3. Формулы для логистической регрессии

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

$$\nabla L = -\frac{1}{m} \left\langle A^T, b \odot k \right\rangle + \lambda x,$$

где k - вектор с компонентами $k^{(i)} = \dfrac{1}{1 + \exp\left(b_i \langle a_i, x \rangle\right)}$.

$$\nabla^2 L = \frac{1}{m} A^T D\left(p \odot k\right) A + \lambda I_n,$$

где $D\left(p \odot k\right)$ - диагональная матрица с $p \odot k$ на диагонали, p - вектор с компонентами $p^{(i)} = \dfrac{1}{1 + \exp\left(-b_i \langle a_i, x \rangle\right)}$.

3.1 Эксперимент: Траектория градиентного спуска на квадратичной функции

Сначала выберем 3 квадратичные матрицы:

$$A_1 = \begin{pmatrix} 2.0 & 0.0 \\ 0.0 & 2.0 \end{pmatrix}; \quad A_2 = \begin{pmatrix} 2.0 & 0.0 \\ 0.0 & 12.0 \end{pmatrix}; \quad A_3 = \begin{pmatrix} 2.0 & 1.0 \\ 1.0 & 2.0 \end{pmatrix}$$

С общим параметром b:

$$b = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$$

Начальные точки для дальнейших экспериментов возьмем такие:

$$x_0^{(1)} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}; \quad x_0^{(2)} = \begin{pmatrix} -5 \\ -3 \end{pmatrix}$$

3.1.1 Сравнение разных стратегий выбора шага

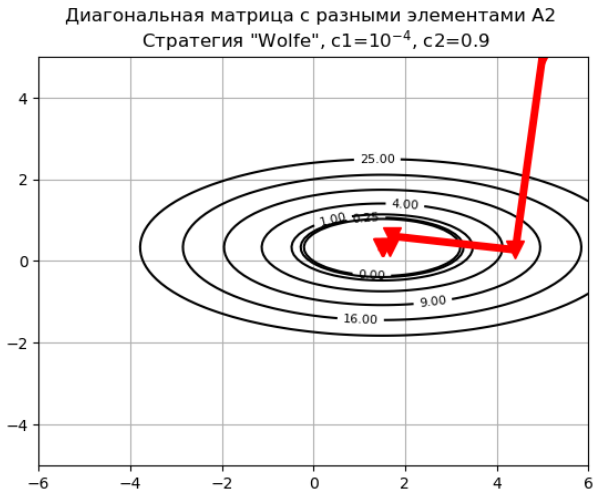


Рисунок 1. Сравнение разных стратегий выбора шага для A2, b = (3, 4), x = (5, 5)

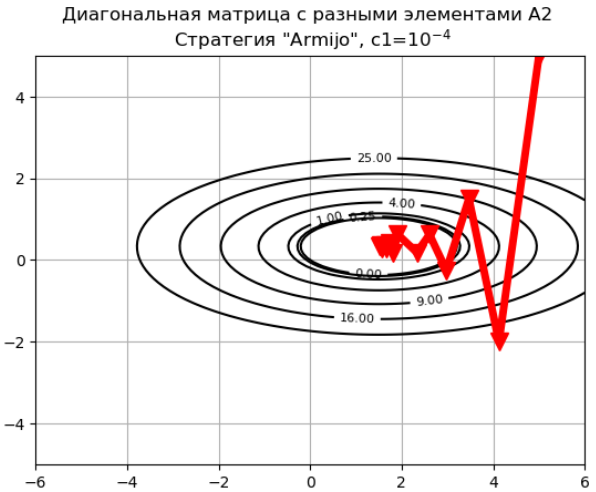


Рисунок 2. Сравнение разных стратегий выбора шага для A2, b = (3, 4), x = (5, 5)

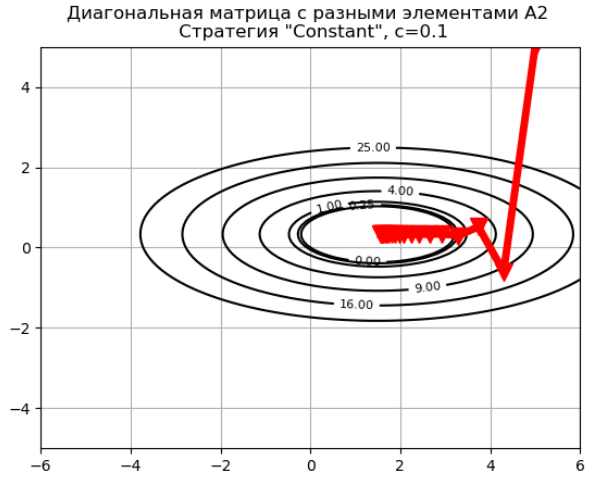


Рисунок 3. Сравнение разных стратегий выбора шага для A2, b = (3, 4), x = (5, 5)

Вывод по эксперименту:

Как можно увидеть на первых трех рисунках, все три метода успешно сходятся к минимуму функции, но делают это по-разному. Стратегия Вульф показывает себя «золотой серединой» с точки зрения баланса между скоростью и точностью. Здесь используется наиболее строгое ограничение на длину шага. Она хорошо подойдет и для более сложных задач, где может потребоваться высокая точность. Стратегия Армихо также эффективна ввиду ограничения на длину шага, но менее точна. Требуется большее количество итераций, чтобы «попасть» в минимум. Константная стратегия наиболее прямолинейная и наименее адаптивная. Если шаг будет маленьким, то это приведет к большому количеству итераций в попытках достичь минимума. Соответственно, при слишком большом шаге может не быть сходимости

3.1.2 Сравнение работы методов для матриц с разной обусловленностью

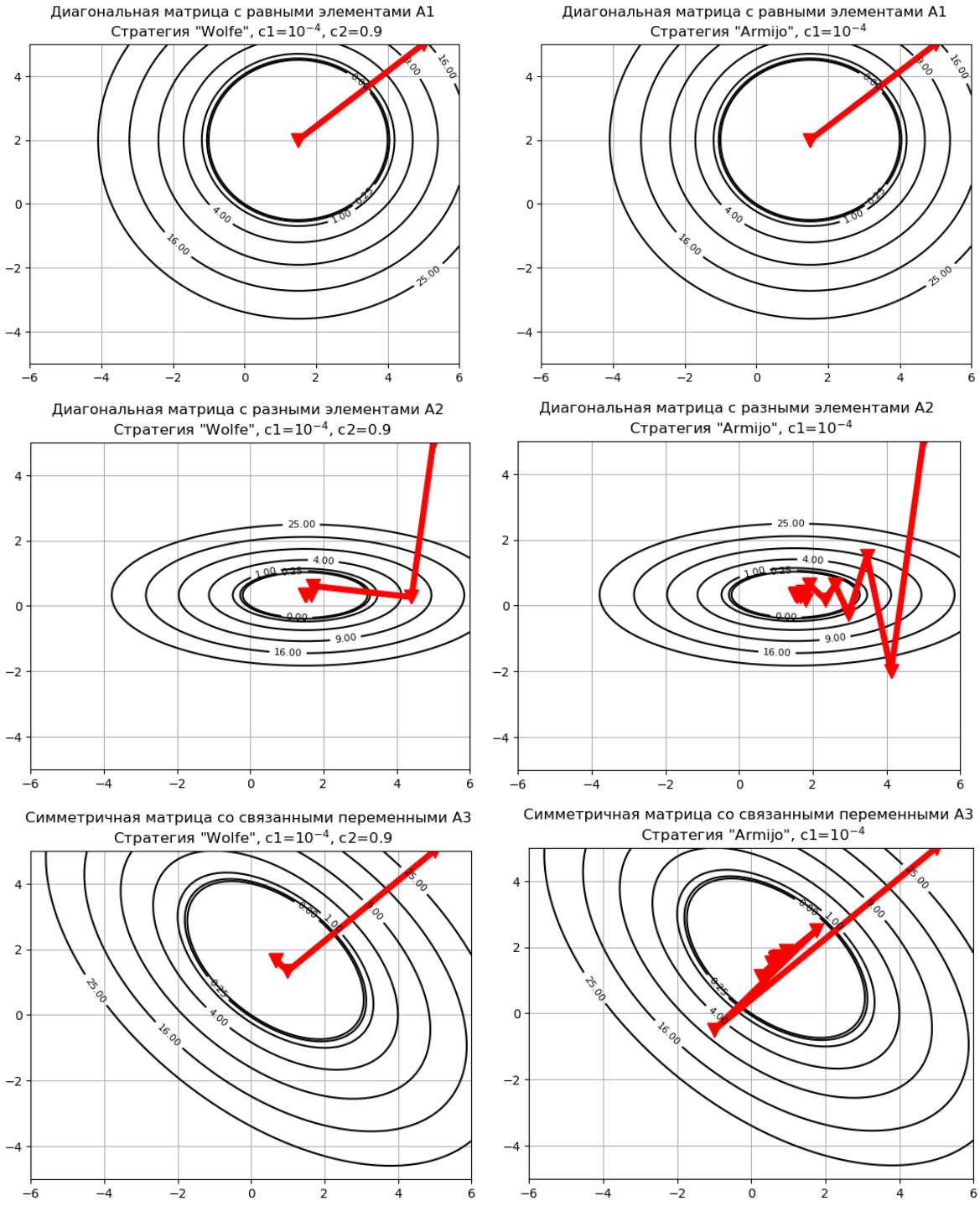


Рисунок 3. Сравнение работы методов для матриц A1, A2, A3

Вывод по эксперименту:

Все методы сходятся к минимуму, но Вульф показывает со всеми матрицами лучшую производительность и стабильность. Армихо показывает большее количество колебаний, особенно с более сложными матрицами: диагональной матрицей с разными элементами и симметричной матрицой со связанными переменными. Чем больше обусловленность – тем больше итераций. Также стоит отметить наклон осей эллипсоидов матрицы А3. Это связано с корреляцией переменных, что приводит к повороту главных осей эллипсоидов относительно координатных осей

3.1.3 Сравнение при различных начальных точках

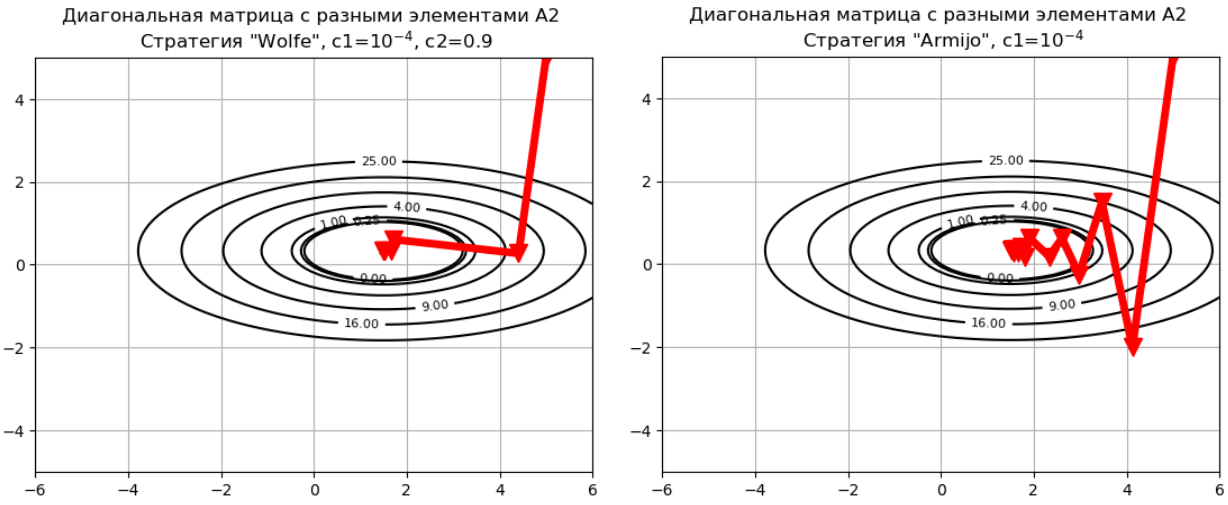


Рисунок 4. Сравнение при x = (5, 5)

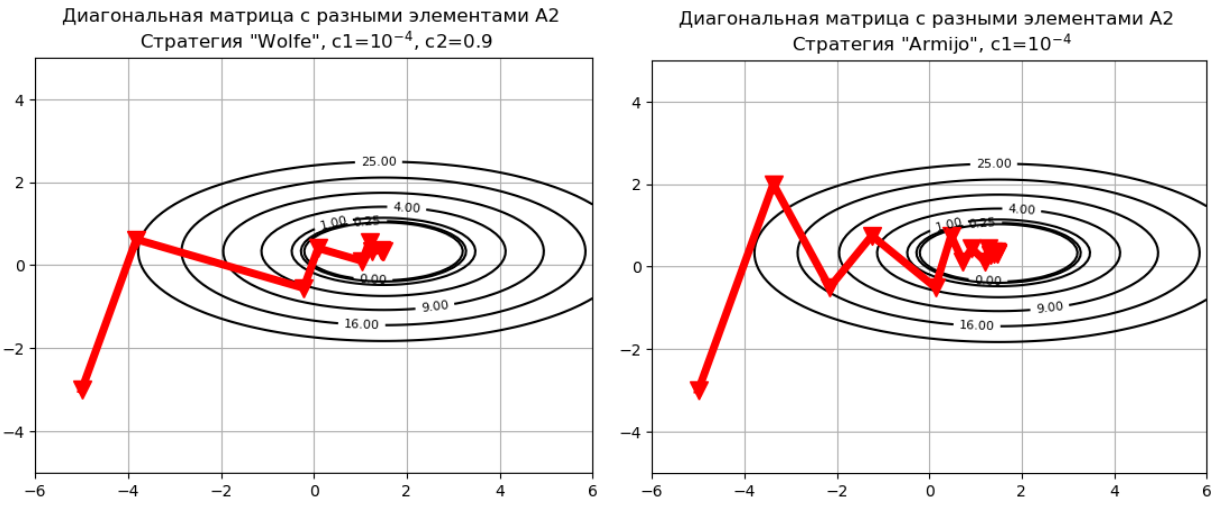


Рисунок 5. Сравнение при x = (-5, -3)

Вывод по эксперименту:

Вульф демонстрирует высокую устойчивость и эффективность независимо от начальной точки. Армихо более чувствителен к выбору начальной точки. При начальной точке x0 = (-5, -3) метод демонстрирует значительное количество колебаний и требует больше итераций для достижения минимума. Можно заключить, что выбор начальной точки важен для эффективности. При удачном выборе начальной точки, которая находится перпендикулярно линиям уровня функции, метод градиентного спуска движется прямо к минимуму. Это минимизирует количество колебаний и обеспечивает более прямолинейный путь к цели. Также начальная точка, которая находится близко к минимуму, позволяет методу быстрее достичь его, так как начальный градиент будет направлен прямо к цели. Это уменьшает количество итераций, необходимых для достижения минимума

3.2 Эксперимент: Зависимость числа итераций градиентного спуска от числа обусловленности и размерности пространства

Задача:

исследовать, 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.

Дано:

стратегия Вульф при c1 = 10^-4, c2 = 0.9, диагональная матрица размерности n x n со случайными значениями на диагонали, минимальное значение равно 1, максимальное = k. Начальное приближение – вектор из единиц x n

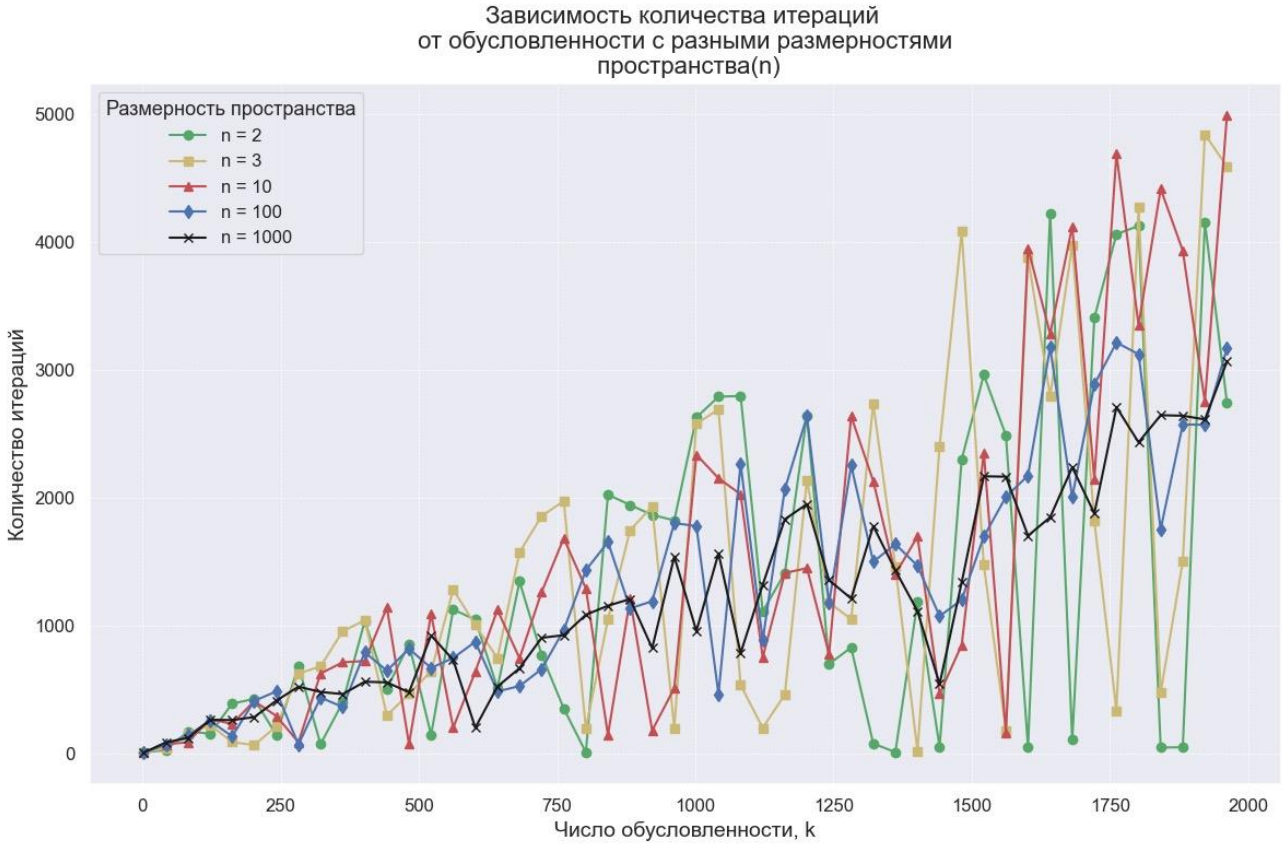


Рисунок 6. Зависимость числа итераций от размерности пространства n и числа обусловленности k

Вывод:

Высокая обусловленность матрицы (разница между максимальным и минимальным собственными значениями) усложняет задачу оптимизации. Градиенты становятся менее информативными, что увеличивает количество итераций, необходимых для достижения сходимости. Хотя увеличение размерности пространства в среднем увеличивает количество итераций, влияние размерности не всегда линейно. Существует множество факторов, включая конкретные значения элементов на диагонали матрицы A, начальные условия и стратегию выбора шага, которые могут влиять на количество итераций. Следовательно, эффект размерности пространства не всегда однозначен и зависит от другий условий

3.3 Эксперимент: Сравнение методов градиентного спуска и Ньютона на реальной задаче логистической регрессии

Задача:

Сравнить методы градиентного спуска и Ньютона на задаче обучения логистической регрессии на реальных данных. В качестве реальных данных использовать следующие три набора с сайта LIBSVM w8a, gisette и real-sim

Дано:

Стратегия Вульф при c1 = 10^-4, c2 = 0.9 из нулевого вектора с коэффициентом регуляризации $\lambda = 1/m$.

Dataset «w8a»:

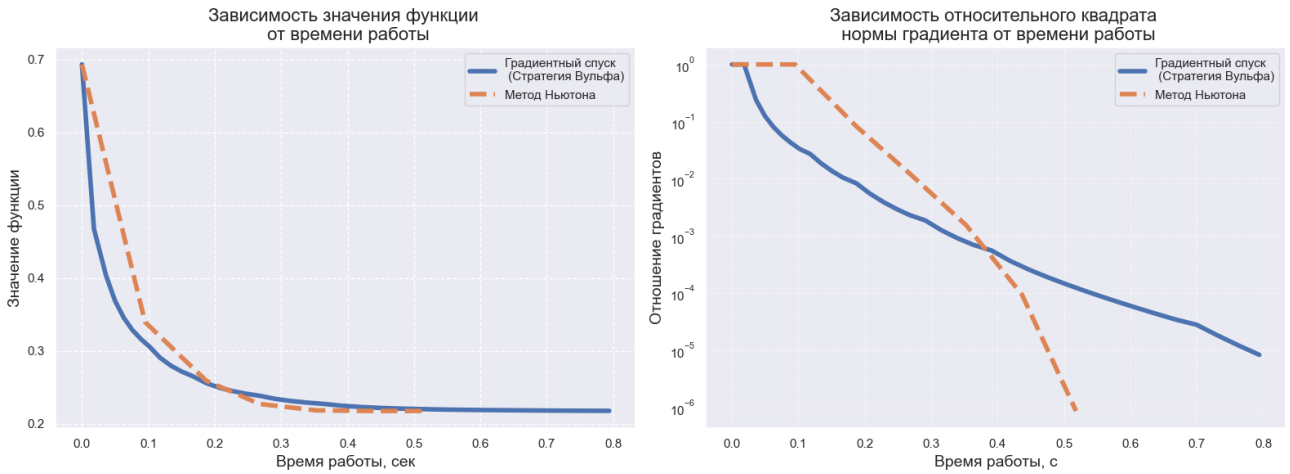


Рисунок 7. Сравнение Wolfe и Newton по скорости сходимости обучения лог. регрессии

Dataset «real-sim »:

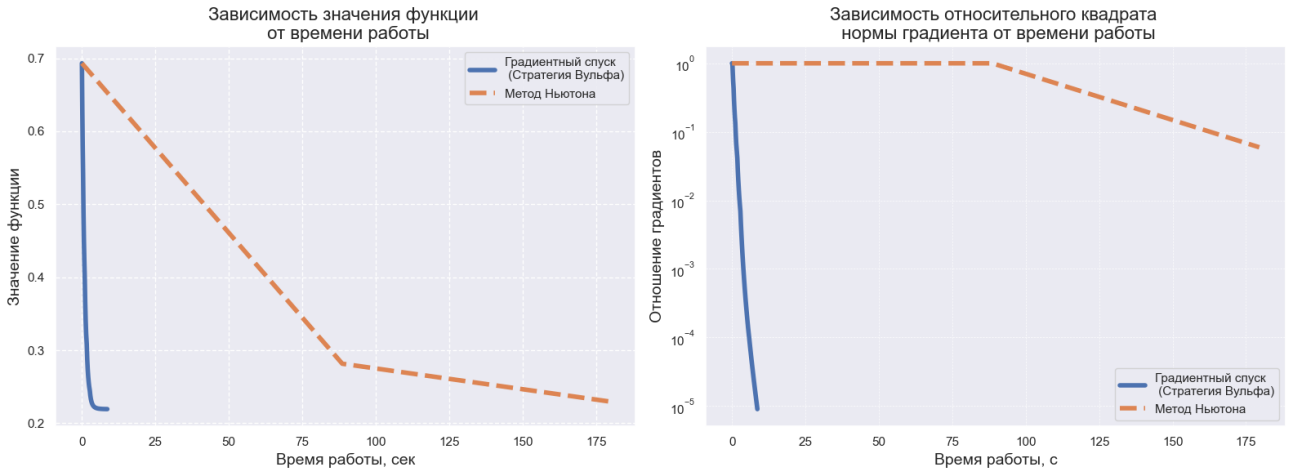


Рисунок 8. Сравнение Wolfe и Newton по скорости сходимости обучения лог. регрессии

Dataset «gisette»:

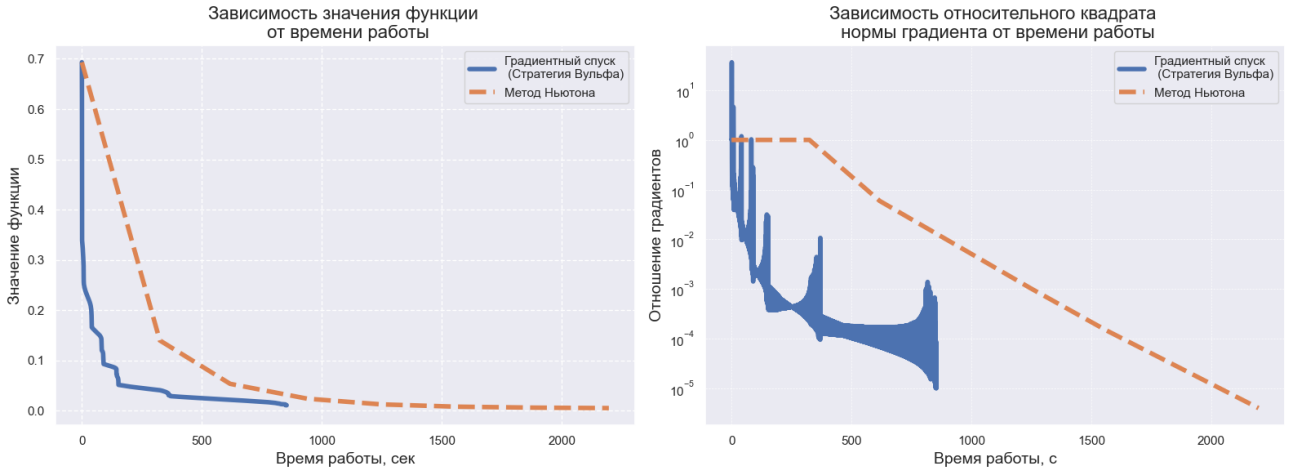


Рисунок 9. Сравнение Wolfe и Newton по скорости сходимости обучения лог. регрессии

Вывод:

Метод Ньютона сходится быстрее и за меньшее число итераций на небольших датасетах, таких как w8a. Это связано с тем, что метод Ньютона использует информацию о втором порядке, что ускоряет сходимость. На больших же датасетах (оставшиеся два) метод Ньютона также сходится за меньшее количество итераций, но требует гораздо больше времени. Это объясняется высокой вычислительной сложностью. Градиентный же спуск хоть и требует больше итераций, имеет меньшую вычислительную сложность $O(n)$ по памяти и $O(nm)$ по времени

Метод Ньютона:

Сложность по памяти: $O(n^2)$

Сложность по времени: $O(n^3 + m^2n)$

Градиентный спуск:

Сложность по памяти: $O(n)$

Сложность по времени: $O(nm)$

3.4 Эксперимент: Стратегия выбора длины шага в градиентном спуске

Задача:

Исследовать, как зависит поведение метода от стратегии подбора шага. Рассмотреть квадратичную функцию и логистическую регрессию с модельными данными (сгенерированными случайно). Рассматривается разное начальное приближение для квадратичной задачи и оптимизируется при различных(трех) стратегиях. Для решения генерируется случайная матрица с n =10, m = 30 с элементами из промежутков [-10, 10], [-1, 1], [0, 1] со случайным вектором b

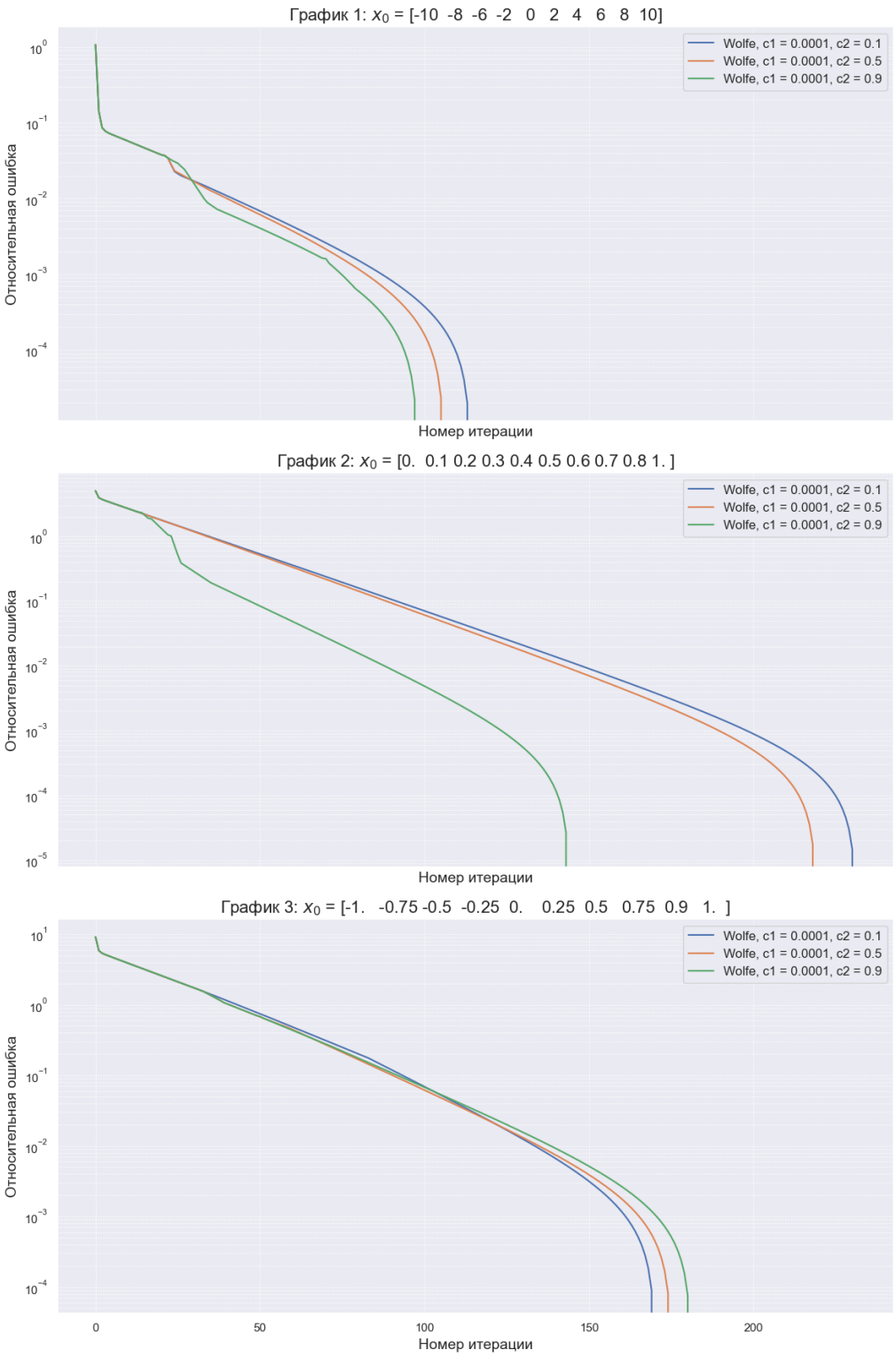


Рисунок 10. Метод Wolfe с разными параметрами для различных начальных точек

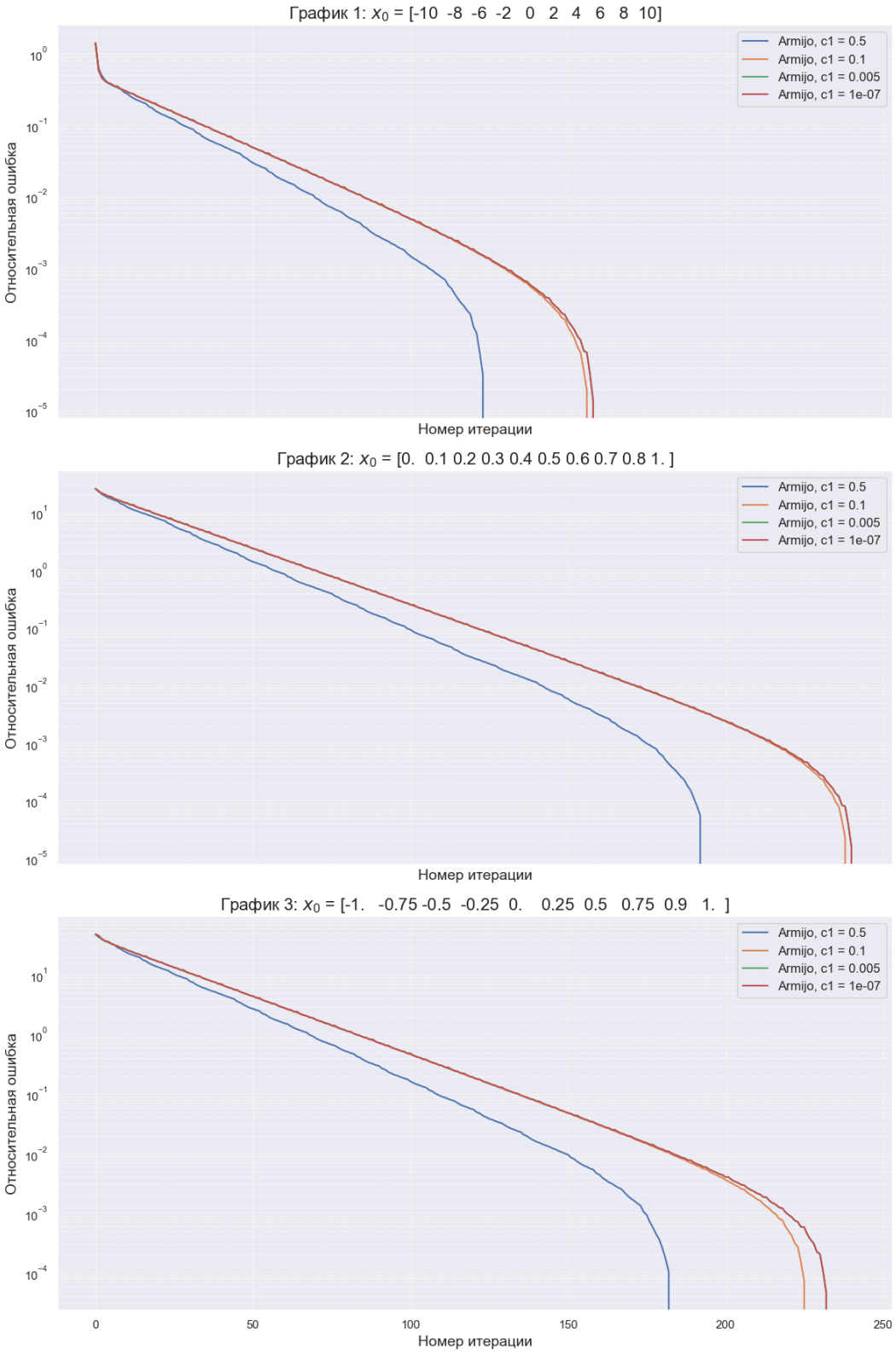


Рисунок 11. Метод Armijo с разными параметрами для различных начальных точек

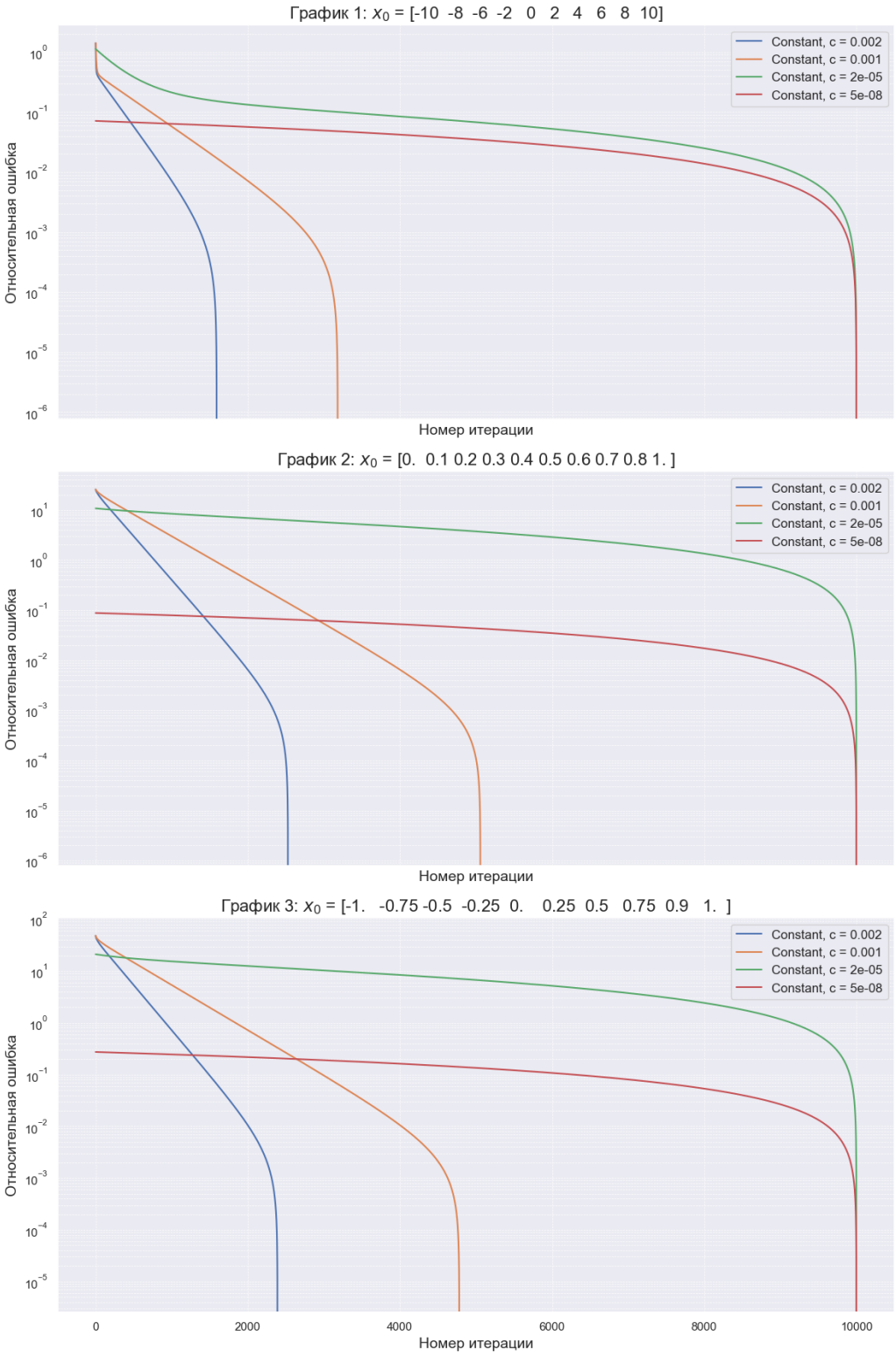
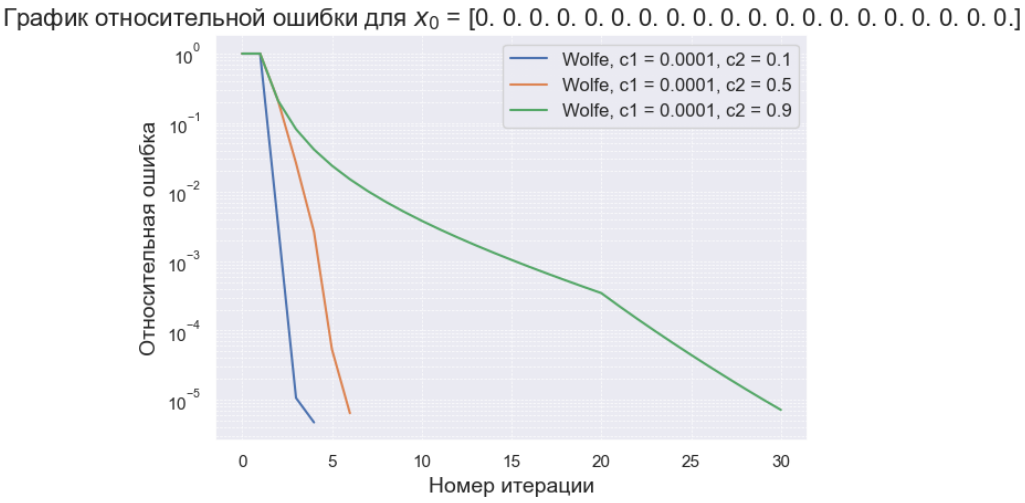
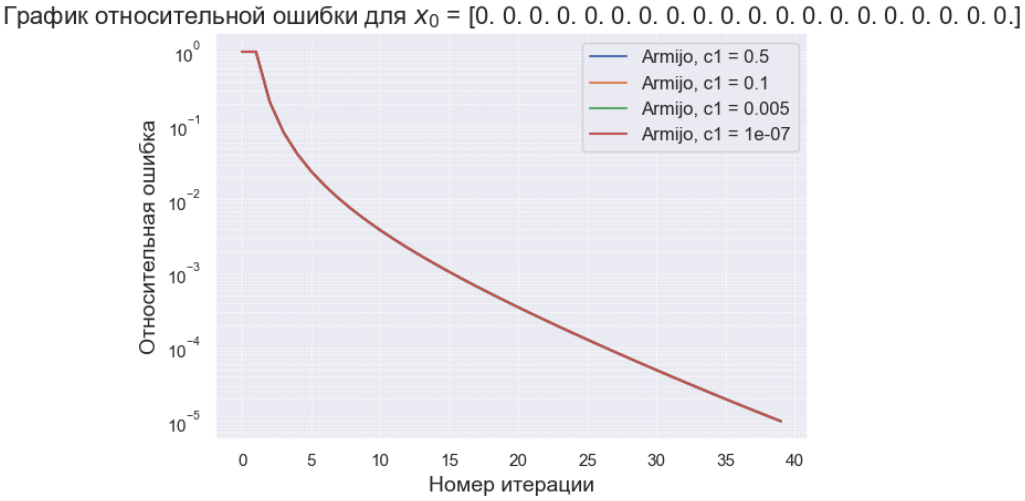


Рисунок 12. Константная стратегия с разными параметрами для различных начальных точек





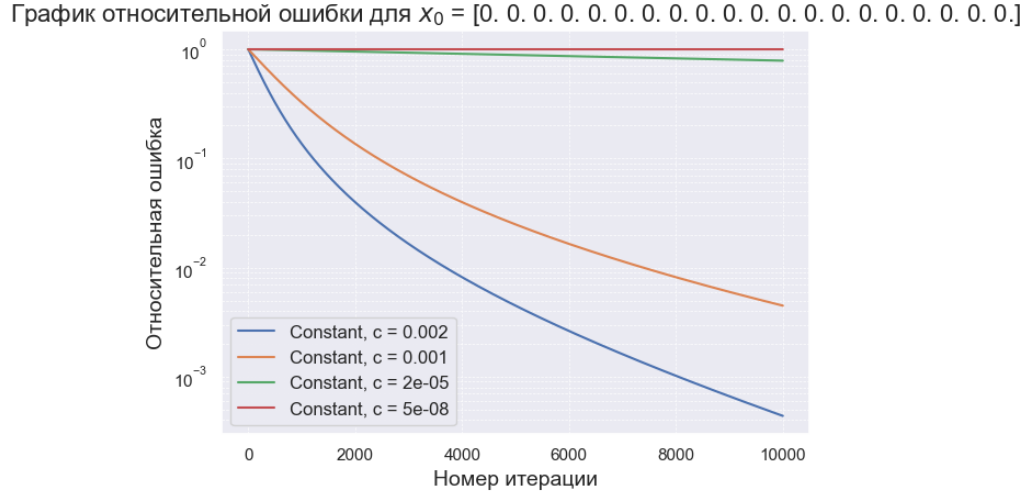


Рисунок 13. Градиентный спуск для задачи логистической регрессии с 3 стратегиями выбора шага

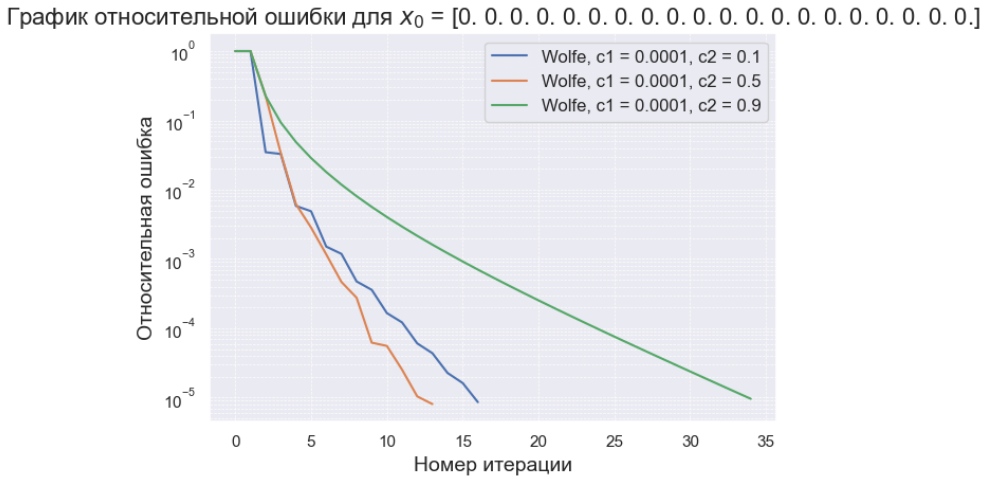
Вывод:

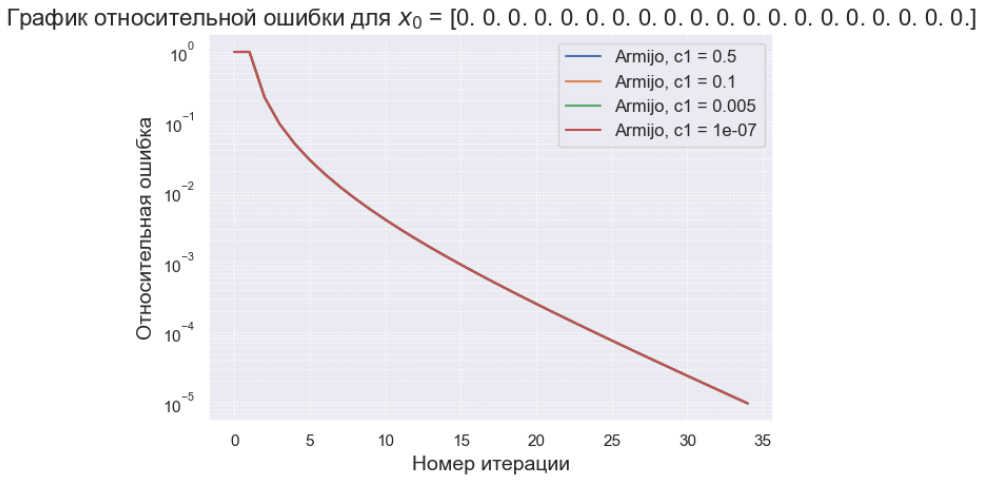
Из экспериментов видно, что константный шаг является наименее эффективной стратегией. Он либо требует много итераций для сходимости, либо вообще не сходится. Вульф стратегия демонстрирует лучшие результаты по сравнению с другими, но её эффективность сильно зависит от начальной точки. Для квадратичных задач, использование больших значений c2 позволяет значительно сократить количество итераций. Для задач логистической регрессии, стратегия Вульфа также показывает лучшие результаты, особенно при использовании маленьких значений c2. Стратегия Армихо хорошо подходит для решения квадратичных задач, обеспечивая стабильные результаты. Начальная точка оказывает одинаковое влияние на количество итераций для всех значений c1. Подводя итоги, стратегия Вульфа является наиболее эффективной среди рассмотренных стратегий. Она обеспечивает быструю сходимость и хорошее решение задач как квадратичной оптимизации, так и логистической регрессии

3.5 Эксперимент: Стратегия выбора длины шага в методе Ньютона

Задача:

Повторить предыдущий эксперимент но для метода Ньютона и решить какая стратегия наилучшая. Те же параметры и матрицы





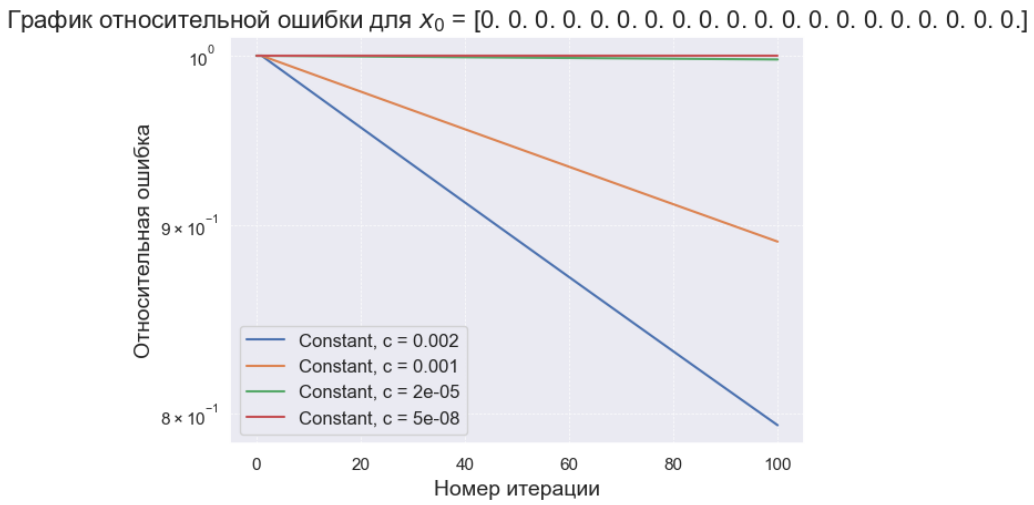


Рисунок 14,15,16. Метод Ньютона для задачи логистической регрессии с 3 стратегиями выбора шага

Вывод:

для метода Ньютона наилучшей стратегией является Вульф при оптимальном выборе параметра c2. Стратегия Армихо также может быть использована, если требуется более стабильная, но медленная сходимость. Константную стратегию лучше избегать из-за её зависимости от параметра и низкой эффективности