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

Отчет по Лабораторной работе 1: Методы градиентного спуска и метод Ньютона

Выполнил:
студент группы ММОАД231С
Широбоков А.А.

Санкт-Петербург, 2024

Содержание

1 Траектория градиентного спуска на квадратичной функции

В данном эксперименте проводится сравнение траекторий градиентного спуска для оптимизации двумерной квадратичной функции следующего вида:

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

При этом выбираются различные матрицы A, векторы b, различные начальные точки x_0 и различные стратегии линейного поиска шага градиентного спуска (Армихо, Вульф, постоянный шаг). В данном случае было рассмотрено два случая с разными A, b, x_0.

Первый случай:

$$A_1 = \begin{pmatrix} 10 & 14 \\ 14 & 20 \end{pmatrix}, \quad b_1 = \begin{pmatrix} 5 \\ 3 \end{pmatrix}, \quad x_{01} = \begin{pmatrix} 4 \\ -6 \end{pmatrix}$$

Результаты получаются следующими:

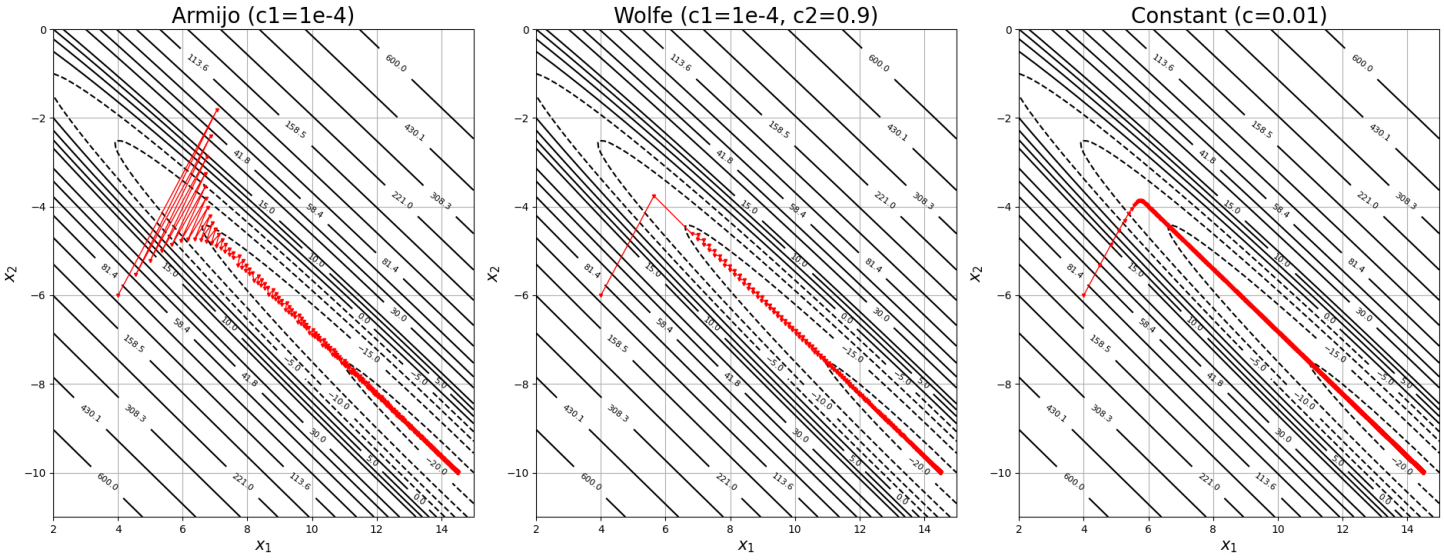

Рис. 1: Линии уровня квадратичной функции и траектория градиентого спуска при различных стратегиях выбора шага для параметров A_1, b_1, x_{01}.

Второй случай:

$$A_2 = \begin{pmatrix} 1.44 & 0 \\ 0 & 1 \end{pmatrix}, \quad b_2 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \quad x_{02} = \begin{pmatrix} 4.5 \\ -2 \end{pmatrix}$$

Результаты получаются следующими:

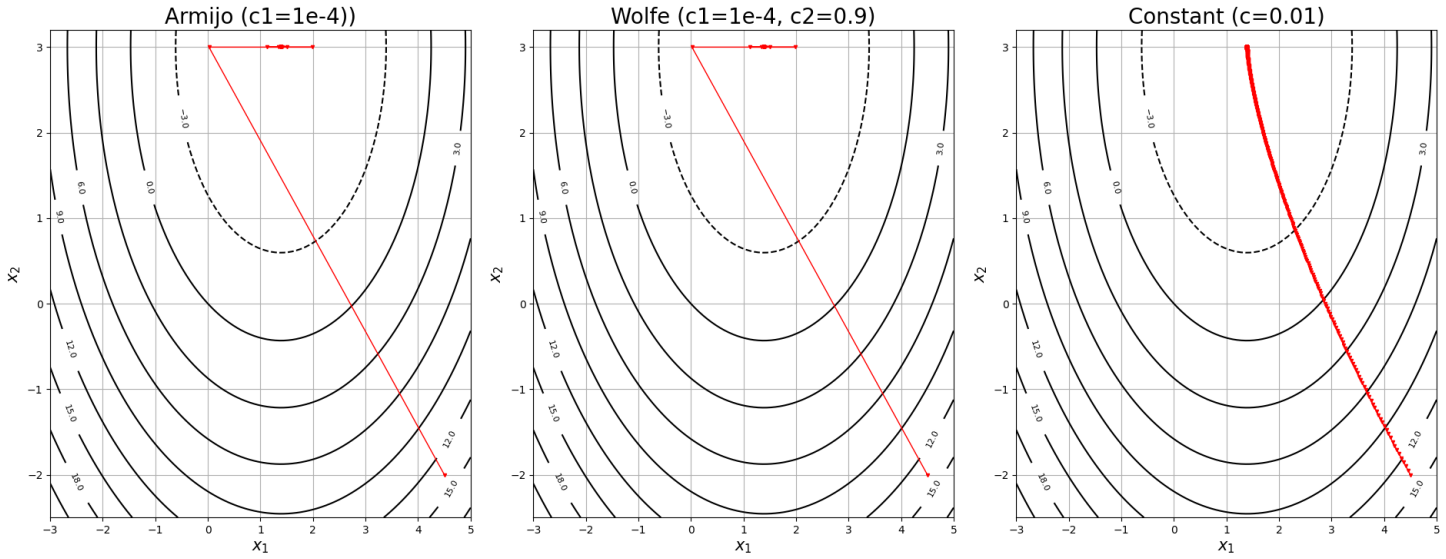

Рис. 2: Линии уровня квадратичной функции и траектория градиентого спуска при различных стратегиях выбора шага для параметров A_2, b_2, x_{02}.

В обоих случаях использовался следующий критерий остановки:

$$\|\nabla f(x_k)\|_2^2 \leq \varepsilon \|\nabla f(x_0)\|_2^2, \qquad \varepsilon = 10^{-15}$$

При этом константы в стратегиях линейного поиска для первого и второго случая выбирались одинаковыми.

Из полученных результатов видно, что чем больше число обусловленности матрицы A (чем более вытянуты линии уровня функции $f(x)$), тем больше шагов требуется градиентному спуску, чтобы добиться указанной точности (для вытянутых линий уровня наблюдаются характерные осцилляции градиентного спуска, связанные с тем, что в каждой точке градиент должен быть ортогонален линиям уровня, и поэтому в плохо обусловленном случае он направлен совсем не в точку оптимума). В данном случае имеем:

$$\mu(A_1) \approx 223 \quad \text{(линии уровня сильно вытянуты)}$$

$$\mu(A_2) = 1.44 \quad \text{(линии уровня близки к окружностям)}$$

Выбор начальной точки тоже заметно влияет на количество итераций, особенно в случае константной стратегии.

Стратегия же выбора шага очень сильно влияет на скорость сходимости - так, наилучший результат показывает стратегия выбора шага с использованием сильных условий

Вульфа, промежуточным являются условия Армихо, и худший вариант наблюдается с постоянным шагом (хотя, конечно, в рамках конкретной задачи можно попытаться так подобрать константы c, c_1, c_2, что результат будет несколько иным). Более того, в данном случае условия Армихо и сильные условия Вульфа дают сходимость при любых параметрах $c_1 \in (0, 0.5), c_2 \in (c_1, 1)$, а вот неудачно выбранный постоянный шаг (например, $c = 1.0$) даёт расходимость градиентного спуска. Кроме того, можно заметить, что в первом случае условия Армихо обеспечивают больший шаг и большие колебания, по сравнению с сильными условиями Вульфа, которые ограничивают шаг за счёт дополнительного условия на кривизну.

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

В данном эксперименте производится оптимизация квадратичной функции, указанной в эксперименте 1. Исследуется зависимость числа итераций градиентного спуска от размерности пространства и числа обусловленности матрицы A.

Для исследования такой зависимости сделано следующее. Рассматриваются различные размерности пространства $n \in \{4, 16, 64, 256\}$. При каждой фиксированной размерности рассматриваются матрицы A с различными числами обусловленности κ в диапазоне от 2 до 1000 с шагом 100. В качестве матрицы $A \in \mathbb{S}^n_{++}$ с данным числом обусловленности κ берётся диагональная матрица такая, что:

$$A_{00} = \frac{1}{\kappa}, \quad A_{n-1,n-1} = 1, \quad A_{ii} \sim U\left(\frac{1}{\kappa}, 1\right), \quad i = 1, ..., n-2$$

Здесь деление на число обусловленности делается для того, чтобы норма матрицы A всегда была единичной (достигнуто в обсуждениях с Данилом Сморчковым). Число обусловленности при этом всё равно получается равным κ. Элементы вектора b генерируются из стандартного нормального распределения. В качестве начальной точки x_0 во всех случаях берётся нулевой вектор. Для каждой пары n, κ замеряется время работы градиентного спуска для достижения погрешности $\varepsilon = 10^{-9}$ (критерий остановки тот же, что в эксперименте 1).

Замечание: в данном случае элементы матрицы A - случайные числа, поэтому для данных n, κ эксперимент проводится многократно, а затем производится усреднение результата. Кроме того, на графиках ниже приводятся именно усреднённые результаты по 5 (для метода Армихо), 15 (для метода Вульфа), 10 (для константной стратегии) итерациям. Для каждой отдельной итерации графики не приведены для лучшего визуального восприятия, уменьшения легенды и снижения нагрузки на глаза :).

Результаты получаются следующими:

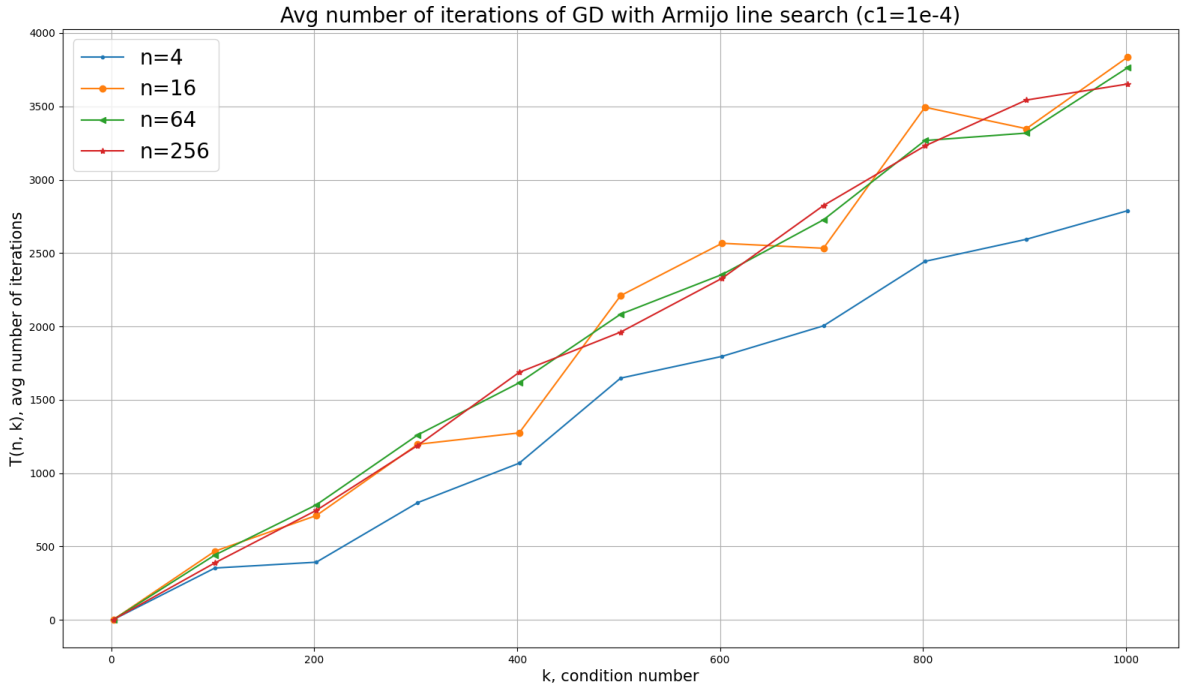

Рис. 3: Зависимость среднего (по 5 запускам) числа итераций градиентного спуска от числа обусловленности матрицы A при различных размерностях пространства с использованием условий Армихо ($c_1 = 10^{-4}$).

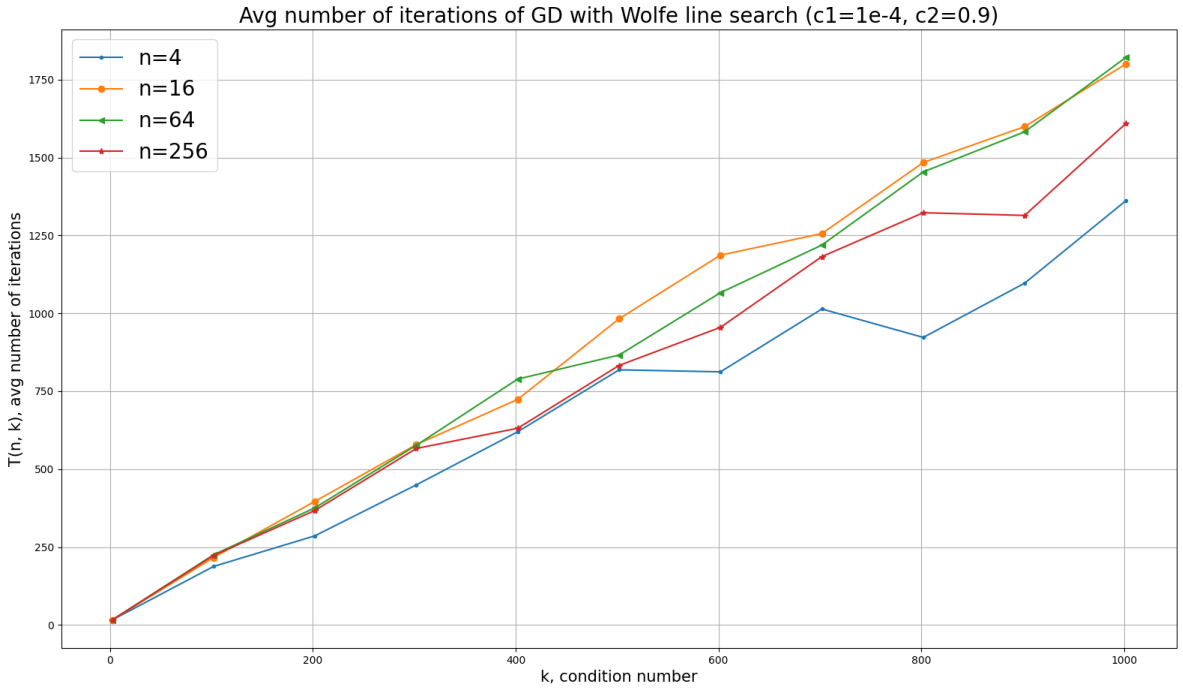

Рис. 4: Зависимость среднего (по 15 запускам) числа итераций градиентного спуска от числа обусловленности матрицы A при различных размерностях пространства с использованием сильных условий Вульфа ($c_1 = 10^{-4}, c2 = 0.9$).

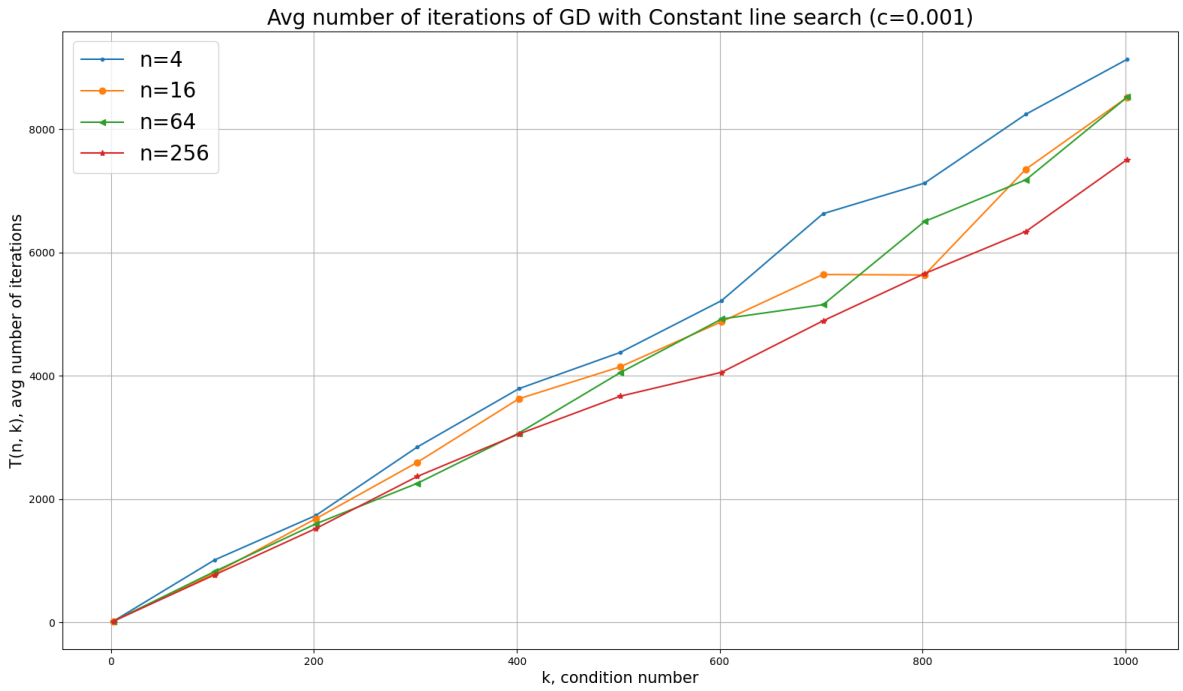

Рис. 5: Зависимость среднего (по 10 запускам) числа итераций градиентного спуска от числа обусловленности матрицы A при различных размерностях пространства с использованием константной стратегии ($c = 0.001$).

Полученные результаты демонстрируют следующее: при увеличении числа обусловленности количество итераций градиентного спуска растёт линейно во всех стратегиях линейного поиска. Увеличение же размерности пространства никак не сказывается на числе итераций градиентного спуска.

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

В данном эксперименте проводится сравнение методов градиентного спуска и Ньютона на задаче обучения логистической регрессии на реальных данных. В качестве реальных данных взяты датасеты w8a, gisette, real-sim с сайте LIBSVM.

Теоретическая часть. В данном случае логистическая регрессия используется для решения задачи бинарной классификации. Для этого минимизируется следующая функция потерь:

$$L(x) = \frac{1}{m}\sum_{i=1}^{m}\ln(1+\exp(-b_i\langle a_i, x\rangle)) + \frac{\lambda}{2}\|x\|_2^2 \longrightarrow \min_{x\in\mathbb{R}^n}$$

где $b_i \in \{-1, 1\}$ - метка класса i-го объекта, $a_i \in \mathbb{R}^n$ - вектор признаков i-го объекта, $x \in \mathbb{R}^n$ - параметры логистической регрессии (вектор нормали разделяющей разделяющей гиперплоскости), λ - коэффициент регуляризации.

В векторно-матричном виде функция потерь может быть переписана следующим образом:

$$L(x) = \frac{1}{m}\langle 1_m, \ln(1_m + \exp(-b \odot Ax))\rangle + \frac{\lambda}{2}\langle x, x\rangle$$

где $\langle ., .\rangle$ - стандартное скалярное произведение, A - матрица объектов-признаков, в строках которой расположены объекты, а в столбцах - их признаки, 1_m - вектор из единиц длины m. Операции exp и ln применяются поэлементно.

Градиент такой функции потерь несложно находится путём поиска дифференциала в стандартной форме:

$$dL(x) = \langle \nabla L(x), dx\rangle$$

Проделав эти несложные действия, находим:

$$\nabla L(x) = -\frac{1}{m}A^T\left(b \odot \sigma(-b \odot Ax)\right) + \lambda x$$

где $\sigma(x) = \frac{1}{1+\exp(-x)}$.

Отсюда видно, что так как градиент - это вектор размерности n, то для его хранения необходимо $O(n)$ памяти. А из самой формулы видно, что для его вычисления, а, следовательно, и для всей итерации градиентного спуска, необходимо $O(mn)$ (умножение Ax за $O(mn)$, затем последовательное применение поэлементных операций за $O(m)$,

а затем ещё одно матрично-векторное произведение за $O(nm)$) операций.

Гессиан данной функции потерь тоже несложно находится путём нахождения второго дифференциала:

$$d^2L(x) = \langle \nabla^2 L(x) dx_1, dx_2 \rangle$$

Опять же проделав это несложное вычисление, получаем:

$$\nabla^2 L(x) = \frac{1}{m} A^T \Sigma A + \lambda I_n$$

где $\Sigma = \operatorname{diag}\left(\sigma(b \odot Ax) \odot (1 - \sigma(b \odot Ax))\right)$ - диагональная матрица размера m, I_n - единичная матрица размера n.

Из формулы для гессиана видно, что для его вычисления необходимо $O(m^2 n)$ операций, так как нужно посчитать матричное произведение трёх матриц размеров $n \times m, m \times m, m \times n$. Соответственно, для хранения этого гессиана требуется порядка $O(n^2)$ памяти. Наконец, для нахождения спуска, нужно решить систему линейных уравнений, что потребует ещё порядка $O(n^3)$ операций. Итого, сложность одной итерации метода Ньютона в данном случае получается порядка $O(n^3) + O(m^2 n)$, а хранение по памяти - $O(n^2)$.

Экспериментальная часть. Проводится минимизация логистической функции потерь для трёх различных датасетов с разными матрицами A и векторами b. Датасеты отличаются количеством объектов и количеством признаков. В датасете w8a имеется 49 749 объектов, каждый с 300 признаками. В датасете gisette - 6000 объектов с 5000 признаками. В датасете real-sim - 72 309 объектов с 20 958 признаками. Однако во всех случаях матрицы A являются разреженными. Так, в датасете w8a ненулевых элементов в матрице A - 579 586, в датасете gisette - 29 729 997, а в датасете real-sim - 3 709 083

Минимизация проводится методом градиентного спуска и методом Ньютона при различных стратегиях линейного поиска. Критерий остановки тот же, что в экспериментах 1 и 2. Для градиентного спуска погрешность $\varepsilon = 10^{-5}$, для метода Ньютона $\varepsilon = 10^{-9}$. В качестве начальной точки во всех случаях берётся нулевой вектор. Коэффициент регуляризации $\lambda = \frac{1}{m}$.

Во всех случаях параметры методов линейного поиска следующие: для константной стратегии берётся $c = 1.0$, для условий Армихо берётся $c_1 = 10^{-4}$, для сильных условий Вульфа берётся $c_1 = 10^{-4}, c_2 = 0.9$.

Результаты получаются следующие:

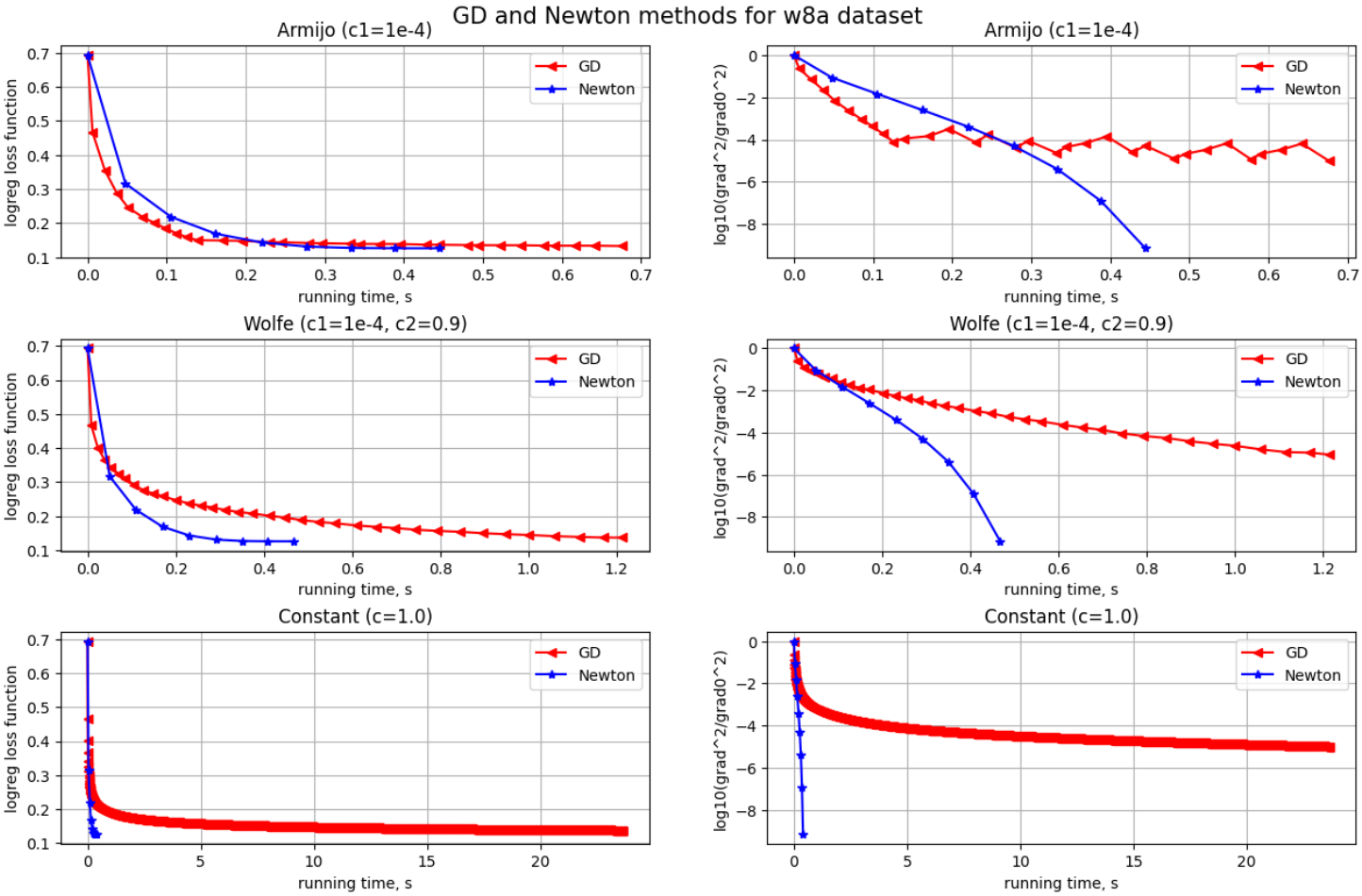

Рис. 6: Датасет w8a. Сравнение методов градиентного спуска и Ньютона при различных стратегиях линейного поиска. Слева: значение логистической функции потерь в зависимости от реального времени работы метода. Справа: логарифм отношения градиентов в зависимости от реального времени работы метода.

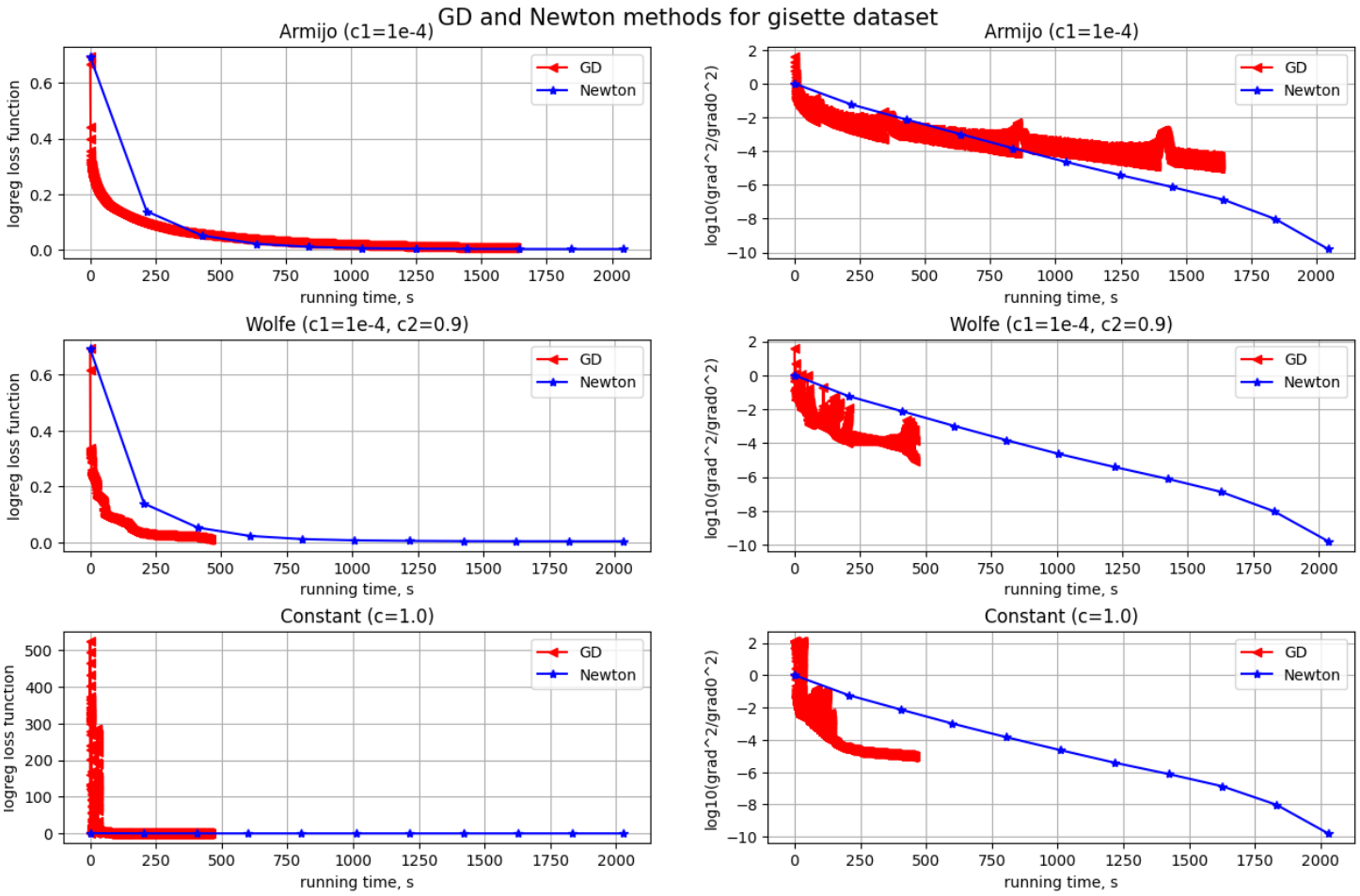

Рис. 7: Датасет gisette. Сравнение методов градиентного спуска и Ньютона при различных стратегиях линейного поиска. Слева: значение логистической функции потерь в зависимости от реального времени работы метода. Справа: логарифм отношения градиентов в зависимости от реального времени работы метода.

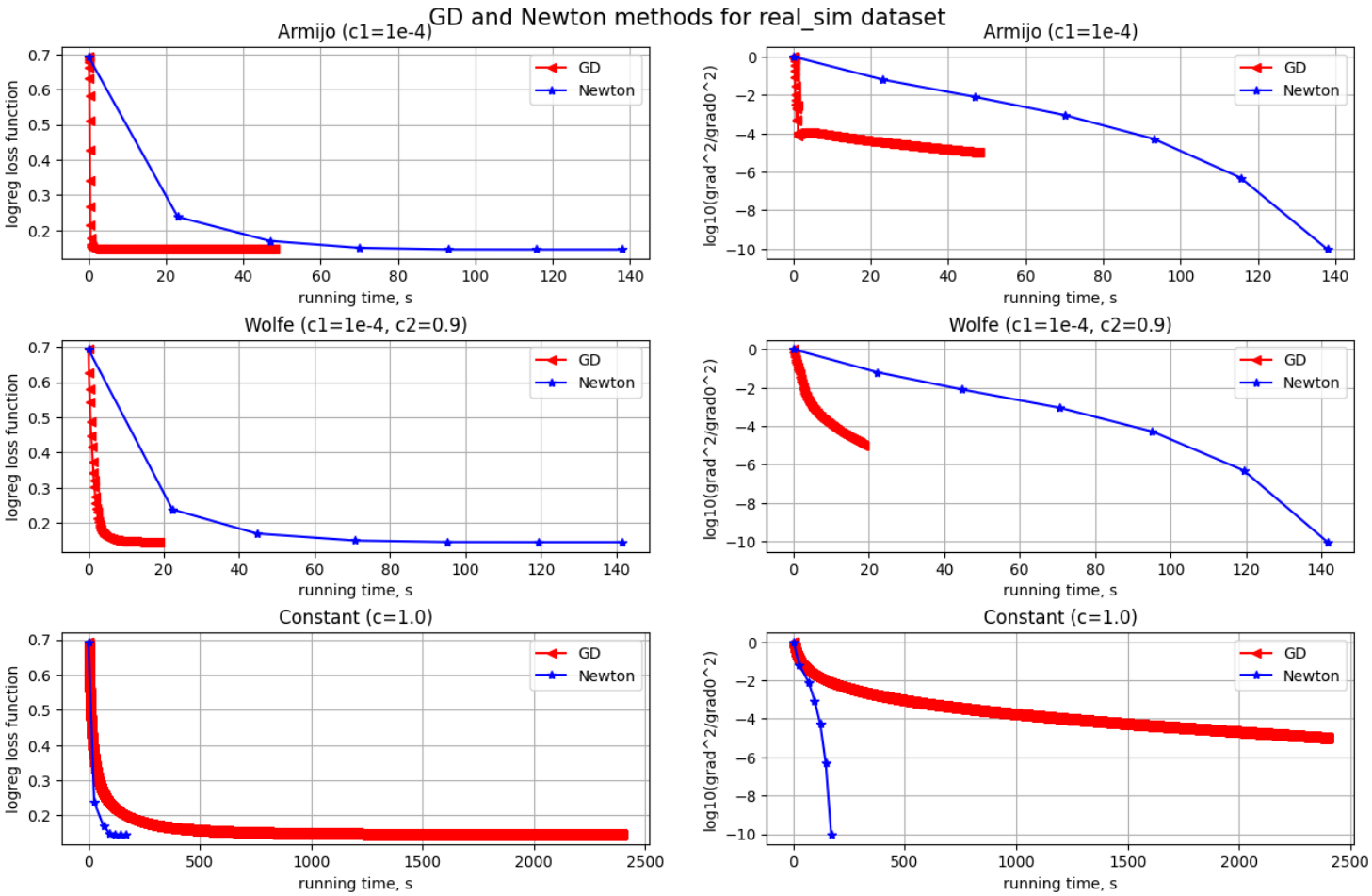

Рис. 8: Датасет real-sim. Сравнение методов градиентного спуска и Ньютона при различных стратегиях линейного поиска. Слева: значение логистической функции потерь в зависимости от реального времени работы метода. Справа: логарифм отношения градиентов в зависимости от реального времени работы метода.

Глобально, из зависимостей логарифма отношения градиентов, мы можем наблюдать следующее: вблизи оптимума градиентный спуск во всех случаях сходится линейно, а метод Ньютона - сверхлинейно.

Так же видно следующее: на не очень большом датасете (w8a) с небольшой размерностью пространства и небольшим количеством ненулевых элементов в разреженной матрице A метод Ньютона работает заметно быстрее градиентного спуска, а вот на больших датасетах (gisette, real-sim) с высокоразмерным пространством признаков и плотной матрицей A (gisette) метод Ньютона работает несравненно медленнее градиентного спуска (за исключением константной стратегии на датасете real-sim).

Однако здесь стоит вспомнить, что в данном случае погрешность для градиентного спуска - 10^{-5}, а для метода Ньютона - 10^{-9}. Если теперь мысленно провести линейную

экстраполяцию зависимостей логарифма отношения градиентов для всех датасетов для метода градиентного спуска, то будет видно, что для достижения точности 10^{-9}, ему понадобится огромное количество времени по сравнению с методом Ньютона. И дело в том, что, как отмечалось, в области оптимума метод Ньютона сходится квадратично, а градиентный спуск - всего лишь линейно. Однако если нам нужна не очень высокая точность, то градиентный спуск может отлично себя показать.

Кроме того, если сравнивать методы линейного поиска, то оптимальным здесь является метод с использованием сильных условий Вульфа - во всех случаях он показывает наилучший результат по времени сходимости методов.

4 Стратегия выбора длины шага в градиентном спуске

В данном эксперименте исследуется зависимость поведения градиентного спуска от стратегии подбора шага: константная стратегия с разными шагами, условия Армихо с разными константами c_1, сильные условия Вульфа с разными константами c_2. Исследуются квадратичная и логистическая функции потерь с модельными данными, сгенерированными случайно.

Для квадратичной функции генерируются случайная матрица A и вектор b размерности 10, а затем запускается градиентный спуск с различными стратегиями линейного поиска шага. Для константной стратегии рассматриваются значения $c \in \{10^{-5}, 7 \cdot 10^{-5}, 5 \cdot 10^{-5}, 3 \cdot 10^{-5}, 2 \cdot 10^{-5}\}$, для условий Армихо значения $c_1 \in \{10^{-10}, 10^{-7}, 10^{-4}, 10^{-1}, 0.49\}$, для сильных условий Вульфа значения $c_2 \in \{0.3, 0.6, 0.9, 0.99\}$. Строится график зависимость логарифма относительной невязки по функции $\log_{10}\left(\frac{|f_k - f^*|}{|f^*|}\right)$, где $f^* = -\frac{1}{2}\langle b, A^{-1}b\rangle$ - аналитическое значение минимума квадратичной функции, от числа итераций метода.

Для логистической функции потерь генерируется случайная матрица A размером 1000x300, а также вектор меток класса b размера 1000. Коэффициент регуляризации $\lambda = 10^{-3}$. Затем запускается градиентный спуск с различными стратегиями линейного поиска шага. Для константной стратегии рассматриваются значения $c \in \{0.19, 7 \cdot 10^{-1}, 5 \cdot 10^{-2}, 2 \cdot 10^{-2}\}$, для условий Армихо значения $c_1 \in \{10^{-10}, 10^{-7}, 10^{-4}, 10^{-1}, 0.49\}$, для сильных условий Вульфа значения $c_2 \in \{0.001, 0.1, 0.99\}$. Строится зависимость логарифма отношения градиентов от числа итераций метода.

Для построения графиков в качестве начальной точки берётся нулевой вектор. Для обоих функций потерь погрешность $\varepsilon = 10^{-9}$.

Получаются следующие результаты:

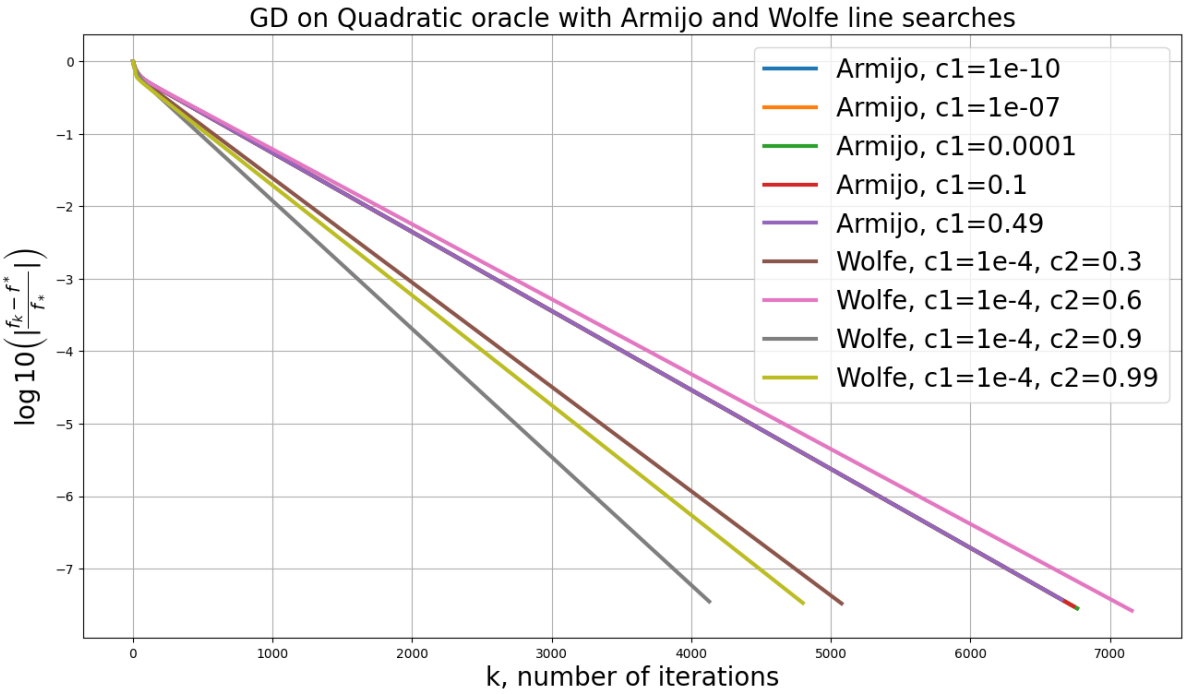

Рис. 9: Квадратичный оракул. Зависимость логарифма относительной невязки функции от числа итераций градиентного спуска при различных значениях констант адаптивных методов подбора шага (Армихо и Вульф).

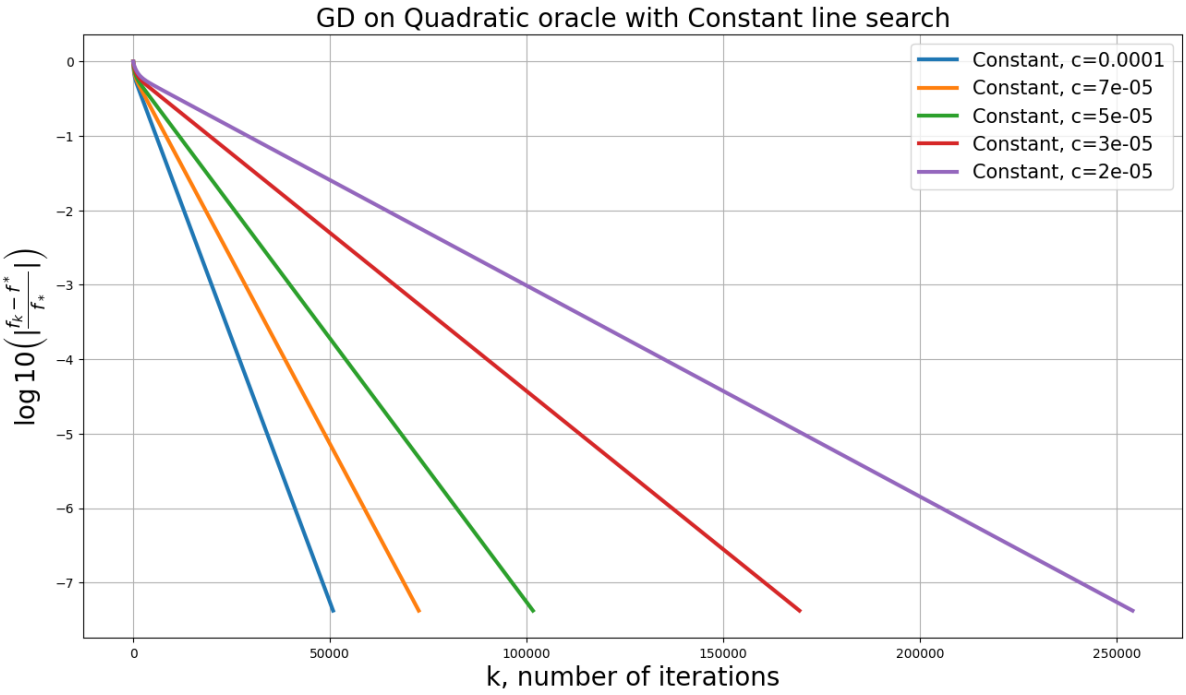

Рис. 10: Квадратичный оракул. Зависимость логарифма относительной невязки функции от числа итераций градиентного спуска для различных константных стратегий выбора шага.

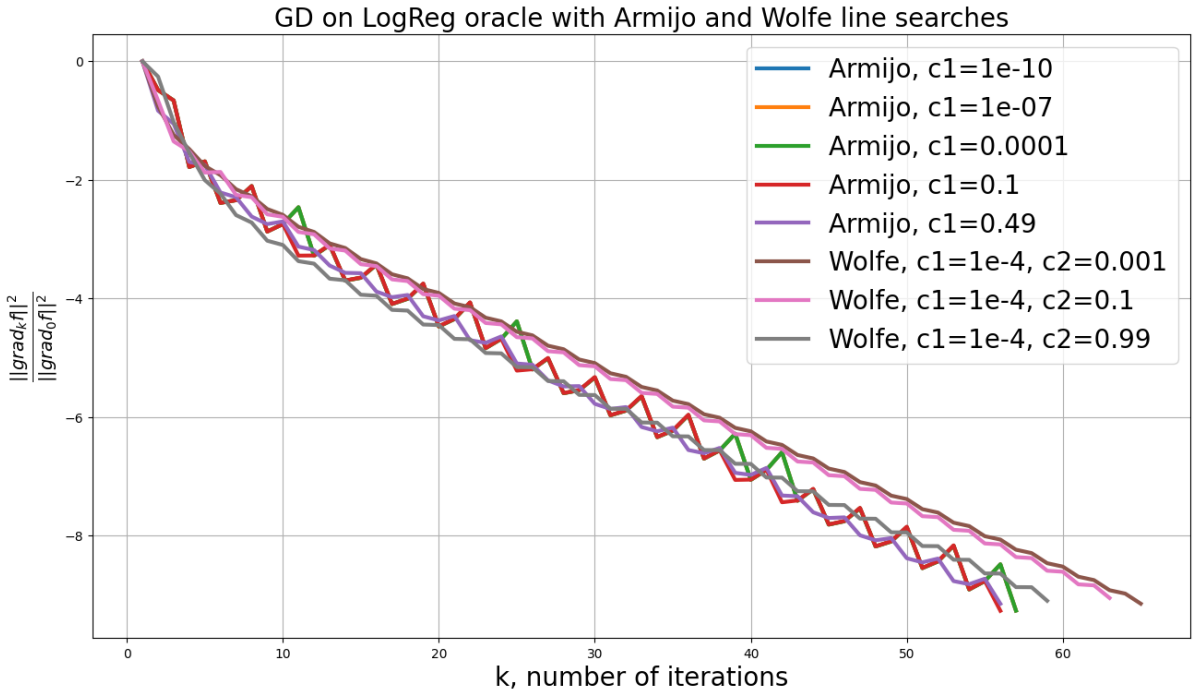

Рис. 11: Логистический оракул. Зависимость логарифма отношения градиентов от числа итераций градиентного спуска при различных значениях констант адаптивных методов подбора шага (Армихо и Вульф).

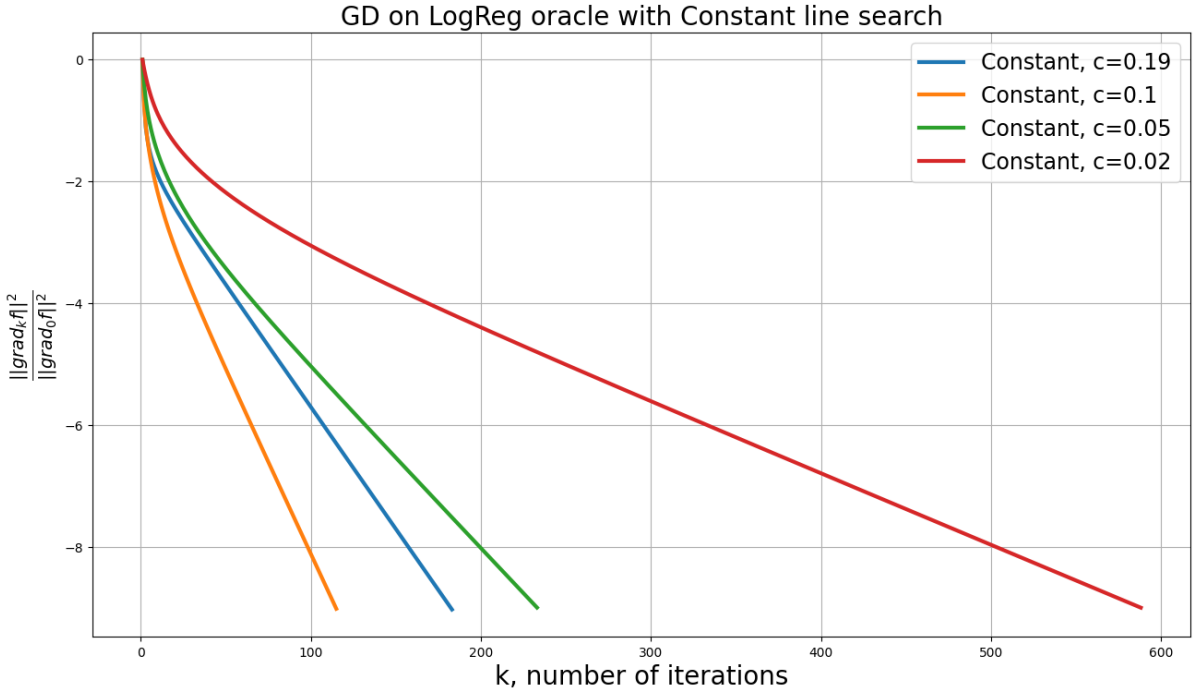

Рис. 12: Логистический оракул. Зависимость логарифма отношения градиентов от числа итераций градиентного спуска для различных константных стратегий выбора шага.

В целом, наблюдаем следующее: во всех случаях градиентный спуск сходится линейно. Кроме того, можем видеть, что чувствительность градиентного спуска к выбору констант в условиях Армихо и сильных условиях Вульфа довольно слабая - варьируя их в достаточно широких диапазонах, мы всё равно получаем одинаковый порядок по количеству итераций до сходимости.

Этого нельзя сказать о константной стратегии - здесь, уменьшая шаг всего лишь в несколько раз, мы на порядок ухудшаем скорость сходимости - константная стратегия очень чувствительна к выбору шага. Более того, если сделать его слишком большим, то градиентный спуск и вовсе разойдётся.

Таким образом, оптимальными стратегиями можно считать условия Армихо и сильные условия Вульфа - время их сходимости довольно близко за тем лишь исключением, что на логистическом оракуле при условиях Армихо возникают небольшие колебания. Тем не менее, это не влияет на сходимость. На квадратичном же оракуле условия Армихо вообще не чувствуют изменений параметра c_1.

Если говорить о начальной точке, то её роль здесь тоже довольно высока - удачный её выбор тоже может сильно повлиять на скорость сходимости алгоритма.

5 Стратегия выбора длины шага в методе Ньютона

В данном эксперименте выполняются те же действия, что и в эксперименте 4 за той лишь разницей, что здесь рассматривается только логистическая функция потерь, которая оптимизируется с помощью метода Ньютона.

Получаются следующие результаты:

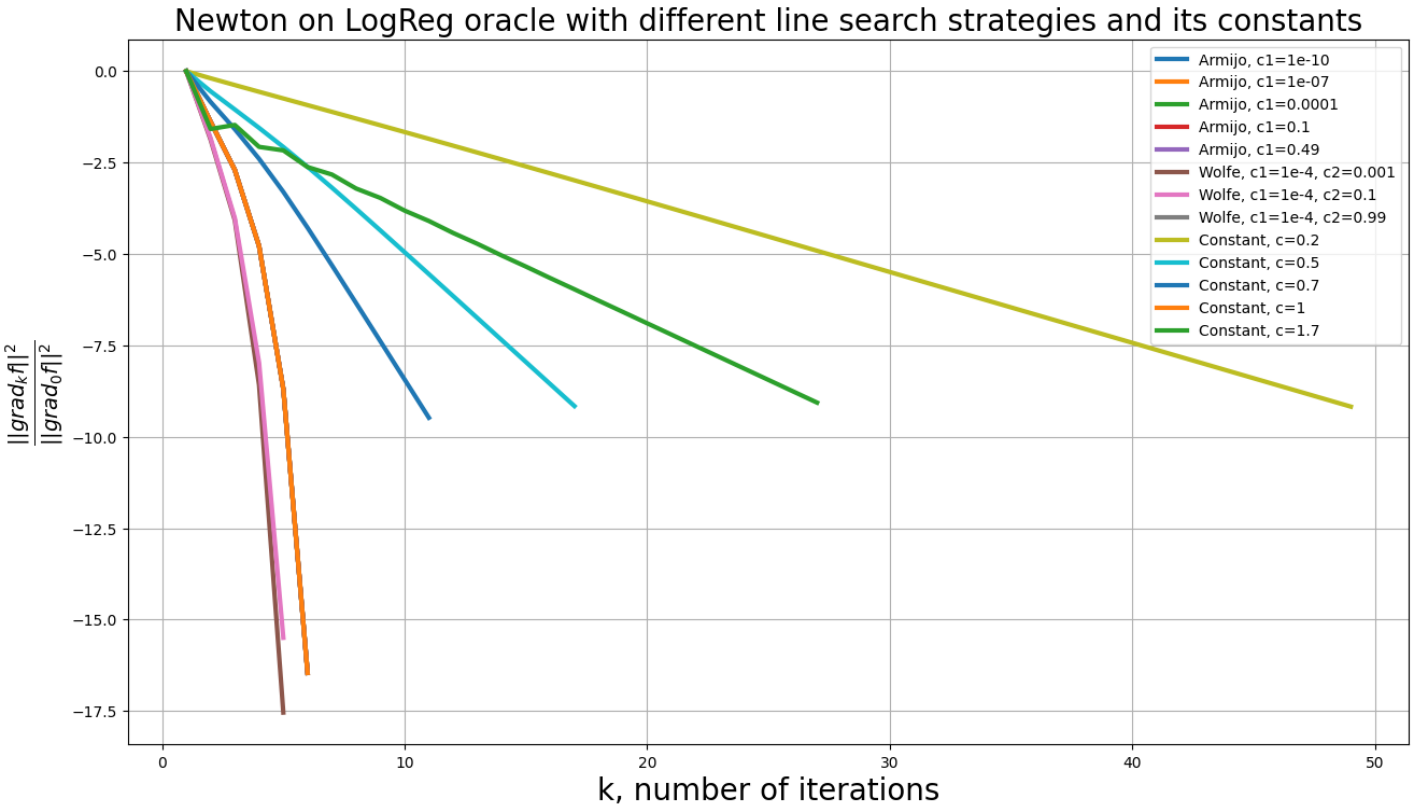

Рис. 13: Логистический оракул. Зависимость логарифма отношения градиентов от числа итераций метода Ньютона для различных стратегий линейного поиска шага.

Из полученного графика мы можем заключить следующее: при константной стратегии выбора шага может наблюдаться как линейная, так и сверхлинейная сходимость метода Ньютона, а может и вообще наблюдаться расходимость. А именно, при $c < 1$ наблюдаем линейную сходимость, при $c = 1$ - сверхлинейную, при $c \gtrapprox 1$ - линейная сходимость, при $c > 1$ - расходимость. То есть, как и в случае с градиентным спуском, мы видим сильную чувствительность метода к выбору шага.

Если же смотреть на адаптивные стратегии выбора шага (Армихо и Вульф), то здесь мы во всех случаях видим сверхлинейную сходимость и очень слабую чувствительность к вариации констант c_1 и c_2 - именно в этом и заключается адаптивность методов - вне области сходимости эти методы сами подбирают нужный шаг (уменьшая его)

для обеспечения своих условий, а в области оптимума единичный шаг удовлетворяет им автоматически и так же обеспечивает квадратичную сходимость.

Таким образом, в данном случае заключаем, что наилучшим выбор шага будет осуществляться при использовании условий Армихо/Вульфа.

Выводы

В лабораторной работе исследованы методы градиентного спуска и Ньютона с различными методами линейного поиска шага и на различных оптимизируемых функциях. Обнаружено, что при плохой обусловленности задачи сходимость градиентного спуска значительно ухудшается, хотя и может быть улучшена за счёт использования адаптивных стратегий подобора шага (например, сильные условия Вульфа). Тем не менее, если требуется не слишком высокая точность, а размер данных очень большой, градиентный спуск с адаптивным выбором шага может служить хорошим начальным приближением для достижения области оптимума, в которой имеет смысл использовать метод Ньютона, который обеспечит квадратичную сходимость и даст нужную точность за приемлемое время работы.