MVA HW3 : Convex Optimization

Amin Dhaou

November 2019

Introduction

Given $x_1, \ldots, x_n \in \mathbb{R}^n$ data vectors and $y_1, \ldots, y_n$ R observations, we are searching for regression parameters $\in R^d$ which fit data inputs to observations y by minimizing their squared difference. In a high dimensional setting (when n $\leq$ d) a '1norm penalty is often used on the regression coefficients w in order to enforce sparsity of the solution (so that w will only have a few non-zeros entries). Such penalization has well known statistical properties, and makes the model both more interpretable, and faster at test time. From an optimization point of view we want to solve the following problem called LASSO (which stands for Least Absolute Shrinkage Operator and Selection Operator)

$$\text{minimize } \frac{1}{2}\|Xw - y\|_2^2 + \lambda\|w\|_1 \quad \text{(LASSO)} \tag{1}$$

Question 1

To get a constrained problem, we use a variable z $\in$ R^n such that we get an equivalent reformulation of our problem :

$$\min_{w,z} \frac{1}{2}\|z\|_2^2 + \lambda\|w\|_1. \quad \text{s.t} \quad z = Xw - y$$

Take $\nu \in \mathbb{R}^n$, we compute the lagrangian as:

$$L(w, z, \nu) = \frac{1}{2}\|z\|_2^2 + \lambda\|w\|_1 + \nu^T(z - Xw + y). \qquad 2005/06/2$$

Because, ω and z are independant in the Lagrangian, we can find the minimal values with respect to z and ω separately. We can therefore separate the two variables:

$$\inf_{w,z} L(w, z, \nu) = \inf_{w,z}\left[\frac{1}{2}\|z\|_2^2 + \lambda\|w\|_1 + \nu^T(z - Xw + y)\right]$$

$$= \inf_z\left[\tfrac{1}{2}\|z\|_2^2 + \nu^T z + \inf_w(\lambda\|w\|_1 - \nu^T Xw)\right] + \nu^T y$$

$$= \inf_z\left[\tfrac{1}{2}\|z\|_2^2 + \nu^T z - \lambda \sup_w(\tfrac{1}{\lambda}(X^T\nu)^T w - \|w\|_1)\right] + \nu^T y$$

The function multiplied by lambda is the conjugate of $\|w\|_1$ that we calculated in the last homework.

$$f : x \mapsto \|x\|_1, f^*(y) = \begin{cases} 0 & \text{if } \|y\|_\infty \leq 1 \\ \infty & \text{otherwise} \end{cases}$$

Hence:
$$\inf_{w,z} L(w, z, \nu) = \begin{cases} \inf_z\left[\frac{1}{2}\|z\|_2^2 + \nu^T z\right] + \nu^T y \text{ if } & \|X^T\nu\|_\infty \leq \lambda \\ -\infty \end{cases}$$

The part of the lagrangian with respect to z is a convex function. We can easily find its minimum by computing its gradian and looking at its critical points.

We find that the minimum is attained at $z = -\nu$, hence $\inf_z\left[\frac{1}{2}\|z\|_2^2 + \nu^T z\right] = -\frac{1}{2}\|\nu\|_2^2$.

Finally, we get:

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

The dual problem of LASSO is therefore

$$\max_\nu -\frac{1}{2}\|\nu\|_2^2 + \nu^T y \quad \text{s.t} \quad \|X^T\nu\|_\infty \leq \lambda$$

We know that $\|X^T\nu\|_\infty \leq \lambda \iff \begin{pmatrix} X^\top \\ -X^\top \end{pmatrix} \leq \lambda \mathbb{I}_{2d}$. Therefore, by equivalence:

$$\max_\nu -\frac{1}{2}\nu^T I_n \nu + y^T \nu \quad \text{s.t} \quad \begin{pmatrix} X^\top \\ -X^\top \end{pmatrix} \leq \lambda \mathbb{I}_{2d}$$

Thus, this lead to the quadratic problem :

$$\min_\nu \nu^T Q \nu + p^T \nu \quad \text{s.t} \quad A\nu \leq \lambda b$$

with

$$Q = \frac{1}{2} I_n, \quad p = -y, \quad A = \begin{pmatrix} X^\top \\ -X^\top \end{pmatrix}, \quad b = \lambda \mathbb{I}_{2d}$$

Question 2

In my code, let g be the function:

$$g : \nu \mapsto t(\nu^T Q \nu + p^T) - \sum_{i=1}^{d} \log(b_i - A_{[i,:]}\nu)$$

where $A_{[i,:]}$ is the i-th line of A.

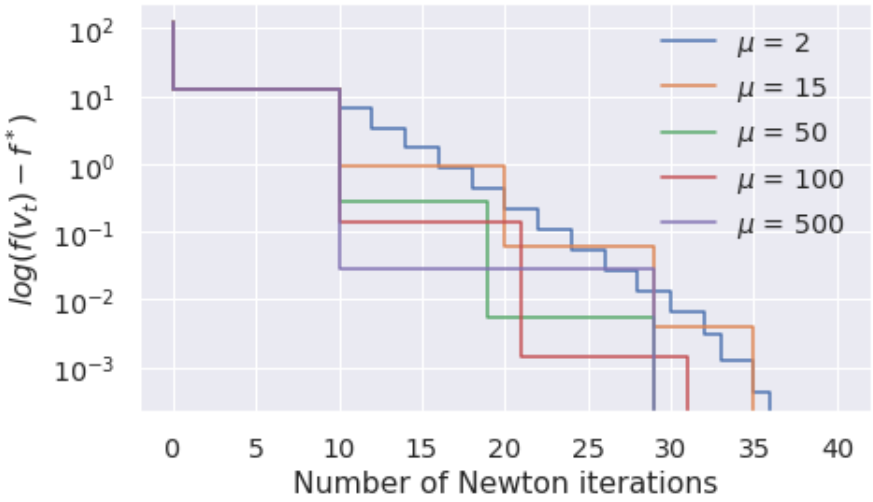

Figure 1: Plot showing the evolution of the gap between $\log(f(v_t) - f^*$ against the number of newton iterations for different μ

In the plot above we show the evolution of the gap between f^* and $f(v_t)$ against the number of iterations of Newton when we run the barrier algorithm. t gives the iteration number when we run the barrier algorithm.
v_t is the last vector we got from the barrier method and a measure of v^* with a precision ϵ and $f(v_t)$ is a measure of f^*

If μ is small, there is a small number of Newton steps per outer iteration (to satisfy the precision criteria in the barrier algorithm) and when μ is large, the algorithm will converge with a bigger number of newton iterations steps and a smaller number of outer iterations. A tradeoff for the number of outer iteration and the number of inner iteration can be found when μ is between 15 and 50 in the plot.

We choose n = 100, d = 15, λ = 10, t = 1, α= 0.1, β =0.5 and we generate a random matrix X from a uniform law between 0 and 1.

We choose $v_0 = 0_n$ and ϵ= 0.001.

Question 3

Because the problem satisfies the slater's conditions, the solutions of the primal and dual satisfies the KKT condition, this implies that:

$$\begin{aligned}\frac{\partial}{\partial u} L\left(u^*, w^*, \nu^*\right) = 0 &\iff u^* = \nu^* \\ &\iff X w^* - y = \nu^* \\ &\iff w^* = \left(X^T X\right)^{\perp} X^T \left(\nu^* + y\right)\end{aligned}$$

We finaly see in the plot below that μ hasn't any influence on w that we got from our algorithm. The plot also show that we get a good approximation of the optimal solution in overall.

A balance between the two is reached for $\mu = 50$, which corresponds to the green curve in Figure 1.

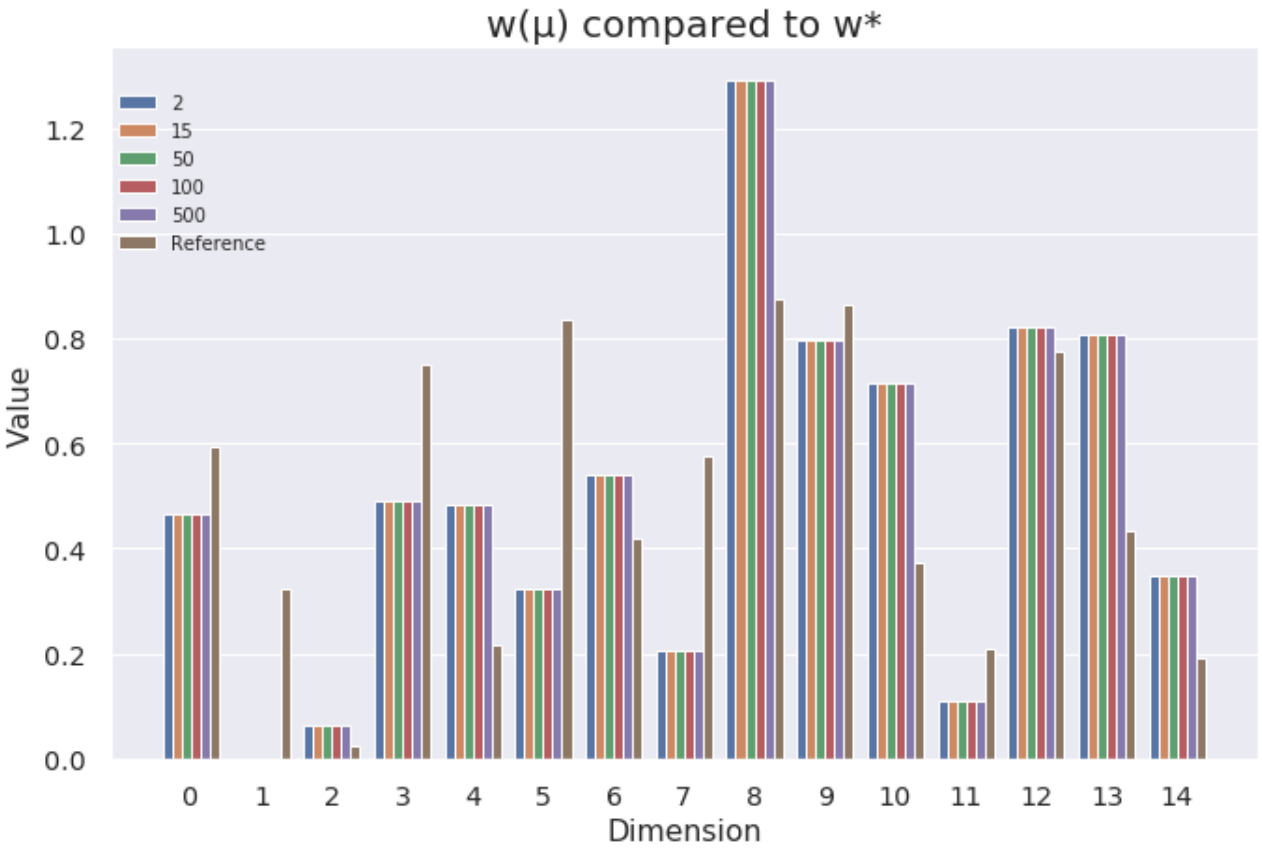

Figure 2: plot comparing the value of $w(\mu)$ for different μ and the optimal w against the dimension