ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Foundations of Data Science
Fall 2024

Assignment date: Friday, January 29, 2025, 9:15
Due date: Friday, January 29, 2025, 12:15

Final Exam – INJ 218

This exam is open book. No electronic devices of any kind are allowed. There are 5 problems. Choose the ones you find easiest and collect as many points as possible. We do not necessarily expect you to finish all of them. Good luck!

Name: _____

Problem 1	/ 12
Problem 2	/ 10
Problem 3	/ 9
Problem 4	/ 12
Problem 5	/ 9
Total	/52

Problem 1 (Donsker-Varadhan to Pinsker inequality – 12 pts)**.** In this problem, we further explore information measures.

Remark: If you refer to class materials, be precise (Theorem or equation numbers, Homework problem identifiers and so on.) Your overall argument must be complete.

Let Z be an arbitrary random variable and let $f(z)$ be an arbitrary function satisfying $0 \leq f(z) \leq b$ for all values z.

(i) [3 pts] Prove that for any distribution Q, we have (recall $0 \leq f(z) \leq b$)

$$\log \mathbb{E}_Q \left[e^{f(Z)} \right] \leq \mathbb{E}_Q \left[f(Z) \right] + \frac{1}{8} b^2, \tag{1}$$

where as in class, the notation $\mathbb{E}_Q \left[e^{f(Z)} \right]$ means that the expectation is taken assuming that Z is distributed according to Q.

HINT: Observe that irrespective of the distribution of Z, we know that the random variable $f(Z) \in [0, b]$. Like in class, use this information to bound the moment generating function of the random variable $f(Z)$.

(ii) [3 pts] Prove that for any distributions P and Q, we have (recall $0 \leq f(z) \leq b$)

$$\mathbb{E}_P[f(Z)] - \mathbb{E}_Q \left[f(Z) \right] \leq D(P \| Q) + \frac{1}{8} b^2. \tag{2}$$

where the KL divergence is computed with respect to the natural logarithm.

(iii) [3 pts] Prove that for arbitary distributions P and Q,

$$\max_f \mathbb{E}_P[f(Z)] - \mathbb{E}_Q[f(Z)] = \frac{b}{2} \|P - Q\|_1, \tag{3}$$

where the maximum is over all functions $f(z)$ satisfying $0 \leq f(z) \leq b$ for all values z.

(iv) [3 pts] Using Parts (ii) and (iii), prove the Pinsker inequality (Example 4.1 in the lecture notes). That is, prove that for arbitary distributions P and Q, we have

$$\|P - Q\|_1 \leq \sqrt{2 D(P \| Q)}, \tag{4}$$

where the KL divergence is computed with respect to the natural logarithm.

(space for problem 1)

(space for problem 1)

(space for problem 1)

Problem 2 (χ^2 Divergence Distance Measure – 10 pts)**.** In class we defined the ℓ_1 distance, the ℓ_2 distance, as well as the KL divergence measure. But there are other distance measures that are important and used in practice.

One of those is the χ^2 divergence distance measure. It is defined as

$$\chi^2(p,q) = \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{q_i}.$$

Recall the definition of the min-max loss for a given distance measure L, an alphabet size of k and assuming we have n iid samples:

$$r_{k,n}^{L} = \min_{q} \max_{p \in \Delta_k} \mathbb{E}_{X^n \sim p}[L(p, q(X^n))].$$

(i) [2 pt] Show that, alternatively, $\chi^2(p,q) = \sum_{i=1}^{k} \frac{p_i^2}{q_i} - 1$.

(ii) [8 pts] Show that for $k \geq 2$ and $n \geq 1$, $r_{k,n}^{\chi^2} \leq \frac{k-1}{n+1}$. [Note: There is also a corresponding lower bound that is fairly close to this upper bound showing that this upper bound is relatively tight, but we will only be concerned with the upper bound.]

HINT: We are looking for an upper bound on the min-max loss. Hence, we are free to consider any particular estimator. The add+1 estimator is your friend. Also, remember that $\frac{\binom{n}{t}}{t+1} = \frac{\binom{n+1}{t+1}}{n+1}$.

(space for problem 2)

(space for problem 2)

(space for problem 2)

Problem 3 (Exponential Families and Conjugate Priors – 9 pts)**.** Let $p_\theta(x) = h(x)e^{\langle \phi(x), \theta \rangle - A(\theta)}$ denote a generic exponential family with sufficient statistics $\phi(x)$ and parameter θ.

Assume that we receive iid samples from this family, call them $\{x_i\}_{i=1}^n$. From these samples, we want to infer the unknown parameter θ via a maximum a-posteriori (MAP) procedure.

In order to apply a MAP procedure we need to define a prior distribution on the parameter θ. Consider the family of prior distributions $q_{\mu,\lambda}(\theta) = K(\mu, \lambda)e^{\langle \theta, \mu \rangle - \lambda A(\theta)}$, parametrized by (μ, λ). Note that this is also an exponential family. However, we have written it in a slightly non-standard form, where $K(\mu, \lambda)$ denotes the normalization constant which is a function of the parameters (μ, λ).

(i) [3 pts] Write down the posterior distribution $p_{\mu,\lambda}(\theta \mid x_1, \cdots, x_n)$ for a fixed set of parameters (μ, λ).

(ii) [3 pts] If you have not already done so in part (i), write the posterior as explicitly and compactly as you can. Justify why we called $q_{\mu,\lambda}(\theta)$ a conjugate prior.

(iii) [3 pts] Derive the MAP estimator of the parameter θ given the samples $\{x_i\}_{i=1}^n$ starting with the posterior derived in (ii). When will the estimate be unique?

(space for problem 3)

(space for problem 3)

(space for problem 3)

Problem 4 (Fano method – 12 pts)**.** In this problem, we will develop a framework to find lower bounds on the estimation error of the minimax distribution estimator. We will use Fano's inequality, which we saw in class. First, recall the minimax distribution estimation problem:

$$r_{k,n}^L = \min_q \sup_{p \in \Delta_k} \mathbb{E}_{X^n \sim p^n}[L(p, q(X^n))].$$

Assume that the loss L is symmetric in its arguments, satisfies the triangle inequality, and that $L(x, x) = 0 \ \forall x$.

(i) [3 pts] Let $\mathcal{P} := \{P_1, \ldots, P_m\}$ be a collection of distributions such that $L(P_i, P_j) \geq \delta > 0$ for $i \neq j$.

Show that

$$\sup_{p \in \Delta_k} \mathbb{E}_{X^n \sim p}[L(p, q(X^n))] \geq \frac{\delta}{2} \max_j \mathbb{P}_{X^n \sim P_j^n} \left[L(P_j, q(X^n)) \geq \frac{\delta}{2} \right].$$

HINT: For a non-negative random variable X we have $\mathbb{E}[X] \geq \epsilon \, \mathbb{P}[X \geq \epsilon]$.

(ii) [3 pts] Now let $V, X^n \sim P_{V,X^n}$ be jointly distributed such that V is uniformly distributed over $[1 : m]$ and $\mathbb{P}[X^n = x^n | V = j] = P_j^n(x^n)$. Define $Z := \arg\min_j L(q(X^n), P_j)$. Show that

$$\mathbb{P}[Z \neq V] \leq \max_j \mathbb{P}_{X^n \sim P_j^n} \left[L(q(X^n), P_j) \geq \frac{\delta}{2} \right].$$

(iii) [3 pts] Use Fano's inequality to show that

$$\max_j \mathbb{P}_{X^n \sim P_j^n} \left[L(p, q(X^n)) \geq \frac{\delta}{2} \right] \geq 1 - \frac{I(X^n; V) + \log 2}{\log m}.$$

HINT: Recall $I(Y; W) = H(Y) - H(Y|W)$.

(iv) [3 pts] Show that

$$I(X^n; V) \leq \frac{1}{m^2} \sum_{i,j \in [1:m]} D(P_i^n \| P_j^n) \leq n \max_{i,j} D(P_i \| P_j),$$

and thus

$$r_{k,n}^L \geq \frac{\delta}{2} \left(1 - \frac{n \max_{i,j} D(P_i \| P_j) - \log 2}{\log m} \right).$$

HINT: $I(W; Y) = D(P_{W|Y} \| P_W | P_Y)$, and the KL divergence is a convex function.

(space for problem 4)

(space for problem 4)

(space for problem 4)

Problem 5 (Dual Basis – 9 pts)**.** Consider a Hilbert space H. Let $G \subseteq H$ be a (Hilbert) subspace of H, exactly like in class. We are given a basis $\{\mathbf{g}_n\}_{n=1}^N$ that spans G but is *not* orthonormal.

(i) [3 pts] Show that there exists a so-called dual basis $\{\tilde{\mathbf{g}}_n\}_{n=1}^N$ that also spans G and has the property that

$$\langle \mathbf{g}_n, \tilde{\mathbf{g}}_m \rangle = \begin{cases} 1, & \text{for } m = n, \\ 0, & \text{for } m \neq n. \end{cases} \tag{5}$$

HINT: Start by considering $\tilde{\mathbf{g}}_1 = \alpha(\mathbf{g}_1 - \sum_{n=2}^N \beta_n \mathbf{g}_n)$. Argue that we can select α and β_n appropriately. No need to give explicit formulas for these coefficients. The more "rigorous" your argument, the more points you will get.

(ii) [3 pts] Show that for any $\mathbf{x} \in H$, the minimum of $\|\mathbf{y} - \mathbf{x}\|$ over all $\mathbf{y} \in G$ is attained by the selection

$$\mathbf{y}^* = \sum_{n=1}^N \langle \mathbf{x}, \tilde{\mathbf{g}}_n \rangle \mathbf{g}_n. \tag{6}$$

HINT: As in class, here $\|\cdot\|$ denotes the Hilbert space norm induced by the inner product.

(iii) [3 pts] Show that for any $\mathbf{x} \in H$, we have

$$\sum_{n=1}^N \langle \mathbf{x}, \tilde{\mathbf{g}}_n \rangle \mathbf{g}_n = \sum_{m=1}^N \langle \mathbf{x}, \mathbf{g}_m \rangle \tilde{\mathbf{g}}_m. \tag{7}$$

(space for problem 5)

(space for problem 5)

(space for problem 5)