Problem Set 9

Sublinear Algorithms

Due Thursday, November 19

Recall from class that, given m samples each from two distributions P and Q over $[n]$, we can distinguish between $P = Q$ and $\|P - Q\|_{TV} \geq \varepsilon$ with $O((n/\varepsilon^2)^{2/3} + \sqrt{n}/\varepsilon^2)$ samples.

1. Let (X, Y) be a pair of random variables drawn from a distribution P_{XY} over $[n] \times [m]$. Let P_X, P_Y be the marginal distributions of X and Y over $[n]$ and $[m]$, respectively. The goal of this question is, given samples of (X, Y) from an unknown distribution, to test if X and Y are mutually independent (i.e., P_{XY} is a product distribution) or ε-far from mutually independent.

   (a) Show how to simulate a sample from $P_X \times P_Y$ using two samples from P.

   (b) Show how to distinguish $P = P_X \times P_Y$ from $\|P - P_X \times P_Y\|_{TV} \geq \varepsilon$ using $O(n^{2/3}m^{2/3}/\varepsilon^2)$ samples of P.

   (c) Show how to distinguish between (X, Y) being independent, and ε-far in total variation distance from *any* independent distribution, with $O(n^{2/3}m^{2/3}/\varepsilon^2)$ samples. (This is sublinear in the number of possible outcomes, nm).

   (d) Now consider the problem of distinguishing between $I(X; Y) = 0$ and $I(X; Y) \geq \varepsilon$. Show that, for any two distributions $(X, Y) \sim P_{XY}$ and $(X', Y') \sim P'_{XY}$ with total variation distance ε, then

   $$I(X; Y) \leq I(X'; Y') + O(\varepsilon \log(mn/\varepsilon)).$$

   Hint: pbhcyr gur qvfgevohgvbaf, naq pbaqvgvba ba gur rirag M gung gurl ner rdhny.

   (e) Show how to distinguish between $I(X; Y) = 0$ and $I(X; Y) \geq \varepsilon$ with $O(\frac{1}{\varepsilon^2} n^{2/3} m^{2/3} \log^{O(1)}(mn/\varepsilon))$ samples.

   (f) [Optional] Improve the dependence on mn and/or ε.