EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Mathematisch-Naturwissenschaftliche Fakultät

Fachbereich Mathematik

Prof. Dr. Andreas Prohl
Dr. Abhishek Chaudhary

Statistical Learning 2

Summer-Semester 2023 **Tübingen, 17.05.2023**

Homework 3

Problem 1. Let $D_n = \{(X_i, Y_i)\}_{i=1}^n$ be an n-sample. This example is meant to again see how relevant a proper choice of the function class $\mathcal{F}_n$ is for the 'least squares (LS)' estimator

$$m_n = \mathrm{argmin}_{f\in\mathcal{F}_n} \frac{1}{n}\sum_{i=1}^n \big|f(X_i) - Y_i\big|^2 . \tag{1}$$

a) Let $X \sim \mathcal{U}[0,1]$ be uniformly distributed. Define

$$Y = \begin{cases} -1 & \text{if } X < \frac{1}{2} \\ 1 & \text{if } X \geq \frac{1}{2} \end{cases} . \tag{2}$$

Consider a related (i.i.d.) n-sample D_n. Define the estimator

$$\widetilde{m}_n(X) = \begin{cases} Y_i & \text{if } X = X_i \\ 1 & \text{else} \end{cases} . \tag{3}$$

Show that $\widetilde{m}_n$ minimizes the above 'empirical LS functional'.

b) Show that the estimator $\{\widetilde{m}_n\}_n$ is *not consistent*, *i.e.*,

$$\mathbb{P}\big[\widetilde{m}_n(X) \neq Y\big] \not\to 0 \qquad (n \uparrow \infty) .$$

<u>**Hint**</u>: Use from *'Statistical learning I'* the 'Bayes classifier'

$$f^*(x) = \begin{cases} 1 & \text{for } \mathbb{P}[Y = 1|X = x] \geq \frac{1}{2} \\ -1 & \text{else} \end{cases} .$$

Problem 2. In the lecture, we discussed the *'uniform LLN'* for a given class of functions $\mathcal{F}_n$ — which follows from the *'uniform exponential estimate'* by Pollard. This exercise now contributes to see its role to verify consistency of the related 'LS estimator' for $\mathcal{F}_n$. For this purpose, we use the following notations:

$$\mathcal{J}(f) := \mathbb{E}\big[|f(\mathbf{X}) - Y|^2\big] \qquad \text{and} \qquad \mathcal{J}_{\mathtt{emp}}(f) := \frac{1}{n}\sum_{i=1}^n \big|f(\mathbf{X}_i) - Y_i\big|^2 \qquad \forall\, f \in \mathcal{F}_n ,$$

and define related minimizers $m, m_h \in \mathcal{F}_n$ via

$$m := \operatorname{argmin}_{f\in\mathcal{F}_n}\mathcal{J}(f) \qquad \text{and} \qquad m_n := \operatorname{argmin}_{f\in\mathcal{F}_n}\mathcal{J}_{\mathtt{emp}}(f)\,.$$

Verify the estimation $(\varepsilon > 0)$:

$$\mathbb{P}\Big[\big|\mathcal{J}(m_n) - \mathcal{J}(m)\big| \geq \varepsilon\Big] \leq \mathbb{P}\Big[\sup_{f\in\mathcal{F}_n}\big|\mathcal{J}(f) - \mathcal{J}_{\mathtt{emp}}(f)\big| \geq \frac{\varepsilon}{2}\Big] \to 0 \qquad (n\uparrow\infty)\,.$$

Hint: Add and subtract '$\mathcal{J}_{\mathtt{emp}}(m_n)$' and '$\mathcal{J}_{\mathtt{emp}}(m)$' when you consider/estimate $\mathcal{J}(m_n) - \mathcal{J}(m)$.

Date of Submission: 12.00 on 24.05.2023.