

Градиентный спуск. Теоремы сходимости в гладком случае (выпуклые, сильно выпуклые, PL). Верхние и нижние оценки сходимости

Даниил Меркулов

Методы оптимизации. МФТИ



Градиентный спуск

Направление локального наискорейшего спуска

Рассмотрим линейное приближение
дифференцируемой функции f вдоль
направления h , где $\|h\|_2 = 1$:

Направление локального наискорейшего спуска

Рассмотрим линейное приближение
дифференцируемой функции f вдоль
направления h , где $\|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha)$$

Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль направления h , где $\|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha)$$

Хотим, чтобы h было направлением убывания:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha) < f(x)$$

Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль направления h , где $\|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha)$$

Хотим, чтобы h было направлением убывания:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha) < f(x)$$

Переходя к пределу при $\alpha \rightarrow 0$:

$$\langle \nabla f(x), h \rangle \leq 0$$

Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль направления h , где $\|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha)$$

Также из неравенства Коши–Буняковского получаем:

$$\begin{aligned} |\langle \nabla f(x), h \rangle| &\leq \|\nabla f(x)\|_2 \|h\|_2 \\ \langle \nabla f(x), h \rangle &\geq -\|\nabla f(x)\|_2 \|h\|_2 = -\|\nabla f(x)\|_2 \end{aligned}$$

Хотим, чтобы h было направлением убывания:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha) < f(x)$$

Переходя к пределу при $\alpha \rightarrow 0$:

$$\langle \nabla f(x), h \rangle \leq 0$$

Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль направления h , где $\|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha)$$

Хотим, чтобы h было направлением убывания:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha) < f(x)$$

Переходя к пределу при $\alpha \rightarrow 0$:

$$\langle \nabla f(x), h \rangle \leq 0$$

Также из неравенства Коши–Буняковского получаем:

$$|\langle \nabla f(x), h \rangle| \leq \|\nabla f(x)\|_2 \|h\|_2$$

$$\langle \nabla f(x), h \rangle \geq -\|\nabla f(x)\|_2 \|h\|_2 = -\|\nabla f(x)\|_2$$

Таким образом, направление антиградиента

$$h = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}$$

представляет собой направление **наискорейшего локального убывания** функции f .

Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль направления h , где $\|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha)$$

Хотим, чтобы h было направлением убывания:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha) < f(x)$$

Переходя к пределу при $\alpha \rightarrow 0$:

$$\langle \nabla f(x), h \rangle \leq 0$$

Также из неравенства Коши–Буняковского получаем:

$$|\langle \nabla f(x), h \rangle| \leq \|\nabla f(x)\|_2 \|h\|_2$$

$$\langle \nabla f(x), h \rangle \geq -\|\nabla f(x)\|_2 \|h\|_2 = -\|\nabla f(x)\|_2$$

Таким образом, направление антиградиента

$$h = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}$$

представляет собой направление **наискорейшего локального убывания** функции f .

Итерация метода имеет вид:

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

Дифференциальное уравнение градиентного потока

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x(t)). \quad (\text{GF})$$

Дифференциальное уравнение градиентного потока

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x(t)). \quad (\text{GF})$$

Дискретизируем его на равномерной сетке с шагом α :

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

Дифференциальное уравнение градиентного потока

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x(t)). \quad (\text{GF})$$

Дискретизируем его на равномерной сетке с шагом α :

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

где $x^k \equiv x(t_k)$ и $\alpha = t_{k+1} - t_k$ — шаг сетки.

Отсюда получаем выражение для x^{k+1} :

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

являющееся точной формулой обновления градиентного спуска.

Открыть в Colab 

Дифференциальное уравнение градиентного потока

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x(t)). \quad (\text{GF})$$

Дискретизируем его на равномерной сетке с шагом α :

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

где $x^k \equiv x(t_k)$ и $\alpha = t_{k+1} - t_k$ — шаг сетки.

Отсюда получаем выражение для x^{k+1} :

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

являющееся точной формулой обновления градиентного спуска.

Открыть в Colab ♣

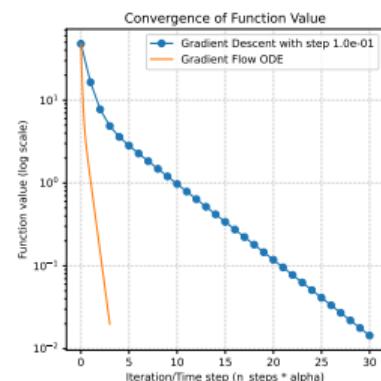
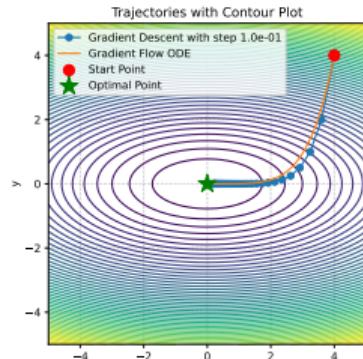
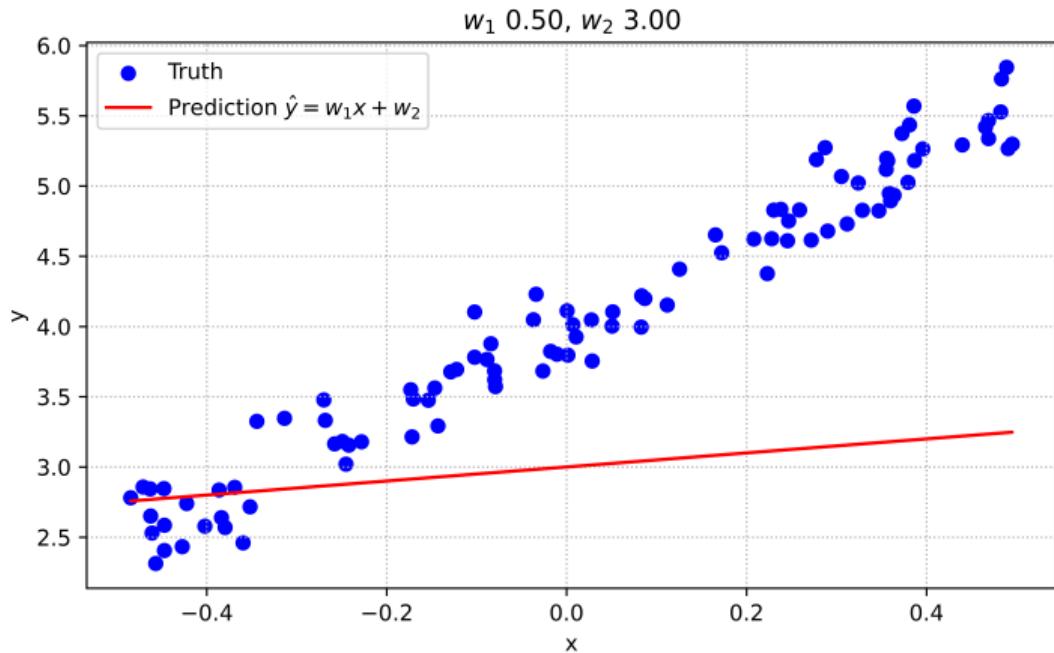
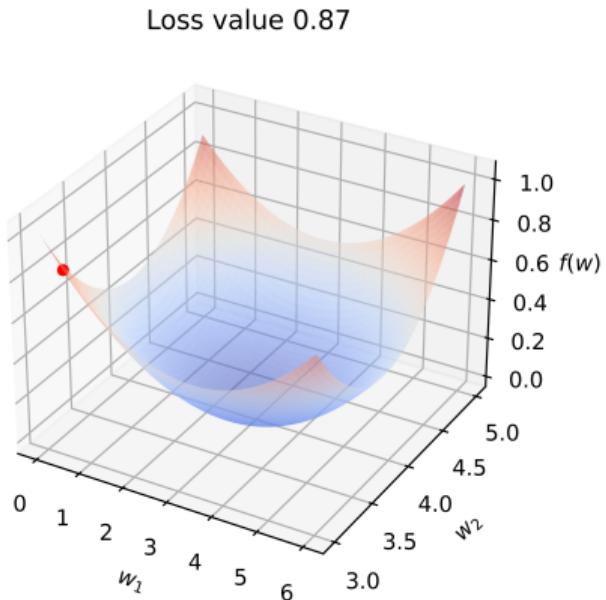


Рис. 1: Траектория градиентного потока

Сходимость градиентного спуска

Существенно зависит от выбора шага α :



Точный линейный поиск (метод наискорейшего спуска)

$$\alpha_k = \arg \min_{\alpha \in \mathbb{R}^+} f(x^k - \alpha \nabla f(x^k))$$

Подход скорее теоретический, чем практический: он удобен для анализа сходимости, но точный линейный поиск часто затруднён, если вычисление функции занимает слишком много времени или стоит слишком дорого.

Интересное теоретическое свойство этого метода заключается в том, что градиенты на соседних итерациях ортогональны. Условие оптимальности по α_k даёт

$$\frac{d}{d\alpha} f(x^k - \alpha \nabla f(x^k)) \Big|_{\alpha=\alpha_k} = 0.$$

Точный линейный поиск (метод наискорейшего спуска)

$$\alpha_k = \arg \min_{\alpha \in \mathbb{R}^+} f(x^k - \alpha \nabla f(x^k))$$

Подход скорее теоретический, чем практический: он удобен для анализа сходимости, но точный линейный поиск часто затруднён, если вычисление функции занимает слишком много времени или стоит слишком дорого.

Интересное теоретическое свойство этого метода заключается в том, что градиенты на соседних итерациях ортогональны. Условие оптимальности по α_k даёт

$$\frac{d}{d\alpha} f(x^k - \alpha \nabla f(x^k)) \Big|_{\alpha=\alpha_k} = 0.$$

Условия оптимальности:

Точный линейный поиск (метод наискорейшего спуска)

$$\alpha_k = \arg \min_{\alpha \in \mathbb{R}^+} f(x^k - \alpha \nabla f(x^k))$$

Подход скорее теоретический, чем практический: он удобен для анализа сходимости, но точный линейный поиск часто затруднён, если вычисление функции занимает слишком много времени или стоит слишком дорого.

Интересное теоретическое свойство этого метода заключается в том, что градиенты на соседних итерациях ортогональны. Условие оптимальности по α_k даёт

$$\frac{d}{d\alpha} f(x^k - \alpha \nabla f(x^k)) \Big|_{\alpha=\alpha_k} = 0.$$

Условия оптимальности:

$$\nabla f(x^{k+1})^\top \nabla f(x^k) = 0$$



Рис. 2: Наискорейший спуск

Открыть в Colab ♣

Сдвиг координат

Рассмотрим следующую задачу квадратичной оптимизации:

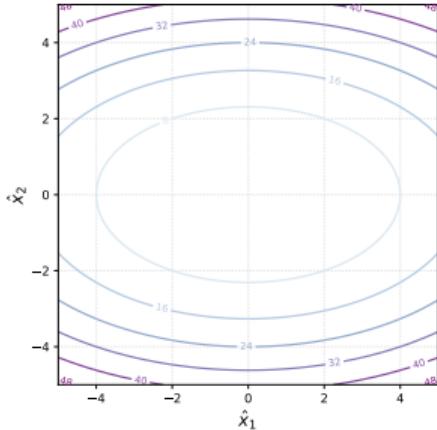
$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

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 влияет на процесс оптимизации.

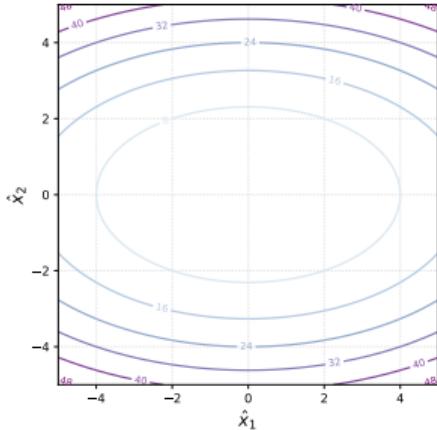


Сдвиг координат

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 влияет на процесс оптимизации.
- Во-вторых, у нас есть спектральное разложение матрицы $A = Q\Lambda Q^T$.

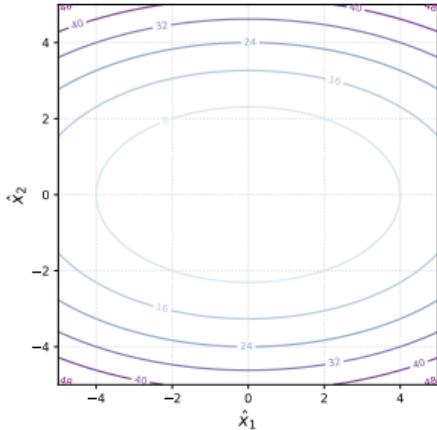
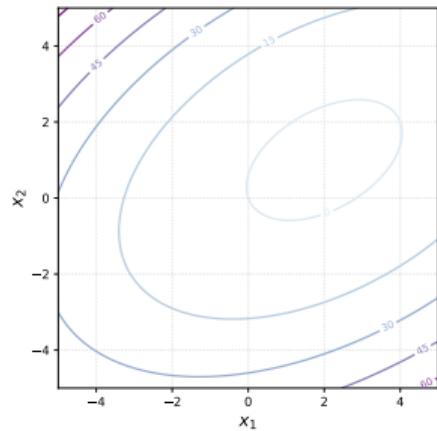


Сдвиг координат

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 влияет 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 $x = Q\hat{x} + x^*$.



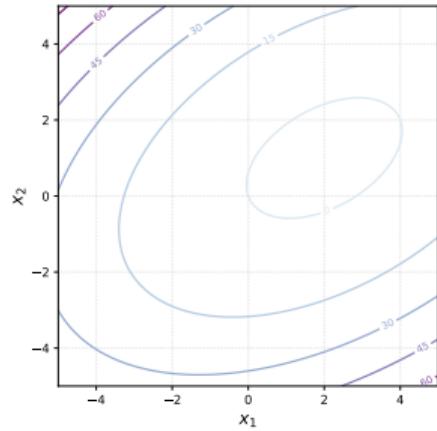
Сдвиг координат

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 влияет 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 $x = Q\hat{x} + x^*$.

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



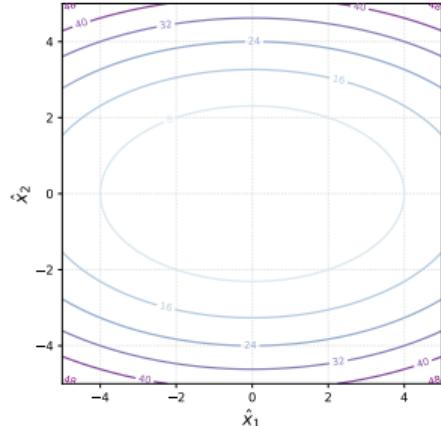
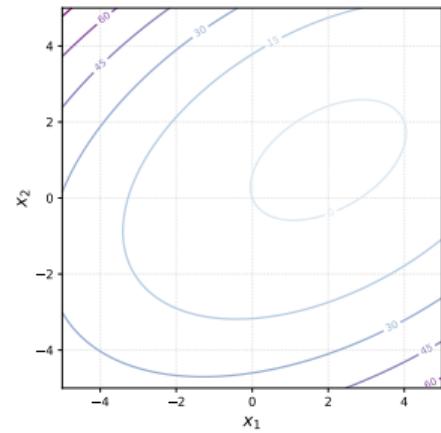
Сдвиг координат

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 влияет 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 $x = Q\hat{x} + x^*$.

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



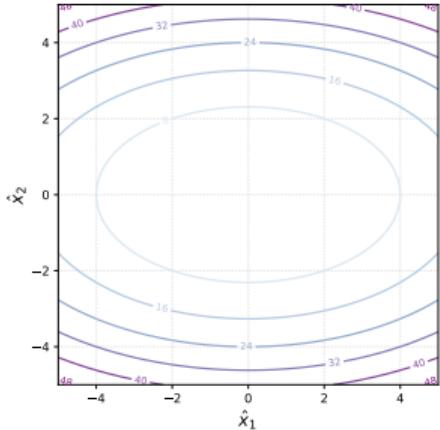
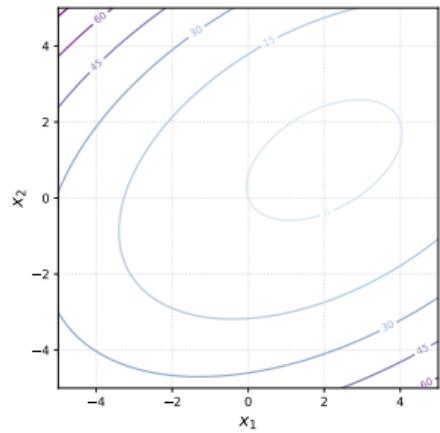
Сдвиг координат

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 влияет 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 $x = Q\hat{x} + x^*$.

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



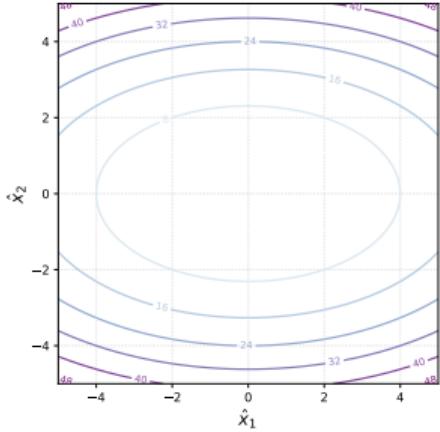
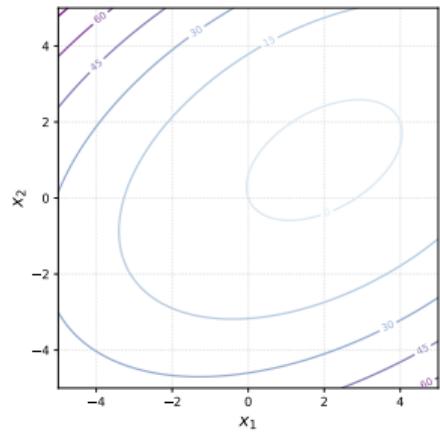
Сдвиг координат

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 влияет 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 $x = Q\hat{x} + x^*$.

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



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 влияет 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 $x = Q\hat{x} + x^*$.

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



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})

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

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

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

$$\begin{aligned}|1 - \alpha \mu| &< 1 \\-1 &< 1 - \alpha \mu < 1\end{aligned}$$

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})

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

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

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

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

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

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

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

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

$$\begin{array}{ll}|1 - \alpha\mu| < 1 & |1 - \alpha L| < 1 \\-1 < 1 - \alpha\mu < 1 & -1 < 1 - \alpha L < 1 \\\alpha < \frac{2}{\mu} & \alpha\mu > 0\end{array}$$

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

$$\begin{array}{ll}|1 - \alpha\mu| < 1 & |1 - \alpha L| < 1 \\-1 < 1 - \alpha\mu < 1 & -1 < 1 - \alpha L < 1 \\\alpha < \frac{2}{\mu} & \alpha < \frac{2}{L} \\ \alpha\mu > 0 & \alpha L > 0\end{array}$$

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

$$\begin{array}{ll}|1 - \alpha\mu| < 1 & |1 - \alpha L| < 1 \\-1 < 1 - \alpha\mu < 1 & -1 < 1 - \alpha L < 1 \\\alpha < \frac{2}{\mu} & \alpha < \frac{2}{L} \\ \alpha\mu > 0 & \alpha L > 0\end{array}$$

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})

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

$$\begin{aligned}x_{(i)}^{k+1} &= (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты} \\x_{(i)}^k &= (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha\end{aligned}$$

Выберем α , 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 $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$\begin{array}{ll} |1 - \alpha \mu| < 1 & |1 - \alpha L| < 1 \\ -1 < 1 - \alpha \mu < 1 & -1 < 1 - \alpha L < 1 \\ \alpha < \frac{2}{\mu} \quad \alpha \mu > 0 & \alpha < \frac{2}{L} \quad \alpha L > 0 \end{array}$$

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})

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

$$\begin{aligned}x_{(i)}^{k+1} &= (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты} \\x_{(i)}^k &= (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha\end{aligned}$$

Выберем α , 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 $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$\begin{array}{ll} |1 - \alpha \mu| < 1 & |1 - \alpha L| < 1 \\ -1 < 1 - \alpha \mu < 1 & -1 < 1 - \alpha L < 1 \\ \alpha < \frac{2}{\mu} & \alpha < \frac{2}{L} \\ \alpha \mu > 0 & \alpha L > 0 \end{array}$$

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})

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

$$\begin{aligned}x_{(i)}^{k+1} &= (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты} \\x_{(i)}^k &= (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha\end{aligned}$$

Выберем α , минимизирующий худший знаменатель прогрессии

$$\begin{aligned}\rho^* &= \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}| \\&= \min_{\alpha} \max \{|1 - \alpha \mu|, |1 - \alpha L|\}\end{aligned}$$

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

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

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

$$\begin{array}{ll}|1 - \alpha \mu| < 1 & |1 - \alpha L| < 1 \\-1 < 1 - \alpha \mu < 1 & -1 < 1 - \alpha L < 1 \\\alpha < \frac{2}{\mu} & \alpha < \frac{2}{L} \\ \alpha \mu > 0 & \alpha L > 0\end{array}$$

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})

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

$$\begin{aligned}x_{(i)}^{k+1} &= (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты} \\x_{(i)}^k &= (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha\end{aligned}$$

Выберем α , минимизирующий худший знаменатель прогрессии

$$\begin{aligned}\rho^* &= \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}| \\&= \min_{\alpha} \max \{|1 - \alpha \mu|, |1 - \alpha L|\} \\&\alpha^* : \quad 1 - \alpha^* \mu = \alpha^* L - 1\end{aligned}$$

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

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

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

$$\begin{array}{ll}|1 - \alpha \mu| < 1 & |1 - \alpha L| < 1 \\-1 < 1 - \alpha \mu < 1 & -1 < 1 - \alpha L < 1 \\\alpha < \frac{2}{\mu} & \alpha < \frac{2}{L} \\ \alpha \mu > 0 & \alpha L > 0\end{array}$$

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

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

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

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

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

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

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

Выберем α , минимизирующий худший знаменатель прогрессии

$$\begin{aligned}\rho^* &= \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}| \\&= \min_{\alpha} \max \{|1 - \alpha \mu|, |1 - \alpha L|\} \\&\alpha^* : \quad 1 - \alpha^* \mu = \alpha^* L - 1\end{aligned}$$

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

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})

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

$$\begin{aligned}x_{(i)}^{k+1} &= (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты} \\x_{(i)}^k &= (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha\end{aligned}$$

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

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

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

$$\begin{aligned}|1 - \alpha \mu| &< 1 \\-1 < 1 - \alpha \mu &< 1 \\\alpha < \frac{2}{\mu} \quad \alpha \mu &> 0\end{aligned}$$

Выберем α , минимизирующий худший знаменатель прогрессии

$$\begin{aligned}\rho^* &= \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}| \\&= \min_{\alpha} \max \{|1 - \alpha \mu|, |1 - \alpha L|\} \\&\alpha^* : \quad 1 - \alpha^* \mu = \alpha^* L - 1 \\&\alpha^* = \frac{2}{\mu + L} \quad \rho^* = \frac{L - \mu}{L + \mu}\end{aligned}$$

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})

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

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

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

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

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

Выберем α , минимизирующий худший знаменатель прогрессии

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

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

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

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

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

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

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

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

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

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

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

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

Выберем α , минимизирующий худший знаменатель прогрессии

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

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

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

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

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

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

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})

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

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

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

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

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

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

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

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

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

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

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

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

Выберем α , минимизирующий худший знаменатель прогрессии

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

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

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

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

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

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

Анализ сходимости

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

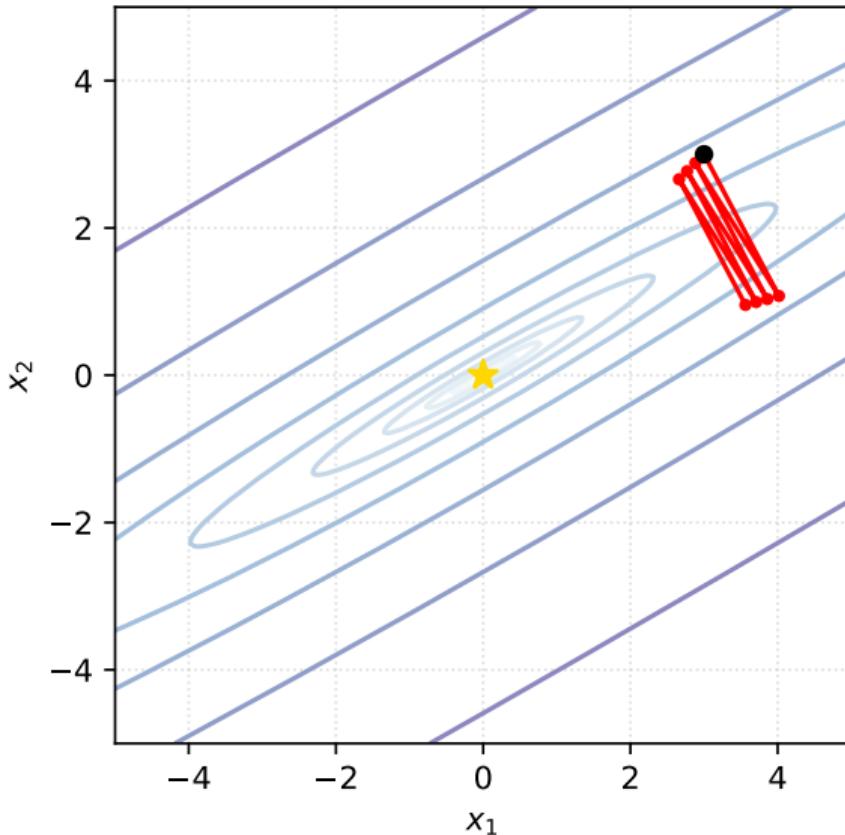
μ	ρ	Итераций до уменьшения ошибки по	Итераций до уменьшения ошибки по
		аргументу в 10 раз	функции в 10 раз
1.1	0.05	1	1
2	0.33	3	2
5	0.67	6	3
10	0.82	12	6
50	0.96	58	29
100	0.98	116	58
500	0.996	576	288
1000	0.998	1152	576

Число обусловленности κ

$$\kappa = 1.0$$



$$\kappa = 100.0$$



PL-функции. Линейная сходимость градиентного спуска без выпуклости

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

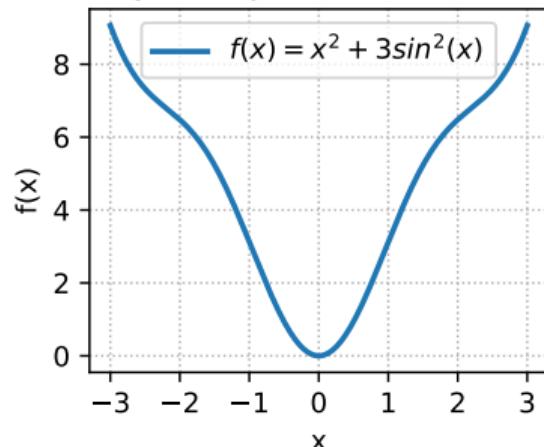
$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x$$

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



PL-функции. Линейная сходимость градиентного спуска без выпуклости

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

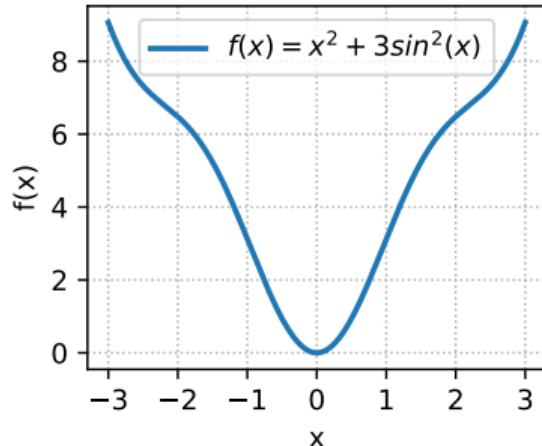
$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x$$

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

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

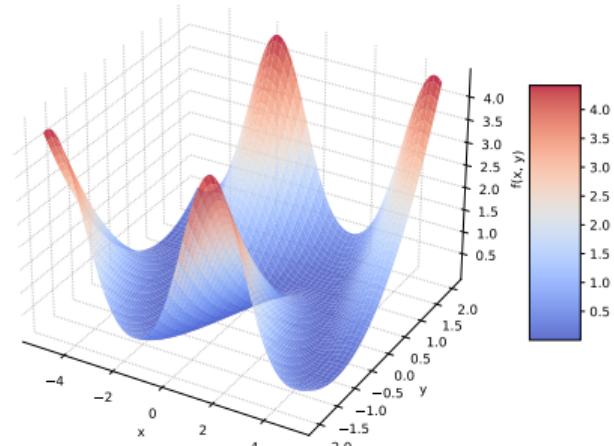
$$f(x) = x^2 + 3 \sin^2(x)$$

Function, that satisfies
Polyak-Lojasiewicz condition



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

Non-convex PL function



Theorem

Рассмотрим задачу

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

и предположим, что f является PL-функцией с константой μ 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 α , удовлетворяющим $0 < \alpha \leq \frac{1}{L}$. Пусть $f^* = \min_{x \in \mathbb{R}^d} f(x)$. Тогда:

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

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.

Любая μ -сильно выпуклая дифференцируемая функция является PL-функцией

Theorem

Если функция $f(x)$ дифференцируема и μ -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 μ -сильно выпуклая дифференцируемая функция является PL-функцией

Theorem

Если функция $f(x)$ дифференцируема и μ -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 μ -сильно выпуклая дифференцируемая функция является PL-функцией

■ Theorem

Если функция $f(x)$ дифференцируема и μ -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 $y = x^*$:

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

Любая μ -сильно выпуклая дифференцируемая функция является PL-функцией

■ Theorem

Если функция $f(x)$ дифференцируема и μ -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 $y = x^*$:

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



Любая μ -сильно выпуклая дифференцируемая функция является PL-функцией

■ Theorem

Если функция $f(x)$ дифференцируема и μ -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 $y = x^*$:

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



Любая μ -сильно выпуклая дифференцируемая функция является PL-функцией

Theorem

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

Доказательство

По критерию сильной выпуклости первого порядка:

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

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

Положим $y = x^*$:

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

Любая μ -сильно выпуклая дифференцируемая функция является PL-функцией

■ Theorem

Если функция $f(x)$ дифференцируема и μ -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 $y = x^*$:

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

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

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 μ -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 μ -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 μ -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 μ -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.

Выпуклый гладкий случай

Выпуклый гладкий случай

Theorem

Рассмотрим задачу

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

и предположим, что f 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 α , удовлетворяющим $0 < \alpha \leq \frac{1}{L}$. Пусть $f^* = \min_{x \in \mathbb{R}^d} f(x)$. Тогда для всех $x^* \in \operatorname{argmin} f$ и всех $k \in \mathbb{N}$ справедливо:

$$f(x^k) - f^* \leq \frac{\|x^0 - x^*\|^2}{2\alpha k}.$$

Анализ сходимости

- Как и раньше, сначала используем гладкость:

(1)

Анализ сходимости

- Как и раньше, сначала используем гладкость:

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

(1)

Анализ сходимости

- Как и раньше, сначала используем гладкость:

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

Анализ сходимости

- Как и раньше, сначала используем гладкость:

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

Анализ сходимости

- Как и раньше, сначала используем гладкость:

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

Анализ сходимости

- Как и раньше, сначала используем гладкость:

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

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{2L} \|\nabla f(x^k)\|^2 \quad \text{если } \alpha = \frac{1}{L}$$

Анализ сходимости

- Как и раньше, сначала используем гладкость:

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

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{2L} \|\nabla f(x^k)\|^2 \quad \text{если } \alpha = \frac{1}{L}$$

Анализ сходимости

- Как и раньше, сначала используем гладкость:

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

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{2L} \|\nabla f(x^k)\|^2 \quad \text{если } \alpha = \frac{1}{L}$$

Обычно для сходящегося градиентного спуска чем больше допустимый шаг, тем быстрее сходимость, поэтому часто берут $\alpha = \frac{1}{L}$.

- После этого используем выпуклость:

(2)

Анализ сходимости

- Как и раньше, сначала используем гладкость:

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

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{2L} \|\nabla f(x^k)\|^2 \quad \text{если } \alpha = \frac{1}{L}$$

Обычно для сходящегося градиентного спуска чем больше допустимый шаг, тем быстрее сходимость, поэтому часто берут $\alpha = \frac{1}{L}$.

- После этого используем выпуклость:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \tag{2}$$

Анализ сходимости

- Как и раньше, сначала используем гладкость:

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

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{2L} \|\nabla f(x^k)\|^2 \quad \text{если } \alpha = \frac{1}{L}$$

Обычно для сходящегося градиентного спуска чем больше допустимый шаг, тем быстрее сходимость, поэтому часто берут $\alpha = \frac{1}{L}$.

- После этого используем выпуклость:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \text{где } y = x^*, x = x^k \tag{2}$$

Анализ сходимости

- Как и раньше, сначала используем гладкость:

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

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{2L} \|\nabla f(x^k)\|^2 \quad \text{если } \alpha = \frac{1}{L}$$

Обычно для сходящегося градиентного спуска чем больше допустимый шаг, тем быстрее сходимость, поэтому часто берут $\alpha = \frac{1}{L}$.

- После этого используем выпуклость:

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle \quad \text{где } y = x^*, x = x^k \\ f(x^k) - f^* &\leq \langle \nabla f(x^k), x^k - x^* \rangle \end{aligned} \tag{2}$$

Анализ сходимости

- Теперь подставляем (2) в (1):

Анализ сходимости

- Теперь подставляем (2) в (1):

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

Анализ сходимости

- Теперь подставляем (2) в (1):

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

Анализ сходимости

- Теперь подставляем (2) в (1):

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

Анализ сходимости

- Теперь подставляем (2) в (1):

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

Пусть $a = x^k - x^*$ и $b = x^k - x^* - \alpha \nabla f(x^k)$.

Анализ сходимости

- Теперь подставляем (2) в (1):

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

Пусть $a = x^k - x^*$ и $b = x^k - x^* - \alpha \nabla f(x^k)$. Тогда $a - b = \alpha \nabla f(x^k)$ и $a + b = 2(x^k - x^* - \frac{\alpha}{2} \nabla f(x^k))$.

Анализ сходимости

- Теперь подставляем (2) в (1):

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

Пусть $a = x^k - x^*$ и $b = x^k - x^* - \alpha \nabla f(x^k)$. Тогда $a - b = \alpha \nabla f(x^k)$ и $a + b = 2(x^k - x^* - \frac{\alpha}{2} \nabla f(x^k))$.

$$f(x^{k+1}) \leq f^* + \frac{1}{2\alpha} [\|x^k - x^*\|_2^2 - \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2]$$

Анализ сходимости

- Теперь подставляем (2) в (1):

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

Пусть $a = x^k - x^*$ и $b = x^k - x^* - \alpha \nabla f(x^k)$. Тогда $a - b = \alpha \nabla f(x^k)$ и $a + b = 2(x^k - x^* - \frac{\alpha}{2} \nabla f(x^k))$.

$$\begin{aligned} f(x^{k+1}) &\leq f^* + \frac{1}{2\alpha} [\|x^k - x^*\|_2^2 - \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2] \\ &\leq f^* + \frac{1}{2\alpha} [\|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2] \end{aligned}$$

Анализ сходимости

- Теперь подставляем (2) в (1):

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

Пусть $a = x^k - x^*$ и $b = x^k - x^* - \alpha \nabla f(x^k)$. Тогда $a - b = \alpha \nabla f(x^k)$ и $a + b = 2(x^k - x^* - \frac{\alpha}{2} \nabla f(x^k))$.

$$\begin{aligned} f(x^{k+1}) &\leq f^* + \frac{1}{2\alpha} [\|x^k - x^*\|_2^2 - \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2] \\ &\leq f^* + \frac{1}{2\alpha} [\|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2] \\ 2\alpha (f(x^{k+1}) - f^*) &\leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2 \end{aligned}$$

Анализ сходимости

- Теперь подставляем (2) в (1):

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

Пусть $a = x^k - x^*$ и $b = x^k - x^* - \alpha \nabla f(x^k)$. Тогда $a - b = \alpha \nabla f(x^k)$ и $a + b = 2(x^k - x^* - \frac{\alpha}{2} \nabla f(x^k))$.

$$\begin{aligned} f(x^{k+1}) &\leq f^* + \frac{1}{2\alpha} [\|x^k - x^*\|_2^2 - \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2] \\ &\leq f^* + \frac{1}{2\alpha} [\|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2] \\ 2\alpha (f(x^{k+1}) - f^*) &\leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2 \end{aligned}$$

- Просуммируем по $i = 0, \dots, k-1$. Большинство слагаемых обнуляется из-за телескопической суммы:

(3)

Анализ сходимости

- Теперь подставляем (2) в (1):

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

Пусть $a = x^k - x^*$ и $b = x^k - x^* - \alpha \nabla f(x^k)$. Тогда $a - b = \alpha \nabla f(x^k)$ и $a + b = 2(x^k - x^* - \frac{\alpha}{2} \nabla f(x^k))$.

$$\begin{aligned} f(x^{k+1}) &\leq f^* + \frac{1}{2\alpha} [\|x^k - x^*\|_2^2 - \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2] \\ &\leq f^* + \frac{1}{2\alpha} [\|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2] \\ 2\alpha (f(x^{k+1}) - f^*) &\leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2 \end{aligned}$$

- Просуммируем по $i = 0, \dots, k-1$. Большинство слагаемых обнуляется из-за телескопической суммы:

$$2\alpha \sum_{i=0}^{k-1} (f(x^{i+1}) - f^*) \leq \|x^0 - x^*\|_2^2 - \|x^k - x^*\|_2^2 \quad (3)$$

Анализ сходимости

- Теперь подставляем (2) в (1):

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

Пусть $a = x^k - x^*$ и $b = x^k - x^* - \alpha \nabla f(x^k)$. Тогда $a - b = \alpha \nabla f(x^k)$ и $a + b = 2(x^k - x^* - \frac{\alpha}{2} \nabla f(x^k))$.

$$\begin{aligned} f(x^{k+1}) &\leq f^* + \frac{1}{2\alpha} [\|x^k - x^*\|_2^2 - \|x^k - x^* - \alpha \nabla f(x^k)\|_2^2] \\ &\leq f^* + \frac{1}{2\alpha} [\|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2] \\ 2\alpha (f(x^{k+1}) - f^*) &\leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2 \end{aligned}$$

- Просуммируем по $i = 0, \dots, k-1$. Большинство слагаемых обнуляется из-за телескопической суммы:

$$2\alpha \sum_{i=0}^{k-1} (f(x^{i+1}) - f^*) \leq \|x^0 - x^*\|_2^2 - \|x^k - x^*\|_2^2 \leq \|x^0 - x^*\|_2^2 \quad (3)$$

Анализ сходимости

- Поскольку на каждой итерации $f(x^{i+1}) \leq f(x^i)$, то

$$kf(x^k) \leq \sum_{i=0}^{k-1} f(x^{i+1})$$

Анализ сходимости

- Поскольку на каждой итерации $f(x^{i+1}) \leq f(x^i)$, то

$$kf(x^k) \leq \sum_{i=0}^{k-1} f(x^{i+1})$$

- Теперь подставим это в (3):

Анализ сходимости

- Поскольку на каждой итерации $f(x^{i+1}) \leq f(x^i)$, то

$$kf(x^k) \leq \sum_{i=0}^{k-1} f(x^{i+1})$$

- Теперь подставим это в (3):

$$2\alpha kf(x^k) - 2\alpha kf^* \leq 2\alpha \sum_{i=0}^{k-1} (f(x^{i+1}) - f^*) \leq \|x^0 - x^*\|_2^2$$

Анализ сходимости

- Поскольку на каждой итерации $f(x^{i+1}) \leq f(x^i)$, то

$$kf(x^k) \leq \sum_{i=0}^{k-1} f(x^{i+1})$$

- Теперь подставим это в (3):

$$2\alpha kf(x^k) - 2\alpha kf^* \leq 2\alpha \sum_{i=0}^{k-1} (f(x^{i+1}) - f^*) \leq \|x^0 - x^*\|_2^2$$

$$f(x^k) - f^* \leq \frac{\|x^0 - x^*\|_2^2}{2\alpha k}$$

Анализ сходимости

- Поскольку на каждой итерации $f(x^{i+1}) \leq f(x^i)$, то

$$kf(x^k) \leq \sum_{i=0}^{k-1} f(x^{i+1})$$

- Теперь подставим это в (3):

$$2\alpha kf(x^k) - 2\alpha kf^* \leq 2\alpha \sum_{i=0}^{k-1} (f(x^{i+1}) - f^*) \leq \|x^0 - x^*\|_2^2$$

$$f(x^k) - f^* \leq \frac{\|x^0 - x^*\|_2^2}{2\alpha k} \leq \frac{L\|x^0 - x^*\|_2^2}{2k}$$

Итог

Градиентный спуск:

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

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

гладкий (не выпуклый)

гладкий и выпуклый

гладкий и сильно выпуклый (или PL)

$$\|\nabla f(x^k)\|^2 \sim \mathcal{O}\left(\frac{1}{k}\right)$$

$$f(x^k) - f^* \sim \mathcal{O}\left(\frac{1}{k}\right)$$

$$\|x^k - x^*\|^2 \sim \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$$

$$k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon}\right)$$

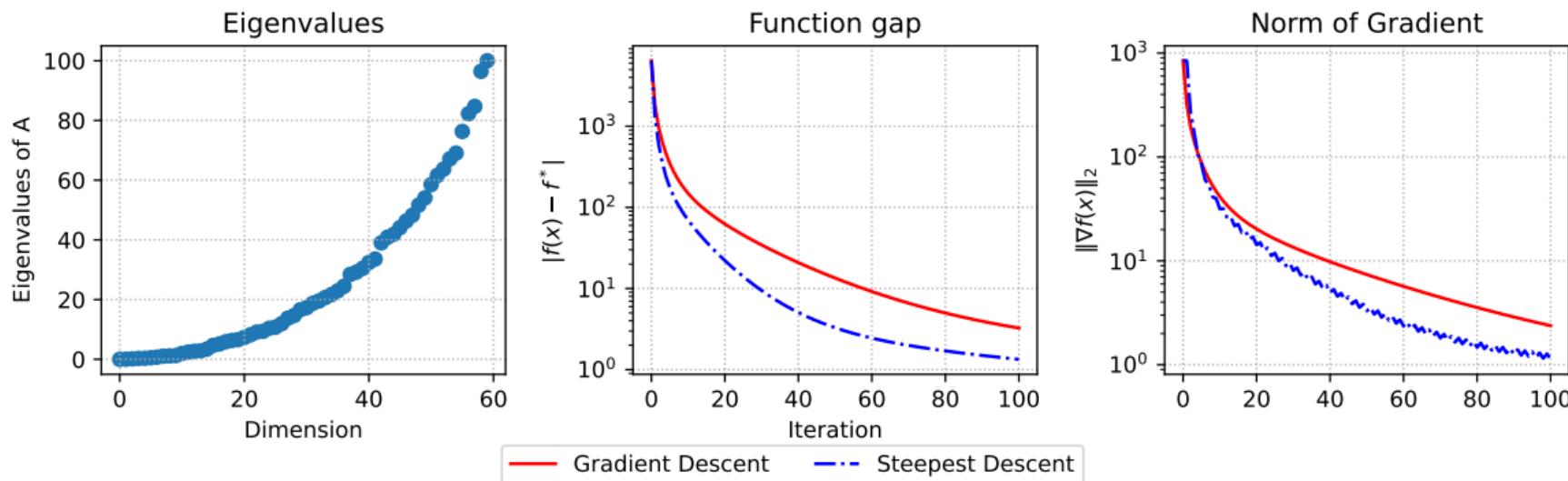
$$k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon}\right)$$

$$k_\varepsilon \sim \mathcal{O}\left(\nu \log \frac{1}{\varepsilon}\right)$$

Численные эксперименты

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

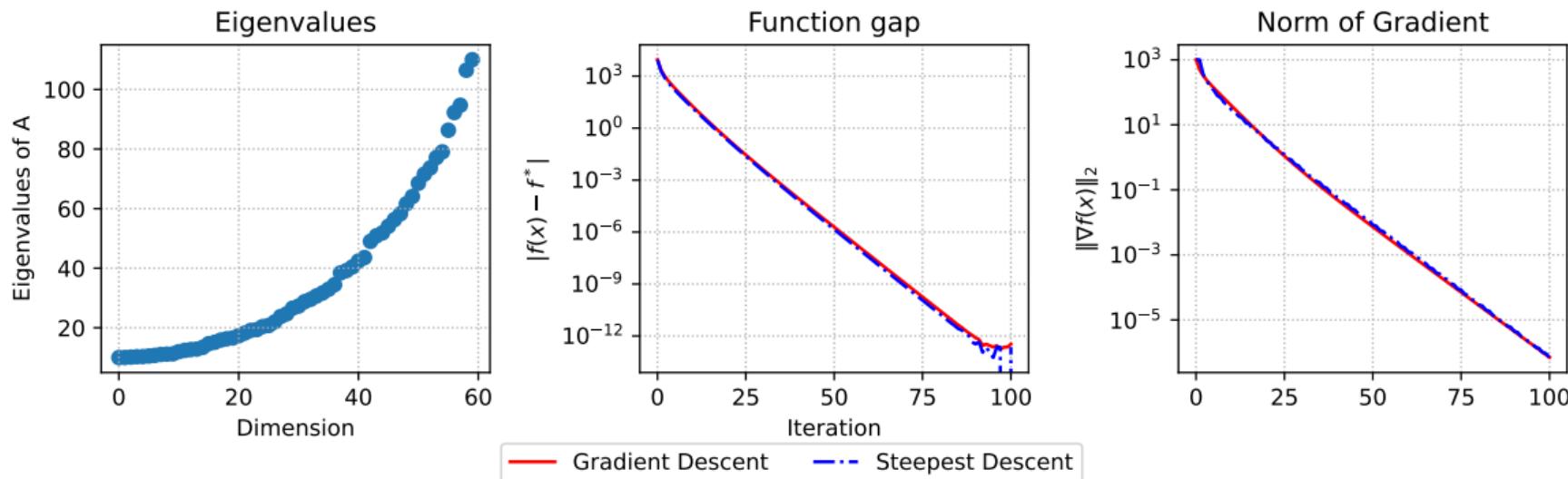
Convex quadratics. $n=60$, random matrix.



Численные эксперименты

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

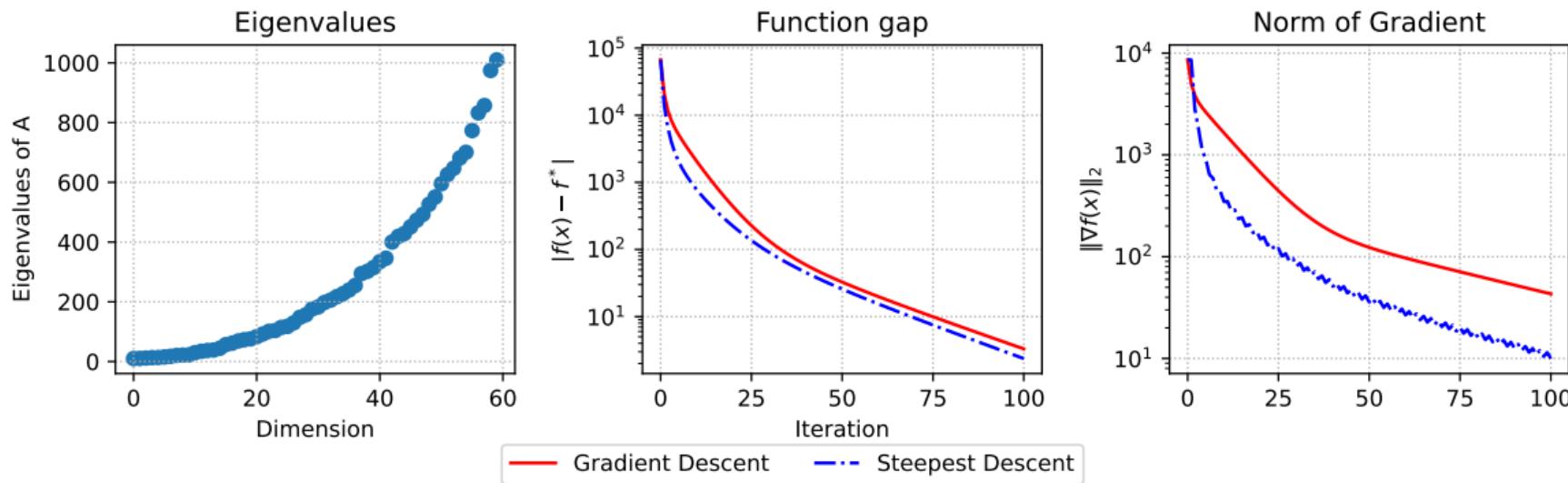
Strongly convex quadratics. n=60, random matrix.



Численные эксперименты

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

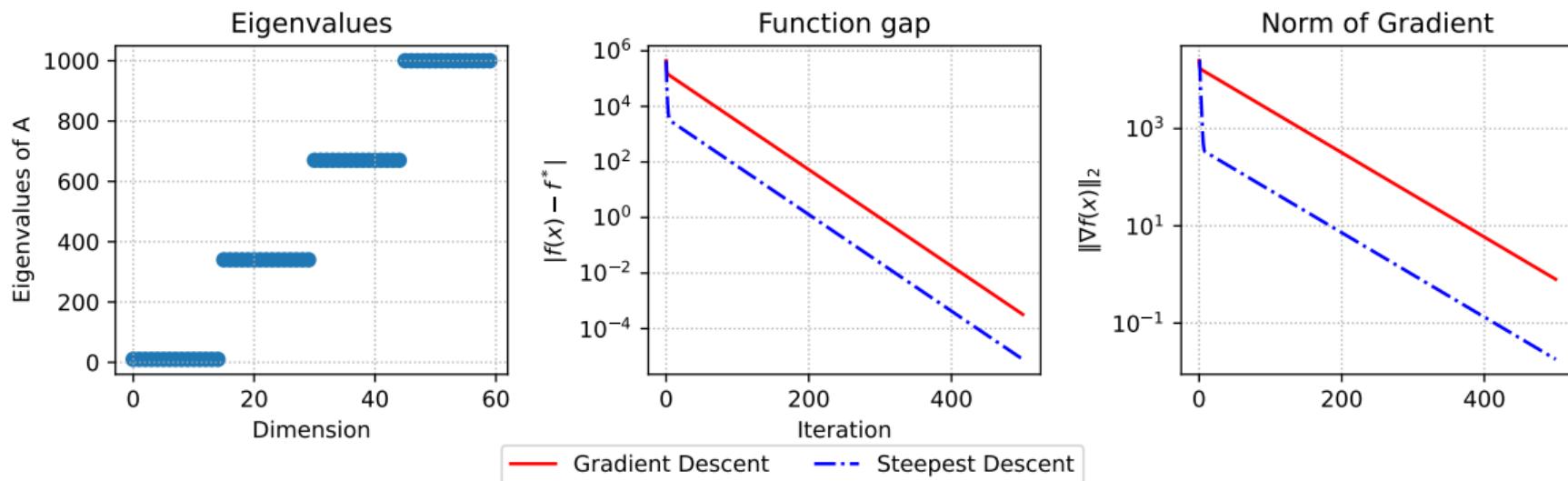
Strongly convex quadratics. n=60, random matrix.



Численные эксперименты

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

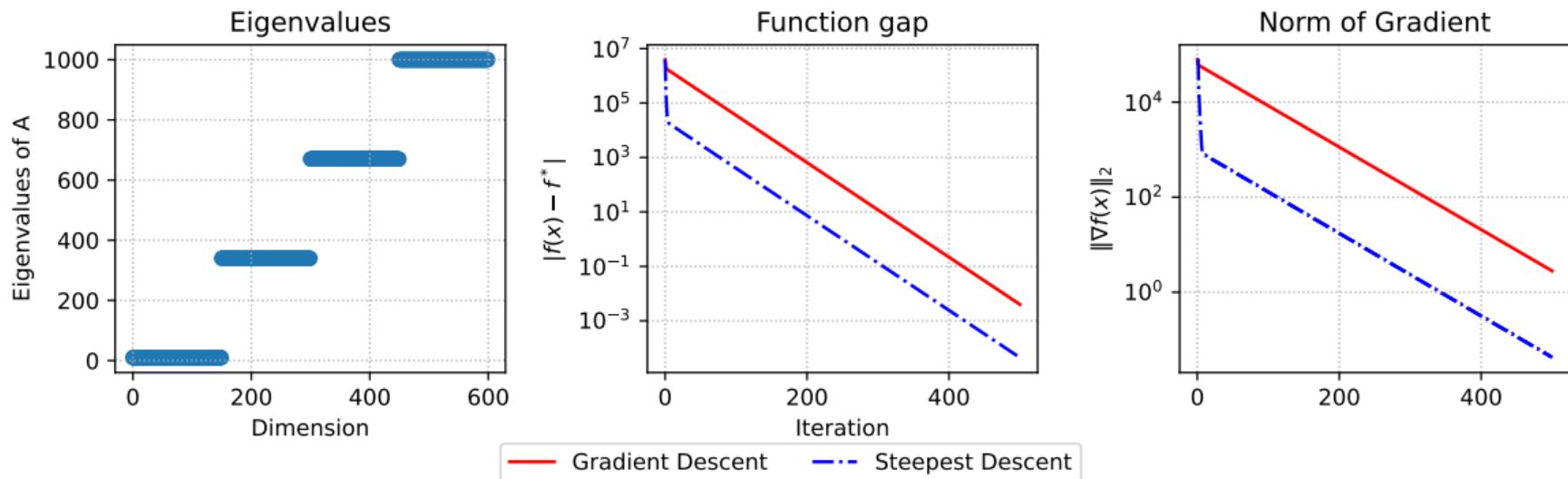
Strongly convex quadratics. $n=60$, clustered matrix.



Численные эксперименты

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

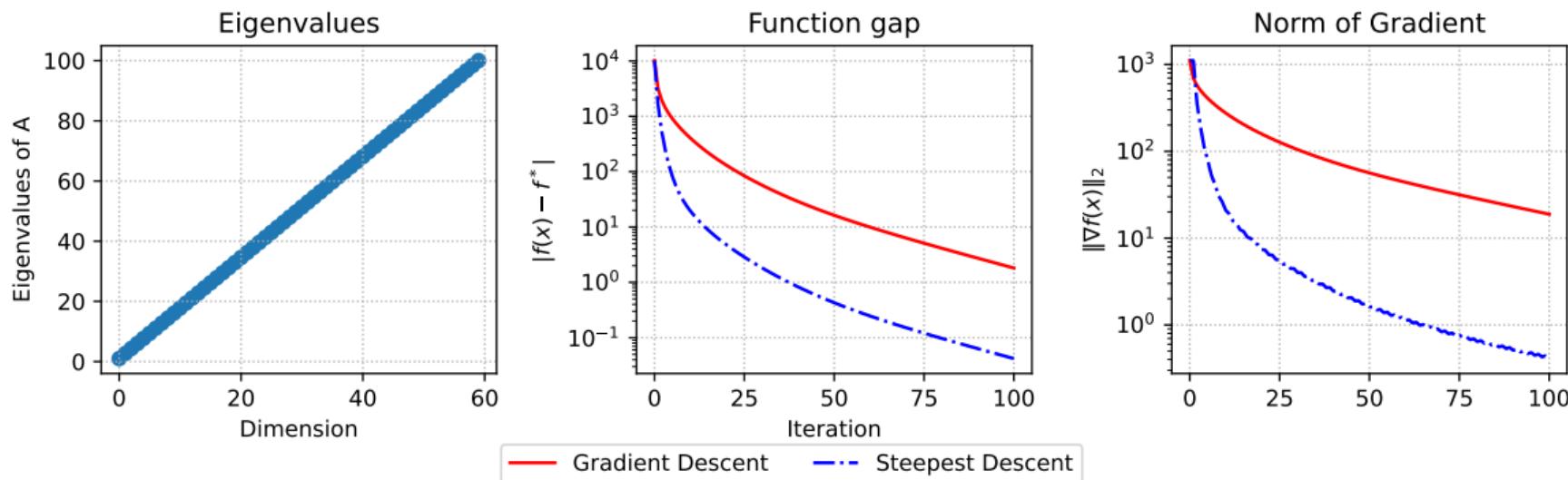
Strongly convex quadratics. $n=600$, clustered matrix.



Численные эксперименты

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

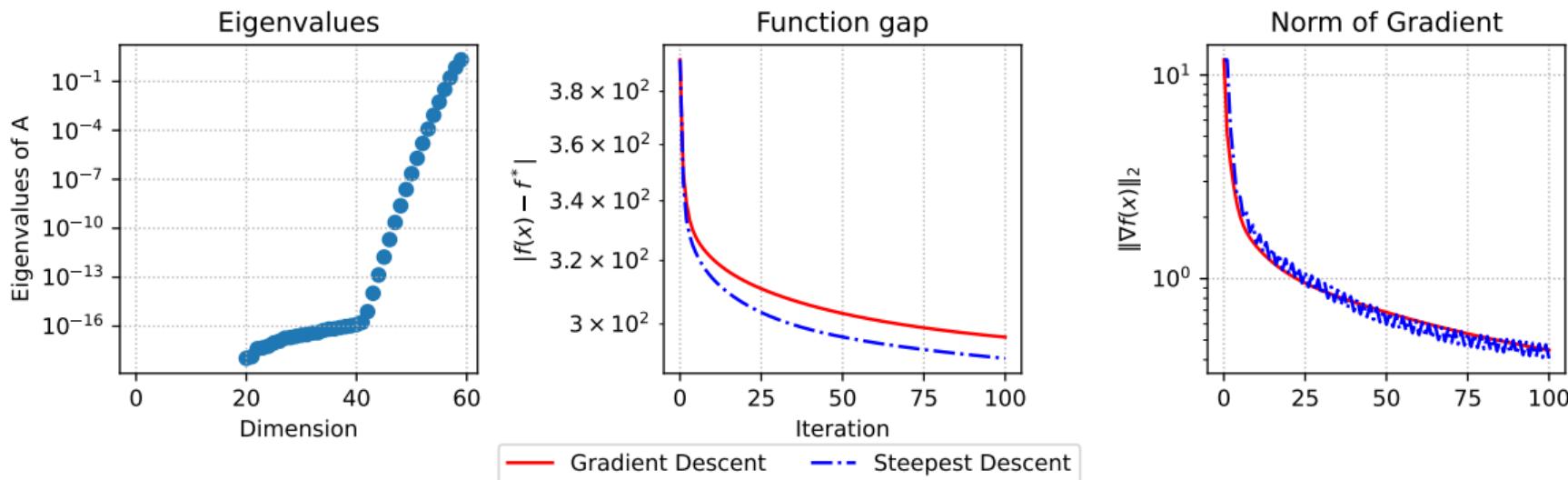
Strongly convex quadratics. $n=60$, uniform spectrum matrix.



Численные эксперименты

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

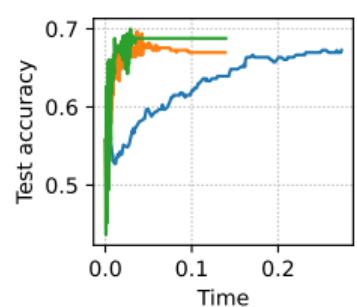
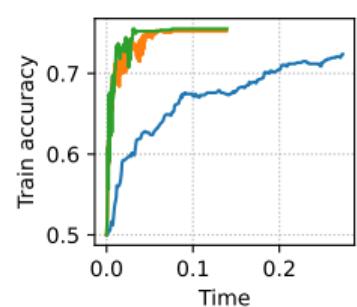
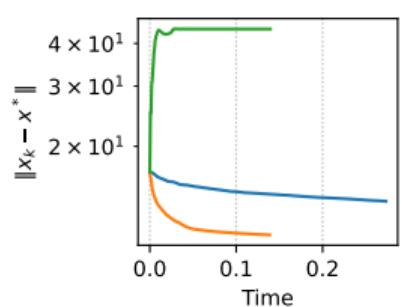
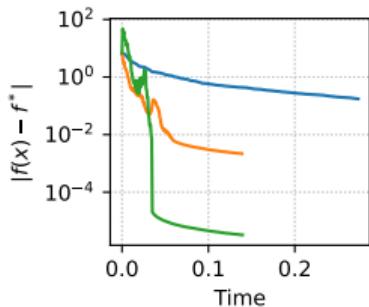
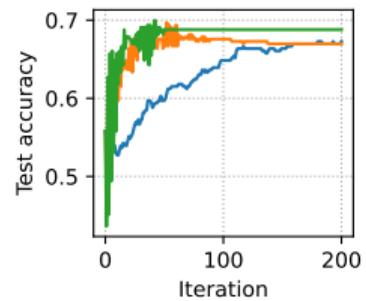
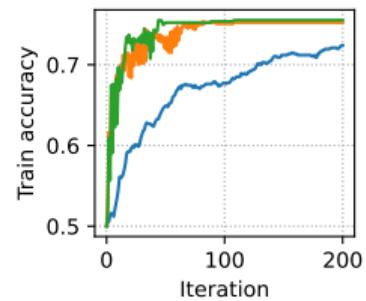
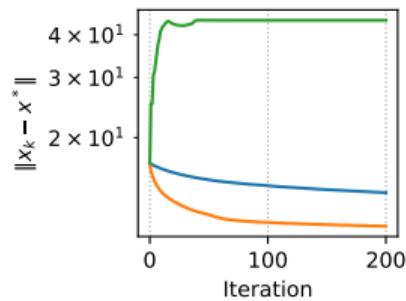
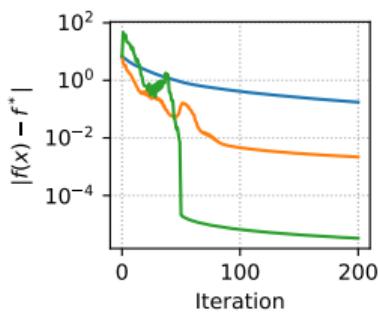
Strongly convex quadratics. n=60, Hilbert matrix.



Численные эксперименты

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

Convex binary logistic regression. mu=0.

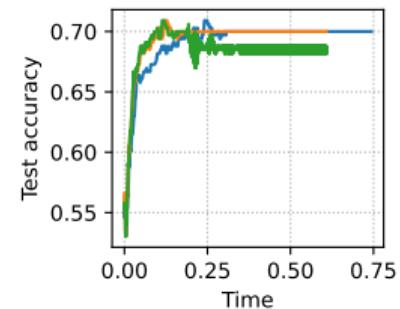
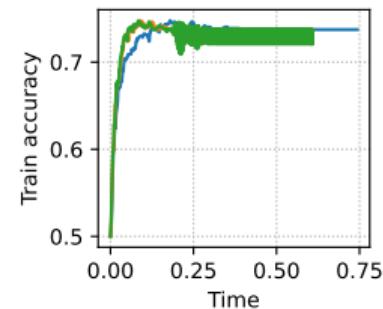
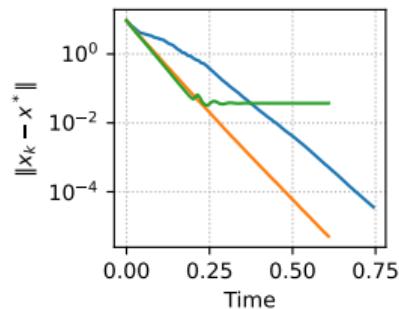
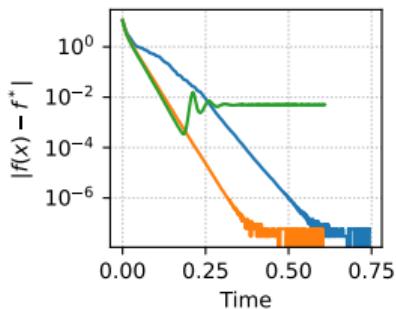
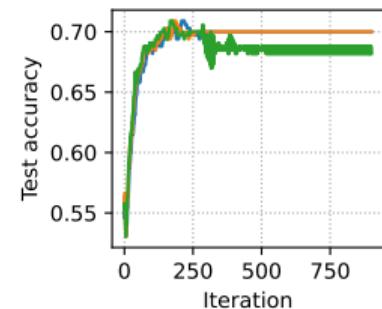
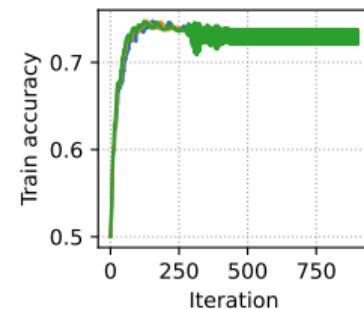
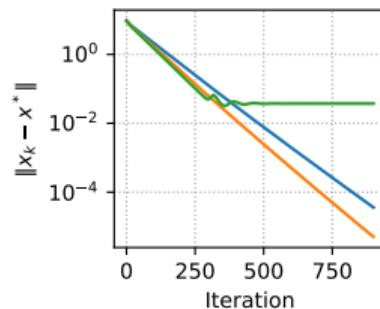
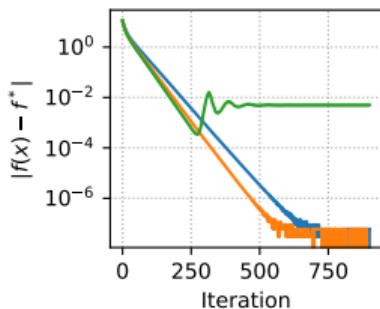


— GD 0.07 — GD 0.9 — GD 10.0

Численные эксперименты

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

Strongly convex binary logistic regression. mu=0.1.

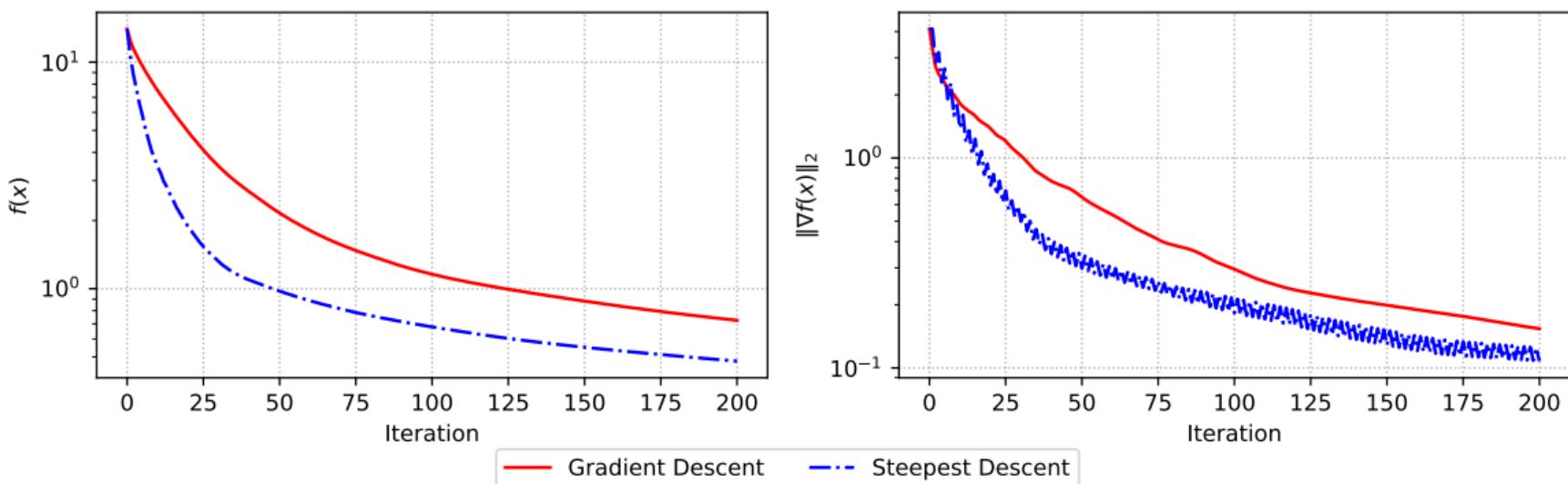


— GD 0.12 — GD 0.14 — GD 0.15

Численные эксперименты

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

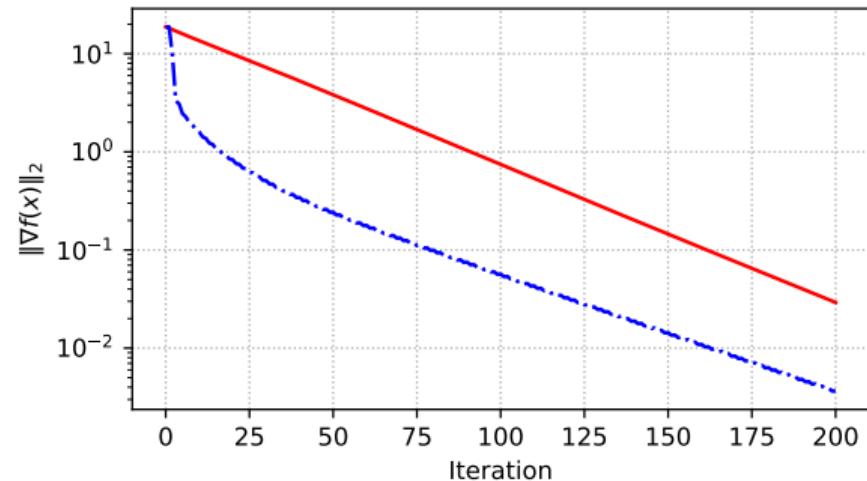
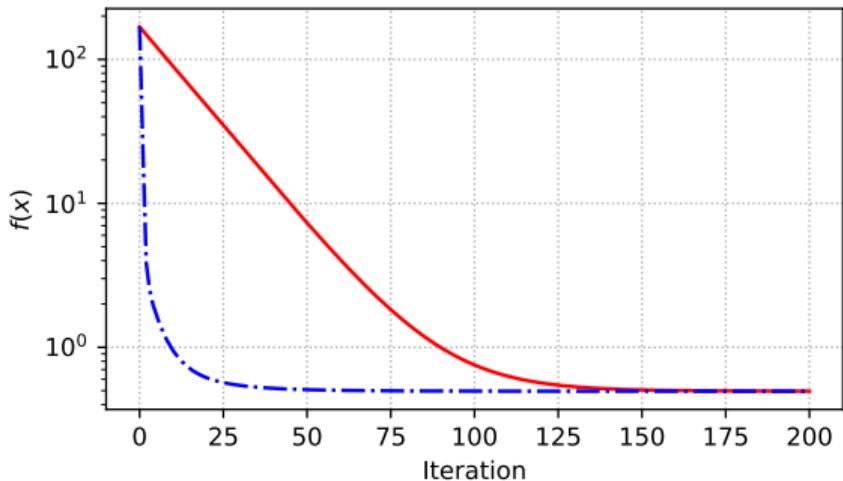
Regularized binary logistic regression. n=300. m=1000. $\mu=0$



Численные эксперименты

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

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



Gradient Descent Steepest Descent