Homework 6

Due: 15 November, 2024

Each problem is graded on the coarse scale of ✓$^+$, ✓, ✓$^-$ and no ✓. It is also assigned a multiplier, denoting the relative importance of the problem. Both correctness and presentation are grading criteria.

Please read and make sure you understand the collaboration policy on the course missive. Extra credit problems are clearly marked below (see course missive for the details of grade calculations).

Remember to prove all your (non-elementary and not shown in class) mathematical claims, unless stated otherwise.

Each group of students should submit only 1 pdf to the corresponding Canvas assignment.

Problem 1

(2 ✓s)

Consider a Bernoulli coin (over $\{0, 1\}$) with bias p. Show the Poissonisation guarantees for this special case, that is, if we sample $\mathrm{Poi}(n)$ many coins, we get:

(a) The number of 0s is distributed as $\mathrm{Poi}(n(1-p))$ and the number of 1s is distributed as $\mathrm{Poi}(np)$

(b) The two quantities are independent

Problem 2

(3 ✓s)

Consider the mean estimation problem under adversarial data corruption (under any of the corruption models mentioned in class). Recall from class that, if there is non-zero corruption that an adversary can make, then, even if we get infinitely many samples, it is impossible to get 0 mean estimation error. The goal of this problem is to show a mean estimation error lower bound in this infinite-sample regime, under data corruption.

1. (2 ✓) Fix a parameter $\eta \in (0, \frac{1}{2})$. Construct two (1-dimensional, $\mathbb{R}$-valued) distributions whose total variation distance is $O(\eta)$ such that

   (a) Both distributions have variance 1 (or variance that goes to 1 as $\eta \to 0$ anyway)

   (b) The difference in mean between them is $\Omega(\sqrt{\eta})$

   (Hint: Construct the distributions on support $\{-1/\sqrt{\eta}, 0, 1/\sqrt{\eta}\}$.)

2. (1 ✓) Show that, under the strong contamination model with corruption budget η, even when the number of samples $n \to \infty$, every algorithm must get estimation error at least $\Omega(\sqrt{\eta})$ with probability at least 1/3.

Problem 3

(3 ✓s)

Consider a special case of identity testing against the known distribution $\mathbf{p}$ over $[n]$, where $\mathbf{p}$ is guaranteed to have a support size upper bounded by $k < n$. (Recall that $\mathbf{p}$ is known to the tester and therefore the true support size $\leq k$ is also known to the tester.) What is the sample complexity of this testing problem? Give and prove upper and lower bounds that are tight up to constant multiplicative factors (that is, big-O tight). You may assume all the results stated in lectures.

Problem 4

(3 ✓s, **Extra credit**)

In this problem, we explore (the easier part of) a different algorithmic approach to identity testing – via an elegant reduction to uniformity testing, due to Oded Goldreich.

To get an intuitive sense of this approach, we will focus on a special case of identity testing, where the known distribution being tested against is *grained*, defined as follows.

Definition 1 (Grained distribution)**.** We say that a distribution $\mathbf{p}$ (over $[n]$, say) is *m-grained* if for all $i \in [n]$, p_i is a multiple of $\frac{1}{m}$.

Reduce the identity testing problem against a m-grained distribution $\mathbf{p}$ to uniformity testing over $[m]$, which takes $O(\sqrt{m}/\epsilon^2)$ samples.

(Hint: think about how, given the explicit description of $\mathbf{p}$, you would modify it into a uniform distribution over m elements. Given that, think about how you would generate samples from this uniform distribution given samples from $\mathbf{p}$.)