Homework 7

Due: 13 November, 2020

Each problem is graded on the coarse scale of $\checkmark^{+}$, $\checkmark$, $\checkmark^{-}$ and no $\checkmark$. 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. The problems for 2000-level credit are clearly marked below: students who are registered for 1000-level credit are welcome to solve and submit these problems for extra credit (see course missive for details).

Remember to prove all your (non-elementary) mathematical claims, unless stated otherwise.

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

Problem 1

(3 $\checkmark$s)

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

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

(b) The two quantities are independent

Problem 2

(2 $\checkmark$s)

In this problem, we design a sampling corrector for the property of *conditional independence* (or equivalently the *Markov* property in the following setting).

Consider a distribution $\mathbf{p}$ over the domain $[n_1] \times [n_2] \times [n_3]$. We say that $\mathbf{p}$ is conditionally independent if, conditioned on its second coordinate, its first and third coordinates are independent. That is, the three coordinates of $\mathbf{p}$ form a Markov chain.

Define $\mathbf{p}|y$ to be the independent distribution over $[n_1] \times [n_3]$ where $\mathbf{p}|y(x, z) = \mathbb{P}_{(X,Y,Z)\leftarrow\mathbf{p}}(X = x|Y = y) \cdot \mathbb{P}_{(X,Y,Z)\leftarrow\mathbf{p}}(Z = z|Y = y)$. Also define $\mathbf{p}_Y$ to be the marginal distribution of $\mathbf{p}$ over $[n_2]$. Lastly, define $\mathbf{p}_{X,Z|Y}$ to be the conditionally independent distribution over $[n_1] \times [n_2] \times [n_3]$ where $\mathbf{p}_{X,Z|Y}(x, y, z) = \mathbf{p}_Y(y) \cdot \mathbf{p}|y(x, z)$. Using the above notation, it follows that $\mathbf{p}$ is conditionally independent if and only if $\mathbf{p} = \mathbf{p}_{X,Z|Y}$.

The following fact is useful for designing a sampling corrector.

Fact 1. *Suppose the distribution* $\mathbf{p}$ *is* ϵ*-close to being conditionally independent, then* $\mathbf{p}$ *is* 4ϵ*-close to* $\mathbf{p}_{X,Z|Y}$.

Use the above fact to design an m-sample (ϵ,4ϵ)-sampling corrector for conditional independence which succeeds with probability 1 and has an expected average sample complexity of $O(n_2)$.

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, **2000-level**)

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 2 (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 "turn it" into a uniform distribution over m elements. Given that, think about how you would generate samples from this uniform distribution given samples from $\mathbf{p}$.)