

Градиентный спуск

Мальковский Н. В.

Санкт-Петербургский академический университет



Общая идея градиентного спуска

$$\text{минимизировать } f(x), x \in \mathcal{D} \subset \mathbb{R}^n. \quad (1)$$

Условия стационарности: если $x^* \in \text{Int } \mathcal{D}$ – точка минимума f на \mathcal{D} , f дифференцируема в x^* , то

$$\nabla f(x^*) = 0_n.$$

Общая идея градиентного спуска

$$\text{минимизировать } f(x), x \in \mathcal{D} \subset \mathbb{R}^n. \quad (1)$$

Условия стационарности: если $x^* \in \text{Int } \mathcal{D}$ – точка минимума f на \mathcal{D} , f дифференцируема в x^* , то

$$\nabla f(x^*) = 0_n.$$

Пусть $x_0 \in \text{Int } \mathcal{D}$. Можно ли понять, где находится точка минимума по $\nabla f(x_0)$?

Общая идея градиентного спуска

$$\text{минимизировать } f(x), x \in \mathcal{D} \subset \mathbb{R}^n. \quad (1)$$

Условия стационарности: если $x^* \in \text{Int } \mathcal{D}$ – точка минимума f на \mathcal{D} , f дифференцируема в x^* , то

$$\nabla f(x^*) = 0_n.$$

Пусть $x_0 \in \text{Int } \mathcal{D}$. Можно ли понять, где находится точка минимума по $\nabla f(x_0)$?

Если немного сдвинуться из x_0 в направлении h , то получаем

$$f(x_0 + th) = f(x_0) + \nabla f(x_0)^T h + o(t).$$

Таким образом, локально выгоднее всего двигаться в направлении $h = -\nabla f(x_0)$.

Общая идея градиентного спуска

Оказывается, при некоторых предположениях на f и $0 < \alpha_k \in \mathbb{R}$ последовательность

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \quad (2)$$

сходится к точке минимума f .

Общая идея градиентного спуска

Оказывается, при некоторых предположениях на f и $0 < \alpha_k \in \mathbb{R}$ последовательность

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \quad (2)$$

сходится к точке минимума f .

Генерирование последовательности x_k по правилу (2) принято называть *градиентным спуском*. Величину α_k принято называть *размером шага* на k -ой итерации.

Общая идея градиентного спуска

Оказывается, при некоторых предположениях на f и $0 < \alpha_k \in \mathbb{R}$ последовательность

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \quad (2)$$

сходится к точке минимума f .

Генерирование последовательности x_k по правилу (2) принято называть *градиентным спуском*. Величину α_k принято называть *размером шага* на k -ой итерации.

В многих случаях легче измерить ∇f в нескольких точках, чтобы получить приближенное значение точки минимума нежели решать систему уравнений $\nabla f(x) = 0_n$.

Основные способы выбора шага

Наиболее распространенными способами выбора последовательности α_k в градиентном спуске являются следующие три:

- Постоянный шаг, $\alpha = \alpha_0 = \alpha_1 = \dots = \alpha_k = \dots$. При аккуратном выборе дает экспоненциальную скорость сходимости для сильно выпуклых функций.
- Последовательность, удовлетворяющая

$$\sum_{i=1} \alpha_k = \infty \quad (3a)$$

$$\sum_{i=1} \alpha_k^2 < \infty. \quad (3b)$$

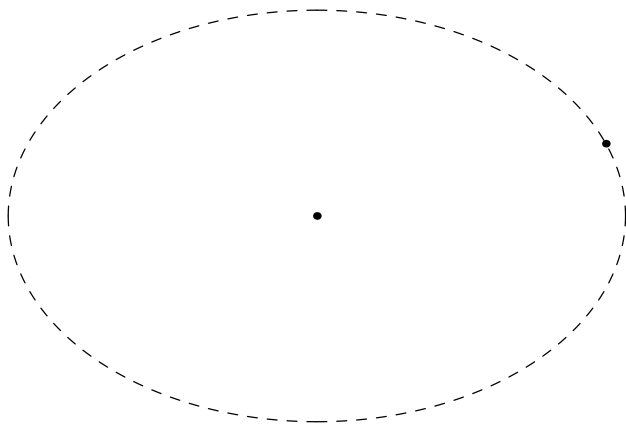
Такая последовательность всегда гарантирует сходимость к точке минимума для выпуклых функций. Проще всего брать $\alpha_k = \frac{c}{k}$.

- Минимум по направлению: α_k выбирается как

$$\alpha_k = \operatorname{argmin}_{\alpha} f(x_k - \alpha \nabla f(x_k)).$$

Обычно используется, если соответствующий минимум можно найти аналитически.

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_0 = 4.000000$$

$$y_0 = 1.000000$$

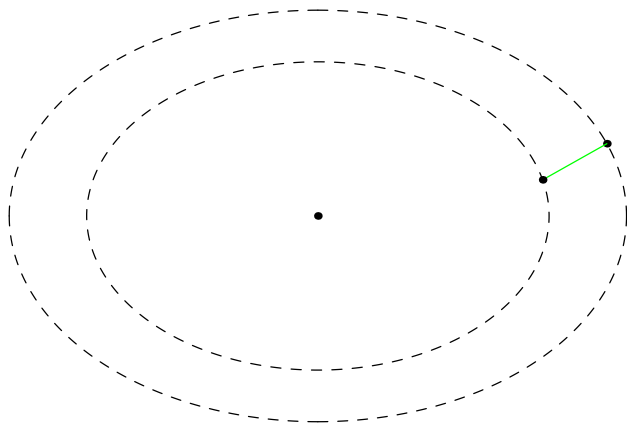
$$\nabla f(\cdot) = (0.888889,$$

$$0.500000)$$

$$\alpha_0 = 1$$

$$\sqrt{x_0^2 + y_0^2} = 4.123106$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_1 = 3.111111$$

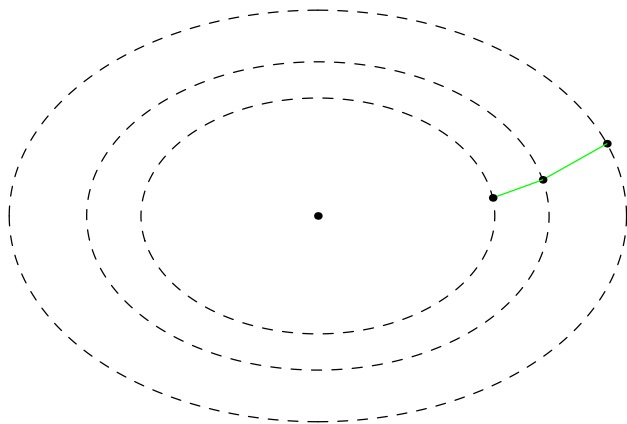
$$y_1 = 0.500000$$

$$\nabla f(\cdot) = (0.691358, \\ 0.250000)$$

$$\alpha_1 = 1$$

$$\sqrt{x_1^2 + y_1^2} = 3.151034$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_2 = 2.419753$$

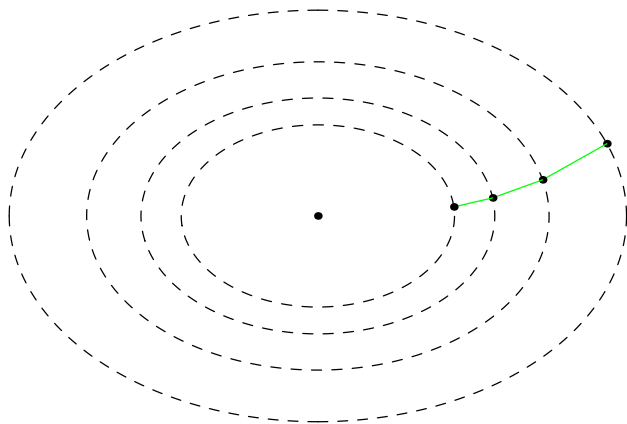
$$y_2 = 0.250000$$

$$\nabla f(\cdot) = (0.537723, \\ 0.125000)$$

$$\alpha_2 = 1$$

$$\sqrt{x_2^2 + y_2^2} = 2.432633$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_3 = 1.882030$$

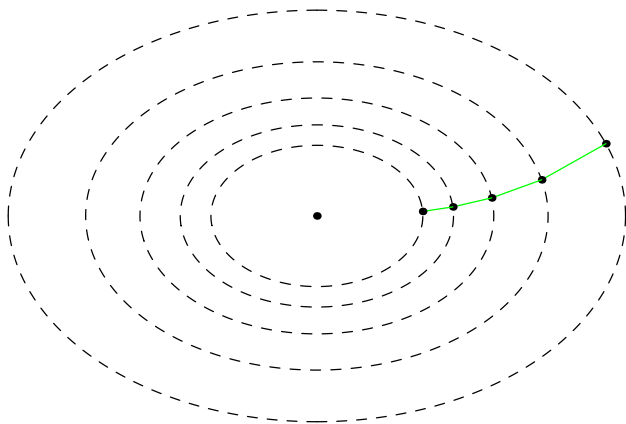
$$y_3 = 0.125000$$

$$\nabla f(\cdot) = (0.418229, \\ 0.062500)$$

$$\alpha_3 = 1$$

$$\sqrt{x_3^2 + y_3^2} = 1.886177$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_4 = 1.463801$$

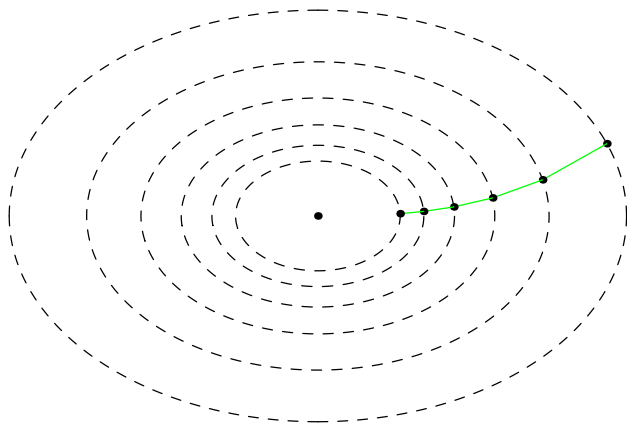
$$y_4 = 0.062500$$

$$\nabla f(\cdot) = (0.325289, \\ 0.031250)$$

$$\alpha_4 = 1$$

$$\sqrt{x_4^2 + y_4^2} = 1.465135$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_5 = 1.138512$$

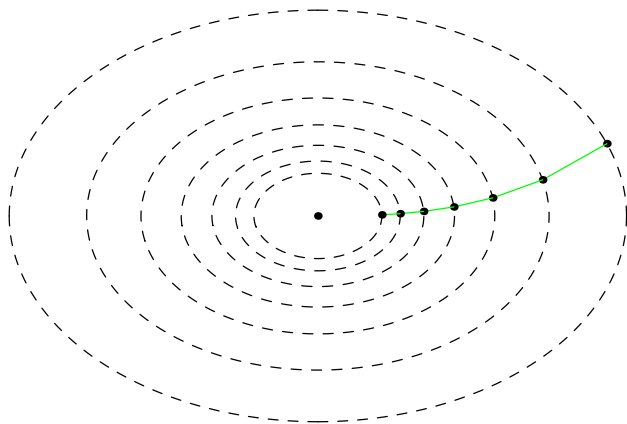
$$y_5 = 0.031250$$

$$\nabla f(\cdot) = (0.253003, \\ 0.015625)$$

$$\alpha_5 = 1$$

$$\sqrt{x_5^2 + y_5^2} = 1.138941$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_6 = 0.885509$$

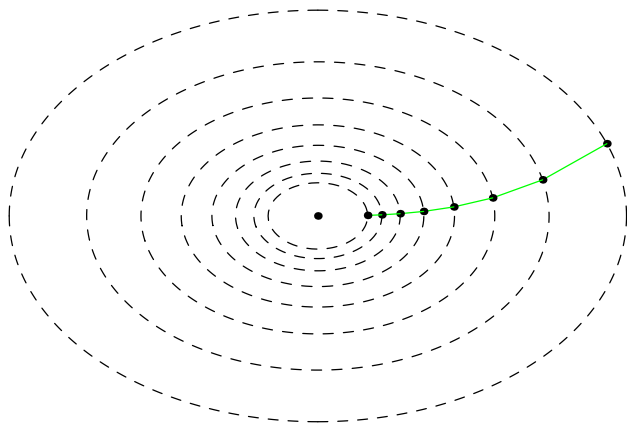
$$y_6 = 0.015625$$

$$\nabla f(\cdot) = (0.196780, \\ 0.007813)$$

$$\alpha_6 = 1$$

$$\sqrt{x_6^2 + y_6^2} = 0.885647$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_7 = 0.688730$$

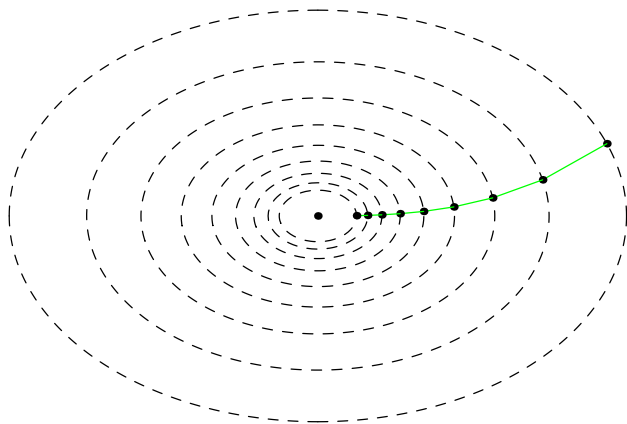
$$y_7 = 0.007813$$

$$\nabla f(\cdot) = (0.153051, \\ 0.003906)$$

$$\alpha_7 = 1$$

$$\sqrt{x_7^2 + y_7^2} = 0.688774$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_8 = 0.535679$$

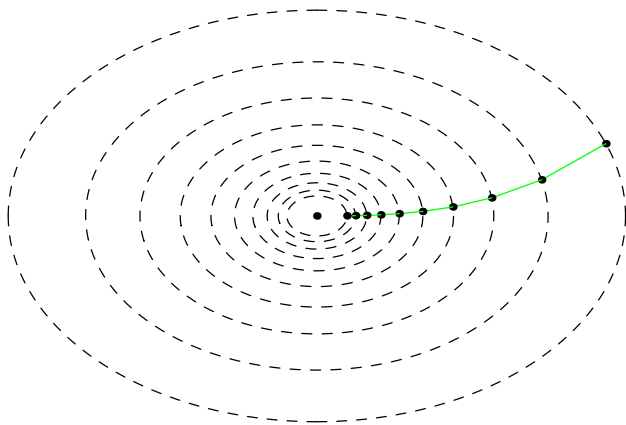
$$y_8 = 0.003906$$

$$\nabla f(\cdot) = (0.119040, \\ 0.001953)$$

$$\alpha_8 = 1$$

$$\sqrt{x_8^2 + y_8^2} = 0.535693$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_9 = 0.416639$$

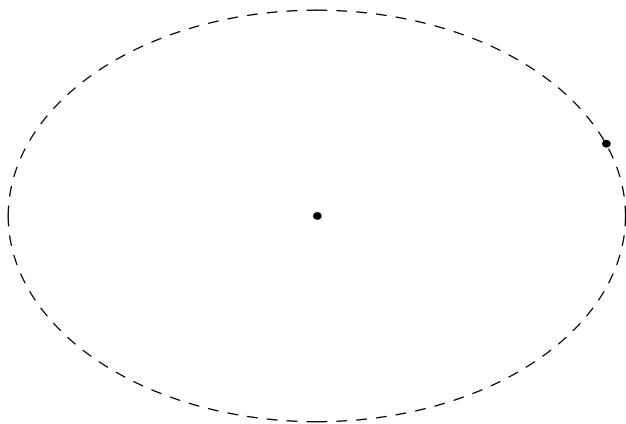
$$y_9 = 0.001953$$

$$\nabla f(\cdot) = (0.092586, \\ 0.000977)$$

$$\alpha_9 = 1$$

$$\sqrt{x_9^2 + y_9^2} = 0.416643$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_0 = 4.000000$$

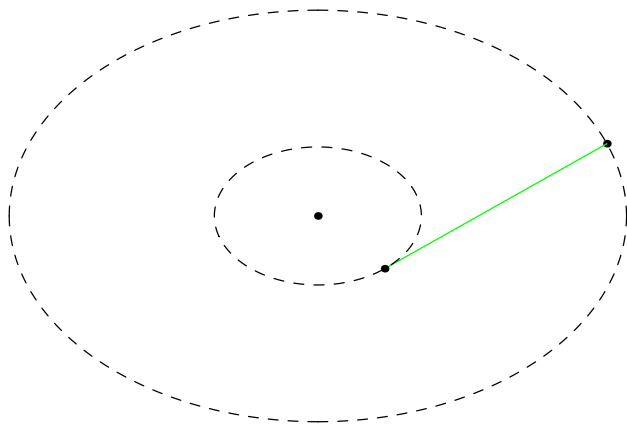
$$y_0 = 1.000000$$

$$\nabla f(\cdot) = (0.888889, \\ 0.500000)$$

$$\alpha_0 = 3.460354$$

$$\sqrt{x_0^2 + y_0^2} = 4.123106$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_1 = 0.924130$$

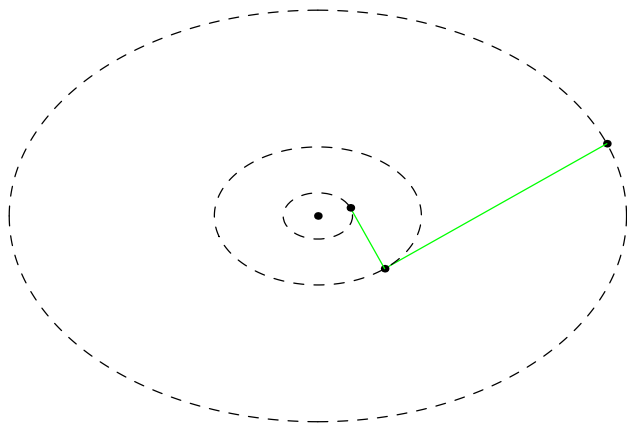
$$y_1 = -0.730177$$

$$\nabla f(\cdot) = (0.205362, \\ -0.365088)$$

$$\alpha_1 = 2.308219$$

$$\sqrt{x_1^2 + y_1^2} = 1.177784$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_2 = 0.450109$$

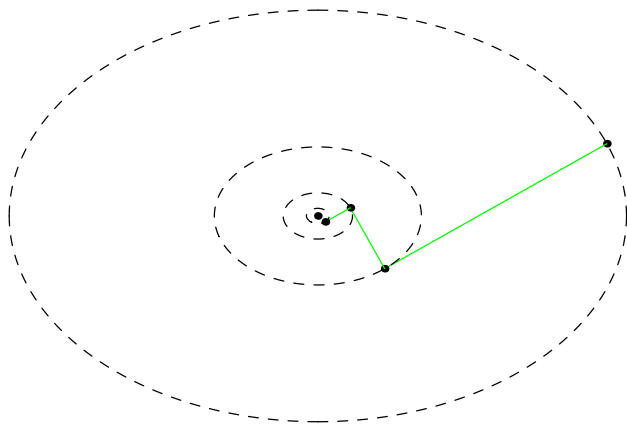
$$y_2 = 0.112527$$

$$\nabla f(\cdot) = (0.100024, \\ 0.056264)$$

$$\alpha_2 = 3.460354$$

$$\sqrt{x_2^2 + y_2^2} = 0.463962$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_3 = 0.103990$$

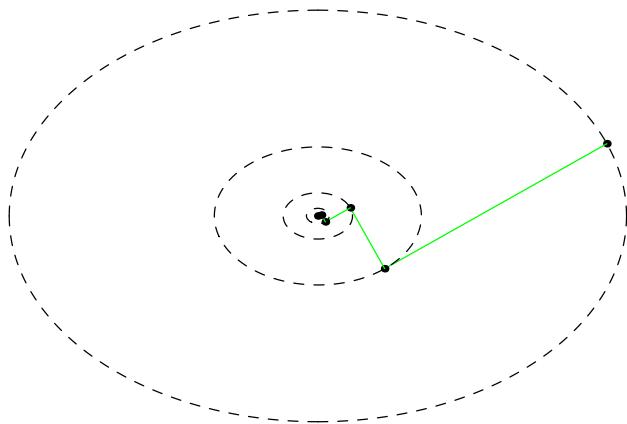
$$y_3 = -0.082165$$

$$\nabla f(\cdot) = (0.023109, \\ -0.041082)$$

$$\alpha_3 = 2.308219$$

$$\sqrt{x_3^2 + y_3^2} = 0.132533$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_4 = 0.050650$$

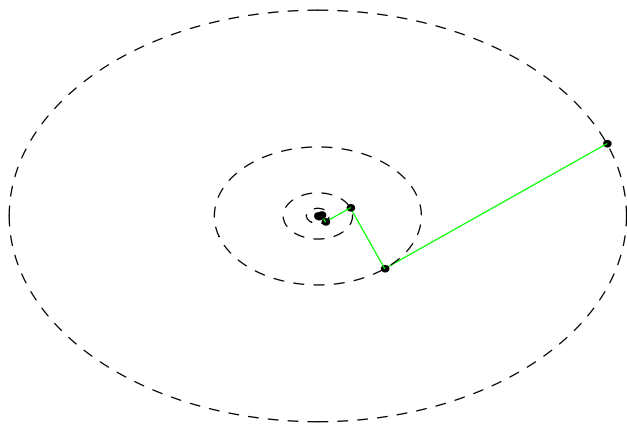
$$y_4 = 0.012662$$

$$\nabla f(\cdot) = (0.011255, \\ 0.006331)$$

$$\alpha_4 = 3.460354$$

$$\sqrt{x_4^2 + y_4^2} = 0.052208$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_5 = 0.011702$$

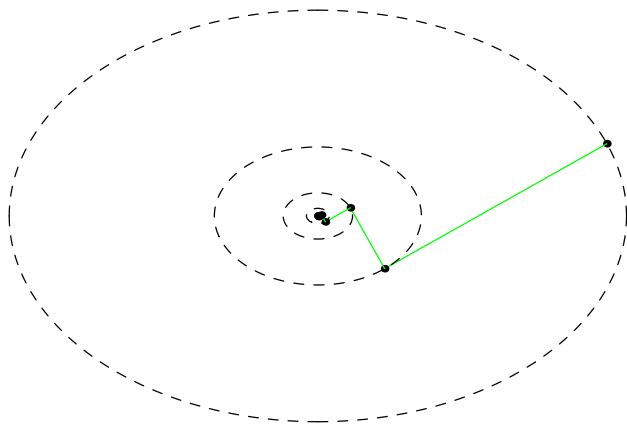
$$y_5 = -0.009246$$

$$\nabla f(\cdot) = (0.002600, \\ -0.004623)$$

$$\alpha_5 = 2.308219$$

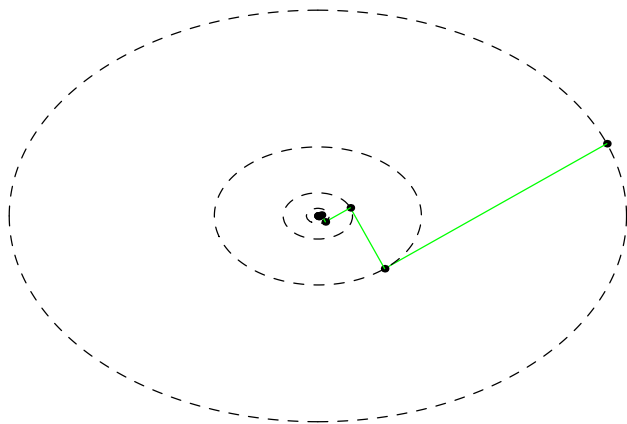
$$\sqrt{x_5^2 + y_5^2} = 0.014914$$

Пример: градиентный спуск для квадратичной функции



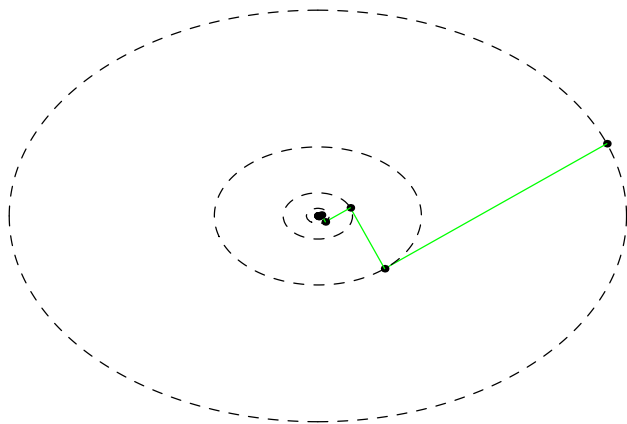
$$\begin{aligned}f(x, y) &= \frac{x^2}{9} + \frac{y^2}{4} \\x_6 &= 0.005699 \\y_6 &= 0.001425 \\\nabla f(\cdot) &= (0.001267, \\&\quad 0.000712) \\\alpha_6 &= 3.460354 \\\sqrt{x_6^2 + y_6^2} &= 0.005875\end{aligned}$$

Пример: градиентный спуск для квадратичной функции



$$\begin{aligned}f(x, y) &= \frac{x^2}{9} + \frac{y^2}{4} \\x_7 &= 0.001317 \\y_7 &= -0.001040 \\\nabla f(\cdot) &= (0.000293, \\&\quad -0.000520) \\\alpha_7 &= 2.308219 \\\sqrt{x_7^2 + y_7^2} &= 0.001678\end{aligned}$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_8 = 0.000641$$

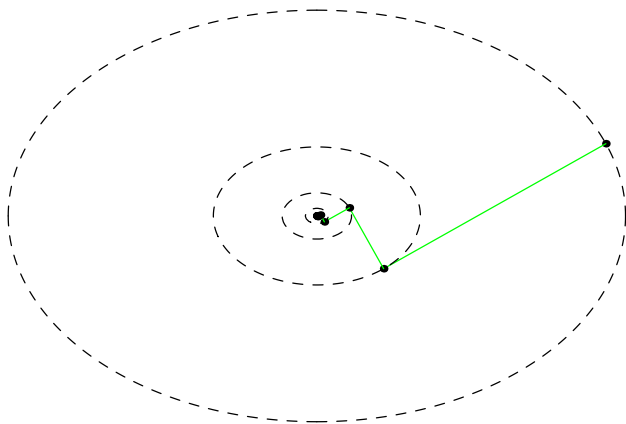
$$y_8 = 0.000160$$

$$\nabla f(\cdot) = (0.000143, \\ 0.000080)$$

$$\alpha_8 = 3.460354$$

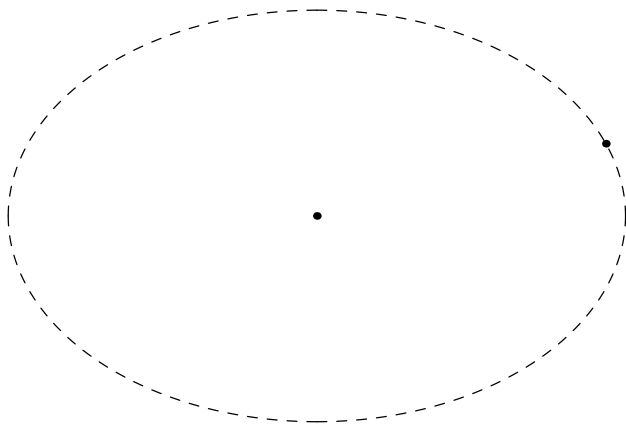
$$\sqrt{x_8^2 + y_8^2} = 0.000661$$

Пример: градиентный спуск для квадратичной функции



$$\begin{aligned}f(x, y) &= \frac{x^2}{9} + \frac{y^2}{4} \\x_9 &= 0.000148 \\y_9 &= -0.000117 \\\nabla f(\cdot) &= (0.000033, \\&\quad -0.000059) \\\alpha_9 &= 2.308219 \\\sqrt{x_9^2 + y_9^2} &= 0.000189\end{aligned}$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_0 = 4.000000$$

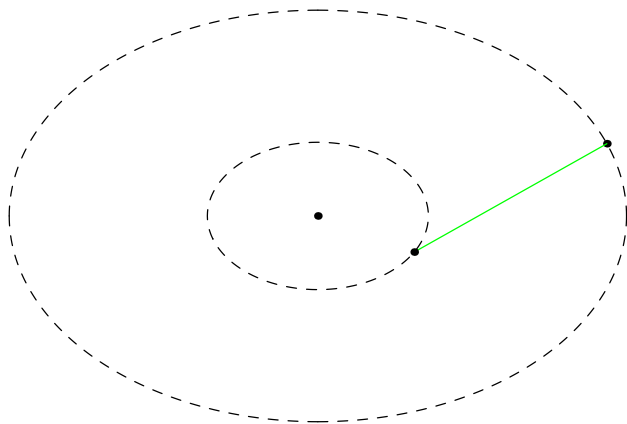
$$y_0 = 1.000000$$

$$\nabla f(\cdot) = (0.888889, \\ 0.500000)$$

$$\alpha_0 = 3/1$$

$$\sqrt{x_0^2 + y_0^2} = 4.123106$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_1 = 1.333333$$

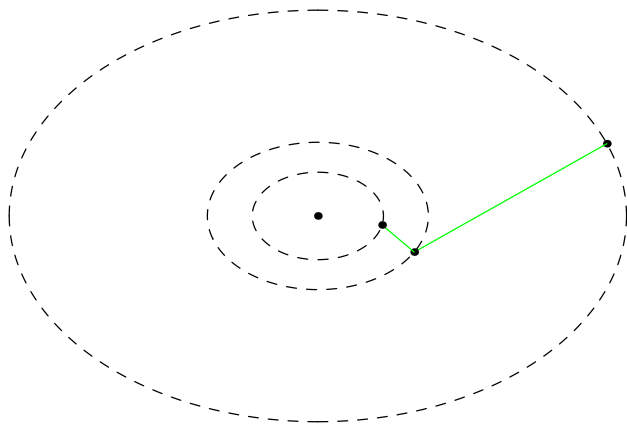
$$y_1 = -0.500000$$

$$\nabla f(\cdot) = (0.296296, \\ -0.250000)$$

$$\alpha_1 = 3/2$$

$$\sqrt{x_1^2 + y_1^2} = 1.424001$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_2 = 0.888889$$

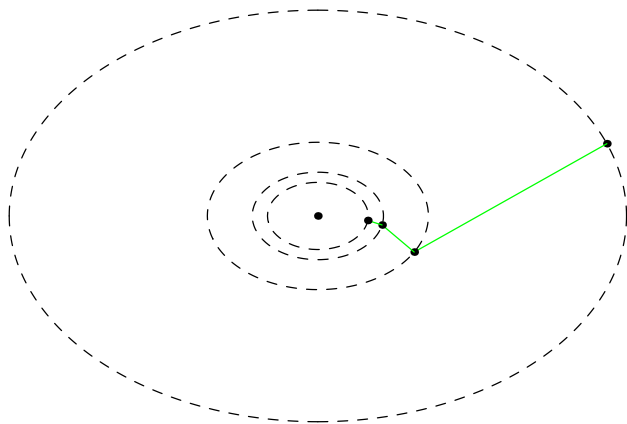
$$y_2 = -0.125000$$

$$\nabla f(\cdot) = (0.197531, \\ -0.062500)$$

$$\alpha_2 = 3/3$$

$$\sqrt{x_2^2 + y_2^2} = 0.897635$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_3 = 0.691358$$

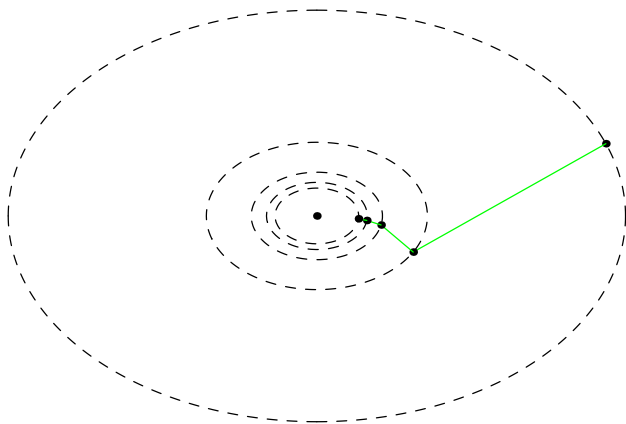
$$y_3 = -0.062500$$

$$\nabla f(\cdot) = (0.153635, \\ -0.031250)$$

$$\alpha_3 = 3/4$$

$$\sqrt{x_3^2 + y_3^2} = 0.694177$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_4 = 0.576132$$

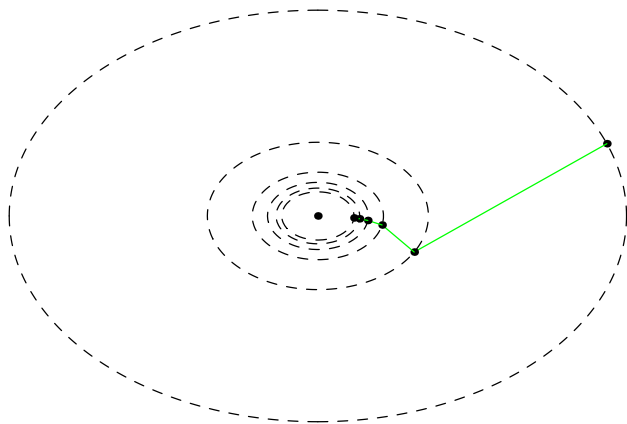
$$y_4 = -0.039063$$

$$\nabla f(\cdot) = (0.128029, \\ -0.019531)$$

$$\alpha_4 = 3/5$$

$$\sqrt{x_4^2 + y_4^2} = 0.577454$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_5 = 0.499314$$

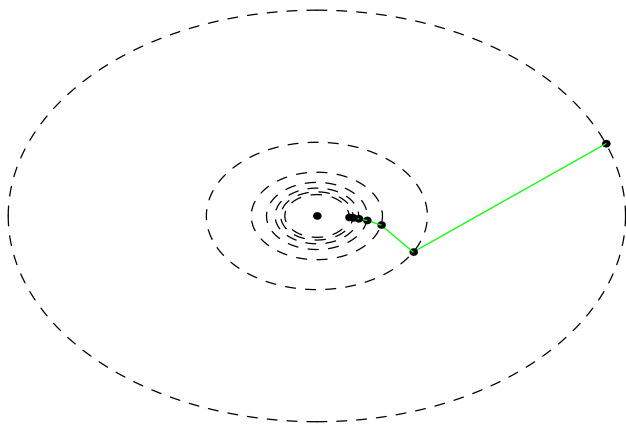
$$y_5 = -0.027344$$

$$\nabla f(\cdot) = (0.110959, \\ -0.013672)$$

$$\alpha_5 = 3/6$$

$$\sqrt{x_5^2 + y_5^2} = 0.500062$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_6 = 0.443835$$

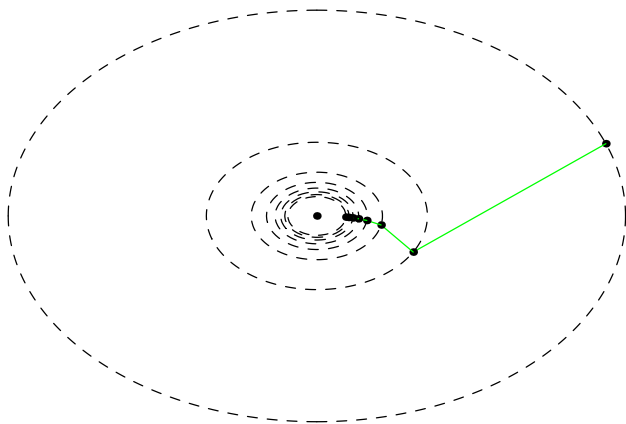
$$y_6 = -0.020508$$

$$\nabla f(\cdot) = (0.098630, \\ -0.010254)$$

$$\alpha_6 = 3/7$$

$$\sqrt{x_6^2 + y_6^2} = 0.444308$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_7 = 0.401565$$

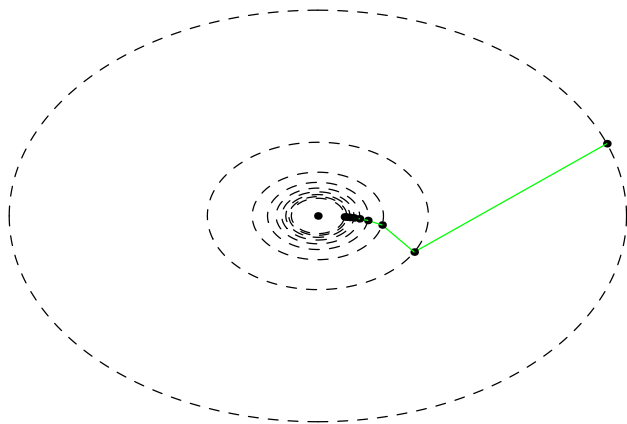
$$y_7 = -0.016113$$

$$\nabla f(\cdot) = (0.089237, \\ -0.008057)$$

$$\alpha_7 = 3/8$$

$$\sqrt{x_7^2 + y_7^2} = 0.401888$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_8 = 0.368101$$

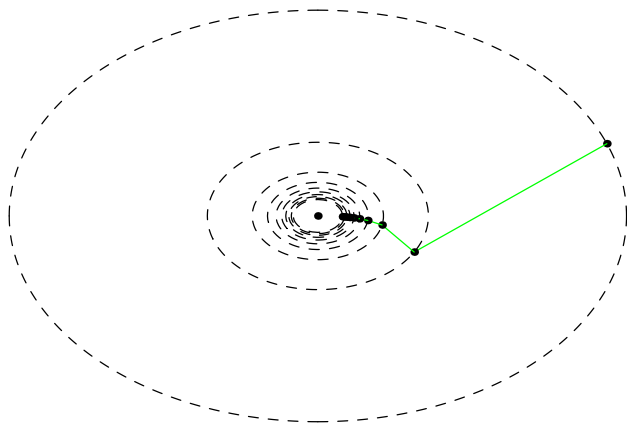
$$y_8 = -0.013092$$

$$\nabla f(\cdot) = (0.081800, \\ -0.006546)$$

$$\alpha_8 = 3/9$$

$$\sqrt{x_8^2 + y_8^2} = 0.368334$$

Пример: градиентный спуск для квадратичной функции



$$f(x, y) = \frac{x^2}{9} + \frac{y^2}{4}$$

$$x_9 = 0.340834$$

$$y_9 = -0.010910$$

$$\nabla f(\cdot) = (0.075741, \\ -0.005455)$$

$$\alpha_9 = 3/10$$

$$\sqrt{x_9^2 + y_9^2} = 0.341009$$

Предположения о минимизируемой функции

В дальнейшем анализе будем полагать, что $f : \mathcal{D} \subset \mathbb{R}^n \rightarrow \mathbb{R}$ выпукла и дифференцируема на \mathcal{D} . Обозначим

$$S_f(x) = \{y \in \mathcal{D} \mid f(y) \leq f(x)\}.$$

Также будут использоваться некоторые из следующих предположений:

Предположения о минимизируемой функции

В дальнейшем анализе будем полагать, что $f : \mathcal{D} \subset \mathbb{R}^n \rightarrow \mathbb{R}$ выпукла и дифференцируема на \mathcal{D} . Обозначим

$$S_f(x) = \{y \in \mathcal{D} \mid f(y) \leq f(x)\}.$$

Также будут использоваться некоторые из следующих предположений:

- Градиент f непрерывен по Липшицу с константой M , т.е.

$$\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\| \quad \forall x, y \in S_f(x_0).$$

Предположения о минимизируемой функции

В дальнейшем анализе будем полагать, что $f : \mathcal{D} \subset \mathbb{R}^n \rightarrow \mathbb{R}$ выпукла и дифференцируема на \mathcal{D} . Обозначим

$$S_f(x) = \{y \in \mathcal{D} \mid f(y) \leq f(x)\}.$$

Также будут использоваться некоторые из следующих предположений:

- Градиент f непрерывен по Липшицу с константой M , т.е.

$$\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\| \quad \forall x, y \in S_f(x_0).$$

- f – сильно выпуклая функция с параметром m на $S_f(x_0)$, т.е.
 $\forall x, y \in S_f(x_0)$

$$(\nabla f(y) - \nabla f(x))^T (y - x) \geq m\|y - x\|^2$$

Предположения о минимизируемой функции

В дальнейшем анализе будем полагать, что $f : \mathcal{D} \subset \mathbb{R}^n \rightarrow \mathbb{R}$ выпукла и дифференцируема на \mathcal{D} . Обозначим

$$S_f(x) = \{y \in \mathcal{D} \mid f(y) \leq f(x)\}.$$

Также будут использоваться некоторые из следующих предположений:

- Градиент f непрерывен по Липшицу с константой M , т.е.

$$\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\| \quad \forall x, y \in S_f(x_0).$$

- f – сильно выпуклая функция с параметром m на $S_f(x_0)$, т.е.
 $\forall x, y \in S_f(x_0)$

$$(\nabla f(y) - \nabla f(x))^T (y - x) \geq m\|y - x\|^2$$

Предположения о минимизируемой функции

В дальнейшем анализе будем полагать, что $f : \mathcal{D} \subset \mathbb{R}^n \rightarrow \mathbb{R}$ выпукла и дифференцируема на \mathcal{D} . Обозначим

$$S_f(x) = \{y \in \mathcal{D} \mid f(y) \leq f(x)\}.$$

Также будут использоваться некоторые из следующих предположений:

- Градиент f непрерывен по Липшицу с константой M , т.е.

$$\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\| \quad \forall x, y \in S_f(x_0).$$

- f – сильно выпуклая функция с параметром m на $S_f(x_0)$, т.е.
 $\forall x, y \in S_f(x_0)$

$$(\nabla f(y) - \nabla f(x))^T (y - x) \geq m\|y - x\|^2$$

Замечание. $S_f(x)$ – выпуклое множество, если f выпукла, более того $S_f(x)$ всегда ограничено, если f сильно выпукла.

Сходимость градиентного спуска (постоянный шаг)

Теорема (Постоянный шаг)

Пусть f дифференцируема и выпукла на \mathcal{D} , $\alpha_k \equiv \alpha \in (0, 1/M]$, градиент f липшицев с константой $M > 0$ на $S_f(x_0)$, f ограничена снизу и существует хотя бы одна точка минимума x^* , тогда для последовательности x_k , генерируемой по правилу (2) $f(x_k)$ убывает и, более того

$$f(x_k) - f(x^*) \leq \frac{1}{2\alpha k} \|x_0 - x^*\|^2.$$

Сходимость градиентного спуска (постоянный шаг)

Док-во. Используя непрерывность по Липшицу и $x_{k+1} - x_k = \alpha \nabla f(x_k)$

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{M}{2} \|x_{k+1} - x_k\|^2 \\ &= -\alpha \left(1 - \frac{\alpha M}{2}\right) \|\nabla f(x_k)\|^2 \end{aligned}$$

Сходимость градиентного спуска (постоянный шаг)

Док-во. Используя непрерывность по Липшицу и $x_{k+1} - x_k = \alpha \nabla f(x_k)$

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{M}{2} \|x_{k+1} - x_k\|^2 \\ &= -\alpha \left(1 - \frac{\alpha M}{2}\right) \|\nabla f(x_k)\|^2 \end{aligned}$$

Таким образом $f(x_k)$ убывает в силу $0 < \alpha < 2/M$, что гарантирует $x_k \in S_f(x_0)$.

Сходимость градиентного спуска (постоянный шаг)

Док-во. Используя непрерывность по Липшицу и $x_{k+1} - x_k = \alpha \nabla f(x_k)$

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{M}{2} \|x_{k+1} - x_k\|^2 \\ &= -\alpha \left(1 - \frac{\alpha M}{2}\right) \|\nabla f(x_k)\|^2 \end{aligned}$$

Таким образом $f(x_k)$ убывает в силу $0 < \alpha < 2/M$, что гарантирует $x_k \in S_f(x_0)$. С другой стороны

$$f(x_k) - f(x^*) \geq f(x_k) - f(x_{k+1}) \geq \alpha \left(1 - \frac{\alpha M}{2}\right) \|\nabla f(x_k)\|^2.$$

Так как это неравенство выполняется при любом $\alpha \in (0, 2/M)$ и любом x_k , то минимизируя по α (минимум при $\alpha = 1/M$) получаем

$$f(x) - f(x^*) \geq \frac{1}{2M} \|\nabla f(x)\|^2 \quad (4)$$

Сходимость градиентного спуска (постоянный шаг)

Вернемся на шаг назад, при условии $\alpha \leq 1/M$

$$\begin{aligned}f(x_{i+1}) &\leq f(x_i) - \alpha \left(1 - \frac{\alpha M}{2}\right) \|\nabla f(x_i)\|^2 \\&\leq f(x_i) - \frac{\alpha}{2} \|\nabla f(x_i)\|^2 \\&\leq f(x^*) + \nabla f(x_i)^T (x_i - x^*) - \frac{\alpha}{2} \|\nabla f(x_i)\|^2 \\&= f(x^*) + \frac{1}{2\alpha} (\|x_i - x^*\|^2 - \|x_i - x^* - \alpha \nabla f(x_i)\|^2) \\&= f(x^*) + \frac{1}{2\alpha} (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2).\end{aligned}$$

Сходимость градиентного спуска (постоянный шаг)

Вернемся на шаг назад, при условии $\alpha \leq 1/M$

$$\begin{aligned}f(x_{i+1}) &\leq f(x_i) - \alpha \left(1 - \frac{\alpha M}{2}\right) \|\nabla f(x_i)\|^2 \\&\leq f(x_i) - \frac{\alpha}{2} \|\nabla f(x_i)\|^2 \\&\leq f(x^*) + \nabla f(x_i)^T (x_i - x^*) - \frac{\alpha}{2} \|\nabla f(x_i)\|^2 \\&= f(x^*) + \frac{1}{2\alpha} (\|x_i - x^*\|^2 - \|x_i - x^* - \alpha \nabla f(x_i)\|^2) \\&= f(x^*) + \frac{1}{2\alpha} (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2).\end{aligned}$$

Суммируя по $i = 0 \dots k - 1$ получаем

$$\begin{aligned}\sum_{i=1}^k (f(x_i) - f(x^*)) &\leq \frac{1}{2\alpha} \sum_{i=1}^k (\|x_{i-1} - x^*\|^2 - \|x_i - x^*\|^2) \\&= \frac{1}{2\alpha} (\|x_0 - x^*\|^2 - \|x_k - x^*\|^2) \leq \frac{1}{2\alpha} \|x_0 - x^*\|^2.\end{aligned}$$

Сходимость градиентного спуска

Так как $f(x_k)$ убывает, то

$$f(x_k) - f(x^*) \leq \frac{1}{k} \sum_{i=1}^k (f(x_i) - f(x^*)) \leq \frac{1}{2\alpha k} \|x_0 - x^*\|^2.$$

Сходимость градиентного спуска (постоянный шаг)

Теорема (Постоянный шаг, сильная выпуклость)

Пусть f дифференцируема и выпукла на \mathcal{D} , $\alpha_k \equiv \alpha \in (0, 2/M)$, f сильно выпукла с константой $m > 0$ на $S_f(x_0)$, градиент f липшицев с константой $M \geq m$ на $S_f(x_0)$, тогда для последовательности x_k , генерируемой по правилу (2) выполняется:

- x_k сходится к единственной точке минимума f на \mathcal{D} , более того для $q = 1 - 2m\alpha + mM\alpha^2$

$$\|x_k - x^*\|^2 \leq \frac{8M}{m^2} q^k (f(x_0) - f(x^*)).$$

- $f(x_k)$ убывает и сходится к $f(x^*)$, более того

$$f(x_k) - f(x^*) \leq q^k (f(x_0) - f(x^*)).$$

Сходимость градиентного спуска (постоянный шаг)

Док-во. Из сильной выпуклости

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|y - x\|^2.$$

Минимизирую правую часть по y (минимум при $y = x - (1/m)\nabla f(x)$) получаем

$$f(y) \geq f(x) - \frac{1}{2m} \|\nabla f(x)\|^2.$$

В частности

$$f(x) - f(x^*) \leq \frac{1}{2m} \|\nabla f(x)\|^2 \quad (5)$$

Сходимость градиентного спуска (постоянный шаг)

Док-во. Из сильной выпуклости

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|y - x\|^2.$$

Минимизирую правую часть по y (минимум при $y = x - (1/m)\nabla f(x)$) получаем

$$f(y) \geq f(x) - \frac{1}{2m} \|\nabla f(x)\|^2.$$

В частности

$$f(x) - f(x^*) \leq \frac{1}{2m} \|\nabla f(x)\|^2 \quad (5)$$

Наконец, вновь воспользовавшись сильной выпуклостью

$$\begin{aligned} 0 \geq f(x^*) - f(x) &\geq \nabla f(x)^T (x^* - x) + \frac{m}{2} \|x - x^*\|^2 \geq \\ & - \|\nabla f(x)\| \cdot \|x^* - x\| + \frac{m}{2} \|x - x^*\|^2, \end{aligned}$$

а значит

$$\|x - x^*\| \leq \frac{2}{m} \|\nabla f(x)\|. \quad (6)$$

Сходимость градиентного спуска (постоянный шаг)

Далее, так как $f(x_k)$ убывает, а f ограничена снизу, то $f(x_k)$ сходится, более того

$$f(x_0) - f(x^*) \geq \sum_{k=0}^{\infty} f(x_k) - f(x_{k+1}) \geq \alpha \left(1 - \frac{\alpha M}{2}\right) \sum_{k=0}^{\infty} \|\nabla f(x_k)\|^2.$$

Сходимость градиентного спуска (постоянный шаг)

Далее, так как $f(x_k)$ убывает, а f ограничена снизу, то $f(x_k)$ сходится, более того

$$f(x_0) - f(x^*) \geq \sum_{k=0}^{\infty} f(x_k) - f(x_{k+1}) \geq \alpha \left(1 - \frac{\alpha M}{2}\right) \sum_{k=0}^{\infty} \|\nabla f(x_k)\|^2.$$

Таким образом ряд $\sum_{k=0}^{\infty} \|\nabla f(x_k)\|^2$ сходится $\Rightarrow \|\nabla f(x_k)\| \rightarrow 0$, в силу (6) $x_k \rightarrow x^*$ и, следовательно $f(x_k) \rightarrow f(x^*)$.

Сходимость градиентного спуска (постоянный шаг)

Далее, так как $f(x_k)$ убывает, а f ограничена снизу, то $f(x_k)$ сходится, более того

$$f(x_0) - f(x^*) \geq \sum_{k=0}^{\infty} f(x_k) - f(x_{k+1}) \geq \alpha \left(1 - \frac{\alpha M}{2}\right) \sum_{k=0}^{\infty} \|\nabla f(x_k)\|^2.$$

Таким образом ряд $\sum_{k=0}^{\infty} \|\nabla f(x_k)\|^2$ сходится $\Rightarrow \|\nabla f(x_k)\| \rightarrow 0$, в силу (6) $x_k \rightarrow x^*$ и, следовательно $f(x_k) \rightarrow f(x^*)$.

Далее, оценим скорость сходимости: вернемся к неравенству

$$f(x_{k+1}) \leq f(x_k) - \alpha \left(1 - \frac{\alpha M}{2}\right) \|\nabla f(x_k)\|^2.$$

Вычитая из обеих частей $f(x^*)$ и используя (5) получаем

$$f(x_{k+1}) - f(x^*) \leq f(x_k) - f(x^*) - \alpha \left(1 - \frac{\alpha M}{2}\right) 2m(f(x_k) - f(x^*))$$

Сходимость градиентного спуска (постоянный шаг)

Таким образом

$$f(x_k) - f(x^*) \leq q(f(x_{k-1}) - f(x^*)) \leq q^k(f(x_0) - f(x^*)).$$

Используя (4) и (6) получаем

$$\|x_k - x^*\|^2 \leq \frac{8M}{m^2} q^k (f(x_0) - f(x^*)). \blacksquare$$

Сходимость (замечания)

Замечание 1. Значение q достигает минимума при $\alpha = 1/M$ и равно $1 - \frac{m}{M}$.

Сходимость (замечания)

Замечание 1. Значение q достигает минимума при $\alpha = 1/M$ и равно $1 - \frac{m}{M}$.

Замечание 2. Очевидным образом, полученные оценки верны в случае, если α_k выбирается как минимум по направлению. К сожалению, улучшения при этом получить не удастся.

Сходимость (замечания)

Замечание 1. Значение q достигает минимума при $\alpha = 1/M$ и равно $1 - \frac{m}{M}$.

Замечание 2. Очевидным образом, полученные оценки верны в случае, если α_k выбирается как минимум по направлению. К сожалению, улучшения при этом получить не удастся.

Замечание 3. Если f дважды дифференцируема, то q можно уменьшить с $\frac{M-m}{M}$ до $\left(\frac{M-m}{M+m}\right)^2$: из условий теоремы $mI \preceq \nabla^2 f(\cdot) \preceq MI$, по формуле Ньютона-Лейбница

$$\nabla f(x_k) = \nabla f(x^*) + \int_0^1 \nabla^2 f(x^* + t(x_k - x^*))(x_k - x^*) dt = A_k(x_k - x^*).$$

Так как $mI \preceq \nabla^2 f(\cdot) \preceq MI$, то $mI \preceq A_k \preceq MI$. Отсюда выводим

$$\|x_k - x^*\| = \|x_{k-1} - x^* - \alpha \nabla f(x_{k-1})\| \leq \|I - \alpha A_{k-1}\| \cdot \|x_{k-1} - x^*\|.$$

Так как A_k – симметричная матрица, у которой все собственные числа лежат на отрезке $[m, M]$, то $\|I - \alpha A_k\| = \max\{|1 - \alpha m|, |1 - \alpha M|\}$, которое достигает минимума при $\alpha = \frac{2}{M+m}$ и равно $\frac{M-m}{M+m}$.

Сходимость градиентного спуска

Теорема (Сходимость градиентного спуска с переменным шагом)

Пусть f непрерывно дифференцируема и выпукла на \mathcal{D} , градиент f ограничен на \mathcal{D} , $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, последовательность x_k генерируется по правилу (2), тогда для $\phi_k = \min_{i=0}^k f(x_k)$ выполняется

$$\phi_k - f(x^*) \rightarrow 0$$

Сходимость градиентного спуска (при условии (3))

Док-во.

$$\begin{aligned} 0 \leq \|x_{k+1} - x^*\|^2 &= \|x_k - x^*\|^2 - 2\alpha_k \nabla f(x_k)^T (x_k - x^*) + \alpha_k^2 \|\nabla f(x_k)\|^2 \\ &\leq \|x_k - x^*\|^2 - 2\alpha_k (f(x_k) - f(x^*)) + \alpha_k^2 \|\nabla f(x_k)\|^2 \\ &\leq \|x_k - x^*\|^2 - 2\alpha_k (\phi_k - f(x^*)) + \alpha_k^2 \|\nabla f(x_k)\|^2. \end{aligned}$$

Сходимость градиентного спуска (при условии (3))

Док-во.

$$\begin{aligned} 0 \leq \|x_{k+1} - x^*\|^2 &= \|x_k - x^*\|^2 - 2\alpha_k \nabla f(x_k)^T (x_k - x^*) + \alpha_k^2 \|\nabla f(x_k)\|^2 \\ &\leq \|x_k - x^*\|^2 - 2\alpha_k (f(x_k) - f(x^*)) + \alpha_k^2 \|\nabla f(x_k)\|^2 \\ &\leq \|x_k - x^*\|^2 - 2\alpha_k (\phi_k - f(x^*)) + \alpha_k^2 \|\nabla f(x_k)\|^2. \end{aligned}$$

Отсюда

$$2\alpha_k (\phi_k - f(x^*)) \leq \alpha_k^2 \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2.$$

Сходимость градиентного спуска (при условии (3))

Док-во.

$$\begin{aligned} 0 \leq \|x_{k+1} - x^*\|^2 &= \|x_k - x^*\|^2 - 2\alpha_k \nabla f(x_k)^T (x_k - x^*) + \alpha_k^2 \|\nabla f(x_k)\|^2 \\ &\leq \|x_k - x^*\|^2 - 2\alpha_k (f(x_k) - f(x^*)) + \alpha_k^2 \|\nabla f(x_k)\|^2 \\ &\leq \|x_k - x^*\|^2 - 2\alpha_k (\phi_k - f(x^*)) + \alpha_k^2 \|\nabla f(x_k)\|^2. \end{aligned}$$

Отсюда

$$2\alpha_k (\phi_k - f(x^*)) \leq \alpha_k^2 \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2.$$

В силу ограниченности $\nabla f(\cdot)$ на \mathcal{D} существует константа C : $\|\nabla f(x_k)\| \leq C$.
Суммируя по k получаем

$$2 \sum_{k=0}^{\infty} \alpha_k (\phi_k - f(x^*)) \leq \sum_{k=0}^{\infty} \alpha_k^2 C^2 + \|x_0 - x^*\|^2.$$

Таким образом ряд в левой части неравенства сходится. Так как ϕ_k убывает, $\phi_k \geq f(x^*)$, а $\sum_{k=1}^{\infty} \alpha_k$ расходится, то $\phi_k \rightarrow f(x^*)$ является необходимым условием сходимости ряда. ■

Сходимость градиентного спуска

Замечание 1. Вместо условия $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$ можно использовать более слабое условие $\alpha_k \rightarrow 0$.

Сходимость градиентного спуска

Замечание 1. Вместо условия $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$ можно использовать более слабое условие $\alpha_k \rightarrow 0$.

Замечание 2. При использовании нормированного шага $\gamma_k = \alpha_k / \|\nabla f(x_k)\|$ имеет место сходимость $x_k \rightarrow x^*$, где x^* – некоторая точка минимума f (при условии существования хотя бы одной).

Сходимость градиентного спуска при $\alpha_k = \mathcal{O}(1/k)$

Замечание 3. Если f сильно выпукла, а $\alpha_k = \mathcal{O}(\frac{1}{k})$, то x_k сходится к x^* со скоростью $\mathcal{O}(\frac{1}{\sqrt{k}})$.

Сходимость градиентного спуска при $\alpha_k = \mathcal{O}(1/k)$

Замечание 3. Если f сильно выпукла, а $\alpha_k = \mathcal{O}(1/k)$, то x_k сходится к x^* со скоростью $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$.

Лемма

Если при $\beta > 1$ для числовой последовательности α_k выполняется $\alpha_k \geq 0$ и

$$\alpha_{k+1} \leq \alpha_k \left(1 - \frac{\beta}{k+1}\right) + \frac{\gamma}{(k+1)^2},$$

то при $D = \max\left\{\frac{\gamma}{\beta-1}, \alpha_0\right\}$ выполняется

$$\alpha_k \leq \frac{D}{k+1}.$$

Сходимость градиентного спуска

Док-во. Докажем по индукции: очевидным образом база верна.

Предположим, что утверждение верно для k , выведем верность для $k + 1$:

$$\begin{aligned}\alpha_{k+1} &\leq \alpha_k \left(1 - \frac{\beta}{k+1}\right) + \frac{\gamma}{(k+1)^2} \leq \frac{D}{k+1} \left(1 - \frac{\beta}{k+1}\right) + \frac{\gamma}{(k+1)^2} \\ &= \frac{D(k+1-\beta)(k+2) + \gamma(k+2)}{(k+1)^2(k+2)} \\ &= \underbrace{\frac{-D + (k+2)(D(1-\beta) + \gamma)}{(k+1)^2(k+2)}}_{<0} + \frac{D}{k+2} \leq \frac{D}{k+2}. \blacksquare\end{aligned}$$

Сходимость градиентного спуска

Док-во. Докажем по индукции: очевидным образом база верна.

Предположим, что утверждение верно для k , выведем верность для $k + 1$:

$$\begin{aligned}\alpha_{k+1} &\leq \alpha_k \left(1 - \frac{\beta}{k+1}\right) + \frac{\gamma}{(k+1)^2} \leq \frac{D}{k+1} \left(1 - \frac{\beta}{k+1}\right) + \frac{\gamma}{(k+1)^2} \\ &= \frac{D(k+1-\beta)(k+2) + \gamma(k+2)}{(k+1)^2(k+2)} \\ &= \underbrace{\frac{-D + (k+2)(D(1-\beta) + \gamma)}{(k+1)^2(k+2)}}_{<0} + \frac{D}{k+2} \leq \frac{D}{k+2}. \blacksquare\end{aligned}$$

Возвращаясь к градиентному спуску, если функция f сильно выпукла с константой m , а градиент липшицев с константой M , то

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^*\|^2 - 2\alpha_k \nabla f(x_k)^T (x_k - x^*) + \alpha_k^2 \|\nabla f(x_k)\|^2 \\ &\leq (1 - 2m\alpha_k) \|x_k - x^*\|^2 + \alpha_k^2 M^2.\end{aligned}$$

Сходимость градиентного спуска

Док-во. Докажем по индукции: очевидным образом база верна.

Предположим, что утверждение верно для k , выведем верность для $k + 1$:

$$\begin{aligned}\alpha_{k+1} &\leq \alpha_k \left(1 - \frac{\beta}{k+1}\right) + \frac{\gamma}{(k+1)^2} \leq \frac{D}{k+1} \left(1 - \frac{\beta}{k+1}\right) + \frac{\gamma}{(k+1)^2} \\ &= \frac{D(k+1-\beta)(k+2) + \gamma(k+2)}{(k+1)^2(k+2)} \\ &= \underbrace{\frac{-D + (k+2)(D(1-\beta) + \gamma)}{(k+1)^2(k+2)}}_{<0} + \frac{D}{k+2} \leq \frac{D}{k+2}. \blacksquare\end{aligned}$$

Возвращаясь к градиентному спуску, если функция f сильно выпукла с константой m , а градиент липшицев с константой M , то

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^*\|^2 - 2\alpha_k \nabla f(x_k)^T (x_k - x^*) + \alpha_k^2 \|\nabla f(x_k)\|^2 \\ &\leq (1 - 2m\alpha_k) \|x_k - x^*\|^2 + \alpha_k^2 M^2.\end{aligned}$$

Из Доказанной леммы следует, что $\|x_k - x^*\| = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ при

$$\alpha_k = \frac{\alpha}{k+1} > \frac{1}{2m(k+1)}.$$