Факультет компьютерных наук
Основная образовательная программа
Прикладная математика и информатика

Отчет по практическому домашнему заданию 3

Курс "методы оптимизации в машинном обучении"

Выполнил Петров Олег Евгеньевич, 193 группа

Содержание

1 Задание 3: вывод вспомогательной функции

$$\begin{cases} \frac{1}{2}\|Ax-b\|^2 + C\langle 1_d, u\rangle \to \min_{x,u} \\ x \leq u \\ x \geq -u \end{cases}$$

Предполагаю, что $A \in \mathbb{R}^{N\times d}$

1.1 Вспомогательная задача

$$\begin{cases} f_\tau(x,u) \to \min_{x,u} \\ Ax = b \end{cases}$$

$f_\tau(x,u) = f_\tau(x_1,\dots,x_d,u_1,\dots,u_d) = f_\tau(v)$

$f_\tau(v) := \tau f(v) + F(v)$

$F(v) = -\sum \log(u_i - x_i) - \sum \log(x_i + u_i)$

$f_\tau(v) = \frac{\tau}{2}\|Ax-b\|^2 + \tau C\langle 1_d, u\rangle - \sum \log(u-x) - \sum \log(x+u) =$

$\frac{\tau}{2}\|Ax-b\|^2 + \tau C\langle 1_d, u\rangle - \langle \log(u-x), 1_d\rangle - \langle \log(x+u), 1_d\rangle$

Гессиан будет выглядеть так:

$$\nabla_v^2 f\tau = \begin{bmatrix} \frac{\partial^2 f\tau}{\partial x_1^2} & \cdots & \frac{\partial^2 f\tau}{\partial x_1 \partial x_d} & \frac{\partial^2 f\tau}{\partial x_1 \partial u_1} & \cdots & \frac{\partial^2 f\tau}{\partial x_1 \partial u_d} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f\tau}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f\tau}{\partial x_d^2} & \frac{\partial^2 f\tau}{\partial x_d \partial u_1} & \cdots & \frac{\partial^2 f\tau}{\partial x_d \partial u_d} \\ \frac{\partial^2 f\tau}{\partial u_1 \partial x_1} & \cdots & \frac{\partial^2 f\tau}{\partial u_1 \partial x_d} & \frac{\partial^2 f\tau}{\partial u_1^2} & \cdots & \frac{\partial^2 f\tau}{\partial u_1 \partial u_d} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f\tau}{\partial u_d \partial x_1} & \cdots & \frac{\partial^2 f\tau}{\partial u_d \partial x_d} & \frac{\partial^2 f\tau}{\partial u_d \partial u_1} & \cdots & \frac{\partial^2 f\tau}{\partial u_d^2} \end{bmatrix} \in \mathbb{R}^{2d\times 2d}$$

В блочном виде:

$$\nabla_v^2 f\tau = \begin{bmatrix} \nabla_x^2 f\tau & \nabla_u(\nabla_x f_\tau) \\ \nabla_x(\nabla_u f_\tau) & \nabla_u^2 f\tau \end{bmatrix}$$

1.2 Вывод гессиана

$F(v) = -\langle \log(u-x), 1_d \rangle - \langle \log(x+u), 1_d \rangle = -\langle \log(u^2 - x^2), 1_d \rangle$
$\tau f(v) = \frac{\tau}{2}\|Ax - b\|^2 + \tau C \langle 1_d, u \rangle$

- $\tau d_x f = \tau (Ax - b)^T A dx$
- $\tau d_u f = \tau C 1_d^T du$

$d_x F = -1_d^T d_x \log(u-x) - 1_d^T d_x \log(x+u) = dx^T \frac{1}{u-x} - dx^T \frac{1}{x+u}$
$d_u F = -1_d^T d_u \log(u-x) - 1_d^T d_u \log(x+u) = -du^T \frac{1}{u-x} - du^T \frac{1}{x+u}$

Дифференциал по x:
$d_x f_\tau(v) = \tau d_x f(v) + d_x F(v) = \tau (Ax-b)^T A dx + d_x^T \frac{1}{u-x} - d_x^T \frac{1}{x+u}$
Дифференциал по u:
$d_u f_\tau(v) = \tau d_u f(v) + d_u F(v) = \tau C 1_d^T du - du^T \frac{1}{u-x} - du^T \frac{1}{x+u}$

Вторые и смешанные дифференциалы:

- $\tau d_x^2 f = d_x(\tau (Ax-b)^T A dx_1) = dx_1^T (\tau A^T A) dx_2$
- $\tau d_u^2 f = d_u(\tau C 1_d^T du_1) = 0$
- $\tau d_u(d_x f) = d_u(\tau (Ax-b)^T A dx) = 0$
- $\tau d_x(d_u f) = 0$

По F:

- $d_x^2 F = d_x(dx_1^T \frac{1}{u-x} - dx_1^T \frac{1}{x+u}) = dx_1^T \frac{dx_2}{(u-x)^2} + dx_1^T \frac{dx_2}{(x+u)^2} =$
  $= dx_1^T \Big(diag(\frac{1}{(u-x)^2}) + diag(\frac{1}{(x+u)^2}) \Big) dx_2$
- $d_u^2 F = d_u(-du_1^T \frac{1}{u-x} - du_1^T \frac{1}{x+u}) = du_1^T \Big(diag(\frac{1}{(u-x)^2}) + diag(\frac{1}{(x+u)^2}) \Big) du_2$
- $d_u(d_x F(v)) = d_u(dx^T \frac{1}{u-x} - dx^T \frac{1}{x+u}) = dx^T \Big(-diag(\frac{1}{(u-x)^2}) + diag(\frac{1}{(x+u)^2}) \Big) du$

- $d_x(d_u F(v)) = d_u(d_x F(v))$

Второй дифференциал по x:

$d_x^2 f_\tau = dx_1^T \Big(\tau A^T A + diag(\frac{1}{(u-x)^2}) + diag(\frac{1}{(x+u)^2}) \Big) dx_2$

Таким образом, гессиан по x:

$\nabla_x^2 f_\tau = \tau A^T A + diag(\frac{1}{(u-x)^2}) + diag(\frac{1}{(x+u)^2})$

Второй дифференциал по u:

$d_u^2 f_\tau == du_1^T \Big(diag(\frac{1}{(u-x)^2}) + diag(\frac{1}{(x+u)^2}) \Big) du_2$

Таким образом, гессиан по u:

$\nabla_u^2 f_\tau = diag(\frac{1}{(u-x)^2}) + diag(\frac{1}{(x+u)^2})$

$\nabla_x \nabla_u f_\tau = \nabla_u \nabla_x f_\tau = -diag(\frac{1}{(u-x)^2}) + diag(\frac{1}{(x+u)^2})$

В итоге:

$$\nabla_v^2 f_\tau = \begin{bmatrix} \tau A^T A + diag(\frac{1}{(u-x)^2}) + diag(\frac{1}{(x+u)^2}) & -diag(\frac{1}{(u-x)^2}) + diag(\frac{1}{(x+u)^2}) \\ -diag(\frac{1}{(u-x)^2}) + diag(\frac{1}{(x+u)^2}) & diag(\frac{1}{(u-x)^2}) + diag(\frac{1}{(x+u)^2}) \end{bmatrix}$$

В общем и целом, система выглядит так:

$$\begin{bmatrix} \nabla_v^2 f_\tau(v_k) \end{bmatrix} \begin{bmatrix} d_x^k \\ d_u^k \end{bmatrix} = \begin{bmatrix} -\nabla_v f_\tau(v_k) \end{bmatrix}$$

1.2.1 Градиент

Ясно, что $\nabla_v f_\tau(v) = [\nabla_x f_\tau(x), \nabla_u f_\tau(u)]$

$\nabla_x f_\tau = \tau A^T(Ax - b) + \frac{1}{u-x} - \frac{1}{x+u}$

$\nabla_u f_\tau = \tau C 1_d - \frac{1}{u-x} - \frac{1}{x+u}$

Тогда:

$$\begin{bmatrix}\nabla_v^2 f_\tau(v_k)\end{bmatrix}\begin{bmatrix}d_x^k \\ d_u^k\end{bmatrix} = \begin{bmatrix}-\nabla_x f_\tau(v_k) \\ -\nabla_u f_\tau(v_k)\end{bmatrix}$$

1.3 Анализ и предложения

Пусть $D_1 = diag(\frac{1}{(u-x)^2})$, $D_2 = diag(\frac{1}{(x+u)^2})$; $D_1, D_2 \in \mathbb{R}^{d\times d}$.

$$\nabla_v^2 f_\tau = \begin{bmatrix}\tau A^T A + D_1 + D_2 & D_2 - D_1 \\ D_2 - D_1 & D_1 + D_2\end{bmatrix}$$

Сложение матриц требует $O(d^2)$ времени. Нужно уметь строить матрицу $A^T A$, что потребует $O(Nd^2)$ времени. Обозначим $M_1 = D_2 - D_1$, $M_2 = D_1 + D_2$ и выпишем систему полностью:

$$\begin{bmatrix}\tau A^T A + M_1 & M_2 \\ M_2 & M_1\end{bmatrix}\begin{bmatrix}d_x^k \\ d_u^k\end{bmatrix} = \begin{bmatrix}-\nabla_x f_\tau(v_k) \\ -\nabla_u f_\tau(v_k)\end{bmatrix}$$

Стоить отметить, что матрица здесь является симметричной. M_1, M_2 являются диагональными. Это значит, что можно эффективно умножать матрицы M друг на друга и считать произведения вида Mv через поэлементное умножение векторов.

$$\begin{bmatrix}M_2 & M_1\end{bmatrix}\begin{bmatrix}d_x^k \\ d_u^k\end{bmatrix} = M_2 d_x^k + M_1 d_u^k = -\nabla_u f_\tau(v_k)$$

$$\begin{bmatrix}\tau A^T A + M_1 & M_2\end{bmatrix}\begin{bmatrix}d_x^k \\ d_u^k\end{bmatrix} = (\tau A^T A + M_1) d_x^k + M_2 d_u^k = -\nabla_x f_\tau(v_k)$$

Можно выразить одно через другое.

Пусть $d_u^k = -M_1^{-1}(M_2 d_x^k + \nabla_u f_\tau(v_k)) = -M_1^{-1} M_2 d_x^k - M_1^{-1}\nabla_u f_\tau(v_k)$.

Подставляем:

$(\tau A^T A + M_1) d_x^k - M_2(M_1^{-1} M_2 d_x^k + M_1^{-1}\nabla_u f_\tau(v_k)) = -\nabla_x f_\tau(v_k)$

$M_1^{-1} M_2 = T$

$(\tau A^T A + M_1) d_x^k - T M_2 d_x^k - T\nabla_u f_\tau(v_k) = -\nabla_x f_\tau(v_k)$

$$(\tau A^T A + M_1 - TM_2)d_x^k = -\nabla_x f_\tau(v_k) + T\nabla_u f_\tau(v_k)$$
$$d_x^k = (\tau A^T A + M_1 - TM_2)^{-1}(-\nabla_x f_\tau(v_k) + T\nabla_u f_\tau(v_k))$$

Матрица $\tau A^T A + M_1 - TM_2 \in \mathbb{R}^{d\times d}$ является симметричной. Для решения СЛУ выше можно запустить какой-нибудь солвер, который работает с симметричными неопределенными матрицами. Подойдет метод MINRES (в силу симметричности). Если угловые миноры невырождены, возможно использование семейства LU-разложений.

Подсчитаем асимптотику:

- Вычисление d_u^k: $O(d)$
- Решение СЛУ и вычисление d_x^k: используя метод MINRES, получим решение за $O(k(d+k))$, где k – число итераций (источник)

Если бы изначальная система размера $2d \times 2d$ решалась через разложение Холецкого, потребовалось бы $O(8d^3)$ времени.

Недостаток: могут быть проблемы с вычислительной устойчивостью.

1.4 Поиск максимального значения длины шага α

По инструкции пункта 1.2 документации, представим функции вида g в аффином виде: $g(v) = q^T v - s$. Это действительно для $g : \mathbb{R}^{2d} \to \mathbb{R}$, однако в конкретном случае имеем $g : \mathbb{R}^{2d} \to \mathbb{R}^d$ (можно расписывать покомпонентно, но проще в конце взять покомпонентный минимум – выражаем солидарность коллеге за идею).

Тогда функция имеет вид $g(v) = Qv - s$, где $Q \in \mathbb{R}^{d\times 2d}, s \in \mathbb{R}^d$. Таким образом,

$$g_1(v) = x - u \implies \quad Q_1 = \begin{bmatrix} diag(1_d) & diag(-1_d) \end{bmatrix}$$
$$g_2(v) = -x - u \implies Q_2 = \begin{bmatrix} diag(-1_d) & diag(-1_d) \end{bmatrix}$$

Введя множества $I_1^k = \{i \mid d_{x_i}^k - d_{u_i}^k > 0\}$ и $I_2^k = \{i \mid d_{x_i}^k + d_{u_i}^k > 0\}$
$\alpha_{\max}^k = \min\{\min_{i \in I_1^k} \frac{s_1 - Q_1 v_k}{Q_1 d_k}, \min_{i \in I_2^k} \frac{s_2 - Q_2 v_k}{Q_2 d_k}\}$
$\alpha_{\max}^k = \min\{\min_{i \in I_1^k} \frac{-x_k + u_k}{d_x^k - d_u^k}, \min_{i \in I_2^k} \frac{-x_k - u_k}{d_x^k + d_u^k}\}$

1.5 Выбор начальной точки (x_0, u_0)

Достаточно взять $u_0 = 1_d$, $x_0 = 0_d$. В этом случае выполняется строгое неравенство $-u < x < u$, что необходимо.

2 Эксперимент

2.1 Дизайн эксперимента

Основная цель эксперимента – выявить, как меняется поведение метода в зависимости от его гиперпараметров и размерности выборки (непосредственна размера выборки и размерности пространства, на котором проводится оптимизация).

Рассматриваются следующие гиперпараметры: скорость γ увеличения величины τ_k, коэффициент регуляризации λ, параметр точности (внутреннего) метода Ньютона ϵ_{inner}.

Эксперимент разбивается на две части. В первой предлагается проанализировать чувствительность метода к выбору γ, ϵ_{inner}. Здесь мы используем набор данных w8a.

Во второй же части требуется исследовать поведение метода в зависимости от и значений m, n (т. что $A \in \mathbb{R}^{m \times n}$). Задавая параметры m, n, мы генерируем случайную систему уравнений.

Предлагается строить графики вида «гарантируемая точность по зазору двойственности против числа итераций/реального времени работы» в логарифмической шкале. Гарантируемая точность по зазору двойственности определяется по формуле (2.3) инструкции.

Используемые значения параметров:

- $\gamma \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$
- $\epsilon_{inner} \in \{1e - i | i \in \{2, 3, \dots, 12\}\}$
- $\lambda \in \{0.1, 0.2, 0.5, 0.7, 1, 2, 5, 7, 10\}$
- $m \in \{100, 500, 1000, 2500, 5000, 10000\}$, $n = 1000$
- $n \in \{100, 500, 1000, 2500, 5000\}$, $m = 10000$

2.2 Результаты

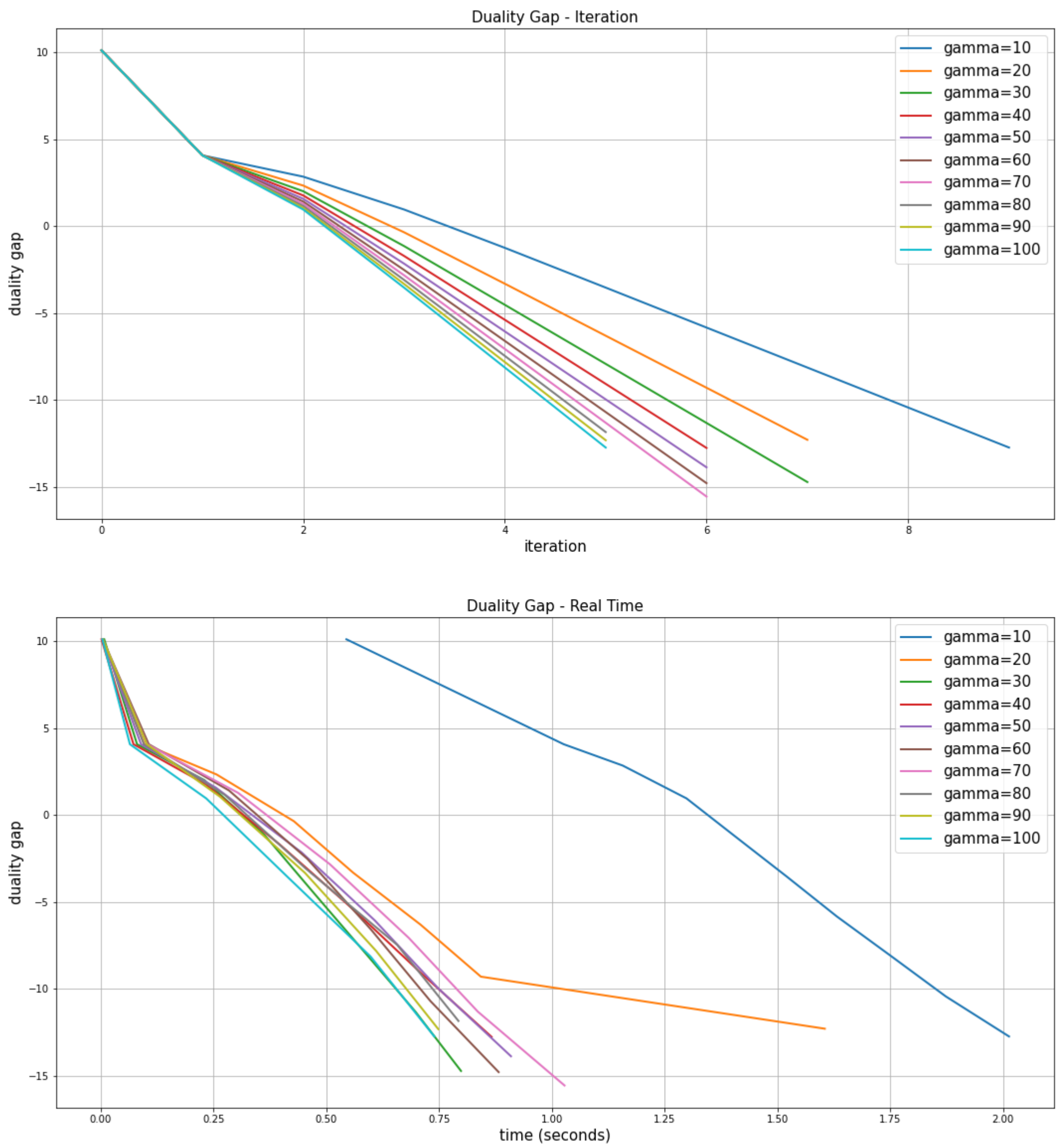


Рис. 1. чувствительность метода к γ

Метод достаточно чувствителен к гиперпарамтру γ: при малых значениях (10, 20) сходимость по числу итераций долгая, желаемая точность может не быть достигнута. При больших значениях (свыше 80) метод сходится быстрее, однако проблема с точностью проявляется еще больше.

По времени выполнения наблюдается значительная разница между небольшими значениями γ (10, 20). При больших значениях разница менее заметна.

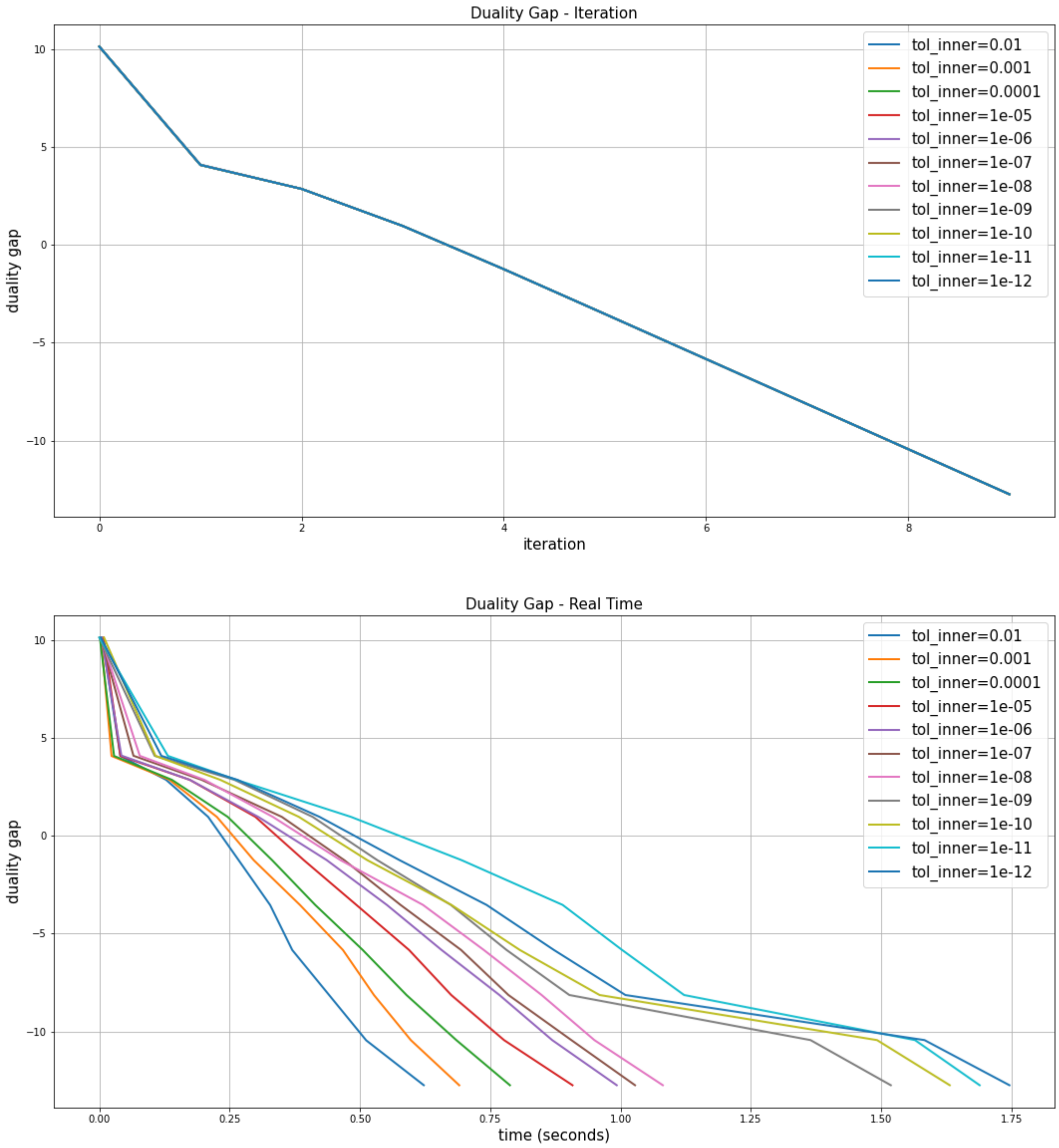


Рис. 2. чувствительность метода к ϵ_{inner}

Выбор ϵ_{inner} не влияет на число итераций метода барьеров, однако влияет на число итераций в методе Ньютона, что влечет за собой линейное увеличение времени выполнения при росте значения параметра.

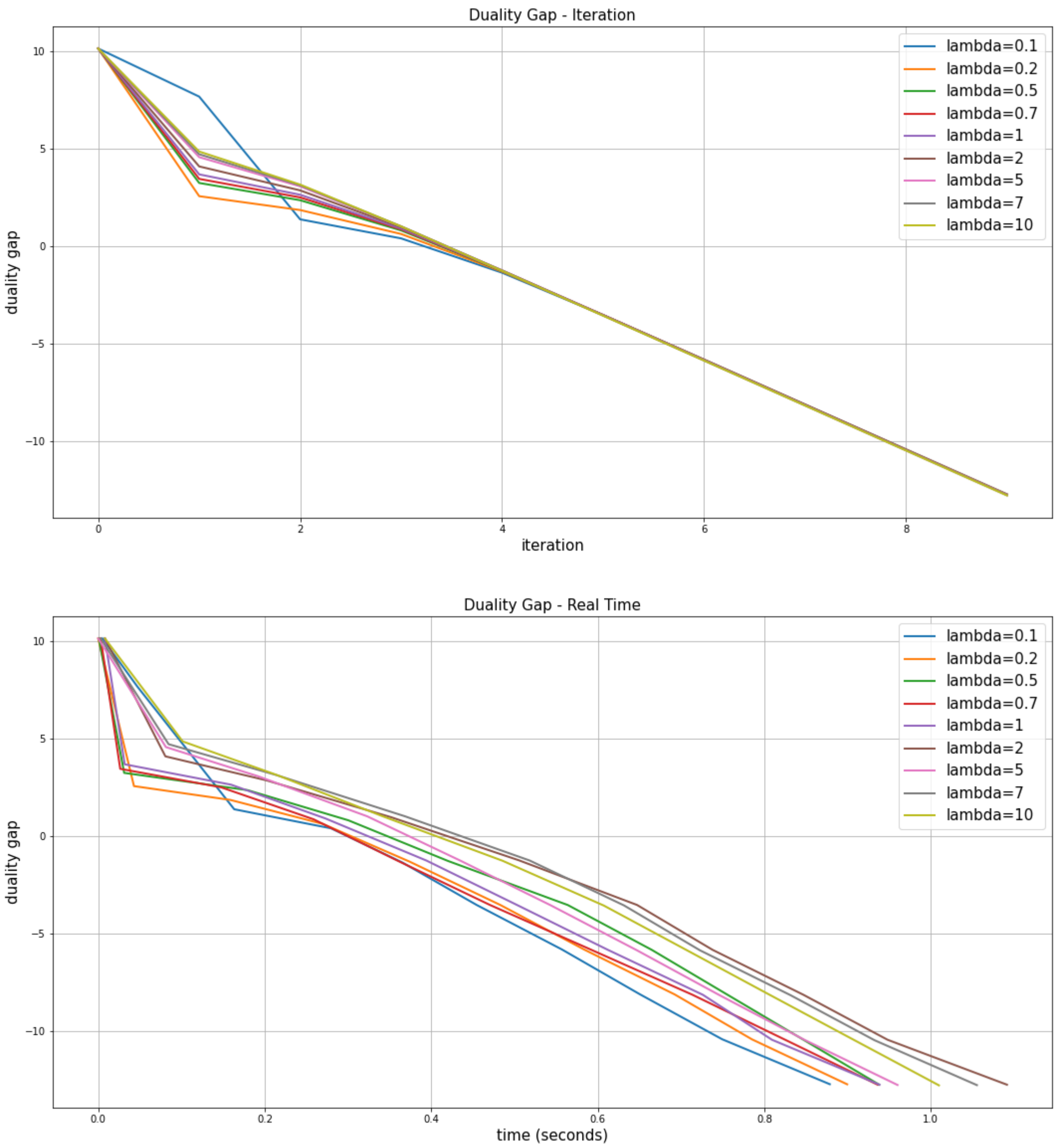
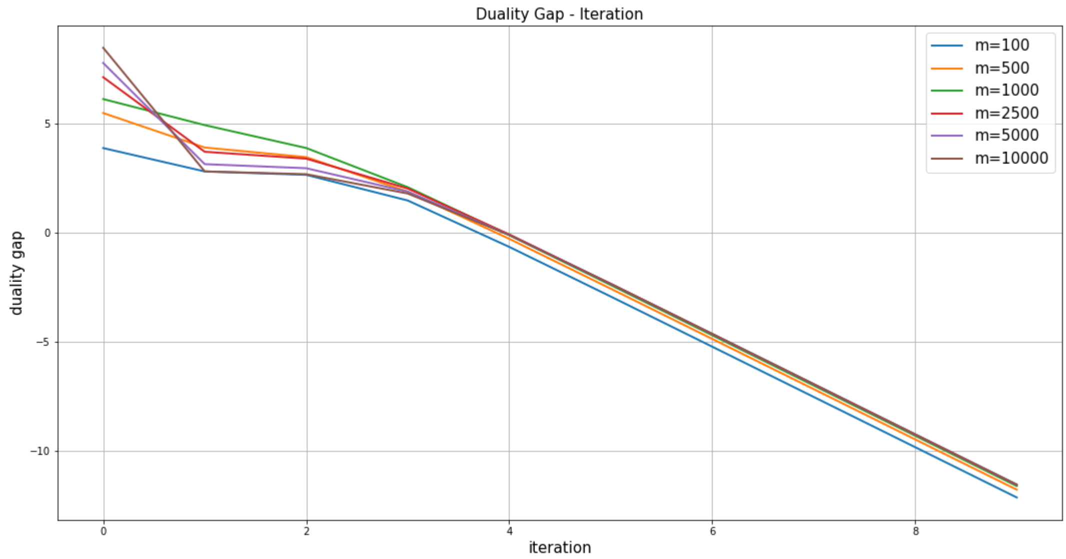


Рис. 3. чувствительность метода к λ

Выбор коэффициента регуляризации абсолютно не влияет на число итераций, если итераций много – видно, как графики для различныз значений собираются в единый пучок. При этом выбор параметра влияет на реальное время выполнения, и чем он меньше, тем быстрее. На графике (рис. 3). Разница между минимальным и максимальным временем выполнения составляет приблизительно 0.2 секунды, что не очень много.



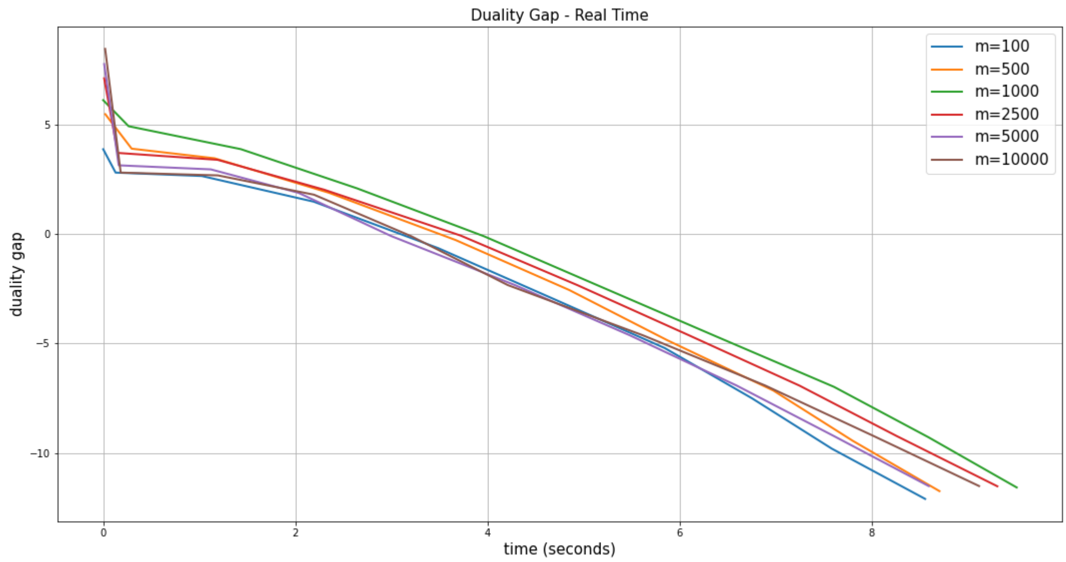
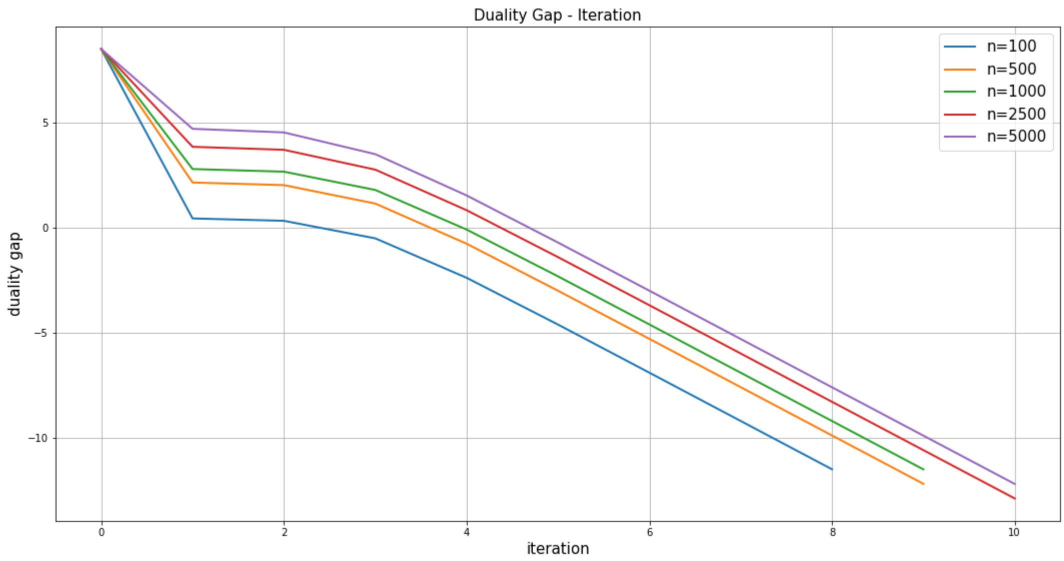


Рис. 4. Зависимость зазора от m при фиксированном $n = 1000$

Параметр m не влияет на число итераций. В незначительной степени влияет на время выполнения: чем меньше m, тем меньше время. Это связано с вычислением $A^T A$, $Ax - b$ и других операци1.



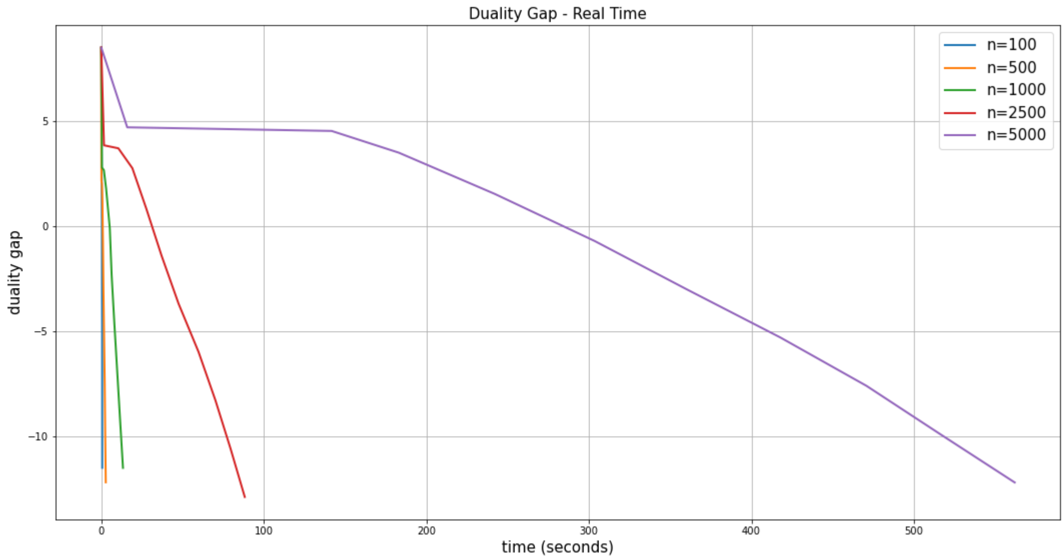


Рис. 5. Зависимость зазора от n при фиксированном $m = 10000$

Размерность оптимизируемого пространства значительно влияет на поведение метода. Во-первых, число итераций положительно линейно по отношению к n: чем меньше n, тем меньше итераций требуется (рис. 4). Во-вторых, зависимость между n и реальным временем выполнения – квадратичная, что отчетливо видно на графике (рис. 4). Это объясняется $O(n^2)$-асимптотикой. В коде представлена реализация метода Ньютона с помощью разложения Холецкого, однако метод, предложенный в пункте 1.3 данного отчета, гипотетически мог бы уменьшить время работы метода при увеличении n.

2.3 Выводы

- Метод чувствителен к γ, требуется подбор этого гиперпараметра для уменьшения числа итераций и времени выполнения. При росте значения гиперпараметра выигрыш в реальном времени работе все менее и менее значителен.

- ϵ_{inner} лучше всего выбирать грубо: не требуется выходить на центральный путь, чтобы хорошо сходиться.

- λ должен выбираться из соображений борьбы с переобучением или отбором признаков. Тем не менее, параметр незначительно влияет на время работы метода.

- Размер выборки m не влияет на число итераций. На время выполнения влияет незначительно. Это говорит о том, что можно брать большие выборки (вытянутые вниз).

- Размерность пространства n сильно влияет на сходимость метода. Вероятно, в случае с большим значением параметра хорошим решением было бы понизить размерность. С $n > 2000$ метод лучше не запускать.