Национальный исследовательский университет
«Высшая школа экономики»
Санкт-Петербургская школа физико-математических и компьютерных наук

Отчет к лабораторной работе № 1
«Методы градиентного спуска и метод Ньютона»

Выполнил:
студент группы МОАД
Сморчков Данил Дмитриевич

Санкт-Петербург
2024 год

Оглавление

1. Траектория градиентного спуска на квадратичной функции

В данном эксперименте необходимо проанализировать траекторию градиентного спуска для нескольких квадратичных функций, различных стратегий линейного поиска и выбора начальной точки. Квадратичная функция задается следующим образом:

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle, \qquad \text{где } A \in \mathbb{S}^n_{++}, b \in \mathbb{R}^n$$

Здесь и во всех следующих экспериментах используется критерий остановки:

$$\left||\nabla f(x_k)|\right|_2^2 \leq \varepsilon \left||\nabla f(x_0)|\right|_2^2,$$

где в данном эксперименте $\varepsilon = 10^{-9}$.

1. Сначала рассмотрим, как на градиентный спуск влияют различные стратегии линейного поиска. В качестве квадратичной функции возьмем хорошо обусловленную функцию с параметрами:

$$A = \begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

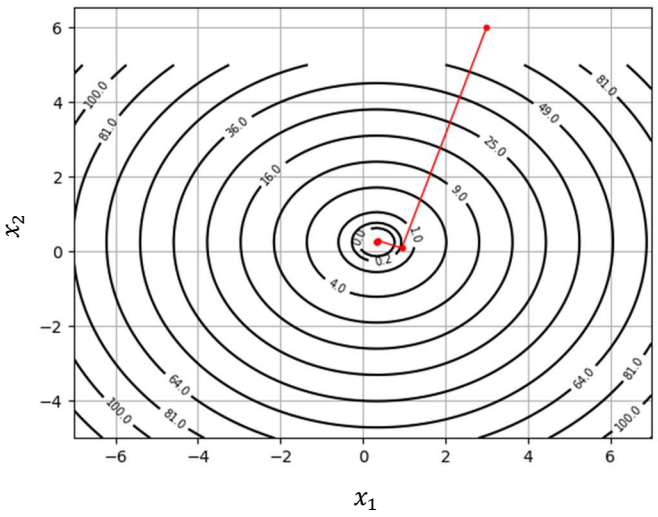
т.е. в данном случае число обусловленности равно $k = \frac{4}{3}$.

Начальную точку зададим следующим образом:

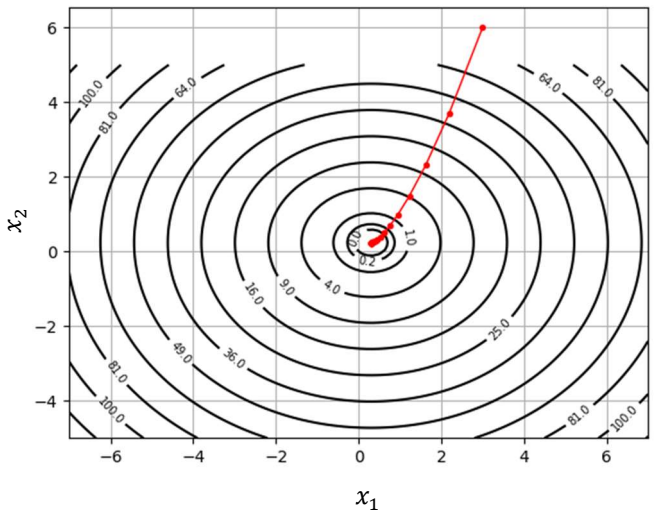
$$x_0 = \begin{pmatrix} 3 \\ 6 \end{pmatrix}$$

Результаты эксперимента:

Вульф $c_1 = 10^{-4}, c_2 = 0{,}9$



Константа $c = 0{,}1$



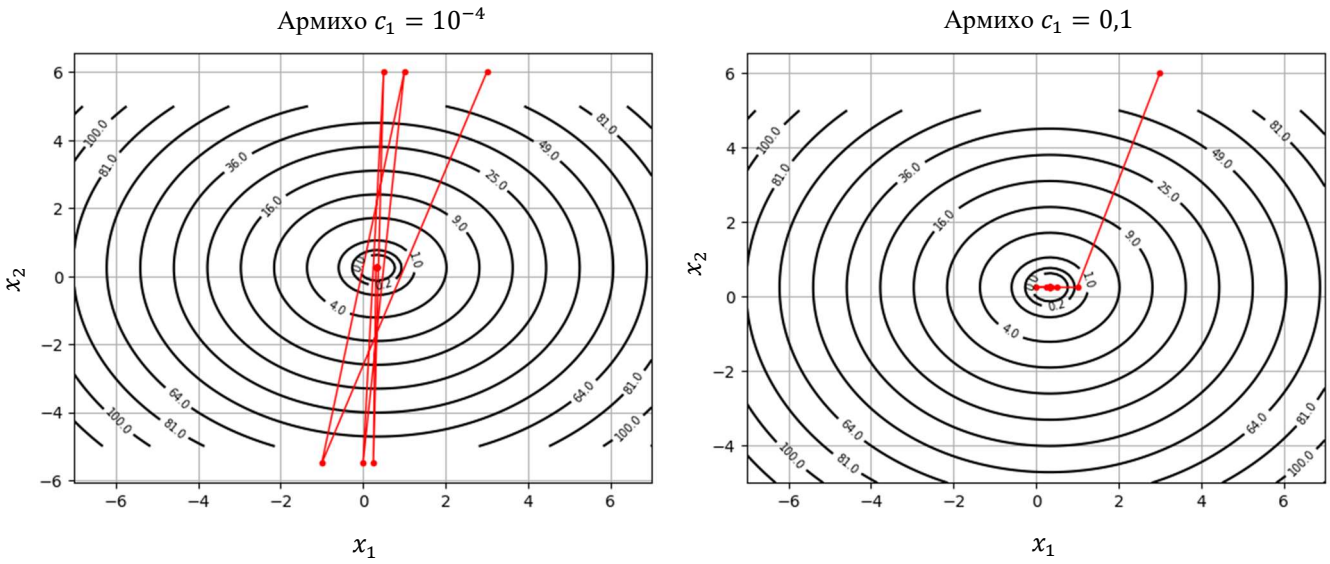


Рис. 1 Траектория градиентного спуска для хорошо обусловленной квадратичной функции с разными подходами линейного поиска

Мы видим, что быстрее всего (по итерациям) сходится GD с использованием стратегии Вульфа, далее по скорости идет стратегия Армихо, и самая медленная сходимость – у константной стратегии. Еще хочу заметить, что в зависимости от параметра c_1 GD с Армихо может вести себя очень по-разному: в данном случае при очень маленьком c_1 точка прыгает с одной стороны функции на другую (при этом, конечно, значение функции уменьшается), выглядит это, как будто никакого спуска не происходит (а может случаться такое, что функция уменьшается, а градиент увеличивается, именно эта ситуация произошла у меня в эксперименте 2, и я потратил какое-то время, чтобы разобраться что происходит). Аналогичная история с константным шагом, только если в случае со стратегией Армихо GD все равно сойдется (в случае данной функции), то с константой он может и вовсе не сходиться (так, например, было для $c = 1$).

2. В случае хорошо обусловленной функции начальная точка имеет значение, однако во всех точках с данным радиусом поведение GD будет примерно одинаковым, то есть в основном зависимость от длины от x_0 до x^*. В случае же плохо обусловленной функции есть большая разница откуда именно начинается GD. Это и будет проиллюстрировано во 2 части.

В качестве квадратичной функции возьмем плохо обусловленную функцию с параметрами:

$$A = \begin{pmatrix} 3 & 0 \\ 0 & 40 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

т.е. в данном случае число обусловленности равно $k = \frac{40}{3}$.

Рассмотрим поведение GD для двух разных начальных точек и метода Вульфа (как самого стабильного из имеющихся):

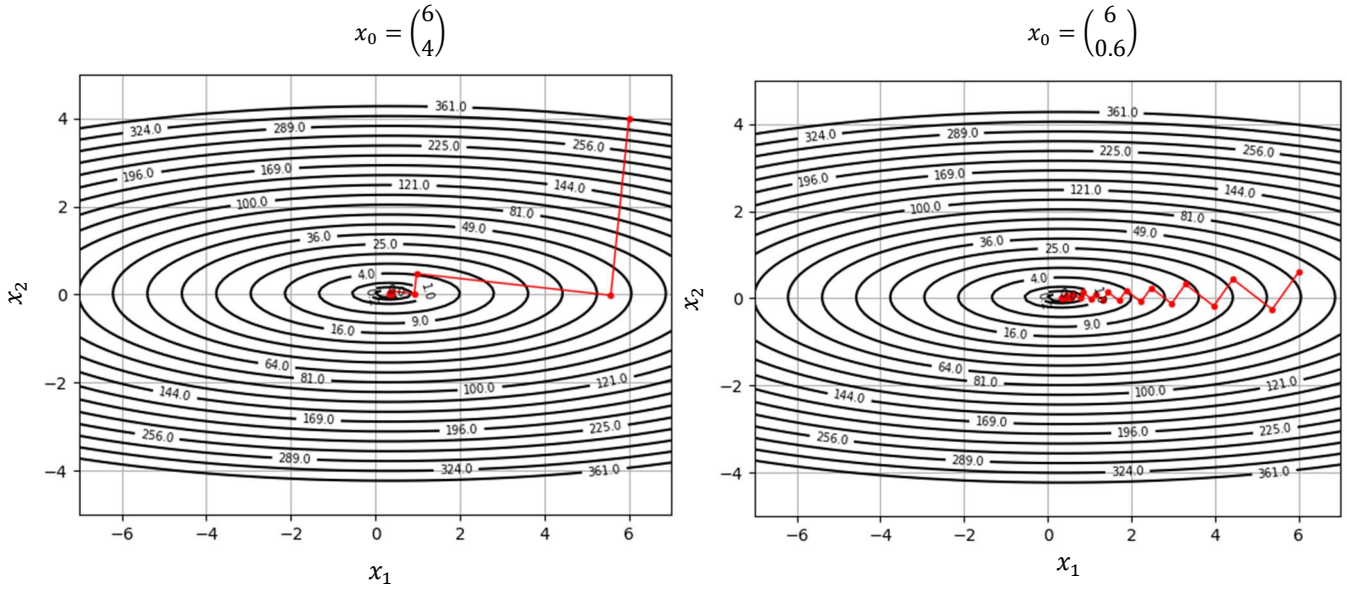


Рис. 2 Траектория градиентного спуска для плохо обусловленной квадратичной функции с разными начальными точками и стратегий линейного поиска Вульфа

Из данных графиков видим, что в случае плохо обусловленной функции точка инициализации метода играет большую роль: так на первом графике точка x_{01} находится дальше, точка на втором графике x_{02}, и аналогичное соотношение для функций $f(x_{01})$ и $f(x_{02})$, Однако в первом случае функция сходится быстрее, из-за того, что во втором случае инициализация произошла внутри «каньона» и на каждой итерации точка «прыгает» по стенкам «каньона», тем самым двигаясь неоптимальным способом. Стоит еще раз подчеркнуть, что данный эксперимент проводился с использованием стратегии Вульфа, который показывает самую стабильную сходимость GD, таким образом, на стратегиях Армихо или константы выбор начальной точки может еще больше повлиять на сходимость.

Зависимость от числа обусловленности описана в следующем эксперменте.

2. Зависимость числа итераций градиентного спуска от числа обусловленности и размерности пространства

В данном эксперименте необходимо исследовать, как зависит число итераций, необходимое градиентному спуску для сходимости, от следующих двух параметров: 1) числа обусловленности $\kappa \geq 1$ оптимизируемой функции и 2) размерности пространства n оптимизируемых переменных.

Эксперимент проводился следующим образом:

Для пространств размерности $n \in \{2, 10, 100, 1000, 10000\}$, генерируется матрица с числом обусловленности k. Для этого сначала определяем диагональ как:

$$a_{11} = 1, \qquad a_{nn} = k, \qquad a_{ii} = randint(1, k),$$

затем создаем диагональную матрицу с данной диагональю и делим её на k для нормировки. Вектор b задаем как случайный вектор из стандартного нормального многомерного распределения размерности n. Ну и начальную точку x_0 задаем как случайный вектор из многомерного нормального распределения с вектором средних значений $\mu = \begin{pmatrix} 13 \\ \vdots \\ 13 \end{pmatrix}$ и матрицей ковариаций $\Sigma = \begin{pmatrix} 3^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 3^2 \end{pmatrix}$.

Из-за того, что матрица и начальная точка генерируются случайным образом, для каждого n и k эксперимент проводится несколько раз и усредняется.

Точность в данном эксперименте $\varepsilon = 10^{-9}$.

1. Вульф $c_1 = 10^{-4}, c_2 = 0.9$, Количество повторений – 20, на графике – 3 из 20.

Рис 3. Число итераций GD для стратегии Вульфа ($c_1 = 10^{-4}, c_2 = 0.9$)

Для лучшей визуализации усредним все графики (20 штук) одного цвета:

Рис 4. Среднее число итераций GD для стратегии Вульфа ($c_1 = 10^{-4}, c_2 = 0.9$)

2. Армихо $c_1 = 0.7$, Количество повторений – 15, на графике – сразу усреднение по 15 повторам.

Рис 5. Среднее число итераций GD для стратегии Армихо ($c_1 = 0.7$)

3. Константная стратегия $c = 0.5$. Количество повторений – 15, на графике – сразу усреднение по 15 повторам.

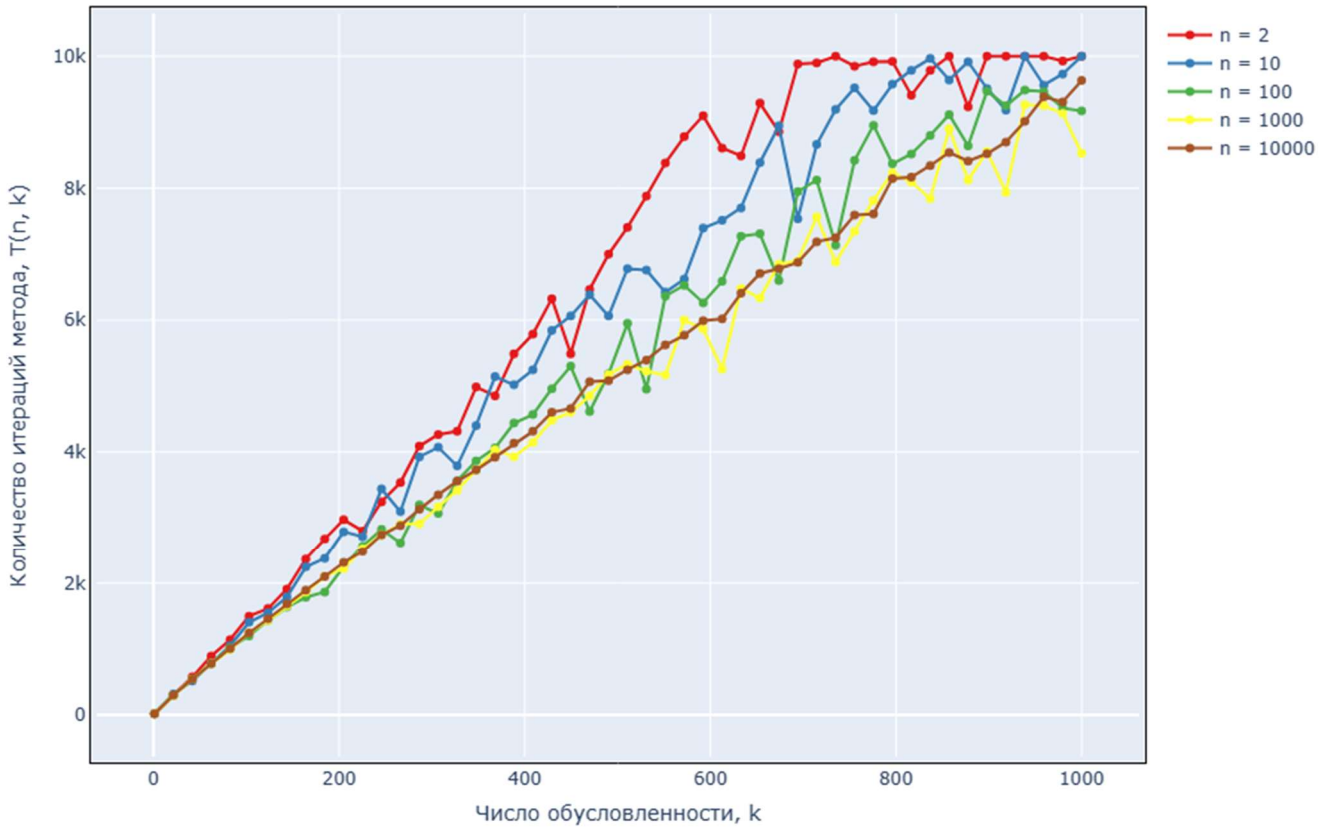


Рис 6. Среднее число итераций GD для константной стратегии ($c = 0.5$)

Из графиков видно, что для размерностей $n > 2$, с увеличением числа обусловленности происходит увеличение количество итераций метода для всех стратегий линейного поиска. Наилучшие результаты были получены с помощью выполнения условий Вульфа, на втором месте – Армихо, хуже всех – константа.

Стоит также отметить, что для адаптивных методов (Армихо и Вульф) задача с размерностью $n = 2$ выбивается из наблюдаемой зависимости количества итераций от числа обусловленности. Об этом стоит помнить, ведь чаще всего для визуализации работы методов многомерной оптимизации используют именно задачи размерности $n = 2$ (к примеру, прошлое упражнение), и полученные результаты могут быть не репрезентативными (не обобщаться на более высокие размерности).

Можно сделать заключение о влиянии размерности на скорость сходимости: именно на этих графиках – чем больше сходимость, тем меньше количество итераций, однако, я думаю, что это заявление ничем не подкреплено (возможно так случайно сгенерировалось). Другое наблюдение – Чем больше размерность, тем более явна линейная зависимость среднего числа итераций GD от числа обусловленности функции, и, в целом, градиентный спуск ведет себя более стабильно на больших размерностях (посмотреть хотя бы график 3).

И последнее, для стратегии Армихо пришлось подбирать константу c_1, так как при $c_1 = 10^{-4}$, получался такой график:

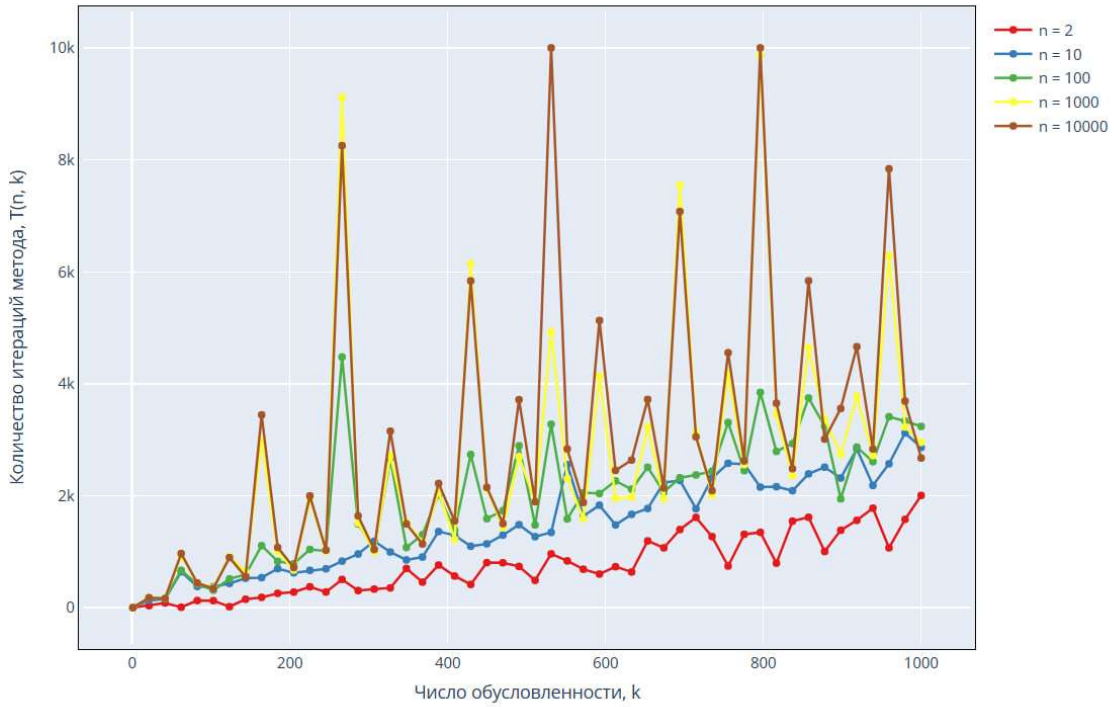


Рис 7. Среднее число итераций GD для стратегии Армихо ($c_1 = 10^{-4}$)

С чем именно связаны эти пики я не понял, но предположу, что из-за маленькой константы, когда точка попадала в такие области, где градиент большой (крутая поверхность), условию Армихо удовлетворяла противоположная точка, и получалось, что точка прыгает по квадратичной функции, практически не меняя значения. Поэтому, когда я наложил более сильные условия ($c_1 = 0.7$), пики исчезли. А это – еще одно преимущество стратегии Вульфа: он «из коробки» (без подбора констант) дает результат лучше, чем Армихо с дополнительной настройкой.

3. Сравнение методов градиентного спуска и Ньютона на реальной задаче логистической регрессии

В данном эксперименте необходимо сравнить методы градиентного спуска и Ньютона на задаче обучения логистической регрессии на реальных данных. В качестве реальных данных используйте следующие три набора с сайта LIBSVM3 w8a, gisette и real-sim.

Начнем с того, что представим функцию потерь (минимизируемую функцию) логистической регрессии, её градиент и гессиан в матричной форме:

Имеем:

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} \ln(1 + \exp(-b_i\langle a_i, x\rangle)) + \frac{\lambda}{2}\left|\left|x\right|\right|_2^2$$

Здесь совершенно нетрудно заметить, что для матричного вида вместо суммы нужно записать произведение единичного вектора на вектор, образованный при замене b_i и a_i на b и A:

$$f(x) = \frac{1}{m}1_m^T * \ln(1 + \exp(-b \odot Ax)) + \frac{\lambda}{2}\left|\left|x\right|\right|_2^2,$$

где $*$ – матричное произведение,

$\odot$ – Адамарово (поэлементное) произведение,

A – матрица наблюдения-признаки,

1_m – вектор-столбец из m единиц,

$\ln(1 + \exp(\ldots))$ применяются к вектору поэлементно.

Градиент такой функции можно найти, взяв дифференциал:

$$df(x) = -\frac{1}{m}1_m^T * \frac{1}{1 + \exp(-b \odot Ax)} \odot \exp(-b \odot Ax) \odot b \odot A * dx,$$

так как $\frac{\exp(x)}{1+\exp(x)} = \sigma(x)$ и $df(x) = \langle \nabla f(x), dx\rangle$, получаем

$$\boxed{\nabla f(x) = -\frac{1}{m}A^T * b \odot \sigma(b \odot Ax) + \lambda x}$$

Далее найдем гессиан. Действия аналогичные поиску градиента, однако нужно взять дифференциал от *фиксированного* первого дифференциала и получить формулу вида:

$$d^2 f(x) = \langle \nabla^2 f(x) dx_1, dx\rangle$$

Проведение вычислений я оставлю у себя на листочке, тем временем ответ получается такой:

$$\boxed{\nabla^2 f(x) = \frac{1}{m}A^T \Sigma A + \lambda I,}$$

где $\Sigma = diag(\sigma(b \odot A) \odot \sigma(-b \odot A))$.

Формулы получили, теперь запишем сложность по скорости и по памяти:

i. Градиентный спуск

Память. Для градиентного спуска необходимо хранить сам градиент $\nabla f(x_k)$, точку x_k, скаляр α_k и градиент с прошлой итерации для остановки $\nabla f(x_{k+1})$ (в моей реализации). Итого $O(n)$.

Скорость. Как мы видим из формулы градиента: самая затратная операцию – произведение матрицы на вектор за $O(nm)$ (2 раза), остальные – линия или константа. В самой итерации градиентного спуска есть еще линейных поиск, но он тоже работает за $O(n)$. Итого $O(nm)$.

ii. Метод Ньютона

Память. Для метода Ньютона необходимо хранить гессиан $\nabla^2 f(x)$ и остальное, как в методе GD. Следовательно, в этот раз $O(n^2)$.

Скорость. Для того, чтобы посчитать гессиан $\nabla^2 f(x)$ нужно перемножить 3 матрицы, что обходится с учетом размеров матриц за $O(nm^2)$. В самой итерации метода Ньютона необходимо решить систему уравнений, что потребует $O(n^3)$ операций. Суммируя, получаем $O(n^3) + O(nm^2)$.

Перейдем к экспериментам:

I. Датасет w8a. Размер датасета: (49749, 300). Для градиентного спуска точность $\varepsilon = 10^{-5}$, для метода Ньютона – $\varepsilon = 10^{-9}$. Для хорошего масштабирования. Рассмотрены три основные стратегии линейного поиска: Вульф, Армихо, Константа; все параметры по умолчанию.

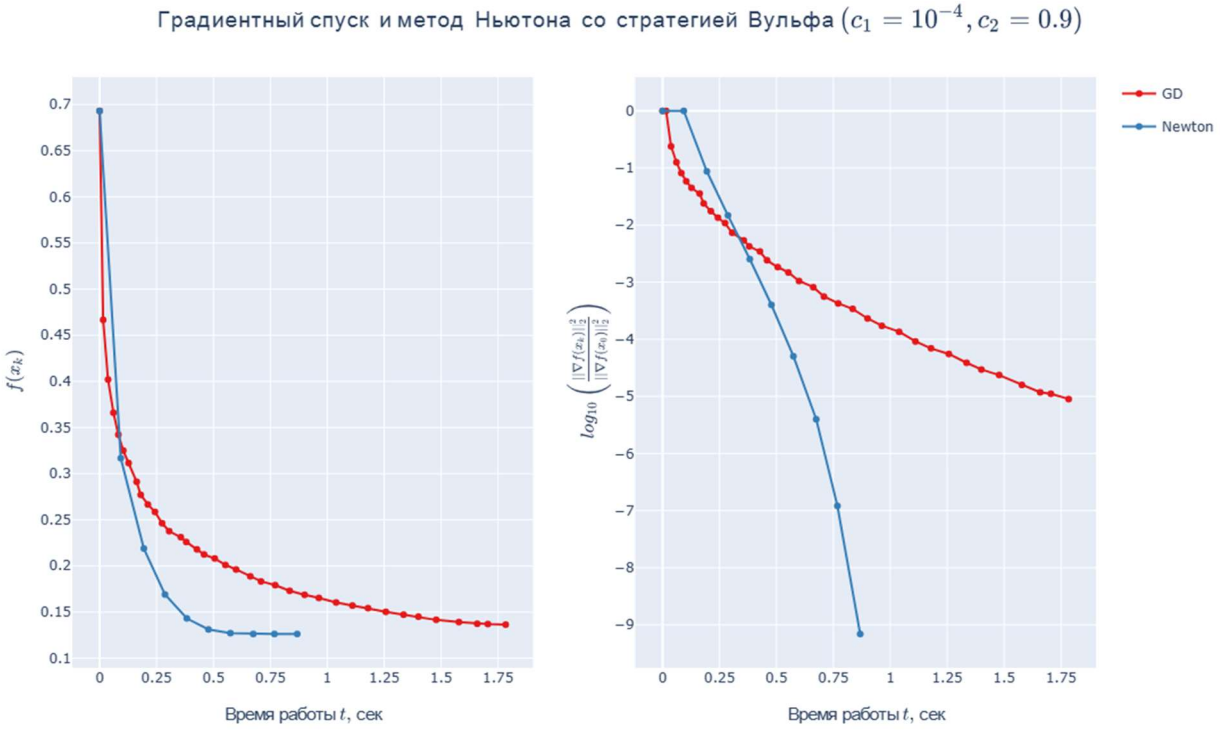


Рис 8. Градиентный спуск и метод Ньютона со стратегией Вульфа ($c_1 = 10^{-4}, c_2 = 0.9$) для датасета w8a

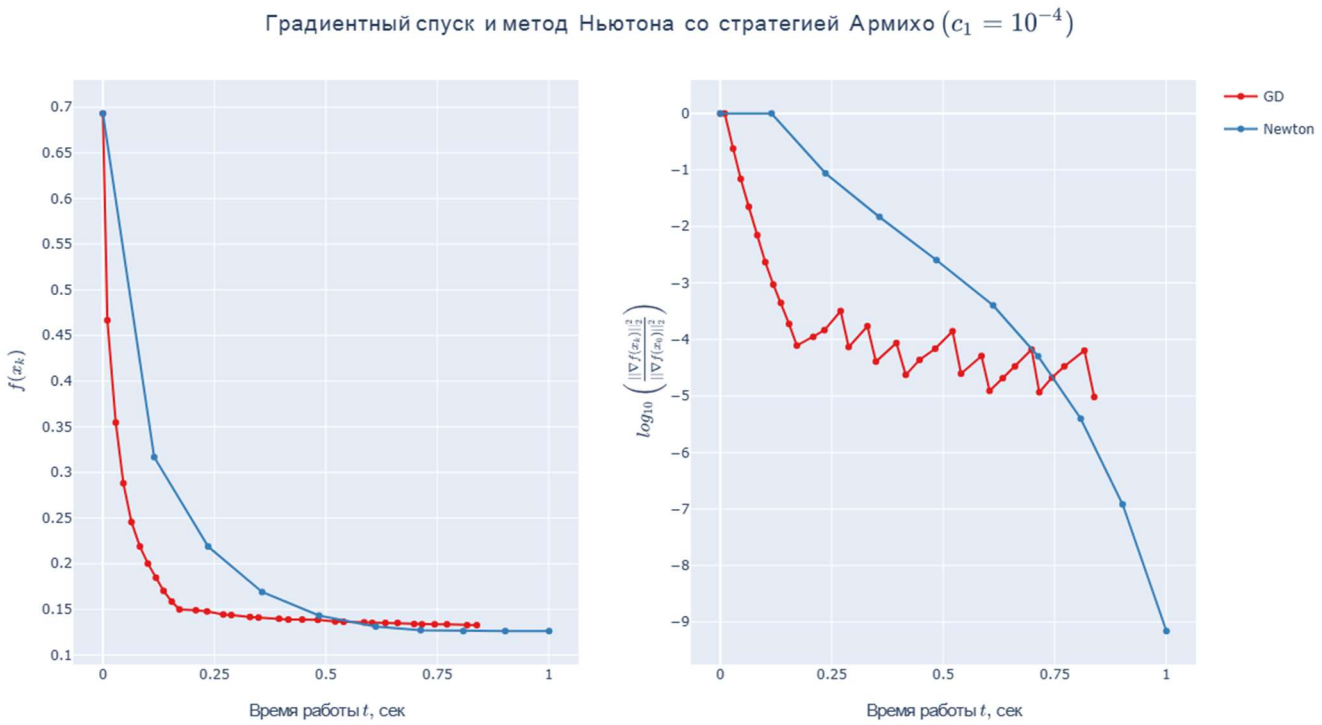


Рис 9. Градиентный спуск и метод Ньютона со стратегией Армихо ($c_1 = 10^{-4}$) для датасета w8a

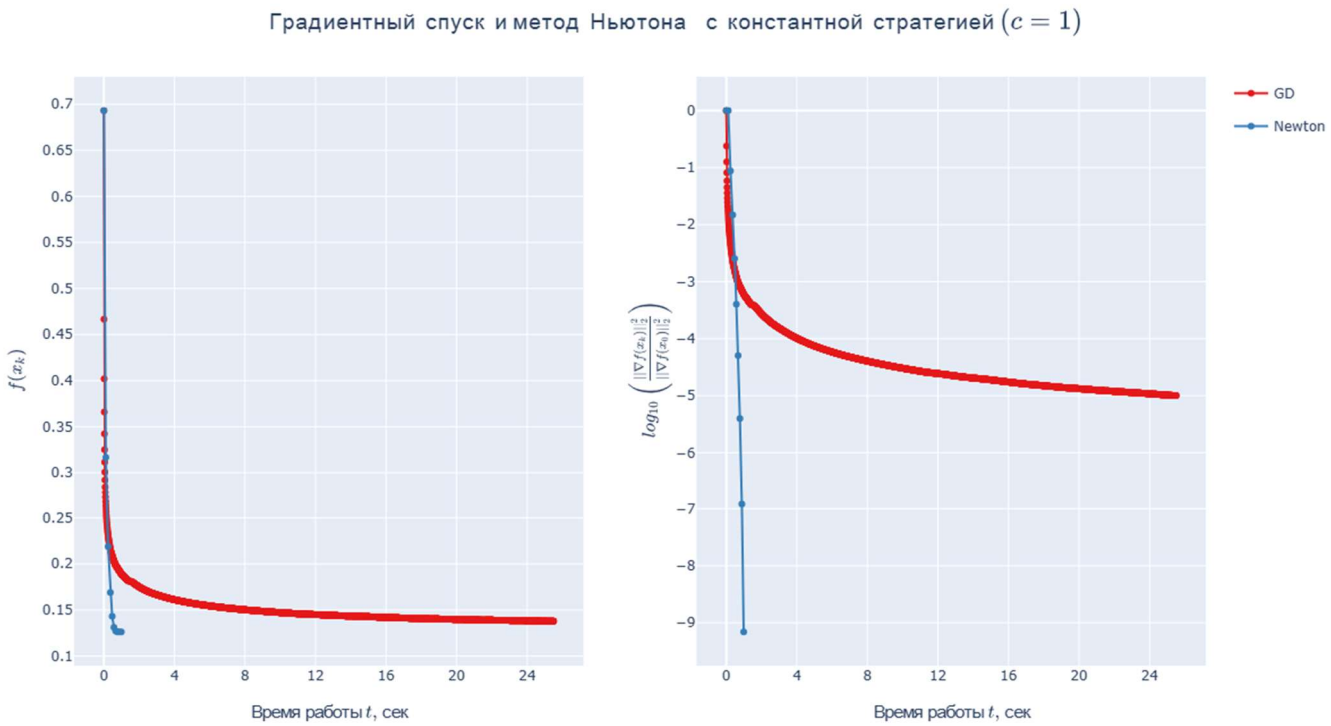


Рис 10. Градиентный спуск и метод Ньютона с константной стратегией (с = 1) для датасета w8a

На графиках показано, что для датасета w8a метод Ньютона сходится быстрее, чем градиентный спуск.

Если рассматривать стратегии линейного поиска, то (как можно было ожидать) стратегия Вульфа оказалась наиболее эффективной и стабильной, хуже всего – константа.

II. Датасет gissete. Размер датасета: (6000, 5000). Для градиентного спуска точность $\varepsilon = 10^{-5}$, для метода Ньютона – $\varepsilon = 10^{-9}$. Для хорошего масштабирования и приемлемого времени работы (хотя, если Вы посмотрите, то поймете, что оно и так не приемлемо:)). Рассмотрены три основные стратегии линейного поиска: Вульф, Армихо, Константа; все параметры по умолчанию.

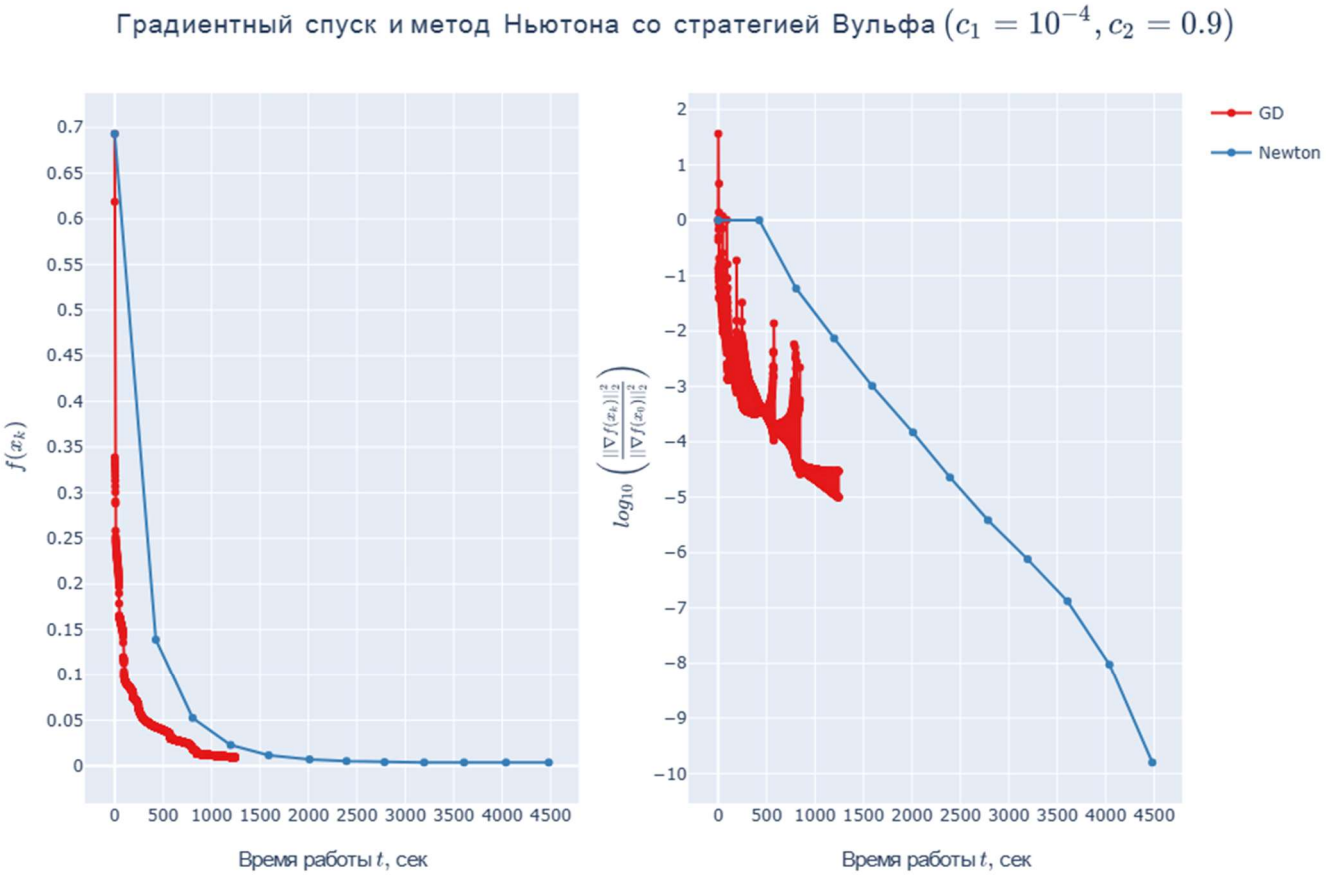


Рис 11. Градиентный спуск и метод Ньютона со стратегией Вульфа ($c_1 = 10^{-4}, c_2 = 0.9$) для датасета gissete

Рис 12. Градиентный спуск и метод Ньютона со стратегией Армихо ($c_1 = 10^{-4}$) для датасета gissete

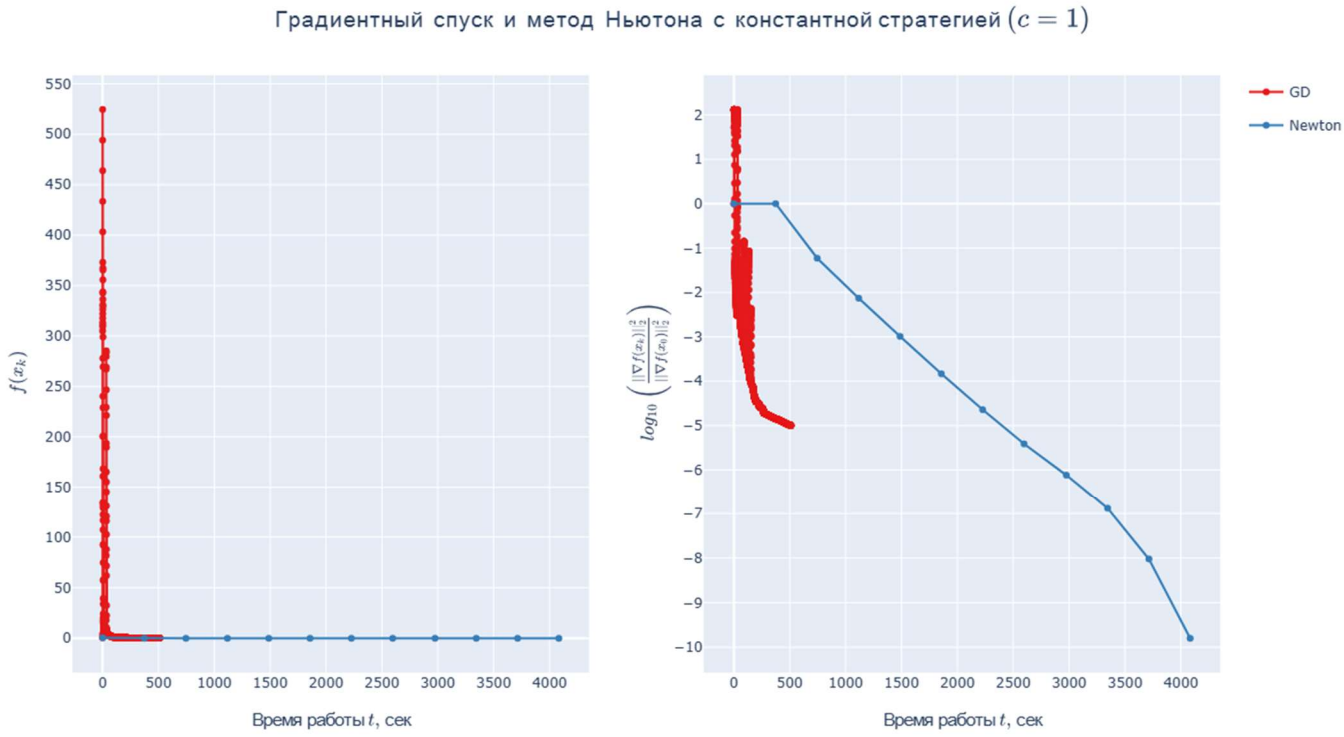


Рис 13. Градиентный спуск и метод Ньютона с константной стратегией (c = 1) для датасета gissete

Gissete достаточно большой датасет, на нем методы работает большое количество времени. Как и в прошлый раз метод Ньютона для точности $\varepsilon = 10^{-9}$ работает быстрее, чем градиентный спуск (если продолжить градиентный метод).

Для линейного поиска картина изменилась: лучше всего показал себя константный метод.

III. Датасет real-sim. Размер датасета: (72309, 20958). Данные разрежены. Для градиентного спуска точность $\varepsilon = 10^{-9}$, для метода Ньютона – $\varepsilon = 10^{-9}$ при выполнении условий Армиха и Вульфа. Для константной стратегии для градиентного спуска точность $\varepsilon = 10^{-5}$, для метода Ньютона – $\varepsilon = 10^{-9}$. Рассмотрены три основные стратегии линейного поиска: Вульф, Армихо, Константа; все параметры по умолчанию.

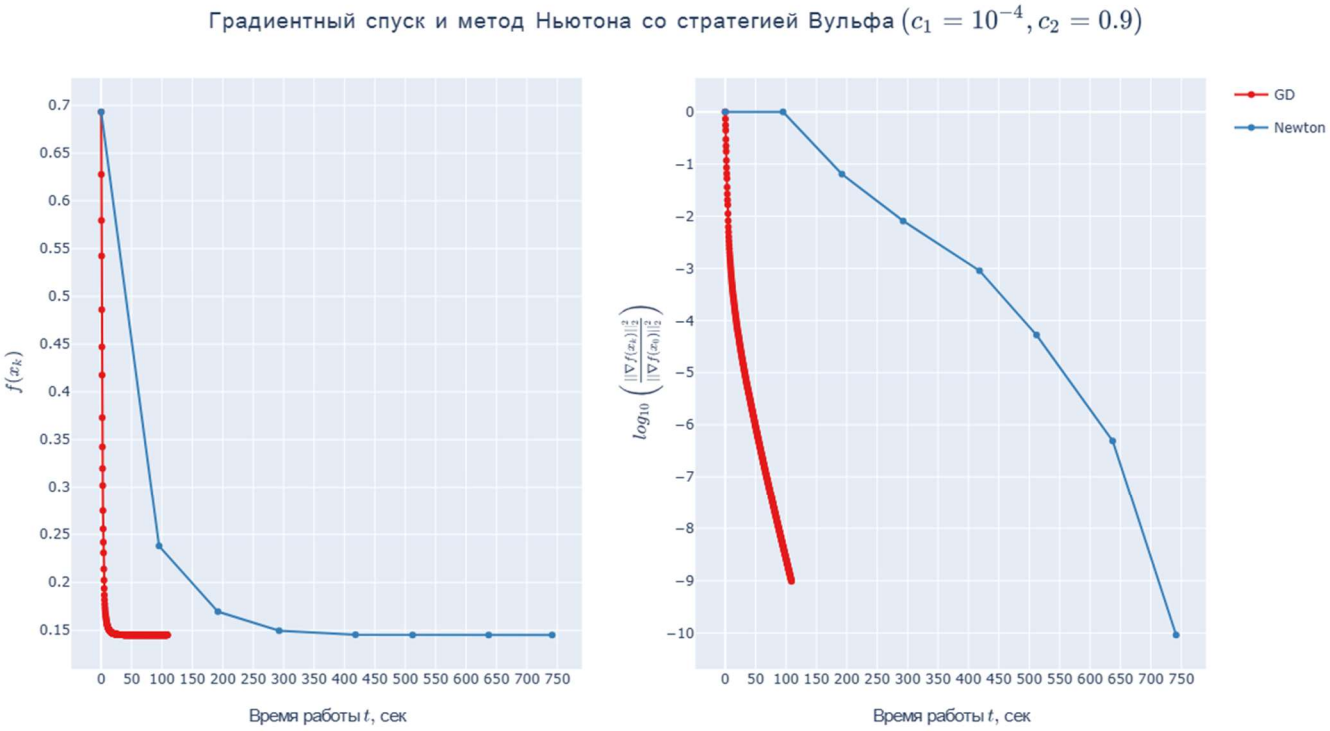


Рис 14. Градиентный спуск и метод Ньютона со стратегией Вульфа ($c_1 = 10^{-4}, c_2 = 0.9$) для датасета real-sim

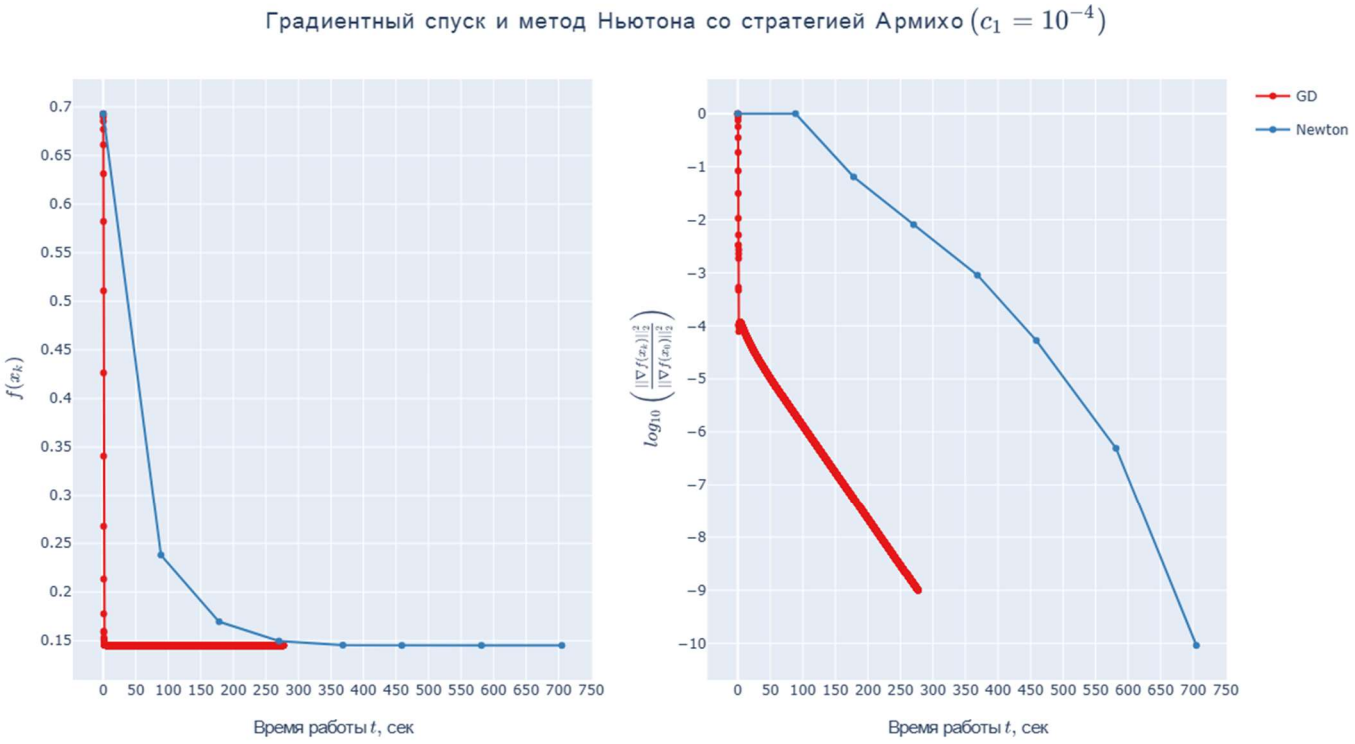


Рис 15. Градиентный спуск и метод Ньютона со стратегией Армихо ($c_1 = 10^{-4}$) для датасета real-sim

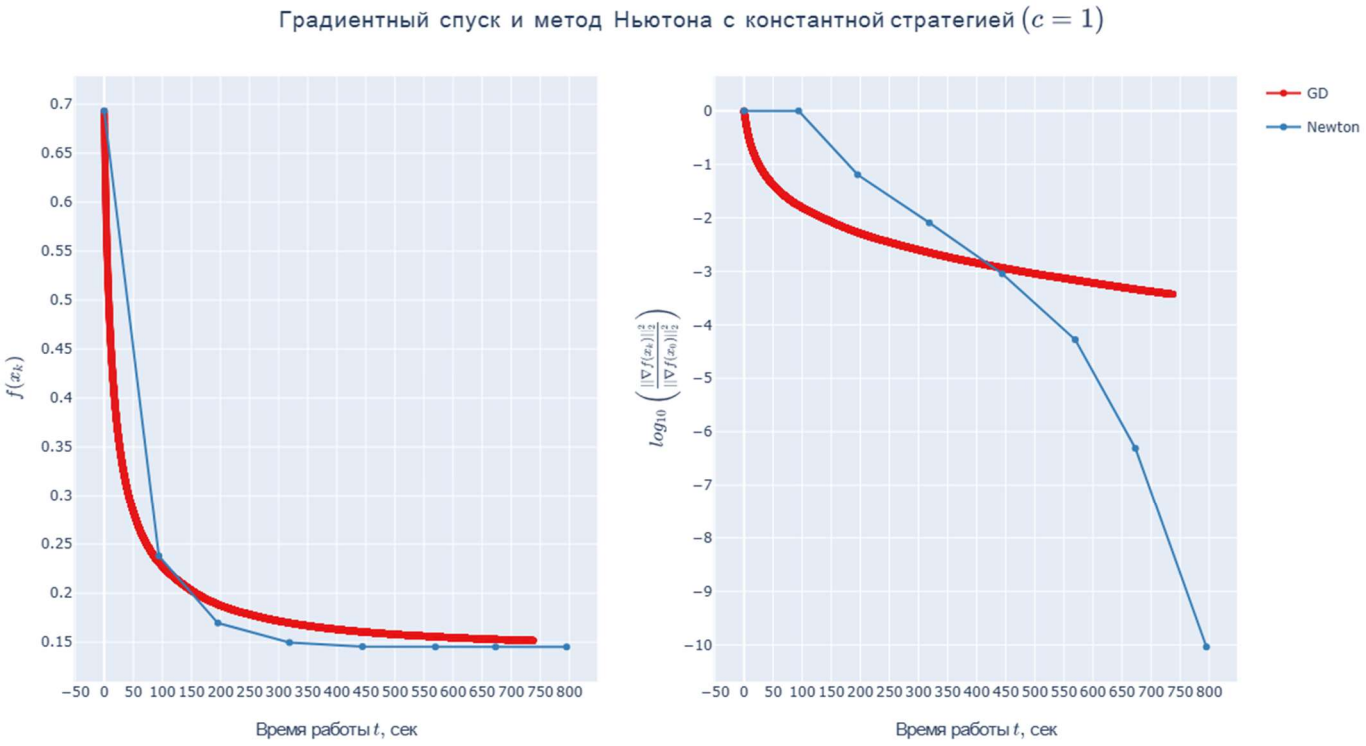


Рис 16. Градиентный спуск и метод Ньютона с константной стратегией (c = 1) для датасета real-sim

В данном случае градиентный метод показал лучший результат на адаптивных стратегиях подбора α по сравнению с методом Ньютона, однако даже не сошелся при константной реализации линейного поиска

Обобщая, градиентный метод сходится обычно сублинейно и в некоторых случаях линейно. Метод Ньютона же практически всегда сходится суперлинейно, однако эта суперлинейность начинает проявлять себя в области оптимума, а в начале градиентный спуск работает гораздо быстрее. Таким образом, если в задаче не нужна большая точность, то градиентный спуск – хорошее решение, если же требуются точные вычисления, то метод Ньютона работает быстрее и эффективнее.

Если говорить о стратегиях линейного спуска, то условия Вульфа практически всегда показывают лучший результат (один раз каким-то чудом константный метод оказался быстрее).

4. Стратегия выбора длины шага в градиентном спуске

В данном эксперименте исследуется, как зависит поведение метода от стратегии подбора шага: константный шаг (различные значения c), бэктрэкинг (условие Армихо) (попробовать различные константы c_1), условия Вульфа (попробовать различные параметры c_2).

Рассматривается квадратичная функция и логистическая регрессия с модельными данным (сгенерированными случайно).

В случае квадратичного оракула генерируется случайная матрица размерности 100 так же, как в упражнение 2 (число обусловленности $k = 10$). Вектор b тоже случайный, начальная точка равна 0. Значение минимума функции вычисляется аналитически (приравнять градиент нулю и подставить точку в функцию):

$$f(x^*) = -\frac{1}{2}\langle b, A^{-1}b\rangle$$

Для стратегии Вульфа коэффициент $c_2 \in \{0.95, 0.9, 0.4, 0.1, 0.05, 0.001,\}$. Для стратегии Армихо – $c_1 \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1},\ 0{,}5,\ 0{,}9\}$. Для константы $c \in \{2, 1, 0.5, 0.1, 0.01, 0.001, 0.0001\}$. Для каждой стратегии строится свой график, на котором разным линиям соответствуют разные коэффициенты.

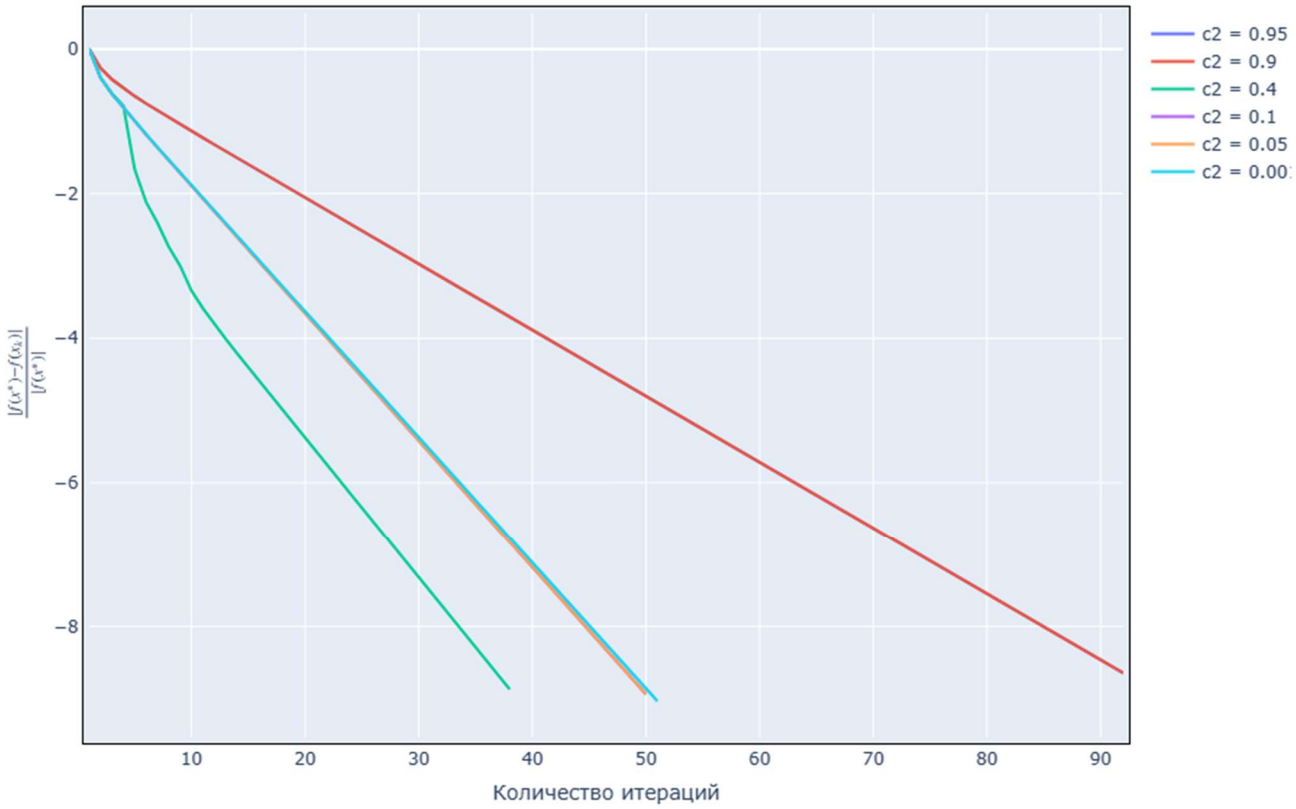


Рис 17. Зависимость логарифма относительной невязки функции от числа итераций для стратегии Вульфа

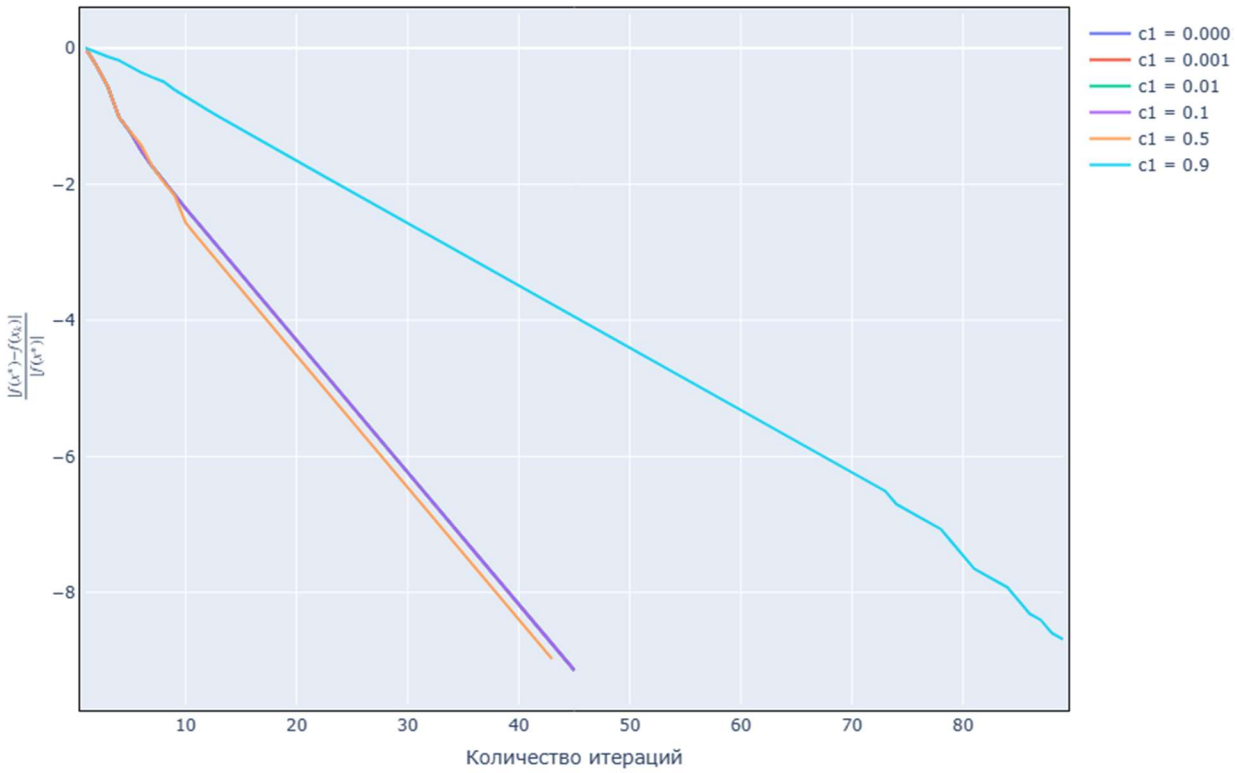


Рис 18. Зависимость логарифма относительной невязки функции от числа итераций для стратегии Армихо

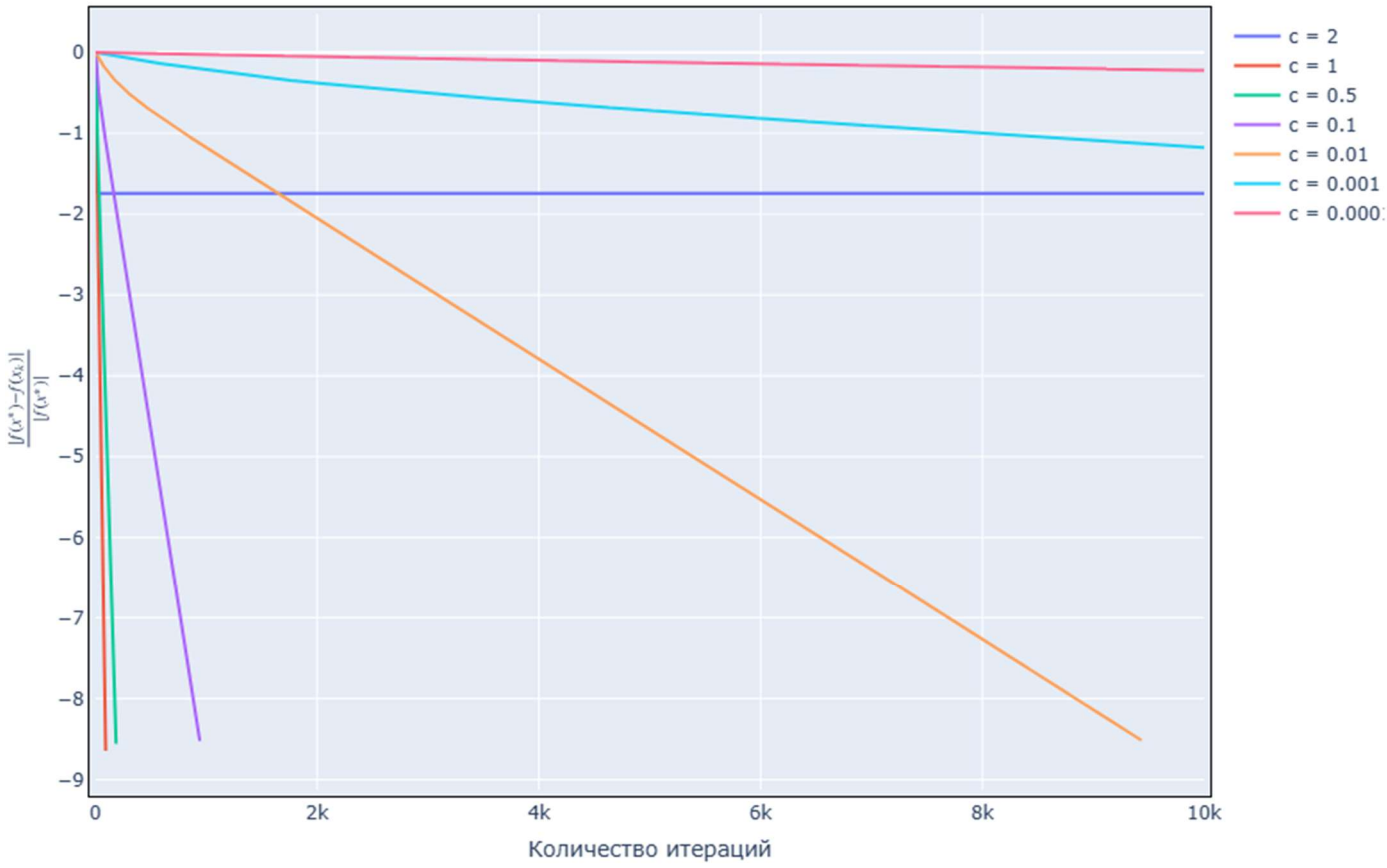


Рис 19. Зависимость логарифма относительной невязки функции от числа итераций для константной стратегии

Для условий Вульфа лучшей константой оказалась $c_2 = 0.4$, все константы, которые больше сходятся медленнее всего, а меньшие константы чуть-чуть проигрывают, хотя вначале и в конце имеют одинаковую скорость. Однако относительно константной стратегии любой выбор c_2 дает хорошие результаты.

Для условий Армихо только спуск с константой $c_1 = 0.9$ плохо выделяется на основе остальных. В остальном результаты схожи с применением условий Вульфа.

Для константного шага поведение спуска очень сильно зависит от выбора константы: взять очень маленькую – метод не успеет сойтись, очень большую – вообще разойдется.

Однако нужно понимать, что конкретные коэффициенты характеры только для данной матрицы с данной начальной точкой.

В случае оракула логистической регрессии генерируется случайная матрица размерности (2000, 500). Вектор b тоже случайный со значениями $\{-1, 1\}$, начальная точка равна 0.

Для стратегии Вульфа коэффициент $c_2 \in \{0.99, 0.9, 0.4, 0.1, 0.05, 0.001\}$. Для стратегии Армихо – $c_1 \in \{10^{-4},\ 10^{-3},\ 10^{-2},\ 10^{-1},\ 0.5,\ 0.9\}$. Для константы $c \in \{10^{-6}, 2 \cdot 10^{-6}, 3 \cdot 10^{-6}\}$. Для каждой стратегии строится свой график, на котором разным линиям соответствуют разные коэффициенты.

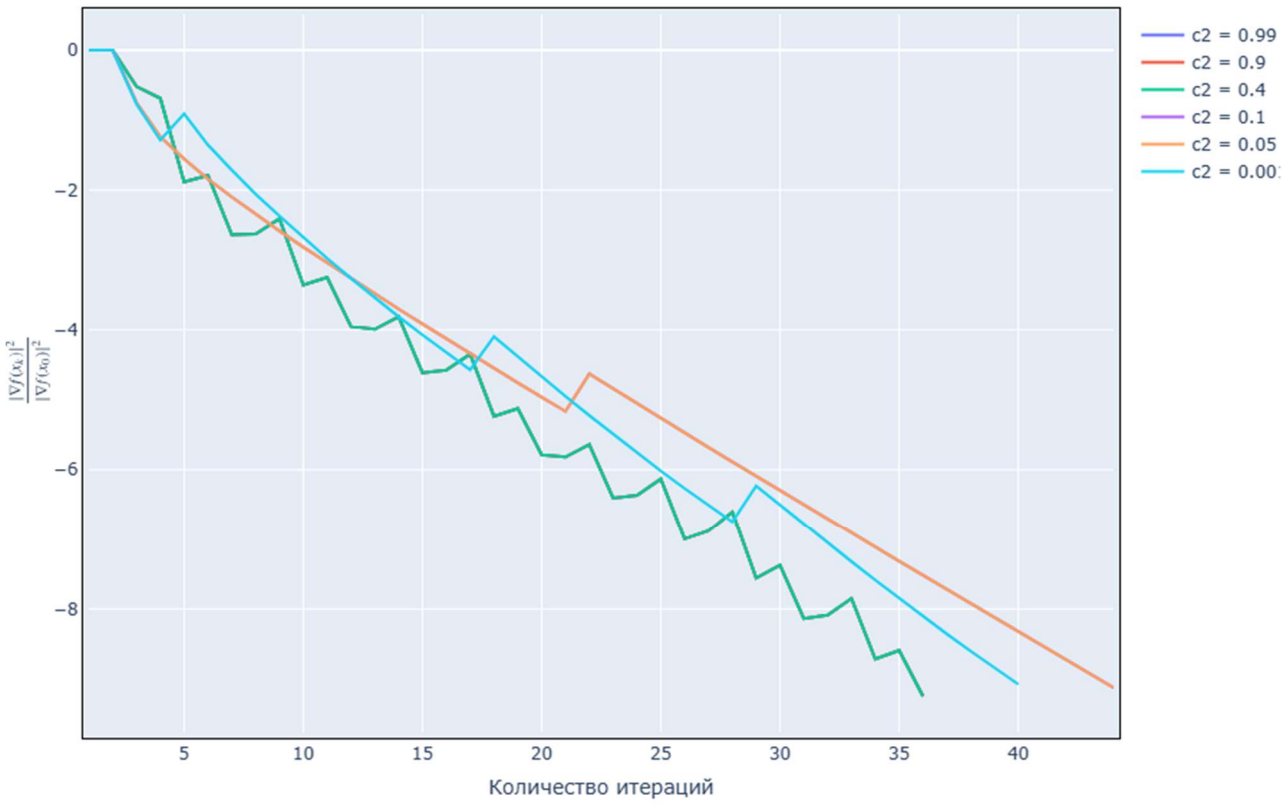


Рис 20. Зависимость относительной невязки градиента от числа итераций для стратегии Вульфа

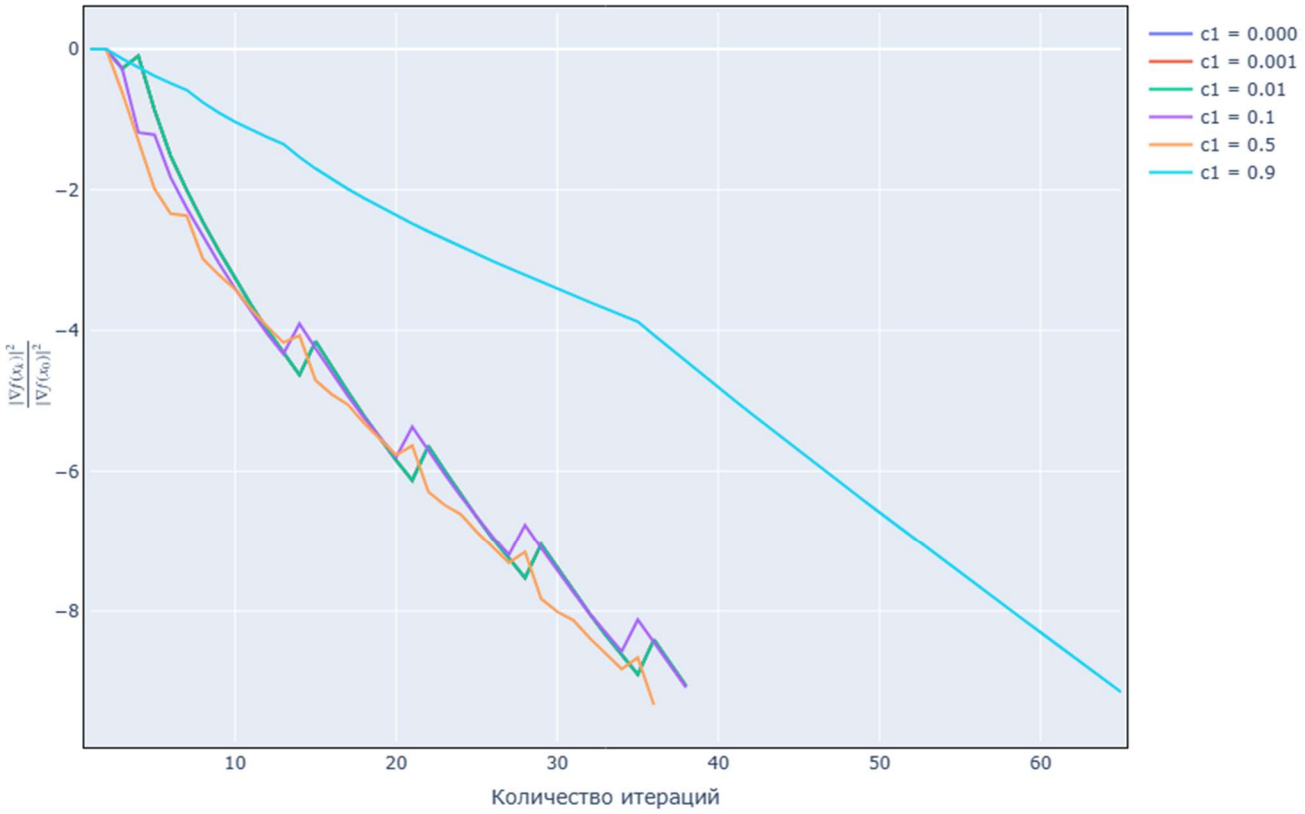


Рис 21. Зависимость относительной невязки градиента от числа итераций для стратегии Армихо

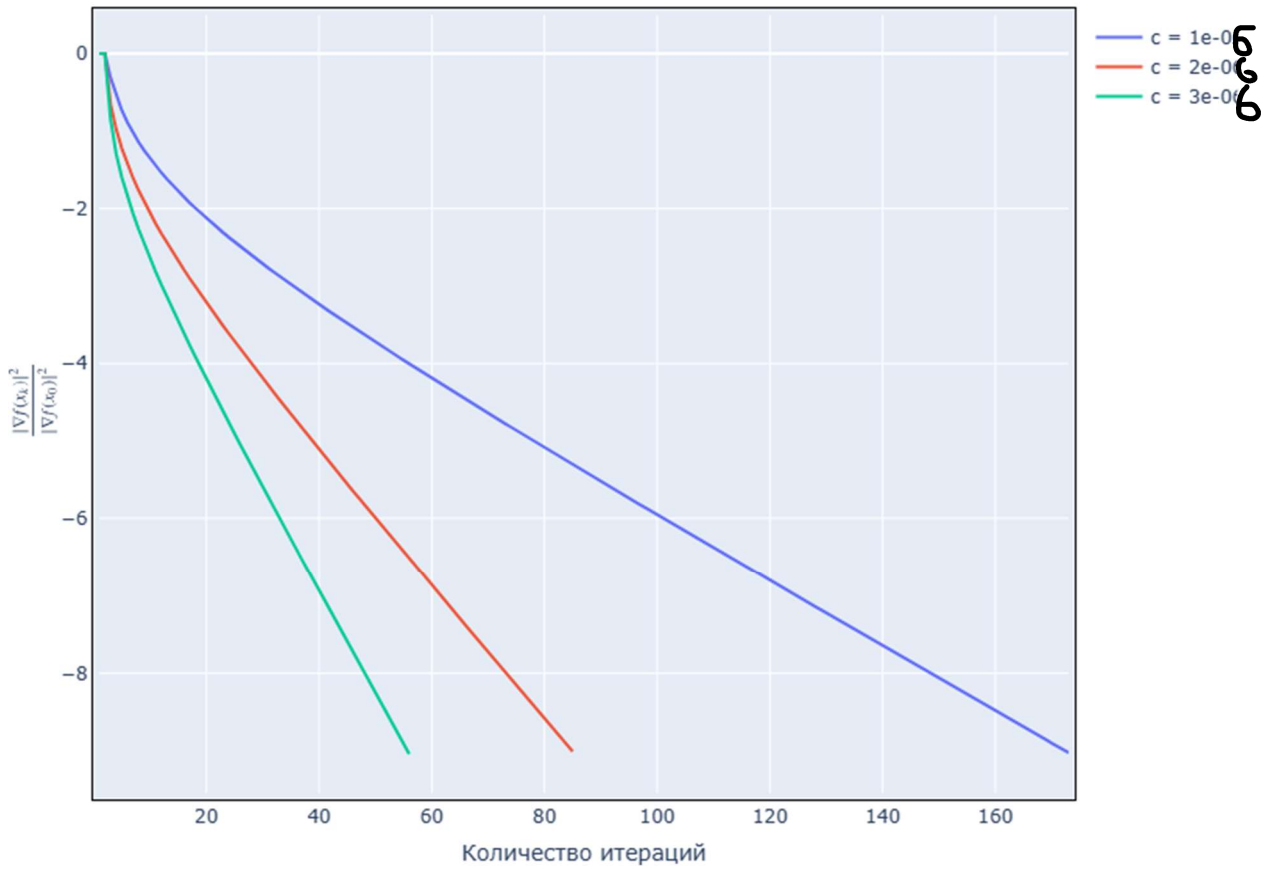


Рис 22. Зависимость относительной невязки градиента от числа итераций для стратегии константы

Для логистической регрессии все выводы аналогичны квадратичной функции, однако поведение сходимости отличается – видно периодическое поведение невязки градиента для всех методов.

Делая вывод, можно сказать, что условия Вульфа и Армихо не очень чувствительны к изменению параметров, чего нельзя сказать о константном методе – тут все очень сильно зависит от выбора шага градиентного спуска.

5. Стратегия выбора длины шага в методе Ньютона

Повторение эксперимента 4 со сравнением стратегий выбора шага на логистической регрессии с модельной выборкой, но для метода Ньютона.

Все параметры и функции такие же, как и в эксперименте 5 для логистической регрессии, меняется только метод поиска:

В случае оракула логистической регрессии генерируется случайная матрица размерности (2000, 500). Вектор b тоже случайный со значениями $\{-1, 1\}$, начальная точка равна 0.

Для стратегии Вульфа коэффициент $c_2 \in \{0.99, 0.9, 0.4, 0.1, 0.05, 0.001\}$. Для стратегии Армихо – $c_1 \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.5, 0.9\}$. Для константы $c \in \{1, 0.5, 0.1, 0.01, 0.001, 0.0001\}$. Для каждой стратегии строится свой график, на котором разным линиям соответствуют разные коэффициенты.

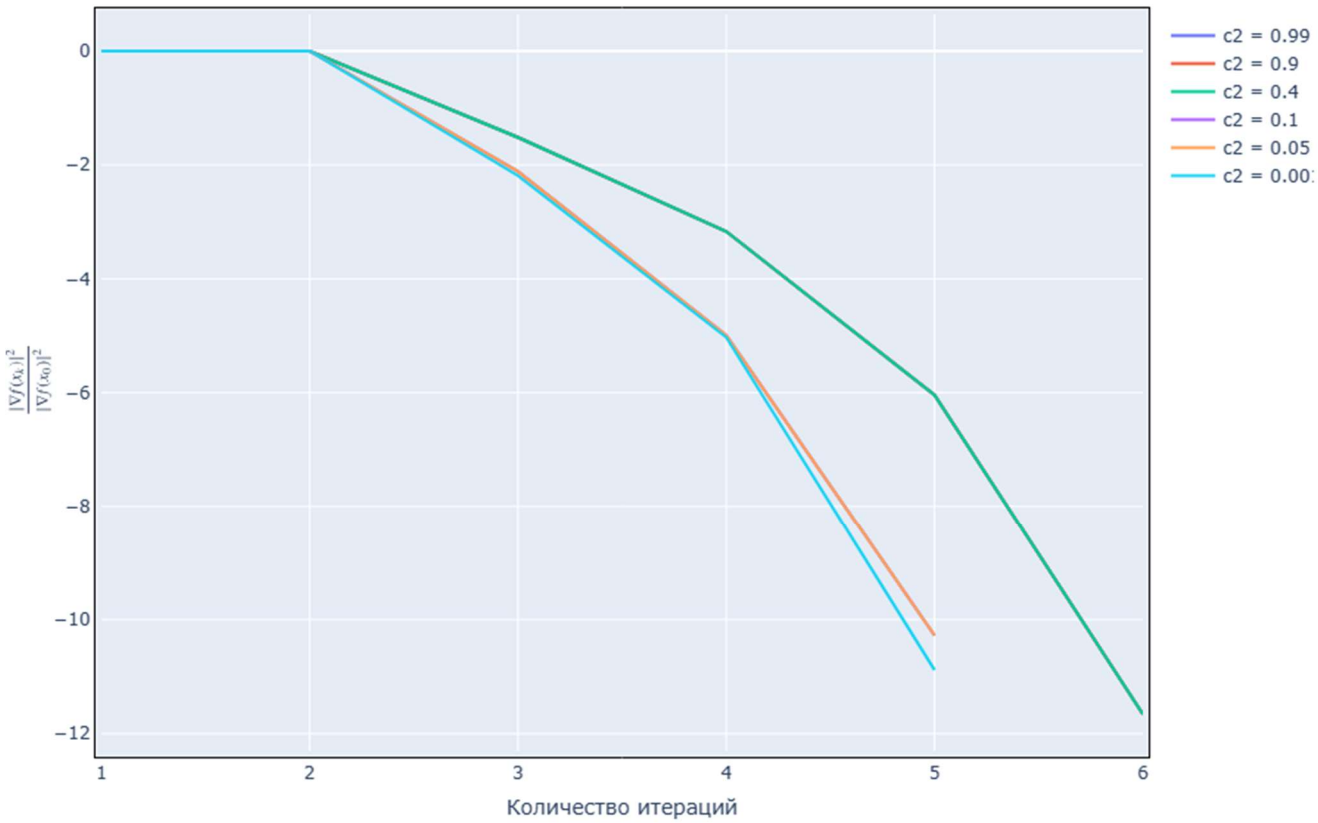


Рис 23. Зависимость относительной невязки градиента от числа итераций для стратегии Вульфа

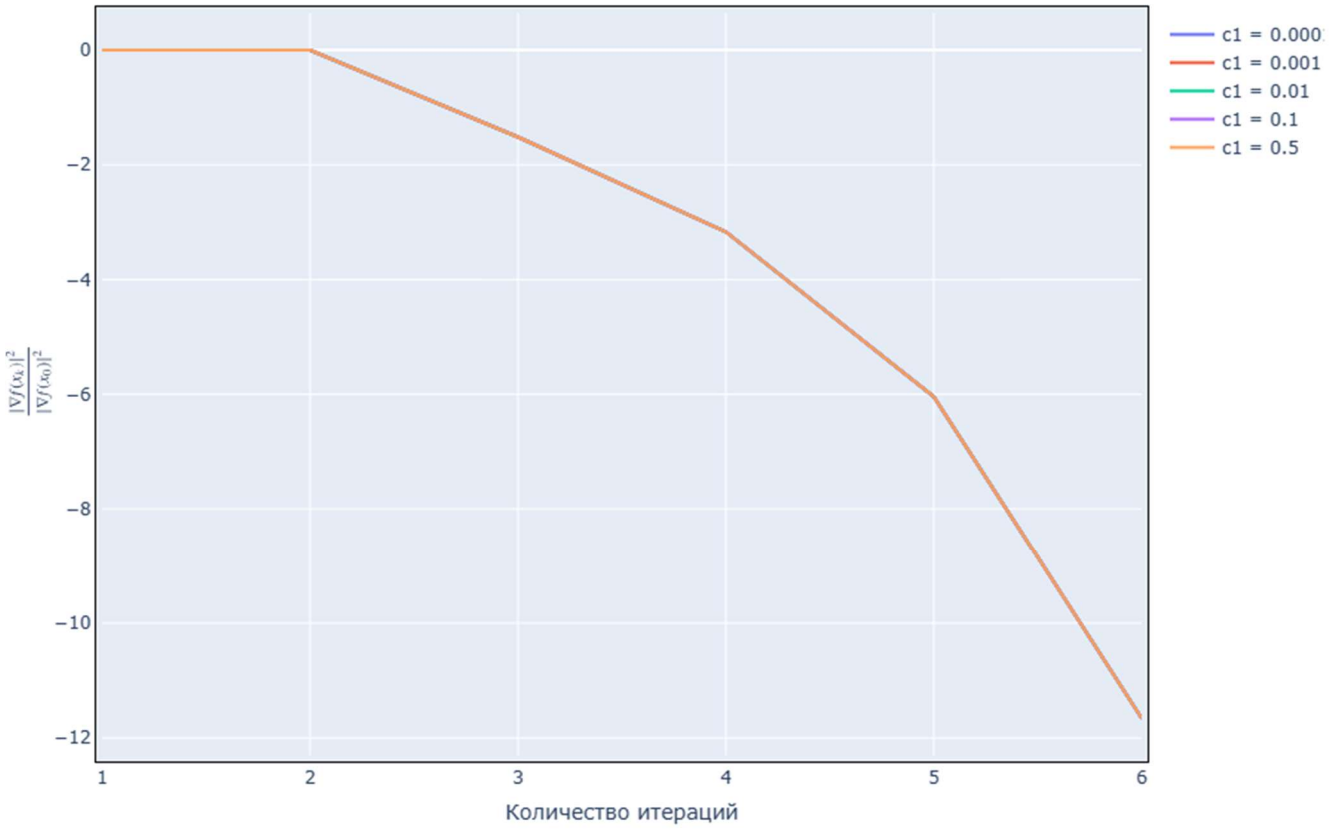


Рис 24. Зависимость относительной невязки градиента от числа итераций для стратегии Армихо

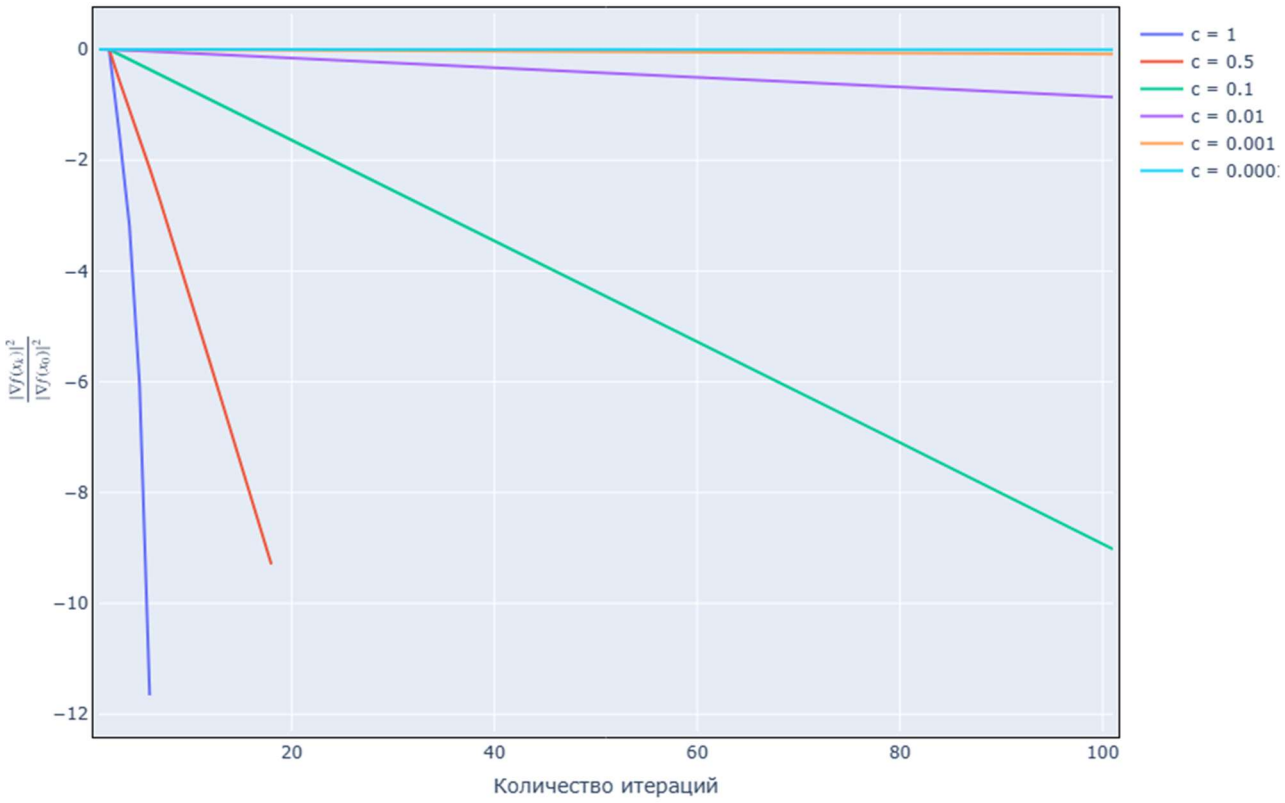


Рис 22. Зависимость относительной невязки градиента от числа итераций для стратегии константы

В данном эксперименте наилучшей стратегией линейного поиска оказались условия Вульфа, однако они очень близки с условием Армихо, оба адаптивных подхода практически не зависят от их параметров, при их использовании метод Ньютона ведет себя как суперлинейный. Константный шаг опять же очень сильно зависит от выбора константы: при $c = 1$ метод сходится суперлинейно, в остальных случаях – линейно или вовсе не сходится за 100 итераций.

Вывод

В данной работе были проведены различные эксперименты для исследования градиентного спуска, метода Ньютона, влияние выбора стратегии линейного поиска и их параметров, а также параметров задачи, такие как размерность и число обусловленности.

Результаты показали, что четкой зависимости от размерности нет. Чем больше число обусловленности, тем в среднем дольше работает градиентный спуск.

Метод Ньютона подходит для точных вычислений, однако, если требуется быстро найти минимум с невысокой точностью, то градиентный спуск работает быстрее метода Ньютона.

При рассмотрении стратегий линейного поиска было неоднократно получено, что условия Вульфа являются самыми эффективными. Самый нестабильный – константный шаг. Однако про влияние параметров сказать трудно, так как для каждой задачи зависимости могут отличаться.