

Ускоренные градиентные методы

МЕТОДЫ ВЫПУКЛОЙ ОПТИМИЗАЦИИ

НЕДЕЛЯ 8

Даня Меркулов
Пётр Остроухов



Даня Меркулов

Оптимизация для всех! ЦУ



Сильно выпуклые квадратичные функции

Сдвиг координат



Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

Сдвиг координат



Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

- Во-первых, без ограничения общности мы можем установить $c = 0$, что не повлияет на процесс оптимизации.



Сдвиг координат



Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

- Во-первых, без ограничения общности мы можем установить $c = 0$, что не повлияет на процесс оптимизации.
- Во-вторых, у нас есть спектральное разложение матрицы $A = Q\Lambda Q^\top$.



Сдвиг координат



Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

- Во-первых, без ограничения общности мы можем установить $c = 0$, что не повлияет на процесс оптимизации.
- Во-вторых, у нас есть спектральное разложение матрицы $A = Q\Lambda Q^\top$.
- Покажем, что мы можем сделать сдвиг координат, чтобы сделать анализ немного проще. Пусть $\hat{x} = Q^\top(x - x^*)$, где x^* — точка минимума исходной функции, определяемая как $Ax^* = b$. При этом $x = Q\hat{x} + x^*$.



Сдвиг координат



Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

- Во-первых, без ограничения общности мы можем установить $c = 0$, что не повлияет на процесс оптимизации.
- Во-вторых, у нас есть спектральное разложение матрицы $A = Q \Lambda Q^\top$.
- Покажем, что мы можем сделать сдвиг координат, чтобы сделать анализ немного проще. Пусть $\hat{x} = Q^\top(x - x^*)$, где x^* — точка минимума исходной функции, определяемая как $Ax^* = b$. При этом $x = Q\hat{x} + x^*$.

$$f(\hat{x}) = \frac{1}{2} (Q\hat{x} + x^*)^\top A (Q\hat{x} + x^*) - b^\top (Q\hat{x} + x^*)$$



Сдвиг координат



Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

- Во-первых, без ограничения общности мы можем установить $c = 0$, что не повлияет на процесс оптимизации.
- Во-вторых, у нас есть спектральное разложение матрицы $A = Q\Lambda Q^T$.
- Покажем, что мы можем сделать сдвиг координат, чтобы сделать анализ немного проще. Пусть $\hat{x} = Q^T(x - x^*)$, где x^* — точка минимума исходной функции, определяемая как $Ax^* = b$. При этом $x = Q\hat{x} + x^*$.

$$\begin{aligned} f(\hat{x}) &= \frac{1}{2} (Q\hat{x} + x^*)^\top A (Q\hat{x} + x^*) - b^\top (Q\hat{x} + x^*) \\ &= \frac{1}{2} \hat{x}^\top Q^\top A Q \hat{x} + \frac{1}{2} (x^*)^\top A (x^*) + (x^*)^\top A Q \hat{x} - b^\top Q \hat{x} - b^\top x^* \end{aligned}$$



Сдвиг координат



Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

- Во-первых, без ограничения общности мы можем установить $c = 0$, что не повлияет на процесс оптимизации.
- Во-вторых, у нас есть спектральное разложение матрицы $A = Q\Lambda Q^T$.
- Покажем, что мы можем сделать сдвиг координат, чтобы сделать анализ немного проще. Пусть $\hat{x} = Q^T(x - x^*)$, где x^* — точка минимума исходной функции, определяемая как $Ax^* = b$. При этом $x = Q\hat{x} + x^*$.

$$\begin{aligned} f(\hat{x}) &= \frac{1}{2} (Q\hat{x} + x^*)^\top A (Q\hat{x} + x^*) - b^\top (Q\hat{x} + x^*) \\ &= \frac{1}{2} \hat{x}^\top Q^\top A Q \hat{x} + \frac{1}{2} (x^*)^\top A (x^*) + (x^*)^\top A Q \hat{x} - b^\top Q \hat{x} - b^\top x^* \\ &= \frac{1}{2} \hat{x}^\top \Lambda \hat{x} + \frac{1}{2} (x^*)^\top A (x^*) + (x^*)^\top A Q \hat{x} - (x^*)^\top A^\top Q \hat{x} - (x^*)^\top A x^* \end{aligned}$$



Сдвиг координат



Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

- Во-первых, без ограничения общности мы можем установить $c = 0$, что не повлияет на процесс оптимизации.
- Во-вторых, у нас есть спектральное разложение матрицы $A = Q\Lambda Q^T$.
- Покажем, что мы можем сделать сдвиг координат, чтобы сделать анализ немного проще. Пусть $\hat{x} = Q^T(x - x^*)$, где x^* — точка минимума исходной функции, определяемая как $Ax^* = b$. При этом $x = Q\hat{x} + x^*$.

$$\begin{aligned} f(\hat{x}) &= \frac{1}{2} (Q\hat{x} + x^*)^\top A (Q\hat{x} + x^*) - b^\top (Q\hat{x} + x^*) \\ &= \frac{1}{2} \hat{x}^\top Q^\top A Q \hat{x} + \frac{1}{2} (x^*)^\top A (x^*) + (x^*)^\top A Q \hat{x} - b^\top Q \hat{x} - b^\top x^* \\ &= \frac{1}{2} \hat{x}^\top \Lambda \hat{x} + \frac{1}{2} (x^*)^\top A (x^*) + (x^*)^\top A Q \hat{x} - (x^*)^\top A^\top Q \hat{x} - (x^*)^\top A x^* \\ &= \frac{1}{2} \hat{x}^\top \Lambda \hat{x} - \frac{1}{2} (x^*)^\top A x^* \end{aligned}$$



Сдвиг координат



Рассмотрим следующую задачу квадратичной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^\top A x - b^\top x + c, \text{ где } A \in \mathbb{S}_{++}^d.$$

- Во-первых, без ограничения общности мы можем установить $c = 0$, что не повлияет на процесс оптимизации.
- Во-вторых, у нас есть спектральное разложение матрицы $A = Q\Lambda Q^T$.
- Покажем, что мы можем сделать сдвиг координат, чтобы сделать анализ немного проще. Пусть $\hat{x} = Q^T(x - x^*)$, где x^* — точка минимума исходной функции, определяемая как $Ax^* = b$. При этом $x = Q\hat{x} + x^*$.

$$\begin{aligned} f(\hat{x}) &= \frac{1}{2} (Q\hat{x} + x^*)^\top A (Q\hat{x} + x^*) - b^\top (Q\hat{x} + x^*) \\ &= \frac{1}{2} \hat{x}^\top Q^\top A Q \hat{x} + \frac{1}{2} (x^*)^\top A (x^*) + (x^*)^\top A Q \hat{x} - b^\top Q \hat{x} - b^\top x^* \\ &= \frac{1}{2} \hat{x}^\top \Lambda \hat{x} + \frac{1}{2} (x^*)^\top A (x^*) + (x^*)^\top A Q \hat{x} - (x^*)^\top A^\top Q \hat{x} - (x^*)^\top A x^* \\ &= \frac{1}{2} \hat{x}^\top \Lambda \hat{x} - \frac{1}{2} (x^*)^\top A x^* \simeq \frac{1}{2} \hat{x}^\top \Lambda \hat{x} \end{aligned}$$



Анализ сходимости



Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

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

Анализ сходимости



Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

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

Анализ сходимости



Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\ &= (I - \alpha^k \Lambda)x^k\end{aligned}$$

Анализ сходимости



Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

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

$$= (I - \alpha^k \Lambda) x^k$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

Анализ сходимости



Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Анализ сходимости

Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

$$\begin{aligned} x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\ &= (I - \alpha^k \Lambda) x^k \end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

Анализ сходимости

Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

$$\begin{aligned} x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\ &= (I - \alpha^k \Lambda) x^k \end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$|1 - \alpha \mu| < 1$$

Анализ сходимости

Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

Анализ сходимости

Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

$$\begin{aligned} x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\ &= (I - \alpha^k \Lambda) x^k \end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

Анализ сходимости

Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$\begin{aligned}|1 - \alpha\mu| &< 1 & |1 - \alpha L| &< 1 \\-1 &< 1 - \alpha\mu < 1 \\ \alpha &< \frac{2}{\mu} & \alpha\mu &> 0\end{aligned}$$

Анализ сходимости

Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

Анализ сходимости

Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L} \quad \alpha L > 0$$

Анализ сходимости

Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

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

$$= (I - \alpha^k \Lambda) x^k$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L} \quad \alpha L > 0$$

Анализ сходимости



Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

Выберем α , минимизирующий худший знаменатель прогрессии

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$\rho^* = \min_{\alpha} \rho(\alpha)$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$|1 - \alpha \mu| < 1$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$\alpha < \frac{2}{L} \quad \alpha L > 0$$

Анализ сходимости



Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

Выберем α , минимизирующий худший знаменатель прогрессии

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$\rho^* = \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}|$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$|1 - \alpha \mu| < 1$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$\alpha < \frac{2}{L} \quad \alpha L > 0$$

Анализ сходимости



Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

Выберем α , минимизирующий худший знаменатель прогрессии

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

$$\begin{aligned}\rho^* &= \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}| \\&= \min_{\alpha} \max \{ |1 - \alpha \mu|, |1 - \alpha L| \}\end{aligned}$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L} \quad \alpha L > 0$$

Анализ сходимости



Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

Выберем α , минимизирующий худший знаменатель прогрессии

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

$$\rho^* = \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}|$$

$$= \min_{\alpha} \max \{ |1 - \alpha \mu|, |1 - \alpha L| \}$$

$$\alpha^* : \quad 1 - \alpha^* \mu = \alpha^* L - 1$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L} \quad \alpha L > 0$$

Анализ сходимости



Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

Выберем α , минимизирующий худший знаменатель прогрессии

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L} \quad \alpha L > 0$$

$$\rho^* = \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}|$$

$$= \min_{\alpha} \max \{|1 - \alpha \mu|, |1 - \alpha L|\}$$

$$\alpha^* : 1 - \alpha^* \mu = \alpha^* L - 1$$

$$\alpha^* = \frac{2}{\mu + L}$$

Анализ сходимости



Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

Выберем α , минимизирующий худший знаменатель прогрессии

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L} \quad \alpha L > 0$$

$$\rho^* = \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}|$$

$$= \min_{\alpha} \max \{ |1 - \alpha \mu|, |1 - \alpha L| \}$$

$$\alpha^* : 1 - \alpha^* \mu = \alpha^* L - 1$$

$$\alpha^* = \frac{2}{\mu + L} \quad \rho^* = \frac{L - \mu}{L + \mu}$$

Анализ сходимости



Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

Выберем α , минимизирующий худший знаменатель прогрессии

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L} \quad \alpha L > 0$$

$$\rho^* = \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}|$$

$$= \min_{\alpha} \max \{ |1 - \alpha \mu|, |1 - \alpha L| \}$$

$$\alpha^* : 1 - \alpha^* \mu = \alpha^* L - 1$$

$$\alpha^* = \frac{2}{\mu + L} \quad \rho^* = \frac{L - \mu}{L + \mu}$$

$$|x_{(i)}^k| \leq \left(\frac{L - \mu}{L + \mu} \right)^k |x_{(i)}^0|$$

Анализ сходимости



Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

Выберем α , минимизирующий худший знаменатель прогрессии

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L} \quad \alpha L > 0$$

$$\rho^* = \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}|$$

$$= \min_{\alpha} \max \{|1 - \alpha \mu|, |1 - \alpha L|\}$$

$$\alpha^* : 1 - \alpha^* \mu = \alpha^* L - 1$$

$$\alpha^* = \frac{2}{\mu + L} \quad \rho^* = \frac{L - \mu}{L + \mu}$$

$$|x_{(i)}^k| \leq \left(\frac{L - \mu}{L + \mu} \right)^k |x_{(i)}^0|$$

$$\|x^k\|_2 \leq \left(\frac{L - \mu}{L + \mu} \right)^k \|x^0\|_2$$

Анализ сходимости



Теперь мы можем работать с функцией $f(x) = \frac{1}{2}x^T \Lambda x$ с $x^* = 0$ без ограничения общности (убрав крышку из \hat{x})

Выберем α , минимизирующий худший знаменатель прогрессии

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\&= (I - \alpha^k \Lambda) x^k\end{aligned}$$

$$x_{(i)}^{k+1} = (1 - \alpha^k \lambda_{(i)}) x_{(i)}^k \quad \text{для } i\text{-й координаты}$$

$$x_{(i)}^k = (1 - \alpha \lambda_{(i)})^k x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha$$

Используем постоянный шаг $\alpha^k = \alpha$. Условие сходимости:

$$\rho(\alpha) = \max_i |1 - \alpha \lambda_{(i)}| < 1$$

...

Помним, что $\lambda_{\min} = \mu > 0$, $\lambda_{\max} = L \geq \mu$.

$$|1 - \alpha \mu| < 1$$

$$-1 < 1 - \alpha \mu < 1$$

$$\alpha < \frac{2}{\mu} \quad \alpha \mu > 0$$

$$|1 - \alpha L| < 1$$

$$-1 < 1 - \alpha L < 1$$

$$\alpha < \frac{2}{L} \quad \alpha L > 0$$

$$\rho^* = \min_{\alpha} \rho(\alpha) = \min_{\alpha} \max_i |1 - \alpha \lambda_{(i)}|$$

$$= \min_{\alpha} \max \{|1 - \alpha \mu|, |1 - \alpha L|\}$$

$$\alpha^* : 1 - \alpha^* \mu = \alpha^* L - 1$$

$$\alpha^* = \frac{2}{\mu + L} \quad \rho^* = \frac{L - \mu}{L + \mu}$$

$$|x_{(i)}^k| \leq \left(\frac{L - \mu}{L + \mu} \right)^k |x_{(i)}^0|$$

$$\|x^k\|_2 \leq \left(\frac{L - \mu}{L + \mu} \right)^k \|x^0\|_2 \quad f(x^k) \leq \left(\frac{L - \mu}{L + \mu} \right)^{2k} f(x^0)$$

Анализ сходимости



Таким образом, имеем линейную сходимость по аргументу со скоростью $\frac{\varkappa-1}{\varkappa+1} = 1 - \frac{2}{\varkappa+1}$, где $\varkappa = \frac{L}{\mu}$ — число обусловленности квадратичной задачи.

\varkappa	ρ	Итераций до уменьшения ошибки по аргументу в	Итераций до уменьшения ошибки по функции в 10
		10 раз	раз
1.1	0.05	1	1
2	0.33	3	2
5	0.67	6	3
10	0.82	12	6
50	0.96	58	29
100	0.98	116	58
500	0.996	576	288
1000	0.998	1152	576

Число обусловленности \mathcal{N}



Случай PL-функций

PL-функции. Линейная сходимость градиентного спуска без выпуклости

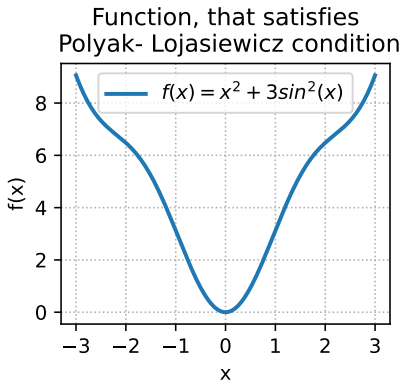
Говорят, что f удовлетворяет условию Поляка-Лоясиевича (PL), если для некоторого $\mu > 0$ выполняется

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x$$

Интересно, что градиентный спуск может сходиться линейно даже без выпуклости.

Следующие функции удовлетворяют условию PL, но не являются выпуклыми. Код

$$f(x) = x^2 + 3 \sin^2(x)$$



PL-функции. Линейная сходимость градиентного спуска без выпуклости

Говорят, что f удовлетворяет условию Поляка-Лоясиевича (PL), если для некоторого $\mu > 0$ выполняется

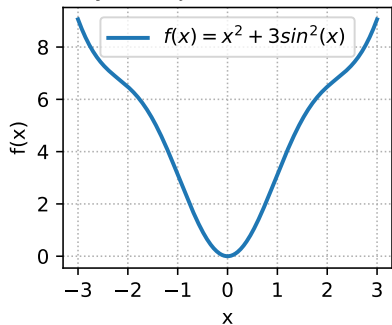
$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x$$

Интересно, что градиентный спуск может сходиться линейно даже без выпуклости.

Следующие функции удовлетворяют условию PL, но не являются выпуклыми.  Код

$$f(x) = x^2 + 3 \sin^2(x)$$

Function, that satisfies
Polyak-Lojasiewicz condition



$$f(x, y) = \frac{(y - \sin x)^2}{2}$$

Non-convex PL function



i Theorem

Рассмотрим задачу

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

и предположим, что f является PL-функцией с константой μ и L -гладкой, для некоторых $L \geq \mu > 0$.

Рассмотрим последовательность $(x^k)_{k \in \mathbb{N}}$, сгенерированную методом градиентного спуска из точки x^0 с постоянным шагом α , удовлетворяющим $0 < \alpha \leq \frac{1}{L}$. Пусть $f^* = \min_{x \in \mathbb{R}^d} f(x)$. Тогда:

$$f(x^k) - f^* \leq (1 - \alpha\mu)^k (f(x^0) - f^*).$$

i Theorem

Рассмотрим задачу

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

и предположим, что f является PL-функцией с константой μ и L -гладкой, для некоторых $L \geq \mu > 0$.

Рассмотрим последовательность $(x^k)_{k \in \mathbb{N}}$, сгенерированную методом градиентного спуска из точки x^0 с постоянным шагом α , удовлетворяющим $0 < \alpha \leq \frac{1}{L}$. Пусть $f^* = \min_{x \in \mathbb{R}^d} f(x)$. Тогда:

$$f(x^k) - f^* \leq (1 - \alpha\mu)^k (f(x^0) - f^*).$$

i Theorem

Если функция $f(x)$ дифференцируема и μ -сильно выпукла, то она является PL-функцией.

Анализ сходимости



Используем L -гладкость вместе с правилом обновления, чтобы записать:

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

Анализ сходимости



Используем L -гладкость вместе с правилом обновления, чтобы записать:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \alpha \|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2} \|\nabla f(x^k)\|^2 \end{aligned}$$

Анализ сходимости



Используем L -гладкость вместе с правилом обновления, чтобы записать:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \alpha \|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2} \|\nabla f(x^k)\|^2 \\ &= f(x^k) - \frac{\alpha}{2} (2 - L\alpha) \|\nabla f(x^k)\|^2 \end{aligned}$$

Анализ сходимости



Используем L -гладкость вместе с правилом обновления, чтобы записать:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \alpha \|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2} \|\nabla f(x^k)\|^2 \\ &= f(x^k) - \frac{\alpha}{2} (2 - L\alpha) \|\nabla f(x^k)\|^2 \\ &\leq f(x^k) - \frac{\alpha}{2} \|\nabla f(x^k)\|^2, \end{aligned}$$

Анализ сходимости



Используем L -гладкость вместе с правилом обновления, чтобы записать:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \alpha \|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2} \|\nabla f(x^k)\|^2 \\ &= f(x^k) - \frac{\alpha}{2} (2 - L\alpha) \|\nabla f(x^k)\|^2 \\ &\leq f(x^k) - \frac{\alpha}{2} \|\nabla f(x^k)\|^2, \end{aligned}$$

Анализ сходимости



Используем L -гладкость вместе с правилом обновления, чтобы записать:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \alpha \|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2} \|\nabla f(x^k)\|^2 \\ &= f(x^k) - \frac{\alpha}{2} (2 - L\alpha) \|\nabla f(x^k)\|^2 \\ &\leq f(x^k) - \frac{\alpha}{2} \|\nabla f(x^k)\|^2, \end{aligned}$$

где в последнем неравенстве использована гипотеза о шаге $\alpha L \leq 1$.

Анализ сходимости



Используем L -гладкость вместе с правилом обновления, чтобы записать:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \alpha \|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2} \|\nabla f(x^k)\|^2 \\ &= f(x^k) - \frac{\alpha}{2} (2 - L\alpha) \|\nabla f(x^k)\|^2 \\ &\leq f(x^k) - \frac{\alpha}{2} \|\nabla f(x^k)\|^2, \end{aligned}$$

где в последнем неравенстве использована гипотеза о шаге $\alpha L \leq 1$.

Теперь используем свойство PL-функции и получаем:

$$f(x^{k+1}) \leq f(x^k) - \alpha \mu (f(x^k) - f^*).$$

Вытя f^* из обеих частей этого неравенства и применив рекурсию, мы получим искомый результат.

Выпуклый гладкий случай

Выпуклый гладкий случай



i Theorem

Рассмотрим задачу

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

Пусть $x^* = \arg \min_{x \in \mathbb{R}^d} f(x)$, а $f^* = f(x^*)$. Предположим, что $f : \mathbb{R}^d \rightarrow \mathbb{R}$ является выпуклой и L -гладкой функцией, для некоторого $L > 0$. Пусть $(x_k)_{k \in \mathbb{N}}$ — последовательность итераций, сгенерированная методом градиентного спуска из точки x_0 с постоянным шагом α , удовлетворяющим $0 < \alpha \leq \frac{1}{L}$.

Тогда для всех $k \in \mathbb{N}$ справедливо:

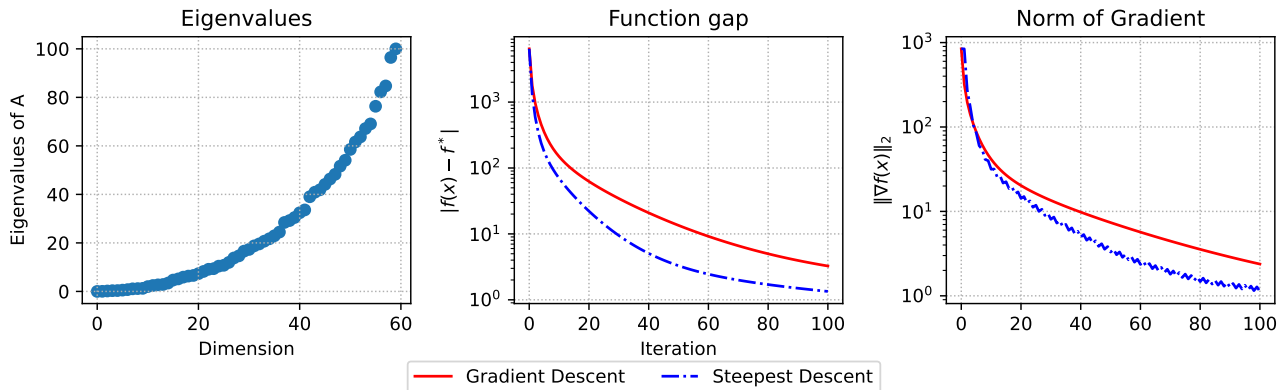
$$f(x_k) - f^* \leq \frac{\|x_0 - x^*\|^2}{2\alpha k}.$$

Численные эксперименты



$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{2} x^T A x - b^T x \right\}, \mu = 0, L = 100.$$

Convex quadratics. $n=60$, random matrix.

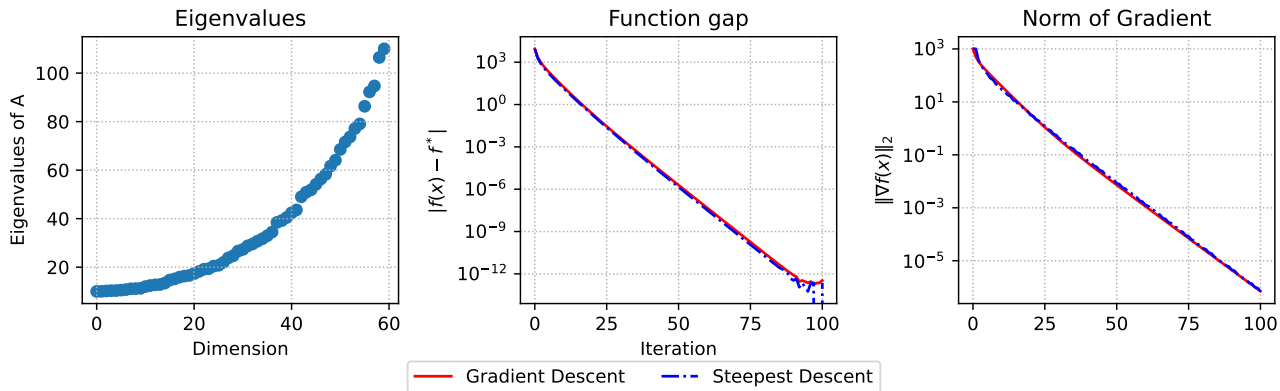


Численные эксперименты



$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{2} x^T A x - b^T x \right\}, \mu = 10, L = 110.$$

Strongly convex quadratics. $n=60$, random matrix.

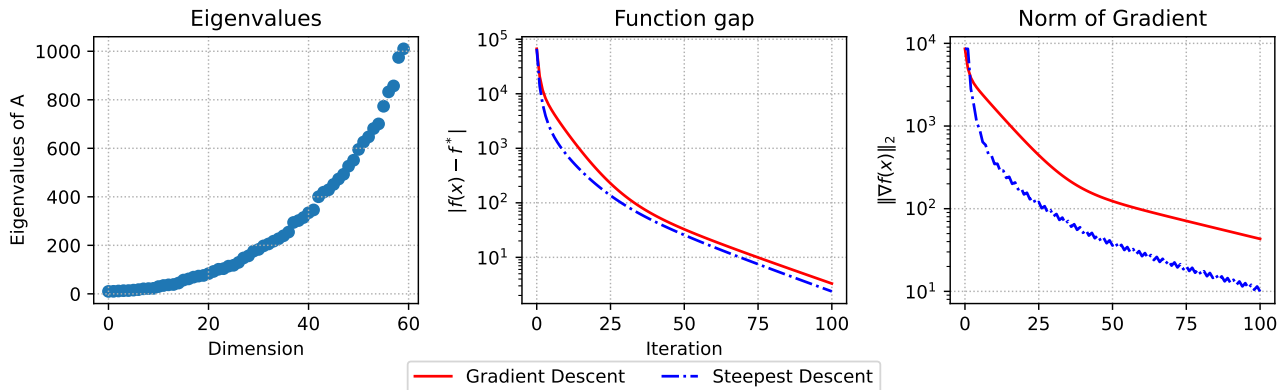


Численные эксперименты



$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{2} x^T A x - b^T x \right\}, \mu = 10, L = 1000.$$

Strongly convex quadratics. $n=60$, random matrix.

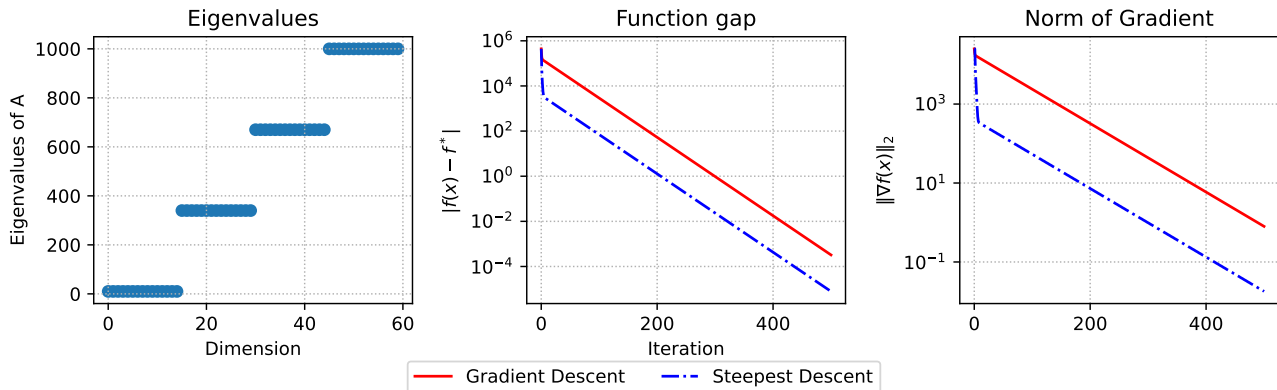


Численные эксперименты



$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{2} x^T A x - b^T x \right\}, \mu = 10, L = 1000.$$

Strongly convex quadratics. $n=60$, clustered matrix.

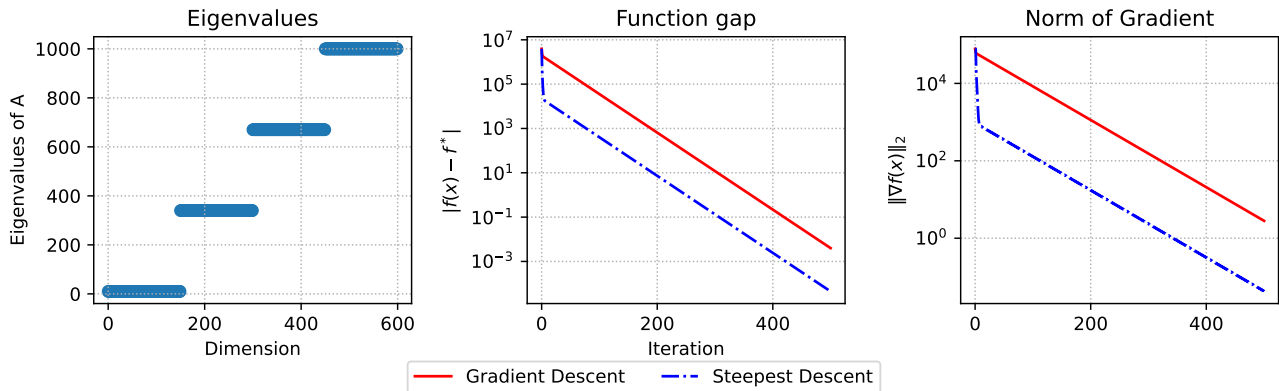


Численные эксперименты



$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{2} x^T A x - b^T x \right\}, \mu = 10, L = 1000.$$

Strongly convex quadratics. $n=600$, clustered matrix.

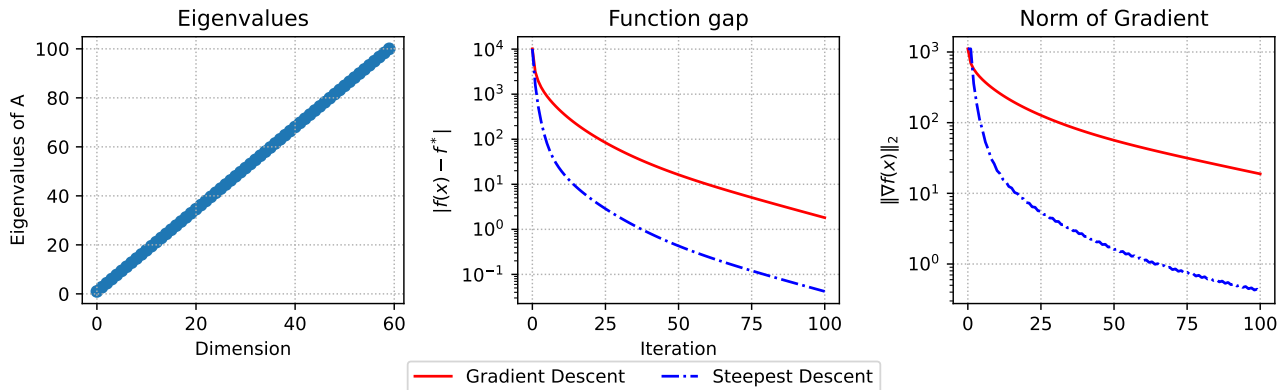


Численные эксперименты



$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{2} x^T A x - b^T x \right\}, \mu = 10, L = 1000.$$

Strongly convex quadratics. $n=60$, uniform spectrum matrix.

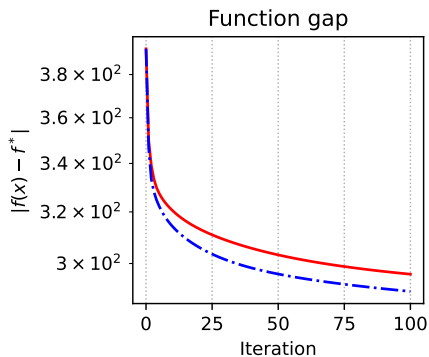
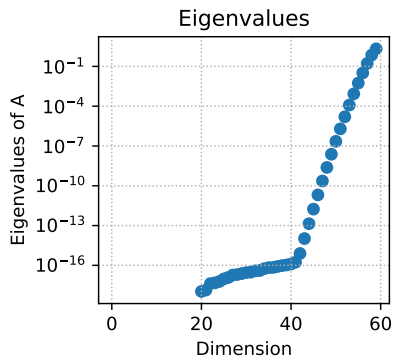


Численные эксперименты

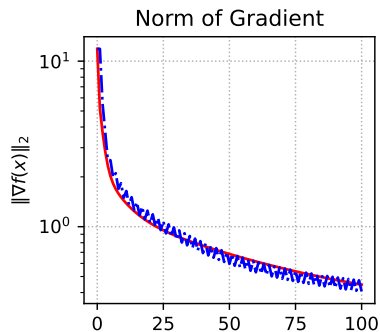


$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{2} x^T A x - b^T x \right\}$$

Strongly convex quadratics. $n=60$, Hilbert matrix.



— Gradient Descent - · - Steepest Descent

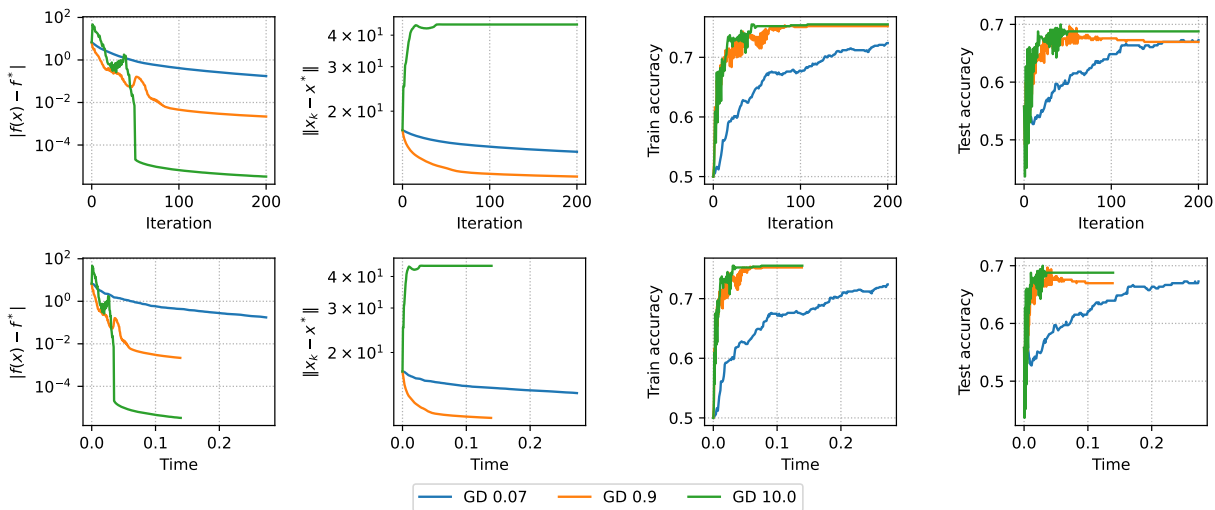


Численные эксперименты



$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{\mu}{2} \|x\|_2^2 + \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-y_i \langle a_i, x \rangle)) \right\}$$

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

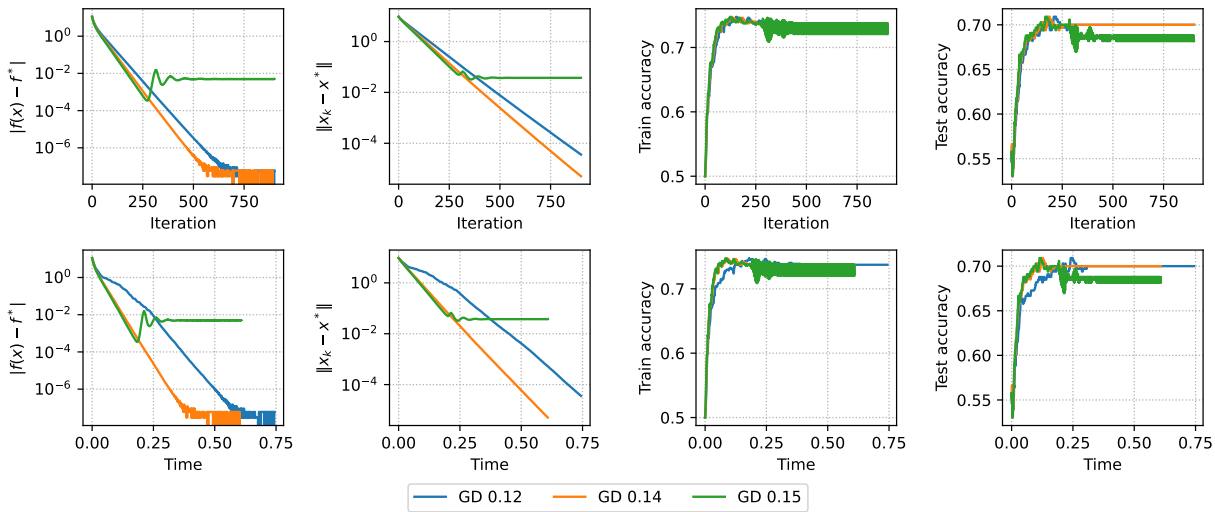


Численные эксперименты



$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{\mu}{2} \|x\|_2^2 + \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-y_i \langle a_i, x \rangle)) \right\}$$

Strongly convex binary logistic regression. $\mu=0.1$.

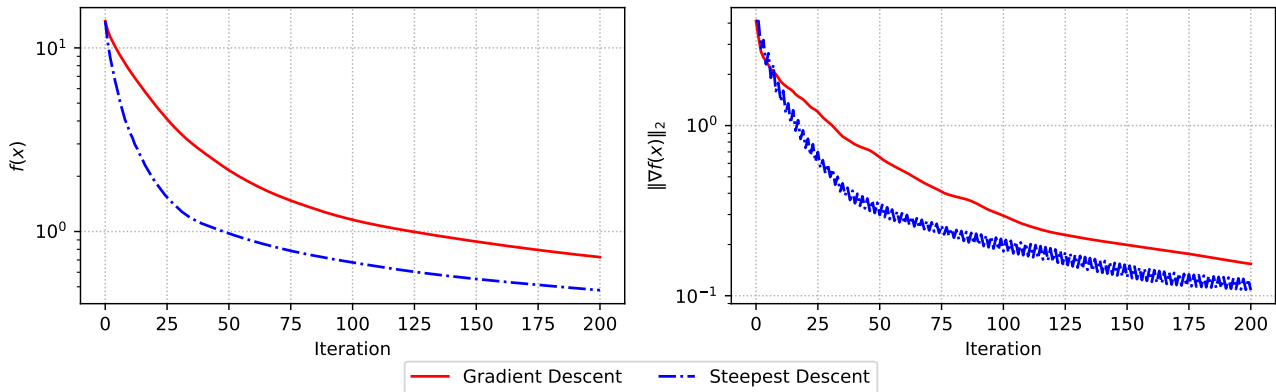


Численные эксперименты



$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{\mu}{2} \|x\|_2^2 + \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-y_i \langle a_i, x \rangle)) \right\}$$

Regularized binary logistic regression. $n=300$. $m=1000$. $\mu=0$

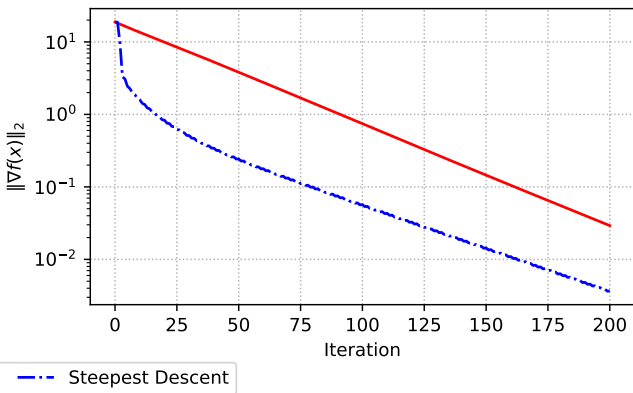
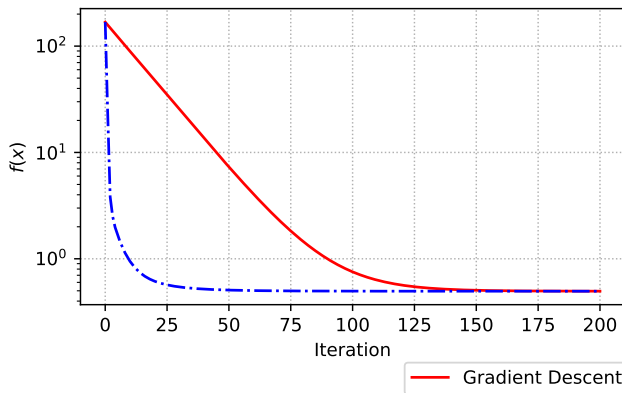


Численные эксперименты



$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{\mu}{2} \|x\|_2^2 + \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-y_i \langle a_i, x \rangle)) \right\}$$

Regularized binary logistic regression. $n=300$. $m=1000$. $\mu=1$



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 & сильно выпуклая
$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\ = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\ x_k - x^*\ ^2 = \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\mu} \log \frac{1}{\varepsilon}\right)$

Сходимость градиентного спуска



Градиентный спуск: $\min_{x \in \mathbb{R}^n} f(x) \quad x_{k+1} = x_k - \alpha_k \nabla f(x_k)$

выпуклая (негладкая)	гладкая (невыпуклая)	гладкая & выпуклая	гладкая & сильно выпуклая
$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\ = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\ x_k - x^*\ ^2 = \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon = \mathcal{O}\left(\log \frac{1}{\varepsilon}\right)$

Для гладкой сильно выпуклой функции мы имеем:

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*).$$

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 & сильно выпуклая
$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\ = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\ x_k - x^*\ ^2 = \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon = \mathcal{O}\left(\log \frac{1}{\varepsilon}\right)$

Для гладкой сильно 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 её 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 & сильно выпуклая
$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\ = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\ x_k - x^*\ ^2 = \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\mu \log \frac{1}{\varepsilon}}\right)$

Для гладкой сильно выпуклой функции мы имеем:

Наконец:

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*).$$

$$\varepsilon = f(x_{k_\varepsilon}) - f^*$$

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

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

Сходимость градиентного спуска



Градиентный спуск: $\min_{x \in \mathbb{R}^n} f(x) \quad x_{k+1} = x_k - \alpha_k \nabla f(x_k)$

выпуклая (негладкая)	гладкая (невыпуклая)	гладкая & выпуклая	гладкая & сильно выпуклая
$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\ = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\ x_k - x^*\ ^2 = \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\mu} \log \frac{1}{\varepsilon}\right)$

Для гладкой сильно выпуклой функции мы имеем:

Наконец:

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*).$$

$$\varepsilon = f(x_{k_\varepsilon}) - f^* \leq \left(1 - \frac{\mu}{L}\right)^{k_\varepsilon} (f(x_0) - f^*)$$

Обратите внимание, что для любого x , поскольку e^{-x} выпуклая и $1 - x$ является её 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 & сильно выпуклая
$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\ = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\ x_k - x^*\ ^2 = \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon = \mathcal{O}\left(\log \frac{1}{\varepsilon}\right)$

Для гладкой сильно выпуклой функции мы имеем:

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*).$$

Наконец:

$$\begin{aligned} \varepsilon = f(x_{k_\varepsilon}) - f^* &\leq \left(1 - \frac{\mu}{L}\right)^{k_\varepsilon} (f(x_0) - f^*) \\ &\leq \exp\left(-k_\varepsilon \frac{\mu}{L}\right) (f(x_0) - f^*) \end{aligned}$$

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

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

Сходимость градиентного спуска



Градиентный спуск: $\min_{x \in \mathbb{R}^n} f(x) \quad x_{k+1} = x_k - \alpha_k \nabla f(x_k)$

выпуклая (негладкая)	гладкая (невыпуклая)	гладкая & выпуклая	гладкая & сильно выпуклая
$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\ = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\ x_k - x^*\ ^2 = \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon = \mathcal{O}\left(\kappa \log \frac{1}{\varepsilon}\right)$

Для гладкой сильно 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 её касательной в точке $x = 0$, мы имеем:

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

Наконец:

$$\begin{aligned} \varepsilon &= f(x_{k_\varepsilon}) - f^* \leq \left(1 - \frac{\mu}{L}\right)^{k_\varepsilon} (f(x_0) - f^*) \\ &\leq \exp\left(-k_\varepsilon \frac{\mu}{L}\right) (f(x_0) - f^*) \\ k_\varepsilon &\geq \kappa \log \frac{f(x_0) - f^*}{\varepsilon} = \mathcal{O}\left(\kappa \log \frac{1}{\varepsilon}\right) \end{aligned}$$

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



- Как правило, это гораздо более нетривиальные результаты - они показывают, что никакой метод не может сходиться быстрее, чем нижняя оценка на выбранном классе функций.

¹Carmon, Duchi, Hinder, Sidford, 2017

²Nemirovski, Yudin, 1979

Нижние оценки



- Как правило, это гораздо более нетривиальные результаты - они показывают, что никакой метод не может сходиться быстрее, чем нижняя оценка на выбранном классе функций.
- Часто, эти результаты получаются путём предъявления конкретной функции из класса, для которой никакой метод не может сходиться быстрее, чем нижняя оценка.

¹Carmon, Duchi, Hinder, Sidford, 2017

²Nemirovski, Yudin, 1979

Нижние оценки

- Как правило, это гораздо более нетривиальные результаты - они показывают, что никакой метод не может сходиться быстрее, чем нижняя оценка на выбранном классе функций.
- Часто, эти результаты получаются путём предъявления конкретной функции из класса, для которой никакой метод не может сходиться быстрее, чем нижняя оценка.
- Для нижних оценок пишут $\Omega(\cdot)$ вместо $\mathcal{O}(\cdot)$.

¹Carmon, Duchi, Hinder, Sidford, 2017

²Nemirovski, Yudin, 1979

Нижние оценки

- Как правило, это гораздо более нетривиальные результаты - они показывают, что никакой метод не может сходиться быстрее, чем нижняя оценка на выбранном классе функций.
- Часто, эти результаты получаются путём предъявления конкретной функции из класса, для которой никакой метод не может сходиться быстрее, чем нижняя оценка.
- Для нижних оценок пишут $\Omega(\cdot)$ вместо $\mathcal{O}(\cdot)$.

¹Carmon, Duchi, Hinder, Sidford, 2017

²Nemirovski, Yudin, 1979

Нижние оценки



- Как правило, это гораздо более нетривиальные результаты - они показывают, что никакой метод не может сходиться быстрее, чем нижняя оценка на выбранном классе функций.
- Часто, эти результаты получаются путём предъявления конкретной функции из класса, для которой никакой метод не может сходиться быстрее, чем нижняя оценка.
- Для нижних оценок пишут $\Omega(\cdot)$ вместо $\mathcal{O}(\cdot)$.

выпуклая (негладкая)	гладкая (невыпуклая) ¹	гладкая & выпуклая ²	гладкая & сильно выпуклая
$f(x_k) - f^* = \Omega\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \Omega\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\ = \Omega\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \Omega\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \Omega\left(\frac{1}{k^2}\right)$ $k_\varepsilon = \Omega\left(\frac{1}{\sqrt{\varepsilon}}\right)$	$f(x_k) - f^* = \Omega\left(\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k}\right)$ $k_\varepsilon = \Omega\left(\sqrt{\kappa} \log \frac{1}{\varepsilon}\right)$

¹Carmon, Duchi, Hinder, Sidford, 2017

²Nemirovski, Yudin, 1979

Нижние оценки



- Как правило, это гораздо более нетривиальные результаты - они показывают, что никакой метод не может сходиться быстрее, чем нижняя оценка на выбранном классе функций.
- Часто, эти результаты получаются путём предъявления конкретной функции из класса, для которой никакой метод не может сходиться быстрее, чем нижняя оценка.
- Для нижних оценок пишут $\Omega(\cdot)$ вместо $\mathcal{O}(\cdot)$.

выпуклая (негладкая)	гладкая (невыпуклая) ¹	гладкая & выпуклая ²	гладкая & сильно выпуклая
$f(x_k) - f^* = \Omega\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \Omega\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \leq i \leq k} \ \nabla f(x_i)\ = \Omega\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \Omega\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \Omega\left(\frac{1}{k^2}\right)$ $k_\varepsilon = \Omega\left(\frac{1}{\sqrt{\varepsilon}}\right)$	$f(x_k) - f^* = \Omega\left(\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k}\right)$ $k_\varepsilon = \Omega\left(\sqrt{\kappa} \log \frac{1}{\varepsilon}\right)$

Например, из таблицы выше следует, что никакой метод первого порядка определённой формы не может сходиться быстрее, чем $\Omega\left(\frac{1}{k^2}\right)$ ($\Omega\left(\frac{1}{\sqrt{\varepsilon}}\right)$ для гладкой выпуклой функции) для гладкой выпуклой функции.

¹Carmon, Duchi, Hinder, Sidford, 2017

²Nemirovski, Yudin, 1979

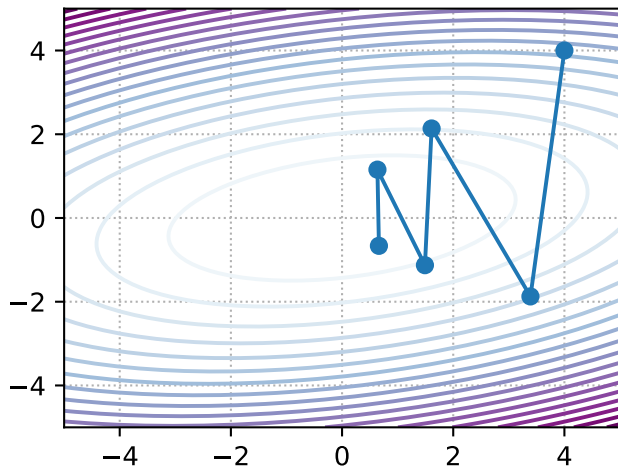
Метод тяжёлого шарика

Колебания и ускорение

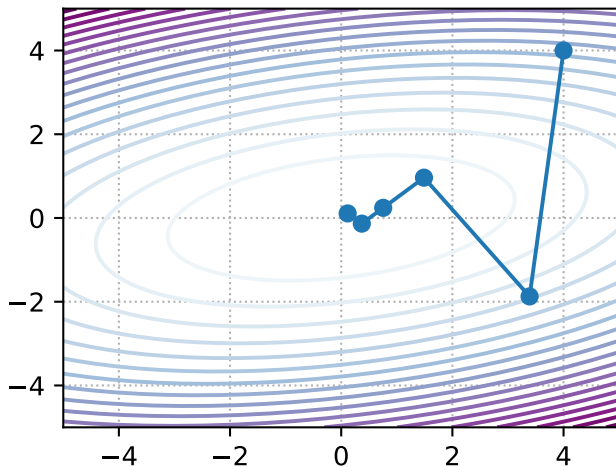


$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}).$$

Gradient Descent



Heavy Ball

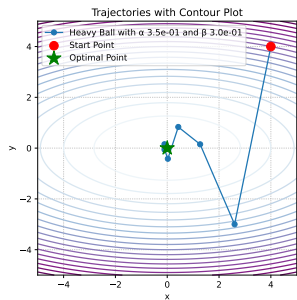
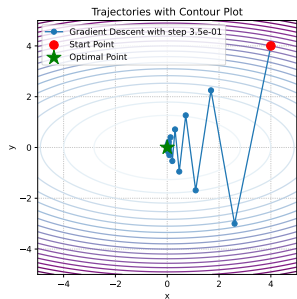


Метод тяжёлого шарика Поляка



Давайте представим идею импульса (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

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



Метод тяжёлого шарика Поляка



Давайте представим идею импульса (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

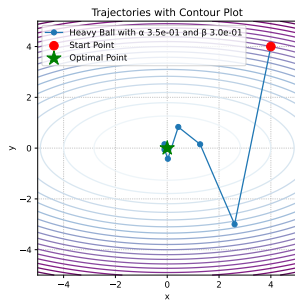
$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}).$$

Давайте теперь подставим в итерацию предыдущую итерацию, т.е.

$$x_k = x_{k-1} - \alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2}), \text{ а так же отметим, что}$$

$$x_k - x_{k-1} = -\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2}):$$

Метод тяжёлого шарика Поляка



Давайте представим идею импульса (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

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

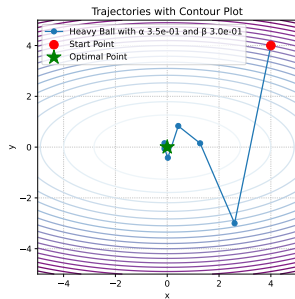
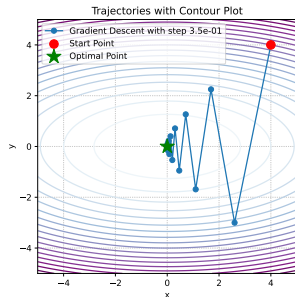
Давайте теперь подставим в итерацию предыдущую итерацию, т.е.

$$x_k = x_{k-1} - \alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2}), \text{ а так же отметим, что}$$

$$x_k - x_{k-1} = -\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2}):$$

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

Метод тяжёлого шарика Поляка



Давайте представим идею импульса (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

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

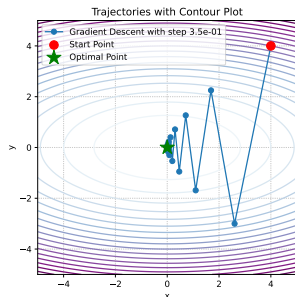
Давайте теперь подставим в итерацию предыдущую итерацию, т.е.

$$x_k = x_{k-1} - \alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2}), \text{ а так же отметим, что}$$

$$x_k - x_{k-1} = -\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2}):$$

$$\begin{aligned} x_{k+1} &= x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}) \\ &= x_k - \alpha \nabla f(x_k) + \beta(-\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})) \end{aligned}$$

Метод тяжёлого шарика Поляка



Давайте представим идею импульса (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

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

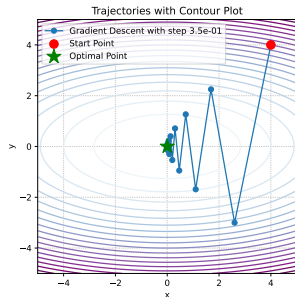
Давайте теперь подставим в итерацию предыдущую итерацию, т.е.

$x_k = x_{k-1} - \alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})$, а так же отметим, что

$x_k - x_{k-1} = -\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})$:

$$\begin{aligned} x_{k+1} &= x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}) \\ &= x_k - \alpha \nabla f(x_k) + \beta(-\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})) \\ &= x_k - \alpha [\nabla f(x_k) + \beta \nabla f(x_{k-1})] + \beta^2(x_{k-1} - x_{k-2}) \end{aligned}$$

Метод тяжёлого шарика Поляка



Давайте представим идею импульса (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

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

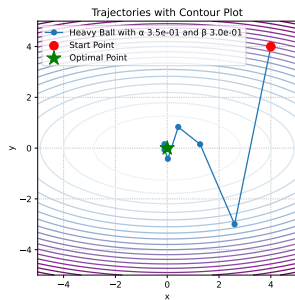
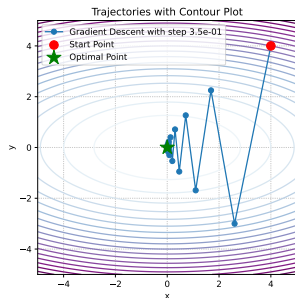
Давайте теперь подставим в итерацию предыдущую итерацию, т.е.

$$x_k = x_{k-1} - \alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2}), \text{ а так же отметим, что}$$

$$x_k - x_{k-1} = -\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2}):$$

$$\begin{aligned} x_{k+1} &= x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}) \\ &= x_k - \alpha \nabla f(x_k) + \beta(-\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})) \\ &= x_k - \alpha [\nabla f(x_k) + \beta \nabla f(x_{k-1})] + \beta^2(x_{k-1} - x_{k-2}) \\ &= x_k - \alpha [\nabla f(x_k) + \beta \nabla f(x_{k-1}) + \beta^2 \nabla f(x_{k-2})] + \beta^3(x_{k-2} - x_{k-3}) \end{aligned}$$

Метод тяжёлого шарика Поляка



Давайте представим идею импульса (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

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

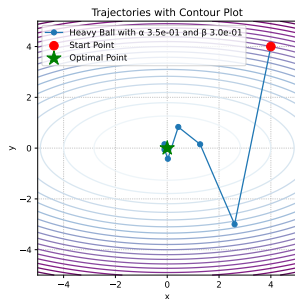
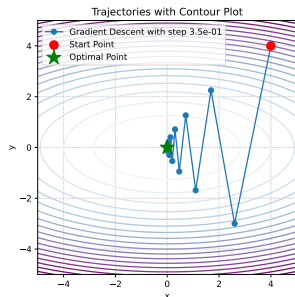
Давайте теперь подставим в итерацию предыдущую итерацию, т.е.

$$x_k = x_{k-1} - \alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2}), \text{ а так же отметим, что}$$

$$x_k - x_{k-1} = -\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2}):$$

$$\begin{aligned} x_{k+1} &= x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}) \\ &= x_k - \alpha \nabla f(x_k) + \beta(-\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})) \\ &= x_k - \alpha [\nabla f(x_k) + \beta \nabla f(x_{k-1})] + \beta^2(x_{k-1} - x_{k-2}) \\ &= x_k - \alpha [\nabla f(x_k) + \beta \nabla f(x_{k-1}) + \beta^2 \nabla f(x_{k-2})] + \beta^3(x_{k-2} - x_{k-3}) \\ &\vdots \end{aligned}$$

Метод тяжёлого шарика Поляка



Давайте представим идею импульса (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

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

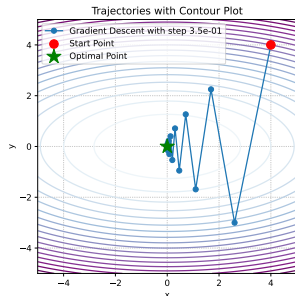
Давайте теперь подставим в итерацию предыдущую итерацию, т.е.

$$x_k = x_{k-1} - \alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2}), \text{ а так же отметим, что}$$

$$x_k - x_{k-1} = -\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2}):$$

$$\begin{aligned} x_{k+1} &= x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}) \\ &= x_k - \alpha \nabla f(x_k) + \beta(-\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})) \\ &= x_k - \alpha [\nabla f(x_k) + \beta \nabla f(x_{k-1})] + \beta^2(x_{k-1} - x_{k-2}) \\ &= x_k - \alpha [\nabla f(x_k) + \beta \nabla f(x_{k-1}) + \beta^2 \nabla f(x_{k-2})] + \beta^3(x_{k-2} - x_{k-3}) \\ &\vdots \\ &= x_k - \alpha [\nabla f(x_k) + \beta \nabla f(x_{k-1}) + \beta^2 \nabla f(x_{k-2}) + \dots + \beta^k \nabla f(x_0)] \end{aligned}$$

Метод тяжёлого шарика Поляка



Давайте представим идею импульса (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

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

Давайте теперь подставим в итерацию предыдущую итерацию, т.е.

$$x_k = x_{k-1} - \alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2}), \text{ а так же отметим, что}$$

$$x_k - x_{k-1} = -\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2}):$$

$$\begin{aligned} x_{k+1} &= x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}) \\ &= x_k - \alpha \nabla f(x_k) + \beta(-\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})) \\ &= x_k - \alpha [\nabla f(x_k) + \beta \nabla f(x_{k-1})] + \beta^2(x_{k-1} - x_{k-2}) \\ &= x_k - \alpha [\nabla f(x_k) + \beta \nabla f(x_{k-1}) + \beta^2 \nabla f(x_{k-2})] + \beta^3(x_{k-2} - x_{k-3}) \\ &\vdots \\ &= x_k - \alpha [\nabla f(x_k) + \beta \nabla f(x_{k-1}) + \beta^2 \nabla f(x_{k-2}) + \dots + \beta^k \nabla f(x_0)] \end{aligned}$$

Таким образом, метод тяжёлого шарика учитывает все предыдущие градиенты с тем меньшим весом, чем старше итерация ($0 \leq \beta < 1$).

Сходимость метода тяжёлого шарика для квадратичной функции

i Theorem

Предположим, что f является μ -сильно выпуклой и L -гладкой квадратичной функцией. Тогда метод тяжёлого 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 тяжёлого 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 тяжёлого шарика, удовлетворяет

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

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 тяжёлого шарика для выпуклой оптимизации, Euhanna Ghadimi et al.

Глобальная сходимость метода тяжёлого шарика⁴



i Theorem

Предположим, что f является гладкой и сильно выпуклой и что

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

Тогда последовательность $\{x_k\}$, генерируемая итерациями метода тяжёлого 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 тяжёлого шарика для выпуклой оптимизации, Euhanna Ghadimi et al.

Итоги по методу тяжёлого шарика



- Обеспечивает ускоренную сходимость для сильно выпуклых квадратичных задач.

Итоги по методу тяжёлого шарика



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

Итоги по методу тяжёлого шарика

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

⁵Provable non-accelerations of the heavy-ball method

Итоги по методу тяжёлого шарика

- Обеспечивает ускоренную сходимость для сильно выпуклых квадратичных задач.
- Локально ускоренная сходимость была доказана в оригинальной статье.
- Недавно⁵ было доказано, что глобального ускорения сходимости для метода не существует.
- Метод не был чрезвычайно популярен до ML-бума.

⁵Provable non-accelerations of the heavy-ball method

Итоги по методу тяжёлого шарика

- Обеспечивает ускоренную сходимость для сильно выпуклых квадратичных задач.
- Локально ускоренная сходимость была доказана в оригинальной статье.
- Недавно⁵ было доказано, что глобального ускорения сходимости для метода не существует.
- Метод не был чрезвычайно популярен до ML-бума.
- Сейчас он фактически является стандартом для практического ускорения методов градиентного спуска, в том числе для невыпуклых задач (обучение нейронных сетей).

⁵Provable non-accelerations of the heavy-ball method

Ускоренный градиентный метод Нестерова

Концепция ускоренного градиентного метода Нестерова



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

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

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

Концепция ускоренного градиентного метода Нестерова



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

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

Давайте определим следующие обозначения

$$x^+ = x - \alpha \nabla f(x)$$

Градиентный шаг

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

Импульс

Тогда мы можем записать:

$$x_{k+1} = x_k^+$$

Градиентный спуск

$$x_{k+1} = x_k^+ + d_k$$

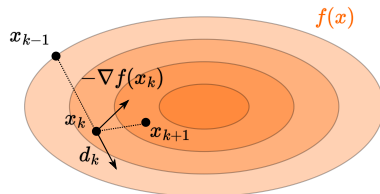
Метод тяжёлого шарика

$$x_{k+1} = (x_k + d_k)^+$$

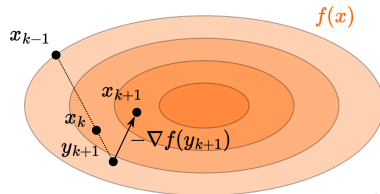
Ускоренный градиентный метод Нестерова

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

Polyak momentum



Nesterov momentum



Сходимость для выпуклых функций



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:
$$x_{k+1} = y_k - \frac{1}{L} \nabla f(y_k)$$

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

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

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

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

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

Ускоренная сходимость для сильно выпуклых функций

i Theorem

Предположим, что $f : \mathbb{R}^n \rightarrow \mathbb{R}$ является μ -сильно 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:
$$x_{k+1} = y_k - \frac{1}{L} \nabla f(y_k)$$

Экстраполяция:
$$y_{k+1} = x_{k+1} - \gamma (x_{k+1} - x_k)$$

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

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

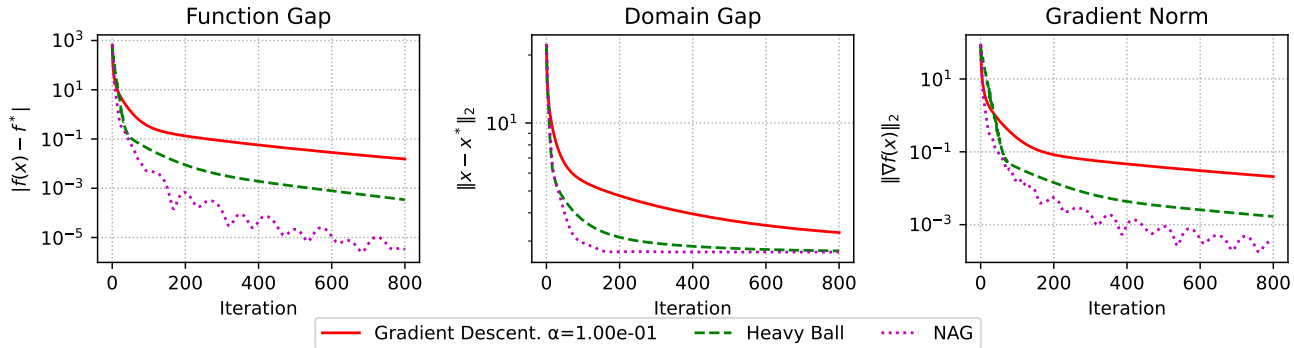
$$f(x_k) - f^* \leq \frac{\mu + L}{2} \|x_0 - x^*\|_2^2 \exp\left(-\frac{k}{\sqrt{\kappa}}\right)$$

Численные эксперименты

Выпуклая квадратичная задача (линейная регрессия)



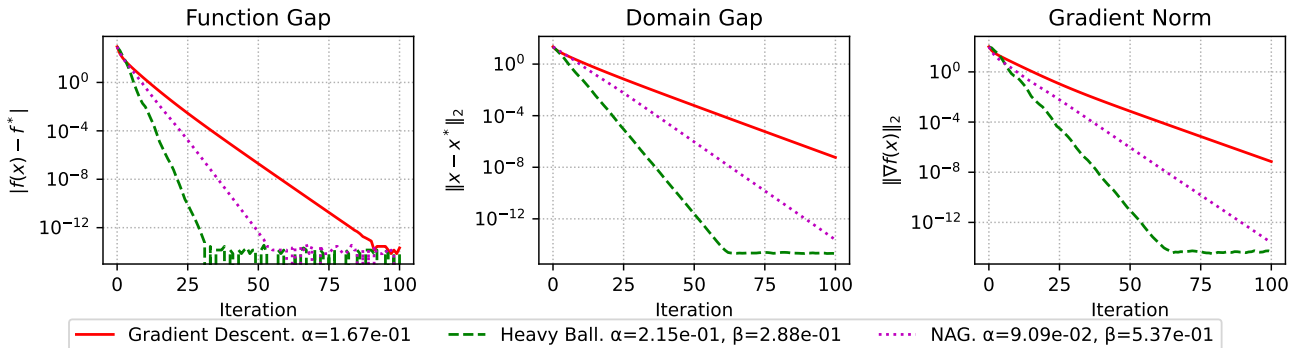
Convex quadratics: $n=60$, random matrix, $\mu=0$, $L=10$



Сильно выпуклая квадратичная задача (регуляризованная линейная регрессия)



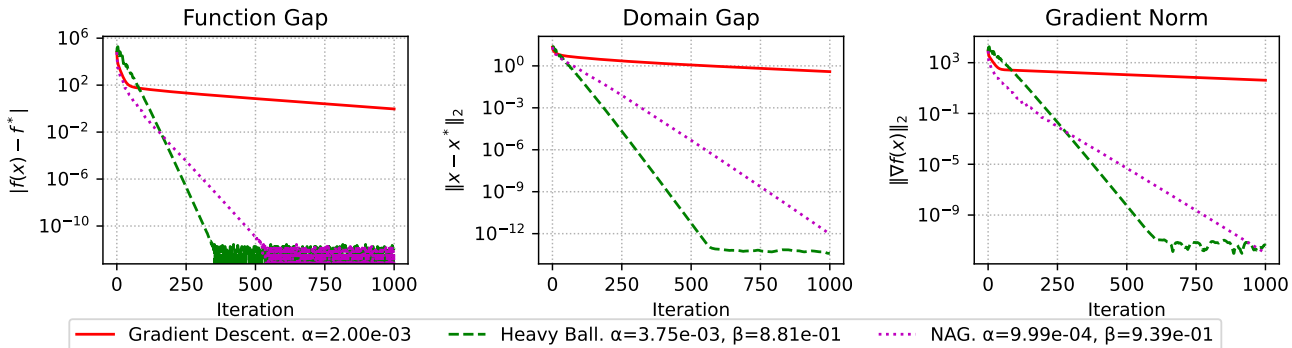
Strongly convex quadratics: $n=60$, random matrix, $\mu=1$, $L=10$



Сильно выпуклая квадратичная задача (регуляризованная линейная регрессия)



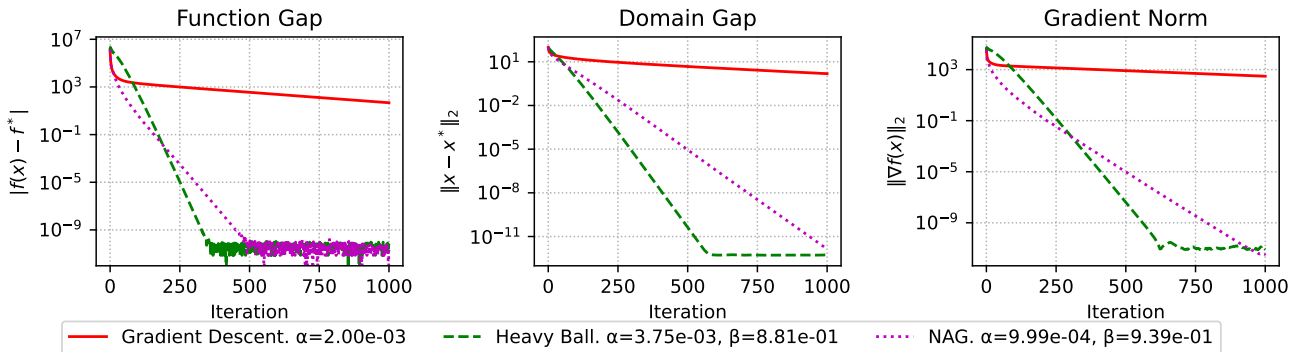
Strongly convex quadratics: $n=60$, random matrix, $\mu=1$, $L=1000$



Сильно выпуклая квадратичная задача (регуляризованная линейная регрессия)



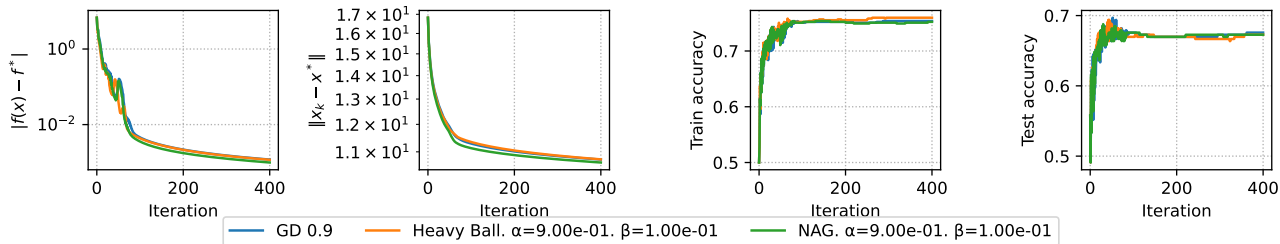
Strongly convex quadratics: $n=1000$, random matrix, $\mu=1$, $L=1000$



Выпуклая бинарная логистическая регрессия



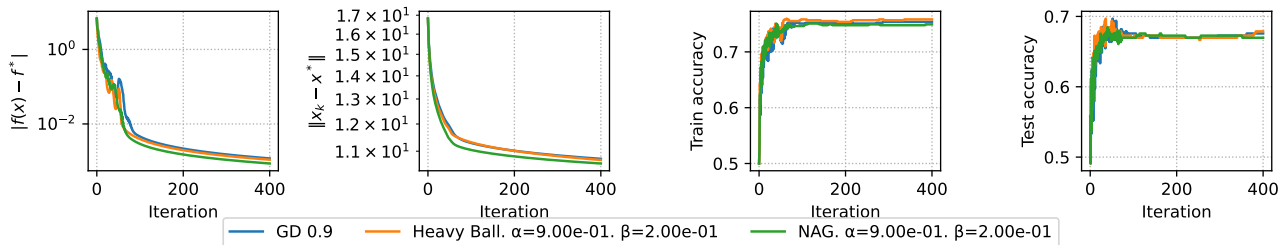
Convex binary logistic regression. $\mu=0$.



Выпуклая бинарная логистическая регрессия



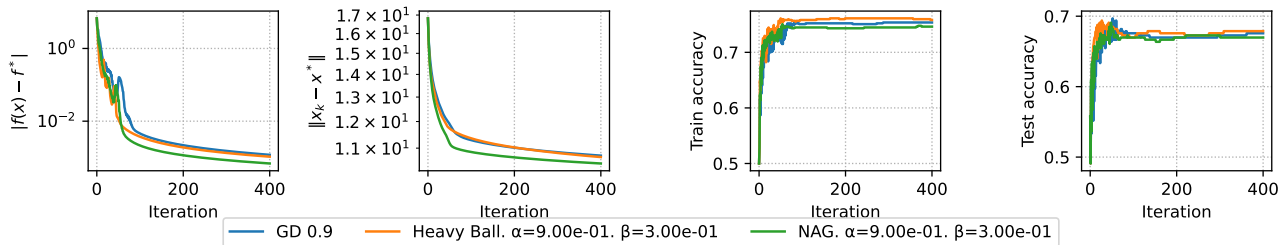
Convex binary logistic regression. $\mu=0$.



Выпуклая бинарная логистическая регрессия



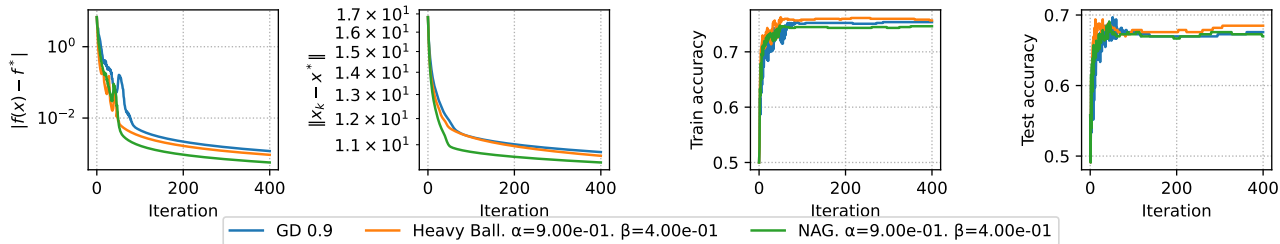
Convex binary logistic regression. $\mu=0$.



Выпуклая бинарная логистическая регрессия



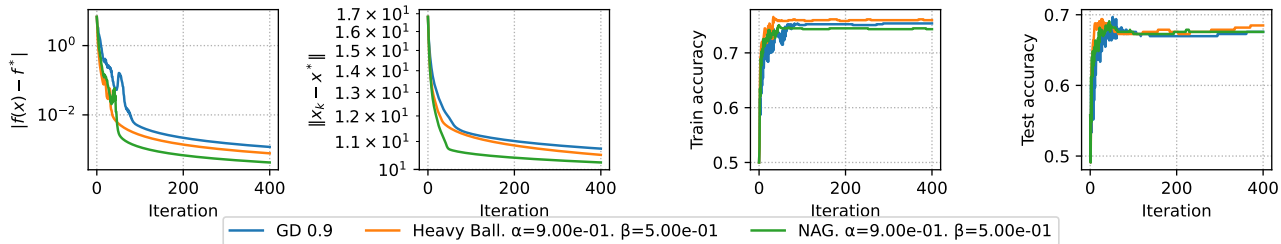
Convex binary logistic regression. $\mu=0$.



Выпуклая бинарная логистическая регрессия



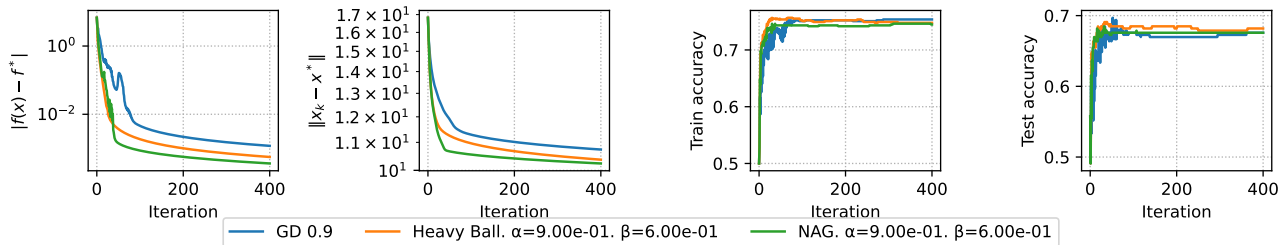
Convex binary logistic regression. $\mu=0$.



Выпуклая бинарная логистическая регрессия



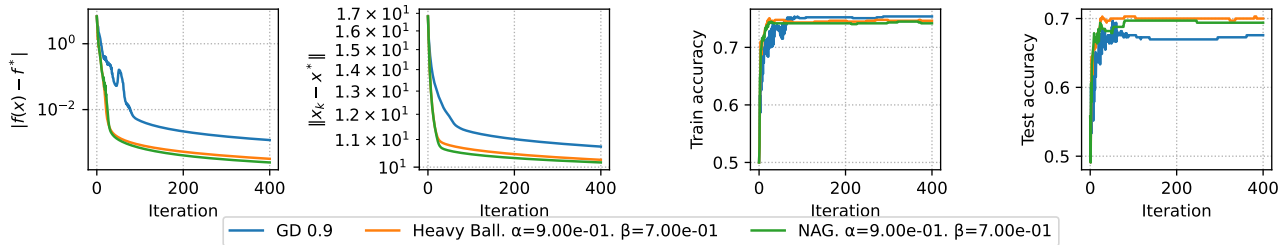
Convex binary logistic regression. $\mu=0$.



Выпуклая бинарная логистическая регрессия



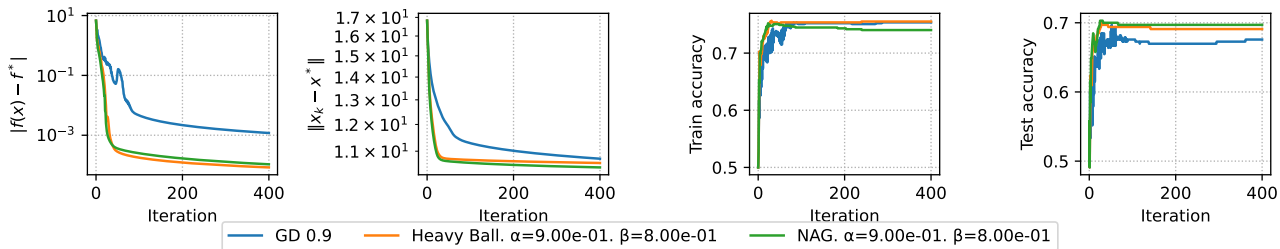
Convex binary logistic regression. $\mu=0$.



Выпуклая бинарная логистическая регрессия



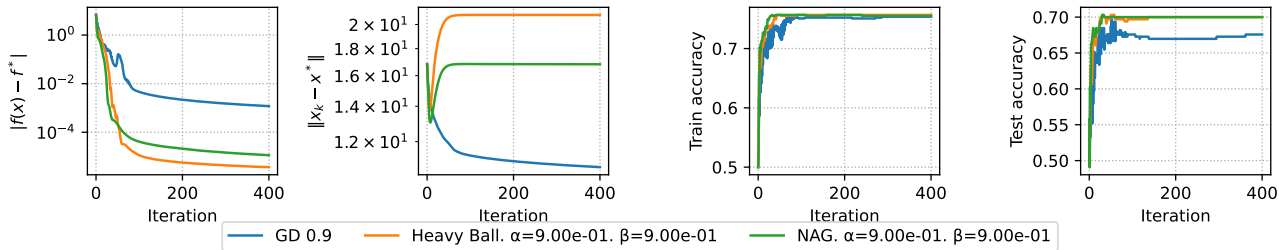
Convex binary logistic regression. $\mu=0$.



Выпуклая бинарная логистическая регрессия



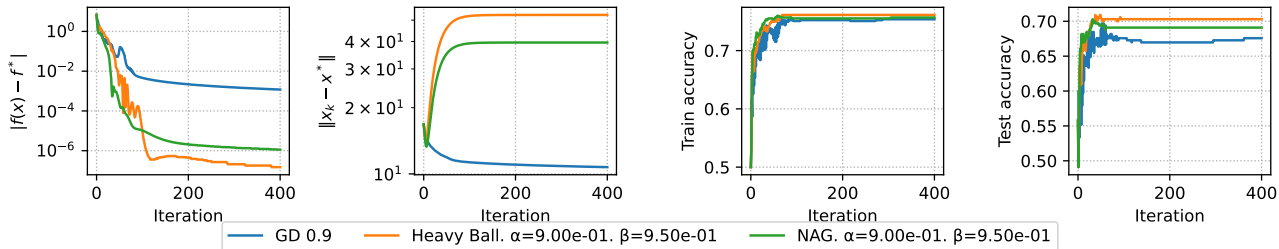
Convex binary logistic regression. $\mu=0$.



Выпуклая бинарная логистическая регрессия



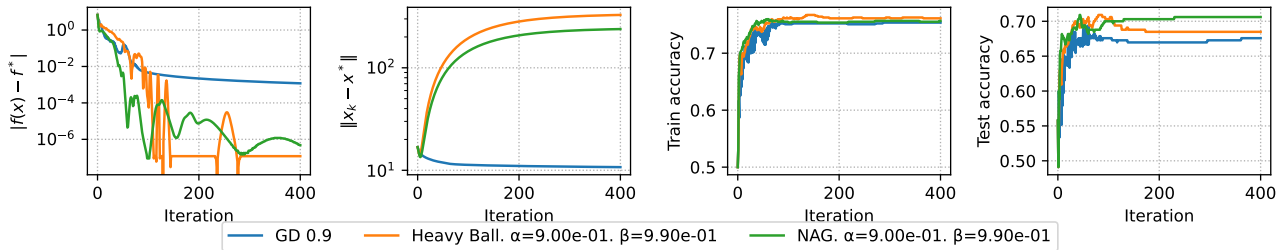
Convex binary logistic regression. $\mu=0$.



Выпуклая бинарная логистическая регрессия



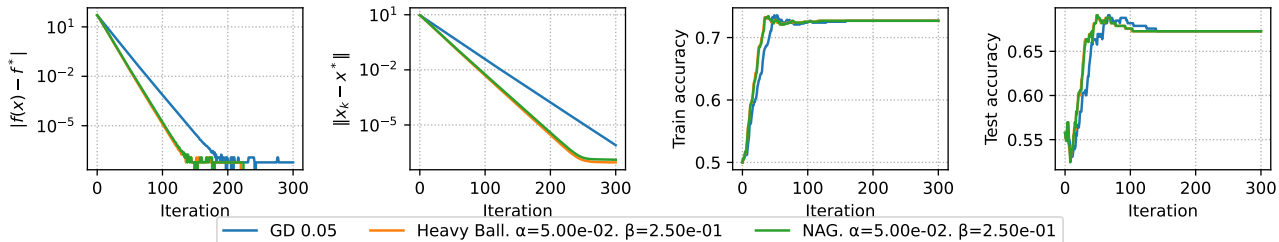
Convex binary logistic regression. $\mu=0$.



Сильно выпуклая бинарная логистическая регрессия



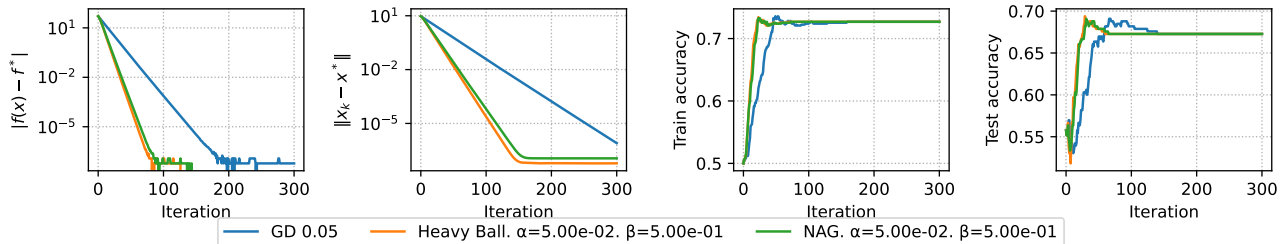
Strongly convex binary logistic regression. $\mu=1$.



Сильно выпуклая бинарная логистическая регрессия



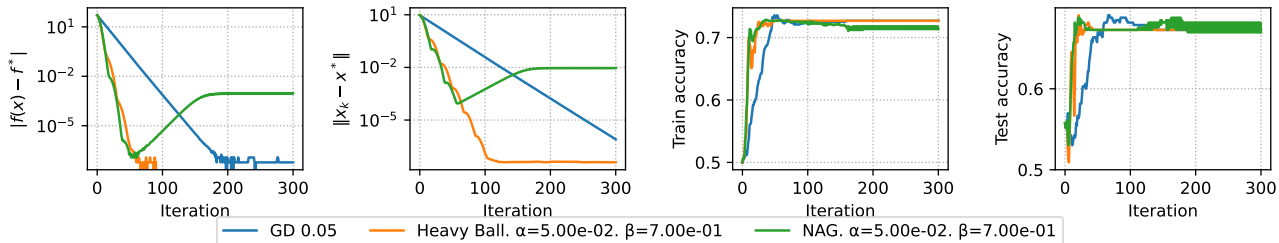
Strongly convex binary logistic regression. $\mu=1$.



Сильно выпуклая бинарная логистическая регрессия



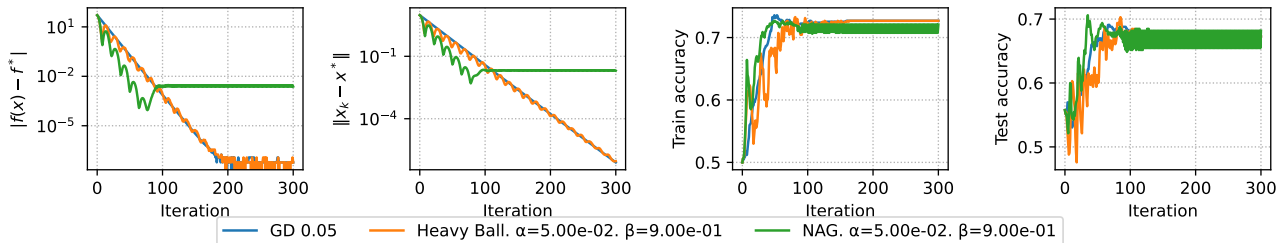
Strongly convex binary logistic regression. $\mu=1$.



Сильно выпуклая бинарная логистическая регрессия



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



Нижние оценки для методов 1 порядка (Источник)



Тип задачи	Критерий	Нижняя оценка	Верхняя оценка	Ссылка (Ниж.)	Ссылка (Верх.)
L -гладкая выпуклая	Зазор оптимальности	$\Omega\left(\sqrt{L} \varepsilon^{-1}\right)$	✓ (точное совпадение)	[1], Теорема 2.1.7	[1], Теорема 2.2.2
L -гладкая μ -сильно выпуклая	Зазор оптимальности	$\Omega\left(\sqrt{\kappa} \log \frac{1}{\varepsilon}\right)$	✓	[1], Теорема 2.1.13	[1], Теорема 2.2.2
Негладкая G -липшицева выпуклая	Зазор оптимальности	$\Omega\left(G^2 \varepsilon^{-2}\right)$	✓ (точное совпадение)	[1], Теорема 3.2.1	[1], Теорема 3.2.2
Негладкая G -липшицева μ -сильно выпуклая	Зазор оптимальности	$\Omega\left(G^2 (\mu \varepsilon)^{-1}\right)$	✓	[1], Теорема 3.2.5	[3], Теорема 3.9
L -гладкая выпуклая (сходимость по функции)	Стационарность	$\Omega\left(\sqrt{\Delta L} \varepsilon^{-1}\right)$	✓ (с точностью до логарифмического множителя)	[2], Теорема 1	[2], Приложение A.1
L -гладкая выпуклая (сходимость по аргументу)	Стационарность	$\Omega\left(\sqrt{\Delta L} \varepsilon^{-1/2}\right)$	✓	[2], Теорема 1	[6], Раздел 6.5
L -гладкая невыпуклая	Стационарность	$\Omega\left(\Delta L \varepsilon^{-2}\right)$	✓	[5], Теорема 1	[7], Теорема 10.15
Негладкая G -липшицева ρ -слабо выпуклая (WC)	Квази-стационарность	Неизвестно	$\mathcal{O}\left(\varepsilon^{-4}\right)$	/	[8], Следствие 2.2
L -гладкая μ -PL	Зазор оптимальности	$\Omega\left(\kappa \log \frac{1}{\varepsilon}\right)$	✓	[9], Теорема 3	[10], Теорема 1

Источники:

- [1] - Lectures on Convex Optimization, Y. Nesterov.
- [2] - Lower bounds for finding stationary points II: first-order methods, Y. Carmon, J.C. Duchi, O. Hinder, A. Sidford.
- [3] - Convex optimization: Algorithms and complexity, S. Bubeck, others.
- [4] - Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions D. Kim, J.A. Fessler.
- [5] - Lower bounds for finding stationary points I, Y. Carmon, J.C. Duchi, O. Hinder, A. Sidford.
- [6] - Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions, D. Kim, J.A. Fessler.
- [7] - First-order methods in optimization, A. Beck. SIAM. 2017.
- [8] - Stochastic subgradient method converges at the rate $\mathcal{O}(\kappa^{-1/4})$ on weakly convex functions, D. Davis, D. Drusvyatskiy.
- [9] - On the lower bound of minimizing Polyak-Lojasiewicz functions, P. Yue, C. Fang, Z. Lin.
- [10] - Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition, H. Karimi, J. Nutini, M. Schmidt.

Обозначения:

- Зазор оптимальности: $f(x_k) - f^* \leq \varepsilon$
- Стационарность: $\|\nabla f(x_k)\| \leq \varepsilon$
- Квази-стационарность: $\|\nabla f_\lambda(x_k)\| \leq \varepsilon$, где $f_\lambda(x) = \inf_{y \in \mathbb{R}^n} \left(f(y) + \frac{1}{2\lambda} \|y - x\|^2\right)$
- Липшицевость функции: $|f(x) - f(y)| \leq G \|x - y\| \forall x, y \in \mathbb{R}^n$
- Липшицевость градиента (L -гладкость): $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| \forall x, y \in \mathbb{R}^n$
- μ -сильная выпуклость: $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \frac{\mu}{2} \lambda(1 - \lambda) \|x - y\|^2$
- ρ -слабо выпуклая функция:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) + \rho \lambda(1 - \lambda) \|x - y\|^2 \forall x, y \in \mathbb{R}^n$$
- Число обусловленности: $\kappa = \frac{L}{\mu}$
- Зазор в начальной точке: $f(x_0) - f^* \leq \Delta$
- Зазор по аргументу: $D = \|x_0 - x^*\|$

Бонус: доказательства сходимости

Любая μ -сильно выпуклая дифференцируемая функция является PL-функцией

Theorem

Если функция $f(x)$ дифференцируема и μ -сильно выпукла, то она является PL-функцией.

Доказательство

По критерию сильной выпуклости первого порядка:

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

Положим $y = x^*$:

$$f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2$$

Любая μ -сильно выпуклая дифференцируемая функция является PL-функцией

i Theorem

Если функция $f(x)$ дифференцируема и μ -сильно выпукла, то она является PL-функцией.

Доказательство

По критерию сильной выпуклости первого порядка:

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

Положим $y = x^*$:

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^T(x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2 \\ f(x) - f(x^*) &\leq \nabla f(x)^T(x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 = \end{aligned}$$

Любая μ -сильно выпуклая дифференцируемая функция является PL-функцией

i Theorem

Если функция $f(x)$ дифференцируема и μ -сильно выпукла, то она является PL-функцией.

Доказательство

По критерию сильной выпуклости первого порядка:

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

Положим $y = x^*$:

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^T(x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2 \\ f(x) - f(x^*) &\leq \nabla f(x)^T(x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 = \\ &= (\nabla f(x) - \frac{\mu}{2}(x^* - x))^T(x - x^*) = \end{aligned}$$

Любая μ -сильно выпуклая дифференцируемая функция является RL-функцией

i Theorem

Если функция $f(x)$ дифференцируема и μ -сильно выпукла, то она является RL-функцией.

Доказательство

По критерию сильной выпуклости первого порядка:

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

Положим $y = x^*$:

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^T(x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2 \\ f(x) - f(x^*) &\leq \nabla f(x)^T(x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 = \\ &= (\nabla f(x) - \frac{\mu}{2}(x^* - x))^T(x - x^*) = \\ &= \frac{1}{2} \left(\frac{2}{\sqrt{\mu}} \nabla f(x) - \sqrt{\mu}(x^* - x) \right)^T \sqrt{\mu}(x - x^*) \end{aligned}$$

Любая μ -сильно выпуклая дифференцируемая функция является RL-функцией

i Theorem

Если функция $f(x)$ дифференцируема и μ -сильно выпукла, то она является RL-функцией.

Доказательство

По критерию сильной выпуклости первого порядка:

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

Положим $y = x^*$:

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^T(x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2 \\ f(x) - f(x^*) &\leq \nabla f(x)^T(x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 = \\ &= (\nabla f(x) - \frac{\mu}{2}(x^* - x))^T(x - x^*) = \\ &= \frac{1}{2} \left(\frac{2}{\sqrt{\mu}} \nabla f(x) - \sqrt{\mu}(x^* - x) \right)^T \sqrt{\mu}(x - x^*) \end{aligned}$$

Любая μ -сильно выпуклая дифференцируемая функция является PL-функцией

i Theorem

Если функция $f(x)$ дифференцируема и μ -сильно выпукла, то она является PL-функцией.

Доказательство

По критерию сильной выпуклости первого порядка:

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

Положим $y = x^*$:

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^T(x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2 \\ f(x) - f(x^*) &\leq \nabla f(x)^T(x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 = \\ &= (\nabla f(x) - \frac{\mu}{2}(x^* - x))^T(x - x^*) = \\ &= \frac{1}{2} \left(\frac{2}{\sqrt{\mu}} \nabla f(x) - \sqrt{\mu}(x^* - x) \right)^T \sqrt{\mu}(x - x^*) \end{aligned}$$

$$\begin{aligned} \text{Пусть } a &= \frac{1}{\sqrt{\mu}} \nabla f(x) \text{ и} \\ b &= \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x) \end{aligned}$$

Любая μ -сильно выпуклая дифференцируемая функция является PL-функцией

i Theorem

Если функция $f(x)$ дифференцируема и μ -сильно выпукла, то она является PL-функцией.

Доказательство

По критерию сильной выпуклости первого порядка:

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

Положим $y = x^*$:

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^T(x^* - x) + \frac{\mu}{2}\|x^* - x\|_2^2 \\ f(x) - f(x^*) &\leq \nabla f(x)^T(x - x^*) - \frac{\mu}{2}\|x^* - x\|_2^2 = \\ &= (\nabla f(x) - \frac{\mu}{2}(x^* - x))^T(x - x^*) = \\ &= \frac{1}{2} \left(\frac{2}{\sqrt{\mu}} \nabla f(x) - \sqrt{\mu}(x^* - x) \right)^T \sqrt{\mu}(x - x^*) \end{aligned}$$

Пусть $a = \frac{1}{\sqrt{\mu}} \nabla f(x)$ и

$$b = \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x)$$

Тогда $a + b = \sqrt{\mu}(x - x^*)$ и

$$a - b = \frac{2}{\sqrt{\mu}} \nabla f(x) - \sqrt{\mu}(x - x^*)$$

Любая μ -сильно выпуклая дифференцируемая функция является PL-функцией

$$f(x) - f(x^*) \leq \frac{1}{2} \left(\frac{1}{\mu} \|\nabla f(x)\|_2^2 - \left\| \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x) \right\|_2^2 \right)$$

Любая μ -сильно выпуклая дифференцируемая функция является PL-функцией

$$f(x) - f(x^*) \leq \frac{1}{2} \left(\frac{1}{\mu} \|\nabla f(x)\|_2^2 - \left\| \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x) \right\|_2^2 \right)$$

$$f(x) - f(x^*) \leq \frac{1}{2\mu} \|\nabla f(x)\|_2^2,$$

Любая μ -сильно выпуклая дифференцируемая функция является PL-функцией

$$f(x) - f(x^*) \leq \frac{1}{2} \left(\frac{1}{\mu} \|\nabla f(x)\|_2^2 - \left\| \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x) \right\|_2^2 \right)$$

$$f(x) - f(x^*) \leq \frac{1}{2\mu} \|\nabla f(x)\|_2^2,$$

Любая μ -сильно выпуклая дифференцируемая функция является PL-функцией

$$f(x) - f(x^*) \leq \frac{1}{2} \left(\frac{1}{\mu} \|\nabla f(x)\|_2^2 - \left\| \sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x) \right\|_2^2 \right)$$

$$f(x) - f(x^*) \leq \frac{1}{2\mu} \|\nabla f(x)\|_2^2,$$

которое является точным условием PL. Это означает, что мы уже имеем доказательство линейной сходимости для любой сильно выпуклой функции.

Сходимость градиентного спуска в выпуклом гладком случае [1/4]



i Theorem

Рассмотрим задачу

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

Пусть $f^* = \min_{x \in \mathbb{R}^d} f(x)$. Предположим, что $f : \mathbb{R}^d \rightarrow \mathbb{R}$ является выпуклой и L -гладкой функцией, для некоторого $L > 0$. Пусть $(x_k)_{k \in \mathbb{N}}$ — последовательность итераций, сгенерированная методом градиентного спуска из точки x_0 с постоянным шагом α , удовлетворяющим $0 < \alpha \leq \frac{1}{L}$.

Тогда для всех $k \in \mathbb{N}$ справедливо:

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

Заметим, что мы здесь никак не упоминаем точку минимума. То есть, это сходимость $\forall x \in \mathbb{R}^d$ (в том числе и до точки минимума).

Сходимость градиентного спуска в выпуклом гладком случае [2/4]



Наш инструментарий:

1. Выпуклость:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall x, y. \quad (1)$$

Сходимость градиентного спуска в выпуклом гладком случае [2/4]



Наш инструментарий:

1. Выпуклость:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall x, y. \quad (1)$$

2. Гладкость:

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

Сходимость градиентного спуска в выпуклом гладком случае [2/4]



Наш инструментарий:

1. Выпуклость:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall x, y. \quad (1)$$

2. Гладкость:

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

3. 3-point identity (по сути, квадрат разности):

$$\|a - b\|^2 = \|a - c - (b - c)\|^2 = \|a - c\|^2 - 2\langle a - c, b - c \rangle + \|b - c\|^2$$

переносим справа все кроме $\|b - c\|^2$ налево и меняем местами все факторы внутри каждого из перенесенных членов:

$$\|b - c\|^2 = \|b - a\|^2 + 2\langle c - a, c - b \rangle - \|c - a\|^2. \quad (3)$$

Сходимость градиентного спуска в выпуклом гладком случае [2/4]



Наш инструментарий:

1. Выпуклость:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall x, y. \quad (1)$$

2. Гладкость:

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

3. 3-point identity (по сути, квадрат разности):

$$\|a - b\|^2 = \|a - c - (b - c)\|^2 = \|a - c\|^2 - 2\langle a - c, b - c \rangle + \|b - c\|^2$$

переносим справа все кроме $\|b - c\|^2$ налево и меняем местами все факторы внутри каждого из перенесенных членов:

$$\|b - c\|^2 = \|b - a\|^2 + 2\langle c - a, c - b \rangle - \|c - a\|^2. \quad (3)$$

- Подставляем в (3) $b \equiv x$, $c \equiv x_{k+1}$, $a \equiv x_k$ и домножаем все на $\frac{1}{2}$:

(4)

Сходимость градиентного спуска в выпуклом гладком случае [2/4]



Наш инструментарий:

1. Выпуклость:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall x, y. \quad (1)$$

2. Гладкость:

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

3. 3-point identity (по сути, квадрат разности):

$$\|a - b\|^2 = \|a - c - (b - c)\|^2 = \|a - c\|^2 - 2\langle a - c, b - c \rangle + \|b - c\|^2$$

переносим справа все кроме $\|b - c\|^2$ налево и меняем местами все факторы внутри каждого из перенесенных членов:

$$\|b - c\|^2 = \|b - a\|^2 + 2\langle c - a, c - b \rangle - \|c - a\|^2. \quad (3)$$

• Подставляем в (3) $b \equiv x$, $c \equiv x_{k+1}$, $a \equiv x_k$ и домножаем все на $\frac{1}{2}$:

$$\frac{1}{2} \|x - x_{k+1}\|^2 = \frac{1}{2} \|x - x_k\|^2 + \langle x_{k+1} - x_k, x_{k+1} - x \rangle - \frac{1}{2} \|x_{k+1} - x_k\|^2 \quad (4)$$

Сходимость градиентного спуска в выпуклом гладком случае [2/4]



Наш инструментарий:

1. Выпуклость:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall x, y. \quad (1)$$

2. Гладкость:

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

3. 3-point identity (по сути, квадрат разности):

$$\|a - b\|^2 = \|a - c - (b - c)\|^2 = \|a - c\|^2 - 2\langle a - c, b - c \rangle + \|b - c\|^2$$

переносим справа все кроме $\|b - c\|^2$ налево и меняем местами все факторы внутри каждого из перенесенных членов:

$$\|b - c\|^2 = \|b - a\|^2 + 2\langle c - a, c - b \rangle - \|c - a\|^2. \quad (3)$$

• Подставляем в (3) $b \equiv x$, $c \equiv x_{k+1}$, $a \equiv x_k$ и домножаем все на $\frac{1}{2}$:

$$\begin{aligned} \frac{1}{2} \|x - x_{k+1}\|^2 &= \frac{1}{2} \|x - x_k\|^2 + \langle x_{k+1} - x_k, x_{k+1} - x \rangle - \frac{1}{2} \|x_{k+1} - x_k\|^2 \\ &= \frac{1}{2} \|x - x_k\|^2 - \alpha \langle \nabla f(x_k), x_{k+1} - x \rangle - \frac{1}{2} \|x_{k+1} - x_k\|^2. \end{aligned} \quad (4)$$

Сходимость градиентного спуска в выпуклом гладком случае [3/4]



- Посмотрим внимательнее на скалярное произведение $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$ и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

Сходимость градиентного спуска в выпуклом гладком случае [3/4]



- Посмотрим внимательнее на скалярное произведение $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$ и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

$$-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle = \alpha (\langle \nabla f(x_k), x - x_k \rangle + \langle \nabla f(x_k), x_k - x_{k+1} \rangle)$$

Сходимость градиентного спуска в выпуклом гладком случае [3/4]



- Посмотрим внимательнее на скалярное произведение $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$ и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

$$\begin{aligned} -\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle &= \alpha (\langle \nabla f(x_k), x - x_k \rangle + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \\ &\stackrel{(1)}{\leq} \alpha (f(x) - f(x_k) + \langle \nabla f(x_k), x_k - x_{k+1} \rangle) \end{aligned}$$

Сходимость градиентного спуска в выпуклом гладком случае [3/4]



- Посмотрим внимательнее на скалярное произведение $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$ и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

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

Сходимость градиентного спуска в выпуклом гладком случае [3/4]



- Посмотрим внимательнее на скалярное произведение $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$ и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

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

- Подставляем это все обратно в (4) и используем условие на размер шага $\alpha \leq \frac{1}{L}$:

Сходимость градиентного спуска в выпуклом гладком случае [3/4]



- Посмотрим внимательнее на скалярное произведение $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$ и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

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

- Подставляем это все обратно в (4) и используем условие на размер шага $\alpha \leq \frac{1}{L}$:

$$\frac{1}{2} \|x - x_{k+1}\|^2 \leq \frac{1}{2} \|x - x_k\|^2 + \alpha (f(x) - f(x_{k+1})) + \left(\frac{\alpha L}{2} - \frac{1}{2} \right) \|x_{k+1} - x_k\|^2$$

Сходимость градиентного спуска в выпуклом гладком случае [3/4]



- Посмотрим внимательнее на скалярное произведение $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$ и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

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

- Подставляем это все обратно в (4) и используем условие на размер шага $\alpha \leq \frac{1}{L}$:

$$\begin{aligned} \frac{1}{2} \|x - x_{k+1}\|^2 &\leq \frac{1}{2} \|x - x_k\|^2 + \alpha (f(x) - f(x_{k+1})) + \left(\frac{\alpha L}{2} - \frac{1}{2} \right) \|x_{k+1} - x_k\|^2 \\ \frac{1}{2} \|x - x_{k+1}\|^2 - \frac{1}{2} \|x - x_k\|^2 &\leq \alpha (f(x) - f(x_{k+1})) + \left(\frac{\alpha L}{2} - \frac{1}{2} \right) \|x_{k+1} - x_k\|^2 \end{aligned}$$

Сходимость градиентного спуска в выпуклом гладком случае [3/4]



- Посмотрим внимательнее на скалярное произведение $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$ и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

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

- Подставляем это все обратно в (4) и используем условие на размер шага $\alpha \leq \frac{1}{L}$:

$$\begin{aligned} \frac{1}{2} \|x - x_{k+1}\|^2 &\leq \frac{1}{2} \|x - x_k\|^2 + \alpha (f(x) - f(x_{k+1})) + \left(\frac{\alpha L}{2} - \frac{1}{2} \right) \|x_{k+1} - x_k\|^2 \\ \frac{1}{2} \|x - x_{k+1}\|^2 - \frac{1}{2} \|x - x_k\|^2 &\leq \alpha (f(x) - f(x_{k+1})) + \left(\frac{\alpha L}{2} - \frac{1}{2} \right) \|x_{k+1} - x_k\|^2 \\ &\stackrel{(\alpha \leq 1/L)}{\leq} \frac{1}{L} (f(x) - f(x_{k+1})). \end{aligned}$$

Сходимость градиентного спуска в выпуклом гладком случае [3/4]



- Посмотрим внимательнее на скалярное произведение $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$ и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

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

- Подставляем это все обратно в (4) и используем условие на размер шага $\alpha \leq \frac{1}{L}$:

$$\begin{aligned} \frac{1}{2} \|x - x_{k+1}\|^2 &\leq \frac{1}{2} \|x - x_k\|^2 + \alpha (f(x) - f(x_{k+1})) + \left(\frac{\alpha L}{2} - \frac{1}{2} \right) \|x_{k+1} - x_k\|^2 \\ \frac{1}{2} \|x - x_{k+1}\|^2 - \frac{1}{2} \|x - x_k\|^2 &\leq \alpha (f(x) - f(x_{k+1})) + \left(\frac{\alpha L}{2} - \frac{1}{2} \right) \|x_{k+1} - x_k\|^2 \\ &\stackrel{(\alpha \leq 1/L)}{\leq} \frac{1}{L} (f(x) - f(x_{k+1})). \end{aligned}$$

- Переносим правую часть влево, левую - вправо и домножаем на L :

Сходимость градиентного спуска в выпуклом гладком случае [3/4]



- Посмотрим внимательнее на скалярное произведение $-\alpha \langle \nabla f(x_k), x_{k+1} - x \rangle$ и воспользуемся сначала выпуклостью (1), а потом – гладкостью (2):

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

- Подставляем это все обратно в (4) и используем условие на размер шага $\alpha \leq \frac{1}{L}$:

$$\begin{aligned} \frac{1}{2} \|x - x_{k+1}\|^2 &\leq \frac{1}{2} \|x - x_k\|^2 + \alpha (f(x) - f(x_{k+1})) + \left(\frac{\alpha L}{2} - \frac{1}{2} \right) \|x_{k+1} - x_k\|^2 \\ \frac{1}{2} \|x - x_{k+1}\|^2 - \frac{1}{2} \|x - x_k\|^2 &\leq \alpha (f(x) - f(x_{k+1})) + \left(\frac{\alpha L}{2} - \frac{1}{2} \right) \|x_{k+1} - x_k\|^2 \\ &\stackrel{(\alpha \leq 1/L)}{\leq} \frac{1}{L} (f(x) - f(x_{k+1})). \end{aligned}$$

- Переносим правую часть влево, левую - вправо и домножаем на L :

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

Сходимость градиентного спуска в выпуклом гладком случае [4/4]



- Берем среднее от левой и правой частей от по всем k от 0 до $N - 1$:

(5)

Сходимость градиентного спуска в выпуклом гладком случае [4/4]



- Берем среднее от левой и правой частей от по всем k от 0 до $N - 1$:

$$\frac{1}{N} \sum_{k=0}^{N-1} (f(x_{k+1}) - f(x)) \leq \frac{L}{2N} \sum_{k=0}^{N-1} (\|x - x_k\|^2 - \|x - x_{k+1}\|)$$

(5)

Сходимость градиентного спуска в выпуклом гладком случае [4/4]



- Берем среднее от левой и правой частей от по всем k от 0 до $N - 1$:

$$\begin{aligned}\frac{1}{N} \sum_{k=0}^{N-1} (f(x_{k+1}) - f(x)) &\leq \frac{L}{2N} \sum_{k=0}^{N-1} (\|x - x_k\|^2 - \|x - x_{k+1}\|^2) \\ &= \frac{L}{2N} (\|x - x_0\|^2 - \|x - x_{k+1}\|^2)\end{aligned}\tag{5}$$

Сходимость градиентного спуска в выпуклом гладком случае [4/4]



- Берем среднее от левой и правой частей от по всем k от 0 до $N - 1$:

$$\begin{aligned}\frac{1}{N} \sum_{k=0}^{N-1} (f(x_{k+1}) - f(x)) &\leq \frac{L}{2N} \sum_{k=0}^{N-1} (\|x - x_k\|^2 - \|x - x_{k+1}\|^2) \\ &= \frac{L}{2N} (\|x - x_0\|^2 - \|x - x_{k+1}\|^2) \\ &\leq \frac{L}{2N} \|x - x_0\|^2.\end{aligned}\tag{5}$$

Сходимость градиентного спуска в выпуклом гладком случае [4/4]



- Берем среднее от левой и правой частей от по всем k от 0 до $N - 1$:

$$\begin{aligned}\frac{1}{N} \sum_{k=0}^{N-1} (f(x_{k+1}) - f(x)) &\leq \frac{L}{2N} \sum_{k=0}^{N-1} (\|x - x_k\|^2 - \|x - x_{k+1}\|^2) \\ &= \frac{L}{2N} (\|x - x_0\|^2 - \|x - x_{k+1}\|^2) \\ &\leq \frac{L}{2N} \|x - x_0\|^2.\end{aligned}\tag{5}$$

- Так как для выпуклых функций (1) градиентный спуск монотонен:

$$\begin{aligned}f(x_k) &\geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle \\ &= f(x_{k+1}) + \alpha \|\nabla f(x_{k+1})\|^2 \\ &\geq f(x_{k+1}),\end{aligned}$$

Сходимость градиентного спуска в выпуклом гладком случае [4/4]



- Берем среднее от левой и правой частей от по всем k от 0 до $N - 1$:

$$\begin{aligned}\frac{1}{N} \sum_{k=0}^{N-1} (f(x_{k+1}) - f(x)) &\leq \frac{L}{2N} \sum_{k=0}^{N-1} (\|x - x_k\|^2 - \|x - x_{k+1}\|^2) \\ &= \frac{L}{2N} (\|x - x_0\|^2 - \|x - x_{k+1}\|^2) \\ &\leq \frac{L}{2N} \|x - x_0\|^2.\end{aligned}\tag{5}$$

- Так как для выпуклых функций (1) градиентный спуск монотонен:

$$\begin{aligned}f(x_k) &\geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle \\ &= f(x_{k+1}) + \alpha \|\nabla f(x_{k+1})\|^2 \\ &\geq f(x_{k+1}),\end{aligned}$$

$$\text{то } \frac{1}{N} \sum_{i=0}^{N-1} (f(x_{i+1}) - f(x)) \geq \min_{i=0, \dots, N-1} f(x_{i+1}) - f(x) = f(x_N) - f(x).$$

Сходимость градиентного спуска в выпуклом гладком случае [4/4]



- Берем среднее от левой и правой частей от по всем k от 0 до $N - 1$:

$$\begin{aligned}\frac{1}{N} \sum_{k=0}^{N-1} (f(x_{k+1}) - f(x)) &\leq \frac{L}{2N} \sum_{k=0}^{N-1} (\|x - x_k\|^2 - \|x - x_{k+1}\|^2) \\ &= \frac{L}{2N} (\|x - x_0\|^2 - \|x - x_{k+1}\|^2) \\ &\leq \frac{L}{2N} \|x - x_0\|^2.\end{aligned}\tag{5}$$

- Так как для выпуклых функций (1) градиентный спуск монотонен:

$$\begin{aligned}f(x_k) &\geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle \\ &= f(x_{k+1}) + \alpha \|\nabla f(x_{k+1})\|^2 \\ &\geq f(x_{k+1}),\end{aligned}$$

то $\frac{1}{N} \sum_{i=0}^{N-1} (f(x_{k+1}) - f(x)) \geq \min_{i=0, \dots, N-1} f(x_{i+1}) - f(x) = f(x_N) - f(x)$. Подставляя это в (5), получаем искомый результат.

Бонус: нижние оценки для градиентных методов

Чёрный ящик



Итерация градиентного спуска:

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

Чёрный ящик



Итерация градиентного спуска:

$$\begin{aligned}x_{k+1} &= x_k - \alpha_k \nabla f(x_k) \\ &= x_{k-1} - \alpha_{k-1} \nabla f(x_{k-1}) - \alpha_k \nabla f(x_k)\end{aligned}$$

Чёрный ящик



Итерация градиентного спуска:

$$\begin{aligned}x_{k+1} &= x_k - \alpha_k \nabla f(x_k) \\&= x_{k-1} - \alpha_{k-1} \nabla f(x_{k-1}) - \alpha_k \nabla f(x_k) \\&\vdots\end{aligned}$$

Чёрный ящик



Итерация градиентного спуска:

$$\begin{aligned}x_{k+1} &= x_k - \alpha_k \nabla f(x_k) \\&= x_{k-1} - \alpha_{k-1} \nabla f(x_{k-1}) - \alpha_k \nabla f(x_k) \\&\vdots \\&= x_0 - \sum_{i=0}^k \alpha_{k-i} \nabla f(x_{k-i})\end{aligned}$$

Чёрный 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

$$\begin{aligned}x_{k+1} &\in x_0 + \text{Lin} \{ \nabla f(x_0), \nabla f(x_1), \dots, \nabla f(x_k) \} && f \text{ — гладкая} \\x_{k+1} &\in x_0 + \text{Lin} \{ g_0, g_1, \dots, g_k \}, \text{ где } g_i \in \partial f(x_i) && f \text{ — негладкая}\end{aligned} \tag{6}$$

Чёрный 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

$$\begin{aligned}x_{k+1} &\in x_0 + \text{Lin} \{ \nabla f(x_0), \nabla f(x_1), \dots, \nabla f(x_k) \} & f &\text{— гладкая} \\x_{k+1} &\in x_0 + \text{Lin} \{ g_0, g_1, \dots, g_k \}, \text{ где } g_i \in \partial f(x_i) & f &\text{— негладкая}\end{aligned} \tag{6}$$

Чтобы построить нижнюю оценку, нам нужно найти функцию f из соответствующего класса, такую, что любой метод из семейства (6) будет работать не быстрее этой нижней оценки.

Гладкий случай



Theorem

Существует L -гладкая и выпуклая функция f , такая, что любой метод (6) для всех 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 L -гладкая и выпуклая функция f , такая, что любой метод (6) для всех k , $1 \leq k \leq \frac{n-1}{2}$, удовлетворяет:

$$f(x_k) - f^* \geq \frac{3L\|x_0 - x^*\|_2^2}{32(k+1)^2}$$

- Какой бы метод из семейства методов первого порядка вы ни использовали, найдётся функция f , на которой скорость сходимости не лучше $\mathcal{O}\left(\frac{1}{k^2}\right)$.

Гладкий случай



i Theorem

Существует L -гладкая и выпуклая функция f , такая, что любой метод (6) для всех k , $1 \leq k \leq \frac{n-1}{2}$, удовлетворяет:

$$f(x_k) - f^* \geq \frac{3L\|x_0 - x^*\|_2^2}{32(k+1)^2}$$

- Какой бы метод из семейства методов первого порядка вы ни использовали, найдётся функция f , на которой скорость сходимости не лучше $\mathcal{O}\left(\frac{1}{k^2}\right)$.
- Ключом к доказательству является явное построение специальной функции f .

Гладкий случай



i Theorem

Существует L -гладкая и выпуклая функция f , такая, что любой метод (6) для всех k , $1 \leq k \leq \frac{n-1}{2}$, удовлетворяет:

$$f(x_k) - f^* \geq \frac{3L\|x_0 - x^*\|_2^2}{32(k+1)^2}$$

- Какой бы метод из семейства методов первого порядка вы ни использовали, найдётся функция f , на которой скорость сходимости не лучше $\mathcal{O}\left(\frac{1}{k^2}\right)$.
- 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 L -гладкая и выпуклая функция f , такая, что любой метод (6) для всех k , $1 \leq k \leq \frac{n-1}{2}$, удовлетворяет:

$$f(x_k) - f^* \geq \frac{3L\|x_0 - x^*\|_2^2}{32(k+1)^2}$$

- Какой бы метод из семейства методов первого порядка вы ни использовали, найдётся функция f , на которой скорость сходимости не лучше $\mathcal{O}\left(\frac{1}{k^2}\right)$.
- 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:
 - а. Нижняя оценка не является точной.

Гладкий случай



i Theorem

Существует L -гладкая и выпуклая функция f , такая, что любой метод (6) для всех k , $1 \leq k \leq \frac{n-1}{2}$, удовлетворяет:

$$f(x_k) - f^* \geq \frac{3L\|x_0 - x^*\|_2^2}{32(k+1)^2}$$

- Какой бы метод из семейства методов первого порядка вы ни использовали, найдётся функция f , на которой скорость сходимости не лучше $\mathcal{O}\left(\frac{1}{k^2}\right)$.
- 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:
 - а. Нижняя оценка не является точной.
 - б. Метод градиентного спуска не является оптимальным для этой задачи.

Гладкий случай



i Theorem

Существует L -гладкая и выпуклая функция f , такая, что любой метод (6) для всех k , $1 \leq k \leq \frac{n-1}{2}$, удовлетворяет:

$$f(x_k) - f^* \geq \frac{3L\|x_0 - x^*\|_2^2}{32(k+1)^2}$$

- Какой бы метод из семейства методов первого порядка вы ни использовали, найдётся функция f , на которой скорость сходимости не лучше $\mathcal{O}\left(\frac{1}{k^2}\right)$.
- 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:
 - а. Нижняя оценка не является точной.
 - б. **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 $A \in \mathbb{R}^{n \times n}$.

Пример, когда $n = 3$:

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

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

- 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 $n = 3$:

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

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

$$x^T A x = 2x_1^2 + 2x_2^2 + 2x_3^2 - 2x_1x_2 - 2x_2x_3$$

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

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

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

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

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

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

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:

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

Наихудшая функция Нестерова

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

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

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

Наихудшая функция Нестерова

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

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

- Гипотеза: $x_i^* = a + bi$ (вдохновлённая физикой). Проверьте, что выполнено 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: $f(x) = \frac{L}{4} \left(\frac{1}{2} x^T A x - e_1^T x \right) = \frac{L}{8} x^T A x - \frac{L}{4} e_1^T x$.
- Оптимальное решение x^* удовлетворяет $Ax^* = e_1$, и решение этой системы уравнений дает:

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

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

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

- И значение функции равно

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

Гладкий случай (доказательство)

- Предположим, что мы начинаем с $x_0 = 0$. Запросив у оракула градиент, мы получаем $g_0 = -\frac{L}{4}e_1$. 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 $g_0 = -\frac{L}{4}e_1$. 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 оракул возвращает градиент $g_1 = \frac{L}{4}(Ax_1 - e_1)$. 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 $g_0 = -\frac{L}{4}e_1$. 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 оракул возвращает градиент $g_1 = \frac{L}{4}(Ax_1 - e_1)$. 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

$$\begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2 \end{bmatrix} \begin{bmatrix} \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow x_2 = \begin{bmatrix} \bullet \\ \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

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

$$x_k = \begin{bmatrix} \bullet & 1 \\ \bullet & 2 \\ \vdots & \vdots \\ \bullet & k \\ 0 & k+1 \\ \vdots & \vdots \\ 0 & n \end{bmatrix}$$

Гладкий случай (доказательство)

- Предположим, что мы начинаем с $x_0 = 0$. Запросив у оракула градиент, мы получаем $g_0 = -\frac{L}{4}e_1$. 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 оракул возвращает градиент $g_1 = \frac{L}{4}(Ax_1 - e_1)$. 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

$$\begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2 \end{bmatrix} \begin{bmatrix} \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow x_2 = \begin{bmatrix} \bullet \\ \bullet \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

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

$$x_k = \begin{bmatrix} \bullet & 1 \\ \bullet & 2 \\ \vdots & \vdots \\ \bullet & k \\ 0 & k+1 \\ \vdots & \vdots \\ 0 & n \end{bmatrix}$$

- 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

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

значение функции в нём $f(\tilde{x}_k)$ будет выше, 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 $\tilde{x}_{k(i)} = 1 - \frac{i}{k+1}$ 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 $\tilde{x}_{k(i)} = 1 - \frac{i}{k+1}$ и $f(\tilde{x}_k) = -\frac{L}{8} \left(1 - \frac{1}{k+1}\right)$.
- Теперь мы имеем:

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

(7)

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 $\tilde{x}_{k(i)} = 1 - \frac{i}{k+1}$ и $f(\tilde{x}_k) = -\frac{L}{8} \left(1 - \frac{1}{k+1}\right)$.
- Теперь мы имеем:

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

(7)

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 $\tilde{x}_{k(i)} = 1 - \frac{i}{k+1}$ и $f(\tilde{x}_k) = -\frac{L}{8} \left(1 - \frac{1}{k+1}\right)$.
- Теперь мы имеем:

$$\begin{aligned} f(x_k) - f(x^*) &\geq f(\tilde{x}_k) - f(x^*) \\ &= -\frac{L}{8} \left(1 - \frac{1}{k+1}\right) - \left(-\frac{L}{8} \left(1 - \frac{1}{n+1}\right)\right) \\ &= \frac{L}{8} \left(\frac{1}{k+1} - \frac{1}{n+1}\right) = \frac{L}{8} \left(\frac{n-k}{(k+1)(n+1)}\right) \end{aligned} \tag{7}$$

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 $\tilde{x}_{k(i)} = 1 - \frac{i}{k+1}$ и $f(\tilde{x}_k) = -\frac{L}{8} \left(1 - \frac{1}{k+1}\right)$.
- Теперь мы имеем:

$$\begin{aligned} f(x_k) - f(x^*) &\geq f(\tilde{x}_k) - f(x^*) \\ &= -\frac{L}{8} \left(1 - \frac{1}{k+1}\right) - \left(-\frac{L}{8} \left(1 - \frac{1}{n+1}\right)\right) \\ &= \frac{L}{8} \left(\frac{1}{k+1} - \frac{1}{n+1}\right) = \frac{L}{8} \left(\frac{n-k}{(k+1)(n+1)}\right) \\ &\stackrel{n=2k+1}{=} \frac{L}{16(k+1)} \end{aligned} \tag{7}$$

Гладкий случай (доказательство)



- Теперь мы ограничиваем $R = \|x_0 - x^*\|_2$:

$$\|x_0 - x^*\|_2^2 = \|0 - x^*\|_2^2 = \|x^*\|_2^2 = \sum_{i=1}^n \left(1 - \frac{i}{n+1}\right)^2$$

Гладкий случай (доказательство)



- Теперь мы ограничиваем $R = \|x_0 - x^*\|_2$:

$$\begin{aligned}\|x_0 - x^*\|_2^2 &= \|0 - x^*\|_2^2 = \|x^*\|_2^2 = \sum_{i=1}^n \left(1 - \frac{i}{n+1}\right)^2 \\ &= n - \frac{2}{n+1} \sum_{i=1}^n i + \frac{1}{(n+1)^2} \sum_{i=1}^n i^2\end{aligned}$$

Гладкий случай (доказательство)



- Теперь мы ограничиваем $R = \|x_0 - x^*\|_2$:

$$\begin{aligned}\|x_0 - x^*\|_2^2 &= \|0 - x^*\|_2^2 = \|x^*\|_2^2 = \sum_{i=1}^n \left(1 - \frac{i}{n+1}\right)^2 \\&= n - \frac{2}{n+1} \sum_{i=1}^n i + \frac{1}{(n+1)^2} \sum_{i=1}^n i^2 \\&\leq n - \frac{2}{n+1} \cdot \frac{n(n+1)}{2} + \frac{1}{(n+1)^2} \cdot \frac{(n+1)^3}{3}\end{aligned}$$

Гладкий случай (доказательство)



- Теперь мы ограничиваем $R = \|x_0 - x^*\|_2$:

$$\begin{aligned}\|x_0 - x^*\|_2^2 &= \|0 - x^*\|_2^2 = \|x^*\|_2^2 = \sum_{i=1}^n \left(1 - \frac{i}{n+1}\right)^2 \\&= n - \frac{2}{n+1} \sum_{i=1}^n i + \frac{1}{(n+1)^2} \sum_{i=1}^n i^2 \\&\leq n - \frac{2}{n+1} \cdot \frac{n(n+1)}{2} + \frac{1}{(n+1)^2} \cdot \frac{(n+1)^3}{3} \\&= \frac{n+1}{3} \stackrel{n=2k+1}{=} \frac{2(k+1)}{3}.\end{aligned}$$

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,

$$k+1 \geq \frac{3}{2} \|x_0 - x^*\|_2^2 = \frac{3}{2} R^2 \quad (8)$$

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,

$$k+1 \geq \frac{3}{2} \|x_0 - x^*\|_2^2 = \frac{3}{2} R^2 \quad (8)$$

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,

$$k+1 \geq \frac{3}{2} \|x_0 - x^*\|_2^2 = \frac{3}{2} R^2 \quad (8)$$

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 (7) и (8), мы получаем:

$$f(x_k) - f(x^*) \geq \frac{L}{16(k+1)} = \frac{L(k+1)}{16(k+1)^2}$$

Гладкий случай (доказательство)



Наконец, используя (7) и (8), мы получаем:

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

Гладкий случай (доказательство)



Наконец, используя (7) и (8), 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 (7) и (8), 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 L -гладкая выпуклая функция f , такая, что любой метод в форме 6 для всех 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 сильно выпуклый случай

Для любого x_0 и любого $\mu > 0$, $\kappa = \frac{L}{\mu} > 1$, существует L -гладкая и μ -сильно выпуклая функция f , такая, что для любого метода в форме 6 выполняются неравенства:

$$\begin{aligned}\|x_k - x^*\|_2 &\geq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x_0 - x^*\|_2 \\ f(x_k) - f^* &\geq \frac{\mu}{2} \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{2k} \|x_0 - x^*\|_2^2\end{aligned}$$

Бонус: ускорение для квадратичных функций

Результат сходимости для квадратичных функций



Предположим, что мы решаем задачу минимизации сильно выпуклой квадратичной функции, с помощью метода градиентного спуска:

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

Результат сходимости для квадратичных функций



Предположим, что мы решаем задачу минимизации сильно выпуклой квадратичной функции, с помощью метода градиентного спуска:

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

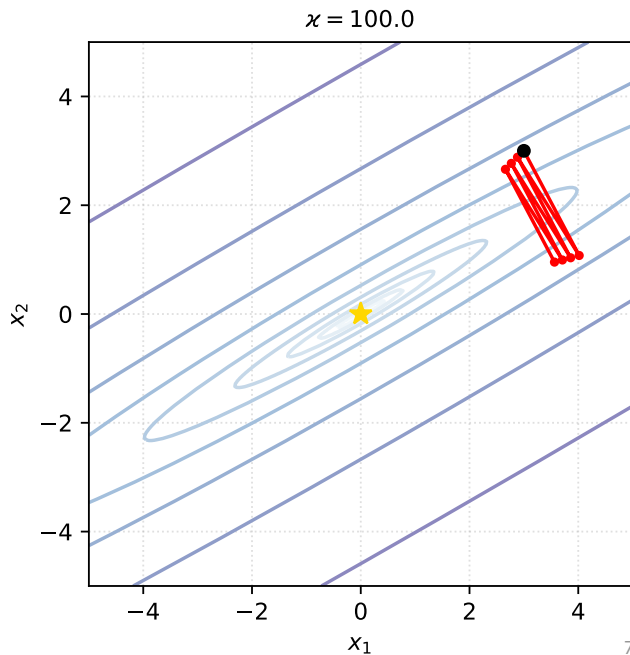
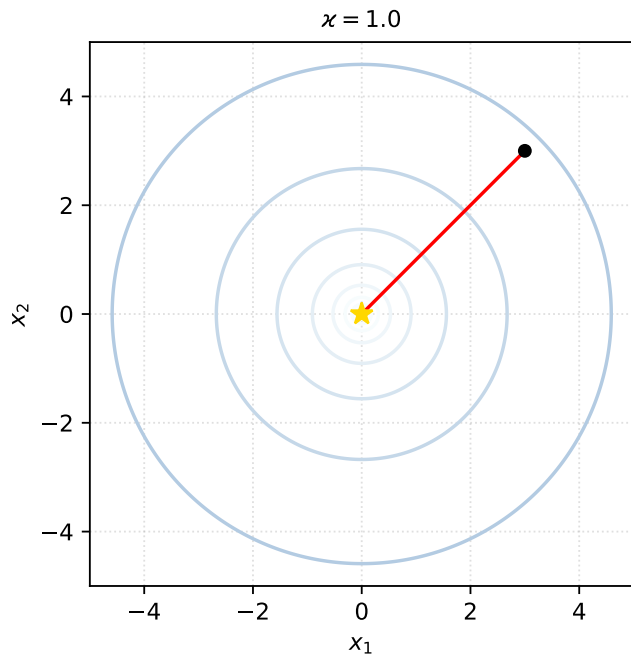
Theorem

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

$$\|x_{k+1} - x^*\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x_0 - x^*\|_2 \quad f(x_{k+1}) - f(x^*) \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} (f(x_0) - f(x^*))$$

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

Число обусловленности \mathcal{N}



Ускорение 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 , а α_k — шаг, который мы определим позже.

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

Ускорение 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 , а α_k — шаг, 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

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

Ускорение 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 , а α_k — шаг, 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

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

Ускорение 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 , а α_k — шаг, 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 постановке задачи: среди всех полиномов, удовлетворяющих $p_k(0) = 1$, мы ищем полином, значение которого как можно меньше отклоняется от нуля на интервале $[\mu, L]$.

Наивное полиномиальное решение



Наивное решение состоит в том, чтобы выбрать равномерный шаг

$\alpha_k = \frac{2}{\mu+L}$. Благодаря этому $|p_k(\mu)| = |p_k(L)|$.

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

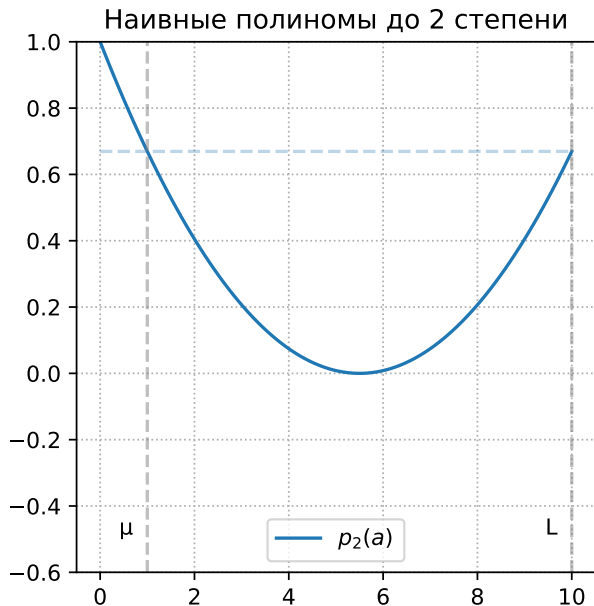
Это точно та же скорость, которую мы доказали в предыдущей лекции для любой гладкой и сильно выпуклой функции.

Давайте посмотрим на этот полином поближе. На правом

рисунке мы выбираем $\mu = 1$ и $L = 10$ 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 $|p_k(\mu)| = |p_k(L)|$.

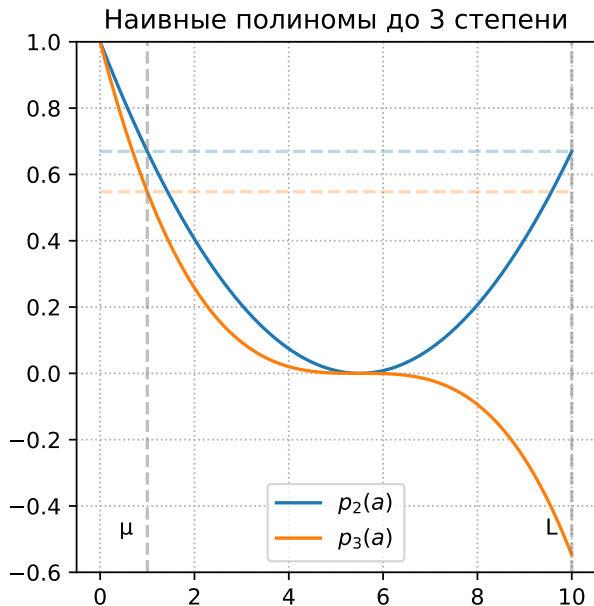
$$\|e_k\| \leq \left(\frac{\kappa-1}{\kappa+1}\right)^k \|e_0\|$$

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

Давайте посмотрим на этот полином поближе. На правом рисунке мы выбираем $\mu = 1$ и $L = 10$ 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 $|p_k(\mu)| = |p_k(L)|$.

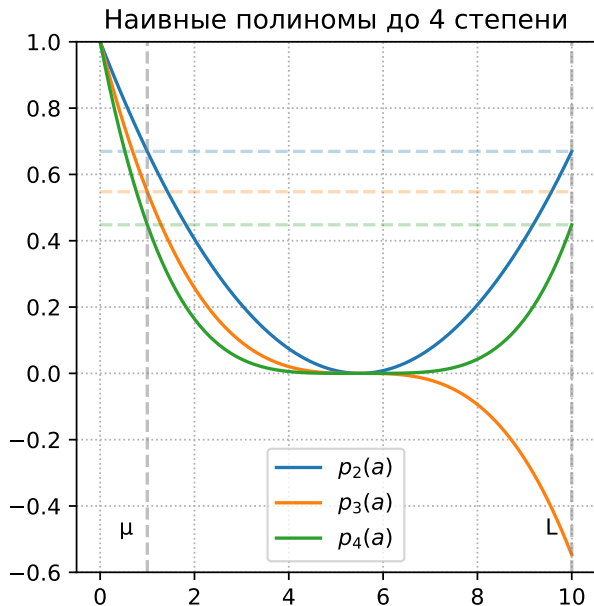
$$\|e_k\| \leq \left(\frac{\kappa-1}{\kappa+1}\right)^k \|e_0\|$$

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

Давайте посмотрим на этот полином поближе. На правом рисунке мы выбираем $\mu = 1$ и $L = 10$ 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 $|p_k(\mu)| = |p_k(L)|$.

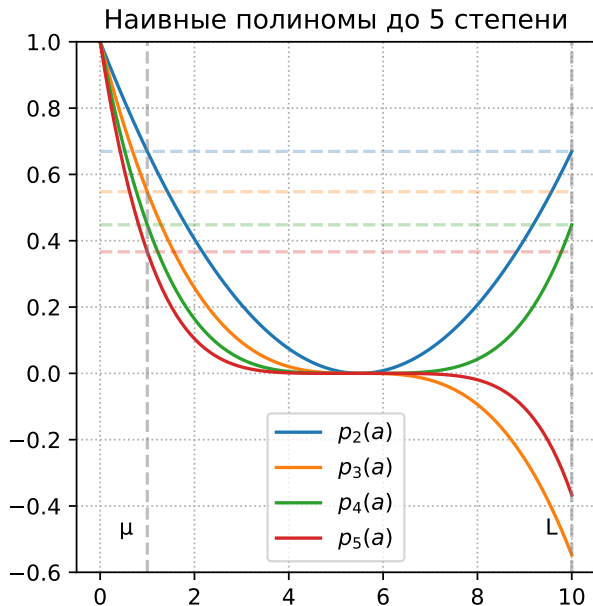
$$\|e_k\| \leq \left(\frac{\kappa-1}{\kappa+1}\right)^k \|e_0\|$$

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

Давайте посмотрим на этот полином поближе. На правом рисунке мы выбираем $\mu = 1$ и $L = 10$ 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 $|p_k(\mu)| = |p_k(L)|$.

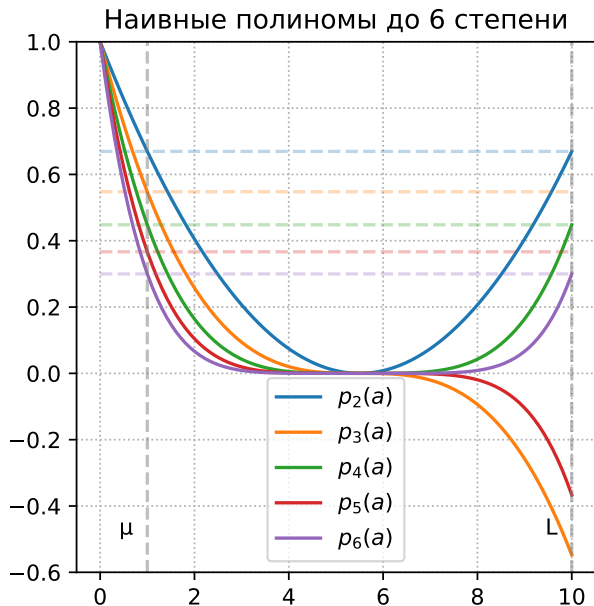
$$\|e_k\| \leq \left(\frac{\varkappa - 1}{\varkappa + 1} \right)^k \|e_0\|$$

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

Давайте посмотрим на этот полином поближе. На правом рисунке мы выбираем $\mu = 1$ и $L = 10$ так, что $\varkappa = 10$.

Следовательно, соответствующий интервал равен $[1, 10]$.

Можем ли мы сделать лучше? Ответ — да.



Полиномы Чебышева



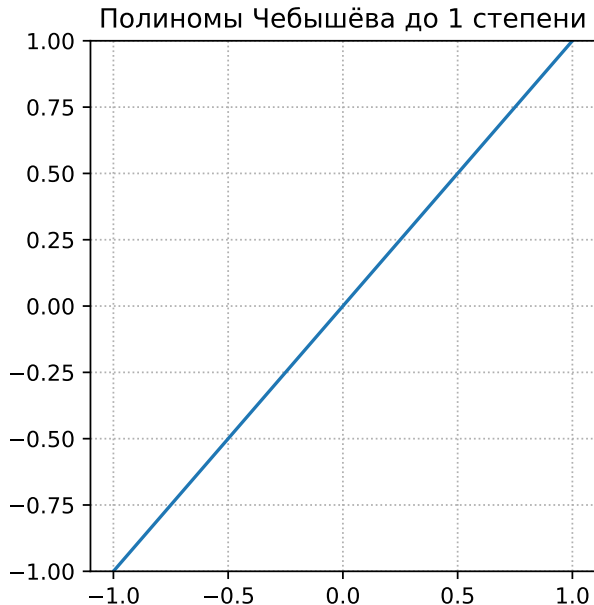
Полиномы Чебышёва дают оптимальный ответ на поставленный вопрос. При соответствующем шкалировании они минимизируют абсолютное значение на заданном интервале $[\mu, L]$, одновременно удовлетворяя нормировочному условию $p(0) = 1$.

$$T_0(x) = 1$$

$$T_1(x) = x$$

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

Давайте построим стандартные полиномы Чебышёва (без масштабирования):



Полиномы Чебышева



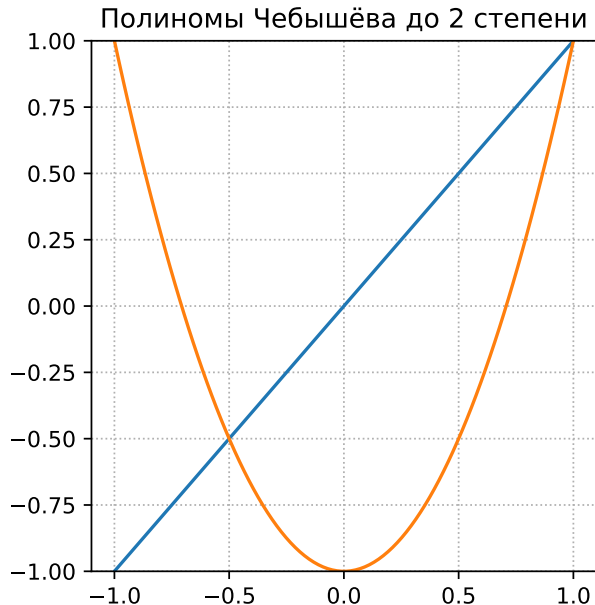
Полиномы Чебышёва дают оптимальный ответ на поставленный вопрос. При соответствующем шкалировании они минимизируют абсолютное значение на заданном интервале $[\mu, L]$, одновременно удовлетворяя нормировочному условию $p(0) = 1$.

$$T_0(x) = 1$$

$$T_1(x) = x$$

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

Давайте построим стандартные полиномы Чебышёва (без масштабирования):



Полиномы Чебышева



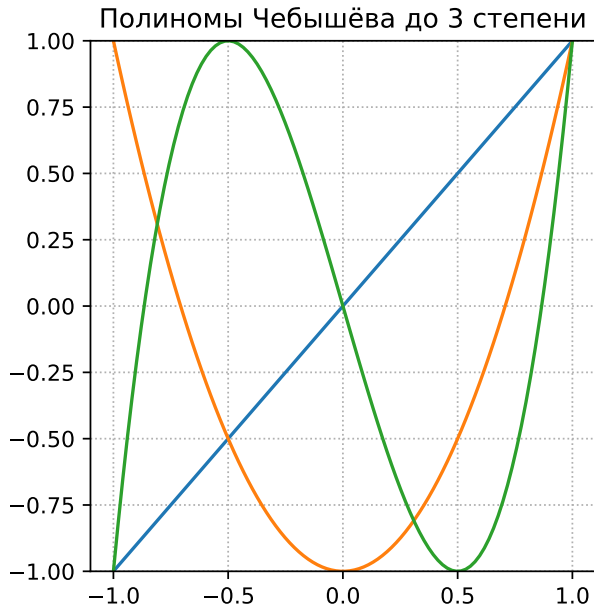
Полиномы Чебышёва дают оптимальный ответ на поставленный вопрос. При соответствующем шкалировании они минимизируют абсолютное значение на заданном интервале $[\mu, L]$, одновременно удовлетворяя нормировочному условию $p(0) = 1$.

$$T_0(x) = 1$$

$$T_1(x) = x$$

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

Давайте построим стандартные полиномы Чебышёва (без масштабирования):



Полиномы Чебышева



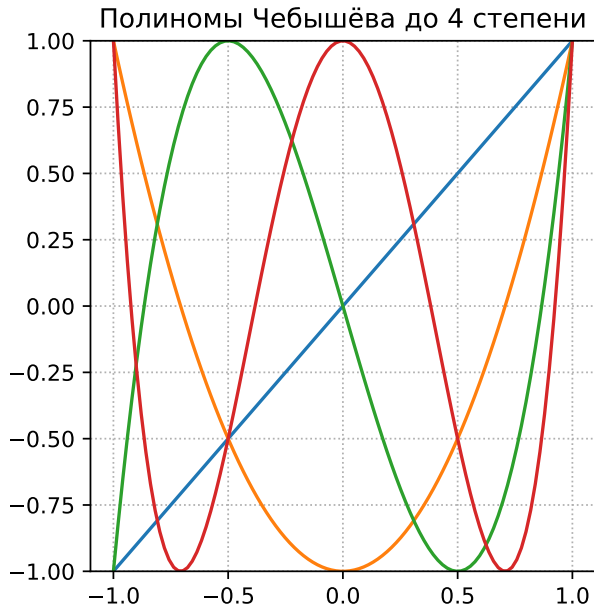
Полиномы Чебышёва дают оптимальный ответ на поставленный вопрос. При соответствующем шкалировании они минимизируют абсолютное значение на заданном интервале $[\mu, L]$, одновременно удовлетворяя нормировочному условию $p(0) = 1$.

$$T_0(x) = 1$$

$$T_1(x) = x$$

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

Давайте построим стандартные полиномы Чебышёва (без масштабирования):



Полиномы Чебышева



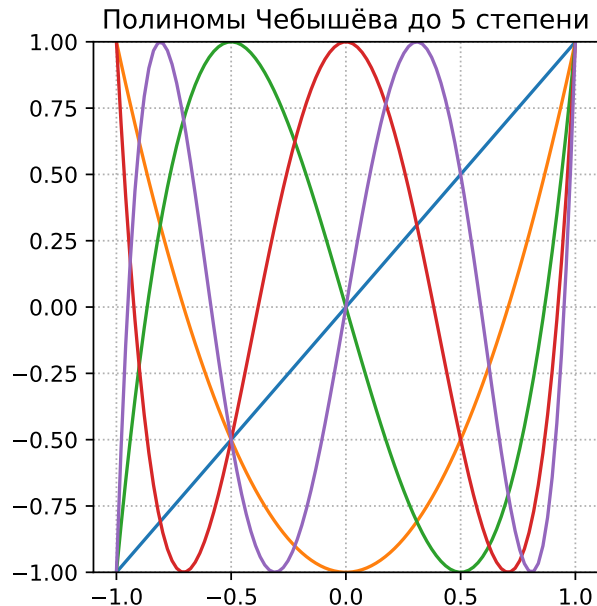
Полиномы Чебышёва дают оптимальный ответ на поставленный вопрос. При соответствующем шкалировании они минимизируют абсолютное значение на заданном интервале $[\mu, L]$, одновременно удовлетворяя нормировочному условию $p(0) = 1$.

$$T_0(x) = 1$$

$$T_1(x) = x$$

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

Давайте построим стандартные полиномы Чебышёва (без масштабирования):



Отшкалированные полиномы Чебышёва



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

Отшкалированные полиномы Чебышёва



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

Отшкалированные полиномы Чебышёва



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

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

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

Отшкалированные полиномы Чебышёва



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

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

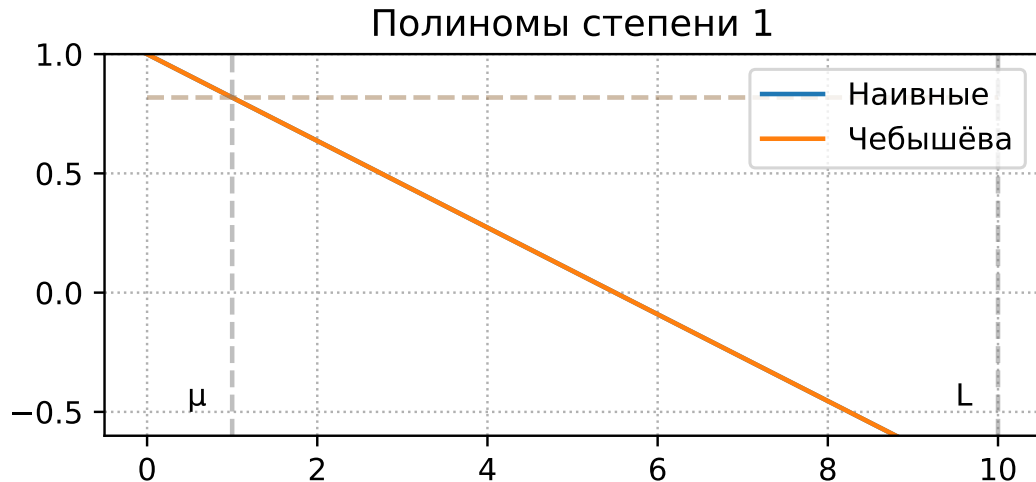
$$\frac{L + \mu}{L - \mu}, \quad \text{что обеспечивает} \quad P_k(0) = T_k\left(\frac{L + \mu - 0}{L - \mu}\right) \cdot T_k\left(\frac{L + \mu}{L - \mu}\right)^{-1} = 1.$$

Построим отшкалированные полиномы Чебышёва

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

и увидим, что они больше подходят для нашей задачи, чем наивные полиномы на интервале $[\mu, L]$.

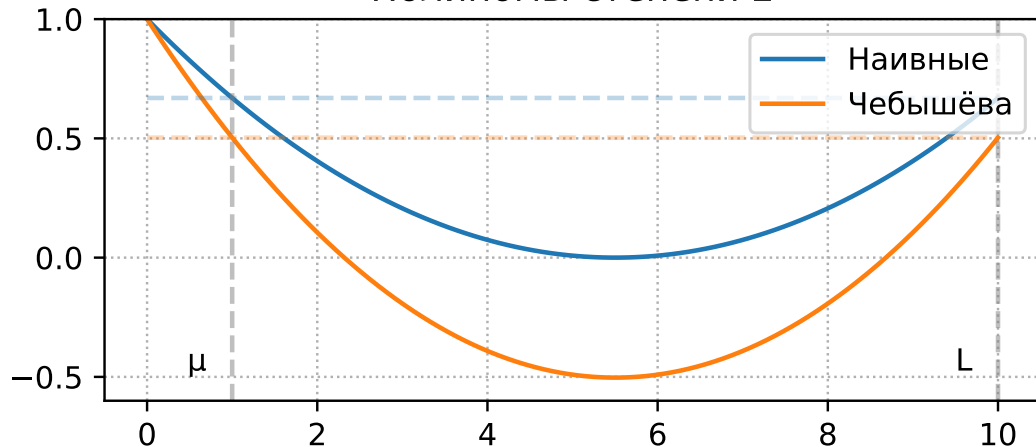
Отшкалированные полиномы Чебышёва



Отшкалированные полиномы Чебышёва



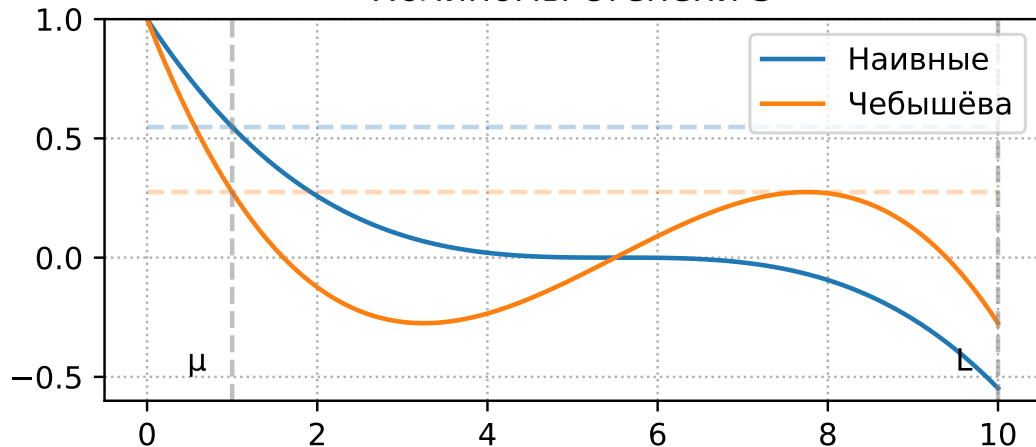
Полиномы степени 2



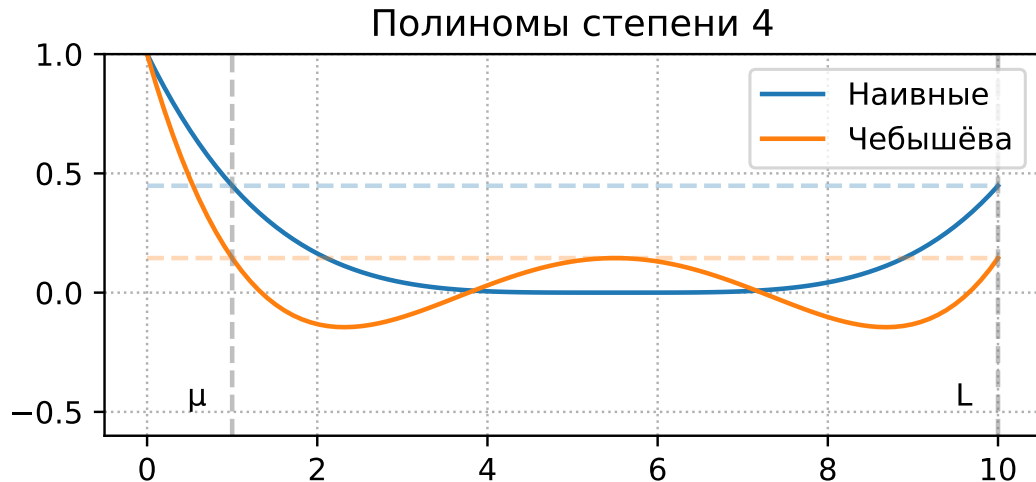
Отшкалированные полиномы Чебышёва



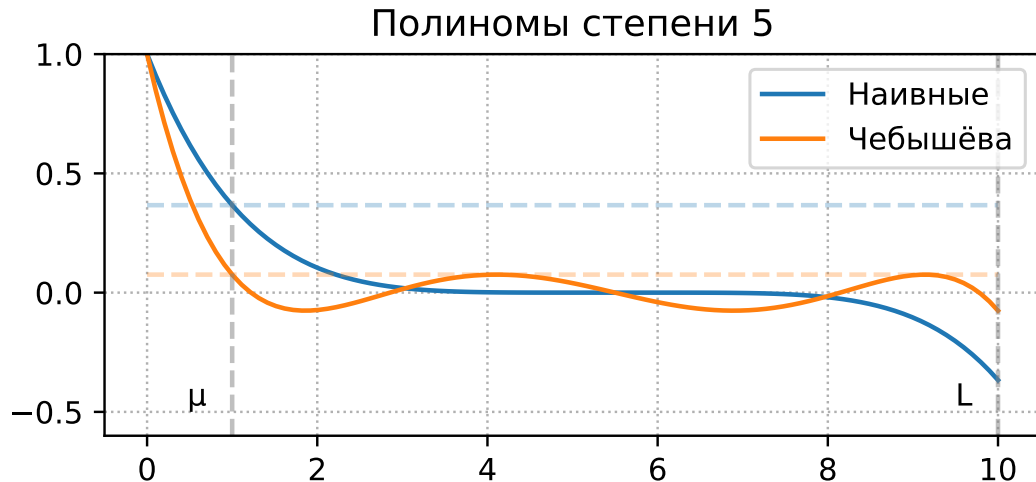
Полиномы степени 3



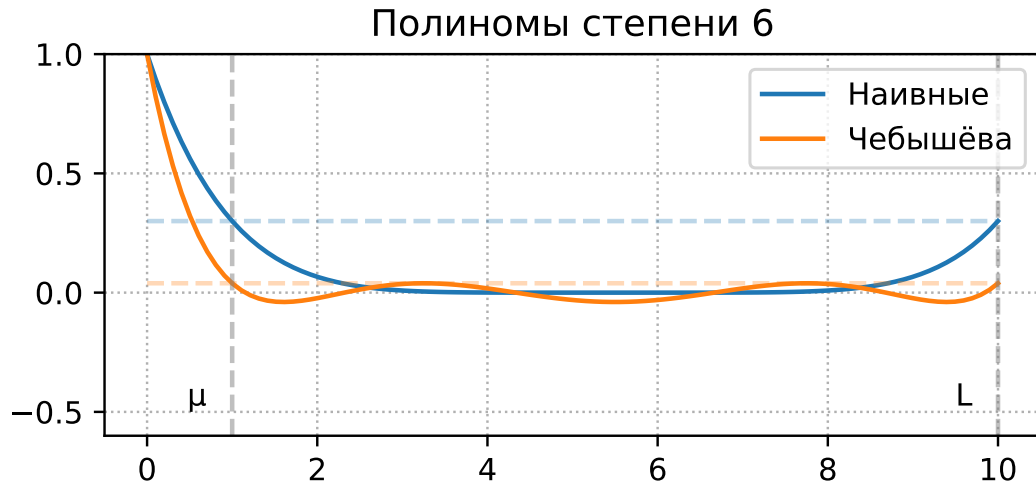
Отшкалированные полиномы Чебышёва



Отшкалированные полиномы Чебышёва



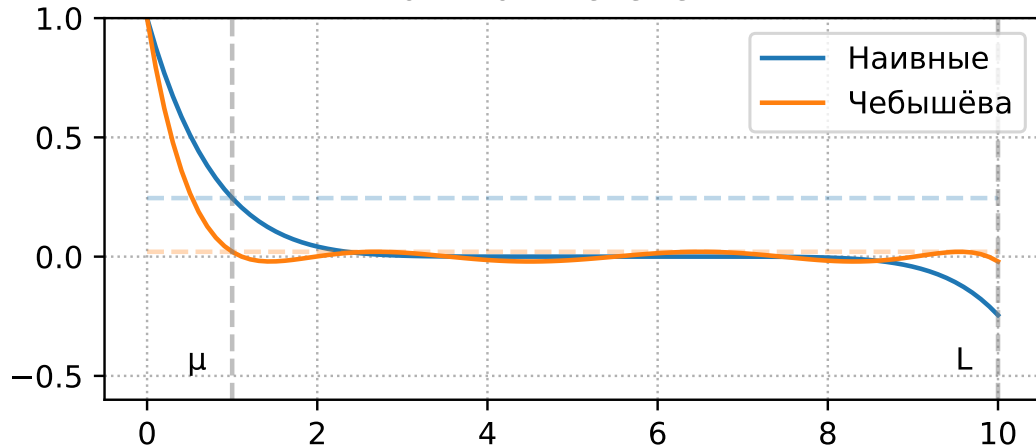
Отшкалированные полиномы Чебышёва



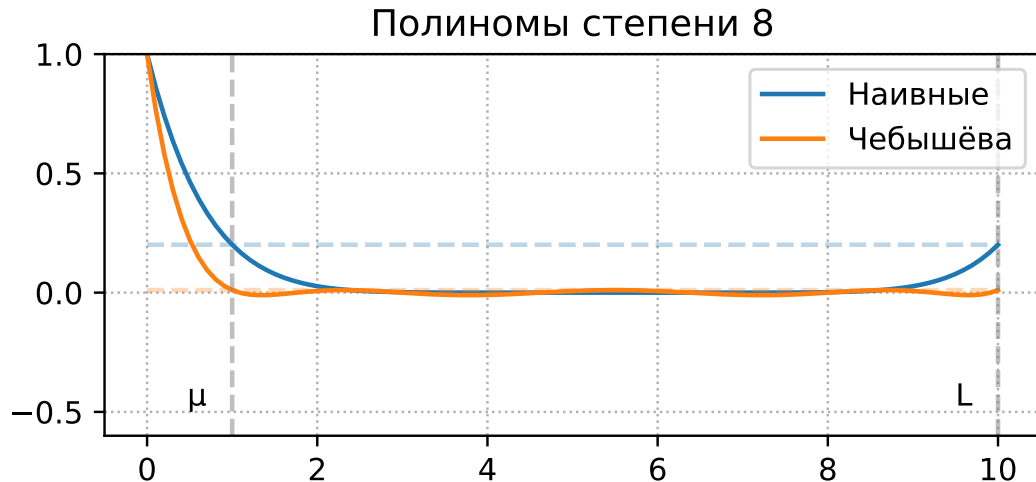
Отшкалированные полиномы Чебышёва



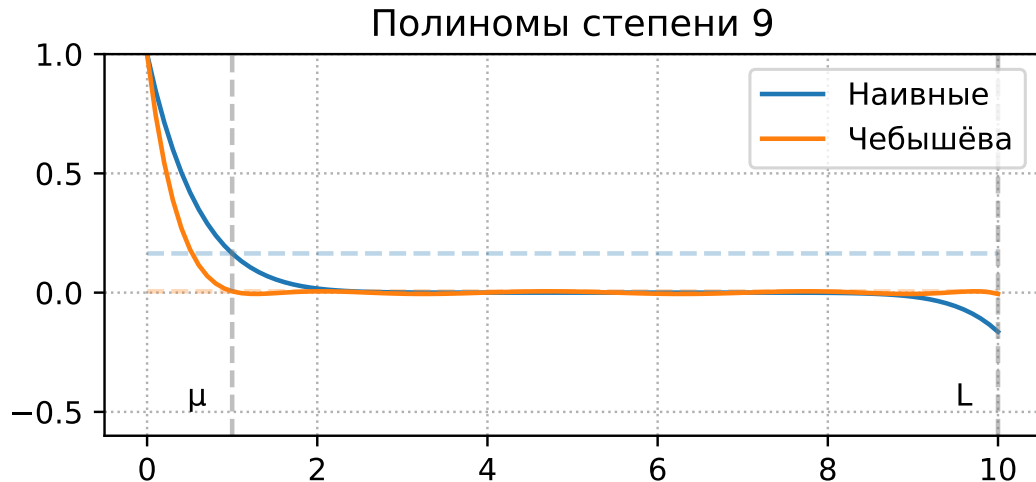
Полиномы степени 7



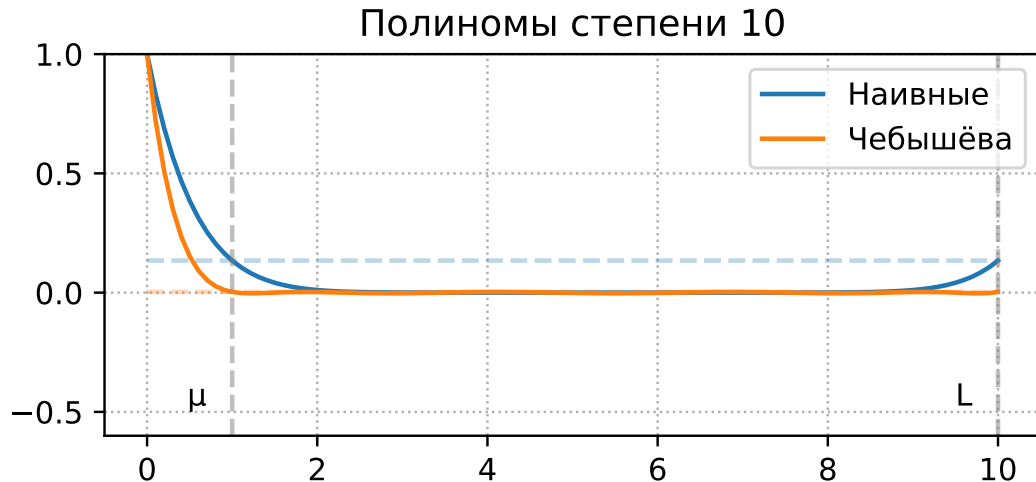
Отшкалированные полиномы Чебышёва



Отшкалированные полиномы Чебышёва



Отшкалированные полиномы Чебышёва



Верхняя оценка для полиномов Чебышёва



Мы можем видеть, что максимальное значение полинома Чебышёва на интервале $[\mu, L]$ достигается на концах отрезка в точках $a = \mu$ и $a = L$. 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 на концах отрезка в точках $a = \mu$ и $a = L$. 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 на концах отрезка в точках $a = \mu$ и $a = L$. 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:

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

Именно в этот момент явно возникнет ускорение. Мы ограничим значение $\|P_k(A)\|_2$ сверху величиной $\left(\frac{1}{1 + \sqrt{\epsilon}}\right)^k$. Для этого детально изучим величину $|T_k(1 + \epsilon)|$.

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

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

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

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

2. Помните, что:

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

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

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

2. Помните, что:

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

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

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

4. Следовательно,

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

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 ускоренного алгоритма. Переформулируя рекурсию в терминах наших отшкалированных полиномов Чебышёва, мы получаем:

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

Ускоренный метод [1/2]



Из-за рекурсивного определения полиномов Чебышёва мы непосредственно получаем итерационную схему ускоренного алгоритма. Переформулируя рекурсию в терминах наших отшкалированных 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:

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

Ускоренный метод [1/2]



Из-за рекурсивного определения полиномов Чебышёва мы непосредственно получаем итерационную схему ускоренного алгоритма. Переформулируя рекурсию в терминах наших отшкалированных 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 ускоренного алгоритма. Переформулируя рекурсию в терминах наших отшкалированных 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:

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

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

Ускоренный метод [1/2]



Из-за рекурсивного определения полиномов Чебышёва мы непосредственно получаем итерационную схему ускоренного алгоритма. Переформулируя рекурсию в терминах наших отшкалированных 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:

$$\begin{aligned} P_k(a) &= T_k\left(\frac{L+\mu-2a}{L-\mu}\right) T_k\left(\frac{L+\mu}{L-\mu}\right)^{-1} \\ T_k\left(\frac{L+\mu-2a}{L-\mu}\right) &= P_k(a) T_k\left(\frac{L+\mu}{L-\mu}\right) \\ T_{k-1}\left(\frac{L+\mu-2a}{L-\mu}\right) &= P_{k-1}(a) T_{k-1}\left(\frac{L+\mu}{L-\mu}\right) \\ T_{k+1}\left(\frac{L+\mu-2a}{L-\mu}\right) &= P_{k+1}(a) T_{k+1}\left(\frac{L+\mu}{L-\mu}\right) \\ P_{k+1}(a)t_{k+1} &= 2\frac{L+\mu-2a}{L-\mu}P_k(a)t_k - P_{k-1}(a)t_{k-1}, \text{ где } t_k = T_k\left(\frac{L+\mu}{L-\mu}\right) \end{aligned}$$

Ускоренный метод [1/2]



Из-за рекурсивного определения полиномов Чебышёва мы непосредственно получаем итерационную схему ускоренного алгоритма. Переформулируя рекурсию в терминах наших отшкалированных 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:

$$\begin{aligned} P_k(a) &= T_k\left(\frac{L+\mu-2a}{L-\mu}\right) T_k\left(\frac{L+\mu}{L-\mu}\right)^{-1} & T_{k-1}\left(\frac{L+\mu-2a}{L-\mu}\right) &= P_{k-1}(a) T_{k-1}\left(\frac{L+\mu}{L-\mu}\right) \\ T_k\left(\frac{L+\mu-2a}{L-\mu}\right) &= P_k(a) T_k\left(\frac{L+\mu}{L-\mu}\right) & T_{k+1}\left(\frac{L+\mu-2a}{L-\mu}\right) &= P_{k+1}(a) T_{k+1}\left(\frac{L+\mu}{L-\mu}\right) \\ P_{k+1}(a) t_{k+1} &= 2 \frac{L+\mu-2a}{L-\mu} P_k(a) t_k - P_{k-1}(a) t_{k-1}, \text{ где } t_k = T_k\left(\frac{L+\mu}{L-\mu}\right) \\ P_{k+1}(a) &= 2 \frac{L+\mu-2a}{L-\mu} P_k(a) \frac{t_k}{t_{k+1}} - P_{k-1}(a) \frac{t_{k-1}}{t_{k+1}} \end{aligned}$$

Ускоренный метод [1/2]



Из-за рекурсивного определения полиномов Чебышёва мы непосредственно получаем итерационную схему ускоренного алгоритма. Переформулируя рекурсию в терминах наших отшкалированных 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:

$$\begin{aligned} P_k(a) &= T_k\left(\frac{L+\mu-2a}{L-\mu}\right) T_k\left(\frac{L+\mu}{L-\mu}\right)^{-1} & T_{k-1}\left(\frac{L+\mu-2a}{L-\mu}\right) &= P_{k-1}(a) T_{k-1}\left(\frac{L+\mu}{L-\mu}\right) \\ T_k\left(\frac{L+\mu-2a}{L-\mu}\right) &= P_k(a) T_k\left(\frac{L+\mu}{L-\mu}\right) & T_{k+1}\left(\frac{L+\mu-2a}{L-\mu}\right) &= P_{k+1}(a) T_{k+1}\left(\frac{L+\mu}{L-\mu}\right) \end{aligned}$$

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

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

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:

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

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:

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

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

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



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 $e_{k+1} = x_{k+1}$.

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 $e_{k+1} = x_{k+1}$.

$$x_{k+1} = P_{k+1}(A)x_0$$

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 $e_{k+1} = x_{k+1}$.

$$x_{k+1} = P_{k+1}(A)x_0 = (I - \alpha_k A)P_k(A)x_0 + \beta_k (P_k(A) - P_{k-1}(A))x_0$$

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:

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

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

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

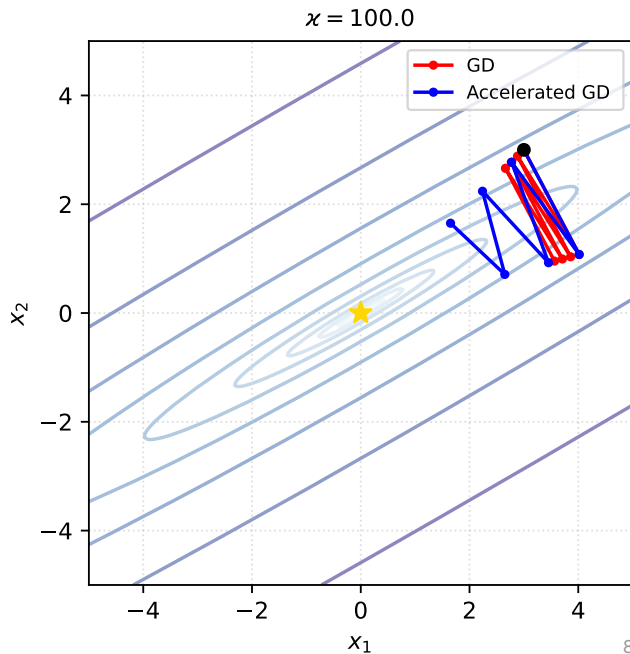
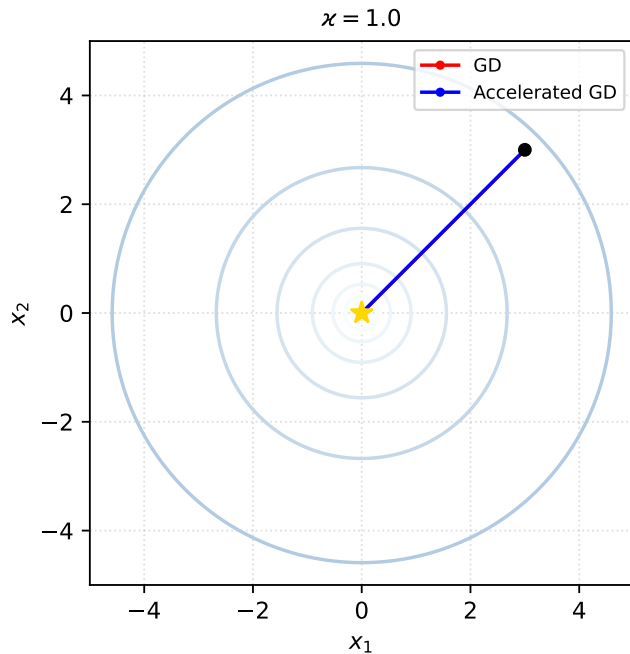
Мы почти закончили :) Помним, что $e_{k+1} = P_{k+1}(A)e_0$. Также обратим внимание, что мы работаем с квадратичной задачей, поэтому мы можем предположить $x^* = 0$ без ограничения общности. В этом случае $e_0 = x_0$ и $e_{k+1} = x_{k+1}$.

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

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



Бонус: анализ сходимости метода тяжёлого шарика

Метод тяжёлого шарика Поляка



Давайте представим идею импульса (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

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



Метод тяжёлого шарика Поляка

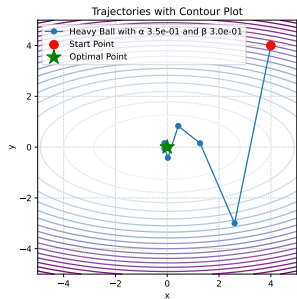
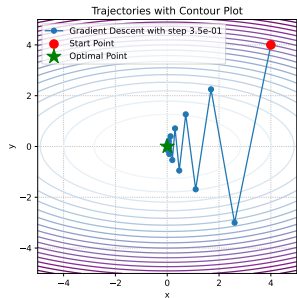


Давайте представим идею импульса (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

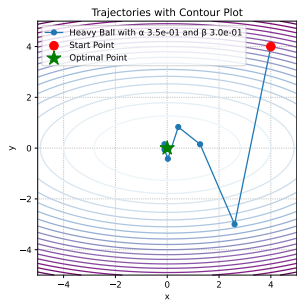
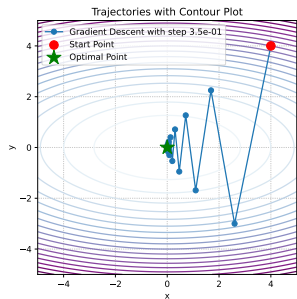
$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}).$$

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 тяжёлого шарика Поляка



Давайте представим идею импульса (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

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

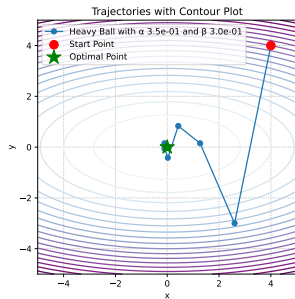
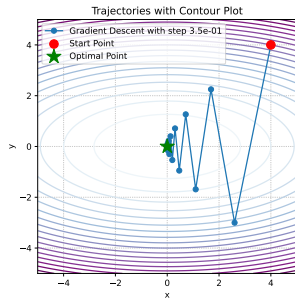
В нашем (квадратичном) случае это

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

Это можно переписать как

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

Метод тяжёлого шарика Поляка



Давайте представим идею импульса (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика 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 введем следующее обозначение: $\hat{z}_k = \begin{bmatrix} \hat{x}_{k+1} \\ \hat{x}_k \end{bmatrix}$. Следовательно, $\hat{z}_{k+1} = M \hat{z}_k$, где матрица итерации M имеет вид:

Метод тяжёлого шарика Поляка



Давайте представим идею импульса (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика 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 введем следующее обозначение: $\hat{z}_k = \begin{bmatrix} \hat{x}_{k+1} \\ \hat{x}_k \end{bmatrix}$. 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 четырьмя блочно-диагональными матрицами размера $d \times d$ внутри. Это означает, что мы можем изменить 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 перегрузку обозначений для простоты.

Сведение к скалярному случаю

Обратим внимание, что M является матрицей $2d \times 2d$ с четырьмя блочно-диагональными матрицами размера $d \times d$ внутри. Это означает, что мы можем изменить 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 перегрузку обозначений для простоты.



Рисунок 3. Иллюстрация перестановки матрицы M

$$\begin{bmatrix} \hat{x}_k^{(1)} \\ \vdots \\ \hat{x}_k^{(d)} \\ \hat{x}_{k-1}^{(1)} \\ \vdots \\ \hat{x}_{k-1}^{(d)} \end{bmatrix} \rightarrow \begin{bmatrix} \hat{x}_k^{(1)} \\ \hat{x}_{k-1}^{(1)} \\ \vdots \\ \hat{x}_k^{(d)} \\ \hat{x}_{k-1}^{(d)} \end{bmatrix} \quad M = \begin{bmatrix} M_1 & & & \\ & M_2 & & \\ & & \dots & \\ & & & M_d \end{bmatrix}$$

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, где λ_i — i -е собственное значение матрицы A , имеем:

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

Сведение к скалярному случаю

Для i -й координаты, где λ_i — i -е собственное значение матрицы A , 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, где λ_i — i -е собственное значение матрицы A , 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-сопряжённую пару, поэтому расстояние до оптимума (в этом случае $\|z_k\|$) обычно не убывает монотонно.

Сходимость метода тяжёлого шарика для квадратичной функции

Мы можем явно вычислить собственные значения M_i :

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

Сходимость метода тяжёлого шарика для квадратичной функции

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. $\beta \geq (1 - \sqrt{\alpha\lambda_i})^2$.

Сходимость метода тяжёлого шарика для квадратичной функции

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. $\beta \geq (1 - \sqrt{\alpha\lambda_i})^2$.

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

Сходимость метода тяжёлого шарика для квадратичной функции

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 $\sqrt{\beta^*}$.