Algorithmic Statistics Problem Set 4

Problem 1 Recall from lecture the definition of the two-community stochastic block model $SBM(d, \epsilon, n)$. In this problem we will consider the clustering problem for the stochastic block model under adversarial corruptions.

Design and analyze a polynomial-time algorithm with the following guarantee. Suppose G_0 is an n-vertex graph drawn from $SBM(d, \epsilon, n)$. Given any G' such that the symmetric difference between the sets of edges in G and G' has size $o(n)$, the algorithm finds $\hat{x}$ such that $\mathbb{E} |\langle \hat{x}, x \rangle| \geq \Omega(n)$, as long as $\epsilon^2 d \geq C$ for a sufficiently-large constant C, where x is the ground-truth clustering of the vertices of G.

You may rely on any polynomial-time primitives presented in lecture.

Problem 2 Moitra book (https://people.csail.mit.edu/moitra/docs/bookexv2.pdf), problem 3-2.

Problem 3 You observe i.i.d. samples (a_i, y_i) for $i = 1, \dots, n$, where

$$a_i \sim \mathcal{N}(0, I_d), \qquad y_i \;=\; \langle a_i, x \rangle^2 + \varepsilon_i,$$

and $\varepsilon_i \sim \mathcal{N}(0, 1)$ are independent. The hidden signal $x \in \mathbb{R}^d$ is an unknown *unit vector*. Your goal is to estimate x up to a global sign, i.e. recover a vector $\hat{x}$ such that $\min\{\|x - \hat{x}\|, \|x + \hat{x}\|\}$ is small. This problem is known as *phase retrieval*.

Design a *spectral estimator* for x.

(a) Construct a symmetric matrix statistic, built only from the samples $\{(a_i, y_i)\}_{i=1}^n$, whose spectrum in the population (i.e. its expectation) reveals the direction of x up to sign. You may find it useful to consider quadratic or bilinear functions of the samples.

(b) Compute the population version of your matrix and show that its top eigenvector spans the same one-dimensional subspace as x.

(c) Define the empirical version of your statistic using the n samples. Prove that, with high probability and for sufficiently large n, the top eigenvector of the empirical matrix yields an estimator $\hat{x}$ satisfying

$$\min\{\|\hat{x} - x\|, \|\hat{x} + x\|\} \;\leq\; C \, (d/n)^c,$$

for some absolute constants $C, c > 0$ (polylogarithmic factors in d, n are acceptable).