Факультет компьютерных наук
Основная образовательная программа
Прикладная математика и информатика

Отчет по практическому домашнему заданию 2

Курс "методы оптимизации в машинном обучении"

Выполнил Петров Олег Евгеньевич, 193 группа

Содержание

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

1.1 Дизайн эксперимента

Основная цель данного эксперимента – проанализировать и сравнить траектории градиентного спуска и метода сопряженных градиентов, определить их зависимость от размерности пространства и числа обусловленности матрицы. Для этого мы для каждой размерности пространства n на всех доступных числах обусловленности k запустим градиентный спуск на сгенерированных данных (seed фиксирован) и посчитаем $T(n, k)$ - число итераций. Повторим процедуру 10 раз и отобразим на графике $T(n, k)$ против k семейств функций для каждого n (выделим каждое семейство разным цветом). Самый первый запуск процедуры характеризуется непрозначными линиями, прозрачность уменьшается по мере повторения процедуры.

Рассматриваем следующие данные:

- $n = 10^i,\ i \in \{1, 2, 3, 4, 5, 6\}$
- k – натуральное число от 10 до 1070 с шагом 50
- Матрица A_k генерируется из случайного вектора $a \sim U_n[1; k]$ размера n, при этом после генерации вектора двум различным позициями присваиваются значения $1,\ k$. После $A_k = \mathrm{diag}(a)$
- Вектор b_k генерируется из многомерного равномерного распределения $U[-k; k]$
- Стартовая точка x генерирутся из симметричного равномерного распределение (см. результаты)

Оба метода используют гиперпараметры по умолчанию. Если для какой-нибудь пары $n,\ k$ метод не сошелся, тогда $T(n, k) = 0$.

1.2 Результаты

Результаты эксперимента, как оказалось, сильно зависят от стартовой точки для GD и вовсе не зависят для CG.

На рисунке 2 $x_0 \sim U_n[-0.05; 0.05]$, на рисунке 3 – $x_0 \sim U_n[-50; 50]$. На рисунке 3 кажется, что при росте n число обусловленности k матрицы A не влияет на сходимость – практически на всем отрезке при $n \geqslant 1000$ число итераций не превышает 50, но и не достигает нуля – значит, метод сходится. На рисунке 2 при любом n имеется тенденция к увеличению числу итераций при параллельном увеличении числа обусловленности. Интересно то, что при росте размерности пространства n графики семейств функций имеют более сглаженный характер. Например, этот эффект хорошо заметен, если сравнивать первые два порядка размерности с двумя последними.

На рисунке 1 $x_0 \sim U_n[-10; 10]$, однако другие параметры распределения показывают абсолютно идентичные результаты. Если размерность пространства небольшая $n = 10$, то число обусловленности матрицы никак не влияет на число итераций. То же видно и с $n = 100$ в аппроксимации, начиная с $\text{cond(A)} = 200$. Имеется в виду то, что число итераций растет по мере роста числа обусловленности матрицы до значения $\text{cond(A)} = 200$. Далее график имеет ярко выраженные колебания, но четкой закономерности на рост числа итераций нет. Далее, чем выше порядок, тем больше число обусловленности влияет на число итераций и тем более сглажен график (тем меньше колебаний). Для $n = 10^4,\ 10^5$ отчетливо прослеживается линейный (или логарифмический) рост.

Стоит отметить, что метод сопряженных градиентов работает за меньшее число итераций, чем градиентный спуск – для $\text{cond(A)} > 1000$ и $n = 10$ CG сходится за 140 итераций, в то время как градиентный спуск – за 500. Кроме того, метод сопряженных градиентов хорошо работает на маленьких значениях размерности.

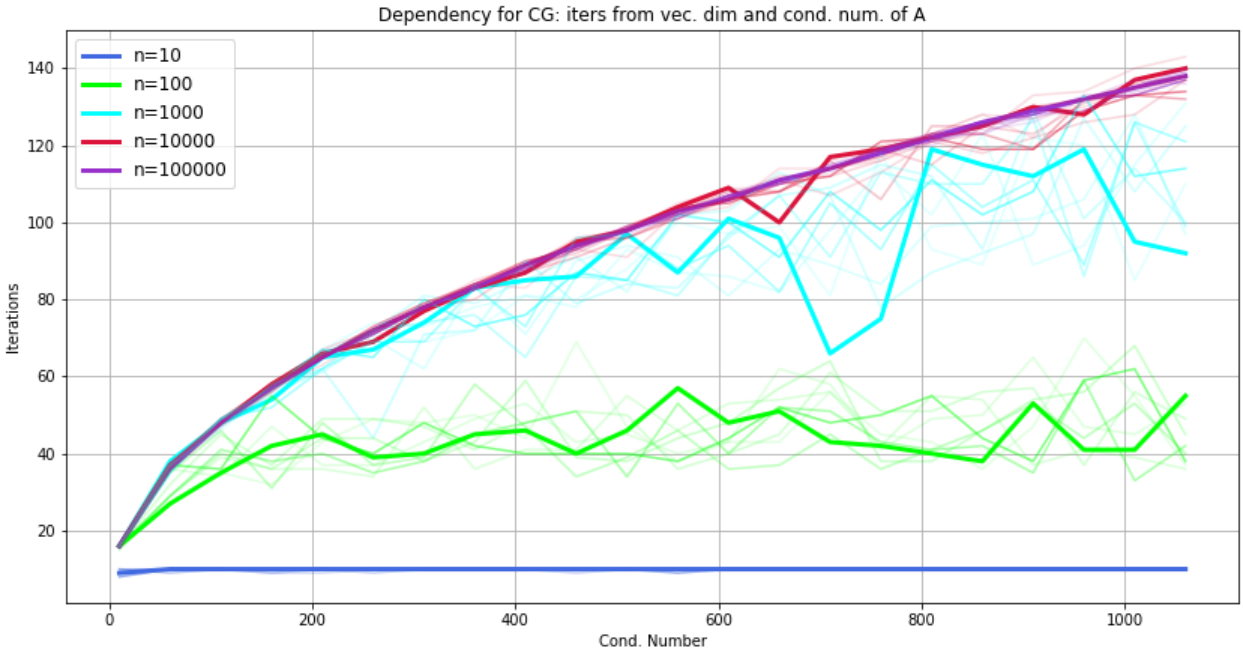

Рис. 1. Зависимость для CG числа итераций от числа обусловленности.

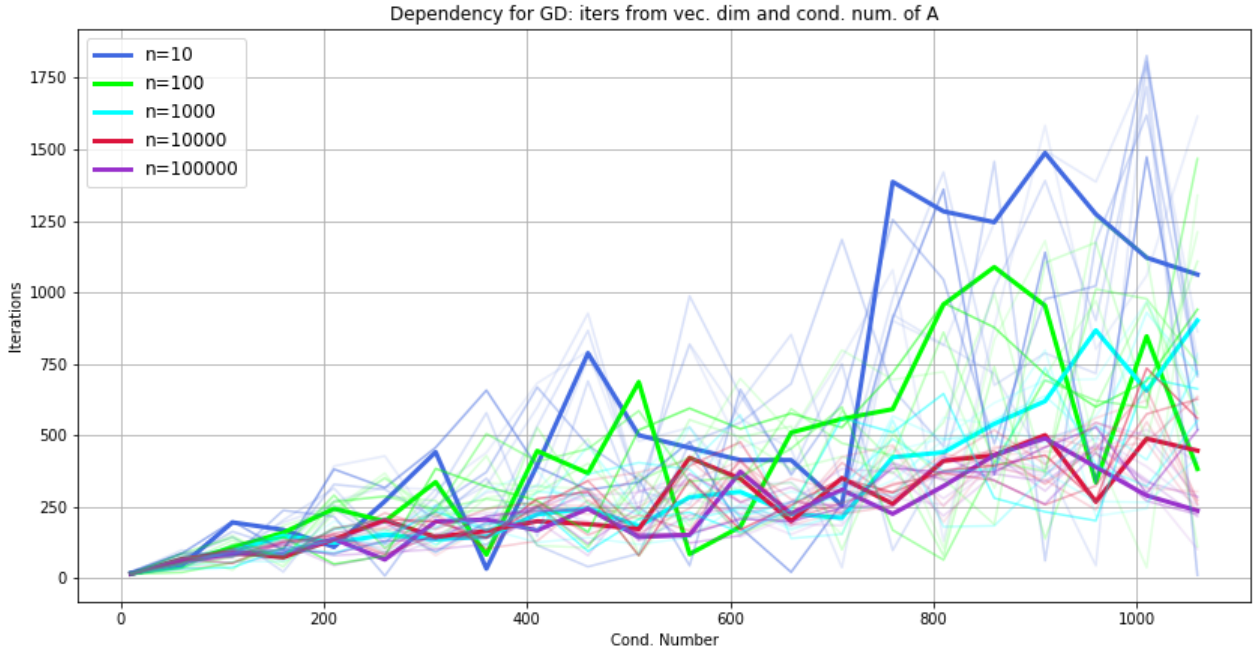

Рис. 2. Зависимость для GD числа итераций от числа обусловленности. x_0 близок к нулю.

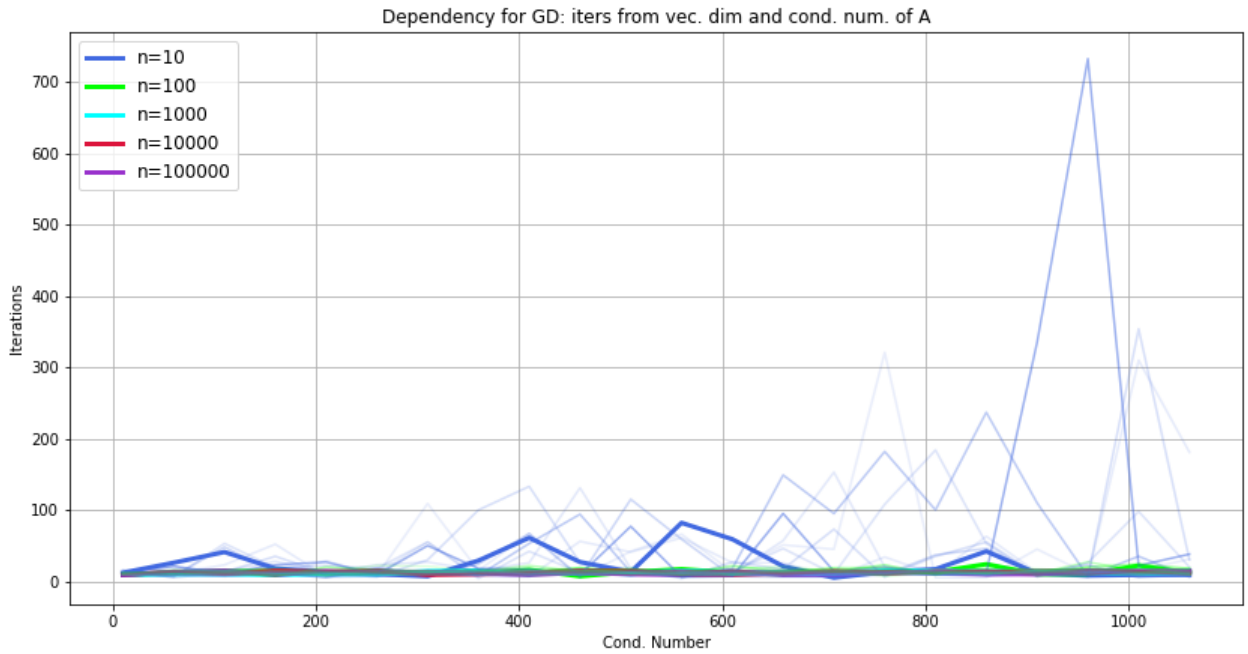

Рис. 3. Зависимость для GD числа итераций от числа обусловленности. x_0 далек от нуля.

1.3 Выводы

- Увеличение размерности пространства сглаживает закономерности роста числа итерация при росте числа обусловленности;
- В отличие от GD, у метода сопряженных градиентов амплитуда колебаний меньше, для низкоразмерных пространств число итераций не зависит от числа обусловленности;
- Сходимость CG не зависит от близости стартовой точки к нулю (для квадратичной функции).

Общий вывод: метод сопряженных градиентов в целом работает лучше, чем метод градиентного спуска, однако CG может работать тольско с симметричной положительно определенной матрицей.

2 Эксперимент: Выбор размера истории в методе L-BFGS

2.1 Дизайн эксперимента

Основная цель данного эксперимента – на задаче логистической регрессии оценить влияние размера хранимой истории ℓ (гиперпараметра) на сходимость метода L-BFGS на реальном наборе данных для бинарной классификации, отрисовать и проанализировать графики зависимости логарифма $\frac{\|\nabla f(x_k)\|^2}{\|\nabla f(x_0)\|^2}$ против номера итерации и реального времени работы метода.

Для размера истории рассматриваются значения $\ell \in \{0, 1, 5, 10, 25, 50, 75, 100\}$, набор данных - news20.binary с сайта LIBSVM, коэффициент регуляризации $= \frac{1}{m}$, где m – размерность пространства параметров (весов), начальная точка нулевая.

2.2 Результаты

На рисунке 4a) отчетливо видно, что чем меньше памяти хранится, тем больше требуется итераций для сходимости, при этом при существенном увеличении ℓ заметно несущественное уменьшение числа итераций. К примеру, увеличивая $\ell = 5$ до $\ell = 10$, мы уменьшаем число итераций всего на 1-2 единицы.

Заметно, что размер истории не может быть больше, чем число итераций на $\ell = 0$. В данном случае это чуть больше 50 итераций, поэтому графики для $\ell = 50,\ 75,\ 100$ совпали.

Рисунок 4b похож на 4a. Примечательно то, что графики для $\ell = 50,\ 75,\ 100$ не совпали полностью, однако они достаточно близки по времени. Заметна та же тенденция – увеличение ℓ начиная со значения 10 не приводит к весомому выигрышу по времени.

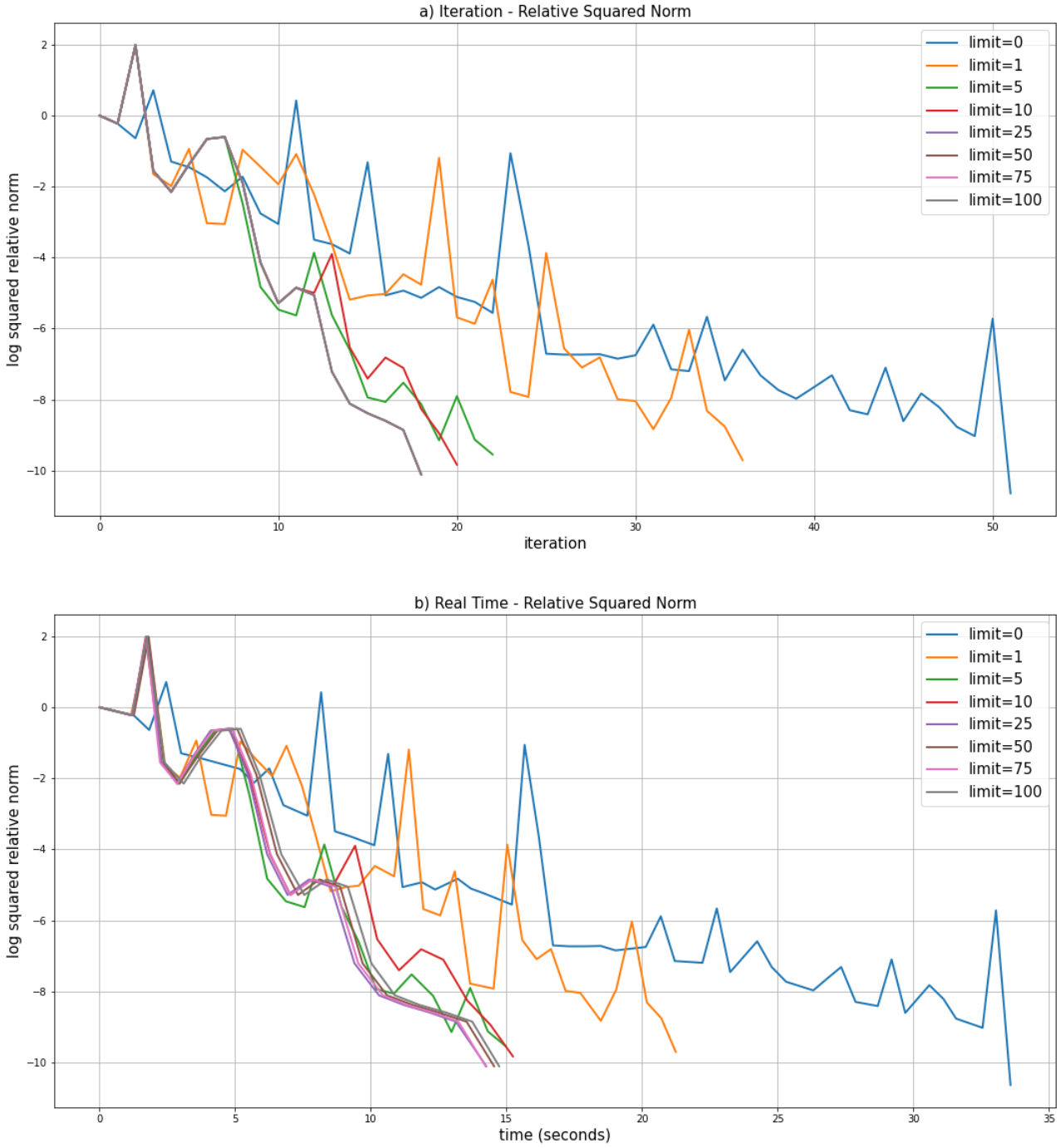

Рис. 4.

2.3 Выводы

Не имеет смысла брать большое ℓ, скорее всего выигрыш по итерациям и времени будет незначителен. В данном случае $\ell = 10,\ 25$ можно считать оптимальными значениями.

3 Эксперимент: Сравнение методов на реальной задаче логистической регрессии

3.1 Дизайн эксперимента

Основная цель эксперимента – взять несколько реальных наборов данных для бинарной классификации, запустить на задаче логистической регрессии метод градиентного спуска, HFN и L-BFGS; для каждого набора данных построить графики значения функции против итерации, значения функции и $\log \frac{\|\nabla f(x_k)\|^2}{\|\nabla f(x_0)\|^2}$ против реального времени работы; на каждом графике отобразить все методы, проанализировать и определить, какой метод лучше и в каких ситуациях.

Используемые наборы: w8a, gisette, real-sim, rcv1.binary, news20.binary; коэффициент регуляризации $\frac{1}{m}$; начальная точка нулевая.

3.2 Результаты

Во-первых, заметим, что на всех наборах данных (рис. 5-9) графики Iteration – Function Value и Real Time – Function Value полностью идентичны. Это говорит о том, что если тот или иной метод быстрее сходится по реальному времени, то он также быстрее сходится по итерациям. Кроме того, на каждом из этих графиков все методы достигают практически одного и того же значения функции. Общая тенденция: HFN работает быстрее всего, L-BFGS чуть хуже и GD – крайне медленно. Например, на рис. 9, набор news20, GD сходится за 400 секунд, в то время как другие два метода менее чем за 25. Примечательно, что чем больше размерность вектора параметров (вторая размерность матрицы), тем сильнее разрыв. Это заметно, если сравнить рис. 8 и рис. 9, наборы rcvl и news20 соответственно: число объектов порядка двадцати тысяч, однако размерность матрицы на рис. 9 существенно больше.

На всех графиках Real Time – Relative Squared Norm (в логарифмической шкале) отчетливо видно, что HFN наилучшим образом приближает

$\|\nabla f(x_{opt})\|$, L-BFGS – чуть лучше, и метод градиентного спуска – посредственно. Кроме того, на рисунке 6 заметно, что колебания GD намного больше по амплитуде колебаний других методов, а график Real Time – Relative Squared Norm говорит о том, что метод градиентного спуска гораздо менее стабилен других методов (возможно, на тех наборах данных, где число объектов мало или сравнимо с числом признаков).

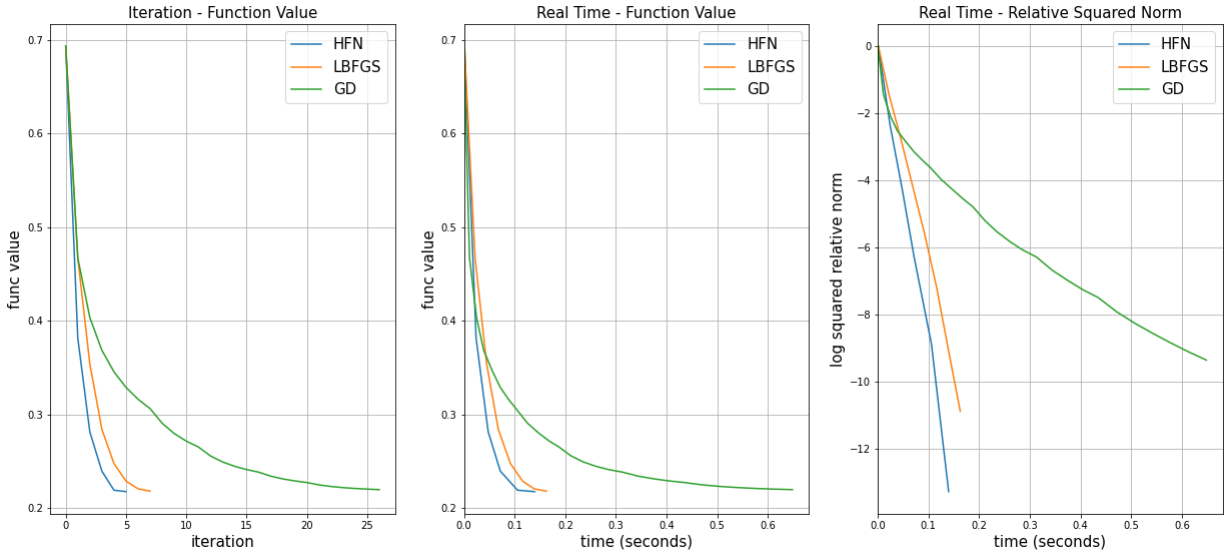

Рис. 5. w8a, размерность матрицы: (49749, 300)

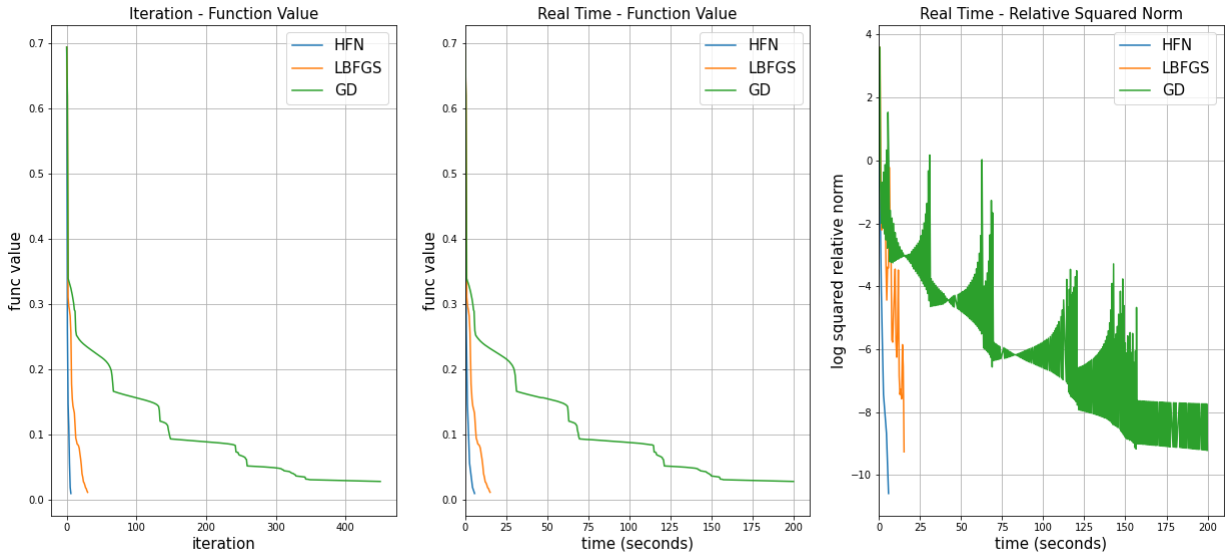

Рис. 6. gisette, размерность матрицы: (6000, 5000)

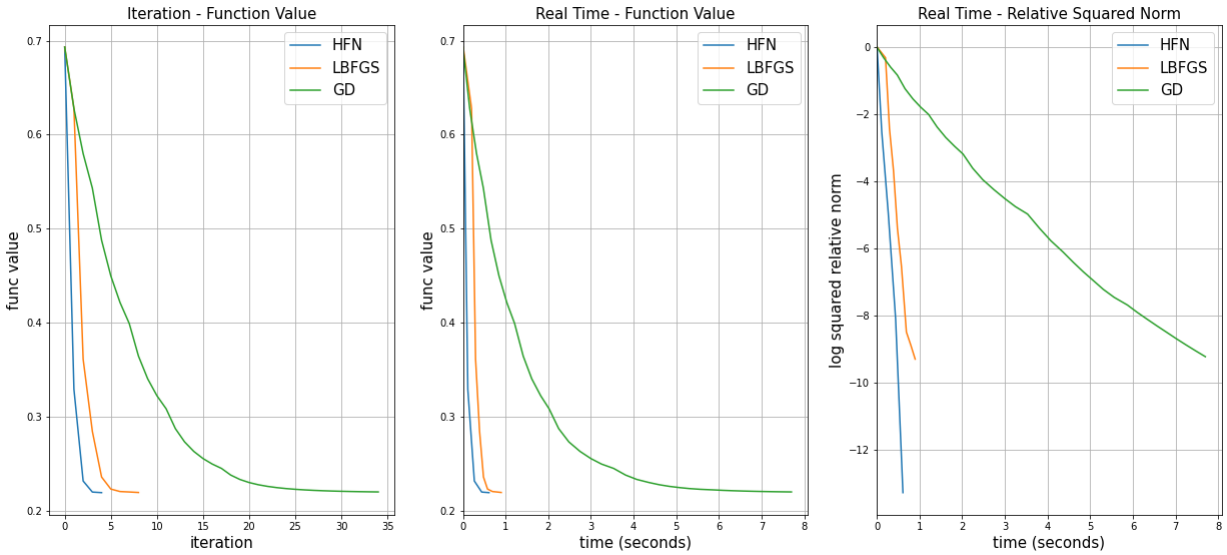

Рис. 7. real-sim, размерность матрицы: (72309, 20958)

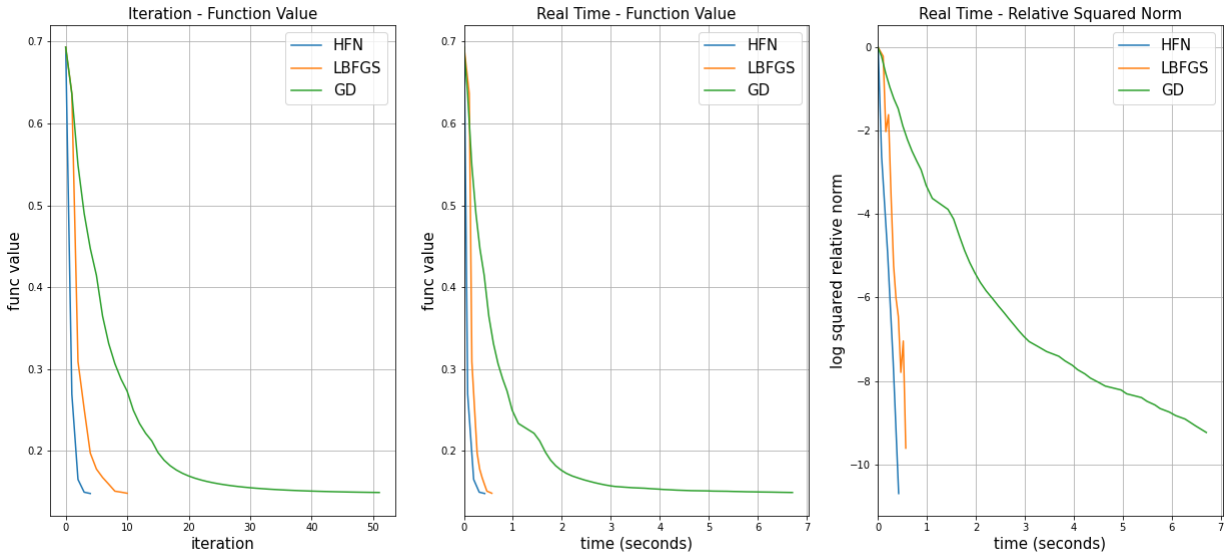

Рис. 8. rcv1, размерность матрицы: (20242, 47236)

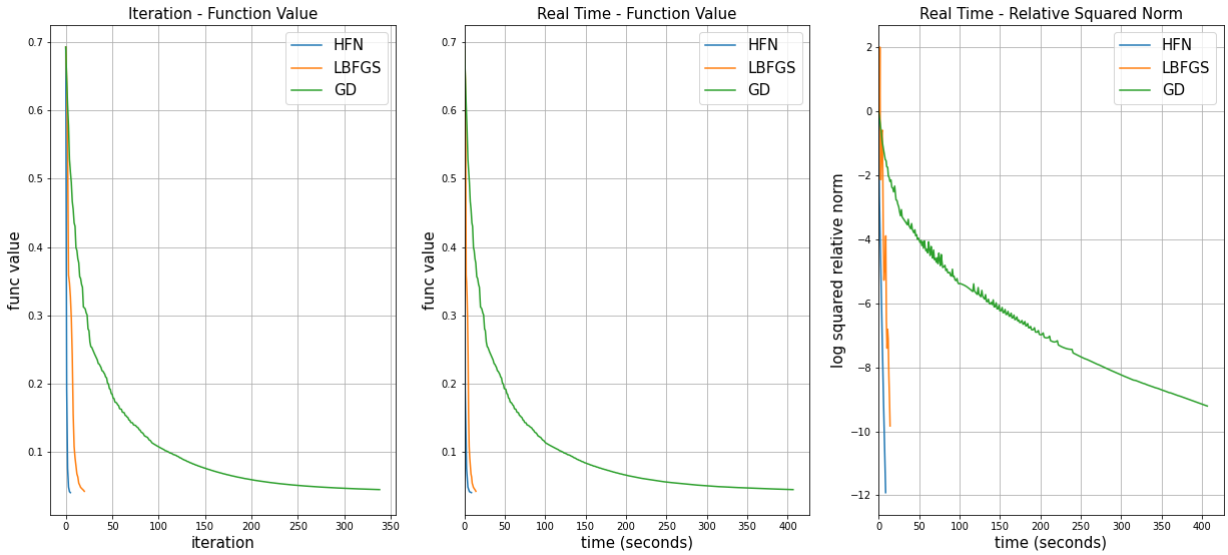

Рис. 9. news20, размерность матрицы: (19996, 1355191)

3.3 Выводы

На всех используемых наборах данных сохраняется одна и та же тенденция: HFN лучше всего приближает норму градиента в оптимальной точке и быстрее всего сходится, L-BFGS – чуть хуже и GD – хуже всех. Не было обнаружено противоречивых ситуаций, поэтому по результатам эксперимента можно сказать, что HFN - лучший выбор.