Math of ML : Exercises 2*

September 15, 2025

Unless otherwise specified, in the following exercises we consider a *fixed* design matrix $X \in \mathbb{R}^{n\times d}$ with rows $x_i^\top$ and assume that it is full-rank (in particular $n \geq d$).

Exercise 1 (Geometric interpretation of least-squares)**.** *Show that the vector of predictions $X\hat{w} = X(X^\top X)^{-1}X^\top y$ is the orthogonal projection of y on* $\mathrm{im}(X) = \{Xw : w \in \mathbb{R}^d\}$.

Exercise 2 (Empirical risk of OLS)**.** *We consider the noisy measurement model $Y_i = x_i^\top w^* + Z_i$ where Z_i are independent, with zero mean $\mathbf{E}Z_i = 0$ and variance $\mathbf{E}[Z_i^2] = \sigma^2$ (the x_i are fixed). What is the expected empirical risk $\mathbf{E}[\hat{\mathcal{R}}_X(\hat{w})]$? Use this answer to propose an estimator of the noise variance σ^2 when $n > d$.*

Exercise 3 (OLS as a maximum-likelihood estimator)**.** *In this exercise, we make the stronger assumption that the i.i.d. noise random variables $Z_1, \dots, Z_n$ all follow the Gaussian distribution $Z_i \sim \mathcal{N}(0, \sigma^2)$. Under the fixed-design Gaussian noise model, letting $Y = (Y_1, \dots, Y_n)$, the likelihood that the observations Y were generated via the well-specified model $Y_i = x_i^\top w + Z_i$ is equal to*

$$L(Y|w, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-(Y_i - x_i^\top w)^2/(2\sigma^2)\Big).$$

The maximum likelihood estimator $(\tilde{w}, \tilde{\sigma})$ is defined as

$$(\tilde{w}, \tilde{\sigma}) = \mathrm{argmax}_{w\in\mathbb{R}^d, \sigma>0} L(Y|w, \sigma^2).$$

In this exercise, you are asked to:

1. *show that $\tilde{w}$ coincides with the OLS estimator;*
2. *compute the maximum likelihood estimator $\tilde{\sigma}^2$ for the variance. Is the estimator unbiased?*

Exercise 4 ((Practical exercise) Tuning ridge parameter)**.** *Recall that the ridge regression estimator is defined by*

$$\hat{w}_\lambda = \mathrm{argmin}_{w\in\mathbb{R}^d} \left\{ \frac{1}{n}\|Xw - y\|_2^2 + \lambda\|w\|_2^2 \right\}.$$

In Section 4 of Lecture 2, we suggested the choice of regularization parameter $\lambda^* = \frac{\sigma\sqrt{\mathrm{tr}\,\Sigma}}{\|w^*\|_2\sqrt{n}}$. In particular, the suggested parameter λ^* scales as inverse square root of the number of samples.

In the language of your choice, perform the following simulation for $n = 500, 600, 700, 800, 900, 1000$:

*Lénaïc Chizat EPFL lenaic.chizat@epfl.ch

1. *Let* $d = 50$.

2. *Choose a fixed design matrix* $X \in \mathbb{R}^{n \times d}$ *by sampling each entry from a Gaussian distribution* $\mathcal{N}(0, 1)$.

3. *Let* $w^* = \mathbf{1}/\sqrt{d} \in \mathbb{R}^d$ *be the ground truth parameter of norm* $\|w^\star\|_2 = 1$.

4. *Fix some large enough grid of* λ *parameter values* $[\lambda_1, \ldots, \lambda_k]$.

5. *For each* $j = 1, \ldots, k$, *compute the expected excess risk of the ridge regression estimator with parameter* λ_j *using the bias-variance decomposition formula (see Lecture 2, Proposition 4.3; see also [Bach, 2024, Proposition 3.7])*

   For this part of the exercise, you may let the variance of the noise be equal to $\sigma^2 = 1$.

6. *Compute* λ_n^{opt} *that minimizes the excess risk among the values of* λ *in the grid* $[\lambda_1, \ldots, \lambda_k]$.

How does the computed solution λ_n^{opt} *compare with* λ^* *obtained theoretically in the lectures? As the number of samples* n *increases, how fast (approximately) does* λ_n^{opt} *decrease as a function of the number of samples* n*? Explain your findings.*

Exercise 5 (A non-linear estimator)**.** *In the previous exercises, we investigated the setting where the design matrix was fixed and the observations followed the model* $Y_i = x_i^\top w^* + Z_i$ *for some zero-mean noise variables* Z_i.

In this exercise, we take a look at the setting where the design is random and the Bayes optimal function is no longer assumed to be linear. Despite no longer assuming that the Bayes optimal function is linear, we may still be interested in bounding the excess risk

$$\mathbf{E}\mathcal{R}(\hat{f}) - \inf_{w \in \mathbb{R}^d} \mathcal{R}(\langle w, \cdot \rangle), \tag{1}$$

where recall that for any function $f : \mathbb{R}^d \to \mathbb{R}$ *we have*

$$\mathcal{R}(f) = \mathbf{E}_{(X,Y)\sim P}(f(X) - Y)^2.$$

In particular, the excess risk (1) *measures how close the estimator* $\hat{f}$ *gets to the best linear explanation of the data.*

The aim of this exercise is to introduce a new technique for bounding the excess risk, called average-stability *analysis, and further, to show how through the calculations performed in Exercise 6 it suggests a certain non-linear predictor due to Forster and Warmuth [2002].*

We shall make the following two assumptions[1] *on the unknown data generating mechanism:*

- *The data samples* (X_i, Y_i) *are generated i.i.d. from a distribution* P *such that we have* $|Y_i| \leq m$ *almost surely for some constant* $m > 0$.

- *For any data sample of size* n $(X_i, Y_i)_{i=1}^n$, *it holds with probability one that the matrix* $\sum_{i=1}^n X_i X_i^\top$ *is invertible (that is, we can compute the OLS estimator).*

This exercise is split into three parts.

[1] Both assumptions can be relaxed. The bounded response variable assumption can be weakened to assuming that the random variable $\mathbf{E}[Y^2|X]$ is almost surely bounded by m^2. We can get rid of the invertibility assumption completely, by replacing matrix inverses by their corresponding Moore-Penrose inverses.

1. *Denote a data sample of $n+1$ points by $S_{n+1} = (X_i, Y_i)_{i=1}^{n+1}$ and let $S_{n+1}^{(-j)}$ denote the sample S_{n+1} without the j-th input-output pair (X_j, Y_j). Show that we may rewrite the expected risk of any estimator $\hat{f} = \hat{f}[S_n]$ as follows:*

$$\mathbf{E}_{S_n} \mathcal{R}(\hat{f}[S_n]) = \mathbf{E}_{S_{n+1}} (\hat{f}[S_n](X_{n+1}) - Y_{n+1})^2.$$

Show that for any estimator $\hat{f} = \hat{f}[(X_1, Y_1), \ldots, (X_n, Y_n)]$ we have

$$\mathbf{E}_{(X_i,Y_i)_{i=1}^n} \mathcal{R}(\hat{f}) = \mathbf{E}_{S_{n+1}} \left[\frac{1}{n+1} \sum_{j=1}^{n} \left(\hat{f}[S_{n+1}^{(-j)}](X_j) - Y_j \right)^2 \right].$$

2. *Let $\hat{w} = \hat{w}[S_{n+1}]$ be the OLS estimator computed on the sample S_{n+1}. Using the previous part of this exercise, show that the excess risk (1) can be upper bounded by*

$$\mathbf{E}\mathcal{R}(\hat{f}) - \inf_{w \in \mathbb{R}^d} \mathcal{R}(\langle w, \cdot \rangle) \leq \mathbf{E}_{S_{n+1}} \left[\frac{1}{n+1} \sum_{j=1}^{n} (\hat{f}[S_{n+1}^{(-j)}](X_j) - Y_j)^2 - (X_j^\top \hat{w}[S_{n+1}] - Y_j)^2 \right].$$

3. *Using part 2 of this exercise, suggest an estimator $\hat{f}$ such that*

$$\mathbf{E}\mathcal{R}(\hat{f}) - \inf_{w \in \mathbb{R}^d} \mathcal{R}(\langle \cdot, w \rangle) \leq \frac{2m^2 d}{n+1}.$$

You may use the fact that for any $j = 1, \ldots, n+1$,

$$X_j^\top \hat{w}[S_{n+1}] = (1 - h_j) X_j^\top \hat{w}[S_{n+1}^{(-j)}] + h_j Y_j,$$

where $h_j \in [0, 1]$ is the j-th leverage score[2] *defined by*

$$h_j = X_j^\top \left(\sum_{i=1}^{n+1} X_i X_i^\top \right)^{-1} X_j.$$

Exercise 6 (Optional: Stability of OLS predictions)**.** *Prove the fact used in part 3 of the previous exercise. That is to say:*

Let the fixed design matrix $X \in \mathbb{R}^{(n+1) \times d}$ contain $n+1$ rows $x_i^\top$ and let $y = (Y_1, \ldots, Y_{n+1}) \in \mathbb{R}^{n+1}$ denote the vector of observed response variables. Let $\hat{w} = (X^\top X)^{-1} X^\top y$ denote the OLS estimator computed on the $(n+1)$ data points.

Consider removing the data point (x_j, Y_j) and computing the OLS estimator

$$\hat{w}_{(-j)} = (X_{(-j)}^\top X_{(-j)})^\top X_{(-j)}^\top y_{(-j)},$$

where $X_{(-j)} \in \mathbb{R}^{n \times d}$ is a matrix obtained by removing the j-th row of X, and $y_{(-j)} \in \mathbb{R}^n$ is a vector obtained by removing the j-th entry of y.

Show that for any $j = 1, \ldots, n+1$ it holds that

$$x_j^\top \hat{w} = (1 - h_j) x_j^\top \hat{w}_{(-j)} + h_j Y_j,$$

where $h_j \in [0, 1]$ is the j-th leverage score *defined by*

$$h_j = x_j^\top (X^\top X)^{-1} x_j.$$

Hint: *use the Sherman-Morrison[3] formula, stating that for any invertible square matrix $\Sigma \in \mathbb{R}^{d \times d}$ and any vector $x \in \mathbb{R}^d$ we have*

$$\left(\Sigma + x x^\top \right)^{-1} = \Sigma^{-1} - \frac{\Sigma^{-1} x x^\top \Sigma^{-1}}{1 + x^\top \Sigma^{-1} x}.$$

[2] https://en.wikipedia.org/wiki/Leverage_(statistics)

[3] https://en.wikipedia.org/wiki/Sherman-Morrison_formula

References

Francis Bach. *Learning theory from first principles*. MIT press, 2024. URL https://www.di.ens.fr/~fbach/ltfp_book.pdf.

Jürgen Forster and Manfred K Warmuth. Relative expected instantaneous loss bounds. *Journal of Computer and System Sciences*, 64(1):76–102, 2002.