



JACEK KORONACKI (Warszawa)

Aproksymacja stochastyczna II. Metody optymalizacji z ograniczeniami

(Praca przyjeta do druku 28.2.1976)

W tej częsci przedstawione zostana iteracyjne metody statycznej optymalizacji stochastycznej stosowane w przypadku, gdy na zbior argumentow funkcji optymalizowanej nałożone sa pewne ograniczenia.

Zgodnie z przyjetą w częsci I⁽¹⁾ terminologią, wymienione zadanie optymalizacji polega na wyznaczeniu punktu $\theta \in C$, $C \subset \mathbb{R}^N$, w którym funkcja rzeczywista f ($f: \mathbb{R}^N \rightarrow \mathbb{R}^1$) osiąga wartość minimalną na C . Zakładac będziemy przy tym, że zbior C ma postać

$$(1) \quad C = \{x \in \mathbb{R}^N: q^i(x) \leq 0, i = 1, 2, \dots, s\},$$

gdzie q^i są funkcjami o wartościach rzeczywistych, $q^i: \mathbb{R}^N \rightarrow \mathbb{R}^1$, $i = 1, \dots, s$.

Oczywiście, również w częsci II obowiazować będą założenia o nieznanosci postaci f i istnieniu zakłóceń losowych nakładających się na pomiary wartości tej funkcji. Ponadto, o f zakładac będziemy co najmniej, że gradient f_x oraz hesjan f_{xx} istnieją i są ciągle na $\mathbb{R}^N$ (por. część I — założenie (Z1); uwaga: wzmiankowane dalej wzory i założenia z częsci I oznaczane będą swoimi oryginalnymi numerami, poprzedzonymi rzymską cyfrą I — np. (I.Z1)). O losowych zakłóceniach nakładających się na wartości f przyjmować będziemy, że mają zerowe wartości oczekiwane i wariancje jednostajnie ograniczone na $\mathbb{R}^N$ (jak w (I.Z2) i (I.Z3)).

W zasadzie, podobnie jak w teorii programowania nieliniowego, a także jak w rozważanym w częsci I zadaniu optymalizacji stochastycznej bez ograniczeń, nie będziemy poszukiwać punktu, w którym f osiąga minimum globalne (na C), za rozwiązanie zaś uznamy punkt $\theta \in C$, spełniający tylko konieczny warunek optymalności. Wiadomo, że takim warunkiem koniecznym (nie najogólniejszym zresztą) jest warunek Kuhna-Tuckera, dający się zapisać następująco ([29], rozdział 1.2 i [36], rozdział 2.4):

$$(2) \quad f_x(\theta) + \sum_{i \in J(\theta)} \lambda_i q_x^i(\theta) = 0,$$

(¹) Str. 21–36 tego tomu.

gdzie $J(\theta)$ oznacza zbiór indeksów i odpowiadających ograniczeniom aktywnym w punkcie θ (tzn. takim, że $q^i(\theta) = 0$), λ_i są natomiast współczynnikami nieujemnymi.

U w a g a 1. W przypadku gdy C nie ma wnętrza, warunek (2) staje się trywialny — spełniony jest dla każdego $x \in C$.

U w a g a 2. Jeżeli funkcje f i q^i , $i = 1, 2, \dots, s$, są wypukłe, to (2) jest również wystarczającym warunkiem optymalności.

Metody optymalizacji przedstawione w rozdziałach II.1–3, są zbieżne z p. 1; dokładnie, każdy punkt skupienia x ciągu przybliżeń $\{X_n\}$, generowanego przez daną metodę, spełnia (2) z p. 1. Podstawę dowodów tej własności stanowi, jak poprzednio, nierówność (I.4).

W rozdziałach II.1 i II.2 przyjmiemy, że wartości q^i oraz q_x^i , $i = 1, \dots, s$, mogą być obliczone dokładnie dla każdego x . W rozdziale II.3 rozważymy natomiast przypadek, gdy zarówno wartości f , jak i wartości funkcji ograniczeń q^i (niekoniecznie wszystkich) obserwowane są z błędami losowymi.

II.1. Metody funkcji kary. Stochastyczny wariant metody (zewnątrznej) funkcji kary (por. [29], [36]) opracowany został przez Kushnera i Sanvicentego [25]. Autorzy ci za funkcję kary przyjęli:

$$(3) \quad p^n(x) = \frac{1}{\delta_n} \sum_{i=1}^s [\max\{0, q^i(x)\}]^2 = \frac{1}{\delta_n} \sum_{i \in J^+(x)} [q^i(x)]^2,$$

gdzie δ_n jest ciągiem liczb dodatnich $\delta_{n+1} < \delta_n$, $\lim_{n \rightarrow \infty} \delta_n = 0$, $J^+(x)$ oznacza zbiór tych i , dla których $q^i(x) \geq 0$, $x \in \mathbb{R}^N$.

Oznaczmy $h^n(x) = f(x) + p^n(x)$. Idea metody Kushnera–Sanvicentego (KS 1), zaczerpnięta z jej deterministycznego pierwowzoru, polega na zastosowaniu do funkcji h^n którejs z iteracyjnych metod optymalizacji stochastycznej bez ograniczeń, obcinanej (przerywanej) zgodnie z żadaną regułą zatrzymania. Powiedzmy, że reguła zatrzymania spowodowała zastopowanie procesu minimalizacji funkcji h^n w punkcie X_j^n ; punkt ten uznajemy wówczas za punkt startowy następnego cyklu, w którym cała procedura zostaje powtórzona z funkcją h^{n+1} zamiast h^n ($X_j^n = X_{n+1} = X_1^{n+1}$, przy czym — ogólnie — X_n oznacza punkt startowy cyklu n -tego). W ten sposób zadanie z ograniczeniami zastąpione zostaje ciągiem zadań bez ograniczeń, ze zmieniającą się funkcją minimalizowaną h^n . Intuicyjnie, zbieżność metody do rozwiązania wynika z tego, że dla każdego $x \notin C$: $p^n(x) > 0$, $p^{n+1}(x) > p^n(x)$ i $\lim_{n \rightarrow \infty} p^n(x) = \infty$, dla każdego zaś $x \in C$: $p^n(x) = 0$; stosowna przy tym organizacja procesu szukania punktu $\theta \in C$ zapewnia

$$(4) \quad \lim_n p^n(X_n) = 0 \quad \text{z p. 1.}$$

Zauważmy tu, że spełnienie związku (4), orzekającego dążenie ciągu przybliżeń X_n do zbioru punktów dopuszczalnych C , jest oczywiste i nie wymaga żadnych

dotatkowych założeń wtedy tylko, gdy w każdym cyklu optymalizacyjnym (tj. dla każdej h^n) wyznacza się z p. 1 punkt, w którym dana funkcja h^n przyjmuje wartość najmniejszą, czyli gdy $X_{n+1} = \arg \min_{x \in R^N} h^n(x)$, z p. 1. W przypadku stochastycznych metod funkcji kary ta ostatnia sytuacja nie może mieć miejsca, ze względu na konieczność obcinania stosowanej w każdym cyklu procedury minimalizacji bez ograniczeń.

Kushner i Sanvicente zaproponowali użycie w każdym cyklu optymalizacyjnym procedury KG (por. (I.16) i komentarz do tego schematu). W rezultacie, kolejny n -ty cykl metody KS1 sam dzieli się na podcykle, indeksowane literą m , w których stosuje się jednowymiarową — także obcinaną — procedurę Kiefera-Wolfowitza:

$$(5) \quad X_{i+1}^{n,m} = X_i^{n,m} - V_m^n a_i^{n,m} [Dh^n(X_i^{n,m}, c_i^{n,m}, V_m^n) + \hat{\mathcal{E}}_i^{n,m}],$$

gdzie

$$\hat{\mathcal{E}}_i^{n,m} = (2c_i^{n,m})^{-1} \mathcal{E}_i^{n,m},$$

$$Dh^n(X_i^{n,m}, c_i^{n,m}, V_m^n) = (2c_i^{n,m})^{-1} [f(X_i^{n,m} + c_i^{n,m} V_m^n) - f(X_i^{n,m} - c_i^{n,m} V_m^n)] + (V_m^n)^T p_x^n(X_i^{n,m}),$$

pozostałe oznaczenia odpowiadają przyjętym w (I.16) z tym, że odnoszą się do m -tego cyklu procedury KG, realizowanej w ramach n -tego cyklu metody KS1.

Zgodnie z tym co powiedziano wyżej, punkt, w którym zastopowany został n -ty cykl metody KS1, stanowi zarazem punkt startowy cyklu następnego ($X_k^{n+1} = X_{n+1} = X_1^{n+1,1}$)⁽²⁾. Odpowiednia reguła zatrzymania n -tego cyklu podana zostanie nieco dalej.

W założeniach metody KS1 uwzględnić należy oczywiście warunki zbieżności procedury KG. W szczególności przyjmuje się, że f jest ograniczona z dołu, funkcje f, q^i są dwukrotnie różniczkowalne oraz że hesjany f_{xx} i $g_{xx}^i, i = 1, 2, \dots, s, g^i(x) = [q^i(x)]^2$, por. (3)⁽³⁾, są ograniczone na R^N .

Pośród założeń związanych już bezpośrednio ze specyfiką zadania z ograniczeniami, najważniejsze to żądanie ograniczonego zbioru C , wypukłości f i ścisłej wypukłości $q^i, i = 1, \dots, s$. Dwa ostatnie z wymienionych wymagań nie występują wprawdzie explicite w [25], stanowią jednak „naturalny” warunek wystarczający dla spełnienia zawartego w tamtej pracy następującego założenia formalnego (por. [25] i tam rys. 3):

(Z1) Dla każdego zwartego zbioru S istnieje stała K_S taka, że jeżeli dla pewnych $x \in S$ i $\lambda_i \geq 0, i \in J^+(x)$, spełniony jest warunek

$$f_x(x) + \sum_{i \in J^+(x)} \lambda_i q_x^i(x) = 0,$$

to $\lambda_i \leq K_S$ dla wszystkich $i \in J^+(x)$.

⁽²⁾ X_k^{n+1} oznacza, że n -ty cykl zatrzymany został w k -tym punkcie podcyklu l -tego. (Zwraca się uwagę, że reguła zatrzymywania podcykli nie ma nic wspólnego z regułą zatrzymania cyklu).

⁽³⁾ W rzeczywistości w [25] założono tylko ograniczoność hesjanów q_{xx}^i . Przy takim jednak założeniu nie można by odwołać się do własności zbieżności procedury optymalizacji bez ograniczeń i, przeto, wykazać zbieżności aktualnie rozważanej metody KS1.

Podamy tu jeszcze jedno założenie, związane z szybkością malenia ciągu δ_n :

$$(Z2) \quad \sum_n \delta_n^2 (\delta_{n+1}^{-1} - \delta_n^{-1}) < \infty.$$

Przyjęcie (Z1) pozwala wykazać, że jeśli dla danego n działania procedury KG nie zatrzymać (tzn. $m = 1, 2, \dots, m \rightarrow \infty$), to — dla pewnego $K < \infty$, niezależnego od n, m, i — warunek

$$(6) \quad p^n(X_i^{n,m}) \leq K \delta_n$$

spełniony byłby nieskończenie wiele razy z p. 1.

Za regułę zatrzymania n -tego cyklu metody KS1 można tedy uznać pierwszy moment, dla którego zachodzi (6). Reguła taka jest dobrze określona, ponieważ liczba kroków iteracyjnych wykonywanych w n -tym cyklu jest z p. 1. skończona.

U w a g a. Ażeby n -ty cykl nie został zatrzymany w przypadku gdy wszystkie kolejne przybliżenia $X_i^{n,m}$ punktu minimalizującego h^n leżą w zbiorze C , należy założyć dodatkowo, że przed momentem zatrzymania wystąpiło przybliżenie nie należące do C . Pewnego komentarza wymaga tu jeszcze fakt, że stała K nie jest z góry znana. Stała ta może być wyznaczona „adaptacyjnie”, tzn. z cyklu na cykl może być zmieniana, na podstawie przebiegu procesu iteracyjnego w cyklach wcześniejszych (przy jednoczesnym założeniu, że każdy cykl jest zatrzymywany, jeżeli ciąg przybliżeń pozostaje „zbyt długo” poza zbiorem C).

Założenie (Z1), w połączeniu z (Z2) i przy podanej regule zatrzymania, zapewnia metodzie KS1 własność (4). (Z1) ingeruje także w ostatnią część dowodu zbieżności, pozwalającą stwierdzić dążenie ciągu $\{X_n\}$ do zbioru punktów spełniających (2). Nie wydaje się przy tym, aby możliwe było istotne osłabienie założenia (Z1). W konsekwencji, nie należy oczekiwać wykazania zbieżności stochastycznych metod funkcji kary bez żądania wypukłości f i q^i . Warto zarazem przypomnieć, iż deterministyczne metody funkcji kary, w których nie zakłada się osiągnięcia minimum h^n w każdym cyklu, również wymagają poczynienia założeń typu (Z1); por. [29], str. 141–144.

Fabian w pracy [5] (p. też [35], tłum. ros., str. 255) podał metodę nie zakładającą ograniczoności hesjanów g_{xx}^i , $i = 1, \dots, s$. Fabian założył tylko ograniczoność hesjanów q_{xx}^i . W jego metodzie dopuszcza się (przy każdym ustalonym n) nieograniczoną zmienność hesjanu funkcji kary $\tilde{p}^n$,

$$(7) \quad \tilde{p}^n(x) = r_n \sum_i \tilde{g}(M_n q^i(x)),$$

przy czym $\tilde{g}(x) = [\max\{0, x\}]e^{-1/x}$, $M_n \nearrow \infty$, $r_n \searrow 0$, $r_n M_n \nearrow \infty$.

Metoda Fabiana działa według schematu (I.2):

$$X_{n+1} = X_n - a_n Y_n,$$

przy czym Y_n jest sumą gradientu $\tilde{p}_x^n(X_n)$ i skończenie różnicowego estymatora gradientu $f_x(X_n)$ ⁽⁴⁾. Dla każdego n wykonuje się zaledwie jeden krok iteracyjny,

(4) Estymator f_x ma postać taką jak w metodzie KW — (I.1').

o długości $a_n \|Y_n\|$; z kroku na krok zatem ulega zmianie funkcja minimalizowana $\tilde{h}^n(x) = f(x) + \tilde{p}^n(x)$. Założenia metody Fabiana — poza dotyczącym ograniczoności tylko hesjanów q_{xx}^i — są podobne do przyjętych przez Kushnera i Sanvicentego (pewnej komplikacji ulegają wszakże warunki typu (I.Z5), wiążące współczynniki r_n, M_n, a_n i c_n).

W metodzie tej nie została przewidziana możliwość prowadzenia ucinanej, iteracyjnej minimalizacji każdej $\tilde{h}^n$. Trudność polega na tym, że dowodu zbieżności nie można tym razem oprzeć na znanych własnościach metod optymalizacji bez ograniczeń (dla danego n hesjan $\tilde{h}^n$ jest nieograniczony!).

Fabian wykazał zbieżność swojej metody i nadto podał oszacowanie asymptotycznej prędkości dążenia $f(X_n)$ do wartości minimalnej na C . Dla $a_n = n^{-5/6}$, $c_n = n^{-1/6}$, $M_n = n^{1/6}$ i $r_n = n^{-1/6}$, $\log n$, rząd prędkości zbieżności $f(X_n)$ wynosi $O(n^{-1/6+\Delta})$, gdzie Δ jest dowolną liczbą dodatnią.

Przedstawione metody funkcji kary wymagały założenia wypukłości — odpowiednio ściślej lub nie — funkcji f i q^i , $i = 1, \dots, s$ (czyniąc w ten sposób (2) koniecznym i wystarczającym warunkiem optymalności, por. uwaga 2). W metodzie KS1 trzeba było zażądać nadto ograniczoności g_{xx}^i . Z drugiej strony, w metodzie Fabiana założenie takie nie zostało wprowadzone, należy jednak oczekiwać, iż jest to metoda wolniejsza od KS1.

Istotną zaletą obydwu metod jest natomiast ich prostota. Wydaje się przy tym, że — przynajmniej w pierwszym etapie szukania punktu optymalnego — można zastosować metodę KS1, bez względu na to, czy hesjany g_{xx}^i są ograniczone na R^N czy nie. Nieograniczoność g_{xx}^i nie powinna mieć wpływu na działanie metody z dala od zbioru punktów dopuszczalnych; może jednak wywołać wyraźne (niezbieżne) oscylacje ciągu przybliżeń w otoczeniu tego zbioru i w takim przypadku metodę KS1 należy zastąpić inną, np. metodą Fabiana lub jedną z metod kierunków dopuszczalnych (por. [12], gdzie podano sposób mierzenia „oscylacyjności” ciągu przybliżeń).

II.2. Metody kierunków dopuszczalnych. Interesujące stochastyczne odpowiedniki metod kierunków dopuszczalnych (por. [29], [36]) opracowane zostały przez Kushnera [21], [24]. W metodach tych o zbiorze C zakłada się, iż jest zwarty i jest domknięciem swego wnętrza. Określające go funkcje q^i są różniczkowalne w sposób ciągły, o f zaś zakłada się, że ma ciągłe pochodne cząstkowe drugiego rzędu. Założenie ograniczoności f_{xx} na R^N nie jest potrzebne, gdyż w metodach kierunków dopuszczalnych wszystkie kolejne przybliżenia punktu optymalnego leżą w zbiorze C , jak przyjęliśmy, zwartym.

Niestety, Kushnerowi nie udało się udowodnić zbieżności jego metod bez założenia wypukłości f . Trzeba jednak podkreślić, że wymaganie to nie wynika ze specyfiki rozważanych metod i wiąże się raczej z trudnościami technicznymi napotkanymi przy szacowaniu nierówności typu (I.4). Wydaje się przeto, że metody kierunków

dopuszczalnych zachowują własność zbieżności do zbioru punktów optymalnych (spełniających (2)) również w przypadku f niewypukłych.

Niżej przedstawiona zostanie idea stochastycznego wariantu, metody Zoutendijka, a właściwie modyfikacji tej ostatniej, dokonanej przez Polaka ([29], str. 164; oryginalna metoda Zoutendijka, pod nazwą metody strefy bezpieczeństwa, opisana jest w [36], rozdział 13.4). W [24] rozważany jest także stochastyczny odpowiednik metody Topkisa i Veinotta ([29], str. 160).

Oznaczmy $q^0(x) = f(x)$ i określmy zbiór indeksów ograniczeń „ ε -aktywnych”

$$J_\varepsilon(x) = \{0\} \cup \{i: q^i(x) \geq -\varepsilon\}.$$

Dla każdego wektora $w \in \mathbf{R}^N$, $w = [w^{(1)} \dots w^{(N)}]^T$, $|w^{(i)}| \leq 1$, $i = 1, \dots, N$, zdefiniujemy

$$(8) \quad \max_{i \in J_\varepsilon(x)} \langle q_x^i(x), w \rangle \equiv \gamma(\varepsilon, x, w),$$

gdzie $\langle \cdot, \cdot \rangle$ oznacza iloczyn skalarny. Oznaczmy dalej

$$(9) \quad \min_w \gamma(\varepsilon, x, w) \equiv \gamma(\varepsilon, x).$$

Nie jest trudno pokazać, że $\gamma(0, x) \leq 0$ i warunek $\gamma(0, \theta) = 0$ jest równoważny koniecznemu warunkowi optymalności (2), [29].

Przez kierunek dopuszczalny rozumie się dowolny wektor w rozwiązujący problem minimaxowy (8)–(9). Wektor w wyznacza w pewnym otoczeniu $x \in C$ kierunek jednoczesnego malenia f i q^i , $i \in J_\varepsilon(x)$. Mówiąc nieścisłe, dzięki uwzględnieniu wszystkich ograniczeń ε -aktywnych, w jest kierunkiem malenia f , przebiegającym wewnątrz C .

Najogólniej, metoda Kushnera polega na wyznaczeniu, dla danego przybliżenia X_n , stochastycznego kierunku dopuszczalnego w_n i określeniu na tym kierunku — drogą ucinanej minimalizacji iteracyjnej — przybliżenia następnego, X_{n+1} . Zakłada się przy tym, że $X_n, X_{n+1} \in C$. Wektor w_n , nazywany dalej krótko kierunkiem dopuszczalnym, jest tak wybierany, aby spełniony był następujący warunek :

(Z3) Niech δ_1 i δ_2 będą funkcjami rzeczywistymi określonymi na $(0, \infty)$, niemającymi i dodatnimi. Niech n_1 będzie funkcją rzeczywistą na $(0, \infty)$, nierosnącą i dodatnią. Niech dalej P_{x_n} oznacza prawdopodobieństwo warunkowe względem X_n . Wówczas, dla każdego $\varepsilon > 0$,

$$P_{x_n} \{ \gamma(\varepsilon, X_n, w_n) \leq -\delta_1(\varepsilon) \} \geq \delta_2(\varepsilon)$$

z p. 1 na zbiorze tych ω , dla których $n \geq n_1(-\gamma(\varepsilon, X_n))$ i $\gamma(\varepsilon, X_n) \leq -\varepsilon$.

Założenie (Z3) wiąże w_n z definicją kierunku dopuszczalnego dla problemu deterministycznego (8)–(9). Warunek ten orzeka, że dla dostatecznie dużego n nierówność $\gamma(\varepsilon, X_n) \leq -\varepsilon$ implikuje $\gamma(\varepsilon, X_n, w_n) < 0$ z prawdopodobieństwem dodatnim, przy czym i to prawdopodobieństwo, i wartość $|\gamma(\varepsilon, X_n, w_n)|$ rosną wraz ze wzrostem ε . Nie jest to założenie ostre; można pokazać, że spełnione jest

np. przy w_n wybieranym losowo, zgodnie z rozkładem równomiernym na powierzchni kostki $\{w \in \mathbb{R}^N: |w^{(i)}| \leq 1, i = 1, \dots, N\}$.

W [24] zaproponowano określanie w_n w oparciu o rozwiązanie stochastycznego wariantu problemu minimaxowego (8)–(9). Dla prostoty założono, że obserwacji podlegają nieobciążone oceny wartości gradientu f_x , nie zaś tylko oceny wartości funkcji f . Niech dany będzie punkt X_n . Przyjmijmy, iż dane są także: rosnący ciąg liczb naturalnych M_j^n oraz liczba $\varepsilon_0 > 0$. Ciąg obserwacji wartości $f_x(X_n)$ ma postać $f_x(X_n) + \hat{\mathcal{E}}^{n,i}$, $i = 1, 2, \dots, E_{X_n}$, $E_{X_n} \hat{\mathcal{E}}^{n,i} = 0$; założmy ponadto, że zakłócenia losowe $\hat{\mathcal{E}}^{n,i}$, $i = 1, 2, \dots$, są ortogonalne (warunkowo, względem X_n). Niech wreszcie estymator wartości $f_x(X_n)$, oparty na M_r^n obserwacjach ma postać $f_x(X_n) + \psi^{n,r}$, gdzie $\psi^{n,r} = (M_r^n)^{-1} \sum_{i=1}^{M_r^n} \hat{\mathcal{E}}^{n,i}$. Przy podanych założeniach algorytm wyznaczania w_n można zapisać następująco:

K r o k 1. Połóż $k = 1$, $\varepsilon_k = \varepsilon_0$.

K r o k 2. Oblicz (na podstawie M_k^n obserwacji) $f_x(X_n) + \psi^{n,k}$ i rozwiąż poniższy stochastyczny wariant problemu (8)–(9):

$$(10) \quad \gamma_k(\varepsilon_k, X_n) = \min_{|w^{(i)}| \leq 1} \max \{ \langle f_x(X_n) + \psi^{n,k}, w \rangle, \max_{i \in J_{\varepsilon_k}^0(X_n)} \{ \langle q_x^i(X_n), w \rangle \} \},$$

gdzie $J_{\varepsilon_k}^0(x) = J_{\varepsilon_k}(x) - \{0\}$.

U w a g a. Zamiast (10) rozwiązuje się zwykle równoważne mu zadanie programowania liniowego, por. [21], [24] i [29], str. 161.

K r o k 3. Jeżeli $\gamma_k(\varepsilon_k, X_n) \leq -\varepsilon_k$, przyjmij za w_n dowolny wektor w rozwiązujący (10) i zatrzymaj działanie algorytmu. W przypadku przeciwnym połóż $\varepsilon_{k+1} = \varepsilon_k/2$, zwiększ k o 1 i wróć do kroku 2.

Ponieważ ciąg $\{M_j^n\}$ jest rosnący ($\lim M_j^n = \infty$) i zakłócenia $\hat{\mathcal{E}}^{n,i}$ są ortogonalne, $\lim_k \psi^{n,k} = 0$ z p. 1 i algorytm wyznaczania w_n musi się zatrzymać po skończonej liczbie kroków, jeśli tylko $\gamma(0, X_n) < 0$, tj. jeśli X_n nie spełnia koniecznego warunku optymalności.

Jeśliby zrezygnować z założenia dostępności nieobciążonych ocen $f_x(X_n)$, odpowiedni estymator tej wartości trzeba by oprzeć na obserwacjach (M_k^n -krotnych) skończonych różnic wartości f . Przypadek taki rozważony został w [20], str. 2–11, 2–12. Zaproponowany tam estymator gradientu $f_x(X_n)$, oparty na M_k^n obserwacjach różnic skończonych, może zostać zapisany w następujący sposób:

$$Y^{n,k} = \frac{1}{2d_n} \left\{ \begin{bmatrix} f(X_n + d_n e^1) - f(X_n - d_n e^1) \\ \vdots \\ f(X_n + d_n e^N) - f(X_n - d_n e^N) \end{bmatrix} + (M_k^n)^{-1} \sum_{i=1}^{M_k^n} \hat{\mathcal{E}}^{n,i} \right\},$$

gdzie e^j , $j = 1, \dots, N$, jest wersorem j -tej osi współrzędnych w $\mathbb{R}^N$, d_n — ustalony współczynnik, $d_n > 0$, $\lim_n d_n = 0$. Przy podanym estymatorze gradientu, wektor w_n — określany jak wyżej — zapewnia spełnienie (Z3). Tym razem jednak może się zdarzyć, iż dla danego d_n i przy $k \rightarrow \infty$ nie istnieje ujemne rozwiązanie problemu (10), mimo że $\gamma(0, X_n) < 0$. Mianowicie, wobec stosowania różnic skończonych, estymator $Y^{n,k}$ nie jest już nieobciążony i może się okazać, że $\lim_{k \rightarrow \infty} \gamma_k(\varepsilon_k, X_n) \geq$

≥ 0 . W praktyce zatem, działanie algorytmu wyznaczania w_n należy przerywać, gdy k osiąga wartość „zbyt dużą”; przyjmując wówczas należy $X_{n+1} = X_n$ i spróbować rozwiązać (10) przy mniejszym obciążeniu estymatora $Y^{n,k}$. Spełnienie przy tym warunku (Z3) wynika stąd, że podana tam nierówność musi zachodzić tylko dla „dostatecznie dużych n ” (przy n zaś dążącym do $+\infty$, znika obciążenie $Y^{n,k}$).

Mając X_n i w_n można przystąpić do minimalizacji f wzdłuż odcinka $[X_n, X_n + \lambda_n^+ w_n]$, gdzie $\lambda_n^+ = \min\{\lambda: X_n + \lambda w_n \notin C, \lambda \geq 0\}$, ucinanej zgodnie z zadaną regułą.

Dla dostatecznie dużych n i $\gamma(\varepsilon, X_n, w_n) \leq -\delta_1(\varepsilon)$ — por. (Z3) — procedura iteracyjna winna spełniać warunek analogiczny do (I.4). Stosowne procedury omówione zostały w [24], przy założeniu, że obserwuje się nieobciążone oceny wartości gradientu f_x ; przypadek dysponowania jedynie ocenami wartości funkcji f rozważony jest w [21]. Na przykład, wykorzystać można jednowymiarową procedurę typu KG — patrz (I.16) — tak jednak zmodyfikowaną, by wszystkie obserwacje wartości f (lub odpowiednio f_x) dokonywane były na odcinku $[X_n, X_n + \lambda_n^+ w_n]$ i punkt X_{n+1} , wyznaczony w ostatnim kroku iteracyjnym, także należał do tego odcinka.

Kushnerowi nie udało się wykazać zbieżności jego metod kierunków dopuszczalnych bez założenia, że liczba iteracji wykonywanych wzdłuż danego w_n rośnie wraz ze wzrostem n (tzn. $\lim_n \mathcal{L}_n = \infty$ z p. 1, gdzie $\mathcal{L}_n$ oznacza liczbę iteracji w n -tym cyklu metody). Nie jest to najprawdopodobniej założenie konieczne, w każdym zaś razie nie ma żadnego uzasadnienia praktycznego. Przeciwnie, wiadomo, że wykonywanie dużej liczby kroków iteracyjnych nie jest racjonalne w sytuacji, gdy X_n leży blisko brzegu C , w_n natomiast — skutkiem istnienia zakłóceń — określony został tak, że $\langle f_x(X_n), w_n \rangle > 0$ (por. [24], uwaga na str. 353 oraz przykłady)⁽⁵⁾.

Reasumując, metody kierunków dopuszczalnych nie wymagają zakładania wypukłości q^i , $i = 1, \dots, s$. Najprawdopodobniej metody te pozostają zbieżne i dla niewypukłych f . Stosują się przeto do szerszej klasy zadań optymalizacji niż metody funkcji kary. Opisana tutaj stochastyczna wersja metody Polaka jest natomiast o tyle bardziej skomplikowana od metod z poprzedniego rozdziału, że dla każdego n potrzebuje (zazwyczaj wielokrotnego) rozwiązania problemu (10).

Dla przedstawionej wersji metody Polaka nie zostały dotąd opracowane zadowalające sposoby doboru ciągów $\{M_n^i\}$ i $\{\mathcal{L}_n\}$. Pewne sugestie dotyczące tych zagadnień — oparte na doświadczeniu numerycznym — znaleźć można w [24].

2.3. Metoda funkcji Lagrange’a. Stochastyczny odpowiednik metody funkcji Lagrange’a ([36], rozdział 10.2), zaproponowany przez Kushnera i Sanvicentego [26], [27], stosuje się do przypadku, gdy zarówno wartości funkcji f , jak i wartości

⁽⁵⁾ W uwadze na str. 353 pracy [24] stwierdzono mylnie, że dyskutowane założenie użyte zostało w [20], w części dowodu twierdzenia 3, dotyczącej przypadku $\langle f_x(X_n), w_n \rangle > 0$. W rzeczywistości założenie to użyte zostało w przypadku funkcji f nierosnącej na $[X_n, X_n + \lambda_n^+ w_n]$ (por. [20], część trzecia dowodu twierdzenia 3) i nigdzie indziej w dowodzie twierdzenia 3 nie było wykorzystywane.

funkcji ograniczeń q^i (niekoniecznie wszystkich) obserwowane są z błędami losowymi. O ocenach $q^i(x)$ — podobnie jak $f(x)$ — zakłada się, że są nieobciążone, przy czym wariancja błędu losowego jest ograniczona na R^N , jednostajnie względem n (por. (I.Z4) i komentarz do tego założenia).

Idea metody funkcji Lagrange'a — jako związana z zasadą dualności (p. [36], rozdział 2.6) — jest szczególnie przydatna w rozważanym przypadku. Metoda kierunków dopuszczalnych nie może być zastosowana wobec niedokładnej znajomości ograniczeń, zastosowanie zaś metod funkcji kary w takiej sytuacji było wprawdzie badane, ale bez zachęcających rezultatów ([5], [34]; znane warianty metody funkcji kary wymagają wówczas, aby wariancje błędów losowych dążyły do zera, gdy $n \rightarrow \infty$).

Funkcja Lagrange'a ma postać

$$(11) \quad L(x, \lambda) = f(x) + \sum_{i=1}^s \lambda^{(i)} q^i(x),$$

gdzie $x \in R^N$, $\lambda = [\lambda^{(1)} \dots \lambda^{(s)}]^T$, $\lambda^{(i)} \geq 0$, $i = 1, \dots, s$.

Przy założeniach, że funkcje f i q^i są wypukłe i mają ciągłe pochodne pierwszego rzędu oraz że istnieje punkt $x \in C$, dla którego $q^i(x) < 0$, $i = 1, \dots, s$, warunkiem koniecznym i wystarczającym optymalności punktu θ jest istnienie punktu siodłowego funkcji Lagrange'a, tzn. istnienie punktu $(\theta, \hat{\lambda})$ o własności:

$$(12) \quad L(\theta, \lambda) \leq L(\theta, \hat{\lambda}) \leq L(x, \hat{\lambda}),$$

dla wszystkich $x \in R^N$, λ , $\lambda^{(i)} \geq 0$, $i = 1, \dots, s$; $\theta \in C$, $\hat{\lambda}^{(i)} \geq 0$, $i = 1, \dots, s$ ([36], rozdział 2.6).

Problem minimalizacji z ograniczeniami sprowadza się tedy do jednoczesnej minimalizacji $L(\cdot, \cdot)$ względem x i maksymalizacji tej funkcji względem λ , $\lambda^{(i)} \geq 0$. Przy tym

$$(13) \quad L_x(x, \lambda) = f_x(x) + \sum_{i=1}^s \lambda^{(i)} q_x^i(x);$$

$$(14) \quad L_\lambda(x, \lambda) = q(x),$$

gdzie $q(x) = [q^1(x) \dots q^s(x)]^T$.

W metodzie Kushnera i Sanvicentego (KS2) — ze względu na stochastyczny charakter zadania — czyni się więcej założeń niż to jest potrzebne dla stwierdzenia równoważności istnienia punktu siodłowego $(\theta, \hat{\lambda})$ i optymalności θ . Przyjmuje się mianowicie dodatkowo, że f i q^i , $i = 1, \dots, s$, mają ciągłe pochodne cząstkowe drugiego rzędu oraz punkt $(\theta, \hat{\lambda})$ leży w kostce (w $R^N \times R^s$) o znanych brzegach $|x^{(i)}| \leq A$, $0 \leq \lambda^{(i)} \leq A$, $A < \infty$ ⁽⁶⁾. Ten ostatni warunek pozwala dla ciągu przy-

⁽⁶⁾ W istocie Kushner i Sanvicente założyli jeszcze, że funkcja f jest ściśle wypukła. Założenie takie nie wydaje się konieczne, upraszcza za to dowód zbieżności; wówczas bowiem punkt θ spełniający (12) jest jedyny.

blużeń punktu optymalnego uzyskać oszacowanie typu (I.4) i, w konsekwencji, dowieść zbieżności metody; nie wydaje się zarazem, aby można go było zastąpić innym, rozszerzającym zakres zastosowań metody KS2.

Punkt $(\theta, \hat{\lambda})$ określa się iteracyjnie, stosując równolegle dwie procedury optymalizacji bez ograniczeń — jedną do minimalizacji L względem x i drugą do maksymalizacji L względem λ . Przy danym (osiągniętym) przybliżeniu (X_n, λ_n) pierwsza procedura wyznacza punkt $\tilde{X}_{n+1}$, druga zaś $\tilde{\lambda}_{n+1}$. Dokładniej, $\tilde{X}_{n+1}$ wyznaczany jest przez procedurę Kiefera–Wolfowitza (I.1) ze skończenie różnicowym estymatorem gradientu $L_x(X_n, \lambda_n)$ zamiast występującego tam estymatora wartości $f_x(X_n)$; w procedurze tej wykorzystuje się zatem skończenie różnicowe estymatory gradientów $f_x(X_n)$ i $q_x^i(X_n)$, $i = 1, \dots, s$, por. (13). Wyznaczenie $\tilde{\lambda}_{n+1}$ jest prostsze, dysponujemy bowiem nieobciążonym estymatorem $L_\lambda(X_n, \lambda_n)$, por. (14). W rezultacie, możemy zastosować wielowymiarową procedurę Robbinsa–Monro (I.3), ze zmienionym znakiem przed $\hat{a}_n$ (tym razem rozważane jest zadanie maksymalizacji, a nie — jak w tamtym przypadku — minimalizacji funkcji):

$$(15) \quad \tilde{\lambda}_{n+1} = \lambda_n + \hat{a}_n [L_\lambda(X_n, \lambda_n) + \hat{\epsilon}_n] = \lambda_n + \hat{a}_n [q(X_n) + \hat{\epsilon}_n].$$

Otrzymany punkt $(\tilde{X}_{n+1}, \tilde{\lambda}_{n+1}) \in \mathbb{R}^N \times \mathbb{R}^s$ jest z kolei rzutowany na kostkę, w której leży $(\theta, \hat{\lambda})$ i dopiero ów rzut uznawany jest za $(n+1)$ -sze przybliżenie rozwiązania:

$$X_{n+1}^{(i)} = \begin{cases} \tilde{X}_{n+1}^{(i)}, & \text{jeżeli } |\tilde{X}_{n+1}^{(i)}| \leq A, \\ A, & \text{jeżeli } \tilde{X}_{n+1}^{(i)} > A, \\ -A, & \text{jeżeli } \tilde{X}_{n+1}^{(i)} < -A, \end{cases}$$

$$\lambda_{n+1}^{(i)} = \begin{cases} \tilde{\lambda}_{n+1}^{(i)}, & \text{jeżeli } 0 \leq \tilde{\lambda}_{n+1}^{(i)} \leq A, \\ A, & \text{jeżeli } \tilde{\lambda}_{n+1}^{(i)} > A, \\ 0, & \text{jeżeli } \tilde{\lambda}_{n+1}^{(i)} < 0. \end{cases}$$

Metoda KS2 — przy założeniach podanych wyżej, a ponadto warunkach (I.Z5) i $a_n = \hat{a}_n$, por. (I.1) i (15) — daje $X_n \rightarrow \theta$, z p. 1. W pracach Kushnera i Sanvicen- tego nie był rozważany problem optymalizacji długości kroku przy przejściu od (X_n, λ_n) do $(\tilde{X}_{n+1}, \tilde{\lambda}_{n+1})$. Oczywiście, ze względu na minimaksowy charakter procesu szukania punktu siodłowego, nie jest tu możliwe skorzystanie wprost z zasad doboru długości kroku iteracyjnego, opisanych w rozdziale I.3.

II.4. Uwagi końcowe. Autorowi tego opracowania nie są znane żadne prace poświęcone badaniu skuteczności metod optymalizacji z ograniczeniami (jeśli nie liczyć wyników numerycznych przedstawionych w [24]). Największe znaczenie miałyby, przynajmniej eksperymentalne, porównanie lokalnych⁽⁷⁾ własności różnych metod optymalizacji (ze względu na podobieństwo problemów, niniejsze uwagi

(7) Tzn. nieasympotycznych, obowiązujących dla n ustalonych.

odnoszą się częściowo także do metod bez ograniczeń; por. komentarze w rozdziale I.4). Porównanie własności nieasymptotycznych jest szczególnie istotne w sytuacji, gdy liczba możliwych kroków iteracyjnych jest skończona, a zatem gdy sumaryczna liczba obserwacji wartości funkcji minimalizowanej jest ograniczona. Należy przy tym pamiętać, że wówczas skuteczniejsze mogą się okazać metody, w których kolejny krok w procesie szukania minimum określany jest na drodze wnioskowania statystycznego. Do metod takich należą przede wszystkim metody oparte na lokalnej analizie funkcji regresji; ich omówieniu poświęcona jest książka Zielińskiego [41] oraz w [15] — rozdział 7.4, 1, 2, 6. Na wnioskowaniu statystycznym opierają się także tzw. metody ewolucyjne ([15], rozdział 7.3.7) oraz metoda istotnej aproksymacji, opracowana przez Kacprzyńskiego ([13]–[15] — rozdział 7.3.2, 3).

Przedmiotem rozważań niniejszego opracowania było zadanie iteracyjnego wyznaczenia punktu, w którym funkcja regresji $f: C \rightarrow \mathbf{R}^1$, $C \subset \mathbf{R}^N$, osiąga minimum *lokalne*. Czytelnikowi zainteresowanemu określeniem *globalnego* minimum funkcji regresji polecić wypada przede wszystkim pracę Zielińskiego [40]. Praca ta ma charakter podstawowy, przedstawione w niej bowiem twierdzenia o zbieżności oraz metody szukania formułowane są przy możliwie najsłabszych założeniach. W szczególności, o funkcji minimalizowanej zakłada się tylko, że jest istotnie ograniczoną funkcją mierzalną, określoną na dowolnej przestrzeni z miarą $\{\mathfrak{X}, \mathfrak{B}, \mu\}$, $\mu(\mathfrak{X}) < \infty$. Udowodnione przez Zielińskiego twierdzenia ogólne mają postać twierdzeń o zbieżności według rozkładu i zbieżności z p. 1.

Ostatnio również, obiecującą metodę wyznaczania minimum globalnego ciągłej funkcji regresji, określonej na podzbiorze przestrzeni $\mathbf{R}^N$, zaproponował Bieluszko [2]. Dzięki przyjęciu dodatkowych jeszcze warunków o regularności funkcji i jej dziedziny metoda ta jest zbieżna według prawdopodobieństwa.

Literatura cytowana

- [1] S. N. Abdelhamid, *Transformation of observations in stochastic approximation*, Ann. Statist. 1 (1973), str. 1158–1174.
- [2] A. Bieluszko, *A method for the global optimization of stochastic systems* (w przygotowaniu).
- [3] A. Dvoretzky, *On stochastic approximation*, Proc. Third Berkeley Symp. Math. Statist. Prob. 1 (1956), str. 39–55.
- [4] V. Fabian, *Stochastic approximation methods*, Czech. Math. J. 10 (1960), str. 123–159.
- [5] — *Stochastic approximation of constrained minima*, Trans. 4-th Prague Conf. Information Th., Decision Functions, Random Processes, (1965), str. 277–290 (wyd. Academia, Praga 1967).
- [6] — *Stochastic approximation of minima with improved asymptotic speed*, Ann. Math. Statist. 38 (1967), str. 191–200.
- [7] — *On the choice of design in stochastic approximation methods*, Ann. Math. Statist. 39 (1968), str. 457–465.
- [8] — *On asymptotic normality in stochastic approximation*, ibidem. 39 (1968), str. 1327–1332.
- [9] — *Stochastic approximation, w Optimizing methods in statistics*, J. S. Rustagi, [ed.], Acad. Press, New York 1971.

- [10] C. C. Heyde, *On martingale limit theory and strong convergence results for stochastic approximation procedures*, Stoch. Processes and their Appl. 2 (1974), str. 359–370.
- [11] J. L. Hodges and E. L. Lehmann, *Two approximations to the Robbins–Monro process*, Proc. Third Berkeley Symp. Math. Statist. Probl. 1 (1956), str. 95–104.
- [12] K. Janáč, *Adaptive stochastic approximations*, Simulation 16, str. 51–58.
- [13] B. Kacprzyński, *O pewnej metodzie rozwiązywania równania regresji*, Arch. Autom. Telemech. 13.2 (1968).
- [14] — *Sekwencyjna estymacja ekstremum funkcji regresji metodą istotnej aproksymacji*, ibidem 13.3 (1968).
- [15] — *Planowanie eksperymentów: Podstawy matematyczne*, WNT, Warszawa 1974.
- [16] J. Koronacki, *Random-seeking methods for the stochastic unconstrained optimization*, Int. J. Control 21 (1975), str. 517–527.
- [17] — *Twierdzenia o zbieżności algorytmów statycznej optymalizacji stochastycznej*, Mat. Stos. 7 (1976), str. 81–89.
- [18] — Abdelhamid i Anbar: *Określenie optymalnej funkcji obserwacji w jednowymiarowych procedurach aproksymacji stochastycznej*, Mat. Stos. 7 (1976), str. 123–128.
- [19] H. J. Kushner, *Stochastic approximation algorithms for the local optimization of functions with non-unique stationary points*, IEEE Trans. Autom. Control, vol. AC-17, no. 5 (1972), str. 646–654.
- [20] — *Stochastic approximation type algorithms for the optimization of constrained and multimode stochastic problems*, CDS Techn. Rep. No. 72-1, Brown Univ. 1972.
- [21] — *Stochastic approximation algorithms for constrained optimization problems*, Ann. Statist. 2 (1974), str. 713–723.
- [22] — and T. Gavin, *Extensions of Kesten's adaptive stochastic approximation methods*, ibidem 1 (1973), str. 851–861.
- [23] — and T. Gavin, *A versatile method for the Monte-Carlo optimization of stochastic systems*, Int. J. Control 18 (1973), str. 963–975.
- [24] — and T. Gavin, *Stochastic approximation type methods for constrained systems: Algorithms and numerical results*, IEEE Trans. Autom. Control AC-19. 4 (1974), str. 349–357.
- [25] — and E. Sanvicente, *Penalty function methods for constrained stochastic approximation*, J. Math. Anal. Appl. 46 (1974), str. 499–512.
- [26] — and E. Sanvicente, *Stochastic approximation methods for constrained systems with observation noise on the systems and constraints*, Prepr. Stoch. Control Symp., str. 211–214, Budapest 1974.
- [27] — and E. Sanvicente, *Stochastic approximation of constrained systems with system and constraint noise*, Automatica 11.4 (1975), str. 375–380.
- [28] M. B. Newelson, R. Z. Chaśminskij, *Стохастическая аппроксимация и рекуррентное оценивание*, Izd. Nauka, Moskwa 1972.
- [29] E. Polak, *Computational methods in optimization: A unified approach*, Acad. Press, New York 1971.
- [30] B. T. Poljak, Ja. Z. Cypkin, *Псевдоградиентные алгоритмы адаптации и обучения*, Avtomatika i Telemekhanika 3 (1973), str. 45–68.
- [31] L. A. Rastrigin, *Стохастические методы поиска*, Izd. Nauka, Moskwa 1968.
- [32] — *Системы экстремального управления*, Izd. Nauka, Moskwa 1974.
- [33] P. Révész, *On the rate of convergence of Kesten's «Accelerated SA»* (w przygotowaniu).
- [34] E. Sanvicente, *Stochastic approximation methods for constrained systems*, Praca doktorska, Brown Univ., 1974.
- [35] M. T. Wasan, *Stochastic approximation*, Cambridge Univ. Press, 1969; tłum. ros. *Стохастическая аппроксимация*, Izd. Mir, Moskwa 1972.

- [36] W. I. Zangwill, *Programowanie nieliniowe*, WNT, Warszawa 1974.
 - [37] R. Zieliński, *Pewna metoda planowania doświadczeń dla estymacji gradientu regresji drugiego stopnia*, Mat. Stos. 2 (1974), str. 101–108.
 - [38] — *A randomized Kiefer–Wolfowitz procedure*, Trans. 7-th Prague Conf. Information Th., Decision Functions, Random Processes, Praga 1974.
 - [39] — *Unbiased estimation of the derivative of a regresion function*, Proc. Symp. to honour Jerzy Neyman, Warszawa 1974.
 - [40] — *Global stochastic approximation*, Diss. Math. (Rozprawy Matematyczne) 147, Warszawa 1977.
 - [41] — *Wybrane zagadnienia optymalizacji statystycznej*, PWN, Warszawa 1974.
-