

Optymalizacja ciągła

4. Metody kierunków poprawy (metoda spadku wzdłuż gradientu)

Wojciech Kotłowski

Instytut Informatyki PP

<http://www.cs.put.poznan.pl/wkotlowski/>

21.03.2019

Plan wykładu

Minimalizacja **różniczkowalnej** funkcji $f(x)$ **bez ograniczeń**:

$$\min_{x \in \mathbb{R}^n} f(x)$$

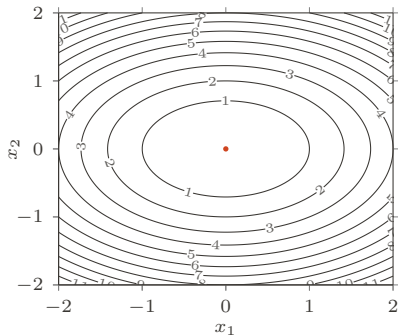
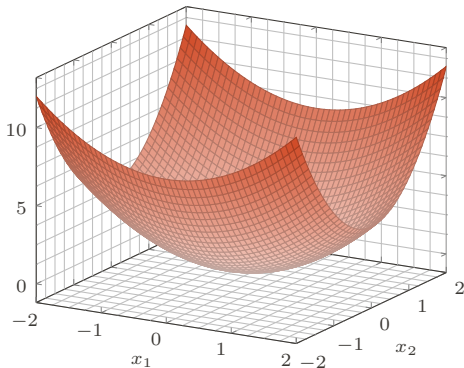
Często będziemy zakładać **wypukłość** $f(x)$

1. Wprowadzenie
2. Metoda spadku wzdłuż gradientu
3. Metoda Newtona-Raphsona (następny wykład)
4. Modyfikacje metody Newtona-Raphsona (następny wykład)
5. Metoda gradientów sprzężonych (następny wykład)

Wprowadzenie

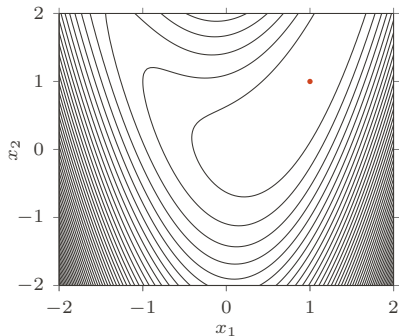
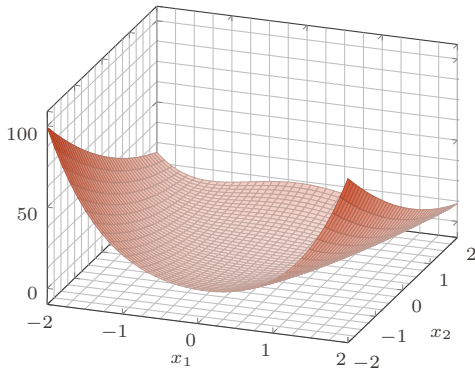
Wykres powierzchniowy a wykres konturowy

$$f(x_1, x_2) = x_1^2 + 2x_2^2$$



Wykres powierzchniowy a wykres konturowy

$$f(x_1, x_2) = 2.5(x_1^2 - x_2)^2 + (1 - x_1)^2$$



Wykres konturowy funkcji kwadratowej

Rozważmy funkcję kwadratową z dodatnio określoną macierzą $\mathbf{A}$:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$$

Wykres konturowy funkcji kwadratowej

Rozważmy funkcję kwadratową z dodatnio określoną macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$$

Przybliżając (dokładnym!) wielomianem Taylora drugiego rzędu wokół minimum $\mathbf{x}^*$ możemy zapisać funkcję jako:

$$f(\mathbf{x}) = f(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^\top \mathbf{A}(\mathbf{x} - \mathbf{x}^*)$$

Wykres konturowy funkcji kwadratowej

Rozważmy funkcję kwadratową z dodatnio określoną macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$$

Przybliżając (dokładnym!) wielomianem Taylora drugiego rzędu wokół minimum $\mathbf{x}^*$ możemy zapisać funkcję jako:

$$f(\mathbf{x}) = f(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^\top \mathbf{A}(\mathbf{x} - \mathbf{x}^*)$$

Przykład: $f(x) = x^2 + 2x + 2$

Wykres konturowy funkcji kwadratowej

Rozważmy funkcję kwadratową z dodatnio określoną macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$$

Przybliżając (dokładnym!) wielomianem Taylora drugiego rzędu wokół minimum $\mathbf{x}^*$ możemy zapisać funkcję jako:

$$f(\mathbf{x}) = f(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^\top \mathbf{A}(\mathbf{x} - \mathbf{x}^*)$$

Przykład: $f(x) = x^2 + 2x + 2$

Minimum w $x^* = -\frac{b}{2a} = -1$, $f(x^*) = 1$, a stąd $f(x) = 1 + (x + 1)^2$

Wykres konturowy funkcji kwadratowej

Rozważmy funkcję kwadratową z dodatnio określoną macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$$

Przybliżając (dokładnym!) wielomianem Taylora drugiego rzędu wokół minimum $\mathbf{x}^*$ możemy zapisać funkcję jako:

$$f(\mathbf{x}) = f(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^\top \mathbf{A}(\mathbf{x} - \mathbf{x}^*)$$

Przykład: $f(x) = x^2 + 2x + 2$

Minimum w $x^* = -\frac{b}{2a} = -1$, $f(x^*) = 1$, a stąd $f(x) = 1 + (x + 1)^2$

Przykład: $f(\mathbf{x}) = x_1^2 + 2x_2^2 - 2x_1 + 4x_2 + 1$

Wykres konturowy funkcji kwadratowej

Rozważmy funkcję kwadratową z dodatnio określoną macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$$

Przybliżając (dokładnym!) wielomianem Taylora drugiego rzędu wokół minimum $\mathbf{x}^*$ możemy zapisać funkcję jako:

$$f(\mathbf{x}) = f(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^\top \mathbf{A}(\mathbf{x} - \mathbf{x}^*)$$

Przykład: $f(x) = x^2 + 2x + 2$

Minimum w $x^* = -\frac{b}{2a} = -1$, $f(x^*) = 1$, a stąd $f(x) = 1 + (x + 1)^2$

Przykład: $f(\mathbf{x}) = x_1^2 + 2x_2^2 - 2x_1 + 4x_2 + 1$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad \mathbf{b} = (-2, 4), \quad \mathbf{x}^* = -\frac{1}{2} \mathbf{A}^{-1} \mathbf{b} = -\frac{1}{2} \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} -2 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Wykres konturowy funkcji kwadratowej

Rozważmy funkcję kwadratową z dodatnio określoną macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$$

Przybliżając (dokładnym!) wielomianem Taylora drugiego rzędu wokół minimum $\mathbf{x}^*$ możemy zapisać funkcję jako:

$$f(\mathbf{x}) = f(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^\top \mathbf{A}(\mathbf{x} - \mathbf{x}^*)$$

Przykład: $f(x) = x^2 + 2x + 2$

Minimum w $x^* = -\frac{b}{2a} = -1$, $f(x^*) = 1$, a stąd $f(x) = 1 + (x + 1)^2$

Przykład: $f(\mathbf{x}) = x_1^2 + 2x_2^2 - 2x_1 + 4x_2 + 1$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad \mathbf{b} = (-2, 4), \quad \mathbf{x}^* = -\frac{1}{2} \mathbf{A}^{-1} \mathbf{b} = -\frac{1}{2} \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} -2 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Mamy $f(\mathbf{x}^*) = -2$, a stąd:

$$f(\mathbf{x}) = -2 + \begin{bmatrix} x_1 - 1 & x_2 + 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 - 1 \\ x_2 + 1 \end{bmatrix} = -2 + (x_1 - 1)^2 + 2(x_2 + 1)^2$$

Wykres konturowy funkcji kwadratowej

$$f(\mathbf{x}) = f(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^\top \mathbf{A}(\mathbf{x} - \mathbf{x}^*)$$

Wykres konturowy funkcji kwadratowej

$$f(\mathbf{x}) = f(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^\top \mathbf{A}(\mathbf{x} - \mathbf{x}^*)$$

Równanie:

$$(\mathbf{x} - \mathbf{x}^*)^\top \mathbf{A}(\mathbf{x} - \mathbf{x}^*) = c$$

jest równaniem **elipsoidy** o środku $\mathbf{x}^*$, której osie główne odpowiadają **wektorom własnym** $\mathbf{A}$, a długości tych osi są odwrotnie proporcjonalne do pierwiastków **wartości własnych**

W szczególności dla $n = 2$ jest to równanie **elipsy** na płaszczyźnie

Wykres konturowy funkcji kwadratowej

$$f(\mathbf{x}) = f(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^\top \mathbf{A}(\mathbf{x} - \mathbf{x}^*)$$

Równanie:

$$(\mathbf{x} - \mathbf{x}^*)^\top \mathbf{A}(\mathbf{x} - \mathbf{x}^*) = c$$

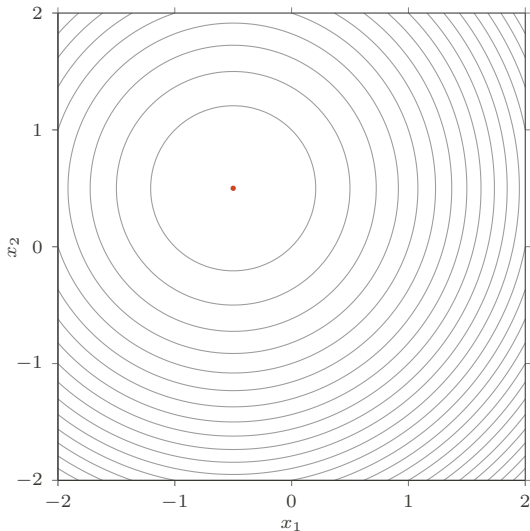
jest równaniem **elipsoidy** o środku $\mathbf{x}^*$, której osie główne odpowiadają **wektorom własnym** $\mathbf{A}$, a długości tych osi są odwrotnie proporcjonalne do pierwiastków **wartości własnych**

W szczególności dla $n = 2$ jest to równanie **elipsy** na płaszczyźnie

Wniosek: poziomice funkcji kwadratowej są elipsoidami

Wykres konturowy funkcji kwadratowej

$$f(x_1, x_2) = x_1^2 + x_2^2 + x_1 - x_2$$



$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Wektory własne:

$$\mathbf{v}_1 = (1, 0), \mathbf{v}_2 = (0, 1)$$

Wartości własne:

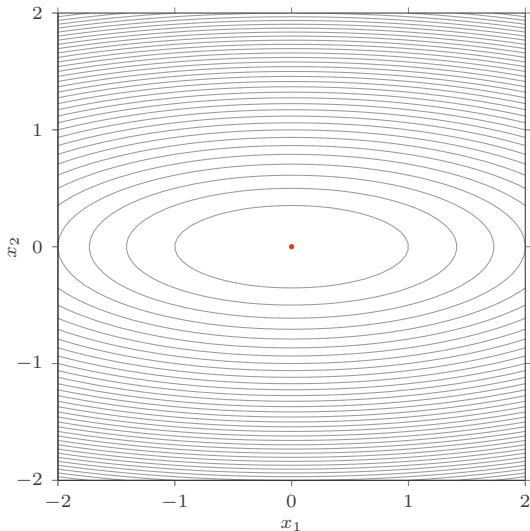
$$\lambda_1 = 1, \lambda_2 = 1$$

Minimum:

$$\mathbf{x}^* = (-0.5, 0.5)$$

Wykres konturowy funkcji kwadratowej

$$f(x_1, x_2) = 0.5x_1^2 + 2x_2^2$$



$$\mathbf{A} = \begin{bmatrix} 0.5 & 0 \\ 0 & 2 \end{bmatrix}$$

Wektory własne:

$$\mathbf{v}_1 = (1, 0), \mathbf{v}_2 = (0, 1)$$

Wartości własne:

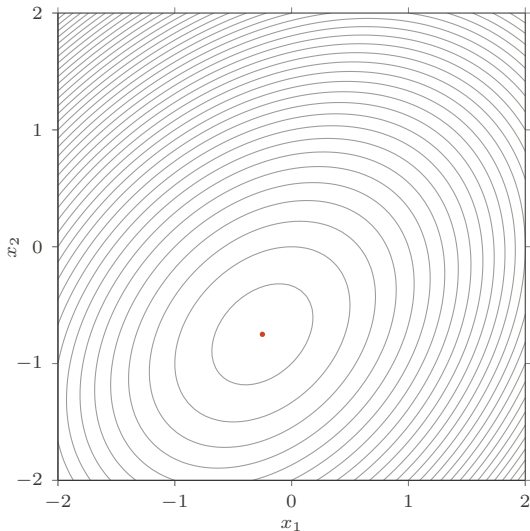
$$\lambda_1 = 0.5, \lambda_2 = 2$$

Minimum:

$$\mathbf{x}^* = (0, 0)$$

Wykres konturowy funkcji kwadratowej

$$f(x_1, x_2) = 1.5x_1^2 - x_1x_2 + 1.5x_2^2 + 2x_2$$



$$A = \frac{1}{2} \begin{bmatrix} 3 & -1 \\ 1 & 3 \end{bmatrix}$$

Wektory własne:

$$v_1 = \frac{1}{\sqrt{2}}(1, 1), v_2 = \frac{1}{\sqrt{2}}(1, -1)$$

Wartości własne:

$$\lambda_1 = 1, \lambda_2 = 2$$

Minimum:

$$x^* = (-0.25, -0.75)$$

Metody spadkowe (kierunków poprawy)

- Iteracyjne metody generujące punkty $\mathbf{x}_1, \mathbf{x}_2, \dots$ w ten sposób aby w każdej iteracji zmniejszać wartość funkcji celu: $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$
- W każdej iteracji $k = 1, 2, \dots$ znajduwany jest **kierunek poprawy** $\mathbf{v}_k$, tzn. taki, że przy dostatecznie małych $\alpha \in \mathbb{R}$:

$$f(\mathbf{x}_k + \alpha \mathbf{v}_k) < f(\mathbf{x}_k)$$

- Kolejny punkt wyznaczany jest jako

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{v}_k,$$

gdzie **długość kroku** α_k dobierana jest aby zapewnić odpowiednio duży spadek funkcji celu

Kierunek poprawy

Przypomnienie: Wzór Taylora

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^\top (\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^\top \nabla^2 f(\boldsymbol{\xi})(\boldsymbol{x} - \boldsymbol{x}_0),$$

gdzie $\boldsymbol{\xi}$ jest punktem między $\boldsymbol{x}_0$ a $\boldsymbol{x}$

Kierunek poprawy

Przypomnienie: Wzór Taylora

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\boldsymbol{\xi})(\mathbf{x} - \mathbf{x}_0),$$

gdzie $\boldsymbol{\xi}$ jest punktem między $\mathbf{x}_0$ a $\mathbf{x}$

Wybierając $\mathbf{x}_0 = \mathbf{x}_k$ oraz $\mathbf{x} = \mathbf{x}_k + \alpha \mathbf{v}_k$ dostajemy:

$$f(\mathbf{x}_k + \alpha \mathbf{v}_k) = f(\mathbf{x}_k) + \alpha \nabla f(\mathbf{x}_k)^\top \mathbf{v}_k + \frac{\alpha^2}{2} \mathbf{v}_k^\top \nabla^2 f(\boldsymbol{\xi}) \mathbf{v}_k$$

Kierunek poprawy

Przypomnienie: Wzór Taylora

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^\top (\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^\top \nabla^2 f(\boldsymbol{\xi})(\boldsymbol{x} - \boldsymbol{x}_0),$$

gdzie $\boldsymbol{\xi}$ jest punktem między $\boldsymbol{x}_0$ a $\boldsymbol{x}$

Wybierając $\boldsymbol{x}_0 = \boldsymbol{x}_k$ oraz $\boldsymbol{x} = \boldsymbol{x}_k + \alpha \boldsymbol{v}_k$ dostajemy:

$$f(\boldsymbol{x}_k + \alpha \boldsymbol{v}_k) - f(\boldsymbol{x}_k) = \alpha \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{v}_k + O(\alpha^2)$$

Kierunek poprawy

Przypomnienie: Wzór Taylora

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\boldsymbol{\xi})(\mathbf{x} - \mathbf{x}_0),$$

gdzie $\boldsymbol{\xi}$ jest punktem między $\mathbf{x}_0$ a $\mathbf{x}$

Wybierając $\mathbf{x}_0 = \mathbf{x}_k$ oraz $\mathbf{x} = \mathbf{x}_k + \alpha \mathbf{v}_k$ dostajemy:

$$f(\mathbf{x}_k + \alpha \mathbf{v}_k) - f(\mathbf{x}_k) = \alpha \nabla f(\mathbf{x}_k)^\top \mathbf{v}_k + O(\alpha^2)$$

Aby kierunek $\mathbf{v}_k$ był kierunkiem poprawy (tzn. gwarantował spadek wartości funkcji przy dostatecznie małych α), **pochodna kierunkowa** wzdłuż $\mathbf{v}_k$ musi być **ujemna**, tzn:

$$\frac{\partial f(\mathbf{x}_k)}{\partial \mathbf{v}_k} = \nabla f(\mathbf{x}_k)^\top \mathbf{v}_k < 0$$

Kierunek poprawy

Przypomnienie: Wzór Taylora

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\boldsymbol{\xi})(\mathbf{x} - \mathbf{x}_0),$$

gdzie $\boldsymbol{\xi}$ jest punktem między $\mathbf{x}_0$ a $\mathbf{x}$

Wybierając $\mathbf{x}_0 = \mathbf{x}_k$ oraz $\mathbf{x} = \mathbf{x}_k + \alpha \mathbf{v}_k$ dostajemy:

$$f(\mathbf{x}_k + \alpha \mathbf{v}_k) - f(\mathbf{x}_k) = \alpha \nabla f(\mathbf{x}_k)^\top \mathbf{v}_k + O(\alpha^2)$$

Aby kierunek $\mathbf{v}_k$ był kierunkiem poprawy (tzn. gwarantował spadek wartości funkcji przy dostatecznie małych α), **pochodna kierunkowa** wzdłuż $\mathbf{v}_k$ musi być **ujemna**, tzn:

$$\frac{\partial f(\mathbf{x}_k)}{\partial \mathbf{v}_k} = \nabla f(\mathbf{x}_k)^\top \mathbf{v}_k < 0$$

Kierunek poprawy musi tworzyć z gradientem **kąt rozwarty** ($\cos \theta < 0$)
Zawsze można dobrać kierunek poprawy **o ile tylko** $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$

Zbieżność metod spadkowych

- Wykorzystujemy własność, że o ile $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$, możemy dobrać kierunek poprawy $\mathbf{v}_k$ i długość kroku α_k tak, aby:

$$f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$$

Zbieżność metod spadkowych

- Wykorzystujemy własność, że o ile $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$, możemy dobrać kierunek poprawy $\mathbf{v}_k$ i długość kroku α_k tak, aby:

$$f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$$

- Zakładając, że funkcja celu $f(\mathbf{x})$ jest ograniczona od dołu i poprawa funkcji w każdej iteracji jest **wystarczająco duża**, wnioskujemy, że punkt zbieżności algorytmu $\mathbf{x}_\infty$ musi mieć własność $\nabla f(\mathbf{x}_\infty) = \mathbf{0}$
(zgrubne uzasadnienie: jeśli $\nabla f(\mathbf{x}_\infty) \neq \mathbf{0}$, moglibyśmy istotnie poprawić funkcję celu, co przeczyłoby temu, że $\mathbf{x}_\infty$ jest punktem zbieżności)

Zbieżność metod spadkowych

- Wykorzystujemy własność, że o ile $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$, możemy dobrać kierunek poprawy $\mathbf{v}_k$ i długość kroku α_k tak, aby:

$$f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$$

- Zakładając, że funkcja celu $f(\mathbf{x})$ jest ograniczona od dołu i poprawa funkcji w każdej iteracji jest **wystarczająco duża**, wnioskujemy, że punkt zbieżności algorytmu $\mathbf{x}_\infty$ musi mieć własność $\nabla f(\mathbf{x}_\infty) = \mathbf{0}$
(zgrubne uzasadnienie: jeśli $\nabla f(\mathbf{x}_\infty) \neq \mathbf{0}$, moglibyśmy istotnie poprawić funkcję celu, co przeczyłoby temu, że $\mathbf{x}_\infty$ jest punktem zbieżności)
- W ogólności nie ma gwarancji, że $\mathbf{x}_\infty$ jest nawet **minimum lokalnym** (może to być **punkt siodłowy**)

Zbieżność metod spadkowych

- Wykorzystujemy własność, że o ile $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$, możemy dobrać kierunek poprawy $\mathbf{v}_k$ i długość kroku α_k tak, aby:

$$f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$$

- Zakładając, że funkcja celu $f(\mathbf{x})$ jest ograniczona od dołu i poprawa funkcji w każdej iteracji jest **wystarczająco duża**, wnioskujemy, że punkt zbieżności algorytmu $\mathbf{x}_\infty$ musi mieć własność $\nabla f(\mathbf{x}_\infty) = \mathbf{0}$
(zgrubne uzasadnienie: jeśli $\nabla f(\mathbf{x}_\infty) \neq \mathbf{0}$, moglibyśmy istotnie poprawić funkcję celu, co przeczyłoby temu, że $\mathbf{x}_\infty$ jest punktem zbieżności)
- W ogólności nie ma gwarancji, że $\mathbf{x}_\infty$ jest nawet **minimum lokalnym** (może to być **punkt siodłowy**)
- Dla funkcji **wypukłych** możemy pokazać, że $\mathbf{x}_\infty$ jest **minimum globalnym** i oszacować **rzęd zbieżności**

Metoda spadku wzdłuż gradientu

Gradient jako kierunek największego wzrostu

Pokazaliśmy wcześniej, że dla jednostkowego wektora $\mathbf{v}_k$:

$$-\|\nabla f(\mathbf{x}_k)\| \leq \nabla f(\mathbf{x}_k)^\top \mathbf{v}_k \leq \|\nabla f(\mathbf{x}_k)\|,$$

a obie skrajne wartości osiągnęte są przez $\mathbf{v}_k = \pm \frac{\nabla f(\mathbf{x}_k)}{\|\nabla f(\mathbf{x}_k)\|}$

Gradient jako kierunek największego wzrostu

Pokazaliśmy wcześniej, że dla jednostkowego wektora $\mathbf{v}_k$:

$$-\|\nabla f(\mathbf{x}_k)\| \leq \nabla f(\mathbf{x}_k)^\top \mathbf{v}_k \leq \|\nabla f(\mathbf{x}_k)\|,$$

a obie skrajne wartości osiągnęte są przez $\mathbf{v}_k = \pm \frac{\nabla f(\mathbf{x}_k)}{\|\nabla f(\mathbf{x}_k)\|}$

Wniosek: gradient jest kierunkiem (lokalnie) **największego wzrostu** funkcji, a ujemny gradient – kierunkiem (lokalnie) **największego spadku**

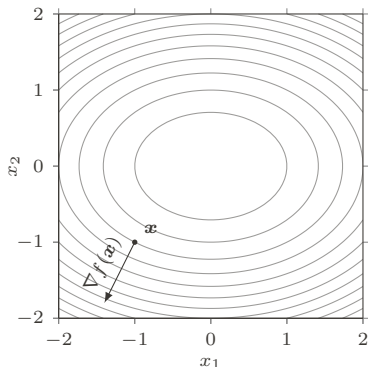
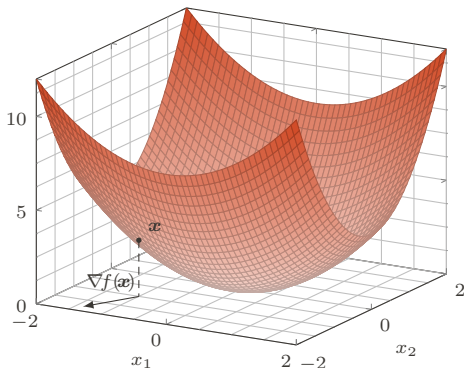
Gradient jako kierunek największego wzrostu

Pokazaliśmy wcześniej, że dla jednostkowego wektora $\mathbf{v}_k$:

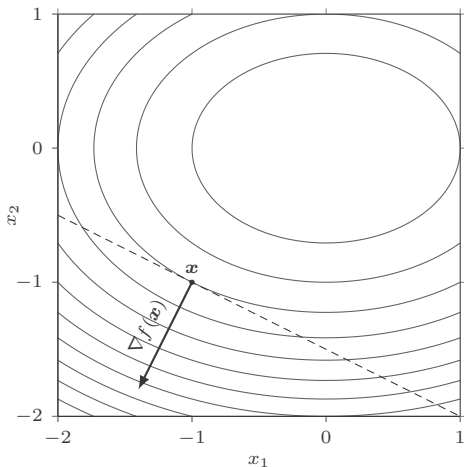
$$-\|\nabla f(\mathbf{x}_k)\| \leq \nabla f(\mathbf{x}_k)^\top \mathbf{v}_k \leq \|\nabla f(\mathbf{x}_k)\|,$$

a obie skrajne wartości osiągane są przez $\mathbf{v}_k = \pm \frac{\nabla f(\mathbf{x}_k)}{\|\nabla f(\mathbf{x}_k)\|}$

Wniosek: gradient jest kierunkiem (lokalnie) **największego wzrostu** funkcji, a ujemny gradient – kierunkiem (lokalnie) **największego spadku**



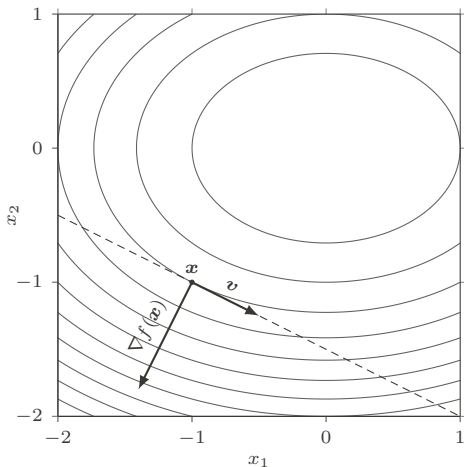
Gradient na wykresie konturowym



Wektor gradientu jest zawsze
prostopadły do poziomicy funkcji

Dlaczego?

Gradient na wykresie konturowym



Wektor gradientu jest zawsze **prostopadły** do poziomicy funkcji

Dlaczego?

Ponieważ pochodna kierunkowa w kierunku v wzdłuż poziomicy musi wynosić zero:

$$\frac{\partial f(x + v)}{\partial v} = \nabla f(x)^\top v = 0$$

Stąd $\nabla f(x) \perp v$

Algorytm spadku wzdłuż gradientu

Rozpoczynając od $\mathbf{x}_1$, w kolejnych iteracjach $k = 1, 2, \dots$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$$

Algorytm spadku wzdłuż gradientu

Rozpoczynając od $\mathbf{x}_1$, w kolejnych iteracjach $k = 1, 2, \dots$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$$

Pytanie: jak dobrać długości kroków α_k ?

Algorytm spadku wzdłuż gradientu

Rozpoczynając od $\mathbf{x}_1$, w kolejnych iteracjach $k = 1, 2, \dots$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$$

Pytanie: jak dobrać długości kroków α_k ?

- **Stała** długość kroku $\alpha_k = \alpha$: najprostszy algorytm, ale...
 - ▶ wartość α trzeba dobrać metodą prób i błędów
 - ▶ nie gwarantuje, że $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$

Algorytm spadku wzdłuż gradientu

Rozpoczynając od $\mathbf{x}_1$, w kolejnych iteracjach $k = 1, 2, \dots$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$$

Pytanie: jak dobrać długości kroków α_k ?

- **Stała** długość kroku $\alpha_k = \alpha$: najprostszy algorytm, ale...
 - ▶ wartość α trzeba dobrać metodą prób i błędów
 - ▶ nie gwarantuje, że $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$
- **Optymalna** długość kroku: algorytm **najszybszego spadku**

Algorytm spadku wzdłuż gradientu

Rozpoczynając od $\mathbf{x}_1$, w kolejnych iteracjach $k = 1, 2, \dots$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$$

Pytanie: jak dobrać długości kroków α_k ?

- **Stała** długość kroku $\alpha_k = \alpha$: najprostszy algorytm, ale...
 - ▶ wartość α trzeba dobrać metodą prób i błędów
 - ▶ nie gwarantuje, że $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$
- **Optymalna** długość kroku: algorytm **najszybszego spadku**
- **Dostatecznie dobra** długość kroku: niekoniecznie optymalna, ale gwarantująca $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$;
 - ▶ najczęściej używana w praktyce

Algorytm najszybszego spadku (Cauchy'ego)

Algorytm dobiera długość kroku α_k , która najbardziej zmniejsza funkcję celu wzdłuż wektora gradientu:

$$\alpha_k = \operatorname{argmin}_{\alpha \geq 0} f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))$$

Algorytm najszybszego spadku (Cauchy'ego)

Algorytm dobiera długość kroku α_k , która najbardziej zmniejsza funkcję celu wzdłuż wektora gradientu:

$$\alpha_k = \underset{\alpha \geq 0}{\operatorname{argmin}} f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))$$

Musimy rozwiązać **jednowymiarowy** problem minimalizacji **funkcji jednej zmiennej**:

$$g(\alpha) = f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))$$

Algorytm najszybszego spadku (Cauchy'ego)

Algorytm dobiera długość kroku α_k , która najbardziej zmniejsza funkcję celu wzdłuż wektora gradientu:

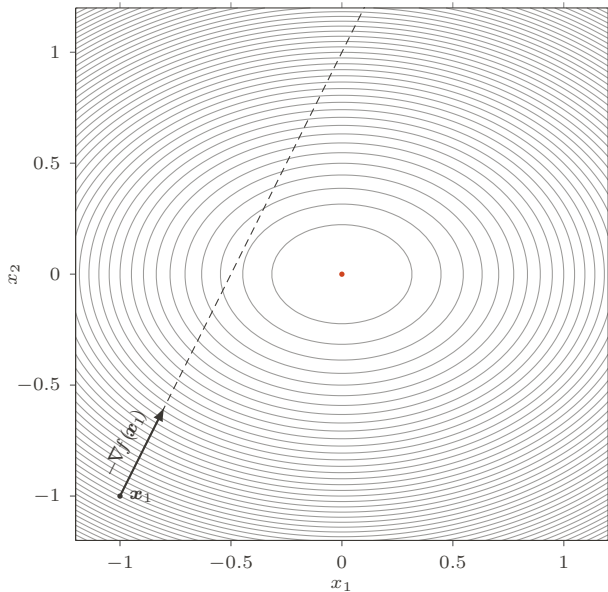
$$\alpha_k = \underset{\alpha \geq 0}{\operatorname{argmin}} f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))$$

Musimy rozwiązać **jednowymiarowy** problem minimalizacji **funkcji jednej zmiennej**:

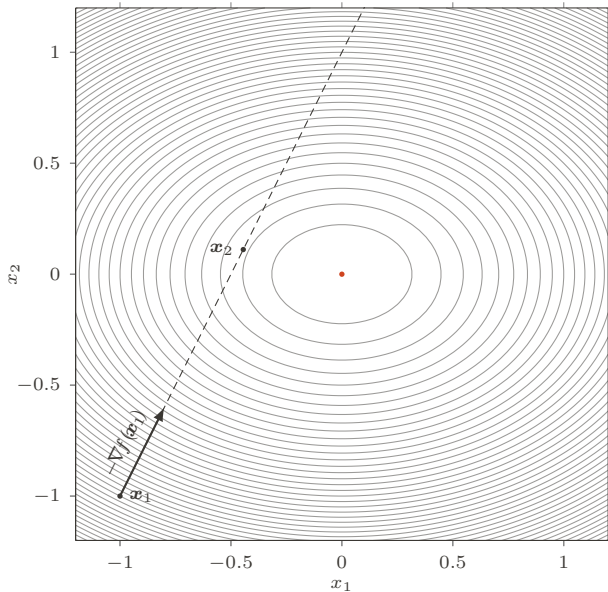
$$g(\alpha) = f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))$$

- Metoda złotego podziału
- Metoda bisekcji
- Metoda Newtona
- ...

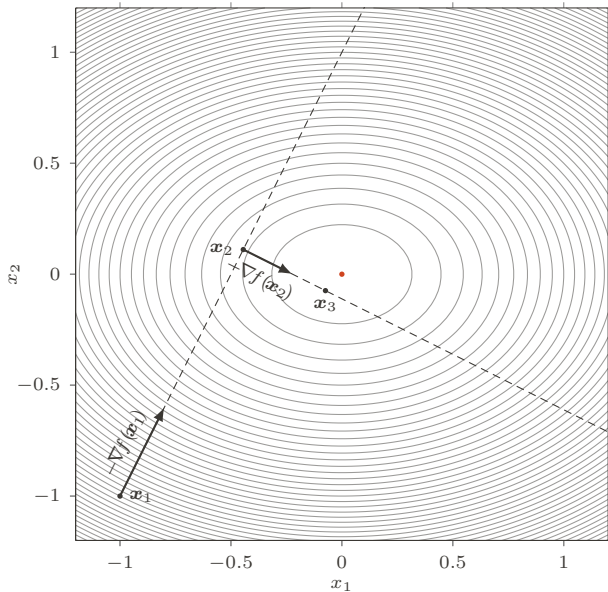
Algorytm najszybszego spadku



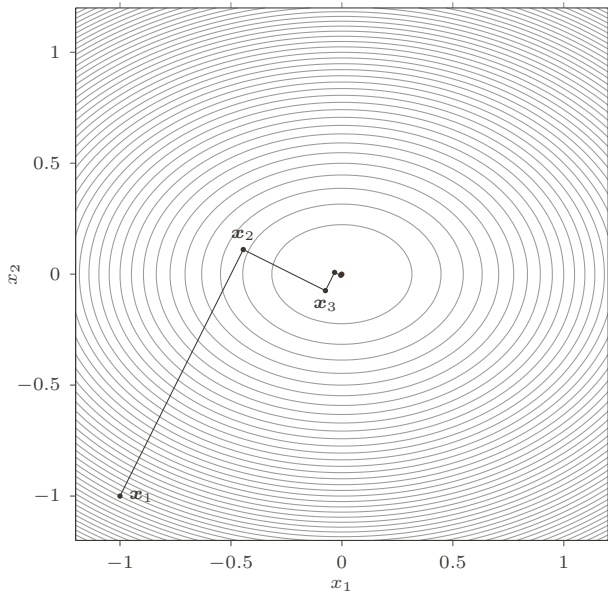
Algorytm najszybszego spadku



Algorytm najszybszego spadku

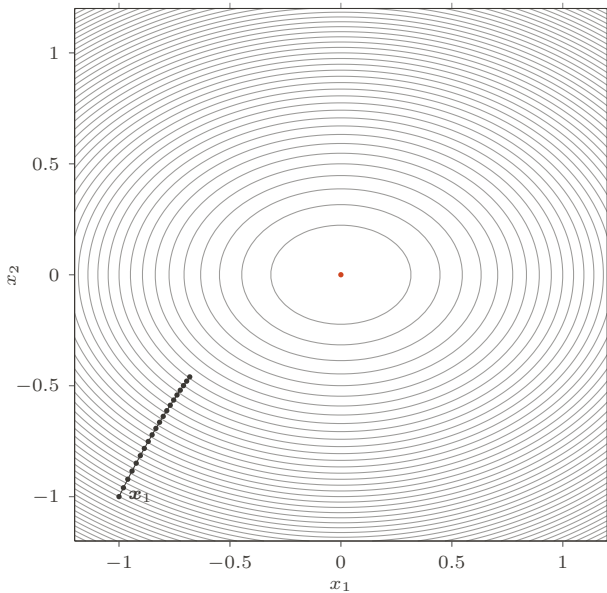


Algorytm najszybszego spadku



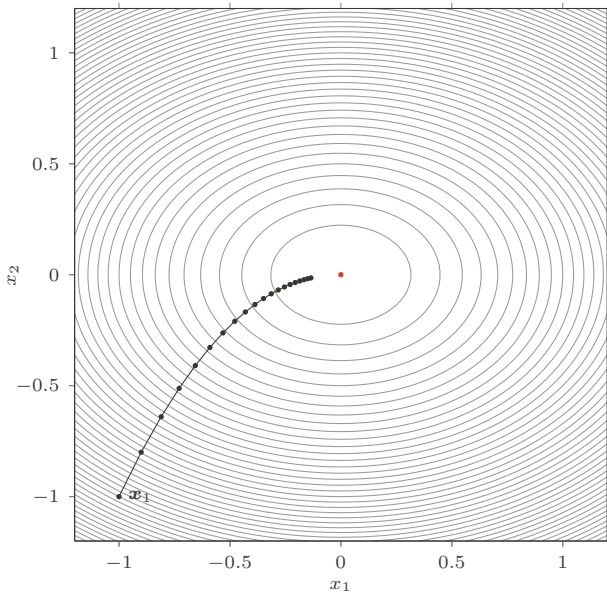
Dla porównania: algorytm ze stałą długością kroku

$$\alpha = 0.01$$



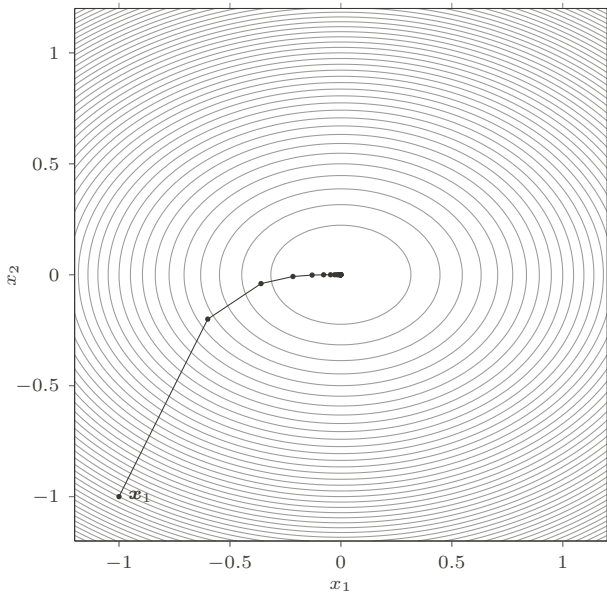
Dla porównania: algorytm ze stałą długością kroku

$$\alpha = 0.05$$



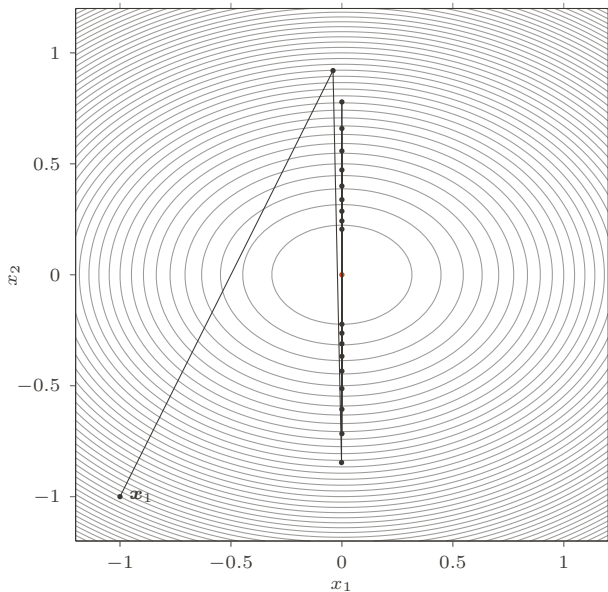
Dla porównania: algorytm ze stałą długością kroku

$$\alpha = 0.2$$



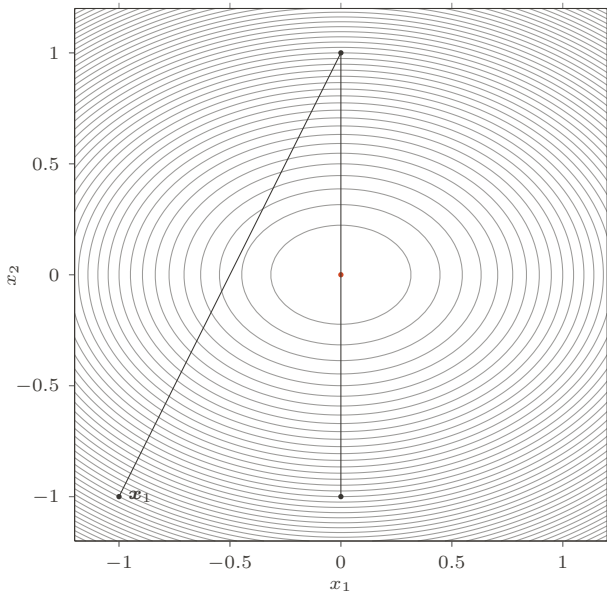
Dla porównania: algorytm ze stałą długością kroku

$$\alpha = 0.48$$



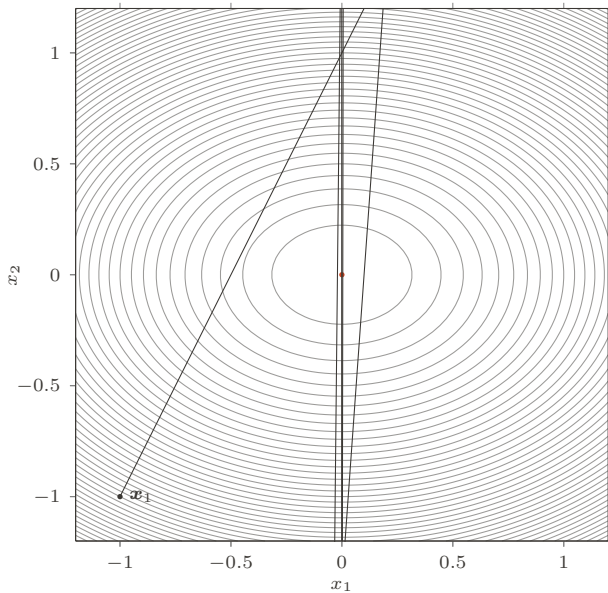
Dla porównania: algorytm ze stałą długością kroku

$$\alpha = 0.5$$



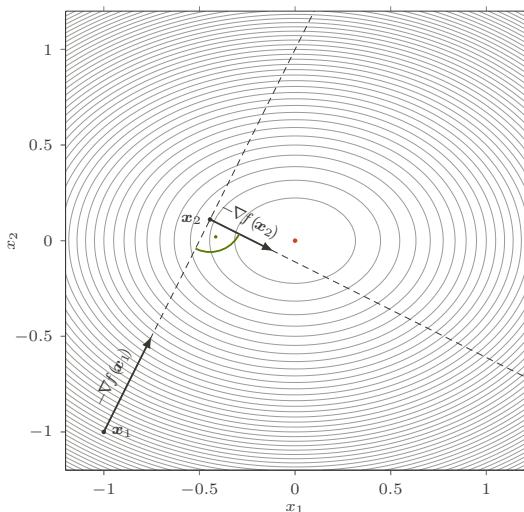
Dla porównania: algorytm ze stałą długością kroku

$$\alpha = 0.6$$



Twierdzenie o zygzakowaniu

W algorytmie najszybszego spadku kolejne kierunki poprawy (gradienty) są do siebie **ortogonalne**



Twierdzenie o zygzakowaniu

W algorytmie najszybszego spadku kolejne kierunki poprawy (gradienty) są do siebie **ortogonalne**

Dowód: Długość kroku α_k w równaniu $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$ uzyskuje się znajdując minimum funkcji jednowymiarowej:

$$g(\alpha) = f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))$$

Twierdzenie o zygzakowaniu

W algorytmie najszybszego spadku kolejne kierunki poprawy (gradienty) są do siebie **ortogonalne**

Dowód: Długość kroku α_k w równaniu $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$ uzyskuje się znajdując minimum funkcji jednowymiarowej:

Pokazaliśmy już, że: $g(\alpha) = f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))$

$$\frac{df(\mathbf{x} + \alpha \mathbf{v})}{d\alpha} = \nabla f(\mathbf{x} + \alpha \mathbf{v})^\top \mathbf{v},$$

dlatego:

$$\frac{dg(\alpha)}{d\alpha} = -\nabla f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))^\top \nabla f(\mathbf{x}_k)$$

Twierdzenie o zygzakowaniu

W algorytmie najszybszego spadku kolejne kierunki poprawy (gradienty) są do siebie **ortogonalne**

Dowód: Długość kroku α_k w równaniu $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$ uzyskuje się znajdując minimum funkcji jednowymiarowej:

Pokazaliśmy już, że: $g(\alpha) = f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))$

$$\frac{df(\mathbf{x} + \alpha \mathbf{v})}{d\alpha} = \nabla f(\mathbf{x} + \alpha \mathbf{v})^\top \mathbf{v},$$

dlatego:

$$\frac{dg(\alpha)}{d\alpha} = -\nabla f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))^\top \nabla f(\mathbf{x}_k)$$

Skoro $g(\alpha)$ ma **minimum** w α_k , musimy mieć $\frac{dg(\alpha)}{d\alpha} = 0$ w $\alpha = \alpha_k$.

Twierdzenie o zygzakowaniu

W algorytmie najszybszego spadku kolejne kierunki poprawy (gradienty) są do siebie **ortogonalne**

Dowód: Długość kroku α_k w równaniu $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$ uzyskuje się znajdując minimum funkcji jednowymiarowej:

Pokazaliśmy już, że: $g(\alpha) = f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))$

$$\frac{df(\mathbf{x} + \alpha \mathbf{v})}{d\alpha} = \nabla f(\mathbf{x} + \alpha \mathbf{v})^\top \mathbf{v},$$

dlatego:

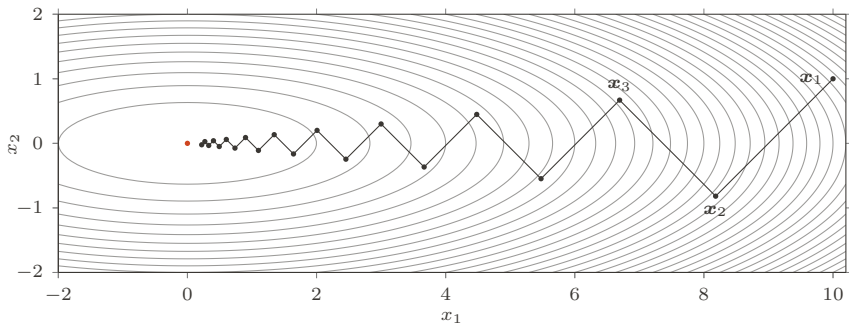
$$\frac{dg(\alpha)}{d\alpha} = -\nabla f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))^\top \nabla f(\mathbf{x}_k)$$

Skoro $g(\alpha)$ ma **minimum** w α_k , musimy mieć $\frac{dg(\alpha)}{d\alpha} = 0$ w $\alpha = \alpha_k$. Ale:

$$0 = \left. \frac{dg(\alpha)}{d\alpha} \right|_{\alpha=\alpha_k} = -\nabla f(\mathbf{x}_{k+1})^\top \nabla f(\mathbf{x}_k)$$

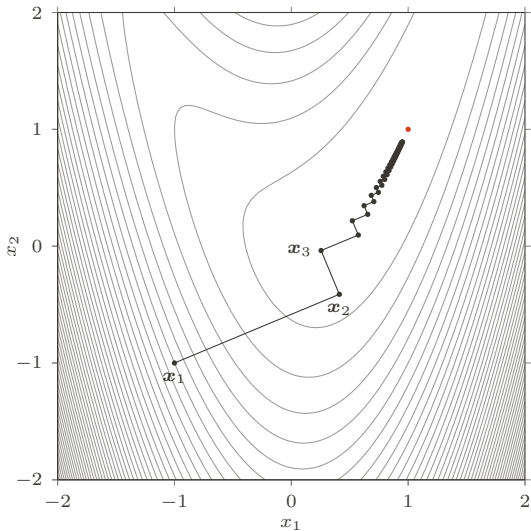
Zygzakowanie – przykład

$$f(x_1, x_2) = x_1^2 + 10x_2^2$$



Zygzakowanie – przykład

$$f(x_1, x_2) = 2.5(x_1^2 - x_2)^2 + (1 - x_1)^2$$



Metoda najszybszego spadku dla funkcji kwadratowej^{*}

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Metoda najszybszego spadku dla funkcji kwadratowej^{*}

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Oznaczmy gradient $\nabla f(\mathbf{x}_k) = 2\mathbf{A}\mathbf{x}_k + \mathbf{b}$ jako $\mathbf{g}_k$, czyli

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k, \quad \text{gdzie } \alpha_k = \operatorname{argmin}_{\alpha \geq 0} \underbrace{f(\mathbf{x}_k - \alpha \mathbf{g}_k)}_{g(\alpha)}$$

Metoda najszybszego spadku dla funkcji kwadratowej^{*}

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Oznaczmy gradient $\nabla f(\mathbf{x}_k) = 2\mathbf{A}\mathbf{x}_k + \mathbf{b}$ jako $\mathbf{g}_k$, czyli

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k, \quad \text{gdzie } \alpha_k = \operatorname{argmin}_{\alpha \geq 0} \underbrace{f(\mathbf{x}_k - \alpha \mathbf{g}_k)}_{g(\alpha)}$$

$$g(\alpha) = (\mathbf{x}_k - \alpha \mathbf{g}_k)^\top \mathbf{A} (\mathbf{x}_k - \alpha \mathbf{g}_k) + (\mathbf{x}_k - \alpha \mathbf{g}_k)^\top \mathbf{b} + c$$

Metoda najszybszego spadku dla funkcji kwadratowej^{*}

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Oznaczmy gradient $\nabla f(\mathbf{x}_k) = 2\mathbf{A}\mathbf{x}_k + \mathbf{b}$ jako $\mathbf{g}_k$, czyli

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k, \quad \text{gdzie } \alpha_k = \operatorname{argmin}_{\alpha \geq 0} \underbrace{f(\mathbf{x}_k - \alpha \mathbf{g}_k)}_{g(\alpha)}$$

$$g(\alpha) = \alpha^2 \mathbf{g}_k^\top \mathbf{A} \mathbf{g}_k - \alpha \mathbf{g}_k^\top \underbrace{(2\mathbf{A}\mathbf{x}_k + \mathbf{b})}_{\mathbf{g}_k} + \text{const}$$

Metoda najszybszego spadku dla funkcji kwadratowej^{*}

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Oznaczmy gradient $\nabla f(\mathbf{x}_k) = 2\mathbf{A}\mathbf{x}_k + \mathbf{b}$ jako $\mathbf{g}_k$, czyli

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k, \quad \text{gdzie } \alpha_k = \operatorname{argmin}_{\alpha \geq 0} \underbrace{f(\mathbf{x}_k - \alpha \mathbf{g}_k)}_{g(\alpha)}$$

$$g(\alpha) = \alpha^2 \mathbf{g}_k^\top \mathbf{A} \mathbf{g}_k - \alpha \|\mathbf{g}_k\|^2 + \text{const}$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Wypukła funkcja kwadratowa z dodatnio określoną macierzą A :

$g(\alpha)$ jest funkcją kwadratową:

$$g(\alpha) = a\alpha^2 + b\alpha + c, \quad \text{dla } b = -\|g_k\|^2, \quad a = g_k^\top A g_k.$$

Oznaczmy gradient

$$x_{k+1} =$$

$$\text{Minimum w } \alpha = \underbrace{-\frac{b}{2a}}_{\alpha \geq 0} = \frac{\|g_k\|^2}{2 \underbrace{g_k^\top A g_k}_{g(\alpha)}}$$

$$g(\alpha) = \alpha^2 g_k^\top A g_k - \alpha \|g_k\|^2 + \text{const}$$

Metoda najszybszego spadku dla funkcji kwadratowej^{*}

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Oznaczmy gradient $\nabla f(\mathbf{x}_k) = 2A\mathbf{x}_k + \mathbf{b}$ jako $\mathbf{g}_k$, czyli

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k, \quad \text{gdzie } \alpha_k = \operatorname{argmin}_{\alpha \geq 0} \underbrace{f(\mathbf{x}_k - \alpha \mathbf{g}_k)}_{g(\alpha)}$$

$$g(\alpha) = \alpha^2 \mathbf{g}_k^\top A \mathbf{g}_k - \alpha \|\mathbf{g}_k\|^2 + \text{const}$$

$$\text{Optymalna długość kroku: } \alpha_k = \frac{\|\mathbf{g}_k\|^2}{2\mathbf{g}_k^\top A \mathbf{g}_k}$$

Metoda najszybszego spadku dla funkcji kwadratowej^{*}

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Oznaczmy gradient $\nabla f(\mathbf{x}_k) = 2A\mathbf{x}_k + \mathbf{b}$ jako $\mathbf{g}_k$, czyli

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k, \quad \text{gdzie } \alpha_k = \operatorname{argmin}_{\alpha \geq 0} \underbrace{f(\mathbf{x}_k - \alpha \mathbf{g}_k)}_{g(\alpha)}$$

$$g(\alpha) = \alpha^2 \mathbf{g}_k^\top A \mathbf{g}_k - \alpha \|\mathbf{g}_k\|^2 + \text{const}$$

$$\text{Optymalna długość kroku: } \alpha_k = \frac{\|\mathbf{g}_k\|^2}{2\mathbf{g}_k^\top A \mathbf{g}_k}$$

Możemy też wyznaczyć spadek funkcji celu:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) = g(\alpha_k) - g(0) = \alpha_k^2 \mathbf{g}_k^\top A \mathbf{g}_k - \alpha_k \|\mathbf{g}_k\|^2$$

Metoda najszybszego spadku dla funkcji kwadratowej^{*}

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Oznaczmy gradient $\nabla f(\mathbf{x}_k) = 2A\mathbf{x}_k + \mathbf{b}$ jako $\mathbf{g}_k$, czyli

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k, \quad \text{gdzie } \alpha_k = \operatorname{argmin}_{\alpha \geq 0} \underbrace{f(\mathbf{x}_k - \alpha \mathbf{g}_k)}_{g(\alpha)}$$

$$g(\alpha) = \alpha^2 \mathbf{g}_k^\top A \mathbf{g}_k - \alpha \|\mathbf{g}_k\|^2 + \text{const}$$

$$\text{Optymalna długość kroku: } \alpha_k = \frac{\|\mathbf{g}_k\|^2}{2\mathbf{g}_k^\top A \mathbf{g}_k}$$

Możemy też wyznaczyć spadek funkcji celu:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) = \frac{1}{4} \frac{\|\mathbf{g}_k\|^4}{\mathbf{g}_k^\top A \mathbf{g}_k} - \frac{1}{2} \frac{\|\mathbf{g}_k\|^4}{\mathbf{g}_k^\top A \mathbf{g}_k}$$

Metoda najszybszego spadku dla funkcji kwadratowej^{*}

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Oznaczmy gradient $\nabla f(\mathbf{x}_k) = 2A\mathbf{x}_k + \mathbf{b}$ jako $\mathbf{g}_k$, czyli

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k, \quad \text{gdzie } \alpha_k = \operatorname{argmin}_{\alpha \geq 0} \underbrace{f(\mathbf{x}_k - \alpha \mathbf{g}_k)}_{g(\alpha)}$$

$$g(\alpha) = \alpha^2 \mathbf{g}_k^\top A \mathbf{g}_k - \alpha \|\mathbf{g}_k\|^2 + \text{const}$$

$$\text{Optymalna długość kroku: } \alpha_k = \frac{\|\mathbf{g}_k\|^2}{2\mathbf{g}_k^\top A \mathbf{g}_k}$$

Możemy też wyznaczyć spadek funkcji celu:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) = -\frac{1}{4} \frac{\|\mathbf{g}_k\|^4}{\mathbf{g}_k^\top A \mathbf{g}_k}$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Funkcja ma **minimum** w punkcie $\mathbf{x}^* = -\frac{1}{2}A^{-1}\mathbf{b}$

Metoda najszybszego spadku dla funkcji kwadratowej*

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Funkcja ma **minimum** w punkcie $\mathbf{x}^* = -\frac{1}{2}A^{-1}\mathbf{b}$

Biorąc reprezentacje funkcji w postaci wielomianu Taylora drugiego rzędu wokół punktu $\mathbf{x}^*$:

Metoda najszybszego spadku dla funkcji kwadratowej*

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Funkcja ma **minimum** w punkcie $\mathbf{x}^* = -\frac{1}{2}\mathbf{A}^{-1}\mathbf{b}$

Biorąc reprezentację funkcji w postaci wielomianu Taylora drugiego rzędu wokół punktu $\mathbf{x}^*$:

$$f(\mathbf{x}) - f(\mathbf{x}^*) = (\mathbf{x} - \mathbf{x}^*)^\top \mathbf{A}(\mathbf{x} - \mathbf{x}^*)$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Funkcja ma **minimum** w punkcie $\mathbf{x}^* = -\frac{1}{2}\mathbf{A}^{-1}\mathbf{b}$

Biorąc reprezentacje funkcji w postaci wielomianu Taylora drugiego rzędu wokół punktu $\mathbf{x}^*$:

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) = (\mathbf{x}_k - \mathbf{x}^*)^\top \mathbf{A}(\mathbf{x}_k - \mathbf{x}^*)$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Wyprowadzamy $\mathbf{g}_k = 2\mathbf{A}\mathbf{x}_k + \mathbf{b}$, a więc z dodatnio określoną macierzą $\mathbf{A}$:

$$\mathbf{x}_k = \frac{1}{2}\mathbf{A}^{-1}\mathbf{g}_k - \frac{1}{2}\mathbf{A}^{-1}\mathbf{b}$$
$$\mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Funkcja ma **minimum** w punkcie $\mathbf{x}^* = -\frac{1}{2}\mathbf{A}^{-1}\mathbf{b}$

Biorąc reprezentację funkcji w postaci wielomiana $\mathbf{x}^* = -\frac{1}{2}\mathbf{A}^{-1}\mathbf{b}$ tego rzędu wokół punktu $\mathbf{x}^*$:

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) = \underbrace{(\mathbf{x}_k - \mathbf{x}^*)^\top}_{\frac{1}{2}\mathbf{A}^{-1}\mathbf{g}_k} \mathbf{A} \underbrace{(\mathbf{x}_k - \mathbf{x}^*)}_{\frac{1}{2}\mathbf{A}^{-1}\mathbf{g}_k}$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Funkcja ma **minimum** w punkcie $\mathbf{x}^* = -\frac{1}{2}A^{-1}\mathbf{b}$

Biorąc reprezentacje funkcji w postaci wielomianu Taylora drugiego rzędu wokół punktu $\mathbf{x}^*$:

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) = \frac{1}{4}(\mathbf{A}^{-1}\mathbf{g}_k)^\top \mathbf{A} \mathbf{A}^{-1} \mathbf{g}_k$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Funkcja ma **minimum** w punkcie $\mathbf{x}^* = -\frac{1}{2}A^{-1}\mathbf{b}$

Biorąc reprezentacje funkcji w postaci wielomianu Taylora drugiego rzędu wokół punktu $\mathbf{x}^*$:

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) = \frac{1}{4}\mathbf{g}_k^\top A^{-1} A A^{-1} \mathbf{g}_k$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Funkcja ma **minimum** w punkcie $\mathbf{x}^* = -\frac{1}{2}\mathbf{A}^{-1}\mathbf{b}$

Biorąc reprezentacje funkcji w postaci wielomianu Taylora drugiego rzędu wokół punktu $\mathbf{x}^*$:

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) = \frac{1}{4} \mathbf{g}_k^\top \mathbf{A}^{-1} \mathbf{g}_k$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

Błąd optymalizacji: różnica między wartością

Funkcja ma **minimum** funkcji celu w x_k a wartością optymalną

Biorąc reprezentację funkcji w postaci wielomianu Taylora drugiego rzędu wokół punktu x^* :

$$E(x_k) = f(x_k) - f(x^*) = \frac{1}{2} g_k^\top A^{-1} g_k$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Funkcja ma **minimum** w punkcie $\mathbf{x}^* = -\frac{1}{2}A^{-1}\mathbf{b}$

Biorąc reprezentację funkcji w postaci wielomianu Taylora drugiego rzędu wokół punktu $\mathbf{x}^*$:

$$E(\mathbf{x}_k) = f(\mathbf{x}_k) - f(\mathbf{x}^*) = \frac{1}{4} \mathbf{g}_k^\top A^{-1} \mathbf{g}_k$$

Biorąc równanie z poprzedniego slajdu:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) = -\frac{1}{4} \frac{\|\mathbf{g}_k\|^4}{\mathbf{g}_k^\top A \mathbf{g}_k}$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Funkcja ma **minimum** w punkcie $\mathbf{x}^* = -\frac{1}{2}A^{-1}\mathbf{b}$

Biorąc reprezentacje funkcji w postaci wielomianu Taylora drugiego rzędu wokół punktu $\mathbf{x}^*$:

$$E(\mathbf{x}_k) = f(\mathbf{x}_k) - f(\mathbf{x}^*) = \frac{1}{4}\mathbf{g}_k^\top A^{-1}\mathbf{g}_k$$

Biorąc równanie z poprzedniego slajdu:

$$\left(f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)\right) - \left(f(\mathbf{x}_k) - f(\mathbf{x}^*)\right) = -\frac{1}{4}\frac{\|\mathbf{g}_k\|^4}{\mathbf{g}_k^\top A \mathbf{g}_k}$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Funkcja ma **minimum** w punkcie $\mathbf{x}^* = -\frac{1}{2}A^{-1}\mathbf{b}$

Biorąc reprezentacje funkcji w postaci wielomianu Taylora drugiego rzędu wokół punktu $\mathbf{x}^*$:

$$E(\mathbf{x}_k) = f(\mathbf{x}_k) - f(\mathbf{x}^*) = \frac{1}{4} \mathbf{g}_k^\top A^{-1} \mathbf{g}_k$$

Biorąc równanie z poprzedniego slajdu:

$$E(\mathbf{x}_{k+1}) - E(\mathbf{x}_k) = -\frac{1}{4} \frac{\|\mathbf{g}_k\|^4}{\mathbf{g}_k^\top A \mathbf{g}_k}$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Funkcja ma **minimum** w punkcie $\mathbf{x}^* = -\frac{1}{2}\mathbf{A}^{-1}\mathbf{b}$

Biorąc reprezentację funkcji w postaci wielomianu Taylora drugiego rzędu wokół punktu $\mathbf{x}^*$:

$$E(\mathbf{x}_k) = f(\mathbf{x}_k) - f(\mathbf{x}^*) = \frac{1}{4}\mathbf{g}_k^\top \mathbf{A}^{-1} \mathbf{g}_k$$

Biorąc równanie z poprzedniego slajdu:

$$\frac{E(\mathbf{x}_{k+1}) - E(\mathbf{x}_k)}{E(\mathbf{x}_k)} = - \frac{\|\mathbf{g}_k\|^4}{(\mathbf{g}_k^\top \mathbf{A} \mathbf{g}_k)(\mathbf{g}_k^\top \mathbf{A}^{-1} \mathbf{g}_k)}$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Funkcja ma **minimum** w punkcie $\mathbf{x}^* = -\frac{1}{2}\mathbf{A}^{-1}\mathbf{b}$

Biorąc reprezentacje funkcji w postaci wielomianu Taylora drugiego rzędu wokół punktu $\mathbf{x}^*$:

$$E(\mathbf{x}_k) = f(\mathbf{x}_k) - f(\mathbf{x}^*) = \frac{1}{4}\mathbf{g}_k^\top \mathbf{A}^{-1} \mathbf{g}_k$$

Biorąc równanie z poprzedniego slajdu:

$$\frac{E(\mathbf{x}_{k+1})}{E(\mathbf{x}_k)} = 1 - \frac{\|\mathbf{g}_k\|^4}{(\mathbf{g}_k^\top \mathbf{A} \mathbf{g}_k)(\mathbf{g}_k^\top \mathbf{A}^{-1} \mathbf{g}_k)}$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Funkcja ma **minimum** w punkcie $\mathbf{x}^* = -\frac{1}{2}A^{-1}\mathbf{b}$

Biorąc reprezentacje funkcji w postaci wielomianu Taylora drugiego rzędu wokół punktu $\mathbf{x}^*$:

$$E(\mathbf{x}_k) = f(\mathbf{x}_k) - f(\mathbf{x}^*) = \frac{1}{4}\mathbf{g}_k^\top A^{-1}\mathbf{g}_k$$

Biorąc równanie z poprzedniego slajdu:

$$E(\mathbf{x}_{k+1}) = \left(1 - \frac{\|\mathbf{g}_k\|^4}{(\mathbf{g}_k^\top A \mathbf{g}_k)(\mathbf{g}_k^\top A^{-1} \mathbf{g}_k)}\right) E(\mathbf{x}_k)$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Wypukła funkcja kwadratowa z **dodatnio określoną** macierzą A :

$$f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

Funkcja ma **minimum** w punkcie $\mathbf{x}^* = -\frac{1}{2}A^{-1}\mathbf{b}$

Biorąc reprezentacje funkcji w postaci wielomianu Taylora drugiego rzędu wokół punktu $\mathbf{x}^*$:

$$E(\mathbf{x}_k) = f(\mathbf{x}_k) - f(\mathbf{x}^*) = \frac{1}{4}\mathbf{g}_k^\top A^{-1}\mathbf{g}_k$$

Biorąc równanie z poprzedniego slajdu:

$$E(\mathbf{x}_{k+1}) = \left(1 - \frac{\|\mathbf{g}_k\|^4}{(\mathbf{g}_k^\top A \mathbf{g}_k)(\mathbf{g}_k^\top A^{-1} \mathbf{g}_k)}\right) E(\mathbf{x}_k)$$

Liniowa zbieżność błędu optymalizacji!

Metoda najszybszego spadku dla funkcji kwadratowej^{*}

Nierówność Kantorowicza:

Niech $\mathbf{A}$ będzie macierzą dodatnio określoną z najmniejszą i największą wartością własną równą, odpowiednio, m i M . Wtedy dla dowolnego niezerowego wektora $\mathbf{x}$:

$$\frac{\|\mathbf{x}\|^4}{(\mathbf{x}^\top \mathbf{A} \mathbf{x})(\mathbf{x}^\top \mathbf{A}^{-1} \mathbf{x})} \geq \frac{4Mm}{(M+m)^2}$$

Metoda najszybszego spadku dla funkcji kwadratowej^{*}

Nierówność Kantorowicza:

Niech $\mathbf{A}$ będzie macierzą dodatnio określoną z najmniejszą i największą wartością własną równą, odpowiednio, m i M . Wtedy dla dowolnego niezerowego wektora $\mathbf{x}$:

$$\frac{\|\mathbf{x}\|^4}{(\mathbf{x}^\top \mathbf{A} \mathbf{x})(\mathbf{x}^\top \mathbf{A}^{-1} \mathbf{x})} \geq \frac{4Mm}{(M+m)^2}$$

$$E(\mathbf{x}_{k+1}) = \left(1 - \frac{\|\mathbf{g}_k\|^4}{(\mathbf{g}_k^\top \mathbf{A} \mathbf{g}_k)(\mathbf{g}_k^\top \mathbf{A}^{-1} \mathbf{g}_k)} \right) E(\mathbf{x}_k)$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Nierówność Kantorowicza:

Niech $\mathbf{A}$ będzie macierzą dodatnio określoną z najmniejszą i największą wartością własną równą, odpowiednio, m i M . Wtedy dla dowolnego niezerowego wektora $\mathbf{x}$:

$$\frac{\|\mathbf{x}\|^4}{(\mathbf{x}^\top \mathbf{A} \mathbf{x})(\mathbf{x}^\top \mathbf{A}^{-1} \mathbf{x})} \geq \frac{4Mm}{(M+m)^2}$$

$$E(\mathbf{x}_{k+1}) \leq \left(1 - \frac{4Mm}{(M+m)^2}\right) E(\mathbf{x}_k)$$

Metoda najszybszego spadku dla funkcji kwadratowej^{*}

Nierówność Kantorowicza:

Niech $\mathbf{A}$ będzie macierzą dodatnio określoną z najmniejszą i największą wartością własną równą, odpowiednio, m i M . Wtedy dla dowolnego niezerowego wektora $\mathbf{x}$:

$$\frac{\|\mathbf{x}\|^4}{(\mathbf{x}^\top \mathbf{A} \mathbf{x})(\mathbf{x}^\top \mathbf{A}^{-1} \mathbf{x})} \geq \frac{4Mm}{(M+m)^2}$$

$$E(\mathbf{x}_{k+1}) \leq \left(\frac{(M+m)^2 - 4Mm}{(M+m)^2} \right) E(\mathbf{x}_k)$$

Metoda najszybszego spadku dla funkcji kwadratowej^{*}

Nierówność Kantorowicza:

Niech $\mathbf{A}$ będzie macierzą dodatnio określoną z najmniejszą i największą wartością własną równą, odpowiednio, m i M . Wtedy dla dowolnego niezerowego wektora $\mathbf{x}$:

$$\frac{\|\mathbf{x}\|^4}{(\mathbf{x}^\top \mathbf{A} \mathbf{x})(\mathbf{x}^\top \mathbf{A}^{-1} \mathbf{x})} \geq \frac{4Mm}{(M+m)^2}$$

$$E(\mathbf{x}_{k+1}) \leq \left(\frac{M^2 + 2Mm + m^2 - 4Mm}{(M+m)^2} \right) E(\mathbf{x}_k)$$

Metoda najszybszego spadku dla funkcji kwadratowej^{*}

Nierówność Kantorowicza:

Niech $\mathbf{A}$ będzie macierzą dodatnio określoną z najmniejszą i największą wartością własną równą, odpowiednio, m i M . Wtedy dla dowolnego niezerowego wektora $\mathbf{x}$:

$$\frac{\|\mathbf{x}\|^4}{(\mathbf{x}^\top \mathbf{A} \mathbf{x})(\mathbf{x}^\top \mathbf{A}^{-1} \mathbf{x})} \geq \frac{4Mm}{(M+m)^2}$$

$$E(\mathbf{x}_{k+1}) \leq \left(\frac{M-m}{M+m} \right)^2 E(\mathbf{x}_k)$$

Metoda najszybszego spadku dla funkcji kwadratowej*

Nierówność Kantorowicza:

Niech $\mathbf{A}$ będzie macierzą dodatnio określoną z najmniejszą i największą wartością własną równą, odpowiednio, m i M . Wtedy dla dowolnego niezerowego wektora $\mathbf{x}$:

$$\frac{\|\mathbf{x}\|^4}{(\mathbf{x}^\top \mathbf{A} \mathbf{x})(\mathbf{x}^\top \mathbf{A}^{-1} \mathbf{x})} \geq \frac{4Mm}{(M+m)^2}$$

$$E(\mathbf{x}_{k+1}) \leq \left(\frac{M-m}{M+m} \right)^{2k} E(\mathbf{x}_1)$$

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \beta^k (f(\mathbf{x}_1) - f(\mathbf{x}^*)), \quad \text{gdzie } \beta = \left(\frac{M-m}{M+m} \right)^2$$

Współczynnik zbieżności liniowej β zależy od największej i najmniejszej wartości własnej macierzy $\mathbf{A}$

Dla $M \gg m$, β jest bliski jedności (bardzo wolna zbieżność!)

Wskaźnik uwarunkowania

Wskaźnikiem uwarunkowania macierzy dodatnio określonej $\mathbf{A}$ nazywamy iloraz jej największej i najmniejszej wartości własnej:

$$\kappa = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}$$

Wskaźnik uwarunkowania

Wskaźnikiem uwarunkowania macierzy dodatnio określonej $\mathbf{A}$ nazywamy iloraz jej największej i najmniejszej wartości własnej:

$$\kappa = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}$$

$$\beta = \left(\frac{M - m}{M + m} \right)^2$$

Wskaźnik uwarunkowania

Wskaźnikiem uwarunkowania macierzy dodatnio określonej $\mathbf{A}$ nazywamy iloraz jej największej i najmniejszej wartości własnej:

$$\kappa = \frac{M}{m}$$

$$\kappa = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}$$

$$\beta = \left(\frac{\frac{M}{m} - 1}{\frac{M}{m} + 1} \right)^2$$

Wskaźnik uwarunkowania

Wskaźnikiem uwarunkowania macierzy dodatnio określonej $\mathbf{A}$ nazywamy iloraz jej największej i najmniejszej wartości własnej:

$$\kappa = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}$$

$$\beta = \left(\frac{\kappa - 1}{\kappa + 1} \right)^2$$

Wskaźnik uwarunkowania

Wskaźnikiem uwarunkowania macierzy dodatnio określonej $\mathbf{A}$ nazywamy iloraz jej największej i najmniejszej wartości własnej:

$$\kappa = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}$$

Twierdzenie: Dla wypukłej funkcji kwadratowej z dodatnio określoną macierzą $\mathbf{A}$:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{x}^\top \mathbf{b} + c,$$

algorytm najszybszego spadku zbiega **liniowo** ze współczynnikiem zbieżności β zależnym od wskaźnika uwarunkowania κ macierzy $\mathbf{A}$:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \beta^k (f(\mathbf{x}_1) - f(\mathbf{x}^*)), \quad \text{gdzie } \beta = \left(\frac{\kappa - 1}{\kappa + 1} \right)^2$$

Im większy wskaźnik uwarunkowania κ , tym wolniejsza zbieżność

Zbieżność algorytmu – przykład

Rozważmy funkcję dla $\gamma > 1$:

$$f(\mathbf{x}) = x_1^2 + \gamma x_2^2, \quad \mathbf{x} = (x_1, x_2),$$

mającą minimum w $\mathbf{x}^* = (0, 0)$

Zbieżność algorytmu – przykład

Rozważmy funkcję dla $\gamma > 1$:

$$f(\mathbf{x}) = x_1^2 + \gamma x_2^2, \quad \mathbf{x} = (x_1, x_2),$$

mającą minimum w $\mathbf{x}^* = (0, 0)$

Funkcję możemy przepisać jako:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}, \quad \text{gdzie } \mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & \gamma \end{bmatrix}$$

Zbieżność algorytmu – przykład

Rozważmy funkcję dla $\gamma > 1$:

$$f(\mathbf{x}) = x_1^2 + \gamma x_2^2, \quad \mathbf{x} = (x_1, x_2),$$

mającą minimum w $\mathbf{x}^* = (0, 0)$

Funkcję możemy przepisać jako:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}, \quad \text{gdzie } \mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & \gamma \end{bmatrix}$$

Wartości własne to $(1, \gamma)$, a stąd:

$$\kappa = \frac{\gamma}{1} = \gamma, \quad \beta = \left(\frac{\gamma - 1}{\gamma + 1} \right)^2$$

Zbieżność algorytmu – przykład

Rozważmy funkcję dla $\gamma > 1$:

$$f(\mathbf{x}) = x_1^2 + \gamma x_2^2, \quad \mathbf{x} = (x_1, x_2),$$

mającą minimum w $\mathbf{x}^* = (0, 0)$

Funkcję możemy przepisać jako:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}, \quad \text{gdzie } \mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & \gamma \end{bmatrix}$$

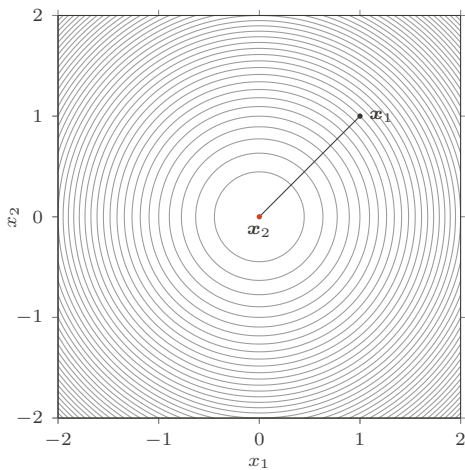
Wartości własne to $(1, \gamma)$, a stąd:

$$\kappa = \frac{\gamma}{1} = \gamma, \quad \beta = \left(\frac{\gamma - 1}{\gamma + 1} \right)^2$$

Przyglądnijmy się zbieżności algorytmu jeśli rozpoczniemy od rozwiązanie startowego $\mathbf{x}_1 = (\gamma, 1)$.

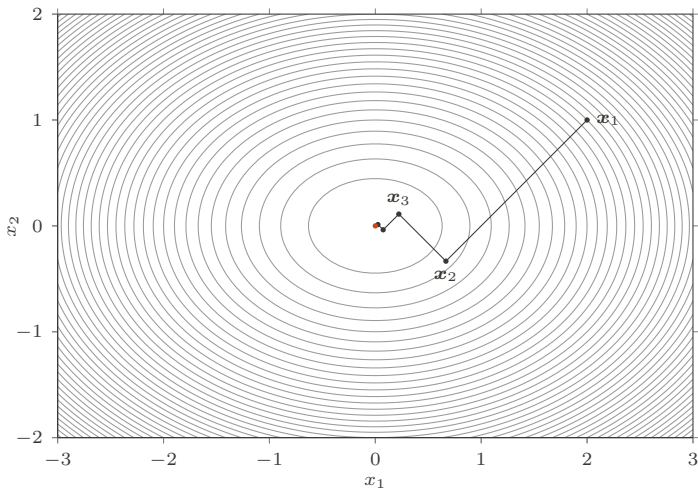
Przykład

$$f(x_1, x_2) = x_1^2 + x_2^2, \quad (\gamma = 1)$$



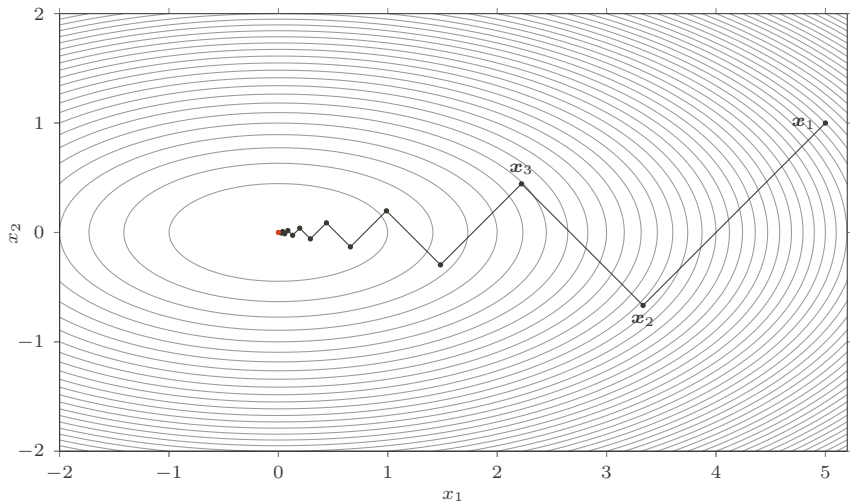
Przykład

$$f(x_1, x_2) = x_1^2 + 2x_2^2, \quad (\gamma = 2)$$



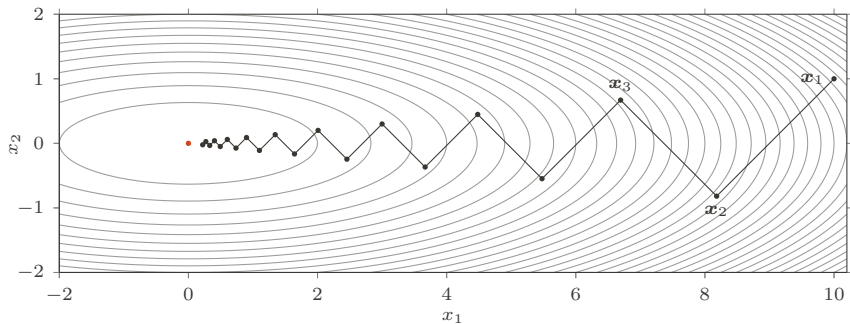
Przykład

$$f(x_1, x_2) = x_1^2 + 5x_2^2, \quad (\gamma = 5)$$



Przykład

$$f(x_1, x_2) = x_1^2 + 10x_2^2, \quad (\gamma = 10)$$



Zbieżność algorytmu dla funkcji wypukłych

Niech $f(x)$ będzie funkcją **wypukłą** i **dwukrotnie różniczkowalną** z ciągłymi pochodnymi. Załóżmy, że hesjan $\nabla^2 f(x^*)$ w minimum x^* jest **dodatnio określony**. Wtedy algorytm najszybszego spadku zbiega liniowo (w sensie błędu optymalizacji) ze ciągiem współczynników zbieżności:

$$\beta_k \xrightarrow{k \rightarrow \infty} \beta = \left(\frac{\kappa - 1}{\kappa + 1} \right)^2$$

gdzie κ jest wskaźnikiem uwarunkowania macierzy $\nabla^2 f(x^*)$

Dobór kroku: reguła Armijo (*backtracking*)

Motywacja:

- Dobór optymalnej długości kroku wymaga rozwiązania problemu optymalizacji jednowymiarowe w każdej iteracji, stąd jest kosztowny obliczeniowo
- Z kolei stała długość nie gwarantuje nawet obniżenia funkcji celu, wymaga dość precyzyjnego strojenia, aby w ogóle zadziałać

Chcemy otrzymać szybką metodę doboru długości kroku, równocześnie gwarantującą istotne obniżenie funkcji celu

Przybliżona minimalizacja

Celem jest przybliżona minimalizacja funkcji jednej zmiennej:

$$\min_{\alpha \geq 0} g(\alpha), \quad \text{gdzie } g(\alpha) = f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))$$

Przybliżona minimalizacja

Celem jest przybliżona minimalizacja funkcji jednej zmiennej:

$$\min_{\alpha \geq 0} g(\alpha), \quad \text{gdzie } g(\alpha) = f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))$$

Pokazaliśmy wcześniej (twierdzenie o zygzakowaniu), że:

$$g'(\alpha) = -\nabla f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))^{\top} \nabla f(\mathbf{x}_k),$$

co oznacza, że $g'(0) = -\|\nabla f(\mathbf{x}_k)\|^2 < 0$

Przybliżona minimalizacja

Celem jest przybliżona minimalizacja funkcji jednej zmiennej:

$$\min_{\alpha \geq 0} g(\alpha), \quad \text{gdzie } g(\alpha) = f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))$$

Pokazaliśmy wcześniej (twierdzenie o zygzakowaniu), że:

$$g'(\alpha) = -\nabla f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))^{\top} \nabla f(\mathbf{x}_k),$$

co oznacza, że $g'(0) = -\|\nabla f(\mathbf{x}_k)\|^2 < 0$

Tym samym funkcja $g(\alpha)$ jest **malejąca** w okolicy $\alpha = 0$, więc wzięcie **bardzo małego** α gwarantuje spadek funkcji, tzn. $g(\alpha) < g(0)$

Przybliżona minimalizacja

Celem jest przybliżona minimalizacja funkcji jednej zmiennej:

$$\min_{\alpha \geq 0} g(\alpha), \quad \text{gdzie } g(\alpha) = f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))$$

Pokazaliśmy wcześniej (twierdzenie o zygzakowaniu), że:

$$g'(\alpha) = -\nabla f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k))^{\top} \nabla f(\mathbf{x}_k),$$

co oznacza, że $g'(0) = -\|\nabla f(\mathbf{x}_k)\|^2 < 0$

Tym samym funkcja $g(\alpha)$ jest **malejąca** w okolicy $\alpha = 0$, więc wzięcie **bardzo małego** α gwarantuje spadek funkcji, tzn. $g(\alpha) < g(0)$

Niestety, spadek funkcji będzie wtedy bardzo niewielki, a zbieżność algorytmu bardzo wolna

Reguła Armijo (*backtracking*)

- Rozważmy funkcję $g(\alpha)$ z $g'(0) < 0$. Celem jest znalezienie $\alpha > 0$, dla którego spadek funkcji względem zera $g(0) - g(\alpha)$ jest **znaczący**

Reguła Armijo (*backtracking*)

- Rozważmy funkcję $g(\alpha)$ z $g'(0) < 0$. Celem jest znalezienie $\alpha > 0$, dla którego spadek funkcji względem zera $g(0) - g(\alpha)$ jest **znaczący**
- Ze wzoru Taylora:

$$g(\alpha) = g(0) + \alpha g'(0) + O(\alpha^2)$$

Reguła Armijo (*backtracking*)

- Rozważmy funkcję $g(\alpha)$ z $g'(0) < 0$. Celem jest znalezienie $\alpha > 0$, dla którego spadek funkcji względem zera $g(0) - g(\alpha)$ jest **znaczący**
- Ze wzoru Taylora:

$$g(\alpha) = g(0) + \alpha g'(0) + O(\alpha^2) \leq g(0) + \gamma \alpha g'(0),$$

dla pewnego $\gamma \in (0, 1)$ i odpowiednio małych α

Reguła Armijo (*backtracking*)

- Rozważmy funkcję $g(\alpha)$ z $g'(0) < 0$. Celem jest znalezienie $\alpha > 0$, dla którego spadek funkcji względem zera $g(0) - g(\alpha)$ jest **znaczący**
- Ze wzoru Taylora:

$$g(\alpha) = g(0) + \alpha g'(0) + O(\alpha^2) \leq g(0) + \gamma \alpha g'(0),$$

dla pewnego $\gamma \in (0, 1)$ i odpowiednio małych α

- Szukamy **jak największego** α spełniającego **regułę Armijo**:

$$g(\alpha) \leq g(0) + \gamma \alpha g'(0), \quad \text{dla ustalonego } \gamma \in (0, 1)$$

Reguła Armijo (*backtracking*)

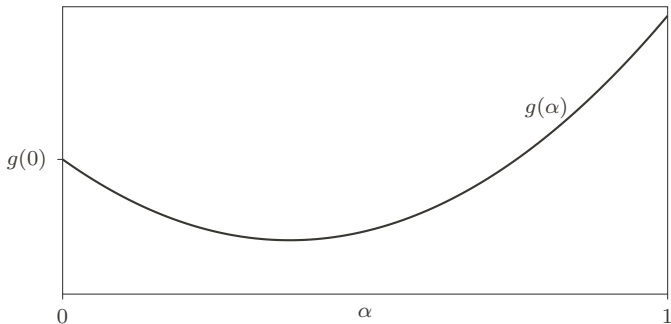
- Rozważmy funkcję $g(\alpha)$ z $g'(0) < 0$. Celem jest znalezienie $\alpha > 0$, dla którego spadek funkcji względem zera $g(0) - g(\alpha)$ jest **znaczący**
- Ze wzoru Taylora:

$$g(\alpha) = g(0) + \alpha g'(0) + O(\alpha^2) \leq g(0) + \gamma \alpha g'(0),$$

dla pewnego $\gamma \in (0, 1)$ i odpowiednio małych α

- Szukamy **jak największego** α spełniającego **regułę Armijo**:

$$g(\alpha) \leq g(0) + \gamma \alpha g'(0), \quad \text{dla ustalonego } \gamma \in (0, 1)$$



Reguła Armijo (*backtracking*)

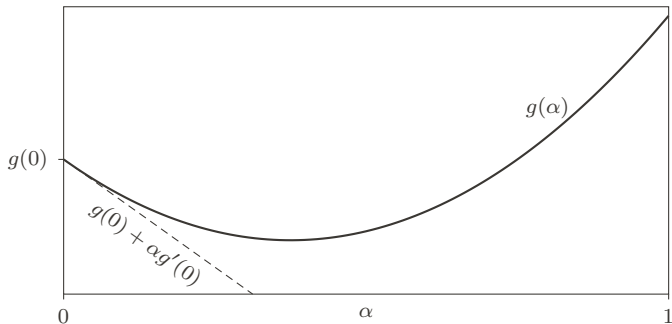
- Rozważmy funkcję $g(\alpha)$ z $g'(0) < 0$. Celem jest znalezienie $\alpha > 0$, dla którego spadek funkcji względem zera $g(0) - g(\alpha)$ jest **znaczący**
- Ze wzoru Taylora:

$$g(\alpha) = g(0) + \alpha g'(0) + O(\alpha^2) \leq g(0) + \gamma \alpha g'(0),$$

dla pewnego $\gamma \in (0, 1)$ i odpowiednio małych α

- Szukamy **jak największego** α spełniającego **regułę Armijo**:

$$g(\alpha) \leq g(0) + \gamma \alpha g'(0), \quad \text{dla ustalonego } \gamma \in (0, 1)$$



Reguła Armijo (*backtracking*)

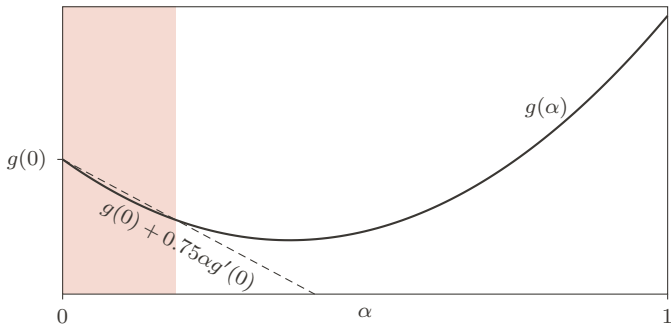
- Rozważmy funkcję $g(\alpha)$ z $g'(0) < 0$. Celem jest znalezienie $\alpha > 0$, dla którego spadek funkcji względem zera $g(0) - g(\alpha)$ jest **znaczący**
- Ze wzoru Taylora:

$$g(\alpha) = g(0) + \alpha g'(0) + O(\alpha^2) \leq g(0) + \gamma \alpha g'(0),$$

dla pewnego $\gamma \in (0, 1)$ i odpowiednio małych α

- Szukamy **jak największego** α spełniającego **regułę Armijo**:

$$g(\alpha) \leq g(0) + \gamma \alpha g'(0), \quad \text{dla ustalonego } \gamma \in (0, 1)$$



Reguła Armijo (*backtracking*)

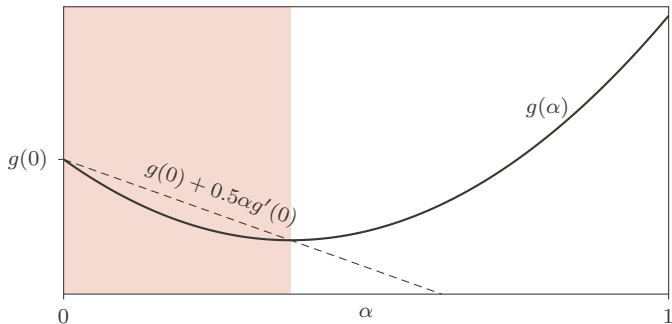
- Rozważmy funkcję $g(\alpha)$ z $g'(0) < 0$. Celem jest znalezienie $\alpha > 0$, dla którego spadek funkcji względem zera $g(0) - g(\alpha)$ jest **znaczący**
- Ze wzoru Taylora:

$$g(\alpha) = g(0) + \alpha g'(0) + O(\alpha^2) \leq g(0) + \gamma \alpha g'(0),$$

dla pewnego $\gamma \in (0, 1)$ i odpowiednio małych α

- Szukamy **jak największego** α spełniającego **regułę Armijo**:

$$g(\alpha) \leq g(0) + \gamma \alpha g'(0), \quad \text{dla ustalonego } \gamma \in (0, 1)$$



Reguła Armijo (*backtracking*)

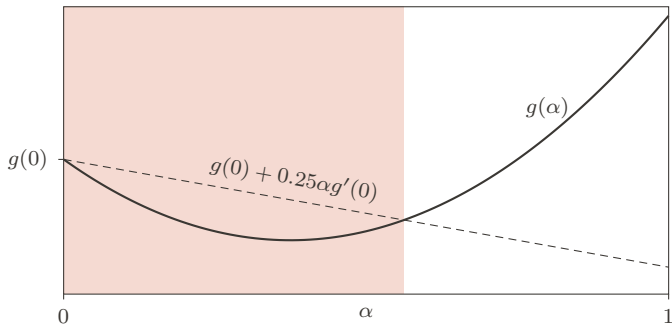
- Rozważmy funkcję $g(\alpha)$ z $g'(0) < 0$. Celem jest znalezienie $\alpha > 0$, dla którego spadek funkcji względem zera $g(0) - g(\alpha)$ jest **znaczący**
- Ze wzoru Taylora:

$$g(\alpha) = g(0) + \alpha g'(0) + O(\alpha^2) \leq g(0) + \gamma \alpha g'(0),$$

dla pewnego $\gamma \in (0, 1)$ i odpowiednio małych α

- Szukamy **jak największego** α spełniającego **regułę Armijo**:

$$g(\alpha) \leq g(0) + \gamma \alpha g'(0), \quad \text{dla ustalonego } \gamma \in (0, 1)$$



Reguła Armijo (*backtracking*)

Rozpoczynamy od dużej wartości $\alpha = \alpha_{\max}$ i redukujemy α mnożąc przez stałą $\eta \in (0, 1)$ tak długo, jak długo nie jest spełniona reguła Armijo

Wejście: Parametry metody: $\alpha_{\max} > 0$, $\gamma \in (0, 1)$ oraz $\eta \in (0, 1)$

Inicjalizuj: $\alpha = \alpha_{\max}$

Tak długo, jak $g(\alpha) > g(0) + \gamma\alpha g'(0)$, **powtarzaj**

$$\alpha \leftarrow \eta\alpha$$

Zwróć długość kroku α

Reguła Armijo (*backtracking*)

Rozpoczynamy od dużej wartości $\alpha = \alpha_{\max}$ i redukujemy α mnożąc przez stałą $\eta \in (0, 1)$ tak długo, jak długo nie jest spełniona reguła Armijo

Wejście: Parametry metody: $\alpha_{\max} > 0$, $\gamma \in (0, 1)$ oraz $\eta \in (0, 1)$

Inicjalizuj: $\alpha = \alpha_{\max}$

Tak długo, jak $f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)) > f(\mathbf{x}_k) - \alpha \gamma \|\nabla f(\mathbf{x}_k)\|^2$, **powtarzaj**
 $\alpha \leftarrow \eta \alpha$

Zwróć długość kroku $\alpha_k = \alpha$

Reguła Armijo (*backtracking*)

Rozpoczynamy od dużej wartości $\alpha = \alpha_{\max}$ i redukujemy α mnożąc przez stałą $\eta \in (0, 1)$ tak długo, jak długo nie jest spełniona reguła Armijo

Wejście: Parametry metody: $\alpha_{\max} > 0$, $\gamma \in (0, 1)$ oraz $\eta \in (0, 1)$

Inicjalizuj: $\alpha = \alpha_{\max}$

Tak długo, jak $f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)) > f(\mathbf{x}_k) - \alpha \gamma \|\nabla f(\mathbf{x}_k)\|^2$, **powtarzaj**
 $\alpha \leftarrow \eta \alpha$

Zwróć długość kroku $\alpha_k = \alpha$

Znalezione α_k gwarantuje więc:

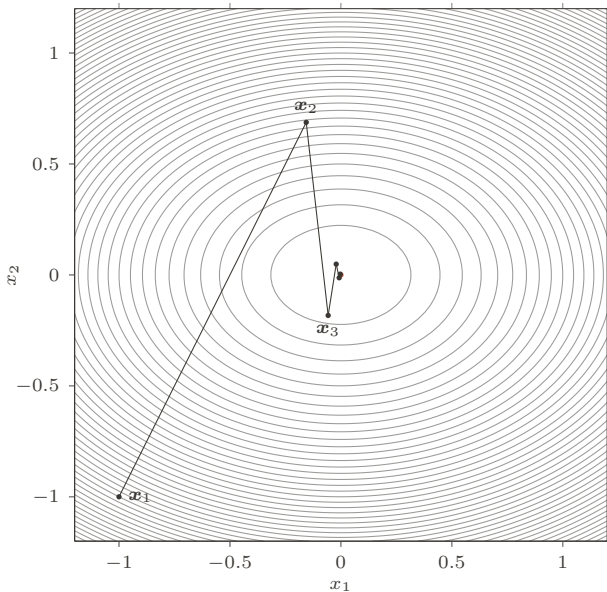
$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \alpha_k \gamma \|\nabla f(\mathbf{x}_k)\|^2$$

Typowe wartości parametrów

- $\alpha_{\max} \simeq 1$
- $\gamma \in [0.01, 0.3]$
- $\eta \in [0.1, 0.8]$ (określa dokładność metody, przetarg między jakością rozwiązania a złożonością obliczeniową)

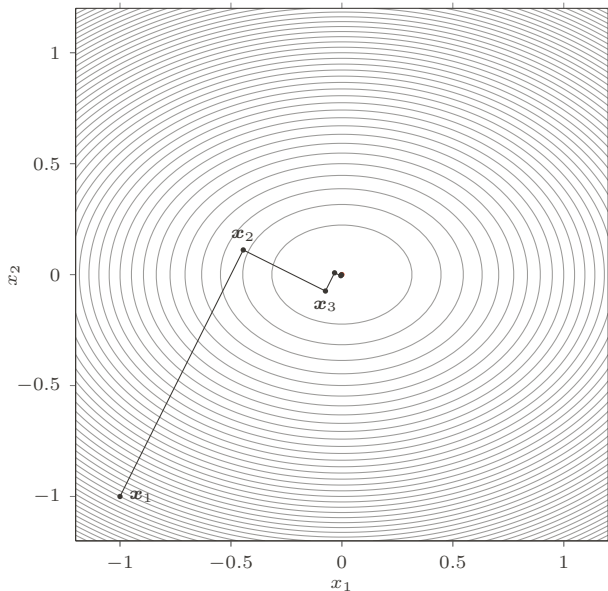
Backtracking dla $\alpha_{\max} = 1$, $\gamma = 0.2$, $\eta = 0.75$

$$f(x_1, x_2) = x_1^2 + 2x_2^2$$



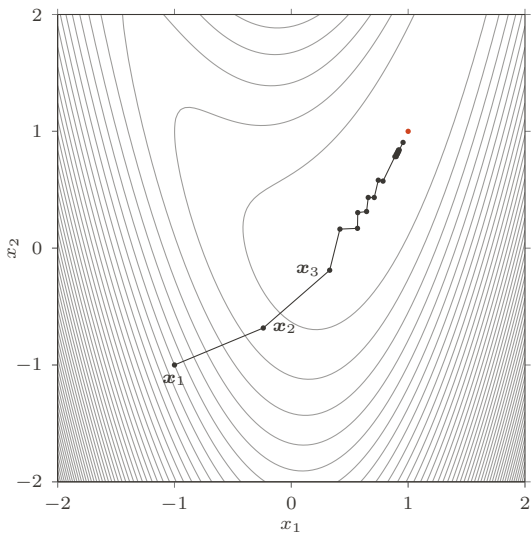
Dla porównania: algorytm najszybszego spadku

$$f(x_1, x_2) = x_1^2 + 2x_2^2$$



Backtracking dla $\alpha_{\max} = 1$, $\gamma = 0.5$, $\eta = 0.75$

$$f(x_1, x_2) = 2.5(x_1^2 - x_2)^2 + (1 - x_1)^2$$



Dla porównania: algorytm najszybszego spadku

$$f(x_1, x_2) = 2.5(x_1^2 - x_2)^2 + (1 - x_1)^2$$

