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

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

Homework 1. Due 09/27/2025, 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 (*). 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 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: $\mathrm{H}^2(P,Q) = \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x$.

Chi squared divergence: $\chi^2(P,Q) = \int \frac{(p(x)-q(x))^2}{q(x)}\, dx = \mathbb{E}_Q\left[\left(\frac{p(X)}{q(X)} - 1\right)^2\right]$.

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 $a, b > 0$?
6. **[3 pts]** $\mathrm{KL}(P,Q) \leq \chi^2(P,Q)$.
   Hint: You may use the inequality $\log(x) \leq x - 1$ for $x > 0$.

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,
$$R_n^\star \geq \frac{1}{4}\Phi\left(\frac{\delta}{2}\right) \exp\left(-\mathrm{KL}(P_0^n, \overline{P})\right)$$

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 PAC lower bounds for normal mean estimation

We are given n independent samples $S = \{X_1, \ldots, X_n\}$, where each X_i is sampled from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with *unknown* mean μ, but known variance σ^2. Let $\epsilon > 0$ be given. We wish to design an estimator $\widehat{\mu} : \mathbb{R}^n \to \mathbb{R}$ which is ϵ close to μ 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. **[2 pts]** *(Upper bound)* Design an estimator $\widehat{\mu}$ for μ 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. **[7 pts]** *(Lower bound)* Next, show that

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

4 Sparse Normal Estimation *

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 k-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:

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

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 $\|\widehat{\theta} - \theta\|_2^2$. In this problem, you will show

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

You may assume that d and/or k is sufficiently large, but $k \in o(d)$.

1. **[8 pts]** *(Lower bound)* Using Fano's method or otherwise, show that $R_n^\star \in \Omega\left(\frac{\sigma^2 k \log(d/k)}{n}\right)$.

   Hint. You may use the following version of the Gilbert-Varshamov bound that applies to sparse binary vectors: For any $\omega, \omega' \in \{0,1\}^d$, let $H(\omega, \omega') = \sum_{i=1}^d \mathbb{1}(\omega_i \neq \omega_i')$ denote the Hamming distance. Then, for all integers k such that $1 \leq k \leq d/8$, there exists a set of N binary vectors $\omega_1, \ldots, \omega_N \in \{0,1\}^d$ such that the following hold: *(i)* $H(\omega_i, \omega_j) \geq k/2$ for all $i, j \in [N]$, *(ii)* $N \geq \left(1 + \frac{d}{2k}\right)^{\frac{k}{8}}$, *(iii)* $|\omega_j|_0 = k$ for all $j \in [N]$.

2. *(Upper bound)* We will study the following estimator $\widehat{\theta}$. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i \in \mathbb{R}^d$ denote the sample mean. Let τ denote the k^{th} largest element of $\{|\overline{X}|_1, \ldots |\overline{X}|_d\}$. The i^{th} element of $\widehat{\theta}$ is given by,

$$\widehat{\theta}_i = \begin{cases} \overline{X}_i & \text{if } |\overline{X}_i| \geq \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

   (a) **[5 pts]** First show that $\|\theta - \widehat{\theta}\|_\infty \leq 2\|\theta - \overline{X}\|_\infty$ a.s.

   (b) **[5 pts]** Let $Z \sim \mathcal{N}(0, I_d)$ be a d-dimensional standard normal vector. Show that $\mathbb{E}[\|Z\|_\infty^2] \in \mathcal{O}(\log(d))$.

   Hint. You may use the fact that the moment generating function MGF of a χ_1^2 distribution is $\mathrm{MGF}(t) = \frac{1}{\sqrt{1-2t}}$ if $t < 1/2$ and $\mathrm{MGF}(t) = \infty$ otherwise.

3. **[3 pts]** Combining the results from parts 2a and 2b, upper bound the risk of the estimator in (1).

Acknowledgement. Problem 4.2 was based on a class project by Jingyun Jia, Jiaqi Tang, and Xinta Yang (2024).

5 Estimating a categorical distribution *

We are given n independent samples $S = \{X_1, \ldots, X_n\}$, where each X_i is sampled from a categorical distribution with d items $\mathrm{Categ}\,(\theta)$. Here $X_i \in \{0,1\}^d$ is a one-hot vector where $X_{i,k} = 1$ means that the i^{th} item selected was k, and $\theta \in \Delta^{d-1} = \left\{\theta' \in \mathbb{R}^d;\ \theta' \geq 0,\ \theta'^\top \mathbf{1} = 1\right\}$.

We wish to estimate the distribution θ in the ℓ_2^2 loss. The minimax risk is,

$$R^\star \stackrel{\Delta}{=} \inf_{\widehat{\theta}} \sup_{\theta \in \Delta^{d-1}} \mathbb{E}_{S \sim \mathrm{Categ}\,(\theta)^n} \left[\|\widehat{\theta}(S) - \theta\|_2^2 \right].$$

[16 pts] Find the minimax rate for this problem, in terms of the number of data n and the number of items d. You may assume that $d \in o(n)$.

Hint. You may find the following inequalities useful: *(i)* $\log(1+x) \leq x$ for all $x > -1$, *(ii)* $\log(1+x) \geq x - x^2/2$ for all $x \geq -0.68$.