Практическое задание 3 по курсу "Методы оптимизации в машинном обучении".
Метод барьеров.

Выполнила студентка группы 191,
Косовская Арина

Содержание

1 Введение

Данное задание посвящено методу барьеров — методу решения условных задач оптимизации с ограничениями вида неравенств:

$$\min_x f(x) \text{ s.t. } g_i(x) \leq 0,\ i = 1, ..., m,$$

где $f, g_1, ..., g_m \in \mathbb{R}^n \to \mathbb{R}$ — дважды непрерывно дифференцируемые функции, $x \in \mathbb{R}^n$ — целевая переменная.

Для решения вспомогательной задачи $f_t(x) := t \cdot f(x) + F(x)$, где $\forall\ t \geq 0$ $f_t : \text{int } Q \to \mathbb{R}$, $Q := \{x \in \mathbb{R}^n : g_i(x) \leq \forall i = 1, ..., m\}$, $F :\ \text{int } Q \to \mathbb{R}$ в данном задании будем пользоваться метод Ньютона. В качестве критерия остановки используется $||\nabla f_{t_k}(y_l)||_2^2 \leq \varepsilon_{\text{inner}} ||\nabla f_{t_k}(x_k)||_2^2$, где $\varepsilon_{\text{inner}} > 0$ — некоторый параметр. Также будет рассматриваться логарифмический барьер $F(x) := -\sum\limits_{i=1}^{m} \ln(-g_i(x))$.

В схеме отслеживания пути будет использоваться линейное увеличение параметров t_k: $t_{k+1} = \gamma t_k$, где $\gamma > 1$ — постоянная константа.

Для решения оптимизационной задачи $\phi(x) := \frac{1}{2} \sum\limits_{i=1}^{m} \left(\langle a_i, x \rangle - b_i\right)^2 +$
$+ \lambda \sum_{j=1}^{n} |x_j| := \frac{1}{2}\|Ax - b\|_2^2 + \lambda \|x\|_1 \to \min\limits_{x \in \mathbb{R}^n}$ будет использоваться критерий остановки $\eta(x, \mu(x)) < \varepsilon$, где $\eta(x, \mu(x))$ — зазор двойственности.

В данном задании будет реализована процедура вычисления зазора двойственности и реализован метод барьеров. Для решения гладкой условной задачи $\min\limits_{x,u} \left\{\frac{1}{2}\|Ax - b\|_2^2 + \lambda \langle 1_n, u \rangle\right\} \quad \text{s.t.} \quad x \preceq u, \quad x \succeq -u$ найдена вспомогательная функция, выписана система линейных уравнений, задающая ньютоновское направление d_k, а также предложен способ ее решения. Для точки (x, u) и направления d_k определена максимально допустимая длина шага α. Проведены соответствующие эксперименты.

2 Теоретическая часть

Рассмотрим гладкую условную задачу
$\min\limits_{x,u} \left\{\frac{1}{2}\|Ax - b\|_2^2 + \lambda \langle 1_n, u \rangle\right\} \quad \text{s.t.} \quad x \preceq u, \quad x \succeq -u$, где $x, u \in \mathbb{R}^n$, $1_n = (1, ..., 1) \in \mathbb{R}^n$. Будем решать ее методом барьеров.

Перепишем ограничения. Из условия $\begin{cases} x - u \preceq 0 \\ x + u \preceq 0 \end{cases} \Rightarrow \begin{cases} x - u \preceq 0 \\ -x - u \preceq 0 \end{cases}$

Пусть $g_1(x,u) := x - u$, $g_2(x,u) = -x - u$. Также определим $g_{1i} = x_i - u_i$, $g_{2i} = -x_i - u_i$.

Тогда множество, задаваемое функциональными ограничениями:
$Q := \{x, u \in \mathbb{R}^n : g_{1i} \leq 0,\ g_{2i} \leq 0\ \forall i = 1, ..., n\}$.

2.1 Вспомогательная функция

Пусть $F : \text{int } Q \to \mathbb{R}$ — логарифмическая барьерная функция, тогда $\forall t$ определим вспомогательную функцию $f_t : \text{int } Q \to \mathbb{R}$ как $t \cdot f(x) + F(x)$.

$$F(x,u) = -\sum_{i=1}^{n} \left(\ln\left(-g_{1i}(x,u)\right) + \ln\left(-g_{2i}(x,u)\right)\right) = -\sum_{i=1}^{n} \ln\left(u_i - x_i\right) - \sum_{i=1}^{n} \ln\left(x_i + u_i\right) =$$

$$= -\langle 1_n, \ln(u - x)\rangle - \langle 1_n, \ln(x + u)\rangle.$$

$$f_t(x,u) = t \cdot \left(\frac{1}{2}\|Ax - b\|_2^2 + \lambda \left\langle 1_n, u\right\rangle\right) - \langle 1_n, \ln(u - x)\rangle - \langle 1_n, \ln(x + u)\rangle.$$

Перейдем к решению вспомогательной задачи. В качестве алгоритма минимазации будем использовать метод Ньютона.

2.2 Система линейных уравнений, задающая ньютоновское направление

Пусть $d_k = (d_k^x, d_k^u)$ — Ньютоновское направление. Вспомним, что оно задается с помощью уравнения $\nabla^2 f_t(x_k, u_k) d_k = -\nabla f_t(x_k, u_k)$.

Далее будем обозначать f'_{tx} — частную производную f_t по аргументу x, f''_{txx} — частную производную второго порядка от f'_{tx} по аргументу x.

$$\nabla f_t = \begin{pmatrix} f'_{tx} \\ f'_{tu} \end{pmatrix} = \begin{pmatrix} tA^T(Ax - b) + \frac{1}{u-x} - \frac{1}{x+u} \\ t\lambda 1_n - \frac{1}{u-x} + \frac{1}{x+u} \end{pmatrix}$$

$$\nabla^2 f_t = \begin{pmatrix} f''_{txx} & f''_{txu} \\ f''_{tux} & f''_{tuu} \end{pmatrix} =$$
$$= \begin{pmatrix} tA^TA + \operatorname{diag}\left(\frac{1}{u-x} \odot \frac{1}{u-x} + \frac{1}{x+u} \odot \frac{1}{x+u}\right) & \operatorname{diag}\left(-\frac{1}{u-x} \odot \frac{1}{u-x} + \frac{1}{x+u} \odot \frac{1}{x+u}\right) \\ \operatorname{diag}\left(-\frac{1}{u-x} \odot \frac{1}{u-x} + \frac{1}{x+u} \odot \frac{1}{x+u}\right) & \operatorname{diag}\left(\frac{1}{u-x} \odot \frac{1}{u-x} + \frac{1}{x+u} \odot \frac{1}{x+u}\right) \end{pmatrix}$$

Тогда получается следующая система уравнений:

$$\begin{pmatrix} tA^TA + \operatorname{diag}\left(\frac{1}{u-x}\odot\frac{1}{u-x}+\frac{1}{x+u}\odot\frac{1}{x+u}\right) & \operatorname{diag}\left(-\frac{1}{u-x}\odot\frac{1}{u-x}+\frac{1}{x+u}\odot\frac{1}{x+u}\right) \\ \operatorname{diag}\left(-\frac{1}{u-x}\odot\frac{1}{u-x}+\frac{1}{x+u}\odot\frac{1}{x+u}\right) & \operatorname{diag}\left(\frac{1}{u-x}\odot\frac{1}{u-x}+\frac{1}{x+u}\odot\frac{1}{x+u}\right) \end{pmatrix} \begin{pmatrix} d^x \\ d^u \end{pmatrix} =$$

$$= \begin{pmatrix} tA^T(Ax-b) + \frac{1}{u-x} - \frac{1}{x+u} \\ t\lambda 1_n - \frac{1}{u-x} + \frac{1}{x+u} \end{pmatrix}$$

$$\begin{cases} \left[tA^TA + \operatorname{diag}\left(\frac{1}{u-x}\odot\frac{1}{u-x}+\frac{1}{x+u}\odot\frac{1}{x+u}\right)\right] d^x + \left[\operatorname{diag}\left(-\frac{1}{u-x}\odot\frac{1}{u-x}+\frac{1}{x+u}\odot\frac{1}{x+u}\right)\right] d^u = \\ = -tA^T(Ax-b) - \frac{1}{u-x} + \frac{1}{x+u} \\ \left[\operatorname{diag}\left(-\frac{1}{u-x}\odot\frac{1}{u-x}+\frac{1}{x+u}\odot\frac{1}{x+u}\right)\right] d^x + \left[\operatorname{diag}\left(\frac{1}{u-x}\odot\frac{1}{u-x}+\frac{1}{x+u}\odot\frac{1}{x+u}\right)\right] d^u = \\ = -t\lambda 1_n + \frac{1}{u-x} - \frac{1}{x+u} \end{cases}$$

2.3 Решение системы

Для решения системы воспользуемся разложением Холецкого. Как обсуждалось на лекции, это устойчивая и быстрая операция. По сравнению с решением системы уравнений в лоб, время, требуемое на выполенние задачи, уменьшится. Тем не менее, все еще приходится вычислять A^TA, что, как известно, затратно по времени.

Для уменьшения затрат по памяти можно выразить d^u из первого уравнения и подставить во второе. Тогда можно решить второе уравнение вновь с помощью разложения Холецкого, а после подставить найденное d^x в первое уравнение и найти d^u. Размерность уменьшится вдвое.

2.4 Максимально допустимая длина шага

Заметим, что функции g_1, g_2, соответствующие условиям задачи, являются

афинными. Действительно, $g_1 = x - u = \langle \begin{pmatrix} 1_n \\ -1_n \end{pmatrix}, x\rangle - 0_n^T,\ g_2 = -x - u =$

$= \langle \begin{pmatrix} -1_n \\ -1_n \end{pmatrix}, x\rangle - 0_n^T$. То есть $q_1 = \begin{pmatrix} 1_n \\ -1_n \end{pmatrix}, s_1 = s_2 = 0_n^T, q_2 = \begin{pmatrix} -1_n \\ -1_n \end{pmatrix}$.

Тогда, согласно теории, $\alpha_l^{\max} = \min\limits_{i \in I_l} -\dfrac{\langle q_i, y_l\rangle}{\langle q_i, d_l\rangle}$.

Для подбора длины шага используется процедура бэктрекинга, а для решения вспомогательной задачи используется метод Ньютона.

Пусть $I_l = \{\langle q_1, d_l\rangle > 0, \langle q_2, d_l\rangle > 0\}$. Так длина шага должна быть положительной и не превосходить 1, учитывая условие выше, получим:

$$\alpha_l^{\max} = \min\left\{1, \theta \min_{i=1\dots n, d^x_{li}-d^u_{li}>0} \frac{u_{li}-x_{li}}{d^x_{li}-d^u_{li}}, \theta \min_{i=1\dots n, -d^x_{li}-d^u_{li}>0} \frac{u_{li}+x_{li}}{-d^x_{li}-d^u_{li}}\right\}$$

В качестве θ выберем 0.99.

2.5 Выбор начальной точки

Так как начальная точка должна лежать внутри множества Q, выберем точку $\begin{pmatrix} \frac{1_n}{2} \\ 1_n \end{pmatrix}$.

3 Эксперименты

3.1 Эксперимент а

. Исследуем чувствительность реализованного метода барьеров для задачи LASSO к выбору параметров γ (скорость увеличения t_k) и ϵ_{inner} (точность для метода Ньютона).

Первым рассмотрим эксперимент, когда исследовалась чувствительность метода к выбору γ. Остальные параметры были выбраны по умолчанию. Для этого был взят набор данных w8a, имеющий размерность (49749, 300). Результаты эксперимента изображены на рисунке 3.1.

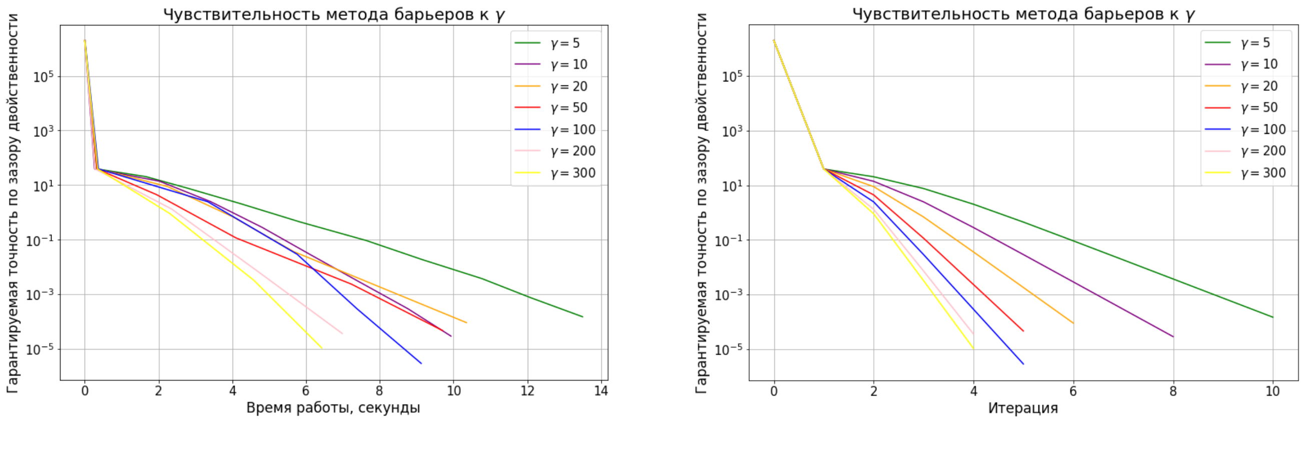

Рис. 3.1: Чувствительность метода барьеров к γ.

Заметим, что чем больше γ, тем меньше время работы и тем меньше зазор двойственности. Для бОльших γ требуется меньшее число итераций для получения высокой гарантированной точности. Тем не менее, все $\gamma \in [10; 300]$ работают за приблизительно одинаковые время и количество итераций, а для $\gamma = 5$ метод работает хуже. Можно сделать предположение, что γ должна выбираться из этого промежутка в зависимости от других параметров.

Вывод: увеличение γ приводит к уменьшению времени работы и уменьшению числа итераций до сходимости.

Теперь проведем эксперимент, в котором исследуется чувствительность метода барьеров к значению ϵ_{inner}. Вновь остальные параметры были выбраны по умолчанию, использовался набор данных w8a. Результаты представлены на рисунке 3.2.

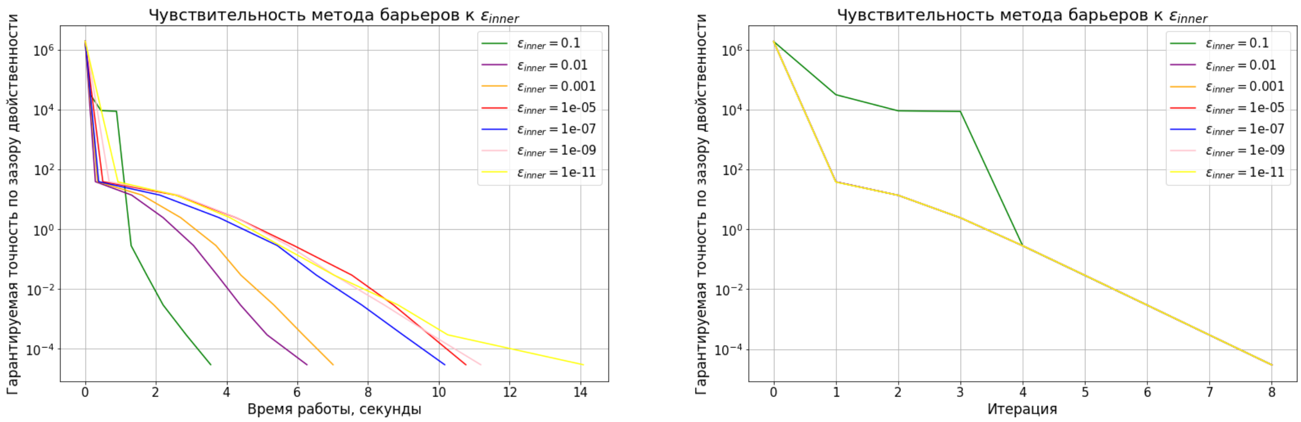

Рис. 3.2: Чувствительность метода барьеров к ϵ_{inner}

Видим, что значение ϵ_{inner} не влияет на количество итераций до сходимости, за исключением наибольшего $\epsilon_{\text{inner}} = 0.1$. При этом же значении параметра достигается наименьшее время работы метода. В целом, чем больше ϵ_{inner}, тем меньше времени требовалось для работы метода барьеров. То есть чем меньше значение ϵ_{inner}, тем дольше совершается итерация (что обусловлено дольшей работой метода метода Ньютона). Также заметим, что выбор ϵ_{inner} не влияет на гарантированную точность по зазору двойственности.

Вывод: для высокой точности и быстрой работы метода достаточно выбрать $\epsilon_{\text{inner}} = 0.1$. При дальнейшем уменьшении ϵ_{inner} на каждую итерацию будет требоваться больше времени, гарантированная точность при этом не меняется.

3.2 Эксперимент б

Перейдем к исследованию поведения метода для различных значений размерности пространства n, размера выборки m и коэффициента регуляризации λ.

Первым проводем эксперимент с размерностью пространства. Для этого сгенерируем случайные A и b так, чтобы размерность пространства была n, m выбралось равным 100. Остальные параметры выбраны по умолчанию. Результаты представлены на рисунке 3.3.

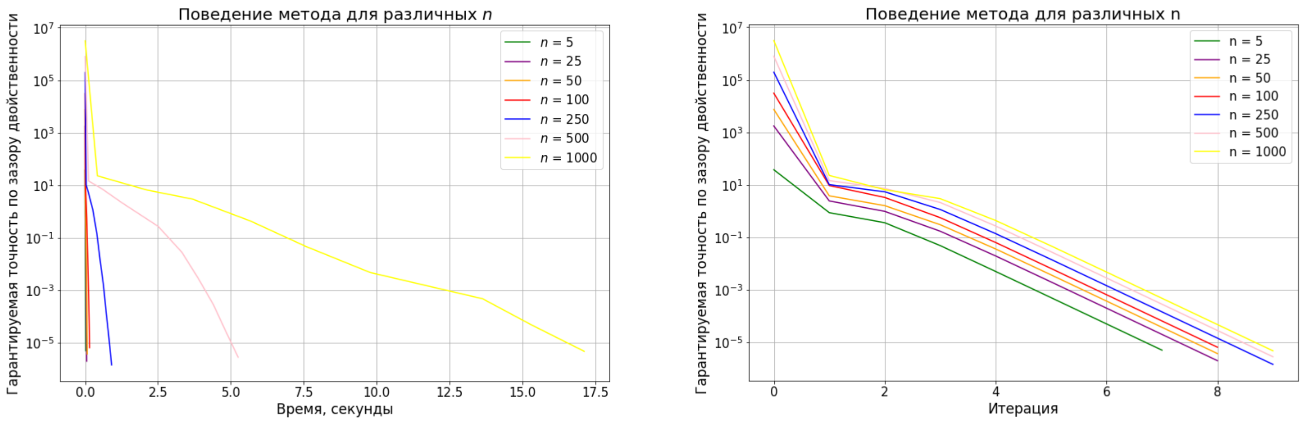

Рис. 3.3: Поведение метода для различных значений размерности пространства n.

При небольших значениях n (не превосходящих 100) метод работает одинаково быстро. При дальнейшем увеличении размерности пространства метод начинает работать дольше, что объяснимо сложностью итерации метода Ньютона и используемой в нем стратегии подбора длины 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.

Вывод: с увеличением размерности пространства увеличивается время выполнения одной итерации метода барьеров.

Теперь рассмотрим поведение метода при изменении размера выборки m. Вновь были сгенерированы случайные A и b соответствующих размеров, остальные параметры были выбраны по умолчанию. $n = 100$. Результаты представлены на рисунке 3.4.

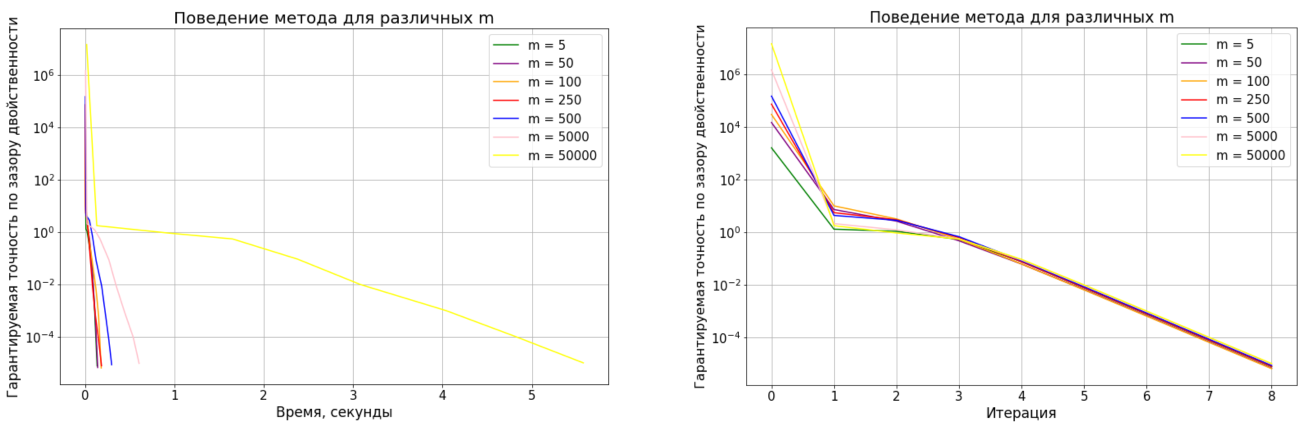

Рис. 3.4: Поведение метода для различных значений размера выборки m.

Число итераций от сходимости не зависит от размера выборки. Явной зависимости времени работы метода от размера выборки при $m \leq 5000$ также не наблюдается. При $m = 50000$ время работы увеличивается до 6 секунд. Объяснить это можно тем, что стоимость вычисления матриц $Ax - b$, $A^T(Ax - b)$ составляет $O(m * n)$, 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).

Вывод: при размерах выборки, меньших 5000, время работы метода одинаковое. При дальнейшем увеличении m время работы метода будет расти значительнее.

Наконец, исследуем поведение метода в зависимости от коэффициента регуляризации. В качестве данных использовался набор w8a. Остальные значения выбраны по умолчанию. Результаты представлены на рисунке 3.5.

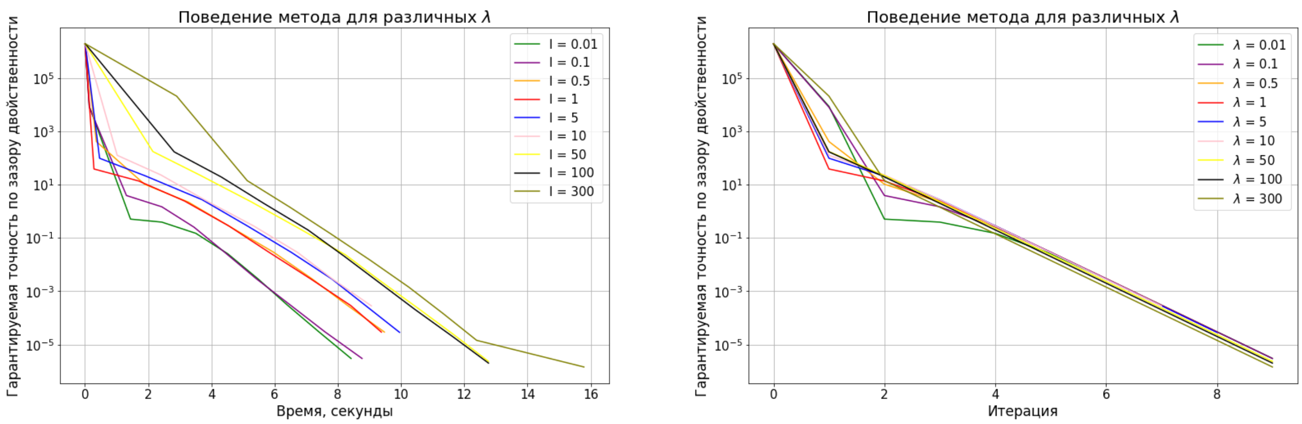

Рис. 3.5: Поведение метода для различных значений коэффициента регуляризации λ.

Вывод: значение коэффициента регуляризации не влияет на число итераций. При увеличении λ немного увеличивается время работы метода.