Convex optimization homework 3

Romain Petit

November 11, 2018

Question 1

(LASSO) is equivalent to :

$$\begin{array}{ll} \underset{w\in\mathbb{R}^d,\ z\in\mathbb{R}^n}{\text{minimize}} & \frac{1}{2}||z||_2^2 + \lambda||w||_1 \\ \text{subject to} & z = Xw - y \end{array} \tag{P}$$

The lagrangian of (P) is :

$$\begin{aligned} \mathcal{L}\colon \mathbb{R}^d \times \mathbb{R}^n \times \mathbb{R}^n &\to \mathbb{R} \\ w, z, \mu &\mapsto \frac{1}{2}||z||_2^2 + \lambda||w||_1 + \mu^T(Xw - y - z). \end{aligned}$$

The dual function is therefore :

$$\begin{aligned} g\colon \mathbb{R}^n &\to \mathbb{R} \\ \mu &\mapsto -\mu^T y + \inf_{w\in\mathbb{R}^d} \lambda||w||_1 + (X^T\mu)^T w + \inf_{z\in\mathbb{R}^n} \frac{1}{2}||z||_2^2 - \mu^T z \end{aligned}$$

$z \mapsto \frac{1}{2}||z||_2^2 - \mu^T z$ is a differentiable convex function, with a unique stationary point $z^* = \mu$, which is therefore its unique minimizer. Its minimum value is therefore $-\frac{1}{2}||\mu||_2^2$.

$\inf_{w\in\mathbb{R}^d} \lambda||w||_1 + (X^T\mu)^T w = -f^*(-X^T\mu)$ with $f : w \mapsto \lambda||w||_1$. Since the conjugate of the l_1 norm is the indicator of the l_∞ norm unit ball, we have $\inf_{w\in\mathbb{R}^d} \lambda||w||_1 + (X^T\mu)^T w$ equal to 0 if $||X^T\mu||_\infty \leq \lambda$, and $-\infty$ otherwise. Therefore $g(\mu) = \begin{cases} -\mu^T y - \frac{1}{2}||\mu||_2^2 & \text{if } ||X^T\mu||_\infty \leq \lambda \\ -\infty & \text{otherwise} \end{cases}$

We can finally write the dual problem :

$$\begin{array}{ll} \underset{\mu\in\mathbb{R}^n}{\text{minimize}} & \frac{1}{2}||\mu||_2^2 + \mu^T y \\ \text{subject to} & ||X^T\mu||_\infty \leq \lambda \end{array} \tag{Q}$$

that can be re-written as a general quadratic problem :

$$\boxed{\begin{array}{ll} \underset{v\in\mathbb{R}^n}{\text{minimize}} & v^T Q v + p^T v \\ \text{subject to} & Av \preceq b \end{array} \quad \text{with } Q = \frac{1}{2}I_n,\ p = y,\ A = \begin{pmatrix} X^T \\ -X^T \end{pmatrix}, \text{ and } b = \lambda \mathbf{1}_{2d}} \tag{QP}$$

Resolution of (QP) by the barrier method

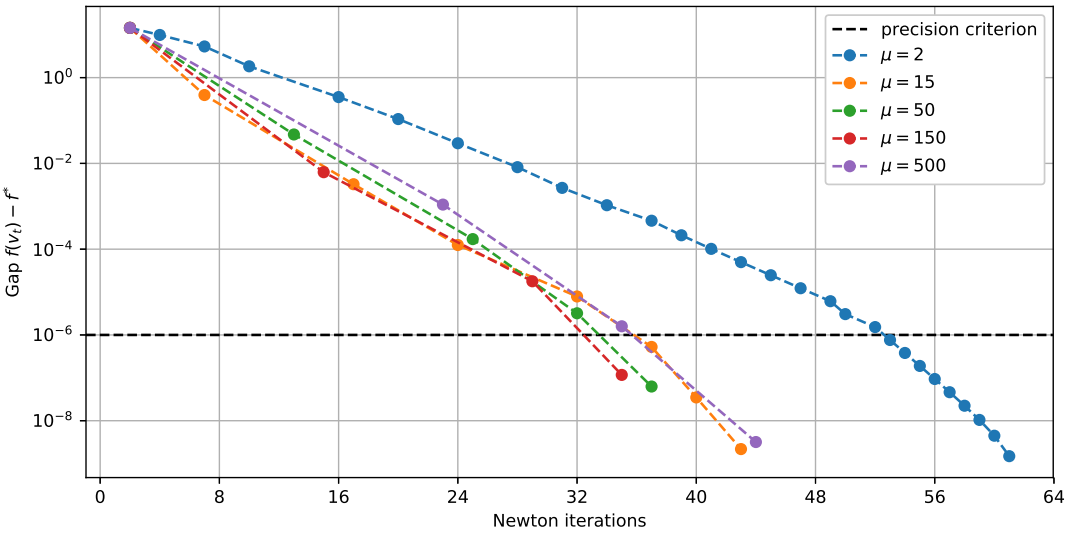

Figure 1: Decrase of the gap between the objective and the final objective

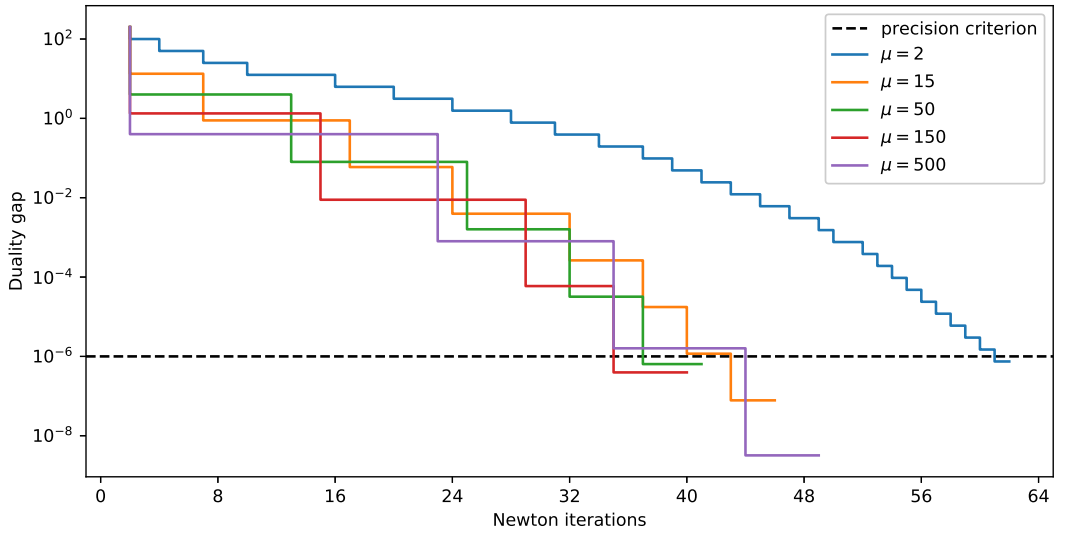

Figure 2: Decrease of the duality gap

To test our implementation, we drew a random data matrix $X \in \mathbb{R}^{n \times d}$ with $n = 10$ and $d = 1000$ ($n \ll d$ regime). We then drew a random sparse vector $w \in \mathbb{R}^d$ and a random normal noise vector $\eta \in \mathbb{R}^n$ with a variance of 10^{-4}, and set $y = Xw + \eta$.

We then ran the barrier method for several values of the barrier method parameter μ, with the following choice of hyper-parameters : $\lambda = 10$, $t_0 = 10$, $\epsilon = \epsilon_{\text{inner}} = 10^{-6}$, $\alpha = 0.01$ and $\beta = 0.5$.
Figure 1 shows the evolution of the gap between the objective and its final value, while figure 2 shows the evolution of the duality gap.

In figure 1, we can see that in some cases, the algorithm does not stop even if $f(v_t) - f(v_{\text{final}}) < \epsilon$. This is due to both the fact that we used the final objective value as a surrogate for the optimal value, and to the fact that $\frac{m}{t}$ only gives an upper bound (which is not necessarily tight) on $f(v_t) - f^*$.

In figure 2, we can however check that we have $\frac{m}{t} < \epsilon$ only for the last iteration, for which convergence is achieved.

Greater values of μ result in more inner iterations at each outer iteration, which can cause the duality gap to be smaller than the criterion by a large margin at the last iteration. A good value for μ seems to be 50 or 150, since this leads to a smaller number of total Newton iterations.

Computation of the LASSO solution

We will now see how solving (QP) by the barrier method allows to find an approximate solution of the LASSO.

The barrier method gives, at each iteration, a dual feasible point such that the duality gap associated to this point and the central point is $\frac{m}{t}$. Since strong duality holds, this dual feasible point is, after convergence, no more than $\frac{m}{t}$ sub-optimal.

Let us now explain how the dual of (QP) is related to the LASSO problem. Its lagrangian is :

$$\begin{aligned}\mathcal{L}\colon \mathbb{R}^n \times \mathbb{R}^d \times \mathbb{R}^d &\to \mathbb{R} \\ v, w_1, w_2 &\mapsto \frac{1}{2}||v||_2^2 + y^T v + w_1^T(X^T - \lambda \mathbf{1}_d) + w_2^T(-X^T - \lambda \mathbf{1}_d).\end{aligned}$$

$v \mapsto \frac{1}{2}||v||_2^2 - (y + Xw_1 - Xw_2)^T v$ is a differentiable convex function, with a unique stationary point, which is therefore its unique minimizer. Its minimum value is therefore $-\frac{1}{2}||y + Xw_1 - Xw_2||_2^2$. The dual function is therefore :

$$\begin{aligned}g\colon \mathbb{R}^d \times \mathbb{R}^d &\to \mathbb{R} \\ w_1, w_2 &\mapsto -\frac{1}{2}||y - X(w_2 - w_1)||_2^2 - \lambda \mathbf{1}^T(w_1 + w_2)\end{aligned}$$

Hence, the dual of (QP) is :

$$\begin{array}{ll} \underset{w_1 \in \mathbb{R}^d,\ w_2 \in \mathbb{R}^d}{\text{minimize}} & \frac{1}{2}||y - X(w_2 - w_1)||_2^2 + \lambda \mathbf{1}^T(w_1 + w_2) \\ \text{subject to} & w_1, w_2 \succeq 0 \end{array} \qquad \text{(QP dual)}$$

We can therefore see than (QP dual) is equivalent to the LASSO problem, and that for each feasible (w_1, w_2) of (QP dual), there is a vector $w = w_2 - w_1$ that yields the same LASSO objective.

In our case, the point v^* returned by the barrier method allows to get the following approximate solution of (QP dual) : $\begin{pmatrix} w_1^* \\ w_2^* \end{pmatrix}_i = -\frac{1}{t(Av^* - b)_i}$. The relation $w^* = w_2^* - w_1^*$ finally allows to recover a solution of the LASSO problem which is not more than ϵ sub-optimal.

Computing w^* for different values of μ allows to verify that the choice of μ does not have any significant impact on w^*.

	ground truth weights	optimal weights
LASSO objective	51.23	25.46
$\|\|w\|\|_1$	5.12	1.46
$\|\|Xw - y\|\|_2$	0.03	4.65
weights' support size	10	4

Table 1: Comparison of ground truth and optimal weights

Table 1 shows that optimal weights are more sparse than ground truth weights. This is due to the fact that we chose a high value of $\lambda = 10$. This therefore yields a smaller size of w's support (and a smaller l_1 norm), but the data fitting term has a greater value than for ground truth weights.