Практическое задание 4 по курсу "Методы оптимизации в машинном обучении". Композитная оптимизация.

Выполнила студентка группы 191,
Косовская Арина

Содержание

1 Введение

Данное задание посвящено композитной оптимизации. Композитная функция определяется как

$$\phi(x) := f(x) + h(x),$$

где функция $f : \mathbb{R}^n \to \mathbb{R}$ — непрерывно дифференцируемая, а функция $h : \mathbb{R}^n \to \mathbb{R}$ — выпуклая (необязательно дифференцируемая) и достаточно простая, то есть для этой функции возможно эффективно вычислить проксимальное отображение.

Для решения оптимизационной задачи будут реализованы процедуры субградиентного метода, проксимального градиентного метода и адаптивный подбор шага по схеме Нестерова, проведены соответствующие эксперименты. В качестве критерия остановки будет использоваться критерий по зазору двойственности.

2 Эксперименты

2.1 Эксперимент 1. Выбор длины шага в субградиентном методе

. В данном эксперименте будет исследована работа субградиентного метода в зависимости от выбора константы α_0 в формуле длины шага: $\alpha_k = \dfrac{\alpha_0}{\sqrt{k+1}}, \alpha_0 > 0$.

Рассматривается $f(x) = 0.5 \cdot ||Ax - b||_2^2$. Тогда элементы матрицы $A \in \mathbb{R}^{200\times100}$ будем генерировать как случайные числа из равномерного распределения $\mathcal{U}(0; 1)$, $b = (0...0) \in \mathbb{R}^{200}$ — вектор, все элементы которого равны нулю. Для одной и той же задачи будем рассматривать различные начальные точки x_0. Будем исследовать зависимость расстояния начальной точки x_0 от точки минимума x^*. Рассмотрим несколько случаев: $||x_0 - x^*||_2^2 = 0.01$, $||x_0 - x^*||_2^2 = 0.1$, $||x_0 - x^*||_2^2 = 1$, $||x_0 - x^*||_2^2 = 5$, $||x_0 - x^*||_2^2 = 10$, $||x_0 - x^*||_2^2 = 15$.

Заметим, что так как A — положительно определенная матрица, $x^* = 0_{100}$. Тогда нужно сгенерировать такие x_0, что $||x_0||_2^2 = 0.01, ||x_0||_2^2 = 0.1, ||x_0||_2^2 = 1, ||x_0||_2^2 = 5, ||x_0||_2^2 = 10, ||x_0||_2^2 = 15$. Сделаем это следующим образом: сгенерируем стомерные векторы, беря элементы из равномерного распределения $\mathcal{U}(0; 1)$, отнормируем их, чтобы получить векторы длиной 1, и умножим на 0.01, 1, 5, 10 и 15 соответственно.

Наконец, описав, как мы будем генерировать данные, проведем эксперимент. Исследуем работу субградиентного метода в зависимости от выбора константы в формуле длины шага. Построим графики зависимости логарифма зазора двойственности от номера итерации для различных $||x_0||_2^2$ и различных α_0. Результаты приведены на рисунке 2.1.

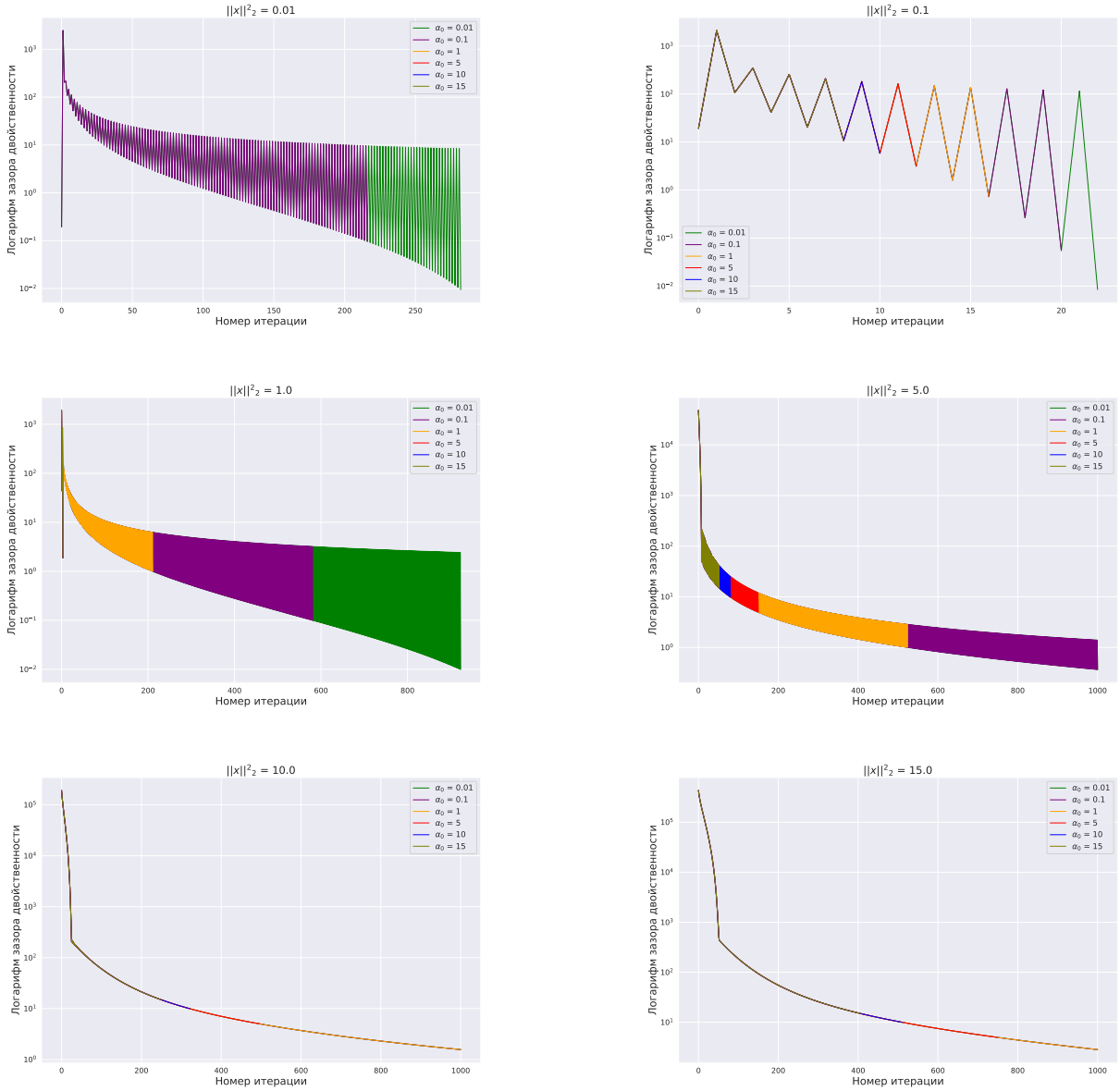

Рис. 2.1: Поведение субградиентного метода для различных значений α_0 в зависимости от начальной точки x_0 в логарифмической шкале.

Заметим, что наименьший зазор двойственности достигается при наименьшем α_0, но при этом требуется наибольшее число итераций до сходимости. Этого следовало ожидать, так как чем меньше константа, тем меньше и точнее можно делать шаг. В случае, если начальная точка находится близко к решению, и константа большая, то не получится сделать достаточно много шагов алгоритма. Также чем меньше значение α_0, тем менее стабилен метод в силу реализации метода, так

как шаг становится меньше. Чем дальше находится начальная точка от решения задачи, тем стабильнее, с меньшим разбросом работает метод.

Результаты при $\alpha_0 = 0.01$ и $\alpha_0 = 0.1$ получаются практически одинаковые: требуется значительно больше итераций до сходимости, чем при больших α_0, но зазор двойственности при $\alpha_0 = 0.01$ и 0.1 значительно меньше, чем при больших α_0, что следует из реализации метода. Исходя из этого, лучше использовать $\alpha_0 = 0.1$ или $\alpha_0 = 0.01$.

Говоря о связи между наилучшим коэффициентом α_0 и начальной точкой x_0, можно заметить, что наилучшее качество получается при $||x_0||_2^2 = 1$. При дальнейшем увеличении расстояния от стартовой точки до оптимальной увеличивается зазор двойственности. Из этого следует, что при использовании субградиентного метода для уменьшения зазора двойственности можно использовать мультистарт.

2.2 Эксперимент 2. Среднее число итераций линейного поиска в схеме Нестерова

.

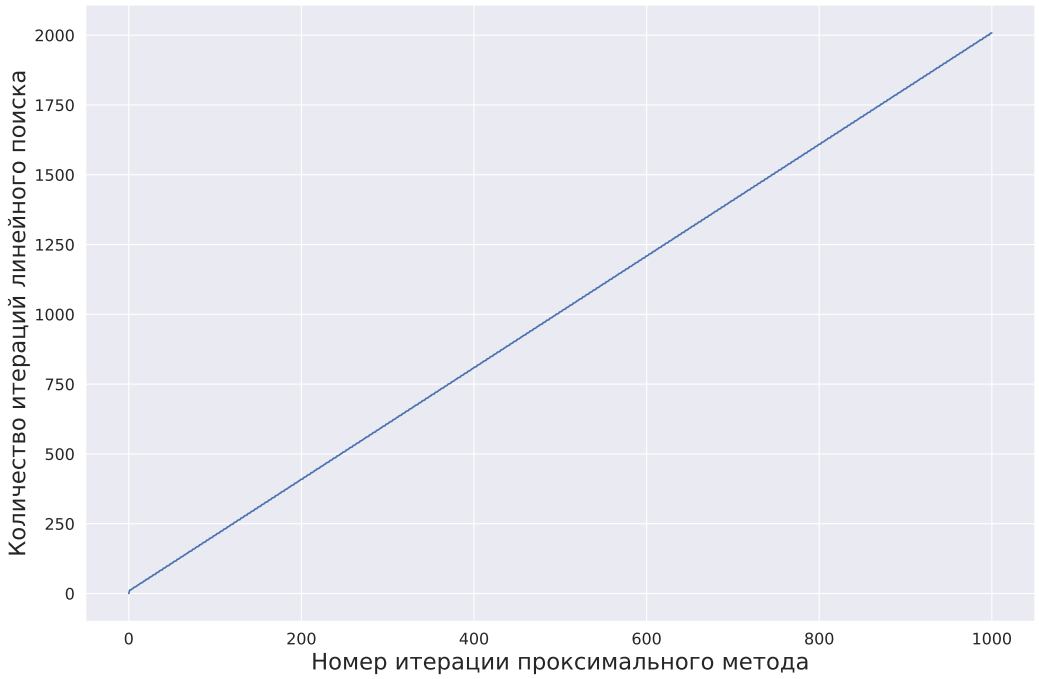

Рис. 2.2: Суммарное число итераций линейного поиска в схеме Нестерова.

В данном эксперименте будет исследовано суммарное число итераций линейного поиска в зависимости от номера итерации для проксимального градиентного

метода при подборе шага по схеме Нестерова.

Рассматривается функция $\phi(x) = f(x) + h(x)$, где функция $f(x) = 0.5 \cdot ||Ax - b||_2^2$ — непрерывно дифференцируемая, а функция $h = \lambda \cdot ||x||_1$ — выпуклая и недеффиринцируемая. Длина шага выбирается по схеме Нестерова, $\lambda = 1$, параметры выбраны по умолчанию.

Для эксперимента будет сгенерирована матрица A, элементы которой выбраны случайно из нормального распределения $\mathcal{N}(0; 1) \in \mathbb{R}^{n,m}$, $b \in \mathbb{R}^n$ — вектор, элементы которого выбраны случайно из нормального распределения, $x_0 \in \mathbb{R}^{>}$ — вектор, состоящий из нулей. Возьмем $n = m = 500$.

Результаты эксперимента представлены на рисунке 2.2.

На графике видна линейная зависимость количества итераций линейного поиска от номера итерации проксимального метода. Действительно, исходя из графика, среднее число итераций линейного поиска примерно равно двум.

2.3 Эксперимент 3. Сравнение методов

. В данном эксперименте требуется сравнить реализованные методы (субградиентный и проксимальный методы) и метод логбарьеров на примере задачи LASSO.

Данные будут генерироваться случайно. Элементы матрицы $A \in \mathbb{R}^{m \times n}$ генерируются как случайные из нормального распределения $\mathcal{N}(0; 1)$, матрицы $b \in \mathbb{R}^n$ — тоже. При фиксировании размера выборки будем использовать размер выборки $m = 500$, размерность пространства $n = 500$ и коэффициент регуляризации $\lambda = 0.1$. В качестве начальной точки x_0 будет браться точка $\frac{1_n}{2}$. В методе логбарьеров в качестве начальной точки рассматривается $\begin{pmatrix} \frac{1_n}{2} \\ 1_n \end{pmatrix}$ (для корректности метода, как было показано в предыдущем домашнем задании). Остальные параметры выбраны по умолчанию.

Сравним методы. Вначале рассмотрим, как меняются график гарантируемой точности по зазору двойственности против числа итераций и график гарантированной точности по зазору двойственности против реального времени работы от размерности пространства n. Для гарантированной точности по зазору двойственности будет использована логарифмическая шкала. Результаты изображены на рисунках 2.3 и 2.4.

Из данных графиков можно сделать вывод, что наименьший зазор двойственности достигается при использовании метода логбарьеров. При этом итерации

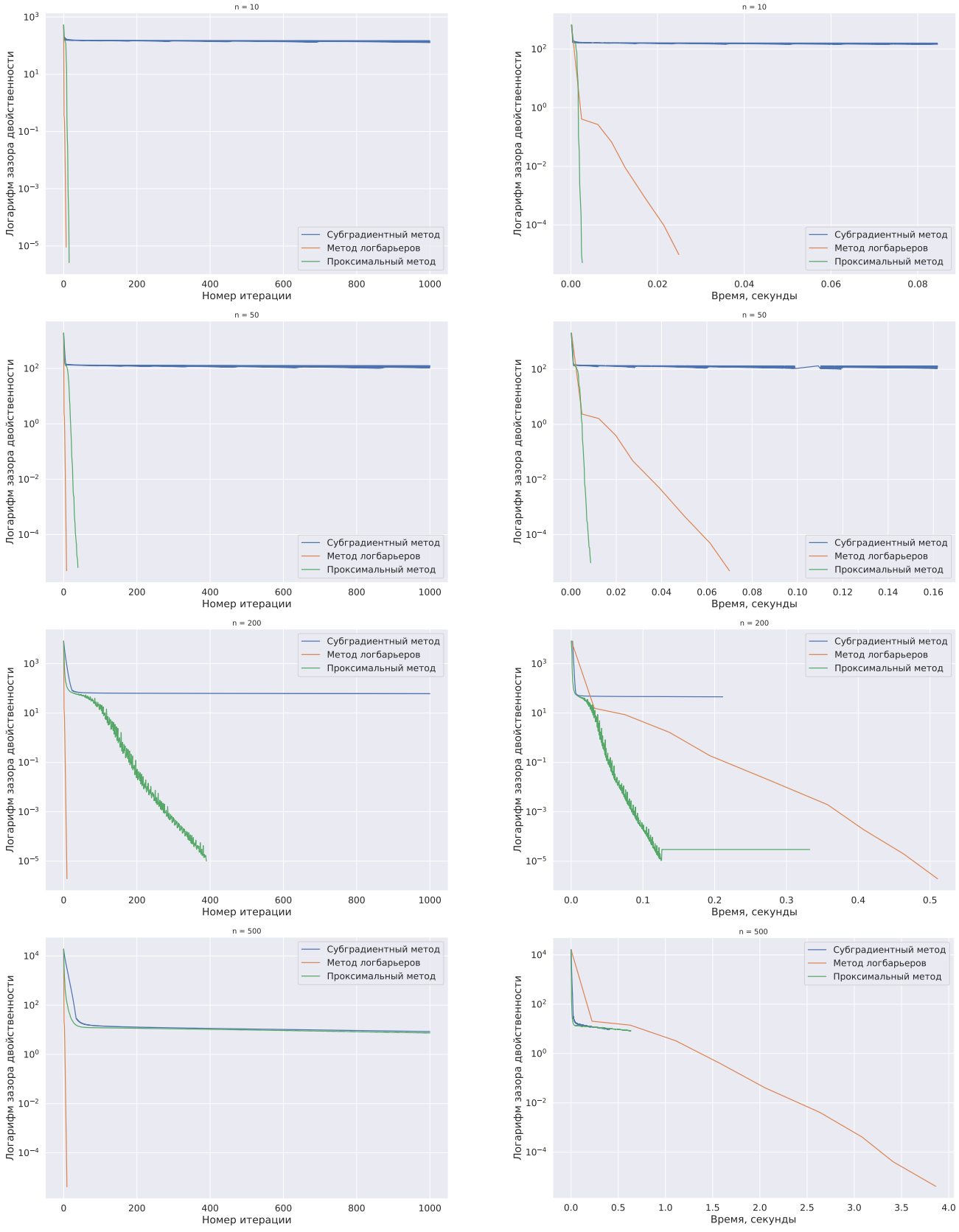

Рис. 2.3: Графики гарантируемой точности по зазору двойственности против числа итераций и и графики гарантированной точности по зазору двойственности против реального времени работы для $n = 10, 50, 200, 500$

данного метода требуют наибольшего количества времени, это связано со сложностью итерации метода Ньютона и используемой в нем стратегии подбора длины

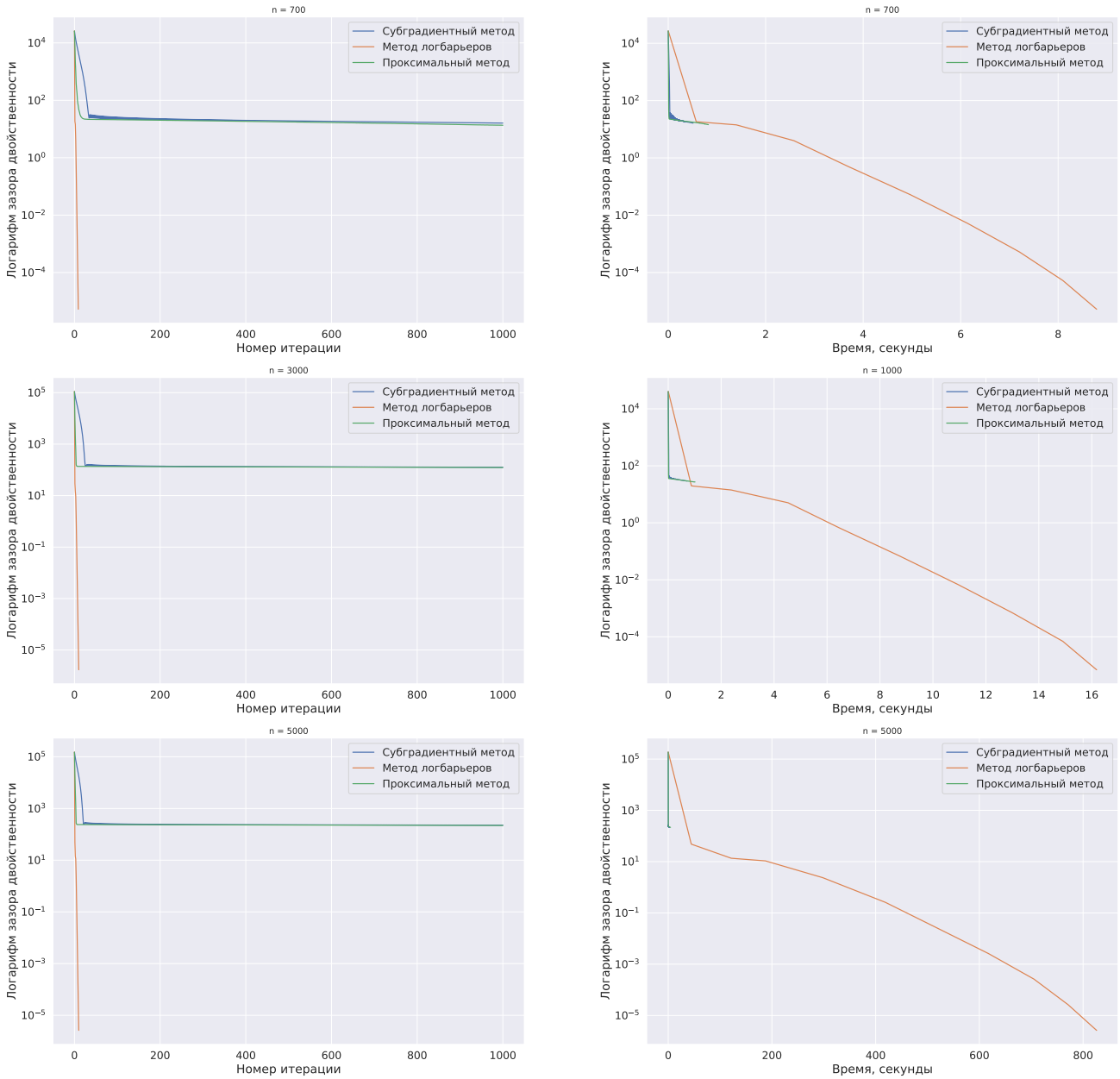

Рис. 2.4: Графики гарантируемой точности по зазору двойственности против числа итераций и графики гарантированной точности по зазору двойственности против реального времени работы для $n = 700, 1000, 5000$

шага. При этом методу требуется приблизительно одинаковое число итераций для различных значений n, что еще раз подтверждает, что при увеличении размерности пространства увеличивается время выполнения итерации метода барьеров, то есть время работы метода Ньютона.

Как обсуждалось на лекции, проксимальный и субградиентные методы медленно сходятся (у проксимального метода линейная скорость сходимости, у субградиентного метода — $O\left(\frac{1}{\sqrt{k}}\right)$, что подтверждается в экспериментах. Поэтому при небольших значениях n проксимальный метод сходится достаточно быстро, в отличии от субградиетного метода. При этом итерации проксимального метода

и метода сопряженных градиентов, в отличии от метода логбарьеров, быстрые. Несмотря на то, что методам требуется гораздо больше итераций до сходимости, время выполнения все равно значительно меньше.

Говоря о сравнении результатов для разных n, можно заметить, что при небольших размерностях пространства (до 50) логарифм зазора двойственности у проксимального метода и метода логбарьеров практически одинаковый. При этом проксимальный метод, как было сказано выше, работает быстрее, а субградиентный сходится значительно медленнее всех методов. При увеличении размерности пространства проксимальный метод начинает сходиться медленнее, ухудшается логарифм зазора двойственности, в особенности в сравнении с методом барьеров. При больших значениях размерности пространства логарифм зазора двойственности для проксимального и субградиентного методов становится значительно больше, чем у метода логбарьеров, но при этом количество секунд до сходимости сильно меньше. Из этого можно сделать вывод, что если не требуется решать задачу очень точно, то можно использовать проксимальный метод для экономии времени, иначе стоит использовать метод логбарьеров.

Теперь рассмотрим графики гарантируемой точности по зазору двойственности против числа итераций и графики гарантированной точности по зазору двойственности против реального времени работы для различных значений m. Они представлены на рисунках 2.5, 2.6.

Даже при небольших размерах выборки метод логбарьеров справляется значительно лучше, если опираться на логарифм зазора двойственности. При небольших значениях m не наблюдается явной зависимости как времени работы метода логбарьеров от размера выборки, так и числа итераций до сходимости. Объяснить это можно тем, что стоимость вычисления матриц Axb, $A^T(Axb)$ составляет $O(m \cdot n)$, а метод Ньютона "стоит"$O(n^3)$. Поэтому при увеличении n время выполнения одной итерации метода барьеров сильно увеличивается, а при увеличении m — нет (только при очень больших значениях m будет более явно наблюдаться линейная зависимость).

Сравнивая метод логбарьеров с реализованными методами в этом задании, можно заметить, что число итераций до сходимости у него меньше, но время работы значительно больше. При увеличении m проксимальный метод начинает работать чуть лучше (и оказывается как наиболее быстрым при $m = 5000$, так и точным: логарифм зазора двойственности у него и у метода логбарьеров практически одинаковые). Субградиентный метод при всех размерах выборки сходится медленнее и менее точно.

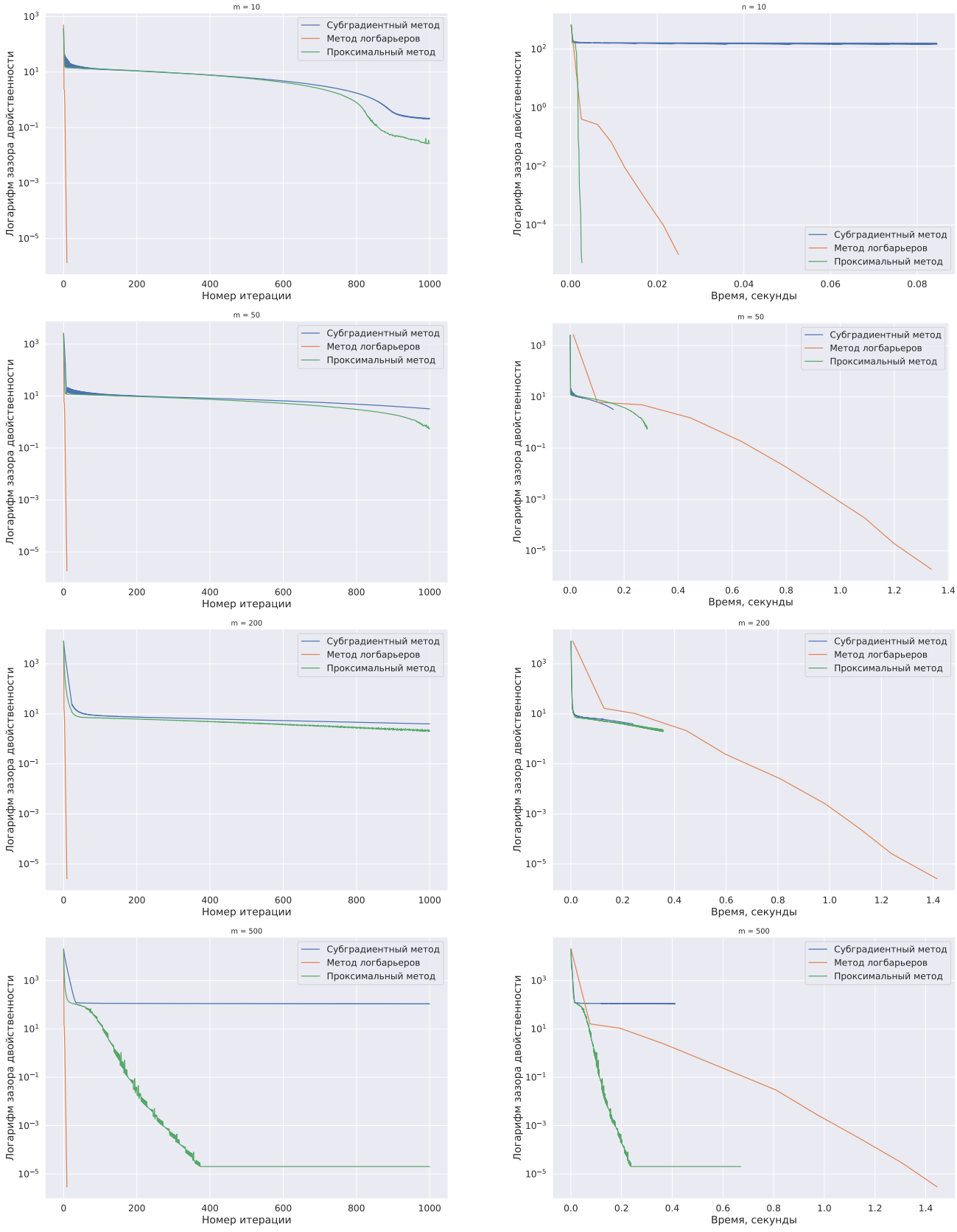

Рис. 2.5: Графики гарантируемой точности по зазору двойственности против числа итераций и и графики гарантированной точности по зазору двойственности против реального времени работы для $m = 10, 50, 200, 500$

Наконец, рассмотрим зависимость от коэффициента регуляризации. Сравнение представлено на рисунках 2.7, 2.8.

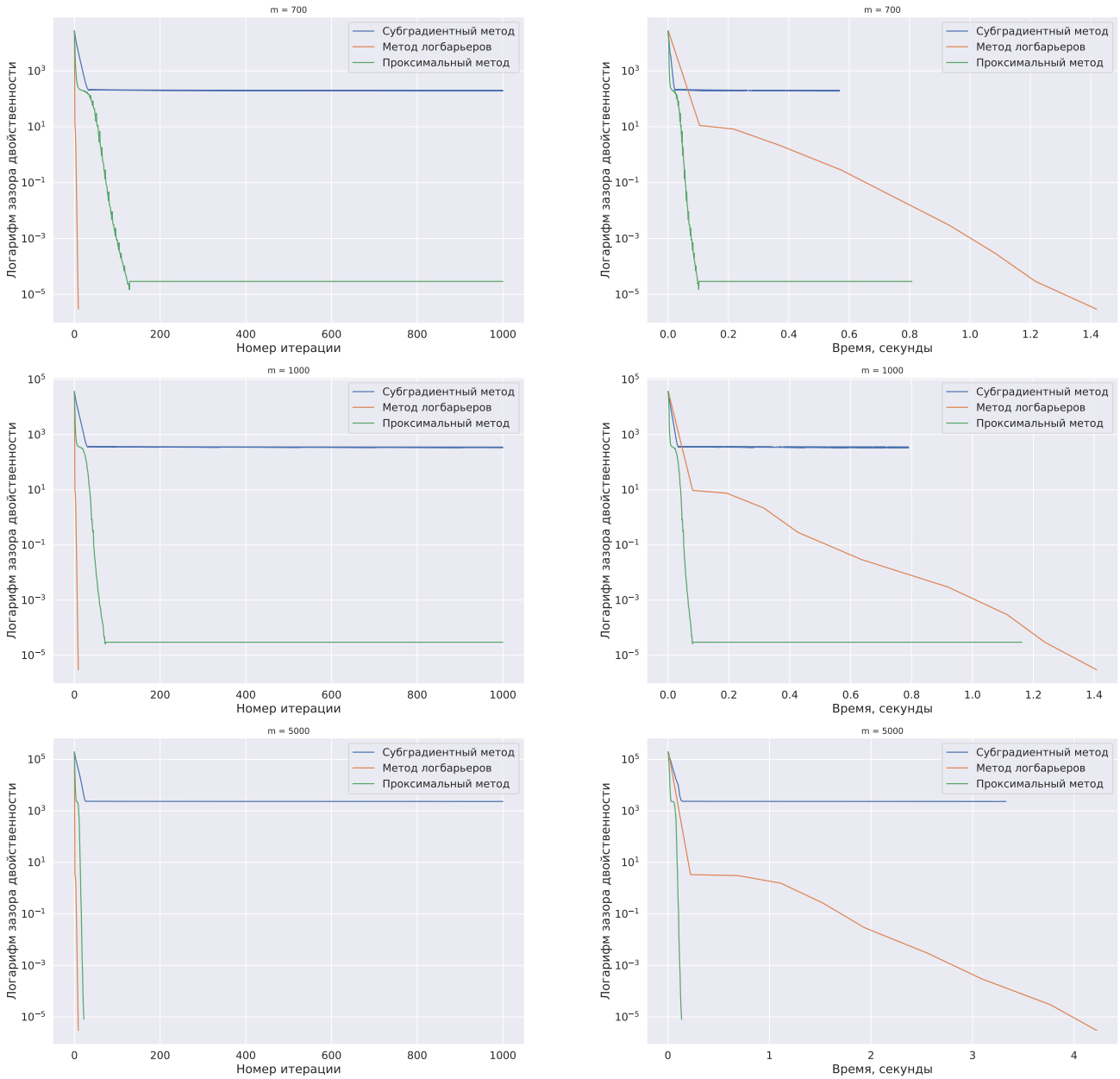

Рис. 2.6: Графики гарантируемой точности по зазору двойственности против числа итераций и и графики гарантированной точности по зазору двойственности против реального времени работы для $m = 700, 1000, 5000$

Как мы знаем с лекций, в решении задачи LASSO часть признаков окажется равна нулю: нулевые веса отвечают исключению соответствующего признака из прогнозирующей модели. Таким образом, чем больше коэффициент регуляризации, тем лучше сходимость субградиентного и проксимального методов. Тем не менее, метод логарифмических барьеров все еще справляется лучше, хоть и медленнее.

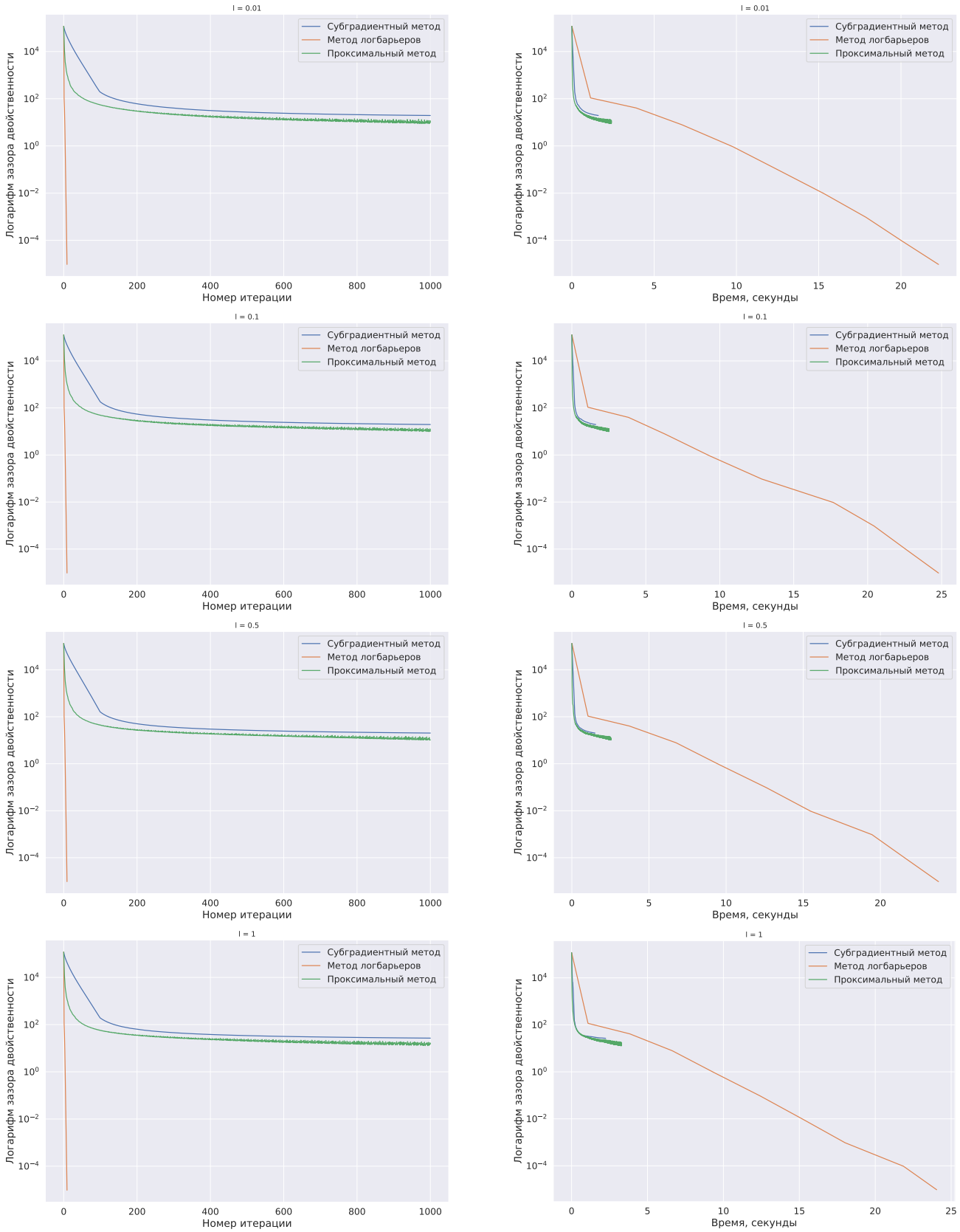

Рис. 2.7: Графики гарантируемой точности по зазору двойственности против числа итераций и и графики гарантированной точности по зазору двойственности против реального времени работы для $\lambda = 0.01, 0.1, 0.5, 1$

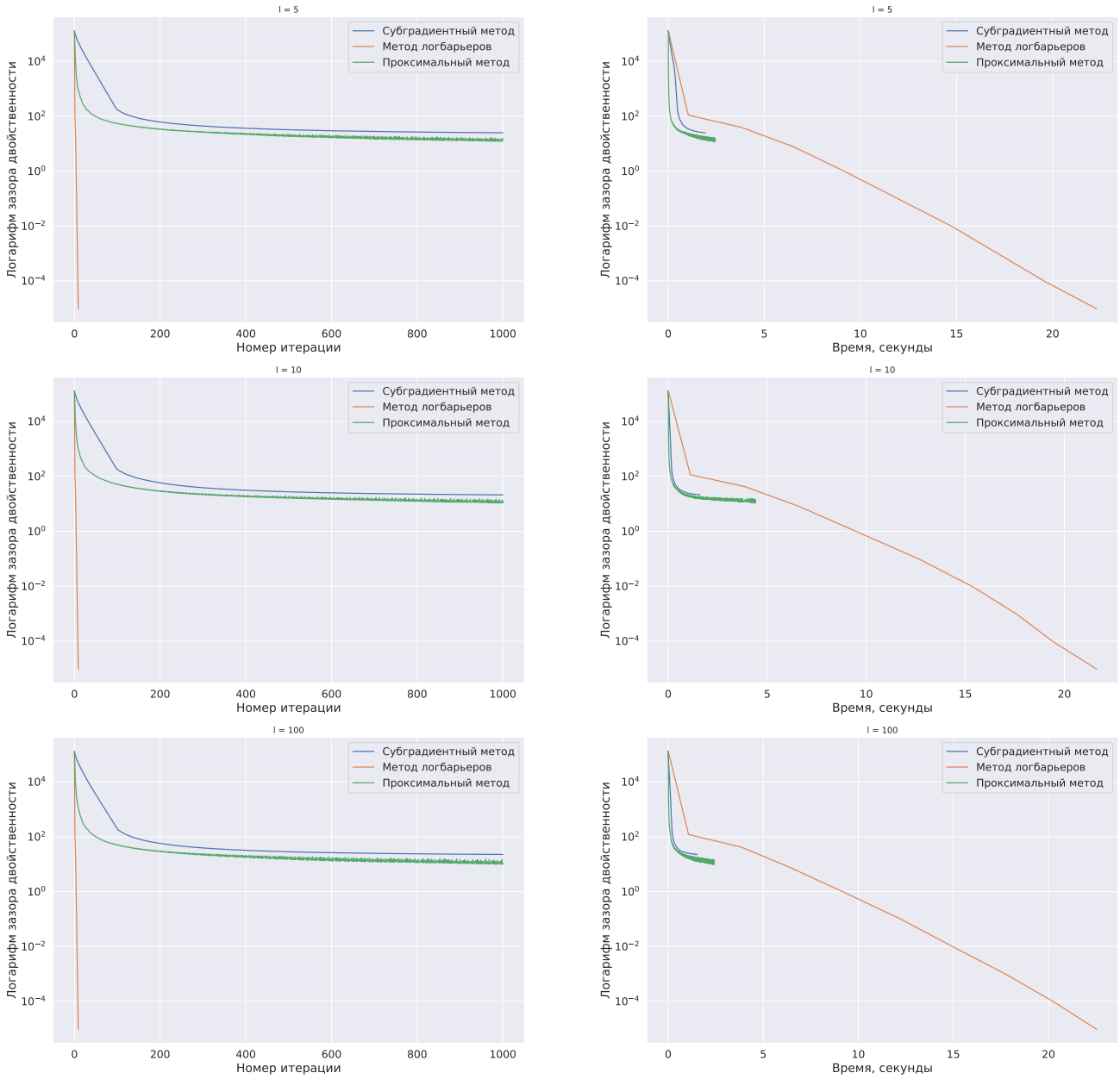

Рис. 2.8: Графики гарантируемой точности по зазору двойственности против числа итераций и и графики гарантированной точности по зазору двойственности против реального времени работы для $\lambda = 5, 10, 100$