EE 290 Mathematics of Data Science - Homework 2

Rules:

1. This assignment is due at 11:59pm on **Tuesday, November 12, 2019**. No late submissions accepted.

2. Please submit the homework solution online via Gradescope. The entry code for gradescope is M4DNZ2. Hand-writing homework is allowed but required to convert to image / pdf file first. You can use any programming language. Submit your code as part of the homework. Tables and figures are preferred to help to illustrate your results. There are several sub-questions for each question. Make sure that it is easy for readers to figure out which sub-questions you are answering.

Problem A) Spiked Wigner Model.

Consider the following rank-one perturbation to a Gaussian random matrix:

$$W = \sqrt{\frac{\mu}{n}}\sigma\sigma^\top + Z \tag{0.1}$$

where $Z = (Z_{ij})$ is a symmetric matrix with $\{Z_{ij} : 1 \le i \le j \le n\}$ being iid $\mathcal{N}(0,1)$, and the membership vector σ is uniformly drawn from the set of all bisections, i.e. $\{\sigma \in \{\pm\}^n : \sum_i \sigma_i = 0\}$.

1. (Detection) Consider the hypothesis testing problem of testing $H_0 : W = Z (i.e. \mu = 0)$ versus $H_1 : W = \sqrt{\frac{\mu}{n}}\sigma\sigma^\top + Z$. Assume that μ is constant. Show that reliable detection (i.e. both Type-I and Type-II error probabilities vanish as $n \to \infty$) is impossible if $\mu < 1$. (Hint: compute the χ^2-divergence using the second moment method).

2. (Correlated recovery) We say an estimator $\hat{\sigma} = \hat{\sigma}(W)$ achieves correlated recovery, if it has an nontrivial overlap with the true partition, i.e. $\mathbb{E}|\langle \sigma, \hat{\sigma}\rangle| = \Omega(n)$ as $n \to \infty$. Instead of using conditional second-moment argument, we show that correlated recovery is impossssible if $\mu < 1$ by a reduction argument:

Suppose correlated recovery is possible. Let's construct a test statistic. Write $\sigma = [\sigma_1, \sigma_2]$, where $\sigma_1 \in \{\pm\}^{(1-\epsilon)n}$ and $\sigma_2 \in \{\pm\}^{\epsilon n}$ with appropriately chosen ϵ. Write W accordingly in a block form $W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$. Apply correlated recovery estimator on W_{11} to obtain $\hat{\sigma}_1$, and compute $y = W_{21}\hat{\sigma}_1$. Under the null, we expect the variance of each coordinate of y is roughly 1: under the alternative, thanks to the correlation between σ_1 and $\hat{\sigma}_1$, we expect the variance of each coordinate is strictly bigger than 1. Make this argument rigorous by analyzing the test statistics $\frac{1}{n}\|y\|_2^2$.

3. (Exact recovery: impossibility) We say an estimator $\hat{\sigma} = \hat{\sigma}(W)$ achieves exact recovery, if $\mathbb{P}(\sigma = \pm\hat{\sigma}) \to 1$ as $n \to \infty$. Show that exact recovery is impossible if $\mu = \sqrt{(2-\epsilon)\log(n)}$ for any fixed $\epsilon > 0$.

(Hint: show that even the maximum likelihood estimator fails in this case.)

4. (Exact recovery: SDP) Consider the following SDP relaxation:

$$\hat{X} = \mathsf{argmax}\{\langle W, X\rangle; X \succeq 0, X_{ii} = 1, \langle X, J\rangle = 0\} \tag{0.2}$$

where J is the all-one matrix. Show that exact recovery is achieved, i.e. $\hat{X} = \sigma\sigma^\top$ with probability tending to one, if $\mu = \sqrt{(2+\epsilon)\log n}$ for any fixed $\epsilon > 0$.

(Hint: do a direct analysis based on two facts (i) $\|Z\|_{op} = O(\sqrt{n})$ with high probability; (ii) the maximum of n iid standard norm is $\sqrt{(2+o(1))\log n}$ with high probability.)

This problem is from S&DS 684: Statistical inference on graphs by Yihong Wu.

Problem B) Experiments on Spiked Wigner Model.

Consider the spiked Wigner model defined in Problem A. Run spectral algorithm discussed in Lecture 9 and SDP in Problem A.4 separately, and plot the block error rate and bit error rate. The pairs of μ, n to run are specified as below:

1. Fix $n = 30$, range μ from 0.5 to 30 with increment 0.5. Plot the bit error rate, for SDP and specral methods, respectively with y-axis being the error rate and x-axis being μ. Repeat it for the block error rate.

Conduct the experiments for $n = 50, 100, 150$ separately and overlay curves corresponding to different n's on the same plot.

2. Fix $n = 30$, range μ from $\sqrt{0.5\log(n)}$ to $\sqrt{30\log(n)}$ with increment $\sqrt{0.5\log(n)}$. Plot the bit error rate, for SDP and specral methods, respectively with y-axis being the error rate and x-axis being $\frac{\mu}{\sqrt{\log(n)}}$. Repeat it for the block error rate.

Conduct the experiments for $n = 50, 100, 150$ separately and overlay curves corresponding to different n's on the same plot.

Repeat each μ, n pair 100 times and compare the empirical results with the prediction of theory. (In all there are 8 figures to plot, which are the combinations of spectral / SDP method, bit error rate / block error rate, two scales of μ.)

Note: Denote σ^j as the true membership vector in j-th experiments, and $\hat{\sigma}^j$ as the output of the algorithm in the j-th experiments. For a total of m experiments, the bit error rate is defined as $e_{bit} = \frac{1}{m}\sum_{j=1}^{m} \min_{\alpha\in\{\pm\}} \frac{n - \langle \alpha\sigma^j, \hat{\sigma}^j\rangle}{2n}$, the block error rate is defined as $e_{block} = \frac{1}{m}\sum_{j=1}^{m} \min_{\alpha\in\{\pm\}} \mathbb{1}(\alpha\sigma^j \neq \hat{\sigma}^j)$.

Problem C) Experiments on symmetric Stochastic Block Model.

Repeat the same procedure in Problem B for the symmetric Stochastic Block Model $SSBM(n, 2, \frac{\mu}{n}, 0)$ on the following μ, n pairs:

1. Fix $n = 30$, range μ from 0.5 to 30 with increment 0.5. Plot the bit error rate, for SDP and specral methods, respectively with y-axis being the error rate and x-axis being μ. Repeat it for the block error rate.

   Conduct the experiments for $n = 50, 100, 150$ separately and overlay curves corresponding to different n's on the same plot.

2. Fix $n = 30$, range μ from $0.5\log(n)$ to $10\log(n)$ with an interval $0.5\log(n)$. Plot the bit error rate, for SDP and specral methods, respectively with y-axis being the error rate and x-axis being $\frac{\mu}{\sqrt{\log(n)}}$. Repeat it for the block error rate.

   Conduct the experiments for $n = 50, 100, 150$ separately and overlay curves corresponding to different n's on the same plot.

(In all there are 8 figures to plot, which are the combinations of spectral / SDP method, bit error rate / block error rate, two scales of μ.)

Compare the empirical results with the results predicted by theory. What is different for the results of two models in Problem B and C?

Bonus problems (5pt each):

Problem D) Comparison of previous methods and widely used spectral clustering algorithms.

Consider the exact recovery setting (only plot the block error rate) and implement the following algorithms:

1. Fiedler's algorithm 2. Leighton-Rao relaxation 3. Arora-Rao-Vazirani relaxation.

Repeat the procedures in B and C, and compare the three algorithms with the previous spectral algorithm and SDP algorithm.

The algorithms can be found in https://people.eecs.berkeley.edu/~luca/books/expanders-2016.pdf.

Problem E) Non-backtracking Spectral methods.

Implement the non-backtracking spectral methods discussed in Lecture 12, repeat the procedures in Problem B and C, and compare with the previous empirical results.