Градиентные методы для задач с ограничениями. Метод проекции градиента. Метод Франк-Вульфа. Метод зеркального спуска

Даня Меркулов

МФТИ

Методы Оптимизации

Методы с ограничениями

Условная оптимизация

Безусловная оптимизация

$$
\min_{x \in \mathbb{R}^n} f(x)
$$

- Любая точка $x_0 \in \mathbb{R}^n$ допустима и может быть решением.

Условная оптимизация

Безусловная оптимизация

$$
\min_{x \in \mathbb{R}^n} f(x)
$$

- Любая точка $x_0 \in \mathbb{R}^n$ допустима и может быть решением.

Условная оптимизация

Безусловная оптимизация

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Любая точка $x_0 \in \mathbb{R}^n$ допустима и может быть решением.

Условная оптимизация

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

- Не все $x \in \mathbb{R}^n$ допустимы и могут быть решением.

Условная оптимизация

Безусловная оптимизация

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Любая точка $x_0 \in \mathbb{R}^n$ допустима и может быть решением.

Условная оптимизация

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

- Не все $x \in \mathbb{R}^n$ допустимы и могут быть решением.
- Решение должно лежать внутри множества S.

Условная оптимизация

Безусловная оптимизация

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Любая точка $x_0 \in \mathbb{R}^n$ допустима и может быть решением.

Условная оптимизация

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

- Не все $x \in \mathbb{R}^n$ допустимы и могут быть решением.
- Решение должно лежать внутри множества S.
- Пример:

$$\boxed{\frac{1}{2}\|Ax - b\|_2^2 \to \min_{\|x\|_2^2 \leq 1}}$$

Условная оптимизация

Безусловная оптимизация

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Любая точка $x_0 \in \mathbb{R}^n$ допустима и может быть решением.

Условная оптимизация

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

- Не все $x \in \mathbb{R}^n$ допустимы и могут быть решением.
- Решение должно лежать внутри множества S.
- Пример:

$$\frac{1}{2}\|Ax - b\|_2^2 \to \min_{\|x\|_2^2 \leq 1}$$

Условная оптимизация

Безусловная оптимизация

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Любая точка $x_0 \in \mathbb{R}^n$ допустима и может быть решением.

Условная оптимизация

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

- Не все $x \in \mathbb{R}^n$ допустимы и могут быть решением.
- Решение должно лежать внутри множества S.
- Пример:

$$\frac{1}{2}\|Ax - b\|_2^2 \to \min_{\|x\|_2^2 \leq 1}$$

Градиентный спуск — отличный способ решения безусловных задач

$$\boxed{x_{k+1} = x_k - \alpha_k \nabla f(x_k)} \tag{GD}$$

Можно ли адаптировать градиентный спуск для задачи с ограничениями?

Условная оптимизация

Безусловная оптимизация

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Любая точка $x_0 \in \mathbb{R}^n$ допустима и может быть решением.

Условная оптимизация

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

- Не все $x \in \mathbb{R}^n$ допустимы и могут быть решением.
- Решение должно лежать внутри множества S.
- Пример:

$$\frac{1}{2}\|Ax - b\|_2^2 \to \min_{\|x\|_2^2 \leq 1}$$

Градиентный спуск — отличный способ решения безусловных задач

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \tag{GD}$$

Можно ли адаптировать градиентный спуск для задачи с ограничениями?

Да. Для этого нужно использовать проекции, чтобы обеспечить допустимость на каждой итерации.

Пример: adversarial white-box attacks

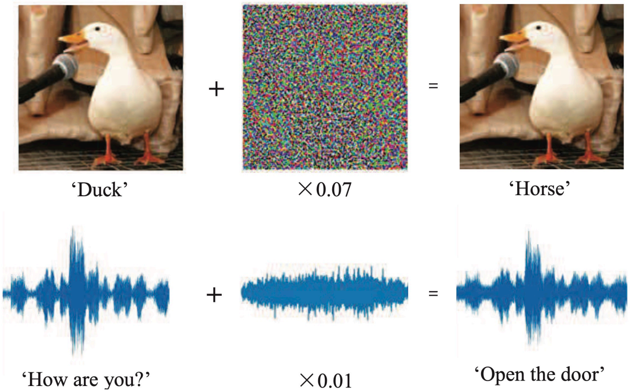


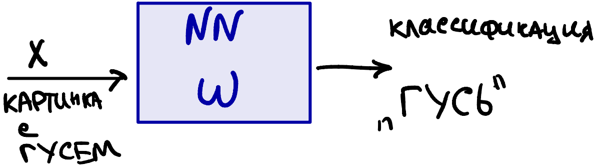
Рис. 1: Источник

- Математически нейронная сеть — это функция $f(w; x)$

$$\min_{\delta} \text{size}(\delta) \quad \text{s.t.} \quad \text{pred}[f(w; x+\delta)] \neq y$$

или

$$\boxed{\max_{\delta} l(w; x+\delta, y) \text{ s.t. size}(\delta) \leq \epsilon,} \; 0 \leq x+\delta \leq 1$$



Пример: adversarial white-box attacks

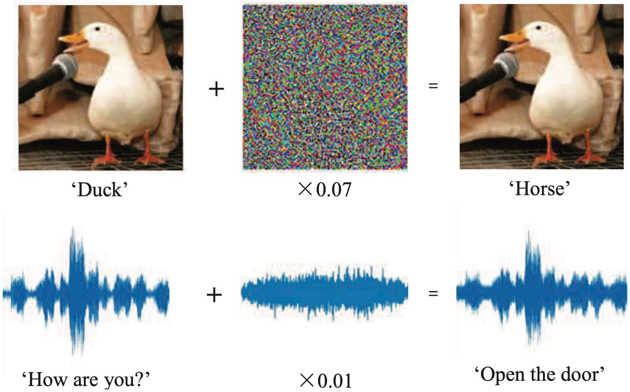


Рис. 1: Источник

- Математически нейронная сеть — это функция $f(w; x)$
- Обычно вход x задан, а веса сети w оптимизируются

$$
\min_{\delta} \text{size}(\delta) \quad \text{s.t.} \quad \text{pred}[f(w; x+\delta)] \neq y
$$

или

$$
\max_{\delta} l(w; x+\delta, y) \text{ s.t. } \text{size}(\delta) \leq \epsilon, \; 0 \leq x+\delta \leq 1
$$

Пример: adversarial white-box attacks

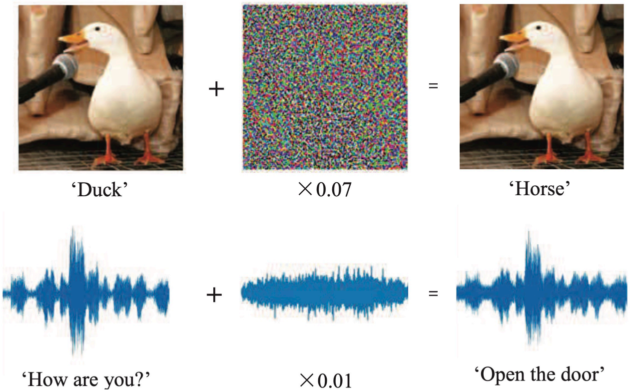


Рис. 1: Источник

- Математически нейронная сеть — это функция $f(w; x)$
- Обычно вход x задан, а веса сети w оптимизируются
- Но можно зафиксировать веса w и оптимизировать x!

$$\min_{\delta} \text{size}(\delta) \quad \text{s.t.} \quad \text{pred}[f(w; x+\delta)] \neq y$$

или

$$\max_{\delta} l(w; x+\delta, y) \text{ s.t. } \text{size}(\delta) \leq \epsilon, \; 0 \leq x+\delta \leq 1$$

Идея метода проекции градиента

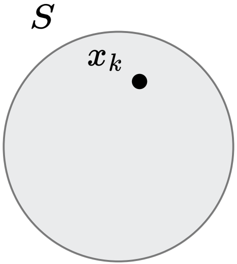


Рис. 2: Предположим, мы стартуем из точки x_k.

Идея метода проекции градиента

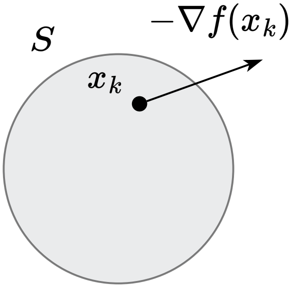


Рис. 3: И движемся в направлении $-\nabla f(x_k)$.

Идея метода проекции градиента

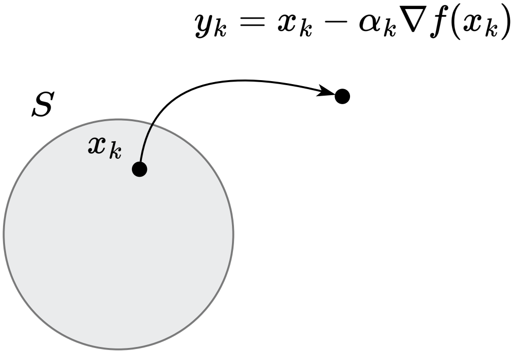


Рис. 4: Иногда мы можем оказаться за пределами допустимого множества.

Идея метода проекции градиента

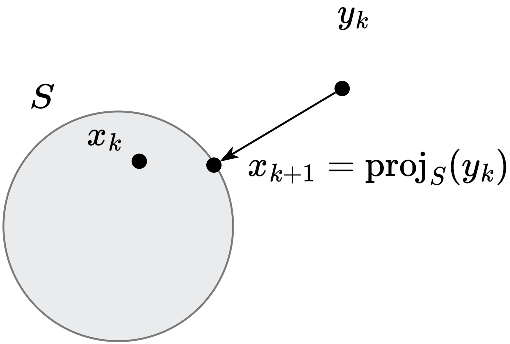


Рис. 5: Решим эту маленькую проблему с помощью проекции!

Идея метода проекции градиента

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

$\Pi_s(y_k)$

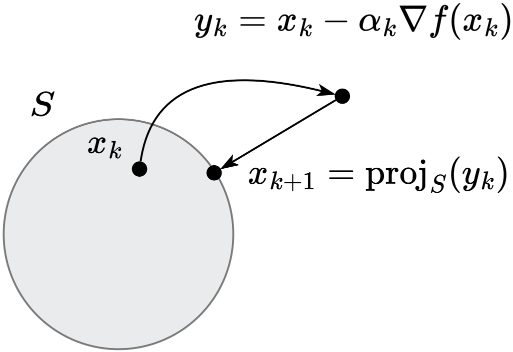


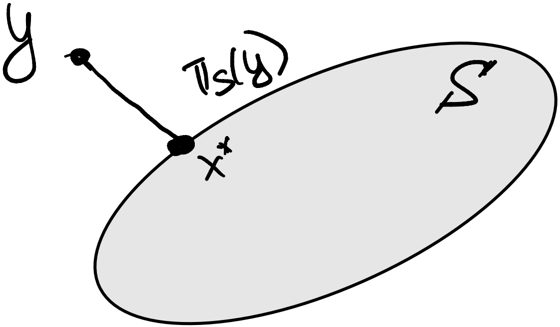
Рис. 6: Иллюстрация метода проекции градиента

Проекция

Проекция

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

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



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 сосредоточимся на евклидовой проекции (возможны и другие варианты) точки $\mathbf{y} \in \mathbb{R}^n$ на множество $S \subseteq \mathbb{R}^n$ — это точка $\text{proj}_S(\mathbf{y}) \in S$:

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

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 сосредоточимся на евклидовой проекции (возможны и другие варианты) точки $\mathbf{y} \in \mathbb{R}^n$ на множество $S \subseteq \mathbb{R}^n$ — это точка $\text{proj}_S(\mathbf{y}) \in S$:

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

- **Достаточные условия 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 сосредоточимся на евклидовой проекции (возможны и другие варианты) точки $\mathbf{y} \in \mathbb{R}^n$ на множество $S \subseteq \mathbb{R}^n$ — это точка $\text{proj}_S(\mathbf{y}) \in S$:

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

- **Достаточные условия существования проекции**. Если $S \subseteq \mathbb{R}^n$ — замкнутое множество, то проекция на множество S существует для любой точки.
- **Достаточные условия 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 сосредоточимся на евклидовой проекции (возможны и другие варианты) точки $\mathbf{y} \in \mathbb{R}^n$ на множество $S \subseteq \mathbb{R}^n$ — это точка $\text{proj}_S(\mathbf{y}) \in S$:

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

- **Достаточные условия существования проекции**. Если $S \subseteq \mathbb{R}^n$ — замкнутое множество, то проекция на множество S существует для любой точки.
- **Достаточные условия единственности проекции**. Если $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, то проекция на множество S единственна для любой точки.
- Если множество открытое, а точка лежит вне этого множества, то её проекция на него 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 сосредоточимся на евклидовой проекции (возможны и другие варианты) точки $\mathbf{y} \in \mathbb{R}^n$ на множество $S \subseteq \mathbb{R}^n$ — это точка $\text{proj}_S(\mathbf{y}) \in S$:

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

- **Достаточные условия существования проекции**. Если $S \subseteq \mathbb{R}^n$ — замкнутое множество, то проекция на множество S существует для любой точки.
- **Достаточные условия единственности проекции**. Если $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, то проекция на множество S единственна для любой точки.
- Если множество открытое, а точка лежит вне этого множества, то её проекция на него может не существовать.
- Если точка принадлежит множеству, то её проекция — это она сама.

Критерий проекции (неравенство Бурбаки-Чейни-Гольдштейна)

Theorem

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

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

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

1. $\text{proj}_S(y)$ - минимизатор дифференцируемой выпуклой функции $d(y, S, \|\cdot\|) = \|x - y\|^2$ на множестве S. Оптимальность:

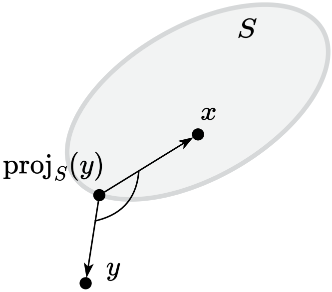


Рис. 7: Угол должен быть тупым или прямым для любой точки $x \in S$

Критерий проекции (неравенство Бурбаки-Чейни-Гольдштейна)

Theorem

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

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

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

1. $\text{proj}_S(y)$ - минимизатор дифференцируемой выпуклой функции $d(y, S, \|\cdot\|) = \|x - y\|^2$ на множестве S. Оптимальность:

$$\nabla d(\text{proj}_S(y))^T(x - \text{proj}_S(y)) \geq 0$$

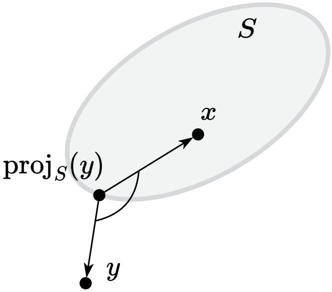


Рис. 7: Угол должен быть тупым или прямым для любой точки $x \in S$

Критерий проекции (неравенство Бурбаки-Чейни-Гольдштейна)

Theorem

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

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

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

1. $\text{proj}_S(y)$ - минимизатор дифференцируемой выпуклой функции $d(y, S, \|\cdot\|) = \|x - y\|^2$ на множестве S. Оптимальность:

$$\nabla d(\text{proj}_S(y))^T(x - \text{proj}_S(y)) \geq 0$$

$$2\left(\text{proj}_S(y) - y\right)^T (x - \text{proj}_S(y)) \geq 0$$

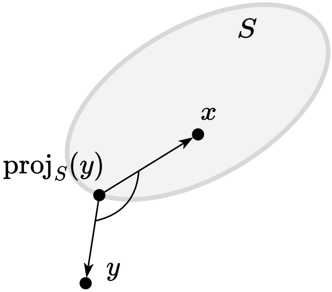


Рис. 7: Угол должен быть тупым или прямым для любой точки $x \in S$

Критерий проекции (неравенство Бурбаки-Чейни-Гольдштейна)

Theorem

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

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

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

1. $\text{proj}_S(y)$ - минимизатор дифференцируемой выпуклой функции $d(y, S, \|\cdot\|) = \|x - y\|^2$ на множестве S. Оптимальность:

$$\nabla d(\text{proj}_S(y))^T(x - \text{proj}_S(y)) \geq 0$$

$$2\left(\text{proj}_S(y) - y\right)^T(x - \text{proj}_S(y)) \geq 0$$

$$\left(y - \text{proj}_S(y)\right)^T(x - \text{proj}_S(y)) \leq 0$$

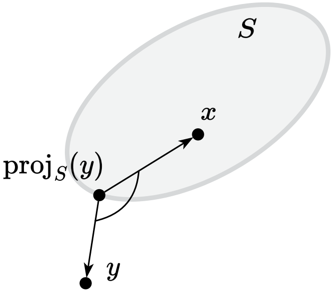


Рис. 7: Угол должен быть тупым или прямым для любой точки $x \in S$

Критерий проекции (неравенство Бурбаки-Чейни-Гольдштейна)

Theorem

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

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

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

1. $\text{proj}_S(y)$ - минимизатор дифференцируемой выпуклой функции $d(y, S, \|\cdot\|) = \|x - y\|^2$ на множестве S. Оптимальность:

$$\nabla d(\text{proj}_S(y))^T (x - \text{proj}_S(y)) \geq 0$$

$$2 \left(\text{proj}_S(y) - y\right)^T (x - \text{proj}_S(y)) \geq 0$$

$$\left(y - \text{proj}_S(y)\right)^T (x - \text{proj}_S(y)) \leq 0$$

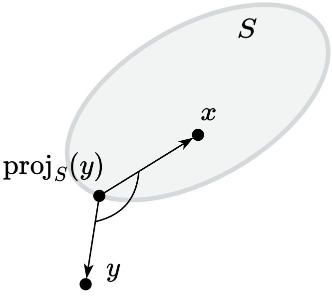


Рис. 7: Угол должен быть тупым или прямым для любой точки $x \in S$

2. Используем теорему косинусов $\boxed{2x^T y = \|x\|^2 + \|y\|^2 - \|x - y\|^2}$ с $\underline{x = x - \text{proj}_S(y)}$ и $\underline{y = y - \text{proj}_S(y)}$.

$\langle x, y \rangle = \|x\| \cdot \|y\| \cdot \cos\alpha$

Критерий проекции (неравенство Бурбаки-Чейни-Гольдштейна)

Theorem

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

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

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

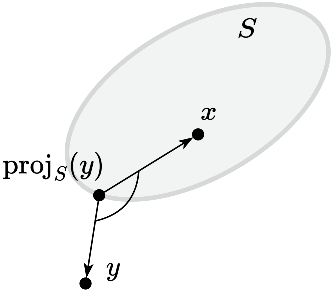


Рис. 7: Угол должен быть тупым или прямым для любой точки $x \in S$

1. $\text{proj}_S(y)$ - минимизатор дифференцируемой выпуклой функции $d(y, S, \|\cdot\|) = \|x - y\|^2$ на множестве S. Оптимальность:

$$\nabla d(\text{proj}_S(y))^T(x - \text{proj}_S(y)) \geq 0$$

$$2\left(\text{proj}_S(y) - y\right)^T(x - \text{proj}_S(y)) \geq 0$$

$$\left(y - \text{proj}_S(y)\right)^T(x - \text{proj}_S(y)) \leq 0$$

2. Используем теорему косинусов $2x^Ty = \|x\|^2 + \|y\|^2 - \|x - y\|^2$ с $x = x - \text{proj}_S(y)$ и $y = y - \text{proj}_S(y)$.

$$\underbrace{0 \geq 2x^Ty}_{\text{по п. 1}} = \|x - \text{proj}_S(y)\|^2 + \|y + \text{proj}_S(y)\|^2 - \|x - y\|^2$$

Критерий проекции (неравенство Бурбаки-Чейни-Гольдштейна)

Theorem

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

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

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

1. $\text{proj}_S(y)$ - минимизатор дифференцируемой выпуклой функции $d(y, S, \|\cdot\|) = \|x - y\|^2$ на множестве S. Оптимальность:

$$\nabla d(\text{proj}_S(y))^T(x - \text{proj}_S(y)) \geq 0$$

$$2\left(\text{proj}_S(y) - y\right)^T(x - \text{proj}_S(y)) \geq 0$$

$$\left(y - \text{proj}_S(y)\right)^T(x - \text{proj}_S(y)) \leq 0$$

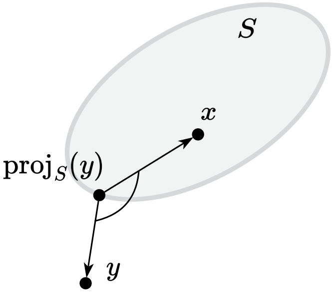


Рис. 7: Угол должен быть тупым или прямым для любой точки $x \in S$

2. Используем теорему косинусов $2x^Ty = \|x\|^2 + \|y\|^2 - \|x - y\|^2$ с $x = x - \text{proj}_S(y)$ и $y = y - \text{proj}_S(y)$.

$$0 \geq 2x^Ty = \|x - \text{proj}_S(y)\|^2 + \|y + \text{proj}_S(y)\|^2 - \|x - y\|^2$$

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

Оператор проекции - нерастягивающий

$L < 1$

сжимающим

- Функция f называется нерастягивающей, если она является L-липшицевой с константой $L \leq 1$ [1]. То есть для любых двух точек $x, y \in \operatorname{dom} f$

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

Это означает, что расстояние между образами точек не превосходит расстояния между самими точками.

[1] Нерастягивающая функция становится сжимающей при $L < 1$.

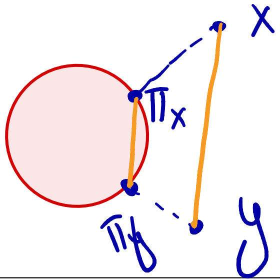
Оператор проекции - нерастягивающий

- Функция f называется нерастягивающей, если она является L-липшицевой с константой $L \leq 1$ [1]. То есть для любых двух точек $x, y \in \text{dom} f$

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

Это означает, что расстояние между образами точек не превосходит расстояния между самими точками.

- Оператор проекции является нерастягивающим:



$$\|\text{proj}(x) - \text{proj}(y)\|_2 \leq \|x - y\|_2.$$

[1] Нерастягивающая функция становится сжимающей при $L < 1$.

Оператор проекции - нерастягивающий

- Функция f называется нерастягивающей, если она является L-липшицевой с константой $L \leq 1$ [1]. То есть для любых двух точек $x, y \in \text{dom} f$

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

Это означает, что расстояние между образами точек не превосходит расстояния между самими точками.

- Оператор проекции является нерастягивающим:

$$\|\text{proj}(x) - \text{proj}(y)\|_2 \leq \|x - y\|_2.$$

- Далее покажем, что из неравенства Бурбаки-Чейни-Гольдштейна следует свойство нерастягиваемости, а именно:

$$\boxed{\langle y - \text{proj}(y), x - \text{proj}(y)\rangle \leq 0 \quad \forall x \in S} \quad \Rightarrow \quad \|\text{proj}(x) - \text{proj}(y)\|_2 \leq \|x - y\|_2.$$

X = PROJ(X)

[1] Нерастягивающая функция становится сжимающей при $L < 1$.

Оператор проекции - нерастягивающий

Сокращённая запись: пусть $\pi = \text{proj}$ и $\pi(x)$ обозначает $\text{proj}(x)$.

Оператор проекции - нерастягивающий

Сокращённая запись: пусть $\pi = \text{proj}$ и $\pi(x)$ обозначает $\text{proj}(x)$.

Начнём с неравенства Бурбаки-Чейни-Гольдштейна БЧГ

$$\langle y - \pi(y), x - \pi(y) \rangle \leq 0 \quad \forall x \in S. \tag{3}$$

Оператор проекции - нерастягивающий

Сокращённая запись: пусть $\pi = \text{proj}$ и $\pi(x)$ обозначает $\text{proj}(x)$.

Начнём с неравенства Бурбаки-Чейни-Гольдштейна

$$\langle y - \pi(y), x - \pi(y) \rangle \leq 0 \quad \forall x \in S. \tag{3}$$

Заменим x на $\pi(x)$ в (3) $\quad x = \pi(x)$

$$\langle y - \pi(y), \pi(x) - \pi(y) \rangle \leq 0. \tag{4}$$

Оператор проекции - нерастягивающий

Сокращённая запись: пусть $\pi = \text{proj}$ и $\pi(x)$ обозначает $\text{proj}(x)$.

Начнём с неравенства Бурбаки-Чейни-Гольдштейна

$$\langle y - \pi(y), x - \pi(y)\rangle \leq 0 \quad \forall x \in S. \tag{3}$$

Заменим x на $\pi(x)$ в (3)

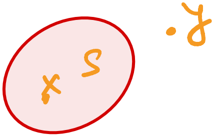
$$\langle y - \pi(y), \pi(x) - \pi(y)\rangle \leq 0. \tag{4}$$

Заменим y на x и x на $\pi(y)$ в (3)

$y := x$ $\quad x := \pi(y)$

$$\langle x - \pi(x), \pi(y) - \pi(x)\rangle \leq 0. \tag{5}$$

Оператор проекции - нерастягивающий



Сокращённая запись: пусть $\pi = \text{proj}$ и $\pi(x)$ обозначает $\text{proj}(x)$.

Начнём с неравенства Бурбаки-Чейни-Гольдштейна

$$\langle y - \pi(y), x - \pi(y) \rangle \leq 0 \quad \forall x \in S. \tag{3}$$

Заменим x на $\pi(x)$ в (3)

$$\langle y - \pi(y), \pi(x) - \pi(y) \rangle \leq 0. \tag{4}$$

Заменим y на x и x на $\pi(y)$ в (3)

$$\langle x - \pi(x), \pi(y) - \pi(x) \rangle \leq 0. \tag{5}$$

(4)+(5) сократят $\pi(y) - \pi(x)$, что нежелательно. Поэтому сменим знак в (5):

$$\langle \pi(x) - x, \pi(x) - \pi(y) \rangle \leq 0. \tag{6}$$

Оператор проекции - нерастягивающий

Сокращённая запись: пусть $\pi = \text{proj}$ и $\pi(x)$ обозначает $\text{proj}(x)$.

Начнём с неравенства Бурбаки-Чейни-Гольдштейна

$$\langle y - \pi(y), x - \pi(y)\rangle \leq 0 \quad \forall x \in S. \tag{3}$$

Заменим x на $\pi(x)$ в (3)

$$\langle y - \pi(y), \pi(x) - \pi(y)\rangle \leq 0. \tag{4}$$

Заменим y на x и x на $\pi(y)$ в (3)

$\langle \pi(x) - x, \pi(x) - \pi(y)\rangle \leq 0$

$$\langle x - \pi(x), \pi(y) - \pi(x)\rangle \leq 0. \tag{5}$$

(4)+(5) сократят $\pi(y) - \pi(x)$, что нежелательно. Поэтому сменим знак в (5):

$y - \pi(y) + \pi(x) - x$

$$\langle \pi(x) - x, \pi(x) - \pi(y)\rangle \leq 0. \tag{6}$$

$$\langle y - \pi(y) + \pi(x) - x, \pi(x) - \pi(y)\rangle \leq 0$$

$$\langle y - x, \pi(x) - \pi(y)\rangle \leq -\langle \pi(x) - \pi(y), \pi(x) - \pi(y)\rangle$$

$$\langle y - x, \pi(y) - \pi(x)\rangle \geq \|\pi(x) - \pi(y)\|_2^2$$

$$\boxed{\|(y - x)^\top(\pi(y) - \pi(x))\|_2 \geq \|\pi(x) - \pi(y)\|_2^2}$$

Оператор проекции - нерастягивающий

$\|\pi(x) - \pi(y)\| \neq 0$

Сокращённая запись: пусть $\pi = \text{proj}$ и $\pi(x)$ обозначает $\text{proj}(x)$.

Начнём с неравенства Бурбаки-Чейни-Гольдштейна

$$\langle y - \pi(y), x - \pi(y)\rangle \leq 0 \quad \forall x \in S. \tag{3}$$

Заменим x на $\pi(x)$ в (3)

$$\langle y - \pi(y), \pi(x) - \pi(y)\rangle \leq 0. \tag{4}$$

Заменим y на x и x на $\pi(y)$ в (3)

$$\langle x - \pi(x), \pi(y) - \pi(x)\rangle \leq 0. \tag{5}$$

(4)+(5) сократят $\pi(y) - \pi(x)$, что нежелательно. Поэтому сменим знак в (5):

$$\langle \pi(x) - x, \pi(x) - \pi(y)\rangle \leq 0. \tag{6}$$

$$\langle y - \pi(y) + \pi(x) - x, \pi(x) - \pi(y)\rangle \leq 0$$

$$\langle y - x, \pi(x) - \pi(y)\rangle \leq -\langle \pi(x) - \pi(y), \pi(x) - \pi(y)\rangle$$

$\|y-x\|\|\pi(y)-\pi(x)\|$

$$\langle y - x, \pi(y) - \pi(x)\rangle \geq \|\pi(x) - \pi(y)\|_2^2$$

$\geq$

$$\|(y - x)^\top(\pi(y) - \pi(x))\|_2 \geq \|\pi(x) - \pi(y)\|_2^2$$

$\|y-x\| \cdot \|\pi(y) - \pi(x)\|$

По неравенству КБШ левая часть ограничена сверху величиной $\|y - x\|_2\|\pi(y) - \pi(x)\|_2$, откуда следует $\|y - x\|_2\|\pi(y) - \pi(x)\|_2 \geq \|\pi(x) - \pi(y)\|_2^2$. Сокращая на $\|\pi(x) - \pi(y)\|_2$, завершаем доказательство.

Пример: проекция на шар

Найдём $\pi_S(y) = \pi$, если $S = \{x \in \mathbb{R}^n \mid \|x - x_0\| \le R\}$, $y \notin S$

Пример: проекция на шар

Найдём $\pi_S(y) = \pi$, если $S = \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq R\}$, $y \notin S$

Построим гипотезу по рисунку: $\pi = x_0 + R \cdot \frac{y - x_0}{\|y - x_0\|}$

Пример: проекция на шар

Найдём $\pi_S(y) = \pi$, если $S = \{x \in \mathbb{R}^n \mid \|x - x_0\| \le R\}$, $y \notin S$

Построим гипотезу по рисунку: $\pi = x_0 + R \cdot \frac{y - x_0}{\|y - x_0\|}$

Проверим неравенство для замкнутого выпуклого множества:
$(\pi - y)^T(x - \pi) \ge 0$

Пример: проекция на шар

Найдём $\pi_S(y) = \pi$, если $S = \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq R\}$, $y \notin S$

Построим гипотезу по рисунку: $\pi = x_0 + R \cdot \frac{y - x_0}{\|y - x_0\|}$

Проверим неравенство для замкнутого выпуклого множества: $(\pi - y)^T(x - \pi) \geq 0$

$$
\begin{aligned}
\left(x_0 - y + R\frac{y - x_0}{\|y - x_0\|} \right)^T \left(x - x_0 - R\frac{y - x_0}{\|y - x_0\|} \right) &= \\
\left(\frac{(y - x_0)(R - \|y - x_0\|)}{\|y - x_0\|} \right)^T \left(\frac{(x - x_0)\|y - x_0\| - R(y - x_0)}{\|y - x_0\|} \right) &= \\
\frac{R - \|y - x_0\|}{\|y - x_0\|^2} \left(y - x_0\right)^T \left(\left(x - x_0\right) \|y - x_0\| - R \left(y - x_0\right) \right) &= \\
\frac{R - \|y - x_0\|}{\|y - x_0\|} \left(\left(y - x_0\right)^T \left(x - x_0\right) - R\|y - x_0\| \right) &= \\
\left(R - \|y - x_0\|\right) \left(\frac{(y - x_0)^T(x - x_0)}{\|y - x_0\|} - R \right)
\end{aligned}
$$

Пример: проекция на шар

Найдём $\pi_S(y) = \pi$, если $S = \{x \in \mathbb{R}^n \mid \|x - x_0\| \le R\}$, $y \notin S$

Построим гипотезу по рисунку: $\pi = x_0 + R \cdot \frac{y - x_0}{\|y - x_0\|}$

Проверим неравенство для замкнутого выпуклого множества: $(\pi - y)^T(x - \pi) \ge 0$

$$
\left(x_0 - y + R\frac{y - x_0}{\|y - x_0\|}\right)^T \left(x - x_0 - R\frac{y - x_0}{\|y - x_0\|}\right) =
$$

$$
\left(\frac{(y - x_0)(R - \|y - x_0\|)}{\|y - x_0\|}\right)^T \left(\frac{(x - x_0)\|y - x_0\| - R(y - x_0)}{\|y - x_0\|}\right) =
$$

$$
\frac{R - \|y - x_0\|}{\|y - x_0\|^2}(y - x_0)^T\left((x - x_0)\|y - x_0\| - R(y - x_0)\right) =
$$

$$
\frac{R - \|y - x_0\|}{\|y - x_0\|}\left((y - x_0)^T(x - x_0) - R\|y - x_0\|\right) =
$$

$$
(R - \|y - x_0\|)\left(\frac{(y - x_0)^T(x - x_0)}{\|y - x_0\|} - R\right)
$$

Первый множитель отрицателен по выбору точки y. Второй множитель также отрицателен, что следует из КБШ:

Пример: проекция на шар

Найдём $\pi_S(y) = \pi$, если $\boxed{S = \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq R\},\ y \notin S}$

Построим гипотезу по рисунку: $\boxed{\pi = x_0 + R \cdot \frac{y - x_0}{\|y - x_0\|}}$

Проверим неравенство для замкнутого выпуклого множества: $(\pi - y)^T(x - \pi) \geq 0$

$$
\left(x_0 - y + R\frac{y - x_0}{\|y - x_0\|}\right)^T\left(x - x_0 - R\frac{y - x_0}{\|y - x_0\|}\right) =
$$

$$
\left(\frac{(y - x_0)(R - \|y - x_0\|)}{\|y - x_0\|}\right)^T\left(\frac{(x - x_0)\|y - x_0\| - R(y - x_0)}{\|y - x_0\|}\right) =
$$

$$
\frac{R - \|y - x_0\|}{\|y - x_0\|^2}(y - x_0)^T\left((x - x_0)\|y - x_0\| - R(y - x_0)\right) =
$$

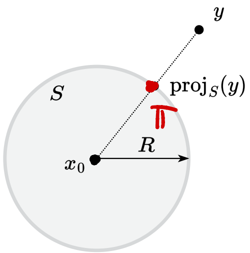
$$
\frac{R - \|y - x_0\|}{\|y - x_0\|}\left((y - x_0)^T(x - x_0) - R\|y - x_0\|\right) =
$$

$$
(R - \|y - x_0\|)\left(\frac{(y - x_0)^T(x - x_0)}{\|y - x_0\|} - R\right)
$$

Первый множитель отрицателен по выбору точки y. Второй множитель также отрицателен, что следует из КБШ:

$$
(y - x_0)^T(x - x_0) \leq \|y - x_0\|\|x - x_0\|
$$

$$
\frac{(y - x_0)^T(x - x_0)}{\|y - x_0\|} - R \leq \frac{\|y - x_0\|\|x - x_0\|}{\|y - x_0\|} -
$$



Пример: проекция на полупространство

Найдём $\pi_S(y) = \pi$, если $S = \{x \in \mathbb{R}^n \mid c^T x = b\}$, $y \notin S$. Построим гипотезу по рисунку: $\pi = y + \alpha c$. Коэффициент α выбирается так, чтобы $\pi \in S$: $c^T \pi = b$, тогда:

Пример: проекция на полупространство

$c^Tx \leq b$

Найдём $\pi_S(y) = \pi$, если $S = \{x \in \mathbb{R}^n \mid c^T x = b\}$, $y \notin S$. Построим гипотезу по рисунку: $\pi = y + \alpha c$. Коэффициент α выбирается так, чтобы $\pi \in S$: $c^T \pi = b$, тогда:

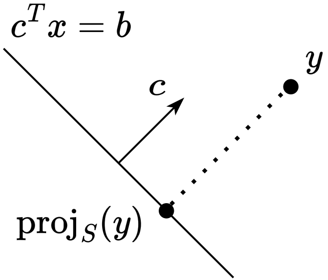


Рис. 9: Гиперплоскость

Пример: проекция на полупространство

Найдём $\pi_S(y) = \pi$, если $S = \{x \in \mathbb{R}^n \mid c^T x = b\}$, $y \notin S$. Построим гипотезу по рисунку: $\pi = y + \alpha c$. Коэффициент α выбирается так, чтобы $\pi \in S$: $c^T \pi = b$, тогда:

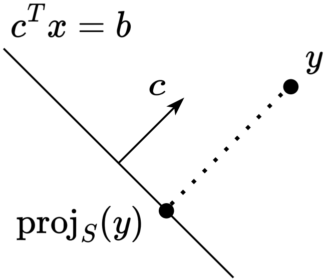


Рис. 9: Гиперплоскость

$$
\begin{aligned}
c^T(y + \alpha c) &= b \\
c^T y + \alpha c^T c &= b \\
c^T y &= b - \alpha c^T c
\end{aligned}
$$

Проверим неравенство для замкнутого выпуклого множества: $(\pi - y)^T(x - \pi) \geq 0$

$$
\begin{aligned}
(y + \alpha c - y)^T(x - y - \alpha c) &= \\
\alpha c^T(x - y - \alpha c) &= \\
\alpha(c^T x) - \alpha(c^T y) - \alpha^2(c^T c) &= \\
\alpha b - \alpha(b - \alpha c^T c) - \alpha^2 c^T c &= \\
\alpha b - \alpha b + \alpha^2 c^T c - \alpha^2 c^T c &= 0 \geq 0
\end{aligned}
$$

Метод проекции градиента (PGD)

Идея

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

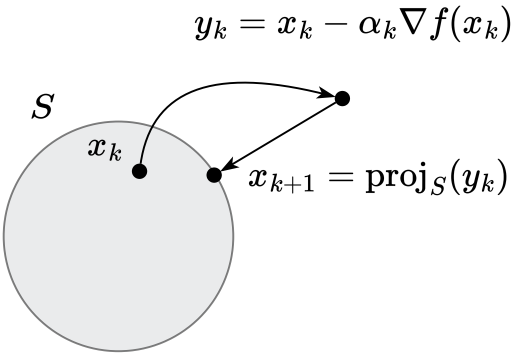


Рис. 10: Иллюстрация алгоритма метода проекции градиента

Инструменты для доказательства сходимости

Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — выпуклая функция с L-липшицевым градиентом. Тогда для любых $x, y \in \mathbb{R}^n$ выполняется следующее неравенство:

$$
f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \le f(y) \text{ или, эквивалентно,}
$$

$$
\|\nabla f(y) - \nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|_2^2 \le 2L\left(f(x) - f(y) - \langle \nabla f(y), x - y \rangle\right)
$$

1. Для доказательства рассмотрим вспомогательную функцию $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$. Она является выпуклой (как сумма выпуклых функций). Легко проверить, что она является L-гладкой по определению, поскольку $\nabla \varphi(y) = \nabla f(y) - \nabla f(x)$ и $\|\nabla \varphi(y_1) - \nabla \varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \le L\|y_1 - y_2\|$.

Инструменты для доказательства сходимости

Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — выпуклая функция с L-липшицевым градиентом. Тогда для любых $x, y \in \mathbb{R}^n$ выполняется следующее неравенство:

$$
f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) \text{ или, эквивалентно,}
$$
$$
\|\nabla f(y) - \nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L \left(f(x) - f(y) - \langle \nabla f(y), x - y \rangle \right)
$$

1. Для доказательства рассмотрим вспомогательную функцию $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$. Она является выпуклой (как сумма выпуклых функций). Легко проверить, что она является L-гладкой по определению, поскольку $\nabla \varphi(y) = \nabla f(y) - \nabla f(x)$ и $\|\nabla \varphi(y_1) - \nabla \varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \leq L\|y_1 - y_2\|$.
2. Теперь рассмотрим свойство липшицевой параболы для гладкой функции $\varphi(y)$:

Инструменты для доказательства сходимости

Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — выпуклая функция с L-липшицевым градиентом. Тогда для любых $x, y \in \mathbb{R}^n$ выполняется следующее неравенство:

$$
\begin{aligned}
f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) \text{ или, эквивалентно,} \\
\|\nabla f(y) - \nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L \left(f(x) - f(y) - \langle \nabla f(y), x - y \rangle \right)
\end{aligned}
$$

1. Для доказательства рассмотрим вспомогательную функцию $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$. Она является выпуклой (как сумма выпуклых функций). Легко проверить, что она является L-гладкой по определению, поскольку $\nabla \varphi(y) = \nabla f(y) - \nabla f(x)$ и $\|\nabla \varphi(y_1) - \nabla \varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \leq L\|y_1 - y_2\|$.
2. Теперь рассмотрим свойство липшицевой параболы для гладкой функции $\varphi(y)$:

$$
\varphi(y) \leq \varphi(x) + \langle \nabla \varphi(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2
$$

Инструменты для доказательства сходимости

Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — выпуклая функция с L-липшицевым градиентом. Тогда для любых $x, y \in \mathbb{R}^n$ выполняется следующее неравенство:

$$
f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) \text{ или, эквивалентно,}
$$

$$
\|\nabla f(y) - \nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L \left(f(x) - f(y) - \langle \nabla f(y), x - y \rangle\right)
$$

1. Для доказательства рассмотрим вспомогательную функцию $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$. Она является выпуклой (как сумма выпуклых функций). Легко проверить, что она является L-гладкой по определению, поскольку $\nabla \varphi(y) = \nabla f(y) - \nabla f(x)$ и $\|\nabla \varphi(y_1) - \nabla \varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \leq L\|y_1 - y_2\|$.
2. Теперь рассмотрим свойство липшицевой параболы для гладкой функции $\varphi(y)$:

$$
\varphi(y) \leq \varphi(x) + \langle \nabla \varphi(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2
$$

$$
\overset{x := y,\, y := y - \frac{1}{L}\nabla \varphi(y)}{} \quad \varphi\left(y - \frac{1}{L}\nabla \varphi(y)\right) \leq \varphi(y) + \left\langle \nabla \varphi(y), -\frac{1}{L}\nabla \varphi(y)\right\rangle + \frac{1}{2L}\|\nabla \varphi(y)\|_2^2
$$

Инструменты для доказательства сходимости

Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — выпуклая функция с L-липшицевым градиентом. Тогда для любых $x, y \in \mathbb{R}^n$ выполняется следующее неравенство:

$$
\begin{aligned}
f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) \text{ или, эквивалентно,} \\
\|\nabla f(y) - \nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L\left(f(x) - f(y) - \langle \nabla f(y), x - y \rangle\right)
\end{aligned}
$$

1. Для доказательства рассмотрим вспомогательную функцию $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$. Она является выпуклой (как сумма выпуклых функций). Легко проверить, что она является L-гладкой по определению, поскольку $\nabla \varphi(y) = \nabla f(y) - \nabla f(x)$ и $\|\nabla \varphi(y_1) - \nabla \varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \leq L\|y_1 - y_2\|$.
2. Теперь рассмотрим свойство липшицевой параболы для гладкой функции $\varphi(y)$:

$$
\begin{aligned}
\varphi(y) &\leq \varphi(x) + \langle \nabla \varphi(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2 \\
\overset{x := y,\, y := y - \frac{1}{L}\nabla \varphi(y)}{\quad} \varphi\left(y - \frac{1}{L}\nabla \varphi(y)\right) &\leq \varphi(y) + \left\langle \nabla \varphi(y), -\frac{1}{L}\nabla \varphi(y) \right\rangle + \frac{1}{2L}\|\nabla \varphi(y)\|_2^2 \\
\varphi\left(y - \frac{1}{L}\nabla \varphi(y)\right) &\leq \varphi(y) - \frac{1}{2L}\|\nabla \varphi(y)\|_2^2
\end{aligned}
$$

Инструменты для доказательства сходимости

3. Из условий оптимальности первого порядка для выпуклой функции $\nabla \varphi(y) = \nabla f(y) - \nabla f(x) = 0$. Можно заключить, что для любого x минимум функции $\varphi(y)$ достигается в точке $y = x$. Следовательно:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

Инструменты для доказательства сходимости

3. Из условий оптимальности первого порядка для выпуклой функции $\nabla \varphi(y) = \nabla f(y) - \nabla f(x) = 0$. Можно заключить, что для любого x минимум функции $\varphi(y)$ достигается в точке $y = x$. Следовательно:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla \varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla \varphi(y)\|_2^2$$

4. Теперь подставим $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$:

Инструменты для доказательства сходимости

3. Из условий оптимальности первого порядка для выпуклой функции $\nabla \varphi(y) = \nabla f(y) - \nabla f(x) = 0$. Можно заключить, что для любого x минимум функции $\varphi(y)$ достигается в точке $y = x$. Следовательно:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Теперь подставим $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$:

$$f(x) - \langle \nabla f(x), x \rangle \leq f(y) - \langle \nabla f(x), y \rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

Инструменты для доказательства сходимости

3. Из условий оптимальности первого порядка для выпуклой функции $\nabla \varphi(y) = \nabla f(y) - \nabla f(x) = 0$. Можно заключить, что для любого x минимум функции $\varphi(y)$ достигается в точке $y = x$. Следовательно:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla \varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla \varphi(y)\|_2^2$$

4. Теперь подставим $\varphi(y) = f(y) - \langle \nabla f(x), y\rangle$:

$$f(x) - \langle \nabla f(x), x\rangle \leq f(y) - \langle \nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

$$f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

Инструменты для доказательства сходимости

3. Из условий оптимальности первого порядка для выпуклой функции $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$. Можно заключить, что для любого x минимум функции $\varphi(y)$ достигается в точке $y = x$. Следовательно:

$$\varphi(x) \le \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \le \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Теперь подставим $\varphi(y) = f(y) - \langle\nabla f(x), y\rangle$:

$$f(x) - \langle\nabla f(x), x\rangle \le f(y) - \langle\nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

$$f(x) + \langle\nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \le f(y)$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 \le 2L\left(f(y) - f(x) - \langle\nabla f(x), y - x\rangle\right)$$

Инструменты для доказательства сходимости

3. Из условий оптимальности первого порядка для выпуклой функции $\nabla \varphi(y) = \nabla f(y) - \nabla f(x) = 0$. Можно заключить, что для любого x минимум функции $\varphi(y)$ достигается в точке $y = x$. Следовательно:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Теперь подставим $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$:

$$f(x) - \langle \nabla f(x), x \rangle \leq f(y) - \langle \nabla f(x), y \rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 \leq 2L\left(f(y) - f(x) - \langle \nabla f(x), y - x \rangle\right)$$

меняем x и y $$\|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L\left(f(x) - f(y) - \langle \nabla f(y), x - y \rangle\right)$$

Инструменты для доказательства сходимости

3. Из условий оптимальности первого порядка для выпуклой функции $\nabla \varphi(y) = \nabla f(y) - \nabla f(x) = 0$. Можно заключить, что для любого x минимум функции $\varphi(y)$ достигается в точке $y = x$. Следовательно:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Теперь подставим $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$:

$$f(x) - \langle \nabla f(x), x \rangle \leq f(y) - \langle \nabla f(x), y \rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 \leq 2L\left(f(y) - f(x) - \langle \nabla f(x), y - x \rangle\right)$$

меняем x и y $$\|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L\left(f(x) - f(y) - \langle \nabla f(y), x - y \rangle\right)$$

Инструменты для доказательства сходимости

3. Из условий оптимальности первого порядка для выпуклой функции $\nabla \varphi(y) = \nabla f(y) - \nabla f(x) = 0$. Можно заключить, что для любого x минимум функции $\varphi(y)$ достигается в точке $y = x$. Следовательно:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Теперь подставим $\varphi(y) = f(y) - \langle \nabla f(x), y\rangle$:

$$f(x) - \langle \nabla f(x), x\rangle \leq f(y) - \langle \nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

$$f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 \leq 2L\left(f(y) - f(x) - \langle \nabla f(x), y - x\rangle\right)$$

меняем x и y $$\|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L\left(f(x) - f(y) - \langle \nabla f(y), x - y\rangle\right)$$

Лемма доказана. На первый взгляд она не имеет очевидной геометрической интерпретации, но мы будем использовать её как удобный инструмент для оценки разности градиентов.

Инструменты для доказательства сходимости

> **Theorem**
>
> Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — непрерывно дифференцируемая функция на $\mathbb{R}^n$. Тогда функция f является μ-сильно выпуклой тогда и только тогда, когда для любых $x, y \in \mathbb{R}^d$ выполняется:
>
> $$\begin{aligned} \text{Сильно выпуклый случай } \mu > 0 \quad & \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2 \\ \text{Выпуклый случай } \mu = 0 \quad & \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0 \end{aligned}$$

Доказательство

1. Приведём доказательство только для сильно выпуклого случая; выпуклый следует из него при $\mu = 0$. Начнём с необходимости. Для сильно выпуклой функции

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

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

$$\text{сумма} \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$$

Инструменты для доказательства сходимости

2. Для достаточности предположим, что $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Используя формулу Ньютона-Лейбница $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x-y)), x - y \rangle dt$:

Инструменты для доказательства сходимости

2. Для достаточности предположим, что $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Используя формулу Ньютона-Лейбница $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$:

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle = \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle$$

Инструменты для доказательства сходимости 💎💎💎💎

2. Для достаточности предположим, что $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Используя формулу Ньютона-Лейбница $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$:

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle = \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle$$

$$\overset{\langle \nabla f(y), x-y \rangle = \int_0^1 \langle \nabla f(y), x-y \rangle dt}{=} \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), (x - y) \rangle dt$$

Инструменты для доказательства сходимости 💎💎💎💎

2. Для достаточности предположим, что $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Используя формулу Ньютона-Лейбница $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$:

$$
\begin{aligned}
f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\
{\scriptstyle \langle \nabla f(y), x-y \rangle = \int_0^1 \langle \nabla f(y), x-y \rangle dt} \quad &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), (x - y) \rangle dt \\
{\scriptstyle y + t(x-y) - y = t(x-y)} \quad &= \int_0^1 t^{-1} \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt
\end{aligned}
$$

Инструменты для доказательства сходимости

2. Для достаточности предположим, что $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x-y\|^2$. Используя формулу Ньютона-Лейбница $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x-y)), x - y\rangle dt$:

$$
\begin{aligned}
& & f(x) - f(y) - \langle \nabla f(y), x - y\rangle &= \int_0^1 \langle \nabla f(y + t(x-y)), x - y\rangle dt - \langle \nabla f(y), x - y\rangle \\
& {\scriptstyle \langle \nabla f(y), x-y\rangle = \int_0^1 \langle \nabla f(y), x-y\rangle dt} & &= \int_0^1 \langle \nabla f(y + t(x-y)) - \nabla f(y), (x - y)\rangle dt \\
& {\scriptstyle y + t(x-y) - y = t(x-y)} & &= \int_0^1 t^{-1}\langle \nabla f(y + t(x-y)) - \nabla f(y), t(x - y)\rangle dt \\
& & &\geq \int_0^1 t^{-1}\mu\|t(x-y)\|^2 dt
\end{aligned}
$$

Инструменты для доказательства сходимости

2. Для достаточности предположим, что $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Используя формулу Ньютона-Лейбница $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x-y)), x - y \rangle dt$:

$$
\begin{aligned}
f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x-y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\
{\scriptstyle \langle \nabla f(y), x-y \rangle = \int_0^1 \langle \nabla f(y), x-y \rangle dt} \qquad &= \int_0^1 \langle \nabla f(y + t(x-y)) - \nabla f(y), (x - y) \rangle dt \\
{\scriptstyle y + t(x-y) - y = t(x-y)} \qquad &= \int_0^1 t^{-1} \langle \nabla f(y + t(x-y)) - \nabla f(y), t(x - y) \rangle dt \\
&\geq \int_0^1 t^{-1} \mu \|t(x-y)\|^2 dt = \mu \|x - y\|^2 \int_0^1 t dt
\end{aligned}
$$

Инструменты для доказательства сходимости

2. Для достаточности предположим, что $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Используя формулу Ньютона-Лейбница $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x-y)), x - y \rangle dt$:

$$
\begin{aligned}
f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x-y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\
&\overset{\langle \nabla f(y), x-y \rangle = \int_0^1 \langle \nabla f(y), x-y \rangle dt}{=} \int_0^1 \langle \nabla f(y + t(x-y)) - \nabla f(y), (x - y) \rangle dt \\
&\overset{y + t(x-y) - y = t(x-y)}{=} \int_0^1 t^{-1} \langle \nabla f(y + t(x-y)) - \nabla f(y), t(x - y) \rangle dt \\
&\geq \int_0^1 t^{-1} \mu \|t(x-y)\|^2 dt = \mu \|x - y\|^2 \int_0^1 t dt = \frac{\mu}{2} \|x - y\|_2^2
\end{aligned}
$$

Инструменты для доказательства сходимости

2. Для достаточности предположим, что $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x-y\|^2$. Используя формулу Ньютона-Лейбница $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x-y)), x - y\rangle dt$:

$$
\begin{aligned}
f(x) - f(y) - \langle \nabla f(y), x - y\rangle &= \int_0^1 \langle \nabla f(y + t(x-y)), x - y\rangle dt - \langle \nabla f(y), x - y\rangle \\
\overset{\langle \nabla f(y), x-y\rangle = \int_0^1 \langle \nabla f(y), x-y\rangle dt}{} &= \int_0^1 \langle \nabla f(y + t(x-y)) - \nabla f(y), (x - y)\rangle dt \\
\overset{y + t(x-y) - y = t(x-y)}{} &= \int_0^1 t^{-1}\langle \nabla f(y + t(x-y)) - \nabla f(y), t(x - y)\rangle dt \\
&\geq \int_0^1 t^{-1}\mu\|t(x-y)\|^2 dt = \mu\|x-y\|^2 \int_0^1 t dt = \frac{\mu}{2}\|x-y\|_2^2
\end{aligned}
$$

Инструменты для доказательства сходимости

2. Для достаточности предположим, что $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Используя формулу Ньютона-Лейбница $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x-y)), x - y \rangle dt$:

$$
\begin{aligned}
f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x-y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\
\overset{\langle \nabla f(y), x-y \rangle = \int_0^1 \langle \nabla f(y), x-y \rangle dt}{} \quad &= \int_0^1 \langle \nabla f(y + t(x-y)) - \nabla f(y), (x - y) \rangle dt \\
\overset{y + t(x-y) - y = t(x-y)}{} \quad &= \int_0^1 t^{-1} \langle \nabla f(y + t(x-y)) - \nabla f(y), t(x - y) \rangle dt \\
&\geq \int_0^1 t^{-1} \mu \|t(x-y)\|^2 dt = \mu \|x - y\|^2 \int_0^1 t dt = \frac{\mu}{2} \|x - y\|_2^2
\end{aligned}
$$

Таким образом, критерий сильной выпуклости выполнен

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

Инструменты для доказательства сходимости

2. Для достаточности предположим, что $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Используя формулу Ньютона-Лейбница $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x-y)), x - y \rangle dt$:

$$
\begin{aligned}
f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x-y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\
\overset{\langle \nabla f(y), x-y \rangle = \int_0^1 \langle \nabla f(y), x-y \rangle dt}{} &= \int_0^1 \langle \nabla f(y + t(x-y)) - \nabla f(y), (x - y) \rangle dt \\
\overset{y + t(x-y) - y = t(x-y)}{} &= \int_0^1 t^{-1} \langle \nabla f(y + t(x-y)) - \nabla f(y), t(x - y) \rangle dt \\
&\geq \int_0^1 t^{-1} \mu \|t(x-y)\|^2 dt = \mu \|x - y\|^2 \int_0^1 t dt = \frac{\mu}{2} \|x - y\|_2^2
\end{aligned}
$$

Таким образом, критерий сильной выпуклости выполнен

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

Инструменты для доказательства сходимости

2. Для достаточности предположим, что $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Используя формулу Ньютона-Лейбница $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x-y)), x - y \rangle dt$:

$$
\begin{aligned}
f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x-y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\
{\scriptstyle \langle \nabla f(y), x-y \rangle = \int_0^1 \langle \nabla f(y), x-y \rangle dt} \quad &= \int_0^1 \langle \nabla f(y + t(x-y)) - \nabla f(y), (x - y) \rangle dt \\
{\scriptstyle y + t(x-y) - y = t(x-y)} \quad &= \int_0^1 t^{-1} \langle \nabla f(y + t(x-y)) - \nabla f(y), t(x - y) \rangle dt \\
&\geq \int_0^1 t^{-1} \mu \|t(x-y)\|^2 dt = \mu \|x-y\|^2 \int_0^1 t dt = \frac{\mu}{2} \|x-y\|_2^2
\end{aligned}
$$

Таким образом, критерий сильной выпуклости выполнен

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

меняем x и y $\quad -\langle \nabla f(x), x - y \rangle \leq -\left(f(x) - f(y) + \frac{\mu}{2} \|x - y\|_2^2 \right)$

Скорость сходимости для гладкого выпуклого случая

Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — выпуклая и дифференцируемая функция. Пусть $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, и существует минимизатор x^* функции f на S; кроме того, предположим, что f является гладкой на S с параметром L. Алгоритм метода проекции градиента с шагом $\frac{1}{L}$ обеспечивает следующую сходимость после $k > 0$ итераций:

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

$f(x) = \frac{1}{12}x^4 + \frac{1}{2}x^2$

$f'' = x^2 + 1$

В точности совпадает с GD

Скорость сходимости для гладкого выпуклого случая

Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — выпуклая и дифференцируемая функция. Пусть $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, и существует минимизатор x^* функции f на S; кроме того, предположим, что f является гладкой на S с параметром L. Алгоритм метода проекции градиента с шагом $\frac{1}{L}$ обеспечивает следующую сходимость после $k > 0$ итераций:

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

$x_{k+1} = \Pi_S(y_k)$

1. Докажем лемму о достаточном убывании, полагая $\boxed{y_k = x_k - \frac{1}{L}\nabla f(x_k)}$ и используя теорему косинусов $2x^T y = \|x\|^2 + \|y\|^2 - \|x - y\|^2$:

Скорость сходимости для гладкого выпуклого случая

Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — выпуклая и дифференцируемая функция. Пусть $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, и существует минимизатор x^* функции f на S; кроме того, предположим, что f является гладкой на S с параметром L. Алгоритм метода проекции градиента с шагом $\frac{1}{L}$ обеспечивает следующую сходимость после $k > 0$ итераций:

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

1. Докажем лемму о достаточном убывании, полагая $y_k = x_k - \frac{1}{L}\nabla f(x_k)$ и используя теорему косинусов $2x^T y = \|x\|^2 + \|y\|^2 - \|x - y\|^2$:

Гладкость:

$$\boxed{f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2}$$

липшицева парабола

Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — выпуклая и дифференцируемая функция. Пусть $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, и существует минимизатор x^* функции f на S; кроме того, предположим, что f является гладкой на S с параметром L. Алгоритм метода проекции градиента с шагом $\frac{1}{L}$ обеспечивает следующую сходимость после $k > 0$ итераций:

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

1. Докажем лемму о достаточном убывании, полагая $y_k = x_k - \frac{1}{L}\nabla f(x_k)$ и используя теорему косинусов $2x^Ty = \|x\|^2 + \|y\|^2 - \|x - y\|^2$:

$\nabla f(x_k) = (x_k - y_k) \cdot L$

$$
\text{Гладкость:} \quad f(x_{k+1}) \leq f(x_k) + \langle \underline{\nabla f(x_k)}, x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2
$$

$$
\text{Метод:} \quad = f(x_k) - L\langle y_k - x_k, x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2
$$

Скорость сходимости для гладкого выпуклого случая

Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — выпуклая и дифференцируемая функция. Пусть $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, и существует минимизатор x^* функции f на S; кроме того, предположим, что f является гладкой на S с параметром L. Алгоритм метода проекции градиента с шагом $\frac{1}{L}$ обеспечивает следующую сходимость после $k > 0$ итераций:

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

1. Докажем лемму о достаточном убывании, полагая $y_k = x_k - \frac{1}{L}\nabla f(x_k)$ и используя теорему косинусов $\boxed{2x^T y = \|x\|^2 + \|y\|^2 - \|x - y\|^2}$:

Гладкость: $$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

Метод: $$= f(x_k) - L\langle y_k - x_k, x_{k+1} - x_k \rangle + \underline{\frac{L}{2}\|x_{k+1} - x_k\|^2}$$

Теорема косинусов: $$= f(x_k) - \frac{L}{2}\left(\|y_k - x_k\|^2 + \|x_{k+1} - x_k\|^2 - \|y_k - x_{k+1}\|^2\right) + \underline{\frac{L}{2}\|x_{k+1} - x_k\|^2}$$

Скорость сходимости для гладкого выпуклого случая

Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — выпуклая и дифференцируемая функция. Пусть $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, и существует минимизатор x^* функции f на S; кроме того, предположим, что f является гладкой на S с параметром L. Алгоритм метода проекции градиента с шагом $\frac{1}{L}$ обеспечивает следующую сходимость после $k > 0$ итераций:

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

1. Докажем лемму о достаточном убывании, полагая $y_k = x_k - \frac{1}{L}\nabla f(x_k)$ и используя теорему косинусов $2x^Ty = \|x\|^2 + \|y\|^2 - \|x-y\|^2$:

$x^T y = \frac{1}{2}(\ldots)$

Гладкость: $$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

Метод: $$= f(x_k) - L\langle y_k - x_k, x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

Теорема косинусов: $$= f(x_k) - \frac{L}{2}\left(\|y_k - x_k\|^2 + \cancel{\|x_{k+1} - x_k\|^2} - \|y_k - x_{k+1}\|^2\right) + \cancel{\frac{L}{2}\|x_{k+1} - x_k\|^2}$$

$$= f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{L}{2}\|y_k - x_{k+1}\|^2$$

$\frac{1}{2L}\|\nabla f(x_k)\|_2^2 \leq f(x_k) - f(x_{k+1}) + \frac{L}{2} \ldots$

Скорость сходимости для гладкого выпуклого случая

2. Пока что мы не получаем немедленного прогресса на каждом шаге. Снова используем теорему косинусов:

$$
\left\langle \frac{1}{L}\nabla f(x_k), x_k - x^* \right\rangle = \frac{1}{2}\left(\frac{1}{L^2}\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L}\nabla f(x_k)\|^2\right)
$$

$= y_k$

$$
\langle \nabla f(x_k), x_k - x^* \rangle = \frac{L}{2}\left(\frac{1}{L^2}\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|y_k - x^*\|^2\right)
$$

Скорость сходимости для гладкого выпуклого случая

2. Пока что мы не получаем немедленного прогресса на каждом шаге. Снова используем теорему косинусов:

$$\left\langle \frac{1}{L}\nabla f(x_k), x_k - x^* \right\rangle = \frac{1}{2}\left(\frac{1}{L^2}\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L}\nabla f(x_k)\|^2\right)$$

$$\langle \nabla f(x_k), x_k - x^* \rangle = \frac{L}{2}\left(\frac{1}{L^2}\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|y_k - x^*\|^2\right)$$

3. Теперь используем свойство проекции: $\boxed{\|x - \text{proj}_S(y)\|^2 + \|y - \text{proj}_S(y)\|^2 \leq \|x - y\|^2 \text{ с } x = x^*, y = y_k:}$

$$\|x^* - \text{proj}_S(y_k)\|^2 + \|y_k - \text{proj}_S(y_k)\|^2 \leq \underline{\|x^* - y_k\|^2}$$

$$\underline{\|y_k - x^*\|^2} \geq \|x^* - x_{k+1}\|^2 + \|y_k - x_{k+1}\|^2$$

PROJ$(y_k) = x_{k+1}$

Скорость сходимости для гладкого выпуклого случая

2. Пока что мы не получаем немедленного прогресса на каждом шаге. Снова используем теорему косинусов:

$$\left\langle \frac{1}{L}\nabla f(x_k), x_k - x^* \right\rangle = \frac{1}{2}\left(\frac{1}{L^2}\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L}\nabla f(x_k)\|^2\right)$$
$$\langle \nabla f(x_k), x_k - x^* \rangle = \frac{L}{2}\left(\frac{1}{L^2}\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|y_k - x^*\|^2\right)$$

3. Теперь используем свойство проекции: $\|x - \text{proj}_S(y)\|^2 + \|y - \text{proj}_S(y)\|^2 \leq \|x - y\|^2$ с $x = x^*, y = y_k$:

$$\|x^* - \text{proj}_S(y_k)\|^2 + \|y_k - \text{proj}_S(y_k)\|^2 \leq \|x^* - y_k\|^2$$
$$\|y_k - x^*\|^2 \geq \|x^* - x_{k+1}\|^2 + \|y_k - x_{k+1}\|^2$$

4. Теперь, используя выпуклость и предыдущую часть:

Скорость сходимости для гладкого выпуклого случая

2. Пока что мы не получаем немедленного прогресса на каждом шаге. Снова используем теорему косинусов:

$$
\left\langle \frac{1}{L}\nabla f(x_k), x_k - x^* \right\rangle = \frac{1}{2}\left(\frac{1}{L^2}\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L}\nabla f(x_k)\|^2\right)
$$
$$
\langle \nabla f(x_k), x_k - x^* \rangle = \frac{L}{2}\left(\frac{1}{L^2}\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|y_k - x^*\|^2\right)
$$

3. Теперь используем свойство проекции: $\|x - \text{proj}_S(y)\|^2 + \|y - \text{proj}_S(y)\|^2 \leq \|x - y\|^2$ с $x = x^*, y = y_k$:

$$
\|x^* - \text{proj}_S(y_k)\|^2 + \|y_k - \text{proj}_S(y_k)\|^2 \leq \|x^* - y_k\|^2
$$
$$
\|y_k - x^*\|^2 \geq \|x^* - x_{k+1}\|^2 + \|y_k - x_{k+1}\|^2
$$

4. Теперь, используя выпуклость и предыдущую часть:

Выпуклость:
$$
\boxed{f(x_k) - f^* \leq \langle \nabla f(x_k), x_k - x^* \rangle}
$$

Скорость сходимости для гладкого выпуклого случая

2. Пока что мы не получаем немедленного прогресса на каждом шаге. Снова используем теорему косинусов:

$$
\left\langle \frac{1}{L}\nabla f(x_k), x_k - x^* \right\rangle = \frac{1}{2}\left(\frac{1}{L^2}\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L}\nabla f(x_k)\|^2\right)
$$

$$
\langle \nabla f(x_k), x_k - x^* \rangle = \frac{L}{2}\left(\frac{1}{L^2}\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|y_k - x^*\|^2\right)
$$

3. Теперь используем свойство проекции: $\|x - \text{proj}_S(y)\|^2 + \|y - \text{proj}_S(y)\|^2 \leq \|x - y\|^2$ с $x = x^*, y = y_k$:

$$
\|x^* - \text{proj}_S(y_k)\|^2 + \|y_k - \text{proj}_S(y_k)\|^2 \leq \|x^* - y_k\|^2
$$

$$
\|y_k - x^*\|^2 \geq \|x^* - x_{k+1}\|^2 + \|y_k - x_{k+1}\|^2
$$

4. Теперь, используя выпуклость и предыдущую часть:

Выпуклость:
$$
\begin{aligned}
f(x_k) - f^* &\leq \langle \nabla f(x_k), x_k - x^* \rangle \\
&\leq \frac{L}{2}\left(\frac{1}{L^2}\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 - \|y_k - x_{k+1}\|^2\right)
\end{aligned}
$$

Скорость сходимости для гладкого выпуклого случая

2. Пока что мы не получаем немедленного прогресса на каждом шаге. Снова используем теорему косинусов:

$$\left\langle \frac{1}{L}\nabla f(x_k), x_k - x^* \right\rangle = \frac{1}{2}\left(\frac{1}{L^2}\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L}\nabla f(x_k)\|^2\right)$$
$$\langle \nabla f(x_k), x_k - x^* \rangle = \frac{L}{2}\left(\frac{1}{L^2}\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|y_k - x^*\|^2\right)$$

3. Теперь используем свойство проекции: $\|x - \text{proj}_S(y)\|^2 + \|y - \text{proj}_S(y)\|^2 \leq \|x - y\|^2$ с $x = x^*, y = y_k$:

$$\|x^* - \text{proj}_S(y_k)\|^2 + \|y_k - \text{proj}_S(y_k)\|^2 \leq \|x^* - y_k\|^2$$
$$\|y_k - x^*\|^2 \geq \|x^* - x_{k+1}\|^2 + \|y_k - x_{k+1}\|^2$$

4. Теперь, используя выпуклость и предыдущую часть:

ТЕЛЕСКОП. СУММА

Выпуклость:
$$f(x_k) - f^* \leq \langle \nabla f(x_k), x_k - x^* \rangle$$
$$\leq \frac{L}{2}\left(\frac{1}{L^2}\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 - \|y_k - x_{k+1}\|^2\right)$$

Сумма для $i = 0, k-1$ $\quad \sum\limits_{i=0}^{k-1}[f(x_i) - f^*] \leq \sum\limits_{i=0}^{k-1}\frac{1}{2L}\|\nabla f(x_i)\|^2 + \frac{L}{2}\|x_0 - x^*\|^2 - \frac{L}{2}\sum\limits_{i=0}^{i-1}\|y_i - x_{i+1}\|^2$

Скорость сходимости для гладкого выпуклого случая

5. Оценим градиенты с помощью неравенства достаточного убывания 7:

Скорость сходимости для гладкого выпуклого случая

5. Оценим градиенты с помощью неравенства достаточного убывания 7:

$$\sum_{i=0}^{k-1} [f(x_i) - f^*] \leq \sum_{i=0}^{k-1} \left[f(x_i) - f(x_{i+1}) + \frac{L}{2} \|y_i - x_{i+1}\|^2 \right] + \frac{L}{2} \|x_0 - x^*\|^2 - \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2$$

Скорость сходимости для гладкого выпуклого случая

5. Оценим градиенты с помощью неравенства достаточного убывания 7:

$$
\sum_{i=0}^{k-1} [f(x_i) - f^*] \le \sum_{i=0}^{k-1} \left[\underline{f(x_i) - f(x_{i+1})} + \frac{L}{2}\|y_i - x_{i+1}\|^2 \right] + \frac{L}{2}\|x_0 - x^*\|^2 - \frac{L}{2}\sum_{i=0}^{i-1}\|y_i - x_{i+1}\|^2
$$

$$
\le \underline{f(x_0)} - \underline{f(x_k)} + \cancel{\frac{L}{2}\sum_{i=0}^{i-1}\|y_i - x_{i+1}\|^2} + \frac{L}{2}\|x_0 - x^*\|^2 - \cancel{\frac{L}{2}\sum_{i=0}^{i-1}\|y_i - x_{i+1}\|^2}
$$

Скорость сходимости для гладкого выпуклого случая

5. Оценим градиенты с помощью неравенства достаточного убывания 7:

$$\sum_{i=0}^{k-1} [f(x_i) - f^*] \leq \sum_{i=0}^{k-1} \left[f(x_i) - f(x_{i+1}) + \frac{L}{2} \|y_i - x_{i+1}\|^2 \right] + \frac{L}{2} \|x_0 - x^*\|^2 - \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2$$

$$\leq f(x_0) - f(x_k) + \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2 + \frac{L}{2} \|x_0 - x^*\|^2 - \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2$$

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

Скорость сходимости для гладкого выпуклого случая

5. Оценим градиенты с помощью неравенства достаточного убывания 7:

$$
\begin{aligned}
\sum_{i=0}^{k-1} [f(x_i) - f^*] &\leq \sum_{i=0}^{k-1} \left[f(x_i) - f(x_{i+1}) + \frac{L}{2}\|y_i - x_{i+1}\|^2 \right] + \frac{L}{2}\|x_0 - x^*\|^2 - \frac{L}{2}\sum_{i=0}^{i-1}\|y_i - x_{i+1}\|^2 \\
&\leq f(x_0) - f(x_k) + \frac{L}{2}\sum_{i=0}^{i-1}\|y_i - x_{i+1}\|^2 + \frac{L}{2}\|x_0 - x^*\|^2 - \frac{L}{2}\sum_{i=0}^{i-1}\|y_i - x_{i+1}\|^2 \\
&\leq f(x_0) - f(x_k) + \frac{L}{2}\|x_0 - x^*\|^2 \\
\sum_{i=0}^{k-1} f(x_i) - kf^* &\leq f(x_0) - f(x_k) + \frac{L}{2}\|x_0 - x^*\|^2
\end{aligned}
$$

Скорость сходимости для гладкого выпуклого случая

5. Оценим градиенты с помощью неравенства достаточного убывания 7:

$$
\begin{aligned}
\sum_{i=0}^{k-1}[f(x_i) - f^*] \leq \sum_{i=0}^{k-1}\left[f(x_i) - f(x_{i+1}) + \frac{L}{2}\|y_i - x_{i+1}\|^2\right] + \frac{L}{2}\|x_0 - x^*\|^2 - \frac{L}{2}\sum_{i=0}^{i-1}\|y_i - x_{i+1}\|^2 \\
\leq f(x_0) - f(x_k) + \frac{L}{2}\sum_{i=0}^{i-1}\|y_i - x_{i+1}\|^2 + \frac{L}{2}\|x_0 - x^*\|^2 - \frac{L}{2}\sum_{i=0}^{i-1}\|y_i - x_{i+1}\|^2 \\
\leq f(x_0) - f(x_k) + \frac{L}{2}\|x_0 - x^*\|^2
\end{aligned}
$$

$$
\sum_{i=0}^{k-1} f(x_i) - kf^* \leq f(x_0) - f(x_k) + \frac{L}{2}\|x_0 - x^*\|^2
$$

$$
\sum_{i=1}^{k}[f(x_i) - f^*] \leq \frac{L}{2}\|x_0 - x^*\|^2
$$

Скорость сходимости для гладкого выпуклого случая

6. Из неравенства достаточного убывания

$$
f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{L}{2}\|y_k - x_{k+1}\|^2,
$$

Скорость сходимости для гладкого выпуклого случая

6. Из неравенства достаточного убывания

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{L}{2}\|y_k - x_{k+1}\|^2,$$

Скорость сходимости для гладкого выпуклого случая

6. Из неравенства достаточного убывания

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{L}{2}\|y_k - x_{k+1}\|^2,$$

используем тот факт, что $\underline{x_{k+1} = \text{proj}_S(y_k)}$. По определению проекции,

$$\|y_k - x_{k+1}\| \leq \|y_k - x_k\|,$$

Скорость сходимости для гладкого выпуклого случая

6. Из неравенства достаточного убывания

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{L}{2}\|y_k - x_{k+1}\|^2,$$

$\leq \frac{L}{2} \cdot \frac{1}{L^2}\|\nabla f(x_k)\|_2^2$

используем тот факт, что $x_{k+1} = \text{proj}_S(y_k)$. По определению проекции,

$$\|y_k - x_{k+1}\| \leq \|y_k - x_k\|,$$

и вспомним, что $y_k = x_k - \frac{1}{L}\nabla f(x_k)$ означает $\boxed{\|y_k - x_k\| = \frac{1}{L}\|\nabla f(x_k)\|}$. Следовательно,

$$\frac{L}{2}\|y_k - x_{k+1}\|^2 \leq \frac{L}{2}\|y_k - x_k\|^2 = \frac{L}{2}\frac{1}{L^2}\|\nabla f(x_k)\|^2 = \frac{1}{2L}\|\nabla f(x_k)\|^2.$$

Скорость сходимости для гладкого выпуклого случая

6. Из неравенства достаточного убывания

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{L}{2}\|y_k - x_{k+1}\|^2,$$

используем тот факт, что $x_{k+1} = \text{proj}_S(y_k)$. По определению проекции,

$$\|y_k - x_{k+1}\| \leq \|y_k - x_k\|,$$

и вспомним, что $y_k = x_k - \frac{1}{L}\nabla f(x_k)$ означает $\|y_k - x_k\| = \frac{1}{L}\|\nabla f(x_k)\|$. Следовательно,

$$\frac{L}{2}\|y_k - x_{k+1}\|^2 \leq \frac{L}{2}\|y_k - x_k\|^2 = \frac{L}{2}\frac{1}{L^2}\|\nabla f(x_k)\|^2 = \frac{1}{2L}\|\nabla f(x_k)\|^2.$$

Подставим обратно в $(*)$:

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{1}{2L}\|\nabla f(x_k)\|^2 = f(x_k).$$

Скорость сходимости для гладкого выпуклого случая 💎💎💎💎💎

6. Из неравенства достаточного убывания

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{L}{2}\|y_k - x_{k+1}\|^2,$$

используем тот факт, что $x_{k+1} = \mathrm{proj}_S(y_k)$. По определению проекции,

$$\|y_k - x_{k+1}\| \leq \|y_k - x_k\|,$$

и вспомним, что $y_k = x_k - \frac{1}{L}\nabla f(x_k)$ означает $\|y_k - x_k\| = \frac{1}{L}\|\nabla f(x_k)\|$. Следовательно,

$$\frac{L}{2}\|y_k - x_{k+1}\|^2 \leq \frac{L}{2}\|y_k - x_k\|^2 = \frac{L}{2}\frac{1}{L^2}\|\nabla f(x_k)\|^2 = \frac{1}{2L}\|\nabla f(x_k)\|^2.$$

Подставим обратно в $(*)$:

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{1}{2L}\|\nabla f(x_k)\|^2 = f(x_k).$$

Таким образом,

$$f(x_{k+1}) \leq f(x_k) \quad \text{для каждого } k,$$

то есть $\{f(x_k)\}$ — монотонно невозрастающая последовательность.

Скорость сходимости для гладкого выпуклого случая

7. Итоговая оценка сходимости Из шага 5 мы уже установили

$$
\sum_{i=0}^{k-1} [f(x_i) - f^*] \leq \frac{L}{2} \|x_0 - x^*\|_2^2.
$$

Скорость сходимости для гладкого выпуклого случая 💎💎💎💎💎

7. Итоговая оценка сходимости Из шага 5 мы уже установили

$$
\sum_{i=0}^{k-1} [f(x_i) - f^*] \leq \frac{L}{2} \|x_0 - x^*\|_2^2.
$$

Скорость сходимости для гладкого выпуклого случая 💎💎💎💎💎

7. Итоговая оценка сходимости Из шага 5 мы уже установили

$$\sum_{i=0}^{k-1}[f(x_i) - f^*] \leq \frac{L}{2}\|x_0 - x^*\|_2^2.$$

Поскольку $f(x_i)$ убывает по i, в частности $f(x_k) \leq f(x_i)$ для всех $i \leq k$. Следовательно,

$$\boxed{k\left[f(x_k) - f^*\right] \leq \sum_{i=0}^{k-1}[f(x_i) - f^*]} \leq \frac{L}{2}\|x_0 - x^*\|_2^2,$$

Скорость сходимости для гладкого выпуклого случая 💎💎💎💎💎

7. Итоговая оценка сходимости Из шага 5 мы уже установили

$$\sum_{i=0}^{k-1}[f(x_i) - f^*] \leq \frac{L}{2}\|x_0 - x^*\|_2^2.$$

Поскольку $f(x_i)$ убывает по i, в частности $f(x_k) \leq f(x_i)$ для всех $i \leq k$. Следовательно,

$$k\left[f(x_k) - f^*\right] \leq \sum_{i=0}^{k-1}[f(x_i) - f^*] \leq \frac{L}{2}\|x_0 - x^*\|_2^2,$$

откуда немедленно получаем

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

!

Это завершает доказательство скорости сходимости $\mathcal{O}(\frac{1}{k})$ для выпуклой L-гладкой функции f в условной задаче оптимизации.

Скорость сходимости для гладкого сильно выпуклого случая

x^* $-\nabla f(x^*)$

Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — μ-сильно выпуклая функция. Пусть $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, и существует минимизатор x^* функции f на S; кроме того, предположим, что f является гладкой на S с параметром L. Алгоритм метода проекции градиента с шагом $\alpha \leq \frac{1}{L}$ обеспечивает следующую сходимость после $k > 0$ итераций:

$x^* \xrightarrow{PGD} x^*$

$$\boxed{\|x_k - x^*\|_2^2 \leq (1 - \alpha\mu)^k \|x_0 - x^*\|_2^2}$$

Доказательство

1. Сначала докажем свойство стационарной точки: $\boxed{\text{proj}_S(x^* - \alpha\nabla f(x^*)) = x^*.}$

   Это следует из критерия проекции и условия оптимальности первого порядка для x^*. Пусть $y = x^* - \alpha\nabla f(x^*)$. Нужно показать, что $\langle y - x^*, x - x^*\rangle \leq 0$ для всех $x \in S$.

$$\boxed{\langle (x^* - \alpha\nabla f(x^*)) - x^*, x - x^*\rangle = -\alpha\langle \nabla f(x^*), x - x^*\rangle \leq 0}$$

   Неравенство выполняется, так как $\alpha > 0$ и $\langle \nabla f(x^*), x - x^*\rangle \geq 0$ — условие оптимальности для x^*.

Скорость сходимости для гладкого сильно выпуклого случая

1. Рассмотрим расстояние до решения, используя свойство стационарной точки:

Скорость сходимости для гладкого сильно выпуклого случая

1. Рассмотрим расстояние до решения, используя свойство стационарной точки:

$$
\|x_{k+1} - x^*\|_2^2 = \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2
$$

Скорость сходимости для гладкого сильно выпуклого случая

1. Рассмотрим расстояние до решения, используя свойство стационарной точки:

$$
\begin{aligned}
\|x_{k+1} - x^*\|_2^2 &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2 \\
\text{свойство стационарной точки} &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - \text{proj}_S(x^* - \alpha \nabla f(x^*))\|_2^2
\end{aligned}
$$

Скорость сходимости для гладкого сильно выпуклого случая

1. Рассмотрим расстояние до решения, используя свойство стационарной точки:

$$
\begin{aligned}
\|x_{k+1} - x^*\|_2^2 &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2 \\
\text{свойство стационарной точки} &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - \text{proj}_S(x^* - \alpha \nabla f(x^*))\|_2^2 \\
\text{нерастягиваемость} &\leq \|x_k - \alpha \nabla f(x_k) - (x^* - \alpha \nabla f(x^*))\|_2^2
\end{aligned}
$$

Скорость сходимости для гладкого сильно выпуклого случая

1. Рассмотрим расстояние до решения, используя свойство стационарной точки:

$$
\begin{aligned}
\|x_{k+1} - x^*\|_2^2 &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2 \\
\text{свойство стационарной точки} &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - \text{proj}_S(x^* - \alpha \nabla f(x^*))\|_2^2 \\
\text{нерастягиваемость} &\leq \|x_k - \alpha \nabla f(x_k) - (x^* - \alpha \nabla f(x^*))\|_2^2 \\
&= \|x_k - x^*\|^2 - 2\alpha \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|_2^2
\end{aligned}
$$

Скорость сходимости для гладкого сильно выпуклого случая

1. Рассмотрим расстояние до решения, используя свойство стационарной точки:

$$
\begin{aligned}
\|x_{k+1} - x^*\|_2^2 &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2 \\
{\scriptstyle \text{свойство стационарной точки}} &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - \text{proj}_S(x^* - \alpha \nabla f(x^*))\|_2^2 \\
{\scriptstyle \text{нерастягиваемость}} &\leq \|x_k - \alpha \nabla f(x_k) - (x^* - \alpha \nabla f(x^*))\|_2^2 \\
&= \|x_k - x^*\|^2 - 2\alpha \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|_2^2
\end{aligned}
$$

2. Теперь используем гладкость из инструментов сходимости и сильную выпуклость:

Скорость сходимости для гладкого сильно выпуклого случая

1. Рассмотрим расстояние до решения, используя свойство стационарной точки:

$$
\begin{aligned}
\|x_{k+1} - x^*\|_2^2 &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2 \\
\text{свойство стационарной точки} &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - \text{proj}_S(x^* - \alpha \nabla f(x^*))\|_2^2 \\
\text{нерастягиваемость} &\leq \|x_k - \alpha \nabla f(x_k) - (x^* - \alpha \nabla f(x^*))\|_2^2 \\
&= \|x_k - x^*\|^2 - 2\alpha \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|_2^2
\end{aligned}
$$

2. Теперь используем гладкость из инструментов сходимости и сильную выпуклость:

$$
\text{гладкость} \quad \|\nabla f(x_k) - \nabla f(x^*)\|_2^2 \leq 2L\left(f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle\right)
$$

Скорость сходимости для гладкого сильно выпуклого случая

1. Рассмотрим расстояние до решения, используя свойство стационарной точки:

$$
\begin{aligned}
\|x_{k+1} - x^*\|_2^2 &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2 \\
\text{свойство стационарной точки} &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - \text{proj}_S(x^* - \alpha \nabla f(x^*))\|_2^2 \\
\text{нерастягиваемость} &\leq \|x_k - \alpha \nabla f(x_k) - (x^* - \alpha \nabla f(x^*))\|_2^2 \\
&= \|x_k - x^*\|^2 - 2\alpha \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|_2^2
\end{aligned}
$$

2. Теперь используем гладкость из инструментов сходимости и сильную выпуклость:

$$
\begin{aligned}
\text{гладкость} \quad & \|\nabla f(x_k) - \nabla f(x^*)\|_2^2 \leq 2L \left(f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle \right) \\
\text{сильная выпуклость} \quad & -\langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle \leq -\left(f(x_k) - f(x^*) + \frac{\mu}{2} \|x_k - x^*\|_2^2 \right) - \langle \nabla f(x^*), x_k - x
\end{aligned}
$$

Скорость сходимости для гладкого сильно выпуклого случая

3. Подставим:

Скорость сходимости для гладкого сильно выпуклого случая 💎💎💎

3. Подставим:

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

Скорость сходимости для гладкого сильно выпуклого случая

3. Подставим:

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

Скорость сходимости для гладкого сильно выпуклого случая

3. Подставим:

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

4. В силу выпуклости f: $f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle \geq 0$. Следовательно, при $\alpha \leq \frac{1}{L}$:

$$
\|x_{k+1} - x^*\|_2^2 \leq (1 - \alpha\mu)\|x_k - x^*\|^2,
$$

что означает линейную сходимость метода со скоростью $1 - \frac{\mu}{L}$.

$$S = \{x \in \mathbb{R}^n : x \geq 0; \mathbb{1}^T x = 1\}$$

Метод Франк-Вульфа

Рис. 11: Маргарит Страус Франк (1927-2024)

Рис. 12: Филип Вульф (1927-2016)

Идея

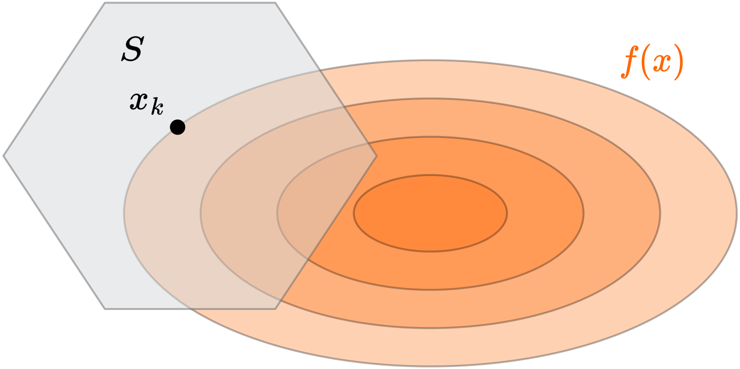


Рис. 13: Иллюстрация алгоритма Франк-Вульфа (условного градиента)

Идея

$$f^{I}_{x_k}(x) = f(x_k) + \nabla f(x_k)^T(x - x_k)$$

$$\underset{x \in S'}{\operatorname{argmin}} f^{I}_{x_k}(x)$$

Рис. 14: Иллюстрация алгоритма Франк-Вульфа (условного градиента)

Идея

$$y_k = \underset{x \in S}{\operatorname{argmin}} f^I_{x_k}(x) = $$
$$= \underset{x \in S}{\operatorname{argmin}} \nabla f(x_k)^T x$$

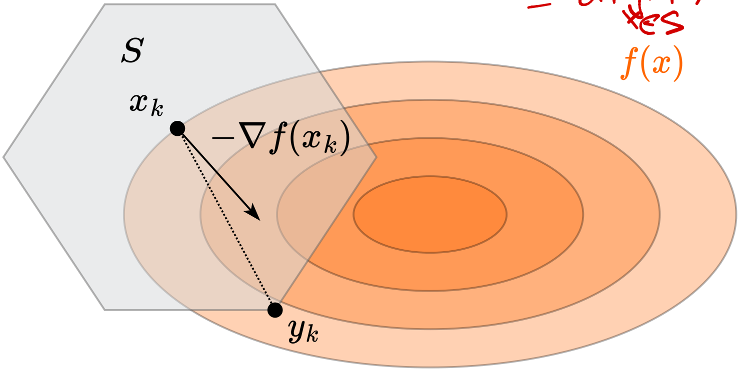


Рис. 15: Иллюстрация алгоритма Франк-Вульфа (условного градиента)

Идея

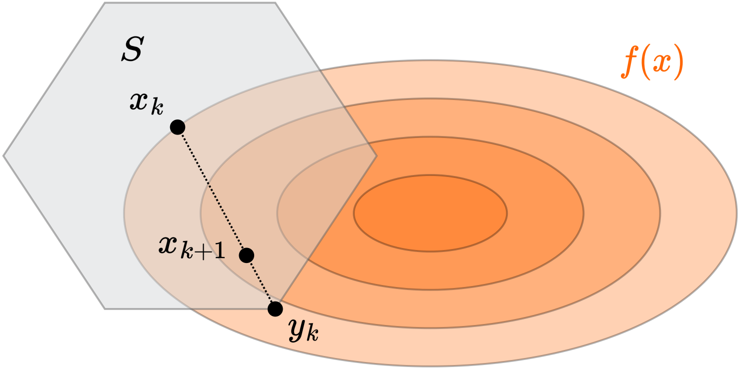


Рис. 16: Иллюстрация алгоритма Франк-Вульфа (условного градиента)

Идея

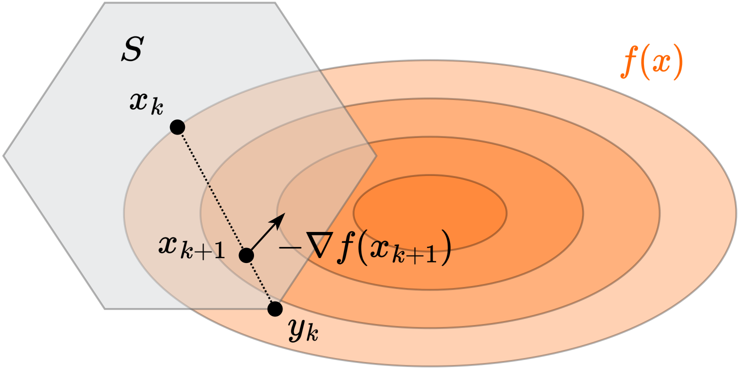


Рис. 17: Иллюстрация алгоритма Франк-Вульфа (условного градиента)

Идея

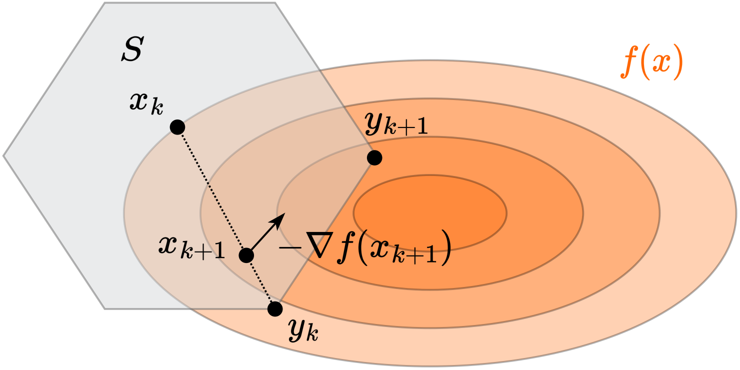


Рис. 18: Иллюстрация алгоритма Франк-Вульфа (условного градиента)

Идея

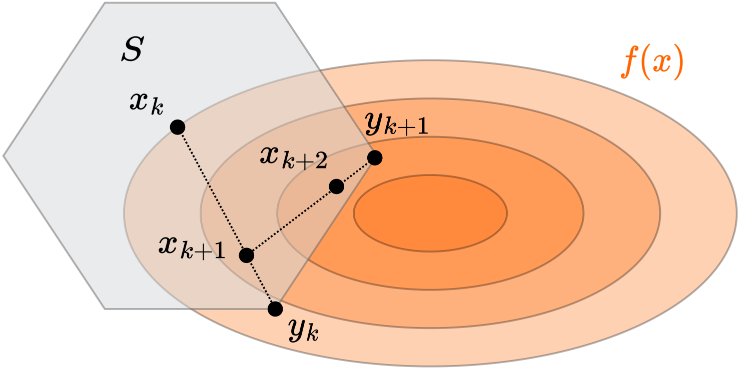


Рис. 19: Иллюстрация алгоритма Франк-Вульфа (условного градиента)

Идея

LMO

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

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

- convex comb.

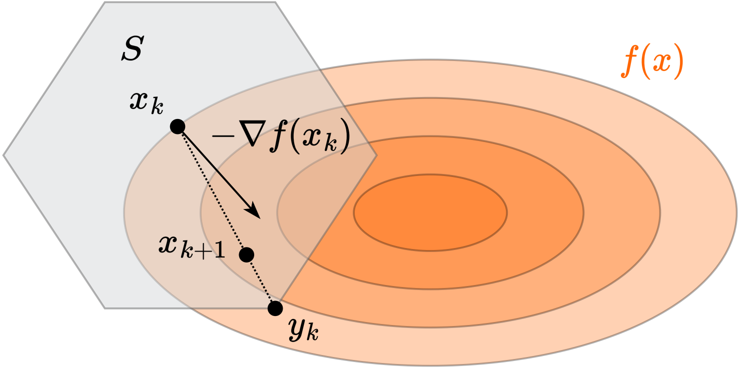


Рис. 20: Иллюстрация алгоритма Франк-Вульфа (условного градиента)

Скорость сходимости для гладкого выпуклого случая

Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — выпуклая и дифференцируемая функция. Пусть $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, и существует минимизатор x^* функции f на S; кроме того, предположим, что f является гладкой на S с параметром L. Алгоритм Франк-Вульфа с размером шага $\boxed{\gamma_k = \frac{k-1}{k+1}}$ обеспечивает следующую сходимость после $k > 0$ итераций:

$$\boxed{f(x_k) - f^* \leq \frac{2LR^2}{k+1}}$$

$k \geq 1 \quad \Big| \ \frac{k}{k+2}$

где $R = \max\limits_{x,y \in S} \|x - y\|$ — диаметр множества S.

LMO может быть дешевле проекции

Скорость сходимости для гладкого выпуклого случая

Theorem

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — выпуклая и дифференцируемая функция. Пусть $S \subseteq \mathbb{R}^n$ — замкнутое выпуклое множество, и существует минимизатор x^* функции f на S; кроме того, предположим, что f является гладкой на S с параметром L. Алгоритм Франк-Вульфа с размером шага $\gamma_k = \frac{k-1}{k+1}$ обеспечивает следующую сходимость после $k > 0$ итераций: $k \geq 1$

$$
f(x_k) - f^* \leq \frac{2LR^2}{k+1}
$$

$y_k = \arg\min\limits_{x \in S} f^I_{x_k}(x)$

где $R = \max\limits_{x,y \in S} \|x - y\|$ — диаметр множества S.

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

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

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

$= \gamma_k x_k - \gamma_k y_k + y_k$

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

1. Из L-гладкости f имеем:

$$
\boxed{f\left(x_{k+1}\right) - f\left(x_k\right) \leq \left\langle \nabla f\left(x_k\right), x_{k+1} - x_k \right\rangle + \frac{L}{2} \left\|x_{k+1} - x_k\right\|^2}
$$

$$
= (1 - \gamma_k) \left\langle \nabla f\left(x_k\right), y_k - x_k \right\rangle + \frac{L(1-\gamma_k)^2}{2} \left\|y_k - x_k\right\|^2
$$

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

Скорость сходимости для гладкого выпуклого случая

2. Из выпуклости f для любого $x \in S$, включая x^*:

$$\boxed{\langle \nabla f(x_k), x - x_k \rangle \leq f(x) - f(x_k)}$$

В частности, для $x = x^*$:

$$\boxed{\langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k)}$$

Скорость сходимости для гладкого выпуклого случая

2. Из выпуклости f для любого $x \in S$, включая x^*:

$$\langle \nabla f(x_k), x - x_k \rangle \leq f(x) - f(x_k)$$

   В частности, для $x = x^*$:

$$\langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k)$$

3. По определению y_k имеем $\boxed{\langle \nabla f(x_k), y_k \rangle \leq \langle \nabla f(x_k), x^* \rangle,}$ поэтому:

$$y_k = \arg\min_{x \in S} \nabla f(x_k)^T x$$

$$\underbrace{\langle \nabla f(x_k), y_k - x_k \rangle \leq \langle \nabla f(x_k), x^* - x_k \rangle}_{} \leq f(x^*) - f(x_k)$$

Скорость сходимости для гладкого выпуклого случая

2. Из выпуклости f для любого $x \in S$, включая x^*:

$$\langle \nabla f(x_k), x - x_k \rangle \leq f(x) - f(x_k)$$

   В частности, для $x = x^*$:

$$\langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k)$$

3. По определению y_k имеем $\langle \nabla f(x_k), y_k \rangle \leq \langle \nabla f(x_k), x^* \rangle$, поэтому:

$$\underline{\langle \nabla f(x_k), y_k - x_k \rangle} \leq \langle \nabla f(x_k), x^* - x_k \rangle \leq \underline{f(x^*) - f(x_k)}$$

4. Объединяя приведённые неравенства:

$$\begin{aligned}\underline{f\left(x_{k+1}\right) - f\left(x_k\right)} &\leq (1 - \gamma_k) \left\langle \nabla f\left(x_k\right), y_k - x_k \right\rangle + \frac{L(1-\gamma_k)^2}{2} \|y_k - x_k\|^2 \\ &\leq \underline{(1 - \gamma_k) \left(f(x^*) - f(x_k)\right) + \frac{L(1-\gamma_k)^2}{2} R^2}\end{aligned}$$

Скорость сходимости для гладкого выпуклого случая

2. Из выпуклости f для любого $x \in S$, включая x^*:
$$\langle \nabla f(x_k), x - x_k \rangle \leq f(x) - f(x_k)$$
   В частности, для $x = x^*$:
$$\langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k)$$
3. По определению y_k имеем $\langle \nabla f(x_k), y_k \rangle \leq \langle \nabla f(x_k), x^* \rangle$, поэтому:
$$\langle \nabla f(x_k), y_k - x_k \rangle \leq \langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k)$$
4. Объединяя приведённые неравенства:
$$\begin{aligned} f\left(x_{k+1}\right) - f\left(x_k\right) &\leq (1 - \gamma_k)\left\langle \nabla f\left(x_k\right), y_k - x_k \right\rangle + \frac{L(1-\gamma_k)^2}{2} \|y_k - x_k\|^2 \\ &\leq (1 - \gamma_k)\left(f(x^*) - f(x_k)\right) + \frac{L(1-\gamma_k)^2}{2} R^2 \end{aligned}$$
5. Перегруппируем слагаемые:
$$\underbrace{f\left(x_{k+1}\right) - f(x^*)}_{\sim \delta_{k+1}} \leq \gamma_k \underbrace{\left(f(x_k) - f(x^*)\right)}_{\sim \delta_k} + (1 - \gamma_k)^2 \frac{LR^2}{2} \qquad \Big| : LR^2$$

Скорость сходимости для гладкого выпуклого случая

$\delta_k \sim f(x_k) - f^*$

γ от $\frac{1}{3}$ до 1

6. Обозначив $\boxed{\delta_k = \frac{f(x_k)-f(x^*)}{LR^2}}$, получаем:

$$\boxed{\delta_{k+1} \leq \gamma_k \delta_k + \frac{(1-\gamma_k)^2}{2} = \frac{k-1}{k+1}\delta_k + \frac{2}{(k+1)^2}}$$

$\gamma_k = \frac{k-1}{k+1}$

$1 - \gamma_k = \frac{2}{k+1}$

Скорость сходимости для гладкого выпуклого случая

6. Обозначив $\delta_k = \frac{f(x_k)-f(x^*)}{LR^2}$, получаем:

$$\delta_{k+1} \leq \gamma_k \delta_k + \frac{(1-\gamma_k)^2}{2} = \frac{k-1}{k+1}\delta_k + \frac{2}{(k+1)^2}$$

7. Докажем по индукции, что $\boxed{\delta_k \leq \frac{2}{k+1}.}$

БАЗА: K=

что даёт нам искомый результат:

$$f(x_k) - f^* \leq \frac{2LR^2}{k+1}$$

Скорость сходимости для гладкого выпуклого случая

6. Обозначив $\delta_k = \frac{f(x_k)-f(x^*)}{LR^2}$, получаем:

$$\delta_{k+1} \leq \gamma_k \delta_k + \frac{(1-\gamma_k)^2}{2} = \frac{k-1}{k+1}\delta_k + \frac{2}{(k+1)^2}$$

$$\delta_2 \leq 0 + \frac{2}{2^2} \leq \frac{1}{2}$$

7. Докажем по индукции, что $\delta_k \leq \frac{2}{k+1}$.

   $k=1$

   $$\delta_1 \leq \frac{2}{2}$$

   - База: $\delta_2 \leq \frac{1}{2} < \frac{2}{3}$

   $$\delta_2 \leq \frac{2}{3}$$

что даёт нам искомый результат:

$$f(x_k) - f^* \leq \frac{2LR^2}{k+1}$$

Скорость сходимости для гладкого выпуклого случая

6. Обозначив $\delta_k = \frac{f(x_k)-f(x^*)}{LR^2}$, получаем:

$$\delta_{k+1} \leq \gamma_k \delta_k + \frac{(1-\gamma_k)^2}{2} = \frac{k-1}{k+1}\delta_k + \frac{2}{(k+1)^2}$$

7. Докажем по индукции, что $\delta_k \leq \frac{2}{k+1}$.
   - База: $\delta_2 \leq \frac{1}{2} < \frac{2}{3}$
   - Предположим $\delta_k \leq \frac{2}{k+1}$

что даёт нам искомый результат:

$$f(x_k) - f^* \leq \frac{2LR^2}{k+1}$$

Скорость сходимости для гладкого выпуклого случая

$O\left(\frac{1}{k}\right)$

6. Обозначив $\delta_k = \frac{f(x_k)-f(x^*)}{LR^2}$, получаем:

$$\delta_{k+1} \leq \gamma_k \delta_k + \frac{(1-\gamma_k)^2}{2} = \frac{k-1}{k+1}\delta_k + \frac{2}{(k+1)^2}$$

7. Докажем по индукции, что $\delta_k \leq \frac{2}{k+1}$.

$\frac{2}{k+1}$

- База: $\delta_2 \leq \frac{1}{2} < \frac{2}{3}$
- Предположим $\delta_k \leq \frac{2}{k+1}$
- Тогда $\delta_{k+1} \leq \frac{k-1}{k+1}\frac{2}{k+1} + \frac{2}{(k+1)^2} = \boxed{\frac{2k}{k^2+2k+1} < \frac{2}{k+2}}$

что даёт нам искомый результат:

$$\boxed{f(x_k) - f^* \leq \frac{2LR^2}{k+1}}$$

$$\frac{2k}{k^2+2k+1} < \frac{2k}{k^2+2k} = \frac{2}{k+2}$$

$$\delta_k \leq \frac{2}{k+1}$$

$$\frac{f_k - f^*}{LR^2} \leq \frac{2}{k+1}$$

$O\left(\frac{1}{k}\right)$

Нижняя оценка для метода Франк-Вульфа [2]

Theorem

Рассмотрим произвольный алгоритм, который обращается к допустимому множеству $S \subseteq \mathbb{R}^n$ только через оракул линейной минимизации (LMO). Пусть диаметр множества S равен R. Существует L-гладкая сильно выпуклая функция $f : \mathbb{R}^n \to \mathbb{R}$, для которой этому алгоритму потребуется не менее

$$\boxed{\min\left(\frac{n}{2}, \frac{LR^2}{16\varepsilon}\right)} \quad \sim O\left(\frac{1}{\varepsilon}\right) \quad O\left(\frac{1}{k}\right)$$

итераций (то есть вызовов LMO), чтобы построить точку $\hat{x} \in S$ с $f(\hat{x}) - \min\limits_{x \in S} f(x) \leq \varepsilon$. Нижняя оценка справедлива как для выпуклых, так и для сильно выпуклых функций.

$$\text{LMO} = \arg\min_{x \in S} \nabla f(x_k)^T x$$

[2] The Complexity of Large-scale Convex Programming under a Linear Optimization Oracle

Нижняя оценка для метода Франк-Вульфа [2]

Theorem

Рассмотрим произвольный алгоритм, который обращается к допустимому множеству $S \subseteq \mathbb{R}^n$ только через оракул линейной минимизации (LMO). Пусть диаметр множества S равен R. Существует L-гладкая сильно выпуклая функция $f : \mathbb{R}^n \to \mathbb{R}$, для которой этому алгоритму потребуется не менее

$$\min\left(\frac{n}{2}, \frac{LR^2}{16\varepsilon}\right)$$

итераций (то есть вызовов LMO), чтобы построить точку $\hat{x} \in S$ с $f(\hat{x}) - \min_{x \in S} f(x) \leq \varepsilon$. Нижняя оценка справедлива как для выпуклых, так и для сильно выпуклых функций.

Набросок доказательства. Рассмотрим следующую задачу оптимизации:

$$\min_{x \in S} f(x) = \min_{x \in S} \frac{1}{2}\|x\|_2^2$$

$$S = \left\{ x \in \mathbb{R}^n \mid x \geq 0, \sum_{i=1}^{n} x_i = 1 \right\}$$

Заметим, что:

- f является 1-гладкой;
- диаметр S равен $R = 2$;
- f сильно выпукла.

[2] The Complexity of Large-scale Convex Programming under a Linear Optimization Oracle

Нижняя оценка для метода Франк-Вульфа [3]

1. Оптимальное решение:

$$x^* = \frac{1}{n}\mathbf{1} = \frac{1}{n}\sum_{i=1}^{n} e_i, \quad \text{и} \quad f(x^*) = \frac{1}{2n},$$

где $e_i = (0, \dots, 0, \underset{\text{позиция } i}{1}, 0, \dots, 0)^\top$ — i-й стандартный базисный вектор.

Нижняя оценка для метода Франк-Вульфа [3]

1. Оптимальное решение:
$$x^* = \frac{1}{n}\mathbf{1} = \frac{1}{n}\sum_{i=1}^{n} e_i, \quad \text{и} \quad f(x^*) = \frac{1}{2n},$$
где $e_i = (0, \dots, 0, \underset{\text{позиция } i}{1}, 0, \dots, 0)^\top$ — i-й стандартный базисный вектор.

2. Оракул линейной минимизации (LMO) на S возвращает вершину e_i. После k итераций метод обнаружит не более k различных базисных векторов $e_{i_1}, \dots, e_{i_k}$. Наилучшая выпуклая комбинация, которую можно составить:
$$\hat{x} = \frac{1}{k}\sum_{j=1}^{k} e_{i_j}.$$

Нижняя оценка для метода Франк-Вульфа [3]

1. Оптимальное решение:

$$x^* = \frac{1}{n}\mathbf{1} = \frac{1}{n}\sum_{i=1}^{n} e_i, \quad \text{и} \quad f(x^*) = \frac{1}{2n},$$

   где $e_i = (0, \dots, 0, \underset{\text{позиция } i}{1}, 0, \dots, 0)^\top$ — i-й стандартный базисный вектор.

2. Оракул линейной минимизации (LMO) на S возвращает вершину e_i. После k итераций метод обнаружит не более k различных базисных векторов $e_{i_1}, \dots, e_{i_k}$. Наилучшая выпуклая комбинация, которую можно составить:

$$\hat{x} = \frac{1}{k}\sum_{j=1}^{k} e_{i_j}.$$

3. Вычислив значение функции в точке $\hat{x}$, получаем:

$$f(\hat{x}) - f(x^*) \geq \frac{1}{2}\left(\frac{1}{\min\{k, n\}} - \frac{1}{n}\right).$$

Нижняя оценка для метода Франк-Вульфа [3]

1. Оптимальное решение:

$$x^* = \frac{1}{n}\mathbf{1} = \frac{1}{n}\sum_{i=1}^{n} e_i, \quad \text{и} \quad f(x^*) = \frac{1}{2n},$$

где $e_i = (0, \dots, 0, \underset{\text{позиция } i}{1}, 0, \dots, 0)^\top$ — i-й стандартный базисный вектор.

2. Оракул линейной минимизации (LMO) на S возвращает вершину e_i. После k итераций метод обнаружит не более k различных базисных векторов $e_{i_1}, \dots, e_{i_k}$. Наилучшая выпуклая комбинация, которую можно составить:

$$\hat{x} = \frac{1}{k}\sum_{j=1}^{k} e_{i_j}.$$

3. Вычислив значение функции в точке $\hat{x}$, получаем:

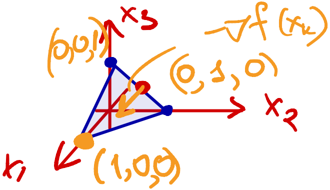
$$f(\hat{x}) - f(x^*) \geq \frac{1}{2}\left(\frac{1}{\min\{k, n\}} - \frac{1}{n}\right).$$

4. Чтобы обеспечить $f(\hat{x}) - f(x^*) \leq \varepsilon$, необходимо (полное доказательство приведено в статье):

$$k \geq \min\left\{\frac{n}{2}, \frac{1}{4\varepsilon}\right\} = \min\left\{\frac{n}{2}, \frac{LR^2}{16\varepsilon}\right\}.$$

[3] The Complexity of Large-scale Convex Programming under a Linear Optimization Oracle

Итоги метода Франк-Вульфа



- Метод не требует проекций, а в некоторых случаях позволяет вычислять итерации в замкнутой форме

$$\min \frac{1}{2} x^T x$$
$$1^T x = 1$$
$$x \geq 0$$

$$y_k = LMO(x_k) = \operatorname*{argmin}_{\substack{x \in \mathbb{R}^n \\ 1^T x = 1 \\ x \geq 0}} \nabla f(x_k)^T \cdot x = \operatorname*{argmin}_{x} \sum_{i=1}^{n} f_i'(x_k) \cdot x_i$$

$$= (0, \ldots 1, \ldots 0)$$

$$\operatorname*{argmin}_{1 \leq i \leq n} f_i'(x_k) - k$$

Итоги метода Франк-Вульфа

- Метод не требует проекций, а в некоторых случаях позволяет вычислять итерации в замкнутой форме
- Глобальная скорость сходимости составляет $O\left(\frac{1}{k}\right)$ для гладких выпуклых функций. Сильная выпуклость не улучшает скорость. Это нижняя оценка для LMO

Итоги метода Франк-Вульфа

- Метод не требует проекций, а в некоторых случаях позволяет вычислять итерации в замкнутой форме
- Глобальная скорость сходимости составляет $O\left(\frac{1}{k}\right)$ для гладких выпуклых функций. Сильная выпуклость не улучшает скорость. Это нижняя оценка для LMO
- По сравнению с методом проекции градиента скорость сходимости хуже, однако стоимость итерации может быть ниже, а решения — более разреженными

Итоги метода Франк-Вульфа

- Метод не требует проекций, а в некоторых случаях позволяет вычислять итерации в замкнутой форме
- Глобальная скорость сходимости составляет $O\left(\frac{1}{k}\right)$ для гладких выпуклых функций. Сильная выпуклость не улучшает скорость. Это нижняя оценка для LMO
- По сравнению с методом проекции градиента скорость сходимости хуже, однако стоимость итерации может быть ниже, а решения — более разреженными
- Недавно было показано, что для сильно выпуклых множеств скорость может быть улучшена до $O\left(\frac{1}{k^2}\right)$ (статья)

Итоги метода Франк-Вульфа

- Метод не требует проекций, а в некоторых случаях позволяет вычислять итерации в замкнутой форме
- Глобальная скорость сходимости составляет $O\left(\frac{1}{k}\right)$ для гладких выпуклых функций. Сильная выпуклость не улучшает скорость. Это нижняя оценка для LMO
- По сравнению с методом проекции градиента скорость сходимости хуже, однако стоимость итерации может быть ниже, а решения — более разреженными
- Недавно было показано, что для сильно выпуклых множеств скорость может быть улучшена до $O\left(\frac{1}{k^2}\right)$ (статья)

" (это типа моментума)

- Если допустить шаги удаления, сходимость становится линейной (статья) в сильно выпуклом случае

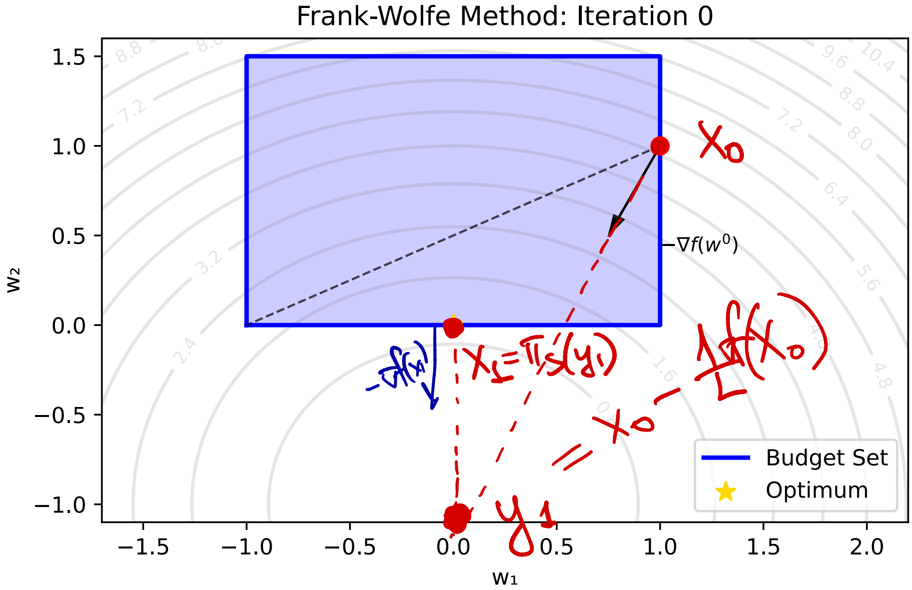
Итоги метода Франк-Вульфа

- Метод не требует проекций, а в некоторых случаях позволяет вычислять итерации в замкнутой форме
- Глобальная скорость сходимости составляет $O\left(\frac{1}{k}\right)$ для гладких выпуклых функций. Сильная выпуклость не улучшает скорость. Это нижняя оценка для LMO
- По сравнению с методом проекции градиента скорость сходимости хуже, однако стоимость итерации может быть ниже, а решения — более разреженными
- Недавно было показано, что для сильно выпуклых множеств скорость может быть улучшена до $O\left(\frac{1}{k^2}\right)$ (статья)
- Если допустить шаги удаления, сходимость становится линейной (статья) в сильно выпуклом случае
- В недавних работах показано обобщение на негладкий случай (статья) со скоростью сходимости $O\left(\frac{1}{\sqrt{k}}\right)$

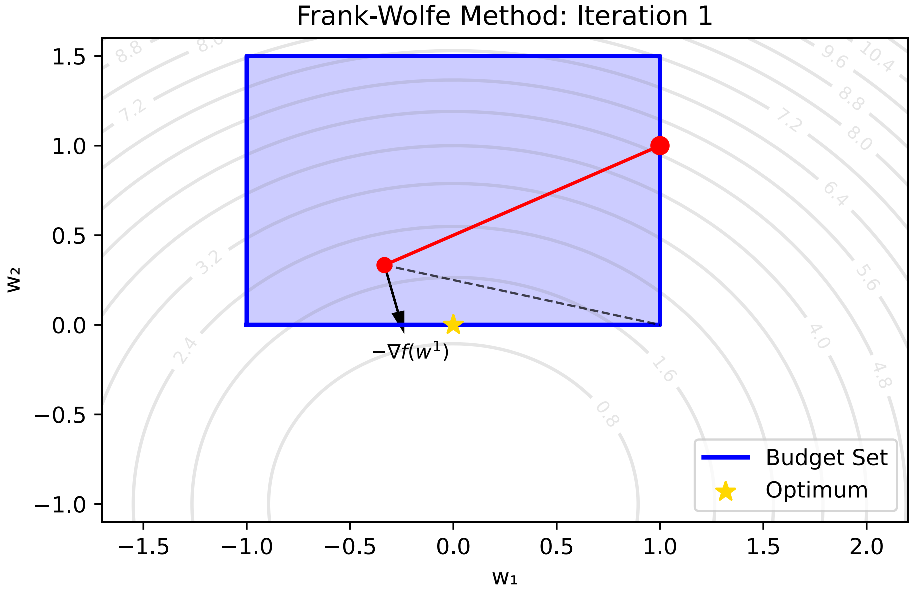
Численные эксперименты

Двумерный пример. Метод Франк-Вульфа

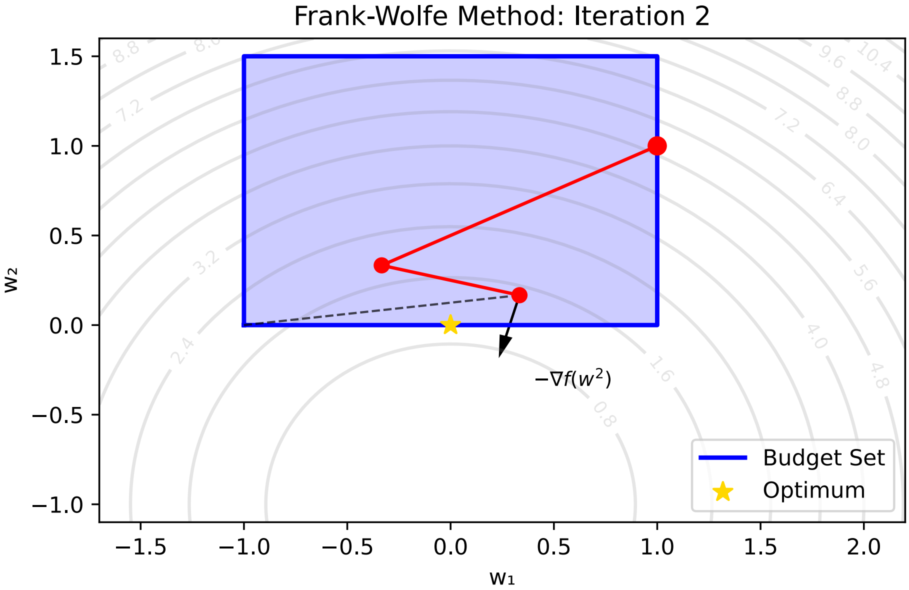
$lB \leq x \leq uB$



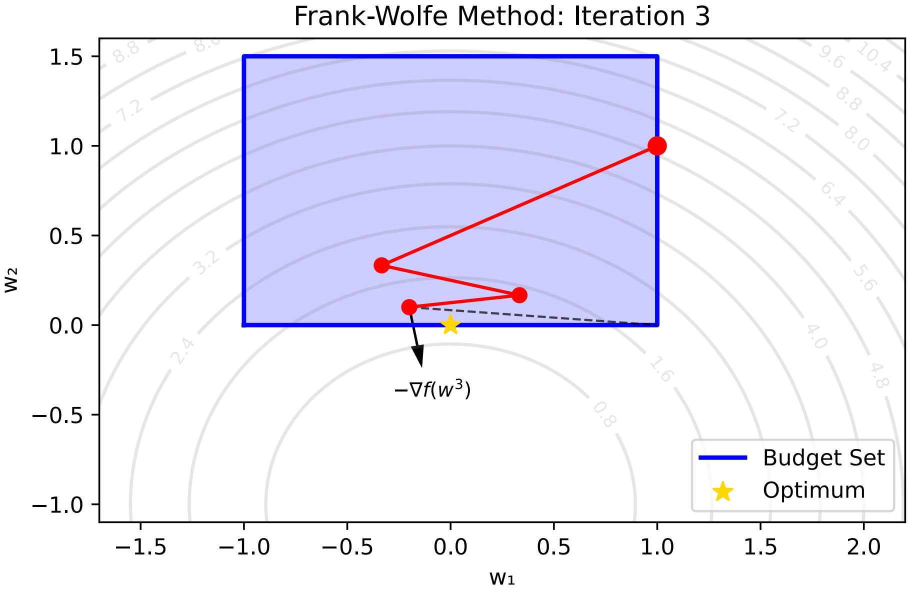
Двумерный пример. Метод Франк-Вульфа



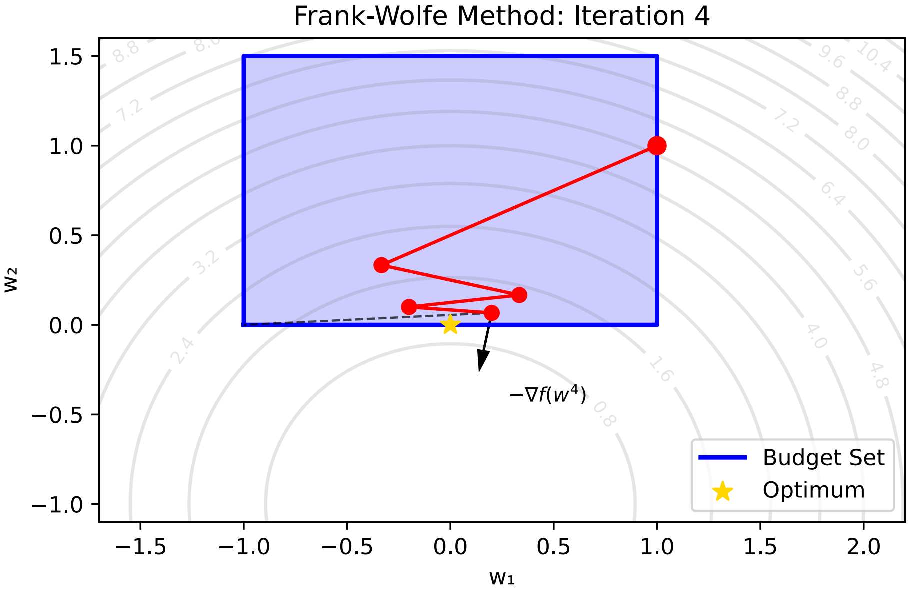
Двумерный пример. Метод Франк-Вульфа



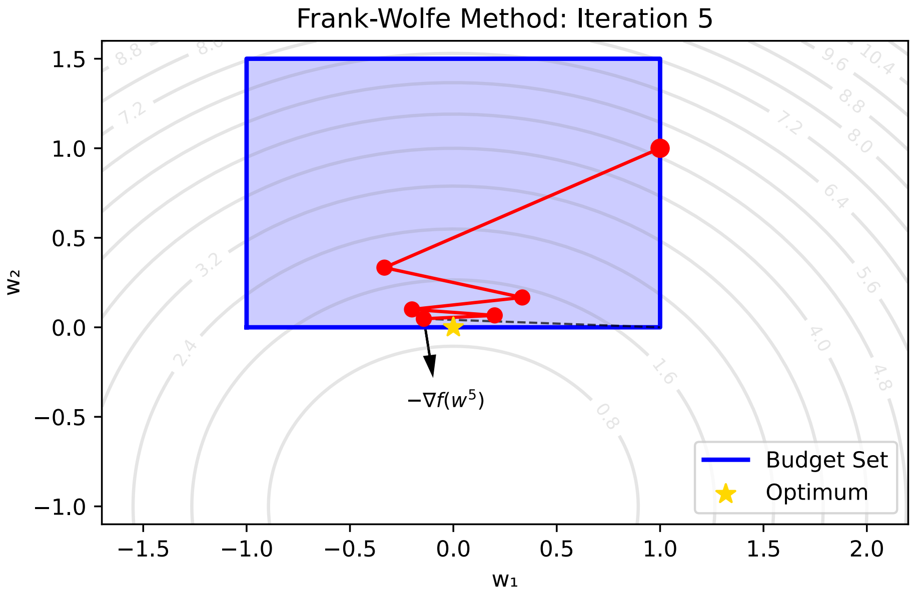
Двумерный пример. Метод Франк-Вульфа



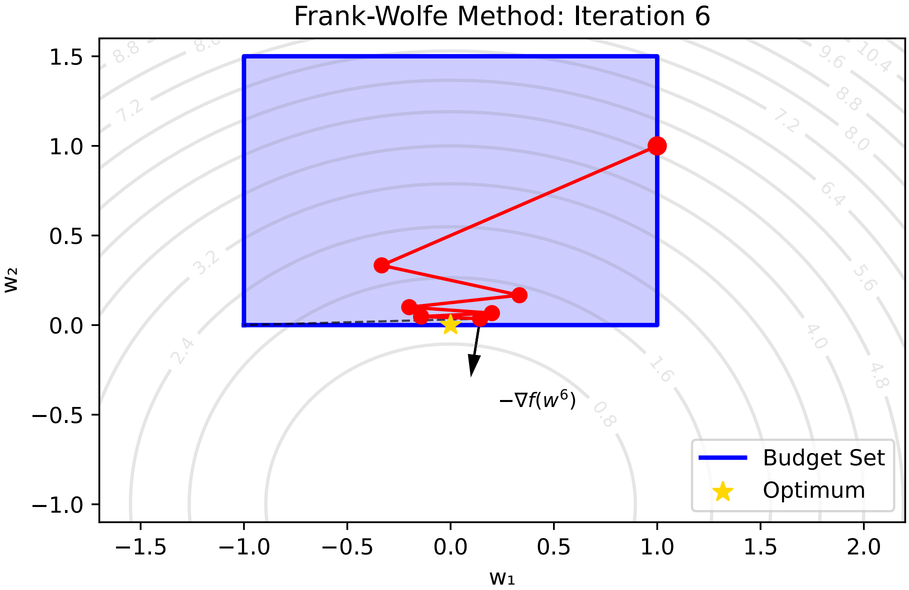
Двумерный пример. Метод Франк-Вульфа



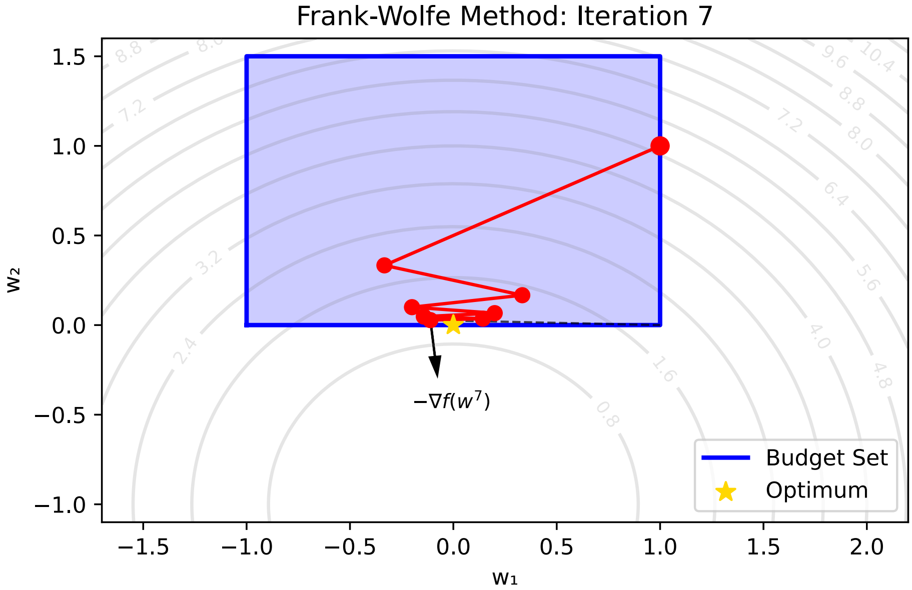
Двумерный пример. Метод Франк-Вульфа



Двумерный пример. Метод Франк-Вульфа



Двумерный пример. Метод Франк-Вульфа



Квадратичная функция. Покоординатные ограничения

$$
\min_{\substack{x \in \mathbb{R}^n \\ -\mathbf{1} \preceq x \preceq \mathbf{1}}} \frac{1}{2} x^\top A x - b^\top x,
$$

$$
A \in \mathbb{R}^{n \times n}, \quad \lambda(A) \in [\mu; L].
$$

Проекция тривиальна:

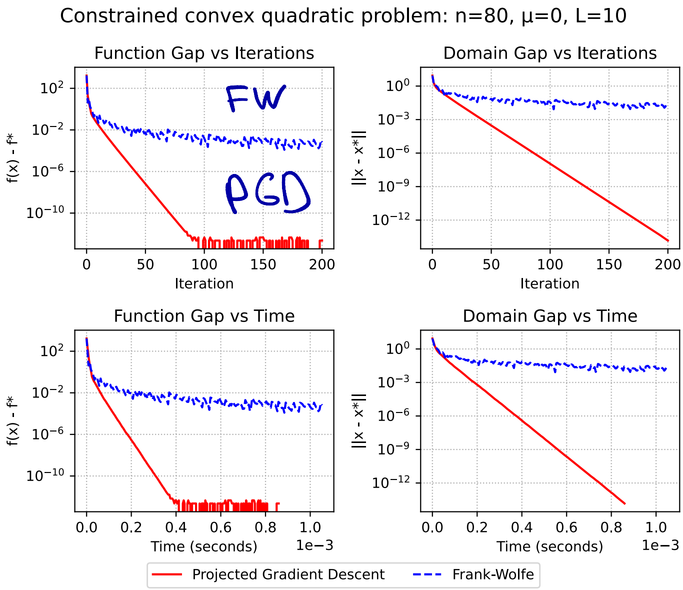
$$
\pi_S(x) = \text{clip}(x, -\mathbf{1}, \mathbf{1}).
$$

или

$$
\pi_S(x) = \max\left(-\mathbf{1}, \min(\mathbf{1}, x)\right).
$$

Оракул линейной минимизации (LMO) для заданного градиента g имеет вид $y = \underset{z \in S}{\text{argmin}} \langle g, z \rangle$.
Поскольку допустимое множество сепарабельно по координатам, решение вычисляется покоординатно:

$$
y_i = \begin{cases} -1, & \text{если } g_i > 0, \\ 1, & \text{если } g_i \le 0. \end{cases}
$$



Квадратичная функция. Покоординатные ограничения

$$\min_{\substack{x \in \mathbb{R}^n \\ -\mathbf{1} \preceq x \preceq \mathbf{1}}} \frac{1}{2} x^\top A x - b^\top x,$$

$$A \in \mathbb{R}^{n \times n}, \quad \lambda(A) \in [\mu; L].$$

Проекция тривиальна:

$$\pi_S(x) = \text{clip}(x, -\mathbf{1}, \mathbf{1}).$$

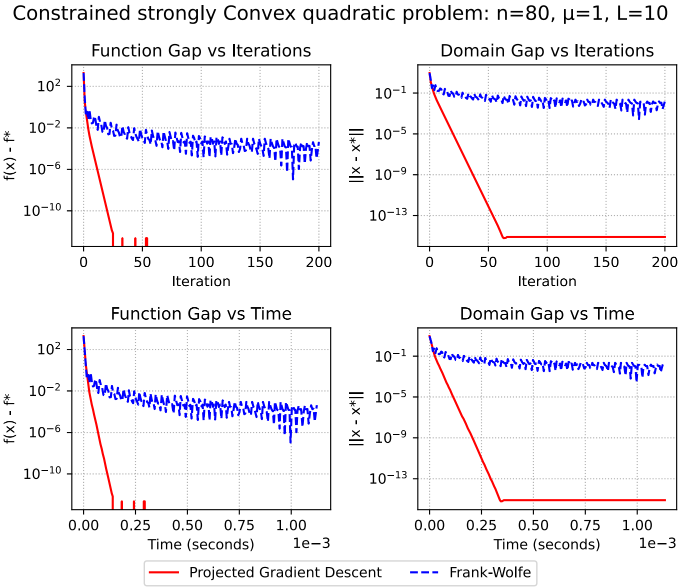
или

$$\pi_S(x) = \max\left(-\mathbf{1}, \min(\mathbf{1}, x)\right).$$

Оракул линейной минимизации (LMO) для заданного градиента g имеет вид $y = \underset{z \in S}{\text{argmin}} \langle g, z \rangle$.

Поскольку допустимое множество сепарабельно по координатам, решение вычисляется покоординатно:

$$y_i = \begin{cases} -1, & \text{если } g_i > 0, \\ 1, & \text{если } g_i \leq 0. \end{cases}$$



Квадратичная функция. Симплексные ограничения (удачная задача с диагональной матрицей)

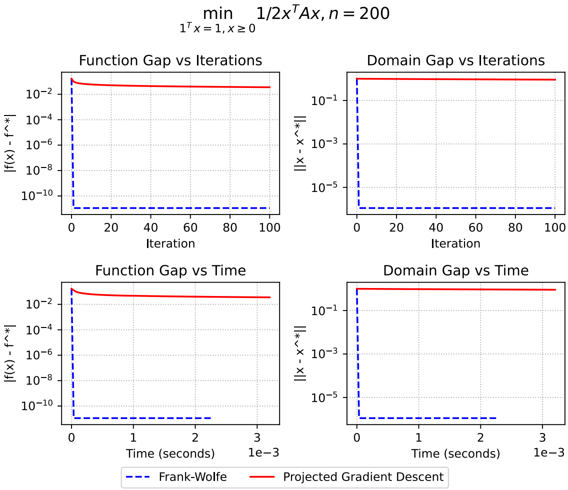
$$\min_{\substack{x \in \mathbb{R}^n \\ x \geq 0, \mathbf{1}^T x = 1}} \frac{1}{2} x^T A x,$$

$$A \in \mathbb{R}^{n \times n}, \quad \lambda(A) \in [0; 100].$$

Метод	Время шага, мс	Время LMO/проекции, мс
PGD	0.0069	0.0167
FW	0.0070	0.0066

Проекция на единичный симплекс $\pi_S(x)$ выполняется за $\mathcal{O}(n \log n)$ или за ожидаемое $\mathcal{O}(n)$ время. [4] LMO для заданного градиента g имеет вид $y = \underset{z \in S}{\operatorname{argmin}} \langle g, z \rangle$. Решение соответствует вершине симплекса:

$$y = e_j \quad \text{где} \quad j = \underset{i}{\operatorname{argmin}}\, g_i.$$



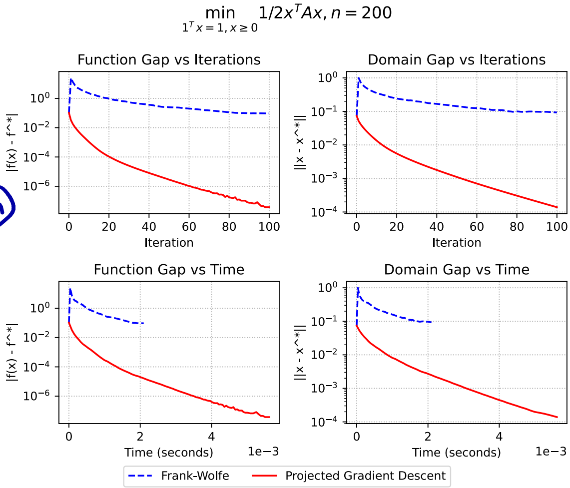
[4] Efficient Projections onto the ℓ_1-Ball for Learning in High Dimensions

Квадратичная функция. Симплексные ограничения

$$
\min_{\substack{x \in \mathbb{R}^n \\ x \geq 0, \mathbf{1}^T x = 1}} \frac{1}{2} x^T A x,
$$

$$
A \in \mathbb{R}^{n \times n}, \quad \lambda(A) \in [0; 100].
$$

Метод	Время шага, мс	Время LMO/проекции, мс	
PGD	0.0069	0.0420	$\mathcal{O}(n \log n)$
FW	0.0069	0.0066	$\mathcal{O}(n)$

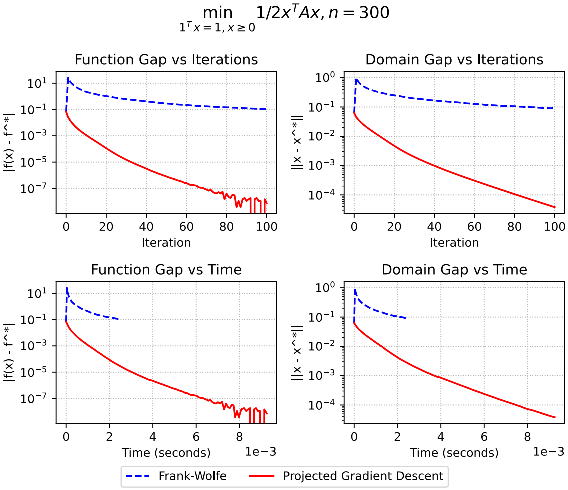


Квадратичная функция. Симплексные ограничения

$$
\min_{\substack{x \in \mathbb{R}^n \\ x \geq 0, \mathbf{1}^T x = 1}} \frac{1}{2} x^T A x,
$$

$$
A \in \mathbb{R}^{n \times n}, \quad \lambda(A) \in [0; 100].
$$

Метод	Время шага, мс	Время LMO/проекции, мс
PGD	0.0068	0.0761
FW	0.0069	0.0070

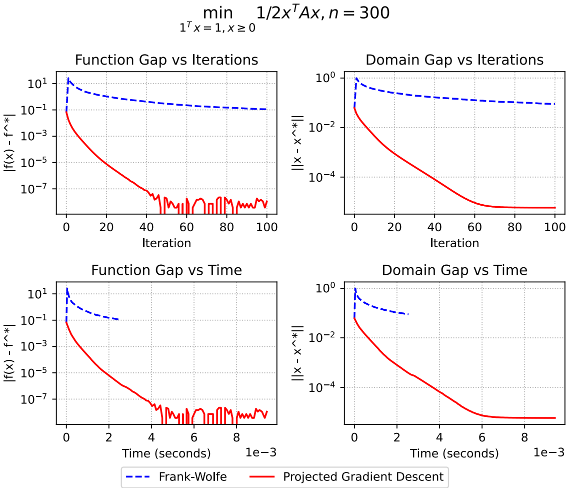


Квадратичная функция. Симплексные ограничения

$$\min_{\substack{x \in \mathbb{R}^n \\ x \geq 0, \mathbf{1}^T x = 1}} \frac{1}{2} x^T A x,$$

$$A \in \mathbb{R}^{n \times n}, \quad \lambda(A) \in [1; 100].$$

Метод	Время шага, мс	Время LMO/проекции, мс
PGD	0.0068	0.0752
FW	0.0067	0.0068



PGD и метод Франк-Вульфа

Ключевое различие между PGD и FW состоит в том, что PGD требует проекции, тогда как FW — лишь оракул линейной минимизации (LMO).

В недавней статье авторы представили следующую сравнительную таблицу сложностей линейной минимизации и проекций на некоторые выпуклые множества с точностью до аддитивной ошибки ϵ в евклидовой норме.

Множество	**Линейная минимизация**	**Проекция**
n-мерный ℓ_p-шар, $p \neq 1, 2, \infty$	$\mathcal{O}(n)$	$\tilde{\mathcal{O}}\left(\frac{n}{\epsilon^2}\right)$
Шар ядерной нормы для $n \times m$ матриц	$\mathcal{O}\left(\nu \ln(m+n) \frac{\sqrt{\sigma_1}}{\sqrt{\epsilon}}\right)$	$\mathcal{O}(m\, n \min\{m, n\})$
Потоковый многогранник на графе с m вершинами и n рёбрами (ограничение пропускной способности на рёбрах)	$\mathcal{O}\left((n \log m)(n + m \log m)\right)$	$\tilde{\mathcal{O}}\left(\frac{n}{\epsilon^2}\right)$ или $\mathcal{O}(n^4 \log n)$
Многогранник Биркгофа ($n \times n$ дважды стохастические матрицы)	$\mathcal{O}(n^3)$	$\tilde{\mathcal{O}}\left(\frac{n^2}{\epsilon^2}\right)$

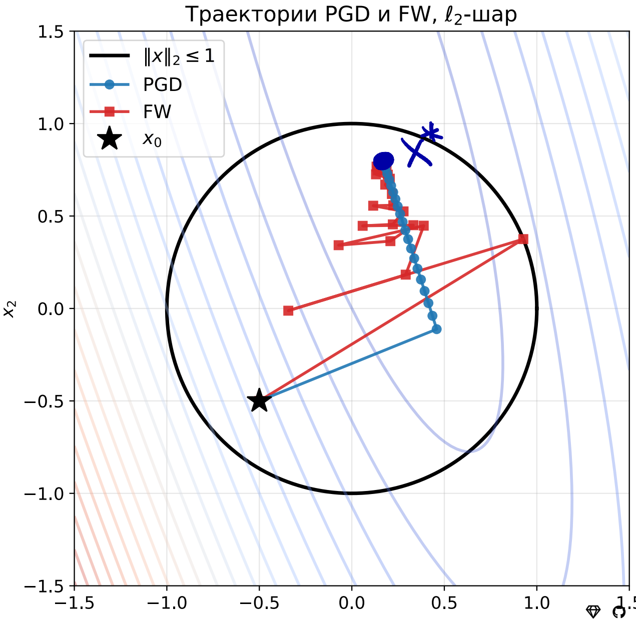
Когда ϵ отсутствует, аддитивной ошибки нет. Обозначение $\tilde{\mathcal{O}}$ скрывает полилогарифмические множители по размерности и полиномиальные множители в константах, связанных с расстоянием до оптимума. Для шара ядерной нормы (спектраэдра) ν обозначает число ненулевых элементов, а σ_1 — наибольшее сингулярное число проецируемой матрицы.

Дополнительные численные эксперименты

Траектории PGD и FW на ℓ_2-шаре

$$\min_{x \in \mathbb{R}^2} \frac{1}{2} x^\top A x - b^\top x, \quad \|x\|_2 \leq 1$$

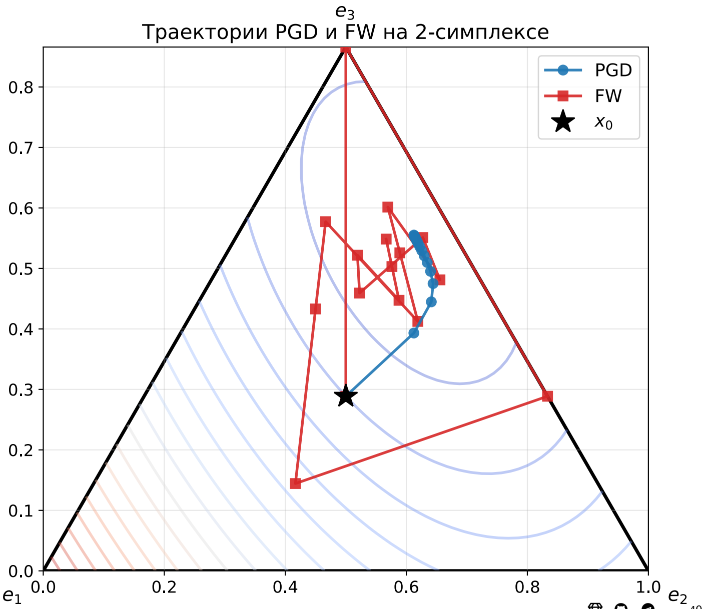
PGD: проекция $\to$ кратчайший путь к границе, затем движение вдоль неё к оптимуму.
FW: каждый шаг — выпуклая комбинация текущей точки и **антиградиентной точки на границе**. Характерный зигзаг вдоль множества.
Оба метода стартуют из $x_0 = (-0.5, -0.5)$.



Траектории PGD и FW на 2-симплексе

$$\min_{\substack{x \in \mathbb{R}^3 \\ x \geq 0,\, \mathbf{1}^\top x = 1}} \quad \frac{1}{2} x^\top A x$$

FW на каждом шаге выбирает одну из **вершин** симплекса (e_1, e_2, e_3), движется к ней $\longrightarrow$ траектория проходит через рёбра. PGD: проекция на симплекс допускает движение в любом направлении $\longrightarrow$ более прямой путь к оптимуму.

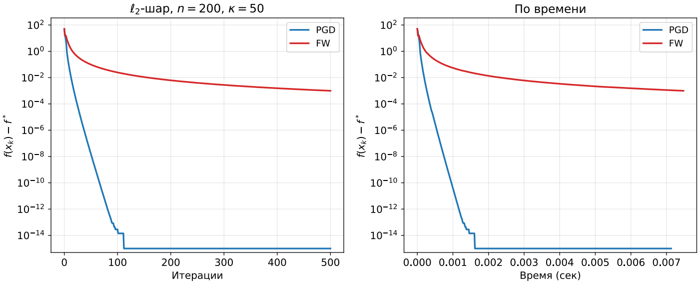


ℓ_2-шар. Проекция и LMO одинаково дёшевы

$$\min_{\|x\|_2 \le 2} \frac{1}{2} x^\top A x - b^\top x$$

- $n = 200,\ \kappa = 50$

сильно вып.



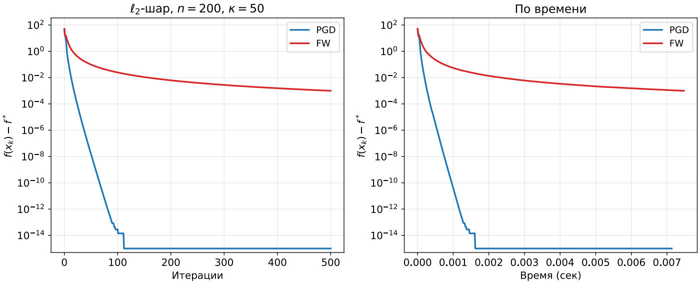
Когда проекция и LMO стоят одинаково, PGD выигрывает за счёт **линейной скорости** при сильной выпуклости.

ℓ_2-шар. Проекция и LMO одинаково дёшевы

$$\min_{\|x\|_2 \le 2} \frac{1}{2} x^\top A x - b^\top x$$

- $n = 200$, $\kappa = 50$
- Проекция: $\pi_S(x) = x \cdot \min\left(1, \frac{R}{\|x\|}\right)$ — $\mathcal{O}(n)$

Когда проекция и LMO стоят одинаково, PGD выигрывает за счёт **линейной скорости** при сильной выпуклости.

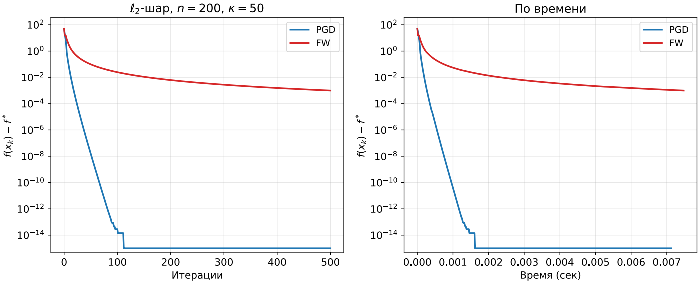


ℓ_2-шар. Проекция и LMO одинаково дёшевы

$$\min_{\|x\|_2 \le 2} \frac{1}{2} x^\top A x - b^\top x$$

- $n = 200,\ \kappa = 50$
- Проекция: $\pi_S(x) = x \cdot \min\left(1, \frac{R}{\|x\|}\right)$ — $\mathcal{O}(n)$
- $\boxed{\text{LMO: } y = -R\frac{g}{\|g\|} \text{ — } \mathcal{O}(n)}$

Когда проекция и LMO стоят одинаково, PGD выигрывает за счёт **линейной скорости** при сильной выпуклости.

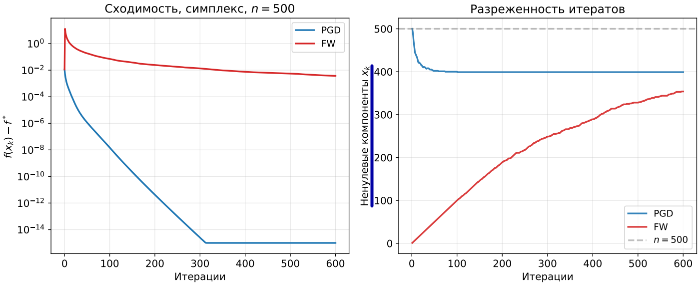


Разреженность итераций FW на симплексе

$$\boxed{\min_{\substack{x\geq 0\\ 1^\top x=1}} \frac{1}{2}x^\top Ax, \quad n = 500}$$

Ключевое свойство FW: итерация x_k есть **выпуклая комбинация** k **вершин** $\Rightarrow$ не более k ненулевых компонент.

- PGD: после проекции число ненулевых компонент скачком достигает ~ 400



Это важно, когда нужно получить разреженное решение.

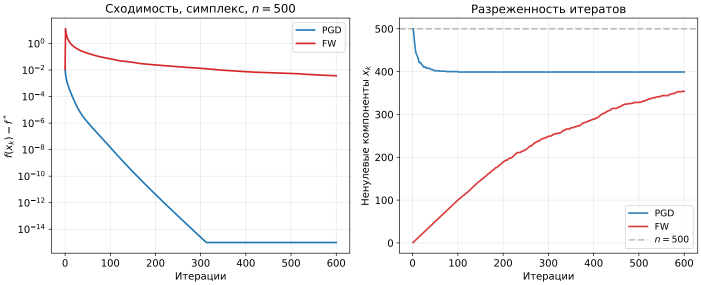
Разреженность итераций FW на симплексе

$$\min_{\substack{x \geq 0 \\ \mathbf{1}^\top x = 1}} \frac{1}{2} x^\top A x, \quad n = 500$$

Ключевое свойство FW: итерация x_k есть **выпуклая комбинация** k **вершин** $\Rightarrow$ не более k ненулевых компонент.

- PGD: после проекции число ненулевых компонент скачком достигает ~ 400
- FW: число ненулевых растёт **линейно** с итерациями

Это важно, когда нужно получить разреженное решение.

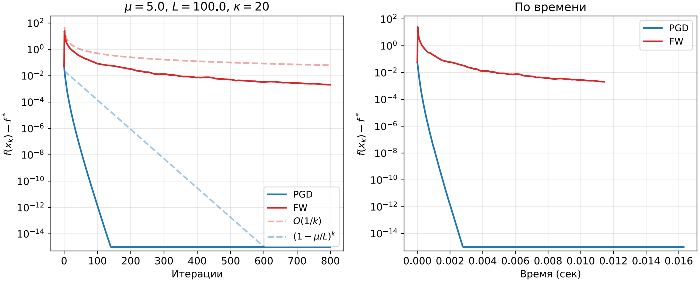


Сильная выпуклость: PGD ускоряется, FW — нет

$$\min_{\substack{x \geq 0 \\ \mathbf{1}^\top x = 1}} \frac{1}{2} x^\top A x$$

- $n = 200,\ \mu = 5,\ L = 100,\ \kappa = 20$

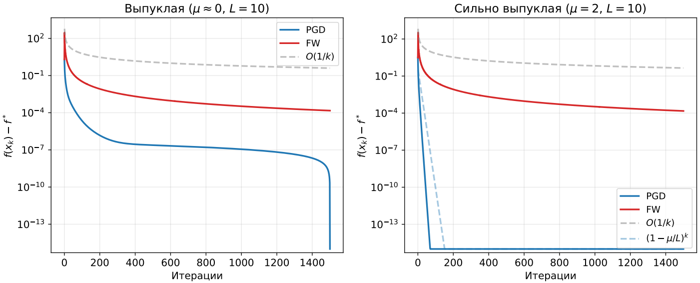
PGD с шагом $\frac{1}{L}$: **линейная сходимость** $(1 - \mu/L)^k$, ведь проекция на выпуклое множество — нерастягивающий оператор.
FW: остаётся $\mathcal{O}(1/k)$ — сильная выпуклость **не помогает** со стандартным шагом $\gamma_k = \frac{2}{k+2}$.



Скорость сходимости: выпуклая vs сильно выпуклая задача

$$\min_{-1 \preceq x \preceq 1} \frac{1}{2} x^\top A x, \quad n = 100$$

- Левая панель: $\mu \approx 0$ — оба метода $\mathcal{O}(1/k)$, но PGD быстрее по константе



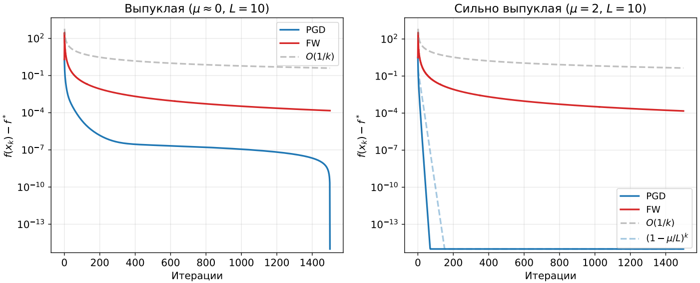
Главный вывод: сильная выпуклость кардинально меняет картину в пользу PGD.

Скорость сходимости: выпуклая vs сильно выпуклая задача

$$\min_{-1 \preceq x \preceq 1} \frac{1}{2} x^\top A x, \quad n = 100$$

- Левая панель: $\mu \approx 0$ — оба метода $\mathcal{O}(1/k)$, но PGD быстрее по константе
- Правая панель: $\mu = 2$ — PGD получает **экспоненциальную** скорость, FW остаётся сублинейным

Главный вывод: сильная выпуклость кардинально меняет картину в пользу PGD.



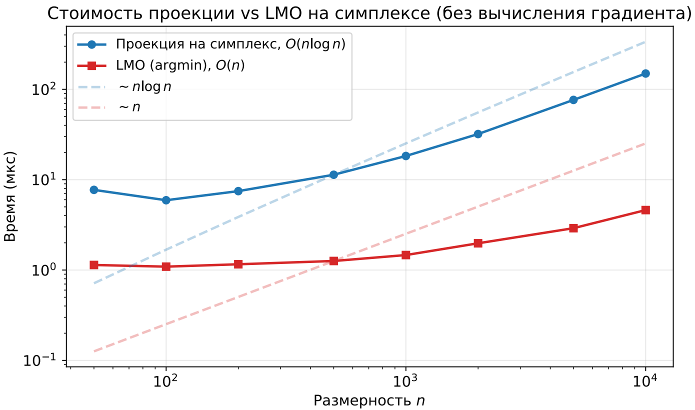
Стоимость итерации: проекция vs LMO на симплексе

Измеряем **только** стоимость проекции и LMO, без вычисления градиента.

- Проекция на симплекс: сортировка + O(n) — всего $\mathcal{O}(n \log n)$

При $n = 10000$ проекция дороже LMO **на порядок**.

В задачах, где градиент считается быстро (разреженная матрица), разница в стоимости шага между PGD и FW определяется именно проекцией/LMO.



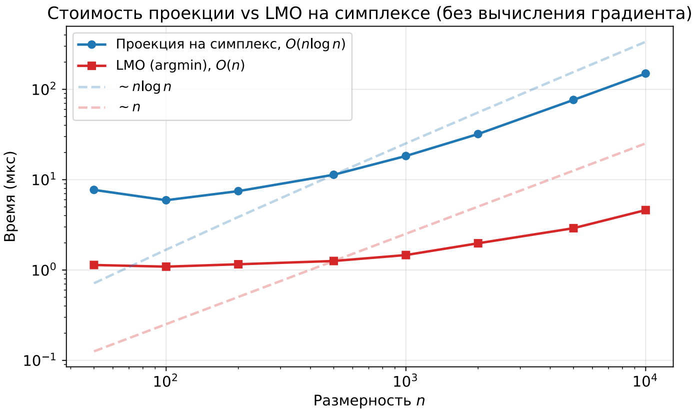
Стоимость итерации: проекция vs LMO на симплексе

Измеряем **только** стоимость проекции и LMO, без вычисления градиента.

- Проекция на симплекс: сортировка + O(n) — всего $\mathcal{O}(n \log n)$
- LMO (argmin): один проход — $\mathcal{O}(n)$

При $n = 10000$ проекция дороже LMO **на порядок**.

В задачах, где градиент считается быстро (разреженная матрица), разница в стоимости шага между PGD и FW определяется именно проекцией/LMO.



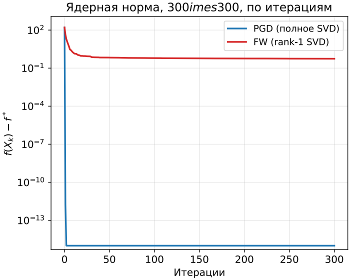
Ядерная норма: когда FW — единственный вариант

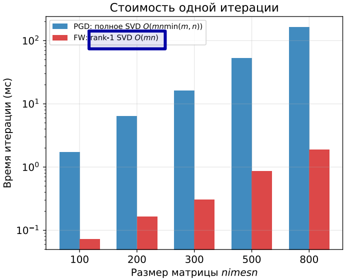
$$\|X\|_* \le R$$

$$\min_{\|X\|_* \le R} \frac{1}{2}\|X - B\|_F^2, \quad X \in \mathbb{R}^{n \times n}$$

- PGD: полное SVD $\mathcal{O}(n^3)$ **на каждой итерации**

По итерациям PGD выигрывает. Но стоимость одной итерации PGD растёт **кубически**: при $n = 800$ полное SVD в **75 раз** дороже rank-1 SVD.
Для матриц размера $10^4 \times 10^4$ и больше полное SVD вычислительно неподъёмно — FW становится единственным разумным методом.





$$\|X\|_* = \sum_{i=1} \sigma_i(X)$$

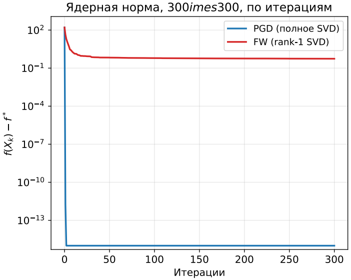
Ядерная норма: когда FW — единственный вариант

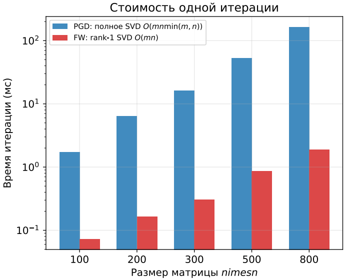
$$\min_{\|X\|_* \le R} \frac{1}{2}\|X - B\|_F^2, \quad X \in \mathbb{R}^{n\times n}$$

- PGD: полное SVD $\mathcal{O}(n^3)$ **на каждой итерации**
- FW: rank-1 SVD (power iteration) $\mathcal{O}(n^2)$

По итерациям PGD выигрывает. Но стоимость одной итерации PGD растёт **кубически**: при $n = 800$ полное SVD в **75 раз** дороже rank-1 SVD.

Для матриц размера $10^4 \times 10^4$ и больше полное SVD вычислительно неподъёмно — FW становится единственным разумным методом.





Ядерная норма: определение и свойства

Ядерная норма (trace norm)

Ядерная норма матрицы $X \in \mathbb{R}^{m \times n}$ определяется как сумма её сингулярных чисел:

$$\boxed{\|X\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(X) = \mathrm{tr}\left(\sqrt{X^\top X}\right)}$$

Ядерная норма: определение и свойства

Ядерная норма (trace norm)

Ядерная норма матрицы $X \in \mathbb{R}^{m\times n}$ определяется как сумма её сингулярных чисел:

$$
\|X\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(X) = \mathsf{tr}\left(\sqrt{X^\top X}\right)
$$

- Ядерная норма — выпуклая оболочка ранга матрицы на единичном спектральном шаре: $\|X\|_* = \mathsf{conv}\left(\mathsf{rank}(X)\right)$ при $\|X\|_{\mathsf{op}} \leq 1$.

Ядерная норма: определение и свойства

Ядерная норма (trace norm)

Ядерная норма матрицы $X \in \mathbb{R}^{m \times n}$ определяется как сумма её сингулярных чисел:

$$\|X\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(X) = \mathsf{tr}\left(\sqrt{X^\top X}\right)$$

- Ядерная норма — выпуклая оболочка ранга матрицы на единичном спектральном шаре: $\|X\|_* = \mathsf{conv}\left(\mathsf{rank}(X)\right)$ при $\|X\|_{\mathsf{op}} \leq 1$.
- Поэтому ограничение $\|X\|_* \leq R$ — стандартный выпуклый релаксант для ограничения на ранг.

Ядерная норма: определение и свойства

Ядерная норма (trace norm)

Ядерная норма матрицы $X \in \mathbb{R}^{m\times n}$ определяется как сумма её сингулярных чисел:

$$
\|X\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(X) = \mathrm{tr}\left(\sqrt{X^\top X}\right)
$$

- Ядерная норма — выпуклая оболочка ранга матрицы на единичном спектральном шаре: $\|X\|_* = \mathrm{conv}\left(\mathrm{rank}(X)\right)$ при $\|X\|_{\mathrm{op}} \leq 1$.
- Поэтому ограничение $\|X\|_* \leq R$ — стандартный выпуклый релаксант для ограничения на ранг.
- Шар ядерной нормы $\mathcal{B}_* = \{X \in \mathbb{R}^{m\times n} \mid \|X\|_* \leq R\}$ — выпуклое компактное множество.

Ядерная норма: определение и свойства

Ядерная норма (trace norm)

Ядерная норма матрицы $X \in \mathbb{R}^{m\times n}$ определяется как сумма её сингулярных чисел:

$$
\|X\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(X) = \text{tr}\left(\sqrt{X^\top X}\right)
$$

- Ядерная норма — выпуклая оболочка ранга матрицы на единичном спектральном шаре: $\|X\|_* = \text{conv}\left(\text{rank}(X)\right)$ при $\|X\|_{\text{op}} \leq 1$.
- Поэтому ограничение $\|X\|_* \leq R$ — стандартный выпуклый релаксант для ограничения на ранг.
- Шар ядерной нормы $\mathcal{B}_* = \{X \in \mathbb{R}^{m\times n} \mid \|X\|_* \leq R\}$ — выпуклое компактное множество.
- Крайние точки $\mathcal{B}_*$ — матрицы ранга 1 вида $R \cdot uv^\top$, где $\|u\|_2 = \|v\|_2 = 1$.

Ядерная норма: определение и свойства

Ядерная норма (trace norm)

Ядерная норма матрицы $X \in \mathbb{R}^{m \times n}$ определяется как сумма её сингулярных чисел:

$$\|X\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(X) = \mathsf{tr}\left(\sqrt{X^\top X}\right)$$

- Ядерная норма — выпуклая оболочка ранга матрицы на единичном спектральном шаре: $\|X\|_* = \mathsf{conv}\left(\mathsf{rank}(X)\right)$ при $\|X\|_{\mathsf{op}} \leq 1$.
- Поэтому ограничение $\|X\|_* \leq R$ — стандартный выпуклый релаксант для ограничения на ранг.
- Шар ядерной нормы $\mathcal{B}_* = \{X \in \mathbb{R}^{m \times n} \mid \|X\|_* \leq R\}$ — выпуклое компактное множество.
- Крайние точки $\mathcal{B}_*$ — матрицы ранга 1 вида $R \cdot uv^\top$, где $\|u\|_2 = \|v\|_2 = 1$.

Ядерная норма: определение и свойства

Ядерная норма (trace norm)

Ядерная норма матрицы $X \in \mathbb{R}^{m \times n}$ определяется как сумма её сингулярных чисел:

$$\|X\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(X) = \text{tr}\left(\sqrt{X^\top X}\right)$$

- Ядерная норма — выпуклая оболочка ранга матрицы на единичном спектральном шаре: $\|X\|_* = \text{conv}\left(\text{rank}(X)\right)$ при $\|X\|_{\text{op}} \leq 1$.
- Поэтому ограничение $\|X\|_* \leq R$ — стандартный выпуклый релаксант для ограничения на ранг.
- Шар ядерной нормы $\boxed{\mathcal{B}_* = \{X \in \mathbb{R}^{m \times n} \mid \|X\|_* \leq R\}}$ — выпуклое компактное множество.
- Крайние точки $\mathcal{B}_*$ — матрицы ранга 1 вида $R \cdot uv^\top$, где $\|u\|_2 = \|v\|_2 = 1$.

Приложения: matrix completion (рекомендательные системы), robust PCA, low-rank matrix recovery, сжатие нейронных сетей.

Итерации PGD и FW на шаре ядерной нормы

Задача: $\min_{\|X\|_* \leq R} f(X)$, где $X \in \mathbb{R}^{m \times n}$.

Итерации PGD и FW на шаре ядерной нормы

Задача: $\min_{\|X\|_* \leq R} f(X)$, где $X \in \mathbb{R}^{m \times n}$.

Итерация PGD

$$
X_{k+1} = \text{proj}_{\mathcal{B}_*}(X_k - \alpha \nabla f(X_k))
$$

1. Вычислить $Y = X_k - \alpha \nabla f(X_k)$

Итерации PGD и FW на шаре ядерной нормы

Задача: $\min_{\|X\|_* \le R} f(X)$, где $X \in \mathbb{R}^{m \times n}$.

Итерация PGD

$$X_{k+1} = \text{proj}_{\mathcal{B}_*}(X_k - \alpha \nabla f(X_k))$$

1. Вычислить $Y = X_k - \alpha \nabla f(X_k)$
2. **Полное SVD:** $Y = U \Sigma V^\top$ — стоимость $\mathcal{O}(mn \cdot \min(m, n))$

Итерации PGD и FW на шаре ядерной нормы

Задача: $\min_{\|X\|_* \leq R} f(X)$, где $X \in \mathbb{R}^{m \times n}$.

Итерация PGD

$$X_{k+1} = \text{proj}_{\mathcal{B}_*}(X_k - \alpha \nabla f(X_k))$$

1. Вычислить $Y = X_k - \alpha \nabla f(X_k)$
2. **Полное SVD:** $Y = U \Sigma V^\top$ — стоимость $\mathcal{O}(mn \cdot \min(m, n))$
3. Мягкое пороговое отсечение (soft thresholding) сингулярных чисел для проекции на $\mathcal{B}_*$:

$$\text{proj}_{\mathcal{B}_*}(Y) = U \cdot \text{diag}(\max(\sigma_i - \lambda, 0)) \cdot V^\top$$

где $\lambda \geq 0$ выбирается так, чтобы $\sum_i \max(\sigma_i - \lambda, 0) = R$.

Итерации PGD и FW на шаре ядерной нормы

Задача: $\min_{\|X\|_* \leq R} f(X)$, где $X \in \mathbb{R}^{m \times n}$.

Итерация PGD

$$X_{k+1} = \text{proj}_{\mathcal{B}_*}(X_k - \alpha \nabla f(X_k))$$

1. Вычислить $Y = X_k - \alpha \nabla f(X_k)$
2. **Полное SVD:** $Y = U \Sigma V^\top$ — стоимость $\mathcal{O}(mn \cdot \min(m, n))$
3. Мягкое пороговое отсечение (soft thresholding) сингулярных чисел для проекции на $\mathcal{B}_*$:

$$\boxed{\text{proj}_{\mathcal{B}_*}(Y) = U \cdot \text{diag}(\max(\sigma_i - \lambda, 0)) \cdot V^\top}$$

где $\lambda \geq 0$ выбирается так, чтобы $\sum_i \max(\sigma_i - \lambda, 0) = R$.

Итерации PGD и FW на шаре ядерной нормы

Задача: $\min_{\|X\|_*\le R} f(X)$, где $X \in \mathbb{R}^{m\times n}$.

Итерация PGD

$$
X_{k+1} = \text{proj}_{\mathcal{B}_*}(X_k - \alpha \nabla f(X_k))
$$

1. Вычислить $Y = X_k - \alpha \nabla f(X_k)$
2. **Полное SVD:** $Y = U\Sigma V^\top$ — стоимость $\mathcal{O}(mn \cdot \min(m,n))$
3. Мягкое пороговое отсечение (soft thresholding) сингулярных чисел для проекции на $\mathcal{B}_*$:

$$
\text{proj}_{\mathcal{B}_*}(Y) = U \cdot \text{diag}(\max(\sigma_i - \lambda, 0)) \cdot V^\top
$$

где $\lambda \ge 0$ выбирается так, чтобы $\sum_i \max(\sigma_i - \lambda, 0) = R$.

Итерация FW

$$
S_k = \arg\min_{\|S\|_*\le R} \langle \nabla f(X_k), S\rangle
$$

$$
X_{k+1} = (1-\gamma_k)X_k + \gamma_k S_k
$$

1. Вычислить $G = \nabla f(X_k)$

Бонус: X_k имеет ранг $\le k$ после k итераций.

Итерации PGD и FW на шаре ядерной нормы

Задача: $\min_{\|X\|_* \leq R} f(X)$, где $X \in \mathbb{R}^{m \times n}$.

Итерация PGD

$$X_{k+1} = \text{proj}_{\mathcal{B}_*}(X_k - \alpha \nabla f(X_k))$$

1. Вычислить $Y = X_k - \alpha \nabla f(X_k)$
2. **Полное SVD:** $Y = U\Sigma V^\top$ — стоимость $\mathcal{O}(mn \cdot \min(m,n))$
3. Мягкое пороговое отсечение (soft thresholding) сингулярных чисел для проекции на $\mathcal{B}_*$:

$$\text{proj}_{\mathcal{B}_*}(Y) = U \cdot \text{diag}(\max(\sigma_i - \lambda, 0)) \cdot V^\top$$

где $\lambda \geq 0$ выбирается так, чтобы $\sum_i \max(\sigma_i - \lambda, 0) = R$.

Итерация FW

$$S_k = \arg\min_{\|S\|_* \leq R} \langle \nabla f(X_k), S \rangle$$

$$X_{k+1} = (1 - \gamma_k) X_k + \gamma_k S_k$$

1. Вычислить $G = \nabla f(X_k)$
2. **Rank-1 SVD** (степенная итерация): найти ведущие сингулярные векторы u_1, v_1 матрицы $-G$ — стоимость $\mathcal{O}(\text{nnz}(G))$

Бонус: X_k имеет ранг $\leq k$ после k итераций.

Итерации PGD и FW на шаре ядерной нормы

Задача: $\min_{\|X\|_* \le R} f(X)$, где $X \in \mathbb{R}^{m \times n}$.

Итерация PGD

$$X_{k+1} = \text{proj}_{\mathcal{B}_*}(X_k - \alpha \nabla f(X_k))$$

1. Вычислить $Y = X_k - \alpha \nabla f(X_k)$
2. **Полное SVD:** $Y = U \Sigma V^\top$ — стоимость $\mathcal{O}(mn \cdot \min(m, n))$
3. Мягкое пороговое отсечение (soft thresholding) сингулярных чисел для проекции на $\mathcal{B}_*$:

$$\text{proj}_{\mathcal{B}_*}(Y) = U \cdot \text{diag}(\max(\sigma_i - \lambda, 0)) \cdot V^\top$$

где $\lambda \ge 0$ выбирается так, чтобы $\sum_i \max(\sigma_i - \lambda, 0) = R$.

ВАЖНО

Итерация FW

$$S_k = \arg \min_{\|S\|_* \le R} \langle \nabla f(X_k), S \rangle$$

$$X_{k+1} = (1 - \gamma_k) X_k + \gamma_k S_k$$

1. Вычислить $G = \nabla f(X_k)$
2. **Rank-1 SVD** (степенная итерация): найти ведущие сингулярные векторы u_1, v_1 матрицы $-G$ — стоимость $\mathcal{O}(\text{nnz}(G))$
3. $S_k = R \cdot u_1 v_1^\top$ — ранг-1 матрица

Бонус: X_k имеет ранг $\le k$ после k итераций.

Итерации PGD и FW на шаре ядерной нормы

Задача: $\min_{\|X\|_* \le R} f(X)$, где $X \in \mathbb{R}^{m\times n}$.

Итерация PGD

$$X_{k+1} = \text{proj}_{\mathcal{B}_*}(X_k - \alpha \nabla f(X_k))$$

1. Вычислить $Y = X_k - \alpha \nabla f(X_k)$
2. **Полное SVD:** $Y = U\Sigma V^\top$ — стоимость $\mathcal{O}(mn \cdot \min(m,n))$
3. Мягкое пороговое отсечение (soft thresholding) сингулярных чисел для проекции на $\mathcal{B}_*$:

$$\text{proj}_{\mathcal{B}_*}(Y) = U \cdot \text{diag}(\max(\sigma_i - \lambda, 0)) \cdot V^\top$$

где $\lambda \ge 0$ выбирается так, чтобы $\sum_i \max(\sigma_i - \lambda, 0) = R$.

Итерация FW

$$S_k = \arg\min_{\|S\|_* \le R} \langle \nabla f(X_k), S\rangle$$

$$X_{k+1} = (1-\gamma_k)X_k + \gamma_k S_k$$

1. Вычислить $G = \nabla f(X_k)$
2. **Rank-1 SVD** (степенная итерация): найти ведущие сингулярные векторы u_1, v_1 матрицы $-G$ — стоимость $\mathcal{O}(\text{nnz}(G))$
3. $S_k = R \cdot u_1 v_1^\top$ — ранг-1 матрица
4. $X_{k+1} = (1-\gamma_k)X_k + \gamma_k S_k$

Бонус: X_k имеет ранг $\le k$ после k итераций.

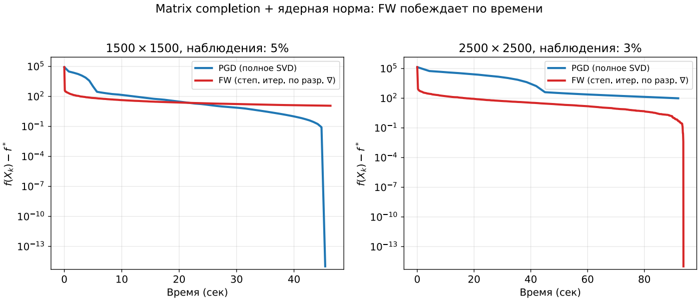
FW побеждает: matrix completion + ядерная норма

Задача: matrix completion с ограничением $\|X\|_* \leq R$:

$$\min_{\|X\|_* \leq R} \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2$$

Ключ: градиент ∇f **разрежен** (ненулевой только на Ω).

- Степенная итерация для rank-1 SVD на разреженном ∇f: $\mathcal{O}(|\Omega|)$



FW компенсирует медленную сходимость $O(1/k)$ огромным числом дешёвых итераций и **побеждает PGD по времени**.

FW побеждает: matrix completion + ядерная норма

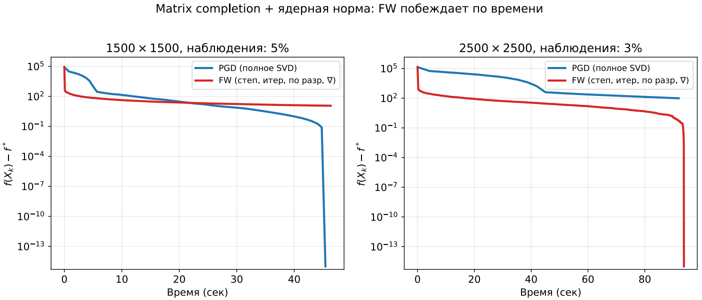
Задача: matrix completion с ограничением $\|X\|_* \leq R$:

$$\min_{\|X\|_* \leq R} \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2$$

Ключ: градиент ∇f **разрежен** (ненулевой только на Ω).

- Степенная итерация для rank-1 SVD на разреженном ∇f: $\mathcal{O}(|\Omega|)$
- Полное SVD плотной матрицы Y: $\mathcal{O}(n^3)$

FW компенсирует медленную сходимость $O(1/k)$ огромным числом дешёвых итераций и **побеждает PGD по времени**.



FW побеждает: matrix completion + ядерная норма

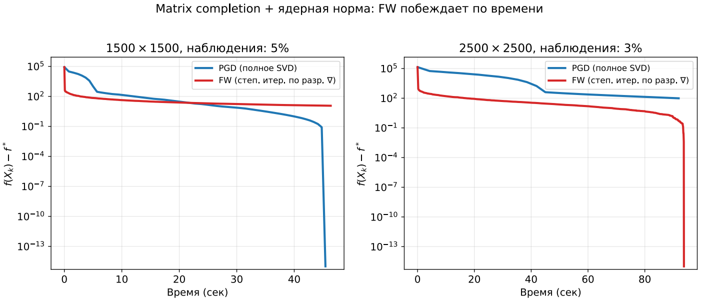
Задача: matrix completion с ограничением $\|X\|_* \le R$:

$$\min_{\|X\|_* \le R} \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2$$

Ключ: градиент ∇f **разрежен** (ненулевой только на Ω).

- Степенная итерация для rank-1 SVD на разреженном ∇f: $\mathcal{O}(|\Omega|)$
- Полное SVD плотной матрицы Y: $\mathcal{O}(n^3)$
- При $n = 2500$, $|\Omega| = 3\% \cdot n^2$: **отношение** $\approx 900\times$

FW компенсирует медленную сходимость $O(1/k)$ огромным числом дешёвых итераций и **побеждает PGD по времени**.



Matrix completion: почему ядерная норма?

Задача восстановления матрицы. Дана матрица $B \in \mathbb{R}^{m \times n}$, наблюдаемая лишь на подмножестве индексов $\Omega \subset [m] \times [n]$. Нужно восстановить всю матрицу, предполагая, что она имеет малый ранг.

Matrix completion: почему ядерная норма?

Задача восстановления матрицы. Дана матрица $B \in \mathbb{R}^{m\times n}$, наблюдаемая лишь на подмножестве индексов $\Omega \subset [m] \times [n]$. Нужно восстановить всю матрицу, предполагая, что она имеет малый ранг.

Некорректная (невыпуклая) постановка

$$\min_X \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2 \quad \text{s.t.} \quad \text{rank}(X) \leq r$$

- Ограничение $\text{rank}(X) \leq r$ — **невыпуклое**

Matrix completion: почему ядерная норма?

Задача восстановления матрицы. Дана матрица $B \in \mathbb{R}^{m \times n}$, наблюдаемая лишь на подмножестве индексов $\Omega \subset [m] \times [n]$. Нужно восстановить всю матрицу, предполагая, что она имеет малый ранг.

Некорректная (невыпуклая) постановка

$$\min_X \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2 \quad \text{s.t.} \quad \text{rank}(X) \leq r$$

- Ограничение $\text{rank}(X) \leq r$ — **невыпуклое**
- NP-трудная задача в общем случае

Matrix completion: почему ядерная норма?

Задача восстановления матрицы. Дана матрица $B \in \mathbb{R}^{m \times n}$, наблюдаемая лишь на подмножестве индексов $\Omega \subset [m] \times [n]$. Нужно восстановить всю матрицу, предполагая, что она имеет малый ранг.

Некорректная (невыпуклая) постановка

$$\min_X \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2 \quad \text{s.t.} \quad \text{rank}(X) \leq r$$

- Ограничение $\text{rank}(X) \leq r$ — **невыпуклое**
- NP-трудная задача в общем случае
- Нет гарантий глобальной оптимальности

Matrix completion: почему ядерная норма?

Задача восстановления матрицы. Дана матрица $B \in \mathbb{R}^{m\times n}$, наблюдаемая лишь на подмножестве индексов $\Omega \subset [m] \times [n]$. Нужно восстановить всю матрицу, предполагая, что она имеет малый ранг.

Некорректная (невыпуклая) постановка

$$\min_X \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2 \quad \text{s.t.} \quad \text{rank}(X) \leq r$$

- Ограничение $\text{rank}(X) \leq r$ — **невыпуклое**
- NP-трудная задача в общем случае
- Нет гарантий глобальной оптимальности

Matrix completion: почему ядерная норма?

Задача восстановления матрицы. Дана матрица $B \in \mathbb{R}^{m \times n}$, наблюдаемая лишь на подмножестве индексов $\Omega \subset [m] \times [n]$. Нужно восстановить всю матрицу, предполагая, что она имеет малый ранг.

Некорректная (невыпуклая) постановка

$$\min_X \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2 \quad \text{s.t.} \quad \text{rank}(X) \leq r$$

- Ограничение $\text{rank}(X) \leq r$ — **невыпуклое**
- NP-трудная задача в общем случае
- Нет гарантий глобальной оптимальности

Выпуклая релаксация

$$\min_X \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2 \quad \text{s.t.} \quad \|X\|_* \leq R$$

- Ядерная норма — **выпуклая оболочка** ранга

Matrix completion: почему ядерная норма?

Задача восстановления матрицы. Дана матрица $B \in \mathbb{R}^{m \times n}$, наблюдаемая лишь на подмножестве индексов $\Omega \subset [m] \times [n]$. Нужно восстановить всю матрицу, предполагая, что она имеет малый ранг.

Некорректная (невыпуклая) постановка

$$\min_X \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2 \quad \text{s.t.} \quad \text{rank}(X) \leq r$$

- Ограничение $\text{rank}(X) \leq r$ — **невыпуклое**
- NP-трудная задача в общем случае
- Нет гарантий глобальной оптимальности

Выпуклая релаксация

$$\min_X \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2 \quad \text{s.t.} \quad \|X\|_* \leq R$$

- Ядерная норма — **выпуклая оболочка** ранга
- Задача выпуклая, можно применять PGD и FW

Matrix completion: почему ядерная норма?

Задача восстановления матрицы. Дана матрица $B \in \mathbb{R}^{m\times n}$, наблюдаемая лишь на подмножестве индексов $\Omega \subset [m] \times [n]$. Нужно восстановить всю матрицу, предполагая, что она имеет малый ранг.

Некорректная (невыпуклая) постановка

$$\min_X \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2 \quad \text{s.t.} \quad \text{rank}(X) \leq r$$

- Ограничение $\text{rank}(X) \leq r$ — **невыпуклое**
- NP-трудная задача в общем случае
- Нет гарантий глобальной оптимальности

Выпуклая релаксация

$$\min_X \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2 \quad \text{s.t.} \quad \|X\|_* \leq R$$

- Ядерная норма — **выпуклая оболочка** ранга
- Задача выпуклая, можно применять PGD и FW
- Гарантии точного восстановления при определённых условиях [5]

[5] Candès, Recht. Exact Matrix Completion via Convex Optimization.

Matrix completion: почему ядерная норма?

Задача восстановления матрицы. Дана матрица $B \in \mathbb{R}^{m \times n}$, наблюдаемая лишь на подмножестве индексов $\Omega \subset [m] \times [n]$. Нужно восстановить всю матрицу, предполагая, что она имеет малый ранг.

Некорректная (невыпуклая) постановка

$$\min_X \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2 \quad \text{s.t.} \quad \text{rank}(X) \leq r$$

- Ограничение $\text{rank}(X) \leq r$ — **невыпуклое**
- NP-трудная задача в общем случае
- Нет гарантий глобальной оптимальности

Выпуклая релаксация

$$\min_X \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2 \quad \text{s.t.} \quad \|X\|_* \leq R$$

- Ядерная норма — **выпуклая оболочка** ранга
- Задача выпуклая, можно применять PGD и FW
- Гарантии точного восстановления при определённых условиях [5]

[5] Candès, Recht. Exact Matrix Completion via Convex Optimization.

Matrix completion: почему ядерная норма?

Задача восстановления матрицы. Дана матрица $B \in \mathbb{R}^{m\times n}$, наблюдаемая лишь на подмножестве индексов $\Omega \subset [m] \times [n]$. Нужно восстановить всю матрицу, предполагая, что она имеет малый ранг.

Некорректная (невыпуклая) постановка

$$\min_X \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2 \quad \text{s.t.} \quad \text{rank}(X) \leq r$$

- Ограничение $\text{rank}(X) \leq r$ — **невыпуклое**
- NP-трудная задача в общем случае
- Нет гарантий глобальной оптимальности

Выпуклая релаксация

$$\min_X \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - B_{ij})^2 \quad \text{s.t.} \quad \|X\|_* \leq R$$

- Ядерная норма — **выпуклая оболочка** ранга
- Задача выпуклая, можно применять PGD и FW
- Гарантии точного восстановления при определённых условиях [5]

Пример из практики: Netflix Prize — предсказание оценок пользователей для фильмов. Матрица ~ 500K $\times$ 17K, известно $\sim 1\%$ записей. Ядерная норма + FW позволяют эффективно решать эту задачу.

[5] Candès, Recht. Exact Matrix Completion via Convex Optimization.

Метод зеркального спуска

Мотивация: зачем нужен зеркальный спуск?

Вспомним итерацию метода проекции градиента:

$$
x_{k+1} = \text{proj}_S \left(x_k - \alpha_k \nabla f(x_k) \right)
$$

Мотивация: зачем нужен зеркальный спуск?

Вспомним итерацию метода проекции градиента:

$$x_{k+1} = \text{proj}_S \left(x_k - \alpha_k \nabla f(x_k)\right)$$

Эта формула опирается на **евклидову норму** $\|\cdot\|_2$, как в шаге градиента, так и в проекции. Но что, если:

- Геометрия задачи **неевклидова**? Например, оптимизация на симплексе $\Delta_n = \{x \geq 0, \mathbf{1}^\top x = 1\}$ — естественная геометрия описывается не ℓ_2, а ℓ_1-нормой.

Мотивация: зачем нужен зеркальный спуск?

Вспомним итерацию метода проекции градиента:

$$
x_{k+1} = \text{proj}_S \left(x_k - \alpha_k \nabla f(x_k) \right)
$$

Эта формула опирается на **евклидову норму** $\|\cdot\|_2$, как в шаге градиента, так и в проекции. Но что, если:

- Геометрия задачи **неевклидова**? Например, оптимизация на симплексе $\Delta_n = \{x \geq 0, \mathbf{1}^\top x = 1\}$ — естественная геометрия описывается не ℓ_2, а ℓ_1-нормой.
- Функция f гладкая относительно **неевклидовой** нормы $\|\cdot\|$?

$$
f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2
$$

Мотивация: зачем нужен зеркальный спуск?

Вспомним итерацию метода проекции градиента:

$$x_{k+1} = \text{proj}_S \left(x_k - \alpha_k \nabla f(x_k)\right)$$

Эта формула опирается на **евклидову норму** $\|\cdot\|_2$, как в шаге градиента, так и в проекции. Но что, если:

- Геометрия задачи **неевклидова**? Например, оптимизация на симплексе $\Delta_n = \{x \geq 0, \mathbf{1}^\top x = 1\}$ — естественная геометрия описывается не ℓ_2, а ℓ_1-нормой.
- Функция f гладкая относительно **неевклидовой** нормы $\|\cdot\|$?

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$$

Мотивация: зачем нужен зеркальный спуск?

Вспомним итерацию метода проекции градиента:

$$x_{k+1} = \text{proj}_S\left(x_k - \alpha_k \nabla f(x_k)\right)$$

Эта формула опирается на **евклидову норму** $\|\cdot\|_2$, как в шаге градиента, так и в проекции. Но что, если:

- Геометрия задачи **неевклидова**? Например, оптимизация на симплексе $\Delta_n = \{x \geq 0, \mathbf{1}^\top x = 1\}$ — естественная геометрия описывается не ℓ_2, а ℓ_1-нормой.
- Функция f гладкая относительно **неевклидовой** нормы $\|\cdot\|$?

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$$

Проблема: PGD с евклидовой проекцией может давать оценку $\mathcal{O}\left(\frac{L_2 R_2^2}{k}\right)$, но L_2 и R_2 могут **зависеть от размерности** n, что делает оценку бессмысленной для задач большой размерности.

Мотивация: пример на симплексе

Рассмотрим задачу на стандартном симплексе $\Delta_n = \{x \in \mathbb{R}^n \mid x \geq 0, \mathbf{1}^\top x = 1\}$:

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

Мотивация: пример на симплексе

Рассмотрим задачу на стандартном симплексе $\Delta_n = \{x \in \mathbb{R}^n \mid x \geq 0, \mathbf{1}^\top x = 1\}$:

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

Пусть f является L_1-гладкой относительно ℓ_1-нормы: $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_1}{2} \|y - x\|_1^2$.

Мотивация: пример на симплексе

Рассмотрим задачу на стандартном симплексе $\Delta_n = \{x \in \mathbb{R}^n \mid x \geq 0, \mathbf{1}^\top x = 1\}$:

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

Пусть f является L_1-гладкой относительно ℓ_1-нормы: $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_1}{2}\|y - x\|_1^2$.

PGD (евклидова геометрия)

- Гладкость: $L_2 \leq n \cdot L_1$ (из $\|x\|_2 \leq \|x\|_1$)

Мотивация: пример на симплексе

Рассмотрим задачу на стандартном симплексе $\Delta_n = \{x \in \mathbb{R}^n \mid x \geq 0, \mathbf{1}^\top x = 1\}$:

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

Пусть f является L_1-гладкой относительно ℓ_1-нормы: $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_1}{2}\|y - x\|_1^2$.

PGD (евклидова геометрия)

- Гладкость: $L_2 \leq n \cdot L_1$ (из $\|x\|_2 \leq \|x\|_1$)
- Диаметр: $R_2 = \sqrt{2}$

Мотивация: пример на симплексе

Рассмотрим задачу на стандартном симплексе $\Delta_n = \{x \in \mathbb{R}^n \mid x \geq 0, \mathbf{1}^\top x = 1\}$:

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

Пусть f является L_1-гладкой относительно ℓ_1-нормы: $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_1}{2}\|y - x\|_1^2$.

PGD (евклидова геометрия)

- Гладкость: $L_2 \leq n \cdot L_1$ (из $\|x\|_2 \leq \|x\|_1$)
- Диаметр: $R_2 = \sqrt{2}$
- **Оценка:**

$$f(x_k) - f^* \leq \frac{n \cdot L_1 \cdot 2}{2k} = \frac{nL_1}{k}$$

Мотивация: пример на симплексе

Рассмотрим задачу на стандартном симплексе $\Delta_n = \{x \in \mathbb{R}^n \mid x \geq 0, \mathbf{1}^\top x = 1\}$:

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

Пусть f является L_1-гладкой относительно ℓ_1-нормы: $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_1}{2}\|y - x\|_1^2$.

PGD (евклидова геометрия)

- Гладкость: $L_2 \leq n \cdot L_1$ (из $\|x\|_2 \leq \|x\|_1$)
- Диаметр: $R_2 = \sqrt{2}$
- **Оценка:**

$$f(x_k) - f^* \leq \frac{n \cdot L_1 \cdot 2}{2k} = \frac{nL_1}{k}$$

Мотивация: пример на симплексе

Рассмотрим задачу на стандартном симплексе $\Delta_n = \{x \in \mathbb{R}^n \mid x \geq 0, \mathbf{1}^\top x = 1\}$:

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

Пусть f является L_1-гладкой относительно ℓ_1-нормы: $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_1}{2}\|y - x\|_1^2$.

PGD (евклидова геометрия)

- Гладкость: $L_2 \leq n \cdot L_1$ (из $\|x\|_2 \leq \|x\|_1$)
- Диаметр: $R_2 = \sqrt{2}$
- **Оценка:**

$$f(x_k) - f^* \leq \frac{n \cdot L_1 \cdot 2}{2k} = \frac{nL_1}{k}$$

Зеркальный спуск (ℓ_1-геометрия)

- Гладкость: L_1 (без множителя!)

Мотивация: пример на симплексе

Рассмотрим задачу на стандартном симплексе $\Delta_n = \{x \in \mathbb{R}^n \mid x \geq 0, \mathbf{1}^\top x = 1\}$:

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

Пусть f является L_1-гладкой относительно ℓ_1-нормы: $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_1}{2}\|y - x\|_1^2$.

PGD (евклидова геометрия)

- Гладкость: $L_2 \leq n \cdot L_1$ (из $\|x\|_2 \leq \|x\|_1$)
- Диаметр: $R_2 = \sqrt{2}$
- **Оценка:**

$$f(x_k) - f^* \leq \frac{n \cdot L_1 \cdot 2}{2k} = \frac{nL_1}{k}$$

Зеркальный спуск (ℓ_1-геометрия)

- Гладкость: L_1 (без множителя!)
- «Диаметр»: $R_1 = \sqrt{2 \ln n}$ (в смысле дивергенции Брэгмана)

Мотивация: пример на симплексе

Рассмотрим задачу на стандартном симплексе $\Delta_n = \{x \in \mathbb{R}^n \mid x \geq 0, \mathbf{1}^\top x = 1\}$:

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

Пусть f является L_1-гладкой относительно ℓ_1-нормы: $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_1}{2}\|y - x\|_1^2$.

PGD (евклидова геометрия)

- Гладкость: $L_2 \leq n \cdot L_1$ (из $\|x\|_2 \leq \|x\|_1$)
- Диаметр: $R_2 = \sqrt{2}$
- **Оценка:**

$$f(x_k) - f^* \leq \frac{n \cdot L_1 \cdot 2}{2k} = \frac{nL_1}{k}$$

Зеркальный спуск (ℓ_1-геометрия)

- Гладкость: L_1 (без множителя!)
- «Диаметр»: $R_1 = \sqrt{2 \ln n}$ (в смысле дивергенции Брэгмана)
- **Оценка:**

$$f(x_k) - f^* \leq \frac{L_1 \cdot 2 \ln n}{2k} = \frac{L_1 \ln n}{k}$$

Мотивация: пример на симплексе

Рассмотрим задачу на стандартном симплексе $\Delta_n = \{x \in \mathbb{R}^n \mid x \geq 0, \mathbf{1}^\top x = 1\}$:

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

Пусть f является L_1-гладкой относительно ℓ_1-нормы: $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_1}{2}\|y - x\|_1^2$.

PGD (евклидова геометрия)

- Гладкость: $L_2 \leq n \cdot L_1$ (из $\|x\|_2 \leq \|x\|_1$)
- Диаметр: $R_2 = \sqrt{2}$
- **Оценка:**

$$f(x_k) - f^* \leq \frac{n \cdot L_1 \cdot 2}{2k} = \frac{nL_1}{k}$$

Зеркальный спуск (ℓ_1-геометрия)

- Гладкость: L_1 (без множителя!)
- «Диаметр»: $R_1 = \sqrt{2 \ln n}$ (в смысле дивергенции Брэгмана)
- **Оценка:**

$$f(x_k) - f^* \leq \frac{L_1 \cdot 2 \ln n}{2k} = \frac{L_1 \ln n}{k}$$

Мотивация: пример на симплексе

Рассмотрим задачу на стандартном симплексе $\Delta_n = \{x \in \mathbb{R}^n \mid x \geq 0, \mathbf{1}^\top x = 1\}$:

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

Пусть f является L_1-гладкой относительно ℓ_1-нормы: $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_1}{2}\|y - x\|_1^2$.

PGD (евклидова геометрия)

- Гладкость: $L_2 \leq n \cdot L_1$ (из $\|x\|_2 \leq \|x\|_1$)
- Диаметр: $R_2 = \sqrt{2}$
- **Оценка:**

$$f(x_k) - f^* \leq \frac{n \cdot L_1 \cdot 2}{2k} = \frac{nL_1}{k}$$

Зеркальный спуск (ℓ_1-геометрия)

- Гладкость: L_1 (без множителя!)
- «Диаметр»: $R_1 = \sqrt{2 \ln n}$ (в смысле дивергенции Брэгмана)
- **Оценка:**

$$f(x_k) - f^* \leq \frac{L_1 \cdot 2 \ln n}{2k} = \frac{L_1 \ln n}{k}$$

Выигрыш: $\frac{n}{\ln n}$ — **экспоненциальный** по размерности!

Мотивация: прямое и двойственное пространство

Идея. Градиент $\nabla f(x)$ живёт в **двойственном** пространстве $(\mathbb{R}^n)^*$ с нормой $\|\cdot\|_*$, а переменная x — в **прямом** пространстве $\mathbb{R}^n$ с нормой $\|\cdot\|$.

Мотивация: прямое и двойственное пространство

Идея. Градиент $\nabla f(x)$ живёт в **двойственном** пространстве $(\mathbb{R}^n)^*$ с нормой $\|\cdot\|_*$, а переменная x — в **прямом** пространстве $\mathbb{R}^n$ с нормой $\|\cdot\|$.

В PGD мы неявно отождествляем прямое и двойственное пространство (что корректно только для ℓ_2):

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

Вычитаем **градиент** (двойственный вектор) из **точки** (прямой вектор) напрямую.

Мотивация: прямое и двойственное пространство

Идея. Градиент $\nabla f(x)$ живёт в **двойственном** пространстве $(\mathbb{R}^n)^*$ с нормой $\|\cdot\|_*$, а переменная x — в **прямом** пространстве $\mathbb{R}^n$ с нормой $\|\cdot\|$.

В PGD мы неявно отождествляем прямое и двойственное пространство (что корректно только для ℓ_2):

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

Вычитаем **градиент** (двойственный вектор) из **точки** (прямой вектор) напрямую.

В зеркальном спуске мы **явно** работаем с двумя пространствами через **зеркальное отображение** $\nabla\omega$:

$$\nabla\omega(y_{k+1}) = \nabla\omega(x_k) - \alpha_k \nabla f(x_k)$$
$$x_{k+1} = \nabla\omega^*(y_{k+1})$$

Шаг делается в двойственном пространстве, затем результат отображается обратно.

Мотивация: прямое и двойственное пространство

Идея. Градиент $\nabla f(x)$ живёт в **двойственном** пространстве $(\mathbb{R}^n)^*$ с нормой $\|\cdot\|_*$, а переменная x — в **прямом** пространстве $\mathbb{R}^n$ с нормой $\|\cdot\|$.

В PGD мы неявно отождествляем прямое и двойственное пространство (что корректно только для ℓ_2):

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

Вычитаем **градиент** (двойственный вектор) из **точки** (прямой вектор) напрямую.

В зеркальном спуске мы **явно** работаем с двумя пространствами через **зеркальное отображение** $\nabla\omega$:

$$\nabla\omega(y_{k+1}) = \nabla\omega(x_k) - \alpha_k \nabla f(x_k)$$
$$x_{k+1} = \nabla\omega^*(y_{k+1})$$

Шаг делается в двойственном пространстве, затем результат отображается обратно.

Функция ω называется **прокс-функцией** (distance-generating function). Она задаёт геометрию через дивергенцию Брэгмана.

Дивергенция Брэгмана

Дивергенция Брэгмана

Пусть $\omega : \mathbb{R}^n \to \mathbb{R}$ — непрерывно дифференцируемая и строго выпуклая функция. **Дивергенция Брэгмана**, порождённая функцией ω, определяется как:

$$V_\omega(x, y) = \omega(x) - \omega(y) - \langle \nabla \omega(y), x - y \rangle$$

Дивергенция Брэгмана

Дивергенция Брэгмана

Пусть $\omega : \mathbb{R}^n \to \mathbb{R}$ — непрерывно дифференцируемая и строго выпуклая функция. **Дивергенция Брэгмана**, порождённая функцией ω, определяется как:

$$V_\omega(x, y) = \omega(x) - \omega(y) - \langle \nabla \omega(y), x - y \rangle$$

Свойства:

- $V_\omega(x, y) \geq 0$ для всех x, y (из строгой выпуклости ω).

Дивергенция Брэгмана

Дивергенция Брэгмана

Пусть $\omega : \mathbb{R}^n \to \mathbb{R}$ — непрерывно дифференцируемая и строго выпуклая функция. **Дивергенция Брэгмана**, порождённая функцией ω, определяется как:

$$V_\omega(x, y) = \omega(x) - \omega(y) - \langle \nabla \omega(y), x - y \rangle$$

Свойства:

- $V_\omega(x, y) \geq 0$ для всех x, y (из строгой выпуклости ω).
- $V_\omega(x, y) = 0 \Leftrightarrow x = y$.

Дивергенция Брэгмана

Дивергенция Брэгмана

Пусть $\omega : \mathbb{R}^n \to \mathbb{R}$ — непрерывно дифференцируемая и строго выпуклая функция. **Дивергенция Брэгмана**, порождённая функцией ω, определяется как:

$$V_\omega(x, y) = \omega(x) - \omega(y) - \langle \nabla \omega(y), x - y \rangle$$

Свойства:

- $V_\omega(x, y) \geq 0$ для всех x, y (из строгой выпуклости ω).
- $V_\omega(x, y) = 0 \Leftrightarrow x = y$.
- **Не симметрична** в общем случае: $V_\omega(x, y) \neq V_\omega(y, x)$.

Дивергенция Брэгмана

Дивергенция Брэгмана

Пусть $\omega : \mathbb{R}^n \to \mathbb{R}$ — непрерывно дифференцируемая и строго выпуклая функция. **Дивергенция Брэгмана**, порождённая функцией ω, определяется как:

$$V_\omega(x, y) = \omega(x) - \omega(y) - \langle \nabla \omega(y), x - y \rangle$$

Свойства:

- $V_\omega(x, y) \geq 0$ для всех x, y (из строгой выпуклости ω).
- $V_\omega(x, y) = 0 \Leftrightarrow x = y$.
- **Не симметрична** в общем случае: $V_\omega(x, y) \neq V_\omega(y, x)$.
- **Не удовлетворяет** неравенству треугольника.

Дивергенция Брэгмана

Дивергенция Брэгмана

Пусть $\omega : \mathbb{R}^n \to \mathbb{R}$ — непрерывно дифференцируемая и строго выпуклая функция. **Дивергенция Брэгмана**, порождённая функцией ω, определяется как:

$$V_\omega(x, y) = \omega(x) - \omega(y) - \langle \nabla \omega(y), x - y \rangle$$

Свойства:

- $V_\omega(x, y) \geq 0$ для всех x, y (из строгой выпуклости ω).
- $V_\omega(x, y) = 0 \Leftrightarrow x = y$.
- **Не симметрична** в общем случае: $V_\omega(x, y) \neq V_\omega(y, x)$.
- **Не удовлетворяет** неравенству треугольника.
- Геометрически: $V_\omega(x, y)$ — разность между $\omega(x)$ и значением касательной к ω в точке y, вычисленным в точке x.

Примеры дивергенций Брэгмана

Евклидова дивергенция

$\omega(x) = \frac{1}{2}\|x\|_2^2$

$$V_\omega(x, y) = \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - \langle y, x - y \rangle = \frac{1}{2}\|x - y\|_2^2$$

Воспроизводит обычный PGD.

Примеры дивергенций Брэгмана

Евклидова дивергенция

$\omega(x) = \frac{1}{2}\|x\|_2^2$

$$
V_\omega(x, y) = \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - \langle y, x - y \rangle = \frac{1}{2}\|x - y\|_2^2
$$

Воспроизводит обычный PGD.

Дивергенция Кульбака-Лейблера (KL)

$\omega(x) = \sum_{i=1}^n x_i \ln x_i$ (негативная энтропия) на Δ_n

$$
V_\omega(x, y) = \sum_{i=1}^n x_i \ln \frac{x_i}{y_i}
$$

Естественна для задач на симплексе.

Примеры дивергенций Брэгмана

Евклидова дивергенция

$\omega(x) = \frac{1}{2}\|x\|_2^2$

$$V_\omega(x, y) = \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - \langle y, x - y\rangle = \frac{1}{2}\|x - y\|_2^2$$

Воспроизводит обычный PGD.

Дивергенция Кульбака-Лейблера (KL)

$\omega(x) = \sum_{i=1}^n x_i \ln x_i$ (негативная энтропия) на Δ_n

$$V_\omega(x, y) = \sum_{i=1}^n x_i \ln \frac{x_i}{y_i}$$

Естественна для задач на симплексе.

Дивергенция Итакура-Саито

$\omega(x) = -\sum_{i=1}^n \ln x_i$ на $\mathbb{R}_{++}^n$

$$V_\omega(x, y) = \sum_{i=1}^n \left(\frac{x_i}{y_i} - \ln \frac{x_i}{y_i} - 1 \right)$$

Используется в обработке сигналов.

Примеры дивергенций Брэгмана

Евклидова дивергенция

$\omega(x) = \frac{1}{2}\|x\|_2^2$

$$V_\omega(x,y) = \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - \langle y, x-y\rangle = \frac{1}{2}\|x-y\|_2^2$$

Воспроизводит обычный PGD.

Дивергенция Кульбака-Лейблера (KL)

$\omega(x) = \sum_{i=1}^n x_i \ln x_i$ (негативная энтропия) на Δ_n

$$V_\omega(x,y) = \sum_{i=1}^n x_i \ln \frac{x_i}{y_i}$$

Естественна для задач на симплексе.

Дивергенция Итакура-Саито

$\omega(x) = -\sum_{i=1}^n \ln x_i$ на $\mathbb{R}^n_{++}$

$$V_\omega(x,y) = \sum_{i=1}^n \left(\frac{x_i}{y_i} - \ln\frac{x_i}{y_i} - 1\right)$$

Используется в обработке сигналов.

Дивергенция Махаланобиса

$\omega(x) = \frac{1}{2}x^\top Q x$, где $Q \succ 0$

$$V_\omega(x,y) = \frac{1}{2}(x-y)^\top Q(x-y)$$

Обобщённая евклидова геометрия.

Сильная выпуклость относительно нормы

Definition

Функция ω называется σ-**сильно выпуклой относительно нормы** $\|\cdot\|$, если для всех x, y из области определения:

$$V_\omega(x, y) \geq \frac{\sigma}{2}\|x - y\|^2$$

Сильная выпуклость относительно нормы

Definition

Функция ω называется **σ-сильно выпуклой относительно нормы** $\|\cdot\|$, если для всех x, y из области определения:

$$V_\omega(x, y) \geq \frac{\sigma}{2}\|x - y\|^2$$

Ключевые примеры:

- $\omega(x) = \frac{1}{2}\|x\|_2^2$ является 1-сильно выпуклой относительно $\|\cdot\|_2$.

Сильная выпуклость относительно нормы

Definition

Функция ω называется **σ-сильно выпуклой относительно нормы** $\|\cdot\|$, если для всех x, y из области определения:

$$V_\omega(x, y) \geq \frac{\sigma}{2}\|x - y\|^2$$

Ключевые примеры:

- $\omega(x) = \frac{1}{2}\|x\|_2^2$ является 1-сильно выпуклой относительно $\|\cdot\|_2$.
- $\omega(x) = \sum_i x_i \ln x_i$ (негативная энтропия) является 1-сильно выпуклой относительно $\|\cdot\|_1$ на симплексе Δ_n.

Сильная выпуклость относительно нормы

Definition

Функция ω называется σ-**сильно выпуклой относительно нормы** $\|\cdot\|$, если для всех x, y из области определения:

$$V_\omega(x,y) \geq \frac{\sigma}{2}\|x-y\|^2$$

Ключевые примеры:

- $\omega(x) = \frac{1}{2}\|x\|_2^2$ является 1-сильно выпуклой относительно $\|\cdot\|_2$.
- $\omega(x) = \sum_i x_i \ln x_i$ (негативная энтропия) является 1-сильно выпуклой относительно $\|\cdot\|_1$ на симплексе Δ_n.
  - Это утверждение известно как **неравенство Пинскера** и является нетривиальным результатом.

Сильная выпуклость относительно нормы

Definition

Функция ω называется σ-**сильно выпуклой относительно нормы** $\|\cdot\|$, если для всех x, y из области определения:

$$V_\omega(x, y) \geq \frac{\sigma}{2}\|x - y\|^2$$

Ключевые примеры:

- $\omega(x) = \frac{1}{2}\|x\|_2^2$ является 1-сильно выпуклой относительно $\|\cdot\|_2$.
- $\omega(x) = \sum_i x_i \ln x_i$ (негативная энтропия) является 1-сильно выпуклой относительно $\|\cdot\|_1$ на симплексе Δ_n.
  - Это утверждение известно как **неравенство Пинскера** и является нетривиальным результатом.

Сильная выпуклость относительно нормы

Definition

Функция ω называется **σ-сильно выпуклой относительно нормы** $\|\cdot\|$, если для всех x, y из области определения:

$$V_\omega(x, y) \geq \frac{\sigma}{2}\|x - y\|^2$$

Ключевые примеры:

- $\omega(x) = \frac{1}{2}\|x\|_2^2$ является 1-сильно выпуклой относительно $\|\cdot\|_2$.
- $\omega(x) = \sum_i x_i \ln x_i$ (негативная энтропия) является 1-сильно выпуклой относительно $\|\cdot\|_1$ на симплексе Δ_n.
    - Это утверждение известно как **неравенство Пинскера** и является нетривиальным результатом.

Именно σ-сильная выпуклость прокс-функции ω относительно нормы $\|\cdot\|$ позволяет дивергенции Брэгмана «измерять расстояния» в геометрии, задаваемой нормой $\|\cdot\|$.

Алгоритм зеркального спуска

Задача: $\min_{x\in S} f(x)$, где f — выпуклая функция, S — замкнутое выпуклое множество.

Алгоритм зеркального спуска

Задача: $\min_{x \in S} f(x)$, где f — выпуклая функция, S — замкнутое выпуклое множество.

Метод зеркального спуска (Mirror Descent)

Вход: $x_0 \in S$, шаги $\{\alpha_k\}$, прокс-функция ω.
Для $k = 0, 1, 2, \ldots$:

1. Вычислить градиент $g_k = \nabla f(x_k)$

Алгоритм зеркального спуска

Задача: $\min_{x \in S} f(x)$, где f — выпуклая функция, S — замкнутое выпуклое множество.

Метод зеркального спуска (Mirror Descent)

Вход: $x_0 \in S$, шаги $\{\alpha_k\}$, прокс-функция ω.
Для $k = 0, 1, 2, \ldots$:

1. Вычислить градиент $g_k = \nabla f(x_k)$
2. Шаг в двойственном пространстве:

$$x_{k+1} = \arg\min_{x \in S} \left\{ \langle g_k, x \rangle + \frac{1}{\alpha_k} V_\omega(x, x_k) \right\}$$

Алгоритм зеркального спуска

Задача: $\min_{x \in S} f(x)$, где f — выпуклая функция, S — замкнутое выпуклое множество.

Метод зеркального спуска (Mirror Descent)

Вход: $x_0 \in S$, шаги $\{\alpha_k\}$, прокс-функция ω.
Для $k = 0, 1, 2, \ldots$:

1. Вычислить градиент $g_k = \nabla f(x_k)$
2. Шаг в двойственном пространстве:

$$x_{k+1} = \arg\min_{x \in S} \left\{ \langle g_k, x \rangle + \frac{1}{\alpha_k} V_\omega(x, x_k) \right\}$$

Алгоритм зеркального спуска

Задача: $\min_{x \in S} f(x)$, где f — выпуклая функция, S — замкнутое выпуклое множество.

Метод зеркального спуска (Mirror Descent)

Вход: $x_0 \in S$, шаги $\{\alpha_k\}$, прокс-функция ω.
Для $k = 0, 1, 2, \ldots$:

1. Вычислить градиент $g_k = \nabla f(x_k)$
2. Шаг в двойственном пространстве:

$$x_{k+1} = \arg\min_{x \in S} \left\{ \langle g_k, x \rangle + \frac{1}{\alpha_k} V_\omega(x, x_k) \right\}$$

Интерпретация: на каждом шаге минимизируем линейное приближение f с регуляризацией дивергенцией Брэгмана вместо $\frac{1}{2\alpha} \|x - x_k\|_2^2$.

Алгоритм зеркального спуска

Задача: $\min_{x\in S} f(x)$, где f — выпуклая функция, S — замкнутое выпуклое множество.

Метод зеркального спуска (Mirror Descent)

Вход: $x_0 \in S$, шаги $\{\alpha_k\}$, прокс-функция ω.
Для $k = 0, 1, 2, \dots$:

1. Вычислить градиент $g_k = \nabla f(x_k)$
2. Шаг в двойственном пространстве:

$$x_{k+1} = \arg\min_{x\in S}\left\{\langle g_k, x\rangle + \frac{1}{\alpha_k} V_\omega(x, x_k)\right\}$$

Интерпретация: на каждом шаге минимизируем линейное приближение f с регуляризацией дивергенцией Брэгмана вместо $\frac{1}{2\alpha}\|x - x_k\|_2^2$.

Частный случай: при $\omega(x) = \frac{1}{2}\|x\|_2^2$ получаем $V_\omega(x, y) = \frac{1}{2}\|x - y\|_2^2$, и итерация принимает вид:

$$x_{k+1} = \arg\min_{x\in S}\left\{\langle \nabla f(x_k), x\rangle + \frac{1}{2\alpha_k}\|x - x_k\|_2^2\right\} = \text{proj}_S(x_k - \alpha_k \nabla f(x_k))$$

то есть стандартный PGD.

Зеркальный спуск: эквивалентные формы итерации

Итерацию зеркального спуска можно записать в нескольких эквивалентных формах:

Зеркальный спуск: эквивалентные формы итерации

Итерацию зеркального спуска можно записать в нескольких эквивалентных формах:

1. Проксимальная форма (основная):

$$x_{k+1} = \arg\min_{x \in S} \{\alpha_k \langle \nabla f(x_k), x \rangle + V_\omega(x, x_k)\}$$

Зеркальный спуск: эквивалентные формы итерации

Итерацию зеркального спуска можно записать в нескольких эквивалентных формах:

1. Проксимальная форма (основная):

$$x_{k+1} = \arg\min_{x \in S} \{\alpha_k \langle \nabla f(x_k), x \rangle + V_\omega(x, x_k)\}$$

2. Двойственная форма (через зеркальное отображение):

Для безусловной задачи ($S = \mathbb{R}^n$), используя условие оптимальности $\nabla_x V_\omega(x, x_k) = \nabla\omega(x) - \nabla\omega(x_k)$:

$$\nabla\omega(x_{k+1}) = \nabla\omega(x_k) - \alpha_k \nabla f(x_k)$$

Зеркальный спуск: эквивалентные формы итерации

Итерацию зеркального спуска можно записать в нескольких эквивалентных формах:

1. Проксимальная форма (основная):

$$
x_{k+1} = \arg\min_{x \in S} \{\alpha_k \langle \nabla f(x_k), x \rangle + V_\omega(x, x_k)\}
$$

2. Двойственная форма (через зеркальное отображение):

Для безусловной задачи ($S = \mathbb{R}^n$), используя условие оптимальности $\nabla_x V_\omega(x, x_k) = \nabla \omega(x) - \nabla \omega(x_k)$:

$$
\nabla \omega(x_{k+1}) = \nabla \omega(x_k) - \alpha_k \nabla f(x_k)
$$

3. Проекция Брэгмана (для условной задачи):

$$
\begin{aligned}
y_{k+1} : \quad & \nabla \omega(y_{k+1}) = \nabla \omega(x_k) - \alpha_k \nabla f(x_k) \\
x_{k+1} &= \arg\min_{x \in S} V_\omega(x, y_{k+1})
\end{aligned}
$$

Зеркальный спуск: эквивалентные формы итерации

Итерацию зеркального спуска можно записать в нескольких эквивалентных формах:

1. Проксимальная форма (основная):

$$x_{k+1} = \arg\min_{x \in S} \{\alpha_k \langle \nabla f(x_k), x \rangle + V_\omega(x, x_k)\}$$

2. Двойственная форма (через зеркальное отображение):

Для безусловной задачи ($S = \mathbb{R}^n$), используя условие оптимальности $\nabla_x V_\omega(x, x_k) = \nabla \omega(x) - \nabla \omega(x_k)$:

$$\nabla \omega(x_{k+1}) = \nabla \omega(x_k) - \alpha_k \nabla f(x_k)$$

3. Проекция Брэгмана (для условной задачи):

$$\begin{aligned} y_{k+1} : \quad & \nabla \omega(y_{k+1}) = \nabla \omega(x_k) - \alpha_k \nabla f(x_k) \\ x_{k+1} &= \arg\min_{x \in S} V_\omega(x, y_{k+1}) \end{aligned}$$

В евклидовом случае ($\omega = \frac{1}{2}\|\cdot\|_2^2$) проекция Брэгмана совпадает с евклидовой проекцией.

Зеркальный спуск на симплексе: Exponentiated Gradient

Важнейший частный случай: $S = \Delta_n$, $\omega(x) = \sum_i x_i \ln x_i$.

Зеркальный спуск на симплексе: Exponentiated Gradient

Важнейший частный случай: $S = \Delta_n$, $\omega(x) = \sum_i x_i \ln x_i$.

Итерация зеркального спуска принимает **замкнутую форму**:

$$
x_{k+1,i} = \frac{x_{k,i} \exp\left(-\alpha_k [\nabla f(x_k)]_i\right)}{\sum_{j=1}^n x_{k,j} \exp\left(-\alpha_k [\nabla f(x_k)]_j\right)}, \quad i = 1, \dots, n
$$

Зеркальный спуск на симплексе: Exponentiated Gradient

Важнейший частный случай: $S = \Delta_n$, $\omega(x) = \sum_i x_i \ln x_i$.

Итерация зеркального спуска принимает **замкнутую форму**:

$$x_{k+1,i} = \frac{x_{k,i} \exp\left(-\alpha_k [\nabla f(x_k)]_i\right)}{\sum_{j=1}^n x_{k,j} \exp\left(-\alpha_k [\nabla f(x_k)]_j\right)}, \quad i = 1, \dots, n$$

Этот метод известен как **Exponentiated Gradient** (EG) или **Multiplicative Weights Update**.

Зеркальный спуск на симплексе: Exponentiated Gradient

Важнейший частный случай: $S = \Delta_n$, $\omega(x) = \sum_i x_i \ln x_i$.

Итерация зеркального спуска принимает **замкнутую форму**:

$$x_{k+1,i} = \frac{x_{k,i} \exp\left(-\alpha_k [\nabla f(x_k)]_i\right)}{\sum_{j=1}^{n} x_{k,j} \exp\left(-\alpha_k [\nabla f(x_k)]_j\right)}, \quad i = 1, \dots, n$$

Этот метод известен как **Exponentiated Gradient** (EG) или **Multiplicative Weights Update**.

Вывод. Из условия оптимальности: $\ln x_{k+1,i} + 1 - \ln x_{k,i} - 1 + \alpha_k g_i + \lambda = 0$, где λ — множитель Лагранжа для ограничения $\sum_i x_i = 1$. Тогда $x_{k+1,i} \propto x_{k,i} \exp(-\alpha_k g_i)$, и нормировка даёт формулу выше.

Зеркальный спуск на симплексе: Exponentiated Gradient

Важнейший частный случай: $S = \Delta_n$, $\omega(x) = \sum_i x_i \ln x_i$.

Итерация зеркального спуска принимает **замкнутую форму**:

$$x_{k+1,i} = \frac{x_{k,i} \exp\left(-\alpha_k [\nabla f(x_k)]_i\right)}{\sum_{j=1}^n x_{k,j} \exp\left(-\alpha_k [\nabla f(x_k)]_j\right)}, \quad i = 1, \dots, n$$

Этот метод известен как **Exponentiated Gradient** (EG) или **Multiplicative Weights Update**.

Вывод. Из условия оптимальности: $\ln x_{k+1,i} + 1 - \ln x_{k,i} - 1 + \alpha_k g_i + \lambda = 0$, где λ — множитель Лагранжа для ограничения $\sum_i x_i = 1$. Тогда $x_{k+1,i} \propto x_{k,i} \exp(-\alpha_k g_i)$, и нормировка даёт формулу выше.

Стоимость: $\mathcal{O}(n)$ — **не требуется** сортировка (в отличие от евклидовой проекции на симплекс, которая стоит $\mathcal{O}(n \log n)$).

Скорость сходимости зеркального спуска

Theorem

Пусть $f : S \to \mathbb{R}$ — выпуклая функция с L-липшицевым градиентом относительно нормы $\|\cdot\|$:

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \quad \forall x, y \in S$$

Пусть прокс-функция ω является σ-сильно выпуклой относительно $\|\cdot\|$, и $R^2 = \max_{x \in S} V_\omega(x, x_0)$. Тогда зеркальный спуск с шагом $\alpha_k = \frac{1}{L/\sigma}$ обеспечивает:

$$f(x_k) - f^* \leq \frac{LR^2}{\sigma k}$$

Скорость сходимости зеркального спуска

Theorem

Пусть $f : S \to \mathbb{R}$ — выпуклая функция с L-липшицевым градиентом относительно нормы $\|\cdot\|$:

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \quad \forall x, y \in S$$

Пусть прокс-функция ω является σ-сильно выпуклой относительно $\|\cdot\|$, и $R^2 = \max_{x\in S} V_\omega(x, x_0)$. Тогда зеркальный спуск с шагом $\alpha_k = \frac{1}{L/\sigma}$ обеспечивает:

$$f(x_k) - f^* \leq \frac{LR^2}{\sigma k}$$

Сравнение с PGD:

	PGD	Зеркальный спуск
Норма	$\|\cdot\|_2$	$\|\cdot\|$ (произвольная)
Гладкость	L_2	L (отн. $\|\cdot\|$)
«Радиус»	$R_2 = \|x_0 - x^*\|_2$	$R^2 = V_\omega(x^*, x_0)$
Оценка	$\frac{L_2 R_2^2}{k}$	$\frac{LR^2}{\sigma k}$

Доказательство сходимости

1. Из L-гладкости f относительно нормы $\|\cdot\|$ при шаге $\alpha = \sigma/L$:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

Доказательство сходимости

1. Из L-гладкости f относительно нормы $\|\cdot\|$ при шаге $\alpha = \sigma/L$:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

Доказательство сходимости

1. Из L-гладкости f относительно нормы $\|\cdot\|$ при шаге $\alpha = \sigma/L$:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

2. Из σ-сильной выпуклости ω: $\frac{L}{2}\|x_{k+1} - x_k\|^2 \leq \frac{L}{\sigma} V_\omega(x_{k+1}, x_k)$. Следовательно:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{\sigma} V_\omega(x_{k+1}, x_k)$$

Доказательство сходимости

1. Из L-гладкости f относительно нормы $\|\cdot\|$ при шаге $\alpha = \sigma/L$:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

2. Из σ-сильной выпуклости ω: $\frac{L}{2}\|x_{k+1} - x_k\|^2 \leq \frac{L}{\sigma} V_\omega(x_{k+1}, x_k)$. Следовательно:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{\sigma} V_\omega(x_{k+1}, x_k)$$

Доказательство сходимости

1. Из L-гладкости f относительно нормы $\|\cdot\|$ при шаге $\alpha = \sigma/L$:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

2. Из σ-сильной выпуклости ω: $\frac{L}{2}\|x_{k+1} - x_k\|^2 \leq \frac{L}{\sigma} V_\omega(x_{k+1}, x_k)$. Следовательно:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{\sigma} V_\omega(x_{k+1}, x_k)$$

3. Из определения итерации (условие оптимальности) и выпуклости f для любого $x \in S$:

$$\alpha_k \langle \nabla f(x_k), x_{k+1} - x \rangle \leq V_\omega(x, x_k) - V_\omega(x, x_{k+1}) - V_\omega(x_{k+1}, x_k)$$

Доказательство сходимости

1. Из L-гладкости f относительно нормы $\|\cdot\|$ при шаге $\alpha = \sigma/L$:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

2. Из σ-сильной выпуклости ω: $\frac{L}{2}\|x_{k+1} - x_k\|^2 \leq \frac{L}{\sigma} V_\omega(x_{k+1}, x_k)$. Следовательно:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{\sigma} V_\omega(x_{k+1}, x_k)$$

3. Из определения итерации (условие оптимальности) и выпуклости f для любого $x \in S$:

$$\alpha_k \langle \nabla f(x_k), x_{k+1} - x \rangle \leq V_\omega(x, x_k) - V_\omega(x, x_{k+1}) - V_\omega(x_{k+1}, x_k)$$

Доказательство сходимости

1. Из L-гладкости f относительно нормы $\|\cdot\|$ при шаге $\alpha = \sigma/L$:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

2. Из σ-сильной выпуклости ω: $\frac{L}{2}\|x_{k+1} - x_k\|^2 \leq \frac{L}{\sigma} V_\omega(x_{k+1}, x_k)$. Следовательно:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{\sigma} V_\omega(x_{k+1}, x_k)$$

3. Из определения итерации (условие оптимальности) и выпуклости f для любого $x \in S$:

$$\alpha_k \langle \nabla f(x_k), x_{k+1} - x \rangle \leq V_\omega(x, x_k) - V_\omega(x, x_{k+1}) - V_\omega(x_{k+1}, x_k)$$

Это **лемма трёх точек** (three-point identity) для дивергенции Брэгмана:

$$V_\omega(x, x_k) - V_\omega(x, x_{k+1}) = \langle \nabla \omega(x_{k+1}) - \nabla \omega(x_k), x - x_{k+1} \rangle + V_\omega(x_{k+1}, x_k)$$

Доказательство сходимости (продолжение)

4. Подставляя $x = x^*$ и используя выпуклость $f(x_k) - f^* \le \langle \nabla f(x_k), x_k - x^* \rangle$:

$$
\begin{aligned}
\alpha_k (f(x_k) - f^*) &\le \alpha_k \langle \nabla f(x_k), x_k - x^* \rangle \\
&= \alpha_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle + \alpha_k \langle \nabla f(x_k), x_{k+1} - x^* \rangle \\
&\le \alpha_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle + V_\omega(x^*, x_k) - V_\omega(x^*, x_{k+1}) - V_\omega(x_{k+1}, x_k)
\end{aligned}
$$

Доказательство сходимости (продолжение)

4. Подставляя $x = x^*$ и используя выпуклость $f(x_k) - f^* \leq \langle \nabla f(x_k), x_k - x^* \rangle$:

$$
\begin{aligned}
\alpha_k (f(x_k) - f^*) &\leq \alpha_k \langle \nabla f(x_k), x_k - x^* \rangle \\
&= \alpha_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle + \alpha_k \langle \nabla f(x_k), x_{k+1} - x^* \rangle \\
&\leq \alpha_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle + V_\omega(x^*, x_k) - V_\omega(x^*, x_{k+1}) - V_\omega(x_{k+1}, x_k)
\end{aligned}
$$

Доказательство сходимости (продолжение)

4. Подставляя $x = x^*$ и используя выпуклость $f(x_k) - f^* \leq \langle \nabla f(x_k), x_k - x^* \rangle$:

$$
\begin{aligned}
\alpha_k(f(x_k) - f^*) &\leq \alpha_k \langle \nabla f(x_k), x_k - x^* \rangle \\
&= \alpha_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle + \alpha_k \langle \nabla f(x_k), x_{k+1} - x^* \rangle \\
&\leq \alpha_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle + V_\omega(x^*, x_k) - V_\omega(x^*, x_{k+1}) - V_\omega(x_{k+1}, x_k)
\end{aligned}
$$

5. Суммируя по $k = 0, \ldots, K-1$ и используя телескопирование:

$$
\sum_{k=0}^{K-1} \alpha_k(f(x_k) - f^*) \leq V_\omega(x^*, x_0) + \sum_{k=0}^{K-1} \left[\alpha_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle - V_\omega(x_{k+1}, x_k)\right]
$$

Доказательство сходимости (продолжение)

4. Подставляя $x = x^*$ и используя выпуклость $f(x_k) - f^* \leq \langle \nabla f(x_k), x_k - x^* \rangle$:

$$
\begin{aligned}
\alpha_k (f(x_k) - f^*) &\leq \alpha_k \langle \nabla f(x_k), x_k - x^* \rangle \\
&= \alpha_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle + \alpha_k \langle \nabla f(x_k), x_{k+1} - x^* \rangle \\
&\leq \alpha_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle + V_\omega(x^*, x_k) - V_\omega(x^*, x_{k+1}) - V_\omega(x_{k+1}, x_k)
\end{aligned}
$$

5. Суммируя по $k = 0, \ldots, K-1$ и используя телескопирование:

$$
\sum_{k=0}^{K-1} \alpha_k (f(x_k) - f^*) \leq V_\omega(x^*, x_0) + \sum_{k=0}^{K-1} \left[\alpha_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle - V_\omega(x_{k+1}, x_k) \right]
$$

Доказательство сходимости (продолжение)

4. Подставляя $x = x^*$ и используя выпуклость $f(x_k) - f^* \leq \langle \nabla f(x_k), x_k - x^* \rangle$:

$$
\begin{aligned}
\alpha_k(f(x_k) - f^*) &\leq \alpha_k \langle \nabla f(x_k), x_k - x^* \rangle \\
&= \alpha_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle + \alpha_k \langle \nabla f(x_k), x_{k+1} - x^* \rangle \\
&\leq \alpha_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle + V_\omega(x^*, x_k) - V_\omega(x^*, x_{k+1}) - V_\omega(x_{k+1}, x_k)
\end{aligned}
$$

5. Суммируя по $k = 0, \ldots, K-1$ и используя телескопирование:

$$
\sum_{k=0}^{K-1} \alpha_k(f(x_k) - f^*) \leq V_\omega(x^*, x_0) + \sum_{k=0}^{K-1} \left[\alpha_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle - V_\omega(x_{k+1}, x_k)\right]
$$

6. Используя шаг 2 для оценки каждого слагаемого и монотонность $f(x_k)$:

$$
K \cdot (f(x_K) - f^*) \leq \frac{L}{\sigma} V_\omega(x^*, x_0) = \frac{LR^2}{\sigma}
$$

откуда $f(x_K) - f^* \leq \frac{LR^2}{\sigma K}$, что и требовалось доказать.

Пример: зеркальный спуск vs PGD на симплексе

Рассмотрим задачу

$$
\min_{x \in \Delta_n} f(x) = \frac{1}{2}\|Ax - b\|_2^2
$$

- PGD: евклидова проекция на Δ_n, $\mathcal{O}(n \log n)$ на шаг

Пример: зеркальный спуск vs PGD на симплексе

Рассмотрим задачу

$$\min_{x \in \Delta_n} f(x) = \frac{1}{2}\|Ax - b\|_2^2$$

- PGD: евклидова проекция на Δ_n, $\mathcal{O}(n \log n)$ на шаг
- MD (EG): мультипликативное обновление, $\mathcal{O}(n)$ на шаг

Пример: зеркальный спуск vs PGD на симплексе

Рассмотрим задачу

$$
\min_{x \in \Delta_n} f(x) = \frac{1}{2}\|Ax - b\|_2^2
$$

- PGD: евклидова проекция на Δ_n, $\mathcal{O}(n \log n)$ на шаг
- MD (EG): мультипликативное обновление, $\mathcal{O}(n)$ на шаг

Пример: зеркальный спуск vs PGD на симплексе

Рассмотрим задачу

$$\min_{x \in \Delta_n} f(x) = \frac{1}{2}\|Ax - b\|_2^2$$

- PGD: евклидова проекция на Δ_n, $\mathcal{O}(n \log n)$ на шаг
- MD (EG): мультипликативное обновление, $\mathcal{O}(n)$ на шаг

Параметр	PGD	MD (EG)
Гладкость	$L_2 = \|A^\top A\|_{\text{op}}$	$L_\infty = \max_{ij} \|A^\top A\|_{ij}$
«Радиус»	$R_2^2 \leq 2$	$R^2 \leq \ln n$
Оценка	$\frac{L_2}{k}$	$\frac{L_\infty \ln n}{k}$

Когда A имеет слабую корреляцию между столбцами: $L_2 \approx n \cdot L_\infty$. Тогда:

- PGD: $\mathcal{O}\left(\frac{nL_\infty}{k}\right)$

Пример: зеркальный спуск vs PGD на симплексе

Рассмотрим задачу

$$\min_{x \in \Delta_n} f(x) = \frac{1}{2}\|Ax - b\|_2^2$$

- PGD: евклидова проекция на Δ_n, $\mathcal{O}(n \log n)$ на шаг
- MD (EG): мультипликативное обновление, $\mathcal{O}(n)$ на шаг

Параметр	PGD	MD (EG)
Гладкость	$L_2 = \|A^\top A\|_{\text{op}}$	$L_\infty = \max_{ij} \|A^\top A\|_{ij}$
«Радиус»	$R_2^2 \leq 2$	$R^2 \leq \ln n$
Оценка	$\frac{L_2}{k}$	$\frac{L_\infty \ln n}{k}$

Когда A имеет слабую корреляцию между столбцами: $L_2 \approx n \cdot L_\infty$. Тогда:

- PGD: $\mathcal{O}\left(\frac{nL_\infty}{k}\right)$
- MD: $\mathcal{O}\left(\frac{L_\infty \ln n}{k}\right)$

Пример: зеркальный спуск vs PGD на симплексе

Рассмотрим задачу

$$\min_{x \in \Delta_n} f(x) = \frac{1}{2}\|Ax - b\|_2^2$$

- PGD: евклидова проекция на Δ_n, $\mathcal{O}(n \log n)$ на шаг
- MD (EG): мультипликативное обновление, $\mathcal{O}(n)$ на шаг

Параметр	PGD	MD (EG)
Гладкость	$L_2 = \|A^\top A\|_{\text{op}}$	$L_\infty = \max_{ij} \|A^\top A\|_{ij}$
«Радиус»	$R_2^2 \leq 2$	$R^2 \leq \ln n$
Оценка	$\frac{L_2}{k}$	$\frac{L_\infty \ln n}{k}$

Когда A имеет слабую корреляцию между столбцами: $L_2 \approx n \cdot L_\infty$. Тогда:

- PGD: $\mathcal{O}\left(\frac{nL_\infty}{k}\right)$
- MD: $\mathcal{O}\left(\frac{L_\infty \ln n}{k}\right)$

Пример: зеркальный спуск vs PGD на симплексе

Рассмотрим задачу

$$\min_{x \in \Delta_n} f(x) = \frac{1}{2}\|Ax - b\|_2^2$$

- PGD: евклидова проекция на Δ_n, $\mathcal{O}(n \log n)$ на шаг
- MD (EG): мультипликативное обновление, $\mathcal{O}(n)$ на шаг

Параметр	PGD	MD (EG)
Гладкость	$L_2 = \|A^\top A\|_{\text{op}}$	$L_\infty = \max_{ij} \|A^\top A\|_{ij}$
«Радиус»	$R_2^2 \leq 2$	$R^2 \leq \ln n$
Оценка	$\frac{L_2}{k}$	$\frac{L_\infty \ln n}{k}$

Когда A имеет слабую корреляцию между столбцами: $L_2 \approx n \cdot L_\infty$. Тогда:

- PGD: $\mathcal{O}\left(\frac{nL_\infty}{k}\right)$
- MD: $\mathcal{O}\left(\frac{L_\infty \ln n}{k}\right)$

Выигрыш: $\frac{n}{\ln n}$ раз!

Пример: зеркальный спуск vs PGD на симплексе

Рассмотрим задачу

$$\min_{x\in\Delta_n} f(x) = \frac{1}{2}\|Ax - b\|_2^2$$

- PGD: евклидова проекция на Δ_n, $\mathcal{O}(n\log n)$ на шаг
- MD (EG): мультипликативное обновление, $\mathcal{O}(n)$ на шаг

Параметр	PGD	MD (EG)
Гладкость	$L_2 = \|A^\top A\|_{\text{op}}$	$L_\infty = \max_{ij} \|A^\top A\|_{ij}$
«Радиус»	$R_2^2 \leq 2$	$R^2 \leq \ln n$
Оценка	$\frac{L_2}{k}$	$\frac{L_\infty \ln n}{k}$

Когда A имеет слабую корреляцию между столбцами: $L_2 \approx n \cdot L_\infty$. Тогда:

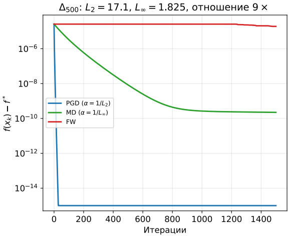
- PGD: $\mathcal{O}\left(\frac{nL_\infty}{k}\right)$
- MD: $\mathcal{O}\left(\frac{L_\infty \ln n}{k}\right)$

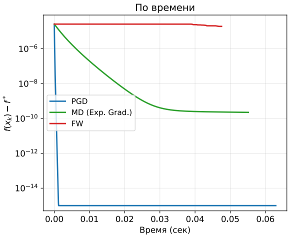
Выигрыш: $\frac{n}{\ln n}$ раз!

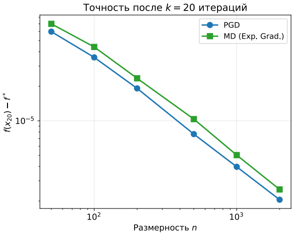
Для $n = 1000$: выигрыш ≈ 145 раз.

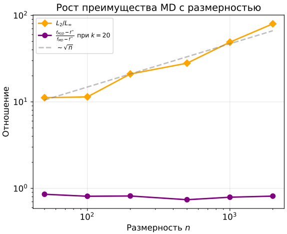
Для $n = 10^6$: выигрыш $\approx 72\,400$ раз.

Численные эксперименты: MD vs PGD vs FW

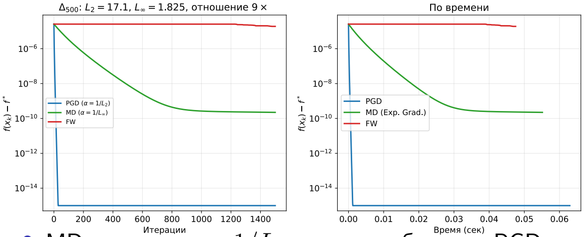


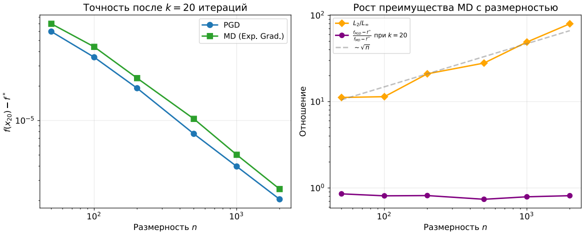




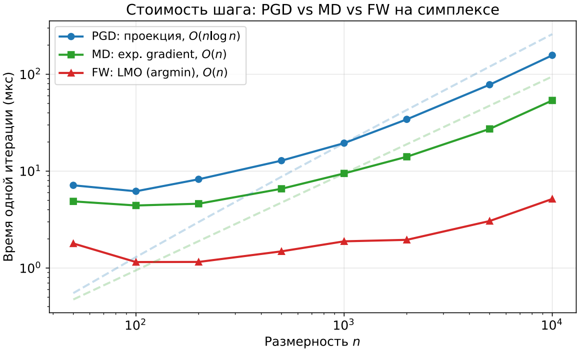


Численные эксперименты: MD vs PGD vs FW

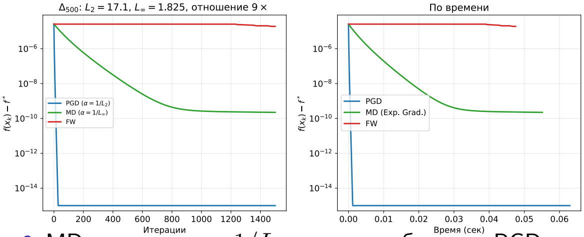


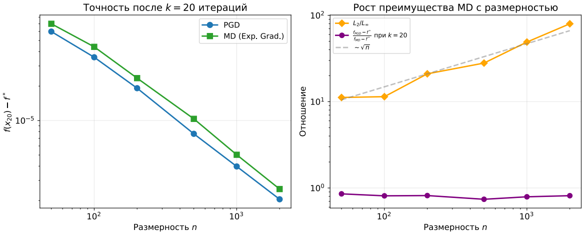


- MD с шагом $\alpha = 1/L_\infty$ сходится быстрее PGD с шагом $\alpha = 1/L_2$

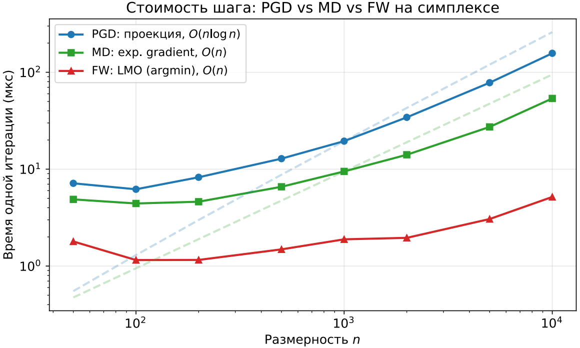


Численные эксперименты: MD vs PGD vs FW

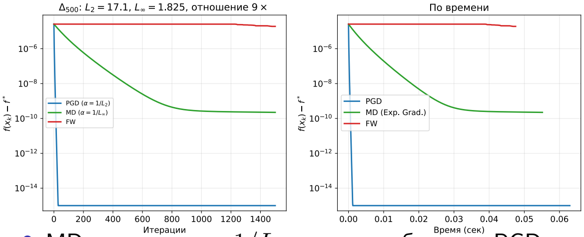




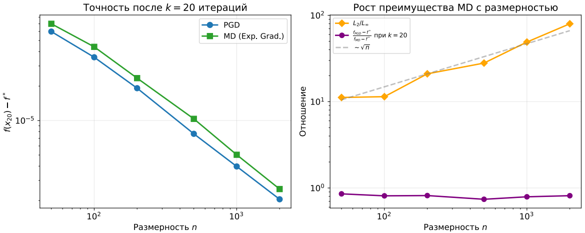
- MD с шагом $\alpha = 1/L_\infty$ сходится быстрее PGD с шагом $\alpha = 1/L_2$
- Отношение L_2/L_∞ растёт с размерностью

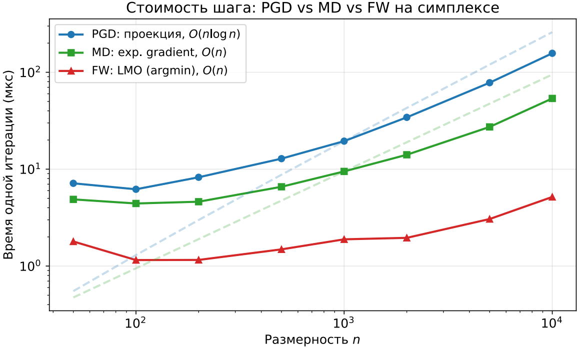


Численные эксперименты: MD vs PGD vs FW



- MD с шагом $\alpha = 1/L_\infty$ сходится быстрее PGD с шагом $\alpha = 1/L_2$
- Отношение L_2/L_∞ растёт с размерностью





- Стоимость шага MD — $\mathcal{O}(n)$, дешевле PGD —

Итоги: зеркальный спуск

- **Обобщение PGD:** заменяем евклидову проекцию дивергенцией Брэгмана, адаптируя геометрию метода к задаче.

Итоги: зеркальный спуск

- **Обобщение PGD:** заменяем евклидову проекцию дивергенцией Брэгмана, адаптируя геометрию метода к задаче.
- **Скорость сходимости:** $\mathcal{O}\left(\frac{LR^2}{\sigma k}\right)$ — та же структура, что у PGD, но с параметрами, адаптированными к «правильной» норме.

Итоги: зеркальный спуск

- **Обобщение PGD:** заменяем евклидову проекцию дивергенцией Брэгмана, адаптируя геометрию метода к задаче.
- **Скорость сходимости:** $\mathcal{O}\left(\frac{LR^2}{\sigma k}\right)$ — та же структура, что у PGD, но с параметрами, адаптированными к «правильной» норме.
- **Когда выигрывает:** задачи, где евклидова геометрия не соответствует структуре множества или функции:

Итоги: зеркальный спуск

- **Обобщение PGD:** заменяем евклидову проекцию дивергенцией Брэгмана, адаптируя геометрию метода к задаче.
- **Скорость сходимости:** $\mathcal{O}\left(\frac{LR^2}{\sigma k}\right)$ — та же структура, что у PGD, но с параметрами, адаптированными к «правильной» норме.
- **Когда выигрывает:** задачи, где евклидова геометрия не соответствует структуре множества или функции:
  - Оптимизация на **симплексе** (ℓ_1-геометрия, негативная энтропия)

Итоги: зеркальный спуск

- **Обобщение PGD:** заменяем евклидову проекцию дивергенцией Брэгмана, адаптируя геометрию метода к задаче.
- **Скорость сходимости:** $\mathcal{O}\left(\frac{LR^2}{\sigma k}\right)$ — та же структура, что у PGD, но с параметрами, адаптированными к «правильной» норме.
- **Когда выигрывает:** задачи, где евклидова геометрия не соответствует структуре множества или функции:
  - Оптимизация на **симплексе** (ℓ_1-геометрия, негативная энтропия)
  - Задачи с **разреженными** решениями

Итоги: зеркальный спуск

- **Обобщение PGD:** заменяем евклидову проекцию дивергенцией Брэгмана, адаптируя геометрию метода к задаче.
- **Скорость сходимости:** $\mathcal{O}\left(\frac{LR^2}{\sigma k}\right)$ — та же структура, что у PGD, но с параметрами, адаптированными к «правильной» норме.
- **Когда выигрывает:** задачи, где евклидова геометрия не соответствует структуре множества или функции:
  - Оптимизация на **симплексе** (ℓ_1-геометрия, негативная энтропия)
  - Задачи с **разреженными** решениями
  - Оптимизация в пространствах большой **размерности**, где ℓ_2-оценки зависят от n

Итоги: зеркальный спуск

- **Обобщение PGD:** заменяем евклидову проекцию дивергенцией Брэгмана, адаптируя геометрию метода к задаче.
- **Скорость сходимости:** $\mathcal{O}\left(\frac{LR^2}{\sigma k}\right)$ — та же структура, что у PGD, но с параметрами, адаптированными к «правильной» норме.
- **Когда выигрывает:** задачи, где евклидова геометрия не соответствует структуре множества или функции:
  - Оптимизация на **симплексе** (ℓ_1-геометрия, негативная энтропия)
  - Задачи с **разреженными** решениями
  - Оптимизация в пространствах большой **размерности**, где ℓ_2-оценки зависят от n
- **Ключевой результат на симплексе:** с прокс-функцией $\omega(x) = \sum_i x_i \ln x_i$ зеркальный спуск даёт оценку $\mathcal{O}\left(\frac{L_\infty \ln n}{k}\right)$ вместо $\mathcal{O}\left(\frac{nL_\infty}{k}\right)$ у PGD — экспоненциальный выигрыш по размерности.

Итоги: зеркальный спуск

- **Обобщение PGD:** заменяем евклидову проекцию дивергенцией Брэгмана, адаптируя геометрию метода к задаче.
- **Скорость сходимости:** $\mathcal{O}\left(\frac{LR^2}{\sigma k}\right)$ — та же структура, что у PGD, но с параметрами, адаптированными к «правильной» норме.
- **Когда выигрывает:** задачи, где евклидова геометрия не соответствует структуре множества или функции:
    - Оптимизация на **симплексе** (ℓ_1-геометрия, негативная энтропия)
    - Задачи с **разреженными** решениями
    - Оптимизация в пространствах большой **размерности**, где ℓ_2-оценки зависят от n
- **Ключевой результат на симплексе:** с прокс-функцией $\omega(x) = \sum_i x_i \ln x_i$ зеркальный спуск даёт оценку $\mathcal{O}\left(\frac{L_\infty \ln n}{k}\right)$ вместо $\mathcal{O}\left(\frac{nL_\infty}{k}\right)$ у PGD — экспоненциальный выигрыш по размерности.
- Выбор прокс-функции ω — ключевое решение при применении метода. Общее правило: ω должна быть сильно выпуклой относительно нормы, в которой f является гладкой.