

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:

- Заранее выбранная последовательность, например $\alpha_k \equiv \alpha > 0$ или $\alpha_k = \frac{\alpha}{n^c}$.
- Точный минимум по направлению:

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

- Аппроксимированный минимум по направлению, α_k вычисляется следующим образом: пусть $\gamma \in (0, 1/2)$, $\beta \in (0, 1)$ – некоторые константы

Function Backtracking line search(f, x_k, γ, β)

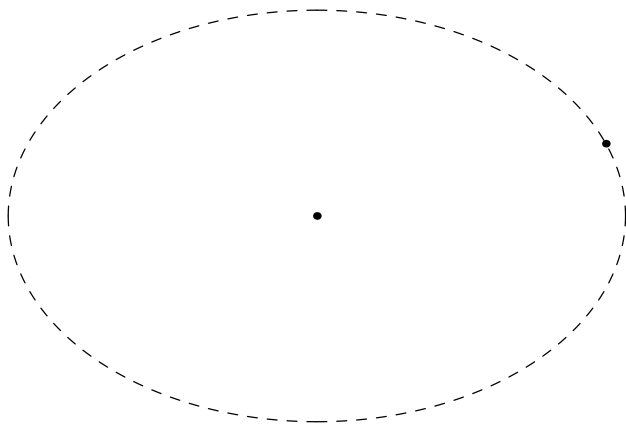
$\alpha_k \leftarrow 1;$

while $f(x_k - \alpha_k \nabla f(x_k)) > f(x_k) - \gamma \alpha_k \|\nabla f(x_k)\|^2$ **do**

$\alpha_k \leftarrow \beta \alpha_k;$

return $\alpha_k;$

Пример: градиентный спуск для квадратичной функции



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

$$x_0 = 4.000000$$

$$y_0 = 1.000000$$

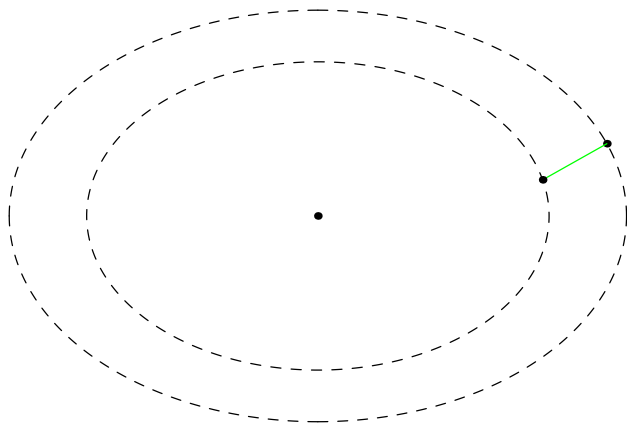
$$\nabla f(\cdot) = (0.888889,$$

 $0.500000)$

$$\alpha_0 = 1$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_1 = 3.111111$$

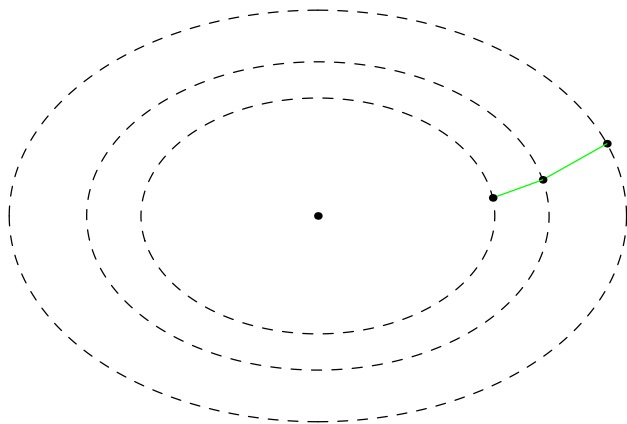
$$y_1 = 0.500000$$

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

$$\alpha_1 = 1$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_2 = 2.419753$$

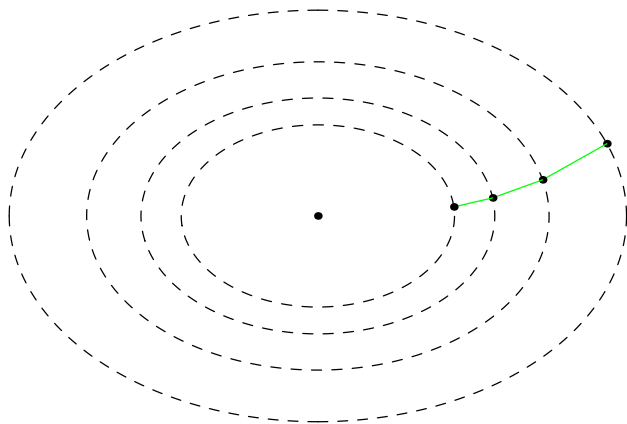
$$y_2 = 0.250000$$

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

$$\alpha_2 = 1$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_3 = 1.882030$$

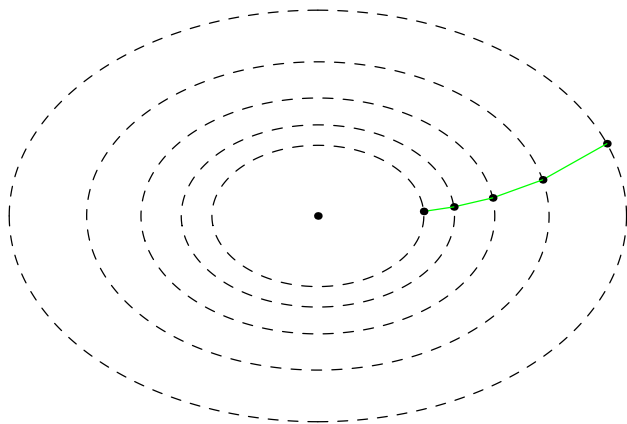
$$y_3 = 0.125000$$

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

$$\alpha_3 = 1$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_4 = 1.463801$$

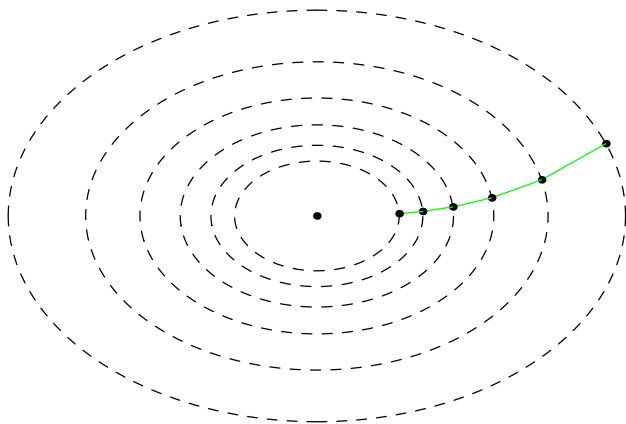
$$y_4 = 0.062500$$

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

$$\alpha_4 = 1$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_5 = 1.138512$$

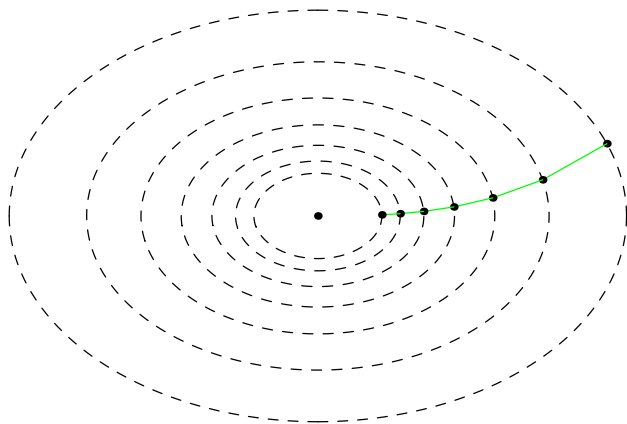
$$y_5 = 0.031250$$

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

$$\alpha_5 = 1$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_6 = 0.885509$$

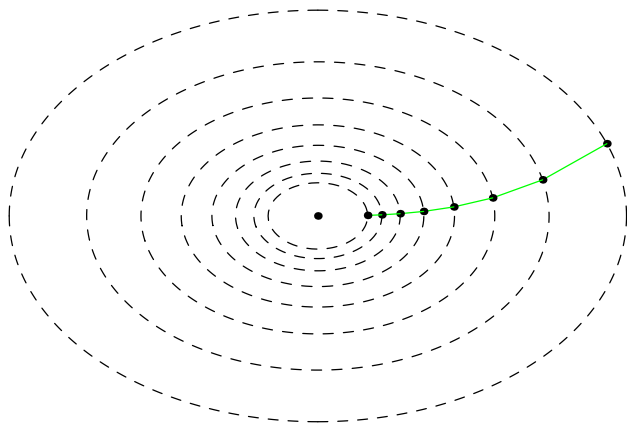
$$y_6 = 0.015625$$

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

$$\alpha_6 = 1$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_7 = 0.688730$$

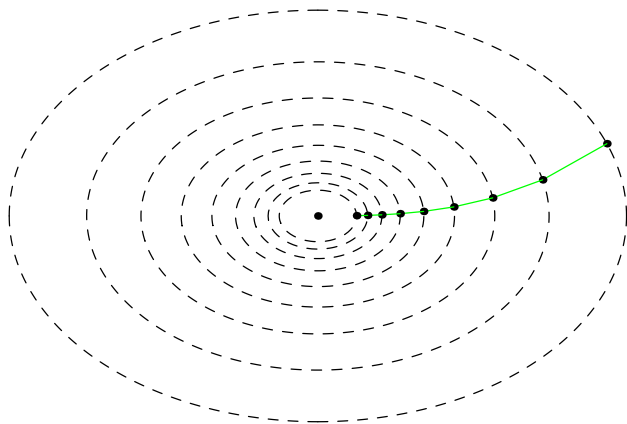
$$y_7 = 0.007813$$

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

$$\alpha_7 = 1$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_8 = 0.535679$$

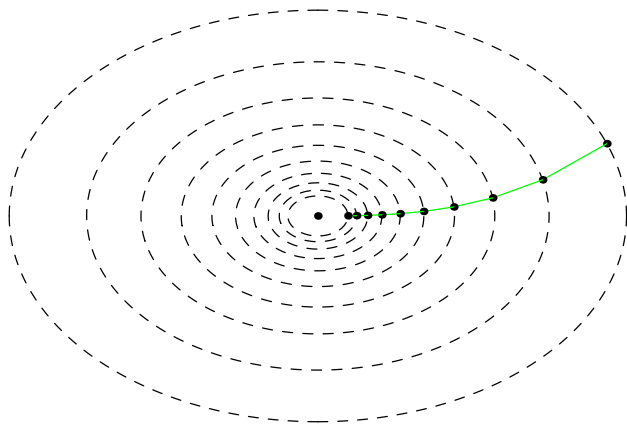
$$y_8 = 0.003906$$

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

$$\alpha_8 = 1$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_9 = 0.416639$$

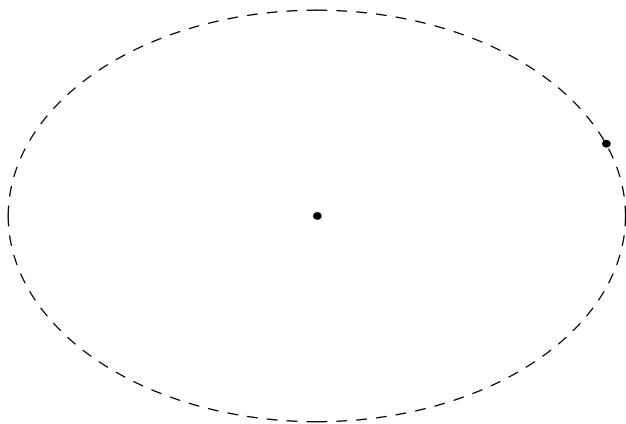
$$y_9 = 0.001953$$

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

$$\alpha_9 = 1$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_0 = 4.000000$$

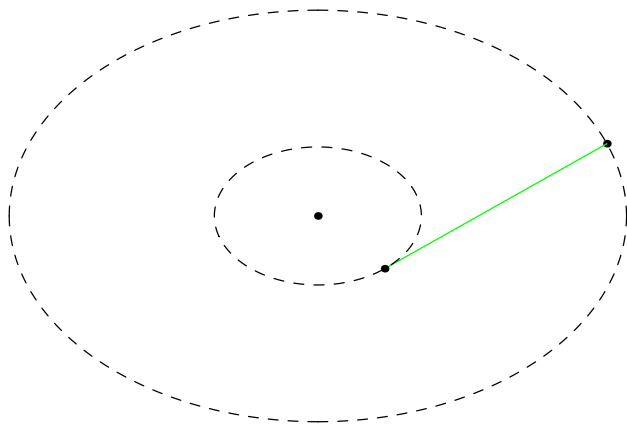
$$y_0 = 1.000000$$

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

$$\alpha_0 = 3.460354$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_1 = 0.924130$$

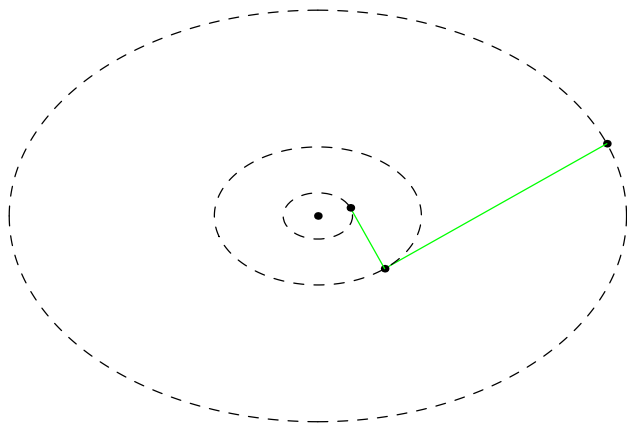
$$y_1 = -0.730177$$

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

$$\alpha_1 = 2.308219$$

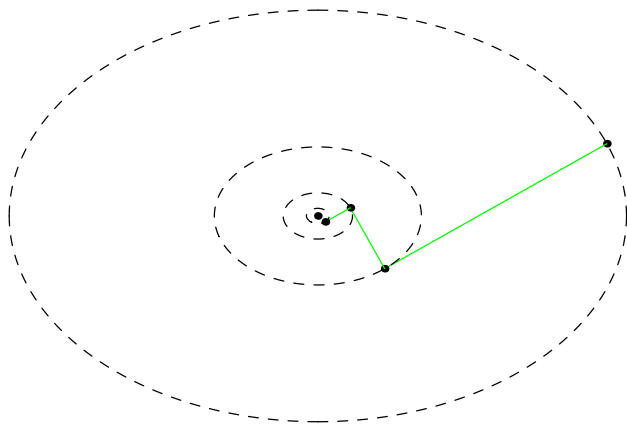
$$\sqrt{x_1^2 + y_1^2} = 1.177784$$

Пример: градиентный спуск для квадратичной функции



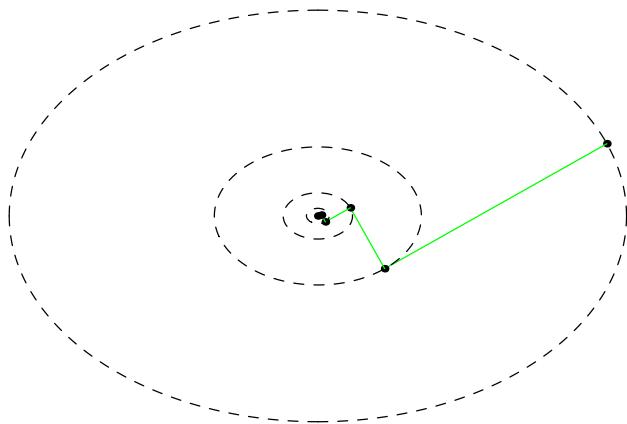
$$\begin{aligned}f(x, y) &= \frac{x^2}{9} + \frac{y^2}{4} \\x_2 &= 0.450109 \\y_2 &= 0.112527 \\\nabla f(\cdot) &= (0.100024, \\&\quad 0.056264) \\\alpha_2 &= 3.460354 \\\sqrt{x_2^2 + y_2^2} &= 0.463962\end{aligned}$$

Пример: градиентный спуск для квадратичной функции



$$\begin{aligned}f(x, y) &= \frac{x^2}{9} + \frac{y^2}{4} \\x_3 &= 0.103990 \\y_3 &= -0.082165 \\\nabla f(\cdot) &= (0.023109, \\&\quad -0.041082) \\\alpha_3 &= 2.308219 \\\sqrt{x_3^2 + y_3^2} &= 0.132533\end{aligned}$$

Пример: градиентный спуск для квадратичной функции



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

$$x_4 = 0.050650$$

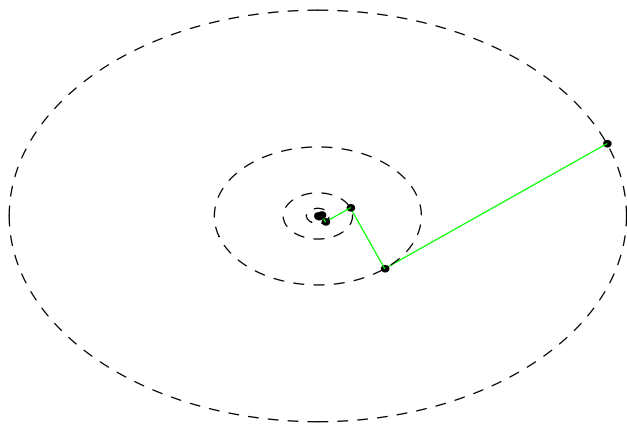
$$y_4 = 0.012662$$

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

$$\alpha_4 = 3.460354$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_5 = 0.011702$$

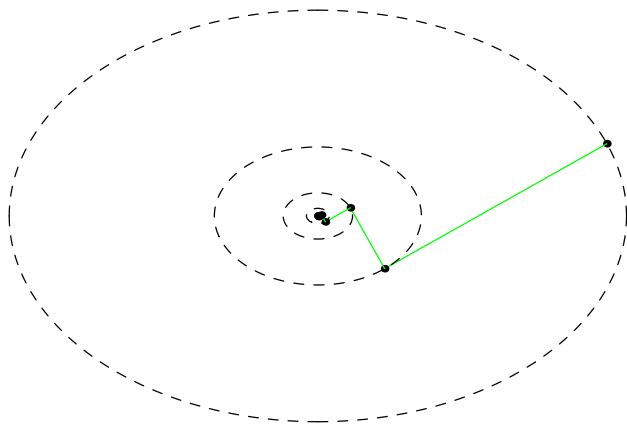
$$y_5 = -0.009246$$

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

$$\alpha_5 = 2.308219$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_6 = 0.005699$$

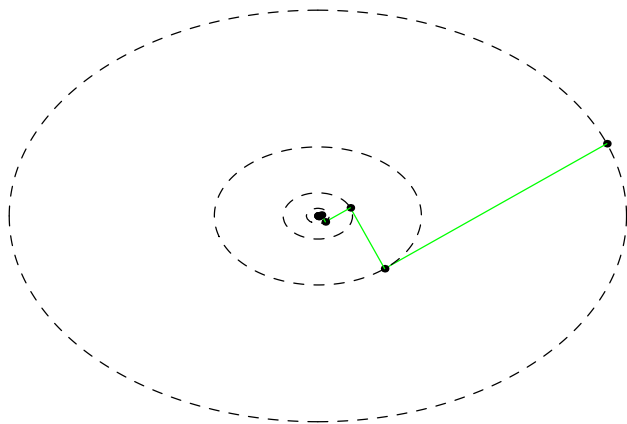
$$y_6 = 0.001425$$

$$\nabla f(\cdot) = (0.001267, \\ 0.000712)$$

$$\alpha_6 = 3.460354$$

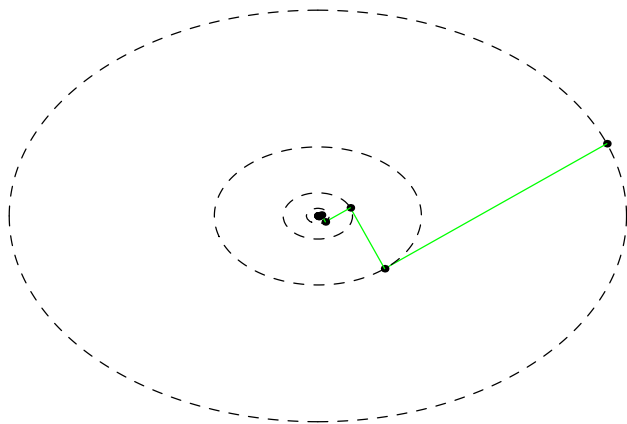
$$\sqrt{x_6^2 + y_6^2} = 0.005875$$

Пример: градиентный спуск для квадратичной функции



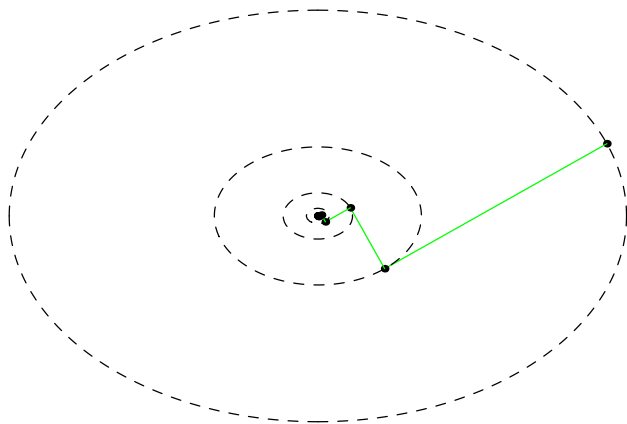
$$\begin{aligned}f(x, y) &= \frac{x^2}{9} + \frac{y^2}{4} \\x_7 &= 0.001317 \\y_7 &= -0.001040 \\\nabla f(\cdot) &= (0.000293, \\&\quad -0.000520) \\\alpha_7 &= 2.308219 \\\sqrt{x_7^2 + y_7^2} &= 0.001678\end{aligned}$$

Пример: градиентный спуск для квадратичной функции



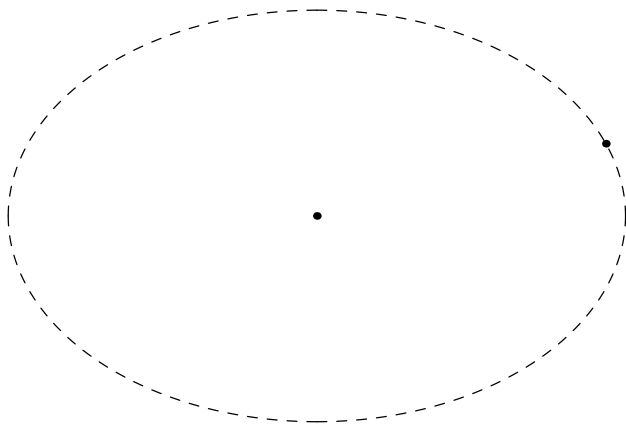
$$\begin{aligned}f(x, y) &= \frac{x^2}{9} + \frac{y^2}{4} \\x_8 &= 0.000641 \\y_8 &= 0.000160 \\\nabla f(\cdot) &= (0.000143, \\&\quad 0.000080) \\\alpha_8 &= 3.460354 \\\sqrt{x_8^2 + y_8^2} &= 0.000661\end{aligned}$$

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



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

$$x_0 = 4.000000$$

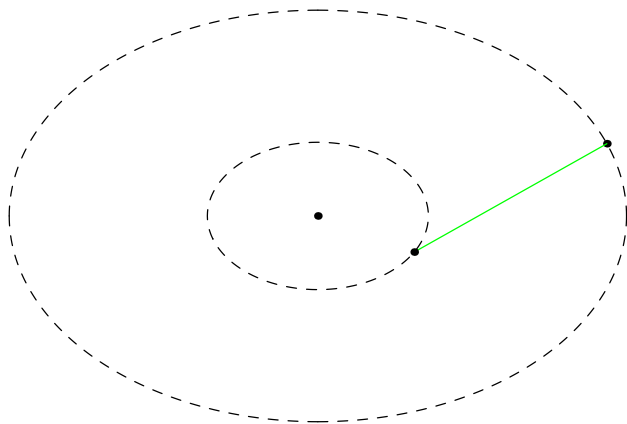
$$y_0 = 1.000000$$

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

$$\alpha_0 = 3/1$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_1 = 1.333333$$

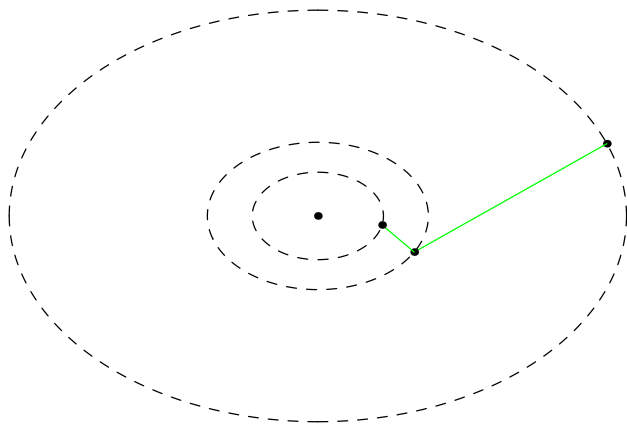
$$y_1 = -0.500000$$

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

$$\alpha_1 = 3/2$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_2 = 0.888889$$

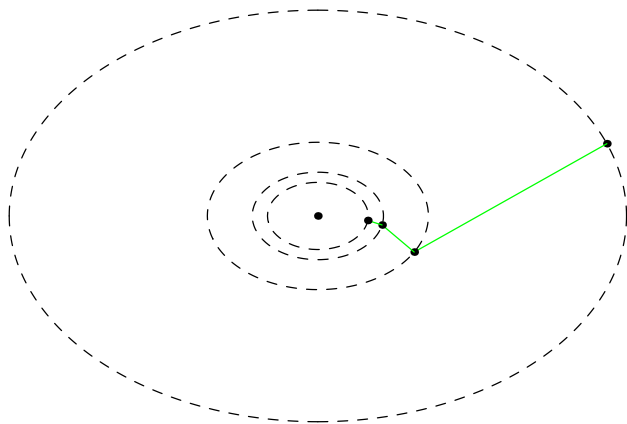
$$y_2 = -0.125000$$

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

$$\alpha_2 = 3/3$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_3 = 0.691358$$

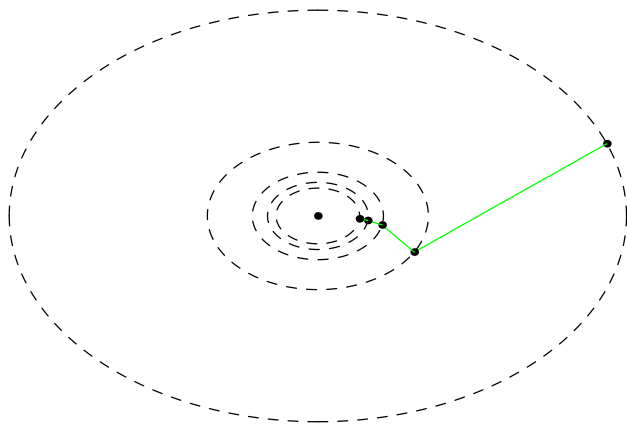
$$y_3 = -0.062500$$

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

$$\alpha_3 = 3/4$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_4 = 0.576132$$

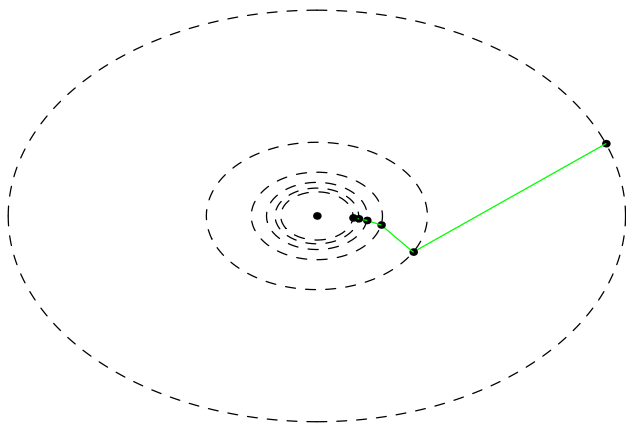
$$y_4 = -0.039063$$

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

$$\alpha_4 = 3/5$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_5 = 0.499314$$

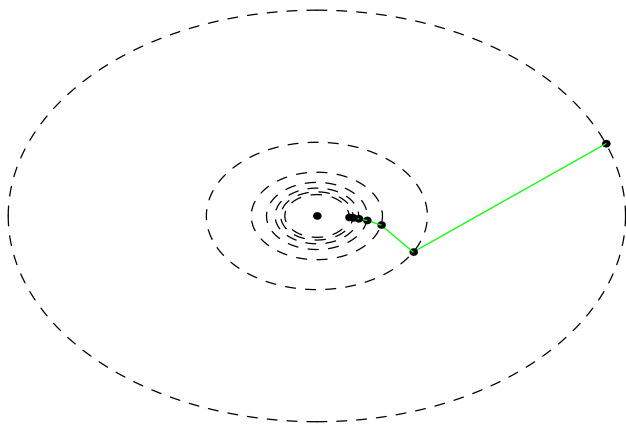
$$y_5 = -0.027344$$

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

$$\alpha_5 = 3/6$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_6 = 0.443835$$

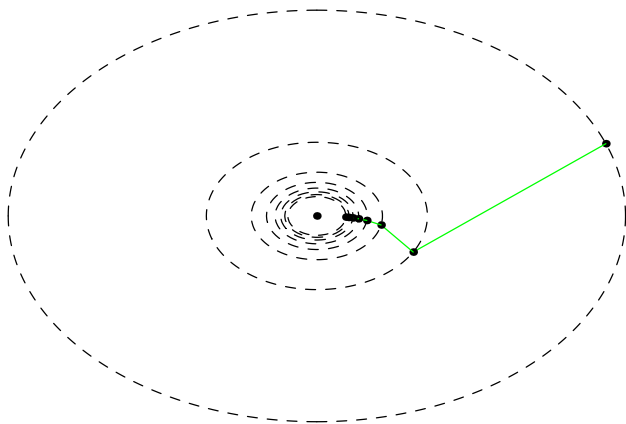
$$y_6 = -0.020508$$

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

$$\alpha_6 = 3/7$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_7 = 0.401565$$

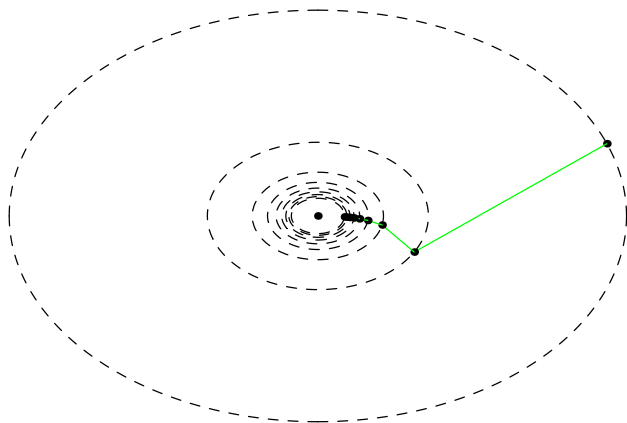
$$y_7 = -0.016113$$

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

$$\alpha_7 = 3/8$$

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

Пример: градиентный спуск для квадратичной функции



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

$$x_8 = 0.368101$$

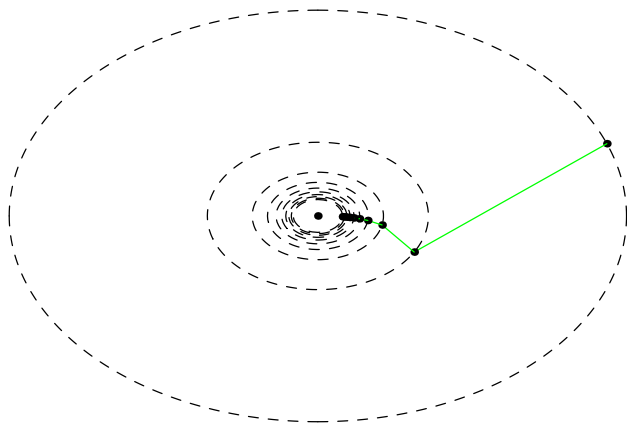
$$y_8 = -0.013092$$

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

$$\alpha_8 = 3/9$$

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

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.

Теорема (Постоянный шаг)

Пусть f выпукла и дифференцируема на \mathcal{D} , градиент f липшицев с константой $M > 0$ на $S_f(x_0)$, f ограничена снизу, существует хотя бы одна точка минимума x^* , $\alpha_k = \alpha \in [0, 1/M]$, 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

$$f(x) - f(x^*) \geq \frac{1}{2M} \|\nabla f(x)\|^2 \quad (3)$$

Сходимость градиентного спуска (постоянный шаг)

Вернемся на шаг назад, при условии $\alpha \leq 1/M$ выполняется $-\alpha + M\alpha^2/2 \leq -\alpha/2$, получаем

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

Сходимость градиентного спуска (постоянный шаг)

Вернемся на шаг назад, при условии $\alpha \leq 1/M$ выполняется $-\alpha + M\alpha^2/2 \leq -\alpha/2$, получаем

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

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

Так как $f(x_k)$ убывает, то

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

Замечание 1. Если использовать минимум по направлению, то величина $f(x_k) - f(x_{k+1})$ увеличиваются, а значит все оценки сохраняются.

Использование *backtracking line search*

Замечание 2. Если использовать аппроксимированный минимум по направлению с параметрами $\gamma \in (0, 1/2)$, $\beta \in (0, 1)$, то учитывая $-\alpha + M\alpha^2/2 \leq -\alpha/2$ при $0 \leq \alpha \leq 1/M$

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

получаем, что условие выхода в *backtracking line search* выполняется для любого $\alpha \in [0, 1/M]$. Так как на каждом шаге α увеличивается в β раз, то *backtracking line search* выдаёт либо 1, либо величину $\alpha_k \geq \beta/M$, что даёт

$$\begin{aligned} f(x_{i+1}) &\leq f(x_i) - \frac{\alpha_k}{2} \|\nabla f(x_i)\|^2 \\ &\leq f(x^*) + \nabla f(x_i)^T (x_i - x^*) - \frac{\alpha_k}{2} \|\nabla f(x_i)\|^2 \\ &= f(x^*) + \frac{1}{2\alpha_k} (\|x_i - x^*\|^2 - \|x_i - x^* - \alpha_k \nabla f(x_i)\|^2) \\ &\leq f(x^*) + \frac{1}{2 \min\{1, \beta/M\}} (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2). \end{aligned}$$

Использование *backtracking line search*

Суммируя по итерациям выводим схожий результат, отличающийся на константу $\alpha \rightarrow \min\{1, \beta/M\}$:

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

Сходимость градиентного спуска ($f(x_k) \rightarrow f(x^*)$)

Теорема (Постоянный шаг, сильная выпуклость)

Пусть f дифференцируема 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), x_k сходится к единственной точке минимума x^* f на \mathcal{D} , $f(x_k)$ убывает и сходится к $f(x^*)$, более того для $q = 1 - 2m\alpha + mM\alpha^2$

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

Сходимость градиентного спуска ($f(x_k) \rightarrow f(x^*)$)

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

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

Сходимость градиентного спуска ($f(x_k) \rightarrow f(x^*)$)

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

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

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

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

Сходимость градиентного спуска ($f(x_k) \rightarrow f(x^*)$)

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 ($f(x_k) \rightarrow f(x^*)$)

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 (5) $x_k \rightarrow x^*$ и, следовательно $f(x_k) \rightarrow f(x^*)$.

Сходимость градиентного спуска ($f(x_k) \rightarrow f(x^*)$)

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 (5) $x_k \rightarrow x^*$ 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 (4) 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 ($f(x_k) \rightarrow f(x^*)$)

Таким образом

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

Сходимость градиентного спуска ($f(x_k) \rightarrow f(x^*)$)

Таким образом

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

Замечание 1. Используя (3) и (5) можно получить

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

Сходимость градиентного спуска ($f(x_k) \rightarrow f(x^*)$)

Таким образом

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

Замечание 1. Используя (3) и (5) можно получить

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

Замечание 2. При использовании *backtracking line search* с параметрами γ, β все выкладки сохраняются при $q = 1 - \min\{2m\gamma, 2\beta\gamma m/M\}$.

Сходимость градиентного спуска ($x_k \rightarrow x^*$)

Лемма (О m -сильно выпуклой M -гладкой функции)

Пусть $f : \mathcal{D} \rightarrow \mathbb{R}$ – сильно выпуклая с параметром m функция, ∇f удовлетворяет условию Липшица с параметром M , т. е.

$$m\|y - x\|^2 \leq (\nabla f(y) - \nabla f(x))^T (y - x) \leq M\|y - x\|^2,$$

тогда для f выполняется

$$(\nabla f(y) - \nabla f(x))^T (y - x) \geq \frac{mM}{m + M} \|y - x\|^2 + \frac{1}{m + M} \|\nabla f(y) - \nabla f(x)\|^2$$

Сходимость градиентного спуска ($x_k \rightarrow x^*$)

Лемма (О m -сильно выпуклой M -гладкой функции)

Пусть $f : \mathcal{D} \rightarrow \mathbb{R}$ – сильно выпуклая с параметром m функция, ∇f удовлетворяет условию Липшица с параметром M , т. е.

$$m\|y - x\|^2 \leq (\nabla f(y) - \nabla f(x))^T (y - x) \leq M\|y - x\|^2,$$

тогда для f выполняется

$$(\nabla f(y) - \nabla f(x))^T (y - x) \geq \frac{mM}{m + M} \|y - x\|^2 + \frac{1}{m + M} \|\nabla f(y) - \nabla f(x)\|^2$$

Док-во. Рассмотрим функцию $g(x) = f(x) - \frac{m}{2}\|x\|^2$. Заметим, что $\nabla g(x) = \nabla f(x) - mx$ и

$$(\nabla g(y) - \nabla g(x))^T (y - x) = (\nabla f(y) - \nabla f(x))^T (y - x) - m\|y - x\|^2,$$

то есть g – выпуклая функция, ∇g удовлетворяет условию Липшица с константой $M - m$.

Сходимость градиентного спуска ($x_k \rightarrow x^*$)

Далее, пусть для некоторого x $\phi(y) = g(y) - \nabla g(x)^T y$. Заметим, что $\nabla \phi(y) = \nabla g(y) - \nabla g(x)$, таким образом ϕ тоже выпукла и $\nabla \phi$ удовлетворяет условию Липшица с константов $M - m$.

Сходимость градиентного спуска ($x_k \rightarrow x^*$)

Далее, пусть для некоторого x $\phi(y) = g(y) - \nabla g(x)^T y$. Заметим, что $\nabla \phi(y) = \nabla g(y) - \nabla g(x)$, таким образом ϕ тоже выпукла и $\nabla \phi$ удовлетворяет условию Липшица с константов $M - m$.

Точка x минимизирует ϕ в силу выпуклости ϕ и $\nabla \phi(x) = 0$, используя (4)

$$\phi(x) \leq \phi(y) - \frac{1}{2(M - m)} \|\nabla \phi(y)\|^2$$

что имеет следующий вид в терминах g

$$g(y) \geq g(x) + \nabla g(x)^T (y - x) + \frac{1}{2(M - m)} \|\nabla g(y) - \nabla g(x)\|^2$$

Сходимость градиентного спуска ($x_k \rightarrow x^*$)

Далее, пусть для некоторого x $\phi(y) = g(y) - \nabla g(x)^T y$. Заметим, что $\nabla \phi(y) = \nabla g(y) - \nabla g(x)$, таким образом ϕ тоже выпукла и $\nabla \phi$ удовлетворяет условию Липшица с константов $M - m$.

Точка x минимизирует ϕ в силу выпуклости ϕ и $\nabla \phi(x) = 0$, используя (4)

$$\phi(x) \leq \phi(y) - \frac{1}{2(M-m)} \|\nabla \phi(y)\|^2$$

что имеет следующий вид в терминах g

$$g(y) \geq g(x) + \nabla g(x)^T (y - x) + \frac{1}{2(M-m)} \|\nabla g(y) - \nabla g(x)\|^2$$

складывая это неравенство с самим собой с переставленными $x \leftrightarrow y$ получаем

$$(\nabla g(y) - \nabla g(x))^T (y - x) \geq \frac{1}{M-m} \|\nabla g(y) - \nabla g(x)\|^2$$

Сходимость градиентного спуска ($x_k \rightarrow x^*$)

Наконец, выражая g через f получаем

$$\begin{aligned}(\nabla g(y) - \nabla g(x))^T(y - x) &= (\nabla f(y) - \nabla f(x))^T(y - x) - m\|y - x\|^2 \\ \|\nabla g(y) - \nabla g(x)\|^2 &= \|\nabla f(y) - \nabla f(x)\|^2 - 2m(\nabla f(y) - \nabla f(x))^T(y - x) \\ &\quad + m^2\|y - x\|^2,\end{aligned}$$

Сходимость градиентного спуска ($x_k \rightarrow x^*$)

Наконец, выражая g через f получаем

$$\begin{aligned}(\nabla g(y) - \nabla g(x))^T(y - x) &= (\nabla f(y) - \nabla f(x))^T(y - x) - m\|y - x\|^2 \\ \|\nabla g(y) - \nabla g(x)\|^2 &= \|\nabla f(y) - \nabla f(x)\|^2 - 2m(\nabla f(y) - \nabla f(x))^T(y - x) \\ &\quad + m^2\|y - x\|^2,\end{aligned}$$

что дает

$$\begin{aligned}(\nabla f(y) - \nabla f(x))^T(y - x) &\geq m\|y - x\|^2 + \frac{1}{M - m}(\|\nabla f(y) - \nabla f(x)\|^2 \\ &\quad - 2m(\nabla f(y) - \nabla f(x))^T(y - x) + m^2\|y - x\|^2)\end{aligned}$$

Сходимость градиентного спуска ($x_k \rightarrow x^*$)

Наконец, выражая g через f получаем

$$\begin{aligned}(\nabla g(y) - \nabla g(x))^T(y - x) &= (\nabla f(y) - \nabla f(x))^T(y - x) - m\|y - x\|^2 \\ \|\nabla g(y) - \nabla g(x)\|^2 &= \|\nabla f(y) - \nabla f(x)\|^2 - 2m(\nabla f(y) - \nabla f(x))^T(y - x) \\ &\quad + m^2\|y - x\|^2,\end{aligned}$$

что дает

$$\begin{aligned}(\nabla f(y) - \nabla f(x))^T(y - x) &\geq m\|y - x\|^2 + \frac{1}{M - m}(\|\nabla f(y) - \nabla f(x)\|^2 \\ &\quad - 2m(\nabla f(y) - \nabla f(x))^T(y - x) + m^2\|y - x\|^2)\end{aligned}$$

$$(\nabla f(y) - \nabla f(x))^T(y - x) \geq \frac{1}{m + M}(Mm\|y - x\|^2 + \|\nabla f(y) - \nabla f(x)\|^2) \quad \blacksquare$$

Сходимость градиентного спуска ($x_k \rightarrow x^*$)

Теорема

Пусть f дифференцируема на \mathcal{D} , $\alpha_k \equiv \alpha \in (0, 2/(M + m))$, 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), x_k сходится к единственной точке минимума x^* f на \mathcal{D} , $f(x_k)$ убывает и сходится к $f(x^*)$, более того для

$$\|x_k - x^*\|^2 \leq \left(1 - \frac{2\alpha m M}{M + m}\right)^k \|x_0 - x^*\|^2$$

Сходимость градиентного спуска ($x_k \rightarrow x^*$)

Теорема

Пусть f дифференцируема на \mathcal{D} , $\alpha_k \equiv \alpha \in (0, 2/(M + m))$, 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), x_k сходится к единственной точке минимума x^* f на \mathcal{D} , $f(x_k)$ убывает и сходится к $f(x^*)$, более того для

$$\|x_k - x^*\|^2 \leq \left(1 - \frac{2\alpha m M}{M + m}\right)^k \|x_0 - x^*\|^2$$

Док-во. Используя доказанную лемму

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

Оптимальность $\alpha = 2/(m + M)$

Замечание. При $\alpha = 2/(m + M)$ параметр сходимости становится

$$1 - \frac{2\alpha m M}{M + m} = 1 - \frac{4\alpha m M}{(M + m)^2} = \left(\frac{M - m}{M + m} \right)^2$$

Оптимальность $\alpha = 2/(m + M)$

Замечание. При $\alpha = 2/(m + M)$ параметр сходимости становится

$$1 - \frac{2\alpha mM}{M + m} = 1 - \frac{4\alpha mM}{(M + m)^2} = \left(\frac{M - m}{M + m}\right)^2$$

Такой выбор α оптимален при условии, что m, M – точные оценки: пусть $f(x) = \frac{1}{2}x^T Ax - b^T x$, $A = A^T$, тогда последовательность (2) принимает вид

$$x_{k+1} = (I - \alpha_k A)x_k - b.$$

Оптимальность $\alpha = 2/(m + M)$

Замечание. При $\alpha = 2/(m + M)$ параметр сходимости становится

$$1 - \frac{2\alpha mM}{M + m} = 1 - \frac{4\alpha mM}{(M + m)^2} = \left(\frac{M - m}{M + m}\right)^2$$

Такой выбор α оптимален при условии, что m, M – точные оценки: пусть $f(x) = \frac{1}{2}x^T Ax - b^T x$, $A = A^T$, тогда последовательность (2) принимает вид

$$x_{k+1} = (I - \alpha_k A)x_k - b.$$

Если $Ax^* = b$, $m, M > 0$ – минимальное и максимальное собственные числа A , то

$$\|x_{k+1} - x^*\| = \|(I - \alpha_k A)(x_k - x^*)\| \leq \max\{|1 - \alpha M|, |1 - \alpha m|\} \|x_k - x^*\|,$$

при этом

$$\min_{\alpha} \max\{|1 - \alpha M|, |1 - \alpha m|\} = \frac{M - m}{M + m},$$

минимум достигается при $\alpha = 2/(m + M)$.

Ссылки на литературу

Нестеров Ю. Е. Методы выпуклой оптимизации // параграфы 1.2.3 и 2.1.5

Boyd S., Vandenberghe L. Convex optimization // параграф 9.3

Поляк Б. Т. Введение в оптимизацию // параграф 1.4

Vandenberghe L. Лекция по градиентному спуску