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

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

Homework 2. Due 10/12/2024, 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 Relationships between divergences

Let P, Q be probabilities with densities p, q respectively. Recall the following divergences we discussed in class

KL divergence: $\mathrm{KL}(P,Q) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\mathrm{d}x$.

Total variation distance: $\mathrm{TV}(P,Q) = \sup_A |P(A) - Q(A)|$.

L_1 distance: $\|P - Q\|_1 = \int |p(x) - q(x)|\mathrm{d}x$.

Hellinger distance: $\mathrm{H}^2(P,Q) = \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x$.

Finally, let $\|P \wedge Q\| = \int \min(p(x), q(x))\mathrm{d}x$ denote the affinity between two distributions. When we have n i.i.d observations, let P^n, Q^n denote the product distributions.

Prove the following statements:

1. **[3 pts]** $\mathrm{KL}(P^n, Q^n) = n\mathrm{KL}(P,Q)$.
2. **[3 pts]** $\mathrm{H}^2(P^n, Q^n) = 2 - 2\left(1 - \frac{1}{2}\mathrm{H}^2(P,Q)\right)^n$.
3. **[3 pts]** $\mathrm{TV}(P,Q) = \frac{1}{2}\|P - Q\|_1$.
   Hint: Can you relate both sides of the equation to the set $A = \{x; p(x) > q(x)\}$?
4. **[3 pts]** $\mathrm{TV}(P,Q) = 1 - \|P \wedge Q\|$.
5. **[3 pts]** $\mathrm{H}^2(P,Q) \leq \|P - Q\|_1$.
   Hint: What can you say about $(a-b)^2$ and $|a^2 - b^2|$ when $a, b > 0$?

2 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 $\mathcal{P}_1, \ldots, \mathcal{P}_N$ 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
$$R^\star = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_S\left[\Phi \circ \rho\left(\theta(P), \widehat{\theta}(S)\right)\right] \geq \Phi\left(\frac{\delta}{2}\right) \inf_{\psi} \max_{j \in [N]} \overline{P}_j(\psi(S) \neq j).$$
2. **[3 pts]** Suppose we observe n i.i.d datapoints $S = \{X_1, \ldots, X_n\}$ drawn from some $P \in \mathcal{P}$. Let $\{P_0, P_1, \ldots, P_N\} \subset \mathcal{P}$ and let $\delta = \min_{j \in \{1,\ldots,N\}} \rho(\theta(P_0), \theta(P_j))$. Let $\overline{P} = \frac{1}{N}\sum_{j=1}^N P_j^n$. Show that,
$$R_n^\star \geq \frac{1}{4}\Phi\left(\frac{\delta}{2}\right) \exp\left(-\mathrm{KL}(P_0^n, \overline{P})\right)$$
   (**correction:** Previously, δ was undefined. Thanks to Guy Zamir for pointing this out. –KK)
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.

3 Lower bounds for estimating a 1-sparse mean

We observe a dataset $S \subset \mathbb{R}^d$ of n i.i.d points drawn from a distribution P belonging to the class $\mathcal{P}$ of d–dimensional normal distributions whose means are at most 1-sparse. For a vector $v \in \mathbb{R}^d$, let $|v|_0 = \sum_{i=1}^d \mathbb{1}(v_i \neq 0)$ denote the number of non-zero elements. Then, $\mathcal{P}$ is defined as follows:

$$\mathcal{P} = \left\{\mathcal{N}(\mu, \sigma^2 I); \quad \mu \in \mathbb{R}^d, |\mu|_0 \leq 1\right\}$$

We wish to design an estimator $\widehat{\theta}$ for the mean $\theta(P) = \mathbb{E}_{X\sim P}[X]$ to minimize the L_2 loss $\|\widehat{\theta} - \theta\|_2^2$.

[6 pts] Show the following lower bound for this problem.

$$R_n^\star \triangleq \inf_{\widehat{\theta}} \sup_{P\in\mathcal{P}_1} \mathbb{E}_S \left[\left\|\widehat{\theta}(S) - \theta(P)\right\|_2^2 \right] \in \Omega\left(\frac{\sigma^2 \log(d)}{n}\right).$$

You do **not** need to prove upper bounds for this question. You may assume that d is sufficiently large.

N.B: It is in fact possible to design an estimator that achieves this rate. Recall that for Gaussian mean estimation in k dimensions, the minimax rate is $\sigma^2 k/n$, whereas for univariate Gaussians, it is σ^2/n. The additional $\log(d/k)$ factor for sparse estimation can be viewed as the price for identifying which of the d coordinates are non-zero.

4 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.
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:

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

[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 $\mathbb{R}^d$.

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.