Mathematical Underpinnings Lab 2 06.03.2024

Task 1 (population risk, empirical risk, excess risk)

We have an i.i.d. sample $(X_i, Y_i)_{i=1}^n$ from the following distribution:

$$X_i \sim \mathcal{N}(0,1)$$
$$\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$$

and $Y_i = X_i + \varepsilon_i$, where $i = 1, 2, \ldots, n$. We denote the joint distribution of (X, Y) as p.

a) Calculate the population risk $\mathbb{E}_p \mathcal{L}(f(X), Y)$ for a quadratic loss function $\mathcal{L}$, a family of linear functions f with a parameter a, i.e. $f(x) = ax$ and the distribution p. Find the argument that minimises the population risk. How the minimum value changes with σ^2? Implement a function computing the population risk.

b) Let $n = 50$ and $\sigma^2 = 1$. Generate a sample $(X_i, Y_i)_{i=1}^n$. Implement a function computing the empirical risk.

c) Visualise the previous results. Compute the argmin and the min of the population risk and empirical risk. Next, compute the value of the population risk at the point minimising the empirical risk. Mark the three points in the plot. Which point corresponds to the ERM estimator?

d) Compute the excess risk. Estimate the unconditional excess risk by generating the data $L = 200$ times and averaging excess risks. How the unconditional excess risk changes with a sample size?

e) (*) Do the tasks a)-c) for $X_i \sim \mathcal{N}(\mu_x, \sigma_x^2)$, $Y_i = a_0 X_i + b_0 + \varepsilon_i$ and $f(x) = ax + b$.

f) (*) (estimation error vs approximation error) Do the tasks a)-c) assuming that the true parameter a^* equals 1.1 and we consider a family of linear functions f with a parameter $a \in \mathbb{Z}$.