Convex Optimization
DM3

Antonin Berthon

November 2018

Question 1

We consider the LASSO problem:

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

with $X = (x_1^T, ..., x_n^T) \in \mathbb{R}^{nxd}$, $Y = (y_1, ..., y_n) \in \mathbb{R}^n$ and $\lambda > 0$.

This optimization problem is equivalent to :

$$\begin{aligned} \min_{u,w} \quad & \frac{1}{2}\|u\|_2^2 + \lambda\|w\|_1 \\ \text{s.t.} \quad & Xw - y - u = 0 \end{aligned}$$

The Lagrangian writes, for $\nu \in \mathcal{R}^n$:

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

To study the function $g(\nu) = \inf_{u,w} \mathcal{L}(u, w, \nu)$, we need to find the conjugates of the functions $s : x \mapsto \frac{1}{2}\|x\|_2^2$ and $h : x \mapsto \lambda\|x\|_1$.

• The function $v : x \mapsto y^T x - \frac{1}{2}\|x\|_2^2$ is concave, so we can maximize it by setting to 0 its gradient : $\nabla v(x) = 0 \Longleftrightarrow x = y$. Therefore: $s^*(y) = \sup_x(y^T x - \frac{1}{2}\|x\|_2^2) = \frac{1}{2}\|y\|_2^2$.

• We know that for $a : x \mapsto \|x\|_1$, $a^*(y) = \begin{cases} 0 & \text{if } \|y\|_\infty \leq 1 \\ \infty & \text{otherwise} \end{cases}$. Since $\lambda > 0$, we have $h(x) = a(\lambda x)$ and $h^*(y) = \sup_x(y^T x - \lambda\|x\|_1) = \lambda \sup_x(\frac{1}{\lambda} y^T x - \|x\|_1) = \lambda a^*(\frac{1}{\lambda} y^T)$. Therefore, $h^*(y) = \begin{cases} 0 & \text{if } \|y\|_\infty \leq \lambda \\ \infty & \text{otherwise} \end{cases}$.

We can now write :

$$\begin{aligned} g(\nu) &= -\sup_{u,w}(-\mathcal{L}(u, w, \nu)) \\ &= -\Big(\sup_u(\nu^T u - \frac{1}{2}\|u\|_2^2) + \sup_w(-\nu^T Xw - \lambda\|w\|_1) + \nu^T y\Big) \\ &= -\Big(s^*(\nu) + h^*(-X^T\nu) + \nu^T y\Big) \end{aligned}$$

Therefore, the dual problem of (LASSO) writes:

$$\begin{aligned} \min_{\nu} \quad & \nu^T Q \nu + p^T \nu \\ \text{s.t.} \quad & Av \preceq b \end{aligned}$$

with:

- $Q = \frac{1}{2} I_n$
- $p = y$
- $A = \begin{pmatrix} X^T \\ -X^T \end{pmatrix}$
- $b = \begin{pmatrix} \lambda \\ \lambda \end{pmatrix}$

Question 2

In the following we call f the objective function of the dual: $f : x \mapsto x^T Q x + p^T x$.
To test the barrier method, we generate random matrices X and observations y with $\lambda = 10$. For a given precision criterion ϵ, the barrier method returns a list of points $(v_t)_{t^{(0)} \leq t \leq t_\epsilon}$ for the sequence of $t = (t^{(0)}, \mu t^{(0)}, \mu^2 t^{(0)}, \dots, \mu^{n_\epsilon} t^{(0)})$ where μ^{n_ϵ} is such that $\frac{m}{\mu^{n_\epsilon - 1} t^{(0)}} \geq \epsilon$ and $\frac{m}{\mu^{n_\epsilon} t^{(0)}} \leq \epsilon$.

Figure 1 shows the evolution of $f(v_t) - f^*$ through the Newton iterations of the barrier method by using $f(v_{t_\epsilon})$ as a surrogate for f^*. The length of each centering step can be seen with the length of each step. We see that when μ is small, the barrier method requires more outer iterations to attain the precision criterion, but each centering step is done in only a few Newton iterations. On the other hand, when μ is too large, the barrier method will converge in fewer outer iterations but each centering step will require more Newton iterations. An appropriate trade-off seems to be $\mu \approx 10 - 20$.

Retrieve w Since the problem satisfies the Slater conditions, the solutions of the primal u^*, w^* and of the dual ν^* satisfy the KKT condition :

$$\begin{aligned} \frac{\partial}{\partial u} \mathcal{L}(u^*, w^*, \nu^*) = 0 &\iff u^* = \nu^* \\ &\iff Xw^* - y = \nu^* \\ &\iff w^* = (X^T X)^{-1} X^T (\nu^* + y) \end{aligned}$$

assuming that $X^T X$ is invertible.
Figure 2 compares the retrieved weight vector w found by the barrier method for different μ, compared to a reference vector obtained by construction of X and y. We see that the parameter μ has no impact on w.

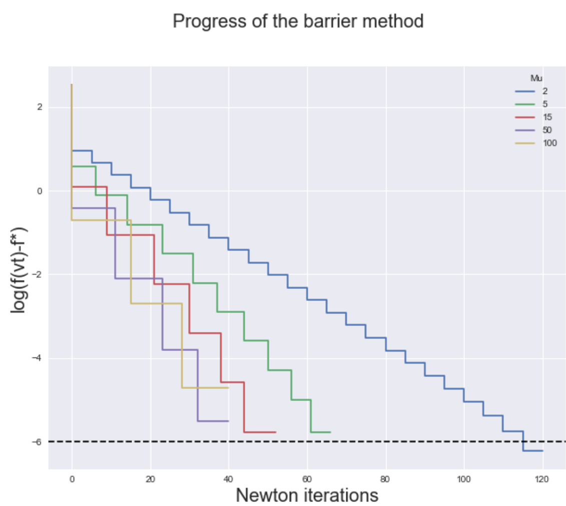

Figure 1: Evolution of the barrier method through the cumulative Newton iterations using $f(v_{t_\epsilon})$ as a surrogate for f^* and for multiple parameters μ, with $n = 200$, $m = 20$, $\epsilon = 10^{-6}$

Figure 2: Retrieved w^* from the dual solution ν^* obtained with the barrier method for different μ, compared to a reference w_{ref}. w_{ref} was obtained by generating y with $y = Xw_{ref} + \epsilon$ where ϵ follows a standard normal distribution. It is clear that the μ used in the barrier method do not impact the solution w^*. Also we see that the solution of the LASSO problem is a sparse approximation of the reference weight vector.