Выпуклость и гладкость. Градиентный спуск

Методы оптимизации

Александр Безносиков

Московский физико-технический институт

12 сентября 2024

Задача оптимизации

- **Вопрос:** какие приложения задач оптимизации знаете/уже встречали?

Задача оптимизации

- **Вопрос:** какие приложения задач оптимизации знаете/уже встречали? Приложений у оптимизации масса: машинное обучение, анализ данных, статистика, финансы, логистика, планирование, управление и многие другие.

Задача оптимизации

- **Вопрос:** какие приложения задач оптимизации знаете/уже встречали? Приложений у оптимизации масса: машинное обучение, анализ данных, статистика, финансы, логистика, планирование, управление и многие другие.
- Оптимизация часто выступает инструментом во многих прикладных задачах, люди пользуются готовыми решениями/пакетами/"черными ящиками", который способны решать задачи оптимизации. Цель курса – как познакомиться с такими "черными ящиками", так и заглянуть внутрь и понять, что лежит внутри (в том числе с токи зрения теории).

Немного истории

- 1847: Коши и градиентный спуск для линейных систем
- 1950ые: линейное программирование (быстро перешло в нелинейное программирование), появление стохастических методов
- 1980ые: появление теории для общих задач.
- 2010ые: задачи оптимизации большого размера, теория стохастических методов

Задача оптимизации

$$\min_{\substack{g_i(x)\&0,\\ i=1,\ldots,m,\\ x\in Q}} f(x) \tag{1}$$

- $Q \subseteq \mathbb{R}^d$ — подмножество d-мерного пространства
- $f : Q \to \mathbb{R}$ — некоторая функция, заданная на множестве Q
- В качестве & берётся $\leq$ либо $=$
- $g_i(x) : Q \to \mathbb{R}$, $i = 1, \ldots, m$ — функции, задающие ограничения

Задача оптимизации

$$\min_{\substack{g_i(x)\&0,\\ i=1,\ldots,m,\\ x\in Q}} f(x) \qquad (1)$$

- $Q \subseteq \mathbb{R}^d$ — подмножество d-мерного пространства
- $f : Q \to \mathbb{R}$ — некоторая функция, заданная на множестве Q
- В качестве & берётся $\leq$ либо $=$
- $g_i(x) : Q \to \mathbb{R}$, $i = 1, \ldots, m$ — функции, задающие ограничения

Вопрос: что можно сказать про эту задач? сложная ли эта задача?

Задачи оптимизации. Первые наблюдения.

1. В общем случае задачи оптимизации могут не иметь решения. Например, задача $\min_{x \in \mathbb{R}} x$ не имеет решения.
2. Задачи оптимизации часто нельзя решить аналитически.
3. Их сложность зависит от вида целевой функции f, множества Q и может зависеть от размерности x.

Задачи оптимизации. Первые наблюдения.

1. В общем случае задачи оптимизации могут не иметь решения. Например, задача $\min_{x \in \mathbb{R}} x$ не имеет решения.
2. Задачи оптимизации часто нельзя решить аналитически.
3. Их сложность зависит от вида целевой функции f, множества Q и может зависеть от размерности x.

Если же задача оптимизации имеет решение, то на практике её обычно решают, вообще говоря, приближённо. Для этого применяются специальные алгоритмы, которые и называют методами оптимизации.

Методы оптимизации

- Нет смысла искать лучший метод для решения конкретной задачи. Например, лучший метод для решений задачи $\min_{x \in \mathbb{R}^d} \|x\|^2$ сходится за 1 итерацию: этот метод просто всегда выдаёт ответ $x^* = 0$. Очевидно, что для других задач такой метод не пригоден.
- Эффективность метода определяется для класса задач, т.к. обычно численные методы разрабатываются для *приближённого* решения множества однотипных задач.
- Метод разрабатывается для класса задач $\Longrightarrow$ метод не может иметь с самого начала полной информации о задаче. Вместо этого метод использует модель задачи, например, формулировку задачи, описание функциональных компонент, множества, на котором происходит оптимизация и т.д.

- Предполагается, что численный метод может накапливать специфическую информацию о задаче при помощи некоторого *оракула*. Под оракулом можно понимать некоторое устройство (программу, процедуру), которое отвечает на последовательные вопросы численного метода.

- Предполагается, что численный метод может накапливать специфическую информацию о задаче при помощи некоторого *оракула*. Под оракулом можно понимать некоторое устройство (программу, процедуру), которое отвечает на последовательные вопросы численного метода.

Вопрос: Какого рода вопросы хочется задавать оракулу?

- Предполагается, что численный метод может накапливать специфическую информацию о задаче при помощи некоторого *оракула*. Под оракулом можно понимать некоторое устройство (программу, процедуру), которое отвечает на последовательные вопросы численного метода.

Вопрос: Какого рода вопросы хочется задавать оракулу?

- Предполагается, что численный метод может накапливать специфическую информацию о задаче при помощи некоторого *оракула*. Под оракулом можно понимать некоторое устройство (программу, процедуру), которое отвечает на последовательные вопросы численного метода.

Вопрос: Какого рода вопросы хочется задавать оракулу?

Примеры оракулов

- **Оракул нулевого порядка** в запрашиваемой точке x возвращает значение целевой функции $f(x)$.
- **Оракул первого порядка** в запрашиваемой точке возвращает значение функции $f(x)$ и её градиент в данной точке $\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right)$.
- **Оракул второго порядка** в запрашиваемой точке возвращает значение и градиент функции $f(x), \nabla f(x)$, а также её гессиан в данной точке $\left(\nabla f^2(x)\right)_{ij} = \frac{\partial^2 f(x)}{\partial x_j \partial x_j}$.

Общая итеративная схема метода оптимизации $\mathcal{M}$

Входные данные: начальная точка x^0 (0 – верхний индекс), требуемая точность решения задачи $\varepsilon > 0$.

Общая итеративная схема метода оптимизации $\mathcal{M}$

Входные данные: начальная точка x^0 (0 – верхний индекс), требуемая точность решения задачи $\varepsilon > 0$.
Настройка. Задать $k = 0$ (счётчик итераций) и $I_{-1} = \varnothing$ (накапливаемая информационная модель решаемой задачи).

Общая итеративная схема метода оптимизации $\mathcal{M}$

Входные данные: начальная точка x^0 (0 – верхний индекс), требуемая точность решения задачи $\varepsilon > 0$.
Настройка. Задать $k = 0$ (счётчик итераций) и $I_{-1} = \varnothing$ (накапливаемая информационная модель решаемой задачи).
Основной цикл

1. Задать вопрос к оракулу $\mathcal{O}$ в точке x^k.

Общая итеративная схема метода оптимизации $\mathcal{M}$

Входные данные: начальная точка x^0 (0 – верхний индекс), требуемая точность решения задачи $\varepsilon > 0$.
Настройка. Задать $k = 0$ (счётчик итераций) и $I_{-1} = \varnothing$ (накапливаемая информационная модель решаемой задачи).
Основной цикл

1. Задать вопрос к оракулу $\mathcal{O}$ в точке x^k.
2. Пересчитать информационную модель: $I_k = I_{k-1} \cup (x^k, \mathcal{O}(x^k))$.

Общая итеративная схема метода оптимизации $\mathcal{M}$

Входные данные: начальная точка x^0 (0 – верхний индекс), требуемая точность решения задачи $\varepsilon > 0$.

Настройка. Задать $k = 0$ (счётчик итераций) и $I_{-1} = \varnothing$ (накапливаемая информационная модель решаемой задачи).

Основной цикл

1. Задать вопрос к оракулу $\mathcal{O}$ в точке x^k.
2. Пересчитать информационную модель: $I_k = I_{k-1} \cup (x^k, \mathcal{O}(x^k))$.
3. Применить правило метода $\mathcal{M}$ для получения новой точки x^{k+1} по модели I_k.

Общая итеративная схема метода оптимизации $\mathcal{M}$

Входные данные: начальная точка x^0 (0 – верхний индекс), требуемая точность решения задачи $\varepsilon > 0$.

Настройка. Задать $k = 0$ (счётчик итераций) и $I_{-1} = \varnothing$ (накапливаемая информационная модель решаемой задачи).

Основной цикл

1. Задать вопрос к оракулу $\mathcal{O}$ в точке x^k.
2. Пересчитать информационную модель: $I_k = I_{k-1} \cup (x^k, \mathcal{O}(x^k))$.
3. Применить правило метода $\mathcal{M}$ для получения новой точки x^{k+1} по модели I_k.
4. Проверить критерий остановки $\mathcal{T}_\varepsilon$. Если критерий выполнен, то выдать ответ $\bar{x}$, иначе положить $k := k + 1$ и вернуться на шаг 1.

Примеры итерационных методов. Градиентный спуск

Рассмотрим задачу оптимизации

$$\min_{x \in \mathbb{R}^d} f(x), \tag{2}$$

где функция $f(x)$ дифференцируема. Предположим, что в любой точке мы можем посчитать её градиент.

Примеры итерационных методов. Градиентный спуск

Рассмотрим задачу оптимизации

$$\min_{x \in \mathbb{R}^d} f(x), \tag{2}$$

где функция $f(x)$ дифференцируема. Предположим, что в любой точке мы можем посчитать её градиент.

Алгоритм 1 Градиентный спуск с постоянным размером шага

Вход: 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: $x^{k+1} = x^k - \gamma \nabla f(x^k)$
4: **end for**

Выход: x^K

Примеры итерационных методов. Градиентный спуск

Рассмотрим задачу оптимизации

$$\min_{x \in \mathbb{R}^d} f(x), \qquad (2)$$

где функция $f(x)$ дифференцируема. Предположим, что в любой точке мы можем посчитать её градиент.

Алгоритм 1 Градиентный спуск с постоянным размером шага

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

4: **end for**

Выход: x^K

Вопрос: в чем Алгоритм 1 отличается от определения общей итеративной схемы?

Примеры итерационных методов. Градиентный спуск

Рассмотрим задачу оптимизации

$$\min_{x \in \mathbb{R}^d} f(x), \qquad (2)$$

где функция $f(x)$ дифференцируема. Предположим, что в любой точке мы можем посчитать её градиент.

Алгоритм 1 Градиентный спуск с постоянным размером шага

Вход: 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:     $x^{k+1} = x^k - \gamma \nabla f(x^k)$
4: **end for**

Выход: x^K

Вопрос: в чем Алгоритм 1 отличается от определения общей итеративной схемы? В итеративной схеме использовался $\mathcal{T}_\varepsilon$.

Критерии останова

- По аргументу: $$\|x^k - x^*\| \leq \varepsilon.$$

  Вопрос: какие проблемы тут видим?

Критерии останова

- По аргументу: $$\|x^k - x^*\| \le \varepsilon.$$

  Вопрос: какие проблемы тут видим?
  - x^* – неизвестно, но можно так
  $$\|x^{k+1} - x^k\| \le \|x^{k+1} - x^*\| + \|x^k - x^*\| \le 2\varepsilon.$$
  Из $\|x^{k+1} - x^k\| \le \|x^k - x^*\| \le \varepsilon/2$, следует $\|x^{k+1} - x^k\| \le \varepsilon$ (в обратную сторону, очевидно, неверно). $\|x^{k+1} - x^k\| \le \varepsilon$ – это скорее практический вариант критерия, который работает, если есть понимание (интуиция), что $\|x^k - x^*\| \to 0$.

Критерии останова

- По аргументу: $$\|x^k - x^*\| \leq \varepsilon.$$

  Вопрос: какие проблемы тут видим?

  - x^* – неизвестно, но можно так
    $$\|x^{k+1} - x^k\| \leq \|x^{k+1} - x^*\| + \|x^k - x^*\| \leq 2\varepsilon.$$
    Из $\|x^{k+1} - x^k\| \leq \|x^k - x^*\| \leq \varepsilon/2$, следует $\|x^{k+1} - x^k\| \leq \varepsilon$ (в обратную сторону, очевидно, неверно). $\|x^{k+1} - x^k\| \leq \varepsilon$ – это скорее практический вариант критерия, который работает, если есть понимание (интуиция), что $\|x^k - x^*\| \to 0$.
  - x^* – не уникально. Тогда можно поменять критерий

Критерии останова

- По аргументу: $$\|x^k - x^*\| \leq \varepsilon.$$

  Вопрос: какие проблемы тут видим?

  - x^* – неизвестно, но можно так
    $$\|x^{k+1} - x^k\| \leq \|x^{k+1} - x^*\| + \|x^k - x^*\| \leq 2\varepsilon.$$
    Из $\|x^{k+1} - x^k\| \leq \|x^k - x^*\| \leq \varepsilon/2$, следует $\|x^{k+1} - x^k\| \leq \varepsilon$ (в обратную сторону, очевидно, неверно). $\|x^{k+1} - x^k\| \leq \varepsilon$ – это скорее практический вариант критерия, который работает, если есть понимание (интуиция), что $\|x^k - x^*\| \to 0$.
  - x^* – не уникально. Тогда можно поменять критерий

- По функции: $$f(x^k) - f^* \leq \varepsilon.$$

  Часто f^* известно, например, для $f(x) = \|Ax - b\|^2$. На практике можно использовать $|f(x^k) - f(x^{k+1})|$.

Критерии останова

- По аргументу: $\|x^k - x^*\| \le \varepsilon.$

  Вопрос: какие проблемы тут видим?
  - x^* – неизвестно, но можно так

    $$\|x^{k+1} - x^k\| \le \|x^{k+1} - x^*\| + \|x^k - x^*\| \le 2\varepsilon.$$

    Из $\|x^{k+1} - x^k\| \le \|x^k - x^*\| \le \varepsilon/2$, следует $\|x^{k+1} - x^k\| \le \varepsilon$ (в обратную сторону, очевидно, неверно). $\|x^{k+1} - x^k\| \le \varepsilon$ – это скорее практический вариант критерия, который работает, если есть понимание (интуиция), что $\|x^k - x^*\| \to 0$.
  - x^* – не уникально. Тогда можно поменять критерий
- По функции: $f(x^k) - f^* \le \varepsilon.$

  Часто f^* известно, например, для $f(x) = \|Ax - b\|^2$. На практике можно использовать $|f(x^k) - f(x^{k+1})|$.
- По норме градиента: $\|\nabla f(x^k)\| \le \varepsilon.$

  Вопрос: когда такой критерий можно использовать?

Критерии останова

- По аргументу: $$\|x^k - x^*\| \leq \varepsilon.$$

  Вопрос: какие проблемы тут видим?
  - x^* – неизвестно, но можно так
  $$\|x^{k+1} - x^k\| \leq \|x^{k+1} - x^*\| + \|x^k - x^*\| \leq 2\varepsilon.$$
  Из $\|x^{k+1} - x^k\| \leq \|x^k - x^*\| \leq \varepsilon/2$, следует $\|x^{k+1} - x^k\| \leq \varepsilon$ (в обратную сторону, очевидно, неверно). $\|x^{k+1} - x^k\| \leq \varepsilon$ – это скорее практический вариант критерия, который работает, если есть понимание (интуиция), что $\|x^k - x^*\| \to 0$.
  - x^* – не уникально. Тогда можно поменять критерий
- По функции: $$f(x^k) - f^* \leq \varepsilon.$$

  Часто f^* известно, например, для $f(x) = \|Ax - b\|^2$. На практике можно использовать $|f(x^k) - f(x^{k+1})|$.
- По норме градиента: $$\|\nabla f(x^k)\| \leq \varepsilon.$$

  Вопрос: когда такой критерий можно использовать? В безусловной оптимизации

Сложность методов оптимизации

- **Аналитическая/Оракульная сложность** — число обращений к оракулу, необходимое для решения задачи с точностью ε.
- **Арифметическая/Временна'я сложность** — общее число вычислений (включая работу оракула), необходимых для решения задачи с точностью ε.

Глобальный и локальный минимумы

Рассмотрим безусловную задачу: $\min_{x \in \mathbb{R}^d} f(x)$.

Глобальный и локальный минимумы

Рассмотрим безусловную задачу: $\min_{x \in \mathbb{R}^d} f(x)$.

Локальный минимум

Точка x^* называется локальным минимумом функции f на $\mathbb{R}^d$ (локальным решением задачи минимизации f на $\mathbb{R}^d$), если существует $r > 0$ такое, что для любого $y \in B_2^d(r, x^*) = \{y \in \mathbb{R}^d \mid \|y - x^*\|_2 \leq r\}$ следует, что $f(x^*) \leq f(y)$.

Глобальный и локальный минимумы

Рассмотрим безусловную задачу: $\min_{x \in \mathbb{R}^d} f(x)$.

Локальный минимум

Точка x^* называется локальным минимумом функции f на $\mathbb{R}^d$ (локальным решением задачи минимизации f на $\mathbb{R}^d$), если существует $r > 0$ такое, что для любого $y \in B_2^d(r, x^*) = \{y \in \mathbb{R}^d \mid \|y - x^*\|_2 \leq r\}$ следует, что $f(x^*) \leq f(y)$.

Глобальный минимум

Точка x^* называется глобальным минимумом функции f на $\mathbb{R}^d$ (глобальным решением задачи минимизации f на $\mathbb{R}^d$), если для любого $y \in \mathbb{R}^d$ следует, что $f(x^*) \leq f(y)$.

Глобальный и локальный минимумы

Рассмотрим безусловную задачу: $\min_{x \in \mathbb{R}^d} f(x)$.

Локальный минимум

Точка x^* называется локальным минимумом функции f на $\mathbb{R}^d$ (локальным решением задачи минимизации f на $\mathbb{R}^d$), если существует $r > 0$ такое, что для любого $y \in B_2^d(r, x^*) = \{y \in \mathbb{R}^d \mid \|y - x^*\|_2 \le r\}$ следует, что $f(x^*) \le f(y)$.

Глобальный минимум

Точка x^* называется глобальным минимумом функции f на $\mathbb{R}^d$ (глобальным решением задачи минимизации f на $\mathbb{R}^d$), если для любого $y \in \mathbb{R}^d$ следует, что $f(x^*) \le f(y)$.

Определение можно обобщить и до локального/глобального минимума на множестве $\mathcal{X}$, т.е. для задачи вида $\min_{x \in \mathcal{X}} f(x)$. Для этого надо брать $y \in B_2^d(r, x^*) \cap \mathcal{X}$ и $y \in \mathcal{X}$ в соответствующих определениях.

Условие оптимальности: общий случай

Теорема об условии оптимальности локального минимума

Пусть x^* – локальный минимумом функции f на $\mathbb{R}^d$, тогда если f дифференцируема, то $\nabla f(x^*) = 0$.

Условие оптимальности: общий случай

Теорема об условии оптимальности локального минимума

Пусть x^* – локальный минимумом функции f на $\mathbb{R}^d$, тогда если f дифференцируема, то $\nabla f(x^*) = 0$.

Легко проверить, что обратное неверно.

Условие оптимальности: общий случай

Доказательство

Пойдем от противного и предположим $\nabla f(x^*) \neq 0$. Разложим в ряд в окрестности локального минимума:

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

где $\lim_{x \to x^*} \frac{o(\|x-x^*\|_2)}{\|x-x^*\|_2} = 0$.

Условие оптимальности: общий случай

Доказательство

Пойдем от противного и предположим $\nabla f(x^*) \neq 0$. Разложим в ряд в окрестности локального минимума:

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

где $\lim_{x \to x^*} \frac{o(\|x - x^*\|_2)}{\|x - x^*\|_2} = 0$.
Рассмотрим $\tilde{x} = x^* - \lambda \nabla f(x^*)$. Цель: выбрать λ, чтобы $\tilde{x}$ попал в нужную окрестность из определения локального минимума.

Условие оптимальности: общий случай

Доказательство

Пойдем от противного и предположим $\nabla f(x^*) \neq 0$. Разложим в ряд в окрестности локального минимума:

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

где $\lim_{x \to x^*} \frac{o(\|x - x^*\|_2)}{\|x - x^*\|_2} = 0$.
Рассмотрим $\tilde{x} = x^* - \lambda \nabla f(x^*)$. Цель: выбрать λ, чтобы $\tilde{x}$ попал в нужную окрестность из определения локального минимума. Понятно, что такое λ можно найти.

Условие оптимальности: общий случай

Доказательство

Пойдем от противного и предположим $\nabla f(x^*) \neq 0$. Разложим в ряд в окрестности локального минимума:

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

где $\lim_{x \to x^*} \frac{o(\|x-x^*\|_2)}{\|x-x^*\|_2} = 0$.
Рассмотрим $\tilde{x} = x^* - \lambda \nabla f(x^*)$. Цель: выбрать λ, чтобы $\tilde{x}$ попал в нужную окрестность из определения локального минимума. Понятно, что такое λ можно найти. Тогда с одной стороны:

$$f(\tilde{x}) \geq f(x^*), \quad \text{и}$$

Условие оптимальности: общий случай

Доказательство

Пойдем от противного и предположим $\nabla f(x^*) \neq 0$. Разложим в ряд в окрестности локального минимума:

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

где $\lim_{x \to x^*} \frac{o(\|x - x^*\|_2)}{\|x - x^*\|_2} = 0$.
Рассмотрим $\tilde{x} = x^* - \lambda \nabla f(x^*)$. Цель: выбрать λ, чтобы $\tilde{x}$ попал в нужную окрестность из определения локального минимума. Понятно, что такое λ можно найти. Тогда с одной стороны:

$$f(\tilde{x}) \geq f(x^*), \quad \text{и}$$

$$\begin{aligned} f(\tilde{x}) &= f(x^*) + \langle \nabla f(x^*), \tilde{x} - x^* \rangle + o(\|\tilde{x} - x^*\|_2) \\ &= f(x^*) - \lambda \|\nabla f(x^*)\|^2 + o(\lambda \|\nabla f(x^*)\|_2) \end{aligned}$$

Условие оптимальности: общий случай

Доказательство

Набросим еще одно ограничение на "малость" λ. Пусть теперь еще выполнено, что $|o(\lambda\|\nabla f(x^*)\|_2)| \leq \frac{\lambda}{2}\|\nabla f(x^*)\|_2^2$. Тогда для подобранного $\lambda > 0$

$$f(\tilde{x}) \leq f(x^*) - \frac{\lambda}{2}\|\nabla f(x^*)\|^2$$

Пришли к противоречию, что x^* – локальный минимум.

Локальный и глобальный минимум

- Наша цель – глобальный минимум (или точка близкая к нему в некотором смысле).
- Заветная мечта – придумать метод решающий все задачи оптимизации. Выглядит нереалистично, но чем черт не шутит.

Класс задач минимизации липшицевых функций

$$\min_{x \in B_d} f(x)$$

- $B_d = \{x \in \mathbb{R}^d \mid 0 \le x_i \le 1, \quad i = 1, \ldots, d\}$
- Функция $f(x)$ является M-липшицевой на B_d относительно ℓ_∞-нормы:

$$\forall x, y \ |f(x) - f(y)| \le M\|x - y\|_\infty = M \max_{i=1,\ldots,d} |x_i - y_i|.$$

Класс задач минимизации липшицевых функций

Наблюдение

Множество B_d является ограниченным и замкнутым, т.е. компактом, а из липшицевости функции f следует и её непрерывность, поэтому задача (17) имеет решение, ибо непрерывная на компакте функция достигает своих минимального и максимального значений. Пусть $f^* = \min_{x \in B_d} f(x)$.

- **Класс методов.** Для данной задачи рассмотрим методы нулевого порядка.
- **Цель:** найти $\bar{x} \in B_d$: $f(\bar{x}) - f^* \leq \varepsilon$.

Метод перебора

Рассмотрим один из самых простых способов решения этой задачи — метод равномерного перебора.

Алгоритм 2 Метод равномерного перебора

Вход: целочисленный параметр перебора $p \geq 1$

1: Сформировать $(p+1)^d$ точек вида $x_{(i_1,\ldots,i_d)} = \left(\frac{i_1}{p}, \frac{i_2}{p}, \ldots, \frac{i_d}{p}\right)^\top$, где $(i_1, \ldots, i_d) \in \{0, 1, \ldots, p\}^d$

2: Среди точек $x_{(i_1,\ldots,i_d)}$ найти точку $\bar{x}$ с наименьшим значением целевой функции f.

Выход: $\bar{x}, f(\bar{x})$

Гарантии

Теорема 1

Алгоритм 2 с параметром p возвращает такую точку $\bar{x}$, что

$$f(\bar{x}) - f^* \leq \frac{M}{2p}, \tag{3}$$

откуда следует, что методу равномерного перебора нужно в худшем случае

$$\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2\right)^d \tag{4}$$

обращений к оракулу, чтобы гарантировать $f(\bar{x}) - f^* \leq \varepsilon$.

Гарантии: доказательство

Доказательство Теоремы 1

Пусть x^* — решение задачи (точка минимума функции f). Тогда в построенной «сетке» из точек найдётся такая точка $x_{(i_1,\ldots,i_d)}$, что $x := x_{(i_1,\ldots,i_d)} \leq x^* \leq x_{(i_1+1,\ldots,i_d+1)} =: y$, где знак «$\leq$» применяется покомпонентно.

Гарантии: доказательство

Доказательство Теоремы 1

Пусть x^* — решение задачи (точка минимума функции f). Тогда в построенной «сетке» из точек найдётся такая точка $x_{(i_1,\ldots,i_d)}$, что $x := x_{(i_1,\ldots,i_d)} \leq x^* \leq x_{(i_1+1,\ldots,i_d+1)} =: y$, где знак «$\leq$» применяется покомпонентно. Во-первых, $y_i - x_i = \frac{1}{p}$ и $x_i^* \in [x_i, y_i]$ для всех $i = 1, \ldots, d$.

Гарантии: доказательство

Доказательство Теоремы 1

Пусть x^* — решение задачи (точка минимума функции f). Тогда в построенной «сетке» из точек найдётся такая точка $x_{(i_1,\ldots,i_d)}$, что $x := x_{(i_1,\ldots,i_d)} \leq x^* \leq x_{(i_1+1,\ldots,i_d+1)} =: y$, где знак «$\leq$» применяется покомпонентно. Во-первых, $y_i - x_i = \frac{1}{p}$ и $x_i^* \in [x_i, y_i]$ для всех $i = 1, \ldots, d$. Кроме того, рассмотрим точки $\hat{x}$ и $\tilde{x}$ такие, что $\hat{x} = \frac{x+y}{2}$ и

$$\tilde{x}_i = \begin{cases} y_i, & \text{если } x_i^* \geq \hat{x}_i, \\ x_i, & \text{иначе.} \end{cases}$$

Гарантии: доказательство

Доказательство Теоремы 1 (продолжение)

Заметим, что $\tilde{x}$ принадлежит «сетке» и $|\tilde{x}_i - x_i^*| \leq \frac{1}{2p}$, а значит, $\|\tilde{x} - x^*\|_\infty \leq \frac{1}{2p}$.

Гарантии: доказательство

Доказательство Теоремы 1 (продолжение)

Заметим, что $\tilde{x}$ принадлежит «сетке» и $|\tilde{x}_i - x_i^*| \leq \frac{1}{2p}$, а значит, $\|\tilde{x} - x^*\|_\infty \leq \frac{1}{2p}$. Поскольку $f(\bar{x}) \leq f(\tilde{x})$ (по определению), получаем

Гарантии: доказательство

Доказательство Теоремы 1 (продолжение)

Заметим, что $\tilde{x}$ принадлежит «сетке» и $|\tilde{x}_i - x_i^*| \leq \frac{1}{2p}$, а значит, $\|\tilde{x} - x^*\|_\infty \leq \frac{1}{2p}$. Поскольку $f(\bar{x}) \leq f(\tilde{x})$ (по определению), получаем

$$f(\bar{x}) - f^* \leq f(\tilde{x}) - f^* \leq M\|\tilde{x} - x^*\|_\infty \leq \frac{M}{2p}.$$

Гарантии: доказательство

Доказательство Теоремы 1 (продолжение)

Заметим, что $\tilde{x}$ принадлежит «сетке» и $|\tilde{x}_i - x_i^*| \leq \frac{1}{2p}$, а значит, $\|\tilde{x} - x^*\|_\infty \leq \frac{1}{2p}$. Поскольку $f(\bar{x}) \leq f(\tilde{x})$ (по определению), получаем

$$f(\bar{x}) - f^* \leq f(\tilde{x}) - f^* \leq M\|\tilde{x} - x^*\|_\infty \leq \frac{M}{2p}.$$

Выписанная выше оценка достигается методом равномерного перебора за $(p+1)^d$ обращений к оракулу. Следовательно, чтобы гарантировать $f(\bar{x}) - f^* \leq \varepsilon$, необходимо взять $p = \left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 1$, т.е. метод сделает $\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2\right)^d$ обращений к оракулу.

Метод перебора: анализ

Вопрос: хороший результат получили или нет?

Метод перебора: анализ

Вопрос: хороший результат получили или нет?

- Предположим $M = 2$, $d = 13$ и $\varepsilon = 0.01$, то есть размерность задачи сравнительно небольшая и точность решения задачи не слишком высокая.

Метод перебора: анализ

Вопрос: хороший результат получили или нет?

- Предположим $M = 2$, $d = 13$ И $\varepsilon = 0.01$, то есть размерность задачи сравнительно небольшая и точность решения задачи не слишком высокая.
- Необходимое число обращений к оракулу: $\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2\right)^d = 102^{13} > 10^{26}$.

Метод перебора: анализ

Вопрос: хороший результат получили или нет?

- Предположим $M = 2$, $d = 13$ И $\varepsilon = 0.01$, то есть размерность задачи сравнительно небольшая и точность решения задачи не слишком высокая.
- Необходимое число обращений к оракулу: $\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2\right)^d = 102^{13} > 10^{26}$.
- Сложность одного вызова оракула не менее 1, но если потребовать, чтобы он обязательно считал, переданную ему точки, то сложность не менее d операции.

Метод перебора: анализ

Вопрос: хороший результат получили или нет?

- Предположим $M = 2$, $d = 13$ И $\varepsilon = 0.01$, то есть размерность задачи сравнительно небольшая и точность решения задачи не слишком высокая.
- Необходимое число обращений к оракулу: $\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2\right)^d = 102^{13} > 10^{26}$.
- Сложность одного вызова оракула не менее 1, но если потребовать, чтобы он обязательно считал, переданную ему точки, то сложность не менее d операции.
- Производительность компьютера: 10^{11} арифметических операций в секунду.

Метод перебора: анализ

Вопрос: хороший результат получили или нет?

- Предположим $M = 2$, $d = 13$ И $\varepsilon = 0.01$, то есть размерность задачи сравнительно небольшая и точность решения задачи не слишком высокая.
- Необходимое число обращений к оракулу: $\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2\right)^d = 102^{13} > 10^{26}$.
- Сложность одного вызова оракула не менее 1, но если потребовать, чтобы он обязательно считал, переданную ему точки, то сложность не менее d операции.
- Производительность компьютера: 10^{11} арифметических операций в секунду.
- Общее время: хотя бы 10^{15} секунд, что больше 30 миллионов лет.

Верхние и нижние оценки

- **Вопрос:** что мы сейчас получили? верхнюю или нижнюю оценку? что такое верхняя оценка?

Верхние и нижние оценки

- **Вопрос:** что мы сейчас получили? верхнюю или нижнюю оценку? что такое верхняя оценка?
- **Верхняя оценка** – гарантии нахождения решения определенным методом из рассматриваемого класса методов (например, методы с оракулом нулевого порядка) для любой задачи из класса (Липшецева целевая функция на кубе).
- **Нижняя оценка** – гарантия, что для любого метода из класса существует «плохая» задача из класса такая, что метод будет сходиться не лучше, чем утверждает нижняя оценка.
- Возникает вопрос: может мы плохо вывели верхнюю оценку (неидеальный анализ), может ли предложить другой метод из рассматриваемого класса, который будет находить приближённое решение существенно быстрее? На этот вопрос и даст ответ нижняя оценка.

Нижняя оценка

Теорема 2

Пусть $\varepsilon < \frac{M}{2}$. Тогда аналитическая сложность описанного класса задач, т.е. аналитическая сложность метода на «худшей» для него задаче из данного класса, составляет по крайней мере

$$\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor \right)^d \quad \text{вызовов оракула.} \tag{5}$$

Нижняя оценка: доказательство

Схема доказательства Теоремы 2

Пусть $p = \left\lfloor \frac{M}{2\varepsilon} \right\rfloor$. Доказываем от противного: предположим, что существует такой метод, который решает задачу за $N < (p^d - 1)$ обращений к оракулу, чтобы решить задачу с точностью ε (по функции).

Нижняя оценка: доказательство

Схема доказательства Теоремы 2

Пусть $p = \left\lfloor \frac{M}{2\varepsilon} \right\rfloor$. Доказываем от противного: предположим, что существует такой метод, который решает задачу за $N < (p^d - 1)$ обращений к оракулу, чтобы решить задачу с точностью ε (по функции). Построим такую функцию, на которой метод не сможет найти ε-решение, при помощи сопротивляющегося оракула: пусть изначально наша целевая функция $f(x)$ всюду равна 0.

Нижняя оценка: доказательство

Схема доказательства Теоремы 2

Пусть $p = \left\lfloor \frac{M}{2\varepsilon} \right\rfloor$. Доказываем от противного: предположим, что существует такой метод, который решает задачу за $N < (p^d - 1)$ обращений к оракулу, чтобы решить задачу с точностью ε (по функции). Построим такую функцию, на которой метод не сможет найти ε-решение, при помощи сопротивляющегося оракула: пусть изначально наша целевая функция $f(x)$ всюду равна 0. Запустим метод, он запросит значение f в N точках, везде получит 0 и выдаст какую-то точку (возможно, отличную от всех предыдущих N, как ответ). В итоге мы в ходе работы алгоритма заглянули в $N + 1 < p^d$ точку.

Нижняя оценка: доказательство

Схема доказательства Теоремы 2

Пусть $p = \left\lfloor \frac{M}{2\varepsilon} \right\rfloor$. Доказываем от противного: предположим, что существует такой метод, который решает задачу за $N < (p^d - 1)$ обращений к оракулу, чтобы решить задачу с точностью ε (по функции). Построим такую функцию, на которой метод не сможет найти ε-решение, при помощи сопротивляющегося оракула: пусть изначально наша целевая функция $f(x)$ всюду равна 0. Запустим метод, он запросит значение f в N точках, везде получит 0 и выдаст какую-то точку (возможно, отличную от всех предыдущих N, как ответ). В итоге мы в ходе работы алгоритма заглянули в $N + 1 < p^d$ точку. Тогда по принципу Дирихле найдётся такой «кубик» $B = \{x \mid \hat{x} \preceq x \preceq \hat{x} + \frac{1}{p}e\}$ (где $\hat{x}$ и $\hat{x} + \frac{1}{p}e$ — точки из «сетки» с шагом p, e — вектор из единиц), который не содержит ни одной из $N + 1$ точки (в том числе и выхода метода).

Нижняя оценка: доказательство

Схема доказательства Теоремы 2 (продолжение)

Пусть x^* — это центр «кубика» B, т.е. $x^* = \hat{x} + \frac{1}{2p}e$.

Нижняя оценка: доказательство

Схема доказательства Теоремы 2 (продолжение)

Пусть x^* — это центр «кубика» B, т.е. $x^* = \hat{x} + \frac{1}{2p}e$. Немного модифицируем функцию $\bar{f}(x) = \min\{0, M\|x - x^*\|_\infty - \varepsilon\}$.

Нижняя оценка: доказательство

Схема доказательства Теоремы 2 (продолжение)

Пусть x^* — это центр «кубика» B, т.е. $x^* = \hat{x} + \frac{1}{2p}e$. Немного модифицируем функцию $\bar{f}(x) = \min\{0, M\|x - x^*\|_\infty - \varepsilon\}$. Функция $\bar{f}(x)$ липшицева с константой M относительно ℓ_∞-нормы и принимает своё минимальное значение $-\varepsilon$ в точке x^*.

Нижняя оценка: доказательство

Схема доказательства Теоремы 2 (продолжение)

Пусть x^* — это центр «кубика» B, т.е. $x^* = \hat{x} + \frac{1}{2p}e$. Немного модифицируем функцию $\bar{f}(x) = \min\{0, M\|x - x^*\|_\infty - \varepsilon\}$. Функция $\bar{f}(x)$ липшицева с константой M относительно ℓ_∞-нормы и принимает своё минимальное значение $-\varepsilon$ в точке x^*. Более того, функция $\bar{f}(x)$ отлична от нуля только внутри куба $B' = \{x \mid \|x - x^*\| \leq \frac{\varepsilon}{M}\}$, который лежит внутри куба B, т.к. $2p \leq \frac{M}{\varepsilon}$.

Нижняя оценка: доказательство

Схема доказательства Теоремы 2 (продолжение)

Пусть x^* — это центр «кубика» B, т.е. $x^* = \hat{x} + \frac{1}{2p}e$. Немного модифицируем функцию $\bar{f}(x) = \min\{0, M\|x - x^*\|_\infty - \varepsilon\}$. Функция $\bar{f}(x)$ липшицева с константой M относительно ℓ_∞-нормы и принимает своё минимальное значение $-\varepsilon$ в точке x^*. Более того, функция $\bar{f}(x)$ отлична от нуля только внутри куба $B' = \{x \mid \|x - x^*\| \le \frac{\varepsilon}{M}\}$, который лежит внутри куба B, т.к. $2p \le \frac{M}{\varepsilon}$. Следовательно, рассмотренный метод на данной функции не может найти ε-решение. Противоречие.

Нижняя оценка: доказательство

Схема доказательства Теоремы 2 (продолжение)

Пусть x^* — это центр «кубика» B, т.е. $x^* = \hat{x} + \frac{1}{2p}e$. Немного модифицируем функцию $\bar{f}(x) = \min\{0, M\|x - x^*\|_\infty - \varepsilon\}$. Функция $\bar{f}(x)$ липшицева с константой M относительно ℓ_∞-нормы и принимает своё минимальное значение $-\varepsilon$ в точке x^*. Более того, функция $\bar{f}(x)$ отлична от нуля только внутри куба $B' = \{x \mid \|x - x^*\| \leq \frac{\varepsilon}{M}\}$, который лежит внутри куба B, т.к. $2p \leq \frac{M}{\varepsilon}$. Следовательно, рассмотренный метод на данной функции не может найти ε-решение. Противоречие.

Итак, в указанном классе у любого метода оценки на скорость сходимости весьма пессимистичные. Возникает вопрос: какие свойства нужно потребовать от класса оптимизируемых функций, чтобы оценки стали более оптимистичными?

Выпуклость: определение

Определение выпуклой функции

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

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

Выпуклость: определение

Определение выпуклой функции

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

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

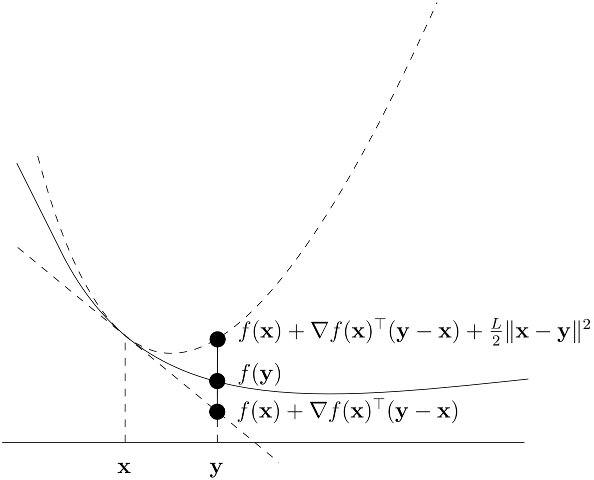
На 5 семинаре будет еще одно определение (эквивалентное в случае дифференцируемых функций).

Определение выпуклой функции

Будем говорить, что она является выпуклой, если для любых $x, y \in \mathbb{R}^d$ и для любого $\lambda \in [0; 1]$ выполнено

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$$

Выпуклость



Ограничение снизу на поведение.

Сильная выпуклость: определение

Определение μ-сильно выпуклой функции

Пусть дана непрерывно дифференцируемая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что она является μ-сильно выпуклой ($\mu > 0$), если для любых $x, y \in \mathbb{R}^d$ выполнено

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

Сильная выпуклость: определение

Определение μ-сильно выпуклой функции

Пусть дана непрерывно дифференцируемая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что она является μ-сильно выпуклой ($\mu > 0$), если для любых $x, y \in \mathbb{R}^d$ выполнено

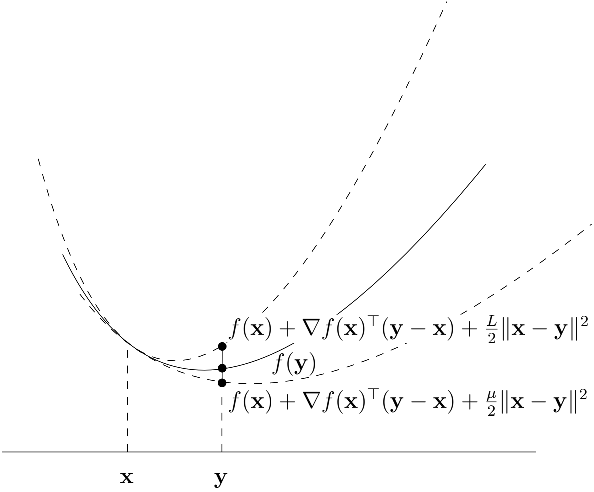
$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2.$$

Определение μ-сильно выпуклой функции

Будем говорить, что она является μ-сильно выпуклой, если для любых $x, y \in \mathbb{R}^d$ и для любого $\lambda \in [0; 1]$ выполнено

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \lambda(1-\lambda)\frac{\mu}{2}\|x - y\|_2^2$$

Сильная выпуклость



Более сильное ограничение снизу на поведение.

Условие оптимальности: выпуклый случай

Теорема об условии оптимальности безусловной выпуклой задачи

Пусть дана выпуклая непрерывно дифференцируемая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$. Если для некоторой точки $x^* \in \mathbb{R}^d$ верно, что $\nabla f(x^*) = 0$, то x^* – глобальный минимум f на всем $\mathbb{R}^d$.

Условие оптимальности: выпуклый случай

Теорема об условии оптимальности безусловной выпуклой задачи

Пусть дана выпуклая непрерывно дифференцируемая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$. Если для некоторой точки $x^* \in \mathbb{R}^d$ верно, что $\nabla f(x^*) = 0$, то x^* – глобальный минимум f на всем $\mathbb{R}^d$.

Доказательство

Запишем определение выпуклости:

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

Условие оптимальности: выпуклый случай

Теорема об условии оптимальности безусловной выпуклой задачи

Пусть дана выпуклая непрерывно дифференцируемая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$. Если для некоторой точки $x^* \in \mathbb{R}^d$ верно, что $\nabla f(x^*) = 0$, то x^* – глобальный минимум f на всем $\mathbb{R}^d$.

Доказательство

Запишем определение выпуклости:

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

В обратную сторону уже доказывали выше для произвольных функций.

Выпуклое множество: определение

Определение выпуклого множества

Множество $\mathcal{X}$ называется выпуклым, если для любых $x, y \in \mathcal{X}$ и для любого $\lambda \in [0; 1]$ следует, что

$$\lambda x + (1 - \lambda) y \in \mathcal{X}.$$

Выпуклое множество: определение

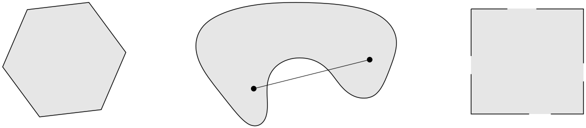
Определение выпуклого множества

Множество $\mathcal{X}$ называется выпуклым, если для любых $x, y \in \mathcal{X}$ и для любого $\lambda \in [0; 1]$ следует, что

$$\lambda x + (1 - \lambda) y \in \mathcal{X}.$$

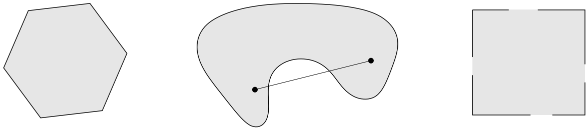
Смысл: вместе с любыми двумя точками множества в множество входит и отрезок с концами в этих точках.
Подробнее на 4 семинаре.

Выпуклое множество: пример



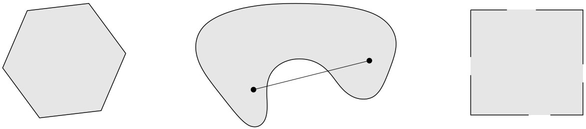
Вопрос: какие множества здесь выпуклые?

Выпуклое множество: пример



Вопрос: какие множества здесь выпуклые? 1 (смотрите на границы 3)

Выпуклое множество: пример



Вопрос: какие множества здесь выпуклые? 1 (смотрите на границы 3)
Вопрос: понятие выпуклости функции можно обобщить на множество $\mathcal{X}$ (необязательно $\mathbb{R}^d$), но важно, чтобы множество $\mathcal{X}$ было выпуклым. Зачем?

Минимумы выпуклых функций

Теорема о минимумах выпуклых функций

Рассмотрим задачу

$$\min_{x \in \mathcal{X}} f(x),$$

где f – выпуклая, $\mathcal{X}$ - выпуклое. Тогда всякий локальный минимум f на $\mathcal{X}$ является и глобальным.

Минимумы выпуклых функций

Доказательство

Пусть x^* – локальный минимум. Рассмотрим точку вида

$$x_\lambda = \lambda x + (1 - \lambda)x^*,$$

где x – произвольная точка из $\mathcal{X}$.

Минимумы выпуклых функций

Доказательство

Пусть x^* – локальный минимум. Рассмотрим точку вида

$$x_\lambda = \lambda x + (1 - \lambda)x^*,$$

где x – произвольная точка из $\mathcal{X}$. **Вопрос:** что можно сказать про x_λ?

Минимумы выпуклых функций

Доказательство

Пусть x^* – локальный минимум. Рассмотрим точку вида

$$x_\lambda = \lambda x + (1 - \lambda)x^*,$$

где x – произвольная точка из $\mathcal{X}$. **Вопрос:** что можно сказать про x_λ?
$x_\lambda \in \mathcal{X}$ в силу выпуклости $\mathcal{X}$.

Минимумы выпуклых функций

Доказательство

Пусть x^* – локальный минимум. Рассмотрим точку вида

$$x_\lambda = \lambda x + (1 - \lambda)x^*,$$

где x – произвольная точка из $\mathcal{X}$. **Вопрос:** что можно сказать про x_λ? $x_\lambda \in \mathcal{X}$ в силу выпуклости $\mathcal{X}$. Подберем $\lambda > 0$ достаточно малым, что x_λ попадает в окрестность, где x^* локальный минимум.

Минимумы выпуклых функций

Доказательство

Пусть x^* – локальный минимум. Рассмотрим точку вида

$$x_\lambda = \lambda x + (1-\lambda)x^*,$$

где x – произвольная точка из $\mathcal{X}$. **Вопрос:** что можно сказать про x_λ? $x_\lambda \in \mathcal{X}$ в силу выпуклости $\mathcal{X}$. Подберем $\lambda > 0$ достаточно малым, что x_λ попадает в окрестность, где x^* локальный минимум.Тогда уже по выпуклости f

$$f(x^*) \leq f(x_\lambda) \leq \lambda f(x) + (1-\lambda)f(x^*).$$

Минимумы выпуклых функций

Доказательство

Пусть x^* – локальный минимум. Рассмотрим точку вида

$$x_\lambda = \lambda x + (1 - \lambda)x^*,$$

где x – произвольная точка из $\mathcal{X}$. **Вопрос:** что можно сказать про x_λ? $x_\lambda \in \mathcal{X}$ в силу выпуклости $\mathcal{X}$. Подберем $\lambda > 0$ достаточно малым, что x_λ попадает в окрестность, где x^* локальный минимум.Тогда уже по выпуклости f

$$f(x^*) \leq f(x_\lambda) \leq \lambda f(x) + (1 - \lambda)f(x^*).$$

Вопрос: что получили?

Минимумы выпуклых функций

Доказательство

Пусть x^* – локальный минимум. Рассмотрим точку вида

$$x_\lambda = \lambda x + (1-\lambda)x^*,$$

где x – произвольная точка из $\mathcal{X}$. **Вопрос:** что можно сказать про x_λ? $x_\lambda \in \mathcal{X}$ в силу выпуклости $\mathcal{X}$. Подберем $\lambda > 0$ достаточно малым, что x_λ попадает в окрестность, где x^* локальный минимум.Тогда уже по выпуклости f

$$f(x^*) \leq f(x_\lambda) \leq \lambda f(x) + (1-\lambda) f(x^*).$$

Вопрос: что получили? $f(x) \geq f(x^*)$. В силу произвольности $x \in \mathcal{X}$ минимум из локального превратился в глобальный.

Минимумы выпуклых функций

Теорема о минимумах выпуклых функций

Рассмотрим задачу

$$\min_{x \in \mathcal{X}} f(x),$$

где f – выпуклая, $\mathcal{X}$ - выпуклое. Тогда множество точек минимума $\mathcal{X}^*$ выпукло.

Минимумы выпуклых функций

Доказательство

Пустое множество и множество из 1 точки выпуклы.

Минимумы выпуклых функций

Доказательство

Пустое множество и множество из 1 точки выпуклы. Пусть теперь $x_1^*, x_2^* \in \mathcal{X}^*$. Рассмотрим $x_\lambda^* = \lambda x_1^* + (1-\lambda)x_2^*$, где $\lambda \in [0;1]$. $x_\lambda^* \in \mathcal{X}$ в силу выпуклости $\mathcal{X}$.

Минимумы выпуклых функций

Доказательство

Пустое множество и множество из 1 точки выпуклы. Пусть теперь $x_1^*, x_2^* \in \mathcal{X}^*$. Рассмотрим $x_\lambda^* = \lambda x_1^* + (1-\lambda)x_2^*$, где $\lambda \in [0;1]$. $x_\lambda^* \in \mathcal{X}$ в силу выпуклости $\mathcal{X}$.
В силу выпуклости функции f:

$$f^* \leq f(x_\lambda^*) \leq \lambda f(x_1^*) + (1-\lambda) f(x_2^*) = f^*.$$

Минимумы выпуклых функций

Доказательство

Пустое множество и множество из 1 точки выпуклы. Пусть теперь $x_1^*, x_2^* \in \mathcal{X}^*$. Рассмотрим $x_\lambda^* = \lambda x_1^* + (1-\lambda)x_2^*$, где $\lambda \in [0;1]$. $x_\lambda^* \in \mathcal{X}$ в силу выпуклости $\mathcal{X}$.
В силу выпуклости функции f:

$$f^* \leq f(x_\lambda^*) \leq \lambda f(x_1^*) + (1-\lambda)f(x_2^*) = f^*.$$

Откуда $f(x_\lambda^*) = f^*$, а значит $x_\lambda^* \in \mathcal{X}^*$.

Минимумы выпуклых функций

Теорема о минимумах выпуклых функций

Рассмотрим задачу

$$\min_{x \in \mathcal{X}} f(x),$$

где f – *сильно* выпуклая, $\mathcal{X}$ - выпуклое. Тогда множество точек минимума $\mathcal{X}^*$ может состоять только из одного элемента.

Минимумы выпуклых функций

Доказательство

От противного: пусть есть $x_1^* \neq x_2^* \in \mathcal{X}^*$. Рассмотрим $x_\lambda^* = \lambda x_1^* + (1-\lambda)x_2^*$, где $\lambda \in (0;1)$. Опять же $x_\lambda^* \in \mathcal{X}$ в силу выпуклости $\mathcal{X}$.

Минимумы выпуклых функций

Доказательство

От противного: пусть есть $x_1^* \neq x_2^* \in \mathcal{X}^*$. Рассмотрим $x_\lambda^* = \lambda x_1^* + (1-\lambda)x_2^*$, где $\lambda \in (0;1)$. Опять же $x_\lambda^* \in \mathcal{X}$ в силу выпуклости $\mathcal{X}$.

Но теперь в силу сильной выпуклости функции f:

$$f^* \leq f(x_\lambda^*) \leq \lambda f(x_1^*) + (1-\lambda) f(x_2^*) - \lambda(1-\lambda)\frac{\mu}{2}\|x_1^* - x_2^*\|_2^2$$
$$= f^* - \lambda(1-\lambda)\frac{\mu}{2}\|x_1^* - x_2^*\|_2^2.$$

Минимумы выпуклых функций

Доказательство

От противного: пусть есть $x_1^* \neq x_2^* \in \mathcal{X}^*$. Рассмотрим $x_\lambda^* = \lambda x_1^* + (1-\lambda)x_2^*$, где $\lambda \in (0;1)$. Опять же $x_\lambda^* \in \mathcal{X}$ в силу выпуклости $\mathcal{X}$.

Но теперь в силу сильной выпуклости функции f:

$$f^* \leq f(x_\lambda^*) \leq \lambda f(x_1^*) + (1-\lambda)f(x_2^*) - \lambda(1-\lambda)\frac{\mu}{2}\|x_1^* - x_2^*\|_2^2$$
$$= f^* - \lambda(1-\lambda)\frac{\mu}{2}\|x_1^* - x_2^*\|_2^2.$$

Последнее слагаемое < 0 в силу выбора $x_1^* \neq x_2^*$ и $\lambda \in (0;1)$. Противоречие.

Минимумы выпуклых функций

Теорема о минимумах выпуклых функций

Рассмотрим задачу

$$\min_{x \in \mathcal{X}} f(x),$$

где f – *сильно* выпуклая, $\mathcal{X}$ - выпуклое. Тогда множество точек минимума $\mathcal{X}^*$ может состоять только из одного элемента.

- На самом деле для сильно выпуклой функции можно доказать, что решение строго единственное (т.е. добавить к предыдущей теореме существование). Это следует из того, что мы снизу всегда подперты параболой. Смотри док-во в конспекте.

Сильная выпуклость: больше фактов

Теорема об еще одном определении сильной выпуклости

Пусть функция $f : \mathbb{R}^d \to \mathbb{R}$ непрерывно дифференцируема на $\mathbb{R}^d$. Тогда функция f является μ-сильно выпуклой тогда и только тогда, когда для любых $x, y \in \mathbb{R}^d$ выполнено

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2.$$

Сильная выпуклость: больше фактов

Теорема об еще одном определении сильной выпуклости

Пусть функция $f : \mathbb{R}^d \to \mathbb{R}$ непрерывно дифференцируема на $\mathbb{R}^d$. Тогда функция f является μ-сильно выпуклой тогда и только тогда, когда для любых $x, y \in \mathbb{R}^d$ выполнено

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2.$$

Теорема о критерии сильной выпуклости

Пусть функция $f : \mathbb{R}^d \to \mathbb{R}$ дважды непрерывно дифференцируема на $\mathbb{R}^d$. Тогда функция f является μ-сильно выпуклой тогда и только тогда, когда для любого $x \in \mathbb{R}^d$ выполнено

$$\nabla^2 f(x) \succeq \mu I.$$

Сильная выпуклость: больше фактов

Теорема об еще одном определении сильной выпуклости

Пусть функция $f : \mathbb{R}^d \to \mathbb{R}$ непрерывно дифференцируема на $\mathbb{R}^d$. Тогда функция f является μ-сильно выпуклой тогда и только тогда, когда для любых $x, y \in \mathbb{R}^d$ выполнено

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2.$$

Теорема о критерии сильной выпуклости

Пусть функция $f : \mathbb{R}^d \to \mathbb{R}$ дважды непрерывно дифференцируема на $\mathbb{R}^d$. Тогда функция f является μ-сильно выпуклой тогда и только тогда, когда для любого $x \in \mathbb{R}^d$ выполнено

$$\nabla^2 f(x) \succeq \mu I.$$

Оба факта доказаны в пособии. Второй пригодится для ДЗ.