Warm-up

Problem 1. Check your understanding: how does the Pearson–Neyman lemma (Lemma 49.1) imply that Alice-Bob game interpretation?

Problem 2. Prove the upper bound of Corollary 50.1 directly, via Hoeffding.

Problem 3. Show that ℓ_2 and ℓ_∞ distances between distributions:

$$\ell_2(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_2 = \sqrt{\sum_{x \in \mathcal{X}} (\mathbf{p}(x) - \mathbf{q}(x))^2}, \quad \ell_\infty(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_\infty = \max_{x \in \mathcal{X}} |\mathbf{p}(x) - \mathbf{q}(x)|$$

do not satisfy the Data Processing Inequality.

Problem 4. Prove Scheffé's lemma. *(Hint: consider the set $S = \{x \in \mathcal{X} : \mathbf{p}(x) > \mathbf{q}(x)\}$.)*

Problem solving

Problem 5. Prove the two "suboptimal" sample complexities for learning distributions. For the second, explain how to get rid of the assumption on $\min_i p_i$ (possibly losing some constant factors in the sample complexity).

Problem 6. Instead of looking at all $\binom{n}{2}$ possible pairs of samples in Algorithm 21 for uniformity testing, describe and analyse the tester which partitions the n samples into $\frac{n}{2}$ (independent) pairs of samples, and use them to estimate $\Pr[X = Y]$. What is the resulting sample complexity?

Problem 7. *This is a programming exercise, to be done in, e.g., a Jupyter notebook.*

a) Write a function which, given two probability distributions represented as two arrays of the same size, computes their total variation distance.

b) Implement the empirical estimator seen in class: given the domain size k and a multiset of n numbers in $\{1, 2, \ldots, k\}$, return the empirical probability distribution over $\{1, 2, \ldots, k\}$.

c) Implement the uniformity testing algorithm (Algorithm 21).

d) Import the Canada's 6/49 lotto dataset (from `https://www.kaggle.com/datasets/datascienceai/lottery-dataset`, available on Ed).

e) Learn the distribution of the first number, from the $n = 3,665$ samples. Plot the result.

f) Test whether the distribution of the "bonus number" is uniform, from the $n = 3,665$ samples, for $\varepsilon \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Report the results.

g) Learn the distribution of the "bonus number", from the $n = 3,665$ samples, and compute the total variation distance between the resulting $\widehat{\mathbf{p}}$ and the uniform distribution on $\{1, 2, \ldots, 49\}$.

Advanced

Problem 8. Consider the following alternative approach to learn a probability distribution over a domain $\mathcal{X}$ of size k:

1. Take n i.i.d. samples from $\mathbf{p}$
2. Compute, for every domain element $i \in \mathcal{X}$, the number n_i of times it appears among the n samples.
3. For every $i \in \mathcal{X}$, let
$$\widehat{\mathbf{p}}(i) = \frac{n_i + 1}{n + k}$$
4. return $\widehat{\mathbf{p}}$

(This is called the *Laplace estimator*. Note that, in contrast to the empirical estimator, it assigns non-zero probability to every element of the domain, even those that do not appear in the samples.)

a) Show that $\widehat{\mathbf{p}}$ is a probability distribution.

b) Define the *chi-squared divergence* between probability distributions as
$$\chi^2(\mathbf{p} \| \mathbf{q}) = \sum_{x \in \mathcal{X}} \frac{(\mathbf{p}(x) - \mathbf{q}(x))^2}{\mathbf{q}(x)}$$
(Note that this is not symmetric, and not bounded!) Show that $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})^2 \leq \frac{1}{4}\chi^2(\mathbf{p} \| \mathbf{q})$ for every $\mathbf{p}, \mathbf{q}$.

c) Show that $\mathbb{E}[\chi^2(\mathbf{p} \| \widehat{\mathbf{p}})] \leq \frac{k-1}{n+1}$.

d) Conclude on the value of n sufficient to learn $\mathbf{p}$ to total variation distance ε using the Laplace estimator.