DE LORE Yonatan
Convex Optimization
Homework 3

1) * Let us derive the dual of the (LASSO) problem which is the following:

$$\min_{w \in \mathbb{R}^d} \frac{1}{2} \|Xw - y\|_2^2 + \lambda \|w\|_1 \qquad \text{(LASSO)}$$

where $X = (x_1, \dots, x_n)^T \in \mathbb{R}^{n \times d}$, $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^{+*}$ are known.

$$\text{(LASSO)} \iff \min_{w \in \mathbb{R}^d, z \in \mathbb{R}^n} \frac{1}{2}\|z\|^2 + \lambda \|w\|_1$$

$$\text{st.} \quad z = Xw - y$$

* The Lagrangian associated to this convex problem with linear equality constraints is the following:

$(w, z, \mu) \in \mathbb{R}^d \times \mathbb{R}^n \times \mathbb{R}^n$

$$\longrightarrow \mathcal{L}(w, z, \mu) = \frac{1}{2}\|z\|^2 + \lambda \|w\|_1 + \mu^T (Xw - y - z)$$

$$= \frac{1}{2}\|z\|^2 - \mu^T z + \lambda \|w\|_1 + \mu^T X w - \mu^T y$$

- $z \longmapsto \frac{1}{2}\|z\|^2 - \mu^T z$ is a convex function which reaches its minimum when $z - \mu = 0 \iff z = \mu$

Thus $\inf_z \frac{1}{2}\|z\|^2 - \mu^T z = \frac{1}{2}\|\mu\|^2 - \|\mu\|^2 = -\frac{1}{2}\|\mu\|^2$

- $\lambda \|w\|_1 + \mu^T X w = -\lambda \left(-\frac{1}{\lambda} \mu^T X w - \|w\|_1 \right)$

$$= -\lambda \left(-\frac{1}{\lambda} (X^T \mu)^T w - \|w\|_1 \right)$$

$(\lambda > 0)$

Thus, $\inf_{\omega} \lambda \|\omega\|_1 + \mu^T X \omega = - \sup_{\omega} \left[-\lambda \left(-\frac{1}{\lambda} (X^T \mu)^T \omega - \|\omega\|_1 \right) \right]$

$$= - f^* \left(-\frac{1}{\lambda} X^T \mu \right)$$

where $f^*(y) = \sup_x \; y^T x - \|x\|_1$ is the conjugate function of the norm-1.
We already computed this conjugate in the previous homework. But for reminder,

$$f^*(y) = \sup_x \left(\sum_{i=1}^d y_i x_i - \sum_{i=1}^d |x_i| \right) \qquad (x \in \mathbb{R}^d)$$

$$= \sup_x \left(\underbrace{\sum_{i=1}^d (y_i - \operatorname{sgn}(x_i)) x_i}_{:= \varphi_y(x)} \right)$$

- if there exists $j \in [\![1, d]\!]$ such that $y_j > 1$, we have considering the sequence $x(p) = \left((+p \, \delta_{ij})_{1 \le i \le m} \right) \; (p \in \mathbb{N})$:

$$\varphi_y(x^{(p)}) = p \underbrace{(y_j - 1)}_{> 0} \xrightarrow[p \to +\infty]{} +\infty$$

- similarly, if there exists j such that $y_j < -1$, considering the sequence $x(p) = \left((-p \, \delta_{ij})_{1 \le i \le m} \right)$, we have:

$$\varphi_y(x^{(p)}) = -p \underbrace{(y_j + 1)}_{< 0} \xrightarrow[p \to +\infty]{} +\infty$$

- finally, if for all $j \in [\![1, d]\!]$, $y_j \in [-1, 1]$,

$$|\varphi_y(x)| \le \sum_{i=1}^d |y_i x_i| - \sum_{i=1}^d |x_i|$$

$$= \sum_{i=1}^d \underbrace{|y_i|}_{\le 1} |x_i| - \sum_{i=1}^d |x_i|$$

$$\le 0$$

As $f_y(0) = 0$, we can conclude that

$$f^*(y) = \begin{cases} 0 & \text{if } |y_i| \leq 1 \ \forall i \in [\![1, d]\!] \\ +\infty & \text{otherwise.} \end{cases}$$

Hence $\inf\limits_{w} \lambda \|w\|_1 + \mu^T X w = \begin{cases} 0 & \text{if } |\frac{-1}{\lambda}(X^T\mu)_i| \leq 1 \ \forall i \in [\![1, d]\!] \\ -\infty & \text{otherwise} \end{cases}$

* Finaly we can conclude that:

$$\inf_{w, z} \mathcal{L}(w, z, \mu) = \begin{cases} -\frac{1}{2}\|\mu\|^2 - \mu^T y & \text{if } |(X^T\mu)_i| \leq \lambda \ \forall i \in [\![1, d]\!] \\ -\infty & \text{otherwise.} \end{cases}$$

The dual of the (LASSO) problem is therefore:

$$\begin{aligned} \max_{\mu \in \mathbb{R}^n} \ & -\frac{1}{2}\|\mu\|^2 - \mu^T y \\ \text{st.} \ & |(X^T\mu)_i| \leq \lambda \quad \forall i \in [\![1, d]\!] \end{aligned}$$

$$\Leftrightarrow \begin{aligned} \min_{\mu \in \mathbb{R}^n} \ & \frac{1}{2}\|\mu\|^2 + y^T \mu \\ \text{st.} \ & (X^T\mu)_i \leq \lambda \quad \forall i \in [\![1, d]\!] \\ & -(X^T\mu)_i \leq \lambda \quad \forall i \in [\![1, d]\!]. \end{aligned}$$

$$\Leftrightarrow \boxed{\begin{aligned} & \min_{v \in \mathbb{R}^n} \ v^T Q v + p^T v \qquad (QP) \\ & \text{st.} \ A v \leq b \\ & \text{where } Q = \tfrac{1}{2} I_n \quad \in \mathbb{R}^{n \times n} \\ & p = y \quad \in \mathbb{R}^n \\ & A = (X^T, -X^T) \quad \in \mathbb{R}^{2d \times n} \\ & b = \lambda \mathbb{1}_{2d} \quad \in \mathbb{R}^{2d} \end{aligned}}$$

where we denoted:

I_p the identity matrix of dimension p

$\mathbb{1}_p$ the vector of ones of size p

(U, V) the concatenation of the colums of U and the colums of V.

---

2) At each centering step of the barrier method, we are solving, for a certain t, the following unconstrained minimization problem:

$$\underset{v \in \mathbb{R}^n}{\text{Min}}\ t(v^T Q v + p^T v) - \sum_{i=1}^{2d} \log(b_i - (Av)_i)$$

As we will use the Newton method to solve it, we need to compute the gradient and the hessian of the objective function we will denote f_t.

$$f_t(v) = t(v^T Q v + p^T v) + \varphi(v)$$
$$\text{where} \quad \varphi(v) = -\sum_{i=1}^{2d} \log(b_i - (Av)_i)$$

$$\varphi(v) = -\sum_{i=1}^{2d} \log(b_i - \sum_{j=1}^{n} A_{ij} v_j)$$

$\star$ Thus, $$\frac{\partial \varphi}{\partial v_k} = -\sum_{i=1}^{2d} \frac{-A_{ik}}{b_i - \sum_j A_{ij} v_j} = \sum_{i=1}^{2d} \frac{A_{ik}}{b_i - (Av)_i}$$

$$= \sum_{i=1}^{2d} (A^T)_{ki} \frac{1}{b_i - (Av)_i}$$

Thus

$$\boxed{\nabla f_t = t(2Qv + p) + A^T \underline{inv}(b - Av)}$$

where $\underline{inv}(x)$ is the vector x to which was applied the inverse function component-wise.

* We also get :

$$\frac{\partial^2 \varphi}{\partial v_k \partial v_l} = \sum_{i=1}^{2d} A_{ik} \left(- \frac{-A_{il}}{(b_i - (Av)_i)^2} \right)$$
$$= \sum_{i=1}^{2d} (A^T)_{ki} \frac{1}{(b_i - (Av)_i)^2} A_{il}$$

Thus we have :

$$\boxed{\nabla_t^2 f = 2tQ + A^T \operatorname{Diag}\left(\left(\frac{1}{(b_i - (Av)_i)^2} \right)_{1 \le i \le 2d} \right) A}$$

where $\operatorname{Diag}(v)$ denotes the diagonal matrix whose diagonal is the vector v

Each Newton iteration will be therefore performed as :

$$v \longleftarrow v - \alpha(v, f_t) \left(\nabla^2 f_t(v) \right)^{-1} \nabla f_t(v)$$

where $\alpha(v, f_t)$ is the stepsize given by the backtracking line search

Experiments and comments

Below, I show results obtained for the following experiment :

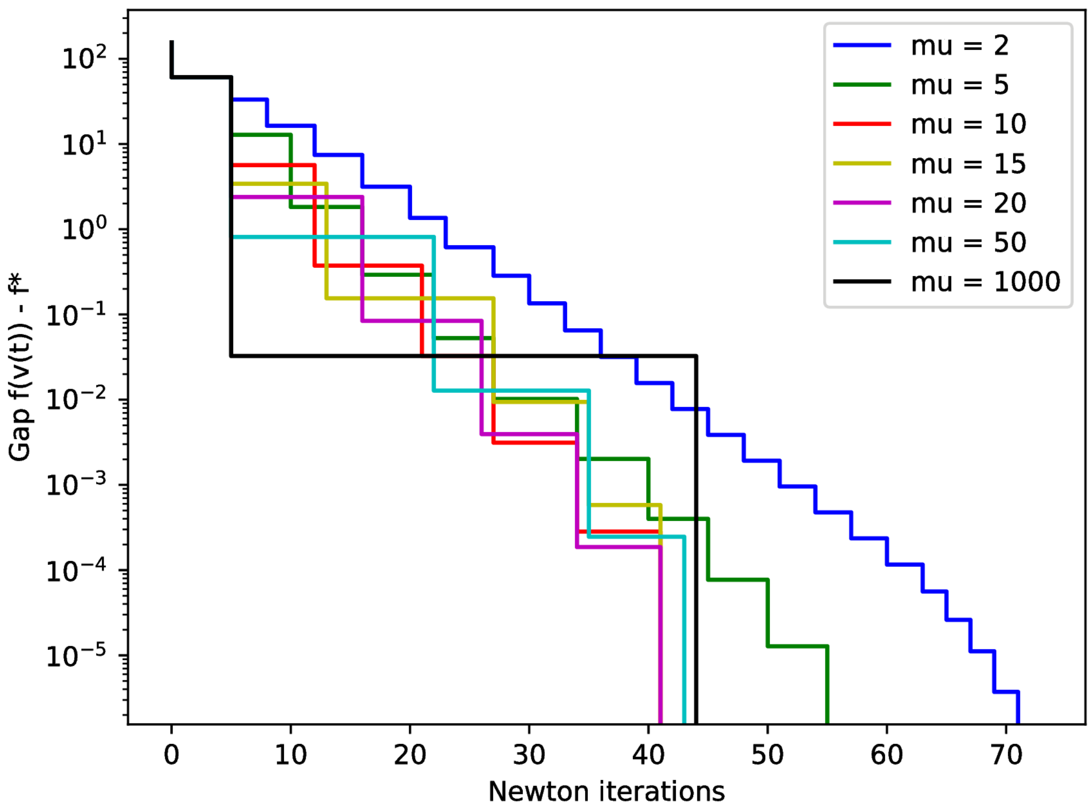
I generated a random matrix X of size (N=50,D=500) (values between -1 and 1), a vector of weights w of size D with 20 % of non-zero random entries (values between -1 and 1), and the vector of observations Y of size N by computing Xw.

I launched the barrier method to solve (QP) (for lambda=10) for various values of mu, on the same data (mu being the multiplicative factor by which we are increasing t after each centering step until convergence, t being the parameter of the barrier method introduced in page 4).
Precision criterion was set to 10^{-4}. Backtracking line search parameters were set to alpha=0.25 and beta=0.5, using the notations of the course.

On the figure below, I am plotting the gap f(v(t)) – f*, where f is the objective function of (QP) ($f=v^TQv + p^Tv$), where v(t) is the estimate of the dual solution at the end of the centering step of parameter t, and where f* is estimated by $f(v_{end})$ with v_{end} the last estimate (for the precision criterion stated above). The length of the steps correspond to the number of Newton iterations needed per centering step.

We can see first that **the larger mu, the larger the number of Newton iterations needed per centering step but the smaller the number of outer iterations (centering steps) required for the precision criterion.**

In addition, we see that **the total number of Newton iterations (ie. for all centering steps) is the smallest for mu between 10 and 20, range which can be seen as achieving a compromise between the number of inner Newton iterations and the number of centering steps.**



However I did not manage to interpret consistently the final subquestion relative to the impact on w.

From the strong duality (convex problem, linear equality constraints), we have that z*=v* (cf. question 1), where the dual variable v is first called mu) or equally said that Xw*= v* + y. Hence one could recover an estimate of the primal solution w* from the estimates of the dual solution v* using the pseudo inverse of X.
However, how to interpret the difference between the true w (the one we used to generate the data) and the ones estimated by the lasso, since the lambda parameter of the lasso, ie. the weight of the L1-regularization, is not necessarily adequate with the number of non-zero entries we chose when generating w ? Therefore, I am not sure about which information we could get (in addition to the first plot) from such an error curb, regarding the choice of mu.