Exercises sheet

Exercise 1: Multi-label Classification

Let us consider a discrete random variables $Y : \Omega \to \mathcal{Y} = \{a_1, \dots, a_K\}$ and a continuous random variable $X : \Omega \to \mathcal{X} = \mathbb{R}^d$.

1. Recall that X has a density given by

$$p_X(x) = \sum_{k=1}^{K} f(x, a_k),$$

for some $f : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$, measurable.

2. Show that the Bayes classifier is

$$g^*(x) \in \arg\max_{a \in \mathcal{Y}} f(x, a).$$

3. In the case of binary classification where $K = 2$ and $a_1 = 0$, $a_2 = 1$, find the Bayes classifier.

Correction

1. The question might be quite technical because X is a continuous random variable and Y is a discrete random variable (to be fully rigorous, one should employ here Radon Nikodym theorem). Let us just say that by definition of joint probabilities:

$$p_X(x) = \sum_{a_k \in \mathcal{Y}} p(X = x, Y = a_k),$$

then we obtain the result denoting $f(x, a_k) \equiv p(X = x, Y = a_k)$.

2. We only saw the binary class example in the course, let us first show that the Bayes rule for the misclassification error $l(z, y) = \mathbb{1}_{z \neq y}$, writes:

$$f^*(x) = \arg\max_{a \in \mathcal{Y}} P(Y = a \mid X = x).$$

Given a decision function $f : \mathcal{X} \to \mathcal{Y}$, let us compute:

$$\begin{aligned}
R(f) &= \mathbb{E}[l(f(X), Y)] = \mathbb{E}[\mathbb{1}_{f(X) \neq Y}] = \mathbb{E}[\mathbb{E}[\mathbb{1}_{f(X) \neq Y} \mid X]] = \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}[\mathbb{1}_{f(X)=a_k} \mathbb{1}_{Y \neq a_k} \mid X]\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}[\mathbb{1}_{f(X)=a_k}] \mathbb{E}[\mathbb{1}_{Y \neq a_k} \mid X]\right] = \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{P}(f(X) = a_k)\left(1 - \mathbb{P}(Y = a_k \mid X)\right)\right] \\
&= \sum_{k=1}^{K} \mathbb{P}(f(X) = a_k) - \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{1}_{f(X)=a_k} \mathbb{P}(Y = a_k \mid X)\right] \\
&\geq \mathbb{E}[1 - \mathbb{P}(Y = f^*(X) \mid X)] = \mathbb{P}(Y \neq f^*(X)) = \mathbb{E}[\mathbb{1}_{Y \neq f^*(X)}] = R(f^*)
\end{aligned}$$

since:

- $\sum_{k=1}^{K} \mathbb{P}(f(X) = a_k) = 1$
- $\forall k \in [K]$: $\mathbb{P}(Y = f^*(X) \mid X) \geq \mathbb{P}(Y = a_k \mid X)$.

The fact that $R(f) \geq R(f^*)$ for any decision function f exactly means that f^* is the Bayes rule. Now, recalling that for any $a \in \mathcal{Y}$ and $x \in \mathbb{R}^d$ such that $p_X(x) > 0$, we have

$$P(Y = a \mid X = x) = \frac{f(x, a)}{p_X(x)},$$

one can conclude that:

$$g^*(x) = \arg\max_{a \in \mathcal{Y}} f(x, a) = \arg\max_{a \in \mathcal{Y}} P(Y = a \mid X = x) p_X(x) = \arg\max_{a \in \mathcal{Y}} P(Y = a \mid X = x) = f^*(x).$$

3. In the binary classification case if $a_1 = 0$ and $a_2 = 1$ we retrieve the Bayes rule given in the course:

$$g^*(x) = \arg\max_{a \in \{0,1\}} f(x, a) = \mathbb{1}_{\{f(x,1) > f(x,0)\}} = \mathbb{1}_{\mathbb{P}(Y=1 \mid X=x) \geq \frac{1}{2}},$$

since $\mathbb{P}(Y = 1 \mid X = x) + \mathbb{P}(Y = 0 \mid X = x) = 1$ and therefore:

$$\mathbb{P}(Y = 1 \mid X = x) \geq \frac{1}{2} \qquad \Longleftrightarrow \qquad \mathbb{P}(Y = 0 \mid X = x) \leq \mathbb{P}(Y = 1 \mid X = x).$$

Exercise 2: Non symmetric classification

We consider the binary classification problem where $Y \sim \mathrm{B}(p)$ and

$$X \mid Y = 0 \sim \mathcal{U}([0, 1/2]),$$
$$X \mid Y = 1 \sim \mathcal{U}([0, 1]).$$

1. Determine the cumulative distribution function (CDF) of X and its density p_X.
2. For any $x \in [0, 1]$, compute $\mathbb{E}[Y \mathbb{1}_{X \leq x}]$.
3. Show that, for any $x \in [0, 1]$,

$$\mathbb{E}[Y 1_{X \leq x}] = \int_0^x \eta^*(u) p_X(u) du,$$

where $\eta_P^*(x) = \mathbb{E}_P[Y \mid X = x]$ is the regression function.

4. Determine the conditional law of Y given $X = x$ and find the form of the Bayes classifier.

Correction

1. X is supported on $[0, 1]$, so its CDF F_X satisfies $F_X(x) = 0$ if $x < 0$ and $F_X(x) = 1$ if $x > 1$. For $x \in [0, 1]$:

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \leq x \mid Y = 0)\mathbb{P}(Y = 0) + \mathbb{P}(X \leq x \mid Y = 1)\mathbb{P}(Y = 1).$$

One can then compute:

- $\forall x \in [0, 1]$: $\mathbb{P}(X \leq x \mid Y = 1) = \int_0^x du = x$
- $\forall x \in [0, \frac{1}{2}]$: $\mathbb{P}(X \leq x \mid Y = 0) = \int_0^x 2du = 2x$ and $\forall x \in [\frac{1}{2}, 1]$: $\mathbb{P}(X \leq x \mid Y = 0) = 1$

Therefore:

- $\forall x \in [0, \frac{1}{2}]$: $F_X(x) = 2x(1 - p) + xp = (2 - p)x$

- $\forall x \in [\frac{1}{2}, 1]$: $F_X(x) = 1 - p + xp$,

Putting everything together, one obtains:

$$F_X(x) = (2-p)x1_{[0,1/2]}(x) + (1-p+xp)1_{[1/2,1]}(x) \quad \text{and} \quad p_X(x) = (2-p)1_{[0,1/2]}(x) + p1_{[1/2,1]}(x),$$

where the density p_X is obtain from a simple differentiation of the cumulative distribution function.

2. The possible values for $Y1_{X\leq x}$ are 1 and 0, therefore:

$$\mathbb{E}[Y\mathbb{1}_{X\leq x}] = \mathbb{P}(Y\mathbb{1}_{X\leq x} = 1) = \mathbb{P}(Y = 1, X \leq x) = \mathbb{P}(X \leq x \mid Y = 1)\mathbb{P}(Y = 1) = xp$$

3. $\mathbb{E}[Y\mathbb{1}_{X\leq x}] = \mathbb{E}[\mathbb{E}[Y\mathbb{1}_{X\leq x} \mid X]] = \mathbb{E}[\mathbb{1}_{X\leq x}\mathbb{E}[Y \mid X]]$
$$= \mathbb{E}[\mathbb{1}_{X\leq x}\eta^*(X)] = \int_0^x \mathbb{1}_{u\leq x}\eta^*(u)p_X(u)du \quad = \int_0^x \eta^*(u)p_X(u)du.$$

4. First note that:

$$\eta^*(x) = \mathbb{P}(Y = 1 \mid X = x),$$

and the Bayes rule of binary classification then writes:

$$g^*(x) = \mathbb{1}_{\eta^*(x)>1/2}.$$

Second, differentiating the two expressions of the mapping $x \mapsto \mathbb{E}[Y\mathbb{1}_{X\leq x}]$ given in questions 2 and 3 provides the identity:

$$\eta^*(x)p_X(x) = p.$$

Thus:

$$\eta^*(x) = \begin{cases} \dfrac{p}{2-p} & \text{if } x \in [0, 1/2] \\ 1 & \text{if } x \in [1/2, 1]. \end{cases}$$

The Bayes classifier then expresses:

$$g^*(x) = \begin{cases} \mathbb{1}_{\frac{p}{2-p}>\frac{1}{2}} & \text{if } x \in [0, 1/2] \\ 1 & \text{if } x \in [1/2, 1]. \end{cases}$$

Exercise 3: Least Squares, Ridge, and Lasso in Dimension 1

Given two random variables $x : \Omega \to \mathbb{R}$ and $\varepsilon : \Omega \to \mathbb{R}$ we consider the random variable:

$$y = \beta^* x + \epsilon \tag{1}$$

for a given $\beta^* \in \mathbb{R}$ that we will try to estimate. The goal of this exercise is to compare, given a data set $((x_1, y_1), \ldots, (x_1, y_1)) \in (\mathbb{R}^2)^n$, the least squares estimator:

$$\hat{\beta}^{(MC)} \in \arg\min_{\beta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n (y_i - \beta x_i)^2,$$

with the ridge estimator

$$\hat{\beta}_\lambda^{(R)} \in \arg\min_{\beta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n (y_i - \beta x_i)^2 + \gamma\|\beta^{(R)}\|^2,$$

and the Lasso estimator

$$\hat{\beta}_\lambda^{(L)} \in \arg\min_{\beta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n (y_i - \beta x_i)^2 + \gamma\|\beta^{(R)}\|_1.$$

1. Write the expression of the least squares estimator $\hat{\beta}^{(MC)}$ in terms of $\{(x_i, y_i), i = 1, \dots, n\}$. Compute the bias and variance of this estimator.

2. Write the minimization problem that the ridge estimator must solve in this framework and compute its bias, variance, and quadratic risk.

3. Give an expression for the point $x^* \in \mathbb{R}$ where the minimum of the following function is reached:

$$f(x) = a|x| + bx^2 + cx, \quad x \in \mathbb{R}, \text{ with } a, b > 0 \text{ and } c \in \mathbb{R}.$$

   Show that:

$$x^* = -\frac{c}{2b}\left(1 - \frac{a}{|c|}\right)_+.$$

4. Compute the solution to the minimization problem for the Lasso estimator.

Correction

1. Let us denote $X = (x_1, \dots, x_n) \in \mathbb{R}^{1\times n}$, $Y = (y_1, \dots, y_n) \in \mathbb{R}^{n\times 1}$ and $E = (\varepsilon_1, \dots, \varepsilon_n) \in \mathbb{R}^{n\times 1}$ (X is a row vector and Y, E are two column vectors). Since $XX^T = \|X\|^2 > 0$, the objective function has a unique minimum:

$$\hat{\beta}^{(MC)} = (XX^T)^{-1}XY = \frac{XY}{\|X\|^2}.$$

   One can then compute the expectation:

$$\mathbb{E}[\hat{\beta}^{(MC)}] = \mathbb{E}\left[\frac{X(X^T\beta^* + E)}{\|X\|^2}\right] = \beta^*$$

   thus the bias is equal to zero.

   Besides, the variance computes:

$$\mathbb{V}[\hat{\beta}^{(MC)}] = \mathbb{V}[\hat{\beta}^{(MC)} - \beta^*] \quad = \mathbb{V}\left[\frac{XE}{\|X\|^2}\right] = \mathbb{E}\left[\frac{XEE^TX^T}{\|X\|^4}\right] \quad = \sigma\mathbb{E}\left[\frac{XX^T}{\|X\|^4}\right] = \sigma\mathbb{E}\left[\frac{1}{\|X\|^2}\right]$$

2. We have seen in the course that the Ridge regression admits a unique solution that writes:

$$\hat{\beta}^{(R)}_\gamma = \frac{1}{n}\frac{XY}{\frac{1}{n}\|X\|^2 + \gamma} = \frac{XY}{\|X\|^2}\left(1 - \frac{\gamma}{\frac{1}{n}\|X\|^2 + \gamma}\right) = \hat{\beta}^{(MC)}\left(1 - \frac{\gamma}{\frac{1}{n}\|X\|^2 + \gamma}\right).$$

   The expectation is:

$$\mathbb{E}\left[\hat{\beta}^{(R)}_\gamma\right] = \mathbb{E}\left[\frac{1}{n}\frac{X(X^T\beta^* + E)}{\frac{1}{n}\|X\|^2 + \gamma}\right] = \mathbb{E}\left[\frac{1}{n}\frac{\|X\|^2\beta^*}{\frac{1}{n}\|X\|^2 + \gamma}\right] = -\mathbb{E}\left[\frac{\gamma}{\frac{1}{n}\|X\|^2 + \gamma}\right]\beta^* + \beta^*$$

   the bias thus expresses:

$$\text{Bias}^{(R)} = -\mathbb{E}\left[\frac{\gamma}{\frac{1}{n}\|X\|^2 + \gamma}\right]\beta^*,$$

   note in passing that $\mathbb{E}\left[\frac{\hat{\beta}^{(MC)}}{\frac{1}{n}\|X\|^2+\gamma}\right] = \mathbb{E}\left[\frac{1}{\frac{1}{n}\|X\|^2+\gamma}\right]\beta^*$ (although $\hat{\beta}^{(MC)}$ and $\|X\|^2$ are two *dependent* random variables). Let us recall the identity $\frac{X(X^T\beta^*+E)}{\frac{1}{n}\|X\|^2+\gamma} - \beta^* = \frac{E-\gamma\beta^*}{\frac{1}{n}\|X\|^2+\gamma}$, the independence between

X and E, that $\mathbb{E}[E] = 0$ and let us denote $\sigma \equiv \mathbb{E}[\epsilon_i^2]$ (then $\mathbb{E}[EE^T] = \sigma I_n$). The variance expresses:

$$\begin{aligned}\mathbb{V}[\hat{\beta}_\gamma^{(R)}] = \mathbb{V}[\hat{\beta}_\gamma^{(R)} - \beta^*] &= \mathbb{E}\left[\left(\frac{\frac{1}{n}XE - \gamma\beta^*}{\frac{1}{n}\|X\|^2 + \gamma}\right)^2\right] - \gamma^2\mathbb{E}\left[\frac{1}{\frac{1}{n}\|X\|^2 + \gamma}\right]^2 \beta^{*2} \\ &= \beta^{*2}\mathbb{E}\left[\left(\frac{\gamma}{\frac{1}{n}\|X\|^2 + \gamma}\right)^2\right] + \frac{\sigma}{n}E\left[\frac{\frac{1}{n}\|X\|^2}{(\frac{1}{n}\|X\|^2 + \gamma)^2}\right] - \gamma^2\mathbb{E}\left[\frac{1}{\frac{1}{n}\|X\|^2 + \gamma}\right]^2 \beta^{*2} \\ &= \gamma^2\beta^{*2}\mathbb{E}\left[\frac{1}{(\frac{1}{n}\|X\|^2 + \gamma)^2}\right] + \frac{\sigma}{n}E\left[\frac{\frac{1}{n}\|X\|^2}{(\frac{1}{n}\|X\|^2 + \gamma)^2}\right] - \gamma^2\mathbb{E}\left[\frac{1}{\frac{1}{n}\|X\|^2 + \gamma}\right]^2 \beta^{*2}\end{aligned} \tag{2}$$

The mean squared error (MSE) for the estimation of β^* (in exercise 4, we will express the mean square error for the estimation of Y) is given by:

$$\text{MSE}(\hat{\beta}^{(R)}) = \mathbb{V}[\hat{\beta}_\gamma^{(R)}] + \text{Bias}^{(R)2} = \frac{\sigma}{n}E\left[\frac{\frac{1}{n}\|X\|^2}{(\frac{1}{n}\|X\|^2 + \gamma)^2}\right] + \gamma^2\mathbb{E}\left[\frac{1}{n^2}\frac{\|X\|^4}{(\frac{1}{n}\|X\|^2 + \gamma)^2}\right]\beta^{*2}$$

3. Note first that f is strictly convex as the sum of strictly convex functions. It has a left and right derivative at every point (equal for $x \neq 0$). Its subdifferential is: Note that f is differentiable on $\mathbb{R}_-^*$ and $\mathbb{R}_+^*$ and we have the expressions:

$$\forall x > 0: \quad f'(x) = \{a + 2bx + c\} \qquad \text{and} \qquad \forall x < 0: \quad f'(x) = \{-a + 2bx + c\}.$$

If the minimum is reached on:

- $x > 0$, then $a + 2bx + c = 0$ and therefore:
$$x = \frac{-a - c}{2b},$$
which implies in particular $c < -a$ since $x, b > 0$.
- $x < 0$, then $-a + 2bx + c = 0$ and therefore:
$$x = \frac{-a - c}{2b},$$
which implies in particular $c > a$ since $x, b < 0$.
- $x = 0$, then $\forall t > 0: f'(-t) \leq 0 \leq f'(t)$, and one can let t tend to zero to obtain the inequality:
$$-a + c \leq 0 \leq a + c \qquad \Longleftrightarrow \qquad -a \leq c \leq a.$$

Thus, relying on the identity (recall that $a > 0$):

$$\left(1 - \frac{a}{|c|}\right)_+ = \begin{cases} \dfrac{a+c}{c} & \text{if } c < -a \\ \dfrac{c-a}{c} & \text{if } c > a \\ 0 \text{ if } -a \leq c \leq a, \end{cases}$$

which allows to retrieve the solution given in the question.

4. The Lasso estimator solves the minimization problem: The estimator minimizes the objective function:

$$\tilde{f}(\beta) = (Y - X^T\beta)^\top(Y - X^T\beta) + \gamma|\beta| = \|Y\|^2 - 2XY\beta + \|X\|^2\beta^2 + \gamma|\beta|.$$

This boils down to minimizing the mapping f introduced in question 3. with $a = \gamma$, $b = \|X\|^2$, $c = -2XY$, plus a constant $Y^\top Y$ (irrelevant for optimization). Therefore, the solution is:

$$\hat{\beta} = \frac{Y^\top X}{\|X\|^2}\left(1 - \frac{\gamma}{2|XY|}\right)_+ = \beta^{(MC)}\left(1 - \frac{\gamma}{2|XY|}\right)_+.$$

Exercise 4: Properties of the Ridge Estimator

We consider here the same model (1) as in the previous exercice but this time $X : \Omega \to \mathbb{R}^d$ and $\beta^* \in \mathbb{R}^d$. The Ridge estimator, given regularizing coefficient $\gamma > 0$, expresses:

$$\hat{\beta}_\gamma^{(R)} = \frac{1}{n}(X^\top X + \gamma I)^{-1} X^\top Y.$$

1. Show that the estimator:

$$\hat{\beta}^{(R')} \equiv \underset{\beta \in \mathbb{R}^d, \|\beta\| \leq M_\gamma}{\arg\min} \left(\sum_{i=1}^{n} (Y_i - X_i \beta)^2 \right),$$

   with $M_\gamma = \frac{1}{n}\|QX^\top Y\|$ and $Q \equiv (X^\top X + \gamma I)^{-1}$ is equal to $\hat{\beta}^{(R)}$.

2. Express the squared norm of the bias of $\hat{\beta}_\gamma^{(R)}$ in terms of the eigenvalues $\lambda_1, \ldots, \lambda_d$ (with multiplicities) of $X^\top X$:

$$B_\gamma^{(R)} := \|\mathbb{E}[\hat{\beta}_\gamma^{(R)}] - \beta^*\|^2.$$

3. Express the variance:

$$V_\gamma^{(R)} = \mathbb{E}\left[\|\hat{\beta}_\gamma^{(R)} - \mathbb{E}[\hat{\beta}_\gamma^{(R)}]\|^2 \right],$$

   in terms of the noise variance σ^2 and the eigenvalues $\lambda_1, \ldots, \lambda_d$.

Correction

1. Let us introduce the mapping $f_0 : \beta \mapsto \sum_{i=1}^n (Y_i - X_i\beta)^2$. The constrained problem ensures that $\|\beta^{(R')}\| \leq M_\lambda = \|\beta_\gamma^{(R)}\|$. Besides since $\beta_\gamma^{(R)}$ satisfies the constraint, on also has the inequality $f_0(\beta^{(R')}) \leq f_0(\beta_\gamma^{(R)})$. Therefore:

$$f_0(\beta^{(R')}) + \gamma\|\beta^{(R')}\|^2 \leq f_0(\beta_\gamma^{(R)}) + \gamma\|\beta_\gamma^{(R)}\|^2$$

   By uniqueness of the solution of the Ridge regression problem, that implies that $\beta^{(R')} = \beta^{(R)}$.

2. We compute:

$$\mathbb{E}[\hat{\beta}_\gamma^{(R)}] = \frac{1}{n}\mathbb{E}[QXY] = \frac{1}{n}\mathbb{E}[QX(X^T\beta^* + E)] = \beta^* - \gamma\mathbb{E}[Q]\beta^*.$$

   Therefore $\text{Bias}_\gamma^2 = \gamma^2 \beta^{*T}\mathbb{E}[Q]^2\beta^*$ and if we denote P, the orthogonal matrix that diagonalizes $X^\top X$ as $P^\top X^\top X P = D$, with $D = \text{diag}(\lambda_1, \ldots, \lambda_d)$, one obtains the form:

$$\text{Bias}_\gamma^2 = \gamma^2 \beta^{*T}\mathbb{E}[P^{-1}(D + \gamma I)^{-1}P]^2\beta^*,$$

   which is not very helpful.

3. Let us compute:

$$\begin{aligned}
\mathbb{V}_\gamma[\beta_\gamma^{(R)}] = \mathbb{V}_\gamma[\beta_\gamma^{(R)} - \beta^*] &= \mathbb{E}\left[\left\| -\gamma Q\beta^* + \frac{1}{n}QXE \right\|^2 \right] - \left\| \mathbb{E}\left[-\gamma Q\beta^* + \frac{1}{n}QXE \right] \right\| \\
&= \gamma^2 \beta^{*T}\mathbb{E}\left[Q^2\right]\beta^* + \frac{\sigma}{n^2}\mathbb{E}\left[\text{Tr}(X^T QQX)\right] - \text{Bias}_\gamma^2 \\
&= \gamma^2 \beta^{*T}\left(\mathbb{E}\left[Q^2\right] - \mathbb{E}\left[Q\right]^2 \right)\beta^* + \frac{\sigma}{n}\mathbb{E}\left[\text{Tr}(Q - \gamma Q^2)\right].
\end{aligned}$$

   One can check that it is possible to retrieve (2) in the case $p = 1$ (then $Q = \frac{1}{\|X\|^2/n + \gamma}$).

Exercise 5: Square of the resolvent

This is a difficult problem (too difficult for a final exam exercise). Let us consider again the model (1):

$$y = x^T \beta^* + \varepsilon,$$

with $x : \Omega \to \mathbb{R}^p$, $\varepsilon : \Omega \to \mathbb{R}$, two independent variables and $\beta^* \in \mathbb{R}^p$ a deterministic vector.

1. Given a train data set $X = ((x_1, y_1), \ldots, (x_n, y_n))$ and a test data (x, y), express the train MSE and the test MSE for the estimation of Y with the Ridge regression as a function of $X = (x_1, \ldots, x_n) \in \mathbb{R}^{p \times p}$, x and β^*. For that, introduce the resolvent matrices $Q \equiv (\gamma I_p + \frac{1}{n} X X^T)^{-1}$ and $Q \equiv (\gamma I_n + \frac{1}{n} X^T X)^{-1}$

2. We will now try to estimate $\mathbb{E}[Q^2]$. Recall from the course the notation:

$$\forall i \in [n] : \qquad \Lambda_i \equiv 1 - \frac{1}{n} x_i^T Q_{-i} x_i$$

   where $Q_{-i} = (\gamma I_p + \frac{1}{n} X_{-i} X_{-i}^T)^{-1}$ and $X_{-i} = (x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n) \in \mathcal{M}_{p,n}$ and the identities:

$$Q = Q_{-i} + \frac{1}{n} \frac{Q_{-i} x_i x_i^T Q_{-i}}{\Lambda_i} \qquad \text{and} \qquad Q x_i = \frac{Q_{-i} x_i}{\Lambda_i}$$

   We further introduced the deterministic matrix:

$$\forall \Delta \in \mathbb{R} : \qquad \tilde{Q}^{\Delta} = \left(\gamma I_p + \frac{\Sigma}{\Delta} \right)^{-1} \qquad \text{with } \Sigma = \mathbb{E}[x_i x_i^T]$$

   and the scalar $\tilde{\Lambda} \in \mathbb{R}$ solution to:

$$\tilde{\Lambda} = \frac{1}{n} \mathrm{Tr}(\Sigma \tilde{Q}^{\tilde{\Lambda}}),$$

   To be able to set concentration results, we assume, as in the course that the matrix X has independent columns and that it is a λ-Lipschitz transformation of a Gaussian vector $Z \sim \mathcal{N}(0, I_q)$. Assuming that $\frac{p}{n}$ and $\|\Sigma\|$ are both bounded with a certain constant independent of p, n, q, first bound without justifications the following probabilities (employing some constants $C, c > 0$ independent with n, p, q):

   - $\mathbb{P}\left(\left| \Lambda_i - \tilde{\Lambda} \right| \geq t \right)$
   - $\mathbb{P}\left(\left| u^T Q_{-i} x_i \right| \geq t \right)$
   - $\mathbb{P}\left(\left| x_i^T Q_{-i} \Sigma \tilde{Q} u \right| \geq t \right)$
   - $\mathbb{P}\left(\left| \frac{1}{n} x_i^T Q_{-i} x_i - \frac{1}{n} \mathrm{Tr}(\mathbb{E}[Q_{-i}] \Sigma) \right| \geq t \right)$

3. Given a deterministic vector $u \in \mathbb{R}^p$ and a deterministic matrix $A \in \mathbb{R}^{p \times p}$, such that $\|u\| \leq 1$, $\|A\| \leq O(1)$, estimate:

$$u^T Q A (Q - \tilde{Q}^{\tilde{\Lambda}}) u,$$

   and deduce that:

$$\mathbb{E}[u^T Q A Q u] = u^T \tilde{Q}^{\tilde{\Lambda}} A \tilde{Q}^{\tilde{\Lambda}} u - \frac{\mathrm{Tr}(\Sigma \tilde{Q} A \tilde{Q})}{\tilde{\Lambda}^2 n} u^T \mathbb{E}[Q \Sigma Q] u + O\left(\frac{1}{\sqrt{n}} \right)$$

4. Playing on the value of A, give an estimate of $\mathbb{E}[Q^2]$ and $\mathbb{E}[Q \Sigma Q]$ and deduce an estimation of the train and test MSE of the Ridge regression.

Correction

1. Let us introduce β solution to the minimization problem:

$$\beta = \underset{\beta \in \mathbb{R}^d}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i \beta)^2 + \gamma \|\beta\|^2$$

then β expresses:

$$\beta = \frac{1}{n} QXY,$$

and one can defined the train and the test MSE as:

$$\text{MSE}_{\text{tr}} = \frac{1}{n} \mathbb{E}\left[\|X^T \beta - Y\|^2\right] \qquad \text{and} \qquad \text{MSE}_{\text{tst}} = \mathbb{E}\left[\|x^T \beta - y\|^2\right],$$

where x is an independent copy of any column of X and $y = x^T \beta^* + \varepsilon$ as described by the model. With the notations we introduced, recall that:

$$QX = X^T \check{Q} \qquad \text{and} \qquad \frac{1}{n} QXX^T = I_p - \gamma Q$$

With the notation $E = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$ and $\eta = \mathbb{E}[\varepsilon^2]$, one can then express as seen during the course:

- $$\begin{aligned} \text{MSE}_{\text{tr}} &= \frac{\gamma^2}{n} \mathbb{E}\left[Y^T \check{Q}^2 Y\right] \\ &= \frac{\gamma^2}{n} \mathbb{E}\left[\beta^* X \check{Q}^2 X^T \beta\right] + \frac{\gamma^2}{n} \mathbb{E}\left[\text{Tr}(E^T \check{Q}^2 E)\right] \\ &= \beta^{*T} \mathbb{E}\left[Q - \gamma^2 Q^2\right] \beta + \frac{\gamma^2 \eta}{n} \text{Tr}(\mathbb{E}[\check{Q}^2]) \end{aligned}$$
- $$\begin{aligned} \text{MSE}_{\text{tst}} &= \mathbb{E}\left[\left(\frac{1}{n} x^T QX(X^T \beta^* + E) - x^T \beta^* - \varepsilon\right)^2\right] \\ &= \mathbb{E}\left[\left(x^T (I_p - \gamma Q)\beta^* - x^T \beta^*\right)^2\right] + \mathbb{E}\left[\left(\frac{1}{n} x^T QXE\right)^2\right] + \mathbb{E}\left[\varepsilon^2\right] \\ &= \mathbb{E}\left[\left(x^T (I_p - \gamma Q)\beta^* - x^T \beta^*\right)^2\right] + \frac{\eta}{n^2} \text{Tr}\left(\mathbb{E}\left[QXX^T Q\right] \Sigma\right) + \eta \\ &= \gamma^2 \mathbb{E}\left[\beta^{*T} Q \Sigma Q \beta^*\right] + \frac{\eta}{n} \text{Tr}\left(\mathbb{E}\left[Q^2 - \gamma Q\right] \Sigma\right) + \eta \end{aligned}$$

2. We can bound from the course:
   - $\mathbb{P}\left(\left|\Lambda_i - \tilde{\Lambda}\right| \geq t\right) \leq Ce^{-c\sqrt{n}t}$
   - $\mathbb{P}\left(\left|u^T Q_{-i} x_i\right| \geq t\right) \leq Ce^{-ct}$
   - $\mathbb{P}\left(\left|x_i^T Q_{-i} \Sigma \tilde{Q} u\right| \geq t\right) \leq Ce^{-ct}$
   - $\mathbb{P}\left(\left|\frac{1}{n} x_i^T Q_{-i} x_i - \frac{1}{n} \text{Tr}(\mathbb{E}[Q_{-i}]\Sigma)\right| \geq t\right) \leq Ce^{-cnt^2} + Ce^{-ct}$.

3. Now, employing as in the course the identities:

$$Qx_i = \frac{Q_{-i} x_i}{\Lambda_i} \qquad \text{and} \qquad Q = Q_{-i} + \frac{Q_{-i} x_i x_i^T Q_{-i}}{\Lambda_i}.$$

and the resolvent identity, one can estimate:

$$\begin{aligned}\mathbb{E}\left[u^TQA\left(Q-\tilde{Q}^{\tilde{\Lambda}}\right)u\right] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[u^TQA\tilde{Q}^{\tilde{\Lambda}}\left(\frac{\Sigma}{\tilde{\Lambda}}-x_ix_i^T\right)Qu\right]\\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{u^TQA\tilde{Q}^{\tilde{\Lambda}}\Sigma Qu}{\tilde{\Lambda}}\right] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{u^TQA\tilde{Q}^{\tilde{\Lambda}}x_ix_i^TQ_{-i}u}{\Lambda}\right]\\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{u^TQ_{-i}A\tilde{Q}^{\tilde{\Lambda}}\Sigma Q_{-i}u}{\tilde{\Lambda}}\right] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{u^TQ_{-i}A\tilde{Q}^{\tilde{\Lambda}}x_ix_i^TQ_{-i}u}{\tilde{\Lambda}}\right]\\ &\quad - \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\frac{u^TQ_{-i}x_ix_i^TQ_{-i}A\tilde{Q}^{\tilde{\Lambda}}x_ix_i^TQ_{-i}u}{\tilde{\Lambda}\Lambda_i}\right] + O\left(\frac{\kappa_z}{\sqrt{n}}\right)\end{aligned}$$

where we replaced a Λ_i with $\tilde{\Lambda}$ and added a small error of order $O(1/\sqrt{n})$ since we know that Λ_i is almost constant as pictured in the previous question. The independence between x_i and Q_{-i} provides:

$$\mathbb{E}\left[\frac{u^TQ_{-i}A\tilde{Q}^{\tilde{\Lambda}}\Sigma Q_{-i}u}{\tilde{\Lambda}}\right] = \mathbb{E}\left[\frac{u^TQ_{-i}A\tilde{Q}^{\tilde{\Lambda}}x_ix_i^TQ_{-i}u}{\tilde{\Lambda}}\right].$$

Besides, the same concentration as the one of Λ_i happens for $\frac{1}{n}x_i^TQ_{-i}A\tilde{Q}^{\tilde{\Lambda}}x_i$ that we can replace with $\frac{1}{n}\mathrm{Tr}(\Sigma\tilde{Q}^{\tilde{\Lambda}}A\tilde{Q}^{\tilde{\Lambda}})$ which finally leads to:

$$\begin{aligned}\mathbb{E}\left[u^TQA\left(Q-\tilde{Q}^{\tilde{\Lambda}}\right)u\right] &= -\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{u^TQ_{-i}x_ix_i^TQ_{-i}u}{\tilde{\Lambda}\Lambda_i}\right]\frac{1}{n}\mathrm{Tr}(\Sigma\tilde{Q}^{\tilde{\Lambda}}A\tilde{Q}^{\tilde{\Lambda}}) + O\left(\frac{\kappa_z}{\sqrt{n}}\right)\\ &= -\frac{1}{n}\mathrm{Tr}(\Sigma\tilde{Q}^{\tilde{\Lambda}}A\tilde{Q}^{\tilde{\Lambda}})\frac{u^T\mathbb{E}[Q\Sigma Q]u}{\tilde{\Lambda}^2} + O\left(\frac{\kappa_z}{\sqrt{n}}\right),\end{aligned}$$

which is exactly the looked for result.

4. Taking $A=\Sigma$ in the above result provides:

$$u^T\mathbb{E}[Q\Sigma Q]u = u^T\tilde{Q}^{\tilde{\Lambda}}\Sigma\tilde{Q}^{\tilde{\Lambda}}u - \frac{\mathrm{Tr}(\Sigma\tilde{Q}\Sigma\tilde{Q})}{\tilde{\Lambda}^2}u^T\mathbb{E}[Q\Sigma Q]u + O\left(\frac{1}{\sqrt{n}}\right),$$

and therefore:

$$u^T\mathbb{E}[Q\Sigma Q]u = \frac{u^T\tilde{Q}^{\tilde{\Lambda}}\Sigma\tilde{Q}^{\tilde{\Lambda}}u}{1+\frac{1}{\tilde{\Lambda}^2 n}\mathrm{Tr}(\Sigma\tilde{Q}\Sigma\tilde{Q})}$$

one can then inject this new value in the initial estimate to finally obtain for any deterministic matrix $A\in\mathbb{R}^{p\times p}$:

$$\mathbb{E}[u^TQAQu] = u^T\tilde{Q}^{\tilde{\Lambda}}A\tilde{Q}^{\tilde{\Lambda}}u - \mathrm{Tr}(\Sigma\tilde{Q}A\tilde{Q})\frac{u^T\tilde{Q}^{\tilde{\Lambda}}\Sigma\tilde{Q}^{\tilde{\Lambda}}u}{\tilde{\Lambda}^2 n+\mathrm{Tr}(\Sigma\tilde{Q}\Sigma\tilde{Q})} + O\left(\frac{1}{\sqrt{n}}\right)$$

Most of the terms appearing in the mean square error formula can now be estimated, one is just left to estimating $\frac{1}{n}\mathbb{E}[\mathrm{Tr}(\check{Q})]$. Note first that one can use the formula:

$$\check{Q} = \frac{1}{\gamma}(\gamma\check{Q}-I_n)+I_n = \gamma I_n - \frac{1}{n\gamma}\check{Q}X^TX = \gamma I_n - \frac{1}{n\gamma}X^TQX$$

and noting that:

$$\frac{1}{n}\mathbb{E}\left[\mathrm{Tr}(X^TQX)\right] = \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[x_i^TQx_i\right] = \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\frac{x_i^TQ_{-i}x_i}{\Lambda_i}\right] = \frac{1}{n\tilde{\Lambda}}\mathrm{Tr}(\Sigma\tilde{Q}^{\tilde{\Lambda}}),$$

one can finally estimate:

$$\begin{aligned}\frac{1}{n}\mathbb{E}[\mathrm{Tr}(\check{Q}^2)] &= \mathbb{E}\left[\mathrm{Tr}(\gamma I_n - \frac{1}{n\gamma}X^TQX)^2\right] = \gamma^2 + \frac{2}{n^2\gamma}\mathbb{E}\left[\mathrm{Tr}(X^TQX)\right] + \frac{1}{n^2\gamma^2}\mathbb{E}\left[\mathrm{Tr}(X^TQXX^TQX)\right] \\ &= \gamma^2 + \frac{2}{n^2\gamma\tilde{\Lambda}}\mathrm{Tr}(\Sigma\tilde{Q}^{\tilde{\Lambda}}) + \frac{1}{n^2\gamma^2}\mathbb{E}\left[\mathrm{Tr}(X^TQ^2X)\right] \\ &= \gamma^2 + \frac{2}{n\gamma\tilde{\Lambda}}\mathrm{Tr}(\Sigma\tilde{Q}^{\tilde{\Lambda}}) + \frac{1}{n\gamma^2\tilde{\Lambda}}\mathrm{Tr}(\Sigma\mathbb{E}\left[Q^2\right]).\end{aligned}$$

One then has all the elements to compute the train and test Mean square error as they are expressed in Question 2.