Metody numeryczne

12. Minimalizacja: funkcje wielu zmiennych

Precyzyjna lokalizacja minimum

P. F. Góra

https://zfs.fais.uj.edu.pl/pawel_gora

10 stycznia 2023

Przypomnienie — daleko od minimum

- **nie minimalizujemy**, tylko podążamy w stronę malejących wartości funkcji
- najczęściej stosowana metoda: metoda najszybszego spadku (*gradient descent*)
- w przypadku Big Data możiwe modyfikacje, typu *Stochastic Gradient Descent*

- w obliczeniach naukowych i inżynierskich najczęściej stosowana metoda Levenberga-Marquardta
  - daleko od minimum zachowuje się jak metoda najszybszego spadku
  - w miarę zbliżania się do minimum wykorzystuje informację o lokalnej krzywiźnie
  - dodatnia określoność hessianu kryterium bycia blisko minimum
  - Levenberg-Marquardt kiepsko nadaje się do Big Data

Strategia precyzyjnej minimalizacji wielowymiarowej

Jeżeli jesteśmy dostatecznie blisko minimum — co stwierdzamy albo na podstawie dodatniej określoności hessjanu, albo wiedząc, że krajobraz daleko od minimum (minimów) ma "strukturę lejka" (*funnel structure*), to znaczy dla dużych wartości argumentów funkcja szybko maleje i ryzyko rozbieżności jest niewielkie — można, a niekiedy wręcz trzeba, zastosować metody, które pozwolą na bardziej precyzyjną, a przy tym możliwie szybką lokalizację minimum. Metody te wykorzystują poznane już metody minimalizowania funkcji jednej zmiennej. Przedstawimy strategię poszukiwania (lokalnego) minimum tej funkcji w postaci ciągu minimalizacji jednowymiarowych.

1. Aktualnym przybliżeniem minimum jest punkt $\mathbf{x}_k$.
2. Wybieramy pewien kierunek poszukiwań $\mathbf{p}_k$.
3. Konstruujemy funkcję g_k:$\mathbb{R} \to \mathbb{R}$

$$g_k(\alpha) = f(\mathbf{x}_k + \alpha \mathbf{p}_k)\,. \tag{1}$$

   Zauważmy, że funkcja $g_k(\alpha)$ jest funkcją jednej zmiennej. Geometrycznie jest to "ślad" $N+1$-wymiarowego wykresu funkcji $f(\mathbf{x})$ przeciętego płaszczyzną zawierającą punkt $\mathbf{x}_k$ i wektor $\mathbf{p}_k$.
4. Znanymi metodami jednowymiarowymi znajdujemy α_{min} takie, że funkcja (1) osiąga minimum. Jest to *minimum kierunkowe* funkcji f.
5.

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_{\mathrm{min}} \mathbf{p}_k\,. \tag{2}$$

6. *goto* 1.

Przykład

Niech $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x_1, x_2) = 2.5x_1^2 + x_1x_2 + x_2^2 - x_1 - x_2$. Przypuśćmy, że dany jest punkt $\mathbf{x}_k = (1, 2)$ oraz kierunek poszukiwań $\mathbf{p}_k = [-1, 1]^T$. Wówczas

$$\begin{aligned} g_k(\alpha) &= f(\mathbf{x}_k + \alpha\mathbf{p}_k) \\ &= 2.5(1-\alpha)^2 + (1-\alpha)(2+\alpha) + (2+\alpha)^2 - (1-\alpha) - (2+\alpha) \\ &= 2.5\alpha^2 - 2\alpha + 5.5 \end{aligned} \tag{3}$$

Jest to funkcja jednej zmiennej, α. Osiąga ona minimum dla $\alpha = \frac{2}{5}$. Zatem *następnym* punktem będzie

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \frac{2}{5}\mathbf{p}_k = \left(1 - \frac{2}{5}, 2 + \frac{2}{5}\right) = (0.6, 2.4) \tag{4}$$

Wystarczy teraz wybrać kolejny kierunek poszukiwań $\mathbf{p}_{k+1}$ etc.

Jak wybierać kierunki poszukiwań?

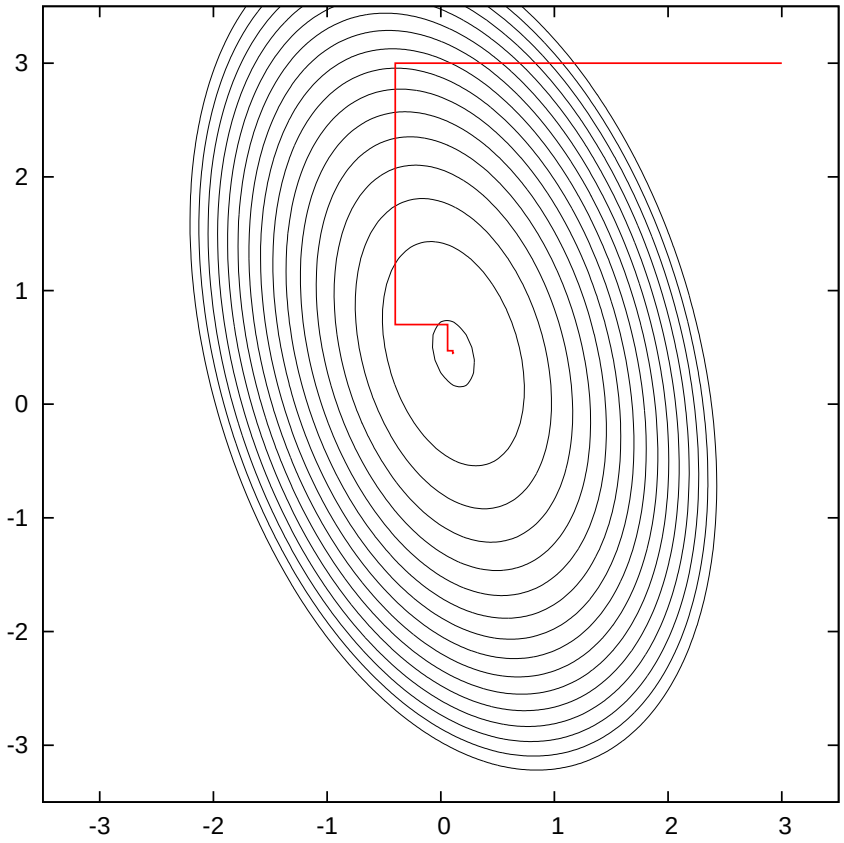
Cały problem sprowadza się zatem do wyboru odpowiedniej sekwencji kolejnych kierunków $\{\mathbf{p}_k\}$.

Następujące strategie wyboru kierunkow poszukiwań są bardzo popularne:

- Minimalizacji po współrzędnych — kolejnymi kierunkami poszukiwań są kierunki równoległe do kolejnych osi układu współrzędnych.
- *Metoda najszybszego spadku*, w której kierunek poszukiwań pokrywa się z minus gradientem minimalizowanej funkcji w aktualnym punkcie.

Okazuje się, że jeśli jesteśmy blisko minimum, nie są to dobre pomysły, gdyż prowadzą do wielu drobnych kroków, które częściowo likwidują efekty osiągnięte w krokach poprzednich. Dlaczego?

Przykład — minimalizacja po współrzędnych

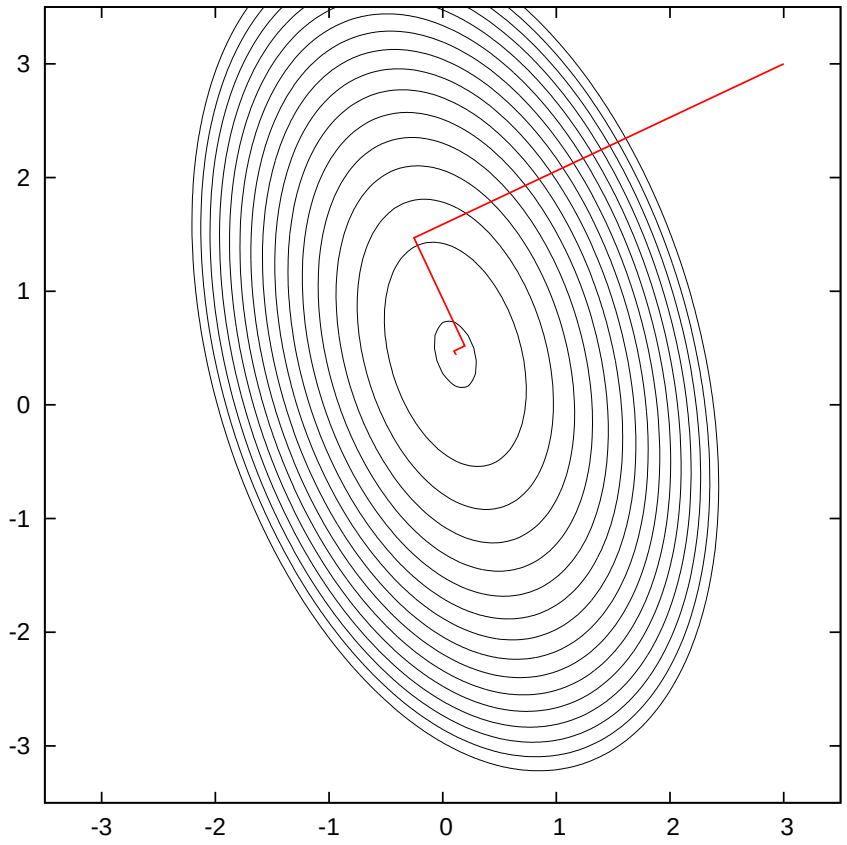


Dużo małych kroków. Osie układu nijak się mają do “naturalnej” geometrii minimalizowanej funkcji.

W metodzie najszybszego spadku co prawda realizujemy najważniejszy pomysł na minimalizację —— zawsze podążamy w kierunku *malejących* wartości (czyli kierunkiem poszukiwań jest minus gradient funkcji w aktualnym punkcie) —— ale czasami poszukiwanie minimów kierunkowych bywa niepotrzebnie kosztowne. W dodatku, jeśli jesteśmy już blisko minimum, okazuje się, że metoda najszybszego spadku także prowadzi do konieczności wykonywania wielu małych kroków, które częściowo niwelują efekty osiągnięte w krokach poprzednich.

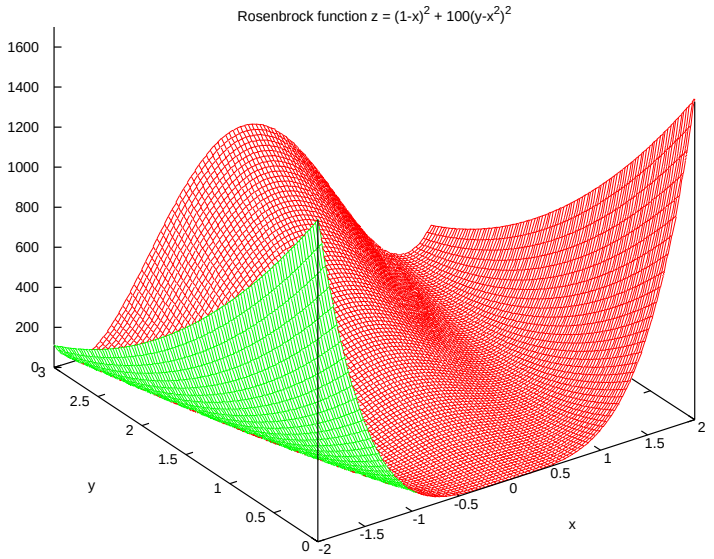
Przy precyzyjnej lokalizacji minimum, **główny**, największy koszt obliczeniowy ponosimy wykonując wiele drobnych kroków końcowych, już w bezpośrednim sąsiedztwie minimum. Gdyby udało nam się zredukować liczbę tych kroków, odnieślibyśmy znaczny zysk na kosztach obliczeniowych.

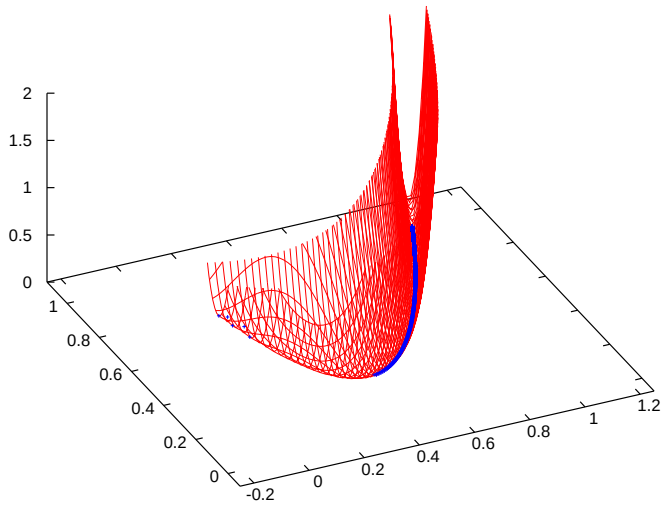
Przykład — metoda najszybszego spadku



Wygląda lepiej, ale kroków prawie tyle samo

Funkcja Rosenbrocka





Funkcja Rosenbrocka $f(x, y) = (1 - x)^2 + 100(y - x^2)^2$ jest przykładem funkcji, którą trudno zminimalizować. W klasycznym zadaniu minimalizuje się tę funkcję startując z punktu $(-3, -4)$. Wartość funkcji w tym punkcie wynosi 16916, zaś gradient to $[-15608, -2600]^T$, należy zatem poruszać się *w kierunku* minus gradientu, ale *o bardzo niewielki ułamek* jego długości. Prawy panel pokazuje końcowy przebieg (niezbyt udanej) minimalizacji funkcji Rosenbrocka za pomocą metody najszybszego spadku.

W pobliżu minimum

Znajdźmy warunek na to, aby f osiągała minimum kierunkowe, czyli aby g_k osiągała minimum:

$$\frac{dg_k}{d\alpha} = \sum_i \frac{\partial f}{\partial x_i}(\mathbf{p}_k)_i = \left(\nabla f|_{\mathbf{x}=\mathbf{x}_{\text{min}}}\right)^T \mathbf{p}_k = 0\,. \tag{5}$$

W minimum kierunkowym gradient funkcji jest prostopadły do kierunku poszukiwań. Zatem w metodzie najszybszego spadku kierunek poszukiwań (lokalny kierunek minimalizacji) co prawda zaczyna się prostopadle do poziomnic funkcji, ale kończy się *stycznie* do poziomnic. Natomiast w minimalizacji po współrzędnych kolejne kierunki poszukiwań, czyli — tutaj — kolejne współrzędne, nie zależą od kształtu minimalizowanej funkcji; taka strategia nie może być optymalna.

Przybliżenie formy kwadratowej

Przypuśćmy, że jesteśmy dostatecznie blisko minimum. Rozwijamy minimalizowaną funkcję w szereg Taylora wokół minimum i otrzymujemy

$$f(\mathbf{x}) \simeq \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x} + f_0\,, \tag{6}$$

gdzie $\mathbf{A}$ jest macierzą drugich pochodnych cząstkowych (hessjanem) obliczanym w minimum. Z definicji minimum, macierz ta jest dodatnio określona, jeśli zaś funkcja jest dostatecznie gładka, macierz ta jest symetryczna. Zatem w pobliżu minimum, funkcja w przybliżeniu zachowuje się jak dodatnio określona forma kwadratowa.

Gradienty sprzężone

W przybliżeniu (6) gradient funkcji f w punkcie $\mathbf{x}_k$ wynosi

$$\nabla f|_{\mathbf{x}=\mathbf{x}_k} = \mathbf{A}\mathbf{x}_k - \mathbf{b}\,. \tag{7}$$

Kolejne poszukiwania odbywają się w kierunku $\mathbf{p}_{k+1}$. Gradient funkcji w pewnym nowym punkcie $\mathbf{x} = \mathbf{x}_k + \alpha\mathbf{p}_{k+1}$ wynosi

$$\nabla f|_{\mathbf{x}} = \mathbf{A}\mathbf{x}_k + \alpha\mathbf{A}\mathbf{p}_{k+1} - \mathbf{b}\,. \tag{8}$$

Zmiana gradientu wynosi

$$\delta\left(\nabla f\right) = \alpha\mathbf{A}\mathbf{p}_{k+1}\,. \tag{9}$$

Punkt $\mathbf{x}_k$ jest minimum kierunkowym w kierunku $\mathbf{p}_k$, a więc gradient funkcji w tym punkcie spełnia warunek (5). *Jeżeli chcemy aby poszukiwania w nowym kierunku nie zepsuły minimum kierunkowego w kierunku $\mathbf{p}_k$, zmiana gradientu musi być prostopadła do starego kierunku poszukiwań*, $\delta\,(\nabla f)^T\,\mathbf{p}_k = 0$. Tak jednak musi być dla *wszystkich* poprzednich kierunków, nie chcemy bowiem naruszyć żadnego z poprzednich minimów kierunkowych. A zatem

$$\mathbf{p}_i^T \mathbf{A} \mathbf{p}_j = 0\,, \quad i > j\,. \tag{10}$$

Metodę wybierania kierunków poszukiwań spełniających (10) nazywamy *metodą gradientów sprzężonych*.

Jak się obejść bez hessjanu?

Z jednego z poprzednich wykładów znamy *algebraiczną* metodę gradientów sprzężonych, wydaje się zatem, iż moglibyśmy jej użyć do skonstruowania ciągu wektorów $\{\mathbf{p}_k\}$. Niestety, nie możemy, nie znamy bowiem macierzy $\mathbf{A}$, czyli hessjanu w minimum. Czy możemy się bez tego obejść?

Twierdzenie 1. *Niech f ma postać* (6) *i niech $\mathbf{r}_k = -\nabla f|_{\mathbf{x}_k}$. Z punktu $\mathbf{x}_k$ idziemy w kierunku $\mathbf{p}_k$ do punktu, w którym f osiąga minimum kierunkowe. Oznaczmy ten punkt $\mathbf{x}_{k+1}$. Wówczas $\mathbf{r}_{k+1} = -\nabla f|_{\mathbf{x}_{k+1}}$ jest tym samym wektorem, który zostałby skonstruowany w algebraicznej metodzie gradientów sprzężonych.*

Dowód. Na podstawie równania (7), $\mathbf{r}_k = -\mathbf{A}\mathbf{x}_k + \mathbf{b}$ oraz

$$\mathbf{r}_{k+1} = -\mathbf{A}(\mathbf{x}_k + \alpha\mathbf{p}_k) + \mathbf{b} = \mathbf{r}_k - \alpha\mathbf{A}\mathbf{p}_k\,. \tag{11}$$

W minimum kierunkowym $\mathbf{p}_k^T \nabla f|_{\mathbf{x}_{k+1}} = -\mathbf{p}_k^T \mathbf{r}_{k+1} = 0$ (por. (5)). Wobec tego mnożąc równanie (11) lewostronnie przez $\mathbf{p}_k^T$, otrzymujemy

$$\alpha = \frac{\mathbf{p}_k^T \mathbf{r}_k}{\mathbf{p}_k^T \mathbf{A}\mathbf{p}_k}\,. \tag{12}$$

Ponieważ w algebraicznej metodzie gradientów sprzężonych $\mathbf{r}_k^T \mathbf{p}_k = \mathbf{r}_k^T \mathbf{r}_k$, otrzymujemy dokładnie takie samo α jak we wzorach na metodę algebraiczną, co kończy dowód. □

Algorytm gradientów sprzężonych

Rozpoczynamy w pewnym punkcie $\mathbf{x}_1$. Bierzemy $\mathbf{r}_1 = \mathbf{p}_1 = -\nabla f|_{\mathbf{x}_1}$.

1. Będąc w punkcie $\mathbf{x}_k$, dokonujemy minimalizacji kierunkowej w kierunku $\mathbf{p}_k$; osiągamy punkt $\mathbf{x}_{k+1}$.
2. Obliczamy $\mathbf{r}_{k+1} = -\nabla f|_{\mathbf{x}_{k+1}}$.
3. Obliczamy (jak w metodzie algebraicznej)

$$\beta = \frac{\mathbf{r}_{k+1}^T \mathbf{r}_{k+1}}{\mathbf{r}_k^T \mathbf{r}_k} \,. \tag{13}$$

4. Obliczamy (jak w metodzie algebraicznej) $\mathbf{p}_{k+1} = \mathbf{r}_{k+1} + \beta \mathbf{p}_k$.

W metodzie gradientów sprzężonych te kroki, które *wymagałyby* znajomości hessjanu w minimum, zastępujemy minimalizacją kierunkową, natomiast te kroki, które *nie wymagają* znajomości hessjanu, wykonujemy tak samo, jak w algebraicznej metodzie gradientów sprzężonych.

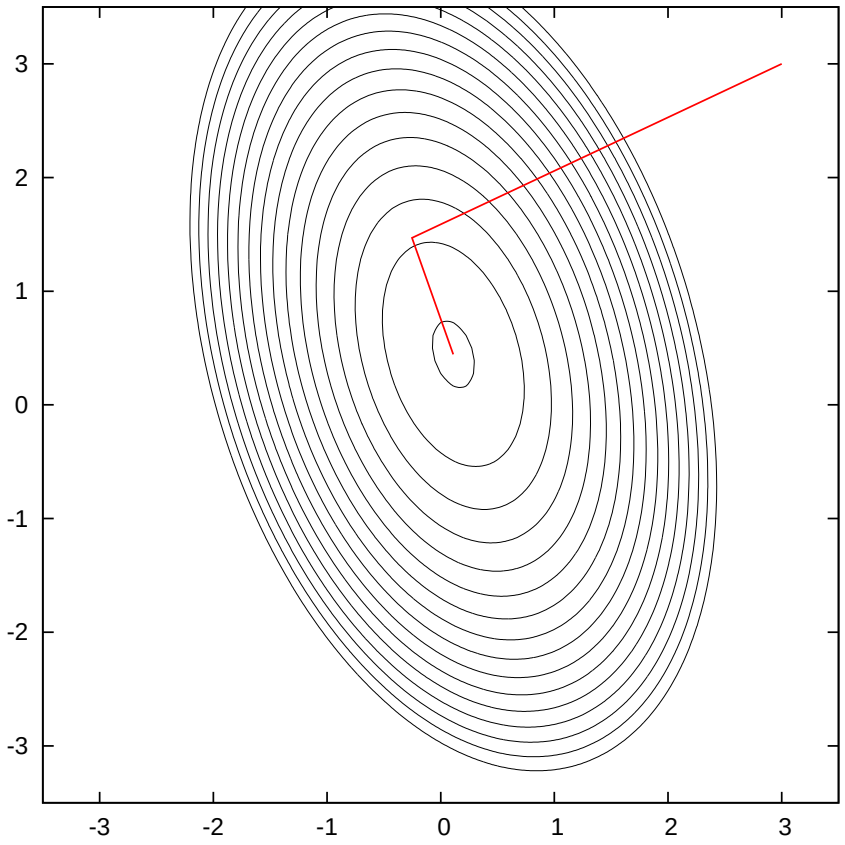
Twierdzenie ze strony 16 gwarantuje, że *formalnie* daje to to samo, co algebraiczna metoda gradientów sprzężonych.

Zamiast używać równania (13), można skorzystać z

$$\beta = \frac{\mathbf{r}_{k+1}^T(\mathbf{r}_{k+1} - \mathbf{r}_k)}{\mathbf{r}_k^T \mathbf{r}_k}\,. \tag{14}$$

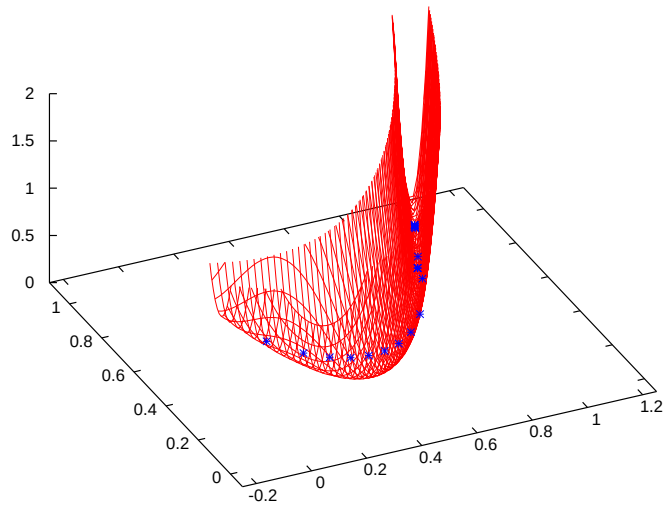
Jeżeli funkcja f ma *ściśle* postać (6), nie ma to znaczenia, gdyż $\mathbf{r}_{k+1}^T \mathbf{r}_k = 0$. Ponieważ jednak f jest tylko w przybliżeniu formą kwadratową, (14) może przyspieszyć obliczenia, gdy grozi stagnacja.

Przykład — metoda gradientów sprzężonych



Tylko dwa kroki! Drugi krok nie jest prostopadły do pierwszego, ale jest z nim sprzężony.

Przykład — funkcja Rosenbrocka



Zastosowanie algorytmu gradientów sprzężonych do minimalizacji funkcji Rosenbrocka. Widać *znaczne* przyspieszenie (mniej punktów pośrednich!) w stosunku do przedstawianej powyżej metody najszybszego spadku.

Uwaga o metodach gradientowych

Analizując metody gradientowe, (podświadomie) myślimy o funkcjach zachowujących się "porządnie". W praktyce jednak często mamy do czynienia z funkcjami, których gradienty są bardzo duże, jeśli chodzi o normę (długość).

Powróćmy do przykładu funkcji Rosenbrocka ze strony 11. Przypuśćmy, że chcemy zminimalizować tę funkcję starując z punktu $(4, 1)$. Wartość funkcji Rosenbrocka w tym punkcie wynosi $f(4, 1) = 22\,509$, a jej pochodne cząstkowe

$$\left.\frac{\partial f}{\partial x}\right|_{(4,1)} = \left[400x^3 - 400xy + 2x - 2\right]_{(4,1)} = 24\,006 \quad (15a)$$

$$\left.\frac{\partial f}{\partial y}\right|_{(4,1)} = \left[200y - 200x^2\right]_{(4,1)} = -3\,000 \quad (15b)$$

a wobec tego minus gradient

$$\mathbf{p} = -\nabla f|_{(4,1)} = [-24\,006, 3\,000] \tag{15c}$$

Jednowymiarowa funkcja (1), którą w tym wypadku minimalizujemy, ma postać

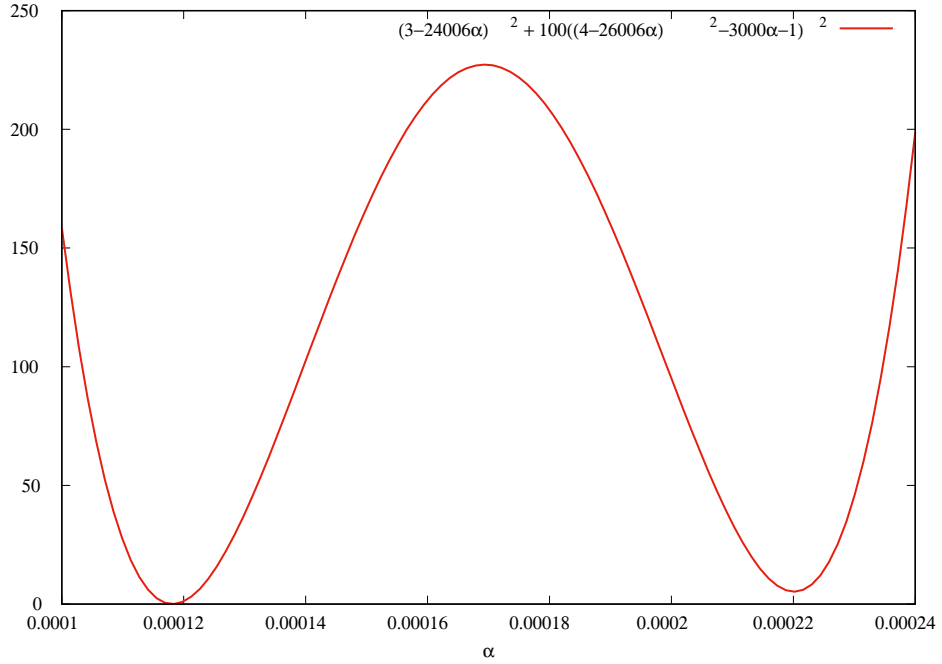
$$g(\alpha) = (3 - 24\,006\alpha)^2 + 100\left((4 - 24\,006\alpha)^2 - 3\,000\alpha - 1\right)^2 \tag{15d}$$

Jest to wielomian czwartego stopnia, o współczynniku wiodącym $100 \cdot (24\,006)^4 \simeq 3.321 \cdot 10^{19}$. Jej przykładowe wartości wynoszą $g(-1) \simeq 3.323 \cdot 10^{20}$, $g(0) = f(4,1) = 22\,505$, $g(1) \simeq 3.319 \cdot 10^{20}$, a więc, aby znaleźć minimum (jedno z dwu minimów) funkcji (15d), należy poruszać się o bardzo małe ułamki gradientu (bardzo małe α).

Być może *w tym przypadku* jako kierunek poszukiwań wygodniejsze byłoby wzięcie znormalizowanego gradientu

$$\mathbf{p} = -\frac{\nabla f|_{(4,1)}}{\left\| \nabla f|_{(4,1)} \right\|} \simeq [-0.992270, 0.124003] \tag{15e}$$

ale w licznych innych przypadkach mogłoby to stanowić utrudnienie.



Funkcja (15d) ma *dwa* blisko położone minima. Z praktycznego punktu widzenia nie ma znaczenia, które z tych minimów osiągniemy w procedurze minimalizacji dwuwymiarowej funkcji Rosenbrocka.

Metoda zmiennej metryki

Powróćmy do wspominanego na poprzednim wykładzie zastosowania metody Newtona. Jeżeli hessjan jest dodatnio określony, krok Newtona prowadzi do spadku wartości funkcji moglibyśmy więc użyć metody Newtona do precyzyjnego określania położenia minimum, gdy już znajdziemy się w jego otoczeniu. Dla funkcji **nie** będących w dobrym przybliżeniu formą kwadratową (6), może to, mimo dodatniej określoności hessjanu, oznaczać konieczność wykonania *wielu* drobnych kroków w końcowej fazie minimalizacji. Praktyczną trudność może stanowić konieczność wielokrotnego obliczania hessjanu, jeżeli obliczanie drugich pochodnych cząstkowych jest kosztowne lub kłopotliwe, w skrajnych przypadkach wręcz niemożliwe.

W takim wypadku warto jest rozważyć skorzystanie z metody zmiennej metryki: Zamiast korzystać z prawdziwego hessjanu obliczanego w każdym kroku, tworzymy ciąg dodatnio określonych przybliżeń hessjanu. Startujemy z jakiegoś $\mathbf{x}_0$. Jako początkowe przybliżenie hessjanu $\mathbf{H}_0$ bierzemy jakąś sensowną macierz symetryczną, dodatnio określoną — jeśli nie mamy lepszego pomysłu, może to być macierz jednostkowa, natomiast jeśli do punktu x_0 doszliśmy korzystając z metody Levenberga-Marquarda, naturalne (i zalecane) jest przyjęcie $\mathbf{H}_0 = \widetilde{\mathbf{H}}$, gdzie ta ostatnia macierz to przybliżony hessjan obliczony w ostatnim kroku metody Levenberga-Marquardta. Następnie wykonujemy iterację:

1. Rozwiązujemy równanie $\mathbf{H}_k \mathbf{p}_k = -\nabla f|_{\mathbf{x}_k}$.
2. Przeprowadzamy minimalizację kierunkową funkcji jednowymiarowej $f(\mathbf{x}_k + \alpha \mathbf{p}_k)$. Niech minimum kierunkowemu odpowiada wartość α_k.
3. $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$.
4. Wyliczamy $\mathbf{y}_k = \nabla f|_{\mathbf{x}_{k+1}} - \nabla f|_{\mathbf{x}_k}$.
5. Wyliczamy nowe przybliżenie $\mathbf{H}_{k+1}$, *na przykład* za pomocą formuły BFGS:

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{\mathbf{y}_k \mathbf{y}_k^T}{\alpha_k \mathbf{p}_k^T \mathbf{y}_k} - \frac{\nabla f|_{\mathbf{x}_k} \left(\nabla f|_{\mathbf{x}_k} \right)^T}{\mathbf{p}_k^T \mathbf{H}_k \mathbf{p}_k} \,. \tag{16}$$

Kończymy gdy $\left\| \mathbf{x}_{k+1} - \mathbf{x}_k \right\| < \varepsilon$, gdzie ε jest zadaną tolerancją.

Własności metody zmiennej metryki

Można pokazać, że jeżeli $\mathbf{H}_0$ jest symetryczna i dodatnio określona, także następne $\mathbf{H}_k$ mają te własności.

Dodatkowo, jeśli f jest formą kwadratową (6), $\mathbf{H}_{i \geqslant N+1} \equiv \mathbf{H}_{\text{min}}$ w arytmetyce dokładnej. W ogólności ciąg $\mathbf{H}_k$ *nie musi* być zbieżny do "prawdziwego" hessjanu, nawet jeśli metoda daje zadowalające numerycznie przybliżenie minimum.

W pewnym uproszczeniu można powiedzieć, że metoda zmiennej metryki ma się do metody gradientów sprzężonych tak, jak metoda Broydena ma się do wielowymiarowej metody Newtona (litera "B" w akronimie BFGS oznacza właśnie Broydena).

Metoda Powella

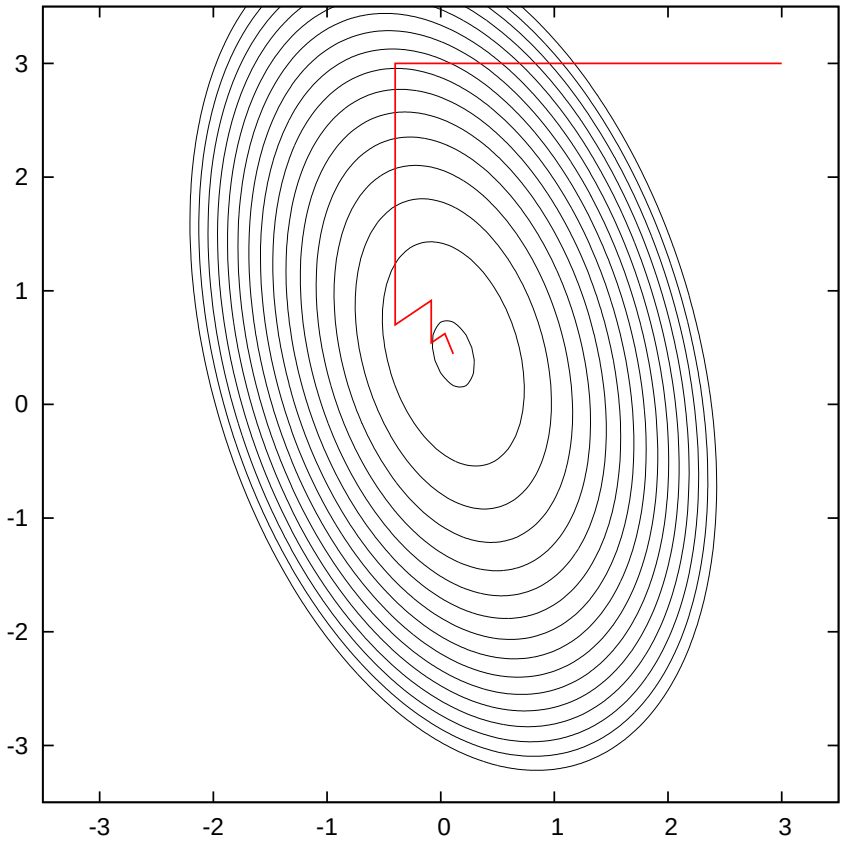
Wszystkie omówione wyżej metody wymagały obliczania pochodnych (gradientu) funkcji $f(\mathbf{x})$. Co jednak zrobić, jeśli obliczenie pochodnych jest kosztowne, niemożliwe lub funkcja jest nieróżniczkowalna? Zasadnicza strategia postępowania — minimalizacja kierunkowa, wybór nowego kierunku etc — pozostaje w mocy, zmienia się tylko sposób wyboru kolejnych kierunków. Metoda Powella polega na konstrukcji kierunków, które z czasem, po wielu iteracjach, stają się sprzężone.

Inicjalizujemy biorąc $\{\mathbf{p}_i\}_{i=1}^N = \{\mathbf{e}_i\}_{i=1}^N$, gdzie $\{\mathbf{e}_i\}_{i=1}^N$ są kolejnymi wersorami (innymi słowy, zaczynamy od minimalizacji po współrzędnych). Następnie

1. Znajdujemy się w pewnym punkcie $\mathbf{X}_0$.
2. Minimalizujemy wzdłuż kolejnych kierunków $\mathbf{p}_i$, osiągając kolejno punkty $\mathbf{X}_i$.
3. Dla $i = 1, \ldots, N-1$: $\mathbf{p}_i = \mathbf{p}_{i+1}$.
4. $\mathbf{p}_N = \mathbf{X}_N - \mathbf{X}_0$.
5. Minimalizujemy wzdłuż (nowego) $\mathbf{p}_N$, oznaczając znaleziony punkt przez $\mathbf{X}_0$.
6. `GOTO 1`

Jeśli badana funkcja jest rozwijalna w szereg Taylora wokół minimum, po N iteracjach powyższej procedury kierunki $\{\mathbf{p}_i\}_{i=1}^N$ stają się sprzężone.

Przykład — metoda Powella



Mniej kroków niż w minimalizacji po współrzędnych. W większej liczbie wymiarów byłoby *jeszcze lepiej*.

Minima wyższych rzędów

W powyższych rozważaniach zakładaliśmy, że w pobliżu minimum funckcja zachowuje się w przybliżeniu jak forma kwadratowa. Co jednak, gdy minimum jest wyższego rzędu, to znaczy gdy najniższy nieznikający rząd rozwinięcia Taylora jest rzędu czwartego (szóstego, ósmego, ...), jak w przykładzie funkcji $f(x,y) = x^4 + y^4$. Takie sytuacje *zdarzają* się w praktyce, na tyle jednak rzadko, że nie warto dla nich opracowywać osobnych metod. Korzystamy wówczas z metod już poznanych, właściwych dla minimów rzędu drugiego. Należy jednak mieć świadomość, że zbieżność w procedurze precyzyjnej lokalizacji minimum będzie wolniejsza, samo zaś minimum będziemy mogli wyznaczyć mniej dokładnie. Wynika to z faktu, że w okolicach minimum wyższego rzędu funkcja jest bardziej "płaska".