Advanced Learning Theory **CS880, Fall 2019**

Problem Set 1: Basics of Distribution Learning and Testing

Due: Monday, October 7, by email

Please title your email "CS880_PS1".

---

1. (Cover-Based Learning) Let $\mathcal{D}$ be a family of probability distributions and $0 < \epsilon < 1$. A set of distributions $\mathcal{D}_\epsilon$ is called an *ϵ-cover* of $\mathcal{D}$ with respect to a metric $d(\cdot,\cdot)$ if for any $P \in \mathcal{D}$ there exists $P' \in \mathcal{D}_\epsilon$ such that $d(P,P') \le \epsilon$. In this problem, you will show that the existence of a small cover for $\mathcal{D}$ under the total variation distance implies the existence of a sample-efficient learning algorithm for $\mathcal{D}$. More specifically, you will be guided through the proof of the following theorem:

   Theorem 1. *Let $\mathcal{D}$ be a distribution family over a discrete domain. For $\epsilon > 0$, let $\mathcal{D}_\epsilon$ be an ϵ-cover of $\mathcal{D}$ of size M, under the total variation distance $\mathrm{d_{TV}}(,)$. There is an algorithm that uses $O(\epsilon^{-2}\log M)$ samples from an unknown distribution $P \in \mathcal{D}$ and, with probability at least $9/10$, outputs a distribution $Q \in \mathcal{D}_\epsilon$ that satisfies $\mathrm{d_{TV}}(P,Q) \le 6\epsilon$.*

   In parts (i)-(iv) you are asked to analyze an algorithm that establishes Theorem 1 and explore its performance. In part (v), you are asked to apply Theorem 1 to obtain a sample-efficient learning algorithm for a natural family of high-dimensional structured distributions.

   (i) Consider the following subroutine to select between two candidate hypotheses:

   > `Choose-Hypothesis`$(H_1, H_2, \epsilon, \delta)$
   > Input: Sample access to discrete distribution P; a pair of hypothesis distributions (H_1, H_2); $\epsilon, \delta > 0$.
   >
   > Let $\mathcal{W}$ be the support of P, $\mathcal{W}_1 := \{w \in \mathcal{W} \mid H_1(w) > H_2(w)\}$, and $p_1 = H_1(\mathcal{W}_1)$, $p_2 = H_2(\mathcal{W}_1)$.
   >
   > (a) If $p_1 - p_2 \le 5\epsilon$, return either H_i. Otherwise:
   >
   > (b) Draw $m = 2\log(1/\delta)/\epsilon^2$ samples $s_1, \ldots, s_m$ from P, and let $\tau = \frac{1}{m}|\{i \mid s_i \in \mathcal{W}_1\}|$ be the fraction of samples that fall inside $\mathcal{W}_1$.
   >
   > (c) If $\tau > p_1 - \frac{3}{2}\epsilon$, return H_1; otherwise,
   >
   > (d) if $\tau < p_2 + \frac{3}{2}\epsilon$, return H_2; otherwise,
   >
   > (e) return either H_i.

   Suppose that $\mathrm{d_{TV}}(P,H_1) \le \epsilon$. Show that if $\mathrm{d_{TV}}(P,H_2) > 6\epsilon$, the probability that `Choose-Hypothesis`$(H_1, H_2, \epsilon, \delta)$ does not output H_1 is $O(\delta)$.

   (ii) Use subroutine `Choose-Hypothesis` for each pair of distributions in an ϵ-cover $\mathcal{D}_\epsilon$ of $\mathcal{D}$ to prove Theorem 1. What is the running time of your algorithm?

   (iii) Adapt your algorithm and its analysis so that it is robust to *model misspecification*. Specifically, show that if P is OPT-close in total variation distance to some distribution in $\mathcal{D}$, then your algorithm outputs a hypothesis distribution Q that satisfies $\mathrm{d_{TV}}(P,Q) = O(\mathrm{OPT}) + \epsilon$ with probability at least $9/10$.

(iv) Is the sample complexity of the algorithm in Theorem 1 information-theoretically optimal (within a constant factor)? Justify your answer.

(v) A binary product distribution is a probability distribution supported on $\{0,1\}^n$ whose probability mass function is a product of n independent Bernoulli random variables. A *mixture* of k binary product distributions $\pi_1, \pi_2, \ldots, \pi_k$ is a distribution π over $\{0,1\}^n$ such that for some $w = (w_1, \ldots, w_k)$ with $w_i \geq 0$ and $\sum_i w_i = 1$ it holds $\pi = \sum_i w_i \pi_i$. Show that the set of mixtures of binary product distributions can be learned to total variation distance ϵ with confidence probability 9/10 using $\mathrm{poly}(n, k, 1/\epsilon)$ samples. What is the running time of your algorithm?

2. (Learning Discrete Distributions under Different Loss Functions) In class we studied learning discrete distributions under the total variation distance metric. In this problem, we explore distribution learning under different loss functions.

(i) The *Kolmogorov distance* between the probability mass functions $p, q : [n] \to [0,1]$ is defined as $\mathrm{d_K}(p,q) := \max_{u \in [n]} |p([1,u]) - q([1,u])|$, where $p([1,u]) := \sum_{k=1}^{u} p(k)$.

(a) Give an algorithm to learn an arbitrary distribution over $[n]$ up to Kolmogorov distance ϵ with probability $1-\delta$ using $O((1/\epsilon^2)\log(1/(\epsilon\delta)))$ samples. How do you explain the fact that your algorithm has sample complexity independent of the domain size n?

(b) Show that any learning algorithm for the aforementioned learning problem requires $\Omega((1/\epsilon^2)\log(1/\delta))$ samples.

(c) [Optional] Show that the upper bound of part (a) can be improved to $O((1/\epsilon^2)\log(1/\delta))$.

(ii) The *chi-squared loss* between distributions p and q supported on $[n]$, is defined as $\chi^2(p,q) := \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i}$. Note that the chi-squared loss is not a metric.

(a) Give a learning algorithm $\mathcal{A}$ with the following performance guarantee: Given iid samples from an unknown arbitrary distribution p over $[n]$, $\mathcal{A}$ outputs a hypothesis distribution q that satisfies $\chi^2(p,q) \leq \epsilon$ with probability at least 9/10. Analyze the sample complexity and runtime of your algorithm.

(b) Prove a sample complexity lower bound for the corresponding learning problem, i.e., a lower bound on the sample complexity of any algorithm for the problem. Is the sample complexity of your algorithm in (a) above information-theoretically optimal, within a constant factor?

(iii) A distribution p on the ordered domain $[n]$ is called *k-piecewise affine* if there exists a partition $\mathcal{I}_k$ of $[n]$ into k intervals $I_1 = [1, i_1], I_2 = [i_1 + 1, i_2], \ldots, I_k = [i_{k-1} + 1, n]$ such that the probability mass function p is an affine function within each interval I_j, $j \in [k]$. That is, for each $j \in [k]$, there exist parameters $a_j, b_j \in \mathbb{R}$ such that $p_i = a_j i + b_j$, for all $i \in I_j$.

(a) Suppose that the partition $\mathcal{I}_k$ is fixed and known. Give an algorithm that learns an arbitrary k-piecewise affine distribution with respect to $\mathcal{I}_k$ under the total variation distance. Show that the sample complexity of your algorithm is optimal, within a constant factor.

(b) Suppose that k is known but the partition $\mathcal{I}_k$ is *unknown*. Prove an upper bound on the sample complexity of learning an arbitrary k-piecewise affine distribution under the total variation distance. Is your upper bound best possible?

3. (Testing Families of Distributions) In this problem, we will explore the complexity of testing simple distribution properties under structural assumptions.

   (i) Let $p : [n] \to [0, 1]$ be a probability mass function that is *promised to be monotone non-increasing* in its ordered domain, i.e., $p_{i+1} \leq p_i$ for all $i \in [n-1]$. Design a uniformity tester for p, i.e., an algorithm that distinguishes with probability at least $1 - \delta$ between the cases that $p = U_n$ and $\mathrm{d_{TV}}(p, U_n) \geq \epsilon$. Show that your algorithm is sample-optimal, up to constant factors.

   (ii) Let $p : [n] \to [0, 1]$ be a probability mass function with the property that each p_i is an integer multiple of $1/n$. Give a uniformity tester for p and a matching sample complexity lower bound.

   (iii) Let $p : [n] \to [0, 1]$ be a probability mass function. We want to test *whether* p is monotone non-increasing over its ordered domain. In more detail, let $\mathcal{M}_n$ be the family of all monotone non-increasing distributions over $[n]$. We want to design an algorithm that distinguishes with probability at least $9/10$ between the cases that $p \in \mathcal{M}_n$ and $\mathrm{d_{TV}}(p, \mathcal{M}_n) \geq \epsilon$. Prove a sample complexity lower bound for this testing problem. [Extra credit for matching upper bound.]

   (iv) Let Bin_n be the family of Binomial distributions on $[n]$. Give a tester for the property Bin_n, i.e., an algorithm that distinguishes between the case that $p \in \mathrm{Bin}_n$ and $\mathrm{d_{TV}}(p, \mathrm{Bin}_n) \geq \epsilon$ with probability at least $2/3$. Prove a matching sample complexity lower bound.

4. (Distribution Testing Under Different Loss Functions) In this problem, we explore distribution testing under different losses.

   (i) Let $p, q : [n] \to [0, 1]$ be two unknown k-piecewise affine distributions over the *known* partitions $\mathcal{I}_k$ and $\mathcal{J}_k$ respectively. Give a tester and a matching sample complexity lower bound for the problem of testing equivalence between p and q (in total variation distance).

   (ii) The *KL-divergence* between the probability mass functions $p, q : [n] \to [0, 1]$ is defined as $\mathrm{d_{KL}}(p, q) := \sum_{i=1}^{n} p_i \ln(p_i/q_i)$. Determine the sample complexity of identity testing between an unknown distribution p and a known distribution q with respect to KL-divergence.

   (iii) The *Hellinger distance* between the probability mass functions $p, q : [n] \to [0, 1]$ is defined as $\mathrm{d_H}(p, q) := (1/\sqrt{2})\sqrt{\sum_{i=1}^{n}(\sqrt{p_i} - \sqrt{q_i})^2}$.

      (a) Determine the sample complexity of uniformity testing with respect to the Hellinger distance. That is, give a uniformity testing algorithm and a lower bound on the sample complexity matching that of your algorithm.

      (b) [Optional] Determine the sample complexity of equivalence testing (between two unknown discrete distributions) with respect to the Hellinger distance.