BARILLER Halvard
halvard.bariller@ens-paris-saclay.fr

Homework # 3
Convex Optimization

11/20/23

1 LASSO: General context

Given $x_1, \dots, x_n \in \mathbb{R}^d$ data vectors and their associated observations $y_1, \dots, y_n \in \mathbb{R}$, we consider the following optimization problem in $w \in \mathbb{R}^d$:

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

where $\lambda > 0$ is a regularization parameter, $X \in \mathbb{R}^{n \times d}$ is the design matrix and $y \in \mathbb{R}^n$ is the vector of associated 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 ℓ_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).

2 Dual Problem

To solve this non-differentiable optimization problem, we reformulate it by deriving its dual, which takes a general Quadratic Program form.

We begin by making the change of variable $z = Xw - y$ and we consider the equivalent problem:

$$\min_{w,z} f(w,z) = \frac{1}{2}\|z\|_2^2 + \lambda \|w\|_1 \quad \text{subject to} \quad z = Xw - y \tag{$LASSO_{eq}$}$$

Deriving from its standard form, we can write the associated Lagrangian:

$$\mathcal{L}(w, z, \alpha) = \frac{1}{2}\|z\|_2^2 + \lambda \|w\|_1 + \alpha^T(Xw - y - z)$$

where $\alpha \in \mathbb{R}^n$ is the vector of Lagrange multipliers.
The dual function is then given by:

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

.

Let $\psi : x \mapsto \frac{1}{2}\|x\|_2^2 - \alpha^T x$. The function ψ is a positive quadratic form, hence strongly convex and $\mathcal{C}^2$-differentiable. There exists a unique minimizer z^* of ψ given by the first order condition:

$$\nabla \psi(z^*) = z^* - \alpha = 0 \quad \iff \quad z^* = \alpha.$$

As a consequence, we have:

$$\inf_z \left(\frac{1}{2}\|z\|_2^2 - \alpha^T z \right) = \psi(z^*) = \psi(\alpha) = \frac{1}{2}\|\alpha\|_2^2 - \alpha^T \alpha = -\frac{1}{2}\|\alpha\|_2^2.$$

The dual function then becomes:

$$\begin{aligned} g(\alpha) &= -\frac{1}{2}\|\alpha\|_2^2 + \inf_w \left(\lambda \|w\|_1 + \alpha^T Xw \right) - \alpha^T y \\ &= -\frac{1}{2}\|\alpha\|_2^2 - \alpha^T y - \sup_w \left(-\alpha^T Xw - \lambda \|w\|_1 \right) \end{aligned}$$

We recognize in the last term the convex conjugate function $h^*(-X^T\alpha) = \sup_w \left(-\alpha^T X w - h(w)\right)$, with $h(w) = \lambda\|w\|_1$. This gives:

$$g(\alpha) = -\frac{1}{2}\|\alpha\|_2^2 - \alpha^T y - h^*(-X^T\alpha).$$

As developed in the lectures and in the previous homework, the convex conjugate of h is given by its dual norm:

$$h^*(u) = \sup_w \left(u^T w - h(w)\right) = \begin{cases} 0 & \text{if} \quad \|u\|_\infty \leq \lambda \\ +\infty & \text{otherwise} \end{cases}$$

Therefore, we have:

$$g(\alpha) = -\frac{1}{2}\|\alpha\|_2^2 - \alpha^T y - h^*(-X^T\alpha) = \begin{cases} -\frac{1}{2}\|\alpha\|_2^2 - \alpha^T y & \text{if} \quad \|X^T\alpha\|_\infty \leq \lambda \\ -\infty & \text{otherwise} \end{cases}$$

The dual problem is then given by:

$$\begin{aligned} \max_\alpha g(\alpha) &= \max_\alpha \left(-\frac{1}{2}\|\alpha\|_2^2 - \alpha^T y\right) \quad \text{subject to} \quad \|X^T\alpha\|_\infty \leq \lambda \qquad (LASSO_{dual}) \\ &= \min_\alpha \frac{1}{2}\alpha^T I_n \alpha + \alpha^T y \quad \text{subject to} \quad \|X^T\alpha\|_\infty \leq \lambda \end{aligned}$$

Finally, by definition of the ℓ_∞-norm, we can write the dual problem with the following Quadratic Program form:

$$\min_{v\in\mathbb{R}^n} v^T Q v + p^T v \quad \text{subject to} \quad Av \preceq b \qquad (QP)$$

where $Q = \frac{1}{2}I_n$, $p = y$, $A = [X^T, -X^T] \in \mathbb{R}^{2d\times n}$ and $b = \lambda\mathbf{1}_{2n}$.

(Remark: we denote by [,] the concatenation.)

3 Barrier method

Reflecting on the (QP) optimization problem, we will proceed to solve it using the barrier method with logarithmic barrier.
Let $t > 0$ be a parameter. We consider the following optimization problem:

$$\min_{v\in\mathbb{R}^n} \quad f_0(v) + \frac{1}{t}\phi(v) = v^T Q v + p^T v - \frac{1}{t}\sum_{i=1}^{2n}\log(b_i - a_i^T v)$$

where ϕ is the (convex) logarithmic barrier s.t. $\mathbf{dom}\ \phi = \{v | Av \prec b\}$, and we denote by $a_1^T, \dots, a_{2n}^T$ the rows of the matrix A.
The barrier method problem is defined as follows:

$$\min_v t f_0(v) + \phi(v) = t(v^T Q v + p^T v) - \sum_{i=1}^{2n}\log(b_i - a_i^T v)$$

Algorithm 1 Barrier method

Require: v (strictly feasible point), $t := t^{(0)} > 0, \mu > 1, \epsilon > 0.$
Ensure:
  while $2n/t \geq \epsilon$ **do**
    $v^*(t) \leftarrow \arg\min\ t f_0(v) + \phi(v)$ ▷ *Centering step*
    $v \leftarrow v^*(t)$
    $t \leftarrow \mu t$
  end while

3.1 Centering step

The centering step is equivalent to solving the following problem:

$$\min_v F(v) = tf_0(v) + \phi(v)$$

with v a given strictly feasible point. To do so, we will use Newton method. We first begin by deriving the gradient and the Hessian of F:

$$\nabla F(v) = (2Qv + p)t + \sum_{i=1}^{2n} \frac{a_i}{b_i - a_i^T v}$$

$$\nabla^2 F(v) = 2Qt + \sum_{i=1}^{2n} \frac{a_i a_i^T}{(b_i - a_i^T v)^2}$$

Algorithm 2 centering_step (Centering step using Newton method)

Require: $Q, p, A, b, t > 0, v_0$ (starting feasible point), $\epsilon > 0, \alpha, \beta$.
Ensure: $(v_i)_{i \geq 0}$

```
λ ← 1
i ← 0
while λ²/2 ≥ ϵ do
    Δv_int ← −∇²F⁻¹(v_i)∇F(v_i)                    ▷ Newton step
    λ² ← ∇F(v_i)ᵀ∇²F(v_i)⁻¹∇F(v_i)                 ▷ Decrement
    s ← backtrack(Δv_int, v_i, α, β)                ▷ Line search
    v_{i+1} ← v_i + sΔv_int                         ▷ Update
    i ← i + 1
end while
```

Algorithm 3 backtrack (Backtracking line search)

Require: Δv (descent direction at v for f), $v, \alpha \in (0, 0.5), \beta \in (0, 1)$.
Ensure: s

```
s ← 1
while f(v + sΔv) > f(v) + αs∇f(v)ᵀΔv do
    s ← βs
end while
```

3.2 Numerical results

We tested the barrier algorithm on randomly generated matrices X and observations y.
The results presented below have been obtained with the following parameter values:

$$\begin{cases} \lambda = 10 \\ \alpha = 0.1 \\ \beta = 0.7 \\ n = 200 \\ d = 200 \\ \epsilon = 10^{-6}, \\ t = 0.2, \\ \mu \in \{2, 5, 10, 15, 30, 50, 100, 500\} \end{cases}$$

The convergence has been computed using the last value obtained with the algorithm as surrogate for f^*.

As expected, we can observe the impact of μ very distinctly: a smaller value of μ lead to few inner iterations (Newton steps in each centering step, c.f. figure 3) but requires more outer iterations (c.f. figure 2).
In figure 1, we visualize the convergence as a function of the total Newton steps that have been required for each value of μ. Once again, we can observe wider constant stages as μ grows (representing the number of Newton steps required in each centering step). We note that the minimum number of steps required is obtained for $\mu = 50$ and $\mu = 500$ (c.f. figures 1 and 4).

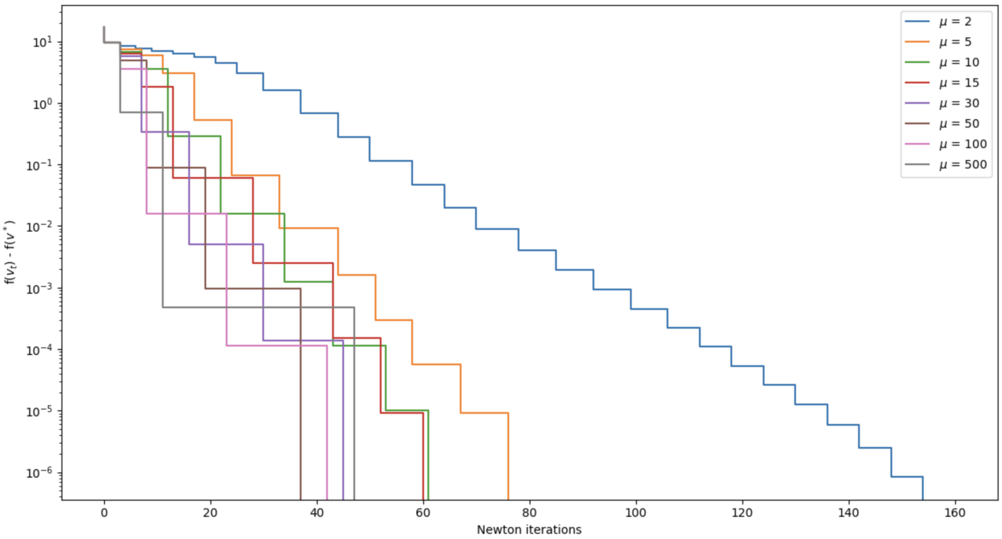

Figure 1: Convergence of the objective function over the totality of Newton steps

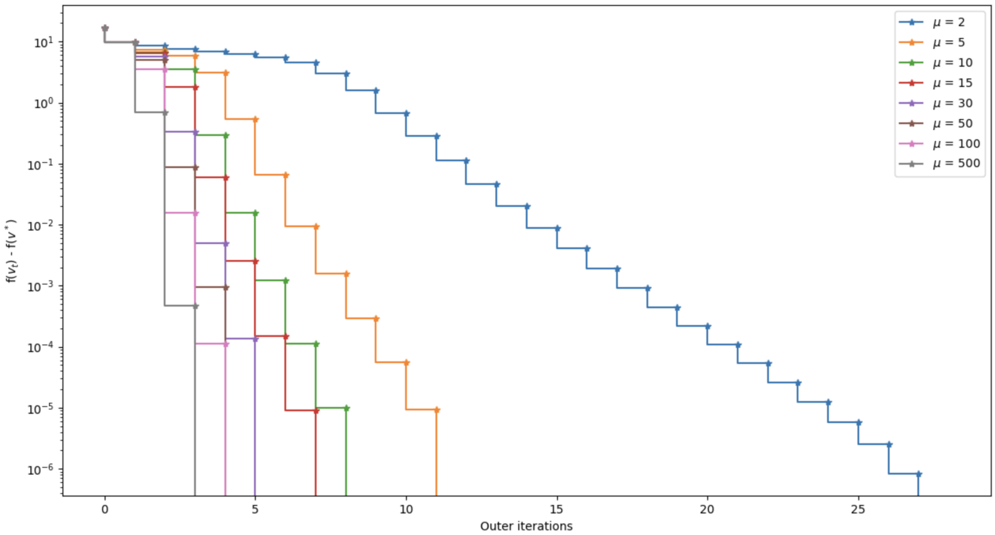

Figure 2: Convergence of the objective function through outer iterations

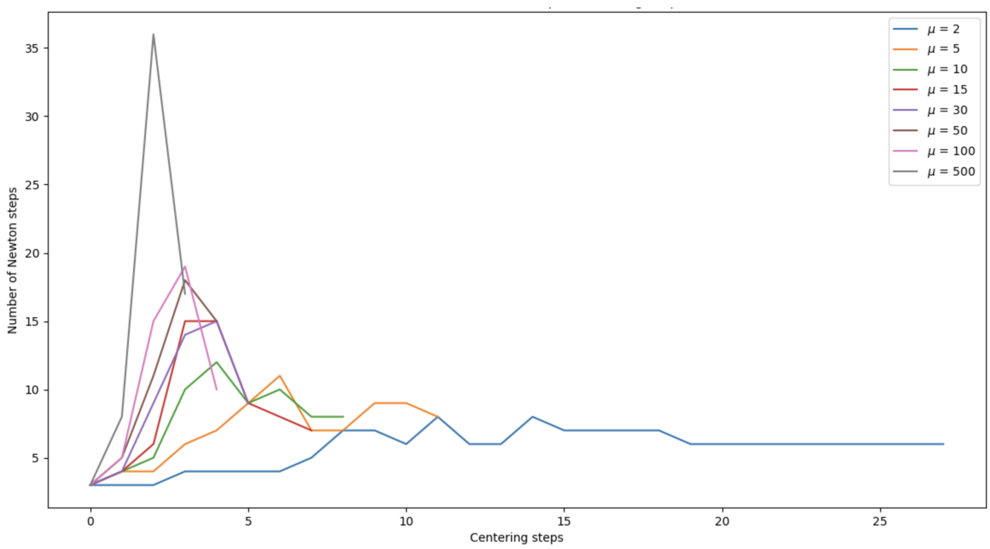

Figure 3: Number of inner iterations per centering step

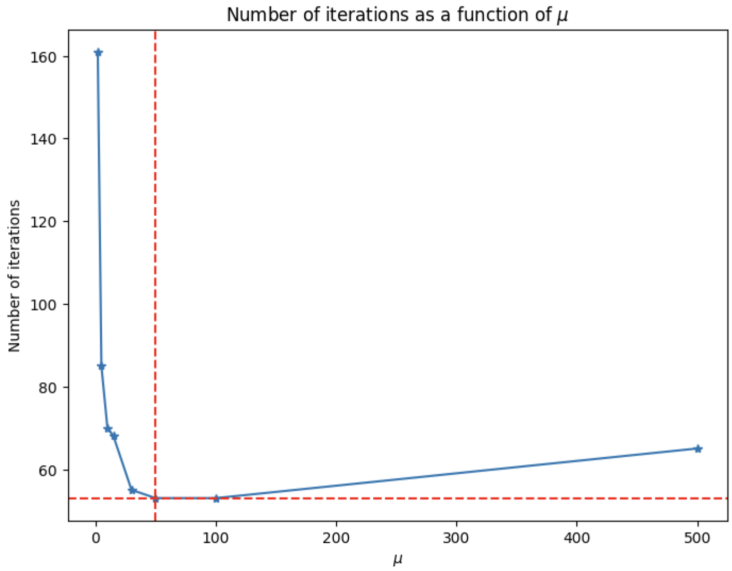

Figure 4: Overall number of Newton steps for μ

As aforementioned, the total number of Newton steps depends on the value chosen for μ. In this example, it seems that choosing $\mu = 50$ will yield the fastest convergence.

The objective being convex and the constraints being affines, we can apply results about strong duality and about the generalized K.K.T. conditions for non-differentiable functions [1, 2] to derive a necessary condition of optimality for $(LASSO_{eq})$. Solving the nullity of the Lagrangian in z, we have:

$$\begin{aligned}\partial_z \mathcal{L}(w, z, \alpha^*) = 0 &\iff \partial_z \left(\frac{1}{2}\|z\|_2^2 + \lambda\|w\|_1 + {\alpha^*}^T(Xw - y - z) \right) = 0 \\ &\iff z - \alpha^* = 0 \\ &\iff z = \alpha^* \\ &\iff Xw - y = \alpha^* \end{aligned}$$

.

We therefore have that any $(LASSO)$ solution will be such that $w = X^+(\alpha^* + y), \quad X^+$ being the Moore-Penrose inverse of the matrix X and $\alpha^* = v^*$.

Using the collection of optimal values $(v_\mu^*)_\mu$ computed using the barrier method, we compared the associated $(w_\mu^*)_\mu$.

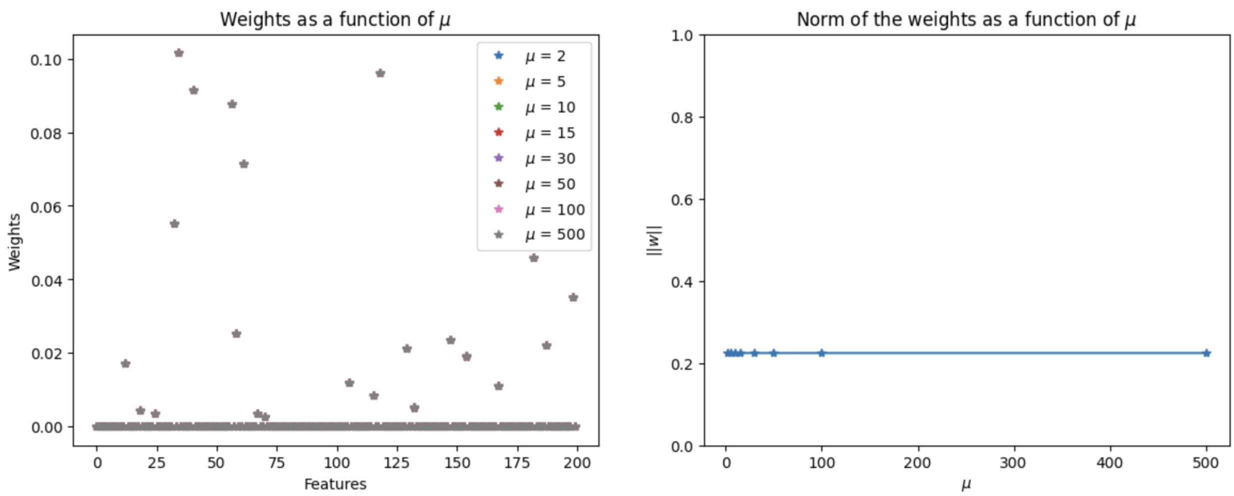

Figure 5: Comparison of the $(w_\mu^*)_\mu$

We first note that the $(w_\mu^*)_\mu$ are unchanged by the value of μ: the element-wise coordinates on the left as well as the l_2-norm on the right do not change. This result is what we expected since the dual optimum is supposed

to be the same regardless of μ.
Finally, we observe that the obtained primal solution is indeed sparse as one could expect in a LASSO problem.

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

[2] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015.