



Проксимальный градиентный метод.

Даня Меркулов

Оптимизация для всех! ЦУ

## Субградиентный метод

$\ell_1$  induces sparsity

$\ell_2$  regularization.  $\|Xw - y\|_2^2 \rightarrow \min_{\|w\|_2 \leq 1}$



$\ell_1$  regularization.  $\|Xw - y\|_2^2 \rightarrow \min_{\|w\|_1 \leq 1}$



@fminxyz

# Субградиентный метод

Субградиентный метод:

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

$$x_{k+1} = x_k - \alpha_k g_k, \quad g_k \in \partial f(x_k)$$

# Субградиентный метод

Субградиентный метод:

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

$$x_{k+1} = x_k - \alpha_k g_k, \quad g_k \in \partial f(x_k)$$

---

выпуклый (негладкий)

$$f(x_k) - f^* \sim \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$$
$$k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$$

---

сильно выпуклый (негладкий)

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

---

# Субградиентный метод

Субградиентный метод:

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

$$x_{k+1} = x_k - \alpha_k g_k, \quad g_k \in \partial f(x_k)$$

выпуклый (негладкий)

$$f(x_k) - f^* \sim \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$$
$$k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$$

сильно выпуклый (негладкий)

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

## i Theorem

Предположим, что  $f$  является  $G$ -липшицевой и выпуклой, тогда субградиентный метод сходится как:

$$f(\bar{x}) - f^* \leq \frac{GR}{\sqrt{k}},$$

где

- $\alpha = \frac{R}{G\sqrt{k}}$

# Субградиентный метод

Субградиентный метод:

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

$$x_{k+1} = x_k - \alpha_k g_k, \quad g_k \in \partial f(x_k)$$

выпуклый (негладкий)

$$f(x_k) - f^* \sim \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$$
$$k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$$

сильно выпуклый (негладкий)

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

## i Theorem

Предположим, что  $f$  является  $G$ -липшицевой и выпуклой, тогда субградиентный метод сходится как:

$$f(\bar{x}) - f^* \leq \frac{GR}{\sqrt{k}},$$

где

- $\alpha = \frac{R}{G\sqrt{k}}$
- $R = \|x_0 - x^*\|$

# Субградиентный метод

Субградиентный метод:

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

$$x_{k+1} = x_k - \alpha_k g_k, \quad g_k \in \partial f(x_k)$$

выпуклый (негладкий)

$$f(x_k) - f^* \sim \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$$
$$k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$$

сильно выпуклый (негладкий)

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

## i Theorem

Предположим, что  $f$  является  $G$ -липшицевой и выпуклой, тогда субградиентный метод сходится как:

$$f(\bar{x}) - f^* \leq \frac{GR}{\sqrt{k}},$$

где

- $\alpha = \frac{R}{G\sqrt{k}}$
- $R = \|x_0 - x^*\|$
- $\bar{x} = \frac{1}{k} \sum_{i=0}^{k-1} x_i$



# Нижние оценки для негладких выпуклых задач

выпуклый (негладкий)

$$f(x_k) - f^* \sim \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$$
$$k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$$

сильно выпуклый (негладкий)

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

## Нижние оценки для негладких выпуклых задач

| выпуклый (негладкий)  | сильно выпуклый (негладкий)  |
|---|--|
| $f(x_k) - f^* \sim \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$ | $f(x_k) - f^* \sim \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon}\right)$ |

- Субградиентный метод является оптимальным для задач выше.

## Нижние оценки для негладких выпуклых задач

| выпуклый (негладкий)  | сильно выпуклый (негладкий)  |
|---|--|
| $f(x_k) - f^* \sim \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$ | $f(x_k) - f^* \sim \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon}\right)$ |

- Субградиентный метод является оптимальным для задач выше.
- Можно использовать метод зеркального спуска (обобщение метода субградиента на, возможно, неевклидову метрику) с той же скоростью сходимости, чтобы лучше согласовать геометрию задачи.

## Нижние оценки для негладких выпуклых задач

| выпуклый (негладкий)  | сильно выпуклый (негладкий)  |
|---|--|
| $f(x_k) - f^* \sim \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$ | $f(x_k) - f^* \sim \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon}\right)$ |

- Субградиентный метод является оптимальным для задач выше.
- Можно использовать метод зеркального спуска (обобщение метода субградиента на, возможно, неевклидову метрику) с той же скоростью сходимости, чтобы лучше согласовать геометрию задачи.
- Однако, мы можем достичь стандартной скорости градиентного спуска  $\mathcal{O}\left(\frac{1}{k}\right)$  (и даже ускоренной версии  $\mathcal{O}\left(\frac{1}{k^2}\right)$ ), если мы будем использовать структуру задачи.

## Проксимальный оператор

# Интуиция проксимального отображения

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x)$$

Явный метод Эйлера:

# Интуиция проксимального отображения

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x)$$

Явный метод Эйлера:

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

Приводит к обычному методу градиентного спуска.

# Интуиция проксимального отображения

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x)$$

Явный метод Эйлера:

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

Неявный метод Эйлера:

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

Приводит к обычному методу градиентного спуска.



# Интуиция проксимального отображения

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x)$$

Явный метод Эйлера:

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

Приводит к обычному методу градиентного спуска.

Неявный метод Эйлера:

$$\begin{aligned} \frac{x_{k+1} - x_k}{\alpha} &= -\nabla f(x_{k+1}) \\ \frac{x_{k+1} - x_k}{\alpha} + \nabla f(x_{k+1}) &= 0 \end{aligned}$$

# Интуиция проксимального отображения

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x)$$

Явный метод Эйлера:

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

Приводит к обычному методу градиентного спуска.

Неявный метод Эйлера:

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

$$\frac{x_{k+1} - x_k}{\alpha} + \nabla f(x_{k+1}) = 0$$

$$\frac{x - x_k}{\alpha} + \nabla f(x) \Big|_{x=x_{k+1}} = 0$$

# Интуиция проксимального отображения

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x)$$

Явный метод Эйлера:

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

Приводит к обычному методу градиентного спуска.

Неявный метод Эйлера:

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

$$\frac{x_{k+1} - x_k}{\alpha} + \nabla f(x_{k+1}) = 0$$

$$\frac{x - x_k}{\alpha} + \nabla f(x) \Big|_{x=x_{k+1}} = 0$$

$$\nabla \left[ \frac{1}{2\alpha} \|x - x_k\|_2^2 + f(x) \right] \Big|_{x=x_{k+1}} = 0$$

# Интуиция проксимального отображения

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x)$$

Явный метод Эйлера:

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

Приводит к обычному методу градиентного спуска.

Неявный метод Эйлера:

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

$$\frac{x_{k+1} - x_k}{\alpha} + \nabla f(x_{k+1}) = 0$$

$$\frac{x - x_k}{\alpha} + \nabla f(x) \Big|_{x=x_{k+1}} = 0$$

$$\nabla \left[ \frac{1}{2\alpha} \|x - x_k\|_2^2 + f(x) \right] \Big|_{x=x_{k+1}} = 0$$

$$x_{k+1} = \arg \min_{x \in \mathbb{R}^n} \left[ f(x) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right]$$

# Интуиция проксимального отображения

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x)$$

Явный метод Эйлера:

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

Приводит к обычному методу градиентного спуска.

Неявный метод Эйлера:

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

$$\frac{x_{k+1} - x_k}{\alpha} + \nabla f(x_{k+1}) = 0$$

$$\frac{x - x_k}{\alpha} + \nabla f(x) \Big|_{x=x_{k+1}} = 0$$

$$\nabla \left[ \frac{1}{2\alpha} \|x - x_k\|_2^2 + f(x) \right] \Big|_{x=x_{k+1}} = 0$$

$$x_{k+1} = \arg \min_{x \in \mathbb{R}^n} \left[ f(x) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right]$$

# Интуиция проксимального отображения

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x)$$

Явный метод Эйлера:

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

Приводит к обычному методу градиентного спуска.

Неявный метод Эйлера:

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

$$\frac{x_{k+1} - x_k}{\alpha} + \nabla f(x_{k+1}) = 0$$

$$\frac{x - x_k}{\alpha} + \nabla f(x) \Big|_{x=x_{k+1}} = 0$$

$$\nabla \left[ \frac{1}{2\alpha} \|x - x_k\|_2^2 + f(x) \right] \Big|_{x=x_{k+1}} = 0$$

$$x_{k+1} = \arg \min_{x \in \mathbb{R}^n} \left[ f(x) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right]$$

! Проксимальный оператор

$$\text{prox}_{f,\alpha}(x_k) = \arg \min_{x \in \mathbb{R}^n} \left[ f(x) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right]$$

# Визуализация проксимального оператора

$$\text{Prox}_f(x) = \underset{x'}{\operatorname{argmin}} \frac{1}{2} \|x - x'\|^2 + f(x')$$

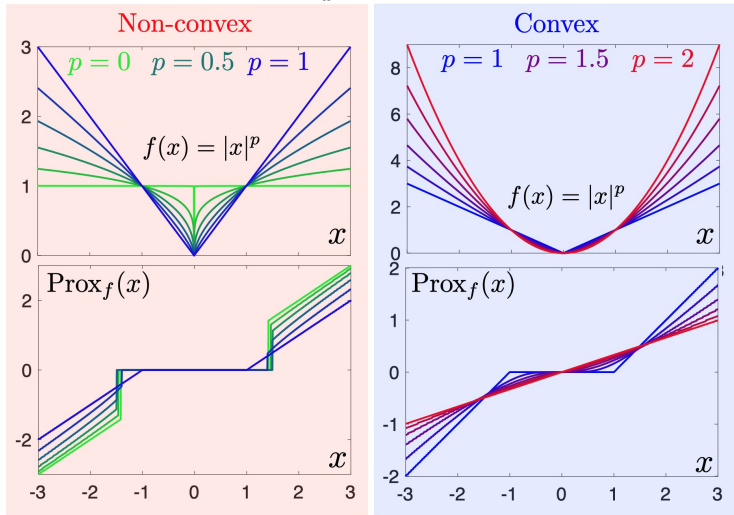


Рис. 1: Источник

# Интуиция проксимального отображения

- GD из метода проксимального отображения. Возвращаемся к дискретизации:



## Интуиция проксимального отображения

- GD из метода проксимального отображения. Возвращаемся к дискретизации:

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

# Интуиция проксимального отображения

- GD из метода проксимального отображения. Возвращаемся к дискретизации:

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

$$(I + \alpha \nabla f)(x_{k+1}) = x_k$$

# Интуиция проксимального отображения

- GD из метода проксимального отображения. Возвращаемся к дискретизации:

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

$$(I + \alpha \nabla f)(x_{k+1}) = x_k$$

$$x_{k+1} = (I + \alpha \nabla f)^{-1} x_k \stackrel{\alpha \rightarrow 0}{\approx} (I - \alpha \nabla f) x_k$$

# Интуиция проксимального отображения

- GD из метода проксимального отображения. Возвращаемся к дискретизации:

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

$$(I + \alpha \nabla f)(x_{k+1}) = x_k$$

$$x_{k+1} = (I + \alpha \nabla f)^{-1} x_k \stackrel{\alpha \rightarrow 0}{\approx} (I - \alpha \nabla f) x_k$$

## Интуиция проксимального отображения

- **GD из метода проксимального отображения.** Возвращаемся к дискретизации:

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

$$(I + \alpha \nabla f)(x_{k+1}) = x_k$$

$$x_{k+1} = (I + \alpha \nabla f)^{-1} x_k \stackrel{\alpha \rightarrow 0}{\approx} (I - \alpha \nabla f) x_k$$

Таким образом, мы получаем обычный градиентный спуск с  $\alpha \rightarrow 0$ :  $x_{k+1} = x_k - \alpha \nabla f(x_k)$ .

- **Метод Ньютона из метода проксимального отображения.** Теперь рассмотрим проксимальное отображение второго порядка приближения функции  $f_{x_k}^{II}(x)$ :

## Интуиция проксимального отображения

- **GD из метода проксимального отображения.** Возвращаемся к дискретизации:

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

$$(I + \alpha \nabla f)(x_{k+1}) = x_k$$

$$x_{k+1} = (I + \alpha \nabla f)^{-1} x_k \stackrel{\alpha \rightarrow 0}{\approx} (I - \alpha \nabla f) x_k$$

Таким образом, мы получаем обычный градиентный спуск с  $\alpha \rightarrow 0$ :  $x_{k+1} = x_k - \alpha \nabla f(x_k)$ .

- **Метод Ньютона из метода проксимального отображения.** Теперь рассмотрим проксимальное отображение второго порядка приближения функции  $f_{x_k}^{II}(x)$ :

$$x_{k+1} = \text{prox}_{f_{x_k}^{II}, \alpha}(x_k) = \arg \min_{x \in \mathbb{R}^n} \left[ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right]$$

## Интуиция проксимального отображения

- **GD из метода проксимального отображения.** Возвращаемся к дискретизации:

$$\begin{aligned}x_{k+1} + \alpha \nabla f(x_{k+1}) &= x_k \\(I + \alpha \nabla f)(x_{k+1}) &= x_k \\x_{k+1} &= (I + \alpha \nabla f)^{-1} x_k \stackrel{\alpha \rightarrow 0}{\approx} (I - \alpha \nabla f) x_k\end{aligned}$$

Таким образом, мы получаем обычный градиентный спуск с  $\alpha \rightarrow 0$ :  $x_{k+1} = x_k - \alpha \nabla f(x_k)$ .

- **Метод Ньютона из метода проксимального отображения.** Теперь рассмотрим проксимальное отображение второго порядка приближения функции  $f_{x_k}^{II}(x)$ :

$$\begin{aligned}x_{k+1} = \text{prox}_{f_{x_k}^{II}, \alpha}(x_k) &= \arg \min_{x \in \mathbb{R}^n} \left[ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right] \\ \nabla f(x_k) + \nabla^2 f(x_k)(x - x_k) + \frac{1}{\alpha}(x - x_k) \Big|_{x=x_{k+1}} &= 0\end{aligned}$$

## Интуиция проксимального отображения

- **GD из метода проксимального отображения.** Возвращаемся к дискретизации:

$$\begin{aligned}x_{k+1} + \alpha \nabla f(x_{k+1}) &= x_k \\(I + \alpha \nabla f)(x_{k+1}) &= x_k \\x_{k+1} &= (I + \alpha \nabla f)^{-1} x_k \stackrel{\alpha \rightarrow 0}{\approx} (I - \alpha \nabla f) x_k\end{aligned}$$

Таким образом, мы получаем обычный градиентный спуск с  $\alpha \rightarrow 0$ :  $x_{k+1} = x_k - \alpha \nabla f(x_k)$ .

- **Метод Ньютона из метода проксимального отображения.** Теперь рассмотрим проксимальное отображение второго порядка приближения функции  $f_{x_k}^{II}(x)$ :

$$\begin{aligned}x_{k+1} &= \text{prox}_{f_{x_k}^{II}, \alpha}(x_k) = \arg \min_{x \in \mathbb{R}^n} \left[ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right] \\&\quad \left. \nabla f(x_k) + \nabla^2 f(x_k)(x - x_k) + \frac{1}{\alpha}(x - x_k) \right|_{x=x_{k+1}} = 0 \\x_{k+1} &= x_k - \left[ \nabla^2 f(x_k) + \frac{1}{\alpha} I \right]^{-1} \nabla f(x_k)\end{aligned}$$



## От проекций к проксимальности

Пусть  $\mathbb{I}_S$  — индикаторная функция для замкнутого, выпуклого множества  $S$ . Возвратимся к ортогональной проекции  $\pi_S(y)$ :

## От проекций к проксимальности

Пусть  $\mathbb{I}_S$  — индикаторная функция для замкнутого, выпуклого множества  $S$ . Возвратимся к ортогональной проекции  $\pi_S(y)$ :

$$\pi_S(y) := \arg \min_{x \in S} \frac{1}{2} \|x - y\|_2^2.$$

## От проекций к проксимальности

Пусть  $\mathbb{I}_S$  — индикаторная функция для замкнутого, выпуклого множества  $S$ . Возвратимся к ортогональной проекции  $\pi_S(y)$ :

$$\pi_S(y) := \arg \min_{x \in S} \frac{1}{2} \|x - y\|_2^2.$$

С использованием следующего обозначения индикаторной функции

$$\mathbb{I}_S(x) = \begin{cases} 0, & x \in S, \\ \infty, & x \notin S, \end{cases}$$

## От проекций к проксимальности

Пусть  $\mathbb{I}_S$  — индикаторная функция для замкнутого, выпуклого множества  $S$ . Возвратимся к ортогональной проекции  $\pi_S(y)$ :

$$\pi_S(y) := \arg \min_{x \in S} \frac{1}{2} \|x - y\|_2^2.$$

С использованием следующего обозначения индикаторной функции

$$\mathbb{I}_S(x) = \begin{cases} 0, & x \in S, \\ \infty, & x \notin S, \end{cases}$$

Перепишем ортогональную проекцию  $\pi_S(y)$  как

$$\pi_S(y) := \arg \min_{x \in \mathbb{R}^n} \frac{1}{2} \|x - y\|^2 + \mathbb{I}_S(x).$$

## От проекций к проксимальности

Пусть  $\mathbb{I}_S$  — индикаторная функция для замкнутого, выпуклого множества  $S$ . Возвратимся к ортогональной проекции  $\pi_S(y)$ :

$$\pi_S(y) := \arg \min_{x \in S} \frac{1}{2} \|x - y\|_2^2.$$

С использованием следующего обозначения индикаторной функции

$$\mathbb{I}_S(x) = \begin{cases} 0, & x \in S, \\ \infty, & x \notin S, \end{cases}$$

Перепишем ортогональную проекцию  $\pi_S(y)$  как

$$\pi_S(y) := \arg \min_{x \in \mathbb{R}^n} \frac{1}{2} \|x - y\|^2 + \mathbb{I}_S(x).$$

Проксимальность: заменим  $\mathbb{I}_S$  на некоторую выпуклую функцию!

$$\text{prox}_r(y) = \text{prox}_{r,1}(y) := \arg \min \frac{1}{2} \|x - y\|^2 + r(x)$$

## Составная оптимизация

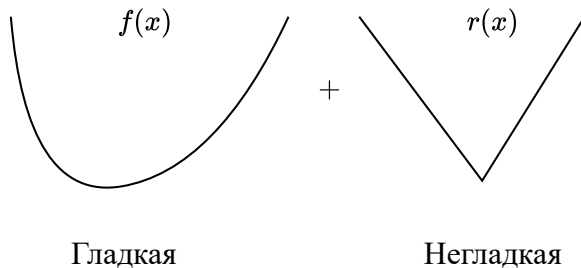
# Регулярные / Составные целевые функции

Многие негладкие задачи имеют вид

$$\min_{x \in \mathbb{R}^n} \varphi(x) = f(x) + r(x)$$

- **Lasso, L1-LS, compressed sensing**

$$f(x) = \frac{1}{2} \|Ax - b\|_2^2, r(x) = \lambda \|x\|_1$$



# Регулярные / Составные целевые функции

Многие негладкие задачи имеют вид

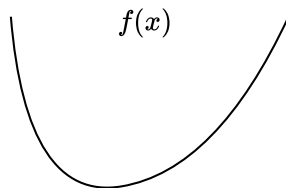
$$\min_{x \in \mathbb{R}^n} \varphi(x) = f(x) + r(x)$$

- **Lasso, L1-LS, compressed sensing**

$$f(x) = \frac{1}{2} \|Ax - b\|_2^2, r(x) = \lambda \|x\|_1$$

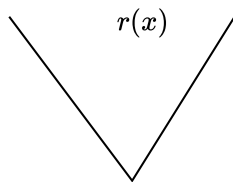
- **L1-логистическая регрессия, разреженная LR**

$$f(x) = -y \log h(x) - (1-y) \log(1-h(x)), r(x) = \lambda \|x\|_1$$



Гладкая

+



Негладкая



# Интуиция проксимального отображения

Условия оптимальности:

$$0 \in \nabla f(x^*) + \partial r(x^*)$$

# Интуиция проксимального отображения

Условия оптимальности:

$$0 \in \nabla f(x^*) + \partial r(x^*)$$

$$0 \in \alpha \nabla f(x^*) + \alpha \partial r(x^*)$$

# Интуиция проксимального отображения

Условия оптимальности:

$$0 \in \nabla f(x^*) + \partial r(x^*)$$

$$0 \in \alpha \nabla f(x^*) + \alpha \partial r(x^*)$$

$$x^* \in \alpha \nabla f(x^*) + (I + \alpha \partial r)(x^*)$$

# Интуиция проксимального отображения

Условия оптимальности:

$$0 \in \nabla f(x^*) + \partial r(x^*)$$

$$0 \in \alpha \nabla f(x^*) + \alpha \partial r(x^*)$$

$$x^* \in \alpha \nabla f(x^*) + (I + \alpha \partial r)(x^*)$$

$$x^* - \alpha \nabla f(x^*) \in (I + \alpha \partial r)(x^*)$$

# Интуиция проксимального отображения

Условия оптимальности:

$$0 \in \nabla f(x^*) + \partial r(x^*)$$

$$0 \in \alpha \nabla f(x^*) + \alpha \partial r(x^*)$$

$$x^* \in \alpha \nabla f(x^*) + (I + \alpha \partial r)(x^*)$$

$$x^* - \alpha \nabla f(x^*) \in (I + \alpha \partial r)(x^*)$$

$$x^* = (I + \alpha \partial r)^{-1}(x^* - \alpha \nabla f(x^*))$$

# Интуиция проксимального отображения

Условия оптимальности:

$$0 \in \nabla f(x^*) + \partial r(x^*)$$

$$0 \in \alpha \nabla f(x^*) + \alpha \partial r(x^*)$$

$$x^* \in \alpha \nabla f(x^*) + (I + \alpha \partial r)(x^*)$$

$$x^* - \alpha \nabla f(x^*) \in (I + \alpha \partial r)(x^*)$$

$$x^* = (I + \alpha \partial r)^{-1}(x^* - \alpha \nabla f(x^*))$$

$$x^* = \text{prox}_{r,\alpha}(x^* - \alpha \nabla f(x^*))$$

# Интуиция проксимального отображения

Условия оптимальности:

$$0 \in \nabla f(x^*) + \partial r(x^*)$$

$$0 \in \alpha \nabla f(x^*) + \alpha \partial r(x^*)$$

$$x^* \in \alpha \nabla f(x^*) + (I + \alpha \partial r)(x^*)$$

$$x^* - \alpha \nabla f(x^*) \in (I + \alpha \partial r)(x^*)$$

$$x^* = (I + \alpha \partial r)^{-1}(x^* - \alpha \nabla f(x^*))$$

$$x^* = \text{prox}_{r,\alpha}(x^* - \alpha \nabla f(x^*))$$

# Интуиция проксимального отображения

Условия оптимальности:

$$0 \in \nabla f(x^*) + \partial r(x^*)$$

$$0 \in \alpha \nabla f(x^*) + \alpha \partial r(x^*)$$

$$x^* \in \alpha \nabla f(x^*) + (I + \alpha \partial r)(x^*)$$

$$x^* - \alpha \nabla f(x^*) \in (I + \alpha \partial r)(x^*)$$

$$x^* = (I + \alpha \partial r)^{-1}(x^* - \alpha \nabla f(x^*))$$

$$x^* = \text{prox}_{r,\alpha}(x^* - \alpha \nabla f(x^*))$$

Которые приводят к методу проксимального градиента:

$$x_{k+1} = \text{prox}_{r,\alpha}(x_k - \alpha \nabla f(x_k))$$

И этот метод сходится со скоростью  $\mathcal{O}(\frac{1}{k})$ !



# Интуиция проксимального отображения

Условия оптимальности:

$$0 \in \nabla f(x^*) + \partial r(x^*)$$

$$0 \in \alpha \nabla f(x^*) + \alpha \partial r(x^*)$$

$$x^* \in \alpha \nabla f(x^*) + (I + \alpha \partial r)(x^*)$$

$$x^* - \alpha \nabla f(x^*) \in (I + \alpha \partial r)(x^*)$$

$$x^* = (I + \alpha \partial r)^{-1}(x^* - \alpha \nabla f(x^*))$$

$$x^* = \text{prox}_{r,\alpha}(x^* - \alpha \nabla f(x^*))$$

Которые приводят к методу проксимального градиента:

$$x_{k+1} = \text{prox}_{r,\alpha}(x_k - \alpha \nabla f(x_k))$$

И этот метод сходится со скоростью  $\mathcal{O}(\frac{1}{k})$ !

**i** Другая форма проксимального оператора

$$\text{prox}_{f,\alpha}(x_k) = \text{prox}_{\alpha f}(x_k) = \arg \min_{x \in \mathbb{R}^n} \left[ \alpha f(x) + \frac{1}{2} \|x - x_k\|_2^2 \right] \quad \text{prox}_f(x_k) = \arg \min_{x \in \mathbb{R}^n} \left[ f(x) + \frac{1}{2} \|x - x_k\|_2^2 \right]$$

## Примеры проксимальных операторов

- $r(x) = \lambda \|x\|_1, \lambda > 0$

$$[\text{prox}_r(x)]_i = [|x_i| - \lambda]_+ \cdot \text{sign}(x_i),$$

который также известен как оператор мягкого порога (soft-thresholding).

## Примеры проксимальных операторов

- $r(x) = \lambda \|x\|_1, \lambda > 0$

$$[\text{prox}_r(x)]_i = [|x_i| - \lambda]_+ \cdot \text{sign}(x_i),$$

который также известен как оператор мягкого порога (soft-thresholding).

- $r(x) = \frac{\lambda}{2} \|x\|_2^2, \lambda > 0$

$$\text{prox}_r(x) = \frac{x}{1 + \lambda}.$$

## Примеры проксимальных операторов

- $r(x) = \lambda \|x\|_1, \lambda > 0$

$$[\text{prox}_r(x)]_i = [|x_i| - \lambda]_+ \cdot \text{sign}(x_i),$$

который также известен как оператор мягкого порога (soft-thresholding).

- $r(x) = \frac{\lambda}{2} \|x\|_2^2, \lambda > 0$

$$\text{prox}_r(x) = \frac{x}{1 + \lambda}.$$

- $r(x) = \mathbb{I}_S(x).$

$$\text{prox}_r(x_k - \alpha \nabla f(x_k)) = \text{proj}_r(x_k - \alpha \nabla f(x_k))$$

# Свойства проксимального оператора

## Theorem

Пусть  $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  — выпуклая функция, для которой  $\text{prox}_r$  определён. Если существует такой  $\hat{x} \in \mathbb{R}^n$ , что  $r(\hat{x}) < +\infty$ , то проксимальный оператор определяется однозначно (т.е. всегда возвращает единственное значение).

**Доказательство:**

# Свойства проксимального оператора

## Theorem

Пусть  $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  — выпуклая функция, для которой  $\text{prox}_r$  определён. Если существует такой  $\hat{x} \in \mathbb{R}^n$ , что  $r(\hat{x}) < +\infty$ , то проксимальный оператор определяется однозначно (т.е. всегда возвращает единственное значение).

## Доказательство:

Проксимальный оператор возвращает минимум некоторой задачи оптимизации.

# Свойства проксимального оператора

## Theorem

Пусть  $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  — выпуклая функция, для которой  $\text{prox}_r$  определён. Если существует такой  $\hat{x} \in \mathbb{R}^n$ , что  $r(\hat{x}) < +\infty$ , то проксимальный оператор определяется однозначно (т.е. всегда возвращает единственное значение).

## Доказательство:

Проксимальный оператор возвращает минимум некоторой задачи оптимизации.

Вопрос: Что можно сказать об этой задаче?

# Свойства проксимального оператора

## i Theorem

Пусть  $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  — выпуклая функция, для которой  $\text{prox}_r$  определён. Если существует такой  $\hat{x} \in \mathbb{R}^n$ , что  $r(\hat{x}) < +\infty$ , то проксимальный оператор определяется однозначно (т.е. всегда возвращает единственное значение).

## Доказательство:

Проксимальный оператор возвращает минимум некоторой задачи оптимизации.

Вопрос: Что можно сказать об этой задаче?

Это сильно выпуклая функция, что означает, что она имеет единственный минимум (существование  $\hat{x}$  необходимо для того, чтобы  $r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2$  принимало конечное значение).



# Свойства проксимального оператора

## i Theorem

Пусть  $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  — выпуклая функция, для которой  $\text{prox}_r$  определен. Тогда, для любых  $x, y \in \mathbb{R}^n$ , следующие три условия эквивалентны:

- $\text{prox}_r(x) = y$ ,

## Доказательство

# Свойства проксимального оператора

## i Theorem

Пусть  $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  — выпуклая функция, для которой  $\text{prox}_r$  определен. Тогда, для любых  $x, y \in \mathbb{R}^n$ , следующие три условия эквивалентны:

- $\text{prox}_r(x) = y$ ,
- $x - y \in \partial r(y)$ ,

## Доказательство

# Свойства проксимального оператора

## i Theorem

Пусть  $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  — выпуклая функция, для которой  $\text{prox}_r$  определен. Тогда, для любых  $x, y \in \mathbb{R}^n$ , следующие три условия эквивалентны:

- $\text{prox}_r(x) = y$ ,
- $x - y \in \partial r(y)$ ,
- $\langle x - y, z - y \rangle \leq r(z) - r(y)$  для любого  $z \in \mathbb{R}^n$ .

## Доказательство

# Свойства проксимального оператора

## i Theorem

Пусть  $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  — выпуклая функция, для которой  $\text{prox}_r$  определен. Тогда, для любых  $x, y \in \mathbb{R}^n$ , следующие три условия эквивалентны:

- $\text{prox}_r(x) = y$ ,
- $x - y \in \partial r(y)$ ,
- $\langle x - y, z - y \rangle \leq r(z) - r(y)$  для любого  $z \in \mathbb{R}^n$ .

## Доказательство

1. Установим эквивалентность между первым и вторым условиями. Первое условие можно переписать как

$$y = \arg \min_{\tilde{x} \in \mathbb{R}^d} \left( r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|^2 \right).$$

Из условий оптимальности для выпуклой функции  $r$ , это эквивалентно:

$$0 \in \partial \left( r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|^2 \right) \Big|_{\tilde{x}=y} = \partial r(y) + y - x.$$

# Свойства проксимального оператора

## i Theorem

Пусть  $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  — выпуклая функция, для которой  $\text{prox}_r$  определен. Тогда, для любых  $x, y \in \mathbb{R}^n$ , следующие три условия эквивалентны:

- $\text{prox}_r(x) = y$ ,
- $x - y \in \partial r(y)$ ,
- $\langle x - y, z - y \rangle \leq r(z) - r(y)$  для любого  $z \in \mathbb{R}^n$ .

## Доказательство

1. Установим эквивалентность между первым и вторым условиями. Первое условие можно переписать как

$$y = \arg \min_{\tilde{x} \in \mathbb{R}^d} \left( r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|^2 \right).$$

Из условий оптимальности для выпуклой функции  $r$ , это эквивалентно:

$$0 \in \partial \left( r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|^2 \right) \Big|_{\tilde{x}=y} = \partial r(y) + y - x.$$

2. Из определения субдифференциала, для любого субградиента  $g \in \partial f(y)$  и для любого  $z \in \mathbb{R}^d$ :

$$\langle g, z - y \rangle \leq r(z) - r(y).$$

В частности, это верно для  $g = x - y$ . Обратно это также очевидно: для  $g = x - y$ , вышеуказанное соотношение выполняется, что означает  $g \in \partial r(y)$ .

# Свойства проксимального оператора

## i Theorem

Оператор  $\text{prox}_r(x)$  является жёстко нестягивающим (FNE):

$$\|\text{prox}_r(x) - \text{prox}_r(y)\|_2^2 \leq \langle \text{prox}_r(x) - \text{prox}_r(y), x - y \rangle$$

и нестягивающим:

$$\|\text{prox}_r(x) - \text{prox}_r(y)\|_2 \leq \|x - y\|_2$$

## Доказательство

1. Пусть  $u = \text{prox}_r(x)$ , и  $v = \text{prox}_r(y)$ . Тогда, из предыдущего свойства:

$$\langle x - u, z_1 - u \rangle \leq r(z_1) - r(u)$$

$$\langle y - v, z_2 - v \rangle \leq r(z_2) - r(v).$$

# Свойства проксимального оператора

## i Theorem

Оператор  $\text{prox}_r(x)$  является жёстко нестягивающим (FNE):

$$\|\text{prox}_r(x) - \text{prox}_r(y)\|_2^2 \leq \langle \text{prox}_r(x) - \text{prox}_r(y), x - y \rangle$$

и нестягивающим:

$$\|\text{prox}_r(x) - \text{prox}_r(y)\|_2 \leq \|x - y\|_2$$

## Доказательство

1. Пусть  $u = \text{prox}_r(x)$ , и  $v = \text{prox}_r(y)$ . Тогда, из предыдущего свойства:

$$\langle x - u, z_1 - u \rangle \leq r(z_1) - r(u)$$

$$\langle y - v, z_2 - v \rangle \leq r(z_2) - r(v).$$

2. Заменим  $z_1 = v$  и  $z_2 = u$  и сложим:

$$\langle x - u, v - u \rangle + \langle y - v, u - v \rangle \leq 0,$$

$$\langle x - y, v - u \rangle + \|v - u\|_2^2 \leq 0.$$

# Свойства проксимального оператора

## i Theorem

Оператор  $\text{prox}_r(x)$  является жёстко нестягивающим (FNE):

$$\|\text{prox}_r(x) - \text{prox}_r(y)\|_2^2 \leq \langle \text{prox}_r(x) - \text{prox}_r(y), x - y \rangle$$

и нестягивающим:

$$\|\text{prox}_r(x) - \text{prox}_r(y)\|_2 \leq \|x - y\|_2$$

## Доказательство

1. Пусть  $u = \text{prox}_r(x)$ , и  $v = \text{prox}_r(y)$ . Тогда, из предыдущего свойства:

$$\langle x - u, z_1 - u \rangle \leq r(z_1) - r(u)$$

$$\langle y - v, z_2 - v \rangle \leq r(z_2) - r(v).$$

2. Заменим  $z_1 = v$  и  $z_2 = u$  и сложим:

$$\langle x - u, v - u \rangle + \langle y - v, u - v \rangle \leq 0,$$

$$\langle x - y, v - u \rangle + \|v - u\|_2^2 \leq 0.$$

3. Что и требовалось доказать после подстановки  $u$  и  $v$ .

$$\|u - v\|_2^2 \leq \langle x - y, u - v \rangle$$



# Свойства проксимального оператора

## i Theorem

Оператор  $\text{prox}_r(x)$  является жёстко нестягивающим (FNE):

$$\|\text{prox}_r(x) - \text{prox}_r(y)\|_2^2 \leq \langle \text{prox}_r(x) - \text{prox}_r(y), x - y \rangle$$

и нестягивающим:

$$\|\text{prox}_r(x) - \text{prox}_r(y)\|_2 \leq \|x - y\|_2$$

## Доказательство

1. Пусть  $u = \text{prox}_r(x)$ , и  $v = \text{prox}_r(y)$ . Тогда, из предыдущего свойства:

$$\langle x - u, z_1 - u \rangle \leq r(z_1) - r(u)$$

$$\langle y - v, z_2 - v \rangle \leq r(z_2) - r(v).$$

2. Заменяем  $z_1 = v$  и  $z_2 = u$  и сложим:

$$\langle x - u, v - u \rangle + \langle y - v, u - v \rangle \leq 0,$$

$$\langle x - y, v - u \rangle + \|v - u\|_2^2 \leq 0.$$

3. Что и требовалось доказать после подстановки  $u$  и  $v$ .

$$\|u - v\|_2^2 \leq \langle x - y, u - v \rangle$$

4. Последний пункт следует из неравенства Коши-Буняковского для последнего неравенства.

# Свойства проксимального оператора

## i Theorem

Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  и  $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  — выпуклые функции. Кроме того, пусть  $f$  непрерывно дифференцируема и  $L$ -гладкая, а для  $r$ ,  $\text{prox}_r$  определена. Тогда,  $x^*$  является решением составной задачи оптимизации тогда и только тогда, когда для любого  $\alpha > 0$ , выполняется:

$$x^* = \text{prox}_{r,\alpha}(x^* - \alpha \nabla f(x^*))$$

## Доказательство

1. Условия оптимальности:

# Свойства проксимального оператора

## i Theorem

Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  и  $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  — выпуклые функции. Кроме того, пусть  $f$  непрерывно дифференцируема и  $L$ -гладкая, а для  $r$ ,  $\text{prox}_r$  определена. Тогда,  $x^*$  является решением составной задачи оптимизации тогда и только тогда, когда для любого  $\alpha > 0$ , выполняется:

$$x^* = \text{prox}_{r,\alpha}(x^* - \alpha \nabla f(x^*))$$

## Доказательство

1. Условия оптимальности:

$$0 \in \nabla f(x^*) + \partial r(x^*)$$

# Свойства проксимального оператора

## i Theorem

Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  и  $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  — выпуклые функции. Кроме того, пусть  $f$  непрерывно дифференцируема и  $L$ -гладкая, а для  $r$ ,  $\text{prox}_r$  определена. Тогда,  $x^*$  является решением составной задачи оптимизации тогда и только тогда, когда для любого  $\alpha > 0$ , выполняется:

$$x^* = \text{prox}_{r,\alpha}(x^* - \alpha \nabla f(x^*))$$

## Доказательство

1. Условия оптимальности:

$$\begin{aligned} 0 &\in \nabla f(x^*) + \partial r(x^*) \\ -\alpha \nabla f(x^*) &\in \alpha \partial r(x^*) \end{aligned}$$

# Свойства проксимального оператора

## i Theorem

Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  и  $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  — выпуклые функции. Кроме того, пусть  $f$  непрерывно дифференцируема и  $L$ -гладкая, а для  $r$ ,  $\text{prox}_r$  определена. Тогда,  $x^*$  является решением составной задачи оптимизации тогда и только тогда, когда для любого  $\alpha > 0$ , выполняется:

$$x^* = \text{prox}_{r,\alpha}(x^* - \alpha \nabla f(x^*))$$

## Доказательство

### 1. Условия оптимальности:

$$\begin{aligned} 0 &\in \nabla f(x^*) + \partial r(x^*) \\ -\alpha \nabla f(x^*) &\in \alpha \partial r(x^*) \\ x^* - \alpha \nabla f(x^*) - x^* &\in \alpha \partial r(x^*) \end{aligned}$$

# Свойства проксимального оператора

## i Theorem

Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  и  $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  — выпуклые функции. Кроме того, пусть  $f$  непрерывно дифференцируема и  $L$ -гладкая, а для  $r$ ,  $\text{prox}_r$  определена. Тогда,  $x^*$  является решением составной задачи оптимизации тогда и только тогда, когда для любого  $\alpha > 0$ , выполняется:

$$x^* = \text{prox}_{r,\alpha}(x^* - \alpha \nabla f(x^*))$$

## Доказательство

1. Условия оптимальности:

$$\begin{aligned} 0 &\in \nabla f(x^*) + \partial r(x^*) \\ -\alpha \nabla f(x^*) &\in \alpha \partial r(x^*) \\ x^* - \alpha \nabla f(x^*) - x^* &\in \alpha \partial r(x^*) \end{aligned}$$

2. Возвратимся к предыдущей лемме:

$$\text{prox}_r(x) = y \Leftrightarrow x - y \in \partial r(y)$$

# Свойства проксимального оператора

## i Theorem

Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  и  $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  — выпуклые функции. Кроме того, пусть  $f$  непрерывно дифференцируема и  $L$ -гладкая, а для  $r$ ,  $\text{prox}_r$  определена. Тогда,  $x^*$  является решением составной задачи оптимизации тогда и только тогда, когда для любого  $\alpha > 0$ , выполняется:

$$x^* = \text{prox}_{r,\alpha}(x^* - \alpha \nabla f(x^*))$$

## Доказательство

1. Условия оптимальности:

$$\begin{aligned} 0 &\in \nabla f(x^*) + \partial r(x^*) \\ -\alpha \nabla f(x^*) &\in \alpha \partial r(x^*) \\ x^* - \alpha \nabla f(x^*) - x^* &\in \alpha \partial r(x^*) \end{aligned}$$

2. Возвратимся к предыдущей лемме:

$$\text{prox}_r(x) = y \Leftrightarrow x - y \in \partial r(y)$$

3. Наконец,

$$x^* = \text{prox}_{\alpha r}(x^* - \alpha \nabla f(x^*)) = \text{prox}_{r,\alpha}(x^* - \alpha \nabla f(x^*))$$

## Теоретические инструменты для анализа сходимости



## Theorem

Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  —  $L$ -гладкая выпуклая функция. Тогда, для любых  $x, y \in \mathbb{R}^n$ , выполняется неравенство:

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) \text{ или, эквивалентно,}$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L (f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$

## Доказательство.

1. Рассмотрим другую функцию  $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$ . Очевидно, это выпуклая функция (как сумма выпуклых функций). И легко проверить, что она является  $L$ -гладкой функцией по определению, так как  $\nabla \varphi(y) = \nabla f(y) - \nabla f(x)$  и  $\|\nabla \varphi(y_1) - \nabla \varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \leq L\|y_1 - y_2\|$ .

## Theorem

Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  —  $L$ -гладкая выпуклая функция. Тогда, для любых  $x, y \in \mathbb{R}^n$ , выполняется неравенство:

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) \text{ или, эквивалентно,}$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L (f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$

## Доказательство.

1. Рассмотрим другую функцию  $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$ . Очевидно, это выпуклая функция (как сумма выпуклых функций). И легко проверить, что она является  $L$ -гладкой функцией по определению, так как  $\nabla \varphi(y) = \nabla f(y) - \nabla f(x)$  и  $\|\nabla \varphi(y_1) - \nabla \varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \leq L\|y_1 - y_2\|$ .
2. Теперь рассмотрим свойство гладкости параболы для функции  $\varphi(y)$ :

## Theorem

Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  —  $L$ -гладкая выпуклая функция. Тогда, для любых  $x, y \in \mathbb{R}^n$ , выполняется неравенство:

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) \text{ или, эквивалентно,}$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L (f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$

## Доказательство.

1. Рассмотрим другую функцию  $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$ . Очевидно, это выпуклая функция (как сумма выпуклых функций). И легко проверить, что она является  $L$ -гладкой функцией по определению, так как  $\nabla \varphi(y) = \nabla f(y) - \nabla f(x)$  и  $\|\nabla \varphi(y_1) - \nabla \varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \leq L\|y_1 - y_2\|$ .
2. Теперь рассмотрим свойство гладкости параболы для функции  $\varphi(y)$ :

$$\varphi(y) \leq \varphi(x) + \langle \nabla \varphi(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2$$



## i Theorem

Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  —  $L$ -гладкая выпуклая функция. Тогда, для любых  $x, y \in \mathbb{R}^n$ , выполняется неравенство:

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) \text{ или, эквивалентно,}$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L (f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$

## Доказательство.

1. Рассмотрим другую функцию  $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$ . Очевидно, это выпуклая функция (как сумма выпуклых функций). И легко проверить, что она является  $L$ -гладкой функцией по определению, так как  $\nabla \varphi(y) = \nabla f(y) - \nabla f(x)$  и  $\|\nabla \varphi(y_1) - \nabla \varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \leq L\|y_1 - y_2\|$ .
2. Теперь рассмотрим свойство гладкости параболы для функции  $\varphi(y)$ :

$$\varphi(y) \leq \varphi(x) + \langle \nabla \varphi(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2$$

$$x := y, y := y - \frac{1}{L} \nabla \varphi(y) \quad \varphi\left(y - \frac{1}{L} \nabla \varphi(y)\right) \leq \varphi(y) + \left\langle \nabla \varphi(y), -\frac{1}{L} \nabla \varphi(y) \right\rangle + \frac{1}{2L} \|\nabla \varphi(y)\|_2^2$$

## i Theorem

Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  —  $L$ -гладкая выпуклая функция. Тогда, для любых  $x, y \in \mathbb{R}^n$ , выполняется неравенство:

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) \text{ или, эквивалентно,}$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L (f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$

## Доказательство.

1. Рассмотрим другую функцию  $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$ . Очевидно, это выпуклая функция (как сумма выпуклых функций). И легко проверить, что она является  $L$ -гладкой функцией по определению, так как  $\nabla \varphi(y) = \nabla f(y) - \nabla f(x)$  и  $\|\nabla \varphi(y_1) - \nabla \varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \leq L\|y_1 - y_2\|$ .
2. Теперь рассмотрим свойство гладкости параболы для функции  $\varphi(y)$ :

$$\varphi(y) \leq \varphi(x) + \langle \nabla \varphi(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2$$

$$x := y, y := y - \frac{1}{L} \nabla \varphi(y) \quad \varphi\left(y - \frac{1}{L} \nabla \varphi(y)\right) \leq \varphi(y) + \left\langle \nabla \varphi(y), -\frac{1}{L} \nabla \varphi(y) \right\rangle + \frac{1}{2L} \|\nabla \varphi(y)\|_2^2$$

$$\varphi\left(y - \frac{1}{L} \nabla \varphi(y)\right) \leq \varphi(y) - \frac{1}{2L} \|\nabla \varphi(y)\|_2^2$$



3. Из условий первого порядка для выпуклой функции  $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$ , мы можем заключить, что для любого  $x$ , минимум функции  $\varphi(y)$  находится в точке  $y = x$ . Следовательно:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$



3. Из условий первого порядка для выпуклой функции  $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$ , мы можем заключить, что для любого  $x$ , минимум функции  $\varphi(y)$  находится в точке  $y = x$ . Следовательно:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Теперь, подставим  $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$ :



3. Из условий первого порядка для выпуклой функции  $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$ , мы можем заключить, что для любого  $x$ , минимум функции  $\varphi(y)$  находится в точке  $y = x$ . Следовательно:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Теперь, подставим  $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$ :

$$f(x) - \langle \nabla f(x), x \rangle \leq f(y) - \langle \nabla f(x), y \rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$





3. Из условий первого порядка для выпуклой функции  $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$ , мы можем заключить, что для любого  $x$ , минимум функции  $\varphi(y)$  находится в точке  $y = x$ . Следовательно:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Теперь, подставим  $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$ :

$$f(x) - \langle \nabla f(x), x \rangle \leq f(y) - \langle \nabla f(x), y \rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$



3. Из условий первого порядка для выпуклой функции  $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$ , мы можем заключить, что для любого  $x$ , минимум функции  $\varphi(y)$  находится в точке  $y = x$ . Следовательно:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Теперь, подставим  $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$ :

$$f(x) - \langle \nabla f(x), x \rangle \leq f(y) - \langle \nabla f(x), y \rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 \leq 2L(f(y) - f(x) - \langle \nabla f(x), y - x \rangle)$$



3. Из условий первого порядка для выпуклой функции  $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$ , мы можем заключить, что для любого  $x$ , минимум функции  $\varphi(y)$  находится в точке  $y = x$ . Следовательно:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Теперь, подставим  $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$ :

$$f(x) - \langle \nabla f(x), x \rangle \leq f(y) - \langle \nabla f(x), y \rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 \leq 2L(f(y) - f(x) - \langle \nabla f(x), y - x \rangle)$$

поменять местами  $x$  и  $y$

$$\|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L(f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$



3. Из условий первого порядка для выпуклой функции  $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$ , мы можем заключить, что для любого  $x$ , минимум функции  $\varphi(y)$  находится в точке  $y = x$ . Следовательно:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Теперь, подставим  $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$ :

$$f(x) - \langle \nabla f(x), x \rangle \leq f(y) - \langle \nabla f(x), y \rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 \leq 2L(f(y) - f(x) - \langle \nabla f(x), y - x \rangle)$$

поменять местами  $x$  и  $y$

$$\|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L(f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$



3. Из условий первого порядка для выпуклой функции  $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$ , мы можем заключить, что для любого  $x$ , минимум функции  $\varphi(y)$  находится в точке  $y = x$ . Следовательно:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Теперь, подставим  $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$ :

$$f(x) - \langle \nabla f(x), x \rangle \leq f(y) - \langle \nabla f(x), y \rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 \leq 2L(f(y) - f(x) - \langle \nabla f(x), y - x \rangle)$$

поменять местами  $x$  и  $y$

$$\|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L(f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$

Лемма доказана. С первого взгляда она не имеет большого геометрического смысла, но мы будем использовать ее как удобный инструмент для оценки разницы между градиентами.



## i Theorem

Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  непрерывно дифференцируема на  $\mathbb{R}^n$ . Тогда, функция  $f$  является  $\mu$ -сильно выпуклой тогда и только тогда, когда для любых  $x, y \in \mathbb{R}^d$  выполняется следующее:

$$\text{Strongly convex case } \mu > 0 \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$$

$$\text{Convex case } \mu = 0 \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$$

## Доказательство

1. Мы докажем только случай сильной выпуклости, случай выпуклости следует из него с установкой  $\mu = 0$ . Начнем с необходимости. Для сильно выпуклой функции

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

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

$$\text{sum} \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$$

## Анализ сходимости

2. Для достаточности мы предполагаем, что  $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$ . Используя теорему Ньютона-Лейбница  $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$ :

## Анализ сходимости

2. Для достаточности мы предполагаем, что  $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$ . Используя теорему Ньютона-Лейбница  $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$ :

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle = \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle$$



## Анализ сходимости

2. Для достаточности мы предполагаем, что  $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$ . Используя теорему Ньютона-Лейбница  $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$ :

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), (x - y) \rangle dt \end{aligned}$$

## Анализ сходимости

2. Для достаточности мы предполагаем, что  $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$ . Используя теорему Ньютона-Лейбница  $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$ :

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ y + t(x - y) - y &= t(x - y) \end{aligned}$$
$$\begin{aligned} &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), (x - y) \rangle dt \\ &= \int_0^1 t^{-1} \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt \end{aligned}$$

## Анализ сходимости

2. Для достаточности мы предполагаем, что  $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$ . Используя теорему Ньютона-Лейбница  $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$ :

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ y + t(x - y) - y &= t(x - y) \\ &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt \\ &\geq \int_0^1 t^{-1} \mu \|t(x - y)\|^2 dt \end{aligned}$$

## Анализ сходимости

2. Для достаточности мы предполагаем, что  $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$ . Используя теорему Ньютона-Лейбница  $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$ :

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ y + t(x - y) - y &= t(x - y) \\ &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt \\ &\geq \int_0^1 t^{-1} \mu \|t(x - y)\|^2 dt = \mu \|x - y\|^2 \int_0^1 t dt \end{aligned}$$

## Анализ сходимости

2. Для достаточности мы предполагаем, что  $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$ . Используя теорему Ньютона-Лейбница  $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$ :

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ y + t(x - y) - y &= t(x - y) \\ &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt \\ &\geq \int_0^1 t^{-1} \mu \|t(x - y)\|^2 dt = \mu \|x - y\|^2 \int_0^1 t dt = \frac{\mu}{2} \|x - y\|_2^2 \end{aligned}$$

## Анализ сходимости

2. Для достаточности мы предполагаем, что  $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$ . Используя теорему Ньютона-Лейбница  $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$ :

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ y + t(x - y) - y &= t(x - y) \\ &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt \\ &\geq \int_0^1 t^{-1} \mu \|t(x - y)\|^2 dt = \mu \|x - y\|^2 \int_0^1 t dt = \frac{\mu}{2} \|x - y\|_2^2 \end{aligned}$$

## Анализ сходимости

2. Для достаточности мы предполагаем, что  $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$ . Используя теорему Ньютона-Лейбница  $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$ :

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ y + t(x - y) - y &= t(x - y) \\ &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt \\ &\geq \int_0^1 t^{-1} \mu \|t(x - y)\|^2 dt = \mu \|x - y\|^2 \int_0^1 t dt = \frac{\mu}{2} \|x - y\|_2^2 \end{aligned}$$

Таким образом, мы получаем критерий сильной выпуклости, удовлетворяющий

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

## Анализ сходимости

2. Для достаточности мы предполагаем, что  $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$ . Используя теорему Ньютона-Лейбница  $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$ :

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ y + t(x - y) - y &= t(x - y) \\ &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt \\ &\geq \int_0^1 t^{-1} \mu \|t(x - y)\|^2 dt = \mu \|x - y\|^2 \int_0^1 t dt = \frac{\mu}{2} \|x - y\|_2^2 \end{aligned}$$

Таким образом, мы получаем критерий сильной выпуклости, удовлетворяющий

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|_2^2 \text{ или, эквивалентно:}$$



## Анализ сходимости

2. Для достаточности мы предполагаем, что  $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$ . Используя теорему Ньютона-Лейбница  $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$ :

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ y + t(x - y) - y &= t(x - y) \\ &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt \\ &\geq \int_0^1 t^{-1} \mu \|t(x - y)\|^2 dt = \mu \|x - y\|^2 \int_0^1 t dt = \frac{\mu}{2} \|x - y\|_2^2 \end{aligned}$$

Таким образом, мы получаем критерий сильной выпуклости, удовлетворяющий

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|_2^2 \text{ или, эквивалентно:}$$

поменять местами  $x$  и  $y$

$$- \langle \nabla f(x), x - y \rangle \leq - \left( f(x) - f(y) + \frac{\mu}{2} \|x - y\|_2^2 \right)$$

## Проксимальный метод градиента. Выпуклый случай

## i Theorem

Рассмотрим проксимальный метод градиента

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k))$$

Для критерия  $\varphi(x) = f(x) + r(x)$ , мы предполагаем:

- $f$  выпукла, дифференцируема,  $\text{dom}(f) = \mathbb{R}^n$ , и  $\nabla f$  является липшицевой с константой  $L > 0$ .
- $r$  выпукла, и  $\text{prox}_{\alpha r}(x_k) = \arg \min_{x \in \mathbb{R}^n} [\alpha r(x) + \frac{1}{2} \|x - x_k\|_2^2]$  может быть вычислен.

Проксимальный градиентный спуск с фиксированным шагом  $\alpha = 1/L$  удовлетворяет

$$\varphi(x_k) - \varphi^* \leq \frac{L \|x_0 - x^*\|^2}{2k},$$

Проксимальный градиентный спуск имеет скорость сходимости  $O(1/k)$  или  $O(1/\varepsilon)$ . Это соответствует скорости градиентного спуска! (Но помните о стоимости проксимальной операции)

# Анализ сходимости

## Доказательство

1. Введем **градиентное отображение**, обозначаемое как  $G_\alpha(x)$ , действующее как “градиентный объект”:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k))$$

$$x_{k+1} = x_k - \alpha G_\alpha(x_k).$$

где  $G_\alpha(x)$  является:

$$G_\alpha(x) = \frac{1}{\alpha} (x - \text{prox}_{\alpha r}(x - \alpha \nabla f(x)))$$

Очевидно, что  $G_\alpha(x) = 0$  тогда и только тогда, когда  $x$  оптимален. Следовательно,  $G_\alpha$  аналогичен  $\nabla f$ . Если  $x$  локально оптимален, то  $G_\alpha(x) = 0$  даже для невыпуклой  $f$ . Это демонстрирует, что проксимальный градиентный метод эффективно объединяет градиентный спуск на  $f$  с проксимальным оператором  $r$ , позволяя ему эффективно обрабатывать недифференцируемые компоненты.

# Анализ сходимости

## Доказательство

1. Введем **градиентное отображение**, обозначаемое как  $G_\alpha(x)$ , действующее как “градиентный объект”:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k))$$

$$x_{k+1} = x_k - \alpha G_\alpha(x_k).$$

где  $G_\alpha(x)$  является:

$$G_\alpha(x) = \frac{1}{\alpha} (x - \text{prox}_{\alpha r}(x - \alpha \nabla f(x)))$$

Очевидно, что  $G_\alpha(x) = 0$  тогда и только тогда, когда  $x$  оптимален. Следовательно,  $G_\alpha$  аналогичен  $\nabla f$ . Если  $x$  локально оптимален, то  $G_\alpha(x) = 0$  даже для невыпуклой  $f$ . Это демонстрирует, что проксимальный градиентный метод эффективно объединяет градиентный спуск на  $f$  с проксимальным оператором  $r$ , позволяя ему эффективно обрабатывать недифференцируемые компоненты.

2. Мы будем использовать гладкость и выпуклость  $f$  для некоторой произвольной точки  $x$ :

# Анализ сходимости

## Доказательство

1. Введем **градиентное отображение**, обозначаемое как  $G_\alpha(x)$ , действующее как “градиентный объект”:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k))$$

$$x_{k+1} = x_k - \alpha G_\alpha(x_k).$$

где  $G_\alpha(x)$  является:

$$G_\alpha(x) = \frac{1}{\alpha} (x - \text{prox}_{\alpha r}(x - \alpha \nabla f(x)))$$

Очевидно, что  $G_\alpha(x) = 0$  тогда и только тогда, когда  $x$  оптимален. Следовательно,  $G_\alpha$  аналогичен  $\nabla f$ . Если  $x$  локально оптимален, то  $G_\alpha(x) = 0$  даже для невыпуклой  $f$ . Это демонстрирует, что проксимальный градиентный метод эффективно объединяет градиентный спуск на  $f$  с проксимальным оператором  $r$ , позволяя ему эффективно обрабатывать недифференцируемые компоненты.

2. Мы будем использовать гладкость и выпуклость  $f$  для некоторой произвольной точки  $x$ :

$$\text{гладкость} \quad f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|_2^2$$

# Анализ сходимости

## Доказательство

1. Введем **градиентное отображение**, обозначаемое как  $G_\alpha(x)$ , действующее как “градиентный объект”:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k))$$

$$x_{k+1} = x_k - \alpha G_\alpha(x_k).$$

где  $G_\alpha(x)$  является:

$$G_\alpha(x) = \frac{1}{\alpha} (x - \text{prox}_{\alpha r}(x - \alpha \nabla f(x)))$$

Очевидно, что  $G_\alpha(x) = 0$  тогда и только тогда, когда  $x$  оптимален. Следовательно,  $G_\alpha$  аналогичен  $\nabla f$ . Если  $x$  локально оптимален, то  $G_\alpha(x) = 0$  даже для невыпуклой  $f$ . Это демонстрирует, что проксимальный градиентный метод эффективно объединяет градиентный спуск на  $f$  с проксимальным оператором  $r$ , позволяя ему эффективно обрабатывать недифференцируемые компоненты.

2. Мы будем использовать гладкость и выпуклость  $f$  для некоторой произвольной точки  $x$ :

$$\text{гладкость} \quad f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|_2^2$$

$$\text{выпуклость} \quad f(x) \geq f(x_k) + \langle \nabla f(x_k), x - x_k \rangle$$

# Анализ сходимости

## Доказательство

1. Введем **градиентное отображение**, обозначаемое как  $G_\alpha(x)$ , действующее как “градиентный объект”:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k))$$

$$x_{k+1} = x_k - \alpha G_\alpha(x_k).$$

где  $G_\alpha(x)$  является:

$$G_\alpha(x) = \frac{1}{\alpha} (x - \text{prox}_{\alpha r}(x - \alpha \nabla f(x)))$$

Очевидно, что  $G_\alpha(x) = 0$  тогда и только тогда, когда  $x$  оптимален. Следовательно,  $G_\alpha$  аналогичен  $\nabla f$ . Если  $x$  локально оптимален, то  $G_\alpha(x) = 0$  даже для невыпуклой  $f$ . Это демонстрирует, что проксимальный градиентный метод эффективно объединяет градиентный спуск на  $f$  с проксимальным оператором  $r$ , позволяя ему эффективно обрабатывать недифференцируемые компоненты.

2. Мы будем использовать гладкость и выпуклость  $f$  для некоторой произвольной точки  $x$ :

$$\text{гладкость} \quad f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|_2^2$$

$$\text{выпуклость} \quad f(x) \geq f(x_k) + \langle \nabla f(x_k), x - x_k \rangle \quad \leq f(x) - \langle \nabla f(x_k), x - x_k \rangle + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2$$



# Анализ сходимости

## Доказательство

1. Введем **градиентное отображение**, обозначаемое как  $G_\alpha(x)$ , действующее как “градиентный объект”:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k))$$

$$x_{k+1} = x_k - \alpha G_\alpha(x_k).$$

где  $G_\alpha(x)$  является:

$$G_\alpha(x) = \frac{1}{\alpha} (x - \text{prox}_{\alpha r}(x - \alpha \nabla f(x)))$$

Очевидно, что  $G_\alpha(x) = 0$  тогда и только тогда, когда  $x$  оптимален. Следовательно,  $G_\alpha$  аналогичен  $\nabla f$ . Если  $x$  локально оптимален, то  $G_\alpha(x) = 0$  даже для невыпуклой  $f$ . Это демонстрирует, что проксимальный градиентный метод эффективно объединяет градиентный спуск на  $f$  с проксимальным оператором  $r$ , позволяя ему эффективно обрабатывать недифференцируемые компоненты.

2. Мы будем использовать гладкость и выпуклость  $f$  для некоторой произвольной точки  $x$ :

$$\text{гладкость} \quad f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|_2^2$$

$$\begin{aligned} \text{выпуклость} \quad f(x) \geq f(x_k) + \langle \nabla f(x_k), x - x_k \rangle &\leq f(x) - \langle \nabla f(x_k), x - x_k \rangle + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2 \\ &\leq f(x) + \langle \nabla f(x_k), x_{k+1} - x \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2 \end{aligned}$$

## Анализ сходимости

3. Теперь мы будем использовать свойство проксимального оператора, которое было доказано ранее:

## Анализ сходимости

3. Теперь мы будем использовать свойство проксимального оператора, которое было доказано ранее:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k)) \quad \Leftrightarrow \quad x_k - \alpha \nabla f(x_k) - x_{k+1} \in \partial \alpha r(x_{k+1})$$

## Анализ сходимости

3. Теперь мы будем использовать свойство проксимального оператора, которое было доказано ранее:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k)) \quad \Leftrightarrow \quad x_k - \alpha \nabla f(x_k) - x_{k+1} \in \partial \alpha r(x_{k+1})$$

$$\text{Так как } x_{k+1} - x_k = -\alpha G_\alpha(x_k) \quad \Rightarrow \quad \alpha G_\alpha(x_k) - \alpha \nabla f(x_k) \in \partial \alpha r(x_{k+1})$$

## Анализ сходимости

3. Теперь мы будем использовать свойство проксимального оператора, которое было доказано ранее:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k)) \quad \Leftrightarrow \quad x_k - \alpha \nabla f(x_k) - x_{k+1} \in \partial \alpha r(x_{k+1})$$

$$\text{Так как } x_{k+1} - x_k = -\alpha G_\alpha(x_k) \quad \Rightarrow \quad \alpha G_\alpha(x_k) - \alpha \nabla f(x_k) \in \partial \alpha r(x_{k+1})$$

$$G_\alpha(x_k) - \nabla f(x_k) \in \partial r(x_{k+1})$$

## Анализ сходимости

3. Теперь мы будем использовать свойство проксимального оператора, которое было доказано ранее:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k)) \quad \Leftrightarrow \quad x_k - \alpha \nabla f(x_k) - x_{k+1} \in \partial \alpha r(x_{k+1})$$

$$\text{Так как } x_{k+1} - x_k = -\alpha G_\alpha(x_k) \quad \Rightarrow \quad \alpha G_\alpha(x_k) - \alpha \nabla f(x_k) \in \partial \alpha r(x_{k+1})$$

$$G_\alpha(x_k) - \nabla f(x_k) \in \partial r(x_{k+1})$$

4. Из определения субградиента выпуклой функции  $r$  для любой точки  $x$ :

## Анализ сходимости

3. Теперь мы будем использовать свойство проксимального оператора, которое было доказано ранее:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k)) \quad \Leftrightarrow \quad x_k - \alpha \nabla f(x_k) - x_{k+1} \in \partial \alpha r(x_{k+1})$$

$$\text{Так как } x_{k+1} - x_k = -\alpha G_\alpha(x_k) \quad \Rightarrow \quad \alpha G_\alpha(x_k) - \alpha \nabla f(x_k) \in \partial \alpha r(x_{k+1})$$

$$G_\alpha(x_k) - \nabla f(x_k) \in \partial r(x_{k+1})$$

4. Из определения субградиента выпуклой функции  $r$  для любой точки  $x$ :

$$r(x) \geq r(x_{k+1}) + \langle g, x - x_{k+1} \rangle, \quad g \in \partial r(x_{k+1})$$

## Анализ сходимости

3. Теперь мы будем использовать свойство проксимального оператора, которое было доказано ранее:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k)) \quad \Leftrightarrow \quad x_k - \alpha \nabla f(x_k) - x_{k+1} \in \partial \alpha r(x_{k+1})$$

$$\text{Так как } x_{k+1} - x_k = -\alpha G_\alpha(x_k) \quad \Rightarrow \quad \alpha G_\alpha(x_k) - \alpha \nabla f(x_k) \in \partial \alpha r(x_{k+1})$$

$$G_\alpha(x_k) - \nabla f(x_k) \in \partial r(x_{k+1})$$

4. Из определения субградиента выпуклой функции  $r$  для любой точки  $x$ :

$$r(x) \geq r(x_{k+1}) + \langle g, x - x_{k+1} \rangle, \quad g \in \partial r(x_{k+1})$$

подставить конкретный субградиент

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



## Анализ сходимости

3. Теперь мы будем использовать свойство проксимального оператора, которое было доказано ранее:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k)) \quad \Leftrightarrow \quad x_k - \alpha \nabla f(x_k) - x_{k+1} \in \partial \alpha r(x_{k+1})$$

$$\text{Так как } x_{k+1} - x_k = -\alpha G_\alpha(x_k) \quad \Rightarrow \quad \alpha G_\alpha(x_k) - \alpha \nabla f(x_k) \in \partial \alpha r(x_{k+1})$$

$$G_\alpha(x_k) - \nabla f(x_k) \in \partial r(x_{k+1})$$

4. Из определения субградиента выпуклой функции  $r$  для любой точки  $x$ :

$$r(x) \geq r(x_{k+1}) + \langle g, x - x_{k+1} \rangle, \quad g \in \partial r(x_{k+1})$$

подставить конкретный субградиент

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

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

## Анализ сходимости

3. Теперь мы будем использовать свойство проксимального оператора, которое было доказано ранее:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k)) \quad \Leftrightarrow \quad x_k - \alpha \nabla f(x_k) - x_{k+1} \in \partial \alpha r(x_{k+1})$$

$$\text{Так как } x_{k+1} - x_k = -\alpha G_\alpha(x_k) \quad \Rightarrow \quad \alpha G_\alpha(x_k) - \alpha \nabla f(x_k) \in \partial \alpha r(x_{k+1})$$

$$G_\alpha(x_k) - \nabla f(x_k) \in \partial r(x_{k+1})$$

4. Из определения субградиента выпуклой функции  $r$  для любой точки  $x$ :

$$r(x) \geq r(x_{k+1}) + \langle g, x - x_{k+1} \rangle, \quad g \in \partial r(x_{k+1})$$

подставить конкретный субградиент

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

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

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

## Анализ сходимости

3. Теперь мы будем использовать свойство проксимального оператора, которое было доказано ранее:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k)) \quad \Leftrightarrow \quad x_k - \alpha \nabla f(x_k) - x_{k+1} \in \partial \alpha r(x_{k+1})$$

$$\text{Так как } x_{k+1} - x_k = -\alpha G_\alpha(x_k) \quad \Rightarrow \quad \alpha G_\alpha(x_k) - \alpha \nabla f(x_k) \in \partial \alpha r(x_{k+1})$$

$$G_\alpha(x_k) - \nabla f(x_k) \in \partial r(x_{k+1})$$

4. Из определения субградиента выпуклой функции  $r$  для любой точки  $x$ :

$$r(x) \geq r(x_{k+1}) + \langle g, x - x_{k+1} \rangle, \quad g \in \partial r(x_{k+1})$$

подставить конкретный субградиент

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

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

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

5. Учитывая приведённую выше оценку, мы возвращаемся к гладкости и выпуклости:

## Анализ сходимости

3. Теперь мы будем использовать свойство проксимального оператора, которое было доказано ранее:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k)) \quad \Leftrightarrow \quad x_k - \alpha \nabla f(x_k) - x_{k+1} \in \partial \alpha r(x_{k+1})$$

$$\text{Так как } x_{k+1} - x_k = -\alpha G_\alpha(x_k) \quad \Rightarrow \quad \alpha G_\alpha(x_k) - \alpha \nabla f(x_k) \in \partial \alpha r(x_{k+1})$$

$$G_\alpha(x_k) - \nabla f(x_k) \in \partial r(x_{k+1})$$

4. Из определения субградиента выпуклой функции  $r$  для любой точки  $x$ :

$$r(x) \geq r(x_{k+1}) + \langle g, x - x_{k+1} \rangle, \quad g \in \partial r(x_{k+1})$$

подставить конкретный субградиент

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

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

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

5. Учитывая приведённую выше оценку, мы возвращаемся к гладкости и выпуклости:

$$f(x_{k+1}) \leq f(x) + \langle \nabla f(x_k), x_{k+1} - x \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2$$

## Анализ сходимости

3. Теперь мы будем использовать свойство проксимального оператора, которое было доказано ранее:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k)) \quad \Leftrightarrow \quad x_k - \alpha \nabla f(x_k) - x_{k+1} \in \partial \alpha r(x_{k+1})$$

$$\text{Так как } x_{k+1} - x_k = -\alpha G_\alpha(x_k) \quad \Rightarrow \quad \alpha G_\alpha(x_k) - \alpha \nabla f(x_k) \in \partial \alpha r(x_{k+1})$$

$$G_\alpha(x_k) - \nabla f(x_k) \in \partial r(x_{k+1})$$

4. Из определения субградиента выпуклой функции  $r$  для любой точки  $x$ :

$$r(x) \geq r(x_{k+1}) + \langle g, x - x_{k+1} \rangle, \quad g \in \partial r(x_{k+1})$$

подставить конкретный субградиент

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

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

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

5. Учитывая приведённую выше оценку, мы возвращаемся к гладкости и выпуклости:

$$f(x_{k+1}) \leq f(x) + \langle \nabla f(x_k), x_{k+1} - x \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2$$

$$f(x_{k+1}) \leq f(x) + r(x) - r(x_{k+1}) - \langle G_\alpha(x_k), x - x_{k+1} \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2$$

## Анализ сходимости

3. Теперь мы будем использовать свойство проксимального оператора, которое было доказано ранее:

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k)) \quad \Leftrightarrow \quad x_k - \alpha \nabla f(x_k) - x_{k+1} \in \partial \alpha r(x_{k+1})$$

$$\text{Так как } x_{k+1} - x_k = -\alpha G_\alpha(x_k) \quad \Rightarrow \quad \alpha G_\alpha(x_k) - \alpha \nabla f(x_k) \in \partial \alpha r(x_{k+1})$$

$$G_\alpha(x_k) - \nabla f(x_k) \in \partial r(x_{k+1})$$

4. Из определения субградиента выпуклой функции  $r$  для любой точки  $x$ :

$$r(x) \geq r(x_{k+1}) + \langle g, x - x_{k+1} \rangle, \quad g \in \partial r(x_{k+1})$$

подставить конкретный субградиент

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

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

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

5. Учитывая приведённую выше оценку, мы возвращаемся к гладкости и выпуклости:

$$f(x_{k+1}) \leq f(x) + \langle \nabla f(x_k), x_{k+1} - x \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2$$

$$f(x_{k+1}) \leq f(x) + r(x) - r(x_{k+1}) - \langle G_\alpha(x_k), x - x_{k+1} \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2$$

$$f(x_{k+1}) + r(x_{k+1}) \leq f(x) + r(x) - \langle G_\alpha(x_k), x - x_k + \alpha G_\alpha(x_k) \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2$$

6. Используя  $\varphi(x) = f(x) + r(x)$  мы можем доказать очень полезное неравенство, которое позволит нам продемонстрировать монотонное убывание итерации:

6. Используя  $\varphi(x) = f(x) + r(x)$  мы можем доказать очень полезное неравенство, которое позволит нам продемонстрировать монотонное убывание итерации:

$$\varphi(x_{k+1}) \leq \varphi(x) - \langle G_\alpha(x_k), x - x_k \rangle - \langle G_\alpha(x_k), \alpha G_\alpha(x_k) \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2$$



6. Используя  $\varphi(x) = f(x) + r(x)$  мы можем доказать очень полезное неравенство, которое позволит нам продемонстрировать монотонное убывание итерации:

$$\begin{aligned}\varphi(x_{k+1}) &\leq \varphi(x) - \langle G_\alpha(x_k), x - x_k \rangle - \langle G_\alpha(x_k), \alpha G_\alpha(x_k) \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2 \\ \varphi(x_{k+1}) &\leq \varphi(x) + \langle G_\alpha(x_k), x_k - x \rangle + \frac{\alpha}{2} (\alpha L - 2) \|G_\alpha(x_k)\|_2^2\end{aligned}$$



6. Используя  $\varphi(x) = f(x) + r(x)$  мы можем доказать очень полезное неравенство, которое позволит нам продемонстрировать монотонное убывание итерации:

$$\varphi(x_{k+1}) \leq \varphi(x) - \langle G_\alpha(x_k), x - x_k \rangle - \langle G_\alpha(x_k), \alpha G_\alpha(x_k) \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2$$

$$\varphi(x_{k+1}) \leq \varphi(x) + \langle G_\alpha(x_k), x_k - x \rangle + \frac{\alpha}{2} (\alpha L - 2) \|G_\alpha(x_k)\|_2^2$$

$$\alpha \leq \frac{1}{L} \Rightarrow \frac{\alpha}{2} (\alpha L - 2) \leq -\frac{\alpha}{2}$$

6. Используя  $\varphi(x) = f(x) + r(x)$  мы можем доказать очень полезное неравенство, которое позволит нам продемонстрировать монотонное убывание итерации:

$$\varphi(x_{k+1}) \leq \varphi(x) - \langle G_\alpha(x_k), x - x_k \rangle - \langle G_\alpha(x_k), \alpha G_\alpha(x_k) \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2$$

$$\varphi(x_{k+1}) \leq \varphi(x) + \langle G_\alpha(x_k), x_k - x \rangle + \frac{\alpha}{2} (\alpha L - 2) \|G_\alpha(x_k)\|_2^2$$

$$\alpha \leq \frac{1}{L} \Rightarrow \frac{\alpha}{2} (\alpha L - 2) \leq -\frac{\alpha}{2} \quad \varphi(x_{k+1}) \leq \varphi(x) + \langle G_\alpha(x_k), x_k - x \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2$$

6. Используя  $\varphi(x) = f(x) + r(x)$  мы можем доказать очень полезное неравенство, которое позволит нам продемонстрировать монотонное убывание итерации:

$$\varphi(x_{k+1}) \leq \varphi(x) - \langle G_\alpha(x_k), x - x_k \rangle - \langle G_\alpha(x_k), \alpha G_\alpha(x_k) \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x_k)\|_2^2$$

$$\varphi(x_{k+1}) \leq \varphi(x) + \langle G_\alpha(x_k), x_k - x \rangle + \frac{\alpha}{2} (\alpha L - 2) \|G_\alpha(x_k)\|_2^2$$

$$\alpha \leq \frac{1}{L} \Rightarrow \frac{\alpha}{2} (\alpha L - 2) \leq -\frac{\alpha}{2} \quad \varphi(x_{k+1}) \leq \varphi(x) + \langle G_\alpha(x_k), x_k - x \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2$$

7. Теперь легко проверить, что когда  $x = x_k$  мы получаем монотонное убывание для проксимального градиентного метода:

$$\varphi(x_{k+1}) \leq \varphi(x_k) - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2$$

8. Когда  $x = x^*$ :



8. Когда  $x = x^*$ :

$$\varphi(x_{k+1}) \leq \varphi(x^*) + \langle G_\alpha(x_k), x_k - x^* \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2$$

8. Когда  $x = x^*$ :

$$\varphi(x_{k+1}) \leq \varphi(x^*) + \langle G_\alpha(x_k), x_k - x^* \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2$$

$$\varphi(x_{k+1}) - \varphi(x^*) \leq \langle G_\alpha(x_k), x_k - x^* \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2$$



8. Когда  $x = x^*$ :

$$\begin{aligned}\varphi(x_{k+1}) &\leq \varphi(x^*) + \langle G_\alpha(x_k), x_k - x^* \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2 \\ \varphi(x_{k+1}) - \varphi(x^*) &\leq \langle G_\alpha(x_k), x_k - x^* \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2 \\ &\leq \frac{1}{2\alpha} [2\langle \alpha G_\alpha(x_k), x_k - x^* \rangle - \|\alpha G_\alpha(x_k)\|_2^2]\end{aligned}$$





8. Когда  $x = x^*$ :

$$\begin{aligned}\varphi(x_{k+1}) &\leq \varphi(x^*) + \langle G_\alpha(x_k), x_k - x^* \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2 \\ \varphi(x_{k+1}) - \varphi(x^*) &\leq \langle G_\alpha(x_k), x_k - x^* \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2 \\ &\leq \frac{1}{2\alpha} [2\langle \alpha G_\alpha(x_k), x_k - x^* \rangle - \|\alpha G_\alpha(x_k)\|_2^2] \\ &\leq \frac{1}{2\alpha} [2\langle \alpha G_\alpha(x_k), x_k - x^* \rangle - \|\alpha G_\alpha(x_k)\|_2^2 - \|x_k - x^*\|_2^2 + \|x_k - x^*\|_2^2]\end{aligned}$$

8. Когда  $x = x^*$ :

$$\begin{aligned}
 \varphi(x_{k+1}) &\leq \varphi(x^*) + \langle G_\alpha(x_k), x_k - x^* \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2 \\
 \varphi(x_{k+1}) - \varphi(x^*) &\leq \langle G_\alpha(x_k), x_k - x^* \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2 \\
 &\leq \frac{1}{2\alpha} [2\langle \alpha G_\alpha(x_k), x_k - x^* \rangle - \|\alpha G_\alpha(x_k)\|_2^2] \\
 &\leq \frac{1}{2\alpha} [2\langle \alpha G_\alpha(x_k), x_k - x^* \rangle - \|\alpha G_\alpha(x_k)\|_2^2 - \|x_k - x^*\|_2^2 + \|x_k - x^*\|_2^2] \\
 &\leq \frac{1}{2\alpha} [-\|x_k - x^* - \alpha G_\alpha(x_k)\|_2^2 + \|x_k - x^*\|_2^2]
 \end{aligned}$$

8. Когда  $x = x^*$ :

$$\begin{aligned}
 \varphi(x_{k+1}) &\leq \varphi(x^*) + \langle G_\alpha(x_k), x_k - x^* \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2 \\
 \varphi(x_{k+1}) - \varphi(x^*) &\leq \langle G_\alpha(x_k), x_k - x^* \rangle - \frac{\alpha}{2} \|G_\alpha(x_k)\|_2^2 \\
 &\leq \frac{1}{2\alpha} [2\langle \alpha G_\alpha(x_k), x_k - x^* \rangle - \|\alpha G_\alpha(x_k)\|_2^2] \\
 &\leq \frac{1}{2\alpha} [2\langle \alpha G_\alpha(x_k), x_k - x^* \rangle - \|\alpha G_\alpha(x_k)\|_2^2 - \|x_k - x^*\|_2^2 + \|x_k - x^*\|_2^2] \\
 &\leq \frac{1}{2\alpha} [-\|x_k - x^* - \alpha G_\alpha(x_k)\|_2^2 + \|x_k - x^*\|_2^2] \\
 &\leq \frac{1}{2\alpha} [\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2]
 \end{aligned}$$

## Анализ сходимости

9. Теперь мы запишем приведенное выше ограничение для всех итераций  $i \in 0, k - 1$  и суммируем их:

## Анализ сходимости

9. Теперь мы запишем приведенное выше ограничение для всех итераций  $i \in 0, k-1$  и суммируем их:

$$\sum_{i=0}^{k-1} [\varphi(x_{i+1}) - \varphi(x^*)] \leq \frac{1}{2\alpha} [\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2]$$

## Анализ сходимости

9. Теперь мы запишем приведенное выше ограничение для всех итераций  $i \in 0, k-1$  и суммируем их:

$$\begin{aligned} \sum_{i=0}^{k-1} [\varphi(x_{i+1}) - \varphi(x^*)] &\leq \frac{1}{2\alpha} [\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2] \\ &\leq \frac{1}{2\alpha} \|x_0 - x^*\|_2^2 \end{aligned}$$

## Анализ сходимости

9. Теперь мы запишем приведенное выше ограничение для всех итераций  $i \in 0, k-1$  и суммируем их:

$$\begin{aligned} \sum_{i=0}^{k-1} [\varphi(x_{i+1}) - \varphi(x^*)] &\leq \frac{1}{2\alpha} [\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2] \\ &\leq \frac{1}{2\alpha} \|x_0 - x^*\|_2^2 \end{aligned}$$

10. Поскольку  $\varphi(x_k)$  является убывающей последовательностью, то:

## Анализ сходимости

9. Теперь мы запишем приведенное выше ограничение для всех итераций  $i \in 0, k-1$  и суммируем их:

$$\begin{aligned} \sum_{i=0}^{k-1} [\varphi(x_{i+1}) - \varphi(x^*)] &\leq \frac{1}{2\alpha} [\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2] \\ &\leq \frac{1}{2\alpha} \|x_0 - x^*\|_2^2 \end{aligned}$$

10. Поскольку  $\varphi(x_k)$  является убывающей последовательностью, то:

$$\sum_{i=0}^{k-1} \varphi(x_k) = k\varphi(x_k) \leq \sum_{i=0}^{k-1} \varphi(x_{i+1})$$



## Анализ сходимости

9. Теперь мы запишем приведенное выше ограничение для всех итераций  $i \in 0, k-1$  и суммируем их:

$$\begin{aligned} \sum_{i=0}^{k-1} [\varphi(x_{i+1}) - \varphi(x^*)] &\leq \frac{1}{2\alpha} [\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2] \\ &\leq \frac{1}{2\alpha} \|x_0 - x^*\|_2^2 \end{aligned}$$

10. Поскольку  $\varphi(x_k)$  является убывающей последовательностью, то:

$$\begin{aligned} \sum_{i=0}^{k-1} \varphi(x_k) = k\varphi(x_k) &\leq \sum_{i=0}^{k-1} \varphi(x_{i+1}) \\ \varphi(x_k) &\leq \frac{1}{k} \sum_{i=0}^{k-1} \varphi(x_{i+1}) \end{aligned}$$

## Анализ сходимости

9. Теперь мы запишем приведенное выше ограничение для всех итераций  $i \in 0, k-1$  и суммируем их:

$$\begin{aligned} \sum_{i=0}^{k-1} [\varphi(x_{i+1}) - \varphi(x^*)] &\leq \frac{1}{2\alpha} [\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2] \\ &\leq \frac{1}{2\alpha} \|x_0 - x^*\|_2^2 \end{aligned}$$

10. Поскольку  $\varphi(x_k)$  является убывающей последовательностью, то:

$$\begin{aligned} \sum_{i=0}^{k-1} \varphi(x_k) &= k\varphi(x_k) \leq \sum_{i=0}^{k-1} \varphi(x_{i+1}) \\ \varphi(x_k) &\leq \frac{1}{k} \sum_{i=0}^{k-1} \varphi(x_{i+1}) \\ \varphi(x_k) - \varphi(x^*) &\leq \frac{1}{k} \sum_{i=0}^{k-1} [\varphi(x_{i+1}) - \varphi(x^*)] \leq \frac{\|x_0 - x^*\|_2^2}{2\alpha k} \end{aligned}$$

## Анализ сходимости

9. Теперь мы запишем приведенное выше ограничение для всех итераций  $i \in 0, k-1$  и суммируем их:

$$\begin{aligned} \sum_{i=0}^{k-1} [\varphi(x_{i+1}) - \varphi(x^*)] &\leq \frac{1}{2\alpha} [\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2] \\ &\leq \frac{1}{2\alpha} \|x_0 - x^*\|_2^2 \end{aligned}$$

10. Поскольку  $\varphi(x_k)$  является убывающей последовательностью, то:

$$\begin{aligned} \sum_{i=0}^{k-1} \varphi(x_k) &= k\varphi(x_k) \leq \sum_{i=0}^{k-1} \varphi(x_{i+1}) \\ \varphi(x_k) &\leq \frac{1}{k} \sum_{i=0}^{k-1} \varphi(x_{i+1}) \\ \varphi(x_k) - \varphi(x^*) &\leq \frac{1}{k} \sum_{i=0}^{k-1} [\varphi(x_{i+1}) - \varphi(x^*)] \leq \frac{\|x_0 - x^*\|_2^2}{2\alpha k} \end{aligned}$$

## Анализ сходимости

9. Теперь мы запишем приведенное выше ограничение для всех итераций  $i \in 0, k-1$  и суммируем их:

$$\begin{aligned} \sum_{i=0}^{k-1} [\varphi(x_{i+1}) - \varphi(x^*)] &\leq \frac{1}{2\alpha} [\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2] \\ &\leq \frac{1}{2\alpha} \|x_0 - x^*\|_2^2 \end{aligned}$$

10. Поскольку  $\varphi(x_k)$  является убывающей последовательностью, то:

$$\begin{aligned} \sum_{i=0}^{k-1} \varphi(x_k) &= k\varphi(x_k) \leq \sum_{i=0}^{k-1} \varphi(x_{i+1}) \\ \varphi(x_k) &\leq \frac{1}{k} \sum_{i=0}^{k-1} \varphi(x_{i+1}) \\ \varphi(x_k) - \varphi(x^*) &\leq \frac{1}{k} \sum_{i=0}^{k-1} [\varphi(x_{i+1}) - \varphi(x^*)] \leq \frac{\|x_0 - x^*\|_2^2}{2\alpha k} \end{aligned}$$

Что является стандартной оценкой  $\frac{L\|x_0 - x^*\|_2^2}{2k}$  с  $\alpha = \frac{1}{L}$ , или, скоростью  $\mathcal{O}\left(\frac{1}{k}\right)$  для гладких выпуклых задач с градиентным спуском!

## Проксимальный градиентный метод. Сильно выпуклый случай

## i Theorem

Рассмотрим проксимальный градиентный метод

$$x_{k+1} = \text{prox}_{\alpha r}(x_k - \alpha \nabla f(x_k))$$

Для критерия  $\varphi(x) = f(x) + r(x)$ , мы предполагаем:

- $f$  является  $\mu$ -сильно выпуклой, дифференцируемой,  $\text{dom}(f) = \mathbb{R}^n$ , и  $\nabla f$  является липшицевой с константой  $L > 0$ .
- $r$  выпукла, и  $\text{prox}_{\alpha r}(x_k) = \arg \min_{x \in \mathbb{R}^n} [\alpha r(x) + \frac{1}{2} \|x - x_k\|_2^2]$  может быть вычислен.

Проксимальный градиентный спуск с фиксированным шагом  $\alpha \leq 1/L$  удовлетворяет

$$\|x_k - x^*\|_2^2 \leq (1 - \alpha\mu)^k \|x_0 - x^*\|_2^2$$

Это точно соответствует скорости сходимости градиентного спуска. Обратите внимание, что исходная задача даже негладкая!

## Доказательство

1. Учитывая расстояние до решения и используя лемму о стационарной точке:



## Доказательство

1. Учитывая расстояние до решения и используя лемму о стационарной точке:

$$\|x_{k+1} - x^*\|_2^2 = \|\text{prox}_{\alpha f}(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2$$





## Доказательство

1. Учитывая расстояние до решения и используя лемму о стационарной точке:

$$\|x_{k+1} - x^*\|_2^2 = \|\text{prox}_{\alpha f}(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2$$

$$\text{лемма о стационарной точке} = \|\text{prox}_{\alpha f}(x_k - \alpha \nabla f(x_k)) - \text{prox}_{\alpha f}(x^* - \alpha \nabla f(x^*))\|_2^2$$



## Доказательство

1. Учитывая расстояние до решения и используя лемму о стационарной точке:

$$\|x_{k+1} - x^*\|_2^2 = \|\text{prox}_{\alpha f}(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2$$

$$\text{лемма о стационарной точке} = \|\text{prox}_{\alpha f}(x_k - \alpha \nabla f(x_k)) - \text{prox}_{\alpha f}(x^* - \alpha \nabla f(x^*))\|_2^2$$

$$\text{нерастяжимость} \leq \|x_k - \alpha \nabla f(x_k) - x^* + \alpha \nabla f(x^*)\|_2^2$$

## Доказательство

1. Учитывая расстояние до решения и используя лемму о стационарной точке:

$$\|x_{k+1} - x^*\|_2^2 = \|\text{prox}_{\alpha f}(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2$$

$$\text{лемма о стационарной точке} = \|\text{prox}_{\alpha f}(x_k - \alpha \nabla f(x_k)) - \text{prox}_{\alpha f}(x^* - \alpha \nabla f(x^*))\|_2^2$$

$$\text{нерастяжимость} \leq \|x_k - \alpha \nabla f(x_k) - x^* + \alpha \nabla f(x^*)\|_2^2$$

$$= \|x_k - x^*\|^2 - 2\alpha \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|_2^2$$



## Доказательство

1. Учитывая расстояние до решения и используя лемму о стационарной точке:

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &= \|\text{prox}_{\alpha f}(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2 \\ \text{лемма о стационарной точке} &= \|\text{prox}_{\alpha f}(x_k - \alpha \nabla f(x_k)) - \text{prox}_{\alpha f}(x^* - \alpha \nabla f(x^*))\|_2^2 \\ \text{нерастяжимость} &\leq \|x_k - \alpha \nabla f(x_k) - x^* + \alpha \nabla f(x^*)\|_2^2 \\ &= \|x_k - x^*\|^2 - 2\alpha \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|_2^2\end{aligned}$$

2. Теперь мы используем гладкость из анализа сходимости и сильную выпуклость:

## Доказательство

1. Учитывая расстояние до решения и используя лемму о стационарной точке:

$$\begin{aligned}
 \|x_{k+1} - x^*\|_2^2 &= \|\text{prox}_{\alpha f}(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2 \\
 \text{лемма о стационарной точке} &= \|\text{prox}_{\alpha f}(x_k - \alpha \nabla f(x_k)) - \text{prox}_{\alpha f}(x^* - \alpha \nabla f(x^*))\|_2^2 \\
 \text{нерастяжимость} &\leq \|x_k - \alpha \nabla f(x_k) - x^* + \alpha \nabla f(x^*)\|_2^2 \\
 &= \|x_k - x^*\|^2 - 2\alpha \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|_2^2
 \end{aligned}$$

2. Теперь мы используем гладкость из анализа сходимости и сильную выпуклость:

$$\text{гладкость} \quad \|\nabla f(x_k) - \nabla f(x^*)\|_2^2 \leq 2L (f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle)$$

## Доказательство

1. Учитывая расстояние до решения и используя лемму о стационарной точке:

$$\begin{aligned}
 \|x_{k+1} - x^*\|_2^2 &= \|\text{prox}_{\alpha f}(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2 \\
 \text{лемма о стационарной точке} &= \|\text{prox}_{\alpha f}(x_k - \alpha \nabla f(x_k)) - \text{prox}_{\alpha f}(x^* - \alpha \nabla f(x^*))\|_2^2 \\
 \text{нерастяжимость} &\leq \|x_k - \alpha \nabla f(x_k) - x^* + \alpha \nabla f(x^*)\|_2^2 \\
 &= \|x_k - x^*\|^2 - 2\alpha \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|_2^2
 \end{aligned}$$

2. Теперь мы используем гладкость из анализа сходимости и сильную выпуклость:

$$\begin{aligned}
 \text{гладкость} \quad \|\nabla f(x_k) - \nabla f(x^*)\|_2^2 &\leq 2L (f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle) \\
 \text{сильная выпуклость} \quad -\langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle &\leq -\left(f(x_k) - f(x^*) + \frac{\mu}{2} \|x_k - x^*\|_2^2\right) - \\
 &\quad -\langle \nabla f(x^*), x_k - x^* \rangle
 \end{aligned}$$

3. Подставим:



3. Подставим:

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &\leq \|x_k - x^*\|^2 - 2\alpha \left( f(x_k) - f(x^*) + \frac{\mu}{2} \|x_k - x^*\|_2^2 \right) - 2\alpha \langle \nabla f(x^*), x_k - x^* \rangle + \\ &\quad + \alpha^2 2L (f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle)\end{aligned}$$





3. Подставим:

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &\leq \|x_k - x^*\|^2 - 2\alpha \left( f(x_k) - f(x^*) + \frac{\mu}{2} \|x_k - x^*\|_2^2 \right) - 2\alpha \langle \nabla f(x^*), x_k - x^* \rangle + \\ &\quad + \alpha^2 2L (f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle) \\ &\leq (1 - \alpha\mu) \|x_k - x^*\|^2 + 2\alpha(\alpha L - 1) (f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle)\end{aligned}$$

3. Подставим:

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &\leq \|x_k - x^*\|^2 - 2\alpha \left( f(x_k) - f(x^*) + \frac{\mu}{2} \|x_k - x^*\|_2^2 \right) - 2\alpha \langle \nabla f(x^*), x_k - x^* \rangle + \\ &\quad + \alpha^2 2L (f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle) \\ &\leq (1 - \alpha\mu) \|x_k - x^*\|^2 + 2\alpha(\alpha L - 1) (f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle)\end{aligned}$$

4. Из выпуклости  $f$ :  $f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle \geq 0$ . Следовательно, если мы используем  $\alpha \leq \frac{1}{L}$ :

$$\|x_{k+1} - x^*\|_2^2 \leq (1 - \alpha\mu) \|x_k - x^*\|^2,$$

что и означает линейную сходимость метода со скоростью не хуже  $1 - \frac{\mu}{L}$ .

## Ускоренный проксимальный градиент – выпуклая функция

### i Ускоренный проксимальный градиентный метод

Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  является **выпуклой** и  $L$ -**гладкой**,  $r : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  является правильной, замкнутой и выпуклой,  $\varphi(x) = f(x) + r(x)$  имеет минимизатор  $x^*$ , и предположим, что  $\text{prox}_{\alpha r}$  легко вычисляется для  $\alpha > 0$ . С любым  $x_0 \in \text{dom } r$  определим последовательность

$$\begin{aligned}t_0 &= 1, & y_0 &= x_0, \\x_k &= \text{prox}_{\frac{1}{L}r}(y_{k-1} - \frac{1}{L}\nabla f(y_{k-1})), \\t_k &= \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2}, \\y_k &= x_k + \frac{t_{k-1} - 1}{t_k}(x_k - x_{k-1}), & k &\geq 1.\end{aligned}$$

Для каждого  $k \geq 1$

$$\varphi(x_k) - \varphi(x^*) \leq \frac{2L \|x_0 - x^*\|_2^2}{(k+1)^2}$$

# Ускоренный проксимальный градиент – $\mu$ -сильно выпуклая функция

## i Ускоренный проксимальный градиентный метод

Добавим, что  $f$  является  $\mu$ -сильно выпуклой ( $\mu > 0$ ).

Установим шаг  $\alpha = \frac{1}{L}$  и фиксированный параметр импульса

$$\beta = \frac{\sqrt{L/\mu} - 1}{\sqrt{L/\mu} + 1}.$$

Генерируем итерации для  $k \geq 0$  (возьмем  $x_{-1} = x_0$ ):

$$\begin{aligned} y_k &= x_k + \beta(x_k - x_{k-1}), \\ x_{k+1} &= \text{prox}_{\alpha r}(y_k - \alpha \nabla f(y_k)). \end{aligned}$$

Для каждого  $k \geq 0$

$$\varphi(x_k) - \varphi(x^*) \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \left(\varphi(x_0) - \varphi(x^*) + \frac{\mu}{2} \|x_0 - x^*\|_2^2\right)$$

## Численные эксперименты

## Квадратичный случай

$$f(x) = \frac{1}{2m} \|Ax - b\|_2^2 + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A \in \mathbb{R}^{m \times n}, \quad \lambda \left( \frac{1}{m} A^T A \right) \in [\mu; L].$$

Linear Least Squares with  $\ell_1$  Regularization (LASSO).

$m=1000$ ,  $n=100$ ,  $\lambda=0$ ,  $\mu=0$ ,  $L=10$ . Optimal sparsity: 0.0e+00



Рис. 2: Гладкий выпуклый случай. Сублинейная сходимость, отсутствие сходимости в области, нет разницы между методом субградиента и проксимальным методом.

## Квадратичный случай

$$f(x) = \frac{1}{2m} \|Ax - b\|_2^2 + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A \in \mathbb{R}^{m \times n}, \quad \lambda \left( \frac{1}{m} A^T A \right) \in [\mu; L].$$

Linear Least Squares with  $\ell_1$  Regularization (LASSO).  
m=1000, n=100,  $\lambda=1$ ,  $\mu=0$ ,  $L=10$ . Optimal sparsity: 2.3e-01



Рис. 3: Негладкий выпуклый случай. Сублинейная сходимость. В начале метод субградиента и проксимальный метод близки.

## Квадратичный случай

$$f(x) = \frac{1}{2m} \|Ax - b\|_2^2 + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A \in \mathbb{R}^{m \times n}, \quad \lambda \left( \frac{1}{m} A^T A \right) \in [\mu; L].$$

Linear Least Squares with  $\ell_1$  Regularization (LASSO).  
m=1000, n=100,  $\lambda=1$ ,  $\mu=0$ ,  $L=10$ . Optimal sparsity: 2.3e-01



Рис. 4: Негладкий выпуклый случай. Если мы возьмем больше итераций, то проксимальный метод сходится с постоянным шагом, что не так для метода субградиента. Разница огромна, в то время как сложность итерации одинакова.



# Бинарная логистическая регрессия

$$f(x) = \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-b_i(A_i x))) + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A_i \in \mathbb{R}^n, \quad b_i \in \{-1, 1\}$$

Binary Logistic Regression with  $\ell_1$  Regularization.  
m=300, n=50,  $\lambda=0.1$ . Optimal sparsity: 8.6e-01

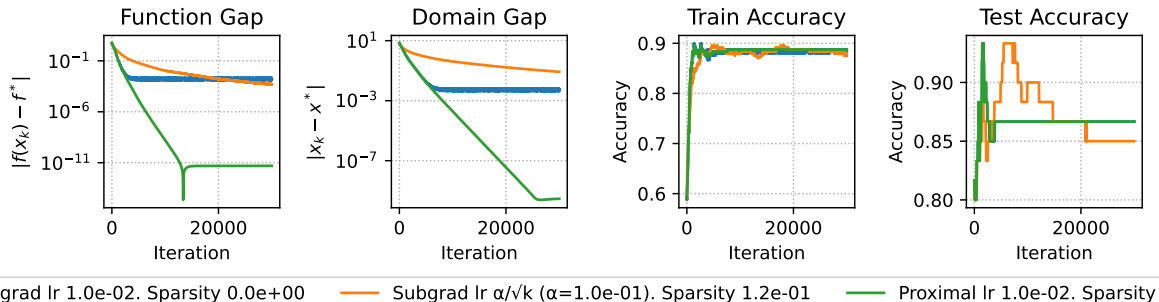
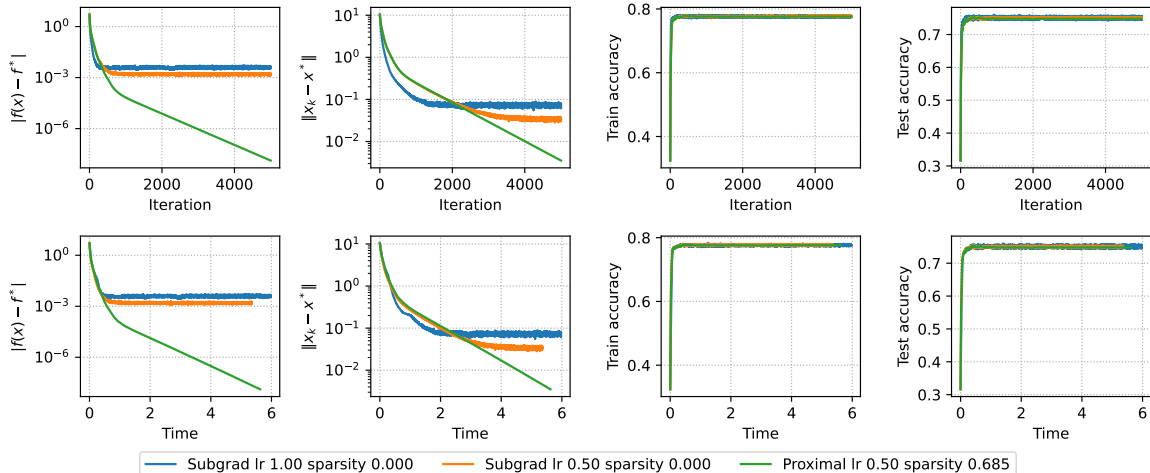


Рис. 5: Логистическая регрессия с  $\ell_1$ -регуляризацией

# Softmax multiclass regression

Convex multiclass regression. lam=0.01.



## Пример: ISTA

### Iterative Shrinkage-Thresholding Algorithm (ISTA)

ISTA является популярным методом для решения задач оптимизации с  $\ell_1$ -регуляризацией, такой как Lasso. Он объединяет градиентный спуск с оператором сжатия для эффективного управления негладким  $\ell_1$ -штрафом.

- **Алгоритм:**

# Пример: ISTA

## Iterative Shrinkage-Thresholding Algorithm (ISTA)

ISTA является популярным методом для решения задач оптимизации с  $\ell_1$ -регуляризацией, такой как Lasso. Он объединяет градиентный спуск с оператором сжатия для эффективного управления негладким  $\ell_1$ -штрафом.

- **Алгоритм:**

- Дано  $x_0$ , для  $k \geq 0$ , повторять:

$$x_{k+1} = \text{prox}_{\lambda\alpha\|\cdot\|_1}(x_k - \alpha\nabla f(x_k)),$$

где  $\text{prox}_{\lambda\alpha\|\cdot\|_1}(v)$  применяет оператор сжатия к каждому компоненту  $v$ .

# Пример: ISTA

## Iterative Shrinkage-Thresholding Algorithm (ISTA)

ISTA является популярным методом для решения задач оптимизации с  $\ell_1$ -регуляризацией, такой как Lasso. Он объединяет градиентный спуск с оператором сжатия для эффективного управления негладким  $\ell_1$ -штрафом.

- **Алгоритм:**

- Дано  $x_0$ , для  $k \geq 0$ , повторять:

$$x_{k+1} = \text{prox}_{\lambda\alpha\|\cdot\|_1}(x_k - \alpha\nabla f(x_k)),$$

где  $\text{prox}_{\lambda\alpha\|\cdot\|_1}(v)$  применяет оператор сжатия к каждому компоненту  $v$ .

- **Сходимость:**

# Пример: ISTA

## Iterative Shrinkage-Thresholding Algorithm (ISTA)

ISTA является популярным методом для решения задач оптимизации с  $\ell_1$ -регуляризацией, такой как Lasso. Он объединяет градиентный спуск с оператором сжатия для эффективного управления негладким  $\ell_1$ -штрафом.

- **Алгоритм:**

- Дано  $x_0$ , для  $k \geq 0$ , повторять:

$$x_{k+1} = \text{prox}_{\lambda\alpha\|\cdot\|_1}(x_k - \alpha\nabla f(x_k)),$$

где  $\text{prox}_{\lambda\alpha\|\cdot\|_1}(v)$  применяет оператор сжатия к каждому компоненту  $v$ .

- **Сходимость:**

- Сходится со скоростью  $O(1/k)$  для подходящего шага  $\alpha$ .

# Пример: ISTA

## Iterative Shrinkage-Thresholding Algorithm (ISTA)

ISTA является популярным методом для решения задач оптимизации с  $\ell_1$ -регуляризацией, такой как Lasso. Он объединяет градиентный спуск с оператором сжатия для эффективного управления негладким  $\ell_1$ -штрафом.

- **Алгоритм:**

- Дано  $x_0$ , для  $k \geq 0$ , повторять:

$$x_{k+1} = \text{prox}_{\lambda\alpha\|\cdot\|_1}(x_k - \alpha\nabla f(x_k)),$$

где  $\text{prox}_{\lambda\alpha\|\cdot\|_1}(v)$  применяет оператор сжатия к каждому компоненту  $v$ .

- **Сходимость:**

- Сходится со скоростью  $O(1/k)$  для подходящего шага  $\alpha$ .

- **Применение:**

# Пример: ISTA

## Iterative Shrinkage-Thresholding Algorithm (ISTA)

ISTA является популярным методом для решения задач оптимизации с  $\ell_1$ -регуляризацией, такой как Lasso. Он объединяет градиентный спуск с оператором сжатия для эффективного управления негладким  $\ell_1$ -штрафом.

- **Алгоритм:**

- Дано  $x_0$ , для  $k \geq 0$ , повторять:

$$x_{k+1} = \text{prox}_{\lambda\alpha\|\cdot\|_1}(x_k - \alpha\nabla f(x_k)),$$

где  $\text{prox}_{\lambda\alpha\|\cdot\|_1}(v)$  применяет оператор сжатия к каждому компоненту  $v$ .

- **Сходимость:**

- Сходится со скоростью  $O(1/k)$  для подходящего шага  $\alpha$ .

- **Применение:**

- Эффективно для восстановления разреженных сигналов, обработки изображений и compressed sensing.



## Пример: FISTA

### Fast Iterative Shrinkage-Thresholding Algorithm (FISTA)

FISTA улучшает сходимость ISTA, включая в неё импульсное слагаемое, вдохновленное методом Нестерова.

- **Алгоритм:**

# Пример: FISTA

## Fast Iterative Shrinkage-Thresholding Algorithm (FISTA)

FISTA улучшает сходимость ISTA, включая в неё импульсное слагаемое, вдохновленное методом Нестерова.

- **Алгоритм:**
  - Инициализируем  $x_0 = y_0$ ,  $t_0 = 1$ .

# Пример: FISTA

## Fast Iterative Shrinkage-Thresholding Algorithm (FISTA)

FISTA улучшает сходимость ISTA, включая в неё импульсное слагаемое, вдохновленное методом Нестерова.

- **Алгоритм:**

- Инициализируем  $x_0 = y_0$ ,  $t_0 = 1$ .
- Для  $k \geq 1$ , обновляем:

$$x_k = \text{prox}_{\lambda\alpha\|\cdot\|_1}(y_{k-1} - \alpha\nabla f(y_{k-1})),$$

$$t_k = \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2},$$

$$y_k = x_k + \frac{t_{k-1} - 1}{t_k}(x_k - x_{k-1}).$$

# Пример: FISTA

## Fast Iterative Shrinkage-Thresholding Algorithm (FISTA)

FISTA улучшает сходимость ISTA, включая в неё импульсное слагаемое, вдохновленное методом Нестерова.

- **Алгоритм:**

- Инициализируем  $x_0 = y_0$ ,  $t_0 = 1$ .
- Для  $k \geq 1$ , обновляем:

$$x_k = \text{prox}_{\lambda\alpha\|\cdot\|_1}(y_{k-1} - \alpha\nabla f(y_{k-1})),$$

$$t_k = \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2},$$

$$y_k = x_k + \frac{t_{k-1} - 1}{t_k}(x_k - x_{k-1}).$$

- **Сходимость:**

# Пример: FISTA

## Fast Iterative Shrinkage-Thresholding Algorithm (FISTA)

FISTA улучшает сходимость ISTA, включая в неё импульсное слагаемое, вдохновленное методом Нестерова.

- **Алгоритм:**

- Инициализируем  $x_0 = y_0$ ,  $t_0 = 1$ .
- Для  $k \geq 1$ , обновляем:

$$x_k = \text{prox}_{\lambda\alpha\|\cdot\|_1}(y_{k-1} - \alpha\nabla f(y_{k-1})),$$

$$t_k = \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2},$$

$$y_k = x_k + \frac{t_{k-1} - 1}{t_k}(x_k - x_{k-1}).$$

- **Сходимость:**

- Улучшает скорость сходимости до  $O(1/k^2)$ .

# Пример: FISTA

## Fast Iterative Shrinkage-Thresholding Algorithm (FISTA)

FISTA улучшает сходимость ISTA, включая в неё импульсное слагаемое, вдохновленное методом Нестерова.

- **Алгоритм:**

- Инициализируем  $x_0 = y_0$ ,  $t_0 = 1$ .
- Для  $k \geq 1$ , обновляем:

$$x_k = \text{prox}_{\lambda\alpha\|\cdot\|_1}(y_{k-1} - \alpha\nabla f(y_{k-1})),$$

$$t_k = \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2},$$

$$y_k = x_k + \frac{t_{k-1} - 1}{t_k}(x_k - x_{k-1}).$$

- **Сходимость:**

- Улучшает скорость сходимости до  $O(1/k^2)$ .

- **Применение:**

# Пример: FISTA

## Fast Iterative Shrinkage-Thresholding Algorithm (FISTA)

FISTA улучшает сходимость ISTA, включая в неё импульсное слагаемое, вдохновленное методом Нестерова.

- **Алгоритм:**

- Инициализируем  $x_0 = y_0$ ,  $t_0 = 1$ .
- Для  $k \geq 1$ , обновляем:

$$x_k = \text{prox}_{\lambda\alpha\|\cdot\|_1}(y_{k-1} - \alpha\nabla f(y_{k-1})),$$

$$t_k = \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2},$$

$$y_k = x_k + \frac{t_{k-1} - 1}{t_k}(x_k - x_{k-1}).$$

- **Сходимость:**

- Улучшает скорость сходимости до  $O(1/k^2)$ .

- **Применение:**

- Особенно полезен для больших задач в машинном обучении и обработке сигналов, где  $\ell_1$ -штраф индуцирует разреженность.

# Пример: задача восстановления матрицы (Matrix Completion)

## Решение задачи Matrix Completion

Задачи matrix completion стремятся заполнить пропущенные элементы частично наблюдаемой матрицы при определенных предположениях, обычно низкого ранга. Это может быть сформулировано в виде задачи минимизации, включающую ядерную норму (сумму сингулярных значений), которая продвигает решения низкого ранга.

- **Формулировка задачи:**

$$\min_X \frac{1}{2} \|P_\Omega(X) - P_\Omega(M)\|_F^2 + \lambda \|X\|_*,$$

где  $P_\Omega$  проецирует на наблюдаемое множество  $\Omega$ , и  $\|\cdot\|_*$  обозначает ядерную норму.



# Пример: задача восстановления матрицы (Matrix Completion)

## Решение задачи Matrix Completion

Задачи matrix completion стремятся заполнить пропущенные элементы частично наблюдаемой матрицы при определенных предположениях, обычно низкого ранга. Это может быть сформулировано в виде задачи минимизации, включающую ядерную норму (сумму сингулярных значений), которая продвигает решения низкого ранга.

- **Формулировка задачи:**

$$\min_X \frac{1}{2} \|P_\Omega(X) - P_\Omega(M)\|_F^2 + \lambda \|X\|_*,$$

где  $P_\Omega$  проецирует на наблюдаемое множество  $\Omega$ , и  $\|\cdot\|_*$  обозначает ядерную норму.

- **Проксимальный оператор:**

# Пример: задача восстановления матрицы (Matrix Completion)

## Решение задачи Matrix Completion

Задачи matrix completion стремятся заполнить пропущенные элементы частично наблюдаемой матрицы при определенных предположениях, обычно низкого ранга. Это может быть сформулировано в виде задачи минимизации, включающую ядерную норму (сумму сингулярных значений), которая продвигает решения низкого ранга.

- **Формулировка задачи:**

$$\min_X \frac{1}{2} \|P_\Omega(X) - P_\Omega(M)\|_F^2 + \lambda \|X\|_*,$$

где  $P_\Omega$  проецирует на наблюдаемое множество  $\Omega$ , и  $\|\cdot\|_*$  обозначает ядерную норму.

- **Проксимальный оператор:**
  - Проксимальный оператор для ядерной нормы включает сингулярное разложение (SVD) и сжатие сингулярных значений.

# Пример: задача восстановления матрицы (Matrix Completion)

## Решение задачи Matrix Completion

Задачи matrix completion стремятся заполнить пропущенные элементы частично наблюдаемой матрицы при определенных предположениях, обычно низкого ранга. Это может быть сформулировано в виде задачи минимизации, включающую ядерную норму (сумму сингулярных значений), которая продвигает решения низкого ранга.

- **Формулировка задачи:**

$$\min_X \frac{1}{2} \|P_\Omega(X) - P_\Omega(M)\|_F^2 + \lambda \|X\|_*,$$

где  $P_\Omega$  проецирует на наблюдаемое множество  $\Omega$ , и  $\|\cdot\|_*$  обозначает ядерную норму.

- **Проксимальный оператор:**
  - Проксимальный оператор для ядерной нормы включает сингулярное разложение (SVD) и сжатие сингулярных значений.
- **Алгоритм:**

# Пример: задача восстановления матрицы (Matrix Completion)

## Решение задачи Matrix Completion

Задачи matrix completion стремятся заполнить пропущенные элементы частично наблюдаемой матрицы при определенных предположениях, обычно низкого ранга. Это может быть сформулировано в виде задачи минимизации, включающую ядерную норму (сумму сингулярных значений), которая продвигает решения низкого ранга.

- **Формулировка задачи:**

$$\min_X \frac{1}{2} \|P_\Omega(X) - P_\Omega(M)\|_F^2 + \lambda \|X\|_*,$$

где  $P_\Omega$  проецирует на наблюдаемое множество  $\Omega$ , и  $\|\cdot\|_*$  обозначает ядерную норму.

- **Проксимальный оператор:**
  - Проксимальный оператор для ядерной нормы включает сингулярное разложение (SVD) и сжатие сингулярных значений.
- **Алгоритм:**
  - Можно применять аналогичные проксимальные методы или ускоренные проксимальные методы; основной вычислительный расход приходится на выполнение SVD.

# Пример: задача восстановления матрицы (Matrix Completion)

## Решение задачи Matrix Completion

Задачи matrix completion стремятся заполнить пропущенные элементы частично наблюдаемой матрицы при определенных предположениях, обычно низкого ранга. Это может быть сформулировано в виде задачи минимизации, включающую ядерную норму (сумму сингулярных значений), которая продвигает решения низкого ранга.

- **Формулировка задачи:**

$$\min_X \frac{1}{2} \|P_\Omega(X) - P_\Omega(M)\|_F^2 + \lambda \|X\|_*,$$

где  $P_\Omega$  проецирует на наблюдаемое множество  $\Omega$ , и  $\|\cdot\|_*$  обозначает ядерную норму.

- **Проксимальный оператор:**
  - Проксимальный оператор для ядерной нормы включает сингулярное разложение (SVD) и сжатие сингулярных значений.
- **Алгоритм:**
  - Можно применять аналогичные проксимальные методы или ускоренные проксимальные методы; основной вычислительный расход приходится на выполнение SVD.
- **Применение:**

# Пример: задача восстановления матрицы (Matrix Completion)

## Решение задачи Matrix Completion

Задачи matrix completion стремятся заполнить пропущенные элементы частично наблюдаемой матрицы при определенных предположениях, обычно низкого ранга. Это может быть сформулировано в виде задачи минимизации, включающую ядерную норму (сумму сингулярных значений), которая продвигает решения низкого ранга.

- **Формулировка задачи:**

$$\min_X \frac{1}{2} \|P_\Omega(X) - P_\Omega(M)\|_F^2 + \lambda \|X\|_*,$$

где  $P_\Omega$  проецирует на наблюдаемое множество  $\Omega$ , и  $\|\cdot\|_*$  обозначает ядерную норму.

- **Проксимальный оператор:**
  - Проксимальный оператор для ядерной нормы включает сингулярное разложение (SVD) и сжатие сингулярных значений.
- **Алгоритм:**
  - Можно применять аналогичные проксимальные методы или ускоренные проксимальные методы; основной вычислительный расход приходится на выполнение SVD.
- **Применение:**
  - Широко используется в рекомендательных системах, восстановлении изображений и других областях, где данные естественно представлены в виде матриц, но частично наблюдаемы.

## Summary

- Если использовать структуру задачи, можно превзойти нижние оценки для неструктурированной постановки.

## Summary

- Если использовать структуру задачи, можно превзойти нижние оценки для неструктурированной постановки.
- Проксимальный метод для задачи с  $L$ -гладкой выпуклой функцией  $f$  и выпуклой функцией  $r$  с вычислимым проксимальным оператором имеет ту же скорость сходимости, что и метод градиентного спуска для  $f$ . Свойства гладкости/негладкости  $r$  на сходимость не влияют.



## Summary

- Если использовать структуру задачи, можно превзойти нижние оценки для неструктурированной постановки.
- Проксимальный метод для задачи с  $L$ -гладкой выпуклой функцией  $f$  и выпуклой функцией  $r$  с вычислимым проксимальным оператором имеет ту же скорость сходимости, что и метод градиентного спуска для  $f$ . Свойства гладкости/негладкости  $r$  на сходимость не влияют.
- Кажется, что если  $f = 0$ , то любая негладкая задача может быть решена таким методом. Вопрос: это правда?

## Summary

- Если использовать структуру задачи, можно превзойти нижние оценки для неструктурированной постановки.
- Проксимальный метод для задачи с  $L$ -гладкой выпуклой функцией  $f$  и выпуклой функцией  $r$  с вычислимым проксимальным оператором имеет ту же скорость сходимости, что и метод градиентного спуска для  $f$ . Свойства гладкости/негладкости  $r$  на сходимость не влияют.
- Кажется, что если  $f = 0$ , то любая негладкая задача может быть решена таким методом. Вопрос: это правда?

## Summary

- Если использовать структуру задачи, можно превзойти нижние оценки для неструктурированной постановки.
- Проксимальный метод для задачи с  $L$ -гладкой выпуклой функцией  $f$  и выпуклой функцией  $r$  с вычислимым проксимальным оператором имеет ту же скорость сходимости, что и метод градиентного спуска для  $f$ . Свойства гладкости/негладкости  $r$  на сходимость не влияют.
- Кажется, что если  $f = 0$ , то любая негладкая задача может быть решена таким методом. Вопрос: это правда?

Если разрешить численно неточный проксимальный оператор, то действительно можно решать любую негладкую задачу оптимизации. Но с теоретической точки зрения это не лучше субградиентного спуска, поскольку для решения проксимальной подзадачи используется вспомогательный метод (например, тот же субградиентный спуск).

- Проксимальный метод является общим современным фреймворком для многих численных методов. Далее развиваются ускоренные, стохастические, приближенные двойственные методы и т.д.

# Summary

- Если использовать структуру задачи, можно превзойти нижние оценки для неструктурированной постановки.
- Проксимальный метод для задачи с  $L$ -гладкой выпуклой функцией  $f$  и выпуклой функцией  $r$  с вычислимым проксимальным оператором имеет ту же скорость сходимости, что и метод градиентного спуска для  $f$ . Свойства гладкости/негладкости  $r$  на сходимость не влияют.
- Кажется, что если  $f = 0$ , то любая негладкая задача может быть решена таким методом. Вопрос: это правда?

Если разрешить численно неточный проксимальный оператор, то действительно можно решать любую негладкую задачу оптимизации. Но с теоретической точки зрения это не лучше субградиентного спуска, поскольку для решения проксимальной подзадачи используется вспомогательный метод (например, тот же субградиентный спуск).

- Проксимальный метод является общим современным фреймворком для многих численных методов. Далее развиваются ускоренные, стохастические, приближенные двойственные методы и т.д.
- Дополнительные материалы: разбиение по проксимальному оператору, схема Дугласа—Рачфорда, задача наилучшего приближения, разбиение на три оператора.