EE378B Inference, Estimation, and Information Processing

Leave-one out arguments

Due on 2/17/2021 *Lecture 8-9 -*

Homework should be submitted via Gradescope, by Monday afternoon (unless Monday is a holiday): the code will be communicated by an announcement on Canvas.

For getting credit for the class, you are required to present solutions of some of these homeworks during the first 15 minutes of class starting on 1/20. Please, sign up for (at least) one slot, and be sure that your explanation lasts 15 minutes (or less). For these presentations, you are free to choose whatever format you prefer (slides, typed notes, handwriting, ...).

This week, the presentations will be:

- Wednesday: Questions (d), (e), (f).

Problem

Consider a graph/labels pairs from the two groups symmetric stochastic block model $(G, \boldsymbol{\sigma}) \sim \mathsf{G}(n, p_n, q_n)$, $G = (V = [n], E)$. We will set the edge probability to be $p_n = (\alpha \log n)/n$, $q_n = (\beta \log n)/n$, $\alpha > \beta > 0$. We showed in class that there exists a finite constant C_* such that, if $\sqrt{\alpha} - \sqrt{\beta} > C_*$, then SDP recovers with high probability the correct labels $\boldsymbol{\sigma}$ (up to a global flip). In this homework we will use a leave-one out technique to derive a lower bound.

> *Alert.* This homework was written from scratch and is somewhat challenging. Indeed a few years ago it would have been a research paper. Do not hesitate to reach out for question!

For the purpose of this homework, an estimator is a map (here $\mathcal{G}_n \cong \{0,1\}^{\binom{n}{2}}$ is the space of graphs over n vertices):

$$\hat{\boldsymbol{\sigma}} : \mathcal{G}_n \to \{+1, -1\}^n , \tag{1}$$
$$G \mapsto \hat{\boldsymbol{\sigma}}(G) . \tag{2}$$

We evaluate such an estimator using the error probability

$$\mathrm{P}_{\mathrm{err},n}(\hat{\boldsymbol{\sigma}}) := \mathbb{P}\big(\hat{\boldsymbol{\sigma}}(G) \notin \{\boldsymbol{\sigma}, -\boldsymbol{\sigma}\}\big) . \tag{3}$$

We will use the following result.

Lemma 1. *Consider $G_{1,m} \sim \mathsf{G}(m, p_m, q_m)$, $G_{2,m} \sim \mathsf{G}(m, \overline{p}_m, \overline{p}_m)$, where $\overline{p}_m = (p_m + q_m)/2$. If $m(p_m - q_m)^2/(p_m + q_m) \to 0$, then*

$$\lim_{m \to \infty} \big\| P_{1,m} - P_{2,m} \big\|_{\mathrm{TV}} = 0 . \tag{4}$$

(a) Consider a simpler problem in which the labels of vertices $\{1, \dots, n-m\}$ have been revealed, denote them by $\boldsymbol{\sigma}_1^{n-m}$, and we want an estimator for vertex n: $\hat{\boldsymbol{\sigma}}^* : \mathcal{G}_n \times \{+1,-1\}^{n-m} \to \{+1,-1\}^m$. Prove that, for any estimator $\hat{\boldsymbol{\sigma}}$,

$$\mathrm{P}_{\mathrm{err},n}(\hat{\boldsymbol{\sigma}}) \geq \inf_{\hat{\boldsymbol{\sigma}}^*} \mathbb{P}\big(\hat{\boldsymbol{\sigma}}^*(G, \boldsymbol{\sigma}_1^{n-m}) \neq \boldsymbol{\sigma}_{n-m+1}^n\big) . \tag{5}$$

(*b*) Consider the modified graph $G_* = (V = [n], E_*)$ such that, if $\min(i,j) \le n - m$, $(i,j) \in E$ if and only if $(i,j) \in E_*$ (and call these edges $E_{*,\le}$), and for $\min(i,j) > n - m$, independently,

$$\mathbb{P}\big((i,j) \in E_* | \boldsymbol{\sigma}, E_{*,\le}\big) = \overline{p}_n\,. \tag{6}$$

Prove that, for $m \le n^{1-\epsilon}$, $\epsilon > 0$:

$$\mathrm{P}_{\mathrm{err},n}(\hat{\boldsymbol{\sigma}}) \ge \inf_{\hat{\boldsymbol{\sigma}}^*} \mathbb{P}\big(\hat{\boldsymbol{\sigma}}^*(G_*, \boldsymbol{\sigma}_1^{n-m}) \ne \boldsymbol{\sigma}_{n-m+1}^n\big) - o_n(1)\,. \tag{7}$$

We will assume hereafter $m \le n^{1-\epsilon}$, $\epsilon > 0$.

(*c*) For each $i \in \{n-m+1, \dots, n\}$, let $N_s(i) := \{j \le n - m : (i,j) \in E_*, \sigma_j = s\}$, for $s \in \{+1, -1\}$. Prove that the infimum in the last display is achieved when $\hat{\boldsymbol{\sigma}}^* = (\hat{\sigma}^*_{n-m+1}, \dots, \hat{\sigma}^*_n)$, with $\hat{\sigma}^*_i$ a function uniquely of $(N_+(i), N_-(i))$.

(*d*) Deduce from the previous point that

$$\lim_{n\to\infty} \mathrm{P}_{\mathrm{err},n}(\hat{\boldsymbol{\sigma}}) = 0 \quad \Rightarrow \quad \lim_{n\to\infty} m \inf_{\hat{\sigma}^*_n} \mathbb{P}\big(\hat{\boldsymbol{\sigma}}^*_n(N_+(n), N_-(n)) \ne \sigma_n\big) = 0\,. \tag{8}$$

(*e*) *(Optional.)* Note that $N_+(n), N_-(n)$ are independent binomial random variables. Prove that there exists a constant $C_0 > 0$ such that:

$$\lim_{n\to\infty} \mathrm{P}_{\mathrm{err},n}(\hat{\boldsymbol{\sigma}}) = 0 \quad \Rightarrow \quad \sqrt{\alpha} - \sqrt{\beta} \ge C_0\,. \tag{9}$$

In other words, exact reconstruction is possible only if $\sqrt{\alpha} - \sqrt{\beta} \ge C_0$.

(*f*) *(Optional.)* What is the best value of the constant C_0 that you get from this argument?