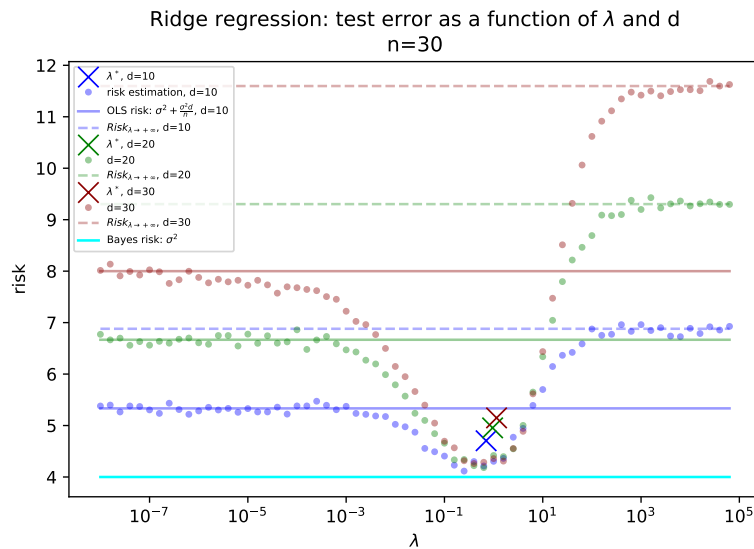


# FTML practical session 9: 2023/05/26



## TABLE DES MATIÈRES

1	Quantitative evaluation of the benefits of Ridge regression	1
1.1	Reminders of the theoretical results	1
1.2	Existence of a benefit	3
1.3	Large bias	3
2	Cross validation	3

## 1 QUANTITATIVE EVALUATION OF THE BENEFITS OF RIDGE REGRESSION

The goal of this exercise is to have a more concrete representations of situations where Ridge regression is useful. We will state some theoretical results, and then find datasets for which these results apply.

### 1.1 Reminders of the theoretical results

We keep the same statistical setting as before (fixed design, linear model). We have seen in the previous classes that the excess risk is  $\frac{\sigma^2 d}{n}$  for OLS (Ordinary least squares)

**Definition 1.** Ridge regression estimator

It is defined as

$$\hat{\theta}_\lambda = \arg \min_{\theta \in \mathbb{R}^d} \left( \frac{1}{n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_2^2 \right) \quad (1)$$

**Proposition.** The Ridge regression estimator is unique even if  $X^T X$  is not inversible and is given by

$$\hat{\theta}_\lambda = \frac{1}{n} (\hat{\Sigma} + \lambda I_d)^{-1} X^T y$$

**Proposition.** We assume the linear model, with fixed design setting, with a Bayes estimator of  $\theta^*$  and a noise with a variance of  $\sigma^2$ . Then, the ridge regression estimator has the following excess risk :

$$E[R(\hat{\theta}_\lambda) - R^*] = \lambda^2 \theta^{*T} (\hat{\Sigma} + \lambda I_d)^{-2} \hat{\Sigma} \theta^* + \frac{\sigma^2}{n} \text{tr}[\hat{\Sigma}^2 (\hat{\Sigma} + \lambda I_d)^{-2}] \quad (2)$$

**Comments :**

- We observe a bias / variance decomposition.
- We consider the bias term B :

$$B = \lambda^2 \theta^{*T} (\hat{\Sigma} + \lambda I_d)^{-2} \hat{\Sigma} \theta^* \quad (3)$$

- The bias B increases when  $\lambda$  increases. It is an approximation error and does not depend on n.
- When  $\lambda = 0$  and  $\hat{\Sigma}$  is invertible (which corresponds to OLS),  $B = 0$ .
- When  $\lambda \rightarrow +\infty$ ,  $B \rightarrow \theta^{*T} \hat{\Sigma} \theta^*$ .
- We consider the variance term V :

$$V = \frac{\sigma^2}{n} \text{tr}[\hat{\Sigma}^2 (\hat{\Sigma} + \lambda I_d)^{-2}] \quad (4)$$

- The variance V decreases when  $\lambda$  increases. It is an estimation error and depends on n
- When  $\lambda = 0$  and  $\hat{\Sigma}$  is invertible (which corresponds to OLS),  $V = \frac{\sigma^2 d}{n}$ .
- When  $\lambda \rightarrow +\infty$ ,  $V \rightarrow 0$ .
- When  $n \rightarrow +\infty$ ,  $V \rightarrow 0$ .

A natural question is whether it is possible to have a lower excess risk with Ridge regression than with OLS, which means an excess risk smaller than  $\frac{\sigma^2 d}{n}$ . We admit the following proposition.

**Proposition.** With the choice

$$\lambda^* = \frac{\sigma \sqrt{\text{tr}(\hat{\Sigma})}}{\|\theta^*\|_2 \sqrt{n}} \quad (5)$$

then

$$E[R(\hat{\theta}_\lambda) - R^*] \leq \frac{\sigma \sqrt{\text{tr}(\hat{\Sigma})} \|\theta^*\|_2}{\sqrt{n}} \quad (6)$$

with

$$\hat{\Sigma} = \frac{1}{n} X^T X \in \mathbb{R}^{d,d} \quad (7)$$

Hence, the convergence to 0 in OLS is in  $\frac{1}{n}$ , while it is in  $\frac{1}{\sqrt{n}}$  for the ridge. However, for the ridge regression, the dependence in the noise is in  $\sigma$ , whereas it is  $\sigma^2$  for OLS. Which one is preferable will depend on the value of the constants, and will not necessarily be the "fast" rate in  $\mathcal{O}(\frac{1}{n})$ .

## 1.2 Existence of a benefit

Find a setting (statistical values, dataset) for which the expected risk is strictly lower for Ridge regression than OLS.

Verify it with a simulation where you compare the test error of OLS and that of Ridge regression with a good choice of the regularization parameter  $\lambda$ .

## 1.3 Large bias

In some settings Ridge performs worse than OLS when  $\lambda$  is too large, as in figure

1. As we have seen, when  $\lambda \rightarrow +\infty$  :

- $V \rightarrow 0$  (variance)
- $B \rightarrow \theta^{*\top} \hat{\Sigma} \theta^*$  (bias)



FIGURE 1 – Ridge and OLS, where Ridge performs bad for  $\lambda \rightarrow +\infty$ , because of the bias becomes large.

Given  $\hat{\Sigma}$ , how could we choose  $\theta^*$  in order to have a high bias when  $\lambda \rightarrow +\infty$ ?  
Implement a simulation in order to observe this large bias.

## 2 CROSS VALIDATION

In practical situations, we have also seen that the quantities involved in the computation of  $\lambda^*$  in 5 are typically unknown. Good values for  $\lambda$  are found by **cross-validation**., with hyperparameter search methods (Gridsearch, Bayesian optimization, etc).

[https://scikit-learn.org/stable/modules/cross\\_validation.html](https://scikit-learn.org/stable/modules/cross_validation.html)

Compare the best hyperparameters found by automated search methods with  $\lambda^*$ , and the test errors obtained with both, for various statistical settings / values of  $d$ ,  $n$ , etc. In some contexts, there might exist some regularization parameter values that are **better** than  $\lambda^*$ , like in figure 3.

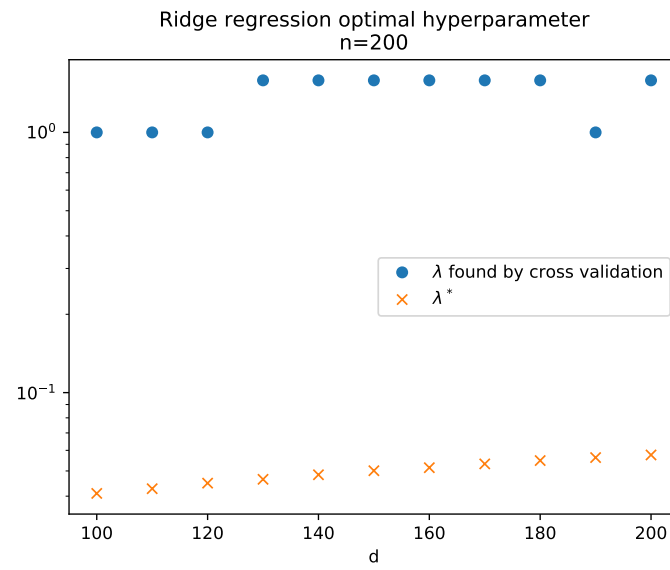


FIGURE 2 – Comparison of  $\lambda^*$  and of the values found by cross-validation,  $n = 200$ .

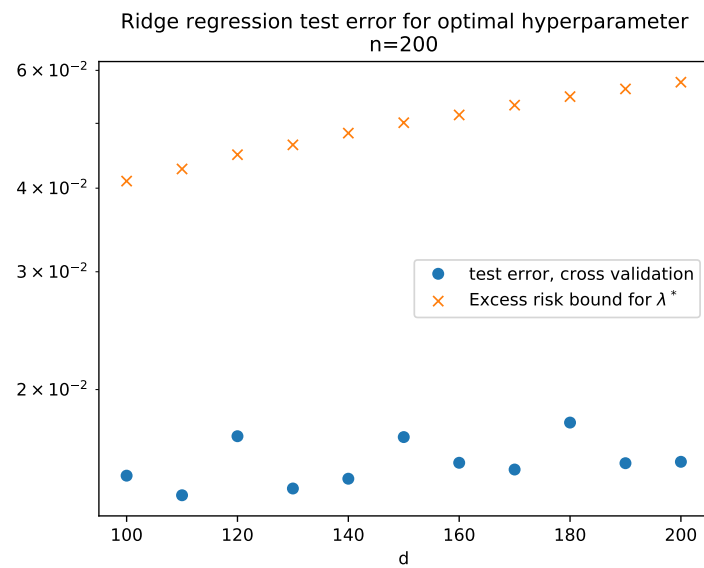


FIGURE 3 – Scores obtained for both parameters,  $n = 200$