Практическое задание 3

Рыбка Елизавета, 474

21 ноября 2017 г.

1 Теория

$$\min_{x,u}\left\{\frac{1}{2}\|Ax-b\|_2^2+\lambda\langle 1_n,u\rangle\right\} \qquad s.t. \quad x\preceq u, \quad x\succeq -u$$

Вспомогательная функция: $f_t(x,u)=tf(x,u)+F(x,u)$.

Перепишем ограничения:

$$g_1: x-u\preceq 0, \quad g_2: -x-u\preceq 0$$
$$g_{1i}: x_i-u_i\le 0, \quad g_{2i}: -x_i-u_i\le 0$$

$$F(x,u)=-\sum_{i=1}^{m}[\ln(-g_{1i}(x,u))+\ln(-g_{2i}(x,u))]=-\sum_{i=1}^{m}[\ln(u_i-x_i)+\ln(x_i+u_i)]$$
$$f_t(x,u)=t(\frac{1}{2}\|Ax-b\|_2^2+\lambda\langle 1_n,u\rangle)-\langle 1_n,\ln(u-x)\rangle-\langle 1_n,\ln(x+u)\rangle \tag{1}$$

Ньютоновское на направление: $d_k=(d_k^x,d_k^u)$
Направление для шага в методе Ньютона задается уравнением: $\nabla^2 f_t(x_k)d_k=-\nabla f_t(x_k)$

$$\nabla f_t=\begin{pmatrix}f'_{t(x)}\\ f'_{t(u)}\end{pmatrix}=\begin{pmatrix}tA^T(Ax-b)-[\frac{-1}{u-x}+\frac{1}{u+x}]\\ t\lambda 1_n-[\frac{1}{u-x}+\frac{1}{u+x}]\end{pmatrix}=\begin{pmatrix}tA^T(Ax-b)-[-c_1+c_2]\\ t\lambda 1_n-[c_1+c_2]\end{pmatrix}$$

Были введены обозначения: $c_1=\frac{1}{u-x}$, $c_2=\frac{1}{u+x}$. Деление покомпонентное.
$c_{1(x)}=\frac{\partial}{\partial x}\frac{1}{u-x}=\wr$покомпонентно$\wr=\frac{\partial}{\partial x_i}\frac{1}{u_i-x_i}=-\frac{-1}{(u_i-x_i)^2}=\wr$ покомпонентно $\wr=\frac{1}{u-x}\odot\frac{1}{u-x}=c_1\odot c_1$
$\odot$ — покомпонентное умножение.

$$\nabla^2 f_t=\begin{pmatrix}f''_{t(xx)} & f''_{t(xu)}\\ f''_{t(xu)} & f''_{t(uu)}\end{pmatrix}=\begin{pmatrix}tA^TA+diag(c_1\odot c_1+c_2\odot c_2) & diag(-c_1\odot c_1+c_2\odot c_2)\\ diag(-c_1\odot c_1+c_2\odot c_2) & diag(c_1\odot c_1+c_2\odot c_2)\end{pmatrix}$$

$$\nabla^2 f_t d=-\nabla f_t$$

$$\begin{pmatrix}tA^TA+diag(c_1\odot c_1+c_2\odot c_2) & diag(-c_1\odot c_1+c_2\odot c_2)\\ diag(-c_1\odot c_1+c_2\odot c_2) & diag(c_1\odot c_1+c_2\odot c_2)\end{pmatrix}\begin{pmatrix}d_x\\ d_u\end{pmatrix}=-\begin{pmatrix}tA^T(Ax-b)-[-c_1+c_2]\\ t\lambda 1_n-[c_1+c_2]\end{pmatrix}$$

Из нижних блоков:

$$\begin{aligned} diag(-c_1 \odot c_1 + c_2 \odot c_2)d_x + diag(c_1 \odot c_1 + c_2 \odot c_2)d_u &= -t\lambda 1_n + c_1 + c_2 \\ (-c_{1i}^2 + c_{2i}^2)d_{xi} + (c_{1i}^2 + c_{2i}^2)d_{ui} &= -t\lambda + c_{1i} + c_{2i} \\ \frac{-t\lambda 1_n + c_1 + c_2}{c_1 \odot c_1 + c_2 \odot c_2} + \frac{c_1 \odot c_1 - c_2 \odot c_2}{c_1 \odot c_1 + c_2 \odot c_2} \odot d_x &= d_u \end{aligned} \tag{2}$$

Подставляем в верхние:

$$\left(tA^TA + 4diag\left(\frac{c_1 \odot c_1 \odot c_2 \odot c_2}{c_1 \odot c_1 + c_2 \odot c_2}\right)\right) d_x = -tA^T(Ax-b) - c_1 + c_2 - (-t\lambda 1_n + c_1 + c_2) \odot \left(\frac{-c_1 \odot c_1 + c_2 \odot c_2}{c_1 \odot c_1 + c_2 \odot c_2}\right) \tag{3}$$

В работе используется нижеописанный метод решения. Уравнение (3) решается с помощью разложения Холецкого (возможно т.к. в левой части уравнения положительно определенная матрица), находим d_x. С помощью выражения (2) находим d_u

Плюсы и минусы

+ Метод Холецкого экономит время по сравнения с, например, методом Гаусса, т.е. решением исходной СЛАУ "в лоб"
+ В исходной СЛАУ ($\nabla^2 f_t d = -\nabla f_t$) матрица была размера $2n \times 2n$, в выденной же (2), (3) имеем дело с матрицей $n \times n$. Это, в свою очередь, приводит к экономии по памяти и по времени.
– Все так же присутствует необходимость вычислять A^TA (но это вообще никуда не деться). В случае плотных матриц необходимо: $O(n^2m)$
– ???

Допустимая длина шага

Во первых, заметим что у нашей задачи ограничения аффинные: $q_1 = \binom{1_n}{-1_n}, s_1 = 0, q_1 = \binom{-1_n}{-1_n}, s_2 = 0$

$$\alpha_l^{max} = \min_{i \in I_l} \frac{-\langle q_i, y_l \rangle}{\langle q_i, d_l \rangle}, \quad I_l = \{1 \le i \le 2 : \langle q_i, y_l \rangle > 0\}, \quad y_l = \begin{pmatrix} x_l \\ u_l \end{pmatrix}$$

$$\alpha_l^+ = \min \frac{u_l - x_l}{d_{xl} - d_{ul}}, \quad \alpha_l^- = \min \frac{u_l + x_l}{-d_{xl} - d_{ul}}$$

$$\alpha_l^{max} = \min\{1, \theta\alpha_l^+, \theta\alpha_l^-\}$$

Начальная точка

На начальную точку $\binom{x_0}{u_0}$ накладывается одно ограничение она должна лежать в допустимом множестве: $\{x \preceq u, \quad x \succeq -u\}$. Разумно взять точку не слишком близко к границе. Для удобства и простоты будем использовать начальную точку: $\binom{x_0}{u_0} = \binom{\frac{1}{2}1_n}{1_n}$

2 Практика: Эксперимент.

В серии экспериментов исследовалась чувствительность метода барьеров для задачи LASSO от различных параметров: $\gamma, \epsilon_{inner}, n, m, \lambda$. Для этого использовалась функция *barriermethodlasso* из модуля *optimization*. Неисследуемые параметры были выставлены на значения по умолчанию предложенные в прототипе, $\lambda^1 = 1$(кроме эксперимента на зависимость от λ), $\theta = 0.9$. Графики строились в логарифмической шкале.

2.1 Зависимость от γ

Эксперимент проводился на датасете *w8a*.
Приведем графики того как γ (скорость изменения t_k) влияет на поведение метода.

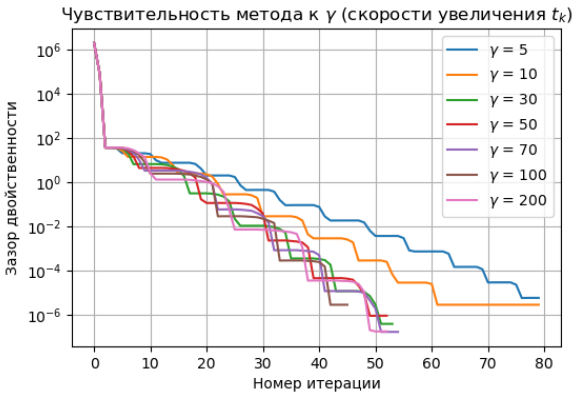


Рис. 1: Поведение метода по количеству шагов (x_k)

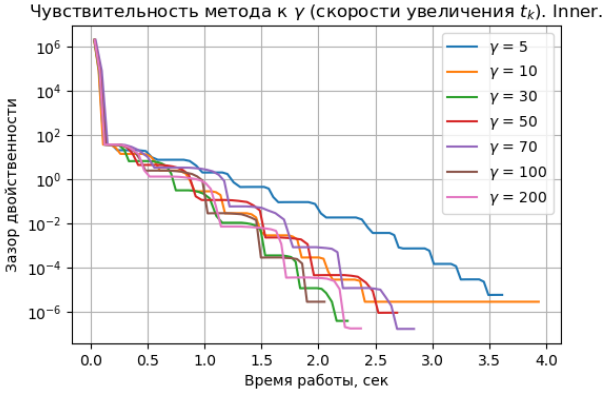


Рис. 2: Поведение метода по времени если смотреть на inner итерации, т.е. учитывая метод Ньютона

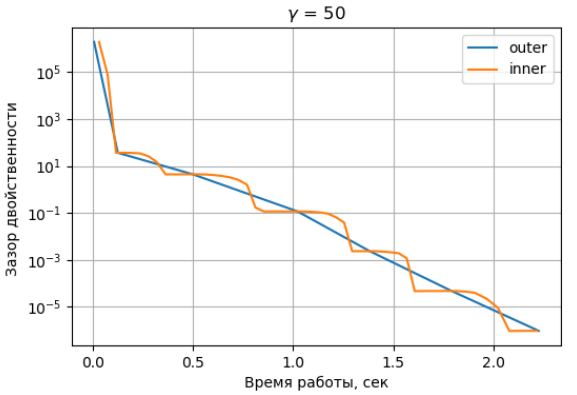


Рис. 3: Пример взаимного расположения inner и outer для одного и того же γ

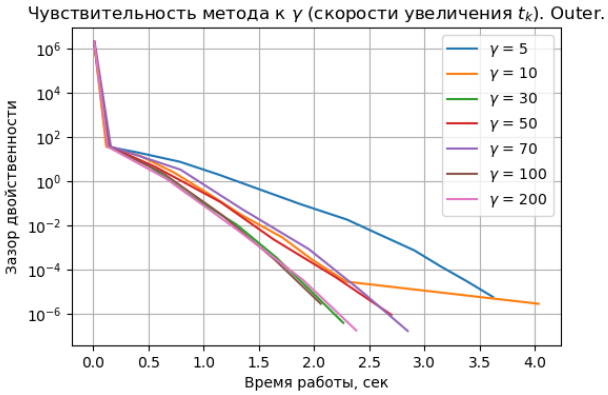


Рис. 4: Поведение метода по времени если смотреть на outer итерации

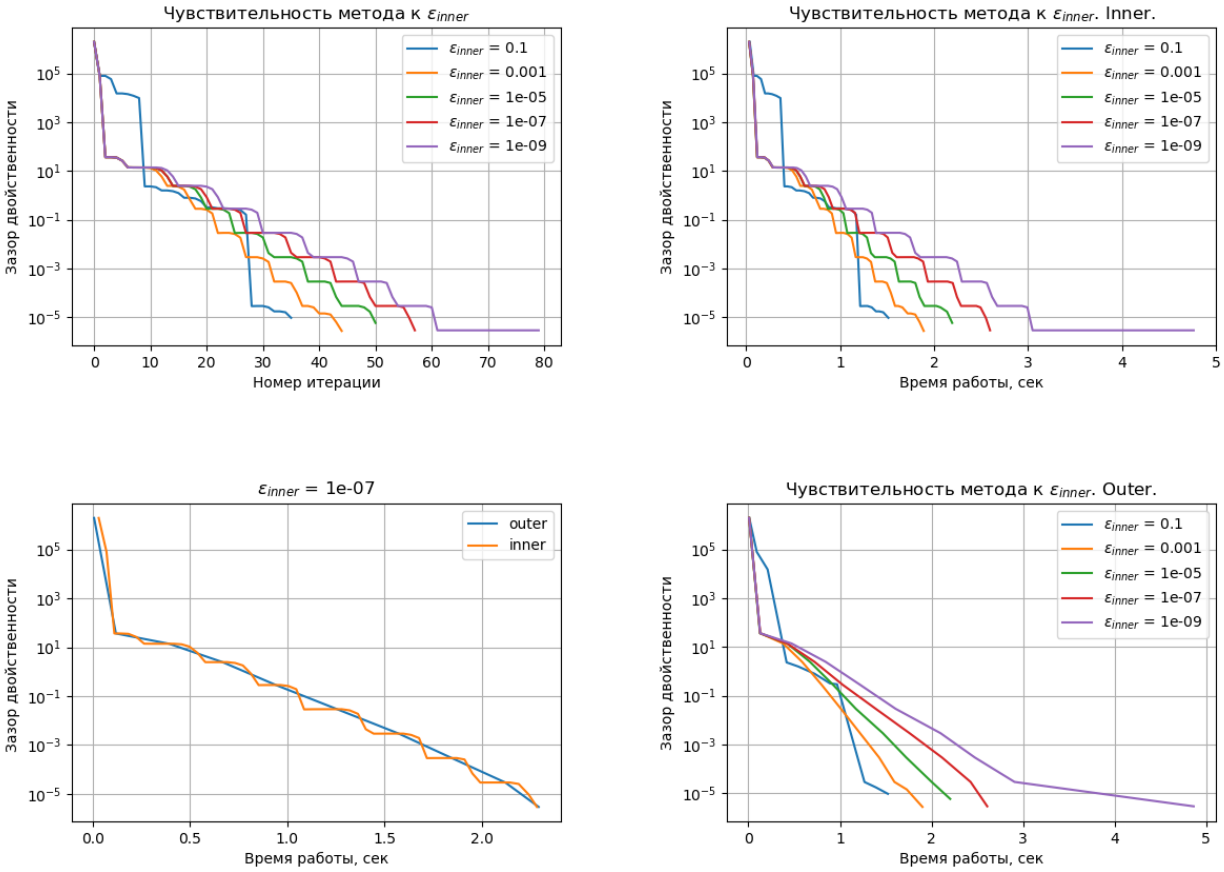
Выводы: Из Рис.2, Рис.4 видно, что сперва увеличение γ сперва приводит к уменьшению времени работы (при фиксированной точности). При $\gamma > 100$ начинается обратный процесс. Из рис. 4 видно, что по времени графики для различных γ (кроме малых значений) образуют 2 пучка, в каждом из которых поведение метода для различных γ практически не отличается. Таким образом, можно заключить, что оптимальное значение γ

[1] regcoef

находиться где-то между 10 и 100 (как и было рекомендовано в условии). Для конкретной задачи значение, видимо, определяется данными и значениями других параметров.

2.2 Зависимость от ϵ_{inner}

Эксперимент проводился на датасете $w8a$.
Приведем графики того как ϵ_{inner} (точность для метода Ньютона) влияет на поведение метода.

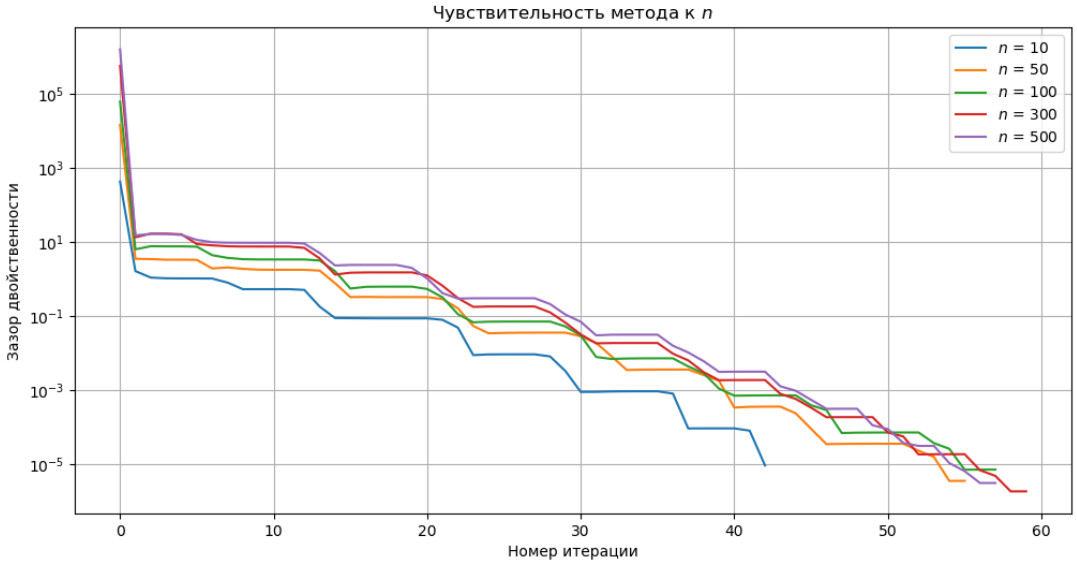


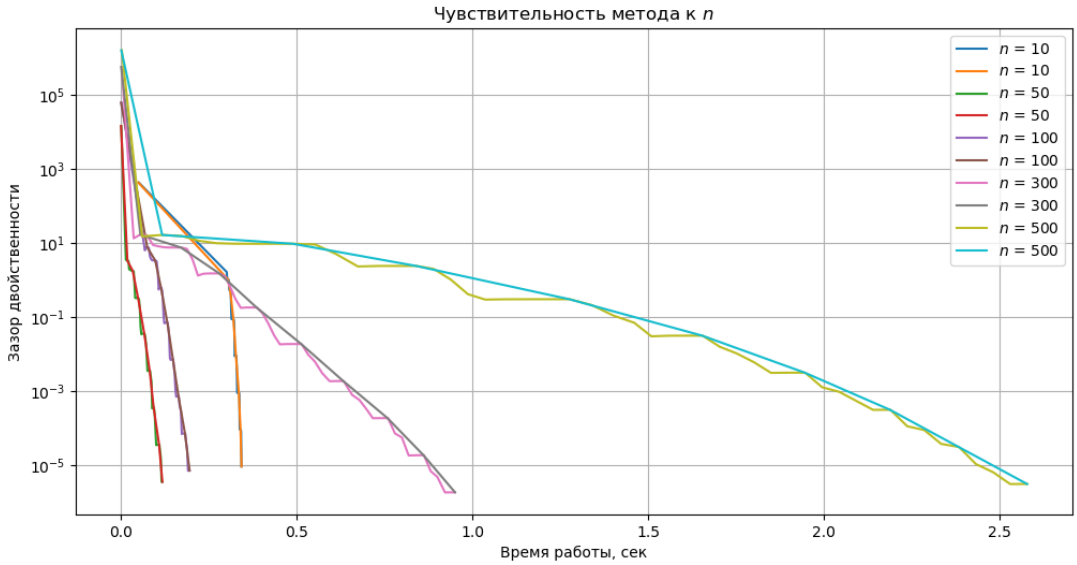
Выводы: Метод сходиться быстрее при меньшей точности $\uparrow$ ($\leftharpoonup\!\!\!\!\frown\!\!\!\!\leftharpoonup$) $\uparrow$. Но как видим при слишком больших значениях $\epsilon_{inner} = 0.1$ Поведение метода начинает принципиально отличаться (еще больше и метод не всегда сходиться), т.е. мы теряем вычислительную устойчивость. Также обратим внимание на то, что при увеличении ϵ_{inner} на 2 порядка время работы изменяется незначительно. Это видимо связано с локально квадратичной сходимостью метода Ньютона. Однако при слишком малых $\epsilon_{inner} = 1e{-}09$ решение последней оптимизационной задачи для метода Ньютона может занимать значительное время. Учитывая все вышесказанное наиболее предпочтительным выглядит значение порядка $\epsilon_{inner} = 1e - 07$.

2.3 Зависимость от n

Эксперимент проводился на случайно генерируемых данных ($np.random$), т. е. компоненты A, b брались случайным образом из $[0, 1)$. Значение $m = 200$ зафиксировано.
Приведем графики того как размерность пространства n влияет на поведение метода по итерациям и по времени.[2]

[2]Т.к графики для разных n различаются достаточно сильно сведем 3 рисунка для зависимости по времени из предыдущего пункта на 1 рисунок



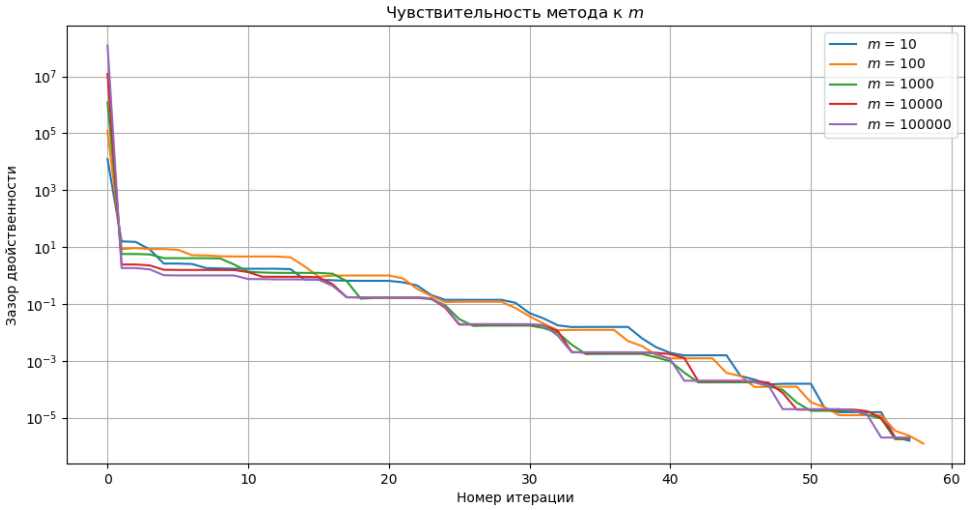


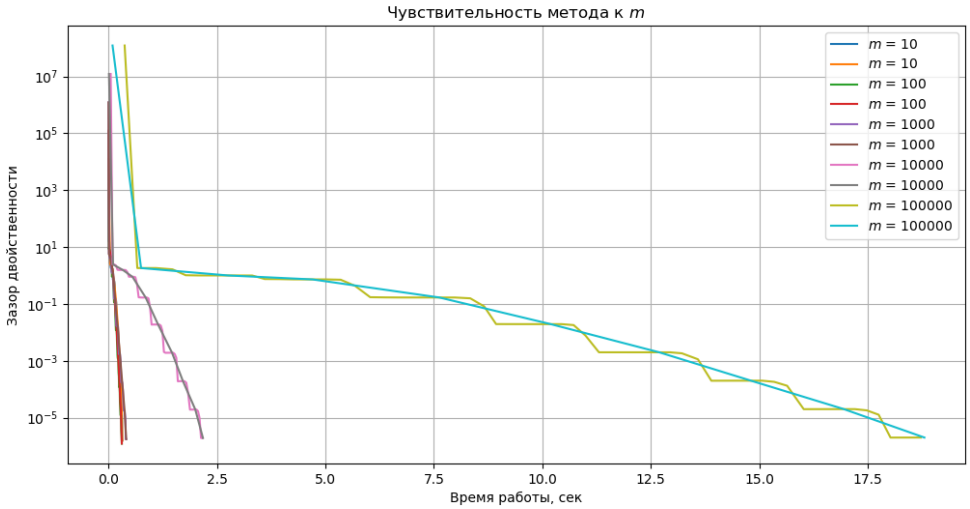
Обратившись к формулам (2),(3), обсудим ожидаемую зависимость от n, m. Вычисление $Ax - b$ и $A^T(Ax - b)$ требует $O(nm)$; A^TA требует $O(n^2m)$; метод Ньютона — $O(n^3)$. Т.е. на практике в разных ситуациях какой-то из вышеуказанных членов оказывает большее влияние на время работы метода и зависимость должна быть между квадратичной и кубической. Выводы: Из графиков видно, что время работы зависит от n^r, $r \in (2; 3)$, что вполне согласуется с вышесказанным.

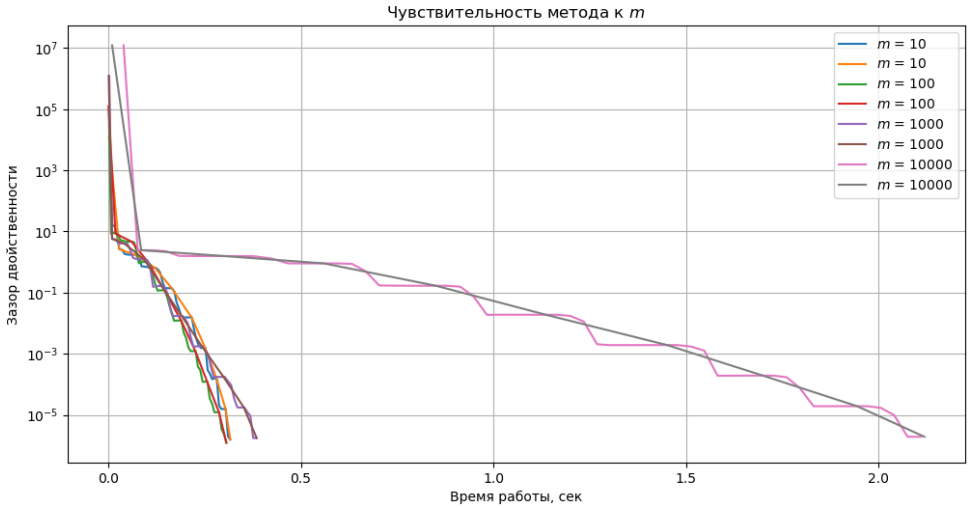
2.4 Зависимость от m

Эксперимент проводился на случайно генерируемых данных ($np.random$). Значение $n = 200$ зафиксировано.
Приведем графики того как размер выборки m влияет на скорость сходимости по итерациям и по времени:[3]

[3]Т.к графики для разных m различаются достаточно сильно сведем 3 рисунка для зависимости по времени из предыдущего пункта на 1 рисунок



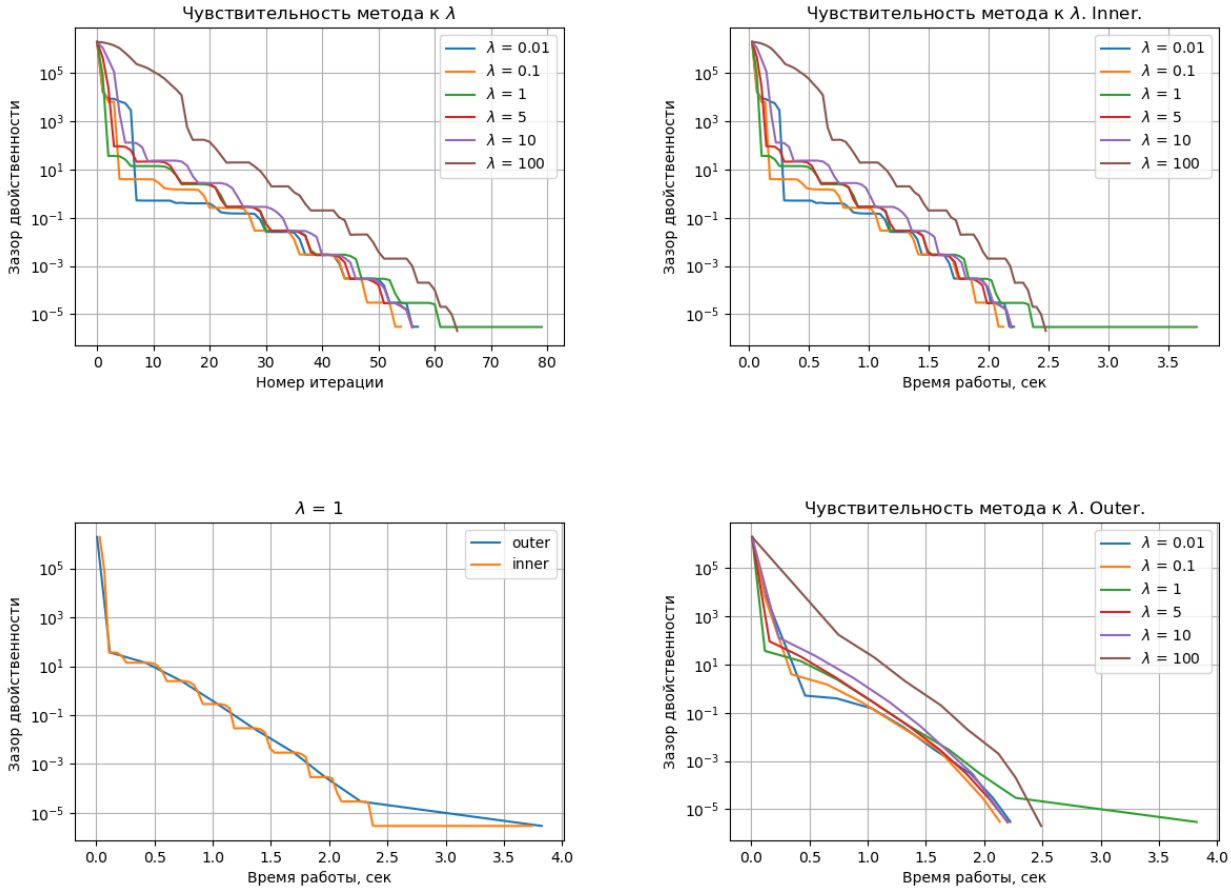




Выводы: По количеству итераций зависимости вообще не наблюдается. По времени работы: m меньше или сравнимых n зависимости практически нет; при больших m (на порядок и более больше n) зависимость наблюдается линейная. Выводы сделанные из графиков согласуются с теоретическими умозаключениями из предыдущего пункта.

2.5 Зависимость от λ

Эксперимент проводился на датасете $w8a$.
Приведем графики того как ϵ_{inner} (коэффициент регуляризации) влияет на поведение метода.



Выводы: При значениях λ от 0.01 до 10 различий в поведении метода не наблюдается, графики практически сливаются. При $\lambda = 100$ график не находится в общем пучке, но не отличается принципиально по поведению, и время работы метода увеличивается слабо. Подобных результат следовало ожидать т.к. значение λ увеличилось на порядок. Все вышесказанное логично (смотри формулы (2),(3)) т.к. λ не влияет на сложность вычислений, от него зависит количество ненулевых регрессоров.