Градиентный спуск. Теоремы сходимости в гладком случае (выпуклые, сильно выпуклые, PL). Верхние и нижние оценки сходимости.

Даня Меркулов

1 Градиентный спуск

1.1 Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль направления h, где $\|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha)$$

Хотим, чтобы h было направлением убывания:

$$f(x + \alpha h) < f(x)$$

$$f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha) < f(x)$$

Переходя к пределу при $\alpha \to 0$:

$$\langle \nabla f(x), h \rangle \leq 0$$

Также из неравенства Коши–Буняковского получаем:

$$\begin{aligned} |\langle \nabla f(x), h \rangle| &\leq \|\nabla f(x)\|_2 \|h\|_2 \\ \langle \nabla f(x), h \rangle &\geq -\|\nabla f(x)\|_2 \|h\|_2 = -\|\nabla f(x)\|_2 \end{aligned}$$

Таким образом, направление антиградиента

$$h = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}$$

представляет собой направление **наискорейшего локального** убывания функции f.

Итерация метода имеет вид:

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

1.2 Дифференциальное уравнение градиентного потока

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x(t)). \tag{GF}$$

Дискретизируем его на равномерной сетке с шагом α:

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

где $x^k \equiv x(t_k)$ и $\alpha = t_{k+1} - t_k$ — шаг сетки.

Отсюда получаем выражение для x^{k+1}:

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

являющееся точной формулой обновления градиентного спуска.

Открыть в Colab ♣

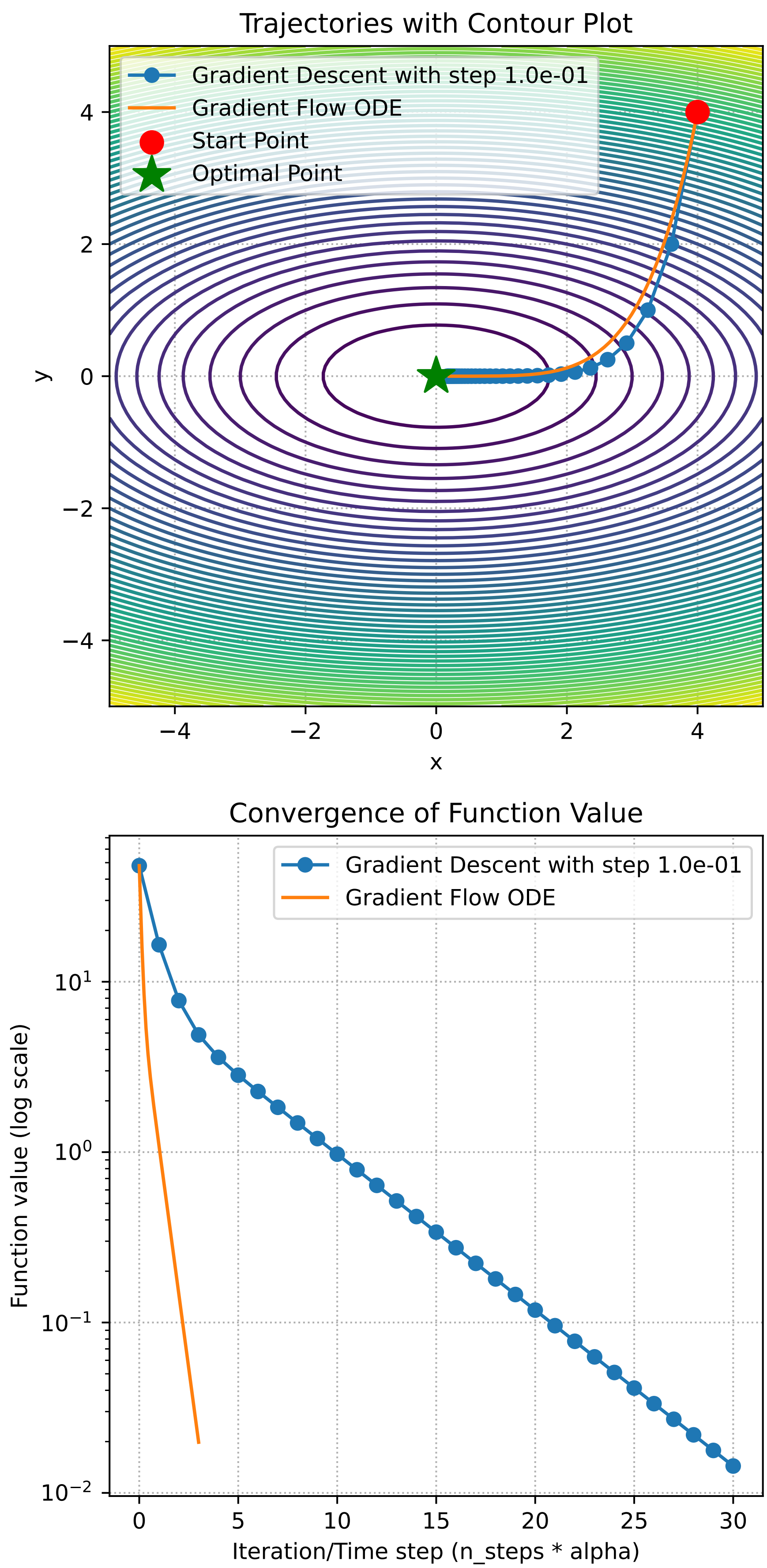
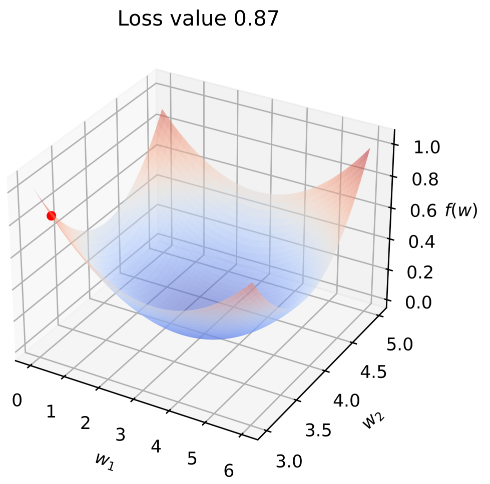
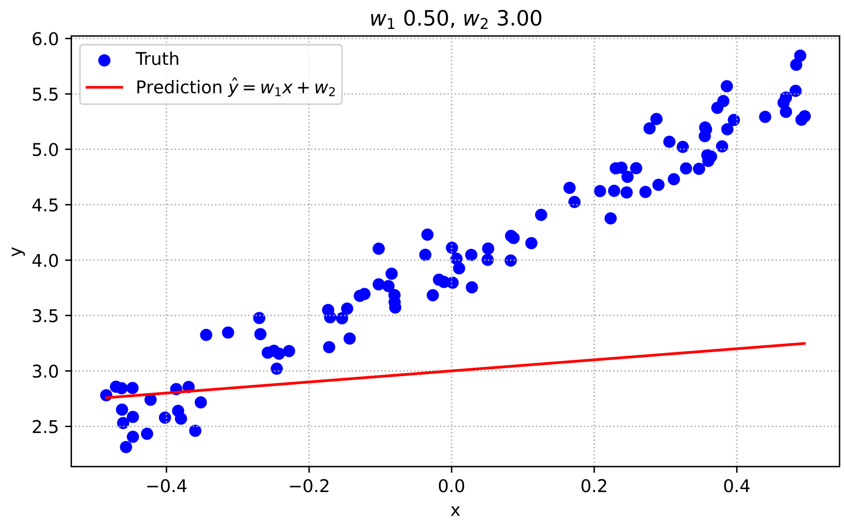


Рисунок 1: Траектория градиентного потока

1.3 Сходимость алгоритма градиентного спуска

Код для построения анимации ниже. Сходимость существенно зависит от выбора шага α:





1.4 Точный линейный поиск (метод наискорейшего спуска)

$$\alpha_k = \arg\min_{\alpha \in \mathbb{R}^+} f(x^k - \alpha \nabla f(x^k))$$

Подход скорее теоретический, чем практический: он удобен для анализа сходимости, но точный линейный поиск часто затруднён, если вычисление функции занимает слишком много времени или стоит слишком дорого.

Интересное теоретическое свойство этого метода заключается в том, что градиенты на соседних итерациях ортогональны. Условие оптимальности по α_k даёт

$$\frac{d}{d\alpha} f(x^k - \alpha \nabla f(x^k))\Big|_{\alpha=\alpha_k} = 0.$$

Условия оптимальности:

$$\nabla f(x^{k+1})^\top \nabla f(x^k) = 0$$

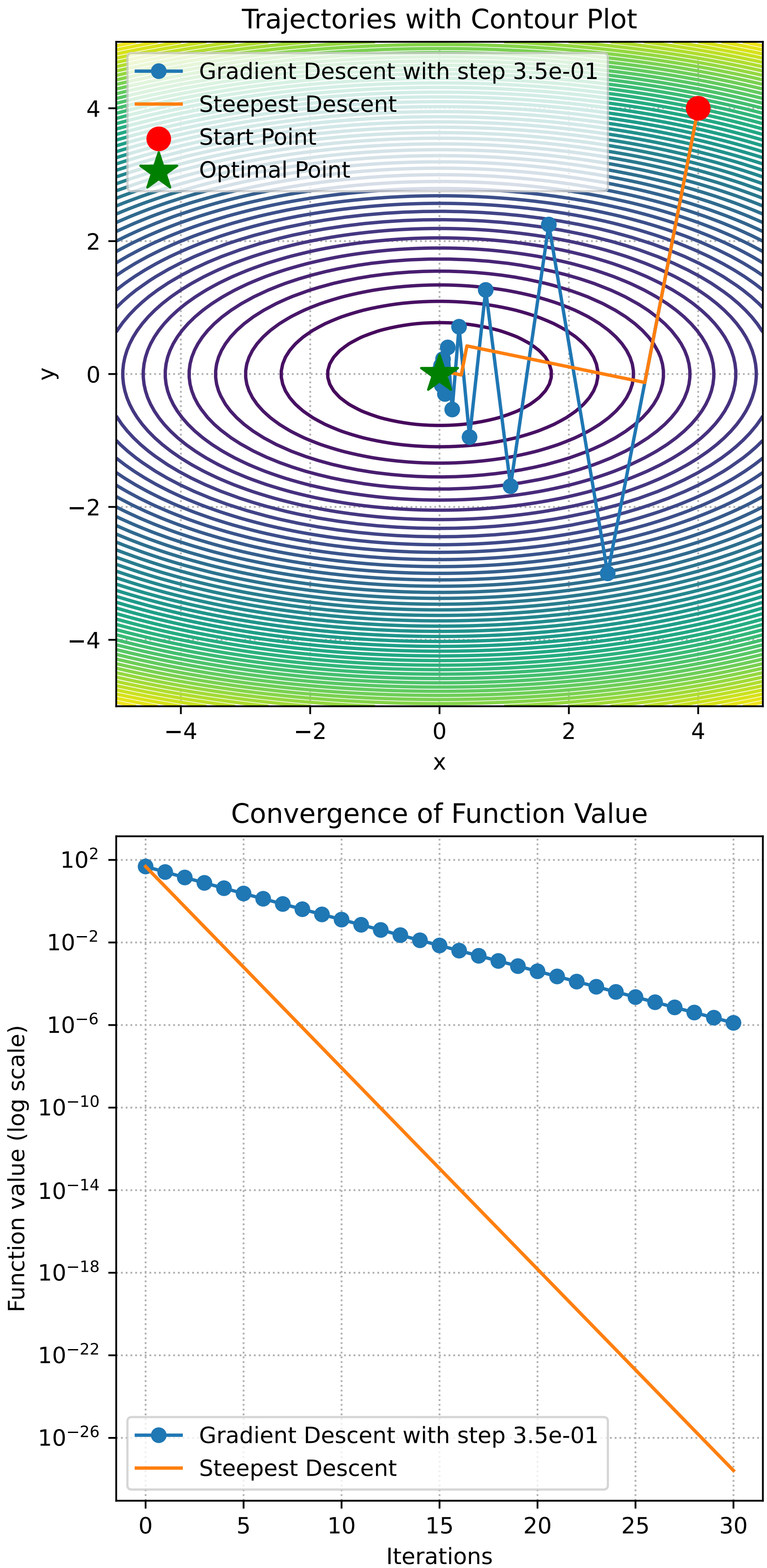


Рисунок 2: Наискорейший спуск

Открыть в Colab ♣

2 Сильно выпуклые квадратичные функции

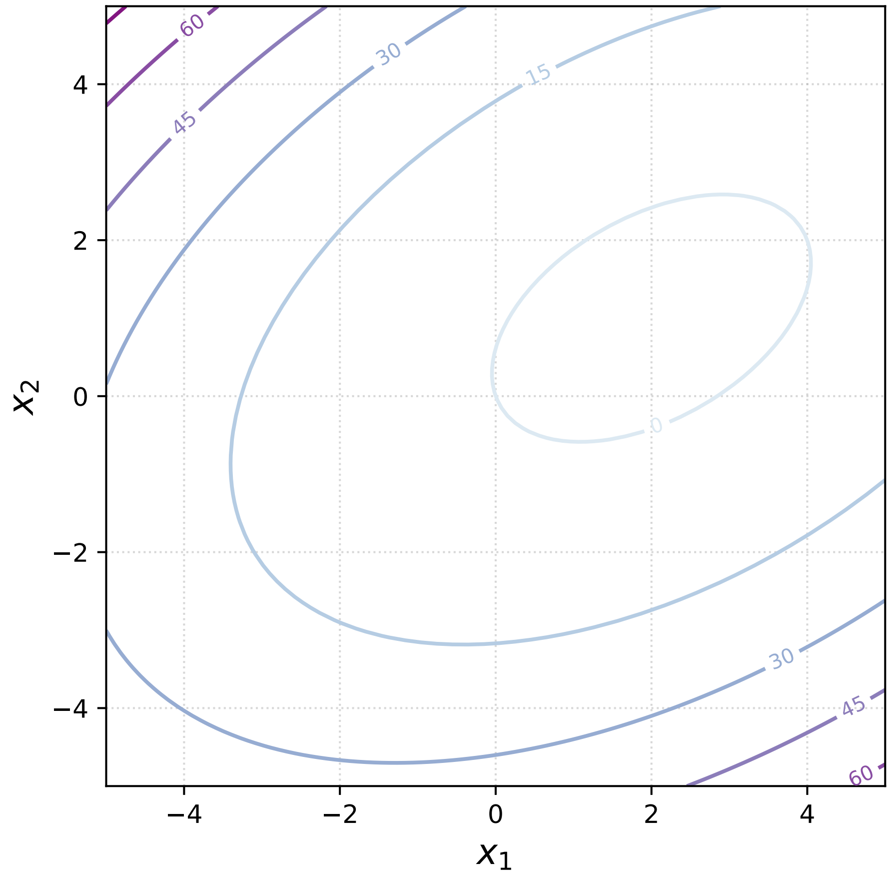
2.1 Сдвиг координат

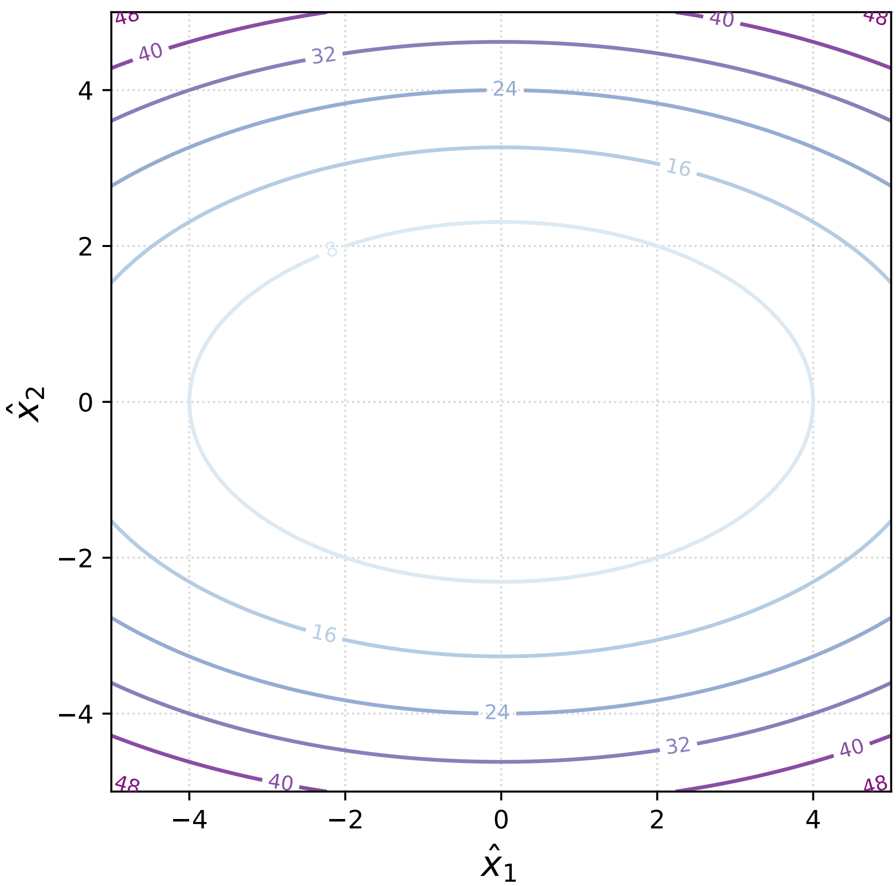
Рассмотрим следующую задачу квадратичной оптимизации:

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

- 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 $x = Q\hat{x} + x^*$.

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


60
45
30
15
0
30
45
60
4
2
0
−2
−4
−4
−2
0
2
4
x_1
x_2


48
40
32
24
16
8
16
24
32
40
48
40
48
40
48
4
2
0
−2
−4
−4
−2
0
2
4
$\hat{x}_1$
$\hat{x}_2$

2.2 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 $\hat{x}$)

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \Lambda x^k \\
&= (I - \alpha^k \Lambda) x^k \\
x_{(i)}^{k+1} &= (1 - \alpha^k \lambda_{(i)})\, x_{(i)}^k \quad \text{для } i\text{-й координаты} \\
x_{(i)}^k &= (1 - \alpha\, \lambda_{(i)})^k\, x_{(i)}^0 \quad \text{при постоянном шаге } \alpha^k = \alpha
\end{aligned}
$$

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

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

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

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

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

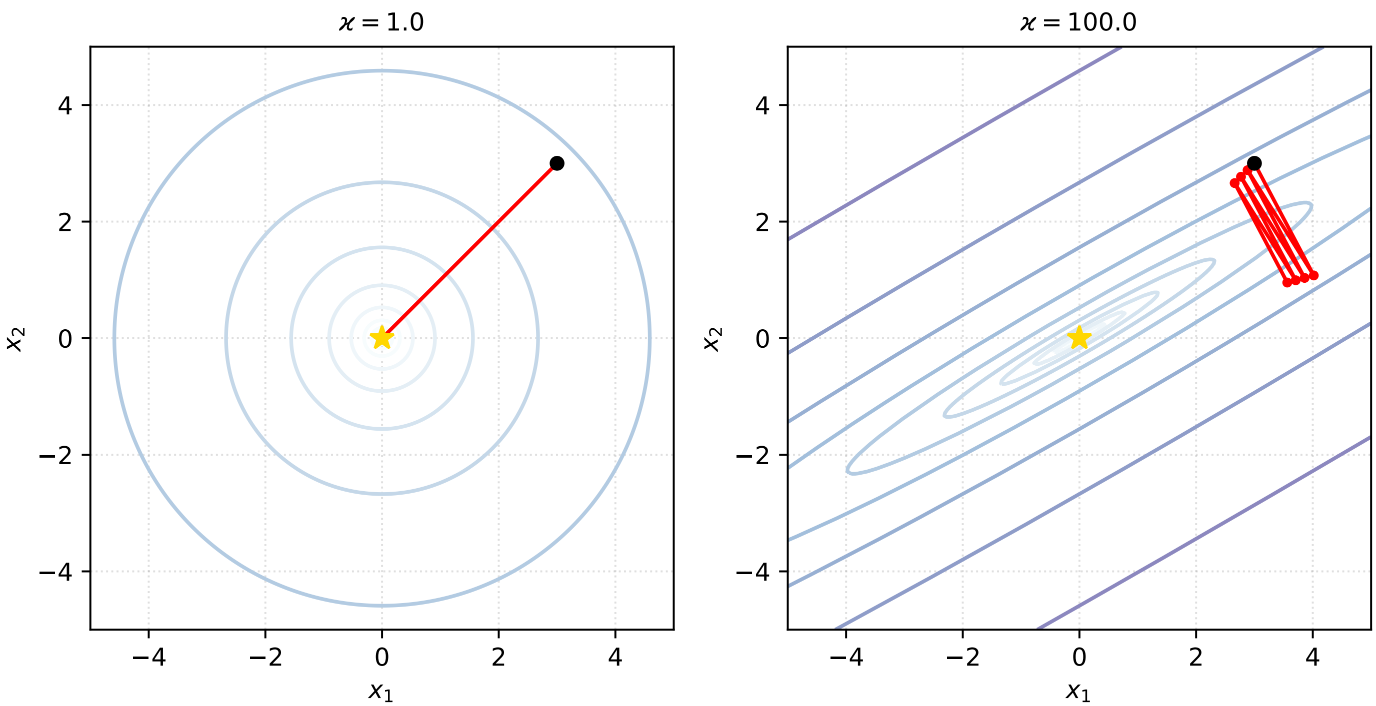
Выберем α, минимизирующий худший знаменатель прогрессии

$$
\begin{aligned}
\rho^* &= \min_\alpha \rho(\alpha) = \min_\alpha \max_i |1 - \alpha\lambda_{(i)}| \\
&= \min_\alpha \{|1 - \alpha\mu|, |1 - \alpha L|\} \\
\alpha^* :& \quad 1 - \alpha^*\mu = \alpha^* L - 1 \\
\alpha^* &= \frac{2}{\mu + L} \quad \rho^* = \frac{L - \mu}{L + \mu} \\
|x_{(i)}^k| &\leq \left(\frac{L - \mu}{L + \mu}\right)^k |x_{(i)}^0| \\
\|x^k\|_2 &\leq \left(\frac{L - \mu}{L + \mu}\right)^k \|x^0\|_2 \quad f(x^k) \leq \left(\frac{L - \mu}{L + \mu}\right)^{2k} f(x^0)
\end{aligned}
$$

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 10 раз	Итераций до уменьшения ошибки по функции в 10 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

2.3 Число обусловленности $\varkappa$



3 Случай PL-функций

3.1 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

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

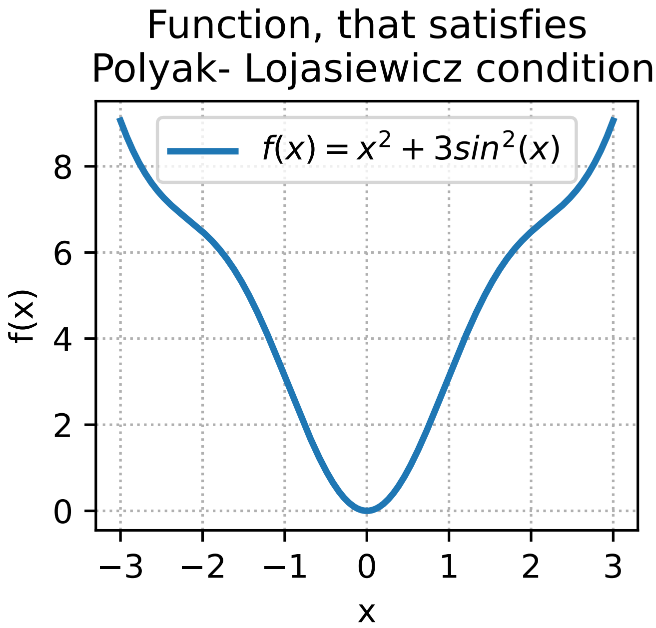


Рисунок 3: PL-функция

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

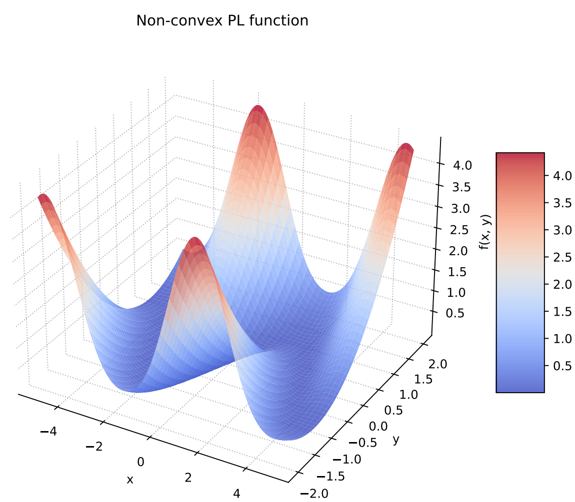


Рисунок 4: PL-функция

3.2 Анализ сходимости

> **Theorem**
>
> Рассмотрим задачу
> $$f(x) \to \min_{x \in \mathbb{R}^d}$$
> 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 $f^* = \min\limits_{x \in \mathbb{R}^d} f(x)$. Тогда:
> $$f(x^k) - f^* \leq (1 - \alpha\mu)^k (f(x^0) - f^*).$$

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

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

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

Теперь используем свойство PL-функции и получаем:

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

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

3.3 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:

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

Положим $y = x^*$:

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

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

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

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

4 Выпуклый гладкий случай

Theorem

Рассмотрим задачу

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

и предположим, что f 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 $0 < \alpha \leq \frac{1}{L}$. Пусть $f^* = \min_{x \in \mathbb{R}^d} f(x)$. Тогда для всех $x^* \in \arg\min f$ и всех $k \in \mathbb{N}$ справедливо:

$$
f(x^k) - f^* \leq \frac{\|x^0 - x^*\|^2}{2\alpha k}.
$$

4.1 Анализ сходимости

- Как и раньше, сначала используем гладкость:

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

  Обычно для сходящегося градиентного спуска чем больше допустимый шаг, тем быстрее сходимость, поэтому часто берут $\alpha = \frac{1}{L}$.

- После этого используем выпуклость:

$$
\begin{aligned}
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \text{ где } y = x^*, x = x^k \\
f(x^k) - f^* \leq \langle \nabla f(x^k), x^k - x^* \rangle
\end{aligned}
\tag{2}
$$

- Теперь подставляем (2) в (1):

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

Пусть $a = x^k - x^*$ и $b = x^k - x^* - \alpha\nabla f(x^k)$. Тогда $a + b = \alpha \nabla f(x^k)$ и $a - b = 2\left(x^k - x^* - \frac{\alpha}{2}\nabla f(x^k)\right)$.

$$
\begin{aligned}
f(x^{k+1}) &\leq f^* + \frac{1}{2\alpha}\left[\|x^k - x^*\|_2^2 - \|x^k - x^* - \alpha\nabla f(x^k)\|_2^2\right] \\
&\leq f^* + \frac{1}{2\alpha}\left[\|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2\right] \\
2\alpha\left(f(x^{k+1}) - f^*\right) &\leq \|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2
\end{aligned}
$$

- Просуммируем по $i = 0, \ldots, k-1$. Большинство слагаемых обнуляется из-за телескопической суммы:

$$
2\alpha \sum_{i=0}^{k-1} \left(f(x^{i+1}) - f^*\right) \leq \|x^0 - x^*\|_2^2 - \|x^k - x^*\|_2^2 \leq \|x^0 - x^*\|_2^2 \tag{3}
$$

- Поскольку на каждой итерации $f(x^{i+1}) \leq f(x^i)$, то

$$
kf(x^k) \leq \sum_{i=0}^{k-1} f(x^{i+1})
$$

- Теперь подставим это в (3):

$$
\begin{aligned}
2\alpha k f(x^k) - 2\alpha k f^* \leq 2\alpha \sum_{i=0}^{k-1}\left(f(x^{i+1}) - f^*\right) \leq \|x^0 - x^*\|_2^2 \\
f(x^k) - f^* \leq \frac{\|x^0 - x^*\|_2^2}{2\alpha k} \leq \frac{L\|x^0 - x^*\|_2^2}{2k}
\end{aligned}
$$

4.2 Итог

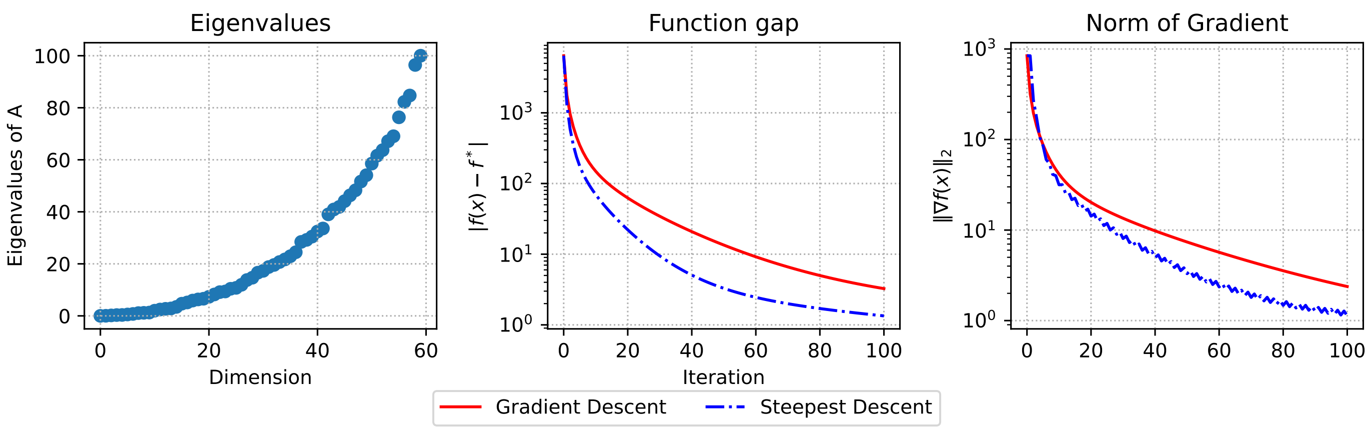
$$\text{Градиентный спуск:} \qquad \min_{x \in \mathbb{R}^n} f(x) \qquad x^{k+1} = x^k - \alpha^k \nabla f(x^k)$$

гладкий (не выпуклый)	гладкий и выпуклый	гладкий и сильно выпуклый (или PL)
$\|\nabla f(x^k)\|^2 \sim \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$f(x^k) - f^* \sim \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\|x^k - x^*\|^2 \sim \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon \sim \mathcal{O}\left(\varkappa \log \frac{1}{\varepsilon}\right)$

4.3 Численные эксперименты

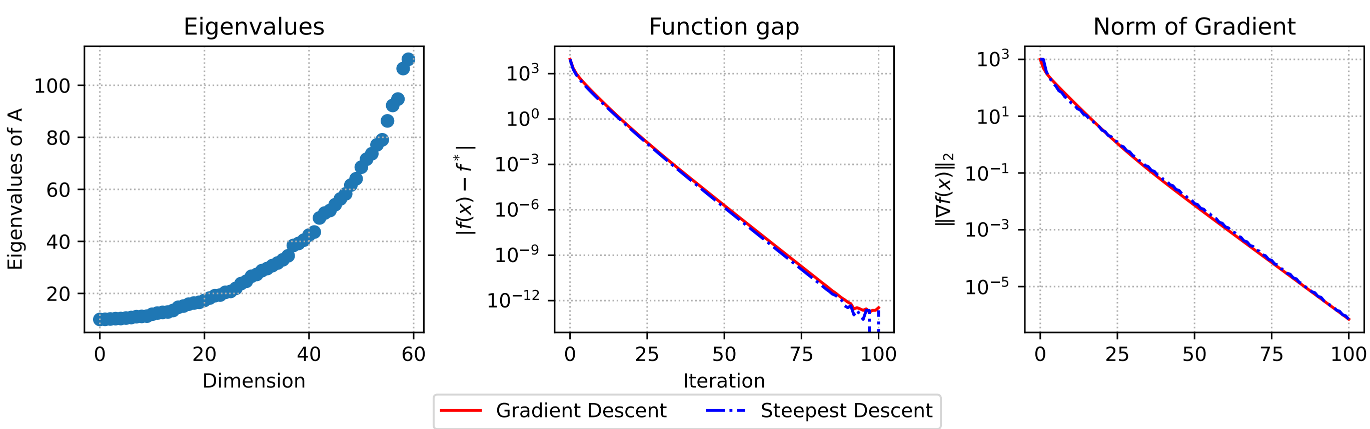
$$f(x) = \frac{1}{2} x^T A x - b^T x \to \min_{x \in \mathbb{R}^n}$$

Convex quadratics. n=60, random matrix.



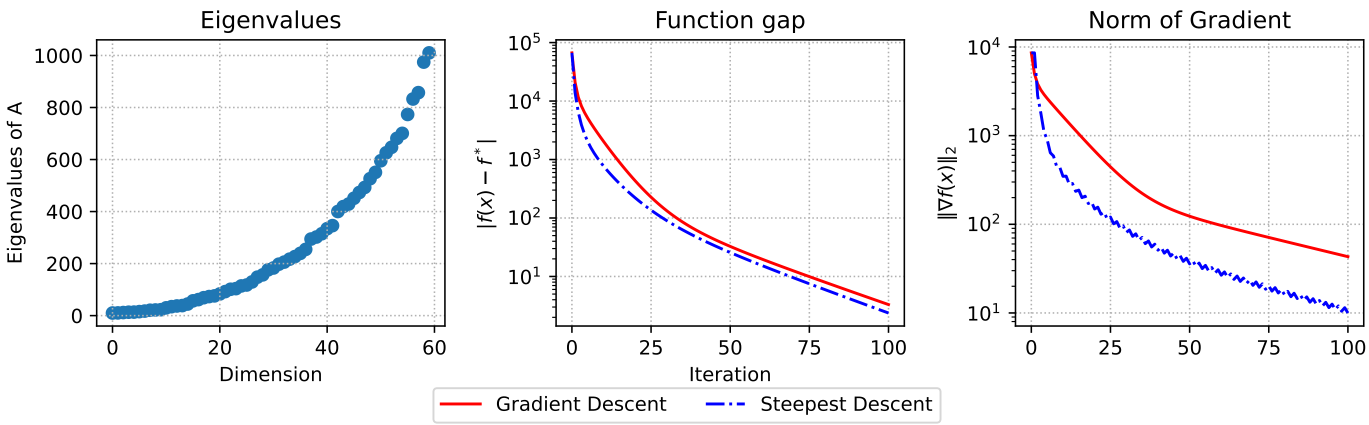
$$f(x) = \frac{1}{2} x^T A x - b^T x \to \min_{x \in \mathbb{R}^n}$$

Strongly convex quadratics. n=60, random matrix.



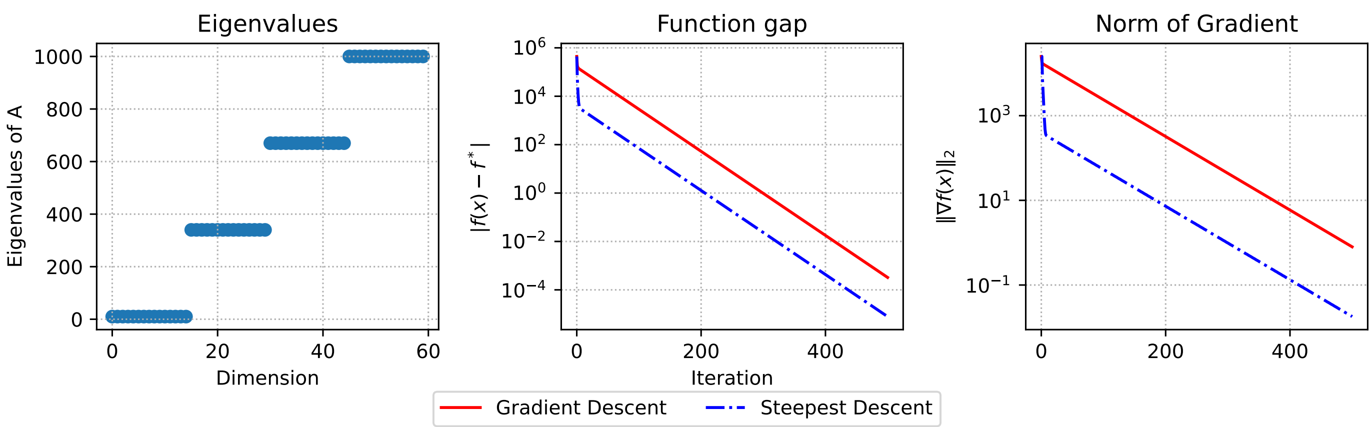
$$f(x) = \frac{1}{2}x^T A x - b^T x \to \min_{x \in \mathbb{R}^n}$$

Strongly convex quadratics. n=60, random matrix.



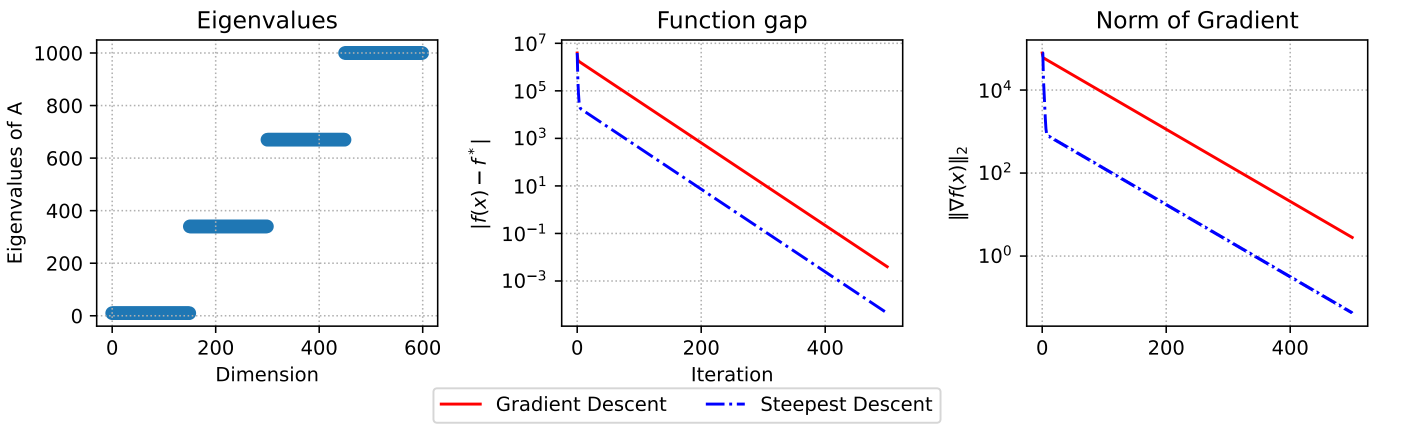
$$f(x) = \frac{1}{2}x^T A x - b^T x \to \min_{x \in \mathbb{R}^n}$$

Strongly convex quadratics. n=60, clustered matrix.



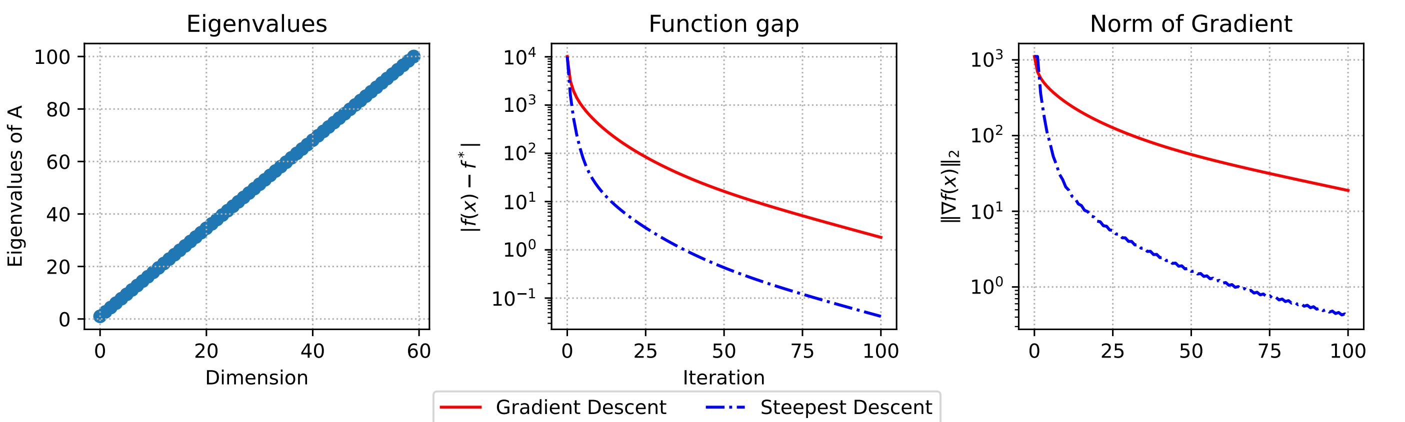
$$f(x) = \frac{1}{2}x^T A x - b^T x \to \min_{x \in \mathbb{R}^n}$$

Strongly convex quadratics. n=600, clustered matrix.



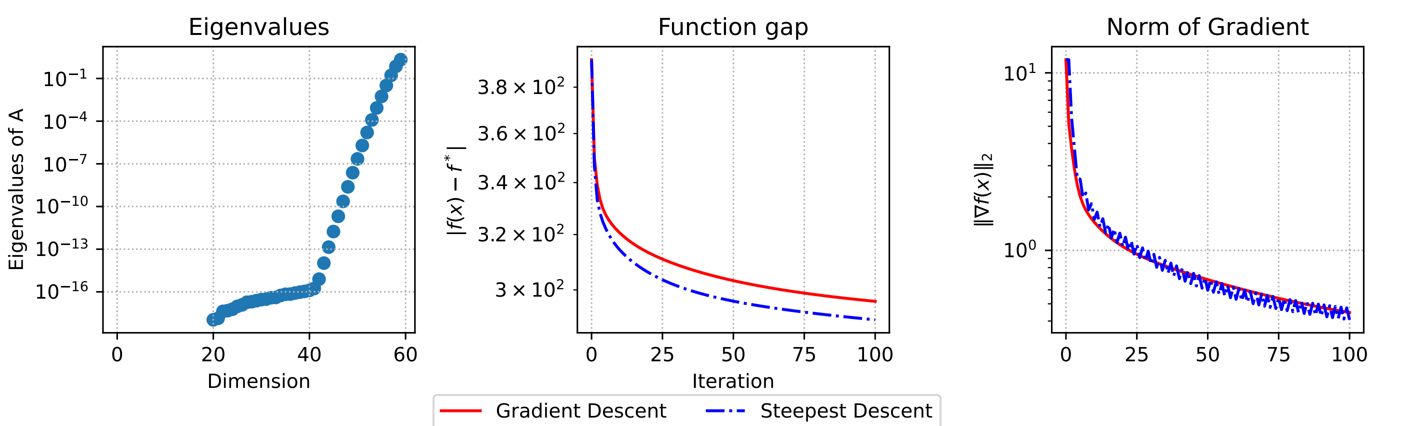
$$
f(x) = \frac{1}{2}x^TAx - b^Tx \to \min_{x \in \mathbb{R}^n}
$$

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



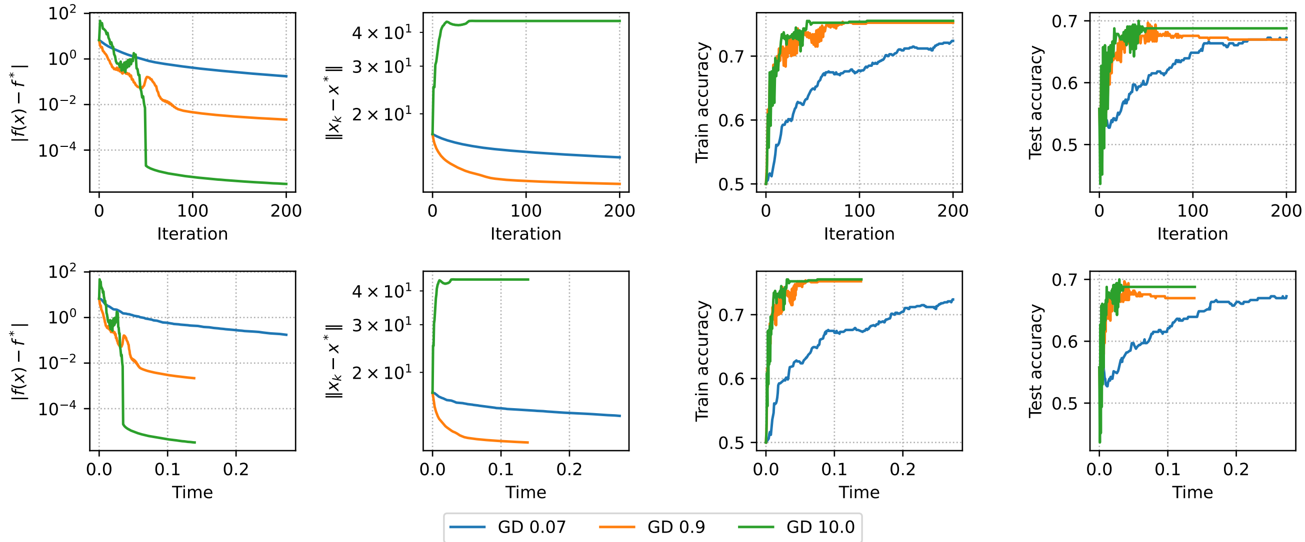
$$
f(x) = \frac{1}{2}x^TAx - b^Tx \to \min_{x \in \mathbb{R}^n}
$$

Strongly convex quadratics. n=60, Hilbert matrix.



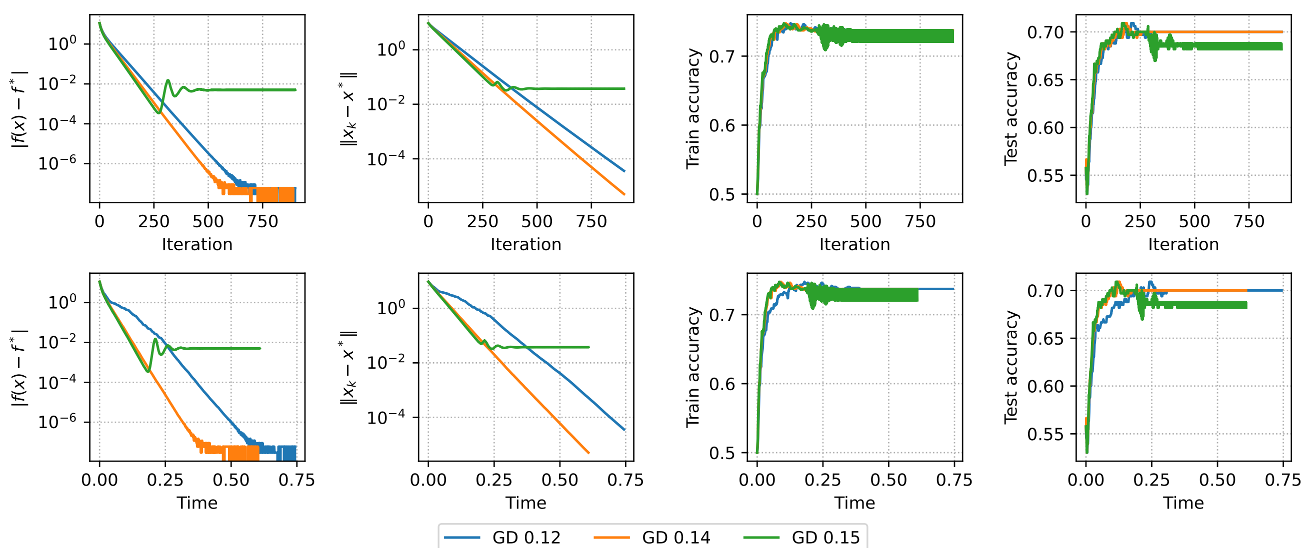
$$f(x) = \frac{\mu}{2}\|x\|_2^2 + \frac{1}{m}\sum_{i=1}^{m} \log(1 + \exp(-y_i\langle a_i, x\rangle)) \to \min_{x\in\mathbb{R}^n}$$

Convex binary logistic regression. mu=0.

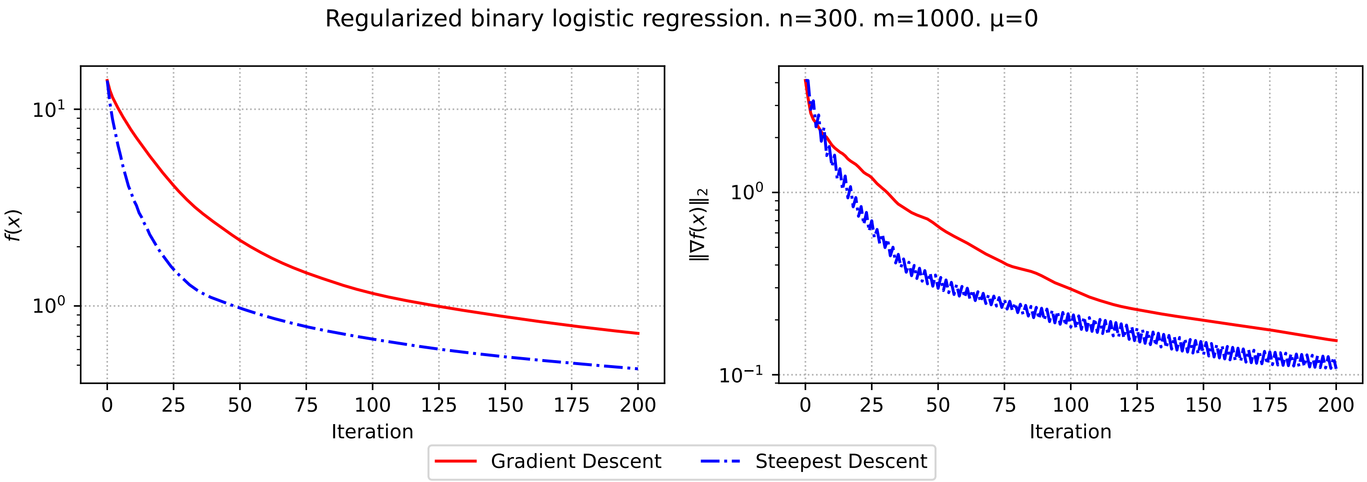


$$f(x) = \frac{\mu}{2}\|x\|_2^2 + \frac{1}{m}\sum_{i=1}^{m} \log(1 + \exp(-y_i\langle a_i, x\rangle)) \to \min_{x\in\mathbb{R}^n}$$

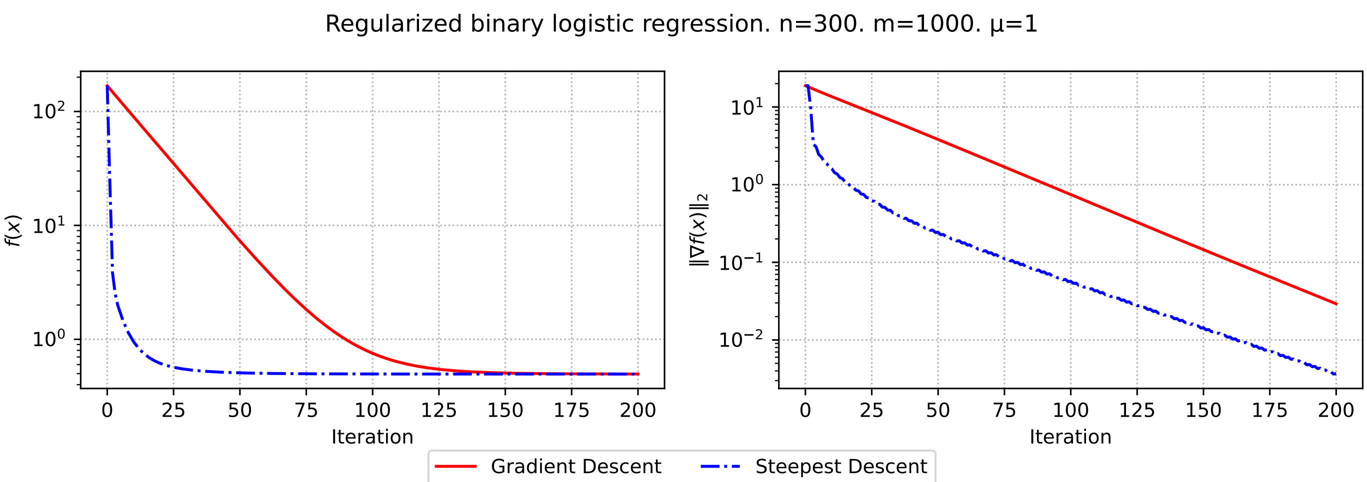
Strongly convex binary logistic regression. mu=0.1.



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



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



5 Задачи

Рассмотрим задачу

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

где $f(x)$ выпукла и L-гладкая. Найдите скорость сходимости градиентного спуска с оптимальным теоретическим шагом $\eta_k = \frac{1}{L}$ *для усредненной точки* и *для лучшей точки*. Другими словами, получите верхние границы на

- $f(\bar{x}_N) - f^*$, where $\bar{x}_N = \frac{1}{N}\sum_{i=0}^{N-1} x_i$,
- $\min_{0\le i\le N-1} f(x_i) - f^*$.

Шаг градиентного спуска

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \left\{ \Psi_k(x) \equiv f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2}\|x - x_k\|_2^2 \right\}$$

Совет

Используйте факт, что $\Psi_k(x)$ является L-строго выпуклой из-за квадратичного регуляризатора.

6 Задачи на дом

6.1 Сходимость градиентного спуска в невыпуклом гладком случае [10 баллов]

Мы не будем делать никаких предположений о выпуклости функции f. Мы покажем, что градиентный спуск достигает ε-стационарной точки x, такой что $\|\nabla f(x)\|_2 \leq \varepsilon$, за $O(1/\varepsilon^2)$ итераций. Важное замечание: вы можете использовать здесь липшицеву параболическую верхнюю оценку:

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|_2^2, \quad \text{for all } x, y. \tag{4}$$

- Подставьте $y = x^{k+1} = x^k - \alpha\nabla f(x^k), x = x^k$ в (Уравнение 4) чтобы показать, что

$$f(x^{k+1}) \leq f(x^k) - \left(1 - \frac{L\alpha}{2}\right)\alpha\|\nabla f(x^k)\|_2^2.$$

- Используйте $\alpha \leq 1/L$, и преобразуйте предыдущий результат, чтобы получить

$$\|\nabla f(x^k)\|_2^2 \leq \frac{2}{\alpha}\left(f(x^k) - f(x^{k+1})\right).$$

- Просуммируйте предыдущий результат по всем итерациям от $1, \ldots, k+1$ чтобы получить

$$\sum_{i=0}^{k} \|\nabla f(x^i)\|_2^2 \leq \frac{2}{\alpha}(f(x^0) - f^*).$$

- Дайте нижнюю оценку сумме в предыдущем результате, чтобы получить

$$\min_{i=0,\ldots,k} \|\nabla f(x^i)\|_2 \leq \sqrt{\frac{2}{\alpha(k+1)}(f(x^0) - f^*)},$$

  что устанавливает желаемую скорость $O(1/\varepsilon^2)$ для достижения ε-стационарности.

6.2 Как сходится градиентный спуск в зависимости от числа обусловленности и размерности. [20 баллов]

Исследуйте, как количество итераций, необходимое для сходимости градиентного спуска, зависит от следующих двух параметров: числа обусловленности $\kappa \geq 1$ функции, которую мы оптимизируем, и размерности n пространства переменных, по которым мы оптимизируем.

Для этого при заданных параметрах n и κ случайно сгенерируйте квадратичную задачу размера n с числом обусловленности κ и запустите на ней градиентный спуск с заранее заданной фиксированной точностью. Измерьте число итераций $T(n, \kappa)$, которое потребовалось методу для сходимости (успешного завершения по критерию останова).

Рекомендация: самый простой способ сгенерировать случайную квадратичную задачу размера n с заданным числом обусловленности κ следующий - удобно взять диагональную матрицу $A \in S_n^{++}$ в виде $A = \text{Diag}(a)$, где диагональные элементы случайно выбираются из интервала $[1, \kappa]$ и удовлетворяют $\min(a) = 1$, $\max(a) = \kappa$. В качестве вектора $b \in \mathbb{R}^n$ можно взять вектор со случайными компонентами. Диагональные матрицы удобны для рассмотрения, поскольку их можно эффективно обрабатывать даже при больших значениях n.

Зафиксируйте определенное значение размерности n. Итерируйте по различным числам обусловленности κ на сетке и постройте зависимость $T(n, \kappa)$ от κ. Поскольку квадратичная задача каждый раз генерируется случайно, повторите этот эксперимент несколько раз. В результате для фиксированного значения n вы должны получить семейство кривых, показывающих зависимость $T(n, \kappa)$ от κ. Изобразите все эти кривые в одном цвете для ясности (например, красный).

Увеличьте значение n и повторите эксперимент. Вы должны получить новое семейство кривых $T(n', \kappa)$ от κ. Изобразите все эти кривые в одном цвете, но отличающемся от предыдущего (например, синий).

Повторите эту процедуру несколько раз для других значений n. В итоге вы должны получить несколько разных семейств кривых - некоторые красные (соответствующие одному значению n), некоторые синие (соответствующие другому значению n), некоторые зеленые и т.д.

Обратите внимание, что имеет смысл перебирать значения размерности n по логарифмической сетке (например, $n = 10$, $n = 100$, $n = 1000$ и т. д.). Используйте следующий критерий остановки: $\|\nabla f(x_k)\|_2^2 \leq \varepsilon \|\nabla f(x_0)\|_2^2$ при $\varepsilon = 10^{-5}$. В качестве начальной точки возьмите $x_0 = (1, \dots, 1)^T$.

Какие выводы можно сделать из полученного рисунка?