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

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

Homework 3. Due 10/26/2024, 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 Lower bounds on the estimation error for prediction problems

In this question, we will develop a framework for lower bounding the excess risk of prediction problems. We will use this to establish a lower bound on the estimation error for binary classificaion in a VC class.

Let $\mathcal{Z}$ be a data space and $\mathcal{H}$ be a hypothesis space. Let $f : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ be the *instance loss*, where $f(h, Z)$ is the loss of hypothesis h on instance Z. Let $F(h, P) = \mathbb{E}_{Z \sim P}[f(h, Z)]$ be the *population loss* of hypothesis h on distribution P, and let $L(h, P) = F(h, P) - \inf_{h \in \mathcal{H}} F(h, P)$ denote the *excess population loss*. Let $\widehat{h}$ be an estimator which maps a dataset to a hypothesis in $\mathcal{H}$. The risk of the estimator is

$$R(\widehat{h}, P) = \mathbb{E}\big[L(\widehat{h}, P)\big] = \mathbb{E}\big[F(\widehat{h}, P)\big] - \inf_{h \in \mathcal{H}} F(h, P).$$

Here, the expectation is taken with respect to the data. The minimax risk is $R^\star = \inf_{\widehat{h}} \sup_{P \in \mathcal{P}} R(\widehat{h}, P)$.

Example: *In classification, $f(h, (X, Y)) = \mathbb{1}(h(X) \neq Y)$ is the 0–1 loss, $F(h, P)$ is usually called the risk of hypothesis h. The infimum $\inf_h F(h, P)$ is attained by the Bayes' optimal classifier, and $L(h, P)$ is the excess risk of hypothesis h. This framework can be used to lower bound the expected excess risk $R(\widehat{h}, P)$ of classification (and regression) problems. When $\widehat{h}$ chooses a hypothesis in some hypothesis class $\mathcal{H}$, then $R(\widehat{h}, P)$ is the estimation error.*

1. **[6 pts]** *(Reduction to testing)* For two distributions P, Q, we define the separation $\Delta(P, Q)$ as,

$$\begin{aligned}\Delta(P, Q) = \sup\big\{\delta \geq 0;\quad & L(h, P) \leq \delta \implies L(h, Q) \geq \delta \;\forall\, h \in \mathcal{H}, \\ & L(h, Q) \leq \delta \implies L(h, P) \geq \delta \;\forall\, h \in \mathcal{H}\big\}.\end{aligned}$$

   A dataset S is drawn from some distribution $P \in \mathcal{P}$. Let $\{P_1, \ldots, P_N\} \subset \mathcal{P}$ such that $\Delta(P_j, P_k) \geq \delta$ for all $j \neq k$. Let ψ be any function which maps S to $[N]$. Show that,

$$R^\star \geq \delta \inf_{\psi} \max_{j \in [N]} P_j(\psi \neq j).$$

We can establish the following statements from the above result when S consists of n i.i.d data points. You do not need to prove them for the homework, but are encouraged to verify that they are true.

> **Le Cam's method:** Let $\{P_0, P_1\} \subset \mathcal{P}$ such that $\Delta(P_0, P_1) \geq \delta$ and $\mathrm{KL}(P_0, P_1) \leq \log(2)/n$. Then, $R_n^\star \geq \delta/8$.
>
> **Local Fano method:** Let $\{P_1, \ldots, P_N\} \subset \mathcal{P}$ such that $\Delta(P_j, P_k) \geq \delta$ and $\mathrm{KL}(P_j, P_k) \leq \log(N)/4n$ for all $j \neq k$. Suppose $N \geq 16$. Then, $R_n^\star \geq \delta/2$.

You may use these results when solving the problems below.

2. *(One sided-threshold classifiers)* Consider a binary classification problem with input in $\mathcal{X} = [0, 1]$ and label in $\{0, 1\}$. We observe $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ drawn i.i.d from some distribution $P \in \mathcal{P}$, where $\mathcal{P}$ consists 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. For any $h \in \mathcal{H}$, let $f(h, (X, Y)) = \mathbb{1}(h(X) \neq Y)$ be the 0–1 loss and let $F(h, P) = \mathbb{E}_{X,Y \sim P}[\mathbb{1}(h(X) \neq Y)]$.

   (a) **[8 pts]** Using Le Cam's method, show that for any estimator $\widehat{h}$ which maps the dataset to a hypothesis in $\mathcal{H}$, there exists some distribution $P \in \mathcal{P}$ such that

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

   (b) **[3 pts]** Note that we have *not* assumed that $h(\cdot) = P(Y = 1|X = \cdot)$ belongs to $\mathcal{H}$ for all $P \in \mathcal{P}$. We have however assumed that the estimator $\widehat{h}$ always chooses some hypothesis in $\mathcal{H}$. In words, explain why this assumption is necessary and where the proof breaks without this assumption.

3. **[15 pts]***(Classification in a VC class)* Let $\mathcal{X}$ be a given input space and let $\mathcal{P}$ be all distributions supported on $\mathcal{X} \times \{0,1\}$. We observe $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ drawn i.i.d from some distribution $P \in \mathcal{P}$. Let $\mathcal{H} \subset \{h : \mathcal{X} \to \{0,1\}\}$ be a hypothesis class with finite VC dimension d. For any $h \in \mathcal{H}$, let $f(h,(X,Y)) = \mathbb{1}(h(X) \neq Y)$ be the 0–1 loss and let $F(h,P) = \mathbb{E}_{X,Y\sim P}[\mathbb{1}(h(X) \neq Y)]$.

   In Homework 1, you showed the following upper bound on the estimation error of the ERM estimator $\widehat{h}_{\mathrm{ERM}}$,
   $$\mathbb{E}\big[F(\widehat{h}_{\mathrm{ERM}}, P)\big] \leq \inf_{h\in\mathcal{H}} F(h,P) + \mathcal{O}\left(\sqrt{\frac{d\log(n)}{n}}\right).$$
   Here, the expectation is with respect to the dataset S. Using the local Fano method, show that this is rate is essentially unimprovable. That is, show that for any estimator $\widehat{h}$ which maps the dataset to a hypothesis in $\mathcal{H}$, there exists some distribution $P \in \mathcal{P}$ such that, for sufficiently large d,
   $$\mathbb{E}\big[F(\widehat{h}, P)\big] \geq \inf_{h\in\mathcal{H}} F(h,P) + \Omega\left(\sqrt{\frac{d}{n}}\right).$$

2 Explore-then-commit for K–armed bandits

In this question, we will upper and lower bound the regret for the explore-then-commit algorithm, described below, on a K–armed stochastic bandit problem.

Algorithm 1 Explore-then-Commit

```
**Given:** the $K$ arms $[K]$, time horizon $T$, number of exploration rounds $m$ $(< T/K)$
Pull each arm $m$ times in the first $mK$ rounds.
Set $A = \mathrm{argmax}_{i\in[K]} \frac{1}{m}\sum_{t=1}^{mK} \mathbb{1}(A_t = i)X_t$.
Pull arm $A$ for the remaining $T - mK$ rounds.
```

Let $\nu = \{\nu_i\}_{i\in[K]}$ be a σ sub-Gaussian K-armed bandit model, i.e each ν_i is a σ sub-Gaussian distribution. Let $\mu_i = \mathbb{E}_{X\sim\nu_i}[X]$ denote the mean of arm i, $\mu^\star = \max_{j\in[K]} \mu_j$ be the highest mean, and let $\Delta_i = \mu^\star - \mu_i$ be the gap between the optimal arm and the i^{th} arm. Assume, without loss of generality, that $\mu_i \in [0,1]$ for all i. Let $R_T(m,\nu)$ denote the regret when we execute the above algorithm on ν with m exploration rounds,
$$R_T(m,\nu) = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right].$$

1. **[5 pts]** *(Gap-dependent bound)* Show that there exists global constants C_1, C_2 such that
   $$R_T(m,\nu) \leq m \sum_{i;\Delta_i>0} \Delta_i + C_1(T - mK) \sum_{\Delta_i>0} \Delta_i \exp\left(\frac{-m\Delta_i^2}{C_2\sigma^2}\right).$$

2. **[6 pts]** *(Gap-independent bound)* Let $\mathcal{P}$ denote the class of all σ sub-Gaussian bandits whose means are bounded between 0 and 1. Show that for a suitable choice of m, say m' (possibly dependent on T and K), that we have
   $$\sup_{\nu\in\mathcal{P}} R_T(m',\nu) \in \tilde{\mathcal{O}}(K^{1/3}T^{2/3}).$$

3. **[15 pts]** *(Lower bound)* Show that the result in part 2 cannot be improved (say via a tighter upper bound analysis) for the explore-then-commit algorithm. That is, show
   $$\inf_{m\in\mathbb{N}} \sup_{\nu\in\mathcal{P}} R_T(m,\nu) \in \Omega(K^{1/3}T^{2/3}).$$

   Hint: One approach is to adopt a similar technique to the proof of the general lower bound for K-armed bandits, but adapt it to the structure of the explore-then-commit algorithm. Your alternatives will need to depend on the specific choice of m to get a tight lower bound. To do so, you should carefully consider the failure cases if m is picked to be too large or too small.

3 The doubling trick

In class, we looked at 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 2 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 Do losses need to be non-negative for Hedge/EXP3?

Suppose we are in a sequential decision-making environment with adversarial losses that are known to be in $[-1, 0]^K$. The losses are bounded, but are negative. If we have full information feedback, we will use the Hedge algorithm, and if we have bandit feedback, we will use EXP3. In this question, we will explore if it is necessary to shift the losses so that they are non-negative before applying either of these algorithms.

1. **[2 pts]** *(Shifting losses in Hedge)* Intuitively, explain if you expect the Hedge algorithm to still work if the losses are negative, but are bounded in $[-1, 0]$.
2. **[2 pts]** Justify your answer in part 1 by showing where the proof we did in class breaks down without the non-negativity assumption, or by arguing that the proof will still carry through.
3. **[2 pts]** *(Shifting losses in EXP3)* Intuitively, explain if you expect EXP3 to still work if the losses are negative, but are bounded in $[-1, 0]$. You may reason using the exploration-exploitation tradeoff.
4. **[2 pts]** Justify your answer in part 3 by showing where the proof we did in class breaks down without the non-negativity assumption, or by arguing that the proof will still carry through.