

Wstęp do metod numerycznych

Minimalizacja funkcji wielu zmiennych

P. F. Góra

<http://th-www.if.uj.edu.pl/zfs/gora/>

2011

Strategia minimalizacji wielowymiarowej

Zakładamy, że metody poszukiwania minimów lokalnych funkcji jednej zmiennej są znane.

Rozważmy funkcję $f: \mathbb{R}^N \rightarrow \mathbb{R}$. Przedstawimy strategię poszukiwania (lokalnego) minimum tej funkcji w postaci ciągu minimalizacji jednowymiarowych.

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

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

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.

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_{\min} \mathbf{p}_k. \quad (2)$$

6. *goto* 1.

Jak wybierać kierunki poszukiwań?

Cały problem sprowadza się zatem do wyboru odpowiedniej sekwencji kolejnych kierunków $\{p_k\}$. Bardzo popularne są metody

- 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 punkcie startowym.

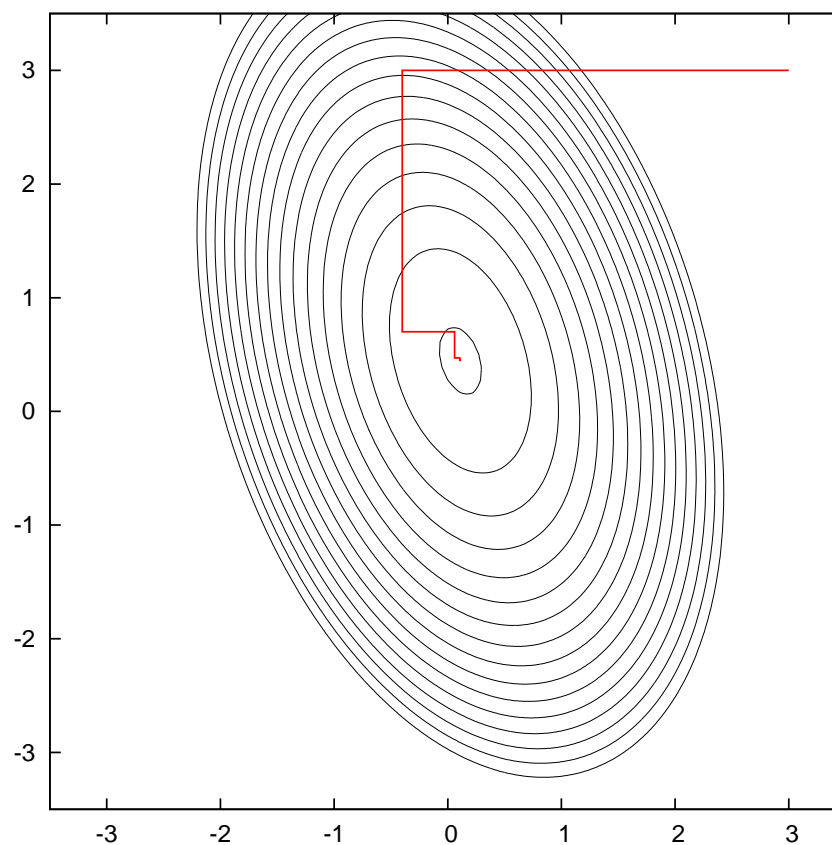
Jeśli jesteśmy blisko minimum nie są to dobre pomysły, gdyż prowadzą do wielu drobnych kroków, które częściowo likwidują efekty osiągnięte w krokach poprzednich. Dlaczego?

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

$$\frac{dg_k}{d\alpha} = \sum_i \frac{\partial f}{\partial x_i} (\mathbf{p}_k)_i = \left(\nabla f|_{\mathbf{x}=\mathbf{x}_{\min}} \right)^T \mathbf{p}_k = 0. \quad (3)$$

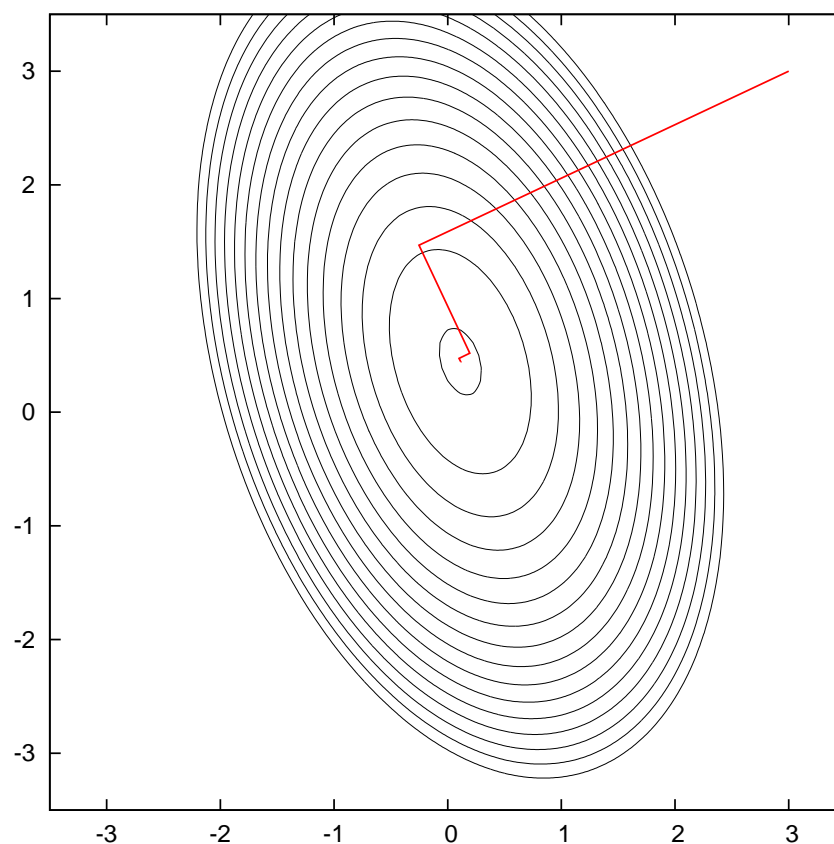
W minimum kierunkowym gradient funkcji jest prostopadły do kierunku poszukiwań. Widać stąd natychmiast, że krok minimalizacji w metodzie najszybszego spadku co prawda zaczyna się prostopadle do poziomnic funkcji, ale kończy się *stycznie* do poziomnic. Dodatkowo w minimalizacji po współrzędnych kolejne kierunki poszukiwań, czyli — tutaj — kolejne współrzędne, nie zależą od kształtu minimalizowanej funkcji; taka strategia nie może być optymalna.

Przykład — minimalizacja po współrzędnych



Dużo małych kroków

Przykład — metoda najszybszego spadku



Wygląda lepiej, ale kroków prawie tyle samo

Przybliżenie formy kwadratowej

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

$$f(\mathbf{x}) \simeq \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x} + f_0, \quad (4)$$

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 (4) gradient funkcji f w punkcie \mathbf{x}_k wynosi

$$\nabla f|_{\mathbf{x}=\mathbf{x}_k} = \mathbf{A}\mathbf{x}_k - \mathbf{b}. \quad (5)$$

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

$$\nabla f|_{\mathbf{x}} = \mathbf{A}\mathbf{x}_k + \alpha\mathbf{A}\mathbf{p}_{k+1} - \mathbf{b}. \quad (6)$$

Zmiana gradientu wynosi

$$\delta(\nabla f) = \alpha\mathbf{A}\mathbf{p}_{k+1}. \quad (7)$$

Punkt \mathbf{x}_k jest minimum kierunkowym w kierunku \mathbf{p}_k , a więc gradient funkcji w tym punkcie spełnia warunek (3). *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

$$\mathbf{p}_i^T \mathbf{A} \mathbf{p}_j = 0, \quad i > j. \quad (8)$$

Metodę wybierania kierunków poszukiwań spełniających (8) 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 konstruowania 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 (4) 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 (5), $\mathbf{r}_k = -\mathbf{A}\mathbf{x}_k + \mathbf{b}$ oraz

$$\mathbf{r}_{k+1} = -\mathbf{A}(\mathbf{x}_k + \alpha \mathbf{p}_k) + \mathbf{b} = \mathbf{r}_k - \alpha \mathbf{A}\mathbf{p}_k. \quad (9)$$

W minimum kierunkowym $\mathbf{p}_k^T \nabla f|_{\mathbf{x}_{k+1}} = -\mathbf{p}_k^T \mathbf{r}_{k+1} = 0$ (por. (3)). Wobec tego mnożąc równanie (9) lewostronnie przez \mathbf{p}_k^T , otrzymujemy

$$\alpha = \frac{\mathbf{p}_k^T \mathbf{r}_k}{\mathbf{p}_k^T \mathbf{A}\mathbf{p}_k}. \quad (10)$$

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)

$$\beta = \frac{\mathbf{r}_{k+1}^T \mathbf{r}_{k+1}}{\mathbf{r}_k^T \mathbf{r}_k}. \quad (11)$$

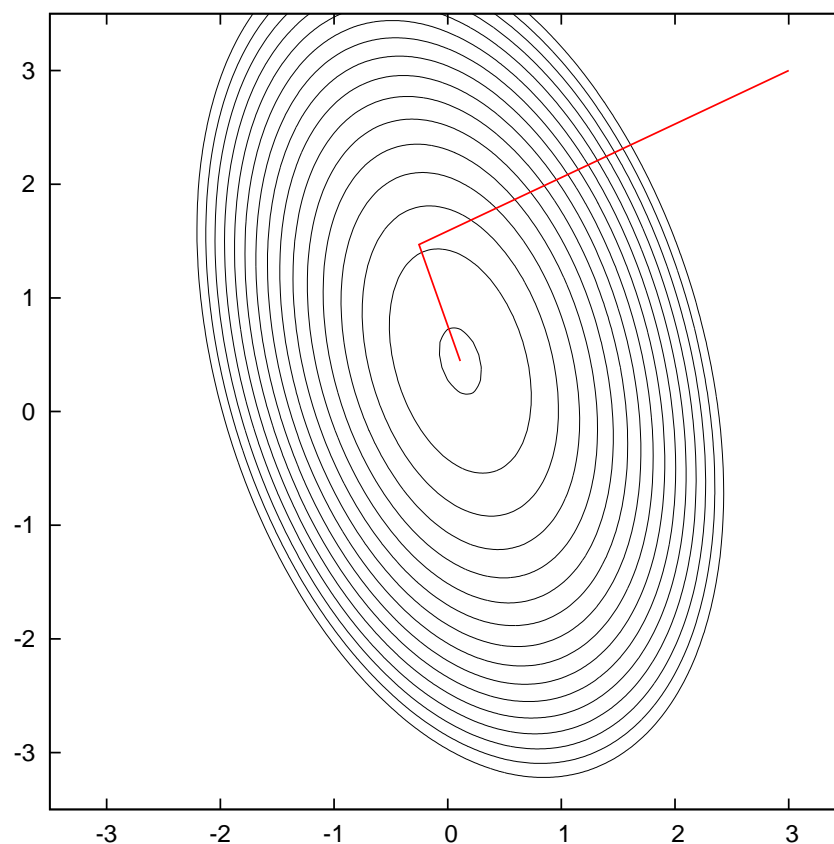
4. Obliczamy (jak w metodzie algebraicznej) $\mathbf{p}_{k+1} = \mathbf{r}_{k+1} + \beta \mathbf{p}_k$.

Zamiast używać równania (11), można skorzystać z

$$\beta = \frac{\mathbf{r}_{k+1}^T (\mathbf{r}_{k+1} - \mathbf{r}_k)}{\mathbf{r}_k^T \mathbf{r}_k}. \quad (12)$$

Jeżeli funkcja f ma *ściśle* postać (4), nie ma to znaczenia, gdyż $\mathbf{r}_{k+1}^T \mathbf{r}_k = 0$. Ponieważ jednak f jest tylko w przybliżeniu formą kwadratową, (12) 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.

Metoda zmiennej metryki

Formalna konieczność rozwiązywania równania $\nabla f = 0$ sugeruje zastosowanie następującego algorytmu opartego na metodzie Newtona:

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \mathbf{H}^{-1} \nabla f|_{\mathbf{x}_i}, \quad (13)$$

gdzie \mathbf{H} oznacza macierz drugich pochodnych cząstkowych funkcji f wyliczanych w punkcie \mathbf{x}_i (aktualnym).

W pobliżu minimum f ma w przybliżeniu postać (4). Wówczas położenie minimum możemy znaleźć formalnie w jednym kroku jako

$$\mathbf{x}_{\min} - \mathbf{x}_i = -\mathbf{A}^{-1} \nabla f|_{\mathbf{x}_i}, \quad (14)$$

gdzie \mathbf{A} jest hessjanem w minimum. Zauważmy, że aby krok Newtona w istocie prowadził do zmniejszenia się wartości funkcji, musi zachodzić $(\mathbf{x} - \mathbf{x}_i)^T \nabla f|_{\mathbf{x}_i} < 0$, czyli

$$-(\mathbf{x} - \mathbf{x}_i)^T \mathbf{A}(\mathbf{x} - \mathbf{x}_i) < 0 \quad (15)$$

co jest spełnione, jeżeli \mathbf{A} jest dodatnio określone.

Wreszcie jeśli $\mathbf{x}_i, \mathbf{x}_{i+1}$ są kolejnymi krokami iteracji prowadzącej do minimalizacji, odejmując odpowiednie równania (14) od siebie dostajemy

$$\mathbf{x}_{i+1} - \mathbf{x}_i = \mathbf{A}^{-1} \left(\nabla f|_{\mathbf{x}_{i+1}} - \nabla f|_{\mathbf{x}_i} \right). \quad (16)$$

Ale my wciąż nie znamy \mathbf{A} .

Celem metody zmiennej metryki jest wobec tego skonstruowanie pewnego *ciągu przybliżeń hessjanu* — zamiast równania (16) mamy

$$\mathbf{x}_{i+1} - \mathbf{x}_i = \mathbf{H}_{i+1} \left(\nabla f|_{\mathbf{x}_{i+1}} - \nabla f|_{\mathbf{x}_i} \right). \quad (17)$$

W powyższym równaniu \mathbf{H}_i oznaczają kolejne przybliżenia hessjanu w minimum. Przybliżenia te muszą spełniać

1. \mathbf{H}_i jest symetryczne i dodatnio określone.
2. Zachodzi

$$\lim_{i \rightarrow \infty} \mathbf{H}_i = \mathbf{A}^{-1}. \quad (18)$$

Jeśli nie mamy żadnego lepszego pomysłu, możemy przyjąć $\mathbf{H}_1 = \mathbb{I}$.
Metoda ma formalnie postać

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \mathbf{H}_i \nabla f|_{\mathbf{x}_i}. \quad (19)$$

Konstruowanie kolejnych \mathbf{H}_i

DFP updating formula:

$$\mathbf{H}_{i+1} = \mathbf{H}_i + \frac{(\mathbf{x}_{i+1} - \mathbf{x}_i)(\mathbf{x}_{i+1} - \mathbf{x}_i)^T}{(\mathbf{x}_{i+1} - \mathbf{x}_i)^T (\nabla f|_{\mathbf{x}_{i+1}} - \nabla f|_{\mathbf{x}_i})} - \frac{\left[\mathbf{H}_i (\nabla f|_{\mathbf{x}_{i+1}} - \nabla f|_{\mathbf{x}_i}) \right] \left[\mathbf{H}_i (\nabla f|_{\mathbf{x}_{i+1}} - \nabla f|_{\mathbf{x}_i}) \right]^T}{(\nabla f|_{\mathbf{x}_{i+1}} - \nabla f|_{\mathbf{x}_i})^T \mathbf{H}_i (\nabla f|_{\mathbf{x}_{i+1}} - \nabla f|_{\mathbf{x}_i})}. \quad (20)$$

BFGS updating formula:

$$\mathbf{H}_{i+1} = \cdots + \left(\nabla f|_{\mathbf{x}_{i+1}} - \nabla f|_{\mathbf{x}_i} \right)^T \mathbf{H}_i \left(\nabla f|_{\mathbf{x}_{i+1}} - \nabla f|_{\mathbf{x}_i} \right) \mathbf{u} \mathbf{u}^T, \quad (21a)$$

$$\mathbf{u} = \frac{\mathbf{x}_{i+1} - \mathbf{x}_i}{(\mathbf{x}_{i+1} - \mathbf{x}_i)^T \left(\nabla f|_{\mathbf{x}_{i+1}} - \nabla f|_{\mathbf{x}_i} \right)} - \frac{\mathbf{H}_i \left(\nabla f|_{\mathbf{x}_{i+1}} - \nabla f|_{\mathbf{x}_i} \right)}{\left(\nabla f|_{\mathbf{x}_{i+1}} - \nabla f|_{\mathbf{x}_i} \right)^T \mathbf{H}_i \left(\nabla f|_{\mathbf{x}_{i+1}} - \nabla f|_{\mathbf{x}_i} \right)}, \quad (21b)$$

gdzie kropki oznaczają całą formułę DFP.

Własności metody zmiennej metryki

Można pokazać, że \mathbf{H}_i modyfikowane tak według formuły DFP, jak i BFGS

1. Jest symetryczne i dodatnio określone, jeśli tylko \mathbf{H}_1 jest symetryczne i dodatnio określone.
2. Spełnia warunek (17).
3. Spełnia (18).
4. Dodatkowo, jeśli f jest formą kwadratową (4), $\mathbf{H}_{i \geq N+1} \equiv \mathbf{A}^{-1}$ w arytmetyce dokładnej.

Metoda Levenberga-Marquardta

Tak metoda gradientów sprzężonych, jak i metoda zmiennej metryki, są dostosowane do przypadku, w którym funkcja jest z dobrym przybliżeniem postaci (4), a więc gdy jesteśmy dostatecznie blisko poszukiwanego minimum. Jednak daleko od minimum metody te są powolne — trudno oczekiwać, że wzory słuszne dla formy kwadratowej będą dobrze działać gdy funkcja formą kwadratową *nie* jest. Daleko od minimum jest sens stosować metodę najszybszego spadku. Powinniśmy zatem mieć metodę, która daleko od minimum zachowuje się jak najszybszy spadek, blisko zaś minimum redukuje się do zmiennej metryki (lub gradientów sprzężonych).

Powróćmy do “algorytmu” opartego na metodzie Newtona (13):

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \mathbf{H}^{-1}(\mathbf{x}_i) \nabla f|_{\mathbf{x}_i} ,$$

Daleko od minimum hessjan nie musi być nawet dodatnio określony, co powoduje, iż krok newtonowski wcale nie musi prowadzić do spadku wartości funkcji (por. (15)). My jednak chcemy aby wartość funkcji w kolejnych krokach spadała. Zmodyfikujmy więc hessjan:

$$\widetilde{\mathbf{H}}_{ii} = (1 + \lambda) \frac{\partial^2 f}{\partial x_i^2} , \quad (22a)$$

$$\widetilde{\mathbf{H}}_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j} , \quad i \neq j , \quad (22b)$$

przy czym $\lambda \geq 0$.

Zauważmy, że zachodzi jedna z dwu możliwych sytuacji: (i) jeśli znajdujemy się w basenie atrakcji minimum, wówczas dla odpowiednio dużego λ macierz (22) stanie się dodatnio określona lub też (ii) jeśli dla żadnego dodatniego λ macierz (22) nie staje się dodatnio określona, znajdujemy się na monotonicznej gałęzi funkcji, poza basenem atrakcji minimum.

Rozpoczynamy z jakimś niewielkim λ , na przykład $\lambda = \lambda_0 = 2^{-10} = 1/1024$. Przypuśćmy, iż aktualnym przybliżeniem minimum jest punkt x_i . Dostajemy zatem...

Algorytm Levenberga-Marquardta

1. Oblicz $\nabla f(\mathbf{x}_i)$.
2. Oblicz $\widetilde{\mathbf{H}}(\mathbf{x}_i)$.
3. Oblicz

$$\mathbf{x}_{\text{test}} = \mathbf{x}_i - \widetilde{\mathbf{H}}^{-1}(\mathbf{x}_i) \nabla f(\mathbf{x}_i). \quad (23)$$

4. Jeżeli $f(\mathbf{x}_{\text{test}}) > f(\mathbf{x}_i)$, to
 - (a) $\lambda \rightarrow 8\lambda$ (można też powiększać o inny znaczny czynnik).
 - (b) Idź do punktu 2.
5. Jeżeli $f(\mathbf{x}_{\text{test}}) < f(\mathbf{x}_i)$, to
 - (a) $\lambda \rightarrow \lambda/8$ (można też zmniejszać o inny znaczny czynnik).
 - (b) $\mathbf{x}_{i+1} = \mathbf{x}_{\text{test}}$.
 - (c) Idź do punktu 1.

Komentarz

Dodatkowo, jeśli $\lambda > \lambda_{\max} \gg 1$, uznajemy, iż znajdujemy się poza basenem atrakcji minimum i algorytm zawodzi. Jeśli natomiast $\lambda < \lambda_{\min} \ll 1$, macierz $\widetilde{\mathbf{H}}$ jest w praktyce równa hessjanowi, a zatem modyfikacja (22) przestaje być potrzebna. Możemy wówczas przenieść się na metodę gradientów sprzężonych lub metodę zmiennej metryki aby wykorzystać ich szybką zbieżność w pobliżu minimum, gdzie funkcja ma postać (4).

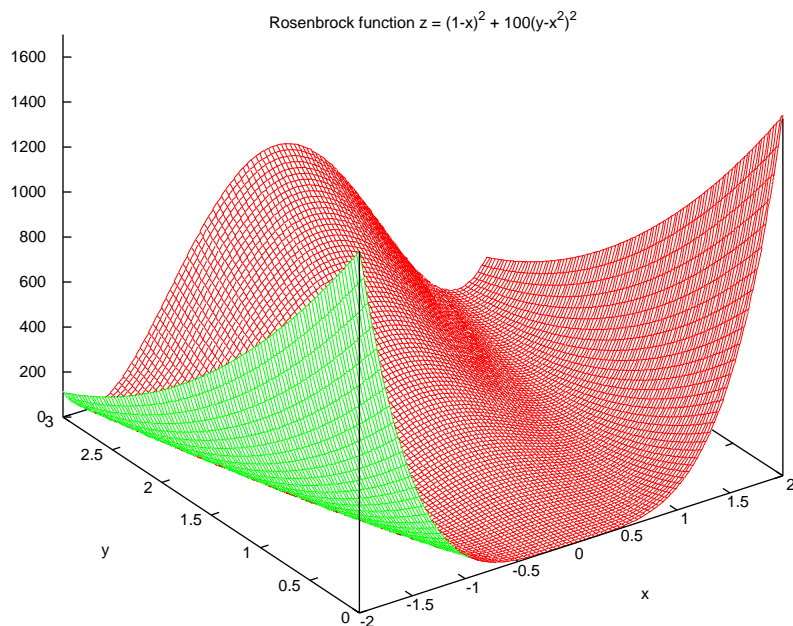
Ponadto w celu przyspieszenia obliczeń, jeżeli $f(\mathbf{x}_{\text{test}}) < f(\mathbf{x}_i)$, możemy *chwilowo* zrezygnować ze zmniejszania λ i modyfikowania $\widetilde{\mathbf{H}}$ i przeprowadzić kilka kroków z tą samą macierzą, a więc korzystając z tej samej faktoryzacji.

Zauważmy, iż przypadek $\lambda \gg 1$ oznacza, iż jesteśmy daleko od minimum. Z drugiej strony jeśli $\lambda \gg 1$, macierz $\widetilde{\mathbf{H}}$ staje się w praktyce diagonalna, a zatem

$$\begin{aligned} \mathbf{x}_{\text{test}} &\simeq \mathbf{x}_i - (1 + \lambda)^{-1} \text{diag} \left\{ \left(\frac{\partial^2 f}{\partial x_1^2} \right)^{-1}, \left(\frac{\partial^2 f}{\partial x_2^2} \right)^{-1}, \dots, \left(\frac{\partial^2 f}{\partial x_N^2} \right)^{-1} \right\} \nabla f(\mathbf{x}_i) \\ &\simeq \mathbf{x}_i - \text{const} \nabla f(\mathbf{x}_i), \end{aligned} \quad (24)$$

o ile drugie pochodne cząstkowe w poszczególnych kierunkach nie różnią się znacznie od siebie. Widać, iż daleko od minimum, gdzie warunek zachowujący raz osiągnięte minima kierunkowe nie zachodzi, algorytm Levenberga-Marquardta zachowuje się prawie jak metoda najszybszego spadku.

Przykład — funkcja Rosenbrocka



Funkcja Rosenbrocka $f(x, y) = (1 - x)^2 + 100(y - x^2)^2$ jest przykładem funkcji, w której daleko od minimum metody gradientów sprzężonych i zmiennej metryki nie zapewniają szybkiej zbieżności.

Metoda Powella

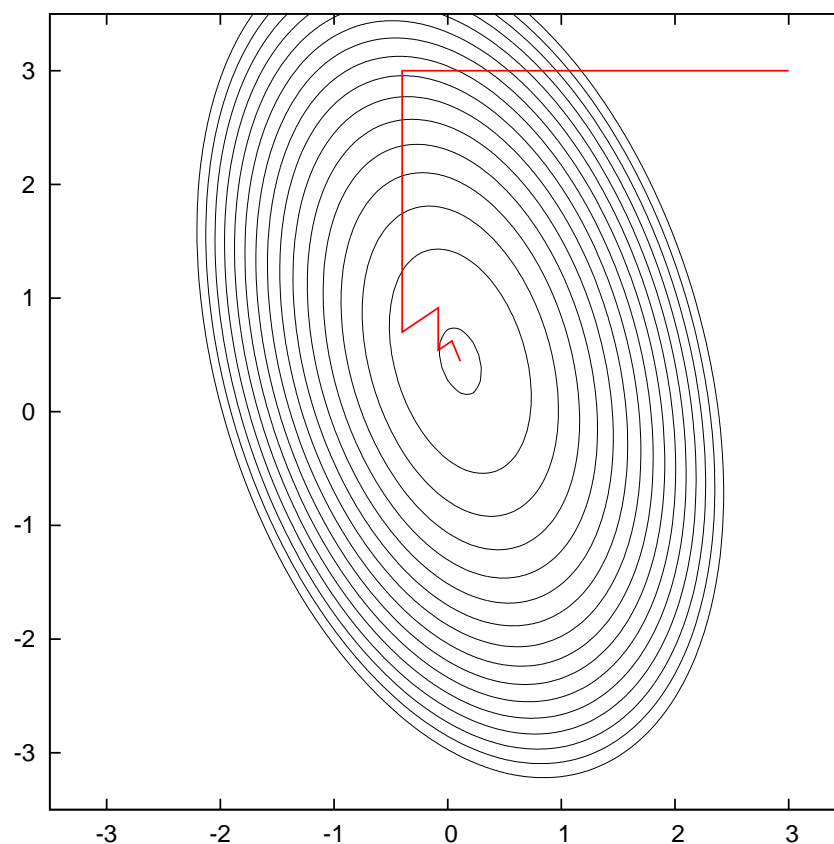
Wszystkie omówione wyżej metody wymagały obliczania pochodnych (gradientu) funkcji $f(\mathbf{x})$. Co jednak zrobić, jeśli obliczenie pochodnych jest kosztowne, niemożliwe lub funkcja jest nieróżniczkowalna? Zasadnicza strategia postępowania — minimalizacja kierunkowa, wybór nowego kierunku etc — pozostaje 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 wer-
sorami (innymi słowy, zaczynamy od minimalizacji po współrzędnych). Na-
stępnie

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 $i = 1, \dots, N - 1$: $\mathbf{p}_i = \mathbf{p}_{i+1}$.
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 Powell



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