Convex Optimization - Homework 3

Realized By: Eya Ghamgui
eya.ghamgui@telecom-paris.fr

1. Dual Problem of Lasso

We have $x_1, ..., x_n \in \mathbb{R}^d$ data vectors and $y_1, ..., y_n \in \mathbb{R}$ observations. We are searching for regression parameters $w \in \mathbb{R}^d$.

The Least Absolute Shrinkage Operator and Selection Operator (LASSO) Problem is the following:

$$\underset{w}{\text{minimize}} \;\; \frac{1}{2}||Xw - y||_2^2 + \lambda ||w||_1 \tag{1}$$

$$\begin{aligned} Where: &- X = (x_1^T, ..., x_n^T) \in \mathbb{R}^{n\times d} \\ &- y = (y_1, ..., y_n) \in \mathbb{R}^n \\ &- \lambda > 0 \;\text{ is a regularization parameter} \end{aligned}$$

Let's introduce an equality constraint to the problem. Thus, the problem (1) is equivalent to:

$$\underset{w,Z}{\text{minimize}} \;\; \frac{1}{2}||Z - y||_2^2 + \lambda ||w||_1$$

$$s.t. \;\; Z - Xw = 0$$

The Lagrangian function is the following:

$$\begin{aligned} L(w, Z, \nu) &= \frac{1}{2}||Z - y||_2^2 + \lambda||w||_1 + \nu^T(Z - Xw) \\ &= (\frac{1}{2}||Z - y||_2^2 + \nu^T Z) + (\lambda||w||_1 - \nu^T Xw) \end{aligned}$$

Let's suppose $D = \{(w, Z) \in \mathbb{R}^d \times \mathbb{R}^n | Z - Xw = 0\}$

The dual function:

$$\begin{aligned} g(\nu) &= \inf_{(w,Z)\in D} \;\; L(w, Z, \nu) \\ &= \inf_{(w,Z)\in D} \;\; \{(\frac{1}{2}||Z - y||_2^2 + \nu^T Z) + (\lambda||w||_1 - \nu^T Xw)\} \\ &= \inf_{(w,Z)\in D} \;\; (\frac{1}{2}||Z - y||_2^2 + \nu^T Z) + \inf_{(w,Z)\in D} \;\; (\lambda||w||_1 - \nu^T Xw) \\ &= \inf_{Z} \;\; (\frac{1}{2}||Z - y||_2^2 + \nu^T Z) + \inf_{w} \;\; (\lambda||w||_1 - \nu^T Xw) \\ &= \inf_{Z} \;\; (\frac{1}{2}||Z - y||_2^2 + \nu^T Z) + \lambda \;\inf_{w} \;\; (||w||_1 - (\frac{X^T\nu}{\lambda})^T w) \\ &= \inf_{Z} \;\; (\frac{1}{2}||Z - y||_2^2 + \nu^T Z) - \lambda \;\sup_{w} \;\; ((\frac{X^T\nu}{\lambda})^T w - ||w||_1) \end{aligned}$$

We have: $\sup\limits_{w} \ ((\frac{X^T\nu}{\lambda})^T w - ||w||_1) = f^*(\frac{X^T\nu}{\lambda})$.

Where $f^*(y)$ is the conjugate function of $f(x) = ||x||_1$. Thus,

$$f^*(y) = \begin{cases} 0 & \text{if } ||y||_\infty \leq 1 \\ +\infty & \text{otherwise} \end{cases}$$

Let's suppose $h(Z) = \frac{1}{2}||Z-y||_2^2 + \nu^T Z$ which is a quadratic function, differentiable.

$$\frac{\partial h(Z)}{\partial Z} = \frac{\partial \frac{1}{2}||Z-y||_2^2 + \nu^T Z}{\partial Z} = \frac{\partial \frac{1}{2}(Z-y)^T(Z-y) + \nu^T Z}{\partial Z} = Z - y + \nu = 0$$

$\implies Z = y - \nu$

$\implies \inf\limits_{Z} \ (\frac{1}{2}||Z-y||_2^2 + \nu^T Z) = \frac{1}{2}||\nu||_2^2 + \nu^T(y-\nu) = \frac{1}{2}||\nu||_2^2 + \nu^T y - ||\nu||_2^2 = -\frac{1}{2}||\nu||_2^2 + \nu^T y$

We have shown that:

$$g(\nu) = \begin{cases} -\frac{1}{2}||\nu||_2^2 + \nu^T y & \text{if } ||\frac{X^T\nu}{\lambda}||_\infty \leq 1 \\ -\infty & \text{otherwise} \end{cases} = \begin{cases} -\frac{1}{2}||\nu||_2^2 + \nu^T y & \text{if } ||X^T\nu||_\infty \leq \lambda \\ -\infty & \text{otherwise} \end{cases}$$

We found that g is a quadratic function with negative coefficients on a convex set: $\{\nu \in \mathbb{R}^n | \ \ ||X^T\nu||_\infty < \lambda\}$. Thus, g is a concave function. We can write the dual problem as:

$$\begin{cases} \underset{\nu}{\text{maximize}} & -\frac{1}{2}||\nu||_2^2 + \nu^T y \\ s.t. & ||X^T\nu||_\infty \leq \lambda \end{cases}$$

$$\iff$$

$$\begin{cases} \underset{\nu}{\text{minimize}} & \frac{1}{2}||\nu||_2^2 - \nu^T y \\ s.t. & ||X^T\nu||_\infty \leq \lambda \end{cases}$$

$$\iff$$

$$\begin{cases} \underset{\nu}{\text{minimize}} & \frac{1}{2}\nu^T\nu - y^T\nu \\ s.t. & ||X^T\nu||_\infty \leq \lambda \end{cases}$$

$$\iff$$

$$\begin{cases} \underset{\nu}{\text{minimize}} & \nu^T(\frac{1}{2}I_n)\nu - y^T\nu \\ s.t. & ||X^T\nu||_\infty \leq \lambda \end{cases}$$

We have: $||X^T\nu||_\infty \leq \lambda \implies$ for all i, $(X^T\nu)_i \leq \lambda$ and $-(X^T\nu)_i \leq \lambda$.

$$||X^T\nu||_\infty \leq \lambda \implies (X^T, -X^T)\nu \leq \begin{pmatrix} \lambda \\ . \\ . \\ . \\ \lambda \end{pmatrix} \in \mathbb{R}^{2d}$$

The dual problem:

$$\begin{cases} \underset{\nu}{\text{minimize}} & \nu^T Q\nu + p^T\nu \\ s.t. & A\nu \preccurlyeq b \end{cases} \qquad \text{(Quadratic Problem)}$$

$$\begin{aligned} Where: & - Q = \frac{1}{2}I_n \succcurlyeq 0 \\ & - A = (X^T, -X^T) \in \mathbb{R}^{2d\times n} \\ & - p = -y \\ & - b = (\lambda, ..., \lambda) \in \mathbb{R}^{2d} \\ & - \nu \in \mathbb{R}^n \end{aligned}$$

2. Implementation

In this question, we will implement the barrier method to solve the previous quadratic problem. In this case, we will use the Newton method to solve the centering step. The centering problem is therefore written as:

$$\text{minimize} \;\; tf_0(x) - \sum_{i=1}^{m} log(-f_i(x))$$

.

Let's denote $f(\nu) = t(\nu^T Q\nu + p^T\nu) + \sum_{i=1}^{2d} log(b_i - A_i\nu)$

Where A_i is the ith line of the matrix A.

$\implies \nabla f(\nu) = t(2Q\nu + p) + \sum_{i=1}^{2d} \frac{A_i^T}{bi - A_i\nu}$

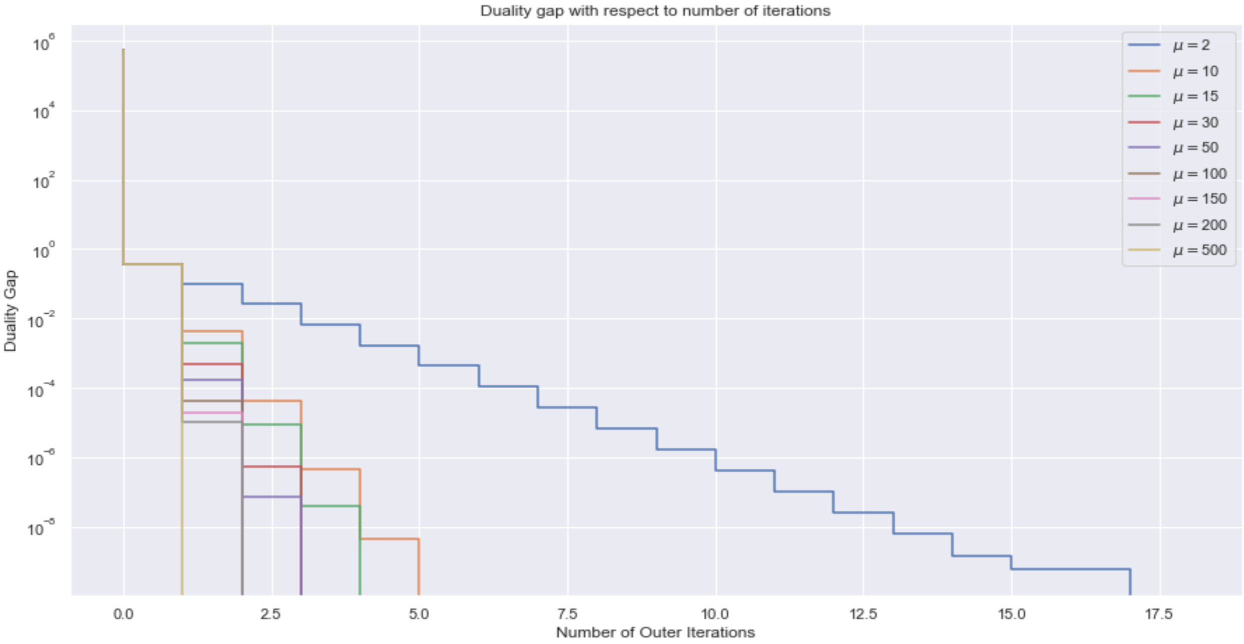
$\implies \nabla^2 f(\nu) = t(2Q) + \sum_{i=1}^{2d} \frac{A_i^T A_i}{(bi - A_i\nu)^2}$

While implementing, we will choose the parameters as follow:
- $\alpha = 0.01 \in (0, \frac{1}{2})$
- $\beta = 0.5 \in (0, 1)$
- $\epsilon = 10^{-5}$ in Newton method
- $\epsilon = 10^{-3}$ in Barrier method
- $t_0 = 1 > 0$
- $\mu \in [2, 10, 15, 30, 50, 100, 150, 200, 500]$

3. Experimental Results

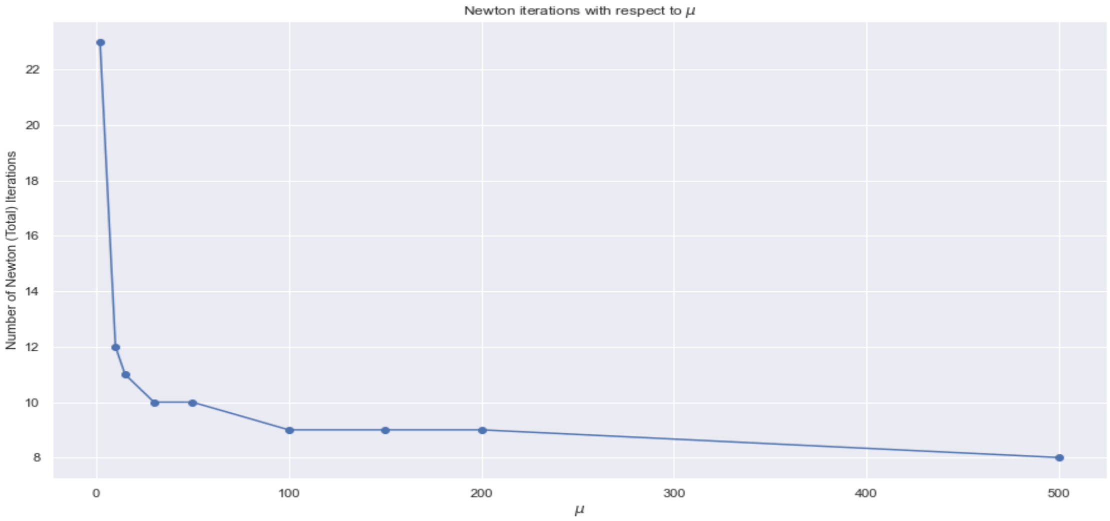
In this part, we generate data randomly with $n = 50$ and $d = 100$. For initialization, we chose $v_0 = 0$ which is a feasible point. We obtained the following results and the objective function has a value arround -546860.712189.

3.1. Duality Gap



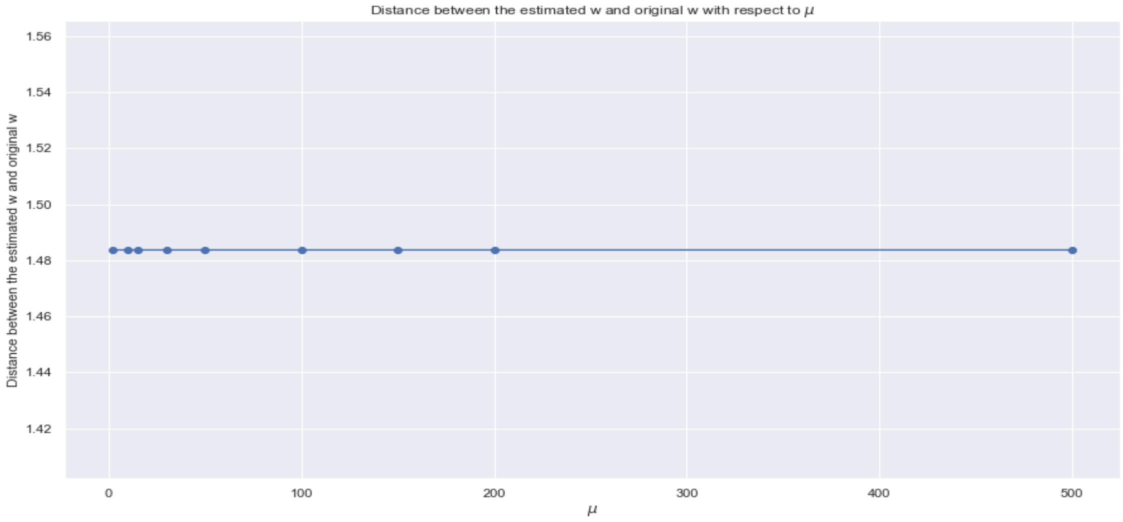
In the previous figure, each curve corresponds to a step function of the duality gap as a function of the number of Newton steps. We can notice that, if μ is small ($\mu = 2$), there is a small number of iterations at each step. However, the outer iteration for this value is too large, equal to 16. Now, if μ is too large ($\mu = 500$), there is a large number of iterations at each step and the outer iteration for this value is very small, equal to 1. To get the best value of μ, we should make a trade-off between the number of inner iterations and the number of outer iteration. I think the best value is between 50 and 100.

3.2. Newton Iterations



From the previous figure, we can say that the total number of iterations decreases when the value of μ increases. When μ is higher, we need less iterations to converge to the optimal solution.

3.3. Impact of μ on w



From the previous figure, we can say that the value of μ has no impact on the estimation of w. The algorithm converges to the same optimal solution for all values of μ. However, the distance between the estimated w and the original w is slightly high around 1.49.