Derivation of the proximal operator relative to the lasso, ridge, group lasso and nuclear norm penalty

Brice Ozenne

March 2, 2017

Contents

1 Subgradient

For non-differentiable convex functions, z sub-gradient of f at x :

$$f(x') \geq f(x) + < z, x' - x > \forall x'' \in \mathbb{R}^p$$

subdifferential of f at x: $\partial f(x) = \{z\}$

1.1 L1 norm

When x is not 0 then $\partial|x| = sign(x) = \pm 1$. For $|h| < 1$ and all $x \in \mathbb{R}$ we have:

$$f(x) = |x| \geq hx = f(0) + h(x - 0)$$

By definition of the subgradient $\partial f(0) = [-1; 1]$

1.2 Euclidean norm

When x is not 0 then: $\partial ||x||_2 = \partial \sqrt{\sum_i x_i^2} = \frac{2(x_1,\ldots,x_n)}{2\sqrt{\sum_i x_i^2}} = \frac{x}{||x||_2}$.

For $||h||_2 < 1$ and all $x \in \mathbb{R}$ we have:

$$f(x) = ||x||_2 \geq ||h||_2 ||x||_2 \geq h^\intercal x = f(0) + h^\intercal(x - 0)$$

By definition of the subgradient $\partial f(0) = \{h | ||h||_2 \leq 1\}$

1.3 Nuclear norm

Denoting the SVD decomposition of x:

$$x = U\Sigma V^\intercal$$

$$\begin{aligned}
||x||_* &= tr(\sqrt{x^\intercal x}) = tr(\sqrt{(U\Sigma V^\intercal)^\intercal (U\Sigma V^\intercal)}) \\
&= tr(\sqrt{V\Sigma U^\intercal U\Sigma V^\intercal}) \\
&= tr(\sqrt{V\Sigma^2 V^\intercal}) \text{ (circularity of the trace)} \\
&= tr(\sqrt{V^\intercal V\Sigma^2}) \\
&= tr(\sqrt{\Sigma^2}) \\
&= tr(|\Sigma|_1)
\end{aligned}$$

We are interested in the derivative of the functionnal $F(x) = tr(|\Sigma(x)|_1)$. According to Watson (1992) (theorem 2) the subdifferential of this functional is:

$$\begin{aligned}
\partial F(x) &= U diag(\partial(tr(|\Sigma|_1)))V^\intercal \\
&= U diag(\sum_j \partial(|\sigma_j|_1))V^\intercal
\end{aligned}$$

2 Proximal operator

$$prox_{\tau g} : \mathbb{R}^p \rightarrow \mathbb{R}^p$$
$$\theta \mapsto \underset{x}{\operatorname{argmin}} \left(g(x) + \frac{1}{2\tau}||x-\theta||_2^2 \right)$$

2.1 One penalty

2.1.1 Lasso

The lasso penalty is $\mathcal{P}_1(\theta) = \lambda_1||\theta||_1$.
The subdifferential of the L1 norm at x is $\partial||x||_1 = s_1(x)$ with:

$$s_1(x) = \begin{cases} -1 \text{ if } x < 0 \\ [-1;1] \text{ if } x = 0 \\ 1 \text{ if } x > 0 \end{cases}$$

The proximal operator can be computed solving:

$$\begin{aligned} prox_{\tau\mathcal{P}_1}(\theta) &= \underset{x}{\operatorname{argmin}} \left(\lambda_1||x||_1 + \frac{1}{2\tau}||x-\theta||_2^2 \right) \\ \partial_x \left(\lambda_1||x||_1 + \frac{1}{2\tau}||x-\theta||_2^2 \right) &= 0 \\ &= \lambda_1 s_1(x) + \frac{1}{\tau}(x-\theta) \\ x &= \theta - \lambda_1\tau s_1(x) \end{aligned}$$

$$x = \begin{cases} \theta - \lambda_1\tau & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ \theta + \lambda_1\tau & \text{if } x < 0 \end{cases} = \begin{cases} \theta - \lambda_1\tau & \text{if } \theta \geq \lambda_1\tau \\ 0 & \text{if } |\theta| \leq \lambda_1\tau \\ \theta + \lambda_1\tau & \text{if } \theta \leq -\lambda_1\tau \end{cases}$$

$$\boxed{prox_{\tau\mathcal{P}_1}(\theta) = sign(\theta)(|\theta| - \lambda_1\tau)^+}$$

2.1.2 Ridge

The ridge penalty is $\mathcal{P}_2(\theta) = \frac{\lambda_2}{2}||\theta||_2^2$.

The proximal operator can be computed solving:

$$\begin{aligned} prox_{\tau\mathcal{P}_2}(\theta) &= \underset{x}{\operatorname{argmin}} \left(\frac{\lambda_2}{2}||x||_2^2 + \frac{1}{2\tau}||x-\theta||_2^2 \right) \\ \partial_x \left(\frac{\lambda_2}{2}||x||_2^2 + \frac{1}{2\tau}||x-\theta||_2^2 \right) &= 0 \\ &= \lambda_2 x + \frac{1}{\tau}(x-\theta) \\ x &= \frac{1}{1+\tau\lambda_2}\theta \end{aligned}$$

$$\boxed{prox_{\tau\mathcal{P}_2}(\theta) = \frac{1}{1+\tau\lambda_2}\theta}$$

2.1.3 Group Lasso

The group lasso penalty is $\mathcal{P}_G(\theta) = \lambda_G ||\theta||_2$ where θ is a vector.
The subdifferential of the L2 norm at x is $\partial ||x||_G = s_G(x)$ with:

$$s_G(x)_j = \left\{ \begin{array}{l} \frac{x_j}{||x||_2} \text{ if } ||x||_2 > 0 \\ h; ||h||_2 < 1 \text{ if } ||x||_2 = 0 \end{array} \right.$$

The proximal operator can be computed solving:

$$\begin{aligned}
prox_{\tau \mathcal{P}_G}(\theta) &= \underset{x}{\operatorname{argmin}} \left(\lambda_G ||x||_2 + \frac{1}{2\tau} ||x - \theta||_2^2 \right) \\
\partial_{x_j} \left(\lambda_G ||x||_2 + \frac{1}{2\tau} ||x - \theta||_2^2 \right) &= 0 \\
&= \lambda_G s_G(x)_j + \frac{1}{\tau}(x_j - \theta_j) \\
x_j &= \theta_j - \lambda_G \tau s_G(x)_j
\end{aligned}$$

$$x_j = \begin{cases} \theta_j - \lambda_G \tau \frac{\theta_j}{||\theta||_2} & \text{if } ||x||_2 > 0 \\ 0 & \text{if } ||x||_2 = 0 \end{cases} = \begin{cases} \theta_j - \lambda_G \tau \frac{\theta_j}{||\theta||_2} & \text{if } ||\theta||_2 \geq \lambda_G \tau \\ 0 & \text{if } ||\theta||_2 \leq \lambda_G \tau \end{cases}$$

$$\boxed{prox_{\tau \mathcal{P}_G}(\theta) = \theta \left(1 - \frac{\lambda_G \tau}{||\theta||_2} \right)^+}$$

2.1.4 Nuclear norm

The nuclear penalty is $\mathcal{P}_G(\theta) = \lambda_N ||\theta||_N = tr(\sqrt{\theta^\intercal \theta})$ where θ is a matrix.
The subdifferential of the nuclear norm in x is $\partial ||x||_N = s_N(x)$ with

$$s_N(x) = U_x diag(s_1(\sigma_x)) V_x{}^\intercal$$

where $x = U_x diag(\sigma_x) V_x$

The proximal operator can be computed solving:

$$\begin{aligned}
prox_{\tau \mathcal{P}_N}(\theta) &= \underset{x}{\operatorname{argmin}} \left(\lambda_N ||x||_N + \frac{1}{2\tau} ||x - \theta||_2^2 \right) \\
\partial_x \left(\lambda_N ||x||_N + \frac{1}{2\tau} ||x - \theta||_2^2 \right) &= 0 \\
&= \lambda_N s_N(x) + \frac{1}{\tau}(x - \theta) \\
x &= \theta - \lambda_G \tau s_N(x) \\
U_x \Sigma_x V_x{}^\intercal &= U \Sigma_\theta V^\intercal - \lambda_G \tau U_x diag(s_1(\sigma_x)) V_x{}^\intercal \\
U_x \left(\Sigma_x + \lambda_G \tau diag(s_1(\sigma_x)) \right) V_x{}^\intercal &= U \Sigma_\theta V^\intercal
\end{aligned}$$

Therefore

$$\Sigma_x + \lambda_G \tau diag(s_1(\sigma_x)) = \Sigma_\theta$$

with $U_x = U$ and $V_x = V$ is a valid solution. See appendix for more on the unicity of the solution.

So

$$\sigma_x = \begin{cases} \sigma_\theta - \lambda_N \tau & \text{if } \sigma_x > 0 \\ 0 & \text{if } \sigma_x = 0 \\ \sigma_\theta + \lambda_N \tau & \text{if } \sigma_x < 0 \end{cases} = \begin{cases} \theta - \lambda_N \tau & \text{if } \sigma_\theta \geq \lambda_N \tau \\ 0 & \text{if } |\sigma_\theta| \leq \lambda_N \tau \\ \theta + \lambda_N \tau & \text{if } \sigma_\theta \leq -\lambda_N \tau \end{cases}$$

$$\boxed{prox_{\tau\mathcal{P}_N}(\theta) = Udiag(sign(\sigma_\theta)(|\sigma_\theta| - \lambda_N\tau)^+V^\intercal}$$

2.2 Combinaison of penalties

2.2.1 Elastic Net

The elastic net penalty is $\mathcal{P}_{12}(\theta) = \lambda_1||\theta||_1 + \lambda_2||\theta||_2^2$.

The proximal operator can be computed solving:

$$\begin{aligned}
prox_{\tau\mathcal{P}_{12}}(\theta) &= \underset{x}{\text{argmin}}\left(\lambda_1||x||_1 + \lambda_2||x||_2^2 + \frac{1}{2\tau}||x-\theta||_2^2\right) \\
\partial_x\left(\lambda_1||x||_1 + \lambda_2||x||_2^2 + \frac{1}{2\tau}||x-\theta||_2^2\right) &= 0 \\
&= \lambda_1 s + \lambda_2 x + \frac{1}{\tau}(x-\theta) \\
x &= \frac{1}{1+\tau\lambda_2}(\theta - \lambda_1\tau s)
\end{aligned}$$

$$\begin{aligned}
x &= \begin{cases} \frac{1}{1+\tau\lambda_2}(\theta-\lambda_1\tau) & \text{if } x>0 \\ 0 & \text{if } x=0 \\ \frac{1}{1+\tau\lambda_2}(\theta+\lambda_1\tau) & \text{if } x<0 \end{cases} = \begin{cases} \frac{1}{1+\tau\lambda_2}(\theta-\lambda_1\tau) & \text{if } \theta\geq\lambda_1\tau \\ 0 & \text{if } |\theta|\leq\lambda_1\tau \\ \frac{1}{1+\tau\lambda_2}(\theta+\lambda_1\tau) & \text{if } \theta\leq-\lambda_1\tau \end{cases} \\
prox_{\tau\mathcal{P}_1}(\theta) &= \frac{1}{1+\tau\lambda_2}(sign(\theta)(|\theta|-\lambda_1\tau)) \qquad (1)
\end{aligned}$$

$$(2)$$

$$\boxed{prox_{\tau\mathcal{P}_1}(\theta) = prox_{\tau\mathcal{P}_2}(prox_{\tau\mathcal{P}_1}(\theta))}$$

2.2.2 Sparse Group lasso

The sparse group lasso penalty is $\mathcal{P}_{G1}(\theta) = \lambda_1||\theta||_1 + \lambda_G||\theta||_2$ where θ is a vector.

The proximal operator can be computed solving:

$$\begin{aligned}
prox_{\tau\mathcal{P}_{G1}}(\theta) &= \underset{x}{\text{argmin}}\left(\lambda_1||\theta||_1 + \lambda_G||\theta||_2 + \frac{1}{2\tau}||x-\theta||_2^2\right) \\
\partial_{x_j}\left(\lambda_1||\theta||_1 + \lambda_G||\theta||_2 + \frac{1}{2\tau}||x-\theta||_2^2\right) &= 0 \\
&= \lambda_1 s_1(x_j) + \lambda_G s_G(x)_j + \frac{1}{\tau}(x_j-\theta_j) \\
x_j &= \theta_j - \lambda_1\tau s_1(x_j) - \lambda_G\tau s_G(x)_j
\end{aligned}$$

If x is not null then:

$$\begin{aligned}
x &= \theta - \lambda_1\tau s_1(x) - \lambda_G\tau\frac{x}{||x||_2} \\
x\left(1+\frac{\lambda_G\tau}{||x||_2}\right) &= \theta - \lambda_1\tau s_1(x) \\
x &= \frac{||x||_2}{||x||_2+\lambda_G\tau}(\theta - \lambda_1\tau s_1(x))
\end{aligned}$$

Taking the L2 norm on both sides:

$$\begin{aligned}
|||x||_2 + \lambda_G \tau| &= ||x||_2 + \lambda_G \tau \\
&= ||x - \lambda_1 \tau s_1(x)||_2 \\
||x||_2 &= ||x - \lambda_1 \tau s_1(x)||_2 - \lambda_G \tau
\end{aligned}$$

So $||x||_2 > 0$ implies $||x - \lambda_1 \tau s_1(x)||_2 \geq \lambda_G \tau$

Then:

$$\begin{aligned}
x &= \frac{||x - \lambda_1 \tau s_1(x)||_2 - \lambda_G \tau}{||x - \lambda_1 \tau s_1(x)||_2} (x - \lambda_1 \tau s_1(x)) \\
&= \left(1 - \frac{\lambda_G \tau}{||x - \lambda_1 \tau s_1(x)||_2}\right) (x - \lambda_1 \tau s_1(x))
\end{aligned}$$

$$prox \tau \mathcal{P}_G(\theta) = prox_{\tau \mathcal{P}_G}(prox_{\tau \mathcal{P}_1}(\theta))$$

2.2.3 others ?

If the penalty is separable: $\mathcal{P}_{gh}(\theta) = \lambda_g g(\theta_g) + \lambda_h h(\theta_h)$ Then

$$\begin{aligned}
prox_{\tau \mathcal{P}_{gh}}(\theta) &= \underset{x}{\operatorname{argmin}} \left(\lambda_g g(\theta_g) + \lambda_h h(\theta_h) + \frac{1}{2\tau} ||x - (\theta_f, \theta_g)||_2^2\right) \\
&= \underset{x}{\operatorname{argmin}} \left(\lambda_g g(\theta_g) + \lambda_h h(\theta_h) + \frac{1}{2\tau} ||x_g - \theta_g||_2^2 + \frac{1}{2\tau} ||x_h - \theta_h||_2^2\right)
\end{aligned}$$

So we can apply independently the proximal operator relative to each penalty.

3 Appendix

3.1 Unicity of the SVD

$A = U_1\Sigma_1 V_1 = U_2\Sigma_2 V_2$

Since U_1, U_2 and $M = {U_2}^\intercal U_1$ are unitary matrices:

$$\begin{aligned}
AA^\intercal &= U_1\Sigma_1{\Sigma_1}^\intercal {U_1}^\intercal = U_2\Sigma_2{\Sigma_2}^\intercal {U_2}^\intercal \\
U_1\Sigma_1{\Sigma_1}^\intercal &= U_2\Sigma_2{\Sigma_2}^\intercal {U_2}^\intercal U_1 \\
{U_2}^\intercal U_1\Sigma_1{\Sigma_1}^\intercal &= \Sigma_2{\Sigma_2}^\intercal {U_2}^\intercal U_1 \\
M\Sigma_1^2 &= \Sigma_2^2 M
\end{aligned}$$

Therefore $det(\Sigma_1) = det(\Sigma_2)$ and $tr(\Sigma_1) = tr(\Sigma_2)$. So if there are 2 or less eigenvalues, Σ_1 andΣ_2 are equal in absolute value. By recurrence if it is true for p eigenvalues we consider a matrix with p+1 eigen vectors. We can use an indicator vector x to project on a subspace of size p:

$$\begin{aligned}
x^\intercal M\Sigma_1^2 x &= x^\intercal \Sigma_2^2 M x \\
M'\Sigma_1^{2'} &= \Sigma_2^{2'} M'
\end{aligned}$$

Using the hypothesis of the recurrence we find that Σ_1 and Σ_2 must coincide in any subspace in absolute value. Therefore they must be equal in absolute value. Denoting ϵ a diagonal matrix filled with -1 and 1 such that:

$$\begin{aligned}
\Sigma_1 &= \Sigma_2\epsilon \\
U_1\Sigma_1 V_1 &= U_2\Sigma_1\epsilon V_2 = U_2\Sigma_1 V_2'
\end{aligned}$$

Because $(\epsilon V_2)^\intercal \epsilon V_2 = {V_2}^\intercal V_2$ and ϵ and V_2 commute since ϵ is diagonal. Therefore we can find U_2 and V_2 such that $\Sigma_1 = \Sigma_2$

Moreover:

$$\begin{aligned}
{U_2}^\intercal U_1\Sigma_1{\Sigma_1}^\intercal &= \Sigma_1{\Sigma_1}^\intercal {U_2}^\intercal U_1 \\
M\Sigma &= \Sigma M
\end{aligned}$$

We know that the eigenvectors of Σ are the canonical basis ($\{e_j; j = 1, \dots, p\}$). But since M and Σ commute, $\{Me_j; j = 1, \dots p\}$ are also eigenvectors. Since Σ has at most p eigen vectors then $\{Me_j; j = 1, \dots p\} \propto \{e_j; j = 1, \dots p\}$ and thus $\{e_j; j = 1, \dots p\}$ are the eigenvector of M. Then M is diagonal so $U_1 = UU_2$ with U diagonal and unitary.

3.2 Commutation and eigen vectors

Let consider A and B two matrices that commutes.

x eigenvector of A with eigenvalue $\lambda \Leftrightarrow Bx$ eigenvector of A:

$$ABx = BAx = \lambda Bx$$

The reciproque is also true:

$$\begin{aligned}
BAx &= ABx = \lambda Bx \\
B^{-1}BAx &= \lambda BB^{-1}x \\
Ax &= \lambda x
\end{aligned}$$

Therefore if A has distinct eigenvalues x and Bx must correspond to the same eigenvector and are thus linearly related. x is then an eigen vector for B. In addition if both are invertible and have n linear independent eigenvectors they must be the same.

References

Watson, G. A. (1992). Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45.