Homework 3

Léo Alberge

November 12, 2018

Given $x_1, ..., x_n \in \mathbb{R}^d$ data vectors and $y_1, ..., y_n \in \mathbb{R}$ observations, we are searching for regression parameters $\omega \in \mathbb{R}^d$ which fit data inputs to observations y by minimizing their squared difference. In a high dimensional setting (when $n \leq\leq d$) a L1 norm penalty is often used on the regression coefficients ω in order to enforce sparsity of the solution (so that ω 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:

$$\underset{\omega}{\text{minimize}} \quad \frac{1}{2}||X\omega - y||_2^2 + \lambda||\omega||_1 \tag{LASSO}$$

Dual problem

Let's derive the form of the dual for this problem. By posing $a = X\omega - y$, (LASSO) is equivalent to:

$$\begin{aligned} \underset{a,\omega}{\text{minimize}} \quad & \frac{1}{2}||a||_2^2 + \lambda||\omega||_1 \\ \text{subject to} \quad & X\omega - y - a = 0 \end{aligned}$$

Note that since Slater conditions apply(There exist a, ω such that $X\omega - y = a$, simply choosing any ω and setting $a = X\omega - y$), strong duality holds. The lagrangian of this problem writes:

$$\mathcal{L}(a, w, v) = \frac{1}{2}||a||_2^2 + \lambda||\omega||_1 + v^T(X\omega - y - a) = ||a||_2^2 - v^Ta + \lambda(||\omega||_1 + \frac{v^TX\omega}{\lambda}) - v^Ty$$

Minimizing $\mathcal{L}(a, w, v)$ w.r.t. (a, w) yields in minimizing separately $\frac{1}{2}||a||_2^2 - v^Ta$ w.r.t. a and $||\omega||_1 + \frac{v^TX\omega}{\lambda}$ w.r.t. w. Yet $\inf_{\omega}(||\omega||_1 + \frac{v^TX\omega}{\lambda}) = -\sup_{\omega}(-\frac{(X^Tv)^T\omega}{\lambda} - ||\omega||_1) = -|| \circ ||_1^*(-\frac{X^Tv}{\lambda}) = \begin{cases} 0 \text{ if } \frac{||X^Tv||_\infty}{\lambda} \leq 1 \\ \text{else } -\infty \end{cases}$ where $*$ denotes the conjugate.(See HW2).
$\frac{1}{2}||a||_2^2 - v^Ta$ is a coercive differentiable convex function thus its minimum is achieved by vanishing the gradient and obtained for $a = v$ and the optimal objective is $-\frac{1}{2}v^Tv$.
Therefore the dual function writes:

$$g(v) = \begin{cases} -\frac{1}{2}v^Tv - v^ty \text{ if } \frac{||X^Tv||_\infty}{\lambda} \leq 1 \\ \text{else } -\infty \end{cases}$$

Therefore the dual yields:

$$\begin{aligned} \underset{v}{\text{maximize}} \quad & v^T(-\frac{1}{2}\mathbb{I})v - v^Ty \\ \text{subject to} \quad & \frac{||X^Tv||_\infty}{\lambda} \leq 1 \end{aligned} \tag{DUAL}$$

It can be formatted as a quadratic problem:

$$\begin{aligned} &\underset{v}{\text{minimize}} && v^T Q v + p^T v \\ &\text{subject to} && Av \preceq b \\ &\text{where} \quad Q = \frac{1}{2}\mathbb{I}, p = y, A = \begin{pmatrix} X^T \\ -X^T \end{pmatrix} \in \mathbf{R}^{2dxn}, &&\text{and } b = \lambda \mathbf{1}_{2d} \end{aligned} \tag{DUAL}$$

Implementation and Results

Generating data

In order to test the implementation, we generate a random design matrix X of size(d=1000, n=10) and a sparse random vector ω with 10 non-null random values. We then simulate y by $X\omega$. The goal is to recover ω. Details can be found in the function **generatedata**.

Centering step

For the centering step of the barrier method we implement the Newton method with a backtracking line search. Note that $v_0 = 0$ is always a feasible point of the problem.
The objective function yields:

$$\phi(t) = t * (v^T Q v + v^T p) - \sum_{i=1}^{m} -log(b_i - A_i v) \tag{Objective function}$$

One needs its gradient and hessian to implement the Newton method.

$$\begin{aligned} &\nabla\phi(t) = t * (2Qv + p) + A^T d \\ &\text{where } d_i = 1/(b_i - A_i v) \end{aligned} \tag{Gradient}$$

$$\nabla^2\phi(t) = t * (2Q) + A^T \text{diag}(d) A \tag{Hessian}$$

Details can be found in the function **centering step** and **backtracking linesearch**.
In the Figure 1, we can clearly observe the behavior of the newton method: linear then quadratic near the optimal value and the higher t is, the more difficult it is to solve the centering step.

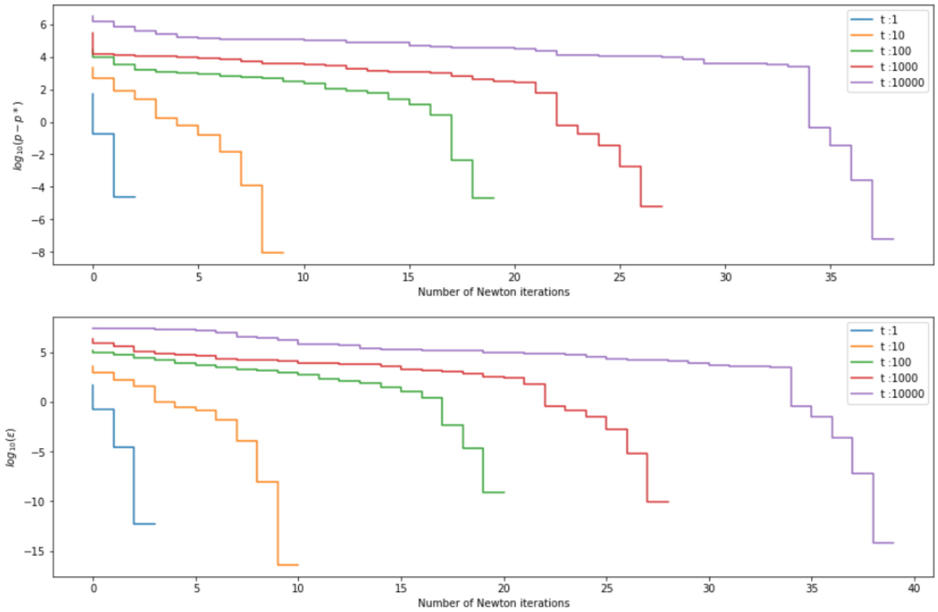

Figure 1: Centering step for different values of t, p^* is obtained for $\epsilon = 10e - 8$.The backtracking line search parameters are $\alpha = 0.25$, $\beta = 0.5$.

Barrier method

Details can be found in the function **barrier method**.
We observe in Figure 2 the influence of μ in terms of number of newton iterations: we see that $\mu = 2$ is not optimal because the convergence is too slow(nearly 100 newton iterations). For μ too big, there is few outer iterations but a lot of inner iterations and we see that we get extra precision($10e - 10$ for $\mu = 500$) which is costly and not necessary. $\mu = 20$ seems to be a fair solution.

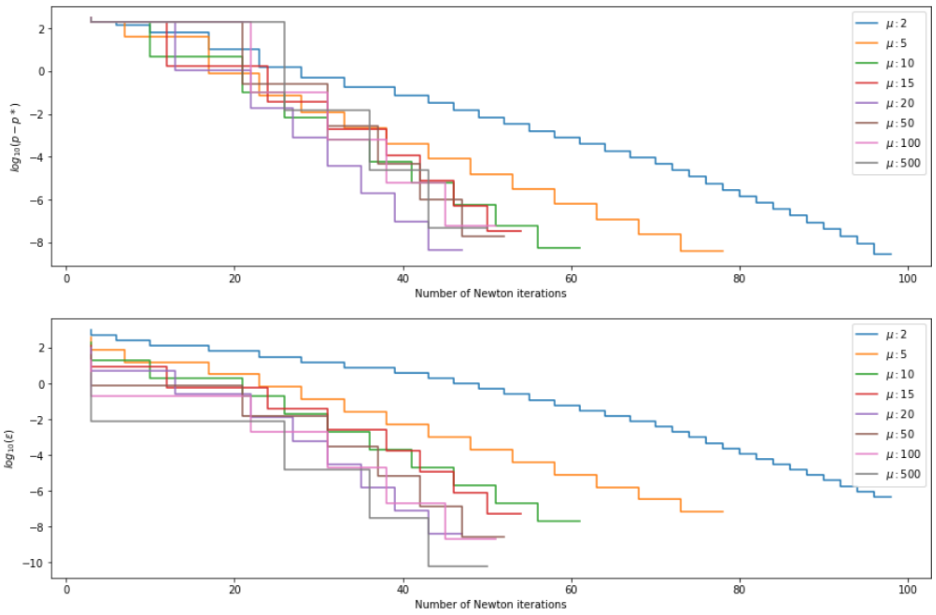

Figure 2: Barrier method step for different values of μ, p^* is obtained for $\epsilon = 10e - 6$, $t_0 = 1$.The Newton backtracking line search parameters are $\alpha = 0.25$, $\beta = 0.5$.

Primal solutions from dual We need to get the primal solution from the dual: one way to do it is to use dual points from central path.
We need to derive the dual of the (DUAL).
The Lagrangian yields:

$$\mathcal{L}(v, \mu) = v^T Q v + v^T p + \mu^T (Av - b)$$

For minimizing $\mathcal{L}(v, \mu)$ w.r.t. v, we vanish the gradient of this coercive convex function.
Therefore the dual function writes for $\mu \geq 0$:

$$g(\mu) = -\frac{1}{2}||p + A^T \mu||_2^2 - \mu^T b$$

Therefore the dual of (DUAL) writes:

$$\underset{\mu \geq 0}{\text{maximize}} \quad -\frac{1}{2}||y + (X, -X)\mu||_2^2 - \lambda \sum_i \mu_i$$

μ is of size (2d), let's note it by stacking to vector of size d μ_1, μ_2. We can rewrite the dual:

$$\underset{\mu_1 \geq 0, \mu_2 \geq 0}{\text{minimize}} \quad \frac{1}{2}||y - X(\mu_2 - \mu_1)||_2^2 + \lambda \sum_i \mu_i$$

And we recall the special form of the (LASSO) where μ_2, μ_1 denote respectively the positive part and the negative part of a real vector μ'.
Let v^* be a solution of (DUAL) then $\mu^* = -\frac{1}{f_i(v^*)}$ for $i = 1...2d$ where $f_i(v) = A_i v - b_i$ is a feasible point of the special LASSO and we get the solution of the prima $\mu_2^* - \mu_1^*$.
The coefficients of the solution of the primal are showed in Figure 3: we indeed get a very sparse solution with only few non null coefficients which was the goal of the LASSO.
We study the influence of μ on the primal solution ω but we can not assess anything. μ seems to have no influence on the primal solution(See Figure 4). We verify numerically that we indeed recover the objective of the primal.

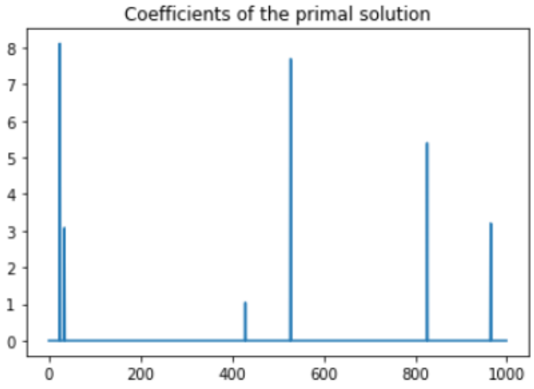

Figure 3: Coefficients of the primal solution

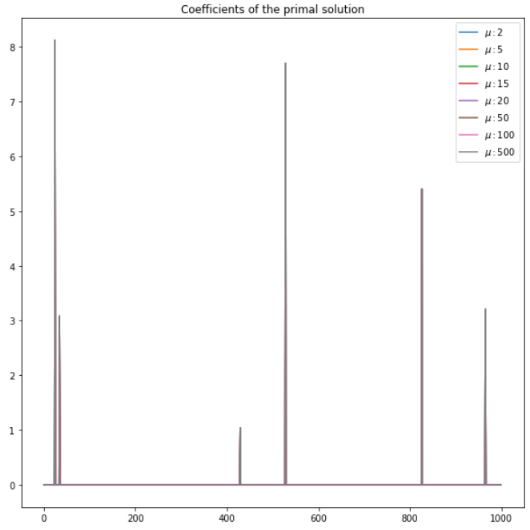

Figure 4: Coefficients of the primal solution for several values of μ