



Градиентный спуск. Теоремы сходимости в  
квадратичном случае. Теоремы сходимости  
в гладком случае (выпуклые, сильно  
выпуклые, PL)

Даня Меркулов

Оптимизация для всех! ЦУ

## Сильно выпуклые квадратичные функции

## Сдвиг координат

Рассмотрим следующую задачу квадратичной оптимизации:

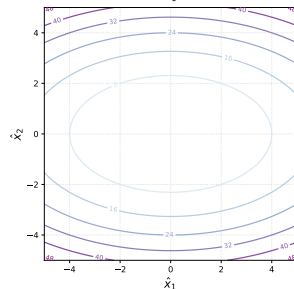
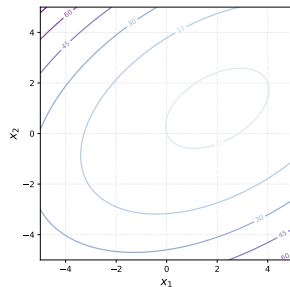
$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

## Сдвиг координат

Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

- Во-первых, без ограничения общности мы можем установить  $c = 0$ , что не повлияет на процесс оптимизации.

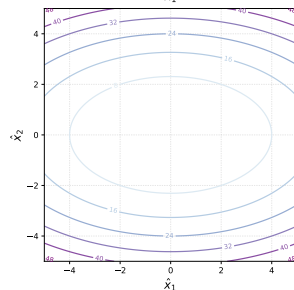
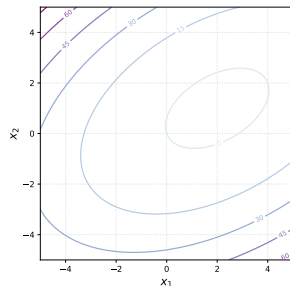


## Сдвиг координат

Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

- Во-первых, без ограничения общности мы можем установить  $c = 0$ , что не повлияет на процесс оптимизации.
- Во-вторых, у нас есть спектральное разложение матрицы  $A = Q\Lambda Q^T$ .



## Сдвиг координат

Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

- Во-первых, без ограничения общности мы можем установить  $c = 0$ , что не повлияет на процесс оптимизации.
- Во-вторых, у нас есть спектральное разложение матрицы  $A = Q\Lambda Q^T$ .
- Покажем, что мы можем сделать сдвиг координат, чтобы сделать анализ немного проще. Пусть  $\hat{x} = Q^T(x - x^*)$ , где  $x^*$  — точка минимума исходной функции, определяемая как  $Ax^* = b$ . При этом  $x = Q\hat{x} + x^*$ .



## Сдвиг координат

Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

- Во-первых, без ограничения общности мы можем установить  $c = 0$ , что не повлияет на процесс оптимизации.
- Во-вторых, у нас есть спектральное разложение матрицы  $A = Q\Lambda Q^T$ .
- Покажем, что мы можем сделать сдвиг координат, чтобы сделать анализ немного проще. Пусть  $\hat{x} = Q^T(x - x^*)$ , где  $x^*$  — точка минимума исходной функции, определяемая как  $Ax^* = b$ . При этом  $x = Q\hat{x} + x^*$ .

$$f(\hat{x}) = \frac{1}{2} (Q\hat{x} + x^*)^\top A (Q\hat{x} + x^*) - b^\top (Q\hat{x} + x^*)$$



## Сдвиг координат

Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

- Во-первых, без ограничения общности мы можем установить  $c = 0$ , что не повлияет на процесс оптимизации.
- Во-вторых, у нас есть спектральное разложение матрицы  $A = Q\Lambda Q^T$ .
- Покажем, что мы можем сделать сдвиг координат, чтобы сделать анализ немного проще. Пусть  $\hat{x} = Q^T(x - x^*)$ , где  $x^*$  — точка минимума исходной функции, определяемая как  $Ax^* = b$ . При этом  $x = Q\hat{x} + x^*$ .

$$\begin{aligned} f(\hat{x}) &= \frac{1}{2} (Q\hat{x} + x^*)^\top A (Q\hat{x} + x^*) - b^\top (Q\hat{x} + x^*) \\ &= \frac{1}{2} \hat{x}^\top Q^\top A Q \hat{x} + \frac{1}{2} (x^*)^\top A (x^*) + (x^*)^\top A Q \hat{x} - b^\top Q \hat{x} - b^\top x^* \end{aligned}$$





## Сдвиг координат

Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

- Во-первых, без ограничения общности мы можем установить  $c = 0$ , что не повлияет на процесс оптимизации.
- Во-вторых, у нас есть спектральное разложение матрицы  $A = Q\Lambda Q^T$ .
- Покажем, что мы можем сделать сдвиг координат, чтобы сделать анализ немного проще. Пусть  $\hat{x} = Q^T(x - x^*)$ , где  $x^*$  — точка минимума исходной функции, определяемая как  $Ax^* = b$ . При этом  $x = Q\hat{x} + x^*$ .

$$\begin{aligned} f(\hat{x}) &= \frac{1}{2} (Q\hat{x} + x^*)^\top A (Q\hat{x} + x^*) - b^\top (Q\hat{x} + x^*) \\ &= \frac{1}{2} \hat{x}^\top Q^\top A Q \hat{x} + \frac{1}{2} (x^*)^\top A (x^*) + (x^*)^\top A Q \hat{x} - b^\top Q \hat{x} - b^\top x^* \\ &= \frac{1}{2} \hat{x}^\top \Lambda \hat{x} + \frac{1}{2} (x^*)^\top A (x^*) + (x^*)^\top A Q \hat{x} - (x^*)^\top A^\top Q \hat{x} - (x^*)^\top A x^* \end{aligned}$$



## Сдвиг координат

Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

- Во-первых, без ограничения общности мы можем установить  $c = 0$ , что не повлияет на процесс оптимизации.
- Во-вторых, у нас есть спектральное разложение матрицы  $A = Q\Lambda Q^T$ .
- Покажем, что мы можем сделать сдвиг координат, чтобы сделать анализ немного проще. Пусть  $\hat{x} = Q^T(x - x^*)$ , где  $x^*$  — точка минимума исходной функции, определяемая как  $Ax^* = b$ . При этом  $x = Q\hat{x} + x^*$ .

$$\begin{aligned} f(\hat{x}) &= \frac{1}{2} (Q\hat{x} + x^*)^\top A (Q\hat{x} + x^*) - b^\top (Q\hat{x} + x^*) \\ &= \frac{1}{2} \hat{x}^\top Q^\top A Q \hat{x} + \frac{1}{2} (x^*)^\top A (x^*) + (x^*)^\top A Q \hat{x} - b^\top Q \hat{x} - b^\top x^* \\ &= \frac{1}{2} \hat{x}^\top \Lambda \hat{x} + \frac{1}{2} (x^*)^\top A (x^*) + (x^*)^\top A Q \hat{x} - (x^*)^\top A^\top Q \hat{x} - (x^*)^\top A x^* \\ &= \frac{1}{2} \hat{x}^\top \Lambda \hat{x} - \frac{1}{2} (x^*)^\top A x^* \end{aligned}$$



## Сдвиг координат

Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

- Во-первых, без ограничения общности мы можем установить  $c = 0$ , что не повлияет на процесс оптимизации.
- Во-вторых, у нас есть спектральное разложение матрицы  $A = Q\Lambda Q^T$ .
- Покажем, что мы можем сделать сдвиг координат, чтобы сделать анализ немного проще. Пусть  $\hat{x} = Q^T(x - x^*)$ , где  $x^*$  — точка минимума исходной функции, определяемая как  $Ax^* = b$ . При этом  $x = Q\hat{x} + x^*$ .

$$\begin{aligned} f(\hat{x}) &= \frac{1}{2} (Q\hat{x} + x^*)^\top A (Q\hat{x} + x^*) - b^\top (Q\hat{x} + x^*) \\ &= \frac{1}{2} \hat{x}^\top Q^\top A Q \hat{x} + \frac{1}{2} (x^*)^\top A (x^*) + (x^*)^\top A Q \hat{x} - b^\top Q \hat{x} - b^\top x^* \\ &= \frac{1}{2} \hat{x}^\top \Lambda \hat{x} + \frac{1}{2} (x^*)^\top A (x^*) + (x^*)^\top A Q \hat{x} - (x^*)^\top A^\top Q \hat{x} - (x^*)^\top A x^* \\ &= \frac{1}{2} \hat{x}^\top \Lambda \hat{x} - \frac{1}{2} (x^*)^\top A x^* \simeq \frac{1}{2} \hat{x}^\top \Lambda \hat{x} \end{aligned}$$



## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k)$$

## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k$$

## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\ &= (I - \alpha^k \Lambda)x^k\end{aligned}$$

## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k$$

$$= (I - \alpha^k \Lambda) x^k$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\ &= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$



## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

$$|1 - \alpha \mu| < 1$$

## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

$$|1 - \alpha \mu| < 1 \qquad |1 - \alpha L| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \qquad \alpha \mu > 0$$

## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

$$|1 - \alpha\mu| < 1$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha\mu < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha\mu > 0$$



## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L} \quad \alpha L > 0$$

## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L} \quad \alpha L > 0$$



## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Выберем  $\alpha$ , минимизирующий худший знаменатель прогрессии

$$\rho^* = \min_{\alpha} \rho(\alpha)$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L} \quad \alpha L > 0$$



## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Выберем  $\alpha$ , минимизирующий худший знаменатель прогрессии

$$\rho^* = \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}|$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L} \quad \alpha L > 0$$



## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k$$

$$= (I - \alpha^k \Lambda) x^k$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Выберем  $\alpha$ , минимизирующий худший знаменатель прогрессии

$$\rho^* = \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}|$$

$$= \min_{\alpha} \max \{ |1 - \alpha \mu|, |1 - \alpha L| \}$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu}$$

$$\alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L}$$

$$\alpha L > 0$$



## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

$$|1 - \alpha\mu| < 1$$

$$-1 < 1 - \alpha\mu < 1$$

$$\alpha < \frac{2}{\mu}$$

$$\alpha\mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L}$$

$$\alpha L > 0$$

Выберем  $\alpha$ , минимизирующий худший знаменатель прогрессии

$$\rho^* = \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}|$$

$$= \min_{\alpha} \max \{|1 - \alpha\mu|, |1 - \alpha L|\}$$

$$\alpha^* : \quad 1 - \alpha^* \mu = \alpha^* L - 1$$

## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k$$

$$= (I - \alpha^k \Lambda) x^k$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu}$$

$$\alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L}$$

$$\alpha L > 0$$

Выберем  $\alpha$ , минимизирующий худший знаменатель прогрессии

$$\rho^* = \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}|$$

$$= \min_{\alpha} \max \{|1 - \alpha \mu|, |1 - \alpha L|\}$$

$$\alpha^* : 1 - \alpha^* \mu = \alpha^* L - 1$$

$$\alpha^* = \frac{2}{\mu + L}$$

## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k$$

$$= (I - \alpha^k \Lambda) x^k$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu}$$

$$\alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L}$$

$$\alpha L > 0$$

Выберем  $\alpha$ , минимизирующий худший знаменатель прогрессии

$$\rho^* = \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}|$$

$$= \min_{\alpha} \max \{|1 - \alpha \mu|, |1 - \alpha L|\}$$

$$\alpha^* : 1 - \alpha^* \mu = \alpha^* L - 1$$

$$\alpha^* = \frac{2}{\mu + L} \quad \rho^* = \frac{L - \mu}{L + \mu}$$

## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k$$

$$= (I - \alpha^k \Lambda) x^k$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu}$$

$$\alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L}$$

$$\alpha L > 0$$

Выберем  $\alpha$ , минимизирующий худший знаменатель прогрессии

$$\rho^* = \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}|$$

$$= \min_{\alpha} \max \{|1 - \alpha \mu|, |1 - \alpha L|\}$$

$$\alpha^* : 1 - \alpha^* \mu = \alpha^* L - 1$$

$$\alpha^* = \frac{2}{\mu + L} \quad \rho^* = \frac{L - \mu}{L + \mu}$$

$$|x_{(i)}^k| \leq \left( \frac{L - \mu}{L + \mu} \right)^k |x_{(i)}^0|$$

## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k$$

$$= (I - \alpha^k \Lambda) x^k$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu}$$

$$\alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L}$$

$$\alpha L > 0$$

Выберем  $\alpha$ , минимизирующий худший знаменатель прогрессии

$$\rho^* = \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}|$$

$$= \min_{\alpha} \max \{|1 - \alpha \mu|, |1 - \alpha L|\}$$

$$\alpha^* : 1 - \alpha^* \mu = \alpha^* L - 1$$

$$\alpha^* = \frac{2}{\mu + L} \quad \rho^* = \frac{L - \mu}{L + \mu}$$

$$|x_{(i)}^k| \leq \left( \frac{L - \mu}{L + \mu} \right)^k |x_{(i)}^0|$$

$$\|x^k\|_2 \leq \left( \frac{L - \mu}{L + \mu} \right)^k \|x^0\|_2$$



## Анализ сходимости

Теперь мы можем работать с функцией  $f(x) = \frac{1}{2}x^T \Lambda x$  с  $x^* = 0$  без ограничения общности (убрав крышку из  $\hat{x}$ )

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k$$

$$= (I - \alpha^k \Lambda) x^k$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг  $\alpha^k = \alpha$ . Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что  $\lambda_{\min} = \mu > 0$ ,  $\lambda_{\max} = L \geq \mu$ .

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu}$$

$$\alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L}$$

$$\alpha L > 0$$

Выберем  $\alpha$ , минимизирующий худший знаменатель прогрессии

$$\rho^* = \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}|$$

$$= \min_{\alpha} \max \{|1 - \alpha \mu|, |1 - \alpha L|\}$$

$$\alpha^* : 1 - \alpha^* \mu = \alpha^* L - 1$$

$$\alpha^* = \frac{2}{\mu + L} \quad \rho^* = \frac{L - \mu}{L + \mu}$$

$$|x_{(i)}^k| \leq \left( \frac{L - \mu}{L + \mu} \right)^k |x_{(i)}^0|$$

$$\|x^k\|_2 \leq \left( \frac{L - \mu}{L + \mu} \right)^k \|x^0\|_2 \quad f(x^k) \leq \left( \frac{L - \mu}{L + \mu} \right)^{2k} f(x^0)$$



## Анализ сходимости

Таким образом, имеем линейную сходимость по аргументу со скоростью  $\frac{\kappa-1}{\kappa+1} = 1 - \frac{2}{\kappa+1}$ , где  $\kappa = \frac{L}{\mu}$  — число обусловленности квадратичной задачи.

$\kappa$	$\rho$	Итераций до уменьшения ошибки по аргументу в 10 раз	Итераций до уменьшения ошибки по функции в 10 раз
1.1	0.05	1	1
2	0.33	3	2
5	0.67	6	3
10	0.82	12	6
50	0.96	58	29
100	0.98	116	58
500	0.996	576	288
1000	0.998	1152	576

# Число обусловленности $\kappa$

$\kappa = 1.0$



$\kappa = 100.0$



## Случай PL-функций

## PL-функции. Линейная сходимость градиентного спуска без выпуклости

Говорят, что  $f$  удовлетворяет условию Поляка-Лоясиевича (PL), если для некоторого  $\mu > 0$  выполняется

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x$$

Интересно, что градиентный спуск может сходиться линейно даже без выпуклости.

Следующие функции удовлетворяют условию PL, но не являются выпуклыми. 📄Код

$$f(x) = x^2 + 3\sin^2(x)$$



## PL-функции. Линейная сходимость градиентного спуска без выпуклости

Говорят, что  $f$  удовлетворяет условию Поляка-Лоясиевича (PL), если для некоторого  $\mu > 0$  выполняется

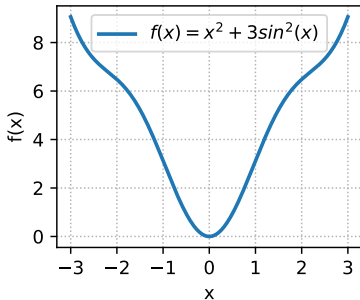
$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x$$

Интересно, что градиентный спуск может сходиться линейно даже без выпуклости.

Следующие функции удовлетворяют условию PL, но не являются выпуклыми. 📄Код

$$f(x) = x^2 + 3\sin^2(x)$$

Function, that satisfies  
Polyak-Lojasiewicz condition



$$f(x, y) = \frac{(y - \sin x)^2}{2}$$

Non-convex PL function



## i Theorem

Рассмотрим задачу

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

и предположим, что  $f$  является PL-функцией с константой  $\mu$  и  $L$ -гладкой, для некоторых  $L \geq \mu > 0$ . Рассмотрим последовательность  $(x^k)_{k \in \mathbb{N}}$ , сгенерированную методом градиентного спуска из точки  $x^0$  с постоянным шагом  $\alpha$ , удовлетворяющим  $0 < \alpha \leq \frac{1}{L}$ . Пусть  $f^* = \min_{x \in \mathbb{R}^d} f(x)$ . Тогда:

$$f(x^k) - f^* \leq (1 - \alpha\mu)^k (f(x^0) - f^*).$$

## i Theorem

Рассмотрим задачу

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

и предположим, что  $f$  является PL-функцией с константой  $\mu$  и  $L$ -гладкой, для некоторых  $L \geq \mu > 0$ . Рассмотрим последовательность  $(x^k)_{k \in \mathbb{N}}$ , сгенерированную методом градиентного спуска из точки  $x^0$  с постоянным шагом  $\alpha$ , удовлетворяющим  $0 < \alpha \leq \frac{1}{L}$ . Пусть  $f^* = \min_{x \in \mathbb{R}^d} f(x)$ . Тогда:

$$f(x^k) - f^* \leq (1 - \alpha\mu)^k (f(x^0) - f^*).$$

## i Theorem

Если функция  $f(x)$  дифференцируема и  $\mu$ -сильно выпукла, то она является PL-функцией.



## Анализ сходимости

Используем  $L$ -гладкость вместе с правилом обновления, чтобы записать:

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2$$

## Анализ сходимости

Используем  $L$ -гладкость вместе с правилом обновления, чтобы записать:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \alpha \|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2} \|\nabla f(x^k)\|^2 \end{aligned}$$

## Анализ сходимости

Используем  $L$ -гладкость вместе с правилом обновления, чтобы записать:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \alpha \|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2} \|\nabla f(x^k)\|^2 \\ &= f(x^k) - \frac{\alpha}{2} (2 - L\alpha) \|\nabla f(x^k)\|^2 \end{aligned}$$

## Анализ сходимости

Используем  $L$ -гладкость вместе с правилом обновления, чтобы записать:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \alpha \|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2} \|\nabla f(x^k)\|^2 \\ &= f(x^k) - \frac{\alpha}{2} (2 - L\alpha) \|\nabla f(x^k)\|^2 \\ &\leq f(x^k) - \frac{\alpha}{2} \|\nabla f(x^k)\|^2, \end{aligned}$$

## Анализ сходимости

Используем  $L$ -гладкость вместе с правилом обновления, чтобы записать:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \alpha \|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2} \|\nabla f(x^k)\|^2 \\ &= f(x^k) - \frac{\alpha}{2} (2 - L\alpha) \|\nabla f(x^k)\|^2 \\ &\leq f(x^k) - \frac{\alpha}{2} \|\nabla f(x^k)\|^2, \end{aligned}$$

## Анализ сходимости

Используем  $L$ -гладкость вместе с правилом обновления, чтобы записать:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \alpha \|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2} \|\nabla f(x^k)\|^2 \\ &= f(x^k) - \frac{\alpha}{2} (2 - L\alpha) \|\nabla f(x^k)\|^2 \\ &\leq f(x^k) - \frac{\alpha}{2} \|\nabla f(x^k)\|^2, \end{aligned}$$

где в последнем неравенстве использована гипотеза о шаге  $\alpha L \leq 1$ .

## Анализ сходимости

Используем  $L$ -гладкость вместе с правилом обновления, чтобы записать:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \alpha \|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2} \|\nabla f(x^k)\|^2 \\ &= f(x^k) - \frac{\alpha}{2} (2 - L\alpha) \|\nabla f(x^k)\|^2 \\ &\leq f(x^k) - \frac{\alpha}{2} \|\nabla f(x^k)\|^2, \end{aligned}$$

где в последнем неравенстве использована гипотеза о шаге  $\alpha L \leq 1$ .

Теперь используем свойство PL-функции и получаем:

$$f(x^{k+1}) \leq f(x^k) - \alpha\mu(f(x^k) - f^*).$$

Вычтя  $f^*$  из обеих частей этого неравенства и применив рекурсию, мы получим искомый результат.

## Выпуклый гладкий случай



# Выпуклый гладкий случай

## i Theorem

Рассмотрим задачу

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

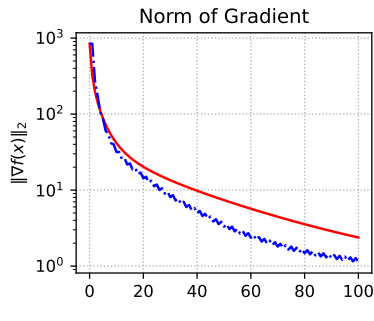
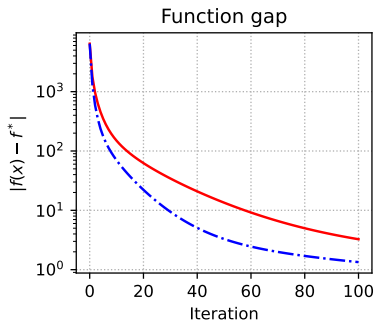
Пусть  $x^* = \arg \min_{x \in \mathbb{R}^d} f(x)$ , а  $f^* = f(x^*)$ . Предположим, что  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  является выпуклой и  $L$ -гладкой функцией, для некоторого  $L > 0$ . Пусть  $(x_k)_{k \in \mathbb{N}}$  — последовательность итераций, сгенерированная методом градиентного спуска из точки  $x_0$  с постоянным шагом  $\alpha$ , удовлетворяющим  $0 < \alpha \leq \frac{1}{L}$ . Тогда для всех  $k \in \mathbb{N}$  справедливо:

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

# Численные эксперименты

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{2} x^T A x - b^T x \right\}, \quad \mu = 0, \quad L = 100.$$

Convex quadratics.  $n=60$ , random matrix.

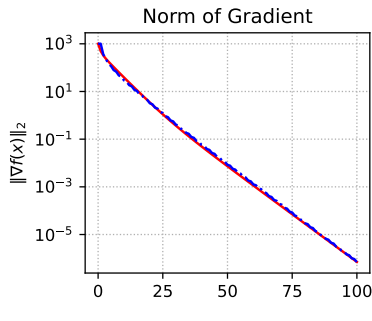
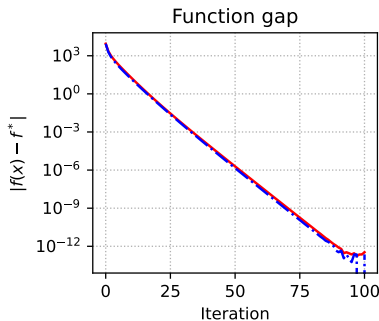
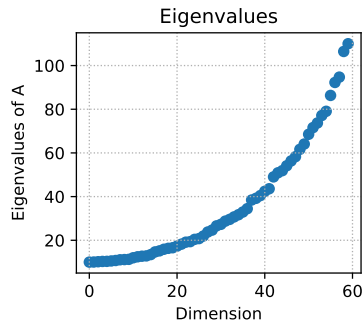


— Gradient Descent    -.- Steepest Descent

## Численные эксперименты

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{2} x^T A x - b^T x \right\}, \quad \mu = 10, \quad L = 110.$$

Strongly convex quadratics.  $n=60$ , random matrix.

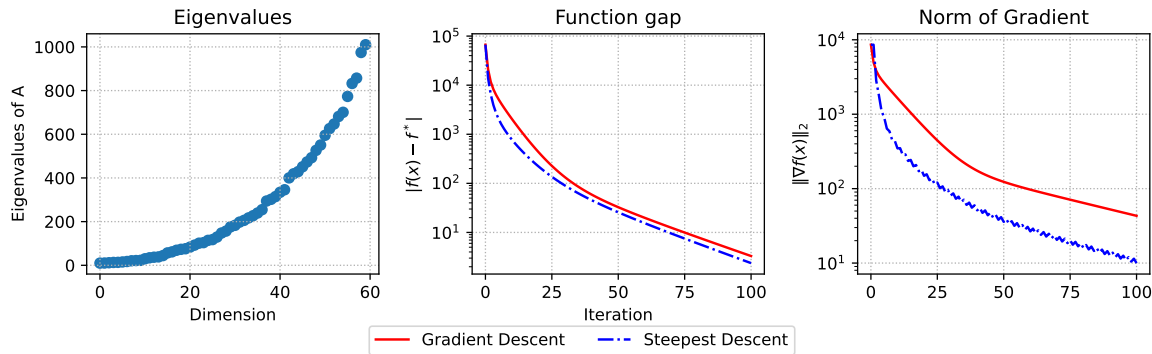


— Gradient Descent    - - - Steepest Descent

# Численные эксперименты

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{2} x^T A x - b^T x \right\}, \quad \mu = 10, \quad L = 1000.$$

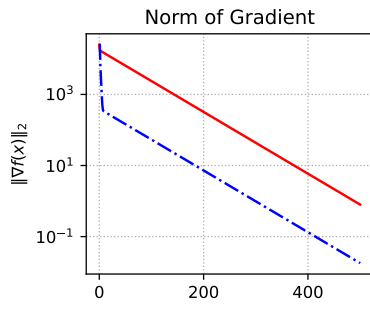
Strongly convex quadratics.  $n=60$ , random matrix.



# Численные эксперименты

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{2} x^T A x - b^T x \right\}, \quad \mu = 10, \quad L = 1000.$$

Strongly convex quadratics.  $n=60$ , clustered matrix.

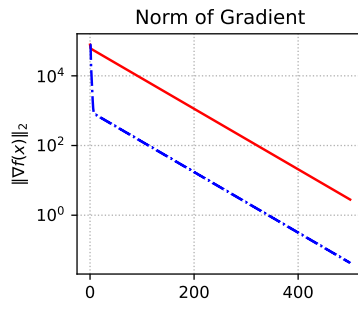
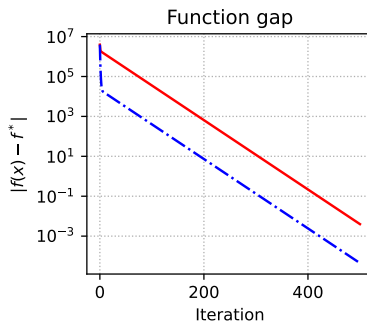
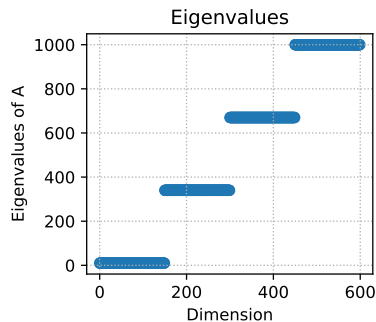


— Gradient Descent    -.- Steepest Descent

# Численные эксперименты

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{2} x^T A x - b^T x \right\}, \quad \mu = 10, \quad L = 1000.$$

Strongly convex quadratics.  $n=600$ , clustered matrix.

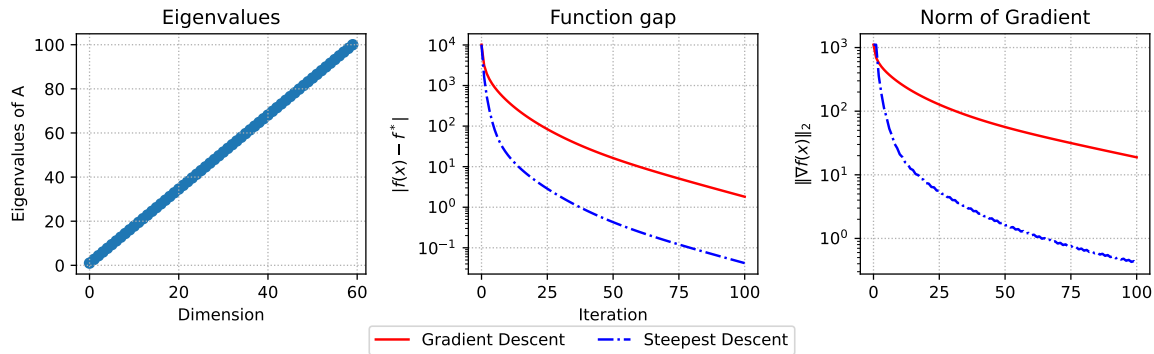


— Gradient Descent    -.- Steepest Descent

# Численные эксперименты

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{2} x^T A x - b^T x \right\}, \quad \mu = 10, \quad L = 1000.$$

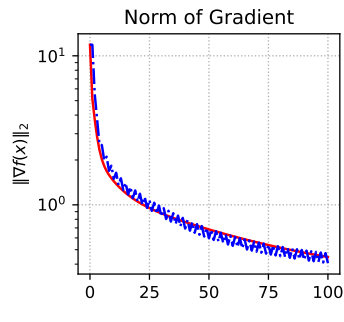
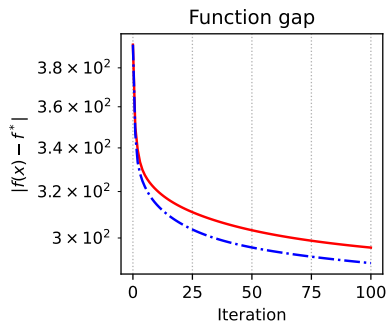
Strongly convex quadratics.  $n=60$ , uniform spectrum matrix.



# Численные эксперименты

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{2} x^T A x - b^T x \right\}$$

Strongly convex quadratics.  $n=60$ , Hilbert matrix.



— Gradient Descent    - - - Steepest Descent



# Численные эксперименты

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{\mu}{2} \|x\|_2^2 + \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-y_i \langle a_i, x \rangle)) \right\}$$

Convex binary logistic regression.  $\mu=0$ .



# Численные эксперименты

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{\mu}{2} \|x\|_2^2 + \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-y_i \langle a_i, x \rangle)) \right\}$$

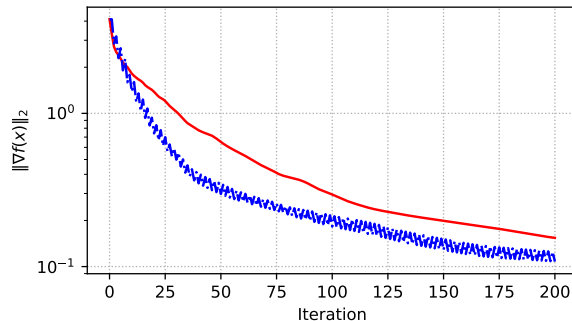
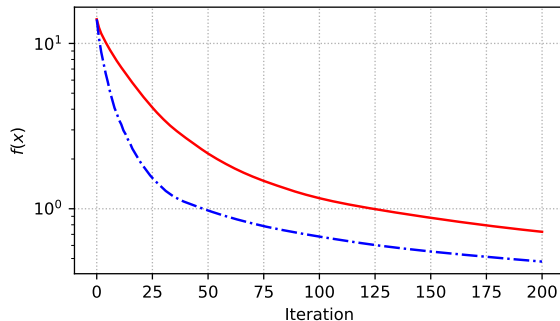
Strongly convex binary logistic regression.  $\mu=0.1$ .



## Численные эксперименты

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{\mu}{2} \|x\|_2^2 + \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-y_i \langle a_i, x \rangle)) \right\}$$

Regularized binary logistic regression.  $n=300$ .  $m=1000$ .  $\mu=0$

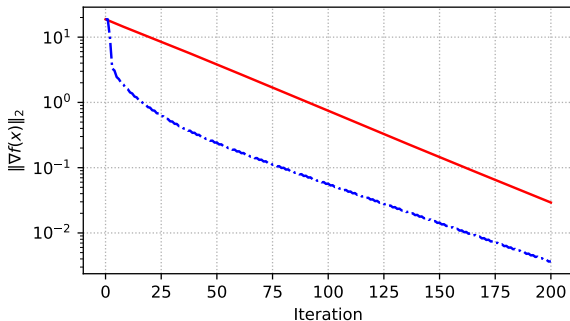
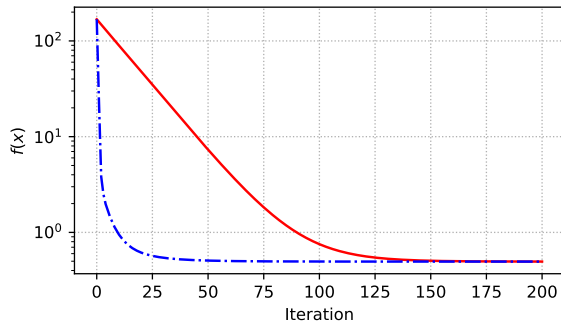


— Gradient Descent    -.- Steepest Descent

## Численные эксперименты

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{\mu}{2} \|x\|_2^2 + \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-y_i \langle a_i, x \rangle)) \right\}$$

Regularized binary logistic regression.  $n=300$ .  $m=1000$ .  $\mu=1$



— Gradient Descent    -.- Steepest Descent

# Сходимость градиентного спуска

Градиентный спуск:

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

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

выпуклая (негладкая)	гладкая (невыпуклая)	гладкая & выпуклая	гладкая & сильно выпуклая
$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\  = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\ x_k - x^*\ ^2 = \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\mu} \log \frac{1}{\varepsilon}\right)$

# Сходимость градиентного спуска

Градиентный спуск:

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

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

выпуклая (негладкая)	гладкая (невыпуклая)	гладкая & выпуклая	гладкая & сильно выпуклая
$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\  = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\ x_k - x^*\ ^2 = \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\mu} \log \frac{1}{\varepsilon}\right)$

Для гладкой сильно выпуклой функции мы имеем:

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*).$$

# Сходимость градиентного спуска

Градиентный спуск:

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

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

выпуклая (негладкая)	гладкая (невыпуклая)	гладкая & выпуклая	гладкая & сильно выпуклая
$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\  = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\ x_k - x^*\ ^2 = \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\mu} \log \frac{1}{\varepsilon}\right)$

Для гладкой сильно выпуклой функции мы имеем:

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*).$$

Обратите внимание, что для любого  $x$ , поскольку  $e^{-x}$  выпуклая и  $1 - x$  является её касательной в точке  $x = 0$ , мы имеем:

$$1 - x \leq e^{-x}$$

# Сходимость градиентного спуска

Градиентный спуск:

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

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

выпуклая (негладкая)	гладкая (невыпуклая)	гладкая & выпуклая	гладкая & сильно выпуклая
$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\  = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\ x_k - x^*\ ^2 = \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\mu} \log \frac{1}{\varepsilon}\right)$

Для гладкой сильно выпуклой функции мы имеем:

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*).$$

Наконец:

$$\varepsilon = f(x_{k_\varepsilon}) - f^*$$

Обратите внимание, что для любого  $x$ , поскольку  $e^{-x}$  выпуклая и  $1 - x$  является её касательной в точке  $x = 0$ , мы имеем:

$$1 - x \leq e^{-x}$$



# Сходимость градиентного спуска

Градиентный спуск:

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

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

выпуклая (негладкая)	гладкая (невыпуклая)	гладкая & выпуклая	гладкая & сильно выпуклая
$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\  = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\ x_k - x^*\ ^2 = \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\mu \log \frac{1}{\varepsilon}}\right)$

Для гладкой сильно выпуклой функции мы имеем:

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*).$$

Наконец:

$$\varepsilon = f(x_{k_\varepsilon}) - f^* \leq \left(1 - \frac{\mu}{L}\right)^{k_\varepsilon} (f(x_0) - f^*)$$

Обратите внимание, что для любого  $x$ , поскольку  $e^{-x}$  выпуклая и  $1 - x$  является её касательной в точке  $x = 0$ , мы имеем:

$$1 - x \leq e^{-x}$$

# Сходимость градиентного спуска

Градиентный спуск:

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

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

выпуклая (негладкая)	гладкая (невыпуклая)	гладкая & выпуклая	гладкая & сильно выпуклая
$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\  = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\ x_k - x^*\ ^2 = \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\mu} \log \frac{1}{\varepsilon}\right)$

Для гладкой сильно выпуклой функции мы имеем:

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*).$$

Обратите внимание, что для любого  $x$ , поскольку  $e^{-x}$  выпуклая и  $1 - x$  является её касательной в точке  $x = 0$ , мы имеем:

$$1 - x \leq e^{-x}$$

Наконец:

$$\begin{aligned} \varepsilon = f(x_{k_\varepsilon}) - f^* &\leq \left(1 - \frac{\mu}{L}\right)^{k_\varepsilon} (f(x_0) - f^*) \\ &\leq \exp\left(-k_\varepsilon \frac{\mu}{L}\right) (f(x_0) - f^*) \end{aligned}$$

# Сходимость градиентного спуска

Градиентный спуск:

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

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

выпуклая (негладкая)	гладкая (невыпуклая)	гладкая & выпуклая	гладкая & сильно выпуклая
$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\  = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\ x_k - x^*\ ^2 = \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon = \mathcal{O}\left(\kappa \log \frac{1}{\varepsilon}\right)$

Для гладкой сильно выпуклой функции мы имеем:

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*).$$

Обратите внимание, что для любого  $x$ , поскольку  $e^{-x}$  выпуклая и  $1 - x$  является её касательной в точке  $x = 0$ , мы имеем:

$$1 - x \leq e^{-x}$$

Наконец:

$$\begin{aligned} \varepsilon &= f(x_{k_\varepsilon}) - f^* \leq \left(1 - \frac{\mu}{L}\right)^{k_\varepsilon} (f(x_0) - f^*) \\ &\leq \exp\left(-k_\varepsilon \frac{\mu}{L}\right) (f(x_0) - f^*) \\ k_\varepsilon &\geq \kappa \log \frac{f(x_0) - f^*}{\varepsilon} = \mathcal{O}\left(\kappa \log \frac{1}{\varepsilon}\right) \end{aligned}$$

# Сходимость градиентного спуска

**Вопрос:** Можно ли добиться лучшей скорости сходимости, используя только информацию первого порядка?

# Сходимость градиентного спуска

**Вопрос:** Можно ли добиться лучшей скорости сходимости, используя только информацию первого порядка?  
**Да, можно.**

## Нижние оценки

- Как правило, это гораздо более нетривиальные результаты - они показывают, что никакой метод не может сходиться быстрее, чем нижняя оценка на выбранном классе функций.

---

<sup>1</sup>Carmon, Duchi, Hinder, Sidford, 2017

<sup>2</sup>Nemirovski, Yudin, 1979

## Нижние оценки

- Как правило, это гораздо более нетривиальные результаты - они показывают, что никакой метод не может сходиться быстрее, чем нижняя оценка на выбранном классе функций.
- Часто, эти результаты получаются путём предъявления конкретной функции из класса, для которой никакой метод не может сходиться быстрее, чем нижняя оценка.

---

<sup>1</sup>Carmon, Duchi, Hinder, Sidford, 2017

<sup>2</sup>Nemirovski, Yudin, 1979

## Нижние оценки

- Как правило, это гораздо более нетривиальные результаты - они показывают, что никакой метод не может сходиться быстрее, чем нижняя оценка на выбранном классе функций.
- Часто, эти результаты получаются путём предъявления конкретной функции из класса, для которой никакой метод не может сходиться быстрее, чем нижняя оценка.
- Для нижних оценок пишут  $\Omega(\cdot)$  вместо  $\mathcal{O}(\cdot)$ .

---

<sup>1</sup>Carmon, Duchi, Hinder, Sidford, 2017

<sup>2</sup>Nemirovski, Yudin, 1979



## Нижние оценки

- Как правило, это гораздо более нетривиальные результаты - они показывают, что никакой метод не может сходиться быстрее, чем нижняя оценка на выбранном классе функций.
- Часто, эти результаты получаются путём предъявления конкретной функции из класса, для которой никакой метод не может сходиться быстрее, чем нижняя оценка.
- Для нижних оценок пишут  $\Omega(\cdot)$  вместо  $\mathcal{O}(\cdot)$ .

---

<sup>1</sup>Carmon, Duchi, Hinder, Sidford, 2017

<sup>2</sup>Nemirovski, Yudin, 1979

## Нижние оценки

- Как правило, это гораздо более нетривиальные результаты - они показывают, что никакой метод не может сходиться быстрее, чем нижняя оценка на выбранном классе функций.
- Часто, эти результаты получаются путём предъявления конкретной функции из класса, для которой никакой метод не может сходиться быстрее, чем нижняя оценка.
- Для нижних оценок пишут  $\Omega(\cdot)$  вместо  $\mathcal{O}(\cdot)$ .

выпуклая (негладкая)	гладкая (невыпуклая) <sup>1</sup>	гладкая & выпуклая <sup>2</sup>	гладкая & сильно выпуклая
$f(x_k) - f^* = \Omega\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \Omega\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\  = \Omega\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \Omega\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \Omega\left(\frac{1}{k^2}\right)$ $k_\varepsilon = \Omega\left(\frac{1}{\varepsilon}\right)$	$f(x_k) - f^* = \Omega\left(\left(\frac{\sqrt{n}-1}{\sqrt{n}+1}\right)^{2k}\right)$ $k_\varepsilon = \Omega\left(\sqrt{n} \log \frac{1}{\varepsilon}\right)$

<sup>1</sup>Carmon, Duchi, Hinder, Sidford, 2017

<sup>2</sup>Nemirovski, Yudin, 1979

## Нижние оценки

- Как правило, это гораздо более нетривиальные результаты - они показывают, что никакой метод не может сходиться быстрее, чем нижняя оценка на выбранном классе функций.
- Часто, эти результаты получаются путём предъявления конкретной функции из класса, для которой никакой метод не может сходиться быстрее, чем нижняя оценка.
- Для нижних оценок пишут  $\Omega(\cdot)$  вместо  $\mathcal{O}(\cdot)$ .

выпуклая (негладкая)	гладкая (невыпуклая) <sup>1</sup>	гладкая & выпуклая <sup>2</sup>	гладкая & сильно выпуклая
$f(x_k) - f^* = \Omega\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \Omega\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\  = \Omega\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \Omega\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \Omega\left(\frac{1}{k^2}\right)$ $k_\varepsilon = \Omega\left(\frac{1}{\sqrt{\varepsilon}}\right)$	$f(x_k) - f^* = \Omega\left(\left(\frac{\sqrt{n}-1}{\sqrt{n}+1}\right)^{2k}\right)$ $k_\varepsilon = \Omega\left(\sqrt{n} \log \frac{1}{\varepsilon}\right)$

Например, из таблицы выше следует, что никакой метод первого порядка определённой формы не может сходиться быстрее, чем  $\Omega\left(\frac{1}{k^2}\right)$  ( $\Omega\left(\frac{1}{\sqrt{\varepsilon}}\right)$  для гладкой выпуклой функции) для гладкой выпуклой функции.

<sup>1</sup>Carmon, Duchi, Hinder, Sidford, 2017

<sup>2</sup>Nemirovski, Yudin, 1979

## Бонус: доказательства сходимости

# Любая $\mu$ -сильно выпуклая дифференцируемая функция является PL-функцией

## Theorem

Если функция  $f(x)$  дифференцируема и  $\mu$ -сильно выпукла, то она является PL-функцией.

## Доказательство

По критерию сильной выпуклости первого порядка:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|_2^2$$

Положим  $y = x^*$ :

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2} \|x^* - x\|_2^2$$

# Любая $\mu$ -сильно выпуклая дифференцируемая функция является PL-функцией

## i Theorem

Если функция  $f(x)$  дифференцируема и  $\mu$ -сильно выпукла, то она является PL-функцией.

## Доказательство

По критерию сильной выпуклости первого порядка:

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2}\|y - x\|_2^2$$

Положим  $y = x^*$ :

$$f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2$$

$$f(x) - f(x^*) \leq \nabla f(x)^T(x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 =$$

# Любая $\mu$ -сильно выпуклая дифференцируемая функция является PL-функцией

## i Theorem

Если функция  $f(x)$  дифференцируема и  $\mu$ -сильно выпукла, то она является PL-функцией.

## Доказательство

По критерию сильной выпуклости первого порядка:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|_2^2$$

Положим  $y = x^*$ :

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2} \|x^* - x\|_2^2 \\ f(x) - f(x^*) &\leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2} \|x^* - x\|_2^2 = \\ &= (\nabla f(x) - \frac{\mu}{2} (x^* - x))^T (x - x^*) = \end{aligned}$$

# Любая $\mu$ -сильно выпуклая дифференцируемая функция является PL-функцией

## i Theorem

Если функция  $f(x)$  дифференцируема и  $\mu$ -сильно выпукла, то она является PL-функцией.

## Доказательство

По критерию сильной выпуклости первого порядка:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|_2^2$$

Положим  $y = x^*$ :

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2} \|x^* - x\|_2^2 \\ f(x) - f(x^*) &\leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2} \|x^* - x\|_2^2 = \\ &= \left( \nabla f(x) - \frac{\mu}{2} (x^* - x) \right)^T (x - x^*) = \\ &= \frac{1}{2} \left( \frac{2}{\sqrt{\mu}} \nabla f(x) - \sqrt{\mu} (x^* - x) \right)^T \sqrt{\mu} (x - x^*) \end{aligned}$$



# Любая $\mu$ -сильно выпуклая дифференцируемая функция является PL-функцией

## i Theorem

Если функция  $f(x)$  дифференцируема и  $\mu$ -сильно выпукла, то она является PL-функцией.

## Доказательство

По критерию сильной выпуклости первого порядка:

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2}\|y - x\|_2^2$$

Положим  $y = x^*$ :

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^T(x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2 \\ f(x) - f(x^*) &\leq \nabla f(x)^T(x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 = \\ &= \left(\nabla f(x) - \frac{\mu}{2}(x^* - x)\right)^T(x - x^*) = \\ &= \frac{1}{2} \left(\frac{2}{\sqrt{\mu}}\nabla f(x) - \sqrt{\mu}(x^* - x)\right)^T \sqrt{\mu}(x - x^*) \end{aligned}$$

# Любая $\mu$ -сильно выпуклая дифференцируемая функция является PL-функцией

## i Theorem

Если функция  $f(x)$  дифференцируема и  $\mu$ -сильно выпукла, то она является PL-функцией.

## Доказательство

По критерию сильной выпуклости первого порядка:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|_2^2$$

Положим  $y = x^*$ :

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2} \|x^* - x\|_2^2 \\ f(x) - f(x^*) &\leq \nabla f(x)^T (x - x^*) - \frac{\mu}{2} \|x^* - x\|_2^2 = \\ &= \left( \nabla f(x) - \frac{\mu}{2} (x^* - x) \right)^T (x - x^*) = \\ &= \frac{1}{2} \left( \frac{2}{\sqrt{\mu}} \nabla f(x) - \sqrt{\mu} (x^* - x) \right)^T \sqrt{\mu} (x - x^*) \end{aligned}$$

$$\begin{aligned} \text{Пусть } a &= \frac{1}{\sqrt{\mu}} \nabla f(x) \text{ и} \\ b &= \sqrt{\mu} (x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x) \end{aligned}$$

# Любая $\mu$ -сильно выпуклая дифференцируемая функция является PL-функцией

## i Theorem

Если функция  $f(x)$  дифференцируема и  $\mu$ -сильно выпукла, то она является PL-функцией.

## Доказательство

По критерию сильной выпуклости первого порядка:

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2}\|y - x\|_2^2$$

Положим  $y = x^*$ :

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^T(x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2 \\ f(x) - f(x^*) &\leq \nabla f(x)^T(x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 = \\ &= (\nabla f(x) - \frac{\mu}{2}(x^* - x))^T(x - x^*) = \\ &= \frac{1}{2} \left( \frac{2}{\sqrt{\mu}} \nabla f(x) - \sqrt{\mu}(x^* - x) \right)^T \sqrt{\mu}(x - x^*) \end{aligned}$$

Пусть  $a = \frac{1}{\sqrt{\mu}} \nabla f(x)$  и

$$b = \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x)$$

Тогда  $a + b = \sqrt{\mu}(x - x^*)$  и

$$a - b = \frac{2}{\sqrt{\mu}} \nabla f(x) - \sqrt{\mu}(x - x^*)$$

Любая  $\mu$ -сильно выпуклая дифференцируемая функция является PL-функцией

$$f(x) - f(x^*) \leq \frac{1}{2} \left( \frac{1}{\mu} \|\nabla f(x)\|_2^2 - \left\| \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x) \right\|_2^2 \right)$$

Любая  $\mu$ -сильно выпуклая дифференцируемая функция является PL-функцией

$$f(x) - f(x^*) \leq \frac{1}{2} \left( \frac{1}{\mu} \|\nabla f(x)\|_2^2 - \left\| \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x) \right\|_2^2 \right)$$
$$f(x) - f(x^*) \leq \frac{1}{2\mu} \|\nabla f(x)\|_2^2,$$

Любая  $\mu$ -сильно выпуклая дифференцируемая функция является PL-функцией

$$f(x) - f(x^*) \leq \frac{1}{2} \left( \frac{1}{\mu} \|\nabla f(x)\|_2^2 - \left\| \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x) \right\|_2^2 \right)$$
$$f(x) - f(x^*) \leq \frac{1}{2\mu} \|\nabla f(x)\|_2^2,$$

## Любая $\mu$ -сильно выпуклая дифференцируемая функция является PL-функцией

$$f(x) - f(x^*) \leq \frac{1}{2} \left( \frac{1}{\mu} \|\nabla f(x)\|_2^2 - \left\| \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x) \right\|_2^2 \right)$$
$$f(x) - f(x^*) \leq \frac{1}{2\mu} \|\nabla f(x)\|_2^2,$$

которое является точным условием PL. Это означает, что мы уже имеем доказательство линейной сходимости для любой сильно выпуклой функции.

# Сходимость градиентного спуска в выпуклом гладком случае [1/4]

## i Theorem

Рассмотрим задачу

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

Пусть  $f^* = \min_{x \in \mathbb{R}^d} f(x)$ . Предположим, что  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  является выпуклой и  $L$ -гладкой функцией, для некоторого  $L > 0$ . Пусть  $(x_k)_{k \in \mathbb{N}}$  — последовательность итераций, сгенерированная методом градиентного спуска из точки  $x_0$  с постоянным шагом  $\alpha$ , удовлетворяющим  $0 < \alpha \leq \frac{1}{L}$ .

Тогда для всех  $k \in \mathbb{N}$  справедливо:

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

**Заметим**, что мы здесь никак не упоминаем точку минимума. То есть, это сходимость  $\forall x \in \mathbb{R}^d$  (в том числе и до точки минимума).



# Сходимость градиентного спуска в выпуклом гладком случае [2/4]

Наш инструментарий:

1. Выпуклость:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall x, y. \quad (1)$$

# Сходимость градиентного спуска в выпуклом гладком случае [2/4]

Наш инструментарий:

1. Выпуклость:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall x, y. \quad (1)$$

2. Гладкость:

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2, \quad \forall x, y. \quad (2)$$

# Сходимость градиентного спуска в выпуклом гладком случае [2/4]

Наш инструментарий:

1. Выпуклость:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall x, y. \quad (1)$$

2. Гладкость:

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2, \quad \forall x, y. \quad (2)$$

3. 3-point identity (по сути, квадрат разности):

$$\|a - b\|^2 = \|a - c - (b - c)\|^2 = \|a - c\|^2 - 2\langle a - c, b - c \rangle + \|b - c\|^2$$

переносим справа все кроме  $\|b - c\|^2$  налево и меняем местами все факторы внутри каждого из перенесенных членов:

$$\|b - c\|^2 = \|b - a\|^2 + 2\langle c - a, c - b \rangle - \|c - a\|^2. \quad (3)$$

# Сходимость градиентного спуска в выпуклом гладком случае [2/4]

Наш инструментарий:

1. Выпуклость:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall x, y. \quad (1)$$

2. Гладкость:

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2, \quad \forall x, y. \quad (2)$$

3. 3-point identity (по сути, квадрат разности):

$$\|a - b\|^2 = \|a - c - (b - c)\|^2 = \|a - c\|^2 - 2\langle a - c, b - c \rangle + \|b - c\|^2$$

переносим справа все кроме  $\|b - c\|^2$  налево и меняем местами все факторы внутри каждого из перенесенных членов:

$$\|b - c\|^2 = \|b - a\|^2 + 2\langle c - a, c - b \rangle - \|c - a\|^2. \quad (3)$$

- Подставляем в (3)  $b \equiv x$ ,  $c \equiv x_{k+1}$ ,  $a \equiv x_k$  и домножаем все на  $\frac{1}{2}$ :

(4)

# Сходимость градиентного спуска в выпуклом гладком случае [2/4]

Наш инструментарий:

1. Выпуклость:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall x, y. \quad (1)$$

2. Гладкость:

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2, \quad \forall x, y. \quad (2)$$

3. 3-point identity (по сути, квадрат разности):

$$\|a - b\|^2 = \|a - c - (b - c)\|^2 = \|a - c\|^2 - 2\langle a - c, b - c \rangle + \|b - c\|^2$$

переносим справа все кроме  $\|b - c\|^2$  налево и меняем местами все факторы внутри каждого из перенесенных членов:

$$\|b - c\|^2 = \|b - a\|^2 + 2\langle c - a, c - b \rangle - \|c - a\|^2. \quad (3)$$

• Подставляем в (3)  $b \equiv x$ ,  $c \equiv x_{k+1}$ ,  $a \equiv x_k$  и домножаем все на  $\frac{1}{2}$ :

$$\frac{1}{2} \|x - x_{k+1}\|^2 = \frac{1}{2} \|x - x_k\|^2 + \langle x_{k+1} - x_k, x_{k+1} - x \rangle - \frac{1}{2} \|x_{k+1} - x_k\|^2 \quad (4)$$

## Сходимость градиентного спуска в выпуклом гладком случае [2/4]

Наш инструментарий:

1. Выпуклость:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall x, y. \quad (1)$$

2. Гладкость:

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2, \quad \forall x, y. \quad (2)$$

3. 3-point identity (по сути, квадрат разности):

$$\|a - b\|^2 = \|a - c - (b - c)\|^2 = \|a - c\|^2 - 2\langle a - c, b - c \rangle + \|b - c\|^2$$

переносим справа все кроме  $\|b - c\|^2$  налево и меняем местами все факторы внутри каждого из перенесенных членов:

$$\|b - c\|^2 = \|b - a\|^2 + 2\langle c - a, c - b \rangle - \|c - a\|^2. \quad (3)$$

• Подставляем в (3)  $b \equiv x$ ,  $c \equiv x_{k+1}$ ,  $a \equiv x_k$  и домножаем все на  $\frac{1}{2}$ :

$$\begin{aligned} \frac{1}{2} \|x - x_{k+1}\|^2 &= \frac{1}{2} \|x - x_k\|^2 + \langle x_{k+1} - x_k, x_{k+1} - x \rangle - \frac{1}{2} \|x_{k+1} - x_k\|^2 \\ &= \frac{1}{2} \|x - x_k\|^2 - \alpha \langle \nabla f(x_k), x_{k+1} - x \rangle - \frac{1}{2} \|x_{k+1} - x_k\|^2. \end{aligned} \quad (4)$$

## Сходимость градиентного спуска в выпуклом гладком случае [3/4]

- Посмотрим внимательнее на скалярное произведение  $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$  и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

## Сходимость градиентного спуска в выпуклом гладком случае [3/4]

- Посмотрим внимательнее на скалярное произведение  $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$  и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

$$-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle = \alpha (\langle \nabla f(x_k), x - x_k \rangle + \langle \nabla f(x_k), x_k - x_{k+1} \rangle)$$



## Сходимость градиентного спуска в выпуклом гладком случае [3/4]

- Посмотрим внимательнее на скалярное произведение  $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$  и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

$$\begin{aligned} -\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle &= \alpha (\langle \nabla f(x_k), x - x_k \rangle + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \\ &\stackrel{(1)}{\leq} \alpha (f(x) - f(x_k) + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \end{aligned}$$

## Сходимость градиентного спуска в выпуклом гладком случае [3/4]

- Посмотрим внимательнее на скалярное произведение  $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$  и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

$$\begin{aligned} -\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle &= \alpha (\langle \nabla f(x_k), x - x_k \rangle + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \\ &\stackrel{(1)}{\leq} \alpha (f(x) - f(x_k) + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \\ &\stackrel{(2)}{\leq} \alpha \left( f(x) - f(x_{k+1}) + \frac{L}{2} \|x_{k+1} - x_k\|^2 \right), \end{aligned}$$

## Сходимость градиентного спуска в выпуклом гладком случае [3/4]

- Посмотрим внимательнее на скалярное произведение  $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$  и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

$$\begin{aligned} -\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle &= \alpha (\langle \nabla f(x_k), x - x_k \rangle + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \\ &\stackrel{(1)}{\leq} \alpha (f(x) - f(x_k) + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \\ &\stackrel{(2)}{\leq} \alpha \left( f(x) - f(x_{k+1}) + \frac{L}{2} \|x_{k+1} - x_k\|^2 \right), \end{aligned}$$

- Подставляем это все обратно в (4) и используем условие на размер шага  $\alpha \leq \frac{1}{L}$ :

## Сходимость градиентного спуска в выпуклом гладком случае [3/4]

- Посмотрим внимательнее на скалярное произведение  $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$  и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

$$\begin{aligned} -\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle &= \alpha (\langle \nabla f(x_k), x - x_k \rangle + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \\ &\stackrel{(1)}{\leq} \alpha (f(x) - f(x_k) + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \\ &\stackrel{(2)}{\leq} \alpha \left( f(x) - f(x_{k+1}) + \frac{L}{2} \|x_{k+1} - x_k\|^2 \right), \end{aligned}$$

- Подставляем это все обратно в (4) и используем условие на размер шага  $\alpha \leq \frac{1}{L}$ :

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

## Сходимость градиентного спуска в выпуклом гладком случае [3/4]

- Посмотрим внимательнее на скалярное произведение  $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$  и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

$$\begin{aligned} -\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle &= \alpha (\langle \nabla f(x_k), x - x_k \rangle + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \\ &\stackrel{(1)}{\leq} \alpha (f(x) - f(x_k) + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \\ &\stackrel{(2)}{\leq} \alpha \left( f(x) - f(x_{k+1}) + \frac{L}{2} \|x_{k+1} - x_k\|^2 \right), \end{aligned}$$

- Подставляем это все обратно в (4) и используем условие на размер шага  $\alpha \leq \frac{1}{L}$ :

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

## Сходимость градиентного спуска в выпуклом гладком случае [3/4]

- Посмотрим внимательнее на скалярное произведение  $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$  и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

$$\begin{aligned} -\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle &= \alpha (\langle \nabla f(x_k), x - x_k \rangle + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \\ &\stackrel{(1)}{\leq} \alpha (f(x) - f(x_k) + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \\ &\stackrel{(2)}{\leq} \alpha \left( f(x) - f(x_{k+1}) + \frac{L}{2} \|x_{k+1} - x_k\|^2 \right), \end{aligned}$$

- Подставляем это все обратно в (4) и используем условие на размер шага  $\alpha \leq \frac{1}{L}$ :

$$\begin{aligned} \frac{1}{2} \|x - x_{k+1}\|^2 &\leq \frac{1}{2} \|x - x_k\|^2 + \alpha (f(x) - f(x_{k+1})) + \left( \frac{\alpha L}{2} - \frac{1}{2} \right) \|x_{k+1} - x_k\|^2 \\ \frac{1}{2} \|x - x_{k+1}\|^2 - \frac{1}{2} \|x - x_k\|^2 &\leq \alpha (f(x) - f(x_{k+1})) + \left( \frac{\alpha L}{2} - \frac{1}{2} \right) \|x_{k+1} - x_k\|^2 \\ &\stackrel{(\alpha \leq 1/L)}{\leq} \frac{1}{L} (f(x) - f(x_{k+1})). \end{aligned}$$

## Сходимость градиентного спуска в выпуклом гладком случае [3/4]

- Посмотрим внимательнее на скалярное произведение  $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$  и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

$$\begin{aligned} -\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle &= \alpha (\langle \nabla f(x_k), x - x_k \rangle + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \\ &\stackrel{(1)}{\leq} \alpha (f(x) - f(x_k) + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \\ &\stackrel{(2)}{\leq} \alpha \left( f(x) - f(x_{k+1}) + \frac{L}{2} \|x_{k+1} - x_k\|^2 \right), \end{aligned}$$

- Подставляем это все обратно в (4) и используем условие на размер шага  $\alpha \leq \frac{1}{L}$ :

$$\begin{aligned} \frac{1}{2} \|x - x_{k+1}\|^2 &\leq \frac{1}{2} \|x - x_k\|^2 + \alpha (f(x) - f(x_{k+1})) + \left( \frac{\alpha L}{2} - \frac{1}{2} \right) \|x_{k+1} - x_k\|^2 \\ \frac{1}{2} \|x - x_{k+1}\|^2 - \frac{1}{2} \|x - x_k\|^2 &\leq \alpha (f(x) - f(x_{k+1})) + \left( \frac{\alpha L}{2} - \frac{1}{2} \right) \|x_{k+1} - x_k\|^2 \\ &\stackrel{(\alpha \leq 1/L)}{\leq} \frac{1}{L} (f(x) - f(x_{k+1})). \end{aligned}$$

- Переносим правую часть влево, левую - вправо и домножаем на  $L$ :

## Сходимость градиентного спуска в выпуклом гладком случае [3/4]

- Посмотрим внимательнее на скалярное произведение  $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$  и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

$$\begin{aligned} -\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle &= \alpha (\langle \nabla f(x_k), x - x_k \rangle + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \\ &\stackrel{(1)}{\leq} \alpha (f(x) - f(x_k) + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \\ &\stackrel{(2)}{\leq} \alpha \left( f(x) - f(x_{k+1}) + \frac{L}{2} \|x_{k+1} - x_k\|^2 \right), \end{aligned}$$

- Подставляем это все обратно в (4) и используем условие на размер шага  $\alpha \leq \frac{1}{L}$ :

$$\begin{aligned} \frac{1}{2} \|x - x_{k+1}\|^2 &\leq \frac{1}{2} \|x - x_k\|^2 + \alpha (f(x) - f(x_{k+1})) + \left( \frac{\alpha L}{2} - \frac{1}{2} \right) \|x_{k+1} - x_k\|^2 \\ \frac{1}{2} \|x - x_{k+1}\|^2 - \frac{1}{2} \|x - x_k\|^2 &\leq \alpha (f(x) - f(x_{k+1})) + \left( \frac{\alpha L}{2} - \frac{1}{2} \right) \|x_{k+1} - x_k\|^2 \\ &\stackrel{(\alpha \leq 1/L)}{\leq} \frac{1}{L} (f(x) - f(x_{k+1})). \end{aligned}$$

- Переносим правую часть влево, левую - вправо и домножаем на  $L$ :

$$f(x_{k+1}) - f(x) \leq \frac{L}{2} (\|x - x_k\|^2 - \|x - x_{k+1}\|^2).$$



## Сходимость градиентного спуска в выпуклом гладком случае [4/4]

- Берем среднее от левой и правой частей от по всем  $k$  от 0 до  $N - 1$ :

(5)

## Сходимость градиентного спуска в выпуклом гладком случае [4/4]

- Берем среднее от левой и правой частей от по всем  $k$  от 0 до  $N - 1$ :

$$\frac{1}{N} \sum_{k=0}^{N-1} (f(x_{k+1}) - f(x)) \leq \frac{L}{2N} \sum_{k=0}^{N-1} (\|x - x_k\|^2 - \|x - x_{k+1}\|^2)$$

(5)

## Сходимость градиентного спуска в выпуклом гладком случае [4/4]

- Берем среднее от левой и правой частей от по всем  $k$  от 0 до  $N - 1$ :

$$\begin{aligned} \frac{1}{N} \sum_{k=0}^{N-1} (f(x_{k+1}) - f(x)) &\leq \frac{L}{2N} \sum_{k=0}^{N-1} (\|x - x_k\|^2 - \|x - x_{k+1}\|^2) \\ &= \frac{L}{2N} (\|x - x_0\|^2 - \|x - x_{k+1}\|^2) \end{aligned} \quad (5)$$

## Сходимость градиентного спуска в выпуклом гладком случае [4/4]

- Берем среднее от левой и правой частей от по всем  $k$  от 0 до  $N - 1$ :

$$\begin{aligned}\frac{1}{N} \sum_{k=0}^{N-1} (f(x_{k+1}) - f(x)) &\leq \frac{L}{2N} \sum_{k=0}^{N-1} (\|x - x_k\|^2 - \|x - x_{k+1}\|^2) \\ &= \frac{L}{2N} (\|x - x_0\|^2 - \|x - x_{k+1}\|^2) \\ &\leq \frac{L}{2N} \|x - x_0\|^2.\end{aligned}\tag{5}$$

## Сходимость градиентного спуска в выпуклом гладком случае [4/4]

- Берем среднее от левой и правой частей от по всем  $k$  от 0 до  $N - 1$ :

$$\begin{aligned}\frac{1}{N} \sum_{k=0}^{N-1} (f(x_{k+1}) - f(x)) &\leq \frac{L}{2N} \sum_{k=0}^{N-1} (\|x - x_k\|^2 - \|x - x_{k+1}\|^2) \\ &= \frac{L}{2N} (\|x - x_0\|^2 - \|x - x_{k+1}\|^2) \\ &\leq \frac{L}{2N} \|x - x_0\|^2.\end{aligned}\tag{5}$$

- Так как для выпуклых функций (1) градиентный спуск монотонен:

$$\begin{aligned}f(x_k) &\geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle \\ &= f(x_{k+1}) + \alpha \|\nabla f(x_{k+1})\|^2 \\ &\geq f(x_{k+1}),\end{aligned}$$

## Сходимость градиентного спуска в выпуклом гладком случае [4/4]

- Берем среднее от левой и правой частей от по всем  $k$  от 0 до  $N - 1$ :

$$\begin{aligned}\frac{1}{N} \sum_{k=0}^{N-1} (f(x_{k+1}) - f(x)) &\leq \frac{L}{2N} \sum_{k=0}^{N-1} (\|x - x_k\|^2 - \|x - x_{k+1}\|^2) \\ &= \frac{L}{2N} (\|x - x_0\|^2 - \|x - x_{k+1}\|^2) \\ &\leq \frac{L}{2N} \|x - x_0\|^2.\end{aligned}\tag{5}$$

- Так как для выпуклых функций (1) градиентный спуск монотонен:

$$\begin{aligned}f(x_k) &\geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle \\ &= f(x_{k+1}) + \alpha \|\nabla f(x_{k+1})\|^2 \\ &\geq f(x_{k+1}),\end{aligned}$$

$$\text{то } \frac{1}{N} \sum_{i=0}^{N-1} (f(x_{k+1}) - f(x)) \geq \min_{i=0, \dots, N-1} f(x_{i+1}) - f(x) = f(x_N) - f(x).$$

## Сходимость градиентного спуска в выпуклом гладком случае [4/4]

- Берем среднее от левой и правой частей от по всем  $k$  от 0 до  $N - 1$ :

$$\begin{aligned}\frac{1}{N} \sum_{k=0}^{N-1} (f(x_{k+1}) - f(x)) &\leq \frac{L}{2N} \sum_{k=0}^{N-1} (\|x - x_k\|^2 - \|x - x_{k+1}\|^2) \\ &= \frac{L}{2N} (\|x - x_0\|^2 - \|x - x_{k+1}\|^2) \\ &\leq \frac{L}{2N} \|x - x_0\|^2.\end{aligned}\tag{5}$$

- Так как для выпуклых функций (1) градиентный спуск монотонен:

$$\begin{aligned}f(x_k) &\geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle \\ &= f(x_{k+1}) + \alpha \|\nabla f(x_{k+1})\|^2 \\ &\geq f(x_{k+1}),\end{aligned}$$

то  $\frac{1}{N} \sum_{i=0}^{N-1} (f(x_{i+1}) - f(x)) \geq \min_{i=0, \dots, N-1} f(x_{i+1}) - f(x) = f(x_N) - f(x)$ . Подставляя это в (5), получаем искомый результат.