Оптимизация на "простых" множествах. Метод проекции градиента. Метод условного градиента

Методы оптимизации

Александр Безносиков

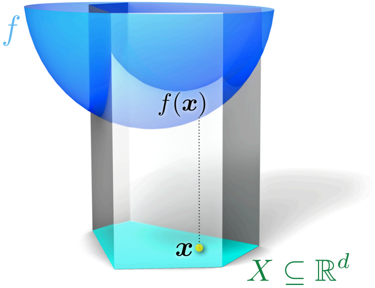
Московский физико-технический институт

17 октября 2024

ФПМИ

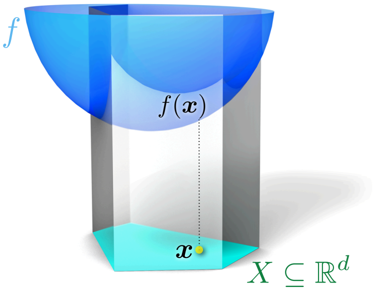
Задача оптимизации с ограничениями

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & x \in \mathcal{X} \end{aligned}$$



Задача оптимизации с ограничениями

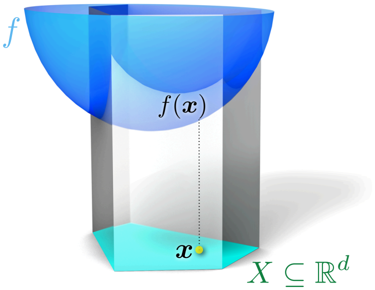
$$\begin{array}{rl} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{X} \end{array}$$



- До этого $\mathcal{X} = \mathbb{R}^d$ – безусловная оптимизация

Задача оптимизации с ограничениями

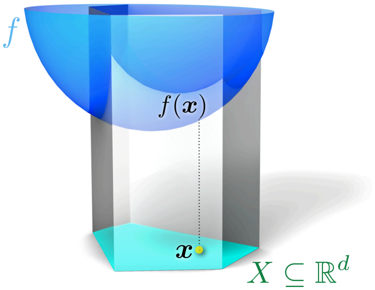
$$\begin{array}{rl} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{X} \end{array}$$



- До этого $\mathcal{X} = \mathbb{R}^d$ – безусловная оптимизация
- Теперь $\mathcal{X} \subsetneq \mathbb{R}^d$ (выпуклое множество)

Задача оптимизации с ограничениями

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & x \in \mathcal{X} \end{aligned}$$



- До этого $\mathcal{X} = \mathbb{R}^d$ – безусловная оптимизация
- Теперь $\mathcal{X} \subsetneq \mathbb{R}^d$ (выпуклое множество)

Вопрос: Зачем нам могут понадобится ограничения?

Условие оптимальности для задачи с ограничениями

Пусть дана выпуклая непрерывно дифференцируемая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$ и выпуклое множество $\mathcal{X}$. Тогда $x^* \in \mathcal{X}$ – глобальный минимум f на $\mathcal{X}$ тогда и только тогда, когда для всех $x \in \mathcal{X}$ выполнено

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

Доказательство

- Достаточное условие.

Доказательство

- Достаточное условие. Пусть $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ для $x \in \mathcal{X}$, тогда воспользуемся определением выпуклости:

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle \geq f(x^*).$$

Откуда следует, что x^* — глобальный минимум на $\mathcal{X}$.

Доказательство

- Достаточное условие. Пусть $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ для $x \in \mathcal{X}$, тогда воспользуемся определением выпуклости:

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle \geq f(x^*).$$

  Откуда следует, что x^* — глобальный минимум на $\mathcal{X}$.
- Необходимое условие.

Доказательство

- Достаточное условие. Пусть $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ для $x \in \mathcal{X}$, тогда воспользуемся определением выпуклости:

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle \geq f(x^*).$$

  Откуда следует, что x^* — глобальный минимум на $\mathcal{X}$.
- Необходимое условие. Пусть x^* — глобальный минимум на $\mathcal{X}$. Будем доказывать, что $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ для любого $x \in \mathcal{X}$. Пойдем от противного, т.е. предположим, что существует $x \in \mathcal{X}$ такой, что $\langle \nabla f(x^*), x - x^* \rangle < 0$.

Доказательство

- Достаточное условие. Пусть $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ для $x \in \mathcal{X}$, тогда воспользуемся определением выпуклости:

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle \geq f(x^*).$$

  Откуда следует, что x^* — глобальный минимум на $\mathcal{X}$.
- Необходимое условие. Пусть x^* — глобальный минимум на $\mathcal{X}$. Будем доказывать, что $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ для любого $x \in \mathcal{X}$. Пойдем от противного, т.е. предположим, что существует $x \in \mathcal{X}$ такой, что $\langle \nabla f(x^*), x - x^* \rangle < 0$. Рассмотрим точки вида

$$x_\lambda = \lambda x + (1 - \lambda) x^*, \text{ где } \lambda \in [0; 1].$$

  В силу выпуклости множества $\mathcal{X}$ точки $x_\lambda \in \mathcal{X}$.

Доказательство

Посмотрим, как ведет себя функция $\phi(\lambda) = f(x_\lambda) = f(\lambda x + (1-\lambda)x^*)$. В частности, заметим, что

$$\frac{d\phi}{d\lambda} = \frac{d}{d\lambda} f(\lambda x + (1-\lambda)x^*) = \langle \nabla f(x^* + \lambda(x - x^*)), x - x^* \rangle.$$

Доказательство

Посмотрим, как ведет себя функция $\phi(\lambda) = f(x_\lambda) = f(\lambda x + (1-\lambda)x^*)$. В частности, заметим, что

$$\frac{d\phi}{d\lambda} = \frac{d}{d\lambda} f(\lambda x + (1-\lambda)x^*) = \langle \nabla f(x^* + \lambda(x - x^*)), x - x^* \rangle.$$

Также заметим, что $\frac{d\phi}{d\lambda}\Big|_{\lambda=0} = \langle \nabla f(x^*), x - x^* \rangle < 0$.

Доказательство

Посмотрим, как ведет себя функция $\phi(\lambda) = f(x_\lambda) = f(\lambda x + (1-\lambda)x^*)$. В частности, заметим, что

$$\frac{d\phi}{d\lambda} = \frac{d}{d\lambda} f(\lambda x + (1-\lambda)x^*) = \langle \nabla f(x^* + \lambda(x - x^*)), x - x^* \rangle.$$

Также заметим, что $\frac{d\phi}{d\lambda}\Big|_{\lambda=0} = \langle \nabla f(x^*), x - x^* \rangle < 0$. **Вопрос:** что это значит?

Доказательство

Посмотрим, как ведет себя функция $\phi(\lambda) = f(x_\lambda) = f(\lambda x + (1-\lambda)x^*)$. В частности, заметим, что

$$\frac{d\phi}{d\lambda} = \frac{d}{d\lambda} f(\lambda x + (1-\lambda)x^*) = \langle \nabla f(x^* + \lambda(x - x^*)), x - x^* \rangle.$$

Также заметим, что $\frac{d\phi}{d\lambda}\big|_{\lambda=0} = \langle \nabla f(x^*), x - x^* \rangle < 0$. **Вопрос:** что это значит? Функция ϕ убывает в окрестности нуля. А значит для достаточно малых $\lambda > 0$ выполнено

$$f(x^* + \lambda(x - x^*)) = \phi(\lambda) < \phi(0) = f(x^*).$$

Доказательство

Посмотрим, как ведет себя функция $\phi(\lambda) = f(x_\lambda) = f(\lambda x + (1-\lambda)x^*)$. В частности, заметим, что

$$\frac{d\phi}{d\lambda} = \frac{d}{d\lambda} f(\lambda x + (1-\lambda)x^*) = \langle \nabla f(x^* + \lambda(x - x^*)), x - x^* \rangle.$$

Также заметим, что $\frac{d\phi}{d\lambda}\Big|_{\lambda=0} = \langle \nabla f(x^*), x - x^* \rangle < 0$. **Вопрос:** что это значит? Функция ϕ убывает в окрестности нуля. А значит для достаточно малых $\lambda > 0$ выполнено

$$f(x^* + \lambda(x - x^*)) = \phi(\lambda) < \phi(0) = f(x^*).$$

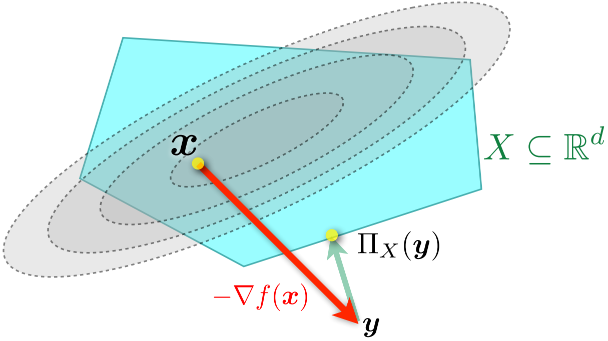
Пришли к противоречию, что x^* — глобальный минимум на $\mathcal{X}$.

Градиентный спуск с проекцией

Вопрос: Как «дешево» адаптироваться под задачи с ограничениями?

Градиентный спуск с проекцией

Вопрос: Как «дешево» адаптироваться под задачи с ограничениями?
Идея: взять градиентный спуск, но всегда держать текущую точку внутри множества $\mathcal{X}$, используя проекцию на $\mathcal{X}$:
$\Pi_{\mathcal{X}}(y) := \text{argmin}_{x \in \mathcal{X}} \|x - y\|_2^2$



Градиентный спуск с проекцией

Алгоритм 1 Градиентный спуск с проекцией

Вход: размеры шагов $\{\gamma_k\}_{k=0} > 0$, 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: $x^{k+1} = \Pi_{\mathcal{X}}\left[x^k - \gamma_k \nabla f(x^k)\right]$

4: **end for**

Выход: x^K

Свойства оператора проекции

Свойство оператора проекции

Для выпуклого замкнутого множества $\mathcal{X}$ и любой точки оператор проекции существует и принимает единственно значение.

Свойства оператора проекции

Свойство оператора проекции

Для выпуклого замкнутого множества $\mathcal{X}$ и любой точки оператор проекции существует и принимает единственно значение.

Доказательство:

Свойства оператора проекции

Свойство оператора проекции

Для выпуклого замкнутого множества $\mathcal{X}$ и любой точки оператор проекции существует и принимает единственно значение.

Доказательство: Следует из того, что задача проекции – это минимизация сильно выпуклой функции на выпуклом множестве. Решение такой задачи существует и единственно.

Свойства оператора проекции

Свойство оператора проекции

Пусть $\mathcal{X} \subseteq \mathbb{R}^d$ выпуклое замкнутое множество, $x \in \mathcal{X}, y \in \mathbb{R}^d$. Тогда

$$\langle x - \Pi_{\mathcal{X}}(y), y - \Pi_{\mathcal{X}}(y) \rangle \leq 0.$$

Свойства оператора проекции

Свойство оператора проекции

Пусть $\mathcal{X} \subseteq \mathbb{R}^d$ выпуклое замкнутое множество, $x \in \mathcal{X}, y \in \mathbb{R}^d$. Тогда

$$\langle x - \Pi_{\mathcal{X}}(y), y - \Pi_{\mathcal{X}}(y) \rangle \leq 0.$$

Доказательство:

Свойства оператора проекции

Свойство оператора проекции

Пусть $\mathcal{X} \subseteq \mathbb{R}^d$ выпуклое замкнутое множество, $x \in \mathcal{X}, y \in \mathbb{R}^d$. Тогда

$$\langle x - \Pi_{\mathcal{X}}(y), y - \Pi_{\mathcal{X}}(y) \rangle \leq 0.$$

Доказательство: Заметим, что $\Pi_{\mathcal{X}}(y)$ минимизирует дифференцируемую выпуклую функцию $d(z) = \|z - y\|_2^2$ на выпуклом множестве $\mathcal{X}$. **Вопрос:** что тогда нам даст условие оптимальности?

Свойства оператора проекции

Свойство оператора проекции

Пусть $\mathcal{X} \subseteq \mathbb{R}^d$ выпуклое замкнутое множество, $x \in \mathcal{X}, y \in \mathbb{R}^d$. Тогда

$$\langle x - \Pi_{\mathcal{X}}(y), y - \Pi_{\mathcal{X}}(y) \rangle \leq 0.$$

Доказательство: Заметим, что $\Pi_{\mathcal{X}}(y)$ минимизирует дифференцируемую выпуклую функцию $d(z) = \|z - y\|_2^2$ на выпуклом множестве $\mathcal{X}$. **Вопрос:** что тогда нам даст условие оптимальности?

$$\langle \nabla d(\Pi_{\mathcal{X}}(y)), x - \Pi_{\mathcal{X}}(y) \rangle \geq 0 \text{ для любой } x \in \mathcal{X}$$

Свойства оператора проекции

Свойство оператора проекции

Пусть $\mathcal{X} \subseteq \mathbb{R}^d$ выпуклое замкнутое множество, $x \in \mathcal{X}, y \in \mathbb{R}^d$. Тогда

$$\langle x - \Pi_{\mathcal{X}}(y), y - \Pi_{\mathcal{X}}(y) \rangle \leq 0.$$

Доказательство: Заметим, что $\Pi_{\mathcal{X}}(y)$ минимизирует дифференцируемую выпуклую функцию $d(z) = \|z - y\|_2^2$ на выпуклом множестве $\mathcal{X}$. **Вопрос:** что тогда нам даст условие оптимальности?

$$\langle \nabla d(\Pi_{\mathcal{X}}(y)), x - \Pi_{\mathcal{X}}(y) \rangle \geq 0 \ \text{ для любой } x \in \mathcal{X}$$

Вопрос: чему равен $\nabla d(z)$?

Свойства оператора проекции

Свойство оператора проекции

Пусть $\mathcal{X} \subseteq \mathbb{R}^d$ выпуклое замкнутое множество, $x \in \mathcal{X}, y \in \mathbb{R}^d$. Тогда

$$\langle x - \Pi_{\mathcal{X}}(y), y - \Pi_{\mathcal{X}}(y) \rangle \leq 0.$$

Доказательство: Заметим, что $\Pi_{\mathcal{X}}(y)$ минимизирует дифференцируемую выпуклую функцию $d(z) = \|z - y\|_2^2$ на выпуклом множестве $\mathcal{X}$. **Вопрос:** что тогда нам даст условие оптимальности?

$$\langle \nabla d(\Pi_{\mathcal{X}}(y)), x - \Pi_{\mathcal{X}}(y) \rangle \geq 0 \text{ для любой } x \in \mathcal{X}$$

Вопрос: чему равен $\nabla d(z)$? $2(z - y)$. Тогда

$$2\langle \Pi_{\mathcal{X}}(y) - y, x - \Pi_{\mathcal{X}}(y) \rangle \geq 0 \text{ для любой } x \in \mathcal{X}.$$

Что и требовалось доказать.

Свойства оператора проекции

Свойство нерасширяемости оператора проекции

Пусть $\mathcal{X} \subseteq \mathbb{R}^d$ выпуклое замкнутое множество, $x_1, x_2 \in \mathbb{R}^d$. Тогда

$$\|\Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2)\|_2 \leq \|x_1 - x_2\|_2.$$

Свойства оператора проекции

Свойство нерасширяемости оператора проекции

Пусть $\mathcal{X} \subseteq \mathbb{R}^d$ выпуклое замкнутое множество, $x_1, x_2 \in \mathbb{R}^d$. Тогда

$$\|\Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2)\|_2 \leq \|x_1 - x_2\|_2.$$

Доказательство:

Свойства оператора проекции

Свойство нерасширяемости оператора проекции

Пусть $\mathcal{X} \subseteq \mathbb{R}^d$ выпуклое замкнутое множество, $x_1, x_2 \in \mathbb{R}^d$. Тогда

$$\|\Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2)\|_2 \leq \|x_1 - x_2\|_2.$$

Доказательство: Из предыдущего свойства для $x \in \mathcal{X}$:

$$\langle x_1 - \Pi_{\mathcal{X}}(x_1), x - \Pi_{\mathcal{X}}(x_1) \rangle \leq 0$$

Свойства оператора проекции

Свойство нерасширяемости оператора проекции

Пусть $\mathcal{X} \subseteq \mathbb{R}^d$ выпуклое замкнутое множество, $x_1, x_2 \in \mathbb{R}^d$. Тогда

$$\|\Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2)\|_2 \leq \|x_1 - x_2\|_2.$$

Доказательство: Из предыдущего свойства для $x \in \mathcal{X}$:

$$\langle x_1 - \Pi_{\mathcal{X}}(x_1), x - \Pi_{\mathcal{X}}(x_1) \rangle \leq 0$$

Подставим $x = \Pi_{\mathcal{X}}(x_2)$:

$$\langle x_1 - \Pi_{\mathcal{X}}(x_1), \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1) \rangle \leq 0$$

Свойства оператора проекции

Свойство нерасширяемости оператора проекции

Пусть $\mathcal{X} \subseteq \mathbb{R}^d$ выпуклое замкнутое множество, $x_1, x_2 \in \mathbb{R}^d$. Тогда

$$\|\Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2)\|_2 \leq \|x_1 - x_2\|_2.$$

Доказательство: Из предыдущего свойства для $x \in \mathcal{X}$:

$$\langle x_1 - \Pi_{\mathcal{X}}(x_1), x - \Pi_{\mathcal{X}}(x_1) \rangle \leq 0$$

Подставим $x = \Pi_{\mathcal{X}}(x_2)$:

$$\langle x_1 - \Pi_{\mathcal{X}}(x_1), \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1) \rangle \leq 0$$

Аналогично:

$$\langle x_2 - \Pi_{\mathcal{X}}(x_2), \Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2) \rangle \leq 0$$

Свойства оператора проекции

Сложим

$$\langle x_1 - \Pi_{\mathcal{X}}(x_1), \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1) \rangle \leq 0$$

и

$$\langle x_2 - \Pi_{\mathcal{X}}(x_2), \Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2) \rangle \leq 0$$

Свойства оператора проекции

Сложим

$$\langle x_1 - \Pi_{\mathcal{X}}(x_1), \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1)\rangle \leq 0$$

и

$$\langle x_2 - \Pi_{\mathcal{X}}(x_2), \Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2)\rangle \leq 0$$

Получим

$$\langle \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1), \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1)\rangle \leq \langle \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1), x_2 - x_1\rangle$$

Свойства оператора проекции

Сложим

$$\langle x_1 - \Pi_{\mathcal{X}}(x_1), \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1) \rangle \leq 0$$

и

$$\langle x_2 - \Pi_{\mathcal{X}}(x_2), \Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2) \rangle \leq 0$$

Получим

$$\langle \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1), \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1) \rangle \leq \langle \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1), x_2 - x_1 \rangle$$

КБШ дает нужный результат

$$\|\Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1)\|_2^2 \leq \|\Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1)\| \cdot \|x_2 - x_1\|$$

Свойства оператора проекции

Свойство оператора проекции

Для x^* – решения условной задачи минимизации выпуклой непрерывно дифференцируемой функции f на выпуклом замкнутом множестве $\mathcal{X}$ справедливо

$$x^* = \Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*))$$

Свойства оператора проекции

Свойство оператора проекции

Для x^* – решения условной задачи минимизации выпуклой непрерывно дифференцируемой функции f на выпуклом замкнутом множестве $\mathcal{X}$ справедливо

$$x^* = \Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*))$$

Доказательство:

Свойства оператора проекции

Свойство оператора проекции

Для x^* – решения условной задачи минимизации выпуклой непрерывно дифференцируемой функции f на выпуклом замкнутом множестве $\mathcal{X}$ справедливо

$$x^* = \Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*))$$

Доказательство: Распишем:

$$\begin{aligned}\Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*)) &= \arg\min_{x \in \mathcal{X}} \|x^* - \gamma \nabla f(x^*) - x\|_2^2 \\ &= \arg\min_{x \in \mathcal{X}} \big[\|x^* - x\|_2^2 + 2\gamma\langle \nabla f(x^*), x - x^*\rangle \\ &\qquad + \gamma^2 \|\nabla f(x^*)\|_2^2\big]\end{aligned}$$

Свойства оператора проекции

Свойство оператора проекции

Для x^* – решения условной задачи минимизации выпуклой непрерывно дифференцируемой функции f на выпуклом замкнутом множестве $\mathcal{X}$ справедливо

$$x^* = \Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*))$$

Доказательство: Распишем:

$$\begin{aligned}\Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*)) &= \arg\min_{x \in \mathcal{X}} \|x^* - \gamma \nabla f(x^*) - x\|_2^2 \\ &= \arg\min_{x \in \mathcal{X}} \left[\|x^* - x\|_2^2 + 2\gamma \langle \nabla f(x^*), x - x^* \rangle \right.\\ &\qquad \left. + \gamma^2 \|\nabla f(x^*)\|_2^2\right]\end{aligned}$$

$$\Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*)) = \arg\min_{x \in \mathcal{X}} \left[\|x^* - x\|_2^2 + 2\gamma \langle \nabla f(x^*), x - x^* \rangle\right]$$

Свойства оператора проекции

С предыдущего слайда:

$$\Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*)) = \arg\min_{x \in \mathcal{X}} \left[\|x^* - x\|_2^2 + 2\gamma \langle \nabla f(x^*), x - x^* \rangle \right]$$

Свойства оператора проекции

С предыдущего слайда:

$$\Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*)) = \arg\min_{x \in \mathcal{X}} \left[\|x^* - x\|_2^2 + 2\gamma \langle \nabla f(x^*), x - x^* \rangle \right]$$

Вопрос: что можем сказать про выражение под $\arg\min$?

Свойства оператора проекции

С предыдущего слайда:

$$\Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*)) = \arg\min_{x \in \mathcal{X}} \left[\|x^* - x\|_2^2 + 2\gamma \langle \nabla f(x^*), x - x^* \rangle \right]$$

Вопрос: что можем сказать про выражение под $\arg\min$? Оба слагаемых неотрицательны.

Свойства оператора проекции

С предыдущего слайда:

$$\Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*)) = \arg\min_{x \in \mathcal{X}} \left[\|x^* - x\|_2^2 + 2\gamma \langle \nabla f(x^*), x - x^* \rangle \right]$$

Вопрос: что можем сказать про выражение под $\arg\min$? Оба слагаемых неотрицательны. **Вопрос:** чему тогда равен $\min_{x \in \mathcal{X}} \left[\|x^* - x\|_2^2 + 2\gamma \langle \nabla f(x^*), x - x^* \rangle \right]$?

Свойства оператора проекции

С предыдущего слайда:

$$\Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*)) = \arg\min_{x \in \mathcal{X}} \left[\|x^* - x\|_2^2 + 2\gamma \langle \nabla f(x^*), x - x^* \rangle \right]$$

Вопрос: что можем сказать про выражение под $\arg\min$? Оба слагаемых неотрицательны. **Вопрос:** чему тогда равен $\min_{x \in \mathcal{X}} \left[\|x^* - x\|_2^2 + 2\gamma \langle \nabla f(x^*), x - x^* \rangle \right]$? 0 и достигается он на x^*.

Доказательства сходимости

- Как и раньше рассматриваем:

$$\|x^{k+1} - x^*\|_2^2 = \|\Pi_{\mathcal{X}}\left[x^k - \gamma_k \nabla f(x^k)\right] - x^*\|_2^2$$

Доказательства сходимости

- Как и раньше рассматриваем:

$$\|x^{k+1} - x^*\|_2^2 = \|\Pi_{\mathcal{X}}\left[x^k - \gamma_k \nabla f(x^k)\right] - x^*\|_2^2$$

- Используем последнее доказанное свойство о стационарной точке:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|\Pi_{\mathcal{X}}\left[x^k - \gamma_k \nabla f(x^k)\right] - x^*\|_2^2 \\ &= \|\Pi_{\mathcal{X}}\left[x^k - \gamma_k \nabla f(x^k)\right] - \Pi_{\mathcal{X}}\left[x^* - \gamma_k \nabla f(x^*)\right]\|_2^2\end{aligned}$$

Доказательства сходимости

- Как и раньше рассматриваем:

$$\|x^{k+1} - x^*\|_2^2 = \|\Pi_{\mathcal{X}}\left[x^k - \gamma_k \nabla f(x^k)\right] - x^*\|_2^2$$

- Используем последнее доказанное свойство о стационарной точке:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|\Pi_{\mathcal{X}}\left[x^k - \gamma_k \nabla f(x^k)\right] - x^*\|_2^2 \\ &= \|\Pi_{\mathcal{X}}\left[x^k - \gamma_k \nabla f(x^k)\right] - \Pi_{\mathcal{X}}\left[x^* - \gamma_k \nabla f(x^*)\right]\|_2^2\end{aligned}$$

- Теперь третье свойство

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

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:

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

- Введем следующий объект дивергенцию Брэгмана, порожденную выпуклой функцией f:

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

Вопрос: почему дивергенция всегда положительна?

Доказательства сходимости

- С предыдущего слайда:

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

- Введем следующий объект дивергенцию Брэгмана, порожденную выпуклой функцией f:

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

Вопрос: почему дивергенция всегда положительна? В силу выпуклости f.

Доказательства сходимости

- Воспользуемся, как раньше 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, в силу неотрицательности дивергенции Брэгмана.

Сходимость

- Метод градиентного спуска с проекцией для L-гладкой и (μ-сильно) выпуклой целевой функции имеет такую же сходимость, что и метод градиентного спуска для аналогичной безусловной задачи оптимизации.
- Получается, что итерационные и оракульные сложности этих методов совпадают.

Сходимость

- Метод градиентного спуска с проекцией для L-гладкой и (μ-сильно) выпуклой целевой функции имеет такую же сходимость, что и метод градиентного спуска для аналогичной безусловной задачи оптимизации.
- Получается, что итерационные и оракульные сложности этих методов совпадают.
- Остается одна проблема — проекция является дополнительной задачей оптимизации, которую нужно отрешивать.

Проекция как отдельная задача

Аналитические решения:

- ℓ_2-шар радиуса 1 с центром в 0:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid \|x\|_2^2 = \sum_{i=1}^{d} x_i^2 \leq 1 \right\} \quad \Pi_{\mathcal{X}}(x) = \min\left\{1, \frac{1}{\|x\|_2}\right\} x$$

Проекция как отдельная задача

Аналитические решения:

- ℓ_2-шар радиуса 1 с центром в 0:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid \|x\|_2^2 = \sum_{i=1}^d x_i^2 \leq 1 \right\} \quad \Pi_{\mathcal{X}}(x) = \min\left\{1, \frac{1}{\|x\|_2}\right\} x$$

- Куб

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid a_i \leq x_i \leq b_i \right\} \quad [\Pi_{\mathcal{X}}(x)]_i = \begin{cases} a_i, & \text{если } x_i \leq a_i \\ x_i, & \text{если } a_i < x_i \leq b_i \\ b_i, & \text{иначе} \end{cases}$$

Проекция как отдельная задача

Аналитические решения:

- ℓ_2-шар радиуса 1 с центром в 0:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid \|x\|_2^2 = \sum_{i=1}^d x_i^2 \leq 1 \right\} \quad \Pi_{\mathcal{X}}(x) = \min\left\{1, \frac{1}{\|x\|_2}\right\} x$$

- Куб

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid a_i \leq x_i \leq b_i \right\} \quad [\Pi_{\mathcal{X}}(x)]_i = \begin{cases} a_i, & \text{если } x_i \leq a_i \\ x_i, & \text{если } a_i < x_i \leq b_i \\ b_i, & \text{иначе} \end{cases}$$

- Линейные ограничения типа равенств

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid Ax = b \right\} \quad \Pi_{\mathcal{X}}(x) = x - A^T(AA^T)^{-1}(Ax - b)$$

Проекция как отдельная задача

Аналитические решения:

- ℓ_2-шар радиуса 1 с центром в 0:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid \|x\|_2^2 = \sum_{i=1}^{d} x_i^2 \leq 1 \right\} \quad \Pi_{\mathcal{X}}(x) = \min\left\{1, \frac{1}{\|x\|_2}\right\} x$$

- Куб

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid a_i \leq x_i \leq b_i \right\} \quad [\Pi_{\mathcal{X}}(x)]_i = \begin{cases} a_i, & \text{если } x_i \leq a_i \\ x_i, & \text{если } a_i < x_i \leq b_i \\ b_i, & \text{иначе} \end{cases}$$

- Линейные ограничения типа равенств

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid Ax = b \right\} \quad \Pi_{\mathcal{X}}(x) = x - A^T(AA^T)^{-1}(Ax - b)$$

- Для некоторых множеств существуют эффективные алгоритмы проекций.

Линейная минимизация

Даже аналитическое решение для проекции может стоить вычислительно дорого. Получается, что квадратичная задача (которой является проекция) дорогая... **Вопрос:** тогда что взамен?

Линейная минимизация

Даже аналитическое решение для проекции может стоить вычислительно дорого. Получается, что квадратичная задача (которой является проекция) дорогая... **Вопрос:** тогда что взамен? Линейная задача:

$$\min_{s \in \mathcal{X}} \langle s, g \rangle$$

Линейная минимизация

Вопрос: Легко ли решить задачу линейной минимизации на выпуклом множестве?

- ℓ_1-шар с радиусом 1 и центром в 0:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid \|x\|_1 = \sum_{i=1}^{d} |x_i| \leq 1 \right\} \; s^* = -\text{sign}(g_i)\text{e}_i, \; i = \underset{j}{\text{argmax}} \, |g_j|$$

Линейная минимизация

Вопрос: Легко ли решить задачу линейной минимизации на выпуклом множестве?

- ℓ_1-шар с радиусом 1 и центром в 0:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid \|x\|_1 = \sum_{i=1}^{d} |x_i| \leq 1 \right\} \; s^* = -\text{sign}(g_i)\text{e}_i, \; i = \underset{j}{\text{argmax}}\, |g_j|$$

- Вероятностный симплекс:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid x_i \geq 0, \; \sum_{i=1}^{d} x_i = 1 \right\} \quad s^* = \text{e}_i, \; \text{где } i = \underset{j}{\text{argmin}}\, g_j$$

Линейная минимизация

Вопрос: Легко ли решить задачу линейной минимизации на выпуклом множестве?

- ℓ_1-шар с радиусом 1 и центром в 0:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid \|x\|_1 = \sum_{i=1}^{d} |x_i| \leq 1 \right\} \quad s^* = -\text{sign}(g_i)\text{e}_i, \; i = \underset{j}{\text{argmax}}\,|g_j|$$

- Вероятностный симплекс:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid x_i \geq 0, \; \sum_{i=1}^{d} x_i = 1 \right\} \quad s^* = \text{e}_i, \; \text{где } i = \underset{j}{\text{argmin}}\, g_j$$

- ℓ_∞-шар радиуса 1 с центром в 0:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid \|x\|_\infty = \max_{i=1,\dots d} |x_i| \leq 1 \right\} \quad s^* = -\sum_{i=1}^{d} \text{sign}(g_i)\text{e}_i$$

Линейная минимизация

Вопрос: Легко ли решить задачу линейной минимизации на выпуклом множестве?

- ℓ_1-шар с радиусом 1 и центром в 0:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid \|x\|_1 = \sum_{i=1}^{d} |x_i| \leq 1 \right\} \quad s^* = -\mathrm{sign}(g_i)\mathrm{e}_i, \; i = \operatorname*{argmax}_j |g_j|$$

- Вероятностный симплекс:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid x_i \geq 0, \; \sum_{i=1}^{d} x_i = 1 \right\} \quad s^* = \mathrm{e}_i, \; \text{где } i = \operatorname*{argmin}_j g_j$$

- ℓ_∞-шар радиуса 1 с центром в 0:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid \|x\|_\infty = \max_{i=1,\ldots d} |x_i| \leq 1 \right\} \quad s^* = -\sum_{i=1}^{d} \mathrm{sign}(g_i)\mathrm{e}_i$$

- Линейная минимизация не панацея, но хорошая альтернатива проекции.

Метод Франк-Вульфа

Алгоритм 2 Метод Франк-Вульфа

Вход: 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: Найти $s^k = \underset{s \in \mathcal{X}}{\operatorname{argmin}} \langle s, \nabla f(x^k) \rangle$

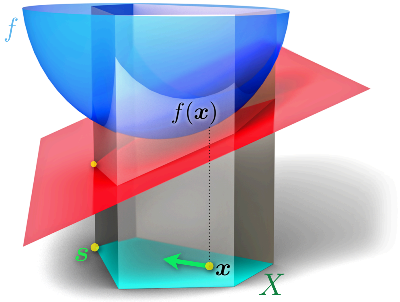
4: $\gamma_k = \frac{2}{k+2}$

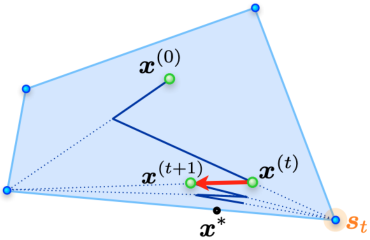
5: $x^{k+1} = (1 - \gamma_k) x^k + \gamma_k s^k$

6: **end for**

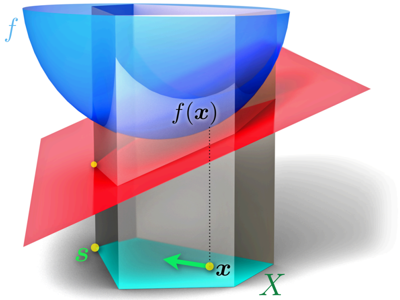
Выход: x^K

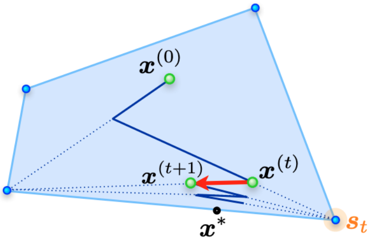
Физика метода Франк-Вульфа





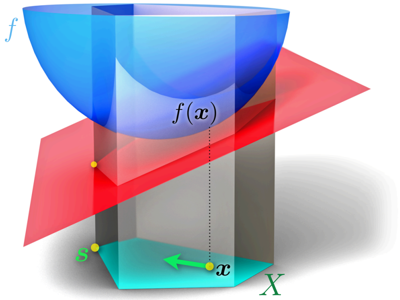
Физика метода Франк-Вульфа

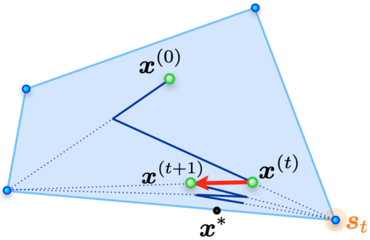




- Линейная минимизация указывает на "уголок".

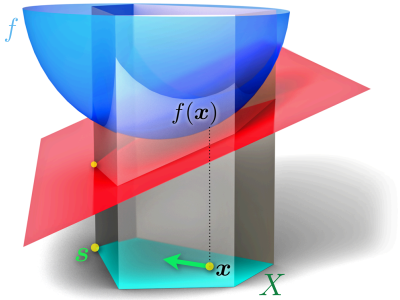
Физика метода Франк-Вульфа

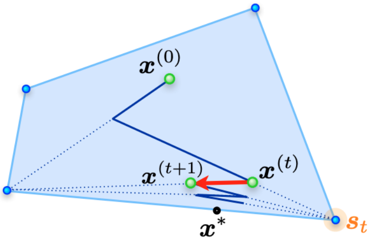




- Линейная минимизация указывает на "уголок".
- **Вопрос:** а что дает $(1 - \frac{2}{k+2})x^k + \frac{2}{k+2}s^k$?

Физика метода Франк-Вульфа





- Линейная минимизация указывает на "уголок".
- **Вопрос:** а что дает $(1 - \frac{2}{k+2})x^k + \frac{2}{k+2}s^k$? Вспомним, как усреднять в режиме онлайн — мы усредняем "уголки", те "уголки", которые ближе к решению, будут выбираться чаще.

Доказательство сходимости

- Сначала пользуемся L - гладкостью

$$\begin{aligned} f(x^{k+1}) =& f(x^k + \gamma_k(s^k - x^k)) \\ \leq& f(x^k) + \gamma_k \langle s^k - x^k, \nabla f(x^k) \rangle + \frac{\gamma_k^2}{2} L \|s^k - x^k\|_2^2 \\ \leq& f(x^k) + \gamma_k \langle s^k - x^k, \nabla f(x^k) \rangle + \frac{\gamma_k^2}{2} L \,\mathrm{diam}(\mathcal{X})^2. \end{aligned}$$

Доказательство сходимости

- Сначала пользуемся L - гладкостью

$$
\begin{aligned}
f(x^{k+1}) =& f(x^k + \gamma_k(s^k - x^k)) \\
\leq& f(x^k) + \gamma_k \langle s^k - x^k, \nabla f(x^k) \rangle + \frac{\gamma_k^2}{2} L \|s^k - x^k\|_2^2 \\
\leq& f(x^k) + \gamma_k \langle s^k - x^k, \nabla f(x^k) \rangle + \frac{\gamma_k^2}{2} L \operatorname{diam}(\mathcal{X})^2.
\end{aligned}
$$

- Принимаем во внимание, что s ищется как решение линейной минимизации:

$$
\begin{aligned}
f(x^{k+1}) - f(x^*) \leq& f(x^k) - f(x^*) + \gamma_k \langle x^* - x^k, \nabla f(x^k) \rangle + \frac{\gamma_k^2}{2} L \operatorname{diam}(\mathcal{X})^2 \\
\leq& f(x^k) - f(x^*) - \gamma_k (f(x^k) - f(x^*)) + \frac{\gamma_k^2}{2} L \operatorname{diam}(\mathcal{X})^2 \\
=& (1 - \gamma_k)(f(x^k) - f(x^*)) + \gamma_k^2 C.
\end{aligned}
$$

Здесь $C = \frac{L \operatorname{diam}(\mathcal{X})^2}{2}$.

Доказательство сходимости

Докажем по индукции следующее:

$$f(x^k) - f(x^*) \leq \frac{\max\{4C; f(x^0) - f(x^*)\}}{k+2} \quad k = 0, 1, \dots$$

Доказательство сходимости

Докажем по индукции следующее:

$$f(x^k) - f(x^*) \leq \frac{\max\{4C; f(x^0) - f(x^*)\}}{k+2} \quad k = 0, 1, \ldots$$

База индукции $k = 0$ следует автоматически. Рассмотрим $k \geq 1$,

$$\begin{aligned} f(x^{k+1}) - f(x^*) \leq & \quad (1-\gamma_k)(f(x^k) - f(x^*)) + {\gamma_k}^2 C \\ = & \quad \left(1 - \tfrac{2}{k+2}\right)(f(x^k) - f(x^*)) + \left(\tfrac{2}{k+2}\right)^2 C \\ \leq & \quad \left(1 - \tfrac{2}{k+2}\right)\tfrac{\max\{4C; f(x^0) - f(x^*)\}}{k+2} + \left(\tfrac{2}{k+2}\right)^2 C \ , \end{aligned}$$

где в последнем шаге мы использовали предположение индукции.

Доказательство сходимости

Осталось немного "поиграть" с выражением в правой части

$$
\begin{aligned}
f(x^{k+1}) - f(x^*) &\leq \frac{\max\{4C; f(x_0) - f(x^*)\}}{k+2}\left(1 - \frac{2}{k+2} + \frac{1}{k+2}\right) \\
&= \frac{\max\{4C; f(x^0) - f(x^*)\}}{k+2}\left(1 - \frac{1}{k+2}\right) \\
&= \frac{\max\{4C; f(x^0) - f(x^*)\}}{k+2} \frac{k+2-1}{k+2} \\
&\leq \frac{\max\{4C; f(x^0) - f(x^*)\}}{k+2} \frac{k+2}{k+3} \\
&= \frac{\max\{4C; f(x^0) - f(x^*)\}}{k+3},
\end{aligned}
$$

что и требовалось доказать.

Сходимость

Теорема о сходимости метода Франк-Вульфа

Пусть дана непрерывно дифференцируемая выпуклая L-гладкая функция $f : \mathbb{R}^d \to \mathbb{R}$, тогда для метода Франк-Вульфа справедлива следующая оценка сходимости

$$f(x^K) - f(x^*) \leq \frac{\max\{2L\,\mathrm{diam}(\mathcal{X})^2, f(x^0) - f(x^*)\}}{K+2}$$

где $\mathrm{diam}(\mathcal{X}) := \max_{x,y\in\mathcal{X}} \|x - y\|$ – диаметр множества $\mathcal{X}$.

Сходимость

- Сублинейная сходимость $1/K$, как и у градиентного спуска для выпуклой L-гладкой функции.

Сходимость

- Сублинейная сходимость $1/K$, как и у градиентного спуска для выпуклой L-гладкой функции.
- Проблема, что для сильно выпуклой целевой функции, линейная сходимость не появится в общем случае. Связано с линейной минимизацией.