Лабораторная работа 1, часть 2: Методы градиентного спуска и метод Ньютона

Выполнила Симакова Елизавета,
студентка 1 курса магистратуры ВБИБ НИУ ВШЭ Санкт-Петербург

20 апреля 2025 г.

Содержание

1 Эксперимент 3.3. Сравнение методов градиентного спуска и Ньютона на задаче логистической регрессии

Описание эксперимента

Целью настоящего эксперимента является сравнение двух методов оптимизации — градиентного спуска и метода Ньютона — при обучении модели логистической регрессии на реальных наборах данных.
Используются три набора данных с сайта LIBSVM. `w8a` - количество ненулевых значений: 579586, `gisette` - количество ненулевых значений: 29729997 и `real-sim` - количество ненулевых значений: 3709083.

Основные параметры эксперимента:

- **Загрузка данных:** формат SVMlight, с помощью функции `load_svmlight_file` из `scikit-learn`, возвращающей разреженную матрицу CSR.
- **Коэффициент регуляризации:** $\lambda = 1/m$, где m — число объектов выборки.
- **Инициализация:** начальная точка $x_0 = 0 \in \mathbb{R}^n$.
- **Параметры методов:** все параметры взяты по умолчанию реализаций из `optimization` и `optimization.newton`.
- **Сбор истории:** включён флаг `trace=True` для записи в каждой итерации:
  - времени работы до текущей итерации;
  - значения целевой функции $f(x_k)$;
  - нормы градиента $\|\nabla f(x_k)\|$.
- **Графики сходимости:**
  1. *Значение функции $f(x_k)$ от реального времени* (сек) для обоих методов на одном графике.
  2. *Относительный квадрат нормы градиента*

$$\frac{\|\nabla f(x_k)\|^2}{\|\nabla f(x_0)\|^2}$$

в логарифмической шкале от времени (сек), также совмещён для двух методов.

Анализ стоимости итерации и памяти

Количество ненулевых элементов:

$$\text{nnz} = \big|\{(i,j) \mid A_{ij} \neq 0\}\big|.$$

Градиентный спуск

Вычислительная сложность

- **Плотная матрица:**

$$
\begin{aligned}
&O(mn) \text{ — } Ax,\\
&O(m) \text{ — } \sigma(Ax) - b,\\
&O(mn) \text{ — } A^\top(\sigma(Ax) - b),\\
&O(n) \text{ — обновление } x.\\
&\text{Итого: } O(mn).
\end{aligned}
$$

- **Разреженная матрица (CSR):**

$$
\begin{aligned}
&O(\text{nnz}) \text{ — } Ax,\\
&O(m) \text{ — } \sigma(Ax) - b,\\
&O(\text{nnz}) \text{ — } A^\top(\sigma(Ax) - b),\\
&O(n) \text{ — обновление } x.\\
&\text{Итого: } O(\text{nnz} + m + n) \approx O(\text{nnz}).
\end{aligned}
$$

Память

- **Плотная матрица:** хранение $A \in \mathbb{R}^{m\times n}$ и $x \in \mathbb{R}^n$ — $O(mn + n)$.
- **Разреженная матрица (CSR):** хранение структуры CSR (nnz, индексов, указателей) и вектор x — $O(\text{nnz} + m + n + n)$.

Метод Ньютона

Вычислительная сложность

- **Плотная матрица:**

$$
\begin{aligned}
&O(mn) \text{ — градиент},\\
&O(mn^2) \text{ — гессиан } A^\top DA,\\
&O(n^3) \text{ — решение системы } H\Delta x = -g,\\
&O(n) \text{ — обновление } x.\\
&\text{Итого: } O(mn^2 + n^3).
\end{aligned}
$$

- **Разреженная матрица (CSR):**

$$
\begin{aligned}
&O(\text{nnz}) \text{ — градиент},\\
&O(\text{nnz}) \text{ — гессиан } A^\top DA,\\
&O(n^3) \text{ — решение разреженной системы},\\
&O(n) \text{ — обновление } x.\\
&\text{Итого: } O(\text{nnz} + n^3),
\end{aligned}
$$

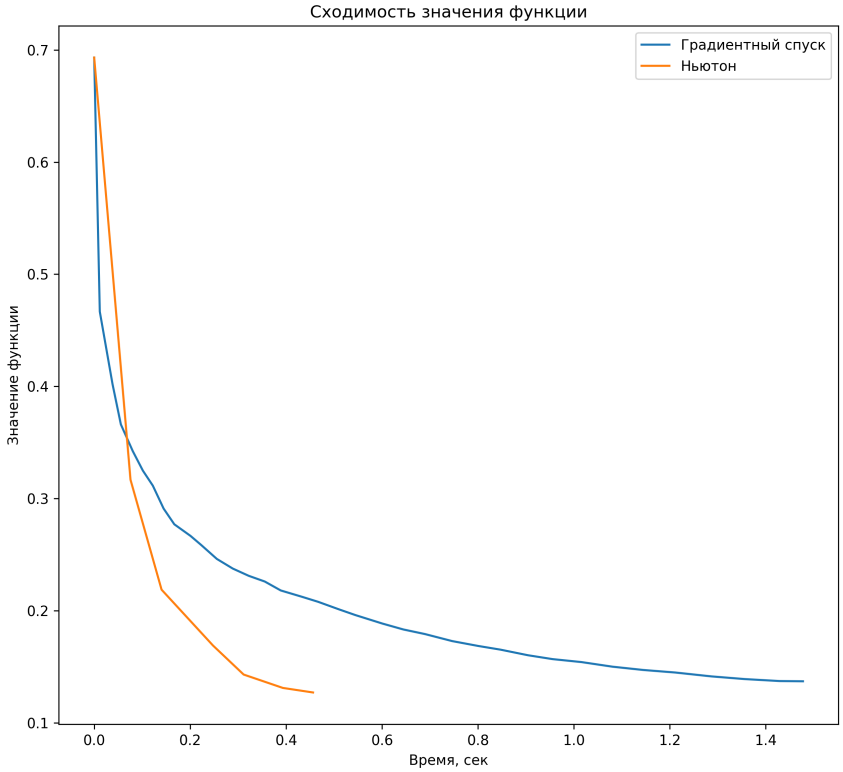
Память

- **Плотная матрица:** хранение A, x— $O(mn + n)$.
- **Разреженная матрица (CSR):** хранение A, x — $O(\text{nnz} + m + n + n)$.

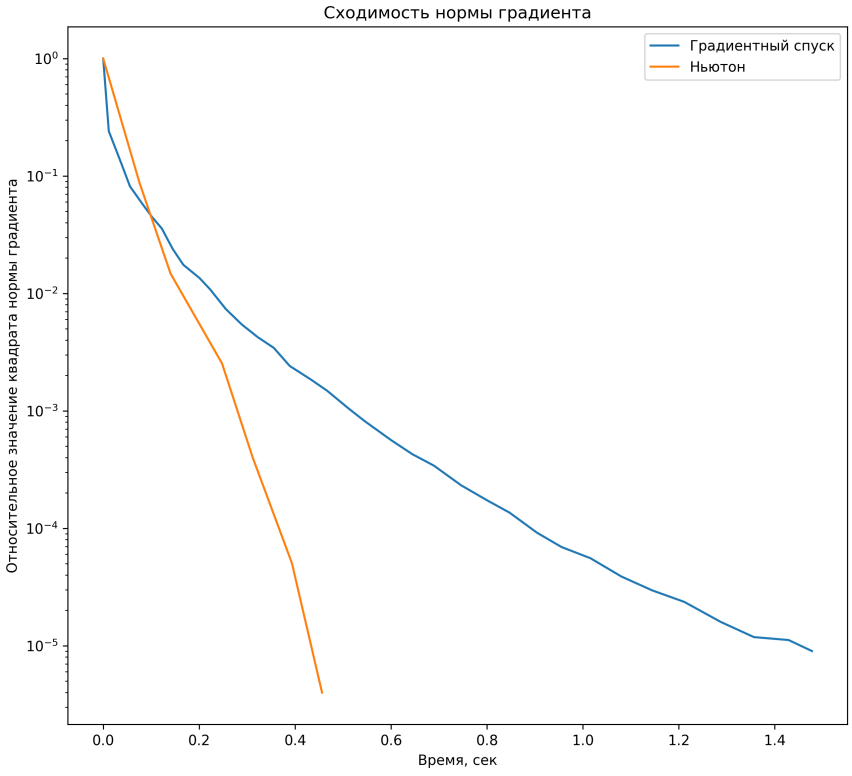
Результаты эксперимента

Для каждого из трёх наборов данных представлены два графика: сходимость значения функции и сходимость нормы градиента.

Набор данных w8a



(a) Зависимость значения функции от времени



(b) Сходимость нормы градиента

Рис. 1: Методы градиентного спуска и Ньютона на датасете `w8a`

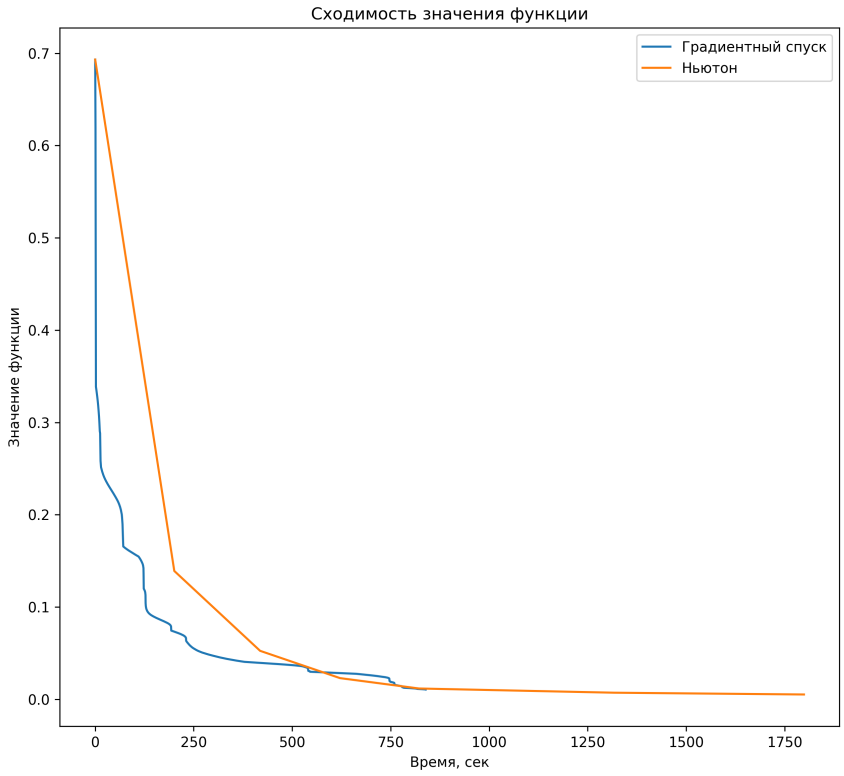
Комментарии (w8a):

(a) *Зависимость значения функции от времени:*
Метод Ньютона достигает заметно более быстрого уменьшения значения функции по сравнению с градиентным спуском. Уже на начальных итерациях видно резкое снижение значения функции при использовании Ньютона; градиентный спуск снижает её плавнее и медленнее.

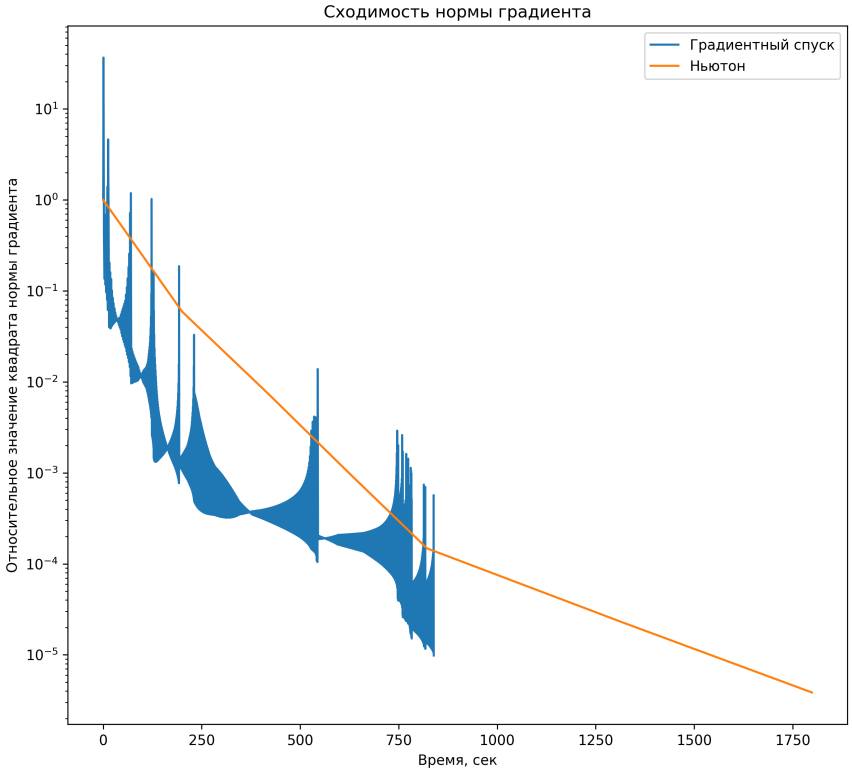
(b) *Сходимость нормы градиента:*
Метод Ньютона демонстрирует экспоненциальное уменьшение нормы градиента и достигает высокой точности быстрее, чем градиентный спуск, у которого норма градиента убывает заметно медленнее.

Вывод: метод Ньютона значительно эффективнее на данном датасете и по скорости уменьшения функции, и по скорости сходимости градиента.

Набор данных gisette



(a) Зависимость значения функции от времени



(b) Сходимость нормы градиента

Рис. 2: Методы градиентного спуска и Ньютона на датасете `gisette`

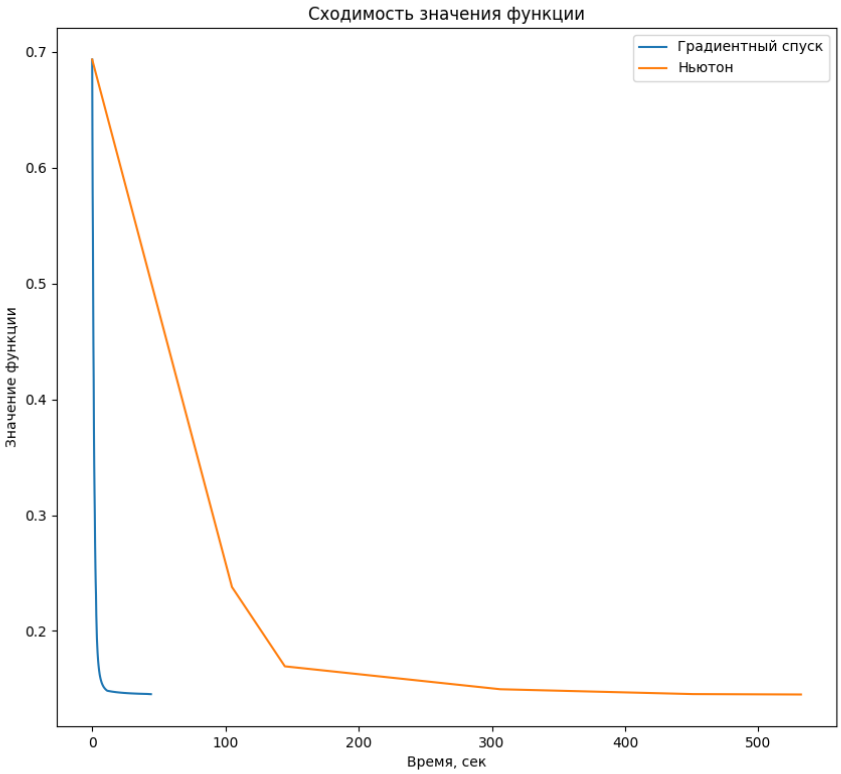
Комментарии (gisette):

(a) *Зависимость значения функции от времени:*
Метод Ньютона демонстрирует более медленный спад значения функции на ранних этапах, однако вскоре догоняет градиентный спуск по значению функции.

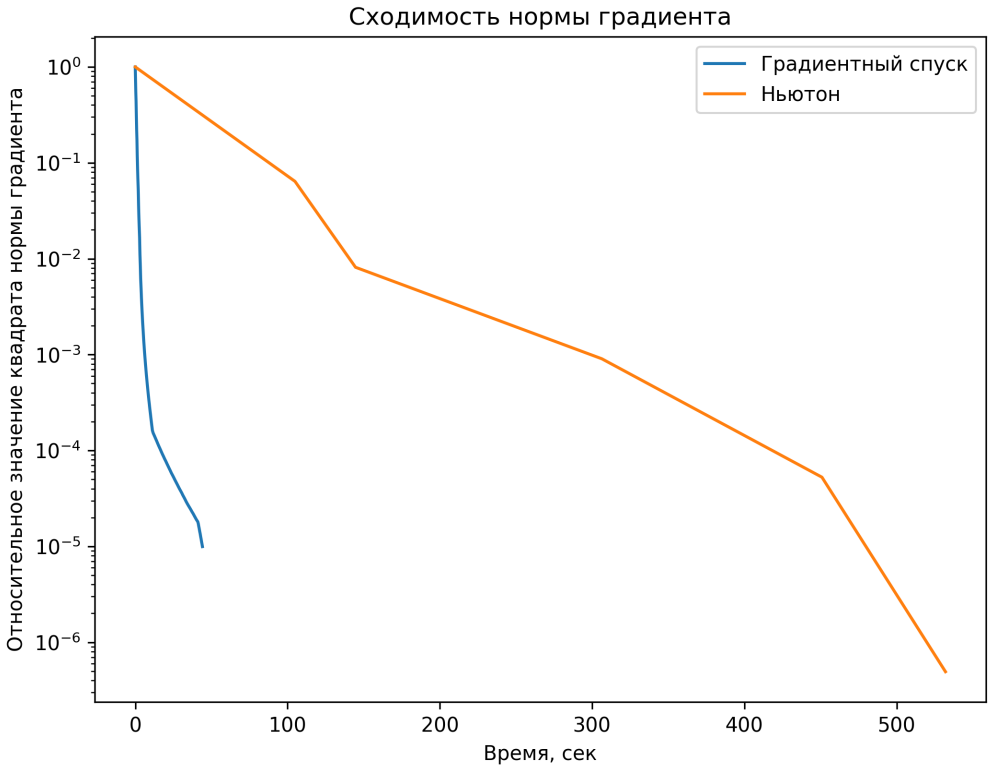
(b) *Сходимость нормы градиента:*
Метод Ньютона демонстрирует плавное убывание нормы градиента. Градиентный спуск показывает более выраженные колебания, но более быструю сходимость.

Вывод: на датасете `gisette` градиентный спуск оказывается эффективнее по норме градиента, однако по значению функции оба метода дают сравнимые результаты при более долгом времени обучения.

Набор данных real-sim



(a) Зависимость значения функции от времени



(b) Сходимость нормы градиента

Рис. 3: Методы градиентного спуска и Ньютона на датасете `real-sim`

Комментарии (real-sim):

(a) *Зависимость значения функции от времени:*
И здесь градиентный спуск показывает более резкое падение значения функции в начале. Однако возможно, что после некоторого числа итераций и метод Ньютона достигает сравнимого уровня, хотя и за большее время.

(b) *Сходимость нормы градиента:*
В норме градиента преимущество градиентного проявляется достаточно отчётливо: он убывает быстрее, в то время как у метода Ньютона наблюдается более медленное снижение.

Вывод: на датасете `real-sim` вновь подтверждается, что градиентный спуск достигает высокой точности за меньшее число итераций, тогда как метод Ньютона требует большего числа итераций.

Заключение

На небольшом объеме данных метод Ньютона быстро снижает значение функции и норму градиента, достигая требуемой точности за значительно меньшее число итераций по сравнению с градиентным спуском. Но так как каждая итерация метода Ньютона требует вычислительную сложность $O(mn^2 + n^3)$, метод оказывается существенно дороже метода градиентного спуска со сложностью итерации $O(mn)$. Для больших наборов данных (например, real-sim) и, в частности, при очень разреженных матрицах градиентный спуск остаётся предпочтительным выбором.

2 Эксперимент 3.4. Стратегия выбора длины шага в градиентном спуске

Описание эксперимента

В данном эксперименте изучается влияние различных стратегий подбора длины шага при классическом методе градиентного спуска на две задачи:

- **Квадратичная функция**

$$f(x) = \tfrac{1}{2}\, x^T A x - b^T x, \quad A = \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

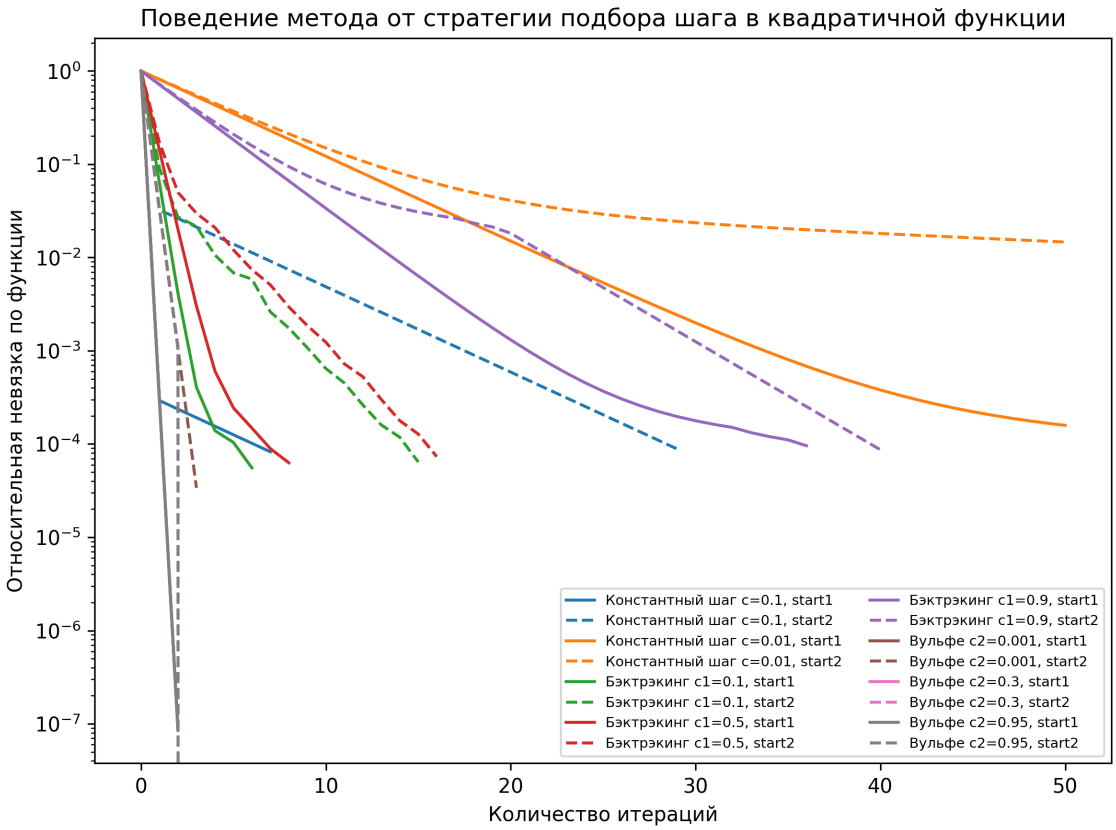
  Начальные точки: $start1_1 = (100,\, -5)^T,\ start2_1 = (-3,\, 3)^T$.

- **Логистическая регрессия** на синтетических данных:
  - $n_{\text{samples}} = 500,\ n_{\text{features}} = 10$.
  - Истинный вектор весов true_w $\sim \mathcal{N}(0, I)$.
  - Метка $y_i \in \{-1, +1\}$ генерируется по модели логистического регрессора с коэффициентом регуляризации $\lambda = 0.1$.
  - Начальные точки: $start1_2 = \mathbf{1},\ start2_2 \sim U[0, 100]^{10}$.

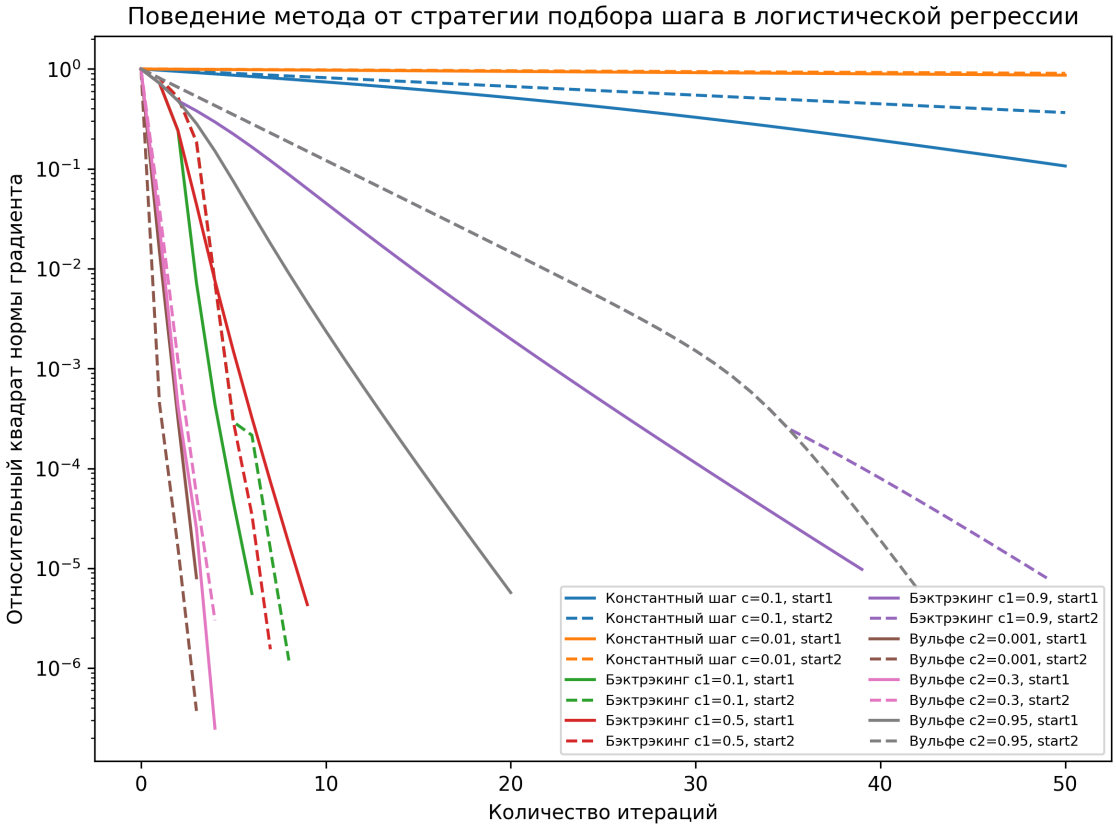
- **Стратегии подбора шага:**
  1. *Постоянный шаг*: $c = 0.1$, $c = 0.01$.
  2. *Бэктрэкинг (Armijo)*: $c_1 = 0.2,\ 0.5,\ 0.8$.
  3. *Условия Вульфе (Wolfe)*: $c_1 = 10^{-4}$, $c_2 = 0.001,\ 0.3,\ 0.95$.

- **Общие настройки:**
  - Максимум итераций: max_iter $= 50$.
  - Сбор истории (`trace=True`): $\{f(x_k)\}$ и $\{\|\nabla f(x_k)\|\}$.
  - Для квадратичной функции строится *относительная невязка по функции* $(f(x_k) - f^*)/(f(x_0) - f^*)$ в логарифмической шкале.
  - Для лог-регрессии строится *относительный квадрат нормы градиента* $\|\nabla f(x_k)\|^2/\|\nabla f(x_0)\|^2$ в логарифмической шкале.
  - Глобальный seed генератора случайных чисел: `np.random.seed(42)`;

Результаты эксперимента

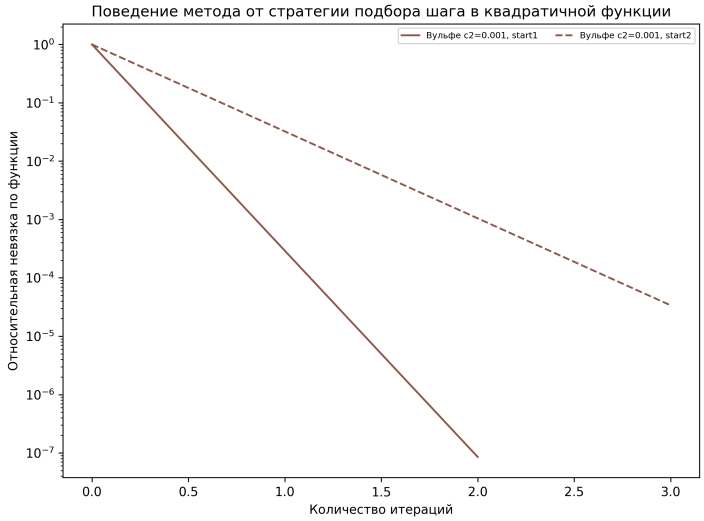


(a) Квадратичная функция: относительная невязка по функции

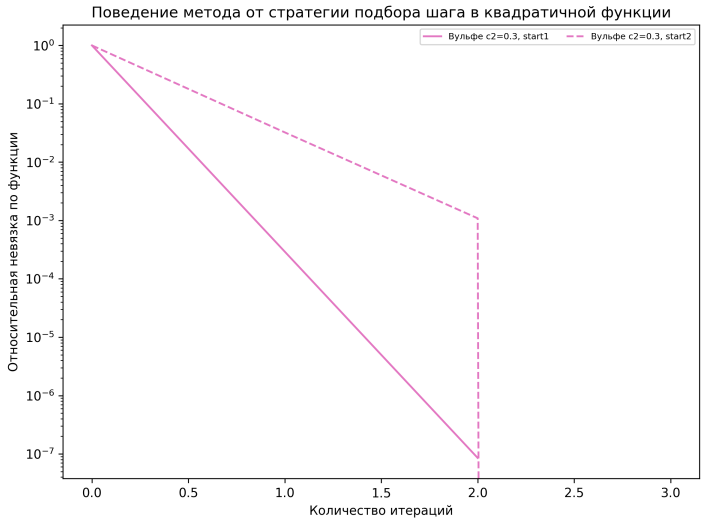


(b) Логистическая регрессия: относительный квадрат нормы градиента

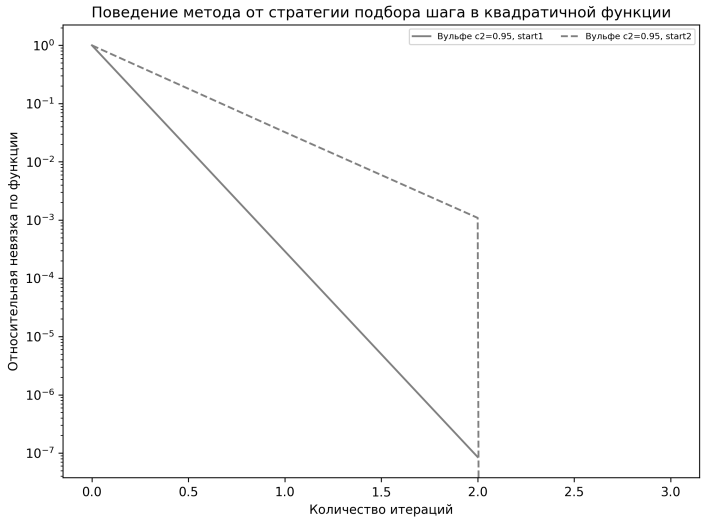
Рис. 4: Сходимость метода градиентного спуска при разных стратегиях подбора шага



(a) Вульфе, $c_2 = 0.001$



(b) Вульфе, $c_2 = 0.3$



(c) Вульфе, $c_2 = 0.95$

Рис. 5: Сходимость метода градиентного спуска для квадратичной функции при разных значениях параметра c_2 в стратегии Вульфе

Комментарии к графику Рис. 4(a). Квадратичная функция: относительная невязка по функции:

1. **Константный шаг**
   - $c = 0.1$:
     - Для точки $start1\_1$ достигается линейная сходимость за 7 итераций. Относительная невязка по функции достигает порядка 10^{-4}.
     - Для точки $start2\_1$ — за 29 итераций.
   - $c = 0.01$: не наблюдается линейной сходимости ни с $start1\_1$, ни с $start2\_1$.
2. **Бэктрекинг (условие Армихо)**
   - Точка $start1\_1$:
     - $c_1 = 0.1$ — 6 итераций;
     - $c_1 = 0.5$ — 8 итераций;
     - $c_1 = 0.9$ — 36 итераций.
   - Точка $start2\_1$:
     - $c_1 = 0.1$ — 15 итераций;
     - $c_1 = 0.5$ — 16 итераций;
     - $c_1 = 0.9$ — 40 итераций.
   - Наилучшая скорость сходимости достигается при $c_1 = 0.1$.
   - Во всех трёх случаях относительная невязка по функции достигает порядка 10^{-4}.
3. **Условия Вульфа**
   - При всех значениях c_2 (0.001, 0.3, 0.95) метод сходится за 2 итерации от точки $start1\_1$ и за 3 итерации от точки $start2\_1$.
   - Относительная невязка по функции достигает порядка 10^{-8} для c_2 (0.3, 0.95) для точки $start1\_1$, порядка 10^{-7} для c_2 (0.001, 0.3, 0.95) для точки $start2\_1$ и порядка 10^{-4} для c_2 (0.001) для точки $start1\_1$.
   - Примечание: более отчётливо траектории можно наблюдать на Рис. 5, поскольку на Рис. 11(a) они накладываются друг на друга.

Комментарии к графику Рис. 4(b). Логистическая регрессия: относительный квадрат нормы градиента:

1. **Константный шаг**
   - $c = 0.01$ и $c = 0.1$: не наблюдается линейной сходимости ни с $start2\_1$, ни с $start2\_2$; после 50 итераций относительный квадрат нормы градиента остаётся выше или порядка 10^{-1}.
2. **Бэктрекинг (условие Армихо)**

- Точка $start1\_2$:
  - $c_1 = 0.1$ — 6 итераций;
  - $c_1 = 0.5$ — 9 итераций;
  - $c_1 = 0.9$ — 39 итераций.
- Точка $start2\_2$:
  - $c_1 = 0.1$ — 8 итераций;
  - $c_1 = 0.5$ — 7 итераций;
  - $c_1 = 0.9$ — 49 итераций.
- Наилучшая скорость сходимости достигается при $c_1 = 0.1$.
- Во всех трёх случаях относительный квадрат нормы градиента остаётся порядка 10^{-5}.

3. **Условия Вульфа**

- Точка $start1\_2$:
  - $c_2 = 0.001$ — 3 итераций;
  - $c_2 = 0.3$ — 4 итераций;
  - $c_2 = 0.95$ — 20 итераций.
- Точка $start2\_2$:
  - $c_2 = 0.001$ — 3 итераций;
  - $c_2 = 0.3$ — 4 итераций;
  - $c_2 = 0.95$ — 42 итераций.
- Наилучшая скорость сходимости достигается при $c_1 = 0.001$.
- Во всех случаях, кроме $c_2 = 0.3$ для точки $start1\_2$ и $c_2 = 0.01$ для точки $start2\_2$, относительный квадрат нормы градиента остаётся порядка 10^{-5}. При $c_2 = 0.3$ для точки $start1\_2$ и при $c_2 = 0.01$ для точки $start2\_2$ достигает порядка 10^{-6}.

Выводы

1. **Константный шаг**.

- При $c = 0.01$ наблюдается очень медленная сходимость как в задачах логистической регрессии, так и в задачах с квадратичной функцией;
- в квадратичной функции $c = 0.1$ показывает приемлемый результат, тогда как в логистической регрессии метод не сходится.

2. **Бэктрэкинг**. При использовании параметров $c_2 = 0.1$ и $c_2 = 0.5$ данный метод дает результат, близкий по качеству к методу Вульфа в задачах логистической регрессии, но требует чуть больше итераций (6–9 итераций).

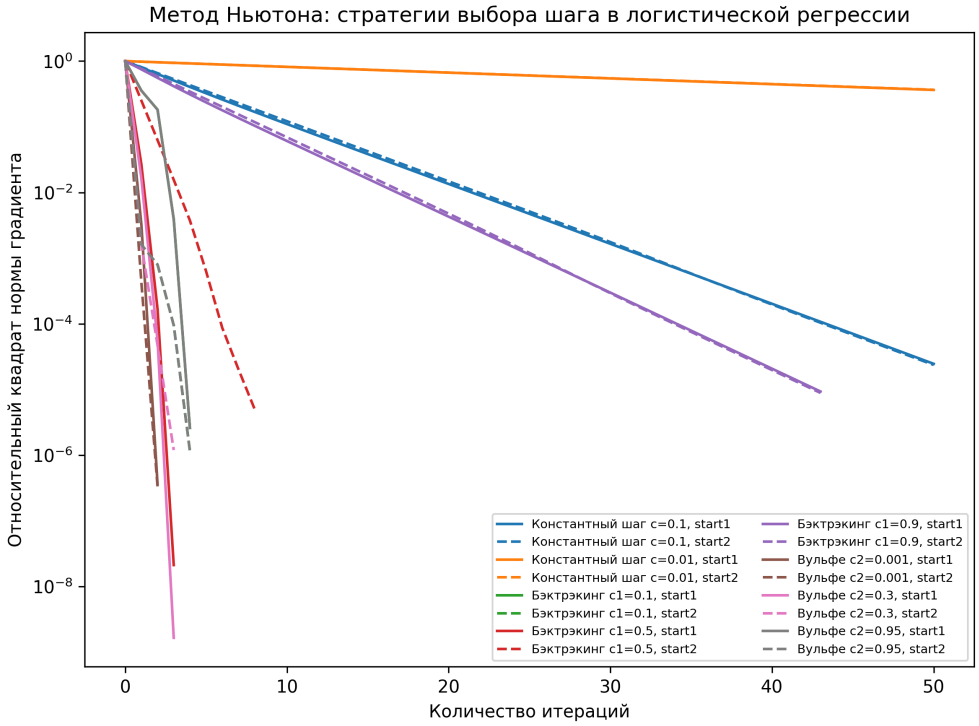
3. **Условия Вульфа (Wolfe)**. Эти условия обеспечивают наиболее быструю и устойчивую сходимость в квадратичной функции для обеих стартовых точек, а в задаче логистической регрессии результаты сопоставимы с результатами метода бэктрэкинга.

Заключение. Константный шаг не рекомендуется использовать в рассматриваемых задачах. Для квадратичной функции оптимальным оказывается метод Вульфа при любых начальных точках и c_2. В задаче логистической регрессии методы бэктрэкинга и Вульфа показывают сопоставимые результаты, но при более точной настройке параметров метод Вульфа может обеспечить более высокую точность, в то время как метод бэктрэкинга может давать плохие результаты при неудачно выбранном c_1.

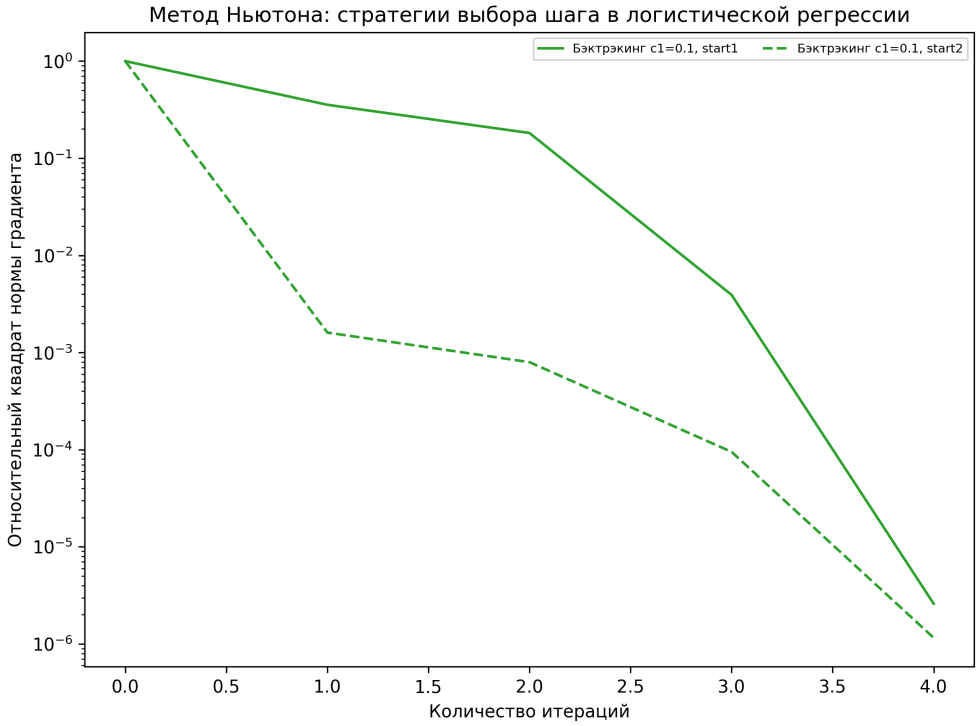
3 Эксперимент 3.5. Стратегия выбора длины шага в методе Ньютона

Описание эксперимента В данном эксперименте изучается влияние различных стратегий подбора длины шага при методе Ньютона в задаче логистической регрессии. В эксперименте были использованы параметры аналогичные эксперименту 3.4.

Результаты эксперимента



(a) Общий график



(b) С выделенной траекторией Бэктрэкинг, c1 = 0.1

Рис. 6: Метод Ньютона в задаче логистической регрессии: относительный квадрат нормы градиента

Комментарии к графику Рис. 6. Метод Ньютона: стратегии выбора шага в логистической регрессии:

1. **Константный шаг**
   - $c = 0.01$ и $c = 0.1$: не наблюдается линейной сходимости ни с $start2\_1$, ни с $start2\_2$; после 50 итераций относительный квадрат нормы градиента остаётся порядка 10^{-1}, при $c = 0.01$ порядка 10^{-4} .
2. **Бэктрекинг (условие Армихо)**
   - Точка $start1\_2$:
     - $c_1 = 0.1$ — 4 итераций (можно проследить на рис.6(b));
     - $c_1 = 0.5$ — 3 итераций;
     - $c_1 = 0.9$ — 43 итераций.
   - Точка $start2\_2$:
     - $c_1 = 0.1$ — 4 итераций (можно проследить на рис.6(b));
     - $c_1 = 0.5$ — 8 итераций;
     - $c_1 = 0.9$ — 43 итераций.
   - Наилучшая скорость сходимости достигается при $c_1 = 0.1$.
   - Наименьшее значение относительного квадрата нормы градиента достигается при $c_1 = 0.5$ для точки $start1\_2$.
3. **Условия Вульфа**
   - Точка $start1\_2$:
     - $c_2 = 0.001$ — 2 итераций;
     - $c_2 = 0.3$ — 3 итераций;
     - $c_2 = 0.95$ — 4 итераций.
   - Точка $start2\_2$:
     - $c_2 = 0.001$ — 2 итераций;
     - $c_2 = 0.3$ — 3 итераций;
     - $c_2 = 0.95$ — 4 итераций.
   - Наилучшая скорость сходимости достигается при $c_1 = 0.001$.
   - Наименьшее значение относительного квадрата нормы градиента достигается при $c_1 = 0.3$ для точки $start1\_2$.

Выводы

Константный шаг не рекомендуется использовать в задаче логистической регрессии, так как даже при применении метода Ньютона не наблюдается сходимость. Методы Вульфа и бэктрэкинга демонстрируют стабильные результаты, за исключением случая бэктрэкинга при $c_1 = 0.9$; для данной задачи рекомендуется выбирать $c_1 < 0.5$.

Метод Вульфа показывает наилучший результат: в точке $start1\_2$ при $c_2 = 0.001$ относительный квадрат нормы градиента достигает порядка 10^{-8}, а при $c_2 = 0.95$ и $c_2 = 0.1$ — порядка 10^{-6}. Таким образом, для достижения более высокой точности рекомендуется использовать метод Вульфа, поскольку при более точной настройке параметров он демонстрирует лучшие результаты, в то время как метод бэктрэкинга может давать плохие результаты при неудачно выбранном c_1.