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: Кудреватых П.С.

Санкт-Петербург 2024

Содержание

1 Траектория градиентного спуска на квадратичной функции

В данном эксперименте оптимизируется функция квадратичного оракула вида:

$$L = \frac{1}{2} \cdot \langle Ax, x \rangle - \langle b, x \rangle$$

методом градиентного спуска с различными параметрами выбора шага.

Построим две квадратичные функции и посмотрим, как будут вести себя различные стратегии в методе градиентного спуска.

1.1 Метод Вульфа

Для первого эксперимента я взял матрицы следующего вида:

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 7 \end{pmatrix} \quad B = \begin{pmatrix} -1 \\ 10 \end{pmatrix} \quad X_0 = \begin{pmatrix} 25 \\ -5 \end{pmatrix}$$

Сперва посмотрим на работу линейного поиска со стратегией **Вульфа** и параметром точности $1e-9$, $c_1 = 1e-4, c_2 = 0.9$.

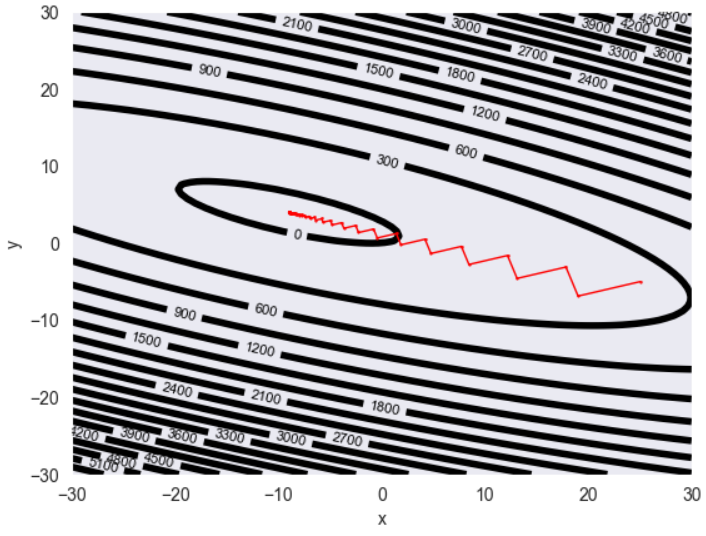

Рис. 1: Траектория градиентного спуска с выбором шага стратегией Вульфа на первом тестовом наборе. Черным изображены линии уровня, красным - траектория.

Повторим эксперимент на другом тестовом наборе:

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 7 \end{pmatrix} \quad B = \begin{pmatrix} 10 \\ -2 \end{pmatrix} \quad X_0 = \begin{pmatrix} -25 \\ -5 \end{pmatrix}$$

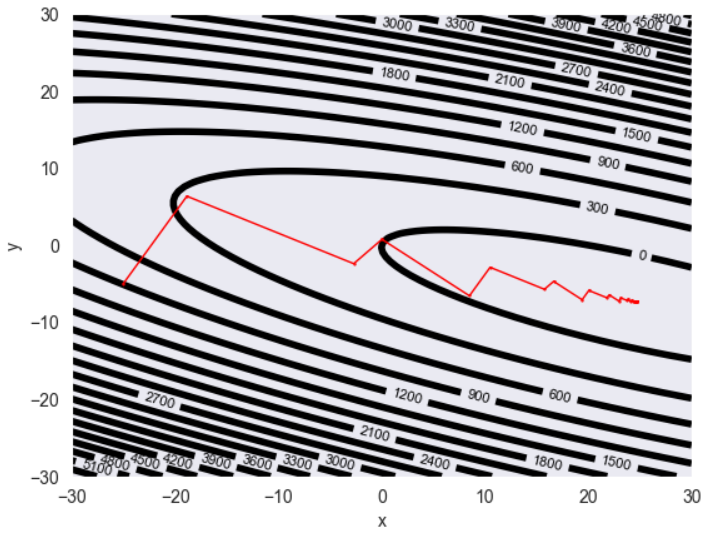

Рис. 2: Траектория градиентного спуска с выбором шага стратегией Вульфа на втором тестовом наборе.

1.2 Метод Армихо

Теперь взглянем на метод Армихо с бектрекингом на первом и втором тестовых наборах соответственно. Начальное значение $\alpha = 1$, $c_1 = 1e - 4$.

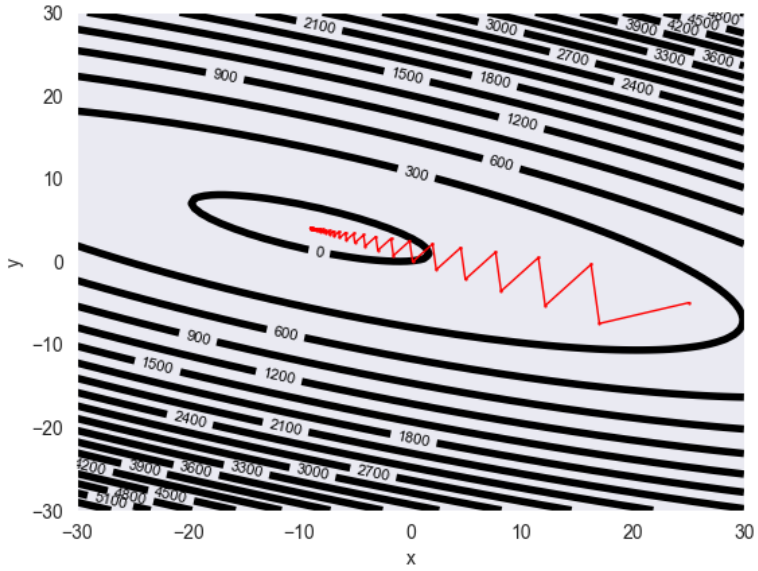

Рис. 3: Траектория градиентного спуска с выбором шага стратегией Армихо на первом тестовом наборе.

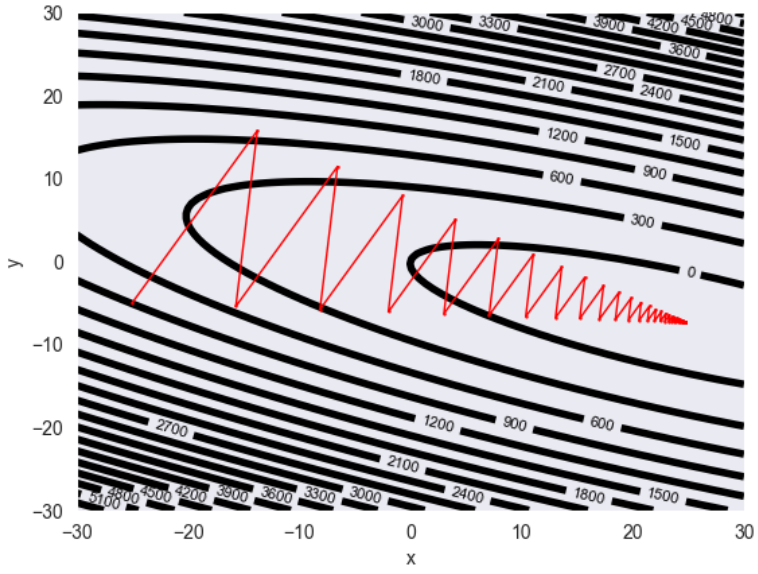

Рис. 4: Траектория градиентного спуска с выбором шага стратегией Армихо на втором тестовом наборе.

Видно, что метод Армихо отличается от метода Вульфа большим выбором

шага. Это неудивительно, ведь в методе Вульфа требуется выполнение не только условия Армихо, но и условия достаточной кривизны, что дает нам дополнительное ограничение на размер шага!

1.3 Постоянный шаг

С постоянным шагом всё непросто. Если выбрать его слишком малым, то сходимость будет слишком долгой. Слишком большим - градиентный спуск грозит разойтись. Для обоих тестовых наборов был выбран шаг $c = 0.2$.

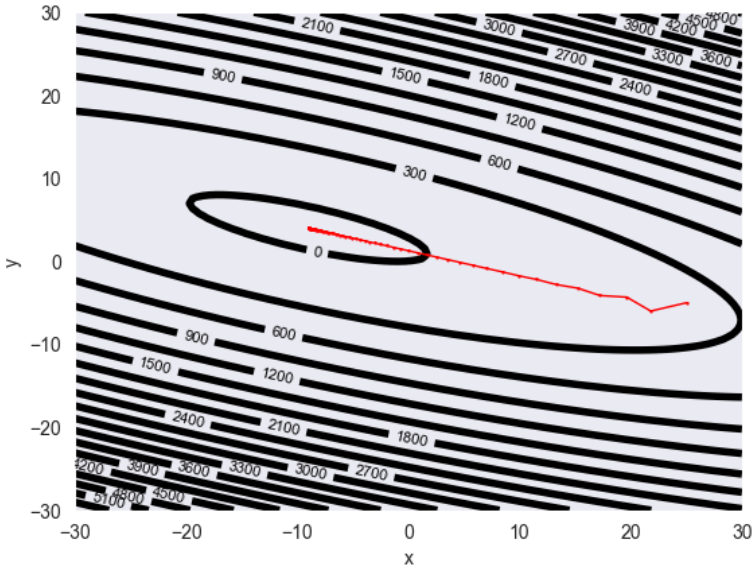

Рис. 5: Траектория градиентного спуска с постоянным шагом на первом тестовом наборе.

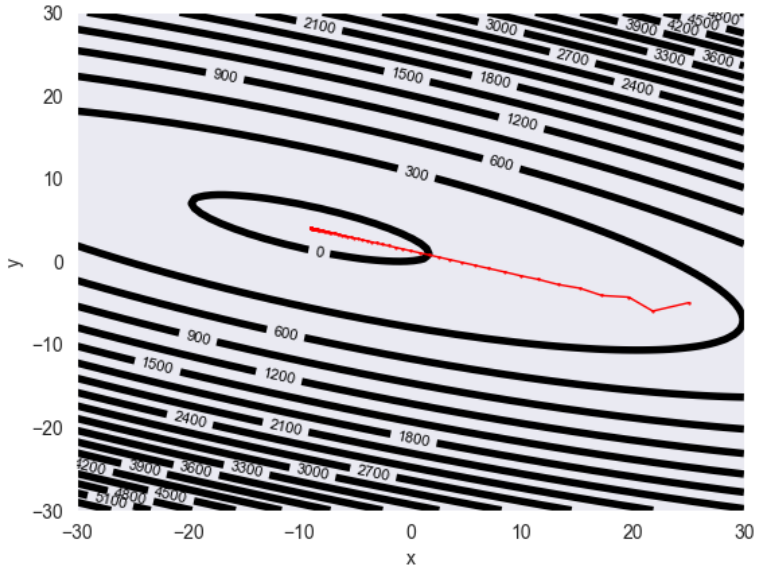

Рис. 6: Траектория градиентного спуска с постоянным шагом на втором тестовом наборе.

1.4 Вывод

В ходе эксперимента на двух тестовых данных была исследована зависимость поведения градиентного спуска от выбора шага. При экспериментах было замечено, что выбор начального приближения очень сильно влияет на количество итераций метода и сходимость в целом. Это, в принципе, логично - чем лучше начальное приближение, тем легче найти минимум функционала. Если задача плохо обусловлена и выберем плохо начальное приближение - получим траекторию «зигзаг» (подсмотрел в лекции ФКН ВШЭ)

Говоря о стратегии выбора шага, лучшей для этих тестов мне показалась стратегия Вульфа: она сходится за наименьшее количество итераций по сравнению с остальными методами. На втором месте метод Армихо - стабильный и сходится. А вот постоянный шаг дает расхождение при значениях $c > 0.2$.

Анализируя траекторию, можно сделать вывод, что при использовании постоянного шага траектория становится гладкой, а при использовании метода Армихо с бектрекингом - довольно шумной. При использовании метода Вульфа шаг становится меньше, а траектория всё еще шумная.

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

В этой задаче продолжается оптимизация функции квадратичного оракула вида:

$$L = \frac{1}{2} \cdot \langle Ax, x \rangle - \langle b, x \rangle$$

Исследуем, как зависит число итераций, необходимое градиентному спуску для сходимости, от следующих двух параметров:
1) числа обусловленности $\kappa \geq 1$ оптимизируемой функции
2) размерности пространства n оптимизируемых переменных

Входные данные для матрицы A будем генерировать случайным образом, по следующему рецепту: возьмем идею о постоении матрицы из методички, а потом поделим ее на число обусловленности, нормировав ее, как предложил Данил Сморчков. Без этого эксперимент с постоянным шагом давал странные результаты. Вектор b возьмем как выборку из нормального распределения для каждой задачи. Начальное приближение я задаю как вектор из единиц с гауссовым нормальным шумом.

Зададим тестировочную сетку параметров, под которые мы будем генерировать тестовые данные. Размерность переберем как степени двойки, число обусловленности - линейно, начиная от 10 и заканчивая 1000. Верхнюю границу числа обусловленности я взял, посоветовавшись с Андреем Широбоковым, так как нужно было найти оптимум между вычислительным временем и размером сетки. Для увеличения точности измерения для каждой «точки» на сетке эксперимент выполнен 10 раз.

2.1 Постоянный шаг

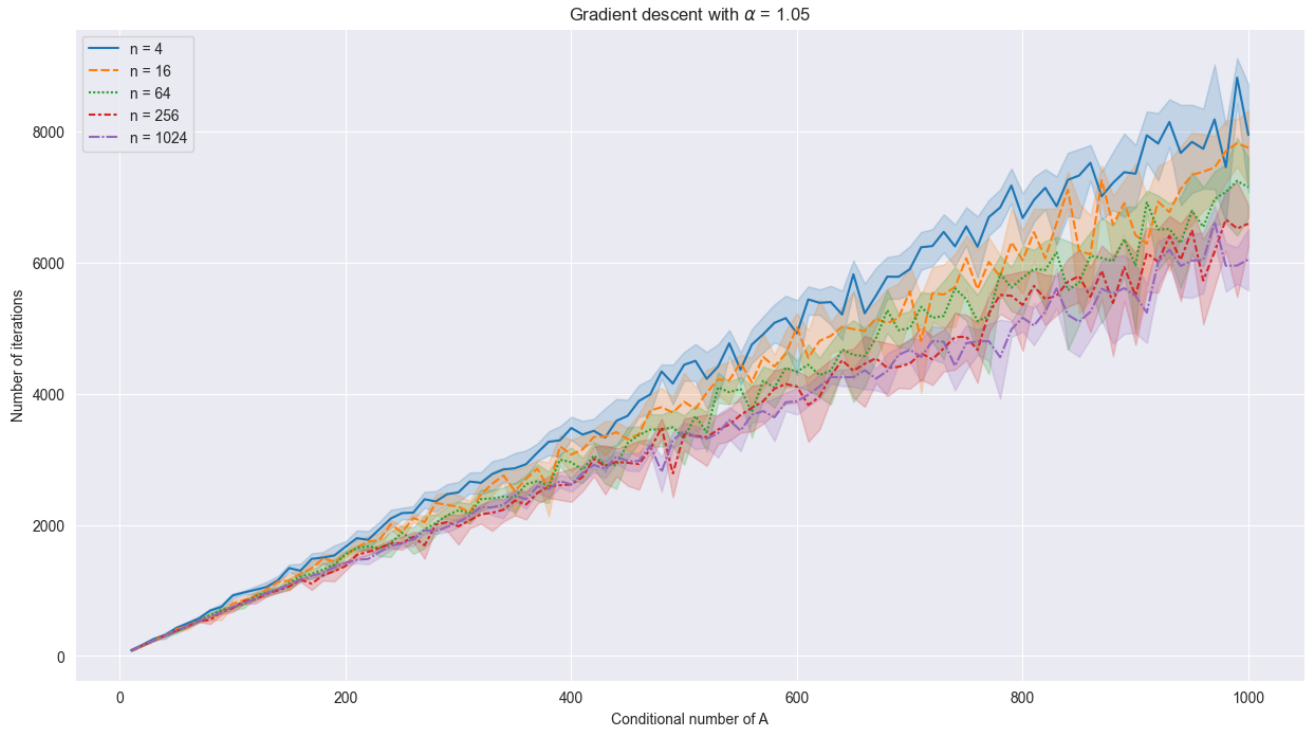

Рис. 7: Зависимость количества шагов от числа обусловленности и размерности пространства для постоянного шага, c = 1.05

Значение шага было подобрано эмпирически, чтобы ни один эксперимент не разошелся в процессе.

Видно, что при увеличении размерности пространства градиентному спуску с постоянным шагом требуется всё меньше итераций для сходимости, но эффект выражен слабо. Также наблюдается четкая линейная зависимость от числа обусловленности при любой размерности пространства: чем хуже обусловлена задача, тем труднее для нее найти оптимум!

2.2 Метод Армихо

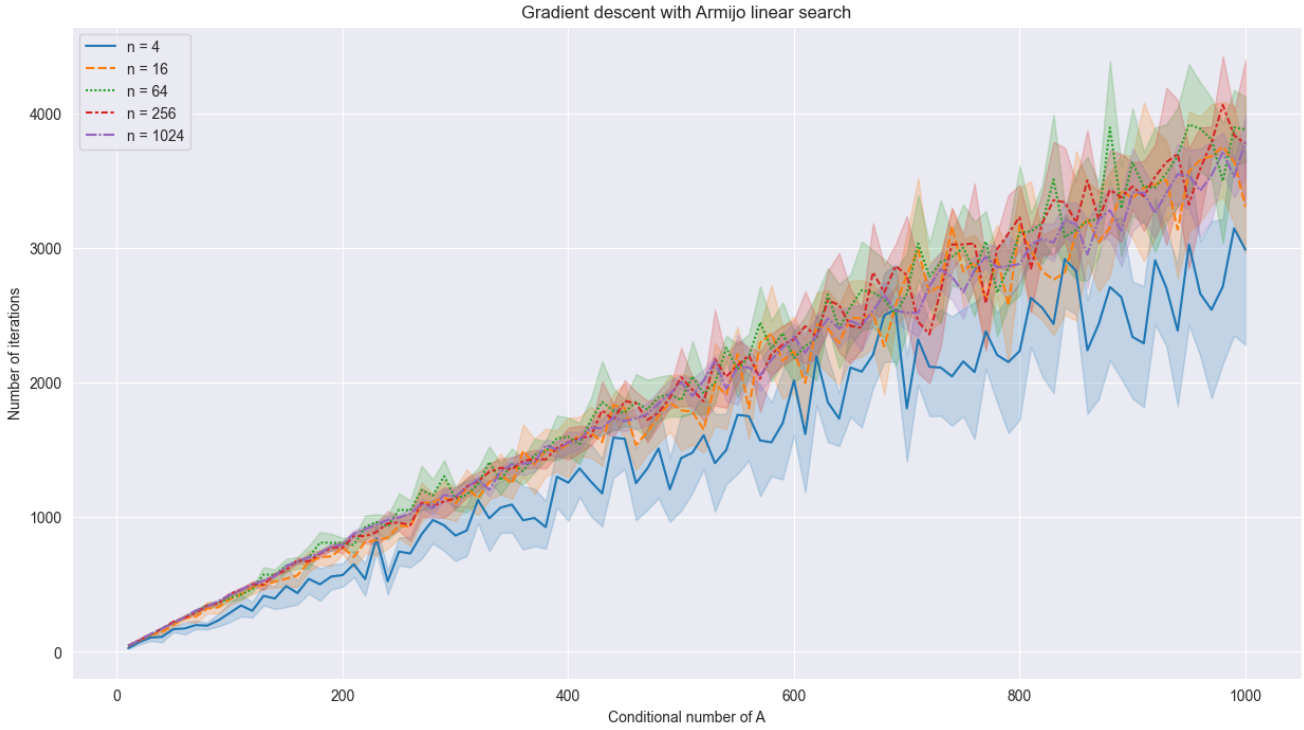

Рис. 8: Зависимость количества шагов от числа обусловленности и размерности пространства для метода Армихо, $c_1 = 1e - 4$

Для метода Армихо ситуация похожа на постоянный шаг: чем хуже обусловлена задача, тем больше итераций требуется методу Армихо для достижения точки оптимума. Открыв лекцию ФКН ВШЭ, я наткнулся, что количество итераций относительно размерности пространства должно расти линейно - как и есть у меня на графике! И, важное замечание: методу Армихо как будто бы всё равно на размерность пространства, или, точнее, количество итераций не увеличивается относительно размерности пространства.

2.3 Метод Вульфа

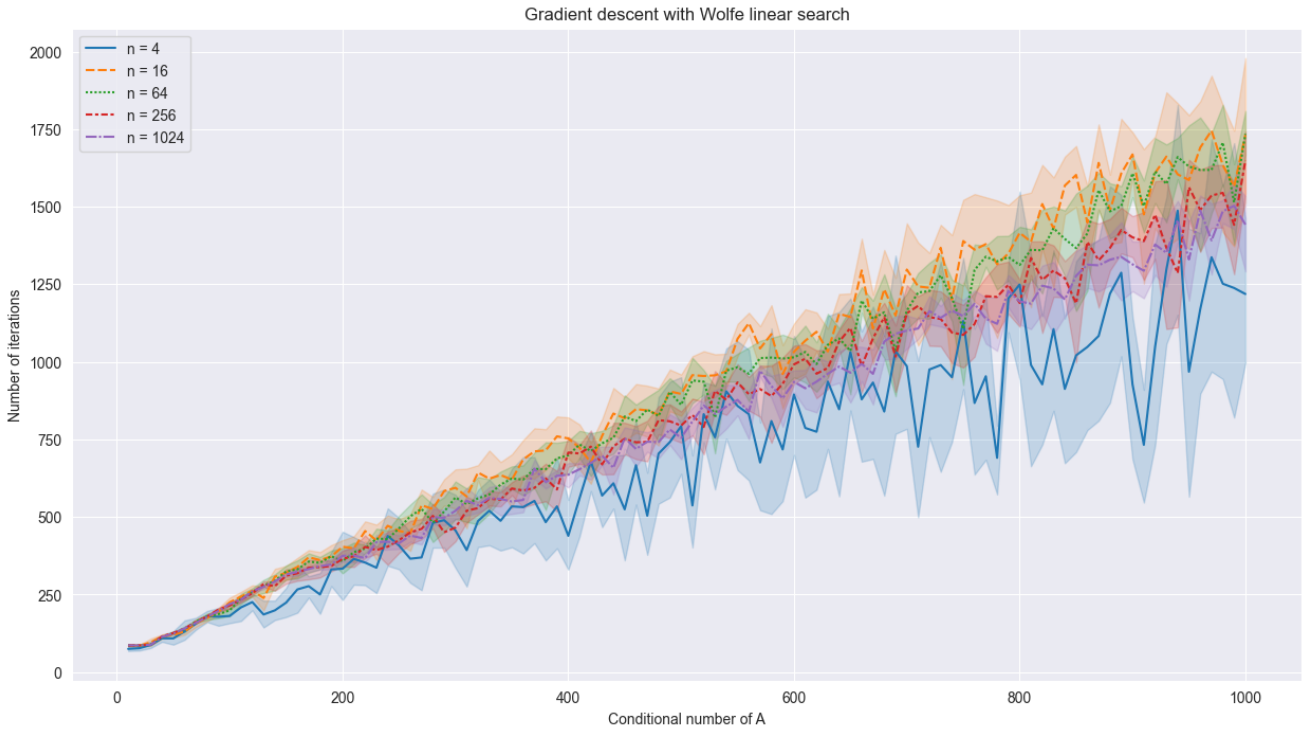

Рис. 9: Зависимость количества шагов от числа обусловленности и размерности пространства для метода Вульфа, $c_1 = 1e - 4, c_2 = 0.9$

Количество итераций при использовании метода Вульфа тоже является линейной функцией от числа обусловленности и не возрастает при увеличении размерности пространства. Хотя видно, что он сходится быстрее метода Армихо почти всегда.

2.4 Вывод

В этом эксперименте мы изучили зависимость количества итераций метода градиентного спуска на квадратичном оракуле с различными методами линейного поиска. Исходя из графиков полученных зависимостей видно, что количество итераций не увеличивается при увеличении размерности пространства для всех трех методов. Для всех трех методов количество итераций линейно растет от числа обусловленности матрицы A.

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

В этом эксперименте нужно сравнить методы градиентного спуска и Ньютона на задаче обучения логистической регрессии на реальных данных. В качестве реальных данных будут использоваться следующие три набора: w8a, gisette и real-sim. Обозначим количество фичей и векторов за за n и m соответственно. Коэффициент взят стандартным образом: $\lambda = \frac{1}{m}$.
Начальной точкой в каждом эксперименте является $\overrightarrow{x_0} = \overrightarrow{0}$.

3.1 Теория

Сперва поговорим про задачу логистической регрессии, которую нам нужно было реализовать чисто векторно. Функцией потерь здесь будет являтся:

$$L(x) = \frac{1}{m} \cdot 1_m \cdot log(1 + exp(-b \odot (A \cdot x)) + \frac{\lambda}{2}||x^2||$$

где 1_m является вектором строчкой из единиц размерности количества точек в датасете, вектор b является вектором-столбцом классов, а $A_{m,n}$ - матрицей признаков, где вектора расположены в строчках. Посчитаем ее градиент:

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

А теперь и Гессиан (подсмотрев в Yandex Handbook):

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

где матрица $S_{m,m}$ является диагональной матрицей с элементами:

$$S = diag \left\{ \frac{1}{1 + exp(b \odot (A \cdot x))} \odot (1 - \frac{1}{1 + exp(b \odot (A \cdot x))}) \right\}$$

Поговорим о методах и стоимости их итераций по времени и по памяти относительно размерность выборки и размерности пространства.

Градиентный спуск использует операцию вычисления градиента, которая требует в среднем $O(mn)$ времени на итерацию и так же $O(n)$ на хранение вектора градиента на каждом шаге.
В свою очередь **Метод Ньютона** использует $O(n^3 + m^2n)$ на вычисление направления, а по памяти - $O(n^2)$ опять же из-за Гессиана.

3.2 Проведение экспериментов

Для каждого из трех тестовых датасетов мы будем обучать логистическую регрессию с регуляризацией двумя способами - методом Градиентного спуска и методом Ньютона с различными стратегиями линейного поиска.

По заданию, точность метода Ньютона была выставлена $1e-9$, а градиентного спуска - $1e-5$, так как он дольше сходится. Для каждого датасета считались нормы градиента на каждом шаге, значение функции потерь и время, затраченное с начала запуска. Константы для методов были взяты стандартно ($c_1 = 1e-4, c_2 = 0.9$), за исключением постоянного шага - там взята $c = 1$ для квадратичной сходимости метода Ньютона.

3.2.1 W8a dataset

Сам по себе датасет маленький, считается быстро. В итоге получились следующие зависимости:

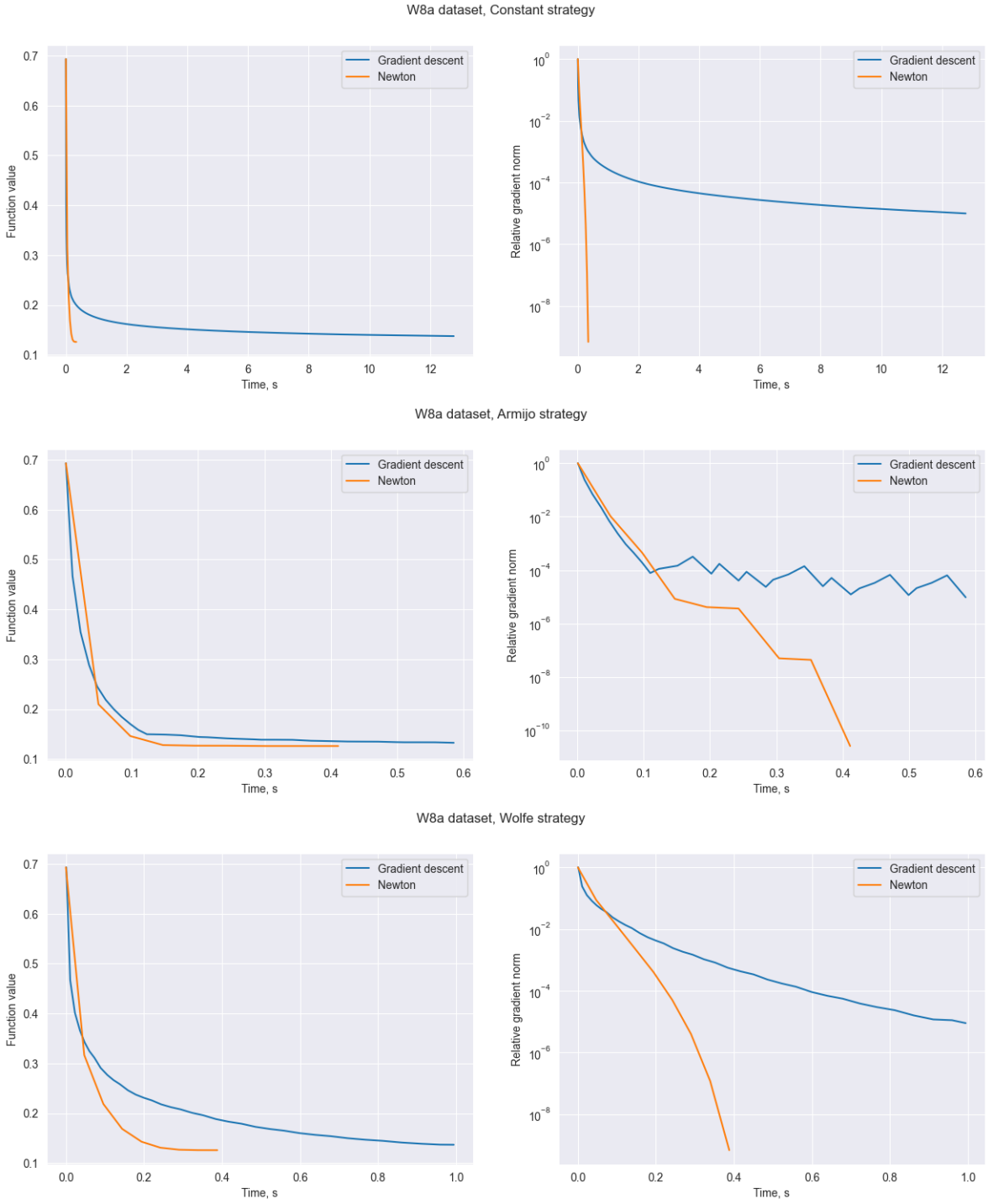

Рис. 10: Зависимость значения функции потерь в линейном масштабе от времени (слева); Зависимость относительного квадрата нормы градиента от времени в логарифмическом масштабе (справа)

В итоге видим, что для этого датасета наиболее быстрым оказывается метод Ньютона, хоть и точность у него выставлена больше. Как говорится в методичке ФКН ФШЭ, метод Ньютона не сильно чувствителен к методу выбора шага, давая оптимальную сходимость. А вот градиентный спуск, напротив - при постоянном шаге он сходится неприлично долго, а вот при адаптивных методах - намного быстрее.

3.2.2 Gisette dataset

Этот датасет является самым большим по весу файла и по времени исполнения и считался довольно долго...

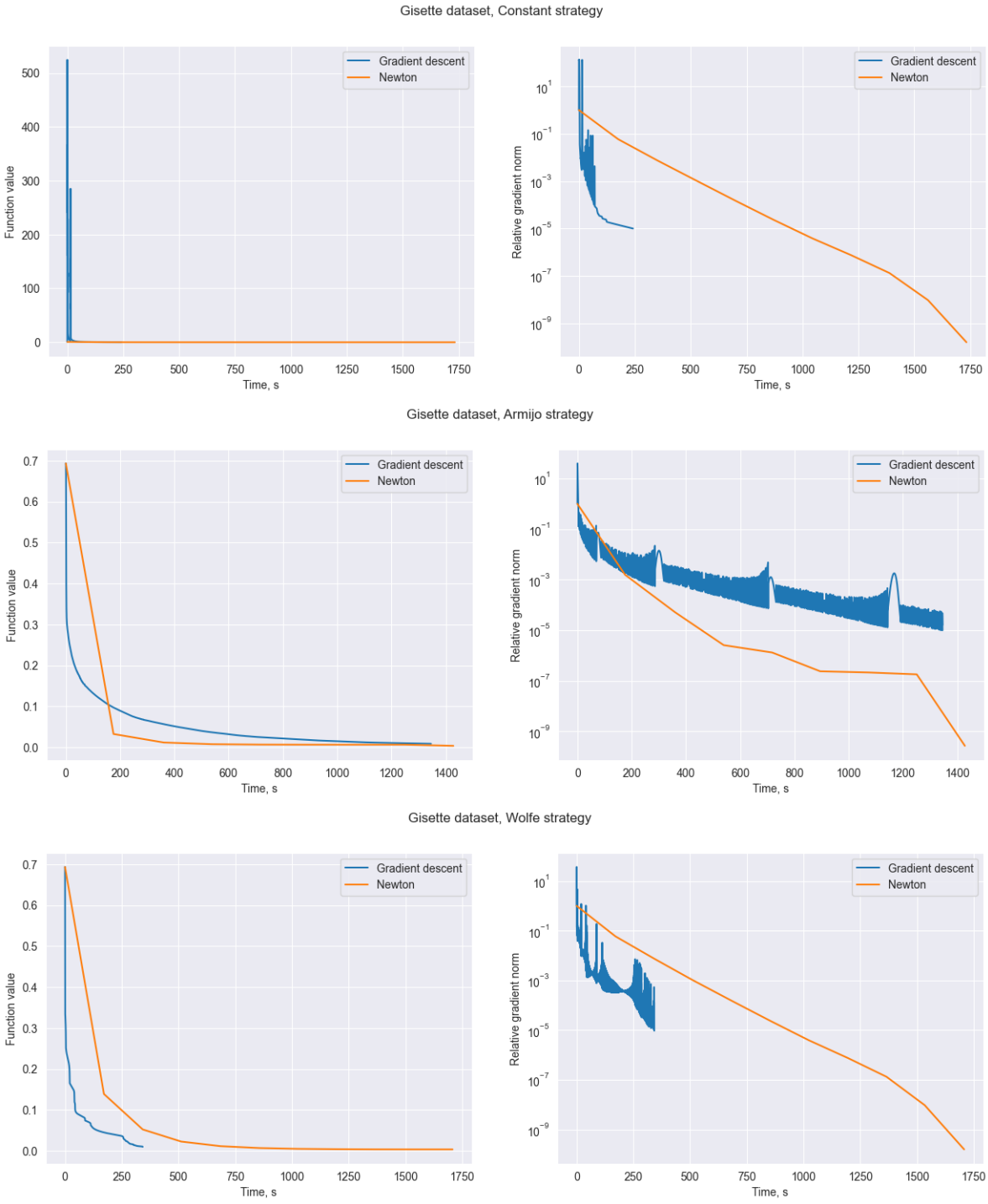

Рис. 11: Зависимость значения функции потерь в линейном масштабе от времени (слева); Зависимость относительного квадрата нормы градиента от времени в логарифмическом масштабе (справа)

На этот датасете проявил себя Градиентный спуск с постоянным шагом или стратегией Вульфа - он оказался **невероятно** быстрым. Но при использовании метода он существенно замедлился. Но есть один минус - градиентный спуск довольно «шумный» по квадрату нормы градиента, и в какой-то момент мне казалось, что он вообще разойдется. Метод Ньютона, напротив, вообще не был шумным, но зато одна его итерация длится порядка 250 секунд - непозволительно много...

3.2.3 Real-sim dataset

Хоть датасет находится на втором месте по весу файла из-за разреженности матрицы, по количеству фичей и точек - определенно на первом.

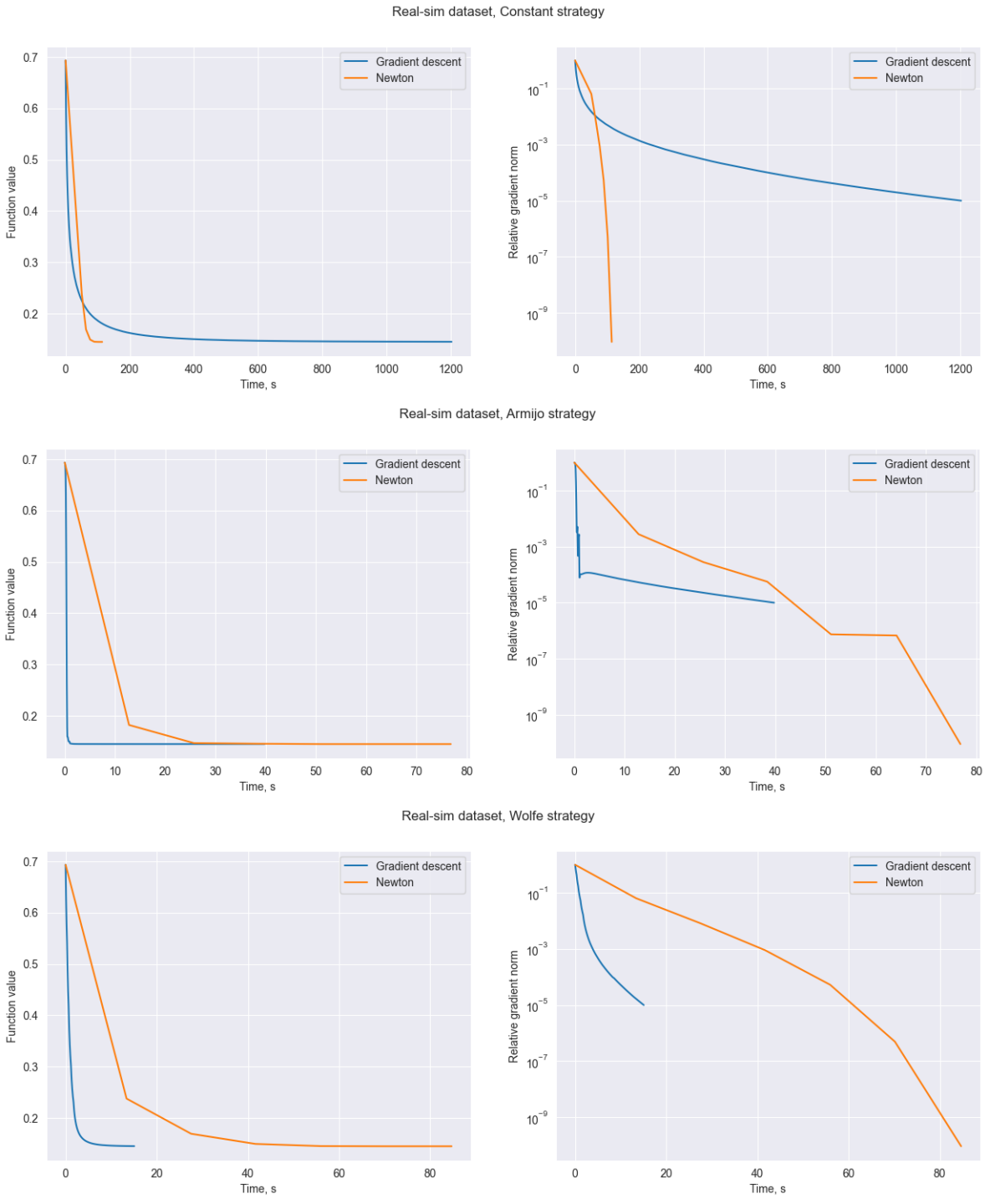

Рис. 12: Зависимость значения функции потерь в линейном масштабе от времени (слева); Зависимость относительного квадрата нормы градиента от времени в логарифмическом масштабе (справа)

Лучшей стратегией для градиентного спуска на этом датасете оказывается метод Вульфа, как и для метода Ньютона.

3.3 Выводы

Для трех датасетов была обучена логистическая регрессия с тремя различным стратегиями выбора шага. Видно, что более простой метод (GD) сходится быстрее в области низкой точности, а более сложный (Ньютон) - наоборот. Также, говоря о сходимости, Ньютон показывает квадатичную сходимость в окрестности решения, а Градиентный спуск - линейную. Как должно быть в теории - так и выходит.

Какой из методов лучше и в каких ситуациях?

Нужно смотреть на ситуацию. Если нам нужна небольшая точность - то стоит выбирать градиентный спуск со стратегией Вульфа и будет счастье. Если же интересует большая точность, то нужно посмотреть на датасет. Если размерность пространства в нем большая, то стоит воспользоваться Градиентным спуском опять же со стратегией Вульфа (Так как она показала себя намного более робастной, чем все остальные). Если маленькая - то определенно метод Ньютона! А можно вообще сделать хитрее: запустить Градиентный спуск с методом Вульфа до какой-то небольшой точности, а потом запустить метод Ньютона тоже со стратегий Вульфа из точки остановки градиентного спуска. В итоге мы довольно быстро пройдем шаги с линейной сходимостью, а потом заимеем квадратичную!

4 Стратегия выбора длины шага в градиентном спуске

Здесь исследуем, как зависит поведение метода градиентного спуска от стратегии подбора шага: константный шаг (попробовать различные значения), бэктрэкинг (попробовать различные константы с), условия Вульфа (попробовать различные параметры c2). В качестве модельных данных будем рассматривать как квадратичного, так и LogReg оракулов, матрицы для которых генерируем случайным образом из стандартного нормального распределения. Также сразу оговорюсь, хоть в задании и было сказано делать всё на одном графике, я сделал отдельные графики для большей читаемости.

4.1 Теория

Здесь нам нужно рисовать кривые сходимости, а именно - относительная невязка от числа итераций. Определим относительную невязку как $|\frac{f(x)-f(x^*)}{f(x^*)}|$. Хорошо, что мы можем посчитать истинное значение минимума ($f(x^*)$) аналитически для квадратичного оракула. Если взять производную и подставить, получим что наше значение будет: $f(x^*) = -\frac{b^T \cdot A^{-1} \cdot b}{2}$.

Для экспериментов я сгенерировал следующие данные:

1. Квадратичный оракул
    - Матрица A размером $2000 \cdot 2000$ из $N_{0,1}$ с диагональным преобладанием, чтобы обратная матрица существовала и задача была довольно хорошо обусловлена
    - Вектор b размером $2000 \cdot 1$ из того же распределения
    - Начальное приближение $\vec{x_0} = 1_n$ - вектор из единиц
2. LogReg оракул
    - Тестовый датасет с данными из $N_{0,1}$ размером в m векторов и n фичей, который был преобразован в разреженную матрицу плотности 0.8.

- Вектор b размером $2000 \cdot 1$ с метками классов: 1 и -1 соответственно
- Начальное приближение $\vec{x_0}$ - вектор из $N_{0,1}$
- Коэффициент регуляризации $\lambda = 1e-3$

Критерием остановки для GD в случае Квадратичного оракула является точность $1e-16$, для LogReg - $1e-12$ соответственно.

4.2 Метод Вульфа

Здесь я подбирал всяческие константы и начальные приближения, но метод Вульфа оказался самым робастным из всех - решает задачу за минимальное количество шагов, какой параметр ему не ставь... Видно, что стратегия выбора константы сильно зависит от задачи, так как судя по квадратичному оракулу лучшим оказость значение 0.3, а на LogReg - вообще $1e-3$!

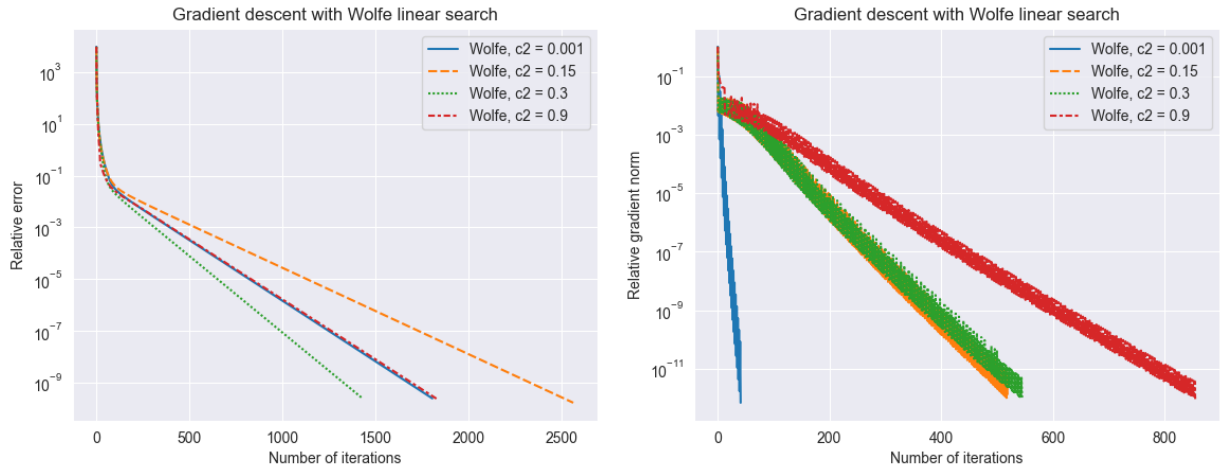

Рис. 13: Зависимости относительной ошибки (в линейном масштабе) и квадрата относительной нормы градиента для стратегии Вульфа

4.3 Метод Армихо

По сравнению с методом Вульфа, Армихо - вообще не почувствовал смены константы. Это видно на графике с относительной ошибкой на Квадратичном

оракуле. С другой стороны, это хорошо, ведь можно поставить значение по умолчанию и не сильно переживать за результат. Градиент шумный, как и в Вульфе, но период осцилляций тут заметно меньше.

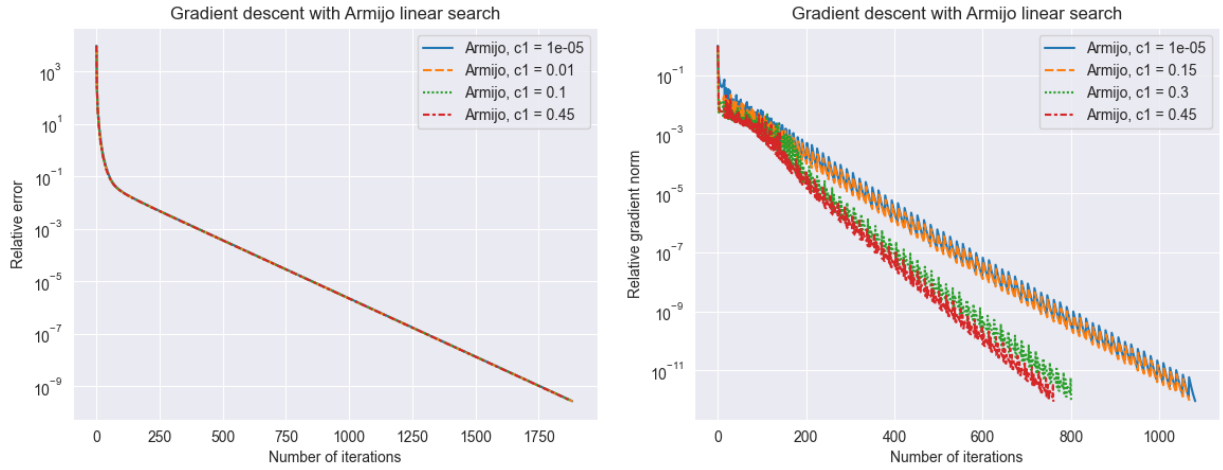

Рис. 14: Зависимости относительной ошибки (в линейном масштабе) и квадрата относительной нормы градиента для стратегии Армихо

4.4 Постоянный шаг

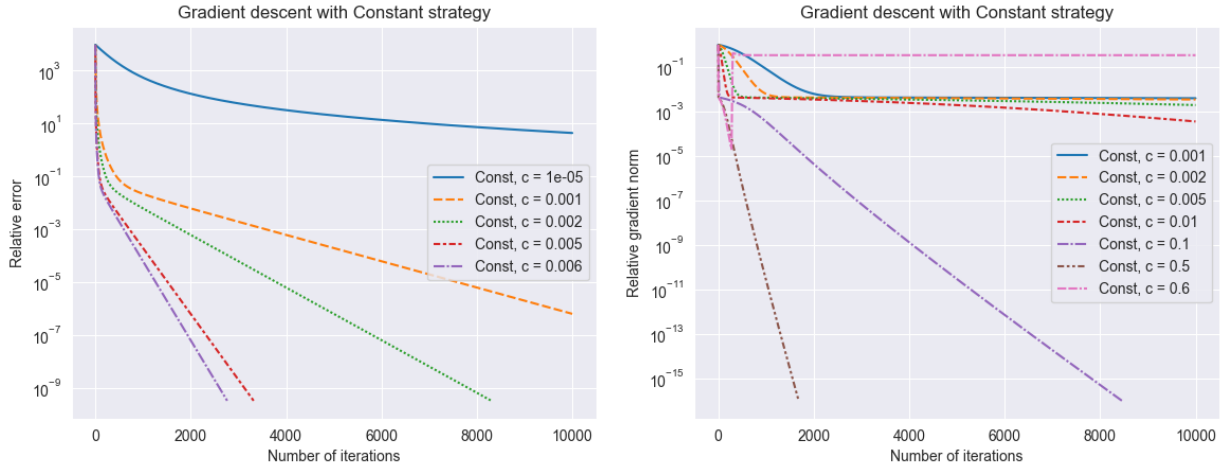

Рис. 15: Зависимости относительной ошибки (в линейном масштабе) и квадрата относительной нормы градиента для постоянного шага

Здесь не получилось найти ту константу, которая бы приводила к бесконечной осцилляции: получалось что метод расходился и такой график был

уже тяжело интерпретируем. Видно, что при выборе различных значений постоянного шага градиентный спуск может как сойтись за требуемое количество итераций, так и нет. Видно, что для каждой задачи нужен конкретный анализ, чтобы найти оптимальное значение шага.

4.5 Выводы

Исходя из полученных результатов можно сделать вывод, что для метода градиентного спуска предпочитаемыми стратегиями являются метод Вульфа и метод Армихо, с приоритетом метода Вульфа. Они обе показали себя довольно устойчивыми к изменению параметров и показали наилучшую сходимость по количеству итераций, затраченных на достижение требуемой точности.

5 Стратегия выбора длины шага в методе Ньютона

Теория этого эксперимента аналогична теории, изложенной выше, за исключением того, что мы будем наблюдать только за LogReg оракулом и относительной нормой градиента функции потерь. Также, начальным приближением в этой задаче был выбран вектор $\vec{x_0}$, полученный как выборка из стандартного нормального распределения. Требуемая точность для метода Ньютона была выставлена равной $1e-16$.

5.1 Метод Вульфа

Видно, что метод Ньютона отлично сочетается со стратегией Вульфа - она уменьшает единичный шаг, если они не удовлетворяют условию, и оставляют как есть, если всё хорошо. Это и дает нам квадратичную сходимость в области решения. На графике мы видим, что есть совсем незначительные отличия в количестве итераций в зависимости от выбора константы. При выборе слишком большой константы квадратичная сходимость набирается за бОльшее количество итераций.

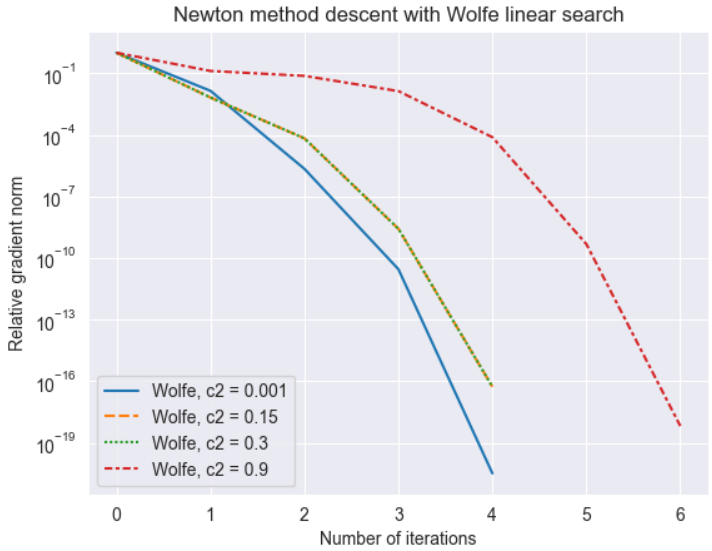

Рис. 16: Зависимость квадрата относительной нормы градиента для стратегии Вульфа

5.2 Метод Армихо

Метод Армихо очень похож по поведению на метод Вульфа, за исключением особенности - при малой константе он способен заставить метод Ньютона делать бесполезные итерации, которые почти не снижают невязку...

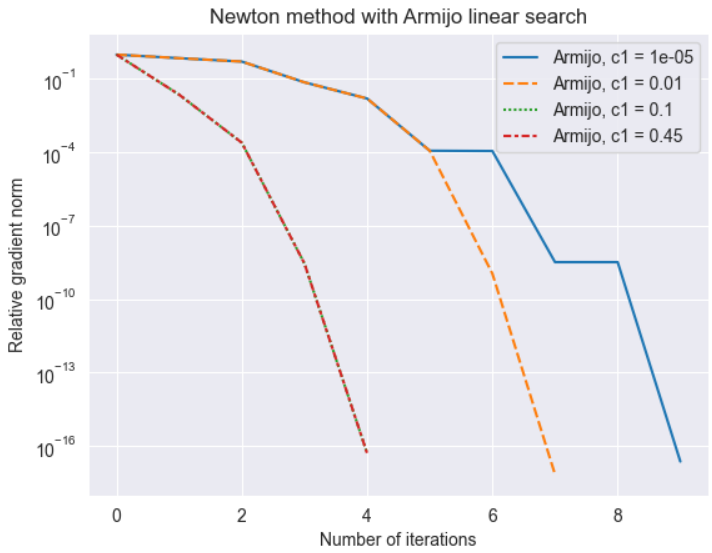

Рис. 17: Зависимость квадрата относительной нормы градиента для стратегии Армихо

5.3 Постоянный шаг

Именно из-за этого примера я начал думать об изменении начального приближения - от него ОЧЕНЬ многое зависит, если не использовать линейный поиск шага. И то, у меня вышло поймать константу, которая хотя бы дает какую-то сходимость, а не колебания. При выборе начальной точки, равной нулю, всё получается и метод сходится к требуемой точности, но я оставил это как пример того, что без линейного поиска полагаться на метод Ньютона выходит не всегда!

5.4 Выводы

Исходя из результатов эксперимента можно сделать вывод, что метод Ньютона лучше всего работает с методом Вульфа. Метод Армихо тоже может быть

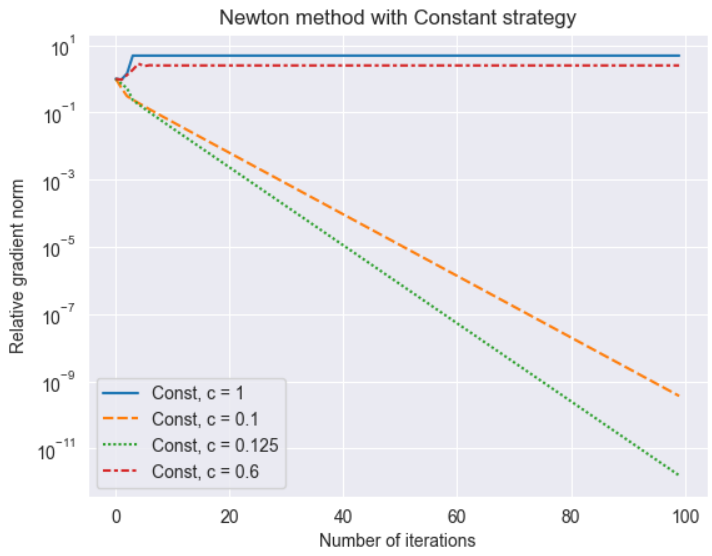

Рис. 18: Зависимость квадрата относительной нормы градиента для постоянного шага

применим, но требует больше итераций. А итерации в методе Ньютона стоят очень дорого, так как нужно вычислять Гессиан. Стратегия с постоянным шагом на тестовой задаче вообще оказалась неприменима в той начальной точке, которую я выбрал. Хотя, на датасете Gisette стратегия с постоянным шагом ничем не уступала остальным. Если говорить о надежности и робастности методов, то я бы выбирал стратегию линейного поиска Вульфа со страндартными параметрами.