CS/ECE/STAT-861: Theoretical Foundations of Machine Learning

University of Wisconsin–Madison, Fall 2025 *Instructor: Kirthevasan Kandasamy*

Homework 2. Due 10/11/2025, 11.59 pm

Instructions:

1. Homework is due on Canvas by 11:59 pm on the due date. Please plan to submit well before the deadline. Refer to the course website for policies on late submission.
2. Homework must be typeset using appropriate software, such as LaTeX. Handwritten and scanned submissions will **not** be accepted.
3. Your solutions will be evaluated on correctness, clarity, and conciseness.
4. Unless otherwise specified, you may use any result we have already proved in class. Clearly state which result you are using.
5. Solutions to some of the problems may be found in the recommended textbook or other resources. Unless stated otherwise, you should attempt the problems on your own. You may **not** search the internet for solutions or use LLM-based tools.
6. If you use any external references, please cite them in your submission.
7. **Collaboration:** You may collaborate in groups of size up to 3 to solve problems indicated by a star ($\star$). You may not collaborate on other problems. If you collaborate, please indicate your collaborators at the beginning of the problem. Even if you collaborate, *you must write the solution in your own words*.

1 Density estimation in a Hölder class ⋆

Let $\mathcal{H}(2, L, B)$, defined below, denote the bounded second order Hölder class in $[0, 1]$. It consists of functions whose derivatives are L-Lipschitz.

$$\mathcal{H}(2, L, B) = \{f : [0, 1] \to [0, B]; \quad |f'(x_1) - f'(x_2)| \leq L|x_1 - x_2| \text{ for all } x_1, x_2 \in \mathbb{R}\}$$

Let $\mathcal{P}$ denote the set of distributions whose densities are in $\mathcal{H}(2, L, B)$. We observe n samples $S = \{X_1, \ldots, X_n\}$ drawn i.i.d from some $P \in \mathcal{P}$ and wish to estimate its density p in the L_2 loss $\Phi \circ \rho(p_1, p_2) = \|p_1 - p_2\|_2^2$. The minimax risk is

$$R_n^\star = \inf_{\widehat{p}} \sup_{p \in \mathcal{H}(2,L,B)} \mathbb{E}_S \left[\|p - \widehat{p}\|_2^2\right].$$

In this question, you will show that the minimax rate[1] for this problem is $\Theta(n^{-4/5})$.

1. **[15 pts]** *(Lower bound)* Using Fano's method, or otherwise, show that $R_n^\star \in \Omega(n^{-4/5})$.
2. **[15 pts]** *(Upper bound)* Design an estimator $\widehat{p}$ for p and bound its risk by $\mathcal{O}(n^{-4/5})$.

   Hint: If you choose to use a kernel density estimator, consider the first order Taylor expansion of p and then apply the Hölder property. You may ignore boundary correction.
3. **[4 pts]** *(High dimensional setting)* In words, briefly explain how you can extend both the upper and lower bounds for density estimation in d dimensions. The d dimensional second-order Hölder class, defined below, consists of functions whose partial derivatives are Lipschitz.

   $$\mathcal{H}(2, L, B) = \left\{f : [0, 1]^d \to [0, B]; \quad \frac{\partial f}{\partial x_i} \text{ is } L\text{–Lipschitz for all } i \in [d]\right\}.$$

   You can focus *only* on the key differences. A detailed proof is not necessary.
4. **[4 pts]** *(Lipschitz second derivatives)* In words, briefly explain how you can extend both the upper and lower bounds if the densities belonged to the third order Hölder class in one dimension, defined below:

   $$\mathcal{H}(3, L, B) = \{f : [0, 1] \to [0, B]; \quad |f''(x_1) - f''(x_2)| \leq L|x_1 - x_2| \text{ for all } x_1, x_2 \in \mathbb{R}\}$$

   Please focus *only* on the key differences. A detailed proof is not necessary.

   Hint: For the upper bound, if you choose to use a kernel density estimator, you may consider a kernel of the form $K(u) = \mathbb{1}(|u| \leq 1/2)(\alpha - \beta u^2)$ for appropriately chosen α, β.

2 Short problems on Chapter 3

1. *(What is wrong with this proof?)* We perform empirical risk minimization (ERM) in a finite hypothesis class $\mathcal{H}$ using an i.i.d dataset S of n points. Let $h^\star \in \operatorname{argmin}_{h \in \mathcal{H}} L(h)$ be an optimal classifier in the class, and let $\widehat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} \widehat{L}(h)$ minimize the empirical loss of the dataset S. Recall that we showed, with probability at least $1 - 2|\mathcal{H}|e^{-2n\epsilon^2}$, we have

   $$L(\widehat{h}) - L(h^\star) \leq L(\widehat{h}) - \widehat{L}(\widehat{h}) + \widehat{L}(h^\star) - L(h^\star) \leq 2\epsilon.$$

   This resulted in the bound $\mathbb{E}[L(\widehat{h}) - L(h^\star)] \in \tilde{\mathcal{O}}\left(\sqrt{\frac{\log(|\mathcal{H}|)}{n}}\right)$.

   A student offers the following alternative proof and claims that it is possible to bound the error without any dependence on $|\mathcal{H}|$:

[1] Recall from class that the minimax rate for a Hölder class of order β is $\mathcal{O}\left(n^{-\frac{2\beta}{2\beta+d}}\right)$ in $\mathbb{R}^d$.

(i) Let $B_1 = \{\widehat{L}(h^\star) - L(h^\star) > \epsilon\}$ denote the bad event that the empirical loss of $h^\star$ is ϵ larger than its true loss. By Hoeffding's inequality we have $\mathbb{P}(B_1) \leq e^{-2n\epsilon^2}$.

(ii) Similarly, Let $B_2 = \{L(\widehat{h}) - \widehat{L}(\widehat{h}) > \epsilon\}$ denote the bad event that the empirical loss of $\widehat{h}$ is ϵ smaller than its true loss. By Hoeffding's inequality we have $\mathbb{P}(B_2) \leq e^{-2n\epsilon^2}$.

As $\widehat{L}(\widehat{h}) \leq \widehat{L}(h^\star)$, we have,

$$L(\widehat{h}) - L(h^\star) \leq L(\widehat{h}) - \widehat{L}(\widehat{h}) + \widehat{L}(h^\star) - L(h^\star) \leq 2\epsilon$$

under the good event $G = B_1^c \cap B_2^c$ which is true with probability at least $1 - 2e^{-2n\epsilon^2}$. This result does not depend on $|\mathcal{H}|$ and even applies to infinite hypothesis classes provided there exists $h^\star$ which minimizes the loss.

[4 pts] Which sentence below best describes the mistake (if any) with this proof? State your answer with an explanation. If you believe there is a mistake, be as specific as possible as to what the mistake is.

(a) Both statement *(i)* and statement *(ii)* are incorrect.

(b) Only statement *(i)* is incorrect. Statement *(ii)* is correct.

(c) Only statement *(ii)* is incorrect. Statement *(i)* is correct.

(d) Both statements are correct. There is nothing wrong with this proof.

2. *(Uniform convergence via finite covers for classification)* In class, we saw that it may not be possible to show uniform convergence for "non-smooth" losses, such as the 0-1 loss for binary classification, via a covering argument. In this question, you will show this rigorously via threshold classifiers.

   Consider the following setting for binary classification, where for all $X, Y \sim P$, we have $X \in [0, 1]$ and $Y \in \{0, 1\}$. Consider the following hypothesis class,

   $$\mathcal{H} = \{h_\theta; h_\theta(x) = \mathbb{1}(x \geq \theta); \theta \in [0, 1]\}.$$

   [7 pts] Prove that[2] for all $\epsilon > 0$, we cannot construct a *finite* $C_\epsilon \subset [0, 1]$ such that the following is true: for every $\theta \in [0, 1]$, there exists $[\theta] \in C_\epsilon$ such that $\forall\, x, y \;\; |\ell(h_\theta, (x, y)) - \ell(h_{[\theta]}, (x, y))| \in \mathcal{O}(\epsilon)$.

3. *(Empirical Rademacher complexity of threshold classifiers)* Consider a binary classification problem where $\mathcal{X} = \mathbb{R}$. Consider the following dataset $S = \{(x_1 = 0, y_1), (x_2 = 1, y_2)\}$, where $x_1 < x_2$ and $y_1, y_2 \in \{0, 1\}$ are arbitrary.

   (a) **[2 pts]** Let $\mathcal{H}_1 = \{h_\theta(x) = \mathbb{1}(x \geq \theta); \theta \in \mathbb{R}\}$ be the hypothesis class of one-sided threshold classifiers. Compute the empirical Rademacher complexity $\widehat{\mathcal{R}}(\mathcal{H}_1, S)$.

   (b) **[2 pts]** Let $\mathcal{H}_2 = \{h_\theta(x) = \mathbb{1}(x \geq \theta); \theta \in \mathbb{R}\} \cup \{h_\theta(x) = \mathbb{1}(x \leq \theta); \theta \in \mathbb{R}\}$ be the class of two-sided threshold classifiers. Compute the empirical Rademacher complexity $\widehat{\mathcal{R}}(\mathcal{H}_2, S)$.

   (c) **[1 pts]** Are the values computed above consistent with the fact that $\mathcal{H}_1 \subset \mathcal{H}_2$?

4. **[6 pts]** *(Reading exercise, VC dimension of linear classifiers)* Consider a binary classification problem where $\mathcal{X} = \mathbb{R}^D$ is the D-dimensional Euclidean space. The class of linear classifiers is given by $\mathcal{H} = \{h_{w,b}(x) = \mathbb{1}[w^\top x + b \geq 0]; w \in \mathbb{R}^D, b \in \mathbb{R}\}$. Prove that the VC dimension of this class is $D + 1$.

   You may read the proof in either SB or MRT, and reproduce it in your own words.

3 On the Rademacher complexity

Recall the following definitions from class. Let $\sigma = (\sigma_1, \ldots, \sigma_n) \in \{-1, +1\}^n$ be n independent Rademacher random variables. For $A \subset \mathbb{R}^n$, denote $\widehat{\mathcal{R}}(A) \triangleq \mathbb{E}_\sigma \left[\sup_{a \in A} \frac{1}{n} \sigma^\top a\right]$. Next, let $\mathcal{Z}$ be a set, let $\mathcal{F} \subset \mathbb{R}^\mathcal{Z}$. The empirical

[2] Recall, in the linear regression example, we first showed that the losses were Lipschitz, *i.e.*, for all x, y, we have $|\ell(h_\theta, (x, y)) - \ell(h_{\theta'}, (x, y))| \leq K\|\theta - \theta'\|_2$. We used this property when constructing a cover *for all* ϵ and proving uniform convergence.

Rademacher complexity of $\mathcal{F}$ for set $S = \{z_1, \ldots, z_n\} \subset \mathcal{Z}$ is, $\widehat{\mathcal{R}}(\mathcal{F}, S) \triangleq \widehat{\mathcal{R}}(\mathcal{F}(S)) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f_i(z_i)\right]$.
The Rademacher complexity of a function class $\mathcal{F}$ with respect to a distribution P is, $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{S \sim P^n}\left[\widehat{\mathcal{R}}(\mathcal{F}, S)\right]$.

Prove the following statements. In each case, you may prove the first statement about $\widehat{\mathcal{R}}$ and then verify that the second statement about $\mathcal{R}_n$ follows from the first.

1. **[2 pts]** Let $A \subset B \subset \mathbb{R}^n$. Then $\widehat{\mathcal{R}}(A) \leq \widehat{\mathcal{R}}(B)$.
   Hence, if $\mathcal{F} \subset \mathcal{G} \subset \mathbb{R}^{\mathcal{Z}}$, we have $\mathcal{R}_n(\mathcal{F}) \leq \mathcal{R}_n(\mathcal{G})$.
2. **[3 pts]** Let $\alpha \in \mathbb{R}$, and let $a \in \mathbb{R}^n$. Let $A \subset \mathbb{R}^n$. Then, $\widehat{\mathcal{R}}(\alpha A + a) = |\alpha| \widehat{\mathcal{R}}(A)$.
   Hence, for any $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$ and $f \in \mathbb{R}^Z$, we have $\mathcal{R}_n(\alpha \mathcal{F} + f) = |\alpha| \mathcal{R}_n(\mathcal{F})$.
3. **[2 pts]** Let $A, B \subset \mathbb{R}^n$. Then, $\widehat{\mathcal{R}}(A + B) = \widehat{\mathcal{R}}(A) + \widehat{\mathcal{R}}(B)$.
   Hence, for any $\mathcal{F}, \mathcal{G} \subset \mathbb{R}^{\mathcal{Z}}$, we have $\mathcal{R}_n(\mathcal{F} + \mathcal{G}) = \mathcal{R}_n(\mathcal{F}) + \mathcal{R}_n(\mathcal{G})$.
4. **[4 pts]** Let $A \subset \mathbb{R}^n$. Let $|A| = \{|a|; a \in A\}$, where $|a| = \{|a_1|, \ldots, |a_n|\}$ is the pointwise absolute value. Then, $\widehat{\mathcal{R}}(|A|) \leq \widehat{\mathcal{R}}(A)$.
   Hence, for any $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$, we have $\mathcal{R}_n(|\mathcal{F}|) \leq \mathcal{R}_n(\mathcal{F})$ where $|\mathcal{F}| = \{|f|; f \in \mathcal{F}\}$ and $|f|(x) = |f(x)|$.
5. **[4 pts]** Let $A', A'' \subset \mathbb{R}^n$. Let $A = \{\max(a', a''); a' \in A', a'' \in A''\}$, where $\max$ is the pointwise maximum of the vectors. Then, $\widehat{\mathcal{R}}(A) \leq \widehat{\mathcal{R}}(A') + \widehat{\mathcal{R}}(A'')$.
   Hence, for any $\mathcal{F}', \mathcal{F}'' \subset \mathbb{R}^{\mathcal{Z}}$, we have $\mathcal{R}_n(\{\max(f', f''); f' \in \mathcal{F}', f'' \in \mathcal{F}''\}) \leq \mathcal{R}_n(\mathcal{F}') + \mathcal{R}_n(\mathcal{F}'')$.

   Hint. : For any $x, y \in \mathbb{R}$, we have $\max(x, y) = \frac{1}{2}(x + y + |x - y|)$.

4 Sauer's lemma for interval classifiers ⋆

1. *(Interval classifiers)* Let $\mathcal{X} = \mathbb{R}$. Consider the class of interval classifiers, given by

$$\mathcal{H} = \{h_{a,b}(x) = \mathbb{1}(a \leq x \leq b); a, b \in \mathbb{R}, a \leq b\}.$$

   (a) **[3 pts]** What is the VC dimension d of this class?

   (b) **[6 pts]** Show that Sauer's lemma is tight for this class. That is, for all n, show that $g(n, \mathcal{H}) = \sum_{i=0}^d \binom{n}{i}$.

2. *(Union of interval classifiers)* Let $\mathcal{X} = \mathbb{R}$. Consider the class of the union of K interval classifiers, given by

$$\mathcal{H} = \{h_{a,b}(x) = \mathbb{1}(\exists k \in \{1, \ldots, K\} \text{ s.t } a_k \leq x \leq b_k); a, b \in \mathbb{R}^k, a_k \leq b_k \forall\, k\}.$$

   (a) **[4 pts]** What is the VC dimension d of this class?

   (b) **[8 pts]** Show that Sauer's lemma is tight for this class. That is, for all n, show that $g(n, \mathcal{H}) = \sum_{i=0}^d \binom{n}{i}$.

   Hint: The following identity, which we used in the proof of Sauer's lemma, may be helpful.

$$\forall\, m > k, \quad \binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1}.$$

3. **[5 pts]** *(Tightness of Sauer's lemma)* Prove the following statement about the tightness of Sauer's lemma when $\mathcal{X} = \mathbb{R}$: For all $d > 0$, there exists a hypothesis class $\mathcal{H} \subset \{h : \mathbb{R} \to \{0, 1\}\}$ with VC dimension $d_{\mathcal{H}} = d$ such that, for all dataset sizes $n > 0$, we have $g(n, \mathcal{H}) = \sum_{i=0}^d \binom{n}{i}$. Note that the hypothesis class $\mathcal{H}$ could depend on d but not on n.

   Hint: There are many ways to solve this. One approach will be to use the results from part 2 which will allow you to prove the results for even d. You should consider a different hypothesis class to show this for odd d.

   An alternative approach is to prove the following more general statement: "For any set $\mathcal{X}$ such that $|\mathcal{X}| \geq d$, there exists a hypothesis class $\mathcal{H}$ of VC dimension d such that for all $n \leq |\mathcal{X}|$, we have $g(n, \mathcal{H}) = \sum_{i=0}^d \binom{n}{i}$".