Метод штрафов. ADMM

Методы оптимизации

Александр Безносиков

Московский физико-технический институт

13 ноября 2025

ФПМИ

Штрафная функция

- Рассмотрим следующую задачу с ограничениями:

$$\min_{x \in \mathbb{R}^d} \quad f(x),$$
$$\text{s.t.} \quad h_i(x) = 0, \;\; i = 1, \ldots m.$$

Штрафная функция

- Рассмотрим следующую задачу с ограничениями:

$$\min_{x \in \mathbb{R}^d} \ f(x), \\ \text{s.t. } \ h_i(x) = 0, \ i = 1, \ldots m.$$

- Возьмем некоторое $\rho > 0$ и немного модифицируем нашу задачу:

$$\min_{x \in \mathbb{R}^d} \ f(x) + \rho \cdot \frac{1}{2} \sum_{i=1}^{m} h_i^2(x), \\ \text{s.t. } \ h_i(x) = 0, \ i = 1, \ldots m.$$

Вопрос: что можно сказать о новой задаче?

Штрафная функция

- Рассмотрим следующую задачу с ограничениями:

$$\min_{x \in \mathbb{R}^d} \; f(x), \quad \text{s.t.} \;\; h_i(x) = 0, \; i = 1, \ldots m.$$

- Возьмем некоторое $\rho > 0$ и немного модифицируем нашу задачу:

$$\min_{x \in \mathbb{R}^d} \; f(x) + \rho \cdot \frac{1}{2} \sum_{i=1}^{m} h_i^2(x), \quad \text{s.t.} \;\; h_i(x) = 0, \; i = 1, \ldots m.$$

Вопрос: что можно сказать о новой задаче? она эквивалентна старой, так как «добавка» равна 0 для x, удовлетворяющих ограничениям.

Штрафная функция

А теперь сделаем вот так:

$$\min_{x \in \mathbb{R}^d} \left[f_\rho(x) = f(x) + \rho \cdot \frac{1}{2} \sum_{i=1}^{m} h_i^2(x) \right].$$

Штрафная функция

А теперь сделаем вот так:

$$\min_{x \in \mathbb{R}^d} \left[f_\rho(x) = f(x) + \rho \cdot \frac{1}{2} \sum_{i=1}^{m} h_i^2(x) \right].$$

Вопрос: осталась ли задача эквивалента исходной?

Штрафная функция

А теперь сделаем вот так:

$$\min_{x \in \mathbb{R}^d} \left[f_\rho(x) = f(x) + \rho \cdot \frac{1}{2} \sum_{i=1}^{m} h_i^2(x) \right].$$

Вопрос: осталась ли задача эквивалента исходной? нет!

Штрафная функция

А теперь сделаем вот так:

$$\min_{x\in\mathbb{R}^d}\left[f_\rho(x) = f(x) + \rho\cdot\frac{1}{2}\sum_{i=1}^m h_i^2(x)\right].$$

Вопрос: осталась ли задача эквивалента исходной? нет!

- f_ρ называют штрафной функцией, а ρ – параметром штрафа.
- Задача с ограничениями стала задачей без ограничений.

Штрафная функция

А теперь сделаем вот так:

$$\min_{x \in \mathbb{R}^d} \left[f_\rho(x) = f(x) + \rho \cdot \frac{1}{2} \sum_{i=1}^{m} h_i^2(x) \right].$$

Вопрос: осталась ли задача эквивалента исходной? нет!

- f_ρ называют штрафной функцией, а ρ – параметром штрафа.
- Задача с ограничениями стала задачей без ограничений.
- Решая новую задачу, можно выйти за пределы множества ограничений.

Штрафная функция

А теперь сделаем вот так:

$$\min_{x \in \mathbb{R}^d} \left[f_\rho(x) = f(x) + \rho \cdot \frac{1}{2} \sum_{i=1}^{m} h_i^2(x) \right].$$

Вопрос: осталась ли задача эквивалента исходной? нет!

- f_ρ называют штрафной функцией, а ρ – параметром штрафа.
- Задача с ограничениями стала задачей без ограничений.
- Решая новую задачу, можно выйти за пределы множества ограничений.
- Предельное ρ:

$$\lim_{\rho \to +\infty} f_\rho =$$

Штрафная функция

А теперь сделаем вот так:

$$\min_{x \in \mathbb{R}^d} \left[f_\rho(x) = f(x) + \rho \cdot \frac{1}{2} \sum_{i=1}^{m} h_i^2(x) \right].$$

Вопрос: осталась ли задача эквивалента исходной? нет!

- f_ρ называют штрафной функцией, а ρ – параметром штрафа.
- Задача с ограничениями стала задачей без ограничений.
- Решая новую задачу, можно выйти за пределы множества ограничений.
- Предельное ρ:

$$\lim_{\rho \to +\infty} f_\rho = \begin{cases} f(x), & x \text{ удовлетворяет ограничениям исходной задачи} \\ +\infty, & \text{иначе} \end{cases}$$

Штрафная функция

А теперь сделаем вот так:

$$\min_{x \in \mathbb{R}^d} \left[f_\rho(x) = f(x) + \rho \cdot \frac{1}{2} \sum_{i=1}^{m} h_i^2(x) \right].$$

Вопрос: осталась ли задача эквивалента исходной? нет!

- f_ρ называют штрафной функцией, а ρ – параметром штрафа.
- Задача с ограничениями стала задачей без ограничений.
- Решая новую задачу, можно выйти за пределы множества ограничений.
- Предельное ρ:

$$\lim_{\rho \to +\infty} f_\rho = \begin{cases} f(x), & x \text{ удовлетворяет ограничениям исходной задачи} \\ +\infty, & \text{иначе} \end{cases}$$

- Есть надежда, что минимизируя f_ρ (решая штрафную задачу) для достаточно большого ρ, мы получим неплохое решение и для исходной задачи.

Штрафная функция: ограничения вида неравенств

- Добавим еще ограничения вида неравенств:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & f(x), \\ \text{s.t.} \quad & h_i(x) = 0, \; i = 1, \ldots m, \\ & g_i(x) \leq 0, \; j = 1, \ldots n. \end{aligned}$$

Штрафная функция: ограничения вида неравенств

- Добавим еще ограничения вида неравенств:

$$
\begin{aligned}
\min_{x \in \mathbb{R}^d} \quad & f(x), \\
\text{s.t.} \quad & h_i(x) = 0, \; i = 1, \ldots m, \\
& g_i(x) \leq 0, \; j = 1, \ldots n.
\end{aligned}
$$

Вопрос: как их запихать в штраф?

Штрафная функция: ограничения вида неравенств

- Добавим еще ограничения вида неравенств:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & f(x), \\ \text{s.t.} \quad & h_i(x) = 0, \; i = 1, \ldots m, \\ & g_i(x) \leq 0, \; j = 1, \ldots n. \end{aligned}$$

Вопрос: как их запихать в штраф?

- С помощью «срезки»:

$$f_\rho(x) = f(x) + \rho \cdot \frac{1}{2} \sum_{i=1}^{m} h_i^2(x) + \rho \cdot \frac{1}{2} \sum_{j=1}^{n} (g_j^+)^2(x),$$

где $y^+ = \max\{y, 0\}$. Активируем штраф только, когда нарушено неравенство.

Свойства решений штрафной задачи

Свойства решений штрафной задачи

Пусть x^* – решение исходной задачи, а x^*_ρ – решение соотвествующей штрафной задачи с $\rho > 0$, тогда

$$f(x^*) \geq f(x^*_\rho).$$

Доказательство:

$$f(x^*) = f_\rho(x^*) \geq \min_{x \in \mathbb{R}^d} f_\rho(x) = f_\rho(x^*_\rho) \geq f(x^*_\rho).$$

Свойства решений штрафной задачи

Предыдущий результат говорит о том, что либо нарушаем ограничения, либо $f(x^*) = f(x^*_\rho)$. Но за счет ρ с этим можно бороться. Следующие два свойства про это.

Свойства решений штрафной задачи

Предыдущий результат говорит о том, что либо нарушаем ограничения, либо $f(x^*) = f(x^*_\rho)$. Но за счет ρ с этим можно бороться. Следующие два свойства про это.

Свойства решений штрафной задачи

С увеличение ρ решения штрафной задачи (если существует) гарантировано не ухудшает степень нарушения ограничений, т.е. для $\rho_1 > \rho_2$ следует, что

$$\sum_{i=1}^m h_i^2(x^*_{\rho_2}) \geq \sum_{i=1}^m h_i^2(x^*_{\rho_1}),$$

где $x^*_{\rho_1}$ и $x^*_{\rho_2}$ – решения соответствующих штрафных задач.

Доказательство

- Пользуясь тем, что $x^*_{\rho_1}$ и $x^*_{\rho_2}$ – решения соответствующих штрафных задач:

$$f(x^*_{\rho_2}) + \rho_1 \cdot \frac{1}{2}\sum_{i=1}^{m} h_i^2(x^*_{\rho_2}) \geq f(x^*_{\rho_1}) + \rho_1 \cdot \frac{1}{2}\sum_{i=1}^{m} h_i^2(x^*_{\rho_1})$$

и

$$f(x^*_{\rho_1}) + \rho_2 \cdot \frac{1}{2}\sum_{i=1}^{m} h_i^2(x^*_{\rho_1}) \geq f(x^*_{\rho_2}) + \rho_2 \cdot \frac{1}{2}\sum_{i=1}^{m} h_i^2(x^*_{\rho_2})$$

- Складываем и делим на $(\rho_1 - \rho_2) > 0$:

$$\sum_{i=1}^{m} h_i^2(x^*_{\rho_2}) \geq \sum_{i=1}^{m} h_i^2(x^*_{\rho_1}).$$

Свойства решений штрафной задачи

Свойства решений штрафной задачи

Пусть функция f и все функции h_i $(i = 1, \dots m)$ являются непрерывными. Пусть X^* множество решений исходной условной задачи оптимизации и для $x^* \in X^*$ множество

$$U = \{x \in \mathbb{R}^d \mid f(x) \leq f(x^*)\}$$

ограничено. Тогда для любого $e > 0$ существует $\rho(e) > 0$ такое, что множество решений штрафной задачи X^*_ρ для любых $\rho \geq \rho(e)$ содержится в

$$X^*_e = \{x \in \mathbb{R}^d \mid \exists x^* \in X^* \ : \ \|x - x^*\|_2 \leq e\}.$$

Свойства решений штрафной задачи

Свойства решений штрафной задачи

Пусть функция f и все функции h_i $(i = 1, \ldots m)$ являются непрерывными. Пусть X^* множество решений исходной условной задачи оптимизации и для $x^* \in X^*$ множество

$$U = \{x \in \mathbb{R}^d \mid f(x) \le f(x^*)\}$$

ограничено. Тогда для любого $e > 0$ существует $\rho(e) > 0$ такое, что множество решений штрафной задачи X^*_ρ для любых $\rho \ge \rho(e)$ содержится в

$$X^*_e = \{x \in \mathbb{R}^d \mid \exists x^* \in X^* \ : \ \|x - x^*\|_2 \le e\}.$$

Ограниченность U нужна для того, чтобы гарантировать, что вне ограничений функция f ведет себя «адекватно» и штрафная функция просто не улетит в $-\infty$. По факту это и гарантирует существование и непустоту X^*_ρ.

Доказательство

- От противного:

Доказательство

- От противного: пусть существует некоторое $e > 0$ и последовательность $\{\rho_i\} \to \infty$, что $X^*(\rho_i)$ не содержится в X_e^*, т.е. существуют $x_i^* \in X^*(\rho_i)$ не лежащие в X_e^*.

Доказательство

- От противного: пусть существует некоторое $e > 0$ и последовательность $\{\rho_i\} \to \infty$, что $X^*(\rho_i)$ не содержится в X^*_e, т.е. существуют $x^*_i \in X^*(\rho_i)$ не лежащие в X^*_e.
- Мы уже знаем, что $f(x^*) \geq f(x^*_\rho)$, а значит все x^*_i лежат в ограниченном множестве.

Доказательство

- От противного: пусть существует некоторое $e > 0$ и последовательность $\{\rho_i\} \to \infty$, что $X^*(\rho_i)$ не содержится в X_e^*, т.е. существуют $x_i^* \in X^*(\rho_i)$ не лежащие в X_e^*.
- Мы уже знаем, что $f(x^*) \geq f(x_\rho^*)$, а значит все x_i^* лежат в ограниченном множестве.
- По теореме Больцано-Вейерштрасса из ограниченной последовательности можно выделить сходящуюся подпоследовательность: $\tilde{x}_i^* \to \tilde{x}^*$. Посмотрим, что мы можем сказать про $\tilde{x}^*$.

Доказательство

- От противного: пусть существует некоторое $e > 0$ и последовательность $\{\rho_i\} \to \infty$, что $X^*(\rho_i)$ не содержится в X_e^*, т.е. существуют $x_i^* \in X^*(\rho_i)$ не лежащие в X_e^*.
- Мы уже знаем, что $f(x^*) \geq f(x_\rho^*)$, а значит все x_i^* лежат в ограниченном множестве.
- По теореме Больцано-Вейерштрасса из ограниченной последовательности можно выделить сходящуюся подпоследовательность: $\tilde{x}_i^* \to \tilde{x}^*$. Посмотрим, что мы можем сказать про $\tilde{x}^*$.
- Опять же по известному факту, что $f(x^*) \geq f(\tilde{x}_i^*)$, можно перейти к пределу и сделать вывод, что $f(x^*) \geq f(\tilde{x}^*)$. **Вопрос:** почему переход к пределу валиден?

Доказательство

- От противного: пусть существует некоторое $e > 0$ и последовательность $\{\rho_i\} \to \infty$, что $X^*(\rho_i)$ не содержится в X^*_e, т.е. существуют $x^*_i \in X^*(\rho_i)$ не лежащие в X^*_e.
- Мы уже знаем, что $f(x^*) \geq f(x^*_\rho)$, а значит все x^*_i лежат в ограниченном множестве.
- По теореме Больцано-Вейерштрасса из ограниченной последовательности можно выделить сходящуюся подпоследовательность: $\tilde{x}^*_i \to \tilde{x}^*$. Посмотрим, что мы можем сказать про $\tilde{x}^*$.
- Опять же по известному факту, что $f(x^*) \geq f(\tilde{x}^*_i)$, можно перейти к пределу и сделать вывод, что $f(x^*) \geq f(\tilde{x}^*)$. **Вопрос:** почему переход к пределу валиден? в силу непрерывности f.

Доказательство

- Уже получили, что $f(x^*) \geq f(\tilde{x}^*)$. Покажем, что дополнительно $\tilde{x}^*$ удовлетворяет исходным ограничениям h. От противного: пусть для какого-то $k = 1, \ldots m$, ограничение h_k не выполняется: $h_k(\tilde{x}^*) \neq 0$.

Доказательство

- Уже получили, что $f(x^*) \geq f(\tilde{x}^*)$. Покажем, что дополнительно $\tilde{x}^*$ удовлетворяет исходным ограничениям h. От противного: пусть для какого-то $k = 1, \dots m$, ограничение h_k не выполняется: $h_k(\tilde{x}^*) \neq 0$.
- В силу непрерывности h_k: можно заметить, что начиная с достаточно большого номера i, выполнено

$$|h_k(\tilde{x}_i^*)| \geq \frac{1}{2}|h_k(\tilde{x}^*)| > 0.$$

Доказательство

- Уже получили, что $f(x^*) \geq f(\tilde{x}^*)$. Покажем, что дополнительно $\tilde{x}^*$ удовлетворяет исходным ограничениям h. От противного: пусть для какого-то $k = 1, \ldots m$, ограничение h_k не выполняется: $h_k(\tilde{x}^*) \neq 0$.
- В силу непрерывности h_k: можно заметить, что начиная с достаточно большого номера i, выполнено

$$|h_k(\tilde{x}_i^*)| \geq \frac{1}{2}|h_k(\tilde{x}^*)| > 0.$$

- **Вопрос:** что в пределе $\tilde{\rho}_i \to +\infty$ с

$$f_{\tilde{\rho}_i}(\tilde{x}_i^*) = f(\tilde{x}_i^*) + \tilde{\rho}_i \cdot \frac{1}{2}\sum_{i=1}^{m} h_i^2(\tilde{x}_i^*)?$$

Доказательство

- Уже получили, что $f(x^*) \geq f(\tilde{x}^*)$. Покажем, что дополнительно $\tilde{x}^*$ удовлетворяет исходным ограничениям h. От противного: пусть для какого-то $k = 1, \ldots m$, ограничение h_k не выполняется: $h_k(\tilde{x}^*) \neq 0$.
- В силу непрерывности h_k: можно заметить, что начиная с достаточно большого номера i, выполнено

$$|h_k(\tilde{x}_i^*)| \geq \frac{1}{2}|h_k(\tilde{x}^*)| > 0.$$

- **Вопрос:** что в пределе $\tilde{\rho}_i \to +\infty$ с

$$f_{\tilde{\rho}_i}(\tilde{x}_i^*) = f(\tilde{x}_i^*) + \tilde{\rho}_i \cdot \frac{1}{2}\sum_{i=1}^{m} h_i^2(\tilde{x}_i^*)?$$

  Улетает в бесконечность. Пришли к противоречию, так как $f_{\tilde{\rho}_i}(\tilde{x}_i^*) \leq f(x^*)$, а значит $\tilde{x}^*$ удовлетворяет ограничениям.

Доказательство

- Получили, что $f(x^*) \geq f(\tilde{x}^*)$ и $\tilde{x}^*$ удовлетворяет ограничениям. **Вопрос:** что это значит?

Доказательство

- Получили, что $f(x^*) \geq f(\tilde{x}^*)$ и $\tilde{x}^*$ удовлетворяет ограничениям. **Вопрос:** что это значит? $\tilde{x}^* \in X^*$.

Доказательство

- Получили, что $f(x^*) \geq f(\tilde{x}^*)$ и $\tilde{x}^*$ удовлетворяет ограничениям. **Вопрос:** что это значит? $\tilde{x}^* \in X^*$.
- Но раз $\tilde{x}^* \in X^*$, то начиная с некоторого номера i элементы $\tilde{x}_i^*$ будут лежать в X_e^* – финальное противоречие, которое завершает доказательство.

Итог по классической штрафной функции

- Условная задача превращена в безусловную.
- Увеличение ρ приближает к исходной задаче.

Итог по классической штрафной функции

- Условная задача превращена в безусловную.
- Увеличение ρ приближает к исходной задаче.
- НО даже при большом ρ будет наблюдаться нарушение ограничений, что подходит не для всех задач.

Итог по классической штрафной функции

- Условная задача превращена в безусловную.
- Увеличение ρ приближает к исходной задаче.
- НО даже при большом ρ будет наблюдаться нарушение ограничений, что подходит не для всех задач.
- И увеличение ρ влечет за собой увеличение обусловленности задачи (как будет расти константа Липшица градиента?). А значит задачу будет сложнее решать.

Двойственный подъем

- Рассмотрим

$$\min_{x \in \mathbb{R}^d} \; f(x),$$
$$\text{s.t.} \; Ax = b,$$

где $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$.

Двойственный подъем

- Рассмотрим

$$\min_{x \in \mathbb{R}^d} \quad f(x),$$
$$\text{s.t.} \quad Ax = b,$$

  где $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$.
- Лагранжиан:

$$L(x, \lambda) = f(x) + \lambda^T (Ax - b).$$

Двойственный подъем

- Рассмотрим

$$\min_{x \in \mathbb{R}^d} \quad f(x),$$
$$\text{s.t.} \quad Ax = b,$$

где $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$.

- Лагранжиан:

$$L(x, \lambda) = f(x) + \lambda^T (Ax - b).$$

- Идея запустить градиентный подъем с шагом α для максимизации двойственной функции $g(\lambda) = \min_{x \in \mathbb{R}^d} L(x, \lambda)$:

$$\lambda^{k+1} = \lambda^k + \alpha \nabla \left(\min_{x \in \mathbb{R}^d} \left[f(x) + \lambda_k^T (Ax - b) \right] \right)$$

Двойственный подъем

- Двойственный подъем:

$$\lambda^{k+1} = \lambda^k + \alpha \nabla \left(\min_{x \in \mathbb{R}^d} \left[f(x) + \lambda_k^T (Ax - b) \right] \right)$$

Двойственный подъем

- Двойственный подъем:

$$\lambda^{k+1} = \lambda^k + \alpha \nabla \left(\min_{x \in \mathbb{R}^d} \left[f(x) + \lambda_k^T (Ax - b) \right] \right)$$

- Чуть-чуть по-другому:

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^d} \left[f(x) + \lambda_k^T (Ax - b) \right] = \arg\min_{x \in \mathbb{R}^d} L(x, \lambda^k)$$

$$\lambda^{k+1} = \lambda^k + \alpha \nabla \left(f(x^{k+1}) + \lambda_k^T (Ax^{k+1} - b) \right)$$

Двойственный подъем

- Двойственный подъем:

$$\lambda^{k+1} = \lambda^k + \alpha \nabla \left(\min_{x \in \mathbb{R}^d} \left[f(x) + \lambda_k^T (Ax - b) \right] \right)$$

- Чуть-чуть по-другому:

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^d} \left[f(x) + \lambda_k^T (Ax - b) \right] = \arg\min_{x \in \mathbb{R}^d} L(x, \lambda^k)$$

$$\lambda^{k+1} = \lambda^k + \alpha \nabla \left(f(x^{k+1}) + \lambda_k^T (Ax^{k+1} - b) \right)$$

или

$$\lambda^{k+1} = \lambda^k + \alpha (Ax^{k+1} - b)$$

Аугментация

- Уже знаем, что такая «добавка» не меняет задачу:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & f(x) + \frac{\rho}{2}\|Ax - b\|_2^2, \\ \text{s.t.} \quad & Ax = b, \end{aligned}$$

  Улучшают физику задачи за счет «регуляризации», в первую очередь трюк для практики.

- Лагранжиан:

$$L_\rho(x, \lambda) = f(x) + \lambda^T (Ax - b) + \frac{\rho}{2}\|Ax - b\|_2^2.$$

Аугментация

- Уже знаем, что такая «добавка» не меняет задачу:

$$\min_{x \in \mathbb{R}^d} \quad f(x) + \frac{\rho}{2}\|Ax - b\|_2^2,$$
$$\text{s.t.} \quad Ax = b,$$

Улучшают физику задачи за счет «регуляризации», в первую очередь трюк для практики.

- Лагранжиан:

$$L_\rho(x, \lambda) = f(x) + \lambda^T(Ax - b) + \frac{\rho}{2}\|Ax - b\|_2^2.$$

- Двойственный подъем:

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^d} L_\rho(x, \lambda^k), \quad \lambda^{k+1} = \lambda^k + \rho(Ax^{k+1} - b)$$

Шаг специально заменен на ρ, чтобы подбирать один параметр для метода.

ADMM

- Чуть более общая задача:

$$\begin{aligned} \min_{x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y}} \quad & f(x) + g(y), \\ \text{s.t.} \quad & Ax + By = c, \end{aligned} \tag{1}$$

где $A \in \mathbb{R}^{n \times d_x}$, $B \in \mathbb{R}^{n \times d_y}$, $c \in \mathbb{R}^n$.

ADMM

- Чуть более общая задача:

$$\begin{aligned} \min_{x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y}} \quad & f(x) + g(y), \\ \text{s.t.} \quad & Ax + By = c, \end{aligned} \tag{1}$$

  где $A \in \mathbb{R}^{n \times d_x}$, $B \in \mathbb{R}^{n \times d_y}$, $c \in \mathbb{R}^n$.
- Аугментация

$$\begin{aligned} \min_{x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y}} \quad & f(x) + g(y) + \frac{\rho}{2}\|Ax + By - c\|_2^2, \\ \text{s.t.} \quad & Ax + By = c, \end{aligned}$$

ADMM

- Чуть более общая задача:

$$\begin{aligned} \min_{x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y}} \quad & f(x) + g(y), \\ \text{s.t.} \quad & Ax + By = c, \end{aligned} \tag{1}$$

  где $A \in \mathbb{R}^{n \times d_x}$, $B \in \mathbb{R}^{n \times d_y}$, $c \in \mathbb{R}^n$.

- Аугментация

$$\begin{aligned} \min_{x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y}} \quad & f(x) + g(y) + \frac{\rho}{2}\|Ax + By - c\|_2^2, \\ \text{s.t.} \quad & Ax + By = c, \end{aligned}$$

- Лагранжиан:

$$L_\rho(x, y, \lambda) = f(x) + g(y) + \lambda^T(Ax + By - c) + \frac{\rho}{2}\|Ax + By - c\|_2^2$$

  Такой Лагранжиан порождает выпукло-вогнутую седловую задачу (более подробно мы обсудим седловые задачи через 2 лекции).

ADMM

- Двойственный подъем, он же Alternating Direction Method of Multipliers (ADMM):

Алгоритм 1 ADMM

Вход: стартовая точка $x^0 \in \mathbb{R}^{d_x}, y^0 \in \mathbb{R}^{d_y}, \lambda^0 \in \mathbb{R}^n$, количество итераций K

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

$x^{k+1} = \arg\min_{x \in \mathbb{R}^{d_x}} L_\rho(x, y^k, \lambda^k)$

$y^{k+1} = \arg\min_{y \in \mathbb{R}^{d_y}} L_\rho(x^{k+1}, y, \lambda^k)$

$\lambda^{k+1} = \lambda^k + \rho\left(Ax^{k+1} + By^{k+1} - c\right)$

end for

Выход: $\frac{1}{K}\sum_{k=1}^{K} x^k, \frac{1}{K}\sum_{k=1}^{K} y^k, \frac{1}{K}\sum_{k=1}^{K} \lambda^k$

ADMM

- Двойственный подъем, он же Alternating Direction Method of Multipliers (ADMM):

Алгоритм 2 ADMM

Вход: стартовая точка $x^0 \in \mathbb{R}^{d_x}, y^0 \in \mathbb{R}^{d_y}, \lambda^0 \in \mathbb{R}^n$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: $\quad x^{k+1} = \arg\min_{x \in \mathbb{R}^{d_x}} L_\rho(x, y^k, \lambda^k)$
3: $\quad y^{k+1} = \arg\min_{y \in \mathbb{R}^{d_y}} L_\rho(x^{k+1}, y, \lambda^k)$
4: $\quad \lambda^{k+1} = \lambda^k + \rho\left(Ax^{k+1} + By^{k+1} - c\right)$
5: **end for**

Выход: $\frac{1}{K}\sum_{k=1}^{K} x^k, \frac{1}{K}\sum_{k=1}^{K} y^k, \frac{1}{K}\sum_{k=1}^{K} \lambda^k$

- Alternating Direction – минимизация по x и y происходит не одновременно, а альтерированно: одна за другой.
- Multipliers – наличие двойственных множителей Лагранжа λ

ADMM

- С доказательством лучше ознакомиться после лекции про седловые задачи.
- В доказательстве будем использовать немного измененную версию:

Алгоритм 3 ADMM

Вход: стартовая точка $x^0 \in \mathbb{R}^{d_x}, y^0 \in \mathbb{R}^{d_y}, \lambda^0 \in \mathbb{R}^n$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: $\quad y^{k+1} = \arg\min_{y \in \mathbb{R}^{d_y}} L_\rho(x^k, y, \lambda^k)$
3: $\quad \lambda^{k+1} = \lambda^k + \rho\left(Ax^k + By^{k+1} - c\right)$
4: $\quad x^{k+1} = \arg\min_{x \in \mathbb{R}^{d_x}} L_\rho(x, y^{k+1}, \lambda^{k+1})$
5: **end for**

Выход: $\frac{1}{K}\sum_{k=1}^K x^k, \frac{1}{K}\sum_{k=1}^K y^k, \frac{1}{K}\sum_{k=1}^K \lambda^k$

- Вид Лагранжиана для удобства:

$$L_\rho(x, y, \lambda) = f(x) + g(y) + \lambda^T(Ax + By - c) + \frac{\rho}{2}\|Ax + By - c\|_2^2$$

Доказательство

- Запишем условие оптимальности для линии 2 алгоритма:

$$\nabla g(y^{k+1}) + B^T \lambda^k + \rho B^T (Ax^k + By^{k+1} - c) = 0$$

Доказательство

- Запишем условие оптимальности для линии 2 алгоритма:

$$\nabla g(y^{k+1}) + B^T\lambda^k + \rho B^T(Ax^k + By^{k+1} - c) = 0$$

- Линия 3 алгоритма:

$$\lambda^{k+1} - \lambda^k = \rho(Ax^k + By^{k+1} - c)$$

Доказательство

- Запишем условие оптимальности для линии 2 алгоритма:

$$\nabla g(y^{k+1}) + B^T\lambda^k + \rho B^T(Ax^k + By^{k+1} - c) = 0$$

- Линия 3 алгоритма:

$$\lambda^{k+1} - \lambda^k = \rho(Ax^k + By^{k+1} - c)$$

- Условие оптимальности для линии 5:

$$\nabla f(x^{k+1}) + A^T\lambda^{k+1} + \rho A^T(Ax^{k+1} + By^{k+1} - c) = 0$$

Доказательство

- Запишем условие оптимальности для линии 2 алгоритма:

$$\nabla g(y^{k+1}) + B^T \lambda^k + \rho B^T (Ax^k + By^{k+1} - c) = 0$$

Доказательство

- Запишем условие оптимальности для линии 2 алгоритма:

$$\nabla g(y^{k+1}) + B^T\lambda^k + \rho B^T(Ax^k + By^{k+1} - c) = 0$$

- Линия 3 алгоритма:

$$\lambda^{k+1} - \lambda^k = \rho(Ax^k + By^k - c)$$

Доказательство

- Запишем условие оптимальности для линии 2 алгоритма:

$$\nabla g(y^{k+1}) + B^T \lambda^k + \rho B^T (Ax^k + By^{k+1} - c) = 0$$

- Линия 3 алгоритма:

$$\lambda^{k+1} - \lambda^k = \rho(Ax^k + By^k - c)$$

- Условие оптимальности для линии 5:

$$\nabla f(x^{k+1}) + A^T \lambda^{k+1} + \rho A^T (Ax^{k+1} + By^{k+1} - c) = 0$$

Доказательство

- Простые алгебраические преобразования дают следующее:

$$\begin{pmatrix} \nabla f(x^{k+1}) + A^T\lambda^{k+1} \\ \nabla g(y^{k+1}) + B^T\lambda^{k+1} \\ -(Ax^{k+1} + By^{k+1} - c) \end{pmatrix} = -\begin{pmatrix} \rho A^T(Ax^{k+1} + By^{k+1} - c) \\ -B^T(\lambda^{k+1} - \lambda^k) + \rho B^T(Ax^k + By^{k+1} - c) \\ \frac{1}{\rho}(\lambda^{k+1} - \lambda^k) + A(x^{k+1} - x^k) \end{pmatrix}$$

- Используем, что $\lambda^{k+1} - \lambda^k = \rho(Ax^k + By^k - c)$:

$$\begin{pmatrix} \nabla f(x^{k+1}) + A^T\lambda^{k+1} \\ \nabla g(y^{k+1}) + B^T\lambda^{k+1} \\ -(Ax^{k+1} + By^{k+1} - c) \end{pmatrix} = -\begin{pmatrix} A^T(\lambda^{k+1} - \lambda^k) + \rho A^T A(x^{k+1} - x^k) \\ 0 \\ \frac{1}{\rho}(\lambda^{k+1} - \lambda^k) + A(x^{k+1} - x^k) \end{pmatrix}$$

Доказательство

- Заметим, что

$$\begin{pmatrix} \nabla f(x^{k+1}) + A^T \lambda^{k+1} \\ \nabla g(y^{k+1}) + B^T \lambda^{k+1} \\ -(Ax^{k+1} + By^{k+1} - c) \end{pmatrix} = \begin{pmatrix} \nabla_x L_0(x^{k+1}, y^{k+1}, \lambda^{k+1}) \\ \nabla_y L_0(x^{k+1}, y^{k+1}, \lambda^{k+1}) \\ -\nabla_\lambda L_0(x^{k+1}, y^{k+1}, \lambda^{k+1}) \end{pmatrix}$$

- Введем $P = \begin{pmatrix} \rho A^T A & 0 & -A^T \\ 0 & 0 & 0 \\ -A & 0 & \frac{1}{\rho} I \end{pmatrix}$ (сразу заметим, что она симметричная и положительно полуопределенная) и получим

$$\begin{pmatrix} \nabla_x L_0(x^{k+1}, y^{k+1}, \lambda^{k+1}) \\ \nabla_y L_0(x^{k+1}, y^{k+1}, \lambda^{k+1}) \\ -\nabla_\lambda L_0(x^{k+1}, y^{k+1}, \lambda^{k+1}) \end{pmatrix} = -P \begin{pmatrix} x^{k+1} - x^k \\ y^{k+1} - y^k \\ \lambda^{k+1} - \lambda^k \end{pmatrix}$$

Доказательство

- Тогда, вводя уже знакомое определение нормы $\|x\|_P = \langle x, Px\rangle$, имеем

$$\left\langle \begin{pmatrix} \nabla_x L_0(x^{k+1}, y^{k+1}, \lambda^{k+1}) \\ \nabla_y L_0(x^{k+1}, y^{k+1}, \lambda^{k+1}) \\ -\nabla_\lambda L_0(x^{k+1}, y^{k+1}, \lambda^{k+1}) \end{pmatrix}, \begin{pmatrix} x^{k+1} - x \\ y^{k+1} - y \\ \lambda^{k+1} - \lambda \end{pmatrix} \right\rangle$$

$$= -\left\langle P \begin{pmatrix} x^{k+1} - x^k \\ y^{k+1} - y^k \\ \lambda^{k+1} - \lambda^k \end{pmatrix}, \begin{pmatrix} x^{k+1} - x \\ y^{k+1} - y \\ \lambda^{k+1} - \lambda \end{pmatrix} \right\rangle$$

$$= \left\| \begin{pmatrix} x^k - x \\ y^k - y \\ \lambda^k - \lambda \end{pmatrix} \right\|_P - \left\| \begin{pmatrix} x^{k+1} - x \\ y^{k+1} - y \\ \lambda^{k+1} - \lambda \end{pmatrix} \right\|_P - \left\| \begin{pmatrix} x^{k+1} - x^k \\ y^{k+1} - y^k \\ \lambda^{k+1} - \lambda^k \end{pmatrix} \right\|_P$$

$$\leq \left\| \begin{pmatrix} x^k - x \\ y^k - y \\ \lambda^k - \lambda \end{pmatrix} \right\|_P - \left\| \begin{pmatrix} x^{k+1} - x \\ y^{k+1} - y \\ \lambda^{k+1} - \lambda \end{pmatrix} \right\|_P$$

Доказательство

- Суммируем по всем k от 0 до $K-1$ и усредняем:

$$\frac{1}{K}\sum_{k=0}^{K-1}\left\langle \begin{pmatrix} \nabla_x L_0(x^{k+1}, y^{k+1}, \lambda^{k+1}) \\ \nabla_y L_0(x^{k+1}, y^{k+1}, \lambda^{k+1}) \\ -\nabla_\lambda L_0(x^{k+1}, y^{k+1}, \lambda^{k+1}) \end{pmatrix}, \begin{pmatrix} x^{k+1} - x \\ y^{k+1} - y \\ \lambda^{k+1} - \lambda \end{pmatrix} \right\rangle$$

$$\leq \| \begin{pmatrix} x^0 - x \\ y^0 - y \\ \lambda^0 - \lambda \end{pmatrix} \|_P - \| \begin{pmatrix} x^K - x \\ y^K - y \\ \lambda^K - \lambda \end{pmatrix} \|_P$$

$$\leq \| \begin{pmatrix} x^0 - x \\ y^0 - y \\ \lambda^0 - \lambda \end{pmatrix} \|_P$$

Доказательство

- Суммируем по всем k от 0 до $K-1$ и усредняем:

$$
\frac{1}{K}\sum_{k=0}^{K-1}\left\langle \begin{pmatrix} \nabla_x L_0(x^{k+1}, y^{k+1}, \lambda^{k+1}) \\ \nabla_y L_0(x^{k+1}, y^{k+1}, \lambda^{k+1}) \\ -\nabla_\lambda L_0(x^{k+1}, y^{k+1}, \lambda^{k+1}) \end{pmatrix}, \begin{pmatrix} x^{k+1} - x \\ y^{k+1} - y \\ \lambda^{k+1} - \lambda \end{pmatrix} \right\rangle
$$
$$
\leq \left\| \begin{pmatrix} x^0 - x \\ y^0 - y \\ \lambda^0 - \lambda \end{pmatrix} \right\|_P - \left\| \begin{pmatrix} x^K - x \\ y^K - y \\ \lambda^K - \lambda \end{pmatrix} \right\|_P
$$
$$
\leq \left\| \begin{pmatrix} x^0 - x \\ y^0 - y \\ \lambda^0 - \lambda \end{pmatrix} \right\|_P
$$

- Дальше остается применить уже знакомые шаги: выпуклость L_0 по (x, y), вогнутость L_0 по λ, а также неравенство Йенсена. В итоге получим

Сходимость ADMM

Сходимость ADMM

Если в задаче (1) функции f и g являются выпуклыми и дружественными с точки зрения вычислений $\arg\min$, то ADMM имеет следующую оценку сходимости для любого $x \in \mathbb{R}^{d_x}$, $y \in \mathbb{R}^{d_y}$, $\lambda \in \mathbb{R}^n$

$$L_0\left(\frac{1}{K}\sum_{k=1}^{K} x^k, \frac{1}{K}\sum_{k=1}^{K} y^k, \lambda\right) - L_0\left(x, y, \frac{1}{K}\sum_{k=1}^{K} \lambda^k\right) \leq \frac{1}{2K}\|z^0 - z\|_P^2,$$

где L_0 – Лагранжиан без аугментации, $P = \begin{pmatrix} \rho A^T A & 0 & -A^T \\ 0 & 0 & 0 \\ -A & 0 & \frac{1}{\rho} I \end{pmatrix}$,

$$z^0 = \begin{pmatrix} x^0 \\ y^0 \\ \lambda^0 \end{pmatrix}$$

ADMM

- ADMM является одним из ключевых и популярных методов оптимизации.

ADMM

- ADMM является одним из ключевых и популярных методов оптимизации.
- Реализован во многих солверах и часто используется, как метод по умолчанию.

ADMM

- ADMM является одним из ключевых и популярных методов оптимизации.
- Реализован во многих солверах и часто используется, как метод по умолчанию.
- Нестандартная формулировка самой задачи, для которой придуман ADMM оказывается вбирает в себя много важных частных случаев. «Непривычная» переменная y часто играет роль вспомогательной переменной.

ADMM

- ADMM является одним из ключевых и популярных методов оптимизации.
- Реализован во многих солверах и часто используется, как метод по умолчанию.
- Нестандартная формулировка самой задачи, для которой придуман ADMM оказывается вбирает в себя много важных частных случаев. «Непривычная» переменная y часто играет роль вспомогательной переменной.
- Здесь штраф – дополнительная модификация для стабилизации и ускорения сходимости. При этом не требуется брать ρ обязательно очень большим.