

## Homework 3 — November 25

Instructor: Alexandre d'Aspremont

Student: Antoine Moulin

Given  $x_1, \dots, x_n \in \mathbb{R}^d$  data vectors and  $y_1, \dots, y_n \in \mathbb{R}$  observations, we are searching for regression parameters  $w \in \mathbb{R}^d$  which fit data inputs to observations  $y$  by minimizing their squared difference. In a high dimensional setting (when  $n \ll d$ ) a  $\ell_1$  norm 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)

$$\min_w \frac{1}{2} \|Xw - y\|_2^2 + \lambda \|w\|_1$$

in the variable  $w \in \mathbb{R}^d$ , where  $X = (x_1^T, \dots, x_n^T) \in \mathbb{R}^{n \times d}$ ,  $y = (y_1, \dots, y_n) \in \mathbb{R}^n$  and  $\lambda > 0$  is a regularization parameter.

1. Derive the dual problem of LASSO and format it as a general Quadratic Problem as follows:

$$\begin{aligned} \min v^T Q v + p^T v \\ \text{s.t. } A v \preceq b \end{aligned}$$

in variable  $v \in \mathbb{R}^n$ , where  $Q \succeq 0$ .

2. Implement the barrier method to solve QP.

- Write a function `v_seq = centering_step(Q, p, A, b, t, v0, eps)` which implements the Newton method to solve the centering step given the inputs  $(Q, p, A, b)$ , the barrier method parameter  $t$  (see lectures), initial variables  $v_0$  and a target precision  $\epsilon$ . The function outputs the sequence of variables iterates  $(v_i)_{i=1, \dots, n_\epsilon}$ , where  $n_\epsilon$  is the number of iterations to obtain the  $\epsilon$  precision. Use a backtracking line search with appropriate parameters.
- Write a function `v_seq = barr_method(Q, p, A, b, v0, eps)` which implements the barrier method to solve QP using precedent function given the data inputs  $(Q, p, A, b)$ , a feasible point  $v_0$ , a precision criterion  $\epsilon$ . The function outputs the sequence of variables iterates  $(v_i)_{i=1, \dots, n_\epsilon}$ , where  $n_\epsilon$  is the number of iterations to obtain the  $\epsilon$  precision.

3. Test your function on randomly generated matrices  $X$  and observations  $y$  with  $\lambda = 10$ . Plot precision criterion and gap  $f(v_t) - f^*$  in semilog scale (using the best value found for  $f$  as a surrogate for  $f^*$ ). Repeat for different values of the barrier method parameter  $\mu = 2, 15, 50, 100 \dots$  and check the impact on  $w$ . What would be an appropriate choice for  $\mu$ ?

**Solution.**

1. In the previous homework, one has shown that for  $u \in \mathbb{R}^n$ ,

$$\|\cdot\|_1^*(u) = \begin{cases} 0 & \text{if } \|u\|_\infty \leq 1 \\ +\infty & \text{otherwise} \end{cases}$$

The problem of LASSO is equivalent to

$$\begin{aligned} \min_w \frac{1}{2} \|z\|_2^2 + \lambda \|w\|_1 \\ \text{s.t. } z = Xw - y \end{aligned} \tag{1}$$

Let  $w \in \mathbb{R}^d, z \in \mathbb{R}^n, \nu \in \mathbb{R}^n$ . The Lagrangian function of the new problem is given by

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

and the dual function is

$$\begin{aligned}g(\nu) &= \inf_{w, z} \mathcal{L}(w, z, \nu) \\ &= y^T \nu + \inf_z \left( \frac{1}{2} \|z\|_2^2 + \nu^T z \right) + \inf_w \left( \lambda \|w\|_1 - (X^T \nu)^T w \right)\end{aligned}$$

The function  $h : z \mapsto \frac{1}{2} \|z\|_2^2 + \nu^T z$  is convex and differentiable. Its gradient is given by  $\nabla h(z) = z + \nu$  and we have  $\nabla h(z) = 0$  iff  $z = -\nu$ . Hence, the minimum of  $h$  is  $\frac{1}{2} \|\nu\|_2^2 - \|\nu\|_2^2 = -\frac{1}{2} \|\nu\|_2^2$ .

We can express the last term using the conjugate of  $\|\cdot\|_1$ :

$$\inf_w \lambda \|w\|_1 - (X^T \nu)^T w = \sup_w \left( \frac{1}{\lambda} X^T \nu \right)^T w - \|w\|_1 = \|\cdot\|_1^* \left( \frac{1}{\lambda} X^T \nu \right)$$

Thus,

$$g(\nu) = y^T \nu - \frac{1}{2} \|\nu\|_2^2 + \|\cdot\|_1^* \left( \frac{1}{\lambda} X^T \nu \right)$$

$\|\cdot\|_1^*$  is the indicator function of the  $\|\cdot\|_\infty$ -unit ball. Hence, the dual problem of LASSO is

$$\begin{aligned}\max_{\nu} & y^T \nu - \frac{1}{2} \|\nu\|_2^2 \\ \text{s.t.} & \left\| \frac{1}{\lambda} X^T \nu \right\|_\infty \leq 1\end{aligned}$$

Let's reformulate the constraint as an affine one:

$$\begin{aligned}\left\| \frac{1}{\lambda} X^T \nu \right\|_\infty \leq 1 & \text{ iff } \forall i \in \llbracket 1, n \rrbracket, -1 \leq \left[ \frac{1}{\lambda} X^T \nu \right]_i \leq 1 \\ & \text{ iff } \forall i \in \llbracket 1, n \rrbracket, \left[ \frac{1}{\lambda} X^T \nu \right]_i \leq 1 \text{ and } \left[ -\frac{1}{\lambda} X^T \nu \right]_i \leq 1 \\ & \text{ iff } A\nu \preceq \lambda \mathbf{1}_{2d}\end{aligned}$$

where  $A = \begin{pmatrix} X^T \\ -X^T \end{pmatrix}$ . Hence, multiplying the problem by  $-1$ , one has:

$$\begin{aligned}\min_{\nu} & \nu^T Q \nu + p^T \nu \quad \text{with } Q = \frac{1}{2} I_n, p = -y, b = \lambda \cdot \mathbf{1}_{2d} \\ \text{s.t.} & A\nu \preceq b\end{aligned}$$

2. The goal of the centering step is to solve the unconstrained problem:

$$\min_{\nu} g_t(\nu) \triangleq t (\nu^T Q \nu + p^T \nu) - \sum_{i=1}^{2d} \log(b_i - [A\nu]_i)$$

One can write:

$$A = \begin{pmatrix} -a_1^T & - \\ \vdots & \\ -a_n^T & - \end{pmatrix}$$

Let's compute the gradient and the Hessian matrix of the objective function:

$$\begin{aligned}\nabla g_t(\nu) &= t(2Q\nu + p) - \sum_{i=1}^{2d} \frac{-a_i}{b_i - a_i^T \nu} \\ &= t(2Q\nu + p) + \sum_{i=1}^{2d} (b_i - a_i^T \nu)^{-1} a_i\end{aligned}$$

Defining  $\phi = \left( (b_1 - a_1^T \nu)^{-1}, \dots, (b_{2d} - a_{2d}^T \nu)^{-1} \right)^T$ , we have

$$\boxed{\nabla g_t = t(2Q\nu + p) + A^T \phi}$$

Besides,

$$\nabla^2 g_t(\nu) = 2t \cdot Q + \sum_{i=1}^{2d} (b_i - a_i^T \nu)^{-2} a_i a_i^T$$

and

$$\boxed{\nabla^2 g_t = 2t \cdot Q + A^T \text{Diag}(\phi)^2 A}$$

For the code, see `HW3.ipynb`.

3. See `HW3.ipynb`.