# Негладкая оптимизация. Проксимальный метод

## Методы оптимизации

Александр Безносиков

Московский физико-технический институт

23 ноября 2023

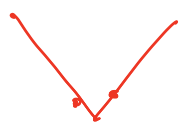
# Негладкие задачи

- **Вопрос:** функция $f(x) = |x|$ выпукла?

# Негладкие задачи

- **Вопрос:** функция $f(x) = |x|$ выпукла? Безусловно. А дифференцируемая и гладкая?

$$\|\nabla f(x) - \nabla f(y)\| = 2 \qquad \cancel{L\|x-y\|}$$



# Негладкие задачи

- **Вопрос:** функция $f(x) = |x|$ выпуклa? Безусловно. А дифференцируемая и гладкая? Нет.
- Получается, что даже довольно простые выпуклые задачи могут быть негладким. До этого мы смотрели только на гладкие задачи.

# Негладкие задачи

- **Вопрос:** функция $f(x) = |x|$ выпукла? Безусловно. А дифференцируемая и гладкая? Нет.
- Получается, что даже довольно простые выпуклые задачи могут быть негладким. До этого мы смотрели только на гладкие задачи.
- Будем рассматривать следующее предположение вместо гладкости (Липшицевости градиента):

**Определение $M$-Липшецевой функции**

Пусть дана функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что она является $M$-Липшицева, если для любых $x, y \in \mathbb{R}^d$ выполнено

$$\underline{|f(x) - f(y)|} \le M \left( \|x - y\|_2 \right).$$

# Негладкие задачи

- **Вопрос:** функция $f(x) = |x|$ выпукла? Безусловно. А дифференцируемая и гладкая? Нет.
- Получается, что даже довольно простые выпуклые задачи могут быть негладким. До этого мы смотрели только на гладкие задачи.
- Будем рассматривать следующее предположение вместо гладкости (Липшицевости градиента):

### Определение $M$-Липшецевой функции

Пусть дана функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что она является $M$-Липшицева, если для любых $x, y \in \mathbb{R}^d$ выполнено

$$|f(x) - f(y)| \leq M\|x - y\|_2.$$

Понятие (и все результаты далее) можно перенести на некоторое ограниченное выпуклое множество $\mathcal{X}$. Связано этом в том числе с тем, что не бывает сильно выпуклых и Липшецевых на $\mathbb{R}^d$ функций.
**Вопрос:** почему?

# Негладкие задачи

- **Вопрос:** функция $f(x) = |x|$ выпукла? Безусловно. А дифференцируемая и гладкая? Нет.
- Получается, что даже довольно простые выпуклые задачи могут быть негладким. До этого мы смотрели только на гладкие задачи.
- Будем рассматривать следующее предположение вместо гладкости (Липшицевости градиента):

**Определение $M$-Липшецевой функции**

Пусть дана функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что она является $M$-Липшицева, если для любых $x, y \in \mathbb{R}^d$ выполнено

$$|f(x) - f(y)| \leq M\|x - y\|_2.$$

Понятие (и все результаты далее) можно перенести на некоторое ограниченное выпуклое множество $\mathcal{X}$. Связано этом в том числе с тем, что не бывает сильно выпуклых и Липшецевых на $\mathbb{R}^d$ функций. **Вопрос:** почему? Линейный и квадратичный рост не сочетаются.

# Субградиент и субдифференциал

Если функция не дифференцируема ~~с~~ в точк~~и~~е, а значит градиента нет. Что может существовать вместо градиент?

# Субградиент и субдифференциал

Если функция не дифференцируема с точки, а значит градиента нет. Что может существовать вместо градиент?

## Субградиент и субдифференциал

Пусть дана выпуклая функция $f : \mathbb{R}^d \to \mathbb{R}$. Вектор $g$ будем называть субградиентом этой функции $f$ в точке $x \in \mathbb{R}^d$, если для любого $y \in \mathbb{R}^d$ выполняется:

$$f(y) \geq f(x) + \langle \overset{\partial f(x)}{g}, y - x \rangle.$$

Множество $\partial f(x)$ всех субградиентов $f$ в $x$ будем называть субдифференциалом.

# Условие оптимальности

**Теорема (условие оптимальности)**

$x^*$ – минимум выпуклой функции $f$ тогда и только тогда, когда

$$0 \in \partial f(x^*).$$

$\Leftarrow \quad 0 \in \partial f(x^*)$

$f(x) \geq f(x^*) + \langle g ; x - x^* \rangle = f(x^*)$

$g \in \partial f(x^*) \qquad g = 0$

$\Rightarrow \quad f(x) \geq f(x^*) \quad \forall x$

$f(x) \geq f(x^*) + \langle 0 ; x - x^* \rangle \quad \Rightarrow \quad 0 \in \partial f(x^*)$

$\forall x$

# Условие оптимальности

**Теорема (условие оптимальности)**

$x^*$ – минимум выпуклой функции $f$ тогда и только тогда, когда

$$0 \in \partial f(x^*).$$

<u>Доказательство:</u>

$\Leftarrow$ Если $0 \in \partial f(x^*)$, то по выпуклости и определению субградиента: $f(x) \geq f(x^*) + \langle 0, x - x^* \rangle = f(x^*)$. Доказано по определению глобального минимума.

# Условие оптимальности

## Теорема (условие оптимальности)

$x^*$ – минимум выпуклой функции $f$ тогда и только тогда, когда

$$0 \in \partial f(x^*).$$

Доказательство:

$\Leftarrow$ Если $0 \in \partial f(x^*)$, то по выпуклости и определению субградиента: $f(x) \geq f(x^*) + \langle 0, x - x^* \rangle = f(x^*)$. Доказано по определению глобального минимума.

$\Rightarrow$ Если $f(x) \geq f(x^*)$ для любых $x \in \mathbb{R}^d$, то для вектора 0 выполнено $f(x) \geq f(x^*) + \langle 0, x - x^* \rangle$ для любого $x \in \mathbb{R}^d$. Доказано по определению субградиента.

# Свойство $M$-Липшицевой функции

**Лемма (свойство $M$-Липшицевой функции)**

Пусть дана выпуклая функция $f : \mathbb{R}^d \to \mathbb{R}$. Тогда функция $f$ является $M$-Липшицевой тогда и только тогда, когда для любого $x \in \mathbb{R}^d$ и $g \in \partial f(x)$ имеем $\|g\|_2 \le M$.

# Доказательство

$\Rightarrow$ Пусть дополнительно к выпуклости функция $f$ еще и $M$-Липшицева, тогда

$$f(x) \geq f(y) + \langle g; x-y \rangle \qquad g \in \partial f(y) \quad \forall x \in \mathbb{R}^d$$

$$x = y + g$$

$$f(x) \geq f(y) + \langle g; g \rangle = f(y) + \|g\|_2^2$$

$$\|g\|_2^2 \leq f(x) - f(y) \leq M\|x-y\|_2 = M\|g\|_2$$

$$\|g\|_2 \leq M$$

# Доказательство

$\Rightarrow$ Пусть дополнительно к выпуклости функция $f$ еще и $M$-Липшицева, тогда

- Рассмотрим $g \in \partial f(x)$, тогда по выпуклости и определению субградиента для любого $y \in \mathbb{R}^d$:

$$f(y) - f(x) \geq \langle g, y - x \rangle.$$

# Доказательство

$\Rightarrow$ Пусть дополнительно к выпуклости функция $f$ еще и $M$-Липшицева, тогда

- Рассмотрим $g \in \partial f(x)$, тогда по выпуклости и определению субградиента для любого $y \in \mathbb{R}^d$:

$$f(y) - f(x) \geq \langle g, y - x \rangle.$$

- Из Липшицевости $f$:

$$M\|y - x\|_2 \geq f(y) - f(x) \geq \langle g, y - x \rangle.$$

# Доказательство

$\Rightarrow$ Пусть дополнительно к выпуклости функция $f$ еще и $M$-Липшицева, тогда

- Рассмотрим $g \in \partial f(x)$, тогда по выпуклости и определению субградиента для любого $y \in \mathbb{R}^d$:

$$f(y) - f(x) \geq \langle g, y - x \rangle.$$

- Из Липшицевости $f$:

$$M\|y - x\|_2 \geq f(y) - f(x) \geq \langle g, y - x \rangle.$$

- Возьмем $y = g + x$, тогда

$$M\|g\|_2 = M\|y - x\|_2 \geq \langle g, y - x \rangle = \|g\|_2^2.$$

# Доказательство

$\Rightarrow$ Пусть дополнительно к выпуклости функция $f$ еще и $M$-Липшицева, тогда

- Рассмотрим $g \in \partial f(x)$, тогда по выпуклости и определению субградиента для любого $y \in \mathbb{R}^d$:

$$f(y) - f(x) \geq \langle g, y - x \rangle.$$

- Из Липшицевости $f$:

$$M\|y - x\|_2 \geq f(y) - f(x) \geq \langle g, y - x \rangle.$$

- Возьмем $y = g + x$, тогда

$$M\|g\|_2 = M\|y - x\|_2 \geq \langle g, y - x \rangle = \|g\|_2^2.$$

Что и требовалось.

# Доказательство

$\Leftarrow$ Пусть дополнительно к выпуклости у функции $f$ все субградиенты равномерно ограничены: $\|g\|_2 \leq M$ для любого $x \in R^d$ и $g \in \partial f(x)$. Тогда

$$f(x) \geq f(y) - \langle g; y-x \rangle \qquad g \in \partial f(y)$$

$$\langle g; y-x \rangle \geq f(y) - f(x)$$

$$\|g\|_2 \|y-x\|_2$$

$$f(y) - f(x) \leq \|g\|_2 \|y-x\|_2$$

$$f(x) - f(y) \leq \|g\|_2 \|y-x\|$$

# Доказательство

$\Leftarrow$ Пусть дополнительно к выпуклости у функции $f$ все субградиенты равномерно ограничены: $\|g\|_2 \leq M$ для любого $x \in R^d$ и $g \in \partial f(x)$. Тогда

- Рассмотрим $g \in \partial f(x)$, тогда по выпуклости и определению субградиента для любого $y \in \mathbb{R}^d$:

$$f(y) - f(x) \leq \langle g, x - y \rangle.$$

# Доказательство

$\Leftarrow$ Пусть дополнительно к выпуклости у функции $f$ все субградиенты равномерно ограничены: $\|g\|_2 \leq M$ для любого $x \in R^d$ и $g \in \partial f(x)$. Тогда

- Рассмотрим $g \in \partial f(x)$, тогда по выпуклости и определению субградиента для любого $y \in \mathbb{R}^d$:

$$f(y) - f(x) \leq \langle g, x - y \rangle.$$

- КБШ:

$$f(y) - f(x) \leq \|g\|_2 \cdot \|x - y\|_2.$$

# Доказательство

$\Leftarrow$ Пусть дополнительно к выпуклости у функции $f$ все субградиенты равномерно ограничены: $\|g\|_2 \leq M$ для любого $x \in R^d$ и $g \in \partial f(x)$. Тогда

- Рассмотрим $g \in \partial f(x)$, тогда по выпуклости и определению субградиента для любого $y \in \mathbb{R}^d$:

$$f(y) - f(x) \leq \langle g, x - y \rangle.$$

- КБШ:

$$f(y) - f(x) \leq \|g\|_2 \cdot \|x - y\|_2.$$

- Пользуемся предположением и получаем:

$$f(y) - f(x) \leq M\|x - y\|_2.$$

# Доказательство

$\Leftarrow$ Пусть дополнительно к выпуклости у функции $f$ все субградиенты равномерно ограничены: $\|g\|_2 \leq M$ для любого $x \in R^d$ и $g \in \partial f(x)$. Тогда

- Рассмотрим $g \in \partial f(x)$, тогда по выпуклости и определению субградиента для любого $y \in \mathbb{R}^d$:

$$f(y) - f(x) \leq \langle g, x - y \rangle.$$

- КБШ:

$$f(y) - f(x) \leq \|g\|_2 \cdot \|x - y\|_2.$$

- Пользуемся предположением и получаем:

$$f(y) - f(x) \leq M\|x - y\|_2.$$

Что и требовалось.

# Субградиентный метод

- Рассматриваем задачу:
$$\min_{x \in \mathbb{R}^d} f(x),$$
где $f$ выпуклая и $M$-Липшицева.

# Субградиентный метод

- Рассматриваем задачу:
$$\min_{x \in \mathbb{R}^d} f(x),$$
где $f$ выпуклая и $M$-Липшицева.
- Простая идея – вместо градиента использовать какой-то субградиент в текущей точке:

**Алгоритм 2** Субградиентный метод

**Вход:** размеры шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций $K$

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $g^k \in \partial f(x^k)$
3: $x^{k+1} = x^k - \gamma g^k$
4: **end for**

**Выход:** $\frac{1}{K}\sum_{k=0}^{K-1} x^k$

- Ничего сверхъестественного:

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 &= \|x^k - \gamma g^k - x^*\|_2^2 \quad \in \partial f(x^k) \\
&= \underline{\|x^k - x^*\|_2^2} - \underbrace{2\gamma\langle g^k, x^k - x^*\rangle}_{2)\ +2\gamma(f(x^*) - f(x^k))} + \underbrace{\gamma^2\|g^k\|_2^2}_{\gamma^2 M^2}
\end{aligned}
$$

1) $\|g^k\|_2^2 \le M^2$

$$\|x^{k+1} - x^*\|_2^2 = \|x^k - x^*\|_2^2 + 2\gamma(f(x^*) - f(x^k)) + \gamma^2 M^2$$

$$2\gamma(f(x^k) - f(x^*)) = \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2 + \gamma^2 M^2$$

$\sum_k \quad \frac{1}{K}$

$$\frac{2\gamma}{K}\sum f(x^k) - f(x^*) \le \frac{\|x^0 - x^*\|_2^2 - \cancel{\|x^K - x^*\|_2^2}}{K} + \gamma^2 M^2$$

# Доказательство сходимости

- Ничего сверхъестественного:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 =& \|x^k - \gamma g^k - x^*\|_2^2 \\ =& \|x^k - x^*\|_2^2 - 2\gamma\langle g^k, x^k - x^*\rangle + \gamma^2\|g^k\|_2^2\end{aligned}$$

- Из $M$-Липшицевости $f$ следует, что субградиентый ограничены:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\gamma\langle g^k, x^k - x^*\rangle + \gamma^2 M^2$$

# Доказательство сходимости

- Ничего сверхъестественного:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 =& \|x^k - \gamma g^k - x^*\|_2^2 \\ =& \|x^k - x^*\|_2^2 - 2\gamma\langle g^k, x^k - x^*\rangle + \gamma^2\|g^k\|_2^2\end{aligned}$$

- Из $M$-Липшицевости $f$ следует, что субградиентый ограничены:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\gamma\langle g^k, x^k - x^*\rangle + \gamma^2 M^2$$

- Из выпуклости и определения субградиента:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\gamma(f(x^k) - f(x^*)) + \gamma^2 M^2$$

# Доказательство сходимости

- С предыдущего слайда:

$$2\gamma(f(x^k) - f(x^*)) \leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2 + \gamma^2 M^2$$

# Доказательство сходимости

- С предыдущего слайда:

$$2\gamma(f(x^k) - f(x^*)) \leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2 + \gamma^2 M^2$$

- Суммируем по всем $k$ и усредняем:

$$\frac{2\gamma}{K} \sum_{k=0}^{K-1} (f(x^k) - f(x^*)) \leq \frac{\|x^0 - x^*\|_2^2}{K} + \gamma^2 M^2$$

# Доказательство сходимости

- С предыдущего слайда:

$$2\gamma(f(x^k) - f(x^*)) \leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2 + \gamma^2 M^2$$

- Суммируем по всем $k$ и усредняем:

$$\frac{2\gamma}{K}\sum_{k=0}^{K-1}(f(x^k) - f(x^*)) \leq \frac{\|x^0 - x^*\|_2^2}{K} + \gamma^2 M^2$$

$$-\frac{\|x^0 - x^*\|_2^2}{2\gamma^2 K} + \frac{M^2}{2} = 0$$

$$\gamma = \frac{\|x^0 - x^*\|}{M\sqrt{K}}$$

- Откуда

$$f\left(\frac{1}{K}\sum x^k\right) - f(x^*) \leq$$

$$\frac{1}{K}\sum_{k=0}^{K-1}(f(x^k) - f(x^*)) \leq \underbrace{\frac{\|x^0 - x^*\|_2^2}{2\gamma K}} + \underbrace{\frac{\gamma M^2}{2}}$$

# Доказательство сходимости

- С предыдущего слайда:

$$\frac{1}{K}\sum_{k=0}^{K-1}(f(x^k) - f(x^*)) \leq \frac{\|x^0 - x^*\|_2^2}{2\gamma K} + \frac{\gamma M^2}{2}$$

- Гладкости нет, поэтому не получится доказать, что $f(x^k) \leq f(x^{k-1})$. Поэтому просто неравенство Йенсена для выпуклой функции:

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f(x^*) \leq \frac{\|x^0 - x^*\|_2^2}{2\gamma K} + \frac{\gamma M^2}{2}$$

# Доказательство сходимости

- С предыдущего слайда:

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f(x^*) \leq \frac{\|x^0 - x^*\|_2^2}{2\gamma K} + \frac{\gamma M^2}{2}$$

- **Вопрос:** как подобрать шаг?

# Доказательство сходимости

- С предыдущего слайда:

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f(x^*) \le \frac{\|x^0 - x^*\|_2^2}{2\gamma K} + \frac{\gamma M^2}{2}$$

- **Вопрос:** как подобрать шаг? минимизировать правую часть по $\gamma$: $\gamma = \frac{\|x^0 - x^*\|_2}{M\sqrt{K}}$. Откуда

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f(x^*) \le \frac{M\|x^0 - x^*\|_2}{\sqrt{K}}$$

- Можно более практично: $\gamma_k \sim \frac{1}{\sqrt{k}}$.

# Сходимость

## Теорема сходимость субградиентного спуска для $M$-Липшицевых и выпуклых функций

Пусть задача безусловной оптимизации с $M$-Липшицевой, выпуклой целевой функцией $f$ решается с помощью субградиентного спуска. Тогда справедлива следующая оценка сходимости

$$\varepsilon \sim f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f(x^*) \leq \frac{M\|x^0 - x^*\|_2}{\sqrt{K}} \qquad \frac{1}{\sqrt{K}}$$

Более того, чтобы добиться точности $\varepsilon$ по функции, необходимо

$$K = O\left(\frac{M^2\|x^0 - x^*\|_2^2}{\varepsilon^2}\right) \text{ итераций.} \qquad \frac{1}{\varepsilon^2} \qquad \frac{1}{K}$$

для гладкого случая: сильно вып. сход. $(1-q)^k$

вып. $\frac{1}{K}$

# Субградиентный метод: итог

- Обобщение градиентного спуска на негладкие задачи.

# Субградиентный метод: итог

- Обобщение градиентного спуска на негладкие задачи.
- Оценки сходимости в выпуклом случае: $\sim \frac{1}{\sqrt{K}}$, в сильно выпуклом случае: $\sim \frac{1}{K}$. **Вопрос:** какие были у градиентного спуска в гладком случае?

# Субградиентный метод: итог

- Обобщение градиентного спуска на негладкие задачи.
- Оценки сходимости в выпуклом случае: $\sim \frac{1}{\sqrt{K}}$, в сильно выпуклом случае: $\sim \frac{1}{K}$. **Вопрос:** какие были у градиентного спуска в гладком случае? $\sim \frac{1}{K}$ и линейная соответственно. Сходимость медленнее.

# Субградиентный метод: итог

- Обобщение градиентного спуска на негладкие задачи.
- Оценки сходимости в выпуклом случае: $\sim \frac{1}{\sqrt{K}}$, в сильно выпуклом случае: $\sim \frac{1}{K}$. **Вопрос:** какие были у градиентного спуска в гладком случае? $\sim \frac{1}{K}$ и линейная соответственно. Сходимость медленнее.
- Может возможно улучшить результат? Например, улучшить анализ или использовать моментум.

# Субградиентный метод: итог

- Обобщение градиентного спуска на негладкие задачи.
- Оценки сходимости в выпуклом случае: $\sim \frac{1}{\sqrt{K}}$, в сильно выпуклом случае: $\sim \frac{1}{K}$. **Вопрос:** какие были у градиентного спуска в гладком случае? $\sim \frac{1}{K}$ и линейная соответственно. Сходимость медленнее.
- Может возможно улучшить результат? Например, улучшить анализ или использовать моментум. В общем случае результат для субградиентного метода является неулучшаемым для выпуклых и сильно-выпуклых задач, т.е. он оптимален.

# Субградиентный метод: итог

- Обобщение градиентного спуска на негладкие задачи.
- Оценки сходимости в выпуклом случае: $\sim \frac{1}{\sqrt{K}}$, в сильно выпуклом случае: $\sim \frac{1}{K}$. **Вопрос:** какие были у градиентного спуска в гладком случае? $\sim \frac{1}{K}$ и линейная соответственно. Сходимость медленнее.
- Может возможно улучшить результат? Например, улучшить анализ или использовать моментум. В общем случае результат для субградиентного метода является неулучшаемым для выпуклых и сильно-выпуклых задач, т.е. он оптимален. **Вопрос:** а что в невыпуклом случае?

# Субградиентный метод: итог

- Обобщение градиентного спуска на негладкие задачи.
- Оценки сходимости в выпуклом случае: $\sim \frac{1}{\sqrt{K}}$, в сильно выпуклом случае: $\sim \frac{1}{K}$. **Вопрос:** какие были у градиентного спуска в гладком случае? $\sim \frac{1}{K}$ и линейная соответственно. Сходимость медленнее.
- Может возможно улучшить результат? Например, улучшить анализ или использовать моментум. В общем случае результат для субградиентного метода является неулучшаемым для выпуклых и сильно-выпуклых задач, т.е. он оптимален. **Вопрос:** а что в невыпуклом случае? С этого мы начинали курс – лучше, чем полный перебор там ничего не придумать.

# Субградиентный метод: итог

- Обобщение градиентного спуска на негладкие задачи.
- Оценки сходимости в выпуклом случае: $\sim \frac{1}{\sqrt{K}}$, в сильно выпуклом случае: $\sim \frac{1}{K}$. **Вопрос:** какие были у градиентного спуска в гладком случае? $\sim \frac{1}{K}$ и линейная соответственно. Сходимость медленнее.
- Может возможно улучшить результат? Например, улучшить анализ или использовать моментум. В общем случае результат для субградиентного метода является неулучшаемым для выпуклых и сильно-выпуклых задач, т.е. он оптимален. **Вопрос:** а что в невыпуклом случае? С этого мы начинали курс – лучше, чем полный перебор там ничего не придумать.
- Можно обобщить на метод с проекцией, а также на произвольную Брэгмановскую постановку (зеркальный спуск).

# Проксимальный оператор

- Поняли, что негладкие задачи «более сложные» по сравнению с гладкими задачами.
- Может быть получится «спрятать под ковер» отсутствие гладкости.

# Проксимальный оператор

- Поняли, что негладкие задачи «более сложные» по сравнению с гладкими задачами.
- Может быть получится «спрятать под ковер» отсутствие гладкости.
- Такую возможность дает проксимальный оператор:

**Определение проксимального оператора**

Для функции $r: \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ проксимальный оператор определяется следующим образом:

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

# Свойства проксимального оператора

**Лемма (свойство проксимального оператора)**

Пусть $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ выпуклая функция, для которой определен $\text{prox}_r$. Если существует такая $\hat{x} \in \mathbb{R}^d$, что $r(x) < +\infty$. Тогда проксимальный оператор однозначно определен (т.е. всегда возвращает единственное уникальное значение).

# Свойства проксимального оператора

## Лемма (свойство проксимального оператора)

Пусть $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ выпуклая функция, для которой определен $\mathrm{prox}_r$. Если существует такая $\hat{x} \in \mathbb{R}^d$, что $r(x) < +\infty$. Тогда проксимальный оператор однозначно определен (т.е. всегда возвращает единственное уникальное значение).

Доказательство: Проксимальный оператор возвращает минимум некоторой задачи оптимизации. Вопрос: что можно сказать про эту задачу?

# Свойства проксимального оператора

## Лемма (свойство проксимального оператора)

Пусть $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ выпуклая функция, для которой определен $\text{prox}_r$. Если существует такая $\hat{x} \in \mathbb{R}^d$, что $r(x) < +\infty$. Тогда проксимальный оператор однозначно определен (т.е. всегда возвращает единственное уникальное значение).

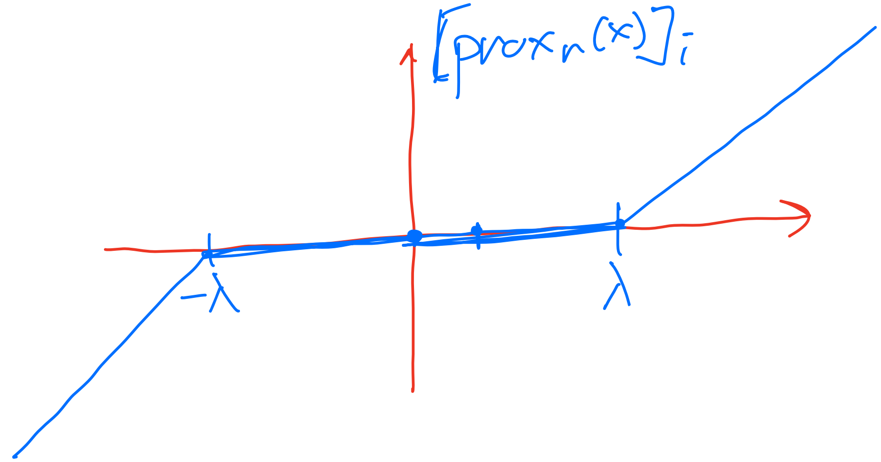
Доказательство: Проксимальный оператор возвращает минимум некоторой задачи оптимизации. Вопрос: что можно сказать про эту задачу? Она сильно выпуклая, а значит имеет строго один уникальный минимум (существование $\hat{x}$ необходимо для того, чтобы $r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|^2$ где-то принимала конечное значение).

# Примеры проксимального оператора

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

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

Такой проксимальный оператор еще называют трешхолдом.



# Примеры проксимального оператора

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

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

Такой проксимальный оператор еще называют трешхолдом.

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

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

# Примеры проксимального оператора

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

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

Такой проксимальный оператор еще называют трешхолдом.

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

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

- $r(x) = \mathbb{I}_{\mathcal{X}}(x)$, где $\mathcal{X}$ – выпуклое множество, и

$$\mathbb{I}_{\mathcal{X}} = \begin{cases} 0, & x \in \mathcal{X} \\ +\infty, & x \notin \mathcal{X} \end{cases}.$$

**Вопрос:** чему равен prox?

$$\text{prox}_r(x) = \operatorname*{argmin}_{\tilde{x} \in \mathbb{R}^d} \left( \mathbb{I}(x) + \frac{1}{2}\|x - \tilde{x}\|_2^2 \right)$$

$$= \operatorname*{argmin}_{\tilde{x} \in X} \left( \|x - \tilde{x}\|_2^2 \right)$$

# Примеры проксимального оператора

- $r(x) = \lambda\|x\|_1$, где $\lambda > 0$. Тогда
$$[\mathrm{prox}_r(x)]_i = [|x_i| - \lambda]_+ \cdot \mathrm{sign}(x_i)$$
Такой проксимальный оператор еще называют трешхолдом.
- $r(x) = \frac{\lambda}{2}\|x\|_2^2$, где $\lambda > 0$. Тогда
$$\mathrm{prox}_r(x) = \frac{x}{1+\lambda}.$$
- $r(x) = \mathbb{I}_{\mathcal{X}}$, где $\mathcal{X}$ – выпуклое множество, и
$$\mathbb{I}_{\mathcal{X}} = \begin{cases} 0, & x \in \mathcal{X} \\ +\infty, & x \notin \mathcal{X} \end{cases}.$$
**Вопрос:** чему равен prox?
$$\mathrm{prox}_r(x) = \mathrm{proj}_{\mathcal{X}}(x).$$

# Примеры проксимального оператора

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

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

  Такой проксимальный оператор еще называют трешхолдом.
- $r(x) = \frac{\lambda}{2}\|x\|_2^2$, где $\lambda > 0$. Тогда

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

- $r(x) = \mathbb{I}_{\mathcal{X}}$, где $\mathcal{X}$ – выпуклое множество, и

$$\mathbb{I}_{\mathcal{X}} = \begin{cases} 0, & x \in \mathcal{X} \\ +\infty, & x \notin \mathcal{X} \end{cases}.$$

  **Вопрос:** чему равен prox?

$$\text{prox}_r(x) = \text{proj}_{\mathcal{X}}(x).$$

- И еще множество других примеров и их комбинаций.

# Свойства проксимального оператора

$2 \Leftarrow\Rightarrow 3$ $\quad \Rightarrow \quad x-y \in \partial r(y)$

$\langle x-y; z-y \rangle \le r(z) - r(y) \quad \forall z$

$\Leftarrow \quad \forall z \quad \langle x-y; z-y \rangle \le r(z) - r(y)$

$\in \partial r(y)$

## Лемма (свойство проксимального оператора)

Пусть $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ выпуклая функция, для которой определен $\mathrm{prox}_r$. Тогда для любых $x, y \in \mathbb{R}^d$ следующие три условия являются эквивалентными:

- $\mathrm{prox}_r(x) = y$,
- $x - y \in \partial \cancel{f}\, r(y)$,
- $\langle x - y, z - y \rangle \le r(z) - r(y)$ для любого $z \in \mathbb{R}^d$.

$1 \Leftarrow\Rightarrow 2$

$\mathrm{prox}_r(x) = \arg\min \left( r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2 \right)$

$0 \in \partial \left( r(\tilde{x}) + \frac{1}{2}\|\tilde{x} - x\|_2^2 \right) \Big|_{\tilde{x} = y}$

$0 \in \partial r(y) + y - x$

$-(y - x) \in \partial r(y)$

# Доказательство

- Первое условие переписывается, как
$$y = \arg\min_{\tilde{x} \in \mathbb{R}^d} \left( r(\tilde{x}) + \frac{1}{2} \|x - \tilde{x}\|_2^2 \right).$$

# Доказательство

- Первое условие переписывается, как
$$y = \arg\min_{\tilde{x} \in \mathbb{R}^d} \left( r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2 \right).$$
- Из условия оптимальности для выпуклой функции $r$ это эквивалентно **вопрос:** чему?

# Доказательство

- Первое условие переписывается, как
$$y = \arg\min_{\tilde{x}\in\mathbb{R}^d} \left( r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2 \right).$$
- Из условия оптимальности для выпуклой функции $r$ это эквивалентно **вопрос:** чему?
$$0 \in \partial \left( r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2 \right)\Bigg|_{\tilde{x}=y} = \partial r(y) + y - x.$$

# Доказательство

- Первое условие переписывается, как
$$y = \arg\min_{\tilde{x} \in \mathbb{R}^d} \left( r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2 \right).$$
- Из условия оптимальности для выпуклой функции $r$ это эквивалентно **вопрос:** чему?
$$0 \in \partial \left( r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2 \right) \Bigg|_{\tilde{x} = y} = \partial r(y) + y - x.$$

Получили эквивалентность первого и второго условий.

# Доказательство

- Первое условие переписывается, как
$$y = \arg\min_{\tilde{x}\in\mathbb{R}^d} \left( r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2 \right).$$
- Из условия оптимальности для выпуклой функции $r$ это эквивалентно **вопрос:** чему?
$$0 \in \partial \left( r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2 \right) \Bigg|_{\tilde{x}=y} = \partial r(y) + y - x.$$
Получили эквивалентность первого и второго условий.
- Из определения субдифференциала, для любого субградиента $g \in \partial f(y)$ и для любого $z \in \mathbb{R}^d$:
$$\langle g, z - y \rangle \leq r(z) - r(y).$$

# Доказательство

- Первое условие переписывается, как
$$y = \arg\min_{\tilde{x}\in\mathbb{R}^d} \left( r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2 \right).$$
- Из условия оптимальности для выпуклой функции $r$ это эквивалентно **вопрос:** чему?
$$0 \in \partial \left( r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2 \right)\Bigg|_{\tilde{x}=y} = \partial r(y) + y - x.$$
Получили эквивалентность первого и второго условий.
- Из определения субдифференциала, для любого субградиента $g \in \partial f(y)$ и для любого $z \in \mathbb{R}^d$:
$$\langle g, z - y\rangle \le r(z) - r(y).$$
В частности справедливо и для $g = x - y$.

# Доказательство

- Первое условие переписывается, как
$$y = \arg\min_{\tilde{x}\in\mathbb{R}^d} \left( r(\tilde{x}) + \frac{1}{2}\|x-\tilde{x}\|_2^2 \right).$$
- Из условия оптимальности для выпуклой функции $r$ это эквивалентно **вопрос:** чему?
$$0 \in \partial \left( r(\tilde{x}) + \frac{1}{2}\|x-\tilde{x}\|_2^2 \right) \Bigg|_{\tilde{x}=y} = \partial r(y) + y - x.$$
Получили эквивалентность первого и второго условий.
- 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 справедливо и для $g = x - y$. В обратную сторону тоже очевидно: для $g = x - y$ выполнено соотношение выше, значит $g \in \partial r(y)$.

# Доказательство

- Первое условие переписывается, как
$$y = \arg\min_{\tilde{x}\in\mathbb{R}^d} \left( r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2 \right).$$
- Из условия оптимальности для выпуклой функции $r$ это эквивалентно **вопрос:** чему?
$$0 \in \partial \left( r(\tilde{x}) + \frac{1}{2}\|x - \tilde{x}\|_2^2 \right) \Bigg|_{\tilde{x}=y} = \partial r(y) + y - x.$$
Получили эквивалентность первого и второго условий.
- Из определения субдифференциала, для любого субградиента $g \in \partial f(y)$ и для любого $z \in \mathbb{R}^d$:
$$\langle g, z - y\rangle \le r(z) - r(y).$$
В частности справедливо и для $g = x - y$. В обратную сторону тоже очевидно: для $g = x - y$ выполнено соотношение выше, значит $g \in \partial r(y)$. Лемма доказана.

# Свойства проксимального оператора

$<x-y; z_1-y> \le r(z_1) - r(y)$

$< y-x; z_2-x> \le r(z_2)-r(x)$

## Лемма (свойство проксимального оператора)

Пусть $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ выпуклая функция, для которой определен $\text{prox}_r$. Тогда для любых $x, y \in \mathbb{R}^d$ выполнено следующее:

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

$u = \text{prox}_r(x)$ $v = \text{prox}_r(y)$

# Доказательство

- Пусть $u = \mathrm{prox}_r(x)$, $v = \mathrm{prox}_r(y)$.

# Доказательство

$y = \mathrm{prox}_r(x)$

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

$$
\begin{aligned}
\langle x - u, z_1 - u \rangle &\le r(z_1) - r(u), \qquad \forall z_1 \\
\langle y - v, z_2 - v \rangle &\le r(z_2) - r(v). \qquad \forall z_2
\end{aligned}
$$

$z_1 = v \qquad z_2 = u$

$$
+\begin{cases}
\langle x-u ; v-u \rangle \le r(v)-r(u) \\
\langle y-v ; u-v \rangle \le r(u)-r(v)
\end{cases}
$$

$\|\mathrm{prox} - \mathrm{prox}\| \le \|x-y\|$

$\langle x-u-y+v ; v-u \rangle \le 0$

$\langle x-y ; v-u \rangle + \|v-u\|_2^2 \le 0$

$\|v-u\|_2^2 \le \langle x-y ; u-v \rangle$

$\|\mathrm{prox}(x)-\mathrm{prox}(y)\|_2^2 \le \langle \mathrm{prox}(x)-\mathrm{prox}(y) ; x-y \rangle \le \|\mathrm{prox}-\mathrm{prox}\|_2 \cdot \|x-y\|_2$

# Доказательство

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

$$\langle x - u, z_1 - u \rangle \le r(z_1) - r(u),$$
$$\langle y - v, z_2 - v \rangle \le r(z_2) - r(v).$$

- Подставляем $z_1 = v$ и $z_2 = u$. Суммируем:

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

# Доказательство

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

$$\langle x - u, z_1 - u\rangle \le r(z_1) - r(u),$$
$$\langle y - v, z_2 - v\rangle \le r(z_2) - r(v).$$

- Подставляем $z_1 = v$ и $z_2 = u$. Суммируем:

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

Откуда

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

# Доказательство

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

$$\langle x - u, z_1 - u \rangle \le r(z_1) - r(u),$$
$$\langle y - v, z_2 - v \rangle \le r(z_2) - r(v).$$

- Подставляем $z_1 = v$ и $z_2 = u$. Суммируем:

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

Откуда

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

А это и требовалось доказать. **Вопрос:** как быстро доказать второе утверждение леммы?

# Доказательство

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

$$\langle x - u, z_1 - u\rangle \le r(z_1) - r(u),$$
$$\langle y - v, z_2 - v\rangle \le r(z_2) - r(v).$$

- Подставляем $z_1 = v$ и $z_2 = u$. Суммируем:

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

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

А это и требовалось доказать. **Вопрос:** как быстро доказать второе утверждение леммы? КБШ.

# Композитная задача

- Рассмотрим следующую задачу:

$$\min_{x \in \mathbb{R}^d} [f(x) + r(x)].$$

# Композитная задача

- Рассмотрим следующую задачу:

$$\min_{x \in \mathbb{R}^d} [f(x) + r(x)].$$

- Такая задача называется композитной.
- Предположим, что $f$ является $L$-гладкой выпуклой функцией, $r$ выпуклой (необязательно гладкой, но) проксимально дружественной функцией.

# Композитная задача

- Рассмотрим следующую задачу:

$$\min_{x \in \mathbb{R}^d} [f(x) + r(x)].$$

- Такая задача называется композитной.
- Предположим, что $f$ является $L$-гладкой выпуклой функцией, $r$ выпуклой (необязательно гладкой, но) проксимально дружественной функцией.
- Получается целевая функция состоит из гладкой и в общем случае негладкой части. Если $r \equiv 0$, то получаем гладкую задачу, которую умеем решать. Если $f \equiv 0$, то получаем негладкую задачу.

# Проксимальный градиентный метод

**Алгоритм 3** Проксимальный градиентный метод

**Вход:** размеры шагов $\{\gamma_{\cancel{k}}\}_{k=0} > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций $K$

1: **for** $k = 0, 1, \ldots, K-1$ **do**

2: Вычислить $\nabla f(x^k)$

3: $x^{k+1} = \mathrm{prox}_{\gamma r}(x^k - \gamma_{\cancel{k}} \nabla f(x^k))$

4: **end for**

**Выход:** $x^K$

$$\arg\min_{x} \left( \gamma r(x) + \frac{1}{2} \| x^k - \gamma \nabla f(x^k) - x \|_2^2 \right)$$

$$\nabla = 0$$

$$\gamma \nabla r(x^{k+1}) + x^{k+1} - x^k + \gamma \nabla f(x^k) = 0$$

$$x^{k+1} = x^k - \gamma \left( \nabla f(x^k) + \nabla r(x^{k+1}) \right)$$

# Проксимальный градиентный метод

**Алгоритм 4** Проксимальный градиентный метод

**Вход:** размеры шагов $\{\gamma_k\}_{k=0} > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций $K$

**for** $k = 0, 1, \ldots, K-1$ **do**
  Вычислить $\nabla f(x^k)$
  $x^{k+1} = \text{prox}_{\gamma r}(x^k - \gamma_k \nabla f(x^k))$
**end for**

**Выход:** $x^K$

- Если $r$ непрерывно дифференцируема, то условие оптимальности для подзадачи подсчета проксимального оператора записывается, как:

$$0 = \gamma \nabla r(x^{k+1}) + x^{k+1} - \gamma \nabla f(x^k).$$

# Проксимальный градиентный метод

**Алгоритм 5** Проксимальный градиентный метод

**Вход:** размеры шагов $\{\gamma_k\}_{k=0} > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций $K$

**for** $k = 0, 1, \ldots, K-1$ **do**
  Вычислить $\nabla f(x^k)$
  $x^{k+1} = \text{prox}_{\gamma r}(x^k - \gamma_k \nabla f(x^k))$
**end for**

**Выход:** $x^K$

- Если $r$ непрерывно дифференцируема, то условие оптимальности для подзадачи подсчета проксимального оператора записывается, как:

$$0 = \gamma \nabla r(x^{k+1}) + x^{k+1} - \gamma \nabla f(x^k).$$

- Откуда получаем так называемую неявную запись метода:

$$x^{k+1} = x^k - \gamma(\nabla f(x^k) + \nabla r({\color{blue}x^{k+1}}))$$

# Сходимость

**Лемма (свойство проксимального оператора)**

Пусть $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ выпуклые функции. Дополнительно предположим, что $f$ является непрерывно дифференцируемой и $L$-гладкой, а для $r$ определен $\mathrm{prox}_r$. Тогда $x^*$ – решение композитной задачи оптимизации тогда и только тогда, когда для любого $\gamma > 0$ выполнено:

$$x^* = \mathrm{prox}_{\gamma r}(x^* - \gamma \nabla f(x^*)).$$

$$\mathrm{argmin}\left(\gamma r(x) + \frac{1}{2}\|x - x^* + \gamma \nabla f(x^*)\|^2\right)$$

$$0 \in \partial\left(\gamma r(x) + \frac{1}{2}\|x - x^* + \gamma \nabla f(x^*)\|^2\right)$$

$$0 \in \gamma \partial r(x^*) + \underline{x^* - x^*} + \gamma \nabla f(x^*) \qquad x = x^*$$

# Доказательство

- Условие оптимальности:

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

# Доказательство

- Условие оптимальности:

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

- Откуда

$$x^* - \gamma \nabla f(x^*) - x^* \in \gamma \partial r(x^*).$$

# Доказательство

- Условие оптимальности:

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

- Откуда

$$x^* - \gamma \nabla f(x^*) - x^* \in \gamma \partial r(x^*).$$

$$x - y \in \gamma \partial r(x^*)$$

- Из свойств проксимального оператора

$$x^* = \mathrm{prox}_{\gamma r}(x^* - \gamma \nabla f(x^*)).$$

А это и требовалось.

# Сходимость

- В итоге имеем следующие свойства:

$$
\|\mathrm{prox}_r(x) - \mathrm{prox}_r(y)\|_2 \leq \|x - y\|_2
$$
$$
x^* = \mathrm{prox}_{\gamma r}(x^* - \gamma \nabla f(x^*)).
$$

**Вопрос:** в доказательстве какого метода уже нам нужны были такие свойства?

# Сходимость

- В итоге имеем следующие свойства:

$$\|\text{prox}_r(x) - \text{prox}_r(y)\|_2 \leq \|x - y\|_2$$
$$x^* = \text{prox}_{\gamma r}(x^* - \gamma \nabla f(x^*)).$$

**Вопрос:** в доказательстве какого метода уже нам нужны были такие свойства? Градиентный спуск с проекцией. Вспомним, что проксимальный оператор включает в себя и оператор проекции.

# Сходимость

- В итоге имеем следующие свойства:

$$\|\mathrm{prox}_r(x) - \mathrm{prox}_r(y)\|_2 \leq \|x - y\|_2$$
$$x^* = \mathrm{prox}_{\gamma r}(x^* - \gamma \nabla f(x^*)).$$

  **Вопрос:** в доказательстве какого метода уже нам нужны были такие свойства? Градиентный спуск с проекцией. Вспомним, что проксимальный оператор включает в себя и оператор проекции.
- Поэтому доказательство будет один в один.

# Доказательства сходимости

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

- Рассматриваем:

$$\|x^{k+1} - x^*\|_2^2 = \|\text{prox}_{\gamma r}(x^k - \gamma_k \nabla f(x^k)) - x^*\|_2^2$$

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

$$= \|\text{prox}_{\gamma r}(x^k - \gamma \nabla f(x^k)) - \text{prox}_{\gamma r}(x^* - \gamma \nabla f(x^*))\|_2^2$$

$$\leq \|x^k - \gamma \nabla f(x^k) - x^* + \gamma \nabla f(x^*)\|_2^2$$

# Доказательства сходимости

- Рассматриваем:

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

- Используем второй свойство с предыдущего слайда:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|\text{prox}_{\gamma f}(x^k - \gamma_k \nabla f(x^k)) - x^*\|_2^2 \\ &= \|\text{prox}_{\gamma f}(x^k - \gamma_k \nabla f(x^k)) - \text{prox}_{\gamma f}(x^* - \gamma_k \nabla f(x^*))\|_2^2\end{aligned}$$

# Доказательства сходимости

- Рассматриваем:

$$\|x^{k+1} - x^*\|_2^2 = \|\mathrm{prox}_{\gamma f}(x^k - \gamma_k \nabla f(x^k)) - x^*\|_2^2$$

- Используем второй свойство с предыдущего слайда:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|\mathrm{prox}_{\gamma f}(x^k - \gamma_k \nabla f(x^k)) - x^*\|_2^2 \\ &= \|\mathrm{prox}_{\gamma f}(x^k - \gamma_k \nabla f(x^k)) - \mathrm{prox}_{\gamma f}(x^* - \gamma_k \nabla f(x^*))\|_2^2\end{aligned}$$

- Теперь первое свойство с предыдущего слайда:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 \leq& \|x^k - \gamma_k \nabla f(x^k) - x^* + \gamma_k \nabla f(x^*)\|_2^2 \\ =& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^* \rangle \\ &+ \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2\end{aligned}$$

# Доказательства сходимости

- С предыдущего слайда:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\gamma_k\langle\nabla f(x^k) - \nabla f(x^*), x^k - x^*\rangle$$
$$+ \gamma_k^2\|\nabla f(x^k) - \nabla f(x^*)\|_2^2$$

# Доказательства сходимости

- С предыдущего слайда:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2$$

- Вспомним такой объект, как дивергенция Брэгмана, порожденную выпуклой функцией $f$:

$f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^* \rangle \geq 0$

$$V_f(x, y) = f(x) - f(y) - \langle \cancel{\nabla f(y), x -} y \rangle \geq 0.$$

$\nabla f(y) = 0$

$\|$

$(x^*)$

# Доказательства сходимости

- Воспользуемся сильной выпуклостью и гладкостью:

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 \\
& - 2\gamma_k \left( f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^* \rangle + \frac{\mu}{2}\|x^k - x^*\|_2^2 \right) \\
& + 2\gamma_k^2 L \left( f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^* \rangle \right) \\
= & (1 - \mu\gamma_k)\|x^k - x^*\|_2^2 + 2\gamma_k \underbrace{(\gamma_k L - 1)}_{\gamma \leq \frac{1}{L}} \underbrace{V_f(x^k, x^*)}_{\geq 0}
\end{aligned}
$$

- Дальше как раньше подбирает $\gamma_k$, пользуемся неотрицательности дивергенции Брэгмана.

# Проксимальный метод: итог

- Проксимальный градиентный спуск для композитной задачи с $L$-гладкой выпуклой функцией $f$ и выпуклой проксимально дружественной функцией $r$ имеет такую же сходимость, что и метод градиентного спуска для функции $f$. Свойства гладкости/негладкости $r$ при этом не влияют.
- Кажется, что положив $f \equiv 0$, с помощью такого метода можно решать любую негладкую задачу.

# Проксимальный метод: итог

- Проксимальный градиентный спуск для композитной задачи с $L$-гладкой выпуклой функцией $f$ и выпуклой проксимально дружественной функцией $r$ имеет такую же сходимость, что и метод градиентного спуска для функции $f$. Свойства гладкости/негладкости $r$ при этом не влияют.
- Кажется, что положив $f \equiv 0$, с помощью такого метода можно решать любую негладкую задачу. **Вопрос:** так ли это?

# Проксимальный метод: итог

- Проксимальный градиентный спуск для композитной задачи с $L$-гладкой выпуклой функцией $f$ и выпуклой проксимально дружественной функцией $r$ имеет такую же сходимость, что и метод градиентного спуска для функции $f$. Свойства гладкости/негладкости $r$ при этом не влияют.
- Кажется, что положив $f \equiv 0$, с помощью такого метода можно решать любую негладкую задачу. **Вопрос:** так ли это? если разрешить считать проксимальный оператор неточно (численно), то и правда можно решать любую задачу негладкой оптимизации.

# Проксимальный метод: итог

- Проксимальный градиентный спуск для композитной задачи с $L$-гладкой выпуклой функцией $f$ и выпуклой проксимально дружественной функцией $r$ имеет такую же сходимость, что и метод градиентного спуска для функции $f$. Свойства гладкости/негладкости $r$ при этом не влияют.
- Кажется, что положив $f \equiv 0$, с помощью такого метода можно решать любую негладкую задачу. **Вопрос:** так ли это? если разрешить считать проксимальный оператор неточно (численно), то и правда можно решать любую задачу негладкой оптимизации. НО это с точки зрения теории не лучше, чем решать задачу субградиентным спуском, потому что при решении подзадачи проксимального используется какой-то вспомогательный метод (например, тот же субградиентный спуск).