ROCZNIKI POLSKIEGO TOWARZYSTWA MATEMATYCZNEGO
SERIA III: MATEMATYKA STOSOWANA II (1974)

R. ZIELIŃSKI (Warszawa)

O optymalizacji statystycznej w R^m

Wstęp. Sformułowanie zadania. W wielu praktycznych sytuacjach pojawiają się zadania wyznaczenia maksimum funkcji, która nie jest explicite znana, chociaż przy każdych ustalonych wartościach argumentów wartość tej funkcji może być oszacowana. To oszacowanie odbywa się najczęściej na podstawie odpowiedniego eksperymentu, którego wynik — przy ustalonych wartościach argumentów funkcji — jest zmienną losową. Ten typ zadań może być ogólnie ujęty w następujący schemat.

Dana jest rodzina zmiennych losowych $\{Y(X)\}$ zależnych od parametru X przebiegającego pewien ustalony zbiór $\mathfrak{X}$. Niech dla każdego X istnieje wartość oczekiwana $EY(X)$; oznaczymy ją przez $F(X)$. Należy wyznaczyć taką wartość parametru X^{opt}, przy której funkcja F osiąga maksimum, przy czym jedynymi dostępnymi informacjami o funkcji F są wartości zmiennych losowych $Y(X)$ zaobserwowane w wybranych punktach $X \in \mathfrak{X}$. O zbiorze $\mathfrak{X}$ będziemy zakładali, że jest pewnym obszarem w przestrzeni R^m,

Zanim dokładniej sformułuję zadanie i przejdę do opisu sposobów podejścia do jego rozwiązania zacytuję przykład dotyczący pewnego praktycznego problemu hutniczego. Przykład ten został po raz pierwszy podany w pracy [11] i od tej pory był wielokrotnie cytowany w różnych pracach omawiających zagadnienia optymalizacji statystycznej.

Badano wytrzymałość stopu na rozerwanie w zależności od ilości składników stopowych (chromu, niklu, molibdenu, wanadu, niobu, manganu oraz węgla). Zadanie polegało na znalezieniu takiej kombinacji tych składników, przy której oczekiwana wytrzymałość na rozerwanie osiąga maksimum.

Zauważmy, że struktura tego przykładowego zadania jest dokładnie taka jak struktura ogólnego zadania sformułowanego na wstępie. Niech bowiem X będzie punktem w siedmiowymiarowej przestrzeni liczb rzeczywistych zdefiniowanym tak, że i-ta współrzędna tego punktu jest równa ilości i-tego składnika w badanym stopie. Mamy więc $X = (x_1, x_2, \ldots, x_7)$, gdzie x_1 oznacza ilość chromu, x_2 — ilość niklu itd. Praktyka wykazuje, że jeżeli weźmiemy dwie próbki stopu o jednakowej zawartości wymienionych wyżej składników i zbadamy wytrzymałość każdej z tych próbek, otrzymamy na ogół różne wyniki. Dzieje się tak dlatego, że na obserwowany rezultat ma wpływ cały szereg innych przyczyn (np. zawartość innych składników, błąd pomiaru); sumaryczny wpływ tych innych, „niekontrolowanych", jak to się często mówi, czynników formalnie opisuje się w ten sposób, że wytrzymałość stopu na rozerwanie przy ustalonym jego składzie X traktuje się jako zmienną losową. W ten sposób z każdym punktem X związana jest pewna zmienna losowa $Y(X)$. Zadanie dobrania takiej kombinacji składników stopowych, przy której oczekiwana wytrzymałość stopu osiąga maksimum może więc być sformułowane jako zadanie wyznaczenia takiego punktu X, w którym wartość oczekiwana zmiennej losowej $Y(X)$ jest największa. Ponieważ zawartość każdego ze składników stopowych wyraża się liczbą z przedziału $(0, 1)$, zbiór $\mathfrak{X}$ w naszym przykładzie jest pewnym podzbiorem zbioru $\{(x_1, x_2, \ldots, x_7)\colon \quad 0 \leqslant x_i \leqslant 1, \sum x_i \leqslant 1\}$.

A oto inne przykłady tego rodzaju. W problemach automatycznego sterowania rozpatruje się taką sytuację, gdy to, co pojawia się na wyjściu obiektu zależy nie tylko od wybieranych przez sterującego parametrów ale również od szeregu innych czynników, których ogólny wpływ traktuje się jako efekt zakłóceń (efekt „szumów"). W ten sposób, z każdym układem parametrów wybranych przez sterującego związana jest pewna zmienna losowa (wyjście obiektu) i zadanie polega na takim wyborze tych parametrów, żeby oczekiwana wartość wynikowej zmiennej losowej (lub, w ogólniejszym przypadku, jakiegoś funkcjonału określonego na stanach i wyjściach obiektu) osiągała wartość ekstremalną. W szczególności w tzw. zagadnieniach kompleksowego sterowania jakością produkcji stawią się zadanie takiego wyboru będących w dyspozycji czynników produkcyjnych, żeby odpowiednio zdefiniowana jakość produktu była możliwie największa. W naukach rolniczych takimi problemami są np. problemy optymalnego nawożenia, zbiór $\mathcal{X}$ może tu być na przykład zbiorem różnych nawozów, obserwowanymi zmiennymi losowymi $Y(X)$ — plony. Przypuszczam, że każdy potrafi natychmiast zacytować tu wiele przykładów „ze swojego własnego podwórka".

Istotnym założeniem w sformułowaniu naszego zadania jest to, że maksymalizowana funkcja F nie jest explicite znana, a informacje jakie możemy o niej uzyskać będą pochodziły z szacowania jej wartości w wybranych punktach zbioru $\mathcal{X}$; liczba punktów, w których możemy dokonać takiego szacowania, jest oczywiście skończona (a praktyk wolałby nawet, żeby była tak mała, jak to jest tylko możliwe). W takiej sytuacji nie ma oczywiście mowy o rozwiązaniu zadania bez bardziej dokładnych założeń o optymalizowanej funkcji.

Regresja „globalna". W wielu praktycznych zadaniach zakłada się, że funkcja F jest elementem pewnej rodziny $\mathcal{F}$, dopuszczającej parametryzację za pomocą skończonego układu liczb rzeczywistych. Niech więc dana będzie rodzina $\mathcal{F} = \{F_\alpha : \alpha \in \mathcal{A}\}$ i niech $\mathcal{A} \subset R^l$ będzie danym zbiorem. Koncepcyjnie zadanie jest teraz proste: na podstawie oszacowanych wartości funkcji w wybranych punktach $X_1, X_2, \ldots, X_n \in \mathcal{X}$ należy „zidentyfikować" parametr α, a następnie znaleźć punkt, w którym funkcja F_α osiąga maksimum. Pewne dodatkowe komplikacje pojawiają się na skutek tego, że dla ustalonego punktu $X \in \mathcal{X}$ wartość funkcji F_α może być oszacowana tylko przez obserwację zmiennej losowej $Y(X)$, niemniej jednak zadanie identyfikacji parametru (tzn. zadanie identyfikacji funkcji) w takiej sytuacji jest dobrze znanym w statystyce matematycznej zadaniem szacowania „współczynników" regresji. Od strony technicznej jednak nawet tak prosto sformułowane zadanie jest trudne i nawet dla „łatwych" klas funkcji nie potrafię przedstawić w pełni zadowalających rozwiązań.

Sformułujmy dokładniej rozważane zadanie. Niech n będzie ustaloną liczbą naturalną, $X_1, X_2, \ldots, X_n$ — ustalonymi (niekoniecznie różnymi) punktami w $\mathcal{X}$. Rozważmy zmienne losowe $Y_1, Y_2, \ldots, Y_n$, gdzie $Y_j = Y(X_j)$. Należy skonstruować funkcję o wartościach w zbiorze $\mathcal{X}$ i argumentach $Y_1, Y_2, \ldots, Y_n$, która będzie przybliżała rozwiązanie optymalne X^{opt}; oznaczymy tę funkcję przez $\hat{X}^{\text{opt}}$. Przybliżenie, o którym wyżej mowa, może być oczywiście rozumiane w różny sposób. Jeżeli wartość oczekiwana zmiennej losowej $\hat{X}^{\text{opt}}$ jest równa X^{opt} mówimy, że $\hat{X}^{\text{opt}}$ jest nieobciążonym estymatorem rozwiązania. W klasie estymatorów nieobciążonych poszukuje się zwykle takiego, który minimalizuje wartość oczekiwaną $E\,\|\hat{X}^{\text{opt}} - X^{\text{opt}}\|^2$ lub wartość oczekiwaną jakoś inaczej zdefiniowanej odległości między $\hat{X}^{\text{opt}}$ i X^{opt}. Czasami jako miarę dokładności estymatora przyjmuje się wartość oczekiwaną różnicy $F(\hat{X}^{\text{opt}}) - F(X^{\text{opt}})$; tu wychodzi się z założenia, że przybliżenie jest dobre, jeżeli jego konsekwencje są takie same lub niewiele gorsze od konsekwencji rozwiązania optymalnego. Rozkład takiej różnicy dla pewnej klasy $\mathcal{F}$ rozważany jest w pracy [6] i w kilku pracach tam cytowanych.

Opisane wyżej podejście do rozwiązania zadania przez specyfikację rodziny $\mathcal{F}$ ma pewne istotne wady z punktu widzenia praktycznych zastosowań. Okazuje się bowiem, że zwykle bardzo trudno jest w zadowalający sposób wyodrębnić klasę funkcji, która dopuszczałaby opisaną wyżej parametryzację i która jednocześnie byłaby dostatecznie „wygodna" dla obliczeń. W analizie regresji rozważa się najczęściej klasę wielomianów, ale aproksymacja funkcji F za pomocą wielomianów na zbiorze $\mathcal{X}$ jest zazwyczaj niezadowalająca z praktycznego punktu widzenia; okazuje się, że wielomiany dobrze spełniają swoją rolę gdy służą do lokalnej aproksymacji funkcji regresji i bardzo źle gdy stosowane są do globalnej aproksymacji tej funkcji na całym obszarze jej określoności. Z kolei, inne niż wielomiany funkcje aproksymujące, nastręczają wiele kłopotów numerycznych.

Metody kolejnych przybliżeń. Aproksymacja stochastyczna. Alternatywne podejście do rozwiązania naszego zadania polega na zastosowaniu metod kolejnych przybliżeń. W ogólnym schemacie mogą one być sformułowane w następujący sposób. Niech $X^0 \in \mathcal{X}$ będzie dowolnym punktem – zerowym przybliżeniem rozwiązania X^{opt}. Jeżeli uzyskano już kolejne przybliżenia $X^0, X^1, \ldots, X^{p-1}$, to przybliżenie X^p konstruuje się na podstawie analizy zachowania się funkcji F w otoczeniu punktu X^{p-1}.

Aby taka procedura prowadziła do rozwiązania wystarcza, żeby funkcja F miała lokalnie pewne pożądane własności, np. żeby była w każdym punkcie odpowiednio wiele razy różniczkowalna. Tego typu założenia mogą być w większości praktycznych zastosowań przyjęte bez najmniejszych zastrzeżeń. Zauważmy jednak, że przy konstrukcji ciągu (X^j) analizę funkcji F w otoczeniu kolejnego przybliżenia przeprowadzamy nie w oparciu o znajomość wartości samej funkcji w wybranych punktach tego otoczenia, lecz w oparciu o zaobserwowane wartości zmiennych losowych $Y(X)$ w tych punktach. W tej sytuacji ciąg (X^j) staje się ciągiem zmiennych losowych i analiza zbieżności metody kolejnych przybliżeń sprowadza się do analizy zbieżności takich ciągów. Dla ilustracji przytoczę jedną z metod budowy ciągu (X^j) w przypadku, gdy $\mathcal{X}$ jest przedziałem liczbowym i funkcja F ma jedno maksimum X^{opt} w tym przedziale. Niech

$$X^{p+1} = X^p + a_p \frac{Y(X^p + \frac{1}{2}c_p) - Y(X^p - \frac{1}{2}c_p)}{c_p}, \tag{1}$$

gdzie (a_p) i (c_p) są pewnymi ciągami liczbowymi. Można dowieść [5], że jeżeli 1) wariancje zmiennych losowych $Y(X)$, $X \in \mathcal{X}$ są wspólnie ograniczone, 2) funkcja F jest rosnąca na lewo od punktu X^{opt}, malejąca na prawo od tego punktu, a jej pochodna spełnia warunek

$$k_0 |X - X^{\text{opt}}| \leqslant F'(X) \leqslant k_1 |X - X^{\text{opt}}|$$

dla pewnych stałych k_0 i k_1 oraz 3) ciągi (a_p) i (c_p) są wybrane tak, że $c_p \to 0$, $\sum a_p = \infty$, $\sum a_p c_p < \infty$, $\sum (a_p/c_p)^2 < \infty$, a_p i c_p są nieujemne, to ciąg $E(X^p - X^{\text{opt}})^2$ jest zbieżny do zera. Inaczej mówiąc, ciąg (X^p) jest zbieżny według średniej do rozwiązania (a więc również zbieżny stochastycznie).

W przypadku, gdy $\mathcal{X}$ jest obszarem w R^m, odpowiednia procedura budowy kolejnego przybliżenia $X^p = (x_1^p, x_2^p, \ldots, x_m^p)$ może być sformułowana na przykład w następujący sposób:

$$x_i^{p+1} = x_i^p + a_p \frac{Y(X^p + \frac{1}{2}c_p e_i) - Y(X^p - \frac{1}{2}c_p e_i)}{c_p} \tag{2}$$

gdzie (a_p) i (c_p) są odpowiednio dobranymi ciągami liczbowymi oraz e_i jest wersorem i-tej osi współrzędnych w R^m. Bywa również stosowana procedura:

$$x_i^{p+1} = x_i^p + a_p \frac{Y(X^p + c_p e_i) - Y(X^p)}{c_p} \tag{3}$$

Dowody zbieżności takich ciągów kolejnych przybliżeń do rozwiązania można znaleźć np. w pracy [17].

Przytoczone wyżej twierdzenie jest typowym twierdzeniem w tzw. teorii aproksymacji stochastycznej. Teoria ta została zapoczątkowana w 1951 roku przez H. Robbinsa i S. Monro w pracy [15], opisującej pewną metodę kolejnych przybliżeń dla rozwiązania równania $F(X) = 0$, gdzie F – funkcja regresji. W 1952 roku teoria ta została rozszerzona przez J. Kiefera i J. Wolfowitza [10] na zadania wyznaczania maksimum funkcji regresji. Teoria aproksymacji stochastycznej zaczęła się szybko rozwijać, a ze względu na jej liczne zastosowania (por. przykłady na początku artykułu) rozwija się nadal i absorbuje uwagę szerokich kręgów specjalistów różnych dziedzin. Odnotujmy, że w pierwszej połowie 1972 roku ukazała się w rosyjskim tłumaczeniu praca M.T. Wasana [17], wydana w 1969 w Cambridge, omawiająca aktualny stan tej teorii; tłumaczenie rosyjskie zostało uzupełnione 30 stronicowym rozdziałem zawierającym przegląd niektórych najnowszych wyników. Z bardziej interesujących rozwinięć tej teorii warto odnotować pracę V. Fabiana z 1965 roku [7] przedstawiającą metody aproksymacji stochastycznej w zastosowaniu do wyznaczania ekstremów warunkowych (czyli do typowych zadań programowania nieliniowego).

Metody kolejnych przybliżeń. Zagadnienia lokalne. Centralnym problemem teorii aproksymacji stochastycznej są graniczne własności ciągów kolejnych przybliżeń. Podstawowe twierdzenia tej teorii odnoszą się do zbieżności takich ciągów (rozpatruje się różne rodzaje zbieżności zmiennych losowych), szybkości tej zbieżności oraz rozkładów granicznych zmiennych losowych X^j, $j = 0, 1, 2, \ldots$ Z praktycznego punktu widzenia istotne znaczenie mają jednak przede wszystkim „lokalne" własności takich ciągów; chodzi mianowicie o to, żeby ciąg (X^j) kolejnych przybliżeń maksimum funkcji F miał taką własność, że ciąg $(F(X^j))$ jest ciągiem rosnącym, a w każdym bądź razie żeby prawdopodobieństwo zdarzenia $\{F(X^{p+1}) > F(X^p)\}$ było duże. Takie żądania wynikają z faktu, że w praktyce dla oszacowania wartości funkcji F w ustalonym punkcie X trzeba wykonywać pewne eksperymenty związane z obserwacją zmiennej losowej $Y(X)$, jak to miało miejsce w rozważanym na początku przykładzie optymalizacji składu stopu lub przykładzie zastosowań rolniczych – w tym ostatnim przypadku jeden eksperyment trwa nieraz cały rok! Chodzi więc po prostu o to, żeby liczba eksperymentów potrzebnych do znalezienia optymalnego rozwiązania nie była zbyt duża. Z tych samych powodów spośród dwóch sposobów (2) i (3) budowy ciągu kolejnych przybliżeń, praktyk wybierze drugi sposób, bo dla wykonania jednego „kroku" iteracyjnego potrzeba tu wykonać $(m+1)$ eksperymentów zamiast $2m$ eksperymentów w (2). Powstaje oczywiście natychmiast pytanie, czy istnieją jeszcze bardziej oszczędne (albo w jakimś innym sensie lepsze) metody budowy kolejnego przybliżenia niż przykładowo zacytowane (2) i (3). Odpowiedź na to pytanie jest twierdząca, a te „najlepsze" metody budowy kolejnego przybliżenia opierają się na lokalnej analizie funkcji regresji w otoczeniu już osiągniętego rozwiązania. Interesującym jest, że takie metody lokalnej analizy funkcji regresji lub, jak to się nieraz mówi, „powierzchni odpowiedzi" („response surface analysis") pojawiły się zupełnie niezależnie od omawianej wyżej teorii stochastycznej aproksymacji. Odnotować przy tym należy, że pojawiły się one dla rozwiązania konkretnych problemów optymalizacyjnych w przemyśle chemicznym. Za początek tego kierunku optymalizacji statystycznej należy uznać pracę

G.E.P. Boxa i K.B. Wilsona z 1951 r. [4]. W Polsce znana jest na ten temat wydana w 1967 roku, monografia W.W. Nalimowa i N.A. Czernowej [13].

Mówiąc bardzo ogólnie, zadaniem lokalnej analizy funkcji regresji w otoczeniu już osiągniętego rozwiązania X^p jest oszacowanie kierunku, w którym należy szukać następnego rozwiązania X^{p+1}, a jeżeli punkt optymalny zostanie już osiągnięty – opisanie zachowania się funkcji regresji w bezpośrednim otoczeniu tego punktu. Za kierunek poszukiwań nowego rozwiązania przyjmuje się najczęściej kierunek gradientu funkcji regresji. Przedstawione wyżej zadania można sformułować jako zadanie oszacowania gradientu, a w przypadku, gdy w osiągniętym punkcie gradient jest równy zeru – przeanalizowanie typu osiągniętego punktu stacjonarnego funkcji.

Już zadanie oszacowania gradientu funkcji regresji stanowi nie banalny problem. Ponieważ funkcja F z założenia nie jest znana, nie możemy oczywiście po prostu jej zróżniczkować i obliczać wartość pochodnych cząstkowych jako składowych gradientu. Jedynymi dostępnymi są tu tylko skończenie-różnicowe oszacowania gradientu, takie jakie możemy łatwo rozpoznać we wzorach (1), (2) i (3). Powstają w związku z tym kłopoty nawet w przypadku, gdy w każdym punkcie $X \in \mathfrak{X}$ możemy dokładnie obliczać wartości funkcji F a nie, jak w naszym przypadku, tylko szacować je za pomocą zmiennych losowych $Y(X)$.

W przypadku, gdy funkcja regresji może być lokalnie z dostateczną dla praktycznych zastosowań dokładnością aproksymowana wielomianem pierwszego stopnia

$$W(X) = a_0 + a_1 x_1 + \ldots + a_m x_m \tag{4}$$

zadanie wyznaczenia gradientu $(a_1, a_2, \ldots, a_m)$ tej funkcji sprowadza się do znanego w statystyce matematycznej zadania szacowania współczynników regresji. Niech $Y_1, Y_2, \ldots, Y_n$ będą zmiennymi losowymi obserwowanymi w niekoniecznie różnych punktach $X_1, X_2, \ldots, X_n$ $(X_j = (x_{1j}, x_{2j}, \ldots, x_{mj}))$ należących do rozpatrywanego otoczenia kolejnego rozwiązania (np. we wzorze (1) są to punkty $X^p + \frac{1}{2} c_p$ oraz $X^p - \frac{1}{2} c_p$). Macierz

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{5}$$

nazywa się *macierzą planu eksperymentu* lub *macierzą planowania*. Jak wiadomo (por. np. [14]), w przypadku, gdy zmienne losowe Y_j, $j = 1, 2, \ldots, n$ są niezależne, a macierz XX^T jest nieosobliwa, nieobciążonym i najefektywniejszym estymatorem $\hat{a} = (\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_m)$ współczynników regresji $a = (a_0, a_1, \ldots, a_m)$ jest

$$\hat{a} = (XX^T)^{-1} XY, \tag{6}$$

gdzie $Y = (Y_1, Y_2, \ldots, Y_n)$. W najprostszym przypadku, gdy zmienne losowe Y_j $(j = 1, 2, \ldots, n)$ mają jednakową wariancję σ^2, macierz kowariancji estymatora $\hat{a}$ jest równa $(XX^T)^{-1}\sigma^2$. (Ogólny przypadek zależnych zmiennych losowych Y_j z dowolną macierzą kowariancji rozważany jest w różnych podręcznikach statystyki matematycznej, np. w książce [14]).

Oszacowania gradientu konstruowane w procesie optymalizacji statystycznej mogą oczy-

wiście tylko w takim sensie przybliżać ten gradient, w jakim zmienna losowa przybliża interesującą nas stałą. Naturalne jest postawienie problemu wyboru macierzy planowania tak, żeby to przybliżenie było w jakimś sensie najdokładniejsze. Tego typu zagadnienia rozważane są w rozwijającej się bardzo intensywnie teorii planowania eksperymentów. Podstawy tej teorii zostały przedstawione w pracy S. Karlina i W.J. Studdena [9]. Aktualny wykład metod planowania doświadczeń znajduje się w wydanej w 1971 roku książce [8], a bardziej popularne omówienie najważniejszych jej wyników pod kątem widzenia zastosowań w przemyśle — w pracy [1].

W praktyce szerokie zastosowanie znalazły tzw. plany sympleksowe; według tych planów zmienne losowe $Y(X)$ obserwuje się w $(m+1)$ punktach $X_1, X_2, \ldots, X_{m+1}$ położonych w wierzchołkach regularnego sympleksu o środku w punkcie wyznaczonym przez osiągnięte rozwiązanie X^p. Takie plany mają kilka bardzo wygodnych z praktycznego punktu widzenia własności, a mianowicie: 1) są to plany minimalne w tym sensie, że dla oszacowania $m+1$ współczynników regresji (4) wykorzystują tylko $m+1$ zmiennych losowych; 2) składowe $\hat{a}_j$ estymatora $\hat{a}$ są niezależnymi zmiennymi losowymi oraz 3) w klasie planów $(m+1)$-punktowych dają najmniejsze wariancje $D^2\hat{a}_j$, przy czym wszystkie te wariancje są jednakowe (por. np. G.E.P. Box [3]).

Przypadek, gdy aproksymacja funkcji regresji za pomocą funkcji liniowej (4) jest za mało dokładna, jest bardziej złożony. Żeby uświadomić sobie konsekwencje szacowania gradientu za pomocą różnic skończonych w takich przypadkach, rozpatrzmy jako przykład funkcję dwóch zmiennych $H(x, y) = x + y - a(x^2 + y^2)$, $a > 0$. Gradient tej funkcji w punkcie $(0, 0)$ jest równy $(1, 1)$, a skończenie-różnicowe oszacowanie tego gradientu za pomocą wzoru

$$\left(\frac{H(x+h, y) - H(x, y)}{h}, \frac{H(x, y+h) - H(x, y)}{h}\right)$$

daje wynik $(1 - ah, 1 - ah)$. Jeżeli więc $h > 1/a$ (czyli gdy a jest duże oraz h niezbyt małe), w wyniku takiego oszacowania otrzymujemy kierunek przeciwny do kierunku gradientu. Można tu oczywiście rozważać inne skończenie-różnicowe oszacowania gradientu, ale nie zmienia to w istotny sposób obrazu.

Przypadek szacowania gradientu funkcji regresji jest analogiczny; gdy aproksymacja tej funkcji za pomocą wielomianu pierwszego stopnia jest zbyt niedokładna, szacowanie gradientu za pomocą takich formuł jak w (1), (2) lub (3) prowadzi do obciążonych estymatorów tego gradientu i w konsekwencji do szukania kolejnego rozwiązania w fałszywym kierunku. Przy odpowiednich założeniach o gładkości funkcji regresji można oczywiście dostatecznie dokładnie aproksymować tę funkcję za pomocą wielomianu odpowiednio wysokiego stopnia, ale z praktycznego punktu widzenia takie postępowanie jest mało przydatne. Wielomian m zmiennych k-tego stopnia ma $\binom{m+k}{k}$ współczynników, oszacowanie takiego wielomianu wymaga więc obserwacji zmiennych losowych $Y(X)$ w co najmniej $\binom{m+k}{k}$ różnych punktach, a taka liczba eksperymentów może być w praktyce po prostu nie do zrealizowania. Niestety, w ogólnym przypadku nie są znane zadowalające estymatory gradientu funkcji regresji nie wymagające szacowania wszystkich $\binom{m+k}{k}$ współczynników. Praktycy najczęściej ograniczają się do wielomianów stopnia drugiego (czasem trzeciego, gdy liczba zmiennych nie jest zbyt duża), a jeżeli taka aproksymacja staje się zbyt niedokładna (weryfikację tej dokładności przeprowadza się za pomocą tzw. testów adekwatności, o czym powiem za chwilę) — po prostu zmniejszają obszar, na którym funkcję regresji aproksymuje się takim wielomianem. Postępowanie takie jest stosunkowo proste. Przypuśćmy, że dla lokalnej aproksymacji funkcji

regresji w otoczeniu kolejnego rozwiązania X^p zastosowano plan $X_1, X_2, \ldots, X_n$. Niech $d = \max_{1 \leqslant j \leqslant n} \|X_j - X^p\|$ będzie średnicą tego planu. Jeżeli okaże się, że aproksymacja funkcji regresji za pomocą wielomianu danego stopnia jest za mało dokładna, konstruuje się nowy plan o średnicy λd $(0 < \lambda < 1)$, biorąc za nowy punkt X_j w tym planie np. punkt $X^p + \lambda(X_j - X^p)$. Jeżeli funkcja regresji jest dostatecznie gładka, to takie postępowanie prowadzi do celu: do lokalnej aproksymacji tej funkcji za pomocą wielomianu danego stopnia.

Ewentualnego wyjaśnienia wymaga sposób weryfikacji dokładności aproksymacji. Idea takiej weryfikacji jest stosunkowo prosta. Rozważmy najpierw przypadek, gdy funkcja $H(x_1, x_2, \ldots, x_m)$ może być w każdym punkcie dokładnie obliczona. Przypuśćmy, że zdecydowaliśmy się aproksymować tę funkcję za pomocą wielomianu danego stopnia $W(x_1, x_2, \ldots, x_m)$. Zmierzyliśmy więc wartości funkcji H w odpowiedniej liczbie punktów $X_1, X_2, \ldots, X_n$ i oszacowaliśmy współczynniki tego wielomianu. Dla weryfikacji dokładności takiego przybliżenia obliczamy wartość funkcji H oraz wartość wielomianu W w pewnych dodatkowych punktach $X_{n+1}, X_{n+2}, \ldots, X_{n+r}$ interesującego nas obszaru aproksymacji. Jeżeli wszystkie różnice $H(X_j) - W(X_j)$, $j = n+1, \ldots, n+r$, są bliskie zeru, aproksymację uważamy za wystarczająco dokładną (kwantyfikacja „bliskości zeru" zależy oczywiście od praktycznych aspektów rozważanego zagadnienia). Taka procedura weryfikacji dokładności aproksymacji może oczywiście budzić zastrzeżenia, ale jest to chyba jedyna procedura dostępna w sytuacji, gdy o funkcji H wiemy tylko tyle, ile jest to możliwe na podstawie obliczania jej wartości w skończonej liczbie punktów.

Przypadek weryfikacji dokładności aproksymacji funkcji regresji $F(X)$ za pomocą danego wielomianu $W(X)$ nie różni się od wyżej opisanego poza pewnymi szczegółami technicznymi wynikającymi stąd, że wartość funkcji F w danym punkcie X jest nam teraz znana tylko za pośrednictwem wartości zmiennej losowej $Y(X)$. W związku z tym sprawdzenie, czy w wybranych punktach X_j wartości $F(X_j) - W(X_j)$ są dostatecznie bliskie zeru sprowadza się do weryfikacji odpowiednich hipotez statystycznych za pomocą odpowiednio skonstruowanych testów. Konstrukcja takiego testu wymaga na ogół wprowadzenia pewnych założeń odnośnie do rozkładów zmiennych losowych $Y(X)$; nie będziemy tego problemu tutaj szczegółowo omawiali ze względu na moc szczegółów technicznych,którymi należałoby się zająć; ogólna idea postępowania jest jasna na podstawie tego co wyżej powiedziano.

Według opisanej wyżej metody kolejnych przybliżeń ciąg $X^0, X^1, X^2, \ldots$ należy konstruować tak długo, aż w kolejno osiągniętym punkcie gradient maksymalizowanej funkcji będzie równy zeru. W przypadku, gdy maksymalizowaną funkcją jest funkcja regresji, której wartości w danych punktach mogą być obserwowane tylko za pośrednictwem pewnych zmiennych losowych, problem znowu się komplikuje. Rozstrzygnięcie, czy w danym punkcie gradient jest zerem może nastąpić tylko na drodze testowania odpowiednich hipotez statystycznych. Zwykle po prostu weryfikuje się hipotezę, że gradient jest równy zeru, a więc hipotezę o współczynnikach regresji. Dyskwalifikacja tej hipotezy prowadzi do wykonania kolejnego kroku iteracyjnego w oszacowanym kierunku gradientu; w przeciwnym przypadku przeprowadza się dokładniejszą analizę funkcji regresji aproksymując ją wielomianami wyższych stopni — najczęściej, o czym już mówiliśmy, wielomianem stopnia drugiego. Taka analiza ma przede wszystkim doprowadzić do oceny osiągniętego punktu stacjonarnego (maksimum? siodło?), a w przypadku stwierdzenia, że osiągnięty punkt jest punktem maksimum (a więc rozwiązaniem zadania) — oszacowanie dokładności uzyskanego rozwiązania oraz dokładniejsze przestudiowanie zachowania się funkcji wokół tego rozwiązania. Typowym podejściem statystycznym do oceny dokładności uzyskanych rozwiązań jest podejście polegające na konstrukcji odpowiednich „przedziałów" ufności. Przypadek funkcji regresji jednej

zmiennej nie nastręcza tu innych niż techniczne kłopotów. Przypadek funkcji wielu zmiennych wymaga zwykle bardziej starannej analizy; nie banalną sprawą jest przy tym wybór kształtu obszaru ufności. Stają się tu aktualne różne kłopoty związane z oceną dokładności estymatorów wielowymiarowych (por. np. [2], [18]); interesujące sposoby konstrukcji wielowymiarowych przedziałów ufności znaleźć można w pracy [16].

Kilka uwag końcowych. Przede wszystkim należy stwierdzić, że praktyka optymalizacji statystycznej jest znacznie bogatsza niż jej teoria. Szczególnie intensywnie rozwijają się w praktyce te metody, które prowadzą do polepszenia osiągniętego już rozwiązania na drodze lokalnej analizy funkcji regresji. Wynika to być może stąd, że z praktycznego punktu widzenia bardziej użyteczne są takie metody, które już dziś mogą dać rozwiązania lepsze od rozwiązań aktualnie znanych niż metody pozwalające co prawda osiągnąć rozwiązanie optymalne, ale wymagające ogromnej pracy teoretycznej związanej przede wszystkim z precyzyjnym sformalizowaniem zadania. Jest to w pewnym sensie naturalne; z własnego doświadczenia obserwujemy na przykład, że produkuje się coraz doskonalsze telewizory, chociaż chyba nikt nie potrafi zdefiniować relacji porządku w zbiorze telewizorów tak, żeby była ona zgodna z „powszechnie odczuwalnymi preferencjami". A jak sformułować zadanie konstrukcji „optymalnego" telewizora? Czy w ogóle takie pytanie ma sens?

Odbiciem tej dominacji praktycznych pomysłów poszukiwania lepszych rozwiązań (często zresztą pomysłów bezsensownych) nad teorią jest sytuacja w piśmiennictwie. Liczba pozycji bibliograficznych poświęcona opisowi różnych „procesów optymalizacyjnych", które w praktyce pozwoliły uzyskać jakieś rozwiązania, wyraża się z pewnością w tysiącach (obfity spis prac tego typu znaleźć można na przykład w cytowanych już pracach [1] i [13]). Liczba prac teoretycznych jest znacznie skromniejsza i dotyczy przede wszystkim różnych matematycznych aspektów aproksymacji stochastycznej (spis takich prac podano w [17], przy czym w rosyjskim wydaniu lista ta została znacznie rozszerzona) oraz teorii planowania doświadczeń (bibliografia znajduje się w [8] i [12]).

Prace cytowane

[1] Ju. P. Adler, *Wwiedienje w planirowanie eksperimenta.* Wyd. „Metalurgia". Moskwa 1969.

[2] T. W. Anderson, *An introduction to multivariate statistical analysis,* New York, London (jest tłumaczenie rosyjskie z 1963 roku).

[3] G. E. P. Box, *Multi-factor designs of first order,* Biometrika 39 (1952), str. 49–57.

[4] G. E. P. Box and K. B. Wilson, *On the experimental attainment of optimum conditions,* J. Roy. Statist. Soc. Ser. B. XIII (1951), str. 1–45.

[5] V. Dupač, *On the Kiefer-Wolfowitz approximation method,* Časopis Pěst. Mat. 82 (1957), str. 47–75.

[6] R. C. Eland t-Johnson, *"Optimal" policy in a maintenance cost problem,* Opns. Res. 15 (1967), str. 813–819.

[7] V. Fabian, *Stochastic approximation of constrained minima,* Trans. 4-th Prague Conf. Inform. Theory, Statist. Decis. Funct., Random Processes. Prague 1967, str. 277–290.

[8] W. W. Fiedorow, *Teoria optimalnogo eksperimenta.* Wyd. Nauka, Moskwa 1971.

[9] S. Karlin and W. J. Studden, *Optimal experimental designs,* Ann. Math. Statist. 37 (1966), str. 783–815.

[10] J. Kiefer and J. Wolfowitz, *Stochastic estimation of the maximum of a regression function,* Ann. Math. Statist. 23 (1952), str. 462–466.

[11] G. H. Li, *Worksheet gives optimum conditions,* Chemical Engineering 65 (1958), str. 4.

[12] W. W. Nalimow (red), *Nowyje idei w planirowanii eksperimenta.* Wyd. Nauka. Moskwa 1969.

[13] W. W. Nalimow i N. A. Czernowa, *Statystyczne metody planowania doświadczeń ekstremalnych,* Warszawa 1967.

[14] C. R. R a o, *Linear statistical inference and its applications,* New York 1965 (jest tłumaczenie rosyjskie z 1968 r.).
[15] H. R o b b i n s and S. M o n r o, *A stochastic approximation method,* Ann. Math. Statist. 22 (1951), str. 400–407.
[16] D. L. W a l l a c e, *Intersection region confidence procedure with an application to the location of the maximum in quadratic regression,* Ann. Math. Statist. 29 (1958), str. 455–475.
[17] M. T. W a s a n, *Stochastic approximation,* Cambridge 1969 (jest tłumaczenie rosyjskie z 1972 r.).
[18] S. S. W i l k s, *Mathematical statistics,* New York 1962 (jest tłumaczenie rosyjskie z 1967 r.).