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

**University of Wisconsin–Madison, Fall 2024** *Instructor: Kirthevasan Kandasamy*

Homework 1. Due 09/27/2024, 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; handwritten and scanned submissions will **not** be accepted. If you typeset your homework using LaTeX, you will receive 5 percent extra credit.
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 try the problems on your own instead of searching for answers. If you used any external references, please cite them in your submission.
6. **Collaboration:** You may collaborate in groups of size up to 3 on this homework. If you collaborate, please indicate your collaborators at the beginning of your homework. In any case, you must write the solution in your own words.

# 1 PAC Learning and ERM

1. **[4 pts]** *(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 $h^\star \in \operatorname{argmin}_{h \in \mathcal{H}} R(h)$ be an optimal classifier in the class, and let $\widehat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} \widehat{R}(h)$ minimize the empirical risk of the dataset $S$. A student offers the following proof and claims that it is possible to bound the estimation error without any dependence on $|\mathcal{H}|$.

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

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

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

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

   *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 risk.*

   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. **[6 pts]** *(PAC bound)* Prove the following result which was presented but not proved in class.

   Let $\mathcal{H}$ be a hypothesis class with finite $\mathrm{Rad}_n(\mathcal{H})$. Let $\widehat{h}$ be obtained via ERM using $n$ i.i.d samples. Let $\epsilon > 0$. Then, there exists universal constants $C_1, C_2$ such that with probability at least $1 - 2e^{-2n\epsilon^2}$, we have

   $$R(\widehat{h}) \leq \inf_{h \in \mathcal{H}} R(h) + C_1 \mathrm{Rad}_n(\mathcal{H}) + C_2 \epsilon.$$

3. **[3 pts]** *(Sample complexity based on VC dimension)* Say $\mathcal{H}$ has a finite VC dimension $d$. Let $\delta \in (0, 1)$. Using the result/proof in part 2 or otherwise, show that there exist universal constants $C_3, C_4$ such that when $n \geq d$, the following bound holds with probability at least $1 - \delta$.

   $$R(\widehat{h}) \leq \inf_{h \in \mathcal{H}} R(h) + C_3 \sqrt{\frac{d \log(n/d) + d}{n}} + C_4 \sqrt{\frac{1}{n} \log\left(\frac{2}{\delta}\right)}.$$

4. **[3 pts]** *(Bound on the expected risk)* The above results show that $R(\widehat{h})$ is small with high probability. Using the results/proofs in parts 2 and 3 or otherwise, show that it is also small in expectation. Specifically, show that there exist universal constants $C_5, C_6$ such that the following bound holds.

   $$\mathbb{E}[R(\widehat{h})] \leq \inf_{h \in \mathcal{H}} R(h) + C_5 \sqrt{\frac{d \log(n/d) + d}{n}} + C_6 \sqrt{\frac{\log(4n)}{n}} + \frac{1}{\sqrt{n}}.$$

   Here, the expectation is with respect to the dataset $S$.

*For parts 2, 3, and 4, of this question, if you can prove a bound that has similar higher order terms but differs in additive/multiplicative constants or poly-logarithmic factors, you will still receive full credit.*

## 2 Rademacher complexity and VC dimension

1. **[5 pts]** *(Empirical Rademacher complexity)* Consider a binary classification problem with the 0–1 loss $\ell(y_1, y_2) = \mathbb{1}(y_1 \neq y_2)$ and where $\mathcal{X} = \mathbb{R}$. Consider the following dataset $S = \{(x_1 = 0, y_1 = 0), (x_2 = 1, y_2 = 1)\}$.

   (a) Let $\mathcal{H}_1 = \{h_a(x) = \mathbb{1}(x \geq a); a \in \mathbb{R}\}$ be the hypothesis class of one-sided threshold functions. Compute the empirical Rademacher complexity $\widehat{\mathrm{Rad}}(S, \mathcal{H}_1)$.

   (b) Let $\mathcal{H}_2 = \{h_a(x) = \mathbb{1}(x \geq a); a \in \mathbb{R}\} \cup \{h_a(x) = \mathbb{1}(x \leq a); a \in \mathbb{R}\}$ be the class of two-sided threshold functions. Compute the empirical Rademacher complexity $\widehat{\mathrm{Rad}}(S, \mathcal{H}_2)$.

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

2. **[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 $d_{\mathcal{H}} = D + 1$.

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

## 3 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) **[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 $g(n, \mathcal{H}) = \sum_{i=0}^{d} \binom{n}{i}$.

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. **[6 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 $g(n, \mathcal{H}) = \sum_{i=0}^{d} \binom{n}{i}$".

# 4 PAC lower bounds for normal mean estimation

We are given $n$ independent samples $S = \{X_1, \dots, X_n\}$, where each $X_i$ is sampled from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with *unknown* mean $\mu$, but known variance $\sigma^2$. Let $\epsilon > 0$ be given. We wish to design an estimator $\widehat{\mu} : \mathbb{R}^n \to \mathbb{R}$ which is $\epsilon$ close to $\mu$ with high probability. In this question, you will show that the minimax risk $R_n^\star$, defined below, satisfies,

$$R_n^\star \triangleq \inf_{\widehat{\mu}} \sup_{\mu \in \mathbb{R}} \mathbb{P}(|\widehat{\mu}(S) - \mu| > \epsilon) = 2\left(1 - \Phi\left(\frac{\epsilon\sqrt{n}}{\sigma}\right)\right).$$

Here, $\Phi(x) = \mathbb{P}_{Z \sim \mathcal{N}(0,1)}(Z < x)$ is the CDF of the standard normal distribution.

1. **[3 pts]** *(Upper bound)* Design an estimator $\widehat{\mu}$ for $\mu$ which satisfies $\sup_{\mu \in \mathbb{R}} \mathbb{P}(|\widehat{\mu}(S) - \mu| > \epsilon) = 2\left(1 - \Phi\left(\frac{\epsilon\sqrt{n}}{\sigma}\right)\right)$.
2. **[6 pts]** *(Lower bound)* Next, show that

$$\inf_{\widehat{\mu}} \sup_{\mu \in \mathbb{R}} \mathbb{P}(|\widehat{\mu}(S) - \mu| > \epsilon) \geq 2\left(1 - \Phi\left(\frac{\epsilon\sqrt{n}}{\sigma}\right)\right).$$

**Hint.** Consider the Bayesian model, where we first sample $\mu \sim \mathcal{N}(0, \tau^2)$ and then sample $n$ i.i.d points $S$ from $\mathcal{N}(\mu, \sigma^2)$. Then, the posterior for $\mu$ conditioned on $S$ follows the normal distribution $\mathcal{N}(\mu_\tau, \sigma_\tau^2)$ where, $\mu_\tau = \frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2} \frac{1}{n} \sum_{i=1}^n X_i$ and $\sigma_\tau^2 = \frac{1}{n/\sigma^2 + 1/\tau^2}$.