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

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

Homework 3. Due 10/25/2025, 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 Lower bounds on the excess risk for binary classification ⋆

In this question, you will derive lower bounds on the excess risk for binary classification. You will first consider a simple setting to build intuition, and then prove a key technical result for finite VC classes, which was omitted in the lectures.

Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a given hypothesis class. For any $h \in \mathcal{H}$, let $\ell(h,(X,Y)) = \mathbb{1}(h(X) \neq Y)$ be the 0–1 loss on instance (X,Y), and let $L(h,P) = \mathbb{E}_{X,Y\sim P}[\mathbb{1}(h(X) \neq Y)]$ be the population loss. In this question, we will consider algorithms $\widehat{h}$ which map a given dataset of n points $S = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ to a hypothesis in $\mathcal{H}$.

1. **[8 pts]** *(One sided-threshold classifiers)* Consider a binary classification problem with input in $\mathcal{X} = [0,1]$ and label in $\{0,1\}$. Let $\mathcal{P}$ be the set of distributions whose marginal $p(x)$ is the uniform distribution on $[0,1]$. Let $\mathcal{H} = \{h_t(\cdot) = \mathbb{1}(\cdot \geq t); t \in [0,1]\}$ be the class of one-sided threshold classifiers. **Using Le Cam's method**, show that, for any algorithm $\widehat{h}$, there exists some distribution $P \in \mathcal{P}$ such that

$$\mathbb{E}_{S\sim P^n}\left[L(\widehat{h}(S),P)\right] \geq \inf_{h\in\mathcal{H}} L(h,P) + \Omega\left(\sqrt{\frac{1}{n}}\right).$$

2. **[10 pts]***(Classification in a VC class)* Now, let $\mathcal{H}$ be a hypothesis class with VC dimension d. In class, we used the local Fano method to show that for any algorithm $\widehat{h}$, there exists some distribution $P \in \mathcal{P}$ such that, for sufficiently large d,

$$\mathbb{E}_{S\sim P^n}\left[L(\widehat{h}(S),P)\right] \geq \inf_{h\in\mathcal{H}} L(h,P) + \Omega\left(\sqrt{\frac{d}{n}}\right).$$

Prove the following result—omitted in class—which provides a lower bound on the separation between the alternatives: You may refer to the proofs from the lecture notes to recall necessary definitions.

Let $P_\omega, P_{\omega'} \in \mathcal{P}'$. Then, $\Delta(P_\omega, P_{\omega'}) \geq \frac{\gamma}{d} H(\omega,\omega')$.

2 Two-layer Neural Networks

(Reading exercise) In this problem, you will improve the bound on the Rademacher complexity of a two-layer neural network we proved in class. You may find it helpful to read section 5.3 of Tengyu Ma's lecture notes before attepting parts 2 and 3.

1. **[6 pts]** *(Warm up, Data-dependent bounds for linear models under ℓ_2 constraints)* Let $\mathcal{H} = \{h_\theta; h_\theta(x) = \theta^\top x; \theta \in \mathbb{R}^d,\ \|\theta\|_2 \leq B_\theta\}$, be the class of linear regressors with bounded ℓ_2 norm. Let $S = \{x_1,\ldots,x_n\} \subset \mathbb{R}^d$ be given. Then, show that $\widehat{\mathcal{R}}(\mathcal{H},S) \leq \frac{B_\theta}{n}\|X\|_F$. Here, $\|X\|_F^2 = \sum_{i=1}^n \sum_{j=1}^d x_{i,j}^2$ is the squared Frobenius norm of the data matrix $X \in \mathbb{R}^{n\times d}$, whose i^{th} row is $x_i^\top$.

2. **[3 pts]** Let $A \subset \mathbb{R}^n$ such that $\mathbf{0} \in A$. Let $\sigma \in \{-1,+1\}^n$ denote n independent Rademacher random variables. Then, show that $\mathbb{E}_\sigma\left[\sup_{a\in A} |\sigma^\top a|\right] \leq 2\mathbb{E}_\sigma\left[\sup_{a\in A} \sigma^\top a\right]$.

   Hint. Letting $\phi(u) = \max(0,x)$, we can write $|u| = \phi(u) + \phi(-u)$.

3. *(Rademacher complexity of a two-layer neural network)* A two-layer neural network, whose input $x \in \mathbb{R}^d$, with ReLU activation can be characterized as follows. It has a hidden layer with m neurons. Let $W \in \mathbb{R}^{m\times d}$ be the weight matrix mapping inputs to the hidden layer. The output of each hidden layer neuron is passed through a ReLU activation function $\phi(x) = \max(x,0)$. A linear layer at the output with parameters $\beta \in \mathbb{R}^m$, maps these activations to a scalar output. Letting $\theta = (\beta, W)$ denote the parameters of the neural network, the output of the neural network is given by $h_\theta(x) = \beta^\top \phi(Wx)$, where, ϕ is applied pointwise to Wx.

   Let us denote $\|\theta\|_{1,2} = \sum_{j=1}^m |\beta_j| \|w_j\|_2$, where $w_j^\top$ is the j^{th} row of W. We will consider the following hypothesis class,

$$\mathcal{H} = \left\{h_\theta;\ h_\theta(x) = \beta^\top \phi(Wx);\ \|\theta\|_{1,2} \leq B_\theta,\ \forall\, j \in [m]\right\}.$$

   We will also assume that for $X \sim P$, we have $\mathbb{E}[\|X\|_2^2] \leq B_x^2$.

(a) **[5 pts]** Let $S = \{x_1, \ldots, x_n\}$ be a dataset of n points. Show that the empirical Rademacher complexity satisfies,

$$\widehat{\mathcal{R}}(\mathcal{H}, S) \leq \frac{B_\theta}{n} \mathbb{E}_\sigma \left[\sup_{w; \|w\|_2 \leq 1} \left| \sum_{i=1}^{n} \sigma_i \phi(w^\top x_i) \right| \right]$$

(b) **[4 pts]** Combine the above result with the results in parts 1 and 2, to show that,

$$\mathcal{R}_n(\mathcal{H}) \leq \frac{2 B_\theta B_x}{\sqrt{n}}.$$

N.B. Note that, unlike the result we showed in class, here, there is no explicit dependence on the number of neurons m.

3 The doubling trick

In class, we studied several algorithms for sequential decision-making problems where the time horizon is assumed to be known ahead of time. However, often, we are interested in *any-time* algorithms, where we execute an algorithm indefinitely, but wish to bound the regret after *any* number of rounds.

The doubling trick is often used to convert a given known time horizon algorithm into an any-time procedure. For such an algorithm $\mathcal{A}$, let $\mathcal{A}_T$ be the version of $\mathcal{A}$ that is executed with time horizon T. The doubling trick proceeds by choosing an arbitrary time horizon T_0 (for simplicity, we may take $T_0 = 1$), and then executing A_{T_0} for T_0 rounds, then A_{2T_0} for $2T_0$ rounds, and proceeding in this fashion, doubling the time horizon each time we finish an execution.

Algorithm 1 The doubling trick

Given: An algorithm $\mathcal{A}$.
Set $T_0 \leftarrow 1$.
Initialize algorithm $\mathcal{A}_{T_0}$.
for $t = 1, 2, \ldots$ **do**
  if $t \geq 2T_0$ **then**
    Set $T_0 \leftarrow 2T_0$.
    Re-initialize algorithm $\mathcal{A}_{T_0}$.
  end if
  Execute round $t - T_0 + 1$ of algorithm $\mathcal{A}_{T_0}$.
end for

In Hedge and EXP3, re-initialization would mean resetting the cumulative losses to 0, and setting the learning rate to $\eta = \sqrt{\log(K)/(KT_0)}$. In UCB, this could mean resetting the mean estimates and confidence intervals.

We wish to bound the regret of this algorithm, which can be written as $R_T \triangleq \sum_{t=1}^{T} r_t$, where r_t is the instantaneous regret. For instance, in the experts problem, this would be the difference between the loss of the action taken and the loss of any fixed action.

Suppose that there exist $\alpha > 0$, $\beta > 0$, and $\gamma \in (0, 1)$ such that the regret of algorithm $\mathcal{A}$ satisfies $R_T(\mathcal{A}_T) \leq \alpha T^\gamma + \beta$ for all *known* time horizons T. Let $\mathcal{A}'$ denote the version of this algorithm modified using the doubling trick.

[6 pts] Show that when executed in an any-time fashion, $\mathcal{A}'$ satisfies $R_T(\mathcal{A}') \in \tilde{\mathcal{O}}(T^\gamma)$ for all T.

4 Optimistic algorithms for structured bandits

1. *(Linear bandits)* Recall the stochastic linear bandit problem, where, on each round a learner chooses an action A_t and observes $X_t = \theta_\star^\top A_t + \epsilon_t$. In class, we studied the LinUCB algorithm which chooses action $A_t =$

$\text{argmax}_{a \in \mathcal{A}} \text{UCB}_{t-1}(a)$ on round t, where $\text{UCB}_{t-1}(a) = \widehat{\theta}_{t-1}^\top a + \beta_t \|a\|_{V_{t-1}^{-1}}$. Here,

$$\widehat{\theta}_t \triangleq \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left(\lambda \|\theta\|_2^2 + \sum_{s=1}^{t} (X_s - \theta^\top A_s)^2 \right) \qquad \text{and,} \quad V_t = \lambda I + \sum_{s=1}^{t} A_s A_s^\top,$$

We showed that UCB_{t-1} is an upper confidence bound for $\theta_\star^\top a$ for all a and all t, for appropriately chosen β_t.

(a) **[2 pts]** *(Expression for $\widehat{\theta}_t$)* Show that, $\widehat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} A_s X_s$.

(b) *(Deriving LinUCB via the optimism principle)* Recall the "optimism under uncertainty principle", which states that we should pretend that the environment is as nice as statistically possible, given the data, and then behave myopically. A common application of this idea for linear bandits is as follows (see for instance, LS Chapter 19): construct a confidence region C_t for $\theta_\star$, and then choose the action which *optimistically* maximizes the expected reward in C_t. That is,

$$A_t = \operatorname*{argmax}_{a \in \mathcal{A}} \max_{\theta \in C_{t-1}} \theta^\top a.$$

In our proof of LinUCB we showed that the following C_t traps $\theta_\star$ with probability at least $1 - 1/T^2$,

$$C_t = \left\{ \theta \in \mathbb{R}^d;\ \|\theta - \widehat{\theta}_t\|_{V_t} \leq \beta_{t+1} \right\}.$$

Here, $\widehat{\theta}_t$ and β_t are as given above. You are encouraged to go through the proof we did in class and identify precisely where we showed this.

[6 pts] Show that the above choice of C_t yields precisely the same action selection rule as LinUCB.

2. *(Reading exercise, GP bandits)* Let $\mathcal{A}$ be an action space and let $f : \mathcal{A} \to \mathbb{R}$ be an (unknown) function defined on $\mathcal{A}$. Consider a bandit algorithm π whose pseudo-regret is,

$$R_T(\pi) = T f(a_\star) - \sum_{t=1}^{T} f(A_t).$$

Here $a_\star \in \text{argmax}_{a \in \mathcal{A}} f(a)$. Recall the following recipe for designing optimistic algorithms for structured bandits, and bounding the pseudo-regret.

(a) **Algorithm:** on each round choose $A_t = \text{argmax}_{a \in \mathcal{A}} \text{UCB}_{t-1}(a)$, where UCB_t is an upper confidence bound for f based on data collected up to round t.

(b) Let G be the good event in which the confidence intervals trap the true function f. Use martingale concentration to upper bound $\mathbb{P}(G^c)$.

(c) Under G, bound the instantaneous pseudo-regret by $f(a_\star) - f(A_t) \leq 2 \times e_{A_t, t-1}$, where $e_{A_t,t-1}$ is the width of the confidence interval of A_t at round $t-1$.

(d) Bound the summation $\sum_{t=1}^{T} e_{A_t,t-1}$.

[7 pts] Read the paper *Gaussian process optimization in the bandit setting* by Srinivas, Krause, Kakade, Seeger, 2010, available at `arxiv.org/pdf/0912.3995`. In particular, focus on the frequentist result in Theorem 3, when f lies in an RKHS. Identify the specific equation, section, lemma, or theorem in the paper that corresponds to each of the four steps outlined above.

N.B. You do not need to understand what an RKHS is to answer this question. My recommendation would be to read sections 2 and 3 to understand the setting and then focus on Theorem 3. You can then go over the proof of Theorem 3 in Appendix A and B. While Appendix A is written in the Bayesian setting, many results (including those relevant to this question) are also applicable in the frequentist setting.

5 Rademacher complexity of monotone functions ⋆

Consider the class of non-decreasing functions bounded in $[0, 1]$,

$$\mathcal{H} = \left\{ h \in [0,1]^{\mathbb{R}_+};\ \text{for all } t_1 < t_2,\ h(t_1) \le h(t_2) \right\} \tag{1}$$

We perform regression via ERM in this class, and wish to bound the excess risk via its Rademacher complexity $\mathcal{R}_n(\mathcal{H})$. In this question, you will bound the Rademacher complexity of this class.

1. **[8 pts]** *(Covering number bound.)* Let $S = \{x_1, \ldots, x_n\} \subset \mathbb{R}$ be a given dataset of n points. Show that

$$N_\infty(\epsilon, \mathcal{H}(S)) \triangleq N(\epsilon, \mathcal{H}(S), \|\cdot\|_\infty) < \begin{cases} (n+1)^{\frac{1}{2\epsilon}} & \text{if, } \epsilon \le 1/2, \\ 1 & \text{if, } \epsilon \ge 1/2. \end{cases}$$

2. **[3 pts]** Recall the following bound on the empirical Rademacher complexity we showed in class, for $p \in [1, \infty]$ and $A \subset [-B, B]^n$:

$$\widehat{\mathcal{R}}(A) \le \inf_{\epsilon > 0} \left(\epsilon + B\sqrt{\frac{2 \log N_p(\epsilon, A)}{n}} \right)$$

   Using this result, show that $\mathcal{R}_n(\mathcal{H}) \in \tilde{\mathcal{O}}(n^{-1/3})$.

3. **[3 pts]** *(Dudley entropy integral)* Recall the bound on the Rademacher complexity, obtained via the Dudley entropy integral, for $p \in [2, \infty]$,

$$\widehat{\mathcal{R}}(A) \le \inf_{\epsilon \ge 0} \left(4\epsilon + \frac{12}{\sqrt{n}} \int_\epsilon^\infty \sqrt{\log(N_p(t, A))}\, \mathrm{d}t \right).$$

   Bound the Rademacher complexity using the above result. Your rate should be better than than the one obtained in part 2 by a polynomial factor.

4. **[7 pts]** *(Slowly decreasing functions)* Suppose that instead of (1), $\mathcal{H}$, defined as follows, consists of unbounded, but slowly decreasing functions:

$$\mathcal{H} = \left\{ h \in \mathbb{R}_+^{\mathbb{R}_+};\ \text{for all } t_1 < t_2,\ h(t_1) \le h(t_2), \quad h(t) \le \alpha t^\beta \right\}.$$

   Here, $\alpha < \infty$ and $\beta \in [0, 1]$ are fixed. Moreover, let us assume that the data distribution P has finite mean, *i.e.*, $\mathbb{E}[X] \le \mu < \infty$ and is η-sub-Gaussian, *i.e.*, $\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] \le e^{\frac{\lambda^2 \eta^2}{2}}$. Bound the Rademacher complexity $\mathcal{R}_n(\mathcal{H})$ of $\mathcal{H}$ with respect to P.

   N.B. You may be able to relax the $\beta \le 1$ condition via a more involved analysis and obtain the same rate up to polylog factors.