Problem 1 is important to look at, to build some familiarity and conceptual understanding of the result. Problems 2 to 4 do not really require you to have read the lecture notes in detail or watched the lecture (just look at the corresponding definitions). They are not difficult, but somewhat on the technical side, and can be skipped, though Problem 2 is good practice and the idea of Problem 4 is important (though the proof is a little tedious).

Problem 5 is good practice, but not crucial and quite long, and can be skipped during the tutorial if you are short on time. Go back to it later on for practice, on your own time.

Problem 6 is just a "sanity check" to see while an obvious approach does not work. Think of it as optional, no need to focus on it in a first pass.

Problem 6 is on the more difficult side, but worth doing to understand why the MG algorithms is a sketching algorithm.

Problem 7 is important, as it applies to a practical (simple) example the algorithms seen in the lecture.

Problem 8 (Advanced) is interesting "to go further", and will help you build some understanding of the techniques used in this area. Worth doing after going through the others.

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.