Методы оптимизации в машинном обучении
Отчет по практическому заданию №4

Коробов Павел, группа 517

20 июня 2020 г.

1 Введение.

В данном задании мы изучим композитную оптимизацию и субградиентный метод на примере модели LASSO-регрессии.

2 Оптимизация модели LASSO-регрессии

Сформулируем задачу LASSO-регрессии. Пусть имеется обучающая выборка $((a_i, b_i))_{i=1}^m$, где $a_i \in \mathbb{R}^n$ – вектор признаков i-го объекта, а $b_i \in \mathbb{R}$ – его регрессионное значение. Задача заключается в прогнозировании регрессионного значения b_{new} для нового объекта, представленного своим вектором признаков a_{new} . Коэффициенты x модели настраиваются с помощью решения следующей оптимизационной задачи:

$$\frac{1}{2}\sum_{i=1}^{m}(\langle a_i, x\rangle - b_i)^2 + \lambda\sum_{j=1}^{n}|x_j| = \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1 \to \min_{x\in\mathbb{R}^n}.$$

Здесь $\lambda > 0$ – коэффициент регуляризации.

2.1 Субградиентный метод

Субградиентный метод – это метод решения безусловной негладкой выпуклой задачи оптимизации

$$\min_{x\in Q}\phi(x),$$

где $\phi : Q \to \mathbb{R}$ – выпуклая функция с ограниченными субградиентами, определённая на выпуклом замкнутом множестве $Q \subset \mathbb{R}^n$. Итерация субградиентного

метода заключается в шаге из текущей точки x_k в направлении произвольного антисубградиента $\phi'(x_k)$, затем выполяется проекция на множество Q. Так как для негладких задач норма $\phi'(x_k)$ не является информативной, мы будем использовать в качестве направления нормированный субградиент.

$$x_{k+1} = P_Q\left(x_k - \alpha_k \frac{\phi'(x_k)}{\|\phi'(x_k)\|}\right)$$

Для сходимости в качестве длин шагов выберем $\alpha_k = \frac{\alpha}{\sqrt{k+1}}$, где $\alpha > 0$. В таком случае мы получим скорость сходимости $O\left(\frac{\log k}{k}\right)$, что является наилучшим возможным вариантом в случае субградиентного метода.

В нашем случае $\phi(x) = \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1$, $Q = \mathbb{R}^n$.

В качестве субградиента можем выбрать

$$\phi'(x) = A^T(Ax - b) + \lambda\,\mathrm{sgn}(x).$$

2.2 Проксимальный градиентный метод

Рассмотрим проксимальный градиентный метод и его модификацию – ускоренный градиентный метод Нестерова.

Задача композитной минимизации представляет собой задачу минимизации специального вида:

$$\min_{x\in Q}\left(f(x) + h(x)\right),$$

где $h : Q \to \mathbb{R}$ – выпуклая замкнутая функция, для которой возможно эффективное вычисление проксимального оператора

$$\mathrm{Prox}_{\lambda h}(x) := \operatorname*{argmin}_{y\in Q}\left\{\lambda h(y) + \frac{1}{2}\|y - x\|^2\right\}$$

для любого $\lambda > 0$, $f : E \to \mathbb{R}$ – дифференцируемая функция с липшецивым градиентом, определенная на открытом нможестве $E \subset Q$.

В нашем случае $h = \|\cdot\|_1$:

$$\mathrm{Prox}_{\lambda\|\cdot\|_1}(x) = \left(\mathrm{sign}\,(x_i)\,[|x_i| - \lambda]_+\right)_{1\le i\le n}.$$

Будем минимизировать следующую модель в точке:

$$F(x) = f(x) + h(x) \approx m_k(x) = \\ = f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{L}{2}\|x - x_k\|^2 + h(x) \to \min_{x\in\mathbb{R}^n}.$$

Отсюда следует, что $x_{k+1} = \text{Prox}_{\frac{\lambda}{L}\|\cdot\|_1}(x_k - \frac{1}{L}\nabla f(x_k))$ при условии, что

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{L}{2}\|x - x_k\|^2,$$

что необходимо для сходимости алгоритма. В силу липшицевости f при при некотором L это неравенство обязательно выполнится. Константа L адаптивно подбирается в ходе работы лагоритма.

Скорость сходимости для функции $f \in C_L^{1,1}$ составляет $O\left(\frac{1}{k}\right)$.

2.3 Ускоренный градиентный метод Нестерова

Ускоренный градиентный метод Нестерова построен с целью достичь оптимальной скорости сходимости, доступной методам первого порядка.

Он строится на основе минимизации следующей модели:

$$\psi_k(x) = \frac{1}{2}\|x - x_0\|^2 + \sum_{i=1}^{k} a_i \left(f(x_i) + \nabla f(x_i)^T (x - x_i) + h(x)\right)$$

Коэффициенты a_i подбираются некоторым образом так, чтобы достичь максимальной возможной скорости сходимости.

Скорость сходимости для функции $f \in C_L^{1,1}$ составляет $O\left(\frac{1}{k^2}\right)$.

3 Эксперименты

В экспериментах параметры, если их значения не указаны явно, взяты стандартными:

- $\alpha_0 = 1$, $\varepsilon = 10^{-2}$, max_iter $= 1000$ для субградиентного метода
- $L_0 = 1$, $\varepsilon = 10^{-5}$, max_iter $= 1000$ для проксимальных методов

Стандартное значение коэффициента регуляризации $\lambda = 1$.

Будем использовать критерий останова по зазору двойственности: $\eta(x, \mu(x)) < \varepsilon$.

3.1 Изучение зависимости сходимости субградиентного метода от начальной длины шага α_0 и начальной точки x_0

Будем перебирать α_0 по сетке: $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 5\}$.

За начальную точку x_0 будем брать нулеой вектор, единичной вектор и вектор семпированный из стандартного нормального распределения.

Будем оценивать алгоритмы на наборах данных bodyfat и housing, взятых с сайта LIBSVM.

Данные	m	n
bodyfat	252	14
housing	506	13

Ниже представлены графики , иллюстрирующие работу методов. Максимальное число итераций равно 10^6.

3.1.1 Набор данных bodyfat

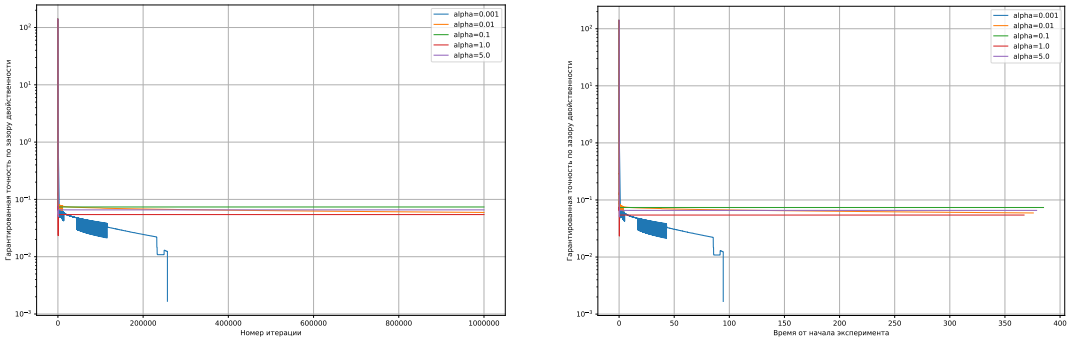


Рис. 1: Поведение субградиентного метода при разных значениях α_0 с начальной точкой $x_0 = 0_n$ на наборе данных bodyfat

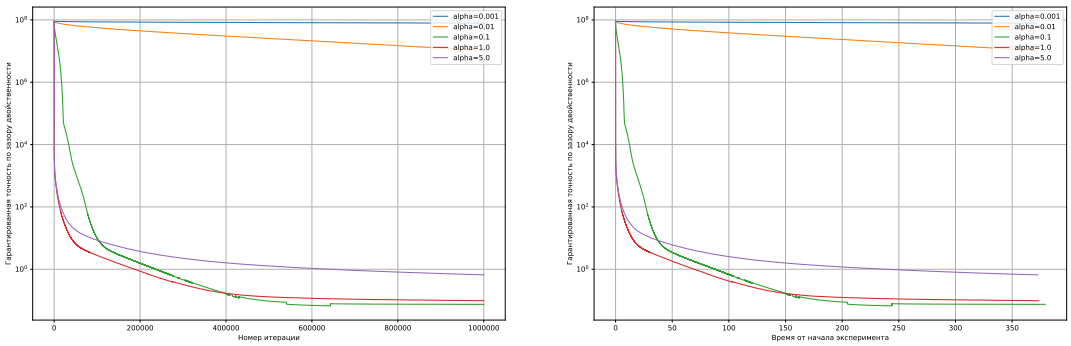


Рис. 2: Поведение субградиентного метода при разных значениях α_0 с начальной точкой $x_0 = 1_n$ на наборе данных bodyfat

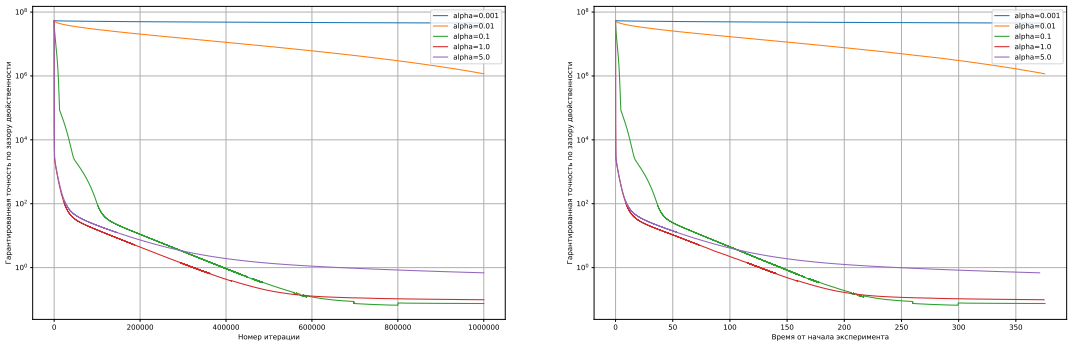


Рис. 3: Поведение субградиентного метода при разных значениях α_0 с начальной точкой $x_0 \sim N(0, I_n)$ на наборе данных bodyfat

Влияние выбора начальной точки явно видно: в случае $x_0 = 0_n$ мы наблюдаем единственный случай сходимости за весь эксперимент ($\alpha = 0.001$), но при этом мы видим «застревание» метода при других значениях α_0. Тем не менее, субградиентный метод всё равно достигает меньших значений зазора двойственности с x_0, чем с другими начальными точками.

При значениях $\alpha_0 \geq 0.1$ на последних двух парах графиков мы имеем примерно одинаковые наблюдения, а при $\alpha_0 \leq 0.01$ зазор двойственности практически не уменьшается.

Значение $alpha_0 = 0.001$ является одновременно и лучшим, и худшим при разном выборе начальной точке. Поэтому можно предположить, что есть некоторая зависимость между длиной шага и начальной точкой.

3.1.2 Набор данных housing

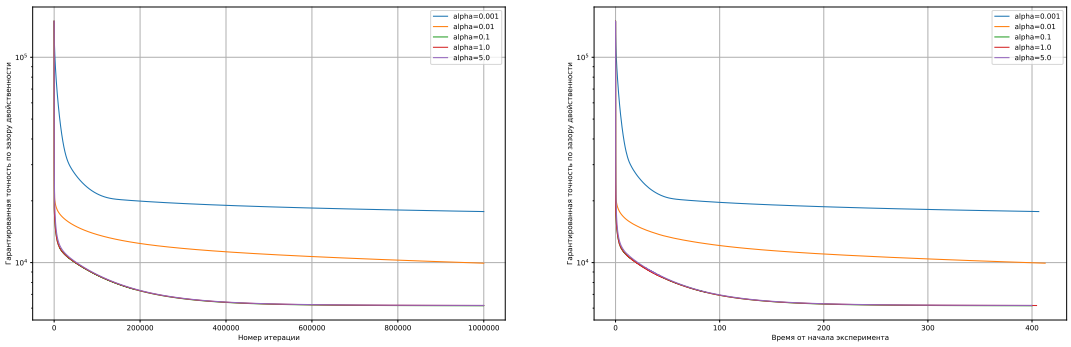


Рис. 4: Поведение субградиентного метода при разных значениях α_0 с начальной точкой $x_0 = 0_n$ на наборе данных housing

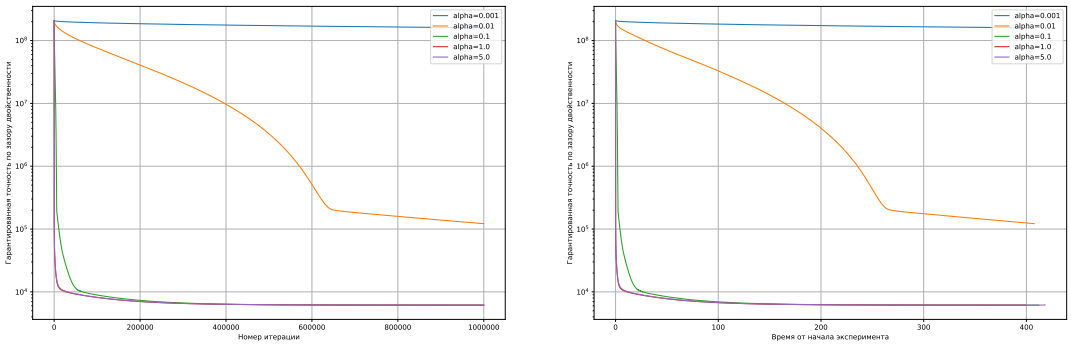


Рис. 5: Поведение субградиентного метода при разных значениях α_0 с начальной точкой $x_0 = 1_n$ на наборе данных housing

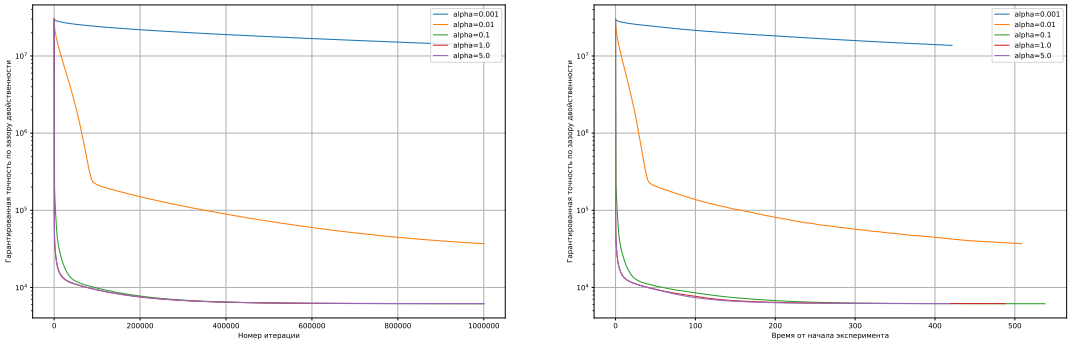


Рис. 6: Поведение субградиентного метода при разных значениях α_0 с начальной точкой $x_0 \sim N(0, I_n)$ на наборе данных housing

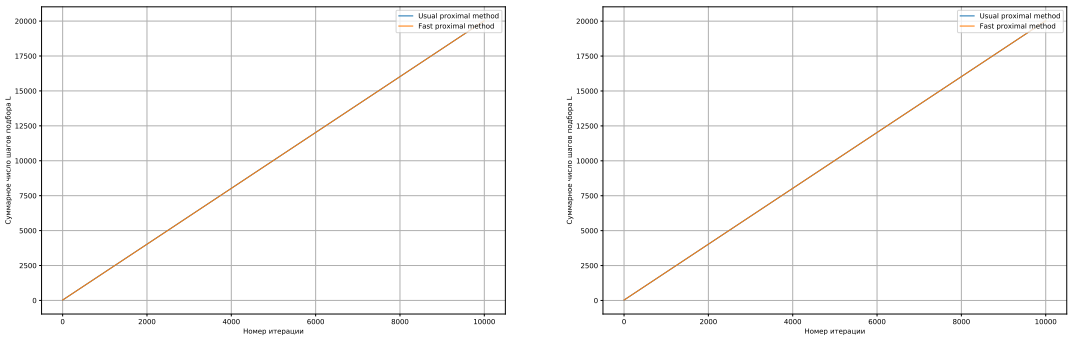
При $\alpha \geq 0.1$ кривые сливаются в одну кривую, находящуюся существенно ниже остальных. На этот раз точка $x_0 = 0_n$ оказалась наоборот самой неудачной. Здесь не прослеживается связь между выбором начальной точки и начальной длиной шага, но выбор начальной точки существенно влияет на работу субградиентного метода сам по себе.

В целом, кажется, что хорошим выбором для начальной длины шага будут значения порядка 10^{-1} или 10^0, но, как мы видели на примере предыдущего набора данных, могут быть и исключения.

3.1.3 Среднее число итераций одномерного поиска в градиентных методах

Проверим, действительно ли число итераций одномерного поиска для подбора L в проксимальных градиентных методах примерно равно 2. Построим графики суммарного числа итераций одномерного поиска для стандартной версии алгоритма и

для его ускоренной версии на наборах данных housing и bodyfat. Максмальное число итераций равно 10000.



(a) Набор данных bodyfat

(b) Набор данных housing

Рис. 7: Суммарное число итераций одномерного поиска для градиентных методов

Графики для обеих версий градиентного метода неразличимы на вид. На обоих наборах данных наблюдается одинаковая картина. Видно, что кривые на графиках выглядят как прямая с коэффициентом наклона равным 2. Это подтверждает, что в среднем делается примерно две итерации одномерного поиска.

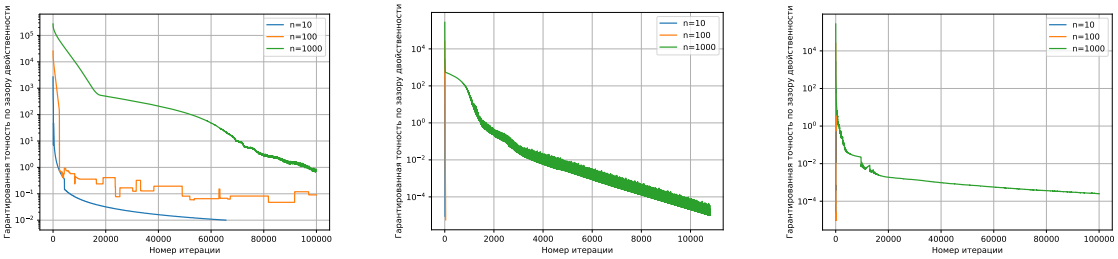
3.2 Зависимость работы методов от размерности пространства n, размера выборки m и коэффициентра регуляризации λ

Будем генерировать компоненты матрицы объектов-признаков и целевого вектора из стандартного нормального распределения.

За начальную точку возьмём $x_0 = 0$.

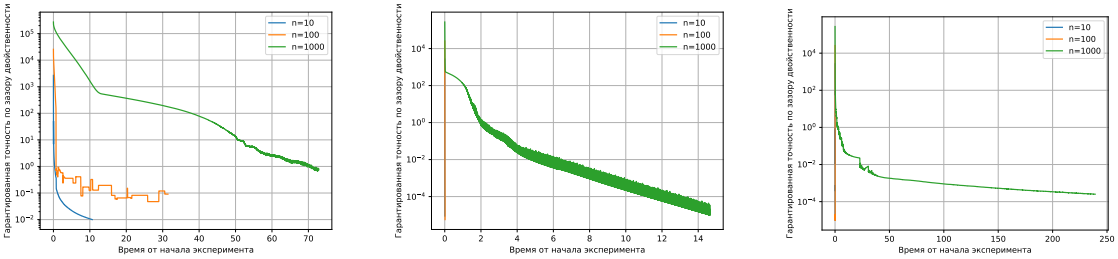
Положим $m = 100$, $n = 100$, $\lambda = 1.0$ как стандартные значения параметров. Будем изменять каждый из них в отдельности и брать оставшиеся равными стандартным значениям.

3.2.1 Зависимость от n



(a) субградиентный метод (b) градиентный методы (c) уск. градиентный метод

Рис. 8: Гарантированная точность по зазору двойственности в зависимости от итерации при разных n



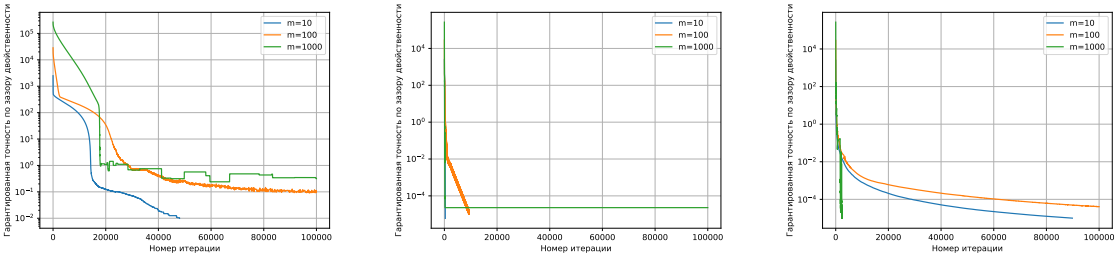
(a) субградиентный метод (b) градиентный метод (c) уск. градиентный метод

Рис. 9: Гарантированная точность по зазору двойственности в зависимости от времени при разных n

Видно, что градиентные методы работают существенно быстрее, чем субградиентный метод, как и ожидается. Как ни странно, ускоренный метод сходится существенно медленнее (обратите внимание, что шкала абсцисс для градиентного метода оканчивается районе 10500 итераций).

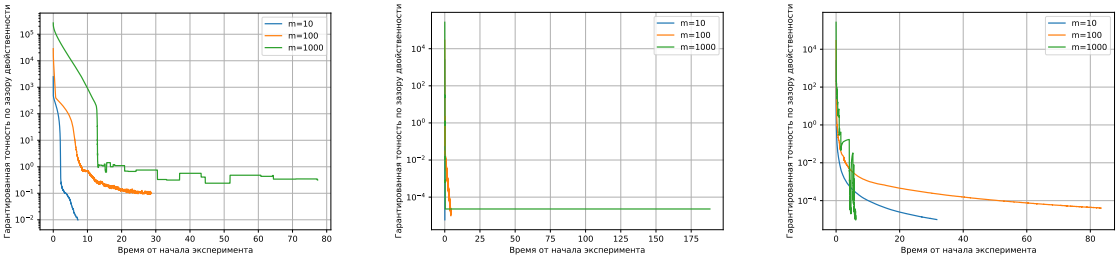
Как и ожидается, чем больше n, тем больше итераций и времени требуется алгоритмам.

3.2.2 Зависимость от m



(a) субградиентный метод (b) градиентный метод (c) уск. градиентный метод

Рис. 10: Гарантированная точность по зазору двойственности в зависимости от итерации при разных m



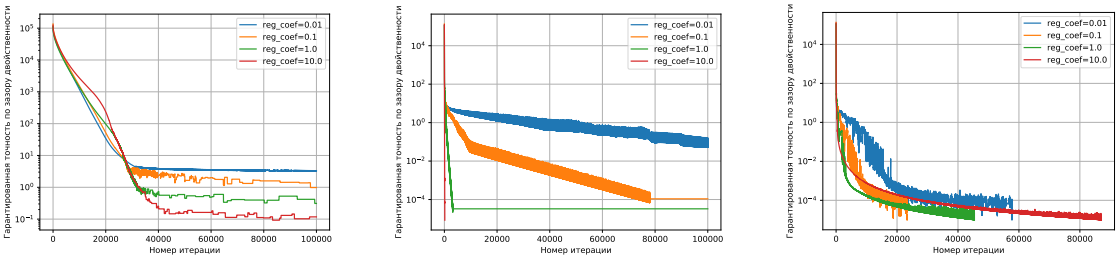
(a) субградиентный метод (b) градиентный метод (c) уск. градиентный метод

Рис. 11: Гарантированная точность по зазору двойственности в зависимости от времени при разных m

Здесь мы видим довольно неоднозначные результаты. В целом, градиентные методы справились лучше субградиентного. При этом градиентный метод по какой-то причине застрял на некотором значении зазора двойственности при $m = 1000$, а ускоренная версия достигла довольно быстрой сходимости. При этом при $m = 100$ ускоренный градиентный метод не сошелся в отличии от обычного градиентного метода, а при $m = 10$ сошелся гораздо позже.

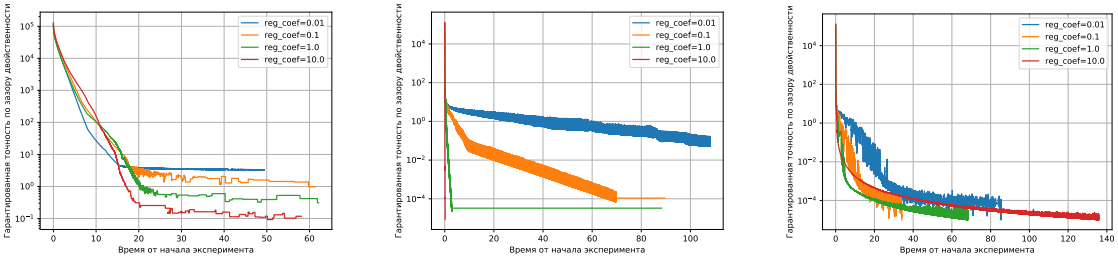
Однозначный рост числа требуемых итераций для достижения определенного значения зазора двойственности виден только для субградиентного метода.

3.2.3 Зависимость от λ



(a) субградиентный метод (b) градиентный метод (c) уск. градиентный метод

Рис. 12: Гарантированная точность по зазору двойственности в зависимости от итерации при разных λ



(a) субградиентный метод (b) градиентный метод (c) уск. градиентный метод

Рис. 13: Гарантированная точность по зазору двойственности в зависимости от времени при разных λ

Чем больше коэффициент регуляризации, тем быстрее методы достигают меньших значений зазора двойственности. Однако, для ускоренного метода не видно такой явной зависимости. Стоит при этом заметить, что при всех значениях коэффициента регуляризации λ, кроме $\lambda = 10$, ускоренный метод сошелся быстрее обычного градиентного метода.

4 Выводы.

В данной работе мы изучили субгрдиентный и градиентный методы. Рассмотрели, как они работают на разных данных, показали, что градиентный метод действительно делает примерно две внутренние итерации одномерного поиска и увидели, как влияют на работу субградиентного метода выбор начальной точки и длины начального шага.