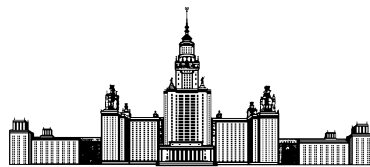


Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной математики и Кибернетики
Кафедра Математических Методов Прогнозирования

Методы оптимизации в машинном обучении

Отчет по практическому заданию №2.

«Продвинутые методы безусловной оптимизации»

Выполнил:
магистр 1 курса 517 группы
Таскынов Ануар

Москва, 2017

Содержание.

1 Постановка задания и описание работы

В данном практическом задании необходимо было реализовать три метода оптимизации:

1. Метод сопряженных градиентом для решения системы линейных уравнений с квадратичной положительно-определенной матрицей.
2. Усеченный (безгессианный) метод Ньютона.
3. Квазиньютоновский метод L-BFGS.

В данном отчете содержатся результаты экспериментов, которые сравнивают между собой данные методы на сгенерированной выборке и на реальных задачах. Были получены соответствующие выводы.

2 Вывод формул

Рассматривается функция:

$$f(x) = \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle,$$

где $x, b \in \mathbb{R}^n$, $A \in \mathbb{S}_{++}^n$. Для этой функции рассматривается задача минимизации на всем пространстве \mathbb{R}^n . Поскольку функция квадратичная с положительно-определенной матрицей, то для произвольного метода спуска $x_{k+1} = x_k + \alpha_k d_k$ можно аналитически посчитать константу $\alpha_k = \arg \min_{\alpha \geq 0} f(x_k + \alpha d_k)$:

$$\begin{aligned} f(x_k + \alpha d_k) &= \frac{1}{2} \langle Ax_k + \alpha Ad_k, x_k + \alpha d_k \rangle - \langle b, x_k + \alpha d_k \rangle = \\ &= \frac{1}{2} \left(\langle Ax_k, x_k \rangle + 2\alpha \langle Ax_k, d_k \rangle + \alpha^2 \langle Ad_k, d_k \rangle \right) - \langle b, x_k \rangle - \alpha \langle b, d_k \rangle = \\ &= \frac{1}{2} \alpha^2 \langle Ad_k, d_k \rangle + \alpha \langle Ax_k - b, d_k \rangle + \left\langle \frac{1}{2} Ax_k - b, x_k \right\rangle. \end{aligned} \quad (1)$$

Видно, что $f(x_k + \alpha d_k)$ как функция от α является квадратичной и можно аналитически найти точку минимума:

$$\alpha_k = \frac{\langle b - Ax_k, d_k \rangle}{\langle Ad_k, d_k \rangle}. \quad (2)$$

3 Эксперимент: Зависимость числа итераций метода сопряженных градиентов от числа обусловленности и размерности пространства

Эксперимент состоял в следующем: нужно было запустить метод сопряженных градиентов на произвольной квадратичной функции и узнать, как зависит число итераций от числа обусловленности κ и размерности пространства n оптимизируемых параметров.

Для этого генерируется случайным образом матрица $A = \text{Diag}(a_1, \dots, a_n)$, где $\min a_i = 1$, $\max a_i = \kappa$, вектор $b \in \mathbb{R}^n$ — вектор со случайными элементами. Результаты экспериментов приведены на Рис. 1.

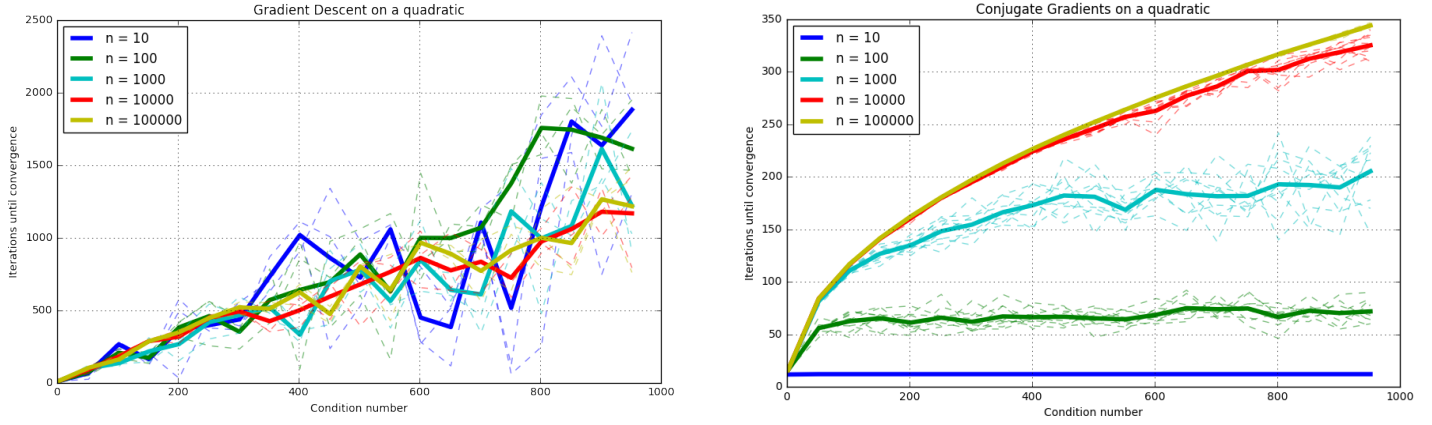


Рис. 1: Зависимость числа итераций от размерности пространства и числа обусловленности. Слева график для градиентного спуска, справа — для метода сопряженных градиентов.

Квадратичная задача генерировалась 10 раз. Пунктиром изображены графики количества итераций в этих 10 запусках, жирной линией — их усреднение.

Видно, что методу сопряженных градиентов при увеличении размерности пространства требуется увеличивается и количество итераций необходимое для сходимости. В отличие от этого в градиентном спуске количество итераций остается одинаковым. Стоит также заметить, что методу сопряженных градиентов в целом нужно меньше итераций для сходимости. Также видно, что для случая $n = 100, 1000$ метод сопряженных градиентов начинает сходиться примерно за одно количество итераций и начинает не зависеть от числа обусловленности, в отличие от градиентного спуска, где с ростом числа обусловленности растет и количество итераций.

4 Эксперимент: Выбор размера истории в методе L-BFGS

В данном эксперименте нужно было запустить метод L-BFGS на реальной задаче логистической регрессии, чтобы проверить как размер истории влияет на скорость сходимости. В качестве параметра регуляризации взято число $\lambda = 1/m$, где m — число объектов в выборке. Стартовая точка $x_0 = 0$. В качестве набора данных был выбран news20.binary. Было построено два графика:

1. Зависимость относительного квадрата нормы градиента $\|\nabla f(x_k)\|_2^2 / \|\nabla f(x_0)\|_2^2$ (в логарифмической шкале) от числа итераций.
2. Зависимость относительного квадрата нормы градиента $\|\nabla f(x_k)\|_2^2 / \|\nabla f(x_0)\|_2^2$ (в логарифмической шкале) от реального времени работы.

Метод L-BFGS на каждой k -й итерации хранит l величин $\{(y_{k-1}, s_{k-1}), \dots, (y_{k-l}, s_{k-l})\}$, где $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$, $s_k = x_{k+1} - x_k$, благодаря которым происходит вычисление следующего направления, то есть метод требует дополнительной памяти $O(ln)$. На каждой итерации производится проход полностью по всей истории l и в каждой данной итерации происходит лишь вычисление скалярных произведений векторов размера n , то есть стоимость одной итерации $O(ln)$.

Результаты эксперимента приведены на Рис. 2. Стоит заметить, что при размере истории 0, метод L-BFGS работает как градиентный спуск. Также при $l = 1$, стоимость итерации всего лишь $O(n)$, что позволяет методу сойтись очень быстро в плане времени, но в плане итераций сходимость будет дольше. По графикам видно, что с увеличением размера истории количество

итераций для сходимости падает, однако из-за стоимости одной итераций по времени каждый метод при $l > 0$ работает примерно одинаково.

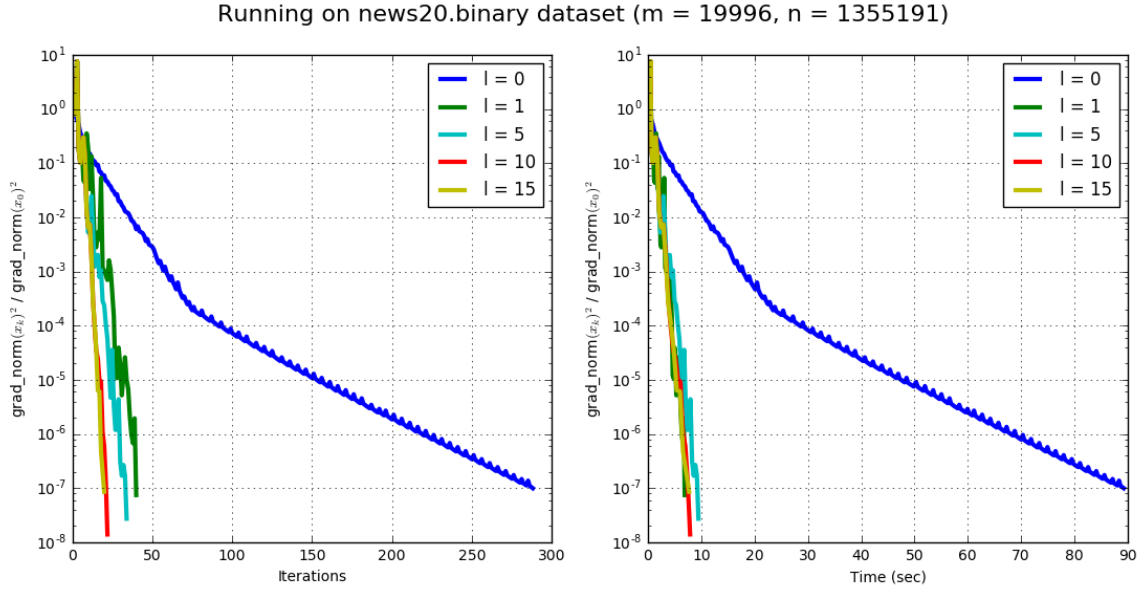


Рис. 2: Зависимость между размером истории и скоростью сходимости.

5 Эксперимент: Сравнение методов на реальной задаче логистической регрессии

В данном эксперименте необходимо сравнить методы градиентного спуска, безгессианного метода Ньютона и L-BFGS на задачах логистической регрессии с l_2 -регуляризатором на датасетах w8a, gisette, news20.binary, real-sim, rcv1.binary. В качестве параметра регуляризации взято число $\lambda = 1/m$, где m — число объектов в выборке. Стартовая точка для всех методов $x_0 = 0$. Параметры методов брались по умолчанию. В качестве сравнения были построены следующие графики:

1. Зависимость значения функции против номера итерации метода.
2. Зависимость значения функции против реального времени работы.
3. Зависимость относительного квадрата нормы градиента $\|\nabla f(x_k)\|_2^2 / \|\nabla f(x_0)\|_2^2$ (в логарифмической шкале) от реального времени работы.

Результаты экспериментов приведены на Рис. 3, 4, 5, 6, 7.

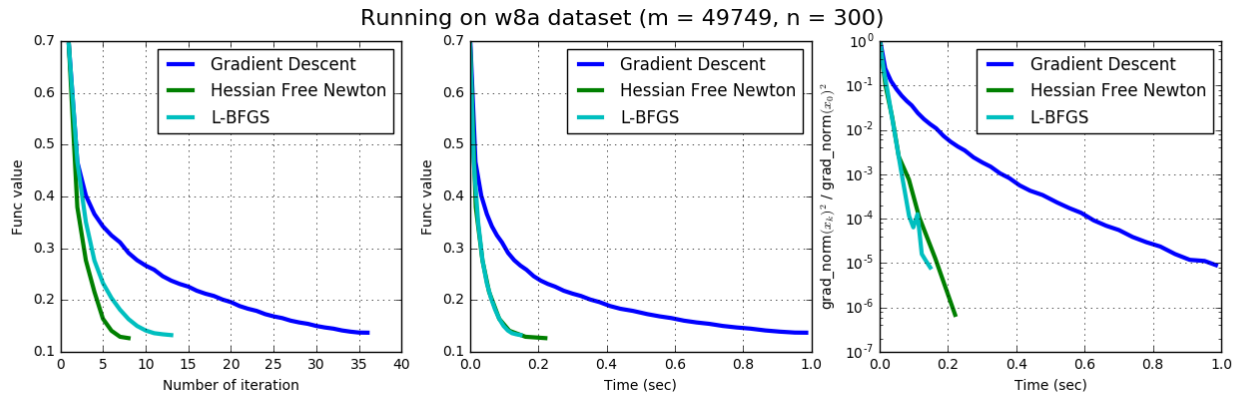
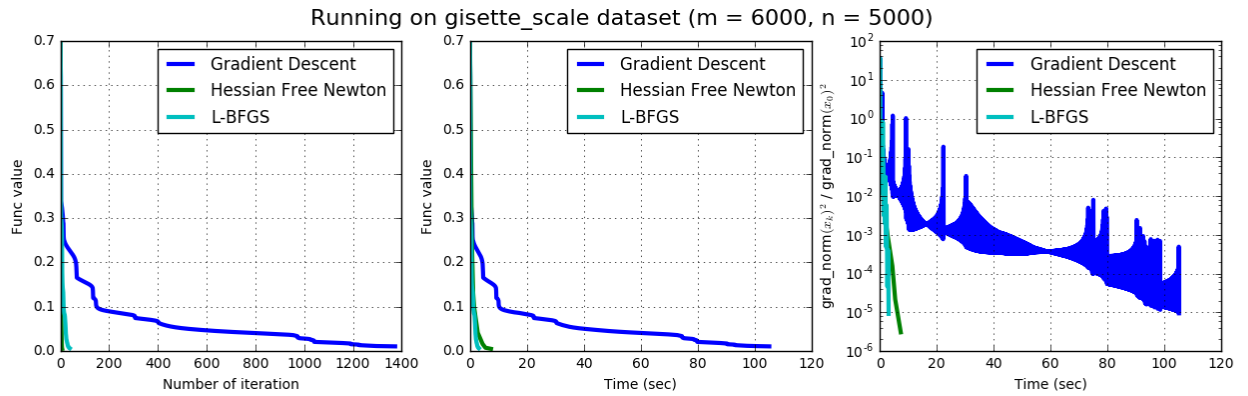


Рис. 3: Датасет w8a.



Датасет gisette.

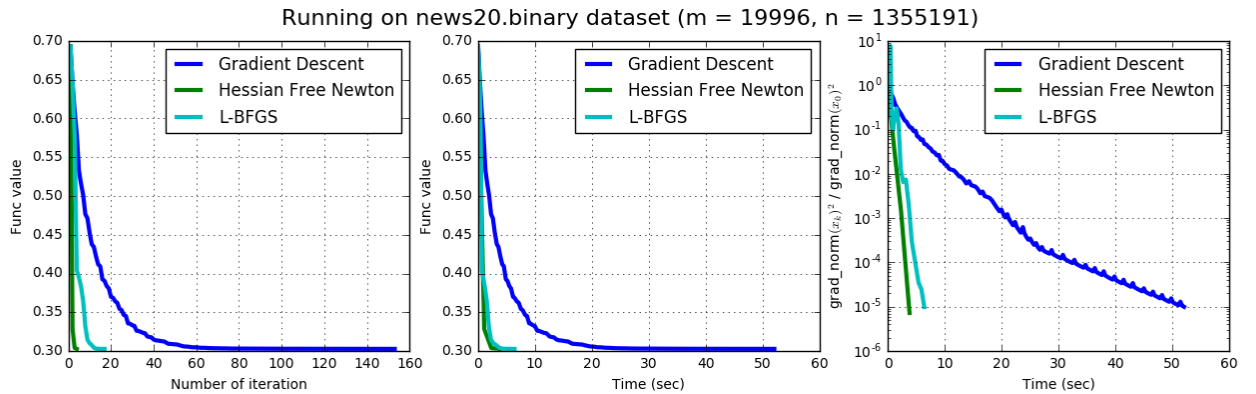


Рис. 4: Датасет news20.binary.

По графикам видно:

1. продвинутые методы HFN и L-BFGS сходятся быстрее по итерациям и по времени в отличие от метода градиентного спуска.
2. В большинстве случаев по итерациям быстрее всего сходится усеченный метод Ньютона, хотя по реальному времени работы данный метод немного уступает L-BFGS. Это можно объяснить тем, что метод HFN использует матрично-векторное умножение на каждой итерации против скалярного произведения в методе L-BFGS. Только на датасете news20.binary HFN работает лучше, чем в L-BFGS во всех трех графиках.

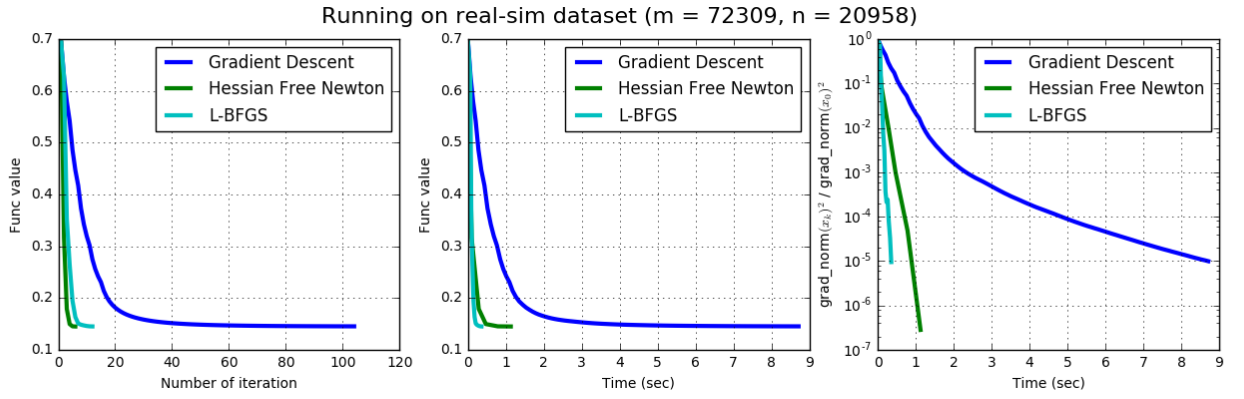


Рис. 5: Датасет real-sim.

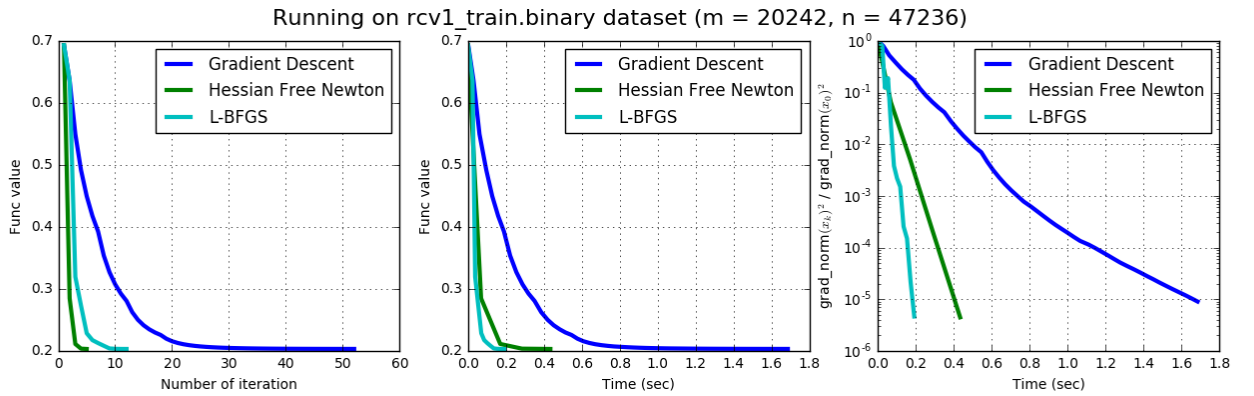


Рис. 6: Датасет rcv1.binary.

6 (Бонусная часть) Эксперимент: Сравнение метода сопряженных градиентов и L-BFGS на квадратичной функции

В данном эксперименте нужно было сравнить метод сопряженных градиентов с методом L-BFGS на квадратичной задаче. Для того чтобы сравнение было честным в качестве шага α_k в методе L-BFGS выбирается шаг, который был выведен ранее. Матрица $A \in \mathbb{S}_{++}^n$, вектор $b \in \mathbb{R}^n$ генерируются также как и в первом эксперименте. Нужно было на изобразить графики сходимости в терминах евклидовой нормы невязки $r_k = Ax_k - b$ (в логарифмической шкале) против номера итерации. Запускалось на двух квадратичных задачах с $\kappa = 100, 10000, n = 6000$. Результаты экспериментов приведены на Рис. 7.

Из графиков видно, что метод L-BFGS с оптимальным шагом работает также как и метод сопряженных градиентов для $l \geq 1$. При $l = 0$ метод L-BFGS работает как градиентный спуск.

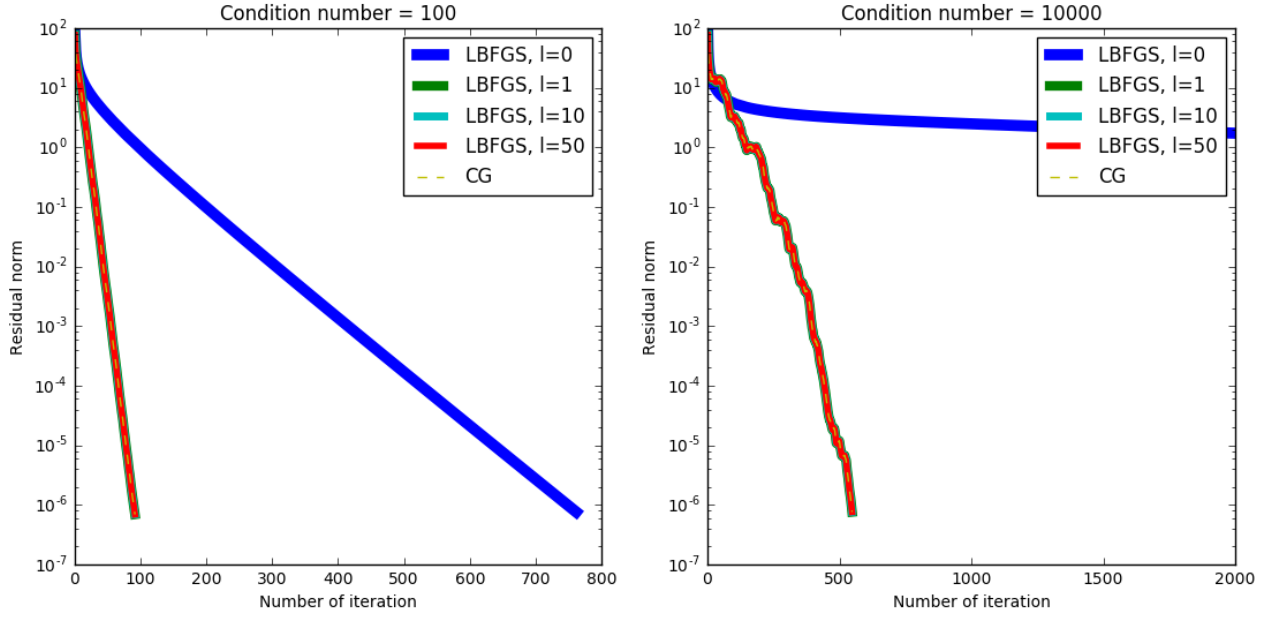


Рис. 7: Сравнение метода сопряженных градиентов (CG) и метода L-BFGS.

7 (Бонусная часть) Эксперимент: Какая точность оптимизации нужна в реальных задачах?

Цель данного эксперимента — узнать какая точность оптимизации нужна в реальных задачах. Для этого были выбраны два датасета: gisette и w8a. Запускался метод оптимизации L-BFGS для логистической задачи с начальной точкой $x_0 = 0$ и коэффициентом регуляризации $\lambda = 1/m$, где m — число объектов в обучающей выборке. В качестве вектора весов бралась итоговая точка x_* , ответы на тестовой выборке считаются по правилу: $\hat{b}_{test} = \text{sgn}(A_{test}x_*)$. В качестве меры качества выбран процент ошибок. Графики построены следующим образом: для каждой точности ϵ запускался метод L-BFGS и находилась процент ошибок на тестовой выборке. Результаты экспериментов приведены на Рис. 8.

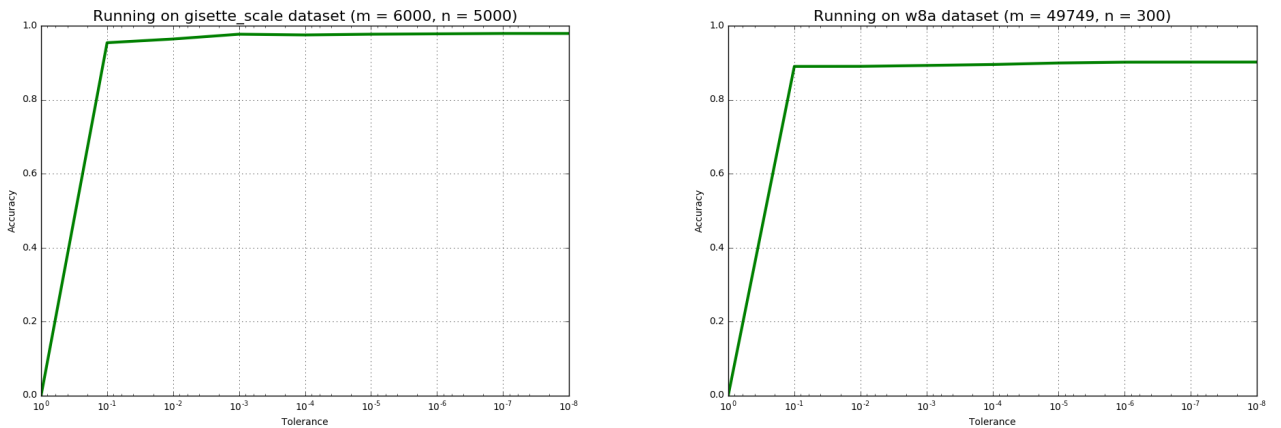


Рис. 8: Зависимость процента ошибок против точности оптимизации. Увеличение точности оптимизации происходит слева направо.

По графикам видно, что при точности $\epsilon = 10^{-3}$ уже достигается хорошее качество на тесто-

вой выборке. Увеличение качества на тестовой выборке по мере увеличения точности незначительно.

8 Выводы

По данному практическому заданию можно сделать следующие выводы:

1. Метод сопряженных градиентов для произвольной квадратической задачи сходится быстрее, чем градиентный метод и с увеличением размерности пространства количество итераций увеличивается. Также метод практически не зависит от числа обусловленности κ .
2. Чем больше размер истории в L-BFGS, тем быстрее метод сходится по итерациям, но реальное время сходимости почти одинаковое.
3. Во всех задачах логистической регрессии усеченный метод Ньютона сходиллся быстрее по итерациям, но по реальному времени работы уступал методу L-BFGS, однако оба данных метода работают быстрее градиентного спуска.
4. Метод L-BFGS и метод сопряженных градиентов работают одинаково для $l > 0$ на произвольной квадратичной задаче с оптимальным шагом α_k .
5. В реальных задачах достаточно выбрать точность оптимизации равной 10^{-3} . Данной точности хватит, чтобы достигнуть желаемого качества на тестовой выборке.