ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Метод Ньютона и квазиньютоновские методы

Методы выпуклой оптимизации

НЕДЕЛЯ 11

Даня Меркулов
Пётр Остроухов

Условные градиентные методы. Метод проекции градиента. Метод Франк–Вульфа.

Семинар

Оптимизация для всех! ЦУ

Повтор лекции. Проекция

Проекция

Расстояние d от точки $\mathbf{y} \in \mathbb{R}^n$ до замкнутого множества $S \subset \mathbb{R}^n$:

$$d(\mathbf{y}, S, \|\cdot\|) = \text{inf}\{\|x - y\| \mid x \in S\}.$$

Мы будем фокусироваться на **евклидовой проекции** (возможны и другие варианты) точки $\mathbf{y} \in \mathbb{R}^n$ на множество $S \subseteq \mathbb{R}^n$. Это точка $\text{proj}_S(\mathbf{y}) \in S$ такая, что

$$\text{proj}_S(\mathbf{y}) = \underset{\mathbf{x} \in S}{\text{argmin}} \; \frac{1}{2}\|x - y\|_2^2.$$

- **Достаточное условие существования проекции.** Если $S \subseteq \mathbb{R}^n$ — замкнутое множество, то проекция на множество S существует для любой точки.

Проекция

Расстояние d от точки $\mathbf{y} \in \mathbb{R}^n$ до замкнутого множества $S \subset \mathbb{R}^n$:

$$d(\mathbf{y}, S, \|\cdot\|) = \inf\{\|x - y\| \mid x \in S\}.$$

Мы будем фокусироваться на **евклидовой проекции** (возможны и другие варианты) точки $\mathbf{y} \in \mathbb{R}^n$ на множество $S \subseteq \mathbb{R}^n$. Это точка $\text{proj}_S(\mathbf{y}) \in S$ такая, что

$$\text{proj}_S(\mathbf{y}) = \underset{\mathbf{x} \in S}{\text{argmin}} \, \frac{1}{2}\|x - y\|_2^2.$$

- **Достаточное условие существования проекции.** Если $S \subseteq \mathbb{R}^n$ — замкнутое множество, то проекция на множество S существует для любой точки.
- **Достаточное условие единственности проекции.** Если $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, то проекция на множество S единственна для любой точки.

Проекция

Расстояние d от точки $\mathbf{y} \in \mathbb{R}^n$ до замкнутого множества $S \subset \mathbb{R}^n$:

$$d(\mathbf{y}, S, \|\cdot\|) = \inf\{\|x - y\| \mid x \in S\}.$$

Мы будем фокусироваться на **евклидовой проекции** (возможны и другие варианты) точки $\mathbf{y} \in \mathbb{R}^n$ на множество $S \subseteq \mathbb{R}^n$. Это точка $\text{proj}_S(\mathbf{y}) \in S$ такая, что

$$\text{proj}_S(\mathbf{y}) = \underset{\mathbf{x} \in S}{\text{argmin}}\ \frac{1}{2}\|x - y\|_2^2.$$

- **Достаточное условие существования проекции.** Если $S \subseteq \mathbb{R}^n$ — замкнутое множество, то проекция на множество S существует для любой точки.
- **Достаточное условие единственности проекции.** Если $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, то проекция на множество S единственна для любой точки.
- Если множество открыто, и точка лежит вне этого множества, то её проекция на это множество может не существовать.

Проекция

Расстояние d от точки $\mathbf{y} \in \mathbb{R}^n$ до замкнутого множества $S \subset \mathbb{R}^n$:

$$d(\mathbf{y}, S, \|\cdot\|) = \inf\{\|x - y\| \mid x \in S\}.$$

Мы будем фокусироваться на **евклидовой проекции** (возможны и другие варианты) точки $\mathbf{y} \in \mathbb{R}^n$ на множество $S \subseteq \mathbb{R}^n$. Это точка $\text{proj}_S(\mathbf{y}) \in S$ такая, что

$$\text{proj}_S(\mathbf{y}) = \underset{\mathbf{x} \in S}{\text{argmin}} \ \frac{1}{2}\|x - y\|_2^2.$$

- **Достаточное условие существования проекции.** Если $S \subseteq \mathbb{R}^n$ — замкнутое множество, то проекция на множество S существует для любой точки.
- **Достаточное условие единственности проекции.** Если $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, то проекция на множество S единственна для любой точки.
- Если множество открыто, и точка лежит вне этого множества, то её проекция на это множество может не существовать.
- Если точка лежит внутри множества, то её проекция — это сама точка.

Проекция

ℹ Theorem

Пусть $S \subseteq \mathbb{R}^n$ — замкнутое и выпуклое множество, и пусть $x \in S, y \in \mathbb{R}^n$. Тогда

$$\langle y - \text{proj}_S(y), \mathbf{x} - \text{proj}_S(y)\rangle \leq 0, \tag{1}$$

$$\|x - \text{proj}_S(y)\|^2 + \|y - \text{proj}_S(y)\|^2 \leq \|x - y\|^2. \tag{2}$$

Нерастягивающее отображение

Отображение (функция) f называется **нерастягивающим**, если оно является L-Липшицевым с константой $L \leq 1$ [1]. То есть для любых двух точек $x, y \in \text{dom} f$ выполнено

$$\|f(x) - f(y)\| \leq L\|x - y\|, \quad \text{где } L \leq 1.$$

Это означает, что расстояние между образами точек не больше (и может быть меньше), чем расстояние между исходными точками.

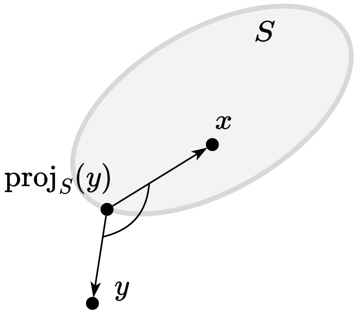

Рисунок 1. Тупой или прямой угол должен получаться для любой точки $x \in S$

[1] Нерастягивающее становится сжимающим, если $L < 1$.

Задачи

Задача. Проекция на неотрицательный ортант

Пусть $\mathcal{S}$ — неотрицательный ортант. Найдите проекцию

$$\text{proj}_{\mathcal{S}}(\mathbf{y}) = \arg\min_{\mathbf{x} \geq 0} \|\mathbf{x} - \mathbf{y}\|_2,$$

где $\mathbf{x} \geq 0$ означает, что $\mathbf{x}$ лежит в неотрицательном ортанге

$$\mathcal{S} = \{\mathbf{x} \mid x_i \geq 0 \ \forall i\}.$$

Что если $\mathcal{S} = \{x \mid l \leq x \leq u\}$ (покоординатные нижние и верхние границы)?

Задача. Проекция на множество 1-Липшицевых матриц

Скажем, что матрица $A \in \mathbb{R}^{m\times n}$ является L-**Липшицевой** (относительно евклидовой нормы), если для любых двух векторов $x, y \in \mathbb{R}^n$ выполнено

$$\|Ax - Ay\|_2 \leq L\|x - y\|_2.$$

Множество всех таких матриц обозначим

$$\mathcal{L}_L = \{A \in \mathbb{R}^{m\times n} \mid A \text{ — } L\text{-Липшицева}\}.$$

Теперь зафиксируем самый простой случай: $L = 1$ и будем мерить расстояние между матрицами в **норме Фробениуса**:

$$\|X - M\|_F^2 = \sum_{i,j}(X_{ij} - M_{ij})^2.$$

Задача. По данной матрице $M \in \mathbb{R}^{m\times n}$ найти матрицу $X \in \mathcal{L}_1$, которая ближе всего к M в норме Фробениуса для случая $L = 1$:

$$\min_{X\in\mathcal{L}_1} \frac{1}{2}\|X - M\|_F^2.$$

Решение

1. Для линейного отображения A константа Липшица по евклидовой норме равна **операторной норме**:

$$\|Ax - Ay\|_2 \leq L\|x - y\|_2 \ \ \forall x, y \quad \Longleftrightarrow \quad \|A\|_2 \leq L,$$

   где $\|A\|_2$ — спектральная норма (наибольшее сингулярное значение).
   В нашем случае $L = 1$, значит

$$\mathcal{L}_1 = \{A \mid \|A\|_2 \leq 1\}.$$

Решение

1. Для линейного отображения A константа Липшица по евклидовой норме равна **операторной норме**:

$$\|Ax - Ay\|_2 \le L\|x - y\|_2 \;\; \forall x, y \quad \Longleftrightarrow \quad \|A\|_2 \le L,$$

   где $\|A\|_2$ — спектральная норма (наибольшее сингулярное значение).
   В нашем случае $L = 1$, значит

$$\mathcal{L}_1 = \{A \mid \|A\|_2 \le 1\}.$$

2. Запишеи сингулярное разложение матрицы M: $M = U\Sigma V^\top$, где $\Sigma =$ diag$(\sigma_1, \dots, \sigma_r),\; \sigma_1 \ge \dots \ge \sigma_r \ge 0$.

Решение

1. Для линейного отображения A константа Липшица по евклидовой норме равна **операторной норме**:

$$\|Ax - Ay\|_2 \leq L\|x - y\|_2 \;\; \forall x, y \quad \Longleftrightarrow \quad \|A\|_2 \leq L,$$

   где $\|A\|_2$ — спектральная норма (наибольшее сингулярное значение).
   В нашем случае $L = 1$, значит

$$\mathcal{L}_1 = \{A \mid \|A\|_2 \leq 1\}.$$

2. Запишеи сингулярное разложение матрицы M: $M = U\Sigma V^\top$, где $\Sigma =$ diag$(\sigma_1, \ldots, \sigma_r), \; \sigma_1 \geq \ldots \geq \sigma_r \geq 0.$
3. Так как норма в минимизируемой функции является инвариантной к умножению на ортогональную матрицу, то можем искать X в виде $X = U\Sigma' V^\top$ и подбираем новые сингулярные числа σ_i' так, чтобы

$$\|X\|_2 = \max_i \sigma_i' \leq 1$$

Решение

1. Для линейного отображения A константа Липшица по евклидовой норме равна **операторной норме**:

$$\|Ax - Ay\|_2 \leq L\|x - y\|_2 \ \ \forall x, y \quad \Longleftrightarrow \quad \|A\|_2 \leq L,$$

где $\|A\|_2$ — спектральная норма (наибольшее сингулярное значение).
В нашем случае $L = 1$, значит

$$\mathcal{L}_1 = \{A \mid \|A\|_2 \leq 1\}.$$

2. Запишеи сингулярное разложение матрицы M: $M = U\Sigma V^\top$, где $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r),\ \sigma_1 \geq \dots \geq \sigma_r \geq 0.$
3. Так как норма в минимизируемой функции является инвариантной к умножению на ортогональную матрицу, то можем искать X в виде $X = U\Sigma' V^\top$ и подбираем новые сингулярные числа σ_i' так, чтобы

$$\|X\|_2 = \max_i \sigma_i' \leq 1$$

4. Тогда минимизируемая функция:

$$\begin{aligned}
&\|X - M\|_F^2 = \langle X - M, X - M\rangle \\
&= \langle U\Sigma' V^\top - U\Sigma V^\top, U\Sigma' V^\top - U\Sigma V^\top\rangle = \\
&= \langle U(\Sigma' - \Sigma)V^\top, U(\Sigma' - \Sigma)V^\top\rangle = \\
&= V(\Sigma' - \Sigma)^T U^T U(\Sigma' - \Sigma)V^\top = \\
&= V(\Sigma' - \Sigma)^2 V^\top = \\
&= \|(\Sigma' - \Sigma)V^\top\|_F^2 = \\
&= \|(\Sigma' - \Sigma)\|_F^2 = \\
&= \sum_i (\sigma_i' - \sigma_i)^2 \to \min .
\end{aligned}$$

Решение

1. Для линейного отображения A константа Липшица по евклидовой норме равна **операторной норме**:

$$\|Ax - Ay\|_2 \le L\|x - y\|_2 \ \ \forall x, y \quad \Longleftrightarrow \quad \|A\|_2 \le L,$$

где $\|A\|_2$ — спектральная норма (наибольшее сингулярное значение).
В нашем случае $L = 1$, значит

$$\mathcal{L}_1 = \{A \mid \|A\|_2 \le 1\}.$$

2. Запишеи сингулярное разложение матрицы M: $M = U\Sigma V^\top$, где $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r),\ \sigma_1 \ge \dots \ge \sigma_r \ge 0.$
3. Так как норма в минимизируемой функции является инвариантной к умножению на ортогональную матрицу, то можем искать X в виде $X = U\Sigma' V^\top$ и подбираем новые сингулярные числа σ_i' так, чтобы

$$\|X\|_2 = \max_i \sigma_i' \le 1$$

4. Тогда минимизируемая функция:

$$\begin{aligned}
&\|X - M\|_F^2 = \langle X - M, X - M\rangle \\
&= \langle U\Sigma' V^\top - U\Sigma V^\top, U\Sigma' V^\top - U\Sigma V^\top\rangle = \\
&= \langle U(\Sigma' - \Sigma)V^\top, U(\Sigma' - \Sigma)V^\top\rangle = \\
&= V(\Sigma' - \Sigma)^T U^T U(\Sigma' - \Sigma)V^\top = \\
&= V(\Sigma' - \Sigma)^2 V^\top = \\
&= \|(\Sigma' - \Sigma)V^\top\|_F^2 = \\
&= \|(\Sigma' - \Sigma)\|_F^2 = \\
&= \sum_i (\sigma_i' - \sigma_i)^2 \to \min .
\end{aligned}$$

5. Задача распадается **покоординатно**:

$$\min_{\sigma_i' \le 1} (\sigma_i' - \sigma_i)^2 \quad \Longrightarrow \quad \sigma_i' = \min(\sigma_i, 1).$$

Задача. Проекция на спектраплекс

Спектраплекс — это спектраэдр, определённый как множество

$$\mathcal{S} := \left\{ X \in \mathbb{S}^n_+ : \operatorname{Tr} X = 1 \right\},$$

где $\mathbb{S}^n_+$ — множество симметричных положительно полуопределённых матриц размера $n \times n$.

Spectraplex = «spectra» + «simplex», в смысле «собственные значения лежат в симплексе».
Спектраплекс — это «полуопределённый» аналог симплекса.

Вопрос. По данной матрице $Z \in \mathbb{R}^{n\times n}$, как найти проекцию Z на множество $\mathcal{S}$?

Иными словами, нужно решить задачу

$$\operatorname*{arg\,min}_{X \succeq 0,\, \operatorname{Tr} X = 1} \frac{1}{2}\|X - Z\|_F^2.$$

Повтор лекции. Метод проекции градиента (PGD)

Идея метода проекции градиента

$$
x_{k+1} = \text{proj}_S\left(x_k - \alpha_k \nabla f(x_k)\right) \quad \Leftrightarrow \quad \begin{aligned} y_k &= x_k - \alpha_k \nabla f(x_k), \\ x_{k+1} &= \text{proj}_S\left(y_k\right). \end{aligned}
$$

Ниже можно найти пример использования этого метода для атаки нейросети (adversarial attack):
Adversarial Attacks.

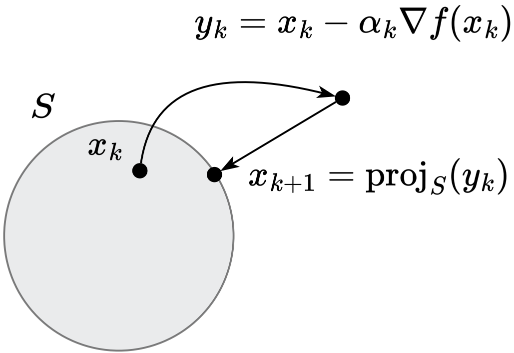

Рисунок 2. Иллюстрация алгоритма метода проекции градиента

Повтор лекции. Метод Франк–Вульфа

Метод Франк–Вульфа (FWM). Идея

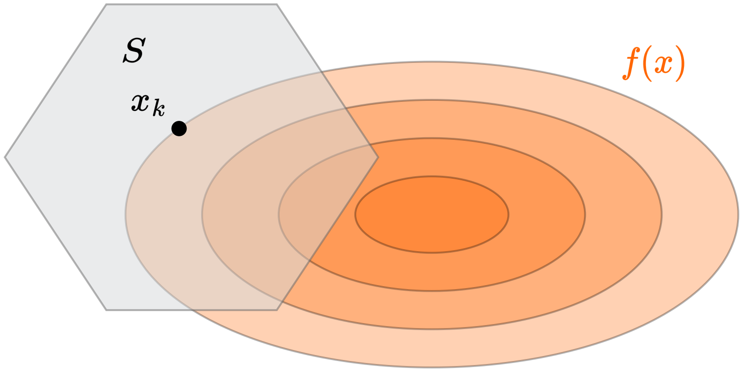

Рисунок 3. Иллюстрация метода Франк–Вульфа (метод условного градиента)

Метод Франк–Вульфа (FWM). Идея

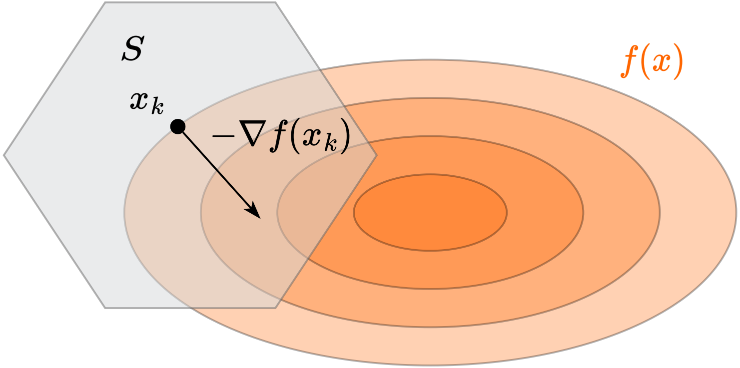

Рисунок 4. Иллюстрация метода Франк–Вульфа (метод условного градиента)

Метод Франк–Вульфа (FWM). Идея

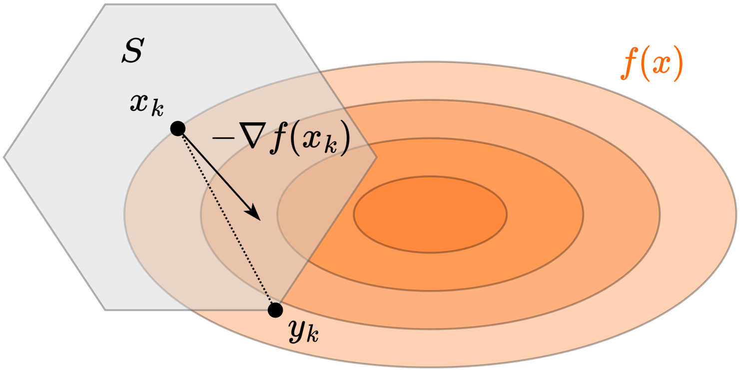

Рисунок 5. Иллюстрация метода Франк–Вульфа (метод условного градиента)

Метод Франк–Вульфа (FWM). Идея

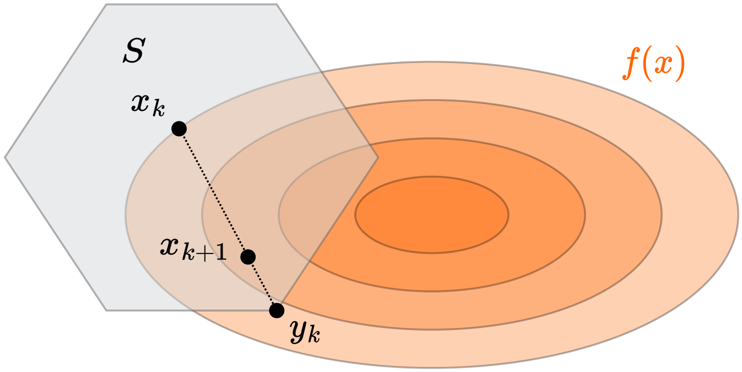

Рисунок 6. Иллюстрация метода Франк–Вульфа (метод условного градиента)

Метод Франк–Вульфа (FWM). Идея

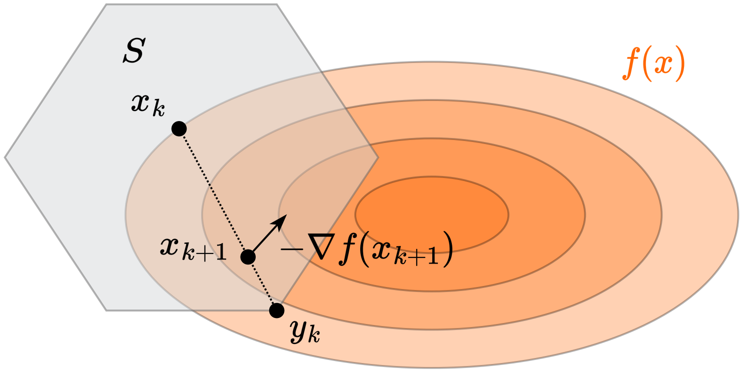

Рисунок 7. Иллюстрация метода Франк–Вульфа (метод условного градиента)

Метод Франк–Вульфа (FWM). Идея

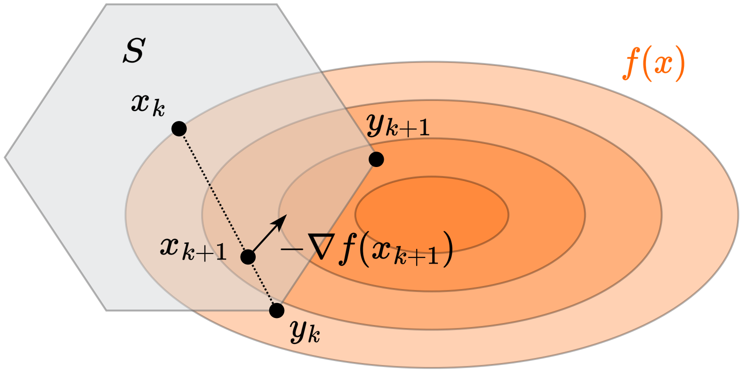

Рисунок 8. Иллюстрация метода Франк–Вульфа (метод условного градиента)

Метод Франк–Вульфа (FWM). Идея

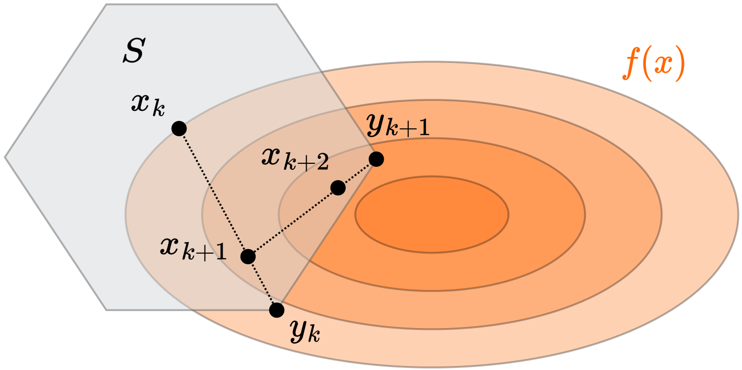

Рисунок 9. Иллюстрация метода Франк–Вульфа (метод условного градиента)

Метод Франк–Вульфа (FWM). Идея

$$y_k = \underset{x \in S}{\text{arg min}}\, f^I_{x_k}(x) = \underset{x \in S}{\text{arg min}} \langle \nabla f(x_k), x \rangle,$$

$$x_{k+1} = \gamma_k x_k + (1 - \gamma_k) y_k.$$

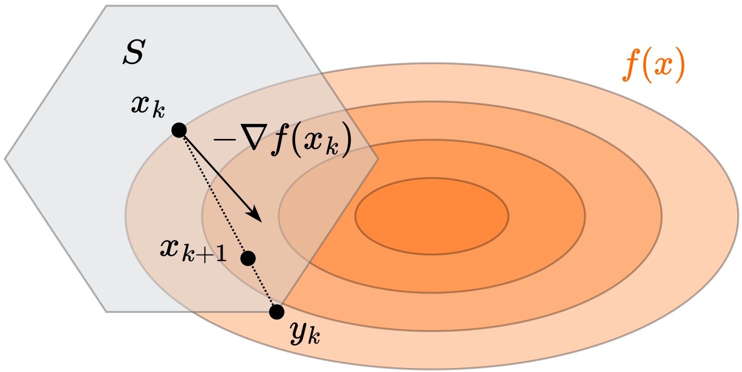

Рисунок 10. Иллюстрация метода Франк–Вульфа (метод условного градиента)

Скорости сходимости

Скорость сходимости в гладком выпуклом случае

ℹ Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — выпуклая L-гладкая функция. Пусть $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, и пусть существует минимизатор x^* функции f на S.

- **Метод проекции градиента** с шагом $\frac{1}{L}$ достигает следующей оценки после итерации $k > 0$:

$$f(x_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2k}.$$

Скорость сходимости в гладком выпуклом случае

Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — выпуклая L-гладкая функция. Пусть $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, и пусть существует минимизатор x^* функции f на S.

- **Метод проекции градиента** с шагом $\frac{1}{L}$ достигает следующей оценки после итерации $k > 0$:

$$f(x_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2k}.$$

- **Метод Франк–Вульфа** достигает следующей оценки после итерации $k > 0$:

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

Особенности метода Франк–Вульфа

- Скорость сходимости метода Франк–Вульфа для μ-сильно выпуклых функций — $\mathcal{O}\left(\frac{1}{k}\right)$.

Особенности метода Франк–Вульфа

- Скорость сходимости метода Франк–Вульфа для μ-сильно выпуклых функций — $\mathcal{O}\left(\frac{1}{k}\right)$.
- В базовой форме метод не работает для негладких функций. Но существуют модификации, которые с этим справляются.

Особенности метода Франк–Вульфа

- Скорость сходимости метода Франк–Вульфа для μ-сильно выпуклых функций — $\mathcal{O}\left(\frac{1}{k}\right)$.
- В базовой форме метод не работает для негладких функций. Но существуют модификации, которые с этим справляются.
- Метод Франк–Вульфа корректно работает для любой нормы.

Бонус: Зеркальный спуск

Метод субградиентного спуска: линейная аппроксимация + проксимальность

Вспомним шаг SubGD с субградиентом g_k:

$$x_{k+1} = x_k - \alpha_k g_k \quad \Leftrightarrow \quad \begin{aligned} x_{k+1} &= \operatorname*{argmin}_x \underbrace{f(x_k) + g_k^\top (x - x_k)}_{\text{линейная аппроксимация } f} + \underbrace{\frac{1}{2\alpha}\|x - x_k\|_2^2}_{\text{проксимальный член}} \\ &= \operatorname*{argmin}_x \alpha g_k^\top x + \frac{1}{2}\|x - x_k\|_2^2. \end{aligned}$$

Идея зеркального спуска: заменить евклидову проксимальность $\|x - x_k\|_2^2$ на другую, более подходящую для задачи меру близости.

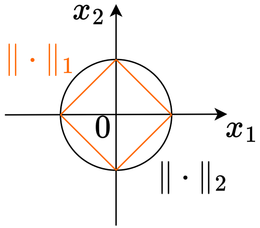

Рисунок 11. $\|\cdot\|_1$ не сферически симметрична

Пример. Плохая обусловленность

Рассмотрим функцию

$$f(x_1, x_2) = x_1^2 \cdot \frac{1}{100} + x_2^2 \cdot 100.$$

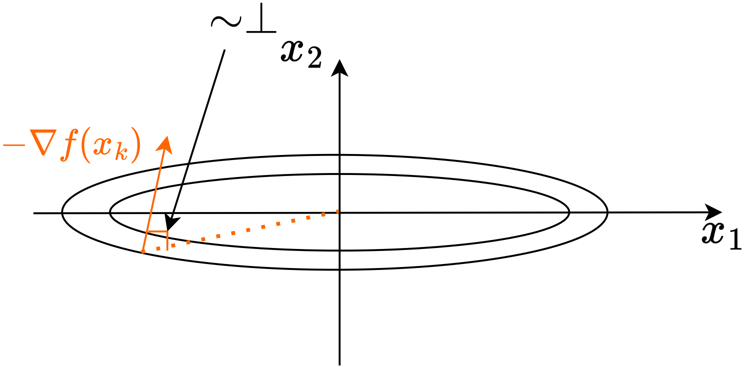

Рисунок 12. Плохо обусловленная задача в норме $\|\cdot\|_2$

Пример. Плохая обусловленность

Пусть мы находимся в точке $x_k = (-10 \quad -0.1)^\top$. Метод градиентного спуска: $x_{k+1} = x_k - \alpha \nabla f(x_k)$, где

$$\nabla f(x_k) = \begin{pmatrix} \frac{2x_1}{100} & 2x_2 \cdot 100 \end{pmatrix}^\top \Bigg|_{(-10 \; -0.1)^\top} = \begin{pmatrix} -\frac{1}{5} & -20 \end{pmatrix}^\top.$$

Проблема: из-за сильной вытянутости линий уровня направление движения $(x_{k+1} - x_k)$ оказывается почти перпендикулярно вектору $(x^* - x_k)$. **Решение:** поменять проксимальный член:

$$x_{k+1} = \underset{x}{\text{argmin}} \; \underbrace{f(x_k) + g_k^\top (x - x_k)}_{\text{линейная аппроксимация } f} + \underbrace{\frac{1}{2\alpha}(x - x_k)^\top I (x - x_k)}_{\text{проксимальный член}}$$

на другой:

$$x_{k+1} = \underset{x}{\text{argmin}} \; \underbrace{f(x_k) + g_k^\top (x - x_k)}_{\text{линейная аппроксимация } f} + \underbrace{\frac{1}{2\alpha}(x - x_k)^\top Q (x - x_k)}_{\text{проксимальный член}},$$

где в этом примере

$$Q = \begin{pmatrix} \frac{1}{50} & 0 \\ 0 & 200 \end{pmatrix}.$$

Более общая идея — заменить квадратичную форму на произвольную функцию $B_\phi(x, y)$, измеряющую «близость» x и y.

Пример. Плохая обусловленность

Найдём x_{k+1} для **нового** алгоритма:

$$\alpha \nabla f(x_k) + \begin{pmatrix} \frac{1}{50} & 0 \\ 0 & 200 \end{pmatrix} (x - x_k) = 0.$$

Решая это уравнение относительно x, получаем

$$x_{k+1} = x_k - \alpha \begin{pmatrix} 50 & 0 \\ 0 & \frac{1}{200} \end{pmatrix} \nabla f(x_k) = (-10 \;\; -0.1)^\top - \alpha(-10 \;\; -0.1)^\top.$$

Наблюдение. Меняя проксимальный член, мы **меняем направление** приращения $x_{k+1} - x_k$.
Иначе говоря, если мы измеряем расстояние «по-новому», мы тем самым **меняем Липшицевость** функции (константу Липшица относительно новой нормы).

> **Question**
>
> Чему равна константа Липшица функции f в точке $(1 \; 1)^\top$ относительно нормы
>
> $$\|z\|_A^2 = z^\top \begin{pmatrix} 50 & 0 \\ 0 & \frac{1}{200} \end{pmatrix} z?$$

Пример. Robust Regression (устойчивая регрессия)

Квадратичная ошибка $\|Ax - b\|_2^2$ очень чувствительна к выбросам.

Вместо этого можно рассматривать

$$\min_x \|Ax - b\|_1.$$

Эта задача тоже **выпуклая**.

Посчитаем константу Липшица L для $f(x) = \|Ax - b\|_1$:

$$\big|\|Ax - b\|_1 - \|Ay - b\|_1\big| \leq L\|x - y\|_2.$$

Для упрощения возьмём $A = I$, $b = 0$, то есть $f(x) = \|x\|_1$.

Возьмём $x = \mathbf{1}_d$, $y = (1 + \varepsilon)\mathbf{1}_d$:

$$\big|\|x\|_1 - \|y\|_1\big| = \big|n - (1 + \varepsilon)n\big| = \varepsilon n \leq L\|x - y\|_2 = L\| - \varepsilon\mathbf{1}_d\|_2 = L\sqrt{n\varepsilon^2} = L\varepsilon\sqrt{n}.$$

Итак, получаем $L = \sqrt{n}$. Видно, что L **зависит от размерности**.

> ℹ Question
>
> Покажите, что если $\|\nabla f(x)\|_\infty \leq 1$, то $\|\nabla f(x)\|_2 \leq \sqrt{n}$.

Литература

Литература

Примеры для зеркального спуска были взяты из лекции.