
Problem Set 4 Distribution Estimation and Property Testing

For the Exercise Sessions of Oct 13-29 — **Due: Tue, November 3, 10am, on Moodle**

1 The Homework

The following three problems are the core of this homework. You work out solutions and turn them in. Problem 1 is suitable for the exercise session October 13. Problem 2 is suitable for October 29.

Problem 1: χ^2 Divergence Distance Measure

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 $\frac{\binom{n}{t}}{t+1} = \frac{\binom{n+1}{t+1}}{n+1}$.

Problem 2: Property Testing: Variance

A colleague claims to have implemented an algorithm which outputs i.i.d. samples distributed according to a discrete distribution P that has unit variance. Your task is to design a statistic to test whether this is indeed true.

Let Δ_k be the set of probability distributions on the alphabet $\mathcal{X} = \{1, \dots, k\}$. Assume that $P \in \mathcal{P} \cup \mathcal{Q}$ with $\mathcal{P} := \{P \in \Delta_k : P \text{ has variance } 1\}$ and $\mathcal{Q} := \{P \in \Delta_k : P \text{ has variance } \in [0, 1 - \epsilon] \cup [1 + \epsilon, \infty)\}$, where $0 < \epsilon < 1$. You are given n samples $\{X_i\}_{i=1}^n$, where the X_i are independent copies sampled according to P .

Remark: For the following three questions we do not ask you to write down a proof (or explicit calculation) that your proposed solution works.

- a) We say that an estimator $e : S^n \mapsto \Pi$ on a sample S^n of length n (ϵ, δ) -learns a parameter $p \in \Pi$ if for any $(\epsilon, \delta) \in (0, 1)^2$, given sufficiently many samples n , we have that $\mathbb{P}(\{|e(S^n) - p| > \epsilon\}) < \delta$. Give a brief explanation (one sentence, no calculations) why the empirical estimator of the second moment $\hat{\mu}_{X^2} := \frac{1}{n} \sum_{i=1}^n X_i^2$ can (ϵ, δ) -learn the second moment in our setting.
- b) First, assume that a genie tells you that X has zero mean. Design a simple test statistic and give a threshold in order to check for the above mentioned unit variance property.
Hint: Use the claim in a).
- c) Now consider the more general case where X can have arbitrary mean. Again, design a simple test statistic and give a threshold.
Hint: You can assume that $\hat{\mu}_X^2$ (ϵ, δ) -learns $\mathbb{E}[X]^2$.

2 Additional Problems

For the next six problems, you do not need to turn in solutions. For the exam, we do expect you know how to solve these problems. So, you may keep these problems around and tackle them in preparation for the Midterm and/or Final exam. Or you may solve them now.

Problem 3: Add- β Estimator

The add- β estimator $q_{+\beta}$ over $[k]$, assigns to symbol i a probability proportional to its number of occurrences plus β , namely,

$$q_i \stackrel{\text{def}}{=} q_i(X^n) \stackrel{\text{def}}{=} q_{+\beta, i}(X^n) \stackrel{\text{def}}{=} \frac{T_i + \beta}{n + k\beta}$$

where $T_i \stackrel{\text{def}}{=} T_i(X^n) \stackrel{\text{def}}{=} \sum_{j=1}^n \mathbf{1}(X_j = i)$. Prove that for all $k \geq 2$ and $n \geq 1$,

$$\min_{\beta \geq 0} r_{k, n}^{l_2^2}(q_{+\beta}) = r_{k, n}^{l_2^2}(q_{+\sqrt{n}/k}) = \frac{1 - \frac{1}{k}}{(\sqrt{n} + 1)^2}$$

Furthermore, $q_{+\sqrt{n}/k}$ has the same expected loss for every distribution $p \in \Delta_k$.

Problem 4: Bayes Risk

Consider the task of parameter estimation, where the parameter $\underline{W} \in \mathbb{R}^d$ is distributed uniformly on the ball $\mathcal{W} = \{\underline{w} \in \mathbb{R}^d : \|\underline{w}\|_2 \leq a\}$. We get samples $\underline{X}_i = \underline{W} + \underline{Z}_i$ with $\underline{Z}_i \sim \mathcal{N}(0, \sigma^2 I_d)$ independent of $\underline{W}$ for $i = 1, \dots, n$, and let $\ell(\underline{w}, \hat{\underline{w}}) = \|\underline{w} - \hat{\underline{w}}\|_2$ be the loss function.

We define the Bayes risk as $R_B = \min_{\hat{\underline{W}}} \mathbb{E}_{\underline{W}, \hat{\underline{W}}}[\ell(\underline{W}, \hat{\underline{W}})]$. Show that the Bayes risk is upperbounded as $R_B \leq \sqrt{\frac{\sigma^2 d}{n}}$.

Hint: Use that the estimator $\hat{\underline{W}} = \frac{1}{n} \sum_i \underline{X}_i$ achieves the desired upperbound.

Problem 5: ℓ_1 versus Total Variation

In class we defined the ℓ_1 distance as

$$\|p - q\|_1 = \sum_{i=1}^k |p_i - q_i|.$$

Another important distance is the total variation distance $d_{\text{TV}}(p, q)$. It is defined as

$$d_{\text{TV}}(p, q) = \max_{S \subseteq \{1, \dots, k\}} \left| \sum_{i \in S} (p_i - q_i) \right|.$$

Show that if p, q are two probability mass vectors (i.e. elements of the simplex) we have that $d_{TV}(p, q) = \frac{1}{2} \|p - q\|_1$.

Problem 6: Poisson Sampling

Assume that we have given a distribution p on $\mathcal{X} = \{1, \dots, k\}$. Let X^n denote a sequence of n iid samples. Let $T_i = T_i(X^n)$ be the number of times symbol i appears in X^n . Then

$$\Pr\{T_i = t_i\} = \binom{n}{t_i} p_i^{t_i} (1 - p_i)^{n-t_i}.$$

Note that the random variables T_i are *dependent*, since $\sum_i T_i = n$. This dependence can sometimes be inconvenient.

There is a convenient way of getting around this problem. This is called *Poisson* sampling. Let N be a random variable distributed according to a Poisson distribution with mean n . Let X^N be then an iid sequence of N variables distributed according to p .

Conditioned on $N = n$, what is the induced distribution of the Poisson sampling scheme?

Show that

1. $T_i(X^N)$ is distributed according to a Poisson random variable with mean $p_i n$.
2. The $T_i(X^N)$ are independent.

Problem 7: Estimating Support Size

You are attending Balelec. You want to estimate how many people are attending. Let this number be m . Here is a very simple algorithm. You walk around randomly. Every 5 minutes you take a picture of the person who is right next to at this moment. Assume that 5 minutes is sufficiently long so that in this manner you sample participants at Balelec with uniform probability. Assume further that during the whole time you do your experiment no person joins or leaves Balelec.

You do this N times, where N is a Poisson random variable with mean $n = 100$. Once you are done you look at the photos. Assume that in total you have encountered $K = 102$ distinct people. Out of those 102, 100 you have seen only once, one you saw twice, and one you saw three times. Give an estimate of the number of people attending Balelec (the support size of the distribution). Call this number $\hat{m}$. We do not expect a number as answer since the estimate might involve an optimization step which might not be trivial to do by hand. Simplify as far as you can and then write down how you would get final answer.

Hint: Follow your own path or answer the question according to the following steps.

1. Assume that there are m people attending Balelec. Take a specific person at Balelec. Call this person "1". Given the procedure outlined above, what is the probability that this person appears c_1 , $c_1 \geq 0$, times on your photos?
2. Now take two specific people. Call them "1" and "2". What is the probability that they appear $\{c_i\}_{i=1}^2$ times on your photos?
3. Now consider all people at Balelec together. Assume as before that each has a specific identity. What is the probability that the m people appear $\{c_i\}_{i=1}^m$ times on your photos?
4. Assume again that m people attend Balelec and also as before that we have the counts $\{c_i\}_{i=1}^m$. But this time we do not know who has what count, i.e., we do not know the identities of the people. All we know is the counts themselves. What is the probability of getting the counts $\{c_i\}_{i=1}^m$? [Note: What we see are the non-zero counts, but since we also assume that we know m , we know in fact all counts.]
5. How can you use the last expression to derive an estimate?