



Нижние оценки для градиентного спуска.
Ускоренный градиентный спуск. Момент.
Ускорение Нестерова

Даня Меркулов

Оптимизация для всех! ЦУ

Сходимость градиентного спуска

Градиентный спуск:

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

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k)$$

выпуклая (негладкая)	гладкая (невыпуклая)	гладкая & выпуклая	гладкая & строго выпуклая (или PL)
$f(x^k) - f^* \sim \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\ \nabla f(x^k)\ ^2 \sim \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$f(x^k) - f^* \sim \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\ x^k - x^*\ ^2 \sim \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon}\right)$

Сходимость градиентного спуска

Градиентный спуск:

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

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k)$$

выпуклая (негладкая)	гладкая (невыпуклая)	гладкая & выпуклая	гладкая & строго выпуклая (или PL)
$f(x^k) - f^* \sim \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$\ \nabla f(x^k)\ ^2 \sim \mathcal{O}\left(\frac{1}{k}\right)$	$f(x^k) - f^* \sim \mathcal{O}\left(\frac{1}{k}\right)$	$\ x^k - x^*\ ^2 \sim \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$
$k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$k_\varepsilon \sim \mathcal{O}\left(n \log \frac{1}{\varepsilon}\right)$

Для гладкой строго выпуклой функции мы имеем:

$$f(x^k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x^0) - f^*).$$

Обратите внимание, что для любого x , поскольку e^{-x} выпуклая и $1 - x$ является ее касательной в точке $x = 0$, мы имеем:

$$1 - x \leq e^{-x}$$

Сходимость градиентного спуска

Градиентный спуск:

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

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k)$$

выпуклая (негладкая)	гладкая (невыпуклая)	гладкая & выпуклая	гладкая & строго выпуклая (или PL)
$f(x^k) - f^* \sim \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$\ \nabla f(x^k)\ ^2 \sim \mathcal{O}\left(\frac{1}{k}\right)$	$f(x^k) - f^* \sim \mathcal{O}\left(\frac{1}{k}\right)$	$\ x^k - x^*\ ^2 \sim \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$
$k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$k_\varepsilon \sim \mathcal{O}\left(\kappa \log \frac{1}{\varepsilon}\right)$

Для гладкой строго выпуклой функции мы имеем:

$$f(x^k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x^0) - f^*).$$

Обратите внимание, что для любого x , поскольку e^{-x} выпуклая и $1 - x$ является ее касательной в точке $x = 0$, мы имеем:

$$1 - x \leq e^{-x}$$

Наконец:

$$\varepsilon = f(x^{k_\varepsilon}) - f^* \leq \left(1 - \frac{\mu}{L}\right)^{k_\varepsilon} (f(x^0) - f^*)$$

$$\leq \exp\left(-k_\varepsilon \frac{\mu}{L}\right) (f(x^0) - f^*)$$

$$k_\varepsilon \geq \kappa \log \frac{f(x^0) - f^*}{\varepsilon} = \mathcal{O}\left(\kappa \log \frac{1}{\varepsilon}\right)$$

Сходимость градиентного спуска

Вопрос: Можно ли добиться лучшей скорости сходимости, используя только информацию первого порядка?

Сходимость градиентного спуска

Вопрос: Можно ли добиться лучшей скорости сходимости, используя только информацию первого порядка?
Да, можно.

Нижние оценки

Нижние оценки

выпуклая (негладкая)	гладкая (невыпуклая) ¹	гладкая & выпуклая ²	гладкая & строго выпуклая (или PL)
$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\mathcal{O}\left(\frac{1}{k^2}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}}\right)$	$\mathcal{O}\left(\frac{1}{k^2}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}}\right)$	$\mathcal{O}\left(\left(1 - \sqrt{\frac{\mu}{L}}\right)^k\right)$ $k_\varepsilon \sim \mathcal{O}\left(\sqrt{n} \log \frac{1}{\varepsilon}\right)$

¹Carmon, Duchi, Hinder, Sidford, 2017

²Nemirovski, Yudin, 1979

Черный ящик

Итерация градиентного спуска:

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) \\&= x^{k-1} - \alpha^{k-1} \nabla f(x^{k-1}) - \alpha^k \nabla f(x^k) \\&\vdots \\&= x^0 - \sum_{i=0}^k \alpha^{k-i} \nabla f(x^{k-i})\end{aligned}$$

Черный ящик

Итерация градиентного спуска:

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) \\&= x^{k-1} - \alpha^{k-1} \nabla f(x^{k-1}) - \alpha^k \nabla f(x^k) \\&\vdots \\&= x^0 - \sum_{i=0}^k \alpha^{k-i} \nabla f(x^{k-i})\end{aligned}$$

Рассмотрим семейство методов первого порядка, где

$$\begin{aligned}x^{k+1} &\in x^0 + \text{Lin} \{ \nabla f(x^0), \nabla f(x^1), \dots, \nabla f(x^k) \} && f - \text{гладкая} \\x^{k+1} &\in x^0 + \text{Lin} \{ g_0, g_1, \dots, g_k \}, \text{ где } g_i \in \partial f(x^i) && f - \text{негладкая}\end{aligned} \tag{1}$$

Черный ящик

Итерация градиентного спуска:

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) \\&= x^{k-1} - \alpha^{k-1} \nabla f(x^{k-1}) - \alpha^k \nabla f(x^k) \\&\vdots \\&= x^0 - \sum_{i=0}^k \alpha^{k-i} \nabla f(x^{k-i})\end{aligned}$$

Рассмотрим семейство методов первого порядка, где

$$\begin{aligned}x^{k+1} &\in x^0 + \text{Lin} \{ \nabla f(x^0), \nabla f(x^1), \dots, \nabla f(x^k) \} && f - \text{гладкая} \\x^{k+1} &\in x^0 + \text{Lin} \{ g_0, g_1, \dots, g_k \}, \text{ где } g_i \in \partial f(x^i) && f - \text{негладкая}\end{aligned} \tag{1}$$

Чтобы построить нижнюю оценку, нам нужно найти функцию f из соответствующего класса такую, что любой метод из семейства 1 будет работать не быстрее, чем нижняя оценка.

Гладкий случай

i Theorem

Существует функция f , которая является L -гладкой и выпуклой, так что любой метод 1 для любого $k : 1 \leq k \leq \frac{n-1}{2}$ удовлетворяет :

$$f(x^k) - f^* \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

Гладкий случай

i Theorem

Существует функция f , которая является L -гладкой и выпуклой, так что любой метод 1 для любого $k : 1 \leq k \leq \frac{n-1}{2}$ удовлетворяет :

$$f(x^k) - f^* \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

- Неважно, какой метод градиентного спуска вы используете, всегда существует функция f , при применении на ней вашего метода градиентного спуска, нижняя оценка скорости сходимости $\mathcal{O}\left(\frac{1}{k^2}\right)$.

Гладкий случай

i Theorem

Существует функция f , которая является L -гладкой и выпуклой, так что любой метод 1 для любого $k : 1 \leq k \leq \frac{n-1}{2}$ удовлетворяет :

$$f(x^k) - f^* \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

- Неважно, какой метод градиентного спуска вы используете, всегда существует функция f , при применении на ней вашего метода градиентного спуска, нижняя оценка скорости сходимости $\mathcal{O}\left(\frac{1}{k^2}\right)$.
- Ключом к доказательству является явное построение специальной функции f .

Гладкий случай

i Theorem

Существует функция f , которая является L -гладкой и выпуклой, так что любой метод 1 для любого $k : 1 \leq k \leq \frac{n-1}{2}$ удовлетворяет :

$$f(x^k) - f^* \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

- Неважно, какой метод градиентного спуска вы используете, всегда существует функция f , при применении на ней вашего метода градиентного спуска, нижняя оценка скорости сходимости $\mathcal{O}\left(\frac{1}{k^2}\right)$.
- Ключом к доказательству является явное построение специальной функции f .
- Обратите внимание, что эта граница $\mathcal{O}\left(\frac{1}{k^2}\right)$ не соответствует скорости градиентного спуска $\mathcal{O}\left(\frac{1}{k}\right)$. Два возможных варианта:

Гладкий случай

i Theorem

Существует функция f , которая является L -гладкой и выпуклой, так что любой метод 1 для любого $k : 1 \leq k \leq \frac{n-1}{2}$ удовлетворяет :

$$f(x^k) - f^* \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

- Неважно, какой метод градиентного спуска вы используете, всегда существует функция f , при применении на ней вашего метода градиентного спуска, нижняя оценка скорости сходимости $\mathcal{O}\left(\frac{1}{k^2}\right)$.
- Ключом к доказательству является явное построение специальной функции f .
- Обратите внимание, что эта граница $\mathcal{O}\left(\frac{1}{k^2}\right)$ не соответствует скорости градиентного спуска $\mathcal{O}\left(\frac{1}{k}\right)$. Два возможных варианта:
 - a. Нижняя оценка не является точной.

Гладкий случай

i Theorem

Существует функция f , которая является L -гладкой и выпуклой, так что любой метод 1 для любого $k : 1 \leq k \leq \frac{n-1}{2}$ удовлетворяет :

$$f(x^k) - f^* \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

- Неважно, какой метод градиентного спуска вы используете, всегда существует функция f , при применении на ней вашего метода градиентного спуска, нижняя оценка скорости сходимости $\mathcal{O}\left(\frac{1}{k^2}\right)$.
- Ключом к доказательству является явное построение специальной функции f .
- Обратите внимание, что эта граница $\mathcal{O}\left(\frac{1}{k^2}\right)$ не соответствует скорости градиентного спуска $\mathcal{O}\left(\frac{1}{k}\right)$. Два возможных варианта:
 - a. Нижняя оценка не является точной.
 - b. Метод градиентного спуска не является оптимальным для этой задачи.

Гладкий случай

i Theorem

Существует функция f , которая является L -гладкой и выпуклой, так что любой метод 1 для любого $k : 1 \leq k \leq \frac{n-1}{2}$ удовлетворяет :

$$f(x^k) - f^* \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

- Неважно, какой метод градиентного спуска вы используете, всегда существует функция f , при применении на ней вашего метода градиентного спуска, нижняя оценка скорости сходимости $\mathcal{O}\left(\frac{1}{k^2}\right)$.
- Ключом к доказательству является явное построение специальной функции f .
- Обратите внимание, что эта граница $\mathcal{O}\left(\frac{1}{k^2}\right)$ не соответствует скорости градиентного спуска $\mathcal{O}\left(\frac{1}{k}\right)$. Два возможных варианта:
 - a. Нижняя оценка не является точной.
 - b. Метод градиентного спуска не является оптимальным для этой задачи.

Наихудшая функция Нестерова

- Пусть $n = 2k + 1$ и $A \in \mathbb{R}^{n \times n}$.

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ 0 & 0 & -1 & 2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 2 \end{bmatrix}$$

Наихудшая функция Нестерова

- Пусть $n = 2k + 1$ и $A \in \mathbb{R}^{n \times n}$.

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ 0 & 0 & -1 & 2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 2 \end{bmatrix}$$

- Обратите внимание, что

$$x^T A x = x_1^2 + x_n^2 + \sum_{i=1}^{n-1} (x_i - x_{i+1})^2,$$

Следовательно, $x^T A x \geq 0$. Также легко увидеть, что $0 \preceq A \preceq 4I$.

Наихудшая функция Нестерова

- Пусть $n = 2k + 1$ и $A \in \mathbb{R}^{n \times n}$.

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ 0 & 0 & -1 & 2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 2 \end{bmatrix}$$

- Обратите внимание, что

$$x^T A x = x_1^2 + x_n^2 + \sum_{i=1}^{n-1} (x_i - x_{i+1})^2,$$

Следовательно, $x^T A x \geq 0$. Также легко увидеть, что $0 \preceq A \preceq 4I$.

Наихудшая функция Нестерова

- Пусть $n = 2k + 1$ и $A \in \mathbb{R}^{n \times n}$.

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ 0 & 0 & -1 & 2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 2 \end{bmatrix}$$

- Обратите внимание, что

$$x^T A x = x_1^2 + x_n^2 + \sum_{i=1}^{n-1} (x_i - x_{i+1})^2,$$

Следовательно, $x^T A x \geq 0$. Также легко увидеть, что $0 \preceq A \preceq 4I$.

Пример, когда $n = 3$:

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

Наихудшая функция Нестерова

- Пусть $n = 2k + 1$ и $A \in \mathbb{R}^{n \times n}$.

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2 \end{bmatrix}$$

- Обратите внимание, что

$$x^T A x = x_1^2 + x_n^2 + \sum_{i=1}^{n-1} (x_i - x_{i+1})^2,$$

Следовательно, $x^T A x \geq 0$. Также легко увидеть, что $0 \preceq A \preceq 4I$.

Пример, когда $n = 3$:

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

Нижняя оценка:

$$\begin{aligned} x^T A x &= 2x_1^2 + 2x_2^2 + 2x_3^2 - 2x_1x_2 - 2x_2x_3 \\ &= x_1^2 + x_1^2 - 2x_1x_2 + x_2^2 + x_2^2 - 2x_2x_3 + x_3^2 + x_3^2 \\ &= x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + x_3^2 \geq 0 \end{aligned}$$

Наихудшая функция Нестерова

- Пусть $n = 2k + 1$ и $A \in \mathbb{R}^{n \times n}$.

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ 0 & 0 & -1 & 2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 2 \end{bmatrix}$$

- Обратите внимание, что

$$x^T A x = x_1^2 + x_n^2 + \sum_{i=1}^{n-1} (x_i - x_{i+1})^2,$$

Следовательно, $x^T A x \geq 0$. Также легко увидеть, что $0 \preceq A \preceq 4I$.

Пример, когда $n = 3$:

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

Нижняя оценка:

$$\begin{aligned} x^T A x &= 2x_1^2 + 2x_2^2 + 2x_3^2 - 2x_1x_2 - 2x_2x_3 \\ &= x_1^2 + x_1^2 - 2x_1x_2 + x_2^2 + x_2^2 - 2x_2x_3 + x_3^2 + x_3^2 \\ &= x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + x_3^2 \geq 0 \end{aligned}$$

Верхняя оценка

$$\begin{aligned} x^T A x &= 2x_1^2 + 2x_2^2 + 2x_3^2 - 2x_1x_2 - 2x_2x_3 \\ &\leq 4(x_1^2 + x_2^2 + x_3^2) \\ 0 &\leq 2x_1^2 + 2x_2^2 + 2x_3^2 + 2x_1x_2 + 2x_2x_3 \\ 0 &\leq x_1^2 + x_1^2 + 2x_1x_2 + x_2^2 + x_2^2 + 2x_2x_3 + x_3^2 + x_3^2 \\ 0 &\leq x_1^2 + (x_1 + x_2)^2 + (x_2 + x_3)^2 + x_3^2 \end{aligned}$$

Наихудшая функция Нестерова

- Определим следующую L -гладкую выпуклую функцию: $f(x) = \frac{L}{4} \left(\frac{1}{2} x^T A x - e_1^T x \right) = \frac{L}{8} x^T A x - \frac{L}{4} e_1^T x$.

Наихудшая функция Нестерова

- Определим следующую L -гладкую выпуклую функцию: $f(x) = \frac{L}{4} \left(\frac{1}{2} x^T A x - e_1^T x \right) = \frac{L}{8} x^T A x - \frac{L}{4} e_1^T x$.
- Оптимальное решение x^* удовлетворяет $Ax^* = e_1$, и решение этой системы уравнений дает:

$$\begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2 \end{bmatrix} \begin{bmatrix} x_1^* \\ x_2^* \\ x_3^* \\ \vdots \\ x_n^* \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \begin{cases} 2x_1^* - x_2^* = 1 \\ -x_i^* + 2x_{i+1}^* - x_{i+2}^* = 0, \quad i = 2, \dots, n-1 \\ -x_{n-1}^* + 2x_n^* = 0 \end{cases}$$

Наихудшая функция Нестерова

- Определим следующую L -гладкую выпуклую функцию: $f(x) = \frac{L}{4} (\frac{1}{2}x^T Ax - e_1^T x) = \frac{L}{8}x^T Ax - \frac{L}{4}e_1^T x$.
- Оптимальное решение x^* удовлетворяет $Ax^* = e_1$, и решение этой системы уравнений дает:

$$\begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2 \end{bmatrix} \begin{bmatrix} x_1^* \\ x_2^* \\ x_3^* \\ \vdots \\ x_n^* \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \begin{cases} 2x_1^* - x_2^* = 1 \\ -x_i^* + 2x_{i+1}^* - x_{i+2}^* = 0, \quad i = 2, \dots, n-1 \\ -x_{n-1}^* + 2x_n^* = 0 \end{cases}$$

- Гипотеза: $x_i^* = a + bi$ (вдохновленная физикой). Проверьте, что удовлетворяется второе уравнение, в то время как a и b вычисляются из первого и последнего уравнений.

Наихудшая функция Нестерова

- Определим следующую L -гладкую выпуклую функцию: $f(x) = \frac{L}{4} (\frac{1}{2}x^T Ax - e_1^T x) = \frac{L}{8}x^T Ax - \frac{L}{4}e_1^T x$.
- Оптимальное решение x^* удовлетворяет $Ax^* = e_1$, и решение этой системы уравнений дает:

$$\begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2 \end{bmatrix} \begin{bmatrix} x_1^* \\ x_2^* \\ x_3^* \\ \vdots \\ x_n^* \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \begin{cases} 2x_1^* - x_2^* = 1 \\ -x_i^* + 2x_{i+1}^* - x_{i+2}^* = 0, \quad i = 2, \dots, n-1 \\ -x_{n-1}^* + 2x_n^* = 0 \end{cases}$$

- Гипотеза: $x_i^* = a + bi$ (вдохновленная физикой). Проверьте, что удовлетворяется второе уравнение, в то время как a и b вычисляются из первого и последнего уравнений.
- Решение:

$$x_i^* = 1 - \frac{i}{n+1},$$

Наихудшая функция Нестерова

- Определим следующую L -гладкую выпуклую функцию: $f(x) = \frac{L}{4} (\frac{1}{2}x^T Ax - e_1^T x) = \frac{L}{8}x^T Ax - \frac{L}{4}e_1^T x$.
- Оптимальное решение x^* удовлетворяет $Ax^* = e_1$, и решение этой системы уравнений дает:

$$\begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2 \end{bmatrix} \begin{bmatrix} x_1^* \\ x_2^* \\ x_3^* \\ \vdots \\ x_n^* \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \begin{cases} 2x_1^* - x_2^* = 1 \\ -x_i^* + 2x_{i+1}^* - x_{i+2}^* = 0, \quad i = 2, \dots, n-1 \\ -x_{n-1}^* + 2x_n^* = 0 \end{cases}$$

- Гипотеза: $x_i^* = a + bi$ (вдохновленная физикой). Проверьте, что удовлетворяется второе уравнение, в то время как a и b вычисляются из первого и последнего уравнений.
- Решение:

$$x_i^* = 1 - \frac{i}{n+1},$$

- И целевое значение равно

$$f(x^*) = \frac{L}{8}x^{*T}Ax^* - \frac{L}{4}\langle x^*, e_1 \rangle = -\frac{L}{8}\langle x^*, e_1 \rangle = -\frac{L}{8} \left(1 - \frac{1}{n+1} \right).$$

Гладкий случай (доказательство)

- Предположим, что мы начинаем с $x^0 = 0$. Запросив у оракула градиент, мы получаем $g_0 = -e_1$. Тогда, x^1 должен лежать на линии, генерируемой e_1 . В этой точке все компоненты x^1 равны нулю, кроме первой, поэтому

$$x^1 = \begin{bmatrix} \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Гладкий случай (доказательство)

- Предположим, что мы начинаем с $x^0 = 0$.

Запросив у оракула градиент, мы получаем $g_0 = -e_1$. Тогда, x^1 должен лежать на линии, генерируемой e_1 . В этой точке все компоненты x^1 равны нулю, кроме первой, поэтому

$$x^1 = \begin{bmatrix} \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

- На второй итерации мы снова запрашиваем у оракула градиент и получаем $g_1 = Ax^1 - e_1$. Тогда, x^2 должен лежать на линии, генерируемой e_1 и $Ax^1 - e_1$. Все компоненты x^2 равны нулю, кроме первых двух, поэтому

$$\begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2 \end{bmatrix} \begin{bmatrix} \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow x^2 = \begin{bmatrix} \bullet \\ \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Гладкий случай (доказательство)

- Предположим, что мы начинаем с $x^0 = 0$.

Запросив у оракула градиент, мы получаем $g_0 = -e_1$. Тогда, x^1 должен лежать на линии, генерируемой e_1 . В этой точке все компоненты x^1 равны нулю, кроме первой, поэтому

$$x^1 = \begin{bmatrix} \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

- На второй итерации мы снова запрашиваем у оракула градиент и получаем $g_1 = Ax^1 - e_1$. Тогда, x^2 должен лежать на линии, генерируемой e_1 и $Ax^1 - e_1$. Все компоненты x^2 равны нулю, кроме первых двух, поэтому

$$\begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2 \end{bmatrix} \begin{bmatrix} \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow x^2 = \begin{bmatrix} \bullet \\ \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Гладкий случай (доказательство)

- Предположим, что мы начинаем с $x^0 = 0$. Запросив у оракула градиент, мы получаем $g_0 = -e_1$. Тогда, x^1 должен лежать на линии, генерируемой e_1 . В этой точке все компоненты x^1 равны нулю, кроме первой, поэтому

$$x^1 = \begin{bmatrix} \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

- На второй итерации мы снова запрашиваем у оракула градиент и получаем $g_1 = Ax^1 - e_1$. Тогда, x^2 должен лежать на линии, генерируемой e_1 и $Ax^1 - e_1$. Все компоненты x^2 равны нулю, кроме первых двух, поэтому

$$\begin{bmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ 0 & -1 & 2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 2 \end{bmatrix} \begin{bmatrix} \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow x^2 = \begin{bmatrix} \bullet \\ \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

- Из-за структуры матрицы A можно показать, что после k итераций все последние $n - k$ компоненты x^k равны нулю.

$$x^{(k)} = \begin{bmatrix} \bullet \\ \bullet \\ \vdots \\ \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ \vdots \\ k \\ k+1 \\ \vdots \\ n \end{matrix}$$

Гладкий случай (доказательство)

- Предположим, что мы начинаем с $x^0 = 0$. Запросив у оракула градиент, мы получаем $g_0 = -e_1$. Тогда, x^1 должен лежать на линии, генерируемой e_1 . В этой точке все компоненты x^1 равны нулю, кроме первой, поэтому

$$x^1 = \begin{bmatrix} \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

- На второй итерации мы снова запрашиваем у оракула градиент и получаем $g_1 = Ax^1 - e_1$. Тогда, x^2 должен лежать на линии, генерируемой e_1 и $Ax^1 - e_1$. Все компоненты x^2 равны нулю, кроме первых двух, поэтому

$$\begin{bmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ 0 & -1 & 2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 2 \end{bmatrix} \begin{bmatrix} \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow x^2 = \begin{bmatrix} \bullet \\ \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

- Из-за структуры матрицы A можно показать, что после k итераций все последние $n - k$ компоненты x^k равны нулю.

$$x^{(k)} = \begin{bmatrix} \bullet \\ \bullet \\ \vdots \\ \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ \vdots \\ k \\ k+1 \\ \vdots \\ n \end{matrix}$$

- Однако, поскольку каждая итерация x^k , произведенная нашим методом, лежит в $S_k = \text{Lin}\{e_1, e_2, \dots, e_k\}$ (т.е. имеет нули в координатах $k+1, \dots, n$), она не может “достичь” полного оптимального вектора x^* . Другими словами, даже если бы мы выбрали лучший возможный вектор из S_k , обозначаемый

$$\tilde{x}^k = \arg \min_{x \in S_k} f(x),$$

его целевое значение $f(\tilde{x}^k)$ будет строго хуже, чем $f(x^*)$.

Гладкий случай (доказательство)

- Поскольку $x^k \in S_k = \text{Lin}\{e_1, e_2, \dots, e_k\}$ и \tilde{x}^k является лучшим возможным приближением к x^* в S_k , мы имеем

$$f(x^k) \geq f(\tilde{x}^k).$$

Гладкий случай (доказательство)

- Поскольку $x^k \in S_k = \text{Lin}\{e_1, e_2, \dots, e_k\}$ и \tilde{x}^k является лучшим возможным приближением к x^* в S_k , мы имеем

$$f(x^k) \geq f(\tilde{x}^k).$$

- Следовательно,

$$f(x^k) - f(x^*) \geq f(\tilde{x}^k) - f(x^*).$$

Гладкий случай (доказательство)

- Поскольку $x^k \in S_k = \text{Lin}\{e_1, e_2, \dots, e_k\}$ и \tilde{x}^k является лучшим возможным приближением к x^* в S_k , мы имеем

$$f(x^k) \geq f(\tilde{x}^k).$$

- Следовательно,

$$f(x^k) - f(x^*) \geq f(\tilde{x}^k) - f(x^*).$$

- Аналогично, для оптимума исходной функции, мы имеем $\tilde{x}_i^k = 1 - \frac{i}{k+1}$ и $f(\tilde{x}^k) = -\frac{L}{8} \left(1 - \frac{1}{k+1}\right)$.

Гладкий случай (доказательство)

- Поскольку $x^k \in S_k = \text{Lin}\{e_1, e_2, \dots, e_k\}$ и \tilde{x}^k является лучшим возможным приближением к x^* в S_k , мы имеем

$$f(x^k) \geq f(\tilde{x}^k).$$

- Следовательно,

$$f(x^k) - f(x^*) \geq f(\tilde{x}^k) - f(x^*).$$

- Аналогично, для оптимума исходной функции, мы имеем $\tilde{x}_i^k = 1 - \frac{i}{k+1}$ и $f(\tilde{x}^k) = -\frac{L}{8} \left(1 - \frac{1}{k+1}\right)$.
- Теперь мы имеем:

$$f(x^k) - f(x^*) \geq f(\tilde{x}^k) - f(x^*)$$

(2)

Гладкий случай (доказательство)

- Поскольку $x^k \in S_k = \text{Lin}\{e_1, e_2, \dots, e_k\}$ и \tilde{x}^k является лучшим возможным приближением к x^* в S_k , мы имеем

$$f(x^k) \geq f(\tilde{x}^k).$$

- Следовательно,

$$f(x^k) - f(x^*) \geq f(\tilde{x}^k) - f(x^*).$$

- Аналогично, для оптимума исходной функции, мы имеем $\tilde{x}_i^k = 1 - \frac{i}{k+1}$ и $f(\tilde{x}^k) = -\frac{L}{8} \left(1 - \frac{1}{k+1}\right)$.
- Теперь мы имеем:

$$\begin{aligned} f(x^k) - f(x^*) &\geq f(\tilde{x}^k) - f(x^*) \\ &= -\frac{L}{8} \left(1 - \frac{1}{k+1}\right) - \left(-\frac{L}{8} \left(1 - \frac{1}{n+1}\right)\right) \end{aligned} \tag{2}$$

Гладкий случай (доказательство)

- Поскольку $x^k \in S_k = \text{Lin}\{e_1, e_2, \dots, e_k\}$ и \tilde{x}^k является лучшим возможным приближением к x^* в S_k , мы имеем

$$f(x^k) \geq f(\tilde{x}^k).$$

- Следовательно,

$$f(x^k) - f(x^*) \geq f(\tilde{x}^k) - f(x^*).$$

- Аналогично, для оптимума исходной функции, мы имеем $\tilde{x}_i^k = 1 - \frac{i}{k+1}$ и $f(\tilde{x}^k) = -\frac{L}{8} \left(1 - \frac{1}{k+1}\right)$.
- Теперь мы имеем:

$$\begin{aligned} f(x^k) - f(x^*) &\geq f(\tilde{x}^k) - f(x^*) \\ &= -\frac{L}{8} \left(1 - \frac{1}{k+1}\right) - \left(-\frac{L}{8} \left(1 - \frac{1}{n+1}\right)\right) \\ &= \frac{L}{8} \left(\frac{1}{k+1} - \frac{1}{n+1}\right) = \frac{L}{8} \left(\frac{n-k}{(k+1)(n+1)}\right) \end{aligned} \tag{2}$$

Гладкий случай (доказательство)

- Поскольку $x^k \in S_k = \text{Lin}\{e_1, e_2, \dots, e_k\}$ и \tilde{x}^k является лучшим возможным приближением к x^* в S_k , мы имеем

$$f(x^k) \geq f(\tilde{x}^k).$$

- Следовательно,

$$f(x^k) - f(x^*) \geq f(\tilde{x}^k) - f(x^*).$$

- Аналогично, для оптимума исходной функции, мы имеем $\tilde{x}_i^k = 1 - \frac{i}{k+1}$ и $f(\tilde{x}^k) = -\frac{L}{8} \left(1 - \frac{1}{k+1}\right)$.
- Теперь мы имеем:

$$\begin{aligned} f(x^k) - f(x^*) &\geq f(\tilde{x}^k) - f(x^*) \\ &= -\frac{L}{8} \left(1 - \frac{1}{k+1}\right) - \left(-\frac{L}{8} \left(1 - \frac{1}{n+1}\right)\right) \\ &= \frac{L}{8} \left(\frac{1}{k+1} - \frac{1}{n+1}\right) = \frac{L}{8} \left(\frac{n-k}{(k+1)(n+1)}\right) \\ &\stackrel{n=2k+1}{=} \frac{L}{16(k+1)} \end{aligned} \tag{2}$$

Гладкий случай (доказательство)

- Теперь мы ограничиваем $R = \|x^0 - x^*\|_2$:

$$\|x^0 - x^*\|_2^2 = \|0 - x^*\|_2^2 = \|x^*\|_2^2 = \sum_{i=1}^n \left(1 - \frac{i}{n+1}\right)^2$$

Заметим, что

$$\begin{aligned}\sum_{i=1}^n i &= \frac{n(n+1)}{2} \\ \sum_{i=1}^n i^2 &= \frac{n(n+1)(2n+1)}{6} \\ &\leq \frac{(n+1)^3}{3}\end{aligned}$$

Гладкий случай (доказательство)

- Теперь мы ограничиваем $R = \|x^0 - x^*\|_2$:

$$\begin{aligned}\|x^0 - x^*\|_2^2 &= \|0 - x^*\|_2^2 = \|x^*\|_2^2 = \sum_{i=1}^n \left(1 - \frac{i}{n+1}\right)^2 \\ &= n - \frac{2}{n+1} \sum_{i=1}^n i + \frac{1}{(n+1)^2} \sum_{i=1}^n i^2\end{aligned}$$

Заметим, что

$$\begin{aligned}\sum_{i=1}^n i &= \frac{n(n+1)}{2} \\ \sum_{i=1}^n i^2 &= \frac{n(n+1)(2n+1)}{6} \\ &\leq \frac{(n+1)^3}{3}\end{aligned}$$

Гладкий случай (доказательство)

- Теперь мы ограничиваем $R = \|x^0 - x^*\|_2$:

$$\begin{aligned}\|x^0 - x^*\|_2^2 &= \|0 - x^*\|_2^2 = \|x^*\|_2^2 = \sum_{i=1}^n \left(1 - \frac{i}{n+1}\right)^2 \\&= n - \frac{2}{n+1} \sum_{i=1}^n i + \frac{1}{(n+1)^2} \sum_{i=1}^n i^2 \\&\leq n - \frac{2}{n+1} \cdot \frac{n(n+1)}{2} + \frac{1}{(n+1)^2} \cdot \frac{(n+1)^3}{3}\end{aligned}$$

Заметим, что

$$\begin{aligned}\sum_{i=1}^n i &= \frac{n(n+1)}{2} \\ \sum_{i=1}^n i^2 &= \frac{n(n+1)(2n+1)}{6} \\ &\leq \frac{(n+1)^3}{3}\end{aligned}$$

Гладкий случай (доказательство)

- Теперь мы ограничиваем $R = \|x^0 - x^*\|_2$:

$$\begin{aligned}\|x^0 - x^*\|_2^2 &= \|0 - x^*\|_2^2 = \|x^*\|_2^2 = \sum_{i=1}^n \left(1 - \frac{i}{n+1}\right)^2 \\&= n - \frac{2}{n+1} \sum_{i=1}^n i + \frac{1}{(n+1)^2} \sum_{i=1}^n i^2 \\&\leq n - \frac{2}{n+1} \cdot \frac{n(n+1)}{2} + \frac{1}{(n+1)^2} \cdot \frac{(n+1)^3}{3} \\&= \frac{n+1}{3} \stackrel{n=2k+1}{=} \frac{2(k+1)}{3}.\end{aligned}$$

Заметим, что

$$\begin{aligned}\sum_{i=1}^n i &= \frac{n(n+1)}{2} \\ \sum_{i=1}^n i^2 &= \frac{n(n+1)(2n+1)}{6} \\ &\leq \frac{(n+1)^3}{3}\end{aligned}$$

Гладкий случай (доказательство)

- Теперь мы ограничиваем $R = \|x^0 - x^*\|_2$:

$$\begin{aligned}\|x^0 - x^*\|_2^2 &= \|0 - x^*\|_2^2 = \|x^*\|_2^2 = \sum_{i=1}^n \left(1 - \frac{i}{n+1}\right)^2 \\&= n - \frac{2}{n+1} \sum_{i=1}^n i + \frac{1}{(n+1)^2} \sum_{i=1}^n i^2 \\&\leq n - \frac{2}{n+1} \cdot \frac{n(n+1)}{2} + \frac{1}{(n+1)^2} \cdot \frac{(n+1)^3}{3} \\&= \frac{n+1}{3} \stackrel{n=2k+1}{=} \frac{2(k+1)}{3}.\end{aligned}$$

- Следовательно,

$$k+1 \geq \frac{3}{2} \|x^0 - x^*\|_2^2 = \frac{3}{2} R^2 \quad (3)$$

Заметим, что

$$\begin{aligned}\sum_{i=1}^n i &= \frac{n(n+1)}{2} \\ \sum_{i=1}^n i^2 &= \frac{n(n+1)(2n+1)}{6} \\ &\leq \frac{(n+1)^3}{3}\end{aligned}$$

Гладкий случай (доказательство)

Наконец, используя (2) и (3), мы получаем:

$$\begin{aligned} f(x^k) - f(x^*) &\geq \frac{L}{16(k+1)} = \frac{L(k+1)}{16(k+1)^2} \\ &\geq \frac{L}{16(k+1)^2} \frac{3}{2} R^2 \\ &= \frac{3LR^2}{32(k+1)^2} \end{aligned}$$

Гладкий случай (доказательство)

Наконец, используя (2) и (3), мы получаем:

$$\begin{aligned} f(x^k) - f(x^*) &\geq \frac{L}{16(k+1)} = \frac{L(k+1)}{16(k+1)^2} \\ &\geq \frac{L}{16(k+1)^2} \frac{3}{2} R^2 \\ &= \frac{3LR^2}{32(k+1)^2} \end{aligned}$$

Это завершает доказательство с желаемой скоростью $\mathcal{O}\left(\frac{1}{k^2}\right)$.

Нижние оценки для гладкого случая

i Гладкий выпуклый случай

Существует функция f , которая является L -гладкой и выпуклой, так что любой метод 1 для любого $k : 1 \leq k \leq \frac{n-1}{2}$ удовлетворяет:

$$f(x^k) - f^* \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

i Гладкий строго выпуклый случай

Для любого x^0 и любого $\mu > 0, \kappa = \frac{L}{\mu} > 1$, существует функция f , которая является L -гладкой и μ -строго выпуклой, так что для любого метода из формы 1 удовлетворяет:

$$\begin{aligned}\|x^k - x^*\|_2 &\geq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x^0 - x^*\|_2 \\ f(x^k) - f^* &\geq \frac{\mu}{2} \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{2k} \|x^0 - x^*\|_2^2\end{aligned}$$

Ускорение для квадратичных функций

Результат сходимости для квадратичных функций

Предположим, что мы имеем задачу минимизации строго выпуклой квадратичной функции, решаемую методом градиентного спуска:

$$f(x) = \frac{1}{2}x^T Ax - b^T x \quad x^{k+1} = x^k - \alpha_k \nabla f(x^k).$$

Theorem

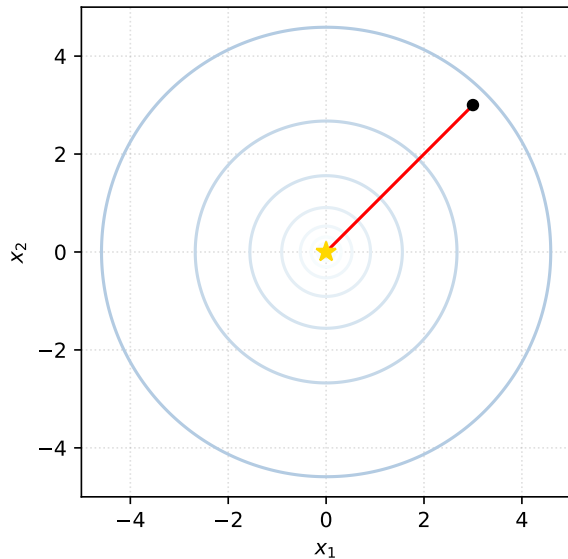
Градиентный спуск с шагом $\alpha_k = \frac{2}{\mu+L}$ сходится к оптимальному решению x^* со следующей гарантией:

$$\|x^{k+1} - x^*\|_2 = \left(\frac{\varkappa - 1}{\varkappa + 1}\right)^k \|x^0 - x^*\|_2 \quad f(x^{k+1}) - f(x^*) = \left(\frac{\varkappa - 1}{\varkappa + 1}\right)^{2k} (f(x^0) - f(x^*))$$

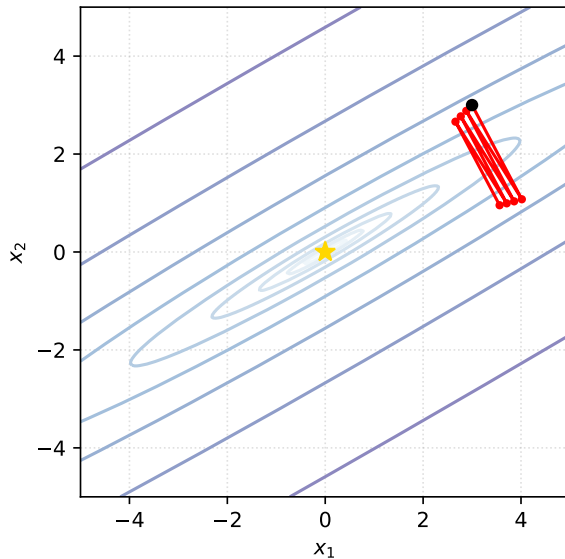
где $\varkappa = \frac{L}{\mu}$ является числом обусловленности A .

Число обусловленности κ

$\kappa = 1.0$



$\kappa = 100.0$



Сходимость из первых принципов

$$f(x) = \frac{1}{2}x^T Ax - b^T x \quad x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$

Пусть x^* будет единственным решением системы линейных уравнений $Ax = b$ и пусть $e_k = \|x_k - x^*\|$, где $x_{k+1} = x_k - \alpha_k(Ax_k - b)$ определяется рекурсивно, начиная с некоторого x_0 , и α_k является шагом, который мы определим позже.

$$e_{k+1} = (I - \alpha_k A)e_k.$$

Полиномы

Вышеуказанный расчет дает нам $e_k = p_k(A)e_0$,
где p_k является полиномом

$$p_k(a) = \prod_{i=1}^k (1 - \alpha_i a).$$

Мы можем ограничить норму ошибки как

$$\|e_k\| \leq \|p_k(A)\| \cdot \|e_0\|.$$

Сходимость из первых принципов

$$f(x) = \frac{1}{2}x^T Ax - b^T x \quad x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$

Пусть x^* будет единственным решением системы линейных уравнений $Ax = b$ и пусть $e_k = \|x_k - x^*\|$, где $x_{k+1} = x_k - \alpha_k(Ax_k - b)$ определяется рекурсивно, начиная с некоторого x_0 , и α_k является шагом, который мы определим позже.

$$e_{k+1} = (I - \alpha_k A)e_k.$$

Полиномы

Вышеуказанный расчет дает нам $e_k = p_k(A)e_0$, где p_k является полиномом

$$p_k(a) = \prod_{i=1}^k (1 - \alpha_i a).$$

Мы можем ограничить норму ошибки как

$$\|e_k\| \leq \|p_k(A)\| \cdot \|e_0\|.$$

Поскольку A является симметричной матрицей с собственными значениями в $[\mu, L]$,

$$\|p_k(A)\| \leq \max_{\mu \leq a \leq L} |p_k(a)|.$$

Это приводит к интересной проблеме: среди всех полиномов, удовлетворяющих $p_k(0) = 1$, мы ищем полином, величина которого как можно меньше в интервале $[\mu, L]$.

Наивное полиномиальное решение

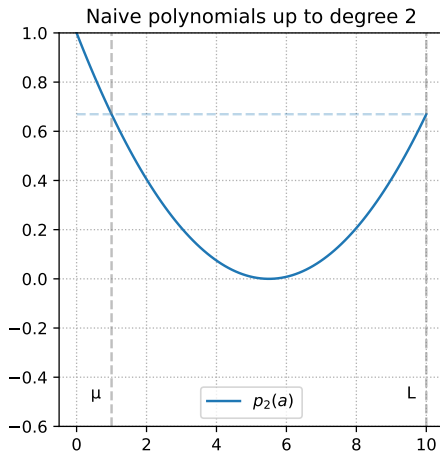
Наивное решение состоит в том, чтобы выбрать равномерный шаг $\alpha_k = \frac{2}{\mu+L}$. Благодаря этому $|p_k(\mu)| = |p_k(L)|$.

$$\|e_k\| \leq \left(1 - \frac{1}{\kappa}\right)^k \|e_0\|$$

Это точно та же скорость, которую мы доказали в предыдущей лекции для любой гладкой и строго выпуклой функции.

Давайте посмотрим на этот полином поближе. На правом рисунке мы выбираем $\alpha = 1$ и $\beta = 10$ так, что $\kappa = 10$. Следовательно, соответствующий интервал равен $[1, 10]$.

Можем ли мы сделать лучше? Ответ да.



Наивное полиномиальное решение

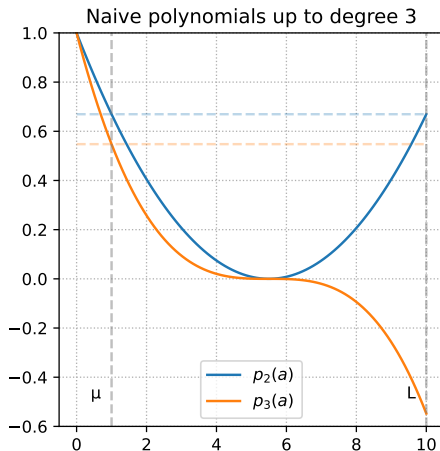
Наивное решение состоит в том, чтобы выбрать равномерный шаг $\alpha_k = \frac{2}{\mu+L}$. Благодаря этому $|p_k(\mu)| = |p_k(L)|$.

$$\|e_k\| \leq \left(1 - \frac{1}{\kappa}\right)^k \|e_0\|$$

Это точно та же скорость, которую мы доказали в предыдущей лекции для любой гладкой и строго выпуклой функции.

Давайте посмотрим на этот полином поближе. На правом рисунке мы выбираем $\alpha = 1$ и $\beta = 10$ так, что $\kappa = 10$. Следовательно, соответствующий интервал равен $[1, 10]$.

Можем ли мы сделать лучше? Ответ да.



Наивное полиномиальное решение

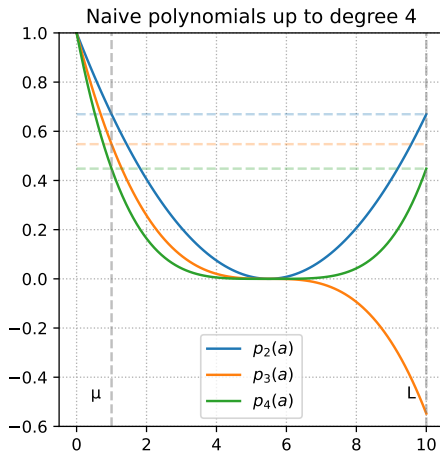
Наивное решение состоит в том, чтобы выбрать равномерный шаг $\alpha_k = \frac{2}{\mu+L}$. Благодаря этому $|p_k(\mu)| = |p_k(L)|$.

$$\|e_k\| \leq \left(1 - \frac{1}{\kappa}\right)^k \|e_0\|$$

Это точно та же скорость, которую мы доказали в предыдущей лекции для любой гладкой и строго выпуклой функции.

Давайте посмотрим на этот полином поближе. На правом рисунке мы выбираем $\alpha = 1$ и $\beta = 10$ так, что $\kappa = 10$. Следовательно, соответствующий интервал равен $[1, 10]$.

Можем ли мы сделать лучше? Ответ да.



Наивное полиномиальное решение

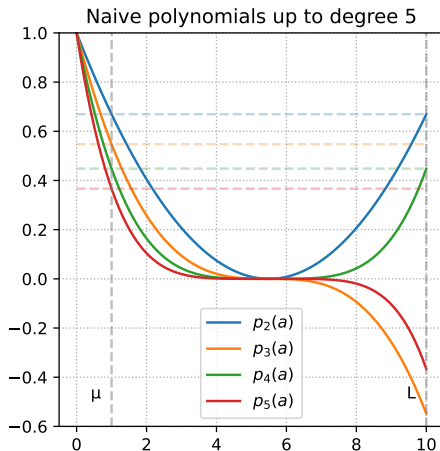
Наивное решение состоит в том, чтобы выбрать равномерный шаг $\alpha_k = \frac{2}{\mu+L}$. Благодаря этому $|p_k(\mu)| = |p_k(L)|$.

$$\|e_k\| \leq \left(1 - \frac{1}{\kappa}\right)^k \|e_0\|$$

Это точно та же скорость, которую мы доказали в предыдущей лекции для любой гладкой и строго выпуклой функции.

Давайте посмотрим на этот полином поближе. На правом рисунке мы выбираем $\alpha = 1$ и $\beta = 10$ так, что $\kappa = 10$. Следовательно, соответствующий интервал равен $[1, 10]$.

Можем ли мы сделать лучше? Ответ да.



Наивное полиномиальное решение

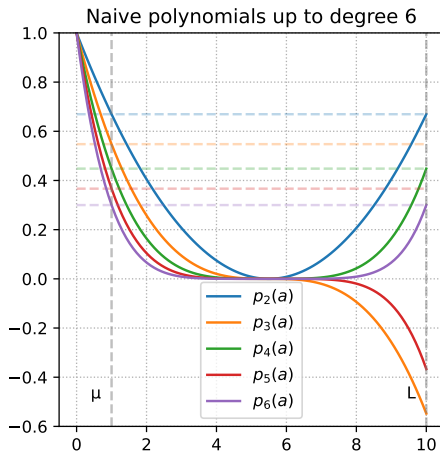
Наивное решение состоит в том, чтобы выбрать равномерный шаг $\alpha_k = \frac{2}{\mu+L}$. Благодаря этому $|p_k(\mu)| = |p_k(L)|$.

$$\|e_k\| \leq \left(1 - \frac{1}{\kappa}\right)^k \|e_0\|$$

Это точно та же скорость, которую мы доказали в предыдущей лекции для любой гладкой и строго выпуклой функции.

Давайте посмотрим на этот полином поближе. На правом рисунке мы выбираем $\alpha = 1$ и $\beta = 10$ так, что $\kappa = 10$. Следовательно, соответствующий интервал равен $[1, 10]$.

Можем ли мы сделать лучше? Ответ да.



Полиномы Чебышева

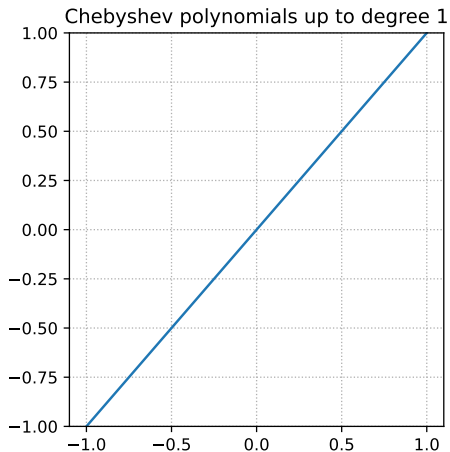
Полиномы Чебышёва дают оптимальный ответ на поставленный вопрос. При соответствующем масштабировании они минимизируют абсолютное значение на заданном интервале $[\mu, L]$, одновременно удовлетворяя нормировочному условию $p(0) = 1$.

$$T_0(x) = 1$$

$$T_1(x) = x$$

$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x), \quad k \geq 2.$$

Давайте построим стандартные полиномы Чебышёва (без масштабирования):



Полиномы Чебышева

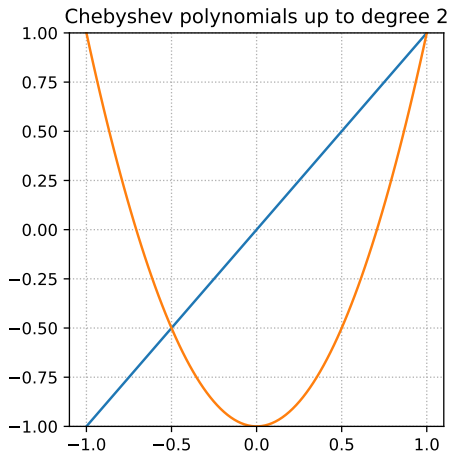
Полиномы Чебышёва дают оптимальный ответ на поставленный вопрос. При соответствующем масштабировании они минимизируют абсолютное значение на заданном интервале $[\mu, L]$, одновременно удовлетворяя нормировочному условию $p(0) = 1$.

$$T_0(x) = 1$$

$$T_1(x) = x$$

$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x), \quad k \geq 2.$$

Давайте построим стандартные полиномы Чебышёва (без масштабирования):



Полиномы Чебышева

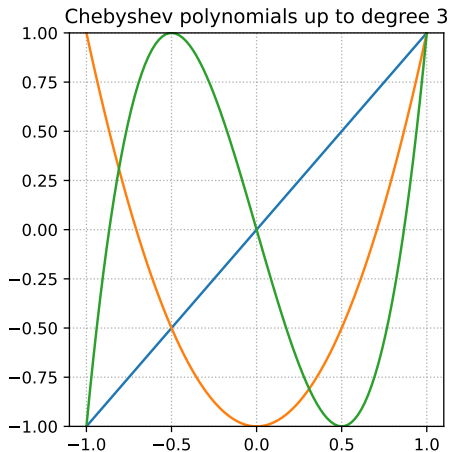
Полиномы Чебышёва дают оптимальный ответ на поставленный вопрос. При соответствующем масштабировании они минимизируют абсолютное значение на заданном интервале $[\mu, L]$, одновременно удовлетворяя нормировочному условию $p(0) = 1$.

$$T_0(x) = 1$$

$$T_1(x) = x$$

$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x), \quad k \geq 2.$$

Давайте построим стандартные полиномы Чебышёва (без масштабирования):



Полиномы Чебышева

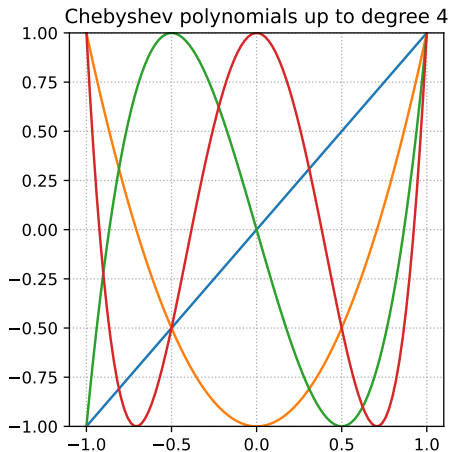
Полиномы Чебышёва дают оптимальный ответ на поставленный вопрос. При соответствующем масштабировании они минимизируют абсолютное значение на заданном интервале $[\mu, L]$, одновременно удовлетворяя нормировочному условию $p(0) = 1$.

$$T_0(x) = 1$$

$$T_1(x) = x$$

$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x), \quad k \geq 2.$$

Давайте построим стандартные полиномы Чебышёва (без масштабирования):



Полиномы Чебышева

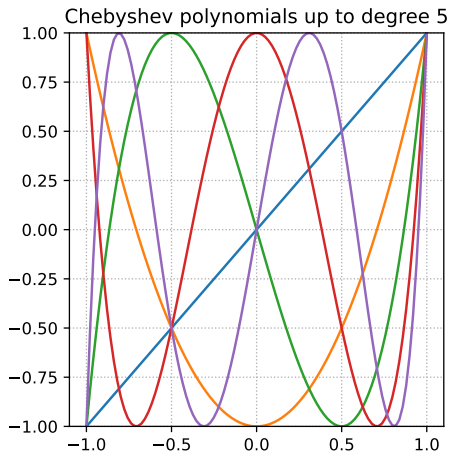
Полиномы Чебышёва дают оптимальный ответ на поставленный вопрос. При соответствующем масштабировании они минимизируют абсолютное значение на заданном интервале $[\mu, L]$, одновременно удовлетворяя нормировочному условию $p(0) = 1$.

$$T_0(x) = 1$$

$$T_1(x) = x$$

$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x), \quad k \geq 2.$$

Давайте построим стандартные полиномы Чебышёва (без масштабирования):



Полиномы Чебышева с масштабированием

Оригинальные полиномы Чебышёва определены на интервале $[-1, 1]$. Чтобы использовать их для наших целей, мы должны масштабировать их на интервал $[\mu, L]$.

Полиномы Чебышева с масштабированием

Оригинальные полиномы Чебышёва определены на интервале $[-1, 1]$. Чтобы использовать их для наших целей, мы должны масштабировать их на интервал $[\mu, L]$.

Мы будем использовать следующее аффинное преобразование:

$$x = \frac{L + \mu - 2a}{L - \mu}, \quad a \in [\mu, L], \quad x \in [-1, 1].$$

Обратите внимание, что $x = 1$ соответствует $a = \mu$, $x = -1$ соответствует $a = L$ и $x = 0$ соответствует $a = \frac{\mu+L}{2}$. Это преобразование гарантирует, что поведение полинома Чебышёва на интервале $[-1, 1]$ отражается на интервале $[\mu, L]$

Полиномы Чебышева с масштабированием

Оригинальные полиномы Чебышёва определены на интервале $[-1, 1]$. Чтобы использовать их для наших целей, мы должны масштабировать их на интервал $[\mu, L]$.

Мы будем использовать следующее аффинное преобразование:

$$x = \frac{L + \mu - 2a}{L - \mu}, \quad a \in [\mu, L], \quad x \in [-1, 1].$$

Обратите внимание, что $x = 1$ соответствует $a = \mu$, $x = -1$ соответствует $a = L$ и $x = 0$ соответствует $a = \frac{\mu+L}{2}$. Это преобразование гарантирует, что поведение полинома Чебышёва на интервале $[-1, 1]$ отражается на интервале $[\mu, L]$

В нашем анализе ошибок мы требуем, чтобы полином был равен 1 в 0 (т.е. $p_k(0) = 1$). После применения преобразования значение T_k в точке, соответствующей $a = 0$, может не быть 1. Следовательно, мы умножаем на обратную величину T_k в точке

$$\frac{L + \mu}{L - \mu}, \quad \text{что обеспечивает} \quad P_k(0) = T_k\left(\frac{L + \mu - 0}{L - \mu}\right) \cdot T_k\left(\frac{L + \mu}{L - \mu}\right)^{-1} = 1.$$

Полиномы Чебышева с масштабированием

Оригинальные полиномы Чебышёва определены на интервале $[-1, 1]$. Чтобы использовать их для наших целей, мы должны масштабировать их на интервал $[\mu, L]$.

Мы будем использовать следующее аффинное преобразование:

$$x = \frac{L + \mu - 2a}{L - \mu}, \quad a \in [\mu, L], \quad x \in [-1, 1].$$

Обратите внимание, что $x = 1$ соответствует $a = \mu$, $x = -1$ соответствует $a = L$ и $x = 0$ соответствует $a = \frac{\mu+L}{2}$. Это преобразование гарантирует, что поведение полинома Чебышёва на интервале $[-1, 1]$ отражается на интервале $[\mu, L]$

В нашем анализе ошибок мы требуем, чтобы полином был равен 1 в 0 (т.е. $p_k(0) = 1$). После применения преобразования значение T_k в точке, соответствующей $a = 0$, может не быть 1. Следовательно, мы умножаем на обратную величину T_k в точке

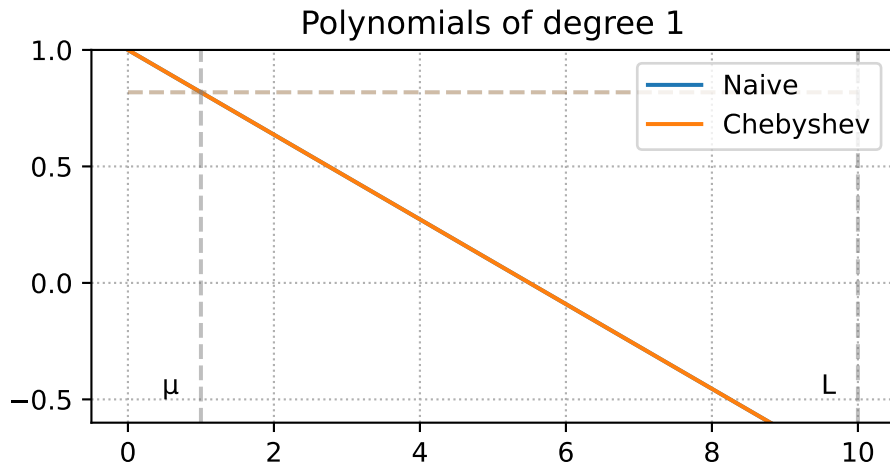
$$\frac{L + \mu}{L - \mu}, \quad \text{что обеспечивает} \quad P_k(0) = T_k\left(\frac{L + \mu - 0}{L - \mu}\right) \cdot T_k\left(\frac{L + \mu}{L - \mu}\right)^{-1} = 1.$$

Давайте построим полиномы Чебышёва с масштабированием

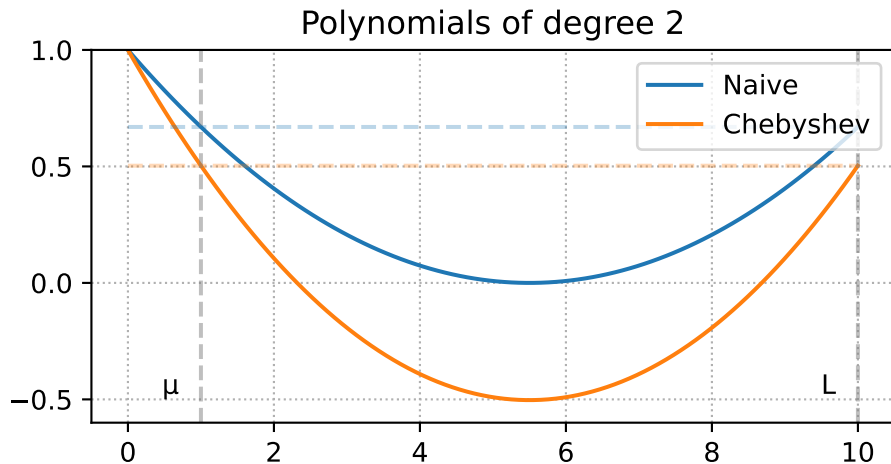
$$P_k(a) = T_k\left(\frac{L + \mu - 2a}{L - \mu}\right) \cdot T_k\left(\frac{L + \mu}{L - \mu}\right)^{-1}$$

и наблюдаем, что они гораздо лучше ведут себя, чем наивные полиномы в отношении величины в интервале $[\mu, L]$.

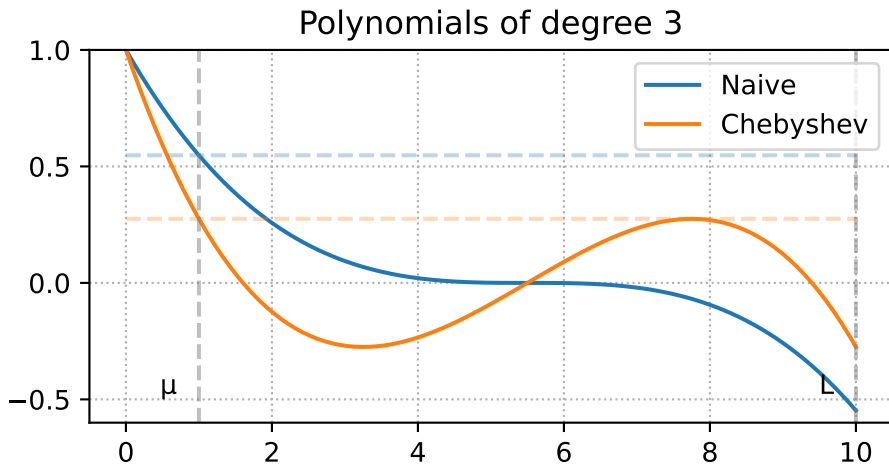
Полиномы Чебышева с масштабированием



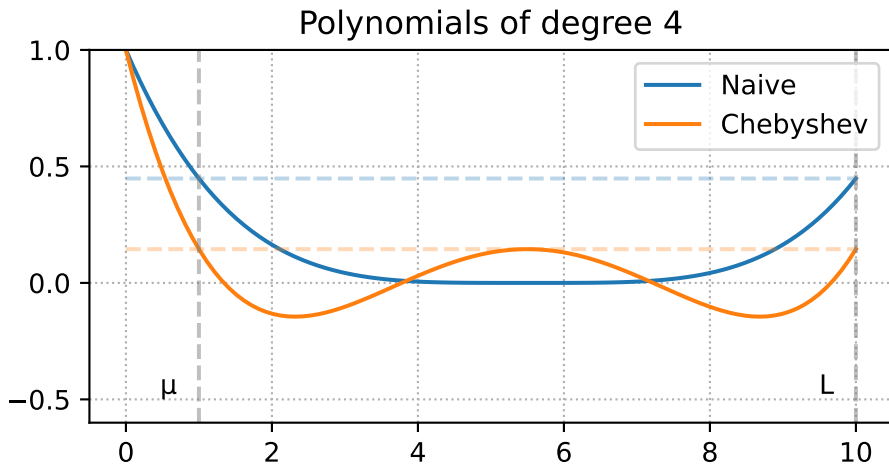
Полиномы Чебышева с масштабированием



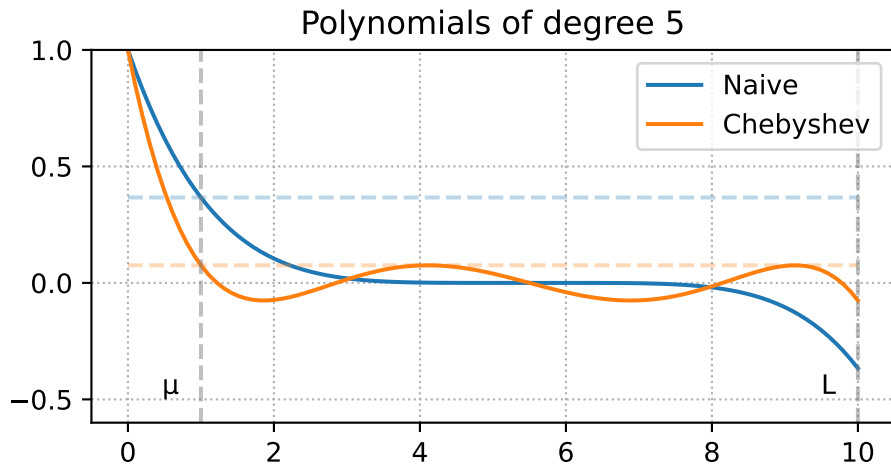
Полиномы Чебышева с масштабированием



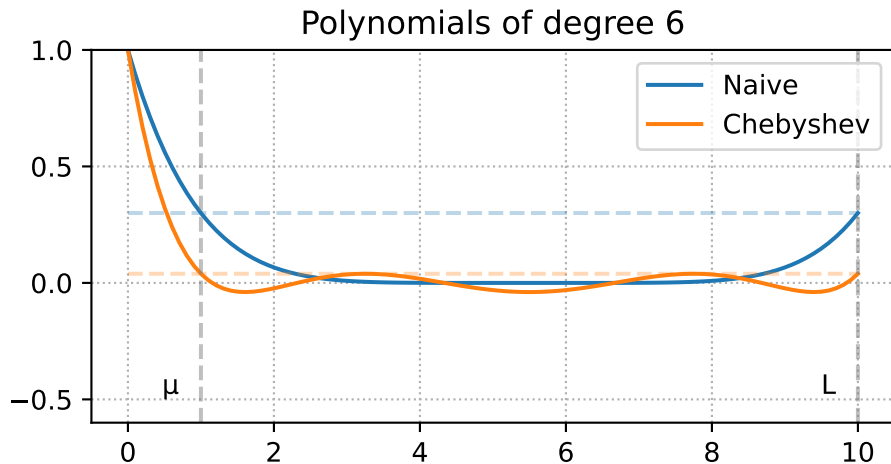
Полиномы Чебышева с масштабированием



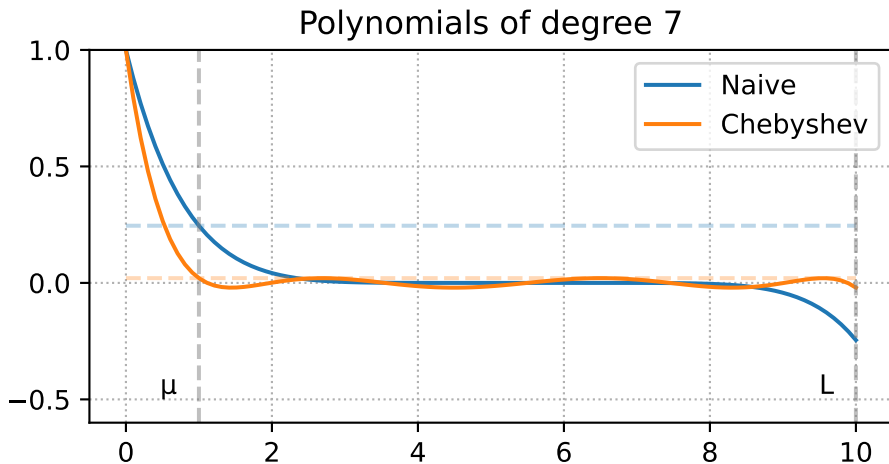
Полиномы Чебышева с масштабированием



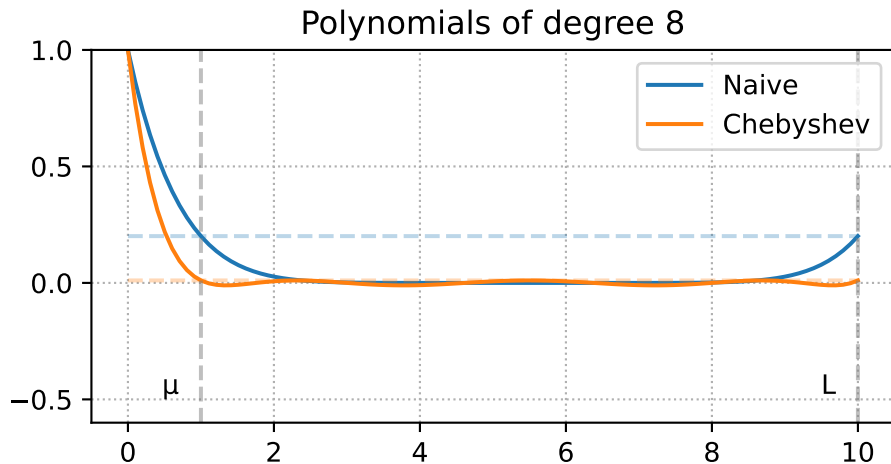
Полиномы Чебышева с масштабированием



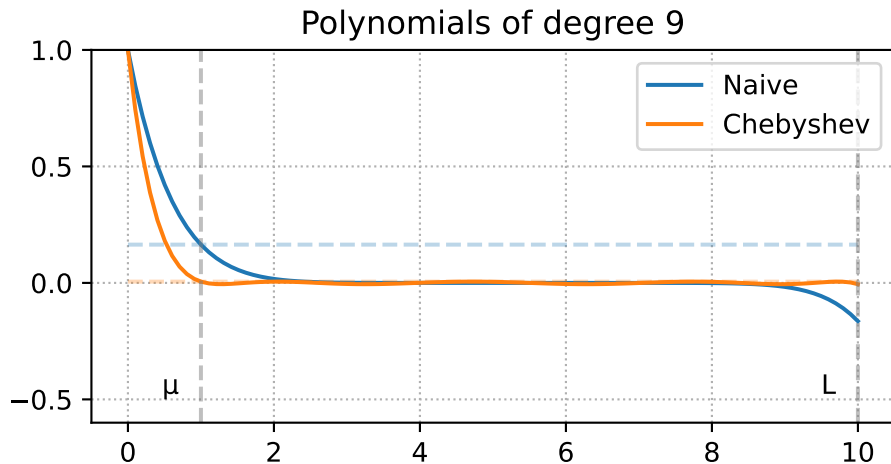
Полиномы Чебышева с масштабированием



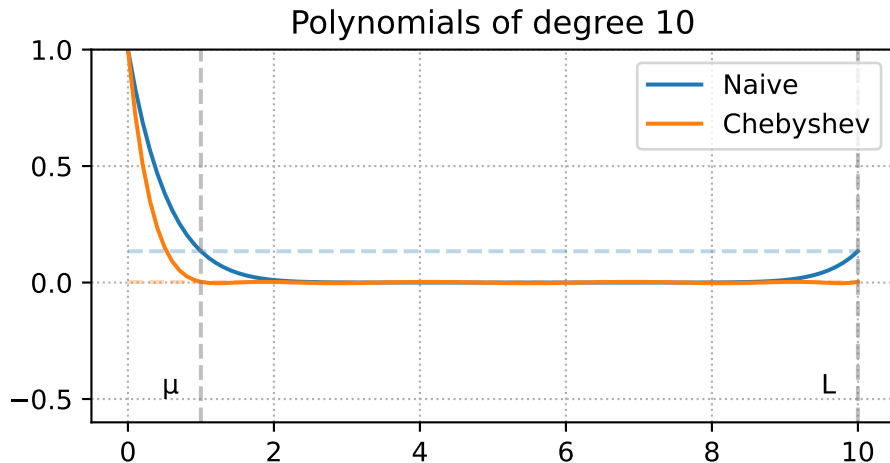
Полиномы Чебышева с масштабированием



Полиномы Чебышева с масштабированием



Полиномы Чебышева с масштабированием



Верхняя оценка для полиномов Чебышёва

Мы можем видеть, что максимальное значение полинома Чебышёва на интервале $[\mu, L]$ достигается в точке $a = \mu$. Следовательно, мы можем использовать следующую верхнюю оценку:

$$\|P_k(A)\|_2 \leq P_k(\mu) = T_k\left(\frac{L + \mu - 2\mu}{L - \mu}\right) \cdot T_k\left(\frac{L + \mu}{L - \mu}\right)^{-1} = T_k(1) \cdot T_k\left(\frac{L + \mu}{L - \mu}\right)^{-1} = T_k\left(\frac{L + \mu}{L - \mu}\right)^{-1}$$

Верхняя оценка для полиномов Чебышёва

Мы можем видеть, что максимальное значение полинома Чебышёва на интервале $[\mu, L]$ достигается в точке $a = \mu$. Следовательно, мы можем использовать следующую верхнюю оценку:

$$\|P_k(A)\|_2 \leq P_k(\mu) = T_k\left(\frac{L + \mu - 2\mu}{L - \mu}\right) \cdot T_k\left(\frac{L + \mu}{L - \mu}\right)^{-1} = T_k(1) \cdot T_k\left(\frac{L + \mu}{L - \mu}\right)^{-1} = T_k\left(\frac{L + \mu}{L - \mu}\right)^{-1}$$

Используя определение числа обусловленности $\kappa = \frac{L}{\mu}$, мы получаем:

$$\|P_k(A)\|_2 \leq T_k\left(\frac{\kappa + 1}{\kappa - 1}\right)^{-1} = T_k\left(1 + \frac{2}{\kappa - 1}\right)^{-1} = T_k(1 + \epsilon)^{-1}, \quad \epsilon = \frac{2}{\kappa - 1}.$$

Верхняя оценка для полиномов Чебышёва

Мы можем видеть, что максимальное значение полинома Чебышёва на интервале $[\mu, L]$ достигается в точке $a = \mu$. Следовательно, мы можем использовать следующую верхнюю оценку:

$$\|P_k(A)\|_2 \leq P_k(\mu) = T_k\left(\frac{L + \mu - 2\mu}{L - \mu}\right) \cdot T_k\left(\frac{L + \mu}{L - \mu}\right)^{-1} = T_k(1) \cdot T_k\left(\frac{L + \mu}{L - \mu}\right)^{-1} = T_k\left(\frac{L + \mu}{L - \mu}\right)^{-1}$$

Используя определение числа обусловленности $\kappa = \frac{L}{\mu}$, мы получаем:

$$\|P_k(A)\|_2 \leq T_k\left(\frac{\kappa + 1}{\kappa - 1}\right)^{-1} = T_k\left(1 + \frac{2}{\kappa - 1}\right)^{-1} = T_k(1 + \epsilon)^{-1}, \quad \epsilon = \frac{2}{\kappa - 1}.$$

Следовательно, нам нужно только понять значение T_k в $1 + \epsilon$. Это то, откуда берется ускорение. Мы будем ограничивать это значение сверху величиной $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$.

Верхняя оценка для полиномов Чебышёва

Чтобы ограничить $|P_k|$ сверху, мы должны ограничить $|T_k(1 + \epsilon)|$ снизу.

Верхняя оценка для полиномов Чебышёва

Чтобы ограничить $|P_k|$ сверху, мы должны ограничить $|T_k(1 + \epsilon)|$ снизу.

1. Для любого $x \geq 1$, полиномы Чебышёва первого рода могут быть записаны как

$$T_k(x) = \cosh(k \operatorname{arccosh}(x))$$
$$T_k(1 + \epsilon) = \cosh(k \operatorname{arccosh}(1 + \epsilon)).$$

Верхняя оценка для полиномов Чебышёва

Чтобы ограничить $|P_k|$ сверху, мы должны ограничить $|T_k(1 + \epsilon)|$ снизу.

1. Для любого $x \geq 1$, полиномы Чебышёва первого рода могут быть записаны как

$$T_k(x) = \cosh(k \operatorname{arccosh}(x))$$
$$T_k(1 + \epsilon) = \cosh(k \operatorname{arccosh}(1 + \epsilon)).$$

2. Помните, что:

$$\cosh(x) = \frac{e^x + e^{-x}}{2} \quad \operatorname{arccosh}(x) = \ln(x + \sqrt{x^2 - 1}).$$

Верхняя оценка для полиномов Чебышёва

Чтобы ограничить $|P_k|$ сверху, мы должны ограничить $|T_k(1 + \epsilon)|$ снизу.

1. Для любого $x \geq 1$, полиномы Чебышёва первого рода могут быть записаны как

$$\begin{aligned}T_k(x) &= \cosh(k \operatorname{arccosh}(x)) \\T_k(1 + \epsilon) &= \cosh(k \operatorname{arccosh}(1 + \epsilon)).\end{aligned}$$

2. Помните, что:

$$\cosh(x) = \frac{e^x + e^{-x}}{2} \quad \operatorname{arccosh}(x) = \ln(x + \sqrt{x^2 - 1}).$$

3. Теперь, пусть $\phi = \operatorname{arccosh}(1 + \epsilon)$,

$$e^\phi = 1 + \epsilon + \sqrt{2\epsilon + \epsilon^2} \geq 1 + \sqrt{\epsilon}.$$

Верхняя оценка для полиномов Чебышёва

Чтобы ограничить $|P_k|$ сверху, мы должны ограничить $|T_k(1 + \epsilon)|$ снизу.

1. Для любого $x \geq 1$, полиномы Чебышёва первого рода могут быть записаны как
4. Следовательно,

$$\begin{aligned}T_k(x) &= \cosh(k \operatorname{arccosh}(x)) \\T_k(1 + \epsilon) &= \cosh(k \operatorname{arccosh}(1 + \epsilon)).\end{aligned}$$

2. Помните, что:

$$\cosh(x) = \frac{e^x + e^{-x}}{2} \quad \operatorname{arccosh}(x) = \ln(x + \sqrt{x^2 - 1}).$$

3. Теперь, пусть $\phi = \operatorname{arccosh}(1 + \epsilon)$,

$$e^\phi = 1 + \epsilon + \sqrt{2\epsilon + \epsilon^2} \geq 1 + \sqrt{\epsilon}.$$

$$\begin{aligned}T_k(1 + \epsilon) &= \cosh(k \operatorname{arccosh}(1 + \epsilon)) \\&= \cosh(k\phi) \\&= \frac{e^{k\phi} + e^{-k\phi}}{2} \geq \frac{e^{k\phi}}{2} \\&= \frac{(1 + \sqrt{\epsilon})^k}{2}.\end{aligned}$$

Верхняя оценка для полиномов Чебышёва

Чтобы ограничить $|P_k|$ сверху, мы должны ограничить $|T_k(1 + \epsilon)|$ снизу.

1. Для любого $x \geq 1$, полиномы Чебышёва первого рода могут быть записаны как
4. Следовательно,

$$\begin{aligned}T_k(x) &= \cosh(k \operatorname{arccosh}(x)) \\T_k(1 + \epsilon) &= \cosh(k \operatorname{arccosh}(1 + \epsilon)).\end{aligned}$$

2. Помните, что:

$$\cosh(x) = \frac{e^x + e^{-x}}{2} \quad \operatorname{arccosh}(x) = \ln(x + \sqrt{x^2 - 1}).$$

3. Теперь, пусть $\phi = \operatorname{arccosh}(1 + \epsilon)$,

$$e^\phi = 1 + \epsilon + \sqrt{2\epsilon + \epsilon^2} \geq 1 + \sqrt{\epsilon}.$$

$$\begin{aligned}T_k(1 + \epsilon) &= \cosh(k \operatorname{arccosh}(1 + \epsilon)) \\&= \cosh(k\phi) \\&= \frac{e^{k\phi} + e^{-k\phi}}{2} \geq \frac{e^{k\phi}}{2} \\&= \frac{(1 + \sqrt{\epsilon})^k}{2}.\end{aligned}$$

5. Наконец, мы получаем:

$$\begin{aligned}\|e_k\| &\leq \|P_k(A)\| \|e_0\| \leq \frac{2}{(1 + \sqrt{\epsilon})^k} \|e_0\| \\&\leq 2 \left(1 + \sqrt{\frac{2}{n-1}}\right)^{-k} \|e_0\| \\&\leq 2 \exp\left(-\sqrt{\frac{2}{n-1}} k\right) \|e_0\|\end{aligned}$$

Ускоренный метод [1/2]

Из-за рекурсивного определения полиномов Чебышёва мы непосредственно получаем итерационную схему ускорения. Переформулируя рекурсию в терминах наших масштабированных полиномов Чебышёва, мы получаем:

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$$

Принимая во внимание, что $x = \frac{L+\mu-2a}{L-\mu}$, и:

$$P_k(a) = T_k\left(\frac{L+\mu-2a}{L-\mu}\right) T_k\left(\frac{L+\mu}{L-\mu}\right)^{-1}$$

$$T_k\left(\frac{L+\mu-2a}{L-\mu}\right) = P_k(a) T_k\left(\frac{L+\mu}{L-\mu}\right)$$

Ускоренный метод [1/2]

Из-за рекурсивного определения полиномов Чебышёва мы непосредственно получаем итерационную схему ускорения. Переформулируя рекурсию в терминах наших масштабированных полиномов Чебышёва, мы получаем:

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$$

Принимая во внимание, что $x = \frac{L+\mu-2a}{L-\mu}$, и:

$$P_k(a) = T_k\left(\frac{L+\mu-2a}{L-\mu}\right) T_k\left(\frac{L+\mu}{L-\mu}\right)^{-1}$$

$$T_k\left(\frac{L+\mu-2a}{L-\mu}\right) = P_k(a) T_k\left(\frac{L+\mu}{L-\mu}\right)$$

$$T_{k-1}\left(\frac{L+\mu-2a}{L-\mu}\right) = P_{k-1}(a) T_{k-1}\left(\frac{L+\mu}{L-\mu}\right)$$

$$T_{k+1}\left(\frac{L+\mu-2a}{L-\mu}\right) = P_{k+1}(a) T_{k+1}\left(\frac{L+\mu}{L-\mu}\right)$$

$$P_{k+1}(a)t_{k+1} = 2\frac{L+\mu-2a}{L-\mu}P_k(a)t_k - P_{k-1}(a)t_{k-1}, \text{ where } t_k = T_k\left(\frac{L+\mu}{L-\mu}\right)$$

$$P_{k+1}(a) = 2\frac{L+\mu-2a}{L-\mu}P_k(a)\frac{t_k}{t_{k+1}} - P_{k-1}(a)\frac{t_{k-1}}{t_{k+1}}$$

Ускоренный метод [1/2]

Из-за рекурсивного определения полиномов Чебышёва мы непосредственно получаем итерационную схему ускорения. Переформулируя рекурсию в терминах наших масштабированных полиномов Чебышёва, мы получаем:

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$$

Принимая во внимание, что $x = \frac{L+\mu-2a}{L-\mu}$, и:

$$P_k(a) = T_k\left(\frac{L+\mu-2a}{L-\mu}\right) T_k\left(\frac{L+\mu}{L-\mu}\right)^{-1}$$

$$T_k\left(\frac{L+\mu-2a}{L-\mu}\right) = P_k(a) T_k\left(\frac{L+\mu}{L-\mu}\right)$$

$$T_{k-1}\left(\frac{L+\mu-2a}{L-\mu}\right) = P_{k-1}(a) T_{k-1}\left(\frac{L+\mu}{L-\mu}\right)$$

$$T_{k+1}\left(\frac{L+\mu-2a}{L-\mu}\right) = P_{k+1}(a) T_{k+1}\left(\frac{L+\mu}{L-\mu}\right)$$

$$P_{k+1}(a)t_{k+1} = 2\frac{L+\mu-2a}{L-\mu}P_k(a)t_k - P_{k-1}(a)t_{k-1}, \text{ where } t_k = T_k\left(\frac{L+\mu}{L-\mu}\right)$$

$$P_{k+1}(a) = 2\frac{L+\mu-2a}{L-\mu}P_k(a)\frac{t_k}{t_{k+1}} - P_{k-1}(a)\frac{t_{k-1}}{t_{k+1}}$$

Поскольку мы имеем $P_{k+1}(0) = P_k(0) = P_{k-1}(0) = 1$, получаем рекуррентную формулу вида:

$$P_{k+1}(a) = (1 - \alpha_k a)P_k(a) + \beta_k (P_k(a) - P_{k-1}(a)).$$

Ускоренный метод [2/2]

Перегруппируя члены, мы получаем:

$$P_{k+1}(a) = (1 + \beta_k)P_k(a) - \alpha_k a P_k(a) - \beta_k P_{k-1}(a),$$
$$P_{k+1}(a) = 2 \frac{L + \mu}{L - \mu} \frac{t_k}{t_{k+1}} P_k(a) - \frac{4a}{L - \mu} \frac{t_k}{t_{k+1}} P_k(a) - \frac{t_{k-1}}{t_{k+1}} P_{k-1}(a)$$

Ускоренный метод [2/2]

Перегруппируя члены, мы получаем:

$$\begin{aligned}P_{k+1}(a) &= (1 + \beta_k)P_k(a) - \alpha_k a P_k(a) - \beta_k P_{k-1}(a), \\P_{k+1}(a) &= 2 \frac{L + \mu}{L - \mu} \frac{t_k}{t_{k+1}} P_k(a) - \frac{4a}{L - \mu} \frac{t_k}{t_{k+1}} P_k(a) - \frac{t_{k-1}}{t_{k+1}} P_{k-1}(a)\end{aligned}$$

$$\begin{cases} \beta_k = \frac{t_{k-1}}{t_{k+1}}, \\ \alpha_k = \frac{4}{L - \mu} \frac{t_k}{t_{k+1}}, \\ 1 + \beta_k = 2 \frac{L + \mu}{L - \mu} \frac{t_k}{t_{k+1}} \end{cases}$$

Ускоренный метод [2/2]

Перегруппируя члены, мы получаем:

$$P_{k+1}(a) = (1 + \beta_k)P_k(a) - \alpha_k a P_k(a) - \beta_k P_{k-1}(a),$$

$$P_{k+1}(a) = 2 \frac{L + \mu}{L - \mu} \frac{t_k}{t_{k+1}} P_k(a) - \frac{4a}{L - \mu} \frac{t_k}{t_{k+1}} P_k(a) - \frac{t_{k-1}}{t_{k+1}} P_{k-1}(a)$$

$$\begin{cases} \beta_k = \frac{t_{k-1}}{t_{k+1}}, \\ \alpha_k = \frac{4}{L - \mu} \frac{t_k}{t_{k+1}}, \\ 1 + \beta_k = 2 \frac{L + \mu}{L - \mu} \frac{t_k}{t_{k+1}} \end{cases}$$

Мы почти закончили :) Помним, что $e_{k+1} = P_{k+1}(A)e_0$. Также обратим внимание, что мы работаем с квадратичной задачей, поэтому мы можем предположить $x^* = 0$ без ограничения общности. В этом случае $e_0 = x_0$ и $e_{k+1} = x_{k+1}$.

$$\begin{aligned} x_{k+1} &= P_{k+1}(A)x_0 = (I - \alpha_k A)P_k(A)x_0 + \beta_k (P_k(A) - P_{k-1}(A))x_0 \\ &= (I - \alpha_k A)x_k + \beta_k (x_k - x_{k-1}) \end{aligned}$$

Ускоренный метод [2/2]

Перегруппируя члены, мы получаем:

$$P_{k+1}(a) = (1 + \beta_k)P_k(a) - \alpha_k a P_k(a) - \beta_k P_{k-1}(a),$$

$$P_{k+1}(a) = 2 \frac{L + \mu}{L - \mu} \frac{t_k}{t_{k+1}} P_k(a) - \frac{4a}{L - \mu} \frac{t_k}{t_{k+1}} P_k(a) - \frac{t_{k-1}}{t_{k+1}} P_{k-1}(a)$$

$$\begin{cases} \beta_k = \frac{t_{k-1}}{t_{k+1}}, \\ \alpha_k = \frac{4}{L - \mu} \frac{t_k}{t_{k+1}}, \\ 1 + \beta_k = 2 \frac{L + \mu}{L - \mu} \frac{t_k}{t_{k+1}} \end{cases}$$

Мы почти закончили :) Помним, что $e_{k+1} = P_{k+1}(A)e_0$. Также обратим внимание, что мы работаем с квадратичной задачей, поэтому мы можем предположить $x^* = 0$ без ограничения общности. В этом случае $e_0 = x_0$ и $e_{k+1} = x_{k+1}$.

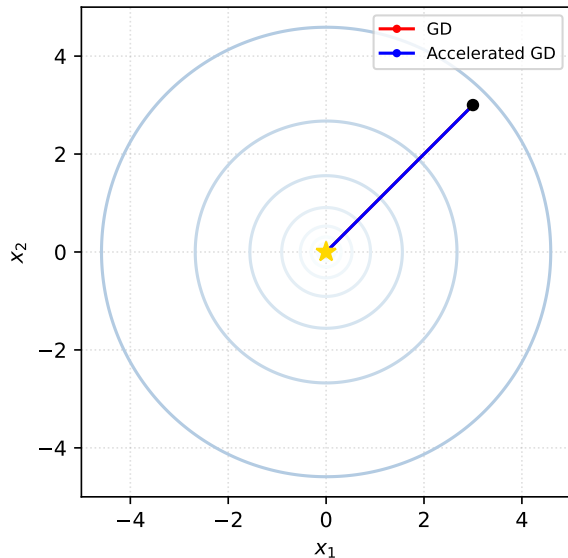
$$\begin{aligned} x_{k+1} &= P_{k+1}(A)x_0 = (I - \alpha_k A)P_k(A)x_0 + \beta_k (P_k(A) - P_{k-1}(A))x_0 \\ &= (I - \alpha_k A)x_k + \beta_k (x_k - x_{k-1}) \end{aligned}$$

Для квадратичной задачи мы имеем $\nabla f(x_k) = Ax_k$, поэтому мы можем переписать обновление как:

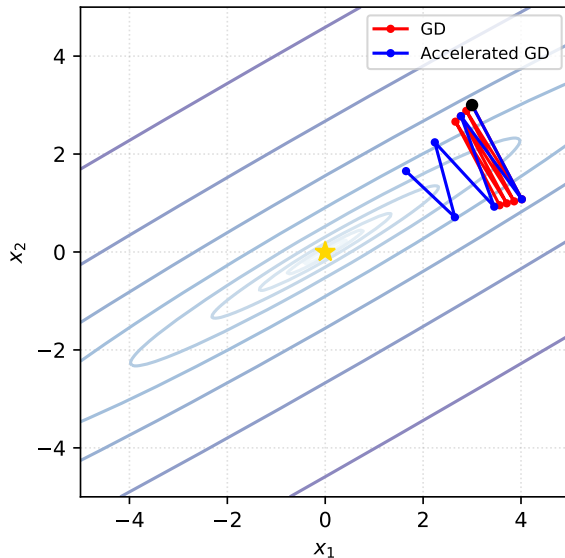
$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k (x_k - x_{k-1})$$

Ускорение из первых принципов

$\kappa = 1.0$



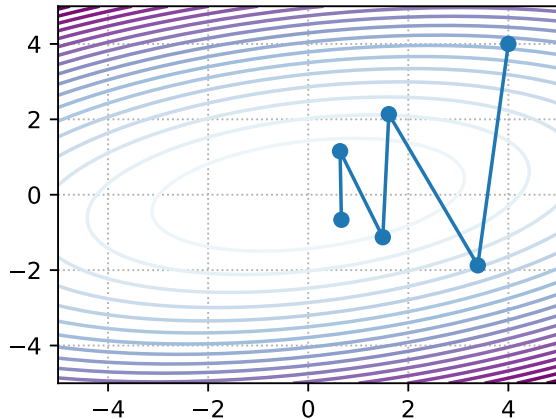
$\kappa = 100.0$



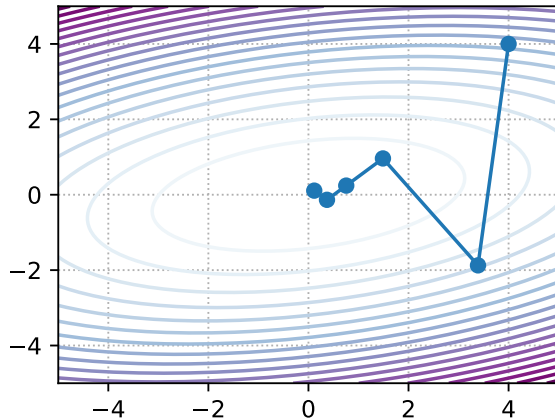
Метод тяжелого шарика

Колебания и ускорение

Gradient Descent



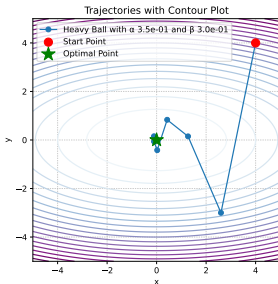
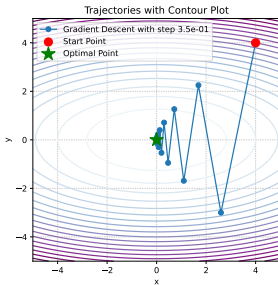
Heavy Ball



Метод тяжелого шарика Поляка

Давайте представим идею импульса, предложенную Поляком в 1964 году. Вспомним, что обновление импульса имеет вид

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}).$$



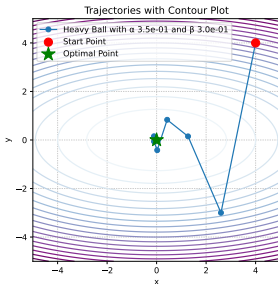
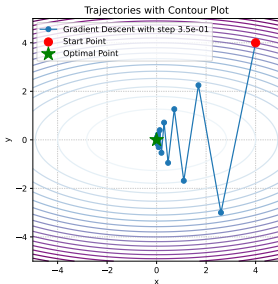
Метод тяжелого шарика Поляка

Давайте представим идею импульса, предложенную Поляком в 1964 году. Вспомним, что обновление импульса имеет вид

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}).$$

В нашем (квадратичном) случае это

$$\hat{x}_{k+1} = \hat{x}_k - \alpha \Lambda \hat{x}_k + \beta(\hat{x}_k - \hat{x}_{k-1}) = (I - \alpha \Lambda + \beta I) \hat{x}_k - \beta \hat{x}_{k-1}$$



Метод тяжелого шарика Поляка

Давайте представим идею импульса, предложенную Поляком в 1964 году. Вспомним, что обновление импульса имеет вид

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}).$$

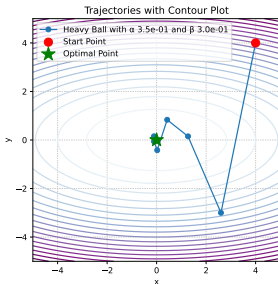
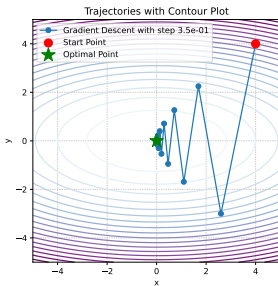
В нашем (квадратичном) случае это

$$\hat{x}_{k+1} = \hat{x}_k - \alpha \Lambda \hat{x}_k + \beta(\hat{x}_k - \hat{x}_{k-1}) = (I - \alpha \Lambda + \beta I) \hat{x}_k - \beta \hat{x}_{k-1}$$

Это можно переписать как

$$\hat{x}_{k+1} = (I - \alpha \Lambda + \beta I) \hat{x}_k - \beta \hat{x}_{k-1},$$

$$\hat{x}_k = \hat{x}_k.$$



Метод тяжелого шарика Поляка

Давайте представим идею импульса, предложенную Поляком в 1964 году. Вспомним, что обновление импульса имеет вид

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}).$$

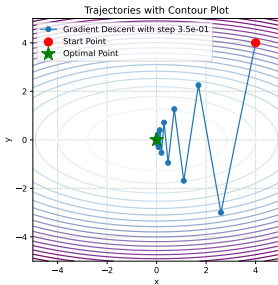
В нашем (квадратичном) случае это

$$\hat{x}_{k+1} = \hat{x}_k - \alpha \Lambda \hat{x}_k + \beta(\hat{x}_k - \hat{x}_{k-1}) = (I - \alpha \Lambda + \beta I) \hat{x}_k - \beta \hat{x}_{k-1}$$

Это можно переписать как

$$\begin{aligned} \hat{x}_{k+1} &= (I - \alpha \Lambda + \beta I) \hat{x}_k - \beta \hat{x}_{k-1}, \\ \hat{x}_k &= \hat{x}_k. \end{aligned}$$

Давайте используем следующую нотацию $\hat{z}_k = \begin{bmatrix} \hat{x}_{k+1} \\ \hat{x}_k \end{bmatrix}$. Следовательно, $\hat{z}_{k+1} = M \hat{z}_k$, где матрица итерации M имеет вид:



Метод тяжелого шарика Поляка

Давайте представим идею импульса, предложенную Поляком в 1964 году. Вспомним, что обновление импульса имеет вид

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}).$$

В нашем (квадратичном) случае это

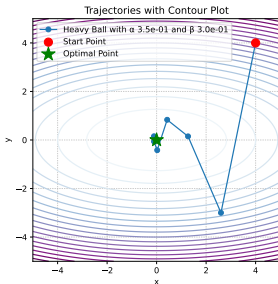
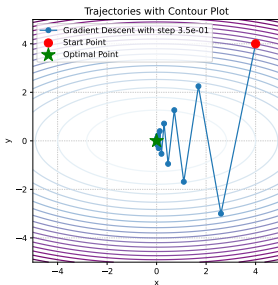
$$\hat{x}_{k+1} = \hat{x}_k - \alpha \Lambda \hat{x}_k + \beta(\hat{x}_k - \hat{x}_{k-1}) = (I - \alpha \Lambda + \beta I) \hat{x}_k - \beta \hat{x}_{k-1}$$

Это можно переписать как

$$\begin{aligned} \hat{x}_{k+1} &= (I - \alpha \Lambda + \beta I) \hat{x}_k - \beta \hat{x}_{k-1}, \\ \hat{x}_k &= \hat{x}_k. \end{aligned}$$

Давайте используем следующую нотацию $\hat{z}_k = \begin{bmatrix} \hat{x}_{k+1} \\ \hat{x}_k \end{bmatrix}$. Следовательно, $\hat{z}_{k+1} = M \hat{z}_k$, где матрица итерации M имеет вид:

$$M = \begin{bmatrix} I - \alpha \Lambda + \beta I & -\beta I \\ I & 0_d \end{bmatrix}.$$



Сведение к скалярному случаю

Обратим внимание, что M является матрицей $2d \times 2d$ с 4 блочно-диагональными матрицами размера $d \times d$ внутри. Это означает, что мы можем переупорядочить порядок координат, чтобы сделать M блочно-диагональной. Обратите внимание, что в уравнении ниже, матрица M обозначает то же самое, что и в обозначении выше, за исключением описанной перестановки строк и столбцов. Мы используем эту небольшую неточность ради ясности.

Сведение к скалярному случаю

Обратим внимание, что M является матрицей $2d \times 2d$ с 4 блочно-диагональными матрицами размера $d \times d$ внутри. Это означает, что мы можем переупорядочить порядок координат, чтобы сделать M блочно-диагональной. Обратите внимание, что в уравнении ниже, матрица M обозначает то же самое, что и в обозначении выше, за исключением описанной перестановки строк и столбцов. Мы используем эту небольшую неточность ради ясности.

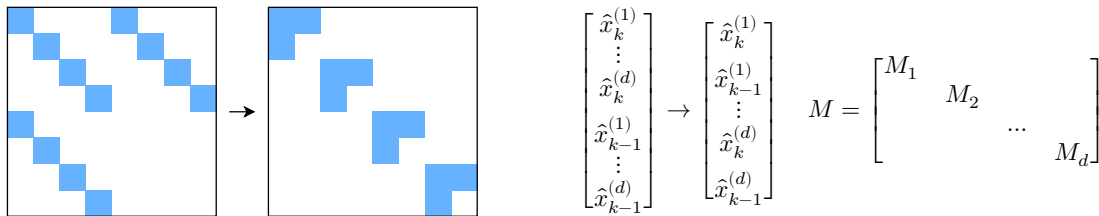


Рис. 1: Иллюстрация переупорядочения матрицы M

где $\hat{x}_k^{(i)}$ является i -й координатой вектора $\hat{x}_k \in \mathbb{R}^d$ и M_i обозначает 2×2 матрицу. Переупорядочение позволяет нам исследовать динамику метода независимо от размерности. Асимптотическая скорость сходимости $2d$ -мерной последовательности векторов \hat{z}_k определяется наихудшей скоростью сходимости среди его блока координат. Следовательно, достаточно исследовать оптимизацию в одномерном случае.

Сведение к скалярному случаю

Для i -й координаты с λ_i как i -е собственное значение матрицы W мы имеем:

$$M_i = \begin{bmatrix} 1 - \alpha\lambda_i + \beta & -\beta \\ 1 & 0 \end{bmatrix}.$$

Сведение к скалярному случаю

Для i -й координаты с λ_i как i -е собственное значение матрицы W мы имеем:

$$M_i = \begin{bmatrix} 1 - \alpha\lambda_i + \beta & -\beta \\ 1 & 0 \end{bmatrix}.$$

Метод будет сходиться, если $\rho(M) < 1$, и оптимальные параметры могут быть вычислены путем оптимизации спектрального радиуса

$$\alpha^*, \beta^* = \arg \min_{\alpha, \beta} \max_i \rho(M_i) \quad \alpha^* = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2}; \quad \beta^* = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^2.$$

Сведение к скалярному случаю

Для i -й координаты с λ_i как i -е собственное значение матрицы W мы имеем:

$$M_i = \begin{bmatrix} 1 - \alpha\lambda_i + \beta & -\beta \\ 1 & 0 \end{bmatrix}.$$

Метод будет сходиться, если $\rho(M) < 1$, и оптимальные параметры могут быть вычислены путем оптимизации спектрального радиуса

$$\alpha^*, \beta^* = \arg \min_{\alpha, \beta} \max_i \rho(M_i) \quad \alpha^* = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2}; \quad \beta^* = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^2.$$

Можно показать, что для таких параметров матрица M имеет комплексные собственные значения, которые образуют комплексно-сопряженную пару, поэтому расстояние до оптимума (в этом случае, $\|z_k\|$), как правило, не будет монотонно уменьшаться.

Квадратичная сходимость метода тяжелого шарика

Мы можем явно вычислить собственные значения M_i :

$$\lambda_1^M, \lambda_2^M = \lambda \left(\begin{bmatrix} 1 - \alpha\lambda_i + \beta & -\beta \\ 1 & 0 \end{bmatrix} \right) = \frac{1 + \beta - \alpha\lambda_i \pm \sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta}}{2}.$$

Квадратичная сходимость метода тяжелого шарика

Мы можем явно вычислить собственные значения M_i :

$$\lambda_1^M, \lambda_2^M = \lambda \left(\begin{bmatrix} 1 - \alpha\lambda_i + \beta & -\beta \\ 1 & 0 \end{bmatrix} \right) = \frac{1 + \beta - \alpha\lambda_i \pm \sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta}}{2}.$$

Когда α и β оптимальны (α^*, β^*), собственные значения являются комплексно-сопряженной парой $(1 + \beta - \alpha\lambda_i)^2 - 4\beta \leq 0$, т.е. $\beta \geq (1 - \sqrt{\alpha\lambda_i})^2$.

Квадратичная сходимость метода тяжелого шарика

Мы можем явно вычислить собственные значения M_i :

$$\lambda_1^M, \lambda_2^M = \lambda \left(\begin{bmatrix} 1 - \alpha\lambda_i + \beta & -\beta \\ 1 & 0 \end{bmatrix} \right) = \frac{1 + \beta - \alpha\lambda_i \pm \sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta}}{2}.$$

Когда α и β оптимальны (α^*, β^*) , собственные значения являются комплексно-сопряженной парой $(1 + \beta - \alpha\lambda_i)^2 - 4\beta \leq 0$, т.е. $\beta \geq (1 - \sqrt{\alpha\lambda_i})^2$.

$$\operatorname{Re}(\lambda^M) = \frac{L + \mu - 2\lambda_i}{(\sqrt{L} + \sqrt{\mu})^2}; \quad \operatorname{Im}(\lambda^M) = \frac{\pm 2\sqrt{(L - \lambda_i)(\lambda_i - \mu)}}{(\sqrt{L} + \sqrt{\mu})^2}; \quad |\lambda^M| = \frac{L - \mu}{(\sqrt{L} + \sqrt{\mu})^2}.$$

Квадратичная сходимость метода тяжелого шарика

Мы можем явно вычислить собственные значения M_i :

$$\lambda_1^M, \lambda_2^M = \lambda \left(\begin{bmatrix} 1 - \alpha\lambda_i + \beta & -\beta \\ 1 & 0 \end{bmatrix} \right) = \frac{1 + \beta - \alpha\lambda_i \pm \sqrt{(1 + \beta - \alpha\lambda_i)^2 - 4\beta}}{2}.$$

Когда α и β оптимальны (α^*, β^*) , собственные значения являются комплексно-сопряженной парой $(1 + \beta - \alpha\lambda_i)^2 - 4\beta \leq 0$, т.е. $\beta \geq (1 - \sqrt{\alpha\lambda_i})^2$.

$$\operatorname{Re}(\lambda^M) = \frac{L + \mu - 2\lambda_i}{(\sqrt{L} + \sqrt{\mu})^2}; \quad \operatorname{Im}(\lambda^M) = \frac{\pm 2\sqrt{(L - \lambda_i)(\lambda_i - \mu)}}{(\sqrt{L} + \sqrt{\mu})^2}; \quad |\lambda^M| = \frac{L - \mu}{(\sqrt{L} + \sqrt{\mu})^2}.$$

И скорость сходимости не зависит от шага и равна $\sqrt{\beta^*}$.

Квадратичная сходимость метода тяжелого шарика

i Theorem

Предположим, что f является квадратичной μ -строго выпуклой L -гладкой квадратичной функцией, тогда метод тяжелого шарика с параметрами

$$\alpha = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2}, \beta = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^2$$

сходится линейно:

$$\|x_k - x^*\|_2 \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x_0 - x^*\|$$

Глобальная сходимость метода тяжелого шарика³

i Theorem

Предположим, что f является гладкой и выпуклой и что

$$\beta \in [0, 1), \quad \alpha \in \left(0, \frac{2(1-\beta)}{L}\right).$$

Тогда последовательность $\{x_k\}$, генерируемая итерациями тяжелого шарика, удовлетворяет

$$f(\bar{x}_T) - f^* \leq \begin{cases} \frac{\|x_0 - x^*\|^2}{2(T+1)} \left(\frac{L\beta}{1-\beta} + \frac{1-\beta}{\alpha} \right), & \text{if } \alpha \in (0, \frac{1-\beta}{L}], \\ \frac{\|x_0 - x^*\|^2}{2(T+1)(2(1-\beta) - \alpha L)} \left(L\beta + \frac{(1-\beta)^2}{\alpha} \right), & \text{if } \alpha \in [\frac{1-\beta}{L}, \frac{2(1-\beta)}{L}), \end{cases}$$

где \bar{x}_T среднее Чезаро последовательности итераций, т.е.

$$\bar{x}_T = \frac{1}{T+1} \sum_{k=0}^T x_k.$$

³Глобальная сходимость метода тяжелого шарика для выпуклой оптимизации, Euhanna Ghadimi et.al.

Глобальная сходимость метода тяжелого шарика⁴

i Theorem

Предположим, что f является гладкой и строго выпуклой и что

$$\alpha \in (0, \frac{2}{L}), \quad 0 \leq \beta < \frac{1}{2} \left(\frac{\mu\alpha}{2} + \sqrt{\frac{\mu^2\alpha^2}{4} + 4(1 - \frac{\alpha L}{2})} \right).$$

Тогда последовательность $\{x_k\}$, генерируемая итерациями тяжелого шарика, сходится линейно к единственному оптимальному решению x^* . В частности,

$$f(x_k) - f^* \leq q^k(f(x_0) - f^*),$$

где $q \in [0, 1)$.

⁴Глобальная сходимость метода тяжелого шарика для выпуклой оптимизации, Euhanna Ghadimi et.al.

Итоги по методу тяжелого шарика

- Обеспечивает ускоренную сходимость для строго выпуклых квадратичных задач

Итоги по методу тяжелого шарика

- Обеспечивает ускоренную сходимость для строго выпуклых квадратичных задач
- Локально ускоренная сходимость была доказана в оригинальной статье.

Итоги по методу тяжелого шарика

- Обеспечивает ускоренную сходимость для строго выпуклых квадратичных задач
- Локально ускоренная сходимость была доказана в оригинальной статье.
- Недавно ⁵ было доказано, что глобальное ускорение сходимости для метода не существует.

⁵Provable non-accelerations of the heavy-ball method

Итоги по методу тяжелого шарика

- Обеспечивает ускоренную сходимость для строго выпуклых квадратичных задач
- Локально ускоренная сходимость была доказана в оригинальной статье.
- Недавно ⁵ было доказано, что глобальное ускорение сходимости для метода не существует.
- Метод не был чрезвычайно популярен до ML-бума.

⁵Provable non-accelerations of the heavy-ball method

Итоги по методу тяжелого шарика

- Обеспечивает ускоренную сходимость для строго выпуклых квадратичных задач
- Локально ускоренная сходимость была доказана в оригинальной статье.
- Недавно ⁵ было доказано, что глобальное ускорение сходимости для метода не существует.
- Метод не был чрезвычайно популярен до ML-бума.
- Сейчас он фактически является стандартом для практического ускорения методов градиентного спуска, даже для невыпуклых задач (обучение нейронных сетей)

⁵Provable non-accelerations of the heavy-ball method

Ускоренный градиентный метод Нестерова

Концепция ускоренного градиентного метода Нестерова

$$x_{k+1} = x_k - \alpha \nabla f(x_k) \quad x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}) \quad \begin{cases} y_{k+1} = x_k + \beta(x_k - x_{k-1}) \\ x_{k+1} = y_{k+1} - \alpha \nabla f(y_{k+1}) \end{cases}$$

Концепция ускоренного градиентного метода Нестерова

$$x_{k+1} = x_k - \alpha \nabla f(x_k) \quad x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$$

Давайте определим следующие обозначения

$$x^+ = x - \alpha \nabla f(x) \quad \text{Градиентный шаг}$$

$$d_k = \beta_k(x_k - x_{k-1}) \quad \text{Импульс}$$

Тогда мы можем записать:

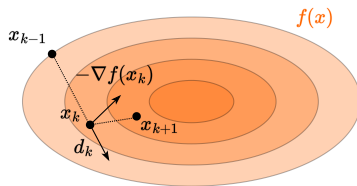
$$x_{k+1} = x_k^+ \quad \text{Градиентный спуск}$$

$$x_{k+1} = x_k^+ + d_k \quad \text{Метод тяжелого шарика}$$

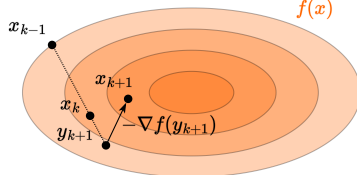
$$x_{k+1} = (x_k + d_k)^+ \quad \text{Ускоренный градиентный метод Нестерова}$$

$$\begin{cases} y_{k+1} = x_k + \beta(x_k - x_{k-1}) \\ x_{k+1} = y_{k+1} - \alpha \nabla f(y_{k+1}) \end{cases}$$

Polyak momentum



Nesterov momentum



Сходимость в общем случае

i Theorem

Предположим, что $f : \mathbb{R}^n \rightarrow \mathbb{R}$ является выпуклой и L -гладкой. Ускоренный градиентный метод Нестерова (NAG) предназначен для решения задачи минимизации, начиная с начальной точки $x_0 = y_0 \in \mathbb{R}^n$ и $\lambda_0 = 0$. Алгоритм выполняет следующие шаги:

Обновление градиента: $y_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$

Экстраполяция: $x_{k+1} = (1 - \gamma_k)y_{k+1} + \gamma_k y_k$

Вес экстраполяции: $\lambda_{k+1} = \frac{1 + \sqrt{1 + 4\lambda_k^2}}{2}$

Вес экстраполяции: $\gamma_k = \frac{1 - \lambda_k}{\lambda_{k+1}}$

Последовательность $\{f(y_k)\}_{k \in \mathbb{N}}$, генерируемая алгоритмом, сходится к оптимальному значению f^* со скоростью $\mathcal{O}\left(\frac{1}{k^2}\right)$, в частности:

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

Сходимость в общем случае

i Theorem

Предположим, что $f : \mathbb{R}^n \rightarrow \mathbb{R}$ является μ -строго выпуклой и L -гладкой. Ускоренный градиентный метод Нестерова (NAG) предназначен для решения задачи минимизации, начиная с начальной точки $x_0 = y_0 \in \mathbb{R}^n$ и $\lambda_0 = 0$. Алгоритм выполняет следующие шаги:

Обновление градиента: $y_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$

Экстраполяция: $x_{k+1} = (1 + \gamma_k)y_{k+1} - \gamma_k y_k$

Вес экстраполяции: $\gamma_k = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$

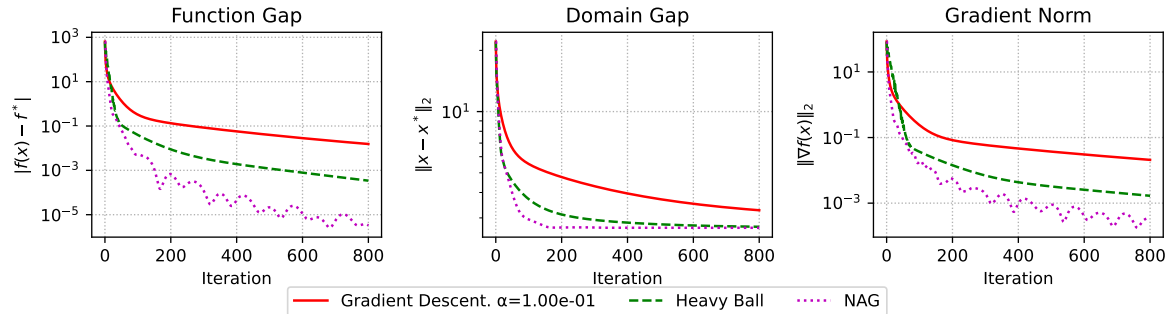
Последовательность $\{f(y_k)\}_{k \in \mathbb{N}}$, генерируемая алгоритмом, сходится к оптимальному значению f^* линейно:

$$f(y_k) - f^* \leq \frac{\mu + L}{2} \|x_0 - x^*\|_2^2 \exp\left(-\frac{k}{\sqrt{\kappa}}\right)$$

Численные эксперименты

Выпуклые квадратичные задачи (т.е. линейная регрессия)

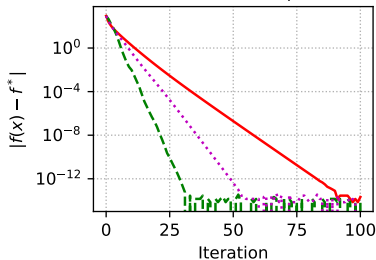
Convex quadratics: $n=60$, random matrix, $\mu=0$, $L=10$



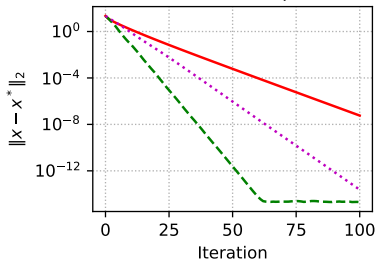
Строго выпуклые квадратичные задачи (т.е. регуляризованная линейная регрессия)

Strongly convex quadratics: $n=60$, random matrix, $\mu=1$, $L=10$

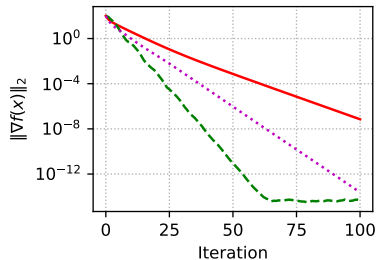
Function Gap



Domain Gap



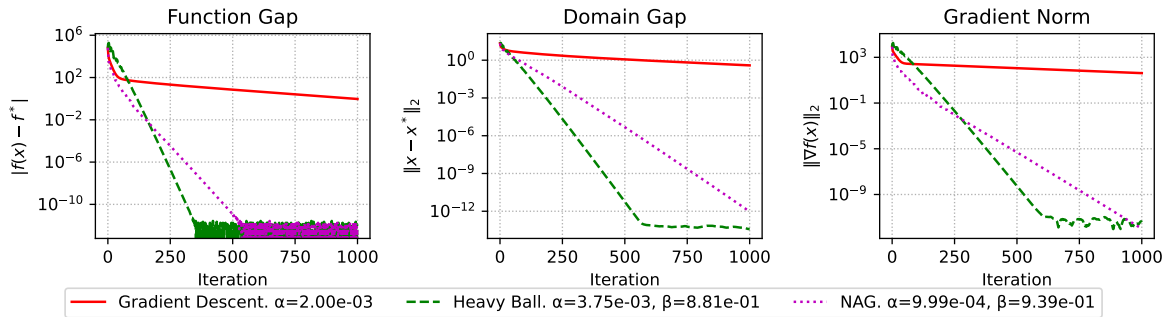
Gradient Norm



— Gradient Descent. $\alpha=1.67\text{e-}01$ - - - Heavy Ball. $\alpha=2.15\text{e-}01$, $\beta=2.88\text{e-}01$ NAG. $\alpha=9.09\text{e-}02$, $\beta=5.37\text{e-}01$

Строго выпуклые квадратичные задачи (т.е. регуляризованная линейная регрессия)

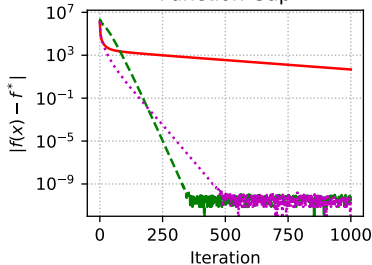
Strongly convex quadratics: $n=60$, random matrix, $\mu=1$, $L=1000$



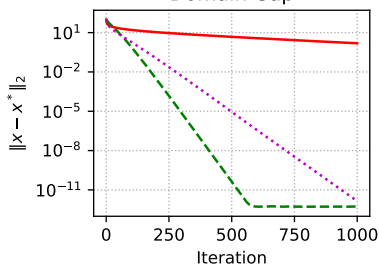
Строго выпуклые квадратичные задачи (т.е. регуляризованная линейная регрессия)

Strongly convex quadratics: $n=1000$, random matrix, $\mu=1$, $L=1000$

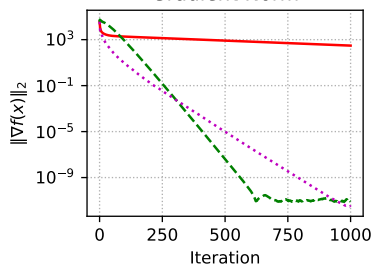
Function Gap



Domain Gap



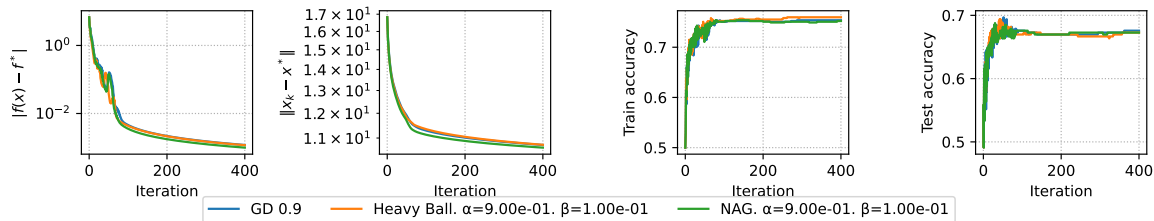
Gradient Norm



— Gradient Descent. $\alpha=2.00\text{e-}03$ - - - Heavy Ball. $\alpha=3.75\text{e-}03, \beta=8.81\text{e-}01$ ····· NAG. $\alpha=9.99\text{e-}04, \beta=9.39\text{e-}01$

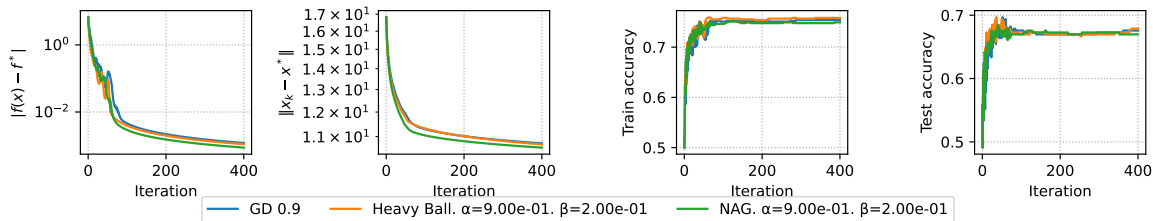
Выпуклая бинарная логистическая регрессия

Convex binary logistic regression. $\mu=0$.



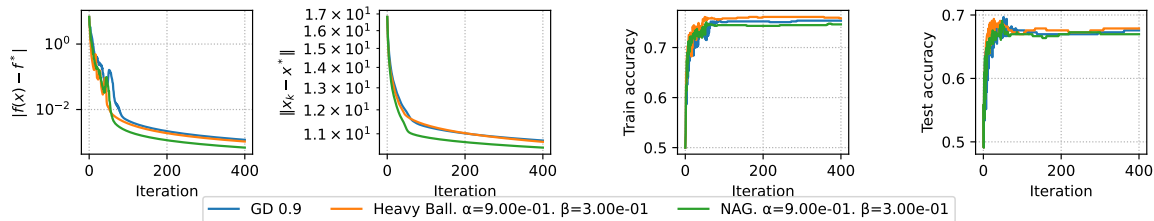
Выпуклая бинарная логистическая регрессия

Convex binary logistic regression. $\mu=0$.



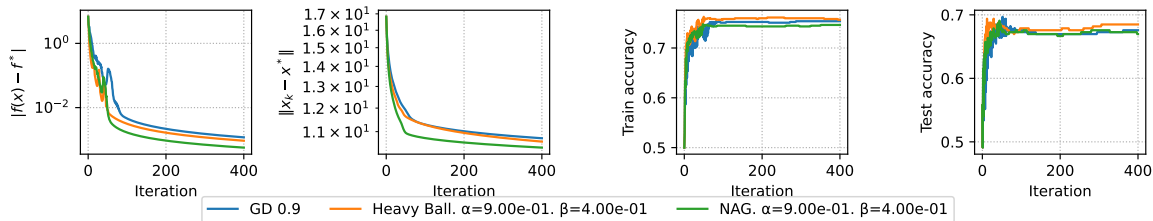
Выпуклая бинарная логистическая регрессия

Convex binary logistic regression. $\mu=0$.



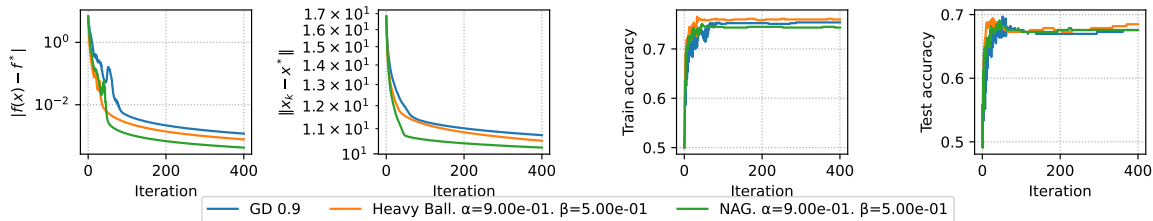
Выпуклая бинарная логистическая регрессия

Convex binary logistic regression. $\mu=0$.



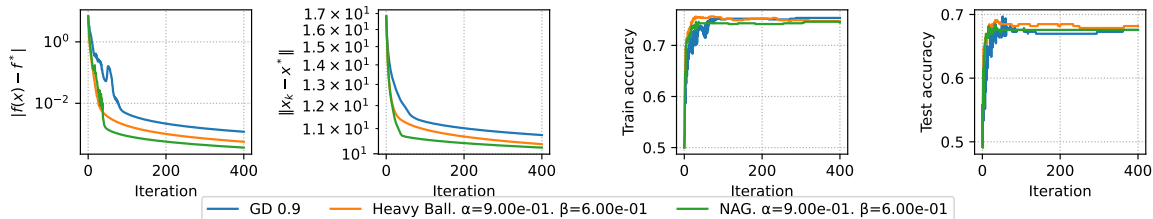
Выпуклая бинарная логистическая регрессия

Convex binary logistic regression. $\mu=0$.



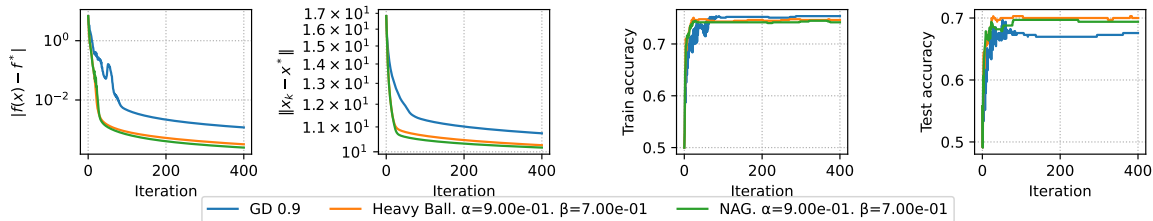
Выпуклая бинарная логистическая регрессия

Convex binary logistic regression. $\mu=0$.



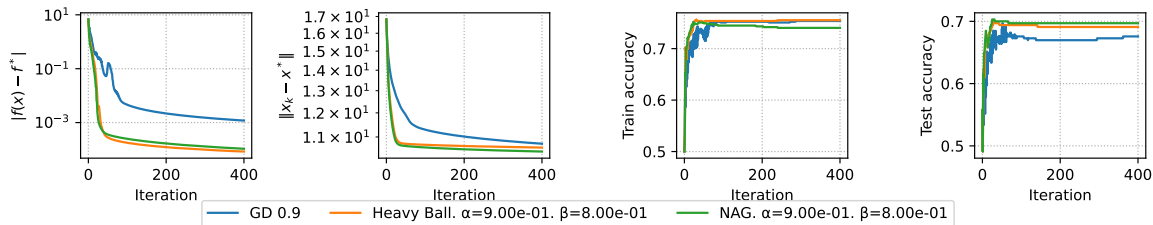
Выпуклая бинарная логистическая регрессия

Convex binary logistic regression. $\mu=0$.



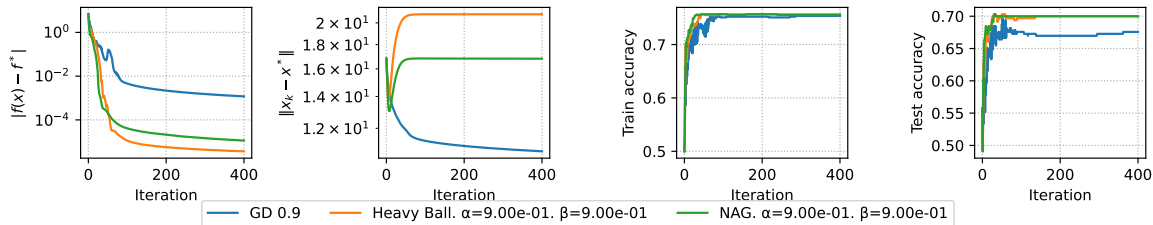
Выпуклая бинарная логистическая регрессия

Convex binary logistic regression. $\mu=0$.



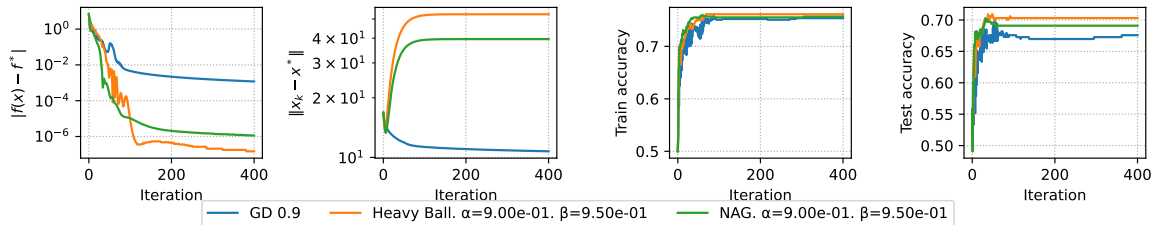
Выпуклая бинарная логистическая регрессия

Convex binary logistic regression. $\mu=0$.



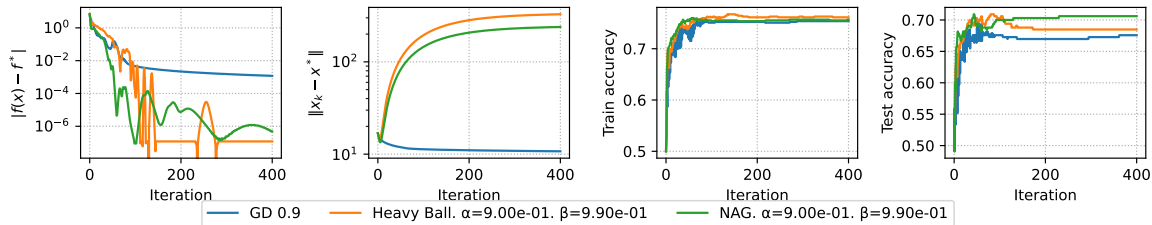
Выпуклая бинарная логистическая регрессия

Convex binary logistic regression. $\mu=0$.



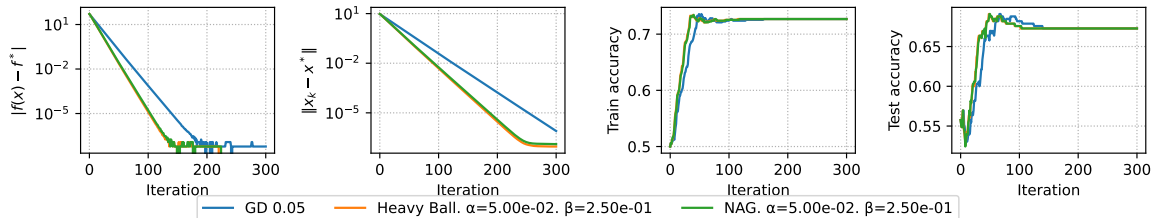
Выпуклая бинарная логистическая регрессия

Convex binary logistic regression. $\mu=0$.



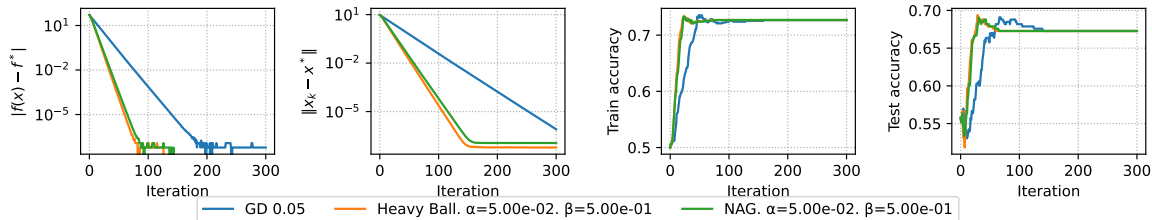
Строго выпуклая бинарная логистическая регрессия

Strongly convex binary logistic regression. $\mu=1$.



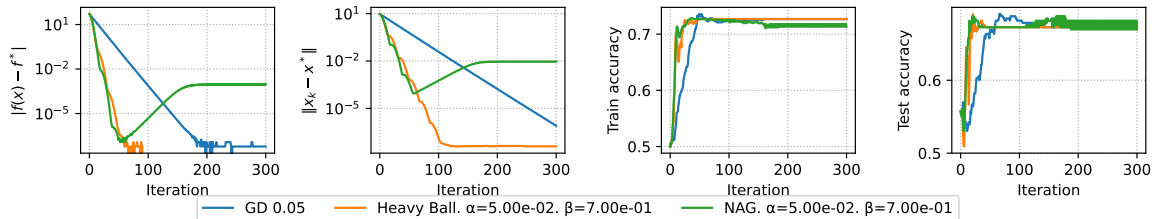
Строго выпуклая бинарная логистическая регрессия

Strongly convex binary logistic regression. $\mu=1$.



Строго выпуклая бинарная логистическая регрессия

Strongly convex binary logistic regression. $\mu=1$.



Строго выпуклая бинарная логистическая регрессия

Strongly convex binary logistic regression. $\mu=1$.

