ROZDZIAŁ 3

Metody minimalizacji różniczkowalnej bez ograniczeń

Gdy wszyscy myślą jednakowo, nikt nie myśli naprawdę.

Platon

3.1 Ogólna postać metod minimalizacji

Przejdziemy teraz do głównego przedmiotu programowania matematycznego, czyli do metod wyznaczania minimum funkcji. Będzie przy tym chodziło zarówno o wyznaczenie wartości minimalnej funkcji celu jak i o wyznaczenie jej minimizera. Podobnie jak przy wyznaczaniu miejsc zerowych funkcji, w ogólnym przypadku nie możemy oczekiwać wzoru na to minimum. Nie możemy również oczekiwać tego, że uda nam się wyliczyć dokładną wartość minimalną i dokładny minimizer. Najczęściej będziemy musieli zadowolić się pewnymi ich przybliżeniami. Niezależnie od tego, jak dobre mają być te przybliżenia, istotne jest to, aby otrzymać je w wyniku skończonego ciągu operacji arytmetycznych. Konstrukcje tych właśnie ciągów operacji arytmetycznych powstają w oparciu o tzw. *metody minimalizacji.* Mają one najczęściej postać pewnych procedur podanych w postaci schematów iteracyjnych lub inaczej algorytmów generujących ciągi kolejnych przybliżeń rozwiązania.

Definicja 3.1.1 Pod pojęciem *algorytmu* rozumiemy proces iteracyjny utworzony przez ciąg tzw. *odwzorowań algorytmicznych* $\mathcal{A}_k : X \longrightarrow 2^X$, gdzie $X \subseteq \mathbb{R}^n$ i przez *obszar startowy* $A_1 \subseteq X$. Dla danego *punktu startowego* $x_1 \in A_1$ algorytm *generuje* ciąg kolejnych punktów $x_k \in X$ spełniających zależność rekurencyjną

$$x_{k+1} \in \mathcal{A}_k(x_k).$$

Jeśli $\mathcal{A}_k = \mathcal{A}$ dla $k = 1, 2, ...$, to mówimy, że ciąg odwzorowań algorytmicznych jest *autonomiczny*

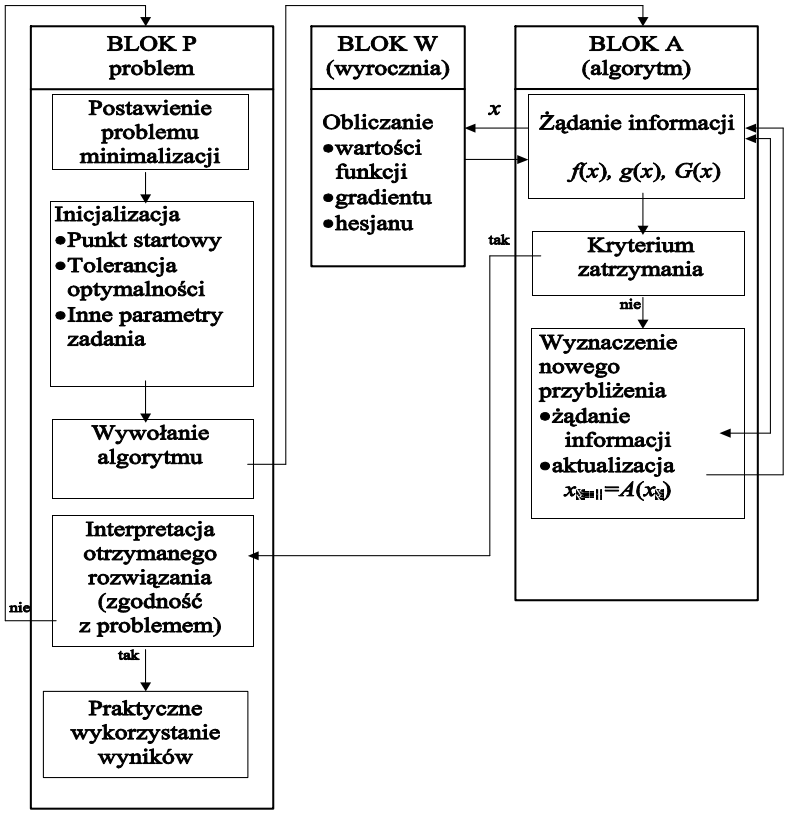
Uwaga 3.1.2

(a) W wielu podręcznikach definicja algorytmu jest węższa, niż podana wyżej (por. [BK82, KS92, SB87]). Przez algorytm rozumie się tam procedurę iteracyjną, która jednoznacznie określa ciąg punktów. W tej sytuacji odwzorowanie algorytmiczne jest funkcją $\mathcal{A}_k : X \longrightarrow X$ natomiast obszar startowy redukuje się do pojedynczego punktu startowego.

(b) Odwzorowanie algorytmiczne podane jest najczęściej w postaci wzoru lub skończonego ciągu wzorów. Wzory te podane są jako skończone ciągi operacji arytmetycznych. Może się jednak zdarzyć, że odwzorowanie algorytmiczne jest wynikiem nieskończonego ciągu takich operacji. Sytuacja taka powstaje na przykład wówczas, gdy we wzorze określającym to odwzorowanie figuruje funkcja zdefiniowana w postaci nieskończonego ciągu operacji arytmetycznych (np. funkcja sin) bądź pochodna funkcji. W praktycznych obliczeniach odwzorowanie algorytmiczne zastępuje w takich przypadkach się jego przybliżeniem, np. sumę nieskończonego szeregu zastępuje się jego sumą częściową, zaś pochodną zastępuje się pewnym ilorazem różnicowym.

(c) Algorytmy (metody) konstruuje się najczęściej w celu wygenerowania ciągu przybliżeń zbieżnego do rozwiązania określonego problemu. Jeśli zbieżność taka zachodzi dla każdego ciągu generowanego przez ten algorytm, to mówimy, że algorytm (metoda) jest zbieżny globalnie do rozwiązania. Może się jednak zdarzyć, że zbieżność do rozwiązania zachodzi dla pewnego zestawu parametrów określających odwzorowania algorytmiczne, bądź dla punktów startowych leżących odpowiednio blisko rozwiązania. W ostatnim przypadku mówimy, że algorytm (metoda) jest zbieżny lokalnie.

Metody minimalizacji różniczkowalnej mają swój specyficzny charakter. Przy ich konstrukcji zakłada się bowiem, że dla dowolnego punktu $x \in \mathbb{R}^n$ będącego pewnym przybliżeniem rozwiązania zadania minimalizacji funkcji f, dysponujemy procedurą, która jest w stanie wyliczyć dokładną wartość funkcji $f(x)$ i dokładny gradient $\nabla f(x)$ (czasem również dokładny hesjan $\nabla^2 f(x)$). Procedura ta nazywa się czasem *wyrocznią* (ang. *oracle*). Metody minimalizacji różniczkowalnej konstruowane są najczęściej zgodnie z następującym schematem.



Rysunek 1. Schemat metod minimalizacji różniczkowalnej

W dalszej części interesować nas będzie głównie blok A powyższego schematu. Niemniej jednak należy mieć świadomość w jaki sposób jest on wbudowany w całość zagadnienia optymalizacji. W szczególności, konstruując jakikolwiek algorytm minimalizacji, należy pamiętać o tym, że najbardziej kosztochłonnymi operacjami w całym schemacie są wyznaczenie wartości funkcji, gradientu i ewentualnie hesjanu dla zadanego przybliżenia rozwiązania problemu. Należy w związku z tym preferować metody, co do których można się spodziewać, że niezbyt często będą wymagać obliczenia tych wartości. Ponieważ często funkcja f nie jest podana *explicite*, mniejsze znaczenie mają metody, w których funkcja badana jest "analitycznie". W dalszej części zakładamy, że dysponujemy jedynie procedurą (wyrocznią), która dla zadanego argumentu x pozwala wyznaczyć dokładną wartość funkcji $f(x)$ i dokładny gradient $\nabla f(x)$ (lub ewentualnie dodatkowo hesjan $\nabla^2 f(x)$).

W metodach przedstawionych w dalszej części rozdziału wyznaczenie nowego przybliżenia x_{k+1} rozwiązania ($x_{k+1} \in \mathcal{A}_k(x_k)$) polega na:

1^0 określeniu *kierunku poszukiwań* s_k,

2^0 określenie *długości kroku* $t_k > 0$ - część tę nazywamy *minimalizacją kierunkową* (ang. *line search*)

3^0 wyznaczenie nowego przybliżenia x_{k+1} rozwiązania zgodnie ze wzorem

$$x_{k+1} = x_k + t_k \frac{s_k}{\|s_k\|}.$$

Po przeskalowaniu wielkości t_k otrzymamy wzór równoważny:

$$x_{k+1} = x_k + t_k s_k. \tag{3.1}$$

W dalszej części będziemy głównie stosować drugi z ostatnich dwóch wzorów, w którym t_k będziemy również nazywać – choć tutaj w sposób nie do końca uprawniony – długością kroku.

Wśród metod minimalizacji różniczkowalnej wyróżniamy *metody spadku*, tj. takie dla których $s_k^T \nabla f(x_k) < 0$. Korzystając z definicji różniczkowalności nietrudno pokazać, że w tym przypadku kierunek s_k jest kierunkiem spadku. Przypomnijmy, że funkcja $q : \mathbb{R} \to \mathbb{R}$ zdefiniowana równością $q(t) = f(x_k + ts_k)$ określa zachowanie funkcji f wzdłuż prostej o równaniu $x = x_k + ts_k$ (porównaj ustęp 1.5.2). Dla funkcji tej zachodzą równości więc $q'(t) = s_k^T \nabla f(x_k + ts_k)$ oraz $q''(t) = s_k^T \nabla^2 f(x_k + ts_k)s_k$, w szczególności $q'(0) = s_k^T \nabla f(x_k)$ oraz $q''(0) = s_k^T \nabla^2 f(x_k)s_k$.

Uwaga 3.1.3 Jeśli H jest macierzą dodatnio określona, to wektor

$$s = -H\nabla f(x)$$

jest kierunkiem spadku funkcji f w punkcie x o ile $\nabla f(x) \neq 0$. Fakt ten wynika prosto z definicji macierzy dodatnio określonej i z uwagi 1.5.26.

Metody minimalizacji kierunkowej dzielimy na:

- metody *dokładnej minimalizacji kierunkowej*, tj. takie w których długość kroku t_k wyznacza się zgodnie ze wzorem

  $$t_k = \arg\min\{f(x_k + ts_k) : t > 0\},$$

  o ile oczywiście takie minimum istnieje.

- metody *przybliżonej minimalizacji kierunkowej*. Celem tych metod jest nie tyle wyznaczenie przybliżonego minimum funkcji f w kierunku s_k, co zapewnienie wystarczającego spadku funkcji w tym kierunku. Wśród metod tych wyróżniamy metody bezgradientowe (m.in. bisekcji, złotego podziału, ciągów Fibonacci'ego) i gradientowe, w których stosowane są między innymi bisekcja, interpolacja kwadratowa, interpolacji sześcienna, ekstrapolacja kwadratowa. Metody przybliżonej minimalizacji kierunkowej będą omówione dokładniej w kolejnym ustępie.

Uwaga 3.1.4 Jeśli długość kroku t_k została wyznaczona metodą dokładnej minimalizacji kierunkowej, to $q'(t_k) = 0$, czyli $s_k^T \nabla f(x_{k+1}) = 0$.

Uwaga 3.1.5 Ponieważ funkcja celu może przyjmować skomplikowaną postać, stosowanie dokładnej minimalizacji kierunkowej jest ograniczone. Jedynie dla nielicznych funkcji celu można podać dokładny wzór na długość kroku realizującego dokładne minimum w zadanym kierunku $s \in \mathbb{R}^n$. Ważnym przykładem takiej funkcji jest funkcja kwadratowa $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \frac{1}{2}x^T Ax + b^T x + c$, gdzie A jest macierzą symetryczną. Przyjmując oznaczenie $q(t) = f(\bar{x} + ts)$ otrzymamy wówczas $q'(t) = s^T \nabla f(\bar{x} + ts) = s^T(A(\bar{x} + ts) + b)$ i $q''(t) = s^T \nabla^2 f(\bar{x} + ts)s = s^T As$. Jeśli więc $s^T As > 0$, co ma miejsce na przykład dla dodatnio określonej macierzy A , to nietrudno zauważyć, że dla

$$\tau = -\frac{s^T(A\bar{x} + b)}{s^T As}$$

mamy $q'(\tau) = 0$ i $q''(\tau) > 0$. Ponieważ q jest funkcją kwadratową, więc dla dodatnio określonej macierzy A osiąga o na swoje minimum globalne w punkcie τ (q jest wówczas funkcją mocno wypukłą). Jeśli ponadto $q'(0) = s^T(A\bar{x} + b) = s^T \nabla f(\bar{x}) < 0$, co ma miejsce w przypadku stosowania metody spadku, to oczywiście $\tau > 0$ i dla $\bar{x} = x_k$ i $s = s_k$ wielkość τ jest długością kroku dla dokładnej minimalizacji kierunkowej.

3.1.1 Kryteria zatrzymania

W metodach minimalizacji, w przypadku stosowania *bezwzględnej tolerancji optymalności* $\varepsilon \geq 0$, najczęściej stosuje się następujące kryteria zatrzymania:

- $f(x_k) - f^* \leq \varepsilon$ (w przypadku, gdy znana jest wartość minimalna funkcji $f^* = \min_{x \in \mathbb{R}^n} f(x)$; kryterium to stosuje się czasem przy testowaniu danej metody dla znanych funkcji testujących),
- $\|x_k - x^*\| \leq \varepsilon$ (niestosowalne bezpośrednio z powodu nieznajomości rozwiązania optymalnego x^*; kryterium to stosuje się czasem przy testowaniu danej metody dla znanych funkcji testujących),
- $\|g_k\| \leq \varepsilon$ (i hesjan G_k dodatnio określony),
- $f(x_k) - f(x_{k+1}) \leq \varepsilon$ (ewentualnie dla kilku kolejnych k; kryterium to stosuje się w przypadku, gdy oczekujemy szybkiej zbieżności),
- $\|x_{k+1} - x_k\| \leq \varepsilon$ (ewentualnie dla kilku kolejnych k; kryterium to stosuje się w przypadku, gdy oczekujemy szybkiej zbieżności),
- $\overline{f}_k - \underline{f}_k \leq \varepsilon$, gdzie $\overline{f}_k$ i $\underline{f}_k$ są odpowiednio górnym i dolnym oszacowaniem wartości optymalnej f^*; najczęściej przyjmuje się $\overline{f}_k = \min_{1 \leq i \leq k} f(x_i)$.

W przypadku stosowania *względnej tolerancji optymalności* $\varepsilon \geq 0$ kryteria powyższe należy odpowiednio zmodyfikować, zastępując błędy bezwzględne przybliżeń ich błędami względnymi. Przykładowo, ostatnie z podanych wyżej kryteriów należy zastąpić przez $\frac{\overline{f}_k - \underline{f}_k}{|\underline{f}_k|} \leq \varepsilon$ o ile $f^* \neq 0$. Czasem stosuje się również kryterium $\frac{\overline{f}_k - \underline{f}_k}{1 + |\underline{f}_k|} \leq \varepsilon$, które jest bardziej odporne na przypadek $f^* = 0$.

Uwaga 3.1.6 W powyższych kryteriach norma $\|\cdot\|$ powinna być dostosowana do konkretnego problemu. Różne współrzędne x_i wektora x mogą bowiem oznaczać różne wielkości i norma wektora x powinna uwzględniać ich skalowanie.

3.1.2 Typy zbieżności

Zbieżność ciągu kolejnych przybliżeń do rozwiązania problemu minimalizacji może być szybsza lub wolniejsza. Stosując jakąkolwiek metodę minimalizacji jesteśmy oczywiście zainteresowani, aby ciąg kolejnych przybliżeń był w miarę szybko zbieżny do rozwiązania. Dana metoda minimalizacji generująca ciąg kolejnych przybliżeń (x_k) do rozwiązania x^* jest więc tym lepsza, im szybciej ciąg $(\|x_k - x^*\|)$ jest zbieżny do zera. Podamy teraz klasyfikację typów zbieżności stosowaną w metodach minimalizacji. W klasyfikacji tej istotne będzie porównanie odległości kolejnego przybliżenia od rozwiązania z odległością bieżącego przybliżenia od tego rozwiązania. Porównanie to można opisać nierównością

$$\|x_{k+1} - x^*\| \leq c_k \|x_k - x^*\|^p, \tag{3.2}$$

gdzie $c_k \geq 0$, $k \in \mathbb{N}$. Oczywiście, bez szkody dla ogólności rozważań, możemy przyjąć, że $x_k \neq x^*$ dla wszystkich $k \geq 1$.

Definicja 3.1.7 Niech ciąg (x_k) elementów przestrzeni $\mathbb{R}^n$ będzie zbieżny do $x^* \in \mathbb{R}^n$. Mówimy, że zbieżność ta jest:

(a) *liniowa* lub *geometryczna*, jeśli istnieje nieujemny ciąg ograniczony (c_k) mający wszystkie punkty skupienia w przedziale $(0,1)$, taki że dla każdego $k \geq 1$ zachodzi nierówność (3.2) dla $p = 1$;

(b) *szybsza od liniowej*, jeśli istnieje nieujemny ciąg (c_k) zbieżny do zera, taki że dla każdego $k \geq 1$ zachodzi nierówność (3.2) dla $p = 1$;

(c) *kwadratowa,* jeśli istnieje nieujemny ciąg ograniczony (c_k), taki że dla każdego $k \geq 1$ zachodzi nierówność (3.2) dla $p = 2$.

Uwaga 3.1.8 Równoważną definicję zbieżności liniowej otrzymamy przyjmując w (a), że ciąg c_k jest stały, $c_k = c \in (0,1)$ i że nierówność (3.2) zachodzi dla prawie wszystkich $k \in \mathbb{N}$. Przypuśćmy teraz, że ciąg $(\frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|^p})$ jest zbieżny. Można wówczas mówić, że zbieżność ciągu (x_k) jest *rzędu* p jeśli

$$\limsup_k \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^p} = c \geq 0.$$

W tym przypadku zbieżność ta jest: liniowa, gdy $p = 1$ i $c \in (0,1)$, szybsza od liniowej, gdy $p = 1$ i $c = 0$, kwadratowa, gdy $p = 2$. Jeśli natomiast dla ciąg (x_k) zbieżnego do x^* zachodzi równość

$$\limsup_k \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 1,$$

to mówimy, że zbieżność ta jest *wolniejsza od liniowej.*

Uwaga 3.1.9 W przypadku, gdy ciąg kolejnych przybliżeń (x_k) jest generowany przez pewien algorytm, to typy zbieżności tego ciągu przenoszą się na ten algorytm. Na przykład mówimy, że algorytm jest zbieżny liniowo, jeśli dowolny ciąg generowany przez ten algorytm jest zbieżny liniowo. Zbieżność algorytmu jest więc wolniejsza od liniowej, jeśli przynajmniej jeden z generowanych przez niego ciągów jest zbieżny wolniej niż liniowo. Algorytmy zbieżne wolniej niż liniowo są najczęściej mało przydatne z praktycznego punktu widzenia.

Uwaga 3.1.10 Typ zbieżności nie zależy od przyjętej normy $\|\cdot\|$, gdyż wszystkie normy w $\mathbb{R}^n$ są sobie równoważne.

Uwaga 3.1.11 Jeśli zbieżność ciągu (x_k) jest kwadratowa, to jest ona również szybsza od liniowej. Jeśli z kolei zbieżność ciągu (x_k) jest szybsza od liniowej, to (przynajmniej zgodnie z definicją 3.1.7) zbieżność ta jest również liniowa. Z punktu widzenia metod minimalizacji najkorzystniejsza wydaje się więc zbieżność kwadratowa. Przy praktycznej realizacji algorytmów zbieżność tę można zaobserwować, gdy po wykonaniu pewnej liczby iteracji odległość kolejnego przybliżenia od rozwiązania tak szybko maleje do zera, że liczba zer po przecinku (cyfr znaczących) dla tej odległości podwaja się z każdą iteracją. Widać to na przykładzie tzw. metody Newtona zastosowanej do minimalizacji funkcji $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x) = x_1^4 + 4\xi_1 x_2 + (1 + x_2)^2$ i dla punktu startowego $x_1 = (\frac{1}{2}, -1)$. Odległości kolejnych przybliżeń od rozwiązania przedstawione w tabeli 1.

k	$\|x_k - x^*\|$
1	$0,7071067811865$
2	$0,6480000000000$
3	$0,0752533710623$
4	$0,0015587238433$
5	$0,0000007181500$
6	$0,0000000000002$

Tabela 1

Z kolei, jeśli zbieżność jest liniowa, to w praktyce wyraża się to tym, że odległość $\|x_k - x^*\|$ maleje do zera w takim tempie, że kolejne zera po przecinku otrzymuje się w mniej więcej tej samej liczbie iteracji. Przykład takiej zbieżności przedstawiono w tabeli 2.

k	$\|x_k - x^*\|$
1	$80,00000000$
15	$0,78192425$
22	$0,00877470$
31	$0,00005440$
41	$0,00000082$
47	$0,00000001$

Tabela 2

Natomiast w tabeli 3. przedstawiono przykład zbieżności wolniejszej od liniowej.

k	$\|x_k - x^*\|$
1	1
10	$0,1$
100	$0,01$
1000	$0,0001$
10000	$0,00001$
100000	$0,000001$

Tabela 3

Jak już wspomnieliśmy w poprzednim ustępie,w praktyce trudno jest stosować kryterium zatrzymania postaci $\|x_k - x^*\| < \varepsilon$, gdyż najczęściej nie znamy rozwiązania x^* problemu minimalizacji. Jeśli natomiast wiemy, że zbieżność jest szybsza od liniowej, to możemy wówczas zastosować kryterium zatrzymania postaci $\|x_{k+1} - x_k\| < \varepsilon$. Zachodzi bowiem następujący lemat, którego prosty dowód pozostawiamy czytelnikowi.

Lemat 3.1.12 *Jeśli ciąg (x_k) jest zbieżny do x^* szybciej niż liniowo, to* $\lim_k \frac{\|x_{k+1}-x_k\|}{\|x_k-x^*\|} = 1$.

3.2 Minimalizacja kierunkowa

Zbliżając się do prawdy oddalamy się od rzeczywistości

S. J. Lec

Przedmiotem tego rozdziału będzie minimalizacja kierunkowa dla *metod spadku*, tj. takich, w których w kolejnych iteracjach niezerowy gradient g_k i kierunek poszukiwań s_k spełniają nierówność

$$s_k^T g_k < 0. \tag{3.3}$$

Przypomnijmy, że minimalizacja kierunkowa polega na doborze długości kroku $t_k > 0$, który zapewnia wystarczający spadek wartości funkcji f w kierunku poszukiwań s_k. Podstawowym, oczywistym żądaniem nakładanym na t_k jest więc zapewnienie spadku, czyli spełnienie warunku

$$f(x_{k+1}) = f(x_k + t_k s_k) < f(x_k).$$

Z różniczkowalności funkcji f wynika, że istnieje wielkość $t_k > 0$ spełniająca powyższy warunek, o ile spełniona jest nierówność (3.3). Określmy funkcję $q : \mathbb{R}_+ \to \mathbb{R}$ następująco:

$$q(t) = f(x_k + ts_k).$$

Oczywiście $q'(t) = s_k^T \nabla f(x_k + ts_k)$, w szczególności $q'(0) = s_k^T g_k < 0$. Nałożone wyżej żądanie oznacza, że $q(t_k) < q(0)$. Będziemy żądać jednak więcej: metoda minimalizacji kierunkowej powinna zapewnić "wystarczający" spadek wartości funkcji. Największy spadek wartości funkcji celu otrzymamy oczywiście stosując dokładną minimalizację kierunkową. Stosowanie jej jest jednak ograniczone. Pamiętajmy bowiem, że dysponujemy jedynie możliwością wyznaczenia w każdym punkcie x wartości funkcji $f(x)$ i gradientu $\nabla f(x)$ (i ewentualnie hesjanu $\nabla f^2(x)$). W metodach minimalizacji kierunkowej musimy więc znaleźć kompromis między dwoma celami: z jednej strony zapewnienie możliwie dużego spadku wartości funkcji, z drugiej zaś zakończenie procedury minimalizacji kierunkowej przy możliwie najmniejszej liczbie wyznaczeń funkcji i gradientu, gdyż do procedury tej odwołujemy się w każdej iteracji metody spadku. Dlatego też dużą uwagę musimy zwrócić na kryterium zatrzymania w tej procedurze. Kryterium to najczęściej nie ma prostej postaci "tak-nie". Dla danej funkcji $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, bieżącego przybliżenia $\bar{x} = x_k \in \mathbb{R}^n$, danego spadkowego kierunku poszukiwań $s = s_k \in \mathbb{R}^n$ i podanej długości kroku t kryterium to ma najczęściej postać pewnego testu składającego się z trzech części:

(O) długości kroku t jest odpowiednia i algorytm minimalizacji kierunkowej zatrzymuje się,

(R) długości kroku t jest za duża i należy ją zmniejszyć (*test prawostronny*),

(L) długości kroku t jest za mała i należy ją zwiększyć (*test lewostronny*).

Metody minimalizacji kierunkowej realizowane są w postaci procedur iteracyjnych, których ogólny schemat przedstawiono poniżej.

Schemat iteracyjny 3.2.1 (ogólna postać minimalizacji kierunkowej)

Wejście: $\bar{x} \in \mathbb{R}^n$ – bieżące przybliżenie rozwiązania, $s \in \mathbb{R}^n$ – kierunek poszukiwań, $t > 0$ – próbkowa długość kroku (można ewentualnie określić inne parametry).

Wyjście: $t > 0$ – długość kroku

Krok 0. (*Inicjalizacja*) Określić końce t_L, t_R przedziału, którym będzie poszukiwana długość kroku (na przykład położyć $t_L = 0$, $t_R = +\infty$).

Krok 1. Przetestować wielkość t.

(a) Jeśli zachodzi (O), to algorytm zatrzymuje się (bieżącej wartości t należy wówczas użyć w celu wyznaczenia wyznaczenia kolejnego przybliżenia rozwiązania w głównej procedurze minimalizacyjnej).

(b) Jeśli zachodzi (R), to położyć $t_R = t$ i przejść do kroku 2.

(c) W przypadku (L) położyć $t_L = t$ i przejść do kroku 2.

Krok 2. Zgodnie z odpowiednią procedurą (interpolacyjną lub ekstrapolacyjną) wybrać nową wartość $t \in (t_L, t_R)$ i przejść do kroku 1.

Uwaga 3.2.2 W przypadku, gdy $t_R < +\infty$ wyboru wielkości t w kroku 2. schematu iteracyjnego 3.2.1 dokonuje się zgodnie z zasadami interpolacji, jeśli zaś $t_R = +\infty$ – zgodnie z zasadami ekstrapolacji. Nietrudno zauważyć, że t_L jest niemalejące zaś t_R nierosnące względem licznika iteracji tego schematu.

Uwaga 3.2.3 Na razie nie został podany ani sposób, w który należy wybrać próbkową długość kroku t, ani postać kryterium zatrzymania (testów (O), (R), (L)), ani też metoda, zgodnie z którą należy wybrać $t \in (t_L, t_R)$,w kroku 2. schematu iteracyjnego 3.2.1. W dalszej części tego ustępu podajemy kilka takich kryteriów i kilka metod wyboru długości kroku $t \in (t_L, t_R)$, które wraz z odpowiednim wyborem kierunku s_k w głównej procedurze minimalizacji zapewniają zbieżność dla szerokiej klasy metod spadkowych.

3.2.1 Wymagania stawiane algorytmowi minimalizacji kierunkowej

Konkretna realizacja schematu iteracyjnego 3.2.1 powinna być tak skonstruowana, aby po wykonaniu skończenie wielu iteracji wyznaczona wartość t spełniała test (O). W tym przypadku mówimy, że odpowiedni algorytm jest *dobrze określony* (implementowalny). Aby to zapewnić, powinien on spełniać następujące warunki:

(w1) Jeśli algorytm prowadzi do nieskończenie wielu ekstrapolacji, to

$$t_L \to +\infty.$$

(w2) Jeśli algorytm prowadzi do nieskończenie wielu interpolacji, to

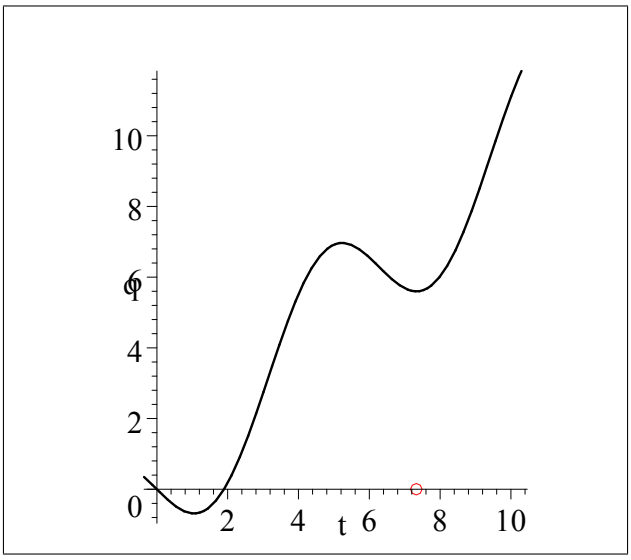
$$t_R - t_L \to 0.$$

(w3) Dla wystarczająco dużych t nie może zajść (L).

(w4) Jeśli $t_R - t_L$ jest wystarczająco małe, to (O) zachodzi dla wszystkich $t \in (t_L, t_R)$.

Ponadto, zgodnie z tym co mówiliśmy na początku tego ustępu, metoda minimalizacji kierunkowej (wybór w schemacie iteracyjnym 3.2.1 wielkości t spełniającej test (O)) powinna zapewniać wystarczający spadek wartości funkcji celu. Przede wszystkim powinien być więc spełniony warunek $f(x_{k+1}) < f(x_k)$, czyli dla wielkości t spełniającej test (O) powinna zajść nierówność $q(t) < q(0)$. Poniżej podajemy przykład kryterium zatrzymania nie posiadającego tej własności.

Przykład 3.2.4 Rozpatrzmy następujące "naiwne" kryterium zatrzymania:
(O) $q'(t) = 0$,
(L) $q'(t) < 0$,
(R) $q'(t) > 0$.
Zastosujmy je do funkcji $q(t) = t - 2\sin t$



Rysunek 2. Naiwne kryterium zatrzymania

Dla $t = \frac{7}{3}\pi$ zachodzi warunek (O) w omawianym kryterium (patrz rysunek 2.). Przyjmując $t_k = t$ otrzymamy $f(x_{k+1}) > f(x_k)$ wbrew wcześniejszemu żądaniu. Na tym przykładzie widzimy więc nieefektywność tego kryterium. Ponadto kryterium to ma jeszcze inną wadę. Nie spełnia ono bowiem warunku (w4).

Samo zmniejszanie się wartości funkcji celu w kolejnych iteracjach jest za małym wymaganiem. Aby to zauważyć, wystarczy rozpatrzyć funkcję $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$ i przyjąć punkt startowy $x_1 = 3$ oraz $s_k = -1$ i $t_k = 2^{-k}$, $k = 1, 2, ...$

Definicja 3.2.5 Niech $f : \mathbb{R}^n \to \mathbb{R}$ będzie klasy C_1, $\bar{x}, s \in \mathbb{R}^n$ będą takie, że $s^T \nabla f(\bar{x}) < 0$ (s jest kierunkiem spadku funkcji f w punkcie $\bar{x}$) i niech implementowalna metoda minimalizacji kierunkowej prowadzi do wyboru wielkości $t > 0$. Metoda ta nazywa się *efektywna*, jeśli istnieje stała $\alpha > 0$ (niezależna od $\bar{x}$ i od s) taka, że

$$f(\bar{x} + ts) < f(\bar{x}) - \alpha \left(\frac{s^T \nabla f(\bar{x})}{\|s\|} \right)^2$$

Zobaczymy później, że efektywna metoda minimalizacji kierunkowej przy odpowiednim wyborze kierunku poszukiwań zapewni przynajmniej stacjonarność każdego punktu skupienia ciągu kolejnych przybliżeń otrzymanego w wyniku realizacji metody spadkowej. Zobaczymy też, że metody minimalizacji kierunkowej, które wprowadzimy w dalszej części tego ustępu będą efektywne.

Uwaga 3.2.6 Wyjaśnimy teraz dlaczego w metodach minimalizacji kierunkowej powinniśmy dążyć raczej do mniejszej liczby wyznaczeń wartości funkcji i gradientu aniżeli do znacznego zmniejszenia wartości funkcji. Z przedstawionej na początku tego ustępu ogólnej postaci metod minimalizacji (schemat iteracyjny 3.2.1) wynika, że czas T wykonywania pojedynczej iteracji głównego algorytmu przedstawia się wzorem $T = T_A + lT_B$, gdzie l jest średnią liczbą wyznaczeń

wartości funkcji/gradientu, T_A – czasem potrzebnym na wyznaczenie kierunku poszukiwań, zaś T_B – czasem potrzebnym na wyznaczenie wartości funkcji/gradientu. Z praktyki wiadomo, że $T_B \gg T_A$ (często przyjmuje się, że $T_B \approx 10T_A$). Widzimy więc, że na łączny czas obliczeń największy wpływ ma liczba wyznaczeń wartości funkcji/gradientu. Powinniśmy zatem przy konstrukcji procedury minimalizacji kierunkowej dążyć do odpowiedniego kompromisu między szybkim zakończeniem procedury a dużym spadkiem wartości funkcji.

3.2.2 Testy prawo- i lewostronne

Zanim przejdziemy do sformułowania konkretnych kryteriów zatrzymania w metodach minimalizacji kierunkowej, podamy przykłady testów prawostronnych i lewostronnych.

Test prawostronny Goldsteina

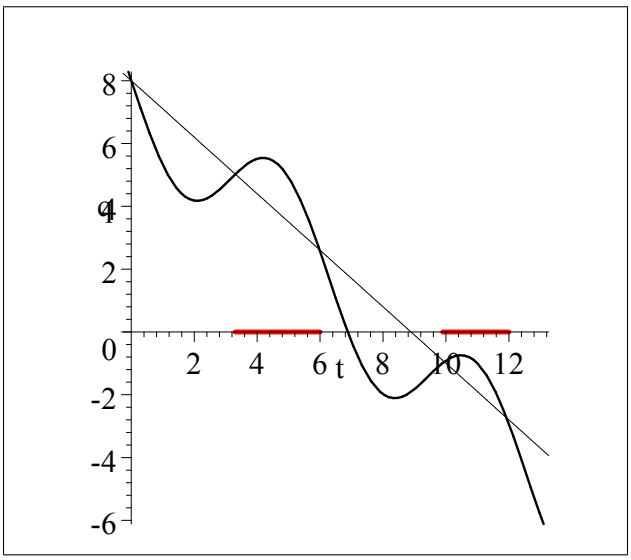
- Niech $m \in (0,1)$. Długość kroku t jest za duża jeśli

$$q(t) > q(0) + mtq'(0) \tag{r1}$$

lub inaczej

$$f(x) > f(x_k) + mts_k^T g_k.$$

Na rysunku 3. przedstawiono wykres przykładowej funkcji q i dla przykładowej stałej $m \in (0,1)$ zaznaczono zbiór wartości t spełniających warunek (r1) w teście prawostronnym Goldsteina.



Rysunek 3. Test prawostronny Goldsteina

Ćwiczenie 3.2.7 Pokazać, że jeśli q jest funkcją kwadratową osiągającą minimum w t^*, to

$$q(t^*) = q(0) + \frac{1}{2}t^* q'(0).$$

W oparciu o tę własność uzasadnić, dlaczego w teście (r1) korzystnie jest przyjąć $m < \frac{1}{2}$.

Test lewostronny Goldsteina

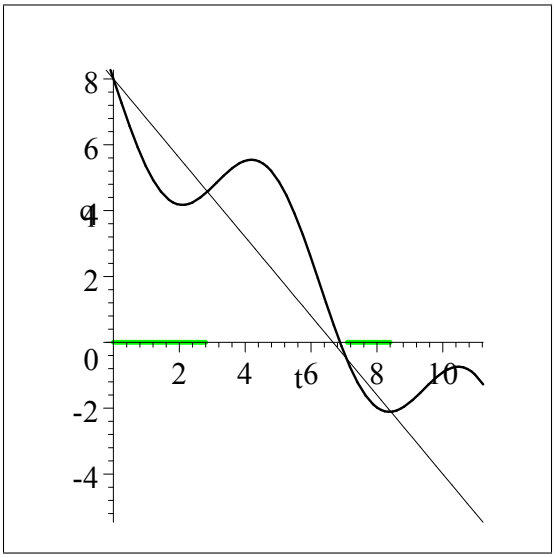
- Niech $m \in (0,1)$ będzie stałą występującą w teście prawostronnym Goldsteina i niech stała m' będzie taka, że $m < m' < 1$. Wówczas długość kroku t jest za mała, jeśli

$$q(t) < q(0) + m' t q'(0) \tag{11}$$

lub inaczej

$$f(x) < f(x_k) + m' t s_k^T g_k.$$

Na rysunku 4. przedstawiono wykres przykładowej funkcji q i dla przykładowej stałej $m' \in (0,1)$ zaznaczono zbiór wartości t spełniających warunek (11) w teście lewostronnym Goldsteina.



Rysunek 4. Test lewostronny Goldsteina

Test lewostronny Wolfe'a

- Niech $m \in (0,1)$ będzie stałą występującą w teście prawostronnym Goldsteina-Price'a i niech stała m' będzie taka, że $m < m' < 1$. Wówczas długość kroku t jest za mała, jeśli

$$q'(t) < m' q'(0), \tag{12}$$

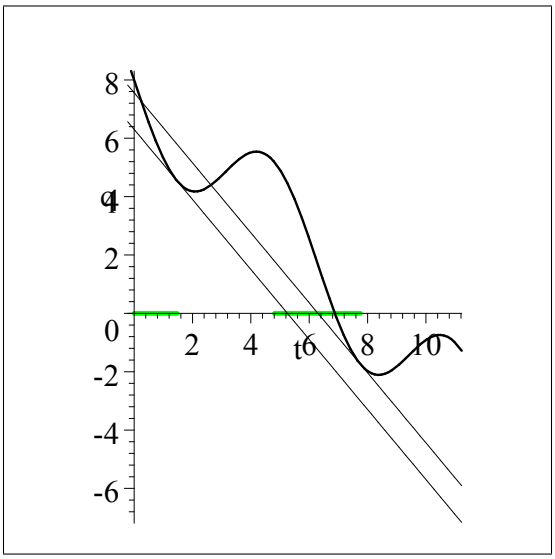
W tym teście wymagane jest dodatkowo wyznaczenie gradientu $\nabla f(x_k + t s_k)$. Należy jednak w tym miejscu podkreślić, że test ten w praktyce może zachowywać się bardzo niestabilnie. W sytuacji, gdy przybliżenie x_k jest już bardzo bliskie rozwiązania x^*, gradient $\nabla f(x_k)$ ma normę bliską zera. Jeśli kierunek poszukiwań s_k ma również małą normę (co ma również miejsce w pobliżu rozwiązania), to zgodnie z nierównością Schwarza mamy

$$|q'(0)| = \left| s_k^T \nabla f(x_k) \right| \leq \|s_k\| \cdot \|\nabla f(x_k)\|.$$

Jeśli więc obie normy po prawej stronie tej nierówności są rzędu 10^{-8}, to oczywiście $|q'(0)| \leq 10^{-16}$. Podobne oszacowanie zajdzie dla $|q'(t)|$. Porównywanie tak małych wielkości może prowadzić do istotnych zaburzeń w trakcie obliczeń numerycznych. Dlatego test ten można

bezpiecznie stosować w sytuacjach gdy wielkość $\|\nabla f(x_k)\|$ nie osiągnęła jeszcze odpowiednio małej wartości.

Na rysunku 5. przedstawiono wykres przykładowej funkcji q i dla przykładowej stałej $m' \in (0,1)$ zaznaczono zbiór wartości t spełniających warunek (l2) w teście lewostronnym Wolfe'a.



Rysunek 5. Test lewostronny Wolfe'a

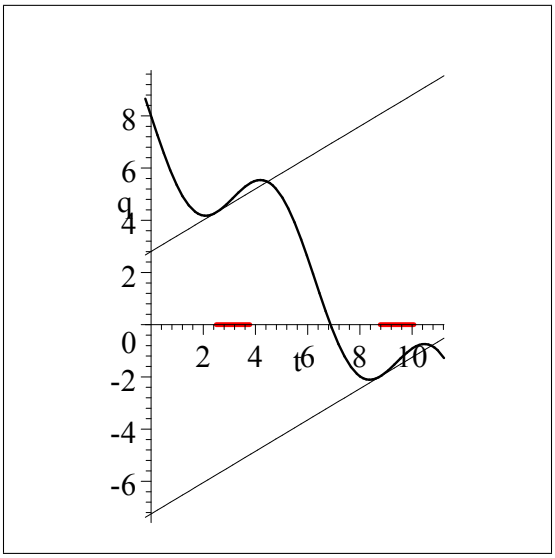
Ćwiczenie 3.2.8 Pokazać, że jeśli funkcja q jest ograniczona z dołu na przedziale $(0, +\infty)$, to testy (r1), (l1) i (l2) są spełnione dla pewnych wartości $t > 0$.

Test prawostronny Wolfe'a

Niech $m \in (0,1)$ będzie stałą występującą w teście prawostronnym Goldsteina i niech stała m' będzie taka, że $m < m' < 1$. Wówczas długość kroku t jest za duża, jeśli zachodzi (r1) lub

$$q'(t) > -m'q'(0), \tag{r2}$$

Na rysunku 6. przedstawiono wykres przykładowej funkcji q i dla przykładowej stałej $m' \in (0,1)$ zaznaczono zbiór wartości t spełniających warunek (r2) w teście prawostronnym Wolfe'a.



Rysunek 6. Test prawostronny Wolfe'a

3.3 Metody minimalizacji kierunkowej

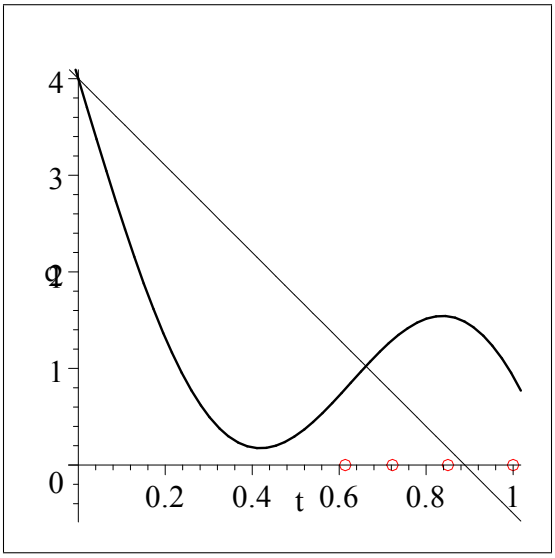
3.3.1 Metody Armijo

Przedstawimy teraz dwie metody Armijo wyznaczenia długości kroku t. Niech dane będą stałe $m \in (0,1)$ i $\beta \in (0,1)$. (Zwykła) *metoda Armijo* wyznacza długość kroku zgodnie ze wzorem

$$t^* = \min_{l \geq 0}\{\beta^l : q(\beta^l) \geq q(0) + m\beta^l q'(0)\}. \tag{3.4}$$

Innymi słowy, metoda Armijo polega na podstawianiu za t kolejnych elementów ciągu malejącego $(\beta^l)_{l\geq 0}$ i sprawdzaniu dla tych wartości t, czy spełniony jest test (r1). Pierwsza spośród kolejnych wartości t dla której test ten nie będzie spełniony jest w metodzie Armijo przyjmowana jako długość kroku t^*. Ponieważ $\beta \in (0,1)$, więc $\beta^l \to 0$ i z definicji pochodnej $q'(0)$ oraz z faktu, że $q'(0) < 0$ wynika prosto, że dla pewnego $l \geq 0$ i dla $t = \beta^l$ zajdzie $\sim$(r1). Oznacza to, że metoda Armijo jest poprawnie określona. Niech maksimum w (3.4) będzie osiągnięte dla $l = l^*$. Przypuśćmy, że dla $t = 1$ zachodzi test (r1) (wielkość t jest za duża). Zauważmy, że wówczas mamy $l^* \geq 1$ i dla $t = \beta^{l^*}$ zachodzi $\sim$(r1).

Na rysunku 7. przedstawiono wykres przykładowej funkcji q i dla przykładowych stałych $\beta, m \in (0,1)$ zaznaczono kolejne wielkości β^l, $l = 1, ..., l^*$ użyte w metodzie Armijo do wyznaczenia długości kroku t^*.



Rysunek 7. Metoda Armijo

Może się jednak zdarzyć, że już dla $t = 1$ (czyli dla $l = 0$) zachodzi $\sim$(r1). Wówczas oczywiście maksimum w (3.4) jest osiągnięte dla $l^* = 0$. Wydaje się rozsądne zwiększenie w tym przypadku długości kroku biorąc pod uwagę również ujemne wykładniki w równości (3.4). Prowadzi to do tzw. *rozszerzonej metody Armijo*. W metodzie tej długość kroku t^* wyznacza się zgodnie z następującą regułą:

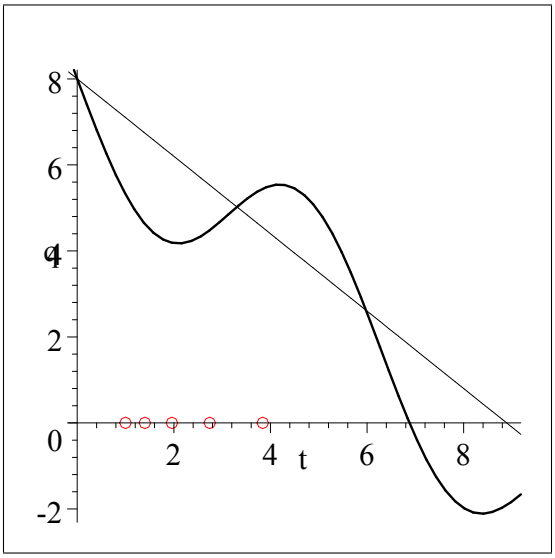
- Jeśli wielkość $t = 1$ jest za duża w teście prawostronnym Goldsteina (r1), to należy postępować jak w zwykłej metodzie Armijo, w przeciwnym przypadku należy przyjąć

$$t^* = \min_{l \geq 0}\{\beta^{-l} : q(\beta^{-l-1}) > q(0) + m\beta^{-l-1} q'(0)\}. \tag{3.5}$$

Innymi słowy, rozszerzona metoda Armijo polega na podstawianiu za t kolejnych elementów ciągu rosnącego $(\beta^{-l})_{l\geq 0}$ i sprawdzaniu dla tych wartości t, czy spełniony jest test (r1). Dla ostatniej wielkości $l = l^*$ wśród serii niespełnionych testów (r1) przyjmujemy $t^* = \beta^{-l}$.

Jeśli funkcja q jest ograniczona z dołu i dla $t = 1$ zachodzi $\sim$(r1), to nietrudno zauważyć, że minimum powyższe jest osiągnięte dla pewnego $l \geq 0$. Rozszerzona metoda Armijo jest więc również dobrze określona. Niech minimum w równości (3.5) będzie osiągnięte dla $l = l^*$. Wówczas dla $t = \beta^{-l^*} > 1$ zachodzi $\sim$(r1). Nietrudno zauważyć, że w obu przypadkach, dla wielkości t^* wyznaczonej zgodnie z jedną z metod Armijo zachodzi $\sim$(r1).

Na rysunku 8. przedstawiono wykres przykładowej funkcji q i dla przykładowych stałych $\beta, m \in (0,1)$ zaznaczono kolejne wielkości β^{-l}, $l = 1, ..., l^*$ użyte w metodzie Armijo do wyznaczenia długości kroku t^*.



Rysunek 8. Rozszerzona metoda Armijo

Praktyczny sposób wyboru stałej $\beta \in (0,1)$ można znaleźć na przykład w podręczniku [DS96, 6.3.2]. W niektórych podręcznikach rozszerzona metoda Armijo jest przedstawiona jako modyfikacja wzoru (3.4), w której maksimum bierze się po wszystkich całkowitych l. Ta postać jest jednak trudno implementowalna, a w przypadku, gdy funkcja f nie jest ograniczona z dołu, maksimum to może nie istnieć.

Chcąc zastosować rozszerzoną metodę Armijo do realizacji schematu iteracyjnego 3.2.1 należy:

(a) wybrać początkową próbkową długość kroku, na przykład $t = 1$, określić parametry $\beta \in (0,1)$ i $m \in (0,1)$,

(b) w kroku 0. przyjąć $t_R = +\infty$ oraz

$$t_L = \begin{cases} 0 & \text{jeśli dla wielkości } t \text{ zachodzi (r1)} \\ t & \text{jeśli dla wielkości } t \text{ zachodzi } \sim \text{(r1)}, \end{cases}$$

(c) określić test (O) w postaci:

(O) wielkość t jest odpowiednia gdy:

$$\begin{cases} \text{dla } t \text{ zachodzi } \sim \text{(r1)} & \text{jeśli } t_L = 0 \\ \text{dla } \beta^{-1}t \text{ zachodzi (r1)} & \text{jeśli } t_L > 0, \end{cases}$$

(d) określić test (R) w postaci:

(R) wielkość t jest za duża, gdy dla t zachodzi (r1),

(e) w przypadku $t_L > 0$ określić test (L) w postaci:

(L) wielkość t jest za mała, gdy dla $\beta^{-1}t$ zachodzi $\sim$(r1),

(f) w kroku 2. określić wartość $t \in (t_L, t_R)$ zgodnie ze wzorem

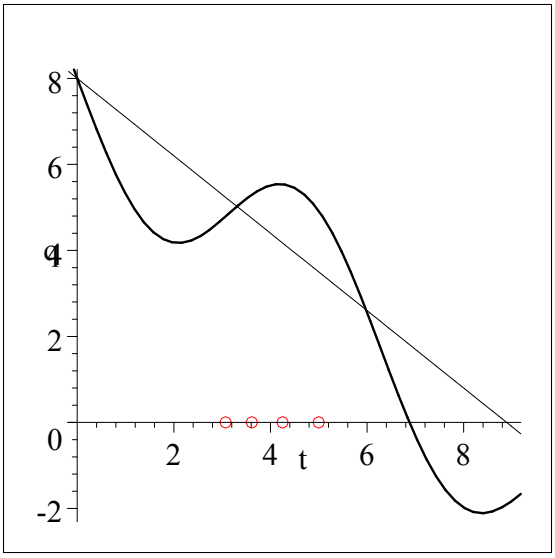
$$t = \begin{cases} \beta t_R & \text{jeśli } t_L = 0 \\ \beta^{-1} t_L & \text{jeśli } t_L > 0. \end{cases}$$

Zauważmy, że nie ma potrzeby określania testu (L) w przypadku $t_L = 0$, gdyż w tym przypadku nie dojdzie do jego wykonania.

Uwaga 3.3.1 Istnieją również warianty metod Armijo, w których wartość $t \in (t_L, t_R)$ w kroku 2. schematu iteracyjnego 3.2.1 określa się metodami interpolacji, np. kwadratowej.

Uwaga 3.3.2 Arbitralne przyjęcie w metodzie Armijo $t = 1$ dla początkowej próbkowej długości kroku może prowadzić do niepotrzebnego wielokrotnego testowania w kroku 1. schematu iteracyjnego 3.2.1, jeśli przyjęta wartość t jest o wiele za duża lub o wiele za mała. Żeby usunąć tę niekorzystną własność metody Armijo, można stosować tzw. *skalowaną metodę Armijo*. Polega ona na tym, że jako początkową próbkową długość kroku przyjmuje się $t = \tau$, gdzie $\tau > 0$ jest wielkością, dla której spodziewamy się, że będzie spełniać test (O). Wielkość τ można wyznaczyć biorąc na przykład pod uwagę długości kroków używane w poprzednich iteracjach głównej procedury minimalizacyjnej. Skalowania można użyć zarówno w zwykłej jak i w rozszerzonej metodzie Armijo.

Na rysunku 9. przedstawiono wykres przykładowej funkcji q i dla przykładowych stałych $\beta, m \in (0, 1)$ i dla stałej skalującej $\tau = 5$ zaznaczono kolejne wielkości β^l, $l = 1, ..., l^*$ użyte w metodzie Armijo do wyznaczenia długości kroku t^*.



Rysunek 9. Skalowana metoda Armijo

3.3.2 Metoda Goldsteina–Price'a

Podstawowym elementem *metody Goldsteina–Price'a* jest kryterium zatrzymania Goldsteina–Price'a, składające się z trzech testów:

(O) $\sim$(r1)$\wedge\sim$(l1)

(R) (r1)

(L) (l1)

Test (O) w kryterium Goldsteina–Price'a ma więc postać

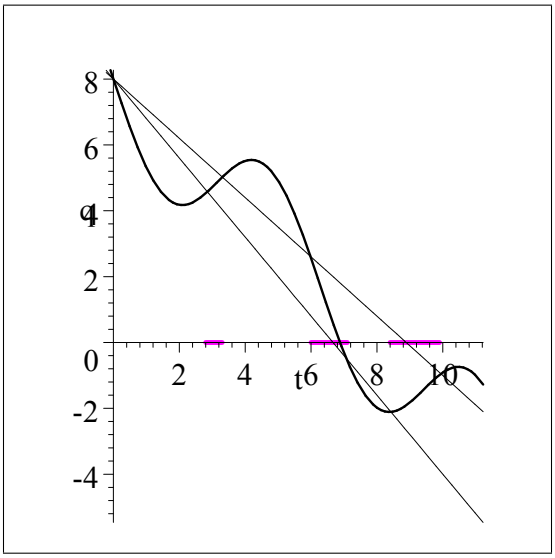
$$q(0) + m'tq'(0) \leq q(t) \leq q(0) + mtq'(0)$$

lub inaczej

$$f(x_k) + m'ts_k^T g_k \leq f(x_k + ts_k) \leq f(x_k) + mts_k^T g_k,$$

gdzie $0 < m < m' < 1$. Chcąc zastosować metodę Goldsteina–Price'a w celu realizacji schematu iteracyjnego 3.2.1 należy przyjąć powyższe testy (O), (R), (L), natomiast wyboru długości kroku $t \in (t_L, t_R)$ w kroku 2. można dokonać którąś z metod interpolacji (dla $t_R < +\infty$) względnie ekstrapolacji (dla $t_R = +\infty$).

Na rysunku 10. przedstawiono wykres przykładowej funkcji q i dla przykładowych stałych m, m', $0 < m < m' < 1$, zaznaczono zbiór wartości t, dla których spełniony jest warunek (O) w kryterium Goldsteina–Price'a.



Rysunek 10. Metoda Goldsteina–Price'a

3.3.3 Metoda Wolfe'a

Podstawowym elementem *metody Wolfe'a* jest kryterium zatrzymania Wolfe'a, składające się z trzech testów:

(O) $\sim$(r1)$\wedge\sim$(l2)

(R) (r1)

(L) $\sim$(r1)$\wedge$(l2)

Warunek (O) w kryterium Wolfe'a ma więc postać

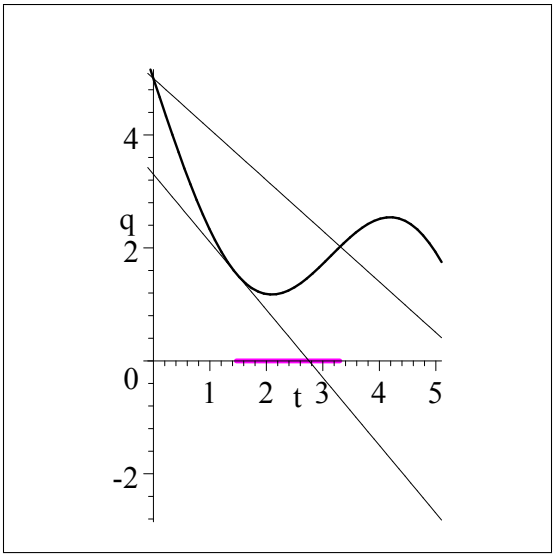
$$q(t) \leq q(0) + mtq'(0) \quad \text{i} \quad q'(t) \geq m'q'(0)$$

lub inaczej

$$f(x_k + ts_k) \leq f(x_k) + mts_k^T g_k \quad \text{i} \quad s_k^T \nabla f(x_k + ts_k) \geq m' s_k^T g_k,$$

gdzie $0 < m < m' < 1$. Chcąc zastosować metodę Wolfe'a w celu realizacji schematu iteracyjnego 3.2.1 należy przyjąć powyższe testy (O), (R), (L), natomiast wyboru długości kroku $t \in (t_L, t_R)$ w kroku 2. można dokonać którąś z metod interpolacji (dla $t_R < +\infty$) względnie ekstrapolacji (dla $t_R = +\infty$).

Na rysunku 11. przedstawiono wykres przykładowej funkcji q i dla przykładowych stałych m, m', $0 < m < m' < 1$, zaznaczono zbiór wartości t, dla których spełniony jest warunek (O) w kryterium Wolfe'a.



Rysunek 11. Metoda Wolfe'a

Ćwiczenie 3.3.3 Pokazać, że w kryteriach Goldsteina–Price'a i Wolfe'a testy (O), (R), (L) wyczerpują wszystkie możliwości.

Ćwiczenie 3.3.4 Pokazać, że jeżeli funkcja q jest ograniczona z dołu na zbiorze $(0, +\infty)$, to testy (L) i (O) w kryterium Wolfe'a zachodzą dla pewnych $t > 0$. **Wskazówka**: skorzystać z twierdzenia Lagrange'a o wartości średniej i z własności Darboux dla pochodnej q'.

3.3.4 Metoda Wolfe'a–Powella

Podstawowym elementem *metody Wolfe'a–Powella* jest kryterium zatrzymania Wolfe'a–Powella, składające się z trzech testów:

(O) $\sim$(r1)$\wedge$ $\sim$(r2)$\wedge$ $\sim$(l2)

(R) (r1)$\vee$(r2)

(L) $\sim$(r1)$\wedge$(l2)

Warunek (O) w kryterium Wolfe'a–Powella ma więc postać

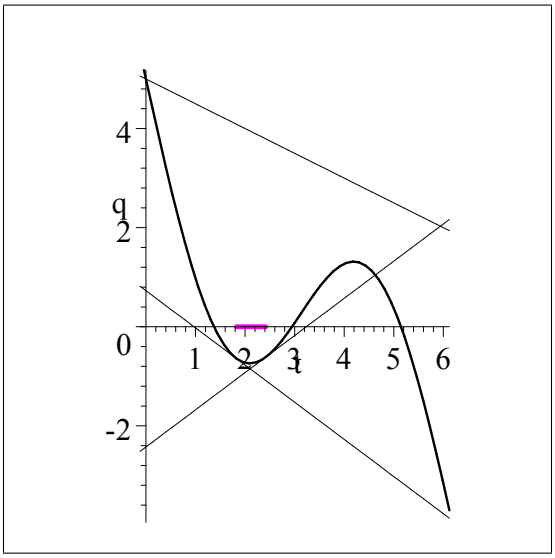
$$q(t) \leq q(0) + mtq'(0) \quad \text{i} \quad \mid q'(t) \mid \leq -m'q'(0) \tag{3.6}$$

lub inaczej

$$f(x_k + ts_k) \leq f(x_k) + mts_k^T g_k \quad \text{i} \quad \mid s_k^T \nabla f(x_k + ts_k) \mid \leq -m' s_k^T g_k,$$

gdzie $0 < m' < m < 1$. Chcąc zastosować metodę Wolfe'a–Powella w celu realizacji schematu iteracyjnego 3.2.1 należy przyjąć powyższe testy (O), (R), (L), natomiast wyboru długości kroku $t \in (t_l, t_R)$ w kroku 2. można dokonać którąś z metod interpolacji (dla $t_R < +\infty$) względnie ekstrapolacji (dla $t_R = +\infty$).

Na rysunku 12. przedstawiono wykres przykładowej funkcji q i dla przykładowych stałych m, m', $0 < m < m' < 1$, zaznaczono zbiór wartości t, dla których spełniony jest warunek (O) w kryterium Wolfe'a–Powella.



Rysunek 12. Metoda Wolfe'a–Powella

Uwaga 3.3.5 W metodach minimalizacji kierunkowej ważnym zagadnieniem jest wybór początkowej wartości t. Nieodpowiedni wybór może prowadzić niepotrzebnie do zbyt wielu wyznaczeń wartości funkcji/gradientu. Podamy teraz racjonalny wybór początkowej wartości t (wspominaliśmy już o tym w uwadze 3.3.2). Gdyby przyjąć, że q jest funkcją kwadratową, to na mocy ćwiczenia 3.2.7 otrzymamy optymalny krok

$$t^* = 2\frac{q(t^*) - q(0)}{q'(0)} = 2\frac{f(x_{k+1}) - f(x_k)}{q'(0)}.$$

Gdyby przyjąć, że spadek funkcji q na przedziale $(0, t^*)$ jest taki sam, jak w poprzedniej iteracji głównego algorytmu, to otrzymamy

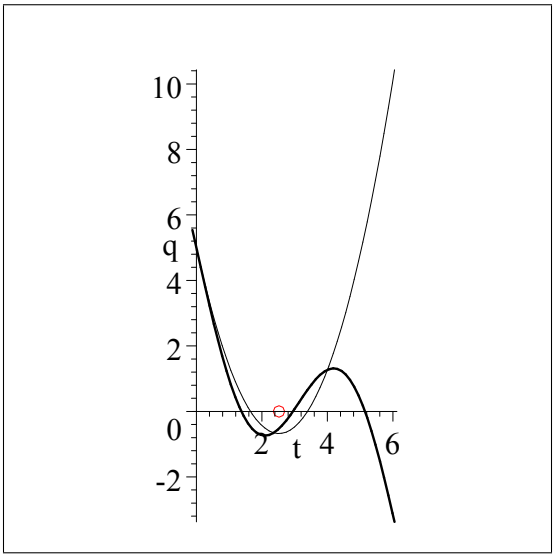
$$t^* = 2\frac{f(x_k) - f(x_{k-1})}{q'(0)},$$

którą to wielkość można przyjąć jako początkową wartość t w algorytmie minimalizacji kierunkowej.

Uwaga 3.3.6 W schemacie iteracyjnym 3.2.1 nie podano szczegółów dotyczących interpolacji i ekstrapolacji. Postępując zgodnie z metodą Armijo dokonuje się prostej interpolacji wybierając $t = \beta t_R$ bądź ekstrapolacji wybierając $t = \beta^{-1} t_L$, gdzie $\beta \in (0,1)$. W przypadku stosowania pozostałych metod minimalizacji kierunkowej istnieje wiele możliwości, wśród których wyróżnić można podział na 2 przy interpolacji ($t = \frac{t_L+t_R}{2}$) i podwajanie przy ekstrapolacji ($t = 2t_L$), interpolacja kwadratowa t_{int} używająca np. $q(t_L), q'(t_L)$ i $q(t_R)$, ekstrapolacja kwadratowa t_{ext} i inne (między innymi interpolacja sześcienna). Punkt t_{int} (względnie t_{ext}) jest wyznaczony w ten sposób, że odpowiedni model (na przykład kwadratowy) osiąga w nim swoje minimum. W przypadku ekstrapolacji należy pamiętać ponadto o wymaganiu $t_L \to +\infty$. Aby temu sprostać, należy w przypadku $t_R = +\infty$ wybrać $\rho > 1$ i wziąć $t = \max\{t_{ext}, \rho t_L\}$. Z kolei w przypadku interpolacji należy pamiętać o wymaganiu $t_R - t_L \to 0$. Aby sprostać z kolei temu wymaganiu należy w przypadku $t_R < +\infty$ wybrać $\rho \in (0, \frac{1}{2})$ i wziąć

$$t = \min\{\max\{t_{int}, (1-\rho)t_L + \rho t_R\}, \rho t_L + (1-\rho)t_R\}.$$

Na rysunku 13. przedstawiono wykres przykładowej funkcji q i dla wartości $t_L = 0$ i $t_R = 4$ wskazano wielkość t uzyskaną metodą interpolacji kwadratowej, w której użyto $q(t_L), q'(t_L)$ i $q(t_R)$.



Rysunek 13. Interpolacja kwadratowa

3.3.5 Bezgradientowe metody minimalizacji kierunkowej

Jeśli nie jesteśmy w stanie wyznaczyć wartości gradientu minimalizowanej funkcji f, czyli nie znamy pochodnej funkcji q, to nie możemy oczywiście zastosować metod minimalizacji kierunkowej opisanych w ustępie 3.3. Jeśli funkcja q jest ciągła, to możemy wówczas zastosować metody przybliżonej minimalizacji kierunkowej oparte wyłącznie o wartości funkcji q. Wśród tych metod można wyróżnić *metodę złotego podziału*. W metodzie tej przedział $[a,b]$, w którym poszukujemy przybliżonego minimum funkcji q dzielimy na dwa przedziały $[a,d]$ i $[c,b]$ zgodnie z tzw. złotym podziałem, czyli w proporcjach $\frac{d-a}{b-a} = \tau$ i $\frac{b-c}{b-a} = \tau$, gdzie $\tau = \frac{\sqrt{5}-1}{2} \cong 0.618$. Opisaną procedurę powtarzamy następnie dla jednego z przedziałów $[a,d]$ i $[c,b]$ w zależności od wartości $q(c)$ i $q(d)$. Procedura ta jest powtarzana tak długo, aż kolejny przedział będzie miał długość mniejszą od zadanej tolerancji optymalności.

Ćwiczenie 3.3.7 Niech $a < b$ i niech $\tau = \frac{\sqrt{5}-1}{2}$. Pokazać, że jeśli $\frac{d-a}{b-a} = \tau$ i $\frac{b-c}{b-a} = \tau$, to $\frac{b-d}{b-c} = \tau$ i $\frac{c-a}{d-a} = \tau$.

Algorytm 3.3.8 (metoda złotego podziału)

Wejście: Funkcja ciągła $q : [a, b] \to \mathbb{R}$

Wyjście: Wielkość $t \in (a, b)$, dla której funkcja q osiąga "przybliżone" minimum.

Krok 0. (*inicjalizacja*)

(a) Wybrać tolerancję optymalności $\varepsilon > 0$.

(b) Położyć $t_L = a$, $t_R = b$.

(c) Położyć $c = \tau t_L + (1 - \tau)t_R$.

(d) Położyć $d = (1 - \tau)t_L + \tau t_R$.

Krok 1. (*kryterium zatrzymania*)

(a) Jeśli $t_R - t_L \leq \varepsilon$, to położyć $t = \frac{t_R + t_L}{2}$ i algorytm zatrzymuje się.

Krok 2. (*aktualizacja przedziału poszukiwań*)

(a) Jeśli $q(c) > q(d)$, to położyć $t_L = c$, $c = d$ i $d = (1 - \tau)t_L + \tau t_R$.
Jeśli $q(c) \leq q(d)$, to położyć $t_R = d$, $d = c$ i $c = \tau t_L + (1 - \tau)t_R$.

(b) Przejść do kroku 1.

3.4 Zbieżność metod spadkowych

Rozważmy teraz metodę spadku (czyli spełniającą nierówność (3.3)). Okazuje się, że nie gwarantuje to jeszcze zbieżności ciągu x_k do rozwiązania niezależnie od zastosowanej metody minimalizacji kierunkowej, gdyż kierunek poszukiwań s_k może okazać się prawie ortogonalny do kierunku najszybszego spadku $-g_k$ lub – co gorsza – przy $k \to \infty$ kąt między tymi dwoma kierunkami może zbiegać do $\pi/2$. W konsekwencji, spadki w kolejnych iteracjach mogą okazać się niewystarczające dla zapewnienia zbieżności ciągu x_k do punktu stacjonarnego funkcji celu. Aby zagwarantować tę zbieżność powinniśmy dodatkowo założyć, że

$$\Theta_k := \sphericalangle(s_k, -g_k) \leq \frac{\pi}{2} - \mu \tag{3.7}$$

dla pewnego $\mu > 0$. Kierunki (s_k) spełniające powyższą zależność nazywają się *kierunkami gradientopodobnymi.* Kąt Θ_k można oczywiście wyznaczyć z zależności

$$\cos\Theta_k = \frac{-s_k^T g_k}{\|s_k\| \cdot \|g_k\|}. \tag{3.8}$$

Warunek (3.7) można wobec tego sformułować w postaci

$$\frac{-s_k^T g_k}{\|s_k\| \cdot \|g_k\|} \geq c > 0, \tag{3.9}$$

gdzie $c = \sin\mu$. Jest jasne, że kierunki gradientopodobne są kierunkami spadku.

Ponieważ kierunki spadku wyznacza się często w postaci $s_k = -H_k g_k$ dla dodatnio określonej macierzy H_k (porównaj uwagę 3.1.3), więc przydatne byłoby kryterium, dzięki któremu potrafilibyśmy stwierdzić, czy dla tak wyznaczonych kierunków spełnione jest założenie (3.7). Kryterium to można sformułować w oparciu o poniższy lemat.

Lemat 3.4.1 *Jeśli dla ciągu macierzy dodatnio określonych H_k, dla pewnych stałych $m, M \in \mathbb{R}$ takich, że $0 < m < M$ i dla dowolnych wektorów $s \in \mathbb{R}^n$ zachodzą nierówności*

$$m\|s\|^2 \leq s^T H_k s \leq M\|s\|^2, \tag{3.10}$$

to dla kierunków spadku $s_k = -H_k g_k$ funkcji różniczkowalnej $f : \mathbb{R}^n \to \mathbb{R}$ spełniony jest warunek (3.9).

Dowód. Z pierwszej nierówności (3.10) otrzymujemy

$$-s_k^T g_k = g_k^T H_k g_k \geq m\|g_k\|^2. \tag{3.11}$$

Z własności macierzy dodatnio określonej i z nierówności (3.10) wynika, że $m \leq \lambda_{\min}(H_k)$ oraz $M \geq \lambda_{\max}(H_k)$. Z kolei dla normy spektralnej zachodzi równość $\|H_k\| = \lambda_{\max}(H_k)$. Zatem z (3.11) i z nierówności Cauchy'ego–Schwarza wynika, że

$$\|s_k\| \cdot \|g_k\| = \|H_k g_k\| \cdot \|g_k\| \leq \|H_k\| \cdot \|g_k\|^2 = \lambda_{\max}(H_k)\|g_k\|^2 \leq M\|g_k\|^2. \tag{3.12}$$

W konsekwencji, po podzieleniu stronami (3.11) i (3.12) otrzymujemy nierówność (3.9) dla $c = m/M$. ■

Możemy już sformułować ogólne twierdzenia podające warunki wystarczające zbieżności metod minimalizacji. Przypomnijmy, że metody spadku składają się z dwóch części: wyznaczenia kierunku poszukiwań i minimalizacji kierunkowej.

Twierdzenie 3.4.2 *Niech $f \in C_1$. Jeśli dla metody spadkowej kierunki poszukiwań są gradientopodobne zaś metoda minimalizacji kierunkowej jest efektywna, to dowolny punkt skupienia ciągu (x_k) generowanego przez tę metodę jest punktem stacjonarnym funkcji f.*

Dowód. Ponieważ metoda minimalizacji kierunkowej jest efektywna, więc istnieje stała $\alpha > 0$ taka, że dla $k = 1, 2, ...$ zachodzą nierówności

$$f(x_{k+1}) = f(x_k + t_k s_k) < f(x_k) - \alpha \left(\frac{s_k^T g_k}{\|s_k\|} \right)^2 . \tag{3.13}$$

Z kolei, ponieważ kierunki poszukiwań są gradientopodobne, więc $\frac{-s_k^T g_k}{\|s_k\| \cdot \|g_k\|} > c$ dla pewnej stałej $c > 0$. W konsekwencji

$$f(x_k) - f(x_{k+1}) > \alpha c^2 \|g_k\|^2. \tag{3.14}$$

Niech teraz x^* będzie punktem skupienia ciągu (x_k). Ponieważ f jest ciągła, więc $f(x^*)$ jest punktem skupienia ciągu $(f(x_k))$. Z nierówności (3.13) wynika, że ciąg $(f(x_k))$ monotonicznie maleje, jest więc zbieżny do $f(x^*)$. Stąd $f(x_k) - f(x_{k+1}) \to 0$ i w konsekwencji $\|g_k\| \to 0$ na mocy nierówności (3.14). Ponieważ $f \in C_1$ więc $\nabla f(x^*) = 0$. ■

Uwaga 3.4.3 Dla zajścia tezy twierdzenia 3.4.2 nie jest potrzebne, aby we wszystkich iteracjach metody spadkowej kierunki s_k były gradientopodobne, zaś długości kroku t_k wyznaczane metodą efektywnej minimalizacji kierunkowej. W istocie wystarczy, aby warunki te były spełnione łącznie dla nieskończenie wielu liczb naturalnych $k = k_i$, $i = 1, 2, ...$, gdzie ciąg (k_i) jest rosnący. Aby to zauważyć wystarczy nieznacznie zmodyfikować dowód. Prawdziwość tej wersji jest o tyle istotna, że w metodach minimalizacji mających charakter heurystyczny dla ich zbieżności wystarczy, że tylko w niektórych iteracjach jednocześnie kierunki będą gradientopodobne i długości kroku efektywne. Mogą się one pojawiać nawet bardzo rzadko i nieregularnie. Należy jednak zadbać o to, aby otrzymywać spadki wartości funkcji celu w kolejnych iteracjach (zgodnie z założeniami twierdzenia metoda ma być spadkowa). Przykładem jest metoda gradientów sprzężonych (opisana w jednym z kolejnych ustępów), w której co n iteracji dokonywany jest tzw. "ciepły restart".

Uwaga 3.4.4 Twierdzenie 3.4.2 nie gwarantuje istnienia punktu skupienia ciągu (x_k) generowanego przez opisaną w nim metodę spadku, a w przypadku istnienia takiego punktu – twierdzenie nie gwarantuje jego jednoznaczności.

Jeśli jednak wiadomo, że podpoziomica $S(f, f(x_{k_0}))$ dla pewnego k_0 jest zbiorem zwartym, to wówczas ciąg (x_k) ma punkt skupienia. Wynika to z faktu, że ciąg ten jest generowany przez metodę spadku, a więc $x_k \in S(f, f(x_{k_0}))$ dla $k \geq k_0$. Jeśli ponadto podpoziomica $S(f, f(x_{k_0}))$ jest zbiorem wypukłym, a funkcja f obcięta do tego zbioru jest ściśle wypukła, to ciąg (x_k) ma dokładnie jeden punkt skupienia. Wynika to z faktu, że funkcja ściśle wypukła ma dokładnie jeden minimizer x^* (porównaj twierdzenie 2.1.3) i z warunku koniecznego rzędu pierwszego osiągania minimum ($\nabla f(x^*) = 0$).

Opisane warunki dotyczące jednoznaczności punktu skupienia ciągu (x_k) są jednak dosyć restrykcyjne i – nawet jeśli są spełnione – są często trudne do sprawdzenia. Trudno bowiem oczekiwać, że będzie spełnione założenie o wypukłości podpoziomicy $S(f, f(x_{k_0}))$ poza przypadkami, gdy sama funkcja f jest wypukła (lub przynajmniej quasi-wypukła).

Natomiast jeśli $f \in C_2$ i hesjan $\nabla^2 f$ jest dodatnio określony punkcie x^* będącym minimizerem globalnym funkcji f (warunek ten jest często spełniony, szczególnie wówczas gdy minimum jest izolowane), to f jest ściśle wypukła w pewnym otoczeniu $B(x^*, \varepsilon)$ punktu x^* oraz dla α dostatecznie bliskiego wartości minimalnej f^* podpoziomica $S(f, \alpha)$ jest zbiorem wypukłym zawartym w $B(x^*, \varepsilon)$. Wówczas, jeśli tylko $x_{k_0} \in S(f, \alpha)$, to z powyższych rozważań wynika, że minimizer x^* jest jedynym punktem skupienia ciągu (x_k). Opisana sytuacja ma miejsce na przykład wtedy, gdy funkcja f jest ściśle wypukła na $\mathbb{R}^n$ i osiąga swoje minimum.

Uwaga 3.4.5 W przypadku, gdy funkcja celu jest mocno wypukła, zbieżność metody spadkowej do (jednoznacznie określonego) minimizera można uzyskać przy nieco słabszych warunkach dotyczących kierunku poszukiwań niż warunek (3.9). Szczegóły można znaleźć w podręczniku [GK99, tw. 4.7].

Uwaga 3.4.6 W przypadku, gdy punkt skupienia ciągu (x_k) generowanego przez metodę spadkową jest minimizerem lokalnym izolowanym, to przy spełnieniu założeń twierdzenia 3.4.2 można uzyskać zbieżność ciągu (x_k) do tego punktu przy niezbyt mocnych założeniach dotyczących długości wektora kierunku ($\|s_k\| \leq \alpha\|\nabla f(x_k)\|$ dla pewnego $\alpha > 0$) i długości kroku ($t_k \leq \tau$ dla pewnego $\tau > 0$). Warunki te są opisane w tzw. twierdzeniu o przechwyceniu [Ber95, propozycja 1.2.5]. Jeśli więc wiadomo, że wszystkie minimizery funkcji celu są izolowane, to przy spełnieniu powyższych warunków zachodzi zbieżność ciągu (x_k) do jednego z minimizerów, w zależności od przyjętego punktu startowego x_1.

Pokażemy teraz, że przedstawione w tym ustępie metody minimalizacji kierunkowej są efektywne przy pewnych ogólnych założeniach dotyczących funkcji celu.

Twierdzenie 3.4.7 *Niech funkcja $f \in C_1$ będzie ograniczona z dołu, $\bar{x} \in \mathbb{R}^n$, zaś gradient funkcji f będzie odwzorowaniem ciągłym w sensie Lipschitza na podpoziomicy $S(f, f(\bar{x}))$ i niech $s \in \mathbb{R}^n$ będzie kierunkiem takim, że $s^T\nabla f(\bar{x}) < 0$. Wówczas następujące metody zastosowane do punktu $\bar{x}$ i kierunku s są efektywne:*

(a) *rozszerzona metoda Armijo,*

(b) *metoda Goldsteina–Price'a,*

(c) *metoda Wolfe'a,*

(d) *metoda Wolfe'a–Powella*

Jeśli ponadto podpoziomica $S(f, f(\bar{x}))$ jest zbiorem zwartym, to

(e) *metoda dokładnej minimalizacji kierunkowej jest efektywna.*

Dowód. Oznaczmy $q(\tau) = f(\bar{x} + \tau s)$.

(a) Przypuśćmy, że dla $\beta \in (0, 1)$ długość kroku $t = \beta^l > 0$ otrzymano rozszerzoną metodą Armijo. Wówczas oczywiście

$$f(\bar{x} + ts) \leq f(\bar{x}) + mts^T\nabla f(\bar{x}) \tag{3.15}$$

i

$$f(\bar{x} + \beta^{-1}ts) \geq f(\bar{x}) + m\beta^{-1}ts^T\nabla f(\bar{x}) \tag{3.16}$$

Ponieważ $s^T\nabla f(\bar{x}) < 0$, więc z pierwszej z powyższych nierówności wynika, że $\bar{x} + ts \in S(f, \bar{x})$. Z nierówności Schwarza i z ciągłości w sensie Lipschitza gradientu funkcji f otrzymujemy więc dla $\tau > 0$

$$\begin{aligned} q'(\tau) - q'(0) &= s^T(\nabla f(\bar{x} + \tau s) - \nabla f(\bar{x})) \\ &\leq \|s\| \cdot \|\nabla f(\bar{x} + \tau s) - \nabla f(\bar{x})\| \\ &\leq L\|s\|^2\tau. \end{aligned}$$

Wobec tego

$$\begin{aligned} q(u) &= q(0) + \int_0^u q'(\tau)d\tau \\ &\leq q(0) + \int_0^u (q'(0) + L\|s\|^2\tau)d\tau \\ &= q(0) + q'(0)u + \frac{1}{2}L\|s\|^2u^2 \end{aligned}$$

Dla $u = \beta^{-1}t$ otrzymujemy więc

$$q(\beta^{-1}t) \leq q(0) + q'(0)\beta^{-1}t + \frac{1}{2}L\|s\|^2\beta^{-2}t^2$$

Z drugiej strony, nierówność (3.16) możemy zapisać w postaci

$$q(\beta^{-1}t) \geq q(0) + m\beta^{-1}tq'(0).$$

Wobec tego

$$q(0) + m\beta^{-1}tq'(0) \leq q(0) + q'(0)\beta^{-1}t + \frac{1}{2}L\|s\|^2\beta^{-2}t^2,$$

czyli

$$\frac{1}{2}L\|s\|^2\beta^{-2}t^2 + (1-m)q'(0)\beta^{-1}t \geq 0$$

Ponieważ $t > 0$, więc z nierówności tej wynika, że

$$t \geq \frac{2(m-1)\beta q'(0)}{L\|s\|^2} = \frac{2(m-1)\beta s^T\nabla f(\bar{x})}{L\|s\|^2}.$$

Podstawiając powyższe oszacowanie do nierówności (3.15) otrzymamy

$$f(\bar{x}+ts) \leq f(\bar{x}) - \frac{2m(1-m)\beta}{L}\left(\frac{s^T\nabla f(\bar{x})}{\|s\|}\right)^2.$$

Nierówność ta oznacza, że rozszerzona metoda Armijo jest efektywna.

(b) Dowód efektywności metody Goldsteina–Price'a jest podobny do dowodu efektywności rozszerzonej metody Armijo i pozostawiamy go czytelnikowi.

(c) Przypuśćmy, że długość kroku $t > 0$ otrzymano metodą Wolfe'a. Wówczas oczywiście

$$f(\bar{x}+ts) \leq f(\bar{x}) + mts^T\nabla f(\bar{x}) \tag{3.17}$$

i

$$s^T\nabla f(\bar{x}+ts) \geq m's^T\nabla f(\bar{x}) \tag{3.18}$$

gdzie $0 < m < m' < 1$. Ponieważ $s^T\nabla f(\bar{x}) < 0$, więc z nierówności (3.17) wynika, że $\bar{x}+ts \in S(f, f(\bar{x}))$. Zapiszmy nierówność (3.18) w postaci

$$(m'-1)s^T\nabla f(\bar{x}) \leq s^T(\nabla f(\bar{x}+ts) - \nabla f(\bar{x})).$$

Na mocy ciągłości w sensie Lipschitza gradientu funkcji f i po skorzystaniu z nierówności Schwarza otrzymujemy więc

$$\begin{aligned} (m'-1)s^T\nabla f(\bar{x}) &\leq \|s\| \cdot \|\nabla f(\bar{x}+ts) - \nabla f(\bar{x})\| \\ &\leq Lt\|s\|^2, \end{aligned}$$

gdzie $L > 0$ jest stałą Lipschitza dla ∇f na zbiorze $S(f, \bar{x})$. Stąd

$$t \geq \frac{(m' - 1)s^T \nabla f(\bar{x})}{L\|s\|^2}.$$

Podstawiając powyższe oszacowanie do nierówności (3.17) otrzymujemy

$$f(\bar{x} + ts) \leq f(\bar{x}) - \frac{m(1 - m')}{L} \left(\frac{s^T \nabla f(\bar{x})}{\|s\|} \right)^2.$$

Nierówność ta oznacza, że metoda Wolfe'a wyboru długości kroku jest efektywna.

(d) Ponieważ długość kroku wyznaczona metodą Wolfe'a–Powella spełnia również test (O) w kryterium Wolfe'a, więc efektywność tej metody wynika bezpośrednio z efektywności metody Wolfe'a.

(e) Przypuśćmy, że długość kroku $t > 0$ otrzymano metodą dokładnej minimalizacji kierunkowej. Postępując podobnie jak w dowodzie punktu (a) otrzymamy dla $\tau > 0$

$$q_1(\tau) \leq q(\tau) \leq q_2(\tau)$$

gdzie

$$q_1(\tau) = q(0) + q'(0)\tau - \frac{1}{2}L\|s\|^2\tau^2$$

zaś

$$q_2(\tau) = q(0) + q'(0)\tau + \frac{1}{2}L\|s\|^2\tau^2.$$

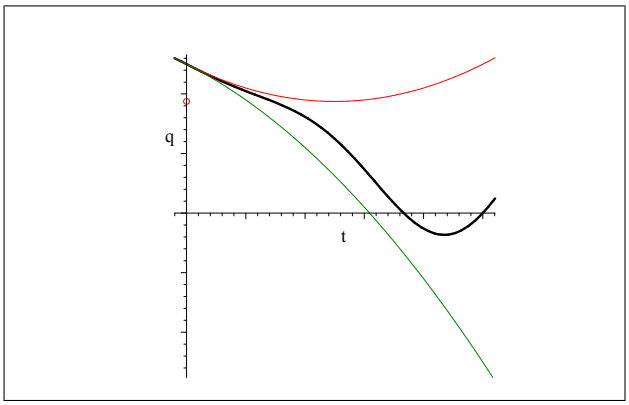
Niech

$$\bar{t} = \operatorname*{argmin}_{t \geq 0} q_2(t) = \frac{-q'(0)}{L\|s\|^2}$$

oraz

$$\bar{q}_2 = q_2(\bar{t}) = q(0) - \frac{(q'(0))^2}{2L\|s\|^2}.$$

Na rysunku 14. przedstawiono wykres przykładowej funkcji q oraz wykresy funkcji q_1 i q_2 oraz punkt $(0, \bar{q}_2)$.



Rysunek 14. Funkcja q i jej ograniczenia

Jest jasne, że

$$q(t) \leq \bar{q}_2$$

czyli

$$f(\bar{x}+ts) \leq f(\bar{x}) - \frac{1}{2L}\frac{(s^T\nabla f(\bar{x}))^2}{\|s\|^2}.$$

Nierówność ta oznacza, że metoda dokładniej minimalizacji kierunkowej jest efektywna.

■

Uwaga 3.4.8 Zwykła metoda Armijo nie jest efektywna. Jeśli jednak zastosujemy do niej skalowanie i przyjmiemy, że w każdej iteracji stała skalująca $\tau_k > -\alpha\frac{s_k^T\nabla f(x_k)}{\|s_k\|}$ dla pewnej stałej $\alpha > 0$ niezależnej od wskaźnika iteracji k, to wówczas skalowana metoda Armijo będzie efektywna. Dowód tego faktu pozostawiamy czytelnikowi.

Wniosek 3.4.9 *Niech $x_1 \in \mathbb{R}^n$ będzie punktem startowym w metodzie spadkowej. Niech funkcja $f \in C_1$ będzie ograniczona z dołu zaś gradient funkcji f będzie odwzorowaniem ciągłym w sensie Lipschitza na podpoziomicy $S(f, f(x_1))$, Jeśli w metodzie tej kierunki poszukiwań są gradientopodobne, zaś minimalizacja kierunkowa jest przeprowadzana rozszerzoną metodą Armijo, metodą Goldsteina–Price'a, metodą Wolfe'a, metodą Wolfe'a–Powella lub metodą dokładnej minimalizacji kierunkowej (o ile podpoziomica $S(f, f(x_1))$ jest zbiorem zwartym), to to dowolny punkt skupienia ciągu (x_k) generowanego przez tę metodę spadkową jest punktem stacjonarnym funkcji f.*

3.5 Metody minimalizacji bez ograniczeń

Smok optymalizacji ma wiele głów i na każdą z nich potrzebny jest właściwy miecz

V. Demyanov

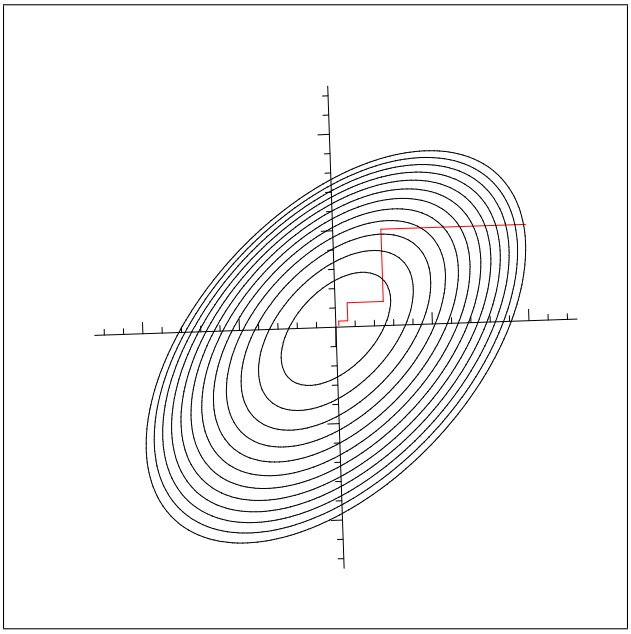
Przedstawimy teraz metody służące rozwiązaniu zadania minimalizacji różniczkowalnej bez ograniczeń

$$\begin{array}{rl} \text{minimalizować} & f(x) \\ \text{względem} & x \in \mathbb{R}^n, \end{array}$$

W większości przypadków metody te polegają na wyznaczeniu kierunku poszukiwań. W niektórych przypadkach minimalizacja kierunkowa jest specjalnie dobrana do danej metody, w innych przypadkach stosuje się jakąkolwiek minimalizację kierunkową z zachowaniem odpowiednich kryteriów zatrzymania.

3.5.1 Metoda Gaussa–Seidela

Metoda Gaussa–Seidela polega na tym, że kolejne kierunki poszukiwań określa się cyklicznie jako wektory równoległe do osi układu współrzędnych,



Rysunek 15. Metoda Gaussa–Seidela

czyli $s_k = \pm e_i$, gdzie $k = i(\operatorname{mod} n)$ oraz $s_k = +e_i$ jeśli $e_i^T g_k < 0$ i $s_k = -e_i$ jeśli $e_i^T g_k > 0$. W przypadku, gdy $e_i^T g_k = 0$, wykonywany jest *krok zerowy*, czyli $x_{k+1} = x_k$. W metodzie Gaussa–Seidela stosuje się dokładną minimalizację kierunkową.

Metoda Gaussa–Seidela jest czasem realizowana w praktyce, jest to jednak metoda bardzo nieefektywna. Metodzie tej brak jest podstaw teoretycznych. Co gorsza, istnieją przykłady, dla których metoda ta nie jest zbieżna do punktu stacjonarnego. Z tych powodów metody tej nie będziemy dalej analizować.

3.5.2 Metody najszybszego spadku

Metody najszybszego spadku polegają na wyborze kierunku poszukiwań będącym kierunkiem najszybszego spadku, czyli $s_k = -g_k$. Widzimy, że dla takiego wybory kierunku mamy $\sphericalangle(s_k, -g_k) = 0$, czyli spełniony jest warunek (3.7), metoda najszybszego spadku jest więc gradientopodobna. Zatem stosując tę metodę łącznie z dowolną efektywną minimalizacją kierunkową (np. rozszerzoną metodą Armijo) otrzymamy na mocy twierdzenia 3.4.2 stacjonarność każdego punktu skupienia procedury minimalizacji.

Jedną z metod najszybszego spadku jest *metoda Cauchy'ego*. Jest to metoda najszybszego spadku z dokładną minimalizacją kierunkową. Ponieważ przy założeniu zwartości zbioru $S(f, f(x_1))$ dokładna minimalizacja kierunkowa jest efektywna, więc dla metody Cauchy'ego otrzymujemy w tym przypadku również stacjonarność każdego punktu skupienia ciągu generowanego przez tę metodę, o ile $f \in C_1$. Z uwagi 3.1.4 wynika, że dla metody Cauchy'ego zachodzi równość $\nabla f(x_k)^T \nabla f(x_{k+1}) = 0$.

Metody najszybszego spadku, w szczególności metoda Cauchy'ego są często używane przez inżynierów. Mimo swojej prostoty i łatwości w zaprogramowaniu, mają one jednak zasadniczą wadę. Już dla funkcji kwadratowej ich zbieżność jest tylko liniowa, co wynika z poniższego twierdzenia.

Twierdzenie 3.5.1 *Niech funkcja $f : \mathbb{R}^n \to \mathbb{R}$ będzie postaci $f(x) = \frac{1}{2}x^T Ax$, gdzie A jest macierzą dodatnio określoną i niech ciąg (x_k) będzie generowany przez metodę najszybszego spadku $x_{k+1} = x_k - t_k \nabla f(x_k)$. Wówczas*

$$\|x_{k+1}\| \leq c(A, t_k)\|x_k\|, \tag{3.19}$$

gdzie

$$c(A, t_k) = \max\{|1 - t_k\lambda_{\max}(A)|, |1 - t_k\lambda_{\min}(A)|\}.$$

Ponadto współczynnik $c(A, t_k)$ jest najmniejszy dla długości kroku $t_k = t_k^* = \frac{2}{\lambda_{\max}(A)+\lambda_{\min}(A)}$ i w tym przypadku zachodzi nierówność

$$\|x_{k+1}\| \leq \frac{\mathrm{cond}(A) - 1}{\mathrm{cond}(A) + 1}\|x_k\|, \tag{3.20}$$

gdzie $\mathrm{cond}(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$ *jest wskaźnikiem uwarunkowania macierzy A.*

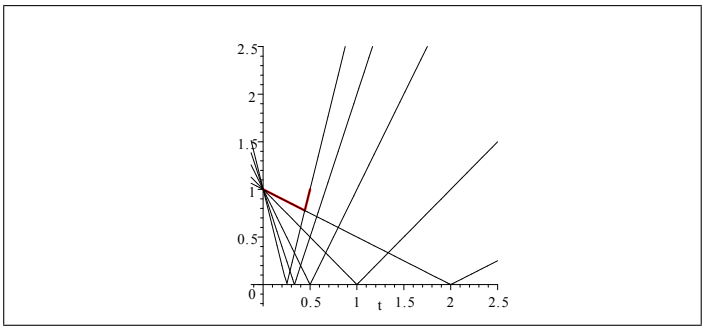
W celu udowodnienia powyższego twierdzenia będziemy potrzebować następującego lematu.

Lemat 3.5.2 *Niech A będzie macierzą symetryczną dodatnio określoną i niech $t \in \mathbb{R}$. Wówczas dla normy spektralnej macierzy $I - tA$ zachodzi nierówność*

$$\|I - tA\| \leq \max\{|1 - t\lambda_{\max}(A)|, |1 - t\lambda_{\min}(A)|\} \tag{3.21}$$

Dowód. Dla $t = 0$ nierówność jest oczywista. Niech więc $t \neq 0$. Niech $\lambda_1, ..., \lambda_n$ oznaczają wartości własne macierzy A, przy czym $0 < \lambda_1 \leq ... \leq \lambda_n$. Zauważmy przede wszystkim, że macierze A oraz $(I-tA)^2$ mają te same wektory własne oraz, że wielkości $(1-t\lambda_i)^2$, $i = 1, ..., n$, są wartościami własnymi macierzy $(I-tA)^2$. Wobec tego, korzystając z własności normy spektralnej macierzy i z prostych obserwacji (patrz rysunek 16.), otrzymujemy

$$\begin{aligned} \|I - tA\| &= \sqrt{\lambda_{\max}(I - tA)^2} \\ &= \max\{|1 - t\lambda_i|, i = 1, ..., n\} \\ &= \max\{|1 - t\lambda_1|, |1 - t\lambda_n|\}. \end{aligned}$$



Rysunek 16. Norma spektralna macierzy $I - tA$

■

Dowód. twierdzenia 3.5.1. Dla metody najszybszego spadku mamy $x_{k+1} = x_k - t_k \nabla f(x_k)$. Dla funkcji kwadratowej postaci $f(x) = \frac{1}{2}x^T Ax$ mamy więc

$$x_{k+1} = x_k - t_k A x_k.$$

Wobec tego, korzystając z lematu 3.5.2 otrzymamy

$$\begin{aligned} \|x_{k+1}\| &= \|(I - t_k A)x_k\| \leq \|I - t_k A\| \cdot \|x_k\| \\ &\leq \max\{|1 - t_k \lambda_{\max}(A)|, |1 - t_k \lambda_{\min}(A)|\} \cdot \|x_k\|. \end{aligned} \tag{3.22}$$

Zauważmy, że funkcja wypukła $h : \mathbb{R} \to \mathbb{R}$ określona wzorem

$$h(t) = \max\{|1 - t\lambda_{\max}(A)|, |1 - t\lambda_{\min}(A)|\}$$

osiąga swoje minimum dla $t = t^* = \frac{2}{\lambda_{\max}(A)+\lambda_{\min}(A)}$ równe

$$h(t^*) = \frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)}.$$

Zatem wybierając długość kroku $t_k = t^*$ otrzymamy oszacowanie

$$\begin{aligned} \|x_{k+1}\| &\leq \frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)} \|x_k\| \\ &= \frac{\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} - 1}{\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} + 1} \|x_k\| \\ &= \frac{\mathrm{cond}(A) - 1}{\mathrm{cond}(A) + 1} \|x_k\| \end{aligned}$$

■

Uwaga 3.5.3 Dla źle uwarunkowanych funkcji kwadratowych, czyli takich, dla których wartość $\mathrm{cond}(A)$ jest duża, współczynnik $c(A, t_k)$ w oszacowaniu (3.19) jest duży. Jest on bliski 1 nawet przy najlepszym doborze długości kroku t_k. Można więc przypuszczać, że zbieżność ciągu (x_k) dla takich funkcji może być bardzo wolna. Przypuszczenie to jest niestety prawdziwe, gdyż oszacowania danego nierównością (3.20) nie da się poprawić. Z dowodów lematu 3.5.2 i twierdzenia 3.5.1 wynika bowiem, że w nierówności (3.20) zachodzi równość gdy x_k jest wektorem własnym macierzy $I - t_k A$ odpowiadającym wartości własnej $\|I - t_k A\|$ i długość kroku $t_k = t^* = \frac{2}{\lambda_{\max}(A)+\lambda_{\min}(A)}$. Stwierdzenie to jest co prawda słuszne dla pojedynczej iteracji, gdyż trudno oczekiwać, że wszystkie kolejne przybliżenia x_k będą omawianymi wektorami własnymi. Poniższy przykład pokazuje jednak, że nawet dla metody Cauchy'ego współczynnik zbieżności geometrycznej może być dowolnie bliski jedności dla funkcji kwadratowej o odpowiednio dużym wskaźniku uwarunkowania.

Przykład 3.5.4 Rozpatrzmy metodę Cauchy'ego dla funkcji kwadratowej $f : \mathbb{R}^n \to \mathbb{R}$ postaci $f(x) = \frac{1}{2}x^T Ax$, gdzie A jest macierzą określoną dodatnio. Niech $\bar{x}$ oznacza pewne przybliżenie minimizera tej funkcji. Zgodnie z metodą Cauchy'ego, kolejne przybliżenie x^+ jest postaci $x^+ = \bar{x} - t\nabla f(\bar{x})$, gdzie długość kroku t jest otrzymana metodą dokładnej minimalizacji kierunkowej. Ponieważ $\nabla f(\bar{x}) = A\bar{x}$, wiec mamy

$$x^+ = \bar{x} - tA\bar{x}.$$

Wielkość t jest dobrana tak, aby $\nabla f(\bar{x})^T \nabla f(x^+) = 0$. Wyliczając wielkość t z tej równości otrzymujemy

$$t = \frac{\bar{x}^T A^2 \bar{x}}{\bar{x}^T A^3 \bar{x}}.$$

W dalszych rozważaniach ograniczymy się do przypadku $n = 2$ i $A = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$. Niech $\bar{x} = (\bar{x}_1, \bar{x}_2)$, przy czym $\bar{x}_1 \neq 0$ i $\bar{x}_2 \neq 0$. Po prostych przekształceniach otrzymamy

$$x^+ = \begin{bmatrix} x_1^+ \\ x_2^+ \end{bmatrix} = \begin{bmatrix} \frac{\lambda_2^2(\lambda_2-\lambda_1)\bar{x}_2^2}{\lambda_1^3\bar{x}_1^2+\lambda_2^3\bar{x}_2^2}\bar{\xi}_1 \\ -\frac{\lambda_1^2(\lambda_2-\lambda_1)\bar{\xi}_1^2}{\lambda_1^3\bar{x}_1^2+\lambda_2^3\bar{x}_2^2}\bar{\xi}_2 \end{bmatrix}.$$

Zatem

$$\frac{x_1^+}{x_2^+} = -(\frac{\lambda_2}{\lambda_1})^2 \frac{\bar{x}_2}{\bar{x}_1}. \tag{3.23}$$

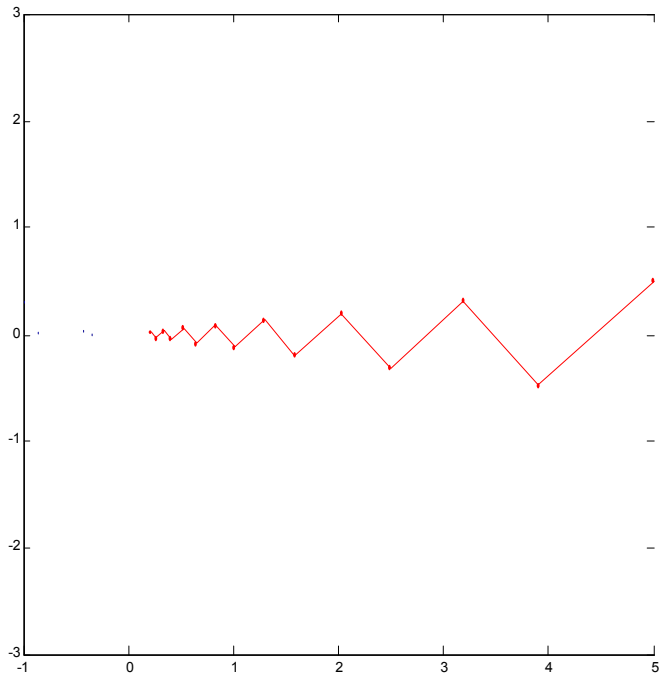
Przypuśćmy teraz, że do punktu x^+ zastosowaliśmy ponownie metodę Cauchy'ego otrzymując punkt $x^{++} = (x_1^{++}, x_2^{++})$. Analogicznie do równości (3.23) otrzymamy

$$\frac{x_1^{++}}{x_2^{++}} = -(\frac{\lambda_2}{\lambda_1})^2 \frac{x_2^+}{x_1^+}$$

(patrz rysunek 17.).W konsekwencji

$$\frac{x_1^{++}}{x_2^{++}} = \frac{\bar{x}_1}{\bar{x}_2}.$$

Wektory $\bar{x}$ i x^{++} są więc liniowo zależne (współliniowe).



Oznaczając $\alpha = \sphericalangle(\bar{x}, \nabla f(\bar{x}))$ otrzymamy

$$\begin{aligned}
\frac{\|x^{++}\|}{\|\bar{x}\|} &= \frac{\|\bar{x} - (\bar{x} - x^{++})\|}{\|\bar{x}\|} \\
&= \frac{\|\bar{x}\| - \|\bar{x} - x^{++}\|}{\|\bar{x}\|} \\
&= 1 - \frac{\|x^{++} - \bar{x}\|}{\|\bar{x}\|} \\
&= 1 - \frac{\|x^{+} - \bar{x}\|}{\|\bar{x}\| \cos\alpha} \\
&= 1 - \frac{\|x^{+} - \bar{x}\| \cdot \|\nabla f(\bar{x})\|}{\bar{x}^T \nabla f(\bar{x})} \\
&= 1 - \frac{t\|A\bar{x}\|^2}{\bar{x}^T A \bar{x}} \\
&= 1 - \frac{\bar{x}^T A^2 \bar{x}}{\bar{x}^T A^3 \bar{x}} \cdot \frac{\|A\bar{x}\|^2}{\bar{x}^T A \bar{x}}
\end{aligned}$$

Przyjmując $\bar{x} = (\lambda_2, \lambda_1)$ dostajemy

$$\frac{\|x^{++}\|}{\|\bar{x}\|} = 1 - \frac{4}{2 + \frac{\lambda_1}{\lambda_2} + \frac{\lambda_2}{\lambda_1}} = \frac{(\mathrm{cond}(A))^{-1} + \mathrm{cond}(A) - 2}{(\mathrm{cond}(A))^{-1} + \mathrm{cond}(A) + 2}.$$

Widzimy, że dla zadań źle uwarunkowanych o odpowiednio dużym wskaźniku uwarunkowania wielkość ta może być dowolnie bliska jedności. Oznacza to, że dla funkcji kwadratowych o źle uwarunkowanym hesjanie metoda Cauchy'ego może być bardzo wolno zbieżna.

W ogólnym przypadku zbieżność dowolnej metody najszybszego spadku (w szczególności metody Cauchy'ego) jest wolniejsza od liniowej. Jest to bardzo niekorzystna własność, wskutek

czego metody najszybszego spadku nie mogą być traktowane jako efektywne metody minimalizacji różniczkowalnej (niektórzy autorzy uważają nawet, że powinno się zabronić używania metod najszybszego spadku). Dla źle uwarunkowanych zadań obserwuje się zatrzymanie tej metody wskutek zadziałania kryterium zatrzymania lub wskutek wykonania dużej liczby iteracji, daleko od punktu stacjonarnego. Czasem, nawet jeśli po wykonaniu niezbyt dużej liczby iteracji wyznaczony zostanie punkt w pobliżu punktu stacjonarnego, to osiągnięcie coraz to lepszej dokładności przybliżenia może wiązać się z wykonaniem niewspółmiernie dużej liczby kolejnych iteracji. Jest to charakterystyczne dla metod posiadających zbieżność wolniejszą od liniowej.

3.5.3 Metody Newtona

Analizując metody najszybszego spadku widzimy, że ich konstrukcja wynika ze stosowania przybliżenia rzędu pierwszego, które otrzymamy rozwijając funkcję celu f we wzór Taylora w otoczeniu bieżącego przybliżenia $\bar{x}$:

$$f(\bar{x}+d) = f(\bar{x}) + d^T \nabla f(\bar{x}) + o(\|d\|).$$

W metodach najszybszego spadku kierunek poszukiwań wybieramy jako kierunek unormowany realizujący najszybszy spadek w punkcie $\bar{x}$ wartości funkcji $\bar{f} : \mathbb{R}^n \longrightarrow \mathbb{R}$,

$$\bar{f}(x) = f(\bar{x}) + (x-\bar{x})^T \nabla f(\bar{x}),$$

będącej modelem liniowym funkcji celu, czyli kierunek będący rozwiązaniem problemu

$$\begin{array}{ll} \text{minimalizować} & \Delta\bar{f}(\bar{x},d) := \bar{f}(\bar{x}+d) - \bar{f}(\bar{x}) = d^T \nabla f(\bar{x}) \\ \text{przy ograniczeniu} & \|d\| = 1. \end{array}$$

Wiadomo, że rozwiązaniem tego problemu jest $d = \nabla f(\bar{x})/\|\nabla f(\bar{x})\|$. Natomiast w *metodach Newtona* kierunek poszukiwań wybieramy w oparciu o przybliżenie rzędu drugiego funkcji celu, które otrzymamy rozwijając ją we wzór Taylora:

$$f(\bar{x}+d) = f(\bar{x}) + d^T \nabla f(\bar{x}) + \frac{1}{2} d^T \nabla^2 f(\bar{x}) d + o(\|d\|^2).$$

Pomijając składnik $o(\|d\|^2)$ i przyjmując $d = x - \bar{x}$ otrzymamy model kwadratowy $\breve{f} : \mathbb{R}^n \longrightarrow \mathbb{R}$,

$$\breve{f}(x) = f(\bar{x}) + (x-\bar{x})^T \nabla f(\bar{x}) + \frac{1}{2}(x-\bar{x})^T \nabla^2 f(\bar{x})(x-\bar{x}).$$

Następnie wyznaczamy kierunek poszukiwań d realizujący najmniejszą wartość funkcji $\breve{f}$ będącej modelem kwadratowym funkcji celu, czyli kierunek d minimalizujący $\Delta\breve{f}(\bar{x},d) := \breve{f}(\bar{x}+d) - \breve{f}(\bar{x})$. Nietrudno zauważyć, że jest to równoważne rozwiązaniu zadania

$$\begin{array}{ll} \text{minimalizować} & h(d) = \frac{1}{2} d^T G d + d^T g \\ \text{względem} & d \in \mathbb{R}^n, \end{array}$$

gdzie $g = \nabla f(\bar{x})$ i $G = \nabla^2 f(\bar{x})$. Aby zadanie to miało jednoznaczne rozwiązanie musimy założyć, że hesjan $G = \nabla^2 f(\bar{x})$ jest dodatnio określony. Korzystając z warunku koniecznego istnienia minimum $\nabla h(d) = 0$ mamy

$$\nabla h(d) = g + Gd = 0, \tag{3.24}$$

lub inaczej

$$\nabla^2 f(\bar{x}) d = -\nabla f(\bar{x}).$$

Ostatnie równanie nosi nazwę *równania Newtona*. Rozwiązaniem tego równania jest kierunek poszukiwań

$$d = -G^{-1} g$$

lub inaczej

$$d = -(\nabla^2 f(\bar{x}))^{-1} \nabla f(\bar{x}),$$

zwany kierunkiem Newtona. W praktyce kierunek ten wyznacza się nie poprzez odwracanie hesjanu G, lecz stosując rozkład Cholesky'ego tej macierzy: $G = LL^T$ i rozwiązując iteracyjnie odpowiedni układ równań. Metoda Newtona służy w istocie wyznaczeniu punktu stacjonarnego gradientu funkcji f, gdyż w istocie kierunek d wyznaczamy z warunku koniecznego istnienia minimum funkcji f.

Klasyczna metoda Newtona

W klasycznej *metodzie Newtona*, mając dane pewne przybliżenie x_k rozwiązania problemu minimalizacji, kolejne przybliżenie x_{k+1} wyznacza się zgodnie z następującym schematem

1^0 wyznaczyć rozwiązanie d_k równania $G_k d = -g_k$, gdzie $G_k = \nabla^2 f(x_k)$ i $g_k = \nabla f(x_k)$,

2^0 położyć $x_{k+1} = x_k + d_k$.

Widzimy więc, że w klasycznej metodzie Newtona nie używa się minimalizacji kierunkowej – wykonuje się mianowicie jednostkowy krok w kierunku d_k.

Zauważmy przede wszystkim, że dla funkcji kwadratowej z dodatnio określonym hesjanem klasyczna metoda Newtona jest zbieżna w jednym kroku, ponieważ dla funkcji kwadratowej f, jej model kwadratowy $\breve{f}$ pokrywa się z f. W ogólnym przypadku otrzymujemy lokalną zbieżność kwadratową, czyli zbieżność kwadratową dla punktu startowego leżącego wystarczająco blisko rozwiązania. Zanim udowodnimy twierdzenie o zbieżności dla metody Newtona podamy pewną własność funkcji różniczkowalnej. Z twierdzenia 1.5.5 wynika że dla odwzorowania różniczkowalnego F wartość $F(x)$ różni się od jej części liniowej $\bar{F}(x)$ określonej dla punktu y o $o(\|x-y\|)$. Poniższy lemat, którego dowód można znaleźć w podręczniku [SB87, ustęp 5.3] podaje oszacowanie funkcji o w przypadku, gdy macierz Jacobiego funkcji F jest odwzorowaniem lipschitzowskim.

Lemat 3.5.5 *Niech F będzie funkcją różniczkowalną w pewnym zbiorze wypukłym $C \subseteq \mathbb{R}^n$, której macierz Jacobiego jest odwzorowaniem γ-lipschitzowskim dla pewnego $\gamma > 0$, czyli dla dowolnych $x, y \in C$ zachodzi*

$$\|\nabla F(y) - \nabla F(x)\| \leq \gamma \|y - x\|.$$

Wówczas dla dowolnych $x, y \in C$

$$\|F(y) - F(x) - \nabla F(x)^T (y - x)\| \leq \frac{\gamma}{2} \|y - x\|^2.$$

Teraz możemy przejść do udowodnienia głównego wyniku tego ustępu.

Twierdzenie 3.5.6 *Niech $f \in C^2$ będzie funkcją osiągającą minimum w punkcie x^* w którym $\nabla^2 f(x^*)$ jest macierzą dodatnio określoną. Niech $\beta, \rho_1 > 0$ będą takie, że*

$$\|\nabla^2 f(x)^{-1}\| \leq \beta \text{ dla dowolnego } x \in B(x^*, \rho_1). \tag{3.25}$$

Załóżmy, że dla pewnych stałych $\gamma, \rho_2 > 0$ zachodzi

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq \gamma \|x - y\| \text{ dla dowolnych } x, y \in B(x^*, \rho_2),$$

czyli $\nabla^2 f(\cdot)$ jest operatorem γ-lipschitzowskim na $B(x^*, \rho_2)$. Niech $0 < \rho \leq \min\{\rho_1, \rho_2\}$ będzie takie, że

$$c := \beta\gamma\rho/2 < 1. \tag{3.26}$$

Wówczas operator T określony równością $T(x) = x - (\nabla^2 f(x))^{-1} \nabla f(x)$ jest c-kontrakcją względem x^* na $B(x^*, \rho)$, tzn. zachodzi

$$\|T(x) - x^*\| \leq c\|x - x^*\| \text{ dla dowolnego } x \in B(x^*, \rho).$$

W konsekwencji ciąg x^k generowany przez metodę Newtona $x_{k+1} = T(x_k)$, gdzie $x_0 \in B(x^*, \rho)$ jest zbieżny do x^*. Ponadto zbieżność ta jest kwadratowa.

Dowód. Zauważmy, że z dodatniej określoności macierzy $\nabla^2 f(x^*)$ i z ciągłości pochodnych cząstkowych rzędu drugiego funkcji f wynika, że istnieją stałe $\beta, \rho_1 > 0$ spełniające (3.25). Niech $\rho \in (0, \min\{\rho_1, \rho_2, 2/(\beta\gamma)\})$. Wówczas oczywiście zachodzi (3.26). Niech $x \in B(x^*, \rho)$. Wówczas z równości $\nabla f(x^*) = 0$, lematu 3.5.5 dla $F = \nabla f$ i $y = x^*$ wynika, że

$$\begin{aligned}
\|T(x) - x^*\| &= \|x - x^* - (\nabla^2 f(x))^{-1}\nabla f(x) = \|x - x^* - (\nabla^2 f(x))^{-1}(\nabla f(x) - \nabla f(x^*))\| \\
&= \|(\nabla^2 f(x))^{-1}[\nabla^2 f(x)(x - x^*) - (\nabla f(x) - \nabla f(x^*))]\| \\
&\leq \|(\nabla^2 f(x))^{-1}\| \cdot \|\nabla^2 f(x)(x - x^*) - (\nabla f(x) - \nabla f(x^*))\| \\
&\leq \frac{\beta\gamma}{2}\|x - x^*\|^2,
\end{aligned}$$

a więc

$$\|T(x) - x^*\| \leq \frac{\beta\gamma}{2}\|x - x^*\|^2. \tag{3.27}$$

Ponieważ $\|x - x^*\| \leq \rho$,więc

$$\|T(x) - x^*\| \leq \frac{\beta\gamma\rho}{2}\|x - x^*\| = c\|x - x^*\|$$

czyli T jest c-kontrakcją względem x^* na $B(x^*, \rho)$. Z twierdzenia 1.8.4 wynika więc, że ciąg x_k generowany przez iterację $x_{k+1} = T(x_k)$ jest zbieżny do x^*. Z nierówności (3.27) dla $x = x_k$ otrzymujemy

$$\|x_{k+1} - x^*\| \leq \frac{\beta\gamma}{2}\|x_k - x^*\|^2,$$

czyli zbieżność jest kwadratowa. ■

Przy pewnych mocniejszych założeniach można uzyskać zbieżność metody Newtona bez zakładania istnienia minimizera. Zainteresowanych odsyłam do [SB87, Twierdzenie 5.3.2].

Uwaga 3.5.7 Niech x^* będzie minimizerem funkcji f, w którym hesjan G^* jest dodatnio określony.

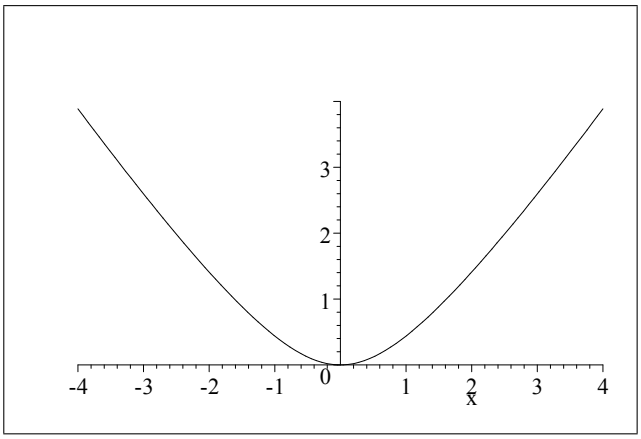
(a) Jeśli x_k jest daleko od rozwiązania x^*, to hesjan G_k nie musi być dodatnio określony (kierunek Newtona nie musi być kierunkiem spadku).

(b) Jeśli nawet hesjan G_k jest dodatnio określony (zgodnie z uwagą 3.1.3 kierunek Newtona jest wówczas kierunkiem spadku), to nie ma gwarancji spadku funkcji – może się zdarzyć, że $f(x_{k+1}) > f(x_k)$, gdyż w klasycznej metodzie Newtona nie ma minimalizacji kierunkowej.

Należy wobec powyższego podkreślić, że metoda Newtona jest tylko lokalnie zbieżna do rozwiązania, tzn. zbieżność jest zagwarantowana jedynie dla punktu startowego leżącego dostatecznie blisko minimizera, w którym hesjan jest dodatnio określony. Dla innych punktów startowych metoda Newtona może okazać się rozbieżna, bądź zbieżna do punktu nie będącego minimizerem (np. do punktu siodłowego).

Przykład 3.5.8 Zwrócimy uwagę na to, że dobór punktu startowego ma duże znaczenie dla zbieżności metody Newtona. Zastosujemy ją do minimalizacji funkcji $f(x) = x \arctan x - \frac{1}{2}\ln(x^2 + 1)$, której wykres przedstawiony jest poniżej. Jest to funkcja wypukła, której jedynym minimizerem jest $x^* = 0$. Metoda Newtona ma postać

$$x_{k+1} = x_k - (f''(x_k))^{-1} f'(x_k) = x_k - (x_k^2 + 1)\arctan x_k.$$



Dla punktów startowych $x_0 = 1$ i $x_0 = 2$ otrzymamy wyniki umieszczone w poniższej tabeli

x_0	1	2
x_1	-0.57080	-3.5357
x_2	0.11686	13.951
x_3	-1.061×10^{-3}	-279.35
x_4	-1.5239×10^{-9}	1.2202×10^{5}

Dla $x_0 = 1$ ciąg x_k jest bardzo szybko zbieżny, ponieważ dla $\rho_1 = \rho_2 = \rho = 1$ można przyjąć $\beta = 2$, $\gamma = 0.61$ i wobec tego $c = 0.6$ spełnia założenie (3.26), natomiast dla $x_0 = 2$ i dla $\rho_1 = \rho_2 = \rho = 2$ można przyjąć $\beta = 5$ i $\gamma = 0.61$ i widzimy, że stała $c = 3$ nie spełnia założenia (3.26).

Zmodyfikowane metody Newtona

W *zmodyfikowanych metodach Newtona* kierunek poszukiwań wyznacza się tak samo jak w klasycznej metodzie Newtona, natomiast następnie dokonuje się minimalizacji kierunkowej. Jeśli hesjan G_k jest dodatnio określony, to kierunek poszukiwań jest kierunkiem spadku (patrz uwaga 3.1.3). Zgodnie z twierdzeniem 3.4.2, dla zagwarantowania zbieżności do rozwiązania wystarczy, aby był spełniony warunek (3.7) oraz metoda minimalizacji kierunkowej była efektywna. Dla spełnienia pierwszego z tych warunków wystarczy na mocy lematu 3.4.1, aby wartości własne hesjanu funkcji celu były ograniczone z dołu i z góry przez pewne dodatnie stałe, zaś dla spełnienia drugiego z nich wystarczy, zgodnie z twierdzeniem 3.4.7, zastosować do minimalizacji kierunkowej na przykład rozszerzoną metodę Armijo. Natomiast jeśli pierwszy z tych warunków nie jest spełniony lub jego sprawdzenie jest kłopotliwe, można zastosować nieco inną modyfikację metody Newtona. W modyfikacji tej w iteracjach, w których nie jest spełniony warunek (3.7), kierunek Newtona zastępuje się kierunkiem najszybszego spadku. Wówczas dla zagwarantowania zbieżności wystarczy tylko zastosować jakąkolwiek efektywną metodę minimalizacji kierunkowej. Zbieżność tak zmodyfikowanej metody Newtona wynika również z twierdzenia 3.4.2.

Metoda Levenberga–Marquardta

Jak zauważyliśmy w poprzednich ustępach metoda Newtona i jej modyfikacje są zbieżne do rozwiązania, jeśli punkt x_k jest dostatecznie blisko rozwiązania x^* i hesjan w punkcie x^* jest dodatnio określony. W przeciwnym wypadku nie mamy gwarancji, że kierunek poszukiwań $s_k = -G_k^{-1} g_k$ jest kierunkiem spadku. Wówczas zamiast hesjanu G_k można użyć macierzy $H_k = G_k + \varepsilon_k I$ dla odpowiednio dobranego $\varepsilon_k > 0$ tak, aby macierz H_k była dodatnio określona. Zauważmy, że wystarczy, aby ε_k było większe niż najmniejsza wartość własna macierzy G_k. W

ten sposób określa się *metodę Levenberga–Marquardta*. W celu zagwarantowania zbieżności do rozwiązania należy postępować podobnie jak opisano to w ustępie 3.5.3.

3.5.4 Metody kierunków sprzężonych

Ogólna postać metod kierunków sprzężonych

Definicja 3.5.9 Niezerowe wektory $s_1, ...s_n$ nazywają się sprzężone (względem dodatnio określonej macierzy G typu $n \times n$), jeśli

$$s_i^T G s_j = 0 \tag{3.28}$$

dla $i \neq j$, $i, j = 1, ..., n$.

Uwaga 3.5.10 Kierunki sprzężone są wektorami ortogonalnymi w przestrzeni $\mathbb{R}^n$ z iloczynem skalarnym $\langle x, y \rangle = x^T G y$.

Ćwiczenie 3.5.11 Podać postać kierunków sprzężonych względem macierzy $G = \mathrm{diag}(1, 2)$.

Rozważmy teraz zadanie minimalizacji funkcji kwadratowej f z dodatnio określonym hesjanem G. *Metoda kierunków sprzężonych* dla tego zadania jest to metoda spadku generująca kierunki poszukiwań będące kierunkami sprzężonymi względem hesjanu G.

Twierdzenie 3.5.12 *Metoda kierunków sprzężonych z dokładną minimalizacją kierunkową zastosowana do funkcji kwadratowej $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ z dodatnio określonym hesjanem G wyznacza minimum po co najwyżej n iteracjach. Dokładniej*

$$x_{i+1} = \mathrm{argmin}\{f(x) : x \in x_1 + \mathrm{Lin}\{s_1, ..., s_i\}\},$$

$i = 1, ..., n$

Dowód. Ponieważ macierz G jest nieosobliwa, więc kierunki sprzężone są liniowo niezależne (pokazać na ćwiczeniach). Pokażemy teraz przez indukcję względem i, $i = 1, ..., n$, że

$$g_{i+1}^T s_j = 0$$

dla $j = 1, ..., i$. Dla $i = 1$ równość ta jest konsekwencją stosowania dokładnej minimalizacji kierunkowej. Przypuśćmy więc, że $g_i^T s_j = 0$ dla $j = 1, .., i-1$. Znowu na mocy stosowania dokładnej minimalizacji kierunkowej mamy $g_{i+1}^T s_i = 0$. Ponadto, ze związku

$$g_{i+1} - g_i = G(x_{i+1} - x_i) = t_i G s_i,$$

prawdziwego dla funkcji kwadratowej (patrz uwaga 1.5.22), mamy na mocy założenia indukcyjnego i na mocy równości (3.28)

$$g_{i+1}^T s_j = (g_i + t_i G s_i)^T s_j = \underbrace{g_i^T s_j}_{=0} + t_i \underbrace{s_i^T G s_j}_{=0} = 0$$

dla $j = 1, ..., i-1$. Tak więc $g_{i+1}^T s_j = 0$ dla $j = 1, ..., i$, czyli funkcja f ma w punkcie x_{i+1} zerowe pochodne kierunkowe w i liniowo niezależnych kierunkach $s_1, ..., s_i$. W konsekwencji, dla dowolnego kierunku $d \in \mathrm{Lin}\{s_1, ..., s_i\}$

$$f'(x_{i+1}, d) = 0$$

(pokazać). Oznaczmy

$$D = x_{i+1} + \mathrm{Lin}\{s_1, ..., s_i\}.$$

Z równości $x_{k+1} = x_k + t_k s_k$, $k = 1, ..., i$, wynika, że

$$D = x_1 + \mathrm{Lin}\{s_1, ..., s_i\}$$

(pokazać). Niech $h = f\,|_D$. Oczywiście funkcja h jest wypukła, gdyż f jest wypukła (hesjan G jest dodatnio określony). Ponieważ zerowanie się wszystkich pochodnych kierunkowych jest warunkiem wystarczającym istnienia minimum dla funkcji wypukłej, więc

$$x_{i+1} = \mathrm{argmin}_{x \in D}\, h(x) = \mathrm{argmin}\{f(x) : x \in x_1 + \mathrm{Lin}\{s_1, ..., s_i\}\}.$$

■

Metoda gradientów sprzężonych dla kwadratowej funkcji celu

Opiszemy tę metodę najpierw dla funkcji kwadratowej f z dodatnio określonym hesjanem G. W tym przypadku inicjalizacja następuje kierunkiem najszybszego spadku:

$$s_1 = -g_1. \tag{3.29}$$

Mając dany kierunek poszukiwań s_k otrzymany w k-tym kroku, kolejny kierunek poszukiwań wyznaczamy zgodnie ze wzorem

$$s_{k+1} = -g_{k+1} + \beta_k s_k, \tag{3.30}$$

gdzie $\beta \in \mathbb{R}$ jest dobrane tak, aby kierunek s_{k+1}był sprzężony do $s_1, ..., s_k$. Pokażemy później, że tak będzie jeśli przyjąć

$$\beta_k = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}. \tag{3.31}$$

Wzory powyższe opisują *metodę gradientów sprzężonych Fletchera–Reevesa* (1964).

Twierdzenie 3.5.13 *Metoda Fletchera–Reevesa z dokładną minimalizacją kierunkową zastosowana do funkcji kwadratowej z dodatnio określonym hesjanem G generuje kierunki sprzężone względem G:*

$$s_i^T G s_j = 0, \tag{3.32}$$

$j = 1, ..., i-1$, $i = 1, ..., n$. *Ponadto dla* $i \leq m = \max\{i : g_i \neq 0\}$ *zachodzą równości*

$$g_i^T g_j = 0, \tag{3.33}$$

$j = 1, ..., i-1$, i

$$s_i^T g_i = -g_i^T g_i \tag{3.34}$$

$i = 1, ..., n$

Dowód. Jeśli $m = 0$, to punkt startowy x_1 jest rozwiązaniem zadania minimalizacji. Niech więc $m \geq 1$. Pokażemy równości występujące w twierdzeniu przez indukcję względem i.

1^0 dla $i = 1$ tylko ostatnia równość wchodzi w grę, która jest oczywista z uwagi na inicjalizację kierunkiem najszybszego spadku.

2^0 Przypuśćmy, że równości w twierdzeniu zachodzą dla pewnego $i < m$. Pokażemy, że zachodzą one również, gdy zastąpimy i przez $i + 1$.

Dla funkcji kwadratowej f zachodzi równość

$$g_{i+1} - g_i = G(x_{i+1} - x_i)$$

(patrz uwaga 1.5.22), więc

$$g_{i+1} - g_i = t_i G s_i. \tag{3.35}$$

Zatem, wobec stosowania dokładnej minimalizacji kierunkowej, mamy

$$0 = g_{i+1}^T s_i = g_i^T s_i + t_i s_i^T G s_i,$$

czyli na mocy założenia indukcyjnego

$$t_i = -\frac{g_i^T s_i}{s_i^T G s_i} = \frac{g_i^T g_i}{s_i^T G s_i}. \tag{3.36}$$

Zatem dla $j < i$ mamy na mocy założenia indukcyjnego i równości (3.35)

$$g_{i+1}^T g_j = g_i^T g_j + t_i s_i^T G g_j = g_i^T g_j - t_i s_i^T G(s_j - \beta_{j-1} s_{j-1}) =$$
$$\underbrace{g_i^T g_j}_{=0} - t_i \underbrace{s_i^T G s_j}_{=0} + t_i \beta_{j-1} \underbrace{s_i^T G s_{j-1}}_{=0} = 0.$$

Natomiast dla $j = i$ mamy po podobnych przekształceniach i po skorzystaniu z równości (3.36)

$$g_{i+1}^T g_j = g_{i+1}^T g_i = g_i^T g_i - t_i s_i^T G s_i + t_i \beta_{i-1} \underbrace{s_i^T G s_{i-1}}_{=0} =$$
$$g_i^T g_i - \frac{g_i^T g_i}{s_i^T G s_i} s_i^T G s_i = 0.$$

Otrzymaliśmy więc

$$g_{i+1}^T g_j = 0 \tag{3.37}$$

dla $j = 1, ..., i$.

Dalej mamy na mocy (3.30), (3.35) i wobec stosowania dokładnej minimalizacji kierunkowej

$$s_{i+1}^T G s_j = -g_{i+1}^T G s_j + \beta_i s_i^T G s_j = \frac{1}{t_j} \underbrace{g_{i+1}^T (g_j}_{=0} - g_{j+1}) + \beta_i s_i^T G s_j =$$
$$-\frac{1}{t_j} g_{i+1}^T g_{j+1} + \beta_i s_i^T G s_j$$

Zatem dla $j = i$ otrzymamy na mocy (3.31) i (3.36)

$$s_{i+1}^T G s_i = -\frac{1}{t_i} g_{i+1}^T g_{i+1} + \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i} s_i^T G s_i =$$
$$-\frac{1}{t_i} g_{i+1}^T g_{i+1} + \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i} \cdot \frac{g_i^T g_i}{t_i} = 0.$$

Natomiast dla $j < i$ otrzymamy po podobnych przekształceniach oraz na mocy (3.37) i założenia indukcyjnego

$$s_{i+1}^T G s_j = -\frac{1}{t_j} \underbrace{g_{i+1}^T g_{j+1}}_{=0} + \beta_i \underbrace{s_i^T G s_j}_{=0} = 0.$$

Otrzymaliśmy więc

$$s_{i+1}^T G s_j = 0$$

dla $j = 1, ..., i$.

Ponadto, z definicji metody gradientów sprzężonych i z uwagi na stosowanie dokładnej minimalizacji kierunkowej, mamy

$$s_{i+1}^T g_{i+1} = (-g_{i+1} + \beta_i s_i)^T g_{i+1} = -g_{i+1}^T g_{i+1} + \beta_i \underbrace{s_i^T g_{i+1}}_{=0} = -g_{i+1}^T g_{i+1},$$

czyli

$$s_{i+1}^T g_{i+1} = -g_{i+1}^T g_{i+1}.$$

Pokazaliśmy więc, że równości w twierdzeniu zachodzą, gdy zastąpić i przez $i+1$, co kończy dowód indukcyjny.

■

Uwaga 3.5.14 Zauważmy, że w przypadku kwadratowej funkcji celu z dodatnio określonym hesjanem dokładna minimalizacja kierunkowa w metodzie gradientów sprzężonych realizowana jest przez bezpośrednie stosowanie wzoru (3.36).

Metoda gradientów sprzężonych dla dowolnej funkcji celu

W przypadku dowolnej funkcji celu bezpośrednie stosowanie wzorów (3.29)-(3.31) nie daje gwarancji otrzymania kierunku spadku. Trudno byłoby również stosować dokładną minimalizację kierunkową. W tej sytuacji stosuje się różne modyfikacje metody Fletchera–Reevesa. Najbardziej popularnym podejściem jest zastąpienie wzoru (3.31) wzorem

$$\beta_k = -\frac{g_{k+1}^T g_{k+1}}{g_k^T s_k} \tag{3.38}$$

(który z uwagi na równość (3.34) jest dla funkcji kwadratowej równoważny wzorowi (3.31)) oraz zastąpienie dokładnej minimalizacji kierunkowej - przybliżoną - z kryterium zatrzymania Wolfe'a–Powella (porównaj warunek (3.6):

$$q(t) \leq q(0) + mtq'(0)$$

i

$$\left|\frac{q'(t)}{q'(0)}\right| \leq m$$

Mamy wówczas

$$g_{k+1}^T s_{k+1} = -g_{k+1}^T g_{k+1} + \beta_k g_{k+1}^T s_k =$$

$$-g_{k+1}^T g_{k+1} + g_{k+1}^T g_{k+1} \cdot \underbrace{\frac{g_{k+1}^T s_k}{-g_k^T s_k}}_{=\frac{q'(t_k)}{-q'(0)} \leq m} \leq (m-1) g_{k+1}^T g_{k+1} < 0,$$

czyli kierunek s_{k+1} jest kierunkiem spadku w punkcie x_{k+1}. Dla zagwarantowania zbieżności metodę tę stosuje się z tzw. ciepłym restartem: co n iteracji powtarzamy opisaną wyżej metodę gradientów sprzężonych dla punktu startowego będącego ostatnio otrzymanym przybliżeniem. Innymi słowy, możemy powiedzieć, że w metodzie gradientów sprzężonych modyfikujemy iteracje o numerach $kn+1$ dla $k = 1, 2, ...$, zastępując wyznaczany tam kierunek kierunkiem najszybszego spadku, czyli

$$s_{kn+1} = -g_{kn+1}$$

Wówczas zbieżność tej zmodyfikowanej metody wynika z uwagi 3.4.3.

3.5.5 Metody quasi-Newtona

Wady metod Newtona:

- konieczność wyznaczania hesjanu G_k w każdym kroku
- konieczność wyznaczania odwrotności hesjanu $(G_k)^{-1}$ w każdym kroku ($\tilde{}n^3$ operacji)
- brak gwarancji dodatniej określoności hesjanu G_k dla podstawowej metody Newtona
- brak gwarancji monotoniczności ($f(x_{k+1}) < f(x_k)$) dla podstawowej metody Newtona
- lokalna zbieżność

Metody quasi-Newtona różnią się od metod Newtona tym, że zamiast odwrotności hesjanu $(G_k)^{-1}$ bierzemy w nich pewne przybliżenie tej odwrotności H_k. Przy konstrukcji macierzy H_k korzystamy z informacji o wartościach funkcji i gradientach uzyskanych w poprzednich iteracjach. Nie korzystamy natomiast z pochodnych cząstkowych rzędu drugiego.

Schemat iteracyjny metod quasi-Newtona

Krok 0 (*inicjalizacja*). Wybrać punkt startowy $x_1 \in \mathbb{R}^n$ i dowolną macierz H_1 dodatnio określona typu $n \times n$, położyć $k = 1$.

Krok 1 (*kierunek poszukiwań*). Wyznaczyć kierunek poszukiwań zgodnie ze wzorem

$$s_k = -H_k g_k.$$

Krok 2 (*minimalizacja kierunkowa*). Wyznaczyć $t_k > 0$ metodą minimalizacji kierunkowej i kolejne przybliżenie x_{k+1} zgodnie ze wzorem

$$x_{k+1} = x_k + t_k s_k.$$

Krok 3 (*aktualizacja macierzy H_k*). Wyznaczyć kolejną macierz H_{k+1} zgodnie ze wzorem

$$H_{k+1} = h(H_k)$$

dla pewnego odwzorowania h.

Krok 4. Położyć $k := k + 1$ i przejść do kroku 1.

Uwaga 3.5.15 Wydaje się rozsądne postawienie metodom quasi-Newtona następujących wymagań:

- aktualizacja macierzy H_k zachowuje dodatnią określoność,
- zaktualizowana macierz H_{k+1} powinna "lepiej" przybliżać $(G_k)^{-1}$ aniżeli H_k,
- zaktualizowana macierz H_{k+1} powinna "niewiele" różnić się od H_k,
- metoda powinna szybko zbiegać do rozwiązania w przypadku kwadratowej funkcji celu.

Oznaczmy

$$d_k = x_{k+1} - x_k$$

i

$$\gamma_k = g_{k+1} - g_k.$$

Z rozwinięcia Taylora dla gradientu w otoczeniu punktu x_k mamy

$$\gamma_k = G_k d_k + o(\|d_k\|).$$

Opuszczając składnik $o(\|d_k\|)$ otrzymamy

$$\gamma_k \approx G_k d_k,$$

czyli hesjan G_k przekształca w przybliżeniu przyrost argumentu d_k w przyrost gradientu γ_k. Jeśli macierz symetryczną H_{k+1} określimy teraz tak, aby

$$H_{k+1}\gamma_k = d_k, \tag{3.39}$$

to kolejny kierunek $s_{k+1} = -H_{k+1}g_{k+1}$ możemy interpretować jako kierunek taki jak w podstawowej metodzie Newtona, w której hesjan w punkcie x_{k+1} został zastąpiony hesjanem w punkcie x_k. Zatem w tym sensie macierz H_{k+1} jest "bliska" odwrotności hesjanu $(G_k)^{-1}$. Równość (3.39) nosi nazwę warunku quasi-Newtona. Dowolna macierz spełniająca ten warunek jest więc rozwiązaniem układu n równań z $\frac{1}{2}n(n+1)$ niewiadomymi (macierz H_{k+1} ma być symetryczna). Oczywiście takich rozwiązań jest wiele. Zauważmy jednak, że w ostatniej iteracji otrzymaliśmy informację o zachowaniu się rzędu drugiego funkcji celu w ustalonym kierunku: przyrost gradientu $\gamma_k = g_{k+1} - g_k$ odpowiadający przyrostowi argumentu $d_k = x_{k+1} - x_k$. Ponieważ zaktualizowana macierz H_{k+1} ma coraz lepiej odzwierciedlać zachowanie rzędu drugiego funkcji celu (ma w pewnym sensie przybliżać odwrotność jej hesjanu) na podstawie bieżącej macierzy H_k i dodatkowo uzyskanej informacji w punkcie x_{k+1} (gradient g_{k+1}), więc rozsądnym wydaje się żądanie, aby H_{k+1} różniła się od H_k o macierz niskiego rzędu.

Ćwiczenie 3.5.16 Pokazać, że macierz A typu $n \times n$ jest macierzą
a) rzędu pierwszego wtedy i tylko wtedy, gdy

$$A = uv^T$$

dla pewnych niezerowych wektorów $u, v \in \mathbb{R}^n$,
b) symetryczną rzędu pierwszego wtedy i tylko wtedy, gdy

$$A = \alpha uu^T$$

dla pewnego niezerowego (liniowo niezależnego) wektora $u \in \mathbb{R}^n$ i pewnej stałej $\alpha \neq 0$,
c) symetryczną rzędu drugiego wtedy i tylko wtedy, gdy

$$A = \alpha uu^T + \beta vv^T$$

dla pewnych liniowo niezależnych wektorów $u, v \in \mathbb{R}^n$ i niezerowych stałych α i β.

Definicja 3.5.17 Mówimy, że macierz H^+ jest (*symetryczną*) *aktualizacją rzędu pierwszego* (*drugiego*) macierzy H, jeśli H^+ różni się od macierzy H o (symetryczną) macierz rzędu pierwszego (drugiego)

Uwaga 3.5.18 Symetryczna aktualizacja rzędu pierwszego ma postać

$$H^+ = H + \alpha uu^T,$$

gdzie u jest wektorem niezerowym, zaś symetryczna aktualizacja rzędu drugiego ma postać

$$H^+ = H + \alpha uu^T + \beta vv^T,$$

gdzie u, v są wektorami liniowo niezależnymi.

Dla uproszczenia zapisu oznaczać będziemy $d = d_k$, $\gamma = \gamma_k$, $H = H_k$ i $H^+ = H_{k+1}$.

Twierdzenie 3.5.19 *Symetryczna aktualizacja rzędu pierwszego zachowująca warunek quasi-Newtona ma postać*

$$H^+ = H + \frac{1}{(d - H\gamma)^T\gamma}(d - H\gamma)(d - H\gamma)^T. \tag{3.40}$$

Uwaga 3.5.20 Aktualizacja dana wzorem (3.40) posiada następujące własności

- spełnia warunek quasi-Newtona (wynika to z powyższego twierdzenia)
- jeśli kolejne kierunki poszukiwań $d_1, ..., d_n$ są liniowo niezależne, to metoda quasi-Newtona z poprawką rzędu pierwszego dana wzorem (3.40) zastosowana do kwadratowej funkcji celu z dodatnio określonym hesjanem G osiąga rozwiązanie po co najwyżej n iteracjach i $H_{n+1} = G^{-1}$.

Niestety, aktualizacja rzędu pierwszego nie zachowuje dodatniej określoności.

Metoda Davidona–Fletchera–Powella (DFP)

Ponieważ istnieje dokładnie jedna symetryczna aktualizacja rzędu pierwszego zachowująca warunek quasi-Newtona, a nie zachowuje ona dodatniej określoności, więc do naszych dalszych rozważań dopuścimy symetryczne aktualizacje rzędu drugiego. Jedną z nich jest tzw. aktualizacja *Davidona–Fletchera–Powella* (*DFP*) dana wzorem

$$H^+_{DFP} = H + \frac{1}{d^T\gamma}dd^T - \frac{1}{\gamma^T H\gamma}H\gamma\gamma^T H \tag{3.41}$$

Metoda quasi-Newtona z aktualizacją DFP nazywa się *metodą DFP*.

Uwaga 3.5.21 Stosując w metodzie DFP minimalizację kierunkową z kryterium zatrzymania Wolfe'a mamy zagwarantowane spełnienie warunku

$$d^T\gamma > 0.$$

Istotnie:

$$d^T\gamma = ts^T(g^+ - g) = t(s^T g^+ - s^T g) = t(q'(t) - q'(0)) \geq$$
$$t(m'q'(0) - q'(0)) = t\underbrace{(m' - 1)}_{<0}\underbrace{q'(0)}_{<0} > 0.$$

Jeśli ponadto H jest dodatnio określona, to oczywiście $\gamma^T H\gamma > 0$ i formuła DFP jest dobrze określona.

Uwaga 3.5.22 Metoda DFP ma następujące własności:

- spełnia warunek quasi-Newtona (pokazać na ćwiczeniach),
- zastosowana do funkcji kwadratowej z dodatnio określonym hesjanem G z dokładną minimalizacją kierunkową wyznacza sprzężone kierunki poszukiwań (a w przypadku $H_1 = I$ – gradienty sprzężone); w konsekwencji wyznacza rozwiązanie po co najwyżej n iteracjach, przy czym $H_{k+1} = G^{-1}$),
- jeśli $d^T\gamma > 0$ (na przykład przy stosowaniu w minimalizacji kierunkowej kryterium zatrzymania Wolfe'a), to metoda DFP zachowuje dodatnią określoność macierzy H_k (będzie pokazane) w konsekwencji $s_k = -H_k g_k$ jest kierunkiem spadku,

- aktualizacja wymaga jedynie ˜$3n^2$ operacji (w przeciwieństwie do odwracania hesjanu, które wymaga ˜n^3 operacji),
- zastosowana do funkcji ściśle wypukłych z dokładną minimalizacją kierunkową metoda DFP jest zbieżna globalnie,
- zbieżność do rozwiązania jest szybsza niż liniowa.

Twierdzenie 3.5.23 *Jeśli $d^T\gamma > 0$, to formuła DFP zachowuje dodatnią określoność macierzy H.*

Dowód. Załóżmy, że macierz H jest dodatnio określona. Niech $H = LL^T$ będzie rozkładem Cholesky'ego macierzy H (który istnieje na mocy dodatniej określoności). Oznaczając $a = L^T s$, $b = L^T\gamma$ mamy na mocy nierówności Schwarza

$$s^T H^+ s =$$

$$(L^T s)^T L^T s + \frac{1}{d^T\gamma} s^T d (s^T d)^T - \frac{1}{(L^T\gamma)^T L^T\gamma}(L^T s)^T L^T\gamma (L^T\gamma)^T L^T s =$$

$$a^T a - \frac{(a^T b)^2}{b^T b} + \frac{(s^T d)^2}{d^T\gamma} \geq a^T a - \frac{a^T a \cdot b^T b}{b^T b} + \frac{(s^T d)^2}{d^T\gamma} =$$

$$\frac{(s^T d)^2}{d^T\gamma} \geq 0.$$

Ponadto przynajmniej jedna z powyższych nierówności jest ostra. Jeśli bowiem pierwsza z nich jest równością, to a i b są liniowo zależne: $b = \alpha a$. Wówczas oczywiście $\gamma = \alpha s$ ($\alpha \neq 0$, bo w przeciwnym wypadku $d^T\gamma = 0$) i

$$s^T d = \frac{1}{\alpha}\gamma^T d \neq 0.$$

■

Metoda Broydena–Fletchera–Goldfarba–Shanno (BFGS)

Inną aktualizacją symetryczną rzędu drugiego jest tzw. *aktualizacja Broydena–Fletchera–Goldfarba–Shanno* (*BFGS*) dana wzorem

$$H^+_{BFGS} = H + (1 + \frac{\gamma^T H\gamma}{d^T\gamma})\frac{1}{d^T\gamma} dd^T - \frac{1}{d^T\gamma}(d\gamma^T H + H\gamma d^T). \tag{3.42}$$

Metoda quasi-Newtona z aktualizacją BFGS nazywa się *metodą BFGS*.

Uwaga 3.5.24 Metoda BFGS ma te same własności co metoda DFP. Ponadto istnieje następujący związek między obydwiema metodami: Jeśli zamienimy rolami d i γ, to formuła BFGS staje się formułą DFP dla $(H^+)^{-1}$, tzn.

$$(H^+_{DFP})^{-1} = H^{-1} + (1 + \frac{\gamma^T H^{-1}\gamma}{\gamma^T d})\frac{1}{\gamma^T d}\gamma\gamma^T - \frac{1}{\gamma^T d}(\gamma d^T H^{-1} + H^{-1} d\gamma^T)$$

oraz formuła DFP staje się formułą BFGS dla $(H^+)^{-1}$, tzn.

$$(H^+_{BFGS})^{-1} = H^{-1} + \frac{1}{\gamma^T d}\gamma\gamma^T - \frac{1}{d^T H^{-1} d} H^{-1} dd^T H^{-1}.$$