CS/ECE/STAT-861: Theoretical Foundations of Machine Learning
University of Wisconsin–Madison, Fall 2023

Homework 2. Due 10/27/2023, 11.00 am

Instructions:

1. Homework is due at 11 am on the due date. Please hand over your homework at the beginning of class. *Please see the course website for the policy on late submission.*
2. I recommend that you typeset your homework using LaTeX. You will receive 5 percent extra credit if you do so. If you are submitting hand-written homeworks, please make sure it is cleanly written up and legible. I will not invest undue effort to understand bad handwriting.
3. You *must* hand in a hard copy of the homework. The only exception is if you are out of town in which case you must let me know ahead of time and email me a copy of the homework by 11 am on the due date. If this is the case, your homework *must* be typeset using LaTeX. Please do *not* email written and scanned copies.
4. Unless otherwise specified, you may use any result we have already proved in class. You do not need to prove them from scratch, but clearly state which result you are using.
5. Solutions to some of the problems may be found as examples or exercises in the suggested textbooks or other resources. You are encouraged to try the problems on your own first before searching for the solution. If you find an existing solution, first read and understand the proof, and then write it in your own words. Please indicate any references you have used at the beginning of your solution when you turn in your homework.
6. **Collaboration:** You are allowed to collaborate on in groups of size up to 3 on each problem. If you do so, please indicate your collaborators at the beginning of your solution.

1 Lower bounds with mixtures

In this question, you will prove a variant of our current framework for proving minimax lower bounds that involve mixtures of distributions.

1. **[5 pts]** We observe data S drawn from some distribution P belonging to a family of distributions $\mathcal{P}$. We wish to estimate a parameter $\theta(P) \in \Theta$ of interest via a loss $\Phi \circ \rho$, where $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a non-decreasing function and $\rho : \Theta \times \Theta \to \mathbb{R}_+$ is a metric. Let $\mathcal{P}_1, \dots, \mathcal{P}_N$ be subsets of $\mathcal{P}$, and let Λ_j denote a prior on $\mathcal{P}_j$. Let $\overline{P}_j$ denote the mixture,
$$\overline{P}_j(S \in A) = \mathbb{E}_{P \sim \Lambda_j}\left[\mathbb{E}_{S \sim P}\left[\mathbb{1}(S \in A)\right]\right].$$
Let $\delta = \min_{j \neq k} \inf_{P \in \mathcal{P}_j, P' \in \mathcal{P}_k} \rho(\theta(P), \theta(P'))$. Let ψ be a function which maps the data to $[N]$ and $\widehat{\theta}$ be an estimator which maps the data to Θ. Then, prove that
$$R^\star = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_S\left[\Phi \circ \rho\left(\theta(P), \widehat{\theta}(S)\right)\right] \geq \Phi\left(\frac{\delta}{2}\right) \inf_{\psi} \max_{j \in [N]} \overline{P}_j(\psi(S) \neq j).$$
(**correction:** The RHS previously said $\inf_{\widehat{\theta}}$ instead of $\inf_{\psi}$. –KK)

2. **[3 pts]** Suppose we observe n i.i.d datapoints $S = \{X_1, \dots, X_n\}$ drawn from some $P \in \mathcal{P}$. Let $\{P_0, P_1, \dots, P_N\} \subset \mathcal{P}$. Let $\overline{P} = \frac{1}{N}\sum_{j=1}^{N} P_j^n$. Show that,
$$R_n^\star \geq \frac{1}{4}\Phi\left(\frac{\delta}{2}\right)\exp\left(-\mathrm{KL}(P_0^n, \overline{P})\right)$$

3. **[2 pts]** Using the result from part 2, briefly explain why using mixtures in the alternatives can *(i)* lead to tighter lower bounds, but *(ii)* are difficult to apply.

2 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 $S = \{X_1, \dots, X_n\}$ 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.
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.
$$\mathcal{H}(2, L, B) = \left\{f : [0, 1]^d \to [0, B]; \quad \frac{\partial f}{\partial x_i} \text{ is } L\text{–Lipschitz for all } i \in [d]\right\}.$$

[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 $\mathbb{R}^d$.

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.

3 Lower bounds on the excess risk 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) **[2 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 $\mathcal{H}$. Briefly, 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,

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

4 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.

Algorithm 1 Explore-then-Commit

```
Given: time horizon $T$, number of exploration rounds $m$ $(< T/K)$
Pull each arm $i \in [K]$ $m$ 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. **[10 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.