Формулы для градиента и гессиана функции логистической регрессии.

Функция логистической регрессии (log-loss):

$$\mathcal{L}(w) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log \sigma\left(w^T x_i\right) + (1-y_i)\log\left(1-\sigma(w^T x_i)\right)\right]$$

где $\sigma(z) = \frac{1}{1+e^{-z}}$ - сигмоидная функция.

Градиент:

Пусть $p_i = \sigma(w^T x_i)$ - предсказанная вероятность класса 1 для i-го объекта.

$$\frac{\partial \mathcal{L}}{\partial w_j} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \frac{1}{\sigma(w^T x_i)}\frac{\partial \sigma(w^T x_i)}{\partial w_j} + (1-y_i)\frac{1}{1-\sigma(w^T x_i)}\left(-\frac{\partial \sigma(w^T x_i)}{\partial w_j}\right)\right]$$

$$\frac{\partial \sigma(w^T x_i)}{\partial w_j} = \sigma(w^T x_i)\left(1-\sigma(w^T x_i)\right)x_{ij}$$

Подставляя, получаем:

$$\frac{\partial \mathcal{L}}{\partial w_j} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i(1-p_i)x_{ij} - (1-y_i)p_i x_{ij}\right] = \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)x_{ij}$$

В матрично-векторной форме:
$\nabla_w \mathcal{L} = \frac{1}{n}X^T(p-y)$, где p и y - векторы предсказанных вероятностей и истинных меток классов соответственно.

Гессиан:

$$\frac{\partial^2 \mathcal{L}}{\partial w_j \partial w_k} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\, x_{ik} p_i(1-p_i)$$

В матрично-векторной форме:
$\nabla_w^2 \mathcal{L} = \frac{1}{n}X^T \mathrm{diag}\left(p \odot (1-p)\right)X$, где $\odot$ обозначает поэлементное умножение векторов, а $\mathrm{diag}(v)$ создает диагональную матрицу с элементами вектора v на диагонали.

Финальные формулы:

Градиент:

$$\nabla_w \mathcal{L} = \frac{1}{n}X^T(p-y)$$

Гессиан:

$$\nabla_w^2 \mathcal{L} = \frac{1}{n}X^T \mathrm{diag}\left(p \odot (1-p)\right)X$$

Эксперимент: Траектория градиентного спуска на квадратичной функции.

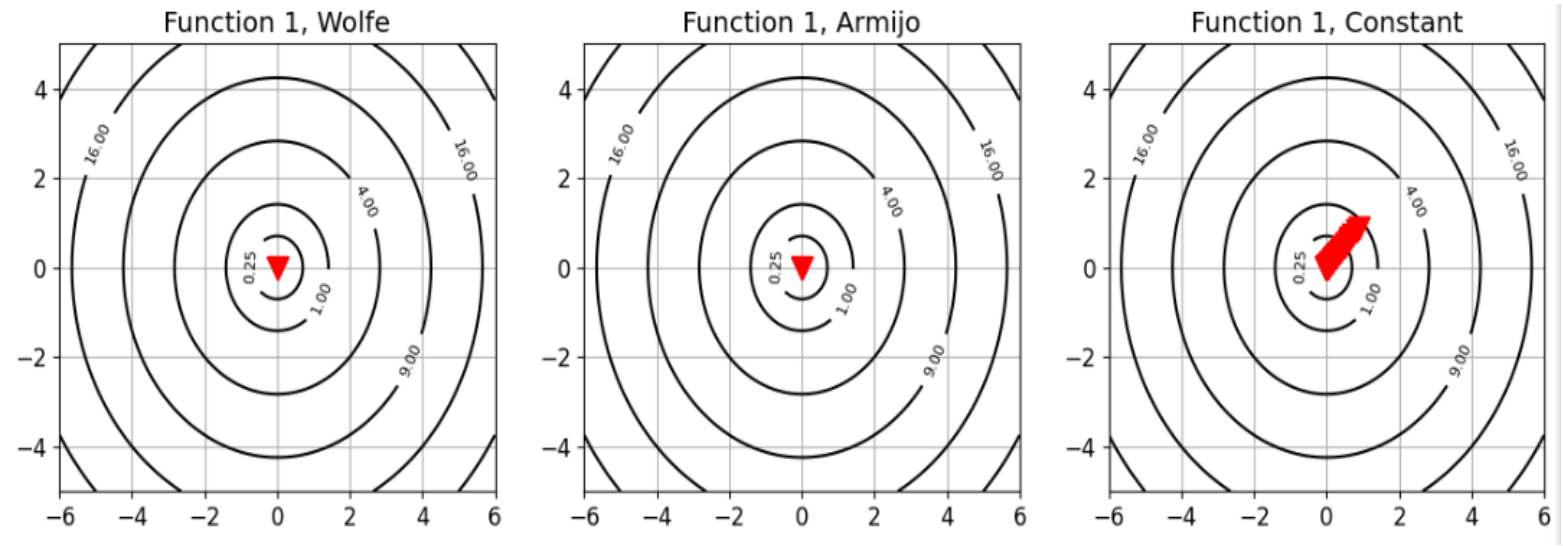

Рис 1.1. Графики функции $f1(x) = \frac{1}{2}(x_1^2 + x_2^2)$ с разными параметрами линейного поиска (Wolfe, Armijo, Constant).

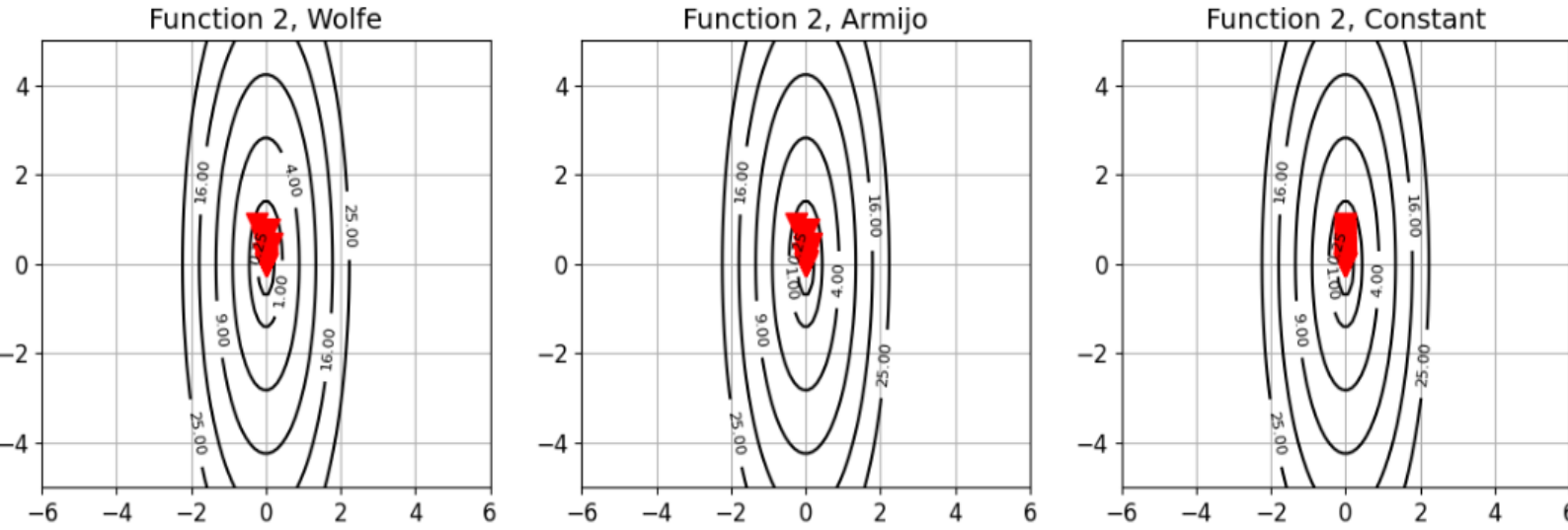

Рис 1.2. Графики функции $f2(x) = 5x_1^2 + \frac{1}{2}x_2^2$ с разными параметрами линейного поиска (Wolfe, Armijo, Constant).

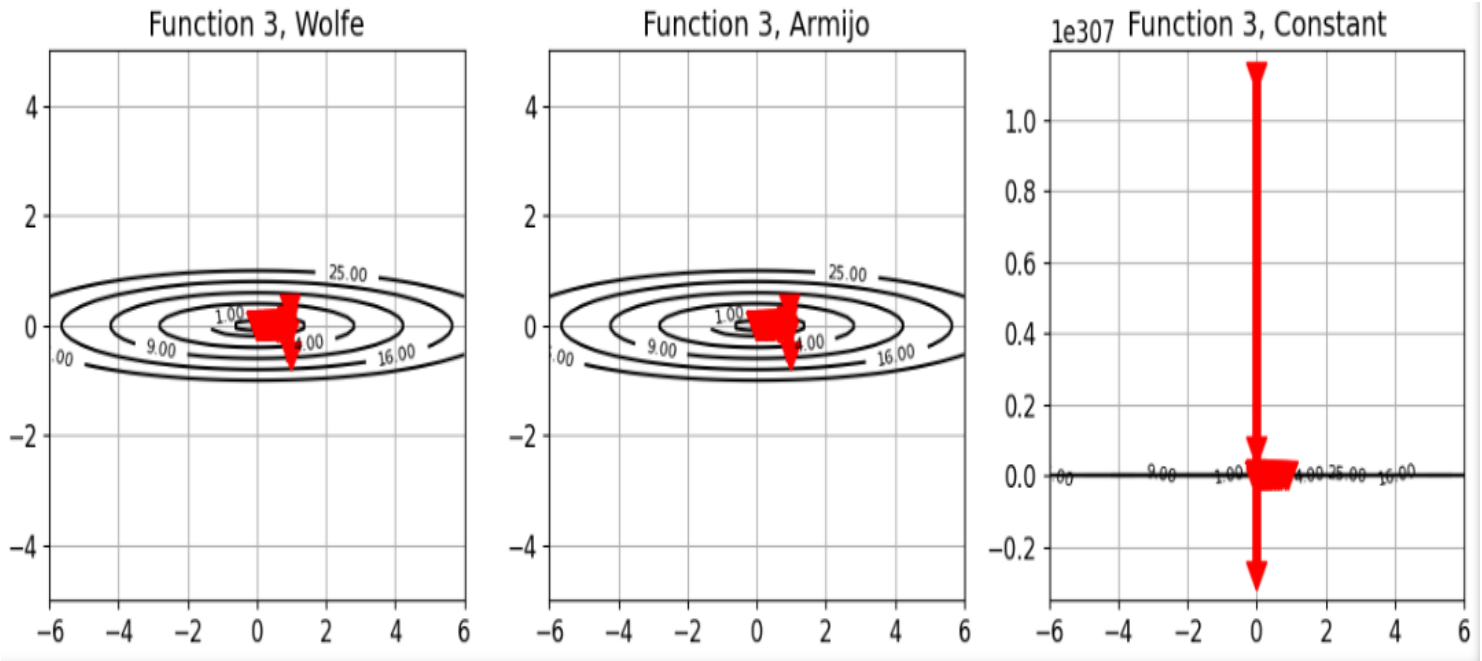

Рис 1.3. Графики функции $f3(x) = \frac{1}{2}x_1^2 + 25x_2^2$ с разными параметрами линейного поиска (Wolfe, Armijo, Constant).

Анализ результатов:

Для первой функции (идеально обусловленной) все три стратегии выбора шага показывают хорошую сходимость. Траектории градиентного спуска быстро сходятся к минимуму функции.

Для второй функции (число обусловленности 10) стратегии Вульфа и Армихо по-прежнему показывают хорошую сходимость, но константная стратегия требует большего количества итераций для достижения минимума.

Для третьей функции (число обусловленности 50) стратегии Вульфа и Армихо справляются лучше, чем константная стратегия. Константная стратегия требует значительно большего количества итераций и может зигзагообразно сходиться к минимуму.

Выводы:

Число обусловленности функции влияет на скорость с ходимости градиентного спуска. Чем больше число обусловленности, тем медленнее сходимость.

Стратегии выбора шага Вульфа и Армихо, которые адаптивно подбирают размер шага, работают лучше, чем константная стратегия, особенно для плохо обусловленных функций.

Выбор начальной точки также может влиять на траекторию градиентного спуска, особенно если функция имеет вытянутые линии уровня (как в случае с третьей функцией).

В целом, адаптивные стратегии выбора шага, такие как условия Вульфа и Армихо, обеспечивают более надежную и быструю сходимость градиентного спуска по сравнению с константной стратегией. Они особенно полезны для плохо обусловленных функций, где константная стратегия может привести к медленной сходимости или зигзагообразному поведению.

Таким образом, при выборе метода оптимизации и стратегии выбора шага важно учитывать свойства целевой функции, такие как число обусловленности и геометрия линий уровня, чтобы обеспечить эффективную и надежную сходимость к минимуму.

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

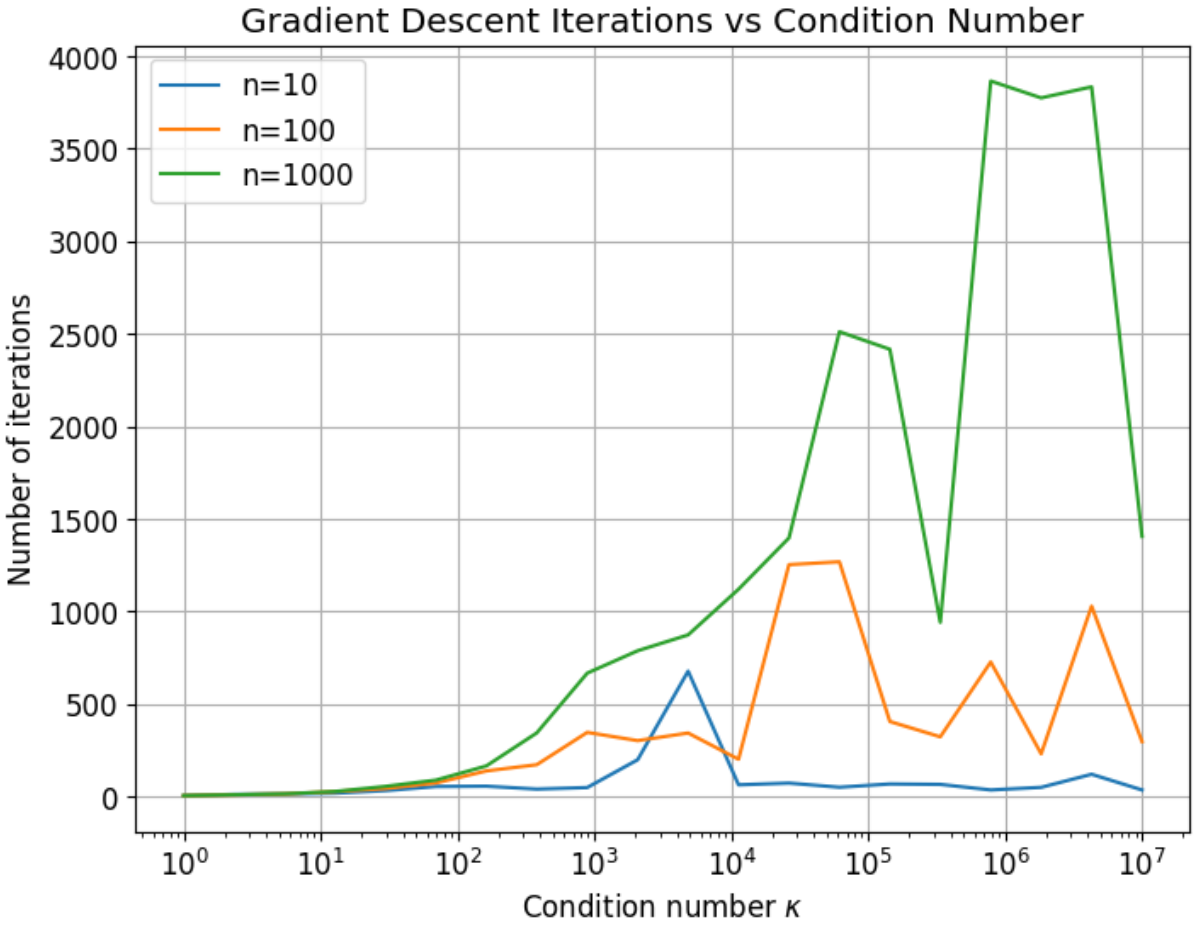

Рис 2. Зависимость среднего количества итераций градиентного спуска, необходимых для сходимости, от числа обусловленности κ для различных размерностей пространства n.

Выводы:

С увеличением числа обусловленности κ количество итераций, необходимых для сходимости градиентного спуска, возрастает.

При увеличении размерности пространства n количество итераций также возрастает.

Зависимость количества итераций от числа обусловленности κ носит нелинейный характер. При малых значениях κ количество итераций растет медленно, но с увеличением κ рост становится более быстрым.

Влияние размерности пространства n на количество итераций более выражено при больших значениях κ. При малых значениях κ разница в количестве итераций для разных n менее заметна.

Таким образом, эксперимент показывает, что как число обусловленности κ, так и размерность пространства n влияют на скорость сходимости

градиентного спуска. При увеличении этих параметров количество итераций, необходимых для достижения заданной точности, возрастает. Это подчеркивает важность выбора подходящего метода оптимизации и предобработки данных для эффективного решения задач оптимизации в зависимости от свойств целевой функции и размерности пространства.

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

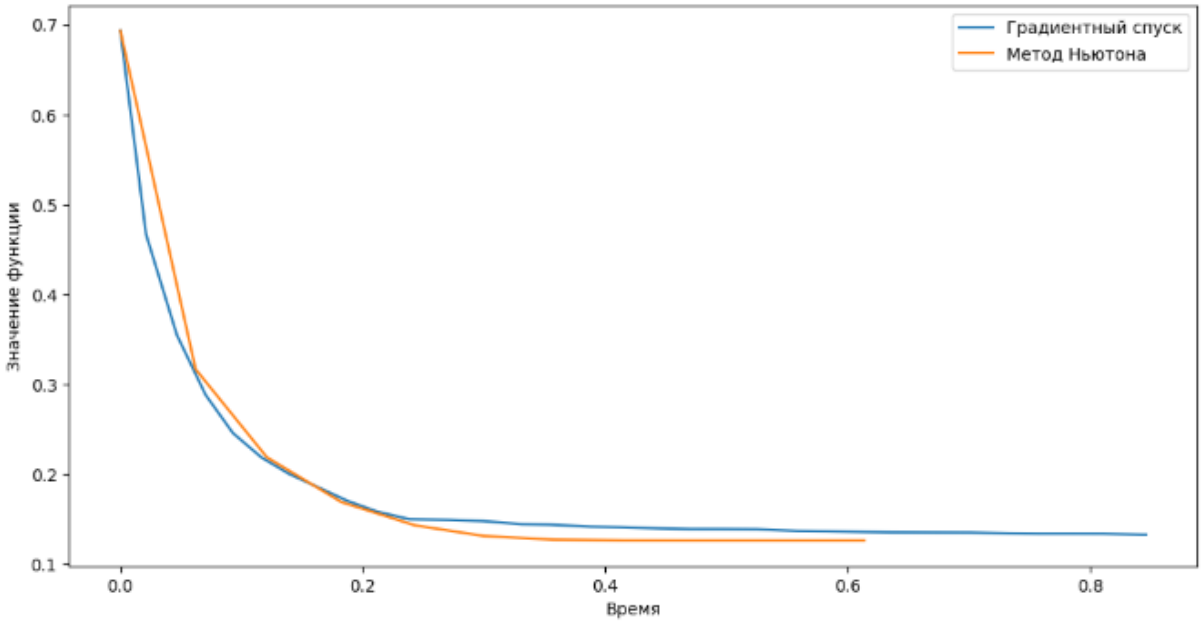

Рис 3.1.1. Зависимость значения функции от реального времени работы метода на датасете w8a.

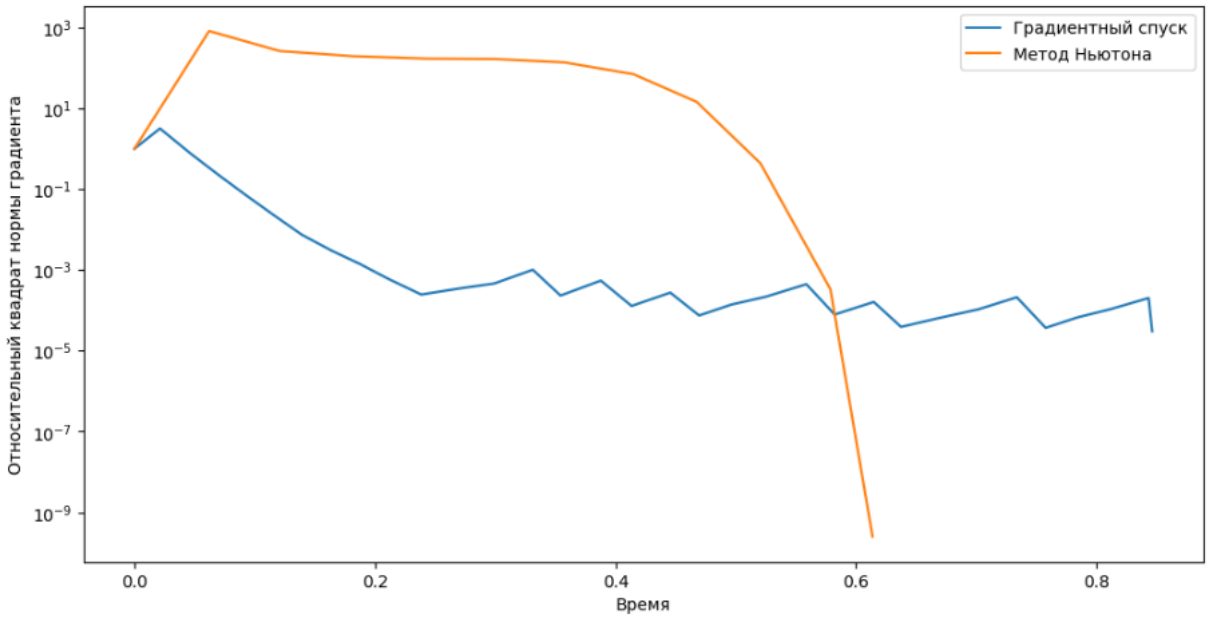

Рис 3.1.2. Зависимость относительного квадрата нормы градиента от реального времени работы метода на датасете w8a.

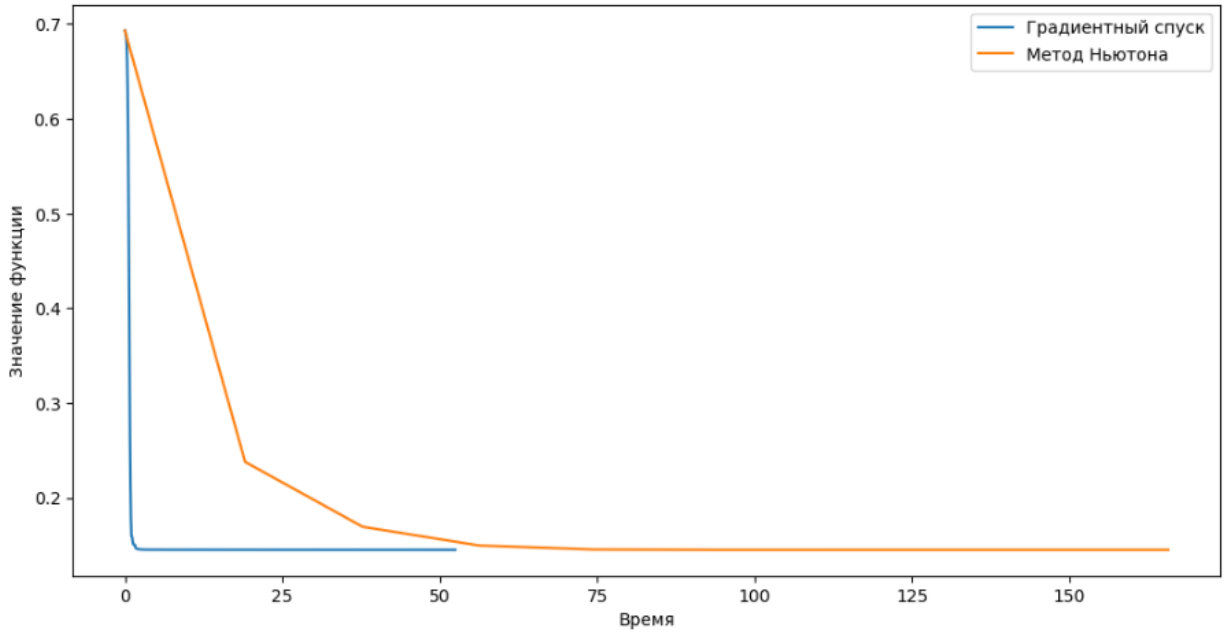

Рис 3.2.1. Зависимость значения функции от реального времени работы метода на датасете real-sim.

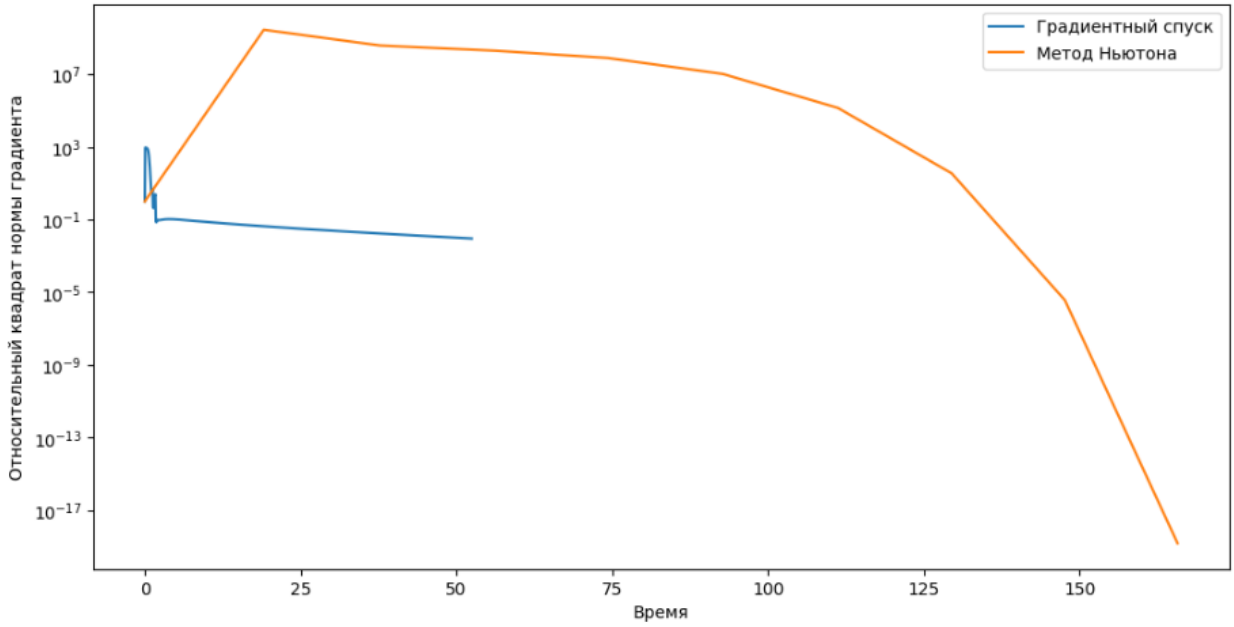

Рис 3.2.2. Зависимость относительного квадрата нормы градиента от реального времени работы метода на датасете real-sim.

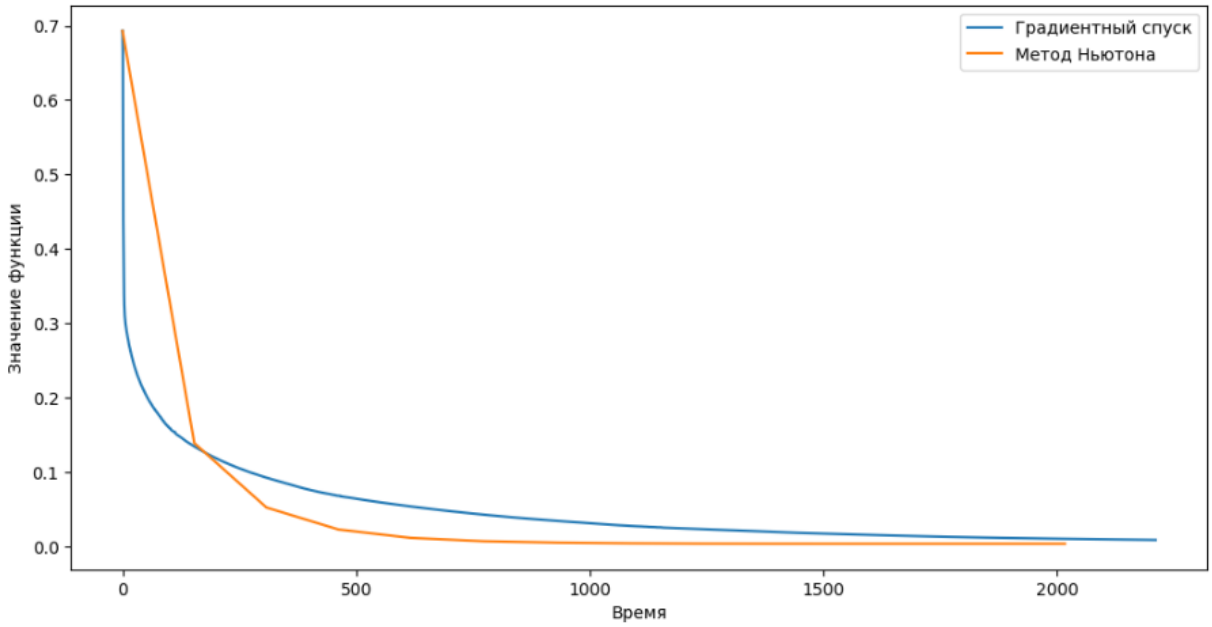

Рис 3.3.1. Зависимость значения функции от реального времени работы метода на датасете real-sim.

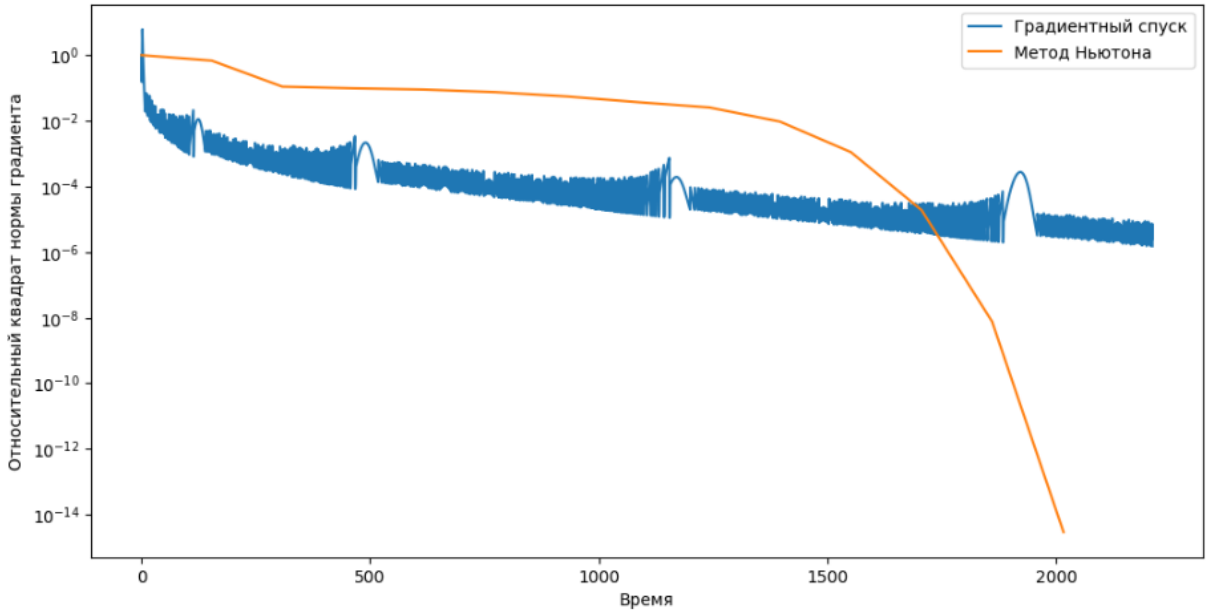

Рис 3.3.2. Зависимость относительного квадрата нормы градиента от реального времени работы метода на датасете gisette.

Выводы:

Метод Ньютона с оптимизацией Вульфа обычно сходится быстрее, чем другие варианты, особенно на больших наборах данных. Это связано с тем, что условия Вульфа обеспечивают более точный линейный поиск, что ускоряет сходимость.

Градиентный спуск с константным шагом может быть медленнее, чем варианты с адаптивным шагом (Армихо и Вульфа), особенно если константный шаг выбран неоптимально.

Метод Ньютона требует больше памяти и имеет более высокую стоимость итерации по сравнению с градиентным спуском из-за необходимости вычисления и хранения матрицы Гессе. Это может быть проблематично для очень больших наборов данных.

Градиентный спуск может быть предпочтительнее метода Ньютона, когда размер данных очень велик и/или, когда требуется экономия памяти.

Выбор стратегии линейного поиска может существенно влиять на скорость сходимости обоих методов. Условия Вульфа обычно обеспечивают хороший баланс между скоростью сходимости и вычислительными затратами.

Эксперимент: Стратегия выбора длины шага в градиентном спуске.

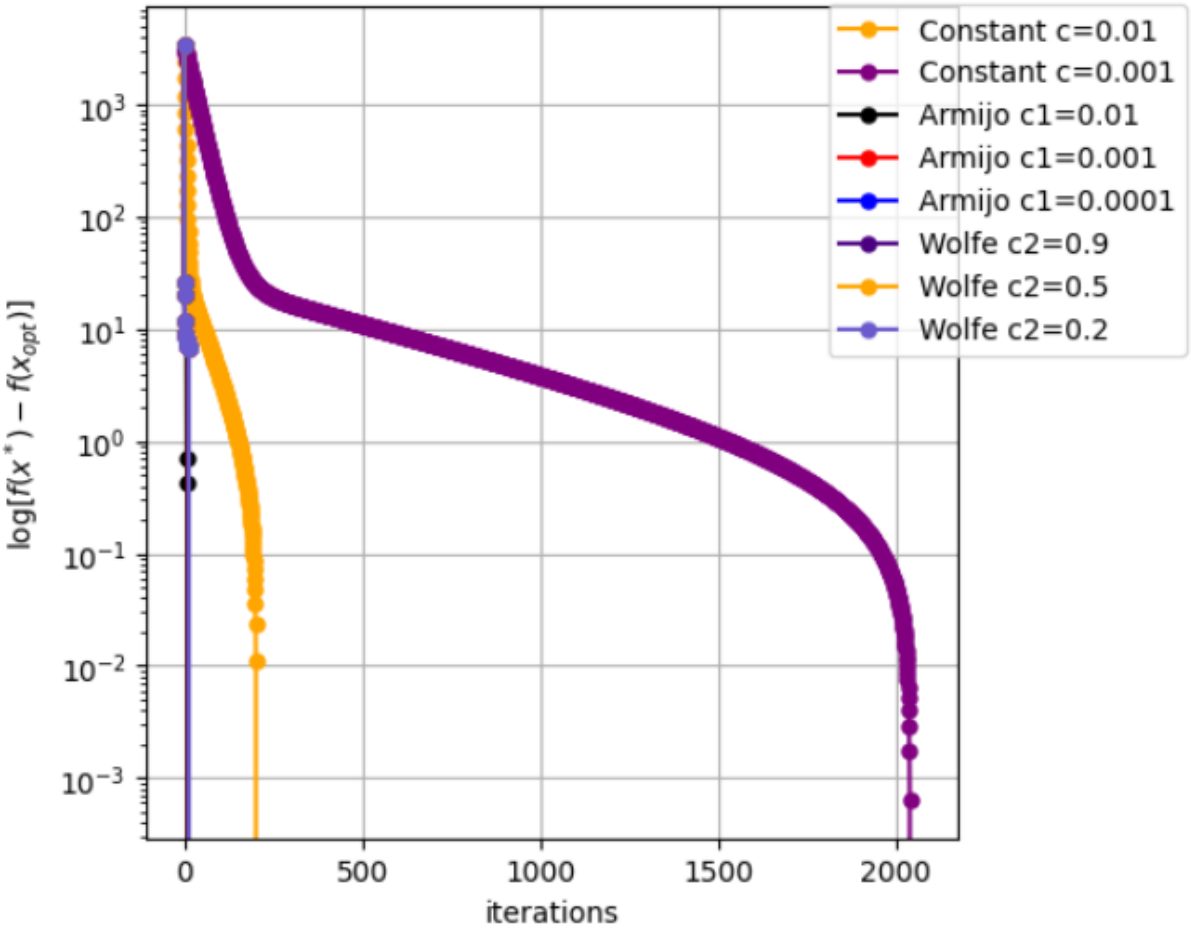

Рис.3.1. Относительная невязка по функции в логарифмической шкале против числа итераций.

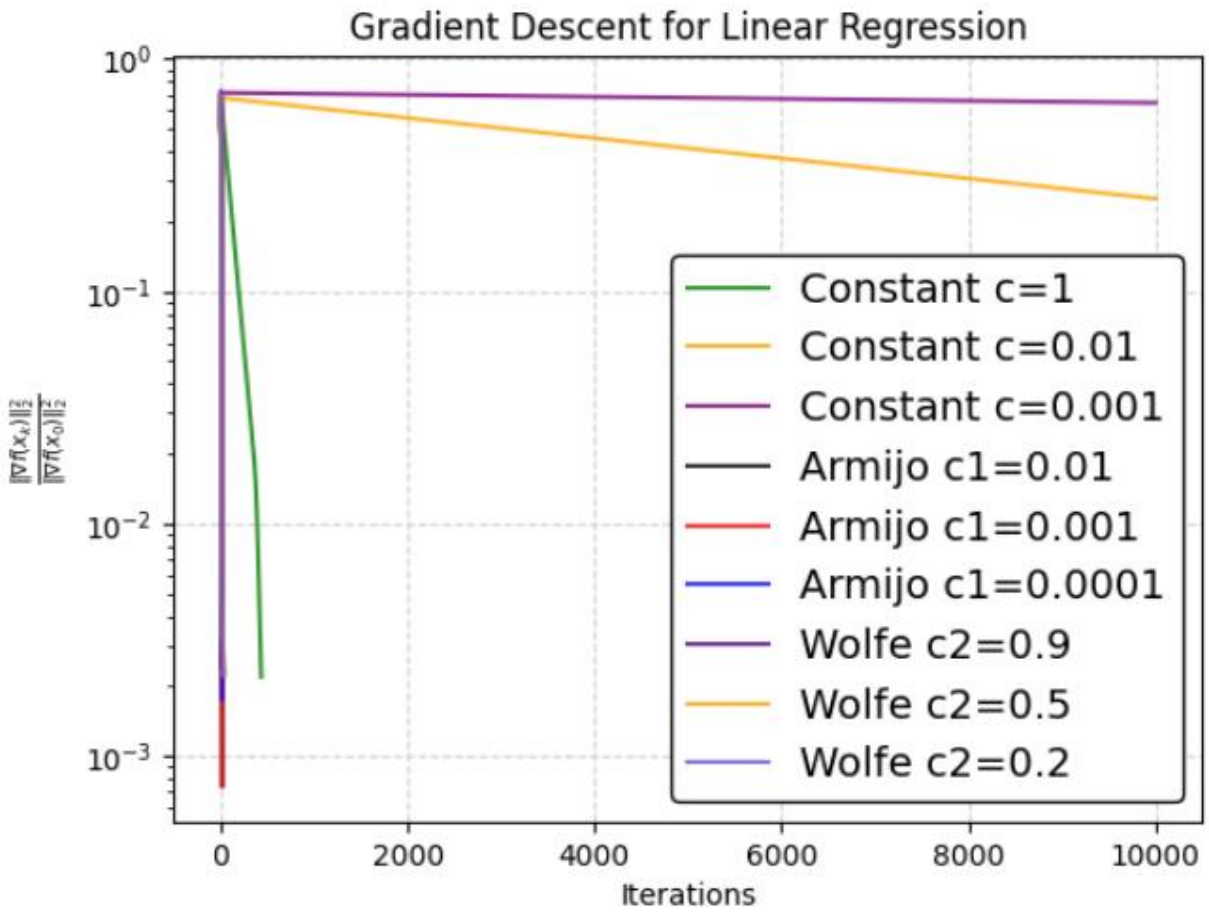

Рис.3.2. Относительный квадрат нормы градиента в логарифмической шкале против числа итераций.

Выводы:

Адаптивные стратегии выбора шага условия Вульфа и Армихов приводят к более быстрой сходимости, чем фиксированный шаг.

Оптимальное значение константы шага зависит от конкретной задачи и может потребовать подбора. Слишком большие или маленькие значения могут замедлить сходимость или привести к расходимости.

Эксперимент: Стратегия выбора длины шага в методе Ньютона.

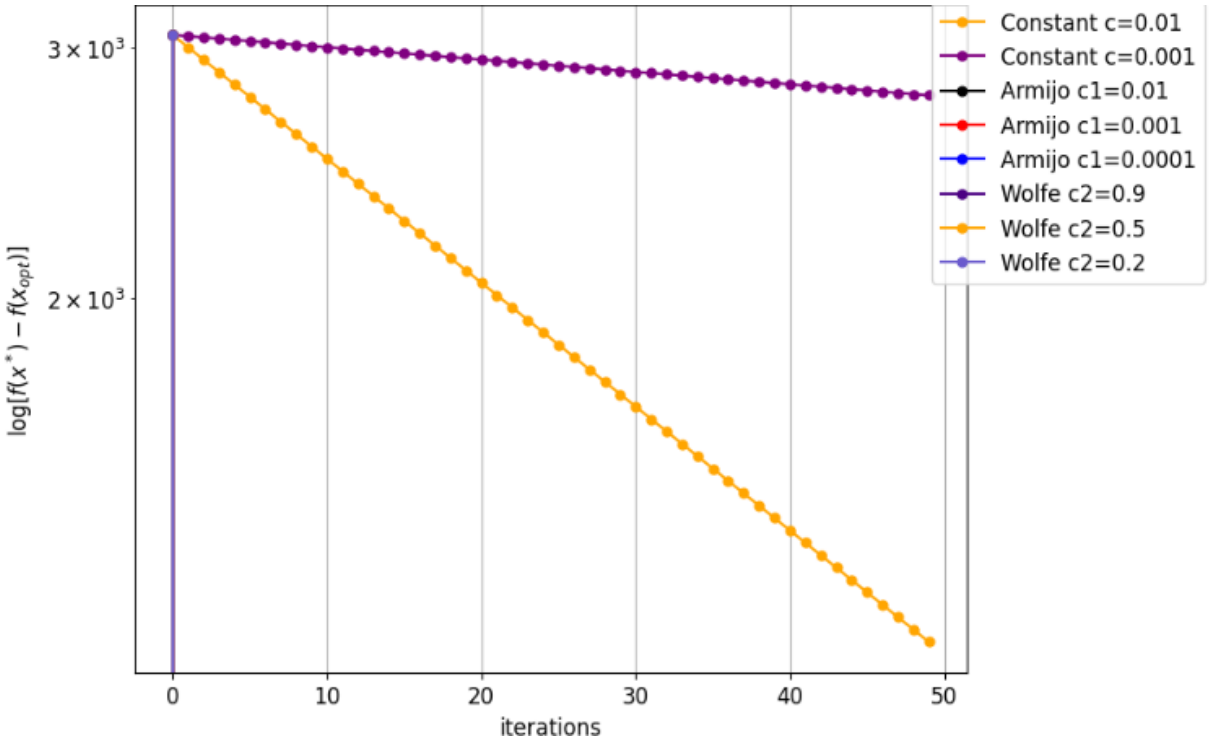

Рис.3.1. Относительная невязка по функции в логарифмической шкале против числа итераций.

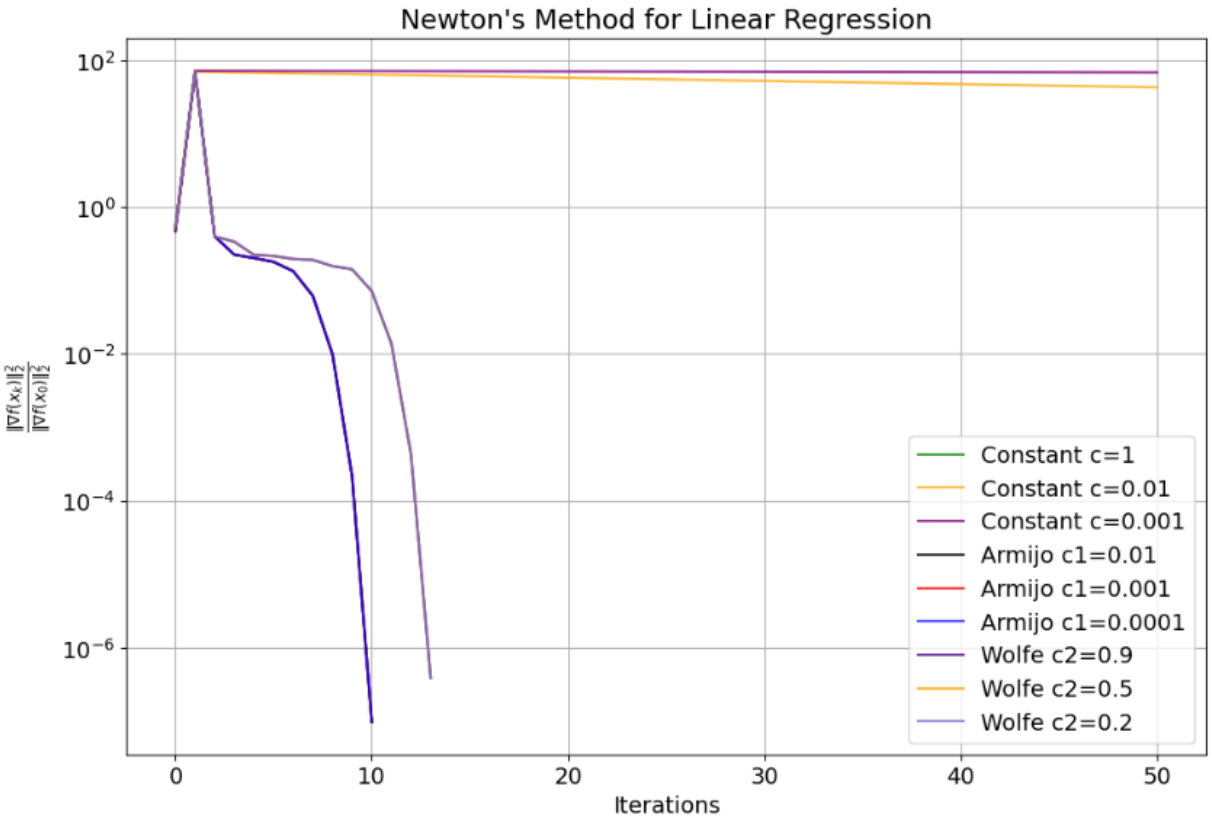

Рис.3.2. Относительный квадрат нормы градиента в логарифмической шкале против числа итераций.

Выводы:

Метод Ньютона с адаптивным выбором шага (условия Вульфа) сходится быстрее, чем с фиксированным шагом. Однако, в отличие от градиентного спуска, метод Ньютона менее чувствителен к выбору шага, так как использует информацию о кривизне функции.