Non-smooth optimization. AdaGrad. RMSProp. ADAM. OASIS. Proximal method

Optimization in ML

Aleksandr Beznosikov

Skoltech

12 December 2023

Негладкие задачи

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

Негладкие задачи

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

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.

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

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

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

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

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 в точке, а значит градиента нет. Что может существовать вместо градиента?

Субградиент и субдифференциал

Если функция не дифференцируема в точке, а значит градиента нет. Что может существовать вместо градиента?

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:

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

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

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

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.

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

Доказательство

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

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

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:

Algorithm 2 Субградиентный метод

Input: 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**

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

Доказательство сходимости

- Ничего сверхъестественного:

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

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:

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

- Откуда

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

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:

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

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

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

- 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

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

Более того, чтобы добиться точности ε по функции, необходимо

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

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.
- Можно обобщить на метод с проекцией.

Подбор шага: AdaGradNorm

- Для субградиентного метода взяли шаг $\gamma = \frac{\|x^0 - x^*\|_2}{M\sqrt{k}}$. **Вопрос:** как заменить его более практично (убрать M, $\|x^0 - x^*\|_2$), но не потерять их физический смысл?

Подбор шага: AdaGradNorm

- Для субградиентного метода взяли шаг $\gamma = \frac{\|x^0 - x^*\|_2}{M\sqrt{k}}$. **Вопрос:** как заменить его более практично (убрать M, $\|x^0 - x^*\|_2$), но не потерять их физический смысл?

$$\gamma_k = \frac{\|x^0 - x^*\|_2}{\sqrt{kM^2}} = \frac{D}{\sqrt{\sum\limits_{t=0}^{k} \|g^t\|_2^2}}$$

Здесь мы дополнительно предполагаем, что $\|x^0 - x^*\|_2 \leq D$, такую D можно найти решая задачу оптимизации на ограниченном множестве (понадобится добавить проекцию) с евклидовым диаметром D.

Подбор шага: AdaGradNorm

- Для субградиентного метода взяли шаг $\gamma = \frac{\|x^0 - x^*\|_2}{M\sqrt{k}}$. **Вопрос:** как заменить его более практично (убрать M, $\|x^0 - x^*\|_2$), но не потерять их физический смысл?

$$\gamma_k = \frac{\|x^0 - x^*\|_2}{\sqrt{kM^2}} = \frac{D}{\sqrt{\sum\limits_{t=0}^{k} \|g^t\|_2^2}}$$

  Здесь мы дополнительно предполагаем, что $\|x^0 - x^*\|_2 \le D$, такую D можно найти решая задачу оптимизации на ограниченном множестве (понадобится добавить проекцию) с евклидовым диаметром D.
- Субградиентный метод с таким шагом называется AdaGradNorm. **Вопрос:** что нужно сделать, чтобы получился AdaGrad и мы приблизились к Adam?

Подбор шага: AdaGrad

- Делаем «индивидуальный» шаг для каждой координаты:

$$\gamma_{k,i} = \frac{D_i}{\sqrt{\sum_{t=0}^{k}(g_i^t)^2}}$$

Здесь введенно $|x_i - x_i^*| \leq D_i$.

Подбор шага: AdaGrad

- Делаем «индивидуальный» шаг для каждой координаты:

$$\gamma_{k,i} = \frac{D_i}{\sqrt{\sum_{t=0}^{k}(g_i^t)^2}}$$

Здесь введенно $|x_i - x_i^*| \leq D_i$.

Algorithm 4 AdaGrad

Input: 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: Для каждой координаты: $\gamma_{k,i} = \frac{D_i}{\sqrt{\sum_{t=0}^{k}(g_i^t)^2}}$

4: Для каждой координаты: $x_i^{k+1} = x_i^k - \gamma_{k,i} g_i^k$

5: **end for**

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

Доказательство сходимости

- Опять ничего сверхъестественного:

$$
\begin{aligned}
|x_i^{k+1} - x_i^*|^2 =& |x_i^k - \gamma_{k,i} g_i^k - x_i^*|^2 \\
=& |x_i^k - x_i^*|^2 - 2\gamma_{k,i} g_i^k (x_i^k - x_i^*) + \gamma_{k,i}^2 (g_i^k)^2
\end{aligned}
$$

Доказательство сходимости

- Опять ничего сверхъестественного:

$$
\begin{aligned}
|x_i^{k+1} - x_i^*|^2 =& |x_i^k - \gamma_{k,i} g_i^k - x_i^*|^2 \\
=& |x_i^k - x_i^*|^2 - 2\gamma_{k,i} g_i^k (x_i^k - x_i^*) + \gamma_{k,i}^2 (g_i^k)^2
\end{aligned}
$$

- Откуда:

$$
g_i^k (x_i^k - x_i^*) = \frac{1}{2\gamma_{k,i}} |x_i^k - x_i^*|^2 - \frac{1}{2\gamma_{k,i}} |x_i^{k+1} - x_i^*|^2 + \frac{\gamma_{k,i}}{2} (g_i^k)^2
$$

Доказательство сходимости

- Опять ничего сверхъестественного:

$$\begin{aligned} |x_i^{k+1} - x_i^*|^2 =& |x_i^k - \gamma_{k,i} g_i^k - x_i^*|^2 \\ =& |x_i^k - x_i^*|^2 - 2\gamma_{k,i} g_i^k (x_i^k - x_i^*) + \gamma_{k,i}^2 (g_i^k)^2 \end{aligned}$$

- Откуда:

$$g_i^k (x_i^k - x_i^*) = \frac{1}{2\gamma_{k,i}} |x_i^k - x_i^*|^2 - \frac{1}{2\gamma_{k,i}} |x_i^{k+1} - x_i^*|^2 + \frac{\gamma_{k,i}}{2} (g_i^k)^2$$

- Суммируем по всем координатам i от 1 до d:

$$\langle g^k, x^k - x^* \rangle = \sum_{i=1}^{d} \left[\frac{1}{2\gamma_{k,i}} |x_i^k - x_i^*|^2 - \frac{1}{2\gamma_{k,i}} |x_i^{k+1} - x_i^*|^2 + \frac{\gamma_{k,i}}{2} (g_i^k)^2 \right]$$

Доказательство сходимости

- Опять ничего сверхъестественного:

$$\begin{aligned}|x_i^{k+1} - x_i^*|^2 =& |x_i^k - \gamma_{k,i} g_i^k - x_i^*|^2 \\ =& |x_i^k - x_i^*|^2 - 2\gamma_{k,i} g_i^k (x_i^k - x_i^*) + \gamma_{k,i}^2 (g_i^k)^2\end{aligned}$$

- Откуда:

$$g_i^k(x_i^k - x_i^*) = \frac{1}{2\gamma_{k,i}}|x_i^k - x_i^*|^2 - \frac{1}{2\gamma_{k,i}}|x_i^{k+1} - x_i^*|^2 + \frac{\gamma_{k,i}}{2}(g_i^k)^2$$

- Суммируем по всем координатам i от 1 до d:

$$\langle g^k, x^k - x^* \rangle = \sum_{i=1}^{d} \left[\frac{1}{2\gamma_{k,i}}|x_i^k - x_i^*|^2 - \frac{1}{2\gamma_{k,i}}|x_i^{k+1} - x_i^*|^2 + \frac{\gamma_{k,i}}{2}(g_i^k)^2 \right]$$

- Из выпуклости и определения субградиента:

$$f(x^k) - f(x^*) \leq \sum_{i=1}^{d} \left[\frac{1}{2\gamma_{k,i}}|x_i^k - x_i^*|^2 - \frac{1}{2\gamma_{k,i}}|x_i^{k+1} - x_i^*|^2 + \frac{\gamma_{k,i}}{2}(g_i^k)^2 \right]$$

Доказательство сходимости

- С предыдущего слайда:

$$f(x^k) - f(x^*) \leq \sum_{i=1}^{d} \left[\frac{1}{2\gamma_{k,i}} |x_i^k - x_i^*|^2 - \frac{1}{2\gamma_{k,i}} |x_i^{k+1} - x_i^*|^2 + \frac{\gamma_{k,i}}{2} (g_i^k)^2 \right]$$

Доказательство сходимости

- С предыдущего слайда:

$$f(x^k) - f(x^*) \leq \sum_{i=1}^{d} \left[\frac{1}{2\gamma_{k,i}} |x_i^k - x_i^*|^2 - \frac{1}{2\gamma_{k,i}} |x_i^{k+1} - x_i^*|^2 + \frac{\gamma_{k,i}}{2} (g_i^k)^2 \right]$$

- Суммируем по всем k и усредняем:

$$\begin{aligned}
&\frac{1}{K} \sum_{k=0}^{K-1} \left(f(x^k) - f(x^*) \right) \\
&\quad \leq \frac{1}{2K} \sum_{k=0}^{K-1} \sum_{i=1}^{d} \left[\frac{1}{\gamma_{k,i}} |x_i^k - x_i^*|^2 - \frac{1}{\gamma_{k,i}} |x_i^{k+1} - x_i^*|^2 + \gamma_{k,i} (g_i^k)^2 \right] \\
&\quad = \frac{1}{2K} \sum_{i=1}^{d} \sum_{k=0}^{K-1} \left[\frac{1}{\gamma_{k,i}} |x_i^k - x_i^*|^2 - \frac{1}{\gamma_{k,i}} |x_i^{k+1} - x_i^*|^2 + \gamma_{k,i} (g_i^k)^2 \right]
\end{aligned}$$

Доказательство сходимости

- Продолжаем:

$$
\frac{1}{K}\sum_{k=0}^{K-1}\left(f(x^k) - f(x^*)\right)
\leq \frac{1}{2K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\left[\left(\frac{1}{\gamma_{k,i}} - \frac{1}{\gamma_{k-1,i}}\right)|x_i^k - x_i^*|^2 + \gamma_{k,i}(g_i^k)^2\right]
$$

Доказательство сходимости

- Продолжаем:

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\left(f(x^k) - f(x^*)\right) \\
&\qquad \leq \frac{1}{2K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\left[\left(\frac{1}{\gamma_{k,i}} - \frac{1}{\gamma_{k-1,i}}\right)|x_i^k - x_i^*|^2 + \gamma_{k,i}(g_i^k)^2\right]
\end{aligned}
$$

- Пользуемся ограниченностью $|x_i^k - x_i^*|^2 \leq D_i^2$:

$$
\frac{1}{K}\sum_{k=0}^{K-1}\left(f(x^k) - f(x^*)\right) \leq \frac{1}{2K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\left[\left(\frac{1}{\gamma_{k,i}} - \frac{1}{\gamma_{k-1,i}}\right)D_i^2 + \gamma_{k,i}(g_i^k)^2\right]
$$

Доказательство сходимости

- Подставляем $\gamma_{k,i}$:

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

$$\leq \frac{1}{2K}\sum_{i=1}^{d}\sum_{k=0}^{K-1}\left[\left(\sqrt{\sum_{t=0}^{k}(g_i^t)^2}-\sqrt{\sum_{t=0}^{k-1}(g_i^t)^2}\right)D_i+\frac{D_i(g_i^k)^2}{\sqrt{\sum_{t=0}^{k}(g_i^t)^2}}\right]$$

$$=\frac{1}{2K}\sum_{i=1}^{d}D_i\left[\sqrt{\sum_{t=0}^{K-1}(g_i^t)^2}+\sum_{k=0}^{K-1}\frac{(g_i^k)^2}{\sqrt{\sum_{t=0}^{k}(g_i^t)^2}}\right]$$

Доказательство сходимости

- С предыдущего слайда:

$$
\frac{1}{K}\sum_{k=0}^{K-1}\left(f(x^k) - f(x^*)\right) \leq \frac{1}{2K}\sum_{i=1}^{d} D_i \left[\sqrt{\sum_{t=0}^{K-1}(g_i^t)^2} + \sum_{k=0}^{K-1}\frac{(g_i^k)^2}{\sqrt{\sum_{t=0}^{k}(g_i^t)^2}}\right]
$$

Доказательство сходимости

- С предыдущего слайда:

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

$$\leq \frac{1}{2K}\sum_{i=1}^{d} D_i\left[\sqrt{\sum_{t=0}^{K-1}(g_i^t)^2}+\sum_{k=0}^{K-1}\frac{(g_i^k)^2}{\sqrt{\sum_{t=0}^{k}(g_i^t)^2}}\right]$$

- Далее нужен технический факт: для любых неотрицательных чисел $\{a_k\}$ выполнено, что

$$\sum_{k=0}^{K-1}\frac{(a_k)^2}{\sqrt{\sum_{t=0}^{k}(a_i^t)^2}}\leq 2\sqrt{\sum_{k=0}^{K-1}(a_i^k)^2}$$

Доказательство сходимости

- Итого:

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(f(x^k) - f(x^*)\right) \leq \frac{3}{2K}\sum_{i=1}^{d} D_i \sqrt{\sum_{t=0}^{K-1}(g_i^t)^2}$$

Доказательство сходимости

- Итого:

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(f(x^k) - f(x^*)\right) \leq \frac{3}{2K}\sum_{i=1}^{d} D_i \sqrt{\sum_{t=0}^{K-1}(g_i^t)^2}$$

- M-Липшицевость функции дает ограниченность компонент субградиента:

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(f(x^k) - f(x^*)\right) \leq \frac{3M}{2\sqrt{K}}\sum_{i=1}^{d} D_i \leq \frac{3M\tilde{D}}{2\sqrt{K}}.$$

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:

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

Свойства проксимального оператора

Лемма (свойство проксимального оператора)

Пусть $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ выпуклая функция, для которой определен 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 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 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

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

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

Вопрос: чему равен prox?

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
$$\mathbb{I}_{\mathcal{X}}(x) = \begin{cases} 0, & x \in \mathcal{X} \\ +\infty, & x \notin \mathcal{X} \end{cases}.$$
Вопрос: чему равен prox?
$$\text{prox}_r(x) = \text{proj}_{\mathcal{X}}(x).$$

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
$$\mathbb{I}_{\mathcal{X}}(x) = \begin{cases} 0, & x \in \mathcal{X} \\ +\infty, & x \notin \mathcal{X} \end{cases}.$$
Вопрос: чему равен prox?
$$\text{prox}_r(x) = \text{proj}_{\mathcal{X}}(x).$$
- И еще множество других примеров и их комбинаций.

Свойства проксимального оператора

Лемма (свойство проксимального оператора)

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

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

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 $z \in \mathbb{R}^d$:
$$\langle g, z - y \rangle \leq r(z) - r(y).$$
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 $z \in \mathbb{R}^d$:
$$\langle g, z - y\rangle \leq r(z) - r(y).$$
В частности справедливо и для $g = x - y$. В обратную сторону тоже очевидно: для $g = x - y$ выполнено соотношение выше, значит $g \in \partial r(y)$. Лемма доказана.

Свойства проксимального оператора

Лемма (свойство проксимального оператора)

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

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

Доказательство

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

Доказательство

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

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

Доказательство

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

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

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

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

Доказательство

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

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

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

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

Откуда

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

Доказательство

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

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

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

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

Откуда

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

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:

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

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

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

Откуда

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

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

Algorithm 5 Проксимальный градиентный метод

Input: размер 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 $\nabla f(x^k)$

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

4: **end for**

Output: x^K

Проксимальный градиентный метод

Algorithm 6 Проксимальный градиентный метод

Input: размер 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 $\nabla f(x^k)$

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

4: **end for**

Output: 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

Algorithm 7 Проксимальный градиентный метод

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

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

2: Вычислить $\nabla f(x^k)$

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

4: **end for**

Output: 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:

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

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 prox_r. Тогда x^* – решение композитной задачи оптимизации тогда и только тогда, когда для любого $\gamma > 0$ выполнено:

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

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

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

- 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:

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

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.

- Поэтому доказательство будет один в один.

Доказательства сходимости

- Рассматриваем:

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

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:

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

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:

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

Доказательства сходимости

- Воспользуемся сильной выпуклостью и гладкостью:

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

- Дальше как раньше подбирает γ_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).