Практическое задание 2 по курсу "Методы оптимизации в машинном обучении". Продвинутые методы безусловной оптимизации.

Выполнила студентка группы 191,
Косовская Арина

Содержание

1 Введение

Данное задание посвящено решению задачи нелинейной безусловной оптимизации: $\min\limits_{x \in \mathbb{R}^n} f(x),\ f(x)$ — достаточно гладкая. Для ее решения мною были реализованы следующие методы:

- метод сопряженных градиентов;
- усеченный метод Ньютона;
- метод L-BFGS.

В линейном поиске для выбора длины шага по умолчанию используется метод Вульфа с константами $c_1 = 1e-4,\ c_2 = 0.9$. В качестве критерия остановки в методе сопряженных градиентов используется условие $||Ax_k - b||_2 \leq \epsilon||b||_2$, где $\epsilon \in (0; 1)$ — заданная относительная точность. В усеченном методе Ньютона для раннего выхода из метода сопряженных градиентов необходимо выполнение $||\nabla^2 f(x_k)d + \nabla f(x_k)||_2 \leq \eta||\nabla f(x_k)||_2$, где $\eta = \min\{0.5, \sqrt{||\nabla f(x_k)||_2}\}$, после чего проверяется, что d_k является направлением спуска, т.е. $\langle \nabla f(x_k), d_k \rangle < 0$ (иначе метод сопряженных градиентов запускается повторно из точки d_k с $\eta_k \ /= 10$. В остальных методах в качестве критерия остановки используется $||\nabla f(x_k)||_2^2 \leq \epsilon||\nabla f(x_0)||_2^2$.

2 Эксперимент 1. Зависимость числа итераций метода сопряженных градиентов от числа обусловленности и размерности пространства.

В данном эксперименте мы исследуем зависимость числа итераций, необходимых для сходимости метода сопряженных градиентов, от числа обусловленности $k \geq 1$ оптимизируемой функции и размерности пространства n оптимизируемой переменной. В качестве оптимизируемой функции мы рассматриваем квадратичную задачу размера n и числа обусловленности k.

Также требуется сравнить полученные результаты и сравнить их с результатами аналогичного эксперимента на градиентном спуске. Для корректности сравнения при запуске метода сопряженных градиентов в качестве относительной точности будем брать $\epsilon_{CG} = \sqrt{\epsilon}\dfrac{||Ax_0 - b||_2}{||b||_2}$, где ϵ — точность, с которой запускается градиентный спуск (мы будем брать значение по умолчанию $\epsilon = 1e-5$).

Перейдем к описанию данных. Случайным образом мы будем генерировать квадратичную задачу размера n и числа обусловленности k. Для этого сначала мы

сгенерируем разреженную диагональную матрицу $A \in S^n_{++}$, на диагонали которой будут стоять числа из промежутка от 1 до k (причем среди данных чисел одно будет равняться 1, а другое — k, остальные значения выбираются случайно из равномерного распределения). Элементы вектора $b \in \mathbb{R}^n$ будут генерироваться случайно из равномерного распределения от 0 до k.

Далее мы будем запускать на сгенерированной квадратичной задаче метод сопряженных градиентов. В качестве x_0 будем брать 0. На каждом запуске будем замерять число итераций $T(n, k)$, необходимых для сходимости (то есть когда был выполнен критерий остановки). Остальные параметры будут выбраны по умолчанию.

Фиксируя значение n, будем случайно генерировать квадратичную функцию и перебирать k (от 1 до 500 с шагом 10). Повторим данный эксперимент 4 раза. У нас получится семейство кривых зависимости $T(n, k)$ от k, построим график, нарисовав кривые одним цветом. Изменив значение n, повторим описанную процедуру: сгенерируем квадратичную функцию, переберем k и повторим 4 раза. Полученные кривые зависимости $T(n, k)$ от k изобразим уже другим цветом, и так далее. Значения n будем перебирать по логарифмической сетке от 10 до 10^6 с шагом 10.

На рисунке 2.1 изображен график зависимости числа итераций метода сопряженных градиентов от числа обусловленности и размерности пространства, на рисунке 2.2 — для градиентного спуска.

Выводы:

При небольшом значении числа обусловленности (меньше 10) для любого размера пространства наблюдается линейная зависимость числа итераций от числа обусловленности. При дальнейшем увеличении в случае, когда размерность пространства равна 10, для метода сопряженных градиентов число итераций не зависит от числа обусловленности, оно не превышает 15. Это объясняется тем, что метод всегда сходится за число итераций, меньшее n (без учета погрешности). При бОльших размерах пространства наблюдается линейная зависимость числа итераций от числа обусловленности (чем больше k, тем больше итераций требуется до сходимости). Число итераций не увеличивается с ростом размерности пространства.

Переходя к сравнению результатов метода сопряженных градиентов и градиентного спуска, стоит отметить, что метод сопряженных градиентов работает стабильнее и быстрее, ему необходимо меньшее число итераций до сходимости (это заметно даже при малых числах обусловленности, при бОльших значеинях разни-

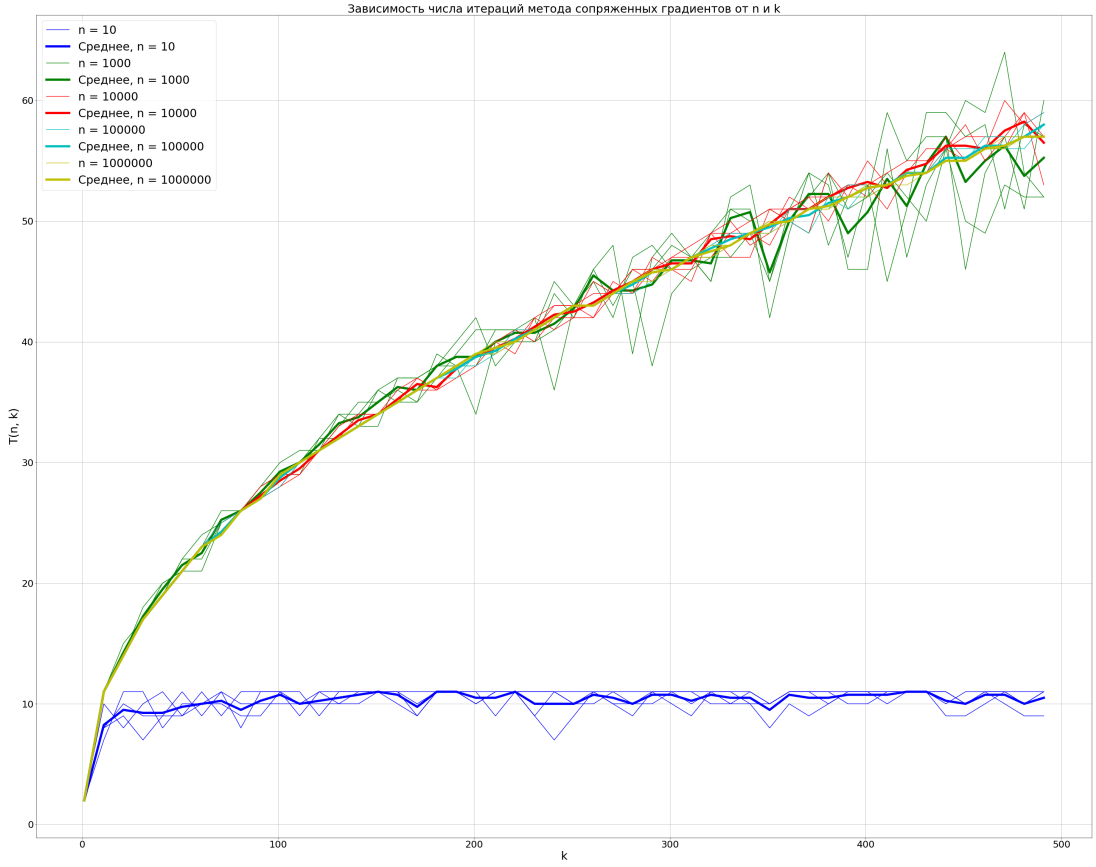


Рис. 2.1: Зависимость числа итераций метода сопряженных градиентов от числа обусловленности и размерности пространства

ца становится еще значительнее). Для метода сопряженных градиентов разброс и дисперсия числа итераций до сходимости значительно меньше.

3 Эксперимент 2. Выбор размера истории в методе L-BFGS

Оценим размер требуемой памяти и сложность итерации метода L-BFGS в зависимости от размера истории l и размерности пространства n без учета сложности оракула. В нашей реализации для вычисления направления поиска мы храним в истории l пар векторов, размер каждого равен n. Поэтому мы тратим $O(n \cdot l)$ памяти. Для вычисления направления используется рекурсивное умножение матрицы на вектор. Рекурсия вызывается число раз, равное размеру истории, то есть l. На каждом ее шаге нужно посчитать скалярное произведение (за $O(n)$ из опре-

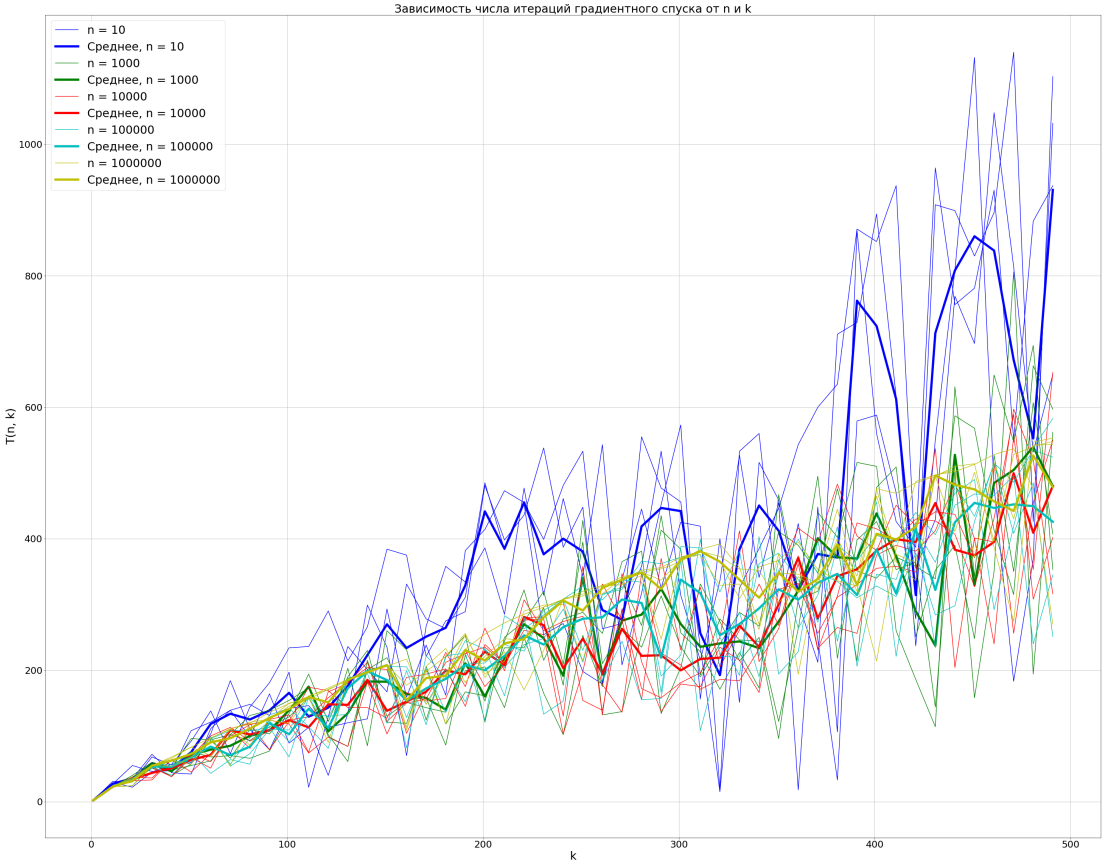


Рис. 2.2: Зависимость числа итераций градиентного спуска от числа обусловленности и размерности пространства

деления операции), значит, для одной итерации требуется $O(n \cdot l)$ памяти.

Далее в эксперименте мы исследуем как влияет размер истории L-BFGS на поведение метода. В качестве тестовой функции использовалась логистическая регрессия с l^2-регуляризатором и набором данных gisette. Коэффициент регуляризации равен $\frac{1}{m}$, начальная точка $x_0 = 0_m$, m — количество признаков. Остальные параметры выбраны по умолчанию.

Рассмотрим несколько вариантов выбора размера истории ($l = 0, l = 1, l = 5, l = 10, l = 50, l = 100$) и построим график зависимости $\frac{||\nabla f(x_k)||_2^2}{||\nabla f(x_0)||_2^2}$ в логарифмической шкале против номера итерации (изображено на рисунке 3.1) и график зависимости $\frac{||\nabla f(x_k)||_2^2}{||\nabla f(x_0)||_2^2}$ в логарифмической шкале против реального времени работы (изображено на рисунке 3.2).

Видим, что больше всего времени и итераций до сходимости требуется методу

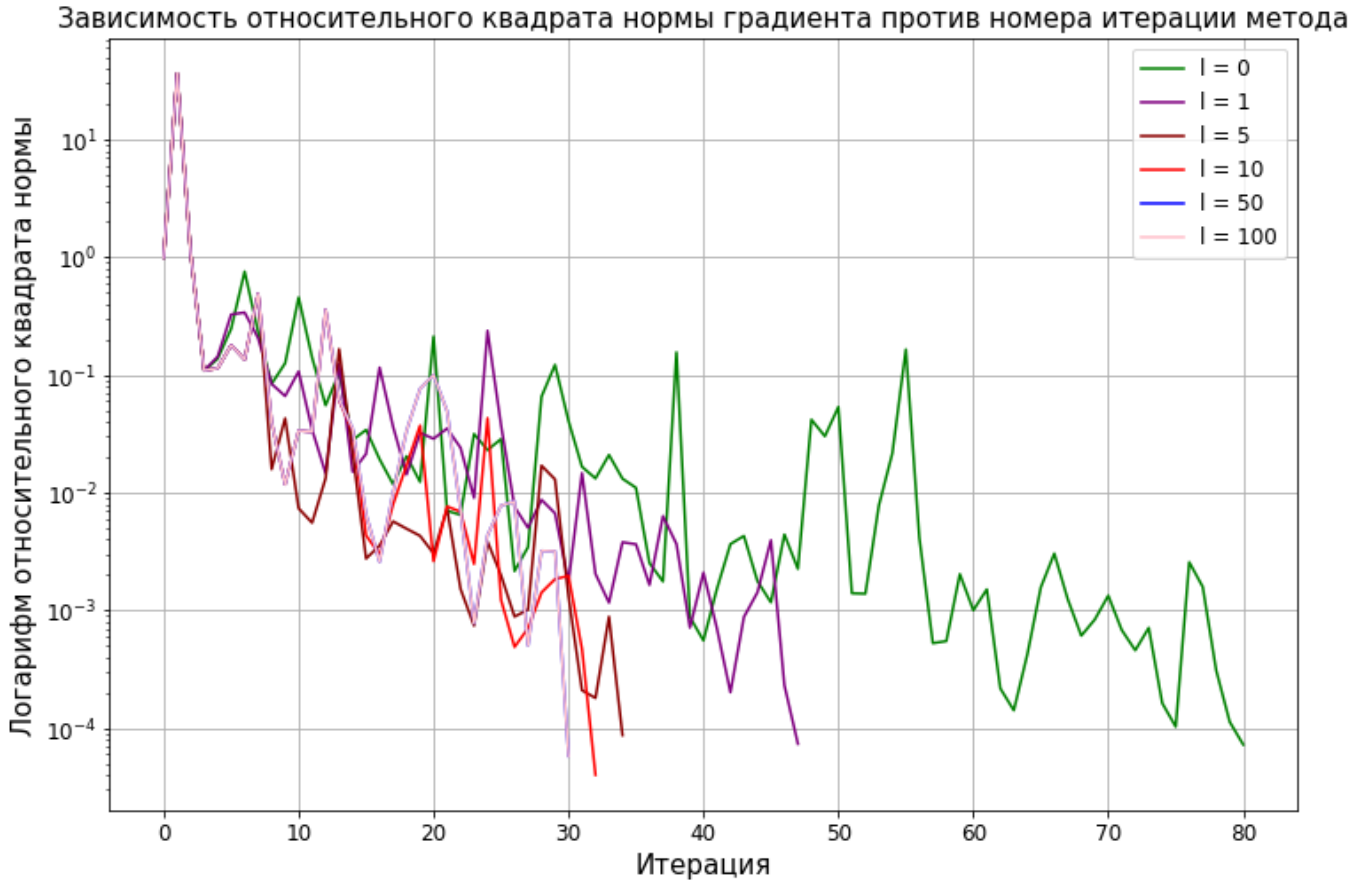


Рис. 3.1: Зависимость относительного квадрата нормы градиента против номера итерации в логарифмической шкале

в случае, когда размер истории равен 0. Как обсуждалось в чате, в данном случае метод представляет собой вариацию градиентного спуска. А, как будет показано в следующем эксперименте, градиентный спуск сходится медленнее L-BFGS.

В случаях, когда размер истории равен 50 и 100, метод работает одинаково, графики совпадают. Объясняется это тем, что сходимость происходит чуть меньше, чем за 20 итераций, и история H_k не заполняется полностью. Этим можно объяснить совпадение графиков на i-х итерациях, где $i \leq l$.

Также стоит отметить, что при размере памяти, равном 10, методу требуется приблизительно столько же времени и итераций до сходимости, сколько и при бОльших размерах. Таким образом, достаточно относительно небольшого размера памяти для эффективной реализации метода.

Выводы:

• Для метода L-BFGS сложность итерации составляет $O(n \cdot l)$, требуется $O(n \cdot l)$ памяти.

• Хуже всего метод работает при размере памяти, равном 0. При размере памяти, большем или равном 10, отличия в скорости работе и количестве итераций метода незначительны.

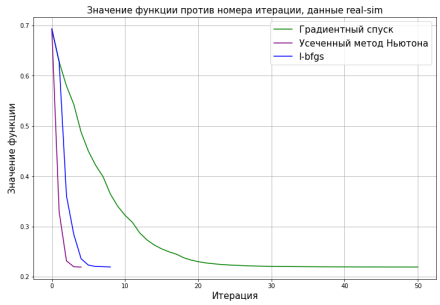
Рис. 3.2: Зависимость относительного квадрата нормы градиента против реального времени работы в логарифмической шкале

4 Эксперимент 3. Сравнение методов на реальной задаче логистической регрессии

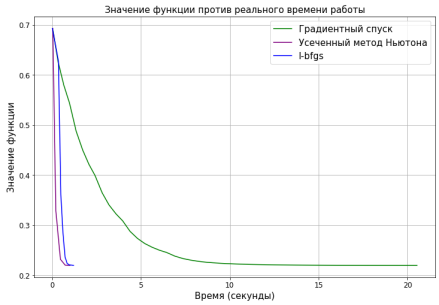
В данном эксперименте мы будем сравнивать усеченный метод Ньютона, метод L-BFGS и градиентный спуск на задаче логистической регрессии. В качестве данных будут использоваться наборы данных с сайта LIBSVM: w8a, gisette, real-sim, news20.binary, rcv1.binary. В качестве коэффициента регуляризации λ и начальной точки x_0 взяты $\frac{1}{m}$ и 0 соответственно. Остальные параметры выбраны по умолчанию. Для считывания данных я воспользовалась кодом, который прислал Владимир Романов в общий чат курса.

На рисунках 4.1, 4.2, 4.3, 4.4, 4.5 изображены графики зависимости значения функции против номера итерации метода, зависимости значения функции против реального времени работы и зависимость относительного квадрата нормы градиента в логарифмической шкале) против реального времени работы.

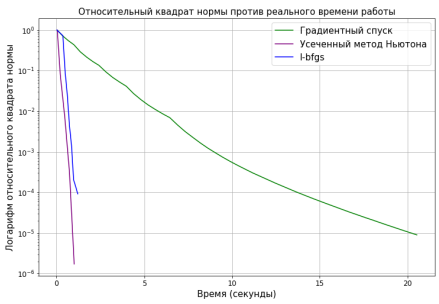
Выводы: на всех наборах данных градиентный спуск работает медленнее и требует бОльшего числа итераций. Быстрее, точнее и стабильнее всех работает усеченный метод Ньютона. Из этого можно сделать вывод, что лучше использо-



(a) значение функции против номера итерации



(b) значение функции против реального времени работы



(c) относительный квадрат нормы против времени

Рис. 4.1: Набор данных real-sim

вать усеченный метод Ньютона.

5 Эксперимент 4. Сравнение метода сопряженных градиентов и L-BFGS на квадратичной функции

В данном эксперименте мы сравним метод сопряженных градиентов и L-BFGS на квадратичной строго выпуклой функции $f(x) = \frac{1}{2}\langle Ax, x - b, x,\ x \in \mathbb{R}^n, b \in \mathbb{R}^n, A \in \mathbb{S}^n_{++}$.

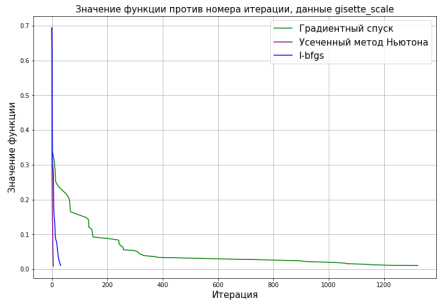
Сгенерируем случайным образом три квадратичные строго выпуклые функции (для первой $n = 3$, для второй $n = 50$, для третьей $n = 500$) и запустим на каждой оба метода из точки $x_0 = 0$. В методе L-BFGS будем перебирать различных размеры истории ($l = 0, 1, 10, 30, 50$). На рисунке 5.1 представлены графики сходимости в терминах евклидовой нормы невязки $r_k = Ax_k - b$ в логарифмической шкале против номера итерации.

В методе L-BFGS использовался точный линейный поиск. Аналитически вычислялась наилучшая длина шага как $\operatorname{argmin}_{\alpha \geq 0} f(x_k + \alpha d_k)$. Поскольку функция квадратичная, $\alpha_k = -\dfrac{\langle \nabla f(x_k), d_k \rangle}{\langle Ad_k, d_k \rangle}$.

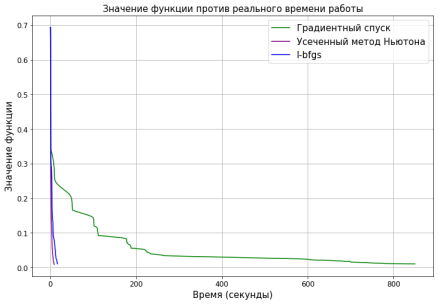
При $n = 3$ графики совпадают при любом значении размера памяти. В остальных случаях метод сопряженных градиентов работает за меньшее число итераций, стабильнее и точнее. Связано это с тем, что метод L-BFGS в случае, если число итераций превышает l, хранит не все данные оптимизации.

Выводы: в случае, если размерность функции небольшая, оба метода справляются за одинаково малое количество итераций и успешно. Иначе стабильнее работает метод сопряженных градиентов, в особенности если значение l малень-

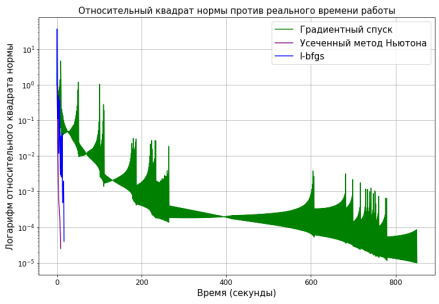
кое.



(a) значение функции против номера итерации

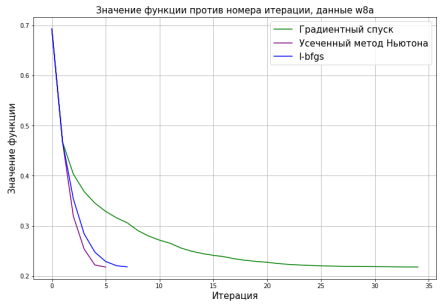


(b) значение функции против реального времени работы

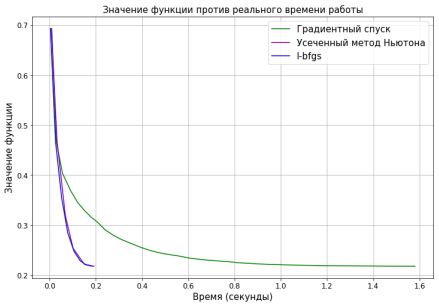


(c) относительный квадрат нормы против времени

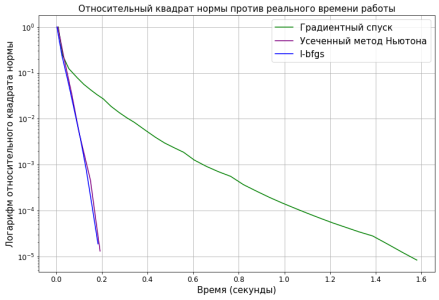
Рис. 4.2: Набор данных gisette



(a) значение функции против номера итерации

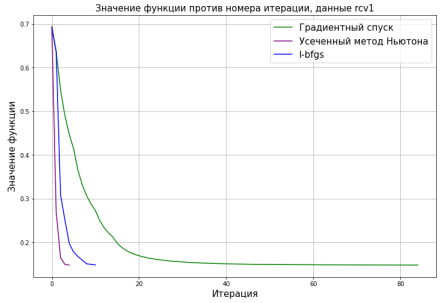


(b) значение функции против реального времени работы

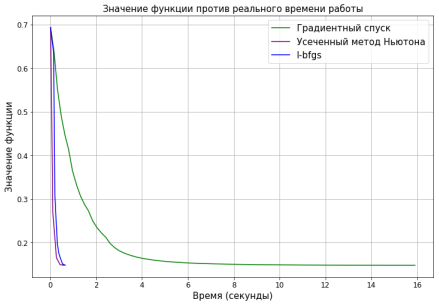


(c) относительный квадрат нормы против времени

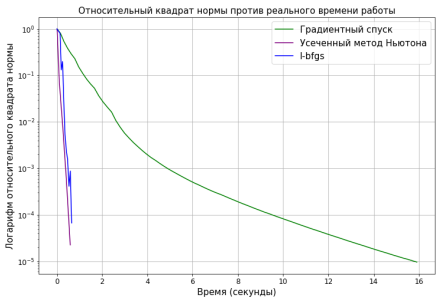
Рис. 4.3: Набор данных w8a



(a) значение функции против номера итерации

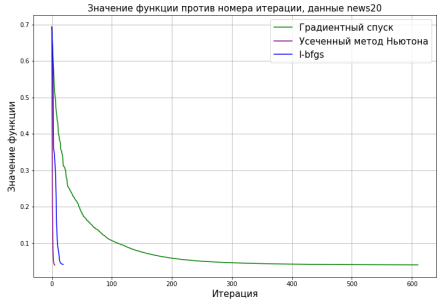


(b) значение функции против реального времени работы

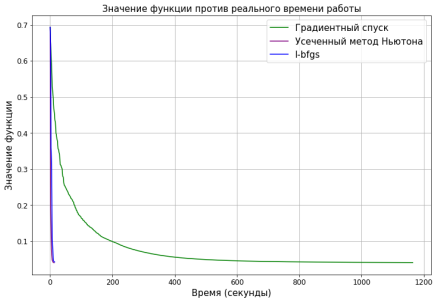


(c) относительный квадрат нормы против времени

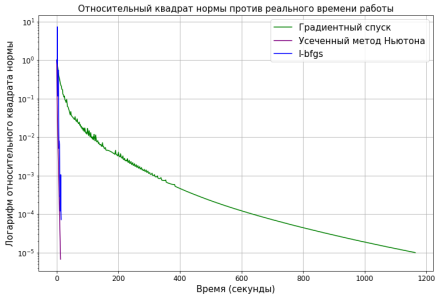
Рис. 4.4: Набор данных rcv1.binary



(a) значение функции против номера итерации

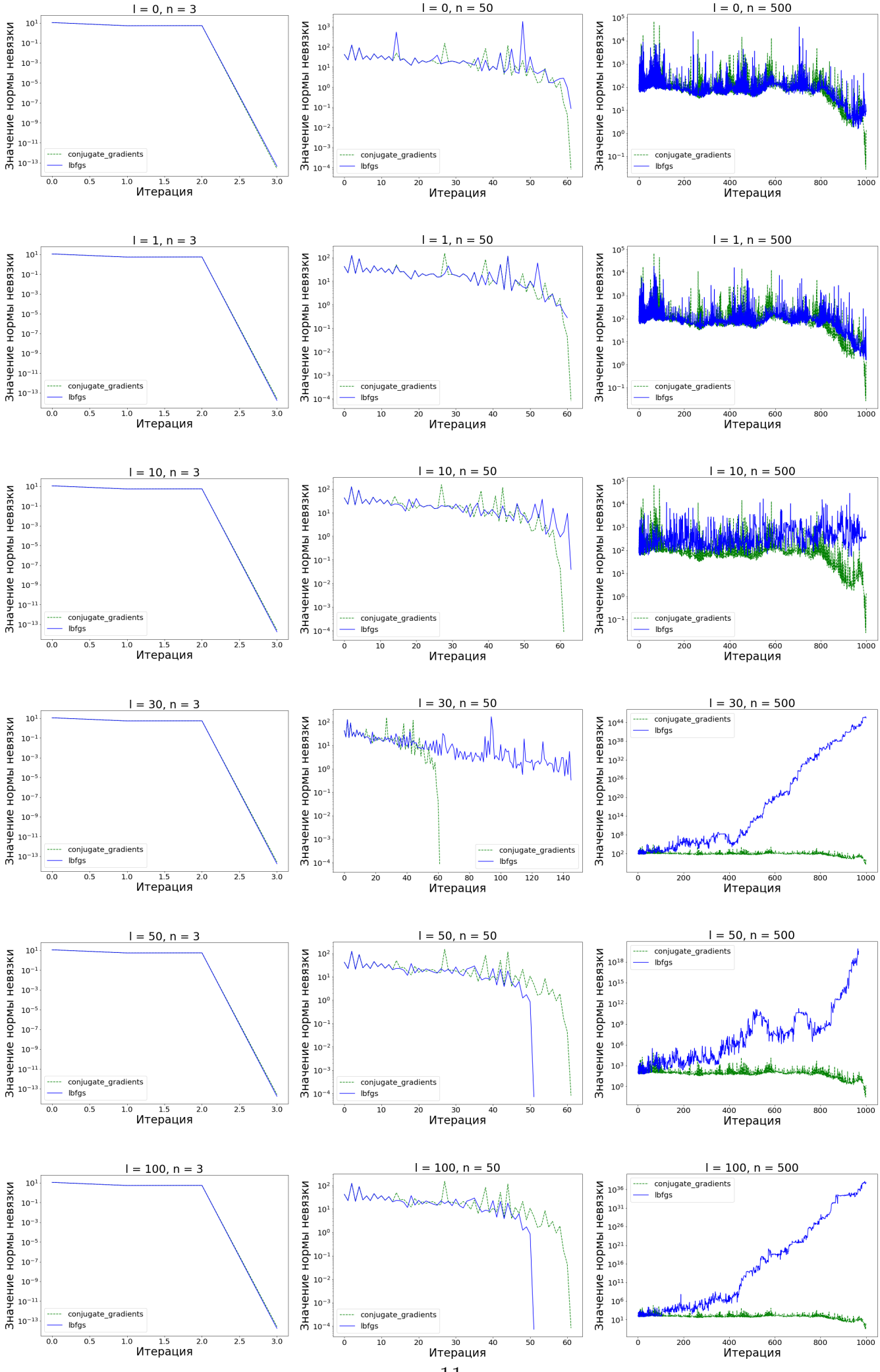


(b) значение функции против реального времени работы



(c) относительный квадрат нормы против времени

Рис. 4.5: Набор данных news20.binary


l = 0, n = 3
l = 0, n = 50
l = 0, n = 500
l = 1, n = 3
l = 1, n = 50
l = 1, n = 500
l = 10, n = 3
l = 10, n = 50
l = 10, n = 500
l = 30, n = 3
l = 30, n = 50
l = 30, n = 500
l = 50, n = 3
l = 50, n = 50
l = 50, n = 500
l = 100, n = 3
l = 100, n = 50
l = 100, n = 500
Значение нормы невязки
Итерация
conjugate_gradients
lbfgs