \mid f(\underline{x}) \leqslant f(\underline{x}^{k-1}); \quad \underline{x} \in R \}$

Dowód zbieżności tej metody można znaleźć w pracach [16], [17].

6.2.5. Metoda Powella

Metoda ta w oryginalnej wersji opracowana została przez Powella [45] dla przypadku ograniczeń równościowych $g_i(\underline{x}) = 0$, a następnie uogólniona przez Wierzbickiego [60], Michalskiego i Szymanowskiego [55] w taki sposób, że można ją bezpośrednio stosować do rozwiązywania ZPN typu (258) i (259). Powstało przy tym kilka odmian tej nowej metody, której jedną, zbadaną i sprawdzoną, przedstawiono w niniejszej pracy.

Podobnie, jak to miało miejsce w poprzednio omawianych algorytmach, istotą metody jest poszukiwanie ekstremum warunkowego przez ciąg kolejnych minimalizacji bezwarunkowych zmodyfikowanej funkcji celu. Jednakże w rozpatrywanych dotychczas przypadkach po przekroczeniu ograniczeń lub też zwiększaniu dokładności obliczeń powiększana była stromość funkcji kary. Powodowało to niekorzystny efekt "rowu", którego skutkiem był zwiększony nakład obliczeń numerycznych. Dla złagodzenia tego zjawiska Powell zaproponował "przesuwanie" funkcji kary w trakcie procesu obliczeń, przy czym przesunięcie to uzależnił od wartości przekroczonych ograniczeń. Stąd też, przyjęto następującą postać modyfikacji funkcji celu

$$F(\underline{x},\underline{\sigma},\underline{\theta}) = f(\underline{x}) + \sum_i \sigma_i \left(g_i(\underline{x}) + \theta_i\right)^2 H\left(g_i(\underline{x}) + \theta_i\right), \qquad (286)$$

gdzie $\sigma_i > 0$, $\underline{\sigma} = \left[\sigma_1, \sigma_2, \ldots \sigma_m\right]^T$ jest wektorem współczynników kary,
$\theta_i \leqslant 0$, $\underline{\theta} = \left[\theta_1, \theta_2, \ldots, \theta_m\right]^T$ - wektorem przesunięć kary, zaś funkcja H posiada własność:

$$H(g_i(\underline{x}) + \theta_i) = \begin{cases} 1 & \text{dla} \quad g_i(\underline{x}) + \theta_i < 0 \\ 0 & \text{dla} \quad g_i(\underline{x}) + \theta_i \geqslant 0 \end{cases} \qquad (287)$$

Przebieg algorytmu jest następujący:

(1) dobierz dane wejściowe takie jak: punkt startowy $\underline{x}^o$, maksymalną wartość przekroczenia ograniczenia w punkcie startowym c, wymaganą dokładność uwzględnienia ograniczeń w chwili zakończenia działania procedury c_{min} oraz dane potrzebne dla wywołania procedury poszukiwania ekstremum bez ograniczeń, a następnie podstaw $k = 0$, $\underline{\sigma}^{(k)} = 1$, $\underline{\theta}^{(k)} = 0$,

(2) dokonaj minimalizacji funkcji (286) oraz uzyskany punkt ekstremalny $\hat{\underline{x}}\left(\underline{\sigma}^{(k)}, \underline{\theta}^{(k)}\right)$ podstaw w miejsce $\underline{x}^o$, a ponadto c w miejsce c^o,

(3) oblicz w punkcie $\hat{\underline{x}}(\sigma^{(k)}, \theta^{(k)})$ wartość ograniczeń $g_i(\hat{\underline{x}})$ dla $i = 1, \ldots, m$ oraz nową wartość c w myśl zasady

$$c = \max_i \left\{ |g_i(\hat{\underline{x}})| \;\middle|\; g_i(\underline{x}) + \theta_i < 0, \quad i = 1, 2, \ldots m \right\} \quad (288)$$

(4) zbadaj czy spełnione zostało kryterium na "minimum" tzn. czy $c < c_{min}$. Jeśli tak, to zakończ działanie procedury, natomiast jeśli nie to,

(5) zbadaj czy po minimalizacji (krok 2) nastąpiło zmniejszenie naruszenia ograniczeń tzn. czy $c < c^o$. Jeśli tak, to przejdź do wykonania kroku 8, natomiast w przeciwnym razie podstaw na miejsce c jego wartość przed minimalizacją tzn. $c = c^o$,

(6) dla $i \in I$ gdzie $I = \left\{ i \;\middle|\; \left(|g_i(x)| > m_1 c^o \right) \wedge \left(g_i(\underline{x}) + \theta_i < 0 \right) \right\}$ zmień wartość parametrów $\sigma_i^{(k)}$ i $\theta_i^{(k)}$ według reguły

$$\begin{aligned} \sigma_i^{(k)} &= m_2 \sigma_i^{(k)} , \\ \theta_i^{(k)} &= \theta_i^{(k)} / m_2 , \end{aligned} \quad (289)$$

przy czym $0 < m_1 < 1$ oraz $m_2 \geqslant 1$ są współczynnikami dobieranymi eksperymentalnie. Powell w swojej procedurze przyjął $m_1 = \frac{1}{4}$ oraz $m_2 = 10$.

(7) podstaw $k + 1$ w miejsce k oraz przyjmując ostatnio wyliczony $\underline{x}^o$ jako nowy punkt startowy powtórz krok 2,

(8) jeśli $k = 0$ lub w $k - 1$ iteracji wykonywany był krok 6, to
(i) zmień wartość $\theta_i^{(k)}$ w myśl zasady

$$\theta_i^{(k)} = \min \left\{ g_i(\underline{x}) + \theta_i ; \; 0 \right\} \quad *) \quad (290)$$

oraz podstaw $\theta^{(k)} = \theta_i^{(k)}$ dla $i = 1, \ldots, m$, a następnie przejdź do wykonania kroku 7,

(ii) natomiast w przeciwnym przypadku zbadaj warunek czy $c \leqslant m_1 c^o$. Jeśli jest on spełniony to wykonaj czynności (i) kroku 8, a jeśli nie przejdź do wykonania kroku 6.

---

*) Zapis $\theta_i = \min \left\{ g_i(\underline{x}) + \theta_i ; \; 0 \right\}$ oznacza, że w miejsce θ_i podstawiamy mniejszą z liczb $g_i(\underline{x}) + \theta_i$; 0.

Pewną modyfikacją niniejszego algorytmu jest wersja zaproponowana przez Wierzbickiego w [60], który w pracy tej ponadto przedstawił interesujący dowód zbieżności omawianej metody.

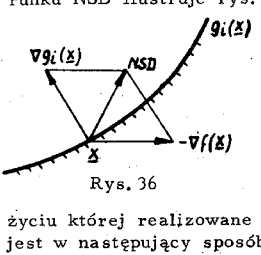
6.3. Metody z zastosowaniem modyfikacji kierunków

Jak już wspomniano istnieją dwie podstawowe grupy metod opartych na koncepcji modyfikacji kierunków poszukiwań w otoczeniu ograniczeń.

Do pierwszej z nich zaliczymy metody, w których z chwilą znalezienia się na ograniczeniu, w wyniku przesuwania się wzdłuż danego kierunku poszukiwań, następuje "odbicie się" od tego ograniczenia w taki sposób, aby w dalszym ciągu pozostać w obszarze dopuszczalnym. Przykładem tej zasady może być metoda Klingmana i Himmelblaua [33] zwana "Multiple Gradient Summation Technique". W metodzie tej po napotkaniu ograniczenia nowy kierunek poszukiwań NSD tworzony jest z kombinacji liniowej gradientu funkcji celu ∇f oraz gradientu funkcji ograniczeń ∇g_i w myśl wzoru

$$NSD = \frac{\nabla g_i(\underline{x})}{|\nabla g_i(\underline{x})|} - \frac{\nabla f(\underline{x})}{|\nabla f(\underline{x})|} . \tag{291}$$

W przypadku pojedynczego ograniczenia sposób tworzenia kierunku NSD ilustruje rys. 36.



Rys. 36

Druga grupa metod wywodzi się od metody Rosena [46], [47] dalej rozwijanej i ulepszonej przez Daviesa [9]. Istotą metody Rosena jest rzutowanie kierunku gradientu na powierzchnię styczną do ograniczeń, a następnie poszukiwanie ekstremum wzdłuż tak wyznaczonego kierunku $\underline{z}$, który okresowo jest uaktualniany. Macierz projekcyjna P_q, przy życiu której realizowane jest wspomniane rzutowanie, tworzona jest w następujący sposób

$$P_q(\underline{x}) \overset{df}{=} I - U_q(\underline{x}) \cdot V_q(\underline{x}) \cdot U_q^T(\underline{x}), \tag{292}$$

gdzie: q oznacza liczbę ograniczeń aktywnych,
I - macierz jednostkową,
U_q - macierz gradientów ograniczeń o postaci

$$U_q(\underline{x}) = \left[\nabla g_1(\underline{x}), \nabla g_2(\underline{x}), \ldots, \nabla g_q(\underline{x})\right] \tag{293}$$

$V_q(\underline{x})$ - macierz odwrotną iloczynu macierzy U_q^T i U_q tzn.

$$V_q(\underline{x}) = \left[U_q^T(\underline{x}) \cdot U_q(\underline{x}) \right]^{-1} \tag{294}$$

Stąd, kierunek $\underline{z}$ styczny do ograniczeń określamy przez

$$\underline{z} = \frac{P_q(\underline{x}) \cdot \nabla f(\underline{x})}{\left\| P_q(\underline{x}) \cdot \nabla f(\underline{x}) \right\|} \cdot \tag{295}$$

Zwróćmy uwagę, że opisana metoda będzie działać prawidłowo tylko pod tym warunkiem, że poszukiwane ekstremum znajduje się na ograniczeniu, a nie we wnętrzu obszaru dopuszczalnego. Dlatego też Rosen przed przystąpieniem do wykonywania normalnej procedury przekształca oryginalne Zadanie Programowania Nieliniowego ZPN do postaci, w której występuje liniowa funkcja celu oraz rozszerzony zbiór nieliniowych ograniczeń nierównościowych typu $g_i(\underline{x}) \geqslant 0$. W celu realizacji tej transformacji do ZPN zostaje wprowadzona dodatkowa zmienna x_{n+1} oraz dodatkowe ograniczenie

$$g_{i+1}(x) = f(\underline{x}) - x_{n+1} \geqslant 0 \tag{296}$$

i tym sposobem, zamiast poszukiwać

$$\min_{\underline{x} \in R} f(\underline{x}), \tag{297}$$

gdzie $R = \left\{ \underline{x} \mid g_i(\underline{x}) \geqslant 0, \quad i = 1, \ldots, m \right\}$, należy wyznaczyć

$$\max_{\underline{x} \in R'} x_{n+1}, \tag{298}$$

gdzie $R' = \left\{ \underline{x} \mid g_i(x) \geqslant 0, \quad i = 1, \ldots, m, m+1 \right\}$.

W rezultacie więc, przekształcone ZNP (298) spełnia postawione wymaganie, gdyż jak wiadomo ekstremum funkcji liniowej leży na ograniczeniu. Dodatkową korzyścią w wprowadzonej transformacji jest to, że gradient funkcji liniowej jest stały, a więc

$$\nabla f^o = \left[0, 0, \ldots, 0, 1 \right], \tag{299}$$

przez co zostaje zmniejszony nakład obliczeń.

W dalszych rozważaniach dla jednolitości zapisu podstawmy $n = n+1$ oraz $m = m+1$, jak również załóżmy, że będziemy poszukiwać maksimum zadania (298).

Przebieg algorytmu metody Rosena jest następujący:

(1) dobierz punkt startowy $\underline{x}^0$ w taki sposób, aby $\underline{x}^0 \in R^0$ gdzie $R^0 = \left\{ \underline{x} \mid g_i(\underline{x}) > 0, \quad i = 1, \ldots, m \right\}$,

(2) startując z punktu $\underline{x}^0$ dokonaj takiego przesunięcia punktu $\underline{x}$ wzdłuż kierunku gradientu ∇f^0, aby znalazł się on w pobliżu któregokolwiek z ograniczeń $g_i(\underline{x})$ tzn. dobierz takie τ żeby punkt $\underline{x}^1 = \underline{x}^0 + \tau \nabla f^0$ należał do B,
gdzie $B = \left\{ \underline{x} \mid \underline{x} \in R, \quad g_i(\underline{x}) = 0, \text{ dla przynajmniej jednego } i \right\}$
oraz podstaw $\underline{x}_\nu = \underline{x}^1$,

(3) określ zbiór ograniczeń aktywnych tzn. zbadaj czy wśród ograniczeń $g_i(\underline{x}) > 0$ istnieje takich q ograniczeń, $1 \leqslant q < m$, dla których $\left\| \underline{w}_q(x_\nu) \right\| < \delta$ gdzie

$$\underline{w}_q(\underline{x}) = \left[g_1(\underline{x}), \; g_2(\underline{x}), \ldots, \; g_q(\underline{x}) \right], \tag{300}$$

natomiast δ jest z góry założoną tolerancją obliczeń,

(4) oblicz w punkcie $\underline{x}_\nu$: macierz odwrotną $V_q(\underline{x}_\nu)$ wzór (294), następnie wektor

$$\underline{r}_q(\underline{x}_\nu) = \left[r_1(\underline{x}_\nu), \ldots, \; r_q(\underline{x}_\nu) \right] = V_q(\underline{x}_\nu) \cdot U_q^T(\underline{x}_\nu) \cdot \nabla f^0, \tag{301}$$

który wyznacza kierunek powrotu do obszaru dopuszczalnego oraz wartość rzutu gradientu na kierunek styczny do przecięcia się ograniczeń w myśl wzoru

$$P_q(\underline{x}_\nu) \cdot \nabla f^0 = \nabla f^0 - U_q(\underline{x}_\nu) \cdot \underline{r}_q(\underline{x}_\nu), \tag{302}$$

(5) oblicz w punkcie $\underline{x}$, wartości współczynników kryterialnych $\beta_1(\underline{x}_\nu)$ oraz $\beta(\underline{x}_\nu)$ według zależności

$$\beta_1(\underline{x}_\nu) = \max_i \left\{ \frac{1}{2} \, r_i(\underline{x}_\nu) \cdot v_{ii}^{-\frac{1}{2}} \right\}, \tag{303}$$

gdzie v_{ii} są diagonalnymi elementami macierzy $V_q(\underline{x}_\nu)$,

$$\beta(\underline{x}_\nu) = \max \left\{ \left\| P_q(\underline{x}_\nu) \; \nabla f^0 \right\| ; \; \beta_1(\underline{x}_\nu) \right\}$$

a następnie zbadaj czy spełnione zostało kryterium na "maksimum" tzn. czy $\beta \leqslant \varepsilon$, gdzie ε jest zadaną dokładnością obliczeń. Jeśli tak, to zakończ działanie procedury, natomiast jeśli nie to

(6) zbadaj czy zbiór ograniczeń aktywnych ulegnie redukcji po wykonaniu kroku w kierunku $\underline{z}$ tzn. sprawdź warunek czy

$\beta > \varepsilon \wedge \left\| P_q(\underline{x}_\nu)\ \nabla f^o \right\| < \beta_1$. Jeśli tak, to przejdź do wykonania kroku 10, natomiast w przeciwnym razie

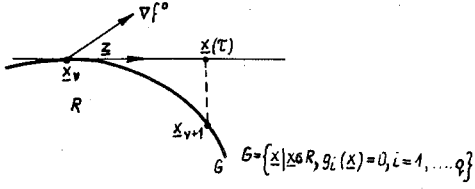
(7) wyznacz w punkcie $\underline{x}_\nu$ kierunek styczny do ograniczeń $\underline{z}(\underline{x}_\nu)$ według (295), a następnie oblicz kolejny punkt $\underline{x}_{\nu+1}$ leżący w obszarze dopuszczalnym w myśl zasady

$$\begin{aligned} \underline{x}(\tau) &= \underline{x}_\nu + \tau\, \underline{z}(\underline{x}_\nu), \\ \underline{x}_{\nu+1}(\tau) &= x(\tau) - U_q(\underline{x}_\nu)\, V_q(x_\nu) \cdot \underline{w}_q(\underline{x}(\tau)), \end{aligned} \tag{304}$$

przy czym τ musi być tak dobrane, aby spełniony był warunek

$$\left\| \underline{w}_q(\underline{x}_{\nu+1}) \right\| \leqslant \delta. \tag{305}$$

Graficzną interpretację opisanych czynności przedstawia rys. 37.



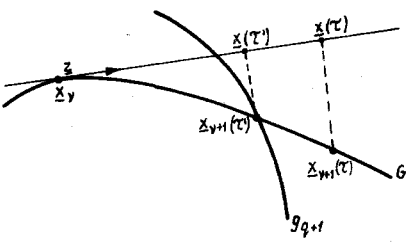
Rys. 37

(8) zbadaj czy w tym nowym punkcie $\underline{x}_{\nu+1}$ spełniane są również ograniczenia uznane poprzednio za nieaktywne tzn. czy

$$g_i(\underline{x}_{\nu+1}) \geqslant 0, \quad \text{dla} \quad i = q+1, \ldots, m. \tag{306}$$

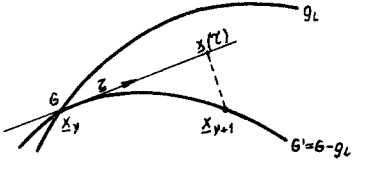
Jeśli tak, to podstaw punkt $\underline{x}_{\nu+1}(\tau)$ w miejsce $\underline{x}_\nu$ oraz przejdź do wykonania kroku 4, natomiast jeśli nie, to

(9) powtarzaj czynności związane z doborem τ zgodnie z wzorami (304) i (305), aż warunek (306) zostanie spełniony, a następnie podstaw znaleziony $\underline{x}_{\nu+1}(\tau')$ w miejsce $\underline{x}_\nu$ oraz rozpocznij procedurę od kroku 3. Przypadek ten został przedstawiony na rys. 38.



Rys. 38

(10) usuń z macierzy U_q wektor ∇g_1 i dla tak utworzonej macierzy U_{q-1} oblicz ponownie macierz V_{q-1}, a następnie $\underline{r} = V_{q-1} U_{q-1} \nabla f^o$, $P_{q-1} \nabla f^o = \nabla f^o - U_{q-1} \underline{r}$ oraz powtórz czynności od kroku 7. Opisaną sytuację przedstawia rys. 39.



Rys. 39

Zwróćmy uwagę, że w omówionej procedurze nie podano sposobu doboru τ występującego w krokach 2, 7 i 9 oraz algorytmu obliczania macierzy odwrotnej V_q, co ma istotny wpływ na efektywność metody. Rosen w swojej oryginalnej procedurze zastosował interpolację sześcienną do wyznaczania τ, natomiast dla określenia V_q posłużył się rekursywnym algorytmem Householdera opisanym w [31].

6.4. Metoda Complex

W przeciwieństwie do metod rozpatrywanych do tej pory, w metodzie Complex nie stosuje się ani modyfikacji funkcji celu, ani też modyfikacji kierunków poszukiwań w otoczeniu ograniczeń. Istota metody polega na utworzeniu w obszarze dopuszczalnym nieregularnego simplexu o k wierzchołkach (przy czym $k > n+1$), który zostaje wpisany w powierzchnię reprezentującą badaną funkcję celu $f(\underline{x})$. Następnie simplex ten, zwany complexem, zostaje tak przekształcany, aby odległość pomiędzy jego wierzchołkami malała przy posuwaniu się w kierunku minimum. Metoda ta została opracowana przez Boxa [5] i jest odpowiednikiem metody Simplex

Neldera i Meada (punkt 5.2.3), stosowanej w przypadku poszukiwania ekstremum bez ograniczeń.

W metodzie Complex, podobnie jak to miało miejsce w metodzie Rosenbrocka punkt 6.2.2 zakłada się odmienną postać ograniczeń, a mianowicie

$$l_i \leqslant x_i \leqslant u_i, \qquad i = 1,2,\ldots n, \tag{307}$$

$$l_j \leqslant X_j(\underline{x}) \leqslant u_j, \qquad j = 1,2,\ldots m, \tag{308}$$

gdzie l_i, l_j, u_i oraz u_j są stałymi lub funkcjami $\underline{x}$, przy czym ograniczenia typu (308) zwane są ograniczeniami funkcyjnymi.

W algorytmie metody Complex stosuje się następujące dwie operacje:

(1) operacja wyliczenia "centroidu" $\underline{c}$ zdefiniowanego jako

$$\underline{c} = \frac{\sum_{\substack{i=1 \\ i \neq h}}^{k} \underline{x}_i^w}{k - 1}, \quad \text{dla} \quad i \neq h, \tag{309}$$

gdzie $\underline{x}_i^w$ oznacza i-ty punkt wierzchołkowy complexu, a h jest indeksem punktu wierzchołkowego $\underline{x}_h^w$, w którym funkcja celu $f(\underline{x})$ osiąga maksimum spośród k punktów complexu tzn. $f(\underline{x}_h^w) = \max$, oraz

(2) operacja "odbicia" punktu $\underline{x}_h^w$ względem $\underline{c}$ określona przez

$$\underline{x}^* = (1 - \alpha)\underline{c} - \underline{x}_h^w \tag{310}$$

gdzie $\alpha > 1$.

Z definicji tej wynika, że punkt $\underline{x}^*$ leży na prostej łączącej punkty $\underline{c}$ i $\underline{x}_h^w$ po stronie przeciwnej $\underline{x}_h^w$ względem $\underline{c}$ w odległości $(\underline{c} - \underline{x}_h^w)\alpha$ -krotnej od $\underline{c}$.

Jako najkorzystniejsze wartości parametrów α i k Box zaproponował $\alpha = 1,3$ oraz $k = 2n$.

Przebieg algorytmu jest następujący:

(1) wyznacz w obszarze dopuszczalnym k punktów complexu w myśl zasady

$$x_i^w = u_i + r_i \left| u_i - l_i \right| \tag{311}$$

gdzie r_i są liczbami pseudo-losowymi wygenerowanymi ze zbioru liczb przypadkowych w przedziale $[0, 1]$.
Jeśli tak wyliczony kolejny punkt $\underline{x}_i^w$ nie spełnia ograniczeń funkcyjnych (308), wówczas przesuń ten punkt w kierunku cent-

rum zaakceptowanych już punktów wierzchołkowych o połowę odległości od tego centrum. Procedurę tę powtarzaj dotąd, aż warunki (308) zostaną spełnione,

(2) określ promień minimalnego koła zawierającego wszystkie punkty complexu - ϱ_{min},

(3) zbadaj czy spełnione zostało kryterium na "minimum" tzn. czy $\varrho_{min} < \varepsilon$. Jeśli tak, to zakończ działanie procedury, natomiast jeśli nie, to

(4) wyznacz spośród k punktów complexu punkt $\underline{x}_h^w$, w którym funkcja celu osiąga wartość maksymalną tzn. wyznacz indeks h taki, że $\left(f\,\underline{x}_h^w\right) = \max$,

(5) oblicz dla $i = 1,2,\ldots,k;\ i \neq h$ centrum complexu $\underline{c}$ według (309), a następnie zbadaj czy znajduje się ono w obszarze dopuszczalnym. Jeśli nie, to wprowadź dodatkowy punkt do complexu, a więc podstaw $k = k+1$ oraz przejdź do wykonania kroku 1, natomiast jeśli tak, to

(6) wykonaj odbicie $\underline{x}^*$ punktu $\underline{x}_h^w$ względem $\underline{c}$ w myśl (310) oraz zbadaj czy punkt $\underline{x}^*$ spełnia ograniczenia (307), (308). Jeśli tak, to przejdź do wykonania kroku 8, natomiast w przeciwnym razie

(7) zbadaj czy zostały naruszone ograniczenia typu (307) czy też (308). Jeśli typu (307) to na miejsce $\underline{x}^*$ podstaw wartość liczbową naruszanego ograniczenia, natomiast jeśli zostało naruszone ograniczenie funkcyjne (308), to przesuń punkt $\underline{x}^*$ w kierunku centrum o połowę odległości $(\underline{c} - \underline{x}^*)$ tzn.

$$\underline{x}^* = \underline{c} - (\underline{c} - \underline{x}^*)/2, \qquad (312)$$

przy czym czynność tę powtarzaj dotąd, aż ograniczenia (308) zostaną spełnione,

(8) zbadaj czy w znalezionym punkcie $\underline{x}^*$ funkcja celu osiąga w dalszym ciągu wartość maksymalną spośród pozostałych punktów complexu. Jeśli tak, to wykonaj operację (312) i powtórz krok 8, natomiast jeśli nie, to podstaw $\underline{x}^*$ w miejsce $\underline{x}_h^w$ i rozpocznij wykonywanie procedury od kroku 2.

Szczegółowy opis omówionego algorytmu można znaleźć w pracach [5], [34], zaś program napisany w ALGOLU w [55].

6.5. Porównanie metod

6.5.1. Wybór metod oraz kryteriów porównawczych

Wśród metod Poszukiwania Ekstremum z Ograniczeniami PEZOG omówionych w poprzednim punkcie, za główne i reprezentacyjne ze względu na zasadę działania można uznać metody: Rosenbrocka (punkt 6.2.2), Carrolla (punkt 6.2.3) oraz Powella (punkt 6.2.5), które zostały oparte o trzy różne koncepcje wprowadzania funkcji kary, następnie metodę Rosena (punkt 6.3) stosującą modyfikację kierunku poszukiwań w sąsiedztwie ograniczeń oraz metodę Complex (punkt 6.4) operującą "ograniczonym simplexem". Dla dokonania porównania rozważanych metod posłużono się więc wymienionymi pięcioma metodami, przy czym w celu zbadania wpływu metod Poszukiwania Ekstremum Bez Ograniczeń PEBOG na efektywność metod z ograniczeniami, rozpatrzono zarówno metodę Carrolla jak i Powella w kilku wersjach korzystających z odmiennych procedur PEBOG. Przyjęto przy tym następujące oznaczenia:

RB - metoda Rosenbrocka

CAR - metoda Carrolla z metodą Rosenbrocka PEBOG

CAZ - metoda Carrolla z metodą Zangwilla PEBOG

CAG - metoda Carrolla z metodą gradientu sprzężonego PEPOG

POW - metoda Powella z metodą Powella PEBOG

POG - metoda Powella z metodą gradientu sprzężonego PEBOG

RS - metoda Rosena

COM - metoda Complex.

Programy w ALGOLU-ZAM, według których przeprowadzono badania wymienionych metod, podane zostały w pracy [55].

Istotnym zagadnieniem przy rozpatrywaniu iteracyjnych metod poszukiwania ekstremum jest dobór właściwego kryterium porównawczego. Oczywistą jest rzeczą, że kryterium takie może być bardzo różnorodnie formułowane w zależności od celu jakiemu ma służyć. Jak już wspomniano w punkcie 5.4.1, w przypadku porównywania metod optymalizacji bez ograniczeń dosyć uniwersalnym kryterium zaproponowanym przez Boxa [4] jest liczba obliczeń wartości funkcji celu jaką należy wykonać dla osiągnięcia żądanej dokładności.

Zaletą tak sformułowanego kryterium jest to, że jest ono dobrą miarą szybkości zbieżności oraz czasu działania procedury, niezależną od typu zastosowanej maszyny cyfrowej, jej organizacji czy też użytego translatora języka programowania.

Box w swoich pracach [4] oraz [5], w których zajmował się porównaniem metod poszukiwania ekstremum z ograniczeniami takich jak: transformacji zmiennych, Rosenbrocka oraz metody Complex, poprzestał tylko na omówionym kryterium. Jednakże podejście takie nie zawsze jest właściwe, gdyż w metodach PEZOG należy również dokonywać obliczeń wartości ograniczeń, których to liczba często nie odpowiada liczbie obliczeń wartości funkcji celu, a ich pracochłonność obliczeń jest także dość znaczna. Przykładem tego rodzaju sytuacji może być metoda Rosena, czy też każda inna metoda, w której uwzględnia się w zmodyfikowanej funkcji celu tylko ograniczenia aktywne. Dlatego też, w niniejszej pracy oprócz kryterium Boxa przyjęto "liczbę obliczeń ograniczeń" jako dodatkowe kryterium porównawcze, które łącznie z poprzednim pozwoli wnioskować o czasie trwania optymalizacji.

6.5.2. Wybór i opis przykładów

Przy doborze przykładów położono głównie nacisk na zbadanie wpływu rodzaju ograniczeń oraz wpływu umiejscowienia ekstremum na efektywność metod. Dla tego celu sformułowano pięć dwuwymiarowych przykładów, w których w różny sposób ukształtowano obszar dopuszczalny oraz ustalono różne położenia punktu ekstremalnego. Natomiast badanie wpływu wymiarowości problemu optymalizacji na efektywność metod przeprowadzono tylko w ograniczonym zakresie na przykładzie pięciowymiarowym zaproponowanym przez Boxa w [5]. Przykłady te mają postać następującą:

1. Problem dwuwymiarowy - obszar ograniczeń wypukły, przy czym występuje jedno minimum globalne położone na krzywoliniowym ograniczeniu (rys. 40).

Znaleźć minimum funkcji

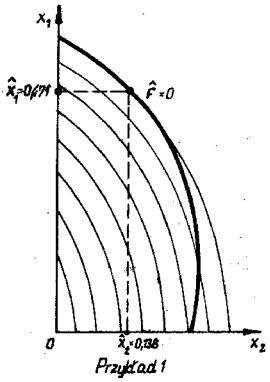
$$f(x_1, x_2) = \frac{1}{2} - \sqrt{\frac{x_1^2 + x_2^2}{x_1^2 + \left(1 - x_2^2\right)}}, \tag{313}$$

przy warunkach

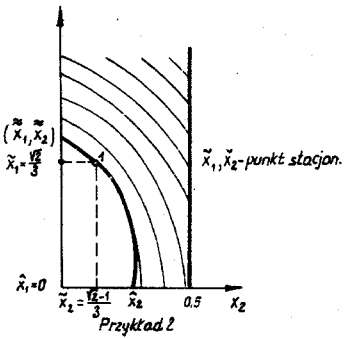
$$g(x_1, x_2) = \frac{1}{4} - \left[\left(x_1 - \frac{\sqrt{2}}{12}\right)^2 + \left(x_2 - \frac{\sqrt{2} - 4}{12}\right)^2\right] \geqslant 0,$$

$$x_1 \geqslant 0, \qquad x_2 \geqslant 0.$$

Szukane rozwiązanie występuje w punkcie $\hat{x}_1 = \frac{\sqrt{2}}{3}$, $\hat{x}_2 = \frac{\sqrt{2} - 1}{3}$, w którym wartość $f(\hat{x}_1, \hat{x}_2) = 0$.



Rys. 40



Rys. 41

2. Problem dwuwymiarowy - obszar ograniczeń wklęsły, w którym występują dwa minima: lokalne i globalne (rys. 41). Minima lokalne $\tilde{\underline{x}}$ i globalne $\hat{\underline{x}}$ położone są w ostrzach, zaś punkt stacjonarny na ograniczeniu krzywoliniowym.

Znaleźć minimum funkcji

$$f(x_1, x_2) = \sqrt{\frac{x_1^2 + x_2^2}{x_1^2 + (1 - x_2)^2}} - \frac{1}{2},$$

przy warunkach

$$g_1(x_1, x_2) = \left(x_1 - \frac{\sqrt{2}}{12}\right)^2 + \left(\frac{\sqrt{2} - 4}{12}\right)^2 - \frac{1}{4} \geqslant 0, \qquad (314)$$

$$g_2(x_1, x_2) = 100 - x_1 \geqslant 0,$$

$$g_3(x_1, x_2) = 0,5 - x_2 \geqslant 0,$$

$$x_1 \geqslant 0, \qquad x_2 \geqslant 0.$$

Rozwiązania powyższego problemu są niejednoznaczne, a więc - w punkcie $\hat{x}_1 = 0$, $\hat{x}_2 = 0,27$ występuje minimum globalne

$$f(\hat{x}_1, \hat{x}_2) = -0,129,$$

- w punkcie $\tilde{x}_1 = 0,569$, $\tilde{x}_2 = 0$ występuje minimum lokalne

$$f(\tilde{x}_1, \tilde{x}_2) = -0,0054,$$

- w punkcie $\tilde{x}_1 = \frac{\sqrt{2}}{3}$, $\tilde{x}_2 = \frac{\sqrt{2} - 1}{3}$ występuje punkt stacjonarny

$$f(\tilde{x}_1, \tilde{x}_2) = 0.$$

3. Problem dwuwymiarowy - ograniczenia liniowe, przy czym występuje jedno minimum leżące na ograniczeniu (rys. 42).

Znaleźć minimum funkcji

$$f(x_1, x_2) = \sqrt{\frac{x_1^2 + x_2^2}{x_1^2 + (1 - x_2)^2}} - \frac{1}{2},$$

przy warunkach

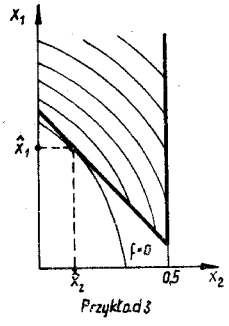
$$g_1(x_1, x_2) = x_1 + x_2 - \frac{2\sqrt{2} - 1}{3} \geqslant 0, \qquad (315)$$

$$g_2(x_1, x_2) = 100 - x_1 \geqslant 0,$$

$$g_3(x_1, x_2) = 0,5 - x_2 \geqslant 0,$$

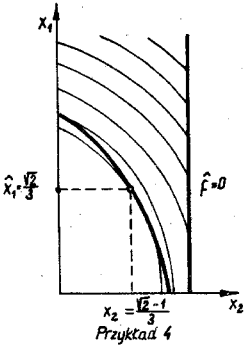
$$x_1 \geqslant 0, \qquad x_2 \geqslant 0.$$

Szukane rozwiązanie występuje w punkcie $\hat{x}_1 = \frac{\sqrt{2}}{3}$, $\hat{x}_2 = \frac{\sqrt{2} - 1}{3}$, w którym wartość $f(\hat{x}_1, \hat{x}_2) = 0$.



Przykład 3

Rys. 42



Przykład 4

Rys. 43

4. Problem dwuwymiarowy - obszar ograniczeń wklęsły, w którym występuje jedno minimum globalne położone na krzywoliniowym ograniczeniu (rys. 43).

Znaleźć minimum funkcji

$$f(x_1, x_2) = \sqrt{\frac{x_1^2 + x_2^2}{x_1^2 + (1 - x_2)^2}} - \frac{1}{2}$$

przy warunkach

$$g_1(x_1, x_2) = \left(x_1 + \frac{\sqrt{2}}{6}\right)^2 + \left(x_2 + \frac{\sqrt{2} + 2}{6}\right)^2 - 1 \geqslant 0,$$

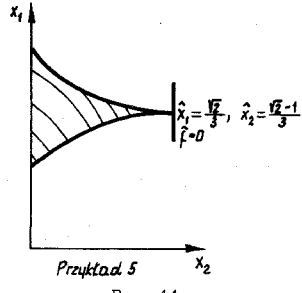
$$g_2(x_1, x_2) = 100 - x_1 \geqslant 0, \qquad (316)$$

$$g_3(x_1, x_2) = 0{,}5 - x_2 \geqslant 0,$$

$$x_1 \geqslant 0, \qquad x_2 \geqslant 0.$$

Szukane rozwiązanie występuje w punkcie $\hat{x}_1 = \frac{\sqrt{2}}{3}$, $\hat{x}_2 = \frac{\sqrt{2} - 1}{3}$, w którym wartość $f(\hat{x}_1, \hat{x}_2) = 0$.

5. Problem dwuwymiarowy - obszar ograniczeń wklęsły, w którym występuje jedno minimum globalne położone w ostrzu (rys. 44).



Rys. 44

Znaleźć minimum funkcji

$$f(x_1, x_2) = \frac{1}{2} - \sqrt{\frac{x_1^2 + x_2^2}{x_1^2 + (1 - x_2)^2}},$$

przy warunkach

$$g_1(x_1, x_2) = x_1^2 + \left(x_2 - \frac{\sqrt{2} - 1}{3}\right)^2 - \frac{2}{9} \geqslant 0,$$

$$g_2(x_1, x_2) = \left(x_1 - \frac{2\sqrt{2}}{3}\right)^2 + \left(x_2 - \frac{\sqrt{2} - 1}{3}\right)^2 - \frac{2}{9} \geqslant 0, \qquad (317)$$

$$g_3(x_1, x_2) = 0,6 - x_1 \geqslant 0,$$

$$g_4(x_1, x_2) = \frac{\sqrt{2} - 1}{3} - x_2 \geqslant 0,$$

$$x_1 \geqslant 0, \qquad x_2 \geqslant 0.$$

Szukane rozwiązanie występuje w punkcie $\hat{x}_1 = \frac{\sqrt{2}}{3}$, $\hat{x}_2 = \frac{\sqrt{2} - 1}{3}$, w którym wartość funkcji $f(\hat{x}_1, \hat{x}_2) = 0$.

6. Problem pięciowymiarowy zaproponowany przez Boxa w [5]. Znaleźć maksimum funkcji celu

$$f(x) = (a_2y_1 + a_3y_2 + a_4y_3 + a_5y_4 + 7840a_6 - 100000\,a_o + - 50800\,ba_7 + k_{31} + k_{32}x_2 + k_{33}x_3 + k_{34}x_4 + - k_{35}x_5)x_1 - 24345 + a_1x_6,$$

gdzie:

$$\begin{aligned}
b &= x_2 + 0,01\,x_3,\\
x_6 &= (k_1 + k_2x_2 + k_3x_3 + k_4x_4 + k_5x_5)\cdot x_1,\\
y_1 &= k_6 + k_7x_2 + k_8x_3 + k_9x_1 + k_{10}x_5,\\
y_2 &= k_{11} + k_{12}x_2 + k_{13}x_3 + k_{14}x_4 + k_{15}x_5,\\
y_3 &= k_{16} + k_{17}x_2 + k_{18}x_3 + k_{19}x_4 + k_{20}x_5,\\
y_4 &= k_{21} + k_{22}x_2 + k_{23}x_3 + k_{24}x_4 + k_{25}x_5,\\
x_7 &= (y_1 + y_2 + y_3)x_1,\\
x_8 &= (k_{26} + k_{27}x_2 + k_{28}x_3 + k_{29}x_4 + k_{30}x_5)x_1 + x_6 + x_7
\end{aligned} \qquad (318)$$

przy ograniczeniach:

$$0 \leqslant x_1;$$

$$1,2 \leqslant x_2 \leqslant 2,4;$$

$$20 \leqslant x_3 \leqslant 60;$$

$$9 \leqslant x_4 \leqslant 9,3;$$

$$6,5 \leqslant x_5 \leqslant 7;$$

$$0 \leqslant x_6 \leqslant 294000;$$

$$0 \leqslant x_7 \leqslant 294000;$$

$$0 \leqslant x_8 \leqslant 277200;$$

przy czym wartość współczynników a_i oraz k_j zaczerpnięto z pracy [5].

Szukane rozwiązanie występuje w punkcie

$$\hat{x}_1 = 4,537; \quad \hat{x}_2 = 2,4; \quad \hat{x}_3 = 60; \quad \hat{x}_4 = 9,3; \quad \hat{x}_5 = 7,$$

w którym wartość funkcji celu $f(\underline{\hat{x}}) = 5\ 280\ 334$.

6.5.3. Wyniki obliczeń

Obliczenia przeprowadzone w ALGOLU na maszynie ZAM 41, przy czym w poszczególnych przykładach dla wszystkich badanych metod przyjmowano te same punkty startowe. Wyjątek stanowi metoda Complex, dla której w przykładzie 3 założono ponadto drugi odmienny punkt startowy. Wyniki obliczeń zostały przedstawione graficznie w postaci dwóch rodzin wykresów:

wartość funkcji = f (ilość obliczeń funkcji)

wartość funkcji = f (ilości obliczeń ograniczeń).

Wykresy te umieszczono w następującej kolejności:

przykład 1 - rys. 45 i 46

przykład 2 - rys. 47 i 48

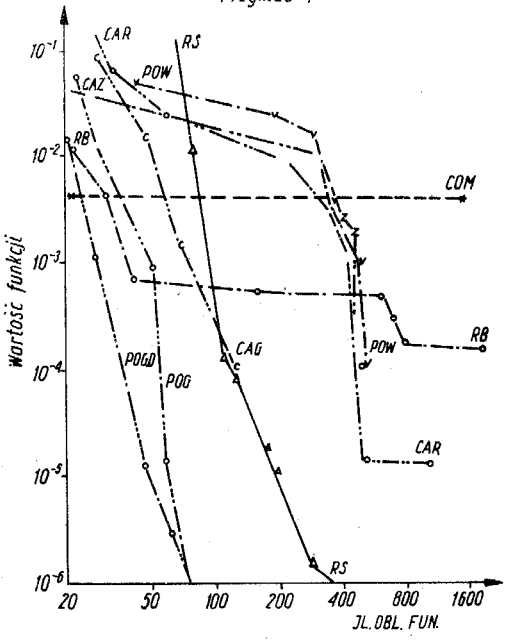
przykład 3 - rys. 49 i 50

przykład 4 - rys. 51 i 52

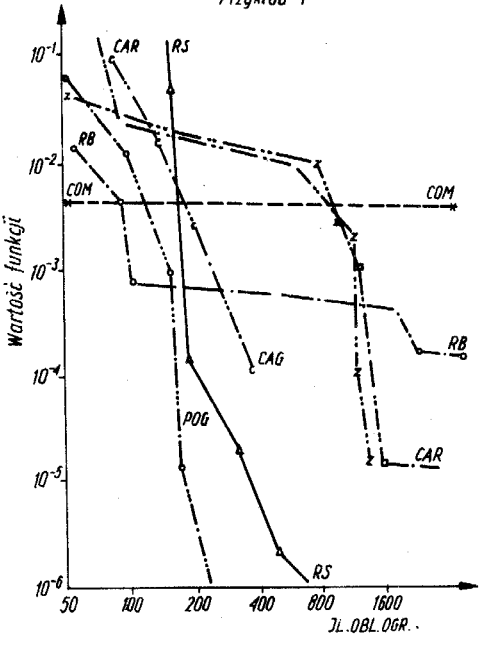
przykład 5 - rys. 53 i 54

przykład 6 - rys. 55 i 56.

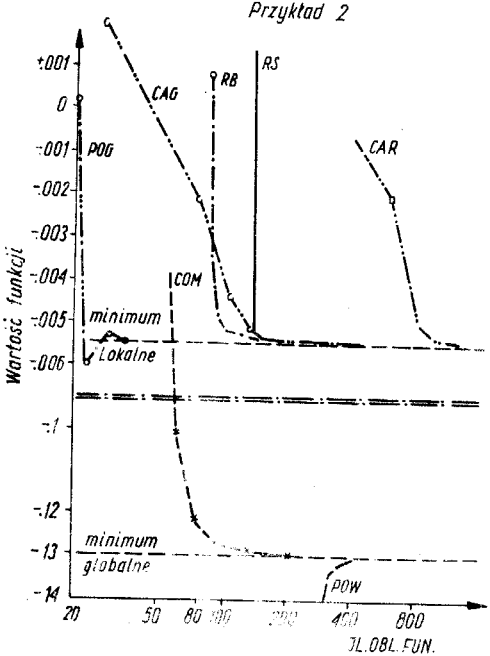
Oznaczenia występujące na tych rysunkach są zgodne z oznaczeniami wprowadzonymi w punkcie 6.5.1.


Przykład 1
CAR
RS
CAZ
POW
RB
COM
CAG
POW
RB
POGD
POG
CAR
RS
Wartość funkcji
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
20
50
100
200
400
800
1600
IL.OBL.FUN.

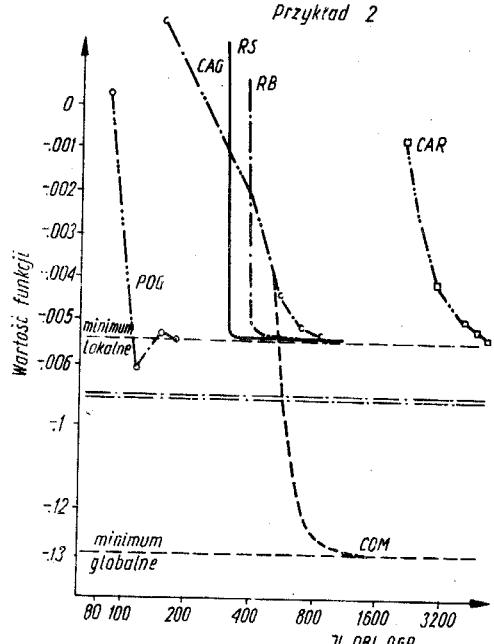
Rys. 45


Przykład 1
CAR
RS
RB
COM
COM
CAG
RB
POG
CAR
RS
Wartość funkcji
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
50
100
200
400
800
1600
IL.OBL.OGR.

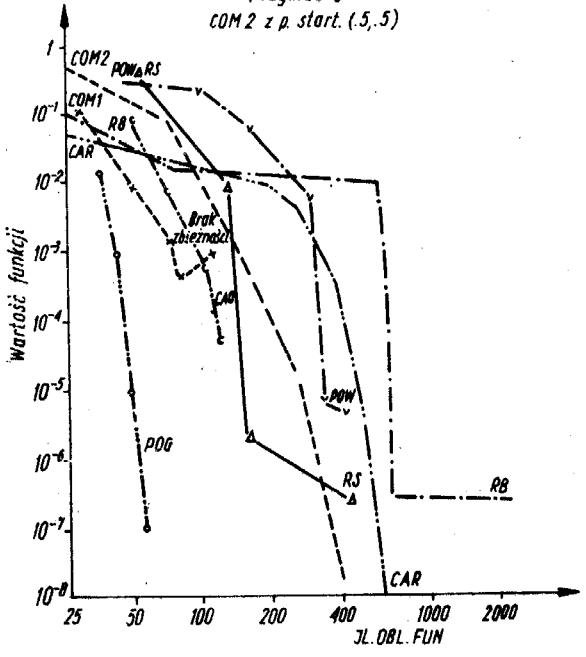
Rys. 46


Przykład 2
+.001
0
-.001
-.002
-.003
-.004
-.005
-.006
-.1
-.12
-.13
-.14
Wartość funkcji
CAG
RB
RS
CAR
POG
COM
minimum
lokalne
minimum
globalne
POW
20
50
80
100
200
400
800
IL. OBL. FUN.

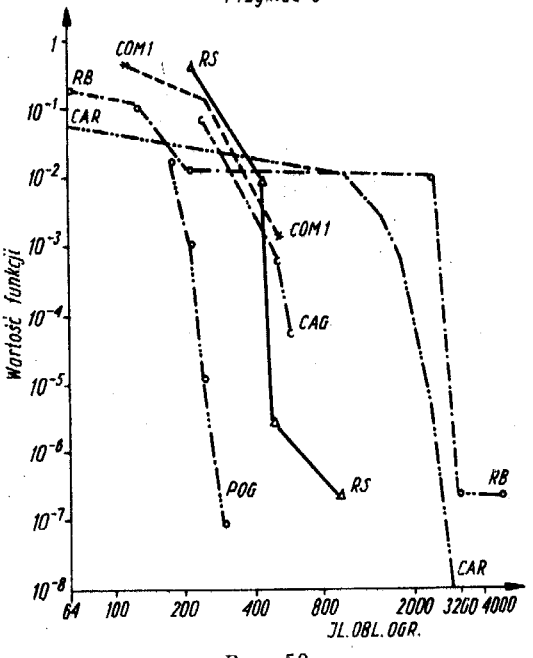
Rys 47


Przykład 2
0
-.001
-.002
-.003
-.004
-.005
-.006
-.1
-.12
-.13
Wartość funkcji
CAG
RS
RB
CAR
POG
minimum
lokalne
minimum
globalne
COM
80
100
200
400
800
1600
3200
IL. OBL. OGR.

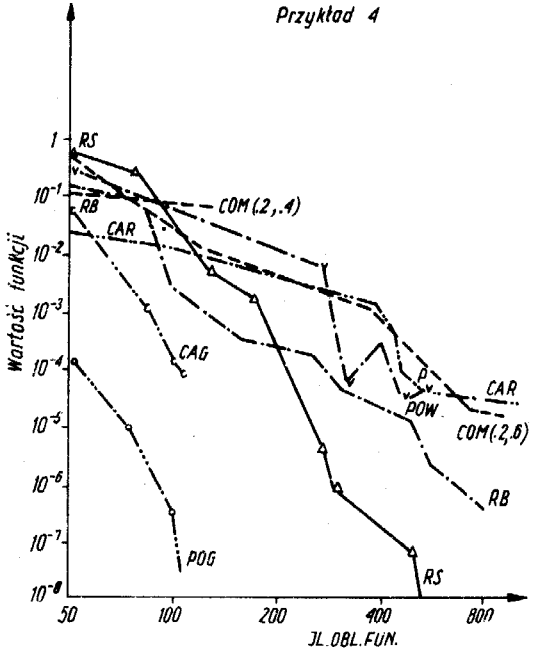
Rys. 48


Przykład 3
COM 2 z p. start. (.5,.5)
COM2
POW
RS
COM1
RB
CAR
Brak zbieżności
CAG
POW
POG
RS
RB
CAR
Wartość funkcji
1
10⁻¹
10⁻²
10⁻³
10⁻⁴
10⁻⁵
10⁻⁶
10⁻⁷
10⁻⁸
25
50
100
200
400
1000
2000
IL.OBL.FUN

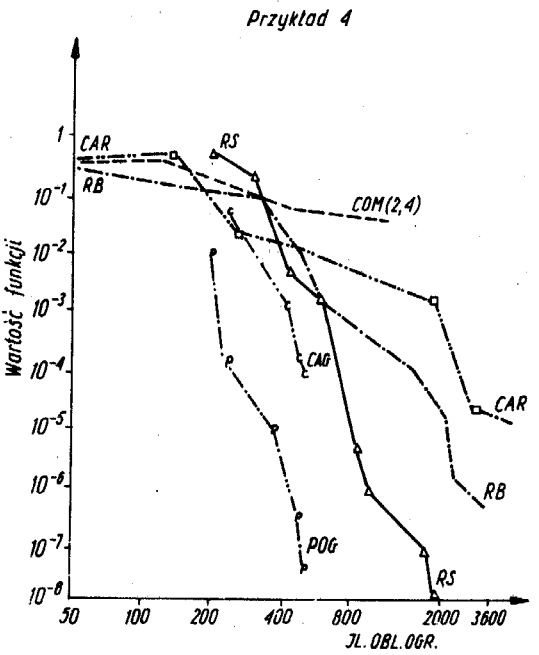
Rys. 49


Przykład 3
COM1
RS
RB
CAR
COM1
CAG
POG
RS
RB
CAR
Wartość funkcji
1
10⁻¹
10⁻²
10⁻³
10⁻⁴
10⁻⁵
10⁻⁶
10⁻⁷
10⁻⁸
64
100
200
400
800
2000
3200
4000
IL.OBL.OGR.

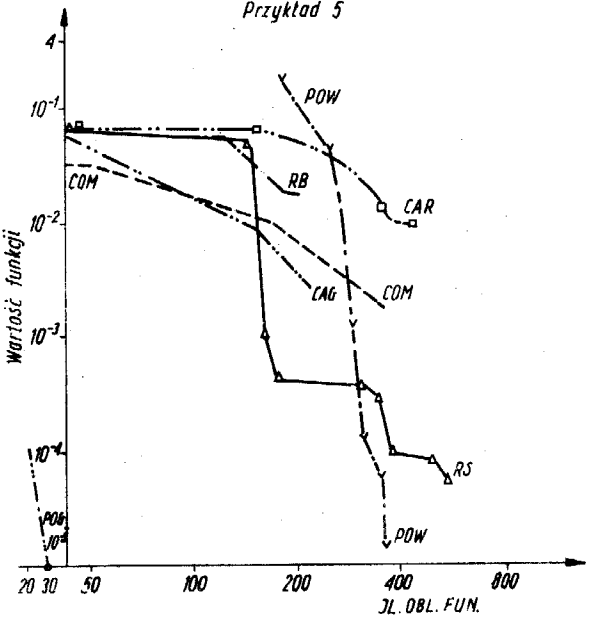
Rys. 50


Przykład 4
Wartość funkcji
1
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
RS
RB
CAR
COM(.2,.4)
CAG
P
POW
CAR
COM(.2,.6)
RB
POG
RS
50
100
200
400
800
IL. OBL. FUN.

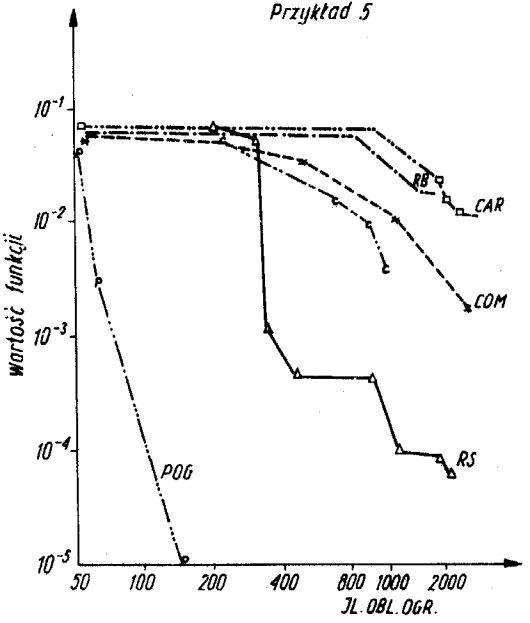
Rys. 51


Przykład 4
Wartość funkcji
1
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
CAR
RS
RB
COM(2,4)
CAG
P
CAR
RB
POG
RS
50
100
200
400
800
2000
3600
IL. OBL. OGR.

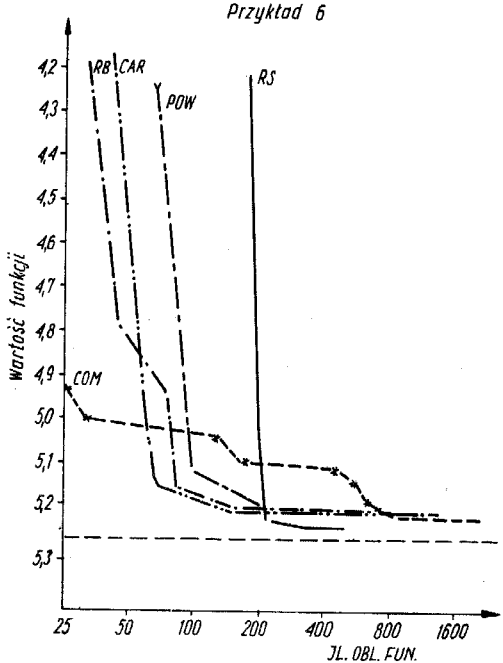
Rys. 52


Przykład 5
Wartość funkcji
4
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
POW
RB
CAR
COM
CAG
COM
RS
POG
POW
20 30 50 100 200 400 800
JL. OBL. FUN.

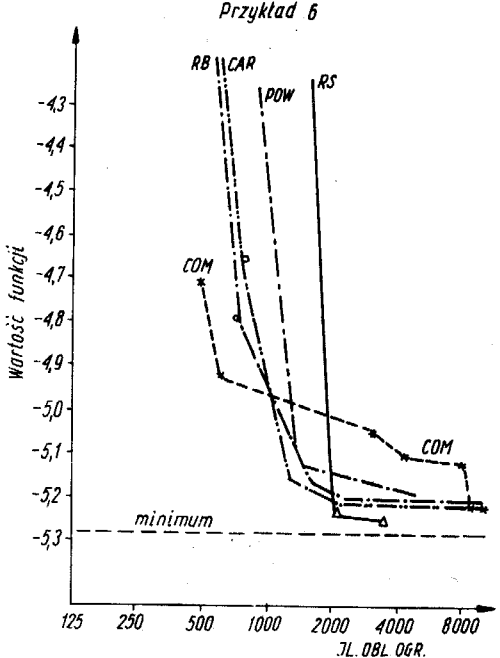
Rys. 53


Przykład 5
Wartość funkcji
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
RB
CAR
COM
POG
RS
50 100 200 400 800 1000 2000
JL. OBL. OGR.

Rys. 54


Przykład 6
RB
CAR
POW
RS
COM
Wartość funkcji
4,2
4,3
4,4
4,5
4,6
4,7
4,8
4,9
5,0
5,1
5,2
5,3
25
50
100
200
400
800
1600
IL. OBL. FUN.

Rys. 55


Przykład 6
RB
CAR
POW
RS
COM
COM
minimum
Wartość funkcji
-4,3
-4,4
-4,5
-4,6
-4,7
-4,8
-4,9
-5,0
-5,1
-5,2
-5,3
125
250
500
1000
2000
4000
8000
IL. OBL. OGR.

Rys. 56

6.5.4. Wnioski

Z uzyskanych rezultatów wynika, że najbardziej efektywną metodą Poszukiwania Ekstremum z Ograniczeniami PEZOG jest zmodyfikowana metoda Powella wykorzystująca metodę gradientu sprzężonego przy wyznaczaniu ekstremum bez ograniczeń. Przewaga tej metody staje się szczególnie widoczna w przypadkach gdy szukane ekstremum znajduje się w ostrzu (przykłady 2 i 5). Dodatkową właściwością wyróżniającą metodę Powella od reszty rozpatrywanych metod jest jej "oscylacyjny" charakter zbliżania się do ekstremum. Właściwość ta została przedstawiona na rys. 47 i 48, natomiast na pozostałych wykresach nie została ona pokazana, gdyż naniesiono na nich jedynie bezwzględne wartości funkcji celu.

W metodach PEZOG posługujących się metodami optymalizacji bez ograniczeń istotny wpływ na ich efektywność ma dobór właściwej metody Poszukiwania Ekstremum Bez Ograniczeń PEBOG. Wpływ ten został zbadany na przykładach metod Powella i Carrolla, które połączono z metodami PEBOG w następujących konfiguracjach: z metodą gradientu sprzężonego - metody POG i CAG, z metodą Powella I - POW, z metodą Rosenbrocka - CAR oraz z metodą Zangwilla - CAZ. Jak wykazały przeprowadzone obliczenia procedury POG i CAG okazały się kilkakrotnie szybsze od pozostałych procedur.

W ten sposób została potwierdzona teza, że im efektywniejszą metodę PEBOG stosuje się do optymalizacji zmodyfikowanej funkcji celu, tym szybciej zbieżna jest sama metoda PEZOG. Wydaje się więc, że celowe byłoby zastosowanie w metodzie Powella metody Davidona zamiast metody gradientu sprzężonego, gdyż jak wiadomo metoda Davidona uważana jest obecnie za jedną z najsilniejszych metod PEBOG. Dla metody Carrolla tego rodzaju próba już została wykonana przez Fletchera i McCanna [19], którzy tę nową metodę nazwali "Acceleration Techniques". Wymagało to jednak znacznej modyfikacji zarówno metody Davidona jak i samej metody Carrolla, w rezultacie której uzyskano bardzo szybko zbieżną metodę. Dla ostatecznego więc porównania metody Powella z metodą Carrolla należałoby zrealizować nowe wersje tych metod przy użyciu algorytmu Davidona, a następnie dokonać odpowiednich obliczeń. Zwróćmy jednak uwagę na dosyć istotną różnicę występującą pomiędzy tymi dwoma metodami. Mianowicie metoda Carrolla w trakcie wykonywania procedury nie narusza zbioru ograniczeń, co oznacza, że bieżący punkt $\underline{x}$ zawsze pozostaje w obszarze dopuszczalnym. Natomiast metoda Powella i wszystkie pozostałe rozpatrywane przez nas metody nie posiadają tej zalety, a więc naruszają one zbiór ograniczeń przy przesuwaniu się w kierunku ekstremum. W wielu praktycznych zastosowaniach przy sterowaniu "on-line" wymagane jest ścisłe przestrzeganie obszaru dopuszczal-

nego, a tym samym zostają wyeliminowane metody, które tego warunku nie spełniają.

Metoda Carrolla w wersji CAG wykazuje bardzo dobrą szybkość zbieżności prawie we wszystkich badanych przykładach, jednakże zbieżność ta wyraźnie maleje przy zwiększaniu dokładności obliczeń. Fakt ten można tłumaczyć tym, że przy dużych dokładnościach tworzone w metodzie powierzchnie stają się coraz bardziej strome, a przez to zmniejszona zostaje efektywność poszukiwania ekstremum bez ograniczeń. W podobny sposób zachowuje się metoda Rosenbrocka odznaczająca się dobrą zbieżnością w początkowej fazie działania, lecz z chwilą wejścia w strefę ograniczeń zbieżność ta znacznie maleje i to tym bardziej im więcej strefa ograniczeń zostaje zwężona przy zwiększaniu dokładności.

Odmienne właściwości od dwóch ostatnio omawianych metod reprezentuje metoda Rosena. Metoda ta okazała się bardzo efektywna dla dużych dokładności obliczeń, natomiast w początkowej fazie jest wyraźnie gorsza tak od metody Carrolla jak i Rosenbrocka, nie mówiąc już o metodzie Powella. Tego rodzaju zachowanie się metody Rosena wypływa stąd, że w pierwszym okresie działania zużywany jest dość duży nakład obliczeń na czynności pomocnicze takie jak: obliczenie macierzy odwrotnej V_q, macierzy projekcyjnej itp. Ponadto nakład ten szybko wzrasta jeśli ograniczenia aktywne są silnie nieliniowe, a więc częściej należy dokonywać projekcji gradientu oraz uaktualniać odpowiednie macierze. Jednakże w bliskim otoczeniu minimum, kiedy zarówno funkcje celu jak i ograniczeń są dobrze aproksymowane formą kwadratową, szybkość zbieżności metody Rosena staje się bardzo duża. Stąd też, wydaje się interesującą koncepcja połączenia dwóch metod: Rosena oraz Rosenbrocka w jedną całość i tym sposobem wyeliminowanie ich niekorzystnych właściwości.

Metoda Complex w porównaniu z poprzednio rozpatrywanymi metodami okazała się najmniej efektywną metodą, przy czym ze wzrostem wymiarowości jej szybkość zbieżności wyraźnie się pogarsza. Posiada ona jednak dość istotną zaletę charakteryzującą się tym, że w rezultacie przekształcania complexu w całym obszarze dopuszczalnym następuje dokładne jego przeszukiwanie, a tym samym istnieje o wiele większe prawdopodobieństwo trafienia w ekstremum globalne niż w lokalne. Taka właśnie sytuacja zaistniała w przykładzie 2 (rys. 47 i 48), gdzie wszystkie metody (z wyjątkiem POW) podążyły w kierunku minimum lokalnego, natomiast metoda Complex wykryła minimum globalne. Wyznaczenie przez metodę POW tego samego minimum globalnego nastąpiło jedynie przypadkowo na skutek znalezienia się w jego otoczeniu po minimalizacji funkcji wzdłuż pierwszego kierunku założonej bazy. W celu wykorzystania wspomnianej zalety, w pewnych przypadkach może okazać się opłacalne zastosowanie w początkowym okresie opty-

malizacji metody Complex dla doboru punktu startowego dla innych metod.

Analizując wpływ rodzaju ograniczeń oraz położenia punktu ekstremalnego na efektywność metod należy stwierdzić, że na niektóre metody wpływ ten jest bardzo silny, zaś na inne o wiele mniejszy. Tak więc, dla metod Carrolla oraz Rosenbrocka najkorzystniejszy jest przypadek gdy ekstremum leży na pojedynczym ograniczeniu oraz kiedy ograniczenie to jest wklęsłe lub liniowe. Natomiast w razie wystąpienia punktu ekstremalnego na przecięciu się ograniczeń tworzących ostrze, szybkość zbieżności tych metod wyraźnie maleje. W odmienny sposób na położenie ekstremum reaguje metoda Complex, która najefektywniejsza okazała się dla minimum znajdującego się w ostrzu. Podobną właściwość posiada również metoda Powella, lecz tak jak to już wspomniano na wstępie, przewyższa ona znacznie pozostałe metody w każdej badanej sytuacji. Najmniej czułą na zmiany położenia ekstremum oraz rodzaj ograniczeń jest metoda Rosena, która we wszystkich przypadkach zachowała swój charakter zbieżności. Jednakże, tego rodzaju ocena nie oddaje w pełni możliwości tej metody, gdyż jak wiadomo jest ona najskuteczniejsza dla zadań optymalizacji z ograniczeniami liniowymi. Brak wyraźnego wzrostu szybkości zbieżności metody Rosena w przykładzie 3 (ograniczenia liniowe) wynika stąd, że obliczeń dokonano przy pomocy procedury specjalnie przystosowanej do rozwiązywania ogólnego Zadania Programowania Nieliniowego ZPN typu A, a nie tylko do zadań z ograniczeniami liniowymi. Warto jednak wspomnieć, że istnieje dość liczna grupa metod optymalizacji, które służą jedynie do tego celu. Do grupy tej zaliczona jest również zmodyfikowana wersja metody Rosena [46], jednakże za najsilniejszą uważana jest obecnie metoda Goldfarba i Lapidusa [24], której koncepcja została oparta na algorytmie Davidona. Ze względu na ograniczony zakres ich zastosowań nie zostały one rozpatrzone w niniejszej pracy.

Istotnym czynnikiem wpływającym na efektywność metod stosujących funkcję kary, jest właściwy dobór kryterium zakończenia działania procedur poszukiwania ekstremum bez ograniczeń PEBOG. Dotyczy to szczególnie metod Powella oraz Carrolla, które w trakcie wykonywania korzystają z tych właśnie procedur. Z przeprowadzonych badań nad tymi dwoma metodami wynikają następujące wnioski.

W metodzie Powella POG przyjęcie mniejszej dokładności obliczeń dla procedury gradientu sprzężonego powoduje wzrost szybkości zbieżności w początkowej fazie, natomiast z chwilą powiększenia dokładności określania minimum warunkowego nakład obliczeń zaczyna bardzo szybko wzastać. Przypadek ten został przedstawiony na rys. 45, krzywa POGD. Odmienny zupełnie przebieg

posiada krzywa szybkości zbieżności (rys. 45, krzywa POG), gdy zwiększona zostaje dokładność obliczeń procedury gradientu sprzężonego. W pierwszym okresie następuje wówczas zmniejszenie szybkości zbieżności metody POG lecz przy powiększaniu dokładności obliczeń szukanego minimum szybkość ta bardzo wzrasta.

W metodzie Carrolla dobór kryterium zakończenia działania procedury PEBOG odgrywa również duże znaczenie. Najlepsze rezultaty uzyskano wtedy, gdy w trakcie przebiegu algorytmu Carrolla wprowadzono zmienną dokładność obliczeń w stosowanej metodzie PEBOG, w zależności od "stromizny" generowanych powierzchni. Tak więc, jeśli w początkowej fazie przebiegu algorytmu utworzone powierzchnie posiadają łagodny charakter wówczas wystarczy przyjmować niewielką dokładność obliczeń dla wyznaczania kolejnych minimów bezwarunkowych. Z chwilą jednak znalezienia się w pobliżu szukanego ekstremum warunkowego, kiedy generowane powierzchnie stają się coraz bardziej ostre, wtedy także musi ulec zaostrzeniu kryterium zbieżności procedury PEBOG.

W pozostałych trzech z rozpatrywanych metod, a więc Rosenbrocka, Rosena oraz Complex wzrost ich efektywności można uzyskać przez zmianę następujących czynników. W metodzie Rosenbrocka przez złagodzenie kryterium według którego zostaje dokonany obrót współrzędnych, w metodzie Rosena przez ulepszenie interpolacji stosowanej do określania położenia punktu na ograniczeniach aktywnych oraz w metodzie Complex przez zmianę parametrów α i k. Zwróćmy przy tym uwagę, że w metodzie Complex, źle dobrana wartość parametru może doprowadzić do niezbieżności metody. Przypadek taki przedstawiono na rys. 49.

Wpływ wymiarowości ZPN na efektywność metod zbadano tylko w ograniczonym zakresie na 5-wymiarowym przykładzie, przy czym porównania nie przeprowadzono dla wszystkich metod. Z otrzymanych krzywych wynika (rys. 55 i 56), że ze wzrostem wymiarowości polepszyła się szybkość zbieżności takich metod jak RB, CAR i POW, natomiast pogorszyła się metoda Complex. Należy więc przypuszczać, że podobnie jak POW i CAR zachowywałyby się i metody POG oraz CAG, a więc uzyskano by dla nich jeszcze lepsze rezultaty.

Reasumując niniejsze rozważania na zakończenie warto podkreślić, że obecnie nie istnieje uniwersalna metoda poszukiwania ekstremum z ograniczeniami, która mogłaby zadowolić wszystkie wymagania stawiane przez użytkowników. Wyboru metody można więc dokonać dopiero gdy zostanie sformułowane konkretne zadanie oraz cel jakiemu ma ono służyć. Oczywiście, że przy obliczeniach optymalizacyjnych przeprowadzonych "off-line" jedynymi ograniczeniami nakładanymi na daną metodę są ograniczenia wynikające z wielkości i wyposażenia komputera, który jest do dyspozycji. Sytuacja ta jednak ulega radykalnej zmianie gdy te same ob-

liczenia mają być dokonywane "on-line". Wówczas oprócz wspomnianych ograniczeń muszą być brane pod uwagę dodatkowe względy takie na przykład jak: szybkość zbieżności, możliwość naruszenia ograniczeń w trakcie obliczeń, niezawodność metody, wielowejściowość itp. W takiej właśnie sytuacji odpowiedź na pytanie, która ze znanych metod optymalizacji powinna być zastosowana przestaje już być sprawą prostą i wymaga głębokiej ich analizy. Stąd też, przedstawiona w niniejszej pracy ocena metod PEZOG może być pomocna w rozstrzygnięciu kwestii słusznego ich wyboru.

LITERATURA DO CZĘŚCI I

[1] Arrow K.J. and Uzawa H. "Studies in Linear and Nonlinear Programming", Stanford University Press, Stanford, California, 1958.

[2] Barnes J.G.P. "An algorithm for solving non-linear equations based an the second method", The Computer Journal, Vol. 8, p.66, 1965.

[3] Beale E.M.L. "On Quadratic Programming" Naval Research Logistics Quarterly, 6, 1959.

[4] Box M.J. "A comparison of several current optimization methods and the use of transformations in constrained problems", The Computer Journal Vol. 9, p.67, 1966.

[5] Box M.J. "A new method of constrained optimization and a comparison with other methods". The Computer Journal Vol. 8, p.42, 1965.

[6] Carroll C.W. "The Created Response Surface Technique for Optimizing Nonlinear Restrained Systems", Operations Research, Vol.9, p.169, 1961.

[7] Charnes A. and Cooper W., "Management Models and industrial Applications of Linear Programming" New York: Wiley 1961.

[8] Courant R. "Variational Methods for the Solution of Problems of Equilibrium and Vibrations", Bull. Am. Math. Soc., 48, 1943.

[9] Davies D. "The Use of Davidon a Method in Nonlinear Programming", IC| Management Services Report MSDH/68/110, 1968.

[10] Davies D. and Swann W.H. "Review of Constrained Optimization", Presented at the conference on "Optimization" at Keele University, March, 1968.

[11] Davidon W.C. "Variable Metric Method for Minimization" A.E.C. Research and Development Report., ANL--5990, 1959.

[12] Davidon W.C. "Variance algorithm for minimization", The Computer Journal, Vol. 10, p. 406, 1968.

[13] Dantzig G.B. "Linear Programming and Extensions" Princeton, 1963.

[14] Findeisen W. "Note on Optimal - Satisfactory Control", IEEE Trans. on Automatic Control, Vol. AC-12, No 5 (Oct. 1967), str. 612-613.

[15] Fiacco A.V. and McCormick G.P. "The sequential unconstrained minimization technique for nonlinear programming, a primal-dual method", Management Science, Vol. 10, p. 360, 1964.

[16] Fiacco A.V. and McCormick G.P. "Nonlinear Programming: sequential unconstrained minimization techniques", J. Wiley, 1968.

[17] Fiacco A.V. and McCormick G.P. "Extensions of SUMT for Nonlinear Programming" Equality Constraints and Extrapolation", Management Science, Vol. 12, p. 816, 1966.

[18] Fletcher R. "Optimization", Academic Press, 1969.

[19] Fletcher R. "Acceleration Techniques for Nonlinear Programming", Presented at the conference on "Optimization" at Keele University, March, 1968.

[20] Fletcher R. "Function minimization without evaluating derivatives - a review", The Computer Journal, Vol. 8, p. 33, 1965.

[21] Fletcher R. and Powell M.J.D. "A rapidly convergent descent method for minimization, The Computer Journal, Vol. 6, p. 163, 1963.

[22] Fletcher R. and Reeves C.M. "Fundation minimization by conjugate gradients", The Computer Journal Vol. 7, p. 149, 1964.

[23] Frank M. and Wolfe P. "An Algorithm for Quadratic Programming", Naval Research Logistics Quarterty, 3, 1956.

[24] Goldfarb D. and Lapidus L.: Conjugate gradient method for nonlinear programming problems with linear constraints", Ind. Engng. Chem. Fundam., Vol. 1, p. 142, 1968.

[25] Hadley G. "Linear Programming", Addison-Wesley, 1962,

[26] Hadley G. "Nonlinear and Dynamic Programming", Addison-Wesley, 1964.

[27] Hestenes M.R. and Stiefel E., "Methods of conjugate gradients for solving linear systems" J. Res. N.R.S., Vol. 49, p. 409, 1952.

[28] Himsworth F.R., Spendley W. and Hext G.R. "The sequential application of simplex desings in optimization an evolutionary operation" Technometrics, Vol. 4, p. 441, 1962.

[29] Hildreth C. "A Quadratic Programming Procedure" Naval Research Logistics Quarterly, 14, 1957.

[30] Houthakker H.S. "The Capacity Method of Quadratic Programming" Econometrica, 28, 1960.

[31] Householder A.S. "Principles of Numerical Analysis", McGraw-Hill, 1953.

[32] Huard P. "Resolution of mathematical programming with non-linear constraints by the method of centres", In: Abadie J., ed., Non-linear programming, North Holland Publishing, p. 207, 1967.

[33] Klingman W.R. and Himmelblau D.M. "Nonlinear programming with the aid of a multiple-gradient summation technique", J. Ass. Coput. Mach. Vol. 11, 1964.

[34] Kowalik J. and Osborne M.R. "Methods for Unconstrained Optimization Problems", American Elsevier Publishing Company, 1968.

[35] Kuhn H.W. and A.W. Tucker "Nonlinear Programming" Proceedings Second Berkeley Symposium an Mathematical Statistics and Probability, 1951, pp. 481-492.

[36] Kulikowski R. Sterowanie w wielkich systemach, WNT 1970.

[37] Lasdon L.S., Mitter S.K. and Waren A.D. "The conjugate Gradient Method for Opital Control Problems" IEE Transactions an Automatic Control, Vol. AC-12, p. 132, 1967.

[38] Moser J. and Courant R. "Calculus of Variations and Supplementary Notes and Exercises", Mimeographed Notes, New York University, 1957.

[39] Nelder J.A. and Mead R. "A simplex method for function minimization", The Computer Journal, Vol. 7, p. 308, 1965.

[40] Pearson J.D. "Variable metric methods of minimization" The Computer Journal, vol. 11, p. 171, 1969.

[41] Powell M.J.D. "An iterative method for finding stacionary values of a function of several variables". The Computer Journal, Vol. 5, p. 147, 1962.

[42] Powell M.J.D. "An efficient method of finding the minimum of a function of several variables without calculating derivatives", The Computer Journal, Vol. 7, p. 155, 1964.

[43] Powell M.J.D. "A method for minimizing a sum of squares of non-linear functions without calculating derivatives", The Computer Journal, Vol. 7, p. 303, 1965.

[44] Powell M.J.D. "On the calculation of orthogonal vectors The Computer Journal, p. 302, 1968.

[45] Powell M.J.D. "A method for non-linear constraints in minimization problems". AERE Harwell Report TP 310, 1967.

[46] Rosen J.B. "The gradient projection method for non-linear programming. Part. I. Linear constraints", J. Soc. Ind. Appl. Math. 8, 1960.

[47] Rosen J.B. "The gradient projection method for non-linear programming. Part. II. Non-linear constraints". J. Soc. Ind. Appl. Math. 9, 1961.

[48] Rosenbrock H.H. and Storey C. "Computational Techniques for Chemical Engineers", Pergamon Press, 1966.

[49] Rosenbrock H.H. "An Automatic Method for finding the Greatest or Value of a Function", The Computer Journal, Vol. 3, p. 175, 1960.

[50] Schmit L.A. and Fox R.L. "Advances in the Integrated Approach to structural Synthesis", A.I.A.A. Sixth Annual Structure and Materials Conference, Palms Springs, California, 1965.

[51] Smith C.S. "The automatic computation of maximum likelihood estimates", N.C.B. Scientific Dept. Report S.C. 846/MR/40, 1962.

[52] Swann W.H. "Report on the development of new direct searching method of Optimization", I.C.I. Ltd, Central Instrument Laboratory Research Note 64/3, 1964.

[53] Szymanowski J. "Algorytmy obliczeniowe dla optymalizacji statycznej", Systemy Automatyki Kompleksowej. Ossolineum, 1969.

[54] Szymanowski J. "Optymalizacja Statyczna". Skrypt OPT, Warszawa, 1970.

[55] Szymanowski J. i inni "Biblioteka Programów Optymalizacji Statycznej", Politechnika Warszawska, 1970.

[56] Szymanowski J. i Brzostek J. "Porównanie Bezgradientowych Metod Optymalizacji Statycznej", Arch. Aut. i Telemech., Nr 1, 1971.

[57] Szymanowski J. i Jastrzębski S. "Porównanie Gradientowych Metod Optymalizacji Statycznej", Arch. Aut. i Telemech. Nr 2, 1971.

[58] Szymanowski J. "Przegląd Metod Poszukiwania Ekstremum z Ograniczeniami", Arch. Aut. i Telemech., Nr 2, 1971.

[59] Szymanowski J. "Porównanie Metod Poszukiwania Ekstremum z Ograniczeniami", Krajowa Konferencja Automatyki, Gdańsk 1971.

[60] Wierzbicki A. "Metoda z przesuwaniem funkcji kary", Krajowa Konferencja Automatyki, Gdańsk 1971.

[61] Wolfe P. "The Simplex Method for Quadratic Programming", Econometrica, 27, 1959.

[62] Zangwill W.J. "Minimizing a function without calculating derivatives", The Computer Journals, Vol. 10, p. 293, 1967.

Część II

OPTYMALIZACJA DYNAMICZNA

7. Wiadomości wstępne

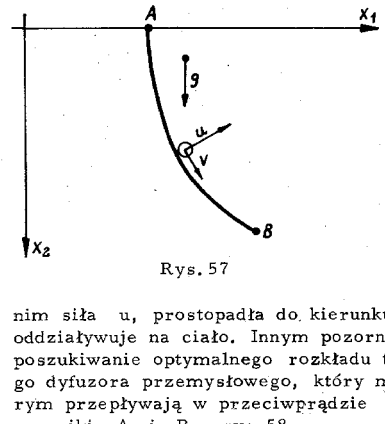
Pod pojęciem optymalizacja rozumiemy ogólnie poszukiwanie rozwiązań najlepszych, optymalnych z określonego punktu widzenia. Rozwiązania te są podstawą dla decyzji, podejmowanych przez człowieka lub zastępujące go urządzenia automatyczne. Oddziaływanie tych decyzji na określony proces fizyczny (lub chemiczny, czy nawet ekonomiczny) zwane jest sterowaniem optymalnym, jeśli podjęte decyzje są optymalne. Stąd też istnieje ścisły związek pomiędzy teorią optymalizacji a teorią sterowania optymalnego i będziemy traktować je łącznie, chociaż można wyobrazić sobie oddzielenie tych pojęć (na przykład optymalizacja decyzji konstrukcyjnych podejmowanych przez projektanta, nie jest bezpośrednio związana ze sterowaniem; natomiast ich realizacja i ewentualne ulepszanie na podstawie zaobserwowanych niedokładności stanowi niewątpliwie przykład sterowania, rozumianego w szerokim sensie).

Zadania optymalizacji dzielimy, jak wiadomo, na dwie podstawowe klasy: optymalizację statyczną, sprowadzającą się do poszukiwania ekstremum (maksimum lub minimum) funkcji oraz optymalizację dynamiczną, sprowadzającą się do poszukiwania ekstremum funkcjonału. Typowe zadanie optymalizacji dynamicznej polega na poszukiwaniu takiego sposobu zmian decyzji w danym przedziale czasu, który zapewni ekstremum pewnego wskaźnika jakości zależnego od przebiegu zmian tej decyzji na całym przedziale. Wskaźnik jakości jest więc funkcjonałem tej decyzji, określonym na danym przedziale czasu, co tłumaczy przymiotnik optymalizacja "dynamiczna".

Prostym przykładem takiego zadania jest problem przestawiania wózka transportowego, którym należy ruszyć, przejechać określony odcinek i zatrzymać w pewnym miejscu. Jak należy manewrować silnikiem wózka, aby przejechać dany odcinek w najkrótszym czasie? Wskaźnikiem jakości jest tu czas od momentu

ruszenia do momentu zatrzymania, zaś decyzją czy też sterowaniem - sposób włączania silnika. Bardziej złożonym przykładem jest problem sterowania rakiety Ziemia-Mars. Należy tu dobrać moment startu i trajektorię rakiety tak, aby przelot wymagał minimum paliwa. Wybór trajektorii jest tu oczywiście związany z wyborem sterowania, czyli sposobem manewrowania silnikami rakiety. Przykłady optymalizacji dynamicznej występują też przy sterowaniu procesów przemysłowych. Wyobraźmy sobie reaktor chemiczny, który napełnia się reagentami, a następnie ogrzewa do uzyskania wymaganej temperatury. Wzrost temperatury powoduje przyspieszenie procesów chemicznych i wydzielania się ciepła reakcji; jednakże nie można grzać reaktora zbyt intensywnie, bo sprawność urządzeń grzejnych maleje ze wzrostem mocy przez nie przenoszonej i straty energetyczne rosną. Jak należy zmieniać w czasie przebieg mocy grzejnej, żeby na ogrzanie reaktora w danym czasie zużyć minimum energii?

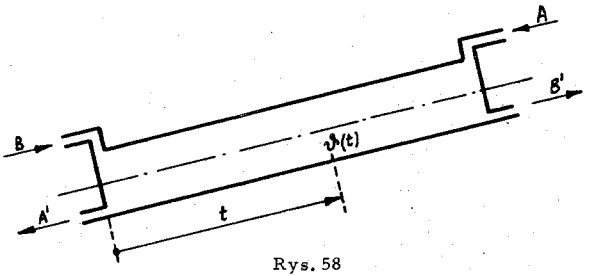
Jako zadania optymalizacji dynamicznej mogą być sformułowania także inne, pozornie odległe problemy. Na przykład w klasycznym zadaniu rachunku wariacyjnego - zagadnieniu brachistochrony - należy tak dobrać tor ciała, ślizgającego się bez tarcia pod wpływem siły ciężkości - rys. 57, by czas, w którym przebywa ono drogę pomiędzy punktami A i B, był najmniejszy. Na pierwszy rzut oka trudno tu wyodrębnić zmienną decyzyjną czyli zmienne w czasie sterowanie, oddziaływujące na ruch ciała. Może być jednak nim siła u, prostopadła do kierunku ruchu ciała, z którą podłoże oddziaływuje na ciało. Innym pozornie odległym problemem jest poszukiwanie optymalnego rozkładu temperatur wzdłuż ogrzewanego dyfuzora przemysłowego, który ma postać długiej rury i w którym przepływają w przeciwprądzie dwa dyfundujące między sobą czynniki A i B - rys. 58.



Rys. 57

Szybkość dyfuzji zależy od koncentracji tych czynników oraz od temperatury, przy czym dla każdej koncentracji istnieje pewna optymalna temperatura, a koncentracja w określonym punkcie dyfuzora zależy od szybkości dyfuzji w innych punktach. Jak dobrać rozkład temperatur, aby uzyskać maksymalną koncentrację w czynniku A wypływającym z dyfuzora? W swej istocie fizycznej jest to problem statyczny, gdyż pytamy o ustalony rozkład temperatur

zakładając, że przepływ czynników przez dyfuzor trwa już dostatecznie długo i ma charakter ustalony. Zmienną, od której zależy tu decyzja (temperatura), jest odległość rozpatrywanego punktu od



Rys. 58

krańca dyfuzora. Z matematycznego punktu widzenia możemy jednak tę zmienną utożsamić z czasem, oznaczając ją przez t, i potraktować zadanie jako problem optymalizacji dynamicznej.

Powyższe przykłady dotyczyły problemu określenia optymalnego sterowania jako funkcji czasu, a więc nie były związane z wyborem sposobu realizacji tego sterowania i jego uzależnienia od bieżących informacji o przebiegu sterowania procesu. Jeśli w pierwszym z rozpatrywanych przykładów założymy, że możemy mierzyć za pomocą odpowiednich urządzeń bieżące położenie i prędkość wózka, oraz zapytamy, jak uzależnić optymalne sterowanie wózka od wyników tych pomiarów, to mamy do czynienia z bardziej złożonym zadaniem - problemem syntezy układu zamkniętego sterowania optymalnego. Podobne zadania można sformułować także dla innych przytoczonych tu przykładów.

Mając do rozwiązania zadanie optymalizacji dynamicznej względnie sterowania optymalnego, należy postępować według naturalnej i wypróbowanej metodyki rozwiązywania takich zadań.

Po pierwsze, należy ustalić dostatecznie dokładny, a jednocześnie nie nadmiernie skomplikowany model matematyczny rozpatrywanego procesu fizycznego (chemicznego, ekonomicznego, czy innego). Sposoby ustalania modelu nie będą rozpatrywane w tej części skryptu, chociaż podamy przykłady arbitralnego wyboru modeli. Będziemy tu też dla uproszczenia zakładać, że chociaż model mógł być ustalony w oparciu o badania statystyczne, to jednak ma on charakter deterministyczny, to znaczy nie występują w tym modelu zmienne i funkcje losowe. Ponadto nie będziemy

tu rozpatrywać modeli o postaci równań różniczkowych cząstkowych (modeli o stałych rozłożonych).

Po drugie należy sprawdzić, czy problem optymalizacji dynamicznej (polegający na wyznaczeniu zależności decyzji od czasu, czyli - innymi słowy - sterowania optymalnego w układzie otwartym) ma rozwiązanie, dające się wyrazić w postaci analitycznej. W tym punkcie wyłania się szereg problemów szczegółowych. Po pierwsze należałoby udowodnić, że rozpatrywany problem posiada w ogóle rozwiązanie, wyrażające się określoną funkcją czasu; ten bardzo istotny i zazwyczaj trudny problem matematyczny nie ma na ogół znaczenia dla poprawnie sformułowanych problemów fizycznych i nie będzie tu dokładniej rozpatrywany. Po drugie należy się zdecydować na wybór jednej z wielu matematycznych metod optymalizacji, które będą tu przedstawione; ponieważ jednak są one w dużej mierze równoważne, zaleca się stosowanie metody o najprostszej notacji i wypróbowanym sposobie stosowania - metodą taką, zdaniem autora, jest zasada maksimum. Po trzecie, klasa zadań optymalizacji dla których istnieją pełne rozwiązania analityczne jest stosunkowo wąska; jednakże należy przeprowadzać jak najpełniejszą analizę każdego zadania i starać się uzyskiwać choćby częściowe czy przybliżone rezultaty analityczne, które ułatwią późniejsze ewentualne obliczenia numeryczne oraz ich interpretację.

Jeśli nie można wyznaczyć rozwiązania na drodze analitycznej, to należy je obliczyć numerycznie za pomocą maszyny cyfrowej. Istnieje wiele metod obliczeniowych optymalizacji dynamicznej, z których podstawowe będą tu dokładniej omówione. Należy się zdecydować na jedną z nich zależnie od postaci zadania; istotną rolę odgrywają tu też takie czynniki, jak przewidywany nakład obliczeń dla uzyskania rozwiązania czy wymagana dokładność rozwiązania (które uzyskuje się w postaci przybliżonej).

Po wyznaczeniu sterowania optymalnego w układzie otwartym (jako funkcji czasu) należy zdecydować jak będzie to sterowanie realizowane w zastosowaniu do konkretnego procesu. Najpierw więc należy sprawdzić czy możliwe jest określenie analitycznej postaci sterowania optymalnego w układzie zamkniętym (nie jako funkcji czasu, lecz w zależności od wyników pomiarów stanu procesu) o podstawowej strukturze, czyli przeprowadzenie syntezy układu zamkniętego sterowania optymalnego. Odpowiedź na to pytanie jest pozytywna tylko dla bardzo wąskiej klasy zadań. Jeśli odpowiedź jest negatywna, to zawsze można zrealizować zamknięty układ sterowania optymalnego, stosując maszynę matematyczną do bieżącego wyznaczania sterowań optymalnych w oparciu o wyniki pomiarów stanu procesu. Jednakże nie można zwykle z góry przewidzieć, że układ zamknięty sterowania będzie zdecydowanie lub w ogóle lepszy od układu otwartego i że zastosowanie maszyny matematycznej jest opłacalne. Co więcej, istnieje szereg wariantów struktury

układu zamkniętego. Należy więc dokonać wyboru struktury układu sterowania. Bierze się tu pod uwagę kilka czynników. Po pierwsze należy ocenić wrażliwość rozważanych struktur, czyli odchylenia od optymalności, jakie mogą wyniknąć w związku z niedokładnością założonego modelu matematycznego rzeczywistego procesu; odchylenia te silnie zależą od wyboru struktury układu. Po drugie, należy ocenić nakład obliczeń niezbędnych do bieżącego wyznaczania sterowania optymalnego w danej strukturze układu; musi on się mieścić w granicach możliwości maszyny matematycznej. Zbyt duży nakład obliczeń może wywołać konieczność uproszczenia modelu matematycznego procesu, a tym samym - zmniejszenie dokładności obliczania sterowania i jakości sterowania. Jeśli uproszczenie modelu jest niedopuszczalne, to nakład obliczeń można zmniejszyć, bądź rozłożyć pomiędzy kilka maszyn przez zastosowanie wielopoziomowej hierarchicznej struktury układu sterowania. Trzecim i zazwyczaj decydującym czynnikiem są względy ekonomiczne - koszty maszyny matematycznej, urządzeń pomiarowych i innych urządzeń układu sterowania, oraz względy czysto techniczne - na przykład możliwość realizacji pewnych pomiarów.

Problemy związane z syntezą układu zamkniętego, wyborem struktury układu, analizą wrażliwości itp. są bardzo obszerne i z braku miejsca nie będą dokładniej omawiane.

8. Metody analityczne optymalizacji dynamicznej

8.1. Sformułowanie problemu i pojęcia podstawowe

Dany jest model dynamiki procesu w postaci jego równań stanu, które w podstawowym przypadku przyjmują postać układu równań różniczkowych zwyczajnych

$$\dot{\underline{x}} = \underline{f}(\underline{x},\ \underline{u},\ t,\ \underline{a}); \qquad \dim \underline{x} = \dim \underline{f} = n;$$
$$\dim \underline{u} = m; \qquad \dim \underline{a} = r; \tag{319}$$

gdzie $\underline{x} = [x_1, \ldots, x_n]'$ jest n-wymiarowym wektorem *) zmiennych, wynikających z decyzji, zwanych współrzędnymi stanu lub krótko stanem procesu,

*) Wszystkie wektory w tej części skryptu - np. $\underline{x}$ - są uważane za wektory kolumnowe, z wyjątkiem wyraźnie oznaczonych wektorów transponowanych - np. $\underline{x}'$.

$\underline{u} = [u_1, \ldots, u_m]'$ jest m-wymiarowym wektorem zmiennych decyzyjnych, zwanych sterowaniem,

$\underline{a} = [a_1, \ldots, a_r]'$ jest wektorem stałych parametrów, których wartości wynikają z identyfikacji modelu procesu,

$\underline{f}$ = dana funkcja wektorowa.

Oprócz podstawowej postaci (319), równania stanu mogą też przyjmować postać równań różnicowych zwyczajnych (model dynamiki i problem nazywamy wówczas dyskretnym w czasie) bądź równań różniczkowych z opóźnionym argumentem (mówimy o opóźnieniu stanu lub opóźnieniu sterowania, jeśli w prawej stronie równań (319) występują odpowiednio zmienne opóźnione $\underline{x}(t - T_o)$ lub $\underline{u}(t - T_o)$, gdzie T_o - czas opóźnienia).

Dany jest model stanu początkowego procesu. W podstawowym wariancie zadania optymalizacji zakłada się zwykle, że znany jest dokładnie stan początkowy procesu w pewnej chwili początkowej t_o

$$\underline{x}(t_o) = \underline{x}_o. \tag{320}$$

Dla procesów z opóźnieniami niezbędna jest dodatkowo znajomość całego przebiegu $\underline{x}(\tau)$ lub $\underline{u}(\tau)$ dla wszystkich τ z przedziału $t_o - T_o \leqslant \tau < t_o$.

Dany jest model ograniczeń procesu. W podstawowym wariancie zadania zakłada się, że ograniczenia te dotyczą tylko sterowania $u(t)$ i mają postać układu nierówności

$$\underline{g}_1(\underline{u}, \underline{a}) \leqslant 0, \tag{321}$$

gdzie $\underline{g}_1$ - dana funkcja wektorowa.

Ograniczenia sterowania zapisuje się też ogólniej

$$\underline{u}(t) \in \Omega \tag{322}$$

gdzie Ω - określony obszar dopuszczalnych wartości $\underline{u}(t)$ w m-wymiarowej przestrzeni sterowań; obszar ten może zależeć od parametrów $\underline{a}$.

Jeśli ograniczenia dotyczą także stanu $\underline{x}(t)$ i mają postać

$$\underline{g}_2(\underline{x}, \underline{u}, t, \underline{a}) \leqslant \underline{0} \tag{323}$$

gdzie $\underline{g}_2$ - dana funkcja wektorowa, to zadanie optymalizacji komplikuje się znacznie.

Dany jest model celu sterowania w postaci zbioru warunków końcowych dla stanu procesu

$$\underline{g}_k\left(\underline{x}(t_k), t_k, \underline{a}\right) = \underline{0}; \qquad \dim \underline{g}_k = p, \tag{324}$$

które muszą być spełnione przez proces w pewnej chwili końcowej t_k; $\underline{g}_k$ jest tu daną funkcją wektorową p-wymiarową. Najprost-

sza postać warunków końcowych polega na podaniu stanu końcowego

$$\underline{x}(t_k) = \underline{x}_k. \tag{325}$$

Jeśli pewne współrzędne stanu x_i nie występują w warunkach (324) lub (325), to mówimy, że są one swobodne; czas końcowy t_k może być także dany z góry lub swobodny.

Warunki końcowe dla procesu tworzą w $(n+1)$-wymiarowej rozszerzonej przestrzeni stanu i czasu pewien twór geometryczny, zwany rozmaitością końcową i oznaczany przez Γ. Rozmaitość końcowa jest punktem, jeśli dane są zarówno wartości $\underline{x}(t_k)$ jak i t_k; jest prostą, jeśli np. jedna z tych wartości jest swobodna, a pozostałe dane; jest hiperpowierzchnią, jeśli dany jest tylko jeden (skalarny) warunek końcowy o postaci $g_k\left(\underline{x}(t_k), t_k, \underline{a}\right) = 0$, czyli jeśli $\dim g_k = p = 1$. Mówimy, że rozmaitość końcowa jest $(n+1-p)$-wymiarowa, gdzie $p = \dim \underline{g}_k$ - liczba niezależnych warunków końcowych. Punkt końcowy jest więc rozmaitością zerowymiarową, prosta końcowa - jednowymiarową, zaś hiperpowierzchnia -n-wymiarową.

Podkreślamy tu, że ilekroć w zadaniu optymalizacji w warunkach końcowych czy równaniach stanu występuje bezpośrednio czas, to posługujemy się raczej przestrzenią rozszerzoną stanu i czasu zamiast samej przestrzeni stanu. Wektor $\underline{\tilde{x}} = \left\{ \underline{x}, t \right\}$ o n+1 składowych nazywamy stanem rozszerzonym.

Dany jest model wskaźnika jakości sterowania w postaci funkcjonału stanu $\underline{x}$ i sterowania $\underline{u}$. W podstawowym wariancie zadania zakłada się, że jest to funkcjonał o postaci

$$Q\left\{\underline{x},\ \underline{u},\ \underline{a}\right\} = f_k\left(\underline{x}(t_k),\ t_k,\ \underline{a}\right) + \int_{t_0}^{t_k} f_o\left(\underline{x},\ \underline{u},\ t,\ \underline{a}\right)dt, \tag{326}$$

gdzie f_o, f_k - dane funkcje skalarne.

Przy takiej postaci funkcjonału jakości mówimy, że mamy do czynienia z problemem Bolzy; jeśli $f_k \equiv 0$ i wskaźnik jakości ma postać całkową, to mamy do czynienia z problemem Lagrange'a; jeśli $f_o \equiv 0$ i wskaźnik jakości jest funkcją stanu końcowego, to mamy do czynienia z problemem Mayera *). Możliwe są jednak także inne postacie funkcjonału jakości.

---

*) Problemy te są wzajemnie równoważne. Na przykład, wprowadzając dodatkową współrzędną stanu x_o o równaniu

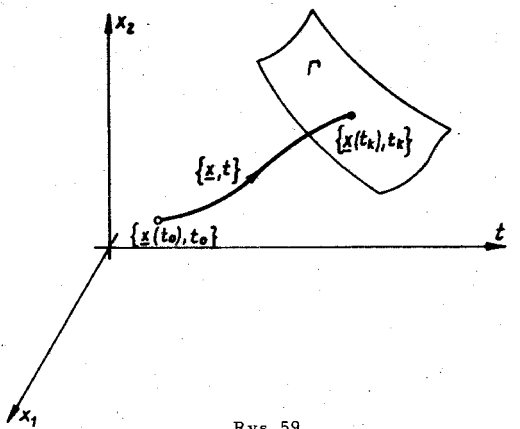
$$\dot{x}_o = f_o\left(\underline{x},\ \underline{u},\ t,\ \underline{a}\right) + \frac{\partial f_k}{\partial \underline{x}}\left(\underline{x},\ t,\ \underline{a}\right) \cdot \underline{f}\left(\underline{x},\ \underline{u},\ t,\ \underline{a}\right) + \frac{\partial f_k}{\partial t}\left(\underline{x},\ t,\ \underline{a}\right)$$

i warunku początkowym

$$x_o(t_o) = f_k\left(\underline{x}(t_o),\ t_o,\ \underline{a}\right)$$

sprowadzamy problem Bolzy do problemu Mayera, gdyż wówczas $Q = x_o(t_k)$. Podstawienie takie jest często stosowane w pracach teoretycznych dotyczących optymalizacji.

Przyjmujemy następujące klasy rozważanych funkcji czasu: zakładamy, że stan $\underline{x}$ jest absolutnie ciągłą funkcją czasu *), zaś sterowanie $\underline{u}$ - funkcją przedziałami ciągłą **). Ciągłą linię, na którą składają się kolejne stany w przestrzeni stanu procesu, nazywamy trajektorią stanu procesu lub krótko - trajektorią - rys. 59.



Rys. 59

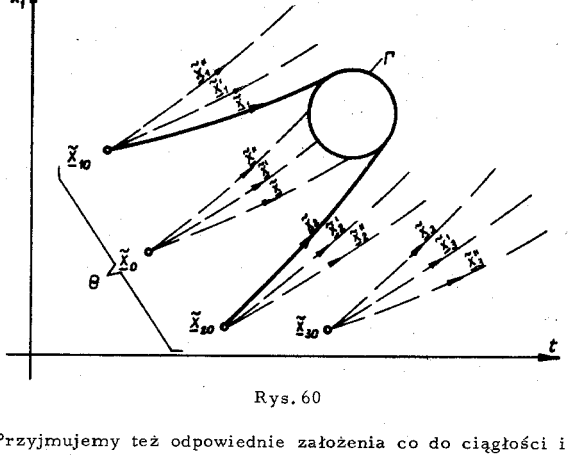
Sterowanie $\underline{u}$ przedziałami ciągłe i o wartościach $\underline{u}(t)$ spełniających dla każdego t z przedziału $t_0 \leqslant t \leqslant t_k$ ograniczenia (321) lub (322) nazywamy sterowaniem dopuszczalnym. Jeśli dodatkowo sterowanie $\underline{u}$ zastosowane do procesu (319) przy warunkach początkowych (320) doprowadza stan procesu do pewnego punktu

*) Funkcja absolutnie ciągła daje się przedstawić jako funkcja górnej granicy całkowania pewnej funkcji przedziałami ciągłej lub mierzalnej ograniczonej. Własność ta jest własnością definicyjną zmiennych stanu, które nie mogą zmieniać się skokowo.

**) Funkcja przedziałami ciągła jest to funkcja ciągła z wyjątkiem skończonej (lub przeliczalnej) liczby punktów, w których jej wartości mogą zmieniać się skokowo. Twierdzenie teorii optymalizacji formułuje się zwykle dla szerszej klasy sterowań, a mianowicie - dla sterowań mierzalnych ograniczonych, których przypadkiem szczególnym są sterowania przedziałami ciągłe. Uogólnienie to nie ma większego znaczenia dla zastosowań teorii optymalizacji.

spełniającego warunki końcowe (324) lub (325), czyli należącego do rozmaitości końcowej Γ, to nazywamy je sterowaniem docelowym.

Zauważmy, że wobec istnienia ograniczeń sterowania i dopuszczalnej nieliniowości równań stanu, nie dla każdego stanu i czasu początkowego $\underline{x}_0$, t_0 istnieje sterowanie docelowe. W przestrzeni stanu i czasu możemy więc wyróżnić obszar stanów sterowalnych docelowo oznaczany tu przez Θ - to jest takich stanów, że przy pewnym sterowaniu dopuszczalnym wychodząca z nich trajektoria osiąga rozmaitość końcową Γ - rys. 60.



Rys. 60

Przyjmujemy też odpowiednie założenia co do ciągłości i różniczkowalności wszystkich funkcji danych w zadaniu. Na ogół zakłada się, że funkcje $\underline{f}$, f_0 są ciągłe względem sterowania, różniczkowalne względem stanu oraz przedziałami ciągłe (a niekiedy znacznie silniej - różniczkowalne) względem czasu, zaś funkcje $\underline{g}_1$, $\underline{g}_2$, $\underline{g}_k$, f_k - różniczkowalne względem sterowań, stanu i czasu. Założenia te będą precyzowane dokładniej w miarę potrzeby.

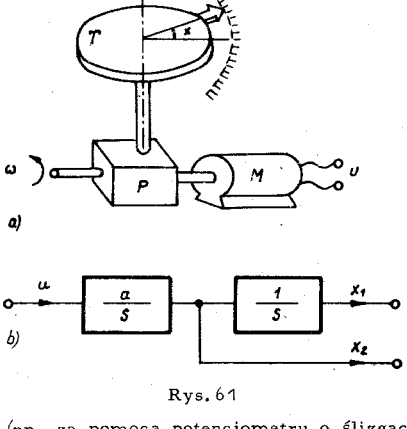
Zadanie optymalizacji dynamicznej, czyli wyznaczenia sterowania optymalnego w układzie otwartym, sprowadza się do znalezienia takiego dopuszczalnego i docelowego sterowania $\hat{\underline{u}}$ jako funkcji czasu t, że wskaźnik jakości ma przy tym sterowaniu wartość minimalną, $Q\{\hat{\underline{x}}, \hat{\underline{u}}, \underline{a}\} = \min Q\{\underline{x}, \underline{u}, \underline{a}\}$, gdzie $\hat{\underline{x}}$ - trajektoria optymalna procesu odpowiadająca sterowaniu $\hat{\underline{u}}$, zaś $\underline{u}$ i $\underline{x}$ - dowolne

sterowanie dopuszczalne i docelowe oraz odpowiadająca mu trajektoria.

Zadanie syntezy podstawowej struktury układu zamkniętego sterowania optymalnego sprowadza się do przedstawienia sterowania optymalnego $\underline{\hat{u}}$ w postaci takiej funkcji stanu $\underline{x}$ i ewentualnie czasu t, zwanej funkcją syntezującą lub algorytmem regulatora optymalnego, że dla każdego zadania optymalizacji o takich samych równaniach stanu, ograniczeniach, warunkach końcowych i wskaźniku jakości, lecz o dowolnych warunkach początkowych $\underline{x}_o$, t_o należących do obszaru sterowalności docelowej, wartość początkowa sterowania optymalnego $\underline{\hat{u}}(t_o)$ jest równa wartości funkcji syntezującej w punkcie $\underline{x}_o$, t_o.

Podkreślamy tu raz jeszcze, że o ile zdanie optymalizacji rozwiązywane jest dla danych warunków początkowych $\underline{x}_o$, t_o i przy zmianie tych warunków staje się innym zadaniem, o tyle zadanie syntezy układu zamkniętego odpowiada wielu zadaniom optymalizacji, przy dowolnych warunkach początkowych, dla których to zadanie ma sens.

Podane wyżej sformułowania i pojęcia mają charakter wysoce abstrakcyjny. Dla ich lepszego zrozumienia zilustrujemy je na dwóch przykładach fizycznych.



Rys. 61

Wyobraźmy sobie - rys. 61a - że chcemy sterować położenie kątowe x pewnej masy bezwładnej, np. tarczy T, napędzając ją poprzez przekładnię P silnikiem M. Pominiemy tu tarcie oraz założymy dla uproszczenia, że silnik rozwija stały moment obrotowy niezależnie od prędkości. Silnik jest nawrotny, a jego moment proporcjonalny do prądu sterującego oznaczonego tu przez u. Założymy ponadto, że możemy w razie potrzeby mierzyć położenie kątowe x (np. za pomocą potencjometru o ślizgaczu napędzanym przez tarczę T) oraz prędkość kątową $\dot{x}$ (np. za pomocą prądnicy tachometrycznej, dołączonej do wału silnika o prędkości obrotowej ω).

Równanie ruchu masy bezwładnej ma więc postać

$$J\ddot{x} = k\,u, \tag{327}$$

gdzie J - łączny moment bezwładności tarczy, przekładni i silnika,

k - współczynnik proporcjonalności.

Wprowadzając współrzędne stanu $x = x_1$ (położenie) i $\dot{x} = x_2$ (prędkość), przepiszemy to równanie w postaci równań stanu

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= a\,u, \end{aligned} \tag{328}$$

gdzie $a = \frac{k}{J}$ jest parametrem, którego wartość należy określić bądź to na podstawie eksperymentalnej identyfikacji, bądź też obliczeniowo, znając dane znamionowe silnika i momenty bezwładności.

Abstrahując od natury fizycznej rozważanego układu, możemy go przedstawić schematycznie jak na rys. 61b gdzie $\frac{1}{s}$ jest transmitancją członu całkującego.

W dowolnej chwili początkowej t_o stan układu tarcza-przekładnia-silnik jest określony, jeśli znamy położenie x_{10} i prędkość x_{20}.

Warunki pracy silnika dopuszczają zastosowanie maksymalnego prądu sterującego U_m. Obszar Ω dopuszczalnych wartości u jest więc odcinkiem

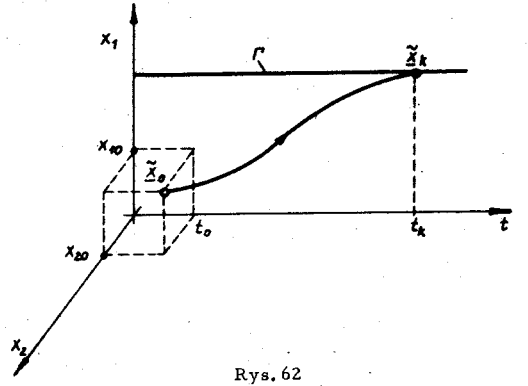
$$-U_m \leq u \leq U_m. \tag{329}$$

Chcemy doprowadzić tarczę w nieokreślonej chwili t_k do położenia x_{1k} i prędkości $x_{2k} = 0$. W przestrzeni o współrzędnych (x_1, x_2, t) - rys. 62 - rozmaitość końcowa Γ jest prostą równoległą do osi t. Oczywiście nie każde sterowanie dopuszczalne jest docelowe przy danym stanie początkowym x_{10}, x_{20}; np. jeśli $x_{20} > 0$, to sterując dopuszczalne $u = U_m$ nigdy nie zatrzymamy silnika i nie osiągniemy $x_{2k} = 0$. Równie oczywiste fizycznie jest natomiast, że dla każdego stanu początkowego znajdziemy sterowanie docelowe, ustawiające tarczę w określonym położeniu; obszar sterowalności docelowej θ pokrywa się z całą przestrzenią.

Jako wskaźnik jakości możemy przyjąć np. czas ustawiania tarczy

$$Q = t_k - t_o = \int_{t_o}^{t_k} 1 \cdot dt \tag{330}$$

Zadanie optymalizacji dynamicznej polega tu na znalezieniu takiego sterowania $\hat{u}$ jako funkcji czasu, żeby przestawić tarczę z danego x_{10}, x_{20} do danego x_{1k}, x_{2k} w minimalnym czasie.



Rys. 62

Zadanie syntezy układu zamkniętego sterowania polega na przedstawieniu $\hat{u}$ w postaci takiej funkcji stanu x_1, x_2, żeby wartości tej funkcji, zastosowane jako sterowania, zapewniały przestawienie tarczy do danego x_{1k}, x_{2k} w minimalnym czasie dla każdego stanu początkowego.

Jako inny przykład wyobraźmy sobie - rys. 63a - że chcemy nagrzewać pewien wsad W w piecu przemysłowym F, za pomocą elektrycznego urządzenia grzejnego G o nastawianej oporności, a tym samym mocy grzejnej. Ze względu na oporność doprowadzeń energii D moc pobierana z sieci P_c jest większa niż moc grzejna P_g. Przyjmując, że prąd płynący w obwodzie grzejnika jest zmienną sterującą u, uzyskamy

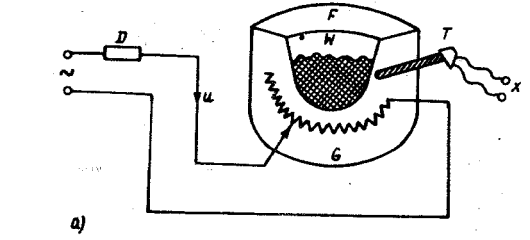
$$P_c = a_1 u,$$
$$P_g = a_1 u - a_2 u^2, \tag{331}$$

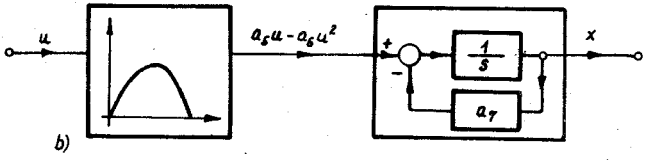
gdzie a_1 - napięcie sieci,
a_2 - oporność doprowadzeń, są parametrami o znanych wartościach,

Oznaczając przez x temperaturę wsadu, uproszczone równanie akumulacji ciepła we wsadzie można zapisać w postaci

$$a_3 \dot{x} = P_g - a_4 x, \tag{332}$$

gdzie a_3x - energia cieplna (mierzona w jednostkach elektrycznych) akumulowana we wsadzie w jednostce czasu,
a_4x - energia oddawana ze wsadu do otoczenia w jednostce czasu.





Rys. 63

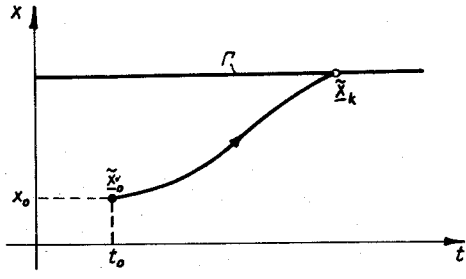
Po przekształceniach uzyskamy równanie

$$\dot{x} = a_5u - a_6u^2 - a_7x, \qquad (333)$$

przy czym $a_5 = \frac{a_1}{a_3}$, $a_6 = \frac{a_2}{a_3}$, $a_7 = \frac{a_4}{a_3}$; temperatura wsadu x jest jedyną współrzędną stanu układu w tak uproszczonym modelu matematycznym*). Schemat blokowy układu odpowiadający równaniu (333) przedstawia rys. 63b.

*) Dokładnie biorąc, akumulację ciepła w piecu należałoby opisać równaniami różniczkowymi cząsteczkowymi. Pewnym uproszczeniem jest zastosowanie osobnych równań różniczkowych zwyczajnych dla ścian pieca, sklepienia, wsadu itp. Równanie (332) jest najsilniejszym uproszczeniem, w którym rozpatruje się tylko średnią dynamikę wsadu, z pominięciem strat ciepła przez promieniowanie i innych temu podobnych zjawisk nieliniowych, przy założeniu pomijalności wpływu temperatury otoczenia itd.

Możemy zakładać, że w chwili początkowej t_o znamy temperaturę początkową x_o (mierzoną np. za pomocą termoelementu T, rys. 63a). Nie musimy tu uwzględniać ograniczeń sterowania u, gdyż stosowanie zbyt dużych wartości u jest i tak nieopłacalne ze względu na wzrost strat energii w doprowadzeniach (wyrażenie a_2u^2 w równaniu (331)). Przyjmiemy, że należy osiągnąć temperaturę końcową x_k w nieokreślonym czasie t_k; rozmaitość końcowa Γ jest więc także prostą równoległą do osi t (rys. 64).



Rys. 64

Przyjmiemy, że na koszty prowadzenia procesu składają się koszt doprowadzonej energii oraz koszt czasu użytkowania pieca. Wskaźnik jakości ma więc postać

$$Q = c_1 \int_{t_o}^{t_k} P_c dt + c_2(t_k - t_o) = \int_{t_o}^{t_k} (a_8 u + a_9) dt, \qquad (334)$$

gdzie c_1 i c_2 - ceny energii i czasu, $a_8 = c_1 a_1$, $a_9 = c_2$.

Zadanie optymalizacji polega na znalezieniu takiego sterowania u jako funkcji czasu (czyli innymi słowy, takiego przebiegu dostarczania energii elektrycznej), aby przy danej temperaturze x_o w chwili t_o uzyskać daną temperaturę x_k w pewnej chwili t_k i żeby wskaźnik jakości (334) miał przy tym wartość minimalną. Zadanie syntezy zamkniętego układu sterowania polega na uzależnieniu optymalnego sterowania od aktualnej temperatury x, tak aby na podstawie pomiaru tej temperatury można było w każdych warunkach wyznaczyć sterowanie optymalne.

Zauważmy, że w obu powyższych zadaniach nie trzeba uzależniać sterowania w układzie zamkniętym od aktualnego czasu t; ponieważ czas końcowy t_k jest w obu zadaniach swobodny, i czas nie występuje jawnie w równaniach procesu, przeto każda chwila początkowa jest równoprawna. W zadaniach tego typu nie

ma w zasadzie potrzeby posługiwania się stanem rozszerzonym $\tilde{\underline{x}} = \{ \underline{x}, t \}$ i wystarczy rozpatrywanie stanu $\underline{x}$.

8.2. Metody rozwiązywania podstawowych wariantów problemu optymalizacji dynamicznej

Istnieje wiele metod rozwiązywania zadań optymalizacji dynamicznej. Np. najprostsze zadania mogą być rozwiązane za pomocą klasycznego rachunku wariacyjnego w oparciu o równania Eulera lub Eulera-Lagrange'a [14]. Ograniczymy się tu do krótkiego przedstawienia trzech nowoczesnych metod podstawowych, opartych na zasadzie optymalności Bellmana, zasadzie maksimum Pontriagina i twierdzeniu Hurwicza o punkcie siodłowym funkcjonału Lagrange'a.

8.2.1. Zasada optymalności. Równanie Hamiltona-Jacobiego-Bellmana

R. Bellman sformułował następującą zasadę optymalności - por. [2].

> Sterowanie optymalne od danej chwili t do chwili końcowej t_k zależy tylko od aktualnego rozszerzonego stanu procesu $\tilde{\underline{x}}(t) = \{ \underline{x}(t), t \}$, a nie zależy od poprzednich stanów $\underline{x}(\tau)$ dla $\tau < t$ (czyli od sposobu, w jaki proces dotarł do stanu aktualnego).

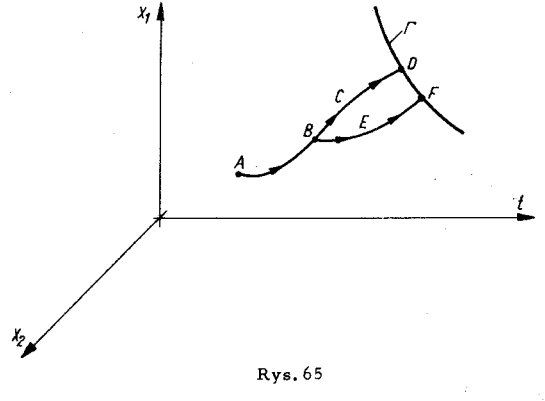
Zasada powyższa obowiązuje oczywiście tylko dla procesów, których zachowanie się jest określone w pełni przez aktualny stan $\underline{x}(t)$. Nie obowiązuje ona np. dla procesów z opóźnieniem stanu lub sterowania *).

Zasadę optymalności można sformułować także w odmienny sposób: Każdy końcowy odcinek trajektorii optymalnej jest sam dla siebie trajektorią optymalną. Istotnie, aktualny stan procesu może być zgodnie z zasadą optymalności uważany za stan początkowy dla nowego zadania optymalizacji, a więc z pierwszego sformułowania wynika drugie. I odwrotnie, jeśli końcowy odcinek jest sam dla siebie trajektorią optymalną, to odpowiadające mu sterowanie optymalne zależy tylko od jego punktu początkowego.

Zasadę optymalności można uzasadnić w prosty sposób (rys. 65).

*) Chyba, że się rozszerzy odpowiednio pojęcie stanu procesu, włączając do niego oprócz $\underline{x}(t)$ także całe przebiegi $\underline{x}(\tau)$ i $\underline{u}(\tau)$ dla τ z przedziału $t - T_o < \tau < t$, gdzie T_o - czas opóźnienia.

Załóżmy, że ABCD jest trajektorią optymalną pewnego procesu. Rozpatrzmy jej odcinek BCD. Jeśli byśmy założyli, że nie jest on sam dla siebie trajektorią optymalną, to istniałaby inna



Rys. 65

trajektoria lepsza, np. BEF. Wówczas jednak trajektoria ABEF byłaby lepsza, niż ABCD (gdyż przyrosty wskaźnika jakości na odcinku AB są takie same, a całkowity wskaźnik jest sumą przyrostów na poszczególnych odcinkach trajektorii), co jest sprzeczne z założeniem, że ABCD jest trajektorią optymalną. Tym samym BCD musi być sam dla siebie trajektorią optymalną.

Dla zadań Lagrange'a z danymi warunkami końcowymi z zasady optymalności wynika, że każdy odcinek trajektorii optymalnej jest sam dla siebie trajektorią optymalną.

Rozpatrzmy teraz zadanie Bolzy poszukiwania minimum funkcjonału *)

$$Q\{\underline{x},\ \underline{u},\ t\} = f_k\left(\underline{x}(t_k),\ t_k\right) + \int_{t_0}^{t_k} f_0(\underline{x},\ \underline{u},\ t)\,dt, \qquad (335)$$

przy równaniach procesu

$$\dot{\underline{x}} = \underline{f}(\underline{x},\ \underline{u},\ t), \qquad (336)$$

*) W zadaniu tym pomijamy w zapisie zależność od parametrów procesu $\underline{a}$, gdyż jest ona istotna tylko dla niektórych zagadnień optymalizacji, np. przy analizie wrażliwości rozmaitych struktur układów sterowania. Będziemy tak postępować stale, zaznaczając zależność od parametrów $\underline{a}$ tylko wtedy, gdy jest ona istotna. Por. [18].

ograniczeniach sterowania

$$\underline{u}(t) \in \Omega \tag{337}$$

warunkach końcowych

$$\underline{g}_k\left(\underline{x}(t_k),\ t_k\right) = \underline{0}; \qquad \left\{ x(t_k),\ t_k \right\} \in \Gamma \tag{338}$$

i dowolnych warunkach początkowych $\underline{x}(t_o) = \underline{x}_o$ należących do obszaru sterowalności docelowej. Zdefiniujemy funkcję jakości optymalnej $P(\underline{x}_o,\ t_o)$.

$$P(\underline{x}_o,\ t_o) = \min_{\underline{u}} Q\left\{ \underline{x},\ \underline{u},\ t \right\}, \tag{339}$$

gdzie $\underline{u}$ jest dowolnym sterowaniem dopuszczalnym i docelowym,
$\underline{x}$ - odpowiadającą mu trajektorią, wychodzącą z punktu $\underline{x}_o$ w momencie t_o;
$P(\underline{x}_o,\ t_o)$ jest to więc minimalna wartość wskaźnika jakości, jaką można uzyskać wychodząc z punktu $\underline{x}_o,\ t_o$.

Zgodnie z zasadą optymalności argumentem funkcji jakości optymalnej P może być dowolny punkt $\left\{ \underline{x},\ t \right\}$ każdej trajektorii optymalnej.

Załóżmy teraz, że $P(\underline{x},\ t)$ jest funkcją różniczkowalną. Istnieją zadania optymalizacji, dla których założenie to nie jest słuszne. Jeśli jednak jest ono słuszne, to obowiązuje zależność

$$P(\underline{x} + \varepsilon \Delta \underline{x},\ t + \varepsilon \Delta t) = P(\underline{x},\ t) + \varepsilon \frac{\partial P(x,\ t)}{\partial \underline{x}} \Delta \underline{x} + \varepsilon \frac{\partial P(\underline{x},\ t)}{\partial t} \Delta t + o(\varepsilon) \tag{340}$$

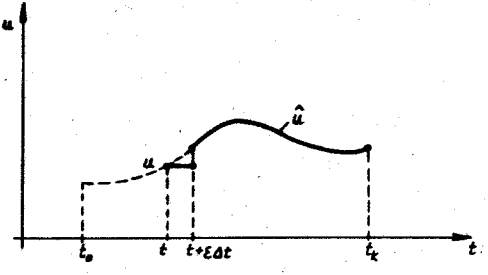
gdzie ε jest liczbą dowolnie małą, $o(\varepsilon)$ - liczbą bardzo małą w porównaniu z ε, $\Delta \underline{x}$ i Δt - dowolnymi przyrostami stanu i czasu, zaś $\frac{\partial P}{\partial \underline{x}}$ jest wektorem wierszowym *) pochodnych cząstkowych funkcji P względem składowych wektora $\underline{x}$, czyli gradientem funkcji P.

Rozpatrzmy teraz sterowanie procesu na odcinku $\left[t,\ t_k\right]$, rozbijając go na dwa odcinki $\left[t,\ t + \varepsilon \Delta t\right]$ i $\left[t + \varepsilon \Delta t,\ t_k\right]$ - rys. 66.

---

*) Umówimy się, że gradient $\frac{\partial P}{\partial \underline{x}}$ jest wektorem wierszowym, zaś $\frac{\partial P}{\partial \underline{x}'}$ - kolumnowym, tak że wyrażenia $\Delta \underline{x}' \frac{\partial P}{\partial \underline{x}'} = \frac{\partial P}{\partial \underline{x}} \Delta \underline{x}$ są iloczynami skalarnymi (sumą iloczynów poszczególnych składowych wektorów $\frac{\partial P}{\partial \underline{x}}$ i $\Delta \underline{x}$).

Załóżmy, że na drugim odcinku proces jest sterowany optymalnie i ma wskaźnik jakości $P(x + \varepsilon\Delta x,\ t + \varepsilon\Delta t)$, zaś na pierwszym odcinku sterowanie ma jakąkolwiek ustaloną wartość $\underline{u}$.



Rys. 66

Przyrost wskaźnika jakości na pierwszym odcinku wynika ze wzoru (335) i wynosi $\varepsilon\Delta t\, f_0(x, u, t) + o(\varepsilon)$, zaś przyrost stanu na pierwszym odcinku wynika ze wzoru (336) i wynosi $\varepsilon\Delta x = \varepsilon\Delta t\, f(x, u, t) + o(\varepsilon)$. Wskaźnik jakości na całym odcinku wynosi

$$Q = P(\underline{x} + \varepsilon\Delta\underline{x},\ t + \varepsilon\Delta t) + \varepsilon\Delta t\, f_0(\underline{x}, \underline{u}, t) + o(\varepsilon) =$$

$$= P(\underline{x}, t) + \varepsilon\Delta t \frac{\partial P(\underline{x}, t)}{\partial t} + \varepsilon\Delta t \left[\frac{\partial P(\underline{x}, t)}{\partial \underline{x}} \underline{f}(\underline{x}, \underline{u}, t) + \right. \tag{341}$$

$$\left. + f_0(\underline{x}, \underline{u}, t)\right] + o(\varepsilon).$$

Nie jest to minimalna wartość wskaźnika jakości, gdyż sterowanie $\underline{u}$ na pierwszym odcinku nie jest optymalne. Ponieważ odcinek ten jest bardzo krótki, przeto możemy założyć, że sterowanie optymalne jest na nim w przybliżeniu stałe w czasie, i znaleźć optymalną wartość $\hat{\underline{u}}$ drogą minimalizacji wskaźnika Q, który przyjmuje wówczas wartość $P(\underline{x}, t)$

$$P(\underline{x}, t) = \min Q =$$

$$= \min_{\underline{u} \in \Omega} \left\{ P(\underline{x}, t) + \varepsilon\Delta t \frac{\partial P(\underline{x}, t)}{\partial t} + \varepsilon\Delta t \left[\frac{\partial P(\underline{x}, t)}{\partial \underline{x}} \underline{f}(\underline{x}, \underline{u}, t) + \right. \right. \tag{342}$$

$$\left. \left. + f_0(\underline{x}, \underline{u}, t)\right] + 0(\varepsilon) \right\}.$$

W tej ostatniej zależności tylko wyrażenie w nawiasie kwadratowym zależy od $\underline{u}$. Możemy więc od obu stron równania odjąć

$P(\underline{x}, t)$, podzielić obie strony przez $\varepsilon \Delta t$ i przejść do granicy przy $\varepsilon \to 0$. Uzyskujemy wtedy równanie Hamiltona-Jacobiego--Bellmana - por. [1] - w postaci

$$\frac{\partial P(\underline{x}, t)}{\partial t} + \min_{\underline{u} \in \Omega} \left[\frac{\partial P(\underline{x}, t)}{\partial \underline{x}} \underline{f}(\underline{x}, \underline{u}, t) + f_o(\underline{x}, \underline{u}, t) \right] = 0. \quad (343)$$

Funkcję H zmiennych $\underline{\psi}$, $\underline{x}$, $\underline{u}$, t o postaci

$$H(\underline{\psi}, \underline{x}, \underline{u}, t) = -f_o(\underline{x}, \underline{u}, t) + \underline{\psi}' \underline{f}(\underline{x}, \underline{u}, t) \quad (344)$$

nazywamy hamiltonianem zadania optymalizacji, zaś funkcję $\tilde{H}$ o postaci

$$\tilde{H}(\tilde{\underline{\psi}}, \tilde{\underline{x}}, \tilde{\underline{u}}) = H(\underline{\psi}, \underline{x}, \underline{u}, t) + \psi_t; \quad \tilde{\underline{\psi}} = \{\underline{\psi}, \psi_t\} \quad (345)$$

hamiltonianem rozszerzonym. Nie sprecyzowaliśmy tu na razie znaczenia zmiennej wektorowej $\underline{\psi}$ i skalarnej ψ_t. Jeśli jednak zastosujemy podstawienie

$$\hat{\underline{\psi}}(\underline{x}, t) = -\frac{\partial P(\underline{x}, t)}{\partial \underline{x}'}; \quad \hat{\psi}_t(\underline{x}, t) = -\frac{\partial P(\underline{x}, t)}{\partial t} \quad (346)$$

to możemy - zmieniając znak obu stron równania (343) i zamieniając przy tym operację poszukiwania minimum na poszukiwanie maksimum - zapisać równanie (343) w postaci

$$\hat{\psi}_t(\underline{x}, t) + \max_{\underline{u} \in \Omega} H(\hat{\underline{\psi}}(\underline{x}, t), \underline{x}, \underline{u}, t) = 0 \quad (347)$$

lub w postaci

$$\max_{\underline{u} \in \Omega} \tilde{H}(\hat{\tilde{\underline{\psi}}}, \hat{\tilde{\underline{x}}}, \underline{u}) = 0. \quad (348)$$

Zauważmy, że w równaniu (343) lub równaniach (347), (348) minimum wyrażenia w nawiasie kwadratowym lub maksimum hamiltonianu zapewnia taka wartość sterowania $\underline{u}$, która jest sterowaniem optymalnym $\hat{\underline{u}}$. Z równania Hamiltona-Jacobiego-Bellmana wynika więc ważna zasada, która może być zresztą udowodniona w niezależny sposób i którą dalej sformułujemy nieco dokładniej, zwana zasadą maksimum:

> sterowanie optymalne $\hat{\underline{u}}$ zapewnia w każdej chwili czasu t z przedziałem (t_o, t_k) maksimum hamiltonianu zadania optymalizacji, przy czym maksimum hamiltonianu rozszerzonego jest równe zeru.

Gdybyśmy więc znali funkcję $\hat{\underline{\psi}}(\underline{x}, t) = -\frac{\partial P(\underline{x}, t)}{\partial \underline{x}'}$, to przez poszukiwanie maksimum hamiltonianu moglibyśmy określić sterowanie optymalne $\hat{\underline{u}}$, i to od razu jako funkcję stanu i czasu, a

więc rozwiązać zadanie syntezy układu zamkniętego sterowania optymalnego. Jeśli znamy rozwiązanie tego zadania, to możemy wyznaczyć trajektorię optymalną procesu $\hat{\underline{x}}$, a następnie sterowanie optymalne $\hat{\underline{u}}$ jako funkcję samego już czasu t, a więc rozwiązać zadanie optymalizacji. Jednakże funkcja $\hat{\underline{\psi}}(\underline{x}, t)$ nie jest z góry znana. Dlatego też postępowanie przy wyznaczaniu sterowania optymalnego na podstawie równania Hamiltona-Jacobiego-Bellmana jest następujące:

1) Rozwiązujemy zadanie poszukiwania maksimum hamiltonianu $H(\hat{\underline{\psi}}, \underline{x}, \underline{u}, t)$ względnie zmiennej $\underline{u} \in \Omega$ przy $\underline{x}, \hat{\underline{\psi}}, t$ traktowanych jako parametry. Jeśli potrafimy rozwiązać to zadanie (będące w gruncie rzeczy zadaniem optymalizacji statycznej) na drodze analitycznej, to uzyskujemy sterowanie optymalne jako taką funkcję zmiennych $\hat{\underline{\psi}}, \underline{x}, t$

$$\hat{\underline{u}} = \underline{\varphi}(\hat{\underline{\psi}}, \underline{x}, t), \tag{349}$$

która po podstawieniu do hamiltonianu zapewnia jego maksimum.

2) Uzyskaną funkcję podstawiamy do równania Hamiltona-Jacobiego-Bellmana, wracając przy tym do oznaczenia $-\frac{\partial P(\underline{x}, t)}{\partial \underline{x}}$ zamiast $\hat{\underline{\psi}}$. Równanie to przyjmuje wtedy postać *)

$$\frac{\partial P(\underline{x}, t)}{\partial t} + \frac{\partial P(\underline{x}, t)}{\partial \underline{x}} \underline{f}\left(\underline{x}, \underline{\varphi}\left(-\frac{\partial P(\underline{x}, t)}{\partial \underline{x}}, \underline{x}, t\right), t\right) + \tag{350}$$

$$+ f_o\left(\underline{x}, \underline{\varphi}\left(-\frac{\partial P(\underline{x}, t)}{\partial \underline{x}}, \underline{x}, t\right), t\right) = 0.$$

Równanie powyższe jest w ogólnym przypadku nieliniowym równaniem różniczkowym cząstkowym rzędu pierwszego. Dla pełnego określenia jego rozwiązania niezbędne jest podanie warunków brzegowych dla funkcji $P(\underline{x}, t)$. Spośród wszystkich funkcji spełniających równanie (350) należy wybrać taką, która spełnia oczywisty (por. wzory (335) i (339), podstawiając $t_o = t_k$) warunek

$$P(\underline{x}_k, t_k) = f_k(\underline{x}_k, t_k), \qquad \{\underline{x}_k, t_k\} \in \Gamma \tag{351}$$

dla wszystkich $\{\underline{x}_k, t_k\}$ należących do rozmaitości końcowej Γ.

---

*) Jest to postać zbliżona do klasycznej postaci równania Hamiltona-Jacobiego, wyprowadzonej zresztą inną drogą przy znacznie silniejszych założeniach. Postać (343) i przedstawiona tu droga wyprowadzenia podane były przez Bellana - zob. np. [2].

Nie będziemy tu omawiali ogólnych metod rozwiązania równania (350), posługując się w przykładzie metodami szczegółowymi *).

3) Jeśli znajdziemy rozwiązanie $P(\underline{x}, t)$ równania (346) przy warunku (351), to możemy wyznaczyć sterowanie optymalne w układzie zamkniętym, obliczając $\frac{\partial P(\underline{x}, t)}{\partial \underline{x}'}$ i podstawiając do zależności (349)

$$\hat{\underline{u}} = \underline{\varphi}\left(-\frac{\partial P(\underline{x}, t)}{\partial \underline{x}'}, \underline{x}, t\right) = \hat{\underline{r}}(\underline{x}, t). \tag{352}$$

Funkcja $\hat{\underline{r}}(\underline{x}, t)$ zwana jest funkcją syntetyzującą podstawowy wariant układu zamkniętego sterowania optymalnego.

4) Podstawiając zależność (352) do równań stanu (336), całkujemy te ostatnie przy danych warunkach początkowych $\underline{x}_o$, t_o. Uzyskujemy w ten sposób optymalną trajektorię procesu $\hat{\underline{x}}$, czyli przebieg stanu jako funkcji czasu przy sterowaniu optymalnym. Podstawiając $\hat{\underline{x}}$ do zależności (352), uzyskujemy z kolei sterowanie $\hat{\underline{u}}$ jako funkcję czasu, a więc rozwiązanie zadania optymalizacji.

Zauważmy, że postępowanie powyższe rozwiązuje najpierw - niejako mimochodem - zagadnienie syntezy układu zamkniętego, a dopiero potem zagadnienie optymalizacji w układzie otwartym. Z postępowaniem tym wiąże się szereg trudności. Przypadki, w których zadanie optymalizacji może być tą drogą w pełni rozwiązane w postaci analitycznej, są raczej rzadkie. Ponadto przy rozwiązywaniu pojawia się szereg wątpliwości co do istnienia i jednoznaczności rozwiązań problemu. W związku z tym hamiltonian $H(\underline{\psi}, \underline{x}, \underline{u}, t)$, który ma jednoznaczne absolutne maksimum względem $\underline{u} \in \Omega$ dla każdych $\{\underline{x}, t\}$ należących do pewnego obszaru $\tilde{X}$, nazywamy normalnym w obszarze $\tilde{X}$, zaś maksymalizujące go sterowanie $\hat{\underline{u}} = \underline{\varphi}(\underline{\psi}, \underline{x}, t)$ nazywamy ekstremalnym. Obowiązuje przy tym twierdzenie:

> Jeśli hamiltonian jest normalny w obszarze $\tilde{X}$ i jeśli istnieje rozwiązanie $P(\underline{x}, t)$ równania (350) przy warunkach (351) w tym obszarze, takie że sterowanie ekstremalne

---

*) Przy odpowiednio silnych założeniach co do różniczkowalności funkcji f_o, $\underline{f}$, $\underline{\varphi}$ równanie to może być rozwiązane metodą charakterystyk Cauchy'ego. Ponieważ metoda ta prowadzi do równań, wykorzystywanych w pełnym sformułowaniu zasady maksimum przy słabszych założeniach, nie omawiamy jej tutaj.

$\hat{u} = \varphi\left(\frac{\partial P(x, t)}{\partial x}, x, t\right)$ jest dopuszczalne, docelowe i nie wyprowadza trajektorii procesu poza obszar $\tilde{X}$, to sterowanie to jest lokalnie optymalne (w porównaniu ze wszystkimi sterowaniami odpowiadającymi trajektoriom w obszarze $\tilde{X}$).

Jako prosty przykład zastosowania przedstawionej wyżej teorii rozpatrzymy problem o skalarnym równaniu stanu

$$\dot{x} = u; \qquad x(t_o) = x_o, \tag{353}$$

bez ograniczeń, o danym czasie końcowym t_k i dowolnym $x(t_k)$ i o wskaźniku jakości

$$Q = \frac{1}{2}\left\{\left[x(t_k)\right]^2 + \int_{t_o}^{t_k} u^2\, dt\right\}. \tag{354}$$

Mamy tu $f_o = \frac{1}{2}u^2$, $f = u$, a więc hamiltonian (344) ma postać

$$H = -\frac{1}{2}u^2 + \hat{\psi}u. \tag{355}$$

W obszarze $\tilde{X}$, stanowiącym całą płaszczyznę zmiennych x, t, hamiltonian ten ma jedyne maksimum względem u z obszaru Ω, stanowiącego całą prostą zmiennej u. Hamiltonian ten jest więc normalny, a jedyne sterowanie ekstremalne ma postać

$$\hat{u} = \varphi(\hat{\psi}, x, t) = \hat{\psi}$$

Możemy teraz napisać równanie Hamiltona-Jacobiego-Bellmana w postaci (350) dla tego problemu

$$\frac{\partial P(x, t)}{\partial t} - \frac{1}{2}\left[\frac{\partial P(x, t)}{\partial x}\right]^2 = 0 \tag{356}$$

oraz warunek brzegowy $\left(f_k = \left[\frac{1}{2}\, x(t_k)\right]^2\right)$

$$P(x, t_k) = \frac{1}{2}x^2. \tag{357}$$

Zauważmy, że równanie (356) może być spełnione m.in. przez dowolną funkcję postaci

$$P_o(x, t) = \frac{(x + b)^2}{2(a - t)}; \quad \frac{\partial P_o(x, t)}{\partial x} = \frac{x + b}{a - t}; \quad \frac{\partial P_o(x, t)}{\partial t} = \frac{(x + b)^2}{2(a - t)^2},$$

gdzie a i b - stałe dowolne.

Ponieważ jednak funkcja $P(x, t)$ nie zależy od t ani t_k dla $t = t_k$, zgodnie z warunkiem (357), więc musi zachodzić $a = c + t_k$, gdzie c - stała dowolna. Dla spełnienia warunku (357) dla dowolnych x potrzeba jeszcze, by $b = 0$, $c = 1$. Funkcja jakości optymalnej ma więc postać

$$P(x, t) = \frac{x^2}{2(1 + t_k - t)}, \tag{358}$$

którą to postać raczej odgadliśmy, niż wyprowadziliśmy na drodze analitycznej. Obliczamy dalej

$$\hat{\psi}(x, t) = - \frac{\partial P(x, t)}{\partial x} = - \frac{x}{1 + t_k - t} \tag{359}$$

oraz

$$\hat{u} = \hat{r}(x, t) = - \frac{x}{1 + t_k - t}. \tag{360}$$

Zależność ta określa układ zamknięty sterowania optymalnego. Dla wyznaczenia sterowania optymalnego w układzie otwartym rozwiązujemy równanie

$$\dot{x} = - \frac{x}{1 + t_k - t}; \quad \frac{dx}{x} = - \frac{dt}{1 + t_k - t};$$

przy warunkach początkowych x_o, t_o, uzyskując optymalną trajektorię

$$x(t) = \frac{x_o}{1 + t_k - t_o} (1 + t_k - t) \tag{361}$$

oraz sterowanie optymalne

$$u(t) = - \frac{x_o}{1 + t_k - t_o}. \tag{362}$$

8.2.2. Wariant podstawowy zasady maksimum

W problemie optymalizacji dynamicznej

$$\dot{\underline{x}} = \underline{f}(\underline{x}, \underline{u}, t); \qquad \underline{x}(t_o) = \underline{x}_o, \tag{363}$$

$$\underline{u} \in \Omega \tag{364}$$

$$\underline{g}_k\left(\underline{x}(t_k), t_k\right) = \underline{0} \Longleftrightarrow \left\{x(t_k), t_k\right\} \in \Gamma, \tag{365}$$

$$Q = f_k\left(\underline{x}(t_k),\ t_k\right) + \int_{t_0}^{t_k} f_o(\underline{x},\ \underline{u},\ t)\,dt. \qquad (366)$$

założymy, że funkcje f_o i $\underline{f}$ są ciągłe względem $\underline{u}$, zaś różniczkowalne względem $\underline{x}$ i t, że funkcje f_k i $g_{k_{\partial g_{ki}}}$ są różniczkowalne względem $\underline{x}(t_k)$, t_k oraz że gradienty - $\frac{}{\partial \underline{x}}$ poszczególnych składowych g_{ki} funkcji g_k są wektorami liniowo niezależnymi *) (mówimy wówczas, że rozmaitość końcowa jest gładka).

Wprowadzimy teraz pojęcie zmiennych sprzężonych (ze stanem układu) $\underline{\psi}$. Są to takie absolutnie ciągłe funkcje czasu, które stanowią rozwiązania równań sprzężonych, czyli liniowych równań różniczkowych o postaci

$$\dot{\underline{\psi}} = \frac{\partial f_o(\underline{x},\ \underline{u},\ t)}{\partial \underline{x}'} - \frac{\partial \underline{f}'_1(\underline{x},\ \underline{u},\ t)}{\partial x'}\,\underline{\psi} \qquad (367)$$

gdzie $\frac{\partial f_o}{\partial \underline{x}'}$ oraz $\frac{\partial \underline{f}'}{\partial \underline{x}'}$ są odpowiednio gradientem funkcji f_o (w formie wektora kolumnowego) oraz macierzą pochodnych cząstkowych składowych f_i funkcji $\underline{f}$ (uzyskaną przez ustawienie obok siebie wektorów kolumnowych $\frac{\partial f_i}{\partial \underline{x}'}$.

Zauważmy, że używając pojęcia hamiltonianu

$$H(\underline{\psi},\ \underline{x},\ \underline{u},\ t) = -f_o(\underline{x},\ \underline{u},\ t) + \underline{\psi}'\underline{f}(\underline{x},\ \underline{u},\ t), \qquad (368)$$

możemy zapisać równania stanu w postaci

$$\dot{\underline{x}} = \frac{\partial H(\underline{\psi},\ \underline{x},\ \underline{u},\ t)}{\partial \underline{\psi}'}, \qquad (369)$$

zaś równania sprzężone - w postaci

$$\dot{\underline{\psi}} = -\frac{\partial H(\underline{\psi},\ \underline{x},\ \underline{u},\ t)}{\partial \underline{x}'}. \qquad (370)$$

Podobnie, wprowadzając dodatkową zmienną sprzężoną z czasem ψ_t o równaniu $\dot{\psi}_t = -\frac{\partial H(\underline{\psi},\ \underline{x},\ \underline{u},\ t)}{\partial t}$, definiując wektory rozszerzone $\tilde{\underline{\psi}} = \{\underline{\psi}, \psi_t\}$; $\tilde{\underline{x}} = \{\underline{x},\ t\}$ i hamiltonian rozszerzony

*) To znaczy, że dla dowolnych, nierównych jednocześnie zeru współczynników a_i zachodzi $\sum_i a_i \frac{\partial g_{ki}}{\partial \underline{x}} \neq 0$.

$\tilde{H} = H + \psi_t$, oraz uwzględniając, że $\dot{t} = 1$, możemy napisać

$$\dot{\underline{x}} = \frac{\partial \tilde{H}(\tilde{\psi}, \underline{x}, u)}{\partial \tilde{\psi}'}; \qquad \dot{\tilde{\psi}} = -\frac{\partial \tilde{H}(\tilde{\psi}, \underline{x}, u)}{\partial \underline{x}'}. \tag{371}$$

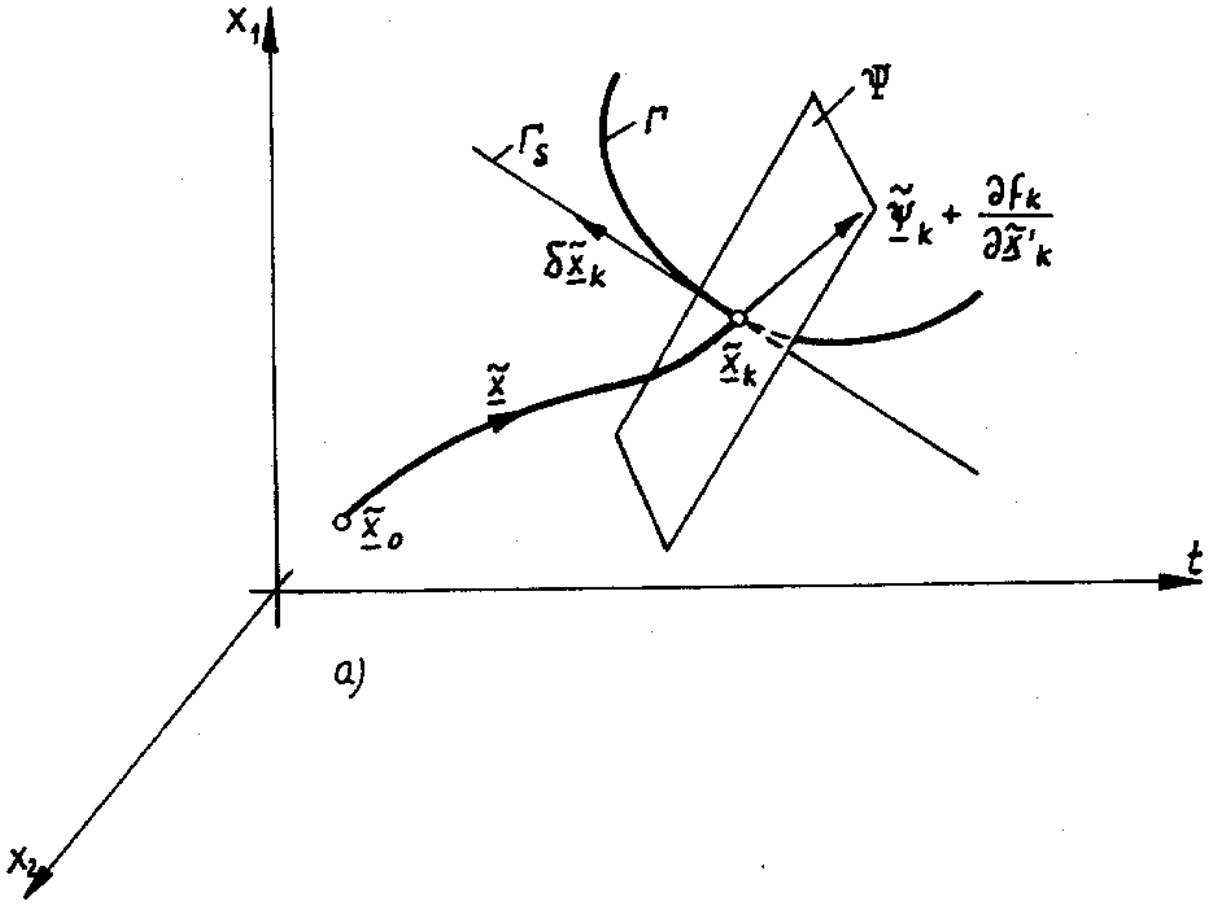
Dla rozwiązania równań sprzężonych określa się zazwyczaj nie warunki początkowe $\psi(t_o)$, lecz warunki końcowe $\psi(t_k)$, i to w dość złożonej formie, zwanej warunkami transwersalności. Wprowadza się przy tym pojęcie dopuszczalnej wariacji $\{\delta \underline{x}_k, \delta t_k\}$ końca trajektorii procesu $\underline{x}(t_k)$, t_k. Wektor $\{\delta \underline{x}_k, \delta t_k\}$ stanowi wariację dopuszczalną, jeśli jest on styczny do rozmaitości końcowej Γ - por. rys. 67; spełnia on wówczas warunek

$$\frac{\partial g_k\left(\underline{x}(t_k), t_k\right)}{\partial \underline{x}(t_k)} \delta \underline{x}_k + \frac{\partial g_k\left(\underline{x}(t_k), t_k\right)}{\partial t_k} \delta t_k = 0. \tag{372}$$

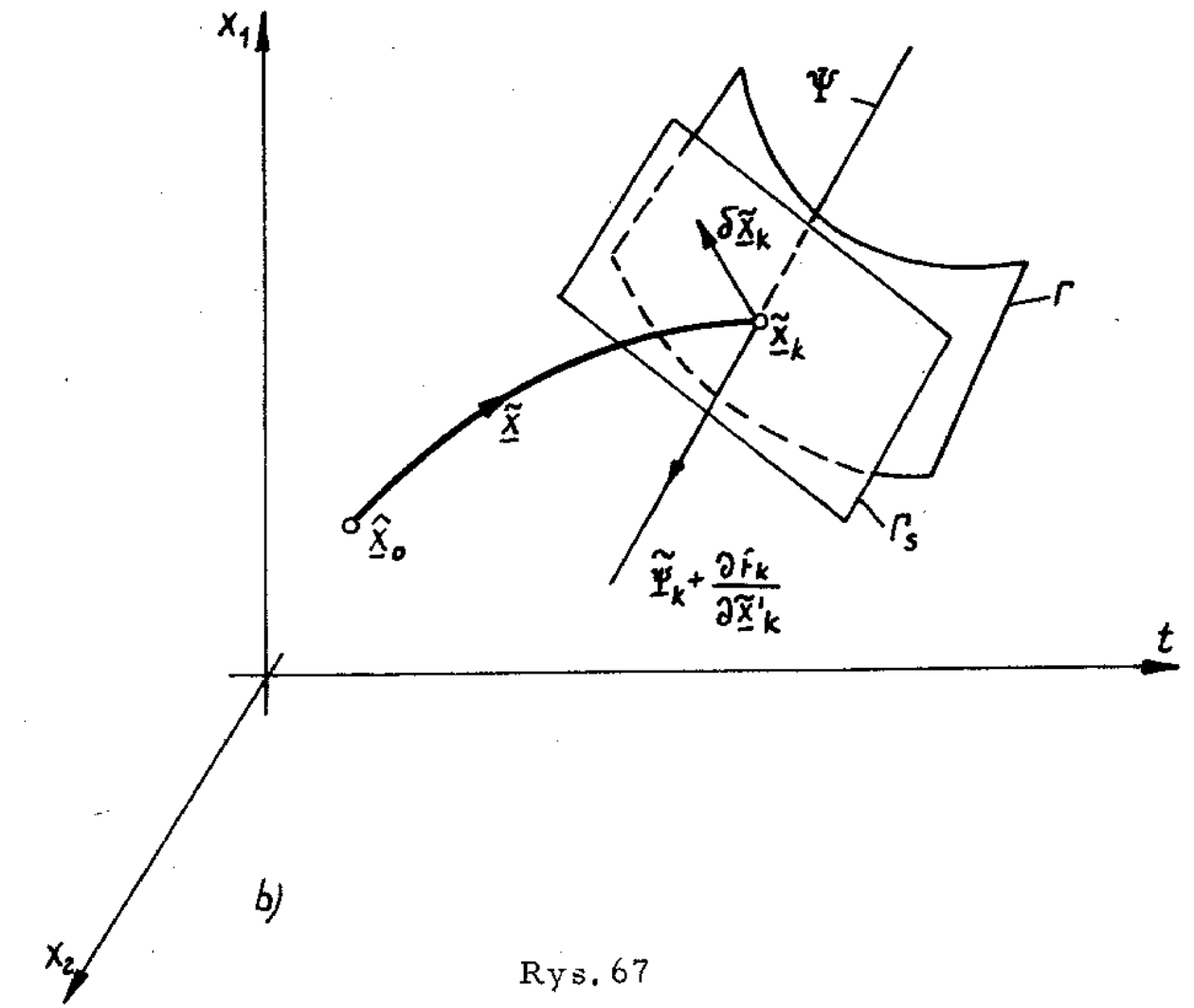
Zbiór wszystkich wariacji dopuszczalnych tworzy rozmaitość liniową Γ_s, styczną w punkcie $\{\underline{x}(t_k), t_k\}$ do rozmaitości Γ. Mówimy, że wektor $\tilde{\psi}(t_k) = \{\psi(t_k), \psi_t(t_k)\}$ spełnia warunki transwersalności, jeśli wektor $\tilde{\psi}(t_k) + \frac{\partial f_k\left(\tilde{\underline{x}}(t_k)\right)}{\partial \tilde{\underline{x}}(t_k)}$ jest normalny do zbioru wariancji dopuszczalnych Γ_s - por. rys. 67; spełnia on wówczas warunek

$$\left[\psi(t_k) + \frac{\partial f_k\left(\underline{x}(t_k), t_k\right)}{\partial \underline{x}(t_k)}\right] \delta \underline{x}(t_k) + \left[\psi_t(t_k) + \frac{\partial f_k\left(\underline{x}(t_k), t_k\right)}{\partial t_k}\right] \delta t_k = 0 \tag{373}$$

Zbiór wszystkich wektorów $\tilde{\psi}(t_k) + \frac{\partial f_k\left(\tilde{\underline{x}}(t_k)\right)}{\partial \tilde{\underline{x}}(t_k)}$ spełniających warunek (373) tworzy rozmaitość liniową Ψ, ortogonalną (prostopadłą) do rozmaitości Γ_s - por. rys. 67a, 67b. Zauważmy, że warunki transwersalności wraz z warunkami końcowymi dają łącznie $(n + 1)$ warunków, dotyczących końcowego punktu trajektorii $\tilde{\underline{x}}(t_k) = \{\underline{x}(t_k), t_k\}$ oraz końcowego wektora sprzężonego $\tilde{\psi}(t_k) = \{\psi(t_k), \psi_t(t_k)\}$. Niech bowiem będzie dane p warunków końcowych czyli p składowych funkcji g_k; rozmaitość końcowa Γ jest wówczas $(n + 1 - p)$-wymiarowa (na przykład na rys. 67b mamy $n + 1 = 3$, $p = 1$, zaś Γ jest dwuwymiarową powierzchnią). Taki sam wymiar ma rozmaitość styczna Γ_s, zaś rozmaitość ortogonalna Ψ jest wówczas p-wymiarowa (na rys. 67b Ψ jest jednowymiarową prostą; na rys. 67a mamy $n + 1 = 3$, $p = 2$, Γ jest linią a Γ_s prostą, zaś Ψ - dwuwymiarową płaszczyzną). Na


x_1
x_2
t
Ψ
Γ
Γ_s
$\delta\tilde{\underline{x}}_k$
$\tilde{\underline{\psi}}_k + \frac{\partial f_k}{\partial \tilde{\underline{x}}'_k}$
$\tilde{\underline{x}}$
$\tilde{\underline{x}}_k$
$\tilde{\underline{x}}_o$

a)


x_1
x_2
t
Ψ
$\delta\tilde{\underline{x}}_k$
Γ
$\tilde{\underline{x}}_k$
$\tilde{\underline{x}}$
Γ_s
$\hat{\underline{x}}_o$
$\tilde{\underline{\psi}}_k + \frac{\partial f_k}{\partial \tilde{\underline{x}}'_k}$

b)

Rys. 67

$(n+1)$-wymiarowy wektor, który musi należeć do rozmaitości p-wymiarowej, nałożone jest $(n+1-p)$-warunków. Tak więc warunki transwersalności (372) i (373) określają łącznie $(n+1-p)$ warunków co do wektora $\underline{\psi}(t_k) + \frac{\partial f_k(\underline{x}(t_k))}{\partial \underline{x}(t_k)}$, czyli uzupełniają liczbę warunków końcowych do $n+1$. Zauważmy dalej, że jeśli którakolwiek współrzędna stanu końcowego $x_i(t_k)$ jest dana z góry, to dopuszczalna wariacja tej współrzędnej $\delta x_{ik} = 0$ i odpowiednia składowa wektora sprzężonego $\psi_i(t_k)$ może być dowolna. Jeśli natomiast współrzędna $x_i(t_k)$ nie występuje w warunkach końcowych (jest swobodna), to δx_{ik} może mieć dowolny znak i wartość; stąd wynika, że odpowiednia składowa wektora transwersalnego musi być równa zeru

$$\psi_i(t_k) + \frac{\partial f_k\ \underline{x}(t_k)}{\partial \underline{x}_i(t_k)} = 0.$$

Jeśli na przykład mamy dane wszystkie współrzędne wektora $\underline{x}(t_k)$, a tylko czas t_k jest swobodny, to nie możemy nic powiedzieć o wektorze $\underline{\psi}(t_k)$, a tylko $\psi_t(t_k) = 0$; jest to częsty przypadek zadań optymalizacji. Inny często spotykany przypadek polega na tym, że dany jest czas końcowy t_k, zaś wektor $\underline{x}(t_k)$ - swobodny. Nie możemy wówczas nic powiedzieć o wartości $\psi_t(t_k)$; natomiast wektor $\underline{\psi}(t_k)$ wynika wówczas z równania

$$\underline{\psi}(t_k) = - \frac{\partial f_k(\underline{x}(t_k))}{\partial \underline{x}(t_k)} . \qquad (374)$$

Pojęcie zmiennych sprzężonych i warunków transwersalności umożliwiają pełne sformułowanie zasady maksimum Pontriagina - por. [15] - dla problemu optymalizacji, określonego równaniami (362), (363), (364), (365):

> Jeśli $\hat{\underline{u}}$ jest dopuszczalnym i docelowym sterowaniem optymalnym, to istnieje taka funkcja czasu $\{\psi, \psi_t\}$ spełniająca równania sprzężone i warunki transwersalności, że w każdej *) chwili t z przedziału (t_o, t_k) sterowanie $\hat{\underline{u}}(t)$ zapewnia maksimum hamiltonianu względem wszystkich $\underline{u} \in \Omega$

*) Jeśli założymy, że sterowania $\underline{u}$ są mierzalnymi ograniczonymi funkcjami czasu, to maksimum hamiltonianu jest zapewnione tylko dla prawie każdej chwili t, czyli z wyjątkiem skończonej lub przeliczalnej liczby punktów.

$$H(\hat{\underline{\psi}}, \hat{\underline{x}}, \hat{\underline{u}}, t) = \max_{\underline{u} \in \Omega} H(\hat{\underline{\psi}}, \hat{\underline{x}}, \underline{u}, t) = M(\underline{\psi}, \hat{\underline{x}}, t), \qquad (375)$$

przy czym maksimum hamiltonianu rozszerzonego jest równe zeru

$$\tilde{M}(\hat{\underline{\psi}}, \hat{\underline{x}}) = M(\hat{\underline{\psi}}, \hat{\underline{x}}, t) + \hat{\psi}_t = \max_{\underline{u} \in \Omega} \tilde{H}(\hat{\underline{\psi}}, \hat{\underline{x}}, \underline{u}) = 0. \qquad (376)$$

W przypadkach szczególnych, gdy zadanie optymalizacji nie zależy w sposób jawny od czasu, mamy $\dot{\psi}_t = -\frac{\partial H}{\partial t} = 0$, a więc ψ_t i maksimum hamiltonianu M są stałe w czasie; jeśli dodatkowo czas t_k jest swobodny, to $\psi_t = 0$ i $M = 0$.

Dowód zasady maksimum przeprowadza się [15] w sposób niezależny od równania Hamiltona-Jacobiego-Bellmana. Pozornie, sformułowane zasady maksimum wynikające z równania Hamiltona-Jacobiego-Bellmana niewiele się różni od sformułowania podanego wyżej; są jednak między nimi istotne różnice.

Po pierwsze, w dowodzie pełnego sformułowania zasady maksimum nie potrzeba zakładać różniczkowalności funkcji $P(\underline{x}, t)$; zasada maksimum obowiązuje także wtedy, gdy funkcja $P(\underline{x}, t)$ nie jest różniczkowalna.

Po drugie, zasada maksimum jest tylko warunkiem koniecznym optymalności sterowania; natomiast równanie Hamiltona-Jacobiego-Bellmana stanowi warunek konieczny, ale także przy dodatkowych założeniach o normalności hamiltonianu - warunek lokalnie dostateczny.

Po trzecie, zmienne sprzężone w zasadzie maksimum są funkcjami czasu, a w równaniu Hamiltona-Jacobiego-Bellmana - funkcjami stanu i czasu. Metoda rozwiązywania zagadnień optymalizacji oparta na tym równaniu prowadzi do wyznaczenia sterowania optymalnego w układzie zamkniętym, a dopiero potem - w układzie otwartym, zaś zasada maksimum umożliwia wyznaczanie sterowania optymalnego od razu w układzie otwartym, przy czym przejście do syntezy układu zamkniętego nie zawsze jest proste.

Związki pomiędzy zasadą maksimum a równaniem Hamiltona-Jacobiego-Bellmana pozwalają na dobrą interpretację zmiennych sprzężonych: wyznaczona w zasadzie maksimum funkcja czasu $\hat{\underline{\psi}}$ jest w każdej chwili równa gradientowi funkcji optymalnej jakości $P(\underline{x})$ względem aktualnego stanu procesu $\underline{x}$, ze zmienionym znakiem (jeśli ten gradient istnieje)

$$\hat{\underline{\psi}} = -\frac{\partial P(\underline{x}(t))}{\partial \underline{x}'} \qquad (377)$$

Można też wykazać, że warunki transwersalności dla $\hat{\underline{\psi}}(t_k)$ są równoważne warunkom brzegowym (347) w zapisie różniczkowym oraz że ogólna metoda charakterystyk Cauchy'ego dla rozwiązywa-

nia równania Hamiltona-Jacobiego-Bellmana prowadzi do układu równań różniczkowych zwyczajnych (367) lub (370), czyli do równań sprzężonych.

Zasada maksimum umożliwia rozwiązanie na drodze analitycznej znacznie szerszej klasy zagadnień optymalizacji, niż równanie Hamiltona-Jacobiego-Bellmana. Metodyka wykorzystania zasady maksimum jest następująca:

1. Formułujemy hamiltonian zadania optymalizacji i poszukujemy jego maksimum. Zakładamy, że potrafimy rozwiązać na drodze analitycznej zadanie poszukiwania maksimum hamiltonianu względem $\underline{u} \in \Omega$ i wyznaczyć sterowanie ekstremalne $\underline{\hat{u}} = \underline{\varphi}(\underline{x}, \underline{\psi}, t)$.

2. Podstawiamy wyznaczone sterowanie ekstremalne do równań stanu i równań sprzężonych, uzyskując układ równań *)

$$\dot{\underline{x}} = \frac{\partial H(\underline{\psi}, \underline{x}, \underline{\varphi}(\underline{\psi}, \underline{x}, t), t)}{\partial \underline{\psi}'} = \underline{f}(\underline{x}, \underline{\varphi}(\underline{\psi}, \underline{x}, t), t), \tag{378}$$

$$\dot{\underline{\psi}} = -\frac{\partial H(\underline{\varphi}, \underline{x}, \underline{\varphi}(\underline{\psi}, \underline{x}, t), t)}{\partial \underline{x}'} = \frac{\partial f_o(\underline{x}, \underline{\varphi}, t)}{\partial \underline{x}'} - \frac{\partial \underline{f}'(\underline{x}, \underline{\varphi}, t)}{\partial \underline{x}'} \underline{\psi}, \tag{379}$$

$$\dot{\psi}_t = -\frac{\partial H(\underline{\psi}, \underline{x}, \underline{\varphi}(\underline{\psi}, \underline{x}, t), t)}{\partial t} = \frac{\partial f_o(\underline{x}, \underline{\varphi}, t)}{\partial t} - \frac{\partial \underline{f}'(\underline{x}, \underline{\varphi}, t)}{\partial t} \underline{\psi} \tag{380}$$

nazywany układem <u>kanonicznym</u>; zauważmy, że równania (379), w odróżnieniu od równań sprzężonych (367), nie muszą być liniowe względem zmiennej $\underline{\psi}$. Zakładamy, że potrafimy rozwiązać równanie kanoniczne na drodze analitycznej; musimy przy tym założyć z góry $(n + 1)$ warunków początkowych dla wektora $\tilde{\underline{\psi}}(t_o) = \{\underline{\psi}(t_o), \psi_t(t_o)\}$, zaś warunki $\underline{x}(t_o)$ są dane. Każde przy dowolnych warunkach początkowych $\underline{x}(t_o) = \underline{x}_o$, $\tilde{\underline{\psi}}(t_o) = \tilde{\underline{\psi}}_c$ rozwiązanie równań (378) - (380) będziemy nazywać <u>ekstremalą problemu</u>; w istocie, rozwiązanie to może odpowiadać trajektorii optymalnej problemu optymalizacji, który różni się od problemu aktualnie rozwiązywanego tylko warunkami końcowymi i ewentualnie początkowymi. Zakładamy, że znamy postać analityczną ekstremal $\underline{x}(\underline{x}_o, \tilde{\underline{\psi}}_o, t)$ i $\tilde{\underline{\psi}}(\underline{x}_o, \tilde{\underline{\psi}}_o, t)$.

3. Podstawiając do warunków końcowych i warunków transwersalności ekstremale $\underline{x}(\underline{x}_o, \tilde{\underline{\psi}}_o, t_k)$ i $\tilde{\underline{\psi}}(\underline{x}_o, \tilde{\underline{\psi}}_o, t_k)$ uzyskujemy $(n + 1)$ równań o $(n + 1)$ niewiadomych $\tilde{\underline{\psi}}_o = \{\underline{\psi}_o, \psi_{to}\}$ oraz dodat-

*) Znak pochodnej cząstkowej w tych równaniach dotyczy tylko bezpośredniej zależności funkcji od $\underline{\psi}$, x lub t, a nie zależności pośredniej przez funkcję $\underline{\varphi}$.

kowej niewiadomej t_k; dodatkowym równaniem jest równanie $\tilde{M} = 0$. Rozwiązania tych równań $\hat{\underline{\psi}}_o$ są właściwymi warunkami początkowymi dla zmiennych sprzężonych, zaś ekstremale $\underline{x}(\underline{x}_o, \hat{\underline{\psi}}_o, t)$, $\underline{\psi}(\underline{x}_o, \hat{\underline{\psi}}_o, t)$ są docelowymi ekstremalami transwersalnymi.

4. Podstawiając do zależności $\hat{u} = \varphi(\underline{\psi}, \underline{x}, t)$ docelowe ekstremale transwersalne uzyskujemy sterowania (jedno lub kilka) spełniając wszystkie warunki zasady maksimum. Jeśli wiemy skądinąd, że problem optymalizacji ma rozwiązanie (np. na podstawie przesłanek fizycznych), i jeśli uzyskamy tylko jedno sterowanie spełniające wszystkie warunki, to jest ono oczywiście sterowaniem optymalnym. Jeśli uzyskamy kilka takich sterowań, to należy je porównać, obliczając dla każdego z nich wskaźnik jakości. Wynika to stąd, że zasada maksimum stanowi tylko warunek konieczny optymalności, a więc pozwala tylko wyznaczyć sterowania "podejrzewane" o optymalność.

W niektórych przypadkach wystarcza przeprowadzenie punktów 1 i 2 powyższej metodyki, bowiem informacje, uzyskane na podstawie tych punktów, wystarczają niekiedy do syntezy zamkniętego układu sterowania.

Jako przykład zastosowania zasady maksimum rozpatrzmy prosty problem sterowania procesu o równaniu

$$\dot{x} = u - x; \qquad x(0) = 0 \qquad (t_o = 0), \tag{381}$$

bez ograniczeń sterowania, z warunkami końcowymi w postaci

$$g_k = x(t_k) - t_k - a = 0; \qquad a = e^2 - 2 \tag{382}$$

i ze wskaźnikiem jakości

$$Q = \frac{1}{2}\int_0^{t_k} u^2 \, dt. \tag{383}$$

Zgodnie z warunkiem (382), pożądany stan końcowy $x(t_k) = t_k + a$ rośnie z czasem; należy więc "dogonić" go możliwie szybko (rys. 68), ale nie zużywając przy tym nadmiernej "energii", na sterowanie, gdyż wskaźnik jakości zależy od u^2.

Hamiltonian problemu ma postać

$$\tilde{H} = -\frac{1}{2}u^2 + \psi(u - x) + \psi_t, \tag{384}$$

a ponieważ

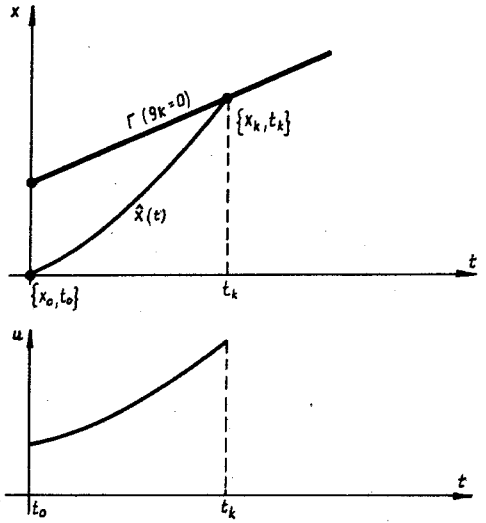
$$\frac{\partial \tilde{H}}{\partial u} = -u + \psi; \qquad \frac{\partial^2 \tilde{H}}{\partial \underline{u}^2} = -1, \tag{385}$$

przeto jedynym sterowaniem ekstremalnym jest

$$\hat{\underline{u}} = \psi, \tag{386}$$

zaś maksimum hamiltonianu ma postać

$$\widetilde{M} = \frac{1}{2}\psi^2 + \psi x + \psi t. \tag{387}$$



Rys. 68

Układ kanoniczny ma postać

$$\begin{aligned} \dot{x} &= \psi - x, \\ \dot{\psi} &= \psi, \\ \dot{\psi}_t &= 0, \end{aligned} \tag{388}$$

zaś ekstremale wyrażają się wzorami

$$\begin{aligned} x &= x_o e^{-t} + \psi_o \operatorname{sht}; \qquad x_o = 0, \\ \psi &= \psi_o e^{t}, \\ \psi_t &= \psi_{to}. \end{aligned} \tag{389}$$

Ponieważ $\frac{\partial g_k}{\partial x_k} = 1$; $\frac{\partial g_k}{\partial t_k} = -1$, przeto równanie rozmaitości stycznej Γ_s (372) ma postać

$$\delta x_k - \delta t_k = 0, \tag{390}$$

(rozmaitość Γ_s pokrywa się tu z rozmaitością Γ, gdyż ta ostatnia jest liniowa). Warunki transwersalności (373) mają natomiast postać

$$\delta x_k \hat{\psi}(t_k) + \delta t_k \hat{\psi}_t(t_k) = 0. \tag{391}$$

Podstawiając $\delta x_k = \delta t_k$ uzyskujemy

$$\delta x_k \left[\hat{\psi}(t_k) + \psi_t(t_k)\right] = 0, \tag{392}$$

a wobec faktu, że δx_k może być niezerowe, zaś $\hat{\psi}_t$ jest stałe i ψ narasta wykładniczo

$$\hat{\psi}_o e^{t_k} + \hat{\psi} t_o = 0. \tag{393}$$

Z warunku, że maksimum rozszerzonego hamiltonianu $\tilde{M}$ jest równe zeru, uzyskany dla $t = 0$ i $x(0) = 0$ (por. 387)

$$\frac{1}{2}\hat{\psi}_o^2 + \hat{\psi} t_o = 0, \tag{394}$$

zaś warunek końcowy (382) po podstawieniu wartości ekstremal przybiera postać

$$\hat{\psi}_o \operatorname{sh} t_k + 2 - e^2 - t_k = 0. \tag{395}$$

Równania (393), (394), (395) zawierają trzy niewiadome $\hat{\psi}_o$, $\hat{\psi}_{to}$, t_k (przy czym wartość $\hat{\psi}_{to}$ nie jest istotna dla wyznaczania sterowania optymalnego). Porównując stronami (393) i (394), uzyskujemy $\hat{\psi}_o = 0$ (rozwiązanie to można odrzucić, gdyż wówczas (395) nie jest nigdy spełnione dla dodatnich t_k lub $\hat{\psi}_o = 2e^{t_k}$); podstawiając to wyrażenie do (395) otrzymujemy równanie przestępne

$$e^{2t_k} - e^2 - t_k + 1 = 0, \tag{396}$$

mające dla dodatnich t_k tylko jedno rozwiązanie $t_k = 1$. Stąd jedyne ekstremale docelowe i transwersalne mają postać

$$\hat{x} = 2e\operatorname{sh} t; \qquad \hat{\psi} = 2ee^t, \tag{397}$$

zaś odpowiadające im sterowanie ekstremalne

$$\hat{u} = 2e \cdot e^{t}, \tag{398}$$

jest sterowaniem optymalnym, zapewniającym minimum wskaźnika jakości

$$Q = e^{2}(e^{2} - 1). \tag{399}$$

Jako drugi przykład rozpatrzymy problem sterowania procesu o równaniach

$$\dot{x}_1 = x_2; \quad \dot{x}_2 = u; \quad x_1(t_o) = x_{10}; \quad x_2(t_o) = x_{20}, \tag{400}$$

przy ograniczonym sterowaniu

$$|u(t)| \leqslant 1, \tag{401}$$

przy warunkach końcowych

$$x_1(t_k) = 0; \quad x_2(t_k) = 0; \quad t_k \text{ - swobodne} \tag{402}$$

oraz wskaźniku jakości, będącym czasem trwania sterowania

$$Q = t_k - t_o = \int_{t_o}^{t_k} 1 \, dt.$$

Jest to problem analogiczny do sformułowanego przykładowo w punkcie 8.1 rys. 61.

Wprowadzamy hamiltonian

$$\tilde{H} = -1 + \psi_1 x_2 + \psi_2 u + \psi_t \tag{403}$$

i poszukując jego maksimum, określamy sterowanie ekstremalne

$$\hat{u} = \operatorname{sign} \psi_2 = \begin{cases} +1, & \text{jeśli} \quad \psi_2 > 0 \\ \text{nieokreślone, jeśli} & \psi_2 = 0 \\ -1, & \text{jeśli} \quad \psi_2 < 0 \end{cases} \tag{404}$$

Z równań sprzężonych

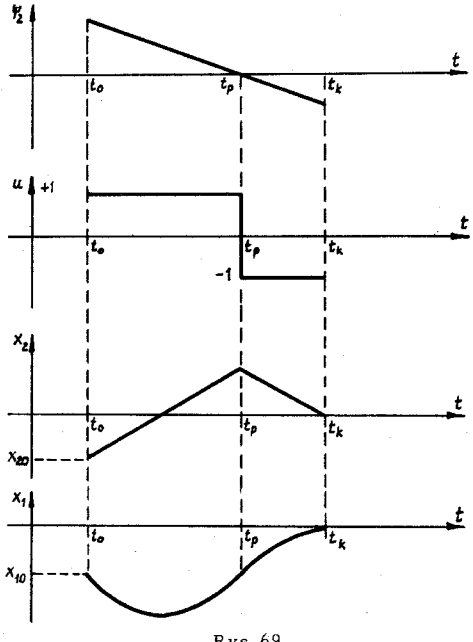
$$\begin{aligned} \dot{\psi}_1 &= -\frac{\partial \tilde{H}}{\partial x_1} = 0, \\ \dot{\psi}_2 &= -\frac{\partial \tilde{H}}{\partial x_2} = -\psi_1, \end{aligned} \tag{405}$$

$$\dot{\psi}_t = -\frac{\partial \tilde{H}}{\partial t} = 0 \qquad (405)$$

oraz z warunku transwersalności $\psi_t(t_k) = 0$ wynikają przebiegi zmiennych sprzężonych

$$\psi_1 = \psi_{10}; \quad \psi_2 = \psi_{20} - \psi_{10}(t - t_o); \quad \psi_t = 0. \qquad (406)$$

Niemożliwe jest przy tym, aby jednocześnie $\psi_{10} = 0$, $\psi_{20} = 0$, gdyż w takim przypadku dla $t = t_o$ wartość maksimum hamiltonianu wyniosłaby $\tilde{M} = -1$, co jest sprzeczne z zasadą maksimum. Zmienna ψ_2 jest więc liniową funkcją czasu, nie równą tożsamościowo zeru, a zatem mogącą zmienić znak tylko raz. Stąd też sterowanie ekstremalne u jako funkcja czasu jest bądź to stałe i równe +1 lub -1, bądź też zmienia znak (ulega przełączeniu) w pewnej chwili t_p - por. rys. 69.



Rys. 69

Rozpatrzmy sterowanie ekstremalne

$$\hat{u}(t) = \begin{cases} +1, & t_o \leqslant t < t_p \\ -1, & t_p \leqslant t \leqslant t_k \end{cases} \tag{407}$$

Całkując równania (400), wyznaczymy odpowiadającą temu sterowaniu trajektorię ekstremalną

$$x_1(t) = \begin{cases} x_{10} + x_{20}(t - t_o) + \frac{1}{2}(t - t_o)^2 & t_o \leqslant t \leqslant t_p \\ x_{10} + x_{20}(t - t_o) + \frac{1}{2}(t_p - t_o)^2 + (t_p - t_o)(t - t_p) - \frac{1}{2}(t - t_p)^2 & t_p \leqslant t \leqslant t_k \end{cases} \tag{408}$$

$$x_2(t) = \begin{cases} x_{20} + t - t_o & t_o \leqslant t \leqslant t_p \\ x_{20} + t_p - t_o - (t - t_p) & t_p \leqslant t \leqslant t_k \end{cases}$$

Trajektoria ta powinna być docelowa, to znaczy spełniać warunki końcowe $x_1(t_k) = 0$, $x_2(t_k) = 0$; warunki te dają nam dwa równania:

$$\begin{aligned} &x_{20} + (t_p - t_o) - (t_k - t_p) = 0, \\ &x_{10} + x_{20}\left[(t_p - t_o) + (t_k - t_p)\right] + \frac{1}{2}(t_p - t_o)^2 + (t_p - t_o)(t_k - t_o) - \frac{1}{2}(t_k - t_o)^2 = 0. \end{aligned} \tag{409}$$

Wyznaczając z pierwszego $(t_k - t_p)$ i podstawiając do drugiego, uzyskujemy równanie kwadratowe

$$(t_p - t_o)^2 + 2x_{20}(t_p - t_o) + x_{10} + \frac{1}{2}x_{20}^2 = 0, \tag{410}$$

w którym odrzucamy pierwiastek ujemny, uzyskując ostatecznie

$$\begin{aligned} t_p - t_o &= -x_{20} + \sqrt{\frac{1}{2}x_{20}^2 - x_{10}}, \\ t_k - t_p &= \sqrt{\frac{1}{2}x_{20}^2 - x_{10}}. \end{aligned} \tag{411}$$

Ponieważ oba te rozwiązania muszą być nieujemne i rzeczywiste, przeto sterowanie ekstremalne o postaci (407) jest docelowe tylko wtedy, gdy

$$\begin{cases} x_{10} \leqslant \frac{1}{2} x_{20}^2 \\ x_{20} \leqslant 0 \end{cases} \quad \text{lub} \quad \begin{cases} x_{10} \leqslant -\frac{1}{2} x_{20}^2 \\ x_{20} \geqslant 0 \end{cases} \tag{412}$$

W przeciwnym przypadku docelowe sterowanie ekstremalne musi mieć wartość -1 dla $t_o \leqslant t < t_p$, zaś $+1$ dla $t_p \leqslant t \leqslant t_k$.

Ponieważ docelowe sterowanie ekstremalne jest jedyne, jest ono jednocześnie sterowaniem optymalnym.

Jak widzimy, w postępowaniu powyższym nie rozwiązywaliśmy jednocześnie układu równań kanonicznych. Ponieważ równania sprzężone są niezależne od czasu i stanu, mogliśmy przeprowadzić tylko ich jakościową analizę, która wystarczyła do wyznaczenia charakteru sterowania optymalnego w układzie otwartym. Wyniki tej analizy wystarczają także dla przeprowadzenia syntezy układu zamkniętego.

Wiemy, że począwszy od momentu przełączenia t_p sterowanie optymalne jest stałe i równe ± 1. Jeśli x_{1p} i x_{2p} są stanem w chwili przełączenia, to dalszy przebieg trajektorii wyrazi się wzorami

$$\begin{aligned} x_1(t) &= x_{1p} + x_{2p}(t - t_p) \pm \frac{1}{2} (t - t_p)^2 , \\ x_2(t) &= x_{2p} \pm (t - t_p) \qquad t_p \leqslant t \leqslant t_k . \end{aligned} \tag{413}$$

Rugując z tych wzorów $(t - t_p)$, otrzymamy równanie trajektorii na płaszczyźnie x_1, x_2

$$x_1 - x_{1p} \mp x_{2p}(x_2 - x_{2p}) \pm \frac{1}{2} (x_2 - x_{2p})^2 = 0. \tag{414}$$

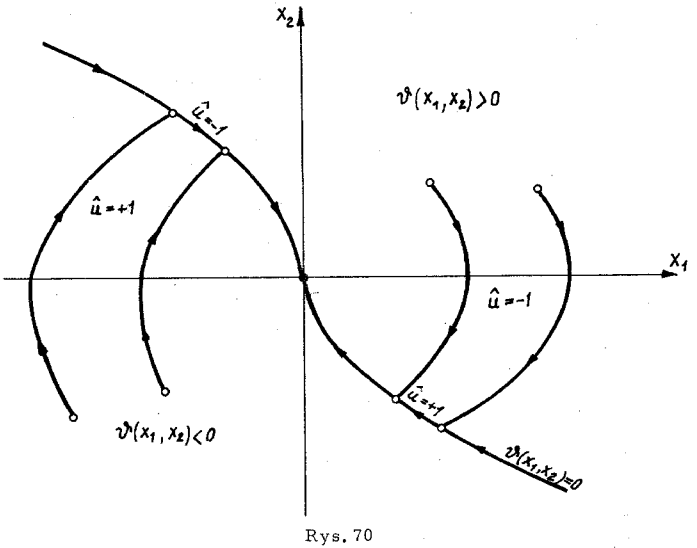
Wybierzemy tu takie trajektorie, które trafiają w punkt końcowy $x_1(t_k) = 0$, $x_2(t_k) = 0$, przy czym oczywiście $u = +1$ dla $x_{2p} < 0$ i przeciwnie, $u = -1$ dla $x_{2p} > 0$. Uzyskujemy wówczas równanie trajektorii docelowych w postaci

$$v(x_{1p}, x_{2p}) = x_{1p} + \frac{1}{2} x_{2p}^2 \operatorname{sign} x_{2p} = 0. \tag{415}$$

Trajektorie te nazywamy linią przełączeń. Dzielą one - por. rys. 70 - płaszczyznę x_1, x_2 na dwa obszary.

Jeśli punkt początkowy x_{10}, x_{20} znajduje się ponad linią przełączeń, to stosując sterowanie $u = +1$ dojdziemy zawsze do

dolnej gałęzi tej linii, gdzie należy zastosować u = -1; odwrotnie, jeśli punkt początkowy znajduje się pod linią przełączeń. Dochodzi-



Rys. 70

my więc do wniosku, że sterowanie optymalne w układzie zamkniętym ma postać

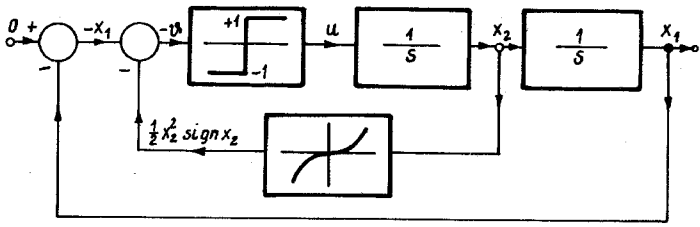
$$\hat{u} = \hat{r}_1(x_1, x_2) = \begin{cases} +1, & \vartheta(x_1, x_2) < 0 \\ -\text{sign } x_2, & \vartheta(x_1, x_2) = 0 \\ -1, & \vartheta(x_1, x_2) > 0. \end{cases} \tag{416}$$

Uwzględniając fakt, że trajektoria po przełączeniu sterowania natychmiast zmienia kierunek - por. rys.70 - można uprościć funkcję syntezującą układ zamknięty

$$\hat{u} = \hat{r}_2(x_1, x_2) = -\text{sign } \vartheta(x_1, x_2). \tag{417}$$

Sterowanie to można zrealizować w układzie o strukturze przedstawionej na rys. 71.

Dokonamy jeszcze analizy funkcji optymalnej jakości $P(x_1, x_2)$ dla powyższego zadania. Funkcja ta nie zależy od t, gdyż $\psi_t = 0$.



Rys. 71

Korzystając ze wzorów (411), (413) oraz z analogicznych wzorów, wyprowadzonych dla przypadku gdy $u = -1$ dla $t_0 \leqslant t < t_p$, uzyskamy

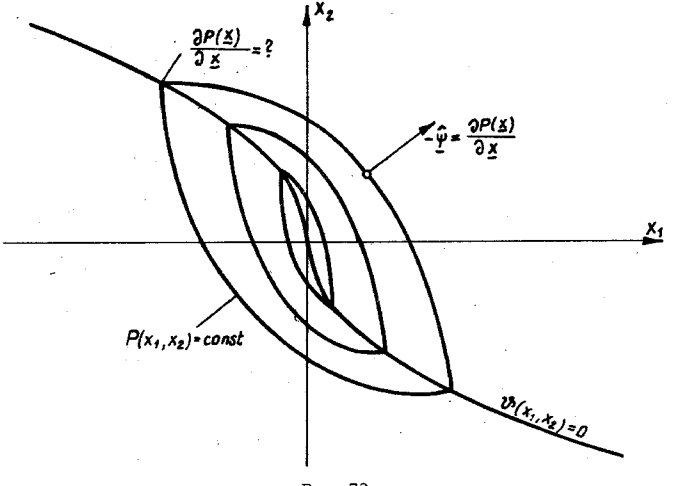
$$P(x_{10}, x_{20}) = \begin{cases} x_{20} + \sqrt{2x_{20}^2 + 4x_{10}}, & \vartheta(x_{10}, x_{20}) > 0 \\ |x_{20}|, & \vartheta(x_{10}, x_{20}) = 0 \\ -x_{20} + \sqrt{2x_{20}^2 - 4x_{10}}, & \vartheta(x_{10}, x_{20}) > 0 \end{cases} \tag{418}$$

Na podstawie tej zależności możemy wyznaczyć równania tzw. izochron minimalnych, czyli miejsca geometrycznego takich wartości początkowych x_{10}, x_{20}, dla których minimalny czas sterowania wartość funkcji $P(x_{10}, x_{20})$ jest stały i równy P:

$$x_{10} = \begin{cases} -\frac{1}{2} x_{20}^2 + \frac{1}{4} (P - x_{20})^2, & \vartheta(x_{10}, x_{20}) > 0 \\ -\frac{1}{2} x_{20} P, & \vartheta(x_{10}, x_{20}) = 0 \\ +\frac{1}{2} x_{20}^2 - \frac{1}{4} (P - x_{20})^2, & \vartheta(x_{10}, x_{20}) < 0 \end{cases} \tag{419}$$

Przebiegi izochron minimalnych przedstawia rys. 72. Jak wiadomo, gradient $\frac{\partial P(\underline{x})}{\partial \underline{x}'} = -\hat{\underline{\psi}}(\underline{x})$ jest wektorem normalnym do linii stałych wartości funkcji $P(x)$. Zauważmy, że gradient ten nie istnieje (nie jest określony) w punktach na linii przełączeń

$\vartheta(x_1, x_2) = 0$, gdzie izochrony mają punkty ostrzowe. W punktach tych funkcja $P(\underline{x})$ nie jest różniczkowalna. Dlatego też ścisłe rozwiązanie powyższego zadania mogło być uzyskane tylko na podstawie zasady maksimum, zaś nie na podstawie równania Hamiltona-Jacobiego-Bellmana.



Rys. 72

8.2.3. Funkcjonał Lagrange'a i wariacje funkcjonału jakości

Dla zadania optymalizacji bez ograniczeń sterowania lub stanu

$$\underline{\dot{x}} = \underline{f}(\underline{x}, \underline{u}, t); \qquad \underline{x}(t_o) = \underline{x}_o, \tag{420}$$

$$\underline{g}_k(\underline{x}(t_k), t_k) = 0, \tag{421}$$

$$Q = f_k\left(\underline{x}(t_k), t_k\right) + \int_{t_o}^{t_k} f_o(\underline{x}, \underline{u}, t)\, dt, \tag{422}$$

funkcjonałem Lagrange'a nazywamy funkcjonał

$$Q_L = f_k\left(\underline{x}(t_k), t_k\right) + \int_{t_o}^{t_k} \left\{ f_o(\underline{x}, \underline{u}, t) + \underline{\lambda}'(t)\left[\dot{\underline{x}} - f(\underline{x}, \underline{u}, t)\right]\right\} dt \tag{423}$$

lub też funkcjonał

$$Q_{L1} = Q_L + \underline{\lambda}_o' \left[\underline{x}(t_o) - \underline{x}_o\right] + \underline{\lambda}_k' \underline{g}_k \left(\underline{x}(t_k),\ t_k\right), \qquad (424)$$

gdzie: $\underline{\lambda}(t)$ - wektor funkcji-mnożników Lagrange'a związanych z równaniami stanu,

$\underline{\lambda}_o$, $\underline{\lambda}_k$ - dodatkowe mnożniki Lagrange'a związane z warunkami początkowymi i końcowymi.

Podobne funkcjonały Lagrange'a można sformułować dla zadań z ograniczeniami sterowania czy nawet stanu, wprowadzając dodatkowe mnożniki Lagrange'a związane z tymi ograniczeniami.

Zauważmy, że wartości funkcjonału Lagrange'a różnią się od wartości pierwotnego funkcjonału jakości tylko wtedy, gdy równania stanu, warunki początkowe i końcowe (i ewentualnie inne ograniczenia czy warunki) nie są spełnione. Z funkcjonałami Lagrange'a związane jest ważne twierdzenie o punkcie siodłowym *), które podamy tu w znacznym uproszczeniu.

> Jeśli pewien funkcjonał (422) ma minimum w punkcie $\hat{x}$, $\hat{u}$ względem zmiennych x, u przy warunkach (420), (421), to istnieje taki funkcjonał Lagrange'a, który ma w tym samym punkcie minimum bezwarunkowe względem zmiennych x, u, zaś maksimum względem zmiennych $\underline{\lambda}$.

Zajmiemy się dokładniej funkcjonałem Lagrange'a o postaci (423). Można wykazać, że funkcje - mnożniki Lagrange'a $\underline{\lambda}(t)$ są identyczne ze zmiennymi sprzężonymi $\underline{\psi}(t)$

$$\underline{\lambda}(t) \equiv \underline{\psi}(t). \qquad (425)$$

Przyjmując powyższą zależność bez dowodu, pokażemy jak na podstawie funkcjonału Lagrange'a można obliczyć przyrosty lub wariacje funkcjonału jakości. Dla uproszczenia postaci warunków końcowych założymy, że czas t_k jest dany, zaś stan $\underline{x}(t_k)$ - swobodny. Jeśli równania stanu są spełnione, to funkcjonał jakości

$$Q = f_k\left(\underline{x}(t_k)\right) + \int_{t_o}^{t_k} f_o(\underline{x},\ \underline{u},\ t)\,dt, \qquad (426)$$

może być zapisany w postaci funkcjonału Lagrange'a

$$Q = Q_L = f_k\left(\underline{x}(t_k)\right) + \int_{t_o}^{t_k} \left\{ f_o(\underline{x},\ \underline{u},\ t) + \underline{\psi}'\left[\dot{\underline{x}} - \underline{f}(\underline{x},\ \underline{u},\ t)\right]\right\} dt. \qquad (427)$$

---

*) Punkt siodłowy pewnej funkcji lub funkcjonału jest to punkt, w którym ta funkcja lub funkcjonał ma minimum względem jednej zmiennej, zaś maksimum względem innej.

Wykorzystując pojęcie hamiltonianu i całkując przez części, zapisujemy funkcjonał Lagrange'a w postaci

$$Q_L = f_k\left(\underline{x}(t_k)\right) + \int_{t_0}^{t_k} \left\{ \underline{\psi}' \cdot \dot{\underline{x}} - H(\underline{\psi}, \underline{x}, \underline{u}, t) \right\} dt =$$

$$= f_k\left(\underline{x}(t_k)\right) + \left[\underline{\psi}'\underline{x}\right]_{t_0}^{t_k} - \int_{t_0}^{t_k} \left\{ \dot{\underline{\psi}}'\underline{x} + H(\underline{\psi}, \underline{x}, \underline{u}, t) \right\} dt. \qquad (428)$$

Załóżmy, że wszystkie funkcje występujące w zadaniu są różniczkowalne względem $\underline{x}$ i $\underline{u}$.

Załóżmy, że dane jest wybrane sterowanie $\underline{u}_1$, niekoniecznie optymalne. Wybierzmy dowolną przedziałami ciągłą funkcję czasu $\delta\underline{u}$, którą nazwiemy wariacją sterowania i utwórzmy nowe sterowanie $\underline{u}_2 = \underline{u}_1 + \varepsilon\delta\underline{u}$, gdzie ε - liczba dowolnie mała. Trajektoria stanu $\underline{x}_1$, odpowiadająca sterowaniu $\underline{u}_1$, zmieni się wówczas na trajektorię $\underline{x}_2 = \underline{x}_1 + \varepsilon\delta\underline{x} + o(\varepsilon)$, przy czym wariację stanu $\delta\underline{x}$ można wyznaczyć przez różniczkowanie równania stanu

$$\delta\dot{\underline{x}} = \frac{\partial \underline{f}(\underline{x}_1, \underline{u}_1, t)}{\partial \underline{x}} \delta\underline{x} + \frac{\partial \underline{f}(\underline{x}_1, \underline{u}_1, t)}{\partial \underline{u}} \delta\underline{u}; \quad \delta\underline{x}(t_o) = \underline{0}. \qquad (429)$$

Funkcjonał jakości zmieni się od wartości Q_1, odpowiadającej $\underline{u}_1$ i $\underline{x}_1$ do wartości

$$Q_2 = Q_1 + \varepsilon\delta Q + o(\varepsilon), \qquad (430)$$

przy czym wariację funkcjonału δQ dogodniej jest wyznaczyć przez różniczkowanie funkcjonału Lagrange'a (428)

$$\delta Q = \left[\frac{\partial f_k\left(\underline{x}_1(t_k)\right)}{\partial \underline{x}(t_k)} + \underline{\psi}'(t_k)\right] \delta\underline{x}(t_k) - \int_{t_0}^{t_k} \left[\frac{\partial H(\underline{\psi}, \underline{x}_1, \underline{u}_1, t)}{\partial \underline{x}} + \dot{\underline{\psi}}'_1\right] \delta x \, dt -$$

$$- \int_{t_0}^{t_k} \frac{\partial H(\underline{\psi}, \underline{x}_1, \underline{u}_1, t)}{\partial \underline{u}} \delta\underline{u} \, dt. \qquad (431)$$

Zauważmy, że jeśli spełnione są warunki transwersalności w postaci (374), to znika pierwszy składnik tej wariacji. Jeśli spełnione są równania sprzężone, to znika pierwszy składnik pod całką i uzyskujemy

$$\delta Q = - \int_{t_0}^{t_k} \frac{\partial H(\underline{\psi}, \underline{x}_1, \underline{u}_1, t)}{\partial \underline{u}} \delta\underline{u} \, dt. \qquad (432)$$

Wektorową funkcję czasu $\underline{\gamma}(t)$, która mnożona skalarnie przez wariację $\delta \underline{u}$ i całkowana w granicach t_0, t_k daje wariację funkcjonału, nazywamy gradientem funkcjonału. Widzimy, że gradient funkcjonału jakości jest równy gradientowi hamiltonianu ze zmienionym znakiem

$$\underline{\gamma} = - \frac{\partial H(\underline{\psi}, \underline{x}, \underline{u}, t)}{\partial \underline{u}'} \tag{433}$$

Jeśli sterowanie jest optymalne, to nie może być poprawione (wariacja δQ nie może być ujemna) przez wariację sterowania o dowolnym przebiegu i znaku. Wynika stąd, że gradient funkcjonału obliczony wzdłuż trajektorii optymalnej musi być równy zeru *)

$$\hat{\underline{\gamma}}(t) = - \frac{\partial H(\hat{\underline{\psi}}, \hat{\underline{x}}, \hat{\underline{u}}, t)}{\partial \underline{u}'} = \underline{0}. \tag{434}$$

Jest to warunek konieczny optymalności (równoważny zresztą wnioskom z zasady maksimum przy założeniu różniczkowalności hamiltonianu). Stąd też nie każde sterowanie i trajektoria, które spełniają ten warunek, muszą być optymalne. Sterowania i trajektorie, spełniające ten warunek, nazwiemy stacjonarnymi. Z pojęciem tym wiąże się ważne twierdzenie, zwane podstawowym lematem rachunku wariacyjnego.

> Jeśli dla pewnego problemu optymalizacji bez ograniczeń, o różniczkowalnym funkcjonale jakości dana jest pewna niestacjonarna trajektoria procesu, to można zawsze skonstruować trajektorię od niej lepszą, to jest zapewniającą zmniejszenie wskaźnika procesu.

Twierdzenie to wynika w sposób oczywisty ze wzoru (432). Jeśli bowiem $\frac{\partial H}{\partial \underline{u}} \neq \underline{0}$, to wystarczy przyjąć $\delta \underline{u} = \frac{\partial H}{\partial \underline{u}'}$, żeby wyrażenie pod całką było zawsze dodatnie, zaś wariacja funkcjonału jakości - ujemna. Twierdzenie to ma dość szerokie zastosowanie w metodach obliczeniowych optymalizacji dynamicznej.

Wariacja funkcjonału określa pierwsze, liniowe przybliżenie przyrostu funkcjonału. Chcąc oszacować dokładniej przyrost funkcjonału, zakładamy, że jest on dwukrotnie różniczkowalny i przedstawiamy go w postaci

$$Q_2 - Q_1 = \varepsilon \delta Q + \frac{1}{2} \varepsilon^2 \delta^2 Q + o(\varepsilon^2), \tag{435}$$

---

*) Oczywiście w przypadku, gdy gradient ten istnieje (funkcjonał, a więc hamiltonian jest różniczkowalny względem $\underline{u}$) i gdy brak jest ograniczeń sterowania (bo tylko wówczas dopuszczalne wariacje $\delta \underline{u}$ są dowolnego znaku).

gdzie $\delta^2 Q$ jest drugą wariacją funkcjonału; można wykazać, że wyraża się ona wzorem

$$\delta^2 Q = \left[\delta \underline{x}'(t_k) \frac{\partial^2 f_k(\underline{x}(t_k))}{\partial \underline{x}' \partial \underline{x}} + \delta \underline{\psi}'(t_k) \right] \delta \underline{x}(t_k) -$$

$$- \int_{t_0}^{t_k} \left[\delta \underline{x}' \frac{\partial^2 H(\ldots)}{\partial \underline{x}' \partial \underline{x}} + \delta \underline{u}' \frac{\partial^2 H(\ldots)}{\partial \underline{u}' \partial \underline{x}} + \delta \underline{\psi} \frac{\partial^2 H(\ldots)}{\partial \underline{\psi}' \partial \underline{x}} + \right. \tag{346}$$

$$\left. + \delta \dot{\underline{\psi}}' \right] \delta \underline{x} \, dt - \int_{t_0}^{t_k} \left[\delta \underline{x}' \frac{\partial^2 H(\ldots)}{\partial \underline{x}' \partial \underline{u}} + \delta \underline{u}' \frac{\partial^2 H(\ldots)}{\partial \underline{u}' \partial \underline{u}} + \right.$$

$$\left. + \delta \underline{\psi}' \frac{\partial^2 H(\ldots)}{\partial \underline{\psi}' \partial \underline{u}} \right] \delta \underline{u} \, dt ,$$

gdzie dla skrócenia oznaczeń symbolem $(\ldots)$ oznaczono zależność od $\underline{x}_1(t_k)$ w przypadku funkcji f_k, zaś od $\underline{\psi}$, $\underline{x}_1$, $\underline{u}_1$, t w przypadku Hamiltonianu.

Symbolami $\frac{\partial^2 f_k}{\partial \underline{x}' \partial \underline{x}}$ oraz $\frac{\partial^2 H}{\partial \underline{x}' \partial \underline{x}}$, $\frac{\partial^2 H}{\partial \underline{\psi}' \partial \underline{x}}$ ... itp. oznaczono tu macierze drugich pochodnych cząstkowych odpowiednich funkcji względem zaznaczonych w mianowniku zmiennych. Wariacja $\delta \underline{u}$ jest tu, jak poprzednio, dowolną funkcją czasu, wariacja $\delta \underline{x}$ wynika z równań (429), które można też zapisać w postaci

$$\delta \dot{\underline{x}} = \frac{\partial^2 H(\ldots)}{\partial \underline{\psi}' \partial \underline{x}} \delta \underline{x} + \frac{\partial^2 H(\ldots)}{\partial \underline{\psi}' \partial \underline{u}} \delta \underline{u}; \quad \delta \underline{x}(t_0) = \underline{0}, \tag{437}$$

zaś wariacja $\delta \underline{\psi}$ odpowiada różnicy pomiędzy dwoma przebiegami zmiennych sprzężonych; jeśli dobierzemy ją tak, by spełniała równania

$$\delta \dot{\underline{\psi}} = - \frac{\partial^2 H(\ldots)}{\partial \underline{x}' \partial \underline{x}} \delta \underline{x} - \frac{\partial^2 H(\ldots)}{\partial \underline{x}' \partial \underline{\psi}} \delta \underline{\psi} - \frac{\partial^2 H(\ldots)}{\partial \underline{x}' \partial \underline{u}} \delta \underline{u}, \tag{438}$$

przy warunkach końcowych

$$\delta \underline{\psi}(t_k) = - \frac{\partial^2 f_k(\ldots)}{\partial \underline{x}' \partial \underline{x}} \delta \underline{x}(t_k), \tag{439}$$

to dwie pierwsze składowe drugiej wariacji funkcjonału (436) są równe zeru. Ponieważ jednocześnie można dobrać sam przebieg $\underline{u}$ tak, by dwie pierwsze składowe pierwszej wariacji funkcjonału

(431) były równe zeru, przeto przyrost funkcjonału (435) można przedstawić w postaci

$$Q_2 - Q_1 = -\varepsilon \int_{t_0}^{t_k} \frac{\partial H(\ldots)}{\partial \underline{u}}\, \delta \underline{u}\, dt - \tag{440}$$

$$-\frac{1}{2}\varepsilon^2 \int_{t_0}^{t_k} \left[\delta \underline{x}' \frac{\partial^2 H(\ldots)}{\partial \underline{x}'\, \partial \underline{u}} + \delta \underline{\psi}' \frac{\partial^2 H(\ldots)}{\partial \underline{\psi}'\, \partial \underline{u}} + \delta \underline{u}' \frac{\partial^2 H(\ldots)}{\partial \underline{u}'\, \partial \underline{u}} \right] \delta \underline{u}\, dt + o(\varepsilon^2).$$

Powyższe wyrażenie na przyrost funkcjonału także znajduje zastosowanie w metodach obliczeniowych optymalizacji.

8.3. Warianty specjalne problemu

8.3.1. Uwagi ogólne

Rozpatrywaliśmy dotychczas wariant podstawowy problemu optymalizacji - gdy równania stanu są równaniami różniczkowymi zwyczajnymi, w których nie występowały opóźnienia, ograniczenia dotyczą co najwyżej sterowania, a wskaźnik jakości jest sumą funkcji stanu końcowego oraz całki z pewnej funkcji stanu i sterowania.

Nie będziemy rozpatrywać wszystkich możliwych wariantów specjalnych. Rozmaite postaci wskaźnika jakości przedyskutowano np. w [16]; ograniczenia stanu i ograniczenia o postaci całkowej w [15] i [16]; opóźnienia stanu w [15], zaś opóźnienia sterowania w [17].

Szczególnie ważny wariant specjalny, który tu rozpatrzymy, dotyczy przypadku dyskretnego w czasie, gdy równania stanu są równaniami różnicowymi zwyczajnymi.

8.3.2. Wariant dyskretny problemu

W wariancie dyskretnym problemu optymalizacji równania stanu mają postać

$$\underline{x}[k+1] = \underline{f}(\underline{x}[k],\ \underline{u}[k],\ k), \tag{441}$$

gdzie k jest czasem dyskretnym, czyli liczbą całkowitą, oznaczającą numer kolejnej chwili czasu (może ona być np. związana z czasem zależnością $t = k \cdot T_p$, przy czym T_p - jednostka dyskretyzacji czasu).

Nawiasy kwadratowe w równaniu (441) podkreślają, że stan $\underline{x}$ i sterowanie $\underline{u}$ są dyskretnymi funkcjami czasu (to znaczy że nie

interesują nas wartości $\underline{x}$ i $\underline{u}$ dla czasów innych, niż np. całkowita wielokrotność jednostki T_p).

Równania tego typu mają między innymi procesy sterowane za pomocą urządzeń cyfrowych (ze względu na dyskretny charakter pracy tak przetworników pomiarowych, jak i urządzeń wykonawczych, sterujących bezpośrednio proces). Często też posługujemy się równaniem tego typu jako przybliżeniem równania różniczkowego dla celów obliczeń numerycznych.

Będziemy zakładali, że dany jest stan początkowy procesu

$$\underline{x}[k_o] = \underline{x}_o, \tag{442}$$

oraz warunki końcowe w postaci danej chwili końcowej k_k i swobodnego stanu końcowego $\underline{x}[k_k]$ (inne postaci warunków końcowych, a zwłaszcza swobodny czas końcowy k_k utrudniają rozwiązanie problemu). Załóżmy ponadto, że ograniczenia dotyczą tylko sterowania

$$g_1(\underline{u}[k]) \leqslant \underline{0}; \qquad \underline{u}[k] \in \Omega \tag{443}$$

rozróżniając przy tym dwa istotne przypadki: gdy ograniczenia są spełnione w sposób silnie nierównościowy ($<$) czyli sterowanie należy do wnętrza obszaru Ω, i gdy ograniczenia są spełnione w sposób równościowy, czyli sterowanie należy do brzegu obszaru Ω.

Wskaźnik jakości procesu ma postać

$$Q = f_k \underline{x}(t_k) + \sum_{k=k_0}^{k_k-1} f_o(\underline{x}[k], \underline{u}[k], k). \tag{444}$$

Zakładamy ponadto, że wszystkie funkcje występujące w zadaniu są różniczkowalne względem $\underline{x}$ i $\underline{u}$. Szukamy takiego sterowania dopuszczalnego (wszystkie sterowania są docelowe ze względu na postać warunków końcowych), które zapewnia minimum wskaźnika Q.

Zauważmy najpierw, że mamy tu w istocie rzeczy do czynienia ze złożonym i specyficznym problemem optymalizacji statycznej. Istotnie, jeśli $\dim \underline{u} = m$ i $\dim \underline{x} = n$, to wskaźnik jakości jest funkcją $m(k_k - k_o)$ zmiennych $u_i[k]$, $k = k_o, \ldots k_k - 1$ oraz $n(k_k - k_o)$ zmiennych $x_i[k]$, $k = k_o + 1, \ldots k_k$. Równania stanu stanowią $n(k_k - k_o)$ dodatkowych warunków między tymi zmiennymi; ponadto dochodzą jeszcze warunki, wynikające z ograniczeń (443). Równoważny problem optymalizacji statycznej ma więc bardzo dużą wymiarowość i nie opłaca się go zwykle rozwiązywać metodami optymalizacji statycznej; znacznie dogodniejsze są tu metody, wynikające z zasady maksimum czy zasady optymalności.

Dla sformułowania zasady maksimum wprowadzamy, jak zwykle, hamiltonian

$$H(\underline{\psi}[k],\ \underline{x}[k],\ \underline{u}[k],\ k) =$$

$$= -f_o(x[k],\ u[k],\ k) + \underline{\psi}'[k]\underline{f}(\underline{x}[k],\ \underline{u}[k],\ k) \qquad (445)$$

oraz równanie sprzężone

$$\underline{\psi}[k-1] = \frac{\partial H(\underline{\psi}[k],\ \underline{x}[k],\ \underline{u}[k],\ k)}{\partial \underline{x}'} \qquad (446)$$

Zauważmy, że w porównaniu z przypadkiem ciągłym w równaniach sprzężonych nie występuje znak minus, a jednocześnie zostaje niejako zmieniony znak upływu czasu - gdyż na podstawie wartości $\underline{\psi}[k]$, $\underline{x}[k]$, $\underline{u}[k]$ wyznaczamy wcześniejszą wartość $\underline{\psi}[k-1]$ *).

Dla założonej postaci warunków końcowych uzyskujemy znaną już postać warunków transwersalności (374), która w przypadku dyskretnym wyraża się wzorem

$$\underline{\psi}[k_k - 1] = -\frac{\partial f_k(\underline{x}[k_k])}{\partial \underline{x}'} . \qquad (447)$$

Zasada maksimum w przypadku dyskretnym ma następujące brzmienie:

Jeśli $\hat{u}$ jest sterowaniem dopuszczalnym i optymalnym, to istnieje taka dyskretna funkcja czasu $\hat{\underline{\psi}}$, że:

a) jeśli $\hat{u}[k]$ należy do brzegu obszaru Ω, to zapewnia ono maksimum hamiltonianu

$$H(\hat{\underline{\psi}}[k],\ \hat{\underline{x}}[k],\ \hat{\underline{u}}[k],\ k) = \max_{\underline{u} \in \Omega} H(\hat{\underline{\psi}}[k],\ \hat{\underline{x}}[k],\ \underline{u}[k],\ k); \qquad (448)$$

b) jeśli $\hat{u}[k]$ należy do wnętrza obszaru Ω, to zapewnia ono punkt stacjonarny hamiltonianu

$$\frac{\partial H(\hat{\underline{\psi}}[k],\ \hat{\underline{x}}[k],\ \hat{\underline{u}}[k],\ k)}{\partial \underline{u}} = \underline{0}. \qquad (449)$$

Zauważmy, że sformułowanie punktu b) jest lokalne i słabe; sterowanie optymalne wcale nie musi zapewniać maksimum hamiltonianu, może to być równie dobrze minimum lub punkt przegię-

*) W istocie rzeczy we wszystkich zadaniach optymalizacji zmieniony kierunek upływu czasu jest naturalnym kierunkiem rozwiązywania równań sprzężonych; widoczne to jest jednak wyraźnie dopiero w przypadku dyskretnym lub w przypadkach z opóźnieniem.

cia; może to być także maksimum lokalne, podczas gdy w zasadzie maksimum obowiązującej dla procesów ciągłych, sterowanie optymalne zapewnia maksimum globalne hamiltonianu.

Zauważmy dalej, że w sformułowaniu zasady maksimum w przypadku dyskretnym nie mówimy nic o zachowaniu się maksimum hamiltonianu (podczas gdy w przypadku ciągłym maksimum hamiltonianu rozszerzonego było stałe i równe zeru). Poza powyższymi różnicami metodyka zastosowania zasady maksimum w przypadku dyskretnym jest analogiczna do metodyki wykorzystywanej w przypadku ciągłym. Ze względu na przeciwny bieg czasu w równaniach sprzężonych, pewną trudność w rozwiązywaniu analitycznym kanonicznego układu równań może stanowić jedynie przekształcenie tego układu do takiej postaci, by można go było rozwiązywać w jednym kierunku czasu.

Rozpatrzmy prosty przykład zadania optymalizacji o równaniach stanu

$$x[k+1] = x[k] + u[k]; \qquad x[0] = 1, \qquad (k_o = 0), \tag{450}$$

gdzie $u[k]$ - nieograniczone, zaś wskaźnik jakości ma postać

$$Q = \frac{1}{2}\left\{x^2[4] + \sum_{k=0}^{3}\left(u^2[k] + x^2[k]\right)\right\}, \tag{451}$$

przy czym $x[4]$ - swobodne $(k_k = 4)$.

Na podstawie hamiltonianu problemu

$$H = -\frac{1}{2}u^2[k] - \frac{1}{2}x^2[k] + \psi[k](x[k] + u[k]), \tag{452}$$

wyznaczamy sterowanie stacjonarne, zapewniające spełnienie warunku $\frac{\partial H}{\partial u} = 0$ (w tym przypadku sterowanie takie jest tylko jedno, i zapewnia jednocześnie maksimum hamiltonianu)

$$\frac{\partial H}{\partial u} = -u[k] + \psi[k]; \qquad \hat{u}[k] = \psi[k]. \tag{453}$$

Równanie sprzężone

$$\psi[k-1] = \frac{\partial H}{\partial x} = -x[k] + \psi[k], \tag{454}$$

przekształcamy do takiej postaci, by można było wyznaczyć $\psi[k+1]$ na podstawie $\psi[k]$

$$\psi[k+1] = \psi[k] + x[k+1]. \tag{455}$$

Podstawiając sterowanie stacjonarne (453) do równania stanu (450), uzyskujemy

$$x[k+1] = x[k] + \psi[k], \tag{456}$$

zaś podstawiając $x[k+1]$ do równania (455) mamy

$$[k+1] = x[k] + 2\psi[k]. \tag{457}$$

Równania (456), (457) stanowią układ równań kanonicznych, które można rozwiązać w normalnym kierunku. Wprowadzimy oznaczenie wektorowe $\underline{z} = \{x, \psi\}$ i zapiszemy ten układ równań w postaci

$$\underline{z}[k+1] = \underline{A}\,\underline{z}[k]; \qquad \underline{A} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \tag{458}$$

Rozwiązanie ogólne takiego liniowego jednorodnego równania różniczkowego ma postać

$$\underline{z}[k] = \underline{A}^k\,\underline{z}[0], \tag{459}$$

przy czym

$$\underline{A}^2 = \begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix}; \qquad \underline{A}^3 = \begin{bmatrix} 5 & 8 \\ 8 & 13 \end{bmatrix}; \qquad \underline{A}^4 = \begin{bmatrix} 13 & 21 \\ 21 & 34 \end{bmatrix} \tag{460}$$

Wśród składowych wektora $\underline{z}[0] = \{\underline{x}[0], \underline{\psi}[0]\}$ nie znamy jeszcze $\psi[0]$. Inaczej mówiąc, równanie (459) określa ekstremale problemu, ale musimy wyznaczyć ekstremalę transwersalną.

Z warunku transwersalności (447) otrzymamy

$$\psi[3] = -x[4], \tag{461}$$

zaś z równań (459), (460)

$$\psi[3] = 8x[0] + 13\psi[0]; \qquad x[4] = 13x[0] + 21\psi[0], \tag{462}$$

co po podstawieniu do (461) i rozwiązaniu względem $\psi[0]$ daje

$$\hat{\psi}[0] = -\frac{21}{34}x[0] = -\frac{21}{34}. \tag{463}$$

Teraz na podstawie (459), (460) łatwo wyznaczymy ekstremalę transwersalną w postaci tabeli

k	0	1	2	3	4
x	1	$\frac{13}{34}$	$\frac{5}{34}$	$\frac{2}{34}$	$\frac{1}{34}$
u =	$-\frac{21}{34}$	$-\frac{8}{34}$	$-\frac{3}{34}$	$-\frac{2}{34}$	-

Ponieważ można udowodnić, że rozwiązanie zadania istnieje, przeto jedyna ekstremala transwersalna musi stanowić trajektorię optymalną.

Zauważmy, że gdyby powyższe zadanie rozwiązywać metodami optymalizacji statycznej, Q byłoby funkcją 8 zmiennych $u[0]$, $u[1]$, $u[2]$, $u[3]$ i $x[1]$, $x[2]$, $x[3]$, $x[4]$ przy czterech warunkach równościowych odpowiadających równaniu stanu na kolejnych etapach.

Dla procesów problemów dyskretnych można również sformułować funkcjonały Lagrange'a oraz obliczać pierwszą wariację funkcjonału, gradient funkcjonału, drugą wariację itp. Na przykład pierwsza wariacja funkcjonału (444) przy warunkach (441) i bez ograniczeń sterowania ma postać

$$\delta Q = \left\{ \underline{\psi}'[k_k - 1] + \frac{\partial f_k(\underline{x}[k_k])}{\partial \underline{x}} \right\} \delta \underline{x}[k_k] -$$

$$- \sum_{K=k_0}^{K=k_k-1} \left\{ \left[\frac{\partial H(\underline{\psi}[k],\ \underline{x}[k],\ \underline{u}[k],\ k)}{\partial \underline{x}} - \underline{\psi}'[k-1] \right] \partial \underline{x}[k] + \right.$$

$$\left. + \frac{\partial H(\underline{\psi}[k],\ \underline{x}[k],\ \underline{u}[k],\ k)}{\partial \underline{u}}\ \delta \underline{u}[k] \right\}, \qquad (464)$$

przy czym zmiana sterowania wynosi $\varepsilon \delta \underline{u}[k]$ - gdzie $\delta \underline{u}[k]$ - dowolna wariacja sterowania, zaś wariację stanu $\delta \underline{x}[k]$ wyznacza się z równań

$$\delta \underline{x}[k+1] = \frac{\partial \underline{f}(\underline{x}[k],\ \underline{u}[k],\ k)}{\partial \underline{x}}\ \delta x[k] + \frac{\partial \underline{f}(\underline{x}[k],\ \underline{u}[k],\ k)}{\partial \underline{u}}\ \delta \underline{u}[k]. \qquad (465)$$

Jeśli teraz tak dobierzemy funkcję $\underline{\psi}[k]$, by spełniała warunki transwersalności (447) i równania sprzężone (446), to wariacja (463) uprości się do postaci

$$\delta Q = - \sum_{K=k_0}^{K_k-1} \frac{\partial H(\underline{\psi}[k],\ \underline{x}[k],\ \underline{u}[k],\ k)}{\partial \underline{u}}\ \delta \underline{u}[k]\ . \qquad (466)$$

Widzimy więc, że - tak samo jak w przypadku ciągłym - gradientem funkcjonału jest gradient hamiltonianu ze zmienionym znakiem. Zmiana niestacjonarnego sterowania w kierunku wzrostu hamiltonianu gwarantuje więc zmniejszanie się wskaźnika jakości (mimo, że w zasadzie maksimum dla procesów dyskretnych sterowanie optymalne niekoniecznie odpowiada maksimum hamilto-

nianu *), Podstawowy lemat o ulepszaniu trajektorii nieoptymalnej i niestacjonarnej ma więc w przypadku dyskretnym takie samo sformułowanie, jak w przypadku ciągłym.

Do szczególnie silnych rezultatów prowadzi zastosowanie zasady optymalności Bellmana dla optymalizacji procesów dyskretnych. Zgodnie z zasadą optymalności, końcowy odcinek trajektorii optymalnej jest sam dla siebie trajektorią optymalną. Rozpatrzmy więc ostatni etap (od k_k-1 do k_k) optymalizacji wskaźnika jakości (444) przy równaniach stanu (441) i ograniczeniu sterowania $\underline{u} \in \Omega$. Wskaźnik jakości na ostatnim etapie wyniesie

$$Q_1 = f_0(\underline{x}[k_k-1],\ \underline{u}[k_k-1],\ k_k-1) + f_k(\underline{x}[k_k]) \quad , \tag{467}$$

przy czym zamiast $\underline{x}[k_k]$ możemy podstawić

$$\underline{x}[k_k] = \underline{f}(\underline{x}[k_k-1],\ \underline{u}[k_k-1],\ k_k-1). \tag{468}$$

Otrzymamy wówczas Q_1 jako funkcję zmiennych $\underline{x}[k_k-1]$, $\underline{u}[k_k-1]$. Jeśli znajdziemy jej minimum względem $\underline{u}[k_k-1] \in \Omega$ dla każdego $\underline{x}[k_k-1]$, to wyznaczymy funkcję jakości optymalnej

$$P(\underline{x}[k_k-1],\ k_k-1) =$$

$$= \min_{\underline{u}[k_k-1] \in \Omega} \Big\{ f_0(\underline{x}[k_k-1],\ \underline{u}[k_k-1],\ k_k-1) +$$

$$+ f_k(\underline{x}[k_k-1],\ \underline{u}[k_k-1],\ k_k-1) \Big\} \tag{469}$$

oraz sterowanie, zapewniające to minimum, czyli sterowanie optymalne w układzie zamkniętym dla ostatniego etapu

$$\underline{\hat{u}}[k_k-1] = \underline{\hat{f}}(\underline{x}[k_k-1],\ k_k-1). \tag{470}$$

Cofnijmy się teraz o jeden etap. Ponieważ na ostatnim etapie dokonaliśmy już optymalizacji, przeto wskaźnik jakości dla dwóch ostatnich etapów możemy wyrazić wzorem

$$Q_2 = f_0(\underline{x}[k_k-2],\ \underline{u}[k_k-2],\ k_k-2) + P(\underline{x}[k_k-1],\ k_k-1), \tag{471}$$

---

*) Sprzeczność ta jest tylko pozorna, jednak jej dokładniejsza interpretacja wykracza poza ramy niniejszego skryptu.

do którego to wzoru znów postawimy

$$\underline{x}\left[k_k-1\right] = \underline{f}(\underline{x}\left[k_k-2\right],\ \underline{u}\left[k_k-2\right],\ k_k-2), \tag{472}$$

przeprowadzimy minimalizację względem $\underline{u}\left[k_k-2\right] \in \Omega$ dla dowolnego $\underline{x}\left[k_k-2\right]$, wyznaczając $P(\underline{x}\left[k_k-2\right],\ k_k-2)$ oraz $\underline{\hat{u}}\left[k_k-2\right] = \underline{\hat{f}}(\underline{x}\left[k_k-2\right],\ k_k-2)$ i tak dalej aż do etapu początkowego k_o. Możemy więc zapisać ogólny przepis optymalizacji, zwany algorytmem programowania dynamicznego, w postaci

$$P(\underline{x}\left[k\right], k) = \min_{\underline{u}\left[k\right] \in \Omega} \left\{ f_o(\underline{x}\left[k\right],\ \underline{u}\left[k\right],\ k) + P\left(\underline{f}(\underline{x}\left[k\right],\ \underline{u}\left[k\right],\ k),\ k+1\right) \right\} \tag{473}$$

gdzie

$$k = k_k-1,\ k_k-2,\ \ldots\ k_o; \quad P(\underline{x}\left[k_k\right],\ k_k) = f_k(\underline{x}\left[k_k\right]) \tag{474}$$

oraz gdzie na każdym etapie wyznaczamy sterowanie optymalne w układzie zamkniętym

$$\underline{\hat{u}}\left[k\right] = \underline{\hat{f}}(x\left[k\right],\ k). \tag{475}$$

Algorytm ten jest bardzo ogólny; zauważmy, że dla jego wprowadzenia skorzystaliśmy tylko z zasady optymalności, a nie zakładaliśmy nic o własnościach funkcji f_o, $\underline{f}_k$, f. Funkcje te mogą więc być nieciągłe, dane w postaci tabel itp. - byleśmy tylko mogli przeprowadzić operację poszukiwania minimum we wzorze (473). Mimo tej ogromnej zalety, algorytm ten nie jest dogodny dla obliczeń numerycznych sterowania optymalnego, gdyż musimy w nim wyznaczać sterowanie optymalne jako funkcję stanu na każdym etapie, a więc przeprowadzać obliczenia dla każdej przewidywanej wartości stanu układu. Prowadzi to do ogromnego nakładu obliczeń, zwłaszcza dla dużych wymiarów stanu n. Algorytm programowania dynamicznego można również stosować dla obliczeń analitycznych, chociaż postępowanie takie może być żmudne, jeśli liczba etapów $k_k - k_o$ jest duża.

Rozpatrzmy dla przykładu jeszcze raz problem, rozwiązany za pomocą zasady maksimum, z równaniami stanu (450) i wskaźnikiem jakości (451). Mamy tu

$$P(x\left[4\right],\ 4) = \frac{1}{2}\, x^2\left[4\right] \tag{476}$$

oraz

$$P(x\left[3\right],\ 3) = \min_{u\left[3\right]} \left\{ \frac{1}{2} \cdot u^2\left[3\right] + \frac{1}{2}\, x^2\left[3\right] + \frac{1}{2}\, (x\left[3\right] + u\left[3\right])^2 \right\} \tag{477}$$

Poszukując minimum, znajdziemy

$$\hat{u}[3] = -\frac{1}{2} x[3]; \qquad P(x[3],\ 3) = \frac{3}{4} x^2\ [3]. \tag{478}$$

Możemy zatem napisać

$$P(x[2],\ 2) = \min_{u[2]} \left\{ \frac{1}{2} u^2[2] + \frac{1}{2} x^2[2] + \frac{3}{4} (x[2] + u[2])^2 \right\} \tag{479}$$

i poszukując minimum, znajdziemy

$$\hat{u}[2] = -\frac{3}{5} x[2]; \qquad P(x[2],\ u[2]) = \frac{4}{5} x^2\ [2]. \tag{480}$$

Postępując tak kolejno, uzyskamy

$$\hat{u}[1] = -\frac{8}{13} x\ [1]; \qquad P(x[1],\ u[1]) = \frac{373}{338} x^2[1], \tag{481}$$

$$\hat{u}[0] = -\frac{21}{34} x[0]; \qquad P(x[0],\ u[0]) = \frac{935}{1156} x^2\ [0]. \tag{482}$$

Obliczone tu sterowanie optymalne zgadza się oczywiście z wynikami, podanymi w tabeli po równaniu (464). Wniosek, że sterowanie optymalne w układzie zamkniętym jest wprost proporcjonalne do stanu procesu, mógł być także uzyskany wcześniej na podstawie zasady maksimum (gdyż sterowanie optymalne jest równe zmiennej sprzężonej (453), równania kanoniczne (456), (457) są liniowe i jednorodne, zaś warunek początkowy $\psi[0]$ proporcjonalny do stanu początkowego $x[0]$ - por. (463)).

9. Metody obliczeniowe optymalizacji dynamicznej

9.1. Uwagi ogólne

Rozwiązania problemów optymalizacji dynamicznej w postaci analitycznej dają się uzyskać tylko dla stosunkowo wąskich klas problemów. Z problemów ogólniejszych należy tu wymienić zagadnienia na minimum czasu sterowania, na minimum wydatku związanego ze sterowaniem czy na minimum wskaźnika jakości o postaci kwadratowej, przy liniowych równaniach procesu sterowanego. Szczegółową dyskusję rozwiązań analitycznych tych problemów można znaleźć w [1].

Dla problemów o ogólniejszej postaci niezbędne jest zastosowanie odpowiednich metod numerycznych. Burzliwy rozwój tych

metod datuje się od roku 1960; w niniejszym skrypcie omówimy podstawowe i wypróbowane metody tego rodzaju. Przed ich omówieniem warto zwrócić uwagę na podstawowe trudności obliczeniowe optymalizacji dynamicznej.

Pierwsza trudność polega na wyznaczaniu chwilowego maksimum hamiltonianu H; jest to trudność optymalizacji statycznej. Nie jest ona jednakże trudnością podstawową; znacznie istotniejsze są trudności dalsze. Dlatego też na ogół zakłada się, że samo wyznaczenie maksimum hamiltonianu jest stosunkowo proste, i że można posługiwać się bądź to postacią analityczną gradientu hamiltonianu $\frac{\partial H}{\partial \underline{u}}$, bądź też nawet postacią analityczną zależności optymalnego rozwiązania $\underline{\hat{u}}$ od zmiennych stanu $\underline{x}$ i sprzężonych $\underline{\psi}$, wynikającą z warunku (375)

$$\underline{\hat{u}} = \varphi(\underline{\psi}, \underline{x}, t). \tag{483}$$

Znacznie istotniejsza - i może największa trudność - związana jest z dwugranicznym charakterem warunków dla stanu początkowego $\underline{x}(t_o)$ oraz końcowych wartości zmiennych sprzężonych $\underline{\psi}(t_k)$, wynikających z warunków transwersalności. Niemożliwe jest bowiem jednoczesne rozwiązanie numeryczne układu równań różniczkowych, z których część ma określone warunki początkowe, a część końcowe. Istotną częścią wszystkich metod numerycznych optymalizacji dynamicznej jest jakiś sposób pokonania czy też obejścia tej trudności. Trudność ta związana jest zresztą z istotą optymalizacji dynamicznej; żeby sterować optymalnie w chwili bieżącej, należy przewidywać, antycypować przyszłe zachowanie się procesu.

Trudność powyższa wzrasta, jeśli warunki końcowe dla procesu mają złożoną postać, i należy korzystać z ogólnej postaci warunków transwersalności (372), (373), nader niedogodnej dla obliczeń numerycznych i możliwej do wykorzystania tylko przy użyciu specyficznych metod wariacyjnych.

Wszelkiego rodzaju ograniczenia nierównościowe, a zwłaszcza ograniczenia stanu, utrudniają rozwiązywanie problemu. Jednakże istnieje cały szereg metod pozwalających uwzględnić te ograniczenia bądź też sformułować równoważne problemy bez ograniczeń. Najbardziej uniwersalna z tych metod polega na wprowadzeniu funkcji kary za przekroczenie ograniczeń [16].

Funkcje kary wprowadza się w sposób następujący: Przypuśćmy, że ograniczone jest sterowanie $\underline{u} \in \Omega$ przy czym obszar określony jest układem nierówności

$$\underline{g}_1(\underline{u}) \leqslant \underline{0}. \tag{484}$$

Funkcją kary $f_{o\varkappa}$ za przekroczenie ograniczeń nazywamy funkcję postaci

$$f_{o\varkappa}(\underline{u}) = \varkappa \underline{g}_1'(\underline{u}) \cdot \underline{1}\{\underline{g}_1(\underline{u})\}\underline{g}_1(\underline{u}) = \begin{cases} 0 & \underline{g}_1(\underline{u}) \leqslant \underline{0} \\ \varkappa \sum_i \underline{g}_{1i}^2(\underline{u}), & g_{1i}(\underline{u}) > 0 \end{cases} \tag{485}$$

przy czym $g_{1i}(\underline{u})$ jest składową funkcji $\underline{g}_1(\underline{u})$, $\underline{1}\{\underline{g}_1(\underline{u})\}$ - macierzą diagonalną o wyrazach na przekątnej równych jedności, gdy $g_{1i}(\underline{u}) > 0$, zaś zeru gdy $g_{1i}(\underline{u}) \leqslant 0$,

$\varkappa$ - współczynnikiem o dostatecznie dużej wartości.

Funkcję kary $f_{o\varkappa}$ dodajemy do funkcji podcałkowej wskaźnika jakości f_o; po rozwiązaniu zadania optymalizacji następuje wtedy zwykle przekraczanie ograniczeń, jednak tylko nieznaczne, gdyż duże przekroczenia są nieopłacalne. Dla uniknięcia przekroczeń można stosować funkcję kary przesuniętą, obliczaną na podstawie przesuniętego ograniczenia o postaci

$$\underline{g}_{1\varepsilon}(\underline{u}) = \underline{g}_1(\underline{u}) + \underline{\varepsilon} \leqslant \underline{0}. \tag{486}$$

Pewne trudności mogą być także związane z opóźnieniami stanu czy sterowania oraz z problemami sformułowanymi w postaci dyskretnej, gdzie konieczne jest rozwiązywanie równań sprzężonych przy odwróconym kierunku biegu czasu, zaczynając od chwili t_k a kończąc w chwili t_o. Jednakże metody obliczeniowe, oparte na podstawowym lemacie rachunku wariacyjnego, z reguły zakładają taki właśnie sposób rozwiązywania równań sprzężonych - por. [17].

Na koniec, metody obliczeniowe optymalizacji dynamicznej są związane zawsze z rozmaitymi metodami numerycznymi rozwiązywania równań różniczkowych, te zaś z kolei - z dyskretyzacją tych równań i rozwiązywaniem równań różniczkowych. Chociaż bardzo istotne, zagadnienia te nie będą omawiane w niniejszym skrypcie, którego celem jest jedynie przedstawienie podstawowych idei metod optymalizacji dynamicznej.

Nakład obliczeń numerycznych związanych z optymalizacją dynamiczną rośnie zazwyczaj z kwadratem lub sześcianem wymiarowości problemu (istotniejsza jest tu zwykle zależność od wymiaru stanu n, niż od wymiaru sterowania m). Szczególnie silna jest zależność nakładu obliczeń od wymiaru stanu w metodzie programowania dynamicznego - jest to zależność wykładnicza, a nie potęgowa; stąd też metoda programowania dynamicznego ma bardzo

ograniczone znaczenie jako algorytm obliczeniowy optymalizacji dynamicznej i nie będzie tu dalej omawiana. Szerszą dyskusję zależności nakładu obliczeń od wymiarowości problemu znaleźć można w [16].

Dla dużych wymiarowości problemu istnieją odpowiednie metody pozwalające na zmniejszenie nakładu obliczeń. Jedną z nich jest jest metoda modelowania łańcuchowego [10], [16], która obliczenia numeryczne zastępuje modelowaniem analogowym. Drugą, bardzo obszerną grupę metod stanowią wielopoziomowe algorytmy obliczeniowe, wyrażające sobą istotę sterowania wielopoziomowego [9], [14], [16].

Na zakończenie tego ogólnego przeglądu podstawowych własności metod obliczeniowych optymalizacji dynamicznej warto jeszcze raz zwrócić uwagę na istotne różnice między optymalizacją dynamiczną a statyczną. Problemy poszukiwania maksimum hamiltonianu muszą być z natury rzeczy znacznie prostsze w porównaniu ze złożonymi problemami optymalizacji statycznej; gdyby wymagały one porównywalnego nakładu obliczeń, to poważniejsze trudności optymalizacji dynamicznej sprawiłyby, że zadanie wyznaczenia optymalnego sterowania byłoby w ogóle nierozwiązalne w rozsądnym czasie. Sam fakt zależności sterowania od czasu sprawia przecież, że wymiarowość problemu dynamicznego jest teoretycznie nieskończenie razy większa od wymiarowości problemu statycznego poszukiwania maksimum hamiltonianu; praktycznie, po dyskretyzacji, jest ona wielokrotnie większa. Gdyby więc podzielić przedział czasu $t_k - t_o$ na N etapów i potraktować problem dynamiczny jako problem statyczny o $N \cdot m$ sterowaniach i $N \cdot n$ więzach równościowych, to nakład obliczeń musiałby wzrastać z kwadratem lub sześcianem liczby N. Tymczasem większość metod optymalizacji dynamicznej wymaga nakładu obliczeń, zależnego proporcjonalnie od liczby etapów dyskretyzacji N, co wynika z wykorzystania jednokierunkowego charakteru związków pomiędzy równaniami procesu na poszczególnych etapach. Dlatego też zazwyczaj bardzo niekorzystne jest zastępowanie metod optymalizacji dynamicznej metodami statycznymi. Wręcz odwrotnie, istnieją przykłady wskazujące, że optymalizacja statyczna pracy szeregu obiektów powiązanych ze sobą jednokierunkowo w przestrzeni może być dokonana bardzo sprawnie za pomocą metod optymalizacji dynamicznej.

Charakter zależności nakładu obliczeń przy optymalizacji dynamicznej od wymiarowości problemu ilustruje następująca tabelka.

	Zastępcze ujęcie statyczne	Programowanie dynamiczne	Inne metody dynamiczne
Charakter zmienności nakładu obliczeń	$[N(n+m)]^{\alpha}$	$N\, m^{\alpha} L^{n}$	$N\,(n+m)^{\alpha}$

gdzie: α - wykładnik zawarty w granicach 2...4,
n - wymiar stanu,
m - wymiar sterowania,
N - liczba etapów dyskretyzacji czasu,
L - liczba kwantów dyskretyzacji stanu.

Dokładniejszą dyskusję powyższych zależności znaleźć można w [16].

9.2. Podstawowe metody obliczeniowe

9.2.1. Podział metod obliczeniowych optymalizacji dynamicznej

Wśród metod obliczeniowych optymalizacji dynamicznej wyróżnić można metody podstawowe, metody specjalne, metody wielopoziomowe i metody łańcuchowe.

Metody podstawowe dzielą się z kolei na dwie grupy: metody bezpośrednie, oparte głównie na podstawowym lemacie rachunku wariacyjnego oraz metody pośrednie, polegające na wyznaczaniu rozwiązania optymalnego na podstawie warunków koniecznych optymalności (np. na podstawie zasady maksimum).

Do grupy metod bezpośrednich należą: metoda gradientu w przestrzeni funkcyjnej [6] metoda gradientu sprzężonego w przestrzeni funkcyjnej [11] oraz metoda drugiej wariacji [7]. Zaliczyć do nich można także metodę programowania dynamicznego [2], która nie będzie tu omawiana.

Do grupy metod pośrednich należą: metoda Newtona w przestrzeni funkcyjnej (metoda Newtona-Raphsona) [7] oraz metody przeszukiwania ekstremal, metoda gradientu oraz metoda Newtona [16]. Metody pośrednie będą omówione tu w skrócie.

Podstawowym celem niniejszej części skryptu jest wyjaśnienie idei i procedur obliczeniowych bezpośrednich metod podstawowych.

Metody specjalne są formułowane dla poszczególnych typów zadań optymalizacji dynamicznej, w których wiadomo jest, że sterowanie jest bądź to przedziałami stałe, bądź przedziałami leżące na ograniczeniach - na przykład dla problemów czaso-optymalnych, - bądź też posiada inne szczególne własności - por. np. [2]. Metody te nie będą tu omawiane.

Metody wielopoziomowe są bardzo obszerne i bogate - por. metody Kulikowskiego [9], Mesarovicia [14] i inne [16]. Dlatego też w niniejszej części skryptu omówiony będzie tylko jeden przypadek metody wielopoziomowej dla nieaddytywnych wskaźników jakości, wprowadzonej przez autora w pracy [16].

Metody modelowania łańcuchowego zapoczątkowane były przez Kurmana [10]; są to metody bardzo specyficzne i nie będą tu omawiane; krótki ich przegląd zawarty jest w [16].

9.2.2. Metoda gradientu w przestrzeni funkcyjnej sterowań

Metoda gradientu formułowana jest zazwyczaj dla zadania optymalizacji, w którym czas końcowy t_k jest dany, zaś współrzędne stanu końcowego - swobodne; możliwe wprawdzie uwzględnienie innej postaci warunków końcowych, jest jednak ono skomplikowane i nie będzie tu dokładniej omawiane. Przypominamy, że wariacja wskaźnika jakości przybiera w tym przypadku postać:

$$\delta Q = \left[\frac{\partial f_k\left(\underline{x}(t_k)\right)}{\partial \underline{x}} + \underline{\psi}'(t_k)\right]\delta \underline{x}(t_k) - \int_{t_0}^{t_k}\left[\underline{\dot{\psi}}' + \frac{\partial H(\underline{\psi}, \underline{x}, \underline{u}, t)}{\partial \underline{x}}\right]\delta \underline{x}\, dt - \int_{t_0}^{t_k}\frac{\partial H(\underline{\psi}, \underline{x}, \underline{u}, t)}{\partial \underline{u}}\,\delta \underline{u}\, dt. \quad (487)$$

Metoda gradientu stanowi bezpośrednie wykorzystanie podstawowego lematu rachunku wariacyjnego i polega na iteracyjnym konstruowaniu coraz lepszych trajektorii, poczynając od pewnego arbitralnie założonego początkowego przebiegu sterowania $\underline{u}^{(1)}$. Jej procedura w i-tej iteracji jest następująca:

a. Znając funkcję czasu $\underline{u}^{(i)}$, przy danych warunkach początkowych $\underline{x}_o$ scałkować równanie stanu procesu

$$\dot{\underline{x}}^{(i)} = \frac{\partial H\left(\underline{\psi}^{(i)}, \underline{x}^{(i)}, \underline{u}^{(i)}, t\right)}{\partial \underline{\psi}} = \underline{f}\left(\underline{x}^{(i)}, \underline{u}^{(i)}, t\right); \quad \underline{x}^{(i)}(t_o) = \underline{x}_o. \quad (488)$$

b. Znając wartość końcową stanu $\underline{x}^{(i)}(t_k)$, obliczyć wartość końcową zmiennych sprzężonych $\underline{\psi}^{(i)}(t_k)$ według wzoru

$$\underline{\psi}^{(i)}(t_k) = -\frac{\partial f_k\left(\underline{x}^{(i)}(t_k)\right)}{\partial \underline{x}'}, \quad (489)$$

Dzięki temu pierwsza składowa wariacji wskaźnika jakości (487) staje się równa zeru.

c. Znając funkcje czasu $\underline{x}^{(i)}$, $\underline{u}^{(i)}$ oraz wartość końcową $\underline{\psi}^{(i)}(t_k)$, scałkować równania sprzężone

$$\dot{\underline{\psi}}^{(i)} = -\frac{\partial H\left(\underline{\psi}^{(i)}, \underline{x}^{(i)}, \underline{u}^{(i)}, t\right)}{\partial \underline{x}'} \quad (490)$$

w kierunku od chwili t_k do chwili t_o (w "odwróconym" kierunku biegu czasu). W ten sposób druga składowa wariacji wskaźnika

jakości (487) staje się równa zeru. Jednocześnie z całkowaniem równań sprzężonych obliczyć gradient

$$\underline{\gamma}^{(i)} = - \frac{\partial H\left(\underline{\psi}^{(i)}, \underline{x}^{(i)}, \underline{u}^{(i)}, t\right)}{\partial \underline{u}'} \tag{491}$$

jako funkcję czasu, stanowiącą gradient funkcjonału jakości. Jeśli funkcja $\underline{\gamma}^{(i)}$ nie jest równa zeru, to wskazuje ona kierunek zmian sterowania $\underline{u}^{(i)}$, które zapewnią zmniejszanie się wskaźnika jakości.

Dla oceny odległości założonego sterowania $\underline{u}^{(i)}$ od sterowania optymalnego $\hat{\underline{u}}$ korzystne jest obliczanie kwadratu normy gradientu $\underline{\gamma}^{(i)}$

$$\left\| \underline{\gamma}^{(i)} \right\|^2 = \int_{t_0}^{t_k} \left(\underline{\gamma}^{(i)} \right)' \underline{\gamma}^{(i)} \, dt \tag{492}$$

d. Zakładając pewną wartość współczynnika długości kroku ρ, zastosować sterowanie

$$\underline{u}_\rho^{(i)} = \underline{u}^{(i)} - \rho \underline{\gamma}^{(i)} \tag{493}$$

i scałkować równania stanu (488) po podstawieniu $\underline{u}_\rho^{(i)}$, wyznaczając odpowiednio trajektorię $\underline{x}_\rho^{(i)}$ oraz obliczyć wskaźnik jakości

$$Q_\rho^{(i)} = f_k\left(\underline{x}_\rho^{(i)}(t_k) \right) + \int_{t_0}^{t_k} f_o\left(\underline{x}_\rho^{(i)}, \underline{k}_\rho^{(i)}, t \right) dt. \tag{494}$$

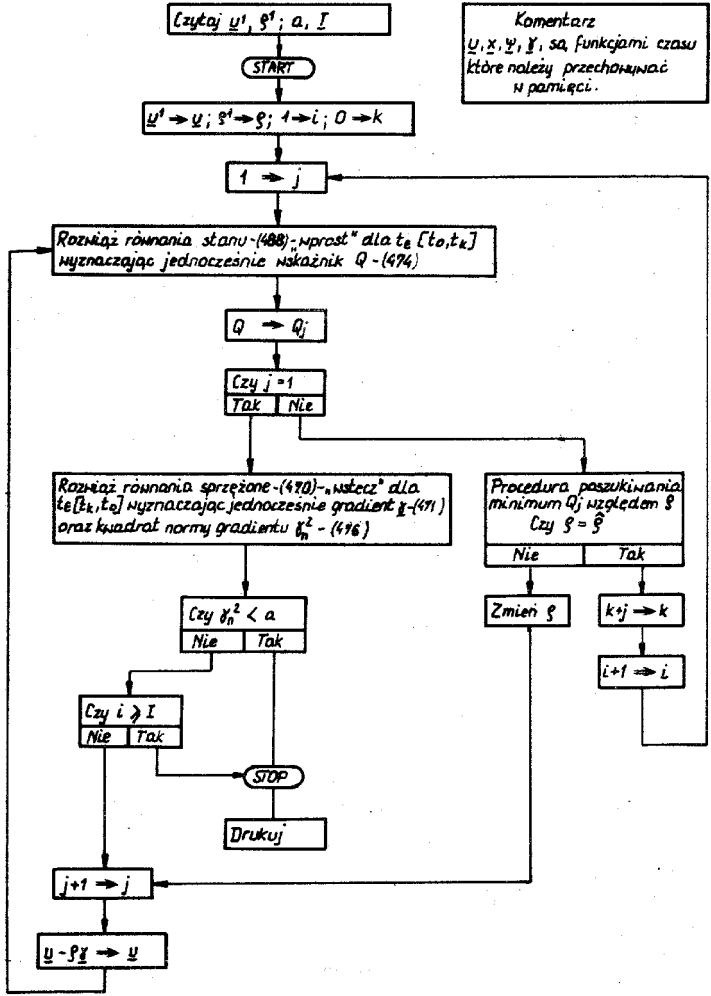
Obliczenie przeprowadzić kilkakrotnie dla różnych wartości ρ, przeprowadzając minimalizację $Q_\rho^{(i)}$ względem ρ i obliczając najkorzystniejsze $\hat{\rho}^{(i)}$.

e. Sterowanie początkowe dla następnej iteracji wynika z równania

$$\underline{u}^{(i+1)} = \underline{u}^{(i)} - \hat{\rho}^{(i)} \underline{\gamma}^{(i)}. \tag{495}$$

Iteryjce powtarza się, dopóki przyrosty $\left| Q_{\rho=0}^{(i-1)} - Q_{\rho=0}^{(i)} \right|$ nie staną się dostatecznie małe lub dopóki wybrana norma gradientu $\underline{\gamma}^{(i)}$ nie stanie się dostatecznie mała, na przykład dopóki nie będzie spełniony warunek

$$\left\| \underline{\gamma}^{(i)} \right\|^2 = \int_{t_0}^{t_k} \left(\underline{\gamma}^{(i)} \right)' \underline{\gamma}^{(i)} \, dt \leqslant a. \tag{496}$$


Czytaj u', ξ', a, I
START
u' → u; ξ' → ξ; 1 → i; 0 → k
1 → j
Rozwiąż równania stanu-(488)-„wprost" dla t_e [t_o, t_k] wyznaczając jednocześnie wskaźnik Q-(474)
Q → Q_j
Czy j = 1
Tak
Nie
Rozwiąż równania sprzężone-(470)-„wstecz" dla t_e [t_k, t_e] wyznaczając jednocześnie gradient γ-(471) oraz kwadrat normy gradientu γ_n^2 - (476)
Procedura poszukiwania minimum Q_j względem ξ
Czy ξ = ξ̂
Nie
Tak
Zmień ξ
k+j → k
i+1 → i
Czy γ_n^2 < a
Nie
Tak
Czy i ≥ I
Nie
Tak
STOP
Drukuj
j+1 → j
u - ργ → u
Komentarz
u, x, y, γ, są funkcjami czasu które należy przechowywać w pamięci.

Rys. 73

Uproszczony schemat działania dla metody gradientu przedstawia rys. 73. W schemacie tym obliczenia przerywa się, gdy spełniony jest warunek (496) lub przekroczona jest dana z góry liczba iteracji I, czyli gdy $i \geqslant I$. Indeks j oznacza liczbę powtórzeń obliczeń równań stanu i wskaźnika w jednej iteracji, indeks k - liczbę powtórzeń obliczeń równań stanu i wskaźnika we wszystkich iteracjach.

Jeśli rozwiązanie zagadnienia optymalizacji istnieje i jest jedyne, to metoda gradientu zapewnia zbieżność do tego rozwiązania. Ogólniej, metoda gradientu zapewnia zbieżność do najbliższej ekstremali transwersalnej zadania (to jest do najbliższej trajektorii, spełniającej warunki konieczne optymalności).

Przedstawiona wyżej procedura gradientowa nie uwzględniała ograniczeń sterowania i stanu; można je uwzględnić wprowadzając do wskaźnika jakości funkcje kary, lub za pomocą innych odpowiednich metod.

W literaturze istnieje cały szereg przykładów konkretnych obliczeń numerycznych metodą gradientu. Przed ich omówieniem podamy jednak przykłady analitycznego zastosowania tej metody dla pewnych prostych zagadnień, o znanych skądinąd rozwiązaniach optymalnych, mające na celu bliższe wyjaśnienie istoty metody.

Przykład 1

Dany jest proces

$$\dot{x} = u; \qquad x(0) = 0; \qquad x(T) \text{ - swobodne}; \qquad T \text{ - dane}, \tag{1.1}$$

o wskaźniku jakości

$$Q = \frac{1}{2}\left\{ \left[X - x(T)\right]^2 + \int_0^T (x^2 + u^2)dt \right\}. \tag{1.2}$$

Hamiltonian problemu ma postać

$$H = -\frac{u^2 + x^2}{2} + \psi u, \tag{1.3}$$

zaś równanie sprzężone

$$\dot{\psi} = x \tag{1.4}$$

oraz warunek transwersalności

$$\psi(T) = X - x(T), \tag{1.5}$$

pozwalają w oparciu o zasadę maksimum wyznaczyć sterowanie optymalne

$$\hat{u} = Xe^{-T}\operatorname{cht} = xe^{-T}\sum_{i=0}^{\infty}\frac{t^{2i}}{(2i)!}, \tag{1.6}$$

trajektorię optymalną oraz minimum wskaźnika jakości

$$x = Xe^{-T}\,\mathrm{sht}; \qquad Q = \frac{X^2}{2}\,(e^{-2T} + 1). \tag{1.7}$$

Dla $X = 2$, $T = 10$ mamy $Q = 1,00$ z dokładnością do 10^{-8}. Zastosujmy teraz dla rozwiązania tego zadania metodę gradientu. Całkując równania stanu (1.1) dla danego $u^{(i)}$ uzyskujemy

$$x^{(i)} = \int_0^t u^{(i)}dt. \tag{1.8}$$

Znając $x^{(i)}(T)$, wyznaczamy $\psi^{(i)}(T)$, i całkując równania sprzężone (1.4) wstecz uzyskujemy

$$\psi^{(i)} = X - \int_0^T u^{(i)}dt - \int_t^T x^{(i)}dt. \tag{1.9}$$

Gradient funkcjonału $\gamma^{(i)}$ ma postać

$$\gamma^{(i)} = \psi^{(i)} - u^{(i)}, \tag{1.10}$$

zaś poprawione sterowanie $u_\rho^{(i)}$ i trajektoria stanu $x_\rho^{(i)}$ wyrażają się wzorami

$$u_\rho^{(i)} = u^{(i)} - \rho\gamma^{(i)}; \qquad x_\rho^{(i)} = x^{(i)} - \rho\int_0^t \gamma^{(i)}dt. \tag{1.11}$$

Podstawiając te wyrażenia do wskaźnika jakości (1.2), uzyskujemy

$$Q_\rho^{(i)} = Q^{(i)} - \Delta Q_\rho^{(i)}, \tag{1.12}$$

gdzie $Q^{(i)}$ jest wartością wskaźnika wynikającą z trajektorii $x^{(i)}$, $u^{(i)}$,

$$Q^{(i)} = \frac{1}{2}\left\{\left[X - x^{(i)}(T)\right]^2 + \int_0^T\left[\left(x^{(i)}\right)^2 + \left(u^{(i)}\right)^2\right]dt, \right. \tag{1.13}$$

zaś poprawa wskaźnika $\Delta Q^{(i)}$ zależy w sposób kwadratowy od ρ

$$-\Delta Q^{(i)} = \frac{1}{2}\left\{-2A^{(i)}\rho + B^{(i)}\rho^2\right\} \tag{1.14}$$

przy czym

$$A^{(i)} = \int_0^T \left(\gamma^{(i)}\right)^2 dt; \quad B^{(i)} = \left(\int_0^T \gamma^{(i)} dt\right)^2 + \int_0^T \left[\left(\gamma^{(i)}\right)^2 + \left(\int_0^t \gamma^{(i)} dt\right)^2\right] dt. \quad (1.15)$$

Jak widać $A^{(i)}$ jest kwadratem normy gradientu $\gamma^{(i)}$. Najkorzystniejszy współczynnik kroku $\hat{\varrho}^{(i)}$ oraz najlepsza poprawka wskaźnika jakości wyrażają się wzorami

$$\hat{\varrho}^{(i)} = \frac{A^{(i)}}{B^{(i)}}; \qquad -\Delta\hat{Q}^{(i)} = -\frac{1}{2} A^{(i)} \cdot \hat{\varrho}^{(i)} \quad (1.16)$$

i wyznaczają dane początkowe dla następującej iteracji

$$Q^{(i+1)} = Q^{(i)} - \Delta\hat{Q}^{(i)}; \quad u^{(i+1)} = u^{(i)} - \hat{\varrho}^{(i)} \gamma^{(i)}. \quad (1.17)$$

Należy tu podkreślić, że w rozważanym przykładzie przeprowadziliśmy minimalizację wskaźnika $Q_{\varrho}^{(i)}$ względem ϱ na drodze analitycznej; w obliczeniach numerycznych stosuje się w tym punkcie oczywiście numeryczne metody poszukiwania minimum funkcji jednej zmiennej.

Załóżmy w pierwszej iteracji $u^{(1)} = 0$. Uzyskamy wtedy dla $X = 2$, $T = 10$:

$$x^{(1)} = 0; \quad Q^{(1)} = 2; \quad \psi^{(1)} = 2; \quad -\gamma^{(1)} = 2 \quad (1.18)$$

oraz

$$A^{(1)} = 40; \quad B^{(1)} = 1770; \quad \hat{\varrho}^{(1)} = 2{,}26 \cdot 10^{-2}, \quad (1.19)$$

$$\Delta\hat{Q}^{(1)} = 0{,}452$$

i na koniec

$$u^{(2)} = 4{,}52 \cdot 10^{-2}; \quad Q^{(2)} = 1{,}548. \quad (1.20)$$

Zauważmy, że w wyniku pierwszej iteracji uzyskaliśmy najlepsze możliwe przybliżenie sterowania optymalnego za pomocą sterowania stałego w czasie. W drugiej iteracji gradient funkcjonału staje się funkcją kwadratową czasu

$$x^{(2)} = 4{,}52 \cdot 10^{-2} \cdot t; \qquad \gamma^{(2)} = -0{,}71 + 2{,}26 \cdot 10^{-2} t^2; \quad (1.21)$$

$$-\gamma^{(2)} = -0,755 + 2,26\cdot 10^{-2}\cdot t^2 \qquad (1.21)$$

i uzyskujemy

$$A^{(2)} = 4,51; \quad B^{(2)} = 47,5; \quad \gamma^{(2)} = 9,5\cdot 10^{-2}; \qquad Q^{(2)} = 0,216, \qquad (1.22)$$

a na koniec

$$u^{(3)} = -2,64\cdot 10^{-2} + 2,15\cdot 10^{-3}\cdot t^2; \qquad Q^{(3)} = 1,332. \qquad (1.23)$$

Zauważmy, że w wyniku drugiej iteracji uzyskaliśmy przybliżenie sterowania optymalnego za pomocą stałej i kwadratowej funkcji czasu - a więc za pomocą funkcji, wchodzących w skład dwóch pierwszych wyrazów szeregu (1.6). Nie jest to jednak przybliżenie najlepsze w tej klasie funkcji; poszukując niezależnie współczynników a_o i a_2 w funkcji

$$u^{(3)} = a_o + a_2 t^2, \qquad (1.24)$$

uzyskalibyśmy przybliżenie lepsze; w metodzie gradientu jednak nie dobieramy tych współczynników niezależnie, a jedynie w kierunku wyznaczonego gradientu.

Podobna sytuacja powtarza się w trzeciej iteracji, gdy uzyskujemy

$$-\gamma^{(3)} = 1,11 - 1,54\cdot 10^{-2}t^2 + 1,79\cdot 10^{-4}t^4;$$

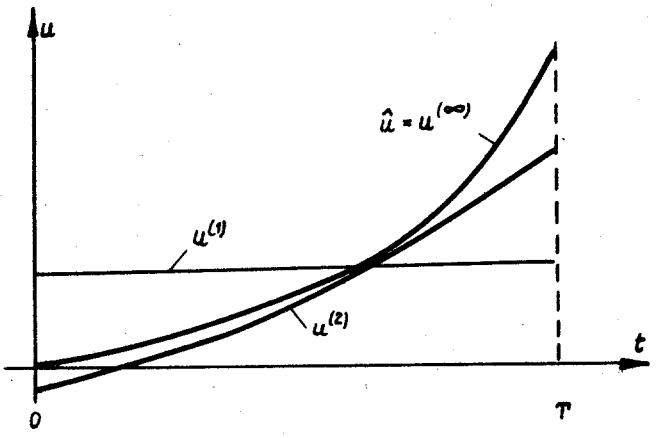
$$u^{(4)} = 7,1\cdot 10^{-3} + 1,7\cdot 10^{-3}t^2 + 5,4\cdot 10^{-6}t^4; \qquad (1.25)$$

$$Q^{(4)} = 1,191$$

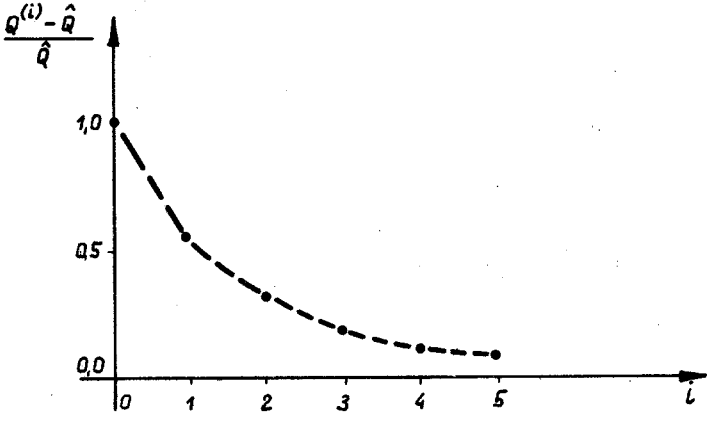
Widoczne jest, że w kolejnych iteracjach sterowanie optymalne jest przybliżone za pomocą kolejnych funkcji, występujących w szeregu (1.6); nie są to jednak najkorzystniejsze z możliwych przybliżeń sterowania optymalnego za pomocą tych funkcji.

Skutkiem tych własności metody gradientu, jej zbieżność jest nader powolna.

Wyniki kolejnych iteracji ilustrują rysunki 74, 75 (koniec przykładu 1).



Rys. 74



Rys. 75

Powolna zbieżność, zwłaszcza w bliskim otoczeniu, sterowania optymalnego jest podstawową wadą metody gradientu. Poza tą wadą ma ona liczne zalety, wynikające w większości z rozwiązywania równań sprzężonych w odwróconym kierunku biegu czasu. Równania te są bowiem niestabilne, a przeciwny kierunek ich rozwiązywania czyni je oczywiście stabilnymi, co zmniejsza trudności oszacowania niezbędnej dokładności obliczeń numerycznych. Dzięki przeciwnemu kierunkowi rozwiązywania równań sprzężonych metodę gradientu można bez trudu zastosować do rozwiązywania problemów z opóźnieniami stanu lub sterowania, w których to problemach w równaniach sprzężonych występują argumenty wyprzedzone lub do problemów sformułowanych w postaci dyskretnej; dla wszystkich tych problemów przeciwny kierunek rozwiązywania równań sprzężonych jest niejako kierunkiem naturalnym, wynikającym z istoty tych problemów - por. [16], [17].

Dobrym przykładem zastosowania metody gradientu, dostępnym w literaturze, jest problem wyznaczania minimalno-czasowej trajektorii statku kosmicznego, sterowanego z Ziemi na Marsa [7]. Proces ruchu statku jest opisywany trzema równaniami różniczkowymi nieliniowymi, z jednym sterowaniem

$$\dot{x}_1 = x_2,$$

$$\dot{x}_2 = \frac{x_3^2}{x_1} - \frac{x_2}{x_1^2} + b \sin u, \tag{497}$$

$$\dot{x}_3 = \frac{-x_2 x_3}{x_1} + b \cos u,$$

$$x_1(t_o),\ x_2(t_o),\ x_3(t_o) \text{ - dane},$$

$$x_1(t_k),\ x_2(t_k),\ x_3(t_k) \text{ - dane},$$

$$Q = \int_{t_o}^{t_k} 1 \, dt = (t_k - t_o) \text{ - minimalne}.$$

Dane warunki końcowe zastąpiono przy rozwiązywaniu swobodnymi, wprowadzając do wskaźnika jakości funkcję kary za odchylenie od danych wartości końcowych. Zakładano pewne wartości t_k, które w trakcie kolejnych iteracji były zmienione zależnie od wartości końcowej hamiltonianu H. Jednokrotna iteracja metody gradientu wymagała około 2 sekund obliczeń na maszynie IBM 7094. Wyznaczenie trajektorii optymalnej z dokładnością do 0,1% wartości minimalnego wskaźnika jakości wymagało ok. 120 iteracji, czyli około 240 sekund.

Innym przykładem zastosowania metody gradientu jest wyznaczanie maksymalnego zasięgu szybowców [6]. W przykładzie tym niestabilność równań sprzężonych jest tak istotna, że zmiany warunków początkowych tych równań przy rozwiązywaniu w zwykłym kierunku biegu czasu powodują setki milionów razy większe zmiany warunków końcowych; okoliczność ta utrudniała ocenę właściwej dokładności obliczeń i wywoływała nadmiary w obliczeniach maszynowych przy zastosowaniu innych metod, niż metoda gradientu. Natomiast metoda gradientu umożliwiała stosunkowo łatwe wyznaczanie rozwiązania optymalnego.

9.2.3. Metoda gradientu sprzężonego w przestrzeni funkcyjnej sterowań

Metoda gradientu sprzężonego stanowi bezpośrednie ulepszenie metody gradientu zwykłego, oparta na pojęciu kierunków sprzężonych przy poszukiwaniu minimum funkcji kwadratowej wielu zmiennych. Wiadomo bowiem, że już w przypadku funkcji kwadratowej wielu zmiennych kierunek gradientu nie jest najkorzystniejszym kierunkiem poszukiwania minimum; lepsze od tego kierunku są oczywiście kierunki osi głównych tworów geometrycznych, odpowiadających stałym wartościom tej funkcji. Kierunki te zwane są właśnie kierunkami sprzężonymi.

Zastosowanie metody gradientu sprzężonego do optymalizacji dynamicznej jest przedstawione szczegółowo w [15]. W pierwszej iteracji stosuje się przy tym metodę gradientu zwykłego. W dalszych iteracjach procedura obliczeniowa jest także bardzo podobna do procedury gradientu zwykłego; różni się od niej tylko dodatkowym wyznaczaniem kierunku sprzężonego oraz poszukiwaniem minimum w tym właśnie kierunku. W związku z powyższym ulegają zmianie punkty c i d procedury.

c. Oblicza się dodatkowo współczynnik:

$$\beta^{(i)} = \frac{\left\| \underline{\gamma}^{(i)} \right\|^2}{\left\| \underline{\gamma}^{(i-1)} \right\|^2} \tag{498}$$

oraz wyznacza się kierunek sprzężony dla i-tej procedury

$$\underline{s}^{(i)} = -\underline{\gamma}^{(i)} + \beta^{(i)} \underline{s}^{(i-1)} \tag{499}$$

W pierwszej iteracji przyjmuje się $\underline{s}^{(1)} = -\underline{\gamma}^{(1)}$.

d. Zakładając pewną wartość współczynnika długości kroku ϱ, wyznaczyć sterowanie

$$\underline{u}_{\varrho}^{(i)} = \underline{u}^{(i)} + \varrho \underline{s}^{(i)} \tag{500}$$

i scałkować równania stanu po podstawieniu $\underline{u}_{\varrho}^{(i)}$, wyznaczając odpowiednie $\underline{x}_{\varrho}^{(i)}$ oraz obliczyć wskaźnik jakości $Q_{\varrho}^{(i)}$ - patrz wzór (494); obliczenia powtórzyć kilakrotnie dla różnych wartości ϱ, przeprowadzając minimalizację $Q_{\varrho}^{(i)}$ względem ϱ i wyznaczając najkorzystniejsze $\hat{\varrho}^{(i)}$.

Punkt e. procedury gradientu pozostaje oczywiście bez zmiany. Nie podajemy tu schematu działania metody, gdyż odpowiednie zmiany w stosunku do rys. 73 są nader oczywiste.

Metoda gradientu sprzężonego zachowuje wszystkie zalety metody gradientu zwykłego, związane z przeciwnym kierunkiem rozwiązywania równań sprzężonych - a więc może być stosowana dla problemów dyskretnych, z opóźnieniami, z silnie niestabilnym charakterem równań sprzężonych itp. Ponadto można udowodnić [11] następujące własności metody gradientu sprzężonego:

A. Mimo, że poszukiwanie minimum nie odbywa się w kierunku gradientu, to istnieje zawsze takie $\varrho > 0$, że $Q_{\varrho}^{(i)} < Q^{(i)}$, jeśli tylko $\| \underline{\gamma}^{(i)} \| \neq 0$; innymi słowy kierunek sprzężony wyznacza zawsze kierunek spadku funkcjonału.

B. Jeśli problem optymalizacji jest liniowo-kwadratowy (to znaczy, jeśli równania stanu oraz ewentualne warunki końcowe są liniowe, zaś wskaźnik jakości, czyli - co równoważne - funkcje f_o, f_k mają postać kwadratową) to począwszy od drugiej iteracji wartości wskaźnika jakości $Q^{(i)}$ uzyskiwane za pomocą metody gradientu sprzężonego są mniejsze, niż uzyskiwane za pomocą metody gradientu zwykłego przy tym samym początkowym $\underline{u}^{(1)}$.

C. Jeśli problem optymalizacji jest liniowo-kwadratowy, to dla każdej i-tej iteracji uzyskuje się minimum wskaźnika jakości względem dowolnej kombinacji liniowej dotychczasowych kierunków sprzężonych - a więc względem dowolnej kombinacji liniowej funkcji czasu $\underline{s}^{(1)}, \ldots \underline{s}^{(i)}$.

Własności B. i C. zapewniają - w przeciwieństwie do zwykłej metody gradientu - szybką zbieżność metody gradientu sprzężonego w otoczeniu rozwiązania optymalnego dla wszystkich problemów dostatecznie regularnych; bardzo szeroka klasa problemów optymalizacji daje się w otoczeniu rozwiązania optymalnego przybliżyć za pomocą problemów liniowo-kwadratowych.

Dla szczególnej klasy problemów kwadratowo-liniowych o dodatnio-półokreślonym wskaźniku jakości, gdy rozwiązania optymalne jako funkcje czasu mają postać wielomianów, metoda gradientu sprzężonego w przestrzeni funkcyjnej sterowań zapewnia osiągnięcie rozwiązania w skończonej liczbie iteracji (własność tę, zwaną zbieżnością drugiego rzędu, ma znacznie obszerniejsza klasa problemów kwadratowych w przypadku optymalizacji statycznej przy zastosowaniu metody gradientu sprzężonego, a także szeregu innych metod; dla problemów optymalizacji dynamicznej rozwiązanych me-

todami gradientowymi zbieżność drugiego rzędu zachodzi dla znacznie węższej klasy problemów).

Podstawowe idee oraz własności metody gradientu sprzężonego wyjaśnia następujący prosty przykład bezpośrednio nawiązujący do przykładu 1.

Przykład 2

Dany jest problem jak w przykładzie 1, którego rozwiązanie analityczne określają wzory (1.6), (1.7). W procedurze iteracyjnej gradientu sprzężonego rozwiązania równań stanu, równań sprzężonych oraz sam gradient wyrażają się wzorami (1.8), (1.9), (1.10). Współczynnik poprawy kierunku $\beta^{(i)}$ wyraża się wzorem

$$\beta^{(i)} = \frac{A^{(i)}}{A^{(i-1)}}; \qquad A^{(i)} = \int_0^T \left(X^{(i)}\right)^2 dt, \tag{2.1}$$

zaś kierunek sprzężony - wzorem

$$s^{(i)} = -\gamma^{(i)} + \beta^{(i)} s^{(i-1)}. \tag{2.2}$$

Sterowanie $u_\varrho^{(i)}$ oraz trajektoria $x_\varrho^{(i)}$ mają postać

$$u_\varrho^{(i)} = u^{(i)} + \varrho s^{(i)}; \qquad x_\varrho^{(i)} = x^{(i)} + \varrho \int_0^t s^{(i)} dt, \tag{2.3}$$

zaś wskaźnik jakości $Q_\varrho^{(i)}$ - postać (1.12). Składowa $Q^{(i)}$ tego wskaźnika wyraża się wzorem (1.13), zaś składowa - $\Delta Q_\varrho^{(i)}$, wzorem

$$-\Delta Q_\varrho^{(i)} = \frac{1}{2}\left\{-2A^{(i)}\varrho + B_s^{(i)} \varrho^2\right\} \tag{2.4}$$

przy czym

$$B_s^{(i)} = \left[\int_0^T s^{(i)} dt\right]^2 + \int_0^T \left\{\left[s^{(i)}\right]^2 + \left[\int_0^t s^{(i)} dt\right]^2\right\} dt. \tag{2.5}$$

Najkorzystniejszy współczynnik kroku $\hat{\varrho}^{(i)}$ oraz najlepsza poprawka wskaźnika $\Delta\hat{Q}^{(i)}$ wyrażają się wzorami

$$\hat{\varrho}^{(i)} = \frac{A^{(i)}}{B_s^{(i)}}; \qquad -\Delta\hat{Q}^{(i)} = \frac{1}{2} A^{(i)} \hat{\varrho}^{(i)} \tag{2.6}$$

i wyznaczają dane początkowe do następnej iteracji według wzoru (1.17).

W pierwszej iteracji, przy założonym $u^{(1)} = 0$ uzyskuje się $x^{(1)} = 0$, $Q^{(1)} = 2$, $-\gamma^{(1)} = s^{(1)} = 2$ - por. wzór (1.18), $A^{(1)} = 40$, $B^{(1)} = 1770$, $\hat{\rho}^{(1)} = 2,26 \cdot 10^{-2}$, $\Delta Q^{(1)} = 0,452$ - wzór (1.19), i na koniec $u^{(2)} = 4,52 \cdot 10^{-2}$, $Q^{(2)} = 1,548$ - wzór (1.20). W drugiej iteracji - według wzoru (1.21) - mamy

$$x^{(2)} = 4,52 \cdot 10^{-2} t; \qquad \psi^{(2)} = -0,71 + 2,26 \cdot 10^{-2} t^2;$$

$$-\gamma^{(2)} = -0,755 + 2,26 \cdot 10^{-2} t^2, \tag{2.7}$$

a następnie

$$A^{(2)} = 4,51; \quad \beta^{(2)} = 0,112; \quad s^{(2)} = -0,52 + 2,26 \cdot 10^{-2} t^2 \tag{2.8}$$

oraz

$$B_s^{(2)} = 24,8; \quad \hat{\rho}^{(2)} = 0,182; \quad \Delta \hat{Q}^{(2)} = 0,411 \tag{2.9}$$

i na koniec

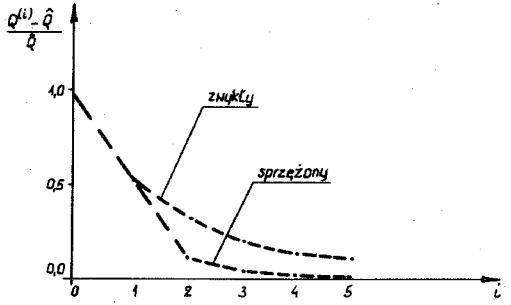
$$u^{(3)} = -5,0 \cdot 10^{-2} + 4,1 \cdot 10^{-3} t^2; \quad Q^{(3)} = 1,127. \tag{2.10}$$

Funkcja $u^{(3)}$ stanowi najlepsze przybliżenie sterowania optymalnego za pomocą funkcji stałej i kwadratowej funkcji czasu, zgodnie z własnością C. metody gradientu sprzężonego. Porównując wyniki (2.10) oraz (1.23) stwierdzimy, że za pomocą gradientu sprzężonego uzyskaliśmy znacznie większą poprawę wartości funkcjonału jakości w drugiej iteracji.

Dalsze rezultaty ilustruje tabelka

i	1	2	3	4	5	6
$Q^{(i)}$	2,000	1,548	1,137	1,053	1,018	1,005

zaś różnice charakteru zbieżności metody gradientu sprzężonego i gradientu zwykłego dla tego przykładu - rys. 76, na któ-

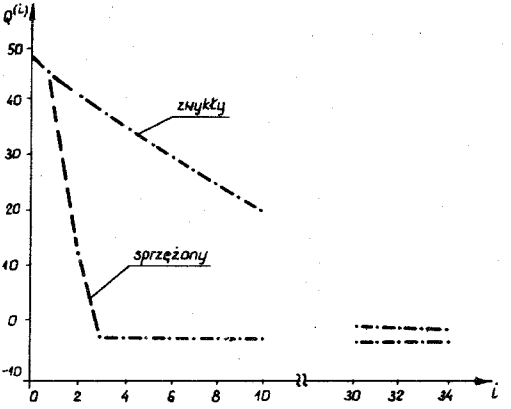


Rys. 76

rym widoczna jest znaczna przewaga metody gradientu sprzężonego.

(koniec przykładu 2).

W artykule [12] podane są między innymi wyniki zastosowania metody gradientu sprzężonego i gradientu zwykłego dla problemu



Rys. 77

$$\dot{x}_1 = x_2 \quad ; \quad x_1(0) = 0,$$

$$\dot{x}_2 = 64 \sin u - 32 \; ; \quad x_2(0) = 0,$$

$$Q = 0{,}002\left[x_1(100) - 10^5\right]^2 + 0{,}05\left[x_2(100)\right]^2 - \int_0^{100} 64 \cos u \, dt \quad (501)$$

Wartości wskaźnika jakości w kolejnych iteracjach dla obu metod przedstawione są na rys. 77.

Tu także widoczna jest znaczna przewaga metody gradientu sprzężonego.

9.2.4. Metoda drugiej wariacji

Metoda drugiej wariacji jest dalszym udoskonaleniem metody gradientu; dlatego też, podobnie jak metodę gradientu, rozpatrzymy ją dla problemu o danym czasie końcowym t_k i swobodnym stanie końcowym $x(t_k)$, bez ograniczeń. I dla tej metody możliwe są odpowiednie modyfikacje, pozwalające na jej zastosowanie dla problemów innych typów; nie będą one jednak tu omawiane.

Przypominamy, że przy odpowiednich założeniach o różniczkowalności funkcji $\underline{f}$, f_0, f_k, przyrost wartości funkcjonału jakości pomiędzy dwoma ostatecznie bliskimi trajektoriami $(\underline{x}, \underline{u})$ i $(\underline{x} + \varepsilon\delta\underline{x}, \underline{u} + \varepsilon\delta\underline{u})$ może być przedstawiony w postaci

$$\Delta Q = Q\left\{\underline{x} + \varepsilon\delta\underline{x}, \; \underline{u} + \varepsilon\delta\underline{u}\right\} - Q\left\{\underline{x}, \; \underline{u}\right\} =$$

$$= \varepsilon\delta Q\left\{\underline{x}, \; \underline{u}, \; \delta\underline{x}, \; \delta\underline{u}\right\} + \frac{1}{2}\varepsilon^2\delta^2 Q\left\{\underline{x}, \; \underline{u}, \; \delta\underline{x}, \; \delta\underline{u}\right\} + \ldots \quad (502)$$

gdzie pierwsza wariacja funkcjonału δQ jest względem wariacji $\delta\underline{x}$, $\delta\underline{u}$ funkcjonałem liniowym i wyraża się wzorem (487), zaś druga wariacja funkcjonału $\delta^2 Q$ jest względem wariacji $\delta\underline{x}, \delta\underline{u}$ funkcjonałem biliniowym, dającym się przedstawić w postaci

$$\delta^2 Q = \left[\delta\underline{x}'(t_k)\frac{\partial^2 f_k(\ldots)}{\partial\underline{x}'\,\partial\underline{x}} + \delta\underline{\psi}'(t_k)\right]\delta\underline{x}(t_k) -$$

$$- \int_{t_0}^{t_k}\left[\delta\underline{x}'\frac{\partial^2 H(\ldots)}{\partial\underline{x}'\,\partial\underline{x}} + \delta\underline{u}'\frac{\partial^2 H(\ldots)}{\partial\underline{u}'\,\partial\underline{x}} + \delta\underline{\psi}'\frac{\partial^2 H(\ldots)}{\partial\underline{\psi}'\,\partial\underline{x}} + \delta\dot{\underline{\psi}}'\right]\delta\underline{x}\,dt -$$

$$- \int_{t_0}^{t_k}\left[\delta\underline{x}'\frac{\partial^2 H(\ldots)}{\partial\underline{x}'\,\partial\underline{u}} + \delta\underline{u}'\frac{\partial^2 H(\ldots)}{\partial\underline{u}'\,\partial\underline{u}} + \delta\underline{\psi}'\frac{\partial^2 H(\ldots)}{\partial\underline{\psi}'\,\partial\underline{u}}\right]\delta\underline{u}\,dt, \quad (503)$$

gdzie dla skrócenia oznaczeń, symbolem $(\ldots)$ zastąpiono zależność od $\underline{x}(t_k)$, t_k w przypadku funkcji f_k, zaś od $\underline{x}$, $\underline{u}$, t w przy-

padku hamiltanianu H. Wariacja $\delta \underline{u}$ jest dowolną przedziałami ciągłą funkcją czasu, zaś wariacja $\delta \underline{x}$ spełnia równanie różniczkowe

$$\delta \dot{\underline{x}} = \frac{\partial^2 H(\ldots)}{\partial \underline{\psi}' \partial \underline{x}} \delta \underline{x} + \frac{\partial^2 H(\ldots)}{\partial \underline{\psi}' \partial \underline{u}} \delta \underline{u}; \quad \delta \underline{x}(t_o) = \underline{0}, \qquad (504)$$

wynikające z linearyzacji równań stanu. Jeżeli dobierzemy teraz wariację $\delta \underline{\psi}$ tak, by spełniała równanie różniczkowe

$$\delta \dot{\underline{\psi}} = - \frac{\partial^2 H(\ldots)}{\partial \underline{x}' \partial \underline{x}} \delta \underline{x} - \frac{\partial^2 H(\ldots)}{\partial \underline{x}' \partial \underline{\psi}} \delta \underline{\psi} - \frac{\partial^2 H(\ldots)}{\partial \underline{x}' \partial \underline{u}} \delta \underline{u} \qquad (505)$$

przy warunkach końcowych

$$\delta \underline{\psi}(t_k) = - \frac{\partial^2 f_k(\ldots)}{\partial \underline{x}' \partial \underline{x}} \delta \underline{x}(t_k), \qquad (506)$$

to dwie pierwsze składowe drugiej wariacji funkcjonału (503) stają się równe zeru. Ponieważ jednocześnie można dobrać tak sam przebieg $\underline{\psi}$, by dwie pierwsze składowe pierwszej wariacji funkcjonału były równe zeru, przeto przyrost funkcjonału można przedstawić w postaci

$$\Delta Q = -\varepsilon \int_{t_o}^{t_k} \frac{\partial H(\ldots)}{\partial \underline{u}} \delta \underline{u} \, dt -$$

$$- \frac{1}{2} \varepsilon^2 \int_{t_o}^{t_k} \left[\delta \underline{x}' \frac{\partial^2 H(\ldots)}{\partial \underline{x}' \partial \underline{u}} + \delta \underline{\psi}' \frac{\partial^2 H(\ldots)}{\partial \underline{\psi}' \partial \underline{u}} + \delta \underline{u}' \frac{\partial^2 H(\ldots)}{\partial \underline{u}' \partial \underline{u}} \right] \delta \underline{u} \, dt + \ldots \qquad (507)$$

Gdyby problem optymalizacji był liniowo-kwadratowy, to rozwinięcie funkcjonału jakości na szereg Taylora nie zawierałoby dalszych składników i najkorzystniejszy przyrost sterowania $\varepsilon \delta \underline{u}$ (odpowiadający dokładnie różnicy pomiędzy sterowaniem optymalnym $\hat{\underline{u}}$ a założonym początkowo $\underline{u}$) mógłby być wyznaczony na drodze minimalizacji wyrażenia ΔQ względem $\varepsilon \delta u$. Mogłyby przy tym wyniknąć - wbrew założeniom - stosunkowo duże wartości przyrostu $\varepsilon \delta \underline{u}$, a mimo to wszystkie podane wyżej wzory pozostawałyby w mocy. Jeśli jednak problem optymalizacji nie jest liniowo-kwadratowy, to należy zadbać o to, by przyrost $\varepsilon \delta \underline{u}$ był dostatecznie mały. Można to osiągnąć przez uzupełnienie przyrostu ΔQ wyrażeniem, zależnym kwadratowo od $\varepsilon \delta \underline{u}$ poprzez diagonalną macierz $\Lambda = \frac{1}{\varrho} I$, gdzie współczynnik ϱ spełnia rolę analogiczną do współczynnika długości kroku w metodzie gradientu (chociaż nie jest mu równoważny)

$$\Delta Q_\Lambda = \Delta Q + \frac{\varepsilon^2}{2}\int_{t_0}^{t_k} \delta \underline{u}' \underline{\Lambda} \delta \underline{u}\, dt = \Delta Q + \frac{\varepsilon^2}{2\varrho}\int_{t_0}^{t_k} \delta \underline{u}'\, \delta \underline{u}\, dt \,. \qquad (508)$$

Wyrażenie ΔQ_Λ można przedstawić w postaci jednej całki, której funkcja podcałkowa zależy liniowo i kwadratowo od $\varepsilon\delta\underline{u}$. Należy przy tym podkreślić, że w wyrażeniu (507) wariacje $\delta\underline{x}$ i $\delta\underline{\psi}$ nie są niezależne od wariacji $\delta\underline{u}$; wręcz przeciwnie, stanowią one wynik operacji liniowych na $\delta\underline{u}$, określonych równaniami (504), (505), (506). Stąd też szukając pochodnej zupełnej wyrażenia typu $\varepsilon^2\, \delta\underline{x}' \cdot \frac{\partial^2 H(\ldots)}{\partial \underline{x}'\, \partial \underline{u}}\, \delta\underline{u}$ względem $\varepsilon\delta\underline{u}'$ uzyskamy w wyniku nie $\varepsilon \frac{\partial^2 H(\ldots)}{\partial \underline{u}'\, \partial \underline{x}}\, \delta\underline{x}$, jak by się mogło wydawać na pierwszy rzut oka, lecz $2\varepsilon \frac{\partial^2 H(\ldots)}{\partial \underline{u}'\, \partial \underline{x}}\, \delta\underline{x}$. Jeśli uwzględniając powyższe uwagi - dla każdej chwili czasu dobrać takie $\varepsilon\delta\underline{u}$, by funkcja podcałkowa w wyrażeniu ΔQ_Λ osiągała wartość minimalną

$$\varepsilon\delta\hat{\underline{u}}_\varrho = -\left[\frac{\partial^2 H(\ldots)}{\partial \underline{u}'\, \partial \underline{u}} - \frac{1}{\varrho}\underline{I}\right]^{-1}\left[\frac{\partial^2 H(\ldots)}{\partial \underline{u}'\, \partial \underline{x}}\varepsilon\delta\underline{x} + \frac{\partial^2 H(\ldots)}{\partial \underline{u}'\, \partial \underline{\psi}}\varepsilon\delta\underline{\psi} + \frac{\partial H(\ldots)}{\partial \underline{u}'}\right], \qquad (509)$$

to uzyska się minimum funkcjonału ΔQ_Λ, a więc - maksymalne zmniejszenie funkcjonału Q według jego przybliżenia za pomocą drugiej wariacji, z ograniczoną za pomocą macierzy $\frac{1}{\varrho}\underline{I}$ zmianą $\varepsilon\delta\underline{u}$. Jeśli bowiem $\varrho \to 0$, to i $\varepsilon\delta\underline{u}_\varrho \to 0$; jeśli $\varrho \to \infty$, to $\varepsilon \cdot \delta\underline{u}_\varrho \to \varepsilon \cdot \delta\underline{u}$ - czyli najkorzystniejszego przyrostu sterowania dla problemu liniowo-kwadratowego.

W metodzie opartej na powyższych rozważaniach, można wyodrębnić trzy warianty:

A) bez ograniczenia długości kroku, przy $\varrho = \infty$;

B) ze stałym, założonym z góry współczynnikiem długości kroku ϱ;

C) z poszukiwaniem najkorzystniejszej wartości współczynnika ϱ.

Przedstawimy tu procedurę jednej iteracji metody drugiej wariacji dla wariantu B); modyfikacje procedury dla innych wariantów są dość oczywiste. Iteracja zaczyna się od arbitralnie wybranego sterowania $\underline{u}^{(1)}$. Na początku i-tej iteracji należy wykonać punkty a. i b. procedury metody gradientu. Punkt c. procedury ulega rozszerzeniu;

c. Scałkować równania sprzężone (490), dla znanych $\underline{u}^{(i)}, \underline{x}^{(i)}, \underline{\psi}^{(i)}$ obliczając gradient - $\underline{\gamma}^{(i)} = \frac{\partial H(\ldots)}{\partial \underline{u}'}$ według wzoru (491)

oraz ewentualnie jego normę (492), służącą dla oceny odległości od rozwiązania optymalnego. Ponadto obliczyć macierze

$$\frac{\partial^2 H(\ldots)}{\partial \underline{\psi}' \partial \underline{u}}\ ;\quad \frac{\partial^2 H(\ldots)}{\partial \underline{u}' \partial \underline{x}}\ ;\quad \frac{\partial^2 H(\ldots)}{\partial \underline{x}' \partial \underline{x}}\ ;\quad \frac{\partial^2 H(\ldots)}{\partial \underline{x}' \partial \underline{\psi}}\ ;\quad \frac{\partial^2 H(\ldots)}{\partial \underline{u}' \partial \underline{\psi}}$$

jako funkcje czasu symbol $(\ldots)$ oznacza tu oczywiście zależność od $\underline{\psi}^{(i)}$, $\underline{x}^{(i)}$, $\underline{u}^{(i)}$, t, zaś macierze $\frac{\partial^2 H(\ldots)}{\partial \underline{x}' \partial \underline{u}}$; $\frac{\partial^2 H(\ldots)}{\partial \underline{\psi}' \partial \underline{u}}$; $\frac{\partial^2 H(\ldots)}{\partial \underline{\psi}' \partial \underline{x}}$ uzyskuje się przez transpozycję oraz macierz $\frac{\partial^2 f_k(\ldots)}{\partial \underline{x}' \partial \underline{x}}$

d. Scałkować równania jednorodne dla wariacji

$$\begin{aligned} \delta \underline{x}_s^{(i)} &= \underline{A}_{11}^{(i)} \delta \underline{x}_s + \underline{A}_{12}^{(i)} \delta \underline{\psi}_s \\ \delta \underline{\psi}_s^{(i)} &= \underline{A}_{21}^{(i)} \delta \underline{x}_s + \underline{A}_{22}^{(i)} \delta \underline{\psi}_s \end{aligned} \qquad (510)$$

gdzie indeks "s" oznacza rozwiązanie swobodne (równania jednorodnego), zaś

$$\begin{aligned} A_{11}^{(i)} &= \frac{\partial^2 H(\ldots)}{\partial \underline{\psi}' \partial \underline{x}} - \frac{\partial^2 H(\ldots)}{\partial \underline{\psi}' \partial \underline{u}} \left[\frac{\partial^2 H(\ldots)}{\partial \underline{u}' \partial \underline{u}} - \frac{1}{\varrho} \underline{I} \right]^{-1} \frac{\partial^2 H(\ldots)}{\partial \underline{u}' \partial \underline{x}}, \\ A_{12}^{(i)} &= - \frac{\partial^2 H(\ldots)}{\partial \underline{\psi}' \partial \underline{u}} \left[\frac{\partial^2 H(\ldots)}{\partial \underline{u}' \partial \underline{u}} - \frac{1}{\varrho} \underline{I} \right]^{-1} \frac{\partial^2 H(\ldots)}{\partial \underline{u}' \partial \underline{\psi}}, \\ A_{21}^{(i)} &= - \frac{\partial^2 H(\ldots)}{\partial \underline{x}' \partial \underline{x}} + \frac{\partial^2 H(\ldots)}{\partial \underline{x}' \partial \underline{u}} \left[\frac{\partial^2 H(\ldots)}{\partial \underline{u}' \partial \underline{u}} - \frac{1}{\varrho} \underline{I} \right]^{-1} \frac{\partial^2 H(\ldots)}{\partial \underline{u}' \partial \underline{x}}, \\ A_{22}^{(i)} &= - \frac{\partial^2 H(\ldots)}{\partial \underline{x}' \partial \underline{\psi}} + \frac{\partial^2 H(\ldots)}{\partial \underline{x}' \partial \underline{u}} \left[\frac{\partial^2 H(\ldots)}{\partial \underline{u}' \partial \underline{u}} - \frac{1}{\varrho} \underline{I} \right]^{-1} \frac{\partial^2 H(\ldots)}{\partial \underline{u}' \partial \underline{\psi}} \end{aligned} \qquad (511)$$

Równania (510) i macierze (511) wynikają oczywiście z podstawienia wzoru (509) do wzorów (504), (505) przy pominięciu składowej, zależnej od gradientu $\frac{\partial H(\ldots)}{\partial \underline{u}'}$.

Równania (510) należy scałkować przy warunkach początkowych $\delta \underline{x}_s^{(i)}(t_o) = \underline{0}$, $\delta \underline{\psi}_s^{(i)}(t_o) = \underline{e}_1 = [1, 0, 0, \ldots, 0]'$, a następnie powtórzyć całkowanie tych równań n-krotnie przy $\delta \underline{\psi}_s^{(i)}(t_o) = \underline{e}_j = [0, 0, \ldots 1, \ldots, 0]'$ aż do $\delta \underline{\psi}_s^{(i)}(t_o) = \underline{e}_n = [0, 0, \ldots, 1]'$. Kolejne wartości rozwiązań $\delta \underline{x}_s^{(i)}(t_k)$ wyznaczają tzw. macierz tranzycyjną

$\underline{\Phi}_{12}^{(i)}(t_k, t_o)$, zaś kolejne wartości rozwiązań $\delta\underline{\psi}_s^{(i)}(t_k)$ - macierz tranzycyjną $\underline{\Phi}_{22}^{(i)}(t_k, t_o)$; macierze te określają wartości końcowe wariacji $\delta\underline{x}_s^{(i)}$ i $\delta\underline{\psi}_s^{(i)}$ przy dowolnej postaci warunków początkowych $\delta\underline{\psi}_s(t_o)$ według wzorów

$$\delta\underline{\psi}_s^{(i)}(t_k) = \underline{\Phi}_{22}^{(i)}(t_k, t_o)\,\delta\underline{\psi}_s^{(i)}(t_o);$$

$$\delta\underline{x}_s^{(i)}(t_k) = \underline{\Phi}_{12}^{(i)}(t_k, t_o)\,\delta\underline{\psi}_s^{(i)}(t_o). \tag{512}$$

e. Scałkować równania pełne dla wariacji

$$\delta\dot{\underline{x}}^{(i)} = \underline{A}_{11}^{(i)}\delta\underline{x}^{(i)} + \underline{A}_{12}^{(i)}\delta\underline{\psi}^{(i)} - \underline{B}_1^{(i)}\underline{\ell}^{(i)},$$

$$\delta\dot{\underline{\psi}}^{(i)} = \underline{A}_{21}^{(i)}\delta\underline{x}^{(i)} + \underline{A}_{22}^{(i)}\delta\underline{\psi}^{(i)} - \underline{B}_2^{(i)}\underline{\ell}^{(i)}, \tag{513}$$

gdzie

$$\underline{B}_1^{(i)} = -\frac{\partial^2 H(\ldots)}{\partial\underline{\psi}'\partial\underline{u}}\left[\frac{\partial^2 H(\ldots)}{\partial\underline{u}'\partial\underline{u}} - \frac{1}{\rho}\underline{I}\right]^{-1},$$

$$\underline{B}_2^{(i)} = \frac{\partial^2 H(\ldots)}{\partial\underline{x}'\partial\underline{u}}\left[\frac{\partial^2 H(\ldots)}{\partial\underline{u}'\partial\underline{u}} - \frac{1}{\rho}\underline{I}\right]^{-1}. \tag{514}$$

Równania (513) i macierze (514) wynikają oczywiście z podstawienia wzoru (509) do wzorów (504), (505) z uwzględnieniem składowej, zależnej od gradientu $\frac{\partial H(\ldots)}{\partial\underline{u}'}$. Równania (513) należy scałkować przy zerowych warunkach początkowych $\delta\underline{x}^{(i)}(t_o) = \underline{0}$, $\delta\underline{\psi}^{(i)}(t_o) = \underline{0}$, obliczając tym samym wartości końcowe rozwiązania szczególnego równań pełnych $\delta\underline{x}_p^{(i)}(t_k)$ i $\delta\underline{\psi}_p^{(i)}(t_k)$.

f. Obliczyć właściwe warunki początkowe $\delta\underline{\psi}^{(i)}(t_o)$, zapewniające spełnienie warunków końcowych (zgodnych ze wzorem (506), posługując się wzorem

$$\delta\underline{\psi}^{(i)}(t_o) = -\left[\underline{\Phi}_{22}^{(i)}(t_k, t_o) + \frac{\partial^2 f_k(\ldots)}{\partial\underline{x}'\partial\underline{x}}\underline{\Phi}_{12}^{(i)}(t_k, t_o)\right]^{-1}\left[\delta\underline{\psi}_p^{(i)}(t_k) + \frac{\partial^2 f_k(\ldots)}{\partial\underline{x}'\partial\underline{x}}\delta\underline{x}_p^{(i)}(t_k)\right] \tag{515}$$

a następnie jeszcze raz scałkować pełne równanie (513) dla wariacji przy tych warunkach początkowych, wyznaczając właściwe wariacje $\delta \underline{x}^{(i)}$, $\delta \underline{\psi}^{(i)}$. Dalej obliczyć wariację $\delta \underline{u}^{(i)}$ według wzoru (515) kładąc $\varepsilon = 1$ oraz nowe sterowanie

$$\underline{u}^{(i+1)} = \underline{u}^{(i)} + \delta \underline{u}^{(i)} . \qquad (516)$$

Ze względu na to, że nie mamy kryterium wyboru sterowania u^1 dla pierwszej iteracji, musimy rozważyć trzy następujące przypadki zależne oczywiście od wybranego u^1:

I gdy macierz $\frac{\partial^2 H}{\partial \underline{u}' \partial \underline{u}}$ jest ujemnie określona,

II gdy macierz $\frac{\partial^2 H}{\partial \underline{u}' \partial \underline{u}}$ jest dodatnio określona,

III gdy macierz $\frac{\partial^2 H}{\partial \underline{u}' \partial \underline{u}}$ nie jest określona;

macierze $\frac{\partial^2 H}{\partial \underline{u}' \partial \underline{u}}$ obliczamy wzdłuż $\underline{u} = \underline{u}^i(t)$, gdzie $u^i(t)$ jest założonym dla i-tej iteracji sterowaniem.

ad I. W tym przypadku wektor $\delta u(t)$ wyznaczony ze wzoru (515) daje w istocie najmniejszą wartość ΔQ_A i ΔQ. Można więc stosować bez zastrzeżeń warianty A, B i C metody drugiej wariacji.

ad II. Ze względu na to, iż wzór (515) dla $\varphi \to \infty$ daje zamiast kierunku poprawy wskaźnika jakości, kierunek jego pogarszania się należy zmienić znak φ, wyznaczyć $\delta \underline{u}$ maksymalizujące przyrost wskaźnika jakości ograniczony do dwu pierwszych wariacji, zmienić na przeciwny wyznaczony kierunek $\delta \underline{u}$, czyli zamiast $\delta \underline{u}$ wziąć $-\delta \underline{u}$ jako kierunek poprawy sterowania. W przypadku tym należy więc wyznaczony w wariancie A kierunek $\delta \underline{u}$ zmienić na przeciwny, w wariancie B zmienić znak założonego φ i wyznaczony kierunek $\delta \underline{u}$ zmienić na przeciwny, w wariancie C wybierać najlepsze φ spośród $\varphi < 0$.

ad III. W tym przypadku δu wyznaczone ze wzoru (515) dla $\varphi \to \infty$ nie jest ani kierunkiem poprawy ani pogarszania się wskaźnika jakości. Wydaje się więc sensownym zaproponować co następuje:

Przy założonym dla iteracji, w której wystąpiła nieokreśloność macierzy $\frac{\partial^2 H}{\partial \underline{u}' \partial \underline{u}}$, sterowaniu wykonać tyle kroków metodą gradientu lub gradientu sprzężonego aż dostaniemy $\frac{\partial^2 H}{\partial \underline{u}' \partial \underline{u}}$ dla nowego sterowania określone dodatnio lub ujemnie. Propozycja ta jakkolwiek istotnie komplikuje algorytm obliczeniowy wydaje się

jednak do przyjęcia, gdyż metoda drugiej wariacji w przypadku I i II jest metodą bez porównania szybciej zbieżną niż metody gradientu i gradientu sprzężonego.

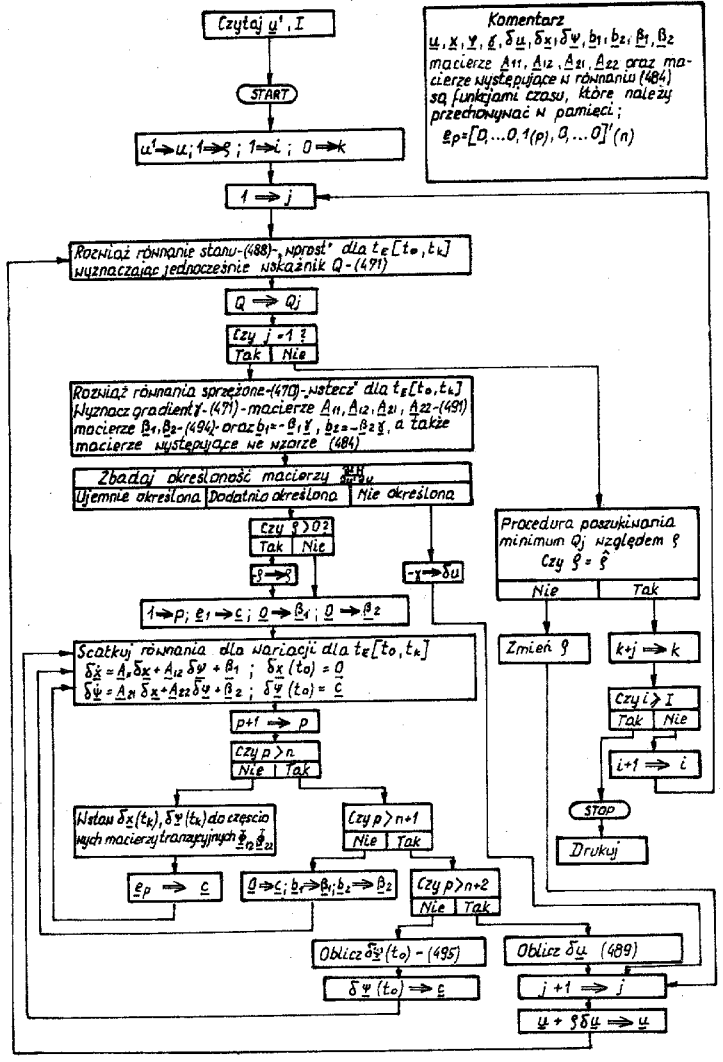
Zauważmy, że w wariancie A metody drugiej wariacji procedura obliczeniowa się nie zmienia, a jedynie w odpowiednich wzorach znika wyrażenie $\frac{1}{\varrho} I$. W wariancie C niezbędne byłoby wielokrotne powtarzanie punktów d, e, f procedury przy różnych ϱ oraz poszukiwanie najkorzystniejszej wartości ϱ. Dlatego też korzystniej jest stosować nieco odmienną postać wariantu z poszukiwaniem najkorzystniejszej długości kroku, a mianowicie wariant D: wyznaczać wariację $\underline{u}^{(i)}$ tak jak w wariancie A (bez ograniczenia długości kroku), zaś następnie poszukiwać najkorzystniejszego współczynnika w w równaniu

$$\underline{u}^{(i)} = \underline{u}^{(i)} + \varrho \delta \underline{u}^{(i)}. \qquad (5\text{.}17)$$

Schemat działania tego wariantu metody drugiej wariacji przedstawiono na rys. 78. Założono tu, że obliczenia kończą się po danej liczbie iteracji I. Widoczne jest, że struktura algorytmu jest podobna do struktury algorytmu gradientu - por. rys. 73. Wariant D metody drugiej wariacji różni się od metody gradientu zwykłego czy sprzężonego tylko tym, że wyznaczony jest możliwie najkorzystniejszy kierunek poszukiwań $\delta u^{(i)}$ - taki, który dla problemu liniowo-kwadratowego byłby kierunkiem idealnym. Jeśli norma gradientu jest niezerowa, to kierunek ten zapewnia zmniejszanie się funkcjonału jakości - a więc - zbieżność metody. Podobnie wariant C i wariant B przy dostatecznie małym ϱ są zawsze zbieżne; wariant A może nie być zbieżny, jeśli problem różni się znacznie od problemu liniowo-kwadratowego.

Jednakże dla większości problemów nieliniowych nawet wariant A jest bardzo szybko zbieżny, co jest główną zaletą metody; jej oczywistą wadą jest złożony program oraz duża pojemność pamięci i nakład obliczeń, niezbędnych dla dokonania jednej iteracji. Ponadto równania sprzężone dla wariacji rozwiązywane są w naturalnym kierunku biegu czasu, co uniemożliwia zastosowanie metody dla problemów z opóźnieniem oraz bardziej złożonych problemów sformułowanych w postaci dyskretnej, a także utrudnia jej zastosowanie we wszystkich przypadkach, gdy równania sprzężone są silnie niestabilne.

Podstawowe idee oraz własności metody drugiej wariacji wyjaśnia następujący prosty przykład, bezpośrednio nawiązujący do przykładów 1 i 2. Ponieważ rozważany problem jest liniowo-kwadratowy, zastosujemy tu wariant A metody, który powinien zapewnić określenie trajektorii optymalnej w ciągu jednej iteracji.


Czytaj u', I
START
u' ⇒ u; 1 ⇒ g ; 1 ⇒ i ; 0 ⇒ k
1 ⇒ j
Rozwiąż równanie stanu-(488)- wprost' dla t∈[t₀, tₖ] wyznaczając jednocześnie wskaźnik Q-(491)
Q ⇒ Qj
Czy j = 1 ?
Tak
Nie
Rozwiąż równania sprzężone-(470)- wstecz' dla t∈[t₀, tₖ] Wyznacz gradienty-(471) - macierze A11, A12, A21, A22-(491) macierze β1, β2-(494)- oraz b1 = -β1 γ, b2 = -β2 γ, a także macierze występujące we wzorze (484)
Zbadaj określoność macierzy ∂²H/∂u²
Ujemnie określona
Dodatnio określona
Nie określona
Czy g > 0 ?
Tak
Nie
g ⇒ g
-γ ⇒ δu
1 ⇒ p; e1 ⇒ c ; 0 ⇒ β1 ; 0 ⇒ β2
Scałkuj równania dla wariacji dla t∈[t₀, tₖ]
δẋ = A11 δx + A12 δψ + β1 ; δx (t₀) = 0
δψ̇ = A21 δx + A22 δψ + β2 ; δψ (tₖ) = c
p+1 ⇒ p
Czy p > n
Nie
Tak
Wstaw δx(tₖ), δψ(tₖ) do części wyr macierzy tranzycyjnych Φ11 Φ21
e_p ⇒ c
(Czy p > n+1
Nie
Tak
0 ⇒ c; b1 ⇒ β1; b2 ⇒ β2
Czy p > n+2
Nie
Tak
Oblicz δψ(t₀) - (495)
δψ (t₀) ⇒ c
Oblicz δu (489)
j+1 ⇒ j
u + g δu ⇒ u
Procedura poszukiwania minimum Qj względem g
Czy g = ĝ
Nie
Tak
Zmień g
k+j ⇒ k
Czy i > I
Tak
Nie
i+1 ⇒ i
STOP
Drukuj
Komentarz
u, x, ψ, g, δu, δx, δψ, b1, b2, β1, β2 macierze A11, A12, A21, A22 oraz macierze występujące w równaniu (484) są funkcjami czasu, które należy przechowywać w pamięci ;
e_p = [0 ... 0, 1(p), 0, ... 0]' (n)

Rys. 78

Przykład 3

Dany jest problem jak w przykładzie 1, którego rozwiązanie analityczne określają wzory (1.6), (1.7). Przypominamy, że hamiltonian problemu i jego gradient mają postać

$$H = -\frac{u^2 + x^2}{2} + \psi u; \qquad -\delta = \frac{\partial H}{\partial u} = \psi - u, \tag{3.1}$$

zaś równania stanu i równania sprzężone

$$\dot{x} = u; \qquad \dot{\psi} = x. \tag{3.2}$$

Zakładając $u^{(1)} = 0$, uzyskuje się w punktach a., b. i c. procedury

$$x^{(1)} = 0; \qquad \psi^{(1)}(T) = X; \qquad \delta^{(1)} = X \tag{3.3}$$

W punkcie c. procedury należy dodatkowo obliczyć macierze drugich pochodnych cząstkowych hamiltonianu. Są one w rozważanym przypadku stałe, niezależne od ψ, x, u oraz numeru iteracji

$$\frac{\partial^2 H}{\partial u^2} = -1; \quad \frac{\partial^2 H}{\partial x^2} = -1; \quad \frac{\partial^2 H}{\partial u\,\partial x} = 0;$$

$$\frac{\partial^2 H}{\partial \psi\,\partial x} = 0; \quad \frac{\partial^2 H}{\partial \psi\,\partial u} = 1 \tag{3.4}$$

i analogicznie, ponieważ $f_k = \frac{1}{2}\left[X - x(T)\right]^2$

$$\frac{\partial^2 f_k}{\partial x^2} = 1. \tag{3.5}$$

W punkcie d. należy wyznaczyć macierze tranzycyjne przez wielokrotne całkowanie równań jednorodnych. Rozważany problem jest pierwszego rzędu i równania jednorodne dla wariacji mają postać

$$\delta\dot{x} = \delta\psi$$

$$\delta\dot{\psi} = \delta x. \tag{3.6}$$

Wystarczy całkować tylko raz dla $\delta x(0) = 0$, $\delta\psi(0) = 1$; w równaniach tych nie występuje czynnik $\frac{1}{\varrho}$, gdyż w wariancie A

metody zakładamy $\varrho = \infty$ i równania (5.10), (5.11) ulegają odpowiednim uproszczeniom. Całkując równania (3.6), otrzymujemy jednoelementowe macierze tranzycyjne - skalary φ_{12}, φ_{22} - oraz rozwiązanie ogólne równań jednorodnych w postaci:

$$\varphi_{12} = \mathrm{sh}T\,; \qquad \varphi_{22} = \mathrm{ch}T\,; \qquad x_s(T) = \psi_s(o)\cdot \mathrm{sh}T;$$
$$\psi_s(T) = \psi_s(o)\cdot \mathrm{ch}T\,, \tag{3.7}$$

niezależnie od numeru iteracji. Widoczne jest, że gdyby całkowanie równań (3.6) odbywało się na drodze numerycznej, zaś wartości parametru T były duże, to rozwiązanie tych równań miałoby silnie niestabilny charakter, i należałoby precyzyjnie określić dokładność obliczeń początkowych wartości wariacji dla osiągnięcia pożądanej dokładności ich wartości końcowych oraz dla uniknięcia nadmiarów w maszynie cyfrowej.

W punkcie e. procedury należy scałkować równania niejednorodne dla wariacji

$$\begin{aligned} \delta\dot{x} &= \delta\psi - \gamma \\ \delta\dot{\psi} &= \delta x \end{aligned} \tag{3.8}$$

i uzyskać rozwiązanie szczególne dla $\delta x(o) = 0$, $\delta\psi(o) = 0$.

Ponieważ $\gamma^{(1)} = \psi^{(1)} - u^{(1)} = X$, więc otrzymujemy

$$\begin{aligned} \delta x_p^{(1)} &= \int_0^t \mathrm{ch}(t-\tau)\cdot X\,d\tau = X\,\mathrm{sh}t\,; \\ \delta\psi_p^{(1)} &= \int_0^t \mathrm{sh}(t-\tau)\cdot X\,d\tau = X(\mathrm{ch}t - 1). \end{aligned} \tag{3.9}$$

W punkcie f. procedury podstawiamy uzyskane wartości $\delta x_p^{(1)}(T)$ i $\delta\psi_p^{(1)}(T)$ do wzoru (1.89), określając właściwą wartość warunku początkowego $\delta\psi^{(1)}(o)$

$$\delta\psi^{(1)}(o) = -(\mathrm{ch}T + \mathrm{sh}T)^{-1}\cdot X(-1 + \mathrm{ch}t + \mathrm{sh}t) = X(e^{-T} - 1), \tag{3.10}$$

a następnie całkujemy równania (3.8) przy tym warunku początkowym oraz $\delta x(o) = 0$, uzyskując

$$\begin{aligned} \delta x^{(1)} &= X(e^{-T} - 1)\,\mathrm{sh}t + X\mathrm{sh}t = Xe^{-T}\mathrm{sh}t, \\ \delta\psi^{(1)} &= X(e^{-T} - 1)\,\mathrm{ch}t + X(\mathrm{ch}t - 1) = Xe^{-T}\mathrm{ch}t - X \end{aligned} \tag{3.11}$$

oraz określając wariację $\delta u^{(1)}$ według wzoru (509) przy $\varrho = \infty$

$$\delta u^{(1)} = \delta\psi^{(1)} + \delta^{(1)} = Xe^{-T}cht. \qquad (3.12)$$

Dane początkowe dla drugiej iteracji

$$u^{(2)} = Xe^{-T}cht; \qquad x^{(2)} = Xe^{-T}sht, \qquad (3.13)$$

stanowią już rozwiązanie optymalne.

W przykładzie powyższym widoczne są zarówno zalety, jak i wady metody drugiej wariacji - z jednej strony bardzo szybka zbieżność, zaś z drugiej - konieczność rozwiązywania równań sprzężonych dla wariacji w naturalnym kierunku biegu czasu.

(koniec przykładu 3).

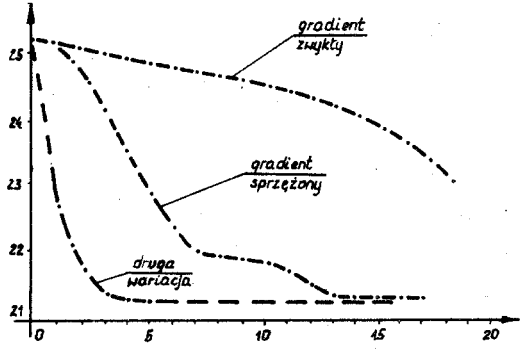
W artykule [11] podane są między innymi wyniki obliczeń numerycznych metodami gradientu zwykłego, sprzężonego i drugiej wariacji dla problemu

$$\dot{x}_1 = \left(1 - x_2^2\right)x_1 - x_2 + u; \quad x_1(o) = 0,$$

$$\dot{x}_2 = x_1 \qquad ; \quad x_2(o) = 3, \qquad (518)$$

$$Q = \int_0^{10}\left(x_1^2 + x_2^2 + u^2\right)dt; \quad f_k = 0.$$

Wartości wskaźnika jakości w kolejnych iteracjach dla tych trzech metod przedstawia rys. 79.



Rys. 79

Widoczna jest znaczna przewaga szybkości zbieżności metody drugiej wariacji nad metodami gradientu.

9.2.5. Metody pośrednie

Pośrednie metody podstawowe stanowią w istocie rzeczy rozmaite metody rozwiązywania równań różniczkowych ekstremal, dla których dane są zarówno warunki początkowe, jak i pewne warunki końcowe. Jeśli bowiem znana jest analityczna zależność sterowania optymalnego od zmiennych stanu i zmiennych sprzężonych w postaci (483), to podstawiając ją do równań stanu i równań sprzężonych uzyskuje się kanoniczne równania ekstremal

$$\begin{aligned} \dot{\underline{x}} &= \underline{f}\left(\underline{x}, \underline{\varphi}(\underline{\psi}, \underline{x}, t), t\right), \\ \dot{\underline{\psi}} &= \frac{\partial f_o}{\partial \underline{x}'}\left(\underline{x}, \underline{\varphi}(\underline{\psi}, \underline{x}, t), t\right) - \frac{\partial \underline{f}'}{\partial \underline{x}'}\left(x, \varphi(\psi, x, t), t\right)\underline{\psi}, \end{aligned} \tag{519}$$

gdzie różniczkowanie funkcji f_o i $\underline{f}$ dotyczy oczywiście tylko ich bezpośredniej zależności od stanu $\underline{x}$, a nie zależności pośredniej, przez funkcję $\underline{\varphi}$. Równania (519) dogodnie jest niekiedy rozpatrywać w ujednoliconym zapisie

$$\underline{z} = \underline{h}(\underline{z}, t); \qquad \underline{z} = \left\{\underline{x}, \underline{\psi}\right\}; \qquad \dim \underline{z} = 2n \tag{520}$$

Warunki początkowe dla $\underline{x}(t_o)$ są dane, $\underline{\psi}(t_o)$ - swobodne. Warunki końcowe są rozmaitej postaci i wynikają z warunków transwersalności. Przeciwnie niż w metodach bezpośrednich, dla metod pośrednich dogodnie jest zakładać, że koszty końcowe nie występują w zadaniu, a więc $f_k \equiv 0$. Jeśli tak nie jest, to można zmodyfikować zadanie pierwotne, przyjmując nową funkcję podcałkową wskaźnika jakości o postaci

$$f_o^m(\underline{x}, \underline{u}, t) = f_o(\underline{x}, \underline{u}, t) + \frac{\partial f_k(\underline{x}, t)}{\partial \underline{x}}\,\underline{f}(\underline{x}, \underline{u}, t) + \frac{\partial f_k(\underline{x}, t)}{\partial t} \tag{521}$$

i pomijając koszty końcowe $f_k\left(\underline{x}(t_k), t_k\right)$. Nowy wskaźnik jakości różni się od pierwotnego jedynie o stałą wartość $f_k\left(\underline{x}(t_o), t_o\right)$, którą łatwo obliczyć z danych zadania. Możliwe jest przy tym rozpatrywanie zarówno zadań ze swobodnym, jak i danym czasem końcowym t_k; ograniczymy się tu jedynie do omówienia problemów z danym t_k, zakładając warunki końcowe dla ekstremal

$$\underline{g}_k\left(\underline{x}(t_k)\right) = \underline{0}; \quad \frac{\partial \underline{g}_k}{\partial \underline{x}}\left(\underline{x}(t_k)\right)\delta \underline{x}_k = \underline{0}; \qquad \dim \underline{g}_k = k \leqslant n, \tag{522a}$$

$$\underline{\psi}'(t_k)\ \delta \underline{x}_k = 0. \tag{522b}$$

Równanie (522a) określa pewną (n-k)-wymiarową rozmaitość dopuszczalnych stanów końcowych. Warunek transwersalności (522b) wymaga, aby końcowy wektor sprzężony był normalny do tej rozmaitości; jest on zatem równoważny (n-k) warunkom, wiążącym składowe wektora $\underline{\psi}(t_k)$ i $\underline{x}(t_k)$. W sumie uzyskuje się zatem n warunków, określających w pełni rozwiązanie równań różniczkowych (519) łącznie z n warunkami początkowymi. Warunki transwersalności (522b) należy na ogół rozwikłać na drodze analitycznej, doprowadzając je łącznie z równaniami (522a) do wspólnej postaci

$$\underline{g}\left(\underline{x}(t_k),\ \underline{\psi}(t_k)\right) = \underline{g}\left(\underline{z}(t_k)\right) = \underline{0}, \qquad \dim\ \underline{g} = n. \tag{523}$$

Zadanie odszukania właściwej ekstremali problemu jest więc następujące: należy rozwiązać równanie ekstremal (519), lub, co równoważne (520), przy warunkach początkowych $\underline{x}(t_o) - \underline{x}_o = \underline{0}$ oraz warunkach końcowych (523). Po wyznaczeniu optymalnych trajektorii stanu $\underline{x}$ i sprzężonej $\underline{\psi}$ określenie sterowania optymalnego według wzoru (480) jest już trywialne.

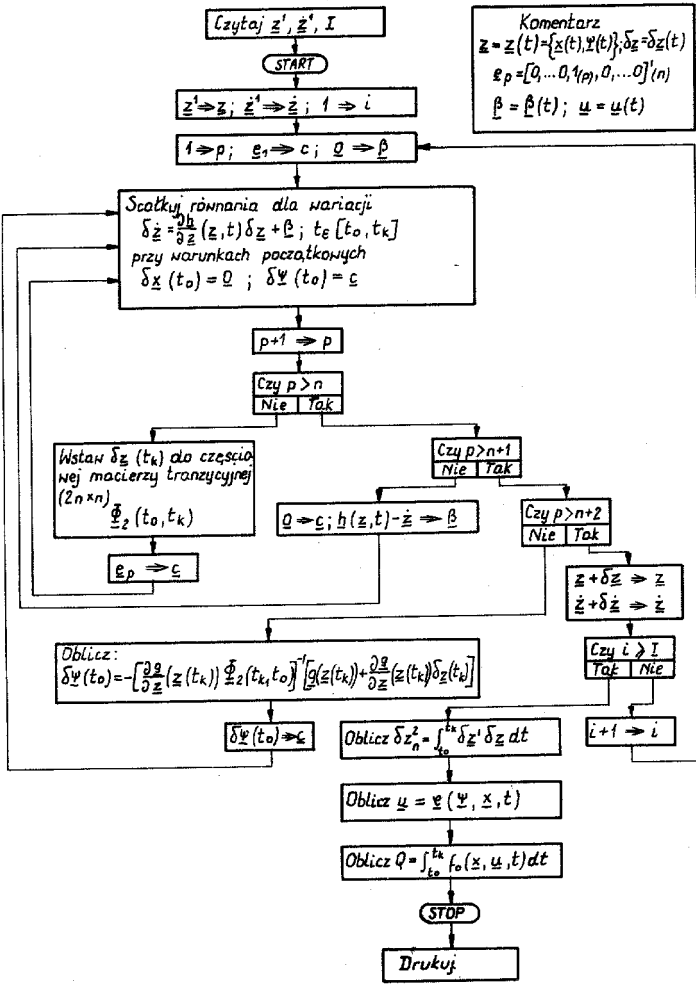
Zadanie powyższe można rozwiązać na wiele sposobów. Jeden z nich polega na wykorzystaniu metody Newtona w przestrzeni funkcyjnej. Formułuje się w niej równania jednorodne dla wariacji, stanowiące linearyzację równań (520)

$$\delta \dot{\underline{z}}_s = \frac{\partial h}{\partial \underline{z}}\ (\underline{z},\ t)\ \delta \underline{z}_s\ , \tag{524}$$

a następnie równania niejednorodne dla wariacji takie, które powinny być spełnione aby przy arbitralnie wybranym przebiegu $\underline{z}^{(i)}$ poprawiony przebieg $\underline{z}^{(i)} + \delta \underline{z}$ spełnił w pierwszym przybliżeniu równanie (520)

$$\delta \dot{\underline{z}} = \frac{\partial h}{\partial \underline{z}}\left(\underline{z}^{(i)},\ t\right) \cdot \delta \underline{z} + \underline{h}\left(\underline{z}^{(i)},\ t\right) - \dot{\underline{z}}^{(i)}. \tag{525}$$

Wyznaczając częściowe macierze tranzycyjne na podstawie równania (524) i znajdując rozwiązanie szczególne równania (525) przy warunkach początkowych zerowych, można tak dobrać warunki początkowe na δ_z, aby przebieg $z + \delta z$ spełnił w pierwszym przybliżeniu także i warunki końcowe (523); metoda postępowania omówiona dokładnie w [16] jest tu analogiczna do opisanej w procedurze obliczeniowej metody drugiej wariacji. Dla lepszego zrozumienia metody Newtona w przestrzeni funkcyjnej zaleca się przeanalizowanie jej schematu działań, przedstawionego na rys. 80, porównując go z rys. 78. W schemacie tym przyjęto zakończenie obliczeń po danej liczbie iteracji I; po zakończeniu iteracji oblicza


Komentarz
$\underline{z} = \underline{z}(t) = [\underline{x}(t), \underline{\Psi}(t)]$; $\delta\underline{z} = \delta\underline{z}(t)$
$\underline{e}_p = [0, \ldots 0, 1_{(p)}, 0, \ldots 0]'_{(n)}$
$\underline{\beta} = \underline{\beta}(t)$; $\underline{u} = \underline{u}(t)$
Czytaj $\underline{z}^1, \dot{\underline{z}}^1, I$
START
$\underline{z}^1 \Rightarrow \underline{z}$; $\dot{\underline{z}}^1 \Rightarrow \dot{\underline{z}}$; $1 \Rightarrow i$
$1 \Rightarrow p$; $\underline{e}_1 \Rightarrow \underline{c}$; $\underline{0} \Rightarrow \underline{\beta}$
Scałkuj równania dla wariacji
$\delta\dot{\underline{z}} = \frac{\partial \underline{h}}{\partial \underline{z}}(\underline{z}, t)\delta\underline{z} + \underline{\beta}$; $t_e [t_0, t_k]$
przy warunkach początkowych
$\delta\underline{x}(t_0) = \underline{0}$; $\delta\underline{\Psi}(t_0) = \underline{c}$
$p+1 \Rightarrow p$
Czy $p > n$
Nie | Tak
Wstaw $\delta\underline{z}(t_k)$ do części macierzy tranzycyjnej $(2n \times n)$
$\underline{\Phi}_z(t_0, t_k)$
$\underline{e}_p \Rightarrow \underline{c}$
Czy $p > n+1$
Nie | Tak
$\underline{0} \Rightarrow \underline{c}$; $\underline{h}(\underline{z}, t) - \dot{\underline{z}} \Rightarrow \underline{\beta}$
Czy $p > n+2$
Nie | Tak
$\underline{z} + \delta\underline{z} \Rightarrow \underline{z}$
$\dot{\underline{z}} + \delta\dot{\underline{z}} \Rightarrow \dot{\underline{z}}$
Oblicz:
$\delta\underline{\Psi}(t_0) = -\left[\frac{\partial \underline{g}}{\partial \underline{z}}(\underline{z}(t_k))\underline{\Phi}_z(t_k, t_0)\right]^{-1}\left[\underline{g}(\underline{z}(t_k)) + \frac{\partial \underline{g}}{\partial \underline{z}}(\underline{z}(t_k))\delta\underline{z}(t_k)\right]$
Czy $i > I$
Tak | Nie
$\delta\underline{\Psi}(t_0) \Rightarrow \underline{c}$
Oblicz $\delta z_n^2 = \int_{t_0}^{t_k} \delta\underline{z}' \delta\underline{z}\, dt$
$i+1 \Rightarrow i$
Oblicz $\underline{u} = \underline{e}(\underline{\Psi}, \underline{x}, t)$
Oblicz $Q = \int_{t_0}^{t_k} f_0(\underline{x}, \underline{u}, t)dt$
STOP
Drukuj

Rys. 80

się sterowanie, wskaźnik jakości oraz normę ostatniej wariacji $\delta \underline{z}$. Punktem wyjścia dla metody Newtona jest więc arbitralnie wybrana absolutnie ciągła funkcja czasu $\underline{z}$, spełniająca dane warunki początkowe, lecz niekoniecznie spełniająca warunki końcowe (523) i oczywiście na ogół nie spełniająca równań ekstremal (520). Następnie obliczana jest odpowiednia poprawka - wariacja $\delta \underline{z}$. Jeśli równania ekstremal (520) i warunki końcowe (523) są liniowe, to sposób obliczania tej wariacji powoduje, że już po pierwszej iteracji funkcja czasu $\underline{z} + \delta \underline{z}$ spełnia te równania i warunki, stanowiąc rozwiązanie problemu; zachodzi to między innymi dla wszystkich liniowo-kwadratowych problemów optymalizacji. Jeśli równanie ekstremal lub warunki końcowe są nieliniowe, to należy dokonywać kolejnych iteracji.

Można udowodnić - por. [7] - że metoda Newtona jest zbieżna i to niezwykle szybko, jeśli tylko założone przybliżenie początkowe jest dostatecznie bliskie rozwiązania. Dla przybliżeń początkowych odległych od rozwiązania metoda Newtona może być rozbieżna, co stanowi jej podstawową wadę. Drugą jej wadę - wspólną z większością metod pośrednich oraz z metodą drugiej wariacji - stanowią omówione wyżej niedogodności, wynikające z rozwiązywania równań sprzężonych w naturalnym kierunku biegu czasu.

Podstawowe idee i własności metody Newtona w przestrzeni funkcyjnej wyjaśnia następujący prosty przykład - nawiązujący do przykładów poprzednich.

Przykład 4

Dany jest problem jak w przykładzie 1, którego rozwiązania analityczne określają wzory (1.6), (1.7). W problemie występuje wprawdzie niezerowa funkcja kosztów końcowych f_k, nie stwarza to jednak komplikacji, gdyż warunki transwersalności mają stosunkowo prostą postać. Sterowanie optymalne, maksymalizujące hamiltonian (1.3) daje się wyrazić w postaci

$$\underline{u} = \varphi(\psi, x) = \psi \tag{4.1}$$

Po podstawieniu tego sterowania do równań stanu (1.1) i sprzężonych (1.4) uzyskuje się równania ekstremal

$$\begin{aligned} \dot{x} &= \psi \\ \dot{\psi} &= x, \end{aligned} \tag{4.2}$$

stanowiące przykład równań (519) lub (520). Warunki transwersalności (1.5) można przepisać w postaci analogicznej do (523)

$$g\big(x(T), \psi(T)\big) = X - x(T) - \psi(T) = 0. \tag{4.3}$$

Należy więc rozwiązać równanie (4.2) przy warunku $x(0) = 0$, tak dobierając $\psi(0)$, by spełniony był warunek (4.3). Ponieważ

równania (4.2) są liniowe, w zasadzie nie jest tu konieczne zastosowanie metody Newtona i wprowadzenie równań dla wariacji; jednakże będziemy tu postępować tak, jak gdybyśmy mieli do czynienia z ogólnym przypadkiem nieliniowych równań ekstremal. Ponieważ funkcja $\underline{h}$ dla równań (4.2) i odpowiednia macierz $\frac{\partial \underline{h}}{\partial \underline{z}}$ mają postać

$$\underline{z} = \{x, \psi\}\ ; \quad \underline{h}(x, \psi) = [\psi,\ x]'; \quad \frac{\partial \underline{h}}{\partial \underline{z}} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{4.4}$$

więc równania jednorodne (524) dla wariacji przyjmują postać

$$\begin{aligned} \delta\dot{x}_s &= \delta\psi_s\,, \\ \delta\dot{\psi}_s &= \delta x_s, \end{aligned} \tag{4.5}$$

(są one w istocie równoważne równaniom ekstremal (4.2), co zachodzi oczywiście dla wszystkich liniowych jednorodnych równań ekstremal), zaś równania niejednorodne (525) dla wariacji wyrażają się wzorem

$$\begin{aligned} \delta\dot{x} &= \delta\psi + \psi^{(1)} - \dot{x}^{(1)}, \\ \delta\dot{\psi} &= \delta x + x^{(1)} - \dot{\psi}^{(1)}, \end{aligned} \tag{4.6}$$

gdzie: $\dot{x}^{(1)}$, $x^{(1)}$, $\dot{\psi}^{(1)}$, $\psi^{(1)}$ - początkowe, arbitralnie założone trajektorie stanu i sprzężone, niekoniecznie spełniające równania (4.2) i (4.3).

Podobnie jak w przykładzie (3) - por. wzór (3.7) wyznaczymy częściowe macierze tranzycyjne dla równania (4.5)

$$\delta x_s(T) = \delta\psi_s(0)\,\mathrm{sh}T; \quad \delta\psi_s(T) = \delta\psi_s(0)\,\mathrm{ch}T. \tag{4.7}$$

Gdybyśmy wyznaczali te macierze na drodze numerycznej, wystąpiłyby omawiane wyżej trudności z rozwiązywaniem niestabilnych równań różniczkowych.

Zakładając na przykład $x^{(1)} = 0$, $\psi^{(1)} = \frac{X}{2}$ (przy czym warunki początkowe dla $x(0)$ są spełnione, ale warunek transwersalności (4.3) nie jest spełniony) uzyskujemy rozwiązanie szczególne równań niejednorodnych dla wariacji (4.6) przy

$$\delta x_p^{(1)}(0) = 0, \qquad \delta\psi_p^{(1)}(0) = 0, \tag{4.8}$$

$$\delta x_p^{(1)} = \frac{X}{2}\,\mathrm{sht}; \quad \delta\psi_p^{(1)} = \frac{X}{2}\,(\mathrm{cht} - 1), \tag{4.8}$$

analogicznie jak w przykładzie 3 - por. wzór (3.9). Wiedząc, że rozwiązanie ogólne równań (4.6) dla $t = T$ wyraża się wzorem

$$\begin{aligned} \delta x^{(1)}(T) &= \delta x_s(T) + \delta x_p^{(1)}(T) = \left[\frac{X}{2} + \delta\psi_s(0)\right]\mathrm{sh}T, \\ \delta\psi^{(1)}(T) &= \delta\psi_s(T) + \delta\psi_p^{(1)}(T) = \left[\frac{X}{2} + \delta\psi_s(0)\right]\mathrm{ch}T - \frac{X}{2}, \end{aligned} \tag{4.9}$$

należy teraz tak dobrać warunek początkowy $\delta\psi_s(0)$, aby trajektorie $x^{(1)} + \delta x^{(1)}$, $\psi^{(1)} + \delta\psi^{(1)}$ spełniały warunek transwersalności (4.3), który można przepisać w postaci

$$X - \left[\frac{X}{2} + \delta\psi_s(0)\right]\mathrm{sh}T - \left[\frac{X}{2} + \delta\psi_s(0)\right]\mathrm{ch}T = 0, \tag{4.10}$$

skąd wynika

$$\delta\psi_s(0) = X\left(e^{-T} - \frac{1}{2}\right). \tag{4.11}$$

Zakładając warunki początkowe $\delta\psi^{(1)}(0) = \delta\psi_s(0)$ i $\delta x^{(1)}(0) = 0$ dla równania pełnego wariacji (4.6), uzyskujemy

$$\delta x^{(1)} = Xe^{-T}\mathrm{sht}; \qquad \delta\psi^{(1)} = Xe^{-T}\mathrm{cht} - \frac{X}{2} \tag{4.12}$$

oraz poprawione trajektorie stanu i sprzężone

$$x^{(2)} = x^{(1)} + \delta x^{(1)} = Xe^{-T}\mathrm{sh}T; \quad \psi^{(2)} = \psi^{(1)} + \delta\psi^{(1)} = Xe^{T}\mathrm{ch}t, \tag{4.13}$$

które spełniają już zarówno równania ekstremal (364), jak i warunki transwersalności (4.3). W tym więc konkretnym przykładzie osiągamy za pomocą metody Newtona rozwiązanie optymalne już po pierwszej iteracji.

(koniec przykładu 4)

Inną grupę metod rozwiązania zagadnienia wyznaczenia właściwej ekstremali problemu optymalizacji stanowią metody przeszukiwania ekstremal. Polegają one na początkowym arbitralnym założeniu warunków początkowych dla zmiennych sprzężonych $\psi(t_0) = \psi_0$, sprawdzeniu niezgodności warunków końcowych, wyrażających się przez wartości funkcji $g(x(t_k), \psi(t_k))$, a następnie na takiej or-

ganizacji przeszukiwania warunków początkowych, aby spełnić dane warunki końcowe. Do organizacji przeszukiwania stosuje się dowolną metodę poszukiwania ekstremum funkcji wielu zmiennych, czyli optymalizacji statycznej; na ogół stosuje się metody bezgradientowe, np. Rosenbrocka. Wprowadza się w tym celu funkcję odchylenia od warunków końcowych, na przykład o postaci

$$K(\underline{\psi}_o) = \underline{g}'\left(\underline{x}(t_k, \underline{\psi}_o), \underline{\psi}(t_k, \underline{\psi}_o)\right) \underline{\Lambda}\, \underline{g}\left(\underline{x}(t_k, \underline{\psi}_o), \underline{\psi}(t_k, \underline{\psi}_o)\right), \qquad (526)$$

gdzie macierz $\underline{\Lambda}$ jest arbitralnie wybraną macierzą diagonalną o elementach dodatnich - na przykład jednostkową - oraz gdzie zaznaczono zależność wartości końcowych stanu i zmiennych sprzężonych od wartości początkowej $\underline{\psi}_o = \underline{\psi}(t_o)$.

Funkcja $K(\underline{\psi}_o)$ w większości przypadków nie jest określona w postaci analitycznej, a jej wartości mogą być wyznaczone jedynie przez numeryczne obliczenie wartości $\underline{x}(t_k, \underline{\psi}_o)$, $\underline{\psi}(t_k, \underline{\psi}_o)$ i podstawienie tych wartości do wzoru (526). Jeśli rozwiązanie problemu optymalizacji istnieje, to dodatnio określona funkcja $K(\underline{\psi}_o)$ posiada minimum globalne, równe oczywiście zeru jeśli $\underline{\psi}_o$ jest poszukiwaną, właściwą wartością początkową zmiennych sprzężonych. Wystarczy więc zastosować jakąkolwiek metodę optymalizacji statycznej dla określenia właściwej wartości $\underline{\psi}_o$.

Charakter zbieżności takich metod przeszukiwania ekstremal zależy więc od charakteru zbieżności odpowiedniej metody numerycznej optymalizacji statycznej. Ponieważ zastępczy problem statyczny jest n-wymiarowy, przeto zbieżność metod poszukiwania ekstremal jest z reguły znacznie szybsza, niż innych metod optymalizacji dynamicznej. Wadą metody jest natomiast konieczność wielokrotnego rozwiązywania równań ekstremal, które w przypadku ogólnym są nieliniowe i zazwyczaj niestabilne. Fakt ten utrudnia ocenę zakresu zmienności współrzędnych stanu i sprzężonych, niezbędną przy maszynowym rozwiązywaniu problemu; ocena tego zakresu jest stosunkowo prosta jedynie w przypadku niestabilnych równań liniowych. W tym więc punkcie metoda Newtona, wymagająca rozwiązywania tylko równań liniowych, ma przewagę nad metodami przeszukiwania ekstremal. Jeśli jednak zagadnienie oceny zakresów zmienności współrzędnych stanu i sprzężonych jest już rozwiązane, to metody przeszukiwania ekstremal wymagają zazwyczaj znacznie mniejszego nakładu obliczeń niż metoda Newtona czy jakiekolwiek inne metody.

Podstawowe idee i własności metody przeszukiwania ekstremal wyjaśnia następujący przykład nawiązujący do przykładów poprzednich.

Przykład 5

Dany jest problem jak w przykładzie 1 i dalszych, którego rozwiązanie analityczne określają wzory (1.6), (1.7). Równania

ekstremal dla tego problemu mają postać - por. wzór (4.2)

$$\dot{x} = \psi$$
$$\dot{\psi} = x, \tag{5.1}$$

zaś warunek końcowy

$$g\ x(T),\ \psi(T) = X - x(T) - \psi(T) = 0. \tag{5.2}$$

Zależność rozwiązań równań ekstremal od warunku początkowego $\psi(0)$ może być np. w tym prostym przykładzie wyznaczona w postaci analitycznej. Przyjmując $x(o) = 0$, $\psi(o) = \psi_o$ uzyskujemy

$$x(T) = \psi_o\ shT; \quad \psi(T) = \psi_o\ cht. \tag{5.3}$$

Przyjmując we wzorze (376) macierz $\underline{A} = 1$, wyznaczamy funkcję $K(\psi_o)$ w postaci

$$K(\psi_o) = (X - \psi_o\ shT - \psi_o\ chT)^2 = (X - \psi_o\ e^T)^2. \tag{5.4}$$

W tak prostym zadaniu znajdujemy natychmiast wartość ψ_o zapewniającą minimum funkcji $K(\psi_o)$

$$\psi_o = Xe^{-T} \tag{5.5}$$

oraz odpowiadającą tej wartości właściwą ekstremalę

$$\hat{x} = Xe^{-T}sht; \quad \hat{\psi} = Xe^{-T}cht, \tag{5.6}$$

stanowiącą rozwiązanie zagadnienia optymalizacji i wyznaczającą sterowanie optymalne $\hat{u} = \hat{\psi}$.

(koniec przykładu 5).

Z przedstawionych wyżej przykładów wynika jasno przewaga metod przeszukiwania ekstremal nad metodą Newtona, a także innymi metodami numerycznymi optymalizacji dynamicznej w przypadku, gdy równania ekstremal są liniowe i pod warunkiem, że ich niestabilność nie nastręcza większych trudności. Oczywiście, przewaga metod przeszukiwania ekstremal nie dotyczy przypadków problemów z opóźnieniem czy nietrywialnych problemów dyskretnych, dla których metody przeszukiwania ekstremal czy metoda Newtona nie mogą być w ogóle zastosowane ze względu na kierunek rozwiązywania równań sprzężonych.

9.3. Przykład dwupoziomowej metody optymalizacji

Idea wielopoziomowego wyznaczania sterowania optymalnego jest dość naturalna. Nakład obliczeń związany z określeniem ste-

rowania optymalnego rośnie zazwyczaj z kwadratem lub sześcianem wymiarowości problemu. Rozbicie (dekompozycja) problemu całościowego na szereg problemów częściowych może więc znacznie obniżyć nakład obliczeń. Niezbędne jest jednak przy tym - z wyjątkiem przypadków trywialnych - zastosowanie algorytmu nadrzędnego (wyższego poziomu), koordynującego wyznaczania sterowania optymalnego dla problemów częściowych tak, aby wyznaczone w końcu sterowanie było również optymalne dla problemu całościowego. Nadrzędny algorytm koordynacji jest zwykle algorytmem iteracyjnym, wymagającym powtarzania rozwiązań problemów częściowych. Oczywistym warunkiem zmniejszenia nakładu obliczeń w metodzie wielopoziomowej w porównaniu z wybraną metodą jednopoziomową jest, by zmniejszenie nakładu obliczeń na skutek dekompozycji przeważało nad dodatkowym nakładem obliczeń związanym z koordynacją.

Niech na przykład problem całościowy, rozwiązywany metodą jednopoziomową, wymaga nakładu obliczeń

$$J_c = J_o \, n^{\alpha} \tag{527}$$

gdzie: J_o - współczynnik,
n - wymiarowość problemu,
α - wykładnik o wartości 2 do 3.

Załóżmy dalej, że problem ten można podzielić na p problemów częściowych o wymiarach $\frac{n}{p}$ (założenie to jest dość silne, gdyż przy dekompozycji często wzrasta łączna wymiarowość problemów częściowych). Rozwiązanie dwupoziomowe tego problemu będzie wymagało nakładu obliczeń

$$J_d = k \, p \left[J_o \left(\frac{n}{p} \right)^{\alpha} + J_1 \right], \tag{528}$$

gdzie: k - ilość powtórzeń procedury koordynacji rozwiązań, zaś
J_1 - nakład obliczeń związany z koordynacją w danej procedurze.

Zazwyczaj J_1 jest pomijalnie małe, a więc zmniejszanie nakładu obliczeń dzięki metodzie dwupoziomowej wyraża się wzorem

$$\frac{J_d}{J_c} \approx k \, p^{-\alpha+1}; \quad \frac{J_d}{J_c} < 1, \quad \text{jeśli} \quad k < p^{\alpha-1}. \tag{529}$$

W literaturze - np. [9], [14] - podano szereg algorytmów wielopoziomowych optymalizacji dynamicznej. Tutaj przedstawiony będzie jedynie pewien szczególny przypadek algorytmu wielopoziomowego. Dotyczy on problemów, w których wskaźnik jakości nie daje się przedstawić w postaci sumy wskaźników jakości problemów częściowych (jest nieaddytywny).

Zakładamy, że problem całościowy daje się podzielić na p problemów częściowych, oznaczonych indeksem β ($\beta = a, \ldots p$), o zmiennych stanu $\underline{x}_\beta$ i sterowaniach $\underline{u}_\beta$. Dla uproszczenia zakładamy, że procesy częściowe mają niezależne równania stanu *)

$$\dot{\underline{x}}_\beta = \underline{f}_\beta(\underline{x}_\beta, \underline{u}_\beta, t); \qquad \beta = a, \ldots p \tag{530}$$

oraz niezależne warunki początkowe i końcowe.

Wskaźnik jakości natomiast zawiera składową, zależną od stanów i sterowań wszystkich procesów częściowych

$$Q = \int_{t_0}^{t_k} \left[\sum_{\beta=a}^{p} f_{o\beta}(\underline{x}_\beta, \underline{u}_\beta, t) + f_{ow}(\underline{x}, \underline{u}, t) \right] dt, \tag{531}$$

gdzie $\underline{x}$ i $\underline{u}$ są całościowymi wektorami stanu i sterowania

$$\underline{x} = \left\{ x_a, \ldots x_\beta, \ldots x_p \right\}; \qquad \underline{u} = \left\{ u_a, \ldots u_\beta, \ldots u_p \right\}. \tag{532}$$

Wskaźnik (531) może być podzielony na szereg wskaźników częściowych przez wprowadzenie zmiennych koordynacyjnych - funkcji czasu $\underline{\pi}_{x\beta}$, $\underline{\pi}_{u\beta}$

$$Q_\gamma = \int_{t_0}^{t_k} \left[f_{o\beta}(\underline{x}_\beta, \underline{u}_\beta, t) + \underline{\pi}'_{x\beta} \cdot \underline{x}_\beta + \underline{\pi}'_{u\beta} \cdot \underline{u}_\beta \right] dt. \tag{533}$$

Całościowe wektory zmiennych koordynacyjnych mogą być zapisane w całości

$$\underline{\pi}x = \left\{ \underline{\pi}_{xa} \cdots \underline{\pi}_{x\beta} \cdots \underline{\pi}_{xp} \right\}; \quad \underline{\pi}u = \left\{ \underline{\pi}_{ua} \cdots \underline{\pi}_{u\beta} \cdots \underline{\pi}_{up} \right\} \tag{534}$$

Algorytm dwupoziomowej optymalizacji dla takiego problemu jest następujący. Przy założonych zmiennych koordynacyjnych $\underline{\pi}_x^{(i)}$, $\underline{\pi}_u^{(i)}$, w i-tej procedurze koordynacji rozwiązuje się za pomocą jakiejkolwiek metody podstawowej p niezależnych problemów częściowych poszukiwania minimum funkcjonałów (533) przy więzach różniczkowych (530) i odpowiednich warunkach krańcowych. W wyniku uzyskuje się trajektorie stanu $x^{(i)}$ i sterowania $u^{(i)}$, które pozwalają na ulepszenie założonych zmiennych koordynacyjnych według wzoru

---

*) Założenie to nie jest konieczne, jednakże nie przyjmując go należałoby tu omówić także i inne metody wielopoziomowe, które należy wykorzystać przy braku tego założenia.

$$\underline{\pi}_x^{(i+1)} = (1-\varrho)\underline{\pi}_x^{(i)} + \varrho \frac{\partial f_{ow}\left(\underline{x}^{(i)}, \underline{u}^{(i)}, t\right)}{\partial \underline{x}'},$$

$$\underline{\pi}_u^{(i+1)} = (1-\varrho)\underline{\pi}_u^{(i)} + \varrho \frac{\partial f_{ow}\left(\underline{x}^{(i)}, \underline{u}^{(i)}, t\right)}{\partial \underline{u}'}, \qquad (535)$$

gdzie ϱ - współczynnik długości kroku.

W pracy [16] wykazano, że algorytm koordynacji (535) jest zbieżny monotonicznie dla dostatecznie małych ϱ, czyli że wartości funkcjonału całościowego (531) maleją przy każdej iteracji, w której $\underline{\pi}_x^{(i+1)} \neq \underline{\pi}_x^{(i)}$ lub $\underline{\pi}_u^{(i+1)} \neq \underline{\pi}_u^{(i)}$; jeśli więc wartości funkcjonału (511) są ograniczone od dołu, to algorytm koordynacji (535) zapewnia dowolnie bliskie zbliżenie się do lokalnego lub globalnego kresu dowolnego funkcjonału.

Szczególnym przypadkiem algorytmu koordynacji (535) jest algorytm

$$\underline{\pi}_x^{(i+1)} = \frac{\partial f_{ow}\left(\underline{x}^{(i)}, \underline{u}^{(i)}, t\right)}{\partial \underline{x}'}; \quad \underline{\pi}_u^{(i+1)} = \frac{\partial f_{ow}\left(\underline{x}^{(i)}, \underline{u}^{(i)}, t\right)}{\partial \underline{u}'}, \qquad (536)$$

który może być także zbieżny monotonicznie - pod warunkiem, że funkcja f_{ow} jest wypukła względem swych argumentów $\underline{x}$ i $\underline{u}$ oraz, że przybliżenie początkowe $\underline{x}^{(1)}, \underline{u}^{(1)}$ nie jest zbyt odległe od rozwiązania całościowego $\underline{x}^{(\infty)}\ \underline{u}^{(\infty)}$

Nie analizując głębiej własności tych algorytmów, możliwości sformułowania algorytmów o liczbie poziomów większej, niż dwa itp. przedstawimy tu tylko poglądowo ich zastosowanie dla prostego przykładu, który może być zresztą rozwiązany bezpośrednio na drodze analitycznej.

Przykład 1

Rozpatrujemy problem optymalizacji o równaniach stanu i warunkach krańcowych

$$\dot{x}_a = k_a u_a - x_a; \quad x_a(o) = 0; \quad x_a(T) \text{ - swobodne},$$
$$\dot{x}_b = k_b u_b - x_b; \quad x_b(o) = 0; \quad x_b(T) \text{ - swobodne} \qquad (1.1)$$

i o całościowym wskaźniku jakości

$$Q = \frac{1}{2}\left\{\left[X_a - x_a(T)\right]^2 + \left[X_b - x_b(T)\right]^2 + \int_0^T (u_a^2 + u_b^2 + u_a u_b)\,dt\right\} \qquad (1.2)$$

Zadanie poszukiwania minimum funkcjonału (6.2) przy więzach (6.1) może być rozwiązane na drodze analitycznej. W tym celu formułujemy hamiltonian

$$H = -\frac{1}{2}(u_a^2 + u_b^2 + u_a u_b) + \psi_a(k_a u_a - x_a) + \psi_b(k_b u_b - x_b), \qquad (1.3)$$

równania sprzężone

$$\dot{\psi}_a = \psi_a; \qquad \dot{\psi}_b = \psi_b, \qquad (1.4)$$

ich rozwiązania

$$\psi_a = \psi_{ao} e^t; \qquad \psi_b = \psi_{bo} e^t, \qquad (1.5)$$

oraz warunki transwersalności

$$\psi_{ao} e^T = X_a - x_a(T); \qquad \psi_{bo} e^T = X_b - x_b(T). \qquad (1.6)$$

Warunki na maksimum hamiltonianu względem sterowań mają postać

$$2\hat{u}_a + \hat{u}_b = 2\psi_a k_a; \qquad 2\hat{u}_b + \hat{u}_a = 2\psi_b k_b. \qquad (1.7)$$

Z warunków tych oraz z postaci zmiennych sprzężonych jako funkcji czasu (6.5) wynika, że można założyć sterowanie optymalne postaci

$$\hat{u}_a = \hat{u}_{ao} e^t; \qquad \hat{u}_b = \hat{u}_{bo} e^t, \qquad (1.8)$$

przy czym

$$\psi_{ao} = \frac{2\hat{u}_{ao} + \hat{u}_{bo}}{2k_a}; \qquad \psi_{bo} = \frac{2\hat{u}_{bo} + \hat{u}_{ao}}{2k_b}. \qquad (1.9)$$

Stosując sterowanie (6.8) do procesów (6.1) uzyskuje się

$$\hat{x}_a = k_a \hat{u}_{ao} \operatorname{sht}; \qquad \hat{x}_b = k_b \hat{u}_{bo} \operatorname{sht}, \qquad (1.10)$$

a więc warunki transwersalności (6.6) można przepisać w postaci:

$$\begin{aligned} 2(k_a \operatorname{sht} + e^T)\hat{u}_{ao} + e^T \hat{u}_{bo} &= 2k_a X_a, \\ 2(k_b \operatorname{sht} + e^T)\hat{u}_{bo} + e^T \hat{u}_{ao} &= 2k_b X_b, \end{aligned} \qquad (1.11)$$

z których wynikają optymalne wartości $\hat{u}_{ao}$, $\hat{u}_{bo}$:

$$\hat{u}_{ao} = \frac{k_a X_a (k_b \operatorname{sh} T + e^T) - k_b X_b \frac{e^T}{2}}{(k_a \operatorname{sh} T + e^T)(k_b \operatorname{sh} T + e^T) - \frac{e^{2T}}{4}}, \tag{1.12}$$

$$\hat{u}_{bo} = \frac{k_b X_b (k_a \operatorname{sh} T + e^T) - k_a X_a \frac{e^T}{2}}{(k_a \operatorname{sh} T + e^T)(k_b \operatorname{sh} T + e^T) - \frac{e^{2T}}{4}}.$$

Wartości te oraz wzory (6.8), (6.10) określają optymalne sterowanie oraz trajektorię procesu.

Zastosujemy teraz dla rozwiązania powyższego zadania opisany poprzednio algorytm dwupoziomowy; oczywiście, postępowanie takie ma jedynie na celu bardziej poglądowe przedstawienie istoty i własności tego algorytmu.

Po dekompozycji rozpatrujemy proces częściowy o równaniu stanu i wskaźniku jakości

$$\dot{x}_a = k_a u_a - x_a; \qquad x_a(o) = 0; \qquad x_a(T) \text{ - swobodne}, \tag{1.13}$$

$$Q_a = \frac{1}{2}\left\{\left[X_a - x_a(T)\right]^2 + \int_0^T (u_a^2 + \pi_a u_a)\,dt\right\} \tag{1.14}$$

oraz drugi proces częściowy o analogicznych równaniach. Załóżmy, że arbitralnie wybrana funkcja koordynacyjna π_a ma postać

$$\pi_a = \pi_{ao} e^t. \tag{1.15}$$

Przy analizie procedury koordynacji okaże się, że postać ta jest zachowywana w kolejnych iteracjach, przy czym wynika ona z założenia początkowego dla pierwszej iteracji $\pi_a^{(o)} = 0$. Możemy teraz rozwiązać w jakikolwiek sposób problemy częściowe.

Hamiltonian problemu częściowego ma postać

$$H_a = -\frac{1}{2}(u_a^2 + \pi_a u_a) + \psi_a (k_a u_a - x_a), \tag{1.16}$$

skąd wynikają równania sprzężone i ich rozwiązania

$$\dot{\psi}_a = \psi_a; \qquad \psi_a = \psi_{ao} e^t \tag{1.17}$$

oraz sterowanie, maksymalizujące hamiltonian

$$\hat{u}_a = \psi_a k_a - \frac{\pi_a}{2} = \left(\psi_{ao} k_a - \frac{\pi_{ao}}{2}\right) e^t = \hat{u}_{ao} e^t, \tag{1.18}$$

przy czym

$$\psi_{ao} = \frac{\hat{u}_{ao} + \frac{\pi_{ao}}{2}}{k_a}. \tag{1.19}$$

Trajektoria stanu ma postać

$$\hat{x}_a = k_a \hat{u}_{ao} \operatorname{sht}, \tag{1.20}$$

zaś z warunku transwersalności

$$\psi_{ao} e^T = X_a - k_a \hat{u}_{ao} \operatorname{sh} T, \tag{1.21}$$

wynika po podstawieniu (6.19) i rozwiązaniu względem $\hat{u}_{ao}$

$$\hat{u}_{ao} = \frac{k_a X_a - \frac{\pi_{ao}}{2} e^T}{k_a \operatorname{sh} T + e^T}. \tag{1.22}$$

Analogicznie, po rozwiązaniu drugiego problemu częściowego uzyskamy

$$\hat{u}_{bo} = \frac{k_b X_b - \frac{\pi_{bo}}{2} e^T}{k_b \operatorname{sh} T + e^T}. \tag{1.23}$$

Przejdźmy teraz do procedury koordynacji rozwiązań częściowych na poziomie nadrzędnym. Wspólna funkcja f_{ow} w całościowym wskaźniku jakości (6.2) ma postać

$$f_{ow} = u_a u_b. \tag{1.24}$$

Zastosujemy drugi wariant algorytmu koordynacji, wyrażający się wzorem (536); jak wiemy, nie musi on być zbieżny dla omawianego problemu, gdyż funkcja f_{ow} nie jest wypukła. Algo-

rytm koordynacji można zapisać w postaci

$$\pi_a^{(i+1)} = \hat{u}_b^{(i)}; \qquad \pi_b^{(i+1)} = \hat{u}_a^{(i)}. \tag{1.25}$$

Z postaci (6.18) optymalnych sterowań dla problemów częściowych wynika, że założenie (6.15) o postaci zmiennych koordynacyjnych π_a, π_b było uzasadnione i że w trakcie koordynacji będą się zmieniać tylko wartości początkowe $\pi_{ao}^{(i)}$, $\pi_{bo}^{(i)}$ zgodnie ze wzorem

$$\pi_{ao}^{(i+1)} = \hat{u}_{bo}^{(i)}; \qquad \pi_{bo}^{(i+1)} = \hat{u}_{ao}^{(i)}. \tag{1.26}$$

Nie jest to oczywiście własność ogólna omawianego algorytmu, a jedynie cecha szczególna rozpatrywanego problemu. Wykorzystajmy teraz wzory (6.22), (6.23). Obowiązuje oczywiście zależność

$$\pi_{bo}^{(i+3)} = \hat{u}_{ao}^{(i+2)} = \frac{k_a X_a - \frac{e^T}{2}\pi_{ao}^{(i+2)}}{k_a \mathrm{sh}T + e^T} = \frac{k_a X_a - \frac{e^T}{2}\hat{u}_{bo}^{(i+1)}}{k_a \mathrm{sh}T + e^T} = \tag{1.27}$$

$$= \frac{k_a X_a - \frac{e^T}{2}\dfrac{k_b X_b - \frac{e^T}{2}\pi_{bo}^{(i+1)}}{k_b \mathrm{sh}T + e^T}}{k_a \mathrm{sh}T + e^T} = \frac{k_a X_a - \frac{e^T}{2}\dfrac{k_b X_b - \frac{e^T}{2}\hat{u}_{ao}^{(i)}}{k_b \mathrm{sh}T + e^T}}{k_a \mathrm{sh}T + e^T},$$

z której wynika równanie różnicowe

$$\hat{u}_{ao}^{(i+2)} - \frac{e^{2T}}{4(k_a \mathrm{sh}T + e^T)(k_b \mathrm{sh}T + e^T)}\hat{u}_{ao}^{(i)} = \tag{1.28}$$

$$= \frac{k_a X_a (k_b \mathrm{sh}T + e^T) - k_b X_b \frac{e^T}{2}}{(k_a \mathrm{sh}T + e^T)(k_b \mathrm{sh}T + e^T)}.$$

Warunki i charakter zbieżności tego równania pozwalają nam ocenić zbieżność algorytmu koordynacji.

Równanie charakterystyczne

$$z^2 - \frac{e^{2T}}{4(k_a shT + e^T)(k_b shT + e^T)} = 0, \qquad (1.29)$$

ma dwa pierwiastki o jednym module.

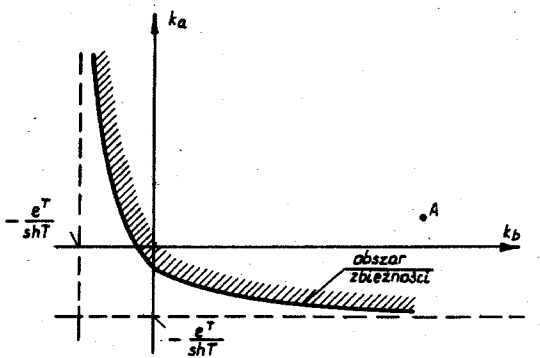
Warunek zbieżności

$$|z_{1,2}| = \frac{e^T}{2\sqrt{(k_a shT + e^T)(k_b shT + e^T)}} < 1, \qquad (1.30)$$

można przekształcić do postaci

$$\left(k_a + \frac{e^T}{shT}\right)\left(k_b + \frac{e^T}{shT}\right) > \frac{1}{4}\frac{e^{2T}}{sh^2 T}. \qquad (1.31)$$

Obszar zbieżności algorytmu na płaszczyźnie parametrów k_a, k_b przedstawia rys. 81.



Rys. 81

Jeśli algorytm jest zbieżny, to granica ciągu $\hat{u}_{ao}^{(i)}$ zgodnie z równaniem (6.28) ma postać identyczną z wyrażeniem (6.12), określającym rozwiązanie całościowe

$$\hat{u}_{ao}^{(\infty)} = \frac{k_a X_a (k_b shT + e^T) - k_b X_b \frac{e^T}{2}}{(k_a shT + e^T)(k_b shT + e^T) - \frac{e^{2T}}{4}}. \qquad (1.32)$$

Charakter zbieżności algorytmu dla przykładowych danych $k_a = 1$, $k_b = 10$, $X_a = 1$, $X_b = 1$, $T = 1$ (danym tym odpowiada punkt A na rysunku 81) oraz $\pi_{ao}^{(1)} = 0$, $\pi_{bo}^{(1)} = 0$ ilustruje następująca tabela

i	1	2	3	4	5
$\hat{u}_{ao}^{(i)}$	0,257	0,0159	0,0243	0,01640	0,01669
$\frac{\hat{u}_{ao}^{(i)} - \hat{u}_{ao}^{(\infty)}}{\hat{u}_{ao}^{(\infty)}}$	+1460%	-3,3%	+48%	-0,11%	+1,55%

Mimo, że wybrany punkt A nie jest bardzo odległy od granicy zbieżności algorytmu na płaszczyźnie parametrów k_a, k_b i mimo znacznego odchylenia przybliżenia początkowego od rozwiązania optymalnego, zbieżność algorytmu jest niezwykle szybka. (koniec przykładu 6).

Przykład powyższy pozwala sądzić, że algorytmy wielopoziomowe mogą być w wielu przypadkach bardzo szybko zbieżne, a tym samym mogą wymagać znacznie mniejszego nakładu obliczeń, niż algorytmy jednopoziomowe.

LITERATURA DO CZĘŚCI II

**[1] Athans M., Falb P.L.: Optimal Control. Mc. Graw Hill, 1966. Sterowanie optymalne. WNT, Warszawa 1969 (w druku).

[2] Bołtiański W.G.: Matiematiczeskije mietody optimalnowo uprawlenija. Nauka 1966.

[3] Bellman R.: Adaptacyjne procesy sterowania. PWN, Warszawa 1965.

[4] Findeisen W., Pułaczewski J., Manitius A.: Multilevel Optimization and Dynamic Coordination of Mass Flow in a Beet Sugar Plant. IV Congress of IFAC, Sess. 66, Warszawa 1969.

[5] Gosiewski A., Wierzbicki A.: Dynamic Optimization of Steel-Making Process in Electric Arc Furnace. IV Congress of IFAC, Sess. 39, Warszawa 1969.

[6] Kelley H.J. Methods of gradients. In G. Leitmann: Optimization techniques. Academic Press, N.Y. 1962.

[7] Kapp. R.E., Mc Gill N., Moyer H.G., Pinkham G.: Several trajectory optimization techniques. Im Balakrishan A.V., Neustadt L.W.: Computing methods in optimization problems. Academic Press, N.Y. 1964.

*[8] Kulikowski R.: Procesy optymalne i adaptacyjne. PWN, Warszawa 1965.

[9] Kulikowski R.: Decentralized Optimization of Large Scale Dynamic Systems. IV Congress of IFAC, Sess. 28. Warszawa 1969.

[10] Kurman K.: Chain Models as Inertialess Optimal Controlles of Multidimensional Processes. IV Congress of IFAC, Sess. 62, Warszawa 1969.

[11] Lasdon L.S., Mitter S.K., Warren A.D.: The conjugate gradient method for optimal control problems. IEEE Trans. on Automatic Control, V-AC12, 1967, No 2.

[12] Makowski K., Majerczyk-Gómułka I.: Wyznaczanie optymalnego sterowania procesami dynamicznymi metodą funkcjonałów Lagrange'a. Archiwum Automatyki i Telemechaniki, t. 13, z. 3, 1968.

*[13] Marcus L., Lee E.B.: Foundations of optimal control theory. J. Wiley, New York, 1967.

[14] Mesarovic M.D., Pearson J.D., Macko D., Takahara Y.: On the synthesis of dynamic multi-level systems. Proc. of the III Congress of IFAC, London 1966.

*[15] Pontriagin L.S., Bołtiański W.G., Gamkrelidze R.W., Miszczenko J.F.: Matiematiczeskaja tieorija optimalnych processow. Fizmatgiz, Moskwa 1961.

*[16] Wierzbicki A.: Zasada maksimum a synteza regulatorów optymalnych. Część I: Warianty i modyfikacje zasady maksimum. Część II: Algorytmy regulatorów optymalnych. Część III: Wrażliwość i struktury regulatorów optymalnych. Część IV: Przykład zastosowania. Archiwum Automatyki i Telemechaniki, t. 13, z. 1, 3, 1968; t. 14, z. 1, 3, 1969.

[17] Wierzbicki A.: Prinzip maksimum dla processow z nietriwielnom zapazdywanijem uprawlienija. W druku. Awtomatika i Telemechanika, 1969.

[18] Wierzbicki A.: Unified approach to the sensitivity analysis of optimal control systems. IV Congress of IFAC, Sess. 68, Warszawa 1969.

Dwiema gwiazdkami zaznaczono literaturę podstawową; jedną gwiazdką - pozycje zalecane dla studiów nad rozszerzeniem materiału.

Część III

OPTYMALIZACJA WIELOPOZIOMOWA

10. Sformułowanie zadania

Rozwiązywanie zadań optymalizacji dla obiektów dynamicznych o dużej wymiarowości lub zadań statycznych o dużej liczbie zmiennych decyzyjnych oraz związków ograniczających wymaga dużych nakładów obliczeniowych. Nakłady te mogą być w wielu przypadkach znacznie zmniejszone jeśli zadanie pierwotne poddamy dekompozycji, to jest podziałowi na kilka zadań częściowych, które byłyby z kolei koordynowane przez wyższy poziom sterowania.

Sposób podziału zadania jest oczywisty jeśli zadania częściowe mogą być sformułowane w taki sposób, że nie mają one zmiennych wspólnych. Oznaczałoby to jednak, że zadanie pierwotne jest w istocie zbiorem zadań nie sprzężonych ze sobą bezpośrednio, poza tym, że składają się one na wspólny wskaźnik jakości. Sytuacja taka nie zachodzi na ogół w praktyce; rozpatrywać będziemy zatem przypadek, gdy zadania częściowe nie są rozdzielne z natury, a ich separacja jest uzyskiwana przez wprowadzenie zmiennych koordynacyjnych. Zmiennymi koordynacyjnymi zajmuje się drugi, tj. wyższy poziom sterowania. Jego zadaniem jest zarówno uzyskanie ekstremum globalnego wskaźnika jakości, jak też zapewnienie dopasowania do siebie wzajemnie wszystkich zadań częściowych, zgodnie z ich istniejącymi w rzeczywistości powiązaniami. Jest oczywiste, że praktyczny wynik w postaci zmniejszenia nakładu obliczeń otrzymamy tylko wtedy, gdy zmiennych koordynacyjnych będzie mniej niż zmiennych decyzyjnych w zadaniu pierwotnym. Przypadek ten zachodzi wówczas, gdy w zadaniu pierwotnym nie wszystkie zmienne są powiązane ze sobą przez każdą z zależności ograniczających. W przypadku, gdy np. ograniczenia liniowe zadania optymalizacji statycznej wyrażone są macierzą (por. 547) oznacza to, że część elementów w tej macierzy jest zerami.

Podstawową strukturą jest, jak wynika z powyższego omówienia, struktura dwupoziomowa. Łącząc ze sobą czyli koordynując

kilka układów dwupoziomowych dodajemy trzeci poziom sterowania i tak dalej. Sens tego postępowania polegać musi jednak zawsze na oszczędności nakładu obliczeniowego.

Rozpatrujemy, dla koncentracji uwagi, zadanie optymalizacji statycznej

$$\max Q(\underline{u}), \tag{537}$$

przy ograniczeniu

$$\underline{u} \in U. \tag{538}$$

Aczkolwiek duże praktyczne znaczenie mają suboptymalne rozwiązania zadania (537), (538), zajmować się będziemy dalej tylko wielopoziomowymi rozwiązaniami ściśle optymalnymi, gdyż tylko takie rozważanie można przeprowadzić całkowicie jednoznacznie.

Wprowadzamy zbiór zmiennych koordynacyjnych $\underline{v}$ i dokonujemy podziału zadania (537), (538) tak, że ma ono postać

$$\max_{\underline{u} \in U} Q(\underline{u}) = \max_{\substack{\underline{u} \in U \\ \underline{v} \in V}} Q(\underline{u}, \underline{v}) = \max_{\substack{\underline{u}^1 \in U^1(\underline{v}) \\ \underline{u}^2 \in U^2(\underline{v}) \\ \underline{u}^N \in U^N(\underline{v}) \\ \underline{v} \in V}} \left[Q_1(\underline{u}^1, \underline{v}),\ Q_2(\underline{u}^2, \underline{v}), \ldots Q_N(\underline{u}^N, \underline{v}) \right] \tag{539}$$

gdzie $\underline{u}^i$ są pewnymi częściami zbioru $\underline{u}$.

Związek (539) wyraża, że przedstawiamy Q jako pewną funkcję N czynników $Q_1, Q_2, \ldots Q_N$, gdzie Q_i zależy od $\underline{u}^i$ oraz (być może) $\underline{v}$, a także, że zmieniamy sformułowanie zależności ograniczających w taki sposób, by powstało N oddzielnych grup ograniczeń na $\underline{u}^1, \underline{u}^2, \ldots$, zależnych od wektora $\underline{v}$ oraz pewne ograniczenie na sam wektor $\underline{v}$.

Rozwiązanie dwupoziomowe zadania (537) z użyciem dekompozycji wyrażonej w (539) jest możliwe, jeżeli $\underline{u}^i$, $i = 1, \ldots N$ są rozłącznymi podzbiorami $\underline{u}$ oraz jeżeli postać (539) pozwala na wykonanie rozłącznej ekstremalizacji względem każdego z $\underline{u}^i$:

$$\max_{\underline{u} \in U} Q(\underline{u}) = \max_{\underline{v} \in V} \left[\max_{\underline{u}^1 \in U^1(\underline{v})} Q_1(\underline{u}^1, \underline{v}),\ \max_{\underline{u}^2 \in U^2(\underline{v})} Q_2(\underline{u}^2, \underline{v}), \ldots \max_{\underline{u}^N \in U^N(\underline{v})} Q_N(\underline{u}^N, \underline{v}) \right]. \tag{540}$$

Po oznaczeniu

$$\max_{\underline{u}^i \in U^i(\underline{v})} Q_i(\underline{u}^i, \underline{v}) = \hat{Q}_i(\underline{v}) \tag{541}$$

związek (540) przyjmie postać

$$\max_{\underline{u} \in U} Q(\underline{u}) = \max_{\underline{v} \in V} \left[\hat{Q}_1(\underline{v}),\ \hat{Q}_2(\underline{v}), \ldots \hat{Q}_N(\underline{v}) \right]. \tag{542}$$

Wzór (541) wyraża zadanie pierwszego poziomu lub zadanie lokalne optymalizacji. Zadanie to jest parametryczne względem $\underline{v}$. Wzór (542) wyraża zadanie drugiego poziomu lub zadanie optymalizacji globalnej. Jest ono wykonywane względem zmiennych $\underline{v}$.

11. Sposoby dekompozycji

Związki (539) i (540) pokazują, że dla danego zadania (537), (538) może istnieć wiele sposobów dekompozycji, wszystkie dopuszczalne z punktu widzenia optymalizacji globalnej. Może być różna liczba utworzonych zadań częściowych, różna liczba wprowadzonych zmiennych koordynacyjnych (wymiar $\underline{v}$), różny sposób przeformułowania związków ograniczających. W zasadzie, najlepszy będzie ten sposób dekompozycji, przy którym nakład obliczeniowy łączny dla całej optymalizacji będzie najmniejszy. Ten jednak punkt widzenia jest trudny do ujęcia analitycznego; omówimy zatem tylko pokrótce niektóre cechy i zasady dokonywania dekompozycji. Rozpatrzymy oddzielnie dekompozycję wskaźnika jakości $Q(\underline{u})$ oraz zależności ograniczających $\underline{u} \in U$.

Wskaźnik jakości $Q(\underline{u})$ pozwala na rozłączną ekstremalizację względem $\underline{u}^i$ ($\underline{u}^i$ są rozłącznymi podzbiorami $\underline{u}$) jeżeli jest on jednego z następujących typów:

a) wyłącznie addytywny

$$Q(\underline{u}) = \sum_{i=1}^{N} Q_i(\underline{u}^i, \underline{v}), \tag{543}$$

b) wyłącznie multiplikatywny

$$Q(\underline{u}) = \prod_{i=1}^{N} Q_i(\underline{u}^i, \underline{v}), \tag{544}$$

z warunkiem, że $Q_i(\underline{u}^i, \underline{v}) \geqslant 0, \quad i = 1, 2, \ldots N,$

c) mieszamy, o rozłącznej części addytywnej i multiplikatywnej

$$Q(\underline{u}) = \prod_{i=1}^{l} Q_i(\underline{u}^i, \underline{v}) + \sum_{i=l+1}^{N} Q_i(\underline{u}^i, \underline{v}), \tag{545}$$

z warunkiem, że $Q_i(\underline{u}^i, \underline{v}) \geqslant 0, \quad i = 1, 2, \ldots l,$

d) mieszany, z nieantagonistycznym wpływem czynników wspólnych części, addytywnej i multiplikatywnej*):

$$Q(\underline{u}) = \prod_{i=1}^{l} Q_i(\underline{u}^i, \underline{v}) + \sum_{i=1}^{i<l} Q_i(\underline{u}^i, \underline{v}) + \sum_{i=l+1}^{N} Q_i(\underline{u}^i, \underline{v}), \qquad (546)$$

z warunkiem, że $Q_i(\underline{u}^i, \underline{v}) \geq 0, \quad i = 1, 2, \ldots l.$

Globalne ograniczenie $\underline{u} \in U$ jest, zgodnie z (539), przekształcone w zbiory ograniczeń $\underline{u}^i \in U^i(\underline{v})$, $\underline{v} \in V$. Dokonuje się tego przez wprowadzenie zmiennych koordynacyjnych $\underline{v}$. Trudno o generalne reguły; zazwyczaj mamy do czynienia z kompromisem między liczbą wprowadzonych zmiennych, a prostotą rozwiązywania utworzonych zadań. Rozpatrzymy pewne przypadki dla ilustracji.

Przyjmijmy, że nałożone na $\underline{u}$ ograniczenia są liniowymi równościami lub nierównościami

$$A\underline{u} \gtreqless \underline{b}, \qquad (547)$$

przy czym macierz A nie jest blokowo-diagonalna, lecz ma szereg elementów zerowych. Wyobraźmy sobie przenumerowanie zmiennych takie, by otrzymana macierz A składała się z bloków, powiązanych ze sobą przez niektóre kolumny:

$$\begin{bmatrix} a_{11} \cdots a_{1l} & & & \\ \cdots \quad \cdots & & & \\ a_{s1} \cdots a_{sl} & & & \\ & a_{(s+1)l} \cdots a_{(s+1)k} & & \\ & \cdots \quad \cdots & & \\ & a_{tl} \quad \cdots \quad a_{tk} & & \\ & & a_{(t+1)k} \cdots a_{(t+1)n} & \\ & & \cdots \quad \cdots & \\ & & a_{rk} \quad \cdots \quad a_{rn} & \end{bmatrix} \begin{bmatrix} u_1 \\ \cdots \\ \cdots \\ u_l \\ \cdots \\ \cdots \\ u_k \\ \cdots \\ \cdots \\ u_n \end{bmatrix} \gtreqless \begin{bmatrix} b_1 \\ \cdots \\ b_s \\ b_{s+1} \\ \cdots \\ b_t \\ b_{t+1} \\ \cdots \\ b_r \end{bmatrix} \qquad (548)$$

*) Zauważmy, że np. wskaźnik $Q(\underline{u}) = Q_1(\underline{u}^1, \underline{v})\, Q_2(\underline{u}^2, \underline{v})\, Q_3(\underline{u}^3, \underline{v}) - Q_2(\underline{u}^2, \underline{v})$ nie może być ekstremalizowany rozłącznie ze względu na $\underline{u}^2$, gdyż sposób w jaki $Q_2(\underline{u}^2, \underline{v})$ wpływa na całość poprzez część addytywną i multiplikatywną jest antagonistyczny.

Macierz w równaniu (548) można przekształcić w macierz blokowo-diagonalną przez usunięcie zmiennych pierwotnych u_l, u_k i wprowadzenie na ich miejsce zmiennych koordynacyjnych v'_l, v''_l, v'_k, v''_k oraz wprowadzenie nowych dwóch ograniczeń równościowych

$$v'_l - v''_l = 0,$$
$$v'_k - v''_k = 0. \tag{549}$$

Wynikiem jest nowy układ związków ograniczających, który zastąpi (548).

Macierz opisująca te ograniczenia jest teraz blokowo-diagonalna, kosztem wzrostu jej wymiaru o dwie kolumny i dwa wiersze:

$$\begin{bmatrix}
a_{11} \cdots a_{1l} & & & & \\
\cdots \quad \cdots & & & & \\
a_{s1} \cdots a_{sl} & & & & \\
 & a_{(s+1)l} \cdots a_{(s+1)k} & & & \\
 & \cdots \quad \cdots & & & \\
 & a_{tl} \quad \cdots a_{tk} & & & \\
 & & a_{(t+1)k} \cdots a_{(t+1)n} & & \\
 & & \cdots \quad \cdots & & \\
 & & a_{rk} \quad \cdots a_{rn} & & \\
1 & -1 & & & \\
 & 1 & -1 & &
\end{bmatrix}
\begin{bmatrix} u_1 \\ \cdots \\ u_{l-1} \\ v'_l \\ v''_l \\ \cdots \\ u_{k-1} \\ v'_k \\ v''_k \\ \cdots \\ u_n \end{bmatrix}
\overset{W}{=}
\begin{bmatrix} b_1 \\ \cdots \\ b_s \\ b_{s+1} \\ \cdots \\ b_t \\ b_{t+1} \\ \cdots \\ b_r \\ 0 \\ 0 \end{bmatrix} \tag{550}$$

Górna część macierzy występującej w (550) opisuje oddzielne zadania częściowe; co więcej, w tym szczególnym przypadku są one również rozdzielnie ograniczone. Na przykład, pierwsze zadanie częściowe ma ograniczenia

$$\begin{aligned}
&a_{11}u_1 + a_{12}u_2 + \cdots + a_{1(l-1)}u_{l-1} \overset{W}{=} b_1 - a_{1l}v'_l \\
&\cdots \\
&a_{s1}u_1 + a_{s2}u_2 + \cdots + a_{s(l-1)}u_{l-1} \overset{W}{=} b_1 - a_{sl}v'_l
\end{aligned} \tag{551}$$

lub w skróconym zapisie

$$\underline{u}^1 \in U^1(v'_l). \tag{551'}$$

Zadanie drugie ma ograniczenia

$$a_{(s+1)(l+1)}u_{l+1} + a_{(s+1)(l+2)}u_{l+2} + \ldots + a_{(s+1)(k-1)}u_{k-1} \gtreqless$$

$$\gtreqless b_{s+1} - a_{(s+1)l}v''_l - a_{(s+1)k}v'_k \tag{552}$$

...

$$a_{t(l+1)}u_{l+1} + a_{t(l+2)}u_{l+2} + \ldots + a_{t(k-1)}u_{k-1} \gtreqless b_t - a_{tl}v''_l - a_{tk}v'_k$$

lub w skróconym zapisie

$$u^2 \in U^2(v''_l, v'_k). \tag{552'}$$

Prawe strony (551), (552) zawierają różne zmienne koordynacyjne, co stanowi właśnie wspomnianą wyżej rozdzielność tych ograniczeń. Fakt ten znacznie ułatwia rozwiązywanie zadań tego rodzaju przy użyciu metod lagranżowskich, jak to omawiamy nieco dalej.

W przypadku nieliniowym ograniczenie $u \in U$ nie może być przedstawione zależnością typu (547). Tym niemniej można uzyskać dekompozycję zadania przez wprowadzenie zmiennych koordynacyjnych na miejsce niektórych zmiennych zadania pierwotnego.

Rozpatrzmy na przykład zadanie o ograniczeniach

$$\begin{aligned} a_{11}u_1u_2 + a_{12}u_2 + \sum_{i=3}^{k} a_{1i}u_i &\geqslant b_1, \\ a_{21}u_1u_2 + \sum_{i=k+1}^{n} a_{2i}u_i &\geqslant b_2. \end{aligned} \tag{553}$$

Usuwamy zmienną u_1 w obu ograniczeniach wprowadzając na to miejsce odpowiednio v'_1, v''_1 oraz dodatkowe wymaganie

$$v'_1 - v''_1 = 0. \tag{554}$$

Rezultatem jest

$$\begin{aligned} (a_{11}v'_1 + a_{12})u_2 + \sum_{i=3}^{k} a_{1i}u_i &\geqslant b_1, \\ a_{21}v''_1 u_2 + \sum_{i=k+1}^{n} a_{2i}u_i &\geqslant b_2. \end{aligned} \tag{555}$$

Ograniczenia są teraz liniowe względem u, lecz nadal sprzężone poprzez zmienną u_2. Połączenie to usuniemy wprowadzając v_2', v_2'' w miejsce u_2. W rezultacie otrzymujemy układ

$$\sum_{i=3}^{k} a_{1i}u_i \geqslant b_1 - (a_{11}v_1' + a_{12})v_2',$$
$$\sum_{i=k+1}^{n} a_{2i}u_i \geqslant b_2 - a_{21}v_1'' v_2'' \tag{556}$$

oraz ograniczenia dla drugiego poziomu

$$v_1' - v_1'' = 0,$$
$$v_2' - v_2'' = 0. \tag{557}$$

Zauważmy, że ograniczenia (556) są rozdzielone i liniowe, ponadto ich prawe strony zawierają różne zmienne koordynacyjne.

Separację ograniczeń (553) można również uzyskać wprowadzając mniejszą liczbę zmiennych, na przykład tylko jedną zmienną v dla zastąpienia iloczynu u_1u_2. Otrzymamy

$$\sum_{i=2}^{k} a_{1i}u_i \geqslant b_1 - a_{12}v, \tag{558}$$

$$\sum_{i=k+1}^{n} a_{2i}u_i \geqslant b_2 - a_{21}v. \tag{559}$$

Prawe strony (558) zawierają teraz tę samą zmienną koordynacyjną.

Dekompozycja ograniczeń $u \in U$ w obu rozpatrywanych przypadkach, liniowym i nieliniowym, opierała się na częściowym odseparowaniu zmiennych już w zadaniu pierwotnym, przez co droga podziału była częściowo wskazana. Sytuacja taka nie zawsze ma miejsce - na przykład może istnieć jedno ograniczenie o postaci

$$h(u_1, u_2, \ldots, u_n) \geqslant b. \tag{560}$$

Dogodny sposób dekompozycji powstaje wówczas przez wprowadzenie zmiennych "agregowanych"

$$v_1 = g_1(u_1, u_2, \ldots, u_1),$$
$$v_2 = g_2(u_{1+1}, u_{1+2}, \ldots, u_k),$$
$$\ldots \tag{561}$$
$$v_N = g_N(u_{k+\alpha}, \ldots, u_n),$$

tak, żeby ograniczenie (249) przybrało postać

$$h(v_1, v_2, \ldots, v_N) \geqslant b. \qquad (560')$$

Ograniczenia równościowe (561) będą stosowane do zadań częściowych, ograniczenie (560') - do zadania drugiego poziomu.

Zawsze należy pamiętać o tym, że zgodnie z (539) dekompozycja $Q(u)$ oraz podział ograniczeń $\underline{u} \in U$ musi się opierać na tych samych rozłącznych podzbiorach $\underline{u}^i \in \underline{u}$. Należy zazwyczaj rozpoczynać próbę podziału od dekompozycji ograniczeń, a następnie sprawdzić czy proponowany podział $\underline{u}$ na $\underline{u}^i$ prowadzi do takiej postaci wskaźnika jakości

$$Q = Q\left[Q_1(\underline{u}^1, \underline{v}),\ Q_2(\underline{u}^2, \underline{v}), \ldots\ Q_N(\underline{u}^N, \underline{v})\right],$$

która pozwala na rozłączną ekstremalizację.

12. Podstawowe metody rozwiązywania

Wzór (541) wyrażał zadania optymalizacji pierwszego poziomu, z parametrycznym wynikiem $\hat{Q}_i(\underline{v})$. Optymalne decyzje dla pierwszego poziomu byłyby również parametryczne względem $\underline{v}$:

$$\hat{\underline{u}}^1(\underline{v}),\ \hat{\underline{u}}^2(\underline{v}), \ldots, \hat{\underline{u}}^N(\underline{v}). \qquad (562)$$

Wzór (542) wyrażał zadanie optymalizacji drugiego poziomu. Rozwiązaniem tego zadania będzie $\underline{v} = \hat{\underline{v}}$, które z kolei określi wartości $\hat{\underline{u}}^1, \hat{\underline{u}}^2, \ldots, \hat{\underline{u}}^N$ zgodnie z (562). Zadania zarówno pierwszego jak drugiego poziomu mogą być rozwiązywane jakąkolwiek odpowiednią dla nich metodą. Jeżeli zadanie pierwszego poziomu (541) oraz zadanie drugiego poziomu (542) są powiązane ze sobą wprost przez zmienną $\underline{v}$, mówimy że rozwiązanie wielopoziomowe jest "bezpośrednie". Przykłady takich rozwiązań podajemy dalej. Jeżeli natomiast, oprócz zmiennych $\underline{v}$, wprowadzamy mnożniki Lagrange'a i przy ich pomocy dokonujemy koordynacji rozwiązań pierwszego i drugiego poziomu, rozwiązanie jest lagranżowskie.

W dziedzinie programowania nieliniowego, Rosen i inni [8] opracowali metody iteracyjne pod nazwą "partition programming", przy pomocy których można sposobem bezpośrednim rozwiązywać zadania częściowo nieliniowe wypukłe o postaci

$$\min\left[Q(\underline{u}, \underline{v}) = \sum_{i=1}^{N} C_i^T(\underline{v})\underline{u}^i + \varphi(\underline{v})\right], \qquad (563)$$

przy ograniczeniach

$$A_i(\underline{v})\underline{u}^i \geqslant \underline{b}_i(\underline{v}), \qquad i = 1, 2, \ldots\ N \qquad (564)$$

Metody te mogą być przydatne w odpowiednich przypadkach.

Zauważmy, że wskaźnik jakości (563) jest liniowy względem $\underline{u}$ oraz nieliniowy względem $\underline{v}$. Ograniczenia (564) są liniowe względem $\underline{u}$, lecz zarówno macierz A jak wektor $\underline{b}$ mogą być funkcjami $\underline{v}$ (porównaj na przykład, zależności (558), (559)). Przyjmuje się, że macierz A jest blokowo diagonalna, tak że ograniczenia zadania są opisane przez układ nierówności (564), w które wchodzą macierze zadań częściowych A_i. Jak już wskazywaliśmy wiele zadań można podzielić i sprowadzić do postaci (564) przez wprowadzenie dodatkowych zmiennych i dodatkowych ograniczeń. Możemy jednak nie mieć wpływu na postać $Q(\underline{u})$ tak, by sprowadzić ten wskaźnik do postaci $Q(\underline{u}, \underline{v})$ wskazanej przez (563).

Metodę lagranżowską zamiast bezpośredniej możemy ze skutkiem zastosować w przypadku następującym. Przyjmijmy, że ograniczenie drugiego poziomu $\underline{v} \in V$ jest zbiorem ograniczeń równościowych liniowych względem elementów $v'_1, v''_1, v'_2, \ldots v''_{N-1}$ wektora koordynacyjnego $\underline{v}$:

$$
\begin{aligned}
& g_1(\underline{v}) = \alpha'_1 v'_1 + \alpha''_1 v''_1 + \alpha'_2 v'_2 + \ldots + \alpha''_{N-1} v''_{N-1} = 0, \\
& \ldots \\
& g_{N-1}(\underline{v}) = \beta'_1 v'_1 + \beta''_1 v''_1 + \beta'_2 v'_2 + \ldots + \beta''_{N-1} v''_{N-1} = 0 .
\end{aligned}
\tag{565}
$$

W szczególnym przypadku, ograniczenia drugiego poziomu mogą mieć po prostu postać (por. 549)

$$
\begin{aligned}
& g_1(\underline{v}) = v'_1 - v''_1 = 0, \\
& g_2(\underline{v}) = v'_2 - v''_2 = 0, \\
& \ldots \\
& g_{N-1}(\underline{v}) = v'_{N-1} - v''_{N-1} = 0.
\end{aligned}
\tag{566}
$$

Dla zadania drugiego poziomu, patrz (542), możemy napisać funkcję Lagrange'a i określić warunki konieczne rozwiązania $\hat{\underline{v}}$ jako warunki konieczne punktu siodłowego

$$
\min_{\underline{\lambda}} \max_{\underline{v}} L(\underline{v}, \lambda) = \min_{\underline{\lambda}} \max_{\underline{v}} \left\{ Q\left[\hat{Q}_1(\underline{v}), \hat{Q}_2(\underline{v}), \ldots \hat{Q}_N(\underline{v})\right] + \underline{\lambda}^T \underline{g}(\underline{v}) \right\} \tag{567}
$$

Załóżmy teraz, że wskaźnik jakości jest addytywny (por. 543) i zależny od rozłącznych części $\underline{v}$ (jeśli w ogóle zależy jawnie od $\underline{v}$):

$$
Q(\underline{u}) = Q_1(\underline{u}^1, v'_1) + Q_2(\underline{u}^2, v''_1, v'_2) + \ldots + Q_N(\underline{u}^N, v''_{N-1}) \tag{568}
$$

oraz że wszystkie ograniczenia w zadaniach częściowych są rozdzielne względem zmiennych koordynacyjnych (por. 551)

$$h_1(\underline{u}^1, \underline{v}'_1) \geq 0,$$

$$h_2(\underline{u}^2, v_1, v_2) \geq 0, \qquad (569)$$

$$\ldots$$

$$h_N(\underline{u}^N, v''_{N-1}) \geq 0.$$

Przyjęliśmy tu, dla uproszczenia rozważań, że każde z zadań częściowych ma jedno (skalarne) ograniczenie. Wobec (569) $\hat{Q}_1(\underline{v})$ staje się $\hat{Q}_1(\underline{v}'_1)$, $\hat{Q}_2(\underline{v})$ staje się $\hat{Q}_2(\underline{v}'_1, \underline{v}''_2)$ i tak dalej, zatem uwzględniając (566), (568) i (569) wyrażenie (567) przybiera postać szczególną

$$\min_{\underline{\lambda}} \max_{\underline{v}} L(\underline{v}, \underline{\lambda}) = \min_{\underline{\lambda}} \left\{ \max_{v'_1} \left[\hat{Q}_1(v'_1) + \lambda_1 v'_1 \right] + \right.$$
$$\left. + \max_{v''_1, v'_2} \left[\hat{Q}_2(v''_1, v'_2) - \lambda_1 v''_1 + \lambda_2 v'_2 \right] + \ldots + \max_{v''_{N-1}} \left[\hat{Q}_N(v''_{N-1}) - \lambda_{N-1} v''_{N-1} \right] \right\} \qquad (570)$$

Wykorzystaliśmy tu zarówno addytywność i rozłączność (568) względem elementów $\underline{v}$, jak charakter zależności (565) wzgl. (566). Postać (570) wskazuje, że maksymalizacja wyrazów w nawiasach kwadratowych względem elementów $\underline{v}$ może być przesunięta do zadań częściowych, które będą miały na przykład postać

$$\max_{v'_1} \left[\hat{Q}_1(v'_1) + \lambda_1 v'_1 \right] = \max_{v'_1} \left[\max_{u^1 \in U^1(v'_1)} Q_1(\underline{u}^1, v'_1) + \lambda_1 v'_1 \right]. \qquad (571)$$

Rozwiązania zadań częściowych będą parametryczne względem $\underline{\lambda}$, a nie względem $\underline{v}$:

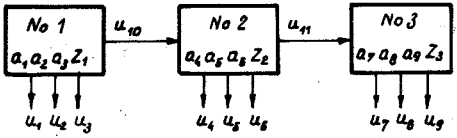
$$\hat{\underline{u}}^1(\lambda_1), \quad \hat{u}'_1(\lambda_1), \quad \hat{\underline{u}}^2(\lambda_1, \lambda_2) \text{ itd.} \qquad (572)$$

Zauważmy, że $\underline{\lambda}$ są to zmienne dodatkowe w zadaniu; zmienne koordynacyjne $\underline{v}$ muszą być użyte również, aby zapewnić potrzebną separację zadań częściowych. Korzyść z użycia $\underline{\lambda}$ polega na tym, że w pewnych przypadkach rozwiązania (572) zawierają mniej parametrów niż rozwiązania (562).

Przykłady

Rozpatrzmy zadanie programowania wypukłego, na przykładzie układu trzech zbiorników wodnych na rzece, rys. 82.

Każdy ze zbiorników zasila trzech odbiorców. Wypływy wody oznaczone są $u_1, u_2, \ldots u_9$. Zapas wody wynosi odpowiednio z_1, z_2, z_3. Zbiornik 1 może oddać ilość wody u_{10} do zbiornika 2,



Rys. 82

zbiornik 2 ilość wody u_{11} do zbiornika 3. Wskaźnik jakości, który należy zminimalizować ma postać

$$Q(\underline{u}) = \sum_{i=1}^{9} (a_i - u_i)^2, \tag{573}$$

a decyzje u_i są ograniczone przez nierówności, wynikające z bilansu wody

$$\begin{aligned} h_1(\underline{u}) &= u_1 + u_2 + u_3 + u_{10} - z_1 \leqslant 0, \\ h_2(\underline{u}) &= u_4 + u_5 + u_6 - u_{10} + u_{11} - z_2 \leqslant 0, \\ h_3(\underline{u}) &= u_7 + u_8 + u_9 - u_{11} - z_3 \leqslant 0 \end{aligned} \tag{574}$$

oraz przez wymaganie, by wszystkie przepływy były dodatnie

$$u_i \geqslant 0, \quad i = 1,2,\ldots 11. \tag{575}$$

Dekompozycja zadania (573), (574) będzie podyktowana raczej przez układ ograniczeń (574), bowiem (573) ma postać czysto addytywną i może być podzielone dowolnie. Jedną z możliwości jest wprowadzenie dwóch zmiennych koordynacyjnych, $v_1 = u_{10}$ oraz $v_2 = u_{11}$, tak by utworzyć trzy zadania częściowe, z ograniczeniami następującymi

$$\begin{aligned} h_1(\underline{u}^1, v_1) &= u_1 + u_2 + u_3 + v_1 - z_1 \leqslant 0, \quad u_i \geqslant 0, \\ h_2(\underline{u}^2, v_1, v_2) &= u_4 + u_5 + u_6 - v_1 + v_2 - z_2, \quad 0 \leqslant u_i, \\ h_3(\underline{u}^3, v_2) &= u_7 + u_8 + u_9 - v_2 - z_3 \leqslant 0, \quad u_i \geqslant 0 \end{aligned} \tag{576}$$

Wskaźnik jakości (573) będzie wówczas podzielony jak następuje

$$Q(u) = Q_1(\underline{u}^1) + Q_2(\underline{u}^2) + Q_3(\underline{u}^3), \tag{577}$$

gdzie

$$\begin{aligned} Q_1(\underline{u}^1) &= \sum_{i=1}^{3} (a_i - u_i)^2, \\ Q_2(\underline{u}^2) &= \sum_{i=4}^{6} (a_i - u_i)^2, \\ Q_3(\underline{u}^3) &= \sum_{i=7}^{9} (a_i - u_i)^2 . \end{aligned} \tag{578}$$

Po rozwiązaniu zadań częściowych, zadanie drugiego poziomu będzie

$$\min_{v_1, v_2} \left[\hat{Q}_1(v_1) + \hat{Q}_2(v_1, v_2) + \hat{Q}_3(v_2) \right]. \tag{579}$$

Ażeby otrzymać $\hat{Q}_1(v_1)$, trzeba rozwiązać zadanie częściowo pierwsze. Weźmy w tym celu wskaźnik jakości $Q_1(\underline{u}^1, v_1)$ oraz pierwszą nierówność z (576) i utwórzmy funkcję Lagrange'a

$$\begin{aligned} L(u_1, u_2, u_3, \lambda) = (a_1 - u_1)^2 + (a_2 - u_2)^2 + (a_3 - u_3)^2 + \\ + \lambda(u_1 + u_2 + u_3 + v_1 - z_1). \end{aligned} \tag{580}$$

Poszukiwanie warunków punktu siodłowego funkcji Lagrange'a daje wyrażenia

$$\begin{aligned} \hat{u}_1 &= a_1 - \Delta a = f_1(v_1), \\ \hat{u}_2 &= a_2 - \Delta a = f_2(v_1), \\ \hat{u}_3 &= a_3 - \Delta a = f_3(v_1), \\ \hat{\lambda} &= 2\,\Delta a, \end{aligned} \tag{581}$$

gdzie

$$\Delta a = \frac{1}{3}(a_1 + a_2 + a_3 + v_1 - z_1), \tag{582}$$

przy czym rozwiązania (581) spełniają warunki konieczne (i dostateczne) Kuhna-Tuckera w obszarze

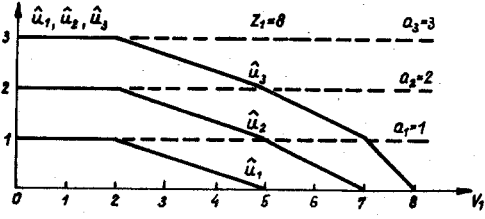
$$0 \leq \hat{u}_1 \leq a_1,$$
$$0 \leq \hat{u}_2 \leq a_2,$$
$$0 \leq \hat{u}_3 \leq a_3, \tag{583}$$
$$z_1 - v_1 \leq a_1 + a_2 + a_3 = \alpha_1$$

lub, inaczej to określając, w obszarze

$$z_1 - \alpha_1 \leq v_1 \leq z_1 - \alpha_1 + 3\,a_1, \tag{584}$$

gdzie a_1 jest najmniejszym spośród a_i tego zadania częściowego.

Wykres rozwiązań (581) podaje rys. 83.



Rys. 83

Zgodnie z (583) rozwiązania (581) są ważne w obszarze $2 \leq v_1 \leq 5$ dla przykładowych wartości a_i, z_1 przyjętych dla rys. 83. Dla $v_1 < 2$ optymalne wartości u_i minimalizujące Q_1 są to oczywiście $u_i = a_i$, a dla $v_1 > 5$ trzeba rozwiązać nowe zadanie optymalizacji

$$\min\left[Q_1(\underline{u}^1) = (a_2 - u_2)^2 + (a_3 - u_3)^2\right] \tag{585}$$

z ograniczeniem $h_1(\underline{u}^1, v_1)$ tym co poprzednio (576) oraz z wartością $\hat{u}_1 = 0$. Rozwiązanie jest łatwe i będzie obowiązywać w przedziale $5 < v_1 \leq 7$. Przy $v_1 = 7$ osiągamy $\hat{u}_2 = 0$, zatem dla $v_1 > 7$ zadanie optymalizacji brzmi

$$\min\left[Q_1(\underline{u}^1) = (a_3 - u_3)^2\right], \tag{586}$$

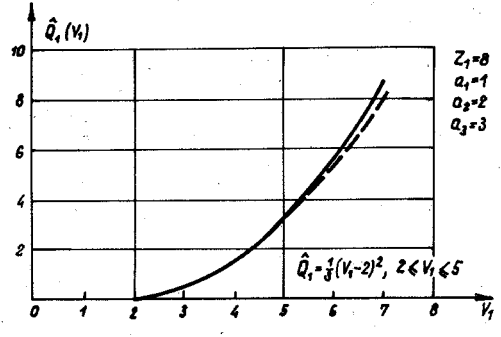
przy ciągle tym samym ograniczeniu $h_1(\underline{u}^1, v_1)$ według (576) oraz z wartościami $\hat{u}_1 = 0$, $\hat{u}_2 = 0$. Wszystkie wyniki $\hat{u}_i(v_1)$ są wykreślone na rys. 83.

Musimy teraz obliczyć $\hat{Q}_1(v_1)$. Wykorzystując (581) dochodzimy po paru przekształceniach do

$$\hat{Q}_1(v_1) = \frac{1}{3}(\alpha_1 + v_1 - z_1)^2, \tag{587}$$

przy czym wzór ten obowiązuje w obszarze (584).

Biorąc poza tym obszarem rozwiązania $\hat{u}_i$ wskazane na rys. 83, możemy przedstawić $\hat{Q}_1(v_1)$, dla rozpatrywanego przykładu, jak pokazano na rys. 84.



Rys. 84

Zauważmy, że $\hat{Q}_1$ jest zerem dla $v_1 < z_1 - \alpha_1$ oraz jest wyrażeniem kwadratowym (587) w przedziale określonym przez (584). Gdybyśmy chcieli uprościć nieco zadanie drugiego poziomu, można by przyjąć (587) jako przybliżenie dla wszystkich $v_1 \geq z_1 - \alpha_1$, otrzymując w ten sposób

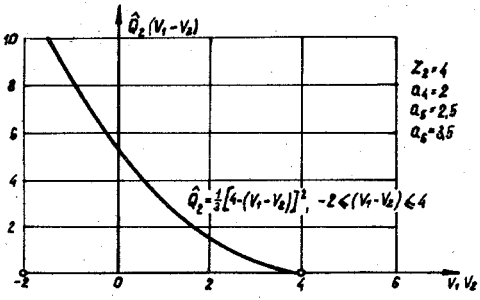
$$\hat{Q}_1(v_1) = \begin{cases} \frac{1}{3}(\alpha_1 + v_1 - z_1)^2 & \text{jeśli } v_1 \geq z_1 - \alpha_1 \\ 0 & \text{w obszarze pozostałym} \end{cases} \tag{588}$$

oraz podobnie dla pozostałych dwóch zadań częściowych:

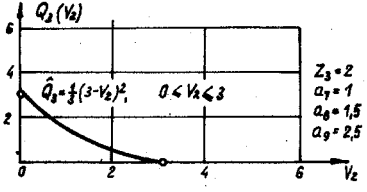
$$\hat{Q}_2(v_1, v_2) = \begin{cases} \frac{1}{3}(\alpha_1 - v_1 + v_2 - z_2)^2 & \text{jeśli } v_1 - v_2 \leq \alpha_2 - z_2 \\ 0 & \text{w obszarze pozostałym} \end{cases} \tag{589}$$

$$\hat{Q}_3(v_2) = \begin{cases} \frac{1}{3}(\alpha_3 - v_2 - z_3)^2 & \text{jeśli } v_2 \leq \alpha_3 - z_3 \\ 0 & \text{w obszarze pozostałym} \end{cases} \tag{590}$$

Odpowiednie wykresy dla wybranych danych liczbowych przedstawiono na rys. 85 i rys. 86.



Rys. 85



Rys. 86

Funkcja drugiego poziomu polega na wykonaniu

$$\min_{v_1, v_2} \left[\hat{Q}_1(v_1) + \hat{Q}_2(v_1, v_2) + \hat{Q}_3(v_2) \right], \tag{591}$$

z ograniczeniami pochodzącymi z bilansów

$$\begin{aligned} v_1 - z_1 &\leq 0, \\ v_2 - v_1 - z_2 &\leq 0 \end{aligned} \tag{592}$$

oraz warunkami na nieujemność

$$v_1 \geq 0, \quad v_2 \geq 0. \tag{593}$$

Zauważmy, że ograniczenia (592) można wyprowadzić wprost z problemu fizycznego lub też otrzymać z (596) przez przyjęcie dla $u_1, u_2, \ldots, u_9$ ich wartości krańcowych $u_i = 0$.

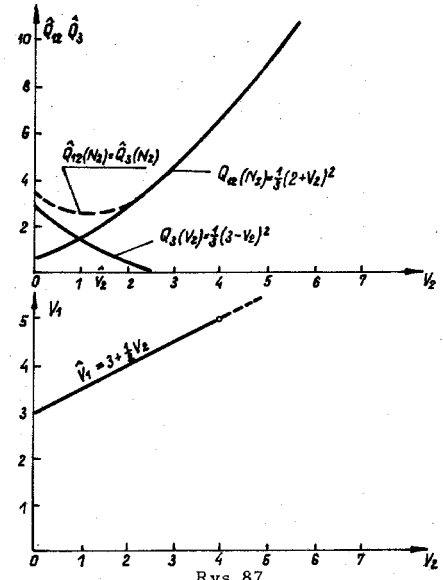
Zadanie (571) z warunkami (592, (593) może być rozwiązane jakąkolwiek odpowiednią metodą. Rozpatrzmy na przykład wprowadzenie trzeciego poziomu sterowania, tworząc na poziomie drugim zadanie parametryczne

$$\min_{v_1}\left[\hat{Q}_1(v_1) + \hat{Q}_2(v_1, v_2)\right] = \hat{Q}_{12}(v_2), \tag{594}$$

z ograniczeniem $0 \leq v_1 \leq z_1$ oraz pozostawiając dla poziomu trzeciego ostateczną optymalizację

$$\min_{\underline{u}} Q(\underline{u}) = \min_{v_2}\left[\hat{Q}_{12}(v_2) + \hat{Q}_3(v_2)\right], \tag{595}$$

z ograniczeniem $0 \leq v_2 \leq z_2 + z_1$.



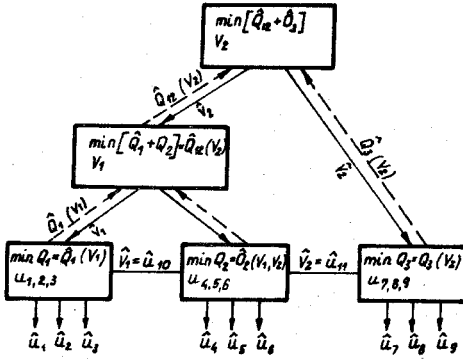
Rys. 87

Obliczenia wskazane przez (594) i (595) mogą być, w rozpatrywanym przykładzie, łatwo przeprowadzone. Wyniki są podane na rys. 87.

Dla decyzji na trzecim poziomie trzeba dodać $\hat{Q}_{12}(v_2)$ do $\hat{Q}_3(v_2)$ wziętego z rys. 86 i stwierdzić położenie minimum tej sumy. Wynik brzmi $v_2 = \frac{4}{3}$, a wskaźnik jakości $\hat{Q}(\underline{u}) = 2\frac{7}{9}$.

Dalej należy wziąć wartość $\hat{v}_2 = \frac{4}{3}$, użyć wykresu $\hat{v}_1(v_2)$ z rys. 87 aby stwierdzić, że $\hat{v}_1 = 3\frac{2}{3}$, zastosować wykres z rys. 83 dla określenia $\hat{u}_1 = \frac{1}{9}$, $\hat{u}_2 = 1\frac{4}{9}$, $\hat{u}_3 = 2\frac{4}{9}$ oraz postąpić podobnie

dla rozwiązań $\hat{u}_4 \div \hat{u}_9$ (odnośnych wykresów nie pokazujemy). Szkic struktury rozwiązania podano na rys. 88.



Rys. 88

Jako drugi przykład rozpatrzmy zadanie, w którym wskaźnik jakości ma postać iloczynową.

Dane jest zadanie maksymalizacji

$$\max\left[Q(\underline{u}) = Q_1(u_1, u_2, u_3) \cdot Q_2(u_3, u_4, u_5)\right], \tag{596}$$

gdzie

$$Q_1(u_1, u_2, u_3) = \alpha^2 - (a_1 - u_1)^2 - (a_2 - u_2)^2 - (a_3 - u_3)^2 \geq 0$$

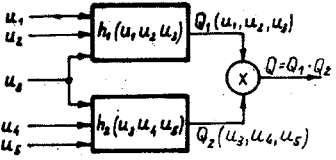
$$Q_2(u_3, u_4, u_5) = \beta^2 - (a_3 - u_3)^2 - (a_4 - u_4)^2 - (a_5 - u_5)^2 \geq 0 \tag{597}$$

z ograniczeniami $u_i \geq 0$ oraz

$$h_1(u_1, u_2, u_3) = z_1 - u_1 - u_2 - u_3 \geq 0 \tag{598}$$

$$h_2(u_3, u_4, u_5) = z_2 - u_3 - u_4 - u_5 \geq 0 \tag{599}$$

Zadanie to może być ilustrowane schematem z rys. 89. Wprowadzając nową zmienną $v = u_3$ w funkcje Q_1, h_1 oraz w funkcje Q_2, h_2 możemy utworzyć dwa zadania częściowe.



Rys. 89

Zadanie częściowe 1

$$\max_{u_1, u_2} Q_1(u_1, u_2, v) = \max_{u_1, u_2} \left[\alpha^2 - (a_1 - u_1)^2 - (a_2 - u_2)^2 - (a_3 - v)^2\right], \quad (600)$$

przy ograniczeniu

$$h_1(u_1, u_2, v) = z_1 - u_1 - u_2 - v \geq 0, \quad (601)$$

z rozwiązaniami parametrycznymi

$$\hat{Q}_1(v), \ \hat{u}_1(v), \ \hat{u}_2(v).$$

Zadanie częściowe 2

$$\max Q_2(u_4, u_5, v) = \max \left[\beta^2 - (a_4 - u_4)^2 - (a_5 - u_5)^2 - (a_3 - v)^2\right] \quad (602)$$

przy ograniczeniu

$$h_2(u_4, u_5, v) = z_2 - u_4 - u_5 - v \geq 0, \quad (603)$$

z rozwiązaniami parametrycznymi

$$\hat{Q}_2(v), \ \hat{u}_4(v), \ \hat{u}_5(v).$$

Rozwiązania zadań częściowych wchodzą w zadanie drugiego poziomu

$$\max \hat{Q}_1(v) \cdot \hat{Q}_2(v) \quad (604)$$

z ograniczeniem $v \geq 0$.

Rozwiązaniem zadania drugiego poziomu będzie $\hat{v}$.

Zadanie częściowe 1 rozwiążemy metodą Kuhna-Tuckera; tworzymy funkcję Lagrange'a

$$L(u_1, u_2, \lambda) = \alpha^2 - (a_1 - u_1)^2 - (a_2 - u_2)^2 - (a_3 - v)^2 + \lambda(z_1 - u_1 - u_2 - v) \quad (605)$$

i stwierdzamy np., że wartości

$$\hat{u}_1 = \frac{1}{2}(z_1 - v + a_1 - a_2),$$
$$\hat{u}_2 = \frac{1}{2}(z_1 - v - a_1 + a_2), \quad (606)$$
$$\hat{\lambda} = a_1 + a_2 - z_1 + v,$$

spełniają warunki Kuhna-Tuckera, czyli są rozwiązaniami zadania w obszarze

$$0 \leq u_1 \leq a_1, \qquad (607)$$

$$0 \leq u_2 \leq a_2,$$

albo, gdy przyjąć, że $a_2 \geq a_1$, w obszarze:

$$z_1 - a_1 - a_2 \leq v \leq z_1 + a_1 - a_2. \qquad (608)$$

W tak określonym obszarze $\hat{Q}_1(v)$ można obliczyć jako

$$\hat{Q}_1(v) = \alpha^2 - (a_1 - \hat{u}_1)^2 - (a_2 - \hat{u}_2)^2 - (a_3 - v)^2, \qquad (609)$$

co daje

$$\hat{Q}_1(v) = \alpha^2 - \frac{1}{2}(a_1 + a_2 - z_1 + v)^2 - (a_3 - v)^2. \qquad (610)$$

Nie rozpatrzyliśmy, dla oszczędności miejsca, rozwiązań właściwych dla obszaru poza określonym przez (608). Zauważmy tylko, że gdy $v \leq z_1 - a_1 - a_2$ rozwiązania optymalne będą $\hat{u}_1 = a_1$, $\hat{u}_2 = a_2$; gdy natomiast $v \geq z_1 + a_1 - a_2$ to $\hat{u}_1 = 0$, a gdy ponadto $v \geq z_1 - a_1 + a_2$ to także $\hat{u}_2 = 0$.

Dla drugiego zadania częściowego utworzymy funkcję Lagrange'a

$$L(u_4, u_5, \mu) = \beta^2 - (a_4 - u_4)^2 - (a_5 - u_5)^2 - (a_3 - v)^2 + \\ + \mu(z_2 - u_4 - u_5 - v), \qquad (609)$$

która ma punkt spełniający warunki Kuhna-Tuckera

$$\hat{u}_4 = \frac{1}{2}(z_2 - v + a_4 - a_5),$$

$$\hat{u}_5 = \frac{1}{2}(z_2 - v - a_4 + a_5), \qquad (610)$$

$$\hat{\mu} = a_4 + a_5 - z_2 + v,$$

w obszarze określonym przez

$$0 \leq u_4 \leq a_4, \qquad (611)$$

$$0 \leq u_5 \leq a_5,$$

albo inaczej, gdy $a_5 \geq a_4$, w obszarze

$$z_2 - a_4 - a_5 \leq v \leq z_2 + a_4 - a_5 . \qquad (612)$$

Wynik optymalizacji brzmi w tym obszarze

$$\hat{Q}_2(v) = \beta^2 - \frac{1}{2}(a_4 + a_5 - z_2 + v)^2 - (a_3 - v)^2. \qquad (613)$$

Ażeby z kolei rozwiązać zadanie drugiego poziomu

$$\max \hat{Q}_1(v) \cdot \hat{Q}_2(v),$$

z ograniczeniem znaku v zastosować trzeba warunki Kuhna-Tuckera; przy danych liczbowych następujących

$$a_1 = a_4 = 2, \qquad z_1 = 8, \qquad \alpha^2 = 100,$$

$$a_2 = a_5 = 4, \qquad z_2 = 8, \qquad \beta^2 = 100,$$

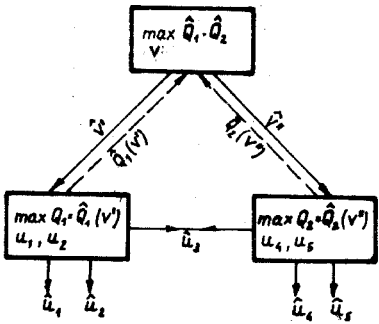
$$a_3 = 3,$$

rozwiązanie brzmi

$$\hat{v} = \hat{u}_3 = \frac{8}{3}$$

oraz w ślad za tym, użytkując (606), (610) otrzymujemy

$$\hat{u}_1 = \frac{5}{3}, \quad \hat{u}_2 = \frac{11}{3}, \quad \hat{u}_4 = \frac{5}{3}, \quad \hat{u}_5 = \frac{11}{3}.$$



Rys. 90

Dwupoziomowa struktura rozwiązywania zadania, jaką tu zastosowaliśmy, przedstawiona jest schematycznie na rys. 90.

Rozwiązanie tego przykładu musiałoby przebiegać inaczej, gdyby istniało trzecie ograniczenie obejmujące wszystkie zmienne decyzyjne, na przykład

$$h_3(u_1, u_2, u_3, u_4, u_5) = z_3 - u_1 - u_2 - u_3 - u_4 - u_5 \geqslant 0 \qquad (614)$$

Trzeba by teraz wprowadzić trzy zmienne koordynacyjne, a mianowicie

$$v_1 = u_1 + u_2, \quad v_2 = u_3, \quad v_3 = u_4 + u_5 \qquad (615)$$

otrzymując jako rozwiązania zadań dolnego poziomu odpowiednie wyrażenia

$$\hat{Q}_1[v_1, v_2], \quad \hat{Q}_2[v_2, v_3] \qquad (616)$$

i formułując następujące zadanie poziomu górnego:

$$\max \hat{Q}_1[v_1, v_2] \cdot \hat{Q}_2[v_2, v_3] \qquad (617)$$

z ograniczeniem $v_i \geqslant 0$ oraz

$$h_3(v_1, v_2, v_3) = z_3 - v_1 - v_2 - v_3 \geqslant 0. \qquad (618)$$

Zadanie (617), (618) można by rozwiązywać przy pomocy warunków Kuhna-Tuckera, utworzywszy funkcję Lagrange'a

$$L(v_1, v_2, v_3, \gamma) = \hat{Q}_1[v_1, v_2] \cdot \hat{Q}_2[v_1, v_2] + \gamma(z_3 - v_1 - v_2 - v_3) \qquad (619)$$

lub też odpowiednią inną metodą. Będzie to jednak bardziej skomplikowane niż w przykładzie przytoczonym poprzednio i korzyści z dekompozycji byłyby niewielkie. Zauważmy, że oprócz 5 zmiennych oryginalnych zadania mamy aż 3 zmienne koordynacyjne.

13. Metody obliczeniowe

Przedstawione w poprzednim punkcie przykłady rozwiązane były na drodze analitycznej. W zadaniach o wymiarowości i stopniu skomplikowania odpowiadających praktyce, rozwiązywanie analityczne nie jest możliwe i sięgać trzeba do metod numerycznych. Stosowane tu metody iteracyjne będą zbliżone do metod szukania ekstremum, stosowanych w optymalizacji statycznej. Nie będziemy rozwijać w niniejszym tekście zagadnień dynamicznych, aczkolwiek

rozwiązywanie ich numeryczne w układach wielopoziomowych jest również możliwe.

Rozpatrzmy zadanie dwupoziomowe, zapisane w ogólnej postaci (patrz (540)):

$$\max_{\underline{v}\in V} Q\left[\max_{\underline{u}^1\in U^1(\underline{v})} Q_1(\underline{u}^1, \underline{v}),\ \max_{\underline{u}^2\in U^2(\underline{v})} Q_2(\underline{u}^2, \underline{v}),\ \ldots\ \max_{\underline{u}^N\in U^N(\underline{v})} Q_N(\underline{u}^N, \underline{v})\right]. \qquad (620)$$

Zadanie nadrzędne polega tu na maksymalizacji wyników zadań lokalnych, uzależnionych od zmiennych koordynacyjnych $\underline{v}$, co zapiszemy

$$\max_{\underline{v}\in V} Q\left[\hat{Q}_1(\underline{v}),\ \hat{Q}_2(\underline{v}),\ \ldots\ \hat{Q}_N(\underline{v})\right]. \qquad (621)$$

Metoda bezpośrednia rozwiązania zadania (601) polegać będzie na zastosowaniu procedury iteracyjnej, w zasadzie dowolnej spośród znanych metod gradientowych lub bezgradientowych, do wykonania maksymalizacji względem $\underline{v}$. W każdym kroku tej procedury trzeba oczywiście wywołań rozwiązanie każdego z zadań częściowych

$$\max_{\underline{u}^i\in U^i} Q_i(\underline{u}^i, \underline{v}), \qquad (622)$$

dla zadanej w danym kroku wartości zmiennych koordynacyjnych $\underline{v}$. Zadania częściowe (602) mogą być rozwiązywane różnymi metodami, a więc również przy pomocy procedur gradientowych lub bezgradientowych.

Rozwiązywanie zadania (621) metodą bezpośrednią jest stosunkowo pracochłonne, ze względu na nakładanie się na siebie procedur szukania ekstremum na dwóch poziomach. Dla powodzenia procedury numerycznej potrzebne jest, by funkcja Q była wklęsła (przy szukaniu minimum - wypukła) względem zmiennych $\underline{v}$. Metoda bezpośrednia nie stawia wymagania addytywności funkcji $Q(Q_1, Q_2, \ldots Q_N)$, która będzie potrzebna dla dwóch następnych metod, omawianych niżej.

Załóżmy, że funkcja $Q(Q_1, Q_2, \ldots Q_N)$ ma postać addytywną:

$$Q(\underline{v}) = \hat{Q}_1(\underline{v}) + \hat{Q}_2(\underline{v}) + \ldots + \hat{Q}_N(\underline{v}). \qquad (623)$$

Przyjmijmy, że na $\underline{v}$ nie są nałożone ograniczenia (jeśli ograniczenia takie będą, można je uwzględnić np. przez dodanie do (623) odpowiedniej funkcji kary). Rozpatrzmy rozwiązywanie zadania drugiego poziomu

$$\max_{\underline{v}} Q(\underline{v}), \qquad (624)$$

przy użyciu którejkolwiek z metod gradientowych; metoda taka opierać będzie postępowanie iteracyjne na wartościach składowych gradientu w poszczególnych krokach:

$$\frac{\partial Q}{\partial v_j} = \frac{\partial \hat{Q}_1(\underline{v}^k)}{\partial v_j} + \frac{\partial \hat{Q}_2(\underline{v}^k)}{\partial v_j} + \ldots + \frac{\partial \hat{Q}_N(\underline{v}^k)}{\partial v_j} . \qquad (625)$$

W wyrażeniu (625) $\frac{\partial \hat{Q}_i(\underline{v}^k)}{\partial v_j}$ oznacza wartość pochodnej rozwiązania i-tego zadania częściowego względem danej zmiennej koordynacyjnej v_j, przy czym mowa tu o rozwiązaniu i-tego zadania dla narzuconej wartości zmiennych koordynacyjnych $\hat{Q}_i(\underline{v}^k)$.

Jeżeli procedury rozwiązywania zadań częściowych

$$\max_{\underline{u}^i} Q_i(\underline{u}^i, \underline{v}^k) = \hat{Q}_i(\underline{v}^k), \qquad (626)$$

będą tego rodzaju, że dadzą w wyniku, oprócz wartości liczbowej $\hat{Q}_i(\underline{v}^k)$, także wartości liczbowe pochodnych $\frac{\partial \hat{Q}_i(\underline{v}^k)}{\partial v_j}$, to procedura gradientowego rozwiązania zadania nadrzędnego (604) będzie miała potrzebne dane liczbowe (625).

Oparcie rozwiązania zadania nadrzędnego na wartościach $\frac{\partial \hat{Q}_i(\underline{v}^k)}{\partial v_j}$ z rozwiązań dolnego poziomu nosi nazwę <u>metody doboru współrzędnych</u> - w nazwie tej "współrzędne" oznaczają zmienne v_j, dobierane wg procedury gradientowej w kolejnych krokach iteracji. W najprostszym wariancie metody gradientowej, zmienne v_j byłyby dobierane następująco

$$v_j^{k+1} = v_j^k + \varrho_j \sum_{i=1}^{N} \frac{\partial \hat{Q}_i(\underline{v}^k)}{\partial v_j} , \qquad (627)$$

gdzie ϱ_j - współczynnik, dodatni przy zadaniu na maksimum, a ujemny przy zadaniu na minimum.

Zwróćmy uwagę, że dla zadania częściowego (626) przy zarzuconej wartości $\underline{v} = \underline{v}^k$ można napisać funkcję Lagrange'a jak następuje

$$L(\underline{u}^i, \underline{v}, \underline{\lambda}_i) = Q_i(\underline{u}^i, \underline{v}) + \underline{\lambda}^{iT}(\underline{v}^k - \underline{v}). \qquad (628)$$

W punkcie $\hat{\underline{u}}^i$, $\hat{\underline{v}}$, $\hat{\underline{\lambda}}^i$, który jest rozwiązaniem zadania, zachodzi jak wiadomo zależność

$$\hat{\lambda}_j^i = \frac{\partial \hat{Q}_i}{\partial v_j^k}, \tag{629}$$

tzn. w punkcie optymalnym mnożniki Lagrange'a są równe pochodnym funkcji celu względem parametrów $\underline{v}^k$, występujących w ograniczeniach równościowych (por. uwagę na str.20). W związku z tym, algorytm (607) bywa zapisywany w postaci

$$v_j^{k+1} = v_j^k + \varrho_j \sum_{i=1}^{N} \hat{\lambda}_j^i . \tag{630}$$

Podobnie jak poprzednio, powodzenie procedury doboru współrzędnych $\underline{v}$ zależy od wklęsłości (wypukłości) funkcji $Q(\underline{v})$. Ze względu na addytywną postać $Q(Q_1, Q_2, \ldots Q_N)$, wklęsłość (wypukłość) wszystkich $Q_i(\underline{v})$ jest warunkiem wystarczającym wklęsłości $Q(\underline{v})$.

Trzecia metoda iteracyjnego rozwiązywania zadań dwzpoziomowych nosi nazwę metody doboru mnożników Lagrange'a.

Dla metody tej potrzebna jest addytywność funkcji celu oraz, dodatkowo, zadania częściowe muszą zależeć od rozłącznych podzbiorów zmiennych koordynacyjnych. Rozłączność tę uzyskamy, gdy w wyrażeniu (623) wprowadzimy nowe zmienne $\underline{v}'$, $\underline{v}''$... $\underline{v}^{(N)}$ oraz dodatkowe ograniczenie równościowe, obowiązujące dla optymalizacji nadrzędnej. Zadanie nadrzędne przybierze zatem postać

$$\max_{\underline{v}', \underline{v}'', \underline{v}^{(N)}} \left[\hat{Q}_1(\underline{v}') + \hat{Q}_2(\underline{v}'') + \ldots + \hat{Q}_N(\underline{v}^{(N)}) \right], \tag{631}$$

z warunkiem

$$\underline{v}' = \underline{v}'' = \underline{v}^{(N)}. \tag{632}$$

Warunek (632) zapisać można w postaci układu równań

$$\begin{aligned} \underline{v}' - \underline{v}'' &= \underline{0}, \\ \underline{v}'' - \underline{v}''' &= \underline{0}, \\ \underline{v}^{(N-1)} - \underline{v}^{(N)} &= 0 \end{aligned} \tag{633}$$

i wówczas dla zadania (631), (633) napiszemy funkcję Lagrange'a

$$L(\underline{v}', \underline{v}'', \ldots \underline{v}^{(N)}, \underline{\lambda}_1, \underline{\lambda}_2, \ldots \underline{\lambda}_{N-1}) = \hat{Q}_1(\underline{v}') + \hat{Q}_2(\underline{v}'') + \ldots +$$

$$+ \hat{Q}_N\left(\underline{v}^{(N)}\right) + \underline{\lambda}_1^T (\underline{v}' - \underline{v}'') + \underline{\lambda}_2^T (\underline{v}'' - \underline{v}''') + \ldots +$$

$$+ \underline{\lambda}_{N-1}^T \left(\underline{v}^{(N-1)} - \underline{v}^{(N)}\right). \qquad (634)$$

Funkcję Lagrange'a (614) można rozdzielić na części, zależne odpowiednio od zmiennych $\underline{v}'$, $\underline{v}''$, ..., zapisując następnie zadanie jej maksymalizacji względem tych zmiennych jako zadania rozłączne, związane wspólnymi parametrami $\underline{\lambda}_i$:

$$\max_{\underline{v}'} \left[\hat{Q}_1(\underline{v}') + \underline{\lambda}_1^T \underline{v}'\right],$$

$$\max_{\underline{v}''} \left[\hat{Q}_2(\underline{v}'') - \underline{\lambda}_1^T \underline{v}'' + \underline{\lambda}_2^T \underline{v}''\right], \qquad (635)$$

$$\ldots$$

Pamiętając, że $\hat{Q}_i$ jest rezultatem maksymalizacji względem $\underline{u}^i$, otrzymamy do rozwiązania numerycznego zadania lokalne

$$\max_{\underline{u}^1, \underline{v}'} \left[Q_1(\underline{u}^1, \underline{v}') + \underline{\lambda}_1^T \underline{v}'\right]$$

$$\max_{\underline{u}^2, \underline{v}''} \left[Q_2(\underline{u}^2, \underline{v}'') - \underline{\lambda}_1^T \underline{v}'' + \underline{\lambda}_2^T \underline{v}''\right], \qquad (636)$$

$$\ldots$$

W zadaniach lokalnych $\underline{\lambda}_i$ są parametrami, które muszą być narzucone przez procedurę rozwiązywania zadania nadrzędnego.

Zwróćmy uwagę, że niezależnie od typu procedury stosowanej do zadań (636), gradientowej czy bezgradientowej, wartości $\underline{v}', \underline{v}''$ będą tu zawsze otrzymywane pośród liczbowych wyników procedury.

Procedurę iteracyjną doboru $\underline{\lambda}_i$, czyli procedurę poziomu nadrzędnego, najdogodniej jest ukształtować jako procedurę gradientową; ponieważ - zgodnie z warunkami optymalizacji metodą Lagrange'a - mnożniki Lagrange'a muszą zapewnić patrz (634)

$$\frac{\partial L}{\partial \underline{\lambda}_1^T} = \underline{v}' - \underline{v}'' = 0,$$

$$\frac{\partial L}{\partial \underline{\lambda}_2^T} = \underline{v}' - \underline{v}'' = \underline{0}, \qquad (637)$$

$$\ldots$$

można je dobierać wg schematu

$$\underline{\lambda}_1^{k+1} = \underline{\lambda}_1^k + \underline{\varrho}_1^T \frac{\partial L}{\partial \underline{\lambda}_1^T} = \underline{\lambda}_1^k + \underline{\varrho}_1^T (\underline{v}' - \underline{v}'')^k , \tag{638}$$

$$\underline{\lambda}_2^{k+1} = \underline{\lambda}_2^k + \underline{\varrho}_2^T \frac{\partial L}{\partial \underline{\lambda}_2^T} = \underline{\lambda}_2^k + \underline{\varrho}_2^T (\underline{v}'' - \underline{v}''')^k ,$$

...

gdzie $\underline{\varrho}_i$ jest współczynnikiem kroku (ujemnym w zadaniach na maksimum).

Algorytm (638) można oczywiście zastąpić inną bardziej skuteczną spośród metod gradientowych. W każdym przypadku, potrzebne dla procedury wartości gradientów $(\underline{v}' - \underline{v}'')^k$, $(\underline{v}'' - \underline{v}''')^k$ są dane przez wyniki zadań (636).

Podobnie jak przy metodzie doboru współrzędnych, skuteczność procedury nadrzędnej będzie zapewniona gdy funkcje $\hat{Q}_i$ będą wklęsłe wzgl. wypukłe względem $\underline{v}$.

Zwróćmy jeszcze uwagę na wymiarowość zmiennych wektorowych $\underline{v}^i$, $\underline{\lambda}_i$, występujących w omawianych zadaniach. Jeśli każde zadanie lokalne $\hat{Q}_i(\underline{v})$ zależy istotnie od całego wektora $\underline{v}$, o wymiarowości np. m, to wymiar każdego z wektorowych mnożników $\underline{\lambda}_i$ wynosi też m. Procedurą (638) objętych jest wówczas znacznie więcej zmiennych, niż wymiar wektora $\underline{v}$. Jest to niekorzystne np. w porównaniu z metodą doboru współrzędnych, patrz (627), ale tam stawialiśmy więcej wymagań odnośnie zadań częściowych: mają one dostarczać wartości pochodnych $\frac{\partial \hat{Q}_i}{\partial v_j}$.

Wymiarowość procedury (638) zmaleje znacznie, gdy nie wszystkie zmienne koordynacyjne z wektora $\underline{v}$ będą wchodzić do każdego z zadań częściowych. Wymiary $\underline{\lambda}_i$ będą wówczas niższe niż wymiar $\underline{v}$.

Weźmy dla ilustracji przykład zadania, sformułowanego przez związki (573), (574), (575) i rys. 82. Dla zastosowania metody mnożników Lagrange'a wprowadzimy zmienne koordynacyjne

$$v_1', \; v_1'' \quad \text{zamiast} \quad u_{10}, \tag{639}$$

$$v_2', \; v_2'' \quad \text{zamiast} \quad u_{11}, \tag{640}$$

otrzymując zadania częściowe (por. (576), (577), (578))

1) $$\min \left[Q_1(\underline{\psi}^1) = \sum_{i=1}^{3} (a_i - u_i)^2 \right], \tag{641}$$

z ograniczeniem

$$h_1(\underline{u}^1,\ v'_1) = u_1 + u_2 + u_3 + v'_1 - z_1 \leqslant 0, \quad u_i \geqslant 0, \tag{642}$$

$$2) \qquad \min\left[Q_2(\underline{u}^2) = \sum_{i=4}^{6} (a_i - u_i)^2\right], \tag{643}$$

z ograniczeniem

$$h_2(\underline{u}^2,\ v''_1,\ v'_2) = u_4 + u_5 + u_6 - v''_1 + v'_2 - z_2 \leqslant 0, \qquad u_i \geqslant 0, \tag{644}$$

$$3) \qquad \min\left[Q_3(\underline{u}^3) = \sum_{i=7}^{9} (a_i - u_i)^2\right], \tag{645}$$

z ograniczeniem

$$h_3(\underline{u}^3,\ v''_2) = u_7 + u_8 + u_9 - v''_2 - z_3 \leqslant 0, \qquad u_i \geqslant 0. \tag{646}$$

Załóżmy, że zadania częściowe rozwiązaliśmy, otrzymując parametryczne wyniki

$$\hat{Q}_1(v'_1), \qquad \hat{Q}_2(v''_1,\ v'_2), \qquad \hat{Q}_3(v''_2). \tag{647}$$

Zadanie drugiego poziomu polega na minimalizacji sumy

$$\min\left[\hat{Q}_1(v'_1) + \hat{Q}_2(v''_1,\ v'_2) + \hat{Q}_3(v''_2)\right], \tag{648}$$

z uwzględnieniem ograniczeń następujących

$$\begin{aligned} v'_1 - v''_1 &= 0, \\ v'_2 - v''_2 &= 0. \end{aligned} \tag{649}$$

Napiszemy funkcję Lagrange'a dla zadania (648), (649)

$$L(v'_1,\ v''_1,\ v'_2,\ v''_2, \lambda_1, \lambda_2) =$$

$$= \hat{Q}_1(v'_1) + \hat{Q}_2(v''_1,\ v'_2) + \hat{Q}_3(v''_2) + \lambda_1(v'_1 - v''_1) + \lambda_2(v'_2 - v''_2) \tag{650}$$

oraz przegrupujemy jej wyrazy, zapisując zarazem że poszukujemy warunków punktu siodłowego

$$\max_{\underline{\lambda}} \min_{\underline{v}} \left[\hat{Q}_1(v'_1) + \lambda_1 v'_1 + \hat{Q}_2(v''_1,\ v'_2) - \lambda_1 v''_1 + \lambda_2 v'_2 + \hat{Q}_3(v''_2) - \lambda_2 v''_2\right]. \tag{651}$$

Wykorzystamy możliwość rozłącznej ekstremalizacji

$$\max_{\lambda_1,\lambda_2}\left\{\min_{v_1}\left[\hat{Q}_1(v'_1)+\lambda_1 v'_1\right]+\min_{v''_1,v''_2}\left[\hat{Q}_2(v''_1,\ v'_2)-\lambda_1 v''_1+\lambda_2 v''_2\right]+\right.$$
$$\left.+\min_{v''_2}\left[\hat{Q}_3(v''_2)-\lambda_2 v''_2\right]\right\} \qquad (6.52)$$

Teraz przypomnijmy, że np. $\hat{Q}_1(v'_1)$ oznaczało rezultat minimalizacji (641) z ograniczeniem (642) możemy zatem pierwszy wyraz w (652) wraz z odpowiednim zadaniem pierwszego poziomu zapisać jako jedną ekstremalizację:

$$\min_{v'_1}\left[\hat{Q}_1(v'_1)+\lambda_1 v'_1\right]=\min_{v'_1,u_1,u_2,u_3}\left[Q_1(u_1,\ u_2,\ u_3)+\lambda_1 v'_1\right], \qquad (653)$$

z ograniczeniem

$$h_1(u_1,\ u_2,\ u_3,\ v'_1)=u_1+u_2+u_3+v'_1-z_1\leqslant 0,\quad u_i\geqslant 0. \qquad (654)$$

Nie jest istotne, czy ograniczenie (654) uwzględnimy w operacji (653) metodą utworzenia funkcji Lagrange'a czy w inny sposób. Istotne jest natomiast, że w nowo powstałym zadaniu częściowym ekstremalizujemy zmodyfikowany wskaźnik jakości

$$Q_1(u_1,\ u_2,\ u_3)+\lambda_1 v'_1, \qquad (655)$$

w porównaniu z samym tylko $Q_1(u_1,\ u_2,\ u_3)$ powstałym z dekompozycji wskaźnika globalnego. Składnik $\lambda_1 v'_1$ stanowi "wartość zmiennej koordynacyjnej". Łatwo zauważyć, że gdy w zadaniu (653) $\lambda_1 v'_1$ jest dodatnie (koszt), to w zadaniu drugim wystąpi $(-\lambda_1 v''_1)$, co oznacza wartość przeciwną (zysk).

Metoda doboru mnożników polegać będzie na zakładaniu wartości mnożników $\lambda_1, \lambda_2, \ldots$ i na ulepszaniu ich w kolejnych iteracjach. Miarą niewłaściwości doboru mnożników są oczywiście niezgodności $v'_1 \neq v''_1$, $v'_2 \neq v''_2$, a algorytm (638) miałby tu konkretną postać

$$\lambda_1^{k+1}=\lambda_1^k+\varrho_1\,(v'_1-v''_1)^k,$$
$$\lambda_2^{k+1}=\lambda_2^k+\varrho_2\,(v'_2-v''_2)^k, \qquad (656)$$
$$\ldots$$

przy czym ϱ_1, ϱ_2 musiałyby być dodatnie (przykład jest zadaniem na minimum).

Rozwiązanie tego samego zadania metodą doboru współrzędnych miałoby przebieg następujący.

Zadania częściowe otrzymują narzucone z góry wartości v_1^k względnie v_2^k; zadanie pierwsze miałoby zatem funkcję Lagrange'a

$$L_1(u_1, u_2, u_3, v_1, \lambda'_1) = Q_1(u_1, u_2, u_3) + \lambda'_1(v_1^k - v_1), \tag{657}$$

a zadanie drugie - funkcję Lagrange'a

$$L_2(u_1, u_2, u_3, v_1, v_2, \lambda''_1, \lambda'_2) =$$

$$= Q_2(u_1, u_2, u_3) + \lambda''_1(v_1^k - v_1) + \lambda'_2(v_2^k - v_2). \tag{658}$$

Rozwiązania zadań (657), (658) muszą dostarczyć, dla celów zadania nadrzędnego, cztery wartości mnożników:

$$\begin{aligned}
\lambda'_1 &= \frac{\partial \hat{Q}_1}{\partial v_1^k}, \\
\lambda''_1 &= \frac{\partial \hat{Q}_2}{\partial v_1^k}, \\
\lambda'_2 &= \frac{\partial \hat{Q}_2}{\partial v_2^k}, \\
\lambda''_2 &= \frac{\partial \hat{Q}_3}{\partial v_2^k},
\end{aligned} \tag{659}$$

a wówczas zadanie nadrzędne może korzystać z iteracji doboru zmiennych v_1, v_2 por. wzory (627), (630):

$$\begin{aligned}
v_1^{k+1} &= v_1^k + \varrho_1(\lambda'_1 + \lambda''_1)^k, \\
v_2^{k+1} &= v_2^k + \varrho_1(\lambda'_2 + \lambda''_2)^k.
\end{aligned} \tag{660}$$

Współczynniki ϱ_1, ϱ_2 muszą tu być ujemne, bowiem zadanie jest na minimum względem v_1, v_2.

LITERATURA DO CZĘŚCI III

[1] Boni P. Wielopoziomowe algorytmy optymalizacji statycznej. Praca magisterska, Instytut Automatyki Politechniki Warszawskiej 1970.

[2] Kulikowski R. Sterowanie w wielkich systemach. WNT, Warszawa 1970.

[3] Lasdon L. Duality and Decomposition in Mathematical Programming. IEEE Trans. on Systems Science and Cybernetics, Vol. SSC-4, July 1968.

[4] Łukasik S., Zieliński S. Porównanie metod optymalizacji statycznej procesu wielostopniowego z recyklem. Archiwum Automatyki i Telemechaniki, Zeszyt 3, 1969.

[5] Mańczak K. Agregacja w pewnych wielopoziomowych zagadnieniach kwadratowego programowania. Archiwum Automatyki i Telemechaniki, Zeszyt 4, 1966.

[6] Pearson J.D. Decomposition, coordination and multilevel systems. IEEE Trans. on Systems Science and Cybernetics, Vol. SSC-2, No 1, 1966.

[7] Pierwozwanskij A.A. Princip diecentralizacyi pri optimizacyi złożnych sistiem. Proceedings IV IFAC Congress, Warszawa 1969.

[8] Rosen J.B., Convex partition programming, w "Recent Advances in Mathematical Programming", R.L. Graven and P.Wolfe, Eds., New York 1963, Mc Graw-Hill.

[9] Zieliński S. O pewnym zastosowaniu teorii sterowania wielopoziomowego. Rozprawa doktorska, Politechnika Warszawska 1968.