EE378B Homework 5 Solution

Due to: Chen Cheng

1 Part (a)

We construct a special estimator $\hat{\boldsymbol{\sigma}}_0^\star$ such that

$$\hat{\boldsymbol{\sigma}}_0^\star(G, \boldsymbol{\sigma}_1^{n-m}) = \begin{cases} \hat{\boldsymbol{\sigma}}_{n-m+1}^n, & \hat{\boldsymbol{\sigma}}_1^{n-m} = \boldsymbol{\sigma}_1^{n-m} \\ -\hat{\boldsymbol{\sigma}}_{n-m+1}^n, & \hat{\boldsymbol{\sigma}}_1^{n-m} = -\boldsymbol{\sigma}_1^{n-m} \\ \mathbf{1}, & \text{else} \end{cases} \tag{1}$$

Note that $\hat{\boldsymbol{\sigma}}_0^\star$ is indeed a function of G and $\boldsymbol{\sigma}_1^{n-m}$ as $\hat{\boldsymbol{\sigma}}$ is a function of G. If $\hat{\boldsymbol{\sigma}}_0^\star(G, \boldsymbol{\sigma}_1^{n-m}) \neq \boldsymbol{\sigma}_{n-m+1}^n$, it implies either $\hat{\boldsymbol{\sigma}}_1^{n-m} \notin \left\{\pm\boldsymbol{\sigma}_1^{n-m}\right\}$ and $\boldsymbol{\sigma}_{n-m+1}^n \neq \mathbf{1}$; or $\hat{\boldsymbol{\sigma}} \notin \{\pm\boldsymbol{\sigma}\}$. Whichever the case, this suggests

$$\{\hat{\boldsymbol{\sigma}} \notin \{\pm\boldsymbol{\sigma}\}\} \supset \left\{\hat{\boldsymbol{\sigma}}_0^\star(G, \boldsymbol{\sigma}_1^{n-m}) \neq \boldsymbol{\sigma}_{n-m+1}^n\right\}, \tag{2}$$

and thus

$$\mathbb{P}_{\text{err},n}(\hat{\boldsymbol{\sigma}}) \geq \mathbb{P}(\hat{\boldsymbol{\sigma}}_0^\star(G, \boldsymbol{\sigma}_1^{n-m}) \neq \boldsymbol{\sigma}_{n-m+1}^n) \geq \inf_{\hat{\boldsymbol{\sigma}}^\star} \mathbb{P}(\hat{\boldsymbol{\sigma}}^\star(G, \boldsymbol{\sigma}_1^{n-m}) \neq \boldsymbol{\sigma}_{n-m+1}^n). \tag{3}$$

2 Part (b)

Let the subgraphs of G and $G_\star$ formed by the vertices $\{n-m+1, \cdots, n\}$ be $\bar{G}, \bar{G}_\star$. We can couple $\bar{G}$ and $\bar{G}_\star$ optimally, such that they are defined on the same probability space and

$$\mathbb{P}(G \neq G_\star) = \mathbb{P}(\bar{G} \neq \bar{G}_\star) = \|\bar{G} - \bar{G}_\star\|_{\text{TV}}. \tag{4}$$

The first equality follows from the construction of $G_\star$. Therefore by part (a) one would naturally have

$$\begin{aligned} \mathbb{P}_{\text{err},n}(\hat{\boldsymbol{\sigma}}) &\geq \inf_{\hat{\boldsymbol{\sigma}}^\star} \mathbb{P}(\hat{\boldsymbol{\sigma}}^\star(G, \boldsymbol{\sigma}_1^{n-m}) \neq \boldsymbol{\sigma}_{n-m+1}^n) \\ &= \inf_{\hat{\boldsymbol{\sigma}}^\star} \mathbb{P}(\hat{\boldsymbol{\sigma}}^\star(G, \boldsymbol{\sigma}_1^{n-m}) \neq \boldsymbol{\sigma}_{n-m+1}^n, G = G^\star) - \mathbb{P}(G \neq G^\star) \\ &= \inf_{\hat{\boldsymbol{\sigma}}^\star} \mathbb{P}(\hat{\boldsymbol{\sigma}}^\star(G^\star, \boldsymbol{\sigma}_1^{n-m}) \neq \boldsymbol{\sigma}_{n-m+1}^n, G = G^\star) - \mathbb{P}(G \neq G^\star) \\ &= \inf_{\hat{\boldsymbol{\sigma}}^\star} \mathbb{P}(\hat{\boldsymbol{\sigma}}^\star(G^\star, \boldsymbol{\sigma}_1^{n-m}) \neq \boldsymbol{\sigma}_{n-m+1}^n) - 2\|\bar{G} - \bar{G}_\star\|_{\text{TV}} \end{aligned} \tag{5}$$

Since $\bar{G}, \bar{G}^\star$ are random graphs on m vertices, by Lemma 1 we have the condition

$$\frac{m(p_n - q_n)^2}{p_n + q_n} = \frac{m(\alpha - \beta)^2 \log n}{n(\alpha + \beta)} \leq \frac{(\alpha - \beta)^2 \log n}{(\alpha + \beta) n^\epsilon} \to 0 \tag{6}$$

verified and thus $\|\bar{G} - \bar{G}_\star\|_{\text{TV}} = o_n(1)$.

3 Part (c)

The infimum is achieved by the Bayes estimator, i.e.,

$\hat{\boldsymbol{\sigma}}_{\text{Bayes}}^\star\left(G_\star, \boldsymbol{\sigma}_1^{n-m}\right)$

$$
\begin{aligned}
&= \operatorname*{argmax}_{\hat{\boldsymbol{\sigma}}^\star} \mathbb{P}\left(\boldsymbol{\sigma}_{n-m+1}^n = \hat{\boldsymbol{\sigma}}^\star \,\middle|\, G_\star, \boldsymbol{\sigma}_1^{n-m}\right) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{\sigma}}^\star} \mathbb{P}\left(G_\star \middle| \boldsymbol{\sigma}_{n-m+1}^n = \hat{\boldsymbol{\sigma}}^\star, \boldsymbol{\sigma}_1^{n-m}\right) \mathbb{P}(\boldsymbol{\sigma}_{n-m+1}^n = \hat{\boldsymbol{\sigma}}^\star) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{\sigma}}^\star} \left(\prod_{n-m+1\le i\le n, \hat{\sigma}_i^\star = 1} \left(\frac{p_n}{1-p_n}\right)^{N_+(i)} \left(\frac{q_n}{1-q_n}\right)^{N_-(i)} \right) \cdot \left(\prod_{n-m+1\le i\le n, \hat{\sigma}_i^\star = -1} \left(\frac{q_n}{1-q_n}\right)^{N_+(i)} \left(\frac{p_n}{1-p_n}\right)^{N_-(i)} \right) \\
&\quad \cdot \prod_{i=n-m+1}^{n} \left((1-p_n)^{\#\{\hat{\sigma}_i^\star \sigma_j = 1 : 1\le j\le n-m, n-m+1\le i\le n\}} \cdot (1-q_n)^{\#\{\hat{\sigma}_i^\star \sigma_j = -1 : 1\le j\le n-m, n-m+1\le i\le n\}} \right),
\end{aligned} \tag{7}
$$

where we use $\mathbb{P}(\boldsymbol{\sigma}_{n-m+1}^n = \hat{\boldsymbol{\sigma}}^\star) = 2^{-m}$. Define $U_l(\boldsymbol{x}) = \#\{x_i = 1 : 1 \le i \le l\}$ for any vector $\boldsymbol{x} \in \{\pm 1\}^l$, then

$$
\begin{aligned}
&\hat{\boldsymbol{\sigma}}_{\text{Bayes}}^\star\left(G_\star, \boldsymbol{\sigma}_1^{n-m}\right) \\
&= \operatorname*{argmax}_{\hat{\boldsymbol{\sigma}}^\star} \left(\prod_{n-m+1\le i\le n, \hat{\sigma}_i^\star = 1} \left(\frac{p_n}{1-p_n}\right)^{N_+(i)} \left(\frac{q_n}{1-q_n}\right)^{N_-(i)} (1-p_n)^{U_{n-m}(\boldsymbol{\sigma}_1^{n-m})} (1-q_n)^{n-m-U_{n-m}(\boldsymbol{\sigma}_1^{n-m})} \right) \\
&\quad \cdot \left(\prod_{n-m+1\le i\le n, \hat{\sigma}_i^\star = -1} \left(\frac{q_n}{1-q_n}\right)^{N_+(i)} \left(\frac{p_n}{1-p_n}\right)^{N_-(i)} (1-q_n)^{U_{n-m}(\boldsymbol{\sigma}_1^{n-m})} (1-p_n)^{n-m-U_{n-m}(\boldsymbol{\sigma}_1^{n-m})} \right)
\end{aligned}
$$

Then we can see conditional on $\boldsymbol{\sigma}_1^{n-m}$, $\hat{\sigma}_{\text{Bayes},i}^\star$ is a function uniquely of $(N_+(i), N_-(i))$.

4 Part (d)

We've seen the optimal $\hat{\sigma}_i^\star$ is only a function of $N_+(i), N_-(i)$ conditioned on $\boldsymbol{\sigma}_1^{n-m}$. Here we assume that the limit is taken conditioned on a sequence of fixed $\boldsymbol{\sigma}_1^{n-m}$ as $n \to \infty$. By part (b) we have

$$
\lim_{n\to\infty} \mathbb{P}(\hat{\boldsymbol{\sigma}}^\star(G_\star, \boldsymbol{\sigma}_1^{n-m}) \ne \boldsymbol{\sigma}_{n-m+1}^n) = 0. \tag{8}
$$

While $\boldsymbol{\sigma}_{n-m+1}^n$ is uniformly distributed in $\{\pm 1\}^m$, the vertices $\{n-m+1, \cdots, n\}$ on the random graph $G_\star$ are equivalent (i.e., the model is invariant under any permutation of the last m vertices). Hence

$$
\mathbb{P}(\hat{\sigma}_{n-m+1}^\star(N_-(n-m+1), N_+(n-m+1)) \ne \sigma_{n-m+1}) = \cdots = \mathbb{P}(\hat{\sigma}_n^\star(N_-(n), N_+(n)) \ne \sigma_n). \tag{9}
$$

Also notice that conditioned on $\boldsymbol{\sigma}_{n-m+1}^n$, $(N_-(i), N_+(i)), i = n-m+1, \cdots, n$ are independent, we have

$$
\begin{aligned}
&\mathbb{P}(\hat{\boldsymbol{\sigma}}^\star(G_\star, \boldsymbol{\sigma}_1^{n-m}) \ne \boldsymbol{\sigma}_{n-m+1}^n) \\
&= \mathbb{E}_{\boldsymbol{\sigma}_{n-m+1}^n}\left[\mathbb{P}(\hat{\boldsymbol{\sigma}}^\star(G_\star, \boldsymbol{\sigma}_1^{n-m}) \ne \boldsymbol{\sigma}_{n-m+1}^n | \boldsymbol{\sigma}_{n-m+1}^n)\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}_{n-m+1}^n}\left[1 - \prod_{i=n-m+1}^{n} \left(1 - \mathbb{P}(\hat{\sigma}_i^\star(N_-(i), N_+(i)) \ne \sigma_i | \boldsymbol{\sigma}_{n-m+1}^n)\right)\right] \\
&\ge \mathbb{E}_{\boldsymbol{\sigma}_{n-m+1}^n}\left[1 - \exp\left(-\sum_{i=n-m+1}^{n} \mathbb{P}(\hat{\sigma}_i^\star(N_-(i), N_+(i)) \ne \sigma_i | \boldsymbol{\sigma}_{n-m+1}^n)\right)\right] \to 0
\end{aligned} \tag{10}
$$

where we used $1 - x \le \mathrm{e}^{-x}$. Next we see that for any $\sigma_i = \sigma_j$, it must hold that

$$
\mathbb{P}(\hat{\sigma}_i^\star(N_-(i), N_+(i)) \ne \sigma_i | \boldsymbol{\sigma}_{n-m+1}^n) = \mathbb{P}(\hat{\sigma}_j^\star(N_-(j), N_+(j)) \ne \sigma_j | \boldsymbol{\sigma}_{n-m+1}^n) \tag{11}
$$

since the model is invariant if we exchange these two vertices. Therefore

$$
\sum_{i=n-m+1}^{n} \mathbb{P}(\hat{\sigma}_i^\star(N_-(i), N_+(i)) \ne \sigma_i | \boldsymbol{\sigma}_{n-m+1}^n)
$$

$$
\begin{aligned}
&= U_m(\boldsymbol{\sigma}_{n-m+1}^n)\mathbb{P}(\hat{\sigma}_n^\star(N_-(n), N_+(n)) \neq \sigma_n | U_m(\boldsymbol{\sigma}_{n-m+1}^n), \sigma_n = 1) \\
&\quad + (m - U_m(\boldsymbol{\sigma}_{n-m+1}^n))\mathbb{P}(\hat{\sigma}_n^\star(N_-(n), N_+(n)) \neq \sigma_n | U_m(\boldsymbol{\sigma}_{n-m+1}^n), \sigma_n = -1)
\end{aligned} \tag{12}
$$

and consequently

$$
\begin{aligned}
&U_m(\boldsymbol{\sigma}_{n-m+1}^n)\mathbb{P}(\hat{\sigma}_n^\star(N_-(n), N_+(n)) \neq \sigma_n | U_m(\boldsymbol{\sigma}_{n-m+1}^n), \sigma_n = 1) \\
&\quad + (m - U_m(\boldsymbol{\sigma}_{n-m+1}^n))\mathbb{P}(\hat{\sigma}_n^\star(N_-(n), N_+(n)) \neq \sigma_n | U_m(\boldsymbol{\sigma}_{n-m+1}^n), \sigma_n = -1) \\
&\leq U_m(\boldsymbol{\sigma}_{n-m+1}^n)\mathbb{P}(\hat{\sigma}_n^\star(N_-(n), N_+(n)) \neq \sigma_n | \sigma_n = 1)\frac{U_m(\boldsymbol{\sigma}_{n-m+1}^n)}{m} \\
&\quad + (m - U_m(\boldsymbol{\sigma}_{n-m+1}^n))\mathbb{P}(\hat{\sigma}_n^\star(N_-(n), N_+(n)) \neq \sigma_n | \sigma_n = -1)\frac{m - U_m(\boldsymbol{\sigma}_{n-m+1}^n)}{m} \\
&\leq 2m\left(\mathbb{P}(\hat{\sigma}_n^\star(N_-(n), N_+(n)) \neq \sigma_n | \sigma_n = 1)\mathbb{P}(\sigma_n = 1) + \mathbb{P}(\hat{\sigma}_n^\star(N_-(n), N_+(n)) \neq \sigma_n | \sigma_n = 1)\mathbb{P}(\sigma_n = -1)\right) \\
&\leq 2m\mathbb{P}(\hat{\sigma}_n^\star(N_-(n), N_+(n)) \neq \sigma_n)
\end{aligned} \tag{13}
$$

which yields

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}_{n-m+1}^n}\left[1 - \exp\left(-\sum_{i=n-m+1}^{n} \mathbb{P}(\hat{\sigma}_i^\star(N_-(i), N_+(i)) \neq \sigma_i | \boldsymbol{\sigma}_{n-m+1}^n)\right)\right] \\
&\geq 1 - \mathbb{E}_{U_m(\boldsymbol{\sigma}_{n-m+1}^n)}\left[\exp\left(-2m\mathbb{P}(\hat{\sigma}_n^\star(N_-(n), N_+(n)) \neq \sigma_n)\right)\right] - o_n(1) \\
&= 1 - \exp\left(-2m\mathbb{P}(\hat{\sigma}_n^\star(N_-(n), N_+(n)) \neq \sigma_n)\right) - o_n(1) \to 0.
\end{aligned} \tag{14}
$$

One must have

$$
m\mathbb{P}(\hat{\sigma}_n^\star(N_-(n), N_+(n)) \neq \sigma_n) \to 0. \tag{15}
$$

5 Part (e)

Without loss of generality, we assume $\sigma_n = 1$. Then by part (c) $\hat{\sigma}_n^\star(N_-(n), N_+(n)) \neq \sigma_n$ if and only if

$$
\left(\frac{p_n}{1-p_n} \cdot \frac{1-q_n}{q_n}\right)^{N_+(n)-N_-(n)} \left(\frac{1-p_n}{1-q_n}\right)^{2U_{n-m}(\boldsymbol{\sigma}_1^{n-m})-(n-m)} < 1, \tag{16}
$$

and $N_+(n) \sim \mathrm{Binom}(U_{n-m}(\boldsymbol{\sigma}_1^{n-m}), p_n)$, $N_-(n) \sim \mathrm{Binom}(n-m-U_{n-m}(\boldsymbol{\sigma}_1^{n-m}), q_n)$ are independent random variables. Equivalently,

$$
N_+(n) - N_-(n) < \left(2U_{n-m}(\boldsymbol{\sigma}_1^{n-m}) - (n-m)\right)\frac{\log\left(\frac{1-q_n}{1-p_n}\right)}{\log\left(\frac{p_n}{1-p_n} \cdot \frac{1-q_n}{q_n}\right)}. \tag{17}
$$

Set $Z_n := \frac{1}{\sqrt{n}}\left(U_{n-m}(\boldsymbol{\sigma}_1^{n-m}) - \frac{n-m}{2}\right)$, we have $Z_n \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1/4)$ by Slutsky and CLT. Then $\hat{\sigma}_n^\star(N_-(n), N_+(n)) \neq \sigma_n$ implies for sufficiently large n

$$
N_+(n) - N_-(n) < \frac{Z_n \log n}{\sqrt{n}} \cdot \frac{2}{\log\frac{\alpha}{\beta}}(\alpha - \beta)\left(1 + o_n(1)\right). \tag{18}
$$

Hence

$$
\begin{aligned}
&\mathbb{P}\left(N_+(n) - N_-(n) < \left(2U_{n-m}(\boldsymbol{\sigma}_1^{n-m}) - (n-m)\right)\frac{\log\left(\frac{1-q_n}{1-p_n}\right)}{\log\left(\frac{p_n}{1-p_n} \cdot \frac{1-q_n}{q_n}\right)}\right) \\
&\geq \mathbb{P}\left(N_+(n) - N_-(n) \leq -\frac{\log^2 n}{\sqrt{n}} \cdot \frac{2}{\log\frac{\alpha}{\beta}}(\alpha - \beta)\left(1 + o_n(1)\right)\right) - o_n(1)
\end{aligned}
$$

$$\geq \mathbb{P}\left(\mathrm{Binom}((n-m)/2, p_n) - \mathrm{Binom}((n-m)/2, p_n) \leq -\frac{\log^2 n}{\sqrt{n}} \cdot \frac{2}{\log \frac{\alpha}{\beta}} (\alpha - \beta)\,(1 + o_n(1)) - \frac{\log^2 n}{\sqrt{n}} \alpha\right) - o_n(1)$$

$$\geq \mathbb{P}\left(\mathrm{Binom}((n-m)/2, p_n) - \mathrm{Binom}((n-m)/2, p_n) \leq -\frac{\log n}{\log \log n}\right) - o_n(1), \tag{19}$$

while the final quantity is lower bounded by

$$\mathbb{P}\left(\mathrm{Binom}((n-m)/2, p_n) - \mathrm{Binom}((n-m)/2, p_n) \leq -\frac{\log n}{\log \log n}\right) \geq \exp\left(-\frac{1}{2}(\sqrt{\alpha} - \sqrt{\beta})^2 \log n + o(\log n)\right) \tag{20}$$

according to Lemma 4 from “Exact recovery in the stochastic block model" [Abbe, Bandeira, Hall ’15]. Hence by part (d) we have for all $\epsilon \in [0, 1)$,

$$m \exp\left(-(\sqrt{\alpha} - \sqrt{\beta})^2 \log n + o(\log n)\right) = \exp\left(-\frac{1}{2}(\sqrt{\alpha} - \sqrt{\beta})^2 \log n + o(\log n) + (1-\epsilon) \log n\right) \to 0. \tag{21}$$

This suggests $\frac{1}{2}(\sqrt{\alpha} - \sqrt{\beta})^2 \geq 1$, i.e.

$$\sqrt{\alpha} - \sqrt{\beta} \geq \sqrt{2}. \tag{22}$$