# 02471 Machine Learning for Signal Processing

## *Solution*

# Exercise 6: Sparsity-aware learning with $\ell_1$

## 6.1 Norms

### Exercise 6.1.1

Solution: We consider the case $p = 0$ and $0 < p < 1$ seperately.

For $p = 0$ we can come with the following counter-example:

Consider the vector $\boldsymbol{x} = [1, 0, \ldots 0]^T$, and choose any non-zero $\alpha \in \mathbb{R}$. Then we get (left hand side of the second property) gives

$$\|\alpha \boldsymbol{x}\|_0 = \|[\alpha, 0, \ldots 0]^T\|_0 = 1$$

But IF the property holds, the result should have been

$$\|\alpha \boldsymbol{x}\|_0 = |\alpha| \|\boldsymbol{x}\|_0 = |\alpha| \ 1 = |\alpha|$$

So clearly the second property is violated, thus $p = 0$ is not a norm.

For $0 < p < 1$ we us show that property three is violated. Consider the two vectors in the $l$ -dimensional space

$$\boldsymbol{x} = [1, 0, \ldots, 0]^T, \quad \boldsymbol{y} = [0, 0, \ldots, 1]^T$$

We will show that for these two vectors the triangle inequality is violated for $p < 1$. Indeed, we have (assuming now the triangle inequality holds)

$$\|\boldsymbol{x} + \boldsymbol{y}\|_p = \left( \sum_{i=1}^{l} |x_i + y_i|^p \right)^{1/p} = (1^p + 1^p)^{\frac{1}{p}} = (2 \cdot 1^p)^{\frac{1}{p}} = 2^{\frac{1}{p}} \leq \|x\|_p + \|y\|_p = 1 + 1 = 2$$

which is violated for $0 < p < 1$.

## 6.2 The regularized least-squares solution

### Exercise 6.2.1

This is solved by direct substitution. If $X^T X = I$ we get

$$\hat{\boldsymbol{\theta}}_{\text{LS}} = \left(X^T X\right)^{-1} X^T \boldsymbol{y} = I^{-1} X^T \boldsymbol{y} = X^T \boldsymbol{y}$$

For Ridge regression we get

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_R &= \left(X^T X + \lambda I\right)^{-1} X^T \boldsymbol{y} \\
&= (I + \lambda I)^{-1} X^T \boldsymbol{y} \\
&= (I(1 + \lambda))^{-1} X^T \boldsymbol{y} \\
&= I^{-1}(1 + \lambda)^{-1} X^T \boldsymbol{y} \\
&= \frac{1}{1 + \lambda} X^T \boldsymbol{y} \\
&= \frac{1}{1 + \lambda} \hat{\boldsymbol{\theta}}_{\text{LS}}
\end{aligned}$$

**Exercise 6.2.2**

This is shown in the book, equation (9.11)–9.13).

**Exercise 6.2.3**

Direct application of the formula will give the following result

$$\hat{\boldsymbol{\theta}}_R = \frac{1}{1+\lambda}[0.7, -0.3, 0.1, -2]^T = \frac{1}{2}[0.7, -0.3, 0.1, -2]^T = [0.35, -0.15, 0.05, -1]^T$$

For the $\ell_1$ norm we get

$$\hat{\boldsymbol{\theta}}_1 = \begin{bmatrix} \mathrm{sgn}(0.7)\left(|0.7| - \frac{1}{2}\right)_+ \\ \mathrm{sgn}(-0.3)\left(|-0.3| - \frac{1}{2}\right)_+ \\ \mathrm{sgn}(0.1)\left(|0.1| - \frac{1}{2}\right)_+ \\ \mathrm{sgn}(-2)\left(|-2| - \frac{1}{2}\right)_+ \end{bmatrix} = \begin{bmatrix} 1\,(0.2)_+ \\ -1\,(-0.2)_+ \\ 1\,(-0.4)_+ \\ -1\,(1.5)_+ \end{bmatrix} = \begin{bmatrix} 1 \cdot 0.2 \\ -1 \cdot 0 \\ 1 \cdot 0 \\ -1 \cdot 1.5 \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0 \\ 0 \\ -1.5 \end{bmatrix}$$

**Exercise 6.2.4**

Ridge regression requires $\lambda \to \infty$ to result in the null vector. For LASSO, $\lambda = 4$.