

# Exercise of Supervised Learning: Regularization Part 1

Yawei Li

yawei.li@stat.uni-muenchen.de

December 8, 2023

# Exercise 1: L0 Regularization

Consider the regression learning setting, i.e.,  $\mathcal{Y} = \mathbb{R}$ , and the feature space  $\mathcal{X} = \mathbb{R}^p$ . Let the hypothesis space be the linear models:

$$\mathcal{H} = \{f(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x} \mid \boldsymbol{\theta} \in \mathbb{R}^p\}.$$

Suppose your loss function of interest is the L2 loss  $L(y, f(\mathbf{x})) = \frac{1}{2}(y - f(\mathbf{x}))^2$ . Consider the  $L_0$ -regularized empirical risk of a model  $f(\mathbf{x} \mid \boldsymbol{\theta})$ :

$$\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_0 = \frac{1}{2} \sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)})^2 + \lambda \sum_{i=1}^p \mathbf{1}_{|\theta_i| \neq 0}.$$

Assume that  $\mathbf{X}^T \mathbf{X} = \mathbf{I}$ , which holds if  $\mathbf{X}$  has orthonormal columns. Show that the minimizer  $\hat{\boldsymbol{\theta}}_{L0} = (\hat{\theta}_{L0,1}, \dots, \hat{\theta}_{L0,p})^T$  is given by

$$\hat{\theta}_{L0,i} = \hat{\theta}_i \mathbf{1}_{\hat{\theta}_i > \sqrt{2\lambda}}, \quad i = 1, \dots, p,$$

where  $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_p)^T = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$  is the minimizer of the unregularized empirical risk. For this purpose, using the following steps:

## Exercise 1 (i)

(i) Derive that

$$\arg \min_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \arg \min_{\boldsymbol{\theta}} \sum_{i=1}^p -\hat{\theta}_i \theta_i + \frac{\theta_i^2}{2} + \lambda \mathbf{1}_{|\theta_i| \neq 0}.$$

Note that  $\theta_i$  is from the minimizer  $\hat{\boldsymbol{\theta}}$  of the unregularized empirical risk :

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{X}^T \mathbf{y}$$

because we assume that  $\mathbf{X}^T \mathbf{X} = \mathbf{I}$ .

## Solution to Exercise 1 (i)

$$\begin{aligned}\arg \min_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) &= \arg \min_{\boldsymbol{\theta}} \frac{1}{2} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \lambda \sum_{i=1}^p I_{|\theta_i| \neq 0} \\&= \arg \min_{\boldsymbol{\theta}} \frac{1}{2} \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}\boldsymbol{\theta} + \frac{1}{2} \boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\theta} + \lambda \sum_{i=1}^p I_{|\theta_i| \neq 0} \\&= \arg \min_{\boldsymbol{\theta}} - \underbrace{\mathbf{y}^T \mathbf{X}}_{\hat{\boldsymbol{\theta}}^T} \boldsymbol{\theta} + \frac{1}{2} \boldsymbol{\theta}^T \underbrace{\mathbf{X}^T \mathbf{X}}_{=\mathbf{I}} \boldsymbol{\theta} + \lambda \sum_{i=1}^p I_{|\theta_i| \neq 0} \\&= \arg \min_{\boldsymbol{\theta}} -\hat{\boldsymbol{\theta}}^T \boldsymbol{\theta} + \frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\theta} + \lambda \sum_{i=1}^p I_{|\theta_i| \neq 0} \\&= \arg \min_{\boldsymbol{\theta}} \sum_{i=1}^p -\hat{\theta}_i \theta_i + \frac{\theta_i^2}{2} + \lambda \sum_{i=1}^p I_{|\theta_i| \neq 0} \quad \triangleleft\end{aligned}$$

## Exercise 1 (ii)

(ii) Note that the minimization problem on the right-hand side of (i) can be written as  $\sum_{i=1}^p g_i(\theta)$ , where

$$g_i(\theta) = -\hat{\theta}_i \theta + \frac{\theta^2}{2} + \lambda \mathbf{1}_{|\theta| \neq 0}.$$

What is the advantage of this representation if we seek to find the  $\theta$  with entries  $\theta_1, \dots, \theta_p$  minimizing  $\mathcal{R}_{\text{reg}}(\theta)$ ?

## Solution to Exercise 1 (ii)

Advantage: we can minimize each  $g_i$  **separately** to obtain the optimal entries  $\theta_1, \dots, \theta_p$ .

## Exercise 1 (iii)

Consider the first case that  $|\hat{\theta}_i| > \sqrt{2\lambda}$  and infer that for the minimizer  $\theta_i^*$  of  $g_i$  it must hold that  $\theta_i^* = \hat{\theta}_i$ .

*Hint:* Show that  $g_i(\theta_i) < 0 = g_i(0)$  and argue that the minimizer must have the same sign as  $\hat{\theta}_i$ . **(Personally I find this hint is not so useful.)**

In other words, if  $|\hat{\theta}_i|$  is larger than the threshold,  $\sqrt{2\lambda}$ , then the optimal  $\theta_i^*$  is the consistent between the regularized and un-regularized empirical risk.

## Solution to Exercise 1 (iii)

We start with computing the  $\arg \min g_i(\theta_i) = \frac{\theta_i^2}{2} - \hat{\theta}_i \theta_i + \lambda \mathbf{1}_{|\theta_i| \neq 0}$ .

- Case 1:  $\theta_i = 0$ . Then,  $g_i(\theta_i) = 0$ .
- Case 2:  $\theta_i \neq 0$ . Then

$$g_i(\theta_i) = \frac{\theta_i^2}{2} - \hat{\theta}_i \theta_i + \lambda,$$

which is a quadratic function, and its minimizer is

$$\theta_i^* = \arg \min_{\theta_i} g_i(\theta_i) = \hat{\theta}_i,$$

and the minimal value of  $g_i$  in this case is

$$g_i(\theta_i^*) = -\frac{\hat{\theta}_i^2}{2} + \lambda.$$

Which optimal  $g_i$  is smaller? Case 1 or Case 2? It depends on  $\lambda$ . Note that we are given with

$|\hat{\theta}_i| > \sqrt{2\lambda}$ . So  $-\frac{\hat{\theta}_i^2}{2} + \lambda < -\frac{2\lambda}{2} + \lambda = 0$ . So the optimal  $g_i$  in Case 2 is smaller.

So for the minimizer of  $g_i$  it holds that  $\theta_i^* = \hat{\theta}_i$ .



## Exercise 1 (iv)

(iv) Derive that  $\theta_i^* = \hat{\theta}_i \mathbf{1}_{|\hat{\theta}_i| > \sqrt{2\lambda}}$ , by using (iii) (and also still considering the case  $|\hat{\theta}_i| > \sqrt{2\lambda}$ ).

**Solution:** In the solution of (iii) we have proven that

$$\theta_i^* = \hat{\theta}_i$$

given that  $|\hat{\theta}_i| > \sqrt{2\lambda}$ . Given this constraint, the optimal  $\theta_i$  can be written as

$$\theta_i^* = \hat{\theta}_i \cdot 1 = \hat{\theta}_i \mathbf{1}_{|\hat{\theta}_i| > \sqrt{2\lambda}}.$$

## Exercise 1 (v)

(v) Consider the complementary case of (iii) and (iv), i.e.,  $|\hat{\theta}_i| \leq \sqrt{2\lambda}$ , and infer that for the minimizer  $\theta_i^*$  of  $g_i$  it must hold that  $\theta_i^* = 0$ .

*Hint:* What is  $g_i(0)$ ? Consider  $\tilde{g}_i(\theta) = \hat{\theta}_i\theta + \frac{\theta^2}{2} + \lambda$  which is the smooth extension of  $g_i$ . What is the relationship between the minimizer of  $g_i$  and the minimizer of  $\tilde{g}_i$ ?

**(We do not need this hint in the solution presented in the subsequent slides)**

## Solution to Exercise 1 (v)

Similarly, we start with computing  $\arg \min g_i(\theta_i) = \frac{\theta_i^2}{2} - \hat{\theta}_i \theta_i + \lambda \mathbf{I}_{|\theta_i| \neq 0}$ .

- Case 1:  $\theta_i = 0$ . Then,  $g_i(\theta_i) = 0$ .
- Case 2:  $\theta_i \neq 0$ . Then

$$g_i(\theta_i) = \frac{\theta_i^2}{2} - \hat{\theta}_i \theta_i + \lambda,$$

We have shown that the minimizer in **this case** is  $\theta_i^* = \hat{\theta}_i$  and  $\min g_i(\theta_i^*) = g_i(\hat{\theta}_i) = -\frac{\hat{\theta}_i^2}{2} + \lambda$ .

Since we consider the constraint  $|\hat{\theta}_i| \leq \sqrt{2\lambda}$ . Then in Case 2

$$g_i(\theta_i^*) \geq -\frac{2\lambda}{2} + \lambda = 0.$$

So the minimal  $g_i$  in Case 2 is **not smaller** than the minimal  $g_i$  in Case 1. **(Plot  $g_i(\theta_i)$  vs.  $\theta_i$ ).**  
Therefore, combining two cases, for the minimizer  $\theta_i^*$  of  $g_i$  it holds that

$$\theta_i^* = 0.$$

## Exercise 2: Regularization

**Directly show the standard solution.**